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

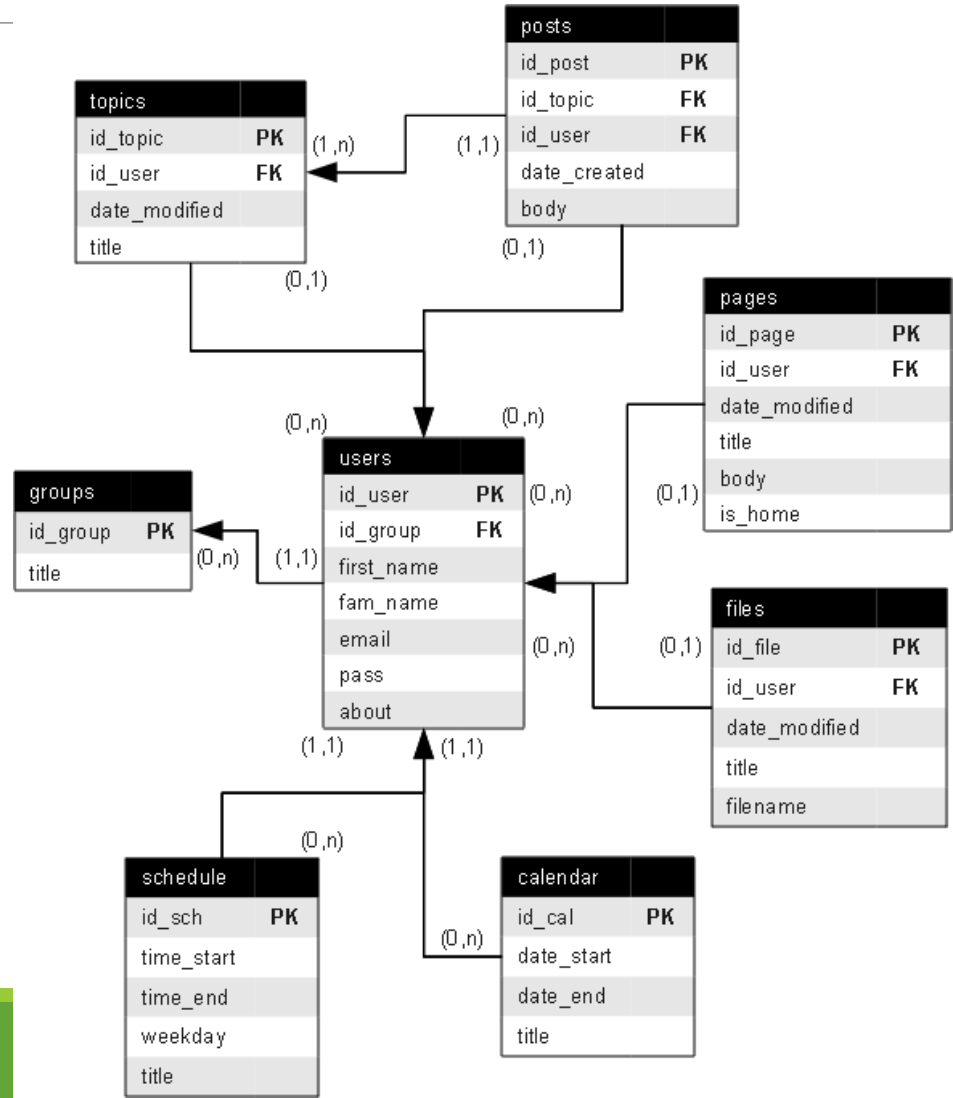
L'ensemble de données de ces objets physiques pourrait commencer par :

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

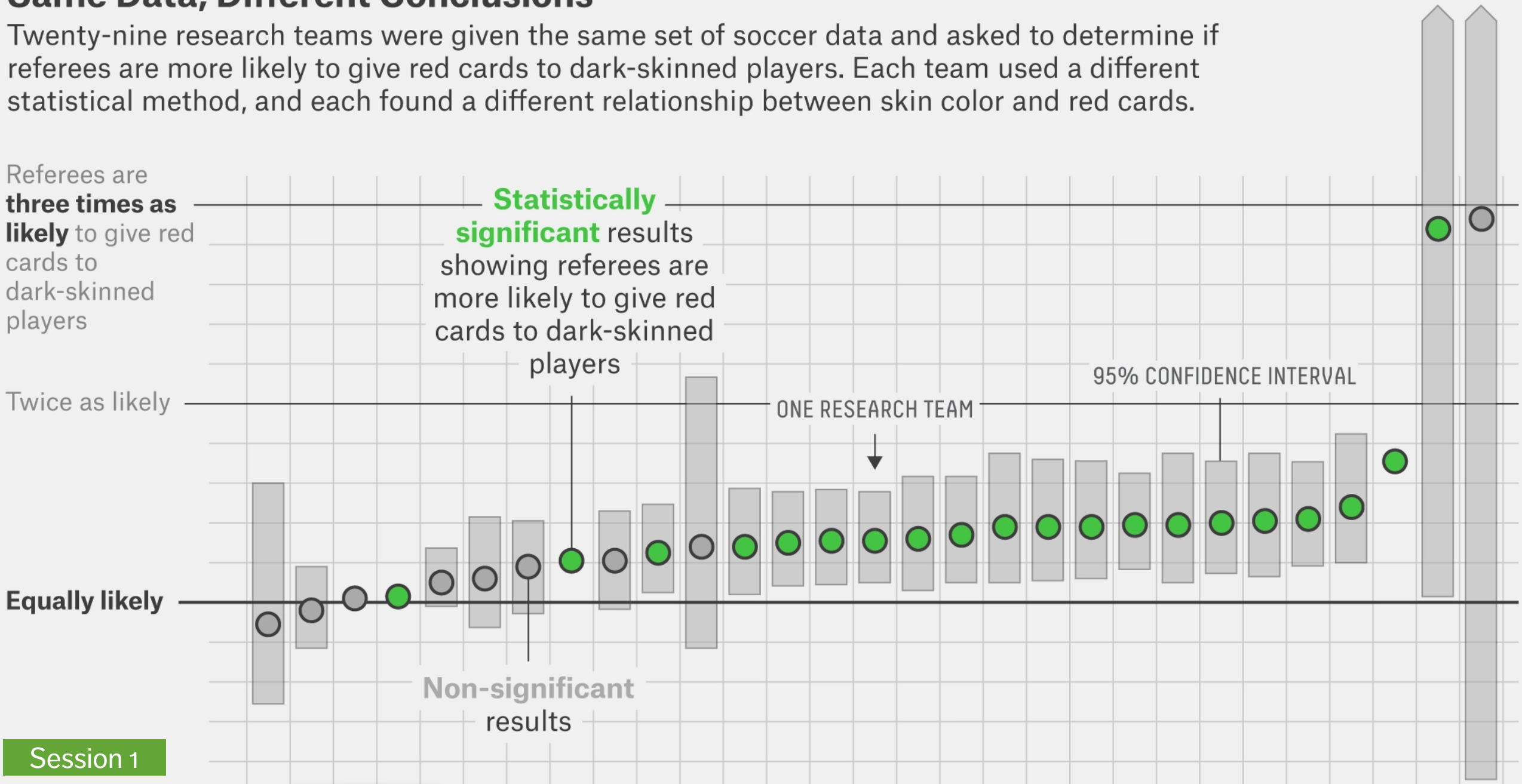
- "AlphaGo vainc le meilleur joueur de Go du monde, marquant la supériorité de l'IA sur l'esprit humain"
- "Une novella écrite par l'IA a presque gagné un prix littéraire"
- "Elon Musk : l'intelligence artificielle peut déclencher une troisième guerre mondiale"
- "L'engouement pour l'I.A. a atteint son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