



Introduction à l'Apprentissage Statistique

Data Science et Actuariat

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming



Big Data



Défis statistiques

D'où vient la hype « Big Data »

- Découverte par le « grand public » de méthodes statistiques fondées sur l'apprentissage statistique
- On parlera aussi de Machine Learning dans leurs applications pratiques

A garder en tête

- Il ne faut pas créer une usine à gaz
- Un modèle statistique est d'autant plus robuste qu'il est simple
- Ces méthodes ne sont pas encore parfaitement adaptées à la gestion de certains types de données

Beaucoup de travail préalable à faire avant un emploi judicieux !

- Ne pas foncer tête baissée dans la mêlée

Big Data, quésaco ?

Définition simplifiée

- Données non traitables en une passe (mémoire vive),
- et dans un temps raisonnable (puissance de calcul) sur une station de travail

Deux époques

- avant 2005 : ordinateurs 32-bits. Taille $n > 10^7$, $p > 100 = 8\text{Go}$
- après 2005, ordinateurs 64-bit. Beaucoup plus de mémoire physique, mais unités de calcul limitées (8 coeurs)

Deux motivations pour deux aspects

- utilisation : description et prévision
- problématique : spatial (volume) et temporel (flux)

Caractérisation des Big Data

On a coutume de parler de Big Data lorsqu'on dispose de données

- en grand **volume** (énorme base de données)
- en grande **variété** (numérique, textes, images, vidéos, ...)
- à grande **vitesse** (fréquence d'arrivée de l'information, évolution des données)

Règle des 3 V ...

... qui doit déboucher sur le 4ème V

- création de Valeur
- par l'exploitation de ces données



Défis pratiques liés aux Big Data

Défi opérationnel, essentiellement informatique

- système d'information, architecture, capacité de stockage, ...
- calculs distribués : Hadoop, Spark, *parallel* en R

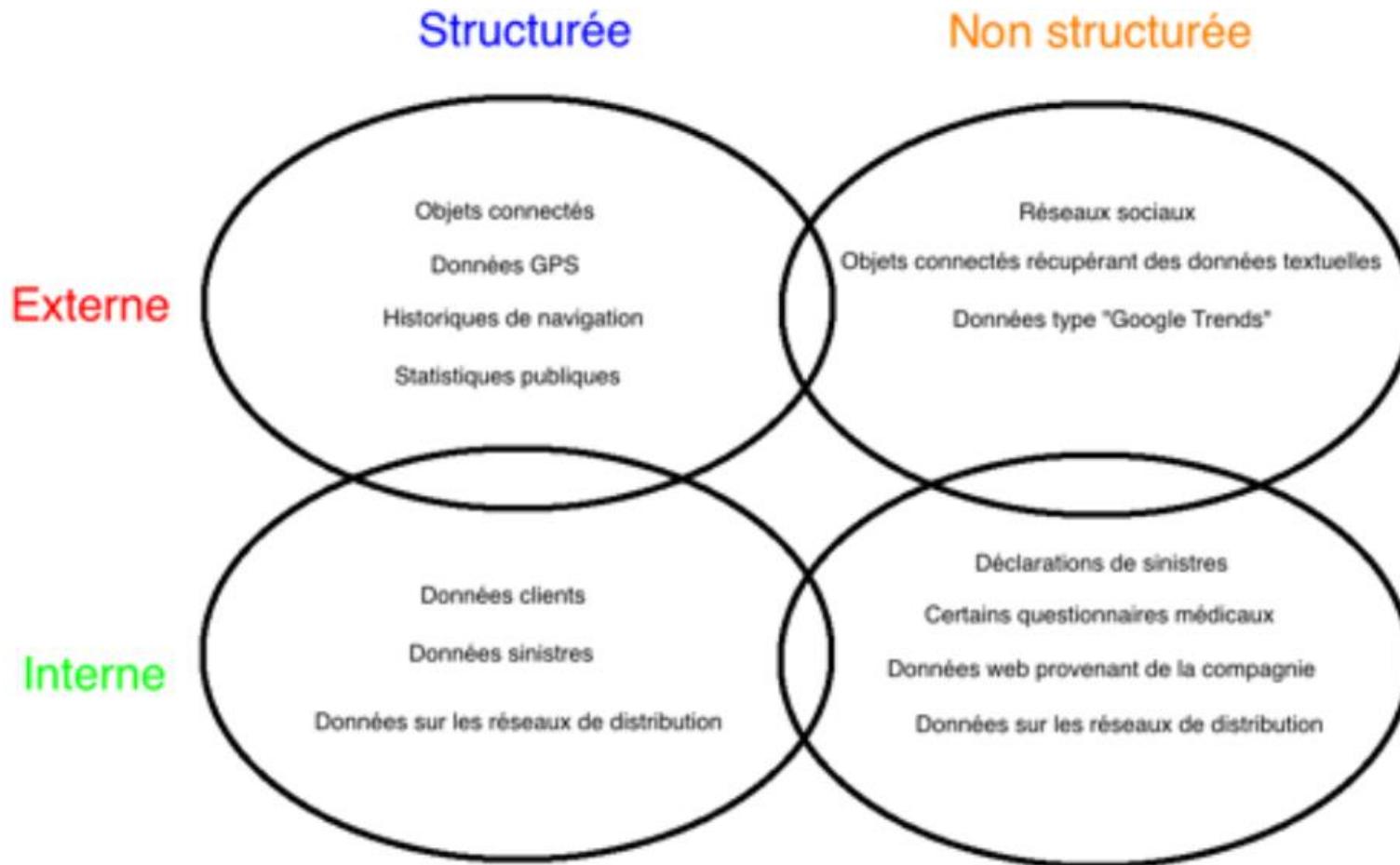
Une réflexion sur la donnée

- qualité de la donnée et gestion de son aspect non-structuré
- Comment homogénéiser des formats différents à l'origine ?
- sélection en fonction de sa pertinence et de sa gestion

Enjeu éthique

- anonymisation principalement (test génétiques en assurance maladie, ...)
- réglementation RGPD

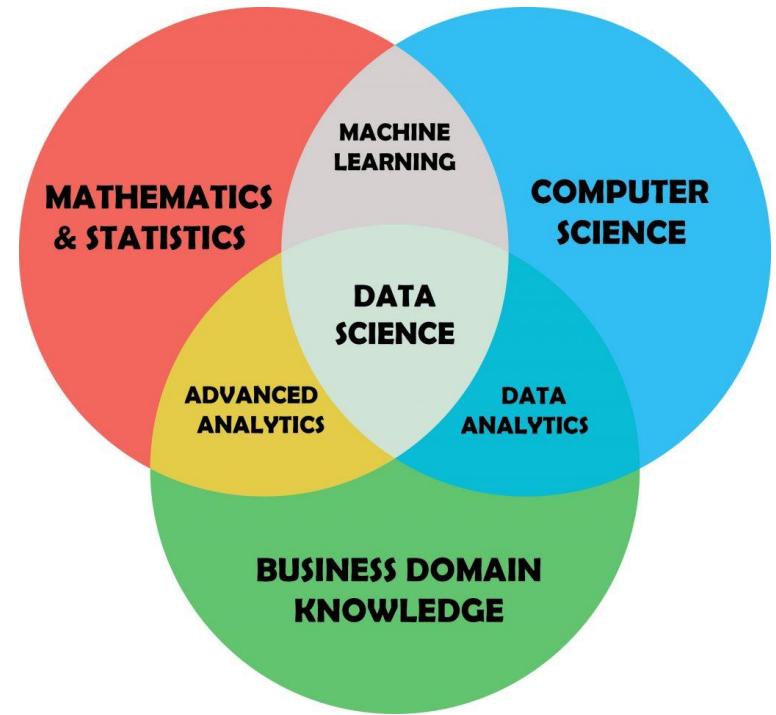
Classification des données



Data scientist

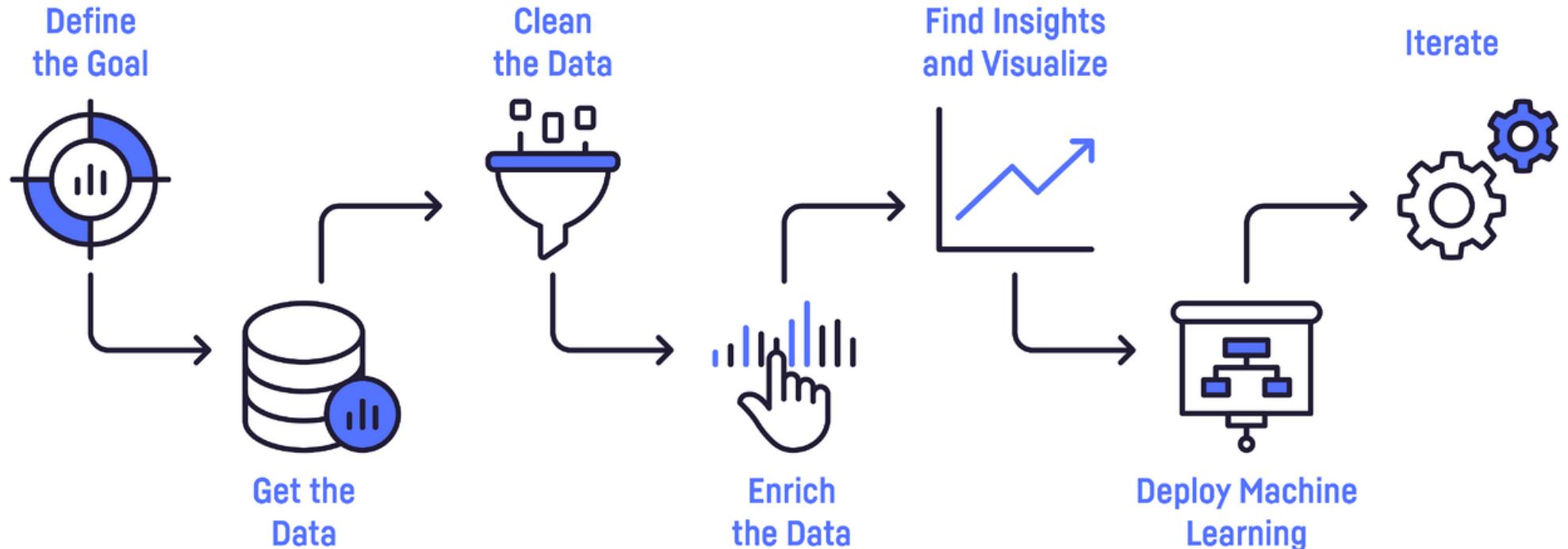
“statistics is the “grammar of data science.” It is crucial to “making data speak coherently.” [...] But it takes statistics to know whether this difference is significant, or just a random fluctuation [...] Statistics plays a role in everything from traditional business intelligence (BI) to understanding how Google’s ad auctions work. Statistics has become a basic skill. [...]”

What differentiates data science from statistics is that data science is a holistic approach. We’re increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”



Mike Loukides, 2010

Etapes d'un projet data



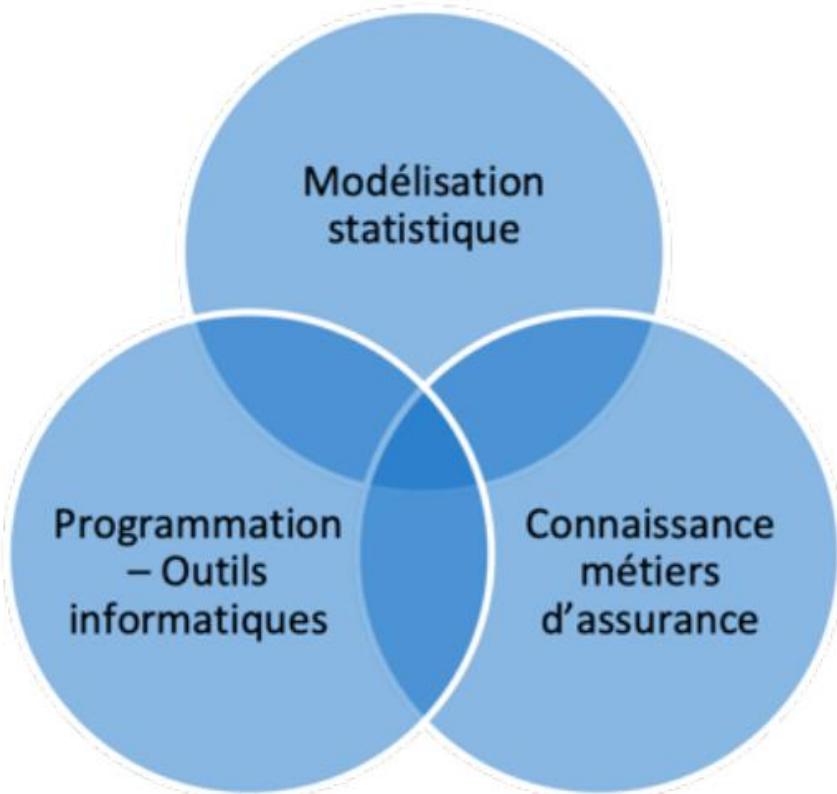
Source : dataiku



L'actuaire : un data scientist

Actuaire = Data scientist de l'assurance

Besoin d'une connaissance assez approfondie de concepts variés



Data et Actuariat

Matière première de l'actuaire = data

- Fonction Clé Actuariat : donne un avis sur la qualité des données !
- Rapport actuariel annuel : en interne pour le Board, mais à disposition sur demande de l'ACPR

Feature engineering

- Construction de nouvelles variables explicatives
- à partir de la base initiale
- et de la connaissance métier (smart data)

Attention : la responsabilité de l'actuaire ne s'arrête pas lorsque l'algorithme rend son résultat !

Feature Engineering

En pratique il s'agit d'augmenter la variété des données d'entrée

Calculer de nouvelles variables à partir de variables existantes

- significatives pour le métier
- ratios, agrégats sur fenêtre glissante, agrégat géographique
- Attention à la corrélation avec les autres variables → l'algorithme peut y être sensible

Obtenir des variables de sources externes

- Open Data structuré (e.g. INSEE)
- ou d'autres types de données (texte, image, ...)



Impact de la data en assurance

Appart en assurance

Un des principaux problèmes de l'assureur (par rapport au banquier) est la faible fréquence de ses interactions avec l'assuré

En général, ils ne se voient que 2 fois

- à la souscription
- à la déclaration du sinistre (s'il a lieu)
- Difficulté pour l'assureur de bien connaître l'assuré !

Technologies liées au Big Data vont augmenter significativement la fréquence de ces interactions et atténuer les particularités de l'assurance

- antisélection
- aléa moral
- inversion du cycle de production

Impact sur la chaîne de valeur

Le Big Data a un impact à plusieurs niveaux pour un assureur, parmi ses tâches « historiques »

- segmentation, tarification (Pay-As-You-Drive, HomeBox)
- provisionnement : micro-level reserving
- détection de fraude (par géolocalisation par exemple)
- UX (déclaration de sinistre, etc.)
- Nouveaux risques (IoT)

Remarque : échelle de temps de l'assurance est parfois beaucoup plus longue que dans d'autres secteurs (attention aux dérives du risque)

Start-up Nation

Shift Techology

- Détection de fraude
- Serie D (Mai 2021) : 220M\$ Funding, Valuation > 1B\$

Luko

- Déclaration de sinistre MRH via photo
- Serie B (Dec 2020) : 50M€ Funding

Zego

- Pay-As-You-Drive pour flotte automobile
- Serie C (Mars 2021) : 150M\$ Funding, Valuation 1,1B\$

Seyna

- IoT : chauffage pendant un hiver froid – compteur Linky
- Serie A (Dec 2019) : 14M€ Funding

Shift Technology



luko



ZEGO



Seyna.



ISFA

Le Big Data : jusqu'où ?

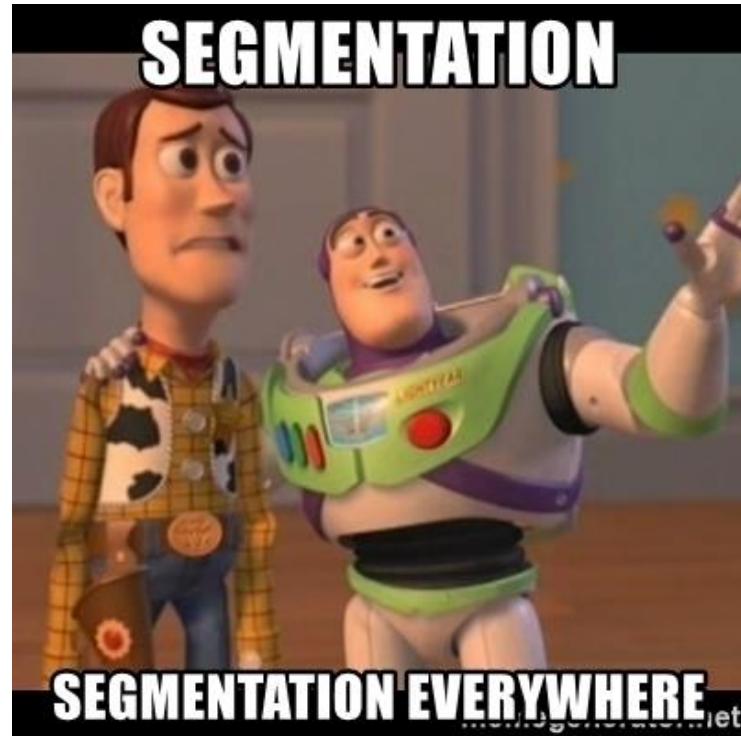
La base de l'assurance est la **mutualisation**

Or l'enjeu principal du Big Data est de mieux comprendre les mécanismes à l'échelle de l'individu !

- « We are moving from an era of private data and public analyses to one of public data and private analyses » (Andrew Gelman)

Il y a donc un risque énorme (surtout en tarification) qui est la **PERTE de MUTUALISATION**

Où arrêter la segmentation ? ...



Contexte réglementaire

Le contexte réglementaire rend le virage Data Science de plus en plus nécessaire

Amendement Bourquin et la Loi Hamon

- Assuré est plus volatile
- Nécessite une plus grande réactivité pour comprendre ses besoins, et le conserver (ou attirer) en portefeuille
- Difficulté : contacts entre assureur/assuré peu nombreux (e.g. emprunteur)

Gender directive

- Interdiction d'utiliser certains critères jugés discriminant pour tarifer (mais pas pour le provisionnement)
- Nécessité de trouver d'autres critères plus complexes pour déterminer le niveau de risque

Impacts sur la vie d'un produit d'assurance

Souscription

- Souscription de plus en plus digitale
- Objectif : améliorer le taux de transformation ...
- ... sans perdre de vue la « valeur-client »

Tarification

- Individualisation de la prime ...
- ... sans porter atteinte au principe de mutualisation

Résiliation (*churn*) ou rachats

- Anticipation du comportement de l'assuré
- Mise en place d'incentives

Impacts sur la vie d'un produit d'assurance

Prévention

- Intervenir pour empêcher le sinistre de se produire
- Changement de la relation avec l'assuré

Déclaration de sinistre

- Simplification de la déclaration par digitalisation
- Tout en luttant contre le risque de fraude

Provisionnement

- Amélioration de l'évaluation des engagements pris
- Atterrissage pour l'estimation des montants de sinistres graves