Data Science*

* : An Introduction to Data Science

Julien Roland

Data Scientist *- Person who is better at statistics than any software engineer and better at software engineering than any statistician. (Josh Wills)*

Data

*"**information**, especially **facts** or **numbers**, collected to be examined and considered and used **to help decision-making**, or information in an electronic form that can be stored and used by a computer"*

(Definition of data from the Cambridge Advanced Learner's Dictionary & Thesaurus © Cambridge University Press)

Data Science

"The scientific study of the creation, validation and transformation of data to create meaning."

See, http://www.datascienceassn.org/code-of-conduct.html.

Data Science

Extract **knowledge/insight** from data to :

- **create or improve** tools, products,...
- **aid decision making**
- **...**

Data Science ?

Isn't Statistics the Science of Data ? What's new ?

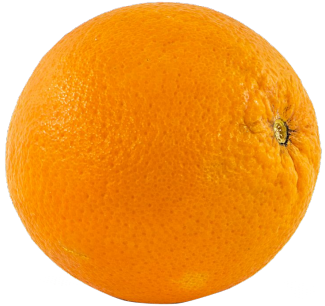
- A massive amount of data (often available in real time) from various sources :
  - Web, mobile devices, sensors,...
- Unstructured data
- Large computing resources are available on demand and at low cost
- The emergence of new usages of data :
  - Recommendation systems (film, music, jobs, products,...)
  - Trading algorithms
  - ...

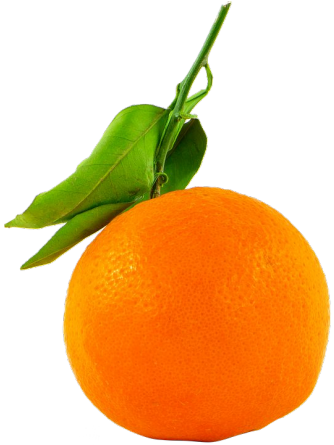
Skills

Beside traditional statistical tools, data science requires **to use and combine techniques** from various related fields such as :

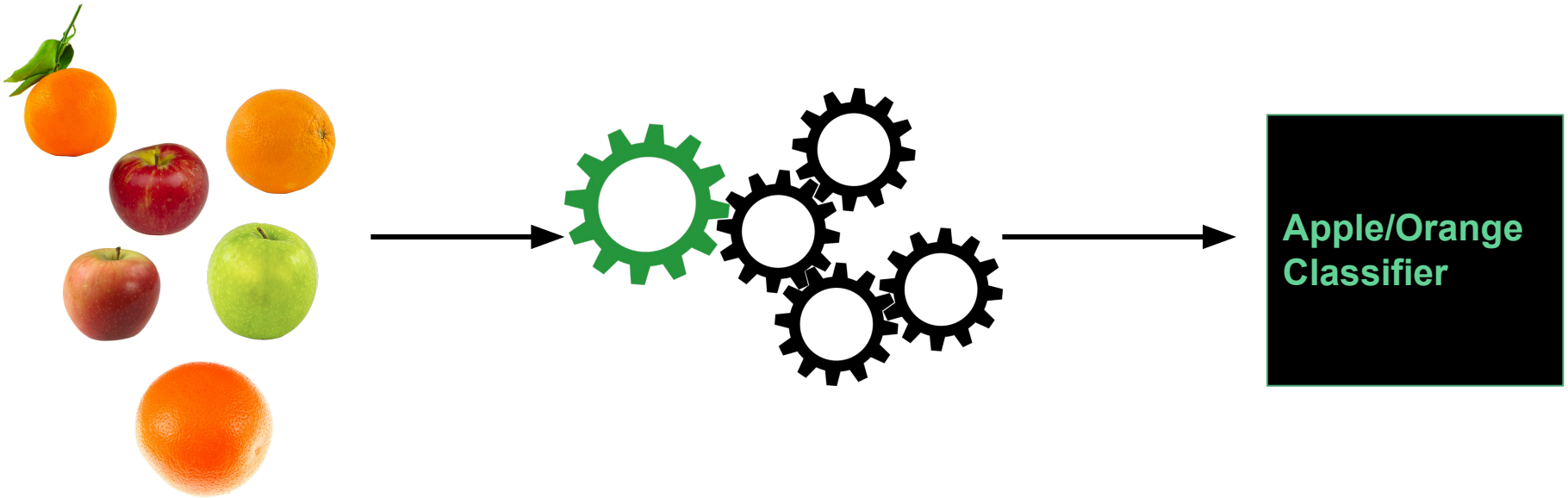
- Visualization
- Machine Learning
- Operations Research
- Database and Storage
- Programming
- Parallel Programming
- Distributed systems
- ...

Example of data : Pictures





Example of tool : A classifier



Training Set **Learning Algorithm** **Model**

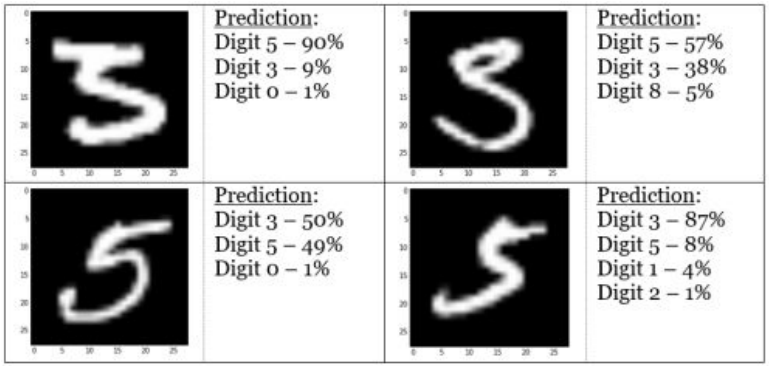
Example of tool : A classifier

Example of tool : A classifier

???

Other usages of classifier

- Recognizing hand-written digits
- Spam filtering
- Computer security : intrusion detection
- Medical diagnosis
- Products recommendation
- ...



Course Organization

Technologies

Technologies, languages, and tools used in this course:

- **NumPy** : n-dimensional array object,
- **SciPy** : scientific computing, including numerical integration and optimization
- **matplotlib** : 2D plotting library
- **Jupyter Notebook** : a web application to create documents that contain live code, equations, visualizations and explanatory text
- **Pandas**: data structures and data analysis
- **scikit-learn** : data mining and data analysis

Why Python ?

- Easy to read, learn, and use

```
print('Hello, world!')
fruits = ['apple', 'orange', 'pear', 'mangos']
for name in fruits:
    print(name)
```

- A Python API is almost always available (Spark, Hadoop, TensorFlow,...)
- Easier to integrate with broader applications
- Is not only useful for statistics
  - System administration,
  - Web programming,...

Course Outline

1. An introduction to Python
2. Data acquisition, Manipulation, and Visualisation
3. Graph and Network Analysis
4. Probability and Statistics with Python
5. Machine Learning Algorithms
6. *Project*

Course Grading

Use the weighted geometric mean with the following weights :

- Project : weight = 2
- Exam : weight = 1
  - About Lectures, Labs, and Projects
  - Multiple Choice and Fill-in-the-Blank

See, https://en.wikipedia.org/wiki/Weighted_geometric_mean.

Course Schedule

- 12/09 : Introduction to Python
  - + NYT Books: Finding the popular best-sellers (HTTP Requests, JSON, XML, XPath,...)
- 19/09 : Introduction to Statistical Learning + Introduction to Numpy and Pandas
- 26/09 : Exploratory Data Analysis
- 03/10 : Classification
- 17/10 : Clustering
- 24/10 : *Free Session*
- 07/11 : Presentations
- 14/11 : Presentations

Course Project

Course Project

- Groups composed of 3 students
- Research paper, book chapter, or white paper
- Illustrative data set
- Python Notebook (in english)
- 20 minutes + 10 minutes for questions

List of Possible Topics

A maximum of 2 groups for each topic :

1. Dynamic (real-time) pricing
2. Recommendation systems
3. Multi-criteria analysis
4. Frequent pattern mining
5. Outlier analysis
6. Spam filtering
7. Data wrangling
8. Mining social networks
9. Deep Learning
9. Malware detection
10. Account hijacking detection
11. Click fraud detection
12. DOS detection
13. (Network) Intrusion detection
14. Weather forecaster
15. Travel time prediction
16. TensorFlow, SparkML, GraphX,...
17. Geographical Information Systems
18. Algorithmic Trading

Data Sets

Examples of datasets :

- RTA Freeway Travel Time Prediction
- KDD Cup 1999 : Computer network intrusion detection
- Movielens, Movietweetings,...

See, for example,

- archive.ics.uci.edu/ml/index.html,
- www.kaggle.com/datasets,
- www.recsyswiki.com/wiki/Category:Dataset,
- http://snap.stanford.edu/data/.

Schedule

- Monday **19/09 (8:00 a.m.)** : Project proposals
  - By e-mail to julien.roland@isen-lille.fr
  - Data set
  - At least one related research paper, book chapter, or white paper
  - Group composition
- Project Progress Meetings (~5 minutes per group)
  - 3/10
  - 24/10
- 7/11 : Presentations
- 14/11 : Presentations

References

References

- **The Elements of Statistical Learning**: Data Mining, Inference, and Prediction, by Trevor Hastie and Robert Tibshirani, Springer, 2009.
- **Data Mining and Analysis**: Fundamental Concepts and Algorithms, by Mohammed J. Zaki and Wagner Meira, Jr., Cambridge University Press, 2014.
- **Doing Data Science**: Straight Talk from the Frontline, by Cathy O'Neil and Rachel Schutt, O'Reilly Media, 2013.
- **Python for Data Analysis**: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012