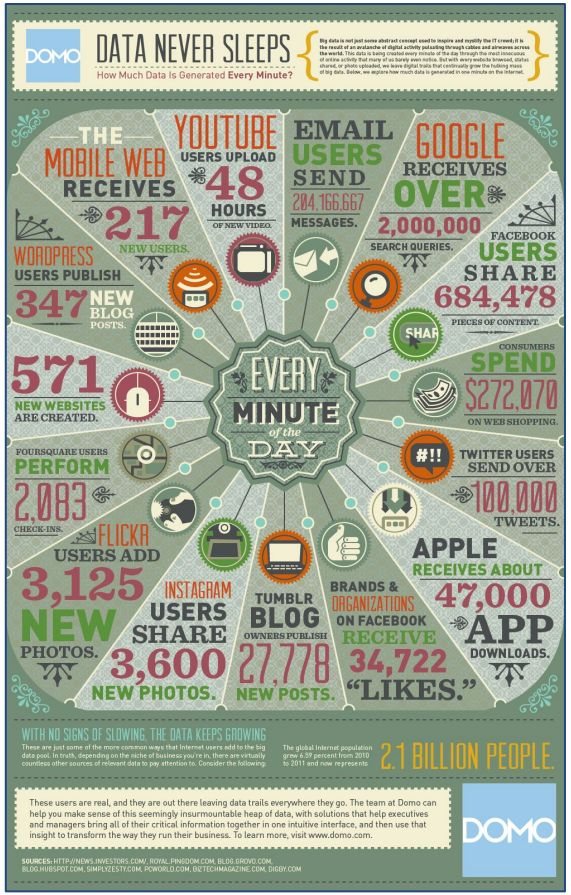
Machine Learning

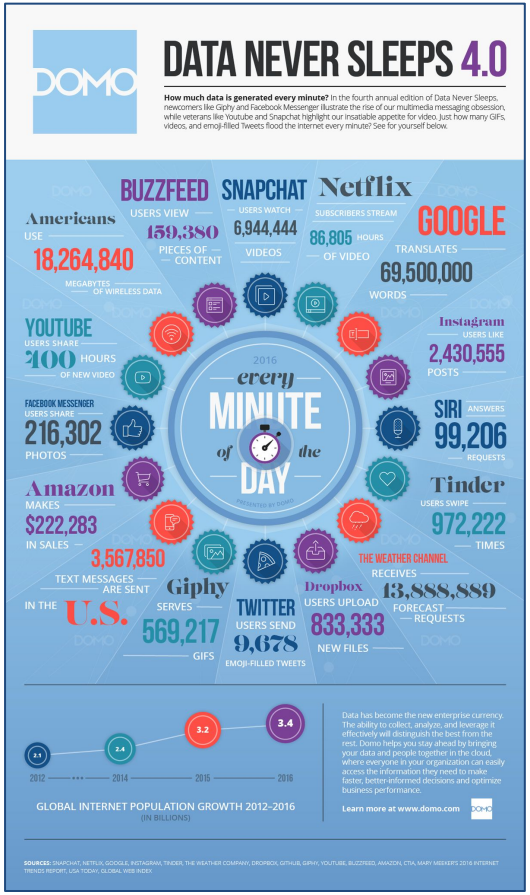
Manuel Simoes
manuel.simoes@cpc-analytics.fr

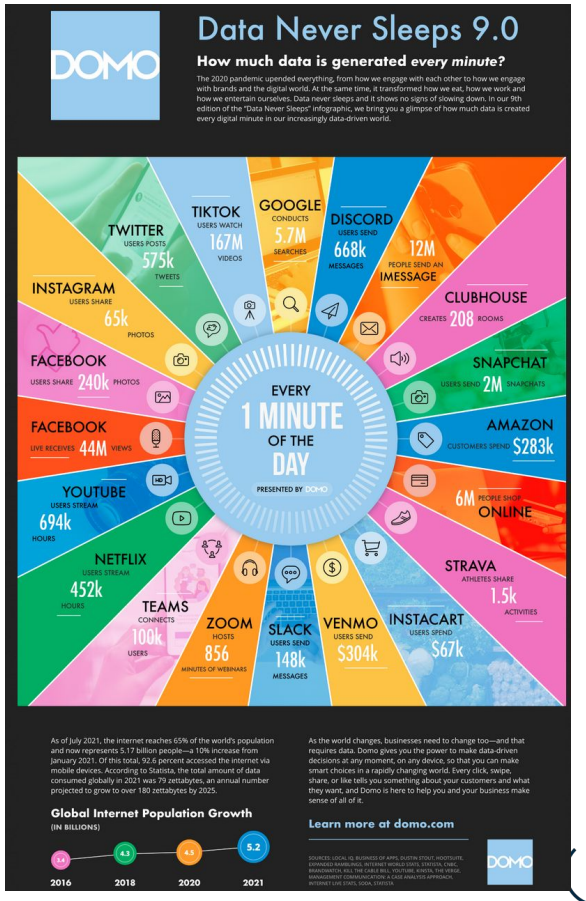
- Introduction -

Un monde de données

Toujours plus de données depuis le Net

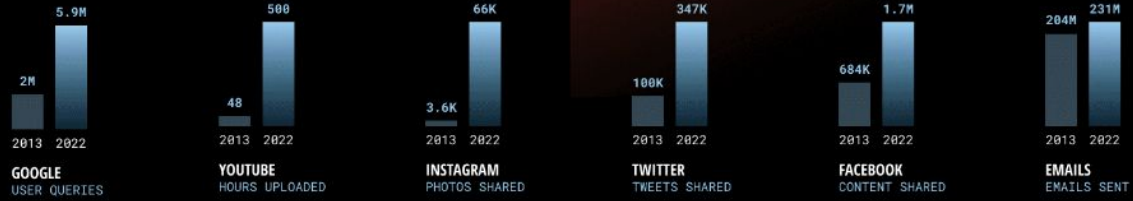


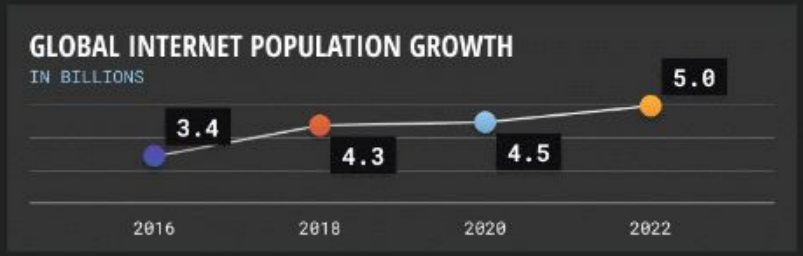




Toujours plus de données depuis le Net

DATA NEVER SLEEPS 1.0 VS. 10.0

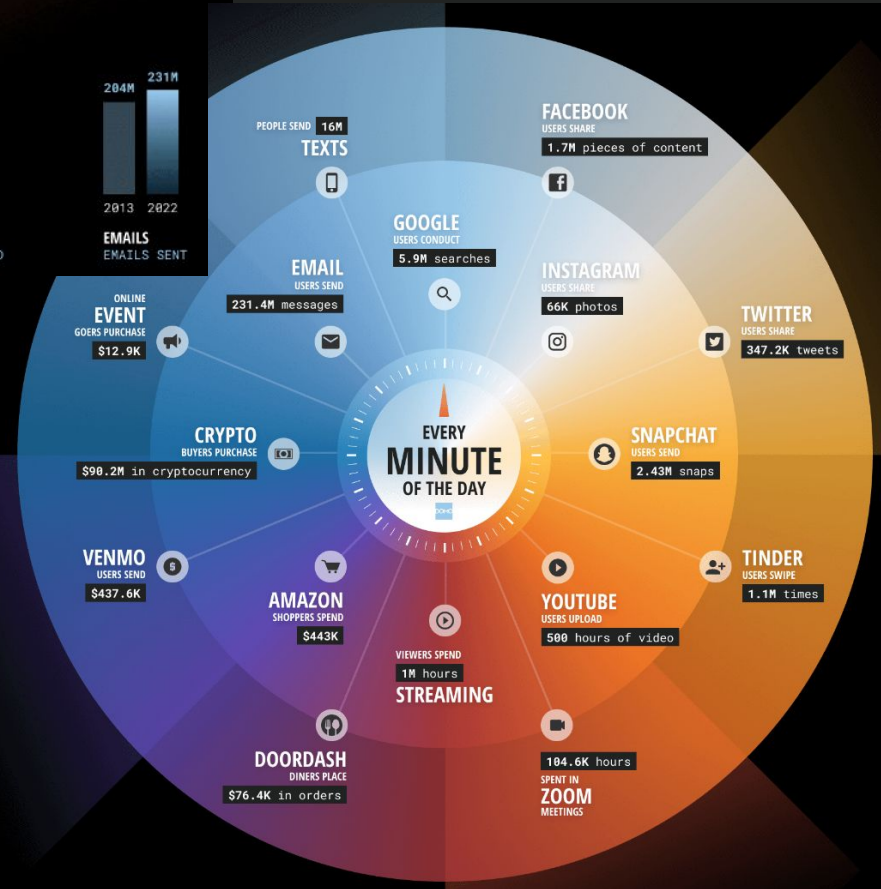




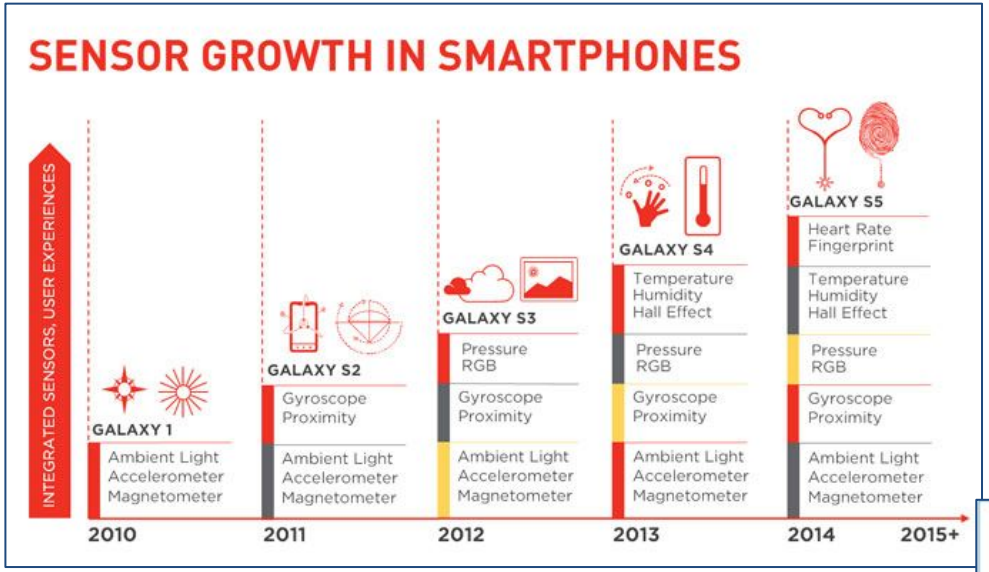
As of April 2022, the internet reaches 63% of the world's population, representing roughly 5 billion people. Of this total, 4.65 billion - over 93 percent - were social media users. According to Statista, the total amount of data predicted to be created, captured, copied and consumed globally in 2022 is 97 zettabytes, a number projected to grow to 181 zettabytes by 2025.

SOURCES

Global Media Insight, Oberlo, Hootsuite, Earthweb, Matthew Woodward.co.uk, Web Tribunal, Deadline.com, Local IQ, Business of Apps, Query Sprout, Young and the Invested, Dating Zest, IBIS World, DoorDash, TechCrunch, Statista, Data Never Sleeps 1.0



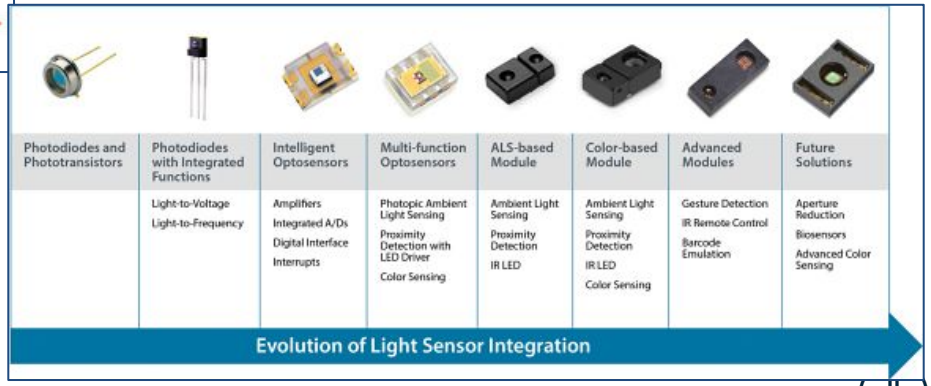
Toujours plus de données depuis les smartphones



Source_: ttps://www.qualcomm.com/news/ong/2014/04/24/behind-sixth-sense-smartphones-snapdrag on-processor-sensor-engine

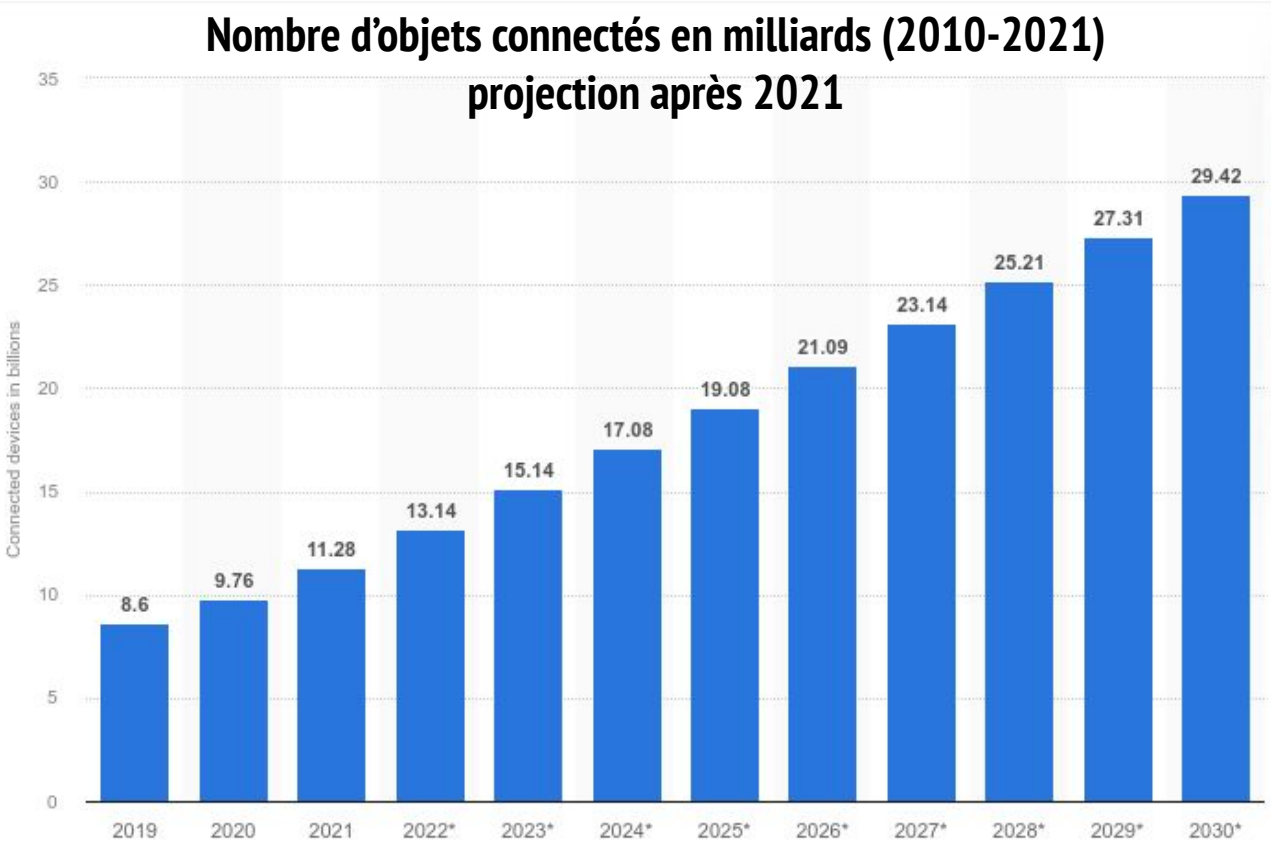
Les téléphones ont également de plus en plus de capteur pour “améliorer la qualité” de l’expérience utilisateur. Ces mêmes capteurs continuent, eux aussi, à s’améliorer. Plus précis, moins énergivore…

Les caméras, mais aussi les contenus de vos échanges, sont également des données pour les applications pour mieux vous connaître (votre influence dans votre réseau, votre humeur…).



Source : https://www.sensorsmag.com/components/smartphone-sensor-evolution-rolls-rapidly-forward

Toujours plus de données depuis votre environnement



Des caméras de surveillance aux maisons intelligentes de nombreux appareils se connectent sur internet.

Les efforts se concentrent aussi bien sur l'augmentation du débit des connexions sans fils que sur la distribution la plus large possible du réseau.

Exemples : machine à café, photocopieuse, porte, fenêtre, voiture automatique...

Source : https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

Différent types de données

Il existe de nombreuses appellations pour différents types de données mais cela est aussi fonction du milieu professionnel dans lequel vous travaillez [les données peuvent être dans différent ensemble en même temps].

Operational Data

Real Time Data

Donnée livrée et à traiter en temps réel. Donnée provenant de capteur ou livré depuis le web/application/utilisateurs. Apprentissage continu...

Translytic Data

Une vision business de la donnée qui est analysée en temps réel et à la demande depuis sa source.

Time-stamped data

BIG DATA

5 V : Volume, vélocité, variété, véracité...

Spatiotemporal data

Machine data

Données générées automatiquement par l'activité et les opérations de mise en connexion des appareils réseaux (incluant ordinateur, smartphone, IOT...)

Dark data

Données non utilisées, non classées ou identifiées détenu par les entreprises.

Unverified outdated data

Genomic data

Information provenant ou relevant des génomes (séquences...)

Umbalanced data

Rapport très défavorable entre 2 ou plusieurs populations dans les données. Exemples Transactions bancaires ok versus Transactions frauduleuses; défaut d'un semi-conducteur...

High-dimensional data

Open data

Données en accès libre, livré par une structure publique, associative ou privée (souvent déjà anonymisé).

Pour un ordinateur toutes les données devront être convertie en valeur (réelle ou entière) mais la provenance des données ainsi que ce que l'on veut en obtenir, va définir l'ensemble dans lequel elles appartiennent.

Source https://www.forbes.com/sites/adrianbridgwater/2018/07/05/the-13-types-of-data/

Les données

Données Structurées

Elles répondent à un schéma précis et connu:

- Bases de données relationnelles
- Fichiers textuels standards (csv, xml, json...)

L'intégrité de ce schéma peut être vérifiée à des contraintes qu'on définit préalablement. L'avantage de ces données est qu'elles sont faciles à manipuler et traiter.

	A	B	C	D	E	F	G	H	I
1	NumFact	Client	Date	Montant	Commercial	Zone	Type client	Ville	CP
2	10001	OBJECTIF PECHE	26/01/2016	4 111,00 €	Marcel DUBROCHET	Sud	A	Foix	09000
3	10002	PECHE DISCOUNT	27/01/2016	3 526,57 €	Gérard CARPE	Sud	B	Foix	09000
4	10003	PECHISSIMO	28/01/2016	1 362,67 €	Gérard CARPE	Est	B	Foix	09000
5	10004	HEXAGONE PECHE	28/01/2016	2 353,01 €	Joël VAIRON	Ouest	A	Toulouse	31000
6	10005	MFD SARL (Manufrance)	29/01/2016	4 281,09 €	Gérard CARPE	Sud	A	Toulouse	31000
7	10006	AU MOHICAN - Maison Perrot Audet	29/01/2016	1 395,12 €	Gérard CARPE	Nord	A	Toulouse	31000
8	10007	GO SPORT	01/02/2016	849,44 €	Joe BARBEAU	Ouest	A	Toulouse	31000
9	10008	LIBERTY-PECHE	01/02/2016	825,33 €	Odile DERAY	Sud	A	Toulouse	31000
10	10009	AQUA LOISIRS	02/02/2016	2 481,90 €	Joël VAIRON	Nord	A	Toulouse	31000
11	10010	AU MARCHE DU PECHEUR	02/02/2016	732,81 €	Odile DERAY	Nord	A	Toulouse	31000
12	10011	AU PECHEUR ROANNAIS	03/02/2016	131,52 €	Joël VAIRON	Nord	C	Toulouse	31000
13	10012	ACQUA PECHE 69	04/02/2016	2 093,15 €	Joël VAIRON	Nord	A	Toulouse	31000
14	10013	PROVENT PECHE EUROPECHE	06/02/2016	1 235,50 €	Fred LABLETTE	Est	A	Foix	09000
15	10014	LAVAL PECHE	07/02/2016	330,01 €	Joël VAIRON	Sud	A	Toulouse	31000
16	10015	SAPEC PECHE CHASSE	07/02/2016	4 552,45 €	René GARDON	Est	A	Foix	09000
17	10016	PINEAU SA	07/02/2016	879,52 €	Marc CHEVESNE	Est	C	Foix	09000
18	10017	PECHE TABAC LOTO MATEL	07/02/2016	2 951,18 €	Gérard CARPE	Est	A	Foix	09000

Données Semi-Structurées

Elles répondent à un schéma précis, mais sans obligation de s'y contraindre.

- Fichiers XML, CSV, JSON
- Bases de données NoSQL

S'affranchir partiellement d'un schéma facilite l'évolution du modèle de données ; en contrepartie, on a moins de contrôle sur l'intégralité des données.

Les données

Données Non-Structurées

Elles sont généralement plus complexe à traiter car elles peuvent prendre des formes différentes :

- Enregistrement médias : photos, images, vidéos et enregistrements audio
- Données textuelles non structurées : e-mails, transcriptions de texte, journaux applicatifs
- Activités sur les réseaux sociaux
- Données remontées par les capteurs électroniques et objets connectés.

	D	E	F	G	H	I
	Montant	Commercial	Zone	Type client	Ville	CP
/2016	4 111,00 €	Marcel DUBROCHET	Sud	A	Foix	09000
/2016	3 526,57 €	Gérard CARPE	Sud	B	Foix	09000
/2016	1 362,67 €	Gérard CARPE	Est	B	Foix	09000
/2016	2 353,01 €	Joël VAIRON	Ouest	A	Toulouse	31000
/2016	4 281,09 €	Gérard CARPE	Sud	A	Toulouse	31000
/2016	1 395,12 €	Gérard CARPE	Nord	A	Toulouse	31000
/2016	849,44 €	Joe BARBEAU	Ouest	A	Toulouse	31000
/2016	825,33 €	Odile DERAY	Sud	A	Toulouse	31000
/2016	2 481,90 €	Joël VAIRON	Nord	A	Toulouse	31000
/2016	732,81 €	Odile DERAY	Nord	A	Toulouse	31000
/2016	131,52 €	Joël VAIRON	Nord	C	Toulouse	31000
/2016	2 093,15 €	Joël VAIRON	Nord	A	Toulouse	31000
/2016	1 235,50 €	Fred LABLETTE	Est	A	Foix	09000
/2016	330,01 €	Joël VAIRON	Sud	A	Toulouse	31000
/2016	4 552,45 €	René GARDON	Est	A	Foix	09000
/2016	879,52 €	Marc CHEVESNE	Est	C	Foix	09000
/2016	2 951,18 €	Gérard CARPE	Est	A	Foix	09000

Une **constitution** est une loi fondamentale ou un ensemble de principes[1] qui fixe l'organisation et le fonctionnement d'un organisme, généralement d'un État ou d'un ensemble d'États[réf. incomplète].
La valeur de la Constitution d'un État varie selon le régime en place, elle a généralement une valeur supérieure à la loi. Elle est à la fois l'acte politique et la loi fondamentale qui unit et régit de manière organisée et hiérarchisée l'ensemble des rapports entre gouvernants et gouvernés au sein de cet État, en tant qu'unité d'espace géographique et humain. La Constitution protège les droits et les libertés des citoyens contre les abus de pouvoir potentiels des titulaires des pouvoirs (exécutif, législatif, et judiciaire).
Si la fiction juridique veut que la Constitution fonde et encadre juridiquement l'État, il est entendu que l'histoire politique la précède et peut lui conférer à la fois sa légitimité circonstanciée et la permanence de son autorité.

La complexité de traitement de ces données tient surtout à leur absence de structure et à leur format souvent binaire (surtout les données médias). Il faut donc les traiter au cas par cas pour en tirer de la valeur ajoutée.

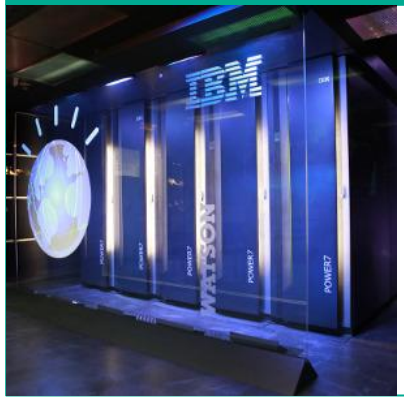
Historiquement, il était difficile pour les algo de traiter les données non structurées, là où performe l'humain. De grands progrès ont été réalisés avec le deep learning.

- Introduction -

L’apprentissage artificielle
&
L’Intelligence Artificielle

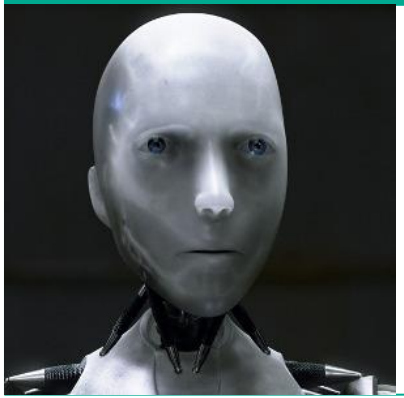
Intelligence artificielle

IA FAIBLE | SPÉCIALISATION



Reconnaissance de visages, jeux…
Milieu actuel du Machine Learning

IA FORTE | « CONSCIENCE »



L'informatique, pas assez fine pour simuler les mécanismes naturels et faire émerger une intelligence ?

Les recherches en intelligence artificielle ont permis de faire émerger d'autres domaines : web, langage objet, ...

Apprentissage artificiel

Une définition de l'apprentissage artificielle donné par *Tom Mitchell* en *1997*

Étant donné :

de l'expérience E,

une classe de tâches T

et une mesure de performance P

On dit d'un ordinateur qu'il apprend si :

sa performance sur une tâche de T

mesurée par P

augmente avec l'expérience E

Cette formulation définit la procédure de mesure de l'apprentissage d'un algorithme

L'intelligence artificielle à travers le jeu

Hypothèse : L'intelligence peut se décomposer en fonctions cognitives élémentaires que l'on peut simuler sur ordinateur

1962 | A. SAMUEL

1997 | BIG BLUE (IBM)

2011 | WATSON (IBM)

2016 | ALPHA GO (GOOGLE)

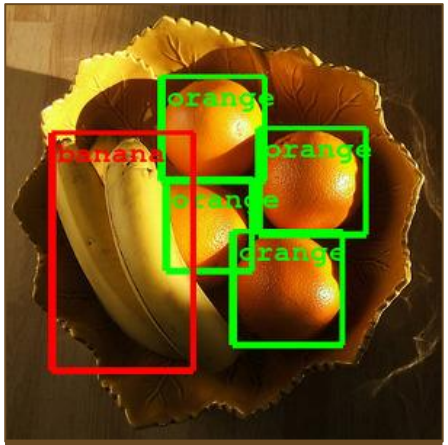
2019 | DEEPMIND

« Donner la capacité aux machines d'apprendre sans les programmer explicitement. » – *Arthur Samuel, 1959*

Quand / Pourquoi utiliser l'IA ?

Navigation sur Mars ...

L'EXPERTISE HUMAINE N'EXISTE PAS

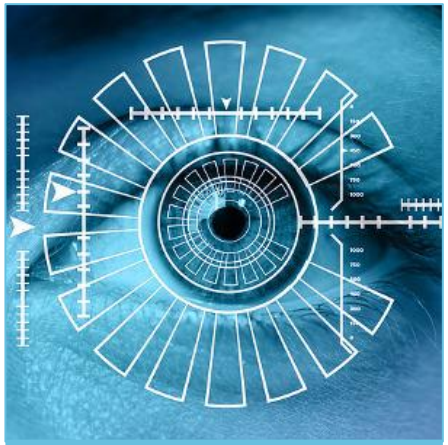


Reconnaissance d'image, de voix ...

INCAPACITÉ À TRANSMETTRE SON EXPERTISE

Trading, routing ...

DES SOLUTIONS QUI VARIENT DANS LE TEMPS

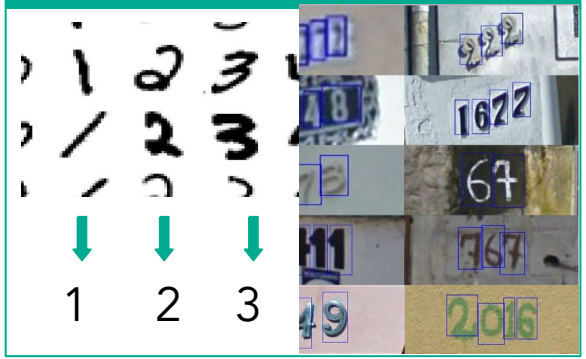


Customisation, biométrique ...

S'ADAPTER AU CAS PARTICULIER

Apprentissage artificiel : exemple

RECONNAISSANCE DE CARACTÈRES



Reconnaître les caractères manuscrit ainsi que ceux présent sur des photos

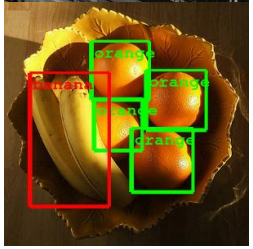
COMPORTEMENT D'UN ROBOT AUTONOME

Navigation « à vue » optimisant l'accomplissement d'une tâche. La distance entre le robot et la terre est trop grande pour qu'un humain conduise le robot en temps réel.

Termes associés à l'apprentissage artificiel

Robotique
conduite automatique, robots ...

Prédictions / prévisions
bourse, pics de pollution …

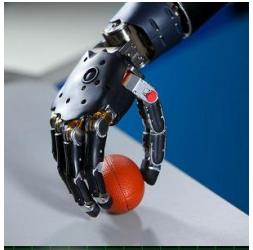


Reconnaissance
visage, voix, écriture, mouvement ...

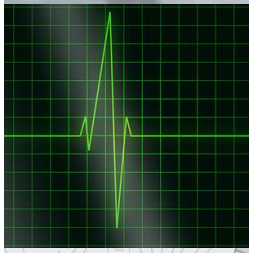
Adaptation
préférences utilisateur, robot sur terrain accidenté ...

Régulation
trafic, chauffage, température du frigo …

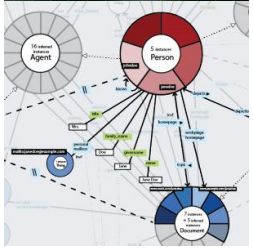
Optimisation
vitesse du métro, voyageur de commerce, recherche ...



Autonomie
robots, prothèses de main ...



Traitement du signal

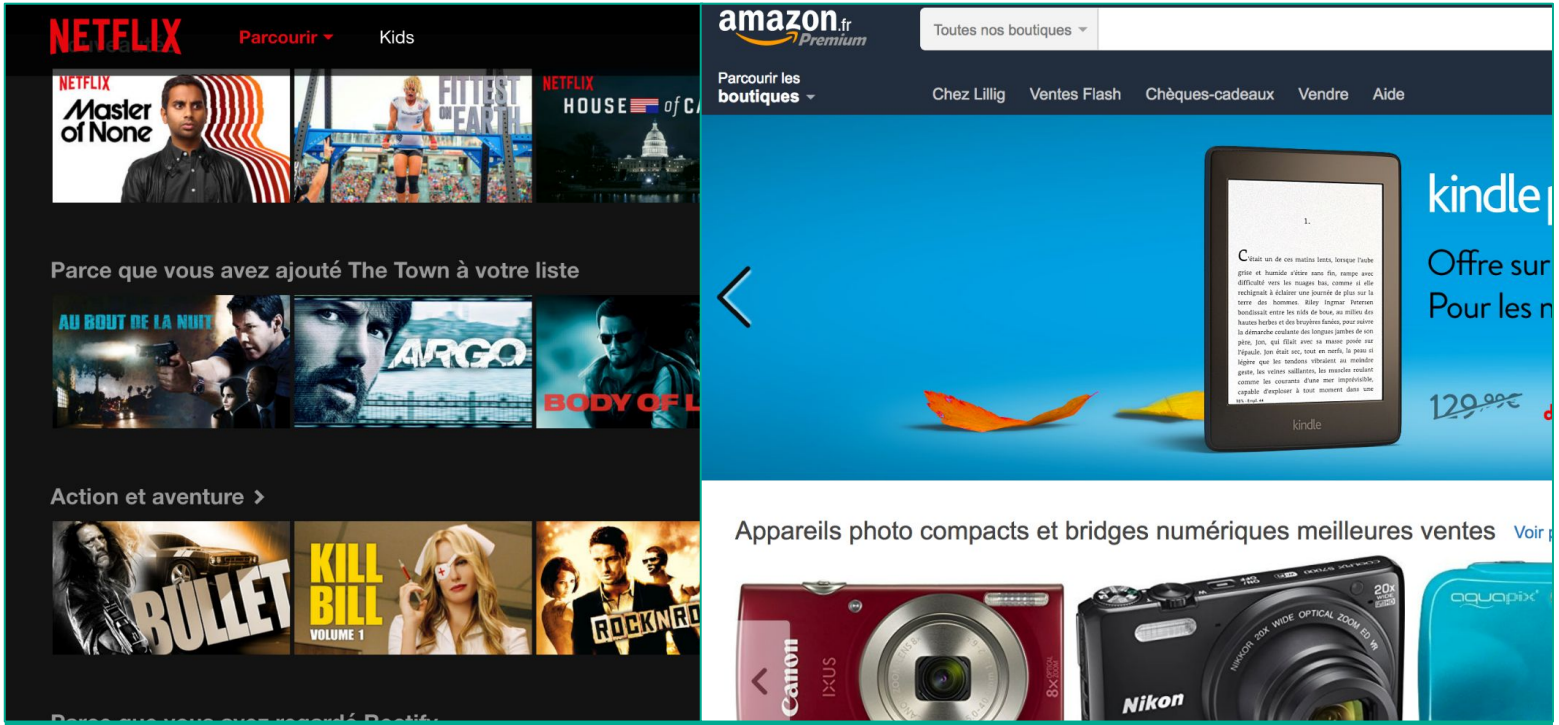


Représentation des connaissances

Apprentissage artificiel : Un Business

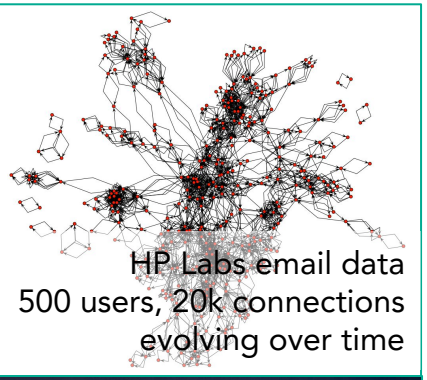
75 % du contenu consommé

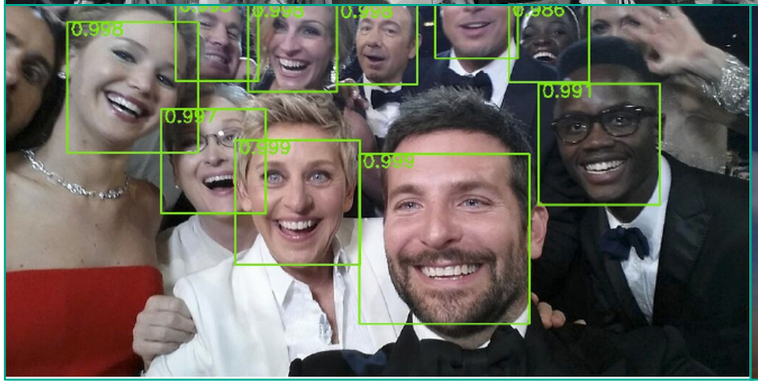
30 % de son chiffre d'affaires (2017)



PERSONNALISER LA RELATION CLIENT

Apprentissage artificiel : exemple





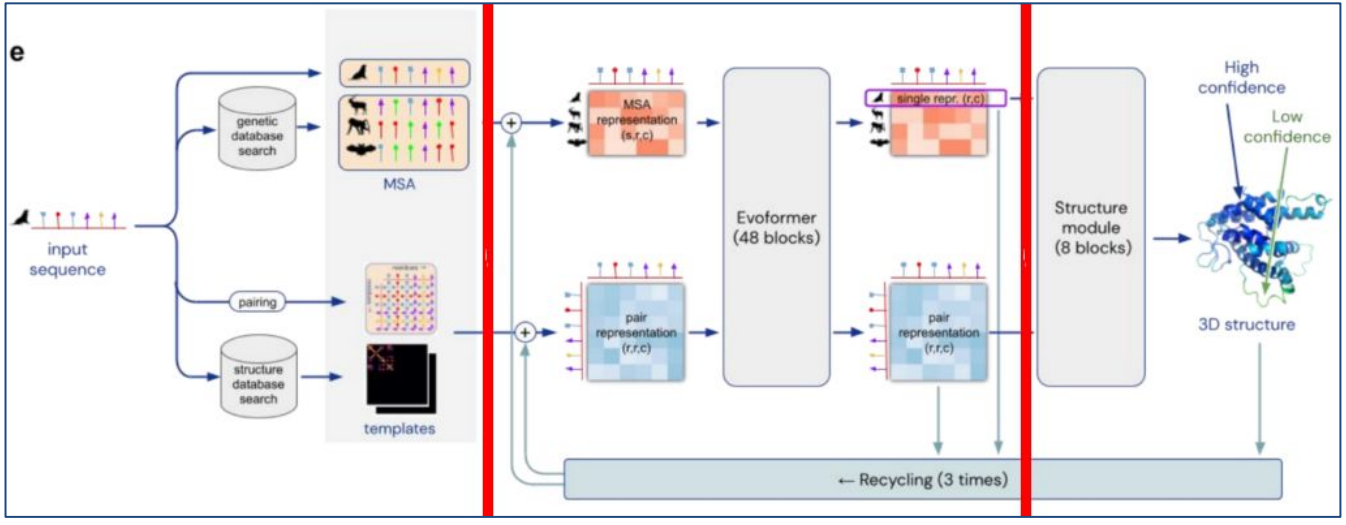
Apprentissage par imitation / démonstration

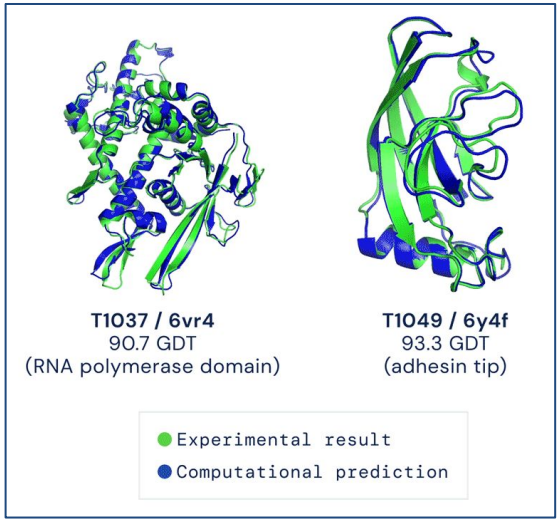
Apprentissage procédural (précision motrice)

Reconnaissance d'objets

AlphaFold1-2 (Google)

AlphaFold1-2 : Détermination de la structure 3D des protéines

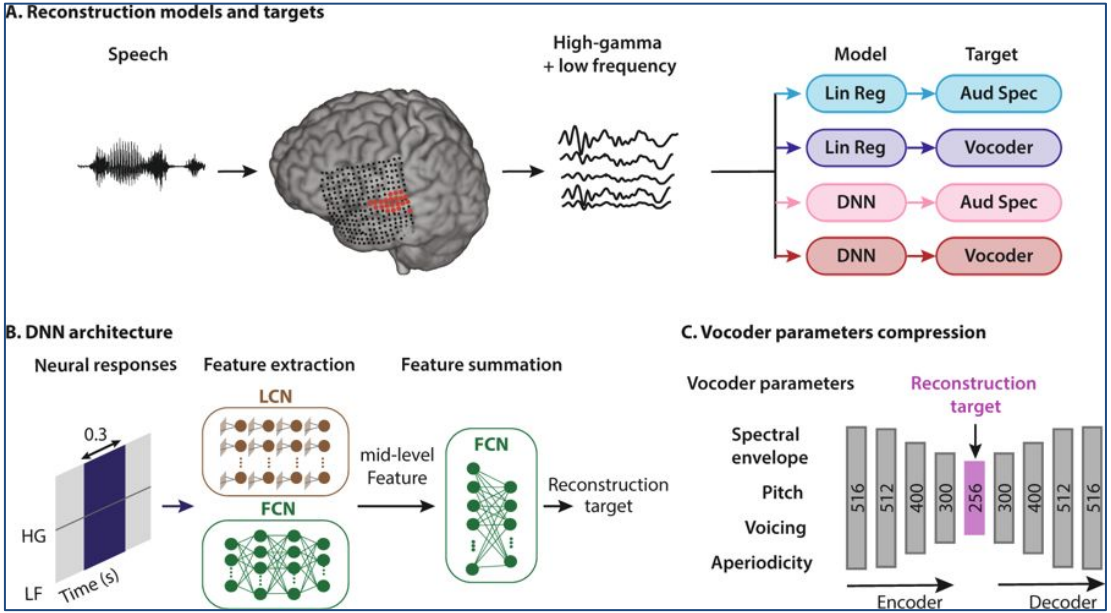




Code source : https://github.com/deepmind/alphafold (utilisation en ligne)
Publication : Highly accurate protein structure prediction with AlphaFold, *Nature* **volume 596**, pages 583–589 (2021).
Autre description de l'algorithme : Oxford Protein Informatics Group

Exemple : Reconstruire la voix depuis le cortex

“Towards reconstructing intelligible speech from the human auditory cortex”



On lit un texte au patient que l’on essaie de reconstituer depuis des électrodes connectées au cortex du cerveau et un système de réseau de neurone artificielle.

Ici un des objectifs affiché est de créer de nouveaux outils de communication pour les muets.

Source : Hassan Akbari, Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta & Nima Mesgarani, 27 janvier 2019 Nature 874 (2019)

Les différents types d'apprentissage artificiel

APPRENTISSAGE SUPERVISÉ / SUPERVISED LEARNING

Nous utilisons des données d'entraînement avec la sortie (la valeur à prédire)

APPRENTISSAGE NON SUPERVISÉ / UNSUPERVISED LEARNING

Les données d'entrainements ne contiennent pas la sortie (la valeur à prédire)

APPRENTISSAGE SEMI-SUPERVISÉ / SEMI-SUPERVISED LEARNING

Certaines données d'entraînement contiennent la valeur à prédire.

APPRENTISSAGE PAR RENFORCEMENT / REINFORCEMENT LEARNING

Continuer à utiliser des données d'entraînement et par intermittence pour renforcer l'apprentissage.

Les différents types d'apprentissage artificiel

L'apprentissage supervisé est une forme d'apprentissage automatique où l'on fournit à l'algorithme un ensemble d'observations (lignes) décrites par des variables explicatives (colonnes/features) ainsi que la variable cible (la réponse / l'étiquette). L'algorithme d'apprentissage supervisé va approcher une fonction de prédiction qui permet un *mapping* depuis les données des observations vers la réponse.

L'apprentissage non supervisé correspond au cas où les observations du jeu de données ne sont pas étiquetées (elles ne disposent pas de leurs variables de sortie). L'algorithme est livré à lui-même et va apprendre les structures et les relations caractérisant le jeu de données. Ces algorithmes d'apprentissage tirent profit de ces jeux de données en combinant les approches supervisée et non supervisée.

L'apprentissage semi-supervisé, dans certains jeux, les données sont partiellement étiquetées : certaines observations disposent d'étiquettes et d'autres non. Généralement, la proportion de ces dernières est beaucoup plus grande. Ce cas est fréquent ; en effet, le processus d'étiquetage requiert une expertise humaine et peut être long et coûteux. Ces algorithmes tirent profit de ces jeux de données en combinant les approches supervisée et non supervisée.

Les différents types d'apprentissage artificiel

L'apprentissage par renforcement (reinforcement learning), le système apprenant (appelé agent) évolue dans un environnement. L'agent effectue des actions et reçoit des récompenses (ou pénalité) selon les cas. Les actions peuvent avoir un poids de récompense différent. L'agent va apprendre une stratégie (politique) qui maximise sa récompense dans une situation donnée. L''ensemble de décisions que l'agent choisit dans les différentes situations représente son apprentissage (ou le modèle d'apprentissage)

Le système AlphaGO de DeepMind est un exemple d'apprentissage par renforcement.

L'apprentissage incrémental (apprentissage en ligne), lors de l'apprentissage incrémental on fournit les données à l'algorithme au fur et à mesure, de façon séquentielle, unité par unité ou bien sous forme de petits lots (mini-batchs). L'algorithme d'apprentissage adapte le modèle prédictif à la volée sans nécessité de redéploiement. Il est à noter qu'une fois le jeu de données utilisé pour l'entraînement, on peut s'en débarrasser.

L'apprentissage en ligne est adapté dans le cas où le modèle prédictif doit s'adapter rapidement aux flux continus de données (comme le prix des actions en bourse) et également lorsque les ressources de calcul sont limitées (stockage, bande passante et puissance de calcul). L'une des limitations de ce mode est qu'il faut le surveiller continuellement. En effet, si les données fournies en continue deviennent de mauvaise qualité et non représentatives, le modèle s'adapte et sa qualité se dégrade . Il est donc important de surveiller continuellement, les données entrant dans l'algorithme d'apprentissage et les performances du modèle.

Les algorithmes

APPRENTISSAGE SUPERVISÉ

- Arbre de décision
- Rule induction
- Instance-based learning
- Réseau bayésien
- Réseaux de neurones
- Support vector machine
- Ensemble de modèles
- Learning theory
- Deep learning
- ...

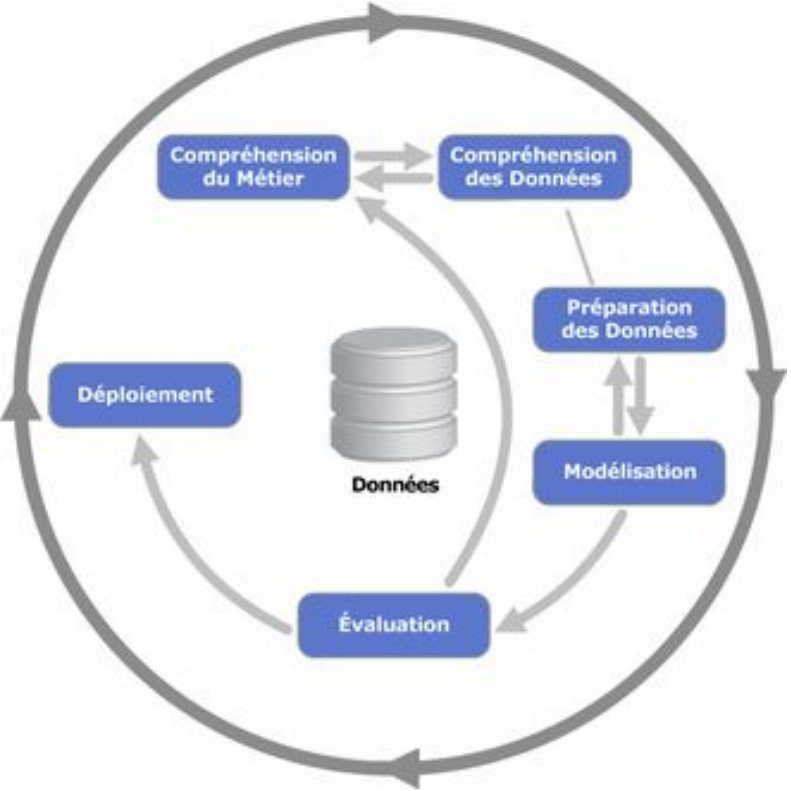
APPRENTISSAGE NON SUPERVISÉ

- Clustering
- Réduction des dimensions

- Introduction -

Sciences des données

Le Machine learning est un outil du **Data Mining**



Cross Industry Standard Process for Data Mining
(CRISP-DM ; Shearer 2000)

Le Data Mining implique l'utilisation d'une quantité importante de science et de technologie ainsi qu'un processus d'analyse, explorative et scientifique. Trois techniques sont communément utilisées,

- Le KKD (Knowledge Discovery in Databases)
- SEMMA (Sample, Explore, Modify, Model, Assess)
- CRISP-DM (Cross Industry Standard Process for Data Mining).

Le processus le plus communément pratiqué, dans les entreprises et les laboratoires est basé sur le
Cross Industry Standard Process for Data Mining (CRISP-DM).

C'est un processus bien défini qui structure chaque problème, assurant une part de cohérence raisonnable, de répétabilité et d'objectivité.

Data Mining

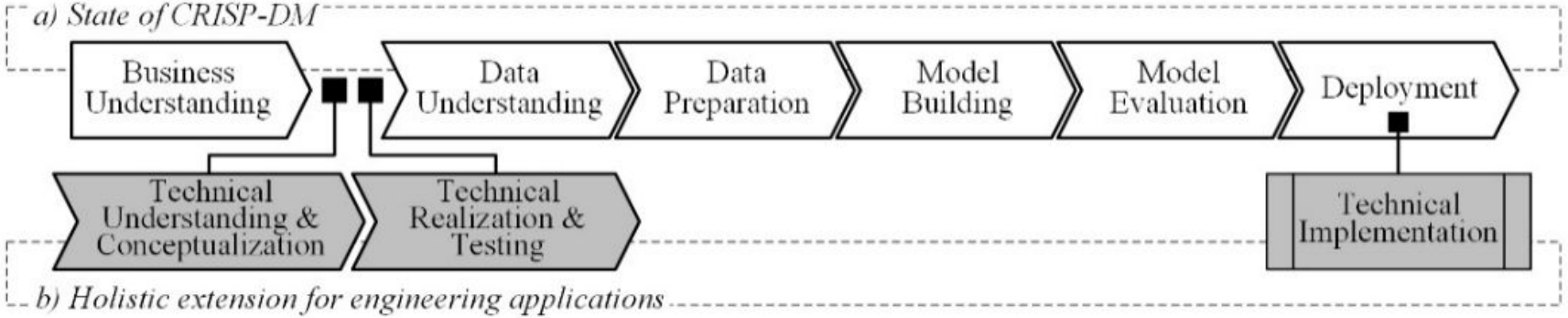
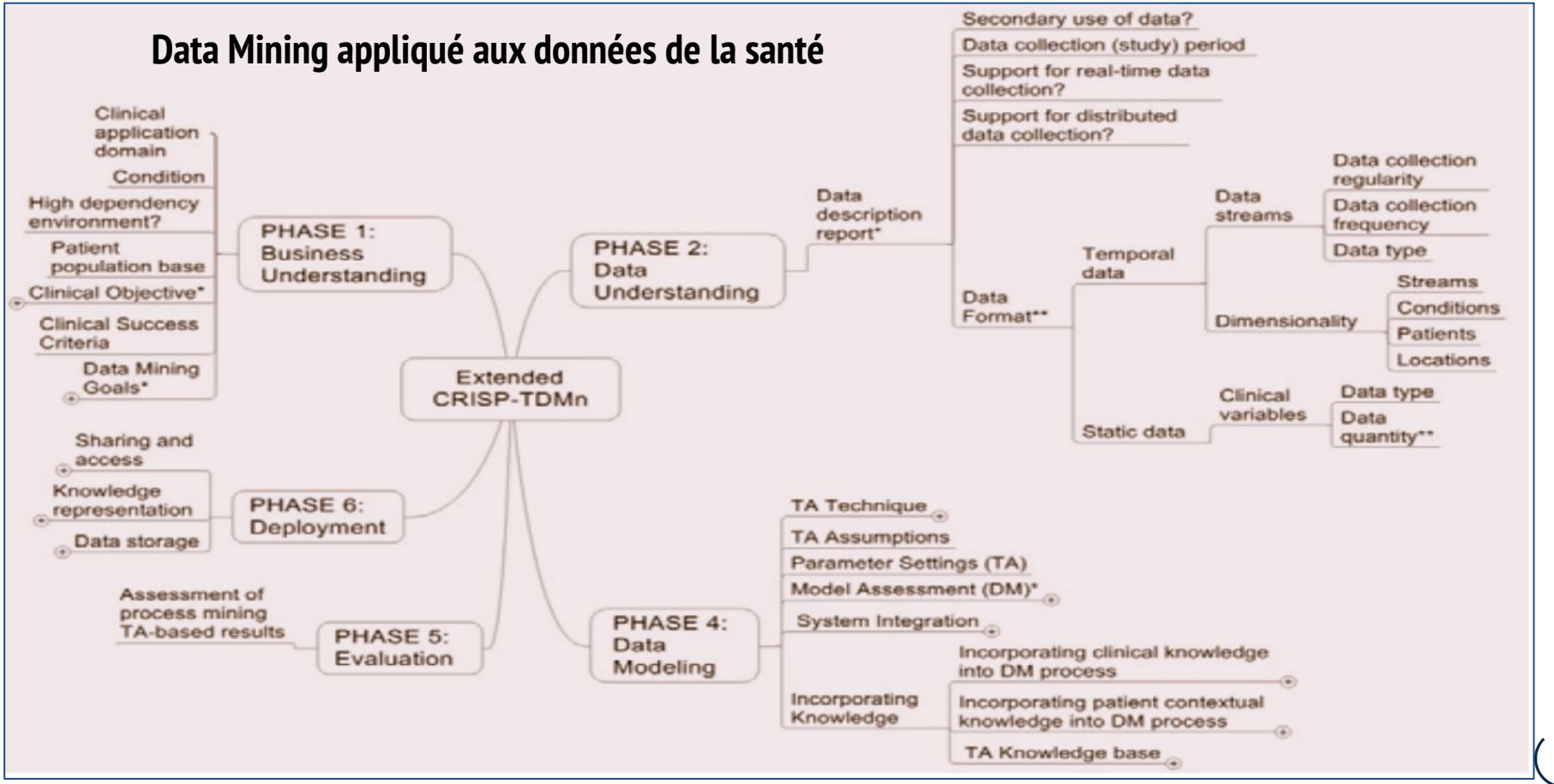


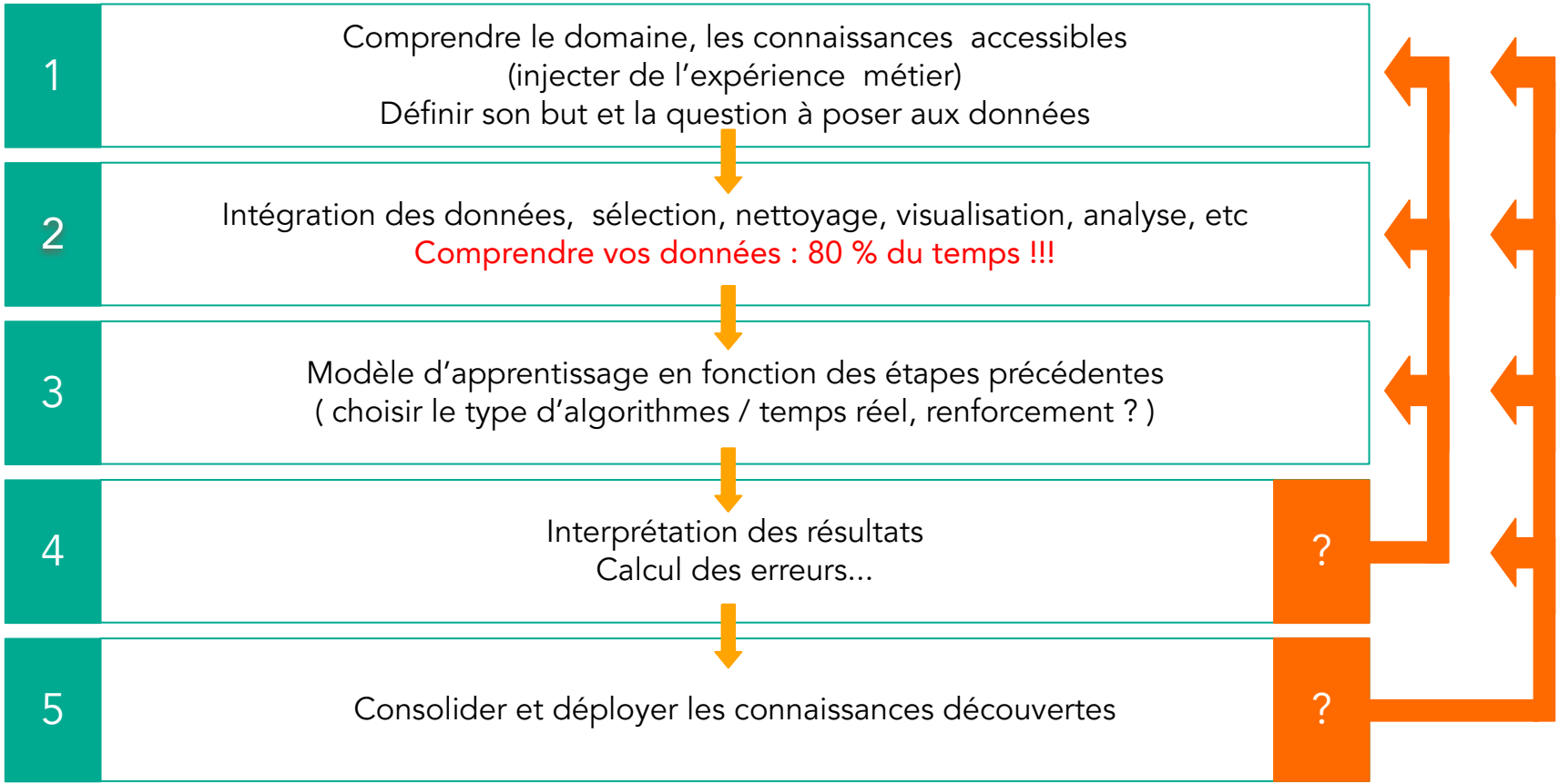
Figure 1. Phases of the reference model of DMME (data mining methodology for engineering applications). (**a**) State of CRISP-DM, (**b**) holistic extension for engineering application.

D'autres publications visent à améliorer les cycles de recherche d'entreprise ou le déploiement ainsi que l'apprentissage continu. Certains travaux visent à ajouter une « extension » au processus de Data Mining basé sur CRISP-DM pour intégrer les contraintes des chercheurs qui prennent en compte l'acquisition (en temps réel) des données et ses connexions aux applications déployées.

Data Mining : Données de la santé



Le Machine Learning dans la pratique

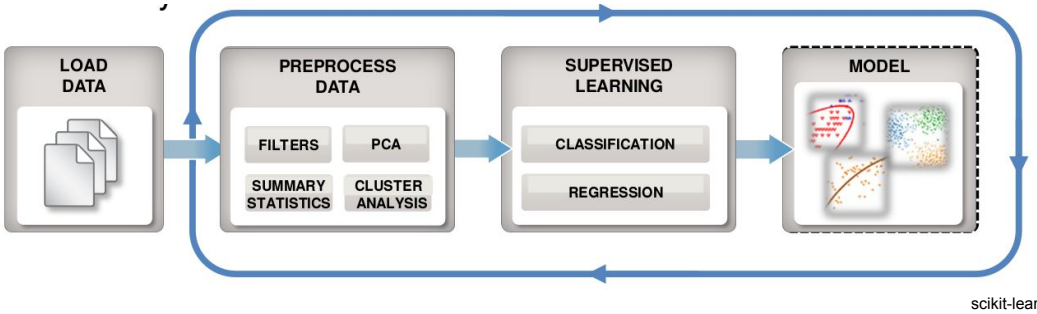


Méthodologie générale

destinées à l'apprentissage 70 % – 80 %	destinées aux tests 20 % – 30 %

DONNÉES « CLIENT »

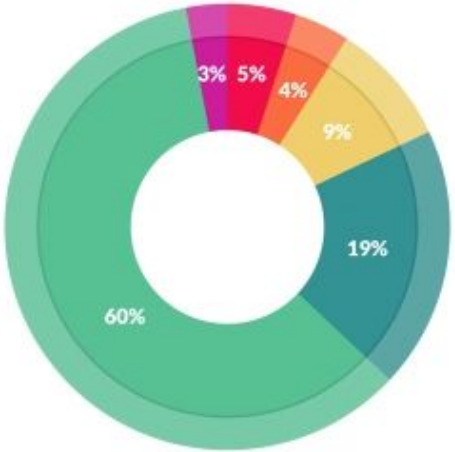
Les prédictions d'un « modèle » sont testées sur des données qui proviennent du même dataset mais qui n'ont jamais été utilisées dans l'étape de l'apprentissage



LEARNING / TRAINING

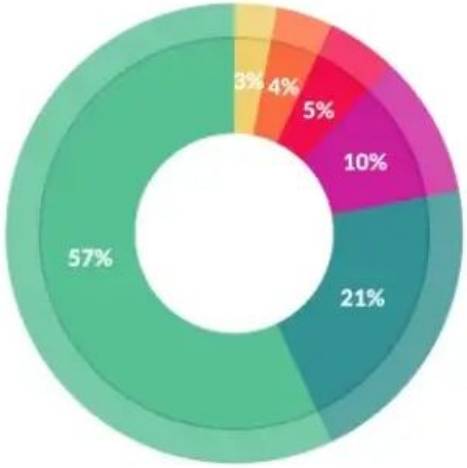
On fait des itérations jusqu'à obtenir le bon modèle

Méthodologie générale



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*



What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/