# CODAPPS

## *Essentials of data for managers*

Clément Levallois

Version 1.0, 30-12-2017

# Table of Contents


= What is data? . . . 1
1. Definition of data . . . 1
2. Examples! . . . 1
3. Three take aways from the examples . . . 1
a. Think about data in a broad sense 1
b. metadata is data, too 2
c. zoom in, zoom out 2
4. Some essential vocabulary to discuss data . . . 2
a. Formats, types, encoding 2
b. Data presented as a table 3
c. Data according to who owns it 4
d. Data: "sociodemo" or "behavior"? 4
5. Finally: data and size . . . 4
= What is big data? . . . 6
Reading list . . . 6
1. Big data is a mess . . . 6
2. The 3 V . . . 6
**V** for Volume 6
**V** for Variety 6
**V** for Velocity 7
A 4th **V** can be added, for Veracity 8
3. What is the minimum size to count as "big data"? It’s all relative . . . 9
1. relative to time 9
2. relative to the industry 9
3. not just about size 9
4. no correlation between size and value 9
5. as an expression, "big data" is evolving 10
4. Where did big data come from? . . . 10
1. Data got generated in bigger volumes because of the digitalization of the economy 10
2. Computers became more powerful 11
3. Storing data became cheaper every year 12
4. The mindset changed as to what "counts" as data 12
5. With open source software, the rate of innovation accelerated 12
6. Hype kicked in 13
6. Big data transforms industries, and has become an industry in itself 14
5. What is the future of big data? . . . 15
1. More data is coming 15
2. Discussions about big data will fuse with AI 15

3. Regulatory frameworks will grow in complexity 16
What is "the cloud"? 17
1. Note on the terminology: what is a server? 17
2. The cloud 19
3. IaaS, PaaS, Saas ... ?? 20
4. Private or public cloud? Hybrid cloud? 20
= The headache of data integration 21
Reading list 21
1. Data: you don't get in on tap 21
2. Sources of fragmentation 24
a. Channels keep diversifying 24
b. Connections between these channels intensify and complexify 24
c. Underlying technologies fragment and keep evolving, across channels 24
d. In the meantime, customers have growing expectations about the quality of service 25
e. Example: A French bank going through the 2010s 25
3. Tools for data integration: DMPs and more 26
a. Data Management Platform (DMP) 26
b. DMP in relation to other information systems 26
= APIs and their business relevance 28
1. Definition of API 28
2. The origin of APIs 28
a. EDI: Electronic Data Interchange 28
b. The emergence of web APIs 29
c. The benefits of a web API compared to an EDI 30
d. REST API? 30
3. Business consequences of APIs 30
a. APIs **opened** software to the world 30
b. APIs **accelerated** software innovation 31
c. APIs **opened** data 31
4. The ecosystem of APIs 31
a. A wealth of APIs 31
b. APIs: a business world of its own 31
their transformation in the age of data 33
1. Definition of CRM 33
2. CRMs - before 34
a) loyalty programs 34
b) Direct mails and coupons 35
3. The digital transformation, 2006-2015 35
a) Until 2006 only half of US and EU households, and 10% of the Chinese population, had Internet broadband access at home: 35
b) Smartphones as we know them appeared just in 2007 37

c) Until 2009 social media was just taking off 37
d) Online retail is growing at a steady pace 38
e) The technoloy for ad campaigns has transformed 39
4. Consequence of this digital transformation: the customer relationship and CRMs have evolved 39
a) CRMs must handle multiple channels (distribution and communication) 39
b) CRMs must handle complex communication patterns, not just "push campaigns" 40
c) CRMs must accomodate multiple, fragmented touchpoints 40
d) CRMs must handle personalized content 40
5. Todays's CRMs must be data-driven . . . . . 41
between marketing and data science . . . . . 42
1. The role of segmentation in marketing . . . . . 42
a. The need for a market fit 42
b. Segmentation and STP 42
2. How to segment, in practice? . . . . . 43
a. Quantitative vs qualitative methods 43
b. Methods for segmentation in data science: "clustering" 44
c. Two classic clustering methods: k-means and hierarchical clustering 44
d. hierarchical clustering 45
e. k-means clustering 46
f. clustering using community detection - via network analysis 46
3. Last notes: clustering, useful beyond segmentation in marketing . . . . . 47
= A primer on network analysis for business . . . . . 48
1. Definitions . . . . . 48
a. Social networks 48
b. Other networks 49
c. How big can networks be? 50
d. How to discuss networks? Some vocabulary 51
2. Networks: what use for business? . . . . . 52
a. Segmentation 52
b. Finding key players 53
c. Understanding how information spreads 54
d. Identifying patterns - for fraud detection, control or intelligence. 54
3. To go further . . . . . 55
= A primer on text mining for business . . . . . 56
1. Definitions . . . . . 56
2. What kind of text? . . . . . 56
3. What can be done with text mining? . . . . . 57
a. Sentiment analysis 57
b. Topic modeling / topic detection 57
c. Semantic disambiguation and Named Entity Recognition 57
d. Automated translation / live translation 57

e. Summarizing 57
4. Ok, amaze me . . . 58
a. Named Entity Recognition 58
b. A demo of sentiment analysis by a research team at Stanford: 58
5. Frontier of text mining: what works, what is hard, what does not work. . . . 58
a. What works: Profiling of individuals on psycho / political / social dimensions 58
b. Printed form (or even pdf) is hard 59
c. Multilingual 59
d. Very informal / colloquial speech 59
e. Detection of irony and sarcasm: progresses but not there yet 60
f. Robust translation 60
g. Reasoning beyond Q&As 60
6. Basic operations in text mining - essential vocabulary . . . 60
a. Tokenization 60
b. Stemming 60
c. Lemmatizing 60
d. Part-of-Speech tagging (aka POS tagging) 61
e. Bag-of-words model 61
f. N-grams 61
7. Types of use of text mining for business . . . 61
a. for market facing activities 61
b. for business management 61
c. for business development 62
= 7 roads to data-driven value creation . . . 63
7 roads to data-driven value creation . . . 63
1. PREDICT . . . 63
Prediction: The ones doing it 63
Prediction: the hard part 64
2. SUGGEST . . . 64
Suggestion: The ones doing it 64
Suggestion: the hard part 65
3. CURATE . . . 65
Curation: The ones doing it 65
Curation: the hard part 66
4. ENRICH . . . 66
Enrichment: The ones doing it 66
Enrichment: the hard part 67
5. RANK / MATCH / COMPARE . . . 67
Ranking / matching / comparing: The ones doing it 67
Ranking / matching / comparing: the hard part 68
6. SEGMENT / CLASSIFY . . . 68

Segmenting / classifying: The ones doing it 68
Segmenting / classifying: the hard part 69
7. GENERATE / SYNTHETIZE(experimental!) 69
Generating: The ones doing it 69
Generating: the hard part 70
Combos! 70
= 3 families of business models centered on data 71
3 families of business models centered on data 71
1. COLLECTING DATA FOR RESALE 71
2. COLLECTING DATA, SELLING TARGETED ADS 72
3. COLLECTING DATA, SELLING PRODUCTS and SERVICES 74
a. business model entirely centered on the product or service enabled by the data 74
b. products and services where the data-enabled featured is a "nice to have", not core to the value proposition 75
= Localization and value creation 77
1. Localization brings interesting new dimensions 77
a. Example - Facebook Local Awareness Feature 77
b. Example - Placemeter 78
c. Example - Data @GrandLyon 78
2. The visual power of maps 79
a. Map: useful metaphors with a political dimension 79
b. Example: how to explore the real estate market in the Netherlands 79
c. Key resources to work with maps 79
3. How to represent “space” in data format? 81
a. The specifity of geospatial data 81
b. Solutions to store and retrieve geospatial data 81
4. Two friends for localization: personalization and real-time 83
5. Ending with a provocation: Challenging the usefulness of location 84
a. Localization is about people and *territories* 84
b. Distributed systems – the end of territories? 85
= Personalization and value creation 86
1. From segmentation to personalization 86
2. Beyond behavior: tracking individual bodies 87
3. The case of Nicholas Felton: constant data monitoring 91
a. The Feltron reports 91
b. Not just Feltron 91
4. Issues, limits 92
a. "personalization" has been blamed for reinforcing "bubbles" or "tribes" views of the world (paying version of the paper, free version here). 92
b. Personalizing the customer relationship, even when effective, is not inherently a good thing 92
c. Does personalization always need technology? 92

a viewpoint on its past, present and future . . . 93
1. Where is data visualization going? . . . 93
2. Before dataviz . . . 93
a. Infoviz 93
b. Infographics 94
c. Business intelligence is still another crowd 96
d. And GIS. 96
e. The scene composed by infovis, infographics, BI and GIS 97
3. The emergence of dataviz . . . 97
a. Individuals possessing an unusually broad mix of skills: 98
b. Twitter based communication around the "#dataviz" hashtag 99
c. A tight knit group across the US and Europe. 99
d. A couple of emblematic projects 100
e. Two aspects where data visualization epitomizes its value: maps and networks. 102
f. If we were looking for 2 defining traits of dataviz 104
4. 2014-2015: The stabilization of dataviz . . . 105
5. 2015 onwards: where is dataviz going? . . . 108
= Essential notions on privacy and data protection . . . 110
1. Privacy: just one aspect of data protection . . . 110
2. When is personal information considered "data"? . . . 110
3. Personal data matters because of privacy . . . 111
4. Evolution of privacy . . . 112
5. Privacy of the consumer and privacy of citizens: the relations between the two . . . 112
6. Conclusion: data protection in business, more than an regulatory obligation . . . 113
= GDPR and data protection globally . . . 115
1. Personal data and privacy: why are the stakes so high? . . . 115
2. Evolution of data protection regulations in the EU . . . 115
a. The Directive on Data Protection by the EU in 1995 115
b. The case of EU citizen data hosted by US-based companies 115
3. Key definitions . . . 116
a. Personal data 117
b. Sensitive data 117
c. Data subject 117
d. Data controller (DC) 117
e. Data processor (DP) 117
4. Four key principles for the rightful processing of personal data . . . 117
a. Prior consent 117
b. Adequacy / legitimate purpose 118
c. Portability 118
d. Safety 118
5. In 2018: the GDPR and what it changes . . . 118

a. Application 118
b. Responsibility 119
c. Penalties 119
d. Consent 119
e. Data breaches 119
f. Data Subjects' Rights 119
g. Privacy by Design 119
h. Data Protection Officer (DPO) Appointment 119
6. Data protection: USA, India, China . . . . . 120
a. U.S.A. 120
b. India 120
c. China 120
= Machine learning, data science and artificial intelligence . . . . . 121
1. Explaining machine learning in simple terms . . . . . 121
a. A comparison with classic statistics 121
b. An illustration: the case of GPUs 123
2. Three families of machine learning . . . . . 124
a. The **unsupervised** learning approach 124
b. The **supervised** learning approach 125
c. The **reinforcement** learning approach 127
d. When is machine learning useful? 131
3. Machine Learning and Data Science . . . . . 131
4. Artificial intelligence . . . . . 133
a. Weak vs Strong AI 133
b. Two videos to understand AI further 133

# = What is data?

# 1. Definition of data

The English term "data" (1654) originates from “datum”, a Latin word for "a given". [1: http://www.etymonline.com/index.php?term=data] "Data" is a single factual, a single entity, a single point of matter.

Using the word "data" to mean "transmittable and storable computer information" was first done in 1946. The expression "data processing" was first used in 1954. [2: http://www.etymonline.com/index.php?term=data]

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

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

# 2. Examples!

| | | |
|---|---|---|
| A date | A color | A grade |
| A relation of friendship | A sound | A hearbeat |
| 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... |

# 3. Three take aways from the examples

### a. Think about data in a broad sense

Data is not just text and figures. 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 (you know individual A, you know individual B, 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 buisness situations and think: "where is the data?"

### b. metadata is data, too

Metadata: this is some data describing some other 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: [3: http://www.newyorker.com/news/news-desk/whats-the-matter-with-metadata]

### 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)

# 4. Some essential vocabulary to discuss data

### a. Formats, types, encoding

- This is a digital **medium** (because it's on screen as opposed to analogic, 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 more structured data-interchange formats (check json or xml).
- The **encoding** of the text is UTF-8. Encoding has to do with the issue: how to represent alphabets and signs from different languages in text? (not even mentioning emojis?). 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 my dataset, it is the way my data is organized. There are many alternative data structures: arrays, sets, dics, maps...
- The tweet is stored as a picture (png file) on my 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. Data presented as a table

### c. Data according to who owns it

- 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. Data: "sociodemo" or "behavior"?

- 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.

# 5. Finally: 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 |

# = What is big data?

## Reading list

Find the reading list for this session on Pinterest: https://fr.pinterest.com/seinecle/what-is-data-what-is-big-data/

# 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

Data generated by a movie-goer:

In a movie theater:

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

On Netflix:

Login to Netflix: age, name, gender, location + preferences for movie genres?
Browsing / purchasing history for movies
Movie title, date and price for the movie
Date and time on movie started / paused / interrupted / finished
Comments / ratings posted
Follow / friends activities
If Netflix account connected to FB: personal info, etc.

*Figure 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.

We have a session devoted to data compliance in this course. You can already start reading the documents for this course:

- Data compliance

# What is "the cloud"?

## 1. Note on the terminology: what is a server?

For this topic, you need to know what a "server" is. A server is simply a computer stripped of everything unessential (screen, mouse, graphic card, sound card, keyboard)…

To illustrate: when Google is calculating what the best results are for your search "what are cheap and delicious restaurants in Lyon", this calculation must be done on a computer, right?

The kind of computer used to do this calculation is **not** this one:

*Figure 12. A desktop computer*

The reason is, we don't need a screen, mouse, desktop, and not even the big box containing the computer itself. They take too much space, consume energy, and there is no need for them.

So when all the unecessary parts are removed, the computer looks like this, and is called a "server":

[server] | *server.jpg*

*Figure 13. A server*

source: https://www.oracle.com/servers/sparc/s7-2/index.html

Take a look at the shape: rectangular and very slim. This makes it easy to stack up servers one on the other. Because for Google and other companies crunching data for their business, a lot of servers are needed, so gaining on space is a real issue.

When many servers are piled up together and put in a big tall box, this is called a **rack** of servers, and look like this:

*Figure 14. A rack of servers*

When all the racks of servers are put in the same room, this is called a "data center" and looks like this:

*Figure 15. A data center*

Look at this video showing a tour of a data center at Google:

▶ https://www.youtube.com/watch?v=XZmGGAbHqa0 *(YouTube video)*

Usually, until 2005 roughly, companies had two options:

- buying their own servers and using them on their premises (at their location).
- paying the services of companies specializing in managing data centers.

Then the "cloud" changed this.

# 2. The cloud

The term "cloud" was made popular by Amazon with their service “Amazon Elastic Compute Cloud” (Amazon EC2) launched in 2006.

This service was new in many ways:

- you can rent servers owned by Amazon, at a distance, when you need them, for a duration that you choose.
- there is an emphasis on ease of use: no need to know the technical details of these servers (how they are plugged, how they are configured…)
- you are just given a login + password and you can start using these servers for your needs.
- it’s "elastic": if you need more servers, or more powerful servers, it’s just possible. No need for signing a new contract or to evaluate whether Amazon has the capacity… it’s dimensioned to be possible.

Let’s compare a situation with or without the cloud:

| **Without the cloud** | **With the cloud** |
|---|---|
| You make a market study for which server to buy<br><br>Get the approval by your finance department to buy it (that’s a fixed asset!)<br><br>Wait for the server to ship<br><br>Install it and configure it<br><br>Maintain it (security, etc.)<br><br>When the job is over: what do you do with your server? That’s a sunk cost.<br><br>If the job happens to need more computing capacity than your server offers: you are stuck with your too-small-server! | On Amazon’ EC2 website, you click to choose a server among those on offer: it is **on demand**<br><br>You run your job on it. Costs are metered precisely.<br><br>When your job is over, you stop the server with a click and pay the bill.<br><br>If the job happens to need more computing capacity, you switch to a bigger server with a single click, or it can be done for you automatically: it is **elastic**. |
| It is a capex | It is an opex |

The cloud can fit in the budget as an operational expense instead of a capital expenditure. Opex are not inherently a better or cheaper option than a capex, but they are easier for a project team or

business unit to fit in their budget. See this blog post, especially the critical comments below the post, to continue this discussion.

# 3. IaaS, PaaS, Saas ... ??

What a company can do with the cloud?

They can use it to run elementary operations, up to more complex ones:

**Infrastructure as a service** (IaaS)

You use servers in the cloud for basic capabilities like storing data, or computing operations.

**Platform as a Service** (Paas)

The cloud is used to provide building blocks of a service: to manage a messenging system, to host apps, ...

**Software as a Service** (Saas)

The cloud is used to host a full software accessible "on demand" through the browser: like Google Drive, Brightspace or SalesForce.

# 4. Private or public cloud? Hybrid cloud?

- Amazon EC2 is an example of a **public cloud**: it is publicly accessible to any customer. Of course, this does not mean that every customer can see what the others are doing on the cloud! Each customer have their private spaces on the cloud.
- Many companies have security requirements which prevent them from accessing public clouds. They need to have their servers on premises. In this case, they can build their own **private cloud**: it is a cloud just like Amazon EC2, except that it is owned, managed and used by the company exclusively - it is not accessible to third parties. But even private, it keeps the basic characteristics of a cloud: on-demand and elastic in particular.
- **Hybrid clouds** are a variety of private clouds: it is a private cloud where some forms of operations can be delegated to a public cloud. For example, operations which are not security sensitive and which need a capacity of computing in excess of what the private cloud of the company can provide.

# = The headache of data integration

## Reading list

Find the reading list for this session on Pinterest: https://fr.pinterest.com/seinecle/what-is-the-cloud/

## 1. Data: you don’t get in on tap

A naive vision of data would get that it’s all fluid. Don’t we talk about "data streams"? Working with data would be as simple as opening a tap and "getting customer data", for instance.

*Figure 16. Data streams, as fluid as water on tap?*

Actually, data is more like a complex patchwork: many different pieces which must be stiched together - and this is hard.

*Figure 17. Data streams make a patchwork*

Take customer data.

It is not a given. Instead, this is a design made of multiple primary data sources:

| | | |
|---|---|---|
| Personal | Preferences | Permissions, exclusions, interests |
| | Demographics | Volunteered & inferred lifestyle attributes |
| Marketing Treatments | Promo History | Offers by type, product, channel, date & campaign |
| | Response History | Responses by offer, channel, date |
| Behavior | Interactions | Summary of sales & service contacts, dispositions |
| | Transactions | Purchases, reservations, cancellations, changes |
| Predictions | Planned Promos | Candidate & optimized offers from recent analytics |
| | Model Scores | Propensity & value estimates |

Source: UNICA Corporation, in *Multichannel Marketing*, by A. Arikan (2008).

*Figure 18. Multiple sources of customer data*

> Analysts often spend 50-80% of their time preparing and transforming data sets before they begin more formal analysis work.

Garrett Grolemund, Data Scientist and Master Instructor at RStudio

Take away: Data is fragmented by nature. It comes from different sources and presented in different formats. **You** (the marketer in collaboration with data scientists) wrangle to *construct* customer profiles by joining and assembling different sources of data into a meaningful synthesis.

(if you are interested, data scientists actually have whole books on the subject of wrangling with the mess of different data sources)

*Figure 19. Data wrangling*

# 2. Sources of fragmentation

## a. Channels keep diversifying

Point of sale, print, TV, radio, outdoor posters, mobile apps, mobile sites, emails, SMS, APIs, social networks, search engines, e-commerce platforms, e-commerce websites, blogs, content channels, …

→ all these channels can provide relevant data.

## b. Connections between these channels intensify and complexify

- Social TV is TV delivered with Internet services,
- User profiles created on one platform are imported on another
- Orders taken online can be picked up on a variety of point of sales
- Ads circulating through one channel replicate on other channels, …

→ It is very complex to trace the "customer journey" on all these channels and to keep an updated view of a customer profile.

It is even more difficult to explore causality (which action on which channel caused which subsequent action by the user?)

## c. Underlying technologies fragment and keep evolving, across channels

Browsers, Cookies, APIs, mobile OS (Android or iOS?), etc… All these different techs evolve and need continuous effort and expertise to integrate.

Example: **did you notice** that on a mobile device, the url of the pages you visit can now start with an AMP url?

Like, to visit the page of the New York Times the url should be http://www.nyt.com but it looks like: **http://google.com/amp/www.nyt.com**

This **http://google.com/amp** prefix is a new tech by Google to accelerate the display of web pages on mobiles. Fine. But then, as a marketing data analyst, how to count visits to:

http://google.com/amp/www.nyt.com

and

http://www.nyt.com/etc

→ It is important to count visits to these two urls as a visit **to the same page**.

In Sept 2017 major services of web data analytics were still struggling with this issue.

This illustrates that to just count visits to a web page (something which should be classic and robust) and integrate this data to a larger analysis, big issues can arise and be hard to fix even in 2017, because of the evolution of techs and standards.

## d. In the meantime, customers have growing expectations about the quality of service

Difficulties posed by data integration do not slow or decrease customers expectations. To the contrary, we see an elevation of expectations. Customers increasingly expect:

- realtime contact
- two-ways interaction (they want to be able to voice their opinion, and get a response)
- seamless experience (no glitch, modern UI, consistence of the UX across channels)
- personalized experience (customization of the message they receive)

## e. Example: A French bank going through the 2010s

*Figure 20. Before - a couple of data sources across a few channels*

*Figure 21. Now - many data sources across a variety of channels*

# 3. Tools for data integration: DMPs and more

## a. Data Management Platform (DMP)

In 2015/2016 a new acronym started to trend: "DMP", standing for **"Data Management Platform"**.

Basically a DMP is an information system dedicated to solving the issues of data integration:

- it can store a large amount of data
- it can receive data from a variety of sources, in a variety of formats
- it offers functions to reconcile records from different data sources and generate a unique identifier for each reconciled entry.
- it offers segmentation / classification functions
- it provides security and analytics capabilities on the data
- it makes this data available for execution by other software.

## b. DMP in relation to other information systems

DMPs are relatively new. They integrate with 3 other information systems in the firm:

- CRM (Customer Relationship Management)
  - This is the software **gathering** data related to customers and sales. It is a major source of **input data** for a DMP.
- ERP (Enterprise Resource Planning)
  - Large software synchronizing information systems from finance, sales, logistics and more.

The CRM can be independent or part of the ERP.

- DSP (Demand Side Platform)
  - piece of software automatizing ad buying. So, the audiences identified in the DMP could be served corresponding ads automatically with a DSP.

How can data circulate across these software and with the external world? The next lesson is devoted to APIs, another important concept.

# = APIs and their business relevance

## 1. Definition of API

API: acronym for "Application Programming Interface"

An API is the way to make software programs “easy to plug and share” with other programs.

Ok so… APIs are a cable? A computer? A USB connection? No.

An API is simply a group of rules (you can also call it a convention, or an agreement…) which programmers follow when writing the part of their code which is in charge of communicating with other software.

These rules are then published (on a webpage for example), so that anyone who needs to connect to the program can learn what rules to follow. That’s it.

So, an API is simply a way to write code to make it easy to interface with other programs? Yes.

Why the fuss then?

Having conventions on how to write a software so that somebody can plug it to its own software is one thing. APIs of this sort are a classic topic in computer science but we are not concerned with this here.

APIs we are going to discuss are about communication between distant computers, in a business context.

Let’s do a bit of history:

## 2. The origin of APIs

Companies which need to exchange data is nothing new. Manufacturers, retailers, banks, … they need to exchange information at regular interval.

Sending invoices, receiving receipts for merchandise, and many other administrative records generated in the course of business.

These receipts, invoices… can be printed and mailed (this solution still exists of course).

With informatics developing in the 1970s and 1980s, a new system emerged: the exchange of information via computers: Electronic Data Interchange (EDI)

### a. EDI: Electronic Data Interchange

EDI is not an exchange of file attachments in emails or via a file transfer on a website, because emails and websites did not exist yet! (emails and the Web were adopted by firms in the late 1990s).

Instead, exchanging data via EDI consisted in using complex electronic tools (like the fax but even more complicated) because:

- each industry has its own protocol to exchange data (one protocol for logistics, one for payments, one for this or that retailer chain, etc.)
- you need a dedicated device or software for each EDI protocol, and these are not given for free
- EDI protocols can vary from one country to another
- EDI protocols are controlled by industry associations which do not adopt innovation quickly

and finally, EDI protocols created "closed systems": a company A can connect to company B via an EDI only if the two have a pre-agreement to use this EDI.

So EDIs are fragmented, complicated to implement, slow to evolve, not cheap and restricts the communication to a "club" of partners who agreed to use it.

EDIs still exist, especially in large B2B industries like transportation (check here), but it lost in popularity in the wider economy because… APIs have arrived.

## b. The emergence of web APIs

In the late 1990s and early 2000s, Internet and the World Wide Web expanded a lot.

More and more servers in different parts of the world needed to be interfaced with each other to exchange data.

It became increasingly convenient to define simple and universal conventions that everyone could learn and follow to standardize these exchanges, for free and easily.

That's what **web APIs** do. They are also often called:

- "**API**" for short
- "**web services**"
- "**REST** API" (see below for that last one).

A web API extends the logic of the APIs we have seen in the beginning of this document, to software communicating via the web.

To recall, an API is a convention followed when writing a software, making this software available to other software.

**NOTE**

> Example: the API of Microsoft PowerpPoint enables the import of Excel tables in pptx documents, because the API of Powerpoint plugs to the API of Excel. In this example Excel and Powerpoint are suposed to be installed on the same computer of course!

**Web** APIs are APIs which enable two pieces of software to communicate, via Internet. **They don't need to be installed on the same computer.**

### c. The benefits of a web API compared to an EDI

Unlike EDI, web APIs drop any industry-specific concern. Web APIs are just a convention to send and receive data over the Internet, without any saying on the content of the data.

The data sent and received can be invoices, webpages, train schedules, audio, video… whatever.

Contrary to EDIs, a company creating a web API can choose to leave its access open (remember that EDIs need the two parties to have a pre-established agreement).

So that a potential client interested in using the web API of a company can set it up in a couple of clicks, instead of waiting weeks or months before a contract is signed and the EDI is setup.

**WARNING**

Saying that APIs are open does not mean an absence of security: communication through APIs can easily be identified and encrypted, as needed.

### d. REST API?

Two popular web API conventions emerged in the 1990s and competed for popularity:

- SOAP (Simple Object Access Protocol)
- REST (Representational State Transfer)

REST became ultimately the most widely adopted, because it uses the same simple principles that webpages use to be transferred over the Internet (the "http" protocol that you see in web page addresses). This is why APIs are often called "REST APIs".

In 2000-2010, it became increasingly easy and natural to adopt the REST convention to make one's software and data available to another computer. This simple evolution to ease interoperability had **immense effects**:

# 3. Business consequences of APIs

### a. APIs opened software to the world

An API transforms a closed software into something that can be plugged to anything other computer or object, as long as it is connected to the Internet.

Fo instance, APIs were a key factor of success for SalesForce in the early 2000s. SalesForce, created in 1999, has a revenue of US$8.39 billion in 2017:

- SalesForce developed a CRM as a SaaS where features of the CRM were **exposed as APIs** (meaning, these features could be plugged to external apps via the REST protocol).
- SalesForce created a PaaS to host apps that could plug to the SalesForce CRM via the APIs developed by SalesForce.

This platform is called Force.com and external developers can put their apps there, as long as they are compatible with the SalesForce API.

SalesForce takes a commission on the sales made by these third party apps hosted on Force.com, but more importantly, the platform creates an **ecosystem** of apps and developers around the SalesForce products which makes it hard for a customer company to switch to a different product.

### b. APIs accelerated software innovation

Thanks to API it became easy to add software blocks together and create new apps, even if the app developers where from different countries, industries, or big and small. Check this amazing story.

### c. APIs opened data

Companies and public organization own many datasets of great business interest. The use of these datasets can be free (for small projects and NGOs) or monetized if the user is an entreprise.

Without APIs, datasets can be made publicly available as docs (eg, Excel spreadsheets) to download but this is not practical (try downloading something like `all_train_schedules_2000_to_2017.xls` ! ).

So, imagine a transportation company like French SNCF which finds it interesting to publish station names, train schedules, etc. because it could be used by other companies to build new services : how can it do it?

The data is on a server of SNCF. Then SNCF adds an API and its documentation, making the data available to anyone who knows about REST APIs (and this is trivial).

Entrepreneurs and programmers in general will be able to access the data via the API and use it, possibly to create new services based on this train information.

## 4. The ecosystem of APIs

### a. A wealth of APIs

To discover new APIs, or to make your APIs easier to discover, the most well known place is the website "Programmable Web": https://www.programmableweb.com/

Searching on this website, you will find APIs ranging from the most business-y use case, to APIs of a more fun and odd sort.

Still, many APIs are not listed on this website, and a google search for "info I need + API" is also a good way to find if the API you'd need exists. Interested in whale sightings? There is an API for that.

### b. APIs: a business world of its own

APIs have become central to the economy. As a result, a large number of services associated to APIs have developed to cater for all the needs of companies that use them.

How to create an API, how to manage the documentation of a large number of APIs, how to connect a wide variety of APIs, how to manage the security of APIs, how to monetize and API…

→ Many large firms and startups now specialize in all these different issues. Here is the 2017

landscape of the main companies active in the API industry:

*Figure 22. The API landscape in 2017*

As managers, you have roles to play in the API economy. Engineers develop the technical part of the APIs (the code itself), but you have the expertise to develop the business aspects of this kind of product. In your job search, don’t hesitate to query job postings with "API" in it, you will probably find positions where you’d apply successfully!

# their transformation in the age of data

## 1. 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 23. 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:

# 2. 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 24. 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.

# 3. 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 25. Home broadband use in the US*

*Figure 26. 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 27. China Internet Users, 2000-2016*

source: Internetlivestats.com

## b) Smartphones as we know them appeared just in 2007

*Figure 28. Steve Jobs presenting the iPhone in 2007*

## c) Until 2009 social media was just taking off

*Figure 29. 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 30. Active consumers on Alibaba, 2012-2017*

*Figure 31. 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)

# 4. 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 " 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.

# 5. 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.
   - 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.

# between marketing and data science

# 1. The role of segmentation in marketing

## a. The need for a market fit

How to market the right product to the right people?

*Figure 32. A product cannot fit 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.

# 2. 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 33. 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 34. Hierarchical clustering*

## e. k-means clustering

*Figure 35. 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 36. 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).

# 3. Last notes: clustering, useful beyond segmentation 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!

# = A primer on network analysis for business

## 1. Definitions

A network is a dataset made of entities and their relations

Scientists use the term "graph" to discuss networks.

*Figure 37. This is a network*

### a. Social networks

As users, we are very familiar with one type of networks - social networks:

*Figure 38. source: http://www.minanacheva.com/getting-visual-with-facebook-data/*

## b. Other networks

It is important to realize that networks cover more than relations between humans:

For example, it is possible to imagine a network made out of cooking recipes. 2 ingredients are connected if they appear frequently in the same recipes.

Scanning all recipes and their ingredients from a website of cooking recipes, this gives:

*Figure 39. source: http://arxiv.org/abs/1111.3919*

Semantic networks are another broad category of networks. The method is the same: we need to find a way to "relate" wors in a text, then we get a network.

The general idea is the same as in cooking recipes: 2 terms of a text will be connected in the network if they frequently appeared in same paragraphs.

*Figure 40. source: http://www.nature.com/nature/journal/v463/n7278/full/463157a.html*

## c. How big can networks be?

With a surge in computing power in the age of big data, and the adequate NOSQL databases (such as Neo4J or OrientDB), we can deal with huge networks:

For example, “The Anatomy of the Facebook Social Graph” (2011)

→ study of 721 million active Facebook users and the 69 billion (!) friendship links connecting them.

A limit is quickly reached in terms of visualization: it is hard to fit millions of nodes on a screen. In the next visualization, we can see a network of 90,000 Swedish speakers and their relations on Twitter. The view is very cluttered.

(open the source for an interactive version)

*Figure 41. source: http://twittercensus.se/graph2015/*

## d. How to discuss networks? Some vocabulary

*Figure 42. Terminology*

# 2. Networks: what use for business?

## a. Segmentation

If a network is made of entities and their relations, then a segment is a subgroup of entities in the network, which has some cohesion or something in common.

This subgroup of nodes in the network is often called a "**community**".

Detecting communities in a network, also called "clustering", consists in finding nodes that have many connections in common.

This is a mathematical and algorithmic procedure, but it is very simple to understand visually:

*Figure 43. segmentation with community detection in networks*

## b. Finding key players

*Figure 44. Key players visualized by resizing nodes*

## c. Understanding how information spreads

A data science company created "Where does my tweet go", which traces how a given tweet spreads through retweets.

The service is now discontinued (Twitter data was too expensive to buy) but the mechanism can be explained:

*Figure 45. WDMTG by MFGLabs*

## d. Identifying patterns - for fraud detection, control or intelligence.

In the following video, we see participants in the money market (short term loans between banks) in Europe.

2 banks are connected if one lends to the other. The pattern of exchanges shifts through years - banks withdraw from the market.

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

(the full study is available here: https://www.dnb.nl/en/binaries/Working%20Paper%20418_tcm47-305800.pdf)

Another example: connecting seemingly unrelated measures of business performance with Oracle BI and Linkurious:

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

# 3. To go further

(if viewing from a screen you can click on the covers to get to the Amazon page)

You can also visit my tutorials on Gephi, the leading software to visualize large graphs:

https://seinecle.github.io/gephi-tutorials/

# = A primer on text mining for business

## 1. Definitions

"Text mining" refers to using computational methods to find interesting information in texts.

Quasi synonyms:

- natural language processing (abbreviated in NLP)

→ "Natural language" insists on the fact that we deal with "everyday language" (possibly with slang, grammar and spelling errors, etc.)

- computational linguistics (name of a scientific discipline)

→ this expression shows that of course, text mining is practised by computer scientists but linguists as well. Some approaches to text mining put a heavy emphasis on grammatical notions (structure of sentences etc.), while other methods ignore grammar and syntax completly and still get good results.

A "corpus" is a collection of documents. The plural form exists: corpora. So data scientists doing text mining refer to their "dataset" or "corpus".

## 2. What kind of text?

- Books
- Tweets
- Product reviews on Amazon
- LinkedIn profiles
- The whole Wikipedia
- Free text answers in the results of a survey
- Tenders, contracts, laws, …
- Print and online media
- Archival material

The list above is to convey that "text" is virtually everywhere, and as obvious as it seems we should not forget about this diversity of sources.

# 3. What can be done with text mining?

## a. Sentiment analysis

Sentiment analysis, also called opinion mining or sentiment detection.

Is this piece of text of a positive or negative tone? This type of text mining is useful to filter a large number of documents and retain only those that need action (eg, addressing negative comments).

Companies are typically interested in sentiment analysis in conjuction with topic modeling: they need to identify which **topic** elates negative sentiments.

For a review of methods for sentiment analysis, see this review article.

## b. Topic modeling / topic detection

This type of tool aims at discovering the main topic or the multiple topics of a corpus. This is an essential function for:

- exploratory analysis: what is this corpus about?
- categorization: is this document about this or that?
- detection / monitoring: send an alert when a document matching this topic is published
- etc.

(want to know more about topic modelling? Again, the digital humanities are a good starting point. Check there, there or there.)

## c. Semantic disambiguation and Named Entity Recognition

How can a software "understand" the difference between "Paris Hilton" and "Hilton Hotel, in Paris"?

Named Entity Recognition (NER) consists in matching an expression with the proper definition, given the context of use. Clever techniques and data sources (like DBPedia) are used to achieve good results.

Semantic disambiguation or "Word sense disambiguation" consists in finding the meaning of an expression when multiple meanings are possible.

## d. Automated translation / live translation

See Google translate but also the Skype Translator.

## e. Summarizing

Shortening a text while keeping its core message intact. See this reference for a short, technical review.

# 4. Ok, amaze me

## a. Named Entity Recognition

A demo for Named Entity Recognition (the demo is for a service delivered as an API):

https://dandelion.eu/semantic-text/entity-extraction-demo/

On the website above, try first:

```
Paris Hilton was visiting Paris.
She was interested in books by Jack London, who has written a book about hobboes in
London.
```

Then:

```
London is bigger than Paris
```

## b. A demo of sentiment analysis by a research team at Stanford:

Visit this page and try writing a sentence with a negative or positive emotion / sentiment:

http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

# 5. Frontier of text mining: what works, what is hard, what does not work.

## a. What works: Profiling of individuals on psycho / political / social dimensions

The current state of text mining makes it **easy** to profile individuals, based on the texts they write on social networks.

Without text mining, we have access to “external”, “cold” states of the individual:

- behavior (eg, clicks on websites, purchases, subscriptions)
- sociodemo attributes (address, gender)
- social networks (but relatively cold ones)

With text mining, there is access to “internal”, “hot” cognitive states of individuals:

- opinions
- intentions
- preferences

- degree of consensus
- social networks (who mentions whom: how, in which context)
- implicit and very private attributes of the author (eg, sexual orientation)

See these following studies:

"Predicting the Political Alignment of Twitter Users" by Conover et al. (2011).

"Political Tendency Identification in Twitter using Sentiment Analysis Techniques" by Pla and Hurtado (2014).

"Private traits and attributes are predictable from digital records of human behavior" by Kosinski et al. (2013).

See this article in the New York Times examining the role of Cambridge Analytica in profiling voters at the service of Donal Trump's campaign in 2016:

https://www.nytimes.com/2017/03/06/us/politics/cambridge-analytica.html

These text mining techniques get even more precise when mixed with network analysis and machine learning.

## b. Printed form (or even pdf) is hard

Printed text is typically harder and slower to analyze, because it needs to be scanned first (the technical term is OCR). The process of OCR introduces errors.

Check this paper in the Digital Humanities Quarterly for a deeper look into this issue, in the context of historical research.

And even when the text is in a digital form, it can be hard to use: extracting text from a pdf is not trivial at all, and this is part of the toolchain of data science.

## c. Multilingual

Many operations in text mining will break when the language changes.

For example, the German language capitalizes nouns. It can be confusing to an algorithm trained on a corpus in English where only names are capitalized: simple nouns could be tagged as first names or family names.

This is just one of many examples. Text mining applications often break, are less efficient and / or are more costly when they handle multiple languages.

## d. Very informal / colloquial speech

Text mining applications will have a relatively easy time on text published by Reuters news, because it is written in a formal style.

It will have a harder time on a Facebook message written by a teenager, peppered with slang,

emojis and spelling shortcuts.

### e. Detection of irony and sarcasm: progresses but not there yet

This project tries to crack the challenge of detecting irony in short texts: https://deepmoji.mit.edu/

Not working perfectly. Irony is hard because it needs contextual knowledge to guess that the real meaning is different from the literal meaning.

### f. Robust translation

Translation remains very imperfect. Again, because the meaning of a sentence or paragraph is crafted from the terms used but also from with the contribution of subtle cues (punctuation, phrasing) which are ignored by current algorithms.

### g. Reasoning beyond Q&As

IBM Watson is a software which beat human players at the TV Game "Geopardy" (and that was in *2011*)

▶ https://www.youtube.com/watch?v=WFR3lOm_xhE *(YouTube video)*

Yet, mining text to produce new "reasoning" in general situations by machines has not made much progresses yet.

This topic (reasoning beyond special tasks) is discussed further in the presentation on artificical intelligence.

# 6. Basic operations in text mining - essential vocabulary

### a. Tokenization

Tokenization is finding terms in a sentence. For example, "I am Dutch" is tokenized into "I", "am", "Dutch".

Trivial? Not so much. Try tokenizing a sentence in Mandarin!

### b. Stemming

With stemming, “liked” and “like” will be reduced to their stem “lik” to facilitate further operations

### c. Lemmatizing

With lemmatizing, “liked”, “like” and “likes” will be grouped to count them as one basic semantic unit

### d. Part-of-Speech tagging (aka POS tagging)

POS detects the grammatical function of the terms used in a sentence, to facilitate translation or other tasks.

See for example the online demo by the Stanford team shown above: POS tagging is used to decompose the sentence.

### e. Bag-of-words model

"Starting the text analysis with a bag-of-words model" means just listing and counting all different words in the text, as a first approach.

### f. N-grams

The text "I am Dutch" is made of 3 words: I, am, Dutch.

But it can also be interesting to look at bigrams in the text: "I am", "am Dutch". Or trigrams: "I am Dutch".

N-Grams is the general approach of considering groups of n terms in a document.

This can reveal interesting things about frequent expressions used in the text.

A good example of how useful n-grams can be: visit the Ngram Viewer by Google: https://books.google.com/ngrams

# 7. Types of use of text mining for business

Three types of use:

- for market facing activities
- for business management
- for business development

### a. for market facing activities

- Refined scoring: propensity scores (including churn), scoring of prospects
- Refined individualization of campaigns: personalized ads, email campaigns, coupons, etc.
- Better community management: getting a clear and precise picture of how customers and prospects perceive, talk about, and engage with your brand / product / industry

### b. for business management

- Organizational mapping: getting a view of the organization through text flows.

Example: getting a view on the activity of a business school through a map of its scientific publications.

- HRM: finding talents in niche industries, based on the mining of profiles
- Marketing research: refined segmentation + targeting + positioning, measuring customer satisfaction, perceptual mapping.

## c. for business development

- Developing adjunct services:
  - product recommendation systems (eg, Amazon's)
  - detection and matching of needs (eg, detection of complaints / mood changes)
  - product enhancements (eg, content enrichment through localization/personalization)
- Developing new products entirely, based on
  - different search engines
  - innovative alert systems / automated systems based on smart monitoring of textual input
  - knowledge databases
  - new forms of content curation / high value info creation + delivery

# = 7 roads to data-driven value creation

## 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

3. Predicting deals

4. Predictive maintenance

### 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

2. Yelp, Tripadvisor, etc… which rank places

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

### Segmenting / classifying: 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 with Aiden aiden.ai
2. wit.ai, the chatbot by FB wit.ai
3. Virtual assistants CX company

4. Image generation

5. Close-to-real-life speech synthesis

6. Generating realistic car models from a few parameters by Autodesk:

A video on the generation of car models by Autodesk:

▶ https://www.youtube.com/watch?v=25xQs0Hs1z0 *(YouTube video)*

### 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 46. Combinations*

# = 3 families of business models centered on data

## 3 families of business models centered on data

> Not a closed list, not a recipe!
>
> Rather, a helpful toolkit for brainstorming on data for business

## 1. COLLECTING DATA FOR RESALE

- Thomson Reuters collecting & selling data from finance, scientific research, medicine and news.
- Nielsen collecting & selling market data
- Twitter collecting & selling tweets through its own branch (GNIP) or partners (DataSift)

- Meteo France or The Weather Company (purchased by IBM) collecting & selling meterological data
- Datavenue, a business service by Orange (French Telecom company) collecting & selling info to cities and companies about car traffic and flows of tourists in public places (museums...) based on the (anomymized) geolocalisation of mobile phones.
- ImDB Inc. collecting data through imdb.com & selling data through pro.imdb.com to various stakeholders of the entertainment industry.

# 2. COLLECTING DATA, SELLING TARGETED ADS

This is the business model of Google and Facebook, as well as many free services offered online.

Users of these services offer their data (likes, emails, social connections, comments, navigation history, etc.). This allows the service to sell this qualified audience to ad providers.

These figures from 2015 show that a very large part of Google's revenues come from advertisement:

*Figure 47. Google revenues as of 2015*

Same for Facebook, which makes most of its $7 billion revenues out of ads, based on the data they collect from Facebook users:

*Figure 48. Facebook revenues as of 2016*

# 3. COLLECTING DATA, SELLING PRODUCTS and SERVICES

## a. business model entirely centered on the product or service enabled by the data

- Nokia Health Mate: Devices which make bodily meausrements (sleep, weight, pulse, steps, etc.) to deliver a service: monitoring your health and well being, with suggestions and nudges to improve it.

*Figure 49. Nokia products collecting data to deliver services*

- Nest thermostat: a device which collects data about your home and your domestic habits to regulate your energy consumption and improve your well being at home (modulating temperature according to your preferences, without your explicit intervention).

*Figure 50. Nest thermostats collect data from their surroundings and users to deliver services*

- Autonomous cars. The service provided by these cars (driving safely and comfortably from one place to another, without a driver) is based on the continuous collection and analysis of various data streams.

This is the reason why car makers and tech companies acknowledge that a car is now defined by its software more than by its mechanical parts (not sure about the origin of this quote, but Elon Musk was cited saying something similar about the Tesla Model S in March 2015)

## b. products and services where the data-enabled featured is a "nice to have", not core to the value proposition

- ABN AMRO (Dutch bank) helps customers benchmark their expenses against similar, average households.

- KLM Meet & Sit: connect to Facebook or LinkedIn to choose a sit next to somebody you'd like to meet.

*Figure 51. KLM adds a nice-to-have with a matching service based on customer profiles*

# = Localization and value creation

## 1. Localization brings interesting new dimensions

Localization relates activities to physical space, in at least 4 different ways:

Place: Where is this activity happening?

Distance: Are these two agents neighbors?

Movement: Is this agent travelling?

(together with **speed** and **acceleration**)

Structure: How are these agents and activities **configured** in space?

### a. Example - Facebook Local Awareness Feature

*Figure 52. Facebook Local Awareness Ad Feature*

"Helping Local Businesses Reach More Customers":

- Target ads to people living in a radius around your store.

- Can also target people who have been recently in this radius.

→ https://www.facebook.com/business/learn/facebook-create-ad-reach-ads

▶ https://www.youtube.com/watch?v=-YE90ygswoU *(YouTube video)*

### b. Example - Placemeter

*Figure 53. Placemeter analyzes pedestrian traffic through video*

“Using computer vision to analyze real life activity”:

- Cameras placed in public places (possibly at the windows of private households)
- Video is treated on the device attached to the camera, not saved.
- measures pedestrian traffic in front of stores to provide "main street analytics"

→ https://www.placemeter.com/

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

### c. Example - Data @GrandLyon

<a href="https://data.grandlyon.com/"> <img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/logo-smart-data-grand-lyon.png" format="png" alt="[align"center"</a>" tmp="false">]

An initiative by the city of Lyon

→ Making data open to foster innovation for citizens and businesses

→ Includes many datasets with geographical relevance

Similar initiatives in large cities:

<a href="https://data.amsterdam.nl/">[amsterdam</a>] <a href="https://www.beijingcitylab.com/"><img src="D:/files/Google

Drive/open/mk99/docs/src/main/asciidoc/images/bjcitylab.jpg" format="jpg" alt="[bjcitylab</a>" width="112.5" tmp="false">] <a href="http://www.milanosmartcity.org/joomla/"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/milano.jpg" format="jpg" alt="[milano</a>" width="112.5" tmp="false">] <a href="http://smartcity.jakarta.go.id/"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/jakarta.png" format="png" alt="[jakarta</a>" width="112.5" tmp="false">] <a href="http://smartcityinnovationlab.com/"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/lisboa.png" format="png" alt="[lisboa</a>" width="112.5" tmp="false">]

# 2. The visual power of maps

## a. Map: useful metaphors with a political dimension

- The visual metaphor of the map is widely understood
- Makes exploration easy: all visible at once, while zoom allows for details as well
- Multiple information cues (colors, symbols, shapes, layers, etc.)
- Keep in mind: maps are always **political**
- Watch this extract from the TV series "The West Wing“, Season 2, Episode 16:

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

## b. Example: how to explore the real estate market in the Netherlands

- Every single building of the Netherlands on a map
- Colored by year of construction
- With role (retail or housing?) and surface highlighted
- Zoomable and draggable

<a href="http://code.waag.org/buildings/"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/waag.png" format="png" alt="[align"center"</a>" tmp="false">]

## c. Key resources to work with maps

*Figure 54. Stamen Design*

- Agency based in San Francisco
- Famous for cutting research in map design

*Figure 55. MapBox*

- Mapbox.com
- SaaS to create interactive maps in web pages and mobile apps.

*Figure 56. Openstreetmap*

- OpenStreetMap

- A crowd sourced open source map of the world. Available through API.

# 3. How to represent “space” in data format?

## a. The specifity of geospatial data

Data is traditionally stored in tables in relational databases, taking this form:

| Customer | Address |
|---|---|
| Customer 1 | 67 Pelikaanstraat, Leiden 2314 CR |
| Customer 2 | 12 Breestraat, Rotterdam 3046 DM |

*Figure 57. A table with two entries*

A table can have millions of rows. How to retrieve information such as "get all customers living in Rotterdam"?

"SQL" (Structured Query Language) is a system to express these kinds of queries.

In the table shown above, a query written in SQL look in the "Address" column and inspect all the text to find if "Rotterdam" is present or not.

This is highly inefficient (slow), and more complex queries would not work.

For example, the table above could not be queried for "get all customers living in a 10 miles radius around Rotterdam".

So how to store geospatial data in a way that makes it easy to retrieve?

## b. Solutions to store and retrieve geospatial data

1. SQL solutions

Even if SQL does not perform well on geospatial data "out of the box", extra modules have been developed to deal with it.

Microsoft SQL server since 2008:

- Possible to store and query “geometric” and “geographic” objects
- Possible to use complex queries on these objects

2. NoSQL solutions

Since ~ 2005, new types of databases have been developed, which don’t follow a table structure in order to facilitate the query of special kinds of data, like geospatial data or network data.

These new databases are called "NoSQL databases"

*Figure 58. the Carto Platform*

Carto (ex CartoDB): specializing in geospatial data + mapping.

*Figure 59. Neo4J, a database for networks*

Neo4J Spatial enables to mix the logics of networks with places in the data, so that you can make such queries on your data:

"Select all streets in the Municipality of NYC where at least 2 of my friends are walking right now."

*Figure 60. GeoJSon and TopoJSon are derivations of the json formats for geospatial data*

GeoJSon and TopoJSon: 2 data formats to represent geometric and geographic data developed for Javascript applications – and beyond.

# 4. Two friends for localization: personalization and real-time

Knowing the person, its location, at a precise time unlocks meaningful push notifications

Push notifications are these alerts sent by an app on your mobile, visible as transient icons.

Gets “push marketing” back on solid foundations:

Push marketing actions only to the right person, at the right place, at the right time (and at the right frequency!)

# 5. Ending with a provocation: Challenging the usefulness of location

## a. Localization is about people and *territories*

- Data is a fungible and universal material (just 0s and 1s)
- Geographical coordinates are perfectly universal (just need a longitude and latitude)

and yet …

The logic of territories is shaping data: there is a geography of data.

Cultural, social, political, linguistic, economic dimensions to data.

→ representations with a supposedly universal and transparent coordinate system blinds us to this fact.

This argument is made by Frederic Martel in his book "Smart": Internet does not flatten everything into one big model. There are several Internets with their geography, politics and sociology.

<a href="https://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Daps&field-keywords=smart+frederic+martel&rh=i%3Aaps%2Ck%3Asmart+frederic+martel"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/smart.jpg" format="jpg" alt="[smart</a>" tmp="false">]

- Data protection: not all countries are equal
- Data handling devices

India and Africa have ++ share of mobile devices

- Data production

Amazon Mechanical Turk is a service of data production through the hiring of a distributed crowd of workers. Tends to "erase distance".

Yet, the geographical distribution of workers on Amazon Mechanical Turk is far from even. The following figure is taken from this study:

Figure 1: The number of workers per country. This map was generated based on geolocating the IP address of 4,983 workers in our study. Omitted are 60 workers who were located in more than one country during the study, and 238 workers who could not be geolocated. The size of the circles represents the number of workers from each country. The two largest are India (1,998 workers) and the United States (866). To calibrate the sizes: the Philippines has 142 workers, Egypt has 25, Russia has 10, and Sri Lanka has 4.

*Figure 61. Distribution of Amazon Mechanical Turk workers*

## b. Distributed systems – the end of territories?

The libertarian dream of the cypher-punks: individuals transact without consideration for their nationality, currency, legal system, political regime.

Organizations, banking, voting systems, … any aggregated human activity could emerge without reference to local territories or institutions. Just groups of individuals transacting voluntarily and securely, without a notion of place or distance.

- Bitcoin: the currency for these transactions?
- Torrent: The exchange platform for numeric goods?
- Etherum: the platform where contracts are made and executed?

<a href="https://www.amazon.com/This-Machine-Kills-Secrets-Whistleblowers/dp/0142180491/ref=sr_1_1?ie=UTF8&qid=1508079962&sr=8-1&keywords=this+machine+kills+secrets"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/cypherpunks.png" format="png" alt="[cypherpunks</a>" tmp="false">]

# = Personalization and value creation

## 1. From segmentation to personalization

Segmentation helps refine the picture from a mass of data to meaningful subgroups of data points.

Why not go down to extreme segmentation: segments the size of an individual?

- Major websites do it (Amazon, Yahoo!, Netflix, etc.)
- Ads providers do it (Facebook)
- News feed do it (Prismatic, Pulse)

Advantages: pinpoint accuracy and relevance Inconvenient: operational complexity

*Figure 62. How is an Amazon page (old version!) personalized*

# 2. Beyond behavior: tracking individual bodies

*Figure 63. The relation between connected objects and personalization*

A list of bodily aspects being measured with examples:

*Table 1. Location*

| **Bodily Measurement** | **Device** | **Company** | **Product** | **Picture** |
|---|---|---|---|---|
| Location | Mobile phone | Samsung, Apple, etc. | Phone | |

*Table 2. Movement*

| **Bodily Measurement** | **Device** | **Company** | **Product** | **Picture** |
|---|---|---|---|---|
| Movement | Wrist band | Nike | Fuelband | |

*Table 3. Gestures*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Gestures | Arm band | Thalmic Labs | Myo | |

*Table 4. Weight, heart rate*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Weight, heart rate | Body scale | Nokia | Smart Body Analyzer | |

*Table 5. Sleep*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Sleep | Undermat | Nokia | Aura | |

*Table 6. Fingerprint*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Fingerprint | Mobile Phone | Apple | iPhone 5 | |

*Table 7. Facial recognition*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Facial recognition | Mobile Phone | Apple | iPhone 8 | |

*Table 8. Emotions*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Emotions | Camera | SightCorp | CrowdSight SDK | |

▶ https://www.youtube.com/watch?v=7V8jrdH5tAQ *(YouTube video)*

*Table 9. Behavior in public places*

| Bodily Measurement | Device | Company | Product | Picture |
|---|---|---|---|---|
| Behavior in public areas | Multiple devices | AGT International | Mega Events Management Solution | |
| Pedestrian traffic | Cameras | Placemeter | Placemeter | |

A description of how AGT monitors large audiences in public events (click on the pic for the full document):

*Figure 64. source: https://www.agtinternational.com/wp-content/uploads/2014/10/AGT_AAG_MegaEvent-02Oct2014-2.pdf*

Video showing how Placemeter monitors pedestrian traffic:

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

# 3. The case of Nicholas Felton: constant data monitoring

## a. The Feltron reports

*Figure 65. Nicholas Felton*

Nicholas Felton is a designer and data artist who produced printed annual reports from 2005 to 2014.

Reports on what?

Reports on his bodily data and social life, which he measures *constantly*

▶ https://vimeo.com/145332585 *(Vimeo video)*

## b. Not just Feltron

Insurance companies are interested in boosting individual health, using connected objects as monitoring devices

<a href="http://www.forbes.com/sites/parmyolson/2014/06/19/wearable-tech-health-insurance/"><img src="D:/files/Google Drive/open/mk99/docs/src/main/asciidoc/images/autodesk.jpg" format="jpg" alt="[autodesk</a>" tmp="false">]

Companies are looking to provide a 360 degree solution to health and well being through constant monitoring:

▶ https://www.youtube.com/watch?v=E9jq6XpZjGo *(YouTube video)*

Monitoring on health is also a B2B market to achieve "corporate welfare". See Nokia's brochure on the topic.

# 4. Issues, limits

These technologies open a vast number of issues: from data privacy to the redefinition of well-being, and the grey boundary between monitoring and surveillance.

A full session of this series is devoted to discussing these issues.

For the moment, let us just repeat cautionary remarks already mentioned in a different session:

### a. "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).

### b. 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).

### c. Does personalization always need technology?

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.

# a viewpoint on its past, present and future

## 1. Where is data visualization going?

by Clement Levallois, adapted from a presentation made at DataStorm 2nd edition, Lisbon, July 15, 2015.

The nice thing about fields in computational science is that they evolve so quickly. Every 6 months or so, there is a new kid on the block in machine learning, mobile app development or text mining.

This makes it hard to stay at the forefront of these fields, and sometimes we might even loose perspective as to the meaning and goals of what was the field we formed an interest in, just a couple of years back.

So I take the opportunity of this talk to reflect on data visualization, which is such a young field. I'd like to explore how data visualization has evolved, why there was a need for it to emerge, where it stands today, and I will try to imagine where it will evolve in the coming years.

> I am more an observer than a participant in this field.
>
> My view is the one of somebody who came to data visualization around 2009 through network visualizations with Gephi, getting my information mostly from the discussions, links and podcasts shared by data visualizers which I follow and interact with on Twitter.

I feel that in the last 6 years, dataviz has evolved in significant ways: it emerged and crystalized into a distinct topic, lived happily through a golden age, and we are today somewhere else. Let's start with the beginning.

## 2. Before dataviz

Before dataviz, there was a couple of established fields of practice dealing with graphics and data. I'll mention 4 of them though there are surely important others:

infoviz, infographics, Business intelligence, and GIS.

### a. Infoviz

Infoviz is "information visualisation" and is the name of an academic field concerned with:

> the study of (interactive) visual representations of abstract data to reinforce human cognition.
>
> — Wikipedia entry "Information visualization", https://en.wikipedia.org/wiki/Information_visualization

The strength of this field is that it is a test bench for many assumptions you'd have on how visualizations can be effective, in terms of "reinforcing human cognition":

- How do colors work?
- What about different scales?
- What is the performance of users at solving given problems when reading graph represented as a matrix, versus graphs represented as networks? Should longitudinal data be presented as time slices or animated movies for best understanding?

Research in information visualization is concerned with providing solid, empirical answers to these important questions.

The issue is, this emphasis on hypothesis testing can be detrimental to the advancing on another fundamental goal: creating visualizations that are **engaging to their audience**, widely adopted and used in a world so full of information (or call it noise) that attention by individuals is becoming the scarcest of resources.

As illustrated by the figure above, and if I were to be a bit tough on infoviz, I'd say you need a PhD to get it: the interfaces they design are not engaging enough, not implemented on the platforms that viewers are now using, and the information displayed hardly enhances human cognition for laymen like me.

## b. Infographics

You could say that infographics is a bit the contrary of infovis: communication agencies doing pretty much what they want to catch the attention of their readers, at the expense of truthfulness and reliability of the data they invoke.

The example below shows how colorful and catchy an infographics can be, and yet completely ineffective at conveying information.

Of course there are excellent infographics and Alberto Cairo, a professor and journalist by trade, reminds us in his book The Functional Art that carefully executed infographics are an excellent way to convey complex information in a limited amount of space.

But my understanding is that it is not in the basic contract of infographics to have a one to one relation with data, there is a license to **illustrate** the data. The reader must trust the source of the infographics much more than in information visualisation: depending on whether this is an established newspaper with a good graphics team or a communication agency doing quick and dirty work, infographics can be trusted or badly misleading.

## c. Business intelligence is still another crowd

The mission is basically to do "excel-level" visualizations in terms of reporting and monitoring business data.

Nothing fancy usually there: bar charts, pie charts (often in 3D as in the illustration above, which is wrong), line charts and progress bars assembled in dashboards, sold by companies more versed in the business side of things than graphical design.

## d. And GIS.

Geographical Information Systems (GIS) may have a claim for the longest tradition in visualizing data.

This is after all their business to draw maps, which is geolocalized data.

It could be that this long tradition was also a curse: because they developed these desktop software that were widely used in the 1990s, the 2000s and still today, they were entrenched in technologies that could not be easily adapted when web technologies opened up richer, more engaging ways to draw maps and to project overlays of data on them.

### e. The scene composed by infovis, infographics, BI and GIS

So the scene is the following: scientists in the field of "information visualisation" in their corner being the guardians of the temple of "proper visualisations", but they have a hard time finding an audience for these graphics.

Infographics in the opposite corner, who have access to crowds of readers everyday in the pages of newspapers and marketing brochures, but with a sense that they don't really show the data - they editorialize it a lot, for good or bad.

At one of the two other corners, we have business intelligence which is a bit scorned upon because of the simplicity of their graphics which does not do justice to the richness of the data, but envied because they have access to relevant, pricey, impactful data.

And GIS which works with data in a way which is universally understood and judged relevant (maps), but with a degree of innovation of this field which remains quite low.

# 3. The emergence of dataviz

Something happened around 2008 and 2009, which changed this statu quo.

A number of javascript charting and drawing libraries were released:

- RaphaelJS (08/08/08)
- the Javascript Infovis Toolkit (2009)
- Protovis (2009)
- Processing.js (2010)
- and D3 (2011), by now the most successful framework for dataviz with web technologies.

Together with the take off of mobiles phones without the Flash and Java plugins (remember: the iPhone was released in 2007 and did not support Flash), the decreasing popularity of the Java plugin even on desktop browsers, you see in 3 years a large technological shift: unification of visualization frameworks on the web using javascript.

The web becomes increasingly a platform in itself (more popular than releasing desktop software), with the release of Google Chrome in 2008 - Javascript and CSS become much less broken than when Internet Explorer was dominant.

For what impact?

It shuffled the cards: with Java came a very rigid way to conceive interfaces: windows, menus and even the fonts had a Java look and feel in the browser.

With Flash, you had a strong history of interaction and design skills, but you could use Flash without coding, so that designs made with Flash could remain pretty much disconnected from the datasets they represented.

All that became thrown into the melting pot of Javascript where everybody had to unlearn their framework and learn on a virgin land.

Data visualization was not the natural offspring of one of the 4 fields I mentioned, it emerged outside of them.

It caused many newcomers to try their hands at these new tools, free from the habits and conventions of the 4 fields we have seen.

These newcomers who created dataviz had a different way to look at things, a different tooling, and different ways to function as a group. This community is remarkable in several aspects:

## a. Individuals possessing an unusually broad mix of skills:

Coding skills for the preparation of the data (Python or R for example), skills in javascript and other scripting language for visual design (ActionScript, Processing), a knowledge of the rules of design and a feel for esthetics, and creativity.

That is what you need to create this:

(live url: http://www.mta.me) (by Alexander Chen, a Creative Director at Google Creative Lab)

## b. Twitter based communication around the "#dataviz" hashtag

In this community, people evaluate each other's works, shared their latest realization chat about past and upcoming conferences but more importantly exhchange info about new frameworks and resources.

(live url: http://neoformix.com/2012/DataVisFieldSubGroups.html)

## c. A tight knit group across the US and Europe.

I identify (this is a non exclusive list of course) Santiago Ortiz, Jerome Cukier, Jer Thorp, Gregor Aisch, Jan Willem Tulp, Lynn Cherny, Nathan Yau from Flowing Data, Kim Rees from Periscopic, Moritz Stefaner, with a couple of established academics like Enrico Bertini, Scott Murray, Jon Schwabish, Alberto Cairo, and in relation with teams at the Guardian and the NYT, and Andy Kirk at VisualisingData as an evangelist and instructor.

They were particularly active in spreading news about dataviz and sharing their critical insights which contributed shaping boundaries for the field.

This is a personal and of course biased observation, a systematic investigation reveals a different picture (see above, and below, which is a zoom on the group where I think we would find most people self identifying as dataviz specialists):

(live url: http://neoformix.com/2012/DataVisField1000_Group2.pdf)

## d. A couple of emblematic projects

### i. OECD Better Life Index by Moritz Stefaner et al

Not infovis, not infographics, just dataviz: simplicity, interaction, access to the data.

(live url: http://www.oecdbetterlifeindex.org/)

### ii. The "Ghost Counties" visualization by Jan Willem Tulp

It shows that a marriage is possible between creativity and esthetics on one hand, and cold hard

data on the other hand (foreclosures per county in the US).

(live url, needs Internet Explorer and the Java plugin: http://www.janwillemtulp.com/eyeo/)

### iii. U.S. Gun Deaths by Periscopic

It illustrates the power of tory telling (through the intro), granularity of the data, and impact.

(live url: http://guns.periscopic.com/?year=2013)

The emergence of data visualisation as a set of practice and professionals was coinciding with the surge in the new importance of data as a driver of value for business.

"Data visualization" became positioned as one powerful lever to extract value from datasets: it possesses both the rigor needed to report objectively on key data features, that you'd find otherwise in information visualisation, and the power to be engaging with the domain specialists or the managers in charge of finding insights in the data.

## e. Two aspects where data visualization epitomizes its value: maps and networks.

### i. Maps

Visualization of geolocalized data and of network data has of course a long history before the birth of data visualization: many software integrated mapping functions from Geographical Information Systems, and network analysis packages also had visualization add-ons.

What data visualization brought was impactful visualizations making engagement with data just stronger, more powerful.

Stamen, an agency with strong ties in the data visualization community, does this kind of maps:

(live url: http://prettymaps.stamen.com/201008/#10.00/38.7250/-9.1500)

This interactive map by Stamen is quite different from your usual GIS mapping!

What this kind of map brings is: interaction, custom-made design, and most of all enhanced **engagement** with the viewers.

### ii. Networks

In terms of networks, a pre-dataviz typical network would look like:

Dataviz brought interaction, web-based interactions:

(live url: http://bl.ocks.org/mbostock/1062288)

This type of visualization is different because:

- you can explore the viz, not just stare at it.
- you can share it - just paste the url.
- it can be developed and modified by a large pool of developers because it is written in javascript, which is the common language of web development.
- there is a strong sense of esthetics and natural feeling using it.

→ it will encourage curiosity, exploration, and just increase 10 folds the time spent on it by the viewers.

## f. If we were looking for 2 defining traits of dataviz

### i. Data is for the viewer to see and play with

There is the assumption that the visualization should not provide you with flat and unverifiable conclusions: it should show the data in a transparent, verifiable form.

Of course there is a narrative and an editorialization of how the data is presented, **but** it always remains possible for the viewer to challenge this editorial view because the data is here for anyone to explore and interact with.

This represents a fundamental break with infographics, which can hide the underlying data by design, or show it with strong bias by carelessness and still be "OK" by pre-dataviz standards.

It is also a break with infovis, where data is indeed there but you might not be enticed to engage with it.

### ii. Custom made, creative act

Because we are in the browser there is no click and point solutions for the visualization of the data.

This departs strongly from GIS where "custom" maps could be done by selecting options in a menu, and also a big change from dashboards in business intelligence where you could drag and drop charts to build a visualization.

The sense of esthetics and the particularity of the datasets makes of each dataviz a craftwork.

One of the best examples of a creative and simple design is this one by Hint.fm:

(live url: http://hint.fm/wind/)

(live url for a worldwide version: http://earth.nullschool.net/)

# 4. 2014-2015: The stabilization of dataviz

Anyhow, industrialization in dataviz came in rapidly, with Tableau becoming the leader for general purpose viz, dashboards reinvented themselves in dataviz-style with Bime, Qlik, Palantir to name a few.

Dataviz became integrated into the business discourse on big data: the Harvard Business Review features in 2012 a blog section on data visualization where Jer Thorp contributed to set perspectives straight on data,

Jer Thorp is the co-founder of the Office For Creative Research, a multi-disciplinary research group exploring new modes of engagement with data. He is and is an adjunct Professor in New York University¹s ITP program. From 2010-2012, he was the Data Artist in Residence at *The New York Times*.

MENU

Harvard Business Review

CUSTOMERS

**Big Data Is Not the New Oil**

by Jer Thorp

NOVEMBER 30, 2012

(live url: https://hbr.org/2012/11/data-humans-and-the-new-oil/)

Nielsen, the leader of market data and market research, worked on its corporate identity to include data visualization, with data-driven visuals custom made by Jan Willem Tulp:

Since 2012 or so, General Electric partners with Fathom, the agency founded by Ben Fry (co-creator of Processing!) to build visualizations relative to their corporate identity, with some impressive realizations:

(live url: http://visualization.geblogs.com/visualization/powering/)

And in 2015, you know dataviz has fully stabilized when you see a panel on dataviz with Chelsea Clinton:

(live url: https://www.youtube.com/watch?v=YFrmQDCpgxs - the panel is with Ben Fry).

So until 2012 and 2013 I'd say that we were in the golden age of #dataviz in terms of discoveries and charting new paths: excited comments on new productions by the NYT, debates around the goals of #dataviz: is it a way to tell stories? To open new worlds? To educate?

New connections made with new comers, new agencies, people meeting for the first time in conferences after exchanging on Twitter for years, new positions, big clients…

And in 2015, things seem to have stabilized and normalized.

The energy has changed. The conversation on Twitter has slowed down a lot. The sense of being pioneers has eroded, because time has passed and because we have indeed tried and explored many low hanging fruits.

Many individuals are now engaged in more industrial, long term projects.

So that's not bad news: dataviz is now mainstream and well established, people are less obliged to enter free competitions and work on long personal projects at weekends and nights to get their name out, that's good.

But I miss a bit the excitement of the previous years when you had one framework or one big personal project published per month, and when you had all these big shots chatting on Twitter about the upcoming developments for dataviz.

# 5. 2015 onwards: where is dataviz going?

So… where is dataviz going? As I said, you have this first exciting phase that passed, and we are now in a stage where processes for the creation of dataviz are more industrialized, commodified, stabilized.

This means that innovation will find other places to erupt. Why? Because the landscape of

technologies keeps changing, and creative minds will seize the opportunity to play and explore these opportunities in places where no "client" is yet waiting for them.

To illustrate possible paths, I like to give the example of the career of Seb Lee-Delisle, who defined himself as a creative coder and now as a digital artist.

I follow his work on Twitter since about 2009. He is not at the heart of the "dataviz" network and does not define himself in regards to this label, but you'd find him on Jeff Clark's map of dataviz in 2012 nonetheless (see map above).

- he was using Adobe Flash as one of his main technologies until 2009, contributing to PaperVision3D, a framework to build 3D games and animations in the Flash Player.
- He plays a bit with Adobe Flex in 2009,
- in 2010,Flash is definitely behind so he moves to HTML5 technologies, using and teaching animated graphics in HTML5 + Javascript
- in 2012, he does the lunar trail project: http://seb.ly/work/lunar-trails/
- in 2013, he does pixelpyros: http://pixelpyros.org/
- in in 2014/2015, he launches workshops on "Stuff that talk to the Internets": http://seb.ly/st4i-stuff-that-talks-to-the-interwebs/

This path, and similar paths followed by others, suggest that:

- The computer screen and even the screen of the mobile phone is becoming less hegemonic as the medium where data can be visualized. Objects, sculptures, buildings, furniture… this is the next frontier to be explored. Not just mapping data on a flat surface, but maybe even actual construction of data objects (see this for a nice example by Moritz Stefaner).
- Interaction is richer than we are used to. When we leave the "screen" environment (desktop or mobile), interactions with the user become more diverse. Not just the hand and the click of the mouse, but the whole body. Not one individual facing an object, but possibly a crowd, possibly moving, possibly gesturing.
- And "data" is in the process of getting an even larger meaning. When you move away from the screen and start connecting to a variety of objects and sensors, and with a variety of people, data takes still other forms: real time measurements from the external physical environment, from the internal (body) environment, from local or distant social interactions as they unfold, all while staying connected to the APIs we are already familiar with… the mix can be bring impactful results.

So, if visualizing data from the Twitter API was the cliché of #dataviz in 2010 - 2015, the next cliché could be the instantaneous 3D printing of data generated from the connected objects and bodies in a home or a workspace.

This is just my vision for dataviz, I'd be happy to discuss it with you now!

**Thank you!**

# = Essential notions on privacy and data protection

## 1. Privacy: just one aspect of data protection

*Figure 66. Defining data protection*

## 2. When is personal information considered "data"?

At the most basic level, anything could count as "data" with possibly a personal character to it, including comments written about somebody in a personal notbook.

In practice, "data" starts to to be considered as such when:

This is information capable of being processed **automatically**

→ Hint: data on computers, not unstructured written notes

**Or** infomation intended to be processed automatically

→ Hint: paper records to be fed in a computer (eg, via scanning), not any pile of paper on your desk.

Or **structured information** that can be used to facilitate the retrieval of specific information on specific individuals

→ Hint: paper records, filing systems, databases

# 3. Personal data matters because of privacy

> Personal data are any anonymous data that can be double checked to identify a specific individual (e.g. fingerprints, DNA, or information such as "the son of the doctor living at 11 Belleville St. in Montpellier does not perform well at school").
>
> — CNIL (French Independent Administrative Authority), https://www.cnil.fr/en/personal-data-definition

Personal data is data that an individual has the right to keep private. **How and why is privacy an issue**?

Privacy is mentioned in the Article 12 of the 1948 Universal Declaration of Human Rights:

> No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks.
>
> — Universal Declaration of Human Rights

This article from the Declaration is found in similar forms in most of the conventions on human rights in the world.

This right to privacy enables individuals to define their identity in relation to the world, by giving each individual the power to control what to keep for themselves, and what to reveal / share with the world.

In 2005, a report on Privacy in the Digital Environment by the Haifa Center of Law & Technology develops:

> The right to privacy is our right to keep a domain around us, which includes all those things that are part of us, such as our body, home, property, thoughts, feelings, secrets and identity.

> The right to privacy gives us the ability to choose which parts in this domain can be accessed by others, and to control the extent, manner and timing of the use of those parts we choose to disclose.

In addition to shaping an individual's own personal sphere and identity, privacy also underpins the development of a relation between the individual and the society she is part of:

→ when an individual's privacy is secured, they don't have to fear that their personal opinions and activities (as simple as reading a newspaper) will endanger them as citizens.

→ this gives liberty to individuals to develop political expressions which do not necessarily conform with the power structure in place. This would be much harder if everyone's political opinions could not be kept private.

# 4. Evolution of privacy

Privacy is a social norm which transforms as societies evolve. Since the 2000s, a couple of tendencies can be identified:

- increasing tracking of the digital traces left by individuals by companies which use these traces for ad targeting and data reselling
- increasing state surveillance through digital means, against security threats and unspecified goals.
- broader public acceptance of new forms of violations to privacy.

For example, TV shows where participants are filmed 24/24 and where they reveal their (real or supposed) intimacy, dates back only from the late 1990s.

*Figure 67. The Truman Show, 1998*

# 5. Privacy of the consumer and privacy of citizens: the relations between the two

Thanks to whistleblowers like Edward Snowden, the extent of privacy breaches by governmental agencies is now better known:

▶ https://vimeo.com/108771171 *(Vimeo video)*

Journalists, academics, activists and NGOs such as the Electronic Frontier Foundation make the case that:

- consumers are insufficently aware and sensitive of how much information is captured in the normal conduct of their lives, just by using mobile phones and apps, web browsing, and increasingly in public places.
- citizens are insufficently aware and sensitive of the breach of their privacy by security agencies of their own country of residence, and by other countries.

Many citizens consider that if they don't break the law, then they have "nothing to hide".

Similarly, consumers might find that bargaining their private data against a free service and some targeted ads, is a good deal.

Sociologist of technology Zeynep Tufekci goes further:

Her argument is that besides "surveillance" and "lack of privacy", companies like Google and Facebook developing a business model based on ad targeting by analytics on personal data, design a **persuasion architecture** which can be used / highjacked for political purposes.

Tufeckci does not argue that Google, Facebook or the likes inherently have anti-democratic purposes, but that:

- they develop of an information architecture which has the potential to shape opinions of crowds,
- they do so without transparency
- some past experiments on voting in the US, and current developments on electronic surveillance in China, show that the power of these technologies has already consequences in the real world.

link to the TEDx conference by Zeynep Tufekci

# 6. Conclusion: data protection in business, more than an regulatory obligation

The collection and treatment of personnal data by businesses has far reaching implication, and should not be considered merely from a legal standpoint by firms.

The topic engages the Corporate social responsibility of the firm.

The nature of the business model itself - profiling consumers in the most specific way - has profound consequences on the design of the environment surrounding individuals.

What are the next steps? Several trends can be identified:

1. some voices question the business model: are personalized ads based on personal data as effective as the market valuation of Facebook suggests? How much is just scam? Some voices warn against the extent of the fraud, as the video below shows (see also here, or here):

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

2. legislation by political authorities to protect the public interest, especially via an obligation for transparency, in the face of more personal data being collected, for a larger variety of purposes. See our related lecture on the GDPR.

3. a deepening of the current model with more personal data being collected, in private spaces (homes) and behavior in relation to the public (crowd management in streets, stadiums, etc.):

*Figure 68. Echo Alexa, a home assistant collecting personal data*

# = GDPR and data protection globally

## 1. Personal data and privacy: why are the stakes so high?

Businesses must of course make sure they comply with the existing rules governing the protection of personal data, to avoid reputation damages and litigation.

But why is the protection of personal data such an important matter in the first place?

After all, don't we all give up routinely much of our personal data to companies like Google or Facebook, without consequences?

We discuss the notions of personal data and privacy in this separate document.

## 2. Evolution of data protection regulations in the EU

### a. The Directive on Data Protection by the EU in 1995

This Directive derives from earlier guidelines adopted by the OECD as far back as 1980 on the Protection of Privacy and Transborder Flows of Personal Data.

These guidelines were adopted by OECD members but were non binding: they were not translated into legislation in the US (source), but were translated into a Directive in the EU.

(full text of the Directive, presentation of the Directive on Wikipedia)

This Directive guarantees and facilitates the free movement of personal data across EU States, by providing a framework valid for all member States for the protection of personal data of EU citizens.

How is handled the issue of EU data owned by non EU companies? For example, what is the level of protection for the personal data of a French individual, owned by a US company on a server located in the US?

→ It is forbidden to export personal data to a non EU-country with a lower level of personal data protection.

### b. The case of EU citizen data hosted by US-based companies

The case is important as major providers of services involving personal data (google search, gmail, gmaps, facebook, etc.) is hosted in the US.

The question is: what it the level of data protection in this case? US-level of protection or EU?

It should not be possible to host EU citizen personal data in the US because the US have much less stringent regulations in these matters. In the US:

- There is a regulatory framework on data protection for data collected / held by the Federal government
- But there is no general framework on data protection outside the Federal government (states level).

To remedy this situation, the Safe Harbor principles is an international agreement between the USA and EU which was put in place in 2000.

The Safe Harbor principles are a series of regulations which US companies can agree to follow if they want to host EU personal data. These rules provide a level of data protection equivalent to the one guaranteed by the 1995 Data Protection Directive in the EU.

In October 2015, Maximillian Schrems (a student in law in Austria) launched a lawsuit against Facebook for failure to protect his personal data against the spying of the NSA in the USA.

The defense of Facebook was to argue that Facebook complied with the Safe Harbor Act. The lawsuit went to the European Court of Justice which ended up declaring invalid the Safe Harbor Act because:

"legislation permitting the public authorities to have access on a generalised basis to the content of electronic communications must be regarded as compromising the essence of the fundamental right to respect for private life"."

The following months were a state of legal uncertainty as the EU data hosted on US servers were so under no legal conditions.

On 2nd February 2016, the EU and the US created a new legal agreement known as the EU-US Privacy Shield. It differs from the Safe Harbor Act in the following:

(source for the following bullet points)

1. Stronger obligations on companies in the US to protect the personal data of Europeans' and stronger monitoring and enforcement by the US Department of Commerce and Federal Trade Commission

2. Access to personal data transferred under the new arrangement by public authorities on the US was scheduled to be subject to clear conditions, limitations and oversight, preventing generalised access

3. Effective protection of EU citizens' rights with several redress possibilities

4. An annual joint review mechanism between the EU and the US

# 3. Key definitions

(source: Dataiku's white paper on GDPR)

### a. Personal data

Any information related to a human being (or data subject) that can be used to directly or indirectly identify that person.

For example: name, photos, email addresses, bank details, posts on social networking websites, medical information, IP addresses, etc.

### b. Sensitive data

A special category of personal data (including personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and data concerning health or sex life) to which additional protections apply.

### c. Data subject

A human being on whom personal data is being collected.

### d. Data controller (DC)

An entity that determines the purposes, conditions, and means of the processing of personal data. When the organization is large enough, a dedicated position of "Data Controller" can be created.

Ex: in France, the DC is in charge of declaring the personal data being processed to the CNIL.

### e. Data processor (DP)

An entity that processes personal data on behalf of the controller (e.g., cloud and datacenter providers).

Until 2017, it is considered that the data processor is just "executing" the mission given by the DC:

- the DP is in charge of proper security measures to ensure data protection against breach, loss…
- but the DP is not liable for the improper collection procedures of personal data set up by the data controller.

Starting in 2018 with the GDPR (see next), the DP is co-responsible with the DC in case of a breach of data privacy.

# 4. Four key principles for the rightful processing of personal data

### a. Prior consent

It is required before collecting personal data in view of processing it:

- Data collection policy should be made clearly available to users

- Opt out should be possible
- Consent should be presented clearly

### b. Adequacy / legitimate purpose

The data collected should be exactly necessary to run the service, not more.

Time out: information should be deleted when service stops. In France, there is a 13 month limit after which consent must be renewed

### c. Portability

→ Information should be available on request

In 2011 Max Scherms requested all his Facebook data. He received 1,200 pages of it.

Thanks to his efforts, now most of social media offer a one-click download of your personal data.

Portability also covers the "right to be forgotten", detailed in this factsheet by the EU.

### d. Safety

All reasonable precautions should be taken against data breaches.

Precautions taken should be scaled to the damage which would result from a breach in security

Basics: define and manage access rights to each relevant aspects of the data.

Users should be told about security breaches potentially affecting their data

# 5. In 2018: the GDPR and what it changes

GDPR stands for "General Data Protection Regulation". It was adopted by the EU on April 14, 2016 and is enforced on **May 25, 2018**.

Its key novelties, compared to the EU Data Protection Directive, are:

(source: Dataiku's white paper on GDPR)

### a. Application

The GDPR applies to any company (regardless of their location, size, and sector) processing the personal data of people residing in the EU.

For example, a US-based company processing the personal data within the United States of EU citizens is required to comply.

## b. Responsibility

Under GDPR, both data controllers **and processors** must comply with the legislation. Under the previous/current Data Protection Directive, only data controllers were held liable for data protection compliance, not data processors.

## c. Penalties

With a maximum fine of up to 4 percent of annual global turnover or €20 million (whichever is greater), penalties for non-compliance are steep.

## d. Consent

Under GDPR, companies will no longer be able to use long, illegible terms and conditions full of legalese; consent for collection and use of personal data must be in plain language and detail the purpose of data processing.

## e. Data breaches

Increased regulation surrounding the disclosure of data breaches; specifically, much quicker reporting is required (within 72 hours).

## f. Data Subjects' Rights

EU data subjects will have expanded rights when it comes to data protection, including:

- the right to be forgotten (have their data erased),
- the right to access (obtain information about exactly what data is being processed where and for what purpose),
- and the right to data portability (receive a copy of the personal data concerning them).

Citizens now also have the right to question and fight decisions that affect them that have been made on a purely algorithmic basis.

## g. Privacy by Design

It will be a legal requirement to consider data privacy on the onset of all projects and initiatives, not as an afterthought.

## h. Data Protection Officer (DPO) Appointment

Controllers and processors whose core business is regular and systematic monitoring of data subjects on a large scale or who deal with special categories of data will be required to appoint a DPO. The DPO may be appointed from within, hired, or contracted, but (among other specific requirements) (s)he must be an expert on data protection law and practices.

# 6. Data protection: USA, India, China

## a. U.S.A.

→ Framework on data protection for data collected / held by the Federal government

→ But no general framework on data protection outside the Fed. gov

## b. India

IT Act of 2000 + IT Rules 2011

→ Focus on **sensitive** personal information:

Passwords, financial information, health condition, sexual orientation, biometric information

→ No need to declare data processing activities to an authority

## c. China

Data protection not enacted in a single piece of legislation.

Except for general laws: National People's Congress Standing Committee Decision concerning Strengthening Network Information Protection.

Rather, sector based pieces of legislation, such as the Regulation on Personal Information Protection of Telecom and Internet Users (MIIT Regulation)

To have a synthetic view of data protection laws in other countries, visit this website by Thomson Reuters.

# = Machine learning, data science and artificial intelligence

## 1. Explaining machine learning in simple terms

### a. A comparison with classic statistics

Let’s compare machine learning to something we would call "regular statistics":

A basic method in statistics is to compute a regression line to identify a trend from a scatter plot.

To illustrate, we take some data about marketing budgets and sales figures in the corresponding period:

*Figure 69. A linear regression*

"Regular statistics" enables, among other things:

1. to find the numerical relation between the 2 series, based on a pre-established formal model (eg, ordinary least squares).

→ we see that sales are correlated with marketing spendings. It is likely that more marketing spending causes more sales.

2. to predict, based on this model:

→ by tracing the line further (using the formal model), we can predict the effect of more marketing

spending

"Regular statistics" is advanced by scientists who:

1. are highly skilled in mathematics

→ their goal is to find the exact mathematical expression defining the situation at hand, under rigorous conditions

→ a key approach is **inference**: by defining a **sample of the data** of just the correct size, we can reach conclusions which are valid for the entire dataset.

2. have no training in computer science / software engineering

→ they neglect how hard it can be to to run their models on computers, in terms of calculations to perform.

→ since they focus on **sampling** the data, they are not concerned with handling entire datasets with related IT issues.

Machine learning does similar things to statistics, but in a slightly different way:

- there is an emphasis on getting the prediction right, not caring for identifying the underlying mathematical model
- the prediction needs to be achievable in the time available, with the computing resources available
- the data of interest is in a format / in a volume which is not commonly handled by regular statistics package (eg: images, observations with hundreds of features)

Machine learning is advanced by scientists who are typically:

1. highly skilled in statistics (the "classic" statistics we have seen above)

2. with a training or exeprience in computer science, familiar with working with unstructured / big data

3. working in environments (industry, military, ...) where the operational aspects of the problem are key determinants (unstructured data, limits on computing resources)

Machine learning puts a premium on techniques which are "computationally adequate":

- which need the minimum / the simplest algebric operations to run: the best technique is worthless if it's too long or expensive to compute.
- which can be run in such a way that multiple computers work in parallel (simultaneously) to solve it.

(footnote: so machine learning, in my opinion, shares the spirit of "getting things done" as was operations research in the early days)

The pursuit of improved models in traditional statistics is not immune to the notion of computational efficiency - it does count as a desirable property - but in machine learning this is

largely a pre-requisite.

## b. An illustration: the case of GPUs

A key illustration of the difference between statistics and machine learning can be provided with the use of graphic cards.

Graphic cards are these electronic boards full of chips found inside a computer, which are used for the display of images and videos on computer screens:

*Figure 70. A graphic card sold by NVidia, a leading manufacturer*

In the 1990s, video gaming developed a lot from arcades to desktop computers. Game developers created computer games showing more and more complex scenes and animations. (see an evolution of graphics, and advanced graphics games in 2017).

These video games need powerful video cards (aka GPUs) to render complex scenes in full details - with calculations on light effects and animations **made in real time**.

This pushed for the development of ever more powerful GPUs. Their characteristics is that they can compute simple operations to change pixel colors, **for each of the millions of pixels of the screen in parallel**, so that the next frame of the picture can be rendered in milliseconds.

Millions of simple operations run in parallel for the price of a GPU (a couple of hundreds of dollars), not the price of dozens of computers running in parallel (can be dozens of thousands of dollars)? This is interesting for computations on big data!

If a statistical problem for prediction can be broken down into simple operations which can be run on a GPU, then a large dataset can be analyzed in seconds or minutes on a laptop, instead of cluster of computers.

To illustrate the difference in speed between a mathematical operation run without / with a GPU:

▶ https://www.youtube.com/watch?v=-P28LKWTzrI *(YouTube video)*

The issue is: to use a GPU for calculations, you need to conceptualize the problem at hand as one that can be:

- broken into a very large series
- of very simple operations (basically, sums or multiplications, nothing complex like square roots or polynomials)
- which can run independently from each other.

Machine learning tyically pays attention to this dimension of the problem right from the design phase of models and techniques, where statistics would typically not consider the issue, or only downstream: not at the design phase but at the implementation phase.

Now that we have seen how statistics and machine learning differ in their approach, we still need to understand how does machine learning get good results, if it does not rely on modelling / sampling the data like statistics does?

Machine learning can be categorized in 3 families of tricks:

# 2. Three families of machine learning

## a. The unsupervised learning approach

This designates all the methods which take a fresh dataset and find interesting patterns in it, **without training on previous, similar datasets**.

The analogy is with a person doing a task for the first time:

→ she learns a new thing by applying clever heuristics, without having been training on the task before.

Example: in your wedding, how to sit people with similar interests at the same tables?

The set up:

- a list of 100 guests, and 3 tastes you know they have for each of them
- 10 tables with 10 sits each.

- a measure of similarity between 2 guests: 2 guests have similarity of 0% if they share 0 tastes, 33% if they share 1 taste, 66% with 2 tastes in common, 100% with three matching interests.
- a measure of similarity at the level of a table: the sum of similarities between all pairs of guests at the table (45 pairs possible for a table of 10).

A possible solution using an unsupervised approach:

- on a computer, assign randomly the 100 guests to the 10 tables.
- for each table:
  - measure the degree of similarity of tastes for the table
  - exchange the sit of 1 person at this table, with the sit of a person at a different table.
  - measure again the degree of similarity for the table: if it improves, keep the new sits, if not, revert to before the exchange

And repeat for all tables, many times, until no exchange of sits improves the similarity. When this stage is achieved, we say the model has "**converged**".

*Figure 71. K-means, an unsupervised learning approach*

## b. The supervised learning approach

Take 50,000 or more observations, or data points, like:

**an image of a cat, with the caption "cat"

**an image of a dog, with the caption "dog"

**another image of a cat, with the caption "cat"

etc....

- you need 50,000 observations of this kind, or more! It is called the **training set**
- this is also called a **labelled dataset**, meaning that we have a label describing each of the observation.

The task is: if we give our computer a new image of a cat without a label, will it be able to guess the label "cat"?

The method:

- take a list of random coefficients (in practice, the list is a vector, or a matrix)
- for each of the 50,000 pictures of dogs and cats:
  - apply the coefficients to the picture at hand (let's say we have a dog here)
  - If the result is "dog", do nothing, it works!
  - If the result is "cat", change slightly the coefficients.
  - move to the next picture
- After looping through 50,000 pictures the parameters have hopefully adjusted and fine tuned. This was the **training of the model**.

Now, when you get new pictures (the **fresh set**), applying the trained model should output a correct prediction ("cat" or "dog").

Supervised learning is currently the most popular family of machine learning.

*Figure 72. A hard test case for supervised learning*

It is called **supervised** learning because the learning is very much constrained / supervised by the intensive training performed:

→ there is limited or no "unsupervised discovery" of novelty.

▶ https://www.youtube.com/watch?v=4HCE1P-m1l8 *(YouTube video)*

Important take away on the supervised approach:

- **collecting *large* datasets for training is key**. Without these data, no supervised learning.
- supervised learning is not good at analyzing situations entirely different from what is in the training set.

## c. The reinforcement learning approach

To understand reinforcement learning in an intuitive sense, we can think of how animals can learn quickly by **ignoring** undesirable behavior and rewarding desirable behavior.

This is easy and takes just seconds. The following video shows B.F. Skinner, main figure in psychology in the 1950s-1970s:

▶ https://www.youtube.com/watch?v=TtfQlkGwE2U *(YouTube video)*

Footnote: how does this apply to learning in humans? On the topic of learning and decision making, I warmly recommend this book by Paul Glimcher, professor of neuroscience, psychology and economics at NYU:

(this is a very hard book to read as it covers three disciplines in depth. The biological mechanisms of decision making it describes can be inspiring to design new computanional approaches.)

*Figure 73. Foundations of Neuroeconomics, Paul Glimcher, 2010*

Besides pigeons, reinforcement learning can be applied to any kind of "expert agents".

Take the case of a video game like Super Mario Bros:

*Figure 74. Mario Bros, a popular video game*

Struture of the game / the task:

- Goal of the task: Mario should collect gold coins and complete the game by reaching the far right of the screen.
- Negative outcome to be avoided: Mario getting killed by ennemies or falling in holes.
- Starting point: Mario Bros is standing at the beginning of the game, doing nothing.
- Possible actions: move right, jump, stand & do nothing, shoot ahead.

Reinforcement learning works by:

1. Making Mario do a new random action ("try something"), for example: "move right"
2. The game ends (Mario moved right, gets hit by a ennemy)
3. This result is stored somewhere:
    - move right = good (progress towards the goal of the game)
    - walking close to an ennemy and getting hit by it = bad
4. Game starts over (back to step 1) with a a combination of
    - continue doing actions recorded as positive
    - try something new (jump, shoot?) when close to a situation associated with a negative outcome

After looping from 1. to 4. thousands of times, Mario completes the game, without any human player:

▶ https://www.youtube.com/watch?v=qv6UVOQ0F44 *(YouTube video)*

Reinforcement learning is perceived as corresponding to an important side of human learning / human intelligence (goal oriented, "trial and error").

### d. When is machine learning useful?

Using machine learning can be a waste of resource, when well known statistics could be easily applied.

Hints that "classic" statistical modelling (maybe as simple as a linear regression) should be enough:

- The dataset is not large (below 50k observations), supervised learning is not going to work
- The data is perfectly structured (tabular data)
- The data points have few features

Cases when "classic" statistics modelling is **necessary**:

- The question is about the relative contribution of independent variables to the determination of an outcome

# 3. Machine Learning and Data Science

Machine learning is a step in the longer chain of steps of data science.

The process was formalized as kdd: "Knowledge Discovery in Databases":

*Figure 75. KDD - knowledge discovery in databases*

More recent representations of the steps in data processing have been suggested, making room for the role of data visualization (see the lecture on the topic):

→ see the version by Ben Fry (source) and this one by Moritz Stefaner:

*Figure 76. data visualization workflow by Moritz Stefaner*

(source)

Machine learning is one of the techniques (along with traditional statistics) that intervenes at the step of "Data mining".

What makes data scientists important is that the steps of this kdd are highly interdependent.

You need indviduals or teams who are not just versed in data mining:

→ because the shape of the data at the collection stage has a huge influence on the kind of techniques, and the kind of software, that can be used to discover knowledge.

The skills of a data scientist are often represented as the meeting of three separate domains:

*Figure 77. The Venn diagram of what is a data scientist*

source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

# 4. Artificial intelligence

## a. Weak vs Strong AI

Weak AI designates computer programs able to outperform humans at complex tasks with a narrow focus (playing chess)

Weak AI is typically the result of applying expert systems or machine learning techniques seen above.

Strong AI is an intelligence that would be general in scope, able to set its own goal, and conscious of itself. Nothing is close to that yet.

So AI is a synonymous with weak AI at the moment.

## b. Two videos to understand AI further

Laurent Alexandre on the social and economic stakes of AI (in French):

▶ https://www.youtube.com/watch?v=rJowm24piM4 *(YouTube video)*

John Launchbury, the Director of DARPA's Information Innovation Office (I2O) in 2017:

▶ https://www.youtube.com/watch?v=-O01G3tSYpU *(YouTube video)*