Big data project

Kick-off

Profs. Gilles Louppe, Bertrand Cornélusse and Pierre Geurts

LIÈGE université

The data science era

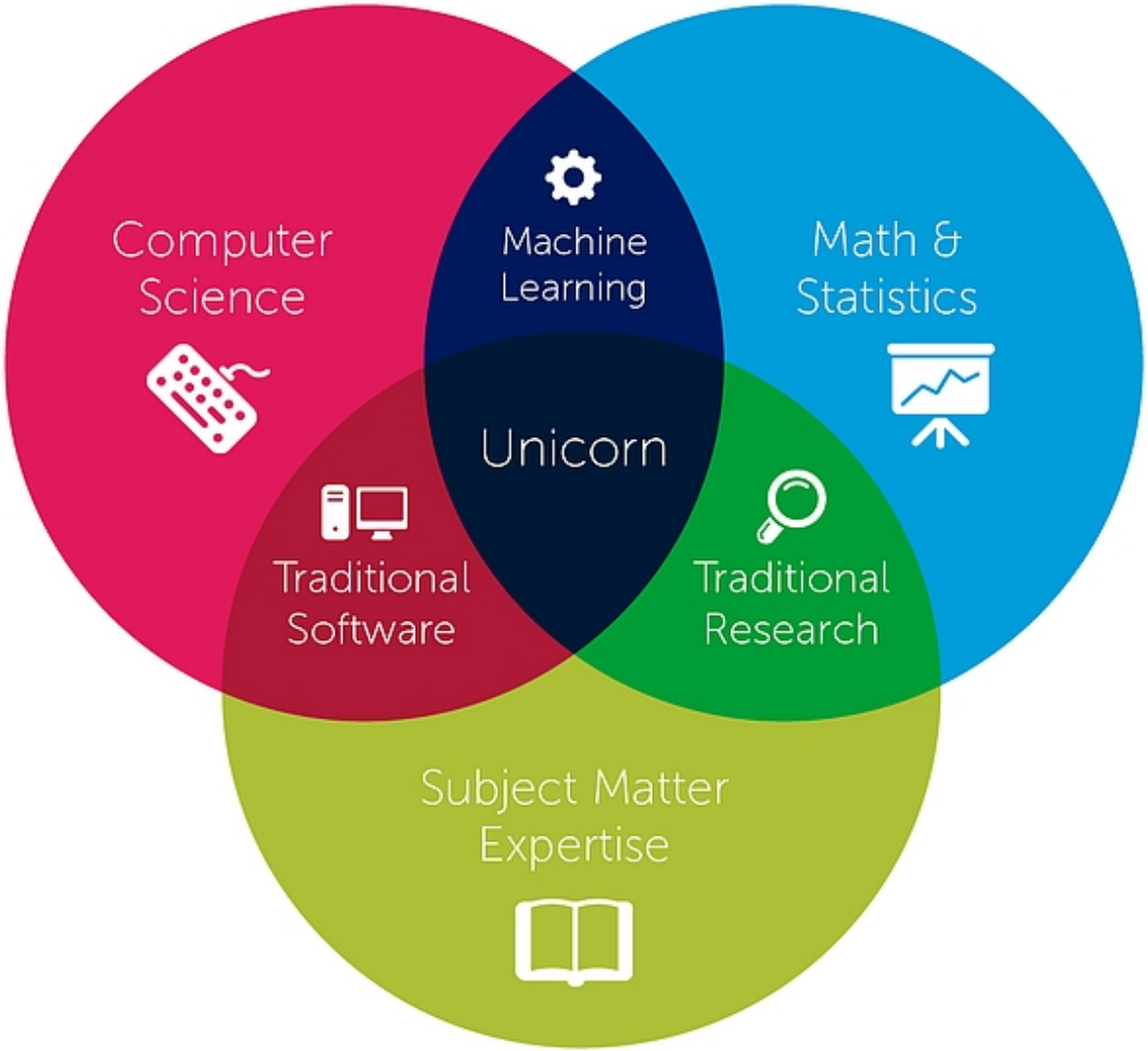
Big data? Data science?

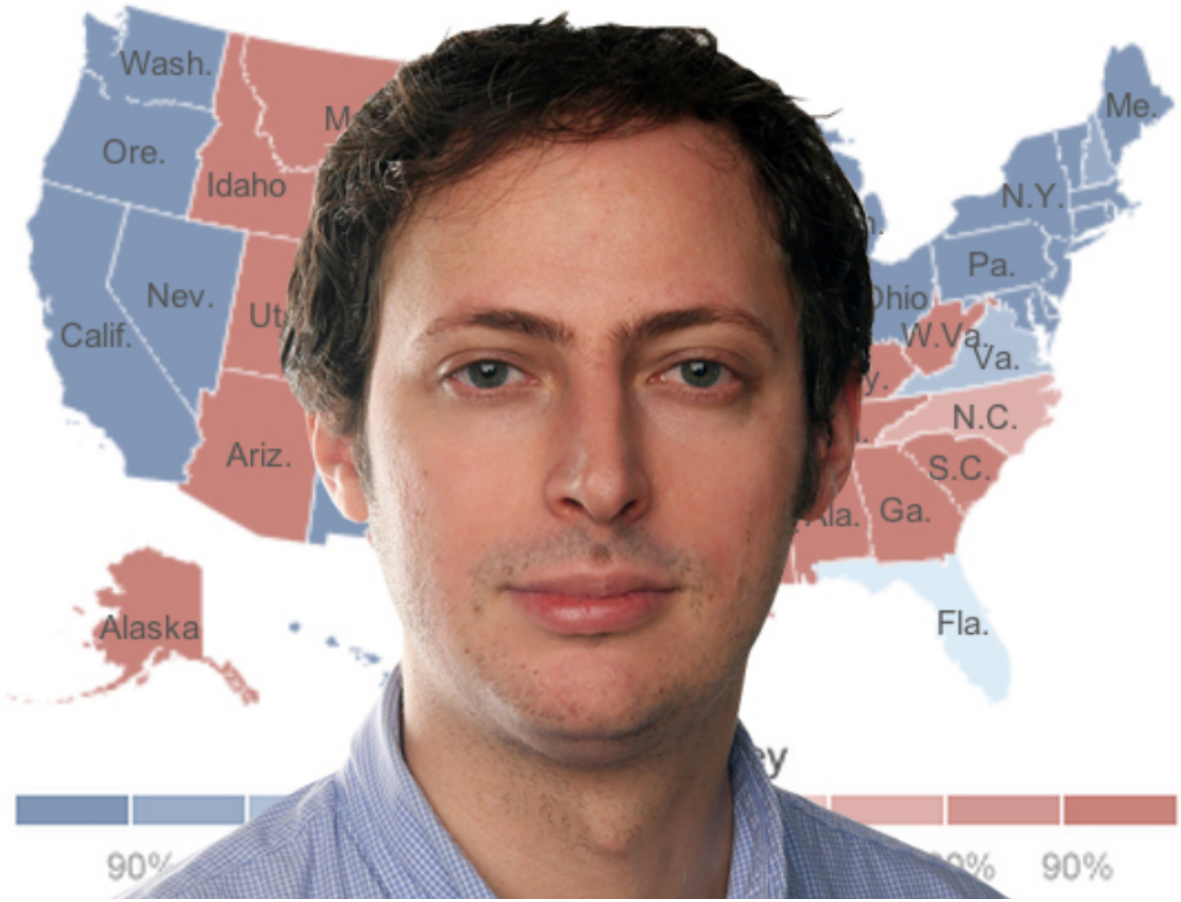
"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

Josh Blumenstock

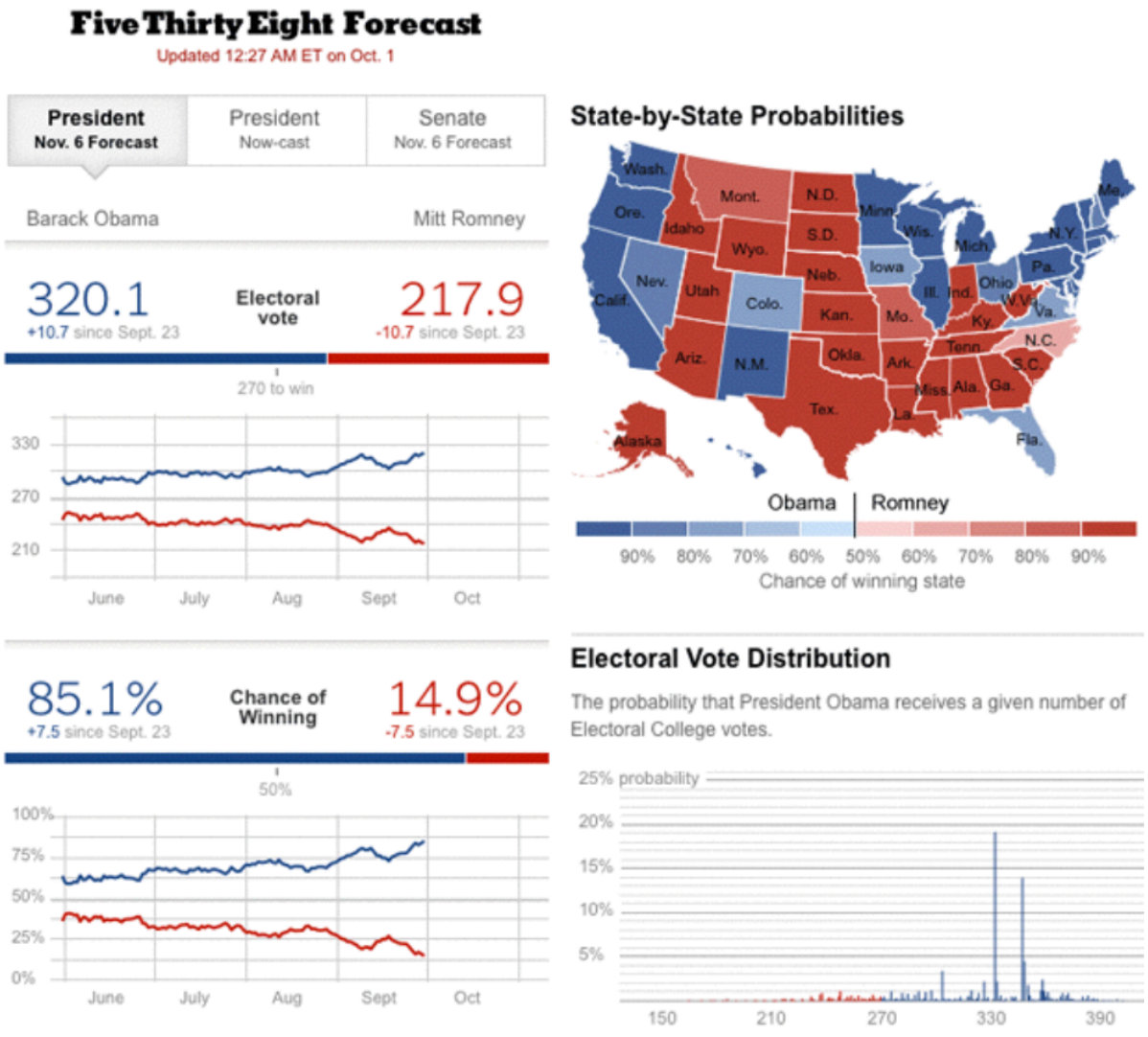
"Data scientist = statistician + programmer + coach + storyteller + artist"

Shlomo Aragmon


Computer Science
Machine Learning
Math & Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise


Wash.
Ore.
Idaho
Nev.
Calif.
Ariz.
Alaska
Me.
N.Y.
Pa.
W.Va.
Va.
N.C.
S.C.
Ga.
Fla.
90%
90%

Nate Silver



"Nate Silver won the election" - Harvard Business Review


analysis
data
knowledge

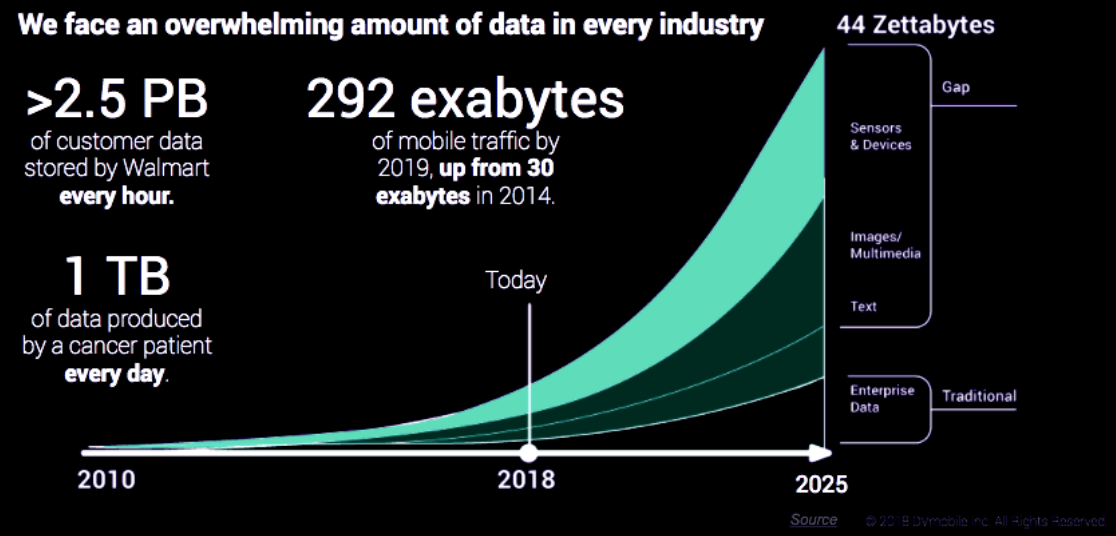
Haven't we be doing data analysis forever?

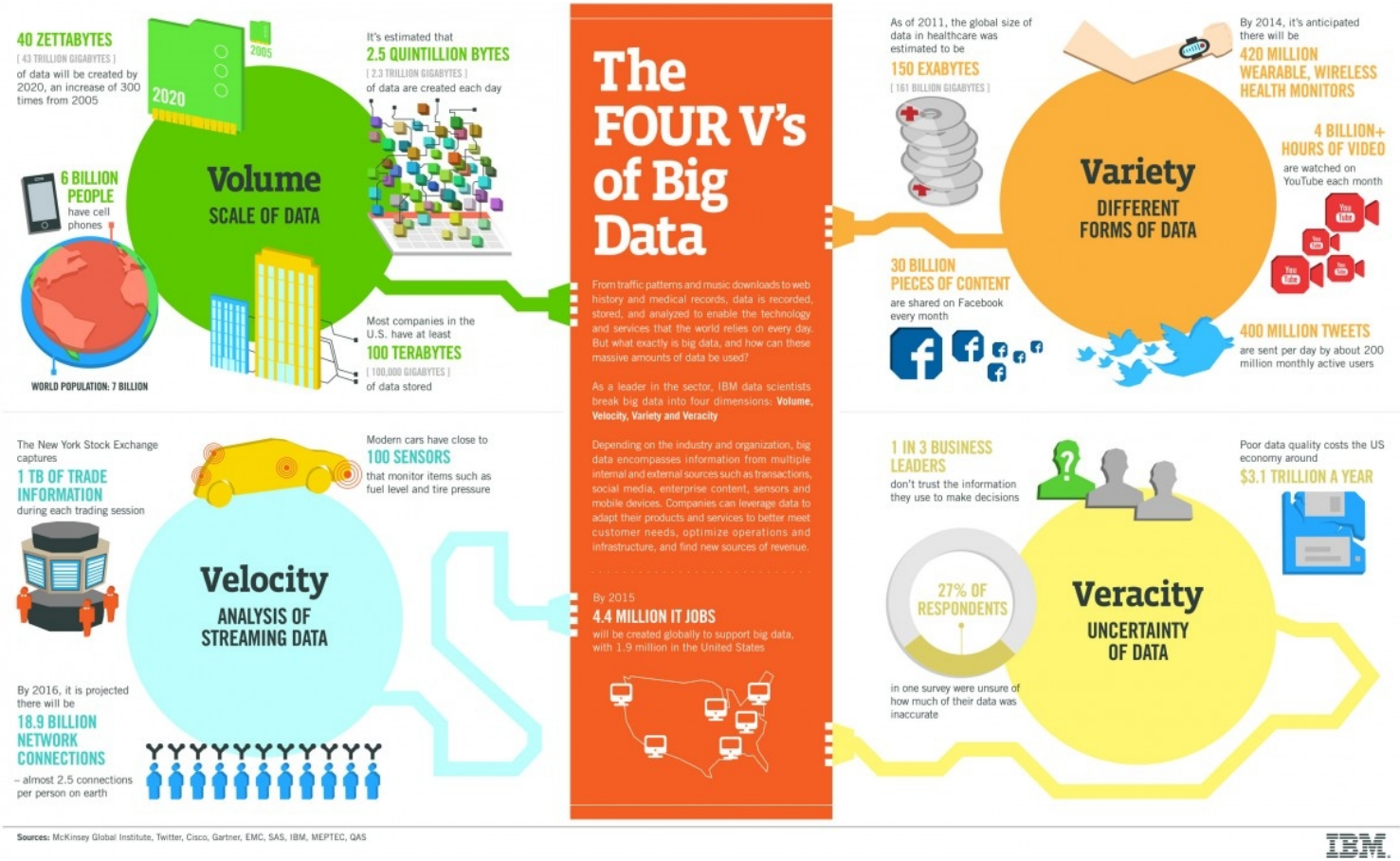
"Every two days now we create as much information as we did from the dawn of civilization up until 2003, according to Schmidt. That's something like five exabytes of data, he says.

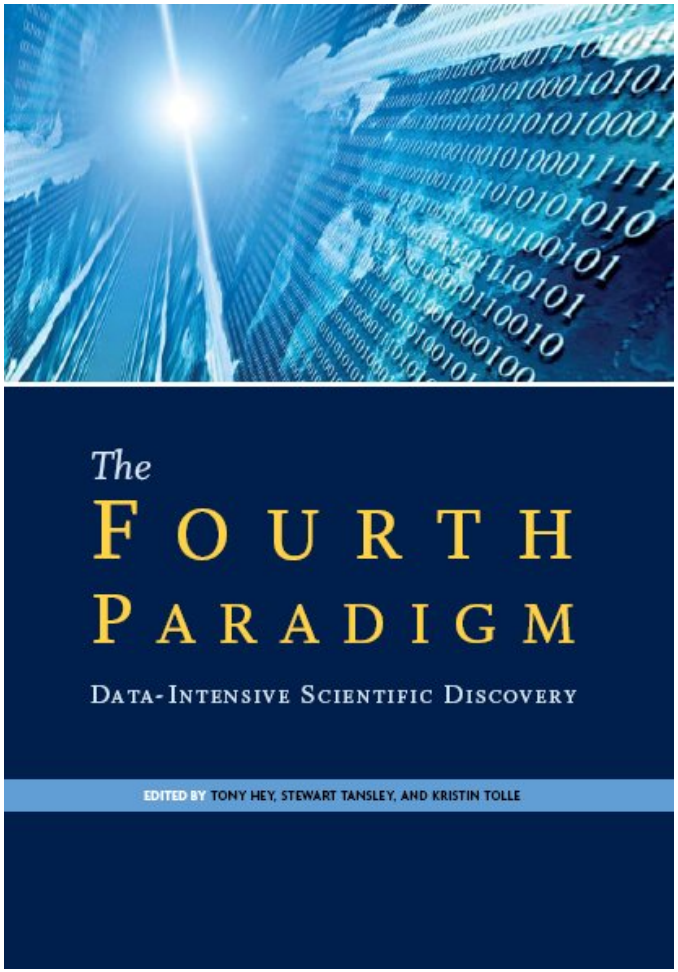
Let me repeat that: we create as much information in two days now as we did from the dawn of man through 2003.

Eric Schmidt, 2010.

1 Zettabyte (ZB) = 1 Trillion Gigabytes (GB)




40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM.



"Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies."

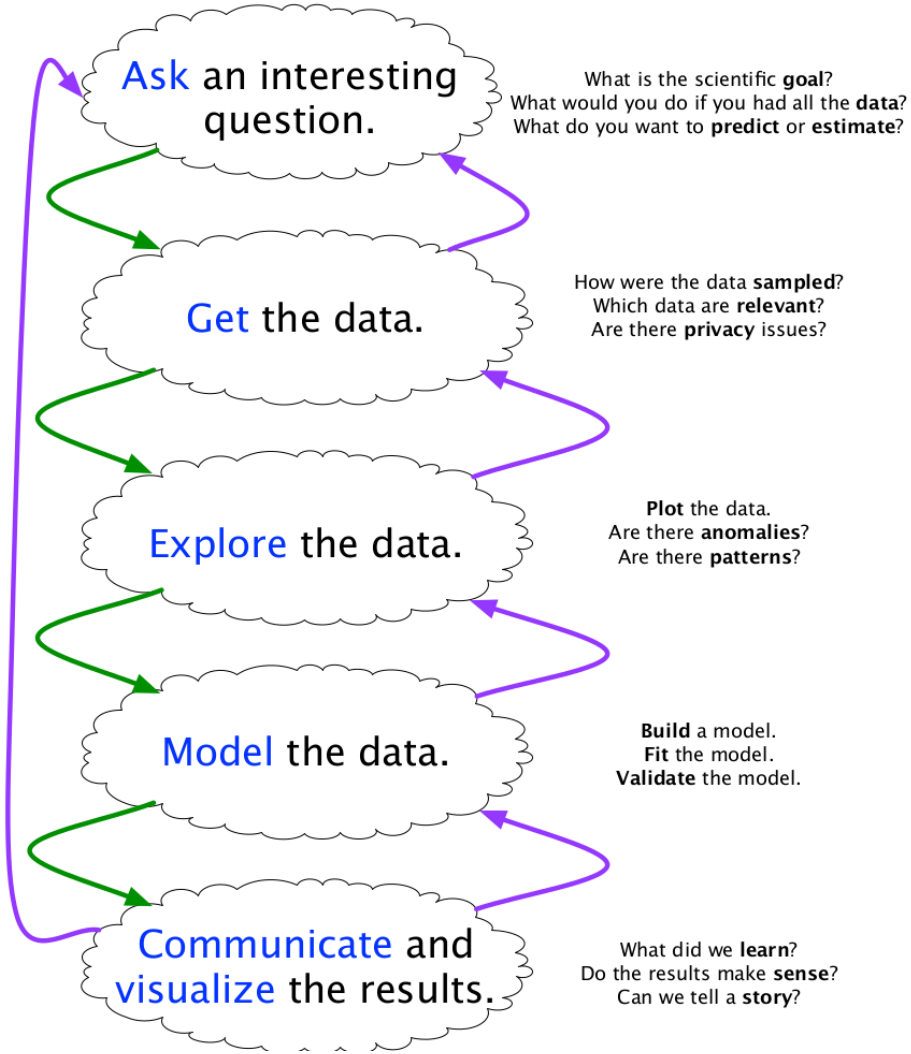
"By 2018, the US could face a shortage of up to 190,000 workers with analytical skills."

McKinsey Global Institute

"The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."

Hal Varian, Chief Economist, Google

The data science pipeline



In practice, the data science process involves several steps:

- Understanding and formalizing the problem
- Defining a model
- Collecting, cleaning and storing data
- Choosing a technology
- Analyzing the results
- Storytelling and visualization
- Iterate

Understanding and formalizing

- What is it that I really want to answer?
- Why do I want an answer to this question?
- Do I understand the problem?

Defining a model

- How do I answer?
- What are my assumptions?
- What statistical model do I consider?
- What algorithm shall I use?

Collecting, cleaning and storing data

- What data do I need for fitting my model?
- How large this data should be?
- Where do I collect this data?
- Is data cleaning necessary?
- How do I store the data?

Choosing a technology

- What tools do I need?
- What technology shall I use?
- Is a laptop enough, or shall I use a large-scale distributed system?
- How do I make my analysis reproducible?

Analyzing the results

- How do I analyze the results of the model?
- How do I assess the significance of the results?
- To what do I compare?
- What are the conclusions?
- Is this convincing?
- Does this corroborate with previous studies or intuition?

Storytelling and visualization

- How do I present my results?
- How do I make interpretable visualizations?
- How do I present my results to a non-technical audience?
- How do I make my results and conclusions as simple as possible, but not simpler?

Iterate, iterate, iterate

- Is this conclusive?
- Am I going in the right direction?
- Shall I go back and define a new model?
- ... or collect new or more data?
- ... or use other tools?

Your project this year

(Pick one!)

Who will win the 2019 French Open?

Is global warming for real?

Is the University an example to follow in terms of electricity consumption?

Organization

Instructors

This project is mentored by:

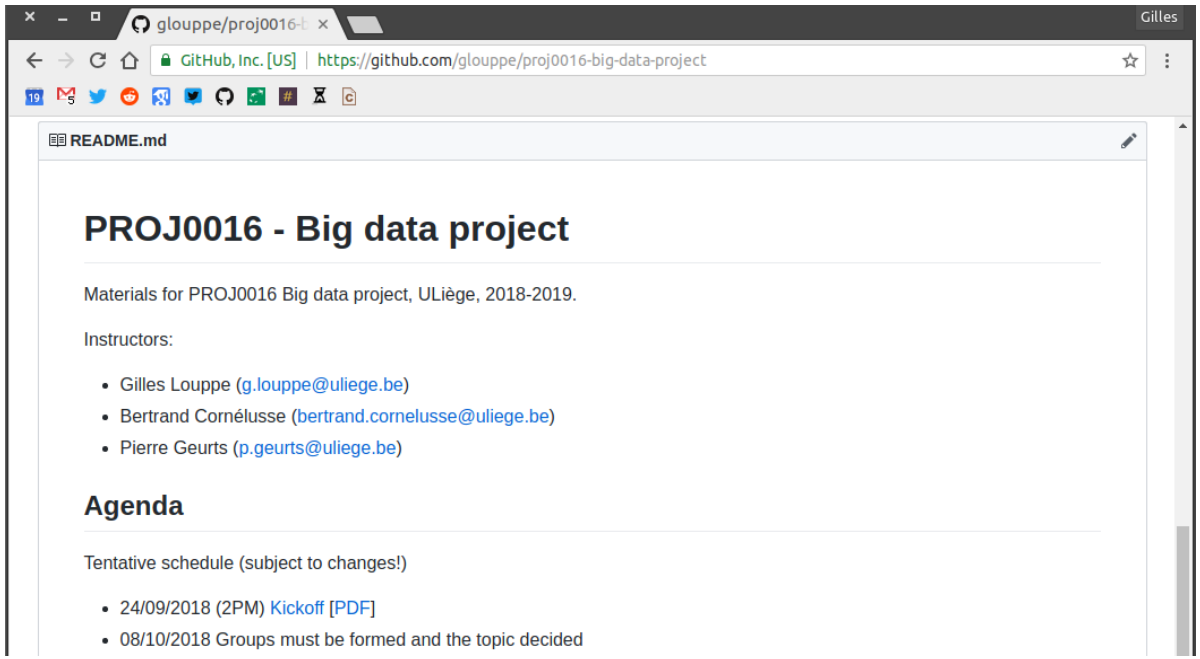
- Prof. Gilles Louppe (g.louppe@uliege.be)
- Prof. Bertrand Cornélusse (bertrand.cornelusse@uliege.be)
- Prof. Pierre Geurts (p.geurts@uliege.be)

Feel free to contact any of us for help!



Materials

Slides and other materials are available at github.com/glouppe/proj0016-big-data-project.



Groups

The project is carried out in groups of 3 students.

The topic should be selected and the groups should be formed by October 8.

- Notify us by email.
- If you are alone, send us an email too!

Reviews

We will meet on every last Monday of the month to review your progress.

- Oral presentation of your ongoing progress.
  - 10mn
  - Q&A
  - Everyone must present at least once
- Short report
  - 4 pages max
  - To be submitted on the Friday before the review day

The goal is to give you feedback on technical progress and project management.

Seminars

The project is complemented by seminars by local and external speakers.

- Topics: big data, data science, visualization, communication, domain-specific presentations, etc.
- Presence at the seminars and intermediate reviews is mandatory.

Final deliverables

The final deliverables of your project should consist in:

- a final comprehensive report of your study
- reproducible scripts for the collection, analysis and visualization of your data

Agenda

- 24/09/2018 (2PM) Kick-off
- 08/10/2018 Groups must be formed and the topic decided
- 29/10/2018 (2PM) Project review #1: Pre-analysis, literature review
- 26/11/2018 (2PM) Project review #2: Data collection #1
- 17/12/2018 (2PM) Project review #3: Data analysis #1
- 25/02/2019 (9AM) Project review #4: Data collection #2
- 25/03/2018 (9AM) Project review #5: Data analysis #2
- 29/04/2018 (9AM) Project review #6: Further improvements.
- 13/05/2018 (9AM) Project public defense and final report

Seminars will be announced later.

Evaluation

The evaluation will be based on:

- the intermediate review meetings (progress achieved, quality of project management) (6x 5%)
- the quality of the final report (15%)
- the quality of the final oral defense (15%)
- the overall study (40%)
    - the originality, methodology, clarity, reproducibility and technological choices of the solution will be mainly assessed.

Let's get to work!