Essentials of data for managers

From the fundamentals to Artificial Intelligence

Clément Levallois

ExpData Press

2018

Essentials of data for managers

From the fundamentals to Artificial Intelligence

Clément Levallois

ExpData Press

Saint-Etienne

ESSENTIALS OF DATA FOR MANAGERS

by Clément Levallois

Published by Peecho, Rokin 75-5, 1012KL Amsterdam, Netherlands

April 2018: first edition

Revision history for the first release:

2018-04-01: first release

From the same author:

Levallois, C. et al, eds. (2015) . *Twitter for Research*. Ecully: EMLYON Press.

Acknowledgments

This book benefitted from many years of teaching and interaction with participants to my courses, at Erasmus University Rotterdam then at em lyon business school in Msc, exec education and MBA. The small world of Twitter is also a great provider of news on the latest technology developments and their societal impacts. Thank you all.

For Manon, Léon and Tristan.

Table of Contents

Preface

A textbook for managers

The target reader for this book is a manager who needs to clearly understand what "data science", "big data", "artificial intelligence" so that they can:

- **leverage** these technologies to improve the efficiency of their existing business,
- **innovate** with new products and services and develop new business guidelines

The promise of this book is to bring you from a starting point with no knowledge of these technical concepts, to a point where you understand the concepts **and** you can develop "data centric" business projects: when "data" contributes to creating value for the customer and all stakeholders.

Is this textbook too technical or too easy for me?

If you are unsure, try this simple test: http://bit.ly/essentials-1-test

→ There are 20 topics you should be comfortable answering. See how you score. If the score is low, this book will be of great use.

Chapter 1. Data, a concept with multiple layers

Definition of data

The English term "data" (1654) originates from "datum", a Latin word for "a given"[1]. "Data" is a single factual, a single entity, a single point of matter.

The word "data" to mean "transmittable and storable computer information" was first used in this sense in 1946. The expression "data processing" was first used in 1954.

Thoughts: the etymology suggests that data is "a given". Can you question this?

Data represents either a single entity, or a collection of such entities ("data points"). We can speak also of datasets instead of data (so a dataset is a collection of data points).

The variety of data sets

A date	A color	A grade
A relation of friendship	A sound	A heartbeat
A user input	A duration	A curriculum vitae

A picture	A longitude and latitude	A price
A number of friends	A temperature	A list of favorite movies
etc…	etc…	etc…

These examples are chosen on purpose to be varied and from unexpected places. They illustrate three principles:

a. Think about data in a broad sense

Data is not just numerical, neither is it "what sits in my spreadsheets". You should train in thinking about data in a broader sense:

- pictures are data
- language is data (including slang, lip movements, etc.)
- relations are data: individual A is known, individual B is known, **but the relationship between A and B is data as well**
- preferences, emotional states… are data

- etc. There is no definitive list, you should train yourself looking at business situations and think: "where is the data?"

b. metadata is data, too

Metadata is a piece of data describing another data.

Example:

```
The bibliographical reference ①
describing
a book ②
```

① the metadata

② the data

→ Data without metadata can be worthless (imagine a library without a library catalogue)

→ Metadata can be informative in its own right, as shown with the NSA scandal (read this article from the New Yorker about NSA and metadata[2]).

c. zoom in, zoom out

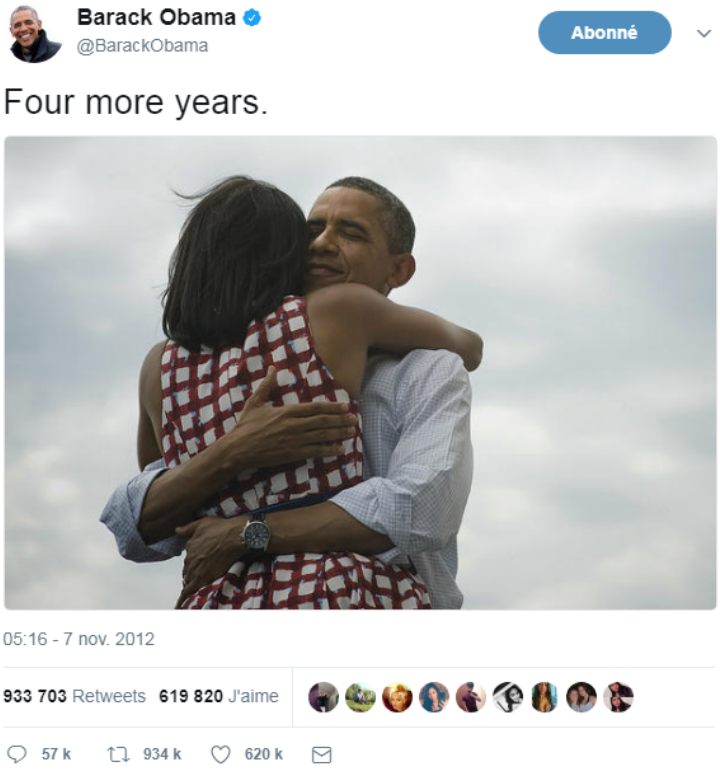
We should remember considering that a data point can be itself a collection of data points:

- a person walking into a building is a data point.
- however this person is itself a collection of data points: location data + network relations + subscriber status to services + etc.

So it is a good habit to wonder whether a data point can in fact be "unbundled" (spread into smaller data points / measurements)

How to describe datasets

a. Formats, types, encoding



- This is a digital **medium** (because it's on screen as opposed to analogical, if we had printed the pic on paper)
- The **type** of the data is textual + image
- The text is formatted in **plain text** (meaning, no special formatting), as opposed to **data-interchange formats** which are formatting marks added to the text to facilitate its readability by software (check csv, json and xml[3]).
- The **encoding** of the text is UTF-8 (one of encodings deriving from the Unicode standard). Encoding deals with the issue: how to represent alphabets, signs (for instance: emojis) and symbols, from different languages, in text? UTF-8 is an encoding which is one of the most universal.
- The tweet is part of a list of tweets. The list represents the **data structure** of the dataset, it is the way the data is organized. There are many alternative data structures: arrays, sets, dics, maps...
- The tweet is stored as a picture (png file) on the hard disk. "png" is the **file format**. The data is **persisted** as a file on disk (could have been stored in a database instead).

b. Tabular data

Tabular data is a common way to handle datasets, by organizing it in lines and columns:

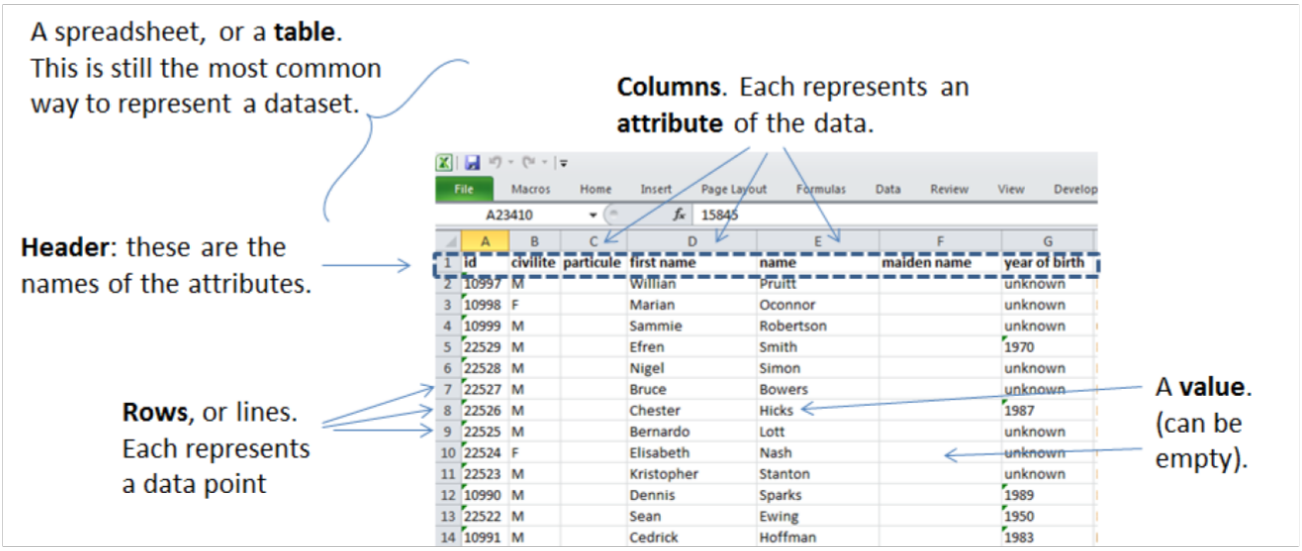


Figure 1. tabular data

c. First party, second party and third party data

- **First party data** : the data generated through the activities of your own organization. Your organization own it, which does not mean that consent from users is not required, when it comes to personal data.
- **Second party data** : the data accessed through partnerships. Without being the generator nor the owner of this data, partners make it available to you through an agreement.
- **Third party data** : the data acquired via purchase. This data is acquired through a market transaction. Its uses still comes with conditions, especially for personal data.

d. Sociodemo data vs behavior data

- Sociodemogaphic or **sociodemio** data refers to information about individuals, describing fundamental attributes of their social identity: age, gender, place of residence, occupation, marital status and number of kids.
- **Behavior data** refers to any digital trace left by the individual in the course of it life: clicks on web pages, likes on Facebook, purchase transactions, comments posted on Tripadvisor…

Sociodemo data is typically well structured or easy to structure. It has a long history of collection and analysis, basically since census exists.

Behavior data allows to go further than sociodemo data: each individual can be characterized by its acts and tastes, well beyond what an age or marital status could define.

But behavior data is typically not well structured and harder to collect.

Data and size

1 bit		can store a binary value (yes / no, true / false…)
8 bits	1 byte (or octet)	can store a single character
~ 1,000 bytes	1 kilobyte (kb)	Can store a paragraph of text
~ 1 million bytes	1 megabyte (Mb)	Can store a low res picture.

~ 1 billion bytes	1 gigabyte (Gb)	Can store a movie
~ 1 trillion bytes	1 terabyte (Tb)	Can store 1,000 movies. Size of commercial hard drives in 2017 is 2 Tb.
~ 1,000 trillion bytes	1 petabyte (Pb)	20 Pb = Google Maps in 2013

Chapter 2. A clarification of big data

Big data is a mess

Figure 2. Facebook post by Dan Ariely in 2013

Jokes aside, defining big data and what it covers needs a bit of precision. Let's bring some clarity.

The 3 V

Big data is usually described with the "3 Vs":

V for Volume

The size of datasets available today is staggering (ex: Facebook had 250 billion pics in 2016).

We should also note that the volumes of data are increasing at an **accelerating rate**. According to sources, 90% of all the data in the world has been generated over the last two years[4] (statement from 2013) or said differently, more data will be created in 2017 than the previous 5,000 years of humanity[5].

V for Variety

"Variety" refers to the fact that "unstructured" data is considered to be increasingly useful, when before the big data phenoenon only structured data was considered worth storing and exploiting. This calls to explain in more details the distinction between unstructured and structured data.

A - Structured data

Structured data refers to data which is formatted and organized according to a well defined set of rules, which makes it **machine readable**. For example, zip codes are a structured dataset because

they follow a precise convention regarding the number of letters and digits composing them, making it easy for an optical reader and software to identify and "read" them. Same with license plates, social security numbers…

But these are simple examples.

What about, for instance, a tax form? If each field of the form is well defined, then the data collected through the form can be said to be "structured". By contrast, a form where the user can write free text (think of a comment on a blog post, or a blank space where users can write a feedback) produces unstructured data: data which does not follow a special convention for its size and content. This is typically much harder for software to process, hence to analyze.

To summarize, think of structured data as as anything that can be represented as well organized tables of numbers and short pieces of text with the expected format, size, and conventions of writing: phonebooks, accounting books, governmental statistics…

Figure 3. A book of accounts showing structured data

B - Unstructured data

Unstructured data refers to datasets made of "unruly" items: text of any length, without proper categorization, encoded in different formats, including possibly pictures, sound, geographical coordinates and what not…

These datasets are much harder to process and analyze, since they are full of exceptions and differences. But they are carry typically rich information: free text, information recorded "in the wild"…

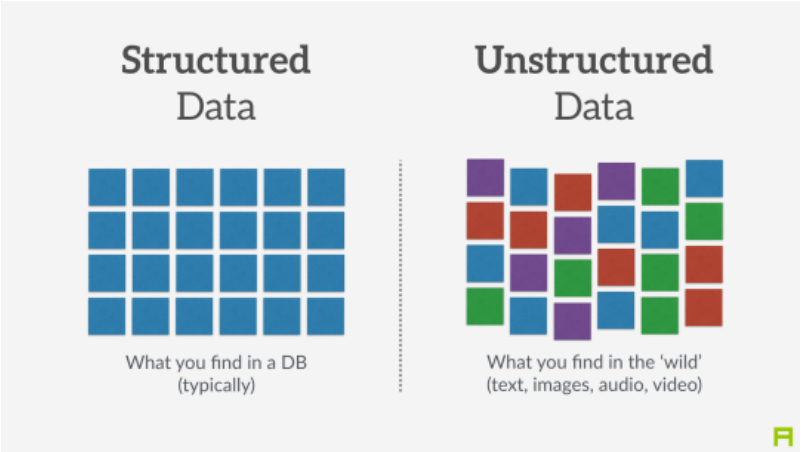


Figure 4. Structured vs unstructured data

V for Velocity

In a nutshell, the speed of creation and communication of data is accelerating[6]:

- Facebook hosts 250 billion pics? It receives 900 million more pictures **per day**
- Examining tweets can be done automatically (with computers). If you want to connect to Twitter to receive tweets in real time as they are tweeted, be prepared to receive in excess of 500 million tweets **per day**. Twitter calls this service the "Twitter firehose"[7], which reflects the velocity of the stream of tweets.
- **Sensor data** is bound to increase speed as well. While pictures, tweets, individual records… are single item data sent at intervals, more and more sensors can send data **in a continuous stream** (measures of movement, sound, etc.)

So, velocity poses challenges of its own: while a system can handle (store, analyze) say 100Gb of data in a given time (day or month), it might not be able to do it in say, a single second. Big data refers to the problems and solutions raised by the velocity of data.

A 4th V can be added, for Veracity

Veracity relates to trustworthiness and compliance: is the data authentic? Has it been corrupted at any step of its processing? Does it comply with local and international regulations?

What is the minimum size to count as "big data"? It’s all relative

There is no "threshold" or "minimum size" of a dataset where "data" would turn from "small data" to "big data".

It is more of a **relative** notion: it is big data if current IT systems struggle to cope with the datasets.

"Big data" is a relative notion… how so?

a. relative to time

- what was considered "big data" in the early 2000s would be considered "small data" today, because we have better storage and computing power today.
- this is a never ending race: as IT systems improve to deal with "current big data", data gets generated in still larger volumes, which calls for new progress / innovations to handle it.

b. relative to the industry

- what is considered "big data" by non tech SMEs (small and medium-sized enterprises) can be considered trivial to handle by tech companies.

c. not just about size

- the difficulty for an IT system to cope with a dataset can be related to the size (try analyzing 2 Tb of data on your laptop…), **but also** related to the content of the data.
- For example the analysis of customer reviews in dozens of languages is harder than the analysis of the same number of reviews in just one language.
- So the general rule is: the less the data is structured, the harder it is to use it, even if it's small in size (this relates to the "V" of variety seen above).

d. no correlation between size and value

- "Big data is often called the new oil"[8], as if it would flow like oil and would power engines "on demand".
- Actually, big data is **created**: it needs work, conception and design choices to even exist (what do I collect? how do I store it? what structure do I give to it?). The human intervention in creating data determines largely whether data will be of value later.
- Example: Imagine customers can write online reviews of your products. These reviews are data. But if you store these reviews without an indication of who has authored the review (maybe because reviews can be posted without login oneself), then the reviews become much less valuable.

Simple design decisions about how the data is collected, stored and structured have a huge impact on the value of the data.

So, in reaction to large, unstructured and badly curated datasets with low value at the end, a notion of "smart data" is sometimes put forward: data which can be small in size but which is well curated and annotated, enhancing its value (see also here[9]).

Where did big data come from?

a. Data got generated in bigger volumes because of the digitalization of the economy

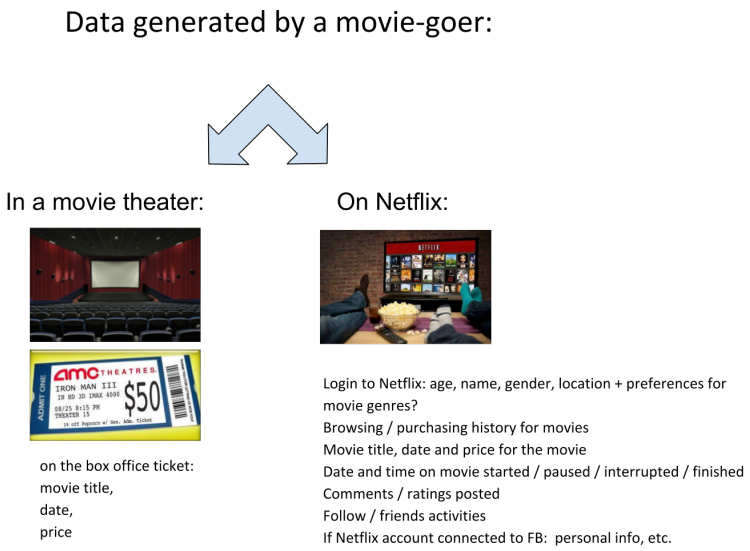
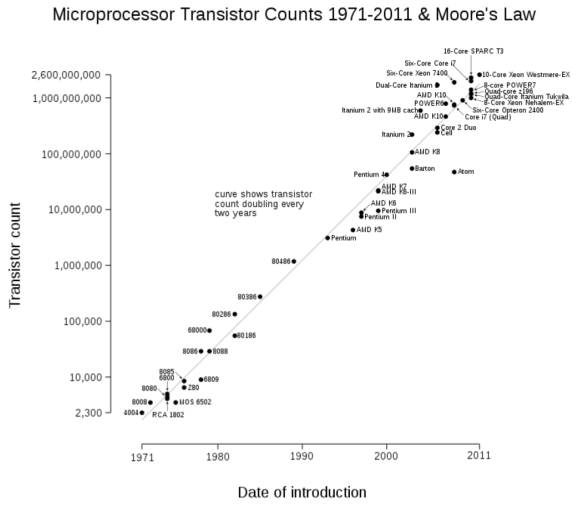


Figure 5. Movie theater vs Netflix

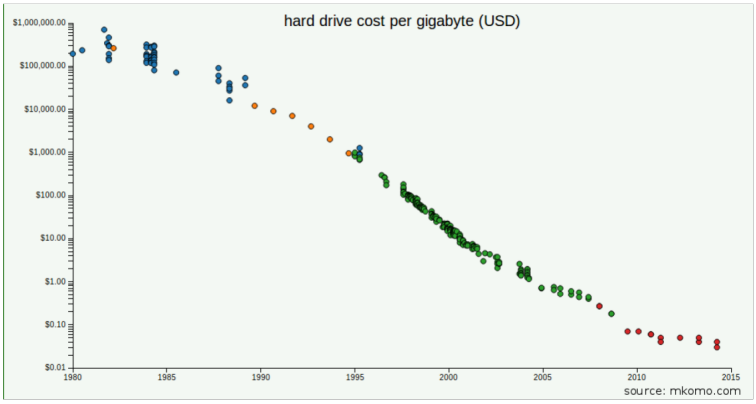
b. Computers became more powerful



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

Figure 6. Moore's law

c. Storing data became cheaper every year



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

Figure 7. Decreasing costs of data storage

d. The mindset changed as to what "counts" as data

- Unstructured data (see above for definition of "unstructured") was usually not stored: it takes a lot space, and software to query it was not sufficiently developped.
- Network data (also known as graphs) (who is friend with whom, who likes the same things as whom, etc.) was usually neglected as "not true observation", and hard to query. Social networks like Facebook made a lot to make businesses aware of the value of graphs (especially social graphs[10]).
- Geographical data has democratized: specific (and expensive) databases existed for a long time to store and query "place data" (regions, distances, proximity info...) but easy-to-use solutions have multiplied recently.

e. With open source software, the rate of innovation accelerated

In the late 1990s, a rapid shift in the habits of software developers kicked in: they tended to use more and more open source software, and to release their software as open source. Until then, most of the software was "closed source": you buy a software **without the possibility** to reuse / modify / augment its source code. Just use it as is.

Open source software made it easy to get access to software built by others and use it to develop new things. Today, all the most popular software in machine learning are free and open source.

See the Wikipedia article for a developed history of open source software: https://en.wikipedia.org/

wiki/History_of_free_and_open-source_software

f. Hype kicked in

The Gartner hype cycle[11] is a tool measuring the maturity of a technology, differentiating expectations from actual returns:

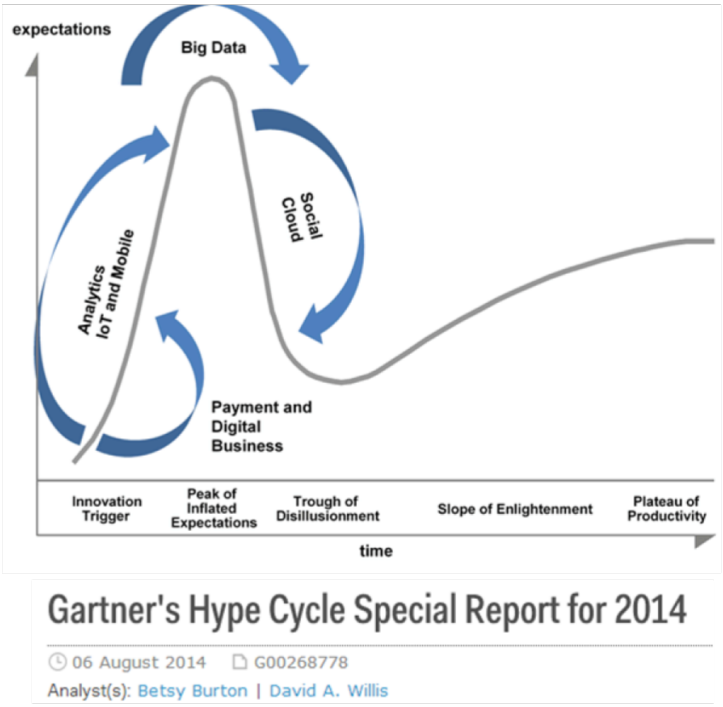


Figure 8. Gartner Hype Cycle for 2014

This graph shows the pattern that all technologies follow along their lifetime:

- at the beginning (left of the graph), an invention or discovery is made in a research lab, somewhere. Some news reporting is done about it, but with not much noise.
- then, the technology starts picking the interest of journalists, consultant, professors, industries… expectations grow about the possibilities and promises of the tech. "With it we will be able to [insert amazing thing here]"
- the top of the bump is the "peak of inflated expectations". All techs tend to be hyped and even over hyped. This means the tech is expected to deliver more than it surely will, in actuality. People get overdrawn.
- then follows the "Trough of Disillusionment". Doubt sets in. People realize the tech is not as powerful, easy, cheap or quick to implement as it first seemed. Newspapers start reporting depressing news about the tech, some bad buzz spreads.
- then: slope of Enlightenment. Heads get colder, expectations get in line with what the tech can actually deliver. Markets stabilize and consolidate: some firms close and key actors continue to grow.

- then: plateau of productivity. The tech is now mainstream.

(all technology can "die" - fall into disuse - before reaching the right side of the graph of course).

In 2014, big data was near the top of the curve: it was getting a lot of attention but its practical use in 5 to 10 years were still uncertain. There were "great expectations" about its future, and these expectations drive investment, research and business in big data.

In 2017, "big data" is still on top of hyped technologies, but is broken down in "deep learning" and "machine learning". Note also the "Artificial General Intelligence" category:

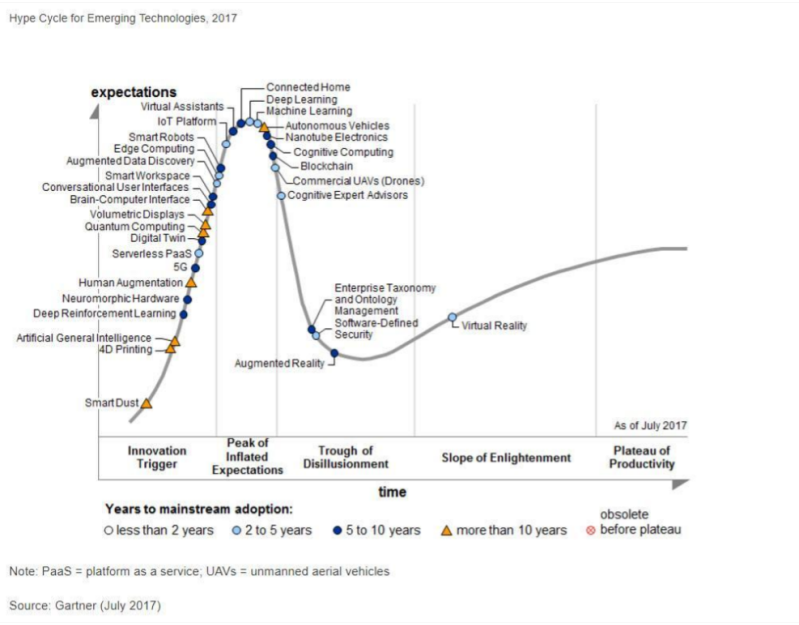


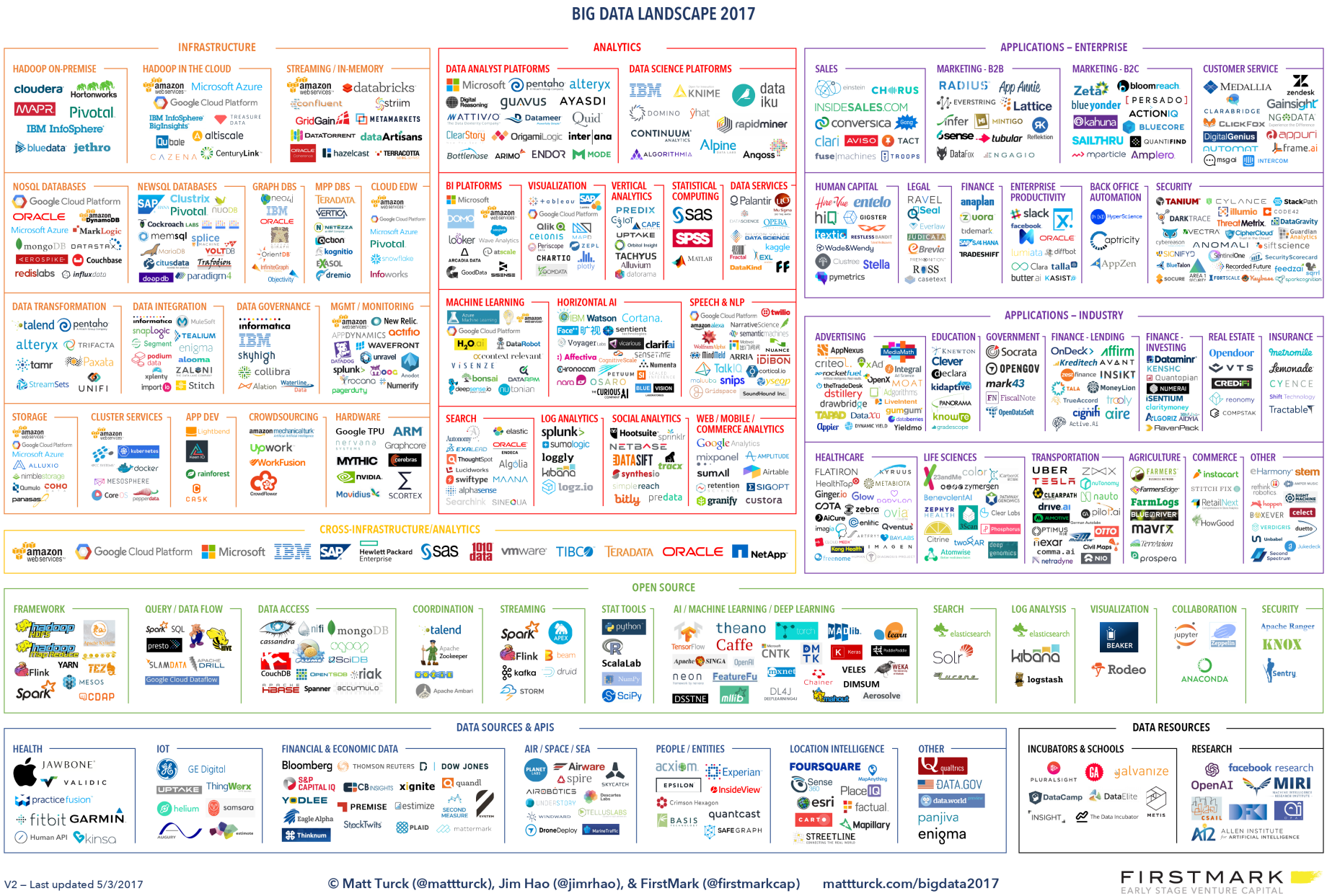
Figure 9. Gartner Hype Cycle for 2017

g. Big data transforms industries, and has become an industry in itself

Firms active in "Big data" divide in many sub-domains: the industry to manage the IT infrastructure for big data, the consulting firms, software providers, industry-specific applications, etc...

Matt Turck, VC at FirstMarkCap[12], creates every year a sheet to visualize the main firms active in these subdomains.

This is the 2017 version:



V2 – Last updated 5/3/2017

 mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Figure 10. Big data landscape for 2017

You can find an high res version of the Big data panorama[13], an Excel sheet version, and a very interesting comment on this website: https://mattturck.com/bigdata2017/

What is the future of big data?

a. More data is coming

The **Internet of things** designates the extension of Internet to objects beyond web pages or emails[14].

The **IoT** is used to **do** things (display information on screen, pilote robots, etc.) but also very much to **collect data** in their environments, through sensors.

Hence, the development of **connected objects** will lead to a tremendous increase in the volume of data collected.

b. Regulatory frameworks will grow in complexity

Societal impacts of big data and AI are not trivial, ranging from racial, financial and medical discrimination to giant data leaks, or economic (un)stability in the age of robots and AI in the workplace.

Public regulations at the national and international levels are trying to catch up with these challenges. As technology evolves quickly, we can anticipate that societal impacts of big data will take center stage.

c. as an expression, "big data" is evolving

- It is interesting to note that "hot" expressions, like "big data", tend to wear out fast. They are too hyped, used in all circumstances, become vague and over sold. For big data, we observe that it is peaking in 2017, while new terms appear:

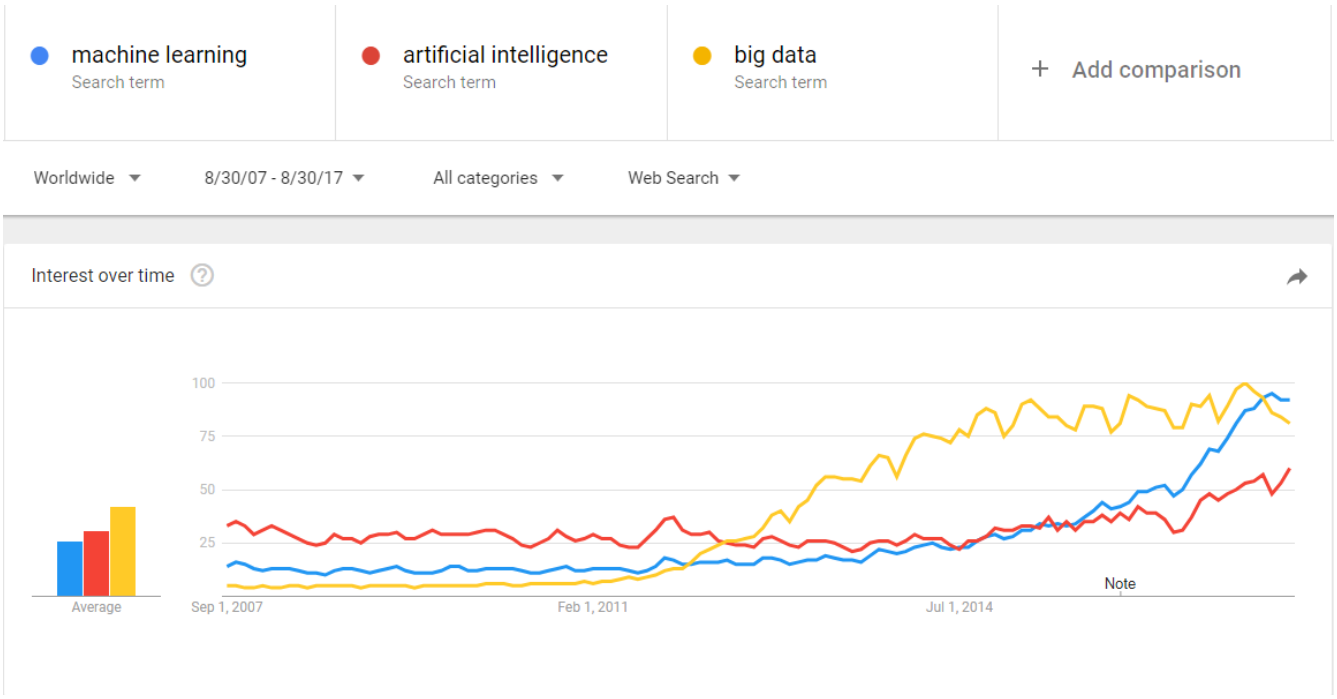


Figure 11. Google searches for big data, machine learning and AI

What are the differences between these terms?

- "Big data" is by now a generic term
- **Machine learning** puts the focus on the scientific and software engineering capabilities enabling to do something useful with the data (predict, categorize, score...)
- **Artificial intelligence** puts the emphasis on human-like possibilities afforded by machine learning. Often used interchangeably with machine learning. AI is fed on data, so the future of big data will intersect with what AI becomes.
- And **data science** ? This is a broad term encompassing machine learning, statistics, and many analytical methods to work with data and interpret it. Often used interchangeably with machine learning. **Data scientist** is a common job description in the field.

h machine learning. **Data scientist** is a common job description in the field.

References

1. *"datum", a Latin word for "a given"*: http://www.etymonline.com/index.php?term=data
2. *NSA and metadata*: http://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata
3. *csv, json and xml*: https://codingislove.com/json-tutorial-indepth/
4. *90% of all the data in the world has been generated over the last two years*: https://www.sciencedaily.com/releases/2013/05/130522085217.htm
5. *more data will be created in 2017 than the previous 5,000 years of humanity*: https://appdevelopermagazine.com/4773/2016/12/23/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity-/
6. *the speed of creation and communication of data is accelerating*: http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/
7. *"Twitter firehose"*: http://support.gnip.com/apis/firehose/
8. *"Big data is often called the new oil"*: https://hbr.org/2012/11/data-humans-and-the-new-oil
9. *here*: https://www.quora.com/After-Big-Data-Smart-Data-is-a-trend-in-2013-So-what-is-Smart-Data-Have-any-clear-definition
10. *social graphs*: https://en.wikipedia.org/wiki/Social_graph
11. *Gartner hype cycle*: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
12. *Matt Turck, VC at FirstMarkCap*: https://twitter.com/mattturck
13. *high res version of the Big data panorama*: https://mattturck.com/bigdata2017/
14. *extension of Internet to objects beyond web pages or emails*: https://seinecle.github.io/IoT4Entrepreneurs/

Index

Topics in Data Science for business: Volume 1 - Fundamentals

Clément Levallois <levallois@em-lyon.com [mailto:levallois@em-lyon.com]> v1.0, April 2018 :icons!: :iconsfont: font-awesome :revnumber: 1.0 :media: prepress :example-caption!: :sourcedir: ../../../main/java :toc: :toclevels: 1 = Preface

A textbook for managers

The target reader for this book is a manager who needs to clearly understand what "data science", "big data", "artificial intelligence" so that they can:

- **leverage** these technologies to improve the efficiency of their existing business,
- **innovate** with new products and services and develop new business guidelines

The promise of this book is to bring you from a starting point with no knowledge of these technical concepts, to a point where you understand the concepts **and** you can develop "data centric" business projects: when "data" contributes to creating value for the customer and all stakeholders.

Is this textbook too technical or too easy for me?

If you are unsure, try this simple test: http://bit.ly/essentials-1-test

→ There are 20 topics you should be comfortable answering. See how you score. If the score is low, this book will be of great use.

Chapter 1. Data, a concept with multiple layers

Definition of data

The English term "data" (1654) originates from "datum", a Latin word for "a given"[1]. "Data" is a single factual, a single entity, a single point of matter.

The word "data" to mean "transmittable and storable computer information" was first used in this sense in 1946. The expression "data processing" was first used in 1954.

Thoughts: the etymology suggests that data is "a given". Can you question this?

Data represents either a single entity, or a collection of such entities ("data points"). We can speak also of datasets instead of data (so a dataset is a collection of data points).

The variety of data sets

A date	A color	A grade
A relation of friendship	A sound	A heartbeat
A user input	A duration	A curriculum vitae

A picture	A longitude and latitude	A price
A number of friends	A temperature	A list of favorite movies
etc…	etc…	etc…

These examples are chosen on purpose to be varied and from unexpected places. They illustrate three principles:

a. Think about data in a broad sense

Data is not just numerical, neither is it "what sits in my spreadsheets". You should train in thinking about data in a broader sense:

- pictures are data
- language is data (including slang, lip movements, etc.)
- relations are data: individual A is known, individual B is known, **but the relationship between A and B is data as well**
- preferences, emotional states… are data

- etc. There is no definitive list, you should train yourself looking at business situations and think: "where is the data?"

b. metadata is data, too

Metadata is a piece of data describing another data.

Example:

```
The bibliographical reference ①
describing
a book ②
```

① the metadata

② the data

→ Data without metadata can be worthless (imagine a library without a library catalogue)

→ Metadata can be informative in its own right, as shown with the NSA scandal (read this article from the New Yorker about NSA and metadata[2]).

c. zoom in, zoom out

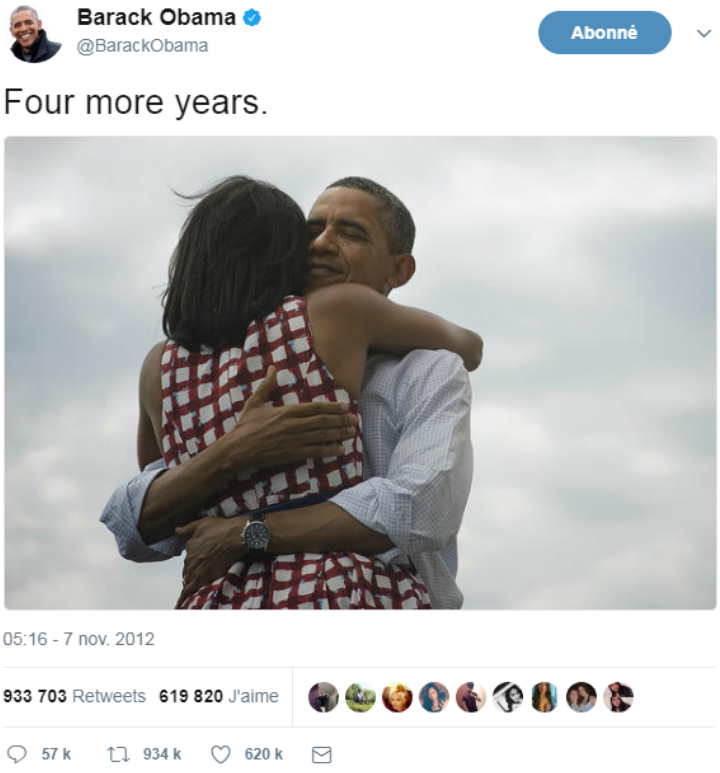
We should remember considering that a data point can be itself a collection of data points:

- a person walking into a building is a data point.
- however this person is itself a collection of data points: location data + network relations + subscriber status to services + etc.

So it is a good habit to wonder whether a data point can in fact be "unbundled" (spread into smaller data points / measurements)

How to describe datasets

a. Formats, types, encoding



- This is a digital **medium** (because it's on screen as opposed to analogical, if we had printed the pic on paper)
- The **type** of the data is textual + image
- The text is formatted in **plain text** (meaning, no special formatting), as opposed to **data-interchange formats** which are formatting marks added to the text to facilitate its readability by software (check csv, json and xml[3]).
- The **encoding** of the text is UTF-8 (one of encodings deriving from the Unicode standard). Encoding deals with the issue: how to represent alphabets, signs (for instance: emojis) and symbols, from different languages, in text? UTF-8 is an encoding which is one of the most universal.
- The tweet is part of a list of tweets. The list represents the **data structure** of the dataset, it is the way the data is organized. There are many alternative data structures: arrays, sets, dics, maps...
- The tweet is stored as a picture (png file) on the hard disk. "png" is the **file format**. The data is **persisted** as a file on disk (could have been stored in a database instead).

b. Tabular data

Tabular data is a common way to handle datasets, by organizing it in lines and columns:

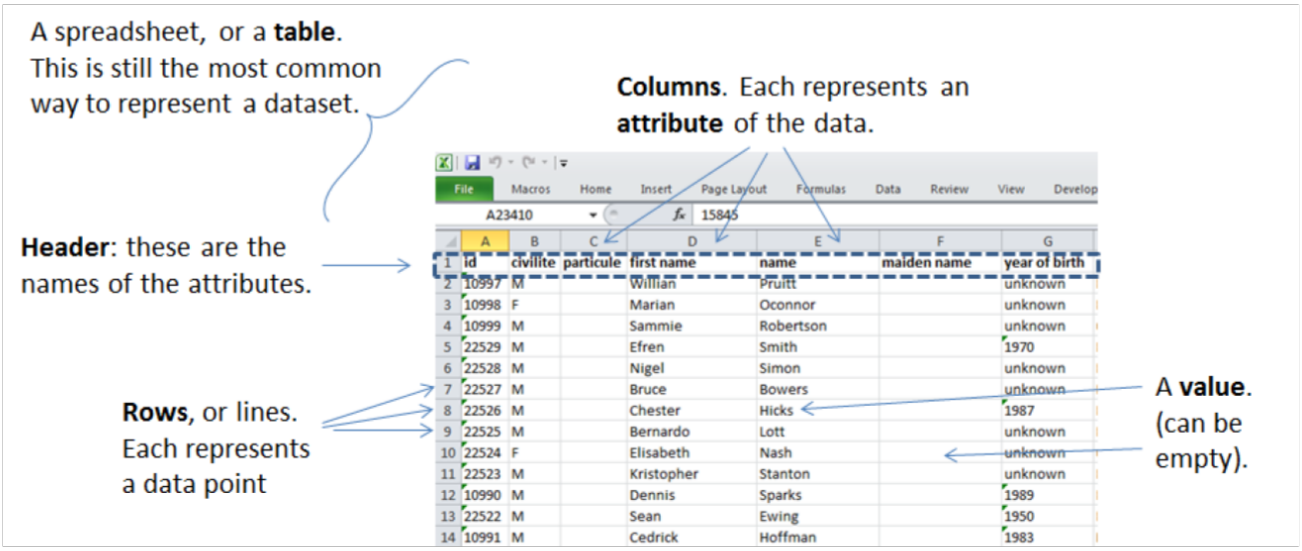


Figure 12. tabular data

c. First party, second party and third party data

- **First party data** : the data generated through the activities of your own organization. Your organization own it, which does not mean that consent from users is not required, when it comes to personal data.
- **Second party data** : the data accessed through partnerships. Without being the generator nor the owner of this data, partners make it available to you through an agreement.
- **Third party data** : the data acquired via purchase. This data is acquired through a market transaction. Its uses still comes with conditions, especially for personal data.

d. Sociodemo data vs behavior data

- Sociodemogaphic or **sociodemio** data refers to information about individuals, describing fundamental attributes of their social identity: age, gender, place of residence, occupation, marital status and number of kids.
- **Behavior data** refers to any digital trace left by the individual in the course of it life: clicks on web pages, likes on Facebook, purchase transactions, comments posted on Tripadvisor…

Sociodemo data is typically well structured or easy to structure. It has a long history of collection and analysis, basically since census exists.

Behavior data allows to go further than sociodemo data: each individual can be characterized by its acts and tastes, well beyond what an age or marital status could define.

But behavior data is typically not well structured and harder to collect.

Data and size

1 bit		can store a binary value (yes / no, true / false…)
8 bits	1 byte (or octet)	can store a single character
~ 1,000 bytes	1 kilobyte (kb)	Can store a paragraph of text
~ 1 million bytes	1 megabyte (Mb)	Can store a low res picture.

~ 1 billion bytes	1 gigabyte (Gb)	Can store a movie
~ 1 trillion bytes	1 terabyte (Tb)	Can store 1,000 movies. Size of commercial hard drives in 2017 is 2 Tb.
~ 1,000 trillion bytes	1 petabyte (Pb)	20 Pb = Google Maps in 2013

Chapter 2. A clarification of big data

Big data is a mess

Figure 13. Facebook post by Dan Ariely in 2013

Jokes aside, defining big data and what it covers needs a bit of precision. Let's bring some clarity.

The 3 V

Big data is usually described with the "3 Vs":

V for Volume

The size of datasets available today is staggering (ex: Facebook had 250 billion pics in 2016).

We should also note that the volumes of data are increasing at an **accelerating rate**. According to sources, 90% of all the data in the world has been generated over the last two years[4] (statement from 2013) or said differently, more data will be created in 2017 than the previous 5,000 years of humanity[5].

V for Variety

"Variety" refers to the fact that "unstructured" data is considered to be increasingly useful, when before the big data phenoenon only structured data was considered worth storing and exploiting. This calls to explain in more details the distinction between unstructured and structured data.

A - Structured data

Structured data refers to data which is formatted and organized according to a well defined set of rules, which makes it **machine readable**. For example, zip codes are a structured dataset because

they follow a precise convention regarding the number of letters and digits composing them, making it easy for an optical reader and software to identify and "read" them. Same with license plates, social security numbers…

But these are simple examples.

What about, for instance, a tax form? If each field of the form is well defined, then the data collected through the form can be said to be "structured". By contrast, a form where the user can write free text (think of a comment on a blog post, or a blank space where users can write a feedback) produces unstructured data: data which does not follow a special convention for its size and content. This is typically much harder for software to process, hence to analyze.

To summarize, think of structured data as as anything that can be represented as well organized tables of numbers and short pieces of text with the expected format, size, and conventions of writing: phonebooks, accounting books, governmental statistics…

Figure 14. A book of accounts showing structured data

B - Unstructured data

Unstructured data refers to datasets made of "unruly" items: text of any length, without proper categorization, encoded in different formats, including possibly pictures, sound, geographical coordinates and what not…

These datasets are much harder to process and analyze, since they are full of exceptions and differences. But they are carry typically rich information: free text, information recorded "in the wild"…

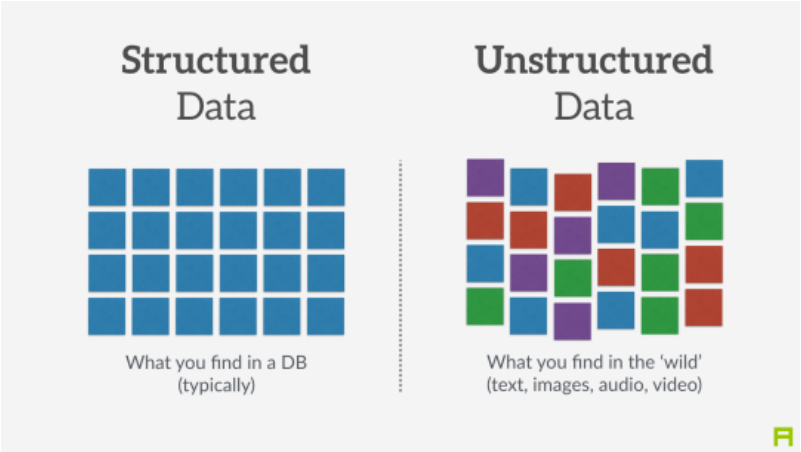


Figure 15. Structured vs unstructured data

V for Velocity

In a nutshell, the speed of creation and communication of data is accelerating[6]:

- Facebook hosts 250 billion pics? It receives 900 million more pictures **per day**
- Examining tweets can be done automatically (with computers). If you want to connect to Twitter to receive tweets in real time as they are tweeted, be prepared to receive in excess of 500 million tweets **per day**. Twitter calls this service the "Twitter firehose"[7], which reflects the velocity of the stream of tweets.
- **Sensor data** is bound to increase speed as well. While pictures, tweets, individual records… are single item data sent at intervals, more and more sensors can send data **in a continuous stream** (measures of movement, sound, etc.)

So, velocity poses challenges of its own: while a system can handle (store, analyze) say 100Gb of data in a given time (day or month), it might not be able to do it in say, a single second. Big data refers to the problems and solutions raised by the velocity of data.

A 4th V can be added, for Veracity

Veracity relates to trustworthiness and compliance: is the data authentic? Has it been corrupted at any step of its processing? Does it comply with local and international regulations?

What is the minimum size to count as "big data"? It’s all relative

There is no "threshold" or "minimum size" of a dataset where "data" would turn from "small data" to "big data".

It is more of a **relative** notion: it is big data if current IT systems struggle to cope with the datasets.

"Big data" is a relative notion… how so?

a. relative to time

- what was considered "big data" in the early 2000s would be considered "small data" today, because we have better storage and computing power today.
- this is a never ending race: as IT systems improve to deal with "current big data", data gets generated in still larger volumes, which calls for new progress / innovations to handle it.

b. relative to the industry

- what is considered "big data" by non tech SMEs (small and medium-sized enterprises) can be considered trivial to handle by tech companies.

c. not just about size

- the difficulty for an IT system to cope with a dataset can be related to the size (try analyzing 2 Tb of data on your laptop…), **but also** related to the content of the data.
- For example the analysis of customer reviews in dozens of languages is harder than the analysis of the same number of reviews in just one language.
- So the general rule is: the less the data is structured, the harder it is to use it, even if it's small in size (this relates to the "V" of variety seen above).

d. no correlation between size and value

- "Big data is often called the new oil"[8], as if it would flow like oil and would power engines "on demand".
- Actually, big data is **created**: it needs work, conception and design choices to even exist (what do I collect? how do I store it? what structure do I give to it?). The human intervention in creating data determines largely whether data will be of value later.
- Example: Imagine customers can write online reviews of your products. These reviews are data. But if you store these reviews without an indication of who has authored the review (maybe because reviews can be posted without login oneself), then the reviews become much less valuable.

Simple design decisions about how the data is collected, stored and structured have a huge impact on the value of the data.

So, in reaction to large, unstructured and badly curated datasets with low value at the end, a notion of "smart data" is sometimes put forward: data which can be small in size but which is well curated and annotated, enhancing its value (see also here[9]).

Where did big data come from?

a. Data got generated in bigger volumes because of the digitalization of the economy

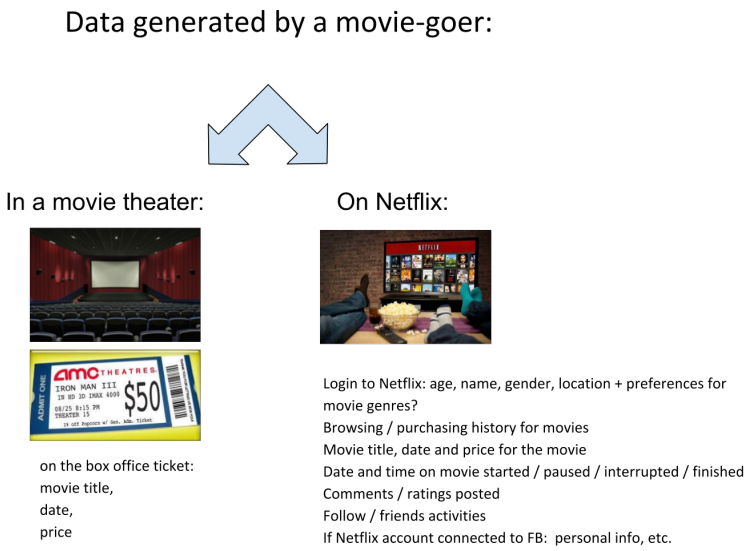
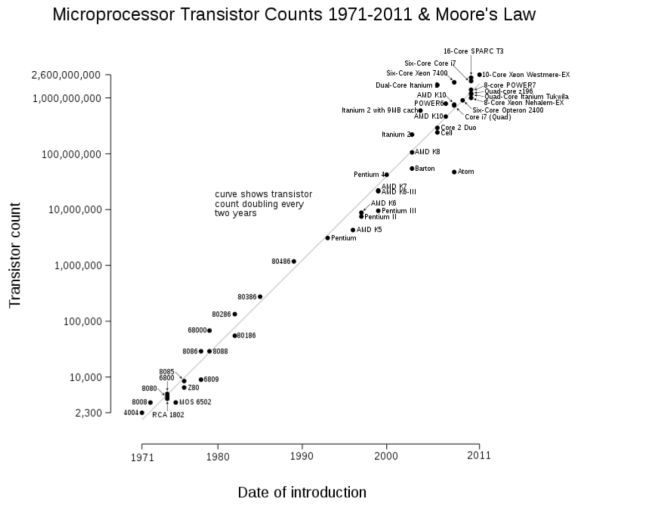


Figure 16. Movie theater vs Netflix

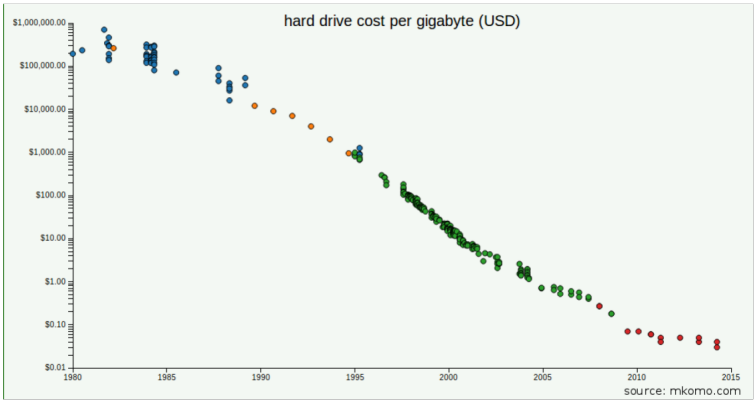
b. Computers became more powerful



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

Figure 17. Moore's law

c. Storing data became cheaper every year



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

Figure 18. Decreasing costs of data storage

d. The mindset changed as to what "counts" as data

- Unstructured data (see above for definition of "unstructured") was usually not stored: it takes a lot space, and software to query it was not sufficiently developped.
- Network data (also known as graphs) (who is friend with whom, who likes the same things as whom, etc.) was usually neglected as "not true observation", and hard to query. Social networks like Facebook made a lot to make businesses aware of the value of graphs (especially social graphs[10]).
- Geographical data has democratized: specific (and expensive) databases existed for a long time to store and query "place data" (regions, distances, proximity info...) but easy-to-use solutions have multiplied recently.

e. With open source software, the rate of innovation accelerated

In the late 1990s, a rapid shift in the habits of software developers kicked in: they tended to use more and more open source software, and to release their software as open source. Until then, most of the software was "closed source": you buy a software **without the possibility** to reuse / modify / augment its source code. Just use it as is.

Open source software made it easy to get access to software built by others and use it to develop new things. Today, all the most popular software in machine learning are free and open source.

See the Wikipedia article for a developed history of open source software: https://en.wikipedia.org/

wiki/History_of_free_and_open-source_software

f. Hype kicked in

The Gartner hype cycle[11] is a tool measuring the maturity of a technology, differentiating expectations from actual returns:

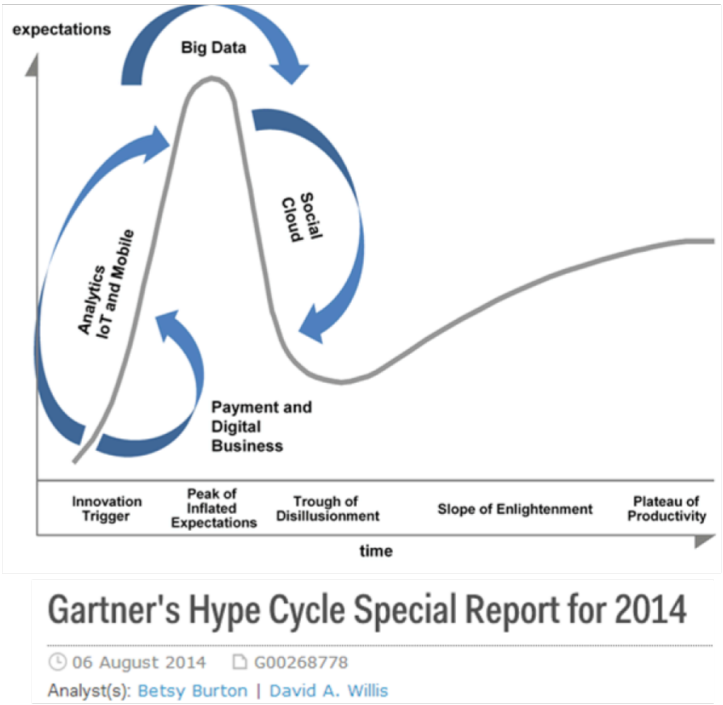


Figure 19. Gartner Hype Cycle for 2014

This graph shows the pattern that all technologies follow along their lifetime:

- at the beginning (left of the graph), an invention or discovery is made in a research lab, somewhere. Some news reporting is done about it, but with not much noise.
- then, the technology starts picking the interest of journalists, consultant, professors, industries… expectations grow about the possibilities and promises of the tech. "With it we will be able to [insert amazing thing here]"
- the top of the bump is the "peak of inflated expectations". All techs tend to be hyped and even over hyped. This means the tech is expected to deliver more than it surely will, in actuality. People get overdrawn.
- then follows the "Trough of Disillusionment". Doubt sets in. People realize the tech is not as powerful, easy, cheap or quick to implement as it first seemed. Newspapers start reporting depressing news about the tech, some bad buzz spreads.
- then: slope of Enlightenment. Heads get colder, expectations get in line with what the tech can actually deliver. Markets stabilize and consolidate: some firms close and key actors continue to grow.

- then: plateau of productivity. The tech is now mainstream.

(all technology can "die" - fall into disuse - before reaching the right side of the graph of course).

In 2014, big data was near the top of the curve: it was getting a lot of attention but its practical use in 5 to 10 years were still uncertain. There were "great expectations" about its future, and these expectations drive investment, research and business in big data.

In 2017, "big data" is still on top of hyped technologies, but is broken down in "deep learning" and "machine learning". Note also the "Artificial General Intelligence" category:

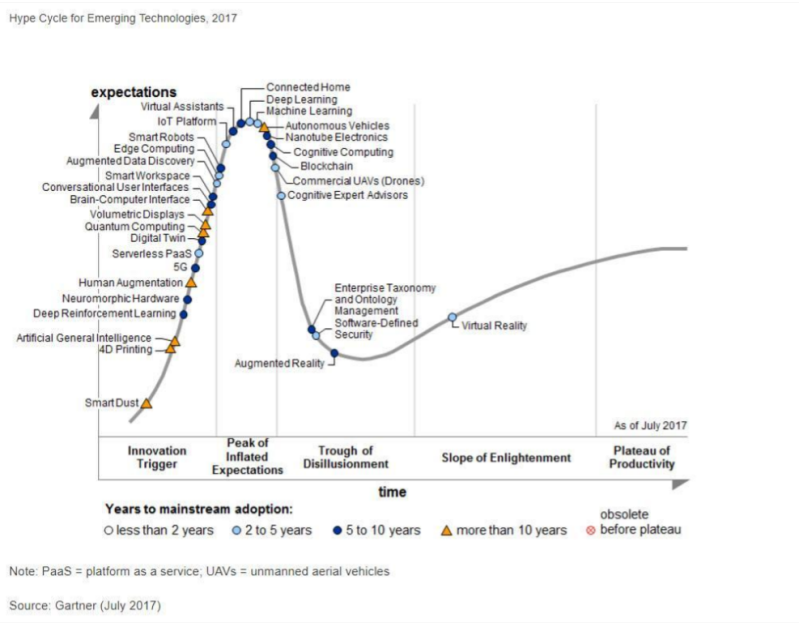


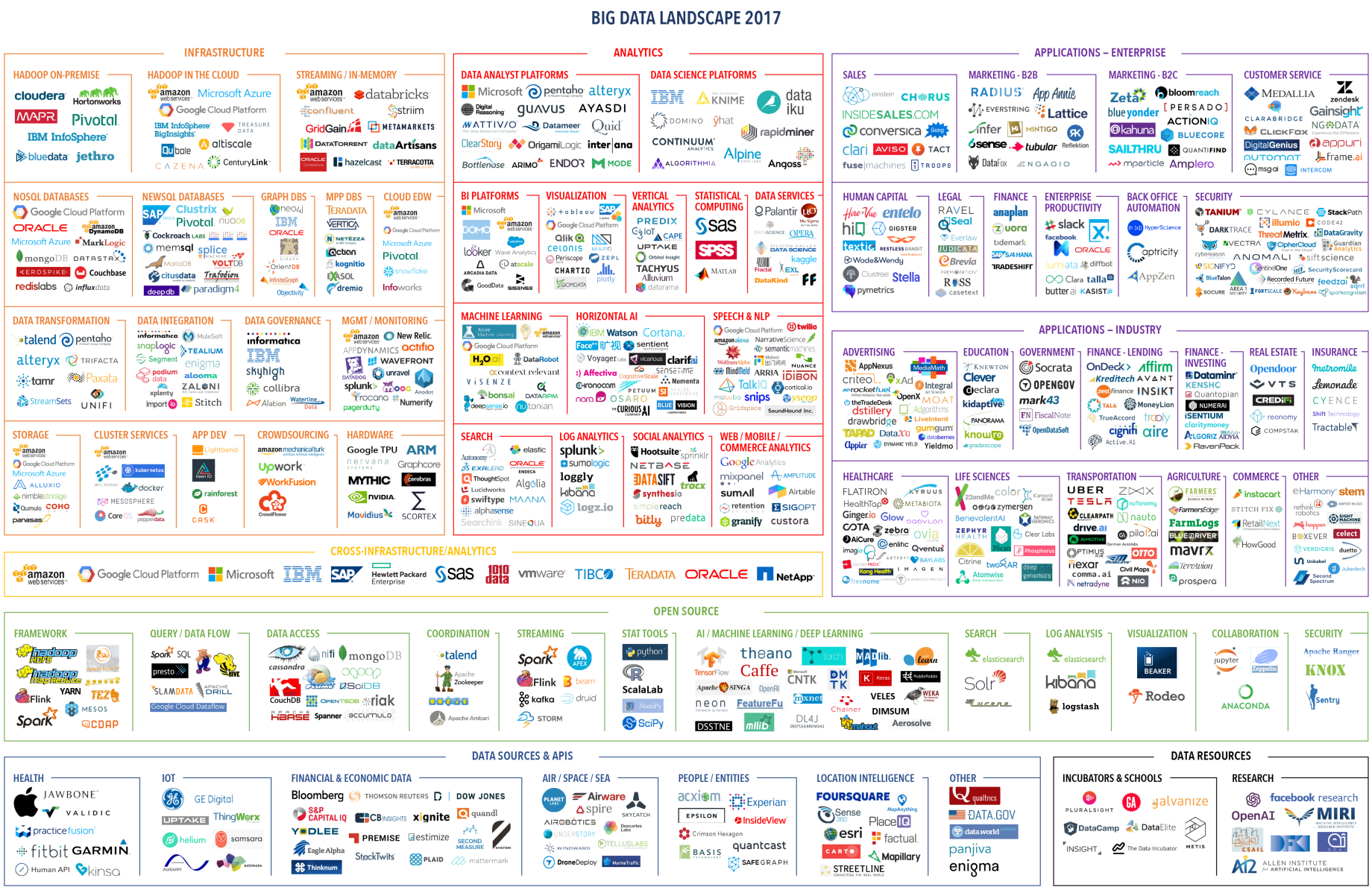
Figure 20. Gartner Hype Cycle for 2017

g. Big data transforms industries, and has become an industry in itself

Firms active in "Big data" divide in many sub-domains: the industry to manage the IT infrastructure for big data, the consulting firms, software providers, industry-specific applications, etc...

Matt Turck, VC at FirstMarkCap[12], creates every year a sheet to visualize the main firms active in these subdomains.

This is the 2017 version:



V2 – Last updated 5/3/2017

 mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Figure 21. Big data landscape for 2017

You can find an high res version of the Big data panorama[13], an Excel sheet version, and a very interesting comment on this website: https://mattturck.com/bigdata2017/

What is the future of big data?

a. More data is coming

The **Internet of things** designates the extension of Internet to objects beyond web pages or emails[14].

The **IoT** is used to **do** things (display information on screen, pilote robots, etc.) but also very much to **collect data** in their environments, through sensors.

Hence, the development of **connected objects** will lead to a tremendous increase in the volume of data collected.

b. Regulatory frameworks will grow in complexity

Societal impacts of big data and AI are not trivial, ranging from racial, financial and medical discrimination to giant data leaks, or economic (un)stability in the age of robots and AI in the workplace.

Public regulations at the national and international levels are trying to catch up with these challenges. As technology evolves quickly, we can anticipate that societal impacts of big data will take center stage.

c. as an expression, "big data" is evolving

- It is interesting to note that "hot" expressions, like "big data", tend to wear out fast. They are too hyped, used in all circumstances, become vague and over sold. For big data, we observe that it is peaking in 2017, while new terms appear:

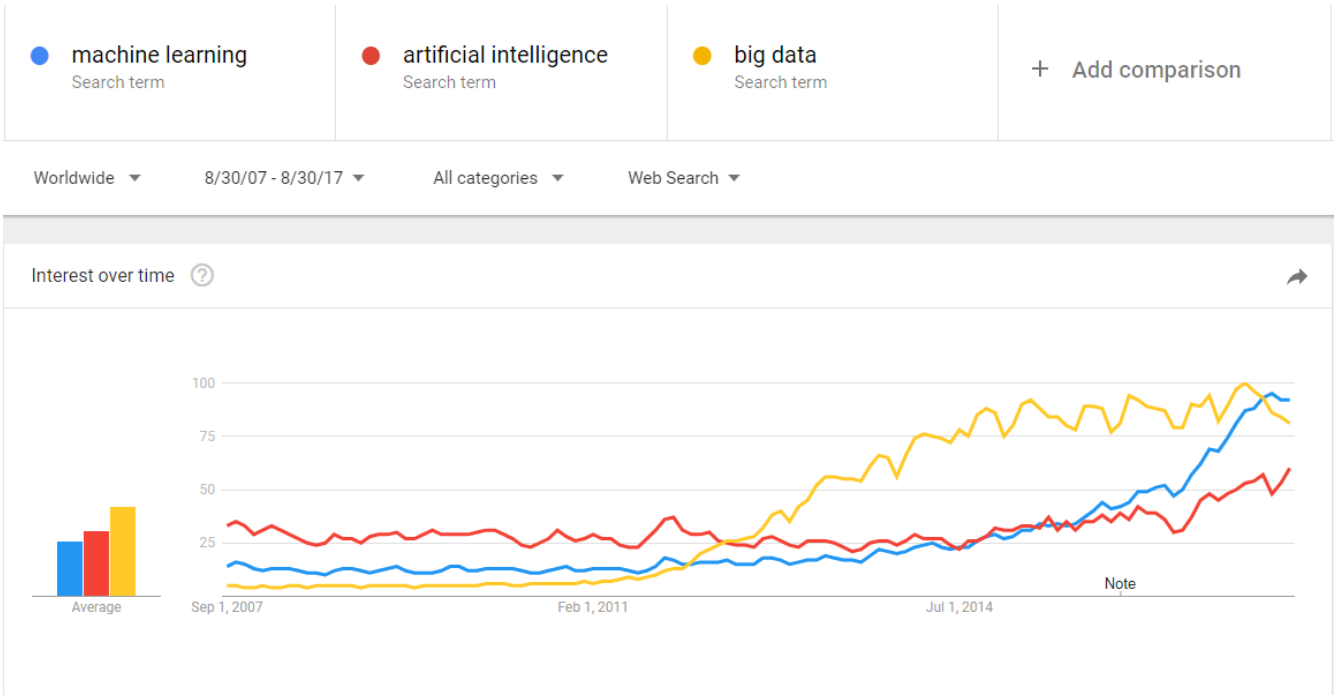


Figure 22. Google searches for big data, machine learning and AI

What are the differences between these terms?

- "Big data" is by now a generic term
- **Machine learning** puts the focus on the scientific and software engineering capabilities enabling to do something useful with the data (predict, categorize, score...)
- **Artificial intelligence** puts the emphasis on human-like possibilities afforded by machine learning. Often used interchangeably with machine learning. AI is fed on data, so the future of big data will intersect with what AI becomes.
- And **data science** ? This is a broad term encompassing machine learning, statistics, and many analytical methods to work with data and interpret it. Often used interchangeably with machine learning. **Data scientist** is a common job description in the field.

h machine learning. **Data scientist** is a common job description in the field.

References

1. *"datum", a Latin word for "a given"*: http://www.etymonline.com/index.php?term=data
2. *NSA and metadata*: http://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata
3. *csv, json and xml*: https://codingislove.com/json-tutorial-indepth/
4. *90% of all the data in the world has been generated over the last two years*: https://www.sciencedaily.com/releases/2013/05/130522085217.htm
5. *more data will be created in 2017 than the previous 5,000 years of humanity*: https://appdevelopermagazine.com/4773/2016/12/23/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity-/
6. *the speed of creation and communication of data is accelerating*: http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/
7. *"Twitter firehose"*: http://support.gnip.com/apis/firehose/
8. *"Big data is often called the new oil"*: https://hbr.org/2012/11/data-humans-and-the-new-oil
9. *here*: https://www.quora.com/After-Big-Data-Smart-Data-is-a-trend-in-2013-So-what-is-Smart-Data-Have-any-clear-definition
10. *social graphs*: https://en.wikipedia.org/wiki/Social_graph
11. *Gartner hype cycle*: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
12. *Matt Turck, VC at FirstMarkCap*: https://twitter.com/mattturck
13. *high res version of the Big data panorama*: https://mattturck.com/bigdata2017/
14. *extension of Internet to objects beyond web pages or emails*: https://seinecle.github.io/IoT4Entrepreneurs/

Index

Topics in Data Science for business: Volume 1 - Fundamentals

Clément Levallois <levallois@em-lyon.com [mailto:levallois@em-lyon.com]> v1.0, April 2018 :icons!: :iconsfont: font-awesome :revnumber: 1.0 :media: prepress :example-caption!: :sourcedir: ../../../main/java :toc: :toclevels: 1

Preface

A textbook for managers

The target reader for this book is a manager who needs to clearly understand what "data science", "big data", "artificial intelligence" so that they can:

- **leverage** these technologies to improve the efficiency of their existing business,
- **innovate** with new products and services and develop new business guidelines

The promise of this book is to bring you from a starting point with no knowledge of these technical concepts, to a point where you understand the concepts **and** you can develop "data centric" business projects: when "data" contributes to creating value for the customer and all stakeholders.

Is this textbook too technical or too easy for me?

If you are unsure, try this simple test: http://bit.ly/essentials-1-test

→ There are 20 topics you should be comfortable answering. See how you score. If the score is low, this book will be of great use.

Chapter 1. Data, a concept with multiple layers

Definition of data

The English term "data" (1654) originates from "datum", a Latin word for "a given"[1]. "Data" is a single factual, a single entity, a single point of matter.

The word "data" to mean "transmittable and storable computer information" was first used in this sense in 1946. The expression "data processing" was first used in 1954.

Thoughts: the etymology suggests that data is "a given". Can you question this?

Data represents either a single entity, or a collection of such entities ("data points"). We can speak also of datasets instead of data (so a dataset is a collection of data points).

The variety of data sets

A date	A color	A grade
A relation of friendship	A sound	A heartbeat
A user input	A duration	A curriculum vitae

A picture	A longitude and latitude	A price
A number of friends	A temperature	A list of favorite movies
etc…	etc…	etc…

These examples are chosen on purpose to be varied and from unexpected places. They illustrate three principles:

a. Think about data in a broad sense

Data is not just numerical, neither is it "what sits in my spreadsheets". You should train in thinking about data in a broader sense:

- pictures are data
- language is data (including slang, lip movements, etc.)
- relations are data: individual A is known, individual B is known, **but the relationship between A and B is data as well**
- preferences, emotional states… are data

- etc. There is no definitive list, you should train yourself looking at business situations and think: "where is the data?"

b. metadata is data, too

Metadata is a piece of data describing another data.

Example:

```
The bibliographical reference ①
describing
a book ②
```

① the metadata

② the data

→ Data without metadata can be worthless (imagine a library without a library catalogue)

→ Metadata can be informative in its own right, as shown with the NSA scandal (read this article from the New Yorker about NSA and metadata[2]).

c. zoom in, zoom out

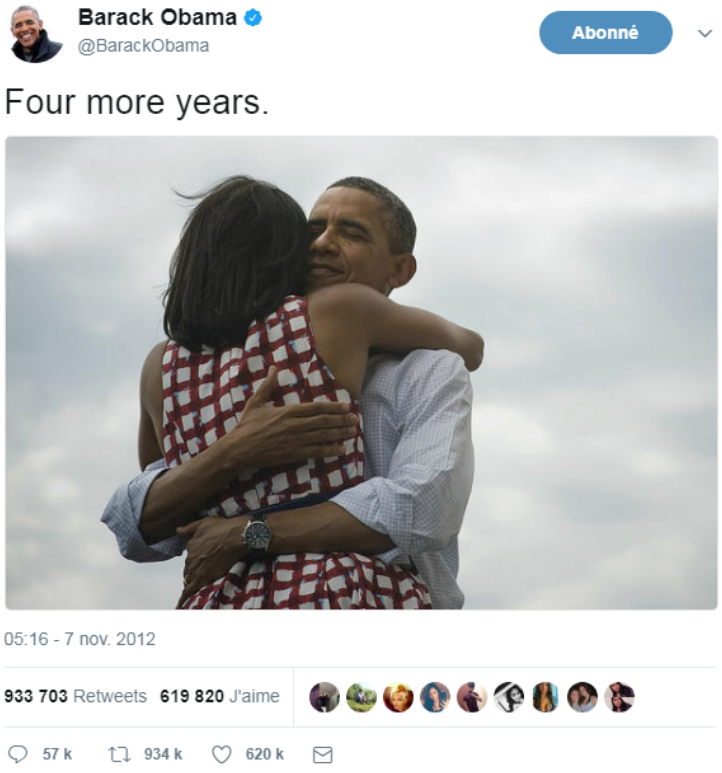
We should remember considering that a data point can be itself a collection of data points:

- a person walking into a building is a data point.
- however this person is itself a collection of data points: location data + network relations + subscriber status to services + etc.

So it is a good habit to wonder whether a data point can in fact be "unbundled" (spread into smaller data points / measurements)

How to describe datasets

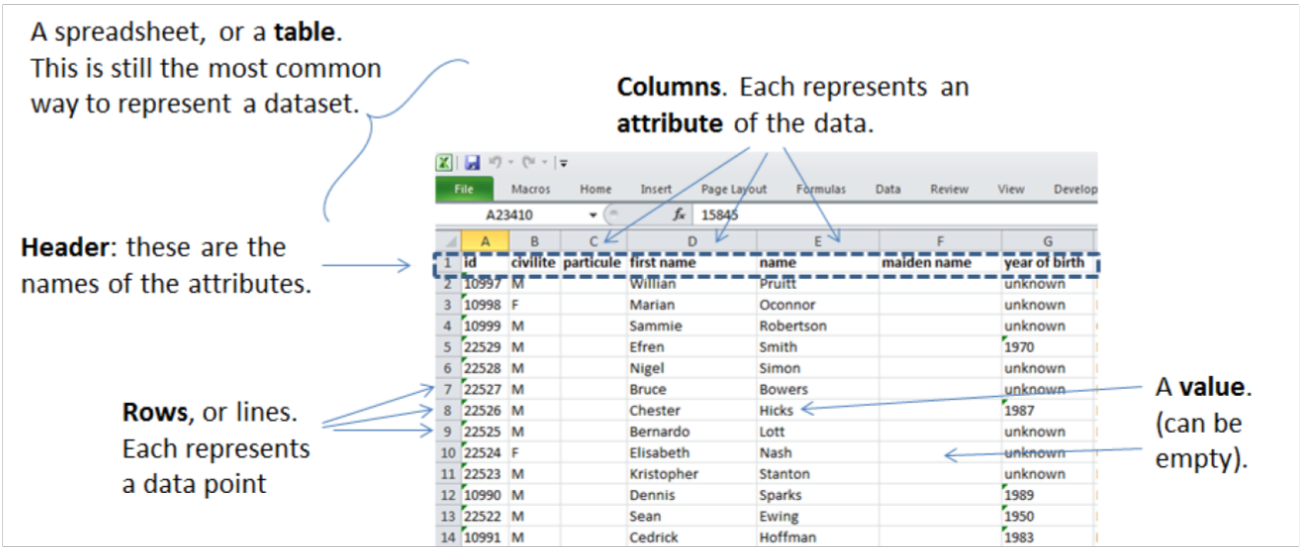
a. Formats, types, encoding



- This is a digital **medium** (because it's on screen as opposed to analogical, if we had printed the pic on paper)
- The **type** of the data is textual + image
- The text is formatted in **plain text** (meaning, no special formatting), as opposed to **data-interchange formats** which are formatting marks added to the text to facilitate its readability by software (check csv, json and xml[3]).
- The **encoding** of the text is UTF-8 (one of encodings deriving from the Unicode standard). Encoding deals with the issue: how to represent alphabets, signs (for instance: emojis) and symbols, from different languages, in text? UTF-8 is an encoding which is one of the most universal.
- The tweet is part of a list of tweets. The list represents the **data structure** of the dataset, it is the way the data is organized. There are many alternative data structures: arrays, sets, dics, maps...
- The tweet is stored as a picture (png file) on the hard disk. "png" is the **file format**. The data is **persisted** as a file on disk (could have been stored in a database instead).

b. Tabular data

Tabular data is a common way to handle datasets, by organizing it in lines and columns:



	A	B	C	D	E	F	G
1	id	civilite	particule	first name	name	maiden name	year of birth
2	10997	M		William	Pruitt		unknown
3	10998	F		Marian	Oconnor		unknown
4	10999	M		Sammie	Robertson		unknown
5	22529	M		Efren	Smith		1970
6	22528	M		Nigel	Simon		unknown
7	22527	M		Bruce	Bowers		unknown
8	22526	M		Chester	Hicks		1987
9	22525	M		Bernardo	Lott		unknown
10	22524	F		Elisabeth	Nash		unknown
11	22523	M		Kristopher	Stanton		unknown
12	10990	M		Dennis	Sparks		1989
13	22522	M		Sean	Ewing		1950
14	10991	M		Cedrick	Hoffman		1983

Figure 23. tabular data

c. First party, second party and third party data

- **First party data** : the data generated through the activities of your own organization. Your organization own it, which does not mean that consent from users is not required, when it comes to personal data.
- **Second party data** : the data accessed through partnerships. Without being the generator nor the owner of this data, partners make it available to you through an agreement.
- **Third party data** : the data acquired via purchase. This data is acquired through a market transaction. Its uses still comes with conditions, especially for personal data.

d. Sociodemo data vs behavior data

- Sociodemogaphic or **sociodemio** data refers to information about individuals, describing fundamental attributes of their social identity: age, gender, place of residence, occupation, marital status and number of kids.
- **Behavior data** refers to any digital trace left by the individual in the course of it life: clicks on web pages, likes on Facebook, purchase transactions, comments posted on Tripadvisor…

Sociodemo data is typically well structured or easy to structure. It has a long history of collection and analysis, basically since census exists.

Behavior data allows to go further than sociodemo data: each individual can be characterized by its acts and tastes, well beyond what an age or marital status could define.

But behavior data is typically not well structured and harder to collect.

Data and size

1 bit		can store a binary value (yes / no, true / false…)
8 bits	1 byte (or octet)	can store a single character
~ 1,000 bytes	1 kilobyte (kb)	Can store a paragraph of text
~ 1 million bytes	1 megabyte (Mb)	Can store a low res picture.

~ 1 billion bytes	1 gigabyte (Gb)	Can store a movie
~ 1 trillion bytes	1 terabyte (Tb)	Can store 1,000 movies. Size of commercial hard drives in 2017 is 2 Tb.
~ 1,000 trillion bytes	1 petabyte (Pb)	20 Pb = Google Maps in 2013

Chapter 2. A clarification of big data

Big data is a mess

Figure 24. Facebook post by Dan Ariely in 2013

Jokes aside, defining big data and what it covers needs a bit of precision. Let's bring some clarity.

The 3 V

Big data is usually described with the "3 Vs":

V for Volume

The size of datasets available today is staggering (ex: Facebook had 250 billion pics in 2016).

We should also note that the volumes of data are increasing at an **accelerating rate**. According to sources, 90% of all the data in the world has been generated over the last two years[4] (statement from 2013) or said differently, more data will be created in 2017 than the previous 5,000 years of humanity[5].

V for Variety

"Variety" refers to the fact that "unstructured" data is considered to be increasingly useful, when before the big data phenoenon only structured data was considered worth storing and exploiting. This calls to explain in more details the distinction between unstructured and structured data.

A - Structured data

Structured data refers to data which is formatted and organized according to a well defined set of rules, which makes it **machine readable**. For example, zip codes are a structured dataset because

they follow a precise convention regarding the number of letters and digits composing them, making it easy for an optical reader and software to identify and "read" them. Same with license plates, social security numbers…

But these are simple examples.

What about, for instance, a tax form? If each field of the form is well defined, then the data collected through the form can be said to be "structured". By contrast, a form where the user can write free text (think of a comment on a blog post, or a blank space where users can write a feedback) produces unstructured data: data which does not follow a special convention for its size and content. This is typically much harder for software to process, hence to analyze.

To summarize, think of structured data as as anything that can be represented as well organized tables of numbers and short pieces of text with the expected format, size, and conventions of writing: phonebooks, accounting books, governmental statistics…

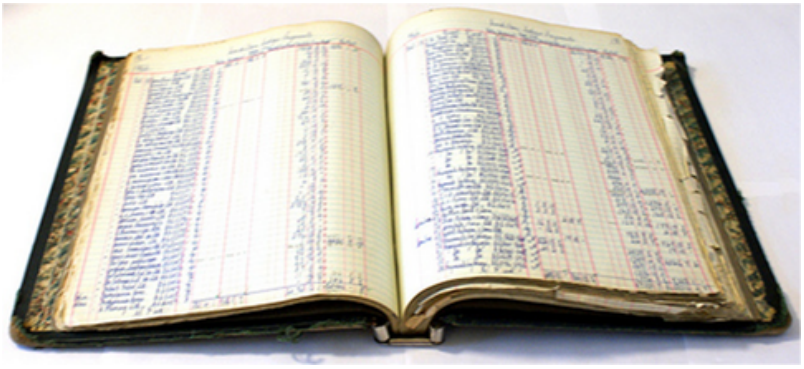


Figure 25. A book of accounts showing structured data

B - Unstructured data

Unstructured data refers to datasets made of "unruly" items: text of any length, without proper categorization, encoded in different formats, including possibly pictures, sound, geographical coordinates and what not…

These datasets are much harder to process and analyze, since they are full of exceptions and differences. But they are carry typically rich information: free text, information recorded "in the wild"…

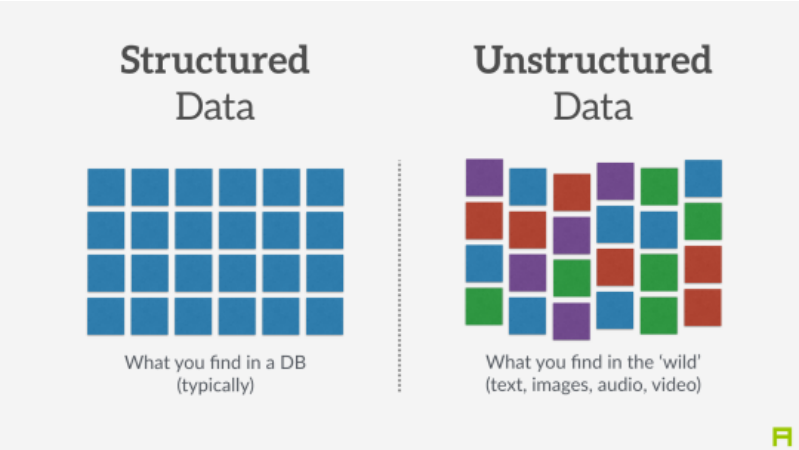


Figure 26. Structured vs unstructured data

V for Velocity

In a nutshell, the speed of creation and communication of data is accelerating[6]:

- Facebook hosts 250 billion pics? It receives 900 million more pictures **per day**
- Examining tweets can be done automatically (with computers). If you want to connect to Twitter to receive tweets in real time as they are tweeted, be prepared to receive in excess of 500 million tweets **per day**. Twitter calls this service the "Twitter firehose"[7], which reflects the velocity of the stream of tweets.
- **Sensor data** is bound to increase speed as well. While pictures, tweets, individual records… are single item data sent at intervals, more and more sensors can send data **in a continuous stream** (measures of movement, sound, etc.)

So, velocity poses challenges of its own: while a system can handle (store, analyze) say 100Gb of data in a given time (day or month), it might not be able to do it in say, a single second. Big data refers to the problems and solutions raised by the velocity of data.

A 4th V can be added, for Veracity

Veracity relates to trustworthiness and compliance: is the data authentic? Has it been corrupted at any step of its processing? Does it comply with local and international regulations?

What is the minimum size to count as "big data"? It’s all relative

There is no "threshold" or "minimum size" of a dataset where "data" would turn from "small data" to "big data".

It is more of a **relative** notion: it is big data if current IT systems struggle to cope with the datasets.

"Big data" is a relative notion… how so?

a. relative to time

- what was considered "big data" in the early 2000s would be considered "small data" today, because we have better storage and computing power today.
- this is a never ending race: as IT systems improve to deal with "current big data", data gets generated in still larger volumes, which calls for new progress / innovations to handle it.

b. relative to the industry

- what is considered "big data" by non tech SMEs (small and medium-sized enterprises) can be considered trivial to handle by tech companies.

c. not just about size

- the difficulty for an IT system to cope with a dataset can be related to the size (try analyzing 2 Tb of data on your laptop…), **but also** related to the content of the data.
- For example the analysis of customer reviews in dozens of languages is harder than the analysis of the same number of reviews in just one language.
- So the general rule is: the less the data is structured, the harder it is to use it, even if it's small in size (this relates to the "V" of variety seen above).

d. no correlation between size and value

- "Big data is often called the new oil"[8], as if it would flow like oil and would power engines "on demand".
- Actually, big data is **created**: it needs work, conception and design choices to even exist (what do I collect? how do I store it? what structure do I give to it?). The human intervention in creating data determines largely whether data will be of value later.
- Example: Imagine customers can write online reviews of your products. These reviews are data. But if you store these reviews without an indication of who has authored the review (maybe because reviews can be posted without login oneself), then the reviews become much less valuable.

Simple design decisions about how the data is collected, stored and structured have a huge impact on the value of the data.

So, in reaction to large, unstructured and badly curated datasets with low value at the end, a notion of "smart data" is sometimes put forward: data which can be small in size but which is well curated and annotated, enhancing its value (see also here[9]).

Where did big data come from?

a. Data got generated in bigger volumes because of the digitalization of the economy

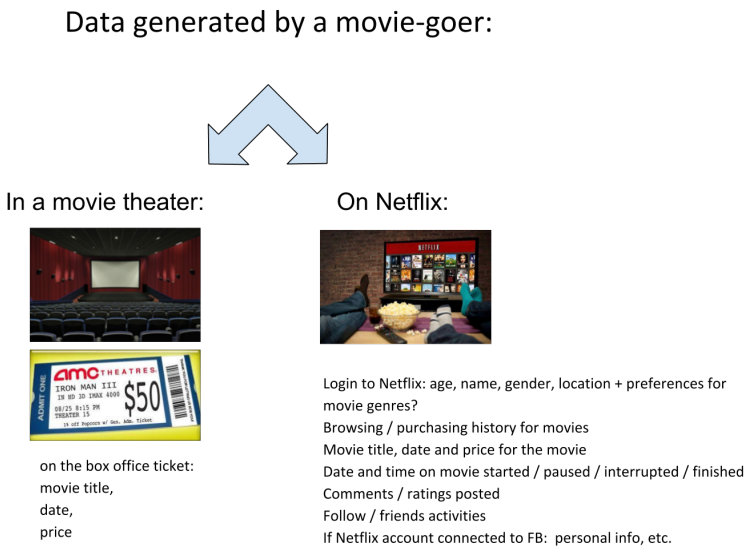
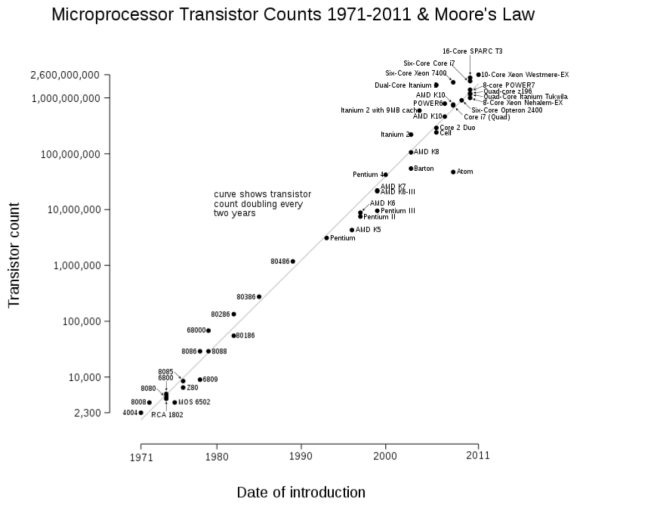


Figure 27. Movie theater vs Netflix

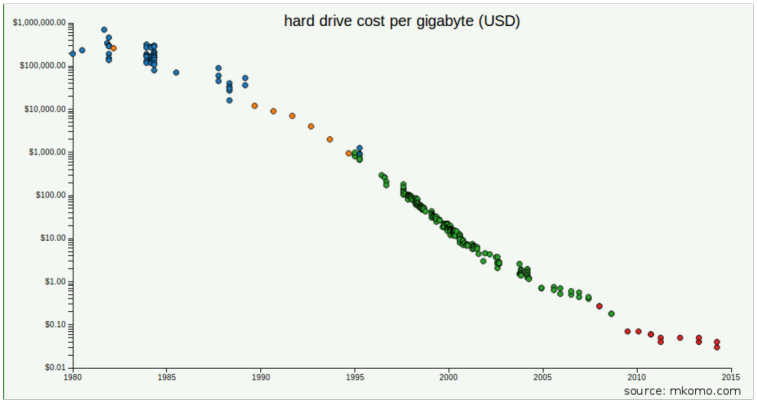
b. Computers became more powerful



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

Figure 28. Moore's law

c. Storing data became cheaper every year



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

Figure 29. Decreasing costs of data storage

d. The mindset changed as to what "counts" as data

- Unstructured data (see above for definition of "unstructured") was usually not stored: it takes a lot space, and software to query it was not sufficiently developped.
- Network data (also known as graphs) (who is friend with whom, who likes the same things as whom, etc.) was usually neglected as "not true observation", and hard to query. Social networks like Facebook made a lot to make businesses aware of the value of graphs (especially social graphs[10]).
- Geographical data has democratized: specific (and expensive) databases existed for a long time to store and query "place data" (regions, distances, proximity info...) but easy-to-use solutions have multiplied recently.

e. With open source software, the rate of innovation accelerated

In the late 1990s, a rapid shift in the habits of software developers kicked in: they tended to use more and more open source software, and to release their software as open source. Until then, most of the software was "closed source": you buy a software **without the possibility** to reuse / modify / augment its source code. Just use it as is.

Open source software made it easy to get access to software built by others and use it to develop new things. Today, all the most popular software in machine learning are free and open source.

See the Wikipedia article for a developed history of open source software: https://en.wikipedia.org/

wiki/History_of_free_and_open-source_software

f. Hype kicked in

The Gartner hype cycle[11] is a tool measuring the maturity of a technology, differentiating expectations from actual returns:

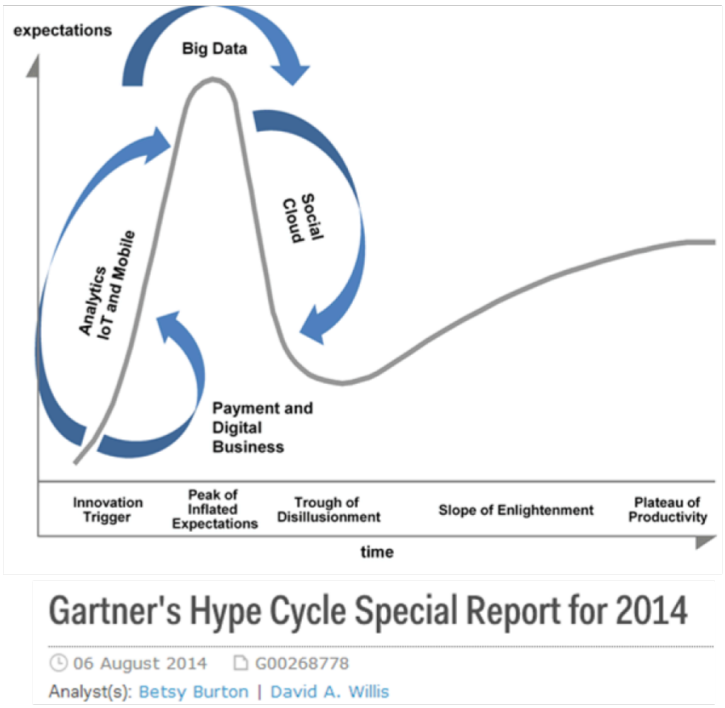


Figure 30. Gartner Hype Cycle for 2014

This graph shows the pattern that all technologies follow along their lifetime:

- at the beginning (left of the graph), an invention or discovery is made in a research lab, somewhere. Some news reporting is done about it, but with not much noise.
- then, the technology starts picking the interest of journalists, consultant, professors, industries… expectations grow about the possibilities and promises of the tech. "With it we will be able to [insert amazing thing here]"
- the top of the bump is the "peak of inflated expectations". All techs tend to be hyped and even over hyped. This means the tech is expected to deliver more than it surely will, in actuality. People get overdrawn.
- then follows the "Trough of Disillusionment". Doubt sets in. People realize the tech is not as powerful, easy, cheap or quick to implement as it first seemed. Newspapers start reporting depressing news about the tech, some bad buzz spreads.
- then: slope of Enlightenment. Heads get colder, expectations get in line with what the tech can actually deliver. Markets stabilize and consolidate: some firms close and key actors continue to grow.

- then: plateau of productivity. The tech is now mainstream.

(all technology can "die" - fall into disuse - before reaching the right side of the graph of course).

In 2014, big data was near the top of the curve: it was getting a lot of attention but its practical use in 5 to 10 years were still uncertain. There were "great expectations" about its future, and these expectations drive investment, research and business in big data.

In 2017, "big data" is still on top of hyped technologies, but is broken down in "deep learning" and "machine learning". Note also the "Artificial General Intelligence" category:

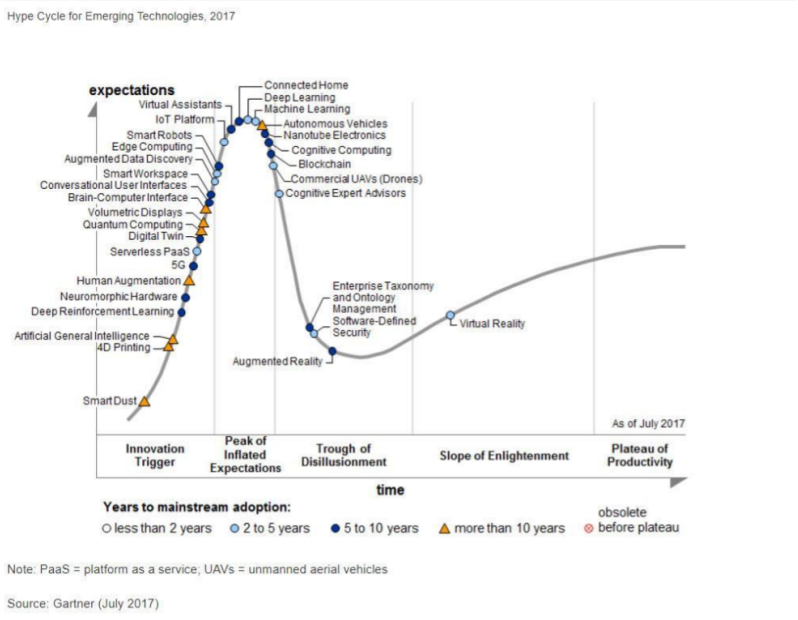


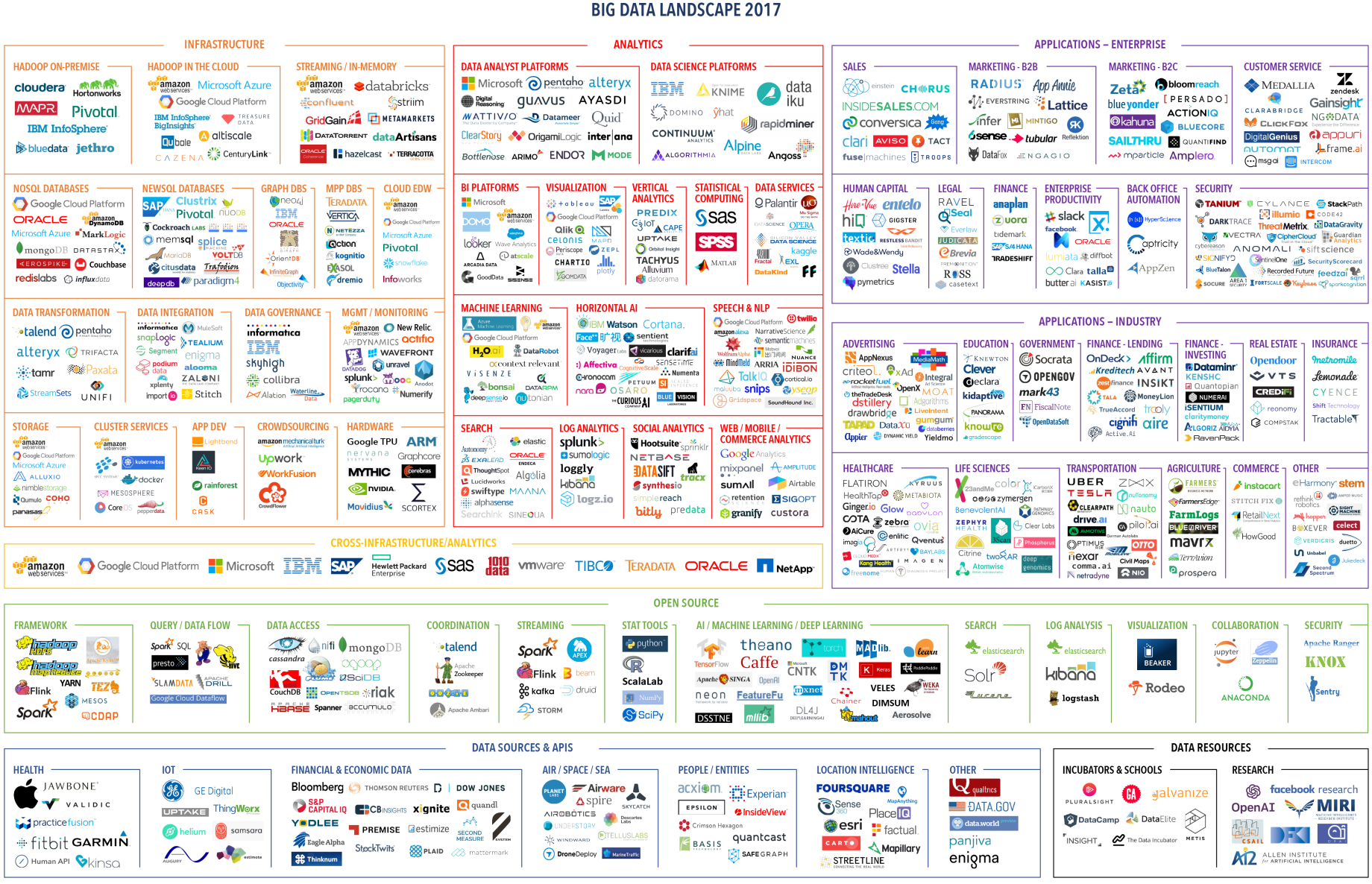
Figure 31. Gartner Hype Cycle for 2017

g. Big data transforms industries, and has become an industry in itself

Firms active in "Big data" divide in many sub-domains: the industry to manage the IT infrastructure for big data, the consulting firms, software providers, industry-specific applications, etc...

Matt Turck, VC at FirstMarkCap[12], creates every year a sheet to visualize the main firms active in these subdomains.

This is the 2017 version:



V2 – Last updated 5/3/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap) mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Figure 32. Big data landscape for 2017

You can find an high res version of the Big data panorama[13], an Excel sheet version, and a very interesting comment on this website: https://mattturck.com/bigdata2017/

What is the future of big data?

a. More data is coming

The **Internet of things** designates the extension of Internet to objects beyond web pages or emails[14].

The **IoT** is used to **do** things (display information on screen, pilote robots, etc.) but also very much to **collect data** in their environments, through sensors.

Hence, the development of **connected objects** will lead to a tremendous increase in the volume of data collected.

b. Regulatory frameworks will grow in complexity

Societal impacts of big data and AI are not trivial, ranging from racial, financial and medical discrimination to giant data leaks, or economic (un)stability in the age of robots and AI in the workplace.

Public regulations at the national and international levels are trying to catch up with these challenges. As technology evolves quickly, we can anticipate that societal impacts of big data will take center stage.

c. as an expression, "big data" is evolving

- It is interesting to note that "hot" expressions, like "big data", tend to wear out fast. They are too hyped, used in all circumstances, become vague and over sold. For big data, we observe that it is peaking in 2017, while new terms appear:

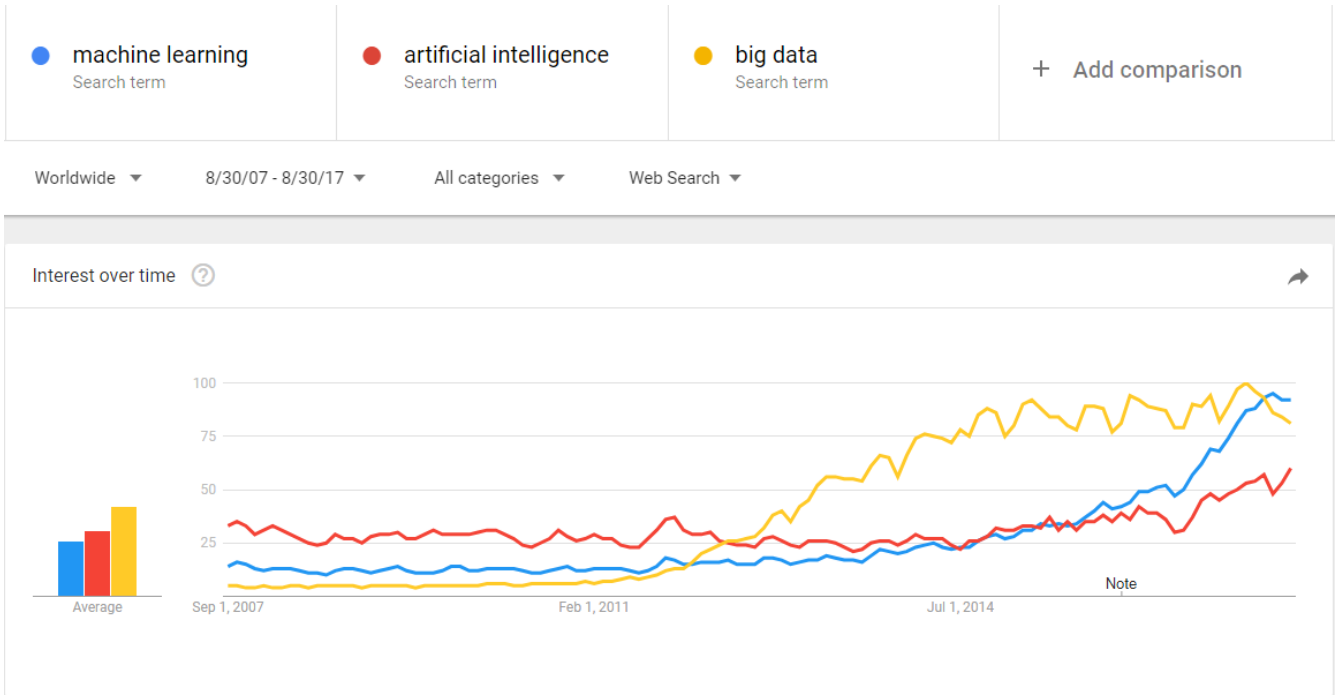


Figure 33. Google searches for big data, machine learning and AI

What are the differences between these terms?

- "Big data" is by now a generic term
- **Machine learning** puts the focus on the scientific and software engineering capabilities enabling to do something useful with the data (predict, categorize, score...)
- **Artificial intelligence** puts the emphasis on human-like possibilities afforded by machine learning. Often used interchangeably with machine learning. AI is fed on data, so the future of big data will intersect with what AI becomes.
- And **data science** ? This is a broad term encompassing machine learning, statistics, and many analytical methods to work with data and interpret it. Often used interchangeably with machine learning. **Data scientist** is a common job description in the field.

h machine learning. **Data scientist** is a common job description in the field.

References

1. *"datum", a Latin word for "a given"*: http://www.etymonline.com/index.php?term=data
2. *NSA and metadata*: http://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata
3. *csv, json and xml*: https://codingislove.com/json-tutorial-indepth/
4. *90% of all the data in the world has been generated over the last two years*: https://www.sciencedaily.com/releases/2013/05/130522085217.htm
5. *more data will be created in 2017 than the previous 5,000 years of humanity*: https://appdevelopermagazine.com/4773/2016/12/23/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity-/
6. *the speed of creation and communication of data is accelerating*: http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/
7. *"Twitter firehose"*: http://support.gnip.com/apis/firehose/
8. *"Big data is often called the new oil"*: https://hbr.org/2012/11/data-humans-and-the-new-oil
9. *here*: https://www.quora.com/After-Big-Data-Smart-Data-is-a-trend-in-2013-So-what-is-Smart-Data-Have-any-clear-definition
10. *social graphs*: https://en.wikipedia.org/wiki/Social_graph
11. *Gartner hype cycle*: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
12. *Matt Turck, VC at FirstMarkCap*: https://twitter.com/mattturck
13. *high res version of the Big data panorama*: https://mattturck.com/bigdata2017/
14. *extension of Internet to objects beyond web pages or emails*: https://seinecle.github.io/IoT4Entrepreneurs/

Index

ABOUT THE AUTHOR

Clément Levallois is *ancien élève* (graduate) de l'école normale supérieure de Cachan, *agrégé* in economics and management, PhD in economics and Associate Professor at em **lyon business school**.

He conducts research projects in data mining, data visualization and network analysis in various fields of social sciences, with scientific publications ranging from *History of Political Economy* to *Nature Reviews Neuroscience*. His teaching activities center on bridging social sciences and management with data science: establishing a common digital culture for students and executive participants.

Clément Levallois is a Java coder and an active supporter of Gephi, the leading open source and free software for network visualization.

His past and current projects can be seen at https://www.clementlevallois.net, and he can be reached on Twitter at @seinecle.

ESSENTIALS OF DATA FOR MANAGERS

From the fundamentals to Artificial Intelligence

Managers can hardly ignore the opportunities afforded by “big data”, an expression often used in relation with “data science” or “artificial intelligence”. But how to find the time to learn these complex notions, for the specific purpose of using them in a business context? This book offers a clear and complete presentation of the concepts and technologies a manager should know in order to conduct business projects where “data” is to play an essential role.

This volume leads you from the fundamentals to the comprehension of the stakes of artificial intelligence. The path starts with an overview of the building blocks of the “data revolution”: data, big data, cloud and APIs. These concepts are essential to be able to understand how the promises of artificial intelligence are connected to big data. Topics covered include:

- Unstructured vs structured data
- GDPR
- The cloud
- Machine learning
- Definition of big data
- Unsupervised and supervised learning
- APIs
- Artificial intelligence
- Data Management Platforms
- Sociodemo vs behavior data
- Business models based on data

This volume is an extended edition of two volumes published separately: “Essentials of Data for Managers, Volume 1: From big data to APIs” and “Essentials of Data for Managers, Volume 2: From artificial intelligence to business applications”.

CLÉMENT LEVALLOIS is *ancien élève* (graduate) l'école normale supérieure de Cachan and Associate Professor at em lyon business school, where he conducts research projects in data mining, data visualization and network analysis in various fields of social sciences. His teaching activities center on bridging managerial and IT cultures for students and executive participants.

ExpData Press