

## Préface

Le big data n'est pas seulement un slogan, mais bien une réalité comme le montre ce livre. Beaucoup d'entreprises et d'organisations dans les domaines de la banque, de l'assurance, du marketing, accumulent des données et n'en ont pas encore tiré tous les bénéfices. En effet, jusqu'alors les statisticiens pouvaient aider à faire parler ces données : corrélations, recherche de composantes principales. Ces méthodes apportaient des informations intéressantes, quelquefois importantes, mais agrégées.

La grande nouveauté, c'est que la puissance des calculateurs autorise maintenant à faire deux choses qui sont radicalement différentes de ce que l'on faisait avant :

- accumuler les données individuelles sur les milliers, voire des millions de clients d'une banque ou d'une compagnie d'assurance, voire même sur ceux qui ne sont pas encore clients et les traiter séparément ;
- utiliser massivement les algorithmes d'apprentissage non supervisés.

Ces algorithmes, qui sont, pour le principe, connus depuis une quarantaine d'années, nécessitent une puissance de calcul qui n'était pas disponible à l'époque et ils ont depuis, beaucoup été améliorés. Ils sont non supervisés, ce qui veut dire qu'à partir d'un vaste ensemble de données comportementales, ils prédisent avec une précision stupéfiante les décisions ultérieures d'un individu sans savoir quels sont les déterminants de son action.

Dans les trois premiers chapitres de ce livre, de grands spécialistes des statistiques appliquées et des big data expliquent d'où viennent les données, comment on les exploite. Le second et le troisième chapitres, en particulier, décortiquent le fonctionnement des algorithmes d'apprentissage qui sont à la base des résultats spectaculaires dans l'utilisation des données massives. Le quatrième et le cinquième chapitres sont consacrés aux applications dans le domaine de l'assurance. Ils sont

absolument fascinants car, écrits par de très grands professionnels, ils montrent que le monde de demain est déjà là.

Il est inutile d'insister sur l'impact économique de ces travaux, les seuls résultats obtenus en détection des fraudeurs payent largement les investissements dans les données massives.

C'est, à ma connaissance, le premier livre qui illustre si bien, dans un contexte professionnel, l'impact et les enjeux réels de ce que certains appellent « la révolution du big data ». C'est pourquoi je crois que ce livre va rencontrer un grand succès dans les entreprises.

Jean-Charles POMEROL  
Président du conseil scientifique de ISTE Editions

## Introduction

Ce livre présente l'état de l'art des méthodes big data appliquées aux problèmes d'assurance. Plus précisément, il s'agit d'un ouvrage multi-auteurs donnant une vue assez complète de cinq aspects importants, chacun d'eux étant exposé par des auteurs bien connus dans les domaines traités, qui ont des profils et expertises complémentaires (*data scientists*, actuaires, statisticiens, ingénieurs). Ceux-ci vont des méthodes classiques de l'analyse des données (incluant les méthodes d'apprentissage du type *machine learning*) jusqu'à l'impact du *big data* sur le marché, présent et futur, de l'assurance.

Le big data, les mégadonnées ou encore les données massives s'appliquent à des ensembles de données tellement vastes que non seulement les méthodes connues de gestion de données mais aussi les méthodes classiques de statistique (par exemple l'inférence) perdent de leur sens ou ne peuvent s'appliquer.

Le développement exponentiel de la puissance des ordinateurs joint au croisement de cette analyse des données et de l'intelligence artificielle permet de générer des méthodes nouvelles d'analyse de base de données gigantesques que l'on trouve notamment dans le domaine de l'assurance présentées dans cet ouvrage.

Le premier chapitre écrit par Romain Billot, Cécile Bothorel et Philippe Lenca (IMT Atlantique, Brest) présente une solide introduction au big data et son application à l'assurance. Ce chapitre insiste sur l'impact des mégadonnées en montrant notamment que des centaines de millions d'individus génèrent ainsi chaque jour quelques milliards de données. La caractérisation classique du big data par les 5V est bien illustrée et enrichie par d'autres V tels les V de variabilité et validité.

Pour remédier à l'insuffisance des techniques classiques de gestion de données, les auteurs développent les méthodes de parallélisation tant des données que des tâches

possibles grâce au développement de l'informatique *via* le parallélisme de plusieurs ordinateurs.

Les principaux outils informatiques, dont Hadoop, sont présentés ainsi que leur lien avec les plateformes spécialisées en solutions décisionnelles et le problème de la migration vers une stratégie orientée donnée. L'application à l'assurance est traitée sur trois exemples.

Le deuxième chapitre écrit par Gilbert Saporta (CNAM, Paris) passe en revue le passage des méthodes classiques d'analyse des données au *big data* montrant ainsi comment le *big data* est redevable à l'analyse des données et à l'intelligence artificielle notamment par l'emploi des méthodes d'apprentissage supervisée ou non. En outre, l'auteur insiste sur les méthodes de validation des modèles prédictifs tant il est vrai que le but final de l'utilisation du *big data* est loin d'être uniquement la constitution de bases de données gigantesques et structurées mais aussi et surtout d'outil de description et de prédiction à partir d'un ensemble de paramètres donnés.

Le troisième chapitre, rédigé par Franck Vermet (EURIA, Brest) a pour but de présenter les méthodes d'apprentissage statistique les plus couramment utilisées en actuariat applicables à de nombreux domaines de l'assurance aussi bien en vie qu'en non-vie. On y retrouve la distinction entre apprentissage supervisé ou non et l'utilisation des réseaux de neurones avec pour chacune des méthodes, notamment les plus utilisées (arbres de décision, rétropropagation du gradient perceptron, *support vector machines*, *boosting staking*, etc.) la présentation rigoureuse et claire de l'algorithme d'utilisation.

Les deux derniers chapitres sont écrits par des professionnels de l'assurance. Dans le chapitre 4, Florence Picard (Institut des actuaires, Paris) décrit le marché de l'assurance aujourd'hui et dans le futur à la lumière du développement du *big data*. Elle illustre son implémentation dans le secteur de l'assurance en présentant de manière détaillée notamment l'impact du *big data* sur les méthodes de gestion, sur le marketing et les nouveaux risques assurables ainsi que sur la sécurité des données. Avec pertinence, elle met en évidence l'émergence de nouvelles techniques managériales renforçant l'importance de la formation continue.

Le cinquième et dernier chapitre a Emmanuel Berthélé (Optimind-Winter, Paris) comme auteur, lui aussi actuaire. Il y présente les principales utilisations du *big data* en assurance et notamment la tarification et l'offre des produits, l'assurance automobile et les *telematics*, l'assurance indicielle, la lutte contre la fraude, la réassurance. Il insiste aussi sur les contraintes réglementaires spécifiques au secteur (Solvency II, ORSA, etc.) et sur la restriction actuelle de l'emploi de certains

algorithmes due à une exigence d'audibilité, restriction qui sera sans doute relâchée dans le futur.

Enfin, une remarque fondamentale émerge de ces deux derniers chapitres mettant en garde les assureurs quant à la préservation du principe de mutualisation qui est le principe fondateur de l'assurance car comme le dit Emmanuel Berthelé :

« Même si le volume de données disponibles et les capacités induites dans l'affinement des tarifs augmentent considérablement, l'individualisation du tarif n'est ni totalement réalisable, ni souhaitable tant pour les assureurs, les assurés que pour la société en général. »

En conclusion, cet ouvrage montre que le *big data* est essentiel pour le développement de l'assurance pour autant qu'on mette les garde-fous nécessaires. Ainsi, il s'adresse clairement aux managers des assurances et des banques ainsi qu'aux étudiants en master de sciences actuarielles, d'informatique, de finance et de statistique et bien sûr aux nouveaux mastères en big data qui foisonnent actuellement.