

Data in practice

Martin Vielvoye 2021

What is data?

Excel don't hurt me.
No more.

Qu'est ce qui est considéré comme de la data?

- La data est une ou plusieurs informations décrivant un objet ou un événement.
- Chaque objets et évènements étant unique, leurs data l'est aussi.
- Chaque informations est réparties en caractéristiques.
- Quelques exemples de caractéristiques :

Identifiant	Mesure
Nom	Pièce
Date	Liens avec d'autres données
Intervale de date	Races
Ratio	...

Les types de data

Qualitative / Catégorique	Relative aux qualités. Souvent des valeurs que l'on ne peut pas mesurer et qui ont des noms. <i>Ex : Couleurs, Secteurs, Rang, Type de Permis, ...</i>	
Quantitative / Numérique	Relative aux quantités. Une valeur sur laquelle on peut placer des numéros.	
	Discrete	Valeurs qui ne peuvent prendre que des quantités complète et fixe. <i>Ex : Année, les chiffres sur un dés, nombre d'étudiant dans une classe, ...</i>
	Continue	Valeurs qui peuvent prendre que des quantités incomplète. <i>Ex : Taille, distance, intervalle de temps, ...</i>

Modèles de donnée

- Le modèle de données est de la méta-information décrivant la façon dont sont représentées les données.
- Il renseigne sur la structure, les champs présent et une description de ces champs.
- Il peut aussi décrire les flux, comment la donnée est relié entre elle ou avec d'autres modèles de données.

Modèles de donnée

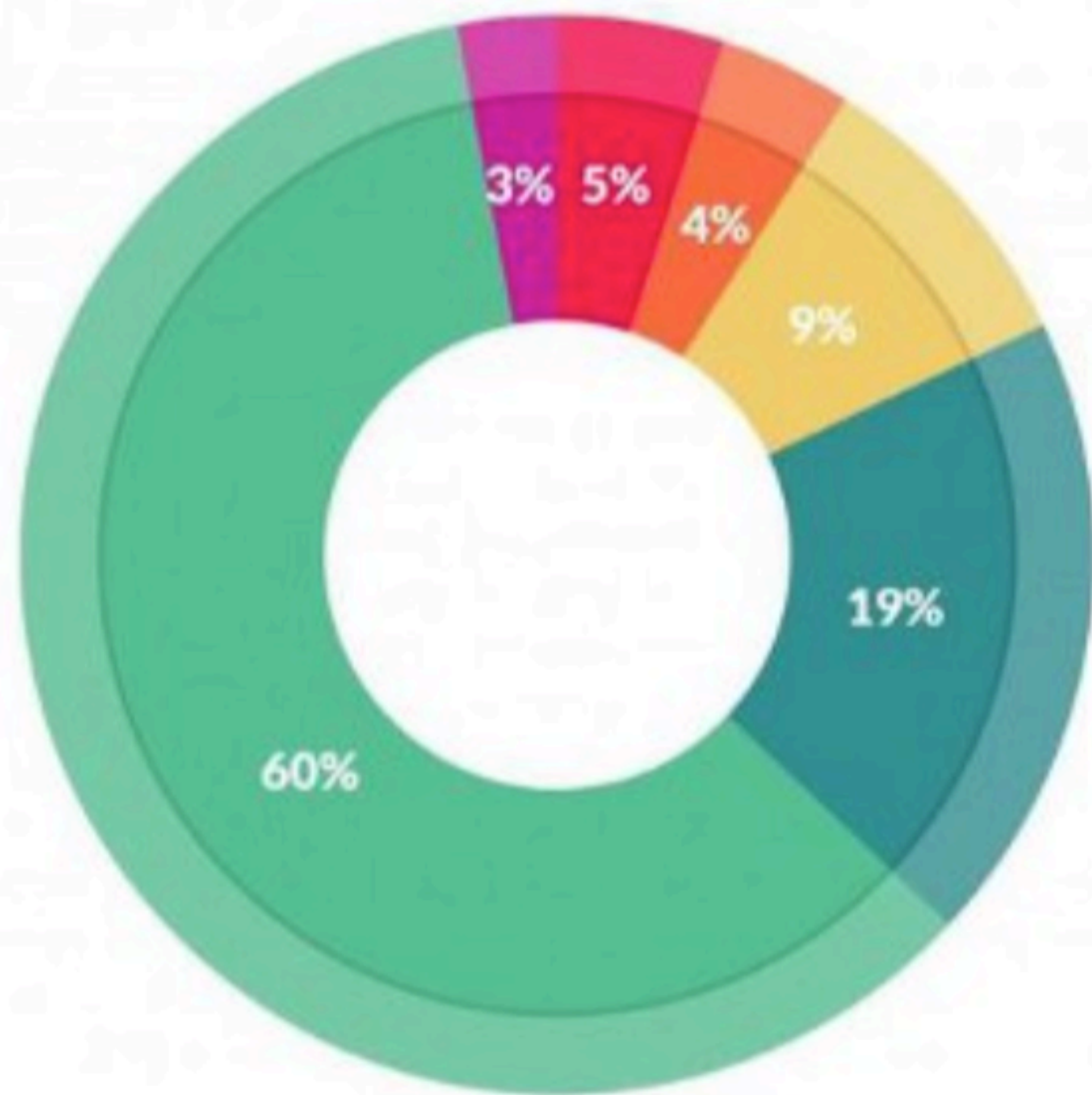
Un dataframe de donnée

ID	biophysical_model	Spikes AUC	Spikes TP @ 0.25
5948c70f-3564-4677-85ff-b3a62e9ab915	NMDA	0.9883157746050288	0.2718369925762123
56151782-6680-41cc-ac72-2d8ede2b4d2b	AMPA	0.9909092158825294	0.23677144211025114
6fb7d3bf-2e43-4c6f-bc76-94d723e02c85	AMPA	0.9913347571054445	0.27531195703680306
6949c9f3-4b03-4883-a5c4-3e9c65aa864d	AMPA	0.9911544832235036	0.26741431053546044

Le modèle de donnée/méta-info

Description	Ce tableau contient des évaluations de performances sur de la donnée test pour différent type d'architectures de réseaux neuronal et 3 types de modèles biophysique.
ID	UUID pour identifier de façon unique la valeur.
biophysical_model	Modèle biophysique de synapse.
Spikes AUC	Area under the Curve of spikes.
Spikes TP @ 0.25	Time-Dependant spike a 0.25.





Répartition du temps de la vie d'un data scientist

●	Construire des datasets d'entrainement	3 %
●	Nettoyage et organisation de la data	60 %
●	Collection de datasets	19 %
●	Récolte de schémas en data-minant	9 %
●	Optimisation d'algorithme	4 %
●	Autre	5 %

Data Life Cycle

I. Data Collection

Collection brute

- Une facilitée de collection et de récupération qui continue de grandir.
- De plus en plus de capteurs et de types de capteurs :
 - Santé.
 - Caméra.
 - Puce électronique (NFC, RFID, ...).
 - Satellite.
- Enregistrement de comportements :
 - Activité sur internet (heat-map).
 - Présence (ou absence) sur les réseaux.
 - Analyse de préférences et de goûts.

Polir la donnée

- Une donnée *non-exploitable* est une donnée qui pourrait tout autant ne pas exister.
- **Illustration** : 100 M de TB de vidéos de sécurité

Version 1

- Aucune videos n'a de titres ou de métadonnées.
- On ne peut faire aucune recherche et on ne se saurait pas par où commencer.
- Trop de videos pour les faire une par une.
- On dit que la donnée est *inexploitable*.

Polir la donnée

- Une donnée *non-exploitable* est une donnée qui n'existe pas.
- **Illustration** : 100 M de TB de videos de sécurité

Version 2

- Les vidéos ont des titres et une date.
- On peut trier par date et effectuer une recherche.
- Trop de vidéos pour les faire une par une.
- Cela prend du temps mais une recherche est possible.

Polir la donnée

- Une donnée *non-exploitable* est une donnée qui n'existe pas.
- **Illustration** : 100 M de TB de videos de sécurité

Version 3

- Les vidéos ont des titres, une date ainsi qu'une description précise des événements présent sur chaque vidéo.
- On peut trier par date et faire une recherche de texte dans les descriptions.
- Une simple lecture de texte permet de retrouver de l'information rapidement.
- La donnée est exploitable malgré la quantité.

Data Labeling

L'Homme nourrissant l'IA

- **L'étiquetage de la data** (*labeling*) est un process permettant de mettre des labels/étiquettes sur de la donnée.
- Elle est essentiel en IA et surtout en machine learning pour indiquer la *cible*.
- Elle ne demande pas beaucoup de compétences techniques mais elle est :
 - Fortement chronophage
 - Source d'erreurs
 - Souvent vu comme une corvée car itérative
- C'est un process qui peu souvent s'externaliser moyennant un coût.

Data Labeling

L'IA nourrissant l'IA

- L'automatisation de l'étiquetage de la donnée est une discipline pleine de promesse et de plus en plus performante sur le marché.
- Il y a plein de façons d'automatiser l'extraction de donnée d'une source.

Version 4

- Les vidéos ont des titres, une date ainsi que une description précise en texte des événements présent sur chaque vidéos.
- Une IA permet de créer un tableau par vidéo avec les informations suivantes :
 - Des informations sur les personnes (noms, tenus, action engagé, sentiment, ...)
 - Des informations sur les objets et les scènes (météo, présence d'objets, ...)
 - Analyser et prédiction d'un événement
 - Etc

Une stratégie autour de la data est **nécessaire**
pour la rendre pertinente et exploitable.

Data Life Cycle

II. Data Exploration

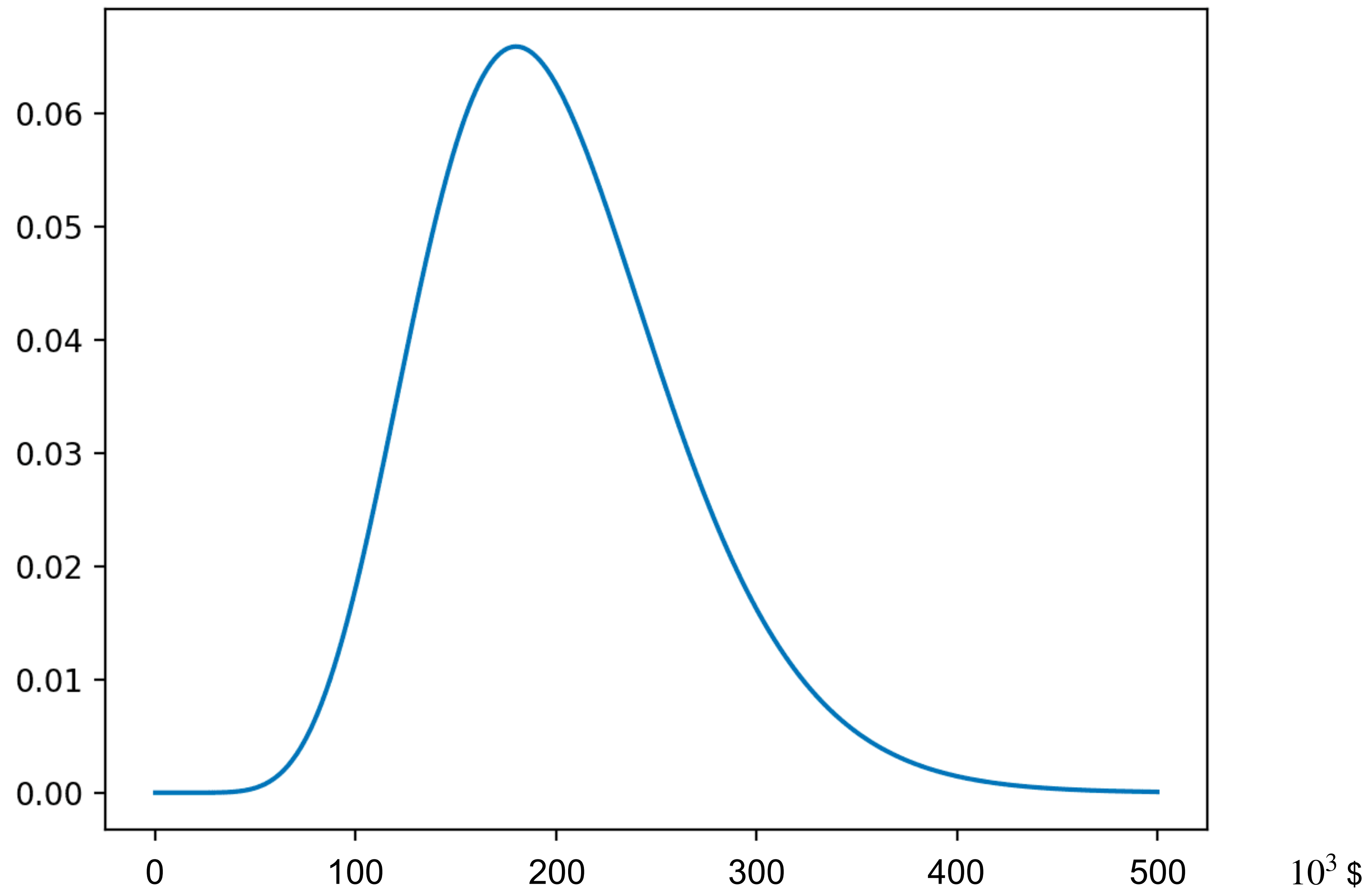
Une première exploration

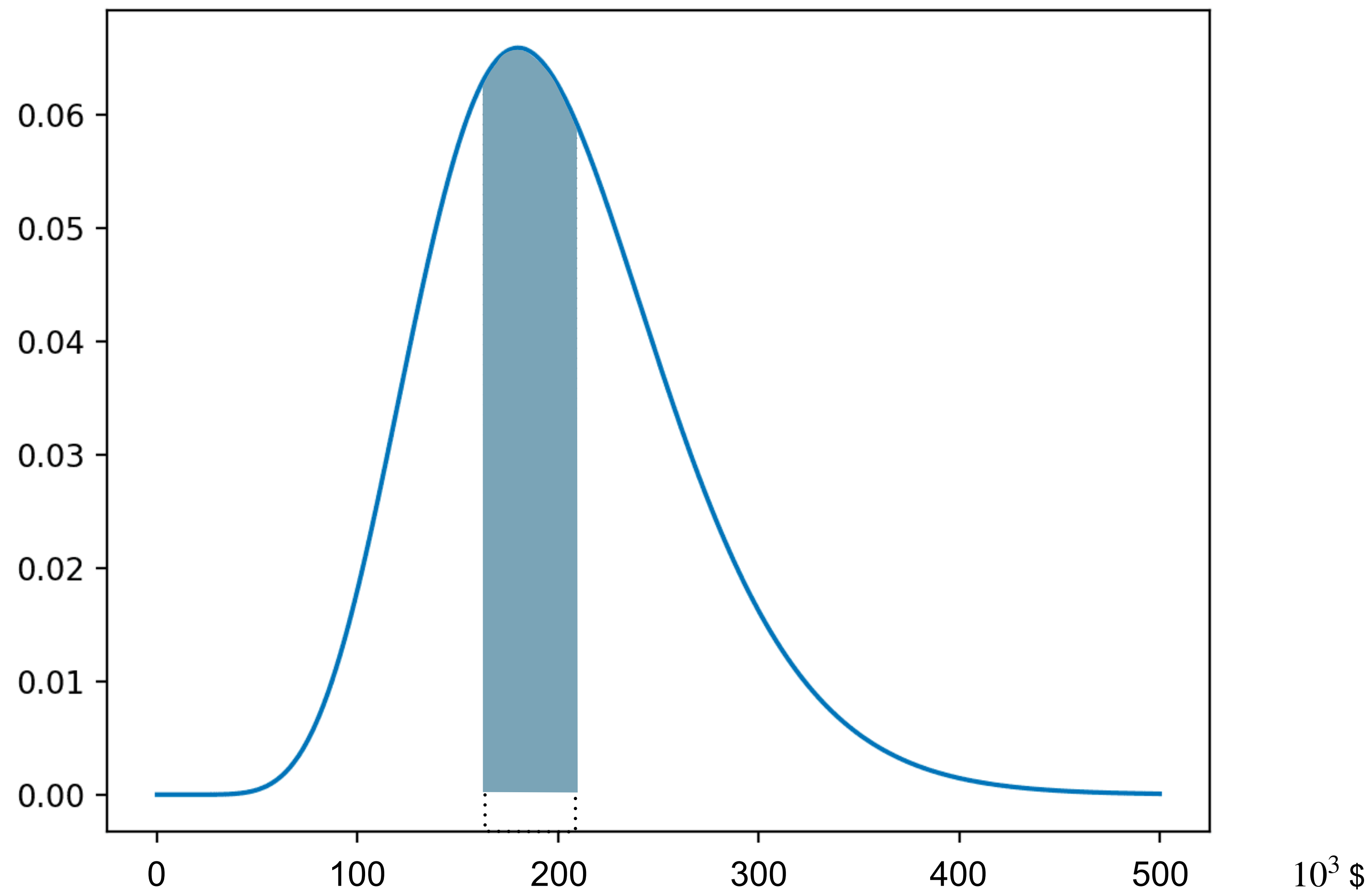
- Une première exploration statistique permet de récupérer un peu de d'instinct sur la donnée exploité.
- Il s'agit de :
 - Comprendre les features qu'on explore et leur contexte
 - Analyser leurs quantités, les valeurs moyennes, leurs répartitions et la distributions des valeurs

Note :

Une formule très pratique avec la librairie *pandas* : `dataframe.describe()`

Il faut explorer avec un nombres importants de métriques et de mesures différentes pour bien comprendre toutes les prismes et facettes d'un jeu de donnée.



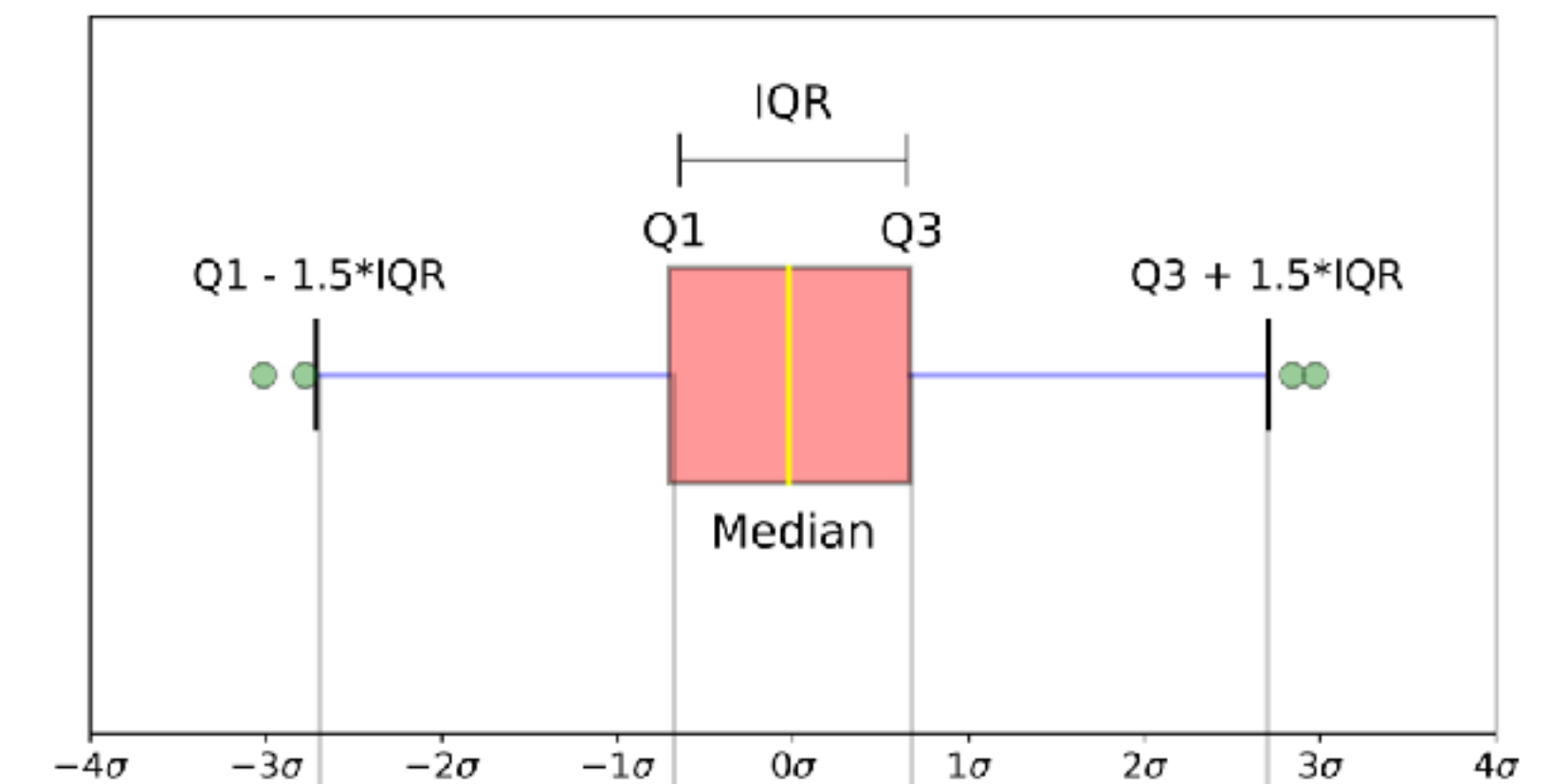


Une première exploration

Boxplot

Comprendre la Boxplot :

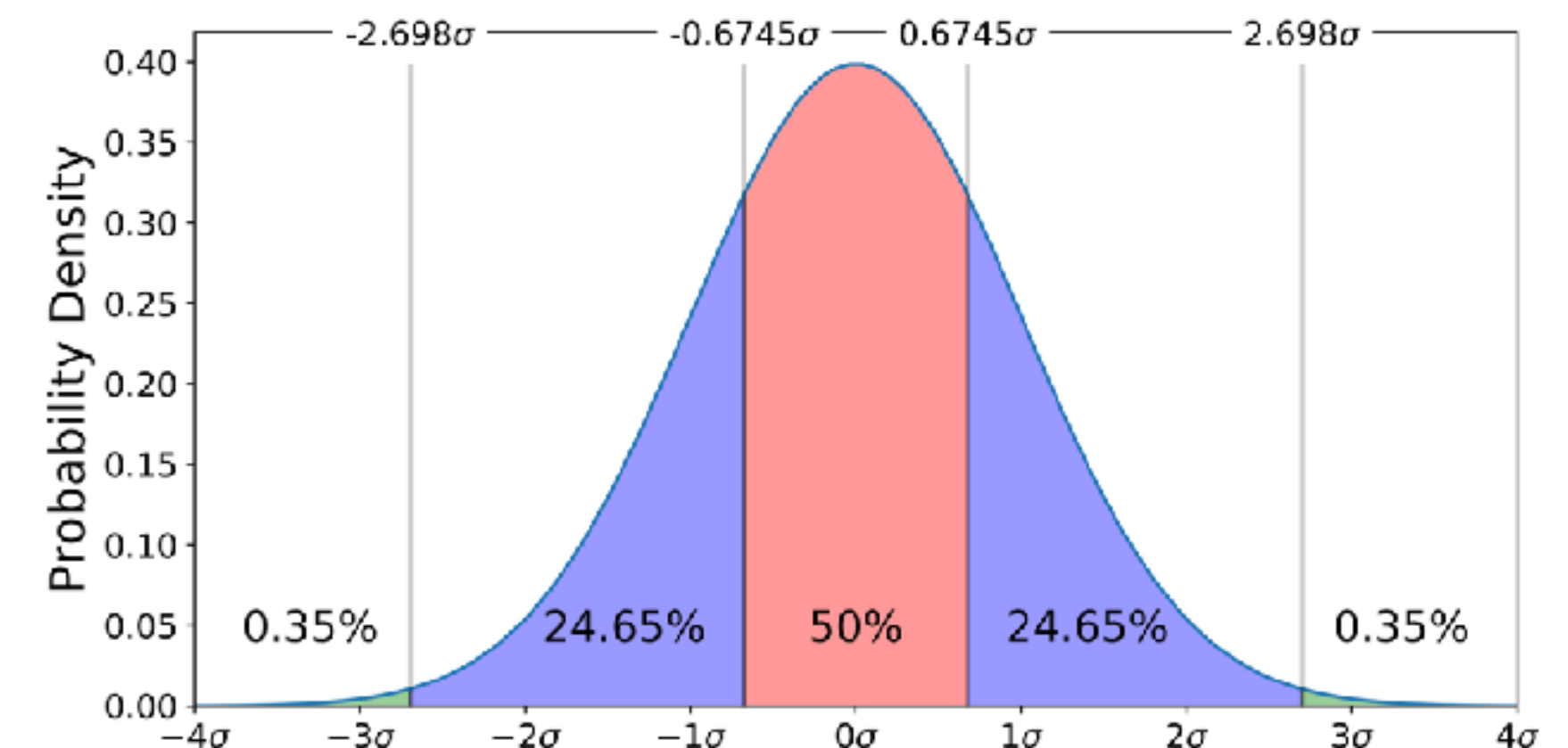
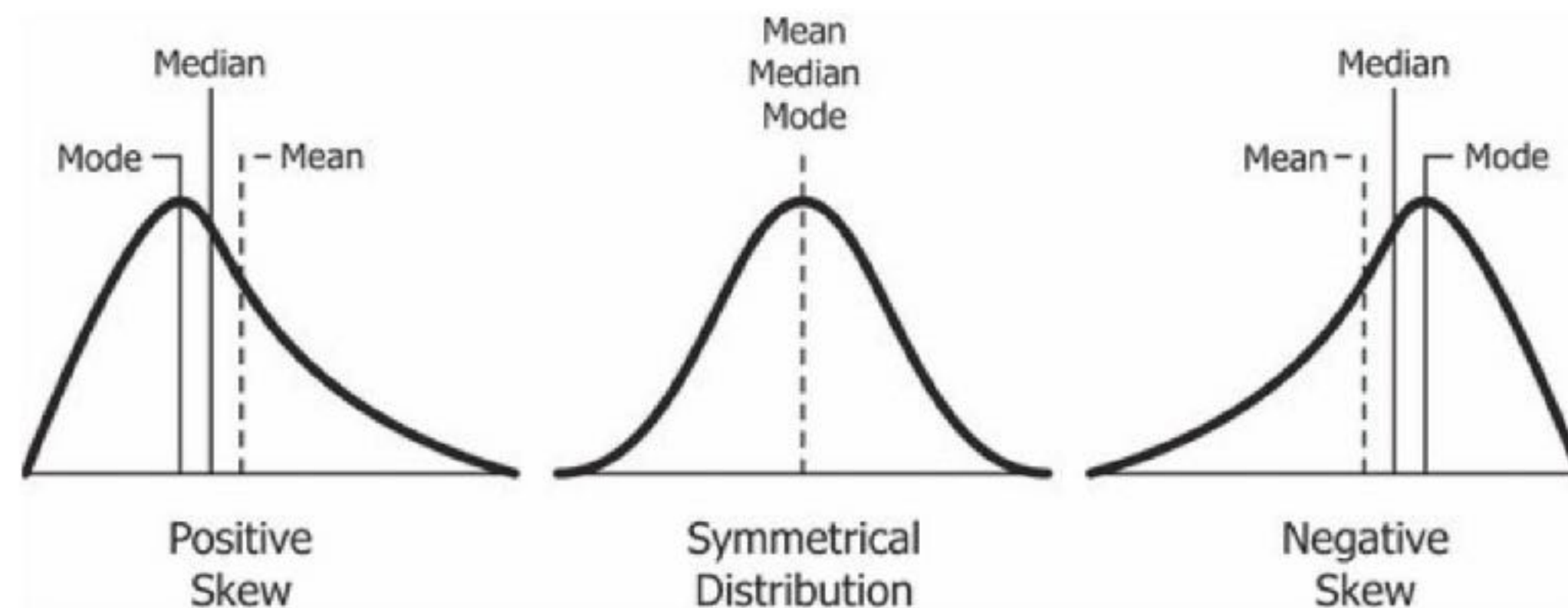
- Q1, Mediane, Q3 \leftrightarrow 25%, 50%, 75%
- 50% des valeurs sont comprises entre Q1 et Q3 (*IQR*).
- Outliers :
 - Toutes données a $\pm Q_1 - 1.5 * (Q_3 - Q_1)$
 - Les outliers sont tellement éloignés de leur distribution que l'on peut les souvent les considérer comme non-représentatifs.

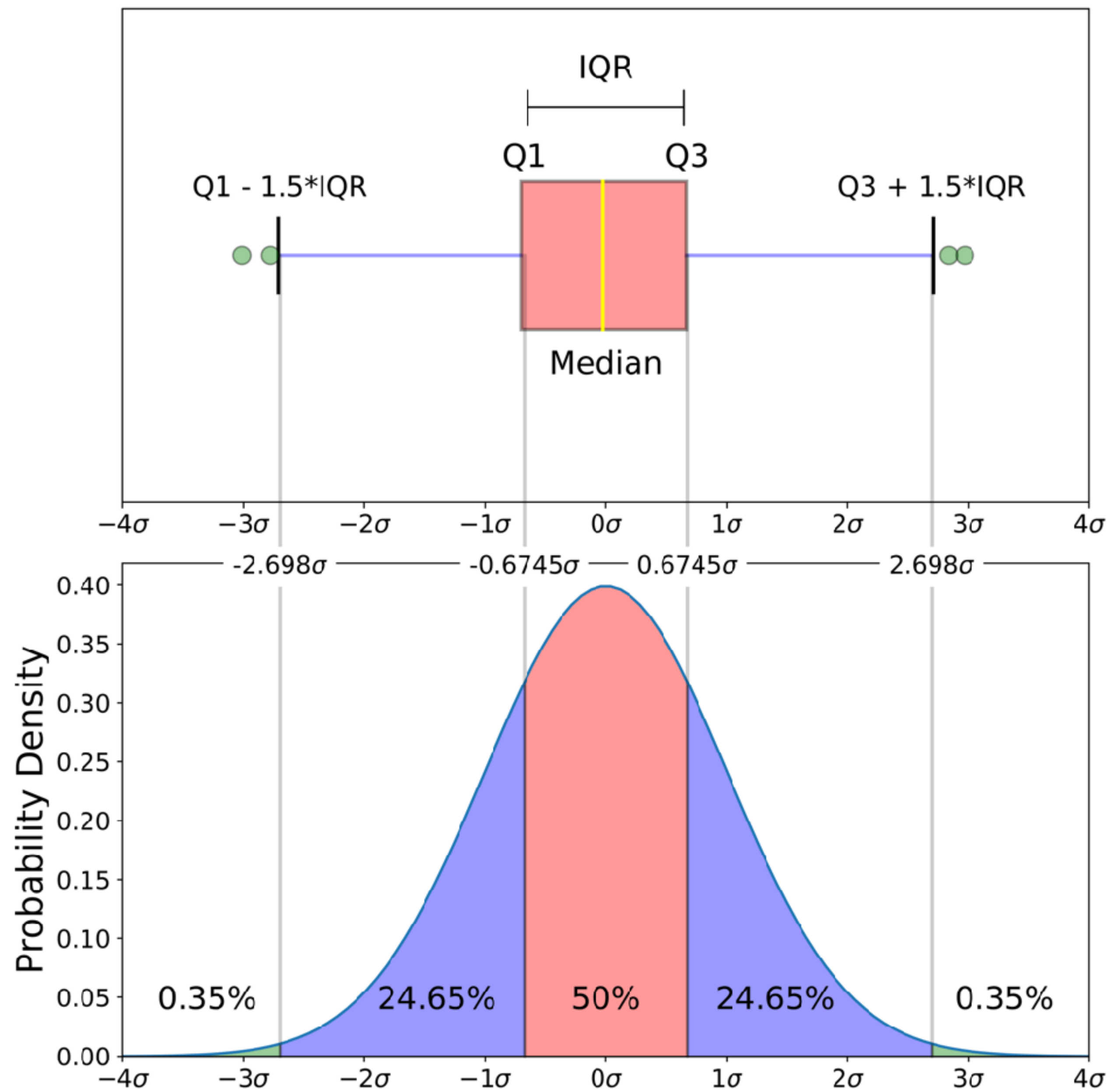


Une première exploration

Distribution & Histogramme

- Deux façons rapides de comprendre les features la répartition des valeurs qualitatifs et quantitatifs d'un jeu de donnée.
- Distribution :
 - Skewness
 - Standard deviation



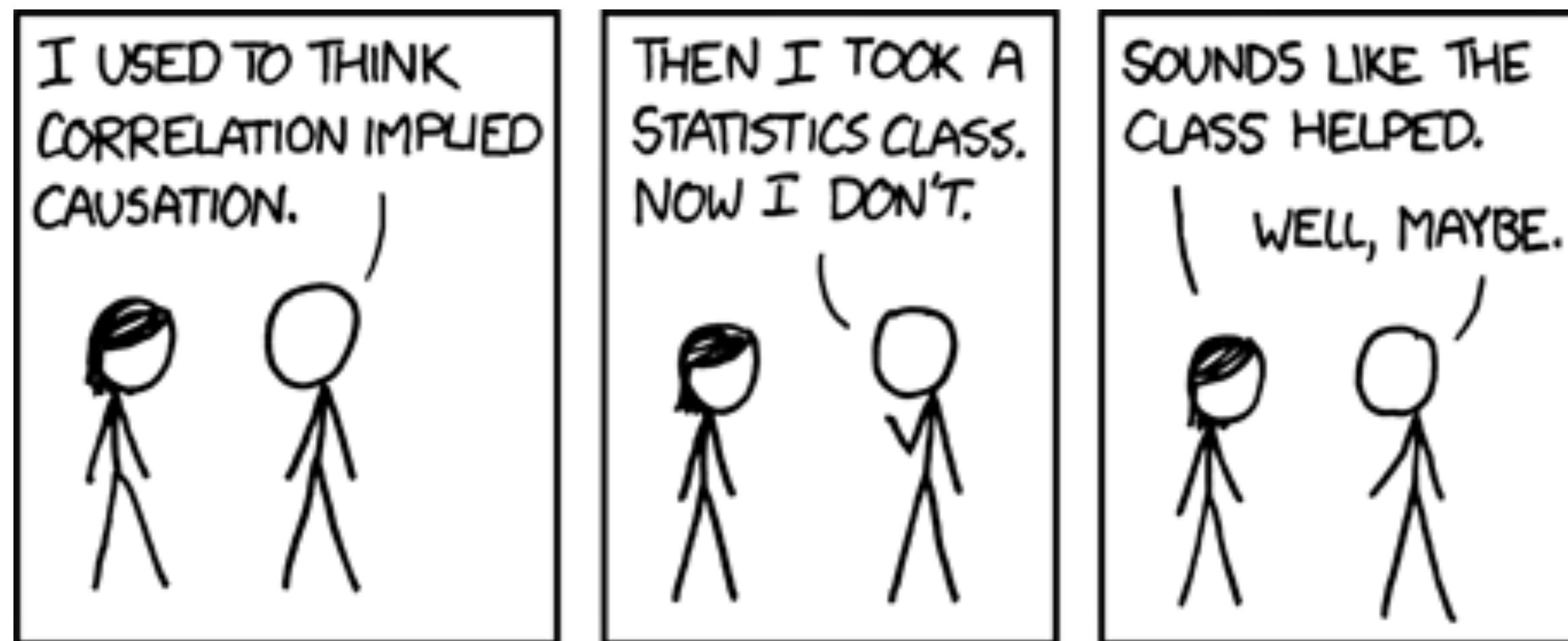


Une première exploration

Corrélation & Causation

- La corrélation est une mesure de statistique symbolisant la dépendance entre deux variables.
- Une dépendance/corrélation permet d'avoir des leviers de prédictions sur les mouvements et comportements entre les données.
- Ex : Quelle est la corrélation entre la taille d'une personne et ...
 - sa taille en inches
 - son IMC
 - sa taille de pied
 - son niveau d'éducation

Correlation ne signifie pas causation



Exemple 1

Corrélation entre la coupe de monde du football et l'augmentation de vente de téléviseurs.

La coupe du monde est bien la cause de la vente de téléviseurs.

(Si la coupe du monde du monde n'a pas lieu, les vente de téléviseurs n'aurait pas augmenté autant)

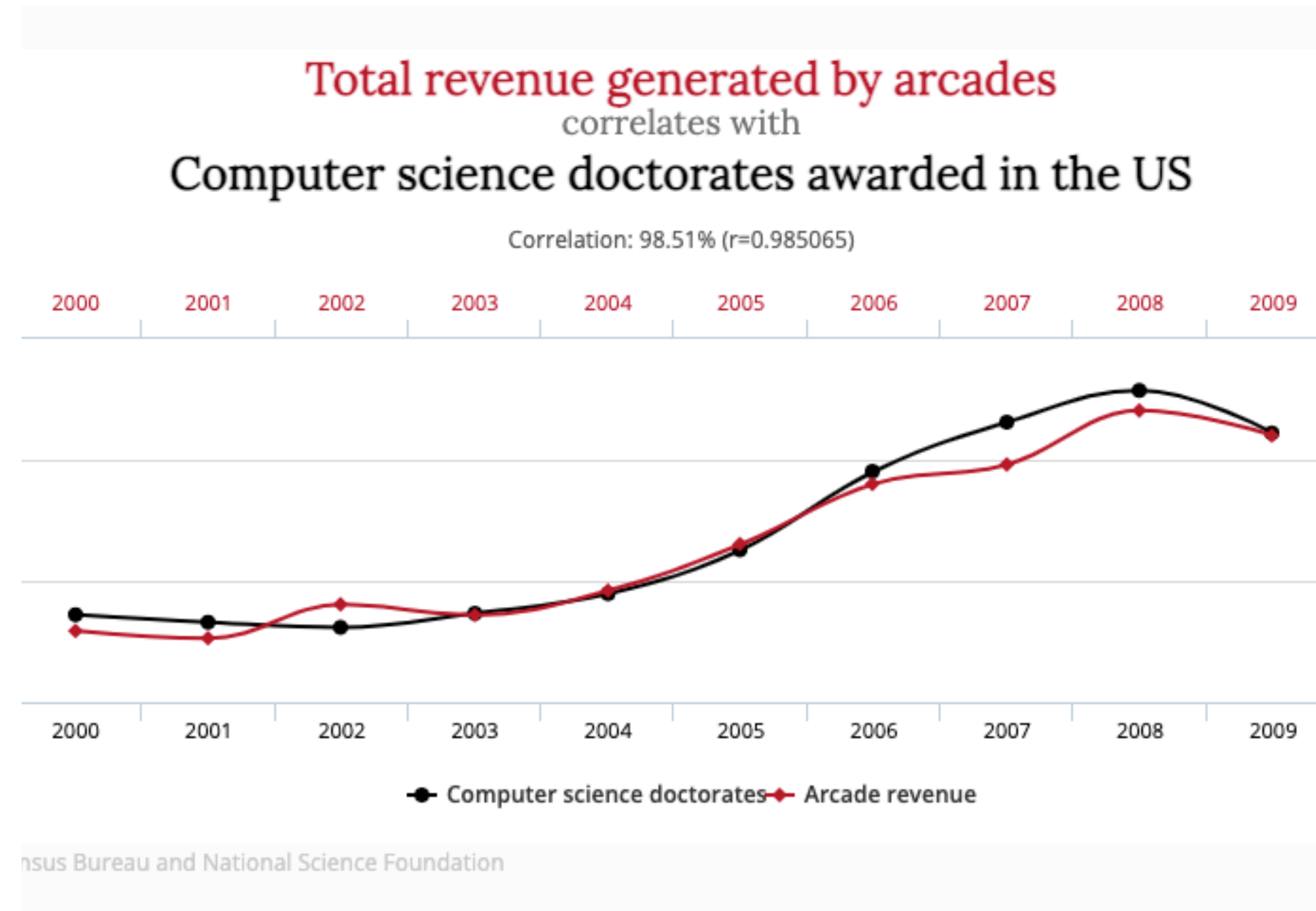
Exemple 2

Corrélation entre le nombre de doctorants en informatique et le revenu des arcades.

La remise de doctorat en informatique est-elle la causation du revenu des arcades?

Ou l'inverse?

Probablement pas.



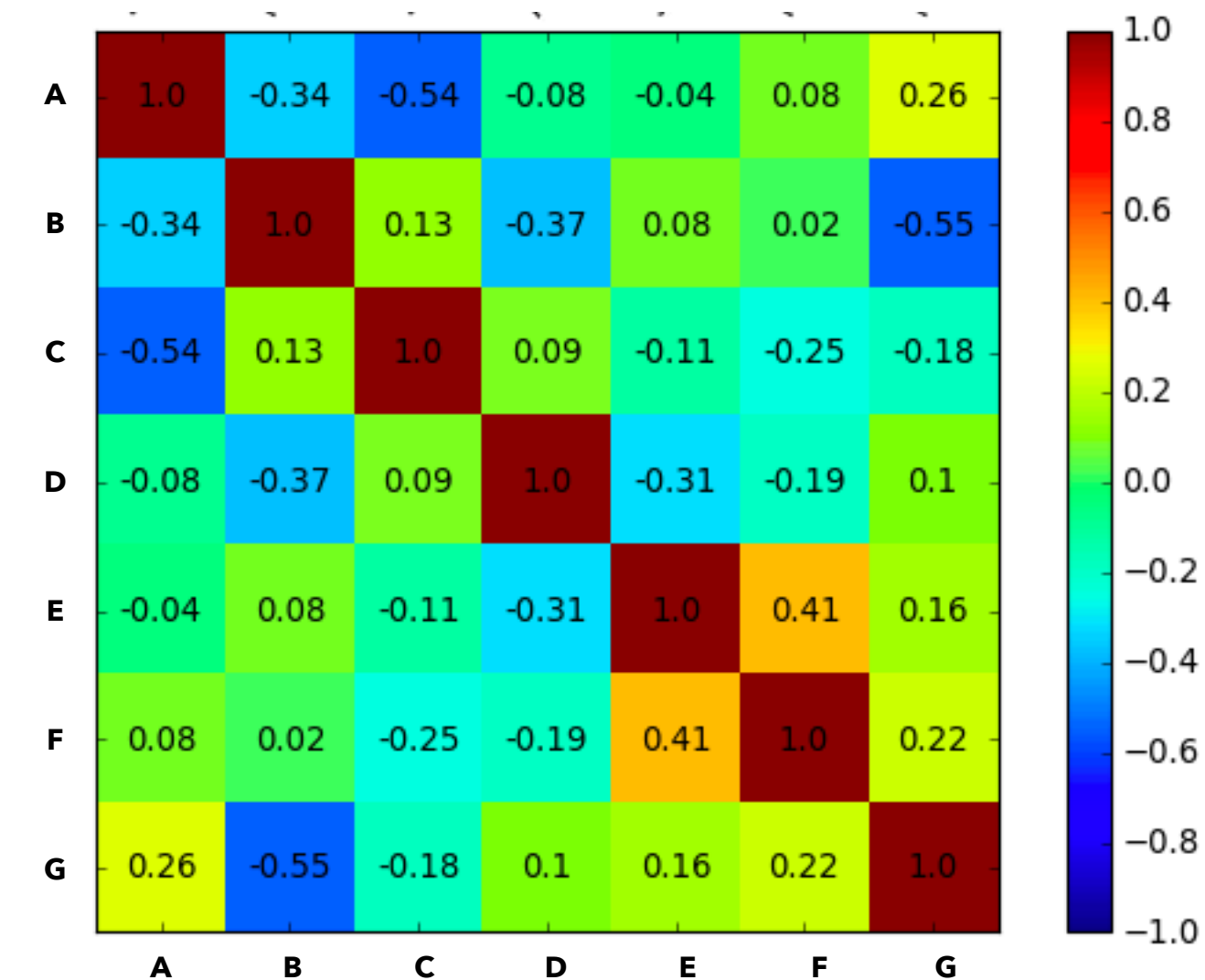
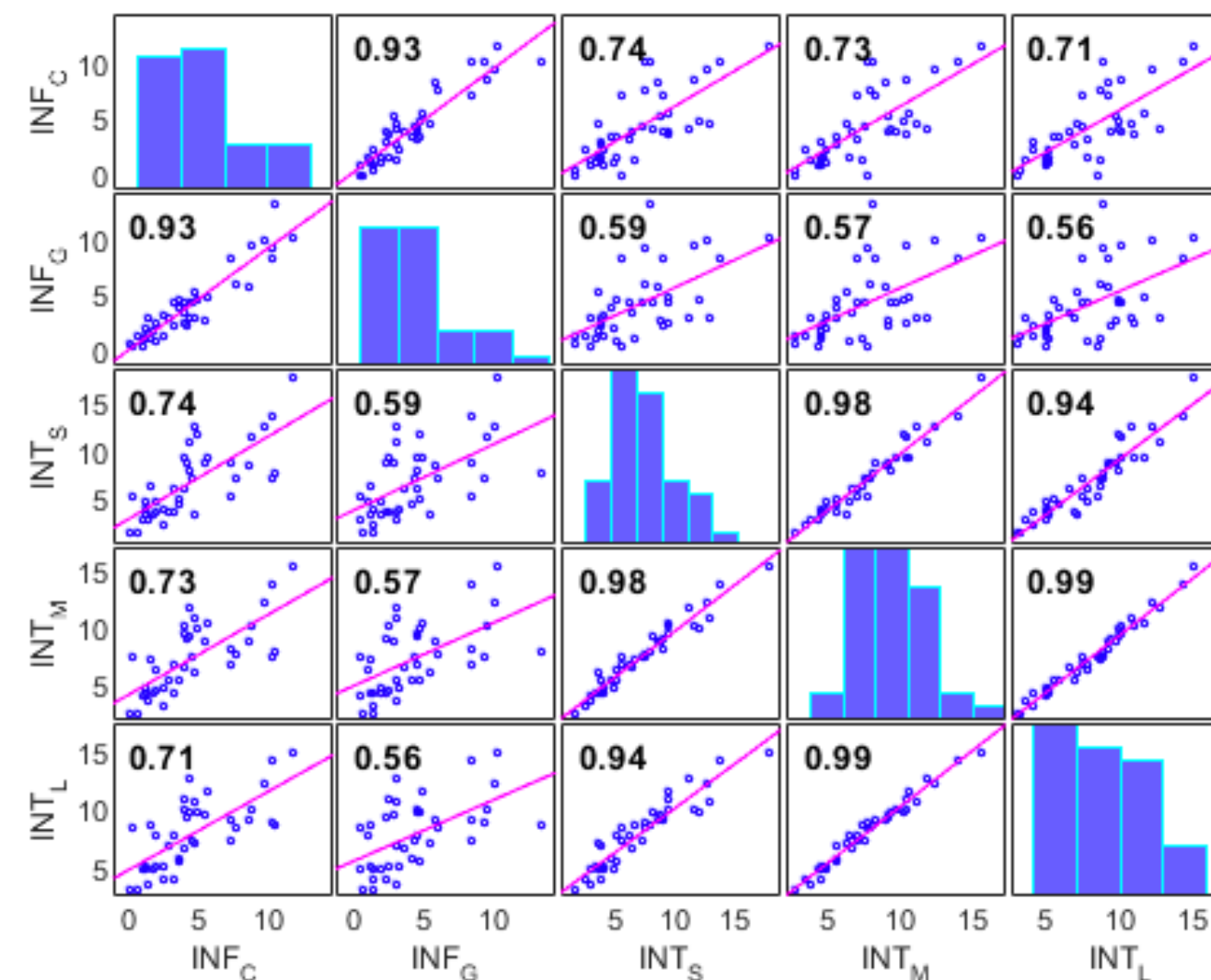
<http://www.tylervigen.com/spurious-correlations>

Une première exploration

Heat-maps & Correlation Matrix

- **Heat-maps** : Comprendre la relation entre les features dans une carte 2-D coloriée.
- **Corrélation Matrix** : Un croisement de toutes les features d'un jeu de données, calculant la corrélation entre chaque pairs de features.

Corrélation
Matrix



Corrélation
Matrix
+
Heat-map

Une première exploration

- Une étude du contexte autour de la donnée est essentiel pour pouvoir comprendre
 - les limites
 - la possible importance des features
 - Les possibles valeurs erronées
- Avoir une bonne compréhension d'un contexte, de la data et de l'objectif finale est essentiel pour savoir
 - Comment et dans quel ordre procéder dans la mise en place
 - Quels modèles et méthodes utiliser pour prédire l'objectif
 - Comment améliorer leurs performances

Data Life Cycle

III. Data Cleaning

Nettoyage nécessaire?

- Il arrive (souvent) que la donnée que l'on reçoit à analyser comporte diverses erreurs.
- Different types d'erreurs :
 - Valeurs manquantes
 - Valeurs erronées
 - Valeurs mal enregistrées
 - Une structure de donnée mal exploitée
- Les erreurs de données peuvent empêcher la compilation de calcul, la manipulation de données ou la stabilisation de convergence d'un modèle.

Valeurs manquante

- Certaines variables sont manquantes dans le jeux de données.

# de chambres	Surface m ²	Cheminée	Année Construction	...	Prix
3	8450	<i>Null</i>	2003	...	208500

- Par quelles valeurs remplacer la valeurs manquante sans impacter notre modèle/prédiction?
 - Par le mode ou la moyenne afin de ne pas impacter les résultats.
 - Si assez de données sont présentes, on peut retirer toute la donnée du dataset.

Valeurs erronées

- Certaines variables sont présentes mais une erreur ou un détail dans le jeux de données empêche l'exploitation de cette donnée.

# de chambres	Surface m^2	Année Construction	Nom	...	Prix
3	8450m	"2003"	@HE20	...	208500
3	9600m	1976	Elysee	...	181500

- Si l'erreur est répétitive, on peut nettoyer toutes les valeurs d'une colonne avec une seule fonction.
- Si l'erreur est unique, il faut nettoyer au cas par cas.

Structure de donnée

- La structure de la donnée n'est pas bien formaté rendant l'exploitation impossible.
- Il faut modifier la structure et réunir les différentes sources de données en une source de données unique.

Tableau 1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

Tableau 1 + Tableau 2

	A	B	C	D	F
0	A0	B0	C0	D0	NaN
1	A1	B1	C1	D1	NaN
2	A2	B2	C2	D2	NaN
3	A3	B3	C3	D3	NaN
2	NaN	B2	NaN	D2	F2
3	NaN	B3	NaN	D3	F3
6	NaN	B6	NaN	D6	F6
7	NaN	B7	NaN	D7	F7

Tableau 2

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Data Life Cycle

IV. Feature Engineering & Data Augmentation

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.



Feature Engineering

Data Transformation

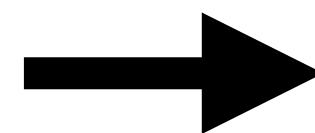
1. Appliquer une transformation logarithmique pour une meilleure répartition de la donnée.
 - $x =: \log(x + 1)$
 - Réduit le nombre d'outliers
 - Répartie mieux la donnée, change les écarts, ...
2. Regrouper la donnée en catégories pour simplifier et réduire la répartition des valeurs.
 - **(0 - 12)** -> Enfant
 - **(12 - 16)** -> Adolescent
 - **(17 - 24)** -> Jeune Adultes
 - **(24 - ...)** -> Adultes

Feature Engineering

One-Hot Encoding

- Le One-Hot Encoding est une façon de convertir les données catégorique en valeur numérique.
- Ce procédé est nécessaire car la plupart des algorithmes d'intelligence artificielle ne prennent en input que des valeurs numériques.

Id	...	Pays
1	...	France
2	...	Grece
3	...	France
4	...	Brésil



Id	...	Pays_France	Pays_Grece	Pays_Bresil
1	...	1	0	0
2	...	0	1	0
3	...	1	0	0
4	...	0	0	1

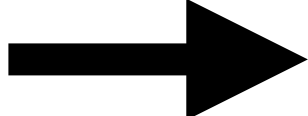
Feature Engineering

Feature extraction

- Le procédé d'extraire plusieurs features à partir d'une.

▸ Exemple : Date


Id	Date			
1	"25-05-1996"			



Id	Jour	Month	Year
1	25	5	1996

▸ Exemple : Interval

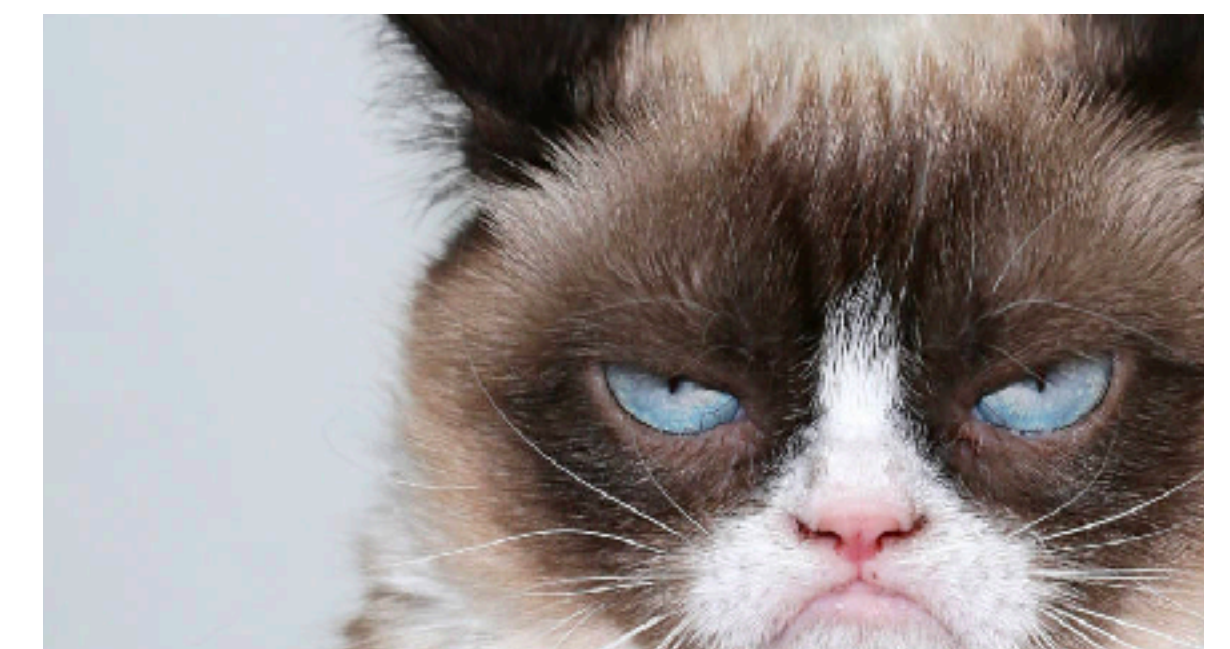
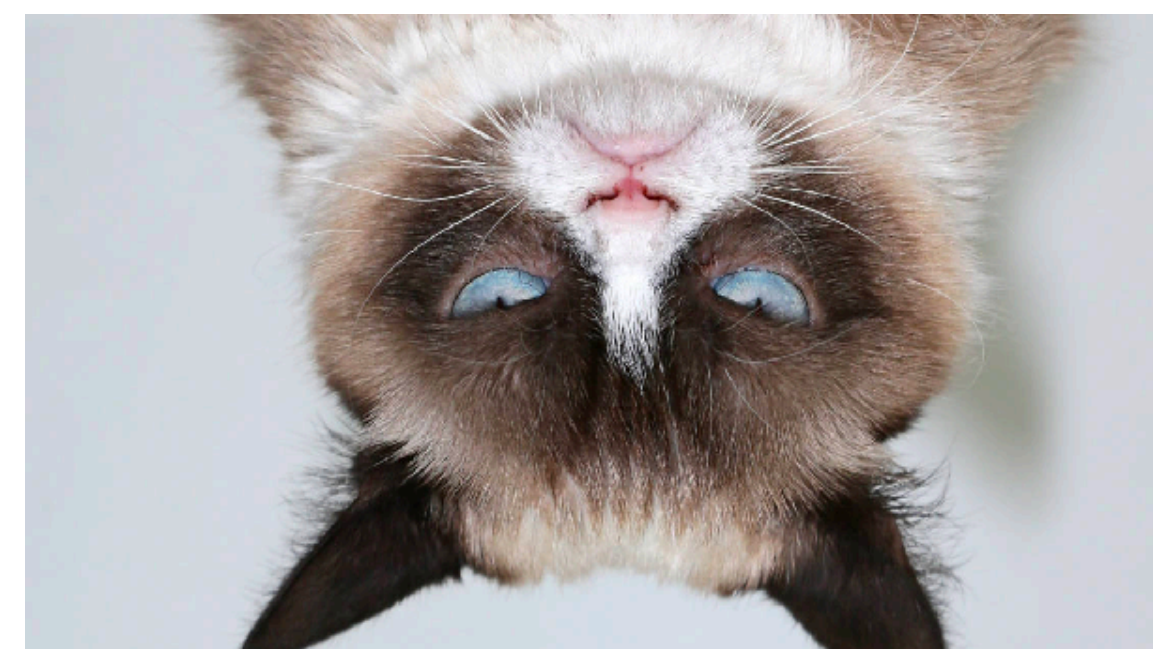
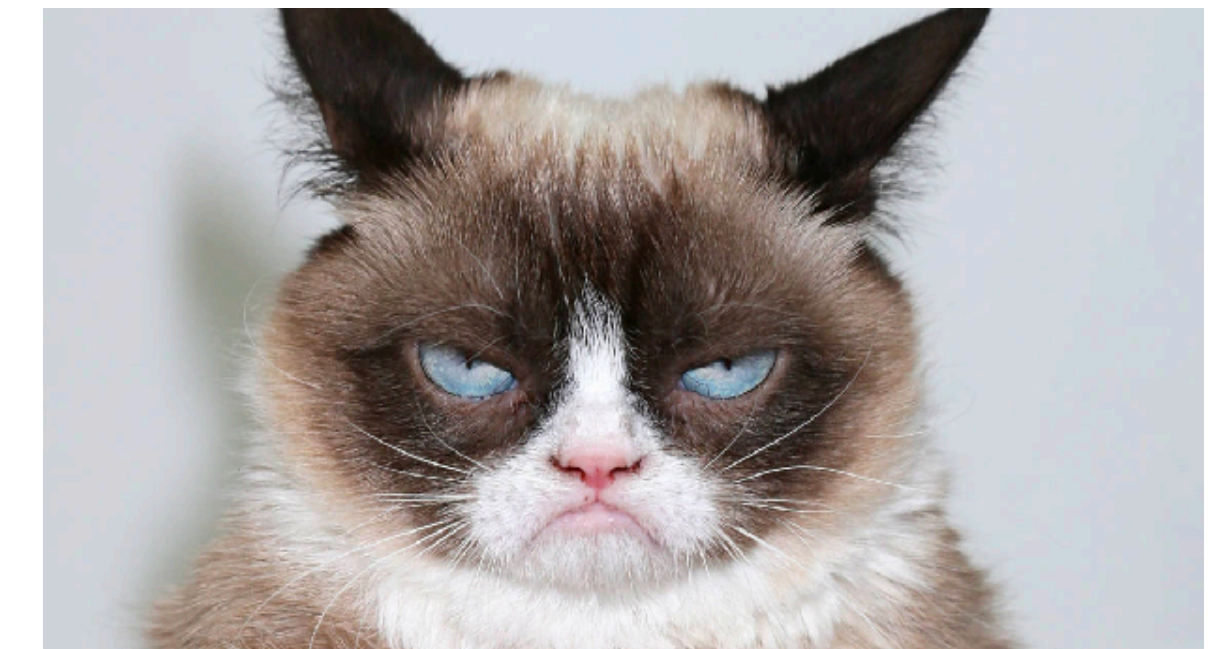
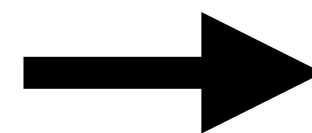
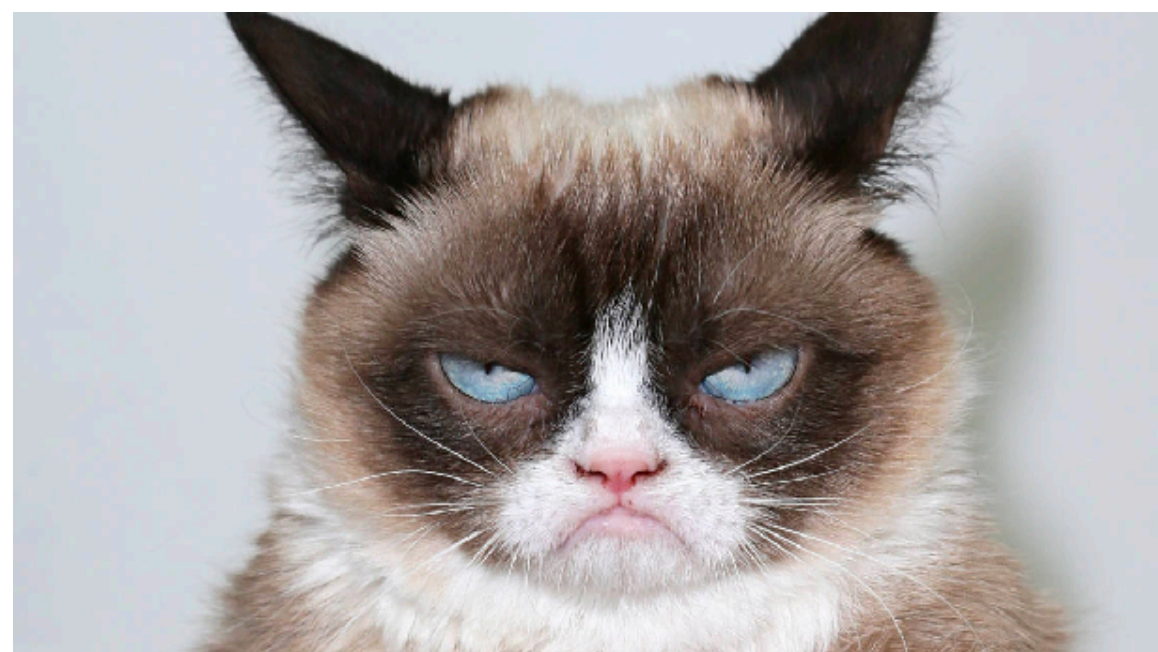
Date Envois	Date Réception		
"25-05-1996"	"27-05-1996"		



Date Envois	Date Réception	Interval (jour)
"25-05-1996"	"27-05-1996"	2

Data Augmentation

- La data augmentation consiste à augmenter le nombres de données dans un jeu de donnée en modifiant légèrement des données déjà existantes.
- Exemple : Images
 - Retourner, filtres de couleurs, inverser, rogner, ...



Biais, Contagion & Ethique

Biais

- Il est impératif de maîtriser le biais dans un jeu de données.
- Aujourd'hui les algorithmes de machine learning ont énormément de mal à sortir de la distribution qu'on leur donne en entraînement. (*Idiosyncrasy*)

Exemple

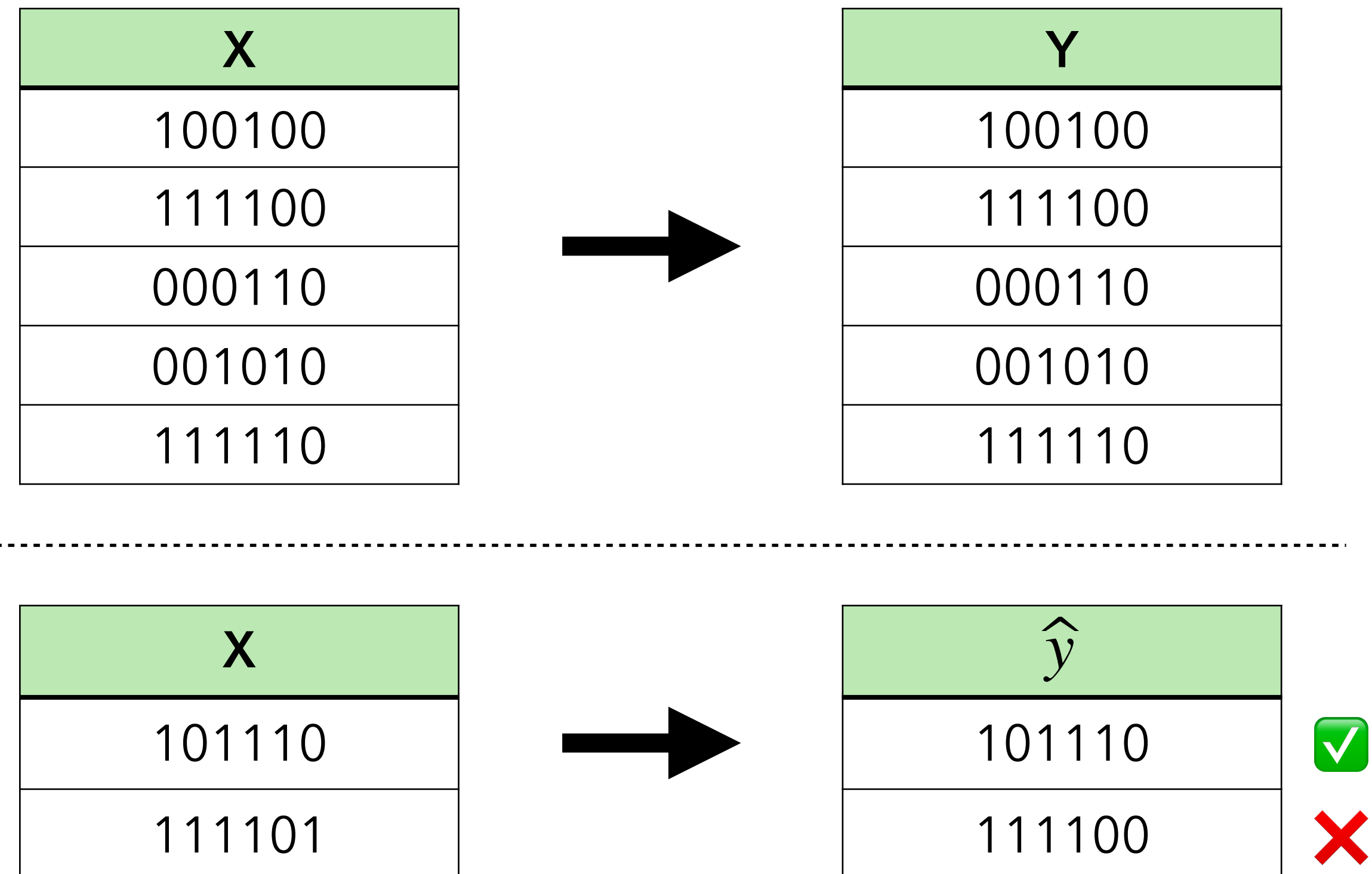
Données d'entraînement

On entraîne un modèle à prédire une simple réflexion.

Note : Aucune valeur d'input ne possède de **1** en dernière position.

Données de test

Le modèle est incapable de prédire le 2ème test car il n'y avait jamais eu de 1 en dernière position. Il n'a donc pas appris la règle de "réflexion".



Contagion

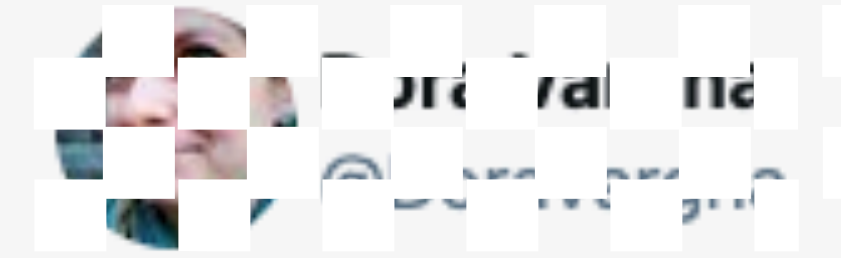
Un exemple en Mars 2021

Google Translate traduit une langue non-genré vers l'anglais, il se retrouve biaisé par sa donnée d'entraînement.

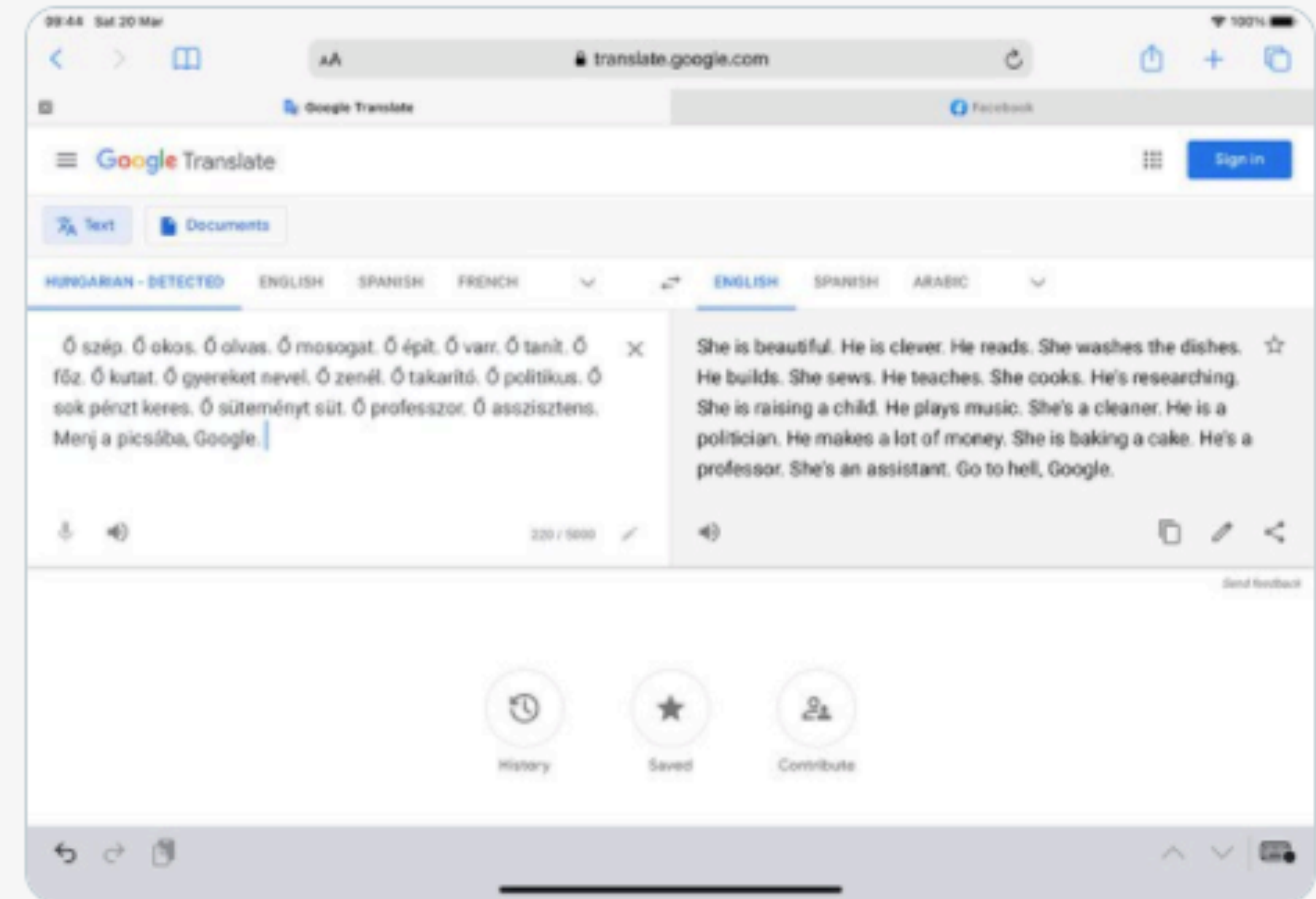
La traduction :

"She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant. Go to hell, Google."

<https://www.thebestsocial.media/blog/google-translate-hungarian/>



Hungarian is a gender neutral language, it has no gendered pronouns, so Google Translate automatically chooses the gender for you. Here is how everyday sexism is consistently encoded in 2021. Fuck you, Google.



Les algorithmes de ML apprennent comme des enfants.
Le résultat dépend de leur éducation et du contexte.

Le biais de donnée fait partie du design d'une application et **DOIT** être pris en compte, testé et validé.

Le biais de donnée fait partie du design d'une application et **DOIT** être pris en compte, testé et validé.

TayBot et Microsoft

CEO -> 🧑 et Apple

Discriminating Law prediction

Machine a Savon Raciste?

https://www.youtube.com/watch?v=YJjv_OeiHmo