Introduction to Machine Learning/Algoritmi Avanzati

Data Science Toolkit, R and Python

Manuel Dalcastagnè
m.dalcastagne@unitn.it

University of Trento

18 February 2019

Section 1

1. Some logistics

Please register to the mailing list of the course:

- Go to http://bit.ly/introml2019
- Login with your UniTN account
- Insert name, last name, course degree and e-mail

Download Anaconda (3.7) at https://www.anaconda.com/distribution/

Book of the course:

- The LION way. Machine Learning plus Intelligent Optimization. Version 3.0. Roberto Battiti and Mauro Brunato. LIONlab, University of Trento, 2017.
- Download PDF at https://intelligent-optimization.org/LIONbook/

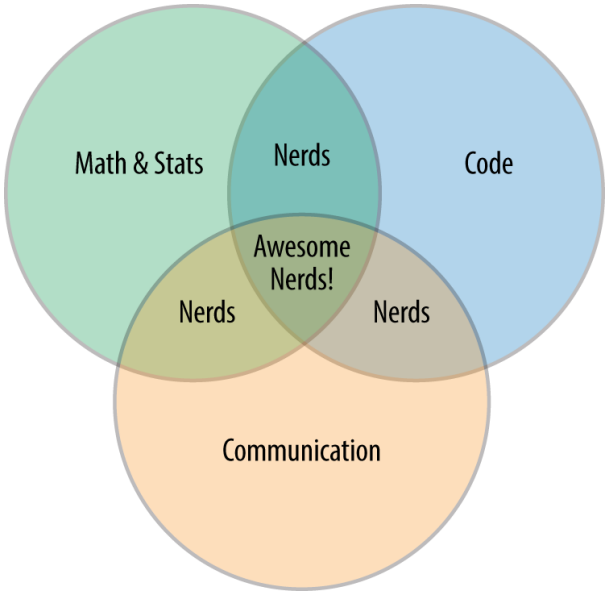
1. Objectives of the course

- Read, clean, normalize and encode raw data
- Select and filter data features
- Use, configure and evaluate different machine learning algorithms
- Compare different machine learning algorithms, analyze results and predictions
- Plot information about results of the analysis
- Create a pipeline to automatize the aforementioned process

Complete syllabus can be found on Esse3

1. Data scientist

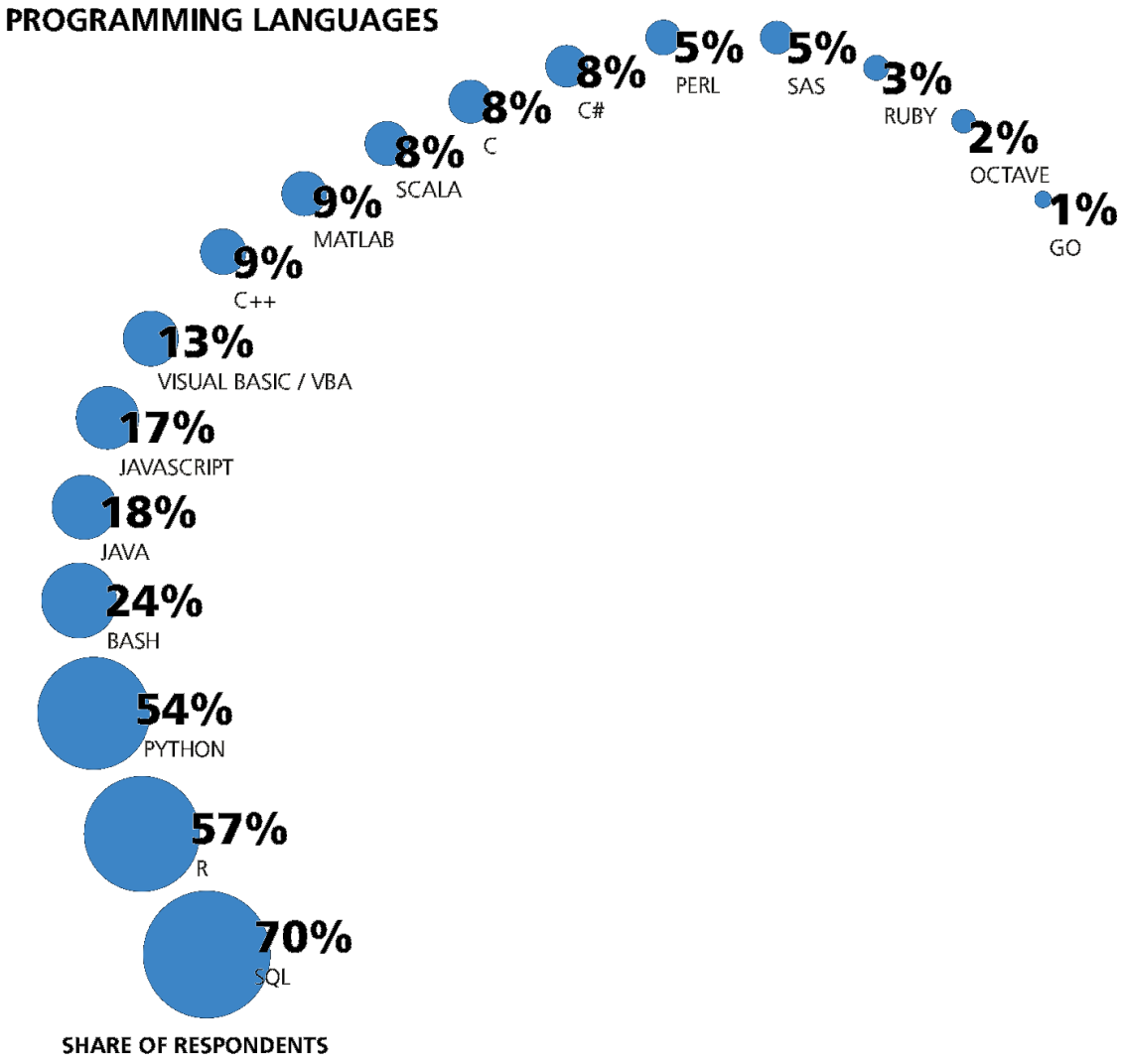
Data scientists? From data magicians to awesome nerds!



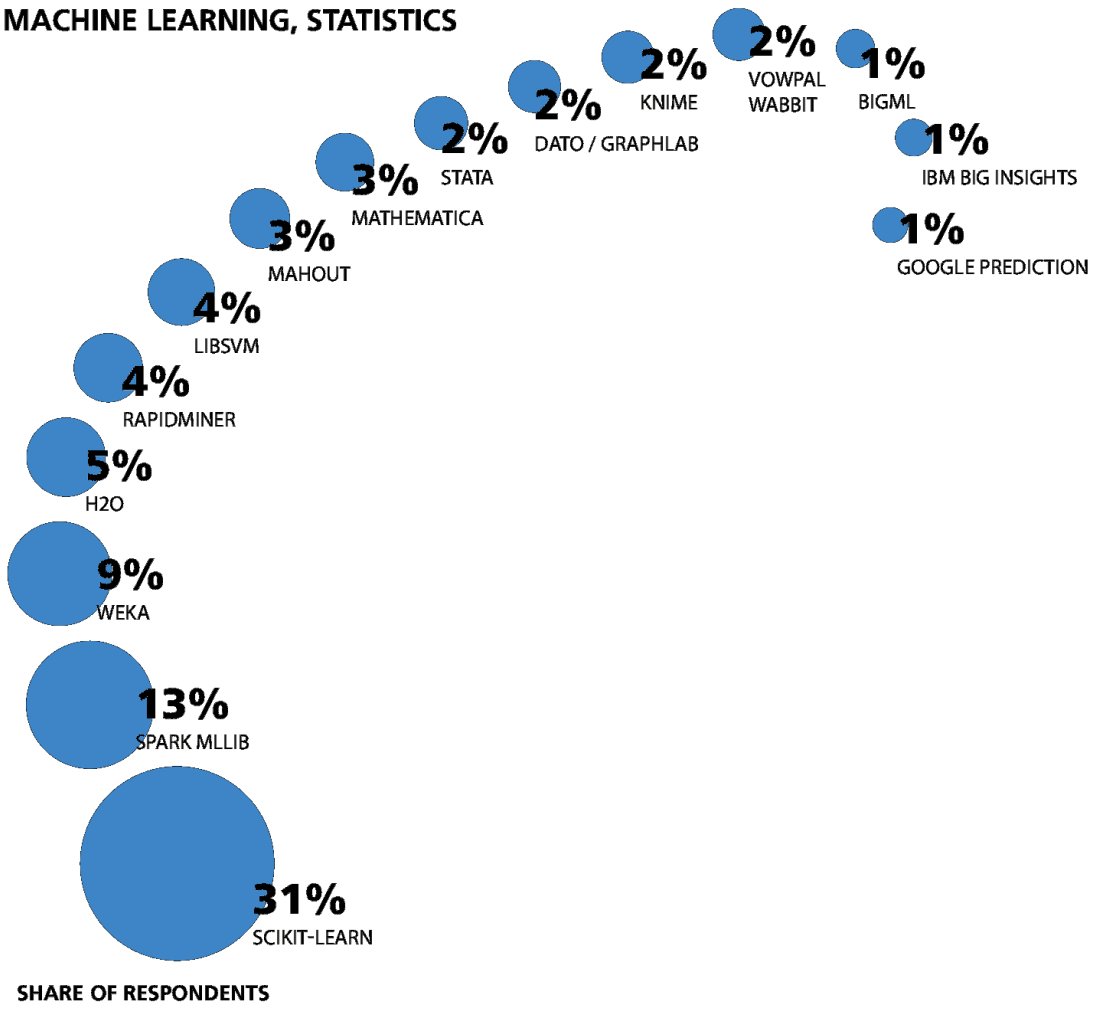
Our definition of a data scientist

A professional who extracts knowledge and builds models from raw data, in order to connect insight to better decisions.

1. Data science tools, programming languages [1]



1. Data science tools, machine learning libraries [1]



1. R vs Python

- Currently, the two most popular programming languages for data science are Python and R
- Both are free and open source
- Python is more general-purpose than R (developed for statistical analysis)
- R is less intuitive and more formal than Python
- Python has a more complete set of machine learning packages
- R has better exploratory and data visualization packages
- Both are easy to install (Anaconda)

Section 2

2. Anaconda info

- Anaconda is a Python/R distribution, a package manager and an environment manager
- it works on Linux, Windows, and Mac OS X
- 11 million users worldwide
- Manage libraries, dependencies, and environments (1,500+ Python/R data science packages)
- Download Anaconda (Python 3.7 version) at https://www.anaconda.com/distribution/

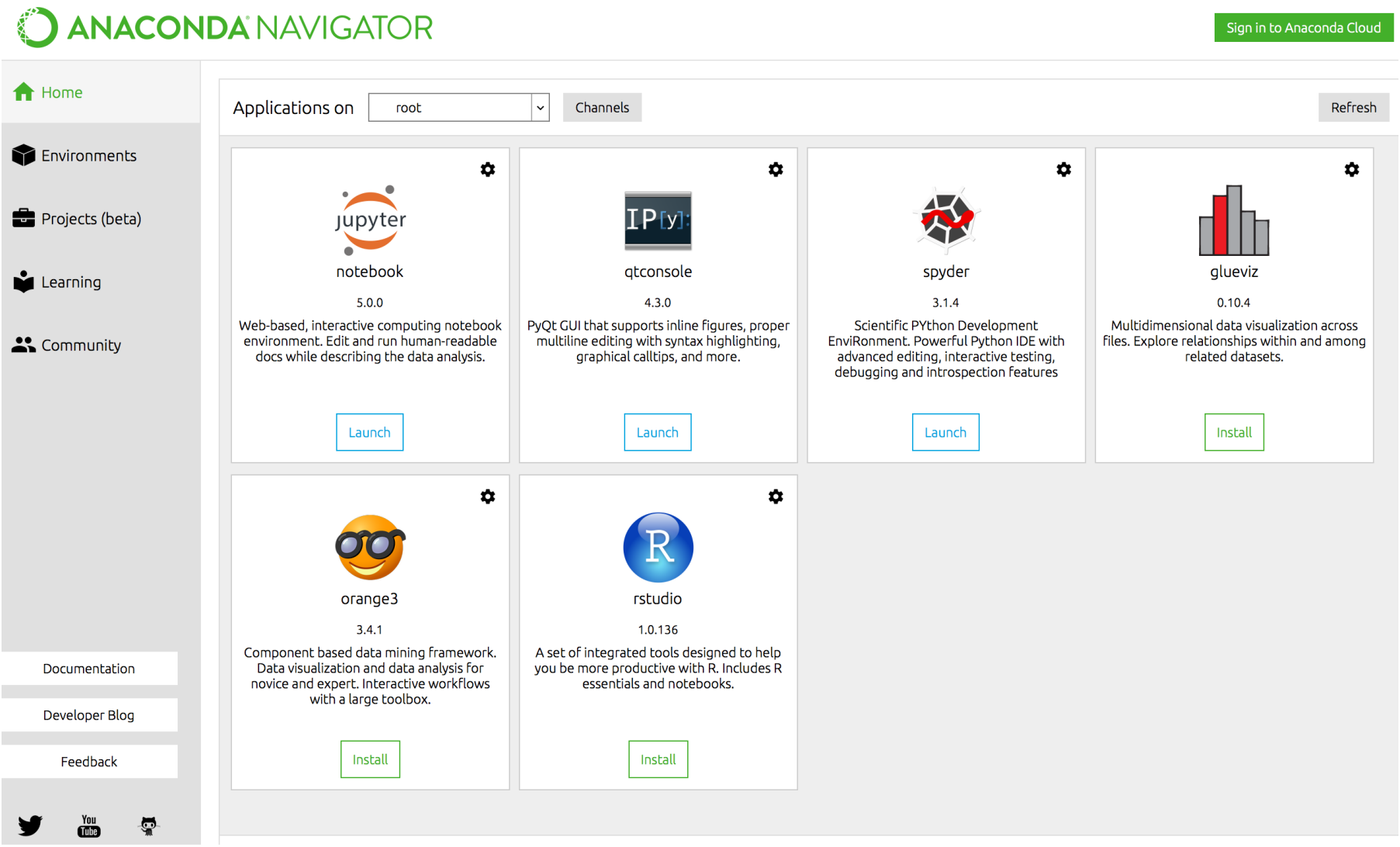
2. Anaconda docs and installation

- Anaconda documentation can be found at https://docs.anaconda.com/anaconda/
- Install instructions and system requirements can be found at https://docs.anaconda.com/anaconda/install/
- To install Anaconda on Linux:
  - Open the terminal and go to the folder where the Anaconda installer has been downloaded
  - Run the following command to start the installation: *bash Anaconda3-5.3.0-Linux-x861_64.sh*
  - Follow the installation process: review and accept license agreement, accept default install location, accept to prepend the install location to your PATH variable, but DO NOT install VS Code
  - Close the terminal and open it again

2. Anaconda Navigator

- Anaconda Navigator is a desktop graphical user interface (GUI) that allows you to:
  - launch programs included in the Anaconda distribution
  - manage independent programming environments and software packages (without using the terminal)
  - access quickly the documentation of main data science packages
- To open Anaconda Navigator, on linux execute the command *anaconda-navigator*
- We are going to use the following Anaconda programs:
  - Spyder Integrated Development Environment (IDE)
  - Jupiter Notebooks

2. Anaconda navigator



2. Spyder

Overview of Spyder:

- Open Spyder from Anaconda Navigator
- Take a tour of the interface (Help/Interactive tours)
- Take a look at the shortcuts (Help/Shortcuts Summary)

Exercise - complete the following tutorial parts (Help/Spyder tutorial):

- First steps with Spyder
- Shortcuts for useful functions
- Debugging - line by line step execution of code

2. Jupiter Notebook

Notebook definition [Cambridge Dictionary]

A book of plain paper or paper with lines, for writing on.

2. Jupiter Notebook

Notebook definition [Cambridge Dictionary]

A book of plain paper or paper with lines, for writing on.

Jupiter Notebook definition [https://jupyter.org/]

An open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

2. Jupiter Notebook

Notebook definition [Cambridge Dictionary]

A book of plain paper or paper with lines, for writing on.

Jupiter Notebook definition [https://jupyter.org/]

An open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

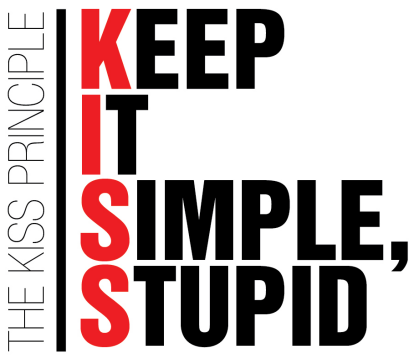
- Jupiter Notebook is good for interactively developing and presenting data science projects
- It supports not only Python, but also many other languages

2. Jupiter Notebook, good practices

- Try to document your code as much as possible
- Use a meaningful naming scheme and code grouping
- Limit line length
- Try to keep the cells of notebooks simple
- Import packages in the first code cell of notebooks

2. Jupiter Notebook, good practices

- Try to document your code as much as possible
- Use a meaningful naming scheme and code grouping
- Limit line length
- Try to keep the cells of notebooks simple
- Import packages in the first code cell of notebooks



2. Anaconda distribution toolkit

Interesting data science packages in Anaconda:

- Numpy
- Matplotlib
- Scipy
- Pandas
- Statsmodels
- Scikit-learn

Section 3

3. About Python

Python is

- Dynamically typed: no need to define the type of variables, function arguments or return types
- Automatically memory managed: no need to explicitly allocate and deallocate memory for variables and data structures
- Interpreted: no need to compile the code. The Python interpreter reads and executes the python code directly
- Slower: the execution of python code can be slow compared to statically compiled languages, such as C++

3. Python resources online

A list of useful resources for Python:

- Tutorial for Python 3 (basic concepts)
  https://docs.python.org/3/tutorial/
- Python Standard Library (list of functions, types, ...)
  https://docs.python.org/3/library/index.html
- Python Language Reference (syntax)
  https://docs.python.org/3/reference/index.html
- Tutorials on the scientific Python ecosystem
  http://scipy-lectures.org/

3. Python contents

Python syntax we are going to learn:

- Variables and types
- Operators (Arithmetic, Boolean, Comparison)
- Basic data structures (Lists, Sets, Tuples, Dictionaries)
- Control flow structures
  - Conditional statements
  - Loops
  - Functions
- Modules imports

References

[1] John King and Roger Magoulas
Data Science Salary Survey: Tools, Trends, What Pays (and What Doesn't) for Data Professionals.
O'Reilly, 2016.