Introduzione ai big data e ai metodi computazionali

Michele Tizzoni
Dipartimento di Sociologia e Ricerca Sociale
Via Verdi 26, Trento
Ufficio 6, 3 piano

UNIVERSITÀ DI TRENTO
FONDAZIONE BRUNO KESSLER

C2S2 Center for Computational Social Science and Human Dynamics


SAY BIG DATA
ONE MORE TIME
memegenerator.net

Chi sono

- Michele Tizzoni (RTD-B dal 2022)
- Email: michele.tizzoni@unitn.it (contattatemi per richieste di colloquio)
- Ufficio: n°6, piano 3, edificio di sociologia
- Aree di ricerca: computational social sciences, epidemiologia computazionale, reti complesse
- Insegnamenti AA 23/24
  - Introduzione ai big data e metodi computazionali (DSRS)
  - Digital epidemiology (DISI)

Obiettivi del corso

- leggere e comprendere la sintassi di un programma scritto in linguaggio Python
- utilizzare Jupyter notebooks
- impostare l'analisi di un dataset con la libreria Pandas
- comprendere il disegno di uno studio basato su metodi computazionali
- acquisire le competenze informatiche di base per la laurea magistrale in Data Science

Materiali

- Le lezioni si baseranno principalmente sui materiali sviluppati dal David Leoni per i seminari di credito Python degli anni passati
  - SoftPython: https://it.softpython.org/
- Le slides saranno disponibili sul mio **repository GitHub**
  - https://github.com/micheletizzoni/Python_for_social_sciences

Esame

- Modalità di esame:
  - **Esame scritto in laboratorio** con alcuni esercizi da risolvere, simili a quelli proposti durante il corso.
  - Alcune lezioni saranno interamente dedicate alla soluzione di esercizi in preparazione dell'esame.
  - L'esame sarà "open book", quindi sarà possibile consultare il libro *Softpython* durante la prova (senza collegamento a Internet)

Perché programmare nelle scienze sociali?

The big picture

- L’immagine digitale del mondo riproduce una copia sempre più fedele del mondo reale, fisico.
- Questo consente di usare metodi computazionali per misurare delle ricorrenze e definire dei nessi causali, attraverso metodi di statistica, machine learning, data mining
- Possiamo mettere in relazione le tracce digitali lasciate dall’attività umana di ogni giorno con misure empiriche ottenute con metodi tradizionali (indagini a campione, osservazioni sul campo)
- Possiamo sviluppare modelli predittivi del comportamento umano su larga scala e testarli quasi in tempo reale

[slide by C. Cattuto]


New York City
1.1 miliardi di chiamate taxi
6 anni
350 Gb di dati
toddwschneider.com

tracce digitali del comportamento umano

[slide by C. Cattuto]

Italia:
97% della popolazione possiede uno smartphone
84% della popolazione accede a internet
6+ ore al giorno online
2 ore al giorno su social media

[slide by C. Cattuto]

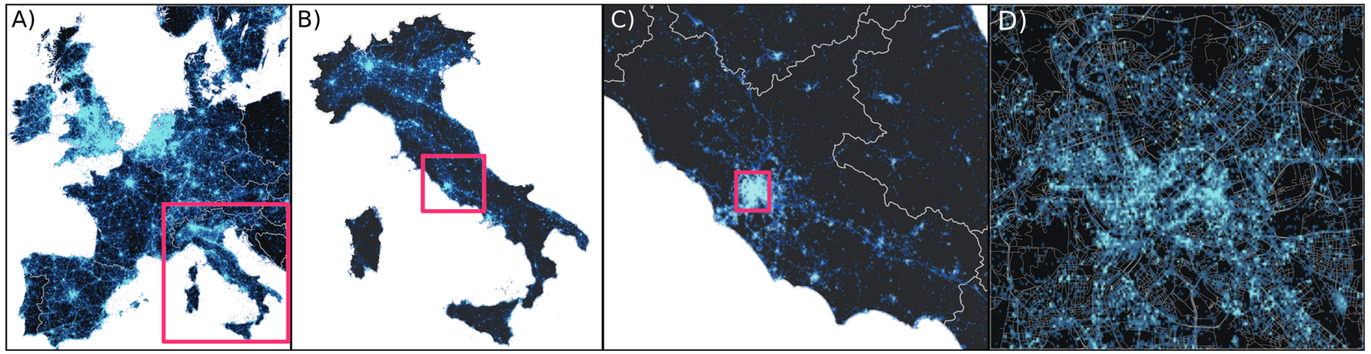
tracce digitali

disponibili come effetto collaterale di attività ordinarie
alto livello di copertura, accesso alle grandi scale,
possibilità di elaborazione automatica

prospettiva storica
orizzonte temporale limitato
riproducibilità limitata
contesto limitato
privacy e protezione dei dati

[slide by C. Cattuto]

Demografia digitale



PLOS ONE

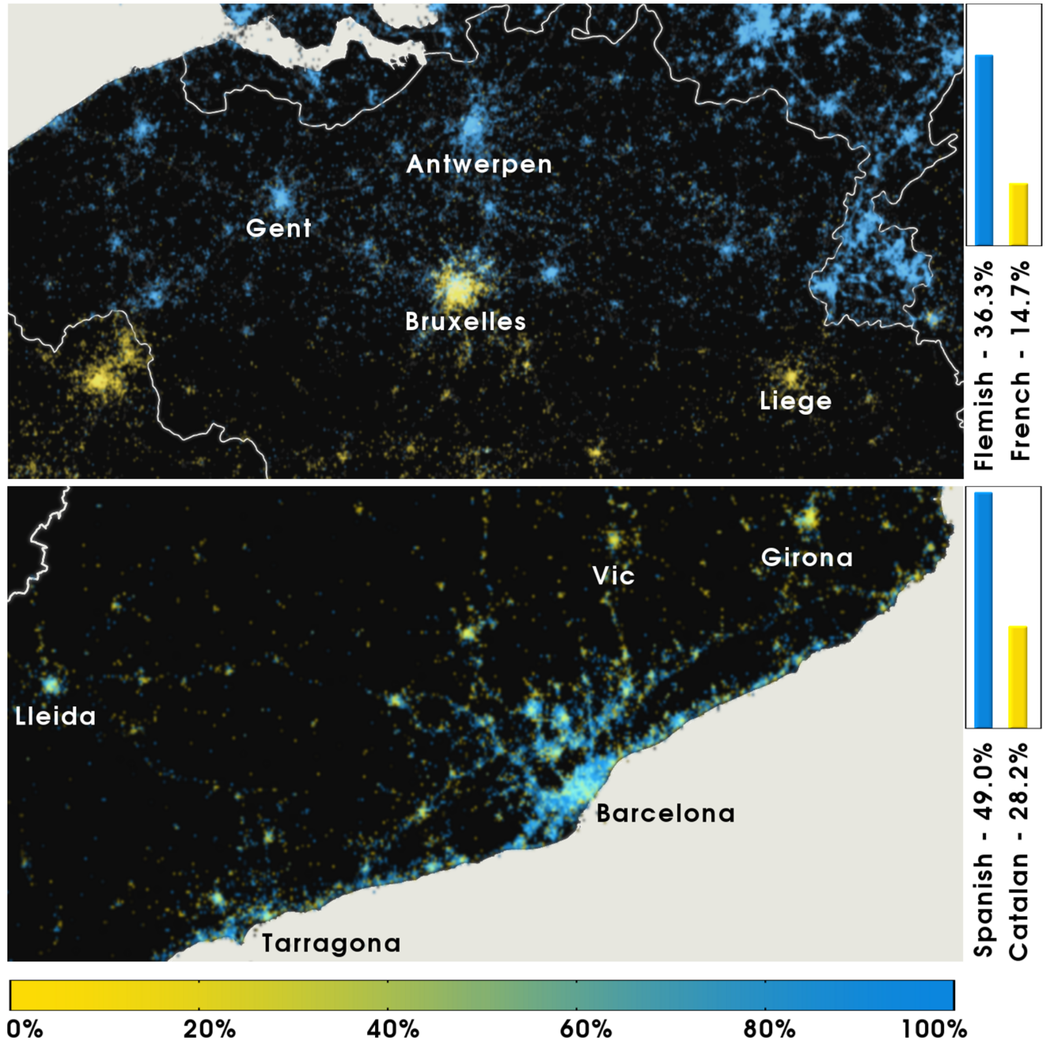
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

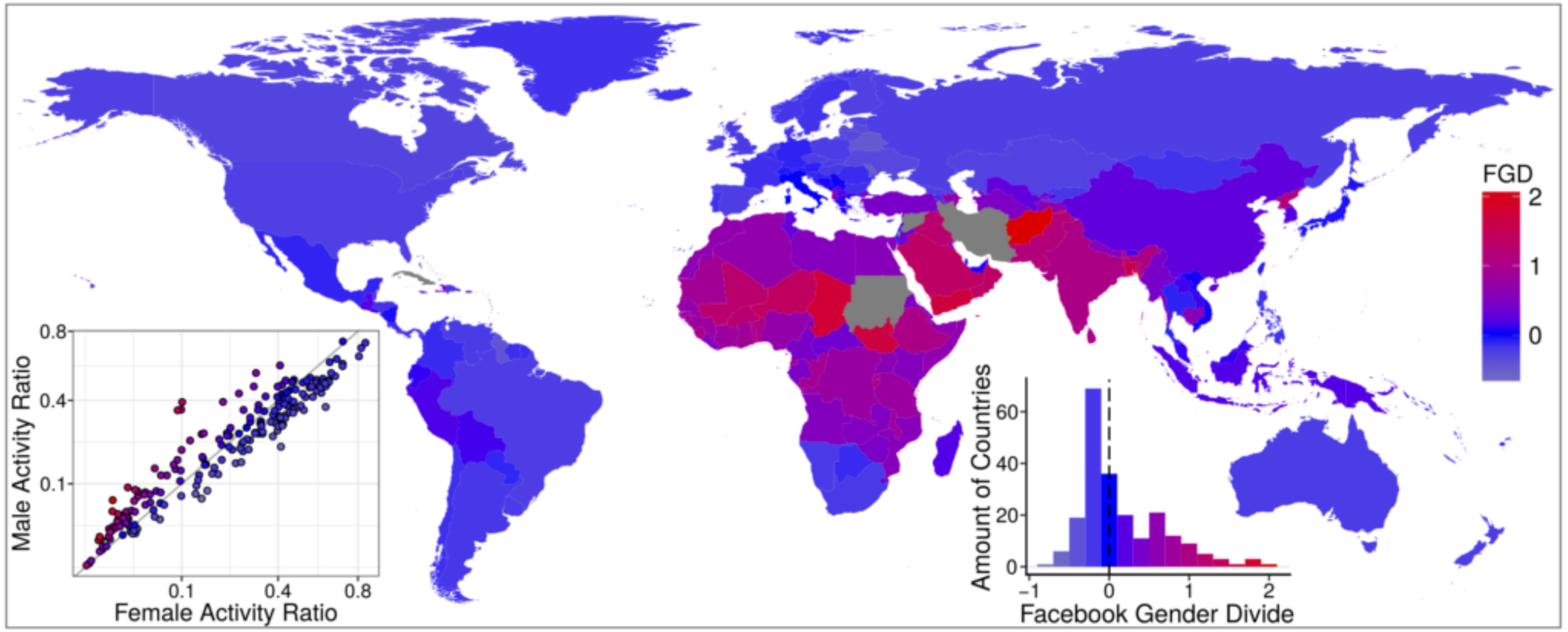
The Twitter of Babel: Mapping World Languages through Microblogging Platforms

Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, Alessandro Vespignani

Published: April 18, 2013 • https://doi.org/10.1371/journal.pone.0061981

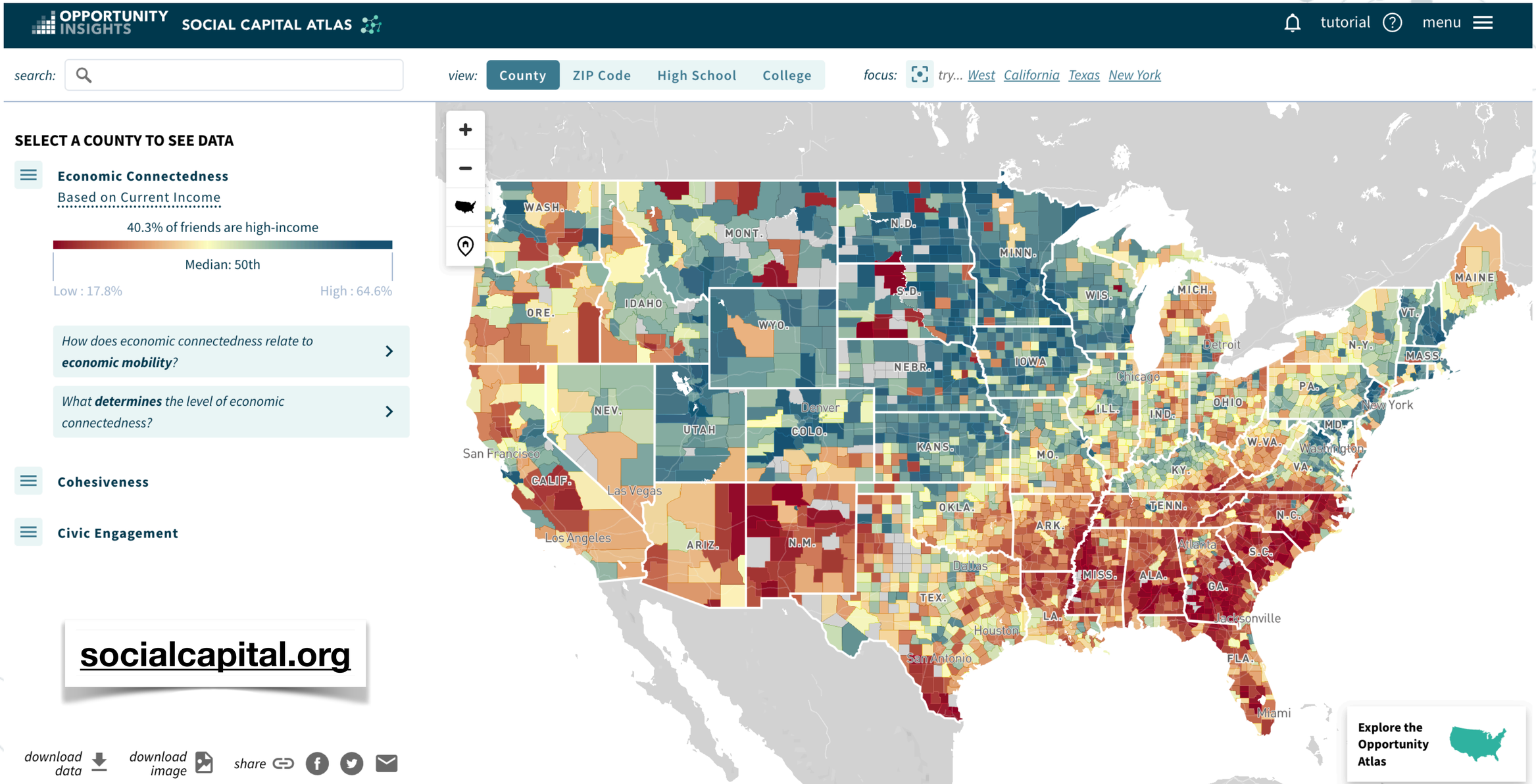


Disuguaglianze di genere

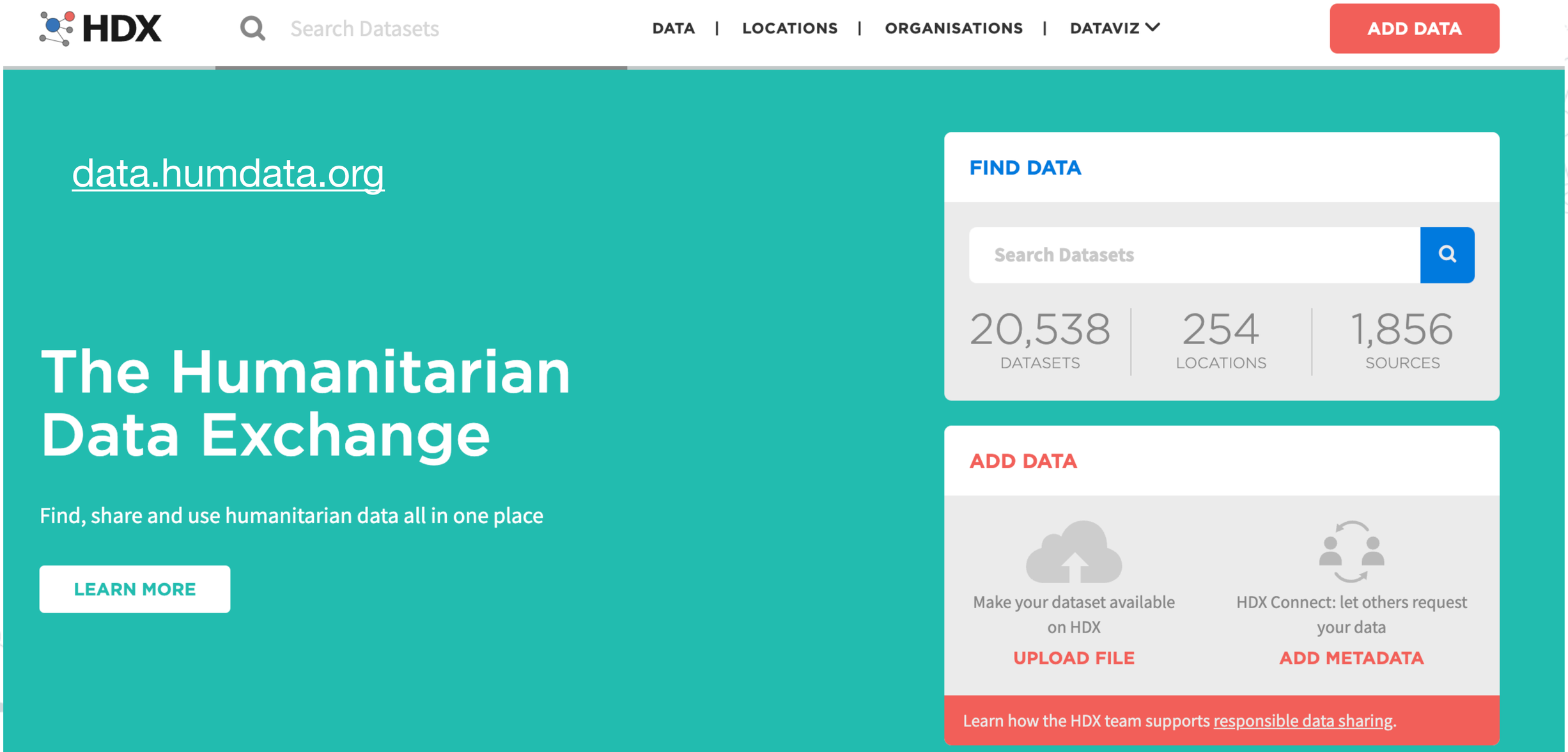


Garcia et al. PNAS 2018

Capitale sociale



Dati nel settore umanitario



La pandemia COVID-19

- Per la prima volta nella storia, durante una pandemia, è stato possibile osservare e misurare il comportamento delle persone su larga scala attraverso l’analisi tracce digitali
- Analisi della mobilità umana con dati di telefonia mobile, geolocalizzati
- Analisi della diffusione di informazione e mis-informazione su social media
- Misura dei cambiamenti nei comportamenti quotidiani (sonno/veglia, socialità) attraverso dati da piattaforme digitali

Mobilità

See how your community is moving around differently due to COVID-19

As global communities respond to COVID-19, we've heard from public health officials that the same type of aggregated, anonymized insights we use in products such as Google Maps could be helpful as they make critical decisions to combat COVID-19.

These Community Mobility Reports aim to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential.

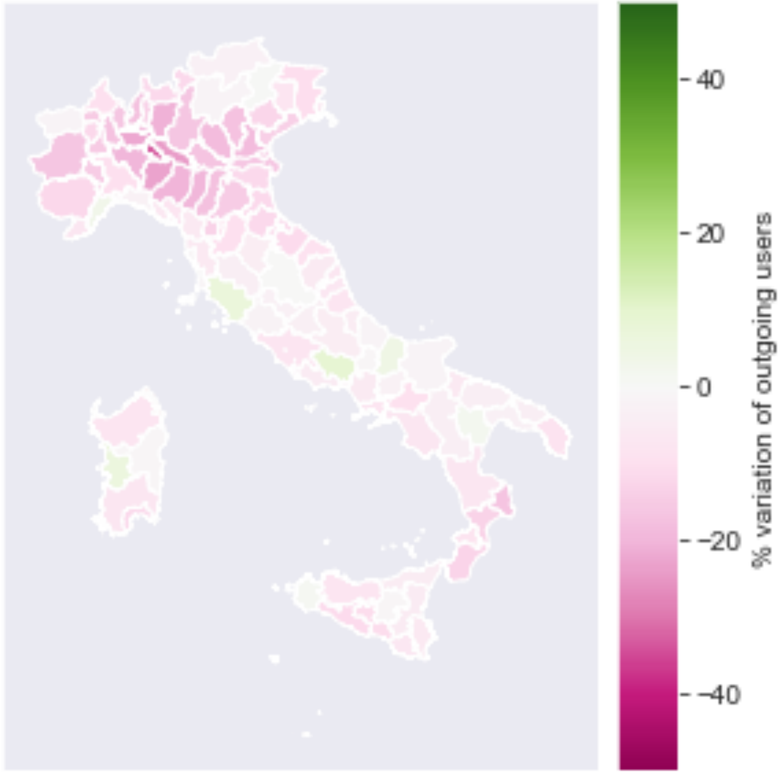
Maps

Mobility Trends Reports

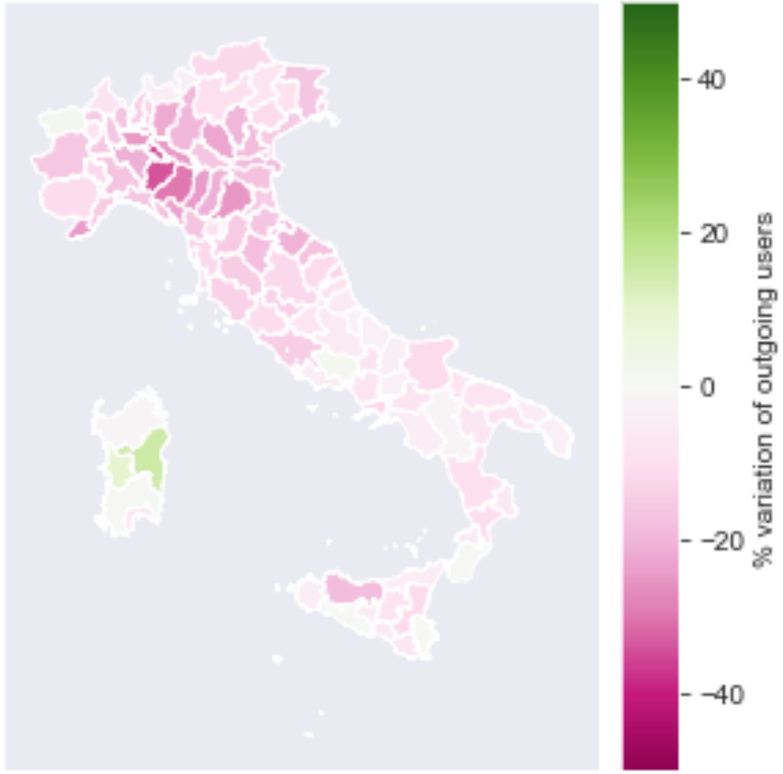
Learn about COVID-19 mobility trends. Reports are published daily and reflect requests for directions in Apple Maps. Privacy is one of our core values, so Maps doesn't associate your data with your Apple ID, and Apple doesn't keep a history of where you've been.

cuebiq

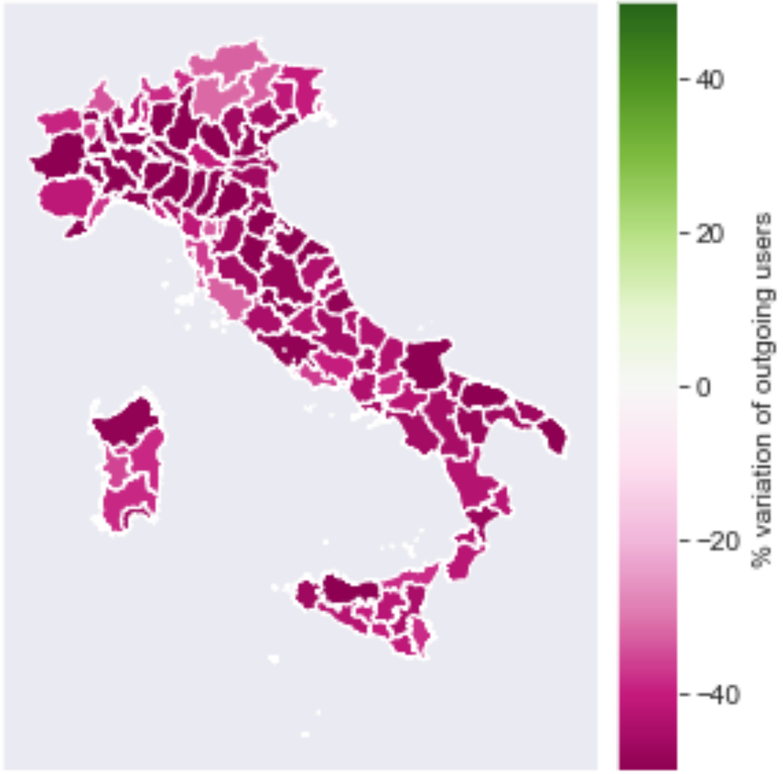
Mobilità



February 21-28

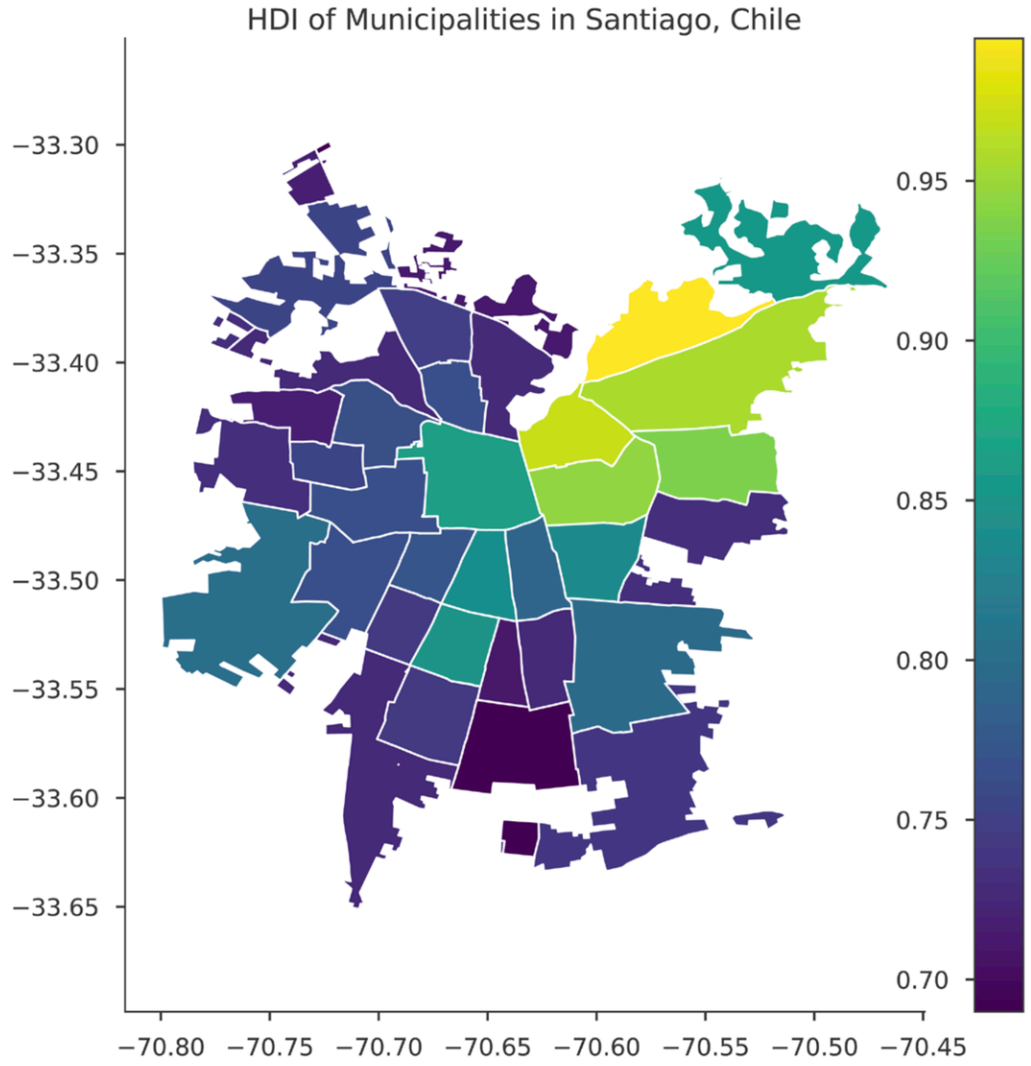


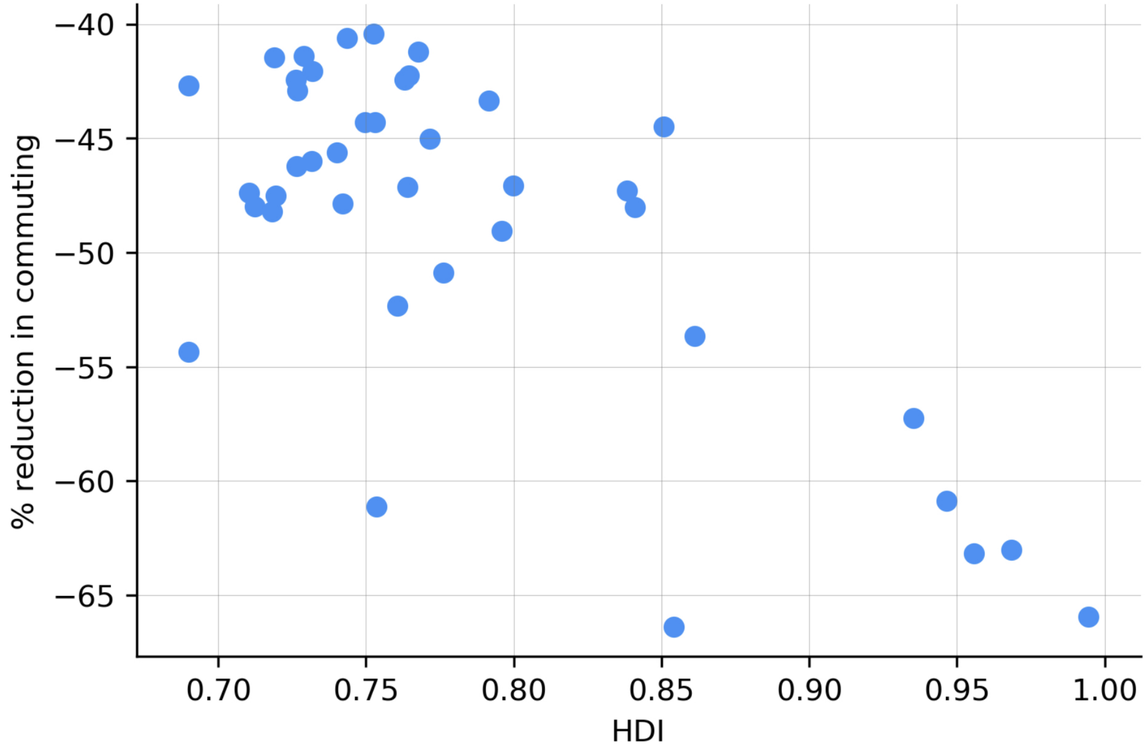
February 29 - March 6



March 7 - 13

Effetti socio-economici





Breve intro a Python

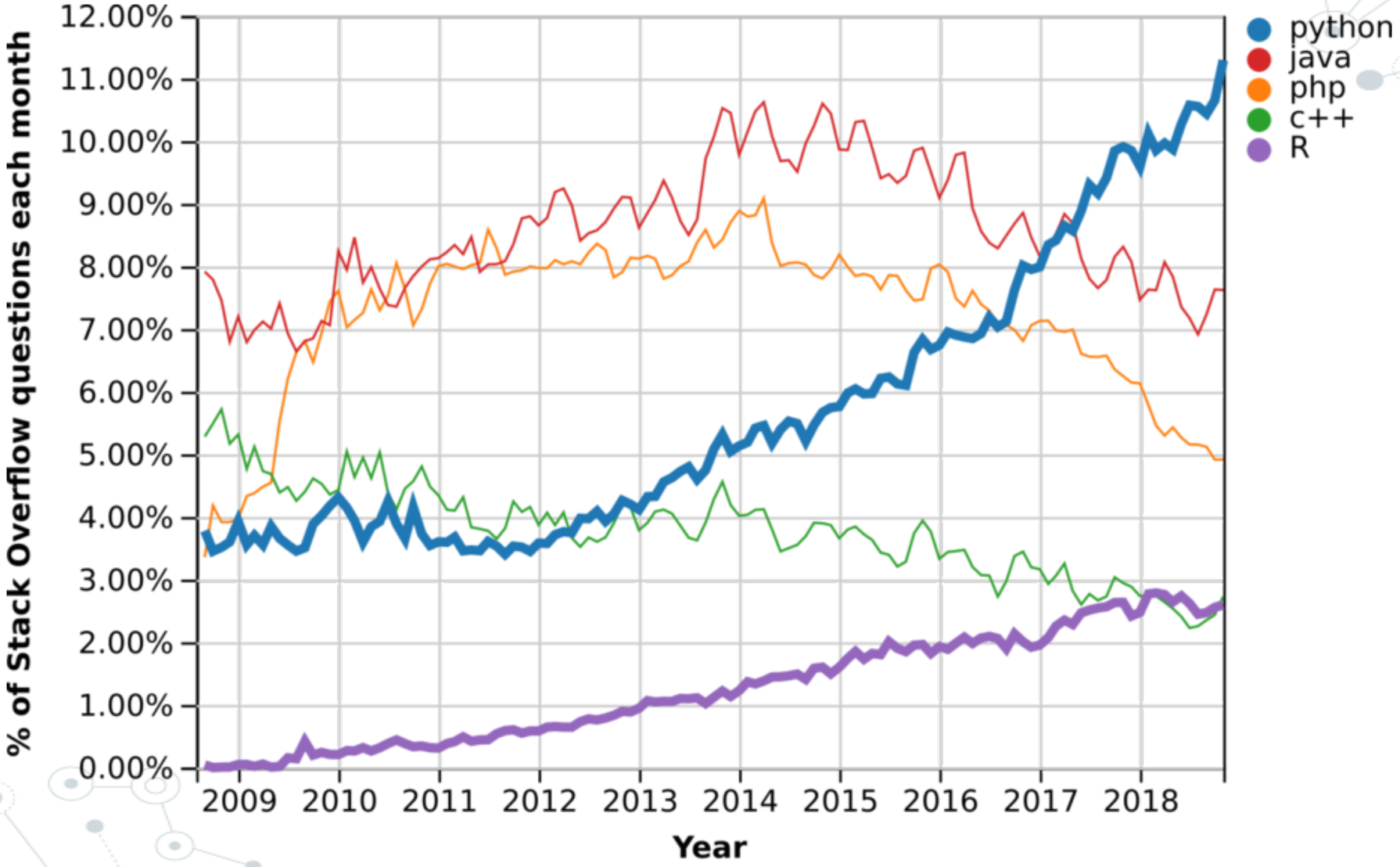
Un po’ di storia

- Python è un linguaggio di programmazione di alto livello, ideato da Guido van Rossum
- 1991 - Python 1.0 discontinued
- 1995 - Guido van Rossum proclamato Benevolent Dictator for Life (BDFL)
- 2000 - Python 2.0 End-of-life: 2020
- 2008 - Python 3.0 Current version 3.11
- 2018 - Guido abbandona la carica di BDFL

The 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.

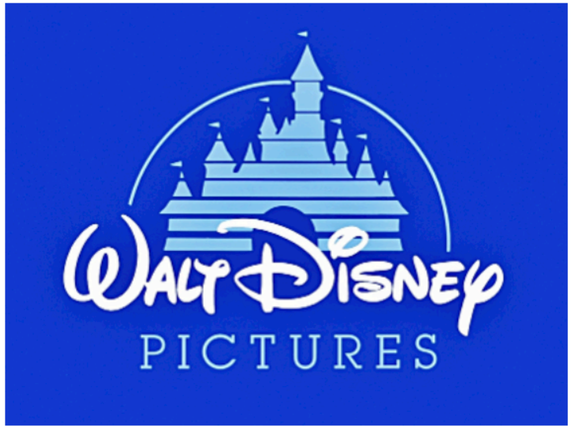
Python trends

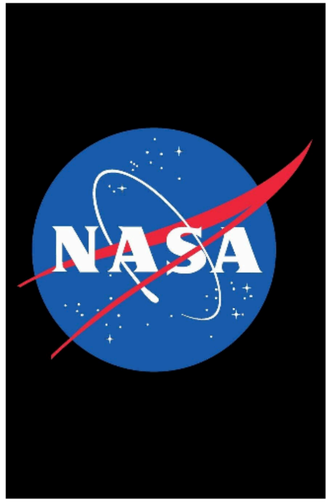


Source: https://insights.stackoverflow.com/trends

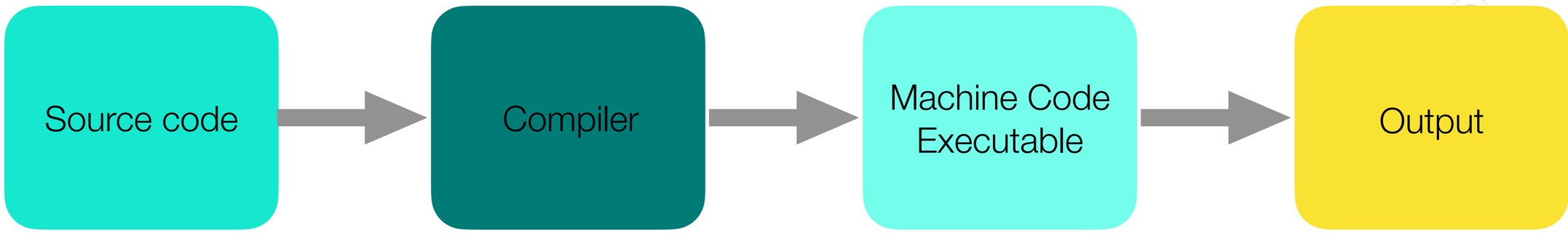
Chi usa Python?

- Qualsiasi azienda di grandi dimensioni e che abbia dei dati da gestire usa Python
- Python è il linguaggio standard nel campo dell'IA



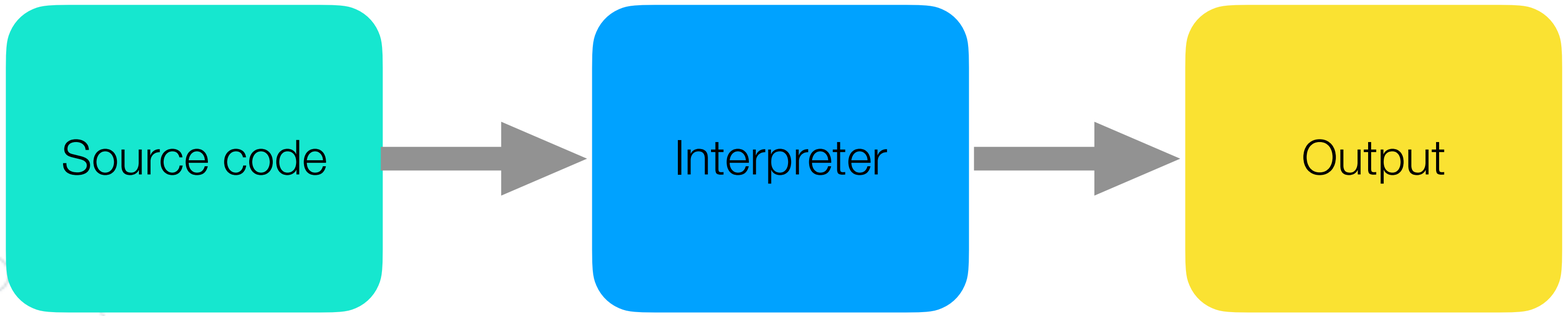


Interpreted vs compiled



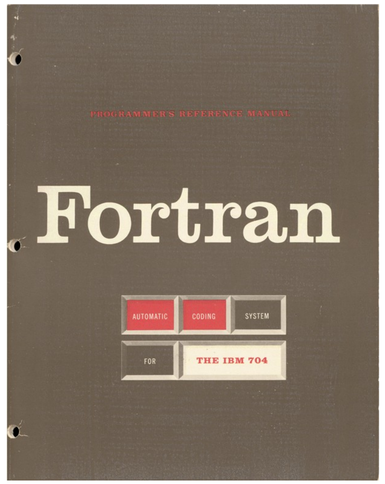
Linguaggi compilati

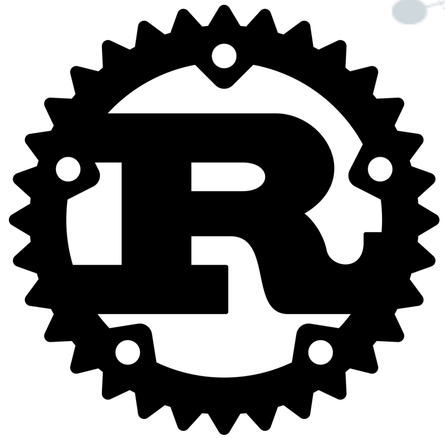
Linguaggi interpretati



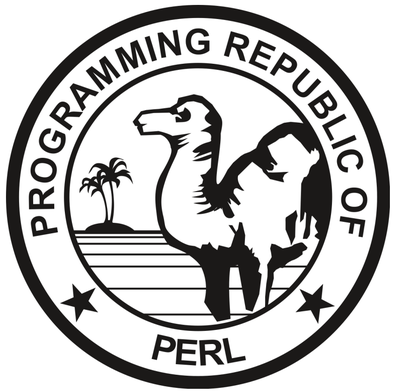
Interpreted vs compiled

Linguaggi compilati





Linguaggi interpretati



Interpreted vs compiled

Compile	Interpreter
Prende come input un intero programma	Prende come input ogni riga di codice, una per una
Genera un oggetto codice intermedio	Non genera un oggetto codice intermedio
Più rapido in esecuzione	Più lento in esecuzione
Debugging più difficile	Debugging più semplice

Python vs R

- General purpose (Web development, etc.)
- In genere, più versatile.
- Più semplice da imparare per beginners.
- Più orientato verso il machine learning, deep learning.

- Molto popolare nelle scienze sociali, economia, biologia.
- Librerie statistiche molto avanzate.
- Ottimo per visualizzazioni.
- Più orientato verso la modellizzazione statistica.

- ChatGPT è un ottimo assistente nel coding

- Per ottenere delle risposte soddisfacenti da ChatGPT è fondamentale conoscere la sintassi del linguaggio che stiamo usando (Python) e formulare una richiesta in modo corretto

- E’ fondamentale conoscere il linguaggio che stiamo usando per valutare la qualità delle risposte

Installazione

ANACONDA®

https://www.anaconda.com/download

jupyter

INSTALL PROJECT DOCUMENTATION BLOG DONATE

jupyter

Jupyter Notebook

The Jupyter Notebook is a web-based interactive computing platform that allows users to author data- and code-driven narratives that combine live code, equations, narrative text, visualizations, interactive dashboards and other media.