PARTIE 01
MÉTHODES DE DONNÉES ET APPLICATIONS

PARTIE 02
CADRE DE PROJET DE DONNÉES

PARTIE 01
MÉTHODES DE DONNÉES ET APPLICATIONS

PARTIE 02
CADRE DE PROJET DE DONNÉES

# ANALYSE DE DONNÉES ET SERVICES FINANCIERS NUMÉRIQUES

MANUEL

mastercard foundation

IFC International Finance Corporation
WORLD BANK GROUP

# REMERCIEMENTS

Le Partenariat pour l'Inclusion Financière d'IFC et de la Fondation MasterCard souhaite remercier les institutions qui ont participé aux études de cas de ce manuel pour leur généreux soutien : Airtel Ouganda, Commercial Bank of Africa, FINCA République Démocratique du Congo, First Access, Juntos, Lenddo, MicroCred, M-Kopa, Safaricom, Tiaxa, Tigo Ghana et Zoona. Ce manuel n'aurait pas été possible sans la participation de ces institutions.

IFC et la Fondation MasterCard souhaitent également remercier tout spécialement les auteurs Dean Caire, Leonardo Camiciotti, Soren Heitmann, Susie Lonie, Christian Racca, Minakshi Ramji et Qiuyan Xu, ainsi que les relecteurs et les contributeurs : Joshua Blumenstock, Sinja Buri, Tiphaine Crenn, Ruth Dueck-Mbeba, Nicolais Guevara, Raza Khan, Joseck Mudiri, Riadh Naouar, Rita Oulai, Laura Pippinato, Max Roussinov, Anca Bogdana Rusu, Matthew Saal et Aksinya Sorokina. Enfin, les auteurs souhaitent remercier tout spécialement Anna Koblanck et Lesley Denyes pour leur important travail d'édition.

Numéro ISBN : 978-0-620-76146-8
Première édition 2017

MANUEL

# ANALYSE DE DONNÉES ET SERVICES FINANCIERS NUMÉRIQUES

mastercard foundation

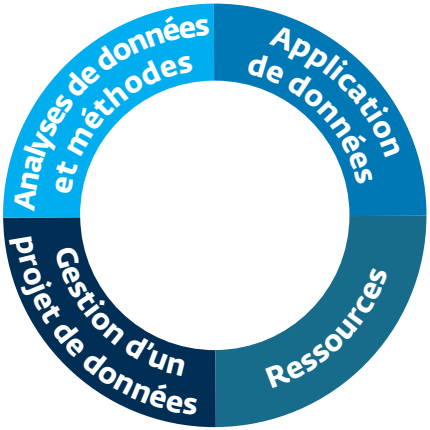
IFC International Finance Corporation
WORLD BANK GROUP

# Avant Propos

Il s'agit du troisième manuel sur les services financiers numériques (SFN) produit et publié par le Partenariat pour l'Inclusion Financière, une initiative conjointe d'IFC et de la Fondation MasterCard pour développer la microfinance et faire progresser les SFN en Afrique subsaharienne. Le premier manuel de la série, le *Manuel sur les canaux de distribution alternatifs et technologies*, fournit un guide complet des différentes technologies financières numériques, avec un accent particulier sur les composantes matérielles et logicielles d'un déploiement réussi. Le deuxième manuel, le *Manuel sur la gestion des risques en matière de canaux de distribution alternatifs*, est un guide sur les risques liés à l'argent mobile et aux services bancaires par agent et offre un cadre de gestion de ces risques. Ce manuel a pour but de fournir des orientations et un soutien utiles sur la façon d'appliquer l'analyse de données au développement et à l'amélioration de la qualité des services financiers.

Ce manuel est conçu pour tout type de prestataire de services financiers offrant ou ayant l'intention d'offrir des services financiers numériques. Les fournisseurs de SFN sont tous les types d'institutions telles que les institutions de microfinance, les banques, les opérateurs de réseaux mobiles, les entreprises de technologie financière et les prestataires de services de paiement. Les canaux, produits et processus à composante technologique génèrent des données extrêmement précieuses sur les interactions des clients ; dans le même temps, des liens avec des ensembles de données externes de plus en plus disponibles peuvent être activés. Le manuel offre un aperçu des concepts de base, identifie les tendances des utilisateurs sur le marché et illustre également une série d'applications pratiques et d'études de cas sur des prestataires de SFN qui tirent de leurs données internes ou externes des opportunités commerciales. Il offre également un cadre pour guider les projets de données des prestataires de SFN qui souhaitent tirer parti d'indications tirées de données afin de mieux répondre aux besoins des clients et d'améliorer les opérations, les services et les produits. Le manuel est conçu comme un premier contact avec les données et l'analyse de données, et suppose que le lecteur n'a aucune connaissance préalable de l'un ou l'autre. On considère cependant que le lecteur comprend les SFN et connait les produits, la fonction des agents, les aspects de la gestion opérationnelle et le rôle des technologies. Le manuel est structuré de la façon suivante :

**Introduction :** Présente le manuel et précise la plateforme et les définitions générales en matière de SFN et d'analyse de données.




**Partie 1 : Méthodes de données et applications**

**Chapitre 1.1 :** Discute de la science des données dans le contexte des SFN et donne un aperçu des types, sources, méthodologies et outils de données utilisés pour obtenir des indications découlant de données.

**Chapitre 1.2 :** Décrit comment appliquer l'analyse de données aux SFN. Ce chapitre résume les techniques utilisées pour dériver des indications sur les marchés à partir de données et décrit le rôle que les données peuvent jouer dans l'amélioration de la gestion opérationnelle des SFN. Le chapitre inclut de grands exemples classiques de la vie réelle et des études de cas sur les leçons tirées par les praticiens sur le terrain. Il se termine par un aperçu de la manière dont les praticiens peuvent utiliser des données pour développer des modèles de notation de risque de crédit fondés sur des algorithmes visant à favoriser l'inclusion financière.

**Partie 2 : Cadre de projet de données**

**Chapitre 2.1 :** Propose un cadre pour la mise en œuvre des projets de données et un guide étape par étape pour résoudre des problèmes commerciaux pratiques en appliquant ce cadre et ainsi tirer parti de la valeur des sources de données existantes et potentielles.

**Chapitre 2.2 :** Fournit un répertoire de sources de données et de ressources technologiques, ainsi qu'une liste d'indicateurs de performance pour évaluer des projets de données. Il inclut également un glossaire qui fournit une description des termes utilisés dans le manuel et la pratique du secteur.

**Conclusion :** Inclut des leçons tirées à ce jour de projets de données, en s'appuyant sur l'expérience d'IFC en Afrique subsaharienne dans le cadre du programme de Partenariat pour l'Inclusion Financière de la Fondation MasterCard.

# TABLE DES MATIÈRES

# ACRONYMES

| | |
|---|---|
| **AO** | Appel d'offres |
| **API** | Interfaces de programmation (Application Programming Interfaces) |
| **AQ** | Assurance qualité |
| **ARPU** | Revenu moyen par utilisateur (Average Revenue Per User) |
| **ARS** | Analyse des réseaux sociaux |
| **BD** | Base de données |
| **CBA** | Commercial Bank of Africa |
| **CBS** | Système bancaire central (Core Banking System) |
| **CDA** | Canal de distribution alternatif |
| **CDO** | Directeur des données (Chief Data Officer) |
| **CDR** | Enregistrements détaillés des appels (Call Detail Records) |
| **CGAP** | Groupe consultatif d'assistance aux pauvres |
| **COT** | Commission sur une transaction (Commission on Transaction) |
| **CRISP-DM** | Processus de norme interprofessionnelle pour l'exploration de données (Cross Industry Standard Process for Data Mining) |
| **CSV** | Valeurs séparées par des virgules (Comma-separated Values) |
| **DN** | Date de naissance |
| **ERC** | Essai randomisé contrôlé |
| **ETL** | Extraction - Transformation - Chargement (Extraction - Transformation - Loading) |
| **FSD** | Approfondissement du secteur financier (Financial Sector Deepening) |
| **FTC** | Commission fédérale du commerce (Federal Trade Commission) |
| **GAB** | Guichet automatique bancaire |
| **GPS** | Géopositionnement par satellite (Global Positioning System) |
| **GRC** | Gestion de la relation client |
| **GSM** | Système mondial de communications mobiles (Global System for Mobile Communications) |

| | |
|---|---|
| **GSMA** | Association du Système mondial de communications mobiles (Global System for Mobile Communications Association) |
| **IA** | Intelligence artificielle |
| **ICP** | Indicateur clé de performance |
| **ICR** | Indicateur clé de risque |
| **IF** | Institution financière |
| **IFC** | Société financière internationale (International Finance Corporation) |
| **IMF** | Institution de microfinance |
| **JSON** | Notation des objets en JavaScript (JavaScript Object Notation) |
| **KCB** | Kenya Commercial Bank |
| **KYC** | Obligation de s'informer sur le client (Know Your Customer) |
| **LBC** | Lutte contre le blanchiment de capitaux |
| **LFT** | Lutte contre le financement du terrorisme |
| **LOS** | Système de constitution de dossier de prêt (Loan Origination System) |
| **MLG** | Modèle linéaire généralisé |
| **MPME** | Micro, petites, et moyennes entreprises |
| **MVP** | Produit minimum viable (Minimum Viable Product) |
| **MVS** | Machine à vecteurs de support |
| **NDA** | Accord de non-divulgation (Non-Disclosure Agreement) |
| **OLA** | Accord au niveau opérationnel (Operating Level Agreement) |
| **ONU** | Nations Unies |
| **ORM** | Opérateur de réseau mobile |
| **P2P** | De personne à personne |
| **PAR** | Portefeuille à risque |

| | |
|---|---|
| **PBAX** | Autocommutateur privé (Private Branch Automatic Exchange) |
| **PDV** | Point de vente |
| **PI** | Propriété intellectuelle |
| **PIN** | Numéro d'identification personnel (Personal Identification Number) |
| **PME** | Petites et moyennes entreprises |
| **PNP** | Prêt non productif |
| **PSF** | Prestataire de services financiers |
| **PSP** | Prestataire de services de paiement |
| **RDC** | République Démocratique du Congo |
| **RVS** | Réseau à vecteurs de support |
| **SEA** | Suivi, Évaluation et Apprentissage |
| **SFN** | Services financiers numériques |
| **SIG** | Système d'information de gestion |
| **SIM** | Module d'identification de l'abonné (Subscriber Identity Module) |
| **SLA** | Accord de niveau de service (Service Level Agreements) |
| **SMS** | Service de messages courts (Short Message Service) |
| **SQL** | Langage de requête structurée (Structured Query Language) |
| **TCP** | Protocole de contrôle de transmission (Transmission Control Protocol) |
| **TIC** | Technologies de l'information et de la communication |
| **TLN** | Traitement du langage naturel |
| **TPS** | Transactions par seconde |
| **UE** | Union Européenne |
| **USSD** | Données de services supplémentaires non structurées (Unstructured Supplementary Service Data) |

# Note de Synthèse

« Laissez les données changer votre façon de voir les choses »
– Hans Rosling

La Société financière internationale (IFC) soutient les institutions qui cherchent à développer des services financiers numériques (SFN) pour développer l'inclusion financière et se consacre à de multiples projets sur un ensemble de marchés grâce à son portefeuille d'investissements et de projets de conseil. À partir de 2017, grâce à son travail avec la Fondation MasterCard et d'autres partenaires, IFC collabore avec les prestataires de SFN à travers l'Afrique subsaharienne pour développer l'inclusion financière par le biais de produits et de services numériques. Les interactions avec les clients ainsi qu'avec le secteur en général, dans la région et au-delà, ont fait apparaitre la nécessité d'un manuel sur la manière d'utiliser le domaine émergent de la science des données pour tirer de la valeur des données issues de ces réalisations. Bien que l'analyse des données offre aux prestataires de SFN une occasion de connaitre des détails précis sur leurs clients et d'utiliser ces connaissances pour offrir des services de meilleure qualité, de nombreux praticiens n'ont pas encore mis en œuvre d'approche systématique axée sur les données pour leurs opérations et organisations. Quelques exemples ont fait l'objet d'une grande attention en raison de leur réussite sur certains marchés, tels que l'intégration de données alternatives pour évaluer le risque de crédit de nouveaux types de clients. Cependant, le potentiel d'utilisation des données va au-delà d'un ou deux cas d'applications spécifiques. Le manque de connaissances, la pénurie de compétences et le malaise causé par une nouvelle approche sont des obstacles courants à l'application des indications provenant de données aux SFN. Ce manuel vise à donner un aperçu des opportunités qu'offrent les données en termes de stimulation de l'inclusion financière, ainsi que des mesures que les praticiens peuvent prendre pour commencer à adopter une approche axée sur les données dans leurs entreprises et à concevoir des projets fondés sur les données pour résoudre des problèmes commerciaux concrets.

Au cours de la dernière décennie, les SFN ont transformé l'offre faite à la clientèle et le modèle économique du secteur financier, en particulier dans les pays en développement. Un grand nombre de personnes à faible revenu, de microentrepreneurs, de petites entreprises et de populations rurales qui n'avaient jusque-là pas accès à des services financiers

formels sont maintenant numériquement bancarisés par le biais d'un ensemble d'anciens et de nouveaux prestataires de services financiers (PSF), notamment des fournisseurs non traditionnels tels que des opérateurs de réseaux mobiles (ORM) et des entreprises émergentes de technologie financière. Cela s'est avéré avoir un impact sur la qualité de vie, comme le montre l'exemple kenyan, où une étude menée par des chercheurs du Massachusetts Institute of Technology (MIT) a montré que l'introduction de services financiers à composante technologique pouvait contribuer à réduire la pauvreté.[1] L'étude estime que, depuis 2008, l'accès aux services d'argent mobile qui permettent aux utilisateurs de conserver et d'échanger de l'argent a augmenté les niveaux de consommation quotidienne par habitant de 194 000 personnes, soit environ deux pour cent des ménages kenyans, ceci ayant pour effet concret de les sortir de l'extrême pauvreté. L'impact le plus important a été ressenti par les ménages dirigés par des femmes, souvent considérés comme particulièrement marginalisés sur le plan économique. Il s'agit d'un bon argument en faveur d'une inclusion financière plus étendue et plus approfondie en Afrique subsaharienne et dans d'autres économies émergentes. Les données et analyses de données peuvent contribuer à atteindre cet objectif.

On estime qu'environ 2,5 quintillions d'octets de données sont produits chaque jour dans le monde.[2] Pour se faire une idée de cette quantité de données, cela représente plus de 10 milliards de DVD haute définition. La plupart de ces données sont récentes - 90 % des données existantes ont été créées au cours des deux dernières années.[3] La révolution des données numériques récente s'étend avec la même intensité dans les pays en développement et dans les pays développés. En 2016, il existait 7,8 milliards d'abonnements de téléphonie mobile dans le monde, dont 74 pour cent se situaient dans des pays en développement.[4] L'abondance de données devrait s'intensifier à l'avenir. À mesure que baissent les couts des smartphones, l'accès à l'Internet mobile devrait passer de 44 pour cent en 2015 à 60 pour cent en 2020. En Afrique subsaharienne, l'utilisation des smartphones devrait passer de 25 pour cent de toutes les connexions en 2015 à 50 pour cent d'ici 2020.[5] Les objets quotidiens sont de plus en plus conçus pour envoyer et recevoir des données, en se connectant et communiquant directement entre eux et via des interfaces utilisateur d'applications de smartphones, connues sous le nom d'Internet des objets.[6] Bien qu'il s'agisse d'un phénomène observé essentiellement dans les pays développés, il existe aussi des exemples issus du monde en développement. En Afrique de l'Est par exemple, il existe des appareils solaires qui produisent des informations sur l'utilisation de l'unité et les remboursements de SFN effectués par le propriétaire. Les données sont ensuite utilisées pour réaliser des évaluations de crédit instantanées qui peuvent au bout du compte générer de nouvelles activités. Pour les prestataires de SFN, les données peuvent être tirées d'un éventail croissant de sources : données transactionnelles, relevés des appels mobiles, enregistrements des centres d'appels, inscriptions des clients et des agents, modèles d'achat de temps de communication, informations de bureau de crédit, publications sur les réseaux sociaux, données géo spatiales et plus encore.

Ces sources émergentes de données ont la capacité d'avoir des répercussions positives sur l'inclusion financière. L'analyse peut améliorer les processus d'entreprise des institutions qui offrent des services aux ménages à faible revenu en leur permettant d'identifier de nouveaux clients et de s'adresser à eux de manière plus efficace. Ainsi, les données peuvent aider les institutions financières (IF) à toucher des personnes nouvelles et jusque-là exclues. Elles renforcent également l'inclusion financière puisque les clients existants utilisent de plus en plus de produits financiers. Dans le même temps, les décideurs politiques et les autres

[1] Suri and Jack, « The Long Run Poverty and Gender Impacts of Mobile Money, » *Science* Vol. 354, Numéro 6317 (2015): 1288-1292.

[2] « The 4 Vs of Big Data », IBM Big Data Hub, consulté le 3 avril 2017, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[3] « The 4 Vs of Big Data », IBM Big Data Hub, consulté le 3 avril 2017, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[4] « The Mobile Economy 2017', GSMA Intelligence

[5] « Global Mobile Trends, » GSMA Intelligence

[6] *Internet des objets*. Dans Wikipedia, l'encyclopédie libre, consulté le 3 avril 2017, https://fr.wikipedia.org/wiki/Internet_des_objets

parties prenantes publiques peuvent maintenant avoir une vision détaillée de l'inclusion financière en examinant l'accès, l'utilisation et d'autres tendances. Ces données concrètes peuvent jouer un rôle dans l'élaboration de futures politiques et stratégies visant à améliorer l'inclusion financière.

La plus grande disponibilité des données représente des défis et des opportunités. Le défi majeur consiste à tirer parti de l'utilité des données tout en respectant la vie privée des personnes. Une grande part des données récemment disponibles sont produites passivement suite à nos interactions avec des services numériques tels que les téléphones mobiles, les recherches sur Internet, les achats en ligne et les transactions stockées électroniquement. Les caractéristiques des individus peuvent être déduites à partir d'algorithmes complexes qui utilisent ces données, tout cela grâce aux progrès en matière de capacité analytique. Ainsi, la vie privée est d'autant plus mise en péril que les générateurs de données primaires n'ont pas conscience des données qu'ils génèrent et de la manière dont elles peuvent être utilisées. En tant que tel, les entreprises et les parties prenantes du secteur public doivent mettre en place les garanties appropriées pour protéger la vie privée. Il doit exister des politiques et des cadres juridiques clairs, tant au niveau national qu'international, qui protègent les producteurs de données contre les attaques par les pirates et les exigences des gouvernements, tout en stimulant l'innovation dans le domaine de l'utilisation des données visant à améliorer les produits et services. Au niveau institutionnel, il doit exister des politiques claires régissant le consentement préalable et l'option de refus de l'utilisation des données, de l'exploration de données, de la réutilisation des données par des tiers, de leur transfert et de leur diffusion.

L'utilisation des données est pertinente pour l'ensemble du cycle de vie d'un client afin de mieux comprendre ses besoins et ses préférences. Il existe trois grandes applications quant aux données dans le domaine des SFN : l'obtention d'indications sur le marché, l'amélioration de la gestion opérationnelle et la notation de risque de crédit. Le manuel fait appel à de nombreuses études de cas afin de montrer comment les praticiens utilisent l'analyse de données. Il est intéressant de noter que l'univers des données est en expansion permanente et que les capacités analytiques s'améliorent également à mesure que progresse la capacité technologique. Ainsi, le potentiel d'utilisation des données dépasse largement les applications décrites dans ce manuel.

Le développement d'indications sur le marché fondés sur les données est essentiel au développement d'une entreprise orientée client. La compréhension des marchés et des clients à un niveau de grande précision permettra aux praticiens d'améliorer les services offerts aux clients et de répondre à leurs besoins les plus importants, générant ainsi une valeur économique. Une entreprise orientée client comprend les besoins et les désirs des clients, en veillant à ce que les processus internes et les processus qui touchent directement la clientèle, les initiatives en matière de marketing et la stratégie de produit reposent sur une science des données qui favorise la fidélisation des clients. Du point de vue des opérations, les données jouent un rôle important dans l'automatisation des processus et la prise de décision, ce qui permet aux institutions de devenir évolutives de façon rapide et efficace. Ici, les données jouent également un rôle important dans le suivi des performances et la génération d'indications sur la façon dont elles peuvent être améliorées. Enfin, l'utilisation répandue de l'Internet et du téléphone mobile est une source de nouvelles données qui permettent aux prestataires de SFN de réaliser une évaluation des risques plus précise des personnes jusque-là exclues qui n'ont pas d'antécédents financiers formels pour appuyer leurs demandes de prêt.

Le manuel décrit les étapes par lesquelles les praticiens peuvent passer afin de comprendre les éléments essentiels requis pour concevoir un projet de données et le mettre en œuvre au sein de leurs propres institutions. Deux outils sont présentés pour guider les chefs de projet à travers ces étapes : L'Anneau des données et son

complément, la Matrice de l'Anneau des données. L'Anneau des données est une liste de contrôle visuelle, dont la forme circulaire a pour centre le « cœur » de tout projet de données en tant qu'objectif stratégique de l'entreprise. Le processus de définition des objectifs est discuté, suivi d'une description des catégories de ressources fondamentales et des structures de conception nécessaires à la mise en œuvre du projet. Ces éléments incluent des ressources directes, telles que les données elles-mêmes, les outils logiciels et le matériel de traitement et de stockage ; ainsi que des ressources indirectes, notamment les compétences, l'expertise dans le domaine et les ressources humaines nécessaires à l'exécution. Cette section décrit également comment ces ressources sont utilisées lors de l'exécution du projet pour affiner les résultats et fournir de la valeur selon une stratégie de mise en œuvre définie.

L'outil complémentaire intègre ces éléments de conception structurelle dans une Matrice, un espace où les chefs de projet peuvent formuler et concevoir les ressources et définitions clés de façon organisée et interconnectée. Les outils permettent de définir les relations interconnectées entre les structures de conception de projet, afin de visualiser la manière dont les éléments sont liés et d'identifier les éventuelles lacunes ou le domaine dans lequel les exigences en ressources nécessitent un ajustement. L'approche de la Matrice sert également d'outil de communication, en fournissant un schéma de conception de projet de haut niveau sur une seule feuille de papier qui peut être mise à jour et discutée tout au long de la mise en œuvre du projet.

Enfin, des tableaux de ressources sont fournis. Le répertoire de données dresse la liste des principales sources de données disponibles pour les praticiens des SFN et un bref aperçu de leur application potentielle à un projet de données. La base de données technologique répertorie les outils essentiels dans le secteur de la science des données et des produits commerciaux de premier plan pour la gestion, l'analyse, la visualisation et les rapports de tableaux de bord de données. Figure également une liste de paramètres pour évaluer les modèles de données qui seraient souvent abordés par des consultants externes ou des fournisseurs d'analyses. Des copies des outils de l'Anneau des données peuvent être téléchargées comme référence ou pour être utilisées.

Le manuel fait appel à de nombreuses études de cas afin d'illustrer les expériences d'un ensemble diversifié de prestataires de SFN lors de la mise en œuvre de projets de données au sein de leurs organisations. Alors que ces praticiens sont principalement basés en Afrique et offrent des SFN à leurs clients sous forme d'argent mobile ou de services bancaires par agent, cela ne veut pas dire que les indications issues des données ne peuvent pas être utilisées par tous types de PSF en utilisant différents modèles économiques. Le fil conducteur de tous ces cas est que les institutions peuvent systématiquement développer leurs capacités en matière de données en commençant par des étapes modestes. Devenir une organisation axée sur les données avec des activités compétitives basées sur les données est un parcours qui nécessite une vision et un engagement à long terme. Il peut être nécessaire de changer certains aspects de la culture organisationnelle et d'améliorer les capacités internes existantes. Il est important de noter que les institutions doivent veiller à ce que les processus par lesquels les données sont recueillies, stockées et analysées respectent la vie privée des individus.

Ce manuel vise à fournir des conseils et un soutien utiles aux prestataires de SFN pour développer l'inclusion financière et améliorer les performances institutionnelles. La science des données offre une opportunité unique aux prestataires de SFN de connaitre leurs clients, agents et commerçants ainsi que d'améliorer leurs processus internes opérationnels et de crédit en utilisant ces connaissances pour offrir des services de meilleure qualité. La science des données exige que les entreprises adoptent de nouvelles compétences et modes de pensée, ce qui peut leur être inconnu. Cependant, ces compétences peuvent être acquises et permettront aux praticiens des SFN d'optimiser à la fois les performances institutionnelles et l'inclusion financière.

# Introduction

Les personnes jusque-là non bancarisées sur les marchés émergents accèdent de plus en plus aux services financiers formels via des canaux numériques. La puissance informatique présente partout, la connectivité omniprésente, le stockage de données de masse et les technologies analytiques évoluées sont exploitées pour fournir des produits et services financiers personnalisés de manière plus efficace et directe à un ensemble croissant de clients ; collectivement, on appelle ces produits et services les Services Financiers Numériques (SFN). Les prestataires de SFN, c'est-à-dire les institutions qui tirent parti des SFN pour fournir des services financiers, comprennent un ensemble d'institutions diversifié notamment les PSF traditionnels, comme les banques et les Institutions de Microfinance (IMF), ainsi que les PSF émergents tels que les ORM, les entreprises de technologie financière et les Prestataires de Services de Paiement (PSP).

Les données sont un terme utilisé pour décrire des informations, des faits ou des statistiques qui ont été recueillis aux fins de tous types d'analyse ou pour servir de référence. Les données existent sous plusieurs formes, telles que les nombres, les images, les textes, l'audio et la vidéo. L'accès aux données est un atout concurrentiel. Cependant, il ne signifie rien sans la capacité de les interpréter et de les utiliser pour améliorer l'orientation vers le client ; en tirer des indications sur le marché et en extraire une valeur économique. Les analyses sont les outils qui permettent de combler l'écart entre les données et les idées. La science des données est le terme donné à l'analyse des données, qui est un processus créatif et exploratoire empruntant des compétences à de nombreuses disciplines, notamment les activités commerciales, les statistiques et l'informatique. Elle a été définie comme « un domaine englobant et multidimensionnel qui utilise les mathématiques, les statistiques et autres techniques évoluées pour trouver des modèles et des connaissances significatifs dans les données collectées ».[7] Les outils de veille économique traditionnels étaient de nature descriptive, alors que les analyses évoluées peuvent utiliser les données existantes pour prédire le comportement futur de la clientèle.

Le caractère interdisciplinaire de la science des données exige que tout projet de données soit réalisé par une équipe qui puisse compter sur différentes gammes de compétences. Elle requiert une contribution du côté technique. Toutefois, elle requiert également une participation de l'équipe commerciale. Comme le montre le graphe ci-dessus, la conversion

[7] « Analytics: What is it and why it matters?, » SAS, consulté le 3 avril 2017, https://www.sas.com/en_za/insights/analytics/what-is-analytics.html

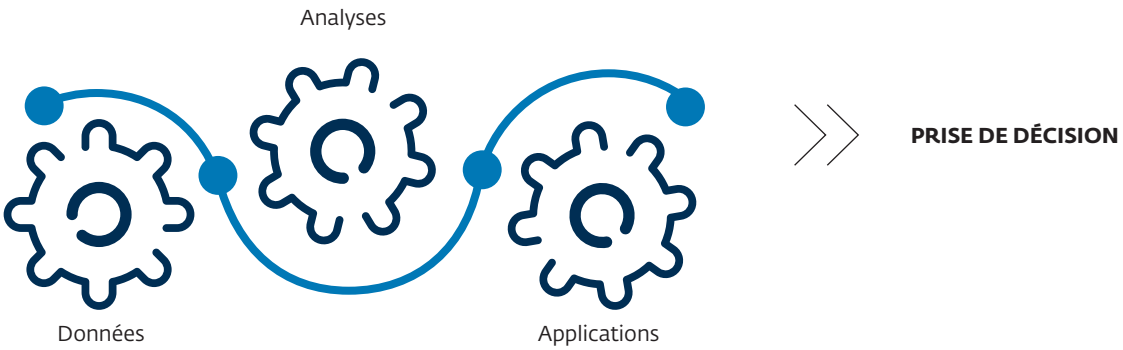



***Figure 1 :*** *La chaine de valeur des données*

des données en valeur pour les entreprises et en inclusion financière est un véritable parcours. La bonne compréhension des sources de données et des outils analytiques ne représente qu'une partie du processus. Ce processus ne saurait être complet sans une conceptualisation des données dans le cadre des strictes réalités commerciales du fournisseur de SFN. En outre, le fournisseur doit intégrer les indications tirées de l'analyse dans ses processus décisionnels.

Pour les prestataires de SFN, l'analyse de données est une opportunité unique. Les prestataires de SFN sont particulièrement actifs sur les marchés émergents et servent de plus en plus les clients qui peuvent ne pas avoir d'antécédents financiers formels tels que les antécédents de crédit. Il peut être particulièrement difficile de desservir ces nouveaux marchés. L'apprentissage des niveaux de préférences et de connaissances des nouveaux types de clients peut nécessiter d'y consacrer davantage de temps et de travail. À mesure que l'utilisation de la technologie numérique et des smartphones se développe sur les marchés émergents, les prestataires de SFN sont particulièrement bien placés pour tirer parti des données et des analyses afin de développer leur clientèle et fournir un service de meilleure qualité. L'analyse des données peut être utilisée à des fins spécifiques telles que la notation de risque de crédit, mais peut également être utilisée de façon plus générale pour améliorer l'efficacité opérationnelle. Quel que soit l'objectif, un fournisseur de SFN qui utilise les données a la capacité d'agir en fonction de données concrètes, plutôt que d'observation anecdotique ou en réaction à ce que font les concurrents sur le marché.

Dans le même temps, il est important de soulever la question de la protection du consommateur et de sa vie privée, car il arrive souvent que les producteurs de données primaires n'aient pas conscience que des données sont recueillies, analysées et utilisées à des fins spécifiques. Une mauvaise confidentialité des données peut entrainer une usurpation d'identité et des pratiques de prêt irresponsables. Dans le contexte du crédit numérique, des politiques sont nécessaires pour s'assurer que les personnes comprennent les implications du partage de leurs données avec les prestataires de SFN et pour s'assurer qu'ils ont accès aux mêmes données que celles auxquelles le fournisseur peut accéder. Afin d'élaborer des politiques, les parties prenantes tels que les prestataires, les décideurs politiques, les organismes de réglementation et d'autres devront se réunir pour discuter des préoccupations en matière de protection de la vie privée, des solutions possibles et de la marche à suivre. Pour ceux qui se consacrent à l'inclusion financière, les prestataires peuvent apprendre aux clients de façon proactive la manière dont les informations sont recueillies, utilisées et s'engager à ne recueillir que les données nécessaires, sans communiquer ces informations à des tiers.


# PARTIE 1

## *Méthodes relatives aux données et applications*

## *Chapitre 1.1 : Données, analyses et méthodes*

La complexité et la diversité croissantes des données produites ont conduit au développement de nouveaux outils et méthodes analytiques pour exploiter ces données et en tirer des indications. Le croisement des données et de leur ensemble d'outils analytiques correspond dans les grandes lignes au domaine émergent de la science des données. Pour les PSF numériques qui cherchent à appliquer des approches axées sur les données à leurs opérations, cette section fournit les connaissances de base pour identifier les ressources et interpréter les opportunités opérationnelles à travers le prisme des données, de la méthode scientifique et de la boite à outils d'analyse.

### Définition des données

Les *données* sont des échantillons de la réalité, enregistrés sous forme de mesures et stockées sous forme de valeurs. La façon dont les données sont classées, leur format, leur structure et leur source déterminent quels types d'outils peuvent être utilisés pour les analyser. Les données peuvent être quantitatives ou qualitatives. Les données quantitatives sont généralement des éléments d'information qui peuvent être mesurés objectivement, par exemple, des enregistrements de transactions. Les données qualitatives sont des éléments d'information sur des qualités et sont généralement plus subjectives. Les sources classiques de données qualitatives sont les entretiens, les observations ou les opinions, et ces types de données sont souvent utilisés pour estimer le sentiment ou le comportement des clients. On classe également les données par format. Au sens le plus immédiat, cela décrit la nature des données : nombre, image, texte, voix ou élément biométrique, par exemple. La numérisation des données est le processus consistant à prendre ces éléments de « réalité » mesurée ou observée et à les représenter sous forme de nombres que les ordinateurs comprennent. Le format des données numérisées décrit la façon dont une mesure donnée est codée numériquement. Il existe de nombreuses façons d'encoder l'information, mais toute information numérisée convertit des choses en nombres qui peuvent faire l'objet d'une analyse, ce qui sert de source d'indication potentielle de la valeur opérationnelle. La classification par format est essentielle car ce format décrit comment transformer l'information numérique en une représentation de la réalité et comment utiliser les bons outils de science des données pour obtenir des indications analytiques.

Pour qu'elles soient à disposition de l'analyse, les données doivent être stockées. Elles peuvent être stockées de façon structurée ou non structurée. Les données structurées ont un ensemble d'attributs et de relations définis lors du processus de conception de la base de données ; ces données suivent une organisation prédéterminée, également appelée un schéma. Dans une base de données structurée, tous les éléments de la base de données ont le même nombre d'attributs selon une séquence spécifique. Les données transactionnelles sont généralement structurées ; elles ont les mêmes caractéristiques et sont enregistrées de la même façon. Les données structurées sont plus faciles à interroger et à analyser. Les données non structurées ne sont pas organisées selon des schémas prédéterminés. Elles peuvent s'accroitre selon plusieurs formes, dans lesquelles des attributs fiables peuvent ou non exister. Cela les rend plus difficiles à analyser, mais c'est un avantage car plus de données sont générées rapidement à partir de nouvelles sources telles que les réseaux sociaux, les e-mails, les applications mobiles et les appareils personnels. Les données non structurées ont l'avantage de pouvoir être enregistrées telles quelles, sans avoir à vérifier si elles respectent les règles de l'organisation. Cela permet de les stocker de manière rapide et souple. Certaines données sont également considérées comme des données semi-structurées. Considérons un tweet de Twitter, par exemple, qui est limité à 140 caractères. Il s'agit d'une structure organisationnelle prédéterminée, et le service est programmé pour vérifier que le moindre tweet satisfait à cette exigence. Cependant, le contenu de ce qui est écrit dans un tweet n'est pas prédéfini et ne correspond à aucune règle ; cette combinaison pratiquement infinie de mots et de lettres illustre bien ce que sont les données non structurées. Dans l'ensemble, les tweets représentent donc des données semi-structurées.

Les données sont également classées par source. Les PSF ont tendance à catégoriser les sources de données en sources traditionnelles ou non traditionnelles ; les sources de données traditionnelles se réfèrent à des sources de données internes telles que les transactions tirées du système principal de gestion des comptes, les enquêtes auprès des clients, les formulaires d'inscription ou les informations démographiques. Les sources de données traditionnelles comprennent également des sources externes telles que les bureaux de crédit. Ce sont habituellement des données structurées. Les données non traditionnelles ou données alternatives, peuvent être structurées, semi-structurées ou non structurées, et ne sont pas toujours liées à l'utilisation des services financiers. Ces types de données sont par exemple les données d'utilisation de services de messages vocaux et courts (SMS) provenant des ORM, d'images satellites, de données géo spatiales, de données de réseaux sociaux, d'e-mails ou d'autres données indirectes. Ces types de sources de données sont de plus en plus utilisés par les PSF pour améliorer ou approfondir la compréhension de la clientèle, ou sont utilisés en association avec des données traditionnelles pour obtenir des indications opérationnelles. Par exemple, une IMF qui souhaite travailler en partenariat avec une coopérative laitière pour accorder des prêts à des producteurs laitiers pourrait utiliser la production de lait comme information indirecte sur les salaires afin d'évaluer la capacité d'octroi de crédits à des agriculteurs qui ne disposent pas d'antécédents de crédits formels.[8]

[8] Transcription de la session « Deploying Data to Understand Clients Better », Symposium de la Fondation MasterCard sur l'inclusion financière 2016, consulté le 3 avril 2017 http://mastercardfdnsymposium.org/resources/

## 1.1_ANALYSES DE DONNÉES ET MÉTHODES

### *Que sont les mégadonnées ?*

*Mégadonnées* est le terme générique habituellement utilisé pour décrire la grande échelle et la nature sans précédent des données qui sont actuellement produites. Les Mégadonnées ont cinq caractéristiques. Les premiers spécialistes des données ont identifié les trois premières caractéristiques énumérées ci-dessous et se réfèrent toujours aujourd'hui aux « trois V ». En développement depuis lors, les caractéristiques des Mégadonnées sont aujourd'hui au nombre de cinq :

1. **Volume:** La quantité de données actuellement produite est en elle-même étourdissante. L'ancienneté de ces données est également de plus en plus réduite, ce qui signifie que la quantité de données de moins d'une minute est en augmentation permanente. On s'attend à ce que la quantité de données dans le monde soit multipliée par 44 entre 2009 et 2020.

2. **Vitesse:** Une grande part des données disponibles est produite et mise à disposition en temps réel. Chaque minute, 204 millions d'e-mails sont envoyés. En conséquence, ces données sont traitées et stockées à très grande vitesse.

3. **Variété:** L'ère du numérique a diversifié les types de données disponibles. Aujourd'hui, 80 % des données générées, sous forme d'images, de documents et de vidéos, ne sont pas structurées.

4. **Véracité:** La véracité signifie la crédibilité des données. Les gestionnaires d'entreprise doivent savoir que les données qu'ils utilisent dans le processus de prise de décision sont représentatives des besoins et des désirs de leurs clients. Il est donc important de s'assurer qu'un processus rigoureux et permanent de nettoyage des données est suivi.

5. **Complexité:** La combinaison des quatre attributs ci-dessus exige des processus analytiques complexes et évolués. Des processus analytiques évolués sont apparus pour traiter ces grands volumes de données.

## Sources de données

Cette section se concentre sur les sources d'information clés que les prestataires de SFN pourraient consulter afin d'obtenir des indications opérationnelles ou sur les marchés. Surtout, une source de données ne doit pas être prise de façon isolée ; la conjonction de multiples sources de données permet souvent d'acquérir une compréhension de plus en plus nuancée des réalités codées par les données. Le chapitre 2.2 sur la collecte et le stockage des données des SFN passe en revue les sources de données traditionnelles et alternatives les plus courantes qui sont à disposition des prestataires de SFN.

### Sources traditionnelles de données

Comme mentionné ci-dessus, de façon traditionnelle, les PSF ont obtenu des données tirées des dossiers des clients, des données transactionnelles et des études de marché primaires. Une grande partie des données pertinentes pour le crédit ont été stockées sous forme de documents (copies papier), et seules les données de base sur l'inscription des clients et les activités bancaires étaient conservées dans des bases de données centralisées. Le défi d'aujourd'hui pour les PSF est de s'assurer que ces types de données traditionnelles sont également stockés sous un format numérique qui facilite l'analyse des données. Cela peut nécessiter une modification de la façon dont les données sont recueillies ou l'utilisation d'une technologie qui convertit les données en format numérique. Bien que de nouvelles technologies soient disponibles pour numériser les données traditionnelles, la numérisation peut représenter une tâche trop importante pour les anciennes données.

#### Données sur les clients et les agents

Les praticiens recueillent une grande quantité d'informations sur leurs clients lors des processus d'inscription et de demande de prêt, à la fois pour des raisons commerciales et pour respecter la réglementation. De même, ils recueillent des informations sur leurs agents dans le cadre du processus de demande et lors des visites de suivi. Pour ces deux catégories, il peut s'agir de variables telles que le sexe, la localisation et le revenu. Certaines de ces données sont vérifiées par des documents officiels, alors que d'autres sont évoquées et saisies lors des entretiens. Dans le cas des emprunteurs, une grande partie de ces informations sur les clients est saisie numériquement dans un Système de constitution de dossier de prêt (LOS) ou un module de constitution de dossier dans le système bancaire central (CBS). Il est surprenant de constater que, souvent, ces informations ne sont toujours disponibles que sur papier ou dans des fichiers numérisés.

#### Tiers

Les bureaux de crédits et les registres sont d'excellentes sources de données objectives et vérifiables. Ils fournissent une vérification de la crédibilité de l'information communiquée par les demandeurs de prêts et peuvent souvent révéler des informations que le demandeur n'est pas enclin à divulguer. La plupart des rapports des bureaux de crédits et des registres publics peuvent maintenant être interrogés en ligne avec accès numérique aux données pertinentes. Il existe cependant une difficulté : tous les marchés émergents ne disposent pas d'une infrastructure d'évaluation du crédit qui fonctionne pleinement.

#### Étude de marché primaire

On a généralement recours à une étude de marché pour mieux comprendre les clients et les segments de marché, suivre les tendances du marché, développer des produits et rechercher les commentaires des clients. Cette étude peut être qualitative ou quantitative, et il peut être utile de comprendre pourquoi et comment les clients utilisent les produits. Les achats anonymes effectués par des enquêteurs représentent une méthode courante d'étude de marché pour vérifier si les agents offrent un bon service à la clientèle ; certains prestataires de SFN recherchent quant à eux des commentaires directs des clients par le biais d'enquêtes qui génèrent un Taux de Recommandation Net permettant d'estimer à quel point les clients sont prêts à recommander un produit ou un service.

#### Données provenant de centres d'appels

Les données provenant de centre d'appels sont une bonne source pour comprendre les problèmes auxquels les clients sont confrontés et quels sont leurs sentiments sur les produits et le service clients d'un prestataire. Les données provenant de centres d'appels peuvent être analysées en classant par catégories les types d'appels et les temps de résolution et en utilisant l'analyse des conversations pour examiner les journaux audio. Les données provenant de centre d'appels sont particulièrement utiles pour comprendre les problèmes auxquels les clients, agents ou commerçants sont confrontés concernant des produits ou une nouvelle technologie qui vient d'être lancée.

## 1.1_ANALYSES DE DONNÉES ET MÉTHODES


*Figure 2 : Formats de données*

### Bases de données transactionnelles

Les données transactionnelles offrent des informations sur les niveaux d'activité et les tendances d'utilisation des produits. De simples comparaisons de transactions en valeur ou en volume peuvent offrir des indications très différentes sur le comportement des consommateurs. Pour les institutions financières telles que les banques ou les IMF, les données sur utilisation des comptes bancaires par les clients (dépôts, débits et crédits) et d'autres services (cartes, prêts, paiements et assurance) sont normalement enregistrées dans le CBS. L'utilisation des comptes et des services bancaires permet une traçabilité objective des données qui peuvent être analysées pour trouver des modèles signalisant différents niveaux de capacité et de sophistication financières. Différents modèles d'utilisation peuvent également être le signe de l'existence de différents niveaux de risque. Pour traiter les demandes de prêt, les institutions financières peuvent exiger des documents de la part d'autres institutions telles que les bureaux de crédits, mais ces documents ont tendance à exister sur papier et sont difficiles à numériser.

### Sources de données alternatives

Comme nos communications et affaires s'effectuent de plus en plus via les téléphones mobiles, les tablettes et les ordinateurs, il existe davantage de sources de données numérisées pouvant donner une indication de la capacité financière et de la réputation des clients. Ces sources peuvent nous indiquer la manière dont les personnes passent leur temps et comment ils dépensent leur argent, où et avec qui ils le font.

### Historique détaillé des appels (CDR) des ORM

Grâce à leurs activités de base, les ORM ont accès aux CDR et aux coordonnées des antennes-relais de téléphonie mobile. Les ORM analysent les CDR pour mener des campagnes de marketing et des promotions ciblées et pour ajuster les prix, par exemple. Au minimum, un CDR comprend 1) les appels vocaux, le temps de conversation, l'utilisation de services de données et les données de SMS sur l'expéditeur, le destinataire, l'heure et la durée, et 2) le temps de communication, des informations sur le rechargement des forfaits de données, notamment le temps, la localisation et la valeur. De plus, ces informations peuvent correspondre à des signaux de l'antenne-relais de téléphonie mobile pour générer les lieux d'activité des clients. Les ORM qui offrent des services d'argent mobile ont accès à la fois aux données du CDR et à celles de la base de données transactionnelle de SFN, et lorsqu'elles sont combinées pour analyse, ces informations prédisent mieux l'activité et les usages des clients que les simples données démographiques. Sur certains marchés, les ORM et les PSF opèrent en partenariat pour tirer parti de la combinaison des données. Les rechargements de temps de communication, par exemple, peuvent être un bon indicateur du revenu discrétionnaire. Les clients qui utilisent leur temps de communication jusqu'à zéro et font régulièrement et souvent de petits rechargements sont susceptibles d'avoir un revenu discrétionnaire moindre que ceux qui rechargent moins souvent, mais pour des montants plus importants.

### Données transactionnelles assistées par agent

Les données du centre d'appels sont particulièrement utiles pour comprendre quels sont les localisations et les agents qui sont les plus actifs pour fournir des indications contribuant à améliorer les performances du réseau d'agents. Pour de nombreux prestataires de SFN, les agents représentent le contact direct avec le client, et le suivi du modèle d'utilisation et de l'activité des agents peuvent donner des indications sur les préférences des clients et les performances de l'agent. Ces informations peuvent être directement enregistrées à partir des téléphones mobiles, des appareils de points de vente (PDV) ou des ordinateurs du point de transaction. Elles peuvent également être indirectement associées, par le biais par exemple des formulaires d'inscription de l'agent, en tenant compte de la nécessité d'être fusionnées dans le pipeline de données transactionnelles pour qu'une analyse puisse être menée.

### Données géo spatiales

Les données géo spatiales correspondent aux données qui contiennent des informations de localisation telles que les coordonnées du système de géo-positionnement par satellite (GPS), les adresses, les villes et autres identifiants géographiques ou de proximité. Ces dernières années, des données géo spatiales très précises ont permis aux prestataires de SFN d'examiner et de croiser des facteurs liés à la demande tels que le niveau d'inclusion financière, la localisation des clients, les niveaux de pauvreté, l'utilisation des données de téléphonie et de données mobiles, avec des facteurs liés à l'offre tels que l'activité des agents, les caractéristiques rurales ou urbaines, la présence d'infrastructures, et autres éléments similaires. Cela peut donner des indications qui peuvent être utiles à des stratégies d'acquisition de clients et de marketing, le développement des agents ou des succursales, et une analyse de la concurrence ou du marché général. Les données géo spatiales peuvent donner des indications plus précises que les indicateurs socio-économiques habituels, qui ne sont généralement disponibles que sous forme agrégée.

### Profils de réseaux sociaux

De plus en plus, les marchés des clients potentiels et existants se développent en ligne et maintiennent une présence sur les sites de réseaux sociaux tels que Facebook, Twitter et LinkedIn. Les données de comportement en ligne peuvent fournir des informations sur les commentaires, les attitudes, les modes de vie, les objectifs des clients et la façon dont les services financiers peuvent jouer un rôle dans leur vie. Les données des réseaux de réseaux sociaux comprennent des données sur les liens sociaux, le trafic créé, et le comportement en ligne, notamment l'heure, le lieu, la fréquence et la séquence d'un site Web ou d'une série de sites Web. Les réseaux sociaux peuvent aussi être le signe du statut socioéconomique d'un individu. Par exemple, les personnes dont le profil LinkedIn a de nombreuses connexions peuvent, en moyenne, représenter un risque plus faible que celles qui n'en ont pas. Ce n'est pas parce que la création d'un compte LinkedIn indique en soi une capacité à payer ses dettes, mais plutôt parce que LinkedIn cible les diplômés et, en moyenne, les diplômés ont des salaires plus élevés que les personnes non diplômées. Les profils publics de réseaux sociaux peuvent également être utiles pour vérifier les coordonnées et les informations personnelles de base sur les clients. Les réseaux sociaux en tant que source de données ont cependant leurs limites. Les PSF ne peuvent généralement avoir accès qu'aux comptes de réseaux sociaux des clients qui donnent leur accord préalable, et il peut être difficile d'obtenir que suffisamment de clients donnent cet accord préalable afin de construire une base de données de taille suffisante pour que l'analyse soit significative. Certains clients peuvent également ne pas être actifs sur les réseaux sociaux, par choix ou selon certaines circonstances. Les données de profil, même lorsqu'elles sont disponibles, peuvent également être biaisées.

## Sources de données opérationnelles

De nombreux processus au sein de l'entreprise sont requis pour exécuter une opération de SFN, chaque service travaillant à la réalisation des tâches et à l'atteinte d'objectifs de performance tout en se basant sur des données provenant de multiples sources. Les sources possibles de données externes et internes sont illustrées dans la figure ci-dessous et énumérées plus en détail au chapitre 2.2. Chaque service à la fois génère et consomme des données dans tout cet écosystème. Voici certaines des sources de données les plus importantes :

### Données du système central

Le système central fournit la majeure partie des données. Le moteur transactionnel est responsable de la gestion du flux de travail des transactions et des interactions, en envoyant autant de données et de métadonnées précises que possible aux bases de données pertinentes. Cela comprend le mouvement des fonds ainsi que les frais et commissions, ainsi que toute règle métier sur les partages de commissions et la réglementation fiscale. Il doit également fournir des pistes entièrement vérifiables de flux de travail des activités non financières telles que les changements de Numéro d'identification personnel (PIN), les demandes de solde, les mini-relevés et les téléchargements de données, ainsi que des fonctions internes telles que les transferts de fonds entre comptes.

### Rapports de systèmes de veille économique

Lorsque des produits de SFN sont nouveaux et qu'il existe un volume relativement faible de données, il est courant pour les entreprises de créer des rapports personnalisés à partir de données brutes en utilisant des outils simples comme Excel. À mesure que l'entreprise et les données se développent, et que l'analyse nécessaire devient plus complexe, cela devient vite ingérable. La plupart des grands systèmes de SFN mettent en place une banque de données qui utilise des systèmes de veille économique pour exploiter de nombreuses sources de données, qui fournissent des rapports de base et offrent la possibilité de personnaliser.

### Historiques techniques

Les historiques techniques constituent une abondante source de données. De plus en plus de fournisseurs de SFN évolués utilisent de manière proactive des tableaux de bord pour veiller en permanence à la santé du système et assurer une détection précoce des défaillances. Il est également courant de voir des moniteurs et alertes sur les performances intégrées au système de surveillance et qui peuvent fournir de précieuses informations. Les prestataires qui n'accèdent à ces données que lorsqu'une analyse de problème spécifique est nécessaire, se privent de données disponibles et utiles.

## Données internes périphériques

### Données d'autocommutateur privé (PBAX)

Le PBAX contrôle les appels entrants dans un centre d'appels, et peut fournir des données sur le volume des appels entrants, le nombre d'appels interrompus avant l'obtention d'une réponse et le temps consacré aux appels. Ces données sont essentielles à une planification efficace des modèles et de la taille des variations, ainsi qu'à la mesure et à l'amélioration de la performance générale de l'équipe.

### Systèmes de gestion des incidents

Le système de gestion des incidents suit le processus de résolution des problèmes de l'activité, et offre une mine d'informations, allant des types de problèmes qui se produisent aux durées de résolution des problèmes.

## Confidentialité des données et protection des consommateurs

Les nouvelles méthodologies d'analyse et de collecte des données soulèvent plusieurs questions relatives aux droits de confidentialité des clients et à la protection des consommateurs. Tout d'abord, comme indiqué plus haut, la plupart des données sont produites et recueillies de façon passive, c'est-à-dire sans que le producteur des données en ait conscience. Parfois, ces données sont partagées avec des tiers à l'insu du producteur de données. Cela peut avoir des conséquences négatives sur la capacité de l'individu à obtenir des prêts ou des assurances. Le problème est aggravé lorsque la personne n'a pas connaissance de ces informations négatives ou n'a pas recours à une contestation des informations négatives. Il n'existe aucune politique standard de consentement préalable concernant le partage des données. Certains prestataires de SFN ayant des applications installées sur les téléphones mobiles de leurs clients peuvent être en mesure d'obtenir des informations sur l'utilisation d'Internet du client et d'autres données, notamment les messages SMS, les contacts et les données de localisation, entre autres.

Étant donnée la diversité des prestataires de SFN, tous les prestataires ne relèvent pas du même régime de surveillance, ce qui conduit à différentes politiques de confidentialité des données s'appliquant à chacun. Certaines des violations aux droits individuels du respect de la vie privée pourraient avoir des répercussions négatives en matière de réputation. Au Kenya, de nombreux prestataires de crédit numériques ont vu le jour pour répondre à la demande de crédit, mais ils opèrent en dehors de la compétence réglementaire de la Banque centrale.[9] Un de ces prestataires a inclus dans ses conditions générales le fait que le prestataire était libre d'afficher les noms des personnes défaillantes sur son site Web et de publier directement sur les pages de réseaux sociaux des personnes défaillantes. Dans des cas comme celui-ci, les clients peuvent ne pas être conscients du fait qu'ils acceptent de céder leurs droits au respect de la vie privée jusqu'à ce qu'il soit trop tard. Cela peut être particulièrement vrai dans les contextes de pays en développement où l'alphabétisation et la sensibilisation à ces questions sont faibles.

En particulier, même dans les pays où le consentement de l'utilisateur est courant, les consommateurs peuvent ne pas comprendre les autorisations qu'ils accordent. À titre d'exemple, les utilisateurs sur les marchés sophistiqués peuvent ne pas avoir conscience de toutes les applications de leur smartphone qui utilisent des données de localisation. Des études montrent que 80 pour cent des utilisateurs de téléphones mobiles s'inquiètent du partage de leurs informations personnelles lorsqu'ils utilisent l'Internet ou des applications mobiles.[10] Malgré tout, 82 pour cent des utilisateurs acceptent les avis de confidentialité sans les lire, car ils ont tendance à être trop long ou à utiliser des termes qui leur sont inconnus. En raison de


***Figure 3 :*** *Exemple de demande d'enregistrement et d'accès aux données d'historique de localisation des utilisateurs via l'application Google Maps*

préoccupations en matière de sécurité et de la volonté affichée des clients d'arrêter d'utiliser des applications qu'ils jugent trop intrusive ou qui offrent une faible sécurité, la plupart des applications offrent de nos jours des moyens simples de donner son consentement préalable ou d'utiliser leur option de retrait.

[9] Ombija and Chege, « Time to Take Data Privacy Concerns Seriously in Digital Lending, » *Blog du Groupe consultatif d'assistance aux plus pauvres*, 24 octobre 24 2016, consulté le 3 avril 2017, https://www.cgap.org/blog/time-take-data-privacy-concerns-seriously-digital-lending

[10] « Mobile Privacy: Consumer research insights and considerations for policymakers, » GSMA

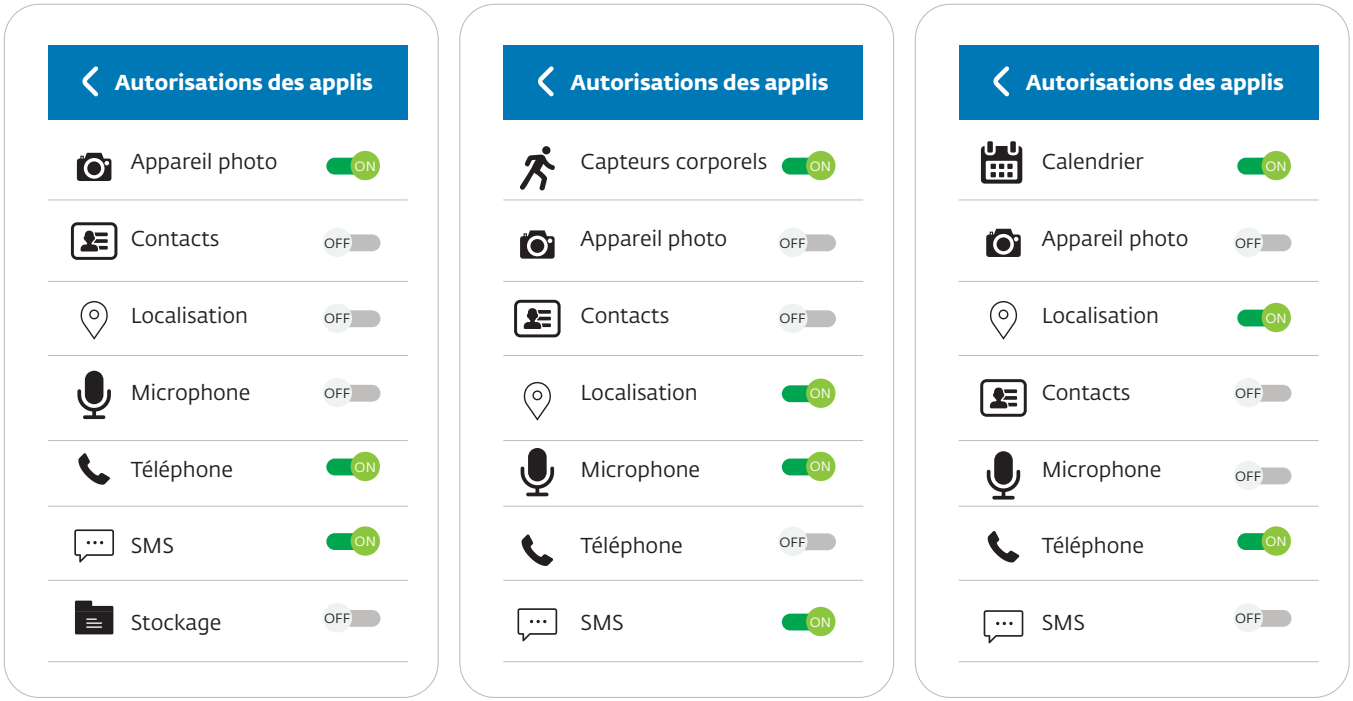



*Figure 4 : Exemples de paramètres d'autorisations d'application de smartphone*

Les lois sur la protection de la vie privée, lorsqu'elles existent, varient considérablement selon la juridiction et encore davantage selon leur degré d'application. Dans le contexte des marchés développés, dans l'Union Européenne (UE), le droit au respect de la vie privée et à la protection des données est fortement réglementé et activement appliqué,[11] alors qu'aux États-Unis il n'existe aucune loi fédérale d'ensemble sur la protection des données. L'UE a adopté une réglementation sur la protection des données en 2016 qui exige que tous les producteurs de données soient en mesure de recevoir en retour les informations qu'elles fournissent aux sociétés, puissent envoyer les informations à d'autres sociétés, et permettent aux sociétés d'échanger les informations entre elles lorsque cela est techniquement possible.[12] Ce genre de règlementation donne un certain pouvoir au consommateur tout en améliorant la concurrence, les consommateurs pouvant maintenant changer de prestataires en gardant leur historique de transactions intact. Aux États-Unis, la Federal Trade Commission (FTC) est l'organisme de réglementation chargé du domaine de la confidentialité des données. Toutefois, le Code des principes d'informations équitables de la FTC ne représente qu'un ensemble de recommandations pour

[11] La réglementation régissant la protection des données dans l'UE inclut la Directive 95/46 CE sur la protection des données et la Directive sur la protection de la vie privée dans le secteur des communications électroniques 02/58 CE (amendée par la Directive 2009/136)

[12] Réglementation (UE) 2016/679 du Parlement européen et du Conseil (2016), consultée le 3 avril 2017, http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

le maintien de pratiques de collecte de données respectant la vie privée et axées sur les consommateurs - il n'est pas exécutoire en vertu de la loi. En l'absence de toute règle de confidentialité fédérale générale, les États-Unis ont mis en place des lois et des réglementations fédérales et par État pour protéger la confidentialité des informations personnelles et la sécurité des données, à la fois de manière générale et de façon sectorielle, que toutes les entreprises concernées doivent respecter.

En ce qui concerne l'Afrique subsaharienne, le Ghana, l'Afrique du Sud et l'Ouganda semblent se démarquer comme les pays proposant les meilleures pratiques régionales. Ce qui distingue ces trois pays est le fait que la réglementation est guidée par un principe d'orientation client et, en tant que telle, la réglementation s'axe sur les principes suivants :

- Donner au consommateur le pouvoir de prendre des décisions pertinentes quant à l'utilisation de ses données personnelles, en particulier en ce qui concerne la prise de décision automatisée
- Stipuler des mécanismes clairs par lesquels le consommateur peut demander une indemnisation
- Accorder au client le « droit à l'oubli »

Les flux transfrontaliers de données constituent une question délicate, car ils peuvent en particulier affecter des sujets relevant de la sécurité nationale. La réglementation dans des pays comme l'Angola, l'Afrique du Sud et la Tanzanie stipule spécifiquement que les données ne peuvent être transférées que vers des pays où la loi prévoit des normes de protection des données personnelles en question identiques ou plus sévères. La Zambie va encore plus loin en interdisant tout transfert off-shore de données qui ne sont pas rendues anonymes.[13] À l'autre extrémité du spectre, le projet de Loi sur la protection des données du Kenya de 2016 a été sévèrement critiqué par les experts car elle n'incluait aucune disposition en matière de compétence extraterritoriale.[14]

Malgré tout, la confidentialité des données des clients est un nouveau domaine de la politique, et des pays tels que le Mozambique et le Zimbabwe se réfèrent encore à la Constitution pour interpréter les droits au respect la vie privée car ils ne disposent pas de projets de loi spécifiques. Dans ce contexte, les marchés émergents se tournent souvent vers les marchés plus établis et les organismes de réglementations pour trouver des indices sur la façon de traiter les problèmes à résoudre.

Étant donné ce contexte, tout en étant conscients des différences entre l'utilisation des technologies dans les pays émergents et sur les marchés développés, les Nations Unies (ONU) ont proposé certaines orientations générales en matière d'élaboration des politiques. L'ONU met l'accent sur la nécessité d'accélérer le développement et l'adoption de normes juridiques, techniques, géo spatiales et statistiques quant aux sujets suivants :

- Ouverture et échange de métadonnées
- Protection des droits de protection des données des personnes physiques[15]

Ainsi, à l'heure actuelle, aucune politique uniforme pour régir les questions de confidentialité des données n'existe. La première étape pour comprendre les implications en matière de confidentialité est d'assurer une discussion au niveau sectoriel impliquant les prestataires de SFN, les organismes de réglementations, les décideurs politiques, les autres parties prenantes du secteur public, les investisseurs et les institutions financières de développement, afin de concevoir des solutions et des normes. En même temps, dans le secteur de l'inclusion financière, les prestataires de SFN doivent reconnaitre que même si les données offrent une occasion d'améliorer le résultat net, elles mettent également en évidence une obligation d'ajouter de la valeur aux clients. Cela peut être réalisé en utilisant les données pour améliorer l'accès aux services financiers. Les prestataires de SFN peuvent tenter d'éduquer les personnes sur la façon dont leurs informations personnelles sont utilisées tout en ne recueillant d'informations que sur ce qui est nécessaire.

[13] « Global Data Privacy Directory, » Norton Rose Fulbright

[14] Francis Monyango, « Consumer Privacy and data protection in E-commerce in Kenya, » *Nairobi Business Monthly*, 1er avril 2016, consulté le 3 avril 2017, http://www.nairobibusinessmonthly.com/politics/consumer-privacy-and-data-protection-in-e-commerce-in-kenya/

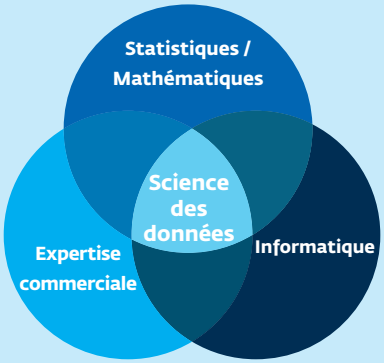
[15] « Un monde qui compte : mobiliser la révolution en matière de données pour le développement durable », Groupe consultatif d'experts indépendants du Secrétaire général des Nations unies sur la révolution des données pour le développement durable

## La science des données : Introduction

La science des données est l'utilisation interdisciplinaire de méthodes, processus et systèmes scientifiques pour extraire des indications et des connaissances de différentes formes de données afin de résoudre des problèmes spécifiques. Elle combine les sciences numériques telles que les statistiques et les mathématiques appliquées, avec l'informatique et l'expertise des entreprises et du secteur. Il s'agit d'une discipline exploratoire et créative, axée sur l'obtention de solutions novatrices à des problèmes complexes par une approche analytique. La science des données se réfère à la méthode scientifique d'analyse : les scientifiques des données se consacrent à la résolution de problèmes en définissant une hypothèse testable et en testant et affinant assidument cette hypothèse pour obtenir des résultats fiables et validés.


*Figure 5 : La méthode scientifique, le processus analytique qui est utilisé de façon similaire pour « la science des données »*

## La science des données




Le terme *scientifique des données* a été inventé en 2008 par DJ Patil et Jeff Hammerbacher pour décrire leurs fonctions chez LinkedIn et Facebook. Ils ont souligné que leurs rôles ne se limitaient pas à faire des calculs et à trouver des modèles dans ces calculs, mais aussi à appliquer un processus créatif et exploratoire visant à établir des liens entre tous ces modèles. « La science des données consiste à utiliser des données complexes pour raconter des histoires », dit Patil, ajoutant qu'il a puisé autant dans le monde du journalisme que dans celui de l'informatique. Pour cette raison, Patil et Hammerbacher ont envisagé un titre alternatif pour leur fonction : *Artiste des données*.

***Figure 6 :*** *La science des données, le croisement de plusieurs disciplines*

Afin de fournir une veille économique, toutes les analyses liées à des données doivent commencer par définir des objectifs commerciaux et identifier les bonnes questions commerciales, ou hypothèses. La méthode scientifique fournit des orientations utiles (voir la figure 5). Il convient de noter que ce n'est pas un processus linéaire. Au lieu de cela, il existe toujours un cycle d'apprentissage et une boucle de rétroaction pour assurer une amélioration progressive. Cela est essentiel pour obtenir des indications qui permettent une prise de décision fiable et fondée sur des données concrètes. Le chapitre 2.1 de ce manuel présente un processus étape par étape de mise en œuvre de projets de données pour les prestataires de SFN, en utilisant la méthodologie de l'Anneau des données.

La science des données facilite l'utilisation de nouvelles méthodes et technologies pour la veille économique, et des indications utiles peuvent être tirées d'ensemble de données, qu'ils soient grands ou petits, et que les données soient traditionnelles ou alternatives. Des ordinateurs plus rapides et des algorithmes complexes augmentent les possibilités d'analyse, mais ne remplacent ni n'écartent les outils et les approches à l'épreuve du temps pour tirer des indications des données visant à résoudre des problèmes commerciaux. Au contraire, il est important de comprendre les forces que les différents outils offrent et de les augmenter de manière appropriée pour obtenir les résultats escomptés en temps voulu et de manière rentable.

La figure 7 donne une description de haut niveau des méthodes d'analyse en veille économique, classées selon leur utilisation opérationnelle et leur complexité relative. De nombreuses catégories et leurs techniques et mises en œuvre associées se chevauchent, mais il est toujours utile de les diviser en quatre principaux cas d'utilisation : *descriptive, diagnostique, prédictive et prescriptives*. Les méthodologies les moins complexes sont souvent de nature descriptive ; elles fournissent une description historique de la performance institutionnelle, des chiffres agrégés et des statistiques synthétiques. Elles sont également moins susceptibles d'offrir un avantage concurrentiel, mais sont néanmoins essentielles pour le suivi des performances opérationnelles et de la conformité réglementaire. À l'opposé, les analyses les plus innovantes et complexes sont prescriptives, optimisées pour la prise de décision et offrent des indications sur les attentes futures. Cette progression contribue également à classer les produits livrables et la stratégie de mise en œuvre d'un projet de données, sujet abordé au chapitre 2.1.

## Cadre analytique de science des données pour la veille économique

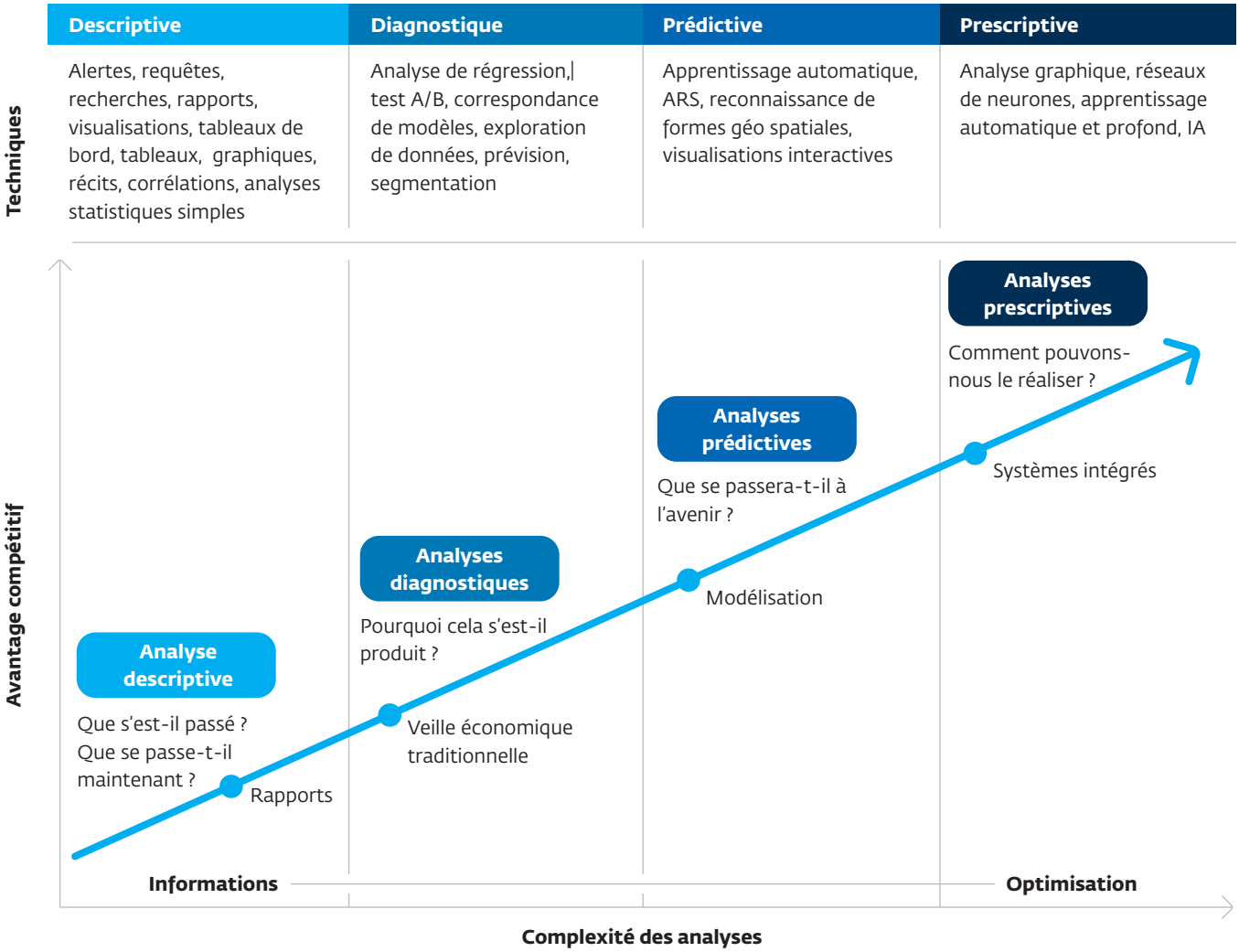



*Figure 7 : Les quatre catégories d'analyse commerciale*

# Méthodes

Les cas d'utilisation d'analyse décrits en figure 7 permettent de déterminer la méthode, le moment, le cout et la complexité des projets de données. Les méthodes suivantes sont généralement incluses dans la boite à outils du scientifique des données, et contribuent à adapter des méthodes générales à des fins d'analyse. Ces méthodes sont particulièrement pertinentes pour des discussions avec des consultants externes ou des fournisseurs de solutions afin d'aider à encadrer ce qu'ils fournissent ou pour évaluer une proposition.

## Analyse descriptive

L'analyse descriptive offre des rapports agrégés de haut niveau sur des historiques et répond aux questions sur ce qui s'est passé. Les Indicateurs clés de performance (ICP) se trouvent également dans cette catégorie.

- **Statistiques descriptives :** également connues sous le nom des statistiques synthétiques, les statistiques descriptives se composent de moyennes, d'additions, de décomptes et d'agrégations. Les statistiques de corrélation qui montrent des relations entre les variables contribuent également à décrire les données.
- **Présentation en tableaux :** Le processus d'agencement des données sous forme de tableau est appelé présentation en tableaux. Les présentations sous forme de tableaux croisés synthétisent les données issues d'une ou plusieurs sources en un format concis pour l'analyse ou la création de rapports, souvent en agrégeant des valeurs. Il s'agit d'une méthode pour segmenter, qui permet de présenter en tableaux les sommes agrégées selon le genre ou la localisation, par exemple, ou d'autres segments intéressants. Excel utilise le terme « tableau croisé dynamique » pour décrire ce type d'analyse.

## Analyses diagnostiques

Trouver les moteurs clés ou comprendre l'évolution de modèles de données constitue une analyse diagnostique. Il s'agit de se demander pourquoi quelque chose est arrivé ; par exemple, se demander pourquoi les modèles de transaction ont changé pour déterminer non seulement s'il existe une corrélation, mais aussi une causalité. L'analyse diagnostique nécessite généralement des méthodes et des protocoles de recherche plus sophistiqués, tel que décrit ci-dessous.

- **Test A/B :** Il s'agit d'une méthode statistique où deux ou plusieurs variantes d'une expérience sont présentées aux utilisateurs au hasard pour déterminer celle qui fonctionne le mieux pour un objectif de conversion donné. Le test A/B permet aux entreprises de tester deux scénarios différents et de comparer les résultats. Il s'agit d'une méthode très utile pour identifier de meilleures stratégies de promotion ou de marketing entre différentes options testées.
- **Régression :** La régression statistique est l'un des types de modélisation les plus élémentaires, et est très puissante. Elle permet une analyse à plusieurs variables pour estimer les relations entre une variable dépendante, habituellement un paramètre d'intérêt commercial, et un ensemble de variables indépendantes avec lesquelles il est en corrélation. L'identification de variables[16] statistiquement significatives peut orienter la stratégie, recentrer les objectifs et estimer les résultats.
- **Segmentation :** La segmentation est une méthode de classification de groupes en sous-groupes en fonction de critères de comportements ou de caractéristiques définis. La segmentation peut aider à identifier les catégories de clients démographiques ou d'utilisation des produits, avec des seuils quantifiés et statistiquement significatifs. Elle est souvent utilisée conjointement avec l'analyse de régression ou des techniques de modélisation plus sophistiquées pour prédire à quel segment un client potentiel non encore identifié pourrait appartenir.
- **Analyses géo spatiales :** Cette méthode groupe des données en fonction de leur localisation sur une carte, ou en lien avec la localisation et la proximité. Elle peut aussi contribuer à identifier des segments de clientèle et des comportements, tels que le lieu d'origine et de destination des envois d'argent, ou les agences que les clients ont tendance à visiter. Combinée avec des techniques plus évoluées, elle peut également permettre à des services fondés sur la localisation de contacter de manière proactive les clients qui sont à proximité de personnes ou de lieux d'intérêt.

[16] *Statistiquement significatif* s'emploie lorsqu'il est probable qu'une relation entre deux ou plusieurs variables soit causée par quelque chose d'autre que le hasard

## Analyses prédictives

Les prévisions permettent une prise de décisions tournées vers l'avenir et des stratégies fondées sur les données. Du point de vue de la science des données, il s'agit sans doute de la catégorie de méthode la plus centrale, car des algorithmes complexes et des calculs puissants sont souvent utilisés pour faire fonctionner ces modèles. Du point de vue commercial, les modèles prédictifs peuvent aboutir sur une meilleure efficacité opérationnelle en identifiant les segments de clients à fortes propensions et en étendant la portée à moindre cout via des campagnes de marketing ciblées. Ils peuvent également contribuer à améliorer l'assistance à la clientèle en anticipant de façon proactive les besoins en termes de services.

- **Apprentissage automatique :** Il s'agit d'un champ d'étude qui crée des algorithmes pour apprendre à partir de données et faire des prédictions sur ces dernières. En particulier, cette méthode permet un processus d'analyse qui identifie des tendances dans les données sans instruction explicite de l'analyste, et permet des méthodes de modélisation pour identifier des variables intéressantes et des facteurs clés de modèles même moins intuitifs. Il s'agit d'une technique plutôt qu'une méthode en elle-même. Les approches fondées sur l'apprentissage automatique sont classées selon les termes « apprentissage supervisé » ou « apprentissage non supervisé » selon qu'il existe une réalité de terrain pour former l'algorithme d'apprentissage ou non ; les méthodes supervisées utilisent la réalité du terrain.
- **Modélisation :** Il existe deux principales méthodes de modélisation : la régression et la classification. Les deux peuvent être utilisées pour faire des prévisions. Les modèles de régression contribuent à déterminer un changement dans une variable de sortie pour des variables d'entrée données ; par exemple, à quel point les notations de crédit augmentent-elle avec le niveau d'éducation ? Les modèles de classification placent les données dans des groupes ou parfois des multigroupes, répondant ainsi à des questions telles que celle de savoir si un client est actif ou inactif, ou la tranche de revenu dans laquelle il se situe. Il existe de nombreux types de techniques de modélisation pour les deux méthodes, avec des détails techniques nuancés. Les approches de modélisation ont tendance à générer beaucoup d'attention, mais il est important de noter que la méthode de modélisation n'est probablement pas une caractéristique importante de conception d'analyse. Habituellement, de nombreux types de modèles sont testés et le meilleur est alors choisi en réponse à des indicateurs de performance prédéfinis. Ou parfois, ils sont associés, créant ainsi une approche d'ensemble. Un consultant doit expliquer pourquoi une approche recommandée est choisie, et non simplement indiquer, par exemple, que la solution se fonde sur une méthode spécifique telle que la très médiatique méthode des « forêts aléatoires ». La décision de la méthode à utiliser pour la modélisation doit prendre en compte l'importance de la capacité d'interpréter la raison pour laquelle les résultats ont été produits par rapport à la précision de la prévision. Les modèles de régression ont tendance à être très transparents et facilement interprétables, par exemple ; alors que la méthode des forêts aléatoires se situe à l'autre extrémité du spectre, offrant de bonnes prévisions mais une compréhension insuffisante de la façon dont elles fonctionnent.

## Analyses prescriptives

Les méthodes de cette catégorie ont tendance à être classées en prédisant ou en classifiant les aspects comportementaux de relations complexes, et elles se composent d'un ensemble de méthodes évoluées décrites ci-dessous. L'intelligence artificielle (IA) et les modèles d'apprentissage profond appartiennent à ce groupe. Cependant, cette classification est mieux encadrée par l'infrastructure attendue nécessaire pour utiliser les résultats d'une analyse, en s'assurant qu'elle offre une valeur opérationnelle. Par exemple, cela pourrait prendre la forme d'un ensemble d'outils de tableau de bord nécessaires pour exécuter une visualisation interactive sur un site Web ou l'infrastructure informatique pour automatiser un modèle de notation de risque de crédit. L'intégration d'un algorithme ou d'un processus fondé sur des données dans un système opérationnel plus général, ou en tant que contrôleur d'accès dans un processus automatisé reposant sur lui pour fournir un service, est ce qui définit un *produit de données*.

# *Leçons du secteur : Google a attrapé la grippe*

***Modélisation prédictive et ajustement de modèle : risques de fiabilité des modèles non supervisés***

*Les chercheurs du moteur de recherche Google se sont demandé s'il pourrait exister une corrélation entre les personnes effectuant une recherche sur les mots tels que « toux », « éternuement » ou « nez qui coule » - les symptômes de la grippe - et la prévalence réelle de la grippe. Aux États-Unis, les données sur la propagation de la grippe sont décalées dans le temps ; les personnes tombent malades et vont chez le médecin, puis le médecin fait son rapport statistique, et ainsi les données enregistrent ce qui s'est déjà produit. Des modèles orientés par les mots d'une recherche pourraient-ils fournir des données en temps réel à mesure que la grippe se propage ? Cette approche de réduction des décalages temporels dans les données est appelée* ***prévision immédiate****. Pour des problèmes tels que la grippe saisonnière, les avantages en matière de santé publique sont évidents. Le modèle a été une réussite et a été rendu public sous le nom Google Suivi de la grippe. La modélisation impressionnante des mégadonnées de Google a été bien décrite dans la revue scientifique* ***Nature*** *en 2008. Six ans plus tard, cependant, l'échec du même modèle a été lui aussi bien décrit dans la revue* ***Science****. Qu'est-il arrivé entre 2008 et 2014 ?*

*Le nombre d'utilisateurs d'Internet a considérablement augmenté au cours de ces six années et les modèles de recherche de 2008 n'étaient pas constants. La question fondamentale était que Google Suivi de la grippe avait été développé en utilisant des techniques d'apprentissage automatique* ***non supervisées*** *: 45 phrases de recherche faisaient fonctionner le modèle, identifiées comme étant des corrélations statistiquement puissantes en 2008. Mais beaucoup de ces termes de recherche étaient en fait des prédicteurs de saison, et les saisons elles-mêmes étaient en corrélation avec la grippe. Lorsque les modèles de grippe survenaient plus tôt ou plus tard qu'en 2008, ces termes de recherche n'étaient plus en corrélation si forte avec la grippe. Si on ajoute l'évolution des données démographiques des utilisateurs, le modèle est devenu peu fiable. Google Suivi de la grippe a été laissé en pilote automatique, en utilisant des méthodes d'apprentissage non supervisées, et les corrélations statistiques se sont affaiblies au fil du temps, incapables de suivre l'évolution des tendances.*

*Lors de l'utilisation de méthodes similaires pour des décisions commerciales ou pour des problèmes de santé publique, il est important de se rappeler que la perte de fiabilité au fil du temps peut présenter des risques importants.*

## La méthode des forêts aléatoires

La *méthode des forêts aléatoires* a généré beaucoup d'enthousiasme dans la science des données, car elle a tendance à faire fonctionner des modèles très précis. Il s'agit d'une forme de *modèle* de classification qui utilise une structure de décision de type arborescence ou de type organigramme combiné à des approches de choix aléatoire pour identifier un chemin optimal entre le résultat désiré et un ensemble de « forêts » de variables d'entrée. Il est important de comprendre que certaines méthodes de modélisation de la science des données sont faciles à comprendre dans un contexte commercial, tandis que d'autres ne le sont pas. La méthode des forêts aléatoires peut, par exemple, générer des modèles très précis, mais sa complexité produit une « boite noire » qui la rend très difficile à interpréter. Cela pourrait être problématique pour un modèle de notation de risque de crédit ; elle pourrait identifier les personnes les plus solvables, compte tenu des données d'entrée, mais pourrait ne pas permettre de décrire ce qui rend ces personnes solvables ou ce qui détermine la recommandation de crédit.

- **Fouille de textes (traitement du langage naturel) :** La fouille de textes est le processus d'obtention d'informations de haute qualité à partir d'un texte. Le texte peut aider à identifier les opinions des clients et les sentiments sur les produits en utilisant des publications de réseaux sociaux, des messages Twitter ou de gestion de la relation client (GRC). Le Traitement du langage naturel (TLN) est une combinaison de linguistique informatique et de méthodes d'IA pour aider les ordinateurs à comprendre des informations textuelles destinées au traitement et à l'analyse.
- **Analyse des réseaux sociaux (ARS) :** Il s'agit du processus d'analyse quantitative et qualitative d'un réseau social. À des fins commerciales, l'ARS peut être utilisée pour limiter le taux de désabonnement, détecter les fraudes et les abus, ou pour déduire des attributs tels que la solvabilité en fonction de groupes de pairs.
- **Traitement des images :** Cette approche utilise des algorithmes informatisés pour effectuer des analyses à des fins de classification, d'extraction de caractéristiques, d'analyse de signal ou de reconnaissance de formes. Les entreprises peuvent l'utiliser pour reconnaitre les personnes sur des images et ainsi contribuer à la détection de fraudes, ou pour détecter des caractéristiques géographiques pertinentes pour le placement d'agent en utilisant des images satellite.

# Outils

La science des données et ses méthodes reposent sur des langages de programmation informatique, ou les algorithmes s'exécutent sur des plateformes de calcul. Les données qui alimentent ces algorithmes sont tirées de bases de données. La boite à outils du scientifique de données comprend également des connaissances pointues sur l'informatique technique et les compétences nécessaires en programmation pour développer et déployer des algorithmes de données. Les spécifications techniques de ces outils se situent au-delà de la portée de l'analyse de données des SFN. Néanmoins, certaines technologies importantes sont mises en évidence pour noter quelques outils que les scientifiques des données sont susceptibles d'utiliser. Les produits de données réussis exigent une combinaison de méthodes, d'outils et de compétences, comme nous le verrons plus loin au chapitre 2.1 : Gestion d'un projet de données.

## Outils matériels

- **Base de données :** La structure des données oriente la solution de base de données appropriée. Les données structurées sont généralement desservies par des *bases de données relationnelles* avec des schémas fixes qui peuvent soutenir la fiabilité intégrale des données, ce qui peut aider les analystes à identifier les anomalies des valeurs de données, ou les empêcher dès le départ d'enregistrer des données erronées. Les bases de données relationnelles

organisent des jeux de données en tableaux qui sont *liés* les uns aux autres par une clé, c'est-à-dire un attribut de métadonnées partagé entre les tableaux. Les solutions de banques de données d'entreprise et le stockage de données de transactions utilisent souvent des bases de données relationnelles. Les produits de premier plan sont notamment Oracle, SQL Server et MySQL. Les données non structurées sont généralement desservies par des bases de données *non relationnelles* qui ne disposent pas de schémas rigides, communément appelées bases de données NoSQL. Elles offrent des avantages en termes d'échelle et de distribution, et sont souvent utilisées pour les mégadonnées et les applications interactives en ligne. À mesure que les grands ensembles de données deviennent encore plus grands, l'espace de disque dur devient limité et le temps de calcul nécessaire pour une recherche augmente. L'avantage des bases de données NoSQL est qu'elles sont conçues pour être *horizontalement évolutive*, ce qui signifie qu'un autre ordinateur, ou deux, ou une centaine, peuvent être facilement ajoutés pour augmenter l'espace de stockage et de puissance de calcul pour y effectuer des recherches. Alors que les solutions relationnelles peuvent également être mises à l'échelle et distribuées, elles sont souvent plus complexes à gérer et à régler lorsque les données sont enregistrées sur de nombreux ordinateurs. Les produits NoSQL de premier plan sont notamment Hadoop, MongoDB et BigTable.

- **Frameworks :** Ce sont des ensembles progiciels qui combinent une solution de stockage de données à une interface de programmation (API) qui intègrent des outils de gestion ou d'analyse dans la base de données. En d'autres termes, il s'agit de solutions à source unique pour gérer et analyser les données. Les produits de premier plan sont notamment Spark et Hive. Hadoop, mentionné ci-dessus, se situe entre une base de données NoSQL et un Framework. Il est utilisé pour gérer et mettre à l'échelle des données distribuées en utilisant une approche de recherche appelée MapReduce, une méthode développée par Google pour stocker et interroger des données à travers ses vastes réseaux de données.
- **Informatique en Cloud :** Les fournisseurs tiers offrent des solutions d'hébergement qui permettent un accès à de la puissance de calcul, du stockage de données et des Frameworks. Il s'agit d'une excellente solution pour les entreprises qui veulent se lancer dans des analyses de données plus sophistiquées, en particulier les mégadonnées, mais n'ont pas la possibilité d'investir dans des serveurs informatiques et d'embaucher des techniciens pour les gérer. Les produits de premier plan sont notamment Amazon Web Services (AWS), Cloudera, Microsoft Azure et IBM SmartCloud.

## Outils logiciels

- **Langages :** « R » et Python sont deux langages de programmation qui sont devenus essentiels à la science des données. Les deux offrent les avantages du prototypage rapide et de l'analyse exploratoire qui peuvent mettre rapidement sur pied des projets de données. Les deux comprennent également des bibliothèques complémentaires conçues pour la science des données, ce qui permet un apprentissage automatique sophistiqué ou des techniques de modélisation avec une relative simplicité de programmation. Les Frameworks et les bases de données ont leurs propres ensembles de langages de programmation. SQL est nécessaire pour les systèmes de bases de données relationnelles, alors que d'autres solutions peuvent nécessiter Java, Scala, Python, ou pour Hadoop, Pig.
- **Conception et visualisation :** Les langages fondamentaux de la science des données comprennent généralement des bibliothèques de visualisation pour aider à explorer les modèles de données et visualiser les résultats finaux. Puisque de nombreux projets de données produisent des tableaux de bord interactifs ou des outils de surveillance fondés sur des données, un certain nombre de fournisseurs offrent des solutions clés en main. Voici des exemples de fournisseurs de produits : IBM, Microsoft, Tableau, Qlik, Salesforce, DataWatch, Platfora, Pyramide et BIME, entre autres, dont certains sont mentionnés dans les études de cas opérationnelles au chapitre 1.2.


# PARTIE 1

## *Chapitre 1.2 : Applications de données pour les prestataires de services financiers numériques*

Ce chapitre couvre les trois principaux domaines dans lesquels l'analyse des données permet aux entreprises d'être orientées client, créant ainsi une meilleure proposition de valeur pour le client et générant une valeur commerciale pour le prestataire de SFN. Il traite d'abord du rôle que les indications tirées de données peuvent jouer pour améliorer la compréhension des clients du prestataire de SFN. Ensuite, il montre comment les données peuvent jouer un plus grand rôle dans les opérations au jour le jour d'un prestataire de SFN typique. Enfin, il aborde l'utilisation des données alternatives en matière d'évaluations et de décisions de crédit. Ces sections présentent un certain nombre de cas d'utilisation pour montrer le potentiel que représente la science des données pour les prestataires de SFN, mais elles ne sont en aucunes façons exhaustives. Les possibilités d'affaires qu'offre la science des données ne sont limitées que par la disponibilité des données et par les méthodes et compétences nécessaires pour faire usage des données. Un certain nombre d'exemples sont présentés ci-dessous pour encourager les prestataires de SFN à initier une réflexion sur la manière dont les données peuvent permettre à leurs opérations existantes d'atteindre le prochain niveau de performance et d'impact.

La figure 8 ci-dessous montre comment les données analytiques peuvent jouer un rôle dans la prise de décision de soutien pour tous les aspects d'une activité de SFN, parallèlement au cycle de vie client et aux tâches opérationnelles correspondantes. À ce titre, les données jouent un rôle clé pour aider les prestataires de SFN à être davantage orientés client. Il va sans dire que toutes les organisations dépendent de la fidélisation de leurs clients. L'orientation client signifie établir une relation positive avec les clients à chaque étape de l'interaction, en vue de favoriser la fidélité, les bénéfices et les activités des clients. Pour l'essentiel, les services orientés client fournissent des produits qui sont fondés sur les besoins, les préférences et les aspirations de leur segment, en intégrant cette compréhension dans les processus opérationnels et la culture d'entreprise.

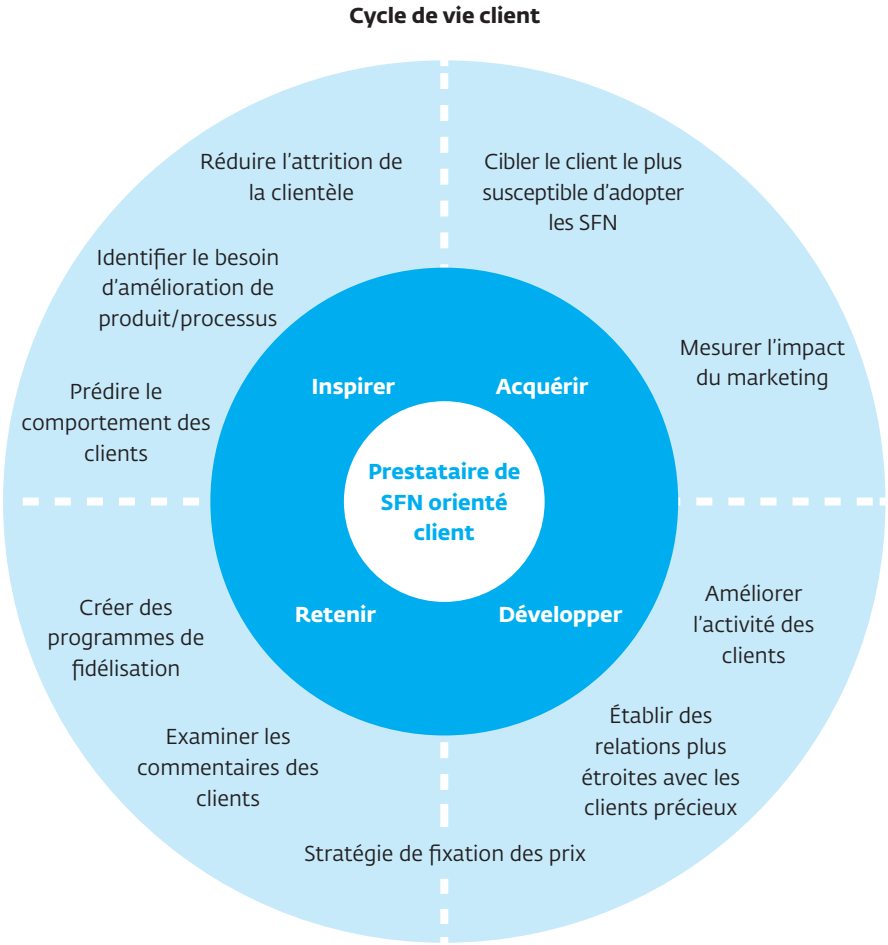



*Figure 8 : Le cycle de vie client*

Répondre aux clients est la clé de l'orientation client. Il est utile de comprendre pourquoi les clients s'en vont et le moment où ils sont les plus susceptibles de s'en aller afin que des mesures appropriées puissent être prises. Certains clients vont inévitablement s'en aller et devenir d'anciens clients. Utiliser l'analyse de données pour comprendre comment ces clients se sont comportés tout au long du cycle de vie client peut aider les prestataires à développer des indicateurs qui alertent l'entreprise lorsque des clients vont probablement s'en aller. Elle peut également donner des indications sur ceux, parmi ces derniers, que le prestataire peut être en mesure de garder et de la façon de les reconquérir.

Les prestataires de SFN pourvoient souvent aux besoins des personnes qui ne bénéficiaient auparavant pas d'accès aux banques ou à d'autres services financiers ainsi que d'autres clients mal desservis. Cela pose des défis particuliers pour les prestataires à mesure qu'ils établissent pour la première fois la confiance et la foi dans un nouveau système pour leurs clients. Ces clients peuvent avoir des revenus irréguliers, être plus sensibles aux chocs économiques et peuvent se caractériser par différentes tendances en matière de dépenses. Enfin, la nécessité de la protection des consommateurs de ce segment est plus importante, car ils pourraient avoir moins accès à l'information, avoir des niveaux inférieurs d'alphabétisation et représenter un risque plus élevé de fraude par rapport à d'autres segments. Les prestataires de SFN doivent d'abord comprendre les besoins particuliers de ces clients et ensuite concevoir des processus opérationnels qui reflètent cette compréhension. Ainsi, la compréhension des clients et l'offre d'une la valeur ajoutée aux clients est cruciale pour les prestataires de SFN, et les données peuvent les aider à être davantage orientés client.

## 1.2.1 Analyses et applications : Indications tirées du marché

Cette section explique comment utiliser les données pour avoir une compréhension plus précise et plus nuancée des clients et des marchés, ce qui peut aider un prestataire à créer des produits et des services qui correspondent aux besoins des clients. Comme cela est décrit dans le chapitre précédent, les prestataires de SFN ont accès à des données précieuses sur les clients sous différentes formes. Ces données peuvent être manipulées et analysées pour donner des indications précises sur le marché. Une telle analyse implique généralement un ensemble diversifié de méthodes, et des données quantitatives et qualitatives. Cette section commence par une étude de cas pour illustrer comment de petites étapes pour intégrer une approche fondée sur les données peuvent apporter une plus grande précision à la compréhension des préférences des clients. Elle est suivie d'une discussion sur la façon dont les données peuvent être utilisées pour comprendre l'interaction des clients avec un produit de SFN en vue d'améliorer l'activité des clients et de réduire l'attrition de la clientèle. Ensuite, elle explique comment utiliser la segmentation des clients pour identifier des groupes spécifiques au sein de la base de clients et comment utiliser ces connaissances pour améliorer le travail de ciblage. Elle est suivie d'une discussion sur la manière dont les prestataires de SFN peuvent exploiter les nouvelles technologies pour prédire le comportement financier et améliorer l'acquisition de clients. Enfin, cette section examine les moyens d'interpréter les commentaires des clients afin d'améliorer les produits et services existants.

# CAS 1

## *Zoona - Tester des stratégies de marketing pour un impact optimal*

***Développer des hypothèses pour créer des messages de marketing efficaces et les tester***

*Zoona est un PSP qui opère en Zambie, au Malawi et au Mozambique, où il compte devenir le principal prestataire de services de transferts d'argent et de comptes d'épargne simples pour le grand public. Le marketing est souvent une activité gourmande en ressources et qui exige d'y consacrer du temps, et il peut être difficile de mesurer son impact. Zoona a traité certains de ces défis en utilisant une approche orientée client pour tester trois stratégies de marketing différentes pour un nouveau produit de dépôt appelé Sunga. Tout d'abord, il a mené un projet pilote sur trois mois du produit Sunga dans une zone, étendant plus tard le projet pilote à trois autres villes pour tester trois stratégies de marketing différentes, tout cela pour identifier l'approche ayant le plus d'impact pour le lancement à l'échelle nationale. La première stratégie était appelée « Gratification instantanée », et elle récompensait tous les clients ouvrant un compte par un bracelet gratuit et offrait une grande chance de recevoir une petite récompense sous forme de remboursement d'argent à chaque fois qu'ils effectuaient un dépôt. Dans la deuxième stratégie, appelée « Loterie », les clients avaient une petite chance de gagner un prix important, avec seulement quatre gagnants sélectionnés sur deux mois. La troisième approche impliquait des ambassadeurs d'ouverture de compte allant dans des zones de haute activité, telles que les marchés, pour inciter les personnes à ouvrir des comptes.*

*Les statistiques du premier mois de ce projet pilote étendu sont présentées ci-dessous. Les chiffres ont été indexés par rapport à la ville pilote initiale, donc le chiffre 1,3 indique des résultats 30 pour cent meilleurs que le projet pilote de référence.*

*L'analyse montre que la méthode de la loterie a eu le moins de succès, alors que le plus grand nombre de comptes ouverts a été obtenu grâce à la stratégie reposant sur les ambassadeurs. Ces comptes ont également reçu des valeurs de dépôt élevées. Zoona a également étudié les taux d'activité des clients, mesurés par le nombre de dépôts par compte. L'approche de la gratification instantanée l'a emportée de loin. Dans la figure 9 ci-dessous, le 24 novembre est la date à laquelle les déposants ont commencé à gagner de petites récompenses sous forme de remboursement à chaque fois qu'ils effectuaient un dépôt sur leurs comptes : la ligne bleue montre que les dépôts progressent de façon importante.*

| Comparaison des stratégies de marketing, tableau de résultats | | |
|---|---|---|
| **INDEXÉ (30 premiers jours)** | **Nb d'inscriptions** | **Valeur du dépôt** |
| Projet pilote | 1.0 | 1.0 |
| P1 : gratification immédiate | 1.4 | 1.9 |
| P2 : Loterie | 1.1 | 1.8 |
| P3 : Ambassadeur | 3.0 | 3.8 |

***Tableau 1 :*** *La comparaison des résultats montre que la stratégie "ambassadeur" permet d' augmenter le volume d'ouverture de comptes de 300% par rapport au début du pilote*

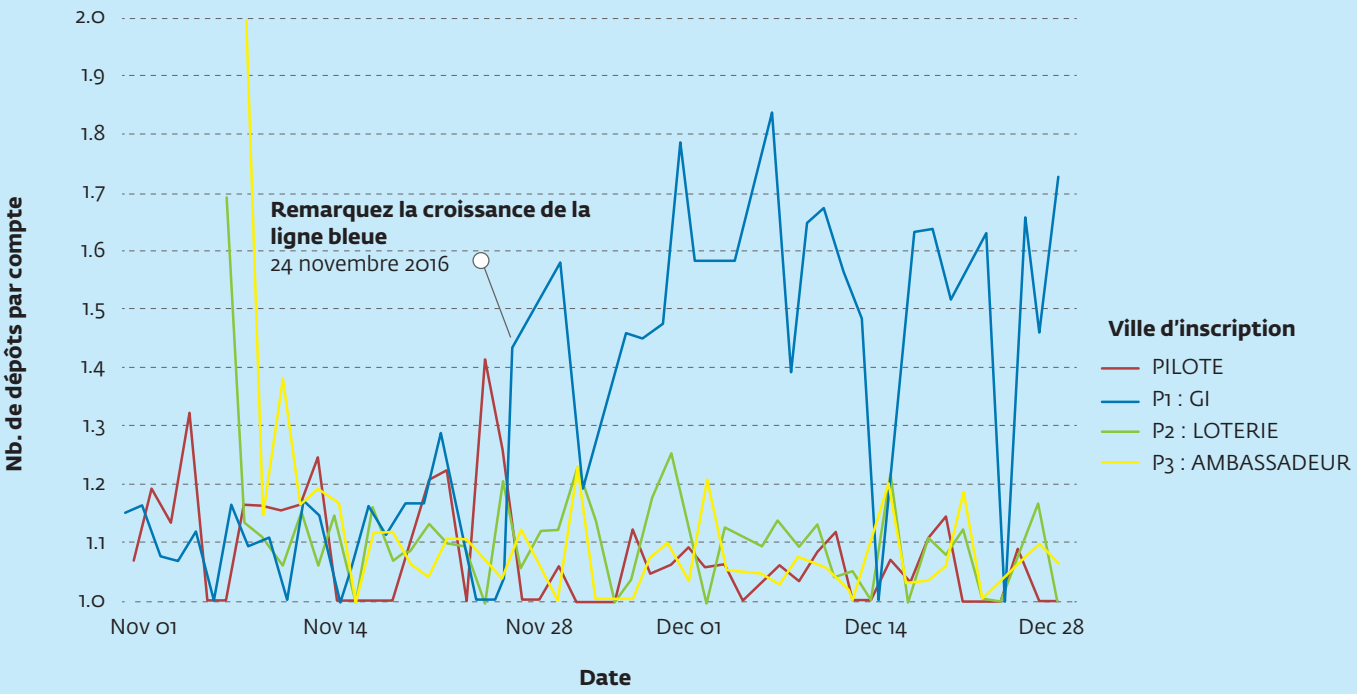



*Figure 9 : Résultats des tests de la campagne de marketing d'incitations des clients*

*Le résultat de l'analyse a été renforcé par des appels de suivi des clients. Les commentaires ont révélé que la gratification instantanée a également fait fonctionner le marketing de bouche-à-oreille, car 88 pour cent des personnes dans le groupe de gratification instantanée ont prévenu une famille ou un ami de l'existence du produit. En conséquence, la stratégie marketing à l'échelle nationale combine maintenant les deux stratégies des « Ambassadeurs » et de la « Gratification instantanée » - la première pour inciter à ouvrir des comptes, et la seconde pour stimuler les niveaux d'activité des clients.*

*Cette étude de cas montre qu'une approche rigoureuse pour tester les stratégies de marketing ne nécessite de méthodes compliquées. Au contraire, une approche et une planification systématiques en répétant rapidement des techniques mesurées par les taux de réponse des clients peut créer des indications mesurables. Cela souligne également l'avantage de combiner les méthodes pour arriver au comportement souhaité du client.*

### Cas d'utilisation : Comprendre l'engagement produit pour les offres de SFN

Comprendre comment un client utilise ou non un produit ou un service est important pour apporter des améliorations à la zone d'opération appropriée afin d'étendre la portée et d'augmenter l'adoption. Les données transactionnelles et les données de profilage des clients fournissent des informations précieuses sur la façon dont les clients interagissent avec un produit au fil du temps. Ce retour d'information peut être utilisé pour créer des messages efficaces pour le produit ou utilisé pour développer des mesures visant à gérer l'interaction des clients avec le produit. Des niveaux élevés d'inscription mais accompagnés de faibles niveaux d'activité impliquent généralement que le cout d'acquisition et de maintien de l'activité des clients est inutilement élevé. Les données transactionnelles, ainsi que les données géo spatiales, peuvent offrir au prestataire des indications sur les niveaux d'activité des clients et des agents. Ces indications peuvent aider le prestataire à effectuer des changements dans toute l'entreprise pour s'aligner sur le comportement et les besoins des clients. Ce type d'analyse peut aider à orienter les stratégies de marketing, les stratégies de recrutement d'agents ou l'adoption de processus pour les agents adoptant les meilleures pratiques, par exemple. La figure 10 fournit une illustration simple de la manière dont les données transactionnelles peuvent être interprétées. Le processus d'analyse des données est également étudié plus en détail au chapitre 2.1.

| Poser une hypothèse | Recueillir des données | Analyser les données | Mesures axées sur les données |
|---|---|---|---|
| • Que s'est-il passé ?<br>• Pourquoi cela s'est-il produit ?<br>• Que se passe-t-il maintenant ? | • Données transactionnelles<br>• Niveaux d'utilisation<br>• Comparaison des comportements entre les groupes<br>• Données de la KYC<br>• Données du CDR | • Analyse statistique simple<br>• Tableaux<br>• Corrélations | • Changer la stratégie en fonction des résultats<br>• Recherche primaire supplémentaire |

*Figure 10 : Le processus d'analyse et d'interprétation des données*

### Amélioration de l'activité des clients

Une analyse transactionnelle simple, comme on l'a vu ci-dessus peut, par exemple, révéler que des clients très actifs sont liés à des agents spécifiques. Pour être en mesure d'agir à partir de cette information, il est nécessaire de savoir pourquoi c'est le cas. Est-ce grâce aux meilleures pratiques adoptées par les agents, à cause de la situation géographique ou du fait d'une autre variable ? À titre d'exemple, des entretiens pourraient être menés afin de mieux comprendre les techniques des agents, et les données géo spatiales pourraient être utilisées pour mieux comprendre l'impact de la localisation sur l'activité des agents et des clients. Des groupes à l'activité très élevée ou très faible indiquent souvent la nécessité d'une étude plus approfondies et de groupes de discussion pour comprendre les raisons qui provoquent cette situation.

### Réduction de l'attrition de la clientèle

En regardant de près les données transactionnelles, on peut trouver des indices sur les raisons pour lesquelles les clients abandonnent le service et comment les retenir. La fréquence avec laquelle les clients interagissent avec un service peut indiquer s'ils viennent d'être acquis, s'ils sont des clients actifs du service, ou s'il est nécessaire de les attirer à nouveau pour qu'ils utilisent le service. Différents messages et canaux sont pertinents pour les clients dans chacune de ces étapes. En général, garder les clients existants est beaucoup moins couteux que d'en acquérir de nouveaux. Un grand nombre de clients qui n'ont effectué aucune transaction est signe d'un ciblage insuffisant au stade de l'acquisition. Un grand nombre de clients qui s'en vont peut indiquer d'autres limites dans l'offre de services, ce qui peut être amélioré par de petites améliorations des produits ou processus.

### Cas d'utilisation : Segmentation

Les segments peuvent être délimités par des marqueurs démographiques, des marqueurs comportementaux tels que des modèles d'utilisation des SFN, des données géographiques, ou d'autres données externes provenant des ORM telles que l'utilisation et l'achat de temps de communication et de données. Comprendre les segments est nécessaire pour découvrir les besoins et les désirs de groupes spécifiques, ainsi que pour concevoir des stratégies de vente et de marketing bien ciblées. Des indications tirées de la segmentation, destinées à développer les perspectives génératrices de revenus dans chaque segment, sont des contributions essentielles pour la feuille de route stratégique d'une institution. La segmentation de la clientèle est un aspect crucial pour devenir une organisation orientée client qui sert correctement ses clients, prend des décisions d'investissement réfléchies et maintient une entreprise en bonne santé.

En principe, bon nombre de prestataires de SFN reconnaissent l'importance de la segmentation. Cependant, dans la pratique, la plupart des prestataires de SFN desservent soit le marché de masse dans des contextes de pays en développement comme un seul segment, ou utilisent une segmentation démographique de base pour comprendre les clients. Il existe deux raisons pour lesquelles l'intégration de la segmentation visant à obtenir des indications sur les clients est limitée. Tout d'abord, les prestataires de SFN aux abois sur des marchés très concurrentiels peuvent être incités, par la réussite de certains produits, à adopter une approche orientée produit, plutôt qu'une approche orientée client, pour leur entreprise. Ainsi, les prestataires de SFN peuvent omettre de penser aux différentes utilisations possibles pour leurs offres en fonction des besoins et préoccupations des clients. Au contraire, ils peuvent choisir de mettre en évidence des cas d'utilisation et des messages pour un produit très particuliers. Ainsi, alors que le produit de transfert d'argent mobile M-Pesa a connu un grand succès au Kenya, les ORM sur d'autres marchés n'ont pas connu la même réussite, ce qui souligne la nécessité d'étudier le comportement et les besoins du marché et des clients, marché par marché, avant le déploiement de produits. En second lieu, on constate un manque de sensibilisation sur la manière de segmenter efficacement la clientèle, et la manière d'utiliser cette analyse de segmentation. Il n'est pas nécessaire que la segmentation soit compliquée ou couteuse. Les praticiens doivent définir clairement les objectifs commerciaux, qui peuvent ensuite guider l'exercice de segmentation.

## Segmentation des clients


*Figure 11 : Exemples de segments de clients de SFN, par activité du produit*

Le cadre suivant présenté par le Groupe consultatif d'assistance aux plus pauvres (CGAP) illustre la façon dont les différents types de segmentation peuvent être utilisés par un praticien en fonction de ses besoins :[17]

| Type de segmentation : | Exemple | Besoins en données | Avantages | Inconvénients |
|---|---|---|---|---|
| **Démographique** | • Rural ou urbain<br>• Homme ou femme<br>• Vieux ou jeune | Informations relevant de l'obligation de s'informer sur le client (KYC) | • Simple<br>• Les données sont faciles à trouver | • Manque d'uniformité au sein des groupes<br>• Moins riche en indications |
| **Comportementale** | • Utilisateurs qui n'ont jamais effectué de transactions ou dormants ou actifs<br>• Épargnants ou enclin à des retraits | • BD transactionnelle | • Les données sont faciles à trouver<br>• Il est facile d'attribuer de la valeur au client | • Manque d'indications sur la vie, les besoins, les aspirations du client<br>• Moins utiles pour les messages de marketing |
| **Démographique et comportementale** | • Étudiants<br>• Travailleurs migrants envoyant de l'argent à la maison | • Inscription et informations relevant de la KYC<br>• BD transactionnelle<br>• Étude de marché primaire | • Attribue de la valeur à un client et donne des indications sur sa vie et ses besoins<br>• Il est plus facile de développer des messages de marketing | • Les données sont relativement plus difficiles à trouver<br>• Il pourrait exister des segments qui se chevauchent |
| **Psychographique** | • Femmes qui veulent un endroit sûr pour épargner<br>• Clients qui pensent que l'accès à l'argent mobile est signe d'un statut plus élevé<br>• Font attention à leur budget | • Données transactionnelles historiques abondantes et significatives<br>• Recherche primaire | • Fortement sensibles aux aspirations des clients<br>• Forte proposition de valeur<br>• Il est plus facile de développer des messages de marketing | • Il est difficile de trouver des données<br>• Il pourrait exister des segments qui se chevauchent<br>• Ceci pourrait être un segment très dynamique, c'est-à-dire que les désirs pourraient évoluer |

*Tableau 2 : Cadre de segmentation des clients du CGAP*

[17] CGAP (2016). Boite à outils de segmentation des clients

# CAS 2

## *Tigo Cash Ghana augmente l'utilisation des portefeuilles d'argent mobile*

***Des modèles de segmentation de clients améliorent l'acquisition et l'activation de clients***

*Tigo Cash a été lancé au Ghana en avril 2011, et est le deuxième plus grand prestataire d'argent mobile en termes d'utilisateurs enregistrés. Malgré des taux d'inscription élevés, obtenir que les clients effectuent diverses transactions par le biais de l'argent mobile reste un défi et un objectif majeurs. Le taux d'inscription des clients, et le maintien des taux d'activité, est resté un objectif majeur après le lancement du service.*

*Une clientèle active en matière de transactions ne représente pas un défi qu'au Ghana ; la GSMA estime que les taux d'activité globaux n'atteignent que 30 pour cent.*

*En 2014, Tigo Cash Ghana a établi un partenariat avec IFC pour effectuer une analyse prédictive visant à identifier les utilisateurs de services vocaux et de données mobiles qui ont une forte probabilité de devenir des utilisateurs actifs de l'argent mobile. Pour ce faire, six mois et près de deux téraoctets de CDR et de données transactionnelles ont été analysés par une équipe de scientifiques des données.*

*Les résultats de l'analyse indiquent qu'il existe des différences entre les clients selon un grand nombre de paramètres d'utilisation du téléphone mobile, la structure des réseaux sociaux et la mobilité individuelle et de groupe.*

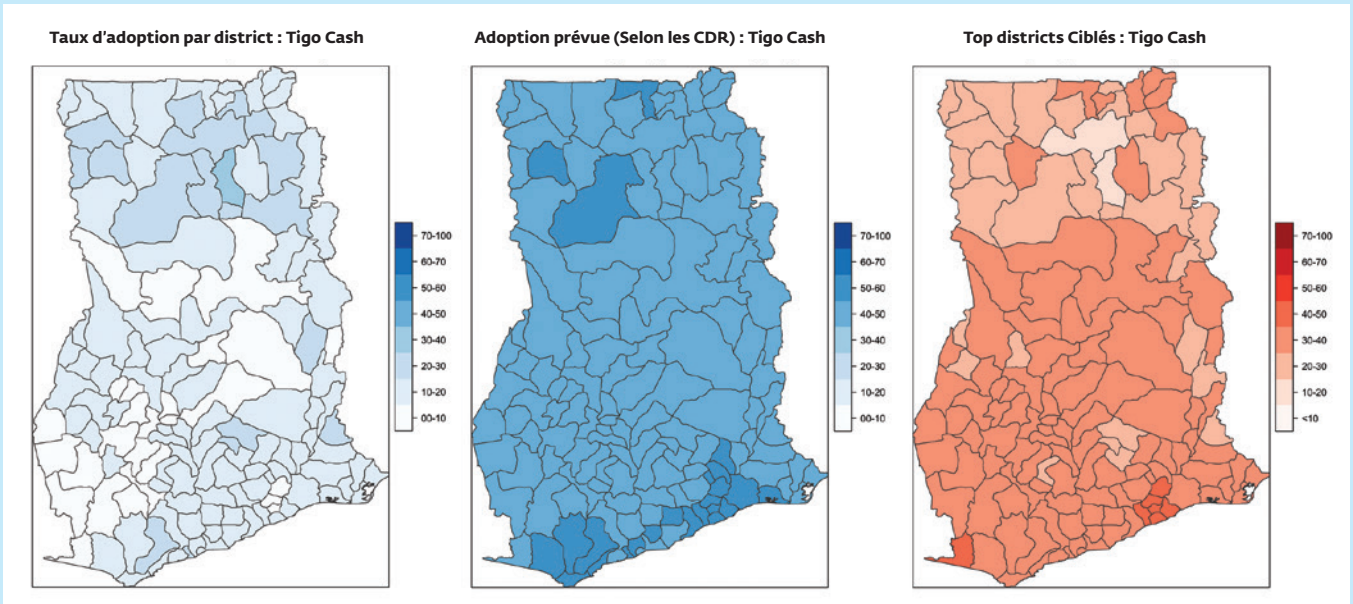



***Figure 12 :*** *Les quartiers actuels, prévus et les plus actifs en termes d'utilisation de l'argent mobile*

*Il existe de fortes différences entre les abonnés aux services vocaux et à ceux composés uniquement de données, les abonnés à l'argent mobile inactifs et les abonnés à l'argent mobile actifs. Une forte corrélation peut être observée entre les grands utilisateurs de services de télécommunications traditionnels et la probabilité que ces utilisateurs deviennent également des utilisateurs d'argent mobile réguliers actifs.*

*Avec l'aide d'algorithmes d'apprentissage automatique, l'équipe de recherche a identifié les profils qui correspondent parmi les clients des services vocaux et ceux composés uniquement de données qui ne sont pas encore des abonnés à l'argent mobile, mais qui sont susceptibles d'en devenir des utilisateurs actifs. L'équipe a également opéré une géocartographie des données (voir la figure ci-dessous) pour une analyse plus approfondie. De plus, l'analyse des CDR et des données transactionnelles a été complétée par des enquêtes non seulement pour comprendre ce qui est arrivé, mais aussi les raisons pour lesquelles cela est arrivé.*

### *Facteurs déterminants de l'adoption de l'argent mobile*

*La nécessité d'intensifier l'éducation de la clientèle et d' adapter les produits sont des éléments qui sont clairement ressortis dans chaque enquête. Seule une faible proportion d'utilisateurs d'argent mobile a rapporté que la non disponibilité des agents les empêchait d'utiliser les services d'argent mobile. Les faibles niveaux d'utilisation étaient plus étroitement liés au manque de sensibilisation des personnes à la proposition de valeur de l'argent mobile ou à l'impression qu'ils ne disposaient pas d'assez d'argent pour utiliser les services.*

### *Nouveaux clients*

*La modélisation prédictive a donné lieu à 70 000 nouveaux utilisateurs d'argent mobile actifs en raison de l'utilisation du modèle unique. Les résultats ont cartographié la réserve d'utilisateurs probables de l'argent mobile et ont identifiés les lieux où les activités de marketing hors médias avaient le meilleur impact. Avoir une idée au préalable du potentiel du marketing dans différentes zones évite une surbudgétisation du personnel de vente et augmente l'efficacité du marketing. L'approche fondée sur les données a permis d'utiliser un moyen plus réfléchi et mieux informé pour cibler les abonnés téléphoniques existants afin qu'ils adoptent l'argent mobile.*

### *L'amélioration des taux d'activité*

*L'utilisation des SMS et un grand volume d'utilisation des services vocaux et des services de données mobiles sont des facteurs clés qui ont été utilisés pour identifier les potentiels utilisateurs d'argent mobile actifs. Ce qui a commencé comme une analyse des CDR a créé une valeur de démonstration de la validité d'un concept et a conduit à une approche fondée sur les données qui a permis à Tigo Cash de dépasser le seuil d'activité de 65 pour cent parmi ses clients de l'argent mobile. La clientèle active est passée de 200 000 avant l'étude à plus d'un million de clients actifs en 90 jours.*

### *Changement de la façon de voir les choses pour les institutions*

*En tant que prestataire d'argent mobile, Tigo Cash est devenu un des services les plus prospères au Ghana. Le résultat de la collaboration est devenu le fondement de tout le travail d'acquisition de clients de Tigo Cash Ghana. Surtout, l'analyse des données a montré la valeur d'une bonne connaissance de ses clients. Tigo Cash Ghana prévoit d'augmenter sa capacité interne en science des données, ainsi que d'améliorer la compréhension de ses clients en menant une recherche primaire supplémentaire. L'objectif est maintenant passé de l'inscription de nouveaux clients, qui seront probablement actifs, à une réflexion prospective sur les moyens de maintenir des niveaux élevés d'activité de façon durable.*

*Une approche institutionnelle à l'acquisition et à la fidélisation des clients peut être fondamentalement modifiée et améliorée en utilisant tout simplement des données existantes afin de prendre des décisions opérationnelles informées.*

### Programmes marketing ciblés

Cibler les bons segments sur le marché, avec les bonnes campagnes de publicité et de marketing, peut augmenter de manière significative l'efficacité d'une campagne en termes d'intérêt suscité et d'utilisation. En utilisant une combinaison de sources de données, les prestataires de SFN peuvent segmenter les données transactionnelles selon des paramètres démographiques afin d'identifier des groupes stratégiques parmi leur clientèle. Des programmes de marketing peuvent être personnalisés pour cibler ces groupes, souvent avec une plus grande efficience et efficacité que l'approche standard. Les prestataires de SFN ont souvent combiné les connaissances sur les segments à des données sur la rentabilité afin de concentrer le travail du marketing sur les segments qui sont susceptibles d'optimiser les profits. De même, d'autres prestataires de SFN ont utilisé le cycle de vie client pour faire les bonnes offres de produit aux bons clients. Le principal défi est de trouver quels sont les groupes de clients à prendre en considération afin de concevoir une campagne de marketing appropriée. Alors que l'univers des données disponibles aux prestataires de SFN augmente chaque jour, en l'absence d'analyse pour faire la lumière sur ce point, lorsque les groupes de clients sont identifiés, les prestataires de SFN peuvent utiliser une recherche primaire pour identifier les segments sur lesquels porter son attention. Toutes les données des clients peuvent être utilisées pour développer des programmes de marketing ciblés. Cependant, les résultats sont susceptibles d'être plus pointus si l'analyse est réalisée sur les membres de segments de clients spécifiques.

### Campagnes de fidélisation et de promotion

Il peut exister des segments de clients qui effectuent un nombre très élevé de transactions sur le canal du SFN. Ces segments peuvent souhaiter des récompenses de fidélité pour des transactions spécifiques telles que les paiements chez certains types de commerçants. Autrement, le prestataire de SFN peut être en mesure d'orienter d'autres segments vers certains types de transactions en proposant des campagnes de promotion. Des transactions spécifiques dans la base de données et les profils des clients contribueraient à identifier quels groupes bénéficieraient de ces campagnes.

### Relations client de grande qualité

La segmentation des clients en fonction de la rentabilité est une application commune du processus de segmentation. On peut en outre évaluer les groupes qui sont susceptibles de devenir important à l'avenir. Les prestataires de SFN peuvent utiliser ces informations afin d'augmenter leur part de marché pour ce groupe et allouer moins de ressources à des groupes moins rentables. Les données nécessaires à ce type d'analyse sont les caractéristiques démographiques des clients, les données transactionnelles et les données concernant la rentabilité des clients.

Ceci est également valable pour l'identification des agents à haut rendement en fonction de la segmentation. En collaborant avec FINCA en République Démocratique du Congo (RDC), IFC a analysé les données de transaction des agents et les formulaires d'inscription en RDC pour montrer que le fait d'être une femme et d'être impliquée dans une entreprise axée sur les services est fortement corrélé avec le fait d'être un agent à meilleur rendement.[18]

### Améliorations de produits ou de processus

Le classement des clients en segments permet également aux prestataires de SFN d'accorder davantage d'attention aux besoins spécifiques d'une cohorte représentative. Dans un grand groupe, ces besoins peuvent disparaitre, mais en faisant attention aux plus petits segments, on permet aux prestataires de

[18] Harten et Rusu Bogdana, « Women Make the Best SFN Agents. » *Note de terrain d'IFC* 5, Partenariat pour l'Inclusion Financière

SFN d'affiner leur objectif et d'étudier des besoins et désirs insatisfaits ou ignorés. Par exemple, dans un groupe de personnes qui n'utilise pas un service, pourraient se trouver les *clients qui ont renoncé*, ou ceux qui ont réalisé des transactions mais ont cessé d'utiliser ce service. Des discussions avec ces utilisateurs pourraient révéler un besoin de réaliser de petites modifications dans le produit ou le processus. Il peut également arriver que les clients d'un segment utilisent la gamme complète de produits offerts par un prestataire de SFN, tandis qu'un autre segment n'utilise qu'un ou deux de ces produits. Dans toutes ces situations, la segmentation donne une indication des études de marché ciblées et du développement de produits visant à accroitre la demande des clients.

### Débouché commercial et produits prioritaires

Une fois l'exercice de segmentation achevé, les prestataires de SFN peuvent évaluer la mesure dans laquelle leur offre de produits répond aux besoins et aux désirs de chaque segment. Ils peuvent estimer quels segments représentent le plus grand débouché au fil du temps et le degré de compétitivité de leur offre au sein de ces segments de croissance essentiels. Ainsi, une analyse fondée sur la segmentation peut jouer un rôle important dans la feuille de route stratégique d'un prestataire de SFN.

La segmentation démographique traditionnelle, qui peut être fondée sur l'âge, le revenu ou la position géographique, est utile, mais l'expérience montre que la segmentation démographique prédit moins bien la future relation d'une institution avec un client que la segmentation fondée sur des caractéristiques comportementales. Le regroupement des clients en fonction de caractéristiques démographiques a tendance à traiter tous les clients d'un groupe comme étant identiques, quel que soit leur niveau d'activité sur le canal. Les critères démographiques peuvent également être de nature statique, lorsque, en particulier dans le monde de l'accès financier par les technologies, le comportement des clients est dynamique et en constante évolution.

L'accès à des bases de données transactionnelles peut faire de la segmentation traditionnelle un outil puissant pour obtenir des indications sur les clients. Avec des données de plus en plus disponibles, de nouveaux outils d'analyse de données et de multiples canaux à la disposition des clients, les prestataires de SFN ont maintenant la possibilité d'utiliser des informations sur les comportements individuels. Ces informations prédisent mieux les besoins et usages financiers des clients. De plus, elles reflètent les évolutions des besoins et des activités des clients. Cependant, les données comportementales peuvent ne pas livrer beaucoup d'informations sur les besoins et les aspirations des clients, ce qui rend difficile la création de messages significatifs pour ces segments.

La réalisation d'un exercice de segmentation de base de données de clients exige des ressources dédiées et un plan détaillé. En particulier, les stratégies de segmentation qui utilisent de multiples sources de données sont les plus efficaces pour décrire de façon utile et précise les groupes de clients. Ainsi, le processus d'élaboration de la segmentation de clients doit intégrer cette approche. L'analyse des données joue un rôle important dans ce processus, car elle permet aux prestataires de SFN de segmenter exactement selon les variables qui jouent un rôle pour motiver l'utilisation et susciter l'intérêt. Ce rapport ne traite que le rôle de l'analyse des données pour faciliter ce processus, mais il est important de noter que ces segments peuvent être créés par le biais de plusieurs types d'études et d'analyse.

# CAS 3

## *Airtel Money - Augmentation de l'activité avec des modèles de prévision de segmentation de clients*

***Le modèle de segmentation par apprentissage automatique apporte une valeur opérationnelle et des indications stratégiques***

*Airtel Money, l'offre de SFN d'Airtel Ouganda, a été lancée en 2012. L'intérêt suscité initialement était faible, avec seulement une fraction de ses 7,5 millions d'abonnés GSM s'inscrivant au service. Les niveaux d'activité étaient également faibles, avec environ 12,5 pour cent d'utilisateurs actifs. IFC et Airtel Ouganda ont collaboré à une étude visant à utiliser des analyses de mégadonnées et une modélisation prédictive visant à identifier les clients GSM existants qui étaient susceptibles de devenir des utilisateurs actifs d'Airtel Money.*

*Le projet a analysé six mois de CDR et de transactions Airtel Money. L'analyse a cherché à segmenter les utilisateurs d'argent mobile très actifs, actifs et non actifs. L'étude a identifié trois catégories distinctes : les niveaux d'activité GSM, les dépenses mobiles mensuelles et la connectivité des utilisateurs. À l'aide de méthodes d'apprentissage automatique, un modèle prédictif a été en mesure d'identifier les utilisateurs actifs potentiels avec une précision de 85 pour cent. Ceci a débouché sur une « haute probabilité » de 250 000 nouveaux clients actifs d'Airtel Money identifiés sur la base d'abonnés GSM qu'Airtel devait atteindre avec un marketing ciblé. L'analyse géo spatiales et du réseau des clients a permis d'identifier de nouvelles zones d'intérêt stratégique, cartographiées par rapport au nouveau potentiel d'intérêt suscité.*

*Le modèle d'apprentissage automatique a identifié certaines variables avec une grande fiabilité statistique, mais elles n'étaient pas très parlantes au sens commercial, par exemple « l'entropie de la durée de la voix ». En conséquence, une analyse supplémentaire a produit des* ***paramètres de règles métier****, ou des indicateurs qui avaient une bonne corrélation avec l'activité potentielle et avaient également de forts liens avec les ICP commerciaux. Chaque mesure avait un seuil numérique pour cibler les clients au-dessus ou en dessous d'un seuil donné. Bien qu'il ne soit pas aussi précis que le modèle sophistiqué, il a fourni un solide « découpage rapide » qui pouvait être utilisé par rapport aux ICP pour évaluer rapidement les attentes.*

*Enfin, l'étude a analysé les zones de mouvement d'argent mobile dans la région. Elle a constaté que 60 pour cent de tous les transferts se produisaient dans une zone d'un rayon de 19 kilomètres autour de Kampala. La compréhension de ce besoin de transferts de fonds à courte distance a également éclairé le travail de marketing d'Airtel Money pour les transferts P2P. De plus, cette analyse de réseau de transactions P2P a identifié d'autres villes et zones rurales avec des zones d'activités qui pourraient guider des engagements stratégiques au-delà de Kampala pour qu'Airtel puisse s'axer sur sa croissance.*

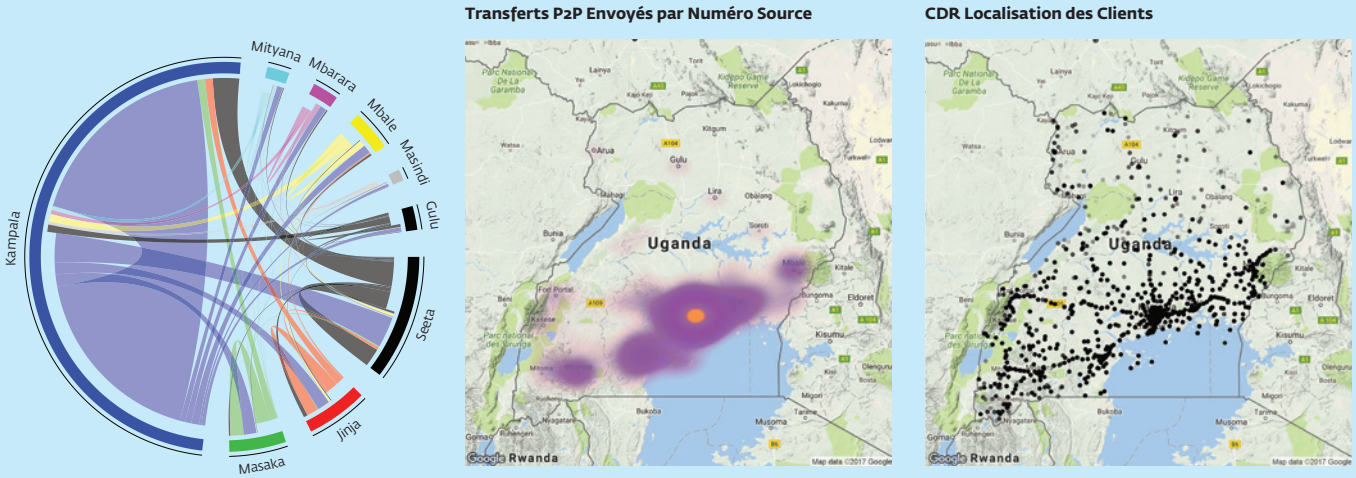



***Figure 13 :*** *Analyse du réseau (à gauche) des flux P2P entre les villes et solidité du canal. Également sur la photo, densité géo spatiales des transactions Airtel Money P2P (centre), par rapport à la distribution de l'utilisation GSM (à droite). Données en 2014.*

*Une analyse de données évoluée peut donner des indications sur des segments de clients actifs et très actifs qui peuvent conduire les modèles de propension à identifier les clients potentiels avec une grande précision. L'analyse du réseau et l'analyse géo spatiales peuvent fournir des indications pour établir les priorités en matière de planification de croissance stratégique.*

## Cas d'utilisation : Prévision du comportement des clients

Une modélisation prédictive est un outil de prise de décision qui utilise l'historique des données des clients pour déterminer la probabilité de résultats futurs. Les prestataires de SFN évaluent les informations multidimensionnelles sur les clients pour définir avec précision leurs caractéristiques qui sont en corrélation avec les résultats souhaités. Dans le cadre de la modélisation, chaque client se voit attribuer une note ou un classement qui calcule la probabilité que le client prenne une certaine décision.

Pour une institution orientée client, la modélisation prédictive peut éclairer sur la façon dont elle comprend leurs besoins et y répond. Il reste toutefois quelques obstacles qui l'empêchent d'être plus largement utilisée. Il a existé un sentiment, qui est en train d'évoluer progressivement chez les prestataires de SFN, que les prestataires connaissent déjà assez bien leur clientèle pour comprendre quels sont les produits et les campagnes de marketing qui fonctionnent. Par ailleurs, certains prestataires de SFN regardent ce qui a fonctionné ailleurs et essayent de reproduire des produits et services similaires sur leurs propres marchés. De nombreux prestataires ne savent également pas exactement comment et par où commencer le processus.

L'analyse prédictive peut aider les praticiens à atteindre les objectifs suivants :

- Acquisition de nouveaux clients
- Création d'une offre de produits optimale
- Identification des cibles de clients et prévision du comportement des clients
- Prévention du désabonnement
- Estimation de l'impact du marketing

### Nouvelle acquisition et identification des cibles

Comme en témoigne la recherche et l'expérience des praticiens, les praticiens ont réussi à abonner un grand nombre de nouveaux clients à leurs services de SFN. Cependant, la transformation de ces clients abonnés en clients actifs reste une tâche difficile que seuls quelques prestataires de SFN ont été en mesure d'accomplir. En moyenne, environ un tiers des clients abonnés n'ont effectué qu'une seule transaction au cours des 90 derniers jours.[19] L'une des raisons invoquées pour ces faibles niveaux d'activité est le ciblage insuffisant au stade de l'acquisition. La plupart des offres de SFN ciblent le vaste marché de masse. À ce titre, ils sont en mesure de voir un grand nombre de clients s'abonner, mais ont connu un succès limité en termes de conversion de ces clients en une clientèle active et génératrice de profits.

L'analyse prédictive pourrait aider à identifier les clients au stade de l'acquisition qui sont bien plus susceptibles de devenir des utilisateurs actifs à l'avenir grâce à une technique statistique appelée modélisation des réponses. Celle-ci utilise les connaissances disponibles sur une clientèle potentielle pour attribuer un score de propension à chaque client potentiel. Plus le score est élevé, plus il est probable que le client devienne un utilisateur actif. Les ORM qui sont des prestataires de SFN ont utilisé ce type de modélisation pour prédire quels membres de leur clientèle de services vocaux et de données sont susceptibles de devenir des utilisateurs actifs de leur service de SFN. Le modèle repose sur l'hypothèse que les clients qui sont susceptibles de dépenser davantage en services vocaux et de données sont aussi susceptibles d'adopter des SFN. À partir des données des CDR, le modèle est capable de prédire avec un fort degré de précision quelle est la probabilité qu'un client devienne un utilisateur actif des SFN.

### Développement des offres de produits optimales

Il existe des modèles prédictifs qui peuvent être utilisés pour découvrir les offres groupées de produits qui sont susceptibles d'être utilisées par les clients. Le modèle identifie donc les segments qui ont tendance à utiliser un seul produit tels que les transferts P2P et d'autres qui font usage de plusieurs produits, tels que les services de dépôt, l'achat de temps de communication et les transferts P2P. Cependant, le deuxième groupe peut ne jamais utiliser le service pour les microprêts. Il s'agit d'une information que

[19] « State of the Industry Report on Mobile Money, » Édition de la décennie 2006 – 2016, GSMA

le prestataire de SFN peut utiliser à des fins de marketing et de développement de produits.

### Prédire le comportement des clients

Cette analyse peut également être utilisée pour comprendre le potentiel de valeur future de chaque client. Cela inclut la valeur du cycle de vie d'un client, la fidélité des clients, les achats et le comportement en termes d'utilisation qui sont prévus, et la réponse attendue aux campagnes et programmes. De même, les prestataires de SFN peuvent augmenter leurs opportunités de montée en gamme et de ventes additionnelles en prédisant l'utilisation future grâce à l' offre de produits et aux modèles actuellement utilisés. La détermination des groupes de produit qui peuvent faire l'objet d'une offre commune grâce à l'analyse de données transactionnelle présente également une opportunité de vente additionnelle. Par exemple, un PSP peut découvrir que les utilisateurs utilisent le portefeuille comme compte de stockage, ce qui indique qu'on peut offrir un service plus efficace à ces clients par une offre de compte d'épargne.

Ces informations peuvent être utilisées pour plusieurs fonctions opérationnelles : la conception de la campagne et du marketing, les projections financières, la répartition des placements des clients et le développement des futurs produits. Ce genre de prévision peut également être utilisé, au niveau de chaque client ou au niveau global, à tout un segment.

Une analyse prédictive complète de la valeur du cycle de vie d'un client nécessite un niveau élevé de clients actifs dans tous les secteurs de produits et de canaux. Cela peut ne pas encore être réaliste pour de nombreux prestataires de SFN. Cependant, à mesure que les organisations grandissent, la capacité de prévoir de futurs modèles et tendances sur les clients ne va pas seulement devenir possible mais impérative pour faire croitre une entreprise prospère. Ainsi, être conscient de cette fonctionnalité peut aider les prestataires de SFN à l'intégrer dans leur processus de prise de décision si besoin est.

### Prévention du désabonnement

*Le désabonnement d'un client* se produit lorsqu'un client se désabonne du service d'un prestataire de SFN. Le cout du désabonnement inclut à la fois les recettes futures perdues qui aurait pu être générées par le client, mais aussi les couts de marketing et d'acquisition liés au remplacement du client perdu. De plus, au moment du désabonnement, les recettes provenant du client peuvent ne pas avoir couvert le cout d'acquisition de ce client. Ainsi, l'analyse du désabonnement des clients a deux objectifs : prédire quels clients vont se désabonner et comprendre quelles mesures de marketing sont susceptibles de convertir un client à haut risque de désabonnement en client fidélisé.

### Estimation de l'impact marketing

Le marketing des SFN a tendance à être gourmand en ressources compte tenu de sa relative nouveauté sur de nombreux marchés. Ceci est accentué par la prise de conscience qu'un produit exige un renforcement de la sensibilisation avant d'obtenir l'acceptation des clients. Sans un outil de mesure de la réussite, les gestionnaires sont obligés de se fier à leur instinct et aux données de vente de haut niveau pour évaluer la valeur de leur travail de marketing. Étant donné que les clients sont désormais en interaction avec les prestataires de SFN sur plusieurs canaux, numériques et autres, il est également difficile d'isoler les effets des campagnes spécifiques, car les clients sont exposés à de nombreux messages à tout moment.

La modélisation prédictive permet de mesurer l'impact du marketing sur le comportement des clients. Selon les données disponibles, l'analyse peut permettre aux prestataires de SFN d'estimer le « lift », ou l'augmentation des ventes qui peut être attribuée au marketing. La modélisation prédictive identifiera comment des mesures de marketing spécifiques peuvent avoir un impact sur le comportement des clients dans tous les segments. Elle peut montrer, par exemple, qu'une certaine mesure prise en marketing ou de la publicité sur un certain canal peut avoir une réponse beaucoup plus marquée dans certains segments que la réponse moyenne de la population.

### Messages de marketing personnalisés

Les sections précédentes ont déjà traité de la façon dont le marketing ciblé peut utiliser une compréhension plus approfondie des segments de clients. *Le marketing personnalisé* est un marketing ciblé à un niveau très personnalisé, dans lequel les besoins et désirs individuels des clients sont anticipés en fonction de leur comportement passé et d'autres informations signalées. De nombreux clients éventuels ont une expérience limitée des services financiers et doutent souvent de leur capacité à leur être utiles. La messagerie personnalisée permet aux prestataires de SFN de « parler » à leurs clients comme s'ils les connaissaient, ce qui permet ainsi aux prestataires de SFN de gagner leur confiance. Les clients peuvent en outre avoir une relation très personnalisée avec leur prestataire. Sur les marchés concurrentiels, des messages personnalisés contribueraient à établir une affinité pour un service plutôt qu'un autre. Les clients sont beaucoup plus susceptibles de répondre à des messages qui répondent à leurs intérêts, plutôt qu'à des messages non-personnalisés qui font référence à une proposition de valeur de SFN de très haut niveau et non spécifique. Enfin, le bon message marketing va inciter le client à prendre des mesures en fonction des messages qu'il reçoit, sans doute parce qu'ils touchent du doigt les besoins sous-jacents du client.

Certains messages personnalisés peuvent ne pas atteindre les objectifs ciblés, car les messages non sollicités peuvent facilement être ignorés, ou pire, peuvent entrainer des associations négatives avec le prestataire de SFN. Ainsi, les messages personnalisés doivent être soigneusement conçus et ciblés afin de garantir qu'ils atteignent les clients qui ont besoin de l'information.

Comment les prestataires de SFN peuvent-ils personnaliser les messages de marketing ?

1. **Recueillir des données et identifier des clients :** Tout d'abord, les prestataires de SFN doivent recueillir des données sur leurs clients. Les sources de ces données comprennent les transactions des clients, les données démographiques, les préférences et les contributions des réseaux sociaux.
2. **Comprendre les clients :** Ensuite, les prestataires de SFN doivent examiner ces données et envisager la segmentation en groupes en fonction de caractéristiques communes.
3. **Développer des messages et interagir avec les clients :** Les prestataires de SFN doivent ensuite créer des messages pour les clients et identifier les canaux appropriés pour transmettre des messages à leur clientèle. L'étape suivante consiste à interagir avec la clientèle grâce à la messagerie.
4. **Tester l'efficacité de la messagerie :** L'impact du message peut être mesuré en utilisant le test A/B. La personnalisation doit être accompagnée de tests pour qu'il soit possible d'évaluer son impact.
5. **Affiner les messages :** Les commentaires des clients et la mesure de l'impact doivent permettre d'affiner les messages.

# CAS 4

## *Juntos offre des messages d'interaction évolutifs et personnalisés avec les clients*

***Sources de données : Les données qualitatives et quantitatives améliorent la segmentation et la sensibilisation***

*Juntos, une société de technologie de la Silicon Valley, a établi un partenariat avec des prestataires de SFN pour établir des relations de confiance avec les utilisateurs finaux, améliorant ainsi les taux généraux d'activité des clients. À l'échelle mondiale, de nombreux prestataires de SFN connaissent une forte inactivité et une faible interaction. Cela décourage les prestataires, dont les investissements peuvent ne pas connaitre un rendement financier suffisant et dont les clients peuvent avoir accès à des services qu'ils n'utilisent pas suffisamment. Juntos offre une solution à ce problème en utilisant des messages personnalisés d'interaction avec des clients fondés sur des stratégies de segmentation basées sur des données qui produisent des résultats quantifiés.*

*Cette approche est fondée sur des données de qualité. Tout d'abord, Juntos conduit des études ethnographiques pour mieux comprendre les clients sur le marché. Les interactions sont toujours guidées par des données quantitatives fournies par le partenaire de SFN, des études comportementales qualitatives effectuées dans le pays et des leçons tirées de l'expérience internationale. Après avoir obtenu une compréhension initiale de l'utilisateur final, Juntos effectue une série d'essais randomisés contrôlés (ECR) avant le lancement complet du produit. Ces expériences contrôlées sont conçues pour tester le contenu, les modèles de timing ou de remise des messages, et identifier l'approche la plus efficace pour interagir avec les clients.*

*Pour commencer, les messages sont envoyés aux utilisateurs, et les utilisateurs peuvent répondre à ces messages. Cela établit la relation de confiance nécessaire. Plus important encore, ces réponses sont reçues par un « chatbot » (un agent conversationnel) automatisé de Juntos qui analyse les résultats selon trois ICP :*

- ***Taux d'engagement** : Quel pourcentage des utilisateurs ont répondu au chatbot ? À quelle fréquence ont-ils répondu ?*
- ***Contenu des réponses** : Quelles étaient les réponses ? Quelles informations ont-ils communiqué ou demandé ?*
- ***Comportement transactionnel** : Est-ce que le comportement transactionnel a changé après avoir reçu des messages pendant une semaine ? Un mois ? Deux mois ?*

*Ces expériences permettent à Juntos de déterminer quels sont les clients inactifs devenus actifs suite à la sensibilisation des messages de Juntos, et de savoir quels messages ont permis une activité plus forte et plus cohérente. Par exemple, un message de commande est envoyé à un groupe d'utilisateurs choisis au hasard : « Vous pouvez utiliser votre compte pour envoyer de l'argent à la maison ! » D'autres pourraient puiser dans les données du service pour inclure le nom du client : « Salut Jean, saviez-vous que vous pouviez utiliser votre compte pour envoyer de l'argent à la maison ? » D'autres données peuvent être intégrées au message : « La dernière fois que vous avez utilisé votre compte, c'était il y a 20 jours. Où voulez-vous envoyer de l'argent aujourd'hui ? » Ce ne sont que des exemples, mais ils montrent comment on peut comparer un message générique avec un message personnalisé avec une incitation qui tient compte du moment. Les données ethnographiques de base de Juntos améliorent la compréhension qualitative des clients, ce qui contribue à établir une hypothèse à propos de laquelle les messages sont susceptibles de résonner, puis de soumettre ces messages à un test statistique.*

*La première question est de savoir si les messages test produisent des résultats statistiquement plus significatifs que les messages génériques. Lorsque la réponse est « oui », il est important d'approfondir les choses, de se poser des questions sur la personne interrogée et de faire des sondages dans tous les segments tels que les segments ruraux ou urbains, masculins ou féminins, correspondant à une tranche de revenu, et selon des modèles d'utilisation, en fusionnant ces informations avec les données ethnographiques sur les opinions des consommateurs.*

*En testant une grande diversité de messages, Juntos est en mesure de segmenter les groupes d'utilisateurs selon les messages qui montrent une amélioration statistiquement significative de l'utilisation au fil du temps. Cela signifie que les messages de fort engagement peuvent être conçus pour tout le monde, des femmes rurales aux jeunes hommes ou aux citadins à revenu élevé. L'approche de Juntos est adaptée à chaque contexte et est affinée en permanence pour s'adapter en souplesse aux clients qui modifient leurs interactions au fil du temps.*

*Recueillir les opinions des clients et les données du marché de manière qualitative permet une meilleure compréhension du comportement des clients, ce qui aide les prestataires à rédiger des messages que les personnes aiment lire. Les tests de l'hypothèse statistique identifient quels messages résonnent le mieux avec des groupes spécifiques, ce qui permet de créer des messages personnalisés pour des publics ciblés.*

## Cas d'utilisation : Comprendre le retour d'information et les analyses des textes des clients

Les prestataires de SFN peuvent aussi extraire des indications utiles sur les préférences et attitudes des clients grâce à de nouvelles techniques fondées sur des algorithmes qu'on appelle fouille de textes, ou analyse de texte. Aujourd'hui, de nombreuses sociétés peuvent accéder à des informations sur ce que les clients aiment ou n'aiment pas le biais des réseaux sociaux, des e-mails, des sites Web, et de transcriptions de conversations avec des centres d'appels. Ces méthodes ont notamment été appliquées dans des contextes de pays développés en Europe et en Amérique du Nord. Toutefois, les prestataires de SFN sur les marchés émergents peuvent également vouloir analyser ces données pour contribuer à la croissance de l'entreprise. L'analyse de texte peut également être faite manuellement. Avec les progrès de la technologie, ces méthodes sont susceptibles de devenir moins chères et plus adaptables aux contextes et langues des pays en développement.

L'application la plus courante pour l'analyse de texte repose sur deux méthodes :

1. **Méthodes de synthèse de texte :** Ces méthodes fournissent un résumé de toutes les informations clés dans un texte. Ce résumé peut être créé soit en n'utilisant que le texte original (approche d'extraction) soit en utilisant du texte qui n'est pas cité dans le texte (approche d'abstraction).
2. **Analyse des opinions :** L'analyse des opinions ou « exploration des opinions » est un outil fondé sur des algorithmes utilisé pour évaluer le langage, parlé et écrit, afin de déterminer si l'expression d'opinion est positive, négative ou neutre et à quel point. Grâce à cette analyse, les prestataires de SFN comprennent ce que les clients pensent de leurs produits, la façon dont ils s'associent à la marque et la façon dont ces attitudes évoluent au fil du temps. Les pics ou creux sont d'un intérêt particulier pour l'analyse des opinions.

À l'heure actuelle, les évaluations tirées de l'analyse de textes peuvent être appliquées à trois domaines :

### Amélioration des produits et services

Les prestataires de SFN pourraient apporter des améliorations rapides aux produits et services s'ils pouvaient avoir un contact direct avec les clients. Les réseaux sociaux, e-mails et autres mécanismes de retour d'information direct sont un excellent moyen de connaitre immédiatement et directement les opinions des clients. Une étude de marché peut ne représenter qu'une source limitée de commentaires des clients dans ce contexte.

### Le marketing de bouche-à-oreille

Le marketing de bouche-à-oreille reste la forme de publicité la plus digne de confiance pour de nombreux clients. Pour les produits et les prestataires de SFN qui ont déjà une large clientèle, motiver des clients satisfaits pour stimuler le marketing de bouche-à-oreille n'est pas difficile. Cependant, pour les nouveaux produits comme les SFN, les prestataires doivent trouver une méthode pour catalyser les niveaux d'éducation parmi les clientèles potentielles, en particulier chez les clients qui montrent de l'enthousiasme et de l'initiative à l'égard du produit au sein de la clientèle cible. En règle générale, les clients sont plus motivés pour passer le mot sur un ou deux cas d'utilisation spécifiques ; ils diffusent rarement un message générique sur la marque. Les fils des réseaux sociaux et autres informations sur le Web peuvent être utilisés pour identifier les leaders d'opinion par leur connectivité, le niveau et la nature des interactions et leur portée potentielle. Ce type d'analyse dépend de données non structurées provenant de réseaux sociaux, de données provenant de sites de critiques et de données provenant de blogs.

### Impact marketing et surveillance des retours d'information

L'exploration des opinions permet aux prestataires de SFN de comprendre le processus de réflexion d'une immense quantité de clients. Grâce à l'analyse des opinions, il est possible de suivre ce que les clients disent sur les nouveaux produits, publicités, services, marques et autres aspects du marketing. Cette analyse peut également être utilisée pour comprendre la manière dont le marché perçoit les produits et services des concurrents. Ces données provenant de réseaux sociaux, blogs, sites de critiques, et autres sites Web dans le domaine social sont également non structurées.

## 1.2.2 Analyses et applications : Gestion des opérations et des performances

L'équipe des opérations est responsable du fonctionnement de la « salle des machines », qui est au cœur de l'entreprise de SFN, car elle effectue une myriade de tâches, notamment : recueillir les données, stocker les données et garantir que leur connectivité est fluide entre les différents systèmes et applications pour l'ensemble de l'environnement informatique du prestataire de SFN ; surveiller en permanence la qualité des données ; accueillir les agents et gérer leurs performances ; veiller à ce que la technologie fonctionne comme prévu ; fournir une assistance à la clientèle ; fournir les informations et les outils nécessaires à l'équipe commerciale, notamment la mesure des performances, la surveillance des risques et l'établissement de procédures réglementaires de déclaration ; la résolution des problèmes ; surveiller efficacement les indicateurs, les exceptions et les anomalies ; gérer les risques ; et veiller à ce que l'entreprise respecte ses obligations réglementaires. Cela ne peut être fait de façon efficace sans avoir accès à des données précises, présentées sous une forme pertinente, facile à lire et en temps voulu.

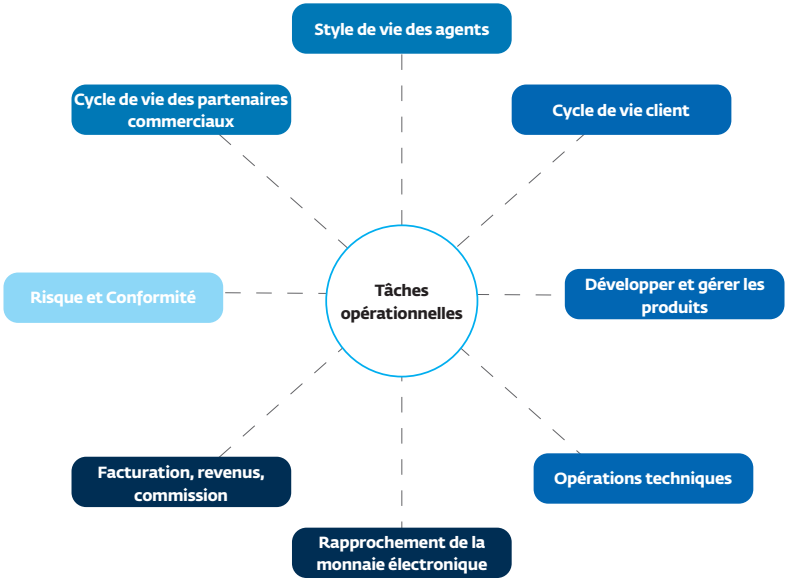



*Figure 14 : Tâches opérationnelles*

Cette équipe joue un rôle important dans la structure organisationnelle, car elle est indépendante des autres fonctions de base et également impliquée dans des activités essentielles de l'entreprise. La nature des responsabilités de l'équipe nécessite des compétences techniques, ainsi qu'une excellente connaissance du volet commercial. Cette combinaison permet des interprétations de données significatives qui peuvent au bout du compte faciliter les processus de prise de décision des acteurs clés de l'entreprise.

Cette section décrit le rôle que les données peuvent jouer dans l'optimisation des opérations au jour le jour d'un prestataire de SFN typique. Elle commence par décrire la façon dont les données peuvent être converties en informations utiles, en donnant des exemples concrets d'application d'analyse des données. Elle inclut quelques conseils sur les meilleures pratiques d'utilisation des données des SFN. À mesure que l'utilisation des tableaux de bord de données devient plus courante, elle donne des indications sur la création et le contenu des tableaux de bord.

### Cas d'utilisation : Visualiser les performances avec des tableaux de bord

On dit souvent qu'une image vaut mieux que mille mots. Ainsi, trouver un moyen graphique de représenter des données est un moyen puissant de communiquer rapidement des informations et des tendances, ce qui est essentiel pour assurer une surveillance constante de la performance des entreprises et pour identifier des risques avant qu'ils ne s'accroissent. Des tableaux de bord bien structurés, adaptés à différents groupes d'utilisateurs, doivent refléter la demande des unités opérationnelles et les aider à prendre des décisions plus informées.

La conversion des données en graphiques et autres formes de visualisation favorise la communication des informations révélées et contribue également à repérer les tendances et anomalies dans les données. Beaucoup de personnes dans l'organisation n'ont pas le temps ou les ressources nécessaires pour analyser les données elles-mêmes ; elles veulent simplement que leurs questions aient des réponses qui les aideront à faire leur travail de manière plus efficace.

Un tableau de bord donne un aperçu des ICP pertinents pour un service ou toute l'entreprise. S'il est rarement nécessaire de prendre des mesures fondées sur les données signalées, les paramètres du tableau de bord sont probablement incorrects. Pour concevoir des tableaux de bord solides, il est important d'intégrer les commentaires des utilisateurs finaux afin de répondre à leurs besoins spécifiques. Sans ce retour d'information, les tableaux de bord pourraient devenir obsolètes et tout le travail consacré à leur création serait perdu. Par conséquent, le développement du tableau de bord est un partenariat entre les équipes opérationnelles et commerciales, ce qui pourrait passer par des répétitions pour faire le tour de la boucle de rétroaction des différentes parties prenantes.

Certains tableaux de bord nécessitent une mise à jour en temps réel. Ainsi, une équipe technique opérationnelle doit agir lors d'alertes déclenchées en temps réel : les responsables de l'assistance à la clientèle évaluent activement les volumes d'appels pour attribuer le travail d'équipe et gérer les incidents, les équipes de gestion des risques sont constamment informées des remboursements qui ne fonctionnent pas, et les équipes de vente peuvent prendre des mesures précoces sur les comptes à faible activité pour activer le client et ne pas laisser le compte en sommeil. Certains de ces tableaux de bord permettraient aux utilisateurs finaux de manipuler les données pour visualiser différentes découpes et segments de données. Souvent, ces types de tableaux de bord sont présentés en direct sur un grand écran dans les locaux de l'équipe pour que tout le monde puisse les voir. Pour le personnel sur le terrain, où l'accès à Internet peut être de qualité variable, des tableaux de bord en ligne peuvent être téléchargés et mis en cache localement pour être utilisés sur le terrain.

D'autres tableaux de bord de gestion fournissent des indications en analysant les données de la veille, de la semaine précédente, du mois précédent ou de l'année précédente, et peuvent donc être livrés de multiples façons, notamment sous forme de rapports, de présentations ou via un portail en ligne. Par conséquent, chaque service et équipe de projet a besoin de tableaux de bord personnalisés selon les objectifs et initiatives du service. Habituellement, au minimum, les solutions de SFN doivent avoir plusieurs tableaux de bord des opérations couvrant les domaines suivants, chacun fournissant un accès en fonction des rôles pour des publics spécifiques :

- **Risque :** Pertes de recettes ; prêts non performants (PNP) ; indications concernant la Lutte contre le blanchiment de capitaux (LBC) ; adéquation des fonds propres ; détection des fraudes
- **Finance :** Perspectives de profits et pertes ; surveillance de la monnaie électronique
- **Marketing :** Indications et tendances sur les clients pour les différentes offres
- **Ventes :** Performance des agents ; performance des commerçants et des émetteurs de facture ; performance de l'équipe de vente
- **Opérations :** Gestion de la liquidité des agents
- **Assistance à la clientèle :** Statistiques et indications provenant du centre d'appels
- **Opérations techniques :** TIndications provenant de l'équipe des opérations techniques

Les outils de gestion de données disponibles sur le marché ont énormément évolué ces dernières années. Des tableaux de bord standard sont souvent livrés dans le cadre de l'offre technologique du fournisseur. Pour obtenir les indications plus précises nécessaires et le faire de manière reproductible, il existe deux approches standard :

1. **Retour au fournisseur :** Un budget est souvent disponible pour que les fournisseurs modifient les tableaux de bord, mais la rivalité des nombreuses demandes des services et des nombreux clients des fournisseurs exigeant de l'attention peut entrainer des problèmes de capacité et des retards.
2. **Utiliser Excel pour manipuler des rapports bruts téléchargés à partir de « cubes de données » du système :** Lorsqu'une question est posée à l'équipe de soutien à la décision de l'entreprise, elle crée un tableau de bord personnalisé et produit un rapport ou une présentation PowerPoint pour tenter d'offrir une réponse. Il s'agit d'une autre forme ad hoc de création de tableau de bord.

La dernière génération d'outils de gestion de données permet d'avoir la liberté d'enquêter sur des domaines d'intérêt sans nécessiter une expertise en manipulation des données. Cependant, les bases de données sous-jacentes doivent être conçues et optimisées pour être capables de déployer et d'utiliser ces types d'outils. Quel que soit le processus de gestion des données ou le système utilisé, voici les points à prendre en compte lors de la création d'un tableau de bord :

1. **Pensez à la réponse « Et alors ? » :** Les résultats doivent avoir une valeur pratique, et pas seulement être « bons à savoir ». De nombreux tableaux de bord ne montrent que l'état actuel de l'entreprise et ne donnent pas le contexte des résultats précédents ou des tendances temporelles.
2. **Choisir à quelle question on doit répondre avant de commencer :** Souvent, les rapports sont un lieu de déversement de toutes les données disponibles, qu'elles soient utiles ou non. Ces types de rapports ne contiennent pas les indicateurs et mesures sources de motivation qui améliorent la performance.
3. **Concevoir le rapport pour raconter une histoire :** Une fois que les bonnes données sont mesurées et recueillies, le rapport doit contenir des informations accrocheuses pour attirer l'attention du lecteur sur les points les plus importants. Présenter de façon visuelle, intéressante et utile.

### Rapports standard des opérations

Afin d'améliorer leurs activités, les prestataires de SFN tentent de trouver les réponses à des questions telles que :

- Quel était le volume et la valeur des transactions ?
- Combien de clients et d'agents étaient actifs ?
- Quel a été le montant de nos recettes ?
- Combien cela représente-t-il par rapport au mois dernier et au budget ?
- Existe-t-il des indicateurs de risque en dehors des limites acceptables ?
- Existe-t-il des transactions inhabituelles récurrentes, des pics d'activité ou des anomalies qui révèlent une activité inhabituelle ?

Le point de départ est de se concentrer sur les ICP, ou des paramètres avec des objectifs quantifiables que la stratégie opérationnelle s'efforce d'atteindre et qui servent de référence pour juger la performance. Les ICP généraux des entreprises doivent être directement liés aux objectifs stratégiques de l'organisation et, par conséquent, déterminer les ICP spécifiques de chaque service. Les données les plus utiles sont celles qui peuvent être converties en informations nécessaires pour prendre des décisions. Avant de créer un rapport, il convient d'identifier exactement ce que l'on veut savoir. Il convient également de confirmer que des mesures seront prises suite à l'obtention des données.

Les ICP des services bien structurés offrent aux équipes opérationnelles des indications avec lesquelles elles peuvent mesurer les performances par rapport aux cibles. Ils aident les équipes à comprendre ce qui se passe sur le terrain et dans quel domaine il existe un potentiel d'amélioration.

Les rapports standards d'ICP sur les principaux moteurs d'activité sont généralement segmentés par zone opérationnelle. Les ICP sur lesquels se concentrer pour chaque domaine opérationnel respectif figurent dans le Tableau 3 ci-dessous.

| Service | Thèmes Prioritaires des ICP |
|---|---|
| **Finance et trésorerie** | Recettes, produits et charges d'intérêts, frais et commissions, montant en dépôt, volume et valeur des transactions, volume de clients et d'agents (actifs), couts indirects et émission de monnaie électronique pour les établissements non bancaires, rapprochement des relevés bancaires |
| **Cycle de vie des partenaires commerciaux** (commerçants, émetteurs de facture, commutateurs, agents, banques partenaires, autres PSP) | Recrutement, niveaux d'activité, résolution de problèmes, gestion des performances, rapprochement et règlement |
| **Gestion du cycle de vie des clients** | Gestion de la KYC, niveaux d'activité, comportement transactionnel, résolution de problèmes (assistance à la clientèle) et gestion des comptes |
| **Opérations techniques** | Suivi des performances des produits, suivi des niveaux de service des partenaires, gestion du changement, intégration des partenaires, résolution des pannes, gestion des incidents et gestion des accès utilisateur |
| **Risque de crédit** | Structure des risques de portefeuille, prêts non performants, pertes liées aux annulations et risques, provisionnements liés aux prêts |
| **Risque opérationnel et de conformité** | Gestion des risques opérationnels, surveillance et suivi des activités suspectes, conformité réglementaire, vérification préalable et enquêtes ad hoc |
| **Cycle de vie (spécifique aux SFN) du réseau d'agents** | Recrutement, niveaux d'activité, gestion du fonds de caisse, résolution des problèmes, gestion des performances, rapprochement et règlement, et audit |
| **Autre** | En fonction de la nature des SFN, d'autres rapports peuvent être nécessaires, par exemple, les organismes octroyant un crédit calculent une cote de crédit, recouvrement de la dette et tâches connexes |

*Tableau 3 : ICP sur lesquels se concentrer par domaine opérationnel*

En fonction de la stratégie commerciale et des objectifs du service, une sélection des données ci-dessus sont présentées en tant qu'ICP de l'entreprise et des services. Ces ICP peuvent, idéalement, être présentées sous forme de tableaux de bord, ou d'une série de rapports. Il est important que chaque service sépare ses données en ICP d'une part et en données justificatives d'autre dans les rapports de gestion, car il existe toujours la tentation d'inclure des données périphériques qui ne sont pas strictement nécessaires pour comprendre la santé de leur service. Cela peut être source de distraction ou conduire à de mauvaises priorités. Les données justificatives sont essentielles pour aider à mieux comprendre ce qui détermine les ICP et à décider comment ils peuvent être optimisés, mais elles ne nécessitent généralement pas d'être signalées à un large public, à moins de vouloir attirer l'attention sur un point particulier. Un bon exemple de cette approche est illustré ci-dessous : l'utilisation de tableaux de bord de données par MicroCred.

# CAS 5

## *MicroCred utilise des tableaux de bord de données pour améliorer ses systèmes de gestion*

***Visualisations et tableaux de bord de données pour le suivi des performances quotidiennes et de la fraude***

*MicroCred est un réseau de microfinance axé sur l'inclusion financière en Afrique et en Asie. Au Sénégal, il exploite une entreprise de microfinance en croissance qui offre des services financiers aux personnes qui n'ont pas accès aux banques ou à d'autres services financiers. La portée a été étendue à l'ensemble du pays en créant un réseau de plus de 500 agents de SFN. Les appareils de PDV des agents peuvent effectuer des transactions de gré à gré pour les paiements de factures et les envois de fonds, et traitent également des dépôts et des retraits sur les comptes MicroCred. La confirmation de la transaction est assurée par la réception d'un SMS. À la fin de 2016, près d'un tiers des clients avaient créé un compte pour utiliser le canal des agents, et plus d'un quart utilisaient activement les points de vente des agents pour effectuer des transactions. Cela a généré d'importantes données sur les opérations et la performance du canal.*

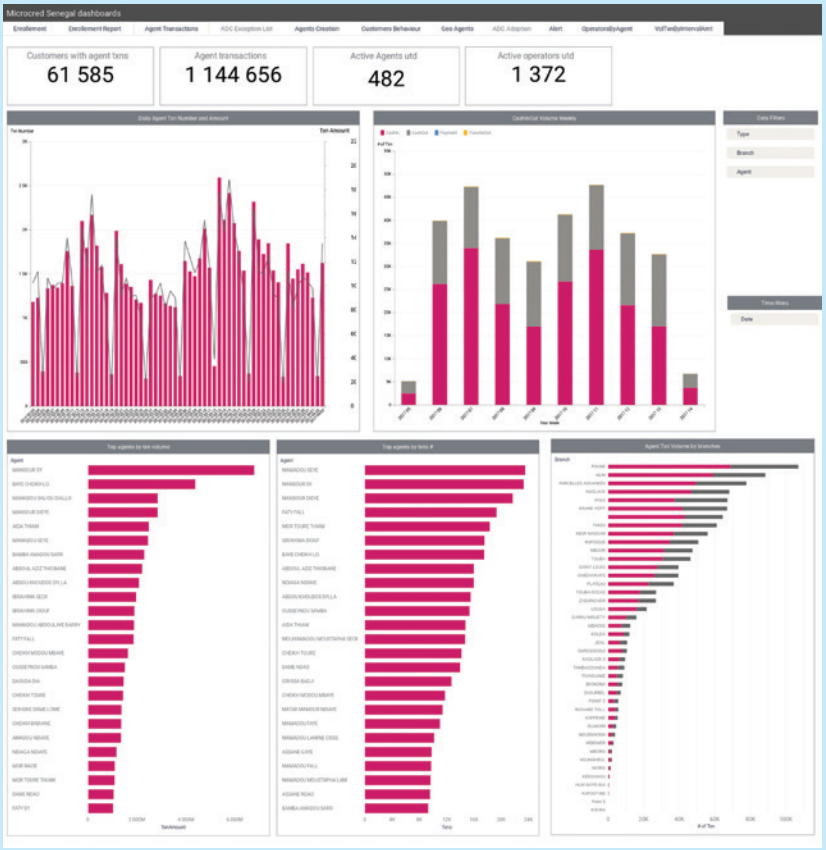



***Figure 15 :*** *Exemple des données des tableaux de bord de MicroCred*

*MicroCred a été un adopteur précoce des systèmes de gestion de données de nouvelle génération, en acquérant et mettant en œuvre BIME, un outil de visualisation pour faciliter l'optimisation des opérations. Il a permis à MicroCred de développer des tableaux de bord interactifs conçus pour répondre à des questions opérationnelles spécifiques.*

*MicroCred utilise le plus souvent deux tableaux de bord :*

### *Tableau de bord des opérations quotidiennes*

*Il permet une visualisation quotidienne des portefeuilles d'épargne et de prêts, en mettant en évidence tout problème. Il présente des données sur une période de trois mois, mais peut être ajusté en fonction des besoins des utilisateurs. Ce tableau de bord utilise des alertes automatisées pour avertir l'équipe des opérations de problèmes potentiels. Dans les rapports, personnalisés pour les équipes opérationnelles, figurent des mesures telles que :*

- *Les ICP de suivi, notamment les volumes de transactions, les commissions et les frais*
- *L'activité de l'agent, avec des alertes pour signaler les agents qui n'effectuent pas de transactions ou qui sont peu performants*
- *Des alertes provoquées par une activité suspecte et une fraude potentielle, telle que l'activité inhabituelle d'un agent ou d'un client*
- *Le suivi des processus d'abonnement aux SFN, en mettant l'accent sur les abonnements infructueux*
- *La répartition géographique des transactions*

### *Tableau de bord stratégique mensuel*

*Il donne une vision à long terme, plus stratégique, et est principalement utilisé par l'équipe de direction pour visualiser des mesures commerciales critiques plus complexes. Il a été développé pour tenir compte des comportements au cours du cycle de vie client, notamment la façon dont l'utilisation du service évolue à mesure que les clients se familiarisent avec la technologie et les services proposés. Il est également possible d'effectuer aisément des analyses ad hoc pour suivre des questions soulevées par les données présentées dans les tableaux de bord. Il s'axe sur :*

- *L'utilisation des agences de MicroCred par rapport aux agents*
- *L'adoption et l'utilisation des SFN par les clients*
- *Le déploiement du canal de SFN*
- *L'évolution des ICP fondamentaux par rapport aux objectifs à long terme*

*Avec des outils de visualisation comme BIME, il est facile de créer des graphiques pour illustrer les données opérationnelles, ce qui permet de repérer les tendances et anomalies plus facilement, et de les communiquer de façon efficace. La mise en œuvre du système de gestion des données a également présenté des difficultés, à la fois techniques et culturelles. MicroCred recommande l'adoption d'une approche étape par étape, en commençant par quelques tableaux de bord de base et en les complexifiant au fil du temps pour obtenir des tableaux de bord plus sophistiqués.*

*Les outils de visualisation et les tableaux de bord interactifs peuvent être intégrés à des systèmes de gestion de données et fournissent des rapports dynamiques et sur mesure utiles pour les opérations, la gestion et le suivi des performances stratégiques.*

### Données utilisées dans les tableaux de bord

Il existe deux principaux niveaux d'enregistrement des données nécessaires à l'élaboration des tableaux de bord : au niveau des transactions et au niveau des clients. Ils servent des objectifs différents, mais les deux sont importants.

### Données sur les transactions

Les données sur les transactions sont caractérisées par une forte fréquence et une forte hétérogénéité. Les prestataires de SFN doivent cependant viser à normaliser la typologie des transactions afin de suivre la rentabilité des produits, de surveiller et d'analyser le comportement des client (et des agents), et de lancer des signaux d'avertissement en cas de mauvaise performance ou de faible activité. Les types de transaction doivent être clairement différenciés et doivent être facilement identifiables dans la base de données, même lorsque les transactions semblent techniquement similaires. Par exemple, une cause fréquente de confusion apparait lorsqu'il existe plusieurs façons de verser des fonds sur un compte client, telles que l'utilisation des P2P entrant, les paiements groupés ou les encaissements, mais que toutes les données sont groupées et simplement signalées comme étant des « dépôts ». Ces trois types de transaction doivent être traités séparément en raison de leur impact très différent sur les recettes - l'un est un cout direct, un autre est une source de revenus et le troisième est d'un cout potentiellement nul - et en raison de leurs implications pour la stratégie de marketing.

### Données sur les clients

Avoir un identifiant client unique est essentiel, surtout quand le tableau de bord puise ses données dans plusieurs applications. Grâce à son intégration de données, les prestataires peuvent contrôler l'intégrité des données afin d'assurer un enregistrement de données de qualité, élément nécessaire au suivi de la concentration du portefeuille, du calcul de la pénétration des produits, de la vente croisée et du suivi du personnel de vente, et de l'analyse d'autres indicateurs importants. Il existe généralement deux grands groupes de données qui doivent être enregistrées au niveau des clients : les données démographiques et financières. Des listes complètes d'indicateurs de données peuvent être consultées dans le chapitre 1.2. La combinaison de données au niveau des transactions et du client peut fournir des indications utiles sur le comportement de certains segments de la clientèle et peut conduire à une gestion optimale des performances.

### Cas d'utilisation : Gestion des performances des agents

La gestion des agents est probablement l'aspect le plus difficile d'une prestation de services financiers réussie car elle nécessite une intervention concrète régulière par une équipe de vente sur le terrain ainsi que l'appui des opérations de back-office. Il peut être difficile de diffuser des informations, car l'équipe et les agents sont dispersés géographiquement, ont différents niveaux de connectivité, et sont souvent équipés d'une technologie assez rudimentaire. Leurs besoins en données sont malgré tout nombreux. Les responsables des relations, les agrégateurs et les agents disposant de plusieurs points de vente à de nombreux endroits ont besoin d'informations sur les performances et la gestion du fonds de caisse. Les employés de la force de vente sur le terrain qui ne passent pas souvent au bureau ont besoin d'accéder aux informations à distance. L'agent a besoin d'informations sur sa propre performance en termes de nombre de transactions et de clients, de volume d'activité, d'efficacité des ventes (conversion), et de rentabilité. Des informations sur la disponibilité des services de réapprovisionnement de fonds, en particulier sur les marchés où les agents peuvent se fournir des services de gestion de fonds de caisse et de trésorerie liés à la monnaie électronique les uns aux autres, seront potentiellement utiles. Sur les marchés où opèrent des partenaires de gestion de trésorerie indépendants, les agents doivent également disposer de données sur les niveaux des fonds de caisse.

La gestion des performances des agents a besoin de données précises, directement liées aux équipes responsables de la gestion des points de vente. Les données sur les performances des agents doivent être

facilement segmentées en correspondant à la structure de l'équipe de vente ; chaque section et chaque individu peut voir ses propres performances. Il s'agit de la base sur laquelle les objectifs de performance peuvent être évalués avec précision et récompensés. Dans l'exemple ci-dessous, les équipes et les personnes responsables de chaque niveau de la hiérarchie des agents, du directeur des ventes aux représentants des ventes de district, ont besoin de données précises et en temps voulu directement liées à leurs responsabilités. Les informations les plus utiles qui peuvent être fournies à l'équipe de vente concernent les agents dont elle est responsable.

### Lacunes du suivi des agents

Il n'existe pas de réponse définitive sur le nombre optimal d'agents nécessaires pour que chaque client ait un accès relativement facile à un agent et que chaque agent ait assez de clients pour générer un revenu acceptable. Les études citent la fourchette de 200 à 600 clients actifs par agent actif comme la situation optimale pour les prestataires de SFN, en fonction des conditions du marché. Une tâche importante de l'équipe de vente est de surveiller les données sur les agents et les clients, en contrôlant la croissance et la localisation des points de vente des agents pour s'assurer qu'ils sont conformes à l'activité des clients.

### Identification des agents les plus performants

Les bons agents doivent être récompensés pour leur travail. Des incitations, notamment des activités de marketing et des commissions exceptionnelles ou des primes liées à la performance, peuvent être basées sur ces données. Il peut être très efficace d'avoir des objectifs par agent personnalisés en fonction des conditions locales du marché, et de disposer d'un moyen de montrer clairement à l'agent ses performances par rapport à ses propres objectifs et ceux de ses pairs. Les objectifs comprennent la liquidité et l'activité des clients. Une caractéristique clé d'un bon agent est qu'il se trouve rarement à court de monnaie électronique ou de fonds de caisse. Les cibles cumulées des agents agrégateurs doivent être fondées sur l'activité de gestion de la liquidité qu'ils se sont engagés par contrat à soutenir, ainsi que les performances de leur équipe d'agents.

### Identification des agents les plus fragiles

Sur la plupart des marchés, environ 80 pour cent des agents sont actifs. Cela signifie que les clients souhaitant effectuer des transactions avec les 20 pour cent restants des agents seront probablement incapables de le faire parce que le fonds de caisse est insuffisant ou qu'un agent est absent. Les agents peu performants doivent soit être amenés à un niveau acceptable ou, si cela se révèle impossible, être écartés du service. Puisque le manque de liquidité en monnaie électronique est fortement corrélé à de mauvaises performances, un paramètre clé souvent utilisé pour analyser les performances des agents est le nombre de jours en situation de « rupture de stock » par mois (c'est-à-dire que les niveaux de fonds de caisse se situent en dessous d'un certain seuil).

Ce type d'analyse de données sur les agents est très efficace, mais assez détaillé et souvent effectué manuellement, ce qui peut être lent et fastidieux. Il peut être efficace de fournir à l'équipe de vente des outils de gestion automatisée des données qu'ils peuvent utiliser sur le terrain, ainsi que des indicateurs personnalisés. L'étude de cas Zoona ci-dessous illustre bien ces points.

# CAS 6

## *Zoona Zambie - Optimisation de la gestion des performances des agents*

***Culture des données : Une approche intégrée axée sur les données des produits, services et rapports***

*Zoona est le principal prestataire de SFN en Zambie. Il offre des transactions de gré à gré via un réseau d'agents Zoona dédiés. Les services des agents comprennent l'inscription des clients, l'envoi et la réception de paiements de transfert de fonds, la fourniture de dépôts et de retraits en espèces sur les comptes et le versement de paiements groupés provenant de tiers, tels que les salaires et les paiements des pouvoirs publics aux personnes. La culture d'entreprise de Zoona est axée sur les données et charge une équipe centralisée d'analystes de données d'affiner constamment la sophistication et l'efficacité de ses services et opérations.*

### *La localisation des agents*

*Zoona a mis au point un simulateur en interne pour déterminer l'emplacement optimal des échoppes des agents. L'approche utilise la méthode de simulations de* ***Monte Carlo***[20] *pour tester des millions de scénarios possibles d'emplacements d'agents afin d'identifier quelles sont les configurations qui maximisent la croissance des entreprises. Des facteurs tels que le nombre de clients desservis par jour par agent existant et les longueurs des files d'attente sont utilisées pour déterminer la demande locale et le potentiel de croissance jusqu'à ce que la saturation soit atteinte. Pour garantir la fiabilité, les scénarios modélisés sont recoupés avec des contributions de l'équipe de vente sur le terrain qui dispose de connaissances locales sur la zone et qui sait quels sont les points de vente soumis à la plus forte pression. Dans les endroits clés, l'équipe utilise également Google Maps et des reconnaissances physiques le long des rues, en observant leur degré d'animation et en localisant des emplacements stratégiques potentiels. Par exemple, des milliers de personnes peuvent arriver à un arrêt de bus, puis se disperser dans toutes les directions ; Zoona cartographie les itinéraires les plus populaires, en créant des zones où les clients potentiels sont susceptibles de se trouver. Zoona cartographie également l'emplacement des concurrents sur ces itinéraires.*

### *Cycle de vie des agents*

*Un agent relativement nouveau sur une route principale peut ne pas être aussi productif qu'un agent expérimenté sur un marché animé, en raison de l'emplacement et du fait que l'agent a développé une clientèle fidèle. Un service de SFN solide a besoin d'agents aux deux endroits - et les objectifs fixés pour chaque agent doivent être réalistes et réalisables. Zoona analyse les données des agents afin de projeter des attentes de performances futures pour les segments des agents, par exemple les segments urbain et rural, en produisant des courbes de « performances au fil du temps » pour chaque agent, jusqu'au niveau du quartier. Cela appuie de bons ICP de gestion d'agent.*

---

[20] *Les simulations de Monte Carlo* tirent des échantillons à partir d'une distribution de probabilité pour chaque variable afin de produire des milliers de résultats possibles. Les résultats sont analysés afin d'obtenir les probabilités de survenue de différents résultats.

### *Gestion des liquidités*

*Les agents ont besoin d'une source pratique de liquidités pour les services liés aux transactions, la proximité de banques ou de guichets automatiques bancaires (GAB) est donc prise en compte dans les scénarios de localisation. La difficulté à réapprovisionner le fonds de caisse peut également être due à une concentration excessive d'agents qui, collectivement, puisent de façon excessive dans les sources de fonds de caisse et sapent la valeur du réseau d'agent local. Les simulations de Zoona examinent les deux scénarios dans le cadre de l'optimisation. En outre, comprenant que le fonds de caisse d'un agent est un facteur clé de sa performance, Zoona expérimente une solution innovante pour recueillir les soldes des fonds de caisse en espèces et en argent électronique pour aider les agents à gérer plus efficacement leur fonds de caisse. Cela fournit aux agents un accès aux outils de gestion de la performance, qui sont développés à l'aide de la boite à outils de visualisation de gestion des données QlikView. Elle fournit à Zoona des données que les agents pourraient souhaiter ne pas signaler.*

*Les analyses peuvent appuyer de nombreux aspects des opérations et du développement de produits : optimisation de placement d'agent, gestion des performances et outils qui créent des incitations pour une communication volontaire des données. Une culture d'entreprise axée sur les données est le moteur de l'intégration.*

### Gestion du back office des agents

L'équipe de back office des agents est responsable de toutes les tâches requises pour mettre en place de nouveaux agents, puis gérer leurs interactions continue en matière de SFN. Souvent, cela inclut également d'obtenir les données requises par l'équipe de vente (ci-dessus). Pour être efficaces, les membres de l'équipe ont besoin d'un grand nombre de données, notamment des rapports standard et un accès aux données pour exécuter des rapports ad hoc se penchant sur des questions spécifiques. En plus de fournir des données sur l'équipe de vente, ils doivent aussi mesurer le temps que prennent leurs nombreux processus d'entreprise afin de veiller à ce que leur équipe ait la capacité d'atteindre les objectifs de niveaux de service internes. Ce qu'on obtient en mesurant les problèmes soulevés par type et par volume, et la mesure du temps de résolution des problèmes, souvent par le biais d'un système de gestion des incidents.

### Back office des partenaires commerciaux

Aux fins de la gestion du back office, divers types de partenaires commerciaux hors agents peuvent être associés. Ceux-ci comprennent les émetteurs de facture et autres PSP, les commerçants, les organisations qui utilisent les SFN à des fins de gestion d'entreprise, notamment la paie et d'autres paiements groupés, et d'autres IF, notamment des banques et des prestataires de SFN. L'équipe de back office de gestion des partenaires commerciaux est responsable de tâches similaires dans le cadre de la gestion des agents, mais avec des exigences réglementaires différentes (et sans besoin d'une gestion de fonds de caisse). Par conséquent, les indicateurs clés dont ils ont besoin sont semblables à ceux des agents, mais avec différents processus d'entreprise et objectifs.

### Optimisation de l'efficacité des agents

Les données peuvent être utilisées de façon plus efficace par des équipes de gestion des agents quand elles disposent d'un accès mobile et en ligne à ces données. Certaines de ces tâches sont notamment les suivantes :

- Planifier la charge de travail
- Vérifier l'aspect intérieur et extérieur des points de vente des agents lors des visites de terrain.
- Mettre à jour ou vérifier l'emplacement et autres informations démographiques du point de vente
- Montrer des statistiques de performances personnalisées à l'agent directement en arrivant
- Montrer la commission reçue à ce jour et pour le mois
- Afficher les recettes obtenues des clients que l'agent dessert
- Leur permettre d'ajouter des photos à la base de données
- Compléter directement les mesures de l'enquête d'assurance qualité (AQ) de base sur les agents
- Notifier que les informations de KYC sont en transit
- Fixer de nouveaux objectifs de performance et de nouvelles incitations
- Soumettre les demandes et questions des agents sur le service directement à l'équipe des opérations
- Noter les emplacements potentiels de nouveaux points de vente d'agent

L'accès à ce type de données peut résulter sur une meilleure motivation et une meilleure réussite des agents, ainsi que sur l'amélioration de la performance globale des activités de SFN. Des questions importantes peuvent être abordées, telles que : « Quel est le montant de fonds de caisse de monnaie électronique dont les agents ont besoin ?» Pour gérer les fonds de caisses en espèces et électronique, il est utile de déterminer quelles sont les périodes les plus actives de la journée, de la semaine et du mois, et de fournir des conseils sur leurs prévisions de besoin de fonds de caisse. Il est également utile que le système ait un dispositif signalant que le fonds de caisse d'un agent passe en dessous d'un niveau minimum, et envoyant un message d'alerte automatique à la personne chargée de la gestion du fonds de caisse de l'agent. Lors d'opérations plus sophistiquées, des algorithmes peuvent être utilisés pour prédire de manière proactive la quantité de fonds de caisse dont chaque agent aura besoin chaque jour et pour les informer du solde de départ optimal, soit avant l'ouverture du point de vente, soit après sa fermeture. Cela peut également être effectué pour le montant de liquidité que l'agent devrait avoir à disposition pour assurer les services de retrait.

# CAS 7

## *FINCA RDC - Portrait d'un agent performant et application pratique des résultats*

**Collecte des données : Ajustement du processus pour obtenir de meilleures indications et une mise en œuvre réussie**

*Avec un taux de pénétration bancaire d'un peu moins de 11 pour cent, la RDC a l'un des taux d'accès aux services financiers les plus faibles d'Afrique. En 2011, l'institution de microfinance FINCA RDC a créé son réseau d'agents, employant des gérants de petites entreprises pour qu'ils offrent les services bancaires de FINCA RDC. Le réseau d'agents a augmenté rapidement, et lorsque la collecte des données des agents a commencé en 2014, il représentait plus de 60 pour cent du total des transactions de FINCA RDC. En 2017, les transactions des agents avaient augmenté pour atteindre 76 pour cent du total des transactions. La croissance se concentrait toutefois principalement dans la capitale du pays, Kinshasa, et dans l'une des plaques tournantes des échanges commerciaux du pays, Katanga. FINCA RDC a cherché à élargir le réseau dans les zones rurales et a donc développé un modèle prédictif pour identifier les critères qui définissent un agent qui réussit. Les résultats ont été intégrés dans des enquêtes de recrutement des agents qui ont permis à FINCA RDC de choisir de bons agents dans des zones en expansion. De plus, la disponibilité d'un réseau d'agents prospère que les clients peuvent utiliser pour rembourser leurs prêts avec commodité favorise la réduction du risque de portefeuille de FINCA RDC.*

*Le modèle prédictif a défini les « agents performants » en termes de nombre, mais aussi de volume de transactions. Les données du modèle linéaire généralisé (MLG) provenaient de trois sources principales :*

- ***Formulaires d'inscription des agents :*** *Ceux-ci fournissent des informations sur les données commerciales et socio-démographiques du chef d'entreprise.*
- ***Formulaires de suivi des agents :*** *Des employés de FINCA RDC suivent régulièrement les agents, en recueillant des informations sur la trésorerie et le fonds de caisse électronique de l'agent, l'état de sa boutique, des données d'opinion sur l'interaction des clients de l'agent et les affichages de la marque du produit FINCA RDC. Ces informations sont ensuite rassemblées dans une note de suivi.*
- ***Données sur les transactions des agents :*** *Ces données sont des informations sur le volume et le nombre de dépôts, de retraits et d'opérations de transfert effectués par chaque agent.*

*La disponibilité des données et la qualité des données ont été les principaux défis dans le développement du modèle de performance des agents. Les données numérisées sont exigées pour des sources habituellement recueillies sur papier, telles que l'inscription de l'agent et les formulaires de suivi. Les données manquantes doivent être réduites au minimum, tant pour garantir des séries de données plus solides que pour permettre la fusion des séries de données en faisant correspondre des champs de métadonnées. Cela nécessite une normalisation des données recueillies par différentes personnes, qui peuvent utiliser différentes méthodes de collecte. Le manque de*

*données cohérentes peut conduire à une réduction significative de l'échantillon, ce qui compromet la précision et la performance de prévision du modèle.*

*Les agents qui réussissent en RDC sont identifiés par les critères statistiquement significatifs suivants : la situation géographique, le secteur de l'activité principale d'un agent, le sexe de l'agent, et s'ils réinvestissent leurs bénéfices. Il est avéré que les femmes agents, par exemple, font 16 pour cent plus de profit sur leurs activités d'agent que leurs homologues masculins ; la valeur des stocks de leur entreprise est 42 pour cent plus élevée. On a également découvert qu'elles réinvestissaient plus d'argent dans les stocks de leur entreprise, plutôt que de le garder sur un compte bancaire qui rapportait peu intérêt. Cela a débouché sur environ 5 pour cent de plus de valeur de transaction moyenne totale par mois.*

*Ces résultats ont été mis en œuvre pour améliorer et rationaliser le processus de sélection d'agents, ce qui a finalement contribué à élargir le réseau à des zones rurales en intégrant des facteurs dans les enquêtes sur les agents et la stratégie de déploiement.*

*En 2016, le réseau d'agents avait augmenté pour représenter 70 pour cent du total des transactions. Le modèle a identifié la localisation comme le critère clé, révélant ainsi une autre possibilité d'étude. Pour l'étude de suivi, FINCA RDC et IFC utiliseront une méthodologie d'ERC pour identifier la localisation optimale des agents.*

*La comparaison des données des profils des agents par rapport aux paramètres des agents peut mettre en évidence les principales caractéristiques qui conduisent à l'amélioration de leurs performances. L'intégration de ces apprentissages en matière de ciblage d'agents et de processus de gestion assure la pleine mobilisation des données au service de la gestion des performances.*

## Cas d'utilisation : Gestion du back office

### Automatisation de processus

Même si les prestataires de SFN consacrent beaucoup d'efforts au développement de l'automatisation frontale (services bancaires mobile et en ligne), certains ont encore du mal à développer des fonctions d'arrière-plan fortement automatisées. Les tâches automatisées qui peuvent aider les opérations de back office, telles que la souscription et la constitution de dossier de prêts, le traitement et le rapprochement automatisés des transactions, ont une immense valeur. Les prestataires se dirigent maintenant vers l'automatisation robotique des processus simples et répétitifs, qui peuvent être réalisés à un bien moindre cout et avec bien plus de précision par des machines que par des humains. Selon AT Kearney, l'automatisation des processus robotiques (RPA, Robotic Process Automation) rend les opérations 20 fois plus rapide que la moyenne des humains et offrent des économies de couts de 25 à 50 pour cent à ceux qui l'adoptent.[21] Différents domaines d'automatisation peuvent généralement être regroupés dans l'automatisation de l'enregistrement des données et du traitement de données.

L'objectif principal de l'enregistrement des données réside dans la numérisation des flux de travail basés sur papier. Nous observons que de nombreux prestataires utilisent encore des formulaires d'inscription papier pour recueillir des informations d'ouverture de compte. Les nombreuses erreurs qui se produisent lors du processus de saisie manuelle font que ces formulaires doivent passer par de multiples boucles de remaniement. Finalement, après être passées par un processus de vérification en plusieurs étapes, les informations clé sont manuellement enregistrées dans le système par le front ou le middle office, ce qui crée une charge de travail supplémentaire pour le personnel et nuit à l'efficacité de l'aménagement du temps. Ces formulaires doivent ensuite être stockés dans un entrepôt physique et conservés pendant un certain temps. La rationalisation et la simplification du processus de collecte de données par l'interface frontale avant et à travers un système de contrôle des données intégré améliorent l'efficacité et réduit les couts de main-d'œuvre. Bien sûr, pour enregistrer les données de manière fiable, l'architecture informatique doit être suffisamment solide pour classer, vérifier et stocker correctement les données.

Le traitement des données peut être automatisé à presque toutes les étapes de la relation client. L'établissement d'étapes de vérification standard peut accélérer les ouvertures de comptes et les modifications de comptes, et les décisions de crédit pour certains segments peuvent être déclenchées par des modèles bien structurés et testés. En outre, des cartes thermiques pratiques peuvent automatiser les décaissements, et des formulaires de demande et de commentaires automatisés peuvent numériser les fermetures de compte. Des analyses évoluées, décrites dans le chapitre précédent et pouvant inclure la génération de pistes pour des campagnes de vente ou la gestion multicanal, peuvent être utilisées pour révéler des débouchés inexploités et des risques dans le portefeuille. Une fois identifiés, des notifications automatiques peuvent être envoyées au personnel du front office ou directement aux clients. Par exemple, pour empêcher le désabonnement, les clients qui s'approchent de l'état d'inactivité peuvent recevoir des messages textuels ou des e-mails de réactivation. Les emprunteurs peuvent recevoir des notifications sur les paiements à venir ou des produits à meilleurs prix disponibles pour le refinancement. Certaines fonctions nécessitant des interventions humaines, telles que l'analyse financière et commerciale et la gestion des relations personnelles, complètent le processus automatisé et en bénéficient.

### Surveillance des risques et de la conformité réglementaire

À la suite de la crise financière de 2008, les organismes de réglementation nationaux ont continuellement durci la réglementation du secteur financier afin de protéger les clients et le secteur en général. L'augmentation des exigences de fonds propres, de liquidités et de transparence a placé un lourd fardeau sur le secteur financier réglementé tout en créant un avantage concurrentiel pour les acteurs non réglementés, tels que les prestataires de technologie financière. En conséquence, les banques doivent prévoir des couts de conformité dans leur budget pour se

[21] 'Robotic Process Automation : Fast, Accurate, Efficient', A.T. Kearney, accédé le 3 avril 2017 https://www.atkearney.com/financial-institutions/ideas-insights/robotic-process-automation

conformer aux exigences réglementaires. La présentation de rapports réglementaires nécessite la mise en commun de données provenant de divers systèmes, notamment la comptabilité financière, le système de comptabilité, la trésorerie, le contrôle de la qualité des actifs et les bases de données de collectes, entre autres. Des simulations régulières de crise nécessitent une infrastructure informatique solide avec une importante capacité de stockage et de traitement de grandes quantités de données. En outre, la conformité à la KYC exige des flux de données concrets pour une prise de décision en temps voulu et en toute sécurité. Les données nécessaires à la mesure et la surveillance du marché, du crédit, de la LBC et des risques de liquidité sont idéalement stockées dans un lieu central unique pour permettre à un prestataire de SFN d'avoir une image complète des risques de la totalité de son portefeuille. Ce lieu central unique permet également au prestataire de SFN d'exécuter des analyses de scénarios et des simulations de crises pour répondre aux exigences réglementaires. La conformité réglementaire entraine des couts directs du fait d'une hausse du cout du capital, ainsi que des couts indirects, tels que l'établissement de processus d'établissement de rapports, l'aménagement de l'emploi du temps du personnel et, dans certains cas, les investissements dans les nouvelles technologies.

### Prévention de la fraude

Les tendances mondiales s'orientant vers le Cloud computing, la gouvernance et la protection des données deviennent de plus en plus importantes. Les prestataires de SFN doivent accorder plus d'attention au comportement des clients en matière de transactions. Ils doivent également être en conformité avec la KYC afin de détecter les activités frauduleuses potentielles- comme le blanchiment d'argent et l'usurpation d'identité - tout en évitant ou en réduisant les risques opérationnels et financiers. De nouvelles interventions et règlementations en matière de cyber sécurité exigeront des prestataires de SFN qu'ils développent et maintiennent des outils visant à se protéger de menaces extérieures et de potentielles activités criminelles. Le maintien et l'agrégation des données appropriées nécessaires pour assurer la prévention de la fraude et les modèles de risques opérationnels peuvent réduire l'exposition des prestataires de SFN. La gestion continue des flux et le traitement des données en temps réel leur permet de détecter les fraudes plus rapidement et avec une plus grande précision, réduisant ainsi les risques potentiels de pertes. Par exemple, si les cartes de crédit ou de débit d'un client sont utilisées depuis un point géographique inhabituel ou à une fréquence inhabituelle, les prestataires de SFN peuvent en alerter le client et éventuellement bloquer le traitement de ces opérations suspectes.

## Suivi des données pour détecter les fraudes

Lorsque les prestataires de SFN offrent des services de P2P, les prestataires peuvent utiliser divers outils pour déterminer si les montants de transactions sont versés frauduleusement sur le compte de quelqu'un d'autre pour éviter les frais. Au lieu d'utiliser leur propre compte et de payer les frais, les clients peuvent effectuer un transfert P2P payé à partir du compte de quelqu'un d'autre. La vitesse de transaction peut donner une indication de base ; si l'argent est déposé sur un compte, puis retiré à nouveau dans une période de temps très court, il y a de bonnes chances qu'il s'agisse d'un dépôt direct. Le lieu de la transaction donne une indication encore meilleure car si la localisation des agents qui effectuent le dépôt et le retrait est à une certaine distance, il est peu probable, voire impossible, que le client ait pu parcourir la distance entre ces points dans l'intervalle séparant les deux transactions. Il doit être possible de créer des alertes pour ce genre de comportement, et des agents qui effectuent un nombre exceptionnellement élevé de dépôts directs peuvent être suivis. Cela ne comptabilisera pas les transactions entre les clients vivant à proximité, donc de nombreux prestataires de SFN ont également recours aux achats anonymes effectués par des enquêteurs pour mieux comprendre les niveaux de dépôt direct.

## Cas d'utilisation : Gestion des interactions des utilisateurs

La gestion des clients dans tout le cycle de vie, l'encouragement à une utilisation accrue et la gestion des nouveaux comportements relèvent de la compétence de l'équipe de marketing. Toutefois, il existe aussi un aspect opérationnel à la gestion des clients qui relève principalement des équipes de service à la clientèle, des risques et techniques. Ces équipes sont chargées de veiller à ce que l'interaction des utilisateurs soit telle que conçue, en détectant et réglant tous les problèmes. Elles sont également responsables de la gestion de l'interaction des utilisateurs pour les clients professionnels et les utilisateurs internes.

À cet égard, il est important de définir l'utilisation et le comportement « normaux » prévus du système de sorte que des prévisions puissent être faites pour la planification technique et commerciale. Des mesures sont généralement définies du haut vers le bas, telles que les objectifs commerciaux mensuels et les objectifs stratégiques. Cela dit, certains indicateurs de résultats doivent être recueillis « du bas vers le haut » tels que les mesures de l'utilisation moyenne d'un service. Comme indiqué précédemment, l'utilisation des moyennes peut être trompeuse, et le comportement peut devoir être décomposé en secteurs, puis regroupé en une « vue moyenne » de l'activité en fonction de laquelle des projets peuvent être établis. Par exemple, l'équipe technique a besoin de connaitre aussi bien le nombre d'opérations par jour que les périodes qui verront probablement une forte activité, pour qu'ils puissent s'assurer que le système peut supporter les pics d'activité.

La définition de modèles de « comportement normal » est un aspect fondamental de la gestion des risques. Les modèles d'activité qui s'écartent des normes convenues, en particulier les données d'utilisation des transactions et des services, doivent être signalés. Ces modèles doivent être examinés pour déterminer si le comportement inhabituel était légitime, ou s'il s'agit d'un cas de fraude potentielle. En plus du comportement des clients et des agents, il est également conseillé de définir ce qu'est une « activité normale » quant aux interactions des employés dans le système. Par exemple, un employé consulte-t-il effectivement davantage de dossiers de clients qu'un employé « normal » ayant la même fonction, ou accède-t-il au système en dehors de ses heures de travail habituelles ? Cette activité anormale pourrait indiquer une activité frauduleuse.

## Améliorations de l'efficacité du service client

Les équipes de service clients dans les centres d'appels sont les employés les plus proches du client des SFN au jour le jour. De fait, ils peuvent signaler de manière précoce toutes les questions importantes susceptibles de survenir. Ils seront souvent les premiers à apprendre une panne du système ou le comportement frauduleux d'un agent, un processus est donc nécessaire pour alerter l'équipe appropriée d'un problème éventuel en fonction des informations (dont l'aspect raisonnable est vérifié) reçues des clients. Ces équipes sont également susceptibles d'entendre parler de problèmes mineurs affectant le service qui empêchent les clients d'effectuer des transactions de façon optimale, tels que le manque d'agents, des limites de transaction restrictives et de courtes interruptions lors de transactions. Il est donc important de recueillir des données statistiques sur les appels reçus, notamment les plaintes et suggestions. Les façons de tirer parti de ce type de données sont illustrées dans le Cas 8 ci-dessous.

Le suivi du nombre d'appels à mesure que le service se développe permet de déterminer le nombre de représentants nécessaires au centre d'appel. Pour certains services à grande utilisation, seule une partie des appels passés arrivent effectivement à une ligne de service clients. Dans ce cas, les tentatives d'appels par rapport aux appels traités constituent un chiffre important, car il indique soit un problème majeur, soit une insuffisance du personnel. Les problèmes de centre d'appels les plus fréquemment rapportés sont les oublis de PIN, les téléphones ou cartes perdus, les transactions envoyées aux mauvais destinataires et la perte de codes de bons promotionnels. Le nombre d'appels qui peuvent être pris dépend de la vitesse du système de back office et la rapidité avec laquelle il peut réagir pour résoudre le problème. Alors que les couts des centres d'appels sont généralement élevés, les données qu'ils fournissent doivent être utilisées pour accélérer le processus de résolution des problèmes et augmenter le nombre d'appels que chaque représentant peut prendre. Ces données peuvent également être utilisées pour améliorer l'expérience utilisateur afin que le client fasse moins d'erreurs.

# CAS 8

## *Safaricom M-Pesa - Utilisation d'ICP pour améliorer le service clients et les produits*

***Utilisation des analyses de données pour identifier les goulots d'étranglement opérationnels et prioriser des solutions***

*M-Pesa au Kenya a été le pionnier des SFN à grande échelle, avec 20,7 millions de clients, une base active sur trente jours de 16,6 millions,[22] et des recettes déclarées en 2016 de 4,5 milliards d'USD.[23] Lorsque Safaricom a lancé le service en 2007, il n'existait pas de modèles ou de meilleures pratiques ; tout a été conçu à partir de zéro. L'amélioration opérationnelle continue a été essentielle au fur et à mesure que le service a été mis à l'échelle.*

*L'intérêt suscité pour le service était étonnamment élevé dès le début, avec plus de 2 millions de clients lors de sa première année, battant les prévisions de 500 pour cent. Cette demande croissante a forcé une rapide mise à l'échelle et les opérations nécessaires pour anticiper de manière proactive les problèmes de mise à l'échelle en matière de technologie et de processus d'entreprise, car une mauvaise expérience client pouvait rapidement éroder la confiance des clients. Des indicateurs fondés sur des données ont permis à l'équipe de planifier et d'orienter les opérations de manière appropriée.*

*Comme l'intérêt suscité par le service était étonnamment élevé dès le début, le nombre d'appels au centre d'appels du service clients a été, dans la même proportion, plus élevé que prévu, ce qui a entrainé un volume élevé d'appels sans réponse. Ce problème a été à l'origine d'un ICP dont l'équipe du service clients avait besoin pour le résoudre et ramener la situation à un niveau acceptable.*

*Le problème a tout d'abord été abordé en recrutant du personnel supplémentaire, mais le recrutement en lui seul n'a pas permis de suivre le rythme de l'augmentation du nombre de clients. Pour identifier les goulots d'étranglement et hiérarchiser les solutions, l'équipe a analysé leurs données. Les historiques de données d'appel des autocommutateurs de résolution des problèmes ont été examinés et il a été constaté ce qui suit :*

- ***Durée de l'appel** : La durée moyenne des appels était de 4,5 minutes, environ deux fois la durée du temps prévu au budget pour chaque appel.*
- ***Questions clés pour une résolution rapide** : Les deux types d'appel clés à aborder pour l'optimisation émanaient de clients ayant oublié leur PIN et de clients qui envoient de l'argent au mauvais numéro de téléphone ; cela représentait 85 à 90 pour cent des appels de longue durée arrivant au centre d'appels.*

*L'analyse a permis de réaliser deux choses. Tout d'abord, les goulots d'étranglement ont été correctement identifiés, en passant des indications clés aux opérations. D'autre part,*

---

[22] Richard Mureithi, « Safaricom announces results for the financial year 2016. » *Hapa Kenya*, 12 mai 2017, consulté le 3 avril 2017, http://www.hapakenya.com/2016/05/12/safaricom-announces-results-for-the-financial-year-2016/

[23] Chris Donkin, « M-Pesa continues to dominate Kenyan market. » *Mobile World Live*, 25 janvier 2017, consulté le 3 avril 2017, https://www.mobileworldlive.com/money/news-money/m-pesa-continues-to-dominate-kenyan-market/

*d'autres problèmes opérationnels ont été découverts, surtout la fréquence à laquelle les clients avaient envoyé de l'argent par erreur ou avaient oublié leur PIN. La gestion en fonction de l'ICP des appels sans réponse a donc apporté des bénéfices opérationnels plus généraux.*

*En utilisant les résultats analytiques, les opérations ont mis en œuvre une stratégie de résolution. Tout d'abord, en dressant une typologie des problèmes longs à résoudre par rapport aux problèmes courts, les problèmes difficiles pouvaient être rapidement identifiés et rapidement passés à une équipe de back office. Cela a réduit les temps d'attente des clients et les goulets d'étranglement, permettant de traiter davantage de clients par jour. En second lieu, les opérations et les équipes de développement de produits ont travaillé pour réduire les délais pour tous les types d'appels. Pour ce faire, l'infrastructure technique et l'interface utilisateur ont été améliorées, atténuant ainsi les problèmes responsables des appels longs. Une série d'initiatives conjointes a réduit l'ICP de durée d'appel et la valeur de l'ICP d'appels sans réponse, les plaçant tous les deux à des niveaux acceptables malgré le nombre de clients qui continuait de croitre au-delà des niveaux prévus.*


*La gestion par le biais d'ICP est un élément essentiel des opérations. L'analyse détaillée des données à l'origine des ICP peut permettre d'identifier les goulots d'étranglement opérationnels, et peut même révéler d'autres facteurs opérationnels qui poussent les indicateurs au-delà des seuils. Comprendre les données qui sont à l'origine d'un ICP peut permettre de mieux l'utiliser.*

## Cas d'utilisation : Données sur les opérations techniques

Par sa nature même, un service de SFN doit être disponible 24 heures sur 24, sept jours sur sept, et il est normalement conçu pour traiter de grands volumes d'interactions du système, à la fois financières et non financières. Pour cette raison, le service doit être surveillé de manière proactive en prenant des mesures préventives pour assurer la disponibilité continue du service. Les données de diagnostic de service sont généralement utilisées pour effectuer cette analyse. Les tableaux de bord de performance technique doivent être mis à jour en temps réel pour montrer l'état de santé du système. Ils doivent être surveillés automatiquement et conçus pour alerter les fonctions et les personnes responsables si un problème potentiel est repéré. Le concept de l'utilisation des données pour « comprendre une situation normale » est utilisé pour détecter de manière proactive des défaillances dans différentes couches du service et des solutions de surveillance automatiques sont mises en place pour détecter les cas de dépassement des paramètres de seuil. Par exemple, si un système de SFN traite normalement un certain nombre de transactions par seconde (TPS) tous les jeudis soir, mais qu'un jeudi, le chiffre est beaucoup plus faible, cela signale qu'il existe probablement un problème qui nécessite d'être résolu.

Les tendances peuvent être utilisées pour prédire des problèmes de performances tout en identifiant des incidents spécifiques ; de fait, l'équipe doit également tenir compte des performances au fil du temps. L'analyse des tendances est essentielle pour planifier les capacités, et des modèles d'utilisation et de croissance du système donnent des indices importants sur les moments où une capacité de système supplémentaire sera nécessaire. Que le système soit externalisé ou en développement interne, il est important que l'équipe technique surveille les niveaux de service et les tendances en termes de capacité, et prévoie des mesures correctives. Les principales données normalement requises comprennent la disponibilité du système, les temps d'arrêt prévus et imprévus, le volume des transactions et la capacité, tant en cas de pic que permanente.

### Transactions et interactions

Une transaction est un mouvement financier d'argent, en général l'acte de débiter un compte et de créditer un autre. Pour y arriver, l'utilisateur doit interagir avec le système. Ces interactions peuvent donner des indications, et sont fréquemment utilisées en développement de produits numériques pour des smartphones et des services Web afin de permettre de mieux comprendre le client.

Les interactions de SFN, même en utilisant des téléphones de base, peuvent être mesurées et peuvent fournir des données utiles sur l'expérience client pour un service. Par exemple, il est possible de mesurer les interactions telles que les « tentatives abandonnées d'effectuer une transaction financière » qui ont permis de diagnostiquer ce qui empêchait alors les clients d'effectuer ces opérations. Un autre exemple est le cas de services à la clientèle qui interagissent avec le système pour le compte d'un client, par exemple en réinitialisant un code PIN oublié. Ces interactions sont rarement mesurées, mais elles peuvent également fournir des indications utiles pour améliorer les opérations du service.

Des services de SFN efficaces établissent une bonne communication entre les équipes commerciales et techniques. L'équipe commerciale doit discuter de manière proactive de ses plans de marketing et des prévisions ainsi que toute activité concurrentielle afin de préparer l'équipe technique à des changements de volume potentiels. Des réunions régulières (au moins tous les trimestres) sont nécessaires pour examiner les dernières prévisions de volumes en fonction des résultats du trimestre précédent et de l'activité de marketing prévue. Cela permet à l'équipe technique d'établir une planification en conséquence. L'équipe technique doit, à son tour, conseiller tous les partenaires qui pourraient être touchés par un changement dans les prévisions. Ceci est particulièrement pertinent pour les partenaires d'ORM, car il a existé plusieurs cas impossibles à gérer de besoins en volume de SMS lors d'opérations de promotion particulièrement réussies. De même, si des changements ou des révisions techniques sont prévus, le marketing doit en être informé et éviter des activités qui pourraient créer une pression supplémentaire sur le système à ce moment-là.

## Leçons tirées de la gestion des opérations et des performances

**Documenter l'avantage que les ventes de temps de communication représentent au niveau commercial :** Les rapports peuvent induire en erreur lorsque les clients utilisent les SFN pour acheter du temps de communication. En fonction de l'activité principale du prestataire de SFN, la vente de temps de communication prépayé peut être soit une source de revenus, soit une réduction des couts.

Pour les organisations qui ne sont pas des ORM, chaque vente de temps de communication rapporte une petite commission, car elles agissent en tant que distributeurs de temps de communication. Ce revenu doit être considéré comme faisant partie des revenus des SFN. Pour les ORM, plutôt que des revenus, cette opération représente une réduction des couts ayant un impact significatif car elle élimine les 2 à 3 pour cent (habituels) des commissions et des couts de distribution. Cependant, bien des ORM n'attribuent pas ces économies de couts à l'entreprise de SFN car elles n'ont pas été prises en compte dans la ligne budgétaire de temps de communication prépayé. Bien que cela puisse être correct en termes comptables, pour évaluer avec précision la valeur des SFN pour l'entreprise, ces économies de couts doivent être incluses dans les comptes de gestion interne des SFN.

**Se méfier des moyennes :** Par nature, les offres de SFN ont tendance à attirer les personnes ayant des ressources limitées qui n'ont pas accès à des banques et les personnes (et les entreprises) plus aisées qui interagissent avec eux. Cela conduit à des volumes très élevés de transactions de faible valeur parallèlement à un petit nombre de transactions de relativement grande valeur. La visualisation des données peut être très efficace pour identifier les cas où l'utilisation des moyennes est inappropriée. Par exemple, la figure 16 montre une courbe de fréquence de distribution typique des valeurs de transaction pour un prestataire de SFN, la majorité des transactions étant de 20 USD. La valeur moyenne des transactions est pourtant de 86 USD, car un nombre relativement faible de transactions de grande valeur biaise la moyenne. Ces moyennes peuvent conduire à une vision erronée et artificiellement élevée de la richesse et de l'activité financière « moyennes » du client.

**Regardez les tendances à plus long terme et les résultats à court terme :** indications beaucoup plus intéressantes qu'un point de données isolé. Les changements doivent être compris dans la durée, car il peut exister un effet saisonnier, par exemple un jour férié, responsable d'un pic d'activité. Ce pic peut être suivi d'un plongeon, puis un retour au statu quo, ce qui est courant aux alentours de Noël. Il peut également y avoir un impact saisonnier ; par exemple, pendant la saison des récoltes, les agriculteurs de cultures de rente obtiennent la majorité de leur revenu annuel et sont beaucoup plus actifs financièrement par rapport à d'autres périodes de l'année. D'autres causes des changements à court terme dans les performances peuvent être dues à l'activité concurrentielle, les conditions météorologiques extrêmes et l'incertitude politique.

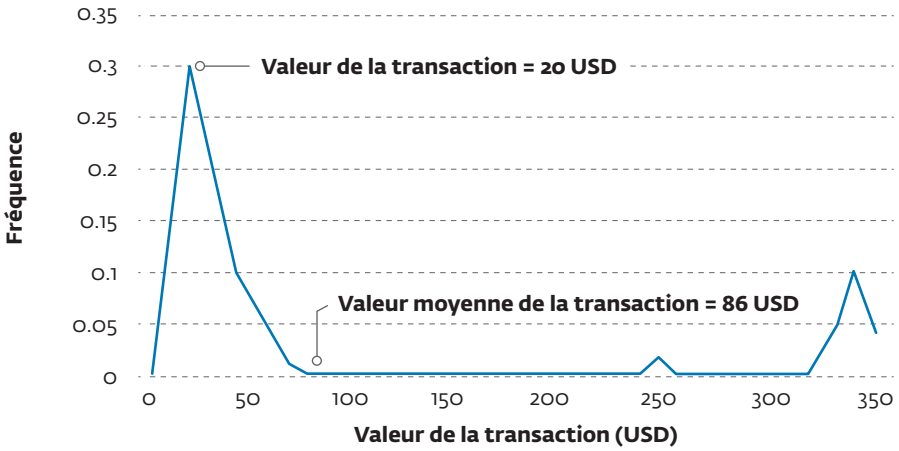



*Figure 16 : Graphique de fréquence des valeurs de transaction montrant que les moyennes peuvent conduire à de mauvaises conclusions*

**Prenez garde aux indicateurs flatteurs :** Les indicateurs flatteurs peuvent donner une bonne impression sur papier, mais ils peuvent donner une vision biaisée de la performance des entreprises. Ils sont faciles à manipuler et ne sont pas nécessairement corrélés aux données qui comptent vraiment, comme l'engagement, le cout d'acquisition et, en fin de compte, les recettes et les profits. Un exemple typique d'indicateur flatteur de SFN est le nombre de clients inscrits, plutôt que de ceux qui sont actifs. De même pour l'indication du nombre total d'agents au lieu du nombre d'agents actifs. Ce n'est qu'en se concentrant sur les véritables ICP et les indicateurs critiques qu'il est possible de bien comprendre la santé de l'entreprise. Si une entreprise s'axe sur des indicateurs flatteurs, elle peut avoir une fausse idée de sa réussite.

**Les données de niveau de service doivent être pertinentes par rapport aux objectifs commerciaux :** Chaque équipe opérationnelle rassemble une profusion de données sur la façon dont le système fonctionne. Cependant, dans le cadre de SFN complexes impliquant plusieurs partenaires, elles peuvent ne pas tenir compte de la performance du service de bout en bout et de son effet sur l'expérience utilisateur. Pour un client, l'indicateur de performance pertinent est la performance des transactions de bout en bout ; la transaction a-t-elle été achevée, et combien de temps cela a-t-il nécessité ? Il est surprenant de voir combien peu de SFN mesurent cette performance de transaction de bout en bout compte tenu de son rôle central dans l'établissement et le maintien de la confiance des clients, établissant ainsi une acceptation du SFN et maintenant la réputation de l'entreprise. La figure 17 illustre le problème posé à un client consistant à payer une facture avec son téléphone. Dans ce cas trois « propriétaires du système » sont impliqués : une ORM fournissant la connectivité, le SFN fournissant la transaction et l'émetteur de facture payé.

Chaque système renvoie ses propres données sur l'efficacité, mais l'expérience client peut être tout à fait différente s'il existe des retards de passation entre les systèmes. Un autre exemple courant est le cas où l'ORM fournit des sessions de données de services supplémentaires non structurées (USSD) avec un délai de temps mort trop court ou une défaillance sous forme de décrochage des USSD, de sorte que certains clients ne peuvent physiquement pas effectuer une transaction dans le délai imparti.

Il devrait être facile dans une relation prestataire-fournisseur de demander des données qui montreront des informations pertinentes, par exemple les décrochages des USSD ou les files d'attente des transactions. Toutefois, le fait qu'il n'existe pas d'accords directs ou complets de niveau de service (SLA), ce qui peut parfois rendre la compréhension précise de l'information impossible, est souvent un problème essentiel pour fournir des SFN.

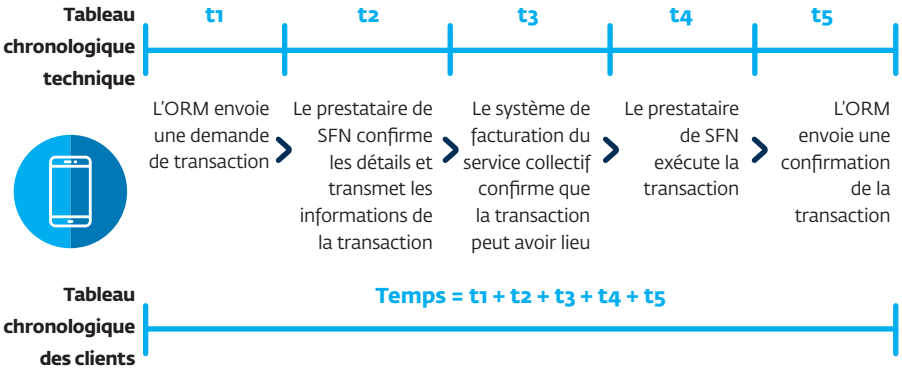



*Figure 17 : Heure de la transaction : Mesures du système par rapport à l'expérience client*

**Filtrer le déluge de données :** Chaque interaction avec un système de SFN peut générer un grand nombre de points de données. Certains d'entre eux sont d'ordre financier, et certains enregistrent quelle interface est utilisée, ou même combien de temps il faut à l'utilisateur pour se repérer dans l'expérience utilisateur. L'intensité des informations recueillies augmente considérablement à mesure que les systèmes ont recours à des interfaces utilisateur plus évoluées, telles que les smartphones. Cela peut conduire à une surcharge d'information et à des « défaillances de filtre », c'est-à-dire, pour résumer, voir les arbres qui cachent la forêt. Ceci, ainsi que les contraintes concernant la sécurisation des ressources nécessaires pour gérer ces nouveaux flux de données, est la raison pour laquelle si peu de ces informations sont utilisées par l'entreprise pour la prise de décision. Rassembler et corréler des informations externes avec des données internes peut conduire à une perte d'indications clés.

# CAS 9

## *M-Kopa Kenya - Modèles économiques innovants et stratégies axées sur les données*

***Une culture d'entreprise fondée sur les données intègre les analyses dans toutes les opérations, tous les produits et tous les services***

*Créée au Kenya en 2011, M-Kopa a commencé en tant que prestataire de systèmes énergétiques domestiques à l'énergie solaire, principalement pour l'éclairage et la recharge des petits objets comme les téléphones portables et les radios. L'entreprise combine des technologies de machine à machine, en utilisant des cartes SIM intégrées avec une solution de SFN de micropaiement, ce qui signifie que la technologie ne peut être suivie et mise à disposition que lorsque le paiement anticipé est reçu. Les clients achètent les systèmes M-Kopa en utilisant des « crédits » via le service d'argent mobile M-Pesa, puis payent pour les systèmes en utilisant M-Pesa jusqu'à ce que le solde soit entièrement payé et que le produit soit acquis. Ces dernières années, l'entreprise s'est étendue à d'autres domaines, notamment la fourniture d'appareils ménagers et de prêts, en utilisant des unités solaires appartenant à la clientèle comme garantie de refinancement. Ces produits sont offerts aux clients qui ont atteint un indicateur de cote de crédit de type « capacité de payer », évalué par leur achat initial de système et le remboursement ultérieur. M-Kopa est maintenant également disponible en Ouganda, en Tanzanie et au Ghana.*

*M-Kopa utilise des données de manière proactive dans toute l'entreprise pour améliorer l'efficacité opérationnelle. Ses bases de données amassent des informations sur les caractéristiques démographiques des clients, le degré de dépendance de l'appareil des clients et le comportement de remboursement. Chaque unité solaire transmet automatiquement des informations d'utilisation des données et de diagnostic des systèmes à M-Kopa, en les informant quand, par exemple, les lumières sont allumées. Tout cela peut être analysé pour améliorer la qualité du service, l'efficacité opérationnelle et la compréhension du comportement des clients.*

### *Gestion des capacités techniques*

*Une analyse de l'utilisation et du comportement de remboursement des clients montre que les utilisateurs préfèrent acheter des crédits à l'avance afin d'obtenir une alimentation électrique fiable pour les jours à venir. En sachant quand les clients sont susceptibles de payer (et combien de temps à l'avance), M-Kopa peut prévoir les attentes et planifier en conséquence, en s'assurant que leurs clients ne seront pas affectés par des interruptions de service annoncées de M-Pesa qui pourraient empêcher ces paiements d'être effectués.*

### *Service clients*

*Les appareils M-Kopa communiquent les données de la batterie quand celle-ci est mise en service et l'analyse des données permet au service clients de vérifier si les unités fonctionnent comme prévu et permettent une maintenance proactive et préventive pouvant être effectuée à distance :*

- *Si un client se plaint qu'il ne reçoit pas la quantité attendue d'électricité, les tableaux de bord des batteries sont utilisés pour diagnostiquer le problème. Par exemple, si la batterie ne se charge pas complètement pendant les heures de la journée.*
- *En dépit de bons contrôles de qualité lors de la fabrication, il existe toujours des variations dans la performance des batteries lorsque les unités sont sur le terrain, qui dépendent de facteurs tels que les modèles d'utilisation ou les conditions environnementales. M-Kopa a créé des algorithmes de maintenance prédictifs pour détecter les performances sous-optimales d'une batterie, ce qui lui permet d'intervenir et de prendre des dispositions pour un remplacement gratuit avant que la « panne » de batterie se produise.*

### *Gestion de l'équipe de vente*

*L'équipe de vente sur le terrain vend des produits et services M-Kopa directement aux clients. Les représentants des ventes utilisent une application sur smartphone pour archiver toutes leurs activités numériquement, en temps réel. Cela permet une compréhension détaillée de leur performance et une rapidité de réaction pour traiter ces problèmes. Les mesures dynamiques de performances en ligne et les classements peuvent être ventilés par personne et sont à la disposition de l'équipe de gestion des ventes et des chefs d'équipe afin d'encourager l'amélioration des performances grâce à la ludification.[24] L'application permet également aux membres de l'équipe de suivre leur commission ainsi que tous bonus et incitations supplémentaires.*

### *Cibler les clients susceptibles de générer des ventes supplémentaires*

*Le comportement de remboursement des clients peut fournir bon nombre d'informations sur la santé financière et la solvabilité. Les données de la batterie montrent à quel point le client dépend d'un service pour l'éclairage, ce qui permet un niveau de compréhension plus approfondi. Ces informations sont utilisées pour identifier et cibler activement les clients existants pour les mises à niveau et les services supplémentaires. M-Kopa partage également ces informations avec les bureaux de crédit pour permettre de fournir aux clients une notation du risque de crédit.*

*Une culture d'entreprise axée sur les données est nécessaire pour intégrer des analyses et des rapports dans l'ensemble de l'entreprise. Cela permet de tirer parti de sources et d'analyses de données dans plusieurs domaines afin d'attirer de nouveaux clients, de gérer des équipes de vente, d'offrir un meilleur service clients et développer de nouveaux produits.*

---

[24] *La ludification* est l'application d'éléments de jeu et de principes de jeu dans des contextes hors-jeu. D'autres exemples dans le cadre des SFN peuvent être consultés dans des études sur le site Web du CGAP : https://www.cgap.org/blog/series/gamification-and-financial-services-poor/

**Interactions des systèmes de stockage :** Il y a seulement quelques années, lorsque du lancement de nombreuses offres de SFN, la saisie et le stockage de données étaient relativement couteux et lourds, et donc des données qui n'étaient pas immédiatement nécessaires pour gérer une entreprise n'étaient pas conservées. La nouvelle technologie permet le stockage de données abondantes à moindre cout. Bien que souvent ignorés, il existe également de nouveaux outils pour analyser des données qui se trouvent sur des historiques de serveurs et permettent, avec les bons outils, d'établir une corrélation entre plusieurs sources de données pour fournir des informations plus intéressantes sur les services. Il est fortement recommandé que les prestataires de SFN recueillent et stockent tout élément de données possible sur toutes les interactions du système, même celles qui ont été refusées par le passé. Bien que cela ne semble pas être utile ou pertinent pour les opérations en cours, elles pourraient bien représenter une valeur à une date ultérieure pour effectuer des analyses de données plus poussées ou une enquête sur une fraude.

Les principes de non-répudiation exigent que ces modifications soient enregistrées en tant qu'événements supplémentaires, plutôt que de tenter de modifier des enregistrements précédemment finalisés. Par exemple, si une commission doit être récupérée auprès d'un agent, cela doit être enregistré explicitement comme étant une activité distincte (mais liée), plutôt que de payer sans aucune mention un plus petit montant, ou simplement modifier le fichier de la commission à payer.

**Combinaison des données pour ajouter des éléments de contexte :** La combinaison de données du prestataire de SFN avec les données des partenaires peut présenter de nombreux avantages opérationnels. Par exemple, lorsqu'il existe une collaboration avec un ORM, il existe aussi des informations sur le lieu où l'expéditeur et le destinataire étaient situés physiquement, la carte SIM utilisée, le type de téléphone utilisé, les potentiels historiques d'appels et les habitudes de recharge des clients. Comme de nombreux marchés ont un strict mandat d'inscription de la carte SIM, les informations de la KYC du client peuvent également être utilisées pour compléter et croiser les dossiers. Si certains de ces paramètres ne sont pas de première importance pour les transactions, ces données sont utiles pour déterminer les anomalies du système ; par exemple, si un client effectue habituellement des transactions à partir d'un téléphone particulier, et que le téléphone a changé, il se peut que la transaction soit frauduleuse. D'autres preuves peuvent être recueillies par des références croisées sur l'endroit où la transaction a eu lieu grâce à l'historique des localisations habituelles du client.

Il peut exister des difficultés à essayer de mettre en corrélation des données provenant de différentes sources, ce qui nécessite un examen au cours du processus de conception de la base de données. Par exemple, même lorsque l'ORM fait partie de la même organisation que le prestataire de SFN, le partage de données peut être un problème car les deux systèmes ne sont pas conçus pour se fournir mutuellement des services d'information. Essayer rétrospectivement de lier les données de télécommunications d'une interaction du système client aux informations sur les transactions financières de SFN n'est pas simple. La raison en est généralement qu'il n'existe pas d'élément commun de données reliant les deux dossiers, et les horloges horodatant l'événement sur les deux systèmes ne sont souvent pas parfaitement synchronisées. De fait, de nombreux systèmes n'effectuent d'activités de combinaisons de données que par exception, le plus souvent dans le cadre d'enquêtes sur la fraude ou au cas par cas. Le contexte supplémentaire fourni par les données combinées peut toutefois ajouter des couches de valeur, en particulier dans le cas d'une surveillance proactive des fraudes. La facilitation de l'association des données pour qu'elles puissent être utilisées dans des activités opérationnelles « normales » mérite d'être examinée, en particulier pour les opérations de SFN plus évoluées.

**Tentatives échouées :** Il est fréquent que les prestataires de SFN conservent les données associées à des transactions réussies, lorsque l'activité demandée a été accomplie. Les transactions ayant échoué peuvent toutefois fournir elles aussi des indications. Les raisons pour lesquelles des transactions particulières ont été refusées peuvent indiquer des besoins très spécifiques, tels que la nécessité de fournir des informations et une éducation ciblées, une défaillance technique ou une lacune dans la conception de services qui doit être modifiée pour offrir une expérience utilisateur plus intuitive.

Pour effectuer ces analyses évoluées, tout élément d'information sur toutes les interactions du système doit être recueilli et stocké, même si son utilité n'est pas immédiatement évidente.

**Source unique de la vérité :** Quand il existe plusieurs systèmes, il est courant d'avoir les mêmes données en double à plusieurs endroits. Cela est souvent dû au fait qu'il est difficile de combiner des sources de données de toute autre manière avec l'infrastructure actuelle. Cette duplication des données peut entrainer des problèmes concernant « la source de vérité », autrement dit, des questions sur la source de données à laquelle on peut se fier lorsque les informations sont contradictoires. Tous les systèmes sont parfois sujets à des erreurs, et lorsqu'il existe un conflit sur les détails d'une transaction ou un débat pour savoir si les fonds ont été transférés, il doit exister un accord clair sur les données auxquelles on peut se fier. Travailler sur ces détails fait partie d'un projet qui combine et compare les sources d'information ; il est également important de comprendre clairement si un enregistrement est définitif ou s'il peut encore être mis à jour. Traiter de façon incorrecte un enregistrement non-définitif comme définitif peut causer des ravages dans l'analyse des données, inspirant ainsi la méfiance quant à l'intégrité de la plateforme.

## 1.2.3 Analyses et applications : Notation du risque de crédit

La notation du risque de crédit peut être largement décrite comme étant l'étude du comportement et des caractéristiques passés des emprunteurs pour prédire le comportement futur d'emprunteurs nouveaux et existants.[25] L'émergence des mégadonnées et les sources et formats de ces données ont permis des approches supplémentaires du processus de notation du risque de crédit. L'intégration de ces sources de données alternatives conduit à des modèles de notation du risque de crédit alternatifs. Cette section se penche sur la façon dont les données façonnent la notation du risque de crédit, et quels types de données fonctionnent mieux pour différents besoins. Les relations fondamentales de notation du risque de crédit sont présentées sous forme de ligne chronologique dans la figure ci-dessous.

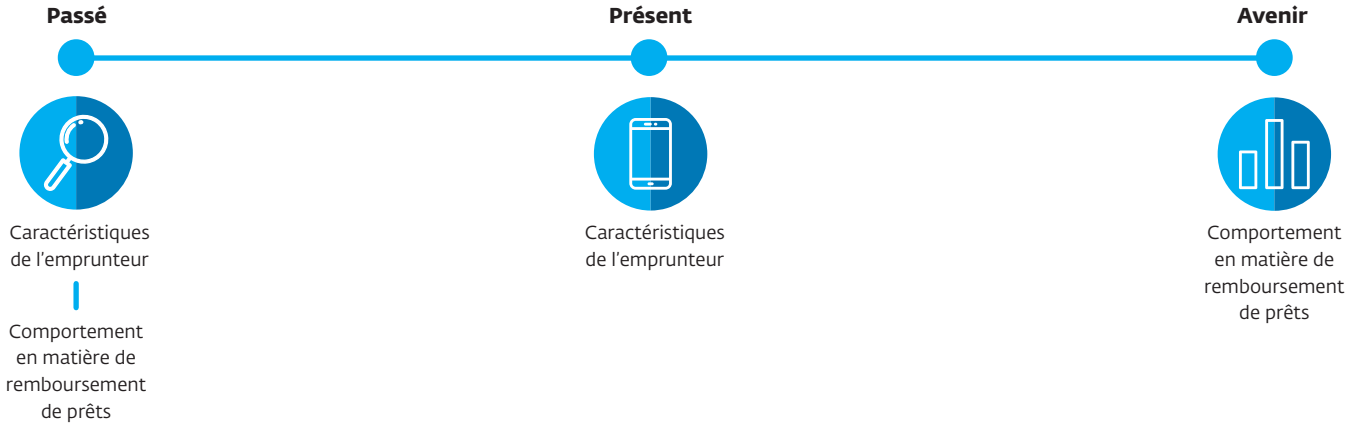



*Figure 18: Définition de la notation du risque de crédit selon une ligne chronologique*

[25] Schreiner, « Credit scoring for microfinance : Can it work? », *Journal of Microfinance/ESR Review*, Vol. 2.2 (2009) : 105-118

## 1.2_APPLICATION DE DONNÉES

Voici les principaux points illustrés à la figure 18 :

1. **Passé :** Les données (ou, en leur absence, l'expérience) sont étudiées pour comprendre quelles caractéristiques de l'emprunteur sont les plus significativement liées au risque de remboursement. Cette étude du passé éclaire le choix des facteurs et indique des pondérations dans la fiche d'évaluation.
2. **Présent :** La fiche d'évaluation (conçue à partir des données sur les caractéristiques passées d'un emprunteur) est utilisée pour évaluer les mêmes caractéristiques des nouveaux demandeurs de prêt. Le résultat est un score numérique qui est utilisé pour situer le demandeur dans un « groupe de risque » ou une fourchette de notations correspondant à une constatation de taux de remboursement similaires.
3. **Avenir :** Le modèle suppose que les nouveaux candidats avec les mêmes caractéristiques que les emprunteurs passés auront le même comportement de remboursement que ces emprunteurs du passé. Par conséquent, le taux de situation de prêt non remboursé constaté dans le passé pour un groupe de risque donné est le taux de situation de prêt non remboursé prévu pour les nouveaux emprunteurs dans ce même groupe de risque.

Un manuel complet peut être écrit sur la notation du risque de crédit, et plusieurs textes approfondis et accessibles ont de fait été publiés sur le sujet au cours de la dernière décennie.[26] Le CGAP a lui aussi publié récemment une introduction à la notation du risque de crédit dans le cadre des services financiers numériques.[27] Aux fins de ce manuel, le reste de cette section sur le crédit s'axe sur :

1. La façon dont les données sont converties en notations de crédit
2. La façon dont les données sont utilisées pour relever les défis d'évaluation du crédit sur les marchés en développement

### Conception de fiches d'évaluation

Les fiches d'évaluations de crédit sont conçues en étudiant un échantillon de données sur des prêts antérieurs qui ont été classés comme « bons » ou « mauvais ». Une définition courante des « mauvais » prêts (ou prêts « de qualité inférieure ») est « 90 jours consécutifs ou plus d'arriérés de paiement »[28] mais pour la conception de fiches d'évaluation, un mauvais prêt doit être décrit comme un prêt (avec le recul) que les institutions financières choisissent de ne pas accorder à l'avenir. Pour chaque nouveau demandeur de prêt, le modèle de notation calcule et compte quel pourcentage des emprunteurs passés ayant la même combinaison de caractéristiques d'emprunteurs étaient « mauvais ».

Il est important de mener des analyses tant sur les bons que sur les mauvais prêts. L'étude des relations de risque en matière de données de crédit consiste simplement à regarder le nombre de bons et de mauvais prêts selon différentes caractéristiques de l'emprunteur. Plus il existe de mauvais prêts en pourcentage du total des prêts pour une caractéristique d'emprunteur donnée, plus le risque est élevé.

Le tableau croisé, ou tableau de corrélation, est un simple outil d'analyse qui peut être utilisé pour créer et gérer des fiches d'évaluation de crédit. Le tableau 4 ci-dessous indique le nombre de bons et mauvais selon les plages de valeurs pour un exemple de champ de données d'ORM ; dans ce cas, il s'agit du temps écoulé depuis l'inscription au réseau mobile. Supposons que l'on s'attende à ce que les candidats ayant une expérience plus longue sur le réseau mobile représentent un plus faible risque (habituellement les antécédents plus longs, que ce soit en matière d'emploi, d'activité, de résidence ou en tant que clients de la banque, sont liés à un risque plus faible).

---

[26] Voir par exemple : Siddiqi, « Credit risk scorecards: developing and implementing intelligent credit scoring », John Wiley and Sons, Vol. 3 (2012).
Anderson, « The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation », Oxford University Press, 2007

[27] « An Introduction to Digital Credit: Resources to Plan a Deployment, » *Consultative Group to Assist the Poor via Slide Share*, 3 juin 2016, consulté le 3 avril 2017, http://www.slideshare.net/CGAP/an-introduction-to-digital-credit-resources-to-plan-a-deployment

[28] Pour les SFN et les microprêteurs, la définition du « mauvais » prêt peut souvent correspondre à une période de défaillance beaucoup plus courte, par exemple 30 ou 60 jours consécutifs d'arriérés. La conception du produit (notamment les pénalités et les frais de retard) et le travail consacré au processus de collecte influenceront le seuil à partir duquel il vaut mieux éviter un client, considéré comme « mauvais ».

| Ligne | | <= 2 mois | > 2 mois et <= 1 an | > 1 an et <= 2 ans | > 2 ans et <= 3 ans | > 3 ans | Total de la ligne |
|---|---|---|---|---|---|---|---|
| A | Bons | 115 | 161 | 205 | 116 | 203 | 800 |
| B | Mauvais | 48 | 48 | 50 | 24 | 30 | 200 |
| C | Taux de « mauvais » | 29.4% | 23.0% | 19.8% | 17.3% | 12.7% | 20.0% |
| D | Total | 163 | 210 | 255 | 140 | 233 | 1,000 |
| E | % du total des prêts | 16.3% | 21.0% | 25.5% | 14.0% | 23.3% | |

*Tableau 4 : Tableau croisé des prêts*

Le tableau 4 peut se lire comme suit :

**Ligne A :** Nombre de bons contrats dans le groupe (colonne)

**Ligne B :** Nombre de mauvais contrats dans le groupe (colonne)

**Ligne C :** Nombre de mauvais contrats (ligne B)/Nombre total de contrats (ligne D)

**Ligne D :** Nombre total de contrats (ligne A + ligne B)

**Ligne E :** Total des contrats dans le groupe (colonne) divisé par tous les contrats (1 000)

Pour effectuer l'analyse, l'étape suivante consiste à rechercher des modèles raisonnables et intuitifs. Par exemple, le taux de « mauvais » à la ligne C du tableau 4 diminue clairement à mesure que le temps écoulé depuis l'inscription au réseau augmente. Cela correspond à l'attente initiale. Pour se représenter facilement le risque de chaque groupe, il suffit d'examiner son taux de « mauvais » par rapport au taux de « mauvais » de 20 pour cent (en moyenne) selon le temps écoulé depuis l'inscription :

- Moins de 2 mois, le taux de « mauvais » est de 29 pour cent, une fois et demie la moyenne.
- Entre 1 an et 2 ans, le taux de « mauvais » est de 19,8 pour cent, ou un risque moyen.
- Plus de 3 ans, le taux de « mauvais » est de 12,7 pour cent, un peu plus de la moitié du risque moyen.

Dans la conception classique de fiches d'évaluation de crédit, les analystes recherchent des modèles simples, notamment la hausse ou la baisse constante des taux de « mauvais », qui sont commercialement raisonnables. Les fiches d'évaluation de crédit ainsi conçues conviennent bien à une utilisation opérationnelle en tant qu'outils commerciaux qui sont à la fois transparents et bien compris par la direction. Une autre approche de la conception de fiche d'évaluation est l'exploration de données, ou l'utilisation d'algorithmes d'apprentissage automatisé plus complexes pour toutes les relations dans un ensemble de données, qu'elles soient comprises par un analyste humain ou non. Bien qu'une approche d'apprentissage automatique pur pourrait entrainer une amélioration des prévisions dans certaines situations, il existe aussi des avantages difficiles à mesurer mais pratiques, pour la gestion des entreprises et des risques, à bien comprendre comment les notations sont calculées.

Des tableaux croisés ou une analyse similaire des prédicteurs simples est la clé de voute des modèles de notation du risque de crédit.[29] La création de tableaux croisés comme ceux illustrés dans l'exemple ci-dessus est facile en utilisant un logiciel de statistiques commercial ou le logiciel gratuit open-source « R ».

[29] En fait, les coefficients de régression logistique peuvent être calculés directement à partir d'un tableau croisé pour une seule variable

## Cas d'utilisation : Conception des fiches d'évaluation

Les points d'une fiche d'évaluation sont des transcriptions des modèles de taux de «mauvais » observés dans des tableaux croisés. Bien qu'il existe de nombreuses méthodes mathématiques pouvant être utilisées pour concevoir des fiches d'évaluation (voir chapitre 1.2.3), différentes méthodes donnent des résultats similaires. Ceci pour la simple raison que la puissance prédictive du modèle de notation statistique ne provient pas du calcul mais de la solidité des données elles-mêmes. Avec des données suffisantes sur les caractéristiques pertinentes de l'emprunteur, des méthodes simples produiront un bon modèle et des méthodes complexes peuvent produire un modèle un peu meilleur. S'il n'existe pas de bonnes données (ou trop peu de données), aucune méthode ne produira de bons résultats. En vérité, la conception de fiches d'évaluation favorise non seulement des modèles simples, mais signifie également qu'un prestataire de SFN axé sur les données doit d'abord se concentrer sur la saisie, le nettoyage et le stockage de données en plus grande quantité et de meilleure qualité.

Le tableau 5 ci-dessus est un autre tableau croisé, cette fois pour le facteur « âge ». Comme le tableau précédent, les taux de « mauvais » dans la ligne C représentent le risque (le taux de « mauvais »), qui diminue à mesure que l'âge augmente.

### Différences des taux de « mauvais »

Une façon très simple de transformer des taux de « mauvais » en points de fiche d'évaluation est de calculer les différences entre les taux de « mauvais ». Comme le montre la ligne G, le taux de « mauvais » pour chaque groupe est soustrait du taux de « mauvais » le plus élevé pour tous les groupes (ici, il est de 30,9 pour cent pour les « 23 ans ou moins »), qui est ensuite multiplié par 100 (pour obtenir des nombres entiers plutôt que des nombres décimaux). Les résultats (indiqués en ligne F) peuvent être utilisés comme points dans une fiche d'évaluation statistique. Dans un tel système de points, le groupe le plus risqué recevra toujours 0 points et le groupe au risque le plus faible (c.-à-d. le groupe avec le taux de « mauvais » le plus bas) recevra le plus de points.

Pour les fiches d'évaluations conçues en utilisant une régression (voir chapitre 1.1), la transformation des coefficients de régression en des points positifs nécessite quelques étapes supplémentaires. Les calculs ne sont pas détaillés ici, mais les résultats du classement sont très similaires, comme le montre la ligne H.

| Ligne | | 23 ans ou moins | 24 à 30 ans | 31 à 47 ans | 48 ans ou plus | Total |
|---|---|---|---|---|---|---|
| A | Bons | 46 | 238 | 374 | 142 | 800 |
| B | Mauvais | 20 | 74 | 82 | 23 | 200 |
| C | Taux de « mauvais » | 30.9% | 23.8% | 18.0% | 14.0% | 20.0% |
| D | Total de la colonne | 66 | 312 | 456 | 166 | 1,000 |
| E | Pourcentage du total des prêts | 6.6% | 31.2% | 45.6% | 16.6% | |
| F | POINTS | 0 | 7 | 13 | 17 | |
| G | Calcul [multiplié par 100] | (0.309 - 0.309) = 0 | (0.309 - 0.238) = 7 | (0.309 - 0.18) = 13 | (0.309 - 0.14) = 17 | |
| H | POINTS LOGIT | 0 | 10 | 21 | 29 | |

*Table 5: Ci-dessus est un autre tableau croisé, cette fois pour le facteur « âge ». Comme le tableau précédent, les taux de « mauvais » dans la ligne C représentent le risque (le taux de « mauvais »), qui diminue à mesure que l'âge augmente.*

## Les facteurs qui obtiennent le plus de points dans les fiches d'évaluation de crédit

Plus les différences de taux de « mauvais » entre les groupes sont importantes, plus un facteur de risque reçoit de points dans une fiche d'évaluation. En utilisant la méthode simple des différences de taux de « mauvais » (décrite ci-dessus), on voit dans le tableau 6 ci-dessous que la « notation de crédit bureau » peut avoir un maximum de 39 points, tandis que la « situation familiale » peut avoir un maximum de huit points. Il existe en effet des différences beaucoup plus importantes entre les taux de « mauvais » les plus élevés et les plus bas pour les antécédents de crédit qu'il n'en existe pour la situation familiale.

| | **Notations de bureau de crédit** | | | | |
|---|---|---|---|---|---|
| Groupe | < 590 Points | 590 - 670 Points | 671 - 720 Points | > 720 Points | Échantillon de taux de « mauvais » |
| Taux de « mauvais » | 39% | 23% | 13% | 0% | 20% |
| POINTS | 0 | 16 | 26 | 39 | |
| | **Situation familiale** | | | | |
| Groupe | Divorcé | Célibataire | Marié | Veuf | Échantillon de taux de « mauvais » |
| Taux de « mauvais » | 25% | 24% | 19% | 17% | 20% |
| POINTS | 0 | 1 | 6 | 8 | |

***Tableau 6 :*** *Exemples de l'importance des facteurs des fiches d'évaluation*

Puisque le classement des risques dans tous les algorithmes est souvent très similaire, de nombreux professionnels préfèrent, dans la pratique, utiliser des méthodes plus simples. L'auteur David Hand, spécialiste de la notation du risque de crédit, a souligné que « les méthodes simples produisent généralement des performances presque aussi bonnes que les méthodes plus sophistiquées, au point où la différence de performance peut être dépassée par d'autres sources d'incertitude qui ne sont en général pas prises en compte ».[30] La pratique généralisée de longue date de régression logistique pour la notation du risque de crédit témoigne de la facilité avec laquelle ces modèles peuvent être présentés sous forme de fiche d'évaluation. Ces fiches d'évaluation sont bien comprises par la direction et peuvent être utilisées pour gérer de façon proactive les risques et les bénéfices des prêts.

[30] David Hand, « Classifier technology and the illusion of progress », Statistical Science, Vol. 21.1 (2006) : 1-14.

## Fiche d'évaluation d'experts

Quand il n'existe pas de données historiques, mais que le prestataire a une bonne compréhension des caractéristiques de l'emprunteur qui déterminent le risque dans le segment, une fiche d'évaluation d'experts peut réussir à établir raisonnablement un classement des risques que représentent des emprunteurs.

Une *fiche d'évaluation d'experts* utilise des points pour classer les emprunteurs en fonction du risque, tout comme une fiche d'évaluation statistique le fait. La principale différence (et elle est de taille) est que sans données antérieures, notamment les données sur les situations de prêt non remboursé, il n'existe aucun moyen pour l'IF de savoir avec certitude si sa compréhension (ou son attente) des relations en matière de risque est correcte.

Par exemple, si l'on sait que l'âge est un facteur de risque significatif pour les prêts à la consommation et que nous avons constaté (dans la pratique) que le risque diminue généralement avec l'âge, on pourrait créer des groupes d'âge semblables à ceux du tableau 5. Dans ce scénario, nous attribuons des points en utilisant un schéma simple où le groupe perçu comme le plus risqué obtient toujours zéro point et le groupe au risque le plus faible obtient toujours 20 points. Dans ce cas, une pondération de la fiche d'évaluation d'experts de la variable « âge » pourrait ressembler au tableau 7. Ces points ne sont pas si différents des points statistiques pour l'âge indiqués aux lignes F et H du tableau 5.

| | 23 ans ou moins | 24 à 30 ans | 31 à 47 ans | 48 ans ou plus |
|---|---|---|---|---|
| POINTS | 0 | 7 | 15 | 20 |

*Tableau 7 : Points d'« experts » pour l'« âge »*

Tant que le classement des risques est correct pour chaque facteur de risque dans une fiche d'évaluation d'experts, la notation d'une fiche d'évaluation d'experts classera le risque des emprunteurs de la même façon qu'une fiche d'évaluation statistique le classe.[31] Cela signifie que les fiches d'évaluations d'experts peuvent représenter un outil utile pour lancer un nouveau produit pour lequel il n'existe pas de données historiques. Elles sont aussi un bon moyen, pour les prestataires de SFN qui ont l'intention de se fonder sur des données, de récolter quelques fruits de la notation - notamment une meilleure efficacité et cohérence - tout en constituant une meilleure base de données.

[31] Habituellement, en utilisant le seul jugement des experts, les prestataires spécifient de manière incorrecte la relation de classement des risques d'un ou plusieurs facteurs. Une fois que les données de performance (remboursement de prêt) sont recueillies, elles peuvent être utilisées pour corriger des relations mal définies, entrainant un meilleur classement des risques du nouveau modèle statistique.

## Choisir un ensemble de facteurs de risque

Bien que les champs de données spécifiques disponibles pour la notation du risque de crédit varient considérablement selon le produit, le segment et le prestataire, généralement, les données du modèle de notations doivent être :

- Très pertinentes
- Faciles à recueillir de manière systématique
- Objectives et non auto déclarées

Certains types de données ont tendance à être de bons prédicteurs de remboursement des prêts pour tous les segments et marchés. La tableau 8 présente quelques-uns de ces modèles ainsi que leurs modèles de risque fréquemment observés.

Le « meilleur » ensemble de prédicteurs à variables simples est assemblé pour former un modèle à plusieurs variables. Bien que cela puisse être fait à l'aide d'un algorithme pour maximiser la prévision, une approche attrayante pour les prestataires de SFN est de choisir une série de facteurs qui, ensemble, créent un profil de risque complet pour l'emprunteur,[31] conformément aux fameux cinq C du crédit : capacité, capital, garantie (« *collatéral* »), *conditions*, et *caractère*. Un tel modèle est facile à comprendre pour les banquiers et la gestion des banques, et est compatible avec les cadres de gestion des risques tels que les Accords de Bâle.

À mesure que chaque facteur prédictif est ajouté à un modèle multifactoriel, son classement des risques s'améliore. Toutefois, après un nombre relativement faible de bons indicateurs (habituellement 10 à 20), l'amélioration apportée par chaque facteur supplémentaire diminue assez fortement. Même si nous choisissons délibérément des facteurs qui ne semblent pas fortement corrélés les uns aux autres, en réalité, bon nombre de facteurs sont corrélés dans une certaine mesure, ce qui provoque la baisse de l'apport des facteurs supplémentaires.

| Type de données | Facteur | Relation de risque |
|---|---|---|
| Comportementales | Achats | Le risque diminue à mesure que le revenu disponible augmente |
| | Dépôts et chiffre d'affaires du compte | Le risque diminue lorsque les dépôts et le chiffre d'affaires augmentent |
| | Antécédents en matière de crédit | Le risque diminue lorsque les antécédents positifs en matière de crédit augmentent |
| | Paiement de factures | Le risque diminue selon la ponctualité des paiements de factures |
| Historiques | Temps passé à la résidence, dans l'emploi, l'entreprise | La stabilité réduit les risques |
| | Ancienneté en tant que client | Les clients ayant une plus longue relation représentent un risque plus faible |
| Démogra-phiques | Âge | Le risque diminue avec l'âge et augmente à nouveau autour de l'âge de la retraite (principalement en raison des risques de santé) |
| | Situation familiale | Les personnes mariées sont plus souvent installées et stables, ce qui réduit le risque |
| | Nombre de personnes à charge | Un nombre croissant de personnes à charge peut augmenter le risque (en particulier pour les personnes seules), mais dans certaines cultures, au contraire, il diminue le risque (plus grand filet de sécurité) |
| | Propriété de la maison | Les propriétaires sont moins risqués que les locataires |

***Tableau 8 :*** *Données qui sont souvent efficaces pour la notation du risque de crédit*

[32] Siddiqi, « Credit risk scorecards: developing and implementing intelligent credit scoring, » *John Wiley and Sons*, Vol. 3 (2012).

Lorsqu'une IF dispose de suffisamment de données, elle doit privilégier les points de données qui :

- Sont objectifs et peuvent être observés directement, plutôt que suscités par le demandeur
- Prouvent des relations au risque de crédit qui confirment un jugement d'expert ou intuitif
- Sont moins couteux à recueillir
- Peuvent être recueillis auprès de la plupart sinon de tous les demandeurs
- N'opèrent pas de discriminations fondées sur des facteurs que l'emprunteur ne peut pas contrôler (c.-à-d. l'âge, le sexe, l'apparence) ou qui sont potentiellement source de division (c.-à-d. la religion, l'origine ethnique, la langue)

### Cas d'utilisation : Les nano-crédits

Puisque les banques doivent déclarer les remboursements de nano-crédits aux bureaux de crédits et banques centrales, les nano-crédits ont fait entrer des millions de personnes qui ne bénéficiaient auparavant pas d'accès aux banques dans le secteur financier formel à travers le monde, en établissant des antécédents de crédit qui sont un tremplin pour ouvrir l'accès à d'autres types de produits de prêt. Cependant, certains craignent que les nano-crédits créent un cycle d'endettement pour les personnes à faible revenu. Plusieurs millions de personnes avec de mauvaises expériences en matière de nano-crédits pourraient se retrouver sur la liste noire de leurs bureaux de crédits locaux, ce qui confirme d'autant plus la nécessité d'une protection des consommateurs.

Cette section examine la façon dont les données sont utilisées pour surmonter quelques-unes des difficultés qui ont longtemps été des obstacles à l'inclusion financière. Ce sont en particulier les données numériques générées par les téléphones portables, l'argent mobile et l'Internet qui permettent à des millions de personnes qui n'ont jamais eu de comptes bancaires ou de prêts bancaires de se faire connaitre par les IF formelles.

Les études de cas qui suivent enquêtent sur la façon dont les ORM, les réseaux sociaux et les données bancaires traditionnelles ont été utilisées pour lancer de nouveaux produits, pour aider davantage d'emprunteurs à devenir éligibles à des prêts formels et pour évaluer les petites entreprises, qui sont moins homogènes que les consommateurs individuels.

### Défi du crédit n° 1 : Vérification des revenus et dépenses

Un important défi du prêt de détail dans les marchés en développement est l'obtention de données fiables sur les flux de trésorerie des nouveaux clients, pour les personnes comme pour les entreprises. Les flux de trésorerie, ou les revenus restants après déduction des frais, sont la principale source de remboursement du prêt et donc un point central des modèles de prêts au détail. Les niveaux de revenu sont également utilisés pour déterminer quel montant de financement un individu peut se permettre.

La croissance de la téléphonie mobile et l'utilisation de l'argent mobile - en particulier en Afrique et en Asie - a créé des historiques numériques vérifiables par des tiers de véritables modèles de paiement, tels que les recharges et les paiements d'argent mobile. Ces données, détenues par les ORM, offrent un aperçu des flux de trésorerie d'un utilisateur de SIM. Les terminaux de PDV et les caisses d'argent mobile peuvent également peindre un tableau un peu plus complet des flux de trésorerie pour les commerçants.

Lorsque vous savez combien d'argent une personne ou société manipule de façon quotidienne, hebdomadaire et mensuelle, vous pouvez mieux estimer quelle taille de prêt elle sera en mesure de rembourser.

Les deux cas suivants examinent comment les données numériques ont permis d'ouvrir d'immenses marchés pour les nano-crédits à la consommation.

# CAS 10

## *M-Shwari lance un marché pour les nano-crédits*

***Solutions de données pour évaluer la solvabilité des emprunteurs sans antécédents de crédit formels***

*La Commercial Bank of Africa (CBA) et l'opérateur de téléphonie mobile Safaricom ont été les premiers à reconnaitre la puissance du téléphone mobile et des données de l'argent mobile.*

*M-Shwari, le premier produit d'épargne et de prêt numérique très prospère, est bien connu des adeptes des entreprises de technologie financière et de l'inclusion financière. Il a accordé de petites limites de crédit sur les téléphones mobiles appelées nano-crédits à des millions d'emprunteurs, en les faisant ainsi entrer dans le secteur financier formel. Des produits similaires ont depuis été lancés dans d'autres régions d'Afrique, et une nouvelle concurrence s'est entassée sur le marché au Kenya. L'histoire de M-Shwari est également une excellente étude d'un exemple d'utilisation de données de façon créative pour faire entrer un nouveau produit sur le marché.*

### *Modélisation de l'inconnu*

*La technologie de notation de risque de crédit examine les caractéristiques et le comportement de remboursement passés de l'emprunteur afin de prévoir le remboursement futur du prêt. Qu'en est-il du cas où il n'existe pas de comportement de remboursement passé ? Les ORM possèdent des données détaillées sur les téléphones mobiles de leurs clients et, dans de nombreux cas, de l'utilisation de l'argent mobile, mais déterminer comment ces données peuvent être utilisées pour prédire la capacité et la volonté de rembourser un prêt sans données sur le paiement des obligations passées est moins clair.*

*Par définition, il n'existe pas de données antérieures spécifiques à un produit qui est nouveau. Une façon d'utiliser encore la notation de risque de crédit avec un nouveau produit est d'utiliser le jugement et les connaissances sur le sujet d'un expert pour concevoir une « fiche d'évaluation d'expert », un outil qui oriente les décisions de prêt fondé sur les classements des risques des emprunteurs. Voir l'encadré de rappel en 84.*

*Une autre façon d'utiliser la notation de risque de crédit avec un nouveau produit est d'étudier un ensemble de données client pertinentes, telles que les données des ORM, en les comparant aux informations de remboursement de prêt, telles que :*

- ***Antécédents généraux en matière de crédit ou rapport de bureau :*** *Cela ne fonctionne que pour les clients qui ont un dossier auprès du bureau.*
- ***Produits de crédit similaires :*** *Un autre produit de crédit suffisamment similaire pour être pertinent par rapport au nouveau produit peut être utilisé comme référence. Bien que le remboursement passé de ce produit puisse être représentatif ou non des remboursements futurs du nouveau produit, il peut représenter une approximation acceptable, ou « indirecte », à des fins de modélisation initiale.*

## 1.2_APPLICATION DE DONNÉES

*La première fiche d'évaluation de M-Shwari a été conçue à partir de données Safaricom et de l'historique de remboursement des clients qui avaient utilisé ses produits de crédit de temps de communication Okoa Jahazi.*[33] *Les deux produits étaient nettement différents, comme le montre Tableau 9 ci-dessous.*

*Le produit M-Shwari a offert aux emprunteurs plus d'argent, de souplesse d'utilisation et de temps pour rembourser. L'hypothèse était que ceux qui avaient utilisé avec succès les très modestes prêts Okao Jahzi représenteraient moins de risques pour le produit de prêt plus important.*

*Le premier modèle de notation de risque de crédit M-Shwari développé avec les données d'Okoa Jahazi,*[34] *accompagné de politiques de limites prudentes et de processus d'entreprise bien conçus, a permis le lancement du produit, qui est rapidement devenu un immense succès.*

*La CBA s'attendait à ce que la fiche d'évaluation fondée sur les données d'Okoa Jahazi soit reconçue le plus rapidement possible en utilisant le comportement de remboursement du produit M-Shwari lui-même. Certains comportements prédictifs de l'utilisation du crédit de temps de communication ne se traduisent pas directement en utilisation de M-Shwari, et des changements appropriés au modèle en fonction des données d'utilisation réelle du produit M-Shwari ont réduit les prêts non productifs de 2 pour cent. M-Shwari continue à mettre à jour sa fiche d'évaluation périodiquement en fonction des nouvelles informations.*

| Produit | Okao Jahzi | M-Shwari |
|---|---|---|
| **Montant** | Le chiffre le plus bas entre le temps de communication dépensé au cours des 7 derniers jours, ou 100 shillings kenyans | 100 à 10 000 shillings kenyans |
| **Objectif** | Utilisé uniquement pour le temps de communication | Utilisé à toute fin |
| **Condition de remboursement** | 72 heures | 30 jours |

*Tableau 9 : Okao Jahzi et la comparaison des produits M-Shwari*

*Le lancement et la conception réussis de M-Shwari montrent qu'il existe des façons d'utiliser des solutions de notation fondées sur les données pour des segments entièrement nouveaux. Il renforce également le bien-fondé général de la notation de risque de crédit qui veut qu'une fiche d'évaluation fasse l'objet d'un travail permanent. Peu importe le degré d'efficacité d'une fiche d'évaluation quant aux données de conception, elle doit être suivie et gérée en utilisant des rapports standards et être affinée à chaque fois qu'il existe des changements importants des risques de marché ou des types de clients qui demandent le produit.*

[33] Cook et McKay, « How M-Shwari works: The story so far », *Groupe consultatif d'assistance aux plus pauvres et Financial Sector Deepening*

[34] Mathias, « *What You Might Not Know*, » Abacus, 18 septembre 2012, consulté le 3 avril 2017, https://abacus.co.ke/okoa-jahazi-what-you-might-not-know/

Le produit de nano-crédit M-Shwari a réussi grâce à la confluence en temps voulu de :

- **L'accès aux données des ORM :** La CBA avait un avantage du premier entrant en raison de son partenariat solide avec Safaricom. Aujourd'hui, Safaricom vend ses données d'ORM à toutes les banques du Kenya.
- **Un produit bien conçu :** Les produits modestes à court terme correspondent mieux à la notation de risque de crédit, en particulier pour les nouveaux produits. Le retour d'information rapide sur la performance de remboursement de la population cible permet la modification du modèle en temps voulu et contrôle le risque.
- **De bons systèmes et les bonnes personnes :** L'équipe de direction de M-Shwari est modeste et flexible, se composant d'une série unique de compétences de gestion et de compétences techniques, ainsi que de systèmes assurant une mise en œuvre sans heurts.
- **Mobilisation des ressources extérieures :** Financial Sector Deepening (FSD) Kenya a soutenu la CBA avec une expertise de modélisation des risques essentielle pour développer le premier modèle de notation et transférer des compétences à l'équipe de M-Shwari.

Alors que l'histoire de la réussite de M-Shwari est source d'inspiration, il existe de nombreux prestataires de SFN qui souhaiteraient entrer dans la sphère du nano-crédit, mais ils pourraient être confrontés à des difficultés. Ces prestataires de SFN peuvent ne pas avoir de relations avec les ORM ou ne pas avoir la capacité interne requise pour concevoir de l'épargne numérique, des produits de prêts et des modèles de notation. Le cas suivant décrit comment les fournisseurs facilitent l'entrée des prestataires de SFN sur le marché de masse des nano-crédits.

# CAS 11

## *Tiaxa, l'approche de nano-crédits clé en main*

***Développement de produits et de services de données par le biais de services d'abonnement externalisés***

*Reconnaissant que de nombreuses IF sur les marchés en développement ne disposent pas des ressources nécessaires pour aborder le marché des SFN en n'utilisant que des ressources internes, Tiaxa propose ses NanoCredits™ brevetés dans le cadre d'une solution « clé en main » qui comprend les éléments suivants :*

- *Conception de produit*
- *Acquisition de clients (fondée sur des modèles de notation propriétaires)*
- *Gestion du risque de crédit de portefeuille*
- *Déploiement de matériel et de logiciels*
- *Gestion du service jour et nuit*
- *Facilité de financement du portefeuille (sur certains marchés africains)*

*Tiaxa réunit les IF et les ORM et forme des partenariats à trois dans lesquels :*

- **Les ORM** *fournissent les données qui définissent leurs modèles de décision de crédit*
- **Les IF** *fournissent les licences de prêt (et la réglementation du secteur financier formel) et le financement nécessaires*
- **Tiaxa** *fournit la solution de produit de nano-crédit de bout en bout*

*En plus de fournir les modèles de conception et de notation des produits fondés sur les données des ORM, dans la plupart des cas, Tiaxa assume et gère le risque de crédit du portefeuille. Le risque de perte est géré en débitant directement les comptes des ORM des emprunteurs pour résoudre le problème de situation de prêt non remboursé, ce qui est divulgué aux emprunteurs dans les conditions générales du produit. Leur modèle économique de partenariat à long terme fonctionne à des conditions qui varient, de l'intéressement aux bénéfices à des modèles de frais par transaction.*

### *Données qui déterminent les modèles de notation de Tiaxa*

*Bien que les ensembles de données des ORM varient selon les pays et les marchés, les ensembles de données qui informent les modèles propriétaires de Tiaxa comprennent habituellement une combinaison des types de données suivantes :*

| Utilisation du GSM | Paie, paiements réguliers | Virements Bancaires | Informations de la KYC | Paiements des services collectifs | Dépôt en espèces |
|---|---|---|---|---|---|
| • Fréquence, montants des recharges<br>• Informations sur la consommation GSM | • Paie, subventions<br>• Flux de trésorerie, besoins en crédit | • Fréquence et valeur<br>• Réception ou envoi ? | • Nom complet<br>• Type de compte<br>• Date d'inscription<br>• Situation quant à la KYC<br>• Date de naissance, région | • Indicateur de flux de trésorerie<br>• Connaissances financières | • Informations sur les flux de trésorerie |

***Tableau 10 :*** *Types de données informant les modèles propriétaires de Tiaxa*

*Tiaxa utilise une série de méthodes d'apprentissage automatique pour réduire des centaines de prédicteurs potentiels en un modèle optimal. Des modèles personnalisés sont conçus pour chaque engagement. Tiaxa a maintenant plus de 60 installations, avec 28 clients, répartis dans 20 pays, en 11 groupes d'ORM, qui, entre eux, représentent plus d'1,5 milliard utilisateurs finaux. Aujourdhui, la société traite plus de 12 millions de nano-crédits par jour dans le monde entier, principalement sous forme de prêts de temps de communication.*

*À mesure que le paysage d'analyse des données évolue, des fournisseurs tiers doivent développer des solutions clé en main qui puisent dans les sources de données internes et apportent de la valeur aux produits existants. Les entreprises qui ne parviennent pas à investir dans l'analyse de données personnalisée ou qui préfèrent une approche attentiste peuvent être en mesure de tirer parti des services d'abonnement à l'avenir en exportant les données à des fournisseurs externes.*

Pour les IF, le choix entre travailler avec des fournisseurs ou directement avec les opérateurs mobiles pour atteindre le segment des nano-crédits ne peut être fait qu'en tenant compte des conditions et des ressources disponibles sur le marché. Certains des avantages et des inconvénients de chaque approche sont présentés ci-dessous.

### Cas d'utilisation : Les données alternatives

Les sources de données alternatives sont prometteuses en matière de vérification d'identité et d'évaluation des risques de base. Un autre moyen utilisé par les prestataires de SFN pour recueillir des données sur de nouveaux candidats est de leur demander de fournir directement les informations. Ces demandes peuvent prendre la forme de :

- Formulaires de demande
- Enquêtes
- « Autorisations » pour accéder aux données des appareils : Cela peut inclure des autorisations pour accéder au contenu des médias, journaux d'appels, contacts, communications personnelles, informations de localisation ou profils de réseaux sociaux en ligne

Ces sources de données en ligne non traditionnelles peuvent être et sont utilisées pour offrir des services de vérification d'identité et des notations de crédit. L'histoire de l'entreprise d'analyse de réseau social Lenddo fournit davantage d'éléments de contexte et une indication de la façon dont les données des réseaux sociaux peuvent ajouter de la valeur dans le processus de crédit.

| Approche | Opportunités. | Défis |
|---|---|---|
| Travailler avec les données des ORM | • Contrôle total des produits<br>• Potentiellement plus rentable | Besoin de compétences internes en matière de :<br>• Développement de produits<br>• Modélisation des risques<br>Besoin de systèmes et de logiciels pour gérer les produits de SFN |
| Travail avec un fournisseur | • Fournit le savoir-faire en matière de produit, de modélisation et de systèmes<br>• Prend les décisions de prêt<br>• A des solutions logicielles prêtes à l'emploi | • Dépendance du fournisseur<br>• Les détails du modèle peuvent ne pas être communiqués<br>• Compétences techniques non transférées |

*Tableau 11 : Travailler avec des ORM ou des fournisseurs : Opportunités et défis*

# CAS 12

## *Lenddo exploite les données de réseaux sociaux pour vérifier l'identité et établir des profils de risque*

***Utilisation de techniques analytiques évoluées et de sources de données alternatives pour les nouveaux produits***

*Les co-fondateurs de Lenddo, Jeffrey Stewart et Richard Eldridge, ont à l'origine eu l'idée de ce service dans le secteur de l'externalisation des processus d'entreprise aux Philippines en 2010. Ils ont été surpris par le nombre d'employés qui leur demandaient régulièrement des avances de salaire et se demandaient pourquoi ces jeunes personnes brillantes, avec un emploi stable, ne parvenaient pas à obtenir de prêts auprès d'IF.*

*Le défi particulier aux Philippines était que le pays n'avait ni bureau de crédit, ni numéros d'identification nationaux. Si les personnes n'utilisaient pas de comptes ou de services bancaires – et moins de 10 pour cent d'entre eux les utilisaient-ils étaient « invisibles » pour les IF formelles et incapables d'obtenir un crédit. En développant leur idée, les fondateurs de Lenddo ont tout de suite remarqué que leurs employés étaient des utilisateurs fervents de la technologie et présents sur les réseaux sociaux. Ces plateformes génèrent de grandes quantités de données, dont l'analyse statistique qu'ils pensaient obtenir pourrait aider à prédire la solvabilité d'un individu.*

*Les demandeurs de prêt de Lenddo donnent l'autorisation d'accéder aux données stockées sur leur téléphone mobile. Les données brutes du demandeur sont consultées, extraites et notées, puis détruites (plutôt que stockées) par Lenddo. Pour un candidat typique, son téléphone peut contenir des milliers de points de données parlantes quant à son comportement personnel :*

- *Trois degrés de connexions sociales*
- *Activité (photos et vidéos affichées)*
- *Membres de groupes*
- *Intérêts et communications (messages, e-mails et tweets)*

*Plus de 50 éléments à travers tous les profils de réseaux sociaux fournissent 12 000 points de données par utilisateur moyen :*

| **Sur les cinq réseaux sociaux :** | **7 900 communications de messages totales et +** |
|---|---|
| • 250 connexions de premier degré et +<br>• 800 connexions de deuxième degré et +<br>• 2 700 connexions de troisième degré et +<br>• 372 photos, 18 vidéos, 13 groupes, 27 intérêts, 88 liens, 18 tweets | • 250 connexions de premier degré et +<br>• 5 200 messages Facebook et +, 1 100 « j'aime » sur Facebook et +<br>• 400 mises à jour de statut Facebook et +, 600 commentaires Facebook et +<br>• 250 e-mails et + |

***Tableau 12 :*** *Moyennes de points de données de réseaux sociaux par utilisateur moyen*

## 1.2_APPLICATION DE DONNÉES

***Utilisation des données***

*La confirmation de l'identité d'un emprunteur est un élément important pour accorder un crédit aux candidats sans antécédents de crédit. L'application pour tablettes de Lenddo demande aux demandeurs de prêt de remplir un court formulaire numérique leur demandant leur nom, date de naissance, numéro de téléphone principal, adresse e-mail principale, école et employeur. Les demandeurs sont ensuite invités à intégrer Lenddo en se connectant et en donnant des autorisations à Facebook. Les modèles de Lenddo utilisent ces informations pour vérifier l'identité des clients en moins de 15 secondes. La vérification d'identité peut considérablement réduire le risque de fraude, qui est beaucoup plus élevé pour les produits de prêt numériques, pour lesquels il n'y a pas de contact personnel lors du processus de souscription. Un exemple de la collaboration de Lenddo avec le plus grand ORM aux Philippines est présenté ci-dessous.*

*Lenddo a travaillé avec un grand ORM pour augmenter la part des forfaits post payés qu'elle pouvait offrir à ses 40 millions d'abonnés aux services prépayés (90 pour cent du total des abonnés). L'admissibilité au forfait post payé dépendait de la réussite de la vérification d'identité, et le processus de vérification existant de Telco exigeait que les clients aillent dans un magasin et présentent leur pièce d'identité, qui était ensuite numérisée et envoyée à un bureau central pour vérification. Le temps moyen pour achever le processus de vérification était de 11 jours.*

*La plateforme d'ARS de Lenddo a été utilisée pour fournir une vérification d'identité en temps réel en quelques secondes en fonction du nom, de la date de naissance et de l'employeur. Cette amélioration de l'expérience client a réduit les fraudes et les erreurs potentielles causées par l'intervention humaine, et a réduit le cout total du processus de vérification.*

*En plus de ses modèles de vérification d'identité, Lenddo utilise une gamme de techniques d'apprentissage automatique pour cartographier les réseaux sociaux et regrouper les demandeurs selon leurs modèles de comportement (d'utilisation). Le résultat final est un LenddoScore™ qui peut être utilisé immédiatement par les IF pour présélectionner les demandeurs ou pour remplir et compléter les propres fiches d'évaluation de crédit d'une IF.*

*Ces algorithmes convertissent un nombre initialement grand de points de données bruts par client en un nombre gérable de caractéristiques et des comportements des emprunteurs avec des relations connues en termes de remboursement de prêts.*

## Cas d'utilisation : La notation de risque de crédit pour les petites entreprises

Les exemples examinés jusqu'à présent ont mis l'accent sur les produits numériques destinés aux consommateurs et aux commerçants du marché de masse. Le flux de données comportementales créées dans les canaux numériques a naturellement généré le plus d'enthousiasme en termes de possibilités d'analyse de données. Cependant, la plupart des IF ont aussi des possibilités étendues de faire un meilleur usage des données en matière d'analyse de crédit et de gestion des risques des produits traditionnels et hors ligne qui comprennent, mais sans s'y limiter :

- Les prêts aux consommateurs
- Les cartes de crédit
- Les prêts et crédits-bails pour les micros, petites, et moyennes entreprises (MPME)
- Les prêts et crédits-bails pour les petits agriculteurs
- La chaine de valeur et le financement de la chaine d'approvisionnement.

Pour ces produits, les IF recueillent de façon classique une profusion de données, mais n'ont pas nécessairement numérisé ou systématisé leur saisie, analyse et stockage. Dans le meilleur des cas, un logiciel de LOS facilite la saisie numérique de données traditionnelles pour favoriser l'analyse des données, notamment la conception de fiches d'évaluation de crédit. À mesure que la chaine de valeur et les paiements de chaine d'approvisionnement se numérisent, il est possible de tirer parti de ces données pour effectuer des prévisions de flux de trésorerie et constituer des notations de crédit.

## Méthodologies de notation de risque de crédit

Les IF ont plusieurs options pour utiliser les données qu'ils recueillent déjà pour modéliser le risque de crédit. Trois des solutions les plus courantes sont de développer des fiches d'évaluation de crédit propriétaires, grâce à une expertise interne, ou en travaillant avec des consultants externes, ou en externalisant la notation de risque de crédit à un fournisseur tiers.

### Développer des fiches d'évaluation de crédit propriétaires

Les banques sur les principaux marchés financiers (par exemple l'Afrique du Sud, l'Amérique du Nord, l'Europe continentale et Singapour) emploient des équipes de grande taille qui développent et entretiennent des modèles, notamment des modèles distincts pour le soutien à la décision de demande, la gestion (comportementale) du portefeuille en cours et le provisionnement. En tant que première étape de développement de modèles internes, les IF peuvent choisir d'utiliser des consultants externes pour mettre sur pied les premiers développements et renforcer les capacités avec du personnel interne par la suite.

De nombreux prestataires de SFN ont des données, des analystes de données, et des spécialistes informatiques en interne capables de gérer leurs propres systèmes de notation. Ces équipes ont toutefois tendance à manquer d'expérience en conception de fiches d'évaluation de crédit. De bons projets d'analyse de données exigent un savoir d'expert pour réussir. Une aide externalisée peut permettre au transfert de connaissances de constituer une expertise en interne dans le cadre de l'appui au projet. Lorsqu'ils travaillent avec des consultants externes, les prestataires de SFN doivent veiller à ce que les outils et les compétences nécessaires soient transférés aux équipes internes de sorte que les fiches d'évaluation puissent être gérées et contrôlées à l'avenir.

## Examen plus approfondi des fiches d'évaluations propriétaires

Un récent projet d'IFC avec une banque en Asie illustre la façon dont le processus peut fonctionner :

1. La banque a partagé ses données de portefeuille passées avec le consultant.
2. Le consultant a préparé les données pour une analyse par le logiciel libre de statistiques « R ».
3. La banque a convoqué un groupe de travail sur la notation de risque de crédit pour qu'il travaille avec le consultant. Dans le cadre d'un atelier, le consultant et le groupe de travail ont analysé et sélectionné des facteurs de risque pour les fiches d'évaluation de prêts aux consommateurs et micro entreprises.
4. La banque a recruté un nouvel analyste pour prendre en charge les fiches d'évaluation (et l'analyste a également participé aux ateliers « R »).
5. Le groupe de travail de notation de risque de crédit et le consultant ont passé en revue les forces et les faiblesses des modèles qui en découlaient pour harmoniser les stratégies d'utilisation avec les objectifs commerciaux et l'appétit pour le risque de la banque.
6. Avec les conseils initiaux du consultant, la banque et son fournisseur de logiciel local ont développé une plateforme logicielle pour déployer la fiche d'évaluation.
7. Le consultant a fourni une assistance à distance en matière de suivi et de gestion de la fiche d'évaluation.

Les avantages et les inconvénients de ces arrangements comprennent :

| **Avantages :** | **Inconvénients :** |
|---|---|
| • La Banque acquiert les compétences nécessaires pour s'approprier les modèles<br>• La Banque a un contrôle total sur ses fiches d'évaluation<br>• Les fiches d'évaluation sont entièrement transparentes | • Cela exige un engagement actif des cadres supérieurs et juniors<br>• Cela nécessite une formation du personnel ou l'intégration de spécialistes de l'analyse de données et de la modélisation des risques<br>• Cela nécessite un logiciel de déploiement supplémentaire, tel qu'un LOS avec une fonctionnalité de notation<br>• Le développement en interne signifie des exigences de maintenance à long terme |

*Tableau 13 : Les avantages et inconvénients des tableaux de bord propriétaires*

### Externaliser la notation de risque de crédit à un fournisseur

La plupart des fournisseurs offrent un développement de modèle personnalisé à l'aide de données de bureau (si elles sont disponibles), des propres données de la banque, ainsi que des données tierces telles que les données de CDR. Normalement, les fournisseurs proposent également un logiciel de déploiement de fiche d'évaluation et s'occupent de la maintenance des modèles pour l'IF. La collaboration avec des fournisseurs de notation de risque de crédit externalise l'expertise en matière de notation et les plateformes logicielles, apportant souvent ainsi de nouvelles données qui seraient autrement inaccessibles. Elle apporte également une expérience internationale et une crédibilité immédiate à la solution de notation.

Voici un exemple de la collaboration de First Access avec une banque en Afrique de l'Est sur le segment des prêts aux petites entreprises, un segment pour lequel les données d'ORM seules ne suffisent pas pour évaluer en intégralité le risque de crédit du demandeur.

# CAS 13

## *First Access : La notation de risque de crédit avec un fournisseur de service complet*

***Externaliser l'expertise en matière de données et travailler avec des partenaires extérieurs***

*Plusieurs IF s'intéressent à la notation de risque de crédit pour accroitre la cohérence et l'efficacité de l'évaluation du crédit pour les petits prêts. Cependant, de moins en moins d'IF sur les marchés en développement ont les compétences en interne pour développer et déployer efficacement des fiches d'évaluations sans aide extérieure.*

*Comme mentionné précédemment, une collaboration avec des fournisseurs de notation de risque de crédit externes externalise l'expertise en matière de notation et les plateformes logicielles, et apporte également souvent une expérience internationale et une crédibilité immédiate à la solution de notation.*

*First Access est l'un des nombreux fournisseurs de notation de risque de crédit, mais surtout l'un des rares à mettre l'accent sur les défis particuliers auxquels sont confrontés les marchés pionniers. Fondée en Juillet 2012, la société a d'abord travaillé largement avec Vodacom Tanzanie, tirant parti de ses données d'ORM pour développer un outil d'auto-décision pour les prestataires de SFN qui dessert des clients à faible revenu sans antécédents de crédit formels. Depuis lors, elle a étendu sa présence à la RDC, au Malawi, au Nigeria, à l'Ouganda et à la Zambie, en concentrant son travail sur les solutions de notation pour le segment des micros et petites entreprises.*

*First Access a collaboré avec une banque en Afrique de l'Est pour développer une fiche d'évaluation pour son activité de (micro) prêts aux petites entreprises, en se consacrant essentiellement aux prêts allant jusqu'à 3 000 USD. La banque prenait en moyenne six jours pour évaluer les demandes de prêt, et en plus de longs délais d'attente, ses PNP étaient en augmentation. Comme beaucoup de banques sur les marchés émergents, elle n'avait aucun outil pour la sélection ou la notation des clients et a donc utilisé un processus pour tous les demandeurs qui frappaient à sa porte.*

*First Access a étudié les historiques de données de portefeuille de la banque pour le segment et a créé un algorithme de notation n'utilisant que les informations disponibles au moment de chaque demande de prêt - sans inclure d'autres données normalement recueillies lors de visites chronophages sur le site de l'entreprise du demandeur, une caractéristique courante d'un processus de souscription de microcrédit. Selon les souhaits de la banque, le modèle a classé les demandeurs en cinq segments de risque.*

*Un « test à l'aveugle » de tous les microcrédits arrivés à échéance décaissés au cours des six derniers mois a indiqué que les notations avaient classé les emprunteurs en fonction du risque, comme l'indiquent les taux de « mauvais » dans le tableau 14 ci-dessous.*

| Segment de risque | A | B | C | D | E |
|---|---|---|---|---|---|
| PAR (Portefeuille à risque) | 1.00% | 3.53% | 9.97% | 22.42% | 26.78% |

***Tableau 14:*** *Classements des emprunteurs de microcrédit en fonction du risque*

*En utilisant l'algorithme de notation, chaque demandeur pouvait être immédiatement noté et affecté à l'un des segments de risque. La banque a ajusté sa procédure d'évaluation de crédit pour offrir une approbation le jour même pour ses clients fidèles dans les segments A et B, qui représentaient 22 pour cent des demandeurs de prêts. Le délai d'approbation de ce groupe de clients a été réduit d'une moyenne de six jours à un jour, ce qui a amélioré l'expérience client ainsi que l'efficacité et la satisfaction du personnel de la banque.*

*Étant donné que les résultats de l'algorithme ont validé en pratique le test à l'aveugle d'origine, la banque élargit l'utilisation de l'algorithme pour effectuer davantage d'approbations et de rejets de prêt le même jour pour les clients réguliers et les nouveaux clients. Les groupes à service accéléré A et B ont augmenté l'efficacité de l'institution en matière de souscription de micro-prêts de 18 pour cent, et les deux groupes ont dépassé les bons résultats de test à l'aveugle, avec un PAR combiné de 1,26 pour cent au lieu des 3 pour cent attendus.*

*La plateforme logicielle First Access permet aux IF de configurer et de gérer leurs propres algorithmes de notation personnalisés et d'utiliser leurs propres données sur leur clientèle et produits de prêt. First Access est en train d'élaborer de nouveaux outils pour sa plateforme afin d'offrir aux IF plus de contrôle et de transparence pour gérer leurs règles de décision, calcul de notation et seuils de risque, avec une surveillance permanente des performances de l'algorithme. Ces fiches d'évaluation d'analyse des performances peuvent permettre aux IF de mieux gérer le risque en réponse aux évolutions du marché.*

*Voici certains avantages et inconvénients de l'externalisation de la notation de risque de crédit à un fournisseur :*

| **Avantages :** | **Inconvénients :** |
|---|---|
| • L'accès à des compétences de modélisation de classe mondiale et à l'expérience internationale | • La banque n'est pas propriétaire du modèle et ne connait habituellement pas le calcul de la notation |
| • La fourniture d'un logiciel de déploiement | • Les couts permanents d'utilisation du modèle et le développement intermittent des modèles |
| • Le délai nécessaire potentiellement plus court pour concevoir et mettre en œuvre une fiche d'évaluation | • Les modèles de tarification pour le développement de notation peuvent ne pas être liés à la souscription des risques |
| • La gestion et le suivi de la fiche d'évaluation et du logiciel | |

*Tableau 15 : Les avantages et inconvénients de la sous-traitance de la notation crédit*

*Une approche externalisée pour le développement de produits de données fournit des solutions rapides et un savoir-faire de bon niveau, mais elle peut aussi signifier des risques de maintenance à long terme, des problèmes de propriété intellectuelle et une exigence que la portée de la conception des projets soit définie en détail dès le départ afin d'assurer des livrables utiles.*

## Accessibilité et le respect de la vie privée

Il existe deux principaux obstacles à l'utilisation des nouvelles formes de données numériques : l'accessibilité et le respect de la vie privée. Pour bénéficier de nouvelles sources de données numériques, les PSF doivent avoir accès à ces données dans un format qui puisse être analysé. Deux des principales façons d'accéder à ces données sont soit d'acheter les données, soit de collaborer avec le fournisseur. Certains ORM, tels que Safaricom au Kenya, vendent des champs de données agrégés prétraités tels que les dépenses moyennes mensuelles ou l'utilisation des appels directement aux PSF. Certains fournisseurs traitent également de grands ensembles de données brutes provenant des ORM, des réseaux sociaux et des données des appareils et les convertissent en profils de clients utilisables et vendables. Les préoccupations concernant le respect de la vie privée ont limité la disponibilité de certaines données, et il n'existe aucune garantie que, par exemple, les données des réseaux sociaux restent une source de données accessible pour les modèles de crédit à l'avenir. Facebook a déjà pris des mesures pour limiter la quantité de données que les services tiers peuvent tirer des profils des utilisateurs,[35] et les données qu'elle rend accessibles par l'intermédiaire de son API ne peuvent être juridiquement utilisées que pour la vérification d'identité. Aux États-Unis, la FTC, qui surveille les règles des données sur le crédit et les consommateurs, a indiqué que les réseaux sociaux risquent d'être soumis à la réglementation des agences d'évaluation des consommateurs si leurs données sont utilisées comme critères pour des prêts.[36]

[35] Seetharaman et Dwoskin, « Facebook's Restrictions on User Data Cast a Long Shadow, » *Wall Street Journal*, 21 septembre 2015

[36] « Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises, » *Site des actualités de la Federal Trade Commission*, 29 novembre 2011, consulté le 3 avril 2017, https://www.ftc.gov/news-events/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep/


# PARTIE 2

## *Cadres de projets de donnees*

## *Chapitre 2.1: Gestion d'un projet de données*

### L'Anneau des données

La gestion de tout projet est complexe et nécessite les bons ingrédients ; une intuition commerciale, l'expérience, des compétences techniques, un travail d'équipe et une capacité à gérer des événements imprévus détermineront la réussite. Il n'existe pas de recette miracle. Cela dit, il existe des moyens d'atténuer les risques et de maximiser les résultats en tirant parti des cadres organisationnels de planification et en appliquant de bonnes pratiques éprouvées. C'est également le cas pour un projet de données. Cette section présente les éléments fondamentaux nécessaires pour planifier un projet de données bien géré à l'aide d'un cadre visuel appelé *l'Anneau des données*.

Les composantes organisationnelles du cadre s'appuient sur les meilleures pratiques du secteur, en identifiant les exigences en ressources générales et les étapes du processus qui sont courantes dans la plupart des projets de données. Il a des points communs avec le Processus de norme interprofessionnelle pour l'exploration de données (CRISP-DM), une approche de processus d'analyse de données qui est devenu célèbre après sa sortie en 1996 et a été largement utilisée au début des années 2000.[37] Son accent mis sur l'exploration de données et les outils informatiques courants il y a deux décennies a entrainé une diminution considérable de l'utilisation de la méthode avec l'avènement des mégadonnées et des techniques de science des données contemporaines. Le site Web d'origine du CRISP-DM a été fermé vers 2014, laissant derrière lui une absence de norme sectorielle spécifique pour les projets de données d'aujourd'hui.

Le cadre d'Anneau des données tire parti des concepts issus des méthodes éprouvées du secteur, avec une approche modernisée correspondant aux technologies et aux besoins des équipes de sciences des données d'aujourd'hui. Il a été développé par Christian Racca

[37] *Processus de norme interprofessionnelle pour l'exploration de données.* Dans Wikipedia, l'encyclopédie libre, consulté le 3 avril 2017, https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining/

et Leonardo Camiciotti[38] comme outil de planification pour aider à déterminer les éléments fondamentaux du projet et réfléchir de façon structurée aux exigences en ressources du projet de données et leurs relations. En collaboration avec les auteurs d'origine et Soren Heitmann, L'Anneau des données et l'outil associé, La Matrice de l'Anneau des données, ont fait l'objet d'une adaptation supplémentaire pour ce manuel. L'idée principale est de fournir un outil qui soutient les chefs de projet tout le long du processus. Ci-dessous figure une liste des manières dont l'outil doit être utilisé :

- **Liste de vérification :** Une liste de vérification ou « liste d'achats », qui permet d'analyser la présence (et les lacunes connexes) des ingrédients nécessaires pour entreprendre un processus fondé sur les données
- **Outil descriptif :** L'Anneau des données est un cadre puissant pour expliquer le processus fondé sur les données (il peut être présenté sous forme de rapport interne, de présentation publique ou de publication scientifique)
- **Miroir de retour d'information continu :** En partant de la définition des objectifs et en terminant par les résultats, chaque cycle d'itération fournit un retour d'information permettant d'affiner le processus et de réévaluer sa conception
- **Outil d'orientation :** Pour préserver *l'orientation* du projet sur les objectifs tout en surveillant des cibles claires

L'approche de l'Anneau des données s'appuie sur l'atténuation des risques et l'amélioration continue ; il est conçu pour éviter les démarrages défectueux, assurer une focalisation sur les objectifs et éviter les scénarios les plus défavorables. Il peut être utilisé comme guide permanent pour définir et affiner les objectifs. Cela permet de garder la phase d'exécution sous contrôle et fournit des résultats de la meilleure manière possible. Le processus de réflexion est circulaire, en demandant aux gestionnaires de réexaminer des questions de planification fondamentale lors de chaque itération, et en affinant, réglant et produisant des résultats. Lorsque des problèmes surviennent, l'idée est d'inciter les gestionnaires à faire le tour de la question, en considérant chaque quadrant de l'anneau comme une source de solution potentielle.

Le schéma de l'Anneau des données est assez complexe, car il représente l'ensemble fondamental des éléments à prendre en compte pour planifier un projet complet. Les chefs de projet peuvent envisager d'imprimer le schéma comme référence visuelle unique pour la conception d'un projet de données. Dans les sections suivantes, chacune de ces structures détaillées sera décomposée étape par étape et traitée. La section conclut en parcourant un cas d'utilisation pour illustrer comment l'Anneau des données peut aussi être utilisé comme outil de planification.

[38] Camiciotti et Christian « Creare valore con i BIG DATA ». *Anneau des données adapté pour le manuel sur les SFN*, Edizioni LSWR (2015) : http://dataring.eu/

## Structures et conception

### Cinq éléments structurels

L'Anneau des données montre l'objectif au centre, entouré de quatre quadrants. Il dispose de cinq éléments structurels : *Objectif*, *Outils*, *Compétences*, *Processus* et *Valeur*. Les quatre quadrants se subdivisent en 10 composantes : *Données*, *Infrastructure*, *Informatique*, *Science des données*, *Activité*, *Planification*, *Exécution*, *Interprétation*, *Ajustement*, et *Mise en œuvre*. Un plan de projet doit viser à intégrer ces composantes et *comprendre leurs interconnexions de manière approfondie*. L'approche organisationnelle de l'Anneau permet aux chefs de projet de définir des ressources et de formuler ces relations ; chaque composante est fournie avec un ensemble de questions de cadre d'orientation, qui sont visuellement alignées à la perpendiculaire de la composante. Ces questions de cadre d'orientation constituent une liste de vérification de la planification des ressources graphiques.

### Objectif : Élément central

La définition d'objectifs clairs est le fondement de tout projet. Mais p résoudre un problème par une solution axée sur les données, sans objectifs quantitatifs et mesurables, présente un fort risque d'échec pour l'ensemble du processus d'analyse des données. Cela se traduit par l'ajout d'une faible valeur ou peut entrainer des interprétations trompeuses.

### Outils et compétences

Les éléments supérieurs de l'Anneau sont axés sur l'évaluation des ressources « pratiques » et « humaines » nécessaires à la mise en œuvre d'un projet de données :

- **Ressources pratiques :** comprennent les données elles-mêmes, les outils logiciels, le matériel de traitement et de stockage
- **Ressources humaines :** comprennent les compétences, l'expertise dans le domaine et les ressources humaines au sens classique pour l'exécution

### Processus et valeur

Les éléments inférieurs de l'Anneau sont axés sur la mise en œuvre et la production de résultat, alors que ces dernières se composent de trois activités concrètes :

1. Planification de l'exécution du projet
2. Génération et manipulation des données - la phase d'exécution
3. Interprétation et réglage des résultats pour mettre en œuvre l'objectif du projet et en extraire la valeur

### Conception circulaire

Un élément central de l'Anneau des données est sa conception circulaire. Elle souligne l'idée d'une amélioration continue et celle d'une optimisation itérative. Ces concepts sont particulièrement essentiels pour les projets de données ; ce sont des éléments établis de conception et de planification de projets correspondant à de bonnes pratiques. Ceci parce que le résultat de tout projet de données est, tout simplement, davantage de données. Prenez par exemple un modèle de notation de risque de crédit. Les données numériques sont saisies : l'âge, le revenu et l'historique des taux de défaut, par exemple. Les résultats sont des notations de crédit, soit davantage de données numériques. Le processus consiste à entrer des données pour en sortir des données.

En fait, ce principe d'entrée et de sortie de données est applicable en permanence dans tout le projet de données. Il peut être appliqué à chaque exploration analytique intermédiaire et test d'hypothèse, au-delà des simples descriptions des conditions de départ et de fin. Le processus circulaire de l'Anneau des données illustre de manière similaire une approche itérative qui vise à affiner, au fur et à mesure des cycles, la compréhension des phénomènes par le prisme de l'analyse des données. Ceci permet une description des causes (données entrantes) et des effets (données sortantes), et l'identification de comportements et de modèles émergents non évidents. Les cinq éléments organisationnels de l'Anneau des données sont conçus pour planifier et atteindre un équilibre entre la spécificité et la flexibilité pendant tout le cycle de vie du projet de données.

En pratique, la planification du projet doit tenir compte de l'élément de chaque anneau sous forme de séquence, en itérant pour suivre le plan général. L'approche circulaire vise à définir les étapes nécessaires pour parvenir à un processus minimum viable. C'est-à-dire quand les

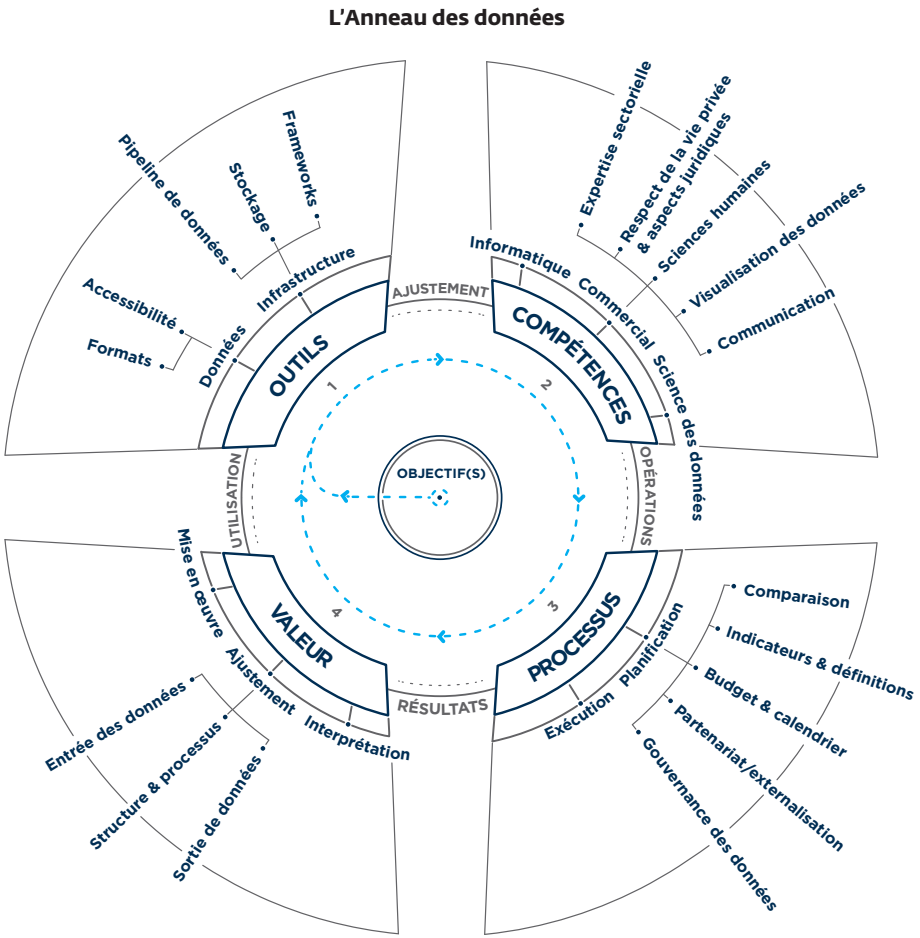



***Figure 19 :*** *l'Anneau des données, un outil de planification visuelle pour les projets de données*

données peuvent être introduites dans le système, être analysées et produire des résultats satisfaisants, puis répétées sans endommager le système ; par exemple, avec un ensemble de données actualisé quelques mois plus tard qui comprend de nouveaux clients. Une fois établi, le projet peut ensuite itérer au prochain niveau pour fournir un produit minimum viable (MVP). Il s'agit du produit de données le plus élémentaire.

Un *produit de données* est un modèle, un algorithme ou une procédure qui prend les données et réintègre de manière fiable les résultats dans l'environnement par le biais d'un processus automatisé. En d'autres termes, ses résultats de sortie sont intégrés dans un contexte opérationnel plus général sans calcul manuel. C'est ce qui constitue un produit de données en dehors d'une analyse particulière. Un produit de données peut être simple-par exemple une visualisation de tableau de bord interactif- mais il existe aussi des produits de données extrêmement complexes, où les notations de crédit sont intégrées à des processus semi-automatisés de prise de décision en matière de prêt, influençant ainsi une nouvelle génération de clients avec des données réinjectées dans le modèle de notation de risque de crédit pour orienter de nouvelles décisions de prêt. Le fait que les produits de données soient consommateurs de leurs propres résultats confirme leur principe circulaire. Le stock de données augmente à chaque itération. Cela met également l'accent sur l'orientation organisationnelle de l'Anneau des données, avec l'objectif placé au centre, qui oriente vers le choix de données à analyser et permet de savoir si le moment est venu ou non de cesser d'itérer et de juger que l'objectif a été atteint.

Commencez petit. Pour les nouveaux projets de données, l'objectif recommandé est un MPV. Il s'agit d'un objectif fondamental et modeste, créé pour tester si un concept de produit axé sur les données est digne d'attention. Une fois atteint, les chefs de projet peuvent prendre en compte les mêmes concepts que l'Anneau des données afin de développer l'échelle du MVP pour en faire un prototype.

## OBJECTIF(S)

L'établissement des objectifs est la première étape de la planification du projet. Le projet doit savoir où il va pour savoir à quel moment il a atteint son but. Dans une certaine mesure, une approche fondée sur le hasard en analyse des données, en particulier lorsqu'il s'agit de structures, de processus et d'organisations complexes, pourrait conduire à des découvertes inattendues et à des trajectoires non planifiées. La découverte est en effet un facteur important pour les projets de données, menant à une exploration et permettant à l'équipe de science des données de « jouer » avec les données. Cela dit, il doit être fait de manière structurée, grâce à des tests d'hypothèses exploratoires, en imitant la méthode scientifique (voir le chapitre 1.1, Méthode scientifique).

L'atteinte de l'objectif signale l'achèvement du projet. Avec une approche itérative, il est particulièrement important de savoir à quoi un projet achevé ressemble pour éviter de se retrouver piégé dans la boucle d'affinement. L'établissement d'indicateurs et de définitions satisfaisants permet d'orienter le projet sur un chemin et émet un signal d'avertissement si le projet commence à s'égarer. Comme pour la gestion opérationnelle, le projet doit à la fois surveiller et évaluer ses ICP pendant tout le processus itératif, en veillant à ce que ces points de référence continuent de servir le projet de la meilleure manière possible.

### Définition de l'objectif

L'objectif est une solution proposée axée sur les données à un problème stratégique afin de produire de la valeur. Les besoins opérationnels du projet sont exprimés par les éléments structurels et les questions d'orientation de l'Anneau des données. Cela se traduit par des besoins précis en ressources, compétences humaines et processus concrets, qui sont tous orientés par les énoncés de problèmes que le projet cherche à résoudre. Il est probable que la déclaration d'objectif et l'énoncé de problème seront définis l'un par rapport à l'autre : vérifiez que l'objectif visé apportera la solution recherchée ; réfléchissez aux nuances du problème stratégique ; ajustez l'un ou l'autre en conséquence. Cela contribue à décomposer de grands problèmes en des problèmes plus distincts, pour avoir l'objectif clair de résoudre un problème clair.

### Énoncé de problème stratégique

L'idée de « résumer le problème avant la solution » contribue à orienter cette approche et permet d'indiquer aux parties prenantes où est la pierre d'achoppement et qui a ce problème. Une fois que l'on a réfléchi au problème, il devient simple de formuler la solution. Voici deux exemples de problèmes stratégiques en matière de SFN :

- **Problème :** Les clients existants ont de faibles taux d'activité du service d'argent mobile
- **Problème :** Les clients potentiels sont exclus de l'accès aux produits de microcrédit

### Énoncé d'objectif

Dans le cadre d'un projet de données, l'objectif est de fournir un processus axé sur les données et un produit avec certaines spécifications. C'est ce qui définit le chemin du projet. Il est également important de savoir si le chemin est bon ; en d'autres termes, si le produit repose sur une hypothèse raisonnable des raisons pour lesquelles cela fonctionne et les résultats sont fiables. Un énoncé d'objectif a deux parties : la spécification du produit et son hypothèse stratégique. Voici deux propositions de solutions par rapport aux énoncés des problèmes précédents :

- **Solution proposée :** un modèle viable minimum de prévision de segmentation des clients pour identifier les utilisateurs actifs à forte propension à augmenter les taux d'activité
- **Solution proposée :** un algorithme de notation de risque de crédit des clients au niveau de la production pour l'émission automatisée de microcrédit

### Processus et spécification de produit

Comme décrit ci-dessus, les deux produits de données illustrés représentent un modèle de prévision de segmentation des clients et un algorithme de notation de risque de crédit des clients. Ceux-ci sont spécifiés par leur échelle, ce qui permet de décrire la « taille » du projet, ou la façon dont il s'intègre dans des systèmes plus généraux.

L'échelle peut être envisagée selon la progression suivante :

- **Processus :** données d'entrée qui produisent des données de résultats de manière fiable par le biais d'un processus automatisé
- **MVP :** un concept et un processus de produit dont les résultats mettent en évidence une valeur essentielle
- **Prototype :** concept de produit avec une mise en œuvre, une facilité d'utilisation et une fiabilité de base
- **Produit :** un concept qui a fait ses preuves avec une mise en œuvre fiable et une proposition de valeur qui a fait ses preuves
- **Production :** un produit systématiquement mis en œuvre et livré aux utilisateurs ou aux clients

Encadrer l'objectif en termes d'échelle contribue à définir à la fois les besoins en ressources et la façon dont les composantes générales du projet doivent bien se combiner. Une validation de principe de MVP pourrait être livrée sur un seul ordinateur portable dans quelques semaines. En comparaison, l'échelle de niveau de production pourrait nécessiter des serveurs de données spéciaux, des experts pour assurer leur maintenance et une supervision juridique pour garantir la sécurité des données. Néanmoins, la production d'un MVP nécessite des ressources pratiques et humaines (c.-à-d. l'infrastructure et les personnes), organisées selon un processus minimum viable. Cela signifie qu'il faut définir des rôles organisationnels ainsi que des relations de gestion et de rapports clairs. Il s'agit de la manière dont une solution axée sur les données pour un problème stratégique est rendue opérationnelle, la manière dont les défis techniques sont identifiés et résolus, et la manière de s'assurer que le produit concret offre une valeur stratégique.

### Hypothèse

Ce que ces produits de données parviennent à accomplir est fonction d'une hypothèse sous-jacente qui n'est implicite que dans ces deux exemples. L'identification des utilisateurs actifs à forte propension repose sur une hypothèse opérationnelle ; il existe une corrélation entre les variables qui définissent ces segments de clientèle et les taux d'activité. Par exemple, les clients ayant un temps de communication vocale élevé ont des taux d'activité plus élevés. Il s'agit d'une hypothèse statistiquement vérifiable et, en fin de compte, il incombe à l'équipe de science des données de le démontrer. Si la corrélation est forte et fiable, cette hypothèse axée sur les objectifs apporte au produit de données de la crédibilité et de la fiabilité. Une hypothèse similaire pourrait être formulée pour un modèle de notation de risque de crédit afin de tester par exemple l'hypothèse suivante : les clients ayant de petits réseaux sociaux ont des taux de défaut de remboursement des prêts plus élevés. La formulation d'une hypothèse ne se limite nullement aux projets de données fondés sur des algorithmes. Un tableau de bord de visualisation correspond également à une hypothèse sur les relations entre les données que l'on cherche à visualiser. Une telle hypothèse peut ne pas être testée statistiquement par des algorithmes, mais la fiabilité de la visualisation implique que ces relations soient cohérentes et valables au fil du temps. Pour cette raison, la visualisation continuera à raconter une histoire significative ou à orienter une prise de décision utile.

Le principe de « recherche reproductible » est devenu important chez les scientifiques des données. La *recherche reproductible* décrit des approches transparentes et reproductibles de l'analyse et la façon dont des résultats sont obtenus dans la première étape de mise à l'échelle du « processus ». En principe, il s'agit de permettre une validation indépendante des résultats, qui peut être pertinente à des fins réglementaires ou d'audit. C'est pourquoi la première étape de l'itération lors de l'utilisation de l'Anneau des données consiste à formuler un processus minimum viable ; il fait en sorte que le projet obtienne des résultats fiables sur lesquels repose la valeur essentielle du produit. Ce processus prend également en charge les produits de données pour voir immédiatement

si et quand les hypothèses deviennent peu fiables, ce qui peut inciter à réajuster les modèles afin d'assurer une fiabilité continue.

### Risques et atténuations des objectifs

L'établissement d'objectifs de projet en termes d'hypothèses formulées, testées et affinées contribue à atténuer les risques courants en matière de projets de données. Les risques d'une mauvaise définition des objectifs sont les suivants :

### Risque : Ne pas poursuivre les objectifs

Le risque principal est l'absence de motivation et d'objectif stratégiques du projet, ou la non définition de véritables objectifs. En d'autres termes, ce risque inclut les motivations pour faire quelque chose de significatif avec les données pour des raisons d'attrait, dans le but d'utiliser des termes populaires à la mode parce que les concurrents le font ou simplement parce qu'ils sonnent comme scientifiquement ou technologiquement solides- alors que les motivations n'ont pas de contrepartie axée sur la valeur. Cette approche pourrait conduire à des résultats inutilisables ou à des dilapidations de budgets car elle représente une occasion manquée de tirer parti de l'analyse pour fournir des résultats axés sur les objectifs qui sont pertinents pour l'organisation. Pour ceux qui sont particulièrement motivés pour faire quelque chose, il n'est pas rare d'embarquer des ressources externes qui sont simplement chargées de découvrir quelque chose d'intéressant. Le risque est d'obtenir des résultats qui non seulement sont inutilisables, mais faux, car une exploration sans fin peut permettre une analyse biaisée ou des résultats forcés pour livrer quelque chose.

**Atténuation :** Savoir ce que le projet vise à accomplir. Si l'équipe veut faire quelque chose, mais ne sait pas par où commencer, elle doit engager des spécialistes des opérations de données pour examiner les données et contribuer à mettre en évidence les types d'indications pertinentes qu'elles pourraient fournir à l'entreprise. L'objectif du projet est généralement validé par la mesure des résultats, mais il est important de remarquer que les tests d'hypothèses se révèlent souvent faux. C'est une bonne chose. Ou on itère et on réussit, ou on accepte que l'indication ne fonctionne pas et on retourne à la phase de conception. C'est une meilleure situation que d'avoir un résultat bon ou intéressant fondé sur de mauvaises données.

### Risque : Manque d'orientation

Les projets sans véritables objectifs comprennent aussi les projets trop généraux, mal définis ou excessivement souples et changeants. L'objectif définit l'orientation et décrit ce qui sera réalisé. Le manque de clarté peut amener les équipes à se distraire ou à analyser des questions auxiliaires, débouchant ainsi sur des résultats auxiliaires. En prenant cela en compte, une certaine souplesse doit exister pour un affinement itératif des objectifs et pour permettre d'explorer et de capitaliser sur une heureuse découverte. Le manque d'orientation peut également résulter d'une incompatibilité entre le problème et la solution. C'est à ce moment-là que le problème stratégique sous-jacent peut ne pas être défini avec précision, ou lorsque la solution pour atteindre l'objectif proposé présente une incohérence logique, telle qu'un lien commercial ou stratégique ténu avec le problème qu'elle est censée résoudre.

**Atténuation :** Définissez des objectifs clairs et précis en intégrant une pertinence commerciale dans chacune des composantes de l'hypothèse problème-produit. Assurez-vous qu'ils peuvent être affinés par une approche itérative et révisez-les au fur et à mesure que le projet progresse. De plus, assurez-vous que les objectifs sont pertinents en permanence à mesure que la stratégie commerciale évolue de manière indépendante. Prévoyez un degré d'exploration et de souplesse dans l'exécution du projet. L'établissement de limites exploratoires est essentiel, car elles permettent d'éviter que les projets s'égarent, tout en permettant une latitude de découverte. Cela est également soutenu par les unités de mesure et les cibles associées spécifiques, ou ICP, tant pour les objectifs intermédiaires que pour l'atteinte de l'objectif global.

### Risque : Non axé sur les données

L'économiste de renom Roland Coase a déclaré : « si vous torturez les données assez longtemps, elles vont avouer ». Le risque oblige les données à révéler ce à quoi on s'attend pour tenter de valider les connaissances, les comportements ou l'organisation souhaités. Passer à une approche fondée sur les données signifie être prêt à observer les faits concrets à mesure qu'ils émergent de l'analyse des

données. En d'autres termes, l'analyse de projets, de processus ou de procédures par des données peut conduire à des résultats qui ne correspondent pas aux croyances, aux réflexions ou à la stratégie actuelles, obligeant ainsi une organisation à opérer un changement profond.

**Atténuation :** S'inspirer de la méthode scientifique pour définir des objectifs de projet assortis de délais et appuyés par des hypothèses qui sont rigoureusement testées. Assurez-vous que la stratégie d'exécution utilise le concept de recherche reproductible pour mieux permettre la possibilité de répétition et la validation indépendante des résultats. De plus, assurez-vous que les promoteurs de projets comprennent parfaitement que la découverte de modèles précieux n'est pas garantie.

### Risque : Manque de pragmatisme

Les objectifs doivent être réalistes quant aux ressources du projet et aux attentes de son promoteur, par exemple concernant les compétences, l'infrastructure ou le budget appropriés.

**Atténuation :** Assurez-vous que l'échelle du produit fait partie intégrante de l'énoncé de l'objectif. Cela contribue à délimiter le projet et à pousser les chefs de projet à faire correspondre les ressources et les exigences. En outre, assurez-vous qu'un spécialiste des technologies de l'information et de la communication (TIC) effectue une évaluation informatique technique de la conception du projet afin de veiller à l'existence d'un pragmatisme entre l'objectif du projet et les outils techniques acquis pour l'atteindre.

## Quadrant 1 : OUTILS

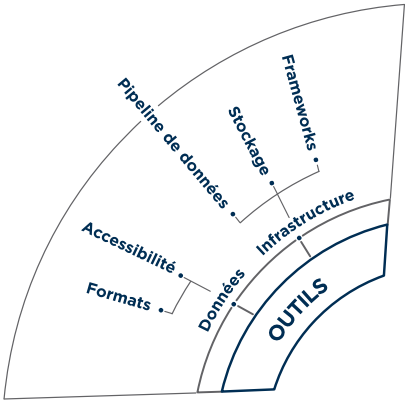



*Figure 20 : Anneau des données Quadrant 1 : OUTILS*

Le monde et ses phénomènes dynamiques peuvent être observés et fragmentés en données. *Autrement dit, les données sont des échantillons de la réalité, enregistrés sous forme de mesures et stockés sous forme de valeurs*. En outre, les systèmes complexes donnent une fausse impression d'un savoir approfondi, intégré dans le comportement collectif des différentes composantes du système. Les composants pris isolément peuvent ne rien révéler, mais des modèles apparaissent lorsqu'on observe l'ensemble du système.

La révolution des données a permis une augmentation exponentielle du volume, de la vitesse et de la variété des données numériques. Cette disponibilité accrue de données numériques permet une plus grande précision de la compréhension des processus, des activités et de leurs interrelations. Pour tirer des connaissances et de la valeur de leur analyse, les données doivent être stockées, décrites de manière appropriée et rendues accessibles. Cela nécessite qu'une infrastructure technique appropriée soit mise en place pour gérer les données, leur accessibilité et leur calcul. Cela permet également d'accéder à l'analyse complète du système et aux modèles très attrayants qui peuvent générer de la valeur. Le premier quadrant de l'Anneau des données demande aux chefs de projet de réfléchir à leurs données et à l'infrastructure technique nécessaire pour les analyser selon deux composantes : les données et l'infrastructure.

### Outils : Données

Les données sont les contributions (et les résultantes) fondamentales d'un projet de données. Les questions d'orientation de l'Anneau des données sont regroupées en deux principes : l'accessibilité et le format. Ce sont des éléments essentiels qui affectent profondément les besoins en ressources et les décisions en matière de processus.

Tout d'abord, il faut savoir comment les données sont décrites, leurs propriétés, et si elles représentent des nombres, du texte, des images ou du son. Il faut savoir aussi si elles sont structurées ou non structurées. Les données doivent également être compréhensibles pour les êtres humains et doivent exister dans un format numérisé et utilisable par une machine. Ces paramètres de base sont pertinents pour les données de toutes tailles et formes. Ce sont là des facteurs critiques pour déterminer la meilleure infrastructure technique à utiliser pour le projet. Voir le chapitre 1 pour une discussion plus détaillée sur les formats de données.

Récemment, le concept de mégadonnées a pris une grande importance. Il s'agit d'un concept utile, mais sa prépondérance a également créé des opinions erronées. En particulier la croyance que la simple disponibilité d'une grande quantité de données peut accroitre les connaissances ou fournir de meilleures solutions à un problème. C'est parfois vrai. Et parfois, ce n'est pas le cas. Bien que les mégadonnées puissent fournir des résultats, il est également vrai que les « petites » données peuvent réussir à atteindre les objectifs du projet. Il est important pour le chef de projet de s'assurer que les bonnes données (et suffisantes) sont disponibles pour la tâche et que les bons outils sont en place.

La définition de « méga » est en constante évolution, donc insister sur le terme lui-même profite rarement à un projet. L'aspect le plus utile du concept de mégadonnées est de comprendre que plus un ensemble de données est important, plus il faudra de temps pour l'analyser. Dans cet esprit, un ensemble de données plus important exige également des capacités d'équipes techniques plus spécifiques et une infrastructure technique plus complexe, plus sophistiquée ou plus couteuse pour la gérer. L'aspect « méga » des données peut également être lié à l'échelle d'un objectif ; on peut parvenir à un MVP avec un simple instantané des données, alors la production peut s'attendre à traiter des données transactionnelles continues à grande vitesse. Il s'agit là d'un élément important du processus de conception du projet ; le *fait d'avoir* des téraoctets de transmission de données en continu ne signifie pas pour autant que l'objectif d'un projet est atteint.

Les questions de cadrage suivantes permettent d'identifier les sources de données et de définir leur étendue en fonction des besoins en ressources du projet. Si les systèmes de données internes ne saisissent pas ce qui est supposé, cela oblige la planification des ressources du projet à effectuer des changements en identifiant les nouvelles ressources de données requises :

- Quelles données sont produites ou collectées par le biais d'activités de base ?
- Comment ces données sont-elles produites (par exemple, quels produits, services, points de contact) ?
- Les données sont-elles stockées et organisées ou passent-elles par le processus ?
- Les données se présentent-elles sous une forme lisible par une machine et prêtes à être analysées ?
- Les données sont-elles propres, ou existe-t-il des irrégularités, des valeurs manquantes ou corrompues ou des erreurs ?
- Les données disponibles sont-elles statistiquement représentatives pour permettre des tests d'hypothèses ?
- Quelle est la relation entre la taille des données et les besoins en matière de performance ?

Ces questions montrent bien le travail nécessaire dans la phase initiale pour réussir à acquérir, nettoyer et préparer le ou les ensembles de données pour une analyse ultérieure. En fonction de la quantité de contrôle disponible dans l'ensemble du processus axé sur les données, cette phase de préparation sera d'une longueur variable, ce qui signifie des couts de projet variables. Une planification inadéquate des données initiales peut entrainer un gonflement des couts au bout du compte ; des révisions pourraient signifier le besoin de choisir une autre infrastructure informatique ou des capacités d'équipe différentes.

### Accessibilité aux données

Les données doivent être consultées pour être utilisées. Cela peut sembler évident, mais cette question est complexe et doit être prise en compte dès le début de chaque processus axé sur les données afin de s'assurer que les résultats sont atteints dans le temps et le budget impartis - ou que des résultats sont même possibles. Le respect de la vie privée du client, la demande et l'octroi d'autorisations d'utilisation de données, et la définition de la propriété et de l'intérêt légal une fois que les autorisations d'accès aux données sont accordées, sont des facteurs qui complexifient l'accessibilité des données, nuisent à son uniformité dans tous les environnements réglementaires et font l'objet de préoccupations éthiques. L'accessibilité des données peut être évaluée selon trois facteurs :

#### Juridique

Des réglementations pourraient empêcher une analyse fondée sur les données excellente et bien conçue d'être réalisée dans son intégralité. Cela interromprait le processus à une phase intermédiaire, il est donc essentiel de connaitre les contraintes juridiques dès le départ.

**La propriété** des données doit être définie, en identifiant qui a la permission de les

analyser pour en tirer des indications. Si des accords de propriété intellectuelle sont en place, ils doivent couvrir les travaux existants et dérivés. Si l'analyse est une étude collaborative, des accords de publication doivent être mis en place, notamment sur la clarté de ce qui constitue une information exclusive et ce qui peut être rendu public.

**L'utilisation éthique** des informations peut également apporter des contraintes juridiques. Les données concernant les personnes, les groupes ou les organisations doivent être traitées avec attention, en prenant la sécurité comme première priorité. Les règles de confidentialité des données peuvent également influencer la façon dont les données peuvent être transférées ou non de leur propriétaire à l'analyste, par exemple en sachant si elles peuvent être envoyées par voie électronique ou par stockage physique. En outre, les règlementations peuvent stipuler des procédures concernant les données quittant les frontières nationales, celles qui sont acheminées via des tiers ou stockées sur des serveurs situés dans des pays particuliers.

### Technologique

Des obstacles peuvent se dresser si le format de données n'est pas aligné avec la technologie choisie pour le traitement et l'analyse des données. Pour prendre un exemple simple, un algorithme de TLN ne peut pas être appliqué de manière significative à des données sous forme d'images. De façon plus pratique, les bases de données sont généralement optimisées pour des types spécifiques de données ; et certaines technologies ne sont pas conçues pour fonctionner ensemble, tout comme l'établissement d'un flux de travail visant à mélanger des produits Apple et Microsoft. Cela peut entrainer des couts et des inefficacités, et peut créer des problèmes supplémentaires à résoudre si l'on tente d'opérer des harmonisations forcées.

**Les données numériques** sont requises pour les analyser à l'échelle et la vitesse d'une machine. Il peut exister des exceptions à la règle, avec différentes nuances, et l'IA repousse ces limites.

**La compatibilité** est nécessaire entre le format de données et la technologie utilisée pour les gérer. Même si les ensembles de données sont numérisés, ils peuvent être isolés et inaccessibles en raison de choix technologiques incompatibles réalisés par différents services d'une même société, d'un même gouvernement ou d'une même organisation. Des systèmes obsolètes pourraient parfois être en place, ce qui pourrait également empêcher les interactions avec des solutions, des langages et des protocoles modernes. La quantité d'effort pour harmoniser l'infrastructure technologique pourrait être une barrière non négligeable du point de vue du cout pour le temps consacré.

### Stratégique

Les parties prenantes pourraient chercher à préserver un avantage concurrentiel en interdisant l'accès à leur actif de données. Cela prend généralement forme d'une des trois façons suivantes : en nécessitant du matériel ou un logiciel spécial pour lire les formats de données propriétaires ; en contrôlant la manière dont les données peuvent être utilisées ; ou en exigeant des redevances de licence particulières. Alors que les facteurs technologiques pourraient offrir une solution - bien que parfois complexe ou inefficace-, les facteurs stratégiques sont encore souvent définis pour s'assurer délibérément que l'accès n'est possible que selon les spécifications du propriétaire des données, ou peut-être que l'accès est totalement refusé.

### Format des données

Les données numériques peuvent être représentées sous de nombreuses formes différentes et un *format de données* décrit les paramètres de données compris par des humains (c'est-à-dire les textes, les images, les vidéos, les données biométriques). Souvent, le format est indiqué par le suffixe de trois ou quatre lettres à la fin d'un fichier informatique. Le format peut également indiquer plus généralement des structures et des bases de données de stockage, par exemple : Oracle, MongoDB et JSON (voir le chapitre 1.1, Définition des données).

Il existe de nombreux formats de données, notamment selon les approches de stockage et de traitement. Le format de données est fortement déterminé par le contexte commercial ou organisationnel et, en particulier, par les personnes responsables de la gestion de la création, du stockage et du traitement des données. Pour les chefs de projet, *le fait d'identifier les problèmes de fragmentation de format et d'incompatibilité* est essentiel pour définir l'harmonisation des données nécessaire à des projets bien conçus. Comprendre les valeurs enregistrées dans un ensemble de données, ainsi que des métadonnées plus générales d'un ensemble de données, permet aux chefs de projet de planifier correctement.

Une valeur de *point de données* se rapporte au contenu intrinsèque d'un enregistrement de données. Ce contenu peut être exprimé sous forme numérique, temporelle ou textuelle, appelée *type de données*. Pour l'analyse de données, le facteur crucial est que ces valeurs sous-jacentes ne soient pas affectées par des erreurs ou des biais systématiques dus à des petits problèmes d'infrastructure ou humains. Généralement, les chefs de projet ne tiennent pas compte de la façon dont les données sont recueillies ou si la méthode de mesure est bien définie. Il est utile de comprendre comment ces mesures sous-jacentes sont effectuées et de s'assurer qu'il existe un transfert approprié des connaissances entre les propriétaires de données et les analystes de données quant aux principaux problèmes de mesure. À titre d'exemple pratique, si un système a été interrompu pendant une mise à jour informatique, cette mise à jour se traduira par une baisse spectaculaire des transactions. Les analystes doivent être conscients de ces informations pour interpréter correctement l'anomalie. Les anomalies des valeurs de données influent grandement sur le processus de nettoyage des données et la planification de projets connexe.

*Les métadonnées* sont des « données sur les données », qui comprennent toutes les informations de base supplémentaires qui enrichissent un ensemble de données et le rendent plus compréhensible. Les colonnes de titre dans une feuille Excel sont des métadonnées (les titres sont eux-mêmes des données textuelles qui décrivent les valeurs dans les lignes suivantes). Par exemple, imaginez un ensemble de données avec les titres, « nom de l'agent » et « volume des transactions, » suivis d'une colonne de nombres sans titre. Ces chiffres sont-ils liés à des valeurs de transaction, peut-être les heures où les opérations ont eu lieu ? Si le projet cherche à visualiser des volumes sur une carte, la localisation de l'agent devient également une exigence de données ; le processus de calcul doit être en mesure de demander à l'ensemble de données de fournir toutes les valeurs de localisation. Si la catégorie des localisations ne se compose pas de métadonnées définies, le processus ne sera pas alors en mesure de trouver de coordonnées GPS à tracer. La solution pourrait être simple, disons, en ajoutant un titre « localisation » à cette colonne sans titre. De cette façon, les équipes de projet peuvent ajouter des informations contextualisées aux ensembles de données et fournir des descriptions plus détaillées des données (par exemple des métadonnées) que le processus d'analyse peut alors questionner et utiliser. En ce sens, les métadonnées ne sont tout simplement qu'un nouvel ensemble de données. Les métadonnées sont particulières car elles sont intrinsèquement liées à l'ensemble de données sous-jacent, qui permet à ce processus de questions-réponse d'avoir lieu. Il ne s'agit que d'un exemple ; les métadonnées ont plus de valeur que de simples titres de colonnes. Même dans Excel, les métadonnées existent *à propos de* la feuille de calcul en cours de constitution, par exemple, la taille du fichier, la date de création et l'auteur sont tous des exemples de métadonnées. Ces métadonnées sous-jacentes permettent la recherche et le tri de fichiers, par exemple, le système d'exploitation peut demander tous les fichiers modifiés de la semaine précédente. Les réponses sont obtenues via les métadonnées du fichier.

Comprendre comment les ensembles de données sont connectés via des métadonnées est un élément clé de la conception de projet et de l'identification des lacunes et des possibilités d'analyse. Les métadonnées permettent d'identifier les domaines dans lesquels les données supplémentaires peuvent être nécessaires pour atteindre les objectifs du projet, et la façon de lier de nouveaux ensembles de données en cas de besoin. Les métadonnées permettent d'identifier des possibilités d'optimisation là où des ensembles de données supplémentaires pourraient déjà exister ; l'obtention sous licence de données tierces peut combler les lacunes et des métadonnées dérivées ou synthétiques pourraient être créées pour contribuer à adapter les ensembles de données du projet au contexte. Pour les chefs de projet, il est important de savoir le moment et l'endroit où les métadonnées sont susceptibles d'exister. Si elles ne font pas partie des ensembles de données de départ, il peut être préférable de demander aux propriétaires de données ces informations, plutôt que de les adapter au contexte dans le cadre du travail sur le projet.

### Outils : Infrastructure

Comme expliqué précédemment, les données sont la contribution (et la résultante) fondamentale d'un projet de données. On appelle infrastructure l'endroit où les données vont et sortent physiquement. Les données sont des informations numériques qui doivent être acquises, stockées, traitées et calculées à l'aide d'outils informatiques s'exécutant sur des ordinateurs virtuels ou physiques.

L'infrastructure technologique doit être adaptée aux objectifs qui se posent en en

termes de *volume*, de *variété* et de *vitesse* des données. Les ressources de l'infrastructure déterminent la facilité d'utilisation des données et influent fortement sur la « puissance » et l'efficacité des algorithmes scientifiques et des modèles mathématiques appliqués. L'infrastructure générique axée sur les données est constituée de ces éléments fondamentaux :

### Pipeline de données

Le pipeline de données est une chaine fonctionnelle d'équipement matériel ou de logiciels où chaque élément reçoit des données d'entrée, les traite, puis les transmet à l'élément suivant. Il représente la manière dont les données sont téléchargées dans le processus analytique ; le pipeline de données comprend le processus de téléchargement, des outils pour calculer les chiffres, la façon dont les chiffres sont téléchargés, et comment ils sont ensuite introduits dans un processus opérationnel. Par exemple, ce pipeline permet l'intégration technique d'un produit de données dans des systèmes d'entreprise plus généraux. Le pipeline doit être prévu pour assurer un processus fiable qui avale des données brutes et produit des résultats utilisables. Le projet doit veiller à ce qu'un schéma ou un diagramme de flux soit écrit pour décrire la mise en œuvre fonctionnelle du pipeline. Le téléchargement initial dans le pipeline marque généralement le début opérationnel d'un projet de données, en commençant par le processus d'extraction-transformation-chargement (ETL) des données. L'ETL est un plan procédural, dans le cadre de la gouvernance des données du projet, qui sera traité de manière plus approfondie plus tard.

### Stockage

Un système de base de données ou de fichiers est appelé stockage, c'est-à-dire l'élément de l'infrastructure destiné à stocker des données. Le stockage affecte la façon dont les données sont enregistrées et récupérées et ces processus d'entrée et de sortie sont essentiels à la conception d'un système performant. Il faut du temps pour écrire des données sur un disque, et quand une requête arrive, il faut du temps pour rechercher la réponse et l'envoyer à l'étape suivante du pipeline de données. Les bons outils de base de données sont souvent orientés par la nature des données elles-mêmes, leur format et leur structure. En outre, la façon dont les données sont utilisées joue un rôle dans le stockage ; un système d'archivage vise à compresser un maximum de données dans un volume aussi peu couteux que possible, alors qu'une base de données transactionnelle garantit la rapidité et la fiabilité de sorte que les clients n'aient pas à attendre. Les cadres guident également le choix des bases de données en fournissant des outils intégrés optimisés pour des solutions et des conceptions de stockage spécifiques.

### Frameworks

Un Framework est un ensemble de solution conçu pour un groupe de problèmes. Techniquement, il s'agit d'un ensemble de bibliothèques prédéfinies et d'outils communs pour permettre d'écrire du code et des programmes plus rapidement et facilement. Dans le domaine des mégadonnées, celles-ci comprennent des plateformes qui recueillent des outils, des bibliothèques et des fonctionnalités afin de simplifier la gestion des données et les processus de manipulation (par exemple, Apache Spark, Apache Hadoop, Hortonworks, Cloudera. Voir le chapitre 2.2.3, Base de données technologique). Il convient de noter qu'un projet peut intégrer plusieurs Frameworks. L'utilisation d'un Framework reconnu est recommandée, car cela évite la nécessité de programmer des outils communs à partir de zéro, ce qui peut représenter d'énormes économies en temps et en couts. Le compromis est que l'approche du projet doit s'adapter à la manière qu'a le Framework de résoudre l'ensemble des problèmes pour lesquels il a été conçu, ce qui peut ou ne peut pas parfaitement répondre aux besoins précis du projet. Et un mauvais choix de Framework risque de mal adapter son approche en termes de solutions face aux problèmes du projet, créant ainsi de l'inefficacité.

Les Frameworks sont généralement conçus à partir de spécifications matérielles, et ils s'exécutent en fin de compte sur des ordinateurs qui font les calculs pour le projet de données. Alors que la puissance de calcul brute est également un élément critique de l'infrastructure du projet, il est préférable de planifier le premier pipeline de données, les besoins de stockage et les Frameworks nécessaires pour répondre aux besoins du projet. Les spécifications informatiques adéquates ont tendance à se mettre en place par la suite. La conception et la gestion de l'infrastructure ne relèvent généralement pas du rôle des chefs de projet, mais ils doivent néanmoins s'assurer que des capacités et des ressources soient disponibles pour répondre aux besoins du projet. C'est la raison pour laquelle une évaluation informatique est spécifiquement indiquée dans le cadre de la gestion des risques et de la définition d'objectifs pragmatiques. Compter sur des équipes informatiques internes ou s'assurer

d'une capacité pertinente de l'équipe du projet de données sont des éléments essentiels pour permettre d'évaluer les exigences d'infrastructure et les besoins techniques, notamment l'évolutivité, la tolérance aux pannes, la distribution ou l'isolement de l'environnement. Ces termes techniques sont utiles pour une infrastructure informatique d'entreprise à grande échelle ; les objectifs de MVP peuvent être obtenus avec beaucoup moins. Même les petits projets de données sont susceptibles d'impliquer l'architecture de l'entreprise dans le pipeline de données. Les données dont un projet a besoin puiseront certainement dans les systèmes de l'entreprise ; l'étendue de cet état de fait doit être bien déterminée, et ses conséquences planifiées et coordonnées avec les équipes informatiques.

## Quadrant 2 : COMPÉTENCES

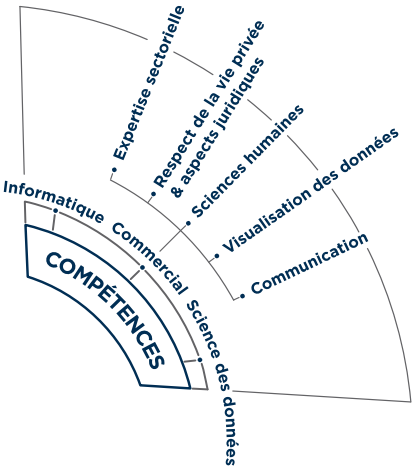



*Figure 21 : Quadrant 2 de l'Anneau des données : COMPÉTENCES*

Les projets fondés sur les données ont besoin de scientifiques des données. Cela dit, « scientifique des données » est un titre relativement vague et général, qui est encore à définir avec précision. Pendant ce temps, le secteur et les médias ont fait un battage médiatique sur les mégadonnées, l'apprentissage automatique et toute une série de technologies, tout en créant une prise de conscience plus générale sur l'immense valeur potentielle des données. Cela a créé une pression incitant à investir dans ces ressources afin de faire face à la concurrence. Il est essentiel pour le gestionnaire de projet axé sur les données d'être conscient que des ensembles très spécifiques de compétences et d'expérience technique sont nécessaires pour établir les exigences d'un projet de données. Ils doivent être conscients d'une manière toute aussi cruciale que bon nombre de ces domaines d'expertise se forment de façon dynamique parallèlement à l'évolution rapide de la technologie. Le deuxième quadrant de l'Anneau des données demande aux chefs de projet de réfléchir aux ressources humaines nécessaires pour réaliser le projet selon trois composantes : l'informatique, la science des données et l'entreprise.

### L'équipe

Le montage de la bonne formule d'ensembles de compétences est un défi pour les chefs de projets de données en raison de l'évolution dynamique de la technologie, des ensembles de données de tailles de plus en plus importantes et des compétences requises pour tirer de la valeur de ces ressources.

**Un scientifique des données représente généralement une équipe de personnes** qui traitent des données. Au-delà d'une seule compétence, cela nécessite généralement une équipe interdisciplinaire d'experts techniques qui interagissent fortement avec toutes les unités – une seule personne ou un groupe – qui gèrent les données, de l'acquisition à la visualisation.

**Les équipes sont dynamiques et travaillent en collaboration**, et il est difficile de suivre le rythme de l'innovation et du développement de nouveaux ensembles de compétences, de l'expertise émergente et de l'hyperspécialisation qui va s'accentuant. L'externalisation des capacités peut permettre d'atteindre le dynamisme nécessaire et les ensembles de compétence adéquats. Vous pouvez également conserver ou constituer une équipe fondamentale de généralistes de la science des données qui peut contribuer à assurer une collaboration réussie au sein de toute une équipe multidisciplinaire de spécialistes des données et des opérations commerciales.

**Une culture d'ouverture, scientifique et axée sur les données est nécessaire**. Une approche scientifique appropriée et une *culture des données* doivent être partagées au sein de l'équipe et, idéalement, dans toute l'entreprise. Parce qu'une fixation d'objectifs réussie repose sur l'imitation de la méthode scientifique et des tests d'hypothèses exploratoires, l'équipe de scientifiques des données doit être menée par un sens de la curiosité et de l'exploration. Le chef de projet doit veiller à ce que la curiosité soit dirigée et maintenue sur la cible.

Les questions d'encadrement suivantes aideront les chefs de projet à identifier les ressources et les besoins :

- Qui est responsable de la gestion des données dans l'entreprise ? De quelle façon ?
- Existe-t-il des collaborations en cours avec des instituts de recherche ou des organismes qualifiés pour réaliser les activités de science des données ?
- Quels sont les canaux de recrutement existants concernant les professionnels spécialisés dans les données ?
- Comment la culture de données est-elle promue dans l'entreprise, et qui est impliqué ?
- En quoi la collaboration multidisciplinaire est-elle favorisée dans la planification et l'exécution du projet ?
- Comment la validité scientifique est-elle assurée dans le choix des algorithmes et des représentations de données mathématiques (modélisation) ? Une personne qualifiée s'assure-t-elle que les résultats sont vrais ?
- Qui garantit que les bonnes pratiques sont en vigueur et que les algorithmes sont programmés de manière efficace ?
- Existe-t-il une collaboration ouverte entre l'équipe spécialisée dans les données et d'autres unités opérationnelles ?

Une équipe complète et hautement interdisciplinaire est difficile à mettre sur pied, et la plupart des entreprises n'auront probablement pas toute l'étendue des compétences pertinentes pour tirer parti de la demande. La compréhension de ces lacunes est généralement la première étape pour être conscient du plein potentiel et de la planification des investissements dans l'externalisation, qui sont considérés comme une partie intégrante de la planification de processus.

## Compétences : Informatique

Les données sont des éléments d'information numériques qui doivent être acquises, stockées, traitées et gérées par des outils informatiques, des langages de programmation et de script, et des bases de données. Par conséquent, les compétences doivent rassembler des connaissances sur les éléments suivants :

### Informatique en Cloud

Lorsque les données sources sont « méga » ou immenses, les outils normaux de programmation et les ressources informatiques locales, telles que les ordinateurs personnels, deviennent rapidement insuffisantes. Les solutions « en nuage » sont une réponse pratique et efficace à ce problème, mais elles signifient la maitrise de connaissances essentielles sur les systèmes de virtualisation, en mettant à l'échelle les paradigmes et la programmation de Frameworks (voir le chapitre 2.2.3, Base de données technologique).

### Langages des scripts

Travailler avec une infrastructure informatique signifie coder. Python ou R sont souvent les meilleures options pour obtenir rapidement des prototypes et explorer des modèles de données. Ce sont des choix probables pour un objectif de MVP et le développement de projets à un stade précoce. Les deux langages de script sont devenus quasi incontournables en tant qu'outils de science de données, et l'équipe doit idéalement être capable de programmer dans les deux langages (voir le chapitre 2.2.3, Base de données technologique).

Certaines infrastructures d'entreprises et exigences de certification pourraient nécessiter des choix différents de codage tels que Scala, Java ou C ++. Cela peut être un problème pour l'échelle d'un objectif ; au-delà du prototypage et de la mise en œuvre dans la production, des solutions de programmation au niveau de l'entreprise seront toujours requises, ainsi que les compétences nécessaires pour la mise en œuvre. Cela signifie aussi probablement qu'un remaniement de code, ou une traduction entre les langages informatiques, peuvent être nécessaires, ainsi que des interactions fortes entre l'équipe de données et l'informatique et les employés de l'ingénierie.

### Bases de données et stockage des données

Le chapitre 1 traite des données structurées ou non structurées. Un projet de données peut puiser dans les deux, qui sont traitées respectivement par des bases de données relationnelles et des bases de données non relationnelles. L'utilisation de ces outils nécessite différents ensembles de compétences. Les données provenant de bases de données transactionnelles de l'entreprise sont susceptibles de provenir de bases de données relationnelles. De plus en plus, même les données internes, telles que les informations biométriques ou de la KYC, peuvent être stockées par les deux solutions, selon la méthode de collecte. Un algorithme de notation de risque de crédit qui vise toutefois à utiliser les données des réseaux sociaux est susceptible de puiser dans des données non structurées provenant de sources de données non relationnelles.

### Contrôle de version et collaboration

Des outils de résolution de problèmes de versions sont cruciaux pour l'organisation de l'évolution du code, de la maintenance et du travail d'équipe et sont donc essentiels pour une bonne planification du projet.

## Compétences : Sciences des données

### Outils scientifiques

Différents contextes nécessiteront un dosage spécifique en fonction des besoins du projet, mais les éléments suivants font partie des grands domaines universitaires auxquels les projets de données sont susceptibles de devoir faire appel :

- Un bagage solide en statistiques : utilisé pour les tests d'hypothèses et la validation des modèles
- Théorie des réseaux : une discipline qui utilise des nœuds et des liens pour représenter mathématiquement des réseaux complexes ; essentiel pour toute donnée de réseau social ou cartographie de transactions de type P2P
- Apprentissage automatique : une discipline qui utilise des algorithmes pour tirer des enseignements de comportements de données sans règle générale prédéfinie explicite ; la plupart des projets qui offrent un modèle ou un algorithme
- Les sciences humaines, le TLN, la science de la complexité et l'apprentissage en profondeur sont aussi des compétences souhaitables qui pourraient jouer un rôle clé dans des domaines spécifiques d'intérêt

### Curiosité et esprit scientifique

L'attitude et les compétences comportementales sont des facteurs essentiels à la réussite d'une équipe de science des données. Les personnes qui cherchent à explorer, fouiller, agréger, intégrer - et donc identifier des modèles et les connexions - obtiendront de meilleurs résultats. En d'autres termes, certaines « compétences de piratage » représentent une valeur ajoutée pour l'équipe de science des données ; autrement dit, l'équipe doit posséder une approche mentale de résolution de problèmes et une motivation interne pour trouver des modèles grâce à une analyse méthodique.

De plus, la validation scientifique est essentielle pour un projet de données, et les scientifiques des données doivent avoir un esprit scientifique. Autrement dit, une approche méthodique pour poser et répondre à des questions et une volonté de tester et de valider les résultats. Chose importante, les membres de l'équipe doivent puiser leur motivation dans les résultats et être ouverts à toute interprétation qu'offre une solide analyse des données, même si les résultats peuvent contredire les attentes initiales. Conformément à la méthode scientifique, cette approche doit se concrétiser sous forme de compétences comportementales, par exemple faire des observations, trouver des questions intéressantes, formuler des hypothèses et développer des prédictions testables.

### Conception et visualisation

Cela nécessite un ensemble de compétences multidisciplinaires en termes de besoins techniques et commerciaux. Sur le plan technique, la visualisation de données ne doit pas exclusivement être considérée comme la dernière partie du projet visant à embellir les résultats. Elle est utile pendante toute l'exploration et tout le prototypage, et est bien intégrée à certains stades périodiques du projet, ce qui en fait un ensemble de compétences de base permettant aux scientifiques des données d'identifier des modèles.

## Compétences : Activité

La définition d'objectifs est essentiellement liée à la fourniture de résultats commercialement pertinents et à la comparaison par rapport aux paramètres et ICP appropriés. Savoir comment faire le lien entre ces indicateurs et l'exécution du projet est l'objet même de la réalisation du projet. Cela nécessite que l'équipe du projet ait une solide connaissance des affaires. Un point de vue commercial clair est essentiel pour l'interprétation des résultats et, au bout du compte, pour utiliser et mettre en œuvre le projet pour créer de la valeur. En matière de compétences, le message clé est qu'un « agent du carrefour » doit jouer les rôles d'intermédiaire entre les données, les spécialistes techniques, la gestion des affaires et la stratégie afin de traduire les indications tirées des données pour les non techniciens ; le rôle de cet intermédiaire est aussi de reformuler les besoins des entreprises sous forme d'algorithmes et de solutions techniques pour les

communiquer à l'équipe. Une expertise appelée opérations de données, qui incarne l'essence de ce rôle, se développe de plus en plus.

### Respect de la vie privée et aspects juridiques

A l'exception des cas où des ensembles de données sont publiés sous licence libre - permettant explicitement l'utilisation, le remaniement et la modification - par exemple en utilisant des initiatives de données ouvertes, les questions liées à la vie privée, à la propriété des données et aux droits d'utilisation à des fins spécifiques ne sont pas négligeables (voir les obstacles juridiques auxquels les données sont confrontées - dans Accessibilité des données à la page 108). Des juristes d'entreprise doivent être consultés pour s'assurer que toutes les préoccupations des parties prenantes sont dûment prises en compte. Cela dit, les problèmes de mégadonnées et de confidentialité représentent un terrain nouveau, et la législation visant à réglementer l'approche des données est encore en cours de développement. De nombreuses sociétés développent aujourd'hui leurs activités fondées sur les données en tirant parti des lacunes juridiques des droits locaux. Cela peut présenter des risques si les lois changent, tout en présentant des opportunités en collaborant pour construire un environnement favorable.

En termes d'ensembles de compétences, les membres de l'équipe du projet doivent tous avoir une certaine conscience juridique de base. Cela permet d'identifier les problèmes potentiels et de mettre en place un dialogue constructif avec les juristes responsables. Des connaissances juridiques sont particulièrement utiles lors de la sécurisation des consultants externes et de la vérification que les accords de non-divulgation (NDA) sont complets, respectent la réglementation, et peuvent être maintenus. Tant d'un point de vue interne qu'externe, les données peuvent également être une source de fraude. Les cas de fraude sont de plus en plus sophistiqués sur le plan technique et axés sur les données. Même si une équipe de science des données cherche des compétences de pirates informatiques pour équilibrer les compétences, il ne faut pas que de vrais pirates soient présents dans ses rangs. Il est essentiel que toute l'équipe soit bien au courant des considérations juridiques et responsables, à la fois juridiquement et moralement, de leur respect.

# *Leçons du secteur : Rendre anonymes des données*

***Confidentialité des données et protection des consommateurs : L'anonymisation des données des utilisateurs est nécessaire et difficile***

*En 2006, America Online (AOL), un prestataire de services Internet, a rendu publiques 20 millions de requêtes de recherche pour étude. Les personnes avaient été rendues anonymes par un nombre aléatoire. Dans un article du* ***New York Times****, les journalistes Michael Barbaro et Tom Zeller décrivent comment le numéro de client 4417749 a été identifié et par la suite interrogé pour leur article. Alors que l'utilisateur 4417749 était anonyme, ses recherches ne l'étaient pas. Il s'agissait d'une internaute passionnée, utilisant des termes de recherche permettant de l'identifier : « doigts engourdis », « hommes célibataires dans la soixantaine », « chien qui urine partout ». Les recherches incluaient des noms de personnes et d'autres informations spécifiques, notamment « paysagistes à Lilburn, Géorgie, États-Unis d'Amérique ». Aucune recherche isolée ne permet d'identifier quelqu'un, mais pour un détective ou un journaliste, il est facile d'identifier les femmes dans la soixantaine avec des chiens mal élevés et des petits jardins agréables à Lilburn, Géorgie. Thelma Arnold a été retrouvée et a affirmé que les recherches étaient les siennes. En termes de relations publiques, ce fut un désastre pour AOL.*

*Une autre violation de données a fait les titres de la presse en 2014 lorsque Vijay Pandurangan, un ingénieur en logiciel, a identifié 173 millions de dossiers de chauffeurs de taxi publiés par la ville de New York pour une initiative de données ouvertes. Les données ont été cryptées en utilisant une technique qui rend mathématiquement impossible l'opération d'une ingénierie inverse sur la valeur chiffrée. L'ensemble de données ne comportait aucune information de recherche comme Arnold, mais les numéros d'immatriculation des taxis cryptés avaient une structure publiquement connue : numéro, lettre, numéro, numéro (par exemple, 5H32). Pandurangan a calculé qu'il n'y avait que 23 millions de combinaisons, donc il a simplement soumis toutes les entrées possibles à l'algorithme de chiffrement jusqu'à ce qu'il produise les résultats correspondants. Compte tenu de la puissance de calcul actuelle, il a pu identifier des millions de chauffeurs de taxi en seulement deux heures.*

*Netflix, une société de films et de médias en ligne, a parrainé et eu recours à la communauté des internautes pour financer un concours mettant au défi les scientifiques des données d'améliorer de 10 pour cent son algorithme interne de prévision des notations des films par les clients. L'une des équipes a pu identifier les habitudes de visionnage de films des utilisateurs cryptés pour le concours. En recoupant les données avec l'Internet Movie Database (IMDB), qui fournit une plateforme de réseaux sociaux pour que les utilisateurs puissent noter les films et écrire leurs propres critiques, les utilisateurs ont été identifiés par les modèles de séries de films notés de façon identique dans les ensembles de données publics d'IMDB et cryptés de Netflix. Netflix a conclu des arrangements à l'amiable pour des procès intentés par les utilisateurs identifiés et a fait l'objet d'enquêtes sur la vie privée des consommateurs lancées par le gouvernement des États-Unis.*

*Rendre anonymes de façon correcte des données est très difficile, car il existe de nombreuses façons de reconstituer les informations. Dans ces exemples, utiliser des références croisées de ressources publiques (Netflix), la force brute et des ordinateurs puissants (taxis de New York), et les techniques de détective à l'ancienne (AOL) ont conduit à des violations de la vie privée. Si des données sont publiées pour des projets de données ouvertes, de recherche ou autres, un grand soin est nécessaire pour éviter les risques d'identification et leurs graves conséquences juridiques et en termes de relations publiques.*

### Sciences humaines et données

Le croisement des compétences en données et des sciences humaines est un nouveau domaine d'activité de recherche et un ensemble de compétences clé pour les équipes de projet. La motivation des entreprises pour un projet de données se résume généralement aux clients, qu'il se rapporte à une augmentation de l'activité, à de nouveaux produits ou à de nouvelles caractéristiques démographiques. Pour interagir avec les clients, il faut savoir quelque chose sur eux. Les compétences en sciences humaines des données permettent d'interpréter les résultats à travers un regard qui cherche à comprendre ce que les utilisateurs font ou ne font pas et pourquoi ; ainsi, les équipes sont en mesure de mieux identifier les modèles de données utiles et d'affiner des modèles autour de variables qui représentent les normes sociales et les activités des clients.

### Expertise sectorielle

L'expérience dans le domaine, la connaissance du marché et l'expertise sectorielle sont tous des termes qui décrivent la relation essentielle entre les résultats du projet et la valeur commerciale. En l'absence d'expertise sectorielle, de mauvaises données peuvent être analysées, des modèles très précis peuvent tester des hypothèses erronées, ou des variables statistiquement significatives qui n'ont aucun lien avec des ICP commerciaux pourraient être choisies. Alors que de nombreux modèles d'apprentissage automatique produisent des « boites noires » ou des cadres d'infrastructure qui utilisent des approches automatisées, il existe des risques importants qu'un projet de données puisse fournir des résultats qui ont l'air fantastiques mais qui, à l'insu de ses concepteurs, sont le produit d'une mauvaise veille économique. Par conséquent, le dialogue permanent avec les experts du secteur doit faire partie intégrante de la conception du projet.

### Communications

Les données racontent une histoire. En réalité, des chiffres précis peuvent raconter quelques-unes des histoires les plus intéressantes d'une manière concise. Les liens entre les communications d'entreprise et les équipes de projet sont un élément important pour l'utilisation des résultats du projet – tout comme le fait d'être en mesure de les mettre en œuvre de la bonne façon, en harmonie avec la stratégie de communication. Il existe aussi une forte relation de communication avec la visualisation des données et la conception, en particulier pour les projets en relation directe avec le public. La visualisation des données est importante pour la communication des résultats intermédiaires et finaux. S'assurer de disposer des compétences de conception visuelle est aussi important que les compétences techniques pour tracer les graphiques, rendre les résultats interactifs ou les offrir au public sur des sites Web. Pour de nombreux projets de données, la visualisation est un livrable fondamental, comme pour les tableaux de bord et de nombreux objectifs du projet visant spécifiquement à orienter la communication de l'entreprise.

## Quadrant 3 : PROCESSUS

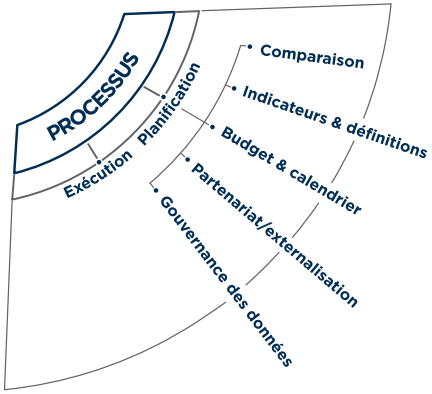



***Figure 22 :*** *Quadrant 3 de l'Anneau des données : PROCESSUS*

Dans les sections précédentes, nous nous sommes intéressés à la moitié supérieure de l'Anneau des données, en nous concentrant sur les exigences pratiques (infrastructure, données et outils) et les exigences humaines (savoir-faire et compétences). Dans cette section, nous passons à la moitié inférieure de l'Anneau des données, qui porte sur le processus de conception et de réalisation d'un projet de données.

Reconnaissant que les entreprises ou les institutions disposent d'approches qui leur sont propres, basées sur une combinaison d'histoire de l'organisation, de culture d'entreprise, de normes en matière d'ICP et de règles de gestion des données, les pratiques suivantes sont considérées comme des bonnes pratiques générales

pour permettre la réalisation de projets fondés sur des données et leurs livrables.

Les projets de données doivent définir leurs livrables, résultats de la Planification et de l'Exécution du projet. Ces résultats sont des intermédiaires entre le Processus et le bloc suivant qui vise à les convertir en valeur commerciale. La liste suivante précise huit éléments communs à de nombreux projets de données. Le cas échéant, ces éléments doivent figurer dans le calendrier de livraison d'un projet, ou être spécifiés dans les termes de référence relatifs à la capacité externalisée.

### Ensemble(s) de données

Les ensembles de données sont toutes les données recueillies ou analysées. En fonction de la taille, de la méthode de collecte et de la nature des données, le format de l'ensemble ou des ensembles de données peut varier. Ceux-ci doivent tous être documentés, en fournissant des informations sur leur emplacement - par exemple sur un réseau ou dans un nuage - et comment y accéder. Les saisies brutes doivent être « nettoyées », un processus abordé dans la section sur l'exécution ci-dessous. Les ensembles de données nettoyés doivent être considérés comme des livrables spécifiques, tout comme les méthodes scriptées ou les étapes méthodologiques utilisées pour nettoyer les données. Enfin, les ensembles de données et les méthodes agrégés pourraient également être considérés comme des livrables spécifiques. Ceux-ci sont nécessaires pour aider les promoteurs de projets à voir ce qui a été fait aux données et éventuellement à détecter les erreurs. En outre, ceux-ci viennent appuyer les projets de suivi ou les analyses dérivées qui se fondent sur des données pré-agrégées nettoyées.

### Questionnaires et outils de collecte

Les projets qui nécessitent une collecte de données primaires, à la fois quantitatives et qualitatives, peuvent exiger l'utilisation ou le développement d'outils de collecte de données, tels que des instruments de sondage, des questionnaires, des données d'identification de localisation, des rapports photographiques, ou encore des discussions de groupes ou des entretiens. Ces instruments doivent être livrés, parallèlement aux données recueillies, ainsi que les textes dans toutes les langues utilisées, et leurs traductions et transcriptions. Ces informations sont nécessaires pour permettre des enquêtes de suivi ou des questions sur la cohérence chronologique, et fournissent également les documents d'audit ou de vérification nécessaires si des questions sont soulevées quant aux méthodes de collecte de données à un stade ultérieur.

### Rapport d'inventaire de données

Il s'agit d'un rapport incluant une synthèse des données utilisées pour l'analyse. Ce rapport inclut le type, la taille et la date des fichiers. Il doit inclure des discussions sur les anomalies ou lacunes majeures observées dans les données, et évaluer si les anomalies sont susceptibles d'être statistiquement biaisées pour présenter des risques d'interprétation. Il peut inclure des graphiques qui représentant les principaux points de données pour les principaux segments, comme les transactions dans le temps, désagrégées par type de produit, afin de montrer les tendances, les pics, les creux et les lacunes. Livré tôt dans le processus d'exécution, le rapport d'inventaire des données est l'occasion de discuter des risques potentiels du projet liés aux données sous-jacentes, ainsi que des stratégies de rectification du cap et de la nécessité d'affiner les données ou d'en acquérir de nouvelles. Il s'avère particulièrement utile pour indiquer les exigences de nettoyage des données et de s'efforcer de régler les anomalies de manière statistiquement non biaisée.

### Dictionnaire de données

Le dictionnaire de données consolide les informations provenant de toutes les sources de données. Il s'agit d'un recueil de la description de tous les éléments de données, comme les tableaux. Cette description inclut généralement le nom du champ de données, son type, son format, sa taille, la définition du champ et, si possible, un exemple de ces données. Les champs de données qui constituent un ensemble doivent lister toutes les valeurs possibles. Par exemple, si un ensemble de données de transaction comporte une colonne appelée « produit » qui indique si une transaction était un rechargement, une transaction pair à pair ou un retrait en espèces, alors le dictionnaire énumérera toutes les valeurs du produit et décrira leurs codes respectifs observés dans les données, telles que TUP, P2P et COT respectivement. Pour les données qui ne sont pas dans un ensemble discret, comme de l'argent, alors une fourchette de valeurs min-max est généralement indiquée, ainsi que son unité d'indicateur, comme le type de devise. Les relations avec d'autres ensembles de

données doivent également être spécifiées, le cas échéant. Par exemple, le champ de données du numéro de compte d'un client peut être présent dans les ensembles de données de transactions de produits, ainsi que dans les ensembles de données de la KYC. La spécification de ce lien permet de comprendre comment les données peuvent être fusionnées, ou d'identifier dans quels domaines des exigences de métadonnées supplémentaires peuvent être nécessaires pour faciliter une telle fusion. Le dictionnaire de données est généralement fourni parallèlement au rapport d'inventaire de données, en appui à une discussion sur la conception stratégique d'un projet, l'évaluation des risques ou les exigences de données supplémentaires dans les premiers temps du projet.

### Analyses exploratoires et journal de bord

Il s'agit d'un ensemble de courbes, de graphiques ou de données sous forme de tableaux récapitulant les principales caractéristiques d'une étude spécifique ou d'un test d'hypothèse. Toutes les statistiques descriptives des données peuvent également être incluses, par exemple les moyennes, médianes ou écarts types. La partie analyse exploratoire de l'identification des tendances et des modèles découverts dans les données est nécessaire pour affiner les hypothèses analytiques, contextualiser les métadonnées ou identifier les « caractéristiques » utilisées dans un modèle. L'analyse exploratoire est effectuée dans le cadre de l'exécution initiale du projet, et elle se poursuit souvent jusqu'à l'achèvement du projet. Les résultats exploratoires viennent généralement appuyer les livrables intermédiaires ou les évaluations des étapes du projet. Ces résultats peuvent également être synthétisés pour faciliter la formulation de l'état et de la progression du projet en mettant en évidence les questions actuellement à l'étude ainsi que les questions qui ont déjà été traitées. Un journal de bord des initiatives à l'étude et des principales constatations se révèle utile à cet égard.

### Graphiques de validation du modèle et indicateurs de performance

Pour les projets de données fondés sur des modèles, il s'agit d'une liste de graphiques présentant les indicateurs de performance les plus pertinentes du modèle prédictif. Voir le chapitre 2.2.4 : Indicateurs des modèles de données pour une liste des 10 meilleurs indicateurs et définitions de la performance. Ces graphiques et indicateurs seront utilisés pour évaluer l'efficacité et la fiabilité du modèle. Les tableaux de validation peuvent inclure les graphiques de gains et de lift, et les indicateurs de performance dépendront du projet particulier. Ces indicateurs peuvent par exemple inclure le test Kolmogorov-Smirnov (KS), la courbe de fonction d'efficacité du récepteur (ROC, Receiver Operating Characteristic) ou le coefficient de Gini. Ces informations sont nécessaires pour évaluer les étapes de réalisation des objectifs. L'approbation du modèle à des fins de production ou de répétition à l'étape suivante doit devrait se fonder sur ces indicateurs.

### Livrables analytiques : Résultats, algorithmes, listes blanches et visualisations

Il s'agit des véritables résultats du projet. Un projet de segmentation de clientèle peut inclure une liste blanche des clients à cibler et les scores de propension associés, ainsi que des informations de géolocalisation possibles pour informer une campagne de marketing. Un algorithme de notation de risque de crédit fournit des ensembles de résultats pour les utilisateurs spécifiés dans les ensembles de données de contrôle et de traitement et le code du modèle lui-même, ou une visualisation incluant des scripts pour tracer les KPI et les animer, et les scripts Web ou autres éléments pour une interface utilisateur. Chaque projet disposera de son propre ensemble de livrables nuancés. Ceux-ci doivent être définis dans le cadre de la conception du processus du projet.

### Rapport d'analyse final et discussion sur le coût-bénéfice de la mise en œuvre

Il s'agit du rapport final présentant les résultats des analyses, qui répond aux questions et se réfère aux objectifs fixés et convenus au début du projet. Celui-ci doit être fourni conjointement aux livrables analytiques. En plus de discuter de la méthodologie, du processus, des conclusions et des solutions aux défis clés, le rapport final doit formuler la proposition de valeur fondamentale des livrables analytiques. Cela peut inclure : les gains d'efficacité et les économies de coûts découlant d'un meilleur marketing fondé sur les données ; les prévisions d'augmentation des opportunités de prêt ;

ou les gains de productivité découlant des tableaux de bord. Le rapport final doit être examiné à la lumière de la stratégie de mise en œuvre du projet, afin de réfléchir au coût-bénéfice de la proposition de valeur dans les livrables analytiques et aux besoins en ressources pour les mettre en œuvre à l'échelle attendue dans le cadre du projet.

### Processus : Planification

Les considérations suivantes sont particulièrement pertinentes en matière de planification de projets de données et pour aider à définir le champ des livrables intermédiaires et finaux.

#### Points de comparaison

Au cours de la planification de la phase d'exécution, il est essentiel de comprendre qui d'autre a rencontré un problème similaire et comment il a pu être abordé et résolu. La littérature scientifique est une véritable mine d'informations et les limites entre la recherche et l'application opérationnelle se chevauchent souvent dans le domaine des données. Du point de vue de la gestion de projet, l'évaluation comparative consiste à analyser les entreprises concurrentes et leurs activités dans le domaine des données, en veillant à ce que le projet soit aligné sur les pratiques et les opérations internes de l'entreprise. Autrement dit, ne réinventez pas la roue.

#### Indicateurs et ICP

Les indicateurs sont les paramètres qui actionnent l'exécution du projet et déterminent si celui-ci est réussi. Par exemple : rejeter une hypothèse nulle à un niveau de confiance de 90 pour cent ; atteindre un taux d'exactitude du modèle de 85 pour cent ; ou un temps de réponse sur une décision de notation de risque de crédit inférieure à deux secondes. La définition préalable des indicateurs évite les risques liés à la post-validation lorsque, en raison de seuils vagues, les chefs de projet fournissent des résultats « satisfaisants ». Cela vise souvent à tenter de justifier l'investissement, ou pire encore, affirmé des résultats à l'encontre des convictions, en insistant sur le fait qu'ils devraient fonctionner. Voir le chapitre 2.2.3: Indicateurs pour l'évaluation des modèles de données, qui fournit une liste des 10 meilleurs indicateurs utilisés dans les projets de modélisation de données. Les indicateurs liés à l'expérience utilisateur sont également importants, mais doivent être propres au contexte du projet. Par exemple, lorsque vous évaluez le temps d'attente acceptable avant qu'un utilisateur obtienne une décision automatisée de notation de risque de crédit, le plus vite est le mieux. Cependant, un ICP doit être préalablement défini pour permettre à l'équipe du projet de livrer un produit bien adapté.

#### Budget et calendrier

La planification et le contrôle de gestion doivent tenir compte de l'état d'ouverture quasi permanent des projets de données. Les objectifs et les cibles montrent un point final, mais jusqu'à ce qu'il soit atteint, un projet de données consiste souvent en des modifications constantes basées sur l'amélioration de la compréhension et de la définition des problèmes. Certains peuvent croire que s'ils le réajustent différemment, la fois suivante ils pourront atteindre 85 pour cent. D'autres pensent pouvoir ajouter de nouvelles données clients pour améliorer le modèle. Cette situation fluide ne contribue pas à l'estimation des budgets, mais les chefs de projet doivent utiliser les paramètres du budget comme instruments pour adapter leur travail, leur engagement et leur espace en vue de tester différentes hypothèses. Les investissements initiaux doivent comprendre ce processus exploratoire et itératif et les risques qui y sont associés. Le concept d'échelle du produit contribue également à atténuer ce risque ; commencez petit, et développez en répétant. Cela risque de provoquer des inefficiences en termes d'échelle et de retravailler le code, mais permet également d'atténuer les risques budgétaires, tels que l'achat de nouveaux ordinateurs pour ensuite constater que l'hypothèse ne tient pas.

La planification du calendrier est associée à des considérations similaires à celles de la planification budgétaire. Encore une fois, le compromis consiste à consacrer suffisamment d'espace à l'exploration et à la recherche en restant centrés sur les objectifs et les indicateurs. Une technique de gestion de projet tirée de l'industrie du logiciel, appelée « méthode agile », est utile dans les projets de données. Cette approche se penche sur la progression du projet par le biais de cycles endogènes dans lesquels la production est une chose mesurable et vérifiable. Cela contribue à intégrer une exploration dans un cycle spécifique.

### Partenariats, externalisation et appel à la contribution collective

Cette question est particulièrement importante du point de vue des ressources du projet. Poser des questions sur la conception de projet quant aux exigences et à leur suffisance contribue à identifier les lacunes que les chefs de projet pourront combler. Surtout, cela ne se limite pas aux ressources humaines. L'informatique en nuage est un outil informatique externalisé. Même les données peuvent venir de l'extérieur, que ce soit en octroyant la licence aux fournisseurs ou en établissant des partenariats qui permettent d'y accéder. L'appel à contribution collective est une technique émergente visant à solliciter des équipes de données complètes en leur donnant des limites exploratoires très larges, généralement dans le but de fournir une créativité pure et des solutions innovantes à un problème fixe, pour une prime fixe. On citera pour exemple Kaggle, qui est un pionnier de premier plan en matière d'expertise en science des données en crowd-sourcing ; ou le service « Mechanical Turk » d'Amazon pour les petites tâches ou enquêtes en crowd-sourcing.

Un élément important à considérer est la propriété intellectuelle. Les droits doivent être spécifiés dans les accords contractuels. Ceci inclut la propriété intellectuelle existante ainsi que la propriété intellectuelle créée dans le cadre du projet. Prenez en compte l'intégralité de la phase de processus et d'exécution le long du pipeline de données. La propriété intellectuelle englobe plus que les résultats livrables finaux ; elle comprend les scripts et les codes informatiques écrits pour procéder à l'analyse, et même les ensembles de données intermédiaires, les agrégats et les segmentations qui viennent alimenter d'autres processus.

### Gestion des données

Il s'agit de la façon dont les données sont utilisées, à quel moment et qui y a accès. La planification de la gestion des données doit tenir compte de la politique générale de l'entreprise, des exigences juridiques et des politiques de communication. L'objet du plan est de permettre l'accès aux données à l'équipe du projet et à ceux qui interviennent dans la livraison, tout en préservant l'équilibre en termes d'exigences de confidentialité des données et de sécurité. Le plan de gestion des données est généralement affecté par l'échelle du projet, les projets plus importants étant susceptibles de comporter davantage de risques que des projets plus petits. Un défi majeur réside dans le fait que l'approche fondée sur la science des données bénéficie d'un accès à autant de données que possible afin de relier les ensembles de données et d'explorer les modèles qui en découlent. Mais en même temps, plus de données et plus d'accès présentent également davantage de risques. La gestion des données de projet doit également spécifier le plan d'ETL. Cela englobe également le transport ou la planification des mouvements physiques ou numériques, qui doivent tenir compte du transit dans des environnements politiques ou réglementaires, par exemple d'une entreprise en Afrique à un fournisseur d'analyse externalisé en Europe. Le plan doit tenir compte des principes suivants :

- **Cryptage :** Les informations sensibles ou d'identification doivent être cryptées, brouillées ou rendues anonymes et rester dans le pipeline de données complet.
- **Autorisations :** L'accès aux ensembles de données doit être défini de manière très précise selon les rôles au sein de l'équipe, ou par point d'accès (c.-à-d. à partir des pare-feux d'entreprise, par opposition aux réseaux externes).
- **Sécurité :** Les ensembles de données placés dans l'environnement d' « expérimentation » du projet doivent disposer de leur propre système de sécurité ou pare-feu, ainsi que d'une capacité à authentifier les accès privilégiés.
- **Connexion :** L'accès et l'utilisation doivent faire l'objet d'un historique et pouvoir être vérifiés, et activés pour permettre l'analyse et l'établissement de rapports.
- **Réglementation :** Le plan doit s'assurer que les exigences réglementaires sont respectées, et des accords de confidentialité ou des contrats juridiques doivent être en place pour couvrir toutes les parties prenantes du projet. Les droits et la confidentialité des clients doivent également être protégés.

### Processus : Exécution

Tout comme l'Anneau des données représente un processus cyclique, la phase d'exécution de nombreux projets de données tend à constituer une sorte de boucle dans la boucle. Ce qu'on appelle généralement l' « analyse de données »

constitue en réalité davantage un ensemble d'étapes progressives et itératives. C'est un parcours d'exploration et de validation d'hypothèses jusqu'à ce qu'un résultat réponde aux indicateurs cibles définis.

La phase d'exécution ressemble beaucoup aux cadres établis pour l'analyse des données, tels que le CRISP-DM ou autres adaptations.[39] Les gestionnaires de projet qui préfèrent utiliser un cadre de processus analytique spécifique, ou dont les projets peuvent être mieux desservis par une approche donnée, peuvent facilement intégrer ces cadres dans la spécification de conception de projet d'Anneau des données ici, dans la phase d'exécution. Les étapes suivantes sont également fournies comme un processus général d'exécution d'analyse des données à utiliser à titre de bonne pratique.

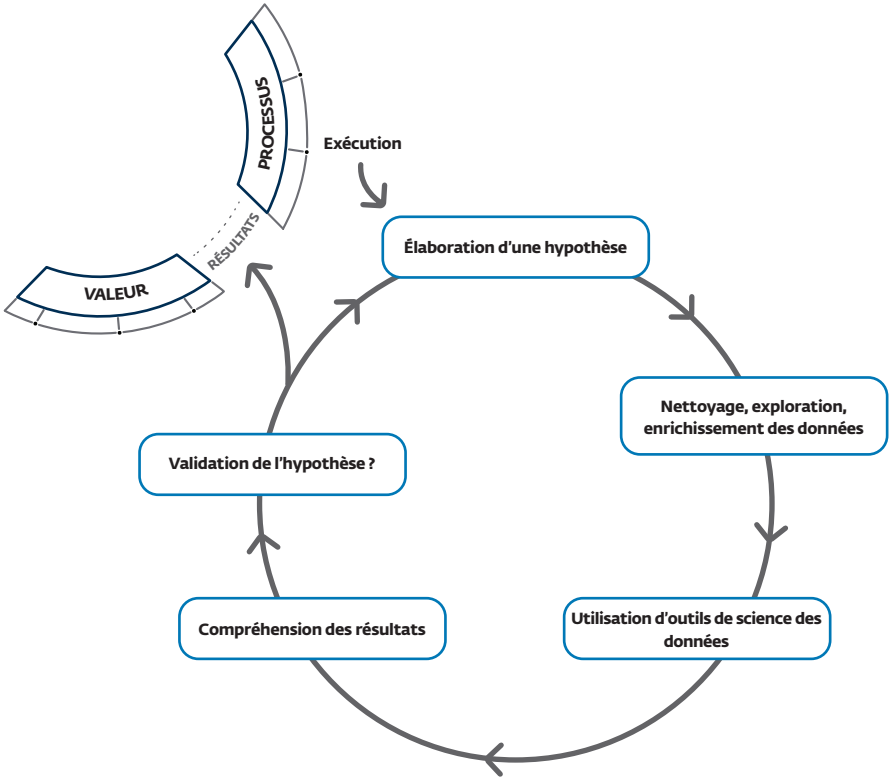



*Figure 23 : Processus d'exécution de l'Anneau des données*

### Nettoyer, explorer et enrichir les données

C'est à cette étape que l'équipe chargée de la science des données commence vraiment à travailler. La probabilité qu'un ensemble de données réponde parfaitement aux besoins de l'étude est faible. Les données devront être nettoyées, ce qui consiste à :

a. **Traiter :** Convertir les données à un format commun, compatible avec les outils de traitement.

b. **Comprendre** : Vérifier les métadonnées et la documentation disponibles pour savoir ce que sont les données.

c. **Valider :** Identifier les erreurs, les champs vides et les mesures anormales.

d. **Fusionner :** Intégrer les descriptions numériques (lisibles par machine) des événements effectués manuellement par des personnes pendant le processus de collecte de données afin de fournir une explication claire de tous les événements.

e. **Combiner :** Enrichir les données par d'autres données, qu'elles proviennent de la même société, du domaine public ou d'ailleurs.

f. **Procéder à une analyse exploratoire :** Utiliser des techniques de visualisation de données pour explorer partiellement les données et les modèles.

g. **Itérer :** Itérer jusqu'à ce que les erreurs soient comptabilisées et qu'un processus soit en place pour passer *efficacement* des données brutes à des données opérationnelles. C'est le processus minimum pour assurer la viabilité.

[39] Les méthodes de processus d'analyse de données connexes comprennent, par exemple : « Knowledge Discovery in Databases Process » (KDD Process) d'Usama Fayyad ; « Sample, Explore, Modify, Model, Assess » (SEMMA) du SAS Institute; « Analytics Solutions Unified Method for Data Mining/Predictive Analytics » (ASUM-DM) d'IBM; « Data Science Team Process » (DSTP) de Microsoft

### Utiliser les outils de la science des données

C'est là que les spécialistes des données appliquent leur expertise. L'apprentissage automatique, l'exploration de données, l'apprentissage profond, le TLN, la science des réseaux, les statistiques ou (habituellement) une combinaison de ce qui précède sont appliqués. Lors de l'élaboration de projets de données incluant des modèles prédictifs, il est nécessaire de mettre en place une stratégie de validation de modèle avant l'exécution du modèle. Cela permet de tester statistiquement les hypothèses du projet. Dans la pratique, l'ensemble de données qui pilote le modèle doit être segmenté en un ensemble « témoin » et un ensemble « traitement » par le biais d'une sélection aléatoire. Une segmentation de 20 pour cent à 80 pour cent constitue une approche courante et basique. Le modèle est testé sur l'ensemble « traitement ». Ensuite, le modèle peut fonctionner sur l'ensemble témoin, et les valeurs prédites du modèle peuvent être comparées aux valeurs connues de l'ensemble témoin. C'est ainsi que les taux de précision sont calculés et qu'une hypothèse peut être testée.

### Compréhension, interprétation et représentation des résultats

L'interprétation des résultats fera l'objet d'une discussion plus approfondie dans la section suivante en termes d'apport de Valeur commerciale. Mais du point de vue du processus, la compréhension des résultats se concentre sur la concordance entre les résultats obtenus et le produit attendu de l'exécution du processus ; l'objectif est également de s'assurer qu'ils sont valides sur le plan informatique (c.-à-d. incluant un contrôle des erreurs arithmétiques ou erreurs de codage). La résultante de tout calcul ou processus analytique, grand ou petit, produira :

- des résultats inutilisables (ou incorrects)
- des résultats insignifiants ou déjà connus
- des résultats utilisables qui viennent alimenter les étapes suivantes
- des résultats inattendus (à étudier avec un nouveau pipeline, de nouvelles données ou une nouvelle approche)

La conception du projet doit reconnaitre ces résultats possibles et être prête à traiter chaque cas. Hormis les résultats inutilisables, toutes les autres catégories de résultats mériteront probablement un travail de présentation ou de rapport afin de les rendre compréhensible à d'autres, et notamment aux membres de l'équipe interne, aux gestionnaires, aux clients et au public en général. Cela signifie habituellement un résumé écrit, un tableau, un graphique ou une animation, ces supports permettant de présenter et d'expliquer les résultats. Les experts en visualisation de données jouent un rôle essentiel dans ce processus, car il n'est pas simplement question d'enjoliver les résultats. La tâche difficile consiste à créer des couches convaincantes, interactives et visuelles pour ajouter de manière succincte des éléments au récit plus général du projet qui doit constituer un énoncé du problème du projet en lui-même.

La phase d'exécution est également l'occasion de réévaluer les plans du projet, en remarquant à nouveau qu'il vaut mieux que les projets de données soient réalisés en utilisant une approche itérative. La phase d'exécution d'un projet est ce qui va permettre de tester le processus de conception et l'approche du projet, en insistant pour une révision lorsque survient un imprévu. Le cadre de l'Anneau des données peut également contribuer à résoudre les problèmes d'exécution pour identifier des solutions ; ses concepts ne se limitent pas à la planification initiale. La trame de l'Anneau des données (discutée au point 2.1 : Application) est conçue dans cette intention, pour fournir un modèle pouvant être constamment mis à jour pour refléter le statut du projet au cours de son exécution.

### Évaluation des indicateurs et étapes suivantes

Ce n'est que par une définition initiale quantitative et précise des objectifs et des indicateurs du projet que l'efficacité du projet peut être jugée. Si les résultats ne sont pas satisfaisants, le processus doit recommencer. Cette étape d'évaluation et d'itération est toujours critique, mais présente des considérations supplémentaires en cas de recours à des entreprises extérieures. Les livrables peuvent être jugés inadéquats malgré la qualité du travail. La responsabilisation des résultats livrés doit être convenue à l'avance, de même que le degré de marge de manœuvre pour continuer à itérer pour obtenir des résultats satisfaisants. Exactement comme le rôle qu'ils jouent dans la première étape, la définition des hypothèses, de cette boucle d'exécution, les gestionnaires de projets de données jouent à nouveau un rôle clé pour s'assurer que les scientifiques restent concentrés sur les principaux objectifs et renforcent les itérations futures.

## Quadrant 4 : VALEUR

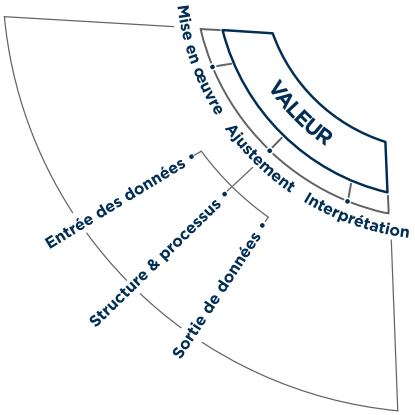



*Figure 24 : Quadrant 4 de l'Anneau des données : VALEUR*

La valeur est la dernière partie de l'Anneau des données ou, sur le plan de la conception, le point de départ des itérations futures afin d'ajouter ou de mettre en œuvre des éléments ou développer la conception. Cette étape explique comment les résultats de l'exécution du processus sont en définitive transformés en « informations », puis en « connaissances et valeurs » qui peuvent être mises en œuvre.

Cette composante de création de valeur des résultats est généralement l'une des différences substantielles entre un projet d'analyse de données traditionnel ou un projet de veille technologique et un processus analytique avancé, en particulier dans la sphère des métadonnées. En effet, les livrables du projet sont rarement définis en termes de rapports écrits, du moins pas exclusivement. Les livrables des projets de données se caractérisent généralement par des tableaux de bord, des modèles prédictifs ou des leviers de prise de décision fondés sur les données, des outils d'automatisation et, dans l'idéal, des idées commerciales puissantes. En d'autres termes, un projet de données se termine rarement par des recommandations. Il tend plutôt à délivrer des modules à mettre en œuvre.

### Valeur : Interprétation

La première étape suivant l'étape d'exécution se concentre sur la compréhension de la proposition de valeur inhérente aux résultats et ce qui peut être nécessaire pour affiner ces résultats ou leurs processus sous-jacents afin de réaliser l'objectif. Un nombre peut tout dire ou ne rien dire, selon l'interprétation que l'on en fait. Comprendre les résultats ne consiste pas en une explication simple des phénomènes. Il s'agit plutôt de placer les résultats dans le contexte commercial et d'embrasser la complexité des opérations réelles. Cela nécessite également une approche transparente et collaborative, pour discuter des résultats avec tous les acteurs du projet, afin de déterminer ce qu'ils veulent dire sous tous les angles. En gardant à l'esprit le rôle des opérations de données (voir Compétences commerciales), il n'est pas rare que les spécialistes des données aient du mal à expliquer la pertinence opérationnelle des résultats aux gestionnaires. Si une constatation importante est faite, sa valeur doit être communiquée avec succès à la direction, qui peut prendre sur une mesure à son sujet.

### Valeur : Ajustement

Comprendre les résultats n'est que la tâche initiale. Les connaissances dérivées des données doivent être transformées en actions concrètes qui se manifestent par des outils, des modèles et des algorithmes. En raison de l'approche itérative et exploratoire d'un projet de données, la première fois qu'un résultat final est atteint, il sera invariablement d'allure grossière et devra faire l'objet d'un ajustement pour en faire un outil opérationnel optimisé. L'ajustement s'axe sur trois domaines :

#### Entrée de données

Le choix et la qualité des données d'entrée peuvent déterminer de manière décisive l'efficacité des algorithmes utilisés pour effectuer l'analyse. Envisagez l'apprentissage automatique, où les algorithmes développent une attitude d'apprentissage après une phase de formation utilisant un sous-ensemble de données. Par conséquent, en travaillant avec les données, les opérations apprennent progressivement à recueillir de meilleures données. L'amélioration des données brutes et la réduction des anomalies, des méthodes de collecte, des saisies manuelles et des erreurs de collecte entraineront des résultats ajustés avec plus de précision au fil du temps.

#### Infrastructure, compétences et processus

Après les premières itérations d'exécution, on disposera d'une meilleure compréhension de l'efficacité de l'équipe allouée au projet, des processus de gestion

des données, ainsi que des outils logiciels et matériels disponibles. En outre, on disposera d'une meilleure compréhension de la manière dont l'ensemble de l'organisation du projet fonctionne. Les inefficacités seront révélées et, comme nous l'avons déjà mentionné, tous les domaines du projet peuvent servir de sources de solutions potentielles. De manière générale, l'ajustement a pour objet que toutes les composantes fonctionnent de mieux en mieux ensemble. Cela se fait par le biais d'une meilleure organisation de l'équipe, d'une communication plus forte, de compétences accrues de l'équipe, et par la technologie, qu'il s'agisse de meilleures méthodes, d'une puissance informatique accrue, ou une combinaison de tout ce qui précède.

### Sortie de données

Enfin, les données de sortie doivent être examinées. Il est important que les résultats de sortie ne soient pas biaisés ou affectés par des erreurs (humaines ou autres), une mauvaise intégration entre différentes étapes du processus ou même des erreurs de codage fréquentes. Souvent, cela veut dire examiner et corriger les données d'entrée. Il convient toutefois de noter que le processus analytique est parfaitement capable d'introduire ses propres anomalies. Il s'agit à la fois d'un contrôle de validation et d'une opportunité d'ajustement. En fin de compte, l'examen des résultats vient appuyer l'organisation et la fiabilité générales, par exemple en veillant à ce qu'une visualisation finale affiche les bons résultats à 100 pour cent du temps et dans toutes les conditions.

## Valeur : Mise en œuvre

### Stratégie de mise en œuvre

Pour avoir un impact réel, la stratégie de mise en œuvre doit être conçue dès le début, dans le cadre de la définition des objectifs. Ce point doit être présent à l'esprit tout au long du processus. Évitez le risque d'obtenir d'excellentes données qui ne peuvent pas être utilisées dans la pratique. Un aspect essentiel de la stratégie de mise en œuvre est de s'assurer de l'adhésion de la direction. On peut supposer que l'attribution de ressources offre un certain niveau d'engagement. Cela dit, parce que les parties prenantes ont été assurées que les processus exploratoires ne produisent pas de résultats garantis, la stratégie de mise en œuvre doit assurer un soutien continu et une forte communication autour des résultats intermédiaires.

Les types d'analyses, tels qu'ils sont décrits au chapitre 1.1, peuvent également être pertinents pour réfléchir à la manière dont les résultats sont utilisés :

- **Analyse descriptive :** résumer ou agréger des informations
- **Analyse diagnostique :** Identifier des sous-ensembles d'informations basés sur des critères spécifiques
- **Analyse prédictive :** se fonde généralement sur des sous-ensembles prédictifs, combinés avec des leviers décisionnels
- **Analyse prescriptive :** entièrement intégrée dans les systèmes automatisés ; fait partie des opérations

Ces descriptions peuvent orienter la stratégie de mise en œuvre, en formulant ce à quoi ressemble le cas d'utilisation. C'est également un élément important pour assurer l'adhésion de la direction. Par exemple, si le cas d'utilisation envisage une automatisation complète, les questions conceptuelles du projet doivent demander que l'infrastructure et les ressources soient suffisantes pour mettre en œuvre un algorithme entièrement automatisé. Si un investissement dans un nouveau centre de données est nécessaire pour exécuter l'algorithme et fournir des décisions de crédit juste-à-temps, il pourrait être difficile d'obtenir l'adhésion requise pour s'assurer que les résultats du projet sont utilisés, alors qu'une stratégie de cas d'utilisation basée sur un petit projet pilote mis en œuvre avec les ressources existantes pourra constituer un cas plus facile.

### Cout-bénéfice

La proposition de valeur anticipée doit être formulée dans la conception initiale. Au début, cela peut l'être en termes généraux, par exemple un gain d'efficacité, une réduction des couts ou la fidélisation des clients. À mesure que le projet se développe et que des résultats sont obtenus et ajustés, la proposition de valeur peut être quantifiée. Une fois l'objectif atteint, cela contribue à définir ce qui a été effectivement obtenu et la valeur que cela représente. Le même processus doit être adopté quant à l'utilisation des résultats. Au début, certaines exigences générales en matière d'infrastructure ou de système peuvent être envisagées. Une fois que le projet a atteint le degré de maturité requis, la valeur doit être estimée par rapport au cout de la mise en œuvre de la solution.

## APPLICATION : Utiliser l'Anneau des données

### Une approche sous forme de matrice

En tant qu'outil de planification, l'Anneau des données se présente sous forme de matrice. Une « matrice » est un outil utilisé pour poser des questions structurées et définir les réponses de manière organisée, en un seul endroit. Les réponses sont simples et descriptives ; même quelques mots suffiront. Il peut encore falloir des semaines pour développer une matrice solide pour piloter la planification du projet, car l'interaction des questions directrices remet en cause la compréhension approfondie des problèmes, des solutions envisagées et des outils permettant de les livrer. Une liste des quatre principales raisons d'adopter une *approche sous forme de matrice* est fournie ci-dessous :

1. obliger le responsable du projet à énoncer une proposition de valeur de projet limpide
2. fournir un autodiagnostic et définir et respecter une stratégie de gestion interne
3. communiquer une représentation complète du processus « sur une seule page »
4. planifier de manière flexible avec un outil capable de redéfinir les éléments à mesure que le projet évolue

Le concept de matrice a été introduit par Alex Osterwalder, qui a développé la Matrice d'affaires. Au cours des dernières années, il est devenu inhabituel de participer à un concours de startups, un concours de projets d'entreprise, un hackathon ou un brainstorming sur le thème de l'innovation sans croiser la matrice d'affaires. Et on observe des individus apposant des post-it colorés sur des affiches de matrice, concentrés sur la difficile tâche de donner une vision schématique concise et complète de leur modèle économique. L'application répandue du cadre parmi les innovateurs et les startups technologiques fournit une base solide pour répondre aux besoins en gestion de projet qu'ont des projets de données novateurs et fondés sur la technologie. Il existe quantité de ressources excellentes proposant des informations supplémentaires sur la matrice d'affaires, mais ce n'est pas une condition préalable pour comprendre ou appliquer l'Anneau des données.

La matrice de l'Anneau des données s'inspire de cette approche, et est appliquée aux exigences spécifiques de la gestion de projet de données, tout en soulignant la nécessité de définir des objectifs clairs et d'appliquer les bons outils et compétences pour permettre la mise en œuvre réussie du projet. Ici, un aperçu étape par étape vient affiner les cinq structures de l'Anneau des données quant à leurs de leurs interrelations. Chacun des blocs centraux de l'anneau représente un élément d'un système dynamique et interconnecté. L'approche itérative et l'application en matrice permettent de les agencer dans un schéma unique pour visualiser les éléments du plan holistique, d'identifier les besoins et les lacunes en matière de ressources et de développer un système harmonieux.

Pour ce faire, une planification itérative est adoptée, dans laquelle un objectif doit d'abord être défini. Une fois que l'objectif est défini, l'approche procède étape par étape en faisant le tour de l'anneau pour formuler les ressources, les relations et les processus nécessaires pour atteindre le but. À cette fin, quatre questions sur la conception du projet sont posées séquentiellement pour chacun des blocs centraux. Les questions relatives à la conception de projet sont les suivantes :

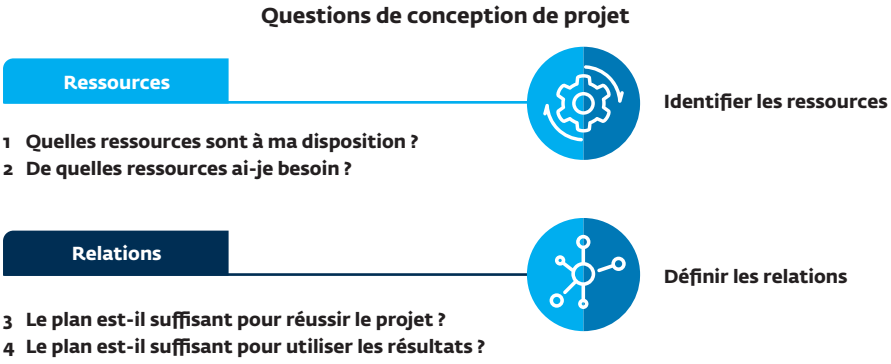



*Figure 25 : Les quatre questions de conception de projet posées par les Matrices d'Anneau des Données*

Avant de fermer cette section, il faut se souvenir de l'erreur la plus courante lors de l'utilisation de ces types d'outils d'entreprise : ne vous concentrez pas trop sur l'achèvement de la matrice. Autrement dit, la matrice de l'Anneau des données- tous comme la matrice d'affaires - n'est qu'un moyen, et non l'objectif lui-même.

## Définir et relier les ressources

### Définir les ressources

Les deux premières questions identifient les besoins en ressources du projet. Ceux-ci sont identifiés en posant séquentiellement la première question directrice : « Quelles sont les données dont je dispose ?... Quelles sont les compétences disponibles pour le projet ?... Quels sont les processus internes déjà en place ? ... » Les questions directrices associées à chaque composante doivent être prises en considération afin de préciser le processus de planification. Il faut ainsi se demander : « Quelle est la valeur dont je dispose ? » Il se peut que vous n'y répondiez pas en termes de résultats déjà atteints, mais au début, cette question peut s'avérer utile et pertinente. Il peut y avoir des méthodes d'ajustement à puiser dans des projets associés, ou peut-être des engagements préexistants de la direction quant au pilotage de la mise en œuvre. Il convient d'en tenir compte dans les ressources de valeur initiales qui sont les éléments moteurs de la planification globale.

Une fois que l'étendue des ressources est définie dans chaque bloc, les questions itèrent ce qui suit :

- De quelles données ai-je besoin ?
- De quelles compétences ai-je besoin ?
- De quel budget, point de comparaison, gestion des données ou plan d'ETL ai-je besoin ?

Ceci est particulièrement critique pour la valeur, car l'exploration de la valeur requise sous-tend la motivation du projet. En outre, la valeur est liée aux ressources obtenues par les résultats analytiques du projet. La planification des besoins de projets en termes de valeur contribue également à définir les livrables intermédiaires et finaux du projet, et notamment l'élaboration de rapports ou le développement de produits de connaissances. Cette approche séquentielle et itérative aide à identifier les lacunes et les exigences d'acquisition à mesure qu'elles se surviennent à chaque étape, en développant progressivement le plan général.

### Relier les ressources

Une fois les ressources spécifiées pour chaque bloc structurel, un plan de projet doit viser à comprendre leurs relations interconnectées de manière approfondie. Les deux dernières questions de conception de projet renvoient à ces relations ; c'est-à-dire, compte tenu des ressources envisagées dans un bloc de catégorie, la nécessité d'explorer si les ressources des autres catégories sont suffisamment reliées entre elles. Si ce n'est pas le cas, les exigences et les liens devront peut-être être ajustés les uns par rapport aux autres. Ces quatre liens sont spécifiés dans la Figure 26 ci-dessous : *ajustement*, *opérations*, *résultats* et *utilisation*. Chaque lien doit être spécifié pour remplir la matrice de l'Anneau des données et formuler un plan de projet holistique. Ceux-ci sont décrits ci-dessous :

**Relations de l'Anneau des données**

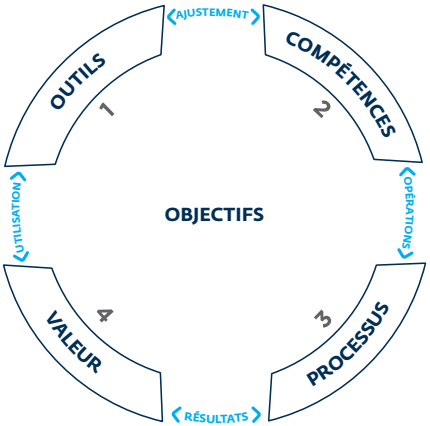



*Figure 26: Mettre en avant les liens entre les ressources dans la matrice de l'Anneau des données*

### AJUSTEMENT : Outils et compétences

Toutes les ressources logicielles et matérielles du projet doivent pouvoir fonctionner ensemble, une relation décrite par *Ajustement*. Cela peut sembler évident, mais l'expérience pratique nous a montré que la phase d'évaluation des ressources est souvent sous-estimée. Différents matériels et logiciels doivent « communiquer » entre eux. Les gens doivent également discuter, non seulement pour communiquer entre eux au sein de l'équipe, mais aussi pour utiliser l'infrastructure technique. La matrice doit spécifier les principaux langages de script et de base de données, ainsi que les méthodes des cadres spécifiques requises pour livrer le projet. Notamment, ces langues doivent être communes entre les équipes et les outils.

Les outils et les compétences doivent également s'adapter à la portée de l'objectif du projet. Le principal risque lié à une évaluation incorrecte des ressources est d'intégrer de force des composantes matériels avancées, des solutions logicielles entièrement développées ou des compétences humaines (par ex. des spécialistes des données) au projet sans intégration adéquate avec les infrastructures existantes et les experts du domaine. L'objectif de départ recommandé pour un processus et un produit qui soient un minimum viable contribue à atténuer ce risque en définissant les objectifs à partir de ressources plus modestes ; l'idée est d'explorer les idées et de tester des concepts de produits. Une fois le processus et le produit éprouvés, il est possible de développer progressivement le processus et le produit en utilisant les ressources pratiques et humaines nécessaires pour passer au niveau suivant.

### OPERATIONS : Compétences et processus

Les opérations du projet représentent le processus par lequel les gens abordent les calculs et l'exploration de données réels nécessaires pour livrer le projet. Ces activités sont pilotées par des questions d'analyse spécifiques et des problèmes opérationnels que l'équipe du projet cherche à résoudre. Par exemple, un projet de notation de risque de crédit se trouvera probablement confronté à un problème opérationnel spécifique pour calculer les variables qui se corrèlent aux taux de défaut de paiement. De même, une visualisation pourrait se trouver confrontée au problème technique consistant à définir la manière de représenter un réseau d'agent sur une carte. Les opérations se penchent sur ce que font les gens. Le bloc Processus définit comment les gens agissent en termes d'exigences de temps, budgétaires, de procédure ou de définition. Les opérations du projet se relient au Compétences dans la mesure où l'identification de solutions viables aux problèmes opérationnels nécessite des connaissances pertinentes sur le sujet. Les opérations de la matrice doivent spécifier les problèmes opérationnels fondamentaux du projet devant être abordés ; ils sont liés par les compétences nécessaires pour les traiter et le processus permettant de les résoudre.

### RÉSULTATS : Processus et Valeur

Les *résultats* informatiques de l'exécution du processus seront transformés en valeur. La matrice doit lister les résultats spécifiques attendus, qu'il s'agisse d'un algorithme, d'un modèle, d'un tableau de bord de visualisation ou d'un rapport analytique. La Valeur est réalisée grâce au processus par lequel les résultats sont interprétés, ajustés et mis en œuvre. Les approches de validation de modèle sont reliées au type de résultats de données du modèle sélectionné. Le choix du modèle est relié par les définitions et les cibles des indicateurs établis dans le Processus, et les mises en œuvre de l'interprétabilité et de l'utilisation commerciales qui créent de la Valeur. Les résultats numériques et leur interprétation sont associés au risque de ne pas pouvoir comprendre correctement les résultats obtenus. Il existe également un risque lors de la conversion de ces résultats en décisions ou leviers commerciaux qui fournissent de la valeur. Pour s'assurer que les résultats sont interprétables pour les besoins commerciaux, la matrice doit considérer ses principaux livrables et peut inclure des ressources supplémentaires qui facilitent l'interprétation des valeurs, comme un rapport analytique final. Il se peut également que des résultats de données supplémentaires ou des modèles supplémentaires doivent être spécifiés pour assurer une relation solide entre les blocs Processus et Valeur.

### UTILISATION : Valeur et outils

La quatrième question de conception de projet porte sur les résultats antérieurs, en vue de réaliser la valeur de l'*Utilisation* du projet. La conception du projet doit être suffisante pour utiliser le résultat du produit de données. Un tableau de bord de visualisation sera exécuté sur un ordinateur, par exemple connecté à un intranet interne ou à Internet. Un serveur Web le mettra en ligne afin que les gens puissent l'utiliser. Les données qu'il visualise seront stockées en un lieu auquel le tableau de bord doit se connecter pour accéder aux données. Le personnel informatique assurera la maintenance de ces serveurs. Ces ressources peuvent ou non être identifiées en fonction de ce qui est nécessaire pour livrer le projet lui-même. La quatrième question de conception de projet contribue à identifier les lacunes en matière de mise en œuvre qui pourraient apparaitre une fois le projet achevé, en veillant à ce que ces considérations soient exposées dans le cadre de la planification préalable du projet. *L'Utilisation* est reliée à la Valeur que le projet fournit avec les Outils nécessaires pour alimenter les données de sortie du projet dans le système de mise en œuvre. Ceci est particulièrement important pour les projets issus de solutions externalisées, où l'étendue des besoins en matière de soutien à la mise en œuvre doit être définie dans le cadre de l'acquisition initiale. L'Utilisation de la matrice doit spécifier comment la stratégie de mise en œuvre est reliée aux outils de mise en œuvre.

# CAS 14

## *Gérer le projet de métadonnées d'Airtel Money*

*Ce cas de gestion de projets s'appuie sur le cas d'Airtel Money Ouganda présenté au chapitre 1.2, Cas 3. Ce projet a été conçu et géré par l'équipe de recherche en Inclusion financière d'IFC basée en Afrique. Le cas d'utilisation ci-dessous passe en revue toutes les questions de conception du projet d'Anneau des données et se penche sur les spécificités de ce projet. Une matrice d'Anneau des données complétée reflète ce processus, en formulant les ressources clés du projet et les relations conceptuelles en une seule visualisation. Bien que cette matrice concerne un projet terminé, le processus d'utilisation d'une approche en matrice est dynamique ; l'écriture et la suppression de composantes présentant un défaut d'alignement entraine de nouvelles considérations en termes de conception et d'exigences. En outre, l'utilisation de post-it constitue une bonne approche, car ceux-ci permettent de procéder facilement à des ajouts et d'apporter de nouveaux éléments conceptuels et de nouveaux éléments de conception tout en permettant un mouvement dans la matrice, jusqu'à obtenir un plan satisfaisant.*

***Définition de l'objectif : Où commence l'Anneau des données***

*Un objectif est une solution pour un problème stratégique, et l'objet du projet est de fournir cette solution. Dans cet exemple, le problème était les faibles taux d'activité d'Airtel Money. IFC a proposé une solution : un modèle pour définir le profil statistique d'un utilisateur actif et faire correspondre ce profil aux non-utilisateurs dans la base d'abonnés GSM existante. Une fois identifiés, ces clients pourraient être efficacement ciblés en tant qu'utilisateurs Airtel Money à fortes propensions. Puisqu'on ne savait pas si cette correspondance de profil était possible, il était important de définir une portée modeste visant à prouver le concept :*

- ***L'Objectif :*** *Développer un modèle de prévision de segmentation de clientèle minimum viable pour identifier les utilisateurs actifs à fortes propensions qui augmenteraient les taux d'activité*
- ***L'Hypothèse :*** *Il existe une corrélation entre l'activité GSM et le comportement de l'activité Airtel Money (c.à.d. que des profils statistiques peuvent être créés et appariés)*

***Identification des ressources***

*IFC ne disposait pas des* ***données*** *d'Airtel au préalable, n'ayant obtenu qu'un engagement à un partenariat de la part d'Airtel sous forme de fourniture d'accès aux données de CDR et aux données des transactions Airtel Money. Bien que IFC et Airtel disposent d'une importante* ***infrastructure*** *informatique pour leurs opérations, celle-ci n'était pas disponible pour pouvoir être réquisitionnée par le projet. L'équipe d'IFC a chargé un spécialiste des opérations de données de gérer le projet, apportant les compétences* ***pertinentes*** *en* ***informatique, science des données*** *et en matière de* ***SFN****. Des spécialistes des SFN d'IFC, des spécialistes de la recherche en inclusion financière et des experts régionaux connaissant le marché local et les comportements des clients ont apporté leur soutien au projet. Lors de la* ***planification*** *du processus, le problème opérationnel était déjà connu : la faible activité d'Airtel Money. L'équipe disposait également de données* ***de comparaisons*** *existantes issues d'un projet de données similaire livré pour Tigo Ghana (voir chapitre 1.2, Cas 2 : Tigo Cash Ghana, Segmentation), qui ont*

*contribué à définir les **indicateurs** de gestion de projet, comme un objectif d'exactitude de 85 pour cent pour le modèle envisagé. Les **définitions** du modèle ont également spécifié à titre de variable dépendante « l'activité sur 30 jours ». Enfin, **un budget** a été attribué dans le cadre du projet de conseil d'IFC, financé par la Fondation Bill et Melinda Gates ; un **calendrier** à six mois a été défini.*

### Exploration des ressources

*Par le biais du **partenariat** du projet IFC-Airtel, l'équipe a négocié **l'accès** à six mois de **données** de CDR et Airtel Money, pour un volume d'environ un téraoctet, à extraire des bases de données relationnelles d'Airtel et fournies au format **CSV**. Il a fallu une **infrastructure** technique de traitement des métadonnées et de compétences en **science des données** pour les analyser. IFC a publié un appel d'offre (AO) concurrentiel pour **externaliser** ces éléments techniques, résultant sur la sélection de Cignifi, Inc. Cignifi a apporté des ressources en infrastructure supplémentaires, avec leurs clusters Hadoop-Hive de mégadonnées, son expérience sectorielle en matière de travail avec des données ORM et CDR, ses compétences en R et Python, son expérience dans les statistiques et l'apprentissage automatique et ses ressources pour la visualisation des données. L'équipe IFC-Airtel-Cignifi a ensuite établi un plan de **gestion des données** et d'ETL répondant aux exigences **juridiques et de confidentialité**. Conformément à ce plan, l'équipe Cignifi a été envoyée à Kampala, en Ouganda, pour travailler avec l'équipe informatique d'Airtel, dans l'objectif de comprendre leurs bases de données internes, définir les exigences associées à l'extraction de données, de crypter et de rendre anonymes les données sensibles, puis de transférer ces données sur un disque dur sécurisé pour les transférer sur les serveurs de Cignifi. Les attentes relatives à la **valeur** du projet étaient spécifiées dans l'OA en vue d'une sortie de données répertoriant les scores de propension des utilisateurs, appelée « liste blanche ». Des analyses supplémentaires ont également été spécifiées, notamment une cartographie de réseau social et une analyse géospatiale.*

### Suffisance du plan : livraison

*L'examen de la suffisance permet d'assurer l'alignement entre toutes les ressources, tous les processus et tous les résultats prévus. Il convient de noter que cela permet d'identifier au préalable les points qu'il est anticipé de devoir affiner lors du processus de mise en œuvre. L'examen participe également à la réévaluation des principaux domaines du processus lorsque des problèmes sont découverts lors de l'exécution analytique et nécessitent d'ajuster le plan.*

*La **gestion des données** prévoit l'affinement attendu ; la phase analytique et d'exécution du projet était de 10 semaines, mais a été planifiée **par rapport** à la date du début de l'acquisition des données, ce qui signifiait que le **calendrier** du projet serait affecté par la date réelle et les éventuels problèmes en matière d'ETL. Le **pipeline de données** présentait également une suffisance incertaine ; la planification du pipeline et l'affectation des ressources techniques étaient impossibles avant que les données finales puissent être analysées et que leur structure soit connue. Il s'agit d'un goulet d'étranglement fréquent. En anticipant ces incertitudes, la **valeur** ajoutée spécifiait une livraison initiale : un « dictionnaire de données » discutant de toutes les descriptions et relations liées aux données acquises, qui serait utilisé pour affiner la suffisance du projet une fois que ces détails seraient connus. C'est à la phase d'**exécution** de tout projet de*

*données que des surprises viennent mettre à l'épreuve les plans du projet. Étant donné qu'il s'agit de quelque chose auquel on peut s'attendre, le projet a également spécifié un livrable précoce sous forme de rapport de données provisoire, fournissant des statistiques descriptives de haut niveau et les résultats de l'analyse exploratoire initiale, les anomalies ou des lacunes observées dans les données. Le rapport de données provisoire doit également inclure tout imprévu susceptible de nécessiter un ajustement stratégique.*

### *Suffisance du plan : mise en œuvre*

*L'objectif de MVP du projet a cherché à tester si l'approche de modélisation était pertinente pour Airtel et le marché des SFN en Ouganda. En ce sens, le plan adopté était suffisant. Le projet devait fournir (a) un rapport final, avec les résultats et les analyses clés (b) une liste blanche : un ensemble de données des millions de clients de téléphonie mobile d'Airtel - par un identifiant crypté - chacun associé à un score de propension quant à la probabilité qu'ils utilisent activement Airtel Money.*

*Le plan adopté n'était pas suffisant dans le sens où les ressources étaient pré-affectées pour utiliser les informations de la liste blanche dans les campagnes de marketing, si l'analyse s'avérait réussie. La stratégie de livraison avait été convenue avec la direction d'Airtel : une réunion finale devait permettre la présentation et la discussion du rapport analytique, et l'équipe informatique d'Airtel devait utiliser la liste blanche et se servir de ses conclusions pour les étapes suivantes.*

### *Exécution du projet : ajustements à la planification*

*Les réalités sur le terrain nécessitent un ajustement du plan du projet. Les difficultés suivantes sont apparues au cours de l'exécution du projet et ont nécessité de réviser le plan pour s'assurer que tous les domaines du projet travaillaient suffisamment à la réalisation des objectifs.*

*Une fois l'ensemble de données initial sécurisé, le processus de **pipeline de données** a révélé des anomalies. D'une manière ou d'une autre, le processus d'extraction insérait des lignes vierges dans les ensembles de données brutes. Si les données pouvaient être transférées avec succès, elles étaient mal interprétées ; de nombreuses lacunes dans les données existaient, quand bien même ce n'était pas le cas. Le processus d'ETL devait être modifié. La correction apportée a révélé une erreur plus significative. L'ensemble de données du premier mois comportait de sérieuses lacunes, et ce problème a nécessité de réviser le plan de **gestion des données** et d'ETL et la conception globale du projet. Le plan de projet initial spécifiait des données d'octobre 2014 à mars 2015. La solution a consisté à rejeter entièrement les données d'octobre et à travailler avec Airtel pour extraire des données pour avril afin de maintenir la série chronologique de six mois nécessaire pour garantir un modèle statistiquement fiable. On a également découvert que, d'après le plan, les **données** elles-mêmes étaient insuffisantes. L'analyse géo spatiale et de réseau exigeant des données de localisation de l'antenne-relais. On a découvert que les ensembles de données Airtel Money n'établissaient pas l'emplacement des transactions effectuées, uniquement le moment où elles étaient réalisées. L'équipe Cignifi a contextualisé ces **métadonnées** en associant de manière créative les horodatages dans les données Airtel Money et les horodatages des appels vocaux pour les utilisateurs correspondants dans les données GSM. L'équipe a utilisé une fenêtre de 30 minutes, ce qui fournissait*

*une coordonnée de localisation qui était fiable dans un rapport distance/temps de 30 minutes à partir de l'emplacement de la transaction Airtel Money. Lors d'une discussion avec l'équipe d'IFC, il a été convenu que ces données étaient acceptables pour pouvoir procéder à l'analyse, même si elles reposaient sur l'hypothèse selon laquelle la plupart des gens, en moyenne, ne parcouraient pas de grandes distances dans la période de 30 minutes entre la réalisation d'une transaction Airtel Money et le passage d'un appel téléphonique.*

*La phase* ***d'ajustement*** *a nécessité un certain nombre de changements significatifs. Les statistiques sommaires des résultats du premier tour semblaient inhabituelles pour les spécialistes des SFN ; Ils ne correspondaient pas aux schémas comportementaux auxquels les experts en sciences sociales étaient habitués. Il a été découvert que les* ***définitions*** *du projet initial avaient donné une définition ambiguë du terme « utilisateur actif » de sorte que l'équipe d'analyse avait modélisé une sortie en termes d'une transaction de SFN dans les 30 jours de la date d'ouverture du compte Airtel Money, plutôt qu'une transaction dans une période de 30 jours sur la totalité de l'ensemble de données. La conception du modèle a ainsi dû être refaite. Ce qui a en fin de compte été profitable, car l'analyse initiale avait également révélé que les opérations de dépôt et de retrait ne fournissaient pas la robustesse statistique souhaitée pour atteindre les* ***indicateurs*** *de précision du projet. L'équipe IFC-Cignifi a accepté de refaire les modèles en utilisant les utilisateurs actifs redéfinis et de se recentrer sur les transactions P2P, considérées comme fournissant la plus grande précision et, surtout, pour définir les scores de propension associés au segment de clientèle générant les recettes les plus élevées. En outre, un modèle supplémentaire a été ajouté pour les « utilisateurs très actifs » ou ceux qui avaient réalisé une opération au moins une fois tous les trente jours sur une période consécutive de trois mois. Bien qu'il s'agisse d'un petit groupe, ces utilisateurs généraient près de 70 % du total des recettes d'Airtel Money ; le modèle supplémentaire visait à identifier ces clients de grande valeur.*

*Enfin, l'****interprétation*** *des résultats a abouti à des livrables de résultats de projet supplémentaires : les règles métier. Comme discuté dans le cas Airtel connexe, les algorithmes d'apprentissage automatique du modèle ont établi un certain nombre de variables significatives qu'il était difficile d'interpréter sur un plan commercial. L'équipe d'IFC a estimé que le livrable à la direction d'Airtel pouvait être renforcé en s'assurant que le modèle et les scores de propension de la liste blanche associés exprimaient le profil statistique des utilisateurs actifs en termes commerciaux, conformes aux ICP pertinents en termes commerciaux. Cignifi a livré trois mesures de segmentation rapide avec des « seuils» permettant de profiler les utilisateurs selon : le nombre d'appels vocaux par mois, le total des recettes des services téléphoniques par mois, et la durée totale des appels téléphoniques mensuels.*

## Une matrice remplie : La conception du projet de Mégadonnées d'Airtel, en utilisant la matrice de l'Anneau des données

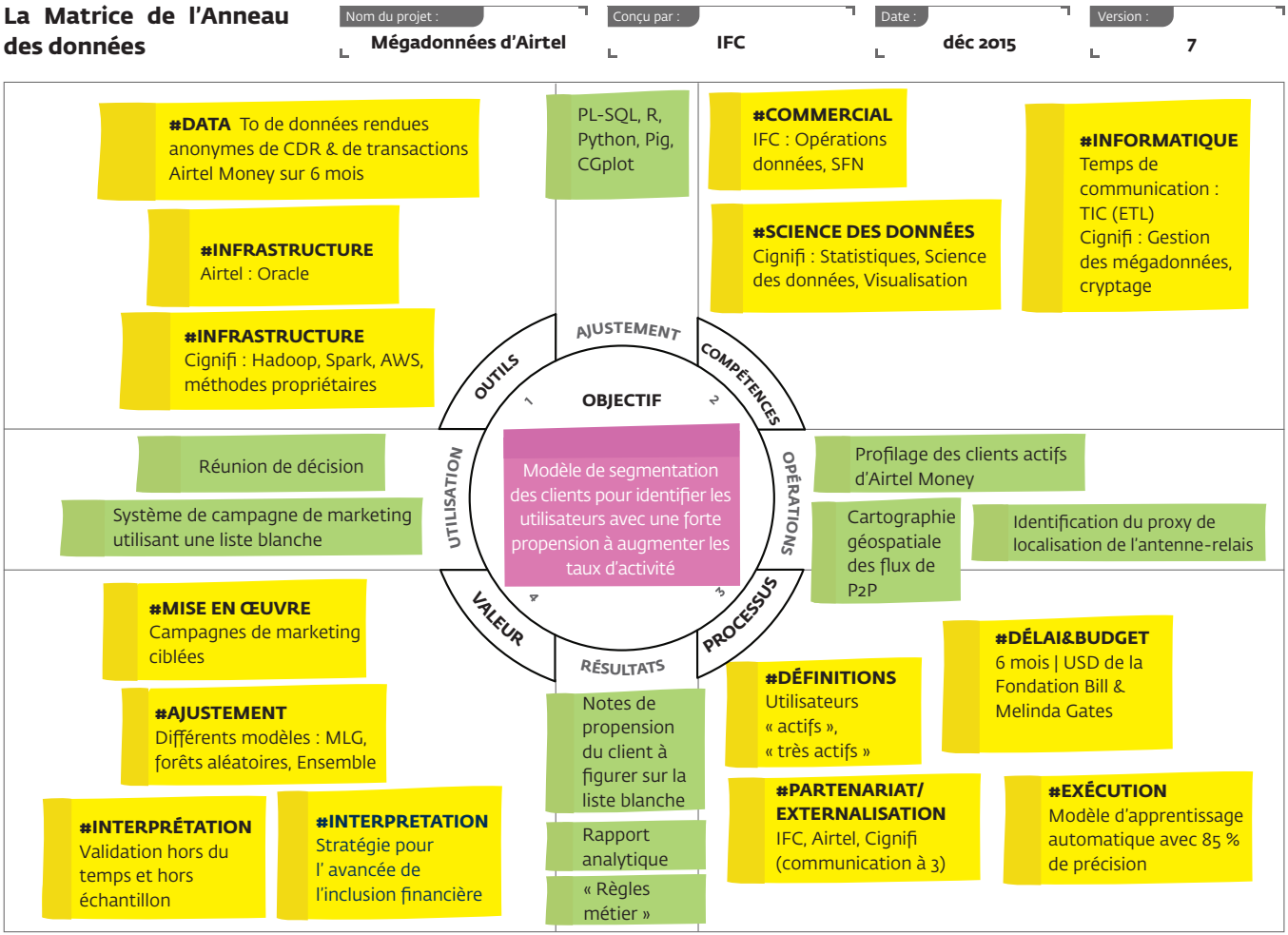



*Figure 27 : Une Matrice de l'Anneau des données remplie pour la phase I du projet de mégadonnées d'Airtel*

**Livraison du projet**

*La liste blanche du modèle a identifié environ 250 000 utilisateurs à propension maximale à cibler comme les utilisateurs actifs d'argent mobile attendus. Dans la liste blanche complète de plusieurs millions d'utilisateurs GSM, les 30 % supérieurs des scores de propension ont prédit que l'intérêt suscité des utilisateurs P2P hautement actifs générerait environ 1,45 milliard de shillings ougandais provenant des transactions P2P ; et 4,68 milliards de shillings ougandais provenant des retraits, soit environ 1,7 million de dollars de recettes annuelles supplémentaires.*

*Les conclusions du projet étaient solides et convaincantes. Toutefois, la stratégie de mise en œuvre n'a été définie que comme seuil décisionnel. La date de livraison a coïncidé avec une campagne de marketing existante, mettant en attente les résultats de la liste blanche. Les abonnés d'Airtel Money ont considérablement augmenté au cours des mois suivants, ce qui a réduit la valeur de la liste blanche puisque de nombreux nouveaux clients ont été intégrés par le marketing habituel. Sur cette période, les abonnés GSM ont également augmenté, fournissant des millions de nouveaux utilisateurs potentiels d'Airtel Money. IFC et Airtel ont accepté de procéder à une analyse de Phase II à la fin de l'année 2016. L'objectif du projet est le même, avec une composante analytique supplémentaire fondée sur la Phase I, conçue pour examiner les schémas d'intérêt suscité et de distribution d'Airtel Money dans le temps et dans l'espace.*


airtel money
Support Me


# PARTIE 2

## *Chapitre 2.2 Ressources*

### 2.2.1 Synthèse des classifications des cas d'utilisation analytiques

| Synthèse des classifications des cas d'utilisation analytiques | | | |
|---|---|---|---|
| **Classification** | **Question traitée** | **Techniques** | **Mise en œuvre** |
| **Analyse descriptive** | • Que s'est-il passé ?<br>• Que se passe-t-il maintenant ? | Alertes, requêtes, recherches, rapports, visualisations statiques, tableaux de bord, tableaux, graphiques, récits, corrélations, analyse statistique simple | Rapports |
| **Analyse diagnostique** | • Pourquoi cela s'est-il produit ? | Analyse de régression, test A/B, filtrage par motif, exploration de données, prévision, segmentation | Veille technologique traditionnelle |
| **Analyse prédictive** | • Que se passera-t-il à l'avenir ? | Apprentissage automatique, ARS, analyse géo spatiale, reconnaissance de formes, visualisations interactives | Modélisation |
| **Analyse prescriptive** | • Que faut-il faire pour parvenir à un résultat donné ? | Analyse graphique, réseaux neuronaux, Apprentissage automatique et profond, IA | Solutions intégrées, décisions automatisées |

## 2.2.2 Répertoire des sources de données

| Source : Systèmes de noyau bancaire et d'ORM | | |
|---|---|---|
| Structure : Données habituellement structurées, utilisant des bases de données relationnelles. | | |
| Format : Données numériques, qui peuvent être extraites sous différents formats pour la production de rapports ou l'analyse. Les données anciennes peuvent inclure des inscriptions papier ou des formulaires d'inscription scannés. | | |
| **Nom** | **Données** | **Exemples** |
| Données de l'émetteur de factures sur les clients | Durée du contrat ; historique de paiement ; types d'achat | Meilleures connaissances marketing ; possibilité de créer une notation de risque de crédit en utilisant des données de facturation |
| Statut de l'inscription du client | Statut de l'inscription (p.ex. actif, inactif, jamais utilisé) | Connaissances marketing ; suivi des performances commerciales ; conformité réglementaire |
| KYC client | Nom, adresse, DN, sexe, revenu | Connaissances marketing ; conformité réglementaire |
| Statut du compte | Type de compte, statut de l'activité (actif, dormant, ancienneté de l'activité, dormant avec solde) | Connaissances marketing ; suivi des performances commerciales ; conformité réglementaire |
| Activité du compte | Solde du compte, vitesse mensuelle, solde moyen quotidien | Connaissances marketing ; notation de risque de crédit ; conformité réglementaire |
| Données de transaction financière (directes) | Volume et valeur des dépôts, des retraits, des paiements de factures, des transferts ou d'autres transactions financières | Suivi des performances commerciales et financières ; conformité réglementaire ; connaissances marketing ; notation de risque de crédit |
| Données de transaction financière (indirectes) | Transactions échouées ; transactions refusées ; canal utilisé ; heure de la journée | Problèmes de performance et de conception de produits ; besoins de formation et de communication |
| Données de monnaie électronique | Fonds de caisse électroniques, rapprochements, transferts de fonds de caisse entre agents | Gestion de la performance des agents ; gestion de la fraude et des risques |
| Activités non financières | Changement de code PIN ; demande de solde ; demande de relevé | Connaissances marketing ; amélioration de l'efficacité ; développement de produits |
| Origine du prêt | Type de prêt, montant du prêt, garantie utilisée, durée, taux d'intérêt | Connaissances marketing ; suivi de la performance du portefeuille ; notation de risque de crédit : nouvelle évaluation de prêt |
| Activité de prêt | Solde du prêt, statut du prêt, source de la transaction de remboursement du prêt | Connaissances marketing ; suivi de la performance du portefeuille ; notation de risque de crédit : nouvelle évaluation de prêt |

| Source : Système d'argent mobile | | |
|---|---|---|
| Structure : Données habituellement structurées, utilisant des bases de données relationnelles. | | |
| Format : Données numériques, qui peuvent être extraites sous différents formats pour la production de rapports ou l'analyse. Les données anciennes peuvent inclure des inscriptions papier ou des formulaires d'inscription scannés. | | |
| **Nom** | **Données** | **Exemples** |
| KYC client | Nom, adresse, DN, sexe, revenu | Connaissances marketing ; conformité réglementaire |
| Statut de l'abonnement | Statut de l'activité (actif, dormant, ancienneté de l'activité, dormant avec solde) | Connaissances marketing ; suivi des performances commerciales ; conformité réglementaire |
| Activité de portefeuille | Solde du portefeuille, vitesse mensuelle, solde moyen quotidien | Connaissances marketing ; notation de risque de crédit ; conformité réglementaire |
| Données de transaction | Volume et valeur des dépôts, des retraits, des paiements de factures, du P2P, des transferts, des rechargements de temps de communication ou autres opérations financières | Suivi des performances commerciales et financières ; conformité réglementaire ; connaissances marketing ; notation de risque de crédit |
| Données de monnaie électronique | Fonds de caisse électroniques, rapprochements, transferts de fonds de caisse entre agents | Gestion de la performance des agents ; gestion de la fraude et des risques |
| **Source : Système de gestion des agents** | | |
| Structure : Données habituellement structurées, utilisant des bases de données relationnelles. | | |
| Format : Données numériques, qui peuvent être extraites sous différents formats pour la production de rapports ou l'analyse. Les données anciennes peuvent inclure des inscriptions sur papier, des formulaires d'inscription scannés ou des rapports de suivi ou de performance des agents. | | |
| **Nom** | **Données** | **Exemples** |
| Activités des agents (direct) | Volume et valeur des transactions des agents ; transfert de fonds de caisse ; dépôt et retrait de fonds de caisse ; solde de fonds de caisse ; jours sans fonds de caisse | Connaissances vente et marketing ; notation de risque de crédit ; gestion de la performance de l'agent |
| Activités de l'agent (indirect) | Changement de code PIN ; demande de solde ; demande de relevé ; créer un nouvel assistant | Connaissances vente et marketing ; gestion de la performance de l'agent |
| Activités du commerçant (direct) | Volume et valeur des transactions du commerçant ; nombre de clients uniques | Connaissances vente et marketing ; notation de risque de crédit ; gestion de la performance marchande |
| Activités du commerçant (indirectes) | Changement de code PIN ; demande de solde ; demande de relevé ; créer un nouvel assistant | Connaissances vente et marketing ; notation de risque de crédit ; gestion de la performance marchande |
| Données techniques du système | Nombre de TPS ; files d'attente de transactions ; temps de traitement | Planification de la capacité ; suivi de la performance par rapport au SLA ; identification des problèmes de performance technique |
| Rapports de visite des agents et des commerçants par le personnel des ventes | Présence de matériel de merchandising ; connaissances des assistants ; volume du fonds de caisse ; peut inclure plus fréquemment des données semi-structurées ou non structurées, par ex. des rapports de suivi sur papier | Indications sur les clients ; gestion de la performance de l'agent |

| Source : Système de gestion de la relation client (GRC) | | |
|---|---|---|
| Structure : Incorpore souvent des données structurées et semi-structurées qui utilisent des bases de données relationnelles ou des systèmes de stockage basés sur des fichiers, tels que des enregistrements vocaux ou des synthèses de problèmes identifiées par catégories structurées. | | |
| Format : Données numériques, en général, bien que les données semi-structurées et non structurées ne soient peut-être pas disponibles dans les rapports (par ex. pour les enregistrements vocaux, le cas échéant). | | |
| **Nom** | **Données** | **Exemples** |
| Enregistrements du centre d'appels | Journal des problèmes, type de problèmes, délai de résolution (peut inclure des données semi-structurées dans les rapports) | Indications sur les clients ; gestion opérationnelle et de la performance ; améliorations du système |
| PBAX | Nombre d'appels du centre d'appels ; durée des appels ; temps d'attente de la file d'attente ; appels abandonnés | Gestion opérationnelle et de la performance |
| Données de retour d'information du service clients | Nombre d'appels ; statistiques sur les types d'appel ; statistiques de résolution des problèmes | Identifier : les problèmes de performance technique et de conception de produit ; les besoins de formation et de communication ; les problèmes liés à un tiers (par ex. agent, émetteur de factures) |
| Données de rétroaction des agents et commerçants | Nombre d'appels d'agents ou de commerçants ; statistiques sur les types d'appel ; statistiques de résolution des problèmes | Identifier : les problèmes de performance technique et de conception de produit ; les besoins de formation et de communication ; les problèmes de clientèle |
| Interactions du canal de communication | Volume des visites sur le site Web, volumes du centre d'appels, enquêtes sur les réseaux sociaux, demandes de chat en direct | Indications sur les clients ; gestion opérationnelle et de la performance ; améliorations du système |
| Données de communication qualitatives | Type de demandes de renseignements, satisfaction du client, examen des réseaux sociaux | Indications sur les clients |
| **Source : Dossiers clients** | | |
| Structure : Incorpore souvent des données structurées, semi-structurées et non structurées, allant de : documents de la KYC qui peuvent inclure diverses informations personnelles selon le type de document ; aux études de marché ou enquêtes clients ; et aux notes de groupes de discussion. | | |
| Format : Une grande diversité de formats peut être utilisée pour stocker des données de dossiers client, notamment les bases de données relationnelles, des systèmes de stockage de fichiers ou les documents numérisés ou papier. | | |
| **Nom** | **Données** | **Exemples** |
| Documents de la KYC | Pièce d'identité ; justificatif de salaire ; justificatif de domicile | Conformité réglementaire ; segmentation démographique et géographique |
| Formulaires d'inscription et de demande | Ouverture de compte de SFN ; demande de prêt | Conformité réglementaire ; segmentation démographique et géographique |
| Recherche qualitative | Entretiens avec les clients ; groupes de discussion | Connaissances marketing et produits |
| Recherche quantitative | Études de sensibilisation et d'usage ; études de sensibilité aux prix ; tests pilotes | Connaissances marketing et produits |

| Source : Dossiers agents et commerçants | | |
|---|---|---|
| Structure : Incorpore souvent des données structurées, semi-structurées et non structurées, allant de : documents de la KYC qui peuvent inclure une diversité d'informations personnelles selon le type de document ; aux études de marchés ou enquêtes auprès du commerçant ; et aux notes de groupes de discussion. | | |
| Format : Une grande diversité de formats peut être utilisée pour stocker des données de dossiers agent ou commerçant, notamment les bases de données relationnelles, des systèmes de stockage de fichiers ou les documents numérisés ou papier. | | |
| Documents de la KYC | Statuts ; déclarations de revenus ; documents de KYC ; relevés bancaires | Conformité réglementaire ; segmentation démographique et géographique |
| Formulaires d'inscription | Enregistrement en tant qu'agent ou commerçant fournissant des SFN | Conformité réglementaire ; segmentation démographique et géographique |
| Recherche qualitative | Entretiens avec les agents ; groupes de discussion | Connaissances ventes, marketing et produits |
| Recherche quantitative | Enquête par des achats anonymes effectués par des enquêteurs | Connaissances ventes, marketing et produits |
| **Source : Partenaires tiers** | | |
| Structure : Le tiers peut prendre toute forme ou structure, selon le contenu, la source et le prestataire qui la fournit. | | |
| Format : Les formats peuvent aller des formats communs .CSV aux API d'accès exclusif et aux méthodes de livraison. | | |
| **Nom** | **Données** | **Exemples** |
| Données de l'émetteur de factures à propos des clients (services publics) | Durée du contrat ; historique de paiement ; types d'achat | Meilleures connaissances marketing ; possibilité de créer une notation de risque de crédit en utilisant des données de facturation |
| Données sur les clients de payeurs des clients (employeur, gouvernement) | Historique de la rémunération ; durée des paiements réguliers | Amélioration des connaissances marketing ; notation de risque de crédit |
| Référentiels d'informations sur le client (par ex. bureau de crédit, listes de surveillance, casiers judiciaires) | Données de la KYC ; notation de risque de crédit ; activité frauduleuse passée | Notation de risque de crédit ; enquêtes sur la fraude ; gestion des risques |
| Données géo spatiales (données satellitaires) | Données démographiques régionales ; densité de population ; topographie ; infrastructures telles que les routes et réseau électriques ; points d'accès financiers | Indications sur le marché ; gestion des agents |
| Médias et Réseaux sociaux | Type et fréquence des activités sur le réseau ; informations personnelles ; nombre de connexions ; type de connexions | Indications sur le marché ; notation de risque de crédit |

## 2.2.3 Indicateurs pour l'évaluation des modèles de données

| Liste des 10 meilleurs indicateurs de performance pour l'évaluation des modèles de données | |
|---|---|
| **Indicateur** | **Définition** |
| Courbe de la fonction d'efficacité du récepteur (ROC) | La courbe ROC est définie comme la courbe entre le taux de vrais positifs et le taux de faux positif. Elle illustre la performance du modèle selon la variation de son seuil de discrimination. Plus la zone entre la courbe ROC et la courbe de référence est importante, meilleure est le modèle. |
| AUC | La zone sous la courbe (AUC) mesure la zone sous la courbe ROC. Elle fournit une estimation de la probabilité que la population soit correctement classée. Elle représente la capacité du modèle à produire un bon classement relatif des instances. Une valeur égale à un constitue un modèle parfait. |
| KS | Le test statistique Kolmogorov-Smirnov (KS) mesure la séparation verticale maximale entre la distribution cumulative des « bons » et des « mauvais ». Cela représente la capacité du modèle à séparer la « bonne » population visée de la « mauvaise » population. |
| Diagramme de lift | Mesure l'efficacité d'un modèle prédictif calculé comme le rapport entre les valeurs prédites positives sur le nombre de positifs dans l'échantillon pour chaque seuil. Plus la zone entre la courbe de lift et la courbe de référence est importante, meilleur est le modèle. |
| Gains cumulés | Mesure l'efficacité d'un modèle prédictif calculé comme le pourcentage de valeurs prédites positives pour chaque seuil. Plus la zone entre la courbe de gains cumulés et la courbe de référence est importante, meilleur est le modèle. |
| Coefficient de Gini | Le coefficient de Gini est lié à l'AUC ; G (i) = 2AUC-1. Il fournit également une estimation de la probabilité que la population soit correctement classée. Une valeur égale à un constitue un modèle parfait. C'est la définition statistique de ce qui influence l'indice économique de Gini pour la distribution des revenus. |
| Exactitude | L'exactitude est la capacité du modèle à faire une prévision correcte. Elle est définie comme le bon nombre de prévisions sur toutes les prévisions réalisées. Cette mesure ne fonctionne que lorsque les données sont équilibrées (c.à.d. une même distribution des bons et des mauvais). |
| Précision | La précision est la probabilité qu'une instance sélectionnée de manière aléatoire soit positive ou bonne. Elle est définie comme le rapport du total des instances positives prédites vraies sur le total des instances positives prédites. |
| Rappel | Le rappel est la probabilité qu'une instance sélectionnée de manière aléatoire soit bonne ou positive. Elle est définie comme le rapport du total des instances positives prédites vraies sur le total des instances positives. |
| Erreur moyenne quadratique (RMSE) | La RMSE est une mesure de la différence entre les valeurs prédites par un modèle et les valeurs effectivement observées. Cette mesure est utilisée dans les prévisions numériques. La RMSE d'un bon modèle doit être faible. |

## 2.2.4 Anneau des données et matrice de l'Anneau des données

Les outils que sont l'Anneau des données et la matrice de l'Anneau des données sont également disponibles au téléchargement sur le site Web du Partenariat pour l'Inclusion Financière, au lien suivant : www.ifc.org/financialinclusionafrica

« La page détachable ci-dessous fournit une copie de l'Anneau des Données  et de la Matrice de l'Anneau des Données »

## 2.2_RESSOURCES

**L'Anneau des données**



*Manuel sur l'analyse de données et les services financiers numériques (ISBN : 978-0-620-76146-8).*

## 2.2_RESSOURCES

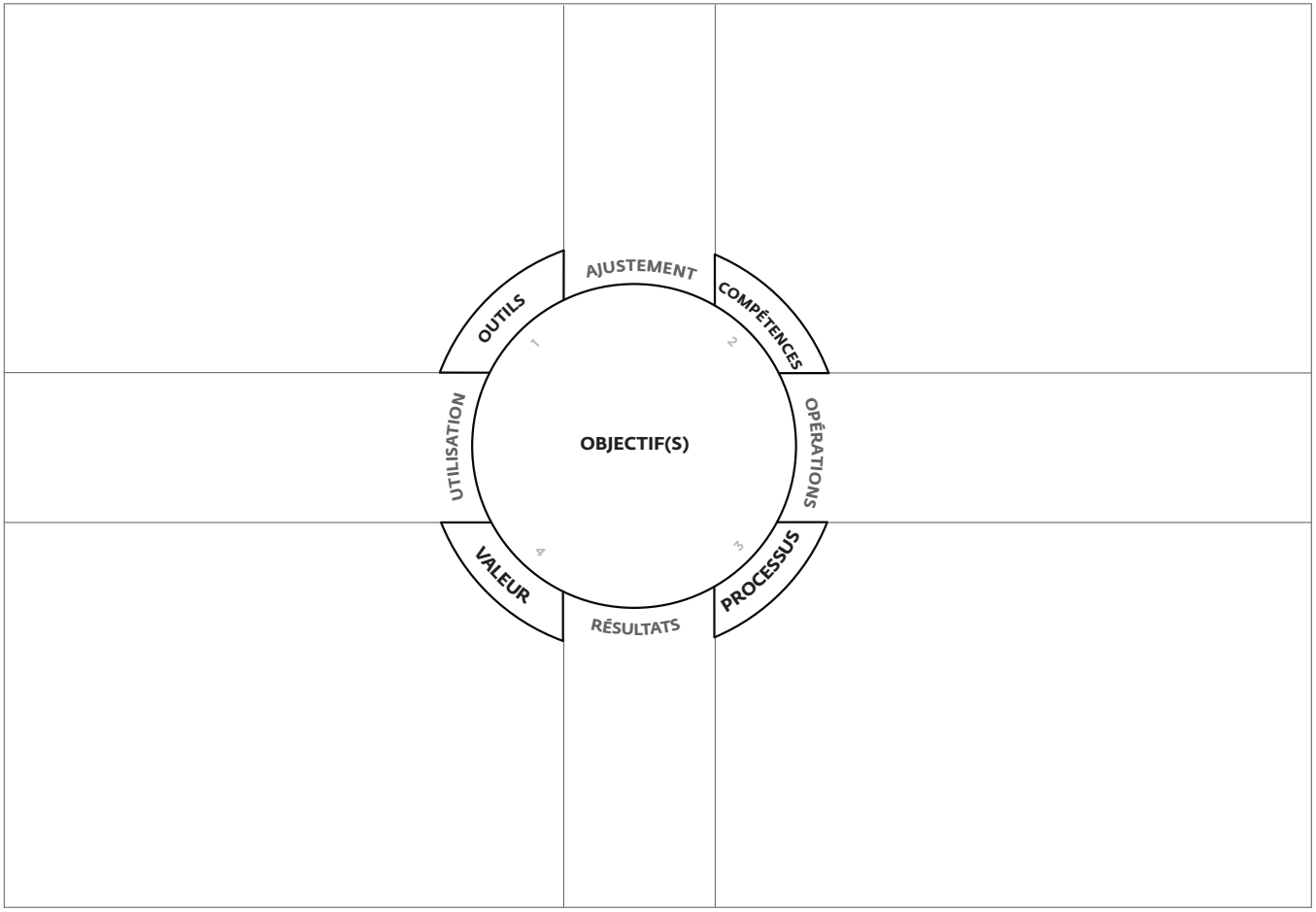
### La Matrice de l'Anneau des données

Nom du projet :

Conçu par :

Date :

Version :





*Manuel sur l'analyse de données et les services financiers numériques (ISBN : 978-0-620-76146-8).*

# Conclusions et leçons tirées

L'univers des données grandit à chaque heure. La capacité analytique de l'informatique est également de plus en plus évoluée et le cout du stockage des données diminue. Le potentiel d'analyse de données décrit dans ce manuel, et dans ces études de cas, met en évidence la façon dont les prestataires de SFN peuvent tirer parti des données, grandes et petites, pour créer de nouveaux services et parvenir à une plus grande efficacité dans leurs opérations actuelles en intégrant des approches fondées sur les données. Les praticiens doivent s'efforcer d'adopter une approche basée sur les données dans leurs activités. Cela apportera davantage de précision à leurs activités et une approche s'appuyant sur des données issues de faits concrets pour la prise de décision.

## Développer une culture fondée sur les données

La culture organisationnelle est essentielle. Les organisations doivent créer un environnement favorable aux données où le pouvoir des données est salué et où les personnes sont habilitées et encouragées à explorer afin de trouver des moyens d'améliorer les résultats. En conséquence, il est nécessaire d'investir dans les compétences des équipes opérationnelles, les outils et les idées afin de valoriser les données. Le leadership organisationnel doit clairement formuler la vision et les normes fondamentales qui formeront la base de son programme de gestion des données. Le leadership doit également s'engager fermement à développer les capacités de l'entreprise en matière de données, tant en termes de vision que de budget. En outre, il est essentiel qu'un service ou un individu soit clairement défini et dispose d'une influence au sein de l'organisation qui anime le processus. Certaines organisations qui sont à un stade plus avancé sur la courbe de maturité ont choisi de créer un poste de haut niveau intitulé directeur des données (DD) ; cette personne travaille en étroite collaboration avec les membres de la direction de l'entreprise pour gérer toutes les stratégies et la gestion liées aux données.

L'organisation doit se pencher sur ses capacités et son expérience actuelles afin de définir clairement l'avenir. Les considérations importantes sont la taille de l'organisation ainsi que les ressources informatiques existantes telles que les compétences et l'expérience. En outre, passer à une approche fondée sur les données impliquera de grands changements pour la culture organisationnelle, en particulier sur la façon dont les données sont partagées et dont les décisions sont prises. L'organisation devra être prête à fournir un soutien continu pendant le changement et doit être préparée à gérer les attentes du personnel et de la direction. Les niveaux actuels de maturité de la gestion des données sont également importants. Le prestataire de SFN souhaitera peut-être examiner les sources de données actuelles, le cadre d'établissement de rapports et l'utilisation des données dans la prise de décision pour se positionner sur la courbe de maturité. Comprendre où l'on se positionne sur l'échelle de maturité de la gestion des données permet au prestataire de développer une feuille de route menant à l'objectif souhaité.

Se fonder sur les données implique également l'examen du savoir-faire du personnel existant et l'évaluation du niveau d'aisance des membres de l'équipe avec la technologie et l'informatique. Le personnel existant peut être formé à gérer les nouvelles technologies. Il est idéalement placé pour appliquer de nouvelles technologies à d'anciens problèmes car ils connaissent déjà l'organisation, son marché et ses défis. En règle générale, le personnel exigera une formation théorique et pratique en gestion de données. Le prestataire de SFN souhaitera peut-être identifier les membres du personnel qui présentent une aptitude et ont la bonne attitude à l'égard de l'adoption de nouvelles pratiques technologiques, puis préparera un plan en vue du développement intensif des compétences.

Peu importe le niveau d'une organisation en termes d'adoption des analyses fondées sur les données, il est possible d'intégrer systématiquement les données dans ses processus et prises de décision. Les praticiens peuvent prendre de petites mesures pour commencer à tester rigoureusement les besoins et les préférences de leurs clients, suivre les performances en interne et comprendre l'impact de leurs activités commerciales. Le plus important est que les objectifs fixés par une organisation pour suivre les performances de l'entreprise soient quantifiables et mesurables.

## Toutes les données sont de bonnes données

L'analyse des données offre aux prestataires de SFN l'opportunité d'acquérir une

compréhension bien plus détaillée de leurs clients. Ces idées peuvent être utilisées pour concevoir de meilleurs processus et procédures qui correspondent aux besoins et aux préférences des clients. L'analyse des données consiste à comprendre les clients, dans l'objectif de tirer une plus grande valeur du produit.

Notamment, combiner les indications offertes par différentes méthodologies et sources de données peut enrichir la compréhension. À titre d'exemple, bien que les données quantitatives puissent donner des indications sur ce qui se passe, les données qualitatives et la recherche permettront d'expliquer pourquoi cela se produit. De même, plusieurs prestataires de SFN ont utilisé une combinaison de modélisation prédictive et d'analyse de géolocalisation pour identifier les domaines cibles sur lesquels ils doivent concentrer leurs efforts de marketing.

Pour le vaste marché de masse que les prestataires de DFS desservent, dans de nombreux cas, il se peut qu'il n'y ait pas d'antécédents financiers formels ou d'historique des données de remboursement à utiliser comme base. Dans ces situations, des données alternatives peuvent permettre aux prestataires de SFN de vérifier les flux de trésorerie par le biais d'informations indirectes, telles que les données des ORM. Ici, les prestataires de SFN ont le choix de travailler directement avec un ORM ou avec un fournisseur. La décision dépend des marchés respectifs ainsi que d'état de préparation de l'établissement. De nombreux prestataires peuvent ne pas avoir le savoir-faire technique pour concevoir des modèles de notation fondés sur des données d'ORM - dans ce cas, le partenariat avec un fournisseur offrant ce service est une bonne option.

## Utilisation de la visualisation des données

Une image vaut mieux que mille mots, ou plutôt qu'une longue série de chiffres. L'utilisation de visualisations pour illustrer graphiquement les résultats des rapports standards de gestion de données peut faciliter la prise de décision et la surveillance. Les représentations graphiques permettent au public d'identifier rapidement les tendances et les valeurs aberrantes. Cela est vrai en ce qui concerne les équipes internes de science des données qui explorent les données, ainsi que pour les communications plus générales, lorsque les tendances et les résultats des données peuvent avoir plus d'impact que les tableaux en visualisant les relations ou des conclusions axées sur les données.

Un graphique ou une courbe est une visualisation de données, au sens le plus élémentaire du terme. Cela dit, la « visualisation » comme concept et comme discipline émergente est beaucoup plus vaste, à la fois en ce qui concerne les outils disponibles et les résultats possibles. Par exemple, une infographie peut être une visualisation de données dans de nombreux contextes, mais ce n'est pas nécessairement une courbe. Dans certains cas, cette portée peut également inclure des médias mixtes. Un exemple de pionnier dans ce domaine serait par exemple Hans Rosling, dont le travail consistant à combiner la visualisation des données avec les histoires interactives sur médias mixtes lui a valu une place dans la liste des 100 personnes les plus influentes du Time.[40] Ces éléments de dynamisme et d'interactivité ont élevé le champ de visualisation des données bien au-delà des graphiques et des courbes, même si le domaine englobe également ces outils plus traditionnels.

La visualisation des données est liée mais distincte des tableaux de bord de données. Un tableau de bord inclura probablement une ou plusieurs visualisations plus à l'écart. Les tableaux de bord sont des points de référence incontournables, qui servent souvent de points d'entrée pour des données plus détaillées ou des outils de génération de rapports. C'est là que les ICP sont visualisés pour fournir des informations instantanées, généralement pour les responsables qui ont besoin d'avoir un aperçu concis du statut opérationnel. Des tableaux de bord simples peuvent être mis en œuvre dans Excel, par exemple. Habituellement, le concept de tableau de bord se réfère à des représentations de données plus sophistiquées, intégrant les idées d'interactivité et de dynamisme qu'englobe le concept plus large de visualisation de données. En outre, des tableaux de bord plus sophistiqués sont susceptibles d'inclure des données en temps réel et des réponses aux requêtes des utilisateurs. Bien que la visualisation des données et les tableaux de bord de données soient intrinsèquement liés et se chevauchent souvent, il est également important de reconnaitre qu'ils sont conceptuellement différents et évalués selon différents critères. Cela permet d'attester que les bons outils sont appliqués pour le bon travail et que les fournisseurs et les produits sont acquis aux fins prévues.

## La science des données est l'art des données

Le chapitre 1 a indiqué que l'histoire du terme « science des données ». Fait intéressant, ceux qui l'ont inventé ont hésité à appeler les spécialistes de la discipline « scientifiques des données » ou « artistes des données ». Si science des données a finalement été choisi, il convient de reconnaitre que la créativité, le design et même la sensibilité artistique restent des éléments critiques

[40] Hans Rosling. In Wikipedia, the Free Encyclopedia, accessed April 3, 2017, https://en.wikipedia.org/wiki/Hans_Rosling

dans ce domaine. Suite à la discussion ci-dessus sur la visualisation des données, le processus consistant à transformer des bits de données en outils informatifs, interactifs, esthétiquement agréables et visuellement intéressants nécessite à la fois des compétences techniques et des idées créatives. En référence à Rosling, le processus consistant à faire de la visualisation des données le personnage principal de ce qui peut être qualifié de performance théâtrale souligne encore l'interaction entre la science des données et l'art des données. Le rôle des scientifiques des données, indépendamment de leur appellation fonctionnelle, consiste à s'appuyer sur les compétences techniques et l'intuition créative pour explorer les schémas, extraire la valeur de ces relations et communiquer leur importance.

Ce dualisme d'organisation structurée et de schémas émergents décrit l'une des complexités globales de nombreux projets de données. D'une part, il est nécessaire d'avoir des objectifs clairs, une architecture définie et une expertise précise pour s'assurer que la livraison de projets se fait dans le respect des échéances et du budget. D'autre part, il est très important de faire preuve d'une flexibilité ouverte pour pouvoir découvrir des modèles, explorer de nouvelles idées, extraire des données pour découvrir les anomalies possibles, tester des hypothèses et concevoir de manière créative des visualisations pour raconter l'histoire des données.

## Le secteur des données dans le monde

Le domaine de la science des données existe depuis moins d'une décennie, le terme lui-même n'ayant véritablement pris de l'importance qu'en 2008 (voir la figure 6 de la partie 1). Depuis, les smartphones sont devenus omniprésents, la puissance informatique a considérablement augmenté et les couts de stockage ont chuté. Les entreprises technologiques ont introduit de nouveaux produits qui ont été rapidement assimilés à la vie quotidienne, tels que Google Maps, le chat vidéo Face Time d'Apple et l'AI domestique d'Amazon, Alexa. Les produits fondés sur les données sont rapidement adoptés dans tous les secteurs, les grands ensembles de données et les outils des sciences des données offrant une valeur innovante sur les marchés établis. Le milieu des années 2000 a vu l'émergence de l'analyse de données connaitre une forte croissance au-delà du secteur de la technologie, en particulier en réalisant des progrès anticipés dans le secteur des biens de consommation courante (BCC), comme dans les épiceries et les grands magasins. Le secteur mondial a changé en l'espace de quelques années, fait résumé par l'observation largement diffusée de Tom Goodwin : « Uber, la plus grande entreprise de taxis au monde, ne possède aucun véhicule. Facebook, le propriétaire de média le plus populaire au monde, ne crée aucun contenu. Alibaba, le plus grand détaillant qui soit, n'a pas de stock. Et Airbnb, le plus grand fournisseur d'hébergement au monde, ne possède aucun bien immobilier. Quelque chose d'intéressant est en train de se produire ». Les solutions fondées sur les données ont permis aux nouveaux entrants de perturber les secteurs établis, et les entreprises technologiques continuent à repousser les limites.

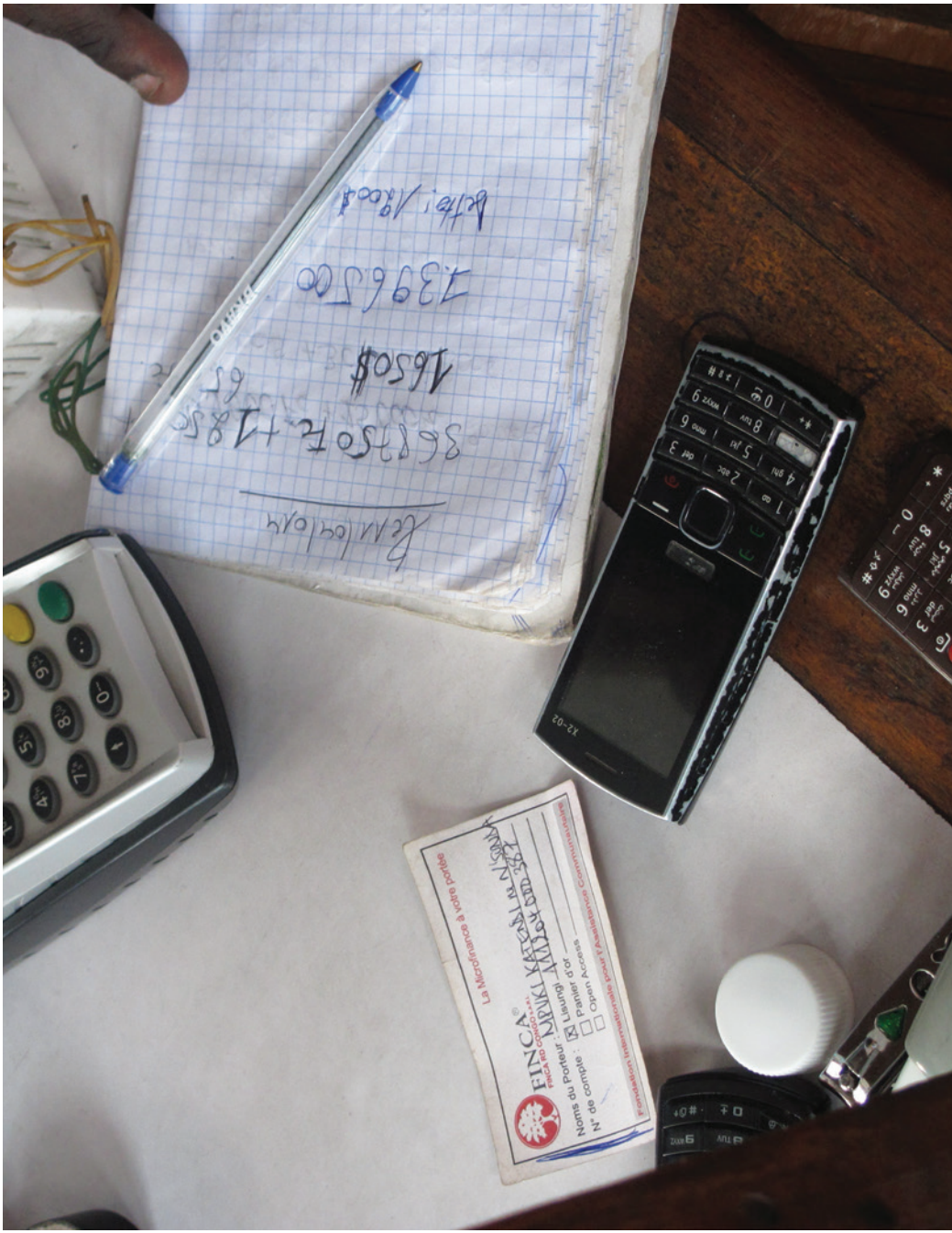
Les méthodes alternatives de notation de risque de crédit trouvent de nouvelles sources de données qui permettent aux produits d'atteindre de nouveaux segments de clients, en s'appuyant souvent sur la technologie des réseaux sociaux. Les stratégies de marketing sont affinées par des tests statistiques A/B rigoureux, qui ont été mis en avant par des sociétés comme Amazon ou Yahoo! pour affiner la conception de leurs sites Web. De plus, l'analyse géographique de la segmentation des clients, la cartographie des flux P2P et l'identification de la localisation optimale des agents bénéficient tous de l'aide de l'analyse géo spatiale et des outils qui fournissent la technologie Google Maps et OpenStreetMap. À mesure que la technologie évolue, les prestataires de SFN peuvent s'attendre à ce que de nouvelles solutions émergent pour mieux comprendre les clients, atteindre de plus grands marchés et fournir des produits et des services adaptés aux besoins des clients.

## Données pour l'inclusion financière

Dans le secteur de l'inclusion financière, les données sont importantes car la clientèle cible n'a souvent pas accès aux banques ou autres services financiers ou subit une exposition limitée et n'est pas familiarisée avec les services financiers. Leurs besoins et leurs modes de dépenses sont variés et divers. Les données permettent aux prestataires de SFN de créer des produits et des services qui reflètent mieux les préférences et les aspirations des clients. Les SFN ont changé l'accès et le caractère abordable des services financiers dans les marchés émergents en répondant aux besoins des clients à faible revenu, améliorant ainsi l'inclusion financière.

Les données permettent d'améliorer l'inclusion financière. Cependant, cela doit être fait en veillant à ce que la protection du consommateur et la confidentialité des données ne soient pas compromises. Des données sont produites et collectées

passivement au moyen d'appareils numériques tels que des téléphones portables et les ordinateurs, entre autres. De nombreuses parties prenantes se sont dites préoccupées par le fait que les ménages à faible revenu, les producteurs primaires de ces données dans le contexte de l'inclusion financière, peuvent ne pas être conscients que ces données sont recueillies, analysées et monétisées. En l'absence de politique uniforme, des normes différentes sont appliquées selon les types de prestataires et il existe certains cas où les droits des consommateurs ont été violés. Avec la prolifération de l'analyse des données, il est essentiel que toutes les parties prenantes - prestataires de SFN, régulateurs, décideurs, institutions de financement du développement et investisseurs - discutent des problèmes liés à la confidentialité des données et à la protection des consommateurs afin de trouver des solutions. Certains praticiens peuvent se sentir obligés d'adopter une nouvelle technologie ou de nouvelles méthodologies pour suivre les tendances qui prévalent ou en raison des mesures prises par leurs concurrents. Inutile de dire que de tels efforts pourraient être invalidés si l'organisation ne dispose pas des compétences techniques nécessaires pour gérer le projet ou n'a pas la capacité d'agir en fonction des indications trouvées. Ainsi, les praticiens doivent identifier les problèmes commerciaux qu'ils essaient de résoudre, évaluer les données et les capacités analytiques dont ils disposent actuellement, puis prendre des décisions quant à la façon de mettre en œuvre un projet de données. L'objectif commercial doit être au cœur de tout projet de gestion de données.



# Glossaire

| Terme | Explication |
|---|---|
| **Accord de niveau de service (SLA)** | Un SLA est la composante du contrat de service entre un prestataire de services et un client. Les SLA fournissent des aspects spécifiques et mesurables liés aux offres de services. Par exemple, les SLA sont souvent inclus dans des accords signés entre les prestataires de services Internet et les clients. Un SLA est également appelé accord au niveau opérationnel (OLA) lorsqu'il est utilisé dans une organisation sans relation prestataire-client établie ou formelle. |
| **Agent** | Une personne ou une entreprise sous contrat chargée de traiter les opérations pour les utilisateurs. Les plus importantes d'entre elles sont les retraits et les dépôts (c'est-à-dire, le chargement de valeur dans le système d'argent mobile, puis sa conversion inverse lors de sa sortie). Dans de nombreux cas, les agents s'occupent également de l'inscription de nouveaux clients. Les agents gagnent généralement des commissions pour la prestation de ces services. Ils fournissent également souvent un service client de première ligne, tel que la formation des nouveaux utilisateurs à la manière d'effectuer des opérations sur leur téléphone. En règle générale, les agents ont d'autres types d'activité, en plus de l'argent mobile. Les agents sont parfois limités par la réglementation, mais les petits commerçants, les institutions de microfinance, les chaines de magasins et les agences de banques servent d'agents sur certains marchés. Certains participants du secteur préfèrent les termes « commerçant » ou « détaillant » pour éviter certaines connotations juridiques du terme « agent » tel qu'il est utilisé dans d'autres secteurs. (GSMA, 2014). |
| **Agent maitre** | Une personne ou une entreprise qui achète de la monnaie électronique à un gros prestataire de SFN et la revend ensuite aux agents, qui la vendent à leur tour aux utilisateurs. Contrairement à un super agent, les agents maitres sont responsables de la gestion de la trésorerie et des exigences en liquidité de valeur électronique d'un groupe particulier d'agents. |
| **Algorithme** | En mathématiques et informatique, un algorithme est une séquence autonome d'actions à réaliser. Les algorithmes effectuent des calculs, traitent des données ou effectuent des tâches de raisonnement automatisé. |
| **Analyse de données** | L'analyse de données fait référence à des techniques et processus qualitatifs et quantitatifs utilisés pour produire de l'information, améliorer la productivité et générer des revenus pour l'entreprise. Les données sont extraites et classées pour identifier et analyser les données et les modèles comportementaux, ainsi que les exigences de l'organisation. |
| **Analyse de la fouille de textes** | La fouille de textes, aussi appelée exploration de données textuelles et à peu près équivalente à l'analyse de texte, est le processus d'obtention d'informations de grande qualité à partir du texte. Des informations de grande qualité sont généralement obtenues par la conception de modèles et de tendances par des moyens tels que l'apprentissage de formes statistiques. La fouille de textes implique généralement de structurer le texte d'entrée, (généralement faire l'analyse grammaticale, parallèlement à l'ajout de certaines caractéristiques linguistiques dérivées et la suppression d'autres, et l'insertion ultérieure dans une BD), de dériver des modèles au sein des données structurées, et d'évaluer et d'interpréter le résultat de sortie. |
| **Analyse des réseaux sociaux (ARS)** | L'analyse des réseaux sociaux, ou ARS, est un processus d'enquête sur les structures sociales grâce à l'utilisation des théories des réseaux et des graphes. Elle définit les structures en réseau en termes de nœuds (chaque acteur, personne ou chose au sein du réseau) et de liaisons ou liens (relations ou interactions) qui les connectent entre eux. |
| **Analyse, méthodologies descriptives** | Les méthodologies analytiques les moins complexes sont de nature descriptive ; elles fournissent des descriptions historiques de la performance institutionnelle, des analyses sur les raisons de ces performances et des informations sur les performances institutionnelles actuelles. Les techniques comprennent des alertes, requêtes, recherches, rapports, visualisations, tableaux de bord, tableaux, graphiques, récits, corrélations, ainsi que des analyses statistiques simples. |
| **Analyse, méthodologies prédictives** | Les analyses prédictives fournissent une analyse beaucoup plus complexe des données existantes afin de faire une prévision. Les techniques comprennent l'analyse de régression, les statistiques à plusieurs variables, le filtrage par motif, l'exploration de données, la modélisation prédictive et la prévision. |
| **Analyse, méthodologies prescriptives** | L'analyse normative va plus loin que les autres types d'analyses - elle fournit des informations pour orienter les décisions optimales pour un ensemble de résultats futurs prévus. Les techniques comprennent l'analyse graphique, les réseaux de neurones, l'apprentissage automatique et profond. |

| | |
|---|---|
| **Antécédents en matière de crédit** | Les antécédents en matière de crédit sont les données enregistrées concernant le remboursement des dettes d'un emprunteur ; un remboursement responsable est interprété comme un antécédent de crédit favorable, alors que la situation de prêt non remboursé ou les défaillances sont des facteurs qui créent un antécédent négatif en matière de crédit. Un rapport de crédit est un dossier sur les antécédents en matière de crédit de l'emprunteur provenant d'un certain nombre de sources, notamment de façon classique les banques, les sociétés de cartes de crédit, les agences de recouvrement et les gouvernements. |
| **Apprentissage automatique** | L'apprentissage automatique est un type d'IA qui fournit aux ordinateurs la possibilité d'apprendre sans être explicitement programmés. L'apprentissage automatique s'axe sur le développement de programmes informatiques qui peuvent changer lorsqu'ils sont exposés à de nouvelles données. |
| **Apprentissage non supervisé** | L'apprentissage non supervisé est une méthode utilisée pour permettre aux machines de classer des objets tangibles et intangibles sans fournir aux machines d'informations préalables sur les objets. Les choses que les machines doivent classer sont diverses, telles que les habitudes d'achat des clients, les comportements de virus, ou les attaques de pirates informatiques. L'idée principale derrière l'apprentissage non supervisé est d'exposer les machines à de grands volumes de données diverses et leur permettre d'apprendre et de déduire à partir des données. Toutefois, les machines doivent d'abord être programmées pour apprendre à partir des données. |
| **Apprentissage supervisé** | L'apprentissage supervisé est une méthode utilisée pour permettre à des machines de classer des objets, des problèmes ou des situations en fonction de données connexes introduites dans les machines. Les machines sont alimentées en données telles que les caractéristiques, les modèles, les dimensions, la couleur et la hauteur des objets, des personnes ou des situations de manière répétitive, jusqu'à ce que les machines soient en mesure d'effectuer des classifications précises. L'apprentissage supervisé est une technologie ou un concept populaire qui est appliqué à des scénarios concrets. L'apprentissage supervisé est utilisé pour fournir des recommandations de produits, segmenter les clients en fonction des données des clients, diagnostiquer une maladie en fonction de symptômes antérieurs, et effectuer bon nombre d'autres tâches. |
| **Architecture de données** | L'architecture de données est un ensemble de règles, de politiques, de normes et de modèles qui régissent et définissent le type de données recueillies et la façon dont elles sont utilisées, stockées, gérées et intégrées au sein d'une organisation et de ses systèmes de BD. Elle offre une approche formelle de création et de gestion des flux de données et de la façon dont elles sont traitées dans tous les systèmes informatiques et applications d'une organisation. |
| **Association du Système mondial de communications mobiles (GSMA)** | L'Association GSM (communément appelée « la GSMA ») est un organisme professionnel qui représente les intérêts des opérateurs de téléphonie mobile dans le monde entier. Environ 800 opérateurs de téléphonie mobile sont membres à part entière de la GSMA et 300 autres sociétés dans l'écosystème mobile plus général sont membres associés. |
| **Canal** | Le point d'accès du client à un PSF, c'est-à-dire avec qui ou avec quoi le client interagit pour accéder à un service ou à un produit financier. |
| **Canal de distribution alternatif** | Canaux qui étendent la portée des services financiers au-delà de l'agence traditionnelle. Ceux-ci comprennent les GAB, les services bancaires par Internet, certaines cartes, les services opérés liés aux appareils au PDV, les services bancaires mobiles, les portefeuilles électroniques et les services-conseils. |
| **Capture de données d'écran** | Il s'agit d'une technique selon laquelle un programme informatique extrait des données provenant d'une sortie lisible par un humain provenant d'une autre source numérique telle qu'un site Web, des rapports ou des écrans d'ordinateur. |
| **Centre d'appels** | Un bureau centralisé utilisé dans le but de recevoir ou de transmettre un grand nombre de demandes d'informations par téléphone. Dans ce contexte, en plus de gérer les plaintes et les requêtes des clients, il peut également être utilisé comme canal de distribution alternatif (CDA) pour améliorer la diffusion et attirer de nouveaux clients par le biais de diverses campagnes promotionnelles. |
| **Commerçant** | Une personne ou une entreprise qui fournit des biens ou des services à un client en échange d'un paiement. |
| **Complexité** | La combinaison des quatre grands attributs de données (volume, vitesse, variété et véracité) exige des processus analytiques évolués. Divers processus analytiques évolués sont apparus pour traiter ces grands ensembles de données. Les processus d'analyse ciblent des types de données spécifiques tels que le texte, l'audio, le web et les réseaux sociaux. Une autre méthodologie qui a reçu une grande attention est l'apprentissage automatique, par lequel un algorithme est créé et entré dans un ordinateur avec des données historiques. Cela permet à l'algorithme de prédire des relations entre des variables apparemment sans rapport entre elles. |
| **Compte actif** | Un compte qui est actif a été utilisé pour au moins une transaction dans la période précédente, généralement de 30 ou 90 jours. Cela n'inclut pas les transactions non financières telles que la modification d'un code PIN. |

| | |
|---|---|
| **Confidentialité des données** | La confidentialité des données, aussi appelée confidentialité de l'information, est l'aspect de l'informatique qui traite de la capacité que possède une organisation ou un individu à déterminer quelles sont les données d'un système informatique qui peuvent être partagées avec des tiers. |
| **Cube de données** | En informatique, il s'agit de données multidimensionnelles, souvent avec le temps comme troisième dimension de colonnes et de lignes. Dans les opérations commerciales, il s'agit un terme générique qui fait référence aux systèmes d'entreprise qui permettent aux utilisateurs de spécifier et de télécharger des rapports de données brutes. Beaucoup incluent des champs de type glisser-déposer pour concevoir une demande d'information ou des agrégations de données simples. |
| **De personne à personne (P2P)** | Transfert de fonds de personne à personne. |
| **Distribution statistique** | La distribution d'une variable est une description du nombre relatif de fois où chaque résultat possible se produira pour un certain nombre d'essais. |
| **Données** | Les données sont un terme générique utilisé pour décrire toute information, fait ou statistique qui a été recueilli pour tout type d'analyse ou à des fins de référence. Il existe de nombreux types de données provenant de nombreuses sources différentes. Les données sont généralement traitées, agrégées, manipulées ou regroupées pour produire des informations qui ont un sens. |
| **Données alternatives** | Données non financières provenant des ORM, des réseaux sociaux et de leurs BD transactionnelles. L'accès à d'autres données alternatives telles que l'historique des paiements et les factures de services collectifs peut également permettre la création de notations de crédit pour les clients qui peuvent être sinon hors d'atteinte du service. |
| **Données de services supplémentaires non structurées (USSD)** | Un protocole utilisé par les appareils mobiles GSM pour communiquer avec les ordinateurs ou le réseau du prestataire de services. Ce canal est pris en charge par tous les combinés GSM et permet une session interactive composée d'un échange de messages dans les deux sens selon un menu d'application défini. |
| **Données géo spatiales** | Informations sur un objet physique qui peuvent être représentées par des valeurs numériques dans un système de coordonnées géographiques. |
| **Données non structurées** | Fait généralement référence à des informations qui ne résident pas dans une BD traditionnelle ligne-colonne. Les fichiers de données non structurées incluent souvent du contenu textuel et multimédia. En voici quelques exemples : messages e-mails, documents de traitement de texte, vidéos, photos, fichiers audio, présentations, pages Web et de nombreux autres types de documents d'entreprise. |
| **Données ouvertes** | Les données ouvertes sont des données auxquelles tout le monde peut accéder, que tout le monde peut utiliser ou partager. |
| **Données périphériques** | Habituellement, les sources de données périphériques les plus utiles sont les données de centre d'appels, les données provenant des GRC (systèmes de gestion des incidents), les informations de la base de connaissances des foires aux questions, des e-mails d'approbation, les programmes d'identification de liste noire et liste blanche, ou des programmes d'identification Excel partagés. |
| **Données qualitatives** | Données qui font des approximations ou caractérisent, mais ne mesurent pas les attributs, caractéristiques ou propriétés d'une chose ou d'un phénomène. Les données qualitatives décrivent, alors que les données quantitatives définissent. |
| **Données quantitatives** | Données qui peuvent être quantifiées et vérifiées, et qui se prêtent à la manipulation statistique. Les données qualitatives décrivent, alors que les données quantitatives définissent. |
| **Données semi-structurées** | Les données semi-structurées sont une forme de données structurées qui ne sont pas conformes à la structure formelle des modèles de données associées à des BD relationnelles ou d'autres formes de tableaux de données. Elles contiennent néanmoins des balises ou d'autres marqueurs pour séparer les éléments sémantiques et appliquer des hiérarchies d'enregistrements et de champs dans les données. |
| **Données structurées** | Les données structurées font référence à toute donnée qui se trouve dans un champ fixe dans un enregistrement ou fichier. Cela inclut les données contenues dans les BD relationnelles. |
| **Données traditionnelles** | Les données traditionnelles se réfèrent aux données internes structurées couramment utilisées (telles que les données transactionnelles) et les données externes (telles que les informations provenant des bureaux de crédits ) qui sont utilisées dans le processus de prise de décision. Elles peuvent inclure des données qui sont générées à partir d'interactions avec des clients tels que des enquêtes, des formulaires d'inscription, le salaire, et des informations démographiques. |

| | |
|---|---|
| **Écart type** | En statistique, l'écart type est une mesure qui est utilisée pour quantifier la quantité de variation ou de dispersion d'un ensemble de valeurs de données. Un écart type faible indique que les points de données ont tendance à être proches de la moyenne de l'ensemble, tandis qu'un écart type élevé indique que les points de données sont répartis sur une série plus large de valeurs. |
| **Entrepôt de données** | Une série d'informations et de données sur une entreprise provenant des systèmes opérationnels et de sources de données externes. Un entrepôt de données est conçu pour appuyer les décisions commerciales en permettant la consolidation et l'analyse de données, ainsi que l'établissement de rapports sur les données à différents niveaux d'agrégation. |
| **Essai randomisé contrôlé (ERC)** | Un essai randomisé contrôlé est une expérience scientifique où les personnes participant à l'essai sont attribuées au hasard à différents contextes d'intervention puis sont comparées par rapport aux autres. La randomisation minimise le biais de sélection lors de la conception de l'expérience scientifique. Les groupes de comparaison permettent aux chercheurs de déterminer les effets de l'intervention par rapport au groupe (de contrôle) sans intervention, tandis que d'autres variables sont maintenues constantes. |
| **Exaoctet (Eo)** | L'exaoctet (Eo) est un multiple de l'unité octet utilisé en information numérique. Selon le Système international d'unités, le préfixe exa indique la multiplication par 1000 (1018) puissance 6. Par conséquent, un Eo est un quintillion d'octets (échelle courte). Le symbole de l'Exaoctet est Eo. |
| **Exploration de données** | L'exploration de données est le processus de calcul de découverte de modèles dans de grands ensembles de données. Il s'agit d'un sous-domaine interdisciplinaire de l'informatique. L'objectif global du processus d'extraction de données est d'extraire des informations à partir d'un ensemble de données et de les transformer en une structure compréhensible pour une utilisation ultérieure. |
| **Fonds de caisse (fonds de caisse d'un agent)** | Le solde de monnaie électronique, ou d'espèces physiques ou d'argent sur un compte bancaire auquel un agent peut immédiatement accéder pour répondre aux demandes des clients désirant acheter (encaisser) ou vendre (décaisser) de la monnaie électronique. |
| **Gestion de données** | La gestion de données est le développement, l'exécution et la supervision de plans, politiques, programmes et pratiques qui contrôlent, protègent, livrent et améliorent la valeur des données et des actifs d'informations. |
| **Historique détaillé des appels (CDR)** | Il s'agit des données enregistrées par un ORM concernant un appel vocal ou un SMS, avec des détails tels que l'origine, la destination, la durée, l'heure, ou le montant facturé pour chaque appel ou SMS. |
| **Hypothèse** | Une hypothèse est une prévision fondée sur des connaissances qui peut être testée. |
| **Indicateur clé de performance (ICP)** | Un ICP est une valeur mesurable qui montre l'efficacité d'une entreprise pour atteindre des objectifs commerciaux clés. Les organisations utilisent des ICP à plusieurs niveaux pour évaluer leur capacité à atteindre les cibles. Les ICP de haut niveau peuvent s'axer sur la performance globale de l'entreprise, tandis que les ICP de faible niveau peuvent s'axer sur des processus dans des services tels que les ventes, le marketing ou un centre d'appels. |
| **Indicateur clé de risque (ICR)** | Un ICR est une mesure utilisée pour indiquer à quel degré une activité est risquée. La différence avec un ICP est que ce dernier est conçu comme une mesure de la qualité avec laquelle quelque chose est fait, alors que le premier indique à quel point quelque chose peut être dommageable si cette chose se produit et quelle est la probabilité qu'elle se produise. |
| **Institution de Microfinance (IMF)** | Une IF spécialisée dans les services bancaires pour les groupes, petites entreprises ou personnes à faible revenu. |
| **Institution financière (IF)** | Un prestataire de services financiers, notamment les coopératives de crédit, les banques, les institutions financières non bancaires, les institutions de microfinance et les PFS mobiles. |
| **Intelligence artificielle (IA)** | L'IA est un domaine de l'informatique qui met l'accent sur la création de machines intelligentes qui fonctionnent et réagissent comme des humains. |
| **Interface de programme d'application (API)** | Une méthode de spécification d'un composant logiciel concernant ses opérations par un accent mis sur un ensemble de fonctionnalités qui sont indépendantes de leur mise en œuvre respective. Les API sont utilisées pour une intégration en temps réel au CBS ou au système d'information de gestion (SIG), qui spécifie la manière dont deux systèmes différents peuvent communiquer entre eux par l'échange de « messages ». Il existe différents types d'API, notamment celles sur le Web, la communication par Protocole de contrôle de transmission (TCP), l'intégration directe à une BD, ou des API propriétaires écrites pour des systèmes spécifiques. |
| **Lac de données** | Un lac de données est un dépôt massif, facilement accessible et centralisé de grands volumes de données structurées et non-structurées. |

| | |
|---|---|
| **Lutte contre le blanchiment de capitaux et Lutte contre le financement du terrorisme (LBC/LFT)** | Les LBC/LFT sont des contrôles juridiques appliqués au secteur financier pour aider à prévenir, détecter et signaler les activités de blanchiment d'argent. Les contrôles de LBC/LFT comprennent les montants maximaux qui peuvent être détenus sur un compte ou transférés entre des comptes pour toute transaction, ou pour tout jour donné. Ils comprennent également les informations financières obligatoires de la KYC pour toutes les transactions supérieures à 10 000 USD, notamment la déclaration de la source des fonds, ainsi que la raison du virement. |
| **Machines à vecteurs de support (MVS)** | Une machine à vecteur de support, ou MVS, est un algorithme d'apprentissage automatique qui analyse les données pour les classer et opérer une analyse de régression. Une MVS est une méthode d'apprentissage supervisé qui examine les données et les trie selon l'une de deux catégories. Une MVS produit en sortie une carte des données triées avec les marges entre les deux catégories les plus éloignées possible. Les MVS sont utilisées dans la catégorisation de textes, le classement d'image, la reconnaissance de l'écriture manuscrite et en sciences. Une machine à vecteur de support est également appelée réseau à vecteurs de support (RVS). |
| **Mégadonnées** | Les mégadonnées sont de grands ensembles de données, dont la taille est mesurée par cinq caractéristiques distinctes : volume, vitesse, variété, véracité et complexité. |
| **Métadonnées** | Les métadonnées décrivent d'autres données. Elles fournissent des informations sur le contenu d'un élément donné. Par exemple, une image peut inclure des métadonnées qui décrivent la taille de l'image, sa profondeur des couleurs, la résolution d'image, le moment où l'image a été créée et autres données. |
| **Méthode scientifique** | Résolution des problèmes en utilisant une approche étape par étape consistant en (1) l'identification et la définition d'un problème, (2) l'accumulation de données pertinentes, (3) la formulation d'une hypothèse, (4) la conduite d'expériences pour tester l'hypothèse, (5) l'interprétation des résultats de manière objective, et (6) la répétition des étapes jusqu'à ce qu'une solution acceptable soit trouvée. |
| **Méthodes de Monte Carlo** | Modèles qui utilisent des approches aléatoires pour modéliser des systèmes complexes en définissant une pondération probabiliste à divers points de décision dans le modèle. Les résultats montrent un modèle de distribution statistique qui peut être utilisé pour prédire la probabilité de certains résultats, compte tenu des entrées dans le système modélisé. Ces modèles sont habituellement utilisés pour des problèmes d'optimisation ou des analyses de probabilités. |
| **Méthodologie non paramétrique** | Une méthode couramment utilisée en statistiques où de petites tailles d'échantillon sont utilisées pour analyser des données nominales. Une méthode non paramétrique est utilisée lorsque le chercheur ne sait rien des paramètres de l'échantillon tiré de la population. |
| **Modèle de notation psychométrique** | La psychométrie fait référence à la mesure des connaissances, capacités, attitudes et traits de personnalité. Dans les modèles de notation psychométriques, les principes psychométriques sont appliqués à la notation de risque de crédit en utilisant des techniques statistiques évoluées pour prévoir la probabilité de défaut d'un demandeur. |
| **Modélisation prédictive** | La modélisation prédictive est un processus qui utilise l'exploration de données et les probabilités pour prévoir des résultats. Chaque modèle est composé d'un certain nombre de prédicteurs, qui sont des variables susceptibles d'influer sur les résultats futurs. Une fois que les données ont été recueillies pour les prédicteurs pertinents, un modèle statistique est formulé. |
| **Monnaie électronique** | La « monnaie électronique » est la valeur stockée détenue sur des cartes ou des comptes tels que les portefeuilles électroniques. En règle générale, la valeur totale de la monnaie électronique émise correspond à des fonds détenus sur un ou plusieurs comptes bancaires. Elle est généralement déposée en fiducie, de sorte que même si le prestataire du service de portefeuille électronique s'avérait défaillant, les utilisateurs peuvent récupérer la valeur totale stockée sur leurs comptes. |
| **Moyenne** | Une moyenne est la somme d'une liste de chiffres divisée par le nombre de chiffres de la liste. En mathématiques et statistiques, on l'appellerait la moyenne arithmétique. |
| **Nettoyage de données** | Le nettoyage de données est le processus de modification des données dans une ressource de stockage donnée afin de s'assurer qu'elles sont précises et correctes. |
| **Notation du risque de crédit** | L'analyse statistique réalisée par les prêteurs et les IF pour accéder à la solvabilité d'une personne. Les prêteurs utilisent la notation de risque de crédit, entre autres, pour prendre une décision quant à l'octroi d'un crédit. La notation de risque de crédit d'une personne est un nombre compris entre 300 et 850, 850 étant la meilleure notation de risque de crédit possible. |
| **Obligation de s'informer sur le client (KYC)** | Les règles relatives à la LBC/LFT qui obligent les prestataires à effectuer des procédures pour identifier un client et qui évaluent la valeur des informations pour la détection, la surveillance et le signalement d'activités suspectes. |
| **Octet** | Il s'agit d'une unité d'information numérique, considérée comme une unité de taille de mémoire. Il se compose de 8 bits, et 1024 octets est égal à 1 kilooctet. |

| | |
|---|---|
| **Opérateur de réseau mobile (ORM)** | Une entreprise qui dispose d'une licence délivrée par un gouvernement pour fournir des services de télécommunications par le biais d'appareils mobiles. |
| **Petites et moyennes entreprises (PME)** | Les petites et moyennes entreprises, ou PME, sont des entreprises non-filiales indépendantes qui emploient moins d'un certain nombre d'employés. Ce nombre est variable selon les pays. |
| **Point de vente (PDV)** | Appareil électronique utilisé pour le traitement des paiements par carte à l'endroit où un client effectue un paiement au commerçant en échange de biens et services. Le terminal de PDV est un appareil matériel (fixe ou mobile) sur lequel s'exécute un logiciel destiné à faciliter la transaction. À l'origine, ces appareils étaient des appareils personnalisés ou des ordinateurs personnels, mais, de façon de plus en plus courante, ce sont des téléphones mobiles, des smartphones et des tablettes. |
| **Portefeuille électronique** | Un compte de monnaie électronique appartenant à un client de SFN et accessible par téléphone portable. |
| **Probabilité** | La probabilité est la mesure de la chance qu'un événement se produise. Une probabilité est quantifiée en un nombre compris entre zéro et un (où « 0 » indique l'impossibilité et « 1 » indique la certitude). Plus la probabilité d'un événement est élevée, plus il est certain que l'événement aura lieu. |
| **Protocole de transfert de fichiers (FTP)** | Le Protocole de transfert de fichiers (FTP) est un protocole client-serveur utilisé pour transférer des fichiers vers un ordinateur hôte ou échanger des fichiers avec un ordinateur hôte. Le FTP est la norme Internet pour le déplacement ou le transfert de fichiers d'un ordinateur à un autre en utilisant les réseaux TCP ou IP. |
| **Revenu moyen par utilisateur (ARPU)** | L'ARPU est une mesure utilisée principalement par les MNO, définie comme la recette totale divisée par le nombre d'abonnés. |
| **Recherche primaire et secondaire** | La recherche primaire porte sur des données originales recueillies selon sa propre approche, souvent une étude ou une enquête. La recherche secondaire utilise les résultats existants d'études et de collecte de données réalisées antérieurement. |
| **Reconnaissance de formes** | En informatique, la reconnaissance de formes est une branche de l'apprentissage automatique qui met l'accent sur la reconnaissance de modèles de données ou de régularités dans les données pour un scénario donné. C'est une sous-division de l'apprentissage automatique et elle ne doit pas être confondue avec une véritable étude d'apprentissage automatique. La reconnaissance de formes peut être soit « supervisée », lorsque l'on trouve des formes déjà connues dans certaines données, ou « non supervisée », lorsque sont découvertes des formes entièrement nouvelles. |
| **Régression linéaire** | Technique mathématique pour trouver la ligne droite qui correspond le mieux aux valeurs d'une fonction linéaire, tracée sur un graphique en nuage de points de données. |
| **Scientifique des données** | Un scientifique des données est une personne, une organisation ou une équipe qui exécute des processus d'analyse statistique, d'exploration et de récupération de données sur une grande quantité de données afin d'identifier des tendances, des chiffres et d'autres informations pertinentes. |
| **Sécurité des données** | La sécurité des données fait référence à des mesures de confidentialité numérique qui sont appliquées pour empêcher un accès non autorisé aux ordinateurs, BD, sites Web et tout autre endroit où les données sont stockées. La sécurité des données protège également les données contre la corruption. La sécurité des données est un aspect essentiel de l'informatique pour les organisations de toute taille et de tout type. |
| **Segmentation du marché** | Le processus de définition et de subdivision d'un grand marché homogène en segments clairement identifiables ayant des besoins, désirs ou caractéristiques de demande similaires. Son objectif est de concevoir un marketing mix qui correspond précisément aux attentes des clients sur le segment ciblé. |
| **Segmentation psychographique** | La segmentation psychographique consiste à diviser le marché en segments fondés sur différents traits de personnalité, valeurs, attitudes, intérêts et modes de vie de consommateur. |
| **Service d'argent mobile, Service financier mobile** | Un SFN qui est fourni par l'émission de comptes virtuels, correspondant à un seul compte bancaire commun, sous forme de portefeuilles électroniques, qui sont accessibles à l'aide d'un téléphone portable. La plupart des prestataires d'argent mobile sont des ORM ou des PSP. |
| **Service de messages courts (SMS)** | Un canal de communication « enregistrement et retransmission » qui implique l'utilisation du réseau de télécommunication et le protocole de message court de pair à pair (SMPP) pour envoyer une quantité limitée de texte d'un téléphone à un autre, ou entre téléphones et serveurs. |
| **Services bancaires électroniques** | La fourniture de produits et de services bancaires par le biais de canaux de distribution numériques. |
| **Services bancaires mobiles** | L'utilisation d'un téléphone portable pour accéder à des services conventionnels. Cela couvre les services opérationnels et non opérationnels, tels que l'affichage d'informations et l'exécution de transactions financières. Parfois appelés « m-banking ». |

| | |
|---|---|
| **Services financiers numériques (SFN)** | Utilisation des moyens numériques pour offrir des services financiers. Les SFN englobent tous les téléphones mobiles, cartes, PDV et les offres de commerce électronique, notamment les services fournis aux clients par l'intermédiaire des réseaux d'agents. |
| **Statistiques paramétriques** | Les statistiques paramétriques sont une branche des statistiques qui suppose que les données d'un échantillon proviennent d'une population qui suit une distribution de probabilité fondée sur un ensemble fixe de paramètres. La plupart des méthodes statistiques élémentaires bien connues sont paramétriques. |
| **Stockage de données** | Le stockage de données est un terme général désignant l'archivage des données, sous des formes électromagnétiques ou autres, destinées à être utilisées par un ordinateur ou un appareil. Différents types de stockage de données jouent des rôles différents dans un environnement informatique. En plus des formes de stockage du matériel de données, il existe maintenant de nouvelles options de stockage de données à distance, telles que le Cloud computing, qui peut révolutionner les façons dont les utilisateurs accèdent aux données. |
| **Super Agent** | Une entreprise, parfois une banque, qui achète de la monnaie électronique en gros à un prestataire de SFN, puis la revend ensuite aux agents, qui à leur tour la vendent aux utilisateurs. |
| **Tableau de bord** | Un tableau de bord de veille économique est un outil de visualisation de données qui affiche l'état actuel de paramètres et d'ICP pour une entreprise. Les tableaux de bord consolident et organisent des chiffres, des indicateurs et parfois des fiches d'évaluation sur un seul écran. |
| **Test A/B** | Le test A/B est une méthode permettant de vérifier deux versions différentes d'un produit ou d'un service afin d'évaluer comment un léger changement dans les attributs d'un produit peut avoir un impact sur le comportement des clients. Ce type d'expérimentation permet aux prestataires de SFN de choisir plusieurs variantes d'un produit ou service, de tester statistiquement le résultat en termes d'intérêt suscité auprès des clients et de comparer les résultats entre les groupes cibles. |
| **Traitement des données** | Le traitement des données est, en général, la collecte et la manipulation d'éléments de données pour produire des informations significatives. En ce sens, il peut être considéré comme un sous-ensemble du traitement de l'information, ou le changement (traitement) de l'information d'une manière quelconque et détectable par un observateur. |
| **Traitement des images** | Le traitement des images est un terme assez général qui fait référence à l'utilisation d'outils d'analyse pour traiter ou améliorer des images. De nombreuses définitions de ce terme spécifient des opérations mathématiques ou des algorithmes comme outils pour le traitement d'une image. |
| **Traitement du langage naturel (TLN)** | Le champ d'étude qui s'axe sur les interactions entre le langage humain et les ordinateurs est appelé Traitement du langage naturel, ou TLN en abrégé. Il se trouve au croisement de l'informatique, de l'IA et de la linguistique informatique. La TLN est un domaine qui couvre la compréhension et la manipulation du langage humain par un ordinateur. |
| **Type de téléphone portable - smartphone** | Un téléphone portable qui a la capacité de traitement pour exécuter la plupart des fonctions d'un ordinateur, doté généralement d'un écran relativement grand et d'un système d'exploitation capable d'exécuter un ensemble complexe d'applications, avec accès à Internet. En plus du service vocal numérique, les smartphones modernes permettent la messagerie textuelle, l'email, la navigation sur le Web, l'utilisation d'appareil photo et de caméra, un lecteur MP3, la lecture de vidéo et des capacités intégrées de transfert de données et de GPS. |
| **Type de téléphone portable - Téléphone à fonctionnalités** | Un téléphone à fonctionnalités est un type de téléphone portable qui a plus de fonctionnalités qu'un téléphone portable de base, mais qui n'est pas équivalent à un smartphone. Les téléphones à fonctionnalités peuvent fournir quelques-unes des fonctionnalités évoluées qu'on trouve sur un smartphone tel qu'un lecteur multimédia portable, un appareil photo numérique, un agenda personnel et l'accès à Internet, mais ne prend habituellement pas en charge d'applications supplémentaires. |
| **Type de téléphone portable - Téléphone de base** | Un téléphone portable de base qui peut envoyer et recevoir des appels, envoyer des messages texte et accéder au canal USSD, mais qui a des fonctionnalités supplémentaires très limitées. |
| **Variété** | L'ère du numérique a diversifié les types de données disponibles. Les données traditionnelles structurées correspondent bien à des BD existantes qui sont destinées à des informations bien définies suivant un ensemble de règles. Par exemple, une transaction bancaire a un horodatage, des montants et un emplacement. Cependant, aujourd'hui, 90 pour cent des données qui sont générées sont « non structurées », ce qui signifie qu'elles se présentent sous la forme de tweets, d'images, de documents, de fichiers audio, d'historiques d'achat des clients et de vidéos. |

# Biographie des auteurs

## DEAN CAIRE

*Spécialiste de la notation de risque de crédit, IFC*

Dean a travaillé au cours des 15 dernières années comme consultant en notation de risque de crédit, 12 ans avec la société DAI Europe, puis comme consultant indépendant. Au cours de cette période, il a aidé des clients de 77 institutions financières dans 45 pays à développer plus de 100 modèles de notation de risque de crédit personnalisés pour les segments suivants : prêts à la consommation (notamment les SFN), locations standard de biens, prêts aux microentreprises, prêts aux petites entreprises (notamment les services aux commerçants en finance numérique), prêts à l'agriculture et location de matériel (notamment sous forme de SFN), micro-prêts à des groupes de solidarité et grands prêts à des sociétés non cotées. Dean cherche à transférer les compétences de développement et de gestion de modèle à des IF homologues afin qu'elles puissent s'approprier pleinement les modèles et les gérer à l'avenir.

## LEONARDO CAMICIOTTI

*Directeur exécutif, TOP-IX Consortium*

Travaillant directement sous la supervision du conseil d'administration, Leonardo est responsable des activités stratégiques, administratives et opérationnelles de TOP-IX Consortium. Il gère le Programme de développement de TOP-IX, qui incite à la création d'entreprises en fournissant un soutien en infrastructure (c'est-à-dire bande passante Internet, Cloud computing et prototypage logiciel) aux start-up et promeut des projets d'innovation dans différents secteurs, tels que les Mégadonnées et les calculs à hautes performances, la fabrication ouverte et les technologies civiques. Il était auparavant chercheur, responsable de la stratégie et de la prospection commerciale et chef d'entreprise chez Philips Corporate Research. Il est diplômé en ingénierie électronique de l'Université de Florence et est titulaire d'un MBA de l'Université de Turin.

## SOREN HEITMANN

*Responsable des opérations, IFC*

Soren dirige le programme de recherche appliquée et de suivi, d'évaluation et d'apprentissage (SEA) intégré du partenariat IFC-Fondation MasterCard. Il travaille au cœur de la recherche et de la technologie fondée sur les données pour inciter à l'apprentissage et l'innovation dans le cadre des projets de SFN d'IFC en Afrique subsaharienne. Auparavant, Soren a dirigé le service d'évaluation des résultats pour l'Unité de vice-présidence sur les risques et l'équipe de Gestion du portefeuille de suivi et d'évaluation régional pour l'Europe et l'Asie centrale d'IFC. Il dispose d'une expérience dans la gestion des bases de données, l'ingénierie logicielle et les technologies Web, qu'il intègre désormais dans son travail en fournissant aux clients d'IFC un appui en matière de gestion de données. Soren est titulaire d'un diplôme en Anthropologie culturelle de l'Université de Boston et d'un Master en Économie du développement de la SAIS de l'Université Johns Hopkins.

## SUSIE LONIE

*Spécialiste des services financiers numériques, IFC*

Susie a passé trois ans au Kenya pour la création et l'opérationnalisation du service de paiement mobile M-PESA, puis elle a facilité son lancement sur plusieurs autres marchés, notamment l'Inde, l'Afrique du Sud et la Tanzanie. En 2010, Susie a été la co-lauréate du Prix de l'innovation de The Economist pour l'innovation sociale et économique pour son travail sur M-PESA. Elle est devenue consultante en SFN indépendante en 2011 et travaille avec des banques, des ORM et d'autres clients sur tous les aspects de la prestation de services financiers aux personnes qui n'ont pas accès aux banques ou autres services financiers sur les marchés émergents, notamment l'argent mobile, les services bancaires par agent, les transferts de fonds internationaux et l'interopérabilité. Susie travaille sur la stratégie, l'évaluation financière, la conception de produits et les exigences fonctionnelles, les opérations, la gestion des agents, l'évaluation des risques, l'évaluation de la recherche, les ventes et le marketing en matière de SFN. Elle a obtenu ses diplômes en ingénierie chimique à Edimbourg et Manchester, au Royaume-Uni.

## CHRISTIAN RACCA

*Ingénieur de conception, TOP-IX Consortium*

Christian gère le programme BIG DIVE de TOP-IX visant à offrir des formations pour les scientifiques des données, des initiatives pédagogiques fondées sur les données pour les entreprises, les organisations et les projets de conseil dans le (vaste) domaine de l'exploitation des données. Après avoir obtenu son diplôme en ingénierie des télécommunications au Politecnico di Torino, Christian a rejoint TOP-IX Consortium, en travaillant sur les flux de données continus et le Cloud computing, et plus tard sur les startups web. Il a été mentor de plusieurs projets sur le modèle économique, le développement de produit et l'architecture de l'infrastructure et a entretenu des relations avec les investisseurs, les incubateurs, les accélérateurs et l'écosystème de l'innovation en Italie et en Europe.

## MINAKSHI RAMJI

*Responsable adjointe des opérations, IFC*

Minakshi mène des projets sur les SFN et l'inclusion financière au sein du Groupe des institutions financières d'IFC en Afrique subsaharienne. Avant cela, elle était consultante à MicroSave, un cabinet de conseil sur l'inclusion financière basé en Inde, où elle était Analyste principale dans leur cabinet des Services financiers numériques. Elle a également travaillé au Centre pour la microfinance chez IFMR Trust, en Inde, qui se spécialise sur les problèmes de la politique d'accès au financement en Inde. Elle est titulaire d'un master en Développement économique de la London School of Economics et d'une licence en Mathématiques du Bryn Mawr College aux États-Unis.

## QIUYAN XU

*Directrice des scientifiques des données, Cignifi*

Qiuyan Xu est la directrice des scientifiques des données chez Cignifi Inc., et dirige l'équipe d'Analyse des mégadonnées. Cignifi est une start-up de technologie financière en pleine croissance à Boston, aux États-Unis, qui a développé la première plateforme analytique éprouvée fournissant des notations de crédit et de marketing pour les consommateurs à l'aide de données sur le comportement des utilisateurs de téléphones portables. Le Docteur Xu dispose d'une expertise dans l'analyse des Mégadonnées, le Cloud computing, la modélisation statistique, l'apprentissage automatique, l'optimisation des opérations et la gestion des risques. Elle a été directrice des analyses chez Liberty Mutual et directrice de la gestion des risques d'entreprise chez Travelers Insurance. Le Docteur Xu est titulaire d'un doctorat en statistiques de l'Université de Californie, Davis et d'un certificat de Gestionnaire des risques financiers de l'Association mondiale des professionnels du risque.

## Le Partenariat pour l'inclusion financière

Le Partenariat pour l'inclusion financière est une initiative conjointe de 37,4 millions d'USD d'IFC et de the Mastercard Foundation visant à développer la microfinance et à faire progresser les services financiers mobiles en Afrique sub-saharienne Le partenariat est également soutenu par la Fondation Bill & Melinda Gates et la Banque autrichienne de développement (OeEB, Oesterreichische Entwicklungsbank AG). Il travaille également avec des institutions de microfinance, des banques, des opérateurs de réseaux mobiles et des prestataires de service de paiement sur le continent pour tester et évaluer les modèles économiques innovants favorables à l'inclusion financière. Le programme inclut un solide volet de partage des connaissances. Ce manuel est le second d'une série de manuels sur la façon de mettre en œuvre avec succès les services financiers numériques, et l'une des nombreuses publications du Partenariat. Pour plus d'informations et pour avoir accès à tous les rapports, veuilles vous rendre sur : www.ifc.org/financialinclusionafrica

## A propos d'IFC

IFC, un membre du Groupe de la Banque mondiale, est la principale institution internationale de développement exclusivement dédiée au secteur privé sur les marchés émergents. Travaillant avec plus de 2 000 entreprises dans le monde, nous utilisons notre capital, notre expertise et notre influence pour créer des opportunités là où elles sont le plus nécessaires. Au cours de l'exercice 2015, nos investissements de long terme dans les pays en développement ont augmenté pour se situer à près de 18 milliards d'USD, aidant le secteur privé à jouer un rôle essentiel dans l'effort mondial visant à mettre fin à l'extrême pauvreté et à favoriser une prospérité partagée. Pour plus d'informations, veuillez-vous rendre sur le site www.ifc.org

## A propos de the Mastercard Foundation

The Mastercard Foundation travaille avec des organisations visionnaires pour fournir un meilleur accès à l'éducation, à la formation en compétences et aux services financiers à des individus vivant dans la pauvreté, essentiellement en Afrique. Étant l'une des fondations indépendantes les plus grandes, son travail est guidé par sa mission consistant à faire progresser les apprentissages et à promouvoir l'inclusion financière pour réduire la pauvreté. Basée à Toronto, au Canada, son indépendance a été établie par MasterCard quand la Fondation fut fondée en 2006. Pour plus d'informations ou pour vous abonner au bulletin d'information de la Fondation, veuillez-vous rendre sur www.mastercardfdn.org

Ce manuel est l'un des trois manuels sur les services financiers numériques publiés par Le Partenariat pour l'inclusion financière, une initiative conjointe d'IFC et de the Mastercard Foundation visant à promouvoir l'inclusion financière. Les deux autres manuels sont également disponibles sur demande à la SFI ou à télécharger sur le site Web du Partenariat : www.ifc.org/financialinclusionafrica :


Le **Manuel Canaux de Distribution Alternatifs et Technologies** fournit un guide pratique, étape par étape, pour la conception de canaux de distribution alternatifs liant les choix technologiques aux processus de l'entreprise.


Le **Manuel Services Financiers Numériques et Gestion des Risques** est concu pour tous types d'institutions financières offrant ou prévoyant d' offrir des services financiers numériques. Ce manuel présente un aperçu des risques connexes et comment appliquer un cadre de gestion des risques pour faire face à ces risques de façon optimale.


## COORDONNÉES

Anna Koblanck

IFC, Sub-Saharan Africa

akoblanck@ifc.org


**www.ifc.org/financialinclusionafrica**


2017

mastercard foundation

IFC International Finance Corporation WORLD BANK GROUP