DataScienceSeed

Data Science, Machine Learning, Artificial Intelligence Meetup a Genova, #1

GENOVA MAKERS VILLAGE

iaml
ITALIAN ASSOCIATION FOR MACHINE LEARNING

MACHINE LEARNING
DATA SCIENCE
Meetup

Filse
Finanziaria ligure per lo sviluppo economico
Incubatore di Genova

BIC
Business
Innovation
Center

A·SIGN
engineering software

rulex
analytics

netalia

Comunità e strumenti open source per data science

by Stefania Delprete

Stefania Delprete

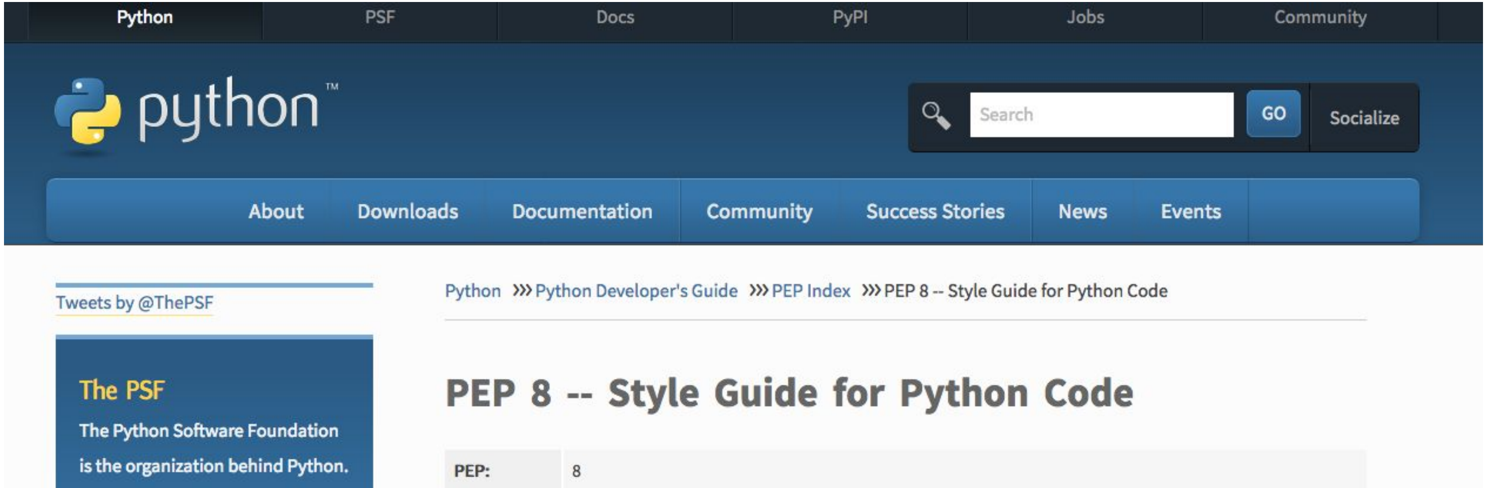
Data Scientist in TOP-IX

/astrastefania su LinkedIn, Twitter, GitHub…

Python, PyCon, EuroPython, EuroSciPy…

Python

- Open Source
- Multi-purpose
- Multi-paradigma
- Leggibile (identazione, PEP8)



The PSF

The Python Software Foundation is the organization behind Python.

Python ››› Python Developer's Guide ››› PEP Index ››› PEP 8 -- Style Guide for Python Code

PEP 8 -- Style Guide for Python Code

PEP:	8

Comunità e conferenze

Comunità di Python su Telegram, Slack…

PyCon 9, Firenze, 19-22 Aprile 2018

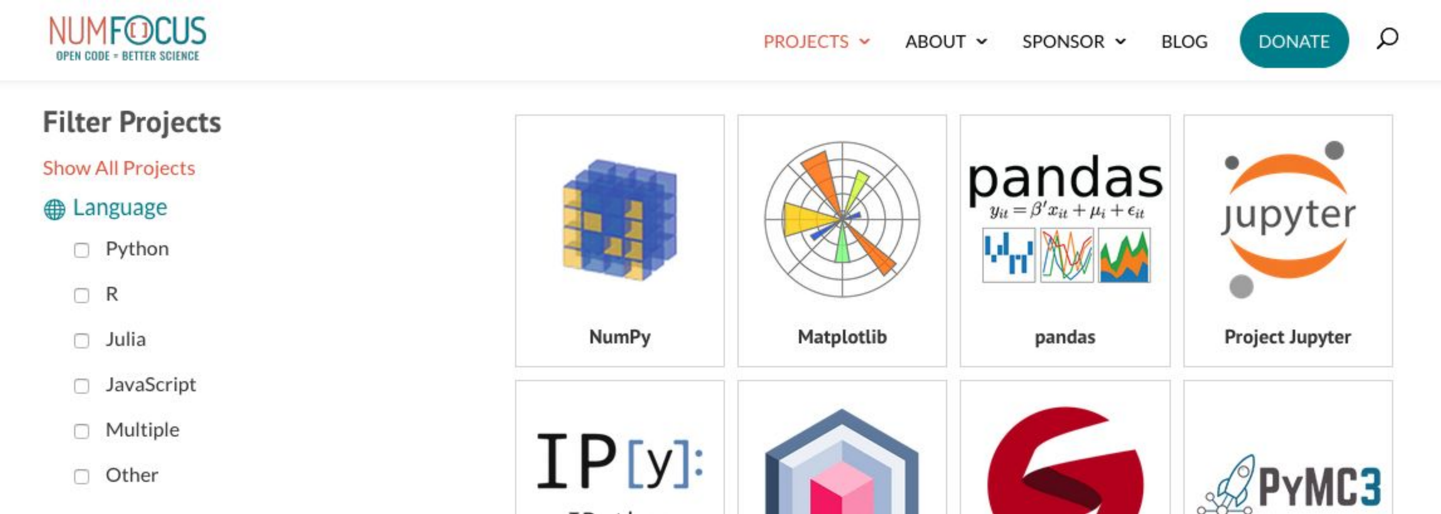
EuroPython, Edimburgo, 23-29 Luglio 2018

EuroSciPy, Trento, 28 Agosto - 1 Settembre 2018

…

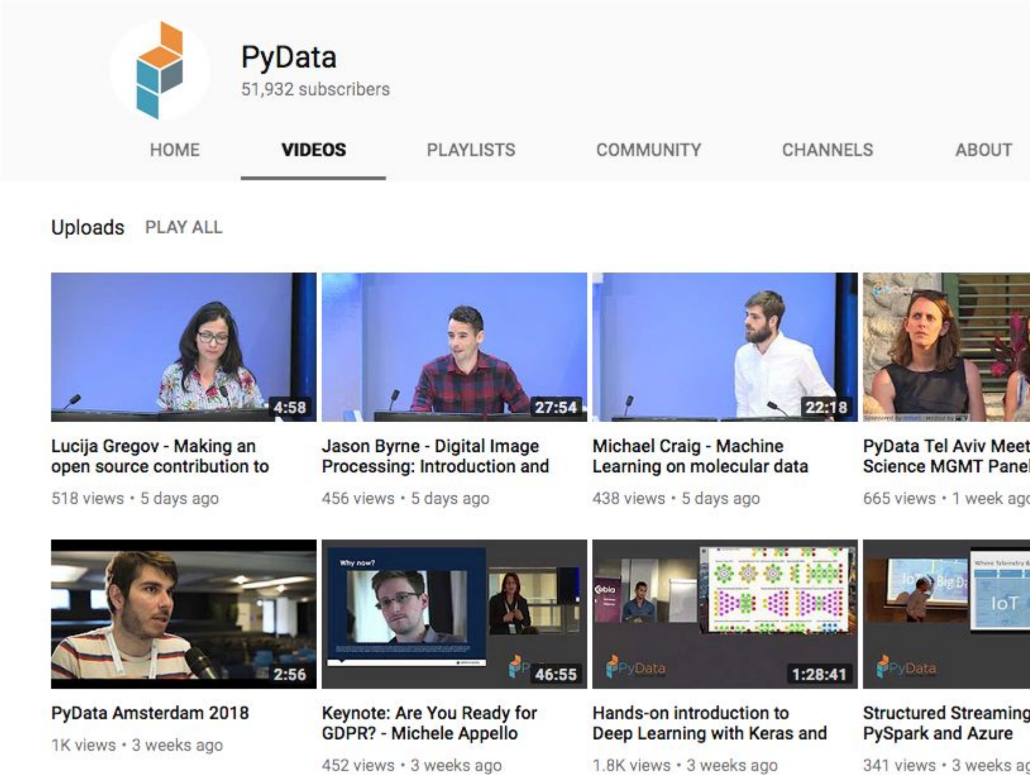
NumFOCUS e PyData

NumFOCUS, 501(c)3 public charity statunitense, sostiene e promuove linguaggi open ad alto livello e progetti a sostegno della comunità scientifica.



NumFOCUS e PyData

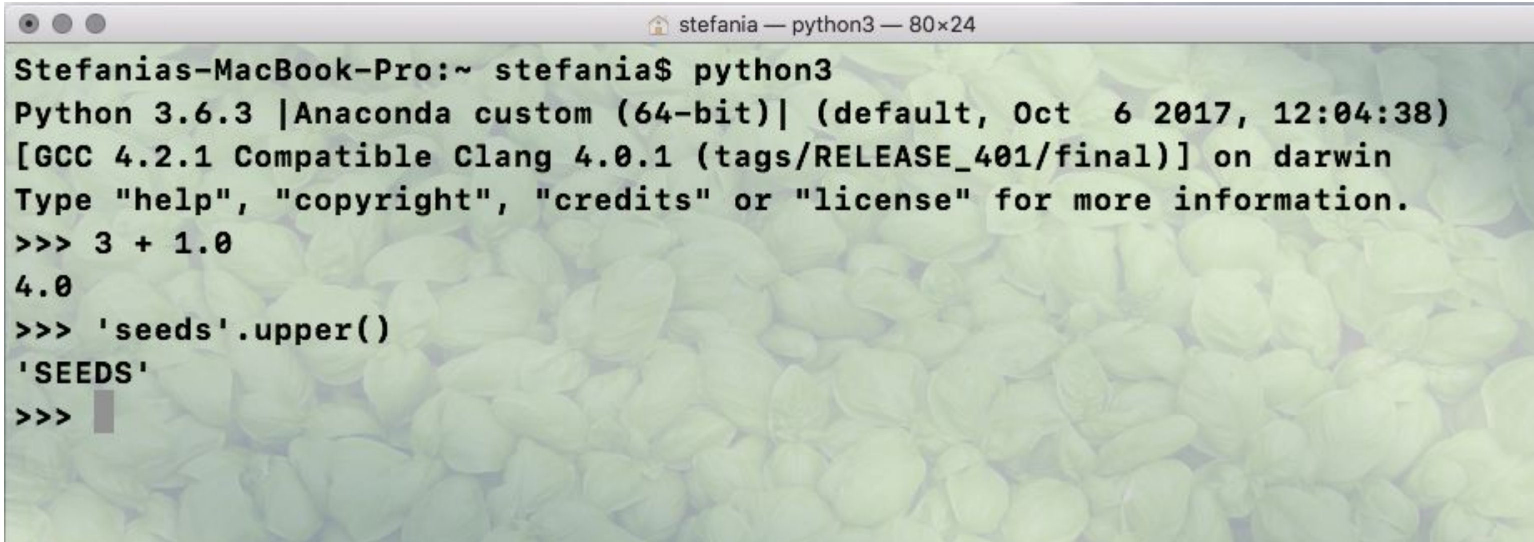
PyData, conferenze dedicate alla divulgazione di progetti di Data Science e Machine Learning con linguaggi open.



Strumenti per Data Science

Da REPL...

Python REPL (Read–Eval–Print Loop), possiamo imparare Python direttamente sul terminale...



```
Stefanias-MacBook-Pro:~ stefania$ python3
Python 3.6.3 |Anaconda custom (64-bit)| (default, Oct  6 2017, 12:04:38)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> 3 + 1.0
4.0
>>> 'seeds'.upper()
'SEEDS'
>>> 
```

… a Jupyter Notebook

IPython, 2001, Fernando Pérez, fisico

- Ottimo strumento per imparare Python, Data Science e Machine Learning
- Espansione ad altri linguaggi (Julia, Python, R...)

NumPy

Numeric, 1995, Jim Hugunin, programmatore

SciPy, 1999 algoritmi e strumenti matematici in Python

NumPy, 2006, Travis Oliphant, data scientist

Libreria Python che comprende conversioni a vettori e matrici, calcoli algebrici, grande gamma di funzioni matematiche.

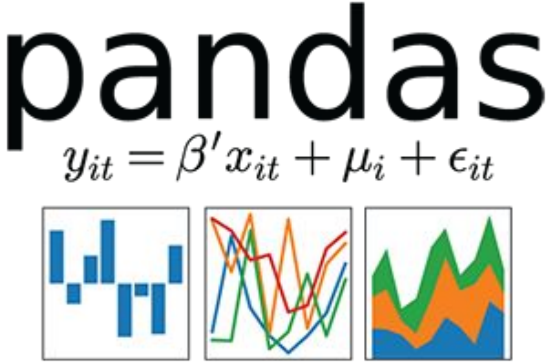
Distribuzione Anaconda

Anaconda, distribuzione di un’ampia collezione di librerie per Data Science e Machine Learning (suo package manager *conda*).

pandas

pandas, 2008, Wes McKinney, statistico

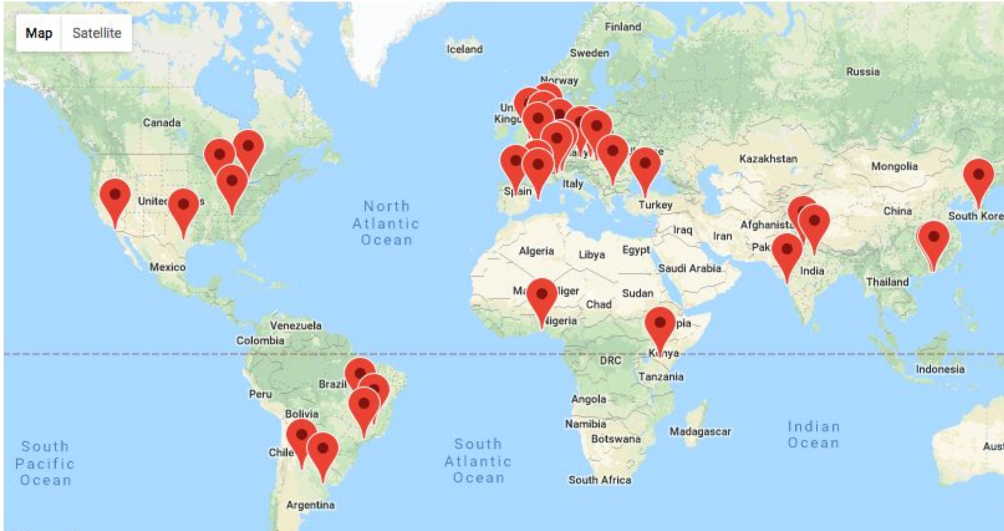
Libreria per manipolazione dei dati, permette di convertire diversi formati in un suo tipo **pandas DataFrame**.



pandas Documentation Sprint

10 Marzo 2018

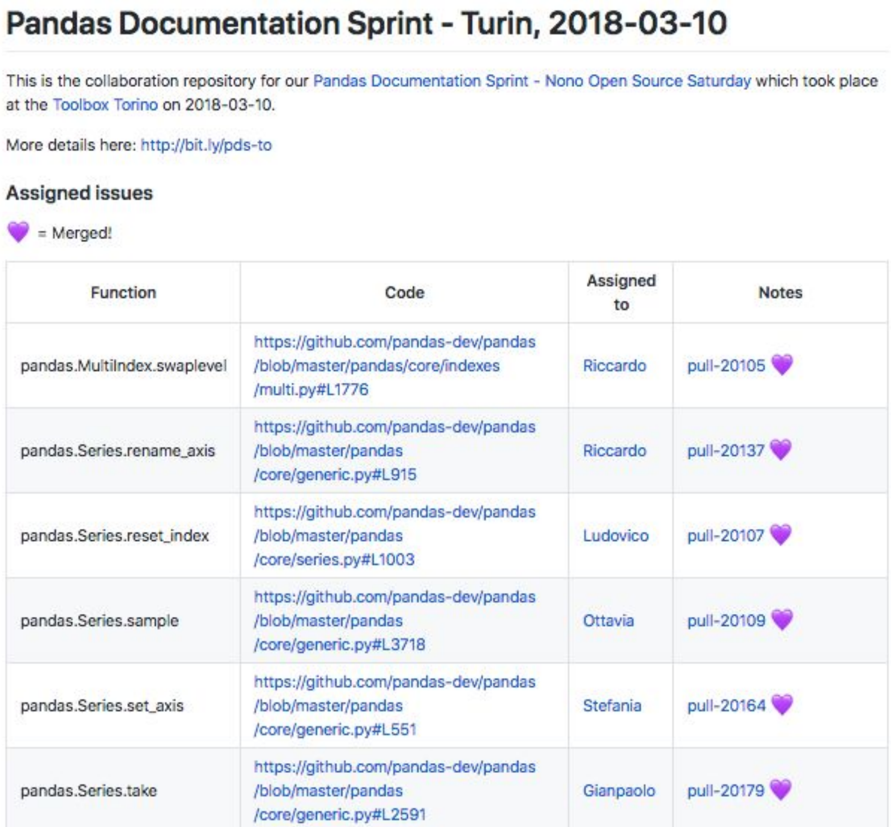
- 500 programmatori
- 30 città
- 6 ore



pandas Documentation Sprint

10 Marzo 2018

Ottima esperienza per iniziare a contribuire nell'open source!



Pandas Documentation Sprint - Turin, 2018-03-10

This is the collaboration repository for our Pandas Documentation Sprint - Nono Open Source Saturday which took place at the Toolbox Torino on 2018-03-10.

More details here: http://bit.ly/pds-to

Assigned issues

💜 = Merged!

Function	Code	Assigned to	Notes
pandas.MultiIndex.swaplevel	https://github.com/pandas-dev/pandas/blob/master/pandas/core/indexes/multi.py#L1776	Riccardo	pull-20105 💜
pandas.Series.rename_axis	https://github.com/pandas-dev/pandas/blob/master/pandas/core/generic.py#L915	Riccardo	pull-20137 💜
pandas.Series.reset_index	https://github.com/pandas-dev/pandas/blob/master/pandas/core/series.py#L1003	Ludovico	pull-20107 💜
pandas.Series.sample	https://github.com/pandas-dev/pandas/blob/master/pandas/core/generic.py#L3718	Ottavia	pull-20109 💜
pandas.Series.set_axis	https://github.com/pandas-dev/pandas/blob/master/pandas/core/generic.py#L551	Stefania	pull-20164 💜
pandas.Series.take	https://github.com/pandas-dev/pandas/blob/master/pandas/core/generic.py#L2591	Gianpaolo	pull-20179 💜

Visualizzare dati

Matplotlib

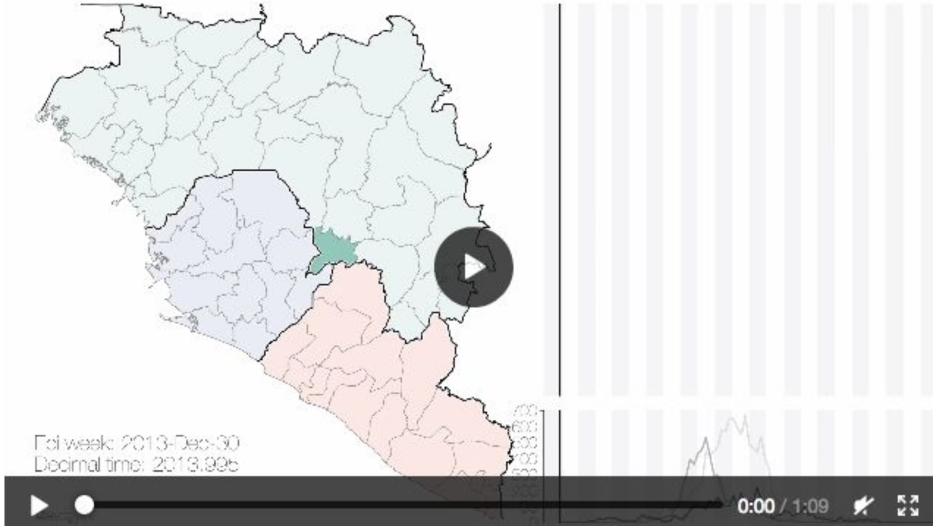
Matplotlib, 2003, John D. Hunter, neurobiologo

Strumento potente e leggero per le maggiori visualizzazioni statistiche.

John Hunter Plotting Contest 2018

Winners
“History of Ebola virus epidemic in West Africa 2013-2015”

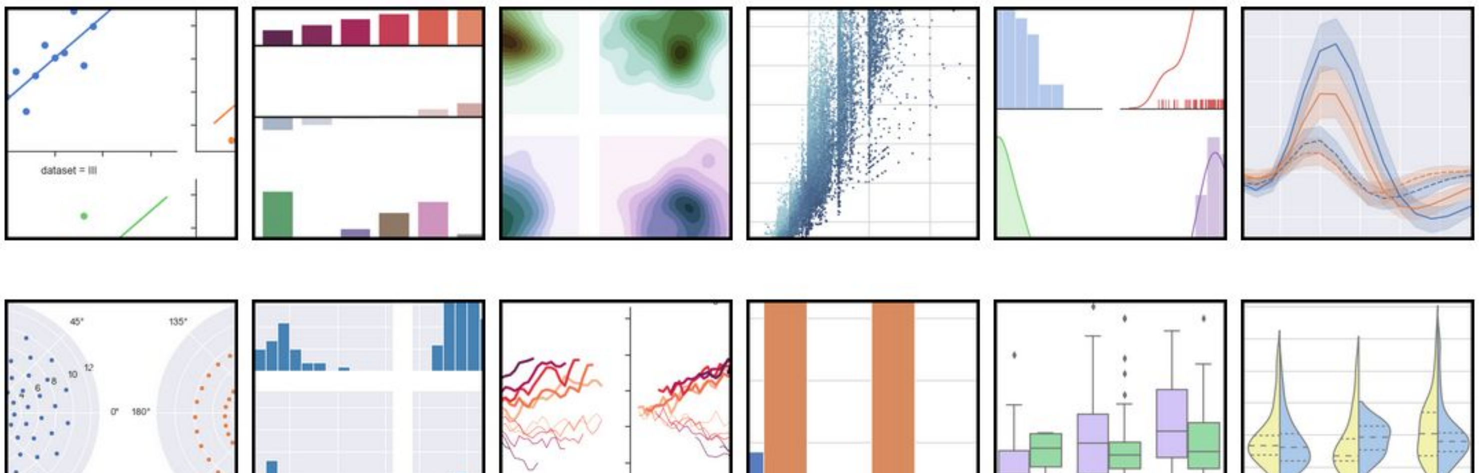
by Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Marc A. Suchard, Philippe Lemey, and Andrew Rambaut.

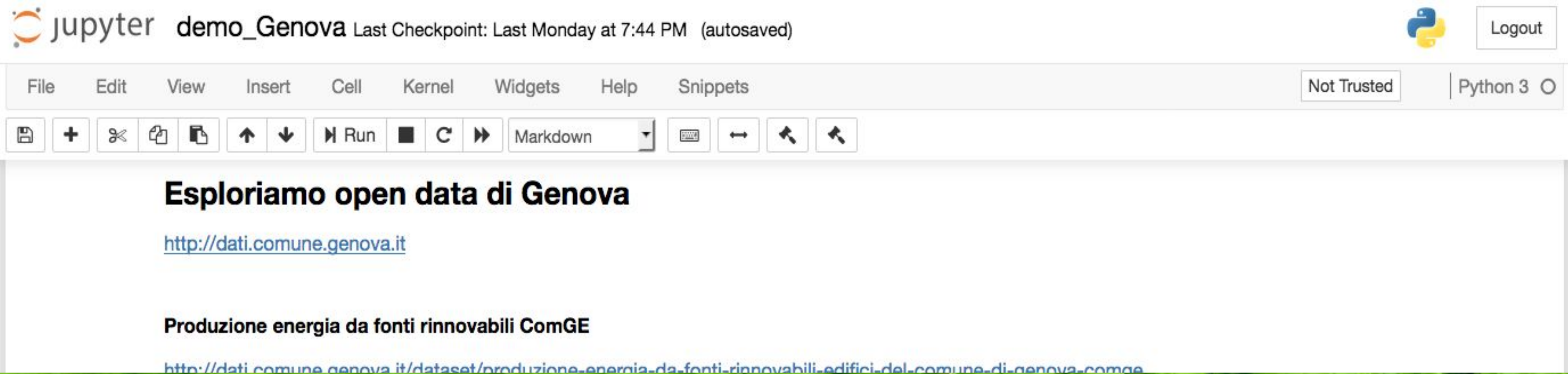


Seaborn

Seaborn, sviluppato sulla base di Matplotlib

Aggregazione di grafici, veloce implementazione di visualizzazioni.





Esploriamo open data di Genova

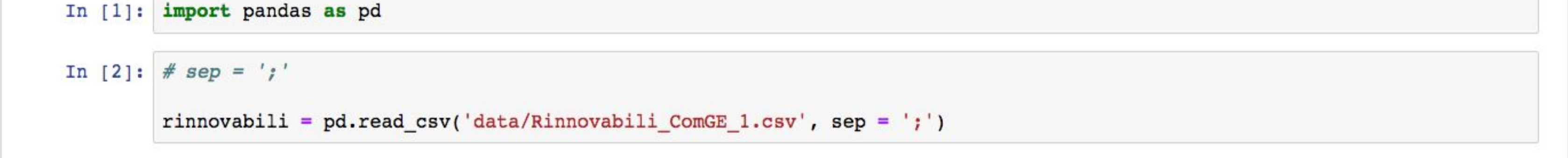
http://dati.comune.genova.it

Produzione energia da fonti rinnovabili ComGE

http://dati.comune.genova.it/dataset/produzione-energia-da-fonti-rinnovabili-edifici-del-comune-di-genova-comge

Demo su Jupyter Notebook

Esempi di pandas, maplotplotlib e seaborn su open data



```
In [1]: import pandas as pd
```

```
In [2]: # sep = ';'

        rinnovabili = pd.read_csv('data/Rinnovabili_ComGE_1.csv', sep = ';')
```

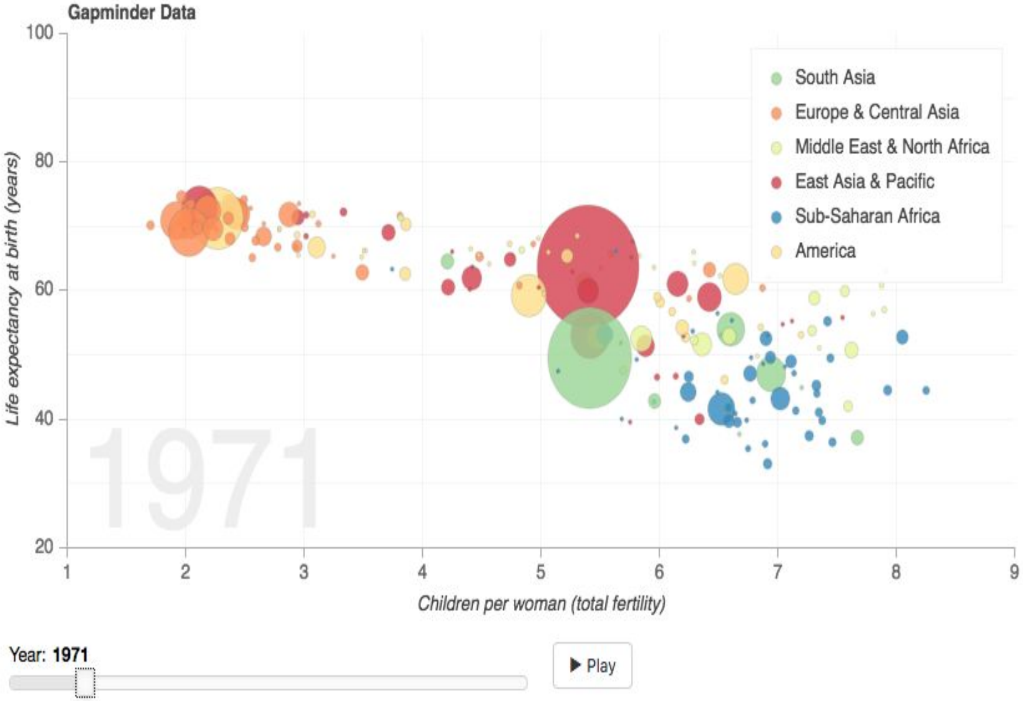
Bokeh

Bokeh, libreria per visualizzazioni interattive ottimizzata per rappresentazioni su web browser.

Permette di realizzare grafici interattivi anche con dataset molto grandi o streaming.

Bokeh

Riproduzione con Bokeh del conosciuto TED talk di Hans Rosling **“The best stats you've ever seen”**



D3.js

D3, libreria di JavaScript per visualizzare dati con HTML, SVG e CSS.

Realizzazione di dashboard interattive e grafici totalmente personalizzati, integrazione sul web.

Machine Learning

Scikit-learn

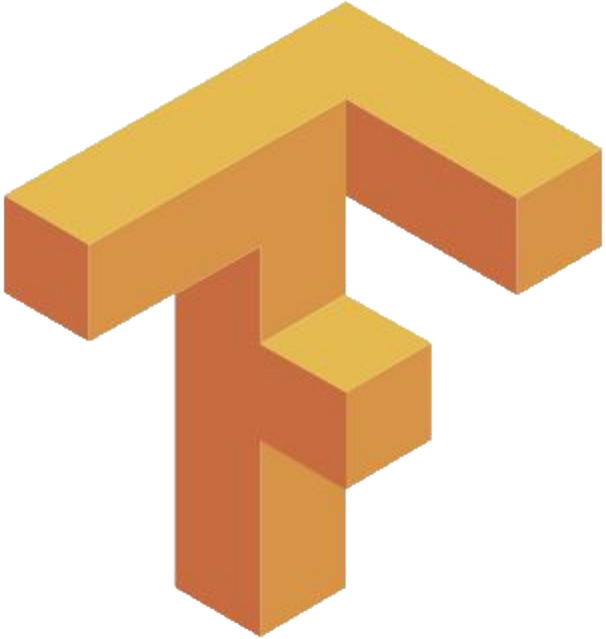
Scikit-learn, 2007,
David Cournapeau, data scientist

Sviluppato su NumPy, SciPy e matplotlib, è la risorsa per eccellenza per fare Machine Learning con Python per la sua gran collezione di algoritmi facilmente implementabili.

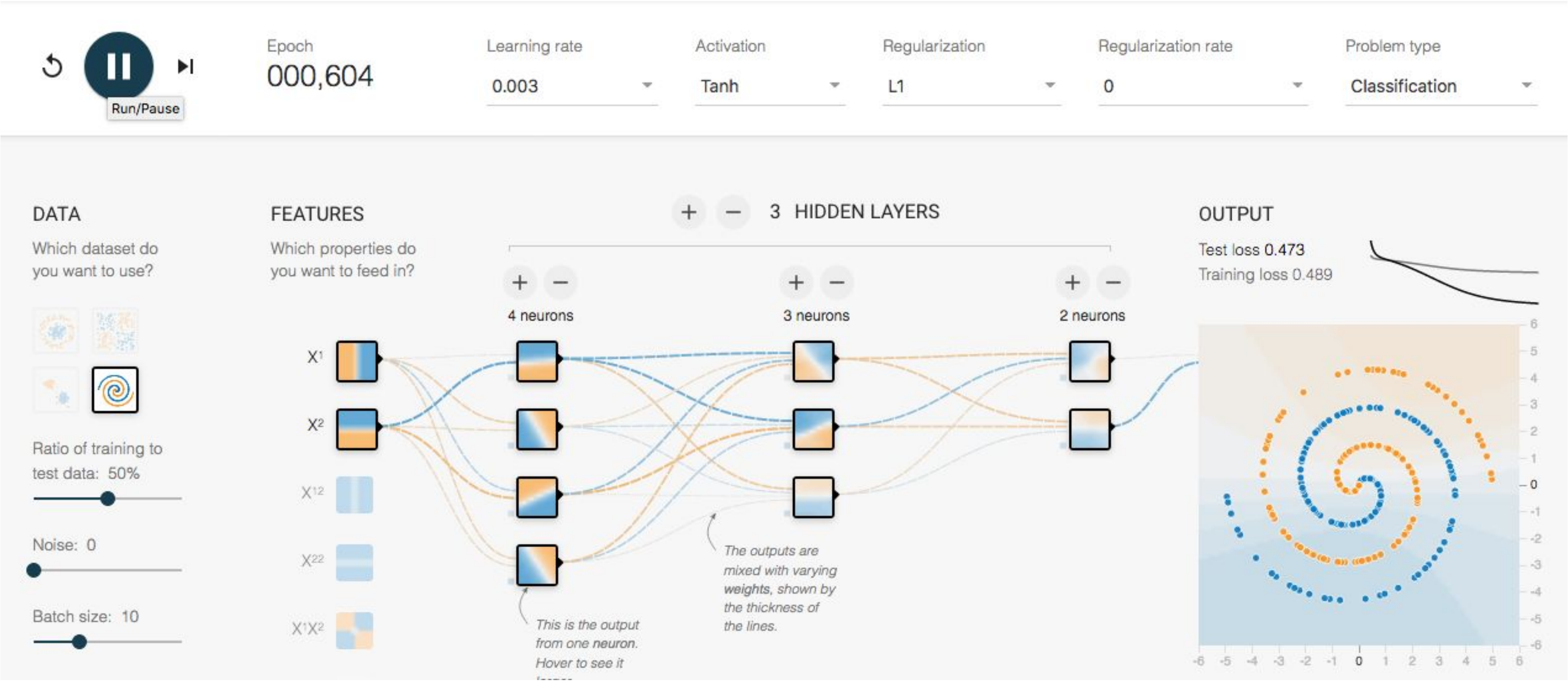
TensorFlow

TensorFlow, 2015, Google Brain

Potente strumento per sviluppare progetti di Machine Learning e reti neurali.



TensorFlow playground



Epoch 000,604

Learning rate 0.003

Activation Tanh

Regularization L1

Regularization rate 0

Problem type Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

FEATURES

Which properties do you want to feed in?

3 HIDDEN LAYERS

4 neurons

3 neurons

2 neurons

OUTPUT

Test loss 0.473

Training loss 0.489

This is the output from one neuron. Hover to see it

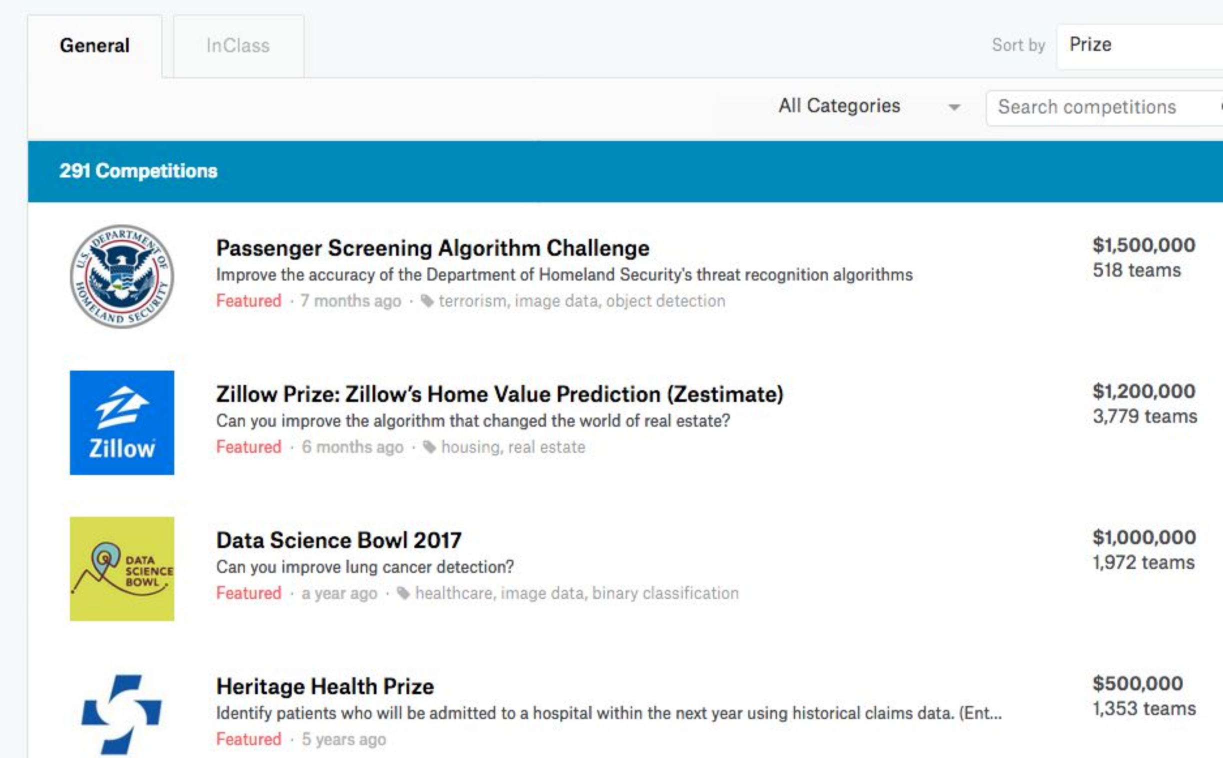
*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

Trovare dati open

Alcuni esempi utili

Kaggle

Kaggle, piattaforma in cui partecipare a sfide mondiali, iniziare nuovi progetti o trovare ottime fonti di dataset.



291 Competitions

Passenger Screening Algorithm Challenge
Improve the accuracy of the Department of Homeland Security's threat recognition algorithms
Featured · 7 months ago · terrorism, image data, object detection
$1,500,000
518 teams

Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?
Featured · 6 months ago · housing, real estate
$1,200,000
3,779 teams

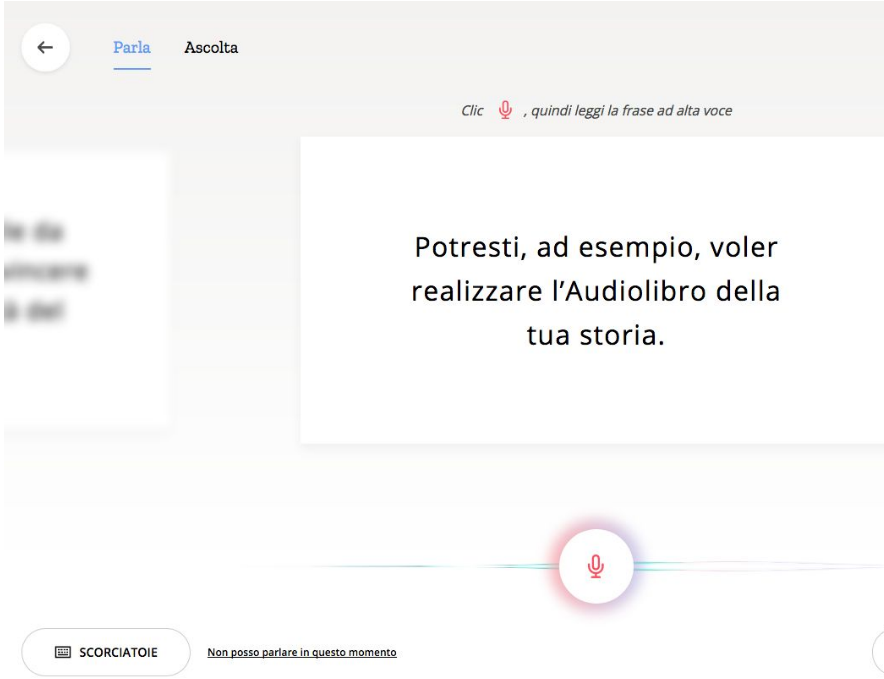
Data Science Bowl 2017
Can you improve lung cancer detection?
Featured · a year ago · healthcare, image data, binary classification
$1,000,000
1,972 teams

Heritage Health Prize
Identify patients who will be admitted to a hospital within the next year using historical claims data. (Ent...
Featured · 5 years ago
$500,000
1,353 teams

Da dataset a common

Common Voice, progetto di Mozilla

(da poco anche in Italiano!), puoi contribuire registrando la tua voce leggendo frasi, validando registrazioni, aggiungendo stringhe al progetto…



Clic , quindi leggi la frase ad alta voce

Potresti, ad esempio, voler realizzare l'Audiolibro della tua storia.

Pubblica Amministrazione

DAF, piattaforma dati italiani dataportal.daf.teamdigitale.it

Genova dati.comune.genova.it

UK data.gov.uk

USA data.gov

Attenzione: puoi trovare dati in vari formati
(testuali, tabelle di vario tipo, dati geospaziali…)

Ulteriori risorse

Tutte le immagini sulle slide sono cliccabili.

Ecco altri collegamenti ad alcuni progetti delle librerie Python menzionate per esplorare il codice e contribuire. ;)

NumPy https://github.com/numpy/numpy

Pandas https://github.com/pandas-dev/pandas

Matplotlib https://github.com/matplotlib/matplotlib

Seaborn https://github.com/mwaskom/seaborn

Grazie, buona esplorazione!

stefania.delprete@top-ix.org

linkedin.com/in/astrastefania
twitter.com/astrastefania