Big data project

Kick-off

Profs. Bertrand Cornélusse, Pierre Geurts and Gilles Louppe

LIÈGE université

Today

- What is data science?
- Solving data science problems
- Course and project organization

The data science era

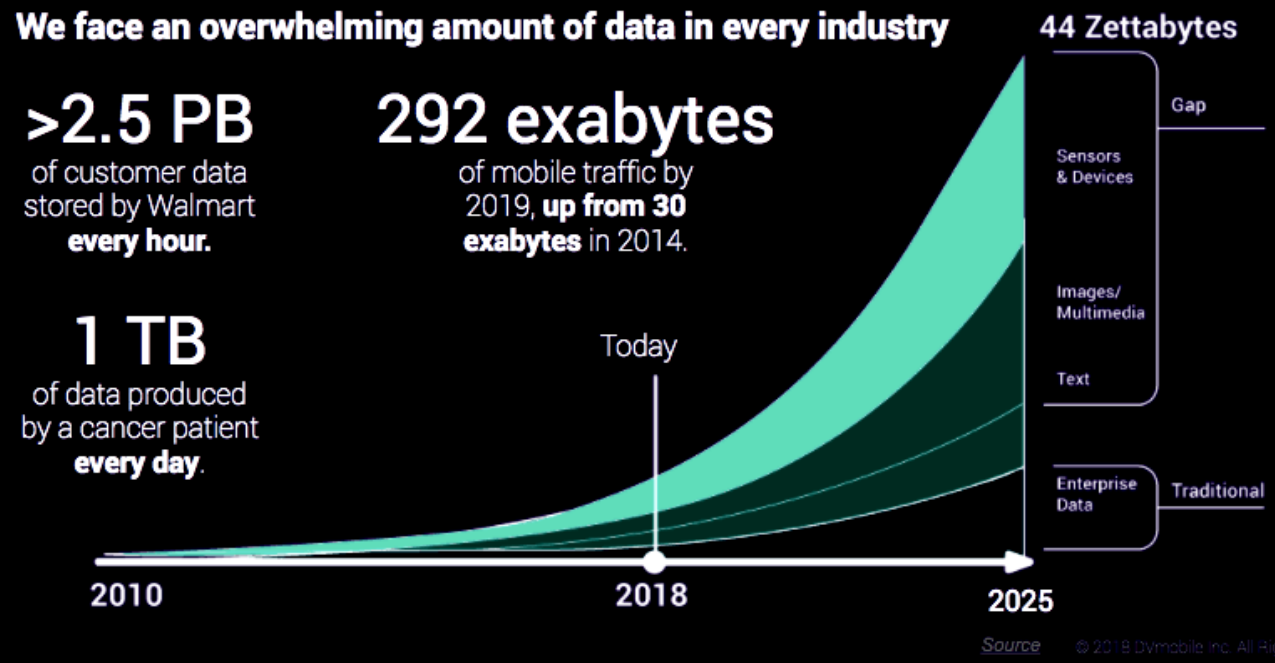

DATA
DATA EVERYWHERE

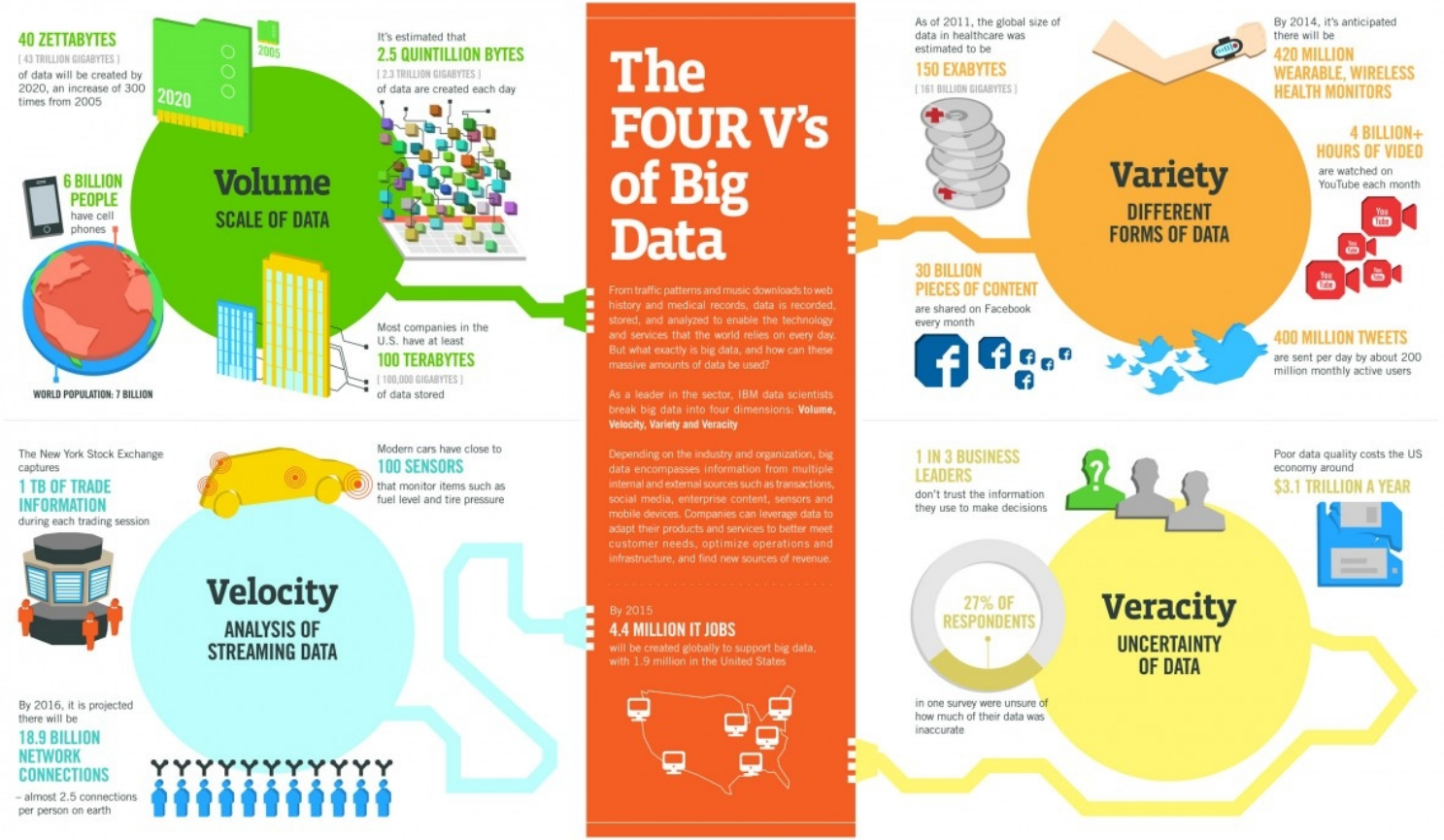
"Every two days now we create as much information as we did from the dawn of civilization up until 2003, according to Schmidt. That's something like five exabytes of data, he says.

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

"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."

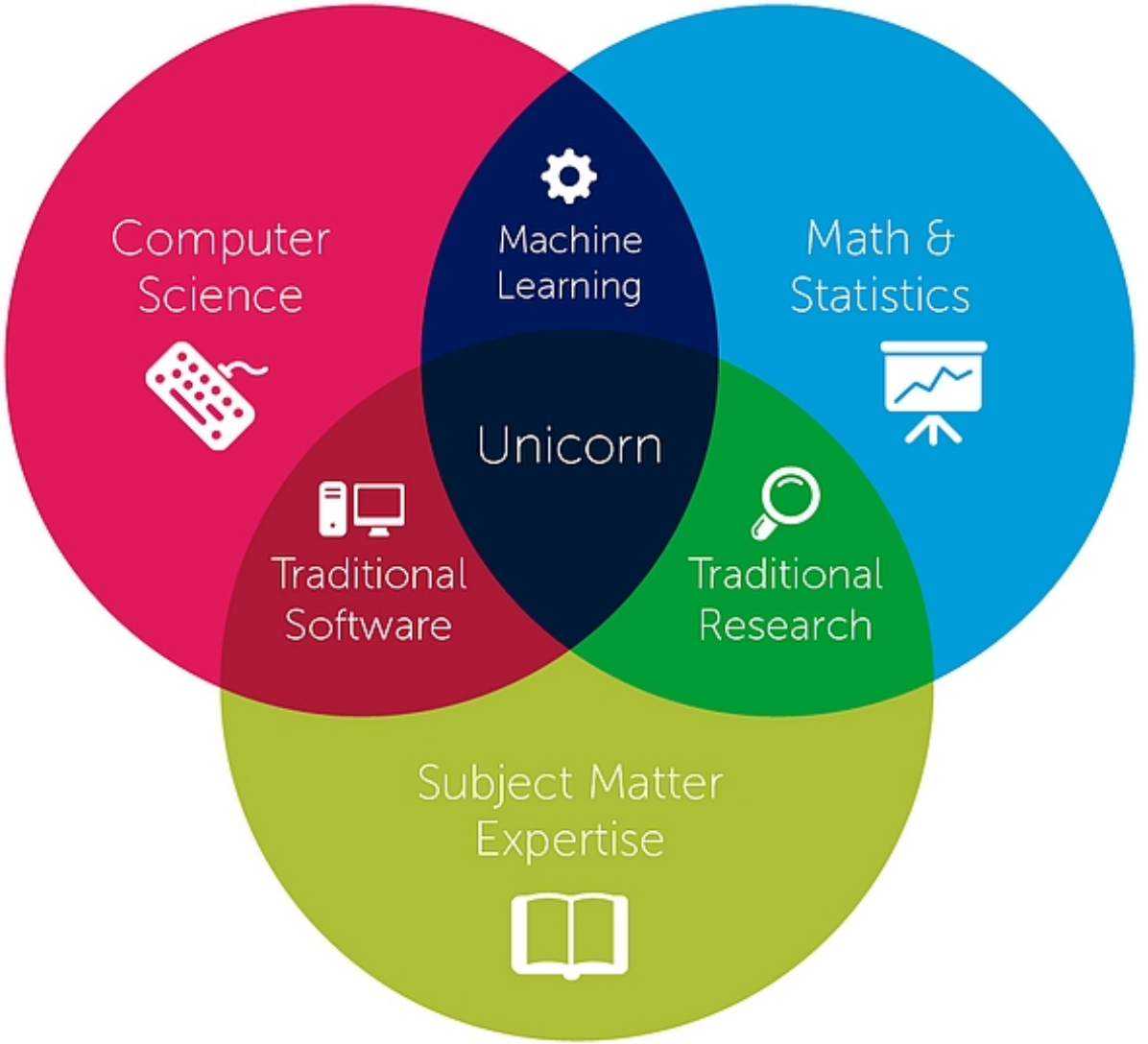
Hal Varian, Chief Economist, Google

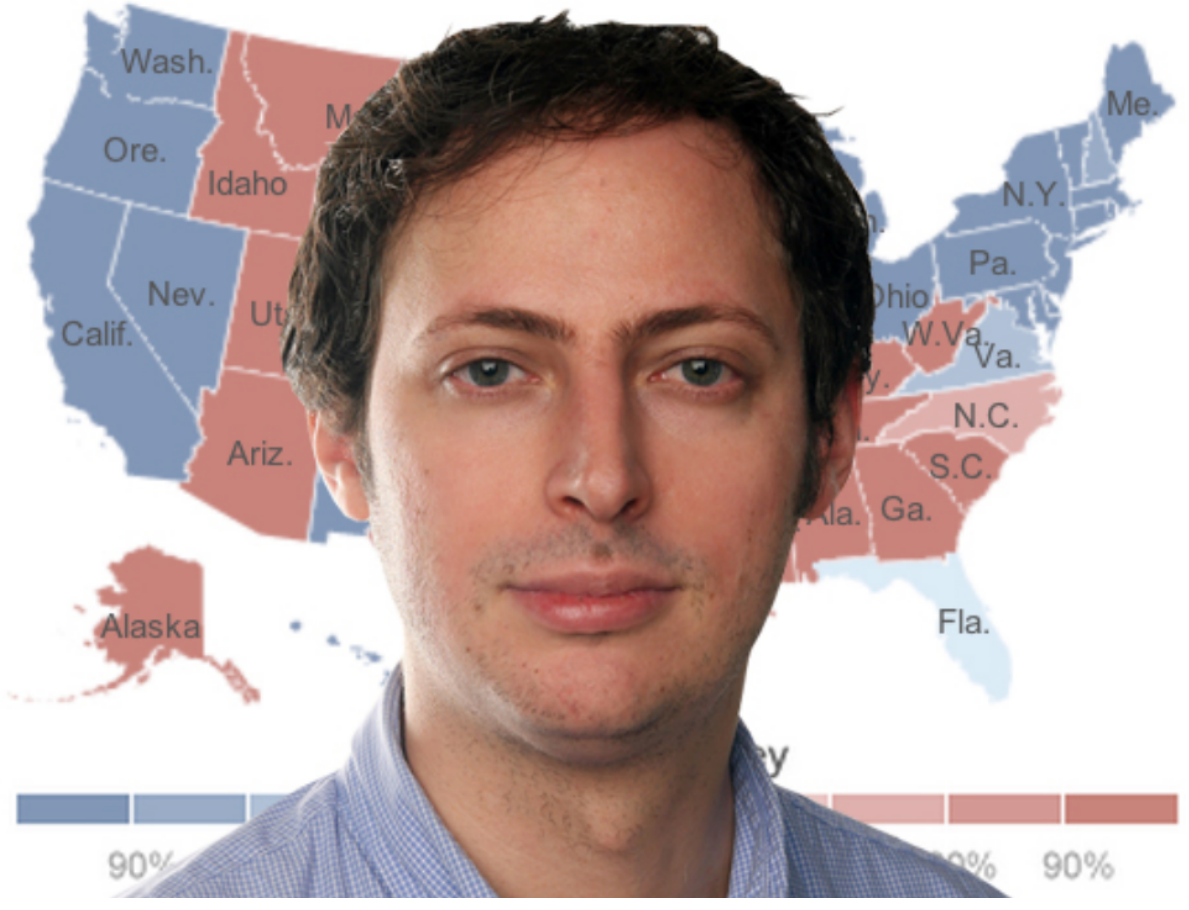
"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

Josh Blumenstock

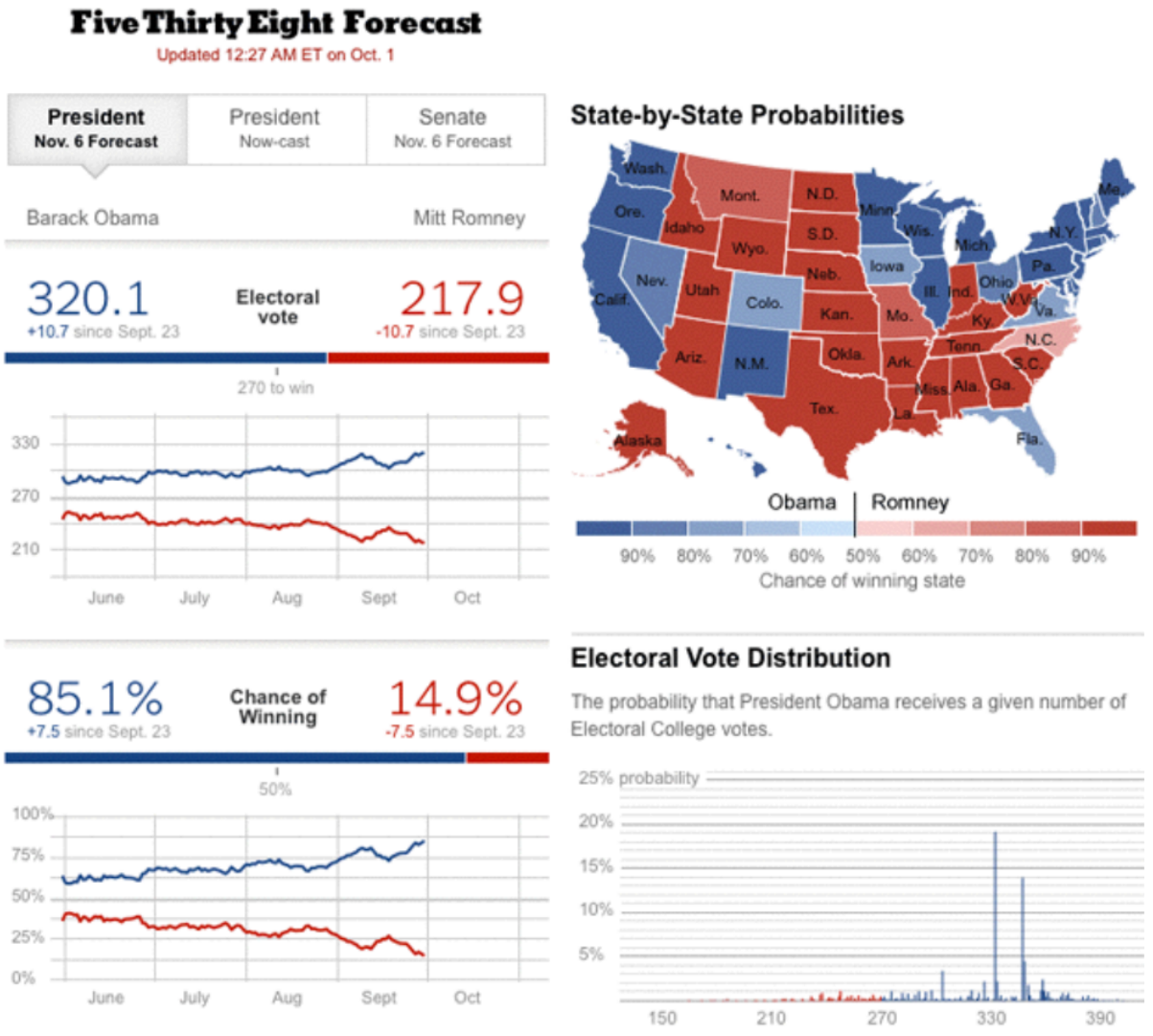
"Data scientist = statistician + programmer + coach + storyteller + artist"

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

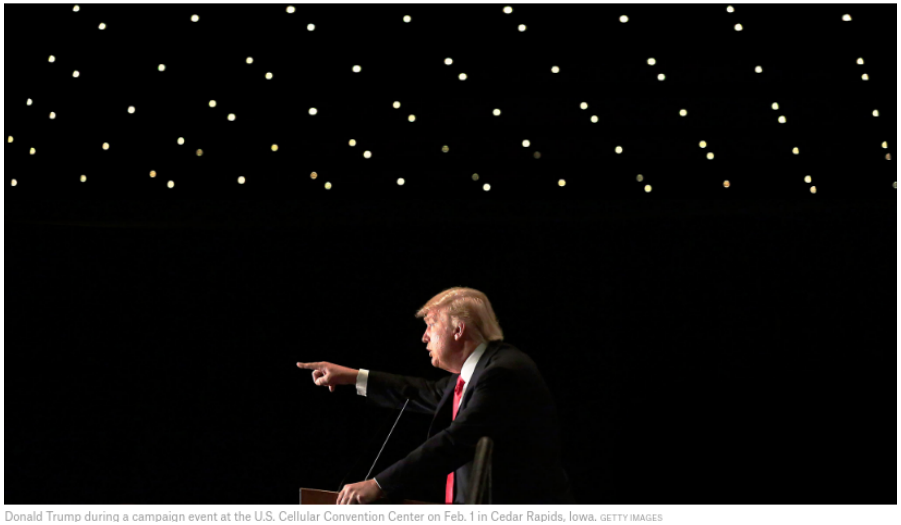
FiveThirtyEight

Politics Sports Science & Health Economics Culture

How I Acted Like A Pundit And Screwed Up On Donald Trump

Trump's nomination shows the need for a more rigorous approach.

By Nate Silver
Filed under 2016 Election
Published May 18, 2016



Donald Trump during a campaign event at the U.S. Cellular Convention Center on Feb. 1 in Cedar Rapids, Iowa. GETTY IMAGES

Our early forecasts of Trump's nomination chances weren't based on a statistical model, which may have been most of the problem.


Cambridge
Analytica
The Power of Big Data and Psychographics
in the Electoral Process


THEY TOOK YOUR DATA.
THEN THEY TOOK CONTROL.
A NETFLIX ORIGINAL DOCUMENTARY
THE GREAT HACK
From the Academy Award® nominated filmmakers behind THE SQUARE
sundance
JULY 24 | NETFLIX

Solving data science problems

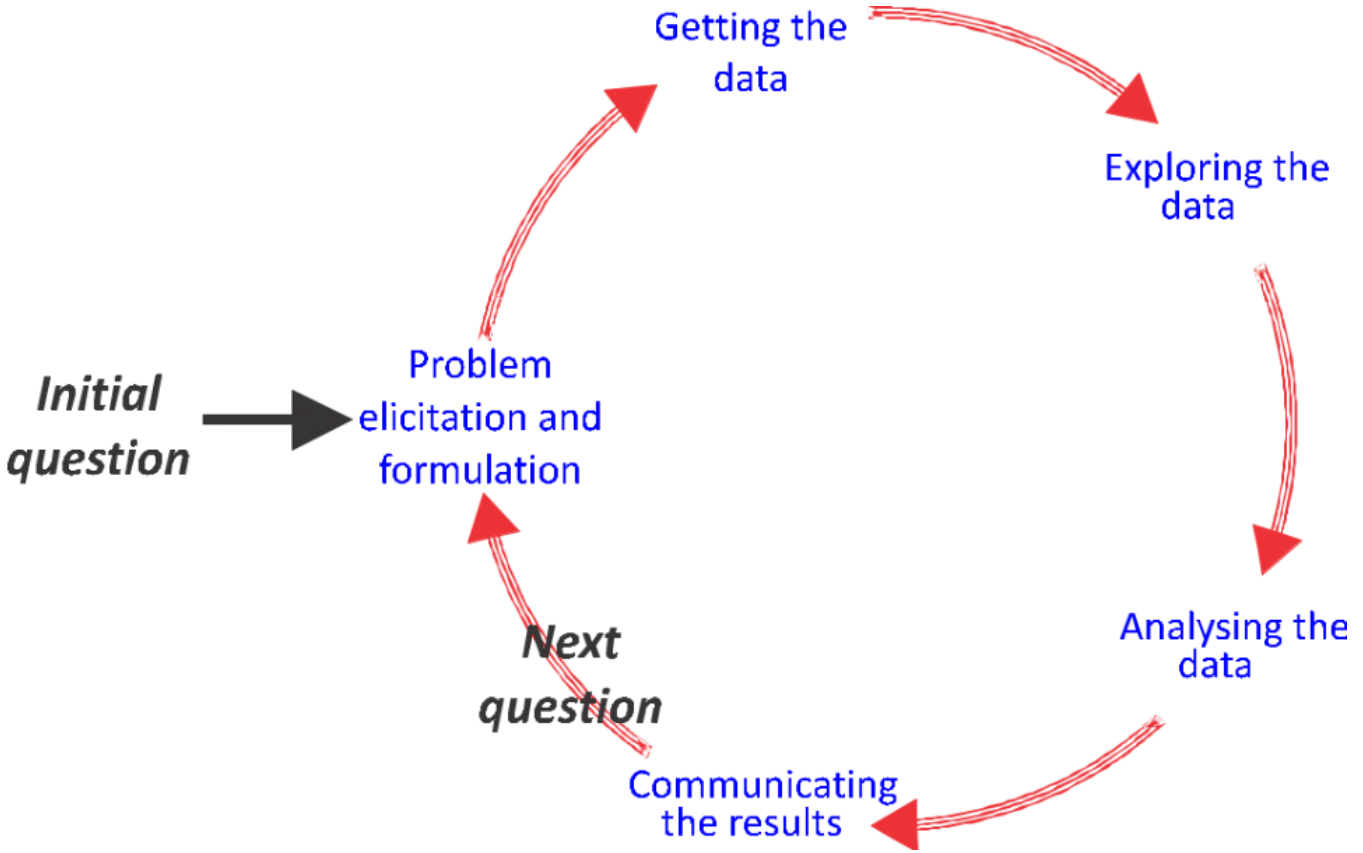


Image credits: International Data Science in Schools Project, 2018.

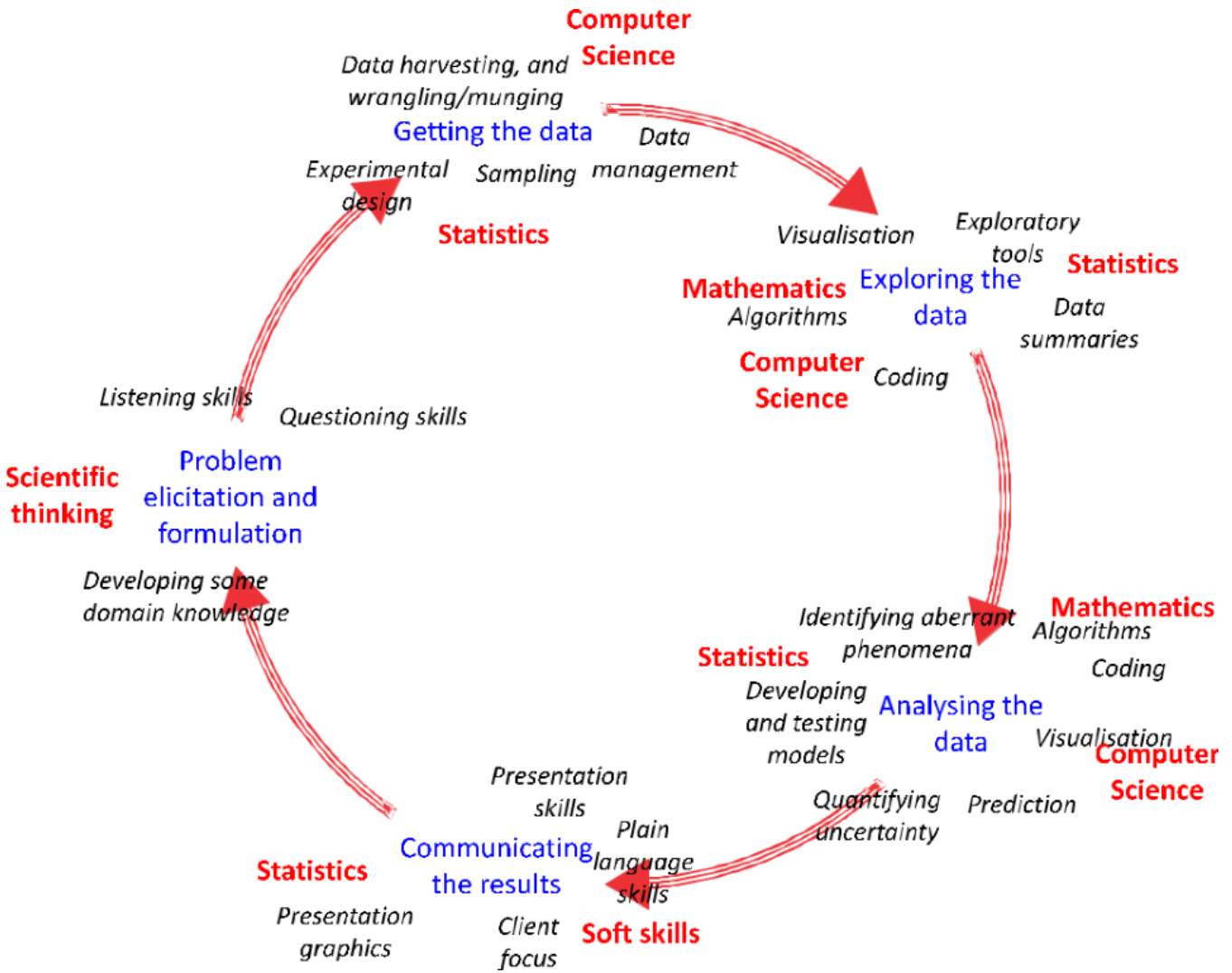


Image credits: International Data Science in Schools Project, 2018.

Understanding and formulating the problem

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

Collecting, cleaning and storing data

- What data do I need?
- How large this data should be?
- Where do I collect this data?
- Do I have to run an experiment to collect data?
- Is data cleaning necessary?
- How do I store the data?

Exploring the data

- How does the data look like?
- What are summary statistics?
- Does this look consistent?

Analyzing the data

- What statistical models should I consider?
- How do I analyze the results of the model?
- How do I assess the significance of the results?
- How do I validate my hypothesis?
- To what do I compare?
- What are the conclusions?
- Is this convincing?
- Am I confident about these results?
- 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?

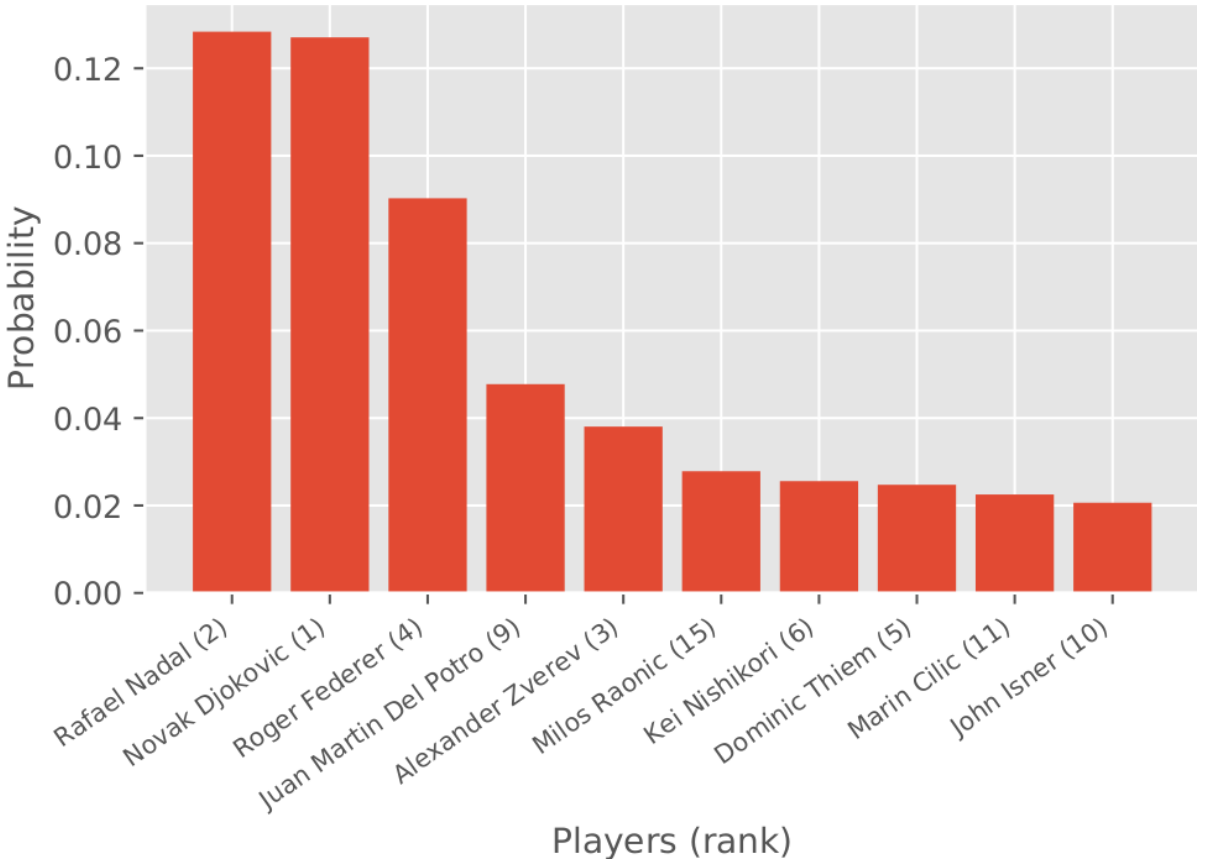
Iterate, iterate, iterate

- 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?

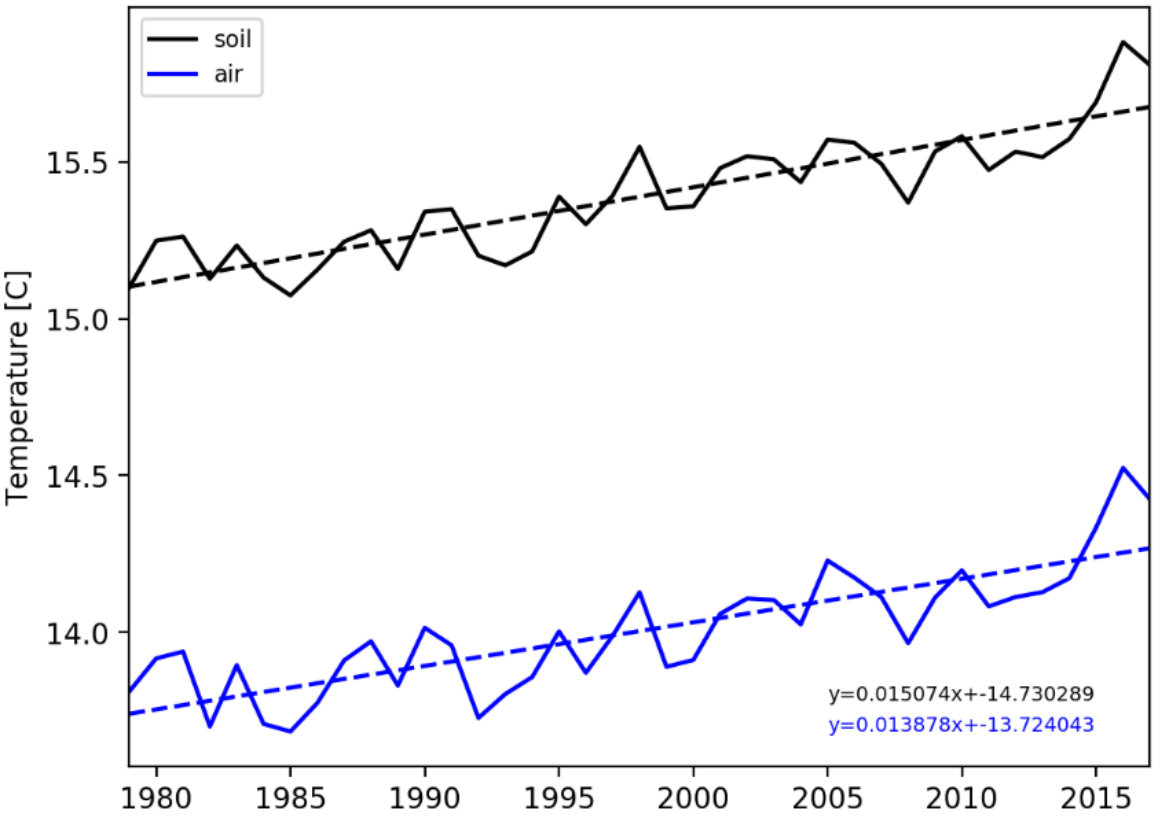
Choosing a technology

- 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?

Past project examples



Predicting the 2019 French Open's winner (Louis, Crasset, Lamborelle)



Is global warming for real? (Nicolas, Mathy, Ivanov)

Your project this year

Proposals to be announced later! (around late October)

(We want you to have followed tutorials first)

Course and project organization

Instructors

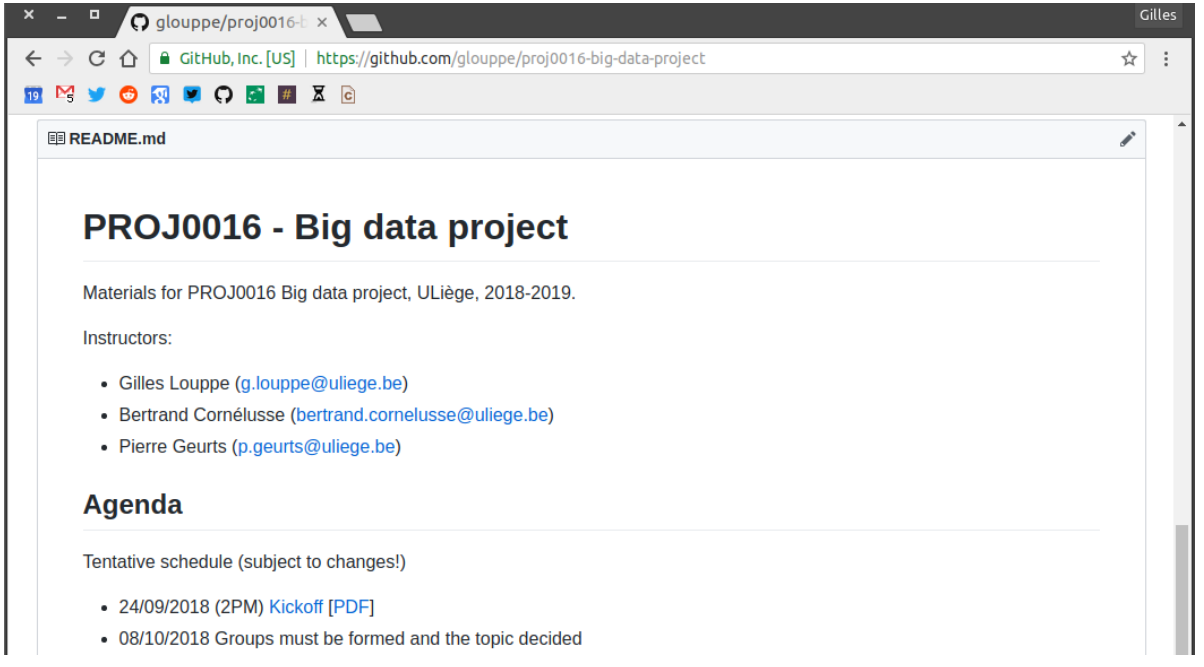
The project is mentored by:

- Prof. Bertrand Cornélusse (bertrand.cornelusse@uliege.be)
- Prof. Pierre Geurts (p.geurts@uliege.be) (Not this year!)
- Prof. Gilles Louppe (g.louppe@uliege.be)
- Jonathan Dumas (jdumas@uliege.be)




Agenda and materials

The agenda and course materials can be found on the Github page of the course, at github.com/glouppe/proj0016-big-data-project.



Seminars and tutorials

The project is complemented by seminars and tutorials 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.

Groups

The project is carried out in groups of 3 students.

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 Thursday before the review day.
- The agenda for the reviews will be announced together with the project proposals.
- The goal is to give you feedback on technical progress and project management.

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

Evaluation

The evaluation will be based on:

- the intermediate review meetings (progress achieved, quality of project management) (30%)
- 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.

We will notify your score individually after each intermediate review stage.

Regularization

- Absence of a group at a review: 0 to this review for the group.
- Delay in submitting a report: -1% per day.
- Absence of a group member at a review: individual 0 for the review.
- Absence of a group member at a seminar: individual -5% on the total of the year.

How to report

- One of the outcomes of the project is to learn how to communicate / report orally and in written form.
- You have to be on time so that we can provide feedback : reports on Thursday 23:59 before review. Hard deadline!
- Create your own template and always use the same for the intermediate review stages.
  - If you make incremental reports, make sure you highlight what is new.
  - If not, introduce a status of your past reports at the beginning.
- Use a strict methodology: follow the steps of slide 16.
  - briefly report what you did for each step.
  - provide approximate percentage of completion for each step.
- Identify which group member is responsible for what.
- For oral presentations, stick to the allocated time slot.

How to work efficiently

- Make small iterations first but try to go/project over all the steps.
- Write what you do. It will help you being on time for your reports.
- Share the work between group members and give precise responsibilities.
    - Make use of collaborative tools! (e.g., Github, Gitlab, Trello, etc)

Let's get to work!