M2 Data Science - Data Camp Syllabus

You will put your basic machine learning and data analysis knowledge to test by

1. **solving practical data science problems** in scientific or industrial applications and by
2. **designing data science workflows**.

Code-submission data challenge

To achieve the first objective, you will participate in a **data challenge at the RAMP site**. The particularity of RAMPs (vs Kaggle) is that you will **submit code**, not predictions. Your code will be inserted into a predictive workflow, trained and tested. A public cross-validation score will be available on a public leaderboard, real time. Your **grade will be a function of the private test score of your best submission**, obtained on a hidden test set. The challenge will also include a **collaborative phase** in which you can access all the submitted solutions, and you will be allowed and **encouraged to reuse each other's code**. Part of your grade will come from your activities from this collaborative phase.

You will be able to **choose from two to five problems** coming from scientific or industrial applications (e.g., brain imaging, astrophysics, biology/chemistry, ad placement, insurance pricing). You can participate in more than one challenges: we will **grade you based on your best performance**.

The starting kit

Each challenge will come with an **open source starting kit** available at https://github.com/ramp-kits, containing

- a **Jupyter notebook** that describes the industrial or scientific prediction problem, does some exploratory analysis and data visualization, and explains the predictive workflow and one or more basic solutions (example),
- a **Python script that parametrizes the challenge** (example),
- public **training and test data** sets, different from the official sets we use to evaluate your submissions at the RAMP site (example), and
- one ore more example submissions (example).

You will be able to test the example submission and all your subsequent submissions before submitting, using a simple command line test script (more information here).

Timeline

- Opening of the challenges: October 27, 12h30-14h30
- Closing of the competitive phase: December 17, 20h
- Closing of the collaborative phase: January 30, 20h

Evaluation

Half of your grade will come from the data challenge, **8/20 from the competitive phase and 2/20 from the collaborative phase**. Selected students will be able to present their solutions to the class, for up to 2/20 bonus grades.

Team project

To achieve the second objective, you will **build a predictive workflow in teams of size three to five**, implement a **data-driven business/science case**. We will give you a set of pointers to existing data sources, but you will be **encouraged to find business/science cases** and data sources on your own. Collaborating with research teams or businesses will also be highly regarded.
You will **submit the projects as RAMP starting kits on github**. We will not ask you here to optimize the solution, rather to **focus on its design** and its match to the business or science case.

The business/science case

Half of your grade (5/10) will come from the **quality of the predictive business or science** case that you will present in the preamble of the Jupyter notebook of your starting kit. The following questions are to guide you in this exercise:

- **What do we want to predict?** How will a good prediction **improve** a key performance indicator (**KPI**) or lead to a scientific result?
- How do we **measure the quality or value of the prediction** in the selected business or science problem? What will be the quantitative score? How does the **quantitative score reflect the quality or value of the prediction**? How does the (possibly asymmetric) prediction error convert into cost or decreased KPI?
- Will the predictor be used as **decision support**, as a part of a **fully automated system**, or only as part of a report or **feasibility study**? **How will an agent use the system**?
- **What data** do we need to develop a predictor? Could you find this data? What were the actual data sources? What other sources (private or public) could be exploited? What were and would be the data collection costs?
- What data **cleaning/tidying steps** were required to obtain clean training data?
- Given the data source(s) and the prediction goal, what is the **workflow and the workflow elements**? Will you need different expertise for the different steps?
- How fast the phenomena underlying the prediction problem change? **How often the model will have to be retrained**? What are the associated costs and risks?

The technical quality

The second half of your grade (5/10) will come from the technical quality of your solution. You kit will have to **pass the ramp_test_submission test**. We will pay close attention to your **validation setup** (Is the validation reasonable? Do you have enough test data to see significant differences between submissions?). We will also grade the **quality of the exploratory analysis** and the **clarity of the technical explanation** of the workflow.

Timeline

- December 18: You should arrive to the data camp week prepared, **having formed teams** and having an approximate **idea about the business/science problem** you would like to tackle and the potential **data sources**.
- December 18 - 22: **The data camp week**, Ecole Polytechnique, Amphi Faure (9h - 17h). We will have **lectures in the morning and guided work and student presentations in the afternoon** (students presenting their solutions will get up to 2 bonus points). The tentative program:
    - Monday: the data science ecosystem, a case study, how to build a data science workflow

    - Tuesday: advanced pandas tutorial, data transformations, tidying data
    - Wednesday: handling categorical features, feature engineering
    - Thursday: the ramp-workflow library, building a workflow, examples, classical regression/classification, feature extraction, time series, multi-objective workflows
    - Friday: wrap-up
- January 30 20h: deadline of submitting the projects.

Prerequisites

The course will require that you develop code in Python. **We strongly suggest that you start preparing.** You should have a complete Python environment setup on your machine on the first day of the course. We recommend to use Anaconda (https://www.continuum.io/downloads). It includes all required libraries. Here are some necessary resources: numpy, pandas, scikit-learn, xarray. You might want to also install: seaborn, hyperopt, and xgboost. Some of the challenges strongly favor deep learning solutions; we will allow submissions both in pytorch and in keras (with tensorflow backend).

The scikit-learn web site is also a great resource to brush up on your ML skills. The following tutorials are recommended to learn more about pandas and scikit-learn:

http://scikit-learn.org/stable/tutorial/basic/tutorial.html
https://github.com/amueller/scipy-2016-sklearn
http://pandas.pydata.org/pandas-docs/stable/tutorials.html

The slack forum

During the challenge and the data camp we will be communicating through slack. The workspace URL is:

https://join.slack.com/t/datacamp2017/signup

You should be able to register if your emails ends with:

- telecom-paristech.fr
- ensae-paristech.fr
- polytechnique.edu
- supelec.fr
- ensae.fr
- edu.ece.fr
- ens-lyon.fr
- ensta-paristech.fr
- u-psud.fr
- eleves.enpc.fr*

* Contact us if you cannot register.

You can also use it for **communicating within and between teams**.

Data sources

The following is a list of data sources that you may use in your team project. Note however that picking a nice data set and setting up a prediction problem is not enough for a good grade: you also have to make a reasonable business or science case.

- Velib data spatial time series.
- RATP data.
- Amazon data sets.
- Kaggle data sets.