

DataScienceSeed

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



The background of the slide is a dense, vibrant green texture of fresh basil leaves. The leaves are small, rounded, and have a slightly serrated edge, typical of Italian basil. They are packed closely together, creating a rich, natural pattern that frames the central text area.

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 screenshot shows the Python.org website. The top navigation bar includes links for Python, PSF, Docs, PyPI, Jobs, and Community. Below this is a dark blue header with the Python logo and a search bar. A secondary navigation bar contains links for About, Downloads, Documentation, Community, Success Stories, News, and Events. The main content area shows a breadcrumb trail: Python >>> Python Developer's Guide >>> PEP Index >>> PEP 8 -- Style Guide for Python Code. The page title is "PEP 8 -- Style Guide for Python Code". On the left, there is a sidebar with a section titled "The PSF" and a description: "The Python Software Foundation is the organization behind Python." Below this, there is a table with two columns: "PEP:" and a value. The first row shows "PEP:" followed by "8".

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.




PROJECTS ▾ ABOUT ▾ SPONSOR ▾ BLOG [DONATE](#) 🔍

Filter Projects


[Show All Projects](#)

 **Language**

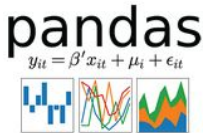
- ☐ Python
- ☐ R
- ☐ Julia
- ☐ JavaScript
- ☐ Multiple
- ☐ Other




NumPy




Matplotlib




pandas





Project Jupyter



IP[y]:



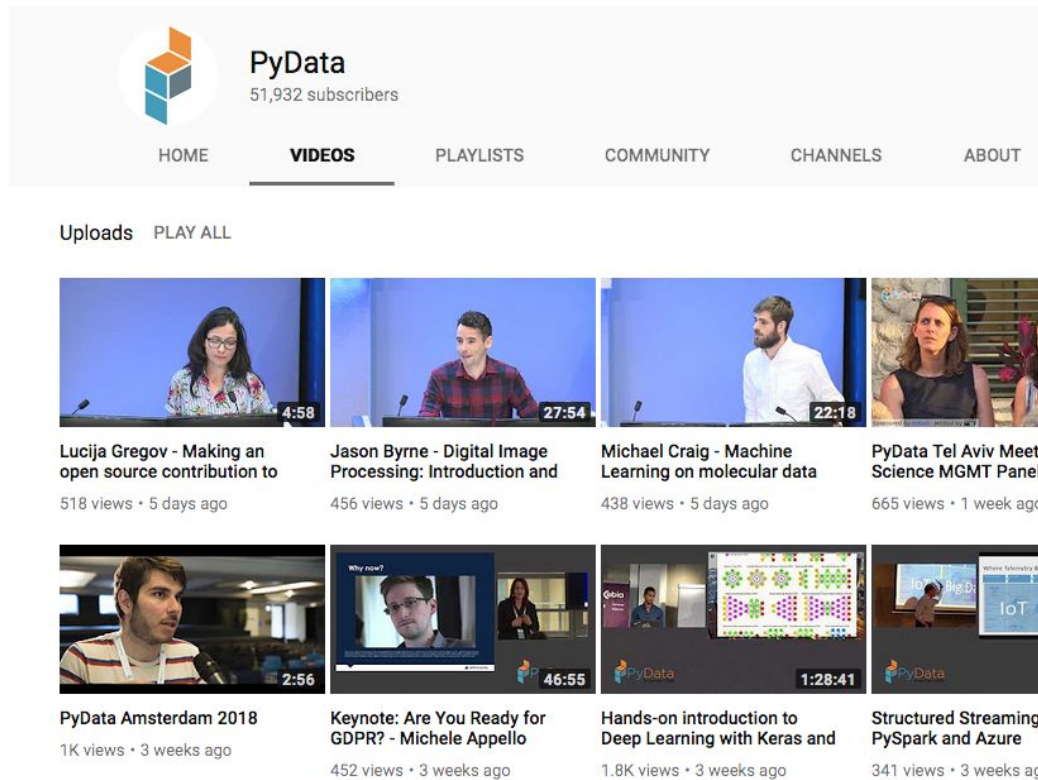




PyMC3

NumFOCUS e PyData

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



PyData
51,932 subscribers

HOME **VIDEOS** PLAYLISTS COMMUNITY CHANNELS ABOUT

Uploads **PLAY ALL**

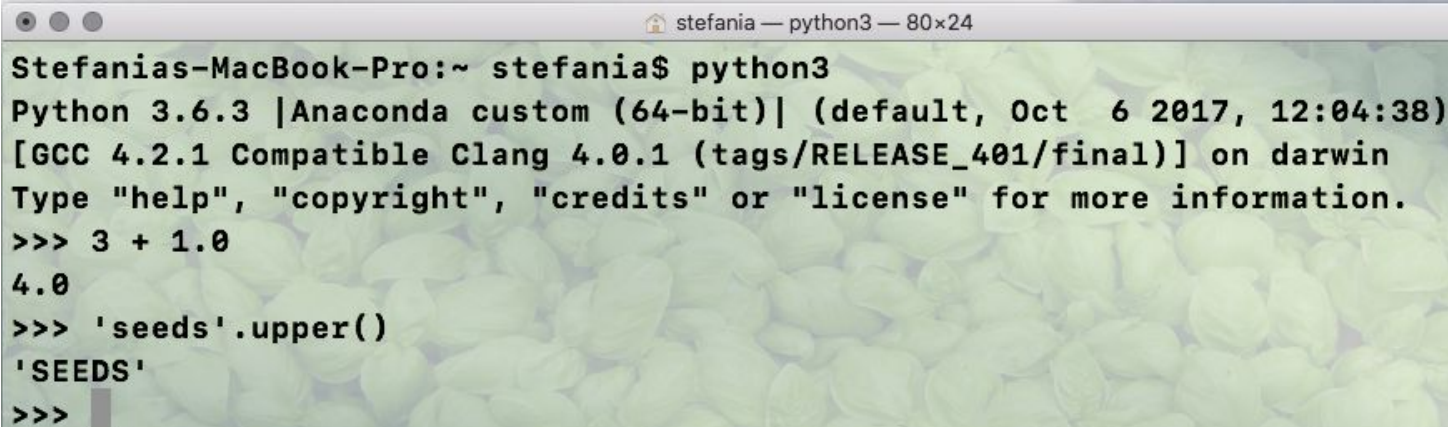
Video Title	Views	Time Ago
Lucija Gregov - Making an open source contribution to	518 views	5 days ago
Jason Byrne - Digital Image Processing: Introduction and	456 views	5 days ago
Michael Craig - Machine Learning on molecular data	438 views	5 days ago
PyData Tel Aviv Meet Science MGMT Panel	665 views	1 week ago
PyData Amsterdam 2018	1K views	3 weeks ago
Keynote: Are You Ready for GDPR? - Michele Appello	452 views	3 weeks ago
Hands-on introduction to Deep Learning with Keras and	1.8K views	3 weeks ago
Structured Streaming PySpark and Azure	341 views	3 weeks ago

The background of the slide is a dense, close-up photograph of fresh green basil leaves. The leaves are vibrant green, with some showing slight variations in shade and texture. They are arranged in a way that creates a repeating pattern across the entire frame.

Strumenti per Data Science

Da REPL...

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

A screenshot of a macOS terminal window titled 'stefania — python3 — 80x24'. The terminal shows the command 'python3' being executed, which displays the Python 3.6.3 version information and the prompt '>>>'. The user enters '3 + 1.0' and the prompt returns '4.0'. Then, the user enters ''seeds'.upper()' and the prompt returns ''SEEDS''. The terminal ends with another '>>>' prompt and a cursor.

```
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

IP[y]:
IPython



- 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*).

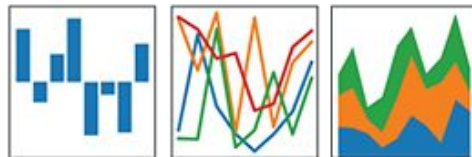


pandas

pandas, 2008, Wes McKinney, statistico

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

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



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 💖

The background of the slide is a dense, close-up photograph of fresh green basil leaves. The leaves are vibrant green, with prominent veins and a slightly wrinkled texture. They are packed closely together, filling the entire frame except for a white rectangular area on the left side where the text is located.

Visualizzare dati

Matplotlib

Matplotlib, 2003, John D. Hunter, neurobiologo

Strumento potente e leggero per le maggiori visualizzazioni statistiche.

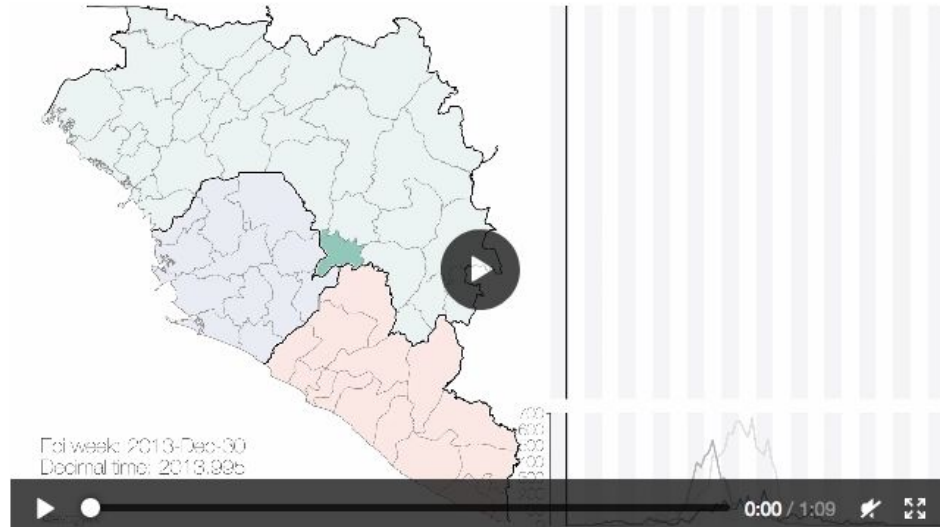


John Hunter Plotting Contest 2018

Winners

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

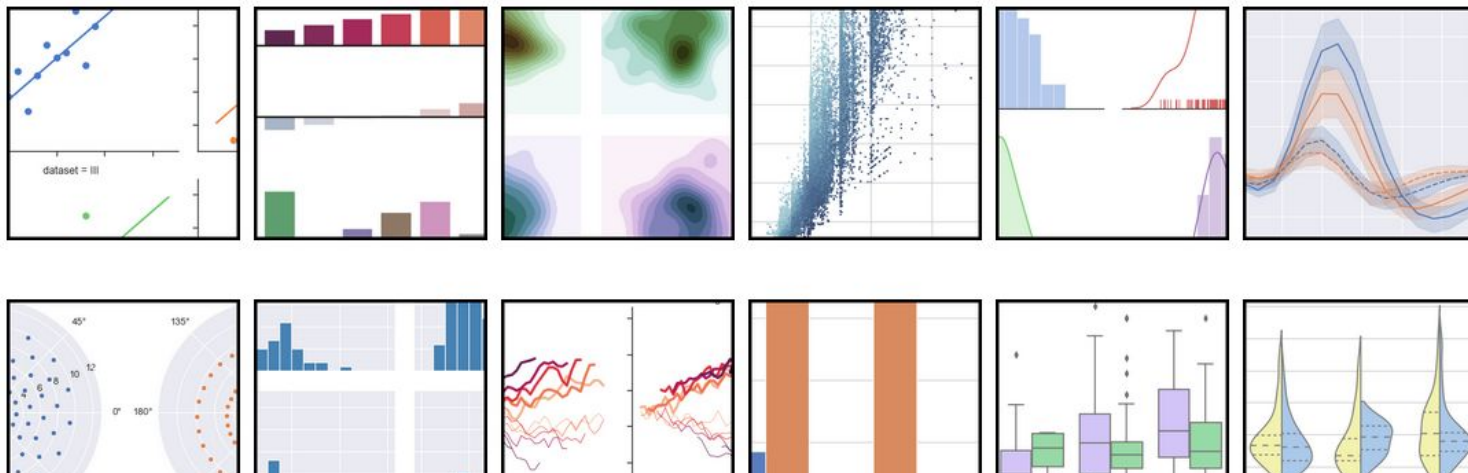
by Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Marc A. Suchard, Philippe Lemey, and Andrew Rambaut.



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, matplotlib 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

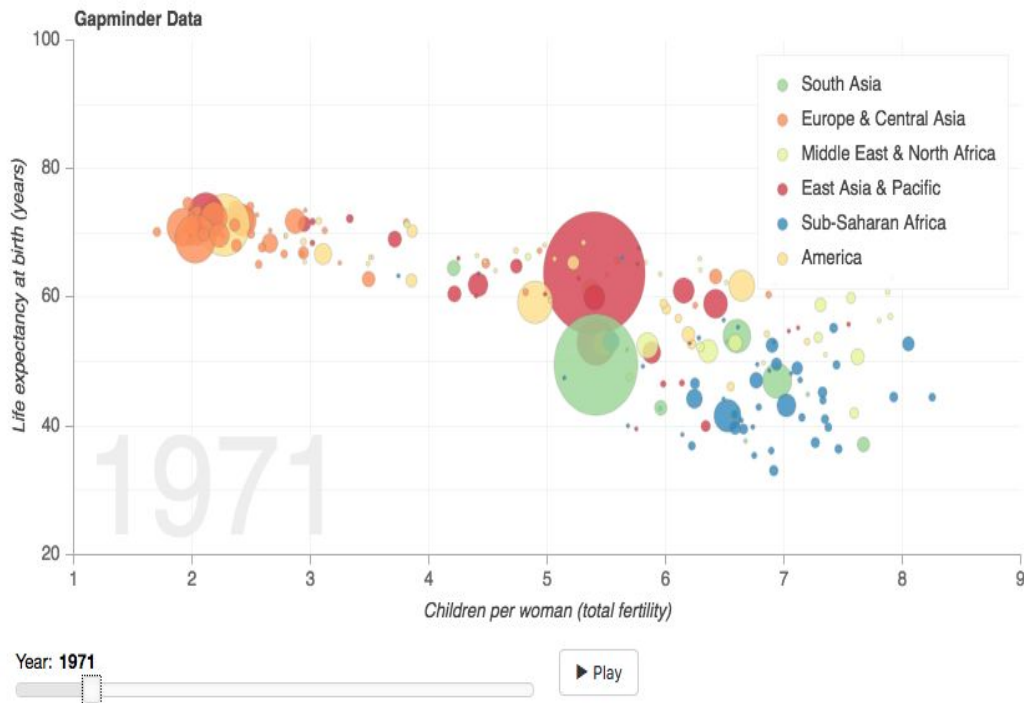
Bokeh, libreria per visualizzazioni interattive ottimizzata per rappresentazioni su web browser.

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



Bokeh

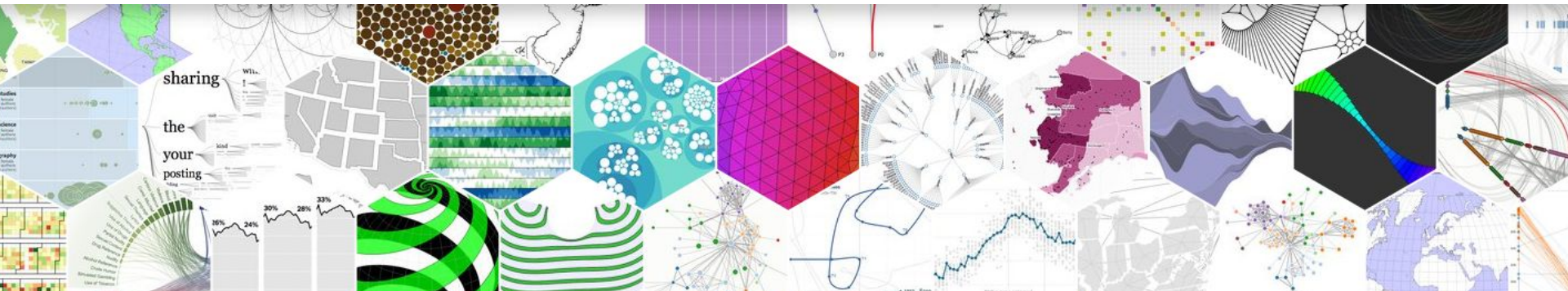
Riproduzione con Bokeh del
conosciuto TED talk di Hans
Rosling “The best stats
you've ever seen”



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.

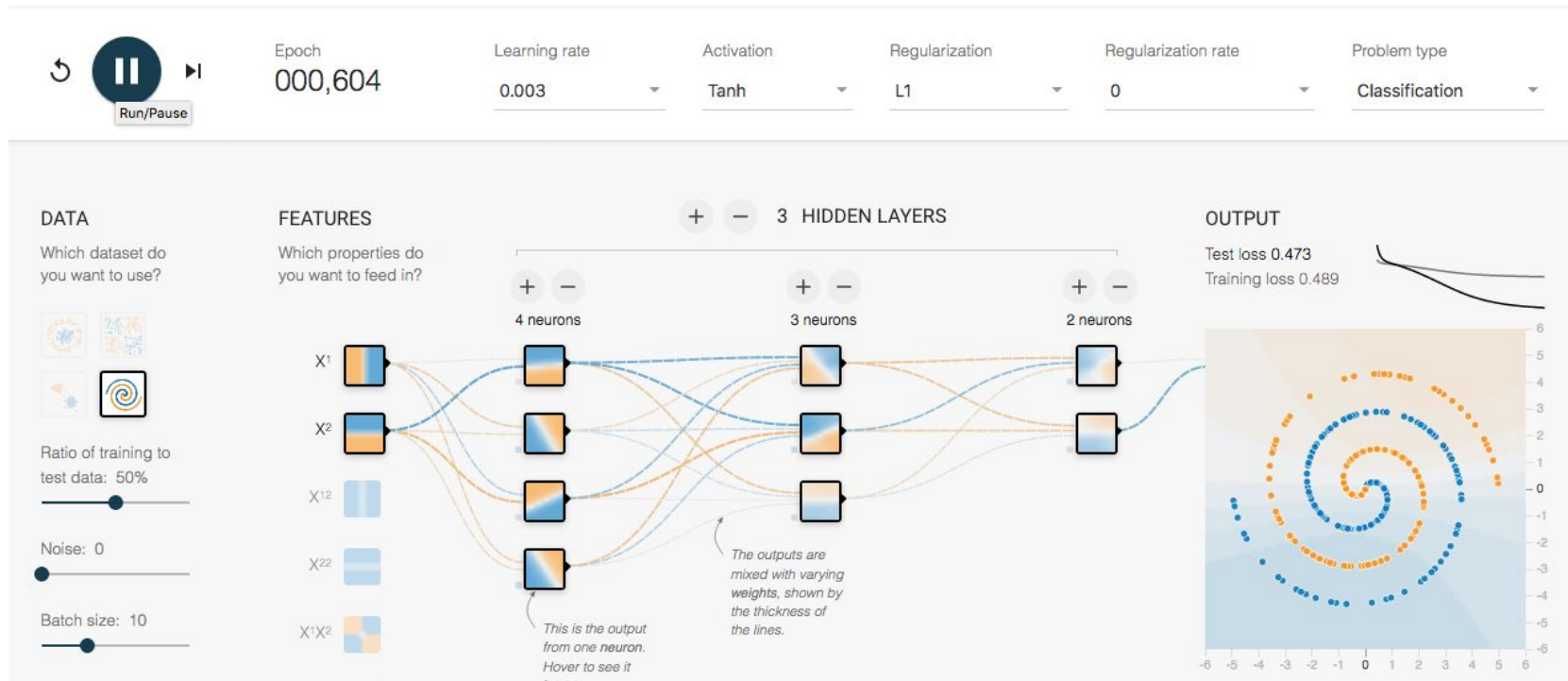
TensorFlow

TensorFlow, 2015, Google Brain

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



TensorFlow playground



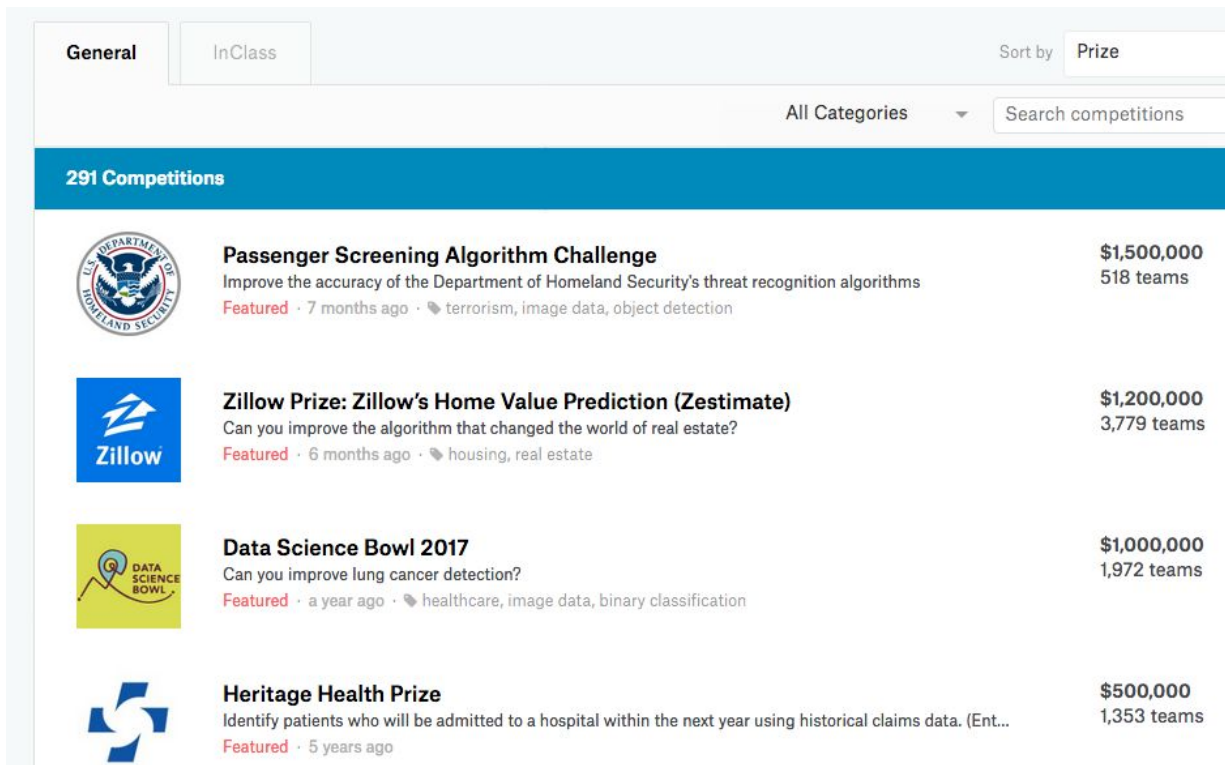
The background of the slide is a dense, close-up photograph of fresh green basil leaves. The leaves are vibrant green, have a slightly serrated edge, and are arranged in a repeating pattern of small, rounded clusters. A white rectangular box is positioned on the left side of the image, containing the main title and subtitle.

Trovare dati open





Alcuni esempi utili

Kaggle

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



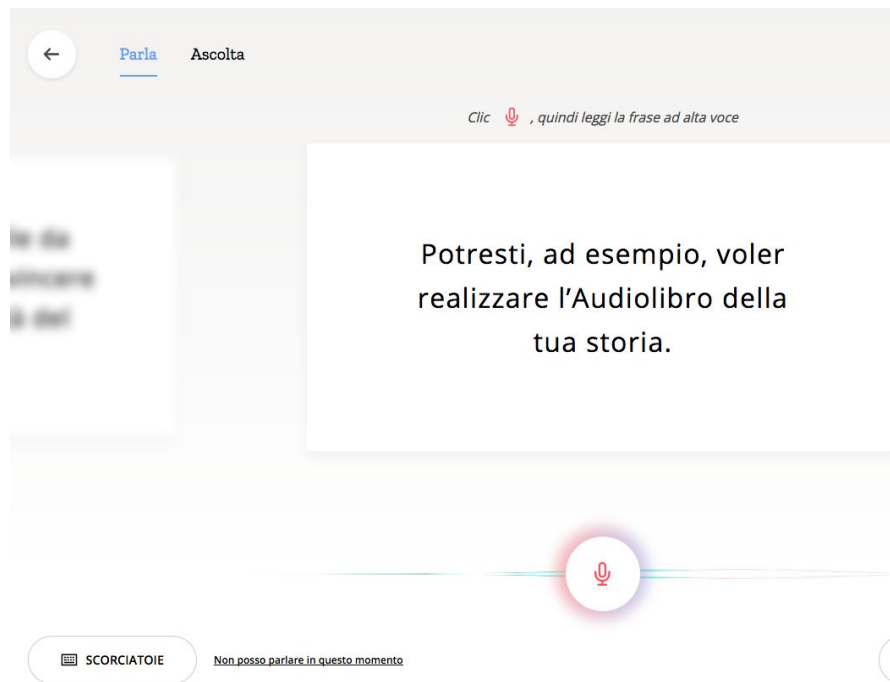
The screenshot shows the Kaggle website interface. At the top, there are tabs for 'General' and 'InClass', and a 'Sort by' dropdown menu set to 'Prize'. Below this is a search bar labeled 'Search competitions'. A blue banner indicates '291 Competitions'. The main content area lists four competitions, each with a logo, title, description, featured status, date, tags, prize amount, and number of teams.

Logo	Competition Title	Description	Featured	Date	Tags	Prize	Teams
	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...



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](https://www.linkedin.com/in/astrastefania)

twitter.com/astrastefania