Big data and marketing

Clément Levallois

2017-16-10

Table of Contents

last modified: 2017-10-15

1. Big data is a mess

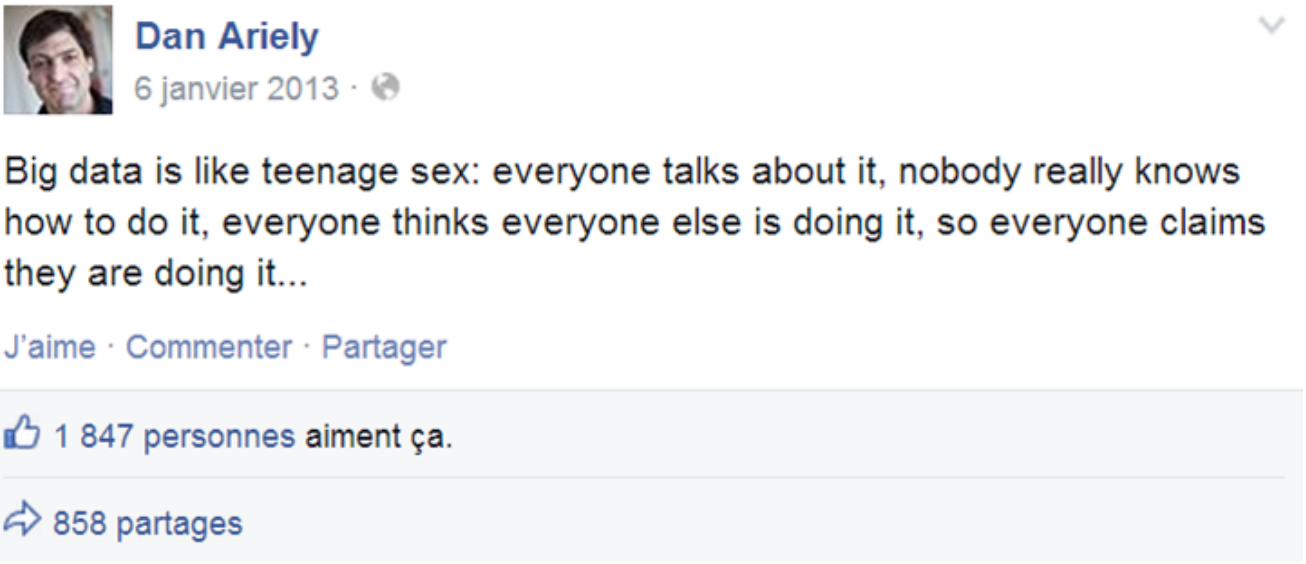


Figure 1. Facebook post by Dan Ariely in 2013

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

2. 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" (statement from 2013) or said differently, "More data will be created in 2017 than the previous 5,000 years of humanity"

V for Variety

This is a bit less intuitive. "Variety" means here that data is increasingly unstructured and messy, and this is an important characteristic of the "big data" phenomenon. To caricature a bit, try to picture a shift from A to B:

A - Structured data:

phonebooks, accounting books, governmental statistics… 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.

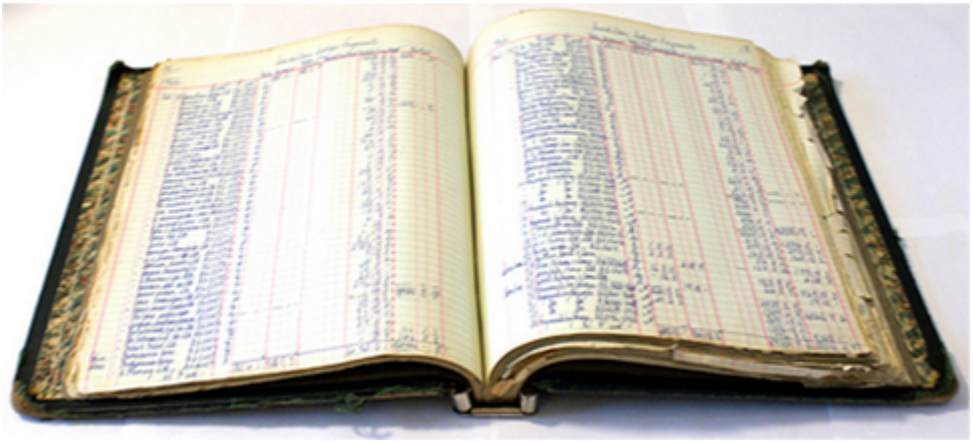


Figure 2. A book of accounts showing structured data

B - Unstructured data:

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…

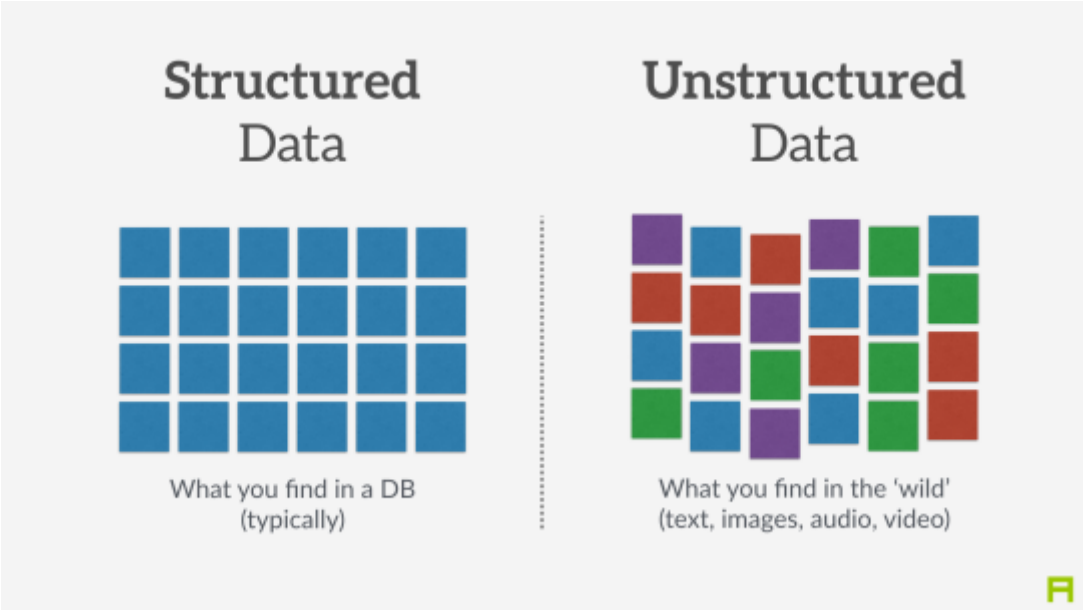


Figure 3. Structured vs unstructured data

V for Velocity

In a nutshell, the speed of creation and communication of data is accelerating (examples taken from here):

- 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 "firehose", which reflects the velocity of the stream of tweets.

Figure 4. The Twitter Firehose

- 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?

We will devote a session of this course to data compliance, which is a broad topic covering data privacy, cybersecurity, and the societal impacts of data.

You can start reading the documents for this course here

3. 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.

(see Wikipedia definition developing on this.)

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

1. 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.

2. relative to the industry

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

3. 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).

4. no correlation between size and value

- Big data is often called "the new oil", 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).

5. 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:

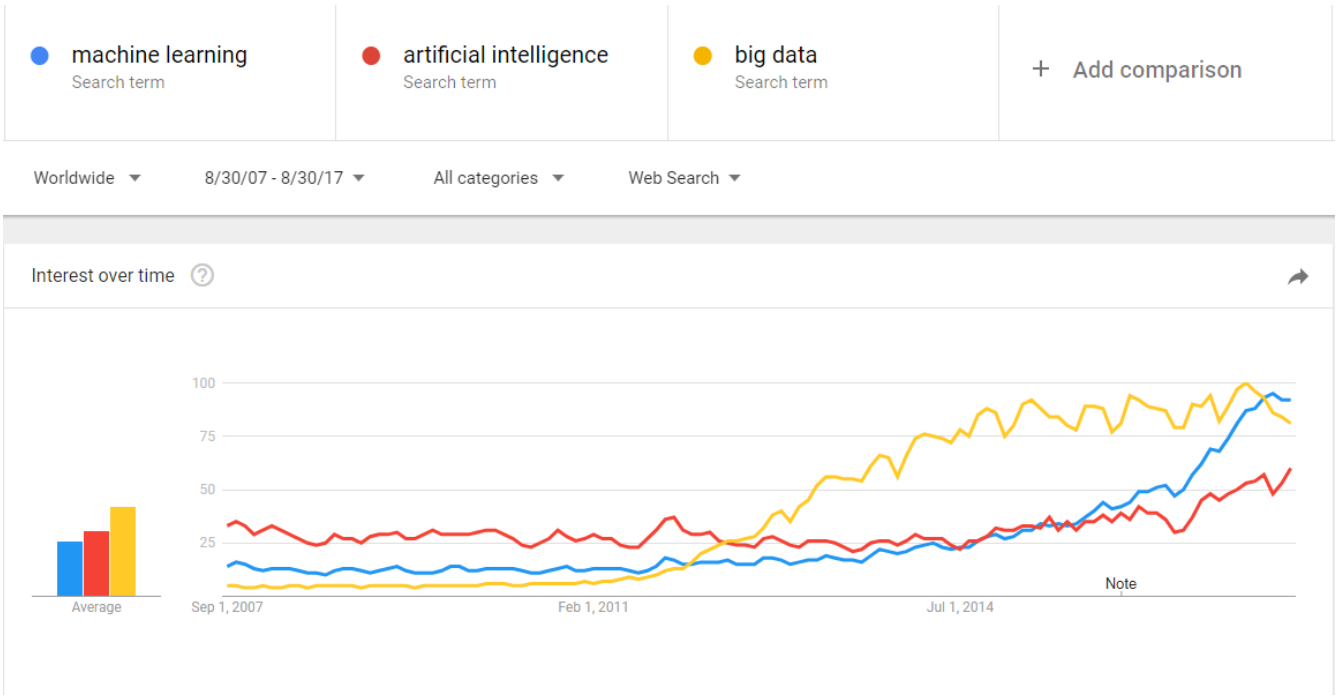


Figure 5. Google searches for big data, machine learning and AI

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.
- And "data science"? This is a broad term encompassing machine learning, statistics, ... and any 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.

4. Where did big data come from?

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

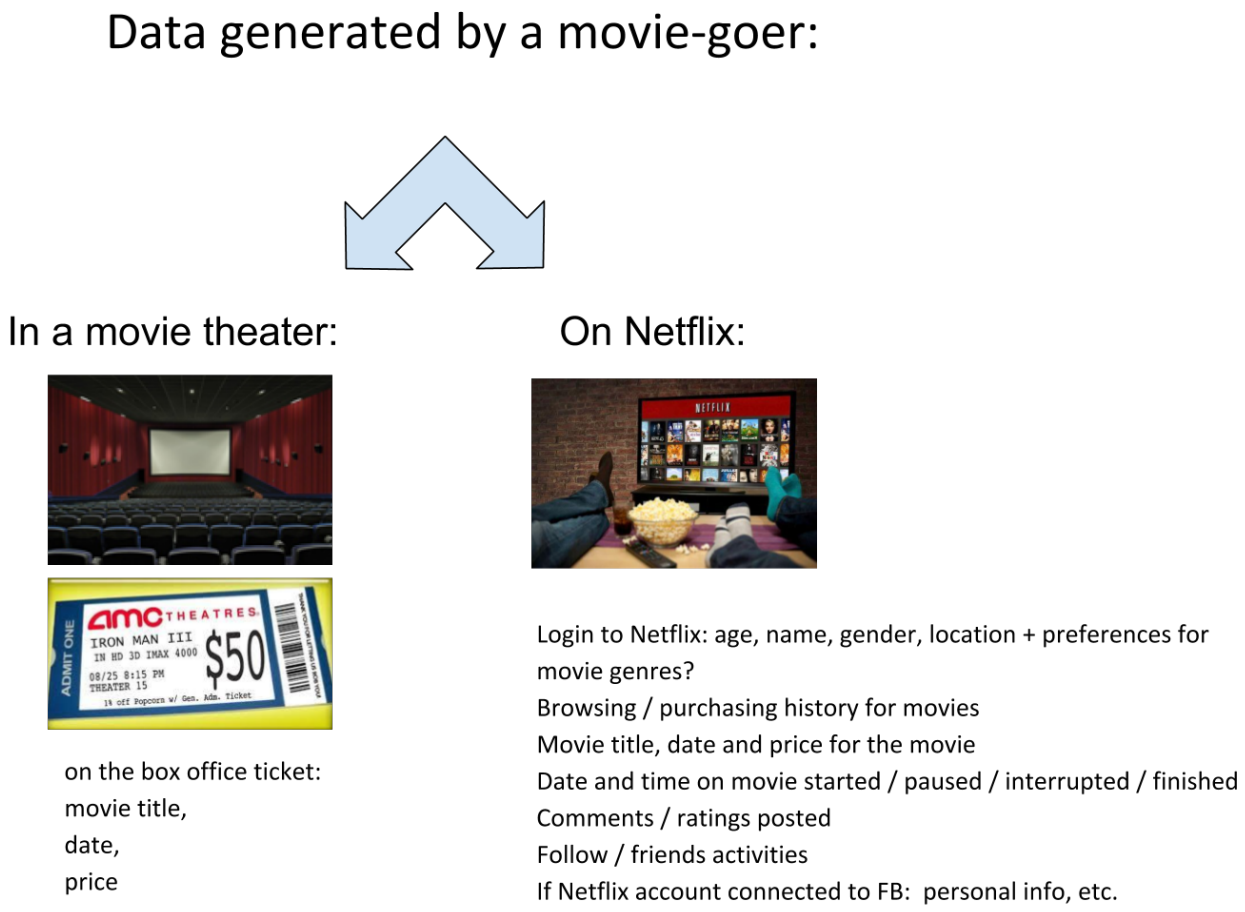
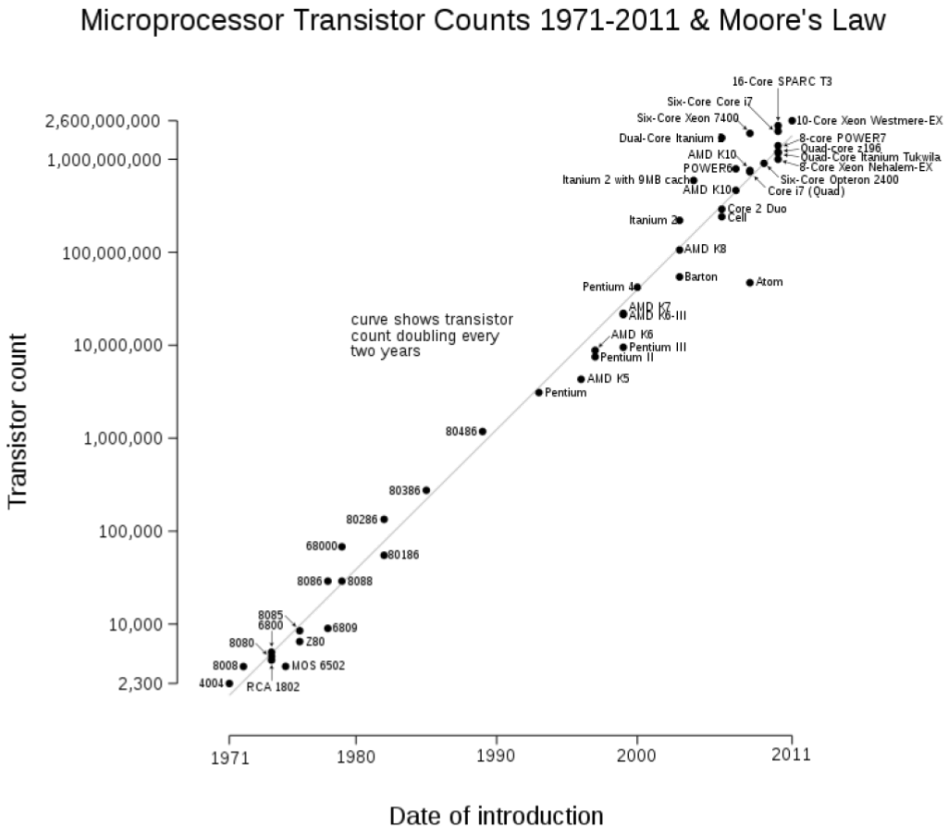


Figure 6. Movie theater vs Netflix

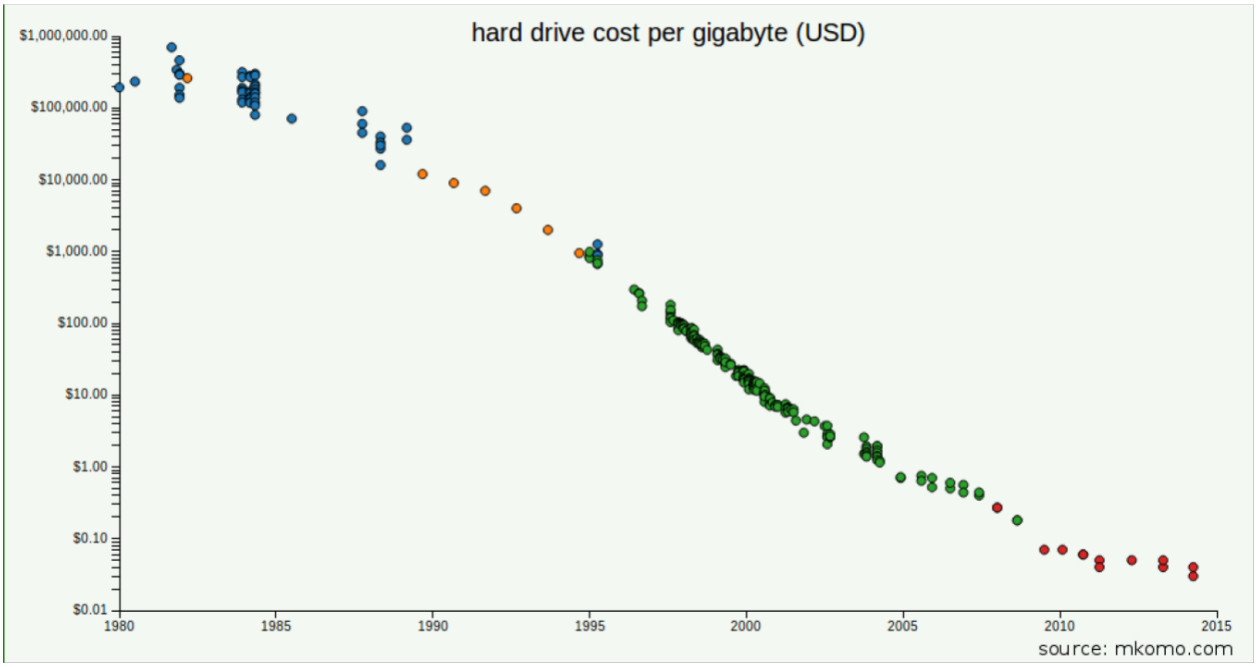
2. Computers became more powerful



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

Figure 7. Moore's law

3. Storing data became cheaper every year



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

Figure 8. Decreasing costs of data storage

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

- Unstructured (see above for definition of "unstructured") textual data 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).
- 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.

5. 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

6. Hype kicked in

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

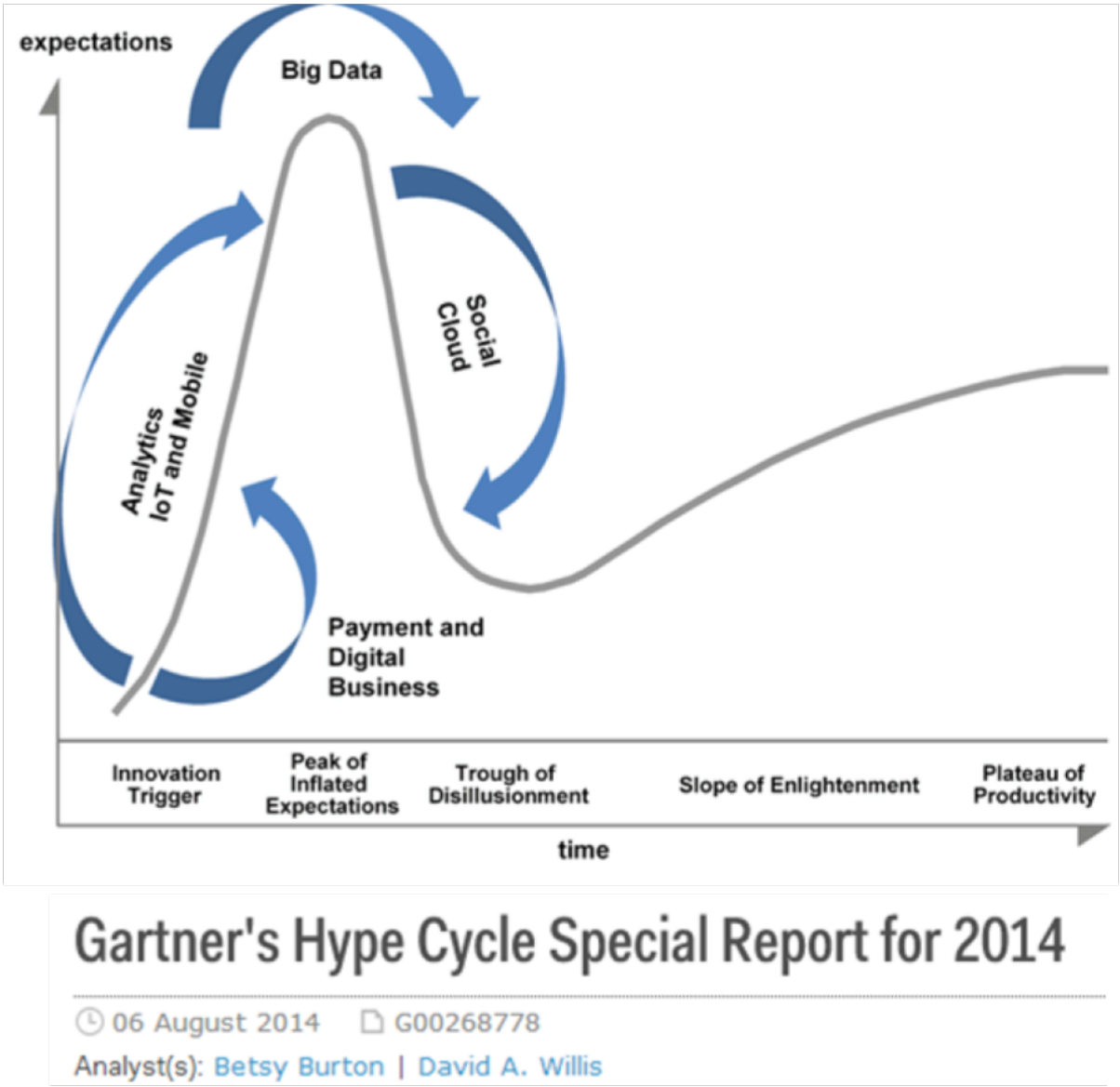


Figure 9. Gartner Hype Cycle for 2014

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:

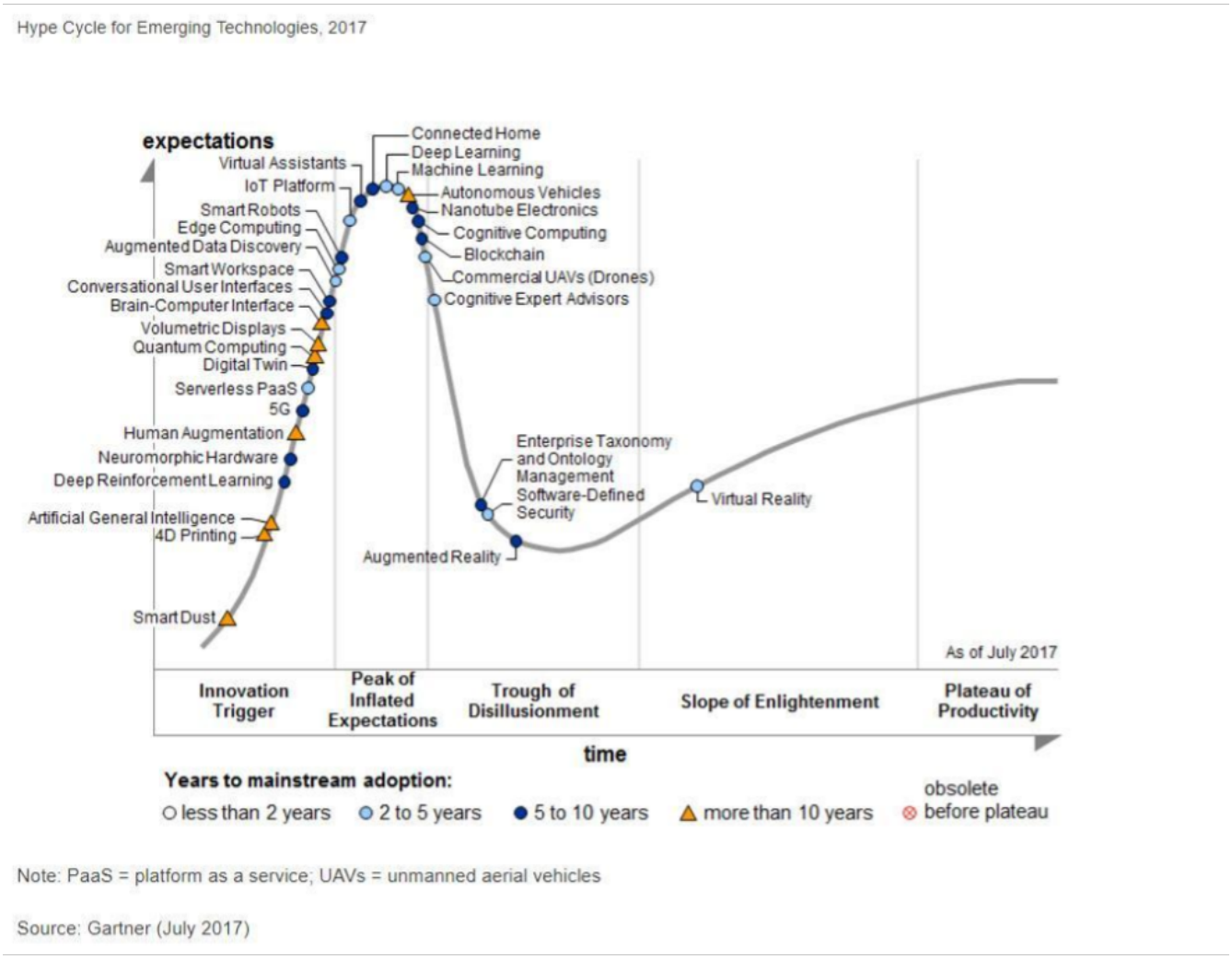


Figure 10. Gartner Hype Cycle for 2017

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

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

→ the field is huge.

Matt Turck, VC at FirstMarkCap, creates every year a sheet to visualize the main firms active in these subdomains. This is the 2017 version:

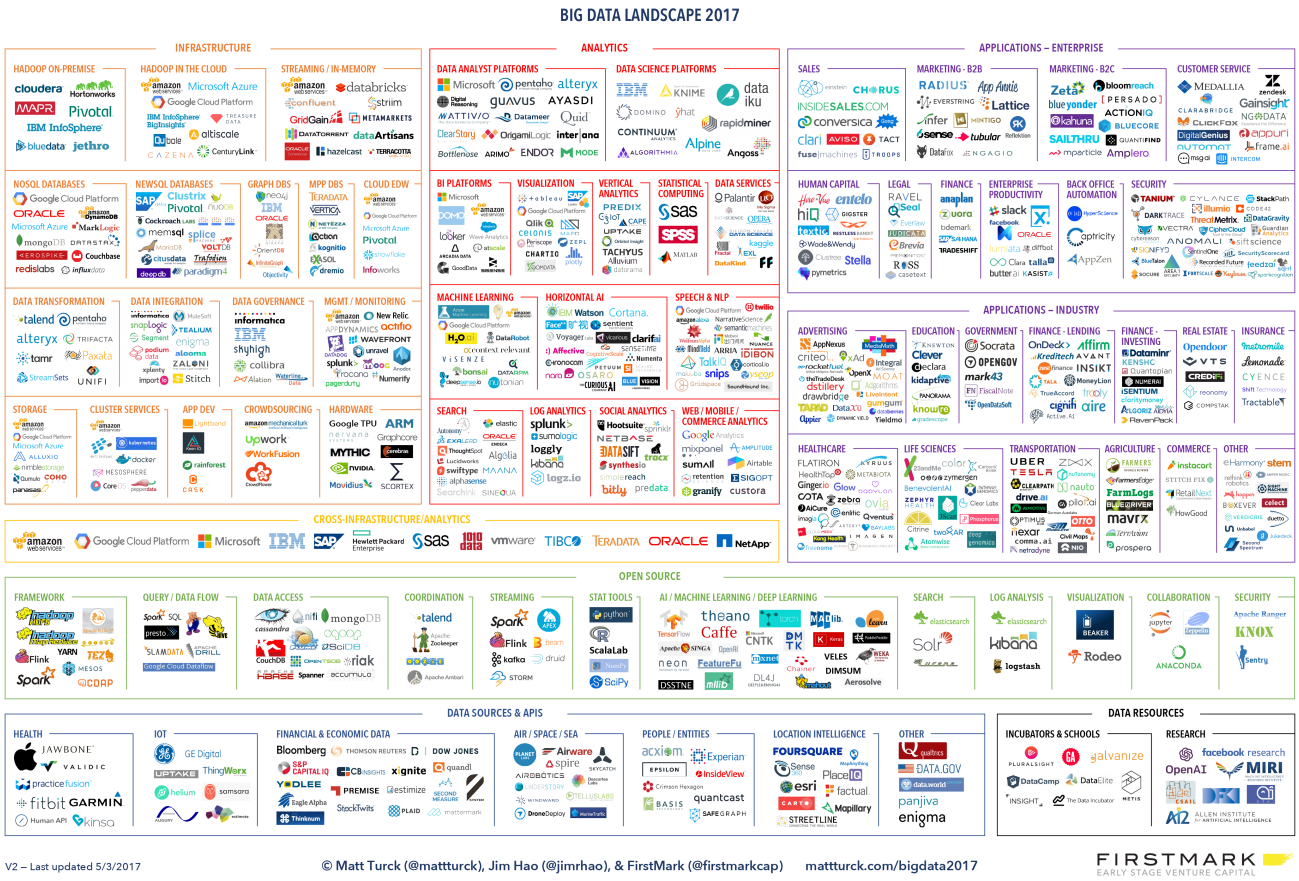


Figure 11. Big data landscape for 2017

You can find a high res version of this pic, an Excel sheet version, and a very interesting comment all here.

5. What is the future of big data?

1. More data is coming

The Internet of things (IoT) designates the extension of Internet to objects, not just web pages and emails (see here for details).

These connected objects are used to **do** things (display stuff on screen, pilote robots, etc.) but also very much to **collect data** in their environments (through sensors).

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

We have a session devoted to IoT later in this course. You can already starting reading the documents for this session:

- Internet of things

2. Discussions about big data will fuse with AI

Enthusiasm, disappointment, bad buzz, worries, debates, promises… the discourse about AI will grow. AI is fed on data, so the future of big data will intersect with what AI becomes.

We have a session devoted to data science / machine learning / AI later in this course. You can already start reading the documents for this course:

- What is data science?
- AI applications in business

3. 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.

6. Definition of CRM

CRM: acronym for "Customer Relationship Management"

A CRM is a software used to manage the commercial relationship between a company and its clients.

A CRM is part of the **information system** (IS) of the firm. The information system designates all software, human resources and procedures devoted to keep track of all info necessary to the business of the firm - from sales to production, etc.

The information system of a firm comprises many other blocks, besides the CRM:

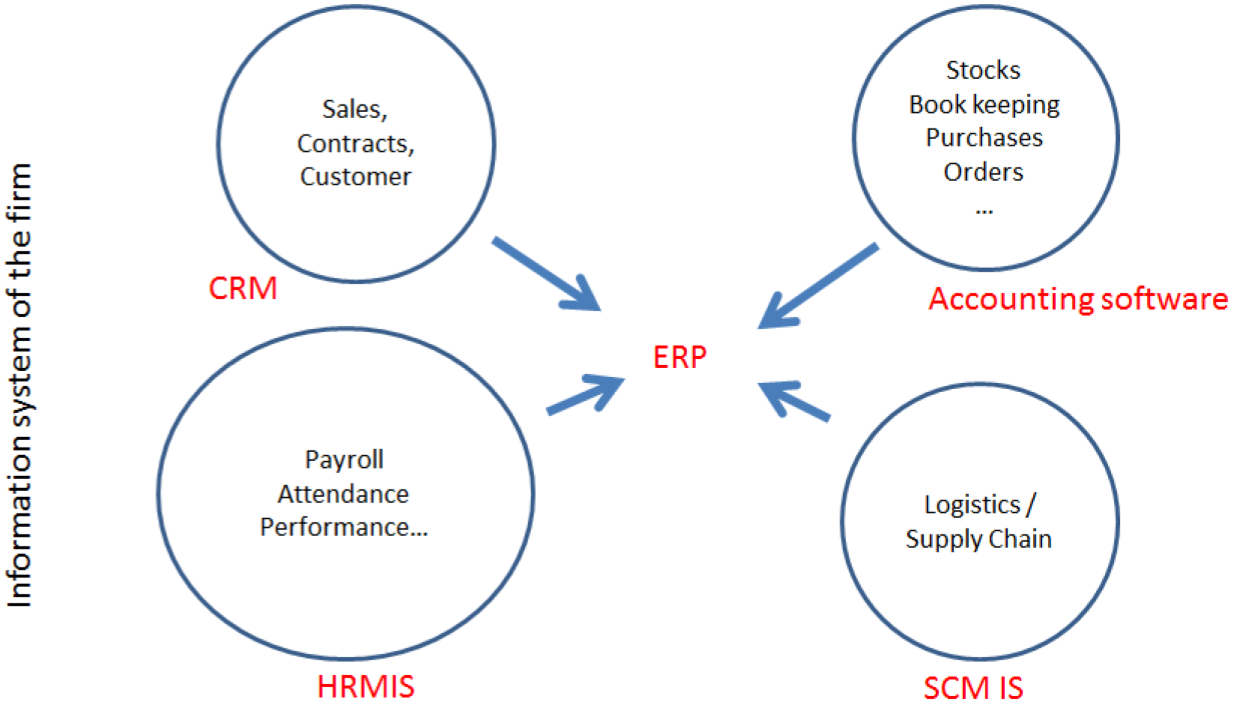


Figure 12. How a CRM integrates in the information system of a firm

Large companies often integrate these different blocks into an **ERP** ("Enterprise Resource Planning"), which is an even larger software able to plug different parts together.

The role of CRMs is evolving, and in this lecture **we make the case that "big data" has transformed CRMs radically**.

To illustrate, we will compare (and caricature a bit) a CRM from 2000 with a CRM of today:

7. CRMs - before

The name of the CRM - Customer **Relationship** Management suggests a kind of rich, personalized and human touch.

In practice, CRMs where used for more practical purposes:

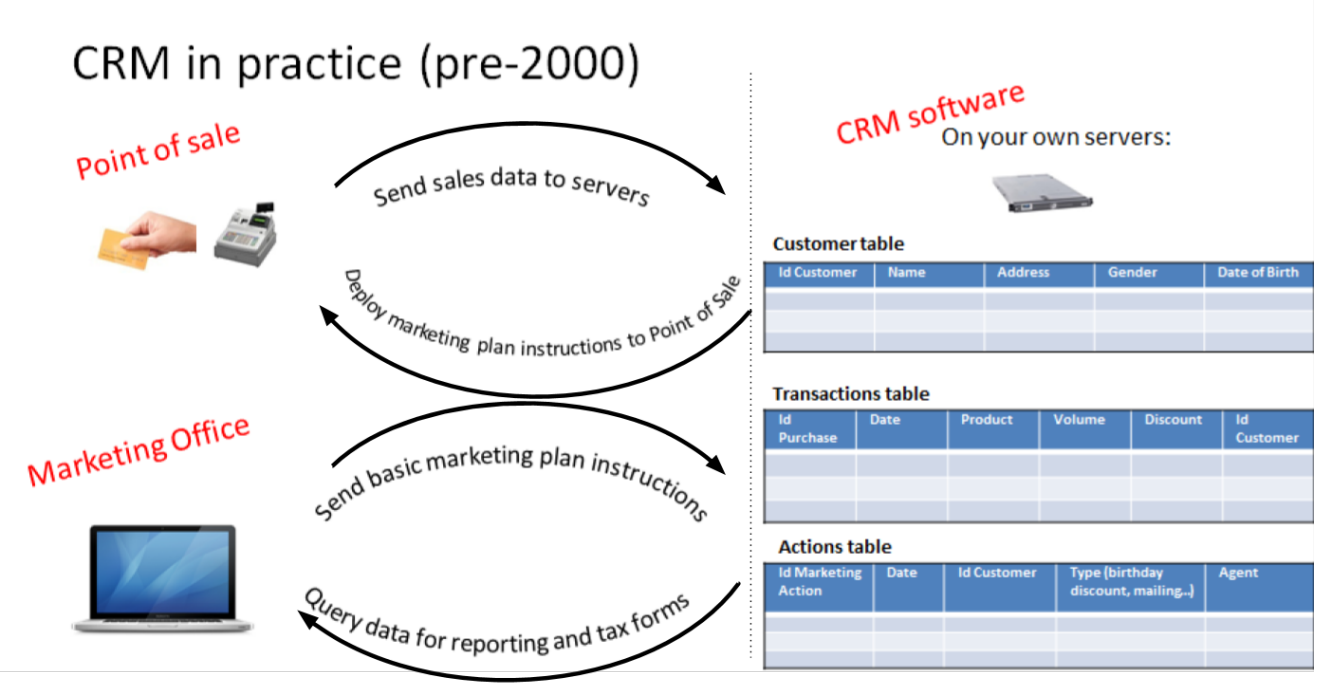


Figure 13. CRMs before the data revolution

We must imagine the CRM software as a tool which **supported the management of sales**, performing these 3 essential functions:

- measuring revenues, through the recording of sales transactions.
- controlling the performance of the sales persons, by registering which cashier, which employee performed the sale, or at least at which location the sale took place.
- recording the VAT ("Value-added tax") collected through sales, which is a legal obligation for tax declaration purposes.

Do you see the customer being catered for in the functions described above? No? Me neither.

The customer was not completely forgotten: CRM are used to run loyalty programs and campaigns:

a) loyalty programs

Loyalty programs afford discounts and special offers to its members.

They increase the share-of-wallet of the company implementing them: the amount of the customer's total spending that a business captures in the products and services that it offers.

A study performed on the loyalty programs run by 7 major supermarket chains in the Netherlands has found that it increased revenues for the supermarket running it:

> On average, a loyalty program enhances the net yearly revenues of a customer by € 163, but the effects vary between € 91 and € 236

source: Leenheer et al. (2007).

Loyalty programs create extra value for the customer as well through the discounts and special offers they bring. But they tend to be limited in their personalization: typically, every customer can enjoy the same offers, even if many of them are irrelevant (discounts on diapers when you don't have a child etc.).

b) Direct mails and coupons

Customers registered in a CRM with their postal address (after joining a loyalty program) can be sent promotional material and coupons.

Using printed material prohibits the customization to the personal needs of the customers, since a printed catalogue is the same for every recipient.

This decreases the efficiency of direct mail campaigns.

8. The digital transformation, 2006-2015

Changes occurring in the past decade have transformed the landscape of the customer relationship. We should realize that:

a) Until 2006 only half of US and EU households, and 10% of the Chinese population, had Internet broadband access at home:

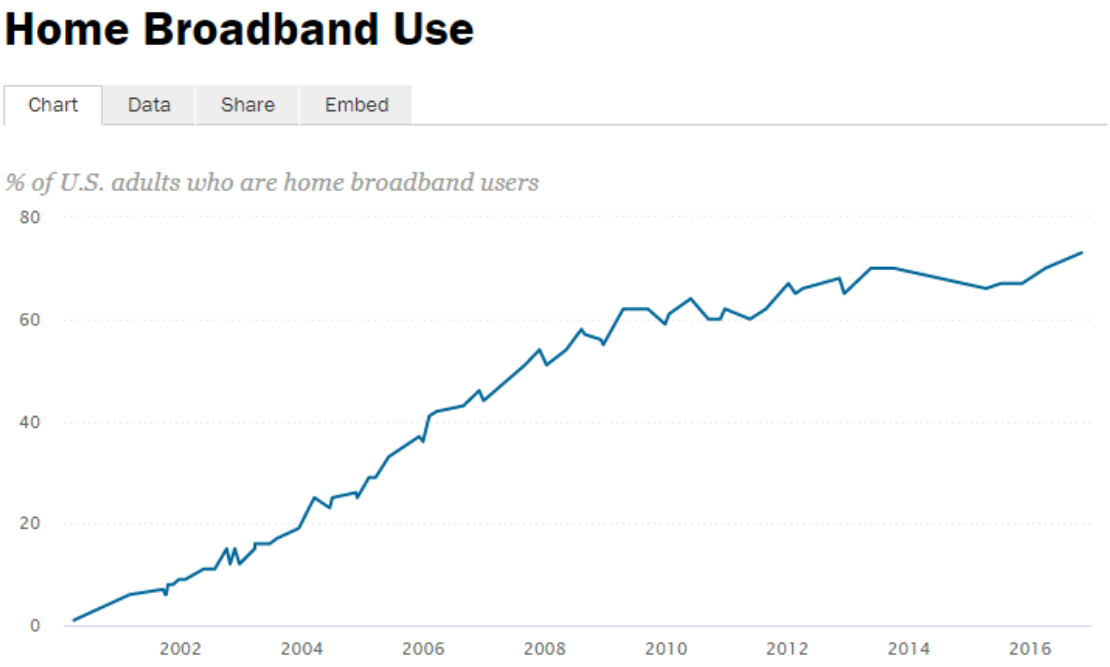


Figure 14. Home broadband use in the US

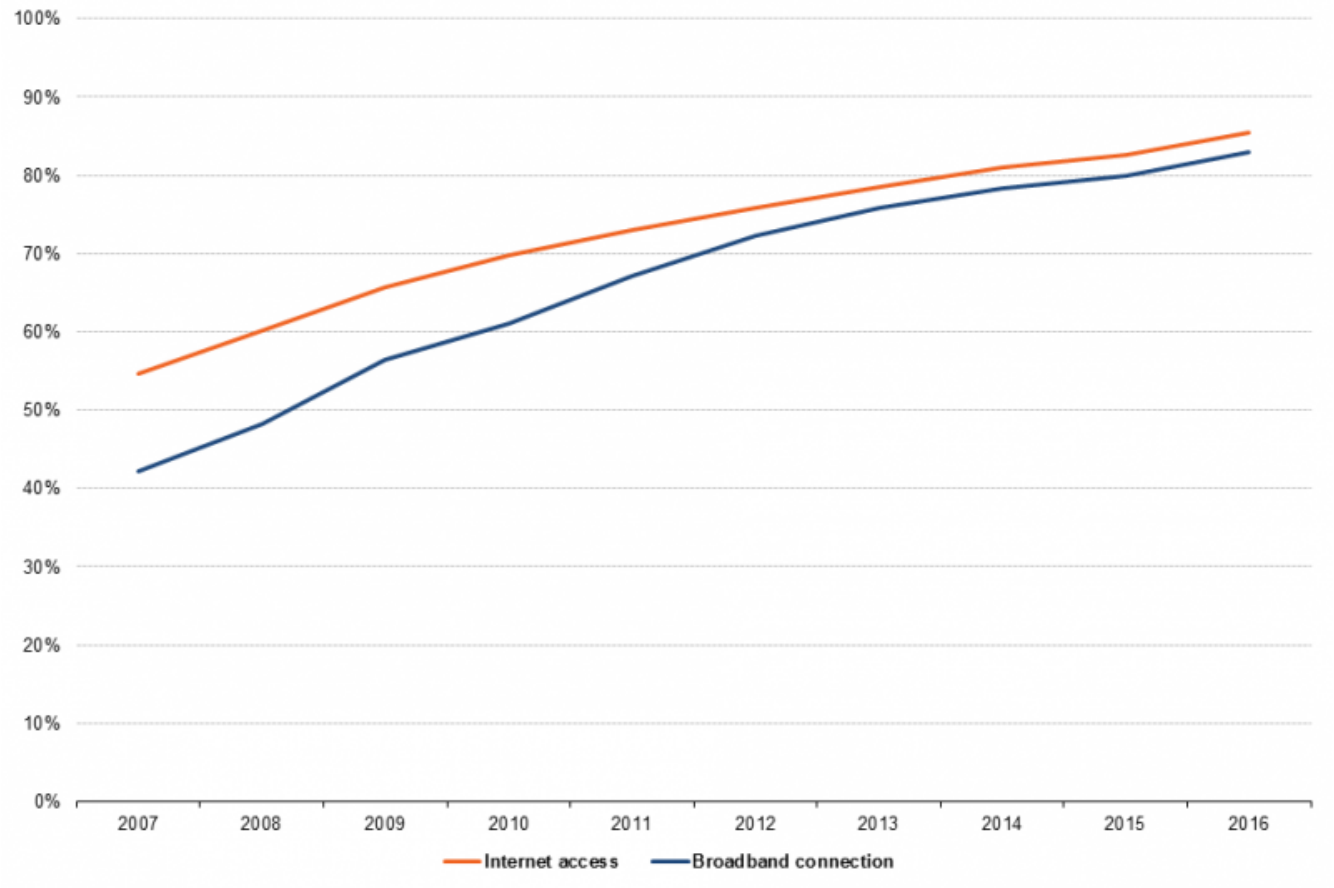


Figure 15. Households with internet access and with broadband connection EU-28, 2007-2016

source: Eurostat

Year	Internet Users**	Penetration (% of Pop)	Total Population	Non-Users (Internetless)	1Y User Change	1Y User Change	Population Change
2016*	**721,434,547**	52.2 %	1,382,323,332	660,888,785	2.2 %	15,520,515	0.46 %
2015*	**705,914,032**	51.3 %	1,376,048,943	670,134,911	4.6 %	30,782,246	0.48 %
2014	**675,131,785**	49.3 %	1,369,435,670	694,303,885	8.2 %	51,100,254	0.51 %
2013	**624,031,531**	45.8 %	1,362,514,260	738,482,729	8.8 %	50,701,259	0.53 %
2012	**573,330,272**	42.3 %	1,355,386,952	782,056,680	11 %	56,979,447	0.53 %
2011	**516,350,825**	38.3 %	1,348,174,478	831,823,653	12.3 %	56,398,548	0.54 %
2010	**459,952,277**	34.3 %	1,340,968,737	881,016,460	19.3 %	74,482,036	0.54 %
2009	**385,470,241**	28.9 %	1,333,807,063	948,336,822	28.6 %	85,638,157	0.54 %
2008	**299,832,084**	22.6 %	1,326,690,636	1,026,858,552	42 %	88,692,052	0.54 %
2007	**211,140,032**	16 %	1,319,625,197	1,108,485,165	52.9 %	73,013,038	0.54 %
2006	**138,126,994**	10.5 %	1,312,600,877	1,174,473,883	24.1 %	26,847,296	0.54 %
2005	**111,279,697**	8.5 %	1,305,600,630	1,194,320,933	17.4 %	16,483,866	0.54 %
2004	**94,795,831**	7.3 %	1,298,573,031	1,203,777,200	18.4 %	14,723,731	0.55 %
2003	**80,072,100**	6.2 %	1,291,485,488	1,211,413,388	35.7 %	21,047,175	0.56 %
2002	**59,024,926**	4.6 %	1,284,349,938	1,225,325,012	75.1 %	25,311,609	0.56 %
2001	**33,713,316**	2.6 %	1,277,188,787	1,243,475,471	49.5 %	11,159,670	0.57 %
2000	**22,553,646**	1.8 %	1,269,974,572	1,247,420,926	152.2 %	13,611,260	0.58 %

Figure 16. China Internet Users, 2000-2016

source: Internetlivestats.com

b) Smartphones as we know them appeared just in 2007

Figure 17. Steve Jobs presenting the iPhone in 2007

c) Until 2009 social media was just taking off

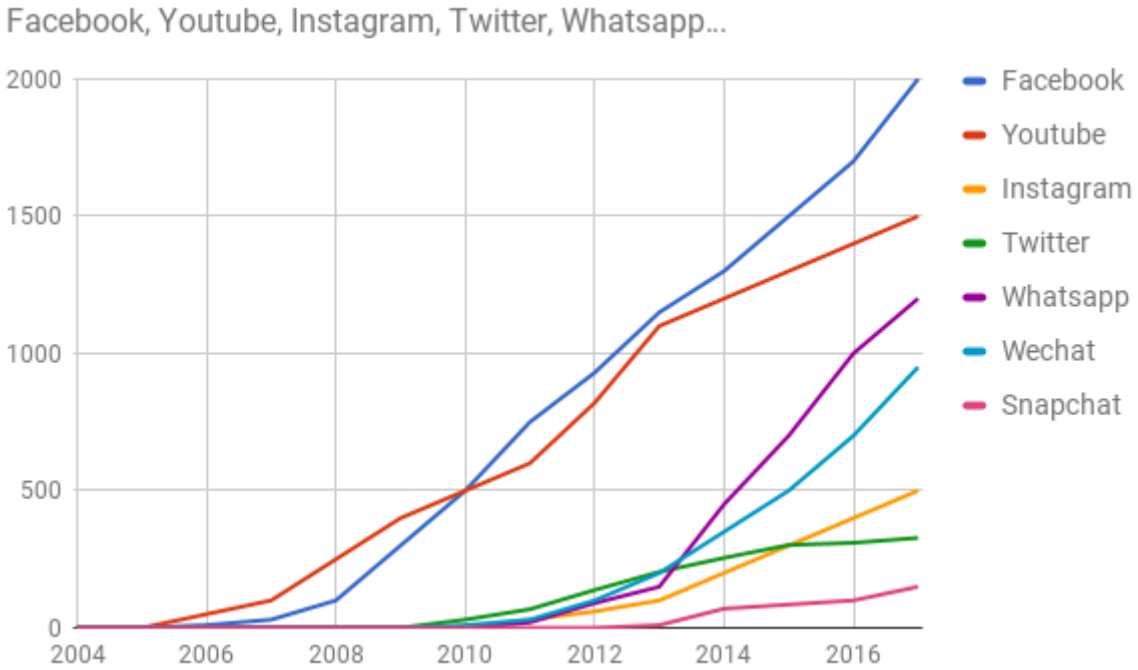


Figure 18. Growth of social media usage, 2004-2017

d) Online retail is growing at a steady pace

Together, Alibaba and Amazon have tripled customers in 5 years, nearing 900 million customers in 2017:

Figure 19. Active consumers on Alibaba, 2012-2017

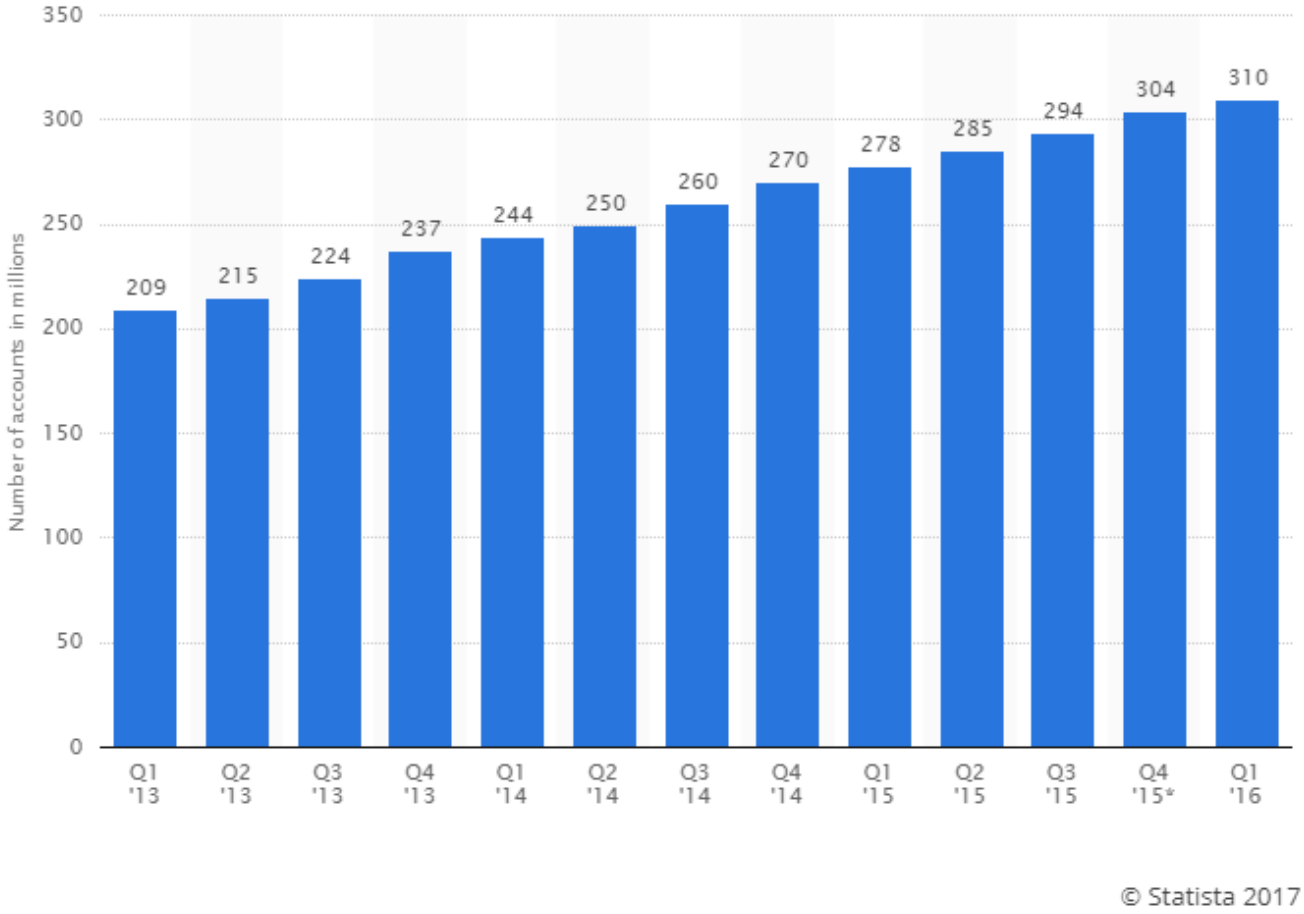


Figure 20. Active consumers on Amazon, 2012-2016

e) The technoloy for ad campaigns has transformed

Three key aspects for ad buying and selling:

- It became programmatic: ad space and ad inventories are bought and sold through automated market places (through SSP, DSP and Ad exchanges).
- Ads are displayed across many channels (with retargeting)
- Ads are personalized (started with Search Engine Advertising showing ads matching search queries, then cookies, then browser fingerprinting (see here) and other techniques)

9. Consequence of this digital transformation: the customer relationship and CRMs have evolved

a) CRMs must handle multiple channels (distribution and communication)

Distribution and communication channels have multiplied and fragmented, and each have their different rules for content generation, data streams and communication modes.

Distribution channels:

- retail stores (as usual)
- ecommerce websites (since 2000s) and mobile apps (since 2010s)
- third party platforms (such as Amazon and Alibaba, taking off since 2010s)
- resellers becoming primary sellers (eg, leboncoin.fr or marktplaats.nl selling cars, housing and jobs) - since 2010s.

Multiplication of distribution channels

→ it becomes increasingly hard to record customers actions (is this customer in my shop the same that clicked on this web page 2 minutes ago?): "click and collect" for example, one example of the broader trend called "https://www.seo.com/blog/phygital-marketing-where-the-physical-and-digital-worlds-converge/[phygital marketing]".

Note how traditional CRMs are unequipped to command and control this variety of distribution channels.

Communication channels:

From brick and mortar + call centers + sms + emails to ...

→ Live chat in websites + Facebook + Twitter + Instagram

b) CRMs must handle complex communication patterns, not just "push campaigns"

Communication used to be mainly "outbound" (company pushing campaigns to customers) and occasionally inbound (customers calling or emailing back).

Three evolutions:

- customers expect their point of view to be heard, without being prompted for it.
- cross customer conversation has spread (without the intervention of companies and brands)
- The high cost of pushing content through ads incentivizes firms to develop inbound communication - this is "inbound marketing".

c) CRMs must accomodate multiple, fragmented touchpoints

- TV, radio, outdoor advertising, in store and outdoor displays: it continues
- mobile phones: operating systems with constantly evolving techs and rules of play (1, 2)
- desktops, tablets, social TVs, but also... watches? cars? homes?

d) CRMs must handle personalized content

- The expectations of customers have elevated: if your company has a Facebook page, it should not just display a catalogue. It should engage (converse) with customers.

- Same with all steps of the customer journey: a CRM should adapt the product (or service) to the profile of the customer.

Several remarks on personalization:

a. "personalization" is the extreme end: one different view for each different customer or prospect.

Micro-segmentation is the step just before: identifying very precise, tiny segments in the population of customers and prospects.

ii. "personalization" has been blamed for reinforcing "bubbles" or "tribes" views of the world (paying version of the paper, free version here).

Content personalization is also blamed for favoring political polarization via an "echo chamber effect": social media tend to show me content I already agree with (paying version of the paper here, free version here).

iii. Personalizing the customer relationship, even when effective, is not inherently a good thing. It has been shown that the Coca-Cola #ShareaCoke campaign is effective at making more children choose a soda with a label to their name, over a healthy drink (paying version of the study here, free version not available).

iv. Personalization through smart CRMs? Companies rated with the best customer service do personalization differently: with humans.

See how Zappos offers a great service to their customers:

▶ https://www.youtube.com/watch?v=vApoQPISmvs *(YouTube video)*

(another impactful version here)

or see (in French) how Trainline makes its customers happy.

10. Todays's CRMs must be data-driven

Explaining the expression "data-driven CRMs":

→ CRMs must turn from a system "supporting the firm's administration needs" to a a system tuned to "plug, host, analyze and push actions from multiple data sources".

To get such a CRM to run in an organization, the right resources must be gathered:

a. Adequate software:
    - the CRM itself - recent enough that it can plug and play with a DMP and a large variety of data sources.
    - a Data Management Platform (**DMP**) as well. The DMP is the software specializing in receiving data streams from a variety of sources and in a variety of formats, and reconciling them (see the lecture on Data Integration).

- a Data Lake to store and query data.
- software bricks for additional analysis, as needed. For example, Dataiku's DSS platform.

b. Adequate human resources:
   - product managers with a tech culture (you), able to design and deploy a marketing strategy in a data intensive environment.
   - data scientists who will implement the strategy.
   - IT engineers to run the pumblery of the software.

c. Adequate organizational culture:
   - This is probably the hardest part: making the top management, and the rest of the organization pay attention and believe in the possibilities afforded by these new way to manage customer relationships.
   - The organization needs to invest and devote enough operational resources to stop doing "business as usual" and develop a data-driven CRM.

11. The role of segmentation in marketing

a. The need for a market fit

How to market the right product to the right people?

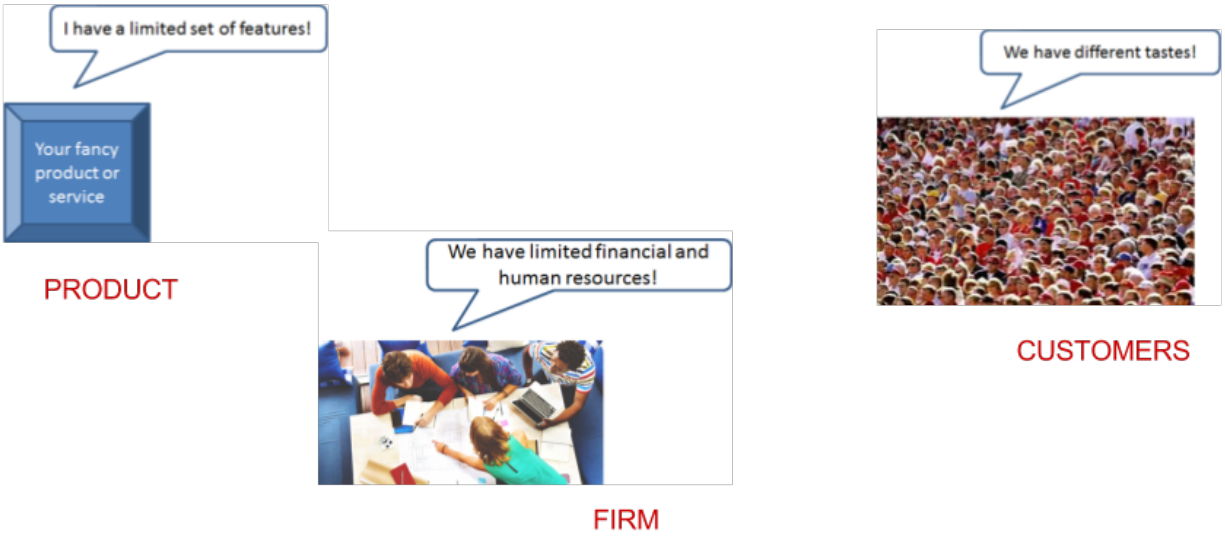


Figure 21. A product cannot meet everybody's expectations

A product cannot have every feature: adding a new feature can conflict with existing features or just hurt the need for simplicity.

Customers have diverging expectations: what is preferred by one customer is considered a nuisance by another.

Firms have limited resources: they cannot create and sell every variety of a product with every feature. They must allocate their scarce resources in the best way possible.

A **market fit** is achieved when there is an alignment between the product, the customers needs and the firms capabilities to deliver.

How to achieve a market fit?

This question is a classic field of study in marketing, and is called "market research". A market fit can be explored and found with the "STP" approach:

b. Segmentation and STP

STP stands for: **Segmentation → Targeting → Positioning**

This a strategy to arrive at a market fit.

Once defined, this strategy will be implemented following a **marketing plan** (for example, following the famous "4Ps").

Let's have a closer look at the "STP":

- SEGMENTATION

First, cut the crowd into segments of customers with similar characteristics / expectations / needs

- TARGETING

Then, evaluate each segment, and select the most attractive one.

- POSITIONING

Creating an offer (a value proposition) corresponding to the segment we target

"Segmenting a market" is the first step of this STP strategy.

It is the key operation where the "anonymous crowd of potential buyers" is analyzed and cut into distinct groups which can be interested in the same kind of product or service.

12. How to segment, in practice?

a. Quantitative vs qualitative methods

Qualitative and quantitative methods can be used for segmentation.

Qualitative methods include surveys, ethnographic observation (online or offline), literature reviews on socio cultural trends, text analysis, interviews, focus groups, and more.

The point of these qualitative methods is to identify typical **usages** and **attitudes** towards the product (aka, **use cases**) among potential customers. Each of these use cases is a potential segment to address.

Qualitative methods are strong at **identifying** segments, and also strong at **understanding** and **explaining** them: we understand better **why** people make their choices.

Qualitative methods are not strong at evaluating the **size** and at **generalizing** beyond the circumstances of the observations made.

In comparison, quantitative methods are relatively stronger at **measuring**, **comparing** and **generalizing**.

Quantitative methods put numbers on things, allowing for an estimation of absolute and relative mangitudes: is this segment a large one? The largest?

And contrary to qualitative methods, quantitative methods are good at projecting a small sample to a larger situation, all while measuring by how much we can be off - that's what statistics do.

But let's not forget about 2 relative weaknesses of quantitative studies compared to qualitative ones:

- correlations are easy to find, **causality** is harder to prove.
- **explanations** (causality + motives → the why?? question) are also hard to establish.

→ we detail **some** of these quantitative methods below.

b. Methods for segmentation in data science: "clustering"

(we use "data science methods" here, but that's very close to "machine learning techniques")

Many overviews of ML techniques tend to have a special category for "clustering techniques" (1, 2, 3).

For example at the bottom right of this chart:

Figure 22. data science techniques grouped in families

source: machinelearningmastery.com

Clustering means "finding groups" in the data. This is analogous to "segmenting" in marketing. So the clustering techniques from machine learning can be used for segmentation.

Does it mean that the other methods shown on the chart above are useless for segmentation?

→ **NO**

For example, Principal Component Analysis (PCA) is classified in the chart as a technique for "dimensionality reduction" (we'll explain this term in an other course), even if it has been used for a long time in marketing (and elsewhere) to segment a dataset (see a great example and tutorial here).

c. Two classic clustering methods: k-means and hierarchical clustering

just to remind you that the goal of this course is to make you familiar and knowledgeable about **what it means to do data science in a business context**, not to turn you in data scientists. Knowing the general principles of k-means and hierarchical clustering is useful if you want to work productively as a business person with data scientists.

The general approach for clustering is:

1. Get data

For example, data on car drivers from consumer panel providing info on their demographics, tastes in car size and design, and pricing preferences

2. Develop measures of association

This means creating a measure of “which customer is similar to which” in terms of their features

For example, families with young children will be roughly similar in terms of demographics, needs and budget.

3. Deal with outliers

Removing car drivers that have extreme values? (the one car driver that needs a race car, etc.)

4. Form segments

Use analytical techniques to create groups of car drivers based on their associations. Also called “clusters” or “communities”.

5. Profile segments and interpret results

Groups have now been found automatically. Identify what these groups mean and how they show a path for action.

d. hierarchical clustering

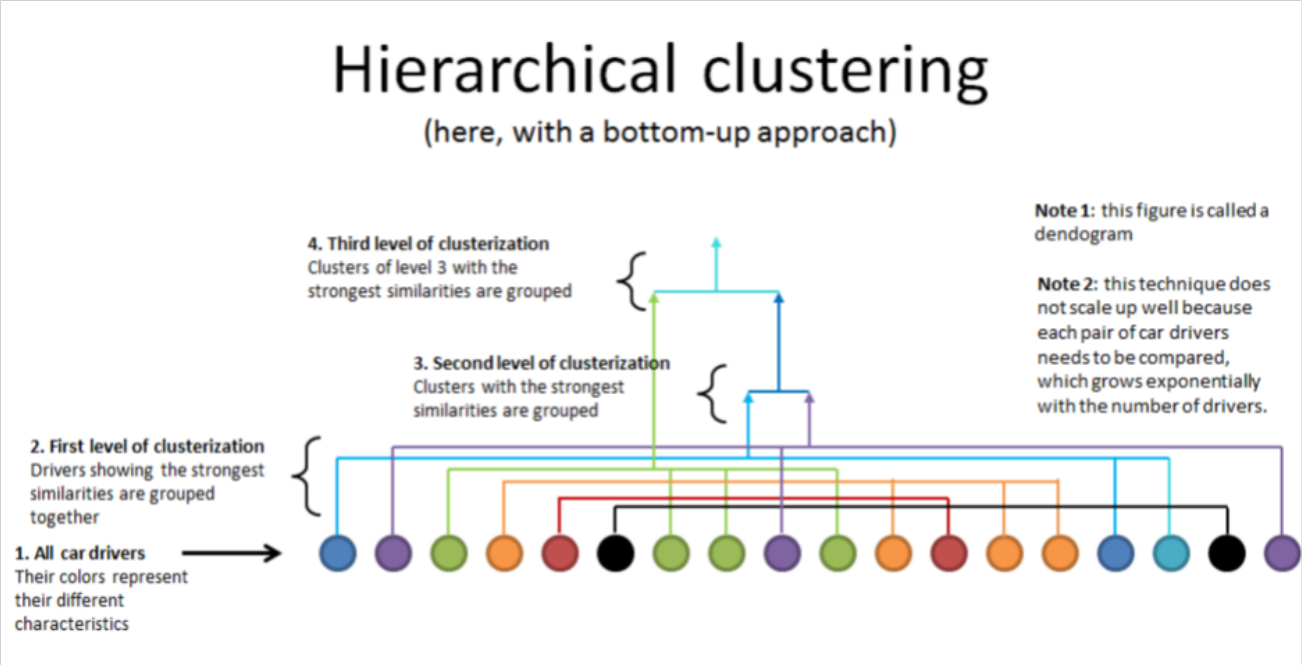


Figure 23. Hierarchical clustering

e. k-means clustering

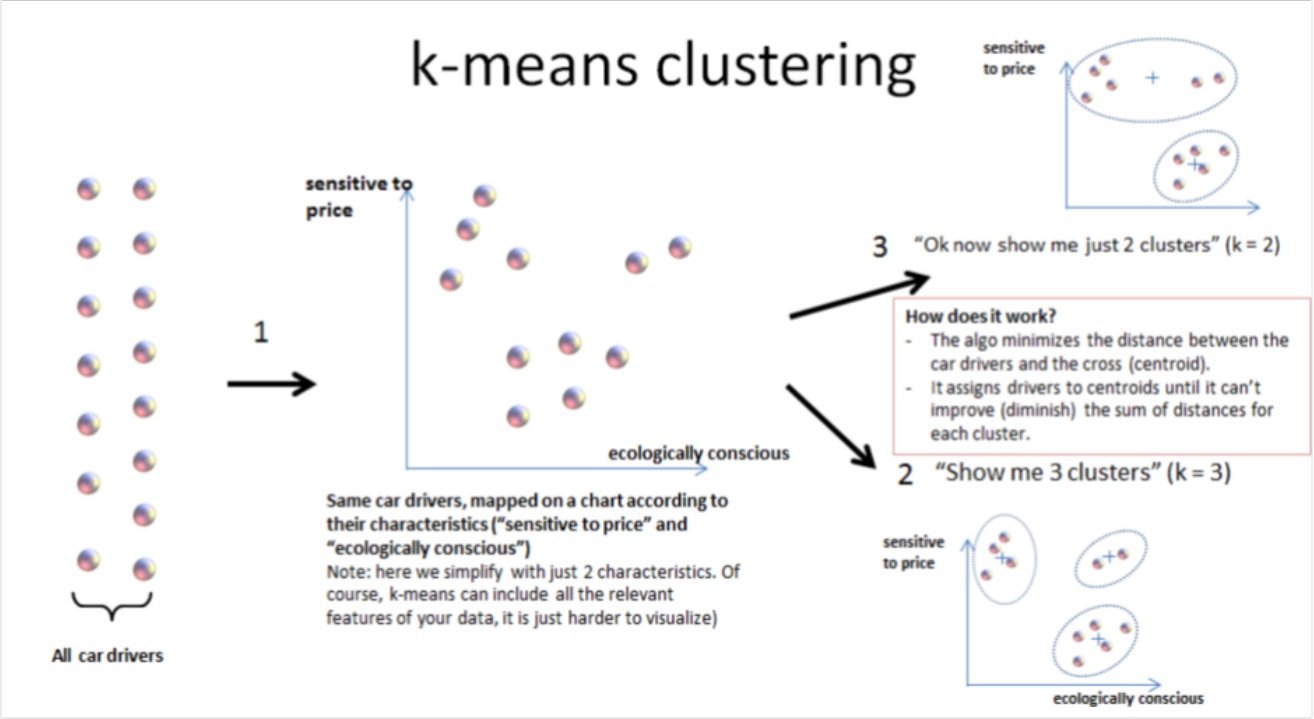


Figure 24. k-means clustering

f. clustering using community detection - via network analysis

This last example of a clustering technique is a bit fancy - not usually represented in ML cheatsheets.

See the lesson on "Network analysis and text mining" for an example of how it can be practised in Gephi.

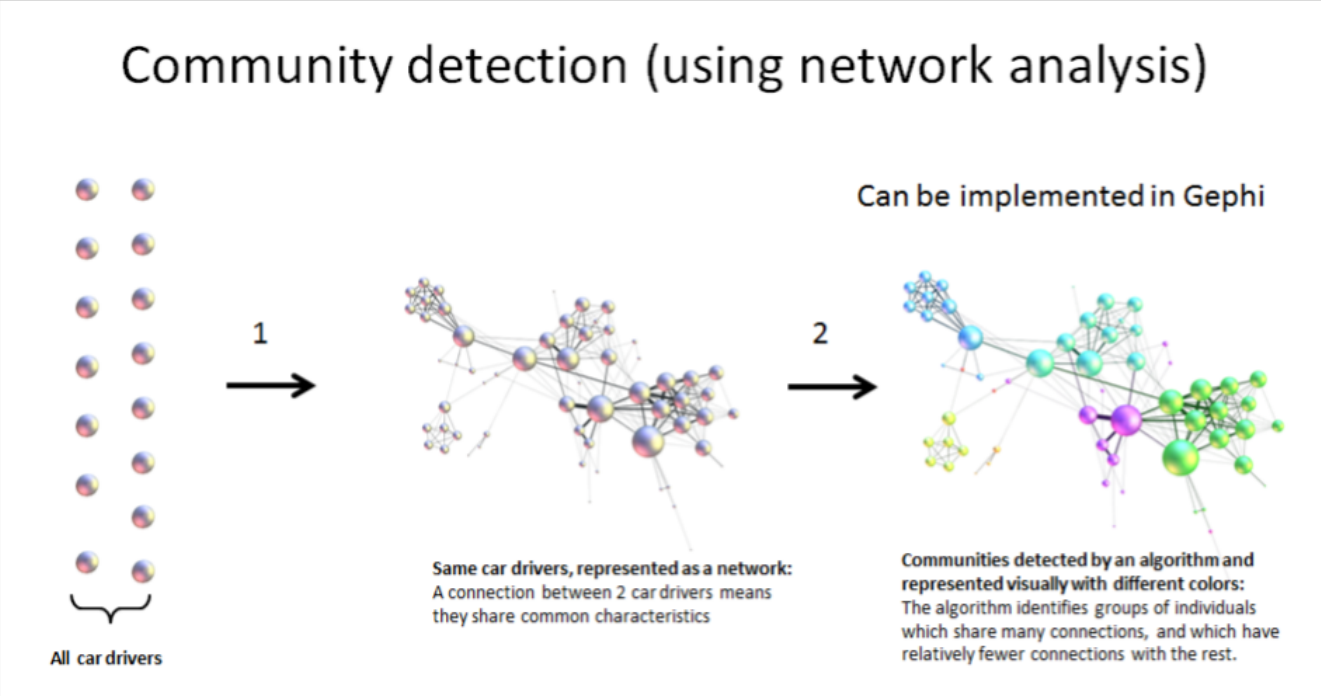


Figure 25. community detection

This clustering example is particularly interesting because the number of clusters found in the dataset is not specified in advance: it "emerges" through the analysis.

(contrary to k-means where the number of clusters is set by the analyst: it is the "k" parameter).

13. Last notes: clustering, useful beyond segmenting in marketing

→ It reveals groups, relations between groups

→ With the network approach, it can even point to the position of single individuals in each group (are they central? Do they bridge to other segments?)

→ Useful for operational marketing (ex: email campaigns), not just strategic product launch!

7 roads to data-driven value creation

Not a closed list, not a recipe!

Rather, these are essential building blocks for a strategy of value creation based on data.

1. PREDICT

Prediction: The ones doing it

1. Predictive churn / default / … (banks / telco)
2. Predicting crime PredPol
3. Predicting deals TILKEE
4. Predictive maintenance CAT

Prediction: the hard part

1. Collecting data (cold start problem)
2. Risk missing the long tail, algorithmic discrimination, stereotyping
3. Neglect of novelty

2. SUGGEST

Suggestion: The ones doing it

1. Amazon's product recommendation system

2. Google's "Related searches..."

3. Retailer's personalized recommendations

Suggestion: the hard part

1. The cold start problem, managing serendipity (see review: paying version, free version not available) and "filter bubble" effects (review: paying version, free version here).

2. Finding the value proposition which goes beyond the simple "you purchased this, you'll like that"

3. CURATE

Curation: The ones doing it

1. Clarivate Analytics curating metadata from scientific publishing

2. Nielsen and IRI curating and selling retail data

3. ImDB curating and selling movie data

Curation: the hard part

1. Slow progress: curation needs human labor to insure high accuracy, it does not scale the way a computerized process would.

2. Must maintain continuity: missing a single year or month hurts the value of the overall dataset disproportionally.

3. Scaling up / right incentives for the workforce: the workforce doing the curation should be paid fairly, which is not the case yet.

4. Quality control

4. ENRICH

Enrichment: The ones doing it

1. Selling methods and tools to enrich datasets 

2. Selling aggregated indicators 

3. Selling credit scores

Enrichment: the hard part

1. Knowing which cocktail of data is valued by the market

2. Limit replicability

3. Establish legitimacy

5. RANK / MATCH / COMPARE

Ranking / matching / comparing: The ones doing it

1. Search engines ranking results Google

2. Yelp, Tripadvisor, etc… which rank places tripadvisor

3. Any system that needs to filter out best quality entities among a crowd of candidates

Ranking / matching / comparing: the hard part

1. Finding emergent, implicit attributes (imagine: if you rank things based on just one public feature: not interesting nor valuable)

2. Insuring consistency of the ranking (many rankings are less straightforward than they appear)

3. Avoid gaming of the system by the users (for instance, companies try to play Google's ranking of search results at their advantage)

6. SEGMENT / CLASSIFY

Generating: The ones doing it

1. Tools for discovery / exploratory analysis by segmentation

2. Diagnostic tools (spam or not? buy, hold or sell? healthy or not?) medimsight

Segmenting / classifying: the hard part

1. Evaluating the quality of the comparison

2. Dealing with boundary cases

3. Choosing between a pre-determined number of segments (like in the k-means) or letting the number of segments emerge

7. GENERATE / SYNTHETIZE(experimental!)

Generating: The ones doing it

(click on the logos to get to the relevant web page)

1. Intelligent BI <img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/aiden.png" format="png" alt="[aiden" tmp="false">]

2. wit.ai, the chatbot by FB <img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/wit.png" format="png" alt="[wit" tmp="false">]

3. Virtual assistants <img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/cx.jpg" format="jpg" alt="[cx" tmp="false">]

4. Image generation <img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/deepart.png" format="png" alt="[deepart" tmp="false">]

5. Close-to-real-life speech synthesis <img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/google.jpg" format="jpg" alt="[google" tmp="false">]

Generating: the hard part

1. Should not create a failed product / false expectations

2. Both classic (think of ) and frontier science: not sure where it's going

Combos!

Figure 26. Combinations

The end

Find references for this lesson, and other lessons, here.

This course is made by Clement Levallois.

Discover my other courses in data / tech for business: http://www.clementlevallois.net

Or get in touch via Twitter: @seinecle