CERN Science for Open Data (CS4OD)

CERN openlab Technical Workshop 2021

Anna Ferrari CERN openlab,

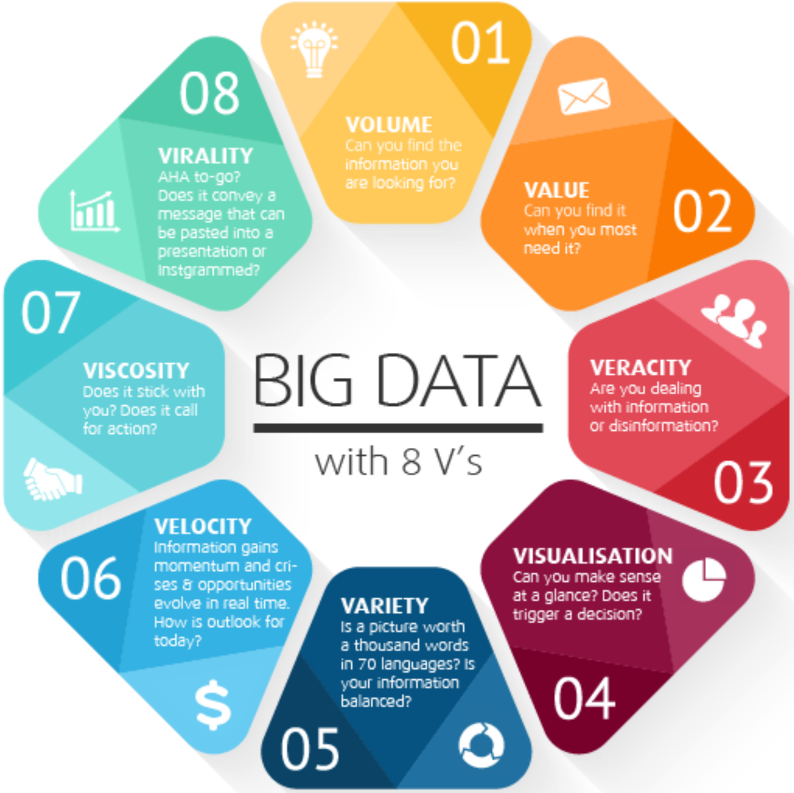
Ivan Knezevic CERN openlab, Alex Ioannidis IT-CDA, José B. G. Lopez IT-CDA

10/03/2021

The Big Data Challenge

Data size

- **Data size** is huge and of high dimensionality
- Data heterogeneity
- Data analysis
- Data overload



BIG DATA

with 8 V's

01 **VOLUME** Can you find the information you are looking for?

02 **VALUE** Can you find it when you most need it?

03 **VERACITY** Are you dealing with information or disinformation?

04 **VISUALISATION** Can you make sense at a glance? Does it trigger a decision?

05 **VARIETY** Is a picture worth a thousand words in 70 languages? Is your information balanced?

06 **VELOCITY** Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?

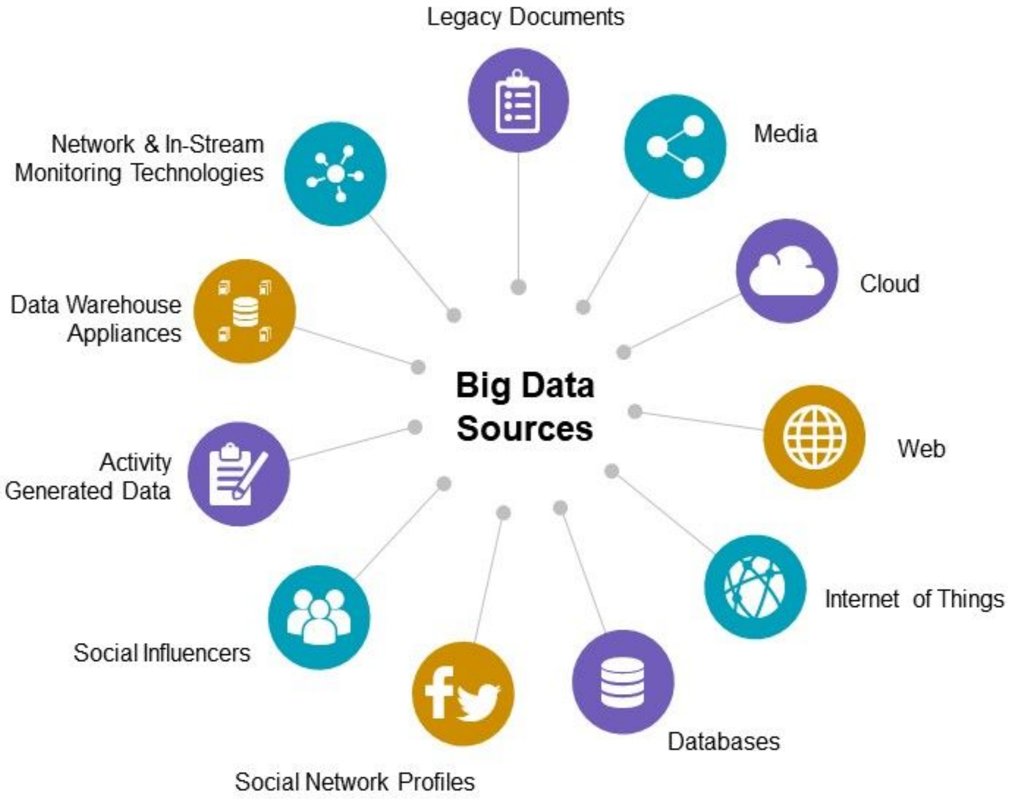
07 **VISCOSITY** Does it stick with you? Does it call for action?

08 **VIRALITY** AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

The Sources Heteogeneity

Data heterogeneity

- **Data size** is huge and of high dimensionality
- **Data heterogeneity** in terms of sources, acquisition, and storage
- **Data analysis**
- **Data overload**



The Analysis Diversity

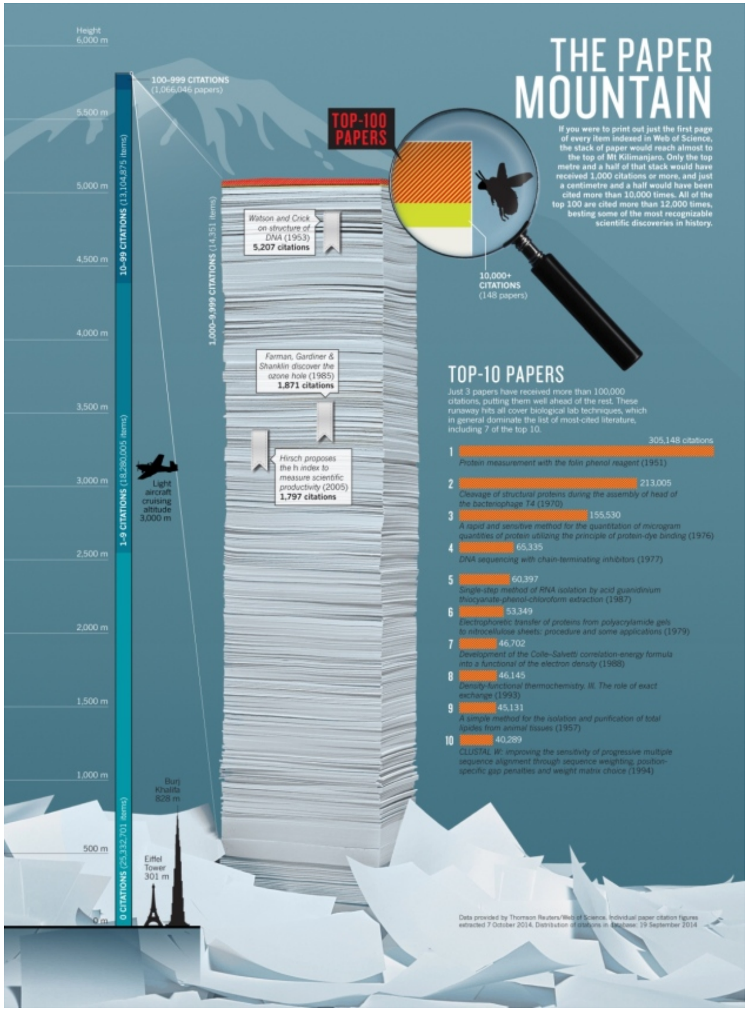
Data analysis

- **Data size** is huge and of high dimensionality
- **Data heterogeneity** in terms of sources, acquisition, and storage
- **Data analysis** differencies in terms of assumptions, models and methods
- **Data overload**

Information vs Knowledge

Data overload

- **Data size** is huge and of high dimensionality
- **Data heterogeneity** in terms of sources, acquisition, and storage
- **Data analysis** differencies in terms of assumptions, models and methods
- **Data overload** and excess of results



THE PAPER MOUNTAIN

If you were to print out just the first page of every item indexed in Web of Science, the stack of paper would reach almost to the top of Mt Kilimanjaro. Only the top metre and a half of that stack would have received 1,000 citations or more, and just a centimetre and a half would have been cited more than 10,000 times. All of the top 100 are cited more than 12,000 times, besting some of the most recognizable scientific discoveries in history.

TOP-10 PAPERS

Just 3 papers have received more than 100,000 citations, putting them well ahead of the rest. These runaway hits all cover biological lab techniques, which in general dominate the list of most-cited literature, including 7 of the top 10.

Height 6,000 m
5,500 m
5,000 m
4,500 m
4,000 m
3,500 m
3,000 m
2,500 m
2,000 m
1,500 m
1,000 m
500 m
0 m
100–999 CITATIONS (1,066,046 papers)
10–99 CITATIONS (13,104,875 items)
1–9 CITATIONS (18,280,005 items)
0 CITATIONS (25,332,701 items)
1,000–9,999 CITATIONS (14,351 items)
TOP-100 PAPERS
10,000+ CITATIONS (148 papers)
Watson and Crick on structure of DNA (1953) 5,207 citations
Farman, Gardiner & Shanklin discover the ozone hole (1985) 1,871 citations
Hirsch proposes the h index to measure scientific productivity (2005) 1,797 citations
Light aircraft cruising altitude 3,000 m
Burj Khalifa 828 m
Eiffel Tower 301 m
305,148 citations
1 Protein measurement with the folin phenol reagent (1951)
2 213,005 Cleavage of structural proteins during the assembly of head of the bacteriophage T4 (1970)
3 155,530 A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding (1976)
4 65,335 DNA sequencing with chain-terminating inhibitors (1977)
5 60,397 Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction (1987)
6 53,349 Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications (1979)
7 46,702 Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density (1988)
8 46,145 Density-functional thermochemistry. III. The role of exact exchange (1993)
9 45,131 A simple method for the isolation and purification of total lipides from animal tissues (1957)
10 40,289 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice (1994)

Data provided by Thomson Reuters/Web of Science. Individual paper citation figures extracted 7 October 2014. Distribution of citations in database: 19 September 2014

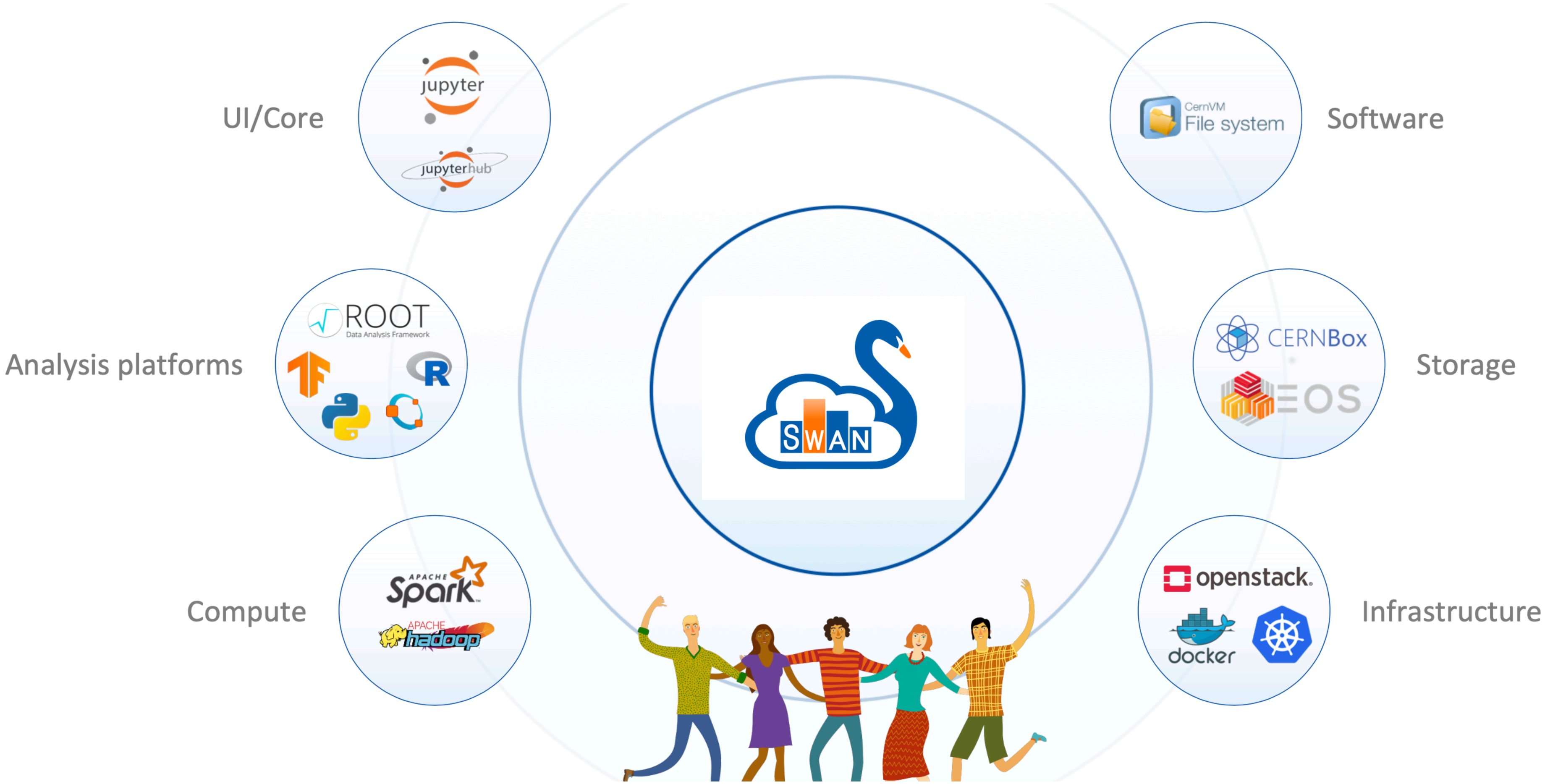
Urgent needs

Overcome barriers

- **Data size**: overcome barriers related to data governance and storage defining **common principles**
- **Data heterogeneity**: overcome barriers of data access defining a **global coordination** of **open data from multi-domain fileds**
- **Data analysis**: overcome barriers of analysis diversity defining **common pipelines and approaches**
- **Data overload**: overcome barriers of excess of information by complying with **results reproducibility and mutli-disciplinary expertises exchange**

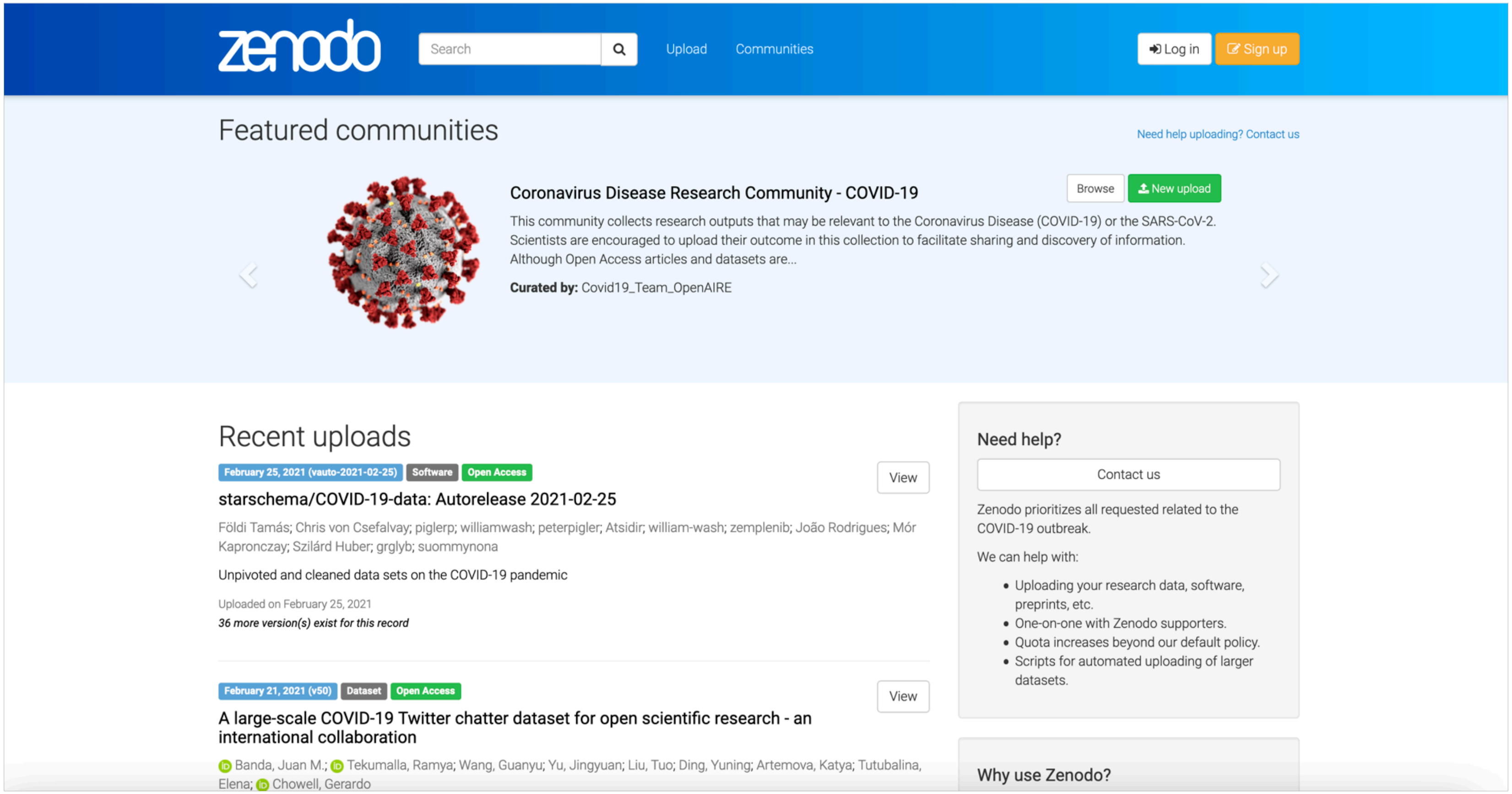
Swan for Data Management

CERN technologies, softwares, tools, infrastructures



Zenodo as Data repository

CERN technologies for data size and heterogeneity



Featured communities

Coronavirus Disease Research Community - COVID-19

This community collects research outputs that may be relevant to the Coronavirus Disease (COVID-19) or the SARS-CoV-2. Scientists are encouraged to upload their outcome in this collection to facilitate sharing and discovery of information. Although Open Access articles and datasets are...

Curated by: Covid19_Team_OpenAIRE

Recent uploads

February 25, 2021 (vauto-2021-02-25) | Software | Open Access

starschema/COVID-19-data: Autorelease 2021-02-25

Földi Tamás; Chris von Csefalvay; piglerp; williamwash; peterpigler; Atsidir; william-wash; zemplenib; João Rodrigues; Mór Kapronczay; Szilárd Huber; grglyb; suommynona

Unpivoted and cleaned data sets on the COVID-19 pandemic

Uploaded on February 25, 2021

36 more version(s) exist for this record

February 21, 2021 (v50) | Dataset | Open Access

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo

Need help?

Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:

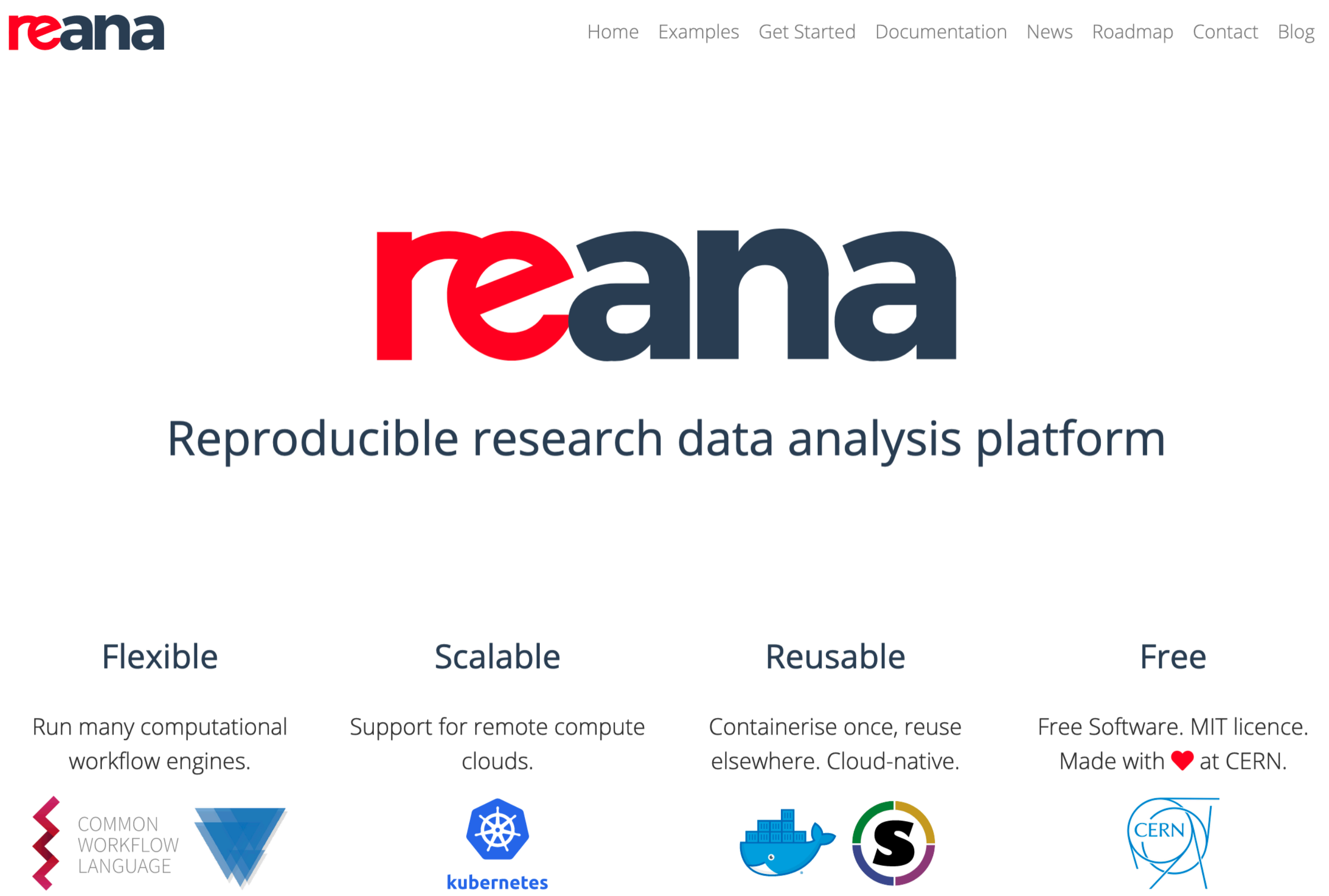
- Uploading your research data, software, preprints, etc.
- One-on-one with Zenodo supporters.
- Quota increases beyond our default policy.
- Scripts for automated uploading of larger datasets.

Why use Zenodo?

Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artifacts.

Reana for results reproducibility

CERN technologies for analysis pipelines definition and results reproducibility



Reproducible research data analysis platform

Flexible

Run many computational workflow engines.

Scalable

Support for remote compute clouds.

Reusable

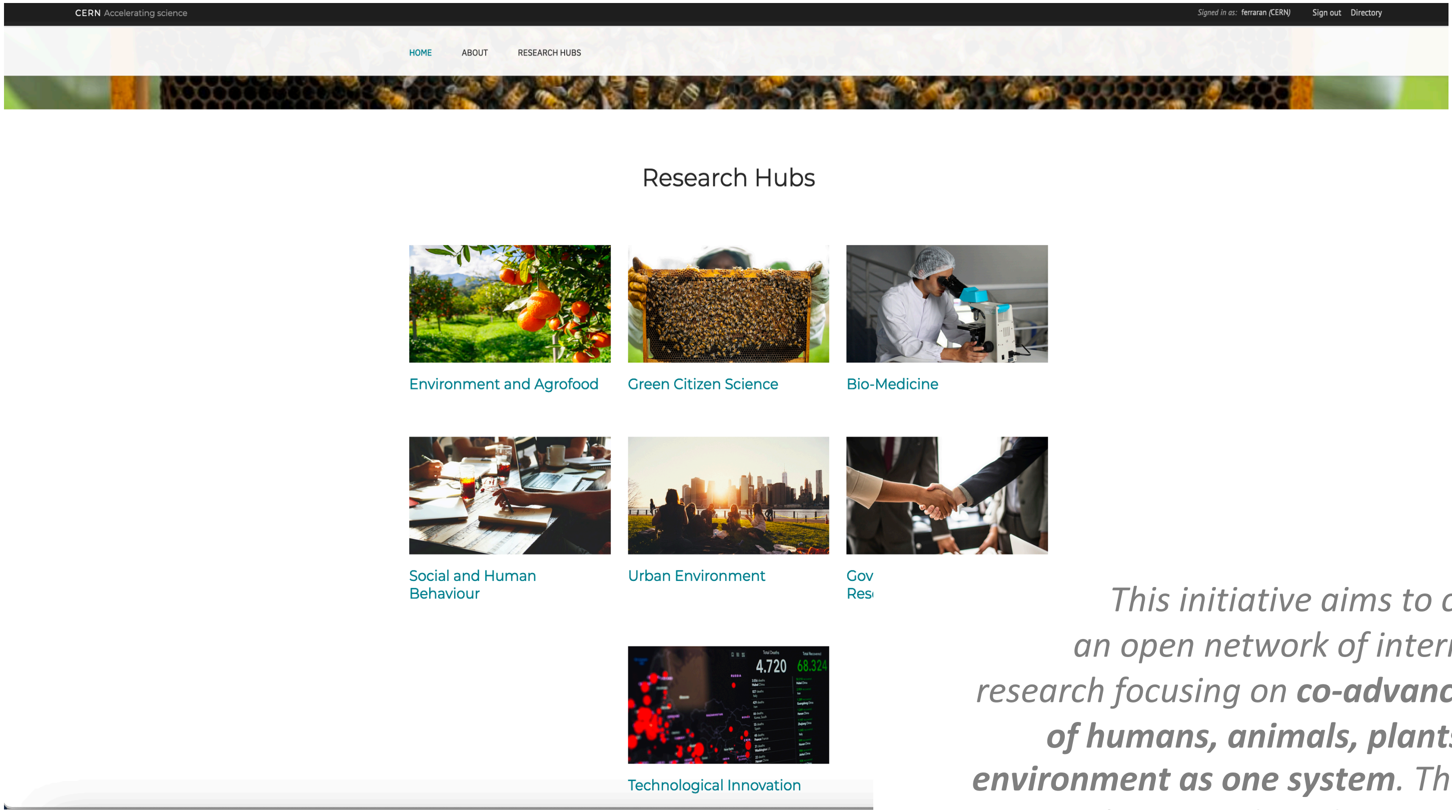
Containerise once, reuse elsewhere. Cloud-native.

Free

Free Software. MIT licence. Made with ❤ at CERN.

Circular Health

The community of multi-disciplinarity



CERN Accelerating science

Signed in as: ferraran (CERN) Sign out Directory

HOME ABOUT RESEARCH HUBS

Research Hubs

Environment and Agrofood

Green Citizen Science

Bio-Medicine

Social and Human Behaviour

Urban Environment

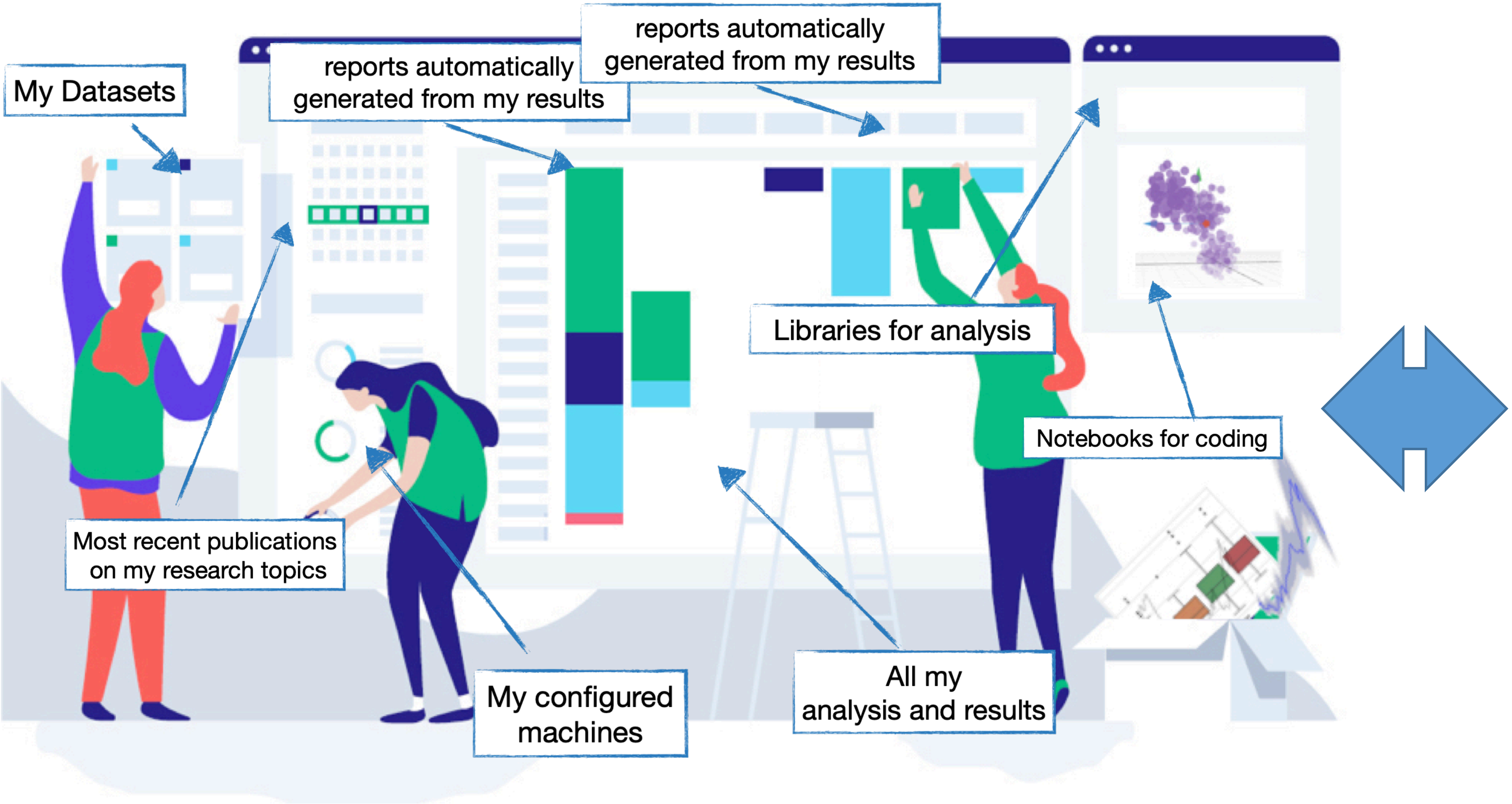
Gov
Res

Technological Innovation

This initiative aims to create an open network of international research focusing on ***co-advancing the health of humans, animals, plants, and the environment as one system****. The main focus is to explore new data driven approaches to funnel research towards the* ***convergence of health into a circular system****.*

CS4OD project

Cross-community platform



My Datasets

reports automatically generated from my results

reports automatically generated from my results

Libraries for analysis

Notebooks for coding

Most recent publications on my research topics

My configured machines

All my analysis and results

Share results

Platform layers

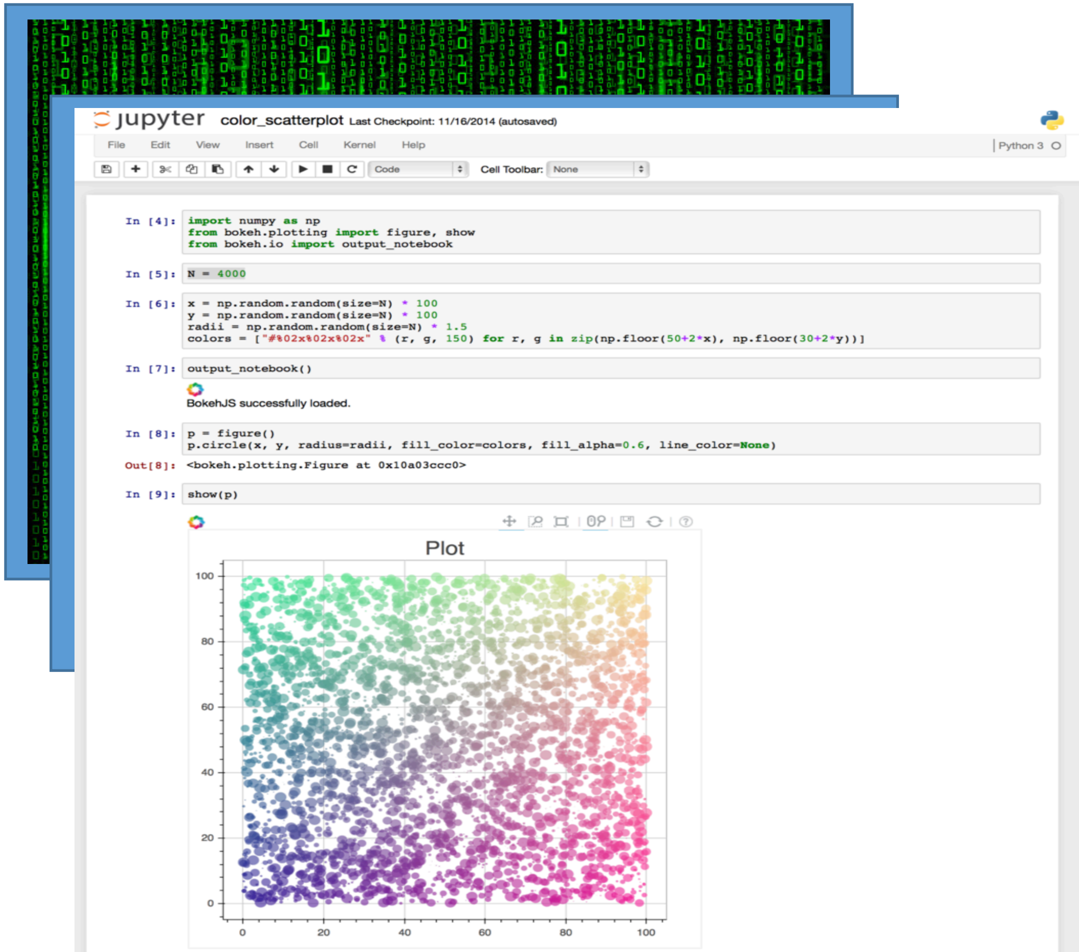
Low level layer

Data Scientist/ Data Engineer:

Data storage,
Homogeneization of data,
Define analysis pipelines,
…

Platform layers

Middle level layer



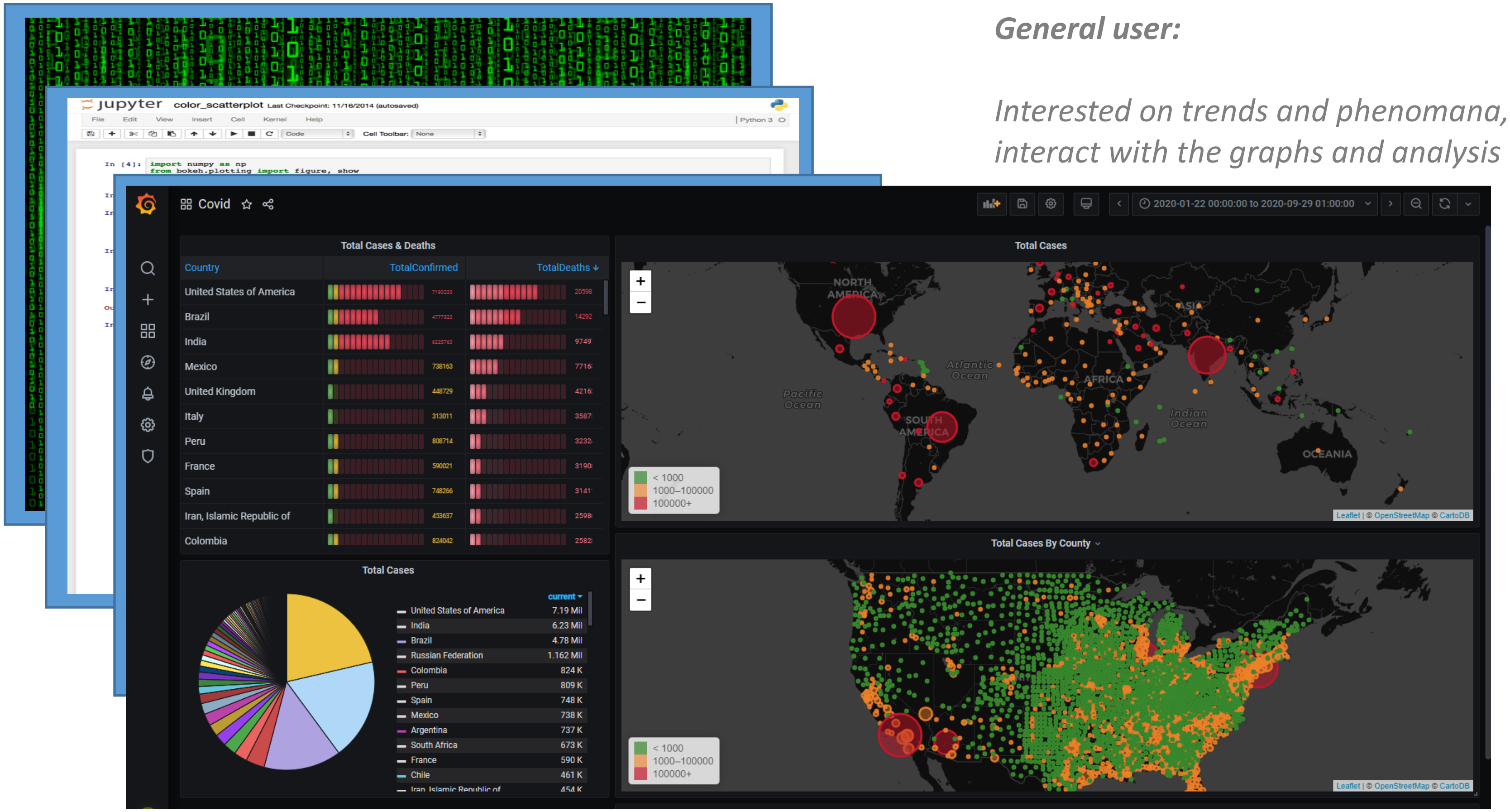
```
In [4]: import numpy as np
        from bokeh.plotting import figure, show
        from bokeh.io import output_notebook

In [5]: N = 4000

In [6]: x = np.random.random(size=N) * 100
        y = np.random.random(size=N) * 100
        radii = np.random.random(size=N) * 1.5
        colors = ["#%02x%02x%02x" % (r, g, 150) for r, g in zip(np.floor(50+2*x), np.floor(30+2*y))]

In [7]: output_notebook()

        BokehJS successfully loaded.

In [8]: p = figure()
        p.circle(x, y, radius=radii, fill_color=colors, fill_alpha=0.6, line_color=None)

Out[8]: <bokeh.plotting.Figure at 0x10a03ccc0>

In [9]: show(p)
```

Researcher:

Use of libraries,
Computation of analysis

Platform layers

High level layer



jupyter color_scatterplot Last Checkpoint: 11/16/2014 (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Code
Cell Toolbar: None
Covid
2020-01-22 00:00:00 to 2020-09-29 01:00:00
Total Cases
NORTH AMERICA
SOUTH AMERICA
Atlantic Ocean
Pacific Ocean
AFRICA
ASIA
Indian Ocean
OCEANIA
< 1000
1000–100000
100000+
Leaflet | © OpenStreetMap © CartoDB
Total Cases By County

```
In [4]: import numpy as np
        from bokeh.plotting import figure, show
```

Total Cases & Deaths

Country	TotalConfirmed	TotalDeaths ↓
United States of America	7190230	20598
Brazil	4777522	14292
India	6225763	9749
Mexico	738163	7716
United Kingdom	448729	4216
Italy	313011	3587
Peru	808714	3232
France	590021	3190
Spain	748266	3141
Iran, Islamic Republic of	453637	2598
Colombia	824042	2582

Total Cases

	current ▾
United States of America	7.19 Mil
India	6.23 Mil
Brazil	4.78 Mil
Russian Federation	1.162 Mil
Colombia	824 K
Peru	809 K
Spain	748 K
Mexico	738 K
Argentina	737 K
South Africa	673 K
France	590 K
Chile	461 K
Iran, Islamic Republic of	454 K

General user:

Interested on trends and phenomana, can interact with the graphs and analysis

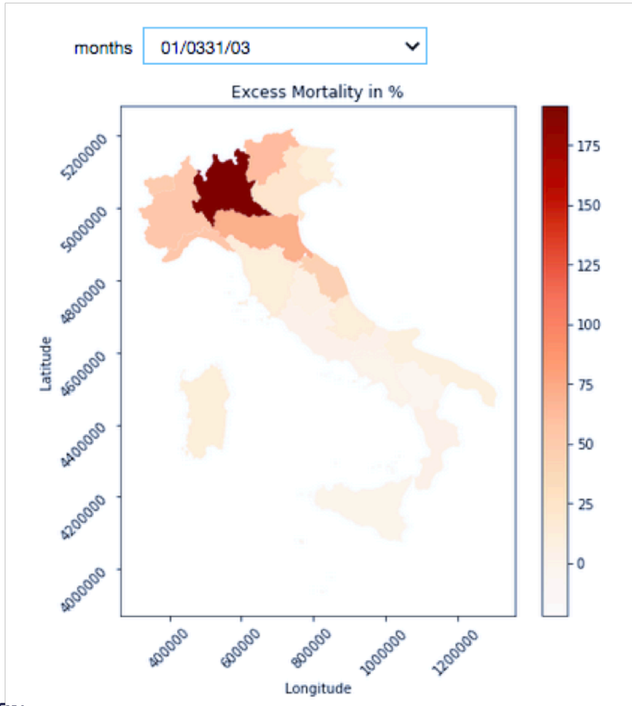
Anna Ferrari - a.ferrari@cern.ch

Use Case: Excess of Mortality

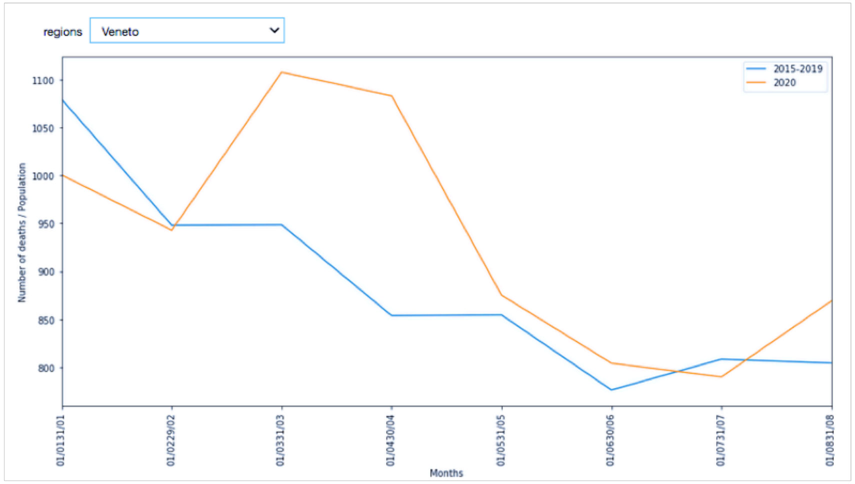
Collaboration with Bocconi University of Milan for implementing a **platform for multi-disciplinary data.**

Piloting on Italian data to estimate and represent graphically:

- All-cause death rates
- Excess mortality



months 01/0331/03



regions Veneto

Proof-of-Concept

Script

Next steps

- **Data Harmonization**: define a data format to be compatible with all european countries
- **Data Flexibility**: increase the flexibility of the functionalities in terms of data management
- **Analysis Flexibility**: increase the flexibility of the functionalities related to the analysis defining new libraries

Thank you

Questions?