



Les principes fondamentaux de la science des données

Instructeur : Patrick Boily



uOttawa

Institut de développement professionnel
Professional Development Institute

Les principes fondamentaux de la science des données

P. BOILY

UNIVERSITÉ D'OTTAWA | FACULTÉ DES SCIENCES | DÉPARTEMENT DES MATHÉMATIQUES ET DES STATISTIQUES
DATA ACTION LAB | IDLEWYLD ANALYTICS

Instructeur – Patrick Boily

Emploi

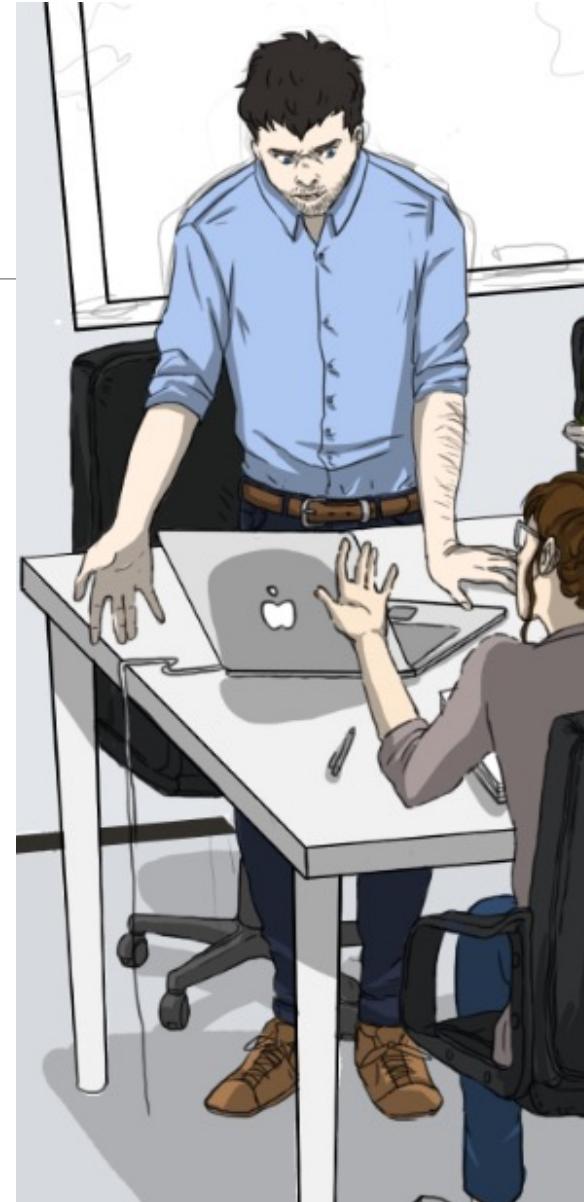
- Professeur Math/Stat [depuis '19, uOttawa]
- Président et conseiller principal [depuis '16, Idlewyld Analytics]
- Directeur et consultant principal ['12 - '19, CQADS, Carleton]
- Fonctionnaire GdC ['08 - '12, ASFC | StatCan | TC | TPSGC]
- 60+ cours universitaires ; 250+ jours d'atelier

Projets

- GAC ; NWMO ; CATSA ; etc.
- 40+ projets

Spécialisation

- Visualisation des données ; nettoyage des données (... malheureusement)
- Application d'un large éventail de techniques à tous les types de données
- Modélisation mathématique/statistique



Matériel de cours

Page Web du cours :

<https://data-action-lab.com/101-dse>

Contact :

pboily@uottawa.ca

Notes de cours :

<https://idlewyldanalytics.com>

Espace de travail Slack :

<https://dspdi.slack.com>

Description du cours

Ce cours offre aux participants l'occasion de maîtriser les connaissances et compétences fondamentales nécessaires à l'analyse des données.

Les participants seront initiés à diverses méthodes de préparation des données, à certaines limites intrinsèques des données et de l'analyse des données, ainsi qu'à des erreurs de pré-analyse facilement évitables.

Après le cours, les participants ont la possibilité de travailler sur un projet guidé, en recevant du feedback de l'instructeur.

Informations supplémentaires

Une exposition à la programmation et aux concepts introduits dans un premier cours universitaire de probabilités et de statistiques serait bénéfique (mais pas nécessaire).

Les participants sont encouragés à apporter un ordinateur portable/personnel sur lequel la version actuelle de R/Rstudio est installée (pour lequel ils peuvent avoir besoin d'une autorisation administrative pour installer des paquets).

Les participants au projet guidé doivent être familiers avec R et/ou Python.

Objectifs d'apprentissage

À la fin de ce cours, les participants seront en mesure de :

- sélectionner des méthodes appropriées pour préparer leurs données à l'analyse
- anticiper les défis et les limites inhérents aux données et aux résultats d'analyse souhaités
- appliquer des stratégies de nettoyage des données à leurs données
- effectuer des analyses simples
- construire de simples pipelines de science des données

Plan du cours

Les aspects techniques et non techniques des données

1. Les compétences quantitatives
- Les logiciels et les outils
- L'approche des "I" multiples
- Les rôles et les responsabilités
- Aide-mémoire

Les bases de la science des données

2. Les préliminaires
3. Les cadres conceptuels
4. L'éthique de la science des données
5. Le flux de travail analytique
6. Les données et les renseignements

Session 1

Session 2

Session 3

Session 4

Plan du cours

La préparation des données

7. La qualité et le traitement des données
8. Les valeurs manquantes
9. Les observations anormales
10. La dimensionnalité et les transformations de données

Miscellanea

11. L'ingénierie des données
12. La gestion des données

Session 1

Session 2

Session 3

Session 4

Problème des champignons vénéneux

Amanita muscaria

Habitat : bois

Taille des branchies : étroites

Odeur : aucune

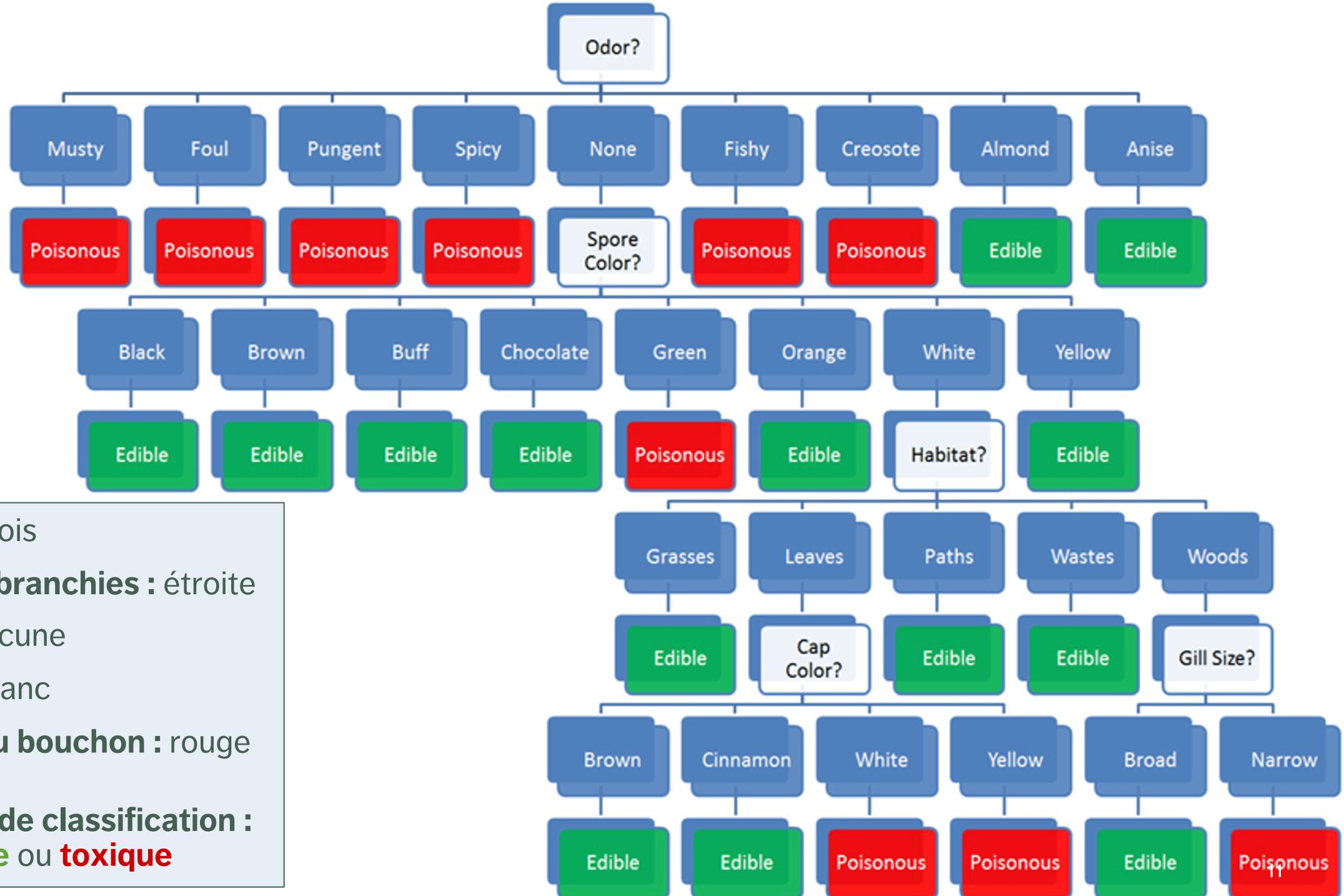
Spores : blanc

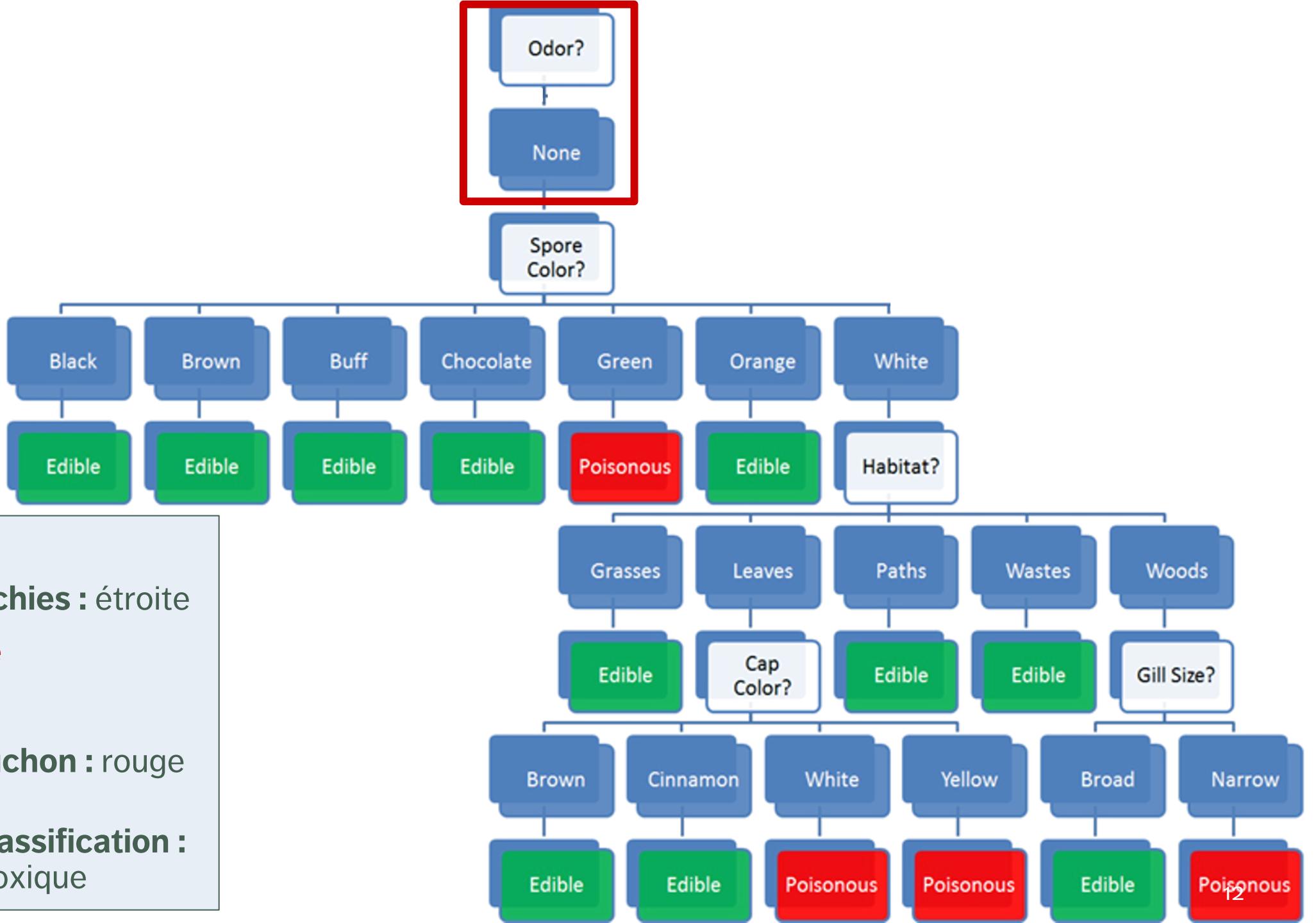
Couleur du chapeau : rouge

Problème de classification :

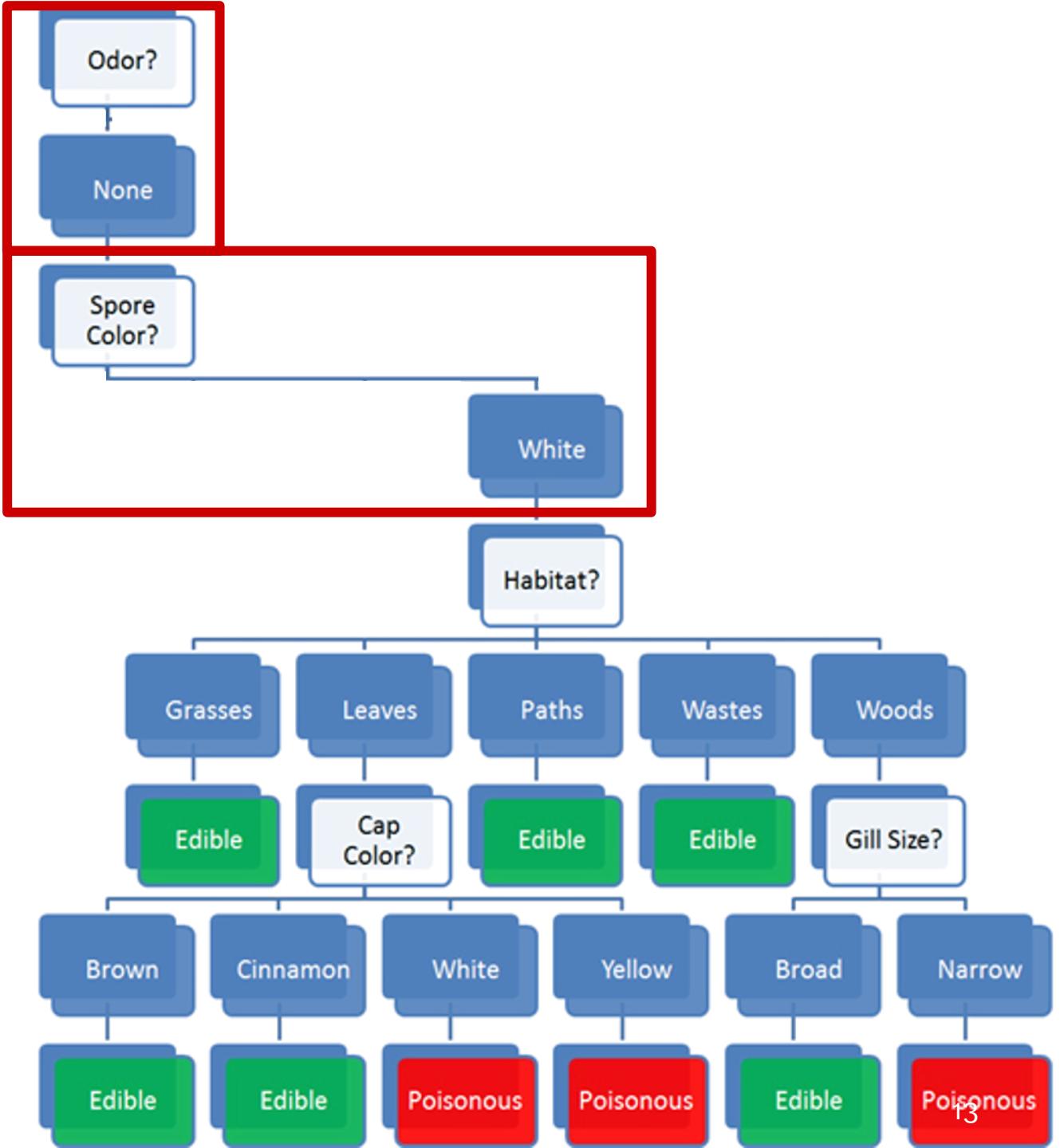
L'*Amanita muscaria* est-elle comestible ou toxique ?







Habitat : bois
Taille des branchies : étroite
Odeur : aucune
Spores : blanc
Couleur du bouchon : rouge
Problème de classification :
comestible ou toxique



Habitat : bois

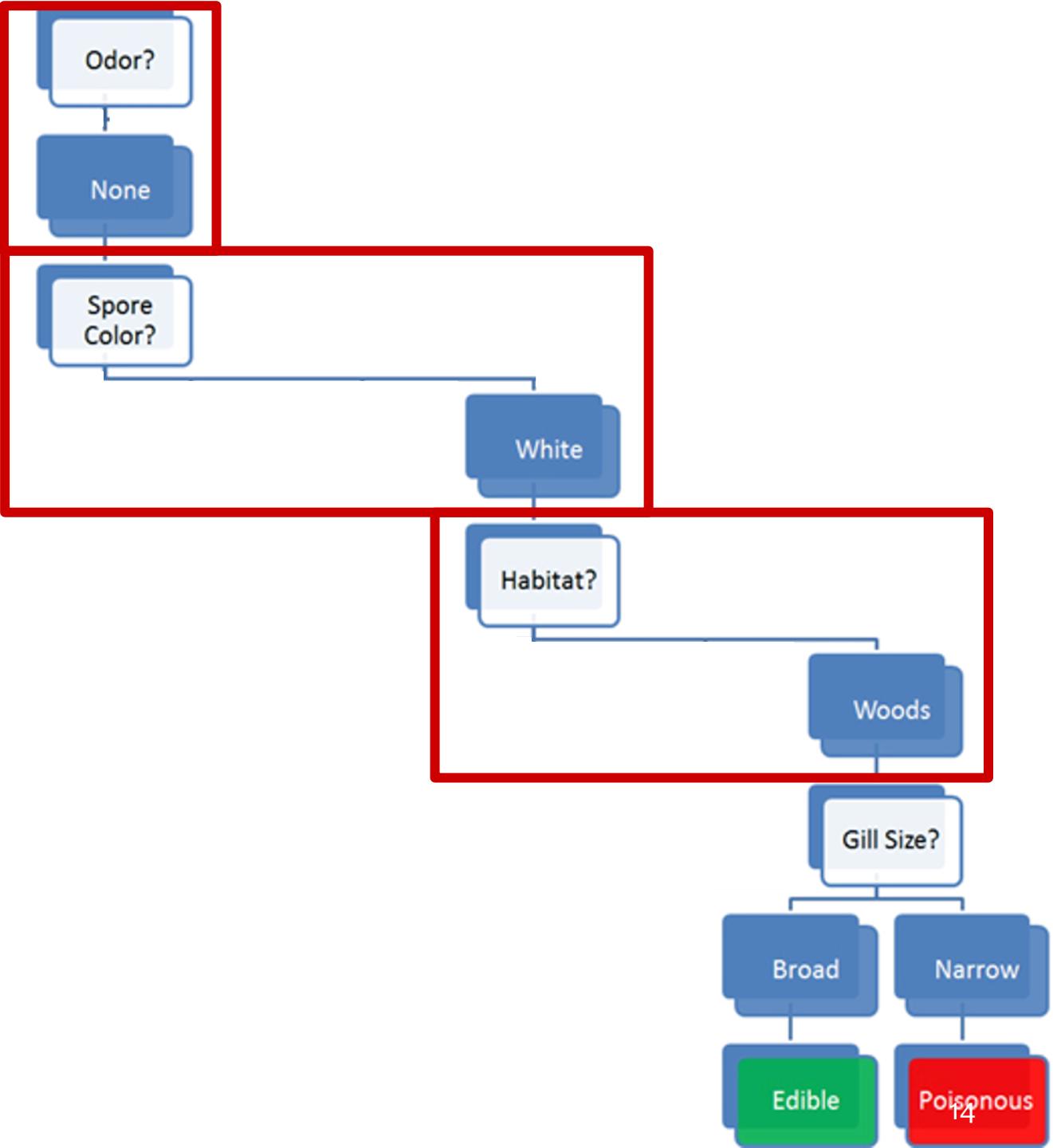
Taille des branchies : étroite

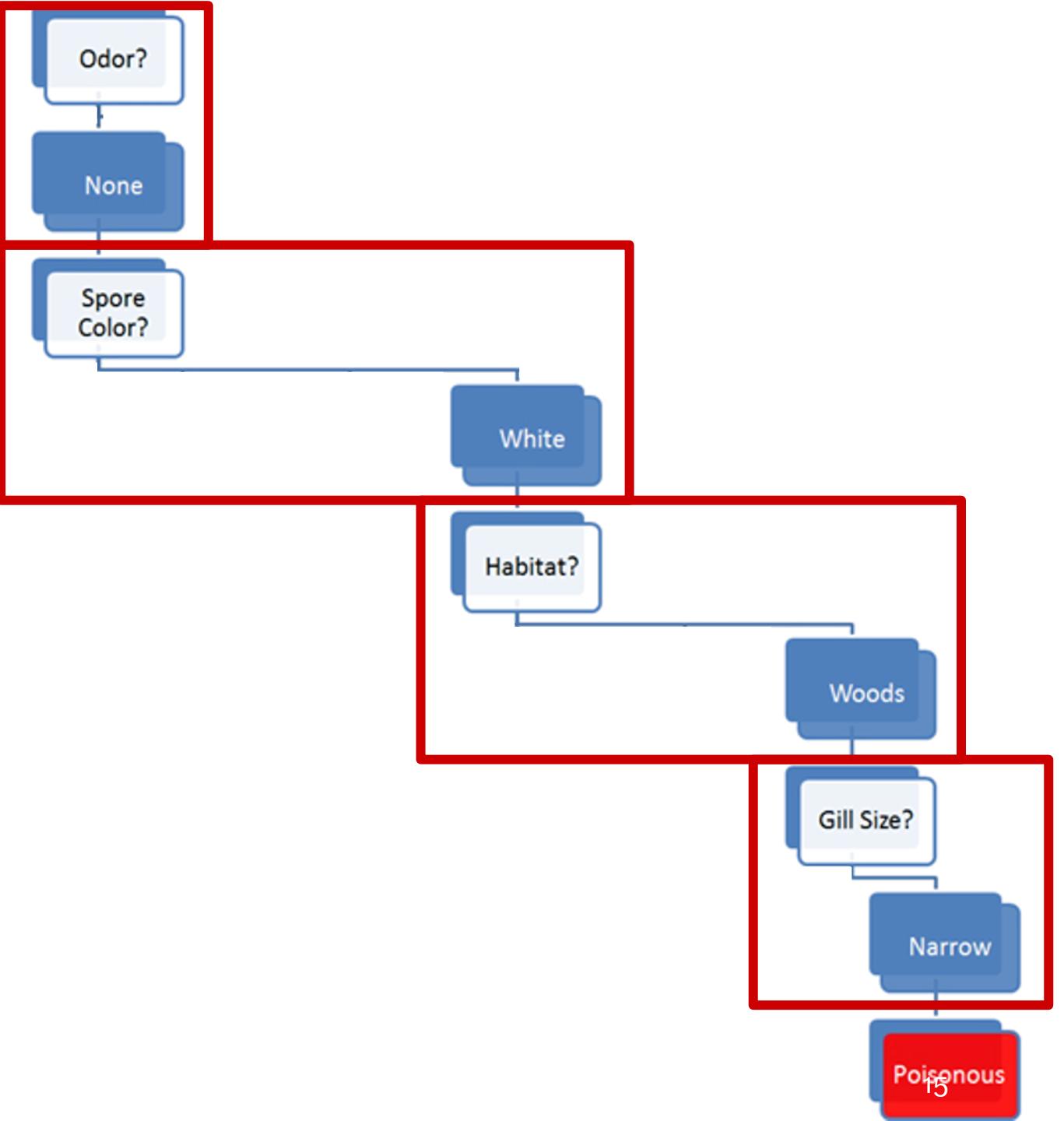
Odeur : aucune

Spores : blanc

Couleur du bouchon : rouge

Problème de classification :
comestible ou toxique





Habitat : bois

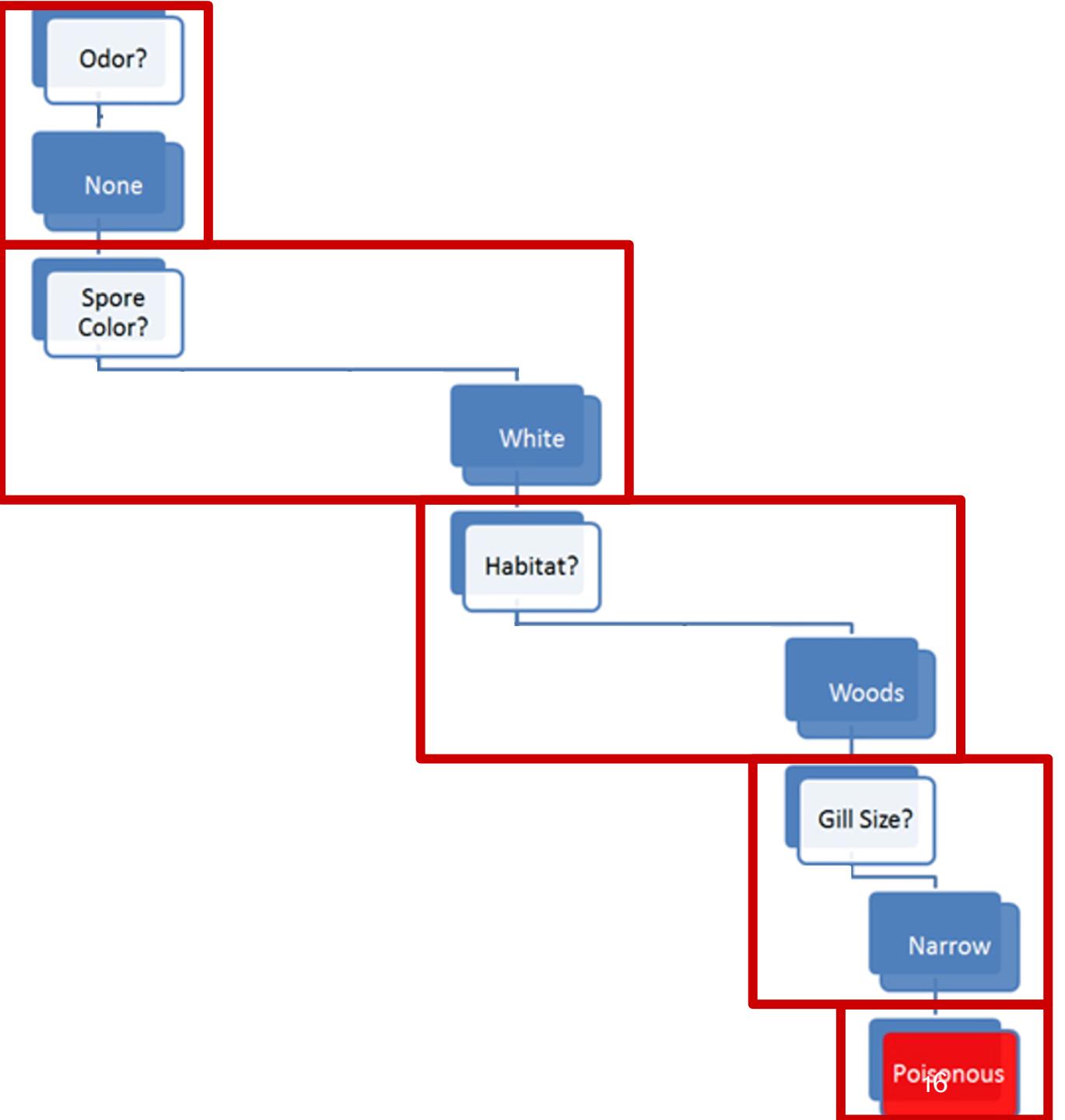
Taille des branchies : étroite

Odeur : aucune

Spores : blanc

Couleur du bouchon : rouge

Problème de classification :
comestible ou toxique



Habitat : bois

Taille des branchies : étroite

Odeur : aucune

Spores : blanc

Couleur du bouchon : rouge

Problème de classification :
comestible ou **toxique**

Discussion

Auriez-vous fait confiance à une prédiction "**comestible**" ?

D'où vient le modèle ?

Qu'auriez-vous besoin de savoir pour faire confiance au modèle ?

Quel est le coût d'une erreur de classification, dans ce cas ?

Cours sœurs

LA VISUALISATION DES DONNÉES ET LES TABLEAUX DE BORD

1. Concepts fondamentaux
2. Les tableaux de bord
3. La mise en récit de données
4. Les visualisations avec ggplot2

INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

1. L'apprentissage statistique
2. La classification
3. Le regroupement
4. Les problèmes et les défis

Session 1

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

Les aspects techniques et non techniques des données

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES



1. Les aspects techniques et non techniques des données

Les compétences quantitatives

Contexte extra-universitaire :

- appliquer des **méthodes quantitatives** à des problèmes (d'affaires) afin d'obtenir des **informations exploitables**
- il est difficile d'avoir une expertise dans **tous les** domaines des mathématiques, des statistiques, de l'informatique, de la science des données, de l'ingénierie des données, etc.

Avec un diplôme en maths/stats, par exemple :

- **expertise** dans 2-3 domaines
- **compréhension décente** des disciplines connexes
- **connaissances de base** du reste domaines

La flexibilité est une alliée, le perfectionnisme... c'est un peu moins évident

Les compétences quantitatives

Suggestions :

- suivre les tendances
- devenir **compétent dans vos domaines non spécialisés**
- savoir **où trouver des renseignements**

Dans de nombreux cas (70 % ?), seuls les fondements (2^e -3^e années de cours obligatoires à uOttawa) suffisent pour répondre aux besoins du gouvernement et de l'industrie.

Focus : assurez-vous de bien **comprendre** les bases et les tremplins.

Dans les autres cas, des connaissances plus sophistiquées sont nécessaires.

Les compétences quantitatives

- échantillonnage et collecte des données
- traitement et nettoyage des données
- visualisation des données
- modélisation mathématique
- méthodes statistiques
- analyse de régression
- modèles de file d'attente
- apprentissage machine
- apprentissage profond
- apprentissage par renforcement
- modélisation stochastique
- optimisation et recherche opérationnelle
- analyse de survie
- analyse bayésienne des données
- détection des anomalies
- réduction de la dimension
- extraction et la prévision des tendances
- cryptographie et théorie du codage
- conception des expériences
- théorie des graphes et des réseaux
- traitement du langage naturel
- etc.

Les logiciels et les outils

Le travail quantitatif moderne requiert généralement de la **programmation** (ou l'utilisation de logiciels de type pointer-cliquer, à tout le moins).

Mais les langages de programmation **vont et viennent**.

Il est important de comprendre non seulement la syntaxe d'un langage particulier, mais aussi le fonctionnement des langages informatiques et de l'infrastructure informatique en général.

ATTENTION : ne vous laissez pas entraîner dans les rivalités de programmation ... tout est plus ou moins équivalent sur le plan fonctionnel !

Les logiciels et les outils

Programmation

- Python, R, C/C++/C#, Perl, Julia, regexps (, Visual Basic ?), Java, Ruby, etc.

Gestion des bases de données

- SQL et variantes, ArangoDB, MongoDB, Redis, Amazon DynamoDB (, Access ?), Big Query, Redshift, Synapse, etc.

Visualisation des données

- ggplot2, seaborn, plot.ly, Power BI, Tableau, D3.js, Google Data Studio, logiciels spécialisés, etc.

Simulations, analyse statistique, analyse des données, apprentissage automatique

- tidyverse, scikit-learn, numpy, pandas, scipy, MATLAB, Simulink, SAS, SPSS, STATA (, Excel ?), Visio, TensorFlow, keras, Spark, Scala, etc.

Mise en page et rapports

- LaTeX, R Markdown, Adobe Illustrator, GIMP (, Word ?, PowerPoint ?), etc.

Les logiciels et les outils

Q : À StatCan, R ou SAS ?

R : StatCan est dans une lente période de transition. L'Agence est mieux équipée pour SAS (avec des options "Big Data", comme SAS Grid).

R n'est pas aussi idéal pour les gros fichiers (par exemple, les données de recensement), il n'est donc pas une option dans de tels cas car il est encore trop lent (à moins que vous ne disposiez de serveurs très puissants). Mais nous préférerions utiliser les paquets R, c'est donc un dilemme.

TL;DR : R est notre avenir, mais SAS est encore très présent. En période de transition, **les analystes qui connaissent les deux sont mieux placés.**

L'approche des “I” multiples

La compétence (ou l'expertise) technique et quantitative est **nécessaire** pour faire un bon travail quantitatif *dans le monde réel*, mais elle **n'est pas suffisante** - les solutions optimales dans le monde réel ne sont pas toujours les solutions académiques ou analytiques optimales.

C'est peut-être la plus grande surprise pour les nouveaux gradés universitaires.

Ce qui fonctionne pour une personne, un projet, etc. peut ne pas fonctionner pour un autre - **méfiez-vous de la tyrannie des succès précédents !**

L'objectif du travail quantitatif inclus la livraison d'**analyses/produits utiles**.

L'approche des “I” multiples

- **intuition**
compréhension du contexte
- **initiative**
établir un plan d'analyse
- **innovation**
nouvelles façons d'obtenir des résultats, au besoin
- **assurance (“insurance”)**
essayer plusieurs approches
- **interprétabilité**
fournir des résultats explicables
- **utilité (“insights”)**
fournir des résultats exploitables
- **intégrité**
rester fidèle aux objectifs et aux résultats
- **indépendance**
auto-apprentissage et auto-enseignement
- **interactions**
des analyses solides grâce au travail d'équipe
- **intérêt**
trouver des résultats intéressants
- **intangibles**
penser "en dehors de la boîte" ;
- **curiosité (“inquisitiveness”)**
ne pas se contenter de poser les mêmes questions à plusieurs reprises

L'approche des “I” multiples

Les analystes ne sont pas seulement jaugés sur leur savoir-faire technique, mais aussi sur leur capacité à **contribuer positivement** au lieu de travail :

- communication
- travail en équipe et capacités multidisciplinaires
- les subtilités sociales et la flexibilité
- intérêts non techniques

Les employeurs choisissent rarement des robots lorsque des êtres humains sont disponibles ; les parties prenantes sont plus susceptibles d'accepter les recommandations quantitatives provenant d'**analystes bien équilibrés**.

Vous devez également évaluer les éventuels employeurs/clients sur ces axes.

Rôles et responsabilités

Une analyste de données ou une scientifique de données (au **singulier**) a peu de chances d'obtenir des résultats significatifs – il y a trop de parties mobiles.

Les projets réussis nécessitent des **équipes** de personnel hautement qualifié qui comprennent les **données**, le **contexte** et les **défis**.

La taille de l'équipe peut varier de quelques personnes à plusieurs dizaines ; il est généralement plus facile de gérer des équipes plus petites (de 1 à 4 membres, par exemple, avec des **chevauchements de rôles**).

Experts du domaine

- font autorité dans un domaine ou un sujet particulier
- guider l'équipe en cas de complications inattendues et de lacunes dans les connaissances

Rôles et responsabilités

Chefs de projet / chefs d'équipe

- comprendre suffisamment le processus pour reconnaître si ce qui est fait a du sens
- fournir des estimations réalistes du temps et des efforts nécessaires à la réalisation des tâches
- agir en tant qu'intermédiaire entre l'équipe et les clients/partenaires
- responsable des livrables du projet.

Traducteurs de données

- avoir une bonne maîtrise des données et du dictionnaire de données
- aider les experts de domaines à transmettre le contexte sous-jacent à l'équipe de science des données

Ingénieurs en données / Spécialistes en bases de données

- travailler avec les clients et les parties prenantes pour acquérir des sources de données utilisables
- peuvent participer aux analyses, mais ne sont pas nécessairement des spécialistes.

Rôles et responsabilités

Analystes de données

- nettoyer et traiter les données
- préparer les visualisations initiales
- avoir une compréhension décente des méthodes quantitatives (au maximum 1 domaine d'expertise)
- effectuer des analyses préliminaires

Scientifiques des données

- travailler avec des données traitées pour construire des modèles sophistiqués
- concentrez-vous sur des informations exploitables
- avoir une bonne compréhension des algorithmes/méthodes quantitatives (2 ou 3 domaines d'expertise)
- peuvent les appliquer à une variété de scénarios de données
- on peut compter sur vous pour rattraper rapidement les nouvelles matières.

Rôles et responsabilités

Ingénieurs en informatique

- concevoir et réaliser des systèmes informatiques et des pipelines
- participent au développement de logiciels et au déploiement de solutions de science des données.

Spécialistes en assurance qualité/contrôle de la qualité AI/ML

- concevoir des plans d'essai pour les solutions qui mettent en œuvre des modèles AI/ML
- aider l'équipe à déterminer si les modèles sont capables d'apprendre

Spécialistes en communication

- communiquer des informations exploitables aux gestionnaires, aux analystes politiques, aux décideurs et aux parties prenantes.
- peuvent participer aux analyses, mais pas nécessairement des spécialistes (souvent des traducteurs de données)
- se tenir au courant des comptes rendus populaires des résultats et développements quantitatifs

Aide-mémoire

1. Les solutions commerciales ne sont pas toujours des solutions académiques.
2. Les données ne soutiennent pas toujours les espoirs et les besoins des parties prenantes.
3. Une communication opportune est essentielle – à l'externe comme à l'interne.
4. Les scientifiques des données doivent être flexibles (dans la limite du raisonnable), et capables d'apprendre quelque chose de nouveau, rapidement.
5. Tous les problèmes ne doivent pas faire appel la science des données.
6. Nous devons tirer des leçons des bonnes comme des mauvaises expériences.

Aide-mémoire

7. Gérez les projets et les attentes.
8. Maintenez un équilibre sain entre vie professionnelle et vie privée.
9. Respectez les parties prenantes, le projet, les méthodes et l'équipe.
10. L'analyse ne consiste pas à montrer à quel point nous sommes intelligents, mais à savoir comment nous pouvons fournir des informations exploitables.
11. Lorsque ce que le client veut est impossible, proposez des alternatives.
12. "Il n'y a pas de repas gratuit."

Lectures suggérées

Les aspects techniques et non techniques des données

Data Understanding, Data Analysis, Data Science Non-Technical Aspects of Data Work

First Principles

- The “Multiple I’s” Approach
- Roles and Responsibilities
- Consulting/Analysis Cheatsheet

Lessons Learned

Exercices

Les aspects techniques et non techniques des données

1. Parmi les compétences quantitatives présentées dans cette section, quelles sont celles que vous possédez ? Lesquelles vous intéressent ? Lesquelles envisagez-vous d'apprendre ?
2. Parmi les compétences informatiques présentées dans cette section, quelles sont celles que vous possédez ? Lesquelles vous intéressent ? Lesquelles envisagez-vous d'apprendre ?
3. Quel rôle en matière de données occupez-vous dans votre organisation ? Pour quel rôle pensez-vous être le mieux placé actuellement ? Quel est le rôle auquel vous aspirez ?
4. Avez-vous rencontré les leçons de l'aide-mémoire dans votre travail ? En avez-vous rencontré d'autres ?

Les bases de la science des données

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

2. Les préliminaires

La dichotomie numérique/analogique

Les humains collectent des données depuis longtemps ; J.C. Scott affirme que la collecte de données est un des principaux catalyseur de l'État-nation.

Historiquement, nous avons vécu dans le **monde analogique** (compréhension fondée sur l'expérience continue de la **réalité physique**).

Nos activités de collecte de données ont été les premiers pas vers une stratégie différente pour comprendre et interagir avec le monde.

Les données nous amènent à conceptualiser le monde d'une manière **plus discrète que continue**.

La dichotomie numérique/analogique

En traduisant nos expériences en chiffres et en catégories, nous créons des frontières **plus nettes** que ce que notre expérience “brute” pourrait suggérer.

Cette stratégie de discrétisation conduit à l'**ordinateur numérique** (série de 1 et 0), qui réussit assez bien à représenter notre monde physique : le **monde numérique** prend une réalité aussi omniprésente et importante que le monde physique.

Ce monde numérique est construit sur le monde physique, mais il **ne fonctionne pas selon les mêmes règles** :

- dans le monde physique, le défaut est **d'oublier** ; dans le monde numérique, c'est de **se souvenir**
- dans le monde physique, le défaut est **privé** ; dans le monde numérique, le défaut est **public**
- dans le monde physique, la copie est **difficile** ; dans le monde numérique, la copie est **facile**

La dichotomie numérique/analogique

La numérisation rend **visibles des** choses **autrefois cachées**.

Les scientifiques des données sont des scientifiques du **monde numérique**. Elles cherchent à comprendre :

- les **principes fondamentaux des données**
- comment ces principes fondamentaux se manifestent dans différents phénomènes numériques

En fin de compte, les données et le monde numérique sont **liés au monde physique**. Ce qui est fait avec les données a des **répercussions** dans le monde physique ; et il est crucial de maîtriser les **principes fondamentaux** et le **contexte** du travail de données avant de se lancer dans les outils et les techniques.

Qu'est-ce qu'une donnée ?

Il est difficile de donner une définition précise des **donnée** (est-ce au singulier ou au pluriel ?).

D'un point de vue linguistique, une **donnée** est "un élément d'information". Les **données** signifient donc "éléments d'information" ou "**collection** d'éléments d'information".

Les **données** représentent le tout (potentiellement plus grand que la somme de ses parties) ou simplement le concept idéalisé.

Est-ce que c'est clair ?

Qu'est-ce qu'une donnée ?

Est-ce que ce qui suit représente des données ?

4,529

“rouge”

25.782

“Y”

Pourquoi ? Pourquoi pas ? Que manque-t-il, le cas échéant ?

L'approche Potter Stewart : "on les reconnaît lorsqu'on le voit".

De manière pragmatique, les données sont des collections d'observations concernant des **objets** et leurs **attributs**.

Objets et attributs

Objet : *pomme*

- **Forme** : sphérique
- **Couleur** : rouge
- **Fonction** : alimentation
- **Lieu** : réfrigérateur
- **Propriétaire** : Jen



Objet : *sandwich*

- **Forme** : rectangle
- **Couleur** : brun
- **Fonction** : alimentation
- **Lieu** : bureau
- **Propriétaire** : Pat



N'oubliez pas : un objet n'est pas simplement **la somme de ses attributs**.

Objets et attributs

Ambiguïtés lorsqu'il s'agit de **mesurer** (et d'**enregistrer**) les attributs :

- l'image d'une pomme est une représentation 2D d'un objet 3D
- la forme générale du sandwich n'est que vaguement rectangulaire (**erreur de mesure ?**)
- insignifiants pour la plupart, mais pas nécessairement pour tous, les objectifs analytiques
- la forme de la pomme = volume, la forme du sandwich = surface (**mesures incompatibles**)
- un certain nombre d'attributs potentiels ne sont pas mentionnés : taille, poids, temps, etc.
- y a-t-il d'autres problèmes ?

Les erreurs de mesure et les listes incomplètes font toujours partie du tableau ; cette collection d'attributs fournit-elle une **description** raisonnable des objets ?



Des objets et attributs aux données

Les **données brutes** peuvent exister dans n'importe quel format.

Un **ensemble de données** représente une collection qui pourraient peut-être introduites dans des algorithmes à des fins d'analyse.

Les ensembles de données se présentent sous la forme de **tableau**, avec des **rangées** et des **colonnes**. Les attributs en sont les **champs** (ou colonnes, variables) ; les objets, les **instances** (ou cas, lignes, enregistrements).

Les objets sont décrits par leur **vecteur de caractéristiques** (signature de l'observation) – la collection d'attributs associés à l'observation d'intérêt.

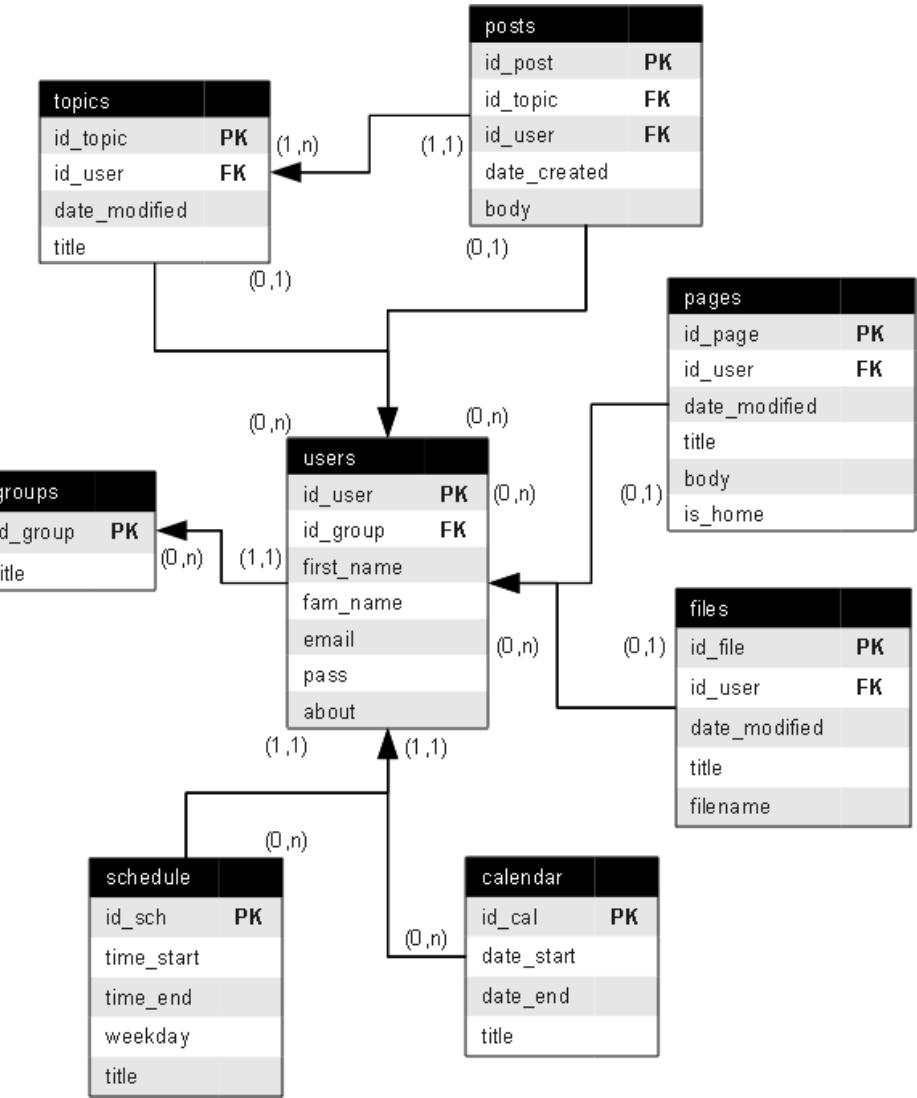
Des objets et attributs aux données

L'ensemble de données de ces objets physiques pourrait commencer par :

ID	shape	colour	function	location	owner
1	spherical	red	food	fridge	Jen
2	rectangle	brown	food	office	Pat
3	round	white	tell time	lounge	school
...

Des objets et attributs aux données

En pratique, on utilise des **banques de données** plus complexes, pour diverses raisons que nous aborderons brièvement à une étape ultérieure.



Les données dans l'actualité

Voici un échantillon de titres de journaux et d'articles mettant en évidence le rôle croissant de la **science des données** (SD), de l'**apprentissage automatique** (AA) et de l'**intelligence artificielle/augmentée** (IA) dans différents domaines de la société.

Bien que ceux-ci démontrent certaines des fonctionnalités/capacités des technologies SD/AA/IA, il est important de rester conscient que les nouvelles technologies sont accompagnées de **conséquences sociales émergentes** (pas toujours positives).

Les données dans l'actualité

- "Les robots sont meilleurs que les médecins pour diagnostiquer certains cancers, selon une étude majeure"
- "Diagnostic assisté par apprentissage profond pour l'imagerie par résonance magnétique du genou : Développement et validation rétrospective de MRNet "
- "Google AI revendique une précision de 99 % dans la détection du cancer du sein métastatique"
- "Des chercheurs trouvent des liens entre le mois de naissance et la santé"
- "Des scientifiques utilisent le suivi GPS sur les chiens sauvages Dhole, une espèce menacée".
- "Ces noms de couleurs de peinture inventés par l'IA sont si mauvais qu'ils sont bons"
- "Nous avons essayé d'enseigner à une IA à écrire des intrigues de films de Noël. L'hilarité s'ensuit. Éventuellement."
- "Un modèle mathématique détermine qui a écrit "In My Life" des Beatles : Lennon ou McCartney ?"

Les données dans l'actualité

- "Des scientifiques utilisent les données d'Instagram pour prévoir les top models du *Fashion Week* de New York"
- "Comment le big data va résoudre votre problème de courriel"
- "L'intelligence artificielle performe mieux que les physiciens pour concevoir des expériences de science quantique".
- "Cette chercheuse a étudié 400,000 tricoteurs et a découvert ce qui transforme un hobby en entreprise"
- "Amazon met au rebut un outil secret de recrutement d'IA qui montrait des préjugés envers les femmes"
- "Des documents de Facebook saisis par des députés enquêtant sur une violation de la vie privée"
- Une entreprise dirigée par des vétérans de Google utilise l'IA pour "pousser" les travailleurs vers le bonheur".
- "Chez Netflix, qui gagne quand c'est Hollywood contre l'algorithme ?"

Les données dans l'actualité

- "AlphaGo vainc le meilleur joueur de Go du monde, marquant la supériorité de l'IA sur l'esprit humain"
- "Une novella écrite par l'IA a presque gagné un prix littéraire"
- "Elon Musk : l'intelligence artificielle peut déclencher une troisième guerre mondiale"
- "L'engouement pour l'I.A. a atteint son apogée, alors quelle sera la prochaine étape ?"

Les opinions sur le sujet sont variées - pour certains, SD/AA/IA fournissent des exemples de **réussites brillantes**, tandis que pour d'autres, ce sont les **échecs dangereux** qui sont au premier plan. Qu'en pensez-vous ?

Êtes-vous du genre à voir le verre à moitié plein ou le verre à moitié vide, cf. données et d'applications ?

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

Referees are
three times as
likely to give red
cards to
dark-skinned
players

Statistically
significant results
showing referees are
more likely to give red
cards to dark-skinned
players

Twice as likely

ONE RESEARCH TEAM

95% CONFIDENCE INTERVAL

Equally likely

Non-significant
results

Session 1

Lectures suggérées

Les préliminaires

*Data Understanding, Data Analysis, Data Science
Data Science Basics*

Introduction

- What is Data?
- From Objects and Attributes to Datasets
- Data in the News
- The Analog/Digital Data Dichotomy

Exercices

Les préliminaires

1. Trouvez des exemples d'articles récents sur "Les données dans l'actualité". S'agit-il de réussites ou d'échecs ? Quelles conséquences sociales pourraient découler des technologies décrites dans ces articles ?

2. Dans quel format les données de votre organisation sont-elles disponibles ? Pouvez-vous y accéder facilement ? Sont-elles mises à jour régulièrement ? Existe-t-il des dictionnaires de données ? Les avez-vous lus ?



3. Les cadres conceptuels

Les cadres conceptuels

Nous utilisons des données pour représenter le monde, mais aussi afin de :

- décrire le monde à l'aide du **langage**
- le représenter en construisant des **modèles physiques**

Fil conducteur : la **représentation** (un objet qui en remplace un autre, qui est utilisé à sa place afin de s'engager indirectement avec l'objet représenté).

“La carte n'est pas le territoire”, c'est vrai, mais nous n'avons pas besoin de beaucoup d'efforts pour utiliser la carte afin de naviguer le territoire.

La transition entre la **représentation** et le **représenté** peut se faire sans heurts, ce qui pose un risque : **confondre données/résultats analytiques et le monde réel**.

Les cadres conceptuels

Meilleure protection : **cadre conceptuel** réfléchi et décrit de manière explicite

- une **spécification des** parties du monde qui sont représentées
- **comment** ils sont représentés
- la **nature de la relation** entre le représenté et le représentant
- **des stratégies appropriées et rigoureuses** pour appliquer les résultats de l'analyse qui est effectuée dans ce cadre de représentation

On pourrait repartir à zéro pour chaque nouveau projet, mais il existe des **cadres de modélisation** qui sont largement applicables à de nombreux phénomènes différents, qui peuvent s'adapter à des cas spécifiques.

Trois stratégies de modélisation

Il y a 3 **stratégies de modélisation** principales (non exclusives) qui peuvent être utilisées pour guider la spécification d'un phénomène ou d'un domaine :

- modélisation **mathématique**
- modélisation **informatique**
- modélisation de **systèmes**

Les deux premiers ont leur propre monde mathématique/numérique, distinct du monde tangible physique étudié par les chimistes, les biologistes, etc :

- utilisés pour décrire des phénomènes du monde réel en **établissant des parallèles** entre les propriétés des objets et en raisonnant par le biais de ces parallèles.

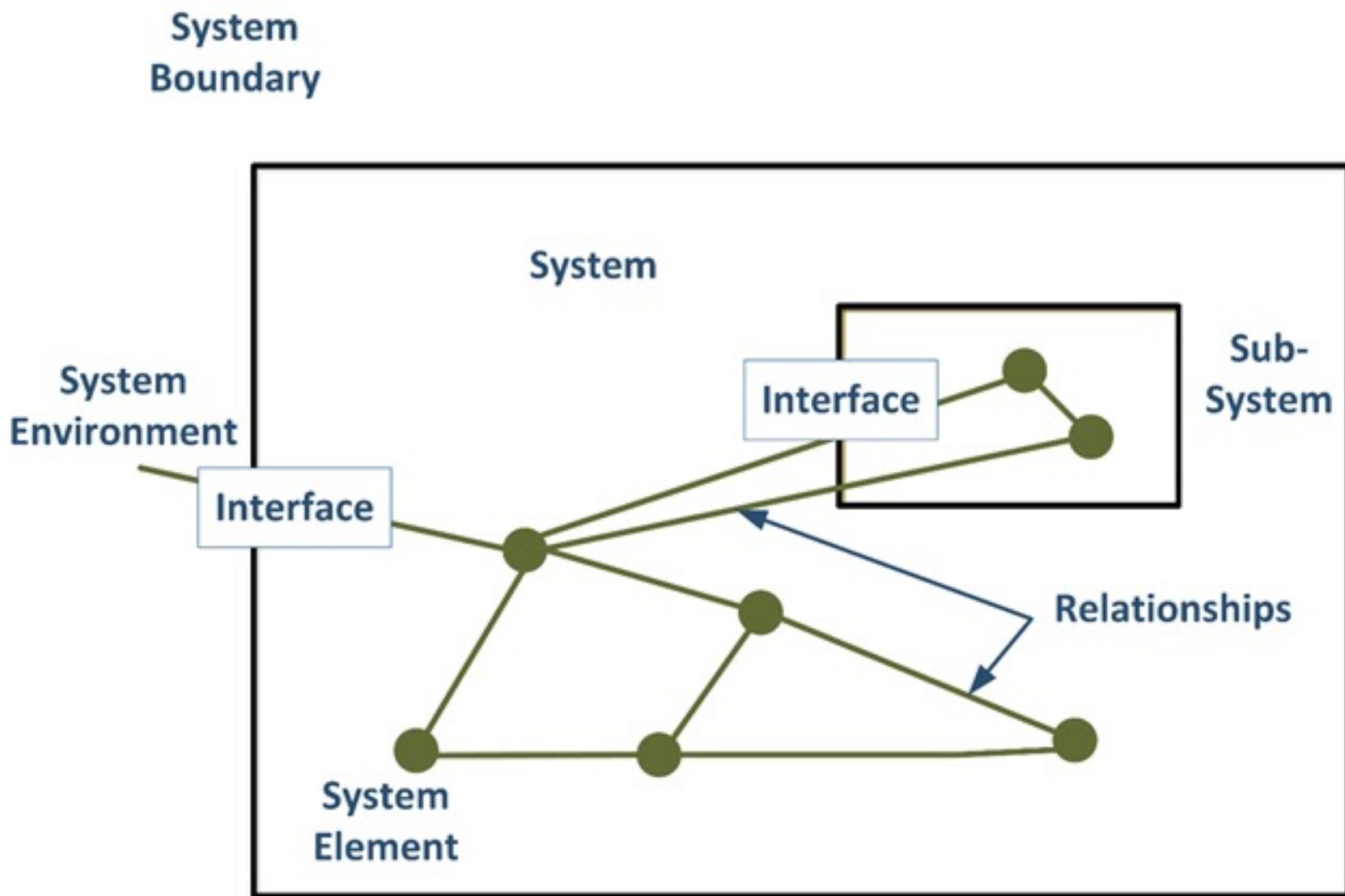
Trois stratégies de modélisation

La théorie générale des systèmes décrit des phénomènes naturels à l'aide d'un **cadre conceptuel commun**, tous étant des systèmes d'objets en interaction.

Lorsque nous sommes confrontés à une nouvelle situation, nous nous demandons :

- quels sont les objets qui semblent les plus pertinents dans les comportements du système ?
- quelles sont les propriétés de ces objets ?
- quels sont les comportements (ou actions) de ces objets ?
- quelles sont les relations entre ces objets ?
- comment les relations entre les objets influencent-elles leurs propriétés et leurs comportements ?

Objectifs : **comprendre le système**, développer une **compréhension commune cohérente**, informer la **collecte de données**, guider l'**interprétation des données**.



La collecte d'information

Il est crucial de parvenir à une **compréhension contextuelle** des données.

Concrètement, comment cette compréhension s'opère-t-elle ?

On peut l'obtenir par le biais :

- **d'excursions sur le terrain**
- des entretiens avec des **experts en la matière**
- de **lectures/visites**
- **d'exploration des données** (le simple fait d'**essayer d'obtenir** ou d'**accéder** aux données peut s'avérer très pénible), etc.

La collecte d'information

Les clients ou les parties prenantes **ne sont pas** des entités **uniformes** – les spécialistes des données des clients et les experts peuvent **ne pas apprécier l'implication** des analystes (externes et/ou internes).

La collecte d'informations donne aux analystes l'occasion de montrer que tout le monde tire dans la même direction, en :

- posant des questions **significatives**
- **s'intéressant véritablement** aux expériences des experts/clients
- reconnaissant la capacité de chacun à contribuer

Un peu de tact peut s'avérer utile lorsqu'il s'agit de recueillir des informations.

Penser en termes de systèmes

Un **système** est composé d'**objets** dont les **propriétés** peuvent changer au fil du temps.

Au sein du système, il y a des **actions/propriétés évolutives**, c-à-d des **processus**.

On comprend comment les différents aspects du monde interagissent ensemble en **découplant des morceaux** correspondant aux aspects et en définissant leurs limites.

Le travail avec d'autres intelligences requiert une **compréhension partagée** de ce qui est étudié.

Les objets eux-mêmes ont diverses propriétés.

Penser en termes de systèmes

Les processus naturels génèrent/détruisent des objets, et modifient les propriétés de ces objets au fil du temps.

Nous **observons**, **quantifions**, et **enregistrons** les valeurs de ces propriétés à des moments précis.

Les observations permettent de **saisir la réalité sous-jacente** avec un degré acceptable de **précision** et d'**erreur**, mais ... **même le meilleur modèle de système ne fournit jamais qu'une approximation de la situation analysée**.

Avec de la chance, de l'expérience, de la prévoyance, ces approximations peuvent être **valables**.

Identifier les lacunes de compréhension

Une **lacune dans les connaissances** est identifiée lorsque nous nous rendons compte que ce que nous pensions savoir sur un système s'avère **incomplet** (ou manifestement faux).

Causes :

- naïveté vis-à-vis de la situation modélisée
- la nature du projet envisagé

Avec **trop de parties mobiles**, des **objectifs irréalistes**, une **distance par rapport au pipeline**, les lacunes en matière de connaissances ne peuvent être évitées (même avec de petits projets bien organisés et faciles à contenir).

Identifier les lacunes de compréhension

Les lacunes en matière de connaissances peuvent survenir **à plusieurs reprises** :

- **nettoyage des** données
- **consolidation des** données
- **analyse des** données
- même pendant la **communication des résultats** (!)

Lorsque vous êtes confronté à un manque de connaissances, **soyez flexible** :

- **revenez en arrière**
- **posez des questions**
- **modifiez la représentation du système** aussi souvent que nécessaire

Il est préférable de combler ces lacunes dès le début du processus (évidemment).

Les modèles conceptuels

Les modèles conceptuels sont construits à l'aide d'outils d'investigation méthodiques :

- **diagrammes**
- **entretiens structurés**
- **des descriptions structurées**, etc.

Les scientifiques des données doivent se méfier des **modèles conceptuels implicites** (lacunes dans les connaissances).

Il est préférable de privilégier le côté du "trop de modélisation conceptuelle", mais n'oubliez pas que "tout modèle est faux ; certains modèles sont utiles" [G.E. Box].

Il est acceptable de construire de meilleurs modèles, de manière itérative.

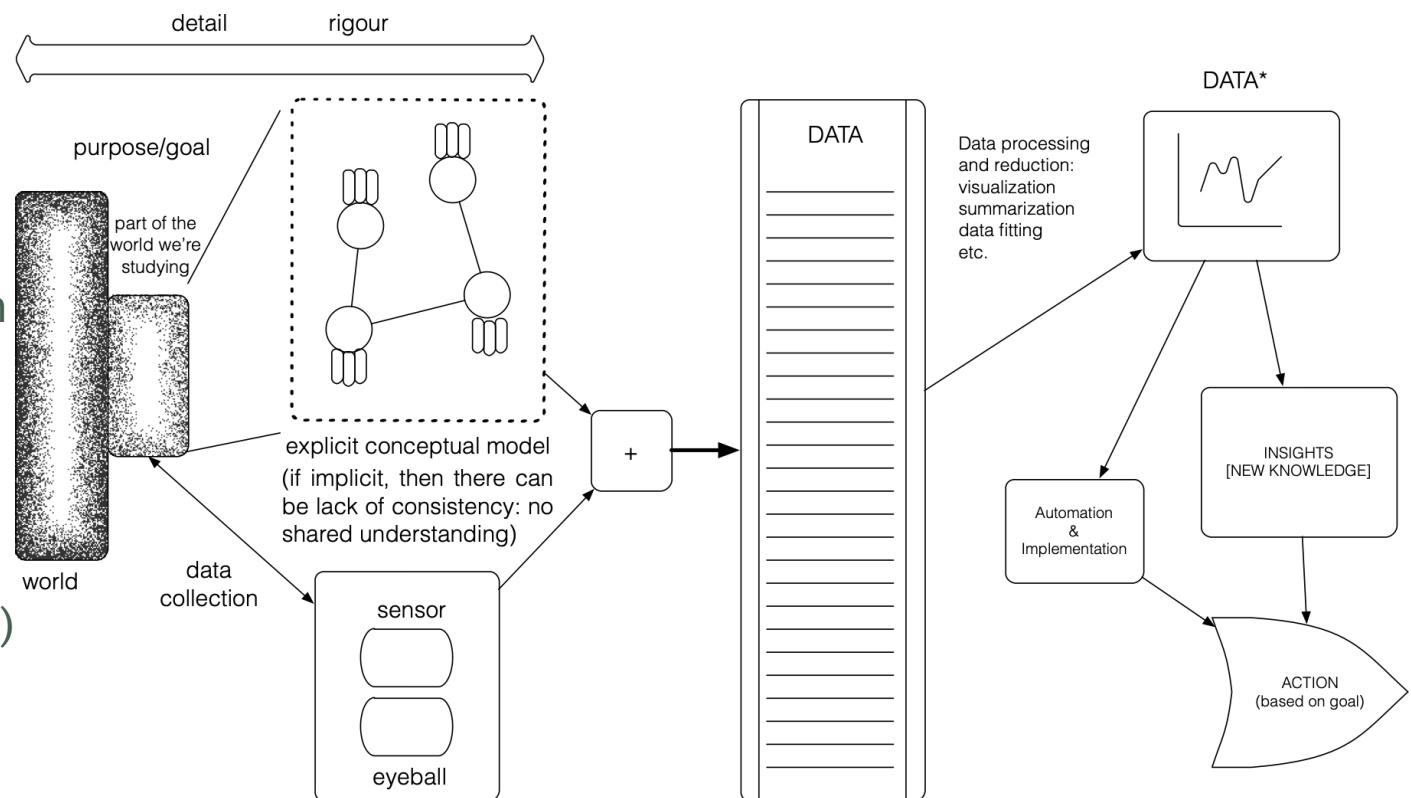
Les modèles conceptuels

Modèles conceptuels :

- ne sont pas mises en œuvre sous forme de modèle d'échelle ou de code informatique
- n'existent que de manière conceptuelle, souvent sous la forme d'un diagramme ou d'une description verbale d'un système – boîtes et flèches, cartes mentales, listes, définitions, etc.

L'accent est mis sur :

- les **états possibles** (pas de comportement spécifique)
- des types d'objets, et non des instances spécifiques ; l'objectif est l'**abstraction**

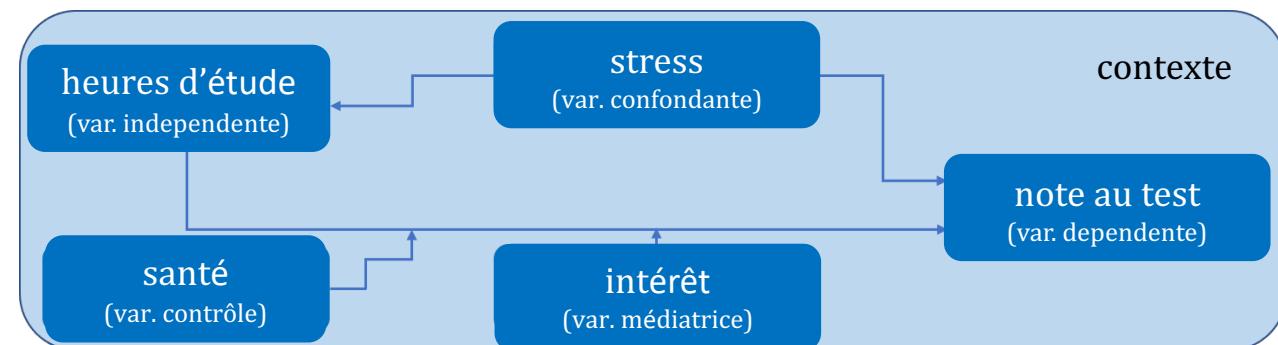


Les modèles conceptuels

En pratique, nous devons d'abord sélectionner un système pour la tâche à accomplir, puis générer un modèle conceptuel qui englobe :

- des **objets pertinents** et **clés** (abstraits ou concrets) ;
- les **propriétés** de ces objets, et leurs valeurs ;
- les **relations entre les objets** (partie-tout, est-un, 1-à-plusieurs, etc.), et
- les **relations entre les propriétés** à travers les instances d'un type d'objet.

Voici un exemple simpliste décrivant une relation supposée entre une **cause présumée** (heures d'étude) et un **effet présumé** (note au test).



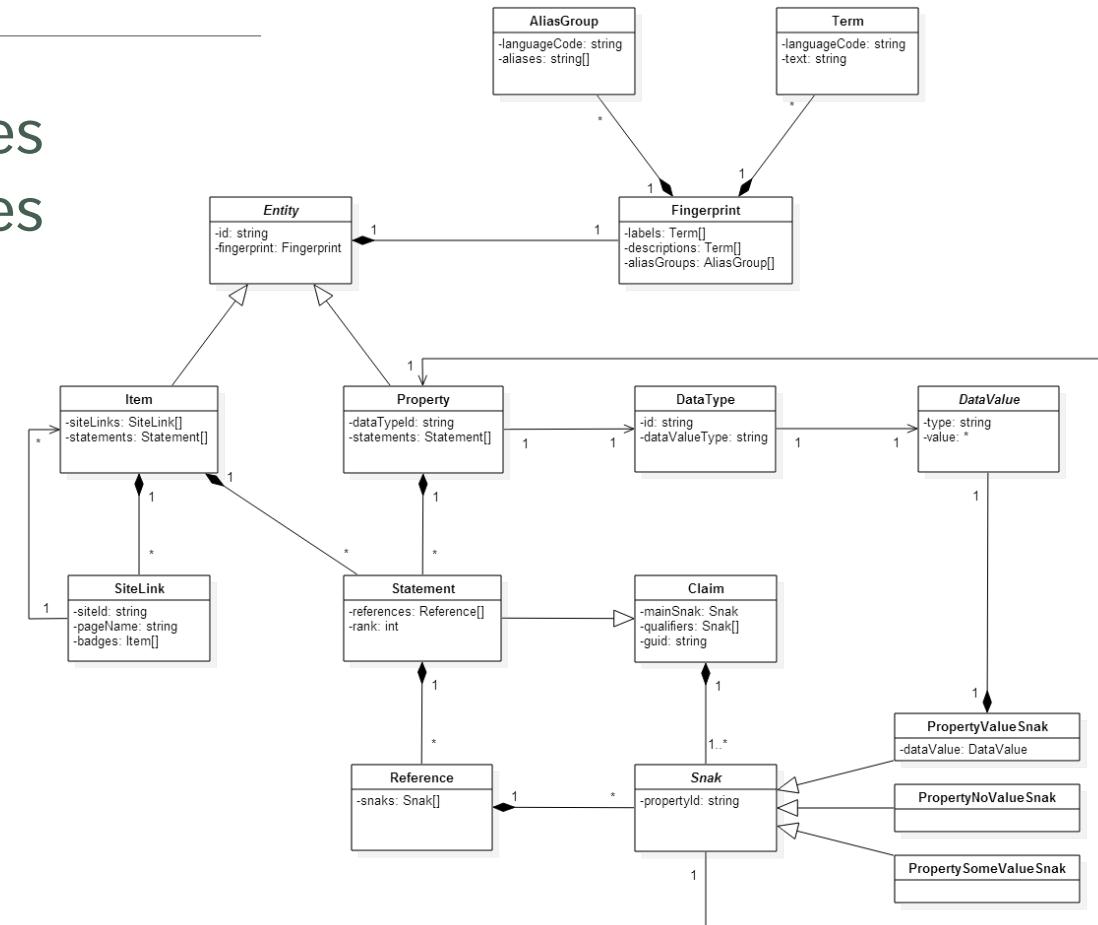
Les modèles conceptuels formels

La modélisation conceptuelle transforme les modèles conceptuels implicites en modèles **explicites et tangibles**.

Elle offre la possibilité d'**examiner** et d'**explorer** les idées et les hypothèses.

Divers efforts ont été déployés pour **formaliser** la modélisation conceptuelle :

- UML (langage universel de modélisation)
- modèles de relations entre entités (ER)



Relier les données au système

Les données collectées et analysées sont-elles **utiles pour comprendre le système** ? On peut mieux répondre à cette question si l'on comprend :

- **comment les** données sont collectées
- la **nature approximative** des données et du système
- ce que les données représentent (observations et caractéristiques)

La combinaison du système et des données est-elle **suffisante** pour comprendre la situation considérée ? Il est difficile de répondre en pratique.

Si les données, le système, et le monde réel ne sont **pas alignés**, tout aperçu des données tiré de la modélisation et de l'analyse pourrait s'avérer inutile.

Les biais cognitifs

Les biais cognitifs ont un impact sur la façon de construire des modèles et de rechercher des schémas dans les données :

- le **biais d'ancrage** nous amène à nous fier trop fortement à la première information que l'on nous donne sur un sujet
- l'**heuristique de disponibilité** décrit notre tendance à utiliser les informations qui nous viennent rapidement et facilement à l'esprit lorsque nous prenons des décisions
- l'**effet “bandwagon”** désigne notre habitude d'adopter des comportements ou des croyances parce que bcp d'autres personnes font de même

- le **biais d'appui du choix** nous mène à considérer nos actions sous un jour positif
- l'**illusion du regroupement** fait référence à notre tendance à voir des schémas dans l'aléatoire
- le **biais de confirmation** décrit notre tendance à remarquer et à accorder plus de crédit aux preuves qui appuient nos croyances existantes
- le **biais de conservation** se produit lorsque nous privilégions les preuves antérieures par rapport aux nouvelles informations
- l'**effet de l'autruche** décrit la façon dont les gens évitent souvent les informations négatives, y compris les commentaires qui les aident à suivre la progression de leurs objectifs

Les biais cognitifs

- le **biais lié aux résultats** consiste à juger une décision en fonction du résultat, plutôt que de la raison pour laquelle elle a été prise
 - l'**excès de confiance** nous pousse à prendre plus de risques dans notre vie quotidienne
 - le **biais pro-innovation** se produit lorsque les partisans d'une technologie sur-évaluent son utilité et sous-évaluent ses limites
 - le **biais de récence** se produit lorsque nous favorisons les nouvelles informations par rapport aux preuves antérieures
 - le **biais du risque zéro** est lié à notre préférence pour la certitude absolue
 - le **biais de survie** est un raccourci cognitif qui se produit lorsqu'un sous-groupe visible ayant réussi est pris pour un groupe entier
 - le **biais de saillance** décrit notre tendance à nous concentrer sur les éléments ou les informations les plus remarquables et à ignorer ceux qui n'attirent pas notre attention.
- Autres biais :**
- sophisme du taux de base, biais de la rationalité limitée, biais de la taille des catégories, effet Dunning-Kruger, effet de cadrage, sophisme de la main chaude, effet IKEA, illusion de validité, corrélations illusoires, etc.

Lectures suggérées

Les cadres conceptuels

Data Understanding, Data Analysis, Data Science Data Science Basics

Conceptual Frameworks for Data Work

- Three Modeling Strategies
- Information Gathering
- Cognitive Biases

Exercices

Les cadres conceptuels

1. Considérez la situation suivante : vous êtes en voyage d'affaires et vous avez oublié de remettre un dessin d'architecture très important (et requis de toute urgence) à votre superviseur avant de partir. Votre bureau enverra un stagiaire pour le récupérer dans votre espace de vie. Comment allez-vous lui expliquer, par téléphone, comment trouver le document ? Si le stagiaire est déjà venu dans votre espace de vie, si son espace de vie est comparable au vôtre, ou si votre conjoint est à la maison, le processus peut être considérablement accéléré, mais avec quelqu'un pour qui l'espace est nouveau (ou une personne ayant une déficience visuelle, par exemple), il est facile de voir comment les choses pourraient se compliquer. Le temps est un facteur essentiel - vous et le stagiaire devez faire le travail **correctement** et le plus **rapidement possible**. Quelle est votre stratégie ?
2. Traduisez les biais cognitifs en contextes analytiques. Quels sont les biais cognitifs auxquels vous, votre équipe et votre organisation êtes les plus sensibles ? Le moins ?

Session 2

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

Data ethics is in each step
of the data product life cycle.



Funding



Motivation

Project
DesignData Collection
& Sourcing

Analysis



Interpretation

Communication
& Distribution

4. L'éthique de la science des données

La nécessité de l'éthique

Dans la plupart des disciplines empiriques, **l'éthique** est introduite tôt dans le processus éducatif et finit par jouer un rôle crucial dans les activités des chercheurs.

Les scientifiques des données qui arrivent dans le domaine par le biais des mathématiques, des statistiques, de l'informatique, de l'économie, ou de l'ingénierie sont toutefois moins susceptibles d'avoir rencontré des comités de recherche éthique ou une **formation formelle en éthique**.

Les discussions sur les questions d'éthique sont souvent **mises de côté** au profit de considérations techniques ou administratives urgentes lorsque les délais sont serrés.

Mais cette échéance est remplacée par une autre échéance, puis par une autre, et ainsi de suite, le résultat final étant que la conversation **peut ne jamais avoir lieu**.

La nécessité de l'éthique

Lorsque la collecte de données à grande échelle devient possible, elle est accompagné d'une mentalité "Far West" : **tout est permis tant que faisable.**

La science des données moderne a des **codes de conduite professionnels**

- décrivant des façons **responsables** de pratiquer la science des données
- légitime plutôt que frauduleuse, éthique plutôt que contraire à l'éthique

Cela confère une **responsabilité supplémentaire** aux scientifiques des données, mais offre une **protection** contre les clients/employeurs qui veulent qu'ils effectuent des analyses de manière douteuse.

La nécessité de l'éthique

L'accent mis sur l'éthique des données récemment ne semble pourtant pas avoir ralenti les brèches :

- Volkswagen
- Whole Foods Markets
- General Motors
- Cambridge Analytica
- Amazon
- Ashley Madison

Qu'est-ce que l'éthique ?

L'éthique fait référence à l'étude et à la définition des **bonnes** et des **mauvaises** conduites :

- en général
- appliqué dans des circonstances spécifiques

L'éthique n'est pas (nécessairement) la même chose que :

- convention sociale
- convictions religieuses
- lois

Qu'est-ce que l'éthique ?

En Occident, les théories éthiques sont utilisées pour encadrer les débats autour des questions éthiques :

- **règle d'or** : faites aux autres ce que vous voudriez qu'ils vous fassent ;
- **conséquentialisme** : la fin justifie les moyens ;
- **utilitarisme** : agir de manière à maximiser l'effet positif ;
- **droits moraux** : agir pour maintenir et protéger les droits et priviléges fondamentaux des personnes affectées par les actions ;
- **justice** : répartir les avantages et les préjudices entre les parties prenantes de manière juste, équitable et impartiale.

Qu'est-ce que l'éthique ?

Il y a une grande variété de codes/cultures éthiques, notamment :

- Confucianisme
- Taoïsme
- Bouddhisme
- Ubuntu
- Te Ara Tika (Maori)
- etc.

Il est facile d'imaginer des contextes dans lesquels l'un de ces éléments serait mieux adapté à la tâche à accomplir – **renseignez-vous**.

L'éthique et science des données

Comment ces théories éthiques peuvent-elles s'appliquer à l'analyse des données ?

- qui, le cas échéant, est **propriétaire des données** ?
- y a-t-il des **limites** à l'utilisation des données ?
- certaines analyses comportent-elles des **biais de valeur** ?
- y a-t-il des catégories qui ne devraient jamais être utilisées dans l'**analyse des données personnelles** ?
- les données doivent-elles être accessibles **publiquement** ?

Les réponses dépendent d'un certain nombre de facteurs. Pour vous donner une idée de certaines des complexités, posons la première question : *qui, le cas échéant, est propriétaire des données* ?

L'éthique et science des données

Est-ce que ce sont les **analystes de données** qui transforment le potentiel des données en informations exploitables ?

Est-ce que ce sont les **collecteurs de données** qui ont une copie et rendent le travail possible ?

Sont-ce les **commendantaires** ou les **employeurs** qui ont rendu le processus viable ?

Dans certains cas, la **loi** peut également intervenir.

Il n'est pas facile de répondre à cette question simple ; il faut s'y prendre au cas par cas.

Vérité cachée : l'**analyse des données ne se limite pas à l'analyse des données**.

L'éthique et science des données

Défi similaire pour les **données ouvertes** (les "pro" et les "anti" ont de solides arguments).

Principe général de l'analyse des données : éviter l'**anecdotique** pour le **general** (se concentrer sur des observations spécifiques peut masquer la vue d'ensemble).

Mais les données **ne sont pas seulement** des marques sur le papier ou des octets sur le "cloud". Les décisions prises sur la base de la science des données peuvent **affecter des gens/la planète de manière négative**. On ne peut ignorer que les individus périphériques et les groupes minoritaires souffrent souvent de manière disproportionnée aux mains des décisions dites "fondées sur l'evidence".

[Principes de PCAP](#) (propriété, contrôle, accès, possession) [des Premières Nations](#).

Les meilleures pratiques

"Ne faites pas de tort" : les données recueillies auprès d'un individu **ne doivent pas être utilisées pour lui nuire.**

Consentement éclairé :

- les individus doivent **accepter la collecte et l'utilisation** de leurs données
- les individus doivent avoir une **réelle compréhension de ce à quoi ils consentent**, et des **conséquences possibles** pour eux et pour les autres.

Respecter la "vie privée" : excessivement difficile à maintenir à l'ère du "scraping" constant de l'Internet pour recueillir des données personnelles.

Meilleures pratiques

Les données doivent être gardées **publiques** (toutes ? la plupart ?).

Opt-In/Opt-Out : le consentement éclairé exige la possibilité de **se désengager**

Anonymiser les données : suppression des champs d'identification des données avant l'analyse.

"Laissez parler les données" :

- pas de sélection à la carte
- l'importance de la validation
- corrélation vs. causalité
- répétabilité

Le bon, la brute, et le truand

Les projets de données pourraient être classés de façon fantaisiste comme **bons**, **mauvais**, ou encore **laids**, soit d'un point de vue technique, soit d'un point de vue éthique (ou les deux).

- les **bons** projets accroissent les connaissances, peuvent aider à découvrir des liens cachés, etc., de la manière la plus inoffensive possible
- les **mauvais** projets peuvent conduire à de mauvaises décisions, qui peuvent à leur tour diminuer la confiance du public et potentiellement nuire à certains individus
- les projets **moches** sont, carrément, des applications peu recommandables ; ils sont mal exécutés d'un point de vue technique, ou mettent beaucoup de personnes en danger ; ces projets (et les approches/études similaires) doivent être évités **à tout prix !**

Le bon, la brute, et le truand

Bons projets (?) :

- P. A. B. Bien Nicholas AND Rajpurkar, “Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet,” *PLOS Medicine*, vol. 15, no. 11, pp. 1–19, 2018, doi: [10.1371/journal.pmed.1002699](https://doi.org/10.1371/journal.pmed.1002699).
- BeauHD, “[Google AI claims 99 percent accuracy in metastatic breast cancer detection](#),” *Slashdot.com*, Oct. 2018.
- Columbia University Irving Medical Center, “[Data scientists find connections between birth month and health](#),” *Newswire.com*, Jun. 2015.

Le bon, la brute, et le truand

Mauvais projets (?) :

- Indiana University, “[Scientists use Instagram data to forecast top models at New York Fashion Week](#),” *Science Daily*, Sep. 2015.
- D. Wakabayashi, “[Firm led by Google veterans uses A.I. to ‘nudge’ workers toward happiness](#),” *New York Times*, Dec. 2018.
- N. Cohn, “[How one 19-year-old illinois man is distorting national polling averages](#),” *The Upshot*, 2016.

Le bon, la brute et le truand

Projets moches (?) :

- J. Dastin, “[Amazon scraps secret AI recruiting tool that showed bias against women](#),” *Reuters*, Oct. 2018.
- I. Johnston, “[AI robots learning racism, sexism and other prejudices from humans, study finds](#),” *The Independent*, Apr. 2017.
- M. Judge, “[Facial-recognition technology affects African-Americans more often](#),” *The Root*, 2016.
- M. Kosinski and Y. Wang, “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images,” *Journal of Personality and Social Psychology*, vol. 114, no. 2, pp. 246–257, Feb. 2018.

Lectures suggérées

L' éthique de la science des données

*Data Understanding, Data Analysis, Data Science
Data Science Basics*

Ethics in the Data Science Context

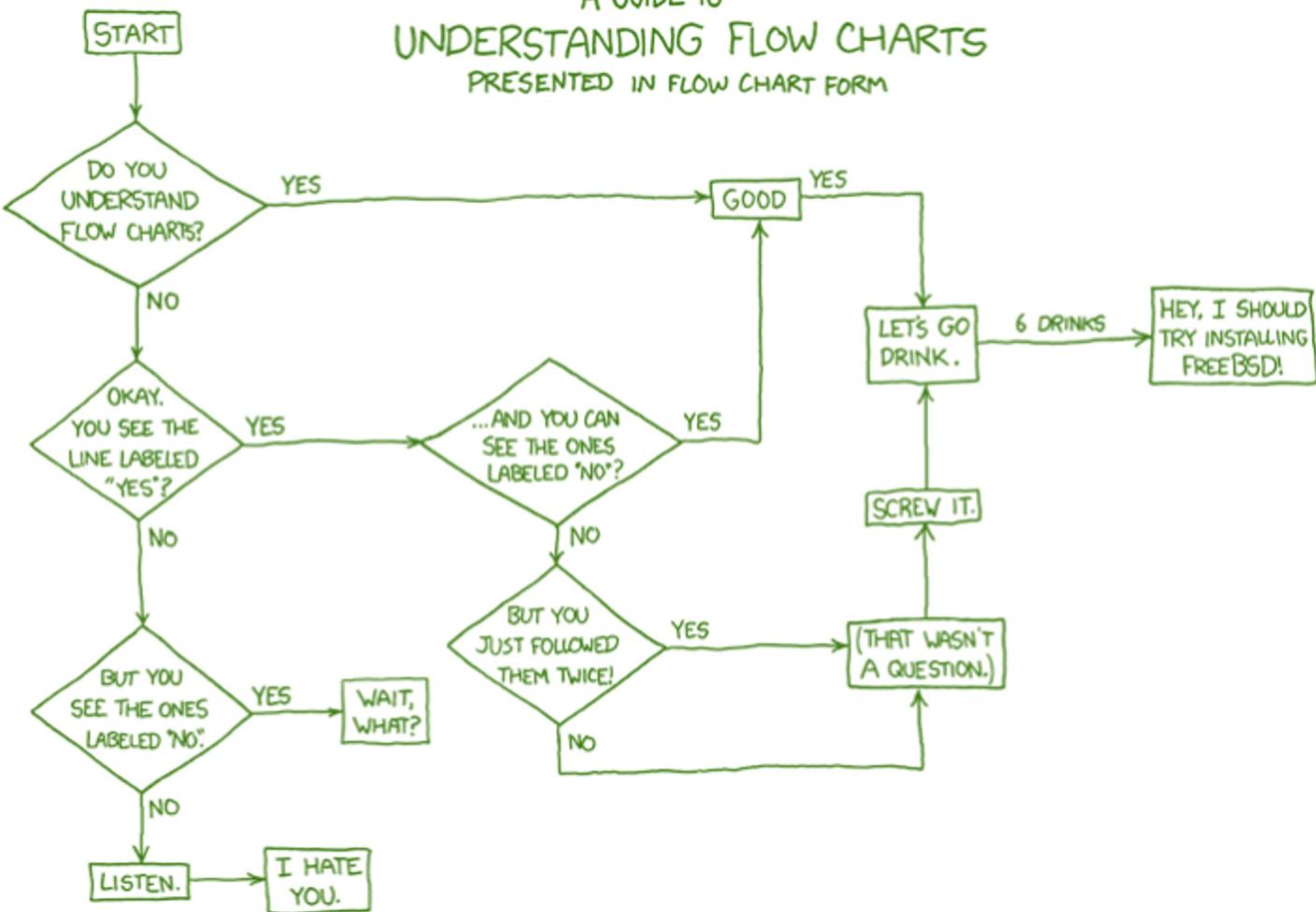
- The Need for Ethics
- What Is/Are Ethics?
- Ethics and Data Science
- Guiding Principles

Exercices

L'éthique de la science des données

1. Faites une recherche sur les récents scandales d'éthique des données impliquant Volkswagen, Amazon, Whole Foods Markets, Cambridge Analytica, Ashley Madison, General Motors ou toute autre organisation. Que s'est-il passé ? Qui a été affecté ? Quelles ont été les conséquences pour le grand public, l'organisation, la communauté des données ? Comment cela aurait-il pu être évité ?
2. Établissez une déclaration d'éthique pour votre travail sur les données. Y a-t-il des domaines sur lesquels vous n'acceptez pas de travailler ?

A GUIDE TO
UNDERSTANDING FLOW CHARTS
PRESENTED IN FLOW CHART FORM



5. Le flux de travail analytique

Le flux de travail analytique

Vous en avez probablement assez des **discussions sur le contexte** et préférerez passer à l'analyse des données proprement dite.

Une dernière chose : le **contexte du projet**.

La science des données ne se résume pas à l'analyse des données ; cela apparaît clairement lorsque l'on examine les étapes typiques d'un **projet de science des données**.

L'analyse a lieu dans un contexte de projet plus large, ainsi que dans le contexte d'une plus grande **infrastructure technique** ou d'un **système pré-existant**.

La méthode “analytique”

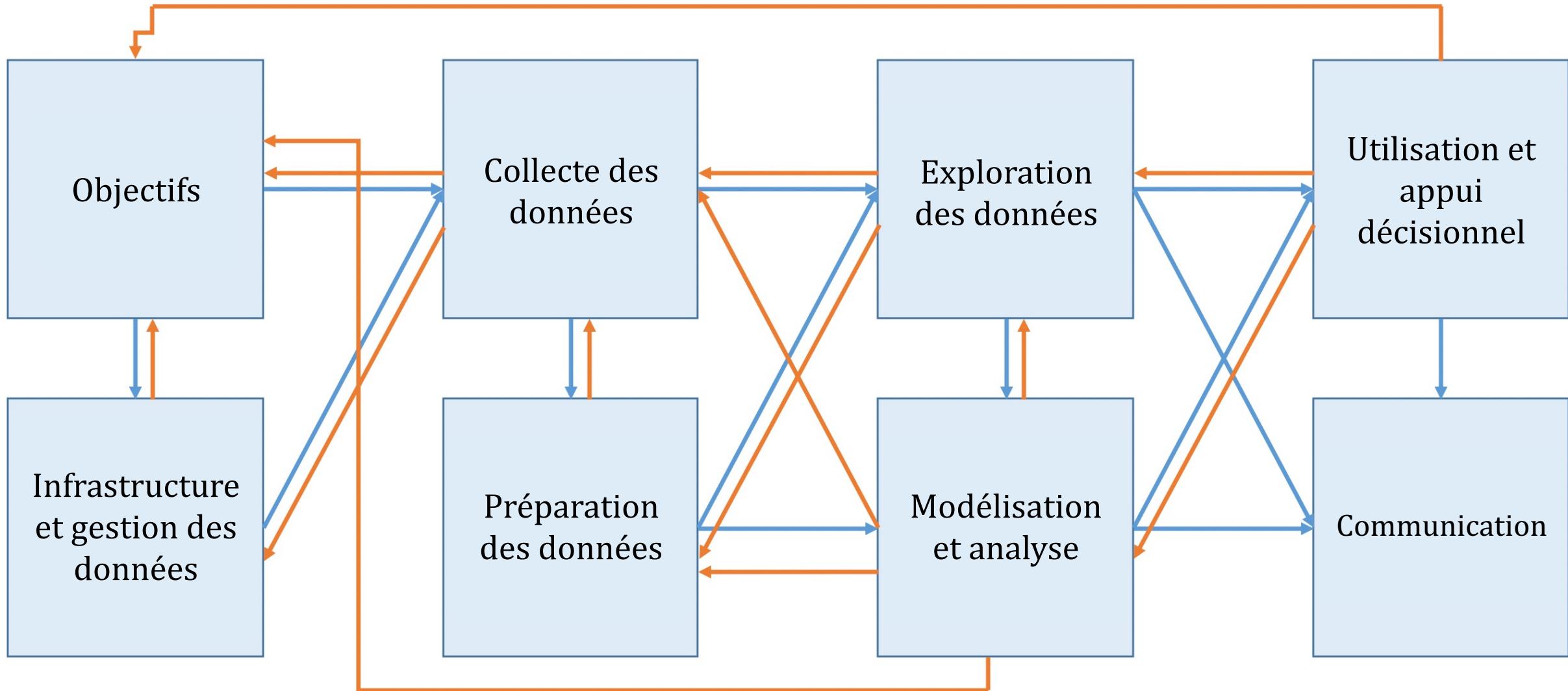
Comme c'est le cas pour la **méthode scientifique**, il existe un guide "étape par étape" pour l'analyse des données

- déclaration d'objectif
- collecte de données
- nettoyage des données
- analyse des données/analytique
- dissémination
- documentation

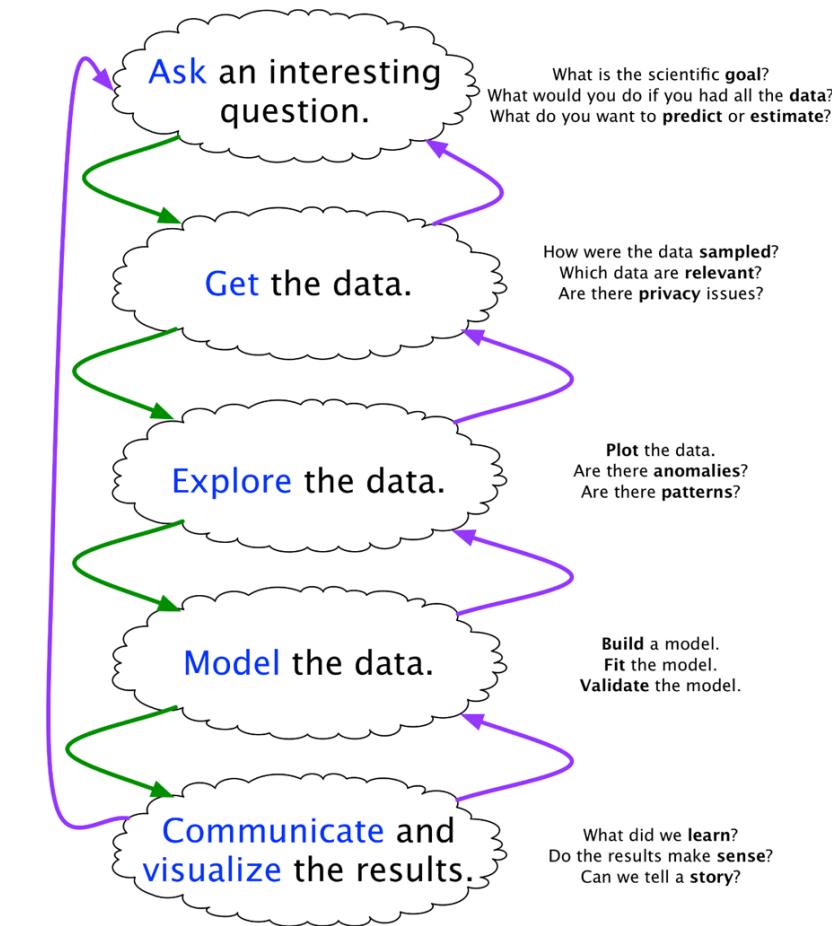
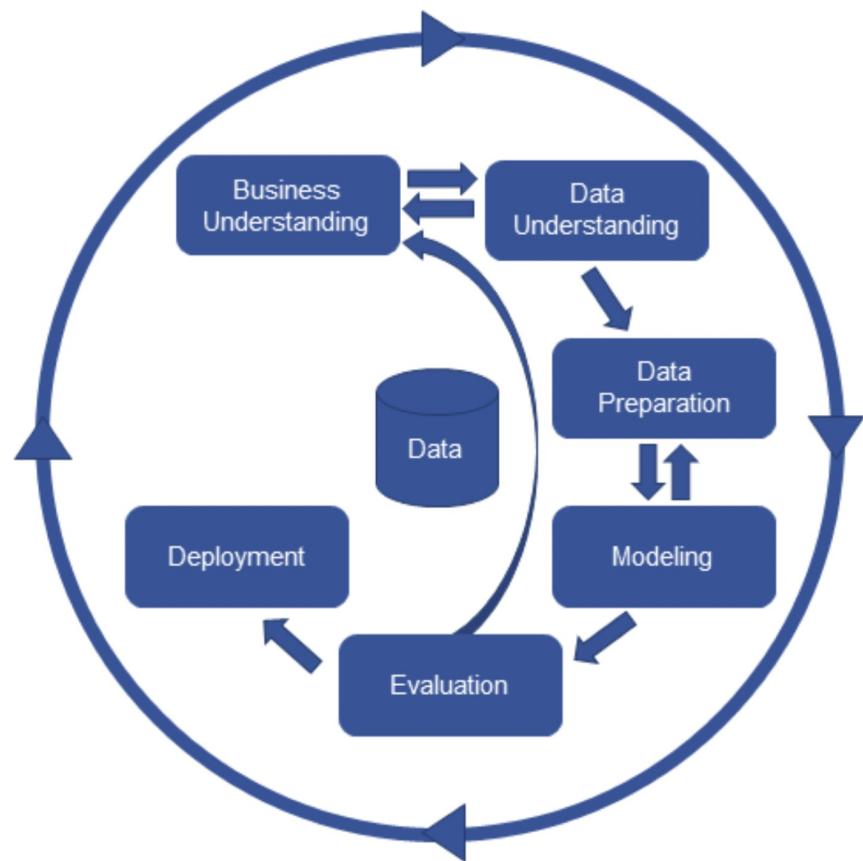
Notez que l'**analyse des données** ne constitue qu'un petit segment de l'ensemble du flux.

En pratique, le processus est souvent **désordonné** ; étapes ajoutées et retirées de la séquence, répétitions, reprises, etc.

Cela a tendance à fonctionner... quand **c'est mené correctement**.



La méthode “analytique”



La méthode “analytique”

En pratique, l'analyse des données est souvent corrompue par :

- le manque de clarté
- remaniement et travail inutile
- transfert aveugle vers TI
- pas d'itération

Les approches ont un noyau commun

- les projets sont **itératifs**
- (souvent) **non séquentiel**.

En aidant les parties prenantes à reconnaître cette **vérité centrale**, il est plus facile pour les scientifiques des données :

- d'obtenir des **informations utiles**

À retenir : il y a beaucoup de choses à prendre en compte avant la modélisation et l'analyse.

- **l'analyse des données ne se limite pas à l'analyse des données**

La collecte de données

Les données entrent dans le **pipeline de la science des données** en étant **collectées**.

Il existe plusieurs façons de procéder :

- les données peuvent être collectées en **un seul passage**
- elle peut être collectée par **lots** (“batches”)
- elle peut être collectée **en continu**

Le **mode d'entrée** peut avoir un impact sur les étapes suivantes, notamment sur la fréquence de **mise à jour des modèles**, des métriques, etc.

Le stockage des données

Une fois recueillies, les données doivent être **stockées**.

Les choix relatifs au stockage (et au **traitement**) doivent refléter :

- la manière dont les données sont recueillies (**mode d'entrée**)
- la quantité de données à stocker et à traiter (**petite ou grande**)
- le type d'accès/de traitement nécessaire (**quelle rapidité, quelle quantité, par qui**)

Les données stockées peuvent devenir **périmées** (*aux sens figuré et littéral*) ; il est recommandé de procéder à des audits réguliers des données.

Le traitement des données

Les données doivent être **traitées** avant de pouvoir être analysées.

Principalement, les **données brutes** doivent être converties dans un format qui **se prête à l'analyse**, en :

- identifiant les entrées **non valides, non fondées, et anormales**
- traitant les **valeurs manquantes**
- **transformant** les variables afin qu'elles répondent aux exigences des algorithmes choisis

L'**analyse** elle-même est presque anti-climatique : il suffit tout simplement d'exécuter les méthodes ou algorithmes sélectionnés sur les données traitées.

La modélisation

Les équipes de SD doivent connaître :

- le nettoyage des données
- les statistiques descriptives et la corrélation
- La probabilité et les statistiques inférentielles
- l'analyse de régression
- la classification et apprentissage supervisé
- le regroupement et appr. non supervisé
- la détection des anomalies et l'analyse des valeurs aberrantes
- les données massives/de hautes dimensions
- la modélisation stochastique, etc.

Cela ne représente qu'une **petite part** de l'analyse (cf. diapo précédente).

Aucun analyste ou scientifique des données ne peut tous les maîtriser (ou même une majorité d'entre eux) ; c'est l'une des raisons pour lesquelles la science des données est une **activité de groupe**.

Évaluation du modèle

Avant d'appliquer les résultats, nous devons d'abord confirmer que le modèle aboutit à des conclusions valables sur le système qui nous intéresse.

Les processus analytiques sont **réducteurs** : les données brutes sont transformées en **résumé numérique**, que nous espérons **lié** au système.

Les méthodologies de SD comprennent une **phase d'évaluation**

- contrôle “d'hygiène analytique” : y a-t-il quelque chose **qui cloche** ?

Méfiez-vous de la **tyrannie des succès précédents** : même si une approche a donné des réponses utiles par le passé, elle peut ne pas toujours le faire.

Le monde réel



Modèle



Théorie

Identification des détails pertinents pour la **description** et la **traduction** des objets du monde réel en variables de modèle

L'analyse de la vie après le modèle

Lorsqu'une analyse ou un modèle est "lâché dans la nature", il prend souvent une vie qui lui est propre. Lorsqu'il cesse inévitablement d'être **actuel**, les SD ne peuvent pas toujours faire grand-chose pour remédier à la situation.

Comment déterminer si le modèle de données actuel est :

- **démodé** ?
- n'est plus **utile** ?
- combien de temps faut-il à un modèle pour réagir à un **changement conceptuel** ?

Des audits réguliers peuvent être utilisés pour répondre à ces questions.

L'analyse de la vie après le modèle

Les SD ont rarement le contrôle total de la **diffusion des modèles**.

- les résultats peuvent être détournés, mal compris, mis de côté, ou ne pas être mis à jour
- les analystes consciencieux peuvent-ils faire quelque chose pour l'empêcher ?

Il n'y a pas de réponse facile : on ne doit pas seulement se concentrer sur l'analyse, mais aussi reconnaître les opportunités qui se présentent pour **éduquer les parties prenantes** sur l'importance des étapes auxiliaires.

En raison de la **déclin analytique**, la dernière étape du processus analytique n'est pas une **impasse**, mais une invitation à retourner au début du processus.

Pipelines de données

Dans le **contexte de la prestation de services**, le processus d'analyse des données est mis en œuvre sous forme de **pipeline de données automatisé** pour permettre des exécutions automatiques.

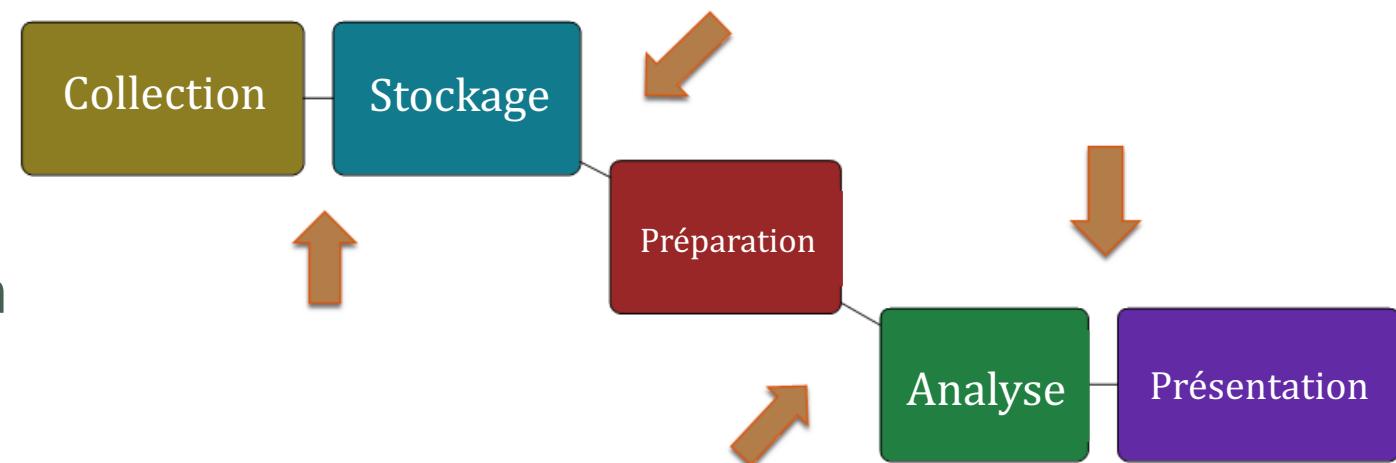
Les pipelines de données se composent généralement de 9 éléments (**5 étapes et 4 transitions**) :

- collecte de données
- stockage de données
- préparation des données
- analyse des données
- présentation des données

Pipelines de données

Chaque composant doit être **conçu** et ensuite **mis en œuvre**.

Généralement, au moins une passe d'analyse des données doit être effectué **manuellement** avant que l'implementation ne soit terminée.



Lectures suggérées

Le flux de travail analytique

Data Understanding, Data Analysis, Data Science Data Science Basics

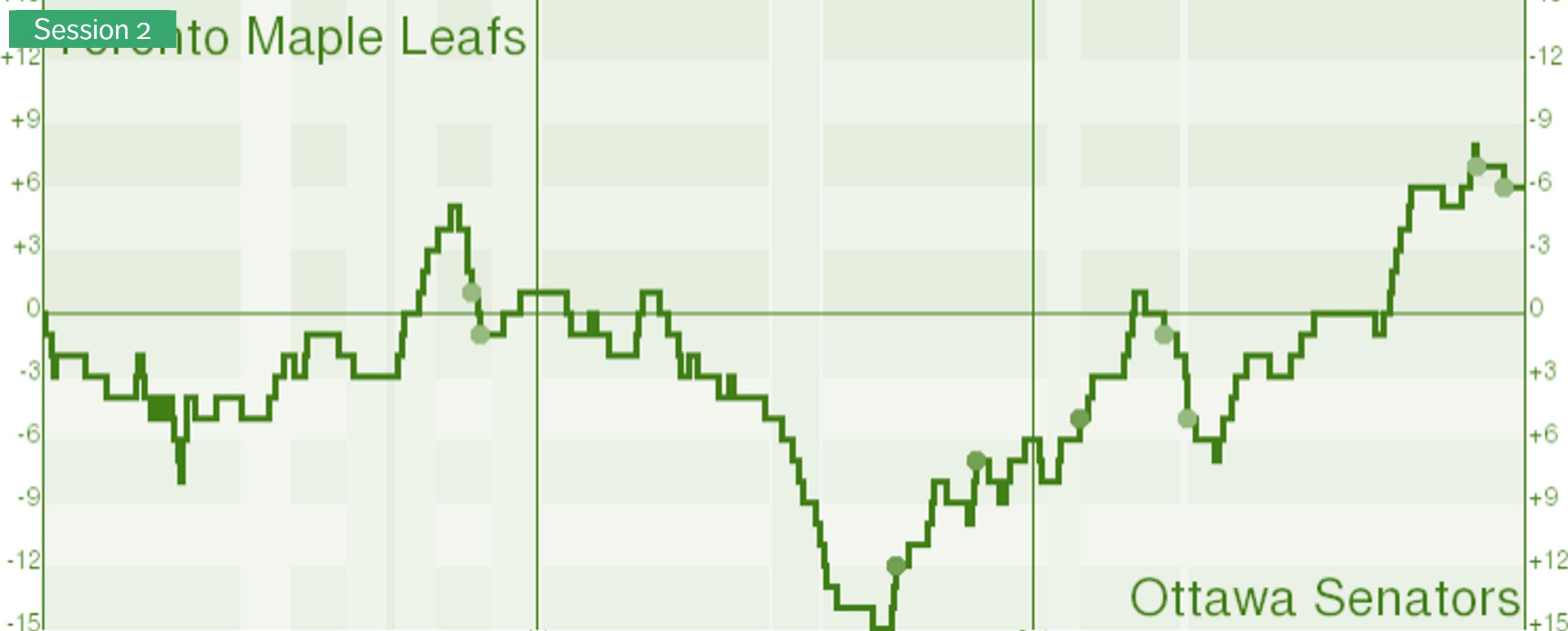
Analytics Workflows

- The "Analytical" Methods
- Data Collection, Storage, Processing, and Modeling
- Model Assessment and Life After Analysis
- Automated Data Pipelines

Exercices

Le flux de travail analytique

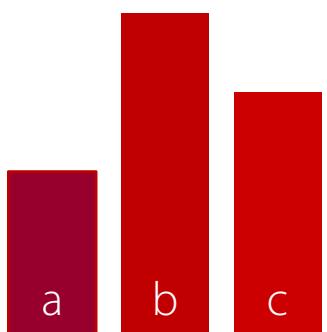
1. Installez [R / RStudio](#) (Posit), et les librairies de la liste fournie par l'instructeur.
2. Testez l'installation à l'aide des exemples du [Programming Primer](#) (sections 2 - 4) pour vous assurer que le logiciel fonctionne comme prévu.



6. Les données et les renseignements

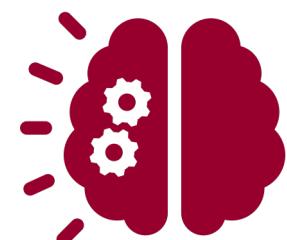
Modes d'analyse

Descriptive



Montrer **ce qui s'est passé**

Diagnostique



Expliquer **pourquoi** quelque chose s'est produit

Préditive



Deviner **ce qui va se passer**

Prescriptive



Suggérer **ce qui devrait se passer**

Valeur faible
Faible difficulté



Valeur élevée
Difficulté élevée



Poser les bonnes questions

La science des données consiste à poser des questions et à y répondre :

- **Analytique** : "Combien de clics ce lien a-t-il obtenu ?"
- **La science des données** : "Sur la base de l'historique des achats précédents de cet utilisateur, puis-je prédire sur quels liens il va cliquer lors de son prochain accès au site ?"

Les modèles d'exploration de données/sciences sont généralement **prédictifs** (et non **explicatifs**) : ils montrent des connexions, mais ne révèlent pas **pourquoi** elles existent.

Attention : toutes les situations ne font pas appel à la science des données, à l'intelligence artificielle, à l'apprentissage automatique, aux statistiques, etc.

Les mauvaises questions

Trop souvent, les analystes posent les **mauvaises questions** :

- des questions **trop larges** ou **trop étroites**
- des questions **auxquelles aucune quantité de données ne pourra jamais répondre**
- les questions pour lesquelles **des données ne peuvent être obtenues**

Dans le meilleur des cas, les parties prenantes reconnaîtront que les réponses ne sont pas pertinentes.

Le **pire scénario** est qu'ils mettent en œuvre par erreur des politiques ou prennent des décisions sur la base de réponses qui n'ont pas été identifiées comme trompeuses ou inutiles.

Feuille de route

Comprendre le problème (opportunité vs problème)

Quelles hypothèses initiales ai-je sur la situation ?

Comment les résultats seront-ils utilisés ?

Quels sont les risques et/ou les avantages de répondre à cette question ?

Quelles questions des parties prenantes pourraient être soulevées en fonction des réponses ?

Ai-je accès aux données nécessaires pour répondre à cette question ?

Comment vais-je mesurer mes critères de "réussite" ?

Le piège du Oui/Non

Exemples de **mauvaises** questions :

- Nos revenus **augmentent-ils** d'une année sur l'autre ?
- La plupart de nos clients appartiennent-ils à **cette catégorie démographique** ?
- **Ce projet a-t-il des** ambitions valables pour l'ensemble du département ?
- Est-ce que notre équipe de succès de la clientèle, qui travaille dur, est **formidable**.
- À quelle fréquence **vérifiez-vous par trois fois** votre travail ?

Exemples de **bonnes** questions :

- Quelle est la **répartition** de nos revenus au cours des trois derniers mois ?
- D'où viennent nos **5** cohortes **les plus** dépensières ?
- Que sont les **différents avantages** de la poursuite de ce projet ?
- Que **sont trois bons et trois mauvais traits de** notre équipe de réussite client ?
- Avez-vous **tendance à** effectuer des tests d'assurance qualité sur vos livrables ?

Liste de contrôle

1. Ai-je évité de créer des questions de type oui/non ?
2. Est-ce que tous les membres de mon équipe/département comprendraient la question, indépendamment de leurs antécédents ?
3. La question nécessite-t-elle plus d'une phrase pour être exprimée ?
4. La question est-elle "équilibrée" ? (champ d'application ni trop large pour une réponse, ni trop restreint au point de n'avoir qu'un impact minime)
5. La question est-elle orientée vers ce à quoi il est plus facile de répondre pour les compétences particulières de mon équipe ?

Contingence/Tableaux croisés

Tableau de contingence : examine la relation entre deux variables catégorielles

Tableau croisé dynamique : un tableau généré en appliquant des opérations (compte, moyenne, etc.) à des variables sur la base d'une autre variable.

Les tableaux de contingence sont des cas particuliers de tableaux croisés dynamiques (“pivot tables”).

	Large	Moyen	Petits
Fenêtre	1	32	31
Porte	14	11	0

Type	N	Signal moy	Signal ET
Bleu	4	4.04	0.98
Vert	1	4.93	N.A.
Orange	4	5.37	1.60

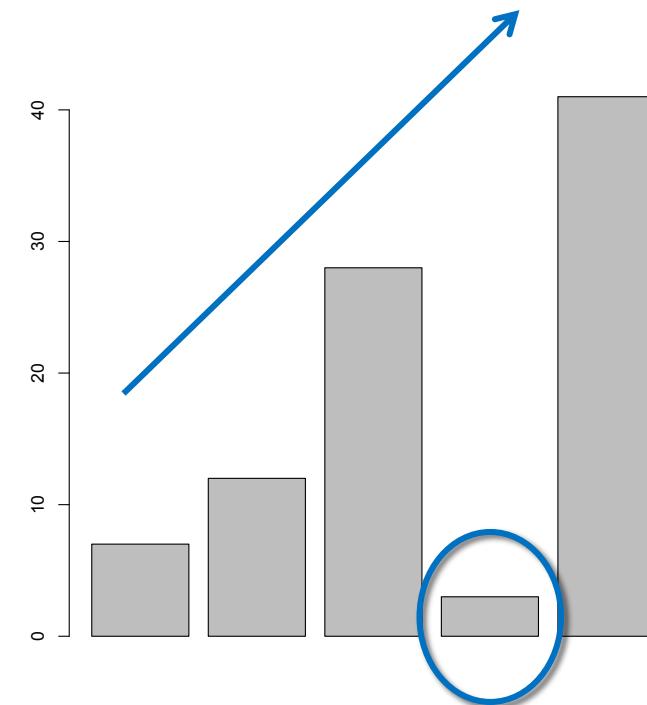
L'analyse par la visualisation

Analyse (au sens large) :

- identifier des modèles ou des structures
- ajouter du sens à ces modèles ou à cette structure en les interprétant dans le contexte du système.

Option 1 : utiliser des méthodes analytiques

Option 2 : visualiser les données et utiliser le pouvoir d'analyse du cerveau (perceptuel) pour tirer des conclusions significatives



Résumés numériques

Dans un premier temps, une variable peut être décrite selon 2 dimensions : la **centralité** et la **dispersion** (l'asymétrie et l'aplatissement sont aussi utilisés).

Les **mesures de centralité** comprennent :

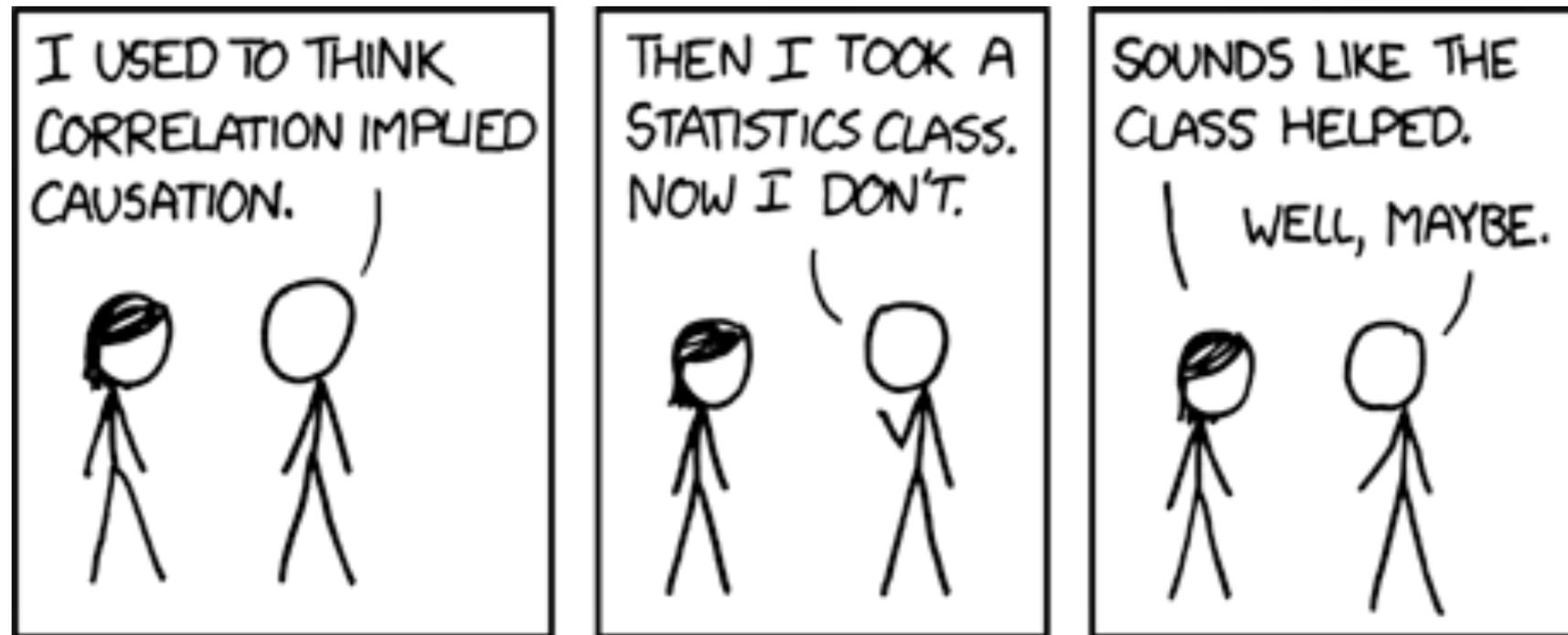
- médiane, moyenne, mode (moins fréquemment)

Les **mesures dispersion** (ou d'étalement) comprennent :

- écart-type (sd), variance, quartiles, écart interquartile (IQR), étendue (moins fréquemment)

La médiane, l'étendue, et les quartiles sont facilement calculés à partir de **listes ordonnées**.

Corrélation



La corrélation n'implique pas la causalité, mais elle agite les sourcils de manière suggestive et fait des gestes furtifs en disant "regardez par là".

Régression linéaire

L'hypothèse de base de la **régression linéaire** est que la variable dépendante peut être approximée par combinaison linéaire des variables indépendantes :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

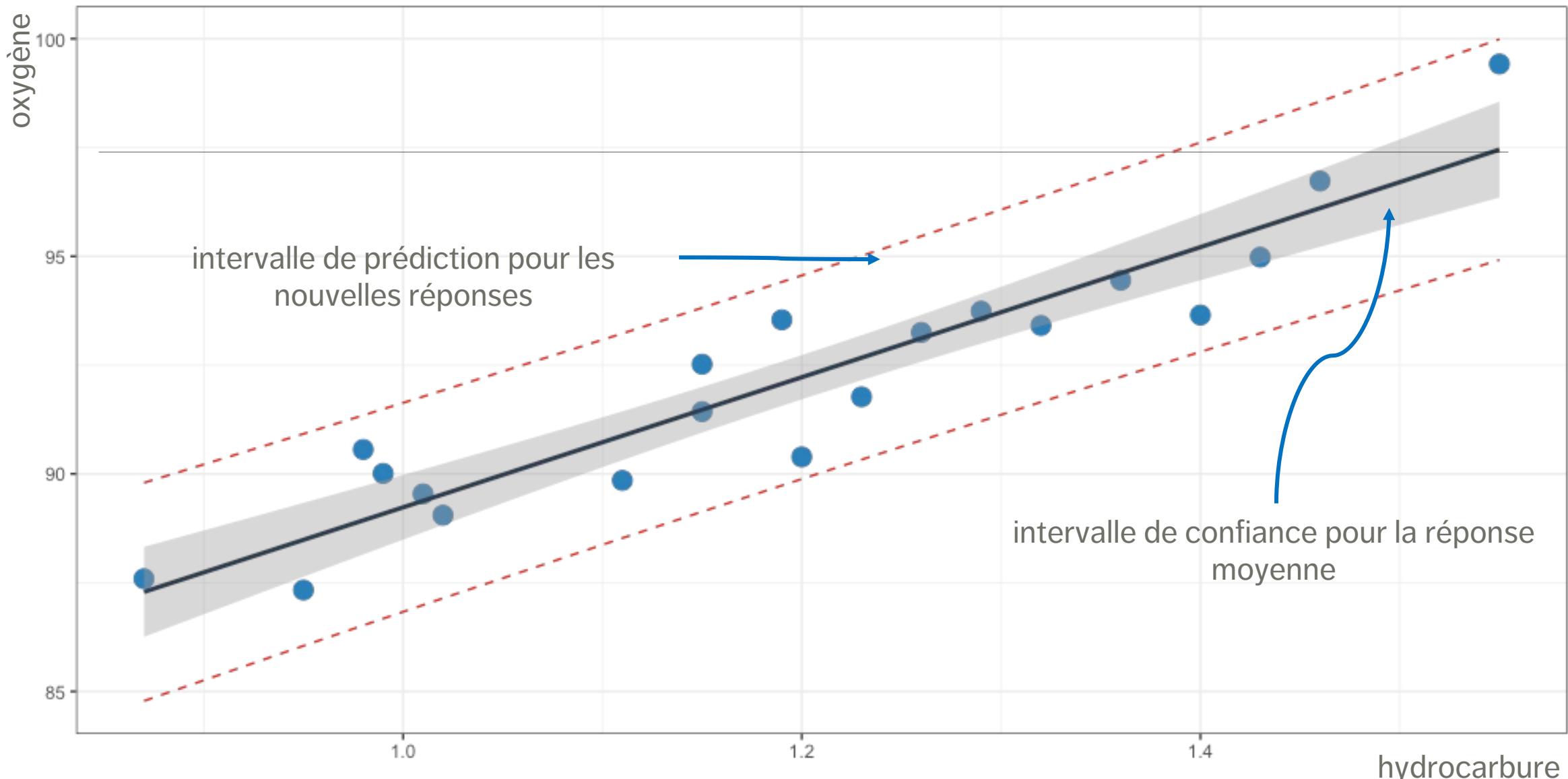
où $\boldsymbol{\beta} \in \mathbb{R}^p$ est déterminé sur la base de l'**ensemble d'apprentissage \mathbf{X}** , et

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \quad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Généralement, les erreurs sont **distribuées selon une normale** :

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

$$\text{oxygène} = 14.95 \times \text{hydrocarbure} + 74.28$$



Tâches d'apprentissage automatique

Classification et estimation de la probabilité de classe : quels clients sont susceptibles d'être des clients réguliers ?

Regroupement (“clustering”) : les clients forment-ils des groupes naturels ?

Règles d'association : quels livres sont couramment achetés ensemble ?

Autres :

profilage et description du comportement ; **prédiction des liens** ; **estimation de la valeur** (combien un client est-il susceptible de dépenser dans un restaurant) ; **mise en correspondance des similarités** (quels clients potentiels sont similaires aux meilleurs clients d'une entreprise ?); **réduction des données** ; **modélisation d'influence**, etc.

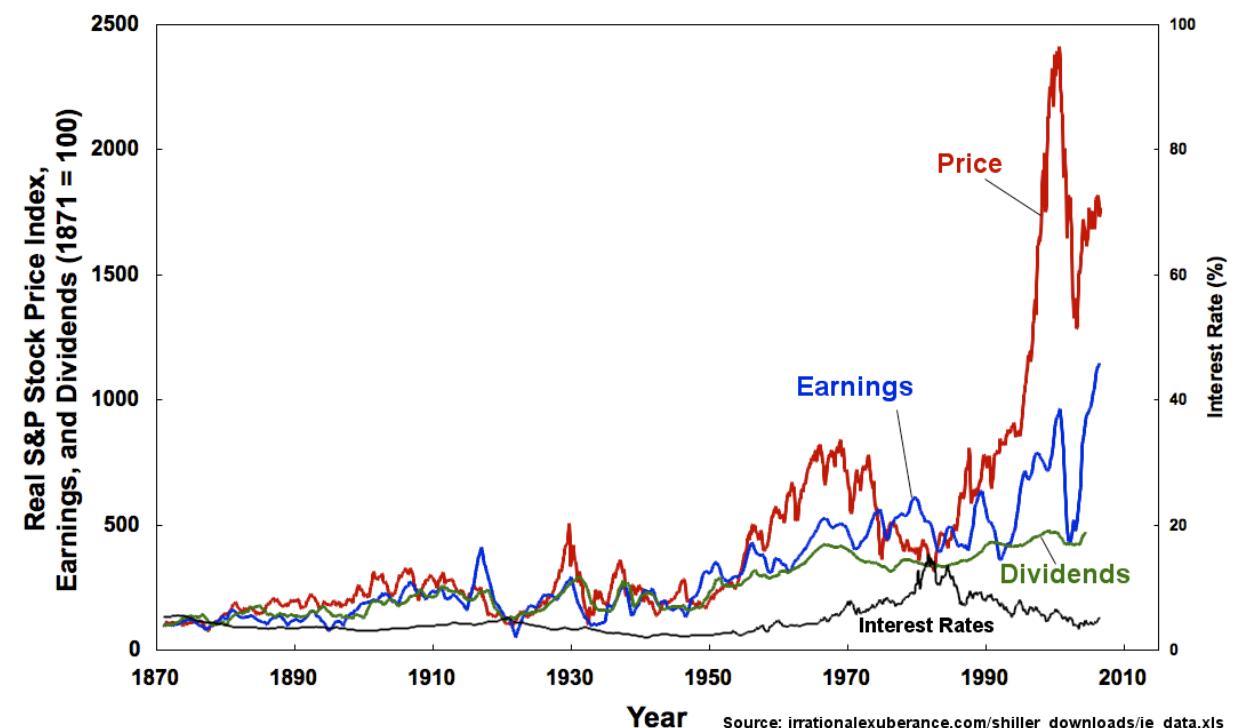
Analyse des séries temporelles

Une **série chronologique** simple :

- a deux variables : temps + 2nd variable
- la deuxième variable est *séquentielle*

Quel est le **comportement** de cette deuxième variable dans le temps ?

Pouvons-nous **prévoir** le **comportement futur** de la variable ?



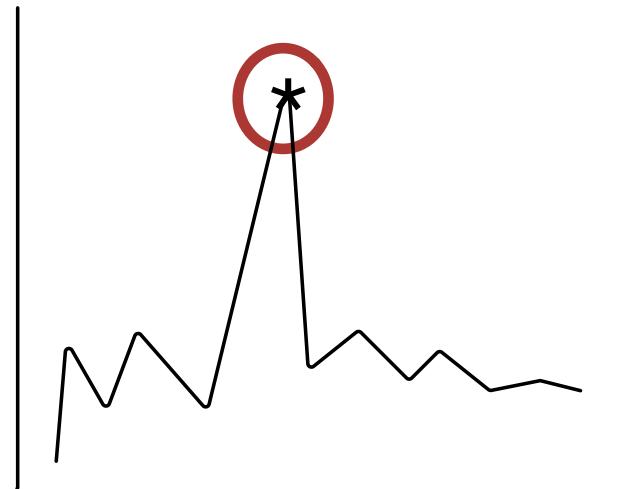
Détection d'anomalies

Anomalie : un événement inattendu, inhabituel, atypique, ou statistiquement improbable.

Ne serait-il pas utile d'avoir un pipeline d'analyse de données qui vous alerte lorsque les choses sortent de l'ordinaire ?

Il y a plusieurs approches analytiques à adopter !

- regroupement
- classification
- techniques d'ensemble, etc.



Lectures suggérées

Les données et les renseignements

Data Understanding, Data Analysis, Data Science Data Science Basics

Getting Insight From Data

- Asking the Right Questions
- Basic Data Analysis Techniques
- Common Statistical Procedures in R
- Quantitative Methods

*Probability and Applications (advanced)

*Introductory Statistical Analysis (advanced)

*Survey Sampling (advanced)

*Regression Analysis (coming soon)

Exercices

Les données et les renseignements

1. Faites l'exercice de la section [Asking the Right Questions](#).
2. Recréez les exemples de [Common Statistical Procedures in R](#).
3. Le fichier [cities.txt](#) contient des informations sur la population des villes d'un pays. Une ville est classée comme "petite" si sa population est inférieure à 75K, comme "moyenne" si elle se situe entre 75K et 1M, et comme "grande" autrement. Localisez et chargez le fichier dans l'espace de travail de votre choix. Combien de villes y a-t-il ? Combien y en a-t-il dans chaque groupe ? Affichez des statistiques démographiques sommaires pour les villes, à la fois globalement et par groupe.

Session 3

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

La préparation des données

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES



7. La qualité et le traitement des données

Le bordel total

"Les données sont désordonnées, vous savez."

"Même après avoir été nettoyées ?"

"Surtout après avoir été nettoyées."

Le nettoyage, le traitement et la manipulation des données sont des aspects essentiels des projets de science des données.

Les analystes peuvent consacrer **jusqu'à 80 % de** leur temps à la **préparation des données**.

La manipulation et le “tidyverse”

Les données “tidy” ont une structure spécifique :

- chaque variable se retrouve dans une seule colonne
- chaque observation se retrouve dans une seule rangée
- chaque type d'unité d'observation dans un seul tableau

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

vs.

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

Fonctionnalité de traitement

Les fonctions de traitement des données doivent permettre à l'analyste de :

- **extraire** un sous-ensemble de **variables** de la trame de données
- **extraire** un sous-ensemble d'**observations** de la trame de données
- **trier** les données selon toute combinaison de variables dans un ordre croissant/décroissant
- **créer de nouvelles variables** à partir de variables existantes
- **créer des tableaux croisés dynamiques**, par groupes d'observation
- **jouer** avec les **banques de données** (jointures, etc.)
- etc.

Le nettoyage des données

Il y a deux approches **philosophiques** de nettoyage/validation des données :

- méthodique
- narrative

L'approche **méthodique** consiste à passer en revue une **liste de contrôle** des problèmes potentiels et à signaler ceux qui s'appliquent aux données.

L'approche **narrative** consiste à **explorer** l'ensemble de données et à essayer de repérer les schémas improbables et irréguliers.

Le nettoyage des données

Méthodique (syntaxe)

- Pour : la liste de contrôle est **indépendante du contexte** ; les pipelines sont **faciles à implémenter** ; les erreurs courantes/observations invalides sont **facilement identifiées**
- Contre : peut s'avérer **chronophage** ; impossible d'identifier de nouveaux types d'erreurs

Narration (sémantique)

- Pour : le processus peut simultanément permettre de **comprendre les données** ; les faux départs sont (au maximum) aussi coûteux que le passage à l'approche méthodique
- Contre : peut manquer d'importantes sources d'erreurs et d'observations invalides pour les données comportant un **nombre élevé de caractéristiques** ; la connaissance du domaine peut biaiser le processus en négligeant les zones intéressantes de l'ensemble de données

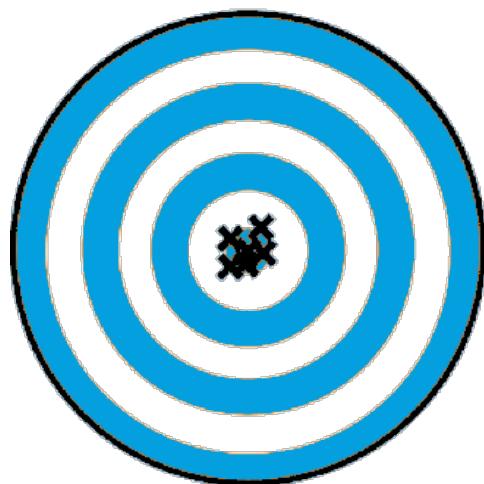
La solidité des données

L'ensemble de données idéal aura le moins de problèmes possible par rapport à ...

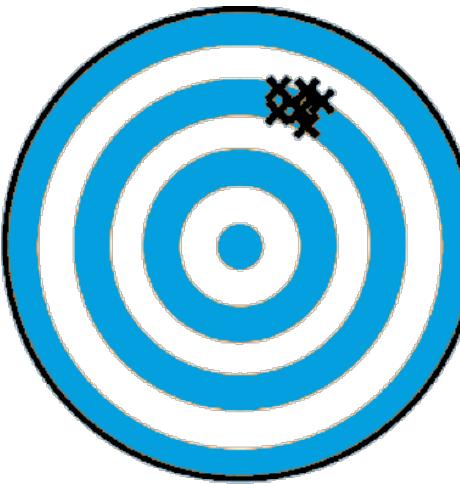
- **validité** : type de données, plage, réponse obligatoire, unicité, valeur, expressions régulières
- **exhaustivité** : observations manquantes
- **exactitude et précision** : liées aux erreurs de mesure et de saisie des données ; diagrammes de cibles (exactitude = biais, précision = erreur standard)
- **cohérence** : observations contradictoires
- **uniformité** : les unités sont-elles utilisées de manière uniforme ?

La vérification des problèmes liés à la qualité des données dès le départ peut vous éviter des maux de tête plus tard dans l'analyse.

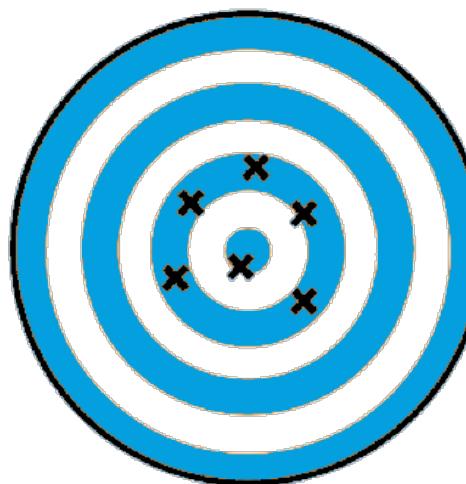
La solidité des données



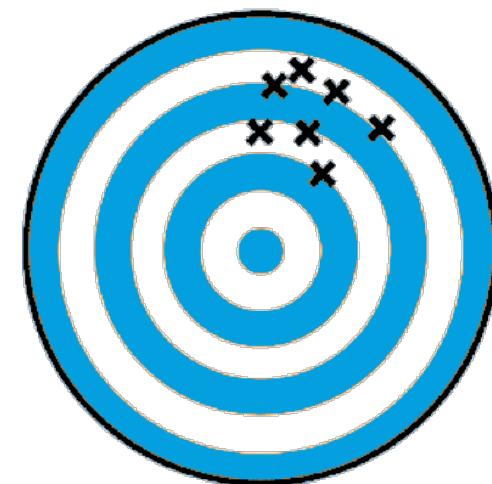
exact et
précis



précis, mais
pas exact



exact, mais
pas précis

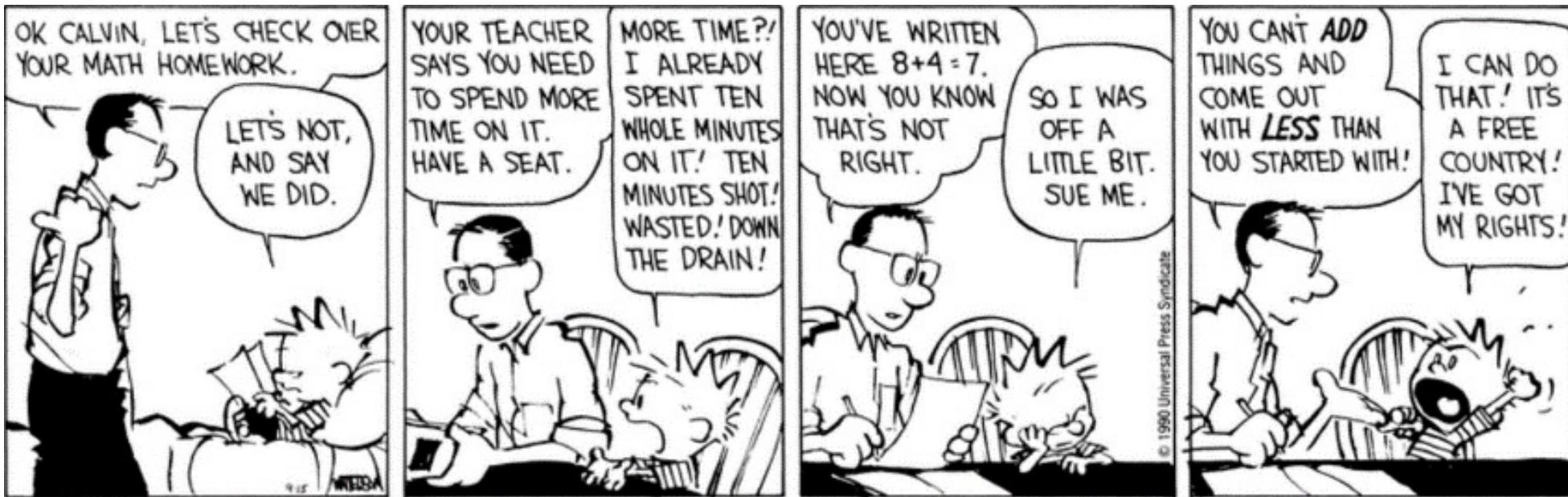


ni exact,
ni précis

Les sources d'erreurs communes

Lorsque vous traitez des ensembles de données **hérités** ou **combinés** (c'est-à-dire des ensembles de données sur lesquels vous n'avez pas contrôle de la collecte et du traitement initial) :

- données manquantes avec un code
- 'NA'/'blank' avec un code
- erreur de saisie de données
- erreur de codage
- erreur de mesure
- entrées dupliquées
- accumulation ("heaping")



© 1990 Universal Press Syndicate

La détection d'entrées non valides

Les entrées potentiellement invalides peuvent être détectées à l'aide de :

- **statistiques descriptives univariées**
compte, étendue, score-z, moyenne, médiane, écart-type, contrôle logique
- **statistiques descriptives multivariées**
tableaux croisés, contrôle logique
- **visualisation des données**
nuage de points, histogramme, etc.

La détection d'entrées non valides

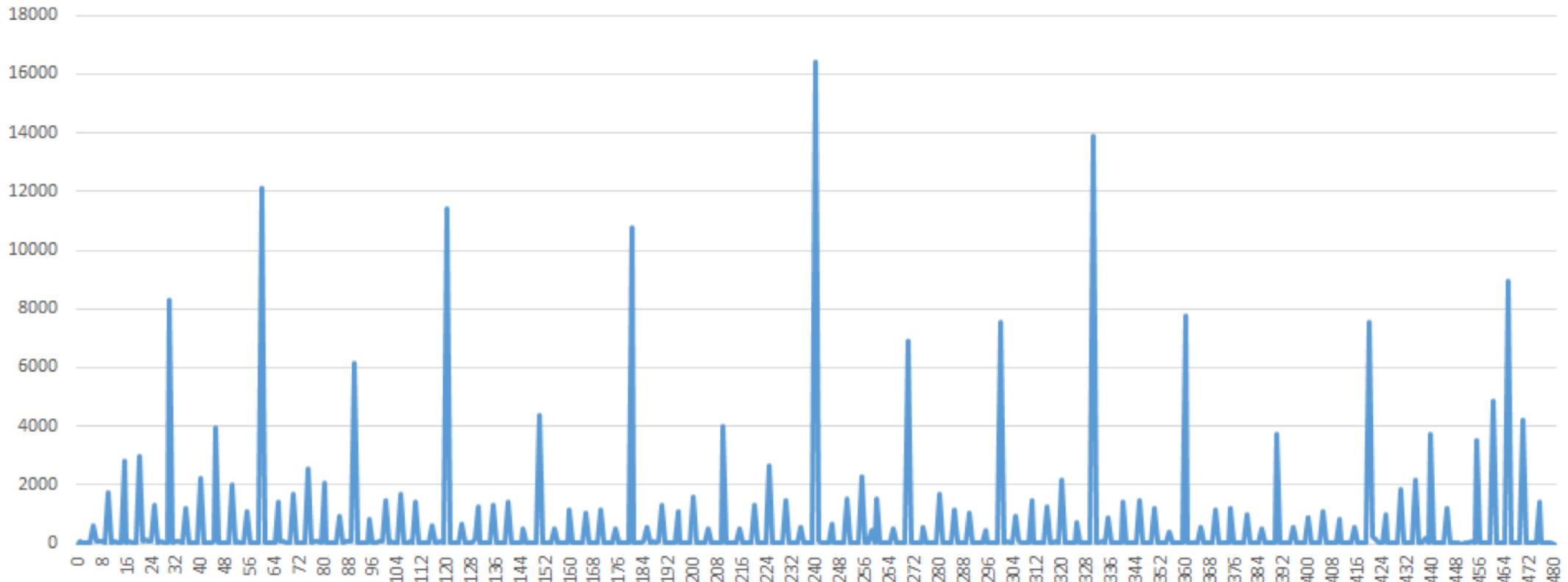
Les tests univariés ne montrent pas toujours **tout ce qui se passe**.

Cette étape pourrait permettre d'identifier les valeurs aberrantes potentielles.

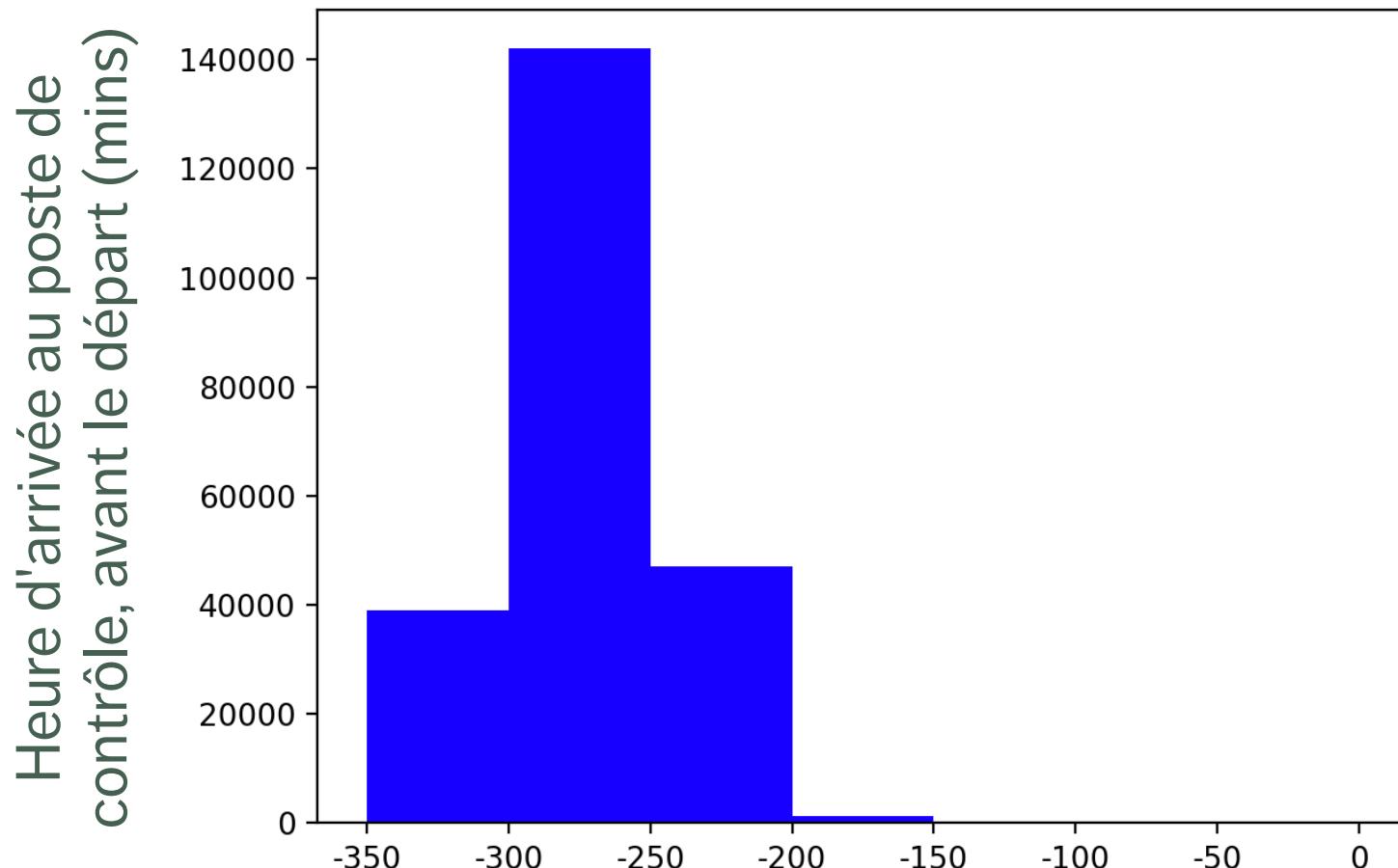
Défaut de détection des entrées non valides \neq toutes les entrées sont valides.

Un petit nombre d'entrées non valides devrait être recodées comme étant "manquantes".

La détection d'entrées non valides



La détection d'entrées non valides



La détection d'entrées non valides

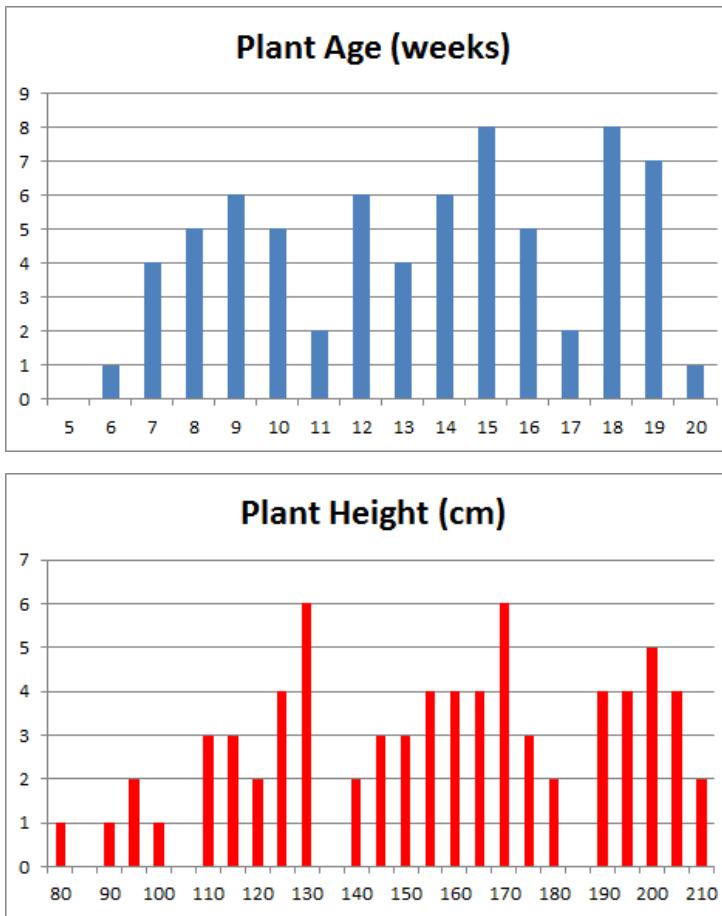
Sex	Male	19
	Female	17
	(blank)	2
	Total	38

Pregnant	Yes	7
	No	27
	99	1
	(blank)	3
	Total	38

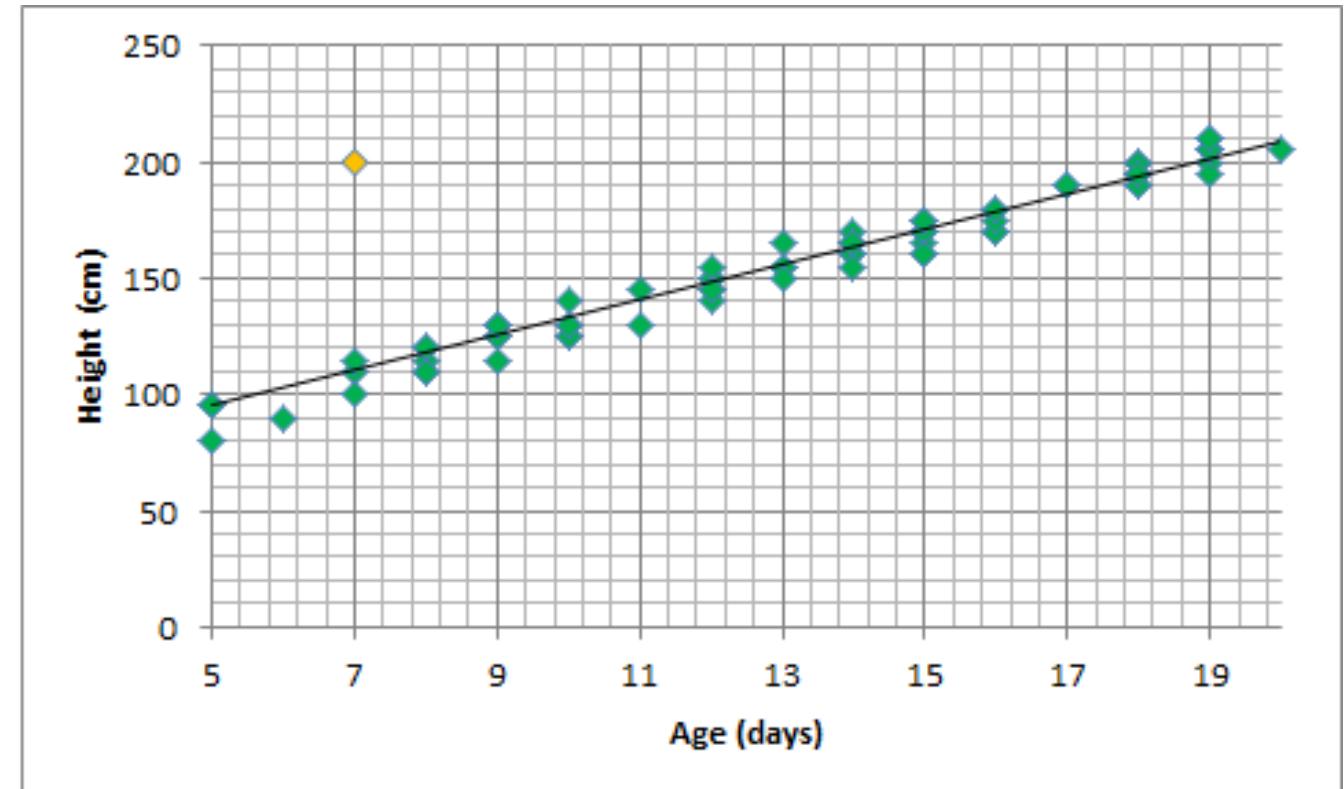
vs.

	Pregnant				Total
	Yes	No	99	(blank)	
Sex	Male	1	17	1	0
	Female	6	9	0	2
	(blank)	0	1	0	1
	Total	7	27	1	3
					38

La détection d'entrées non valides

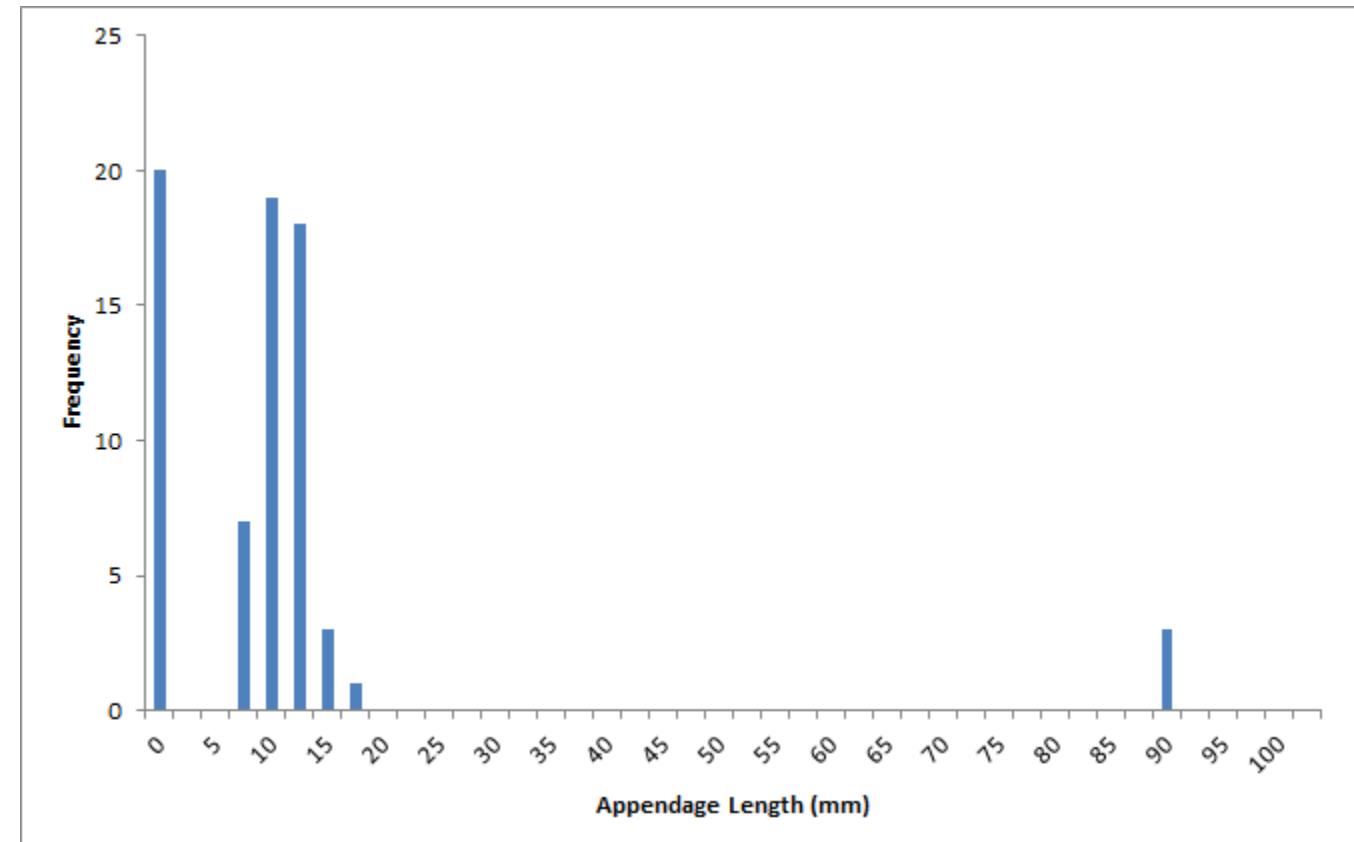


VS.



La détection d'entrées non valides

<i>Appendage length (mm)</i>	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71



Lectures suggérées

La qualité et le traitement des données

Data Understanding, Data Analysis, Data Science **Data Preparation**

Introduction

General Principles

- Approaches to Data Cleaning
- Pros and Cons
- Tools and Methods

Data Quality

- Common Error Sources
- Detecting Invalid Entries

Exercices

La qualité et le traitement des données

1. Recréez les exemples du [Tidyverse](#).
2. Transformez le fichier [cities.txt](#) en ensemble de données “tidy”.
3. L'ensemble de données trouvé dans le fichier [cities.txt](#) semble-t-il être de bonne qualité (est-il “sain” ? comporte-t-il des entrées invalides ?)
4. Créez une liste d'éléments qui pourraient être utilisés dans une liste de contrôle de nettoyage méthodique des données. Utilisez des données que vous avez rencontrées dans le passé comme source d'inspiration (données numériques, catégorielles, textuelles).

Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				NA
Bruce	37	14	63		1	veggie		n/a
Steve	83		77	7	1	chicken		None
Clint	27	9	118	9		shrimp	3	empty
Wanda	19	7	52	2	2	shrimp		-
Natasha	26	4	162	5	3			

8. Les valeurs manquantes

Les types d'observations manquantes

Les champs vierges existent en 4 versions :

- **non-réponse**
une observation était attendue mais aucune n'a été saisie
- **problème de saisie des données**
une observation a été enregistrée mais n'a pas été saisie dans l'ensemble de données
- **entrée invalide**
une observation a été enregistrée mais a été considérée comme non valide et a été supprimée
- **blanc attendu**
un champ a été laissé vide, mais c'est normal

Les types d'observations manquantes

Trop de valeurs manquantes des trois premiers types peut indiquer des **problèmes dans le processus de collecte des données**.

Trop de valeurs manquantes du quatrième type peut indiquer une **mauvaise conception du questionnaire**.

Trouver les valeurs manquantes peut vous aider à traiter d'autres problèmes de science des données.

L'imputation

Les méthodes d'analyse ne s'accommodeent pas facilement des observations manquantes :

- **écartier** l'observation manquante
 - non recommandé, à moins que les données manquantes soient MCAH
 - acceptable dans certaines situations (e.g., un petit nombre de valeurs manquantes dans un ensemble de données massives)

- trouver une **valeur de remplacement (imputation)**
 - principal inconvénient : nous ne savons jamais quelle aurait été la vraie valeur
 - mais cela demeure souvent la meilleure option disponible

Les mécanismes de valeurs manquantes

Manquant complètement au hasard (MCAH)

- l'absence de l'élément est indépendante de sa valeur ou des variables auxiliaires
- **exemple :** une surtension électrique supprime aléatoirement une observation dans l'ensemble de données

Manquant au hasard (MAH)

- l'absence d'un article n'est pas complètement aléatoire ; elle peut être expliquée par des variables auxiliaires avec des informations complètes.
- **exemple :** si les femmes sont moins susceptibles de vous dire leur âge que les hommes pour des raisons sociétales, mais pas à cause des valeurs d'âge elles-mêmes

Les mécanisme de valeurs manquantes

Ne manquant pas au hasard (NMAH)

- la raison de la non-réponse est liée à la valeur de l'item (également appelée **non-réponse non-ignorable**)
- **exemple** : si les consommateurs de drogues illicites sont moins susceptibles d'admettre leur consommation de drogues que les abstinents...

En général, le mécanisme manquant **ne peut pas être déterminé** avec certitude ; on devra émettre des hypothèses (l'expertise du domaine aide).

Les méthodes d'imputation

- suppression par liste
- imputation par la moyenne ou par la valeur la plus fréquente
- imputation par la régression ou la corrélation
- imputation par la régression stochastique
- report de la dernière observation
- report en arrière de l'observation suivante
- imputation par les k voisins les plus proches
- imputation multiple
- etc.

Les méthodes d'imputation

Suppression par liste : supprimer les unités avec au moins 1+ valeurs manquantes

- **hypothèse** : MCAH
- **Contre** : peut introduire un biais (si non MCAH), réduction de la taille de l'échantillon, augmentation de l'erreur standard

Imputation moyenne/la plus fréquente : remplacer les valeurs manquantes par la valeur moyenne/la plus fréquente.

- **hypothèse** : MCAH
- **contre** : distorsions de la distribution (pic à la moyenne) et des relations entre les variables

Les méthodes d'imputation

Imputation par régression/corrélation : remplacer les valeurs manquantes par des valeurs ajustées en se basant sur des variables avec des informations complètes.

- **hypothèse** : MAH
- **contre** : réduction artificielle de la variabilité, surestimation de la corrélation

Imputation par régression stochastique : imputation par la régression/la corrélation avec ajout d'un terme d'erreur aléatoire

- **hypothèse** : MAH
- **contre** : risque accru d'erreur de type I (faux positifs) en raison de la faible erreur-type

Les méthodes d'imputation

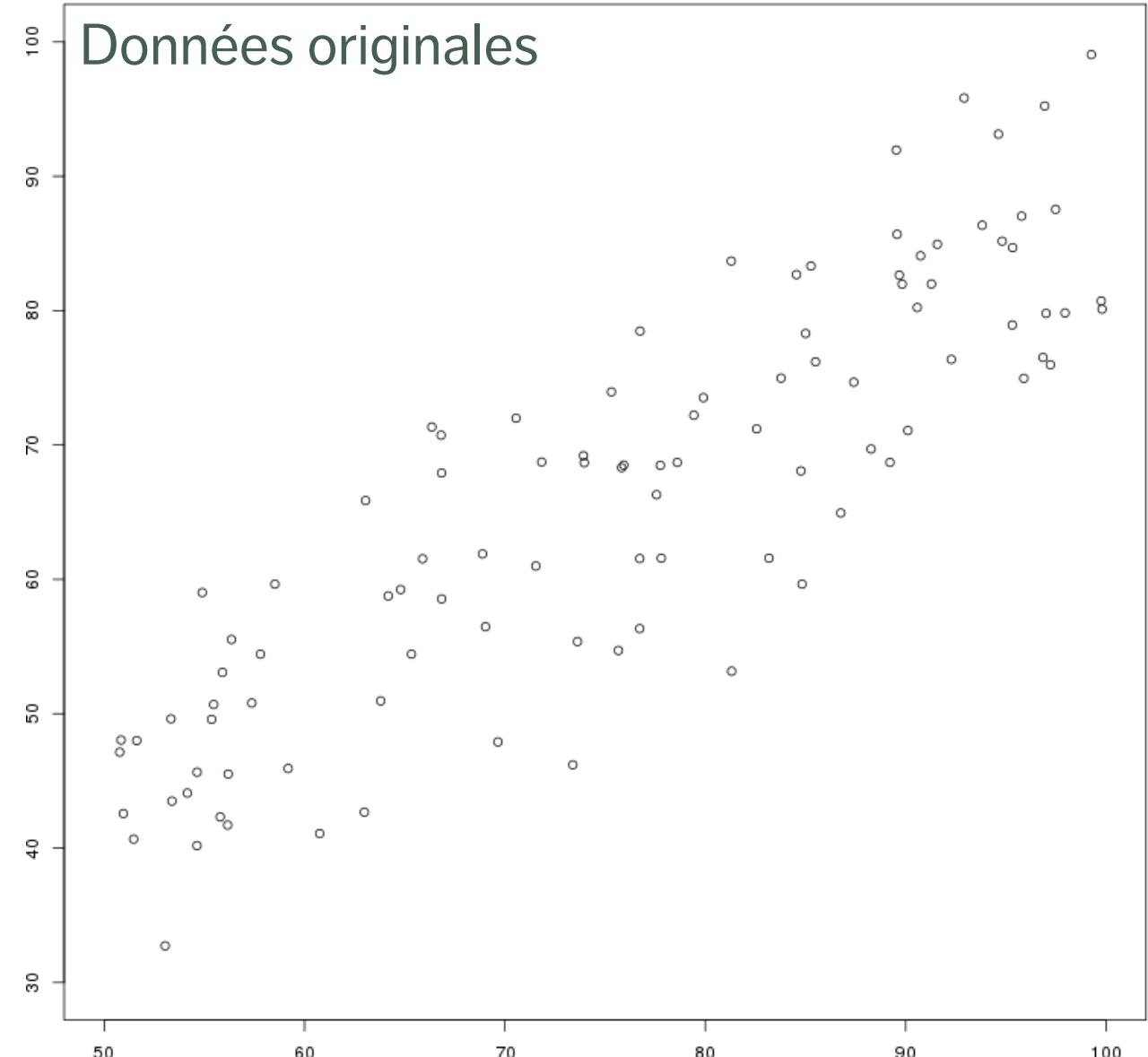
Dernière observation reportée : remplacer les valeurs manquantes par les dernières valeurs précédentes (dans une étude longitudinale)

- **hypothèse** : MCAH, les valeurs ne varient pas beaucoup au fil du temps
- **contre** : peut être trop "généreux", selon la nature de l'étude

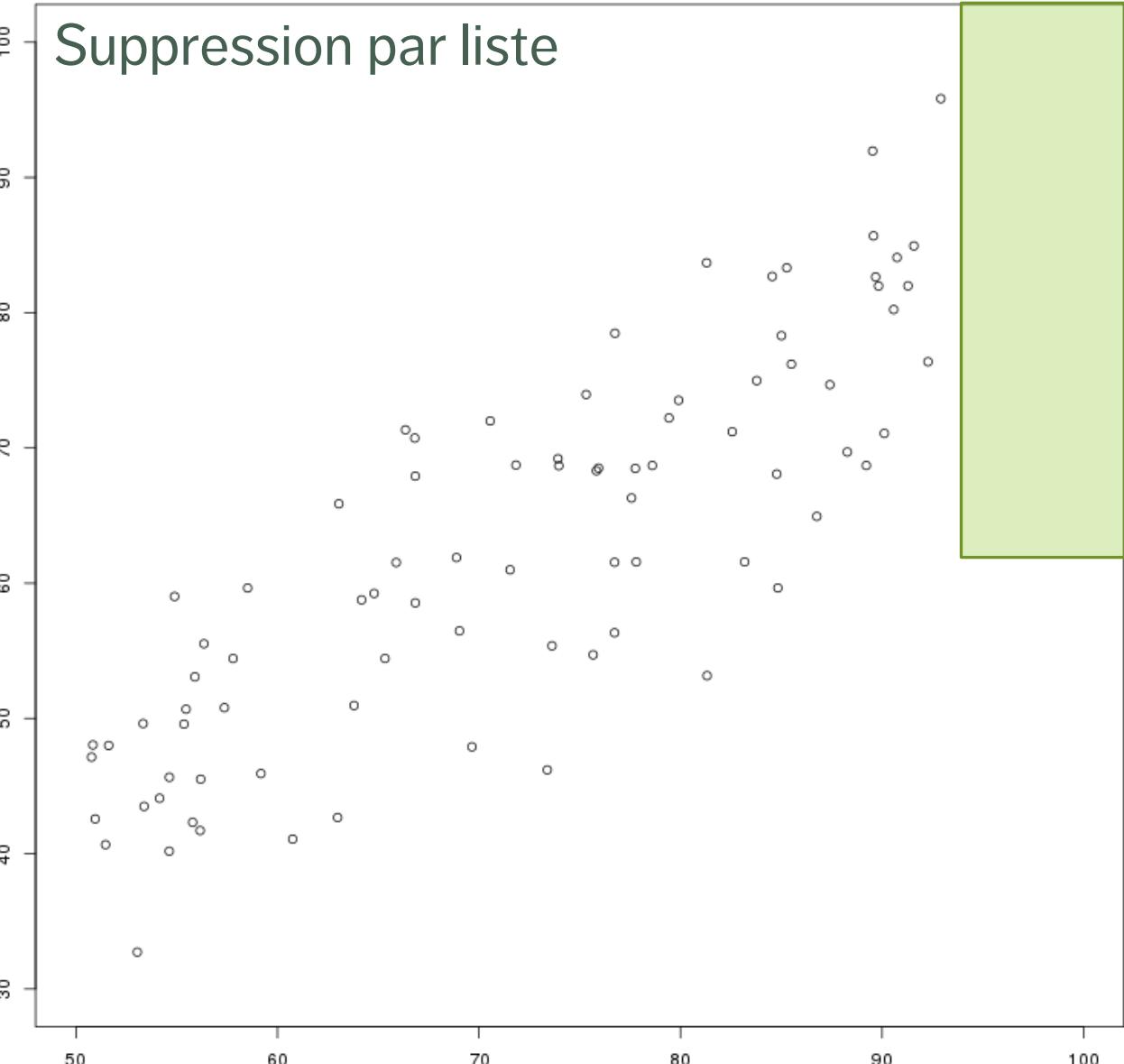
imputation par le plus proche voisin (k NN) : remplacer l'entrée manquante par la moyenne du groupe des k cas complets les plus similaires

- **hypothèse** : MAH
- **contre** : difficile de choisir une valeur appropriée de k ; distorsion possible dans la structure des données

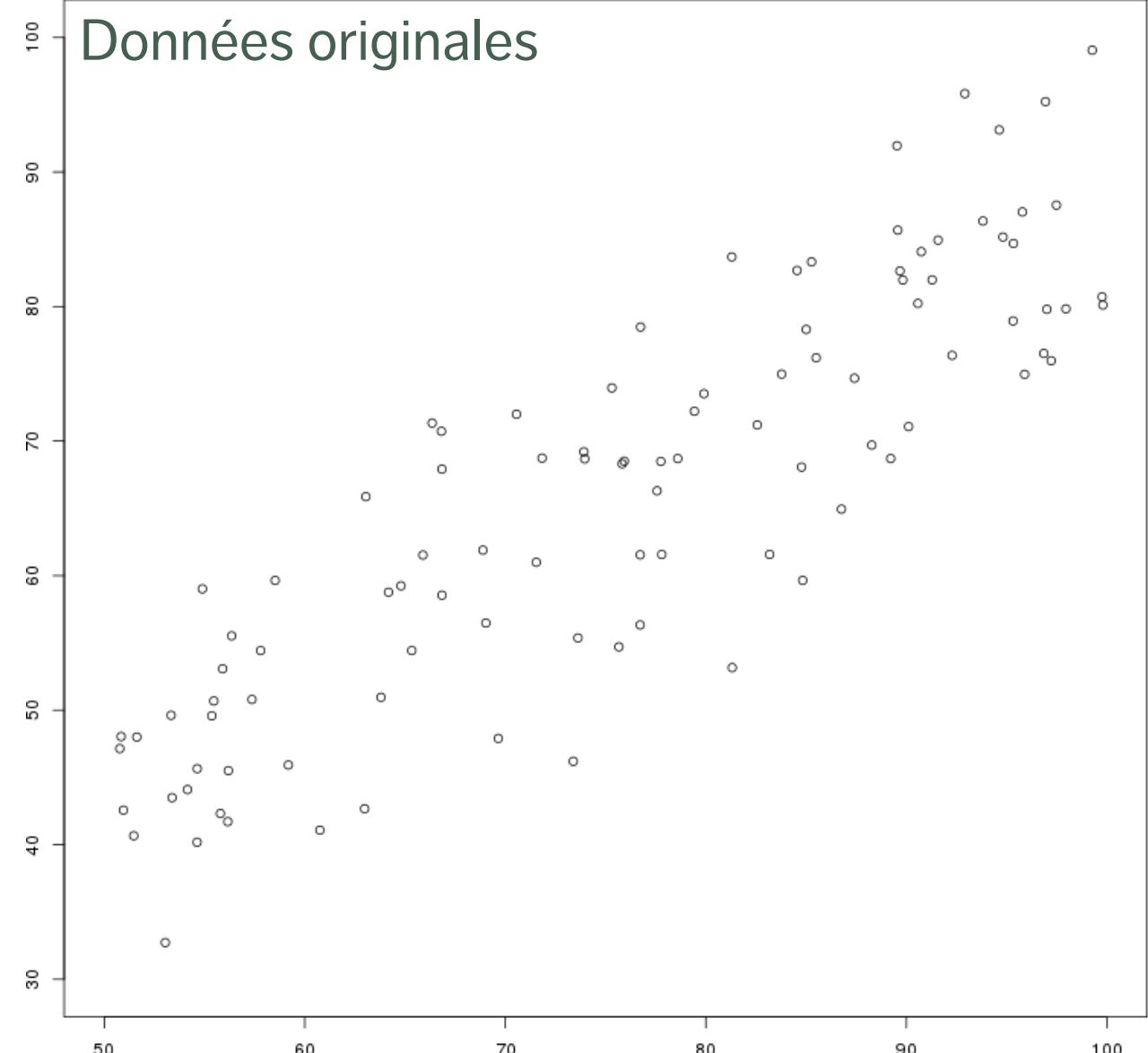
Données originales



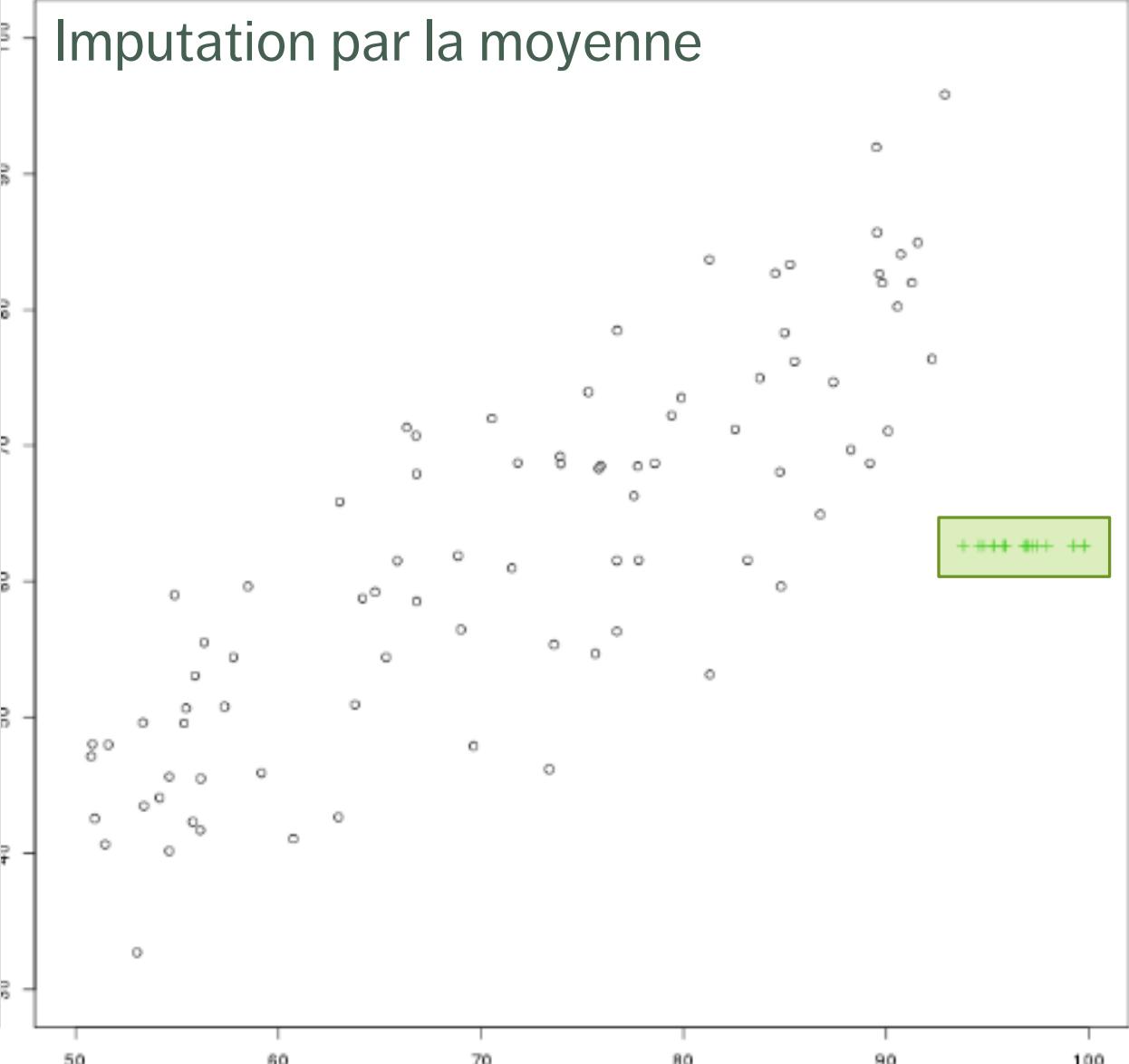
Suppression par liste



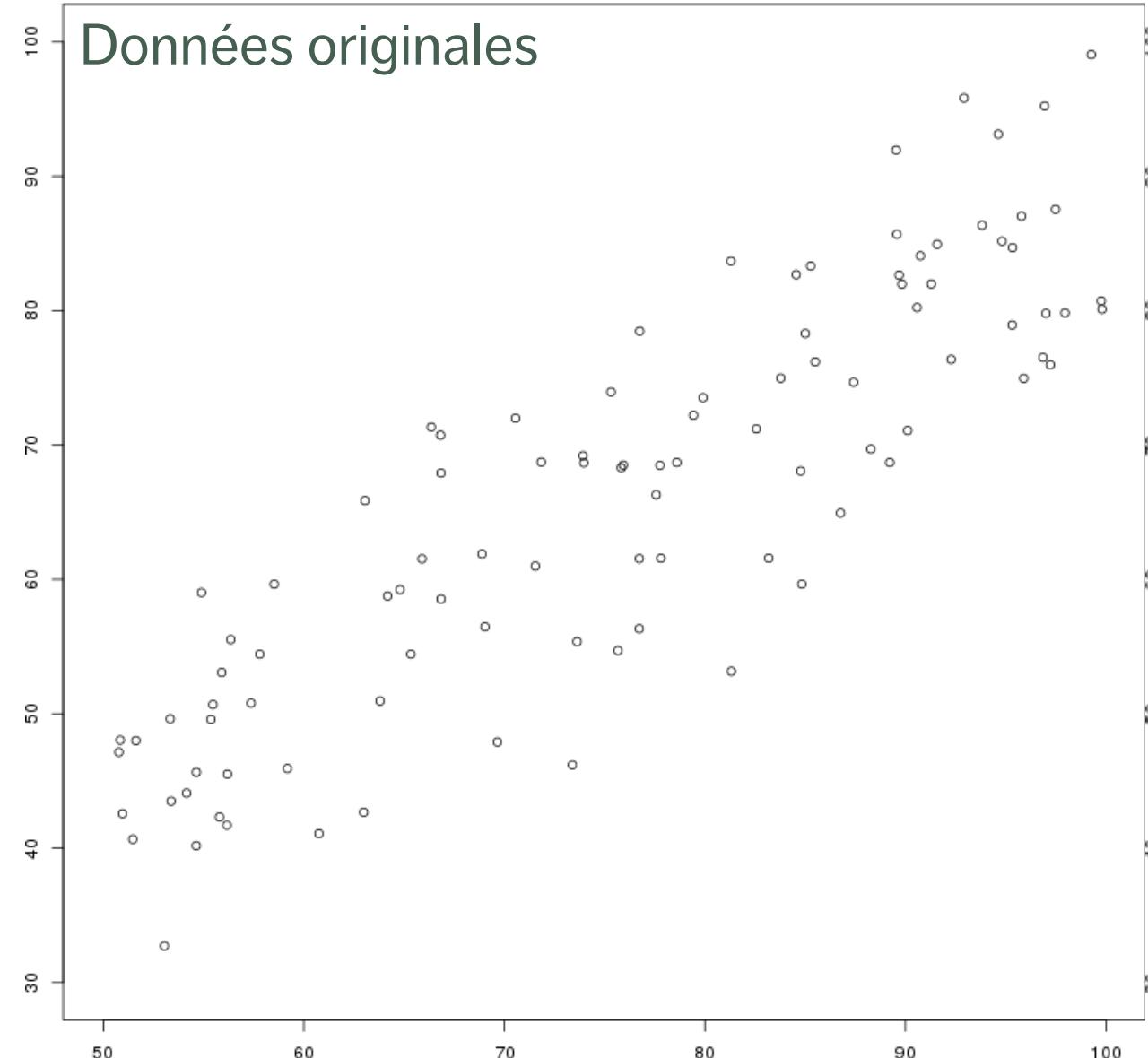
Données originales



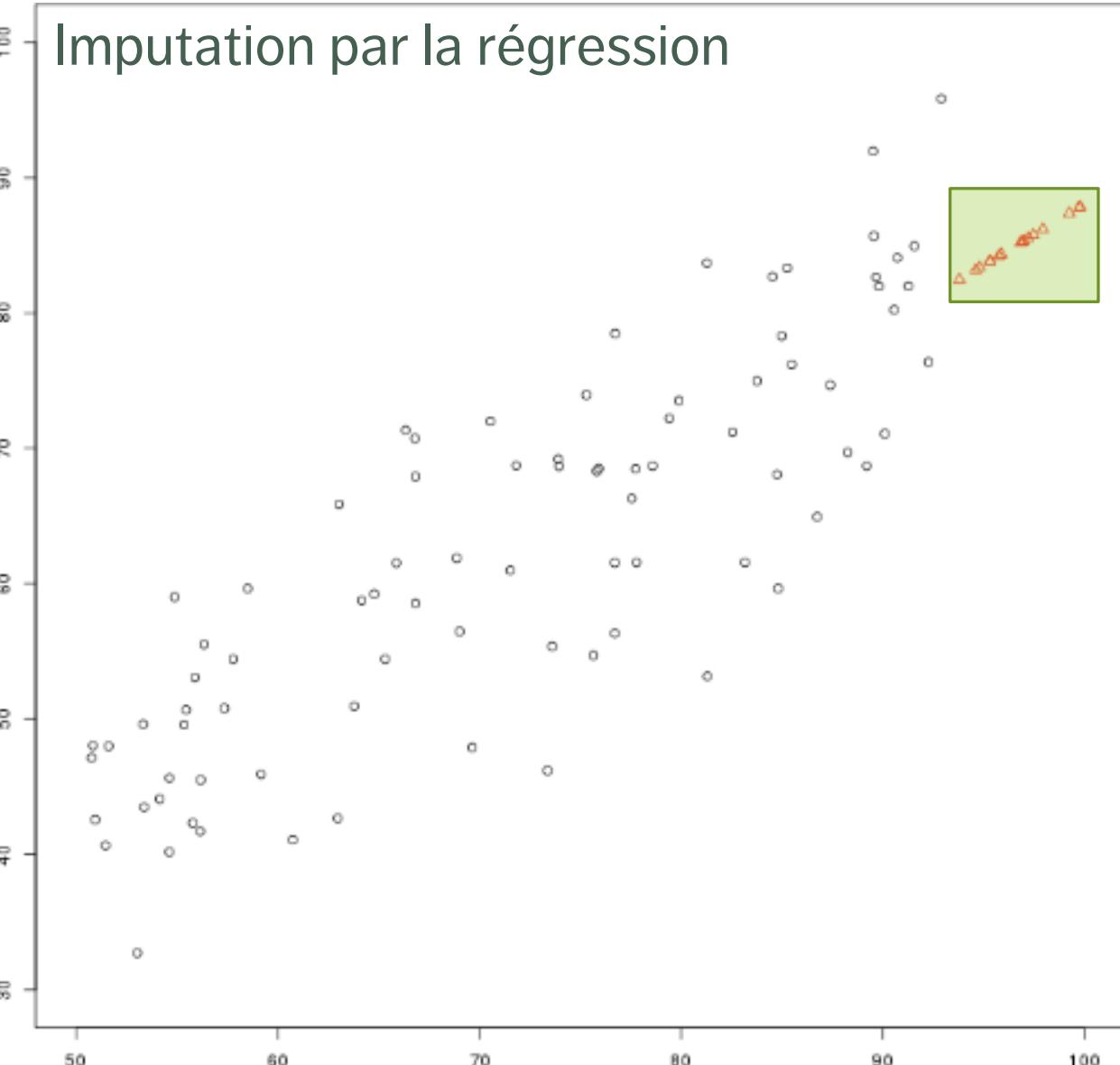
Imputation par la moyenne



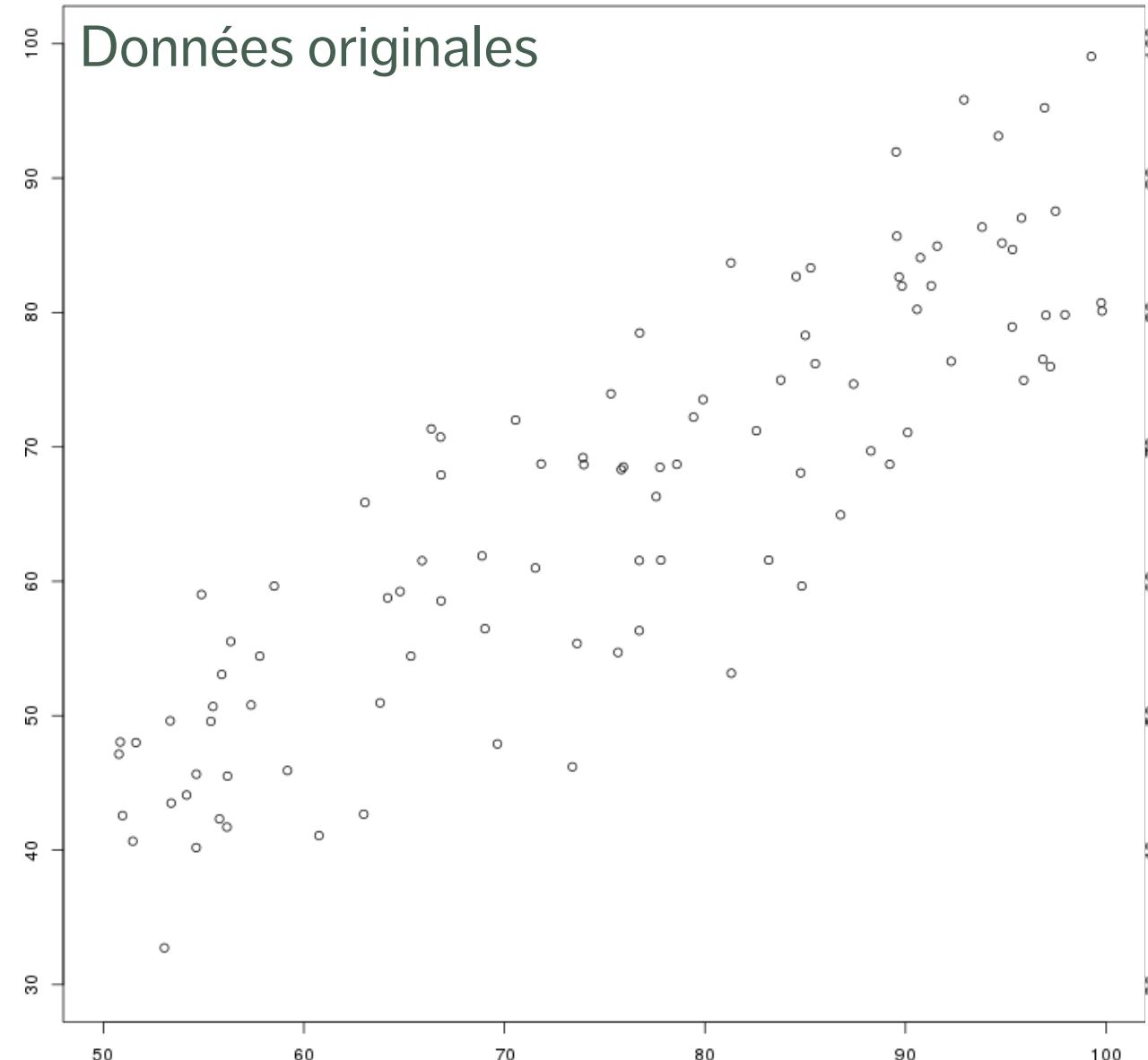
Données originales



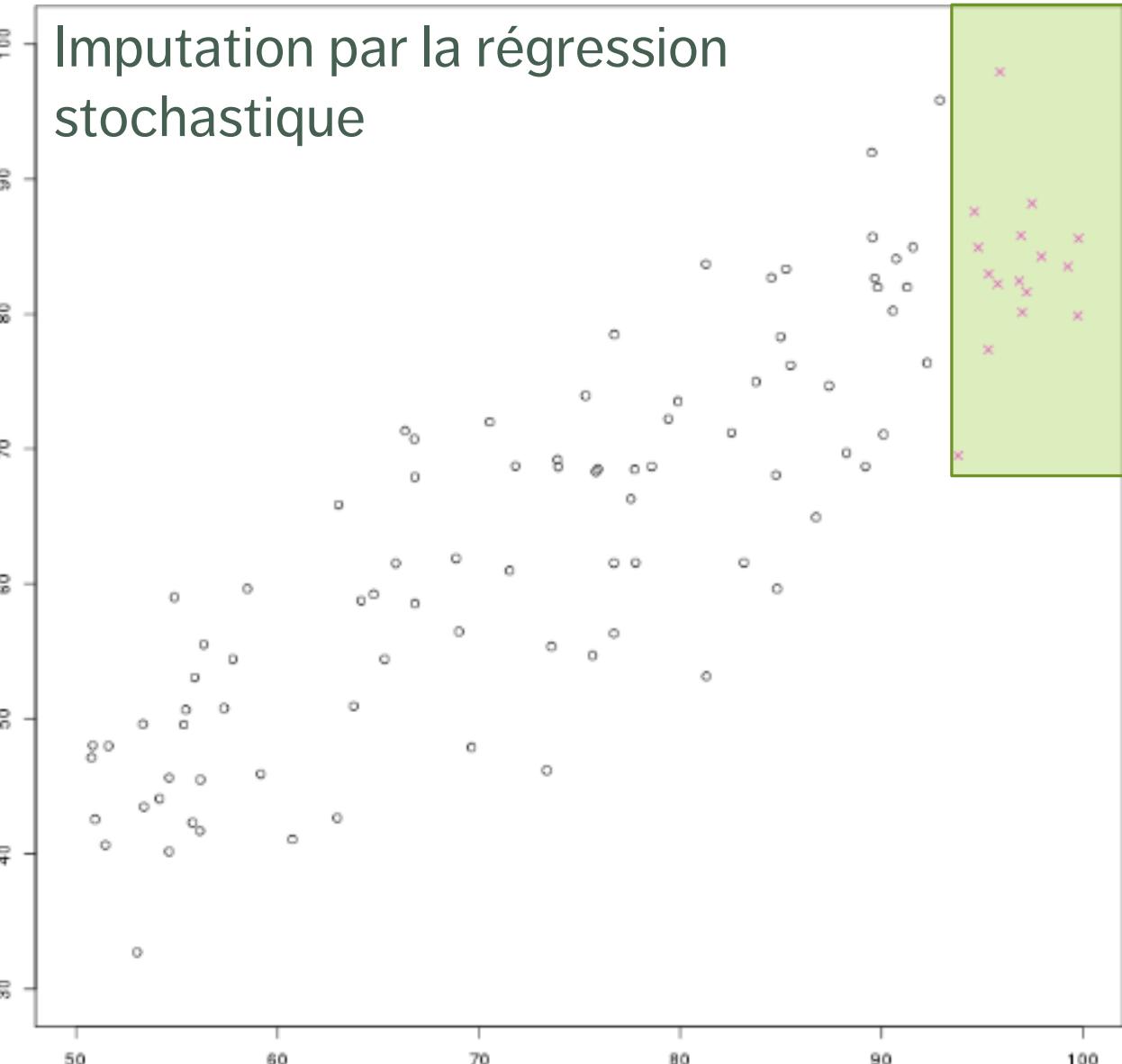
Imputation par la régression



Données originales



Imputation par la régression stochastique



L'imputation multiple

Les imputations augmentent le “bruit” (l’incertitude) dans les données.

Dans le cas de l'**imputation multiple**, l’effet de ce bruit peut être mesuré en consolidant les résultats de l’analyse à partir de plusieurs répétitions de la procédure d’imputation sur l’ensembles de données manquantes

Étapes :

1. l’imputation répétée crée m versions de l’ensemble de données
2. chacun de ces m ensembles de données est analysé, ce qui donne m résultats
3. les m résultats sont regroupés en un seul résultat pour lequel la moyenne, la variance et les intervalles de confiance sont connus

L'imputation multiple

Avantages

- **flexible** ; peut être utilisé dans diverses situations (MCAH, MAH, voire NMAH dans certains cas)
- tient compte de l'**incertitude** des valeurs imputées
- assez facile à mettre en œuvre

Inconvénients

- m peut devoir être assez **grande** lorsqu'il y a plusieurs valeurs manquantes dans de nombreuses caractéristiques, ce qui ralentit les analyses
- si le résultat de l'analyse n'est pas une valeur unique mais un objet mathématique compliqué, cette approche a peu de chances d'être utile

À retenir

Les valeurs manquantes **ne peuvent pas être simplement ignorées.**

Le mécanisme manquant **ne peut généralement pas être déterminé** avec certitude.

Les méthodes d'imputation fonctionnent mieux lorsque les valeurs sont **MCAH** ou **MAH**; les méthodes d'imputation ont tendance à produire des estimations biaisées.

Dans l'imputation simple, les données imputées sont traitées comme les données réelles ; l'**imputation multiple** peut contribuer à réduire le bruit.

L'imputation stochastique est-elle la meilleure solution ? Dans notre exemple, oui - mais ... faites attention au **théorème du “No-Free Lunch”** !

Lectures suggérées

Les valeurs manquantes

*Data Understanding, Data Analysis, Data Science
Data Preparation*

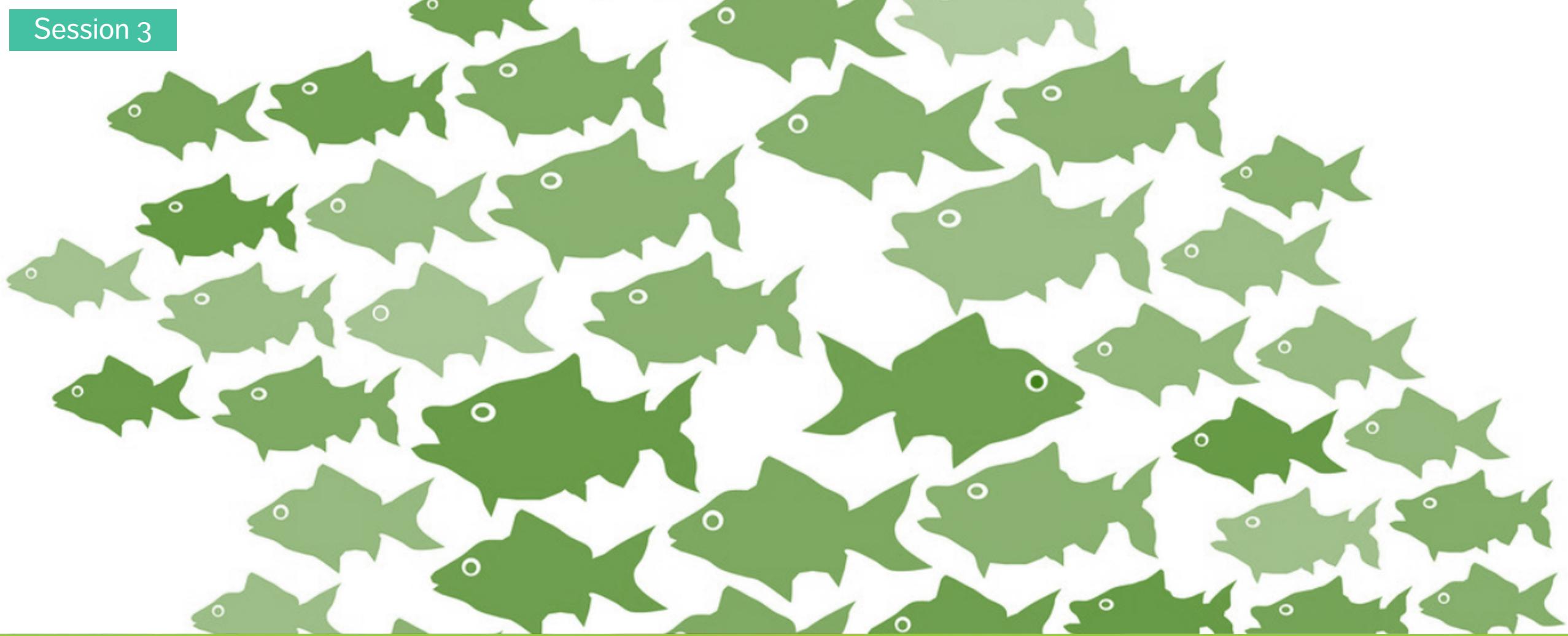
Missing Values

- Missing Value Mechanisms
- Imputation Methods
- Multiple Imputation

Exercices

Les valeurs manquantes

1. Recréez les exemples de Imputation Methods.
2. Recréez le processus d'imputation des valeurs manquantes (nettoyage des données) utilisé dans Example: Algae Bloom.
3. Effectuez l'imputation k NN sur l'ensemble de données des grades avec différentes valeurs de k .
4. Effectuez une imputation multiple sur l'ensemble de données grades en utilisant la régression stochastique afin d'estimer la pente et l'ordonnée de la ligne de meilleur ajustement.



9. Les observations anormales

Les observations anormales

En pratique, une **observation anormale** peut se présenter comme

- un "**mauvais**" **objet/mesure** : artefacts de données, fautes, valeurs mal imputées, etc. ;
- une **observation mal classée** : selon les modèles de données existants, l'observation aurait dû être étiquetée différemment ;
- une observation dont les mesures se trouvent dans les **queues de distribution** d'un nombre suffisamment grand d'éléments ;
- un **inconnu inconnu** : un type d'observation totalement nouveau dont l'existence était jusqu'alors insoupçonnée.

Les observations anormales

Une observation peut être anormale dans un contexte, mais pas dans un autre

- un homme adulte de 1.80 m se situe dans le 86^e percentile pour les hommes canadiens (grand, mais pas inhabituel).
- en Bolivie, le même homme serait dans le 99.9^e percentile (très grand, inhabituel)

La détection des anomalies soulève des **questions intéressantes** pour les analystes et les experts en la matière : dans ce cas, pourquoi existe-t-il un écart aussi important entre les deux populations ?

Les valeurs aberrantes (“outliers”)

Les **observations aberrantes** sont des observations qui sont **atypiques** par rapport aux :

- autres caractéristiques à même l'unité (“*within units*”), et
- valeurs des caractéristiques des autres unités (“*between-units*”)

Les valeurs aberrantes sont des observations qui **ne ressemblent pas aux autres cas** ou qui **contredisent des dépendances** ou des règles **connues**.

Une étude minutieuse est nécessaire pour déterminer si ces valeurs aberrantes doivent être conservées ou supprimées de l'ensemble de données.

La détection des anomalies

Les valeurs aberrantes peuvent être anormales par rapport à une variables de l'unité, ou en combinaison.

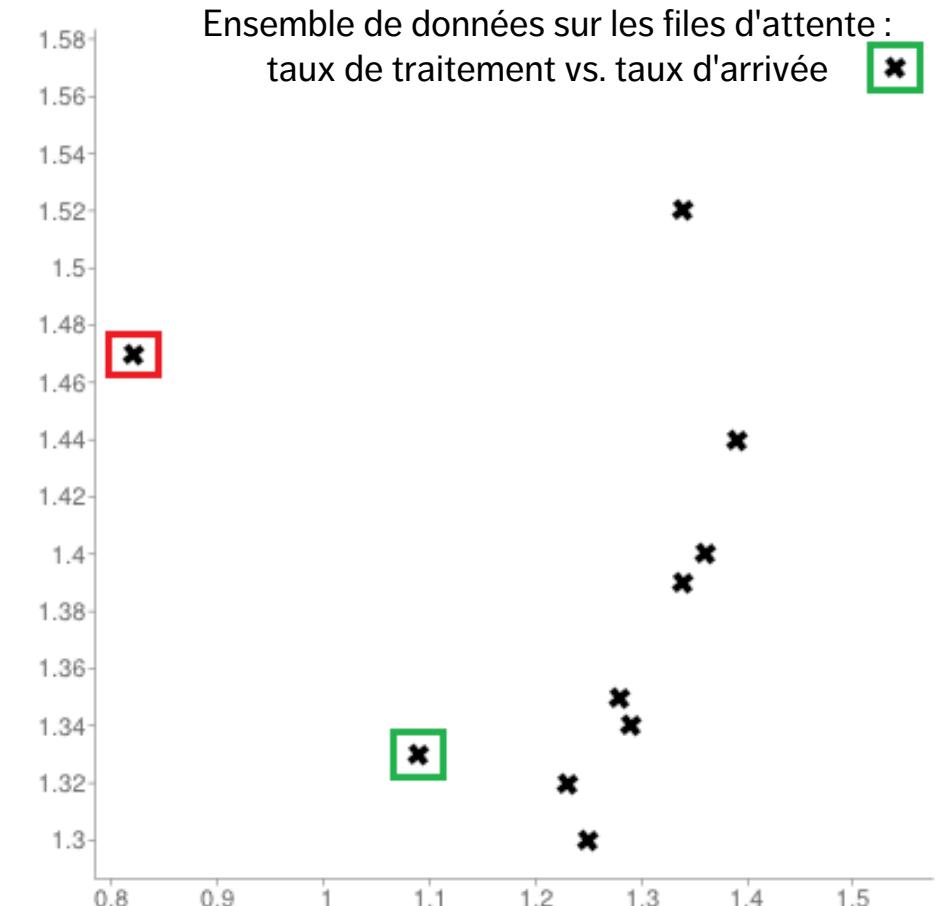
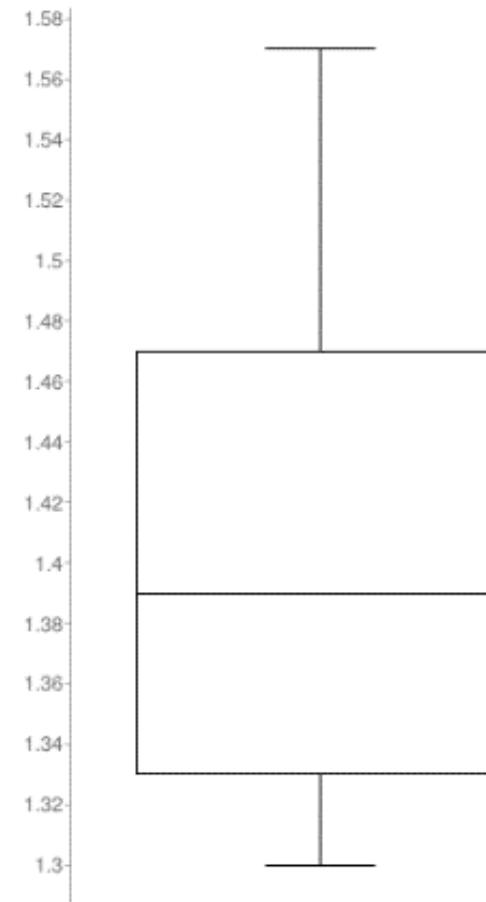
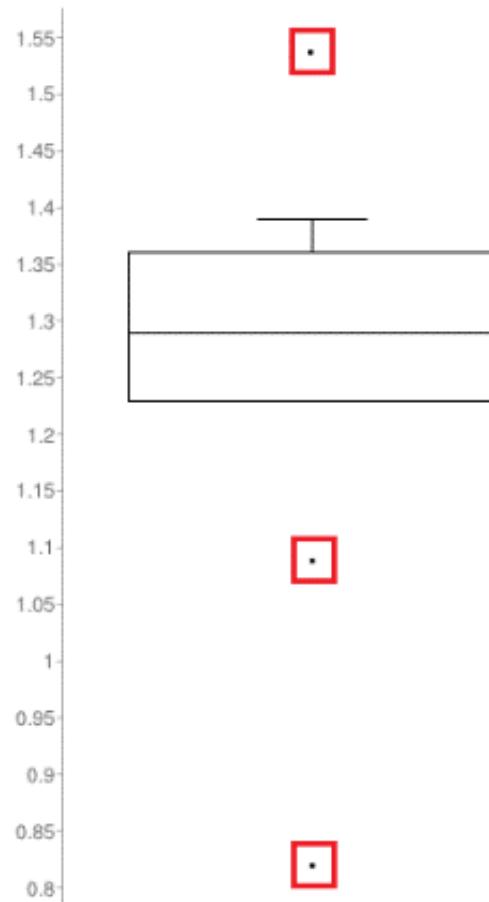
Les anomalies sont par définition **peu fréquentes**, et généralement entourées d'**incertitude en** raison de la petite taille des échantillons.

Il est difficile de différencier les anomalies du bruit ou des erreurs de saisie.

Les limites entre les unités normales et déviante peuvent être **foues**.

Les anomalies liées à des activités malveillantes sont généralement **déguisées**.

La détection des anomalies



La détection des anomalies

Il y a de nombreuses méthodes pour identifier les observations anormales ; **aucune d'entre elles n'est infaillible** et il faut faire preuve de discernement

Les méthodes graphiques sont faciles à mettre en œuvre et à interpréter.

- **observations périphériques**

box-plots, nuages de points, matrices de nuages de points, tour 2D, distance de Cooke, tracés qq normaux

- **données influentes**

un certain niveau d'analyse doit être effectué (effet de levier)

Attention : si les observations anormales ont été retirées de l'ensemble de données, des unités auparavant "régulières" peuvent devenir anormales !

Algorithmes de détection d'anomalie

Les **méthodes supervisées** utilisent un historique d'observations anormales étiquetées :

- l'expertise du domaine est requise pour étiqueter
- tâche de classification ou de régression
- problème d'occurrence rare

		Prédictions	
		Normales	Anormales
Réalité	Normales	<i>VN</i>	<i>FP</i>
	Anormales	<i>FN</i>	<i>VP</i>

Les **méthodes non supervisées** n'utilisent pas d'informations externes :

- méthodes et tests traditionnels
- problème de regroupement ou de règles d'association

Les algorithmes de détection

Le coût des erreurs de classification est souvent supposé être symétrique, ce qui peut conduire à des résultats **techniquement corrects, mais inutiles**.

Par exemple, la grande majorité des passagers aériens (99.99%+) n'emportent pas d'armes ; un modèle qui prédit qu'aucun passager ne fait passer une arme en fraude serait précis à 99.99%+, mais il passerait à côté du problème.

Pour l'**agence de sécurité**, le coût de penser à tort qu'un passager :

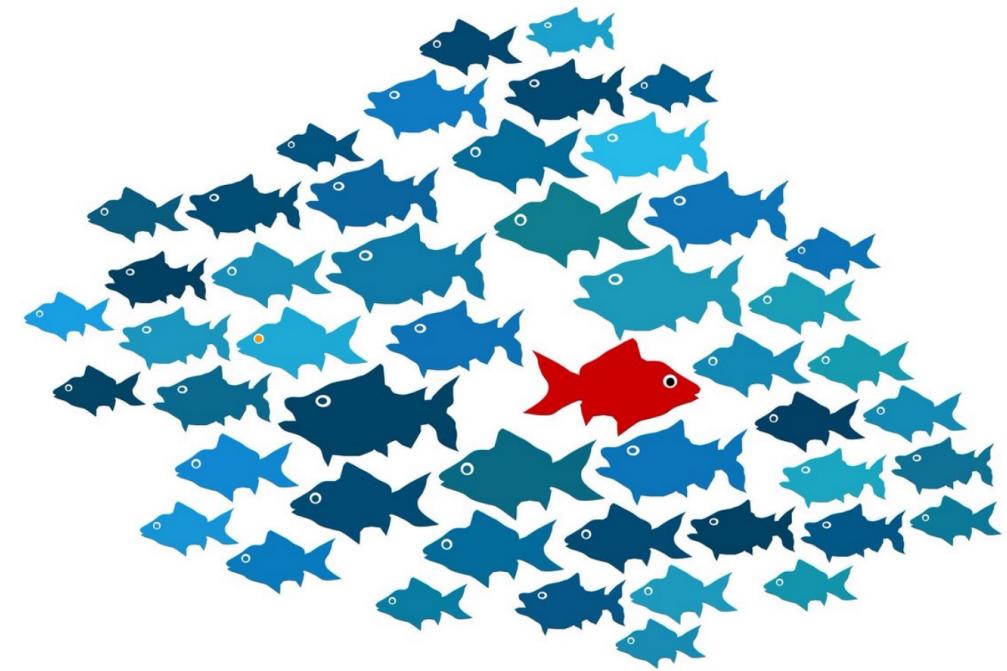
- introduit clandestinement une arme \Rightarrow coût d'une seule fouille
- ne fait pas passer une arme en fraude \Rightarrow catastrophe (potentiellement)

Les personnes injustement visées auront un point de vue différent à ce sujet !

Les algorithmes de détection

Si tous les participants à un atelier à l'exception d'un seul, peuvent visionner les conférences par vidéo, cette personne, cette connexion Internet et cet ordinateur sont **anormaux**, car ils ne se comportent pas comme les autres.

Mais cela **NE SIGNIFIE PAS** nécessairement que le comportement différent est celui qui nous intéresse...



Tests simples de valeurs aberrantes

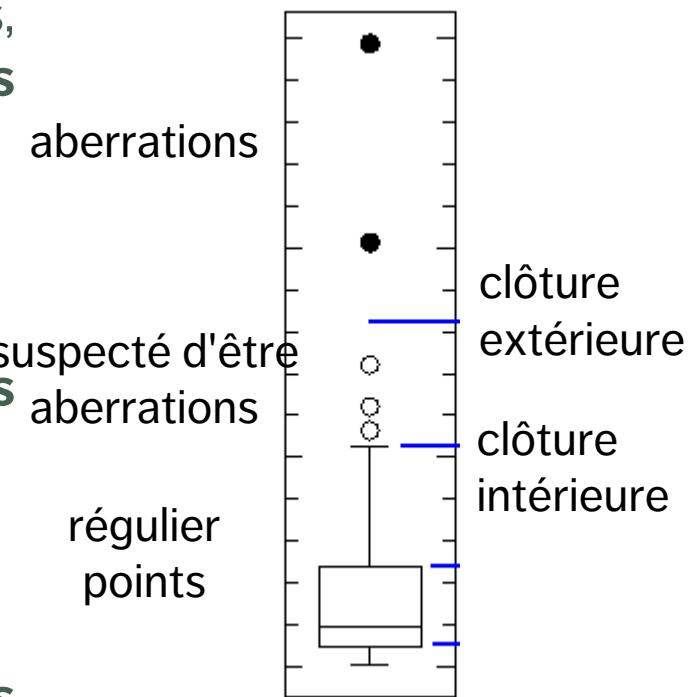
Test Boxplot de Tukey : pour les données normalement distribuées, les observations régulières se situent généralement entre les **clôtures intérieures** :

$$Q_1 - 1.5 \times (Q_3 - Q_1) \text{ et } Q_3 + 1.5 \times (Q_3 - Q_1).$$

Les **valeurs aberrantes suspectes** se situent entre les **clôtures intérieures** et les **clôtures extérieures** :

$$Q_1 - 3 \times (Q_3 - Q_1) \text{ et } Q_3 + 3 \times (Q_3 - Q_1).$$

Les **valeurs aberrantes** se trouvent au-delà des **clôtures extérieures**.



Tests simples de valeurs aberrantes

Le **test Q de Dixon** est utilisé dans les sciences expérimentales pour trouver des valeurs aberrantes dans des ensembles de données (extrêmement) petits (validité douteuse).

La **distance de Mahalanobis** (liée à l'effet de levier) peut être utilisée pour trouver des valeurs aberrantes multidimensionnelles (lorsque les relations sont linéaires).

Autres tests simples :

- **Grubbs** (univarié)
- **Tietjen-Moore** (pour un nombre spécifique de valeurs aberrantes)
- **écart généralisé extrême studentisé** (pour un nombre inconnu de valeurs aberrantes)
- **chi-deux** (les valeurs aberrantes affectant la qualité de l'ajustement)

Test sophistiqués des valeurs aberrantes

- **DBSCAN, OR_h, et LOF** (détection non supervisée des valeurs aberrantes)
- méthode **rang-puissance** (détection supervisée des valeurs aberrantes)
- méthodes **basées sur la distance** ou la **densité** (avec des mesures de distance exotiques)
- **autoencodeurs et erreur de reconstruction** (méthode d'apprentissage profond)
- méthodes **d'occurrences rares** (suréchantillonnage, sous-échantillonnage, CREDOS, PN, SHRINK, SMOTE, DRAMOTE, SMOTEBost, RareBoost, MetaCost, AdaCost, CSB, SSTBoost, etc.)
- **AVF**, algorithmes **Greedy** (données catégoriques)
- **PCA, DOBIN** et autres méthodes de **projection** (pour les données à haute dimension)
- méthodes **subspatiales** et méthodes d'**ensemble**

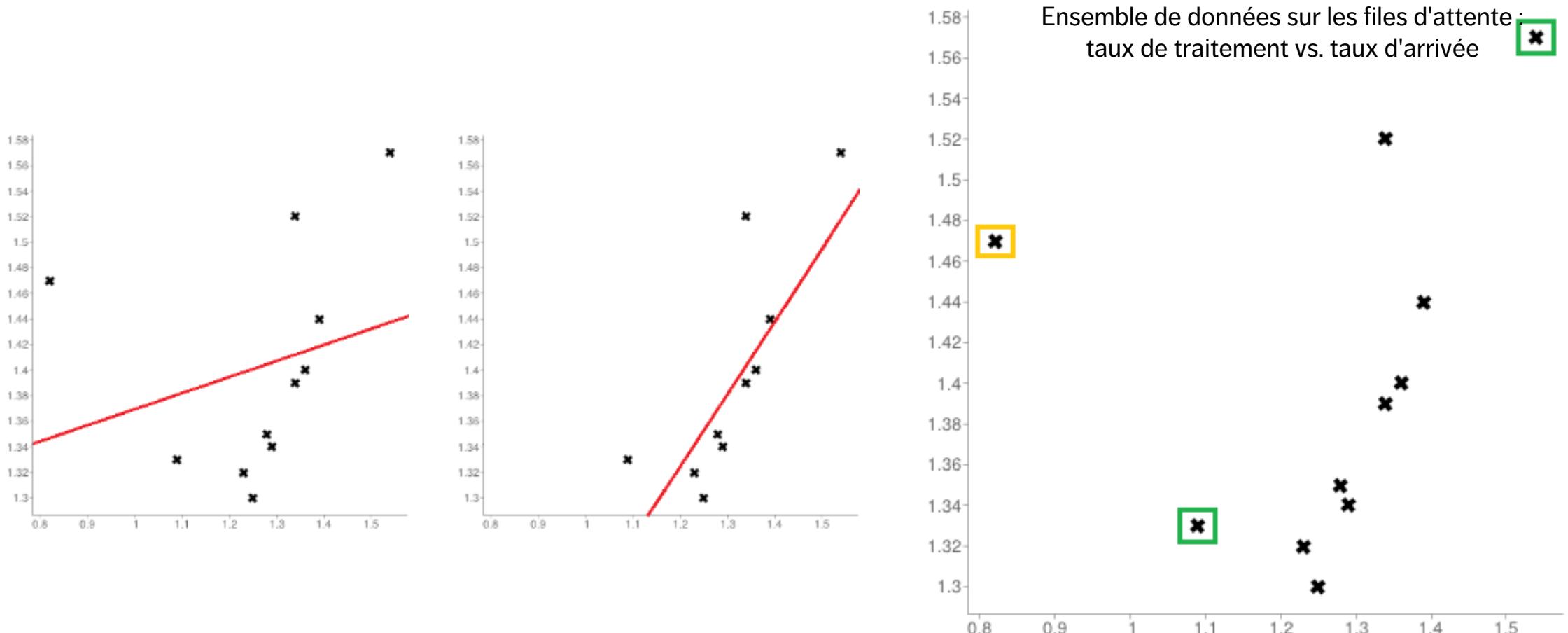
Observations influentes

Les **observations influentes** sont des observations dont l'absence entraîne des résultats d'analyse **nettement différents**.

Lorsque des observations influentes sont identifiées, des **mesures correctives** (telles que des transformations de données) peuvent être nécessaires pour minimiser leurs effets indus.

Les valeurs aberrantes peuvent être des observations influentes ; les observations influentes ne sont pas nécessairement des valeurs aberrantes (et *vice-versa*).

Observations influentes



Remarques

L'identification des observations influentes est un **processus itératif** car les différentes analyses doivent être exécutées à plusieurs reprises.

L'identification et la suppression entièrement automatisées des observations anormales **ne sont PAS recommandées**.

Utilisez des transformations de données si les données **ne sont PAS normalement distribuées**.

Le fait qu'une observation soit une valeur aberrante ou non dépend de **divers facteurs** ; les observations qui finissent par être influentes dépendent de **l'analyse spécifique à effectuer**.

Lectures suggérées

Les observations anormales

*Data Understanding, Data Analysis, Data Science
Data Preparation*

Anomalous Observations

- Anomaly Detection
- Outlier Tests
- Visual Outlier Detection

*Anomaly Detection and Outlier Analysis (avancé)

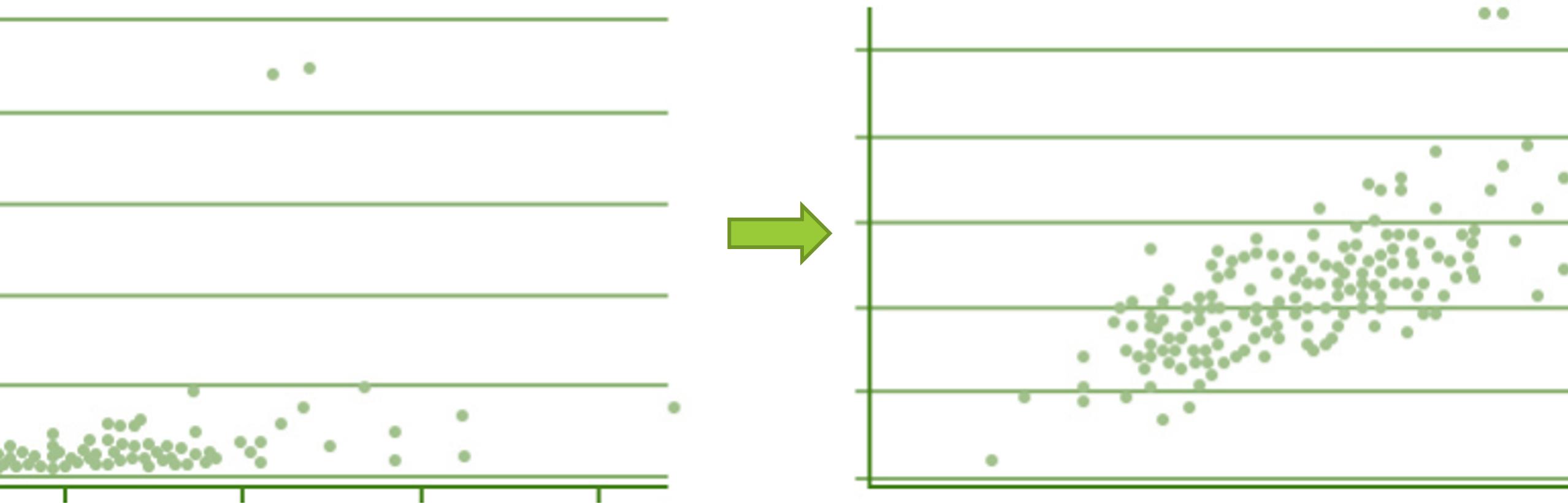
Exercices

Les observations anormales

1. Recréez le processus de détection des anomalies utilisé dans [Example: Algae Bloom](#).
2. Trouvez les observations anormales dans les ensembles de données [cities.txt](#) et [grades](#) (le cas échéant).
3. Trouvez les observations anormales dans un ensemble de données de votre choix.

Session 4

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES



10. La dimensionnalité et les transformations de données

La dimensionnalité des données

En analyse des données, la **dimension** est le nombre d'attributs qui sont rassemblés dans un ensemble de données (le **nombre de colonnes**).

Nous pouvons considérer le nombre de variables utilisées pour décrire chaque objet (ligne) comme un vecteur décrivant cet objet : la dimension est simplement la **taille** de ce vecteur.

(**Remarque** : le terme "dimension" est utilisé différemment dans les contextes de "business intelligence")

Dimensionnalité élevée et “Big Data”

Les ensembles de données peuvent être “massifs” de différentes manières :

- trop grand pour la **gestion** (ne peut être stockée, accédée, manipulée correctement en raison du nombre d'observations, du nombre de caractéristiques, de la taille globale)
- les dimensions peuvent aller à l'encontre des **hypothèses de modélisation** (# de caractéristiques > # d'observations)

Exemples :

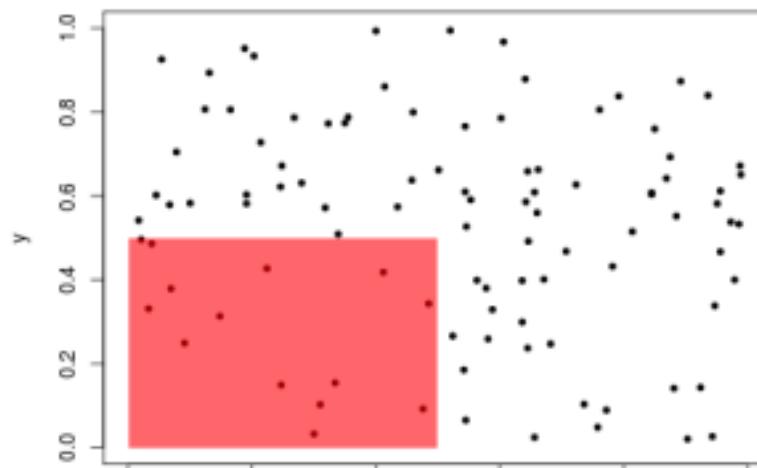
- plusieurs capteurs enregistrant plus de 100 observations par seconde dans une vaste zone géographique sur une longue période = **données massives**
- dans la *matrice terme-document* d'un corpus (colonnes = termes, rangées = documents), le nombre de termes est généralement beaucoup plus élevé que le nombre de documents, ce qui conduit à des **données éparses**

Le fléau de la dimensionnalité

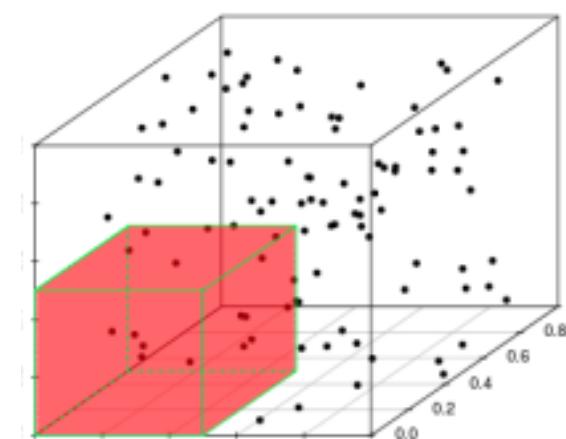
42% des données sont capturées



14% des données sont capturées



7% des données sont capturées



$N = 100$ observations, uniformément distribuées sur $[0,1]^d, d = 1, 2, 3$.
% des observations capturées par $[0,0.5]^d, d = 1, 2, 3$.

L'échantillonnage d'observations

Question : est-ce que toutes les données doivent être utilisée ?

Si les rangées sont sélectionnées au hasard (avec/sans remise), l'échantillon résultant peut être **représentatif** de l'ensemble des données.

Inconvénients :

- si le signal d'intérêt est rare, l'échantillonnage peut le noyer complètement
- si l'agrégation se produit en fin de parcours, l'échantillonnage affectera nécessairement les chiffres (passagers vs. vols)
- sur un fichier massif, même les opérations simples (e.g., trouver le # d'instances) peuvent être coûteuses – utilisez des **informations préalables sur la structure de l'ensemble** !

La sélection de caractéristiques

La suppression des variables **non pertinentes/redondantes** est une tâche courante du traitement des données.

Motivations :

- les outils de modélisation ne les gèrent pas bien (inflation de la variance, etc.)
- réduction de la dimension (# variables \gg # observations)

Approches :

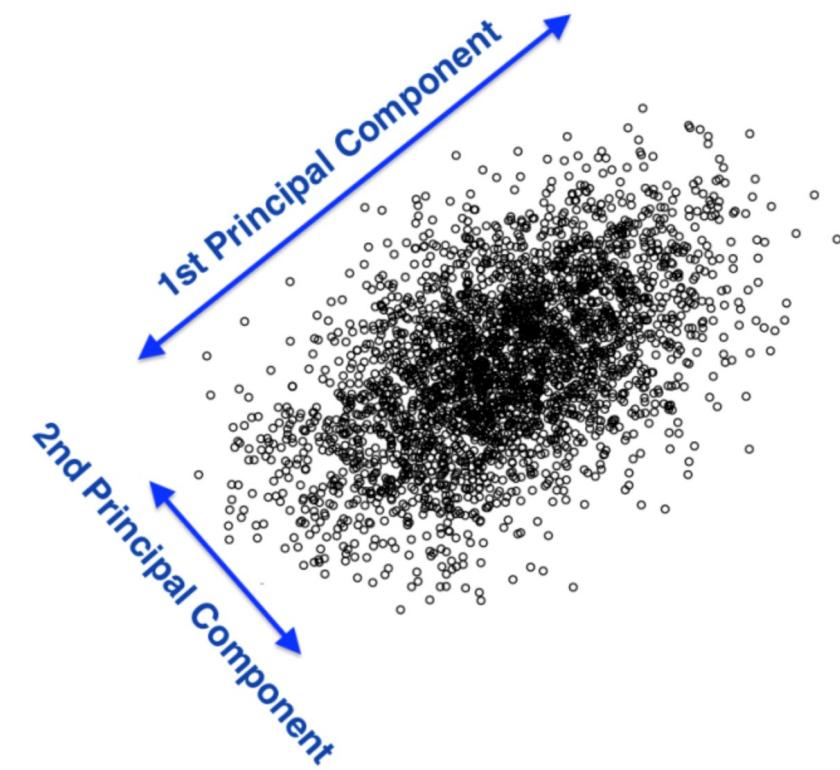
- filtre vs. enveloppe (“filter” vs. “wrapper”)
- non supervisé vs. supervisé

La réduction de dimension : ACP

Motivation : contenu nutritionnel des aliments

Quelle est la meilleure façon de différencier les produits alimentaires ? La teneur en vitamines, en matières grasses, ou en protéines ? Un peu de tout ?

L'analyse en composantes principales (ACP) peut être utilisée pour trouver les combinaisons de variables le long desquelles les observations sont **les plus répartis** (réduction de la dimension).



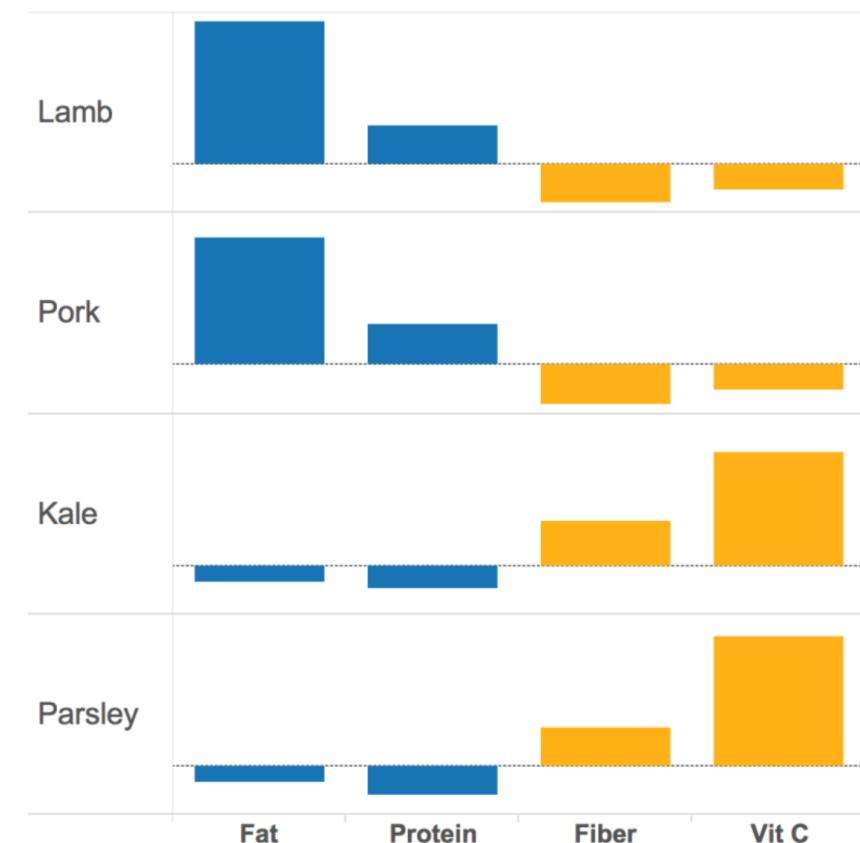
La réduction de dimension : ACP

La présence de nutriments semble être **corrélée** entre les différents aliments.

Dans un (petit) échantillon, les niveaux de *graisses* et de *protéines* semblent en phase, tout comme ceux des *fibres* et de la *vitamine C*.

Dans un ensemble de données plus vaste, les corrélations sont $r = 0.56$ et $r = 0.57$

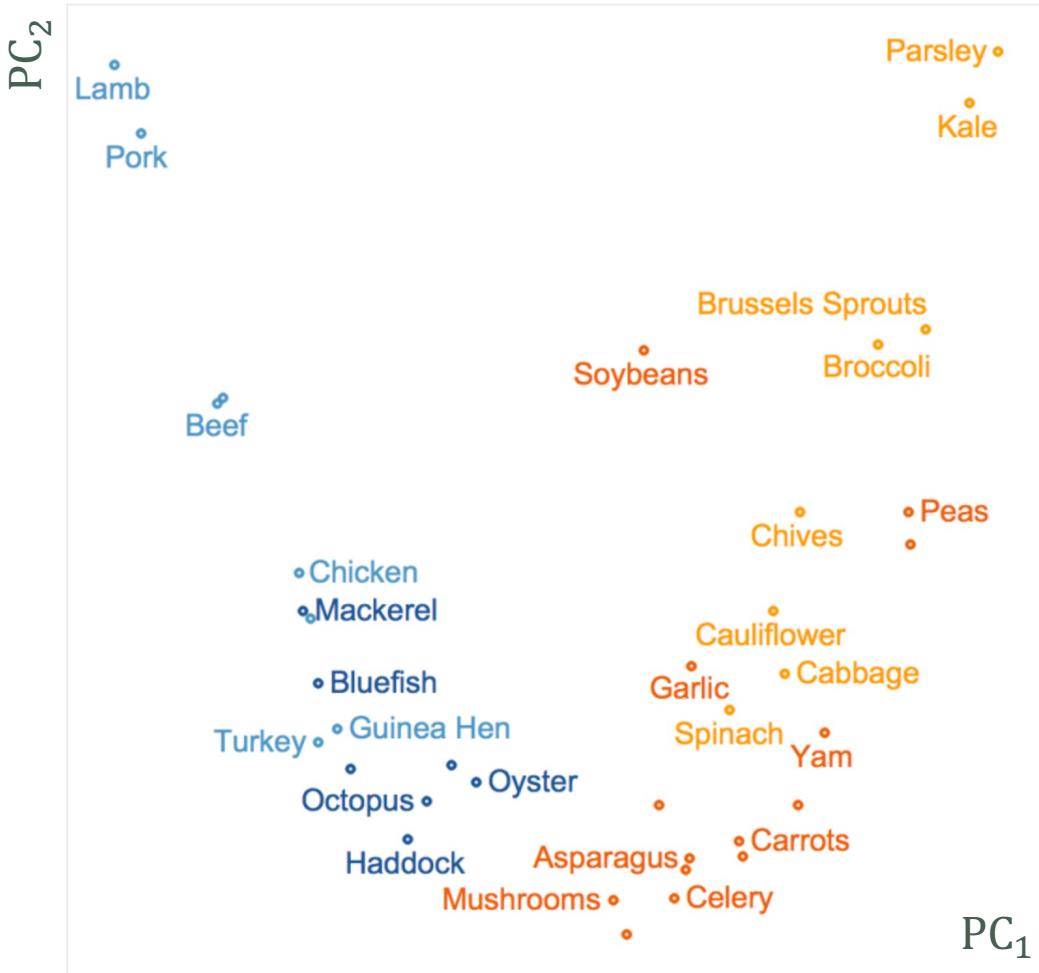
2 variables **dérivées** peuvent-elles expliquer cela ?



$$PC_1 = -0.45 \times \text{Fat} - 0.55 \times \text{Protein} + 0.55 \times \text{Fiber} + 0.44 \times \text{Vitamin C}$$

$$PC_2 = 0.66 \times \text{Fat} + 0.21 \times \text{Protein} + 0.19 \times \text{Fiber} + 0.70 \times \text{Vitamin C}$$

La différenciation ACP



différencie les légumes des viandes ; différencie 2 **sous-catégories** au sein de celles-ci :

- les **viandes** sont concentrées sur la gauche (PC_1 faibles)
- les **légumes** sont concentrés sur la droite (PC_1 élevé)
- les **fruits de mer** ont une plus faible teneur en *matières grasses* (PC_2 faible) et sont concentrés en bas
- les **légumes non feuillus** ont une teneur plus faible en *vitamine C* (PC_2 faible) et sont également regroupés en bas

Les transformations communes

Les modèles exigent parfois que certaines hypothèses relatives aux données soient respectées (normalité des résidus, linéarité, etc.).

Si les données brutes ne répondent pas aux exigences, nous pouvons soit :

- abandonner le modèle
- tenter de **transformer** les données

La deuxième approche nécessite une **transformation inverse** pour pouvoir tirer des conclusions sur les **données d'origine**.

Les transformations communes

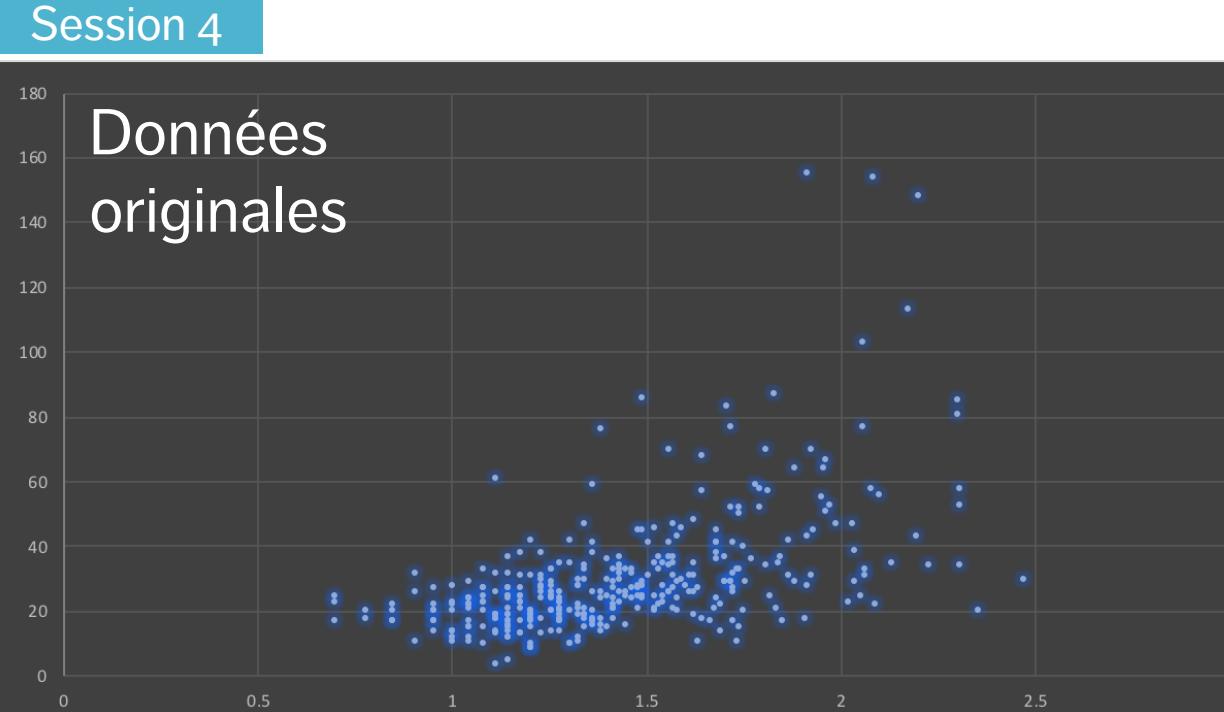
Dans le contexte de l'analyse des données, les transformations sont **monotones** :

- logarithmique
- racine carrée, inverse, puissance :
- exponentielle
- Box-Cox, etc.

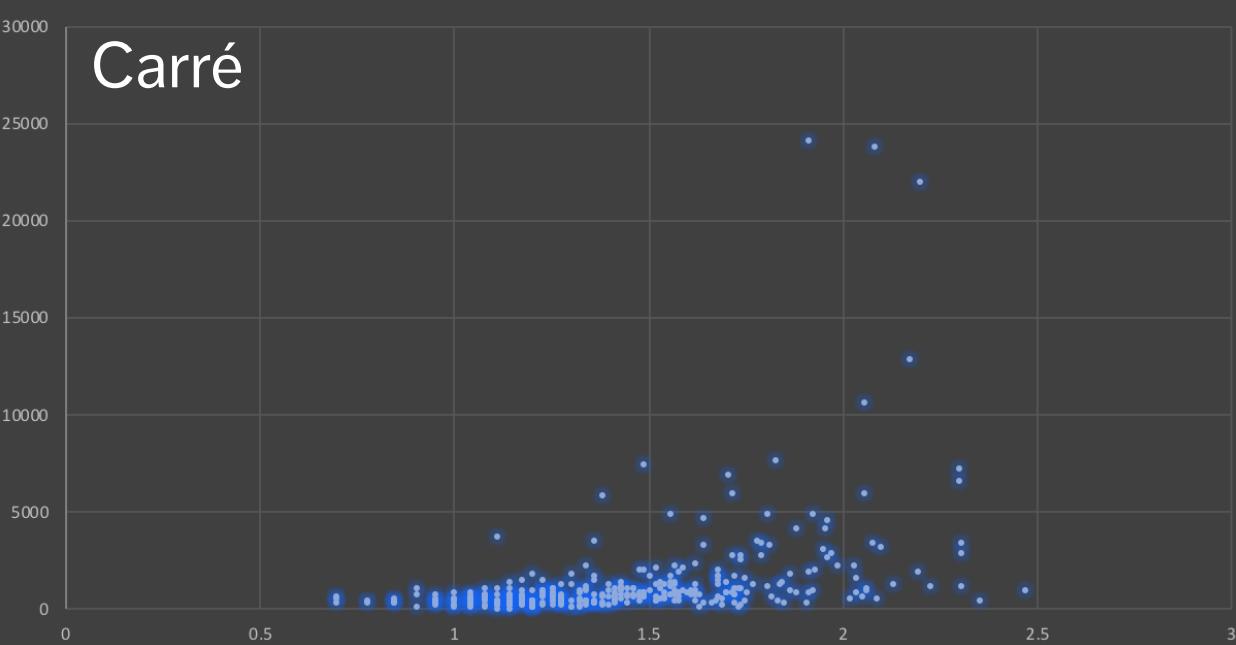
Les transformations sur les prédicteurs X peuvent atteindre la linéarité, mais à un prix (les corrélations ne sont pas préservées, par exemple).

Les transformations sur la réponse Y peuvent aider avec la non-normalité et la variance inégale des termes d'erreur.

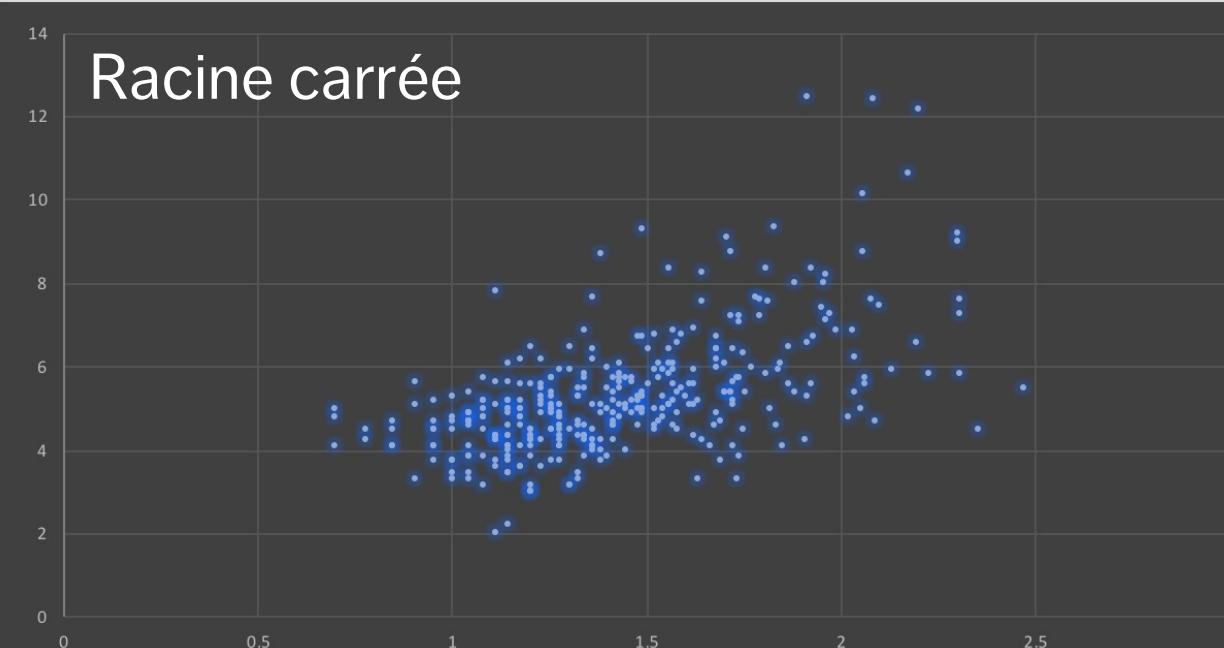
Données originales



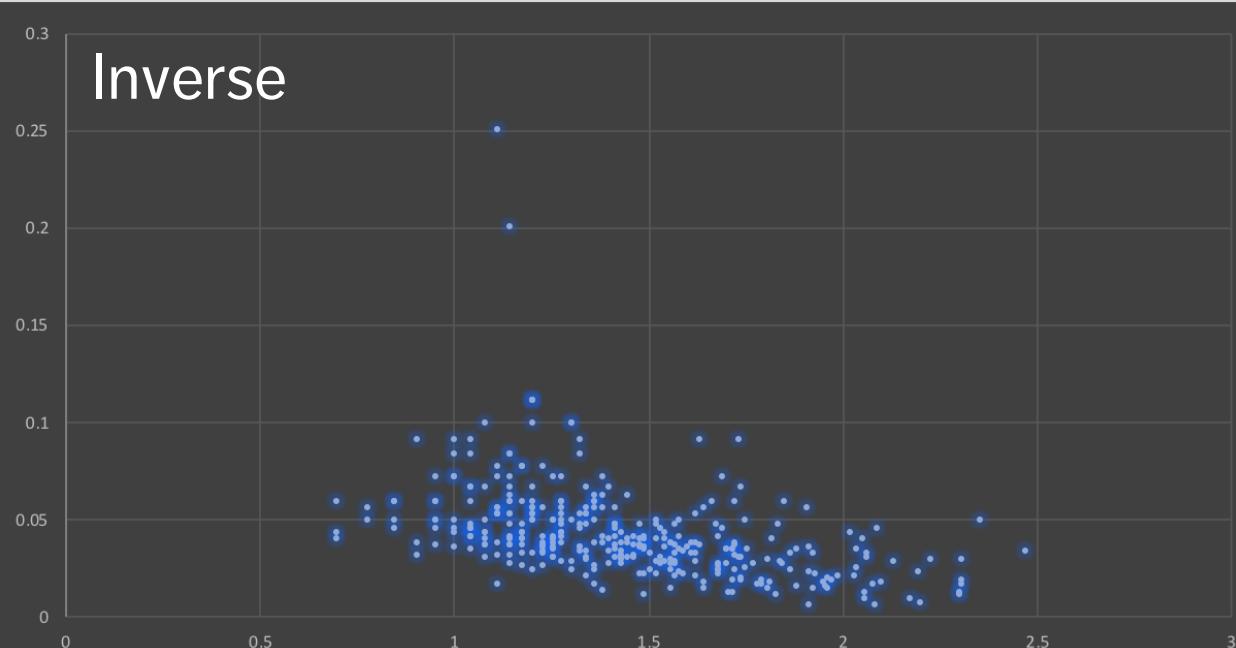
Carré



Racine carrée



Inverse



La transformation de Box-Cox

Supposons le modèle habituel $Y_j = \sum_i \beta_i X_{j,i} + \varepsilon_j$ avec soit

- des résidus asymétriques ;
- une variance non constante, et/ou
- une tendance non linéaire.

La **transformation de Box-Cox** $Y_j \mapsto Y_j'(\lambda)$ suggère un choix : sélectionnez λ qui maximise la log-vraisemblance correspondante

$$Y_j'(\lambda) = \begin{cases} \text{gm}(Y) \times \ln(Y_j), & \lambda = 0 \\ \lambda^{-1} \text{gm}(Y)^{1-\lambda} \times (Y_j^\lambda - 1), & \lambda \neq 0 \end{cases}$$

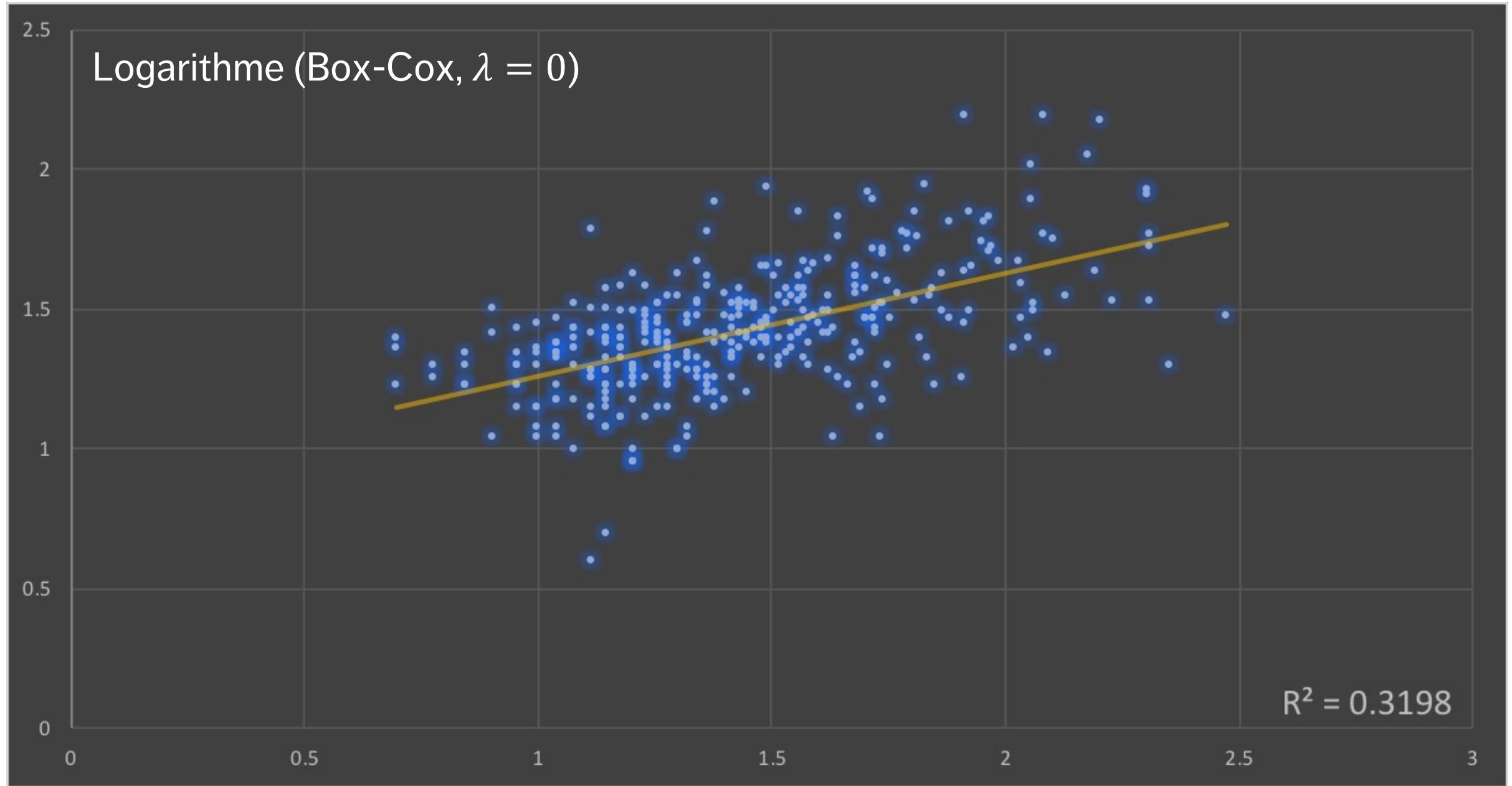
La transformation de Box-Cox

La procédure fournit un **guide** pour sélectionner une transformation.

Des justifications théoriques/pratiques peuvent exister pour un choix de λ .

Une analyse résiduelle est encore nécessaire pour s'assurer que le choix était approprié.

Mieux vaut travailler avec (ou interpréter) les données **transformées**.



La mise à l'échelle

Les variables numériques peuvent avoir différentes **échelles** (e.g., des poids et des hauteurs).

La variance d'une variable à grande échelle est généralement supérieure à celle d'une variable à petite échelle, ce qui peut introduire un biais.

La **standardisation** crée une variable avec une moyenne 0 et un écart-type 1 :

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

La **normalisation** crée une variable dans l'intervalle [0,1]: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

La discrétisation

Pour réduire la complexité des calculs, il peut être nécessaire de remplacer une variable numérique par une variable **ordinale** (e.g., passer de la *taille* à "petit", "moyen", "grand").

L'**expertise de domaine** peut être utilisée pour déterminer les limites des bacs (bien que cela puisse introduire un biais inconscient dans les analyses).

En absence d'une telle expertise, on peut fixer les limites de sorte que soit :

- les bacs contiennent chacun le même nombre d'observations
- les bacs ont tous la même largeur
- la performance d'un certain outil de modélisation est maximisée

La création de variables

Il peut être nécessaire d'introduire de nouvelles variables :

- des **relations fonctionnelles** d'un certain sous-ensemble de caractéristiques disponibles
- pour imposer l'**indépendance des observations**
- pour imposer l'**indépendance des caractéristiques**
- pour simplifier l'analyse en examinant des **résumés agrégés** (en analyse de texte)

Dépendances temporelles → analyse des séries chronologiques (décalages ?)

Dépendances spatiales → analyse spatiale (voisins ?)

Lectures suggérées

La dimensionnalité et les transformations de données

*Data Understanding, Data Analysis, Data Science
Data Preparation*

Data Transformations

- Common Transformations
- Box-Cox Transformations
- Scaling
- Discretizing
- Creating Variables

*Feature Selection and Dimension Reduction (advanced)

Exercices

La dimensionnalité et les transformations de données

1. En utilisant [Example: Algae Bloom](#) comme base, mettez à l'échelle, discrétisez et créez de nouvelles variables à partir de l'ensemble de données `algae blooms`.
2. Mettez à l'échelle, discrétisez et créez de nouvelles variables à partir des ensembles de données `grades` et [cities.txt](#).
3. Mettez à l'échelle, discrétisez et créez de nouvelles variables à partir d'un ensemble de données de votre choix.

Miscellanea

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

11. L'ingénierie des données

Contexte

L'un des défis de la science des données : mettre de grandes quantités de données dans des formats pouvant être **lus** par des algorithmes.

L'ingénierie des données est liée au traitement de ces données.

Après le traitement, les scientifiques des données développent des **preuves de concept** ; les ingénieurs IA/AA les traduisent en **modèles déployables**.

L'ingénierie des données existe depuis un certain ; avec l'essor du “**cloud computing**”, l'expertise dans ce domaine devient aussi recherchée que celle en analyse de données (du moins, dans certains cercles).

Rôles et responsabilités (reprise)

Ingénieurs en données (ID)

- recevoir des données d'une source
- structurer, distribuer et stocker les données dans des lacs et des entrepôts de données
- créer des outils et des modèles de données que les SD utilisent

Ingénieurs AA

- déployer de modèles de données
- combler les écarts entre ID et SD
- faire passer des idées de validation de concept à grande échelle

Scientifiques des données

- recevoir des données procurées/fournies par l'ID
- extraire la valeur des données
- construire des modèles prédictifs de preuve de concept
- mesurer et améliorer les résultats
- construire des modèles analytiques

Rôles et responsabilités (reprise)

Dans les petites organisations, l'ingénierie et la science des données sont généralement **regroupées** dans sous un même toit.

Les grandes entreprises disposent d'ingénieurs de données **spécialisés**, qui construisent des **pipelines de données** et gèrent des **entrepôts de données** (en les alimentant en données et en créant des schémas de table pour assurer le suivi des données stockées).

En général, ID \neq SD.

Les pipelines de données

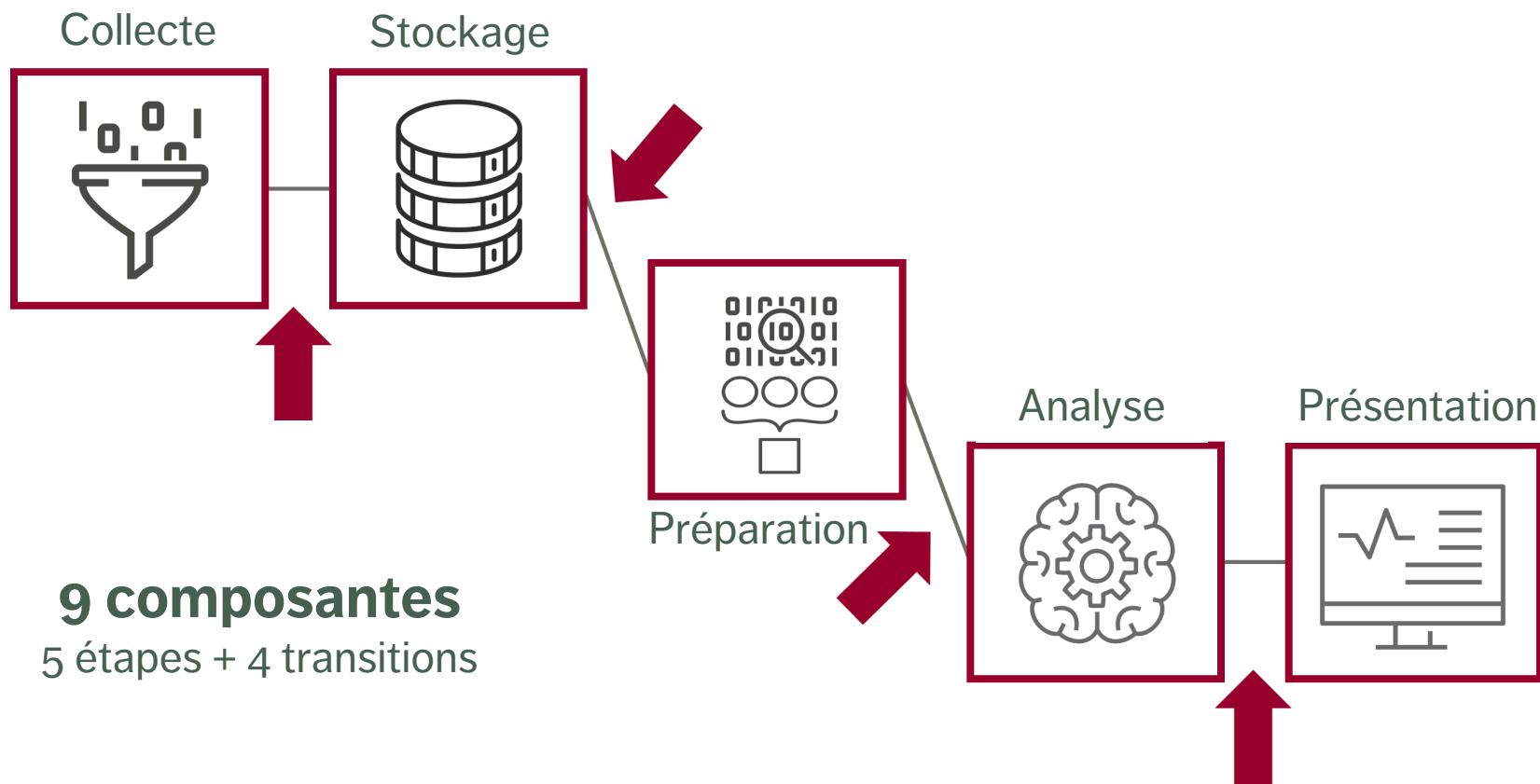
Ingénierie des **données**

- les opérations qui créent des **interfaces** et des **mécanismes** pour le flux/l'accès à l'information
- mise en place d'une **infrastructure de données**, préparation des données pour une analyse plus poussée par des SD

Les données peuvent provenir de nombreuses **sources** (et types de sources), et dans une variété de formats et de tailles.

Transformer tout cela en un processus que les SD peuvent utiliser et dont ils peuvent tirer du sens est connu sous le nom de **construction d'un pipeline de données**.

Les pipelines de données



Les pipelines de données

Principal défi en matière d'ingénierie des données :

- construire un pipeline qui **s'exécute en temps réel** (ou presque) **à chaque fois qu'il est sollicité**
- afin que les utilisateurs obtiennent des **informations actualisées** avec des **délais minimaux**

Les pipelines conceptuels sont transmis aux ingénieurs AA pour le **déploiement** et la **production**. Certains des travaux entourant cette tâche comprennent :

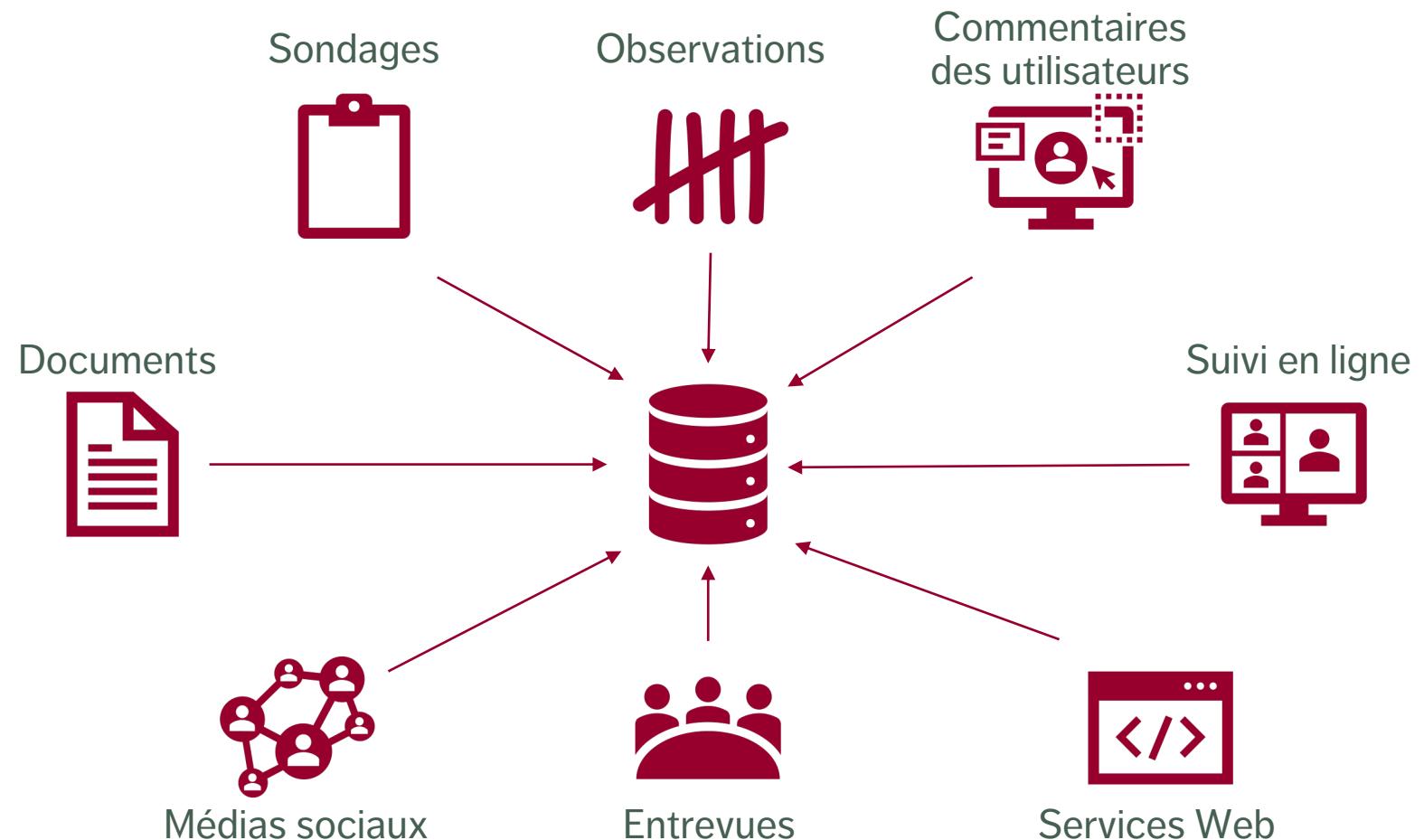
- contrôles de la qualité des données
- optimisation de la performance des requêtes
- la création d'un écosystème d'intégration/livraison continue pour les changements de modèles
- ingestion des données provenant de diverses sources dans le modèle de données
- transfert des techniques d'AA et de SD aux systèmes distribués

Les pipelines de données

Thèmes communs (opérations/framework/tâches/sources) pour les étapes du pipeline :

- **collecte de données** : applications, applications mobiles, microservices, dispositifs de l'Internet des objets (IoT), sites web, instrumentation, journalisation, capteurs, données externes, contenu généré par l'utilisateur, etc.
- **le stockage des données** : Gestion des données de référence (MDM), entrepôt, lac de données, etc.
- **intégration/préparation des données** : ETL, intégration de données en flux, etc.
- **analyse des données** : apprentissage automatique, analyse prédictive, tests A/B, expériences, intelligence artificielle (IA), apprentissage profond, etc.
- **livraison et présentations** : tableaux de bord, rapports, microservices, notifications push, email, SMS, etc.

La collecte de données



ETC – Extraire



ETC – Transformer

Structure



Types de données



Agrégation



Nettoyage



Rejoindre



Regroupement



Extraire

Transformer

Charger

ETC – Transformer



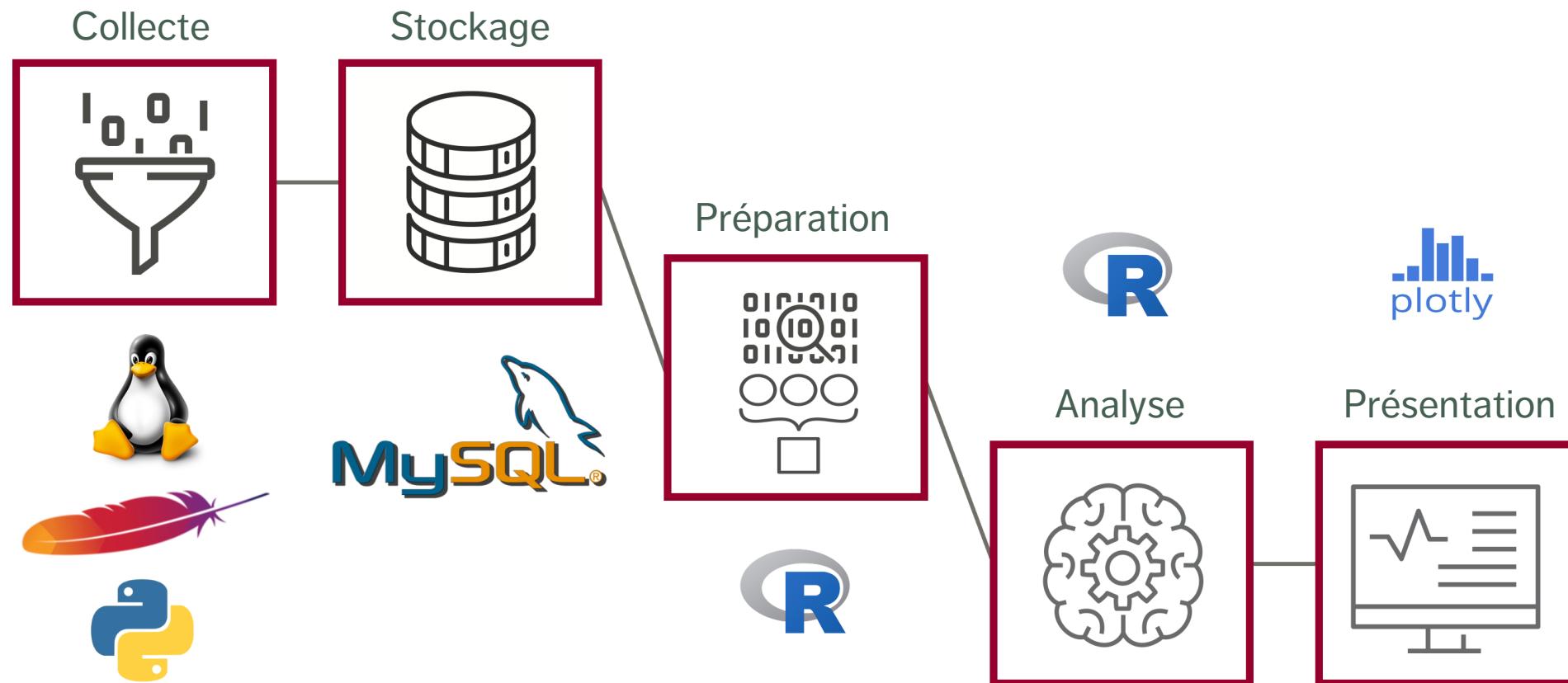
Google Cloud

Extraire

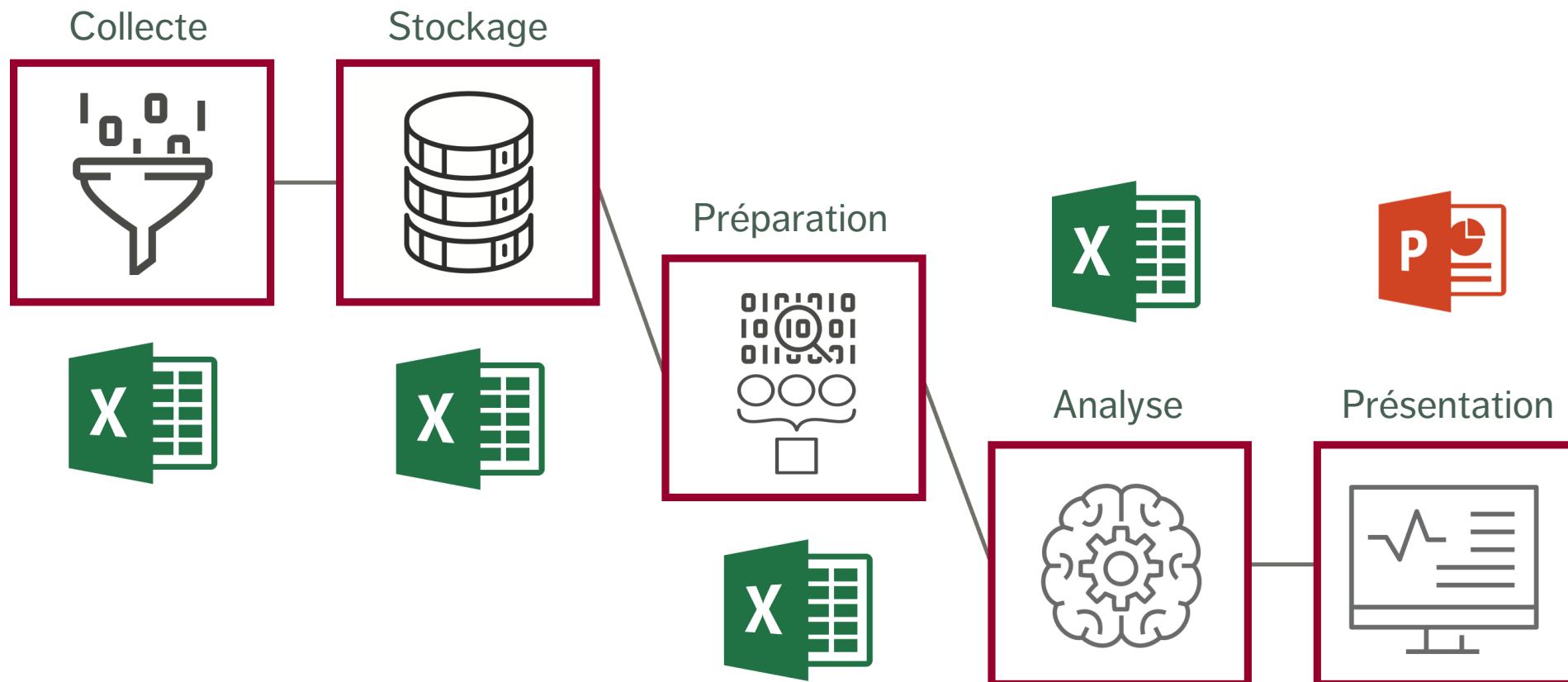
Transformer

Charger

Un pipeline de données “Open Source”



Un pipeline de données GdC (?)



Les outils de pipelines de données

Les pipelines permettent aux utilisateurs de diviser les tâches importantes en une série de petites étapes séquentielles, ce qui peut aider à **optimiser** chaque étape.

E.g., si vous utilisez TensorFlow pour la composante d'analyse d'un pipeline DL qui consiste en un seul grand script, **tout**, de la collecte des données à la présentation, doit utiliser TensorFlow, ce qui peut ne pas être optimal.

Les outils de pipeline de données sélectionnent le meilleur cadre/langage pour chaque composante/tâche du pipeline :

- Luigi (Spotify)
- Airflow (AirBnB)
- scikit-learn
- pandas/tidyverse
- etc.

Les outils d'ingénierie des données

Il est peu probable qu'un ID puisse maîtriser tous les outils d'ingénierie de données possibles, mais les équipes ID ont une plus grande **couverture** :

- **bases de données analytiques** (Big Query, Redshift, Synapse, etc.)
- **ETC** (Spark, Databricks, DataFlow, DataPrep, etc.)
- **moteurs de calcul évolutifs** (GKE, AKS, EC2, DataProc, etc.)
- **orchestration de processus** (AirFlow/Cloud Composer, Bat, Azure Data Factory, etc.)
- **déploiement et mise à l'échelle de plateforme** (Terraform, outils personnalisés, etc.)
- **outils de visualisation** (Power BI, Tableau, Google Data Studio, D3.js, ggplot2, etc.)
- **programmation** (tidyverse, numpy, pandas, matplotlib, scikit-learn, scipy, Spark, Scala, Java, SQL, T-SQL, H-SQL, PL/SQL, etc.)



La gouvernance des données

La gouvernance des données englobe :

- les **personnes** ;
- les **processus**, et
- les **technologie de l'information**

On l'utilise pour créer un traitement **cohérent/approprié** des données d'une organisation à travers l'entreprise.

Elle fournit la base, la stratégie, et la structure pour garantir que les données sont gérées comme un **actif** et transformées en informations **significatives**.

La gouv. des données

Objectifs :

- création d'une culture de données libre service
- établir des règles internes pour leur utilisation
- mettre en œuvre les exigences de conformité
- améliorer les communications
- augmenter la valeur des données
- réduire les coûts associés aux données
- gérer continuellement les risques
- assurer une existence continue



Lectures suggérées

L'ingénierie des données

*Data Understanding, Data Analysis, Data Science
Data Engineering and Management*

Background and Context

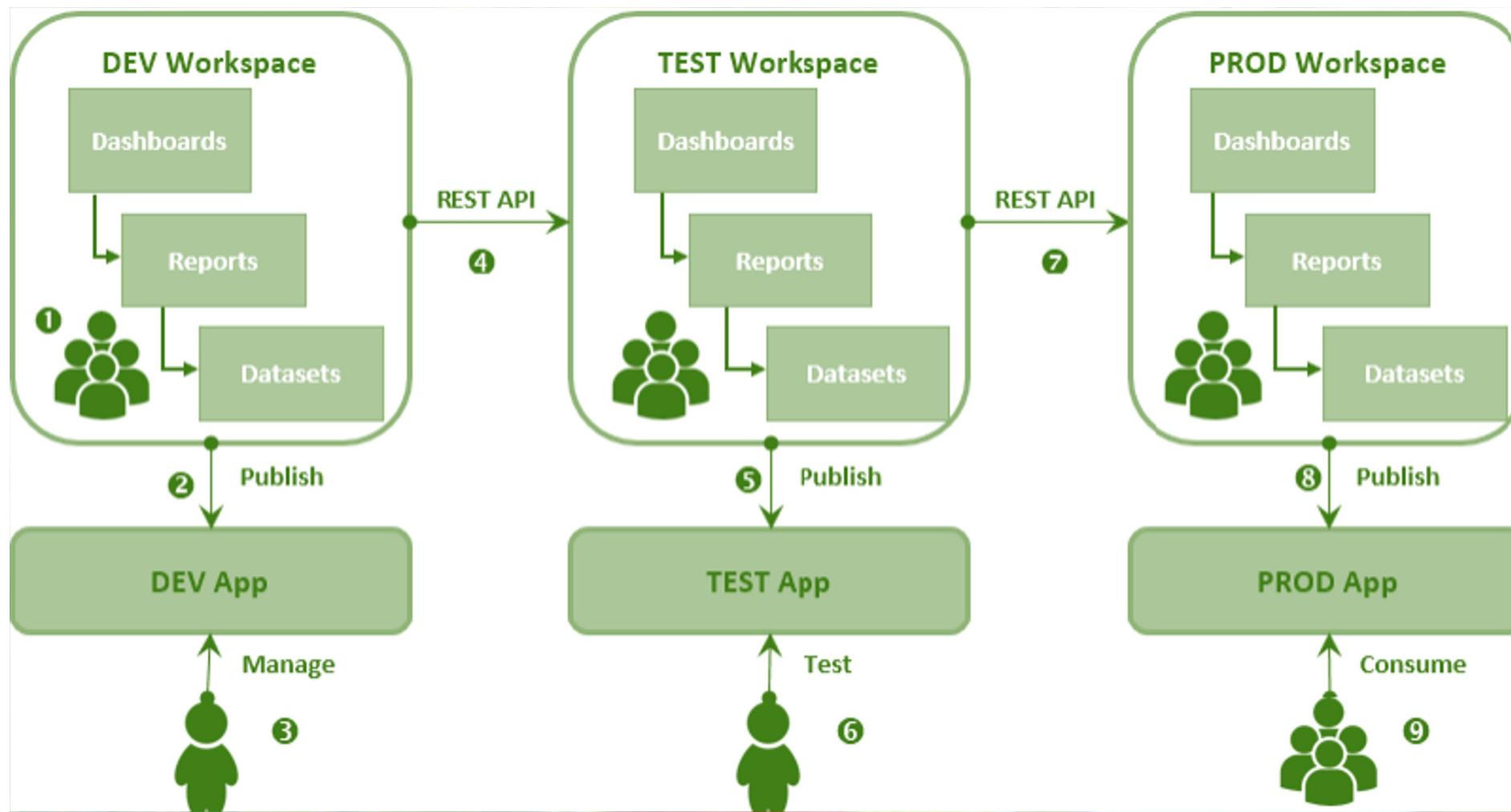
Data Engineering

- Data Pipelines
- Automatic Deployment and Operations
- Scheduled Pipelines and Workflows
- Data Engineering Tools

Exercices

L'ingénierie des données

1. À quoi ressemble votre pipeline de science des données (ou celui de votre organisation) ? Pourrait-il être amélioré ?
2. Identifiez des cas où vous avez rencontré des problèmes liés à la disponibilité, la facilité d'utilisation, la cohérence, l'intégrité, la qualité, la sécurité ou la fiabilité des données.
3. Complétez tous les exercices précédents que vous n'avez pas eu l'occasion de terminer.



12. La gestion des données

Quelques concepts fondamentaux

Les **données** et les **connaissances** doivent être structurées de manière à pouvoir être :

- stockées et accessibles
- modifiables et ajoutables
- extraites utilement et efficacement (extraire - transformer - charger)
- exploitées par des **humains** et des **ordinateurs** (programmes, bots, IA)

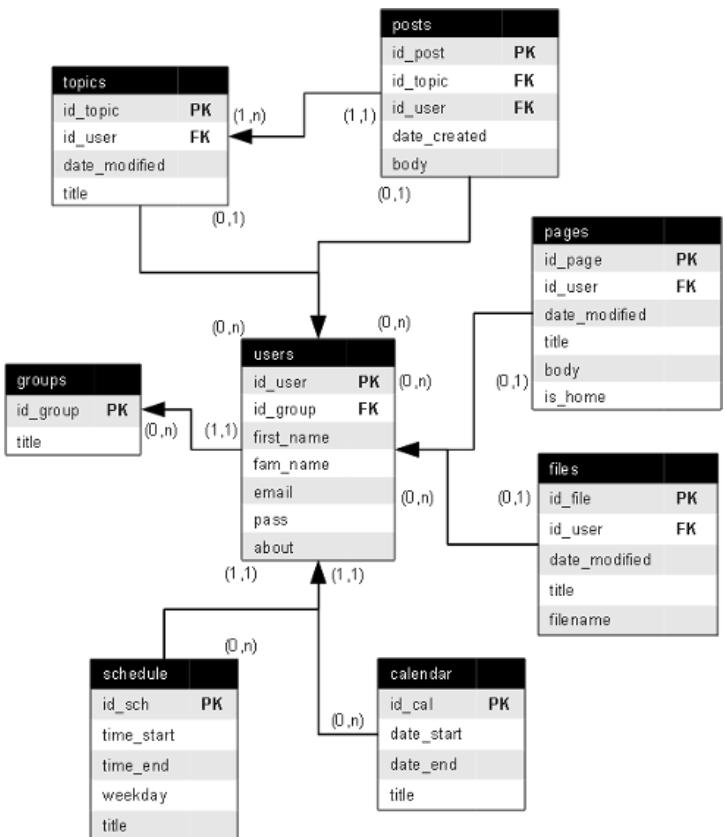
La modélisation des données

Les modèles de données sont des descriptions **abstraites/logiques** d'un système, utilisant des termes qui sont implémentables en tant que structure d'un type de logiciel de gestion des données.

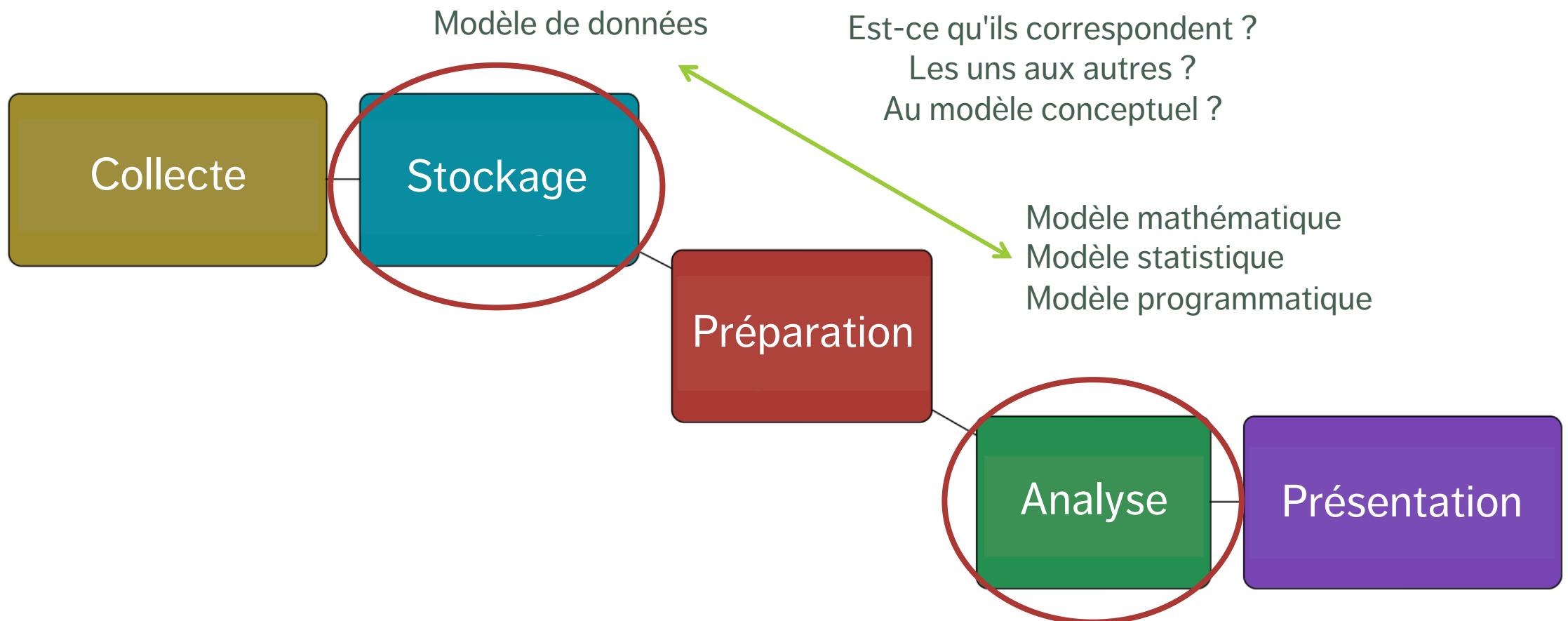
Cela se trouve à mi-chemin entre un **modèle conceptuel** et une **implémentation de banque de données**.

Les données elles-mêmes concernent les **instances** – le modèle, quant à lui, concerne les **types d'objets**.

Une autre option à envisager : les **ontologies**.



Un pipeline de données automatisé



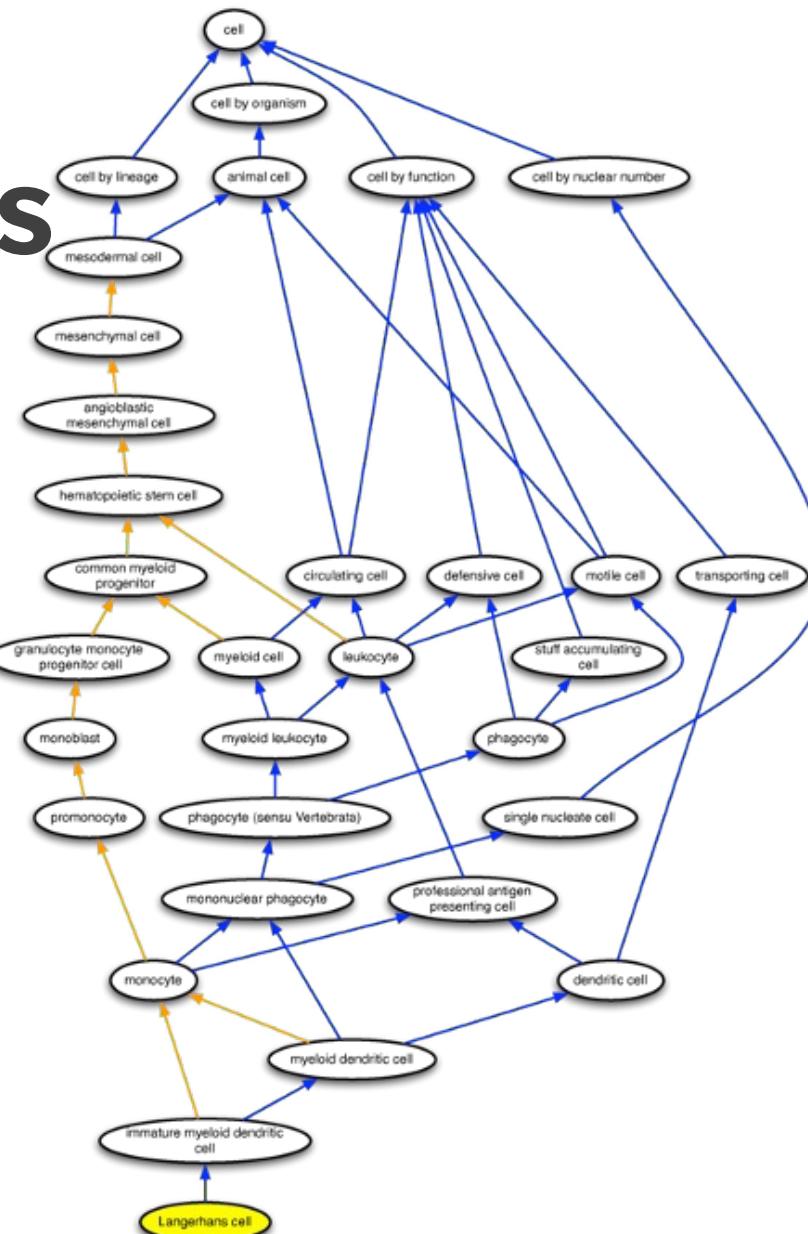
Métadonnées contextuelles

Nous perdons quelque chose lorsque nous passons de notre modèle conceptuel à un modèle de type spécifique – p. ex. le modèle de données ou de connaissances.

Une façon de conserver le contexte est de fournir des **métadonnées** (riches, si possible) – des données sur nos données!

Les métadonnées sont essentielles lorsqu'il s'agit de mettre en œuvre des stratégies pour travailler d'un ensemble de données à l'autre.

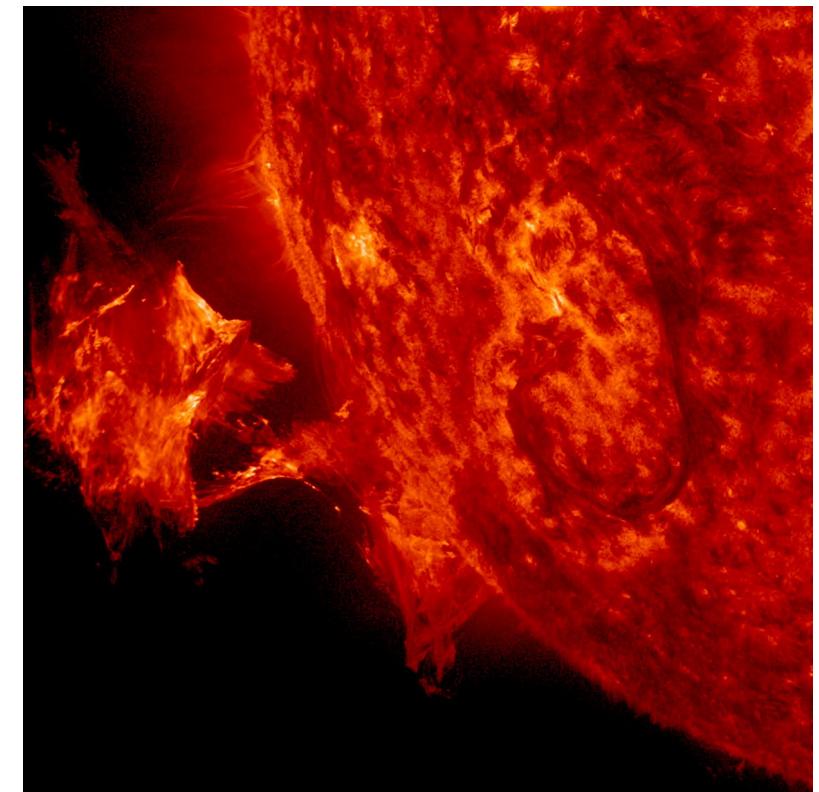
Les **ontologies** peuvent aussi jouer un rôle ici!



Les données (non) structurées

La disponibilité croissante de données non structurées et de grands objets binaires (**blob**) est l'une des principales motivations de certains des nouveaux développements dans les types de bases de données et autres stratégies de stockage de données :

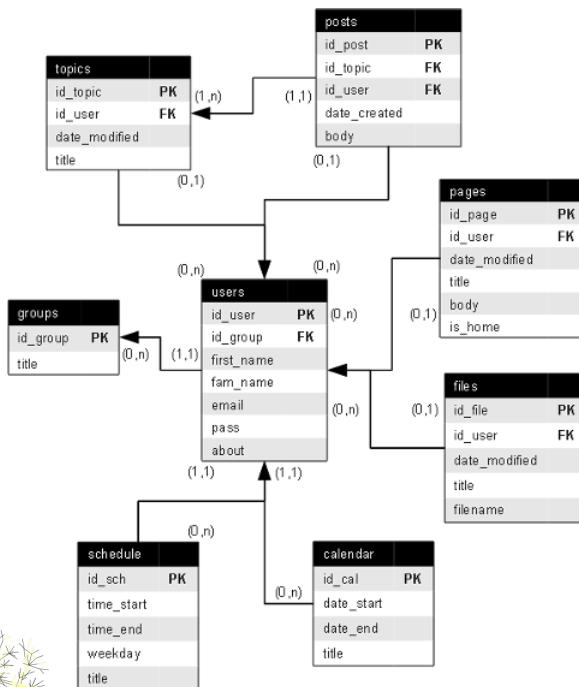
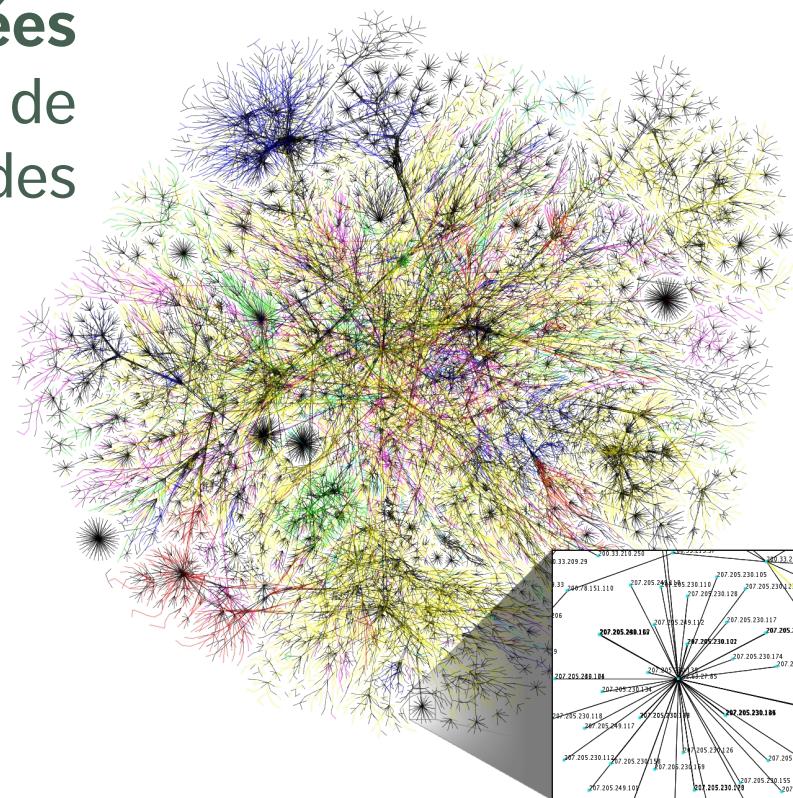
- **données structurées** : étiquetées, organisées, discrètes, selon une structure limitée et prédéfinie
- **données non structurées** : non organisées, pas de modèle de données à structure spécifique prédéfinie
- **données blob** : grand objet binaire – images, audio, multimédia



La modélisation des données

Différentes options sont actuellement populaires en termes de **données fondamentales** et de stratégies de modélisation ou de structuration des **connaissances**:

- paires valeur-clé (e.g., JSON)
- triples (e.g., RDF)
- bases de données graphiques
- bases de données relationnelles
- feuilles de calcul



Les mémoires et les bases de données

Base de données relationnelle :

- largement soutenue, bien comprise, fonctionne bien pour de nombreux types de systèmes et de cas d'utilisation. Base toutefois difficile à changer une fois mise en œuvre; ne gère pas bien les liens.

Magasins de clés-valeurs :

- peuvent prendre n'importe quel type de données; nul besoin de beaucoup de renseignements sur la structure ; si vous avez beaucoup de valeurs manquantes, ces mémoires ne prendront pas de place ; peuvent toutefois être désordonnées et mystérieuses; difficile d'y trouver des données.

Bases de données graphiques :

- rapides et intuitives si vous utilisez des données fortement axées sur les liens; pourraient être la seule option si vos données sont ainsi parce que les bases de données traditionnelles peuvent ralentir énormément ; sont souvent trop spécialisées ; pas encore supportées à grande échelle.

Les fichiers “plats” et les feuilles de calcul

Pour :

- très efficace si vous recueillez des données une seule fois, sur un type particulier d'objet
- certains types d'analyse exigent que vous ayez toutes les données en un seul endroit
- facile à lire dans un logiciel et à effectuer des opérations sur l'ensemble des données

Cons :

- très difficile de gérer l'intégrité des données si l'on collecte continuellement des données
- pas idéal pour les données de systèmes impliquant de multiples types d'objets et de relations
- il peut être très difficile d'effectuer des opérations d'interrogation de données

Quelques outils et mots-clés

- MongoDB, ArangoDB
- Magasin de documents
- JSON, YAML
- API, GraphQL
- Données interreliées
- Web sémantique
- Langage d'ontologie Web (OWL)
- Protégé
- SQL, etc.

La mise en œuvre du modèle

Pour mettre en œuvre votre modèle de données/connaissances, il faut avoir accès à un **logiciel de stockage et de gestion des données**.

Cela peut constituer un défi pour les particuliers : ces logiciels fonctionnent généralement sur des **serveurs**.

Les serveurs sont utiles car ils permettent à plusieurs utilisateurs d'accéder **simultanément** à une même base de données, à partir de différents programmes clients, mais il est difficile de "jouer" avec les données.

C'est là que **SQLite** entre en jeu.

Le rôle du logiciel de gestion des données

Les logiciels de gestion des données offrent aux utilisateurs un moyen facile d'interagir avec leurs données.

Il s'agit essentiellement d'une interface entre les **personnes** et les **données**.

Grâce à cette interface, les utilisateurs peuvent :

- ajouter des données à leur collection de données
- extraire des sous-ensembles de données de leur collection en fonction de certains critères
- supprimer ou modifier des données dans leur collection

Un peu de terminologie

Auparavant :

- base de données
- entrepôt de données
- mini-entrepôt de données
- système de gestion des données
- (SQL)

Maintenant :

- lac de données
- bassin de données
- marais de données ?
- cimetière de données ?
- (NoSQL)

De plus en plus : on fait une distinction entre l'**entrepot de données** et le **logiciel de gestion des données**.

Du modèle de données à la mise en œuvre

Une fois que le mode de données (logique) est achevé :

1. **instancier le modèle** dans le logiciel choisi (par exemple, créer des tables dans MySQL)
2. **télécharger/charger les données**
3. **interroger les données** :
 - les bases de données relationnelles traditionnelles utilisent le **langage de requête structuré** (SQL : Structured Query Language)
 - d'autres utilisent des langages de requête différents (AQL, moteurs sémantiques, etc.) ou s'appuient sur des programmes informatiques sur mesure (par exemple, écrits en R, Python)

La gestion des bases de données

Une fois les données collectées, il faut aussi les **gérer**.

Fondamentalement, cela signifie que la base de données doit être **maintenue**, afin que les données soient

- **précises**
- **exactes**
- **cohérentes**
- **complètes**

Ne laissez pas votre lac de données se transformer en marais de données !

Services en nuage (Cloud Services)



1. Stocker de **grandes** quantités de données
2. Exécutez des processus coûteux et avancés en **cliquant sur un bouton**
3. **Flexible** et **évolutif**
4. Permettre le traitement des données **en code bas**

Nuage vs. accès local

Nuage (Cloud)



sans intervention manuelle

paiement à la consommation

propriétaire douteux

Accès local (On-Premise)



auto-entretenu

tous les coûts sont absorbés

sécurité entièrement contrôlée

Lectures suggérées

La gestion des données

Data Understanding, Data Analysis, Data Science **Data Science Basics**

Getting Insight From Data

- Structuring and Organizing Data

Data Engineering and Management

Data Management

- Databases
- Data Modeling
- Data Storage

Reporting and Deployment

- Reports and Products
- Cloud and On-Premise Architecture

Exercices

La gestion des données

1. Votre organisation possède-t-elle des données ? Si oui, sont-elles hébergées localement ou sur le cloud ? Comment y accède-t-on ? Comment sont-elles structurées ?
2. Complétez tous les exercices précédents que vous n'avez pas eu l'occasion de terminer.

Exercices et projets guidés

LES BASES DE LA SCIENCE DES DONNÉES

Entre les sessions

Session 1 à Session 2

- terminez les exercices de la session 1
- téléchargez les ensembles de données depuis le site web
- lisez le [Programming Primer](#) (sections 1 - 4)
- installez [R / RStudio](#) (Posit)
- installez les librairies R suivantes : dplyr, xts, knitr, tidyverse, ggplot2, pastecs, Hmisc, e1071, psych, quantmod, ggm, kerndwd, MASS, DMwR, ROCR, car, forcats, corrplot

De la session 2 à la session 3

- terminez les exercices de la session 2

De la session 3 à Session 4

- terminez les exercices de la session 3

Après la session 4

- terminez les exercices de la session 4
- essayez les projets guidés

Projet guidé I

Sélectionnez un projet de données qui vous intéresse et fournissez une ébauche de planification pour celui-ci, en abordant les sujets abordés dans ce cours. Les questions suivantes peuvent vous aider :

1. Quelles sont les questions associées au projet ?
2. Quel est le modèle conceptuel de la situation sous-jacente ?
3. Quel type d'ensemble(s) de données existe(nt) qui pourrai(en) vous aider à répondre à ces questions ?
4. Y a-t-il des limites aux données ou à l'analyse ?
5. Devez-vous collecter de nouvelles données pour traiter ces questions ?
6. Comment les données sont-elles stockées/accédées ? Quelles sont les exigences en matière d'infrastructure ?
7. À quoi ressemblent les produits livrables ?
8. Comment les succès seraient-ils quantifiés/qualifiés ?
9. Quels sont vos délais et votre disponibilité ?
10. Quelles sont les compétences requises pour travailler sur ce projet ?
11. Travaillez-vous sur ce projet seul ou au sein d'une équipe ?
12. Quel serait le coût du lancement et de la réalisation de ce projet ?
13. À quoi ressemble le pipeline d'analyse des données ?
14. Quels logiciels et méthodes d'analyse seront utilisés ?

Projet guidé II

Rédigez un article discutant de certaines des questions éthiques entourant l'utilisation de l'intelligence artificielle, de la science des données et des algorithmes d'apprentissage automatique.

Établissez une liste des 3 principes éthiques les plus importants auxquels l'utilisation de tels algorithmes devrait se conformer. Expliquez pourquoi vous avez choisi chacun de ces principes.

Décrivez (au moins) 2 cas réels d'utilisation de l'I.A./S.D./A.A. dans le secteur public, le secteur privé, ou le milieu universitaire, lorsque les principes éthiques que vous avez choisis ont été violés. Discutez de la manière dont le non-respect de vos principes éthiques a occasionné (ou pourrait occasionner) des dommages à des personnes, des organisations, des pays, etc.

Suggérez comment les projets discutés ci-dessus auraient pu être modifiés afin que leur utilisation des algorithmes I.A./S.D./A.A. respecte les principes éthiques que vous avez choisis.

Projet guidé III

Ce projet utilise l'[outil Gapminder](#) (il y a aussi une version [hors-ligne](#))

1. Prenez le temps d'explorer l'outil. Dans la version en ligne, le point de départ par défaut est un graphique à bulles montrant l'espérance de vie en 2020, ainsi que le revenu par personne, par pays (la taille des bulles étant associée à la population totale). Dans la version hors ligne, sélectionnez l'option "Bubbles".
2. Pouvez-vous identifier les catégories de variables disponibles, ainsi que certaines des variables? [Vous devrez peut-être fouiller un peu].
3. Pourquoi pensez-vous que Gapminder ait choisi l'espérance de vie et le revenu par personne comme variables par défaut ?
4. Remplacez l'espérance de vie par le nombre de bébés par femme. Observez et discutez des changements par rapport au graphique par défaut.
5. Formulez quelques questions auxquelles vous pourriez répondre avec les données par défaut.
6. Formulez quelques questions auxquelles vous pourriez répondre en utilisant certaines des autres variables.
7. À quel moment du “flux de travail de la science des données” pensez-vous que des visualisations de cette nature pourraient être utiles ?
8. Ces visualisations permettent-elles de bien comprendre le système étudié (la Terre géopolitique) ?

Projet guidé III (suite)

9. Quelles sont, selon vous, les sources de données de l'ensemble de données sous-jacent? [Vous devrez peut-être fouiller sur Internet pour y répondre].
10. Toutes les variables et mesures sont-elles dignes de confiance? Comment pouvez-vous le déterminer?
11. L'ensemble de données sous-jacent est-il structuré ou non structuré?
12. Fournissez un modèle de données ("data model") potentiel pour l'ensemble de données sous-jacent.
13. Quels sont les types des 4 variables par défaut (espérance de vie, revenu, population, régions)?
14. Jouez un peu avec les graphiques. Pouvez-vous trouver des paires de variables qui sont positivement corrélées? Négativement corrélées? Non corrélées?
15. Parmi les variables qui sont corrélées, certaines vous semblent-elles présenter une relation dépendante-indépendante? Comment pouvez-vous identifier de telles paires?
16. Pouvez-vous fournir une estimation visuelle de la moyenne, de la médiane, et de l'étendue de diverses variables numériques?
17. Pouvez-vous estimer à vue d'œil le mode des variables catégorielles?
18. Pouvez-vous identifier des moments spéciaux (points temporels particuliers) dans les données, où un changement à longue haleine se produit, par exemple?
19. L'outil et son jeu de données sous-jacent sont-ils utilisables ? De quels facteurs dépend votre réponse ?

Projet guidé III (suite)

20. Pensez-vous qu'il pourrait y avoir des problèmes avec les valeurs rapportées ? Par exemple, sélectionnez la Suède et les États-Unis dans le menu de cases à cocher à droite et suivez leur parcours de 1799 à 2018/2020. À partir de quel moment les valeurs sont-elles raisonnables ? À votre avis, que se passe-t-il au début de la série chronologique ?
21. Suivez l'Érythrée pendant la même durée. Recherchez la date d'indépendance de ce pays (vis-à-vis de l'Éthiopie). A votre avis, que représentent les mesures antérieures à cette date ?
22. Suivez l'Autriche pendant la même durée. Recherchez la chronologie historique des frontières du pays (Autriche-Hongrie, Anschluss, frontières modernes, etc.). Qu'est-ce que cela implique pour les mesures rapportées ?
23. Suivez la Finlande pendant la même durée. Que se passe-t-il en 1809 ? Cela vous apprend-il quelque chose sur la façon dont les données sont codées dans l'ensemble de données ?
24. Désélectionnez tous les pays et laissez la simulation se dérouler de 1799 à 2018/2020. Pouvez-vous identifier des cas où un grand sous-ensemble d'observations se comporte de manière inattendue ? Si oui, pensez-vous que cela est dû à des problèmes de nettoyage/de traitement des données ?
25. Continuez à explorer l'ensemble de données. Vous pouvez modifier les variables affichées ou utiliser d'autres méthodes de visualisation. Globalement, pensez-vous que l'ensemble de données est fiables ? L'utiliseriez-vous pour effectuer des analyses ? Quelles sont ses forces et ses faiblesses ?

Projet guidé IV

Sélectionnez un ensemble de données dans la liste ci-dessous (ou tout autre ensemble qui vous intéresse) :

- [GlobalCitiesPBI.csv](#)
- [2016collisionsfinal.csv](#)
- [sondages_us_election_2016.csv](#)
- [HR_2016_Census_simple.xlsx](#)

Pour votre/vos ensemble(s) de données :

1. Créez un "dictionnaire de données" pour expliquer les différents champs et variables. Pouvez-vous trouver une source pour ces ensembles de données ?
2. Dressez une liste des questions auxquelles vous aimeriez obtenir des réponses sur ces données.
3. Étudiez les variables individuelles (au moyen de graphiques simples, de statistiques univariées, etc.)
4. Répétez le processus avec des paires de variables (par le biais de graphiques simples, de distributions conjointes, d'interactions entre variables, etc.)
5. Faites-vous confiance à l'ensemble de données, ou non ? Justifiez votre réponse. Si vous ne faites pas confiance à l'ensemble de données, signalez les entrées potentiellement invalides, les observations anormales, les valeurs manquantes ou les valeurs aberrantes. Comment ces entrées doivent-elles être traitées ?
6. Votre analyse suggère-t-elle que certaines des variables devraient être transformées ? L'une des questions que vous avez élaborées à l'étape 2 soutient-elle de telles transformations ? Si c'est le cas, transformez les données de manière appropriée.

Références

LES BASES DE LA SCIENCE DES DONNÉES

Références

- C. C. Aggarwal, *Data Classification: Algorithms and Applications*. CRC Press, 2015.
- C. C. Aggarwal, *Data Mining: The Textbook*. Cham: Springer, 2015.
- C. C. Aggarwal, C. K. Reddy, *Data Clustering: Algorithms and Applications*. CRC Press, 2014.
- P. Boily, *Data Understanding, Data Analysis, and Data Science*. Data Action Lab, 2022.
- D. Brin, *The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom?* Perseus, 1998.
- Coursera, “[Introduction to Data Engineering](#).”

Références

- T. H. Davenport and D. J. Patil, “[Data Scientist: The Sexiest Job of the 21st Century](#),” *Harvard Business Review*, Oct. 2012.
- T. Hastie, R. Tibshirani, and J. Friedman, [The Elements of Statistical Learning: Data Mining, Inference, and Prediction](#), 2nd ed. Springer, 2008.
- H. Konsek, “[Automating Data Pipelines: Types, Use Cases, Best Practices](#).” Soft Kraft.
- J. Kunigk, I. Buss, P. Wilkinson, and L. George, *Architecting Modern Data Platforms: A Guide to Enterprise Hadoop at Scale*. O'Reilly Media, 2018.
- T. Malaska and J. Seidman, *Foundations for Architecting Data Solutions: Managing Successful Data Projects*. O'Reilly Media, 2018.

Références

C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

T. Orchard and M. Woodbury, *A missing information principle: theory and applications*. University of California Press, 1972.

R. W. Paul and L. Elder, *Understanding the Foundations of Ethical Reasoning*, 2nd ed. Foundation for Critical Thinking, 2006.

What is Data Engineering? Everything You Need to Know in 2022. phData, 2022.

F. Provost and T. Fawcett, *Data Science for Business*. O'Reilly, 2015.

Références

T. Raghunathan, J. Lepkowski, J. Van Hoewyk, and P. Solenberger, “A multivariate technique for multiply imputing missing values using a sequence of regression models,” *Survey Methodology*, vol. 27, no. 1, pp. 85–95, 2001.

D. B. Rubin, *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

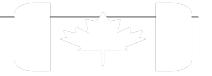
R. Schutt and C. O’Neill, *Doing Data Science: Straight Talk from the Front Line*. O'Reilly, 2013.

simplystatistics.org, “[An interactive visualization to teach about the curse of dimensionality](#).”

S. van Buuren, *Flexible imputation of missing data*. CRC Press, 2012.

A. Watt, [*Database Design*](#). BCCampus, 2014.

La gouvernance des données au sein du GdC



Point central de référence pour le GdC (Gouvernement à l'ère numérique) :

- [Plans stratégiques, politiques, normes et lignes directrices relatives aux services numériques du gouvernement](#)

Rapport au greffier du Conseil privé :

- [Feuille de route de la Stratégie de données pour la fonction publique fédérale](#)

Secrétariat du Conseil du Trésor (élection) :

- [Politique sur les services et le numérique](#)
- [Plan stratégique du GdC pour la gestion de l'information et la technologie de l'information de 2017 à 2021](#)
- [Plan stratégique des opérations numériques de 2018 à 2022](#)
- [Stratégie d'adoption de l'informatique en nuage du gouvernement du Canada : Mise à jour de 2018](#)

Innovation, Sciences et Développement économique Canada :

- [La Charte numérique du Canada en action: un plan par des Canadiens, pour les Canadiens](#)