



UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI INFORMATICA

BIG DATA AND DIGITAL METHODS

LECTURE 3

BIG DATA: CASE STUDIES

Dario Malchiodi

A FIRST TALE: THE H1N1 VIRUS

- Discovered in 2009, combining elements of viruses causing the bird and the swine flu.
- No vaccine readily available, thus the only option was to slow its spread.
- Need to (accurately and timely) localize infected people.

THE STANDARD WAY TO GO

- Doctors inform the CDC (Centers for Disease Control) about all flu cases
- Two-weeks lag
 - people tend to wait before going to the doctor
 - relaying info to central organizations takes time
 - CDC tabulates time only weekly
- Result: a poor timeliness!

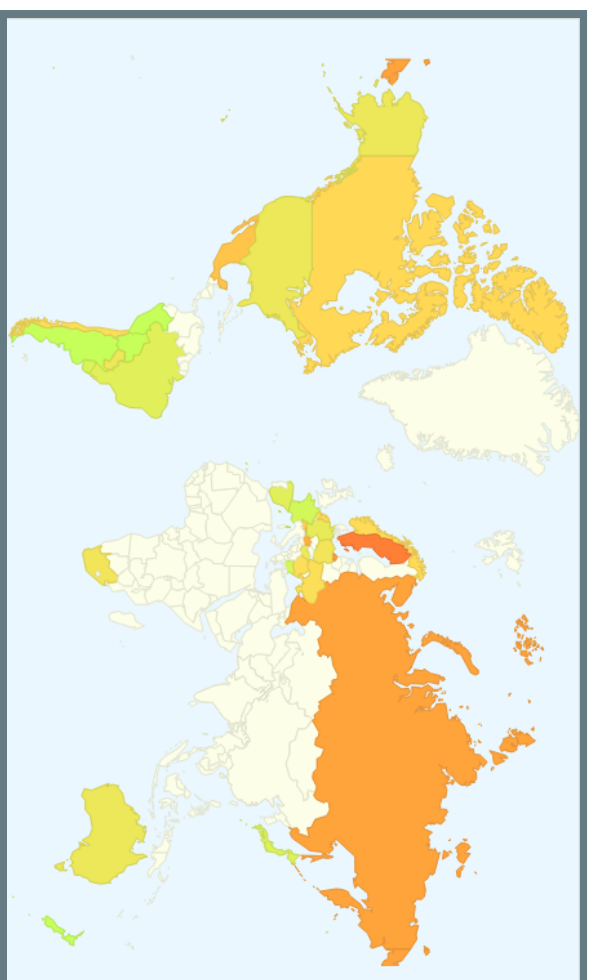
ACCIDENTALLY...

- Engineers at Google just published in *Nature* a paper about predicting the winter flu spread down to specific US states using a completely new approach
- No doctors involved, no medical tests involved: they analyzed search queries
- With prediction obtainable almost in real time (*nowcasting*)

HOW DID THEY DO?

- Informally: finding out which terms ill people types in the Google search box
- Technically: crossing the most frequent (geolocalsed) search terms between 2003 and 2008 with the CDC historical data about seasonal flu, in search for correlations
- Result: 45 search terms predicting the localization of winter flu during 2007 and 2008 with results comparable to those of CDC
- Not a new idea, but Google had a massive amount of data (three billions search queries per day) to process

GOOGLE FLU TRENDS



- Had a good prediction performance until early 2014
- Afterwards it started to overestimate flu cases
- Currently not operating: Google has released data and model predictions for research purpose

ANOTHER TALE: FARECAST

- In 2003 Oren Etzioni discovered that several passengers in his flight had spent considerably less than he had, although he bought his ticket online some months earlier than them
- So he became determined to figure out a way to know if an online flight offer was or not a good deal
- Without knowing the “internal” factors to the flight process itself (most factors influencing the price are known only by airlines)...
- ...he created a web site suggesting customers if the price of a ticket was likely to increase or decrease
- which evolved in a startup called Farecast (later acquired by Microsoft)

HOW DID HE DO?

- Informally, he scaled out the the informal survey made during his flight
- Technically, he “scraped” all ticket prices from the airlines web sites over a 41-day period...
- ...and built a model predicting the likelihood of a price increase or decrease

WHY THESE TWO TALES?

- In both cases, what made the difference was the availability of a very big amount of data
- That would not have been possible ten (or even five) years earlier, because of
 - too high price of computing power
 - too high price of storage
 - the availability of data
- More important, people started to think differently about data

A MINDSHIFT

- Data was no more seen as a “static” information to be collected, computed and discarded (after a plane landed or a user got his search results)
- It became a raw material of business, to be reused to create economic value
- The amount of processable data was crucial to this process

A MINDSHIFT

- Thanks to the diffusion of computers, laptops, cell phones, sensors...
- ...a wide amount of data has become available, and that amount is growing fast
- This change of scale led also to a qualitative change
- Big data refers not only to the use of massive quantities of information, but also to new, specific technologies and new fields (starting from genomics and astronomy)

BIG DATA

There is no rigorous definition (and a lot of hype, too)

- Big enough to not fit the memory of a computer
- Big enough to not fit the secondary storage
- Big because we consider all data instead of a sample
- Big because at a smaller scale we wouldn't get the results we expect:
 - extract new insights
 - create new forms of value
 - changing markets, organizations, citizenship

LOTS OF DATA AROUND

- in 2000, in some weeks telescopes could collect more data than had been amassed in the entire history of astronomy
- the first genome scan in 2003 took a decade to sequence the three billion base pairs; today the same task takes no more than five days
- Google processes 24 PBytes each day (thousands of times the printed material in the U.S. Library of Congress)
- Facebook gets 10 millions photos uploaded each hour, and three billions of “like” and comments daily
- Twitter gets more than 400 millions tweets daily (with a yearly growth rate of 200%)
- users upload over one hour of video on Youtube each second

A GROUND COMPARISON

Q: how much information has been produced and stored ever?

A: Martin Hilbert tried to evaluate the amount of information contained in books, paintings, (e)mails, vynils, photographs, TV broadcasts and so on

- In 2013: 1200 Exabytes, less than 2% in analog form
- In 2007: 300 Exabytes (roughly one billion movies), 7% in analog form
- In 2000: 75% of info in analog form (paper, film, vinyl, tapes...)
- In 1986 pocket calculators amounted to 40% of the world's computing power (more than PCs)

A GROUND COMPARISON

1200 Exabytes are roughly equivalent to

- 52 layers of books covering the whole U.S. surface
- 5 piles of CD-ROMs stretching up to the Moon
- 320 times the quantity of information in the Alexandria library per living person

INFORMATION GROWTH RATE

- Currently, it doubles every three years
- four times faster than the world economy (processing power grows nine times faster)
- Gutenberg printing press:
 - in the first 50 years: 8 million books
 - more than the documents written by scribes since Constantinople's foundation some 1200 years earlier
- Thus, 3 years vs 50 years

WHY IS BIG IMPORTANT?

- Because at that scale we can do things that weren't impossible using a smaller amount of data
- In other words, something simply changes
- An analogy (Peter Norvig): pictures
 - Iconic horse in a cave painting
 - then paintings, drawings
 - then photographs
 - and suddenly, a qualitative change: movies!
- Another (inverse) analogy: macro- and micro-scale in physics

WHY IS BIG IMPORTANT?

Using big data

- Google could speed up the process of identifying the prevalence of flu (faster than official sources) and without requiring actual visits of patients to doctors
- Farecast could predict the price volatility of airplane ticket, giving customers a substantial economic power

SEVERAL OTHER APPLICATIONS

Some already in use:

- Spam filters adapting to changes in junk style
- Dating sites finding similarities between users on the basis of correlation with previous successful matches
- Autocorrect function in smartphones adapting to our typing habits

Other ones start to show up

- from autobraking cars till Google car
- IBM's Watson computer beating humans in *Jeopardy!*

PREDICTIONS

- Most of the big data technology is about predictions
 - With which probability a mail message containing the word “vi6gra” is spam?
 - Should the automatic car break or simply slow in order to avoid hitting that person on the pedestrian walk?
 - Is “teh” an English word or should it be replaced by something else?
 - Which sites are most relevant for this search query?
 - Which books should be recommended to this user?
 - Which Facebook users this guy could know?
- Predictions made on the basis of collected data
- System improving itself over time

MACHINE LEARNING

- Computer science $\supset$ Artificial intelligence $\supset$ Machine learning
- Not related into teach computers how to «think» like a human being
- data + answers $\rightarrow$ prediction

FOR INSTANCE



- How do I discriminate between pictures of Sheldon Cooper and pictures of Leonard Hofstadter?
- I don't know how, but I can do that!
- possible strategy: machine learning

POSSIBLE STRATEGY: LEARNING

- collect several pictures of both characters
- label them as S (heldon) or L (leonard)
- feed a sample of labeled pictures to a learning algorithm
- get a *classifier* turning a generic picture into S or L
- key facts:
 - how do I test this classifier?
 - approximate results

TOOLS FOR LEARNING

- Bayes classifiers
- Neural networks
- Support Vector Machines
- and several other approaches
- note that there's more to learning than classification

THREE SHIFTS IN DATA ANALYSIS

- Far more data than usual, often *all of the data*
- Being satisfied with approximate results rather than aiming at exact ones
- Getting to know *what* rather than *why*

MORE

- before the advent of big data... sampling (analog technology, impossibility of storing all information,...)
- using all data → getting more details
- but also being able to work at different granularities

MORE

But also «new» kinds of information starts deserving attention:

- location of a person
- vibration of an engine

With interesting effects

- ability to predict new things
- less importance to «instinctive» managerial decisions
- democratization of tools and techniques for data-crunching (although not of data!)

MESSY

- big data → loose up the desire for exactitude
- tradeoff between sampling error and measurement error
 - if I have few data, I want them to be as much precise as possible
 - lower sampling error, higher measurement error
- a small grocery store may count the money in the cash register up to the penny
- it's impossible to do the same for a country's GDP
- thus we'll be satisfied with a sense of general direction

GOOD ENOUGH

- suppose you find out that a combination of aspirin and orange juice induces remission of a given cancer in 90% of patients...
- ...but you don't know why this work.
- Not knowing the exact cause of this health improvement has a secondary importance: those patients lived!
- in other words: content yourself in knowing that something happens (and use that information), rather than knowing why it happens
- for instance: using FareCast we don't know what actually affects the price of airline tickets; we (approximately) know whether or not that price will rise... a good enough information for customers

