

Définition du big data

Clément Levallois

2017-31-07

Table of Contents

1. Le Big data est difficile à définir	1
2. Les 3 V	1
V pour le volume	1
V pour Variety	1
V pour Velocity	3
Un 4ème V peut être ajouté, pour Véracité ou Valeur	4
3. Quelle est la taille minimale à considérer comme "Big Data"? Tout est relatif	4
a. relative au temps	5
b. par rapport à l'industrie	5
c. pas seulement sur la taille	5
d. pas de corrélation entre la taille et la valeur	5
4. D'où vient le big data?	6
a. La numérisation de l'économie a généré de nouveaux volumes de données	6
b. Les ordinateurs sont devenus plus puissants	6
c. Le stockage des données est devenu moins cher chaque année	7
d. L'état d'esprit a changé sur ce qui "compte" comme données	8
e. Le logiciel open source accélère l'innovation	9
f. Les promesses et attentes exagérées sur le big data	9
g. Le Big Data transforme les industries et est devenu une industrie en soi	11
5. Quel est l'avenir du Big Data?	13
a. Plus de données arrivent	13
b. Les cadres réglementaires vont augmenter en complexité	13
c. en tant qu'expression, "big data" évolue	13
Pour aller plus loin	14



1. Le Big data est difficile à définir

image::ariely.png [align = "center", title="Message de Facebook de Dan Ariely en 2013", book = "keep"]

Blagues à part, la définition du big data demande un peu de précision. Apportons une certaine clarté.

2. Les 3 V

Les big data sont généralement décrites avec le "3 Vs":

V pour le volume

La taille des jeux de données disponibles aujourd'hui est stupéfiante (ex: Facebook avait 250 milliards de photos en 2016). Les volumes de données augmentent à un rythme **accéléré**. Selon des sources, **90% de toutes les données dans le monde a été généré au cours des deux dernières années** (déclaration de 2013) ou dit différemment, **plus de données être créé en 2017 que dans les 5000 ans de l'histoire de l'humanité**.

V pour Variety

"Variété" fait référence au fait que les données "non structurées" sont considérées comme de plus en plus utiles, alors qu'avant le big data, seules les données structurées valaient la peine d'être stockées et exploitées. Cela appelle à expliquer plus en détail la distinction entre données non structurées et structurées.

A - Données structurées

Données structurées se réfère à des données formatées et organisées selon un ensemble de règles bien défini, ce qui les rend **lisibles par une machine**. Par exemple, les codes postaux sont un ensemble de données structuré car ils suivent une convention précise concernant le nombre de lettres et de chiffres qui les composent, ce qui facilite l'identification et leur « lecture » par un lecteur optique et un logiciel. Pareil avec les plaques d'immatriculation, les numéros de sécurité

sociale ... Mais ce sont des exemples simples.

Qu'en est-il, par exemple, d'un formulaire d'impôt? Si chaque champ du formulaire est bien défini, alors les données collectées à travers le formulaire peuvent être dites « structurées ». En revanche, une forme où l'utilisateur peut écrire du texte libre (pensez à un commentaire sur un article de blog, ou un espace vide où les utilisateurs peuvent écrire un commentaire) produit des données non structurées : données qui ne suivent pas une convention spéciale pour leur taille et leur contenu. C'est généralement beaucoup plus difficile à traiter par le logiciel, donc à analyser.

Pour résumer, pensez aux données structurées comme à tout ce qui peut être représenté comme des tableaux de nombres bien organisés et de petits textes avec le format, la taille et les conventions d'écriture attendues: annuaires, livres comptables, statistiques gouvernementales ...

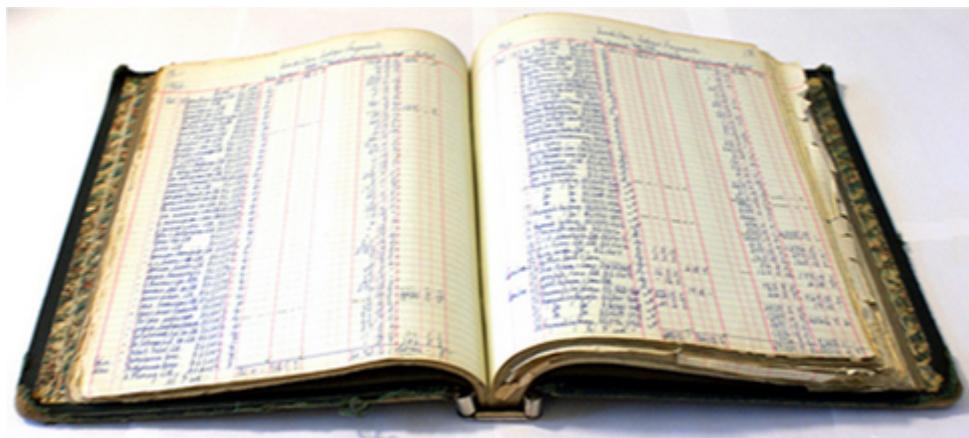


Figure 1. Un livre de comptes montrant des données structurées

B - Données non structurées

Données non structurées se réfère à des ensembles d'éléments "indisciplinés": texte de n'importe quelle longueur, sans catégorisation appropriée, codé dans différents formats, y compris éventuellement des images, du son, des coordonnées géographiques...

Ces ensembles de données sont beaucoup plus difficiles à traiter et à analyser, car ils sont pleins d'exceptions et de différences. Mais ils sont porteurs d'informations généralement riches: texte libre, informations enregistrées "dans la nature" ...

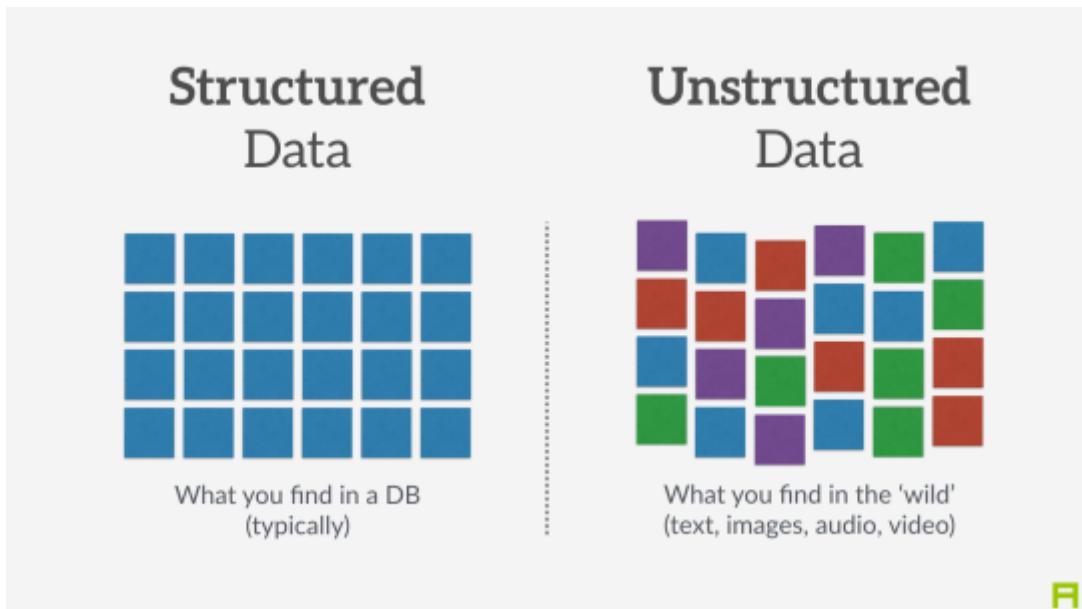


Figure 2. Données structurées vs données non structurées



V pour Velocity

En un mot, [la vitesse de création et de communication des données s'accélère](#) : - Facebook héberge 250 milliards de photos? Il reçoit 900 millions de photos supplémentaires **par jour** - Examiner les tweets peut être fait automatiquement (avec des ordinateurs). Si vous voulez vous connecter à Twitter pour recevoir des tweets en temps réel pendant qu'ils sont tweetés, préparez-vous à recevoir plus de 500 millions de tweets **par jour**. Twitter appelle ce service le "[Twitter firehose](#)", qui reflète la vélocité du flux de tweets.



Figure 3. Le firehose de Twitter

- **Les données de capteur** (données de capteur sont également susceptibles d'augmenter la vitesse. Alors que les images, les tweets, les enregistrements individuels ... sont des données mono-élément envoyées à intervalles réguliers, de plus en plus de capteurs peuvent envoyer des données **en flux continu** (mesures de mouvement, son, etc.)

Ainsi, la vitesse pose des défis qui lui sont propres : alors qu'un système peut gérer (stocker, analyser) 100 Go de données dans un temps donné (jour ou mois), il peut ne pas être capable de le faire en une seconde. Le big data fait référence aux problèmes et aux solutions que soulève la vitesse des données.

Un 4ème V peut être ajouté, pour Véracité ou Valeur

La véracité concerne la fiabilité et la conformité : les données sont-elles authentiques? Ont-elles été corrompues à n'importe quelle étape de leur traitement? Est-ce conforme aux réglementations locales et internationales?

3. Quelle est la taille minimale à considérer comme "Big Data"? Tout est relatif

Il n'y a pas de «seuil» ou de «taille minimale» d'un ensemble de données où les données passeraient de «petites données» à «grandes données». Il s'agit plutôt d'une notion **relative** : ce sont des

données massives si les systèmes informatiques actuels ont du mal à gérer les jeux de données qu'elles doivent traiter. "Big data" est donc une notion relative.

a. relative au temps

- Ce qui était considéré comme «Big Data» au début des années 2000 serait considéré comme « Small Data » aujourd’hui, car nous avons une meilleure puissance de stockage et de calcul.
- C'est une course sans fin : à mesure que les systèmes informatiques s'améliorent pour faire face aux « big data actuels », les données sont générées dans des volumes encore plus importants, ce qui nécessite de nouveaux progrès / innovations pour les gérer.

b. par rapport à l'industrie

- Ce qui est considéré comme « big data » par les PME non technologiques (petites et moyennes entreprises) peut être considéré comme insignifiant par les entreprises technologiques.

c. pas seulement sur la taille

- la difficulté pour un système informatique de faire face à un ensemble de données peut être liée à la taille (essayez d'analyser 2 Tb de données sur votre ordinateur portable ...), **mais aussi** liées au contenu des données.
- Par exemple, l'analyse des avis clients dans des dizaines de langues est plus difficile que l'analyse du même nombre de commentaires dans une seule langue.
- Donc, la règle générale est la suivante : moins les données sont structurées, plus elles sont difficiles à utiliser, même si elles sont de petite taille (cela concerne le « V » de la variété vu plus haut).

d. pas de corrélation entre la taille et la valeur

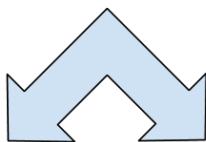
- "[Les big data sont souvent appelées le nouvel or noir](#)", comme si elles coulaient comme du pétrole et qu'on pouvait en servir à la pompe, tout simplement.
- En fait, le big data est **créé** : il faut du travail, un effort de conception et des choix à faire pour que les données viennent à exister (que dois-je collecter, comment le stocker, quelle structure lui donner?). L'intervention humaine dans la création de données détermine en grande partie si les données seront utiles plus tard.
- Exemple: Imaginons que des clients puissent écrire des critiques en ligne de vos produits. Ces avis sont des données. Mais si ces avis sont stockés sans indiquer qui est l'auteur de la critique (peut-être parce que les avis peuvent être publiés sans se connecter), les avis deviennent beaucoup moins utiles.

Les décisions de conception simples sur la façon dont les données sont collectées, stockées et structurées ont un impact énorme sur la valeur des données. Ainsi, en réaction à des ensembles de données volumineux, non structurés et mal organisés et de faible valeur, on avance parfois la notion de « [données intelligentes](#) » smart data : des données de petite taille mais bien organisées et [annotées](#), qui en valorisent la valeur.

4. D'où vient le big data?

a. La numérisation de l'économie a généré de nouveaux volumes de données

Data generated by a movie-goer:



In a movie theater:



on the box office ticket:
movie title,
date,
price

On Netflix:



Login to Netflix: age, name, gender, location + preferences for movie genres?

Browsing / purchasing history for movies

Movie title, date and price for the movie

Date and time on movie started / paused / interrupted / finished

Comments / ratings posted

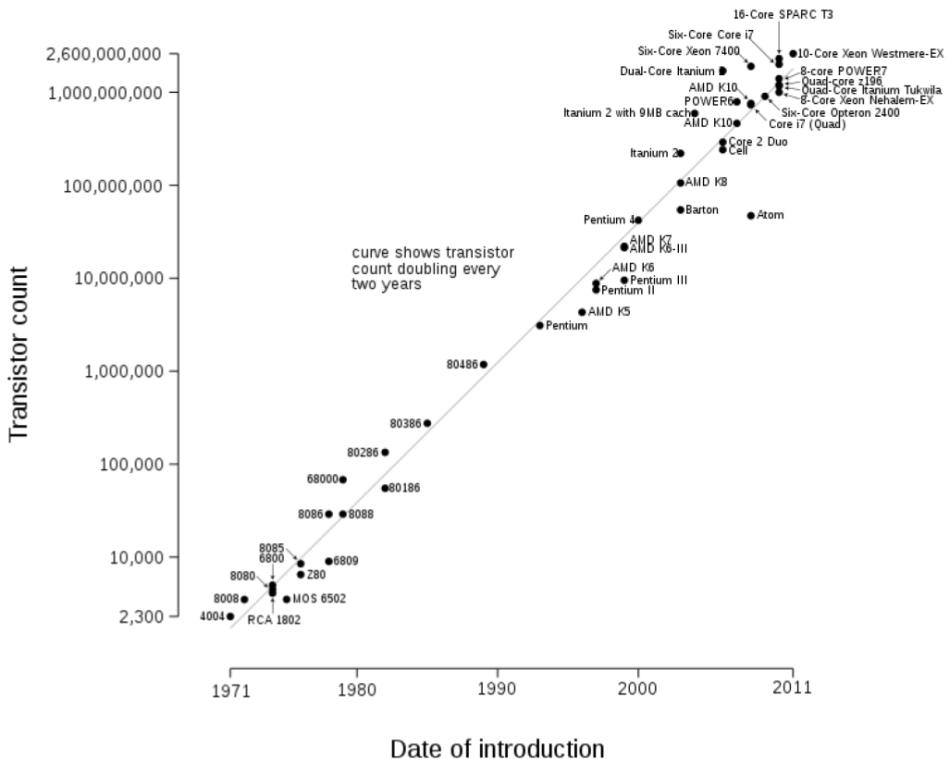
Follow / friends activities

If Netflix account connected to FB: personal info, etc.

Figure 4. Cinéma vs Netflix

b. Les ordinateurs sont devenus plus puissants

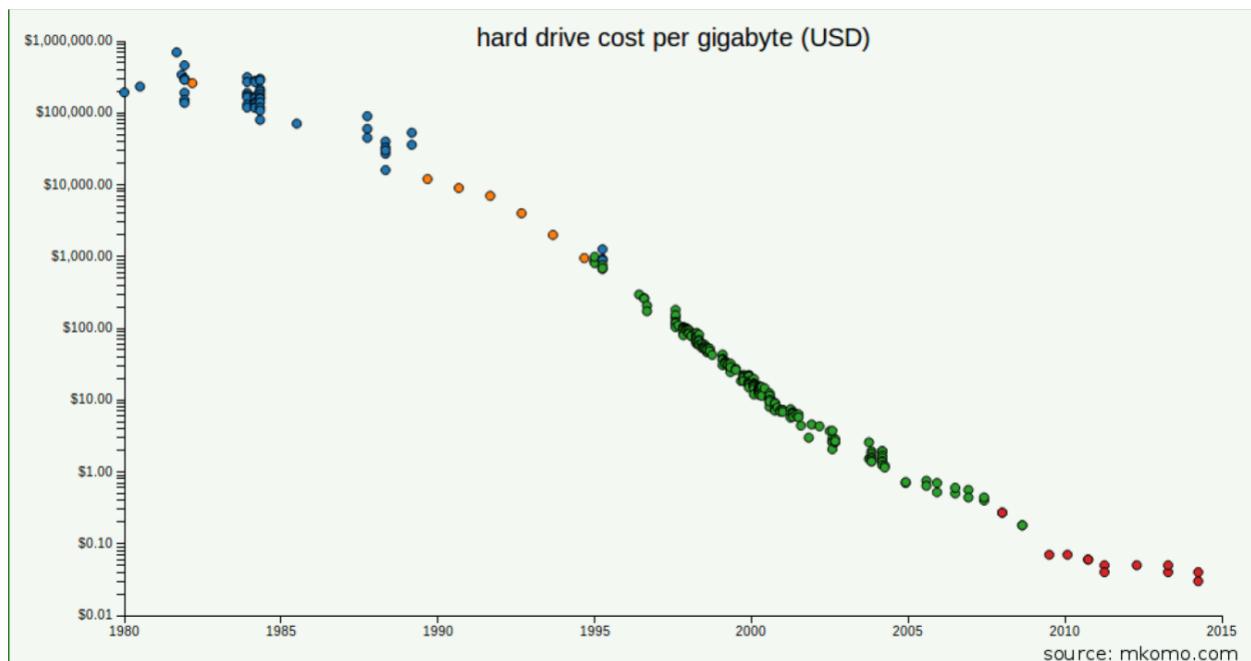
Microprocessor Transistor Counts 1971-2011 & Moore's Law



source: https://en.wikipedia.org/wiki/Moore%27s_law

Figure 5. La loi de Moore

c. Le stockage des données est devenu moins cher chaque année



source: <http://www.mkomo.com/cost-per-gigabyte>

Figure 6. Réduction des coûts de stockage des données

d. L'état d'esprit a changé sur ce qui "compte" comme données

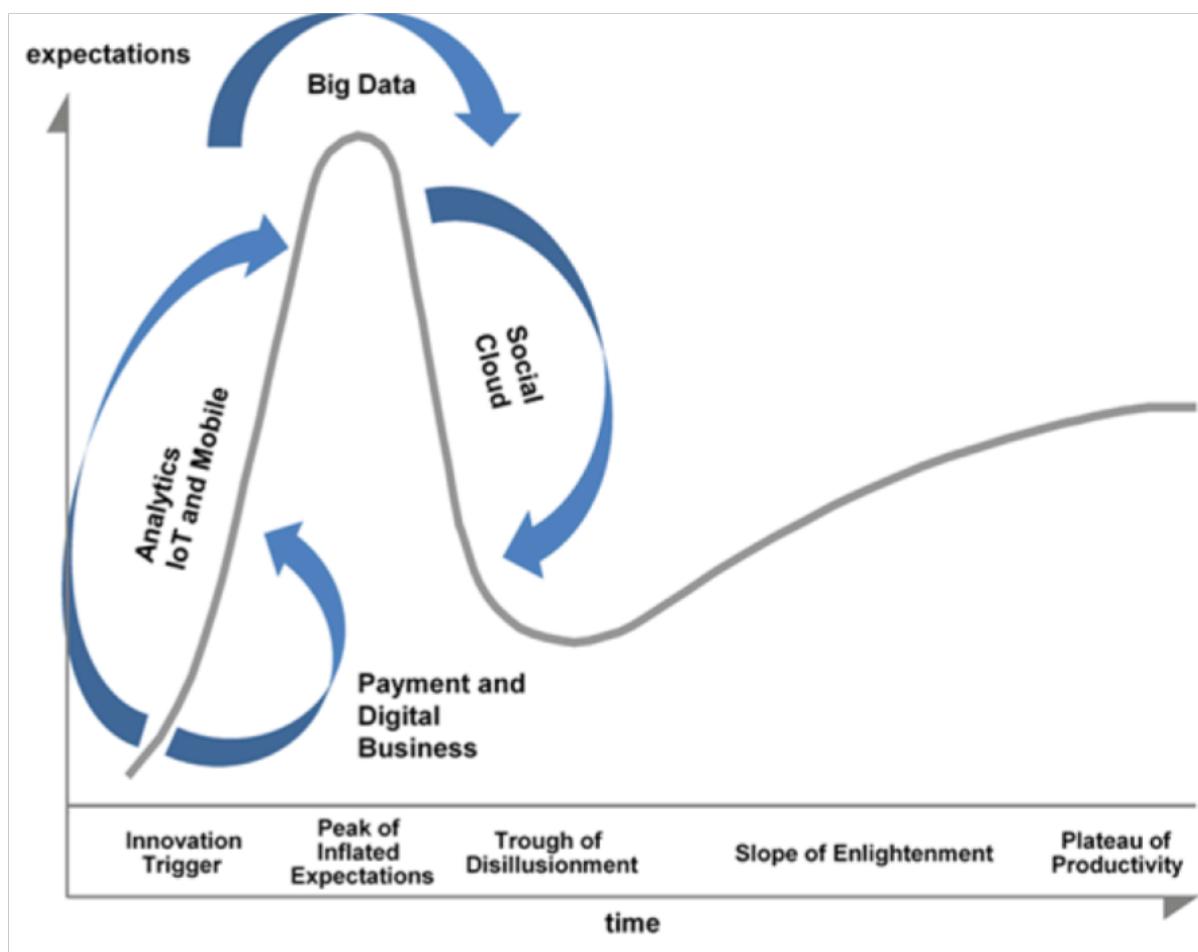
- Les données non structurées (voir ci-dessus pour la définition de "non structuré") n'étaient généralement pas stockées : cela prend beaucoup de place, et les logiciels pour les interroger n'étaient pas suffisamment développés.
- Les données de réseau (également appelées "graphs") (qui est un ami avec qui, qui aime les mêmes choses que qui, etc.) étaient généralement négligées car difficiles à interroger. Les réseaux sociaux comme Facebook ont fait beaucoup pour sensibiliser les entreprises à la valeur des graphs (en particulier les [graphs sociaux](#)). [Neo4J](#) ou [Titan](#) sont des fournisseurs de bases de données spécialisés dans le stockage et l'analyse de données réseau.
- Les données géographiques se sont démocratisées : des bases de données spécifiques (et coûteuses) ont longtemps existé pour stocker et interroger des "données de lieu" (régions, distances, informations de proximité ...) mais des solutions simples à utiliser se sont récemment multipliées, comme [Carto](#).

e. Le logiciel open source accélère l'innovation

À la fin des années 1990, les développeurs de logiciels ont rapidement changé d'habitudes : ils avaient tendance à utiliser de plus en plus de logiciels libres et à publier leurs logiciels en tant que logiciels libres. Jusque-là, la plupart des logiciels étaient "à source fermée": vous achetez un logiciel **sans possibilité** de réutiliser / modifier / augmenter son code source. Vous ne pouvez que l'utiliser tel quel. * L'open source facilite l'accès aux logiciels construits par d'autres, il est possible d'utiliser ces logiciels libres pour développer de nouvelles choses. Après plusieurs décennies, **le logiciel open source s'est banalisé**.

f. Les promesses et attentes exagérées sur le big data

Le [Gartner hype cycle](#) est un outil qui mesure la maturité d'une technologie, en différenciant les attentes des rendements réels:



Gartner's Hype Cycle Special Report for 2014

⌚ 06 August 2014 ⚒ G00268778

Analyst(s): [Betsy Burton](#) | [David A. Willis](#)

Figure 7. Cycle Gartner Hype pour 2014

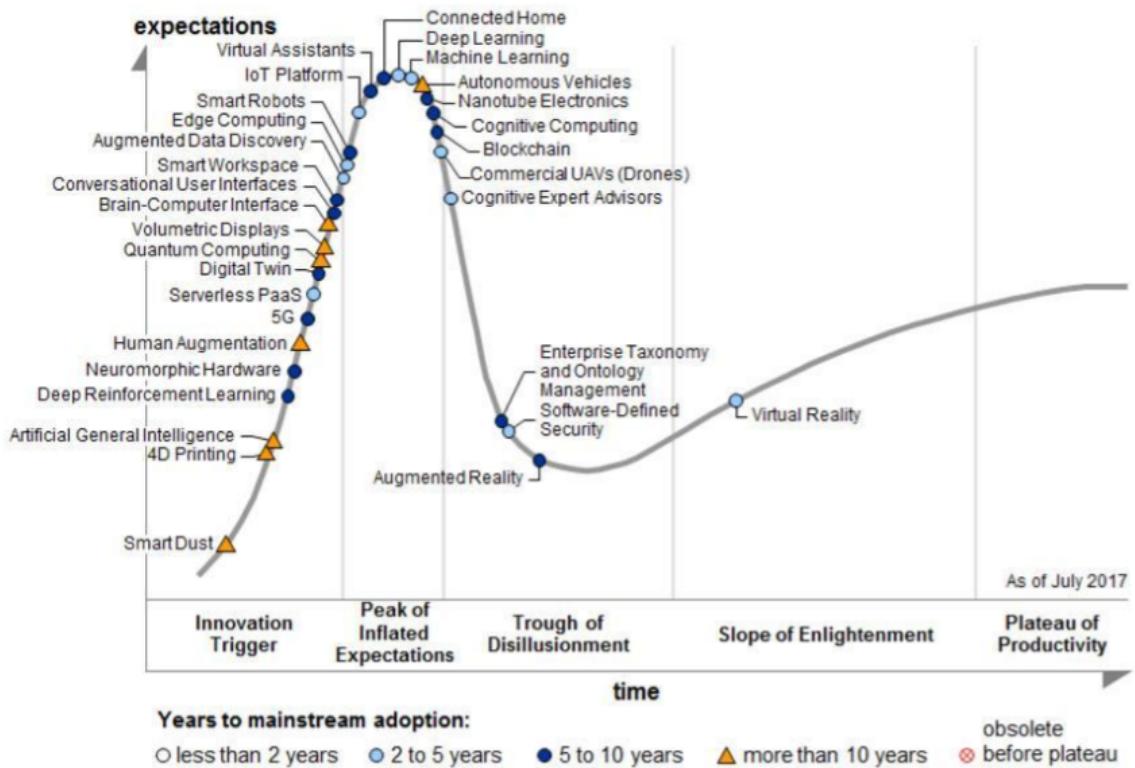
Ce graphique montre le modèle que toutes les technologies suivent au cours de leur vie:

- au début (à gauche du graphique), une invention ou découverte est faite dans un laboratoire de recherche, quelque part. Des reportages sont faits à ce sujet, mais cela fait peu de bruit.
- alors, la technologie commence à capter l'intérêt des journalistes, des consultants, des professeurs, des industriels ... les attentes grandissent quant aux possibilités et aux promesses de la technologie. "Avec cela nous pourrons [insérer quelque chose d'étonnant ici]"
- le sommet de la courbe est le «pic des attentes gonflées». Toutes les techniques et innovations ont tendance à être exagérées dans leur promesses, et même surexagérées. Cela signifie que la technologie devrait fournir plus qu'elle ne le fera sûrement, en réalité. Les gens se sont emballés.
- Puis suit le "creux de la désillusion". Le doute s'installe. Les gens se rendent compte que la technologie n'est pas aussi puissante, facile, bon marché ou rapide à mettre en œuvre qu'elle semblait au premier abord. Les journaux commencent à rapporter des nouvelles déprimantes sur la technologie, et quelques mauvaises rumeurs.
- enfin: la pente des lumières. Les têtes se refroidissent, les attentes s'alignent sur ce que la technologie peut réellement fournir. Les marchés se stabilisent et se consolident : certaines entreprises ferment et des acteurs clés continuent de se développer.
- alors: plateau de productivité. La technologie est maintenant normalisée, elle est utilisée de façon courante pour des usages précis.



Toute technologie peut «mourir» - tomber en désuétude - avant d'atteindre le côté droit du graphique bien sûr.

En 2014, les big data étaient proches du sommet de la courbe: elles retenaient beaucoup d'attention mais leur utilisation pratique en 5 à 10 ans était encore incertaine. Il y avait de « grandes attentes » quant à leur avenir, et ces attentes stimulent l'investissement, la recherche et les affaires dans le Big Data. En 2017, le «big data» est toujours au top des technologies hype, mais se décompose en «deep learning» et en «machine learning». Notez également la catégorie "Intelligence artificielle générale":



Note: PaaS = platform as a service; UAVs = unmanned aerial vehicles

Source: Gartner (July 2017)

Figure 8. Gartner Hype Cycle pour 2017

g. Le Big Data transforme les industries et est devenu une industrie en soi

Les entreprises actives dans les «Big data» se divisent en plusieurs sous-domaines: l'industrie de la gestion de l'infrastructure informatique pour les big data, les cabinets de conseil, les fournisseurs de logiciels, les applications métiers, etc ... Matt Turck, VC chez FirstMarkCap, crée chaque année une feuille pour visualiser les principales entreprises actives dans ces sous-domaines. Ceci est la version 2017:



Figure 9. Paysage de données pour 2017

Vous trouverez une [version haute résolution de ce panorama Big data](#), une version Excel et un commentaire très intéressant sur ce site : <https://mattturck.com/bigdata2017/>

5. Quel est l'avenir du Big Data?

a. Plus de données arrivent

L'**Internet des objets** désigne l' [extension d'Internet aux objets, au-delà des pages web ou des emails](#). L' **IoT** * **est utilisé pour faire** des choses (affichage d'informations à l'écran, robots pilotes, etc.) mais aussi beaucoup pour **collecter des données** dans leurs environnements, via des capteurs. Ainsi, le développement des **objets connectés** conduira à une augmentation considérable du volume de données collectées.

b. Les cadres réglementaires vont augmenter en complexité

Les impacts sociétaux du big data et de l'IA ne sont pas banals, allant de la discrimination raciale, financière et médicale à des fuites géantes de données, ou au déséquilibre économique à l'ère des robots et de l'IA sur le lieu de travail. Les réglementations publiques aux niveaux national et international tentent de rattraper ces défis. À mesure que la technologie évolue rapidement, nous pouvons anticiper que les impacts sociétaux des big data occuperont une place centrale.

c. en tant qu'expression, "big data" évolue

- Il est intéressant de noter que les expressions "à la mode", comme "big data", ont tendance à s'user rapidement. Elles sont sur-utilisées, mentionnées en toutes circonstances, deviennent vagues et trop vendues. Pour les données volumineuses, nous observons qu'on atteint un sommet en 2017, alors que de nouveaux termes apparaissent:

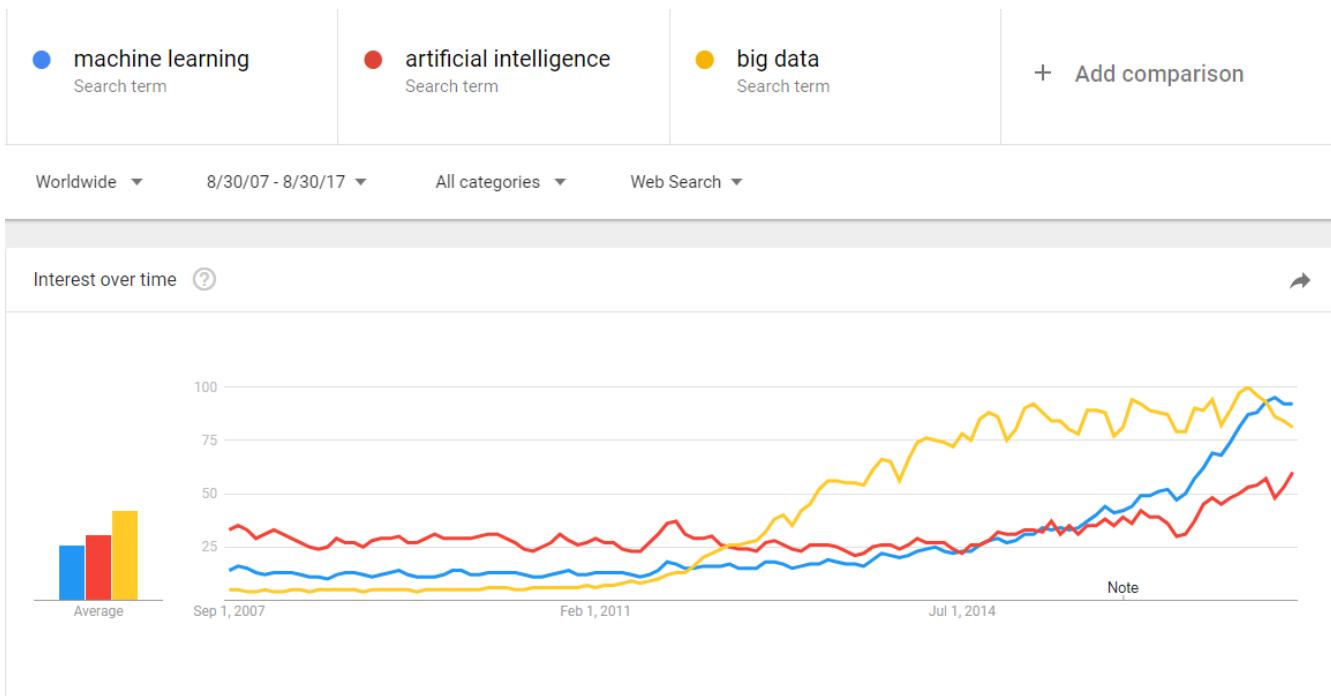


Figure 10. Google searches for big data, machine learning and AI

Quelles sont les différences entre ces termes?

- "Big Data" est maintenant un terme générique.
- **Machine learning** met l'accent sur les capacités de génie scientifique et logiciel permettant de faire quelque chose d'utile avec les données (prédir, catégoriser, marquer ...)
- **Intelligence artificielle** met l'accent sur les possibilités "quasi-humaines" offertes par l'apprentissage automatique. Le terme est souvent utilisé de manière interchangeable avec l'apprentissage automatique. L'intelligence artificielle est alimentée par des données, de sorte que l'avenir des big data se recoupe avec ce que deviendra l'IA.
- Et * data science * ? C'est un terme général englobant l'apprentissage automatique, les statistiques et de nombreuses méthodes analytiques pour travailler avec les données et les interpréter. Souvent utilisé de manière interchangeable avec l'apprentissage automatique. **Data scientist** est une description d'emploi devenue commune, y compris en français.

Pour aller plus loin

Retrouvez le site complet : [here](#).



Clement Levallois

Découvrez mes autres cours et projets : <https://www.clementlevallois.net>

Ou contactez-moi via Twitter: [@seinecle](#)