



# CERN Science for Open Data (CS4OD)

*CERN openlab ExaHealth 2021*

Anna Ferrari\*, Ivan Knezevic\*, Diamantis Patsidis\*, Alberto Di Meglio\*

Alexandros Ioannidis#, Ines P. P. Da Cruz#, Nihal E. Yuceturk#, José B. G. Lopez#

Tomas Roun#, Tim Smith#

\*CERN-IT openlab

# CERN-IT-CDA

18/10/2021

1

# The Big Data Challenge

## Data size

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



Source: <https://www.m-brain.com/technology/>

# 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** differences in terms of assumptions, models and methods
- **Data overload**





# Urgent needs

## *Overcome barriers*

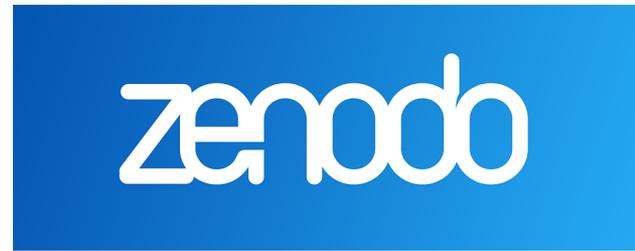
- **Data size:** overcome barriers related to data governance and storage defining **common principles**
- **Data heterogeneity:** overcome barriers of data access defining a **global coordination of open data from multi-domain fileds**
- **Data analysis:** overcome barriers of analysis diversity defining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

# CERN technologies

*Services and tools*

The Zenodo logo consists of the word "zenodo" in a white, lowercase, rounded sans-serif font, centered on a solid blue rectangular background.The reana logo features the word "reana" in a bold, lowercase sans-serif font. The "re" is colored red, and the "ana" is colored dark blue.The CERNBox logo features a white stylized atomic symbol on the left and the text "CERNBox" in a white, uppercase sans-serif font on the right, all set against a solid blue rectangular background.

# What is Zenodo?



Cross-domain digital repository for the long tail of research.

Computer science,  
Biodiversity, Humanities,  
Chemistry, ...

.pdf, .zip, .gz, .h5, .avi,  
.tiff, .png, .ipynb, .r, ...

- Launched in May 2013, by the European Commission's OpenAIRE project & CERN
- Hosted at the CERN datacenter



## Data availability

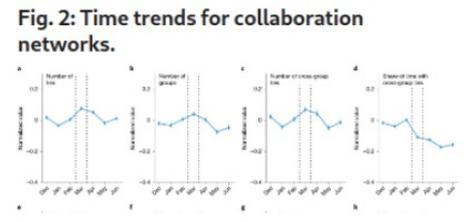
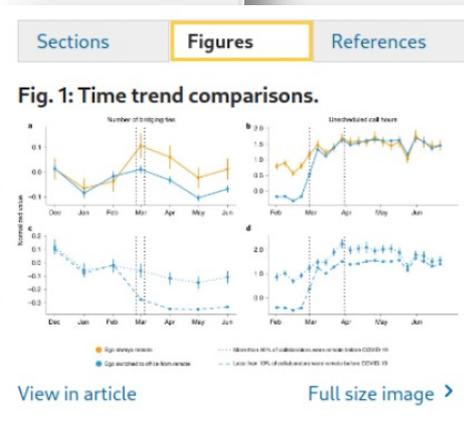
An anonymized version of the data is retained indefinitely for scientific and academic purposes and is publicly available due to employee privacy and other legal restrictions. The data are available from the authors on reasonable request and with permission from Microsoft Corporation.

## Code availability

The code supporting this study is retained indefinitely for scientific and academic purposes. The code is not publicly available due to employee privacy and other legal restrictions. The code is available from the authors on reasonable request and with permission from Microsoft Corporation.

## References

1. Bloom, N. A. *Working From Home and the Future of U.S. Economic Growth Under COVID* (2020); <https://www.youtube.com/watch?v=jtdFIZx3hyk>
2. Brynjolfsson, E. et al. *COVID-19 and Remote Work: Evidence from Enterprise Data*. Technical Report (National Bureau of Economic Research, 2020).
3. Barrero, J. M., Bloom, N. & Davis, S. 60 minutes a day: how Americans use time at home. Working Paper (Univ. Chicago Business School, 2020); [https://bfi.uchicago.edu/content/uploads/2020/09/BFI\\_WP\\_20200901.pdf](https://bfi.uchicago.edu/content/uploads/2020/09/BFI_WP_20200901.pdf)
4. Dingel, J. I. & Neiman, B. How many jobs can be done at home? *Public Econ.* **189**, 104235 (2020).



# Upload

50GB\* for each dataset  
All file formats accepted

# Describe

Rich but flexible metadata  
Based on DataCite schema  
Reserve DOI before publishing

# Publish

Citable DOI  
Export formats

# Zenodo communities

The image shows three overlapping screenshots of Zenodo community pages. The top screenshot is for 'LORY - Lucerne Open Repository', the middle for 'TWISTx Proceedings', and the bottom for 'Knowledge Junction'. Each page displays a search bar, navigation tabs (Upload, Communities, Log in, Register), and a list of recent uploads with 'View' buttons.

Projects, Subjects, Institutes, Nations, Conferences, ...

This screenshot shows a Zenodo community page for 'Software Carpentry'. It includes the community name, a description, and a recent upload titled 'Software Carpentry: Using Databases and SQL' by Adam Carnes, Andrea Paule, Sarah Schwickel, Peter J. Smyth, Pauline, et al.

**Want your own community?** [Sign Up](#)

It's easy. Just sign-up and create a new community.

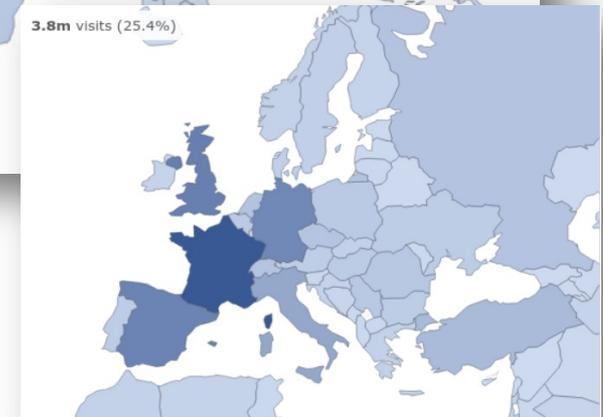
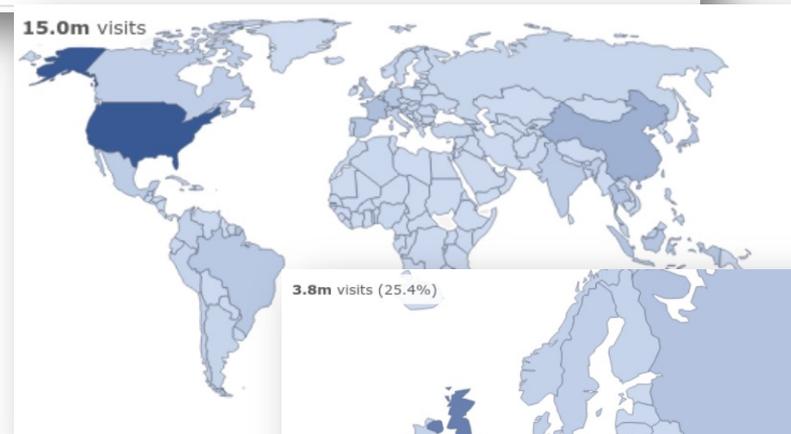
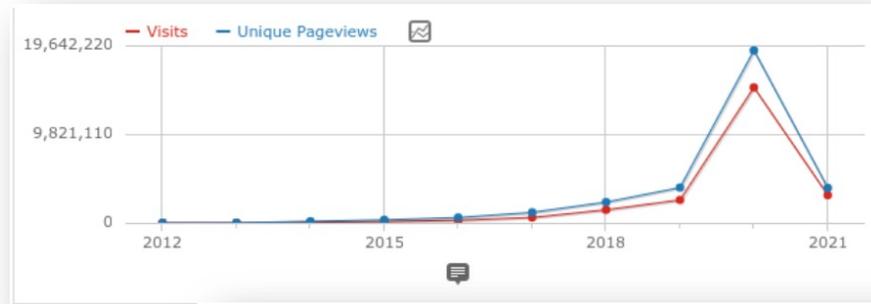
- **Curate** – accept/reject what goes in your community collection.
- **Export** – your community collection is automatically exported via OAI-PMH
- **Embed** – not custom embed link to send to

Accept  Reject

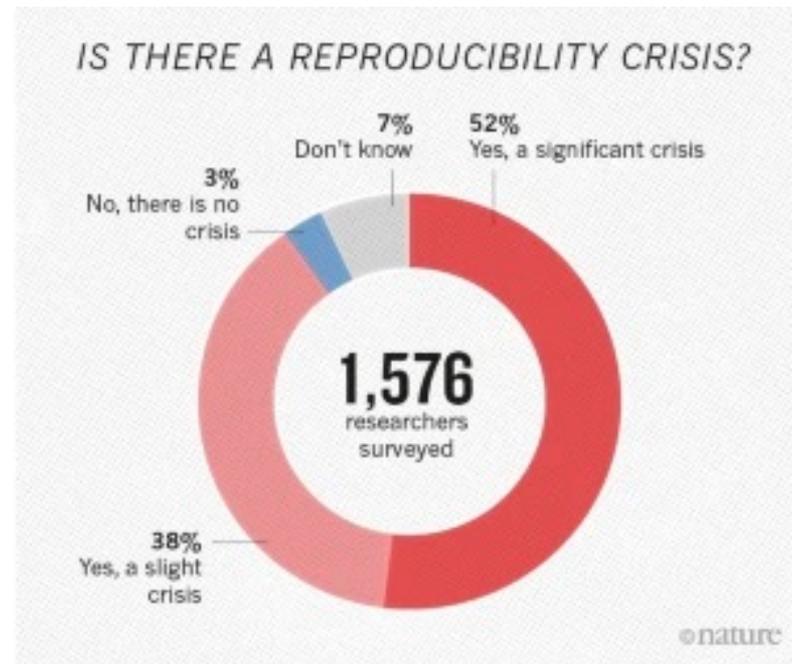
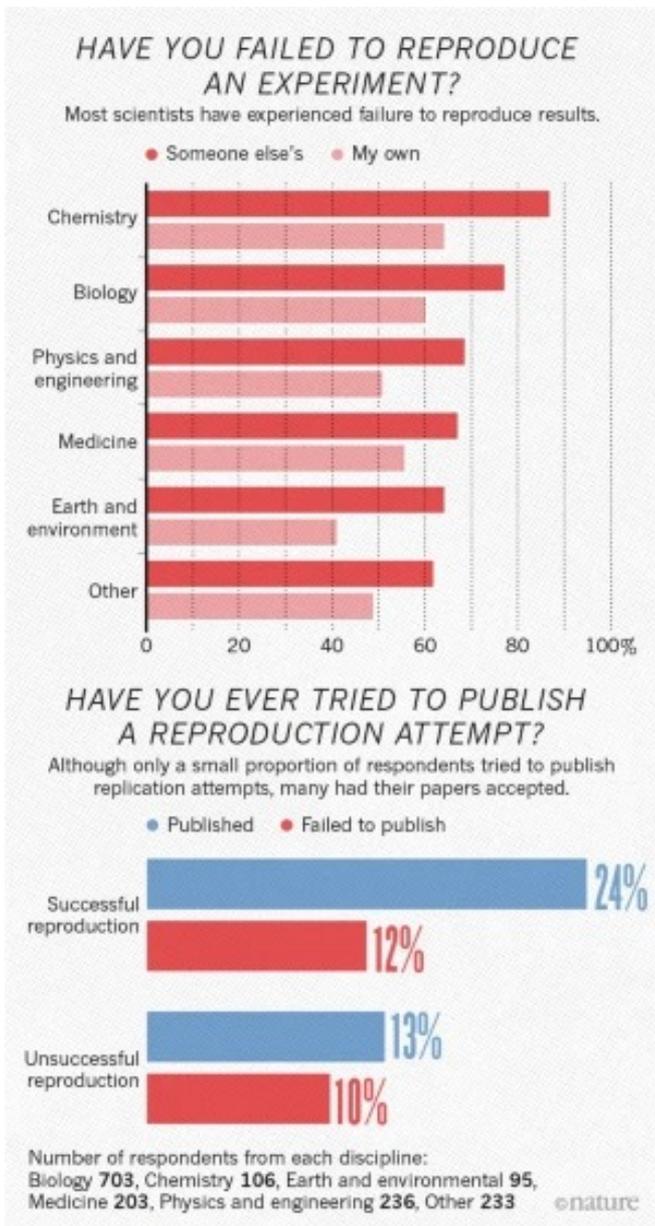
This screenshot shows the EFSA website news page. A yellow arrow points from the 'Knowledge Junction' community page in the previous block to the EFSA logo on this page. The news article is titled 'EFSA to share data on open-access platform' and includes a photo of a keyboard with a green 'Access' key.

# Zenodo in numbers

- **~2.2m records**
  - 1.1m text
  - 710k images
  - 140k software
  - 140k datasets
- **~800TB data, ~7m files**
- **15m visitors/year**
  - 3.8M from EU



# Reproducibility Crisis



Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016). <https://doi.org/10.1038/533452a>

# What can be done?

- Where is the **Data**?
  - Local, cloud storage (AWS, Google Drive, OwnCloud, ...)
- Where is the **Code**?
  - GitLab, GitHub, local, ...
- What is the **Environment**?
  - My laptop, virtual machine, computing cluster...
- What is the **Workflow**?
  - *"I remember it..."*, *bash script*, *README*, ...

# What is REANA? **reana**

Reproducible research data analysis platform

## Flexible

Run many computational workflow engines.



## Scalable

Support for remote compute clouds.



## Reusable

Containerise once, reuse elsewhere. Cloud-native.

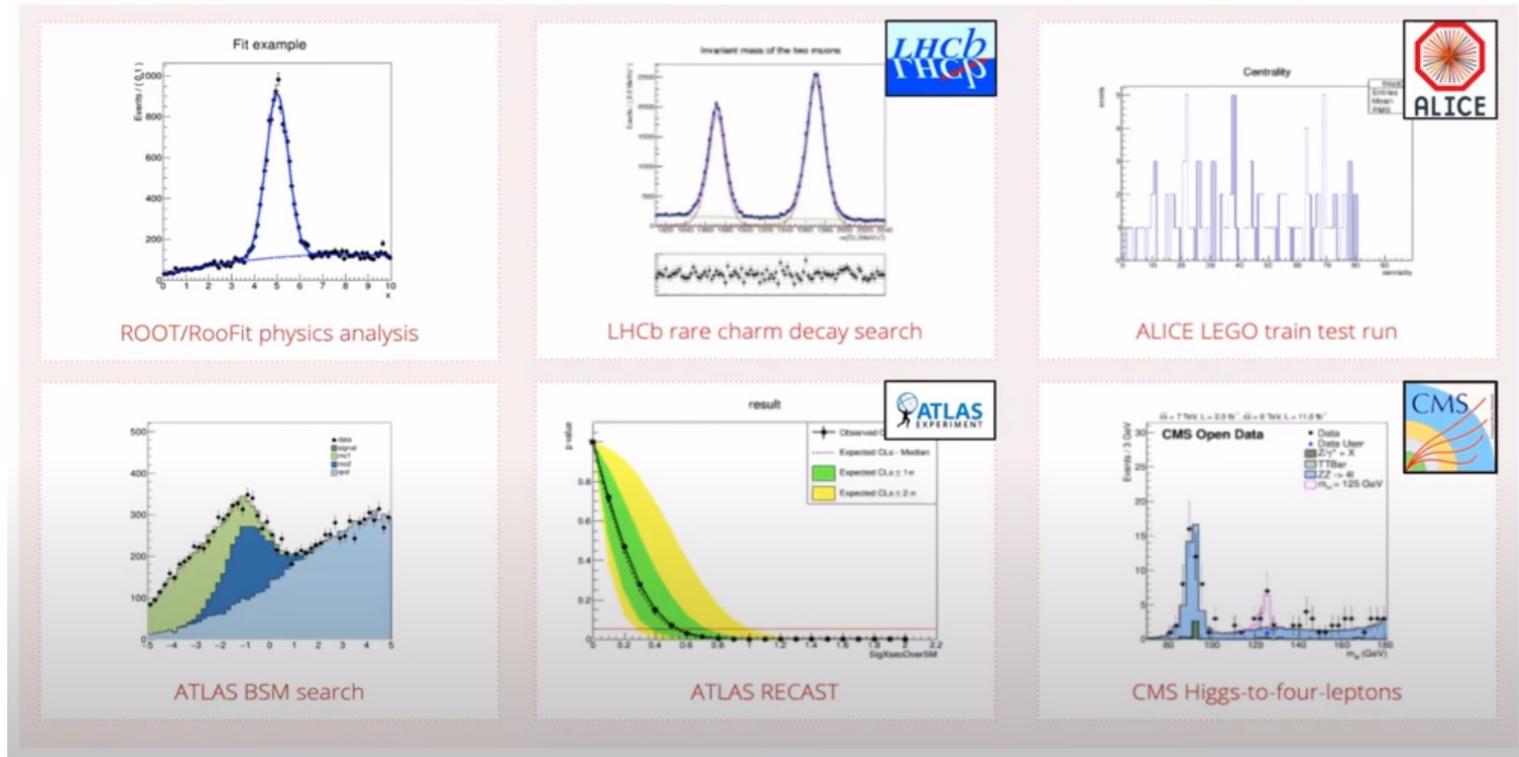


## Free

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



# Examples from HEP...



But why not other disciplines too?  
Medical sciences, Earth sciences, ...

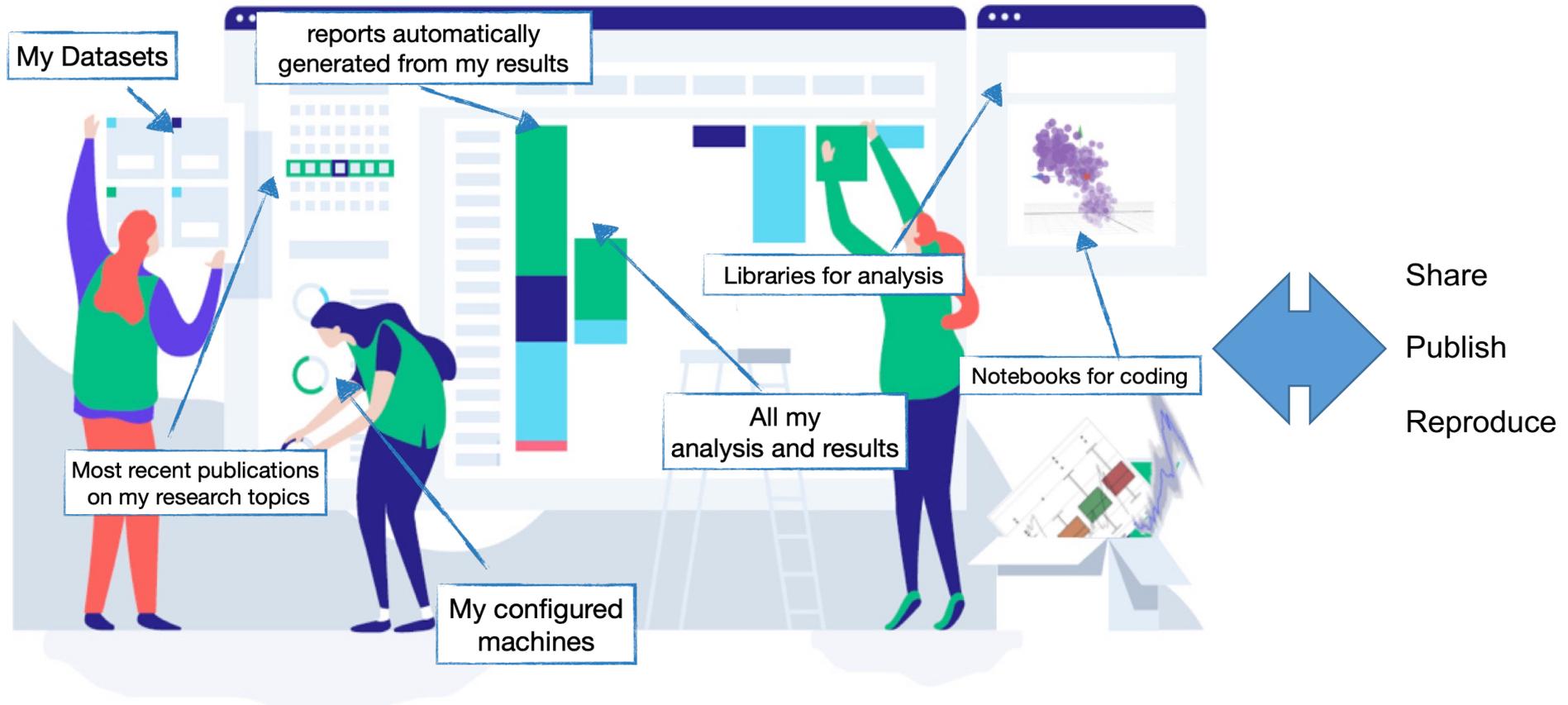
# CERN technologies

*Services and tools*

The Zenodo logo consists of the word "zenodo" in a white, lowercase, rounded sans-serif font, centered on a solid blue rectangular background.The reana logo features the word "reana" in a bold, lowercase, sans-serif font. The "re" is colored red, and the "ana" is colored dark blue.The CERNBox logo features a white stylized atomic symbol (three intersecting orbits) to the left of the text "CERNBox" in a white, uppercase, sans-serif font, all set against a solid blue rectangular background.

# CS40D Project

*Platform for end-to-end analysis best practices*



# Create a new project

## Create Project

Title \*

Project title is required.

Description \*

Project description is required. It will be displayed on the index page of the project.

Tags

Domains

SDGs

Sustainable Development Goals

Zenodo Communities

Search for Zenodo communities

CREATE PROJECT

# Upload Datasets and Notebooks

- Local
- Zenodo

## Create Dataset

Title \*

Dataset title is required.

Description \*

Dataset description is required.

Author/Creator \*

Dataset author/creator is required.

Affiliation

Author/Creator's affiliation

ORCID

Author/Creator's ORCID

Drag 'n' drop a file here, or click to select a file

CREATE DATASET

Updated a month ago



Eurostat dataset weekly deaths from 2019

Eurostat

VIEW EDIT

Updated a month ago



Comuni Italiani 2018 - Coordinate geografiche

Matteo Henry Chinaski

VIEW EDIT

# Create Preview, Visualization and Analysis

## Preview and Concatenation



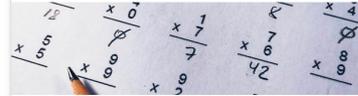
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

## Charts



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

## Analysis



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.



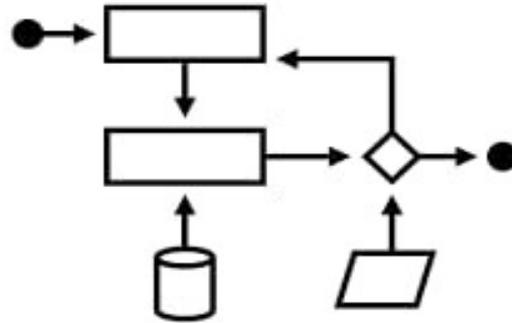
# Your Workspace

The screenshot displays the JupyterLab workspace interface. At the top, a menu bar includes 'File', 'Edit', 'View', 'Run', 'Kernel', 'Tabs', 'Settings', and 'Help'. Below the menu, a toolbar contains icons for file operations. The left sidebar features a file browser for an 'Untitled Folder /' with a table listing 'Datasets', 'Notebooks', and 'Visualizations', all marked as 'seconds ago'. The main area is a 'Launcher' panel with the following sections:

- Untitled Folder**
- Notebook**: A button with the Python logo and 'Python 3' text.
- Console**: A button with a terminal icon and 'Console' text.
- Python 3**: A button with the Python logo and 'Python 3' text.
- Other**: A section containing four buttons: 'Terminal' (with a '\$\_' icon), 'Text File' (with a list icon), 'Markdown File' (with an 'M' icon), and 'Show Contextual Help' (with a help icon).

The bottom status bar shows '0' and '0' next to icons, and the word 'Launcher' is visible in the bottom right corner of the interface.

# Make your analysis reproducible



```
version: 0.3.0
inputs:
  files:
    - code/worldpopulation.ipynb
    - data/World_historical_and_predicted_populations_in_percentage.csv
  directories:
    - workflow/cwl
  parameters:
    input: workflow/cwl/worldpopulation_job.yml
workflow:
  type: cwl
  file: workflow/cwl/worldpopulation.cwl
outputs:
  files:
    - outputs/plot.png
```

# Target User

*Data Scientist/ Data Engineer*

*Data Scientist/ Data Engineer:*

*Data storage and management*

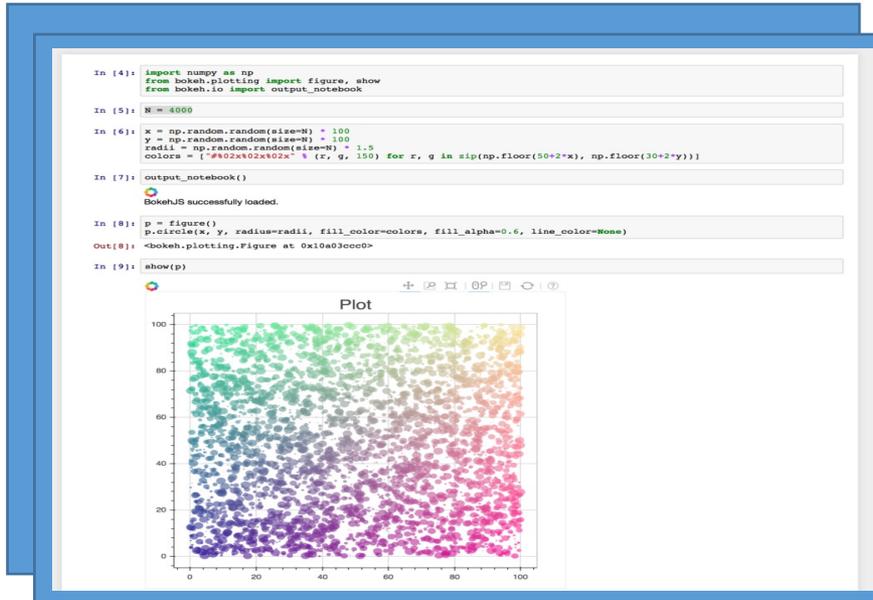
*Homogenisation of data*

*Define analysis pipelines*

*Setting up environments*

# Target User

Researcher



*Researcher:*

*Use of libraries,  
Computation of analysis*

# Target User

*Data Enthusiastic*



**Data Enthusiastic:**

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

<https://cs4od.web.cern.ch/survey/>

# Next steps

- **Extension of features:**
  - Improve platform UI for usability
  - Increase functionalities of analysis and visualizations
  - Improve guidelines for best practices
- **Implementation additional tools and pipelines for target users:**
  - Data Scientists and Engineers
  - Researches
  - Data Enthousiastic
- **Investigation and impelmentation of additional best practices:**
  - based on the new features invetstigate best practices
  - Transfer best practice to the UI

# Thank you



**Alberto di Meglio**

*Head of CERN openlab*



**Tim Smith**

*Head of Collaboration, Devices and Applications @ CERN*



**Anna Ferrari**

*Senior Fellow @ CERN  
openlab*



**Ivan Knezevic**

*COAS @ CERN openlab*



**Alexandros  
Ioannidis**

*Zenodo Service Manager @  
CERN-IT*



**Jose Benito  
Gonzalez Lopez**

*Digital Repositories Manager  
@ CERN-IT*



**Nihal Ezgi  
Yuceturk**

*Junior Fellow @ CERN-IT*



**Ines Pinto  
Pereira da Cruz**

*Junior Fellow @ CERN-IT*



**Tomáš Roun**

*Junior Fellow @ CERN-IT*



**Diamantis  
Patsidis**

*Summer Student @ CERN  
openlab*

# Thank you

*Questions?*