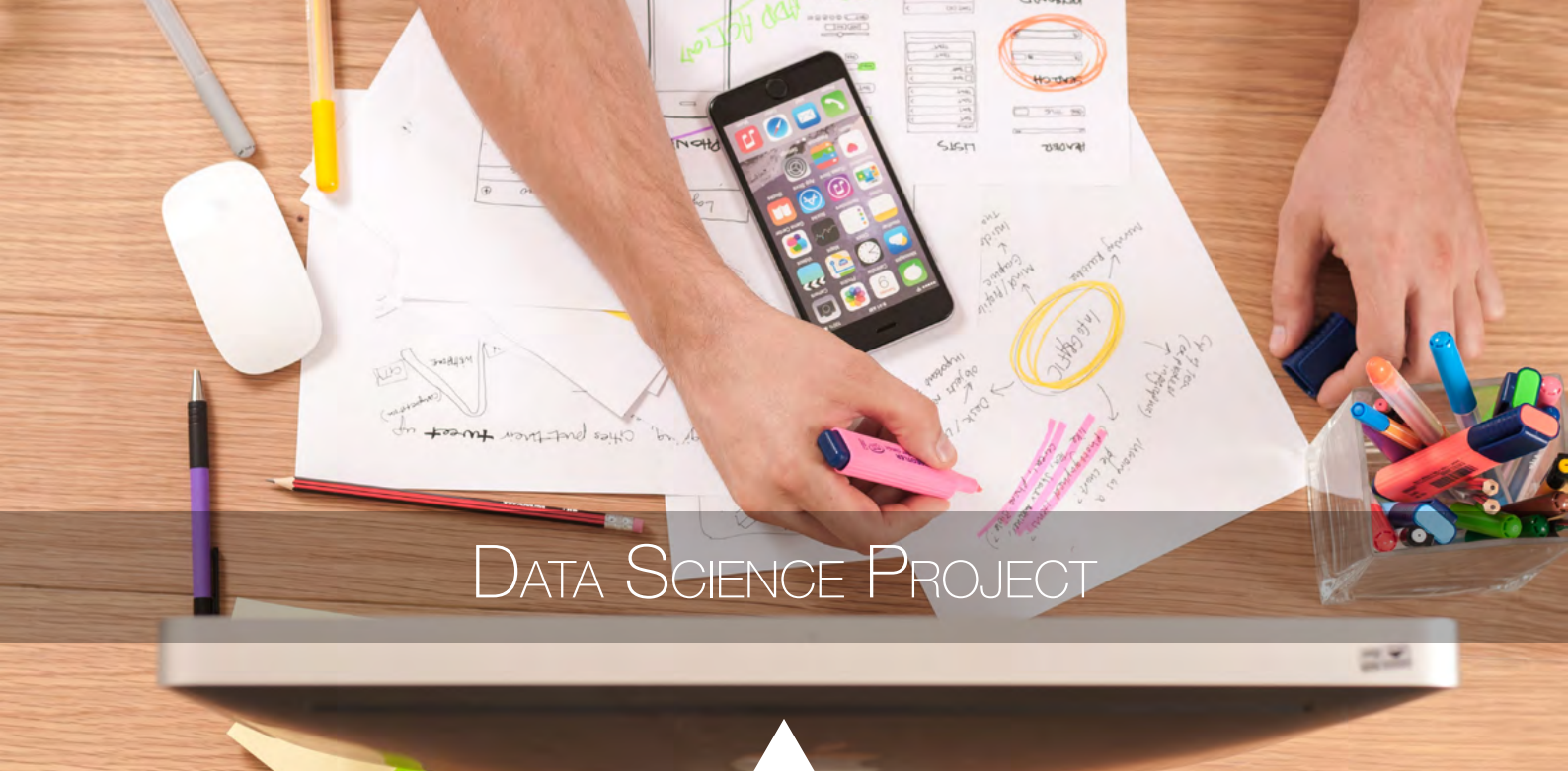

Data Science Project

Data Rocket Science Project Goal

The project's aims is to explore a question regarding a given topic. The topics for this session includes:

- Game of Thrones Books
- Shanghai extract of dianping.com
- Physical sensing with datacanvas.org data
- Cities in Litterature with google Ngrams or Gutemberg

You will choose one of these topics and explore it. You will use the data science techniques introduced in the python course to open, extract, filter and modelize relevant data. Following various hypothesis you will explain, you will provide visualizations, lists and statistics indicators to prove or disprove it. You will also foresee how to make the analysis better and account for it in both a presentation type format and a blog post. The expected total amount of work is 40h and can be done by up to 2 person.

Project Evaluation

To be evaluate you need to provide a zip file of a directory which name is your ***firstname-lastname*** and containing :

- a sub-directory ***scripts/*** containing
  - all your scripts and data files if any
  - a ***authors.txt*** file with the two firstname and lastname of your pair
- a sub directory ***presentation/*** containing
  - a presentation file in the **PDF format** containing exactly 10 slides divided as follow
    - 1 front slide
    - 6 slides to explain and expose your analysis results
    - 2 slides to explain how one could go further in the exploration
    - 1 slide to present a potential business idea based on these data/analysis
- a sub directory ***blog/*** containing
  - A 1500 words blog post style article written in markdown format describing your exploration, choices, hypothesis and results. No need to explain what tools you used extensively, except if you used a tool not covered in the intensive training or that is very hard to figure out.
  - All necessary illustrations in high definition (1600+ px wide, pdf format / jpeg / png lossless compression)

The zip file is to be sent to *fabien.pfaender@me.com, jessica.kohler@utseus.com*
with the title 'Data Science LCI'
If the zip file is too big, you can filex or wetransfer.com or any other way you can think of.

Project Deadline - Sunday 5th of June 0:00pm

See page 2 for details

Project Description

Load a CSV file in python

Loading a csv file a a basic action library included in every python distribution (including Anaconda): https://docs.python.org/2/library/csv.html

Classification

To perform a regression task you will find here a complete example of a linear regression done in python using librairy scikit-learn : http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#example-linear-model-plot-ols-py

Clustering

To perform a clustering in a dataset using an implementation of the K-Mean algorith from the scikit-learn, you will find here an example of a simple use of this algorithm using the library: http://scikit-learn.org/stable/tutorial/statistical_inference/unsupervised_learning.html. .

Help

Google and stackoverfow all full with a huge panel of examples you can learn from. Use this to get some inspiration of how to present and better your result.

Visualization

All examples usually use matplotlib library in python to draw plots. You can also use cartodb.com to create a map visualisation of the data by sending a file in CSV format. D3.js provide an interactive way to explore data and Gephi can visualize networks.
Excel software is also a good tool to visualize simple distribution. Excel can import csv files though the [file -> import] menu. You can generate csv from any array using python with the link provided above.