Les données: pierre angulaire de l'intelligence artificielle

Guillaume Chicoisne

IVADO

HEC Montréal
Polytechnique Montréal
Université de Montréal

TL;DR

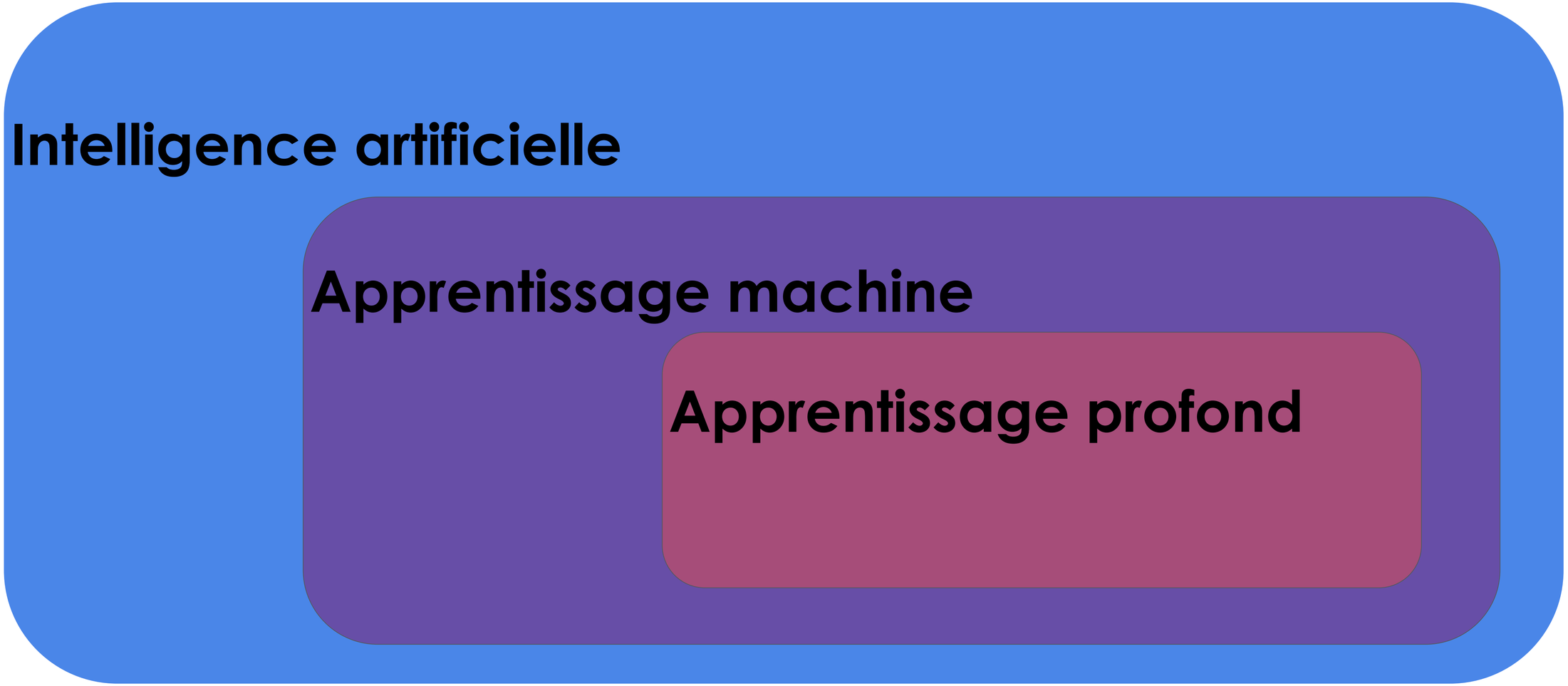
- L’intelligence artificielle actuelle se construit sur les données
- Manipuler et comprendre les données sont deux compétences différentes
- Le monde des données n’est pas le monde réel

Intelligence artificielle et intelligence naturelle

Définition?

- C'est compliqué…
- Autant définir l'intelligence naturelle...

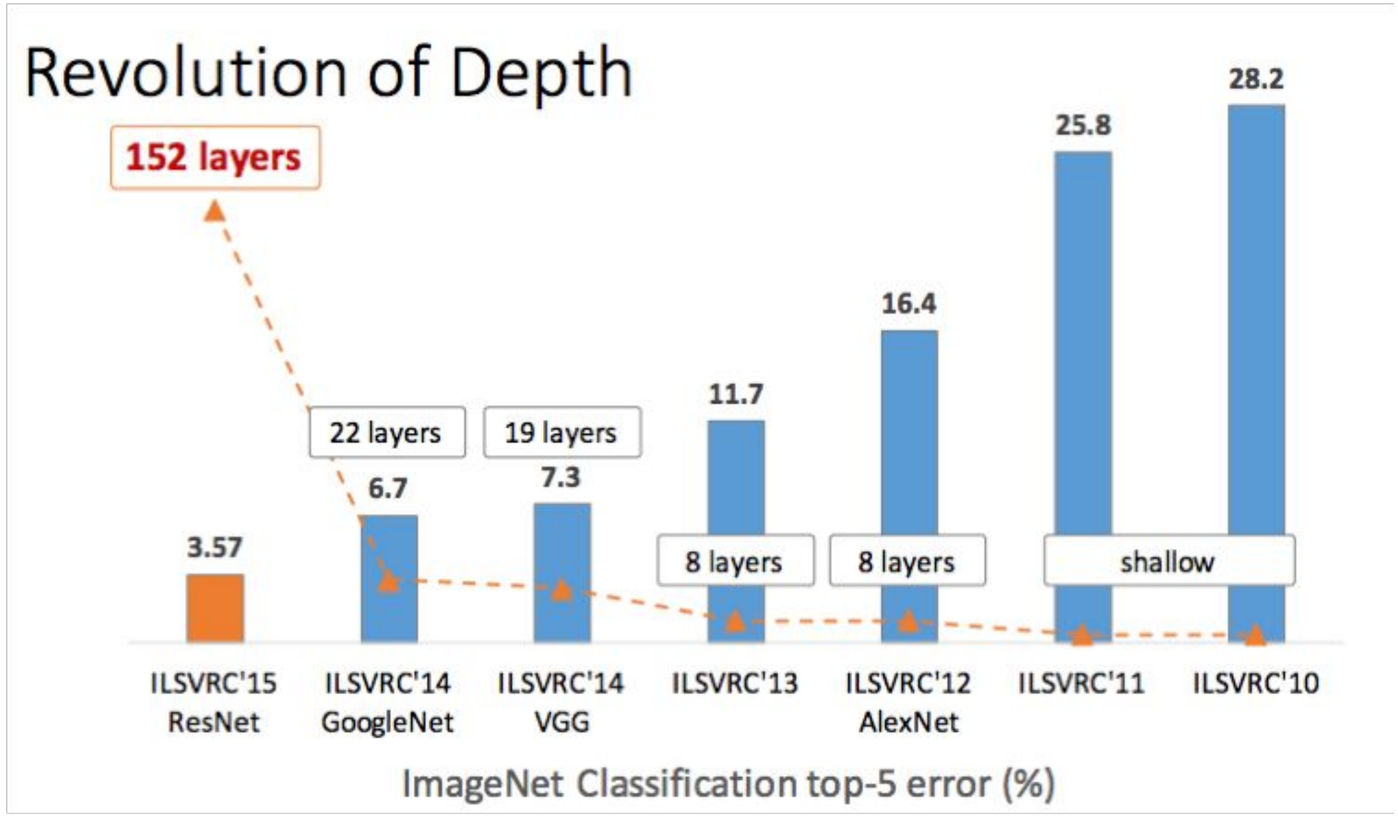
Intelligence artificielle ou *deep learning*?



La nouvelle vague de l'IA

+ Des **données** de plus en plus nombreuses et accessibles
+ Des **outils** capables d'utiliser une grande quantité de données
+ De la **puissance** de calcul

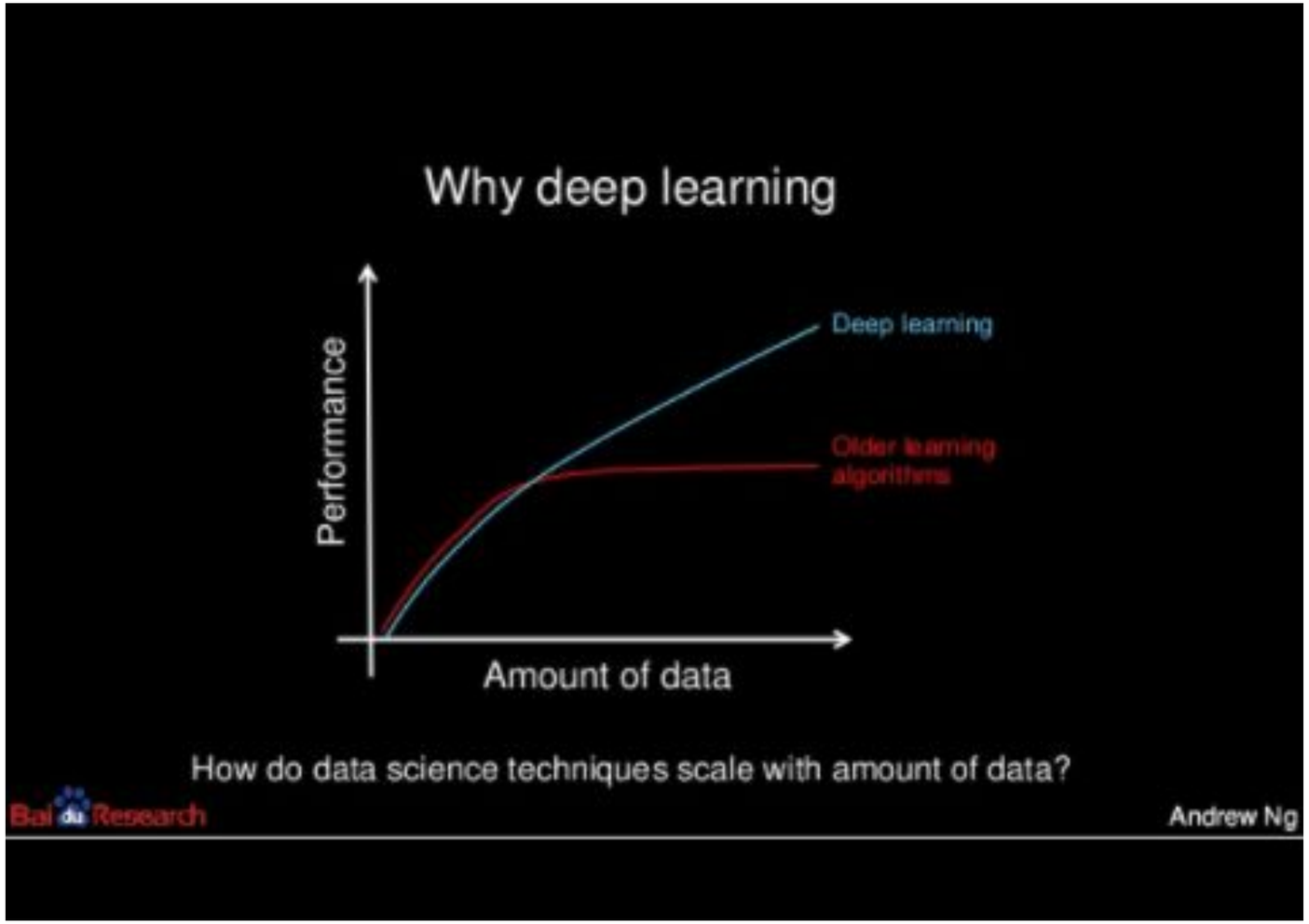
= résultats sur des tâches très différentes: reconnaissance de parole, analyse d'image, ...



La nouvelle vague de l'IA

+ Des **données** de plus en plus nombreuses et accessibles
+ Des **outils** capables d'utiliser une grande quantité de données
+ De la **puissance** de calcul

= résultats sur des tâches très différentes: reconnaissance de parole, analyse d'image, ...



Le nouveau pétrole?

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

- L'accès est parfois difficile et coûteux
- La valeur augmente avec le raffinage
- Raffiner coûte cher
- Avantage économique à la concentration des ressources
- ...

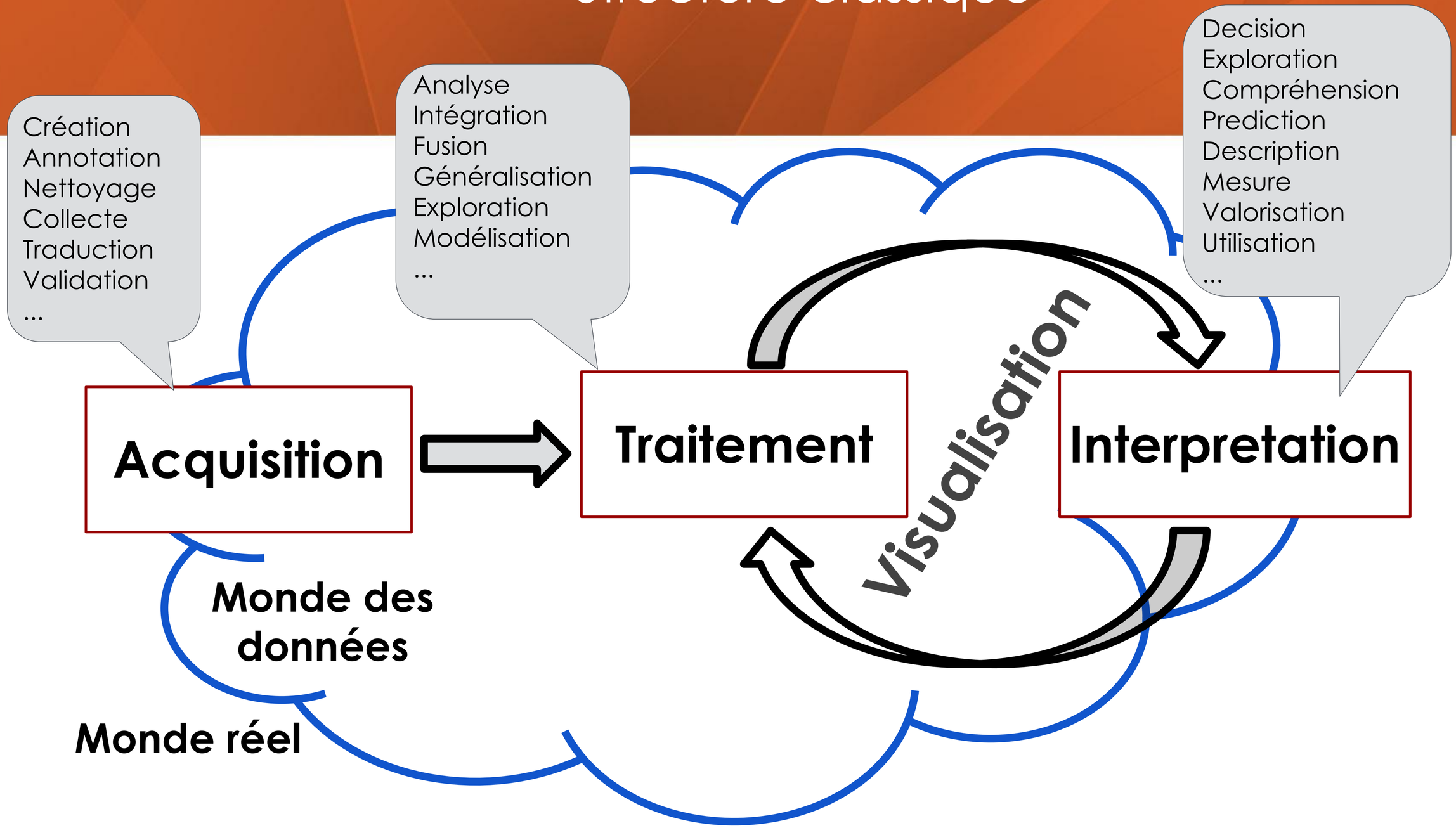
Données?

Les données sont une façon de numériser le monde.

- Ce qui nous intéresse n'est pas forcément mesurable
- Ce qui est facilement mesurable n'est pas forcément ce qui nous intéresse
- Les mesures indirectes doivent générer d'autant plus de méfiance

Le monde des données n'est pas le monde réel

Structure classique



Créer, acquérir, accéder
(le 80% laborieux)

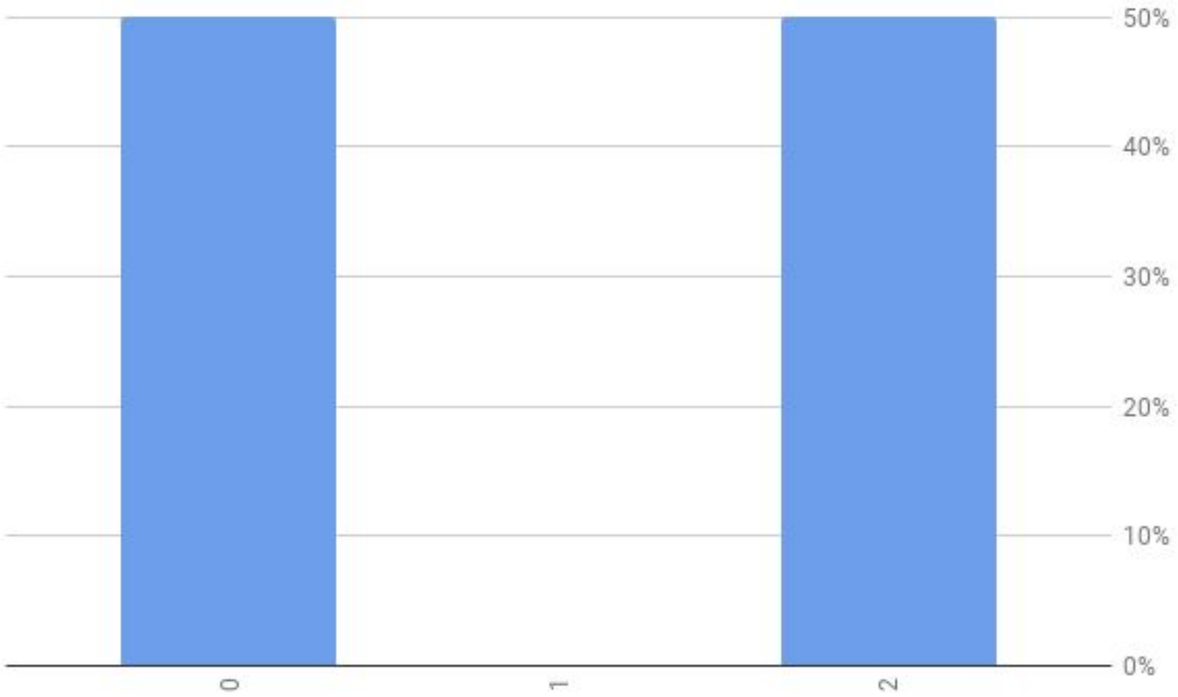
(un banquier et son client)

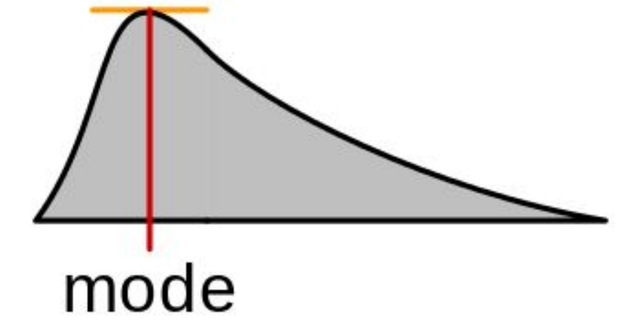
- C'est quoi l'problème mes dollars m'appartiennent, j'suis supposé pouvoir les retirer facilement?!
- Non!
- Quoi?
- Vos fémurs aussi vous appartiennent, pis vous pouvez pas les retirer facilement!

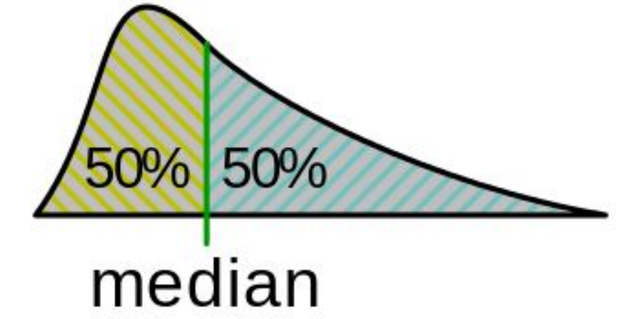
(François Pérusse, Philosophe québecois)

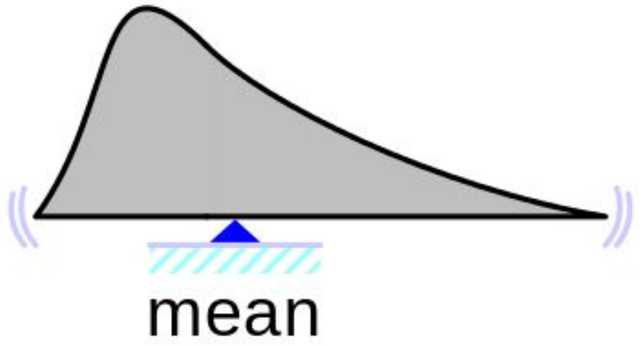
Intermède statistique

- Moyenne? Mode? Médiane?
- Correlation?

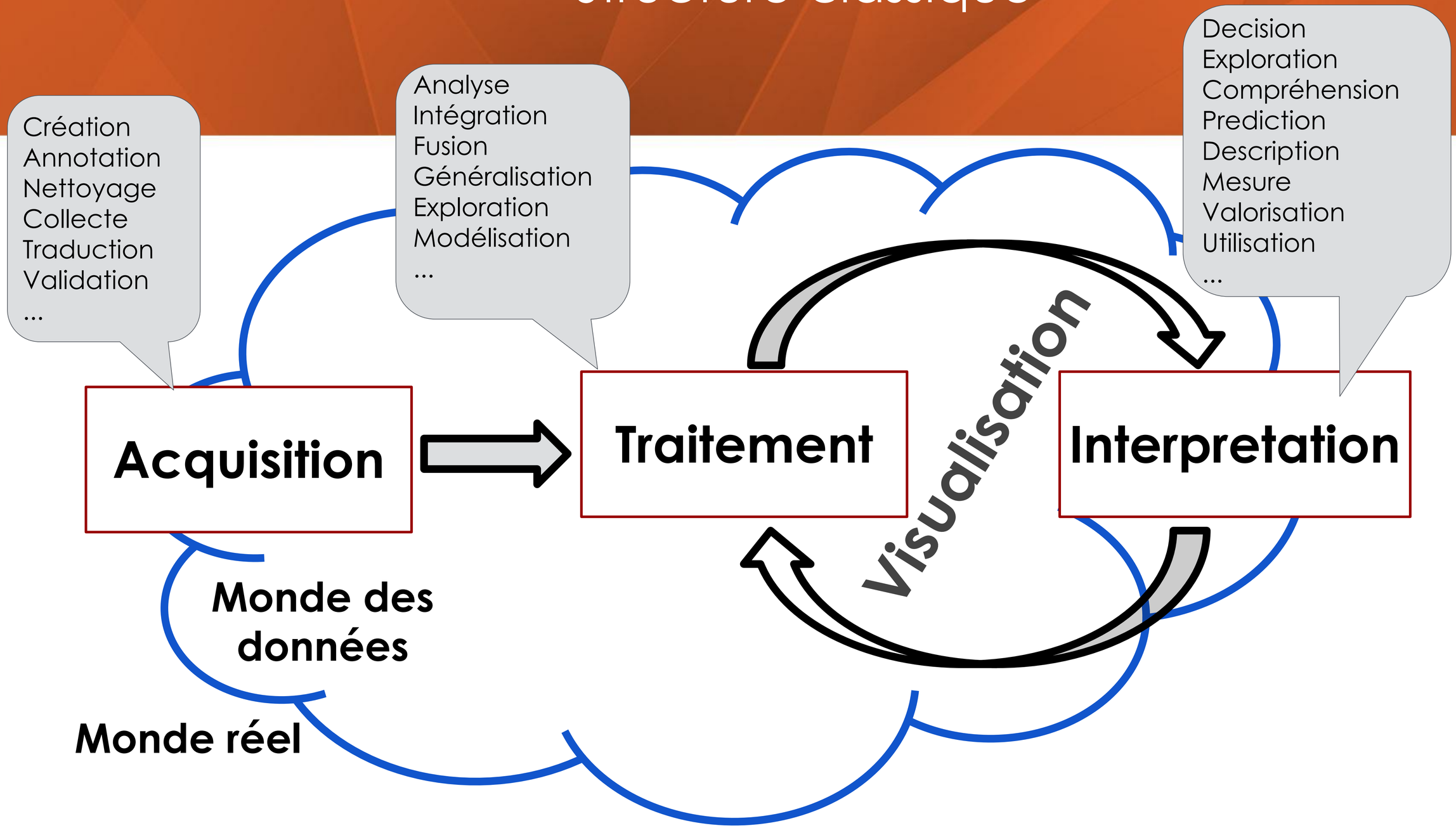






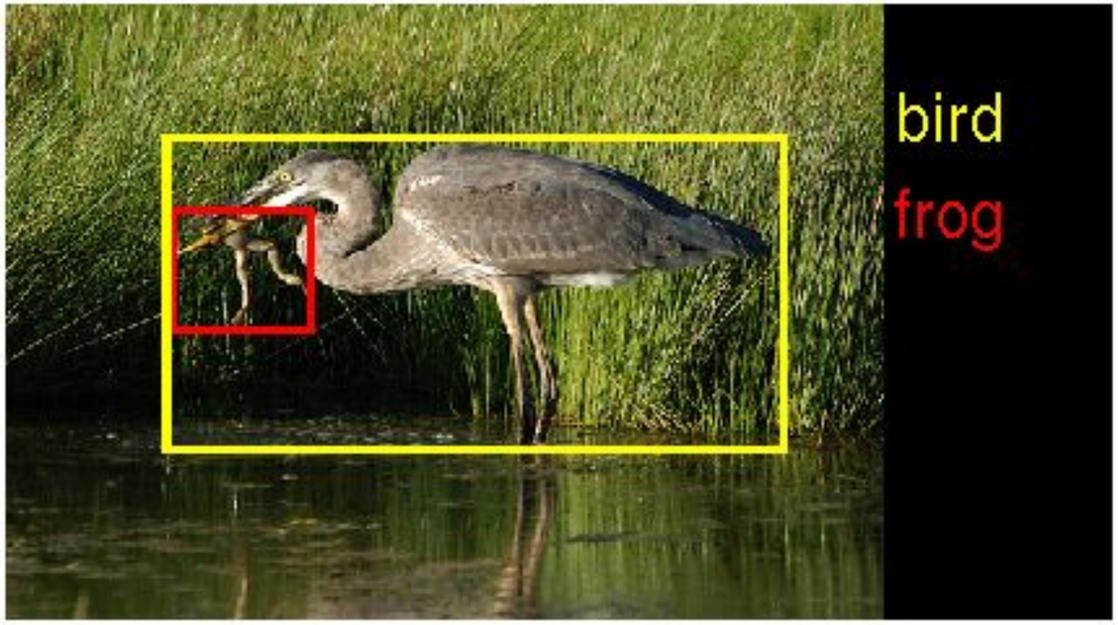


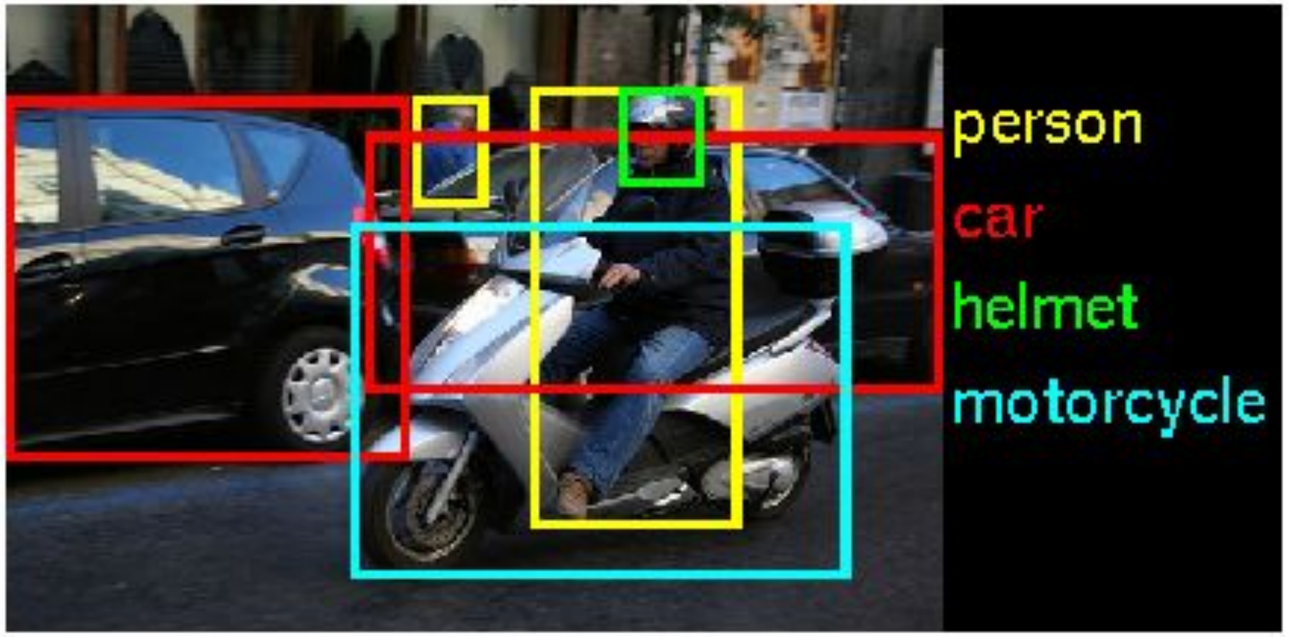
Structure classique



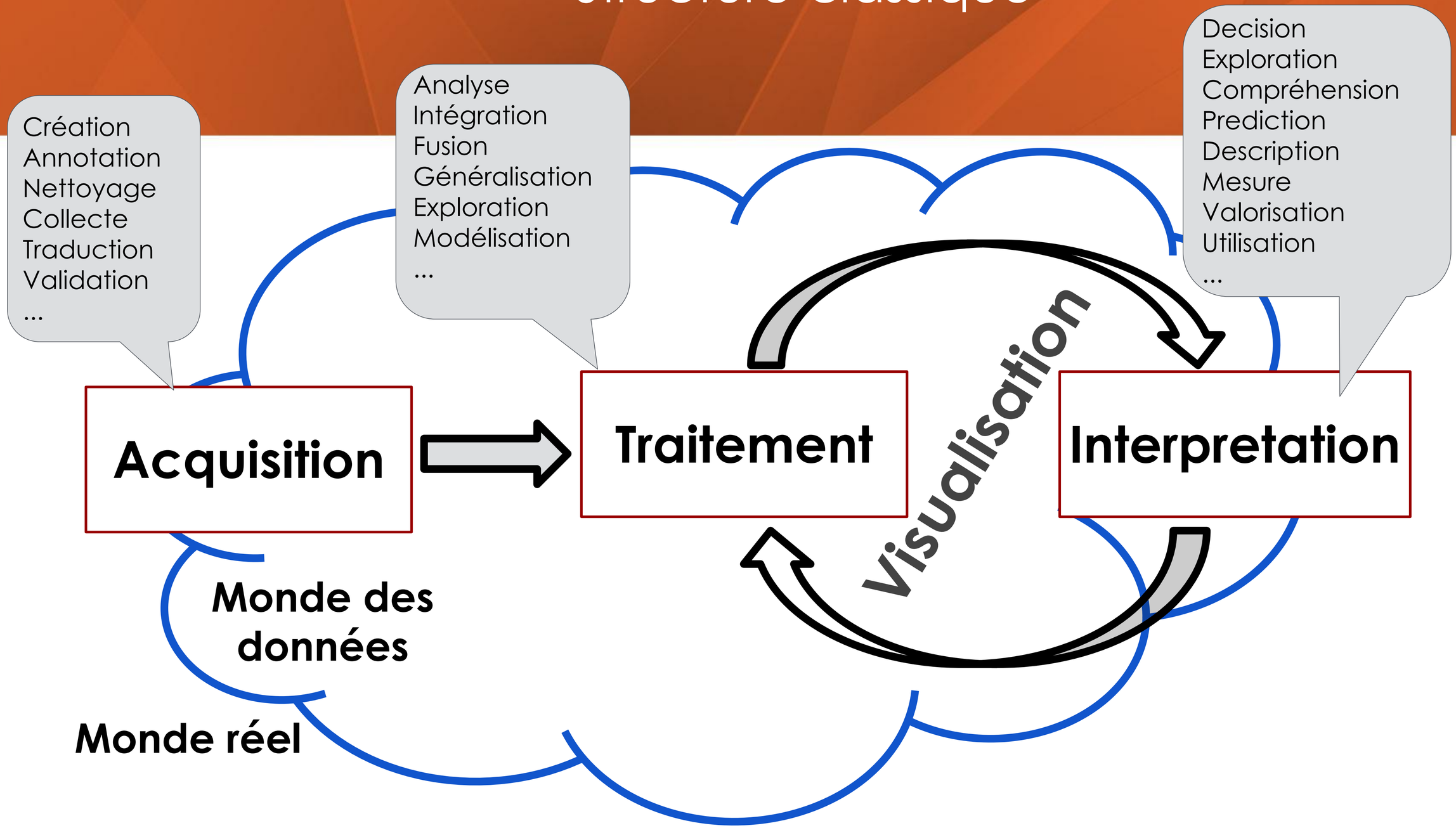
Traitement

- Grand choix de méthodes, depuis les modèles statistiques simples, jusqu'à l'apprentissage profond.
- Les méthodes les plus simples sont parfois les plus adaptées



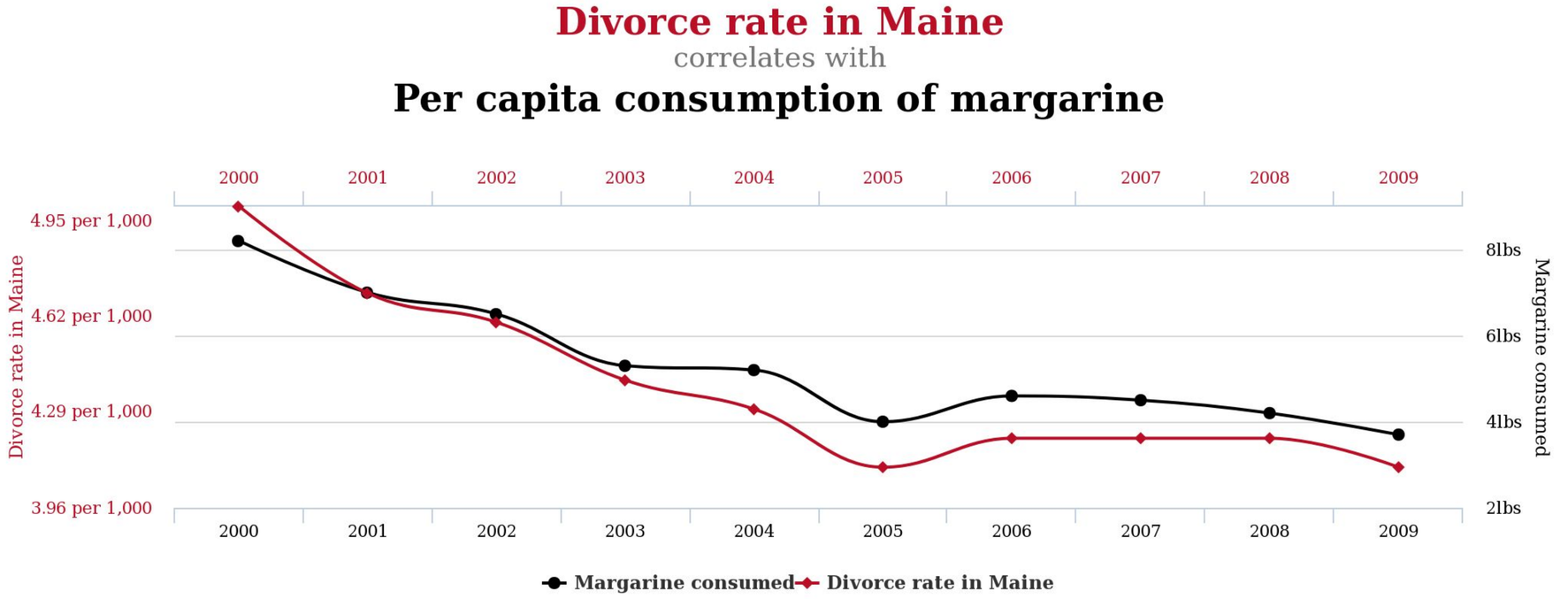


Structure classique

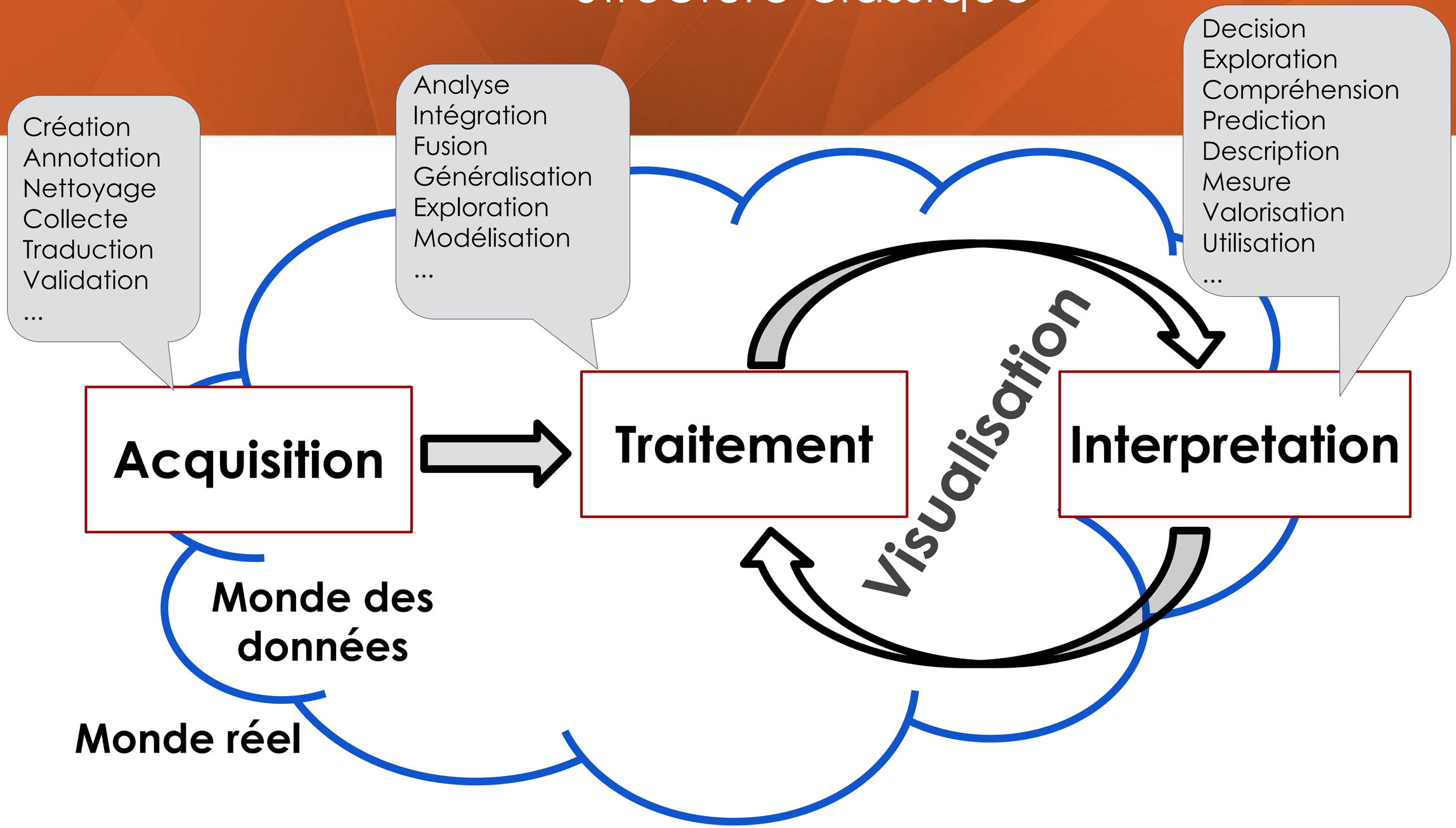


Interprétation

- Confronter les analyses au monde réel:
  - Prendre ou appuyer une décision, mesurer une progression, explorer …
  - Valider les résultats!

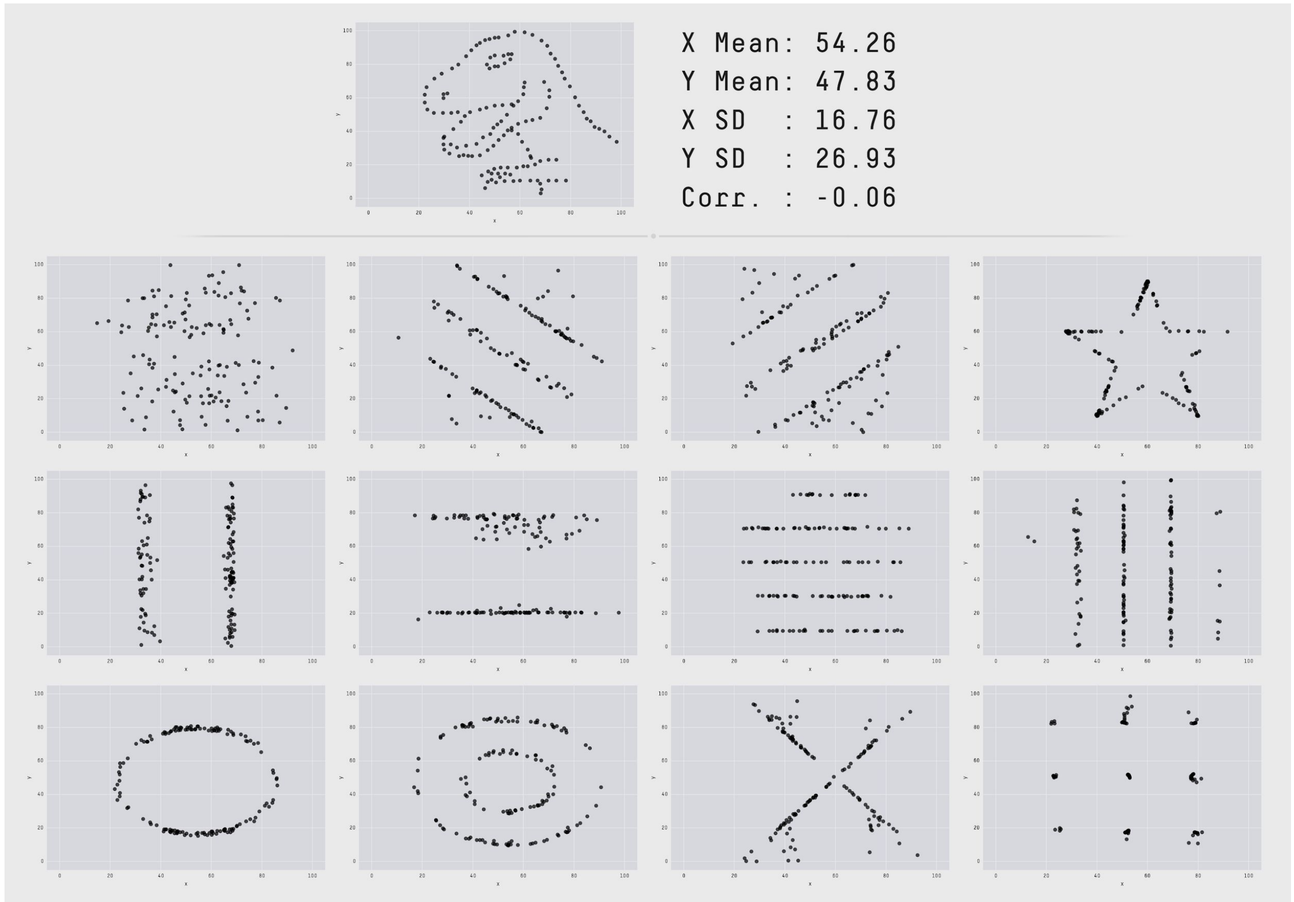


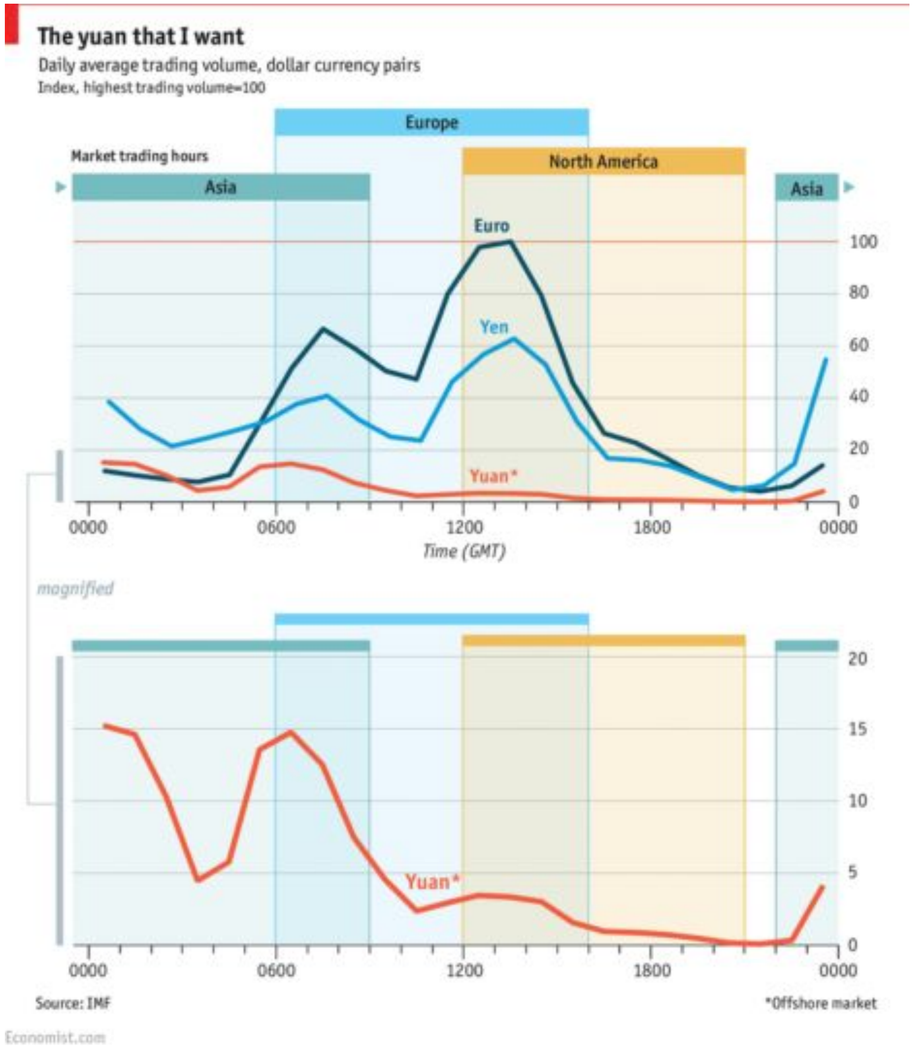
Structure classique



Visualisation

- Permet de “voir” où on en est
- Peut être un objectif en soi





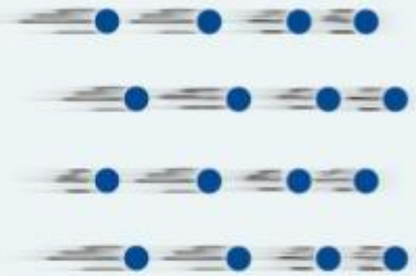
Données massives - une définition parmi d'autres

Volume

Données brutes d'origine

Terabytes à exabytes de données disponibles

Vitesse



Données dynamiques

Analyse en temps réel, décision en une fraction de seconde

Variété

Données hétérogènes

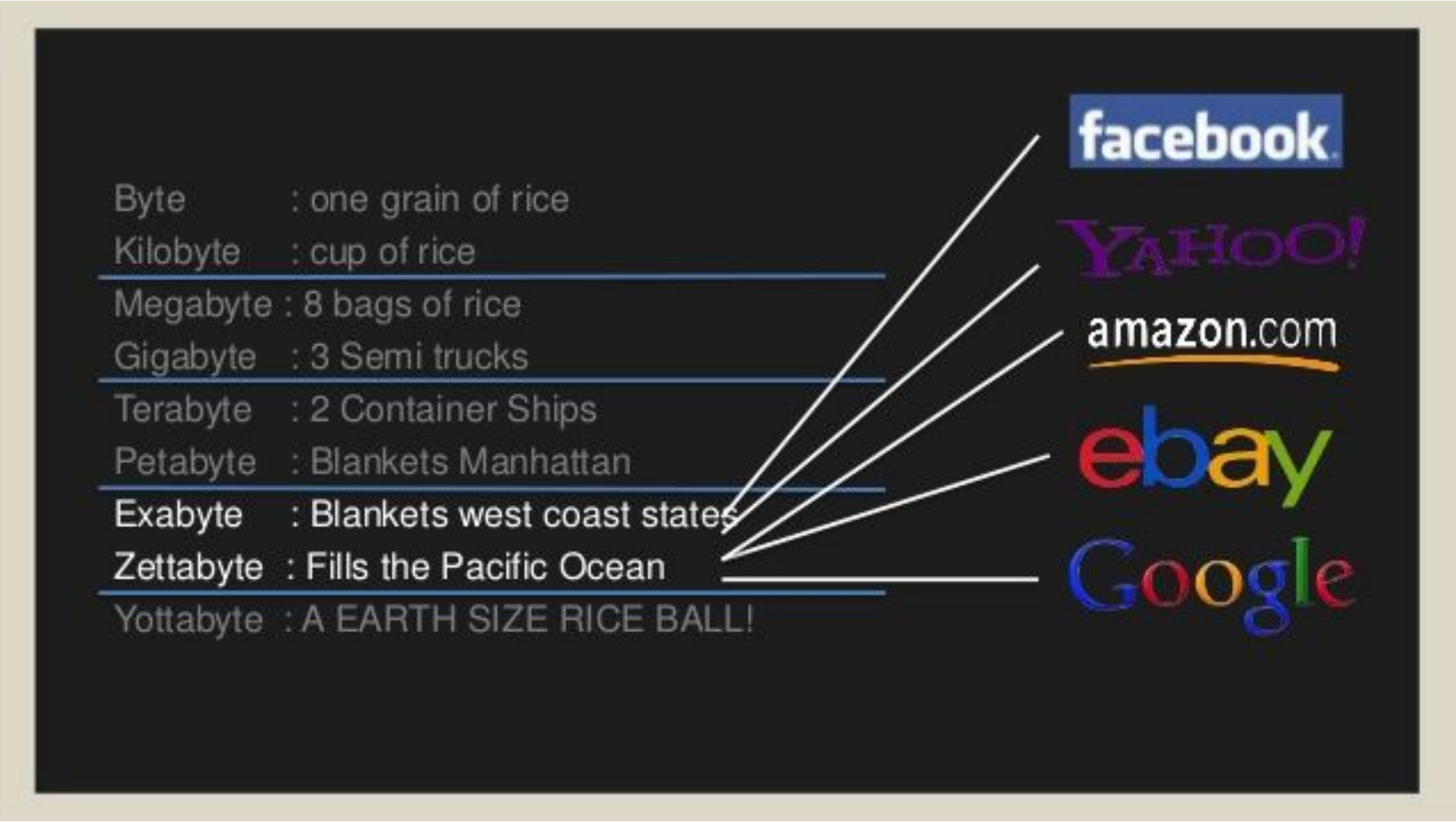
Format structuré, non structuré, texte, multimédia

Véracité

Données incertaines

Cohérence, fiabilité, qualité et prédictibilité des données

Volume

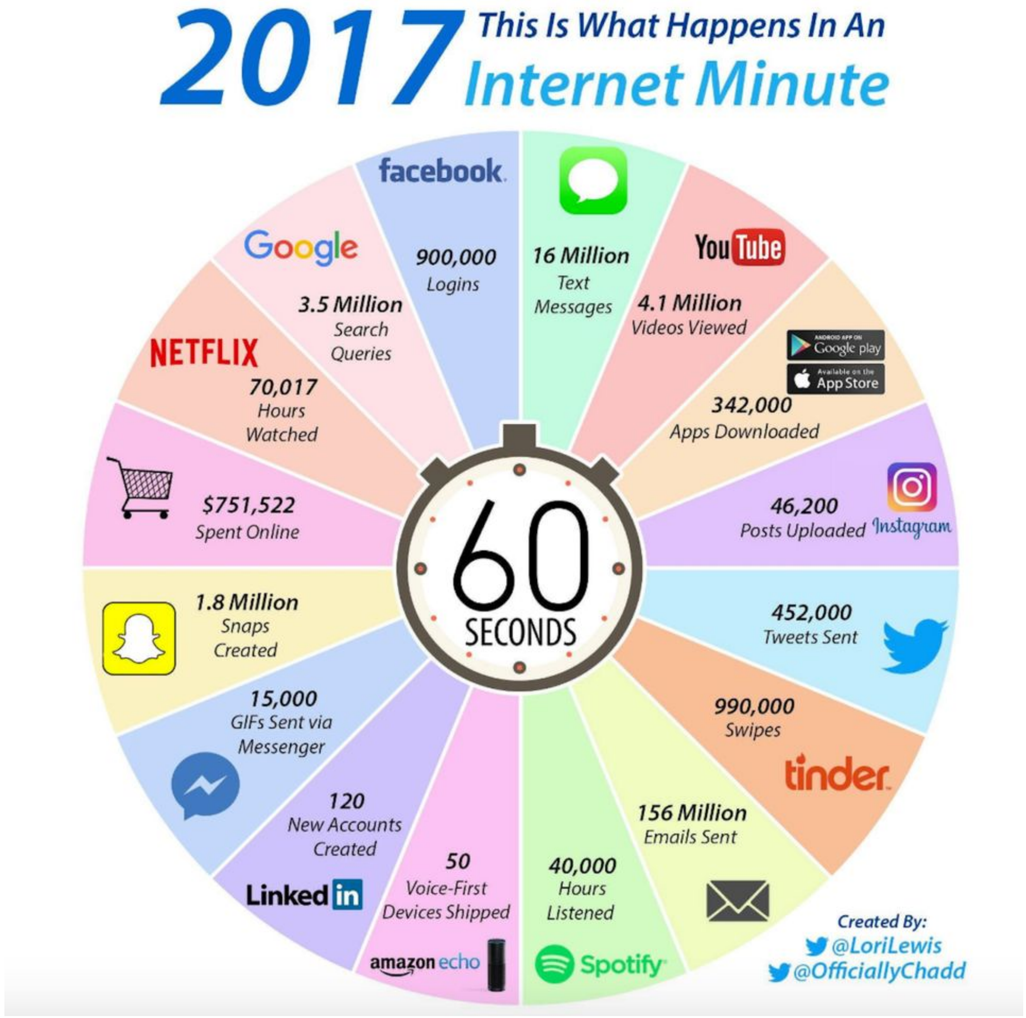


N.B.: La perfection est inatteignable. Un test médical valide à 99,9% appliqué la population du Canada (35 millions) va se tromper pour 35,000 personnes

Volume

- Un **mega**octet: une minute de musique
- Un **giga**octet: **une heure de film**
- Un **tera**octet: **40 jours de films** non-stop
- Un **peta**octet: **un siècle** devant un écran
- Un **exa**octet: **100 000 ans**
    - (et pourtant: c'est la durée de films regardés **chaque jour** sur YouTube en 2017)
- Un **zetta**octet: **100 milions d'années**
    - (ancêtre commun homme/souris)
- Un **yota**octet: **100 milliards d'années**
    - (penser à rechagrer la tablette en attendant l'apparition de l'univers)

Vitesse



Et ça ne compte pas les objets connectés…

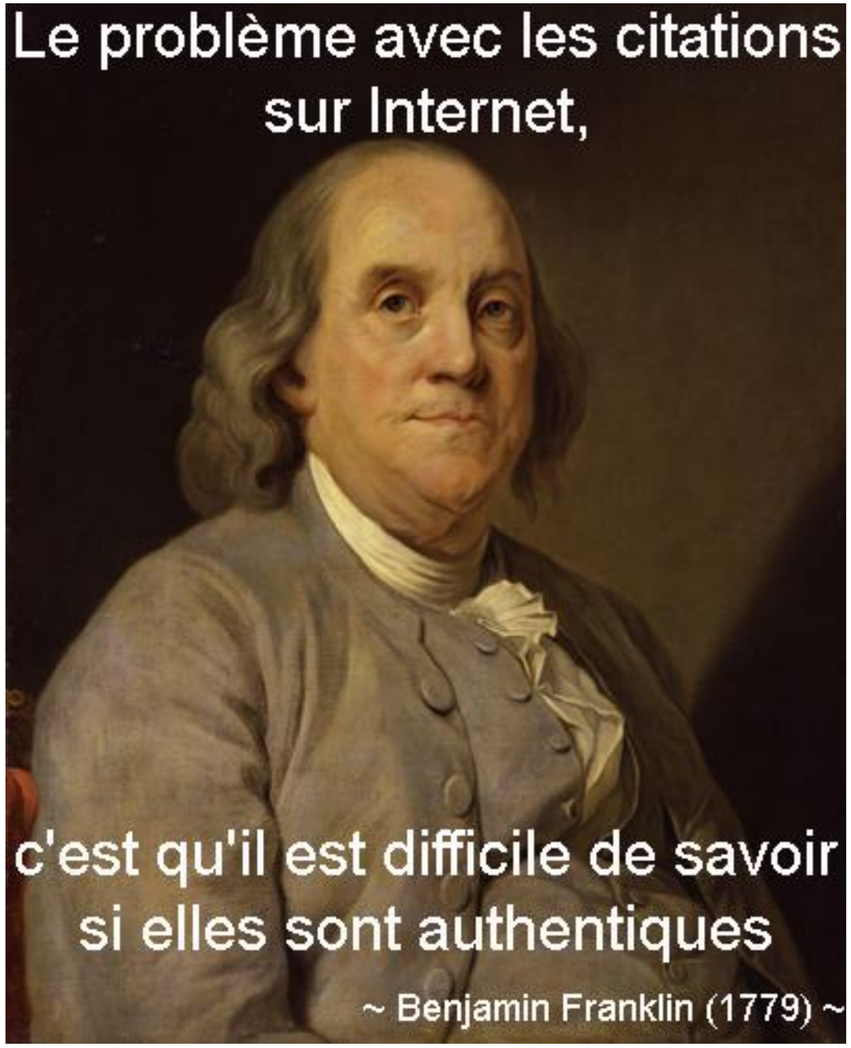
Variété

Un GPS, une liaison cellulaire, des ordinateurs, le WIFI, un réseau local, de la musique, des films, des jeux, … et des mises à jour logiciel!

Véracité

- Pas seulement les données fausses ou trompeuses
    - Données anciennes, données “rien à voir”, données trop indirectes
    - Données biaisées
    - Erreur, bruit, ...
- Pas toujours malicieux
    - Fake news vs. protection de la vie privée

(rappel: une IA ne peut pas être meilleure que ses données)



Un autre V: la Valeur

Coûts

- Acquisition
  - Capteurs, achat, plateforme, ...
- Traitement
  - Puissance de calcul, experts, ...
- Stockage et accès
- ...

Revenus

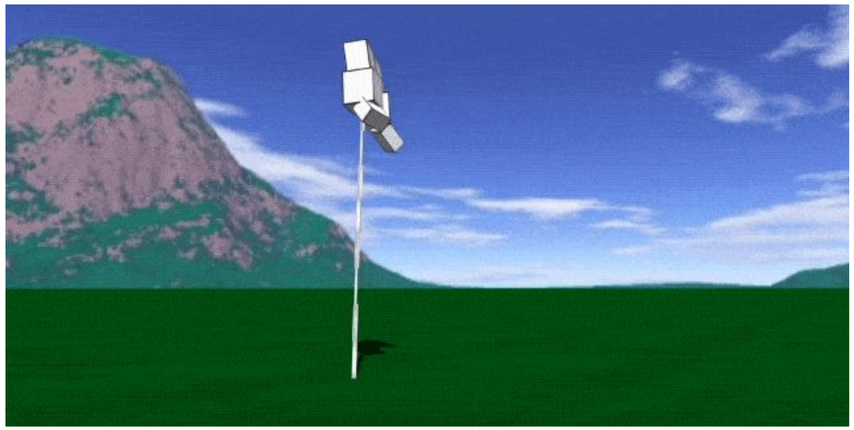
- Prise de décision
- Publicité
- Revente des données
- Valeur future espérée
- ...

Approches basées sur les données

- Attention:
    - Prolongation de l'historique, risque de limiter l'innovation
    - le système optimise ce qu'on lui demande d'optimiser
- Mais le système a une intelligence *artificielle*
    - ce qui lui permet d'explorer des solutions "non-humaines"

...pour le meilleur

et pour le pire...



Accès aux données: pas seulement des défis techniques

Aspects techniques: volume de données, réseau rapide, accès aux sources, distribution des données et du calcul, consommation, etc.

Aspects non-techniques: vie privée, droits d'utilisation, conformité, confidentialité, sécurité… **Responsabilité.**

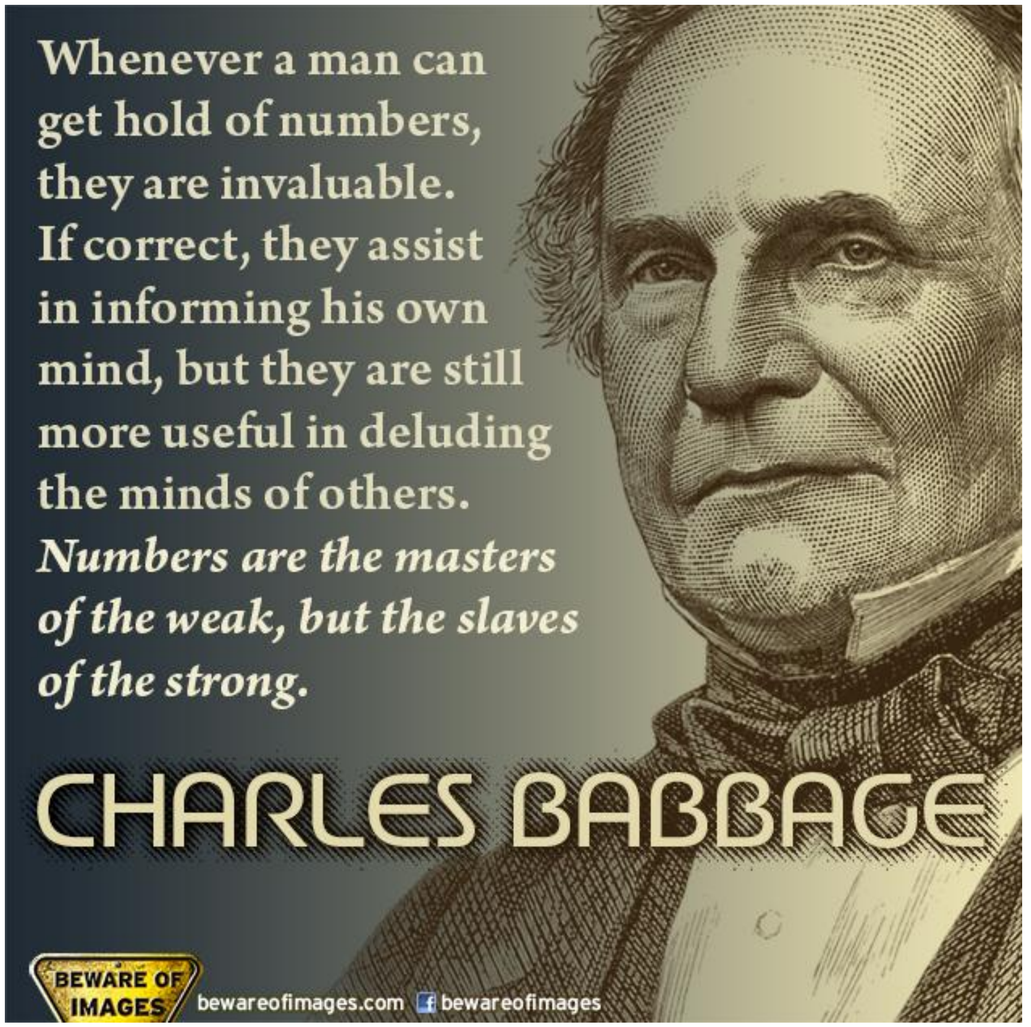
Trois notions parmi d'autres:

- Quasi-identificateurs
- *Mi data es su data...*
- Sécurité: attaques, défaillances, accidents, bévues, …

Les données n'ont pas d'odeur...

Chaque fois qu'un homme peut obtenir des chiffres, ils sont inestimables: s'ils sont corrects, ils aident à informer son propre esprit, mais ils sont encore plus utiles pour tromper l'esprit des autres. *Les nombres sont les maîtres des faibles, mais les esclaves des forts.*
Charles Babbage (1864)

(traduction: google translate… et réseaux de neurones profonds)



La nouvelle vague de l'IA

26 mars à l'UPop!
(Guillaume)

+ Des **données** de plus en plus nombreuses et accessibles
+ Des **outils** capables d'utiliser une grande quantité de données
+ De la **puissance** de calcul

= résultats sur des tâches très différentes: reconnaissance de parole, analyse d'image, ...

9 et 23 avril à l'UPop!
(Bruno, Margaux, Guillaume)

Et avec ces grands pouvoirs viennent… de grandes responsabilités: **que faire** avec ces outils?

7 mai à l'UPop!
(Martin)

Merci!

Besoin de détails?

Guillaume.chicoisne@ivado.ca
ivado.ca

IVADO

HEC Montréal
Polytechnique Montréal
Université de Montréal

Sources

- https://www.eff.org/ai/metrics
- https://www.economist.com/technology-quarterly/2017-05-01/language
- https://minghsiehee.usc.edu/2017/04/the-machines-are-coming/
- https://medium.com/@Lidinwise/the-revolution-of-depth-facf174924f5
- https://en.wikipedia.org/wiki/File:Visualisation_mode_median_mean.svg
- http://www.leparisien.fr/automobile/voiture-autonome-un-blesse-dans-une-tesla-en-pilotage-automatique-30-09-2016-6163713.php
- https://www.tubmanchev.com/chevrolet-spark/
- http://www.polecat.com/wp-content/uploads/2016/10/20161008_woc112_1-466x540.png Yuan
- https://www.economist.com/blogs/graphicdetail/2016/05/daily-chart-18 extreme-droite
- http://www.businessinsider.com/everything-that-happens-in-one-minute-on-the-internet-2017-9
- https://en.wikipedia.org/wiki/Simpson%27s_paradox#/media/File:Simpson%27s_paradox_continuous.svg
- https://www.happybeertime.com/wp-content/uploads/2014/07/Benjamin_Franklin_final_quote.jpg
- https://www.washingtonpost.com/news/the-switch/wp/2018/02/01/google-parent-alphabet-reports-soaring-ad-revenue-despite-youtube-backlash/?utm_term=.36d000e466ef
-