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 catalyste 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 reconnait 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 ?

[R. Magritte, La Trahison des images]

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

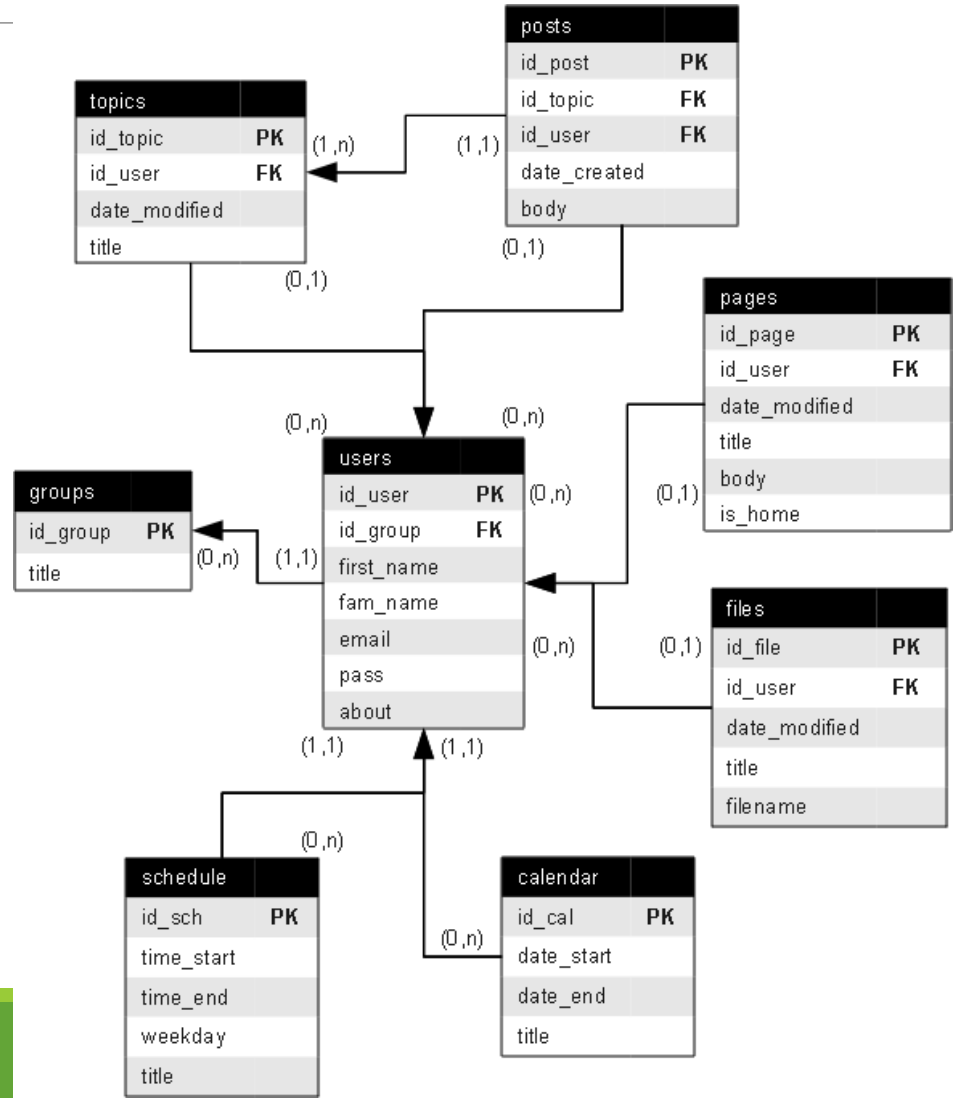
L'ensemble de données de ces objets physiques pourrait commencer par :

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
...	...	...	...	...	...

[Wooptoo, "Entity - relationship model". Wikimedia]

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

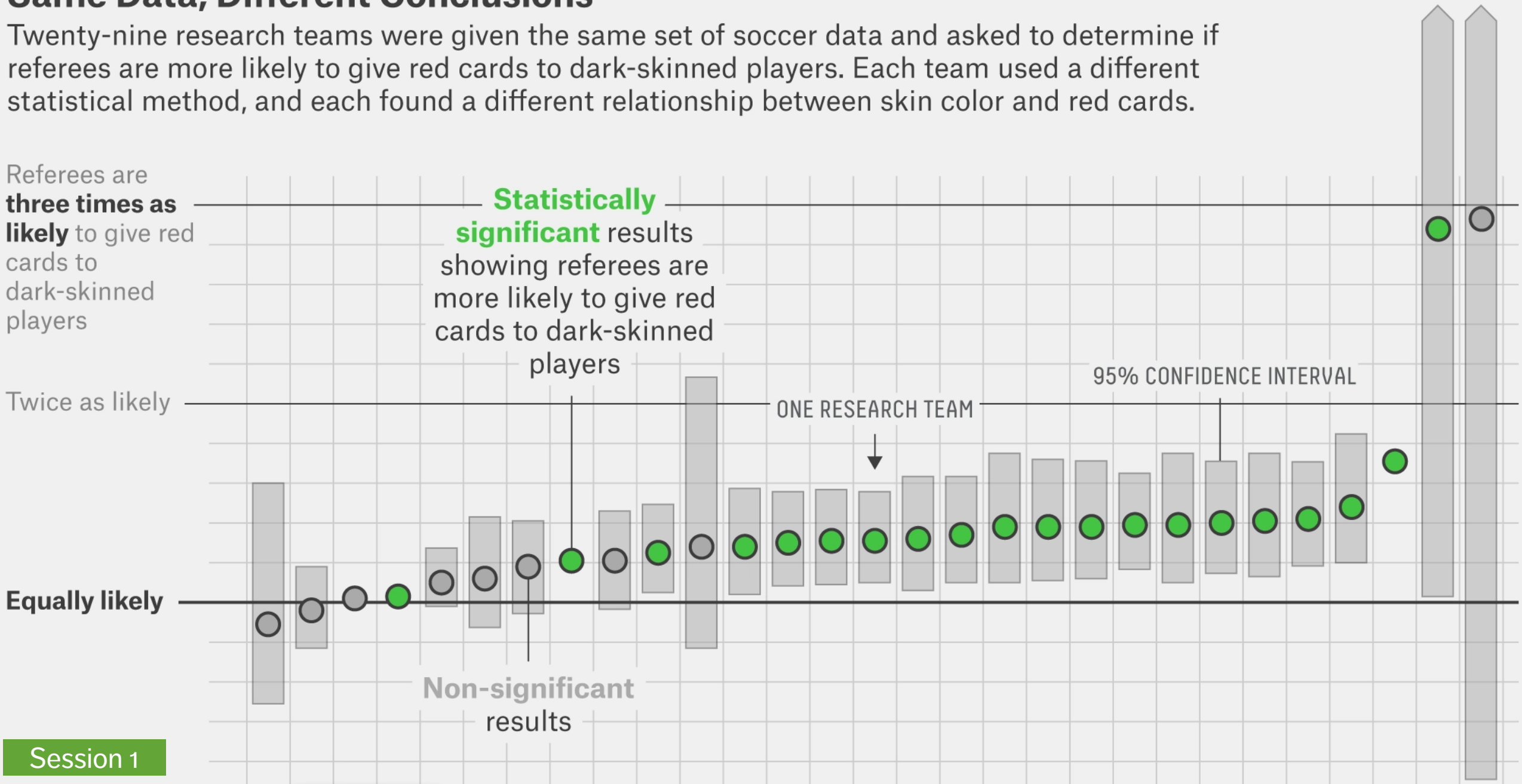
- "AlphaGo vainc le meilleur joueur de Go du monde, marquant la supériorité de l'IA sur l'esprit humain"
- "Une novella écrite par l'IA a presque gagné un prix littéraire"
- "Elon Musk : l'intelligence artificielle peut déclencher une troisième guerre mondiale"
- "L'engouement pour l'I.A. a atteint son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



Session 1

De : Science isn't broken - It's just a hell of a lot harder than we give it credit for. **[Christie Aschwanden, 2015]**

Lectures conseillées

Les préliminaires

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics

14.1 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

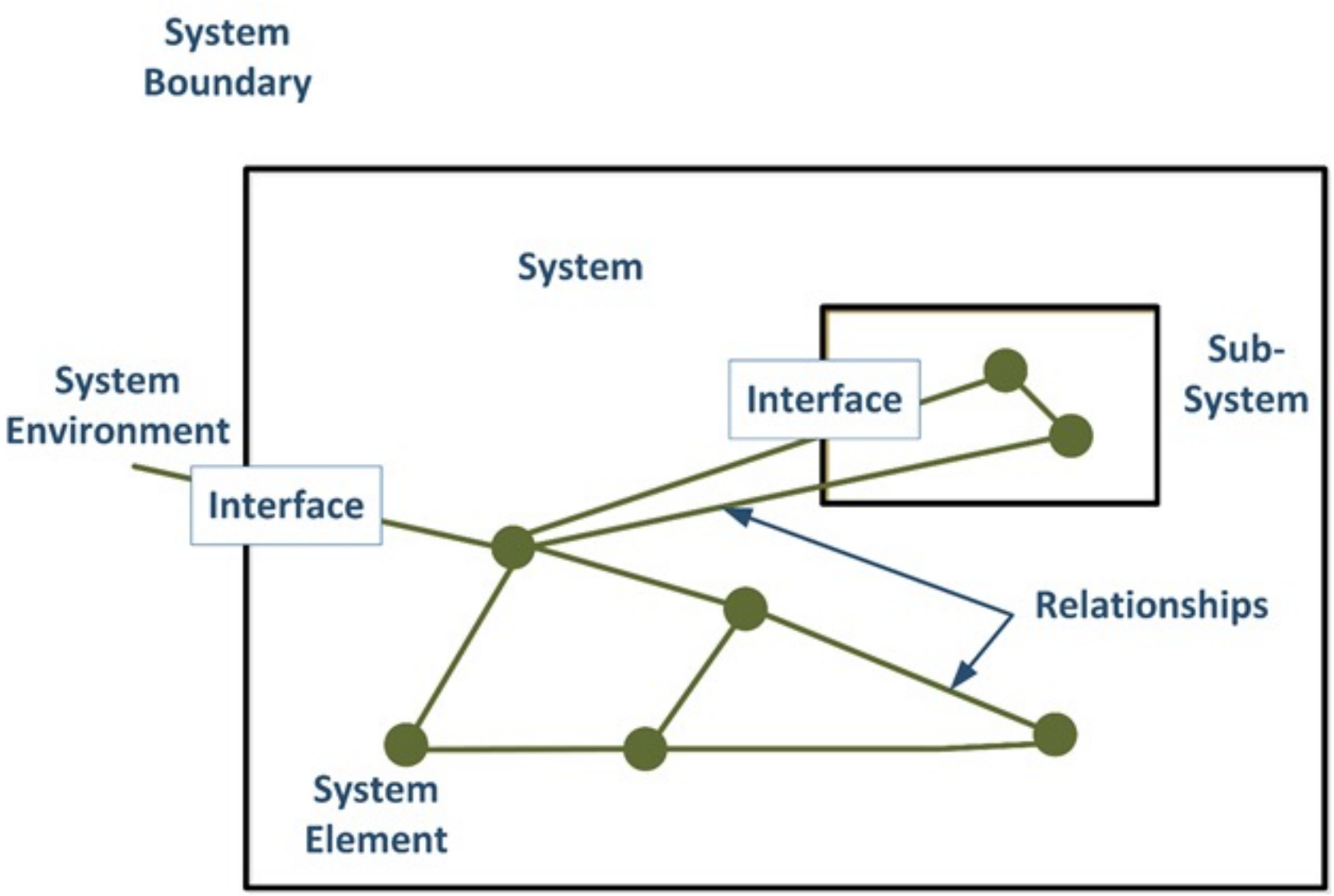
La théorie générale des systèmes décrit des phénomènes naturels à l'aide d'un **cadre conceptuel commun**, tous étant des systèmes d'objets en interaction.

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**.

[Deepak, Théorie générale des systèmes, Wikimedia]



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
- reconnaîssant 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 **découpant 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.

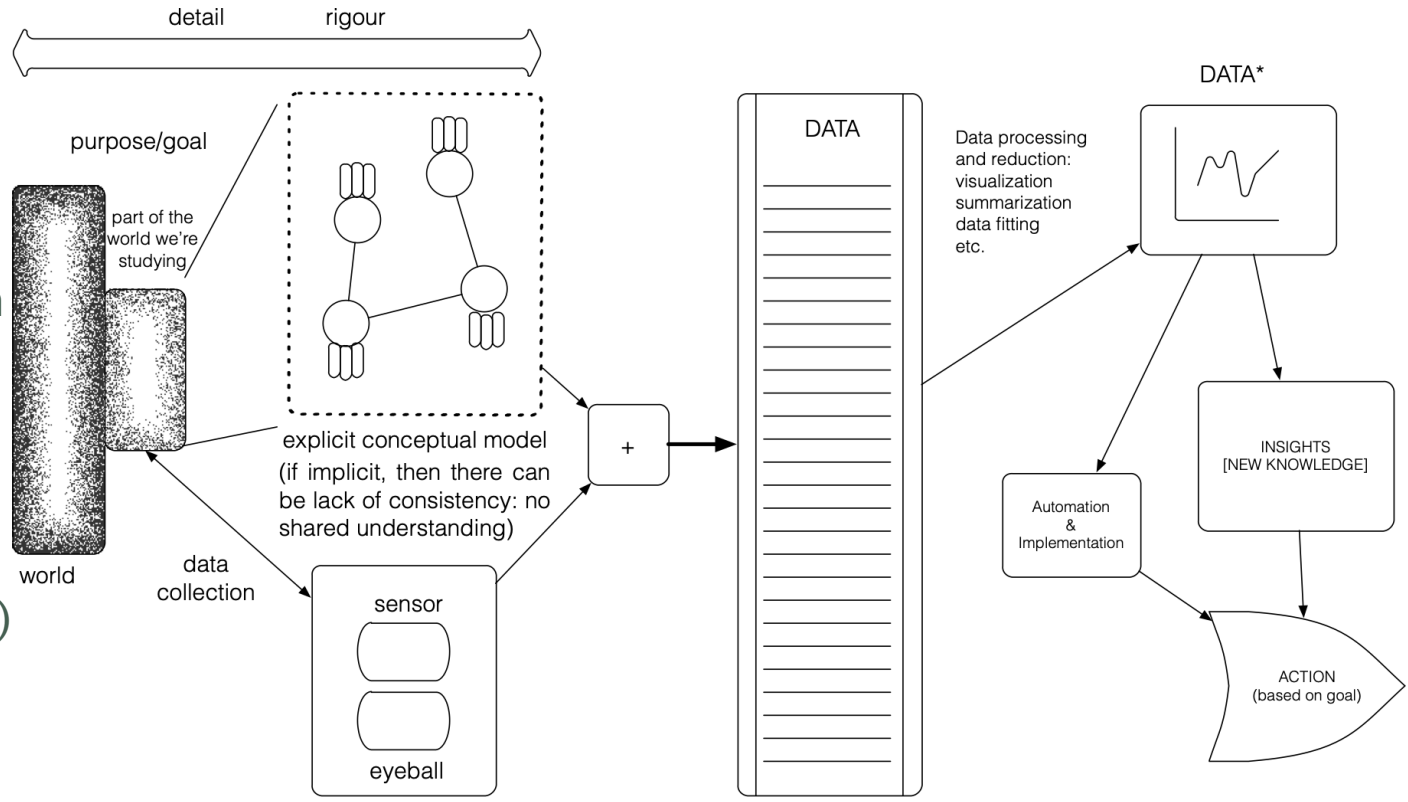
Les modèles conceptuels

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, definitions, 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**

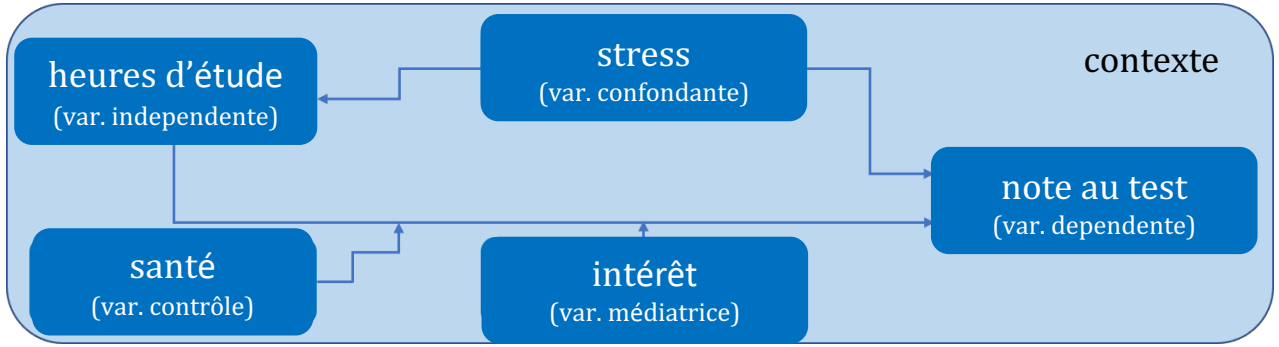


Les modèles conceptuels

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).



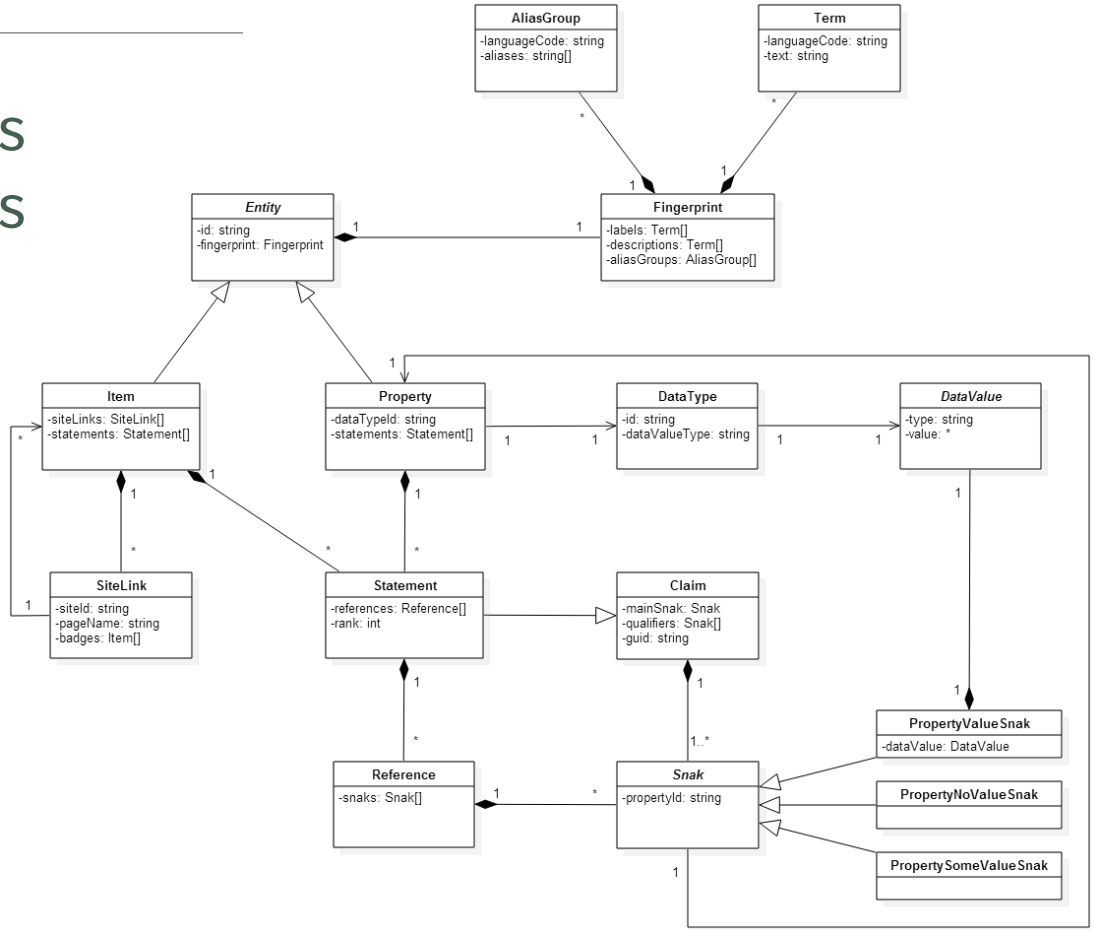
Les modèles conceptuels formels

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 appui 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 entire
- 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 conseillées

Les cadres conceptuels

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics
    14.2 Conceptual Frameworks
    - 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 ?

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 privilèges 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 **commenditaires** 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.
- 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 conseillées

L'éthique de la science des données

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics

14.3 Ethics in the Data Science Context

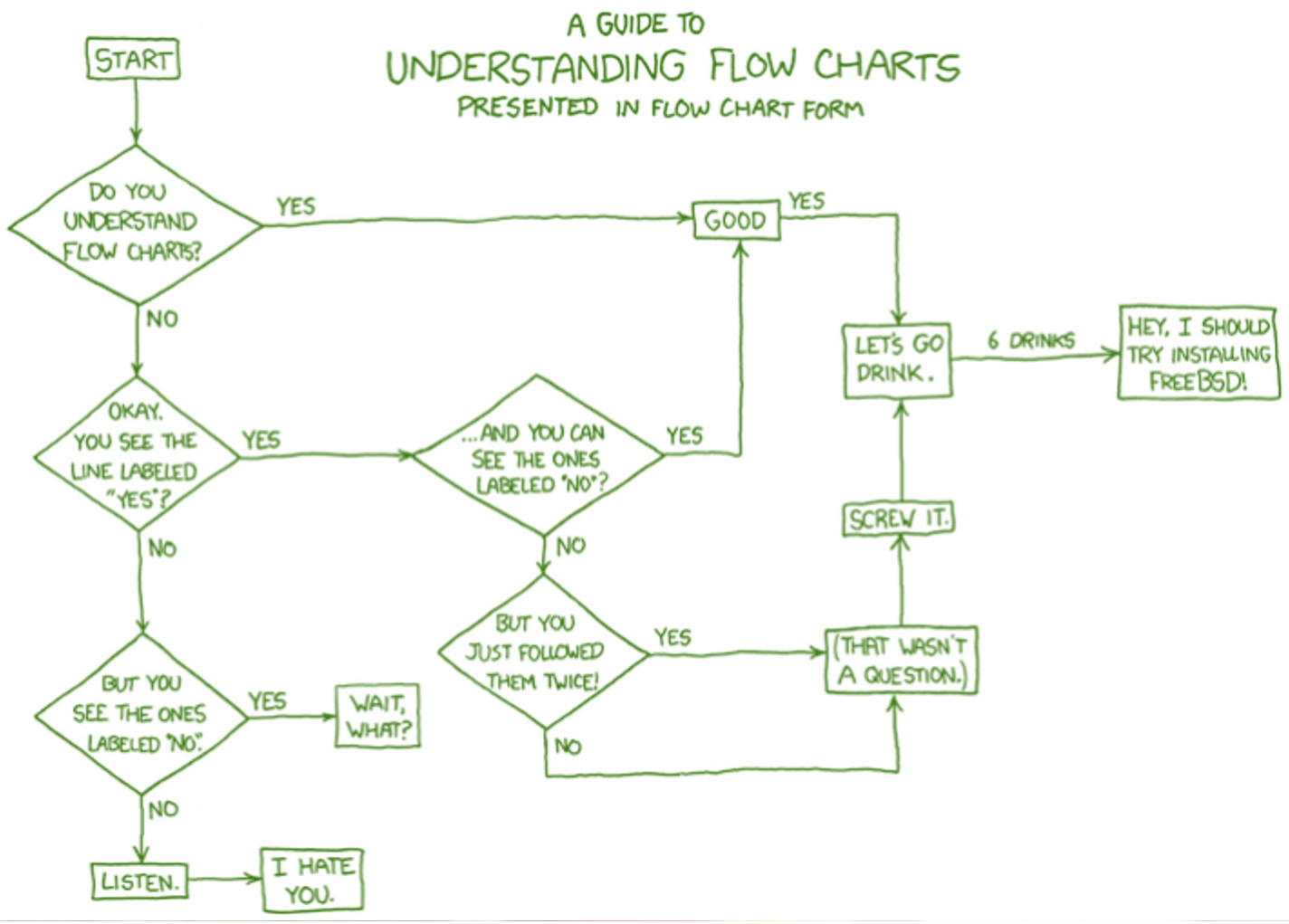
- The Need for Ethics
- What Is/Are Ethics?
- Ethics and Data Science
- Guiding Principles

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 ?



5. Le flux de travail analytique

Le flux de travail analytique

Vous en avez probablement assez des **discussions sur le contexte** et préféreriz 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”

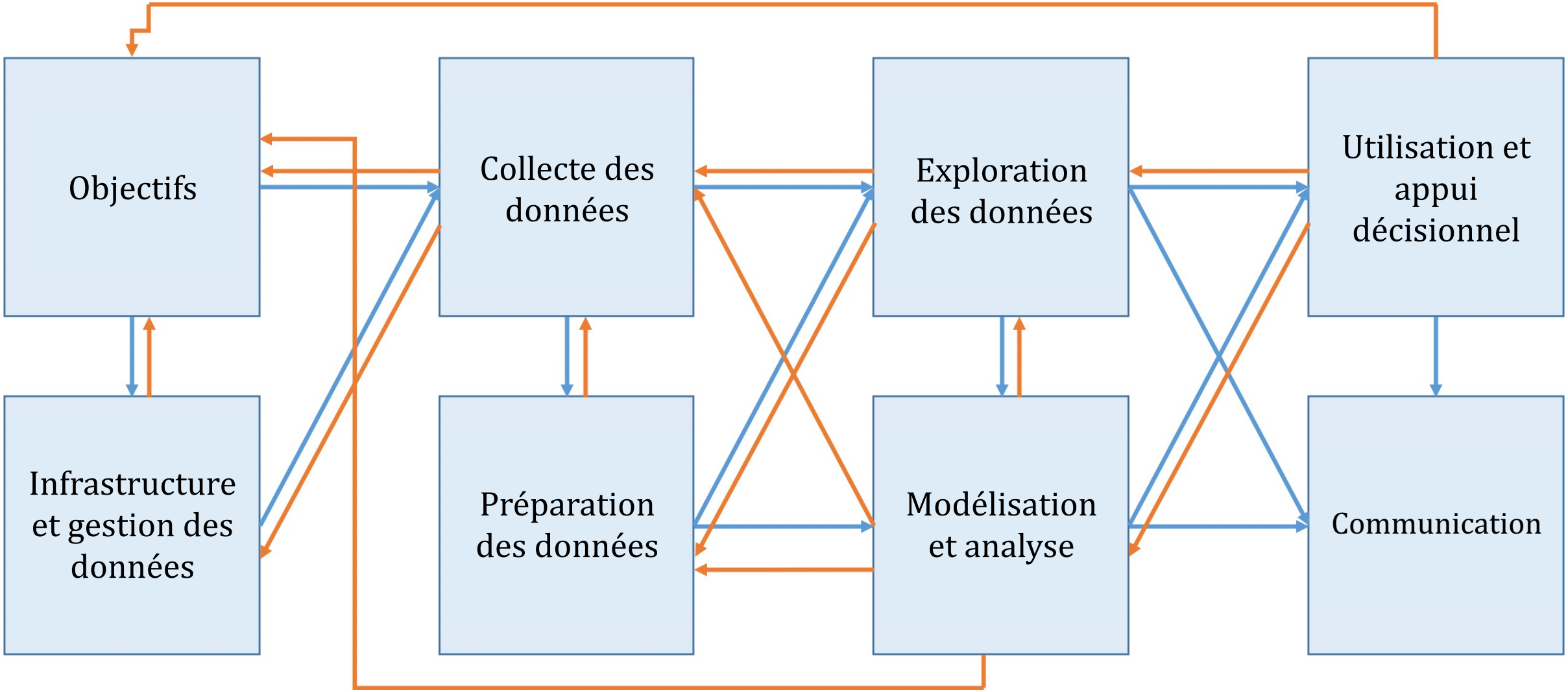
Comme c’est le cas pour la **méthode scientifique**, il existe un guide "étape par étape" pour l'analyse des données

- 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.

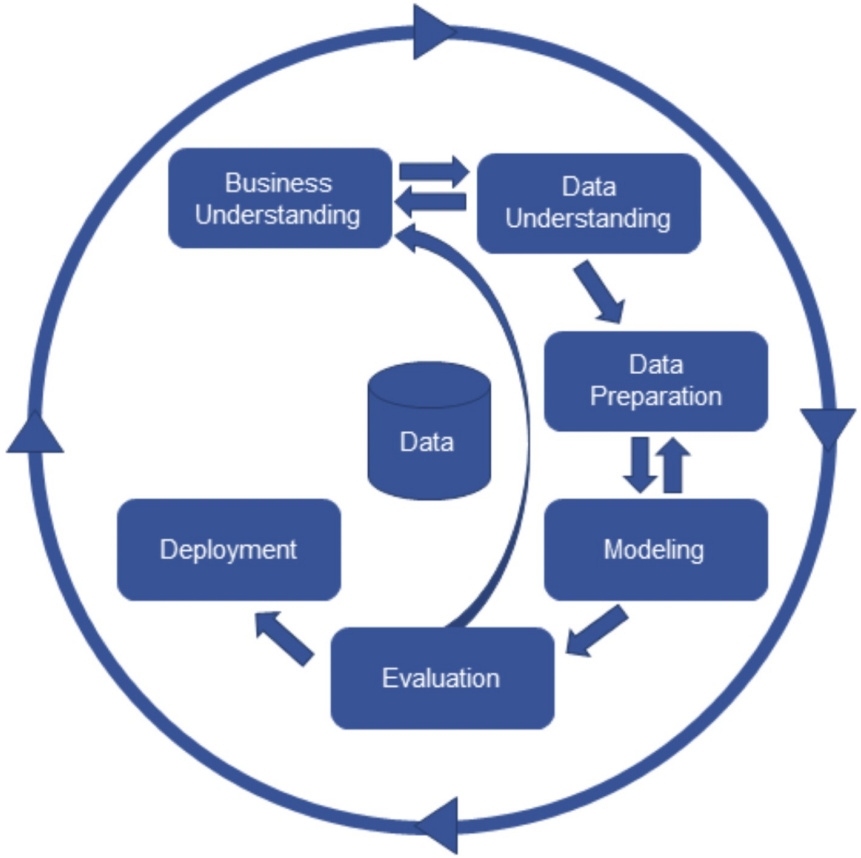
En pratique, le processus est souvent **désordonné** ; étapes ajoutées et retirées de la séquence, répétitions, reprises, etc.

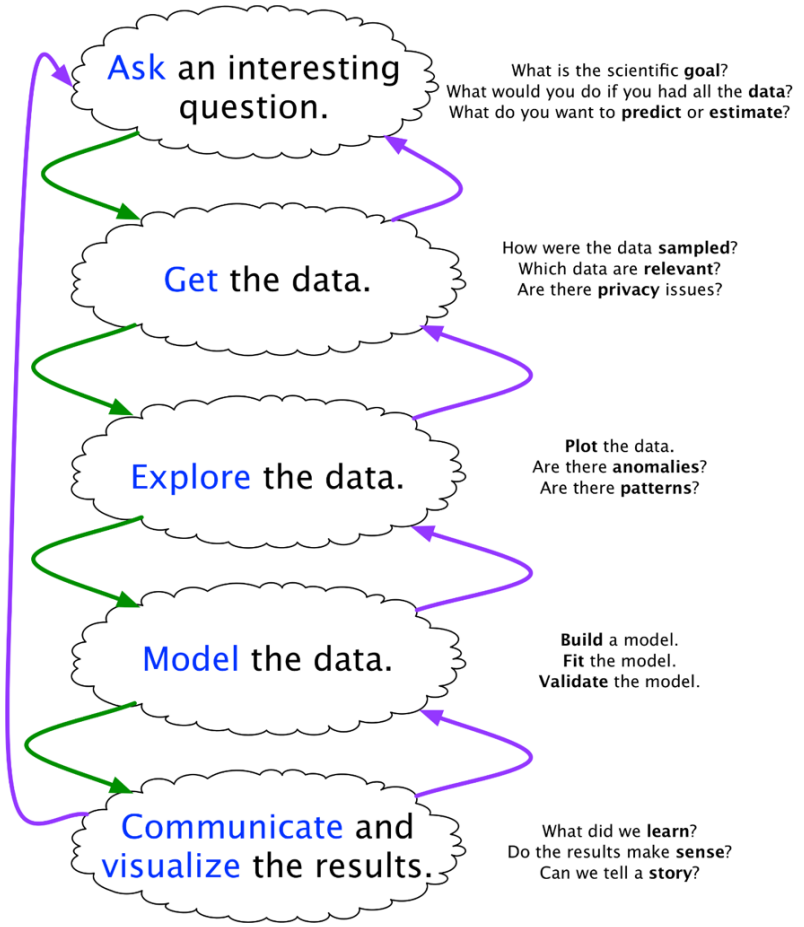
Cela a tendance à fonctionner…
quand **c'est mené correctement**.


Objectifs
Collecte des données
Exploration des données
Utilisation et appui décisionnel
Infrastructure et gestion des données
Préparation des données
Modélisation et analyse
Communication

[Cadre CRISP-DM (à gauche) ; flux de travail de science des données de Blitzstein et Pfister (à droite)].

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 équipe de SD doivent connaître :

- le nettoyage les 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.

[http://dailycatdrawings.tumblr.com ; #309]

Le monde réel

Théorie

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

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.

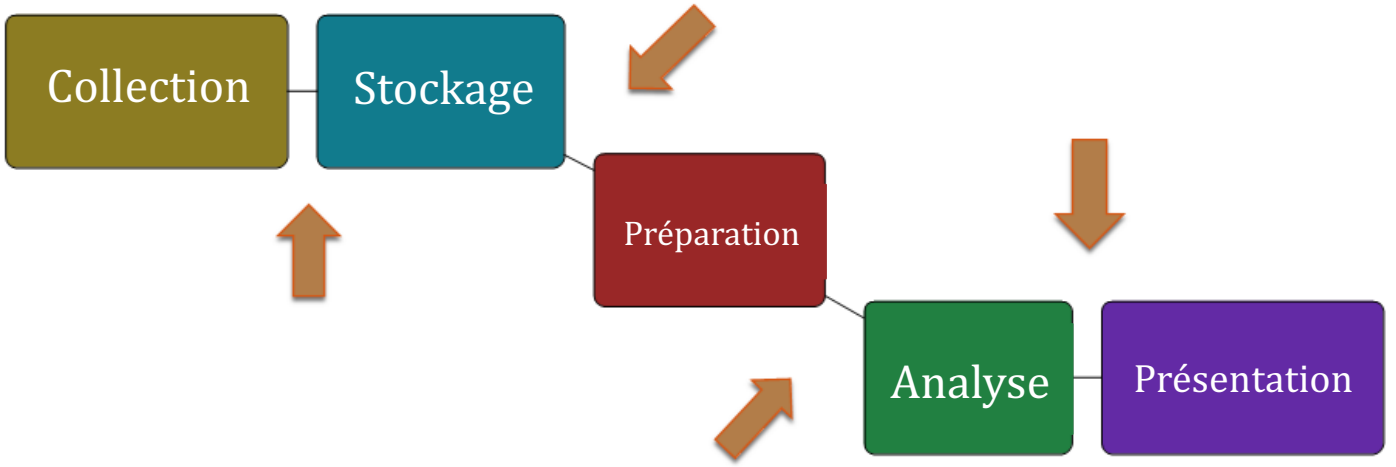
Les pipelines de données se composent généralement de 9 éléments (5 **étapes** et 4 **transitions**) :

- 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 **manuelle-ment** avant que l'implementation ne soit terminée.



Lectures conseillées

Le flux de travail analytique

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

14. Data Science Basics

14.4 Analytics Workflows

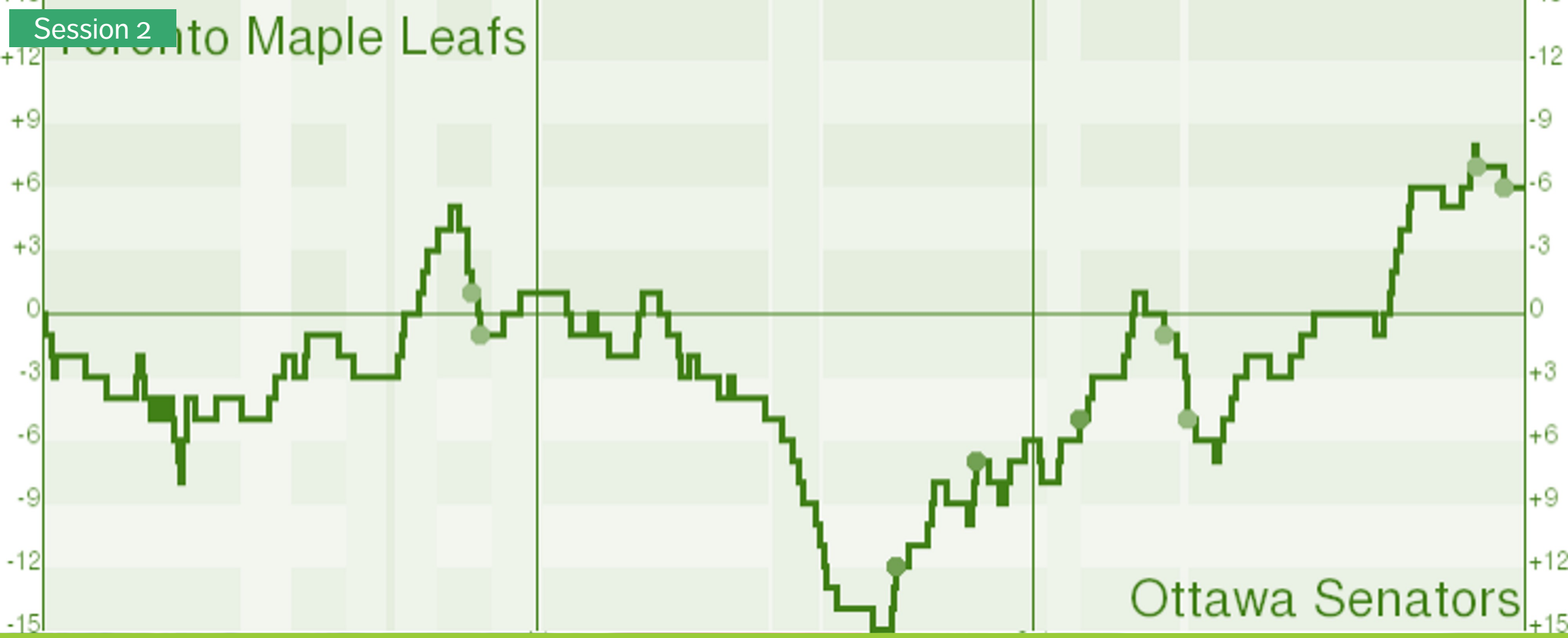
- The “Analytical” Method
- Data Collection, Storage, Processing, and Modeling
- Model Assessment and Life After Analysis
- Automated Data Pipelines

Exercices

Le flux de travail analytique

1. Installez R / RStudio (Posit), et les librairies de la liste fournie par l'instructeur.

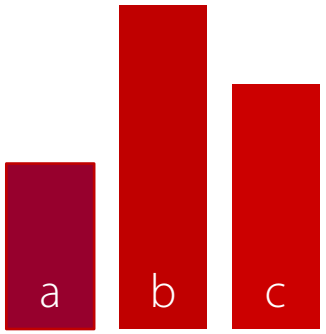
2. Testez l'installation à l'aide des exemples du Programming Primer (sections 2 - 4) pour vous assurer que le logiciel fonctionne comme prévu.



6. Les données et les renseignements

Modes d'analyse

Descriptive



Montrer **ce qui** s'est passé

Diagnostique

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

Prédictive

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

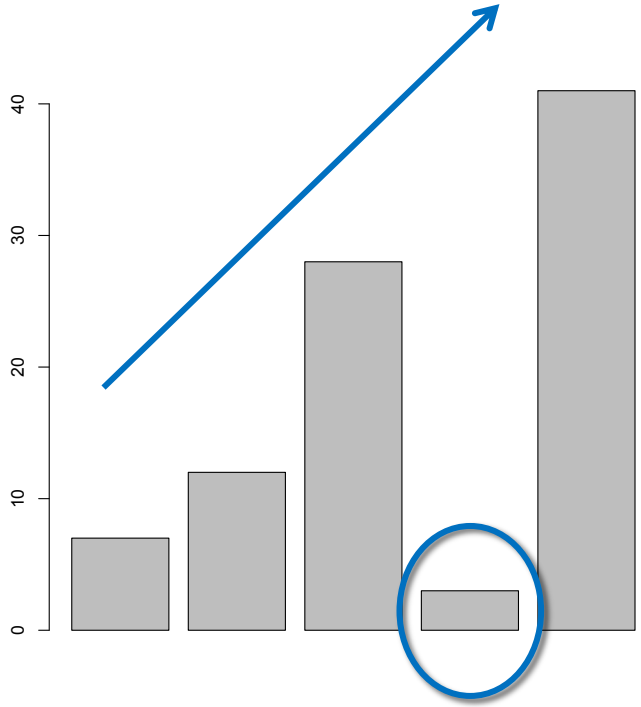
L'analyse par la visualisation

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)

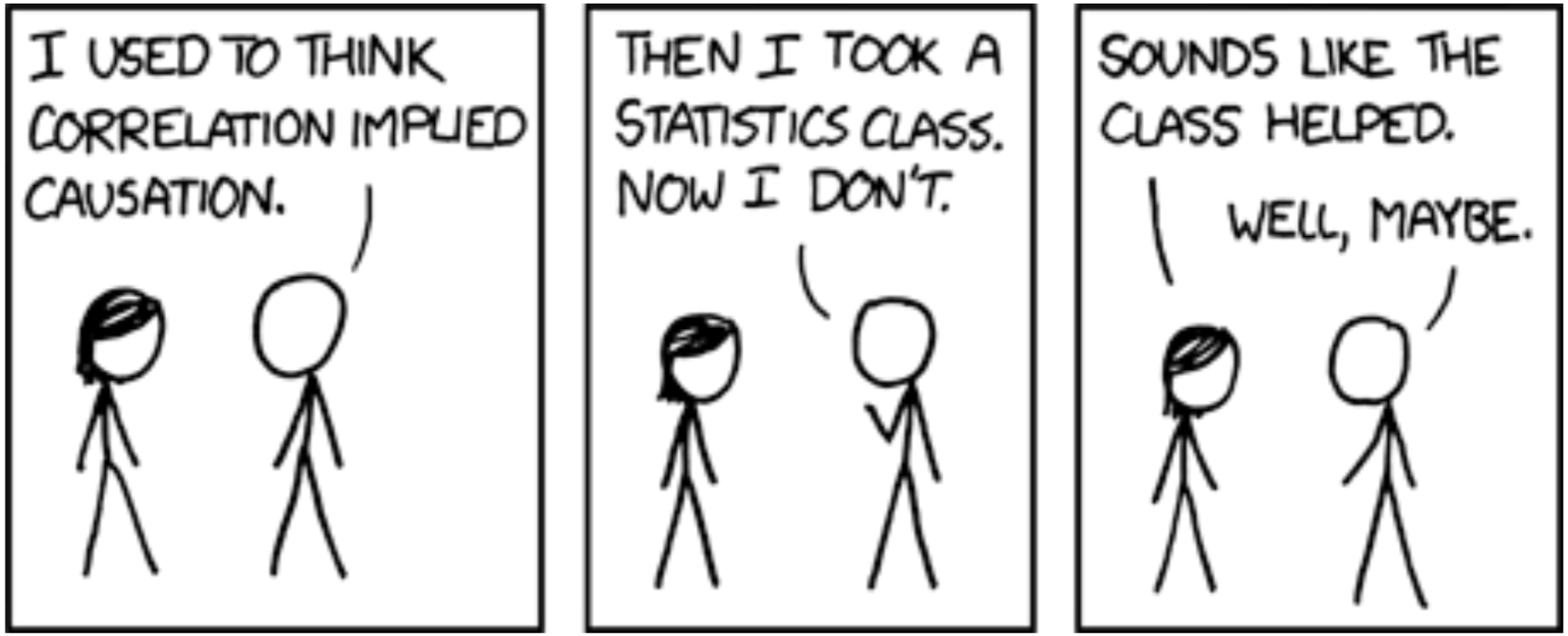
Les **mesures dispersion** (ou d'étalement) comprennent :

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

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

[https://xkcd.com/552]

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 :

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

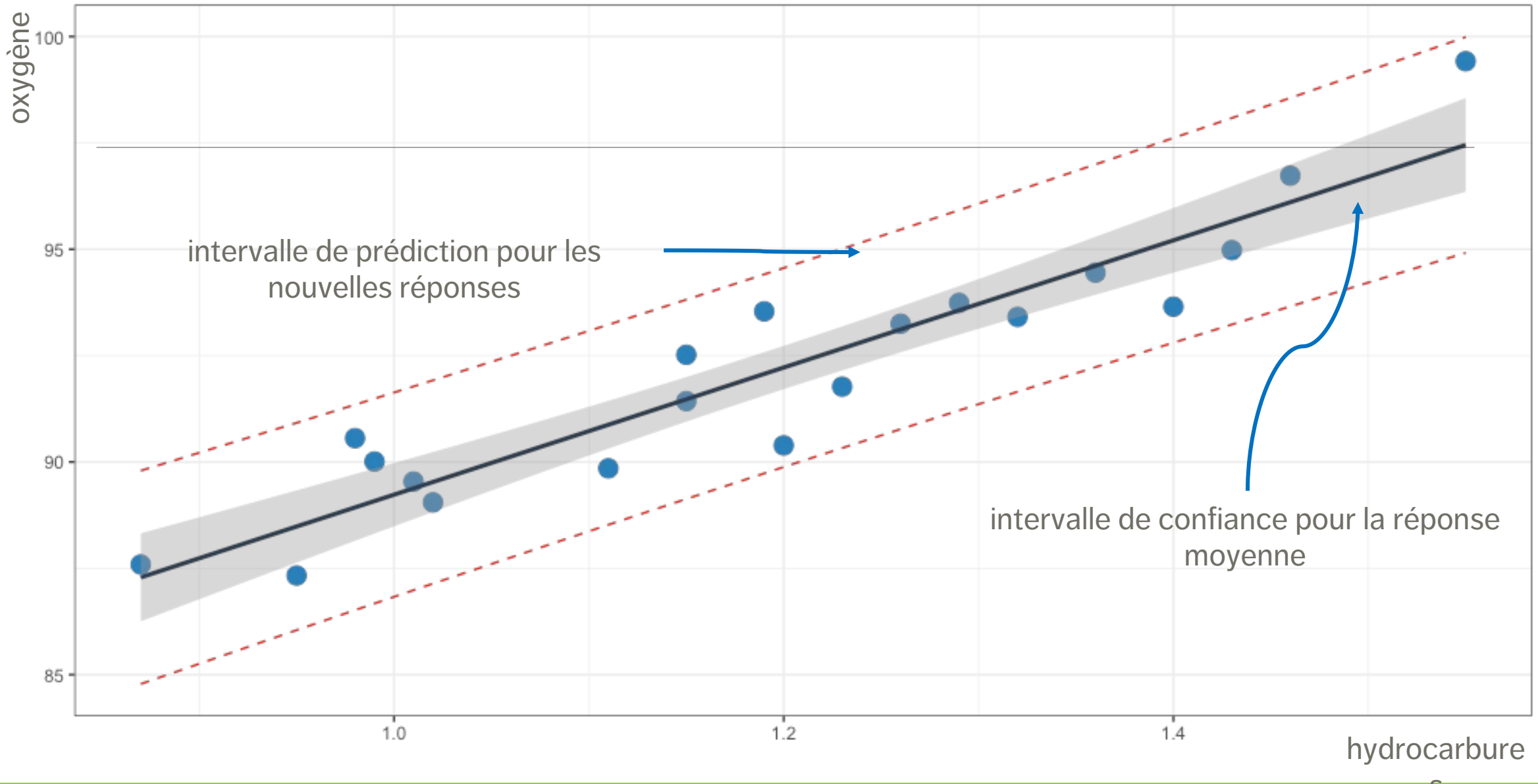
où $\boldsymbol{\beta} \in \mathbb{R}^p$ est déterminé sur la base de l'**ensemble d'apprentissage $\mathbf{X}$**, et

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

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.

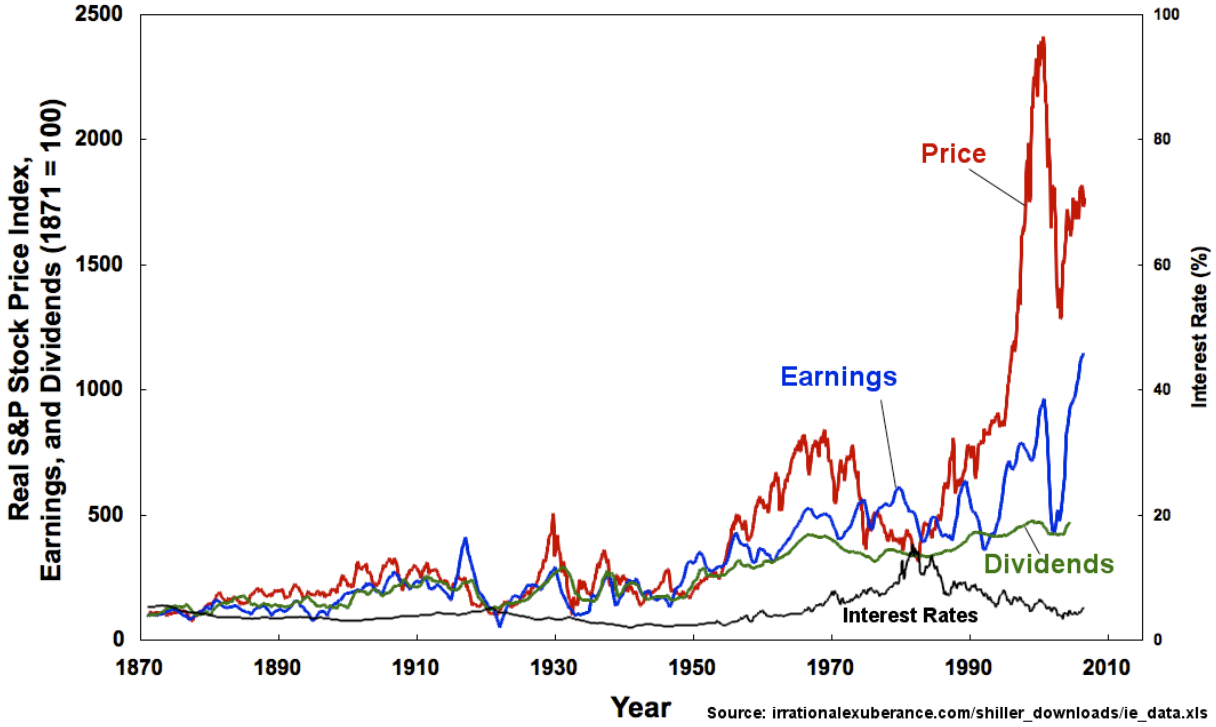
Analyse des séries temporelles

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 ?



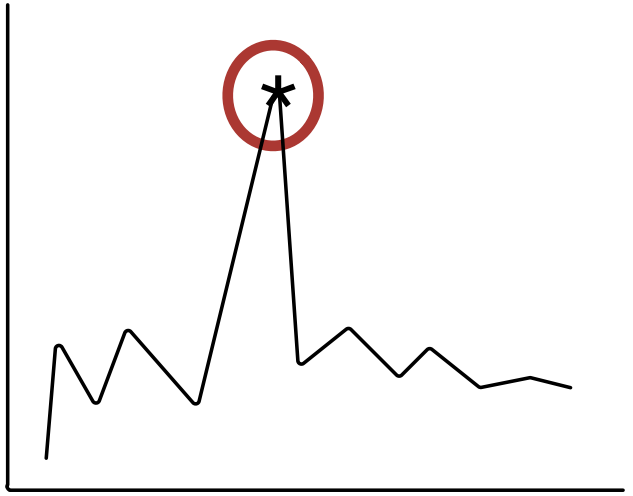
Détection d'anomalies

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 conseillées

Les données et les renseignements

Data Understanding, Data Analysis, Data Science

Volume 2: Fundamentals of Data Insight

14. Data Science Basics
    - 14.5 Getting Insight From Data
        - Asking the Right Questions
        - Basic Data Analysis Techniques
        - Common Statistical Procedures in R
        - Quantitative Methods

Volume 1: Prelude to Data Understanding

6. Probability and Applications
7. Introductory Statistical Analysis
8. Classical Regression Analysis
9. Times Series and Forecasting
10. Survey Sampling Methods
11. The Design of Experiments

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.