Big Data en Santé

La place de la donnée dans la recherche médicale de demain

Nicolas SAVY

Institut de Mathématiques de Toulouse

Atelier 2020 de la plateforme Ethique et Biosciences
« Usage éthique des Big Data en Biosciences »
Volet 1 : « Définition des Big Data: mise en commun et partage »

Jeudi 12 Mars 2020

Médecine de demain…

- I.A. symbolique basé sur la logique
- I.A. numérique basé sur la donnée
  - Prédiction collective
  - Prédiction individuelle

Comprendre ?

- Rôle central de la donnée
  - « Big Data » ?

Dossier de l'INSERM : Intelligence artificielle et santé - **06.07.18**

D’où vient le Big Data ?

1930-1970	hO (10^2)		Expérience planifiée avec n ~ 30 individus p < 10 variables	Statistique inférentielle
1970s	kO (10^3)	Généralisation des outils informatiques	n augmente et p augmente	Analyse de données
1980s	MO (10^6)		n augmente et p augmente	Modèles non-paramétriques ou fonctionnels
1990s	GO (10^9)		Les données non planifiées	Data mining
2000s	TO (10^{12})	Bio-technologies "omiques"	Le nombre de variables explose (10^4; 10^6)	machine learning + statistics = statistical learning
2010s	PO (10^{15})	e-commerce géo-localisation	Le nombre d’individus n explose	Apprentissage supervisée ou non

Source : Statistique, Apprentissage, Big-Data-Mining - Philippe Besse, www.wikistat.fr

D’où vient le Big Data ?

- « Big » c’est quand cela ne rentre pas ma machine…

- Le « Big Data » quand
  - les données ne peuvent être stockées sur un seul ordinateur
    - données réparties
  - les traitements vont nécessiter plusieurs machines
    - calculs distribués

- L’appellation « Big Data » a été introduite par Cox et Ellsworth (NASA), au congrès SIGGRAPH en 1997.

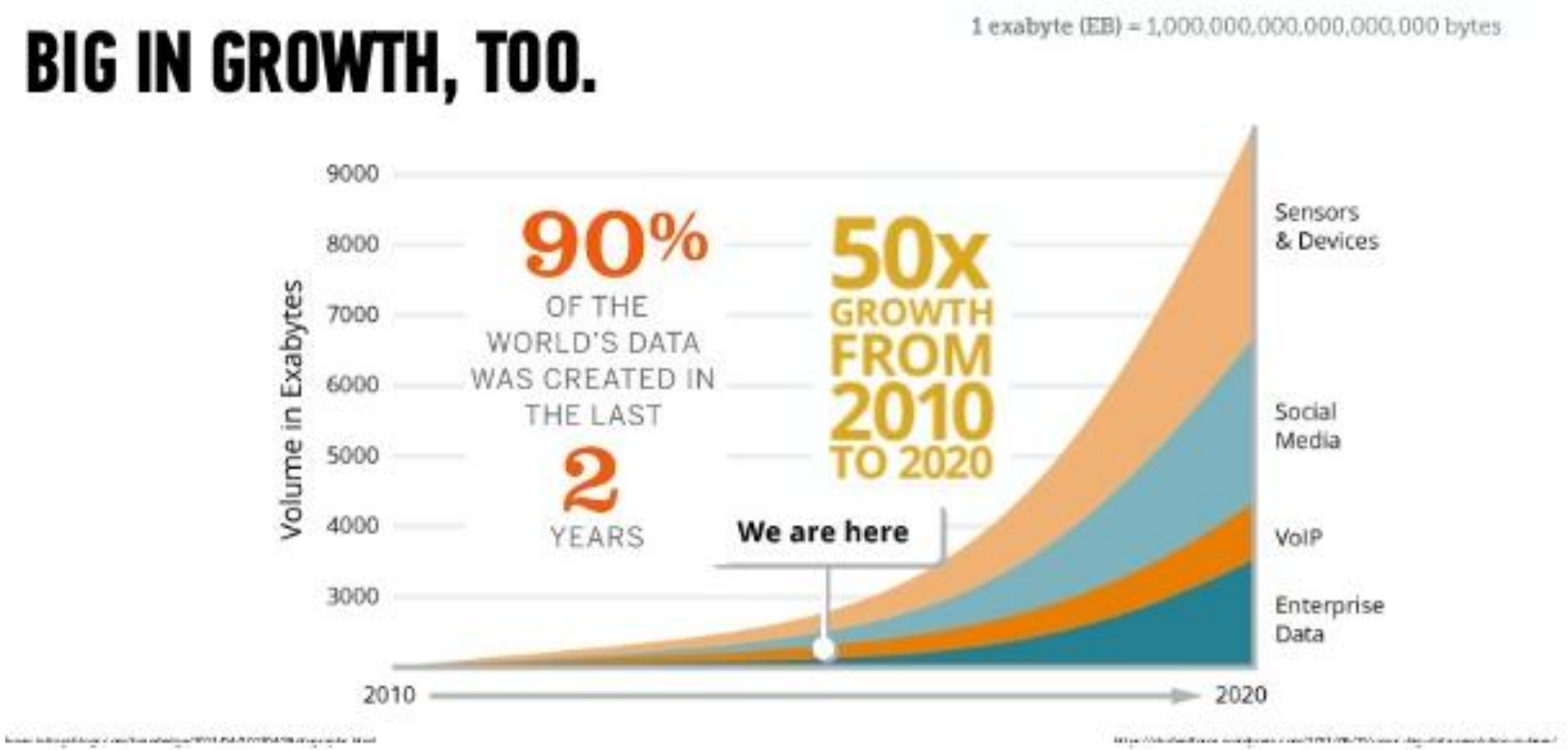
D’où vient le Big Data ?

- Le « Big Data » est une nouvelle manière de voir et d’analyser le monde.
- Nouveaux ordres de grandeur concernant
  - la capture
  - la recherche
  - le partage
  - le stockage
  - l’analyse
  - la présentation des données.

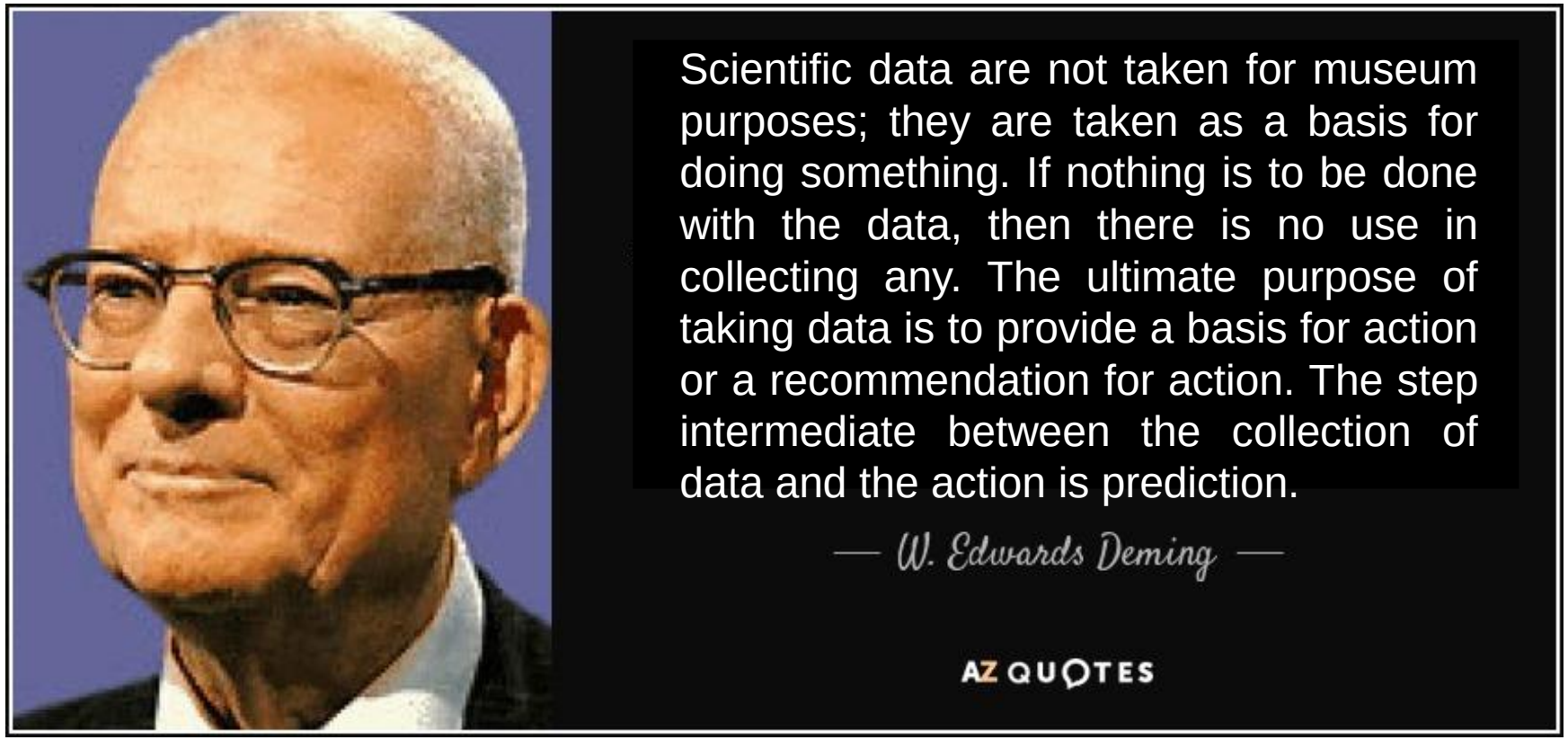
- Usuellement caractérisé par les trois « V »
  - **Volume** : quantité de données
  - **Variété** : hétérogénéité des données (venant de diverses sources, non-structurées, organisées, Open…)
  - **Vélocité** : fréquence de création, collecte et partage de ces données

- Auxquels on peut ajouter d’autres « V »
  - **Valorisation** : business
  - **Visualisation**
  - **Veracité** : justesse de l’information, qualité des données

Vous avez dit volumineux ?



Pourquoi stocker des données ?



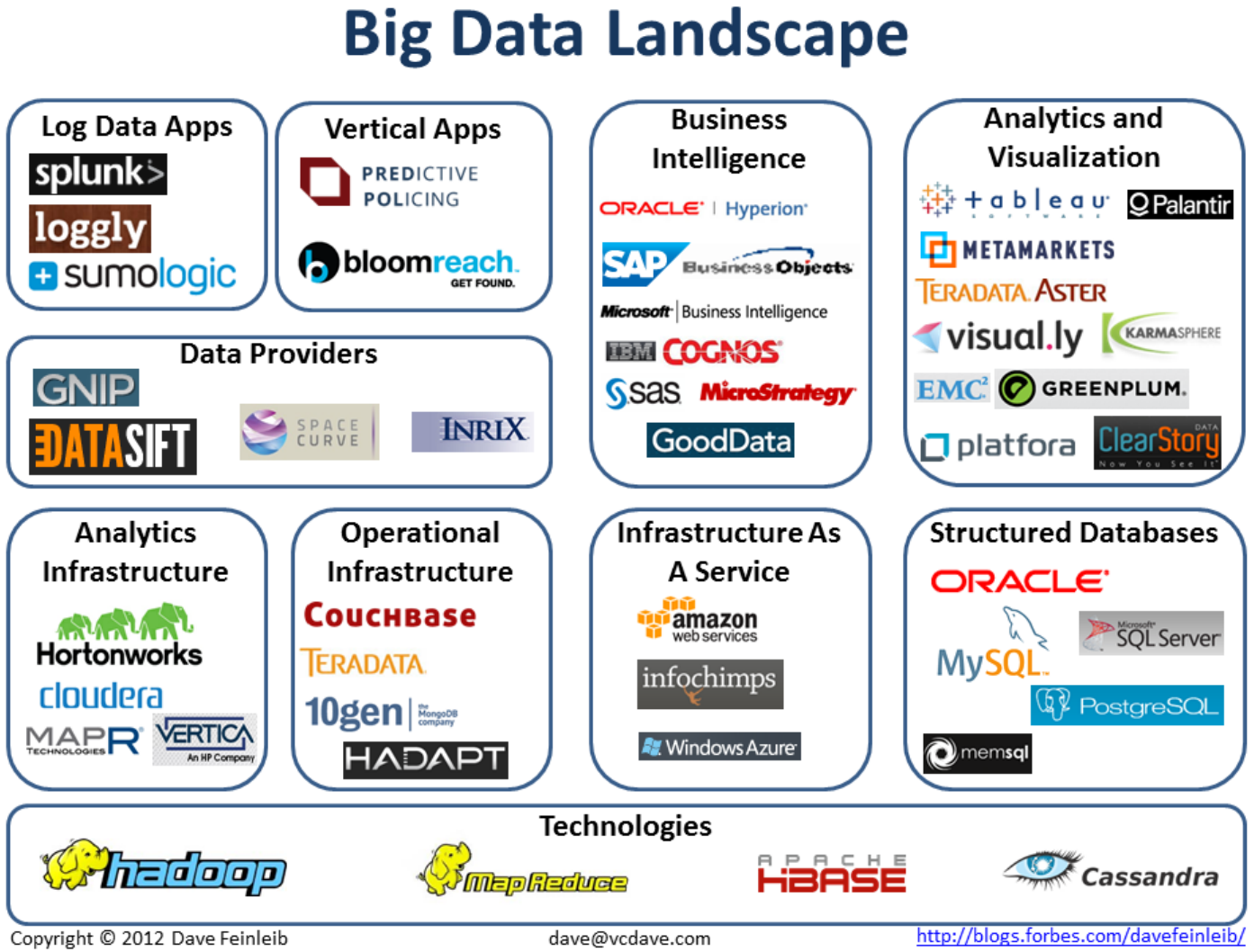
« *Les données scientifiques ne sont pas collectées pour des musées ; elles sont collectées comme une base pour une action. S'il n'y a rien à faire avec des données, il n'y a aucune utilité à les collecter. Le but ultime de la collecte des données est de fournir une base pour l'action, ou une recommandation pour une action. L'étape intermédiaire entre la collecte des données et l'action est la prédiction.* »

Les technologies du « Big Data»

- Qui dit nouveau monde dit nouvelles technologies… et inversement.

- Deux familles de technologies ont facilité la croissance du Big Data :
  - **Les technologies de stockage**
    - Portée par le déploiement du Cloud Computing.
  - **Les technologies de traitement ajustées**
    - le développement de nouvelles bases de données adaptées aux données non-structurées (Hadoop)
    - la mise au point de modes de calcul à haute performance (MapReduce).

Les technologies du « Big Data»



Les technologies de traitement du « Big Data»

Le « Big Data » quand

- les données ne peuvent être stockées sur un seul ordinateur
  - données réparties

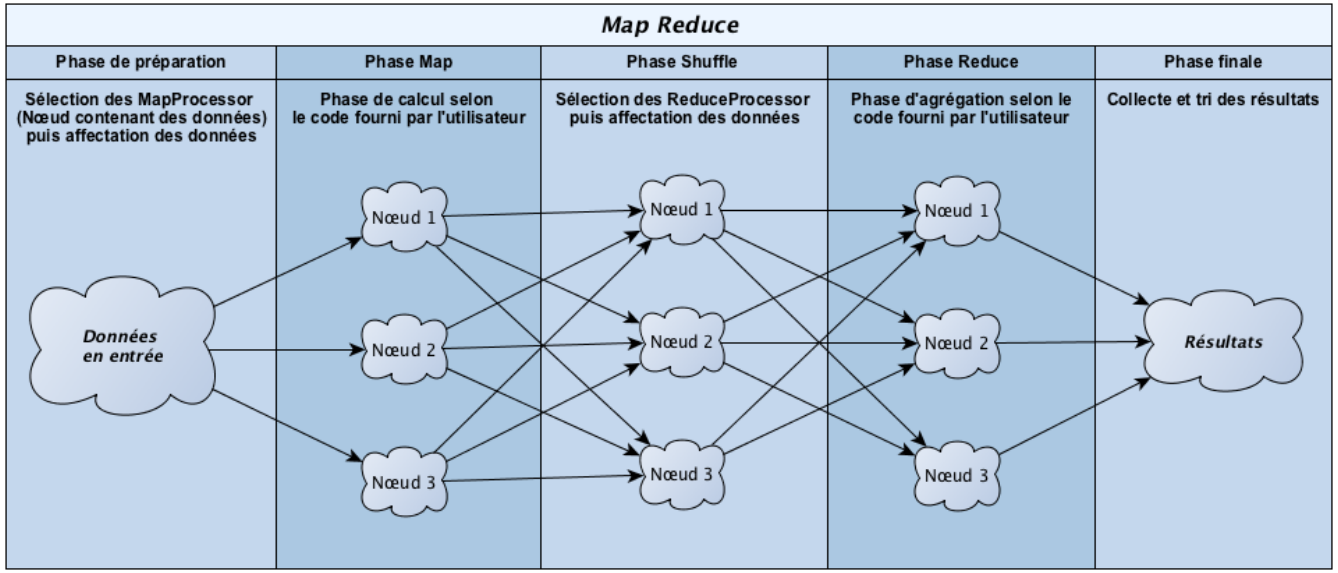
- les traitements vont nécessiter plusieurs machines
  - calculs distribués

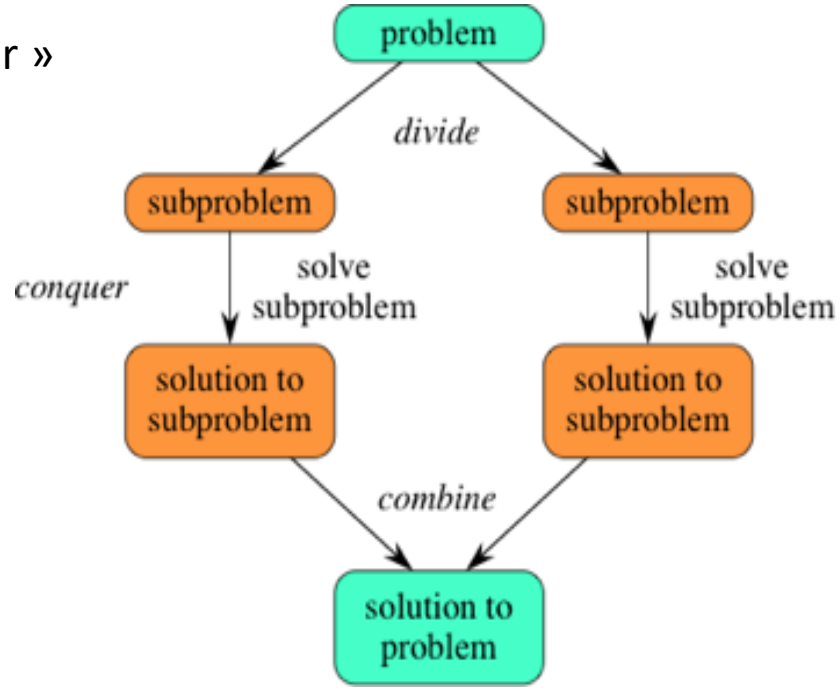
- Technologie de traitement repose sur deux principes :
  - **MapReduce**
  - la **parallélisation**

Les technologies de traitement du « Big Data»

MapReduce (Google 1998)

- Version (très) améliorée d'un algorithme « Divide and Conquer »





Divide and Conquer

Algorithme PGCD (Euclide -300)
Algorithme de tri (Babylone -220)
Transformée de Fourier Rapide (Gauss 1805)
…

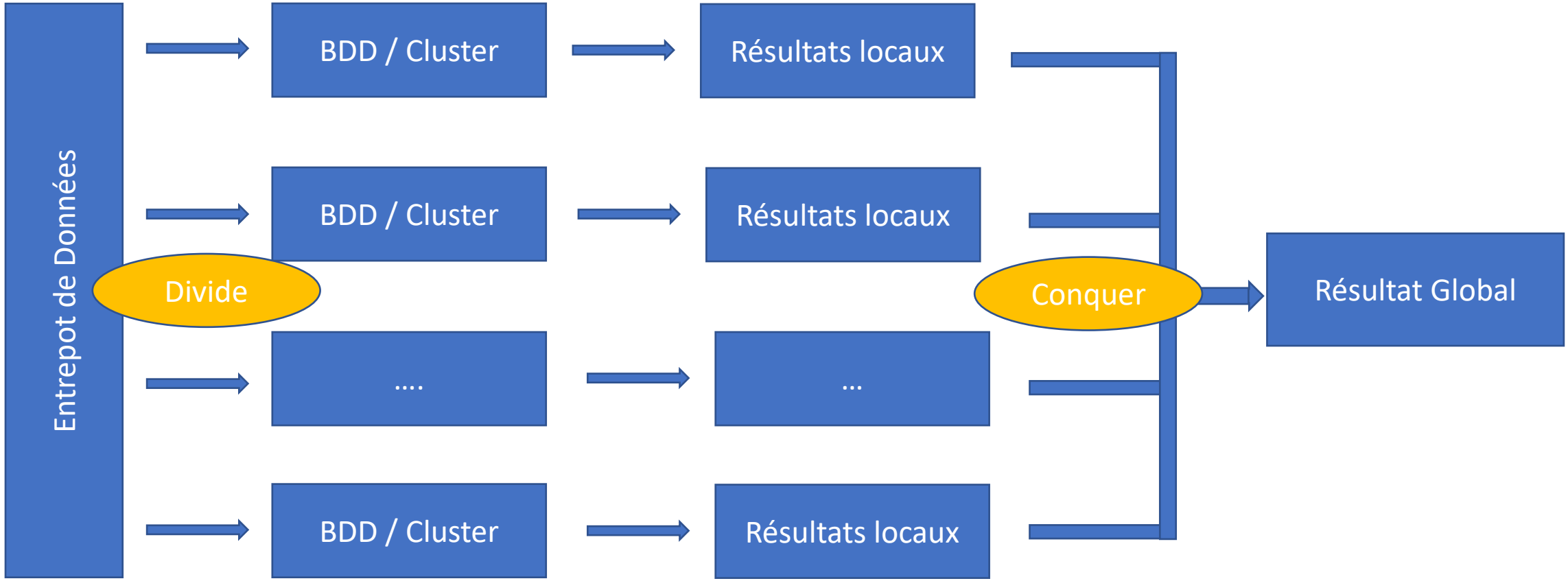
Les technologies de traitement du « Big Data»

Parallélisation

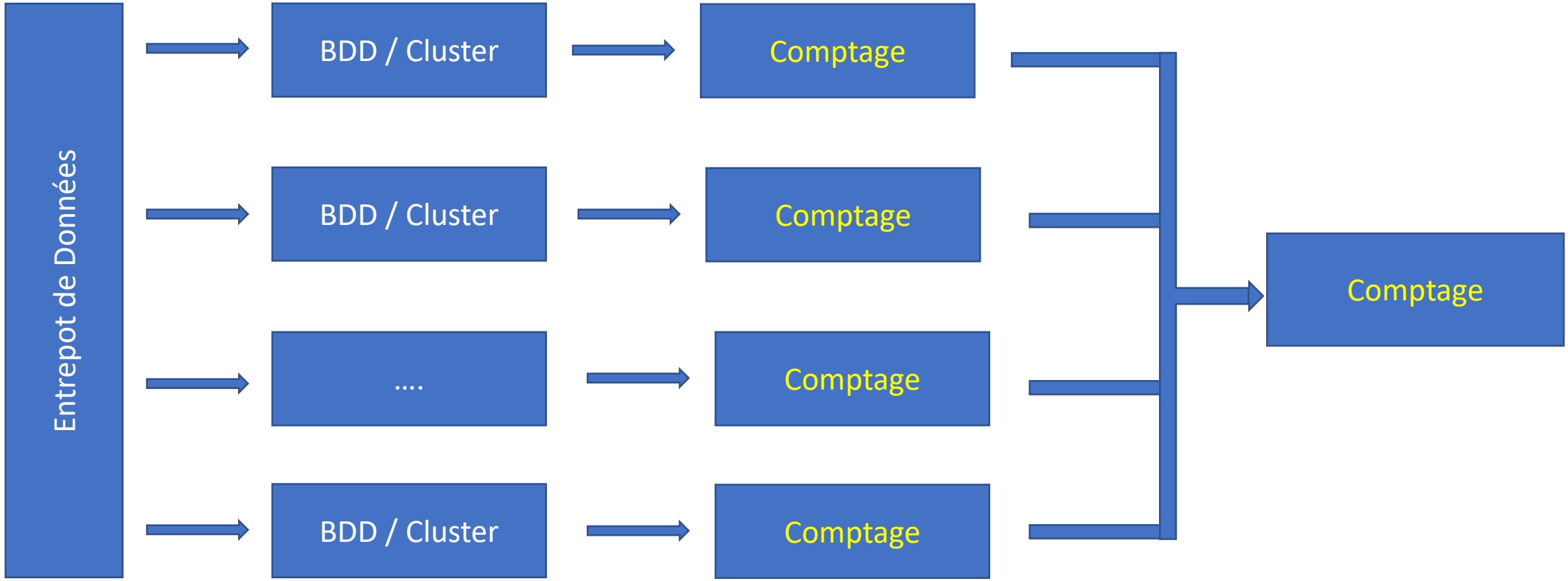
- Traitement des informations de manière simultanée
- Utilisation des différents processeurs d'un ordinateur

CalMip - Olympe - 13 464 cœurs
1,3 Pétaflops
1 365 000 000 000 000 opérations par seconde

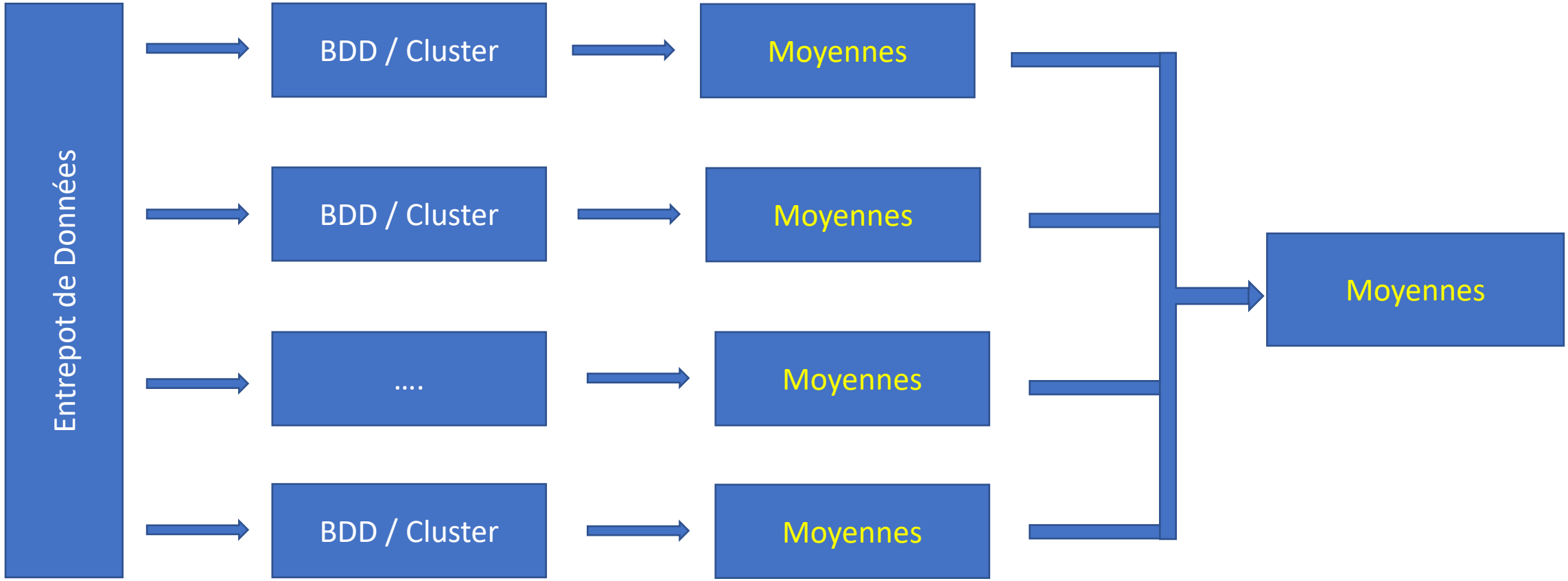
Les technologies de traitement du « Big Data»



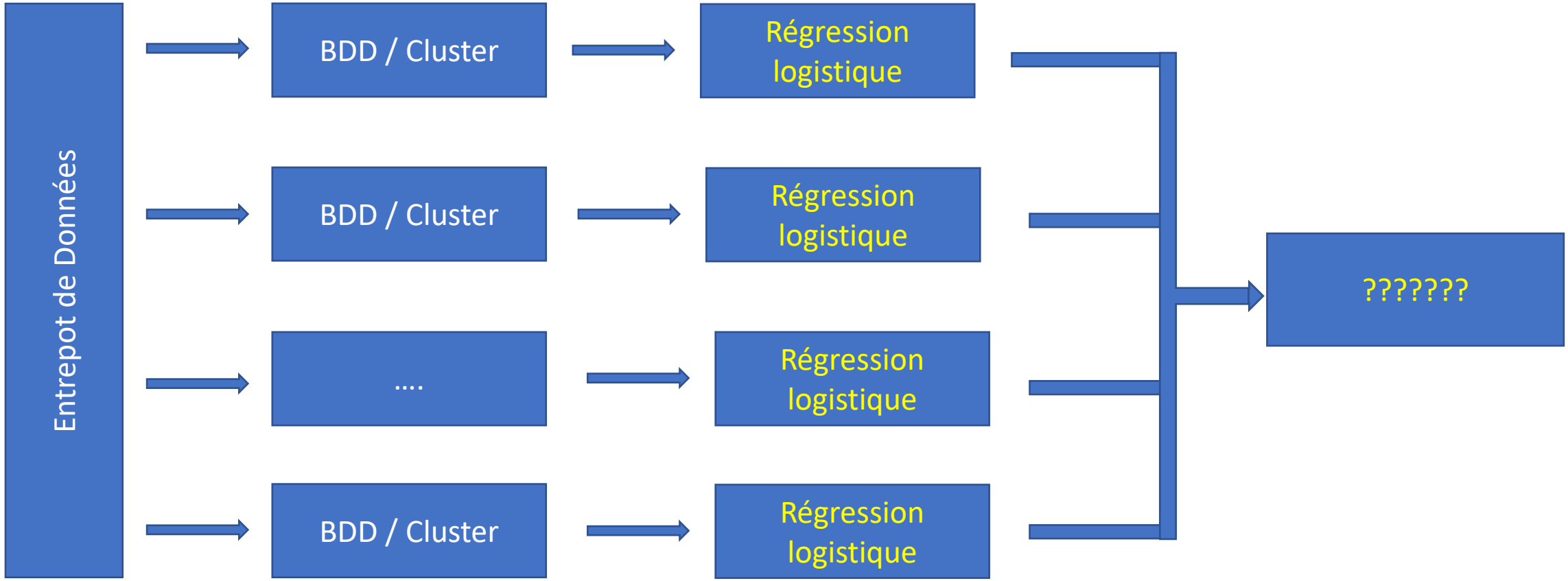
Les technologies de traitement du « Big Data»



Les technologies de traitement du « Big Data»



Les technologies de traitement du « Big Data»



Modèles et « Big Data»

- **Parallélisation**

  Dans le contexte « Big Data » on ne peut pas utiliser tout l'arsenal statistique usuel.

- **Données massive (Volume)**

  Tout est significatif !!!

  - Un modèle de régression même déplorable aura un R^2 « significatif »
  - La plupart des modèles classiques sont rejetés puisque le moindre écart devient significatif
  - Intervalle de confiance réduit à un singleton

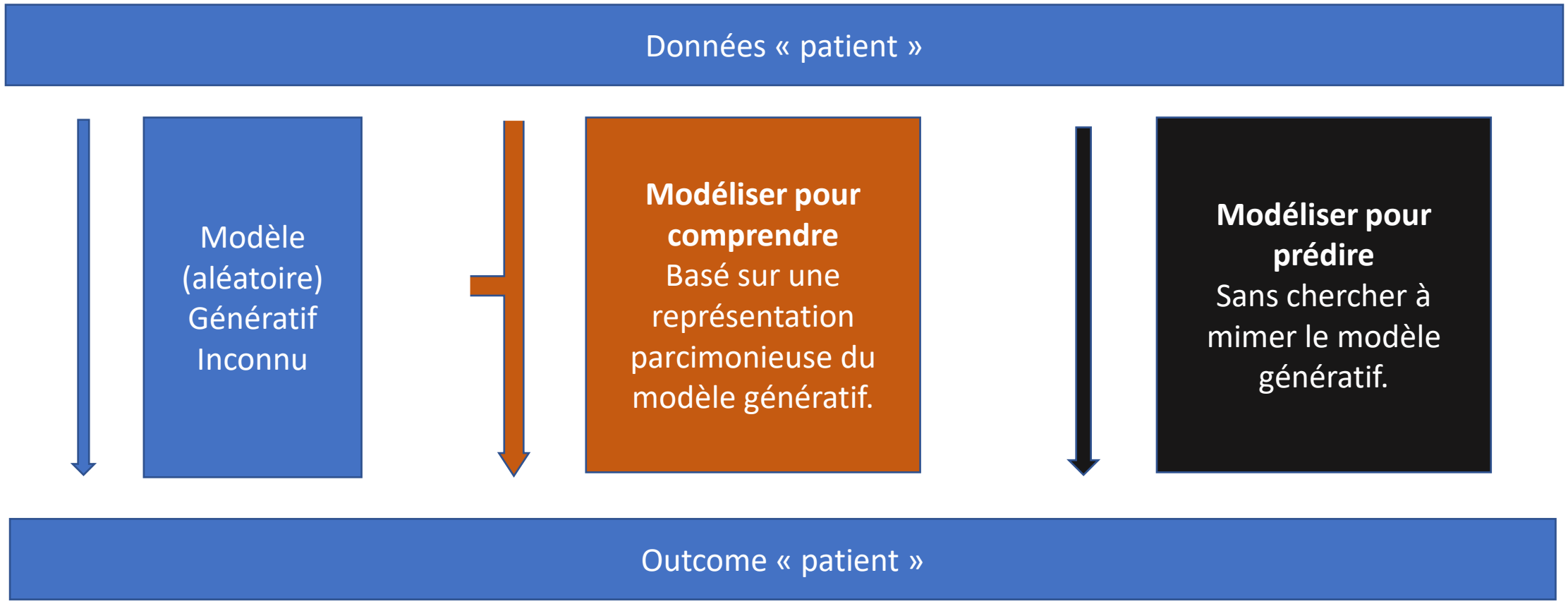
  - Si $n=10^6$ un coefficient de corrélation égal à 0,002 est significativement différent de 0 !!!

  - Pas de question d'inférence statistique dans ce contexte

  - Pas de modèle pour comprendre dans ce contexte…

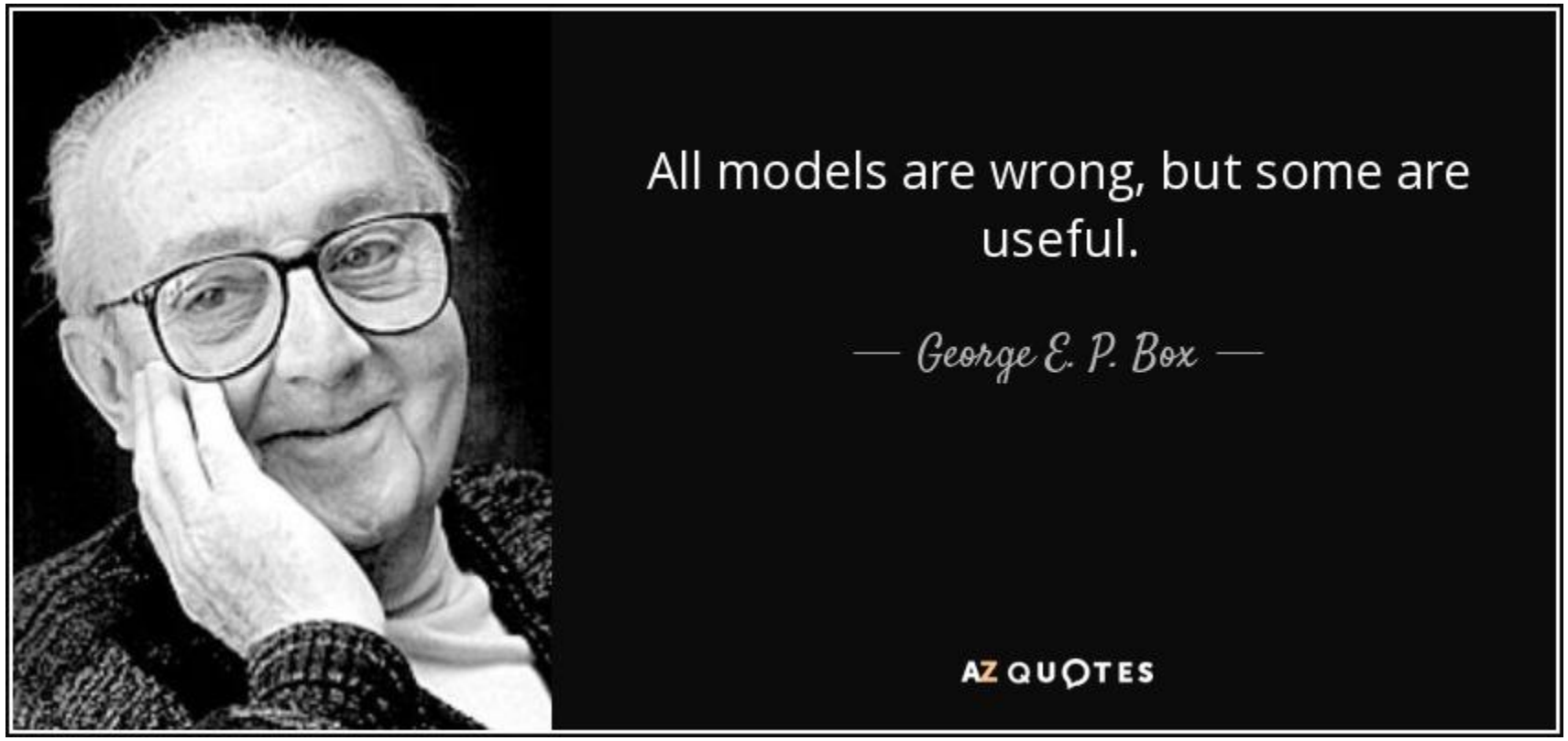
Gilbert Saporta – Quelle statistique pour le Big Data ? - 2017

Modèles et « Big Data»

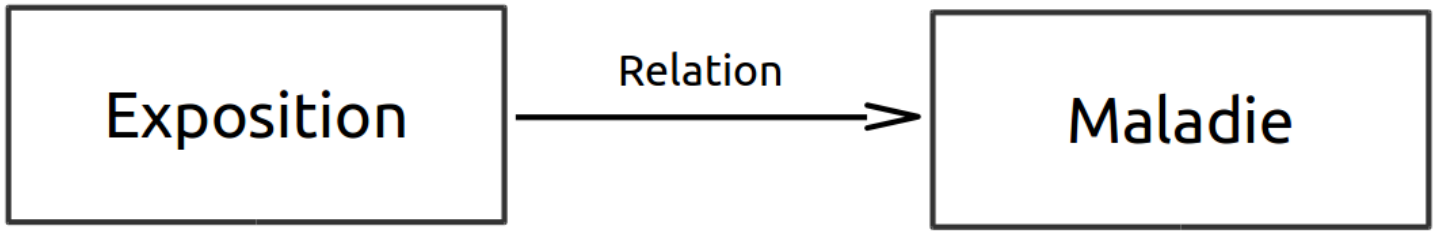


- Un bon modèle ne donne pas nécessairement des prédictions précises
- Un bon modèle de prédiction n'est pas forcément un bon modèle statistique

Modèles et « Big Data»



Modéliser pour Comprendre



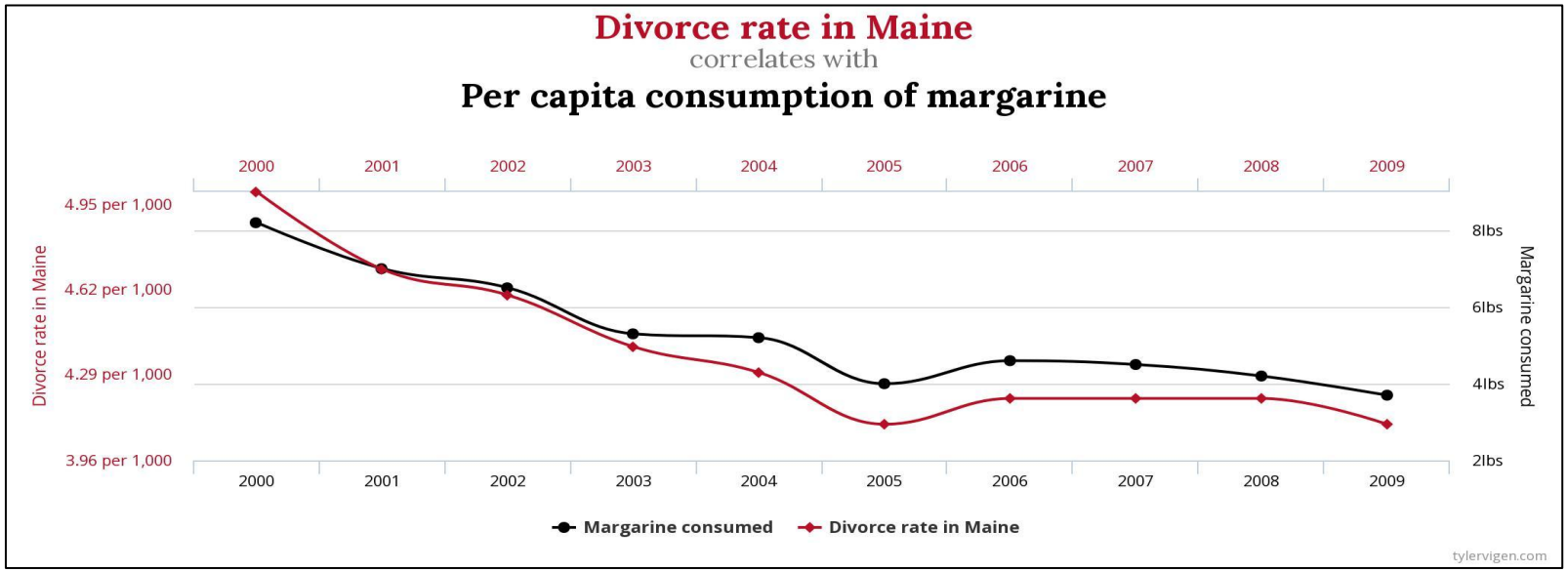
Médicament
Traitement
Intervention chirurgical
Dispositif médical
Outil diagnostique
Outil pronostique
Variable d'exposition
...

Innocuité
Efficacité
Qualité de vie
Guérison /
Complication
Douleur
Diagnostic / Pronostic
Evolution de la maladie
…

Objectif : Montrer que la relation est **causale**.

Modéliser pour Comprendre

- **Mise en évidence d'une relation causale ?**
  - Approche Statistique ne suffit pas
  - Mise en évidence d'une **association**



Correlation : 99.26 %

http://tylervigen.com/spurious-correlations

Modéliser pour Comprendre

- **Principe de la causalité** - Platon
  - « Sans l'intervention d'une cause, rien ne peut être engendré »

- Les **critères de Sir Bradford-Hill (1965)**.
  - Groupe de conditions minimales pour fournir une preuve adéquate d'une relation causale entre deux évènements

- **Evidence-Based Medicine** (Evaluation basée sur des Essais cliniques)

- **Méthodologie** précisée dans un protocole spécifiant :
  - Mode de recueil des données
  - Déroulement de la recherche
  - Méthode d'analyse des données

Modéliser pour Comprendre

- Recherche de la parcimonie
  - Rasoir d'Ockham :

    Pluralitas non est ponenda sine necessitate
    (les multiples ne doivent pas être utilisés sans nécessité)

  - Sélection des variables (régularisation)
    - LASSO, Ridge, ElasticNet
- Paramètres interprétables en termes cliniques
- Représentativité de l'échantillon d'étude

Guillaume d'Ockham
(v. 1285 - 9 avril 1347)

Grandes dimensions (p grand, n grand)

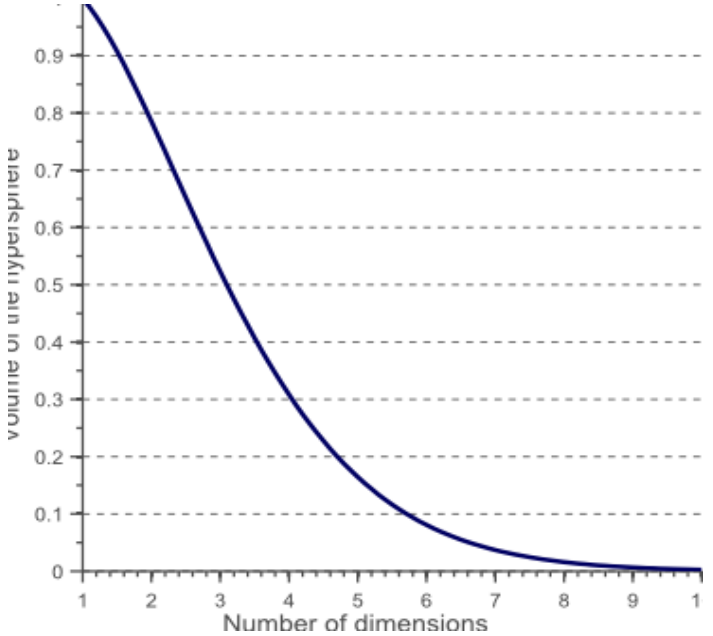
- Aucun modèle simple ne peut représenter de grandes masses de données.

- Représentativité : Fléau de la dimension
  Concept introduit en 1961 par Bellman

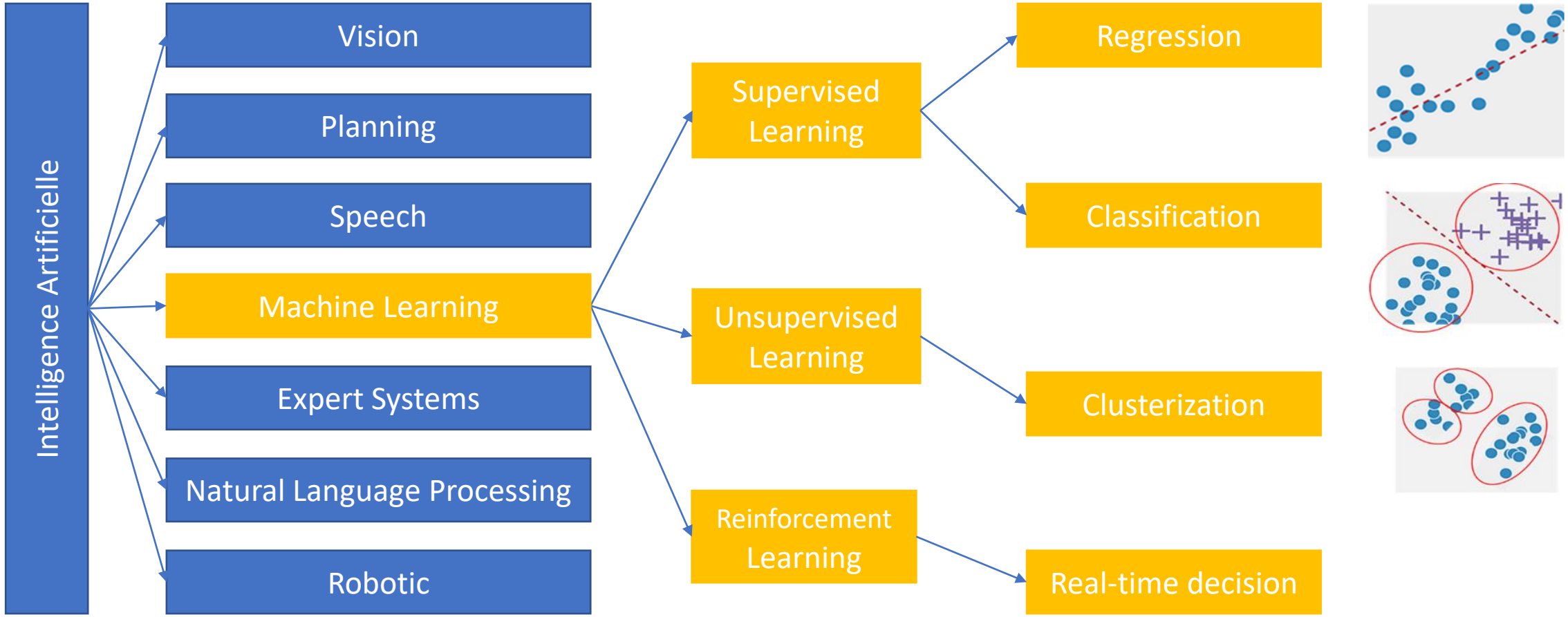
Dans un espace à « grande dimension », tous les individus sont des « outliers ».

Exemple:

Pour réaliser dans $[0,1]^{10}$ une couverture équivalente à celle des 100 points dans [0,1], il faut 10^{20} observations



Intelligence artificielle et Machine learning



Modéliser pour prédire

- Il en faut un peu plus pour prédire que pour comprendre…

- Capacité prédictive sur de nouvelles observations
- **Validation interne** (génération de nouveaux individus à partir des données de travail)
  - Validation Croisée
  - Bootstrap

- **Validation Externe** (constitution d'un jeu de données issu de la même population)

- **Data splitting**
  - Particulièrement adapté aux données massives
  - Attention à la représentativité de l'échantillon…

Modéliser pour prédire

Dilemme Biais Variance

Biais :
Erreur « systématique » de prédiction.

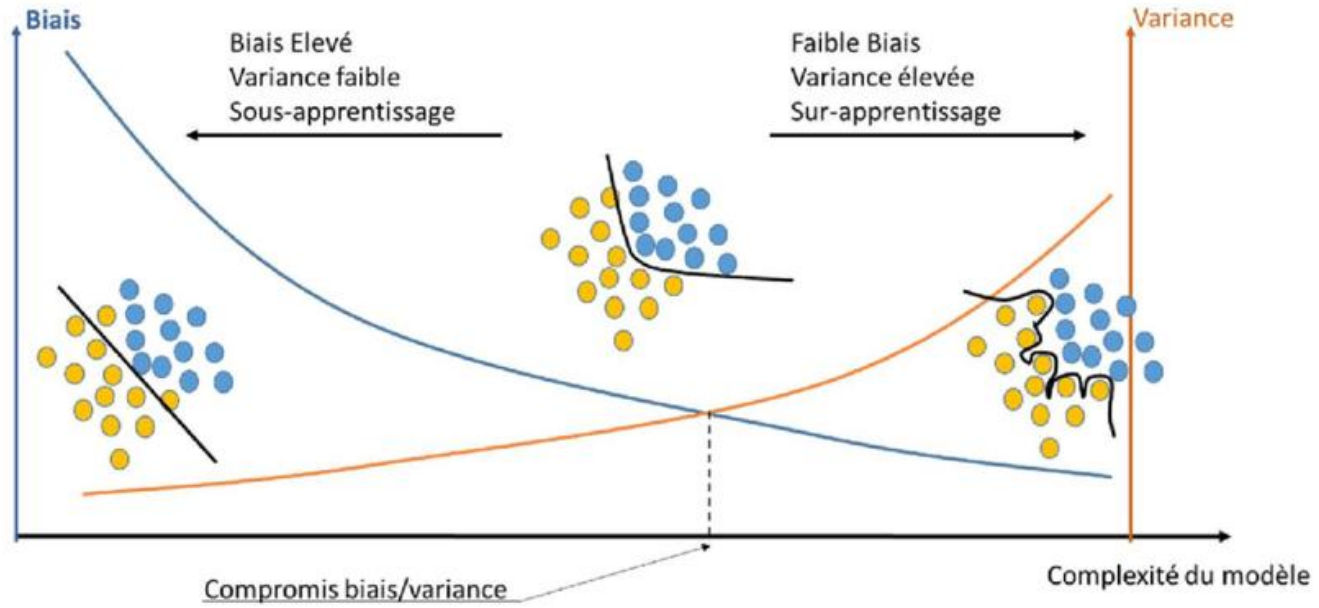
Le modèle n'est pas assez précis pour capturer la diversité des valeurs.

Sous-apprentissage

Variance :
Erreur dûe à la sensibilité du modèle.

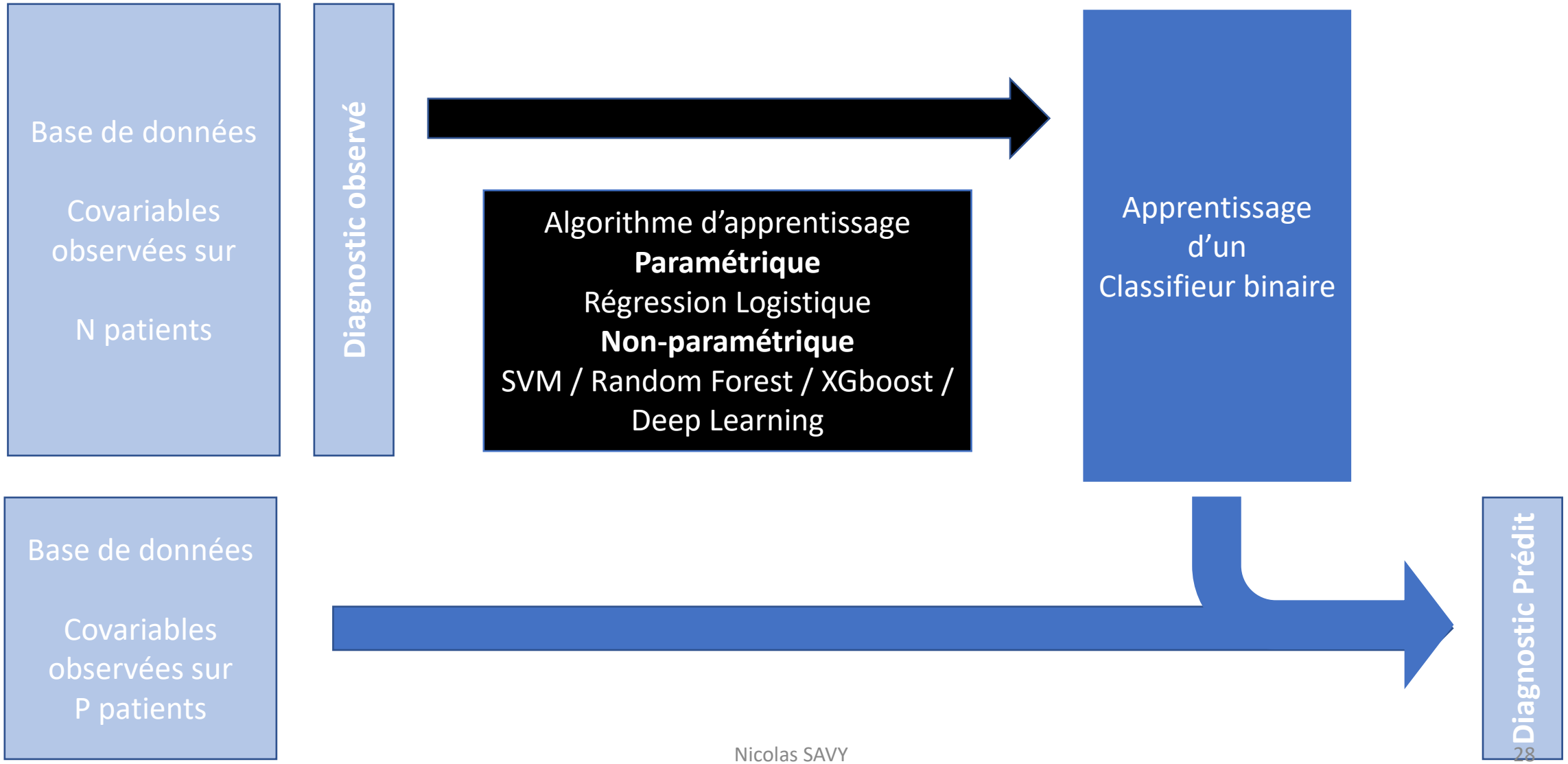
Le modèle est trop précis sur les données se comporte de manière instable sur de nouvelles données.

Sur-apprentissage



P. Scalart – Univ. Rennes I

Modéliser pour prédire



Modéliser pour prédire

- Les modèles « usuels » ne sont pas loin… **approche paramétrique**
- Modélisation de la probabilité d'avoir un diagnostic positif conditionnellement aux observations des covariables:

$$P[D = 1|(X_1, X_2, \dots, X_k)] = g(\alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k)$$

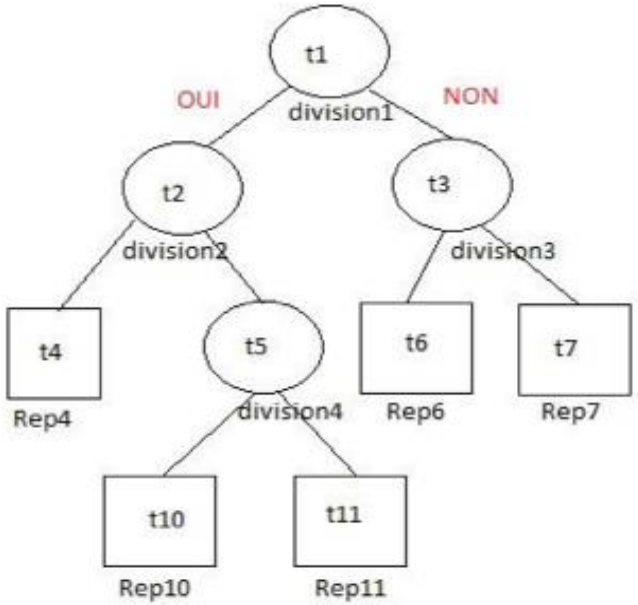
- **Estimation** des paramètres $(\alpha_1, \alpha_2, \dots, \alpha_k)$
- Sous de bonnes **hypothèses** (linéarité), il existe un lien entre le coefficient α_i et l'influence de la variables X_i sur le diagnostic : **risque relatif**
- Un score peut être construit à partir des coefficients

Modéliser pour prédire

Approche Non-paramétrique

Arbre de décision : Algorithme CART

(Breiman, Friedman, Olshen et Stone (1984)



- Comment définir les divisions successives ?
- Quand arrêter le principe de division ?
- Comment définir les réponses ?

- Très peu robuste (à l'ordre notamment)
- Sur apprentissage
- Algorithme très gourmand temps calcul
- Résultats biaisés (variables continues notamment)

➢ **Extensions**

- Random forest (Breiman, 2001)
- Bagging (**B**ootstrap **agg**regat**ing**) (Breiman, 1994)
- XGBoost algorithme associé au Bagging
- …

Modéliser pour prédire

Apprentissage « Paramétrique »

Classifieur est un modèle paramétrique

Classifieur est déterministe

Paramètres ont une interprétation clinique (Risque relatif)

Modèle repose sur des hypothèses (linéarité du risque)

Estimation des paramètres demande moins d'observations

Lien covariable – diagnostic s'exprime en termes de probabilités

On en déduit un score

Apprentissage « Non-Paramétrique »

Le classifieur est un algorithme

Classifieur est (ou peut-être) aléatoire

Pas de paramètres donc pas d'interprétation clinique

Pas d'hypothèse (données explorent les possibilités)

Qualité du modèle directement liée au nombre d'observations

Juste un classement de l'importance des covariables dans le diagnostic

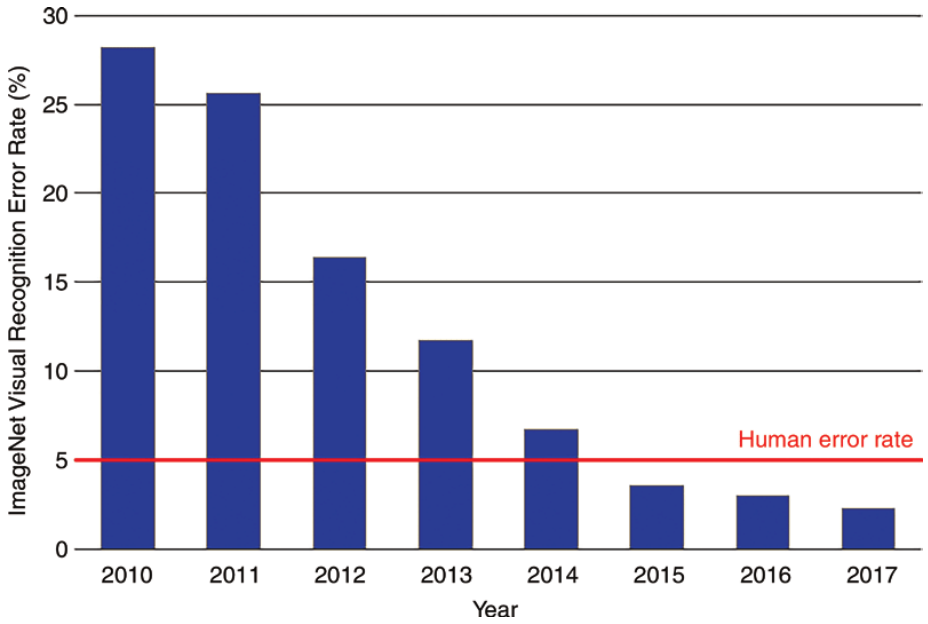
On en déduit une prédiction

Modéliser pour prédire

Deep Learning

Algorithme **« star »** du Machine Learning

- Résultats remarquables notamment en reconnaissance d'images
- « ImageNet »
  - Large Scale Visual Recognition Challenge
  - Nb de vignettes annotés : 14,000,000
  - AlexNet (2012) used Deep Learning
  - Nb de paramètres : 62,378,344



- Problème méthodologique
  - Base américaine
  - Résultats transposables à une autre population ?

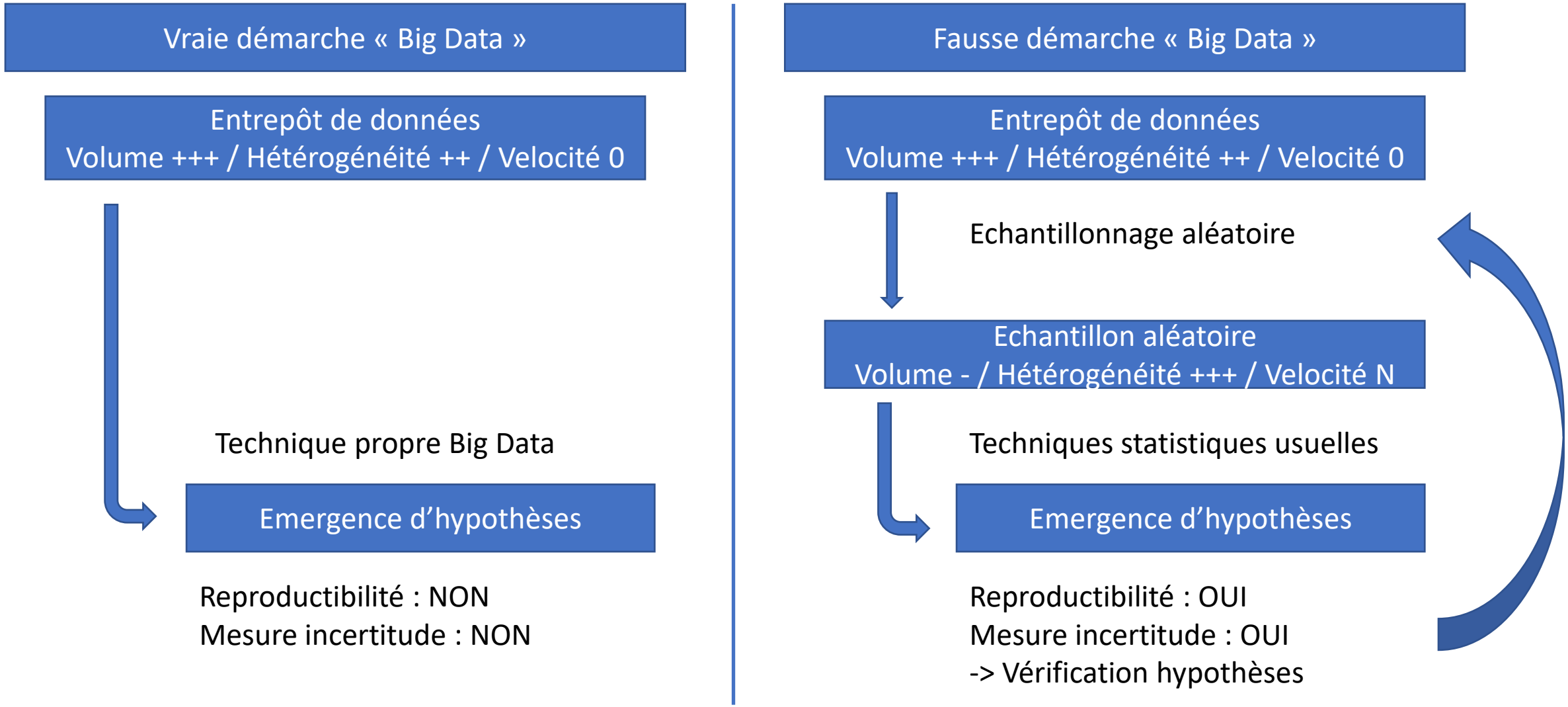
Modéliser pour prédire

Adversary example

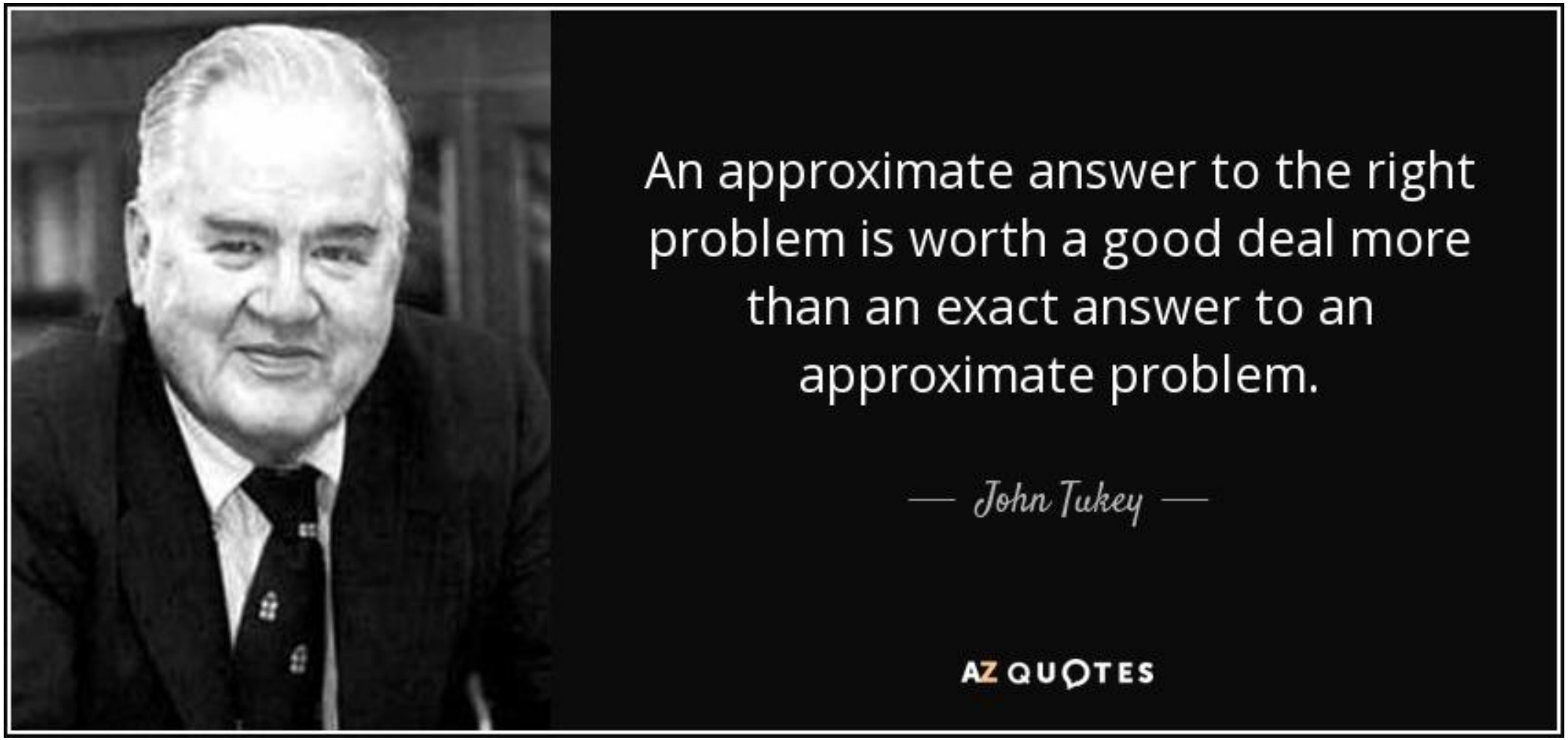
Très difficile d'identifier une « attaque » car non-linéarité

Les Algorithmes peuvent être « piratés »

Modèles et « Big Data»



Tukey


An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.
— John Tukey —
AZ QUOTES

Cinq exemples d'« innovations »…

« Innovations » pour le patient …

Passage de la **Médecine « Factuelle »** à la **Médecine « 6P »**

- **Personnalisée**
- **Préventive**
- **Prédictive**
- **Participative**
- **Preuve (service médical rendu)**
- **Parcours**

N.B. : Le volume (en terme de nombre d'individus)
Grande précision pour la prévision de **comportements moyens** (loi des grands nombres)
Part irréductible d'aléa reste attachée à la prévision d'un comportement individuel

Heureusement …

« Innovations » pour le praticien …

- Performances des outils de diagnostic

DeepMind de Google a créé une IA qui détecterait le cancer du sein mieux que les radiologues humains

Business Insider 2 Jan 2020, 16:11 Tech 1 799

- Résultats dépendent de la base d'apprentissage

Quand le logiciel de recrutement d'Amazon discrimine les femmes

Par **Les Echos**

Publié le 13 oct. 2018 à 12h04 | Mis à jour le 13 oct. 2018 à 12h34

« Innovations » pour la société …

- Identification de « mécanismes » pertinents

Powerful antibiotic discovered using machine learning for first time

Team at MIT says halicin kills some of the world's most dangerous strains

The Guardian – 20/02/2020

- Identification de « mécanismes » alarmants

PIMPON – Remonter aux prescripteurs des alertes pour les interactions médicamenteuses dangereuses
Projet lauréat de l'AO 2019 « Health Data Hub »
Le projet vise à mobiliser les données du SNDS pour estimer la prévalence réelle des complications liées aux interactions médicamenteuses afin d'identifier des alertes nécessitant une mise en valeur particulière du fait de leur impact.

« Innovations » pour l'industriel …

- Utilisation des données « Real-Word » pour la documentation produit

BRIEF

Pfizer wins expanded Ibrance approval using real world data

BiopharmaDive – 05/04/2019

BUSINESS | HEALTH CARE | HEALTH

Drugmakers Turn to Data Mining to Avoid Expensive, Lengthy Drug Trials

Pfizer, Johnson & Johnson and Amgen try to win drug approvals by analyzing vast data sets of electronic medical records

THE WALL STREET JOURNAL.

English Edition ▾ | February 27, 2020 | Print Edition | Video

- Utilisation des données « Real-Word » pour l'optimisation de design expérimentaux

"In Silico Clinical Trials": a way to improve drug development?

Nicolas Savy* Philippe Saint-Pierre† Stéphanie Savy‡ Sylvia Julien§ Emmanuel Pham¶

Proceedings JSM – 15/10/2019

Take Home Message…

Les **données planifiées** doivent rester le standard pour l'utilisation de modèles explicatifs

- Demande une méthodologie finement pensée
- Expertise du Biostatisticien

Les **données massives** nourrissent les modèles d'apprentissage

- Demande une approche spécifique (machine learning)
- Expertise du Data Scientist

A big data-analyst is an expert at producing misleading conclusions from huge datasets.
It is much more efficient to use a statistician, who can do the same with small ones.

Stephen Senn.

Dans tous les cas, nécessaire **évaluation** des outils avec une attention particulière à la méthodologie.

Quand on me présente quelque chose comme un progrès, je me demande avant tout s'il nous rend plus humains ou moins humains…

Georges Orwell

Merci de votre attention

- https://doi.org/10.1016/S0764-4469(00)00153-0