Python for Data Science

SoBigData

Giulio Rossetti, KDD Lab ISTI-CNR
giulio.rossetti@gmail.com @GiulioRossetti

How many people here know a programming language?

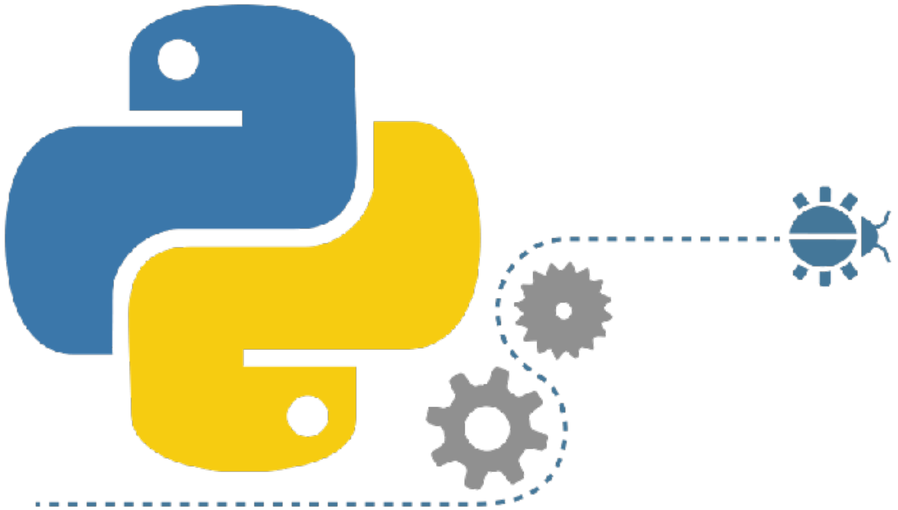
How many people here know a programming language?

How many people here have used Python?

About Python

Python is an interpreted high-level programming language

- Named after Monty Python
- Created by Guido Van Rossum
- First released in 1991
- Support multiple programming paradigms
  (imperative, functional, OO, procedural)
- Coincise and elegant syntax
- Designed to be *“easy to learn”*
- Latest release 3.6
  (support to Python 2.7 will be discontinued in 2020)



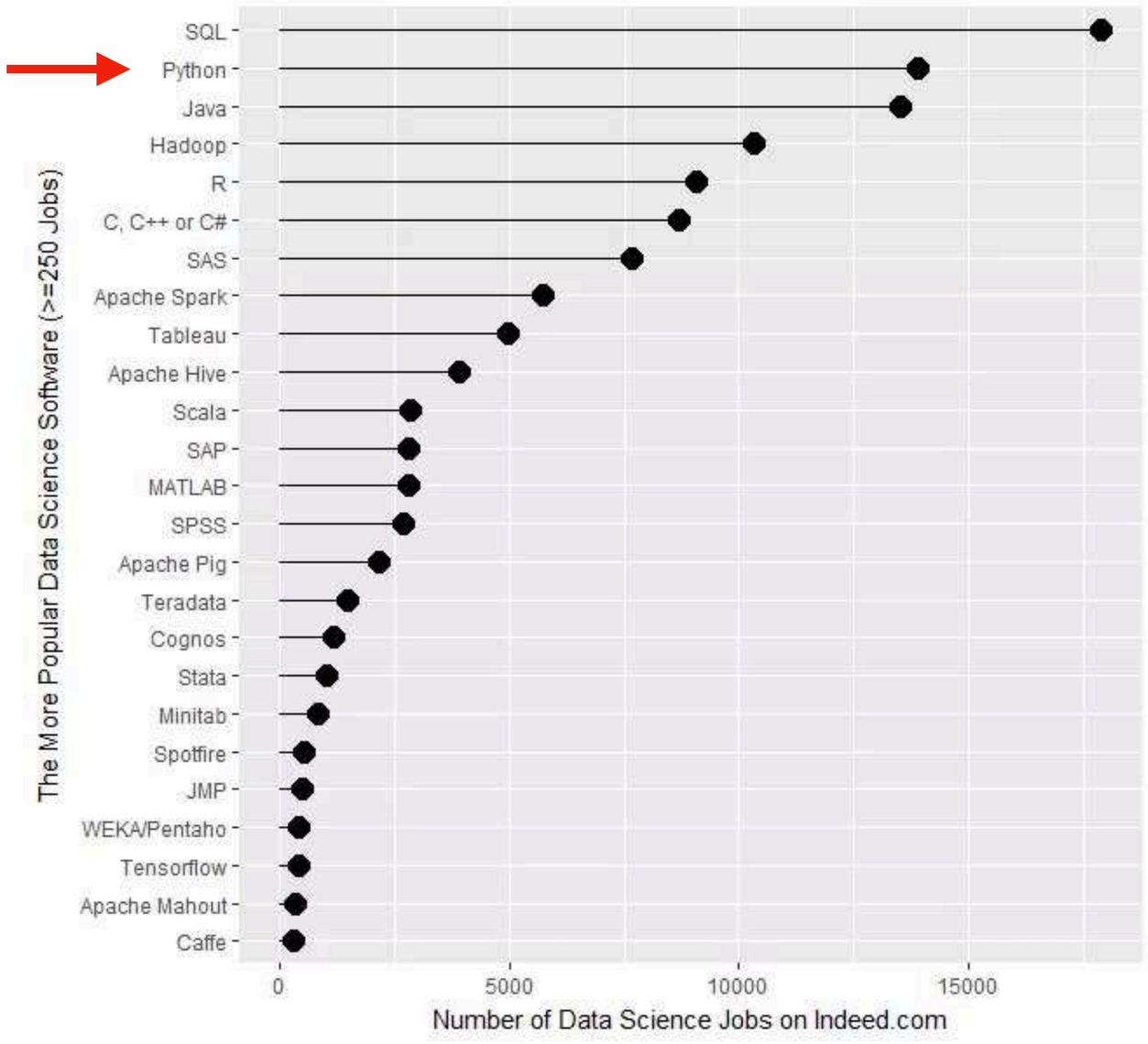
Zen of Python

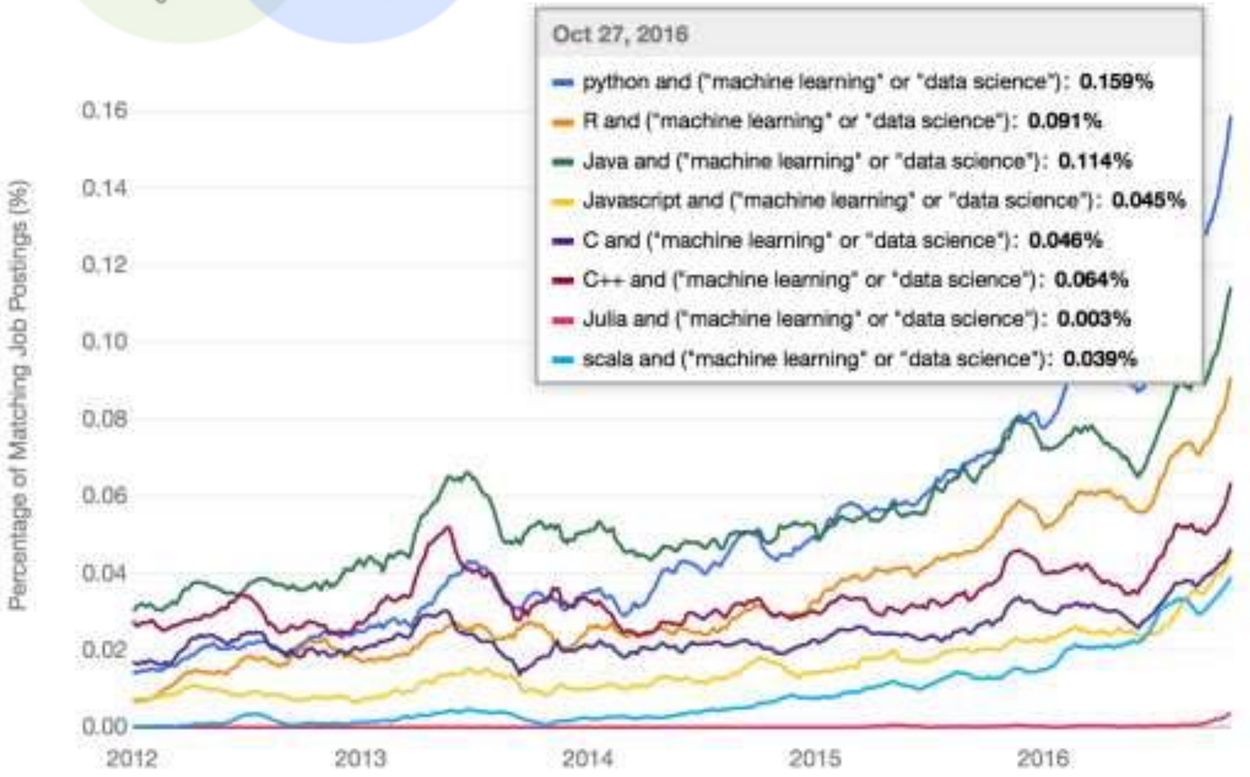
Beautiful is better than ugly
Explicit is better than implicit
Simple is better than complex
Complex is better than complicated
Readability counts
…

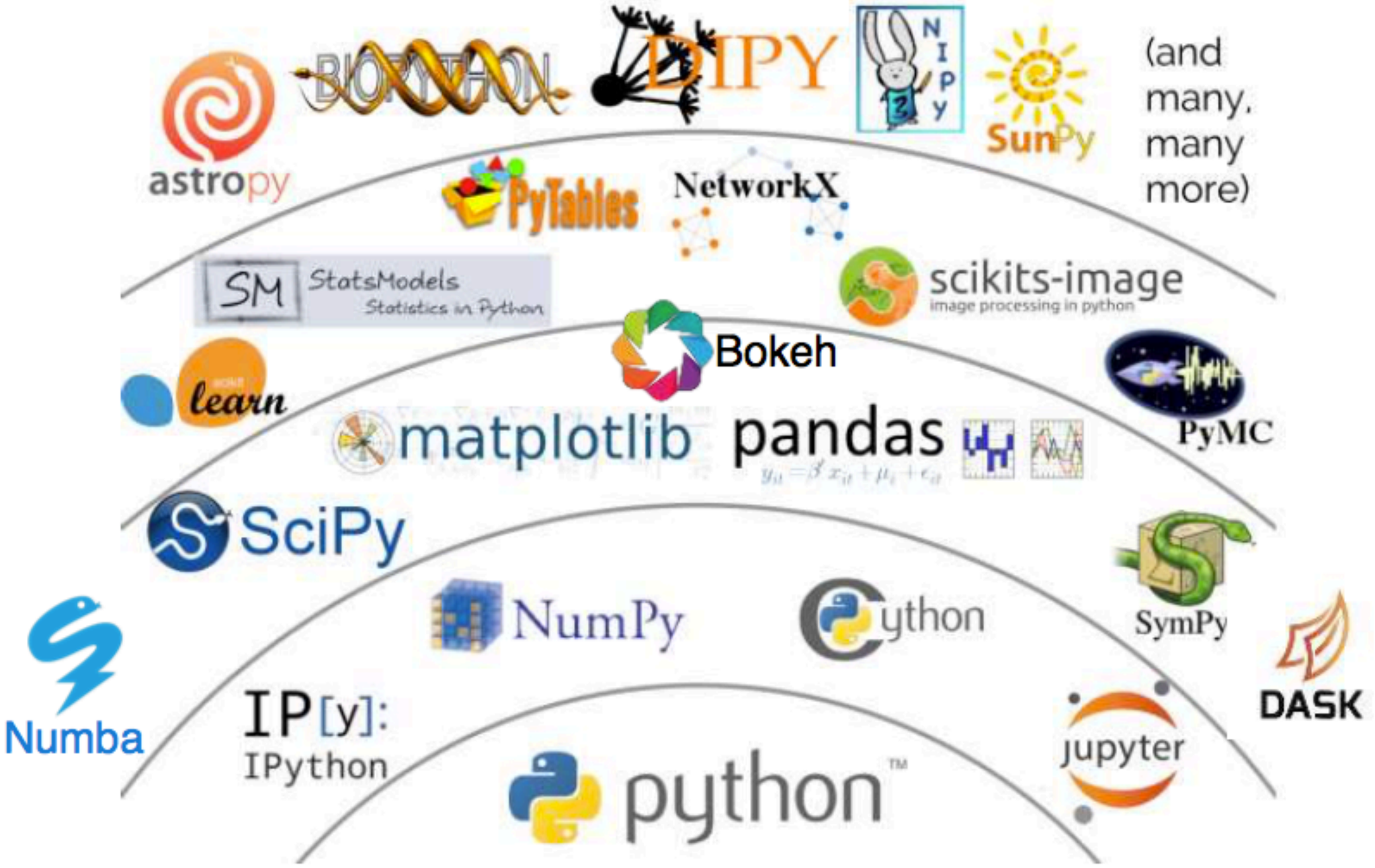
Python is a general-purpose language,
it can be used to build just about anything!

Documentation, tutorials available at:
www.python.org

Data Science is growing with Python leading the way






astropy
BIOPYTHON
IPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM StatsModels Statistics in Python
scikits-image image processing in python
Bokeh
learn
matplotlib
pandas
PyMC
SciPy
NumPy
ython
SymPy
DASK
Numba
IP[y]: IPython
jupyter
python™


astropy
BIOPYTHON
IPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM StatsModels Statistics in Python
scikits-image
image processing in python
Bokeh
learn
matplotlib
pandas
PyMC
SciPy
NumPy
Cython
SymPy
DASK
Numba
IP[y]: IPython
jupyter
python™


astropy
BIOPYTHON
IPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM StatsModels Statistics in Python
scikits-image
image processing in python
Bokeh
learn
matplotlib
pandas
PyMC
SciPy
NumPy
ython
SymPy
DASK
Numba
IP[y]: IPython
python
jupyter


astropy
BIOPYTHON
IPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM StatsModels Statistics in Python
scikits-image
image processing in python
Bokeh
learn
matplotlib
pandas
PyMC
SciPy
NumPy
ython
SymPy
DASK
Numba
IP[y]: IPython
python
jupyter


astropy
BIOPYTHON
IPY
NIPY
SunPy
(and many, many more)
PyTables
NetworkX
SM StatsModels Statistics in Python
scikits-image
image processing in python
learn
Bokeh
matplotlib
pandas
PyMC
SciPy
NumPy
ython
SymPy
DASK
Numba
IP[y]: IPython
jupyter
python™

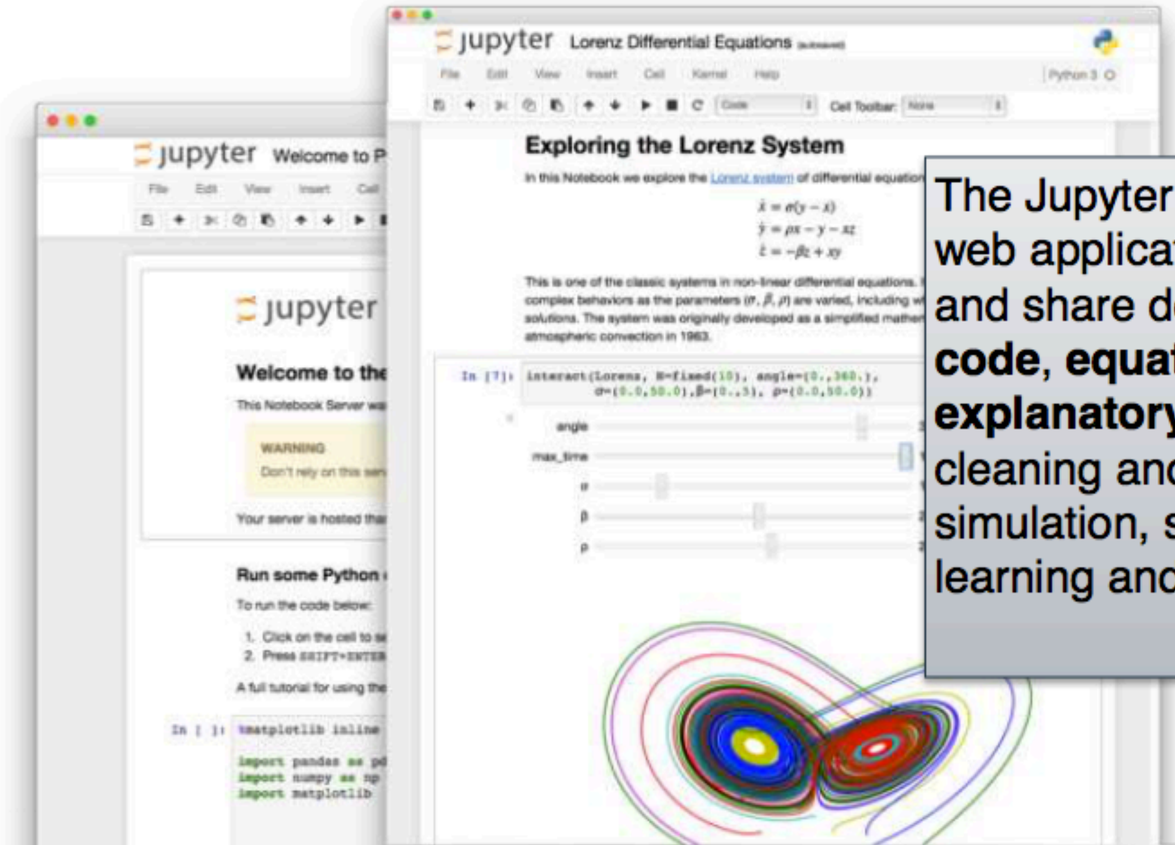
Course organisation and materials

This course will be completely hands-on!

To do so we will use Jupyter notebooks
(available within Anaconda)

Notebooks allow to:

- *Interactively* execute Python code
- Visualise results *step-by-step*



The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live **code**, **equations**, **visualizations** and **explanatory text**. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

Course Outline

The course is organised in 3 sessions

Course Outline

The course is organised in 3 sessions

Session 1: Python Basics

Learn the basic concepts behind modern programming languages. We'll talk about:

- Variables & data structures
- Flow control
- I/O and exceptions

Course Outline

The course is organised in 3 sessions

Session 1: Python Basics

Learn the basic concepts behind modern programming languages. We'll talk about:

- Variables & data structures
- Flow control
- I/O and exceptions

Session 2: Data manipulation in Pandas

Learn how to load, access and transform datasets. We'll talk about:

- DataFrames
- Slicing, Indexing, Grouping…
- …

Course Outline

The course is organised in 3 sessions

Session 1: Python Basics

Learn the basic concepts behind modern programming languages. We'll talk about:

- Variables & data structures
- Flow control
- I/O and exceptions

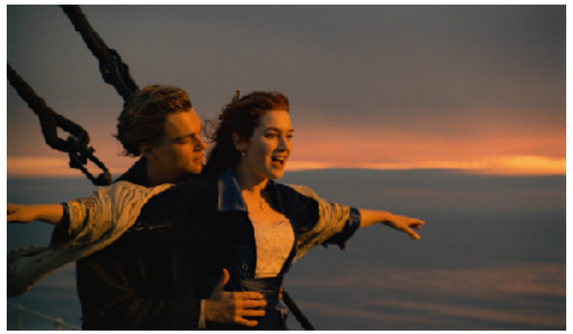
Session 2: Data manipulation in Pandas

Learn how to load, access and transform datasets. We'll talk about:

- DataFrames
- Slicing, Indexing, Grouping…
- …

Session 3: Case Study

Who is more likely to survive to the infamous Titanic shipwreck?



Let’s begin!

Install **Anaconda**
(Python 3 version!)
https://www.anaconda.com/download/

Anaconda is a python distribution for data scientists.

It provides out-of-the box support for:

- Data modeling
- Analysis
- Visualisation
- …

Python Pandas

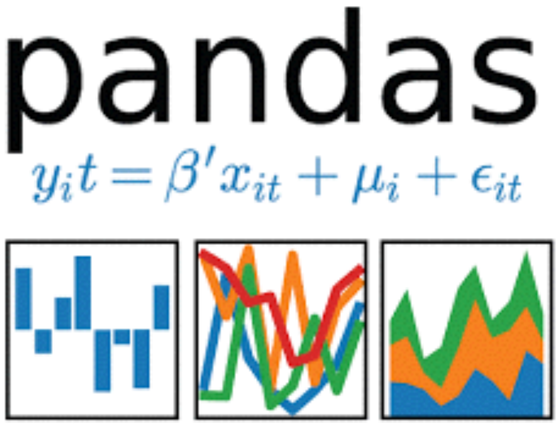
SoBigData

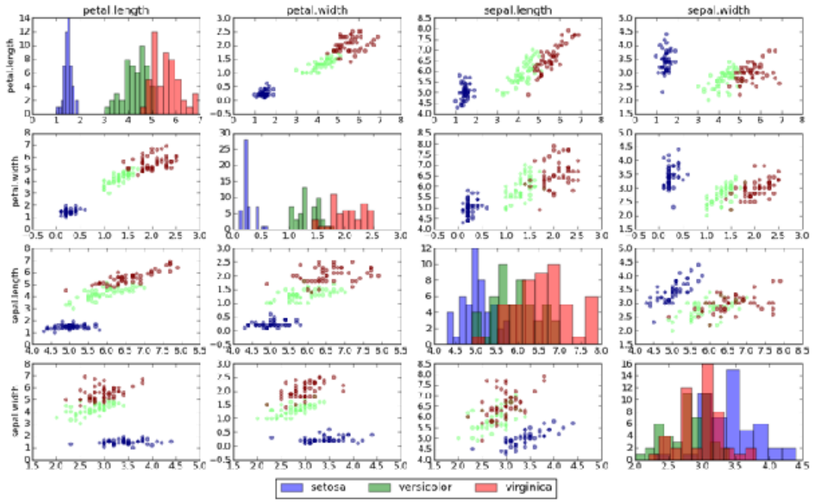
Giulio Rossetti, KDD Lab ISTI-CNR
giulio.rossetti@gmail.com @GiulioRossetti

Pandas?

Pandas is a Python package designed to simplify data-wrangling

- http://pandas.pydata.org
- R's data.frame for Python
- Excellent performance
- It supports:
  - Data preparation
  - Data transformation
  - Data cleaning
  - Data visualisation





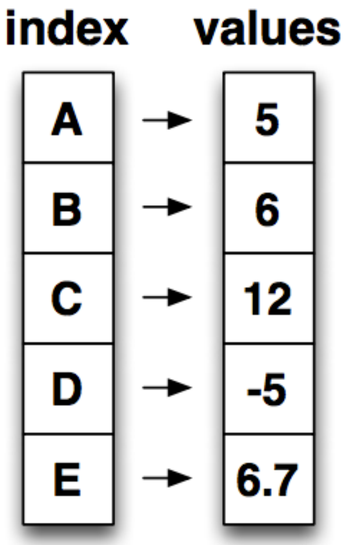
	sepal-length	sepal-width	petal-length	petal-width	class_
50	7.0	3.2	4.7	1.4	Iris-versicolor
102	7.1	3.0	5.9	2.1	Iris-virginica
105	7.6	3.0	6.6	2.1	Iris-virginica
107	7.3	2.9	6.3	1.8	Iris-virginica
109	7.2	3.6	6.1	2.5	Iris-virginica
117	7.7	3.8	6.7	2.2	Iris-virginica
118	7.7	2.6	6.9	2.3	Iris-virginica
122	7.7	2.8	6.7	2.0	Iris-virginica
125	7.2	3.2	6.0	1.8	Iris-virginica
129	7.2	3.0	5.8	1.6	Iris-virginica
130	7.4	2.8	6.1	1.9	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica
135	7.7	3.0	6.1	2.3	Iris-virginica

Data Structures

Pandas allows to model data as Series and DataFrames

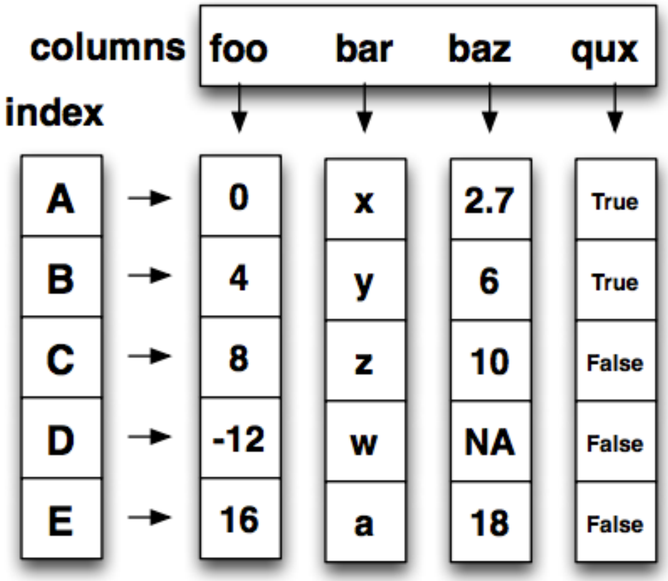
Series

- Indexed arrays
- Index labels need not be ordered



DataFrame

- 2-dimensional arrays
- Row and Column index
- Each column can have different type
- Mutable sizes (insert/delete)



Case Study
Titanic Shipwreck

SoBigData

Giulio Rossetti, KDD Lab ISTI-CNR
giulio.rossetti@gmail.com @GiulioRossetti

Titanic dataset: a Kaggle Competition

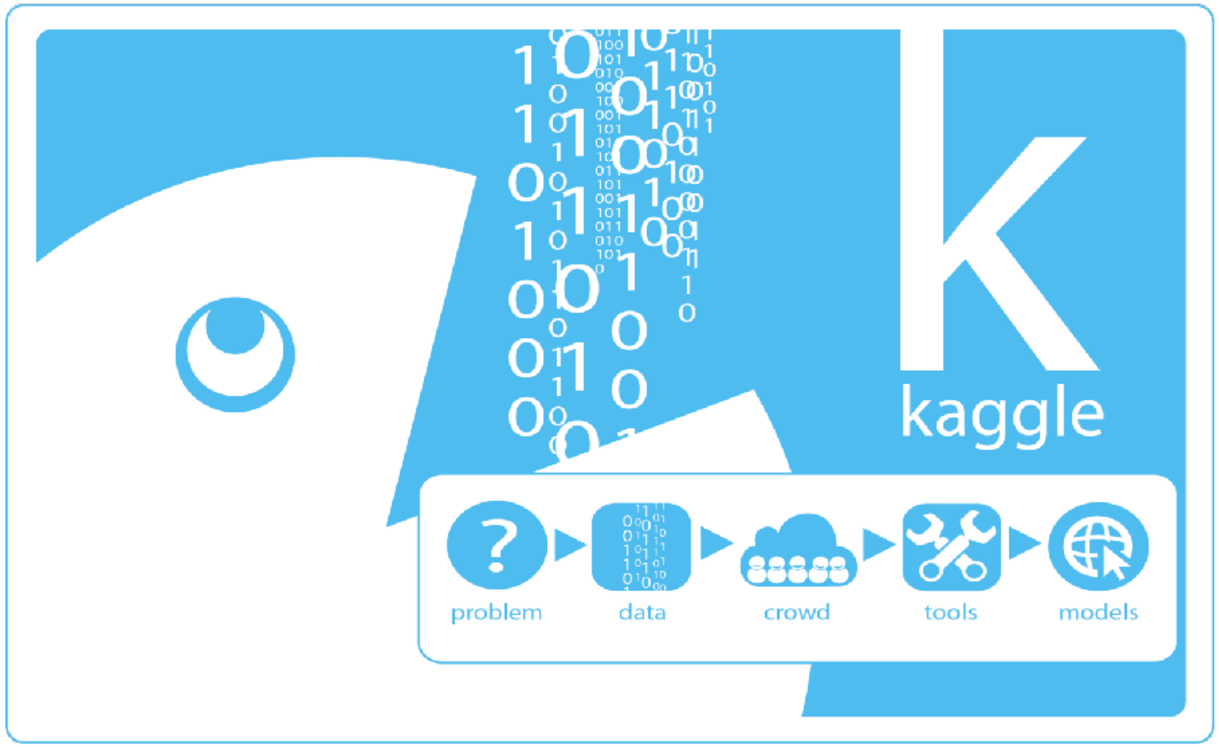
Kaggle is a platform for predictive modelling and analytics competitions

- http://kaggle.com

The Titanic challenge requires to:

- Analyse embarking record data
- Predict who will survive to the infamous shipwreck

Using Python and Pandas we will try to address such challenge with a (simple) approach.



https://www.kaggle.com/c/titanic