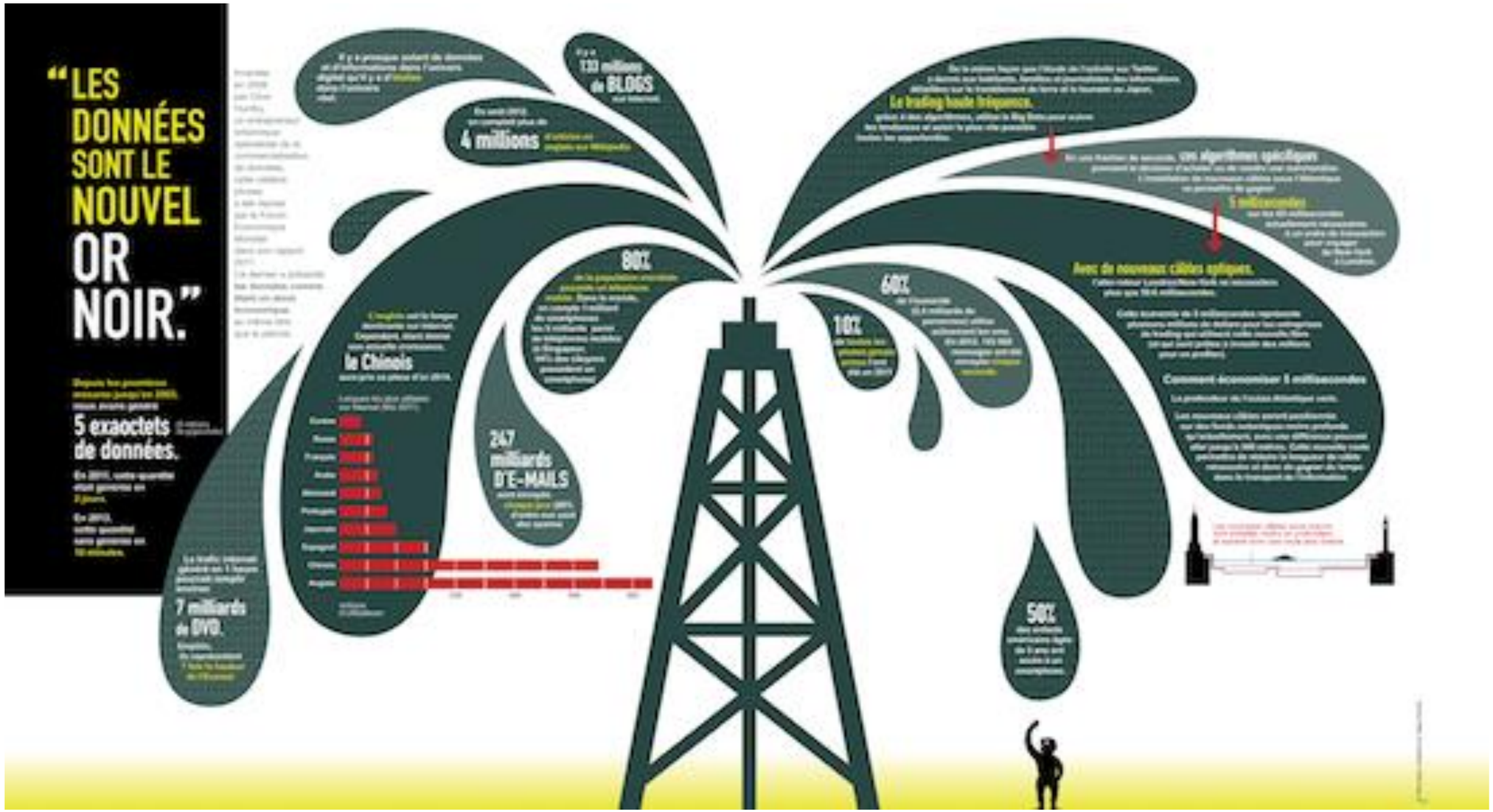
Big Data Analytics

« Pilotage de la performance pour une bonne gouvernance des entreprises »

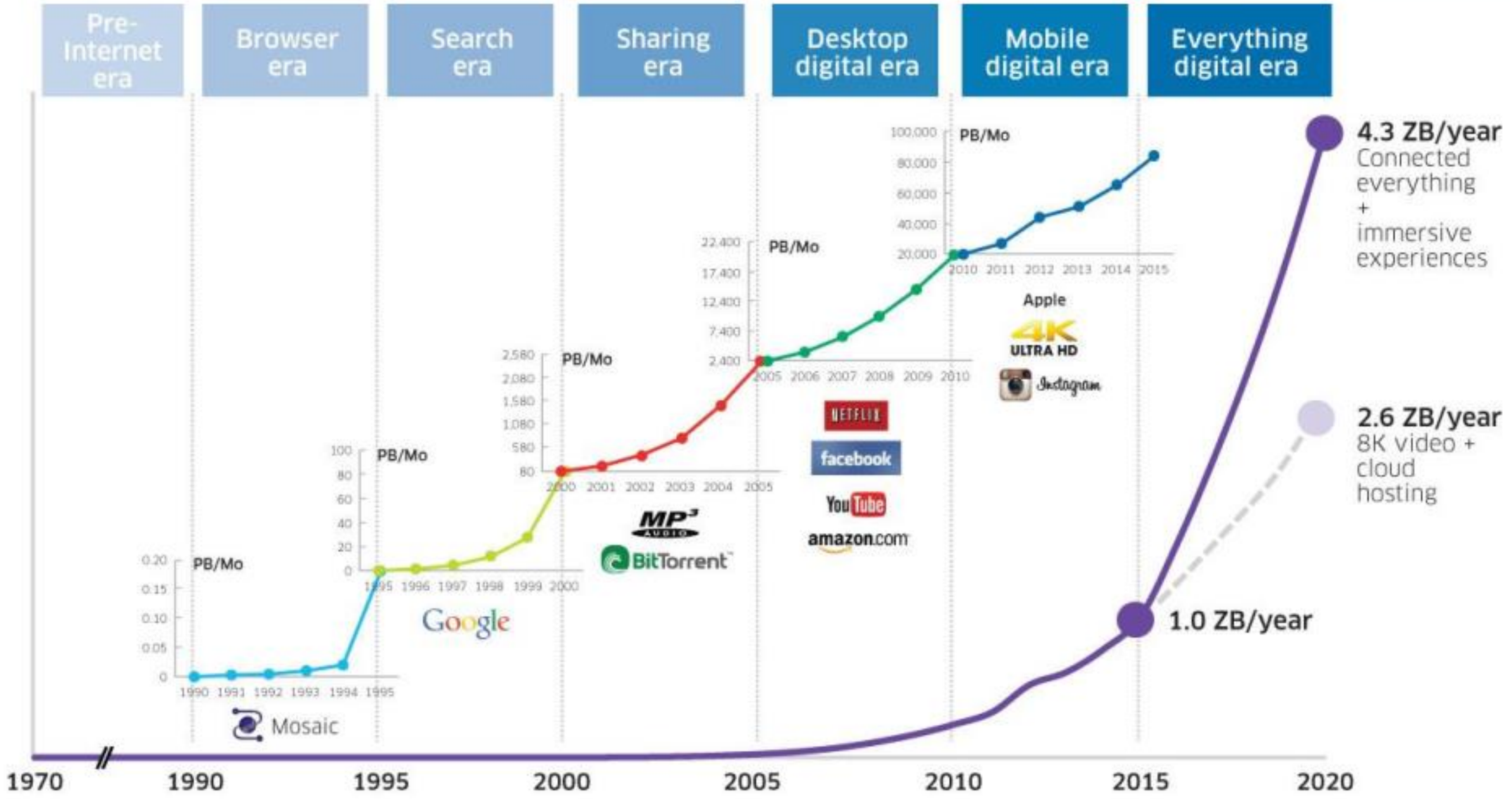
CHAPITRE 1 - Introduction

Les données : nouvel or noir



Les données : nouvel or noir

1 ZB = 10^{21} Bytes



Data transfer projections to 2020 (Source: Bell Labs)

esprit

La science historique des données : la statistique

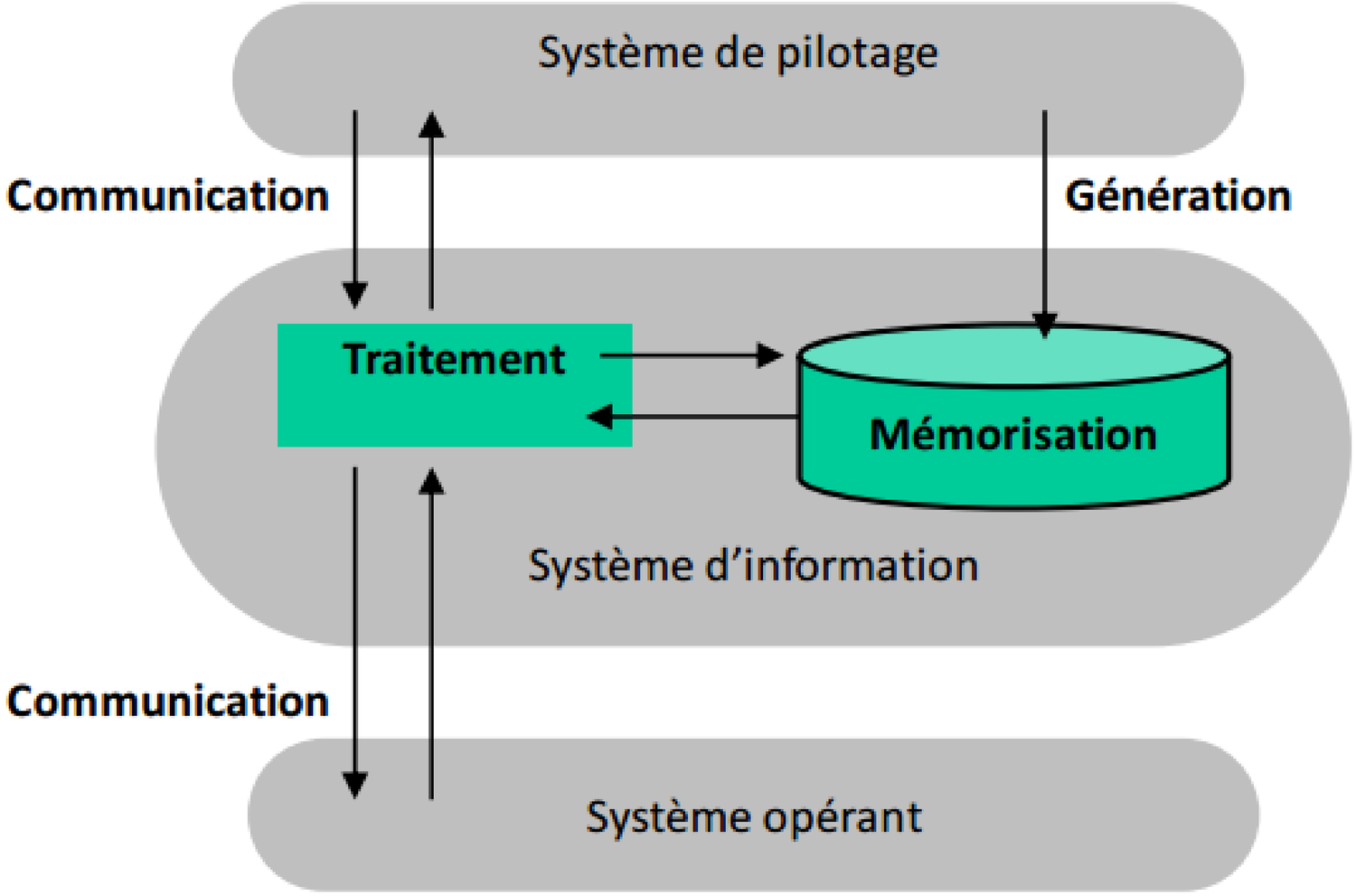
- La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

Source : Wikipedia

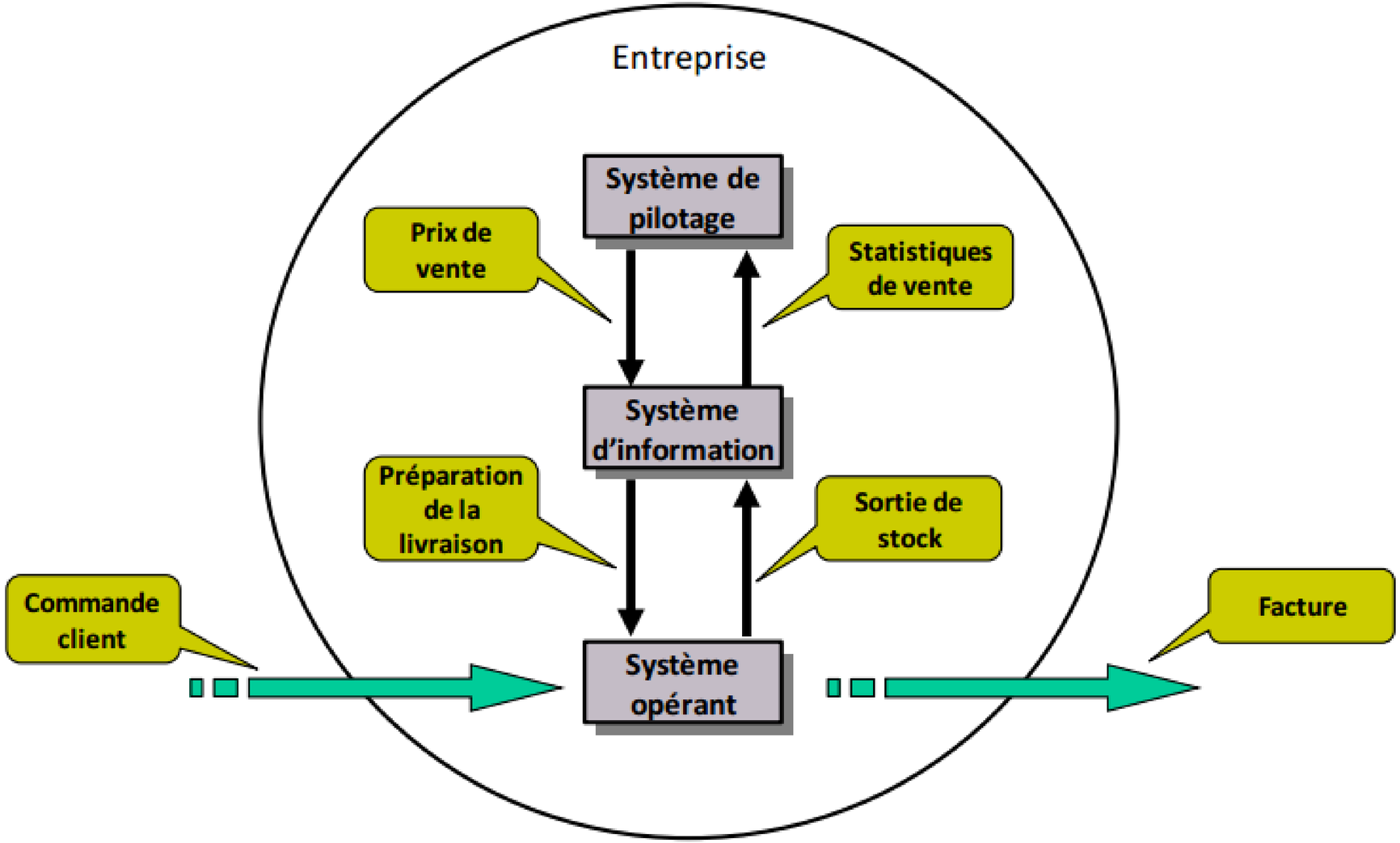
- La statistique est utilisée dans presque tous les domaines de l'activité humaine : sciences sociales, économie, médecine, biologie, industrie ...

- Mais aujourd'hui, Open Data, Big Data, Big Science... changent les règles du jeu.

Analyse systémique d'une entreprise



Analyse systémique d'une entreprise



Qu’est ce que la gouvernance

- La mise en œuvre d'un ensemble de dispositifs (règles, normes, protocoles, conventions, contrats…) pour assurer une meilleure coordination des parties prenantes d'une organisation, chacune détenant une parcelle de pouvoir, afin de prendre des décisions consensuelles et de lancer des actions concertées

PILOTER.ORG
Performance
Management Décision

« Good governance is the single most important way to end poverty and support development. »

Kofi Annan

esprit

Qu’est ce que la gouvernance des entreprises ?

- Le gouvernement (ou gouvernance d'entreprise - *« corporate governance »*) d'entreprise désigne le système formé par l'ensemble des processus, réglementations, lois et institutions destinés à cadrer la manière dont l'entreprise est dirigée, administrée et contrôlée.

- En fonction des objectifs qui gouvernent l'entreprise, ce système est appelé à réguler les relations entre les nombreux acteurs impliqués ou parties prenantes (*« stakeholders »*).

Haidar, Jamal Ibrahim, "Investor protections and economic growth"
Economics Letters, Elsevier

Bonne gouvernance

Les 4 éléments d’une bonne gouvernance

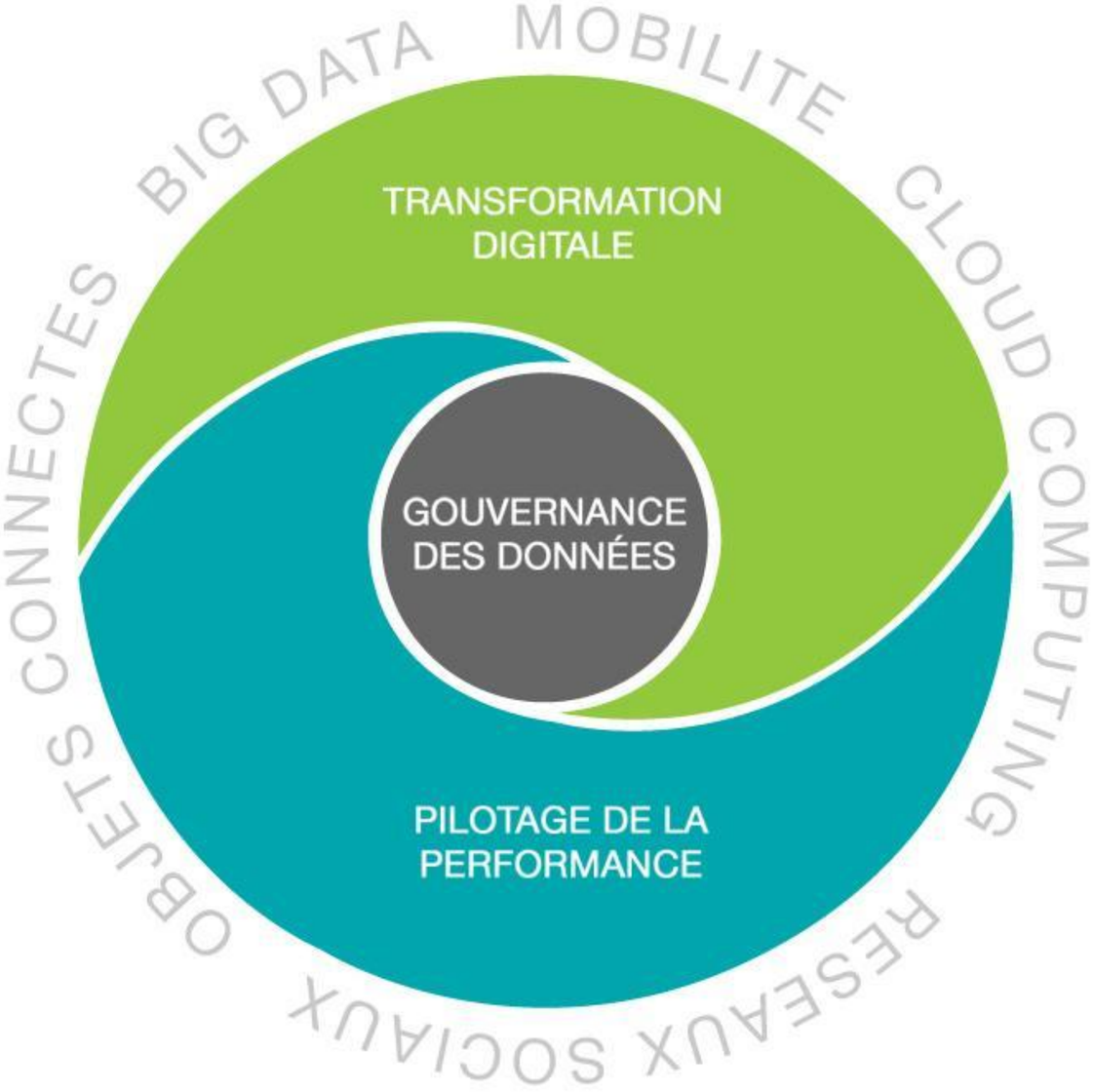
Comprendre

Tenir à jour

Protéger

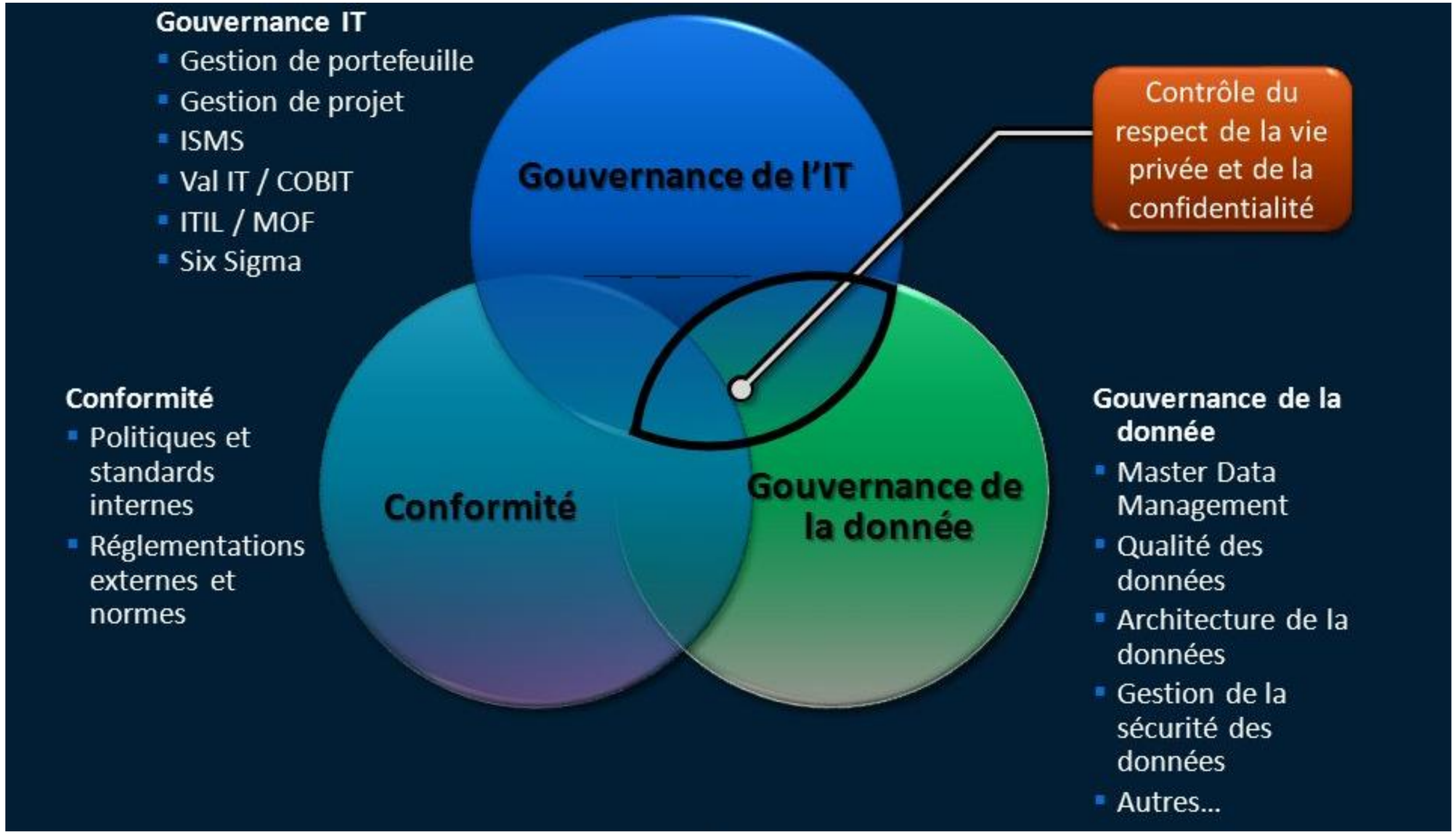
Intégrer

Gouvernance des entreprise et transformation digitale



Gouvernance IT, conformité et gouvernance des données

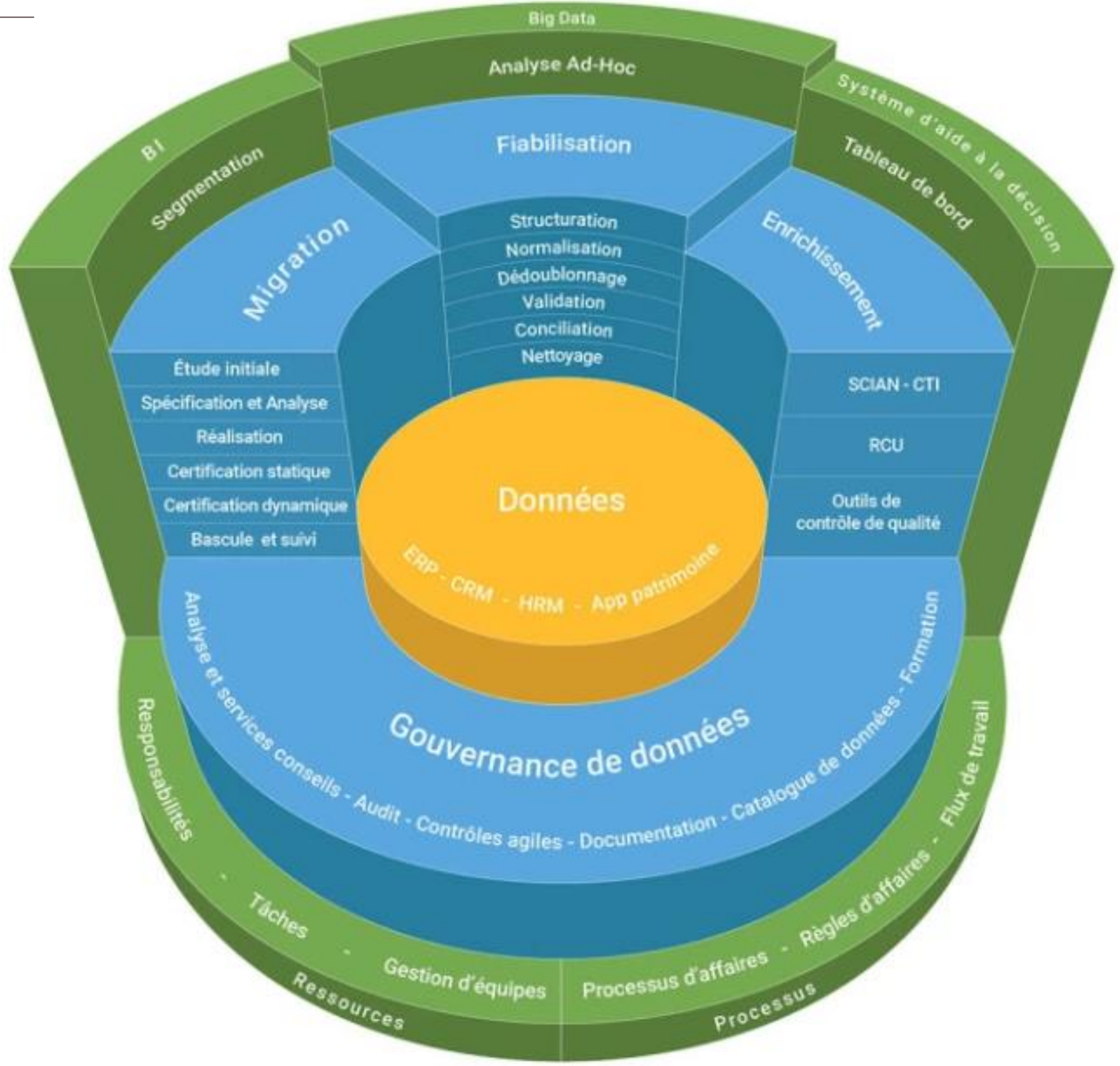
Microsoft



esprit

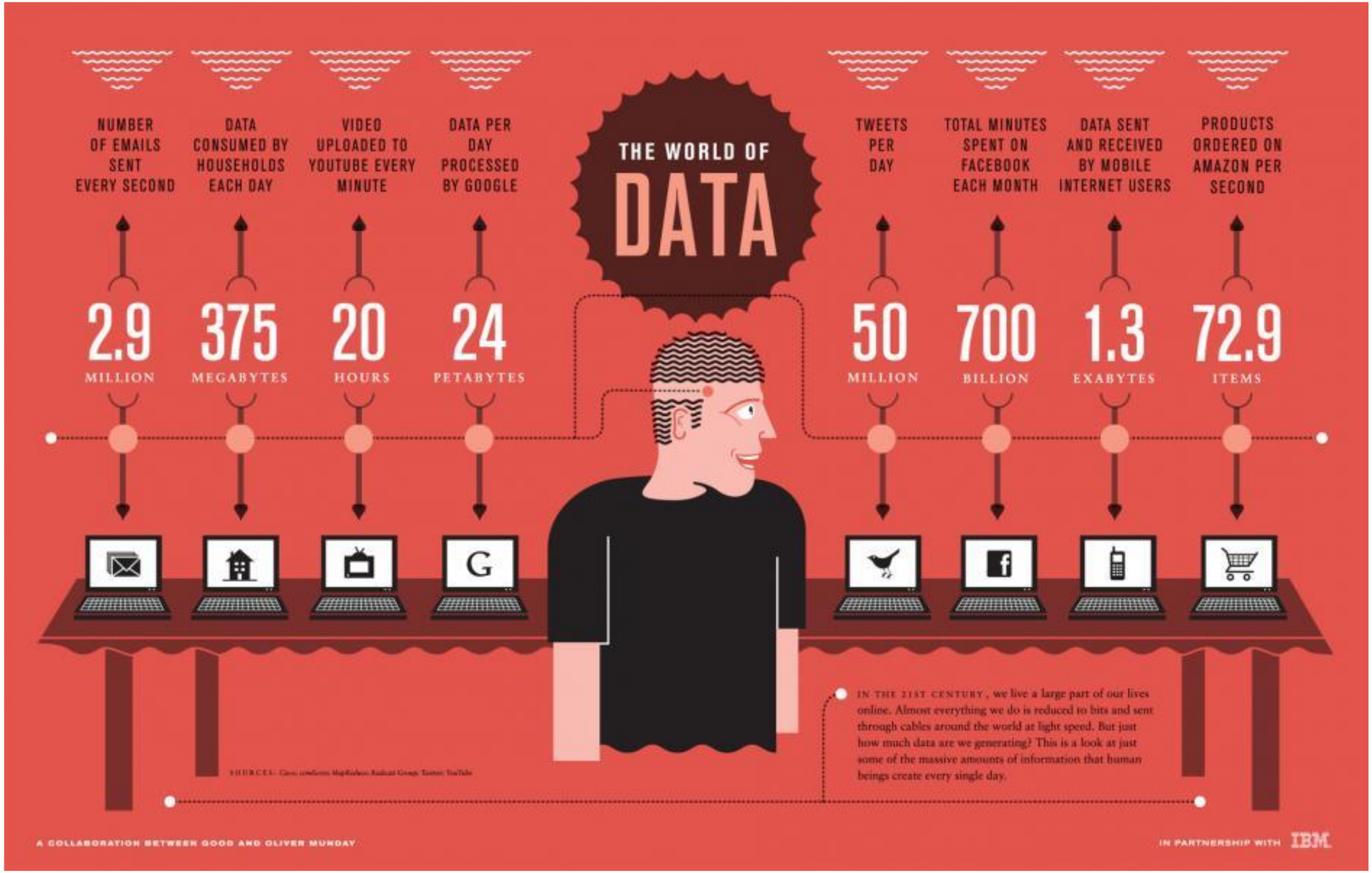
Gouvernance de données

Maîtrisez-vous vos données ?



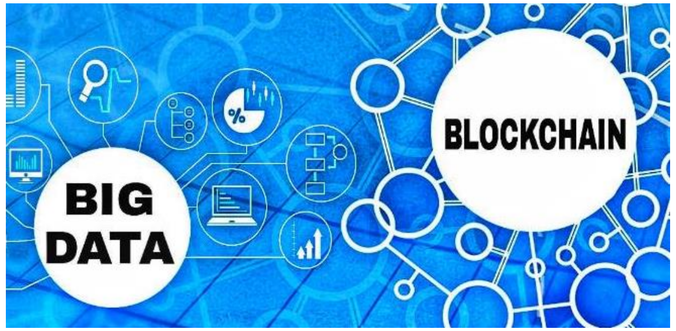
esprit

Gouvernance de données


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.
A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY
IN PARTNERSHIP WITH IBM

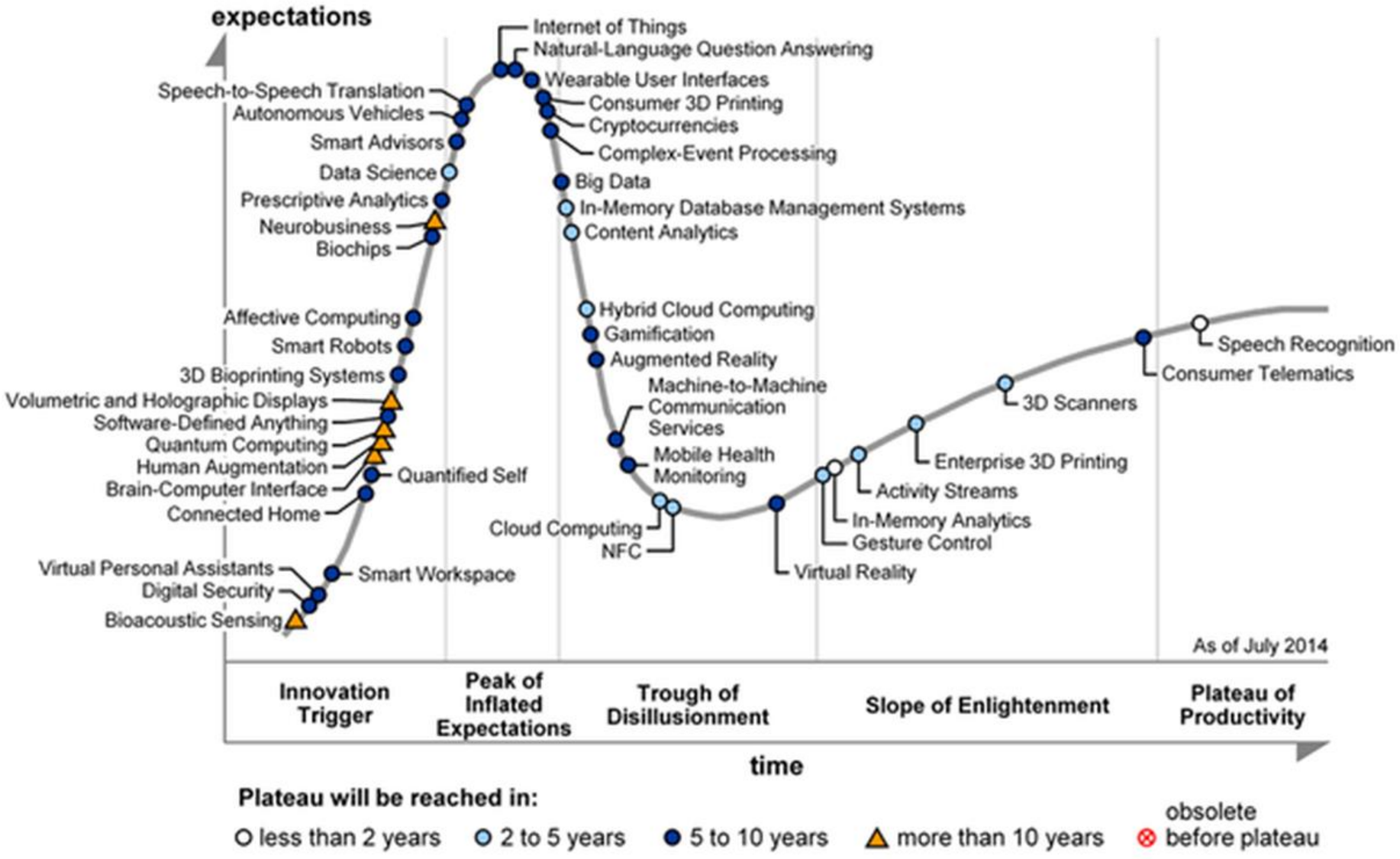
esprit

Big Data Landscape



esprit

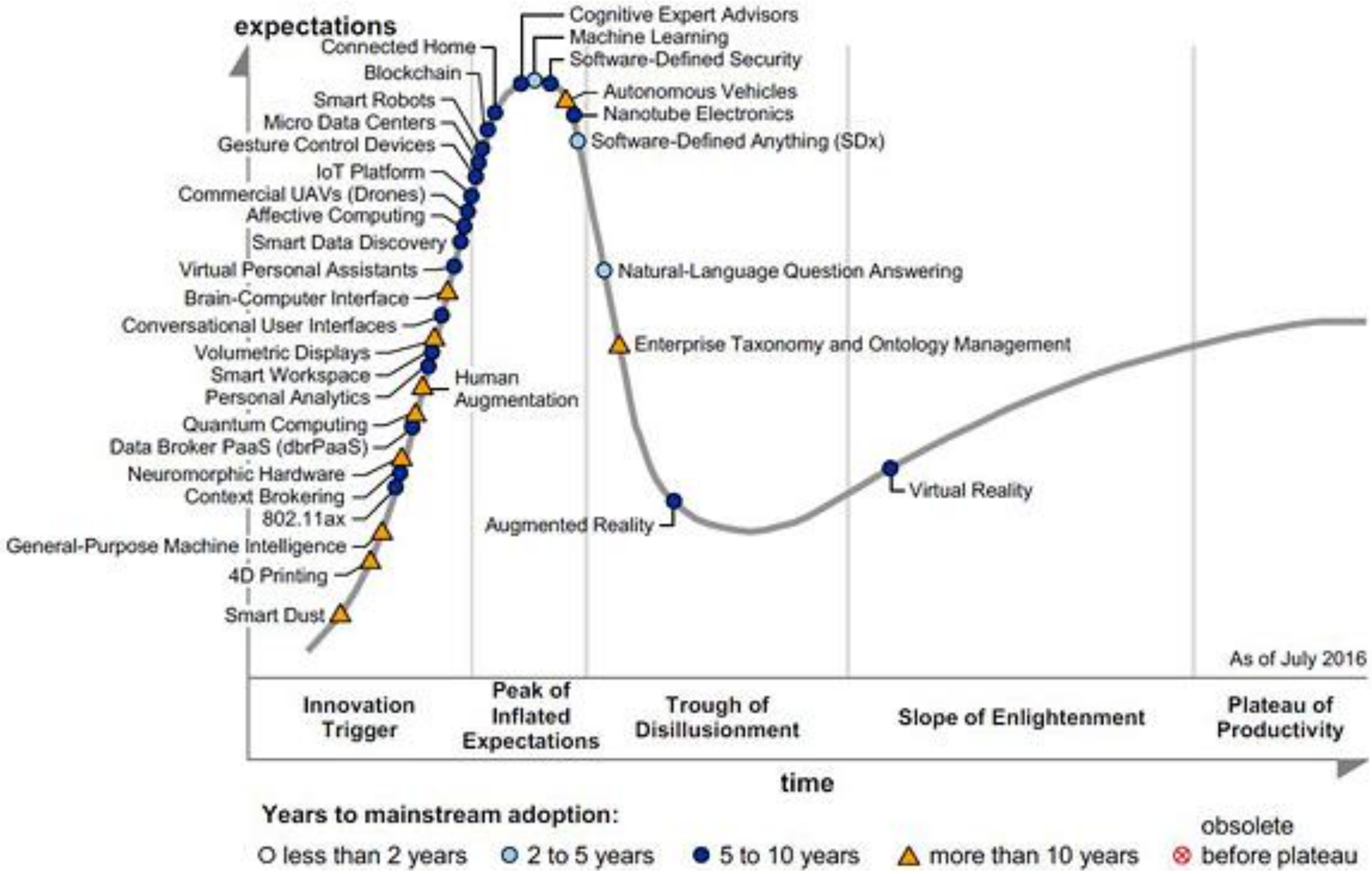
Gartner Hype Cycle – 2014



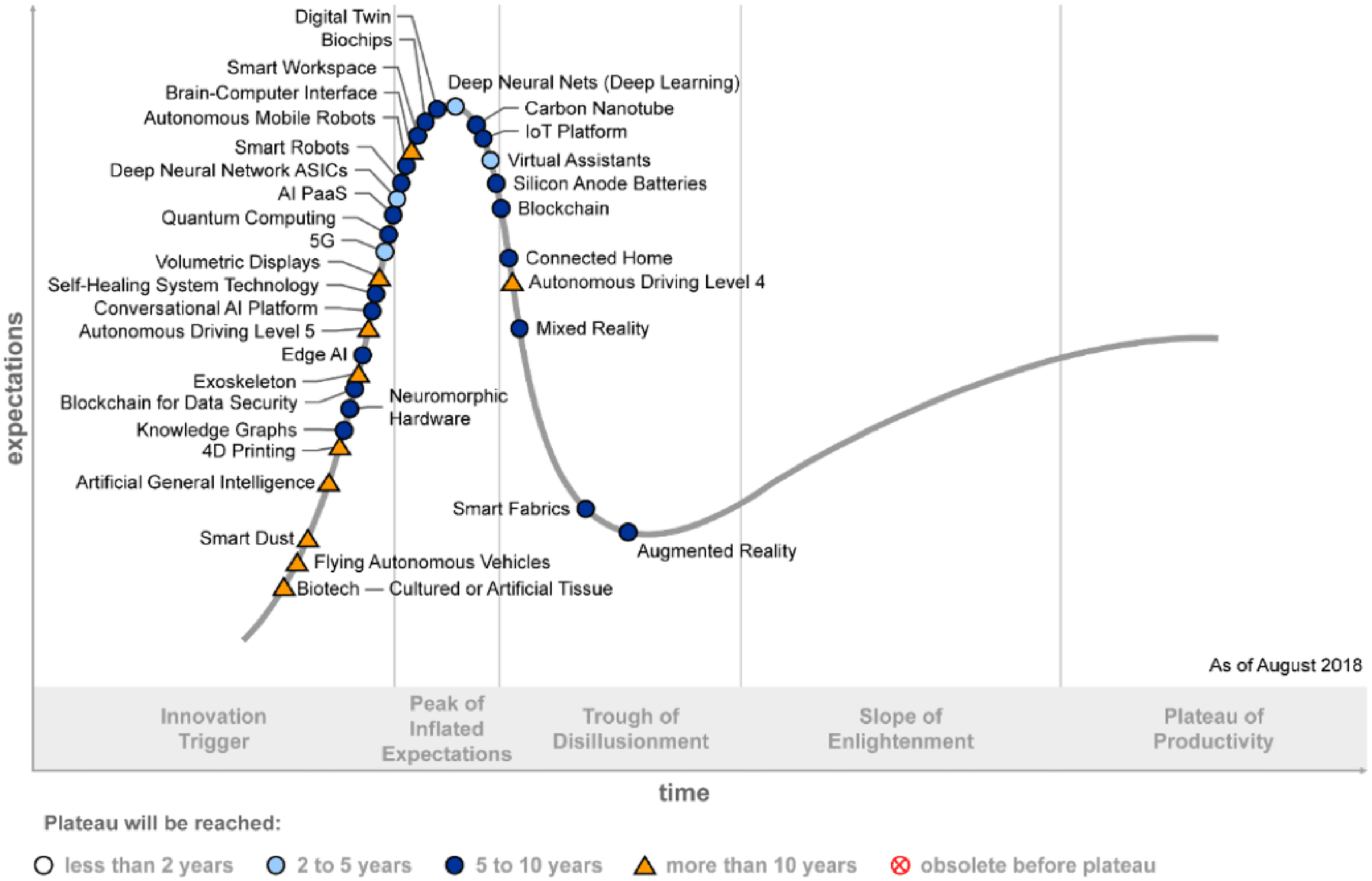
Gartner

esprit

Gartner Hype Cycle – 2016



Gartner Hype Cycle – 2018



© 2018 Gartner, Inc.

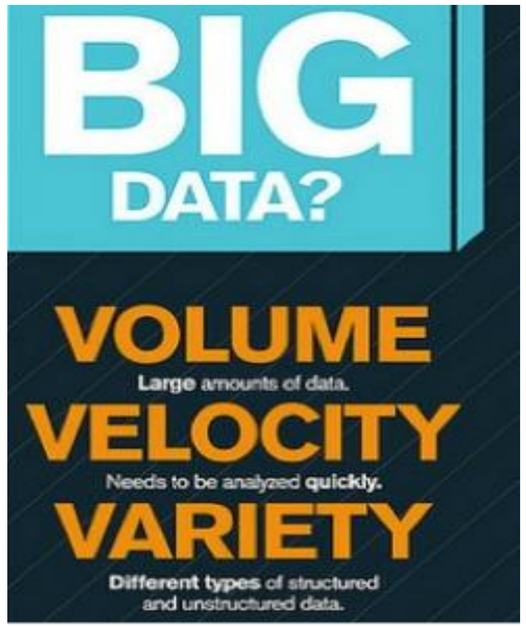
Qu'est-ce que le Big Data ?

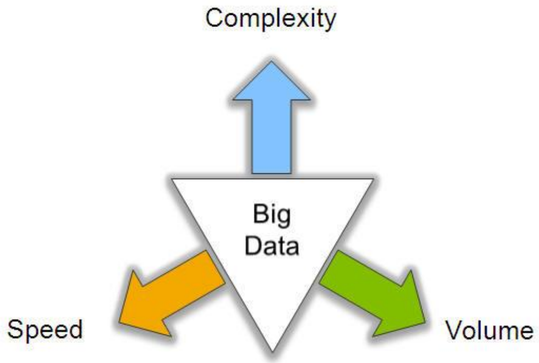
- **Big data** désigne une collection d'ensembles de données d'une certaine taille et complexité tel qu'il devient difficile de les traiter à l'aide d'outils de gestion de base de données ou d'applications de traitement de données classiques.

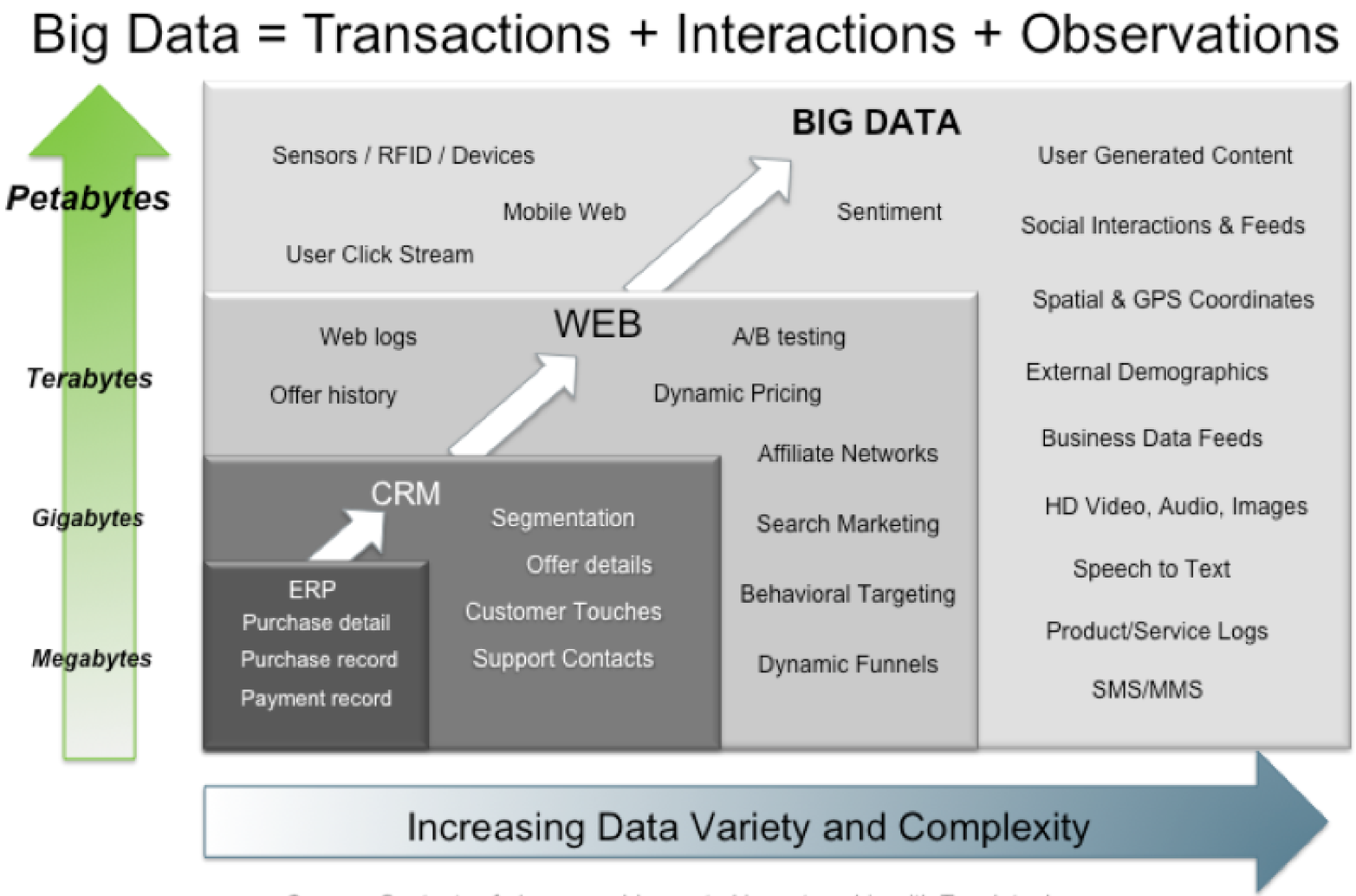
- Les défis incluent la capture, la conservation, le stockage, la recherche, le partage, le transfert, l'analyse et la visualisation.

esprit

Big Data: 3V's





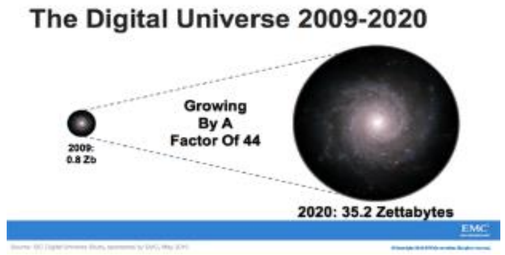


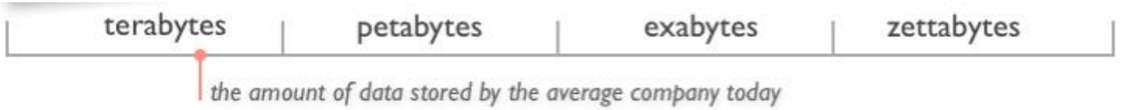
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

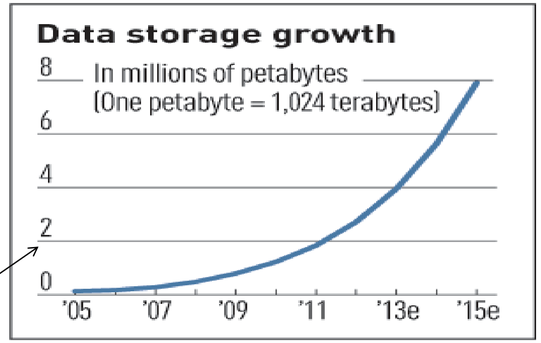
Volume (Scale)

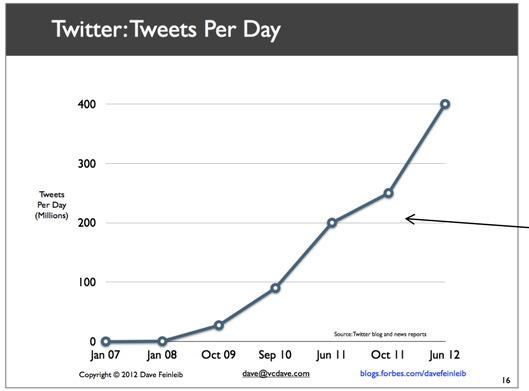
1 ZB = 10^{21} Bytes

- **Volume de données**
  - augmentation * 44 de 2009 à 2020
  - De 0.8 zettaoctets à 35 zettaoctets
- Le volume de données augmente de façon exponentielle







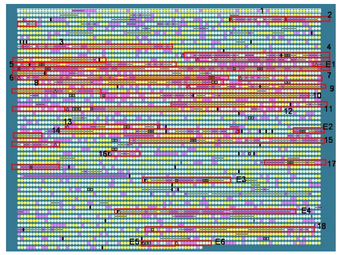


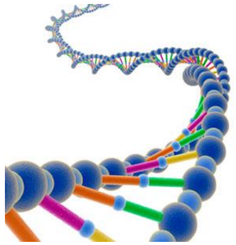
Exponential increase in collected/generated data

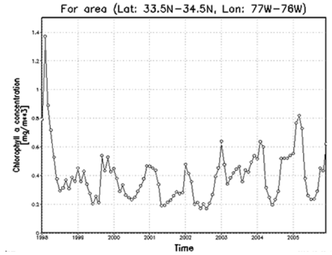
Indicateurs d'utilisation

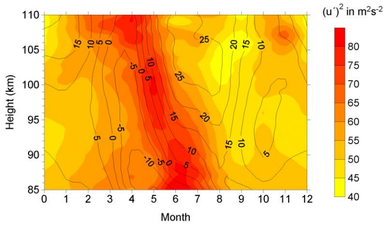
Variété (Complexité)

- Données relationnelles (tables / transaction / données existantes)
- Données texte (Web)
- Données semi-structurées (XML)
- Données graphiques
  - Réseau social, Web sémantique (RDF),…
- Données en streaming
  - Vous ne pouvez numériser les données qu'une seule fois
- Une seule application peut générer / collecter de nombreux types de données
- Big Data Publique (météo, finance, etc.)



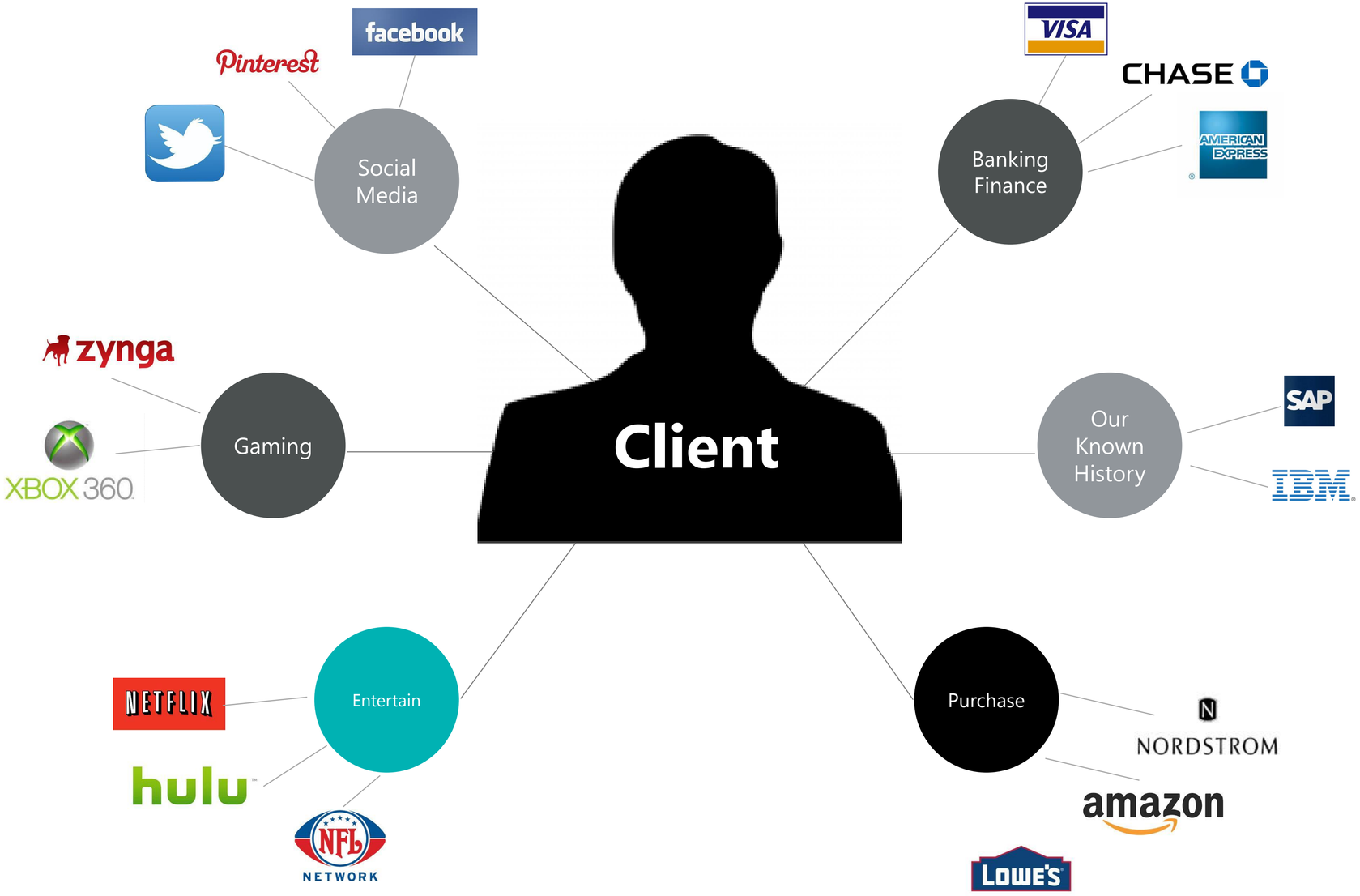






Pour extraire de la connaissance → tous ces types de données doivent être liées entre elles

Une vue unique pour le client



esprit

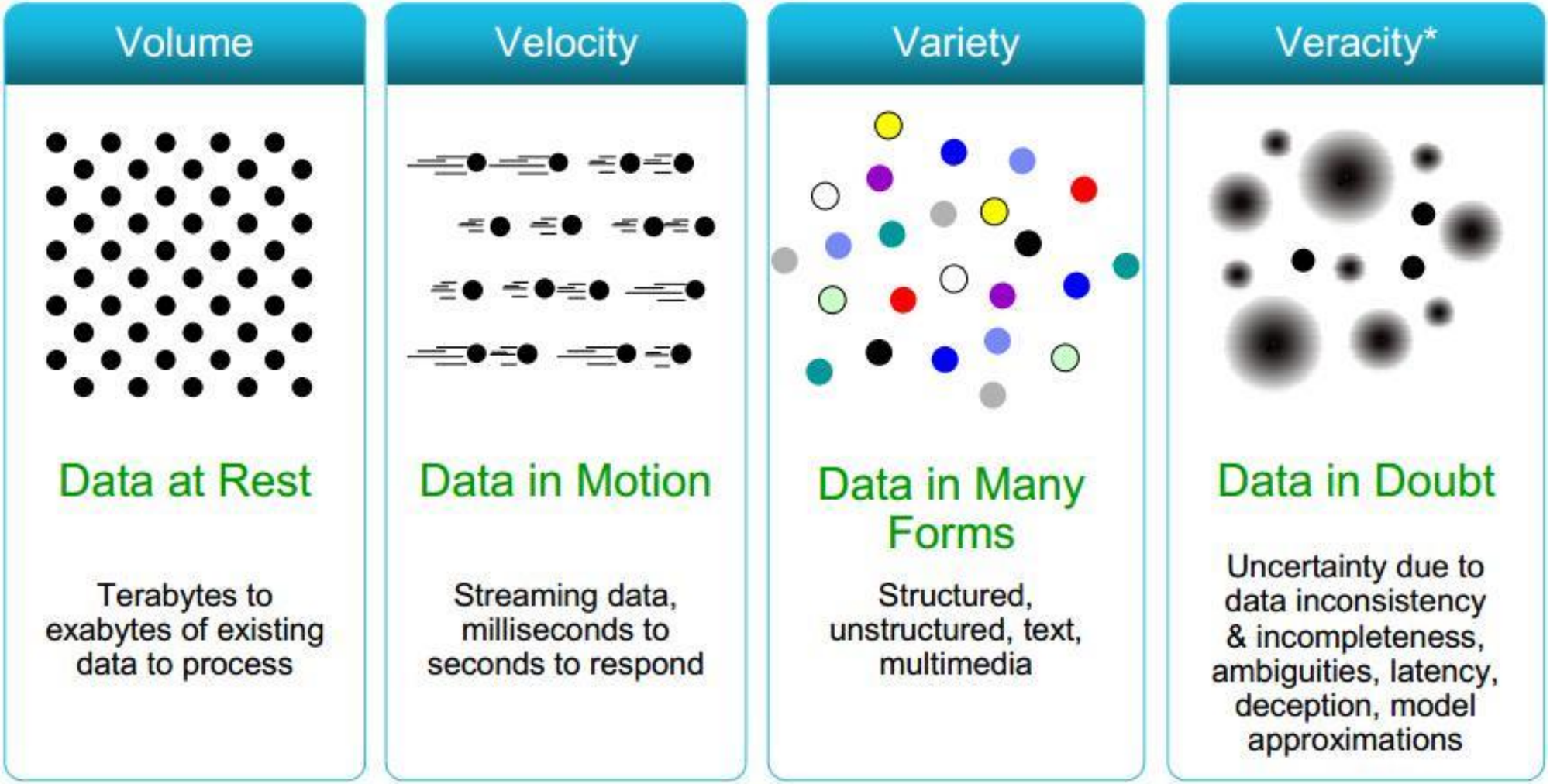
Vélocité (Vitesse)

- Les données sont générées rapidement et doivent être traitées rapidement
- Analyse de données en ligne
- Décisions tardives ➔ Opportunités perdues

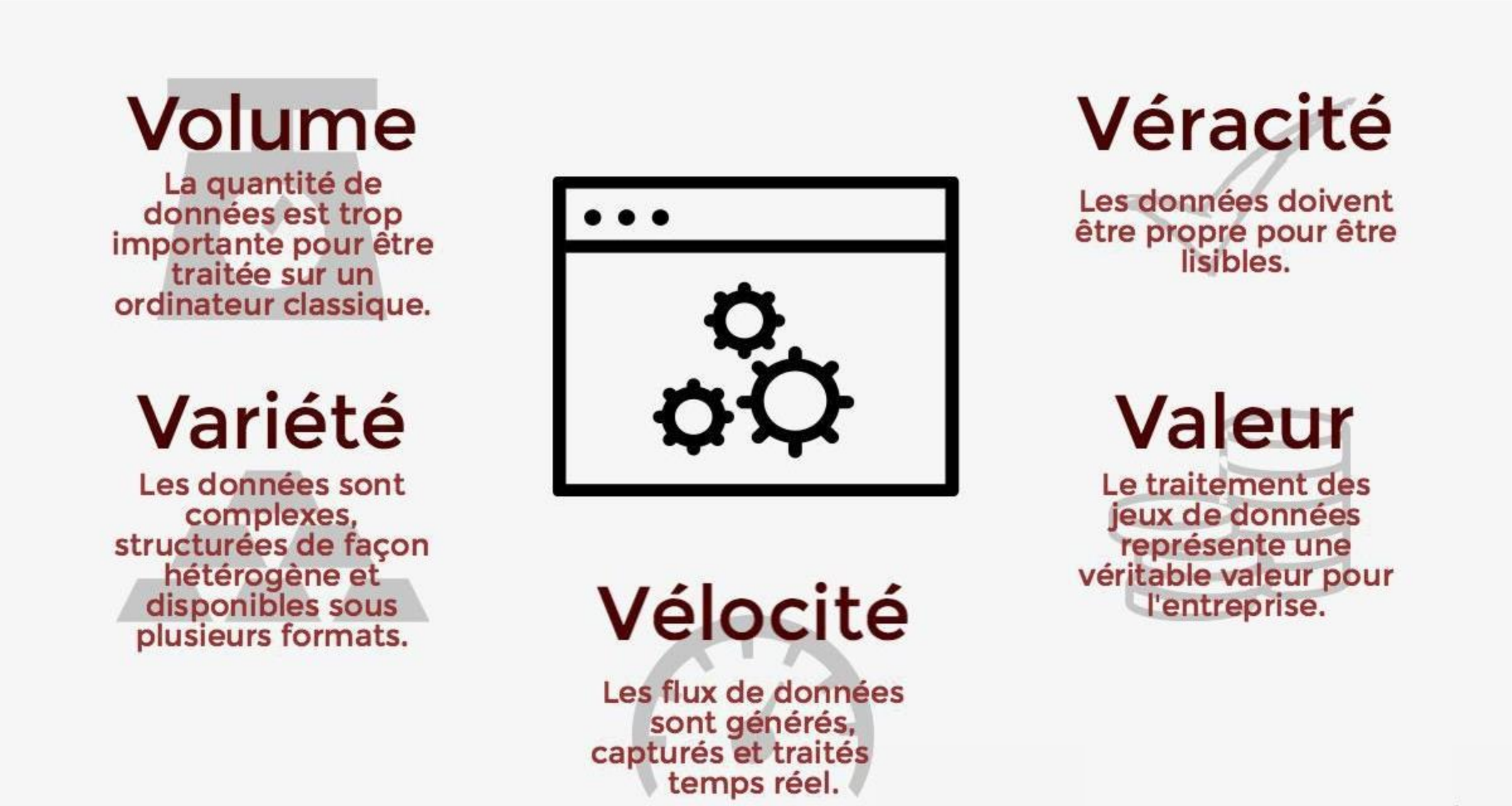
- **Exemples**
  - **E-Promotions:** En fonction de votre emplacement actuel, de l'historique de vos achats, de ce que vous aimez ➔ envoyez des promotions immédiates des magasins voisins
  - **Surveillance de santé :** capteurs surveillant vos activités et votre corps ➔ toute mesure anormale nécessite une réaction immédiate

Analyse en temps réel / Exigences de la prise de décision

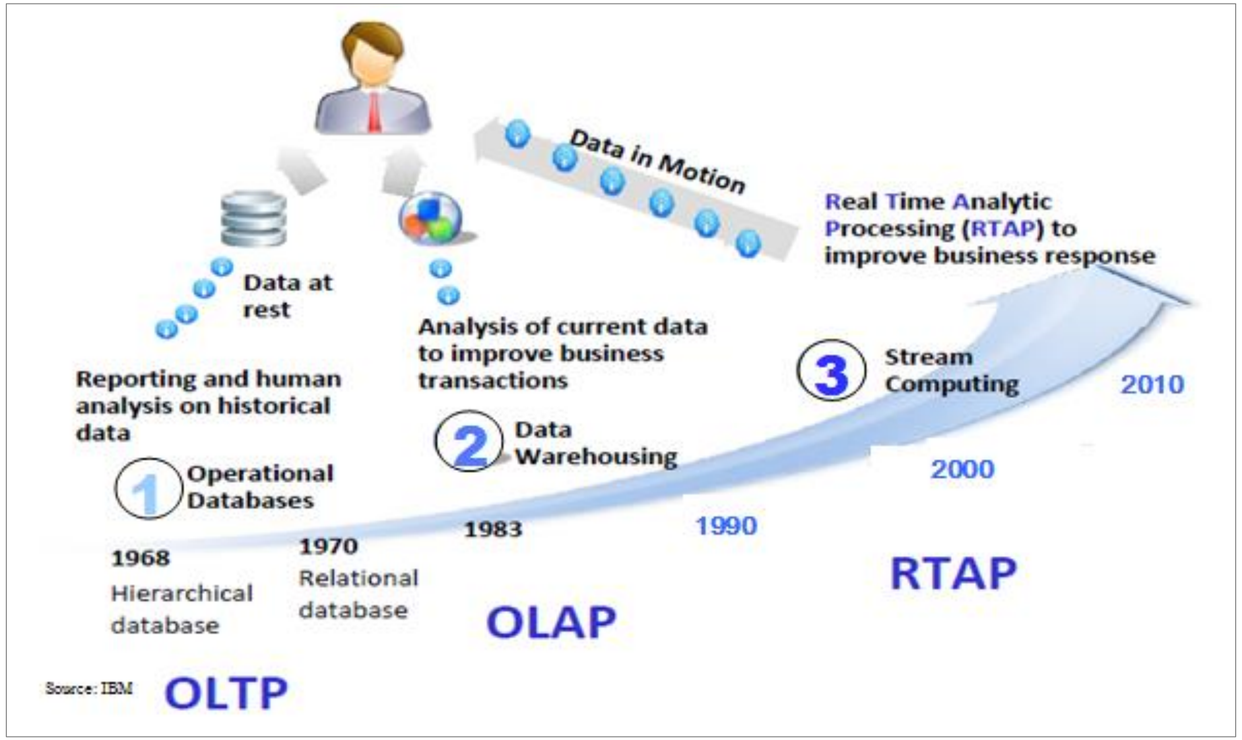
Certains le font 4V



D’autres le font 5V



Exploiter le Big Data



- **OLTP :** Online Transaction Processing (DBMSs : BD de l'ERP, BD d'une application classique, …)
- **OLAP :** Online Analytical Processing (Data Warehousing : Business Intelligence DWH et Cubes OLAP)
- **RTAP :** Real-Time Analytics Processing (Big Data Architecture & technology : Analyse en temps réel)

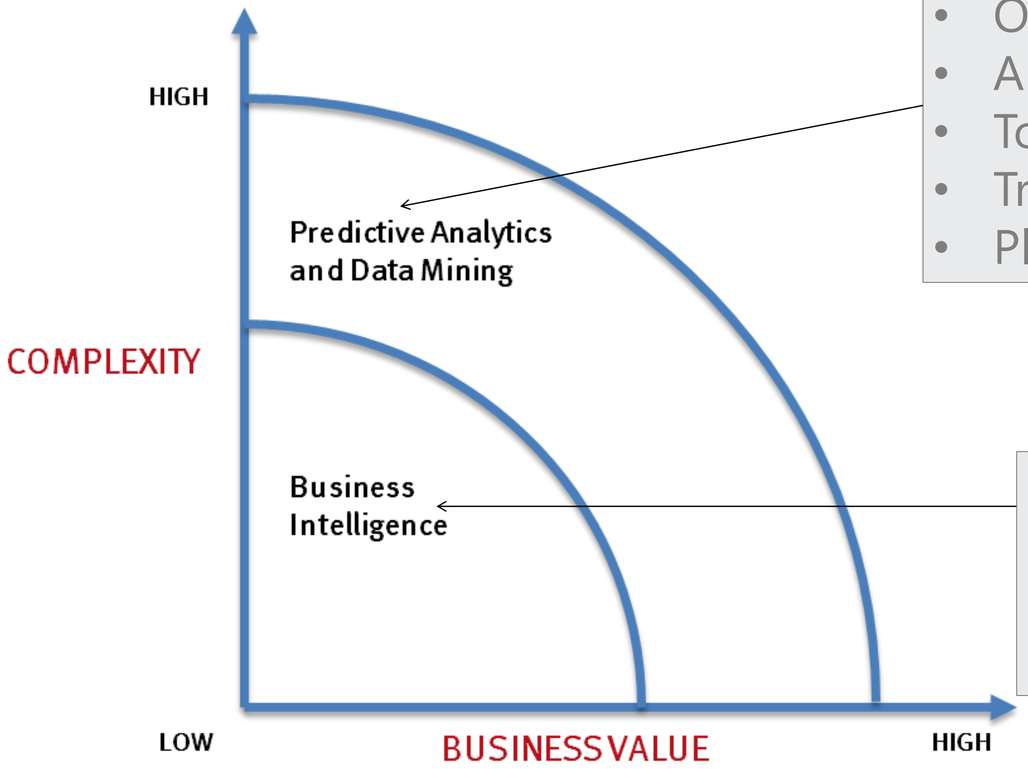
Le modèle a changé…

- **Le modèle de génération / consommation de données a changé**

Ancien modèle : peu d'entreprises génèrent des données, toutes les autres consomment des données

Nouveau modèle : toutes les entreprises / individus génèrent et consomment des données

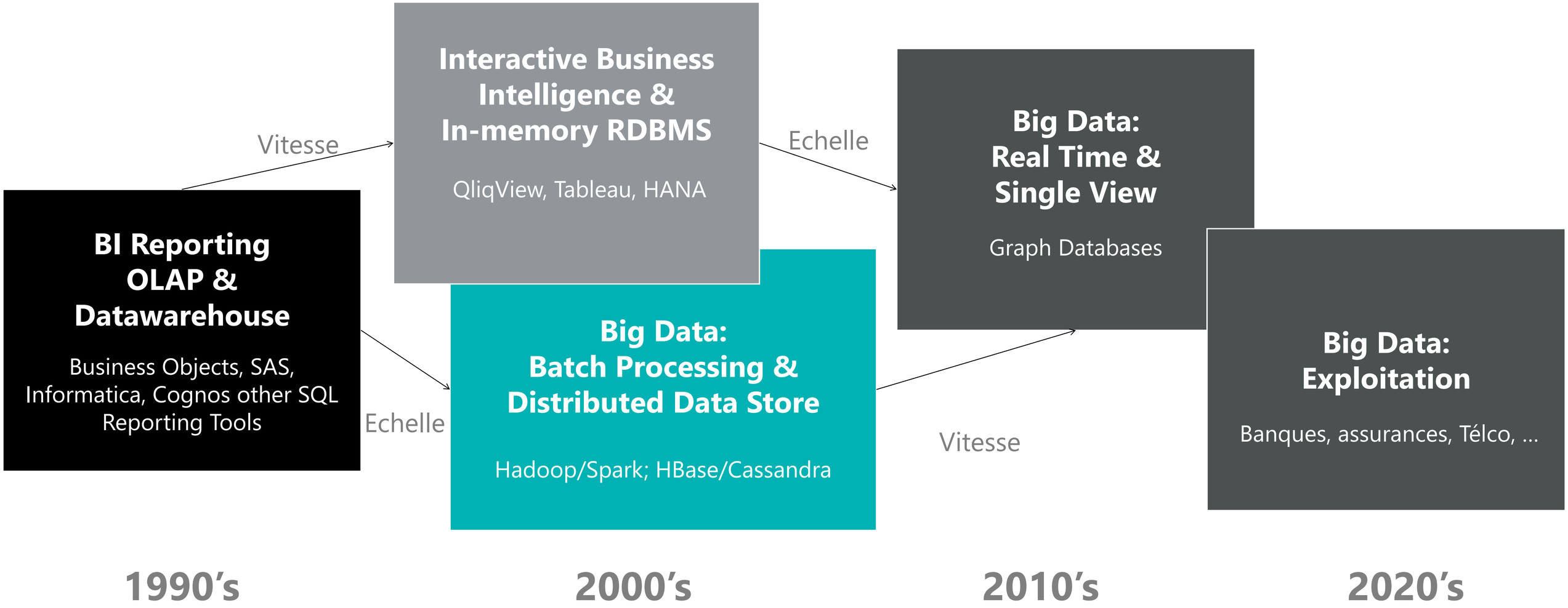
Qu'est-ce qui motive le Big Data ?



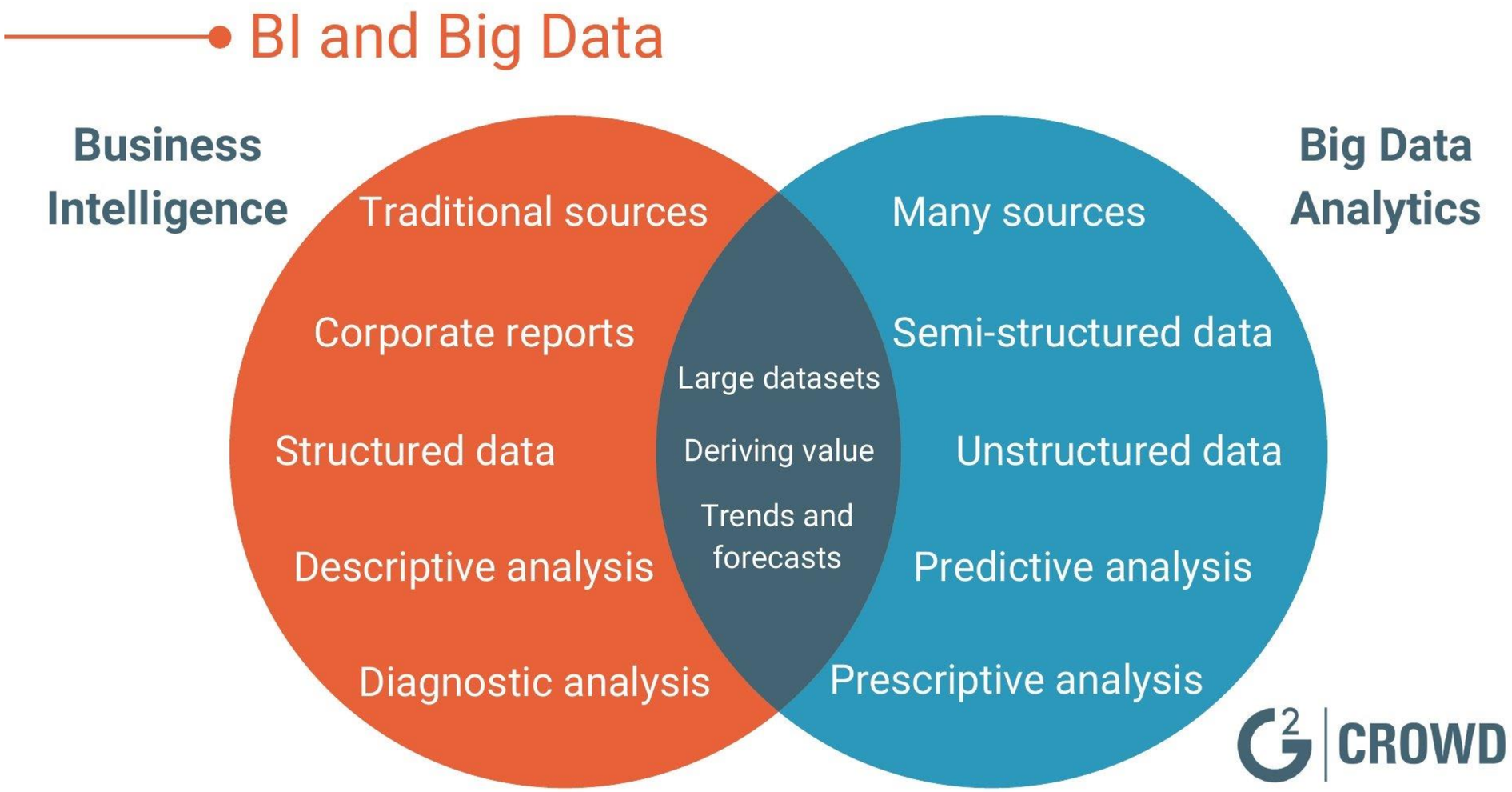
- Optimisations et analyses prédictives
- Analyse statistique complexe
- Tous types de données et de différentes sources
- Très grands ensembles de données
- Plus de temps réel

- Requêtes et rapports ad hoc
- Techniques d'exploration de données
- Données structurées, sources typiques
- Ensembles de données de taille petite à moyenne

Evolution de la *Business Intelligence*



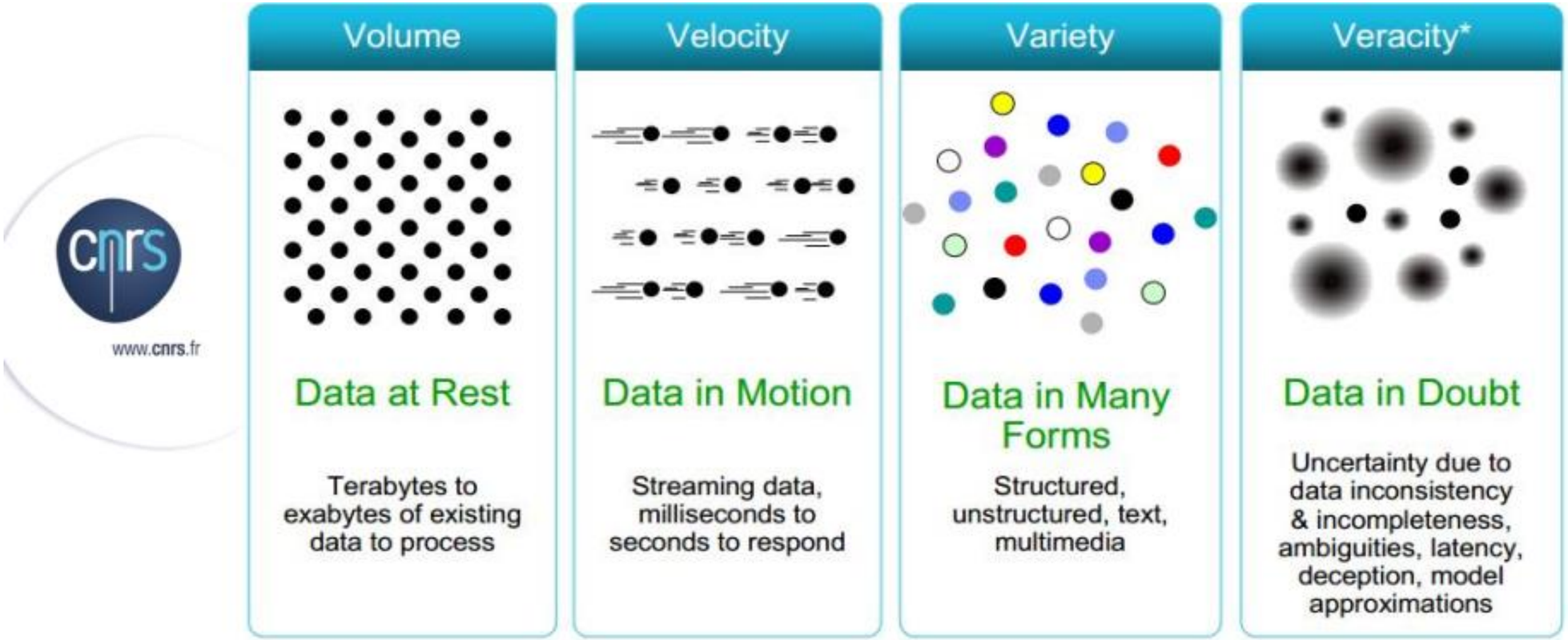
BI Vs Big Data



esprit

Processus BI classique

Complexité multidimensionnelle du Big Data



- Nouvelles archi. de stockage
- Nouvelles archi. d'interopérabilité

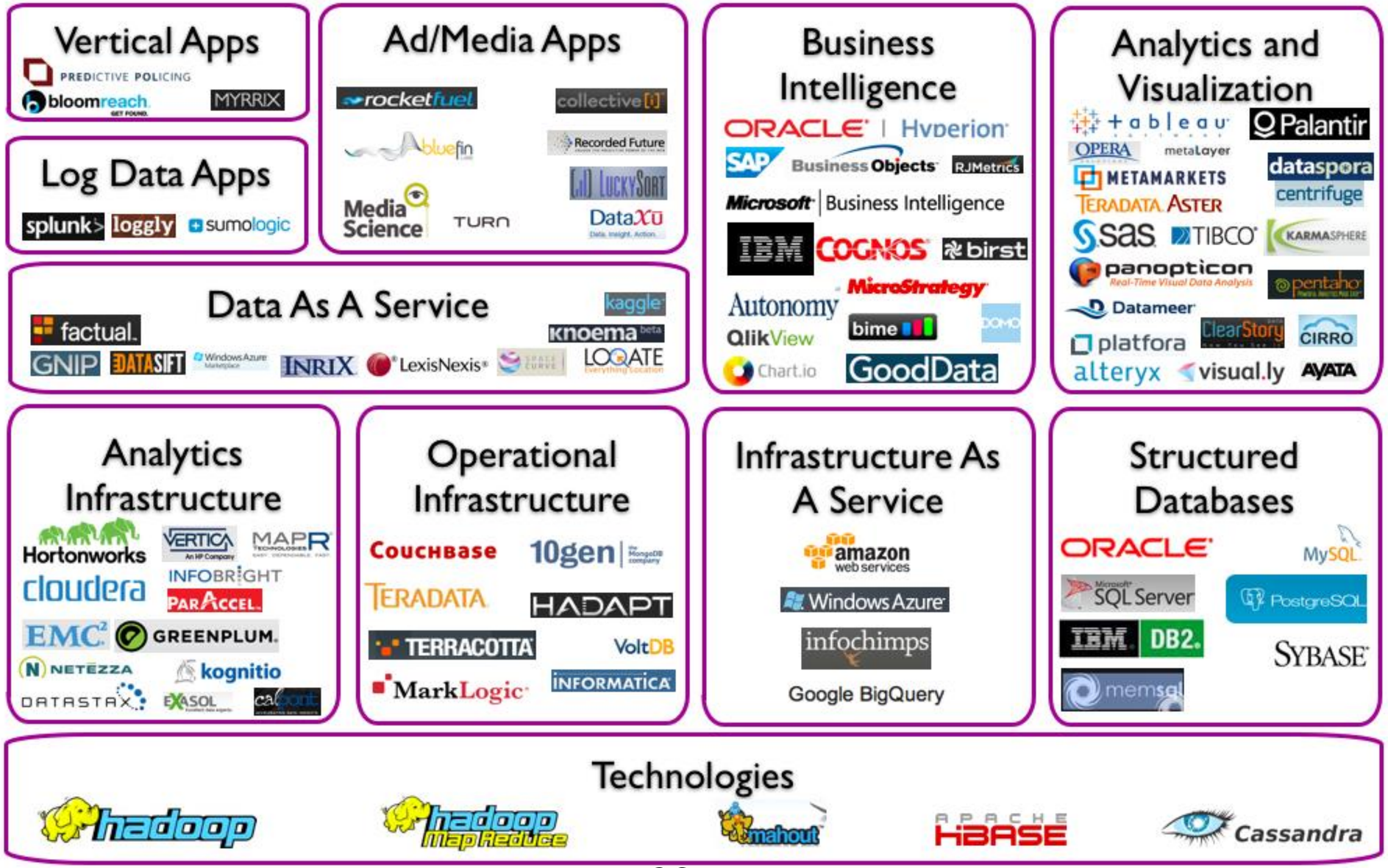
- Défi pour les réseaux de communication
- Nouveaux modèles de calcul sur des flux

- Nettoyage et transformation
- Fusion de données

Nouveaux modèles de qualité (données & processus de traitement)

[Src: Bouzeghoub, Mastodons: Une approche interdisciplinaire des Big Data]
http://www.datasciencecentral.com/profiles/blogs/data-veracity

Big Data – Vue d'ensemble



esprit

Qu’est ce que Hadoop ?

- Framework 100% open source, écrit en Java et géré par la fondation Apache,
- Capable de stocker et traiter un grand nombre de donnés,
- Relie plusieurs serveurs banalisés entre eux pour travailler en parallèle,
- Utilisé par des entreprises ayant de très fortes volumétries de données à traiter :
  - Facebook, Twitter, eBay, Amazon, …
  - Et des centaines d’entreprises à travers le monde

Qu’est ce que Hadoop ?

- Offre une grande flexibilité,
  - En termes de données, de reconstitution, de tolérance aux pannes, …

- Performances évolutives en fonction du nombre de machines constituant le cluster :
  - Plus le nombre de nœuds est élevé moins le temps d’exécution des jobs est court!

- Fonctionne sur le principe des grilles de calcul :
  - répartir l’exécution d’un traitement intensif de données sur plusieurs nœuds ou grappes de serveurs.

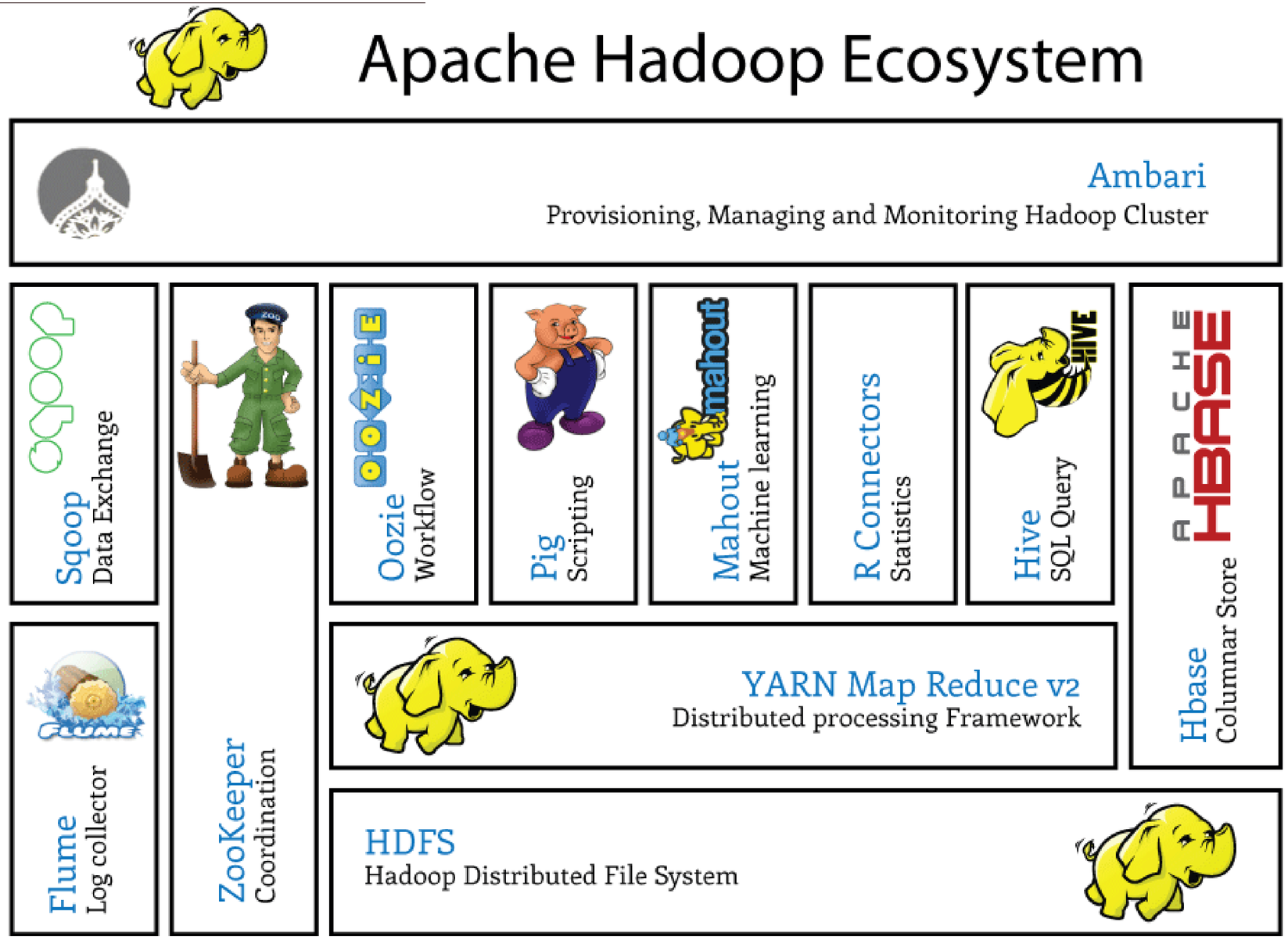
Le nom “Hadoop” était initialement celui d’un éléphant en peluche, jouet préféré du fils de Doug Cutting.

Qu'est-ce que Hadoop ?

- Plate-forme logicielle qui permet d'écrire et d'exécuter facilement des applications qui traitent de grandes quantités de données. Il comprend :
  - MapReduce – moteur de calcul hors ligne
  - HDFS – Système de fichiers distribué de Hadoop
  - HBase (pre-alpha) – accès aux données en ligne
- Yahoo! est le contributeur le plus important
- Voici ce qui le rend particulièrement utile
  - **Evolutif :** Il peut stocker et traiter de manière fiable des pétaoctets (2^{50} octets).
  - **Economique :** Il distribue les données et le traitement sur des clusters couramment disponibles (en milliers).
  - **Efficace :** En distribuant les données, il peut les traiter en parallèle sur les nœuds où ils se trouvent.
  - **Fiable :** Il gère automatiquement plusieurs copies de données et re-déploie automatiquement les tâches de calcul en fonction des défaillances.

esprit

Ecosystème Hadoop



Big Data Analytics

Qu'est-ce que le Data Analytics ?

- Le Data Analytics, abrégé par DA, est une science consistant à examiner des données brutes, dans le but de tirer des conclusions à partir de ces informations.

- Le Data Analytics est utilisé dans de nombreuses industries afin de permettre aux entreprises et aux organisations de prendre de meilleures décisions.

- Dans le domaine scientifique, il est utilisé pour vérifier des théories ou pour réfuter des modèles existants.

Catégories du Data Analytics ?

- EDA : Exploratory Data Analysis (analyse de donnée exploratoire)
  - Permet de découvrir de nouveaux éléments dans les données.

- CDA : Confirmatory Data Analysis (analyse de données confirmatoire)
  - Permet de prouver si des hypothèses existantes sont vraies ou fausses.

- QDA : Qualitative Data Analysis (analyse de données qualitative)
  - Utilisée dans les sciences sociales pour tirer des conclusions de données non numériques telles que les mots, les photographies ou la vidéo.

esprit

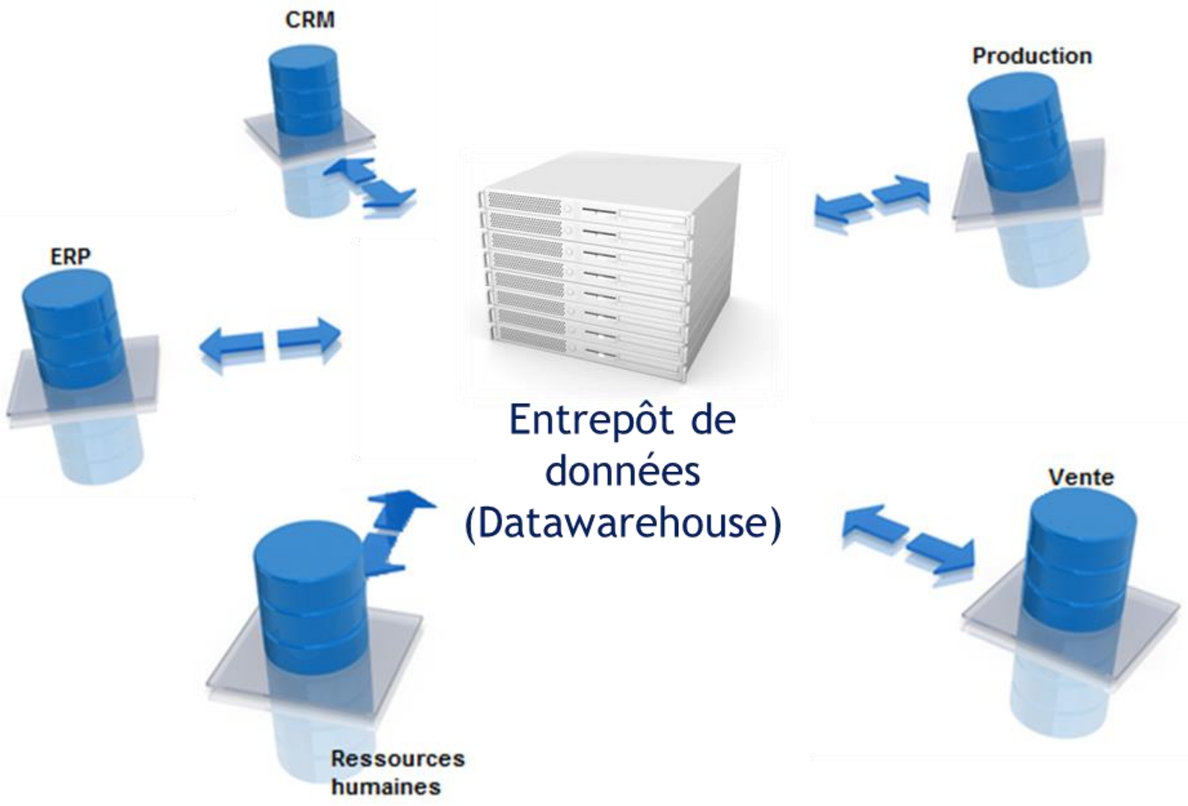
Exemples d’application

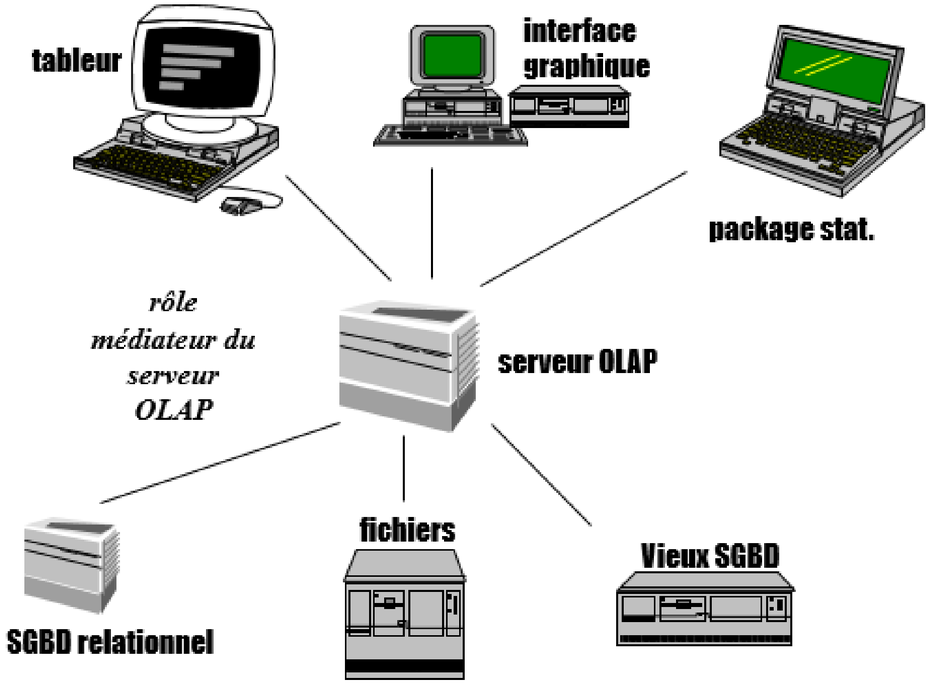
- Les banques et les entreprises de cartes de crédit analysent les transactions et les dépenses pour empêcher les fraudes et les usurpations d’identité.

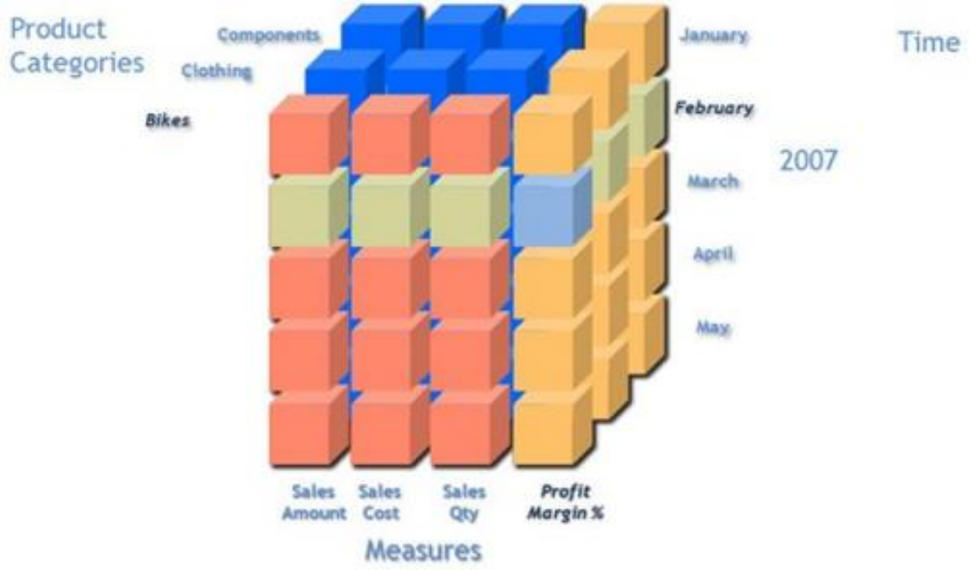
- Les sociétés de e-commerce analysent le trafic d’un site web ou les cycles de navigation pour déterminer quels consommateurs sont plus ou moins enclins à acheter un produit ou un service en fonction de leurs précédents achats ou des pages qu’ils ont consultées.

- Les opérateurs téléphoniques analysent l’historique d’usage de services et le trafic généré par les connexions data pour améliorer le processus de découverte et d’exposition des services.

Data Analytics en BI







Big Data Analytics, pourquoi ?

Big Data Analytics, c’est quoi ?

edureka!

“Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights”

esprit

Etapes du Big Data Analytics

Identifier un problème

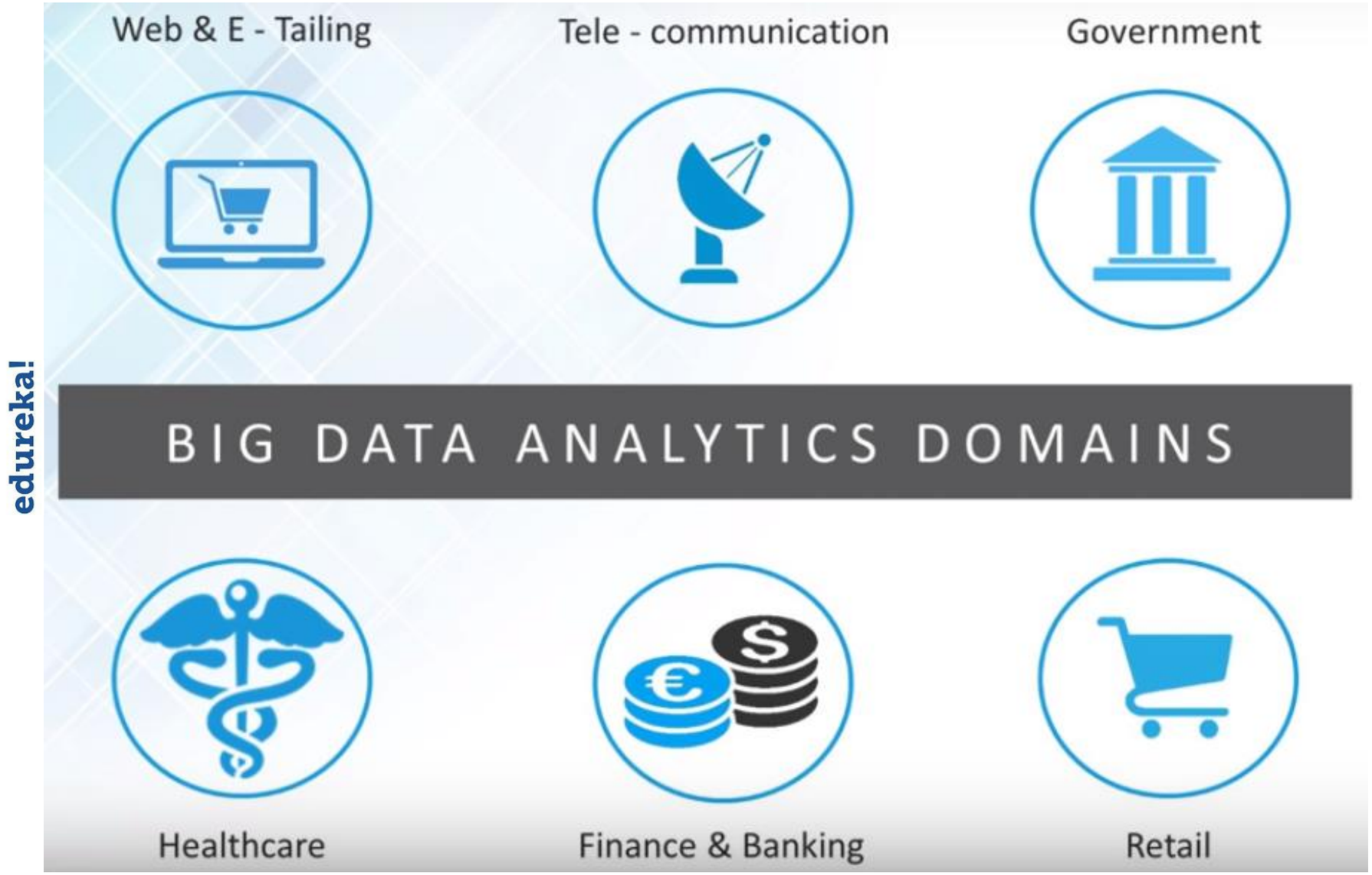
Concevoir les besoins
(en données)

Pré-traitement des données

Effectuer des analyses
sur des données

Visualiser les données

Domaines d’application du Big Data Analytics



esprit

Outils du Big Data Analytics

Hadoop fournit une solution évolutive pour stocker et traiter d'énormes ensembles de données de manière parallèle et distribuée

Apache Hive est un outil de « data warehousing » qui nous permet d'effectuer des analyses de données massives (Big Data Analytics) à l'aide du langage de requête « Hive Query Language », très similaire à SQL.

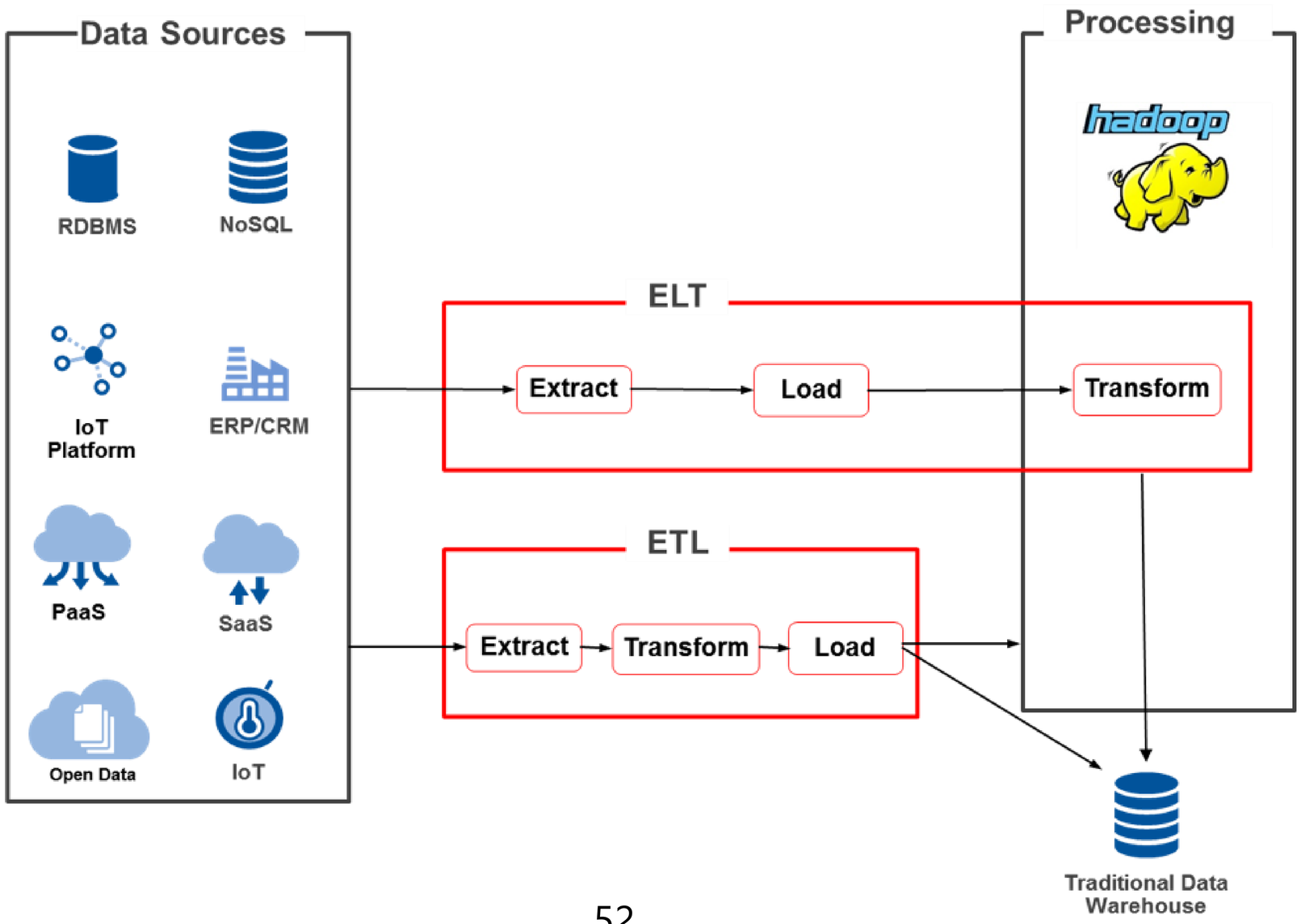
Apache Pig est une plateforme utilisée pour analyser de grands ensembles de données en les représentant sous forme de flux de données.

Apache Spark est un moteur de traitement de données « in-memory » qui permet d'exécuter efficacement du streaming, du machine learning ou une charge de travail et qui nécessite un accès itératif et rapide à des jeux de données.

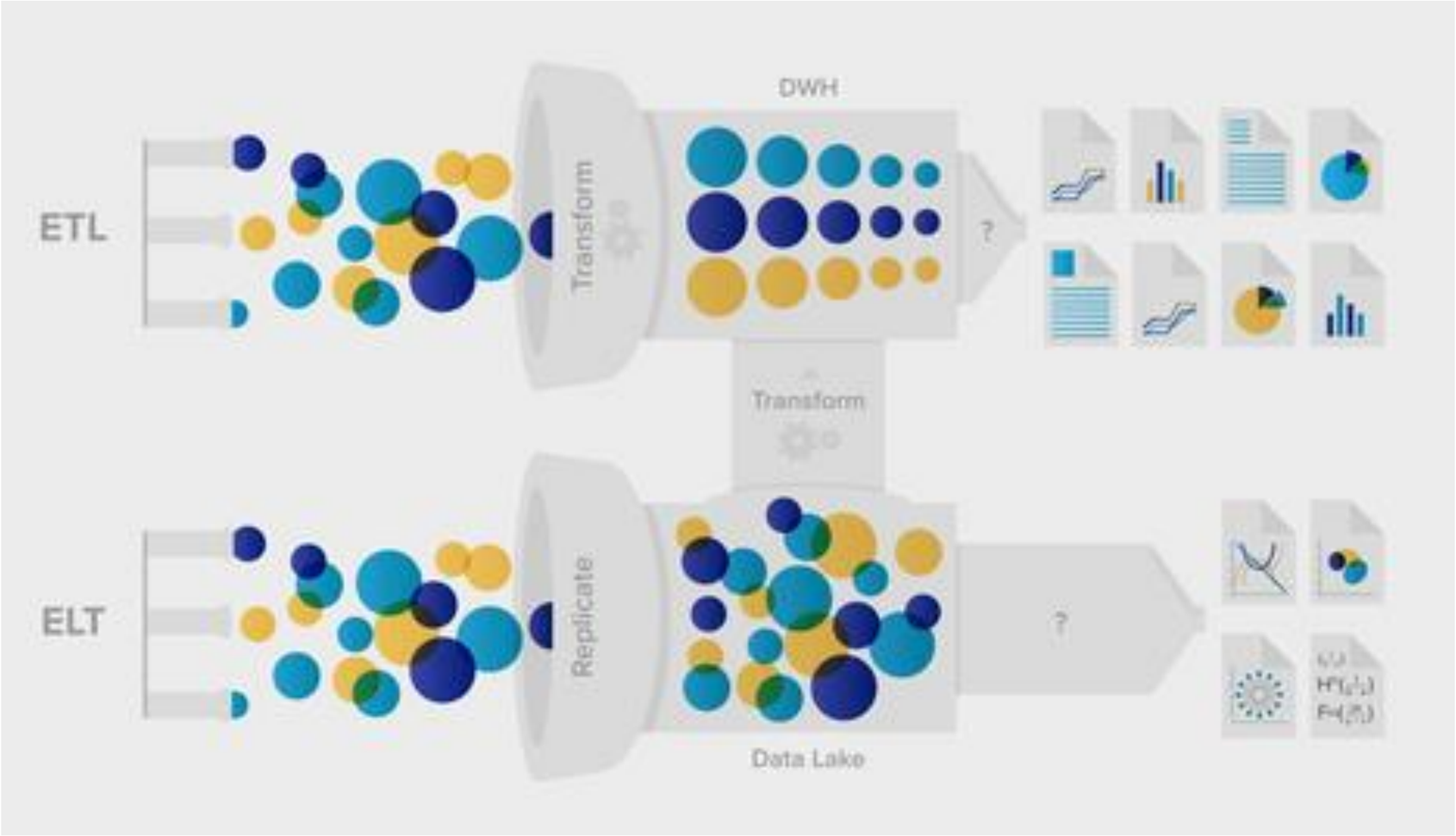
Apache Hbase est une base de données NoSQL qui nous permet de stocker facilement des données non structurées et fournit un accès en lecture / écriture en temps réel.

Data Lake

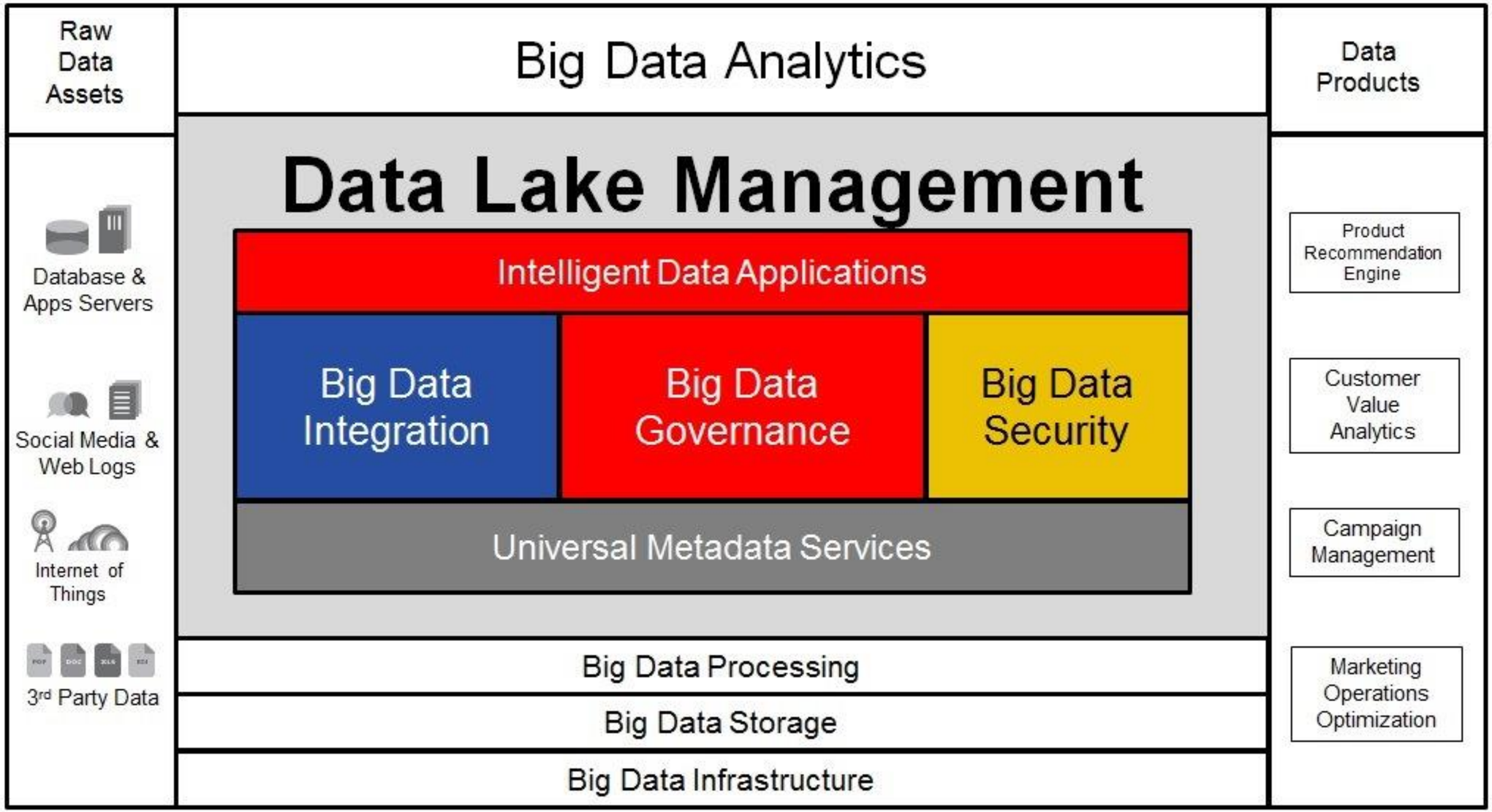
Data Lake : Collecte de données



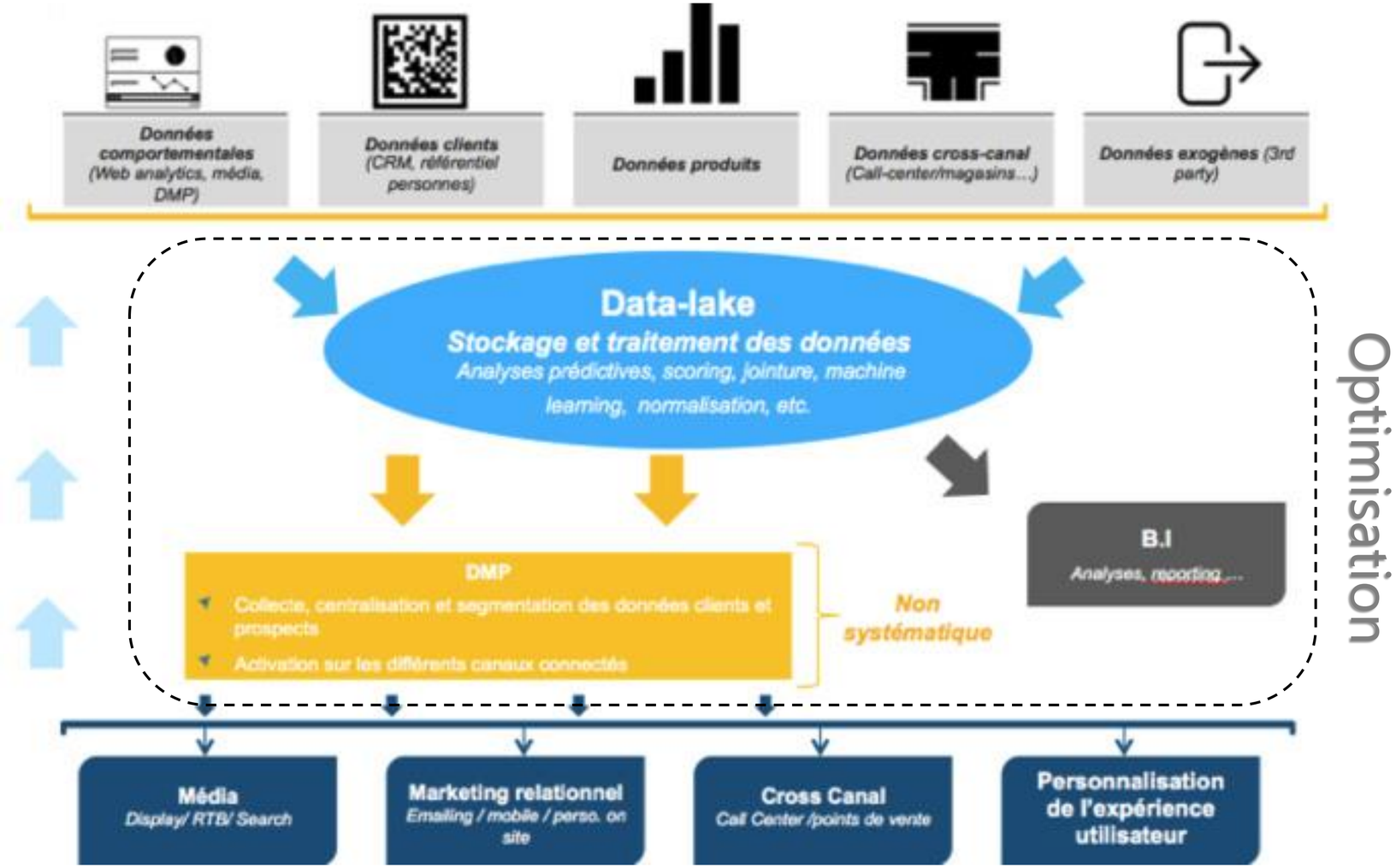
Data Lake : Collecte de données



Gestion d'un Data Lake



Data Lake : Collecte de données



Optimisation

Comment optimiser le processus de collecte de données dans un Data Lake ?

Optimisation Data Lake

1. Ajouter du contexte ajoute de la valeur :
   - Afin de donner de la pertinence aux informations provenant de sources de données non traditionnelles, un contexte doit être ajouté.
2. La gouvernance des données n’est pas mono-threadée :
   - L'avènement du lac de données a créé un vaste référentiel. Cela nécessite un modèle de gestion des données capable de prendre des décisions concernant les données (signification, utilisation, règles utilisateurs, …).
3. La conservation des données plus que la préparation
   - La conservation des données joue un rôle essentiel pour aider les lacs de données à tenir leurs promesses. Appliquée correctement, elle utilise une approche ascendante pour transformer toute donnée brute en information qui produit des analyses utiles et facilement consommée.

esprit

Optimisation Data Lake

4. Assurer une sécurité et une gouvernance solides de données :
   - Les données plus sensibles concernant les clients et les informations d'entreprise impliquant des transactions et des risques doivent être mieux conservées et mieux comprises.

5. Faciliter la consommation du lac de données
   - Les Data Lakes doivent prendre en charge plusieurs scénarios d'utilisation. Pour rendre le lac de données accessible, il est important de considérer les différents utilisateurs.

esprit

Machine Learning

Classification supervisée / non supervisée

Intelligence artificielle

- **Supervisé** : apprendre à classer un nouvel individu parmi un ensemble de classes prédéfinies: on connaît les classes à priori.
- **Non-supervisé :** le nombre et la définition des classes n'étant pas données à priori

Classification supervisée

- L'apprentissage/classification supervisé consiste à apprendre une fonction de prédiction à partir d'exemples annotés

Principe : On dispose d'éléments déjà classés
Exemple : articles en rubrique cuisine, sport, culture…
On veut classer un nouvel élément
Exemple: lui attribuer un nom parmi cuisine, sport, culture.

- Dans notre fichier immobilier, si on veut classer les biens, on dispose déjà des classes (s+1, s+2, s+3, …)
- Si un nouvel élément (un nouveau appartement par exemple) se rajoute à la liste, on peut le classer selon cette logique

Applications : Vision par ordinateur, Reconnaissance de formes, Reconnaissance de l'écriture manuscrite, Reconnaissance vocale, Traitement automatique de la langue, Bio-informatique, …

Classification non supervisée

- L'apprentissage/classification non supervisé consiste à apprendre sans superviseur.
- Il s'agit d'extraire des classes ou groupes d'individus présentant des caractéristiques communes.

Principe : On dispose d'éléments non classés
Exemple : une fleur
On veut les regrouper en classes
Exemple: si deux fleurs ont la même forme, elles sont en rapport avec une même plante correspondante

- Dans le contexte immobilier, on ne dispose pas de classes des profils d'acheteurs par exemple. On veut faire une classification pour mieux détecter le besoin des acheteurs.
- Si un nouvel élément (un nouvel acheteur) se présente, on peut le classer selon cette logique et observer le comportement des acheteurs de la même classe pour anticiper son besoin

Applications : le partitionnement de données, l'estimation de densité de distribution, …

esprit

Algorithmes Machine Learning

think big data

regression

Ordinary Least Squares Regression (OLSR)
Linear Regression
Logistic Regression
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Jackknife Regression

regularization

Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least-Angle Regression (LARS))

instance based

also called cake-based, memory-based

k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)

dimesionality reduction

Principal Component Analysis (PCA)
Principal Component Regression (PCR)
Partial Least Squares Regression (PLSR)
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Discriminant Analysis (LDA, MDA, QDA, FDA)

bayesian

Naive Bayes
Gaussian Naive Bayes
Multinomial Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Bayesian Network (BN)
Hidden Markov Models
Conditional random fields (CRFs)

decision tree

Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5 and C5.0 (different versions of a powerful approach)
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
M5
Random Forests
Conditional Decision Trees

clustering

Single-linkage clustering
k-Means
k-Medians
Expectation Maximisation (EM)
Hierarchical Clustering
Fuzzy clustering
DBSCAN
OPTICS algorithm
Non Negative Matrix Factorization
Latent Dirichlet allocation (LDA)

deep learning

Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders

associated rule

Apriori
Eclat
FP-Growth

ensemble

Logit Boost (Boosting)
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (blending)
Gradient Boosting Machines (GBM)
Gradient Boosted Regression Trees (GBRT)
Random Forest

neural networks

Self Organizing Map
Perceptron
Back-Propagation
Hopfield Network
Radial Basis Function Network (RBFN)
Backpropagation
Autoencoders
Hopfield networks
Boltzmann machines
Restricted Boltzmann Machines
Spiking Neural Networks
Learning Vector quantization (LVQ)

...and others

Support Vector Machines (SVM)
Evolutionary Algorithms
Inductive Logic Programming (ILP)
Reinforcement Learning (Q-Learning, Temporal Difference, State-Action-Reward-State-Action (SARSA))
ANOVA
Information Fuzzy Network (IFN)
Page Rank
Conditional Random Fields (CRF)

esprit

Familles d'algorithmes en ML

- **Régression :**
  - Ce type d'algorithmes supervisés va trouver une valeur continue (un nombre réel) qui est la prédiction de la valeur d'une nouvelle observation donnée.
  - Exemple : prédiction de prix de maison en fonction de ses caractéristiques (superficie, nb pièces, …).
- **Classification :**
  - Ces algorithmes vont classifier une donnée dans une catégorie. Il s'agit également d'un algorithme supervisé.
  - Exemple : classifier un mail en Spam ou non, décider si une tumeur est maligne ou bénigne.
- **Clustering :**
  - Il s'agit d'une famille d'algorithmes non supervisés. Par conséquent, les données n'ont pas d'étiquettes (Non-labeled Data). L'algorithme va regrouper les données par similarité.
  - Exemple : on fournit un ensemble de photos de plantes (sans qu'on dise de quels plantes il s'agit). L'algorithme va regrouper les photos des roses ensemble et celle des tulipes ensemble etc…

NB : on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification.

esprit

Régression

- Ensemble de méthodes statistiques très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres,

- Faire des prédictions à partir de données en apprenant la relation entre les caractéristiques de ces données et certaines réponses observées à valeurs continues,

- Fondamentalement, la classification consiste à prédire un label et la régression, à prédire une quantité.

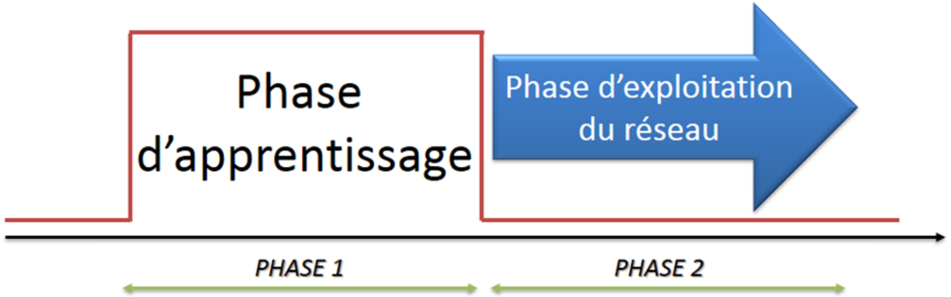
- La régression vous permet de faire des prédictions à partir de données en apprenant la relation entre les caractéristiques de vos données et certaines réponses observées à valeurs continues
  - Exemple d'application : prédiction du prix des actions

esprit

Régularisation

- Un processus consistant à ajouter de l'information à un problème pour éviter le surapprentissage (« overfitting »),
  - Cette information prend généralement la forme d'une pénalité envers la complexité du modèle.
  - Une méthode généralement utilisée est de pénaliser les valeurs extrêmes des paramètres, qui correspondent souvent à un surapprentissage.
- C'est une forme de régression qui contraint / régularise ou réduit les estimations de coefficient vers zéro,
- Décourage l'apprentissage d'un modèle plus complexe ou plus flexible, afin d'éviter le risque de surapprentissage.
- Dans notre exemple immobilier, pour prédire le montant de vente d'un bien, on ne considère pas les valeurs extrêmes de vente (les plus chers ou les moins chers) selon un certain seuil à partir de l'échantillon de données que nous considérons.
  - On dit par exemple que les seuils de prix d'un bien de superficie entre 100m² et 120m² est de 150MD et 600MD pour éviter le surapprentissage ce qui affectera les résultats de prédiction par la suite.
  - Risque de surapprentissage : risque que le modèle développé s'ajuste trop bien aux données fournies (créer des classes très pointues de biens par exemple)

Instance based



- L'apprentissage « *Instance based* » est une famille d'algorithmes d'apprentissage qui, au lieu d'effectuer une généralisation explicite, compare les nouvelles instances aux instances vues lors de la phase d'apprentissage stockées en mémoire.

- Il s'appelle « *instance based* » parce qu'il construit des hypothèses directement à partir des instances de la phase d'apprentissage. Cela signifie que la complexité des hypothèses peut croître avec les données

- Avantage : capacité à adapter le modèle à des données inexistante à l'avance (nouvelles données).
  - Les apprenants peuvent stocker une nouvelle instance ou éliminer une ancienne.

Instance based

Exemple d'application : « *Nearest Neighbor Classification* »

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
12-30	mild	rain	high	false	yes
tomorrow	mild	sunny	normal	false	yes

Dimensionality reduction

- Consiste à prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension.
  - Pour que l'opération soit utile il faut que les données en sortie représentent bien les données d'entrée.
  - Avantage : les données de plus petites dimension peuvent être traitées plus rapidement

- Exemple d'application :
  - Les données multimédia : Un grand volume de données multimédia et un nombre important de dimensions nécessaires pour les décrire → utiliser des techniques de réduction de dimensions afin de pouvoir, d'une part, extraire et résumer les connaissances souvent inconnues des données, et, d'autre part, indexer et retrouver rapidement les documents.

esprit

Bayesian

- Des algorithmes d'apprentissage supervisé utilisés pour la classification.
- On calcule le résultat en se basant sur plusieurs variables.
- L'application du théorème de Bayes sur plusieurs variables rend le calcul complexe.
  - Pour contourner cela, une approche consiste à prendre en considération ces variables indépendamment les unes des autres.
- Particulièrement utiles pour les problématiques de classification de texte.
  - Exemple : filtre anti-spam

esprit

Bayesian

- Exemple : classification de fruits
  - Supposons qu'on ait un jeu de données sur 1000 fruits. On dispose de trois types : Banane, Orange, et "autre". Pour chaque fruit, on a 3 caractéristiques :
    - Si le fruit est **long** ou non
    - S'il est **sucré** ou non
    - Si sa couleur est **jaune** ou non

→ L'idée du jeu est de prédire le type d'un fruit (orange, banane ou autre) qu'on n'a pas encore vu. Ceci en se basant sur ses caractéristiques :
- Il est jaune
- Il est long
- Il est sucré

Jeu de données

Type	Long	petit (¬ long)	sucré	¬ sucré	Jaune	Pas jaune	Total
Banane	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Autre fruit	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Pour savoir s'il s'agit d'une banane, ou d'une orange ou d'un autre fruit, il faut qu'on calcule les trois probabilités suivantes:
- P(Banane|long, jaune, sucre) : La probabilité qu'il s'agisse d'une banane sachant que le fruit est long, jaune et sucré
- P(Orange|long, jaune, sucre) : La probabilité qu'il s'agisse d'une orange sachant que le fruit est long, jaune et sucré
- P(Autre fruit|long, jaune, sucre) : La probabilité qu'il s'agisse d'un autre fruit sachant que ce dernier est long, jaune et sucré

→ Le type du fruit "inconnu" qu'on cherche à classifier sera celui où on a la plus grande probabilité.

Arbre de décision (Decision tree)

- C'est une technique d'apprentissage supervisé :
  - on utilise un ensemble de données pour lesquelles on connaît la valeur de la variable-cible afin de construire l'arbre (données étiquetées), puis on extrapole les résultats à l'ensemble des données de test.
- Son but est de créer un modèle qui prédit la valeur d'une variable-cible depuis la valeur de plusieurs variables d'entrée.
- Chaque feuille (ou nœud terminal de l'arbre) représente soit une valeur de la variable-cible, soit une distribution de probabilité des diverses valeurs possibles de la variable-cible.
- La combinaison des valeurs des variables d'entrée est représentée par le chemin de la racine jusqu'à la feuille.

esprit

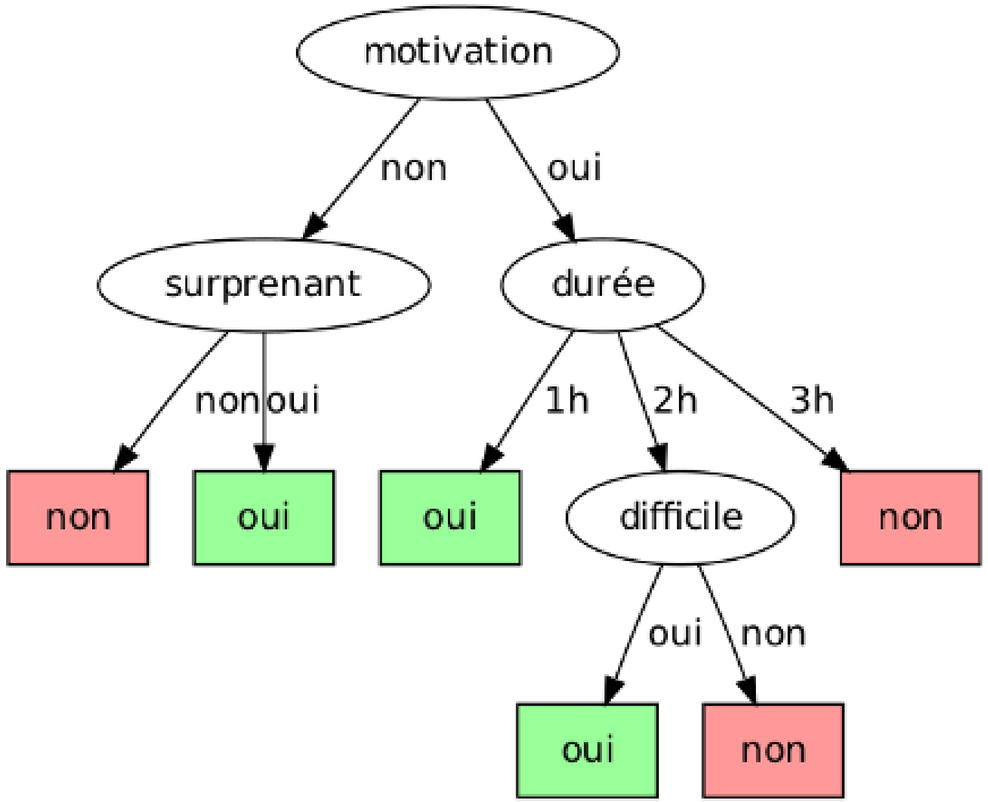
Arbre de décision (Decision tree)

- Il existe deux principaux types d'arbre de décision en fouille de données :
  - Les arbres de classification (« *Classification Tree* ») : permettent de prédire à quelle classe la variable-cible appartient, dans ce cas la prédiction est une étiquette de classe,
  - Les arbres de régression (« *Regression Tree* ») : permettent de prédire une quantité réelle (par exemple, le prix d'une maison ou la durée de séjour d'un patient dans un hôpital), dans ce cas la prédiction est une valeur numérique.

esprit

Arbre de décision (Decision tree)

- Exemple : arbre de décision pour répondre à la question « cette formation est elle intéressante ? »

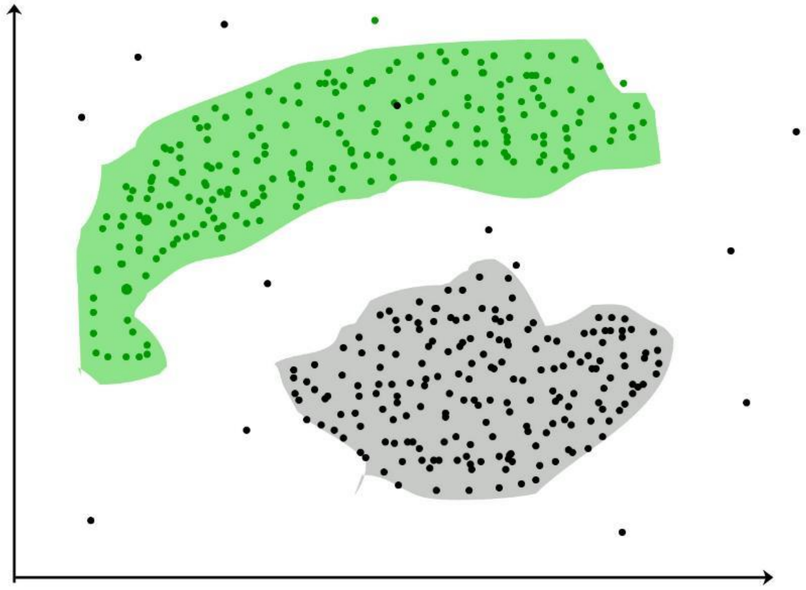


Clustering

- Faisant partie de l'apprentissage non supervisé, le Clustering consiste à diviser la population ou les points de données en un certain nombre de groupes, de sorte que les points de données du même groupe soient plus similaires aux autres points de données du même groupe et différents des points de données des autres groupes.

- Le clustering est très important car il détermine le groupement intrinsèque entre les données non étiquetées présentes.

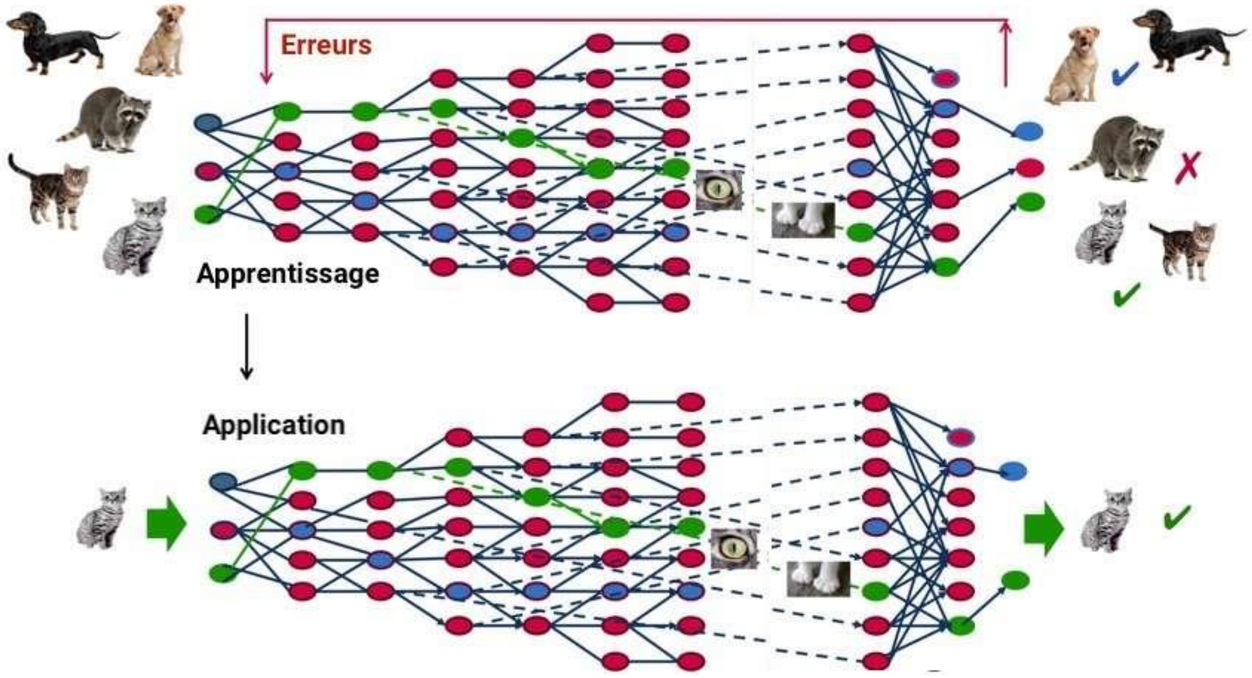
- Dans notre exemple des données immobilières, si on veut créer des groupes d'acheteurs pour améliorer l'expérience client, on a recourt au clustering pour le faire (par exemple l'algorithme K-means).



Deep Learning

- Le Deep Learning ou apprentissage profond est un type d'intelligence artificielle dérivé du Machine Learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même, contrairement à la programmation où elle se contente d'exécuter à la lettre des règles prédéterminées.

- Le Deep Learning s'appuie sur un réseau de neurones artificiels s'inspirant du cerveau humain.



À travers un processus d'autoapprentissage, le Deep Learning est capable d'identifier un chat sur une photo. À chaque couche du réseau neuronal correspond un aspect particulier de l'image

Deep Learning

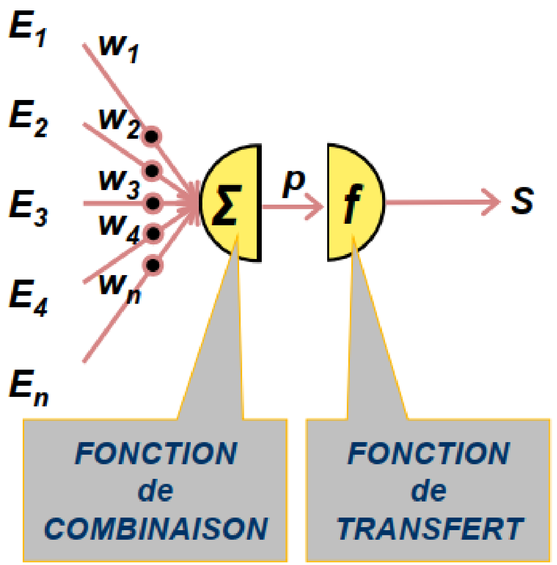
- Le « *deep learning* » peut, par exemple, aider à :
  - mieux reconnaître des objets hautement déformables ;
  - analyser les émotions révélées par un visage photographié ou filmé ;
  - analyser les mouvements et position des doigts d'une main, ce qui peut être utile pour traduire les langues signés ;
  - améliorer le positionnement automatique d'une caméra, etc. ;
  - poser, dans certains cas, un diagnostic médical (ex. : reconnaissance automatique d'un cancer en imagerie médicale), ou de prospective ou de prédiction (ex. : prédiction des propriétés d'un sol filmé par un robot) ;
  - reproduire une œuvre artistique à partir d'une photo à l'ordinateur.

Deep Learning

- ***Conduite automatisée :***
  - Les chercheurs du secteur automobile ont recours au Deep Learning pour détecter automatiquement des objets tels que les panneaux stop et les feux de circulation. Le Deep Learning est également utilisé pour détecter les piétons, évitant ainsi nombre d'accidents.
- ***Aérospatiale et défense :***
  - Le Deep Learning sert à identifier des objets à partir de satellites utilisés pour localiser des zones d'intérêt et identifier quels secteurs sont sûrs ou dangereux pour les troupes au sol.
- ***Recherche médicale :***
  - À l'aide du Deep Learning, les chercheurs en cancérologie peuvent dépister automatiquement les cellules cancéreuses. Des équipes de l'Université de Californie à Los Angeles (UCLA) ont conçu un microscope qui génère un ensemble de données de grande dimension afin d'entraîner une application de Deep Learning à identifier avec précision des cellules cancéreuses.
- ***Automatisation industrielle :***
  - Le Deep Learning participe à l'amélioration de la sécurité des employés travaillant à proximité d'équipements lourds, en détectant automatiquement les situations dans lesquelles la distance de sécurité qui sépare le personnel ou les objets des machines est insuffisante.
- ***Électronique :***
  - Le Deep Learning est utilisé pour la reconnaissance audio et vocale. Par exemple, les appareils d'assistance à domicile qui répondent à votre voix et connaissent vos préférences fonctionnent grâce à des applications de Deep Learning.

Réseau de neurones

Système nerveux	Système de calcul
Neurone	Processeur
Dendrite	Fonction de combinaison
Corps du neurone	Fonction de transfert
Axone	Elément de sortie
Synapse	Poids



Les éléments constitutifs du neurone artificiel :

- Les entrées "E" du neurone proviennent soit d'autres éléments "processeurs", soit de l'environnement.
- Les poids "W" déterminent l'influence de chaque entrée.
- La fonction de combinaison "p" combine les entrées et les poids.
- La fonction de transfert calcule la sortie "S" du neurone en fonction de la combinaison en entrée

- Exemples d'applications :
  - Diagnostique d'une maladie,
  - Conduite autonome,
  - Reconnaissance de formes.

Règle d'association (associated rules)

- Une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre deux ou plusieurs variables stockées dans de très importantes bases de données.
- Les règles d'association ont plusieurs applications et sont largement utilisée pour aider à découvrir les corrélations des ventes dans les données transactionnelles ou dans les ensembles de données médicales.
- Exemple d'application : découvrir les corrélations entre les produits achetés dans les magasins (constater que les clients qui achètent des tomates achètent aussi des pâtes)

esprit

Ensembles

- Les méthodes d'ensemble utilisent plusieurs algorithmes d'apprentissage pour obtenir de meilleures performances prédictives que celles obtenues avec l'un des algorithmes d'apprentissage (combinaison selon les hypothèses).

- Exemples d'applications : Télédétection, Cartographie de la couverture terrestre, Détection de changement, Détection de Malware, Détection d'intrusion, Reconnaissance faciale, Reconnaissance des émotions, ...

esprit

« You can have data without information, but you cannot have information without data »

Daniel Keys Moran
Computer programmer and science fiction author

esprit