Foster Provost
Tom Fawcett

Data science pour l'entreprise

Principes fondamentaux pour développer son activité

O'REILLY®

EYROLLES

Le guide de référence sur les data sciences

Cet ouvrage traite de façon détaillée mais non technique les principes fondamentaux de la data science. Tout au long d'un processus de « raisonnement orienté données », il vous guidera pour acquérir des connaissances utiles et extraire une valeur économique des données que vous collectez. L'apprentissage de la data science vous permettra de comprendre les nombreuses techniques de data mining utilisées aujourd'hui. Ces principes sous-tendent tous les processus et stratégies de data mining qui servent à résoudre des problèmes d'entreprise.

« Ce livre est bien plus qu'une introduction à l'analyse de données. C'est un guide essentiel pour ceux d'entre nous (nous tous ?) qui ont entièrement fondé leur entreprise sur l'ubiquité des données et la nécessité, aujourd'hui, de la prise de décision orientée données. » – Tom Phillips, PDG, Dstillery ; ex-Directeur de Google Search and Analytics.

« Les auteurs de ce livre, tous deux experts en data science avant même que la discipline soit nommée ainsi, présentent ici un sujet complexe en le rendant accessible à tous les niveaux. Cet ouvrage est une première du genre : il se concentre sur les concepts de la data science tels qu'ils doivent être appliqués aux problèmes concrets des entreprises. Il est rempli de captivants exemples réels qui illustrent les problèmes courants auxquels les entreprises sont confrontées : l'attrition client, le marketing ciblé, et même une analyse des données sur les whiskies ! Ce livre se distingue par le fait qu'il n'est pas un traité d'algorithmique. Les auteurs ont pour objectif d'aider le lecteur à comprendre les concepts sous-jacents de la data science, mais également et surtout ils expliquent comment aborder un problème de data science et mettre au point une solution qui marche. Si vous avez besoin d'un aperçu complet de la data science, ou si vous êtes un data scientist en herbe qui veut maîtriser les bases de la discipline, ce livre est un indispensable pour vous. » – Chris Volinsky, Directeur, Statistics Research, AT&T Labs, Gagnant du Netflix Challenge à 1 M$.

À qui cet ouvrage s'adresse-t-il ?

- Aux personnes issues du monde de l'entreprise qui envisagent de travailler avec des data scientists, de gérer des projets orientés data science ou d'investir dans des entreprises spécialisées en data science
- Aux développeurs qui mettront en œuvre des solutions de data science
- Aux data scientists en devenir

Au sommaire

Foster Provost est professeur et membre du corps enseignant à la NYU Stern School of Business, où il enseigne en Business Analytics, Data Science et dans les cursus de MBA. Ses recherches, primées très largement, sont lues et citées. Avant de rejoindre la NYU, il a travaillé pendant cinq ans comme chercheur en data science pour Verizon. Pendant la dernière décennie, le professeur Provost a cofondé avec succès plusieurs entreprises de data science.

Tow Fawcett est titulaire d'un doctorat en apprentissage machine et a travaillé dans des équipes de R&D en entreprise pendant plus de deux décennies (GTE Laboratories, NYNEX/Verizon Labs, HP Labs, etc.). Ses publications sont aujourd'hui des classiques de la littérature de data science à la fois du point de vue méthodologie (par exemple, pour l'évaluation des résultats du data mining) et du point de vue applicatif (par exemple, la détection des fraudes et le filtrage de pourriels).

www.editions-eyrolles.com

Data science pour l'entreprise

DANS LA MÊME COLLECTION

H. BEN REBAH, B. MARIAT. – **API HML 5 : maîtrisez le Web moderne !**
N°67554, 2018, 294 pages.

B. BARRÉ. – **Concevez des applications mobiles avec React Native.**
N°67563, 2018, 220 pages.

R. RIMELÉ. –**HTML 5 – Une référence pour le développeur web.**
N°14365, 3e édition, 2017, 832 pages.

P. FICHEUX. – **Linux embarqué – Mise en place et développement.**
N°67484, 2017, 220 pages.

K. NOVAK. – **Débuter avec LINUX.**
N°13793, 2017, 522 pages.

P. MARTIN, J. PAULI, C. PIERRE DE GEYER. – **PHP 7 avancé.**
N°14357, 2016, 732 pages.

L. BLOCH, C. WOLFHUGEL, A. KOKOS, G. BILLOIS, A. SOULLIÉ, T. DEBIZE. – **Sécurité informatique.**
N°11849, 2016, 648 pages.

R. GOETTER. – **CSS 3 Flexbox.**
N°14363, 2016, 152 pages.

B. PHILIBERT. – **Bootstrap 3 : le framework 100 % web design.**
N°14132, 2015, 318 pages.

C. CAMIN. – **Développer avec Symfony2.**
N°14131, 2015, 474 pages.

H. GIRAUDEL, R. GOETTER. – **CSS 3 : pratique du design web.**
N°14023, 2015, 372 pages.

SUR LE MÊME THÈME

J. GRUS. – **Data science par la pratique.**
N°11868, 2017, 308 pages.

E. BIERNAT, M. LUTZ. – **Data science : fondamentaux et études de cas.**
N°14243, 2015, 312 pages.

W. MCKINNEY. – **Analyse de données en Python.**
N°14109, 2015, 488 pages.

M.-R. AMINI. – **Apprentissage machine, de la théorie à la pratique**.
N°13800, 2015, 272 pages.

Foster Provost
Tom Fawcett

Data science pour l'entreprise

EYROLLES

ÉDITIONS EYROLLES
61, bd Saint-Germain
75240 Paris Cedex 05
www.editions-eyrolles.com

Traduction autorisée de l'ouvrage en langue anglaise intitulé
Data Science for Business
de Foster Provost, Tom Fawcett (ISBN : 9781449361327),
publié par O'Reilly Media.

Table des matières

Avant-propos

L'ouvrage *Data science pour l'entreprise* s'adresse à différents types de lecteurs :

- aux personnes issues du monde de l'entreprise qui envisagent de travailler avec des *data scientists*, de gérer des projets orientés *data science*, ou d'investir dans des entreprises spécialisées en data science ;
- aux développeurs qui mettront en œuvre des solutions de data science ;
- et aux data scientists en devenir.

Ce livre ne se focalise pas sur les algorithmes, et ne peut donc remplacer un livre spécialisé sur ce sujet. Nous avons délibérément évité l'approche axée sur les algorithmes. Il existe un ensemble relativement petit de concepts et principes fondamentaux qui sont à la base des techniques d'extraction de connaissances à partir des données. Ces concepts sont les *fondements* de nombreux algorithmes bien connus de *data mining*. En outre, ils sous-tendent l'analyse de problèmes d'entreprise orientés données, ainsi que la création et l'évaluation de solutions de data science, ainsi que l'évaluation des stratégies et propositions de data science en général. Nous avons donc centré notre ouvrage sur ces principes généraux, plutôt que sur des algorithmes spécifiques. Et lorsque la description de détails procéduraux était nécessaire, nous les avons présentés à l'aide de textes et de graphiques, qui, selon nous, sont plus compréhensibles qu'une liste détaillée des différentes étapes d'un algorithme.

La lecture de ce livre ne nécessite aucun prérequis avancé en mathématiques. Cependant, la nature même de son sujet le rend quelque peu technique, l'objectif étant de transmettre une compréhension approfondie des principes de la data science, et pas seulement d'en donner un aperçu général. Nous avons donc tenté de restreindre les explications mathématiques au maximum et privilégié une présentation la plus conceptuelle possible.

Des collègues en entreprise nous ont fait savoir que ce livre leur a été indispensable pour aider à uniformiser la compréhension du sujet à travers les équipes commerciales, de développement et de data science. Cette observation est basée sur un petit échantillon, et nous sommes curieux de voir à quel point elle est générale (voir chapitre 5 !). Dans l'idéal, notre objectif est d'écrire un livre que n'importe quel data scientist offrirait à ses collaborateurs des équipes commerciales et de développement, en leur disant : « Si vous voulez vraiment implémenter

des solutions de data science de haut niveau pour des problèmes d'entreprise, alors nous devons tous connaître ce livre. »

Des collègues nous disent aussi que le livre a révélé son utilité dans des cas inattendus : pour préparer des entretiens avec des candidats data scientists. Les besoins des entreprises en data scientists sont en forte croissance. En conséquence, de plus en plus de chercheurs d'emploi se présentent comme des data scientists. Tout candidat data scientist devrait comprendre les principes fondamentaux présentés dans ce livre. Nous avons même envisagé plus ou moins sérieusement la rédaction d'un livret complémentaire qui se serait intitulé « Remarques de Cliff pour l'entretien des data scientists ».

Notre approche conceptuelle de la data science

Nous présentons dans ce livre une sélection des principaux concepts fondamentaux de la data science. Certains de ces concepts sont le sujet même d'un chapitre, tandis que d'autres sont introduits plus naturellement tout au long des discussions (et ne sont donc pas nécessairement traités comme des concepts fondamentaux). Les concepts couvrent le processus dans son ensemble, de l'élaboration du problème, en passant par l'application de techniques de data science, jusqu'à l'exploitation des résultats pour améliorer la prise de décision. Ces mêmes concepts sont également à la base de nombreuses méthodes et techniques de traitement analytiques.

On distingue trois types de concepts :

1. Certains concepts déterminent comment la data science s'intègre dans l'organisation et dans le paysage concurrentiel. Entre autres, il y a ceux qui déterminent comment attirer, structurer et soutenir des équipes de data science, ceux qui indiquent comment penser la manière dont la data science apportera un avantage compétitif ; et des concepts tactiques pour gérer au mieux les projets orientés data science.
2. D'autres sont des méthodes générales pour penser de manière orientée données. Ces concepts aident à identifier les données et les méthodes les plus appropriées. Ils englobent à la fois le *processus de data mining* et l'ensemble des *tâches de data mining les plus complexes*.
3. Enfin, les concepts généraux, qui recouvrent l'ensemble des tâches de data science et leurs algorithmes, permettent d'extraire des connaissances à partir des données.

Un exemple de concept fondamental est celui qui permet de mesurer la similarité entre deux entités décrites par les données. Cette information permet d'accomplir diverses tâches bien spécifiques. Elle peut, par exemple, être utilisée tout simplement pour *trouver* des clients ressemblant à un client en particulier. Elle est le noyau de divers algorithmes prédictifs qui calculent une valeur cible, telle que la quantité prévisionnelle de ressources utilisées par un client ou la probabilité qu'un client réponde à une offre. Elle est aussi à la base des techniques de *partitionnement (clustering)*, qui regroupent les entités selon leurs caractéristiques communes sans objectif fixe. Le concept de similarité est par ailleurs fondamental pour l'*extraction d'information*, qui permet d'extraire des documents ou des pages web pertinentes en réponse à une requête. Enfin, il est à la base de nombreux algorithmes de *recommandation*. Dans un ouvrage classique orienté algorithmes, ces différentes tâches seraient présentées chacune dans

un chapitre, sous différents noms et avec leurs aspects communs enfouis sous des quantités de détails algorithmiques ou de propositions mathématiques. Dans ce livre, nous nous concentrons plutôt sur les concepts communs, et nous présentons les diverses tâches et algorithmes spécifiques comme des instances de ces concepts.

Comme autre exemple, dans l'évaluation de l'utilité d'un scénario, on peut aussi citer la notion de *lift* – à quel point un scénario est plus fréquent que le hasard –, une notion très récurrente en data science. Elle permet en effet d'évaluer différents types de scénarios dans différents contextes. Les algorithmes de ciblage publicitaire sont évalués en calculant le lift qu'ils obtiennent pour la population cible. La valeur de lift est ainsi utilisée pour estimer le poids d'un scénario comme indicateur pour ou contre une conclusion. Elle aide donc à déterminer si une co-occurrence (un lien) dans les données est significative, ou si elle est au contraire une simple conséquence de la popularité intrinsèque du produit.

Une présentation des data sciences qui se concentre sur de tels concepts fondamentaux peut non seulement aider le lecteur, mais aussi faciliter la communication entre les différentes parties prenantes de l'entreprise et les data scientists. Elle fournit par ailleurs un vocabulaire commun et permet donc aux différents interlocuteurs de mieux se comprendre. Par ailleurs, les concepts partagés conduisent à des discussions plus approfondies qui peuvent faire émerger des points critiques qui n'auraient pas été découverts autrement.

À l'attention de l'enseignant

Ce livre a servi avec succès comme manuel pour une grande variété de cours de data science. Du point de vue historique, ce livre est né de l'élaboration des cours multidisciplinaires de Data science par Foster à la Stern School de NYU *(New York University)*, à partir de l'automne 2005[1]. Le cours était à l'origine prévu pour les étudiants de MBA et de MSIS, mais il a attiré des étudiants d'autres départements de l'université. L'aspect le plus intéressant de ce cours n'était pas seulement qu'il plaisait aux étudiants de MBA et de MSIS, pour qui il était conçu ; il profitait aussi aux étudiants dotés de connaissances avancées en apprentissage automatique *(machine learning)* et dans d'autres disciplines techniques. L'une des raisons était que l'étude des concepts fondamentaux et autres problèmes sous-jacents des algorithmes n'était pas prévue dans leur cursus.

À la NYU, ce livre est aujourd'hui utilisé comme support de cours dans différents cursus liés aux data sciences : les cursus MBA et MSIS originaux, en licence de business analytics, dans le nouveau cursus du MS Data Science de NYU/Stern, et dans le cours d'introduction à la Data science du nouveau MS Data Science de NYU. De plus, avant même sa publication, ce livre a été adopté par plus de vingt autres universités réparties dans neuf pays (pour le moment), par des écoles de commerce, dans des cursus d'informatique et pour des introductions plus générales à la data science.

Visitez régulièrement le site web du livre (voir le lien ci-après) pour savoir comment obtenir du matériel pédagogique, dont des diapos de cours, des exemples de questions et de pro-

1. Bien sûr, chaque auteur a l'impression claire qu'il a fait la majorité du travail dans ce livre.

blèmes pour des devoirs à la maison, des exemples d'instructions de projet basés sur des règles de travail du livre, des questions d'examen, et d'autres choses à venir.

> Nous conservons sur le site du livre une liste régulièrement actualisée des universités qui l'ont adopté (http://www.data-science-for-biz.com/). Pour la consulter, cliquez sur le lien *Who's Using It* (« Qui l'utilise ») en haut de la page.

Autres compétences et concepts

Outre les principes fondamentaux des data sciences, un data scientist en exercice doit maîtriser beaucoup d'autres concepts et compétences. Ces compétences et concepts seront abordés aux chapitres 1 et 2. Nous encourageons le lecteur intéressé à consulter le site web du livre pour trouver des ressources qui lui permettront d'acquérir ces compétences et concepts supplémentaires, par exemple, la programmation en Python, l'utilisation de la ligne de commande Unix, les fichiers de données, les formats de données les plus courants, les bases de données et les requêtes, les architectures *big data* et les systèmes tels que MapReduce et Hadoop, la visualisation des données, et d'autres sujets connexes.

Sections et notations

En plus des occasionnelles notes de bas de page, vous rencontrerez parfois des paragraphes de commentaires secondaires. Il s'agit en fait de notes de bas de page plus développées. Nous y introduisons des informations que nous considérons comme utiles et intéressantes, mais qui seraient trop longues pour une note de bas de page et trop digressives par rapport au texte principal.

> PRÉCISIONS TECHNIQUES PRÉALABLES **Remarque sur les sections à astérisque**
>
> Les précisions mathématiques occasionnelles sont reléguées dans des sections optionnelles dont les titres sont préfixés par un astérisque et dont le contenu est précédé d'un paragraphe tel que celui-ci. Ces sections pour lecteurs avertis contiennent plus de détails mathématiques et/ou techniques que partout ailleurs dans le livre. Le livre est rédigé de telle manière que vous puissiez ignorer ces sections sans que cela n'affecte la continuité du texte, même si nous rappelons parfois au lecteur qu'il peut y trouver plus de détails sur certains points.

Les passages de texte tels que « (Smith et Jones, 2003) » font référence à une entrée de la bibliographie (en l'occurrence, l'article ou le livre de Smith et Jones qui date de 2003). Il en va de même pour « Smith et Jones (2003) ». Toutes les références bibliographiques du livre sont rassemblées en fin d'ouvrage.

Dans ce livre, nous tentons de limiter les développements mathématiques au minimum, et nous avons simplifié le peu de mathématiques que nous présentons sans que cela n'introduise de confusion. Voici quelques explications concernant notre choix de simplification, à l'attention des lecteurs qui possèdent des connaissances techniques avancées.

1. Nous évitons les notations Sigma (Σ) et Pi (∏), généralement utilisées dans les manuels pour représenter respectivement une somme ou un produit. Nous employons à la place des équations à ellipses telles que :

$$f(x) = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

   Dans les sections techniques, marquées par un astérisque, il nous est arrivé d'utiliser Sigma et Pi lorsque cette notation elliptique nous a semblé trop lourde. Nous supposons que les personnes qui liront ces sections sont assez familières avec les notations mathématiques et n'y trouveront donc aucune difficulté.

2. Les manuels de statistiques prennent généralement soin de distinguer une valeur réelle de son estimation en la surmontant d'un accent circonflexe. Par exemple, la valeur mesurée d'une probabilité est notée $\mathbb{P}$ tandis que sa valeur estimée est notée $\hat{p}$. Dans le présent ouvrage, nous parlons pratiquement toujours de valeurs estimées à partir des données, mettre des chapeaux partout rendrait donc nos équations verbeuses et disgracieuses. Toute information doit être considérée comme une valeur estimée à partir des données, sauf indication contraire.

3. Nous supprimons les variables superflues pour simplifier la notation lorsque nous estimons qu'elles sont compréhensibles. Par exemple, lorsque nous représentons une fonction de classification de manière mathématique, concrètement, nous manipulons des prédicats de décision qui s'appliquent à des vecteurs de caractéristiques. Voici comment on exprimerait cela formellement :

$$\hat{f}_R(\mathbf{x}) = x_{Age} \times -1 + 0{,}7 \times x_{Balance} + 60$$

   À la place, nous avons opté pour une forme plus lisible :

$$f_R(\mathbf{x}) = Age \times -1 + 0{,}7 \times Balance + 60$$

   sachant que **x** est un vecteur et que *Age* et *Balance* en sont des composantes.

Du point de vue typographique, nous avons tenté de rester cohérents, en notant les attributs ou mots-clés des données tels que `sepal_width` avec une police à chasse fixe. Par exemple, dans le chapitre sur le *text mining* (exploration de textes), un mot comme *'discussing'* désigne un mot dans un document, alors que `discuss` serait le token correspondant dans les données.

Les conventions typographiques suivantes sont utilisées dans ce livre :

- *Italique*

  Indique les termes nouveaux, les URL, les adresses e-mail, les noms et extensions de fichiers.

- `Police à chasse fixe`

  Indique les listings de code et met en valeur les éléments de code cités dans le texte, tels que les noms de variables ou de fonctions, les bases de données, les types de données, les variables d'environnement et les mots-clés.

- *`Police à chasse fixe en italique`*

  Indique les éléments de texte qui devraient être remplacés par des valeurs fournies par l'utilisateur ou par le contexte.

Nous avons enfin parsemé le livre d'astuces et d'avertissements spéciaux en rapport avec le contexte.

Utilisation de citations

En plus d'être une introduction à la data science, cet ouvrage est destiné à faciliter les discussions sur le sujet ainsi que le travail au jour le jour sur le terrain. Vous pourrez bien sûr répondre à des questions en citant des passages ou des exemples de ce livre. Nous vous demandons seulement, mais n'exigeons pas, de nous citer en référence dans ces cas-là en mentionnant le titre, l'auteur, l'éditeur et l'ISBN du présent ouvrage.

Remerciements

Merci à tous les collègues et autres personnes qui nous ont fourni des idées précieuses, des commentaires, des critiques, des suggestions et des encouragements suite à des discussions et aux nombreux manuscrits préliminaires. Au risque d'en oublier certains, permettez-nous de remercier spécialement : Panos Adamopoulos, Manuel Arriage, Josh Attenberg, Solon Barocas, Ron Bekkerman, Josh Blumenstock, Ohad Brazilay, Aaron Brick, Jessica Clark, Nitesh Chawla, Peter Devito, Vasant Dhar, Jan Ehmke, Theos Evgeniou, Justin Gapper, Tomer Geva, Daniel Gillick, Shawndra Hill, Nidhi Kathuria, Ronny Kohavi, Marios Kokkodis, Tom Lee, Philipp Marek, David Martens, Sophie Mohin, Lauren Moores, Alan Murray, Nick Nishimura, Balaji Padmanabhan, Jason Pan, Claudia Perlich, Gregory Piatetsky-Shapiro, Tom Philipps, Kevin Reilly, Maytal Saar-Tshechansky, Evan Sadler, Galit Shmueli, Roger Stein, Nick Street, Kiril Tsemekhman, Craig Vaughan, Chris Volinsky, Wally Wang, Geoff Webb, Debbie Yuster et Rong Zheng. Nous tenons également à remercier l'ensemble des étudiants des cours de Foster : Data Mining pour le Business Analytics, Data Science en Pratique, Introduction aux Data Sciences, et du séminaire Recherche en Data Science. Les questions et problèmes qu'ils ont soulevés lors de leur utilisation des précédentes versions de ce livre nous ont été d'une utilité considérable pour son amélioration.

Merci à tous les collègues qui ont fait découvrir les data science et qui nous ont appris à l'enseigner au fil des ans. Merci surtout à Maytal Saar Tshechansky et à Claudia Perlich. Il y a quelques années, Maytal a gracieusement prêté à Foster ses notes de cours de data mining. L'arbre de classification du chapitre 3 (un grand merci pour la représentation visuelle des « corps ») est en grande partie inspiré de ses idées et exemples. Il en va de même pour la représentation visuelle du chapitre 4 dans laquelle nous comparons le partitionnement de l'espace des instances avec des arbres et avec des fonctions discriminantes linéaires. L'exemple « Will David Respond » du chapitre 6 est également basé sur l'un de ses exemples, et probablement beaucoup d'autres choses que nous avons oubliées. Ces dernières années, Claudia a par ailleurs enseigné avec Foster des chapitres complémentaires des cours de Data Mining pour le Business Analytics et Introduction aux Data Sciences. Et elle lui en a beaucoup appris sur les data sciences sur le terrain (et bien plus).

Merci à David Stillwell, Thore Graepel et Michal Kosinski de nous avoir fourni les données Like de Facebook pour certains de nos exemples. Merci à Nick Street pour les données sur les noyaux de cellules et pour nous avoir autorisés à utiliser l'image des noyaux de cellules au chapitre 4. Merci à David Martens de nous avoir aidés pour la représentation visuelle des données de localisation des appareils mobiles. Merci à Chris Volinsky de nous avoir prêté les don-

nées de ses travaux sur Netflix Challenge. Merci à Sonny Tambe d'avoir très tôt partagé ses résultats sur les technologies big data et sur la productivité. Merci à Patrick Perry de nous avoir montré l'exemple du call center bancaire que nous utilisons au chapitre 12. Et merci à Geoff Webb de nous avoir permis d'utiliser le système de data mining de l'association Magnun Opus.

Par-dessus tout, nous remercions nos familles pour leur amour, leur patience et leurs encouragements.

Un grand nombre de logiciels open source ont été utilisés pour la préparation de ce livre et de ses exemples. Les auteurs souhaitent donc remercier les développeurs et contributeurs de :

- Python et Perl ;
- Scipy, Numpy, Matplotlib et Scikit-Learn ;
- Weka ;
- le Machine Learning Repository de l'université de Californie, à Irvine (Bache et Lichman, 2013).

Pour finir, nous encourageons les lecteurs à consulter régulièrement notre site web (http://www.data-science-for-biz.com/) pour les éventuels nouveaux chapitres, mises à jour des supports, errata, addenda et diapositives d'accompagnement.

— Foster Provost et Tom Fawcett

1

Le raisonnement orienté données

Ne faites pas de trop petits rêves, car ils n'ont pas
la puissance de toucher les cœurs des hommes.

— Johann Wolfgang von Goethe

D'importants investissements ont été réalisés ces dernières années dans l'infrastructure des entreprises afin d'augmenter leur capacité à collecter leurs données. Pratiquement tous les aspects de l'activité d'une entreprise sont maintenant ouverts à la collecte de données et souvent eux-mêmes employés au service de la collecte de données : opérations, fabrication, gestion de la chaîne d'approvisionnement, comportement des clients, performances des campagnes de marketing, procédures de gestion de flux, etc. Simultanément, les informations concernant les événements extérieurs à l'entreprise, tels que les tendances des marchés, les actualités des entreprises et les activités des concurrents, sont aujourd'hui largement accessibles. Cette disponibilité des données a provoqué un intérêt croissant pour les méthodes d'extraction d'informations utiles et de connaissances à partir des données, le domaine des data sciences.

L'ubiquité des opportunités offertes par les données

Avec les vastes quantités de données aujourd'hui disponibles, les entreprises de pratiquement tous les secteurs concentrent maintenant leurs efforts sur l'exploitation des données pour gagner en compétitivité. Dans le passé, les entreprises pouvaient employer des équipes de statisticiens, de modélisateurs et d'analystes pour éplucher les jeux de données manuellement, mais le volume et la diversité des données dépassent de loin les capacités de l'analyse manuelle. De plus, les ordinateurs actuels sont bien plus puissants, les réseaux sont devenus omniprésents, et de nou-

veaux algorithmes ont été développés, qui nous permettent de connecter différents ensembles de données afin d'obtenir des analyses plus étendues et plus approfondies. La convergence de ces différents phénomènes a conduit à une exploitation de plus en plus généralisée des principes de data science et des techniques de data mining au sein des entreprises.

Le secteur qui exploite le plus les techniques de data mining est probablement celui du marketing, où elles sont appliquées à des tâches telles que le ciblage marketing, la publicité en ligne et les recommandations de vente additionnelle. Le data mining y est en effet employé dans la gestion globale des relations avec les clients afin d'analyser leurs comportements, l'objectif étant de minimiser l'attrition et de maximiser la valeur client. Dans le domaine de la finance, le data mining sert pour la notation de crédit, pour les échanges commerciaux, ainsi que dans les opérations de détection de fraudes et de gestion du personnel. Et pour les grandes marques de distribution telles que Walmart et Amazon, le data mining est présent à tous les niveaux, du marketing à la gestion de la chaîne d'approvisionnement. Beaucoup d'entreprises se sont distinguées du point de vue stratégique par leur recours à la data science, au point, parfois, d'en faire leur activité principale.

Les principaux objectifs de ce livre sont de vous aider à discerner les problèmes d'entreprise du point de vue des données et à comprendre des principes qui vous permettront d'extraire des connaissances utiles de ces données. Le raisonnement orienté données se fonde sur une structure fondamentale et des principes qui doivent être maîtrisés. Parfois l'intuition, la créativité, le bon sens et les connaissances du domaine doivent être pris en compte. Les modes de raisonnement orientés données fournissent une structure et des principes qui vous serviront de cadre pour analyser de tels problèmes de manière méthodique. Plus vous maîtriserez le raisonnement orienté données, plus vous développerez des intuitions qui vous diront quand et comment recourir à la créativité et aux connaissances métier.

Tout au long des deux premiers chapitres de ce livre, nous aborderons en détail divers sujets et techniques relatifs à la data science et au data mining. Les termes « data science » et « data mining » sont souvent utilisés indifféremment, mais le premier a pris une ampleur considérable à mesure que des individus et des organisations tentent de capitaliser sur le battage médiatique qui l'entoure actuellement. En substance, la data science englobe un ensemble de principes essentiels qui régissent l'extraction de connaissances à partir des données. Le data mining désigne, quant à lui, l'extraction de connaissances à partir des données à l'aide de techniques qui se fondent sur ces principes. Le terme « data science » est souvent employé dans un sens plus général que « data mining », mais les techniques de data mining fournissent quelques exemples parmi les plus évidents des principes des data sciences.

Tout au long de ce livre, nous décrirons un certain de nombre de principes fondamentaux des data sciences. Nous illustrerons chacun de ces principes avec au moins une technique de data mining qui le caractérise. Les principes en question sont généralement à la base de beaucoup de techniques, c'est pourquoi nous avons choisi, dans ce livre, de mettre l'accent sur les principes de base plutôt que de décrire des techniques spécifiques. Cependant nous n'insisterons pas sur la différence entre data science et data mining, sauf lorsque cela risque d'entraver la compréhension des concepts présentés.

Remarque

La compréhension des data sciences est nécessaire même si vous n'avez pas l'intention de les utiliser vous-même. Le raisonnement orienté données vous permet en effet d'analyser les projets de data mining qui vous sont proposés. Si par exemple un employé, un consultant ou une éventuelle cible d'investissement vous propose d'améliorer un logiciel d'entreprise en faisant de l'extraction de connaissances à partir des données, vous devriez être capable d'évaluer méthodiquement sa proposition afin de décider si elle est raisonnable ou mauvaise. Cela ne signifie pas que vous pourrez prédire la rentabilité du projet, car dans le cas des projets de data mining, cela nécessite le plus souvent d'essayer, mais cela devrait vous permettre de détecter les défauts les plus flagrants, les hypothèses irréalistes et les éléments manquants.

Voyons maintenant deux courtes études de cas où l'analyse des données a mené à l'extraction de modèles prédictifs.

Exemple : l'ouragan Frances

Observons un exemple tiré d'un article du *New York Times* datant de 2004 :

> L'ouragan Frances était en route, traversant les Caraïbes, menaçant de frapper de plein fouet la côte atlantique de la Floride. Les habitants se réfugièrent alors en altitude, mais loin de là, à Bentonville, en Arkansas, des cadres des magasins Walmart décidèrent que la situation leur offrait une excellente occasion pour tester un de leurs plus récents outils basé sur l'analyse de données… la technologie prédictive.
>
> Une semaine avant l'arrivée de la tempête sur les côtes, Linda M. Dillman, directrice du système d'information de Walmart, réclama à son équipe de lui fournir des prévisions basées sur les informations connues au sujet de l'ouragan Charley, qui avait frappé plusieurs semaines auparavant. Munie des billions d'octets d'historiques d'achat stockés dans l'entrepôt de données de Walmart, elle se dit que l'entreprise pouvait « commencer à prédire ce qui allait se produire, au lieu d'attendre que cela se produise », selon ses dires. (Hays, 2004)

Demandez-vous *pourquoi* des prédictions basées sur les données seraient utiles dans un tel scénario. Elles pourraient permettre de savoir que les personnes se trouvant sur le chemin de l'ouragan auront tendance à acheter plus d'eau en bouteille. Cela peut sembler évident et avons-nous vraiment besoin de recourir à la data science pour nous en rendre compte ? Mais il pourrait être utile d'anticiper le *volume d'augmentation* des ventes dû à l'ouragan, afin de s'assurer que les magasins Walmart concernés soient suffisamment approvisionnés. L'exploration des données pourrait par ailleurs révéler qu'un certain DVD a été en rupture de stock sur le chemin de l'ouragan – peut-être l'était-il cette semaine dans tous les Walmart du pays, et pas seulement aux alentours du cœur de l'ouragan. Ces prédictions peuvent être utiles d'une certaine manière, mais elles seraient probablement plus générales que Mme Dillman le pensait.

Il serait plus intéressant de découvrir des modèles de comportement dus à l'ouragan qui n'étaient pas évidents. Pour cela, les analystes devraient examiner l'énorme volume de données Walmart décrivant des situations passées similaires (telles que Charley) afin d'identifier dans les environs de l'ouragan des demandes de produits *inhabituelles*. Avec les modèles pré-

dictifs obtenus, l'entreprise pourrait alors anticiper des demandes exceptionnelles de certains produits en approvisionnant les magasins concernés avant l'arrivée de l'ouragan.

Et c'est bien ce qui s'est passé. Le *New York Times* (Hays, 2004) raconte en effet que « … les experts ont exploré les données et ont découvert que les magasins auraient effectivement besoin de certains produits, et pas seulement les lampes torches habituelles. "Nous ne savions pas que juste avant l'arrivée d'un ouragan, les ventes de Pop-Tart aux fraises augmentaient, jusqu'à sept fois plus que leur taux de vente habituel", déclara Mme Dillman lors d'une interview récente. "Et le produit le plus vendu juste avant l'ouragan était la bière."[1] »

Exemple : prédiction de l'attrition client

Comment ces analyses de données sont-elles effectuées ? Observons un deuxième scénario, plus classique, et comment il pourrait être traité du point de vue de l'analyse de données. Cet exemple servira comme support pour illustrer plusieurs des problèmes soulevés dans ce livre et fournira donc un cadre commun de référence.

Supposons que vous venez d'obtenir un excellent emploi d'analyste de données à MegaTelCo, une des plus grandes entreprises de télécommunications aux États-Unis. Ils rencontrent un problème majeur de rétention client dans le secteur du sans fil. Dans la région mi-atlantique des États-Unis, 20 % des utilisateurs de téléphones mobiles les quittent à l'expiration de leur abonnement, et il devient de plus en plus difficile d'attirer de nouveaux clients. Le marché des téléphones mobiles étant aujourd'hui saturé, l'énorme croissance du marché du sans fil est en déclin. Les entreprises de télécommunications sont aujourd'hui engagées dans des batailles sans fin pour attirer les clients des autres tout en conservant les leurs. Le passage d'un client d'une entreprise à une autre est appelé *attrition*, et cela coûte cher à tous les niveaux, car pendant qu'une entreprise investit dans des programmes d'incitation à l'achat pour attirer certains clients, une autre perd en revenus lorsqu'un client la quitte.

On a fait appel à vous pour fournir une explication au problème rencontré par MegaTelCo et pour mettre au point une solution. Attirer de nouveaux clients coûte beaucoup plus que de retenir les clients actuels, une grande partie du budget marketing a donc été allouée à la prévention de l'attrition. Le service marketing a d'ailleurs déjà conçu une offre spéciale de rétention. Votre travail consistera maintenant à élaborer un programme précis, étape par étape, dans lequel vous devrez indiquer à l'équipe de data science comment utiliser les vastes données de MegaTelCo pour sélectionner les clients qui seront ciblés par l'offre spéciale de rétention avant l'expiration de leur contrat.

Choisissez soigneusement les données qui pourront vous être utiles et réfléchissez bien à la manière dont elles pourraient servir. En particulier, comment MegaTelCo devra sélectionner l'ensemble des clients cibles de leur offre de rétention afin de réduire au mieux le taux d'attrition avec un budget de rétention défini. Répondre à cette question est beaucoup plus compliqué qu'il n'y paraît à première vue. Nous y reviendrons à maintes reprises tout au long du

1. Évidemment ! Qu'est-ce qui va le mieux avec les Pop-Tart aux fraises qu'une bonne bière bien fraîche ?

livre, et nous améliorerons notre solution à mesure que notre compréhension des concepts fondamentaux des data sciences grandira.

> **Remarque**
>
> En réalité, la rétention client a été l'une des principales causes d'utilisation des technologies de data mining, en particulier dans les secteurs des télécommunications et des finances. Ces derniers furent parmi les premiers et les plus grands utilisateurs des technologies de data mining, pour des raisons que nous évoquerons plus loin.

Data science, ingénierie et prise de décision orienté données

La data science implique des principes, des processus et des techniques qui permettent de comprendre des phénomènes via l'analyse (automatisée) de données. Dans ce livre, nous verrons que l'objectif ultime de la data science est l'aide à la prise de décision, car elle représente généralement un intérêt immédiat pour les entreprises.

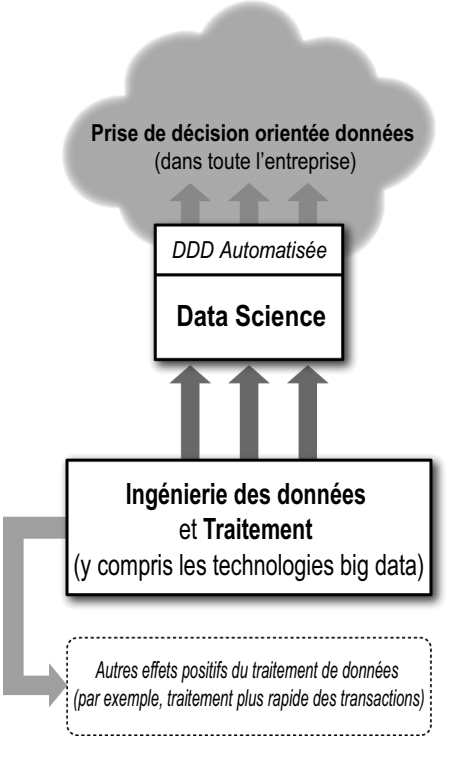

Figure 1–1
La data science dans le contexte de divers processus basés sur les données dans une organisation.

La figure 1-1 situe la data science dans le contexte de divers autres processus étroitement liés et également basés sur les données dans une organisation. Elle distingue la data science d'autres aspects du traitement des données qui font l'objet d'une attention croissante dans les entreprises. Commençons par le haut.

La prise de décision orientée données (*data-driven decision-making*, DDD) désigne la pratique qui consiste à fonder des décisions sur l'analyse de données plutôt que sur l'intuition pure. Par exemple, une responsable marketing pourrait choisir ses publicités exclusivement en fonction de sa longue expérience dans le domaine et sur son intuition quant à ce qui marchera. Ou bien, elle pourrait baser ses choix sur l'analyse de données concernant le comportement des clients vis-à-vis des différents types de publicités. Elle pourrait aussi combiner les deux approches. DDD n'est pas une activité du tout-ou-rien, et différentes entreprises s'y engagent à des degrés plus ou moins grands.

Les bénéfices de la prise de décision orientée données ont été démontrés de manière définitive. L'économiste Erik Brynjolfsson et ses collègues du MIT et de la Penn's Wharton School ont mené une étude sur la façon dont la DDD peut affecter les performances d'une entreprise (Brynjolfsson, Hitt et Kim, 2011). Ils ont développé une méthode de mesure de la DDD qui permet de classer les entreprises selon leur degré d'utilisation des données dans la prise de décision pour l'ensemble de l'entreprise. Ils ont ainsi démontré que du point de vue statistique, plus une entreprise a recours à l'analyse de données, plus elle est productive, voire dominante, pour un large éventail de facteurs surprenants possibles. Et les différences de résultats ne sont pas faibles. Un écart-type de plus sur l'échelle de la DDD coïncide avec une augmentation de productivité de 4 à 6 %. La DDD est également corrélée avec un retour sur investissement, une rentabilité des capitaux propres, une utilisation des actifs et une valeur marchande plus élevés, et la relation semble être causale.

Nous nous intéresserons dans ce livre à deux principales catégories de prise de décision :

- les décisions qui nécessitent des découvertes dans les données ;
- les décisions qui se répètent, en particulier à grande échelle, et pour lesquelles le processus de prise de décision peut donc profiter des plus petites améliorations dans l'exactitude des décisions orientées données.

L'exemple Walmart que nous avons vu plus haut est un problème de type 1 : Linda Dillman voulait extraire des connaissances permettant à Walmart de se préparer au mieux en vue de l'arrivée imminente de l'ouragan Frances.

En 2012, la compagnie Target, le principal concurrent de Walmart, a fait parler d'elle avec un cas de prise de décision orientée données, également un problème de type 1 (Duhigg, 2012). Comme la plupart des marques de grande distribution, Target s'intéresse aux comportements d'achat de ses clients, à ce qui les motive et ce qui peut les influencer. Les clients ont tendance à suivre des routines d'achat qui peuvent être très difficiles à rompre. Mais les décideurs de Target savaient que l'arrivée d'un bébé dans une famille représente un moment particulier où les gens changent leurs habitudes d'achat de manière significative. Comme le disent les analystes de Target, « À partir du moment où nous réussirons à leur faire acheter nos couches, ils commenceront à acheter tout le reste chez nous également ». Les compagnies de grande distribution le savent et rivalisent donc pour essayer de vendre aux nouveaux parents des pro-

duits liés aux bébés. Et étant donné que la plupart des registres de naissances sont publics, les entreprises se procurent facilement ces informations afin d'envoyer des offres spéciales aux nouveaux parents.

Target voulait cependant prendre de l'avance sur ses concurrents : s'ils pouvaient *prédire* que les gens *attendent* un bébé, cela leur donnerait l'avantage de pouvoir devancer leurs concurrents en faisant des offres à leurs clients. À l'aide des techniques de data science, Target analysa donc les historiques d'achat de ses clientes qui se déclarèrent effectivement enceintes par la suite (avant leur grossesse) et réussit à extraire des informations permettant de prédire quelles clientes étaient enceintes. Par exemple, les femmes enceintes changent souvent leur régime alimentaire, leur garde-robe, leur régime vitaminé, etc. Ces indicateurs pourraient être extraits de données historiques, être assemblés en modèles prédictifs et déployés ensuite lors de campagnes de marketing. Nous parlerons des modèles prédictifs plus en détail tout au long de ce livre. Pour le moment, nous dirons simplement qu'un modèle prédictif résume en grande partie la complexité du monde, en se concentrant sur un ensemble d'indicateurs qui sont corrélés d'une certaine manière avec une valeur cible (qui va-t-on perdre, qui achètera, qui est enceinte, etc.). Il est important de noter que dans les exemples Walmart et Target, l'objectif de l'analyse des données n'était pas de tester une hypothèse simple. Au lieu de cela, les données furent explorées dans l'espoir de découvrir des informations utiles[2].

Notre exemple concernant l'attrition client est un problème de DDD de type 2. MegaTelCo a des centaines de millions de clients, chacun étant un déserteur potentiel. Chaque mois des dizaines de millions de clients arrivent au terme de leur contrat, leur probabilité de désertion augmente donc dans un futur proche. Si nous pouvons améliorer notre capacité à estimer à quel point il serait profitable de cibler un client donné, alors nous pourrons récolter des bénéfices considérables en appliquant cette connaissance à ces millions de clients. Cette logique s'applique à beaucoup d'autres secteurs dans lesquels l'application des techniques de data science et de data mining fut des plus intenses : le marketing direct, la publicité en ligne, la notation de crédit, le trading financier, la gestion des services d'assistance à la clientèle, la détection des fraudes, le classement des résultats de recherche, la recommandation de produits, etc.

Le schéma de la figure 1-1, page 13, montre que la data science sert de socle à la prise de décision orientée données, mais qu'elle se recoupe également avec elle. Ce schéma met en évidence le fait, souvent négligé, que de plus en plus, en entreprise, les prises de décision sont faites de manière *automatique* par des ordinateurs. Différents secteurs ont adopté la prise de décision automatique à différents degrés. Les secteurs de la finance et des télécommunications furent parmi les premiers, en grande partie en raison de la précocité de leur développement de réseaux de données et de leurs traitements informatiques à grande échelle, qui leur permirent d'amasser et de modéliser des données à grande échelle, et ensuite d'appliquer leurs modèles à la prise de décision.

2. Target réussit si bien que cette histoire souleva des questions d'éthique sur le déploiement de telles techniques. Les problèmes d'éthique et de vie privée sont intéressants et très importants, mais nous laissons cette discussion pour un autre moment et un autre endroit.

Dans les années 1990, la prise de décision automatisée bouleversa complètement la gestion des crédits bancaires et des crédits à la consommation. Et toujours dans les années 1990, les banques et les entreprises de télécommunications mirent également en œuvre des systèmes de grande envergure pour guider les décisions concernant la surveillance des fraudes basée sur les données. À mesure que les systèmes de vente étaient informatisés, les décisions commerciales furent elles aussi automatisées. Entre autres exemples célèbres, nous pouvons citer les programmes de récompense des casinos Harrah et les systèmes automatisés de recommandation d'Amazon et de Netflix. Aujourd'hui, nous assistons à une révolution dans le secteur de la publicité, qui est due en grande partie à une augmentation considérable du temps que les consommateurs passent sur Internet, et à la capacité à prendre des décisions publicitaires en ligne (littéralement) instantanées.

Traitement des données et big data

Il convient ici de revenir sur un autre point. Une grande partie du traitement des données n'est pas de la data science, malgré l'impression que l'on pourrait tirer des médias. L'ingénierie et le traitement des données sont des éléments essentiels des tâches de data science, mais ils sont plus généraux. Aujourd'hui, par exemple, un grand nombre de compétences, de systèmes et de technologies sont souvent considérés à tort comme appartenant à la data science. Pour comprendre la data science et les entreprises orientées données, il est important de pouvoir les distinguer. La data science nécessite de disposer de données et tire souvent profit de l'ingénierie des données que les technologies de traitement des données permettent d'effectuer, mais ces technologies n'appartiennent pas à la data science en soi. Elles contribuent à la data science, comme le montre la figure 1-1, page 13, mais elles sont aussi utiles pour beaucoup d'autres choses. Les technologies de traitement des données sont essentielles pour beaucoup de tâches orientées données qui n'impliquent pas l'extraction de connaissances ou la prise de décision orientée données, telles que le traitement efficace de transactions, le traitement modernisé des systèmes web, et la gestion de campagnes publicitaires en ligne.

Les technologies big data (telles que Hadoop, HBase et MongoDB) attirent aujourd'hui l'attention des médias de manière considérable. Le terme *big data* désigne simplement des jeux de données qui sont trop volumineux pour les systèmes traditionnels de traitement de données, et qui nécessitent donc de recourir à de nouvelles technologies de traitement. Les technologies big data, comme les technologies traditionnelles, sont utilisées dans une grande variété de tâches, y compris en ingénierie des données. Parfois, elles sont même utilisées pour *mettre en œuvre* des techniques de data mining. Le plus souvent, cependant, les célèbres technologies big data sont utilisées pour effectuer le traitement des données en *support* aux techniques de data mining et autres activités de data science, comme le montre la figure 1-1.

Nous avons évoqué plus haut l'étude de Brynjolfsson démontrant les avantages de la prise de décision orientée données. Une autre étude, menée par l'économiste Prasanna Tambe, de la NYU's Stern School, avait pour objectif de mesurer à quel point les technologies big data favorisent les grandes entreprises (Tambe, 2012). Il constate qu'après avoir pris en compte différents facteurs de confusion, l'exploitation de technologies big data coïncide avec une forte croissance de la productivité. En particulier, l'utilisation des technologies big data à

hauteur d'un écart-type de plus coïncide avec une productivité supérieure de 1 à 3 % ; et l'utilisation des technologies big data à hauteur d'un écart-type de moins coïncide avec une productivité inférieure de 1 à 3 %. Cela conduit à des différences de productivité potentiellement très élevées entre les entreprises qui se situent à chacun des extrêmes.

Du big data 1.0 au big data 2.0

Une manière de voir la situation des technologies big data est d'établir une analogie avec l'adoption par les entreprises des technologies Internet. Au Web 1.0, les entreprises se sont préoccupées d'acquérir les technologies Internet de base afin de pouvoir établir leur présence sur le Web, créer des sites e-commerce et augmenter l'efficacité de leur activité. Nous pouvons donc considérer que nous sommes aujourd'hui dans l'ère du big data 1.0. Les entreprises se soucient en effet de se doter de moyens pour traiter de très larges volumes de données, le plus souvent pour consolider leurs activités actuelles, par exemple pour augmenter leur efficacité.

Une fois que les entreprises avaient incorporé complètement les technologies Web 1.0 (et, au passage, avaient provoqué une réduction des prix des technologies sous-jacentes), elles ont commencé à regarder plus loin. Elles ont par exemple commencé à se demander comment le Web pourrait leur servir, et comment il pourrait transformer leurs activités courantes. C'est alors que nous sommes entrés dans l'ère Web 2.0, où de nouveaux systèmes et entreprises tirèrent profit de la nature interactive du Web. Les changements apportés par cette évolution dans la manière de penser sont omniprésents ; les plus évidents sont l'intégration des composantes de réseautage social, et la montée de la voix du consommateur (et citoyen) individuel.

Une phase big data 2.0 s'ensuivra très probablement. Une fois que les entreprises auront entièrement acquis la capacité de traiter aisément de gros volumes de données, elles se demanderont : « Que puis-je faire aujourd'hui que je ne pouvais pas faire auparavant, ou que je pourrais faire mieux que par le passé ? » Ce sera alors l'âge d'or de la data science. Et les principes et techniques que nous introduisons dans ce livre seront appliqués à plus large échelle et de manière plus approfondie qu'ils ne le sont aujourd'hui.

Remarque

Notons que durant l'ère du Web 1.0, certaines entreprises pionnières avaient commencé à appliquer des idées du Web 2.0 bien avant les autres entreprises. Amazon est un excellent exemple, il a intégré la « voix » du consommateur très tôt, dans l'évaluation des produits, dans les critiques de produits (et jusqu'à l'évaluation des critiques de produits). De même aujourd'hui, certaines entreprises sont déjà dans le big data 2.0. Et Amazon, encore une fois, est à l'avant-garde, avec des recommandations orientées données tirées de grandes masses de données. D'autres exemples existent aussi. Les annonceurs digitaux doivent traiter de larges volumes de données (il n'est pas inhabituel de disposer de milliards d'avis sur les publicités chaque jour) tout en maintenant un débit de production très élevé (les systèmes d'enchères en temps réel prennent des décisions en dizaines de millisecondes). Nous devrions donc rechercher dans ces secteurs et dans d'autres secteurs similaires des idées de progrès du big data et de la data science qui seront ensuite adoptées par d'autres secteurs.