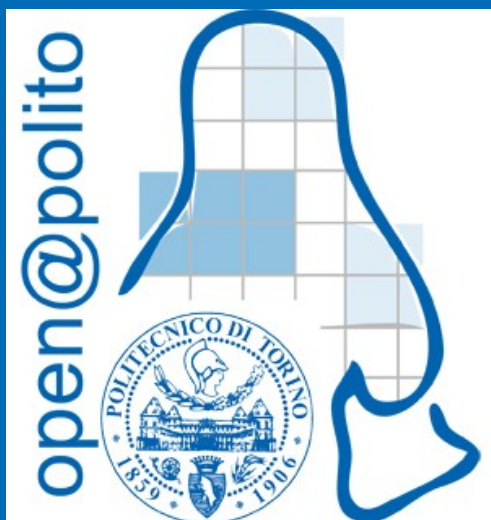




Data science e comunità open source

Con il supporto di:



Stefania Delprete

16/12/2021

AGENDA

- ▶ Conosciamoci
- ▶ Dati, open data e data science
- ▶ Strumenti open source (e perché usarli)
- ▶ Alcune comunità in cui sono coinvolta
- ▶ Domande? :)

Stefania Delprete



- Data Scientist in **TOP-IX**, tutor in **BIG DIVE**
- Coinvolta in comunità di Altruismo Efficace ed open source
- Giro mensilmente con giochi per il **MathsJam**
- Appassionata di fisica, neuroscienze, sogni e jazz

 **@astrastefania**

DATI?



Quali tipi di dati ti vengono in mente?

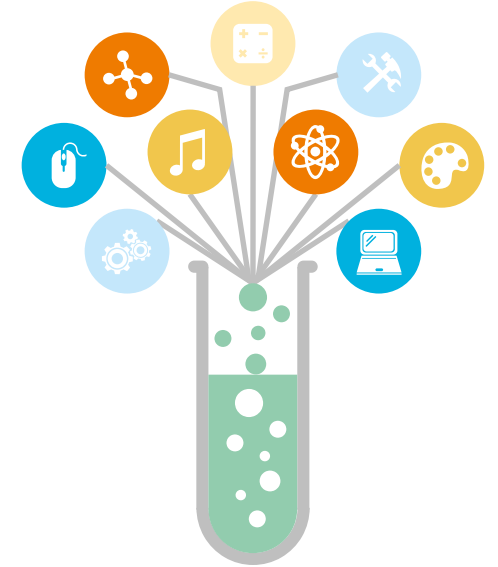
Da dove potresti acquisirli?



Esistono sorgenti aperte?

DATI DI TUTTI I TIPI

- ▶ Numerici (misure, metriche, ...)
- ▶ Testuali (web scraping, sondaggi, ...)
- ▶ Immagini (satellitari, archivi, ...)
- ▶ Audio (registrazioni, musica, ...)
- ▶ Video (telecamere di sicurezza, ...)



ESEMPI DI FONTI DI OPEN DATA

- ▶ Torrette dell'ARPA, portali regionali e nazionali
- ▶ Twitter API, portali Open Science
- ▶ Immagini da ESA, NASA
- ▶ Progetti come Common Voice
- ▶ *Segui la tua curiosità!*

DATA SCIENCE



Come analizzereste tali dati?

A quali domande vorresti rispondere?



Dove vuoi arrivare?

DATA SCIENCE

- ▶ Acquisizione dei dati
- ▶ Preparazione e pulitura dei dati
- ▶ Analisi statistica e rappresentazioni grafiche
- ▶ Osservazioni di trend, risposte a domande iniziali
- ▶ Eventuale creazione e implementazioni predittive

STRUMENTI OPEN SOURCE



Quali linguaggi di programmazione conosci?

Con cosa realizzeresti tali analisi?



Python

[Python](#)[PSF](#)[Docs](#)[PyPI](#)[Jobs](#)[Community](#)

[Donate](#)

[GO](#)

[Socialize](#)

[About](#)[Downloads](#)[Documentation](#)[Community](#)[Success Stories](#)[News](#)[Events](#)

```
# Python 3: List comprehensions
>>> fruits = ['Banana', 'Apple', 'Lime']
>>> loud_fruits = [fruit.upper() for fruit in fruits]
>>> print(loud_fruits)
['BANANA', 'APPLE', 'LIME']

# List and the enumerate function
>>> list(enumerate(fruits))
[(0, 'Banana'), (1, 'Apple'), (2, 'Lime')]
```

12345

Compound Data Types

Lists (known as arrays in other languages) are one of the compound data types that Python understands. Lists can be indexed, sliced and manipulated with other built-in functions. [More about lists in Python 3](#)

Python is a programming language that lets you work quickly and integrate systems more effectively. [>>> Learn More](#)

Jupyter Notebook

I Jupyter Notebook, una volta IPython Notebook, sono diventati uno standard per ogni buona partenza nello studio e sviluppo della data science.



Permettono di eseguire celle di codice alternate a documentazione scritta in Markdown.

NumPy

NumPy è la libreria computazionale scientifica per eccellenza.

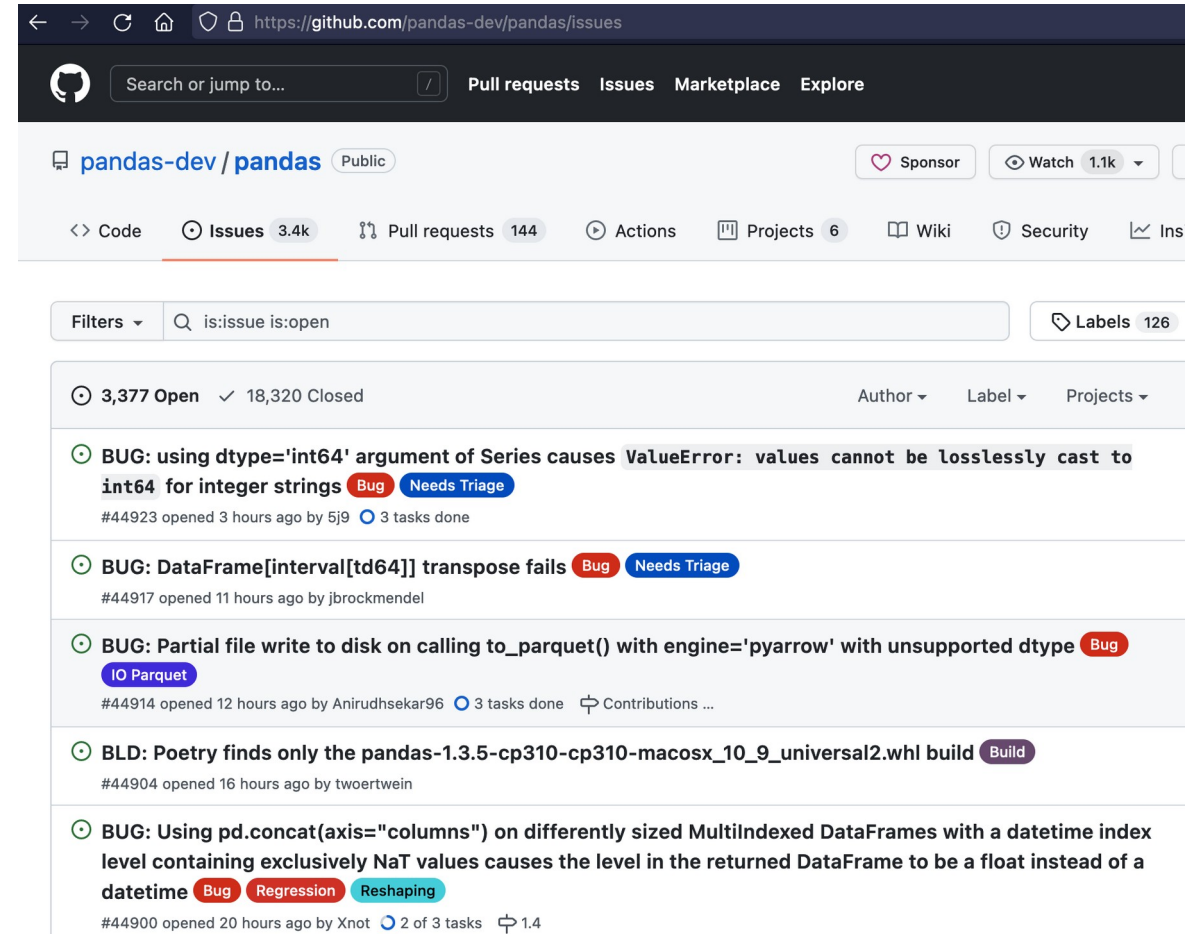
Ottima per la creazione di vettori, matrici, tensori; generazione di numeri random; ...

NumPy è anche alla base di librerie per lo sviluppo di reti neurali.



Pandas

Libreria per importare e manipolare dataset e far prime analisi statistiche



Pandas

Nel 2018 qui a Torino abbiamo partecipato ad una sprint mondiale per contribuire alla documentazione!



Screenshot of a GitHub repository page for `astrastefania/pandas-sprint-Turin`. The page shows a merge pull request #9 from `gmacario/patch-3` on Mar 17, 2018, with 34 commits. The repository contains files: `LICENSE` (Create LICENCE, 4 years ago), `README.md` (20164 got merged too, 4 years ago), and `pandas-docs-sprint-intro.ipynb` (Add Pandas Sprint intro, 4 years ago).

The `README.md` file content is displayed below:

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>

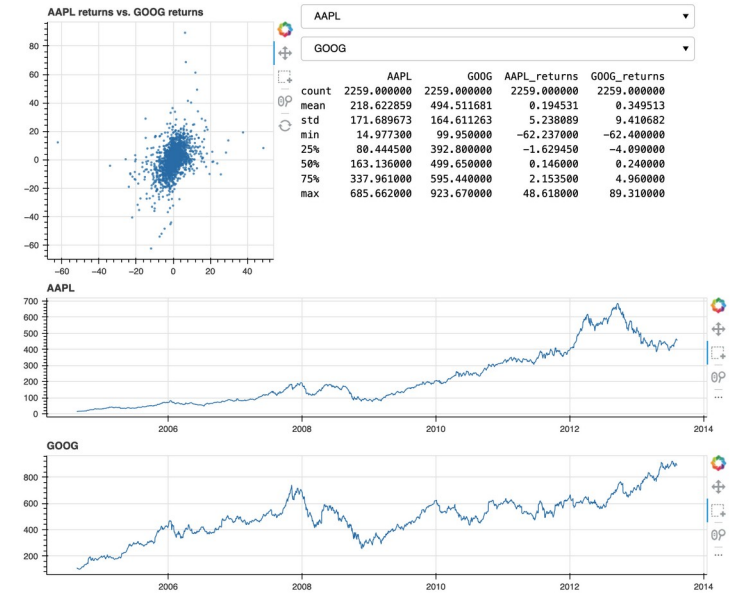
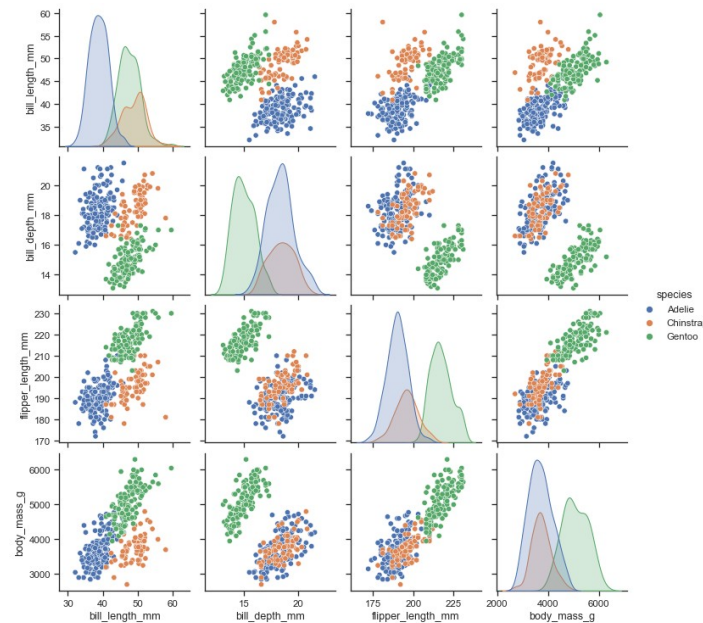
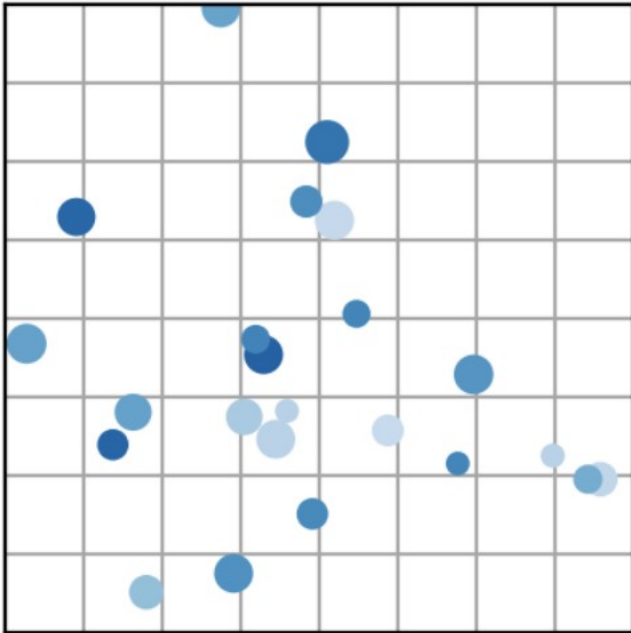
Assigned issues

♥ = Merged!

Function	Code	Assigned to	Notes
<code>pandas.MultiIndex.swaplevel</code>	https://github.com/pandas-dev/pandas/blob/master/pandas/core/indexes/multi.py#L1776	Riccardo	<code>pull-20105</code> ♥

Matplotlib, Seaborn, Bokeh

Librerie per eccellenza per la rappresentazione grafica in Python



Scikit-learn, Tensor Flow, Keras...

Scikit-learn da
accesso ad algoritmi
di machine learning e
ad una straordinaria
documentazione...

 [Install](#) [User Guide](#) [API](#) [Examples](#) [More ▾](#)

scikit-learn

Machine Learning in Python

[Getting Started](#) [Release Highlights for 1.0](#) [GitHub](#)

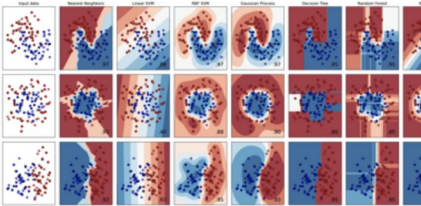
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...



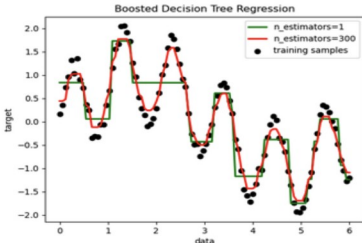
Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



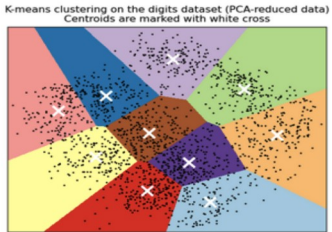
Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...



Examples

Dimensionality reduction

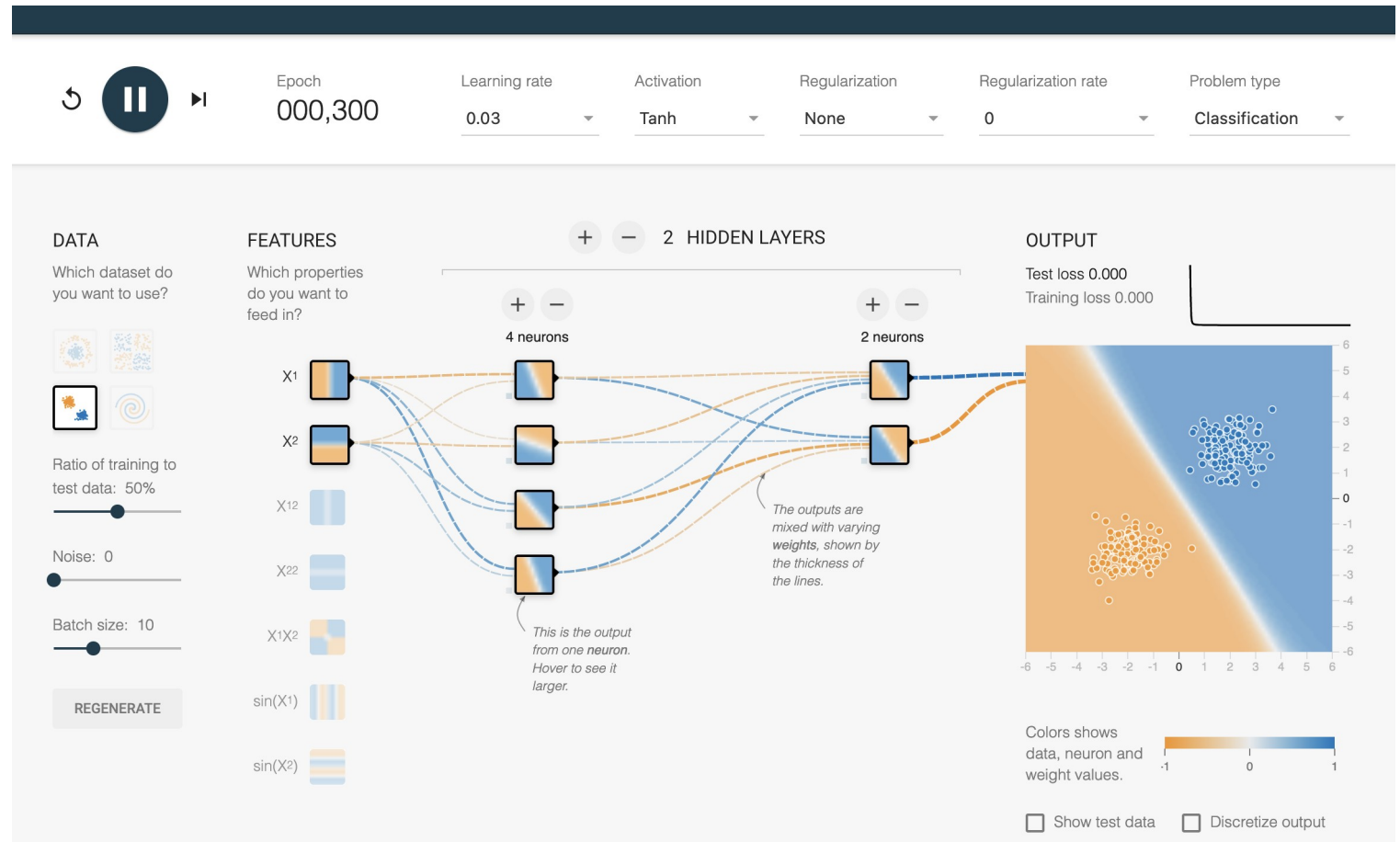
Model selection

Preprocessing

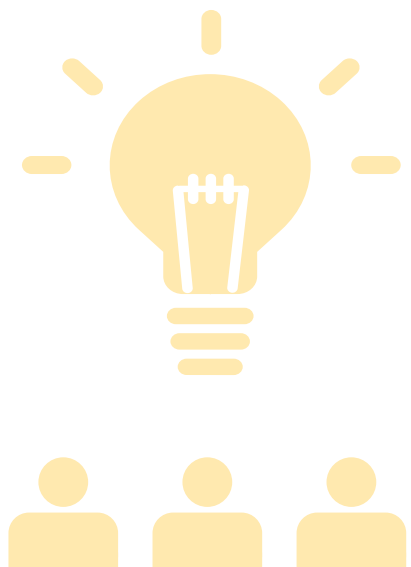
Reti neurali con TensorFlow, Keras...

TensorFlow e Keras sono due fra le librerie per reti neurali.

TensorFlow Playground permette di avere un'idea dell'implementazione di una rete neurale



RIUNIAMOCI!



Quali comunità informatiche conosci?

Vorresti farne parte?



NumFOCUS

[ABOUT ▾](#)[PROJECTS ▾](#)[PROGRAMS ▾](#)[SUPPORT ▾](#)[BLOG](#)[DONATE](#)

Better tools to build a better world.

From Netflix to NASA, researchers use our open source tools to solve the most challenging problems.

[View Our Impact](#)

NumFOCUS



OPEN CODE • BETTER SCIENCE

ABOUT ▾ PROJECTS ▾ PROGRAMS ▾ SUPPORT ▾ BLOG [DONATE](#) 🔍

Show All Projects

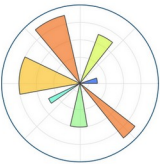
🌐 Language

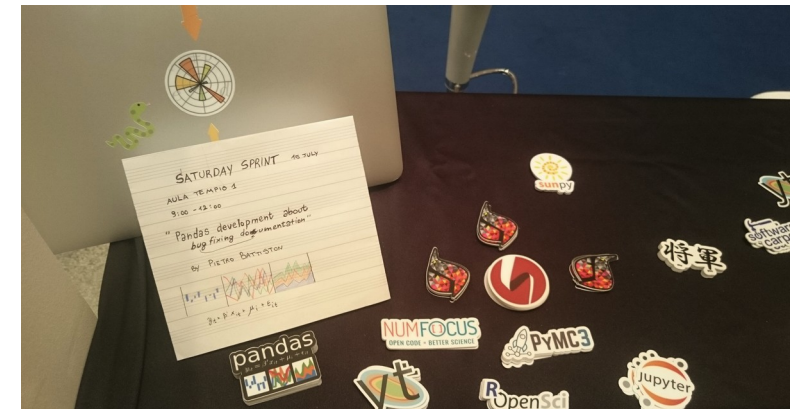
- ☐ Python
- ☐ R
- ☐ Julia
- ☐ JavaScript
- ☐ Other
- ☐ Multiple

⚙️ Features

- ☐ Browser Interactivity
- ☐ Assistive Technology
- ☐ Data Wrangling
- ☐ Modeling
- ☐ Visualization
- ☐ High Performance Computing
- ☐ Big Data
- ☐ Statistical Computing
- ☐ Numerical Computing
- ☐ Data Mining
- ☐ Text Processing

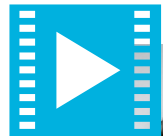
Click on the project logo for detailed information about each project.

 NumPy	 Matplotlib	 pandas	 Project Jupyter
 IPython	 SciPy	 interact	 Stan
 PyMC	 Julia	 JuMP	 PyTables
 將軍			 vt

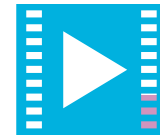


Python e PyCon

PyCon Italia, conferenza annuale della comunità di Python, che si svolge quasi sempre a Firenze, spesso include una track **PyData**.



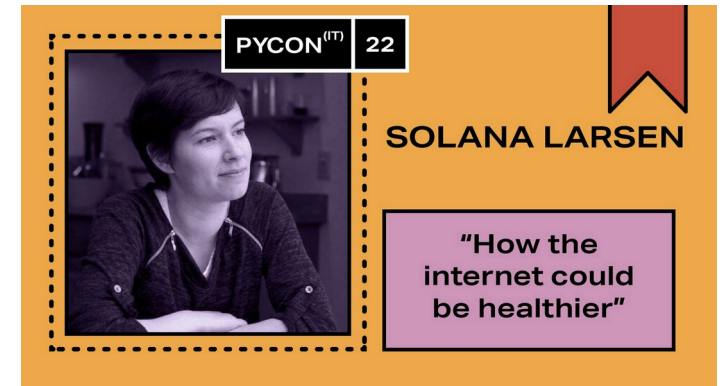
Quest'anno in versione ibrida online come **PyFest**



La prossima PyCon...

La prossima PyCon Italia è già in preparazione...

...e la call per i nuovi speaker è già aperta!



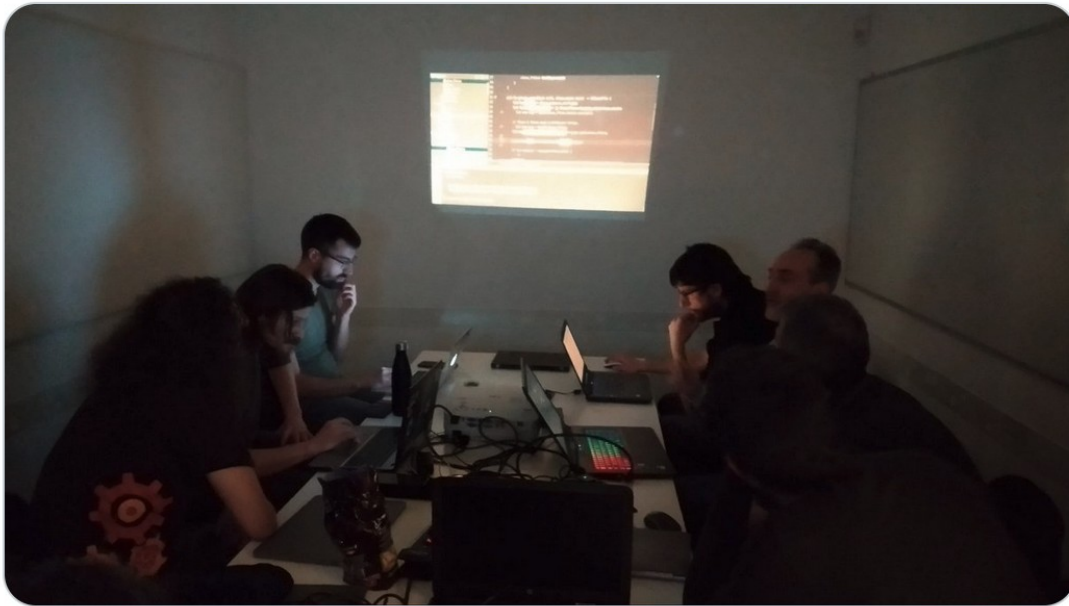
Mozilla

Il mio incontro con Mozilla per me partì in quel di Berlino...



Mozilla

...per poi continuare in quel di Torino e con **Mozilla Italia**

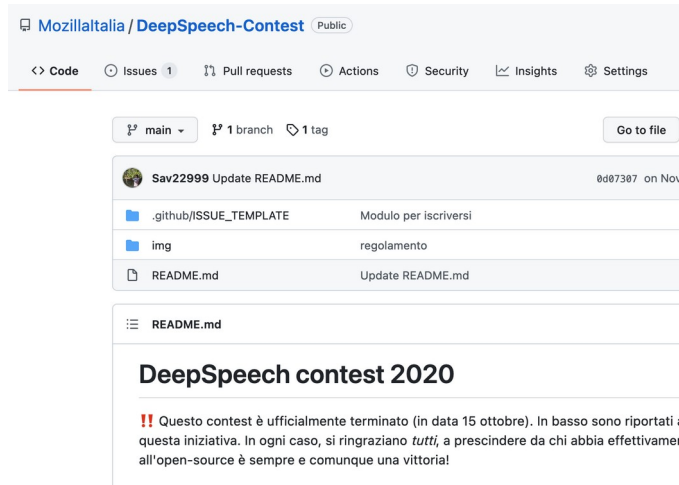


Gruppo di studio di Rust



Common Voice in italiano!

DeepSpeech e Common Voice



Mozilla Common Voice è un'iniziativa per insegnare alle macchine come parlano le persone nella vita reale.

La voce è naturale, la voce è umanità. Per questo l'idea di creare una tecnologia vocale utilizzabile dalle nostre macchine ci affascina così tanto. Per creare sistemi basati sulla tecnologia vocale, però, è necessaria un'altissima quantità di campioni.

Buona parte dei dati usati dalle grandi aziende non è accessibile alla maggioranza delle persone. Secondo noi questo soffoca l'innovazione. Per questa ragione abbiamo sviluppato Common Voice, un progetto di riconoscimento vocale aperto a tutti.

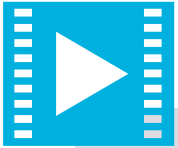
[ULTERIORI INFORMAZIONI](#)

Mozilla Italia



- ▶ Supporto e divulgazione di prodotti Mozilla
- ▶ Grande comunità di localizzazione
- ▶ Sprint per contribuire a Common Voice
- ▶ Contest per creare progetti con DeepSpeech
- ▶ Eventi locali, speriamo di tornare a rivederci!

Linux Day 2021



E per finire....

Quest'anno anche per il **Linux Day** il tema centrale è stato proprio sono stati i dati “**Dati, a chi?**”



DOMANDE?



Grazie per l'attenzione!

STEFANIA DELPRETE

stefania.delprete@top-ix.org

 @astrastefania