A Network Tour of Data Science (NTDS)

Projects

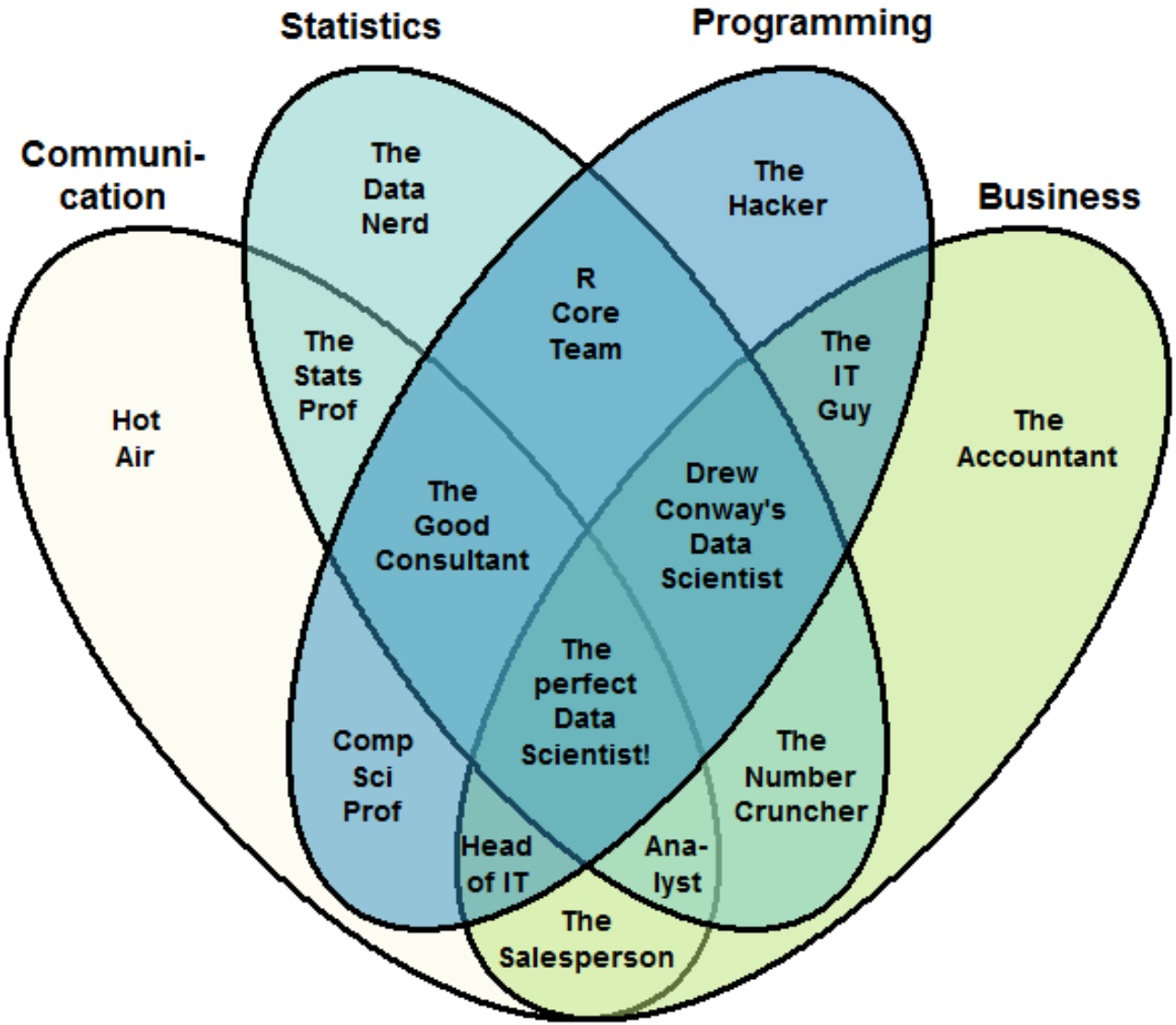
Teachers
Pierre Vandergheynst
Pascal Frossard

Assistants
Michaël Defferrard
Effrosyni Simou
Hermina Petric Maretić

EPFL LTS2 & LTS4 laboratories

November 13, 2017

Data Scientist



Project

1. Define a problem.
   - Form groups of 3 or 4 students (4 preferred).
   - Write a short but convincing proposal.

2. Solve it.
   - Use the concepts learned in class.
   - Follow the Data Science process.

3. Handle your solution for grading.
   - Jupyter notebook as report.
   - Oral presentation.

Problem

Find a problem you want to solve.
Think about your interests: scientific, hobbies, or otherwise.

Use **graphs** (A Network Tour) and **real data** (of Data Science).

- Network Science: study networks! Tasks: analysis of properties, generative models, epidemics, etc.
- Spectral Graph Theory: use the eigendecomposition of the graph Laplacian! Tasks: study of network properties, clustering, visualization, etc.
- Graph Signal Processing: analyze signals defined on graphs! Tasks: information diffusion (e.g., for matrix completion and recommendation), denoising, semi-supervized learning, etc.[1]

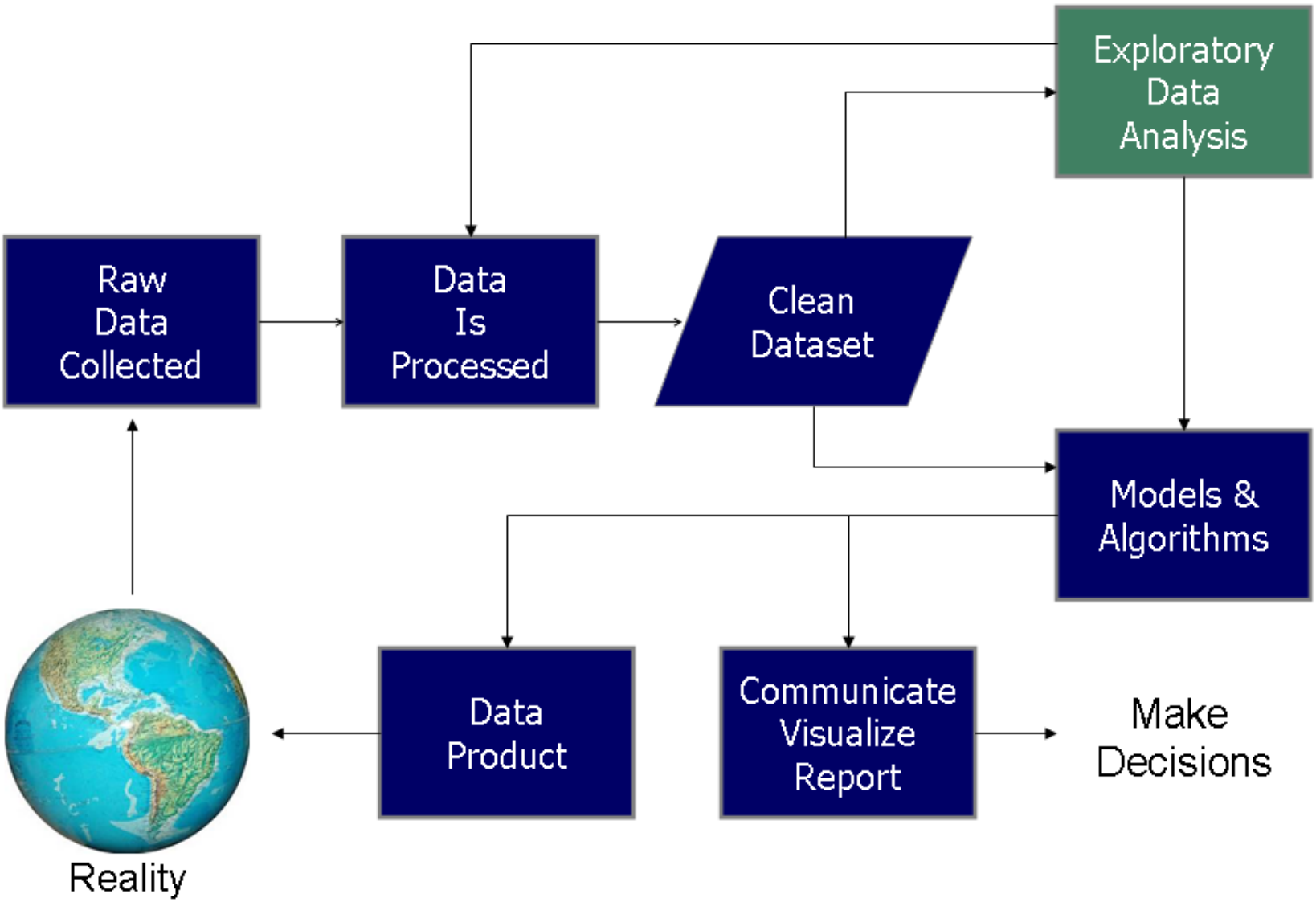
[1] You can take a look at the PyGSP tutorials.

Data

- Your own data, e.g., from your research.
- Call a web API or scrap a website (as assignments 1 and 3). Social websites are a wealth of information.[2]
- From challenges, e.g., on kaggle or crowdAI.
- Some (list of) datasets.
  - Free Music Archive (FMA).
  - Wikipedia graph & visits. See e.g., this blog post.
  - Stanford Large Network Dataset Collection (SNAP).
  - Citation networks (Cora, arXiv, PubMed). See e.g., arXiv viz.
  - The Network Repository.
  - A list of datasets for network analysis.
  - Awesome Public Datasets.
  - Swiss open data.
- More: transportation, communication, neural (artificial or biological), energy networks.
- Any other. Discuss with us!

[2] Twitter, Facebook, GitHub, Pinterest, Stack Overflow, YouTube, LinkedIn, Instagram, Tumblr, last.fm, reddit, etc.

Data Science Process



Structure

The structure of the notebook shall follow the Data Science process seen during the lab sessions.

1. Data acquisition: from the web, a database, a flat file, etc. This includes cleaning the data.

2. Data exploration: some exploratory analysis to describe properties of the data and understand the content.

3. Data exploitation: use the data to solve a task, to infer knowledge, to draw conclusions. The concepts or algorithms taught in class must be used.

4. Conclusion: discuss the results and summarize your findings. What did we learn from the data and the project?

Practical aspects

- Please isolate code blocks in functions and put those in a separate Python module.
- Your notebook should be clean and legible. They are akin to a report.
- You can take inspirations from the notebooks seen during the lab sessions.
- Look at last year projects.

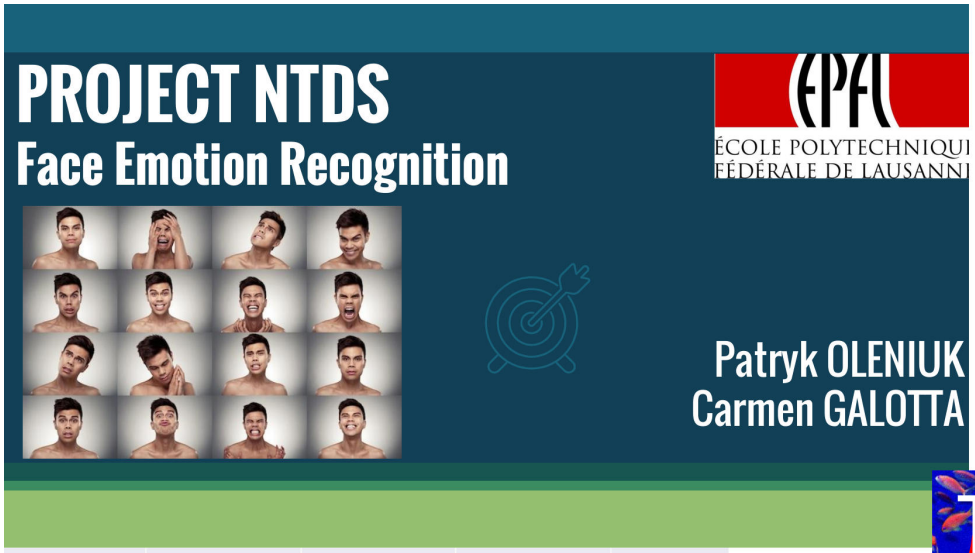
Rules

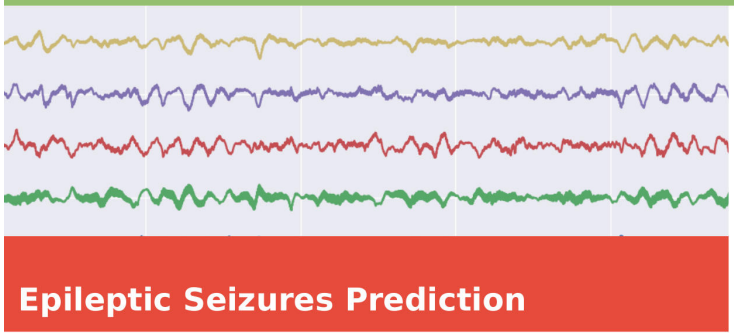
- The project includes graph and network data aspects, and more generally falls under the scope of the class.
- Form groups of 3 or 4 students. No less, no more.
  - One member of the group uploads the deliverables.
  - The names of all members should appear clearly.
- The project can be shared with another course (e.g., data visualization or ADA). State it clearly and specify what gets graded for which course.
- The project should follow the data acquisition, exploration and exploitation workflow.
- Data must not be synthetic. While manually collecting data is optional, i.e., the use of datasets is allowed, it is a plus.
- Each member of a team shall contribute equally to the project.

Organization

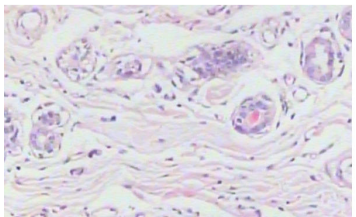
1. Proposal: define the problem and explain your plan.
   - Single page document.
   - Organize yourselves in groups of 3 or 4 students.
   - Deadline: Tuesday, November 28, 2017. Upload on Moodle.
   - Not graded. Discussion with TAs will follow.
2. Report: your solution, using the theory seen in class and the practical skills trained during labs.
   - Jupyter notebook with text, math, code, analysis and results.
   - The notebook will be posted on the course git repository, on GitHub. You can use it for your portfolio!
   - Deadline: Friday, January 12, 2017. Upload on Moodle.
   - Graded (project accounts for 50% of class grade).
3. Presentation: impress us with your work!
   - Presentation of 15 minutes followed by 5 minutes of questions.
   - Each group member must talk.
   - Register for January 24 or 25 (once exams are scheduled).
   - Graded (project accounts for 50% of class grade).

Have fun! Questions?





Sophie du Bois



BREAST CANCER

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Open Source Software Support
A Network Tour of Data Science

Matthaios Olma
Pavlos Nikolopoulos
Stefanos Skalistis