# 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

# 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

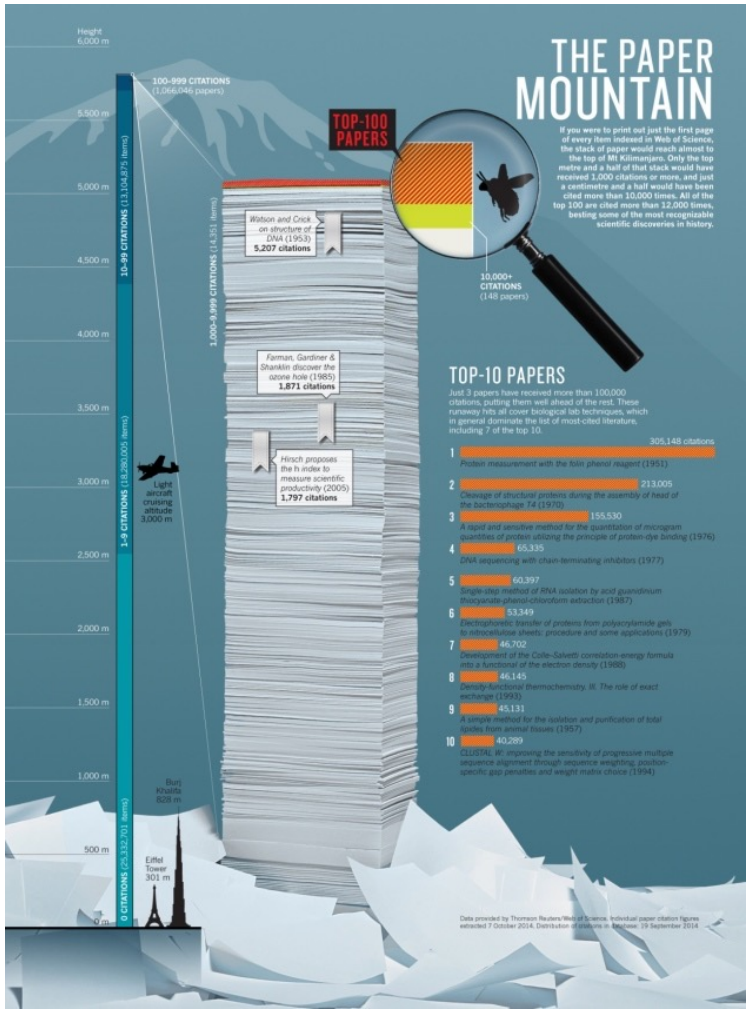
*Data analysis*

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



# 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*

zenodo

reana

SWAN

CERNBox

# What is Zenodo?

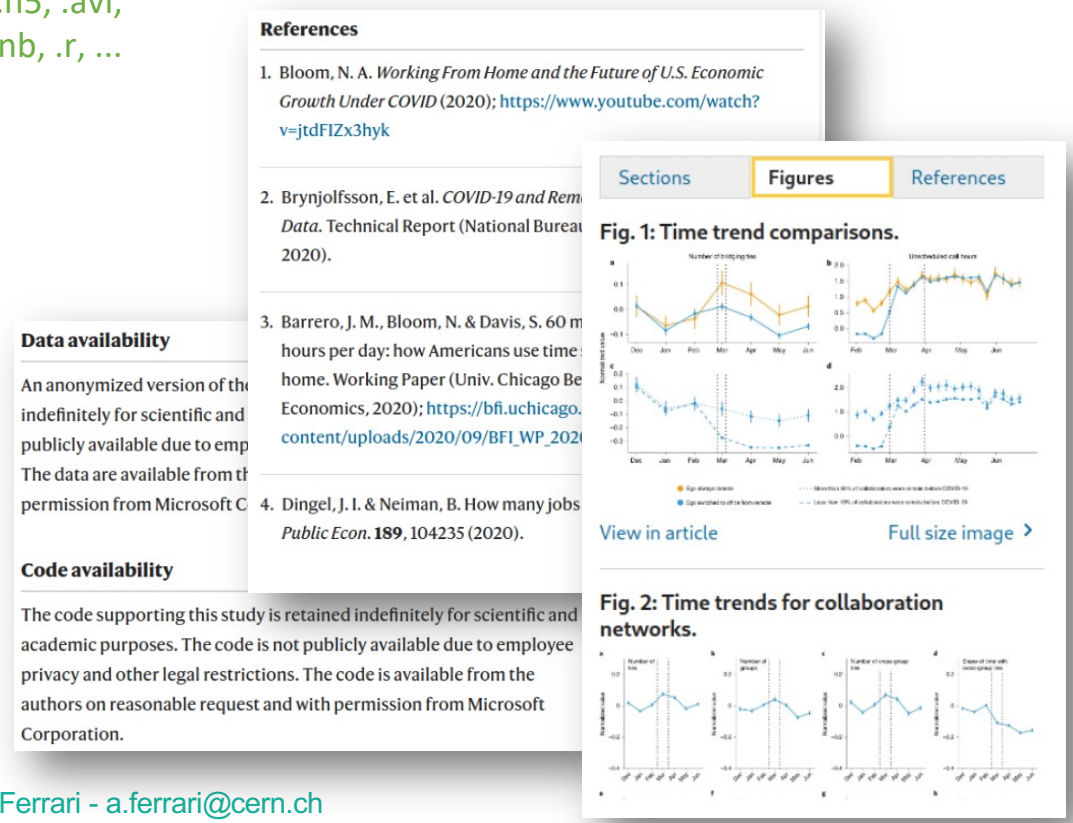
zenodo

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 Commision's OpenAIRE project & CERN
- Hosted at the CERN datacenter



**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 Rem Data*. Technical Report (National Bureau 2020).
3. Barrero, J. M., Bloom, N. & Davis, S. 60 m hours per day: how Americans use time home. Working Paper (Univ. Chicago Be Economics, 2020); https://bfi.uchicago. content/uploads/2020/09/BFI_WP_202
4. Dingel, J. I. & Neiman, B. How many jobs *Public Econ.* **189**, 104235 (2020).

**Data availability**

An anonymized version of th indefinitely for scientific and publicly available due to emp The data are available from th permission from Microsoft C

**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.

Sections | Figures | References

**Fig. 1: Time trend comparisons.**

View in article | Full size image

**Fig. 2: Time trends for collaboration networks.**

European Commission

CERN

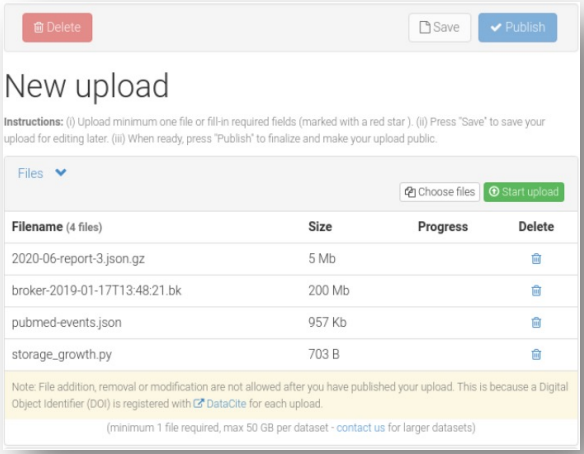
OpenAIRE

CERN openlab

Anna Ferrari - a.ferrari@cern.ch

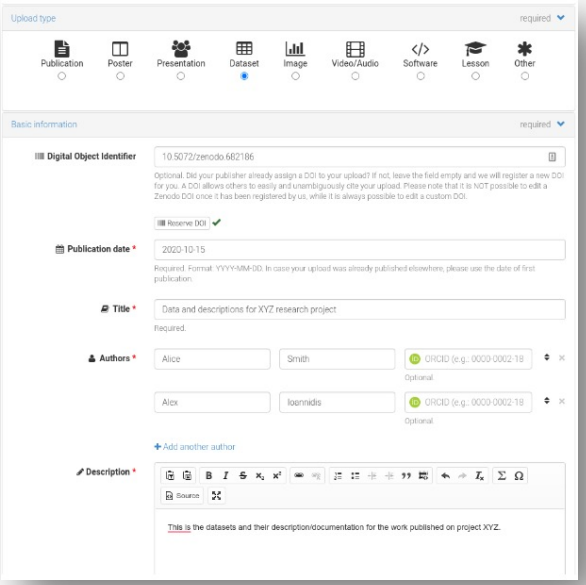
# Upload

50GB* for each dataset
All file formats accepted



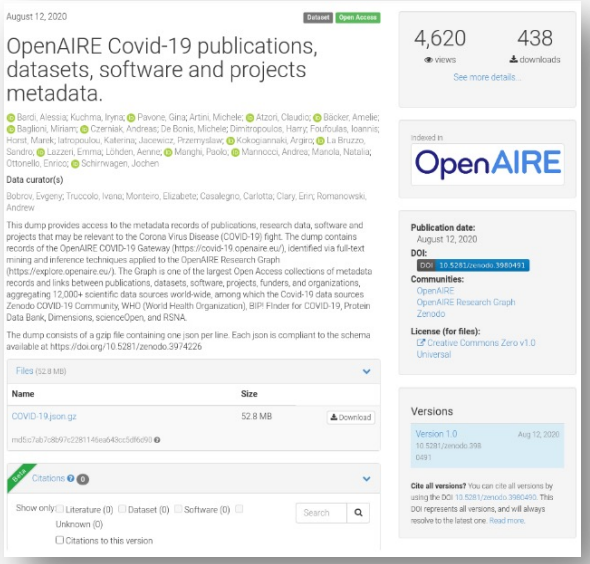

# Describe

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

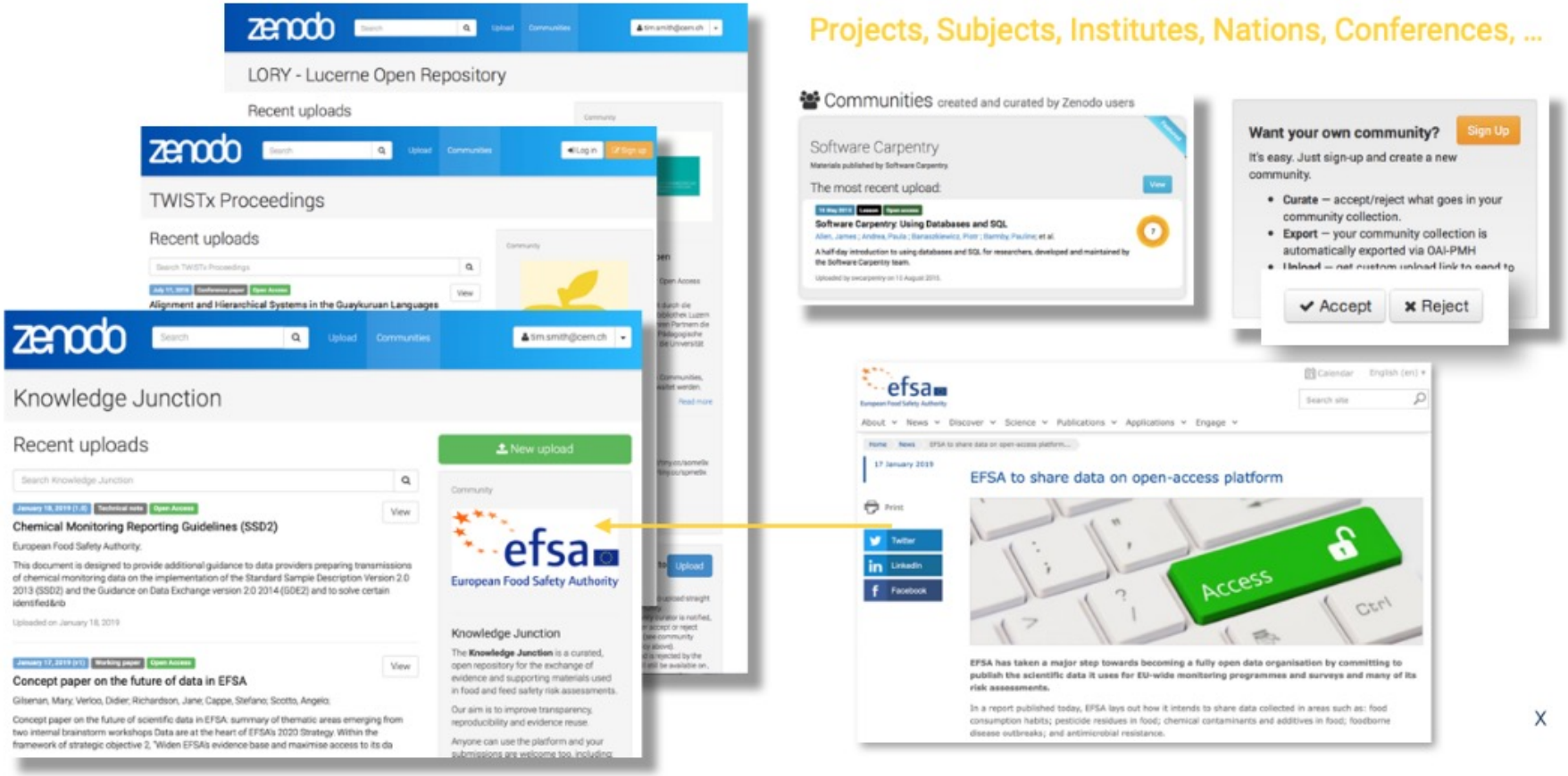



# Publish

Citable DOI
Export formats

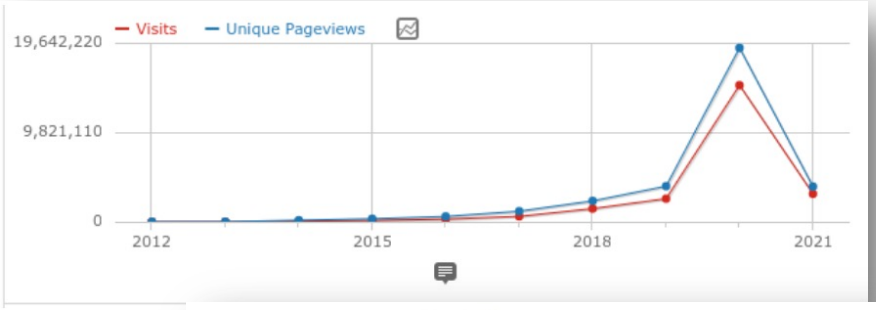


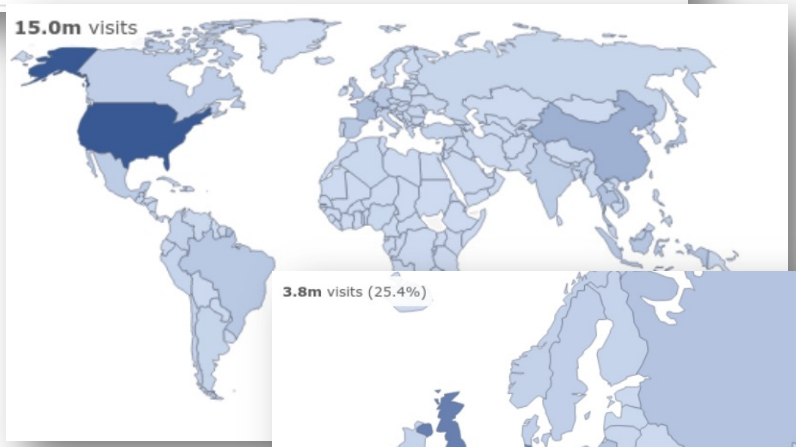
# Zenodo communities

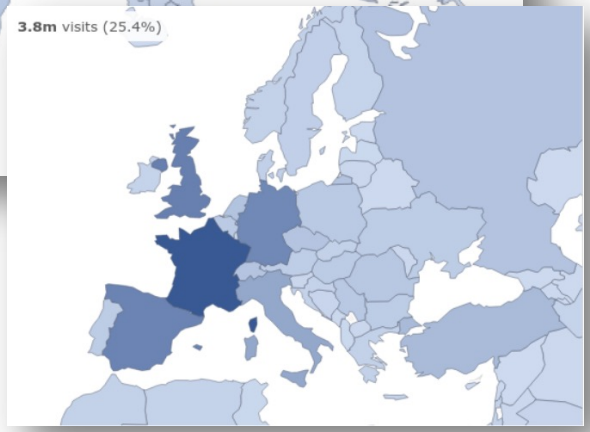


# Zenodo in numbers

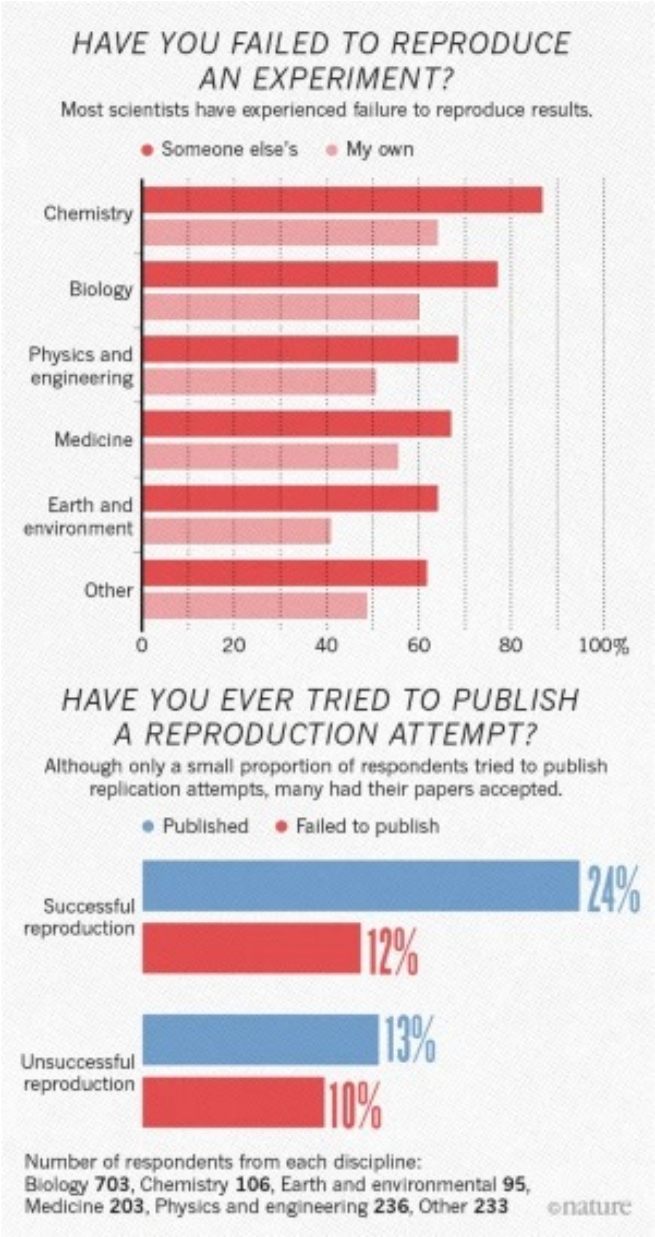
- **~2.2m** records
  - 1.1m text
  - 710k images
  - 140k software
  - 140k datasets
- **~800TB** data, ~7m files
- **15m** visitors/year
  - 3.8M from EU

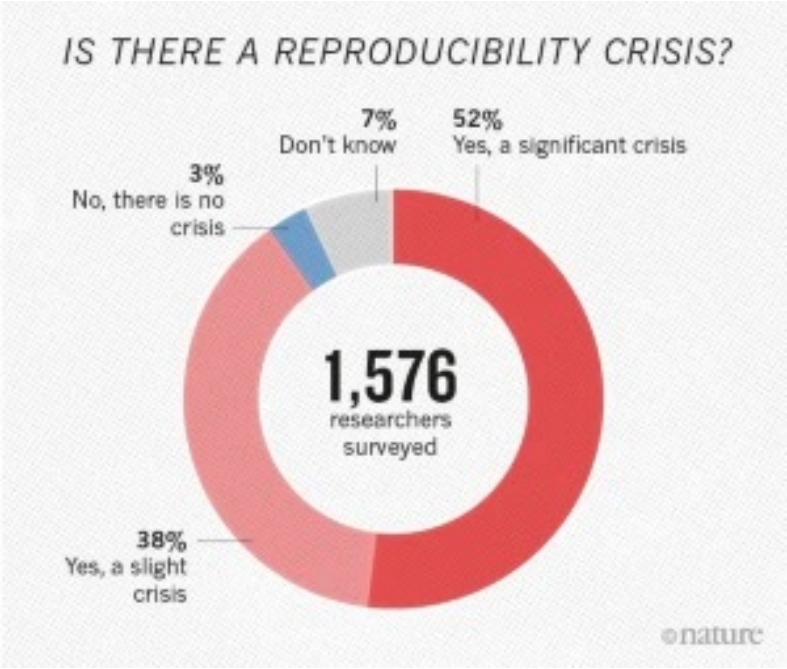


# Reproducibility Crisis




*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?

Reproducible research data analysis platform

## Flexible

Run many computational workflow engines.

COMMON WORKFLOW LANGUAGE

## Scalable

Support for remote compute clouds.

kubernetes HTCondor slurm workload manager

## Reusable

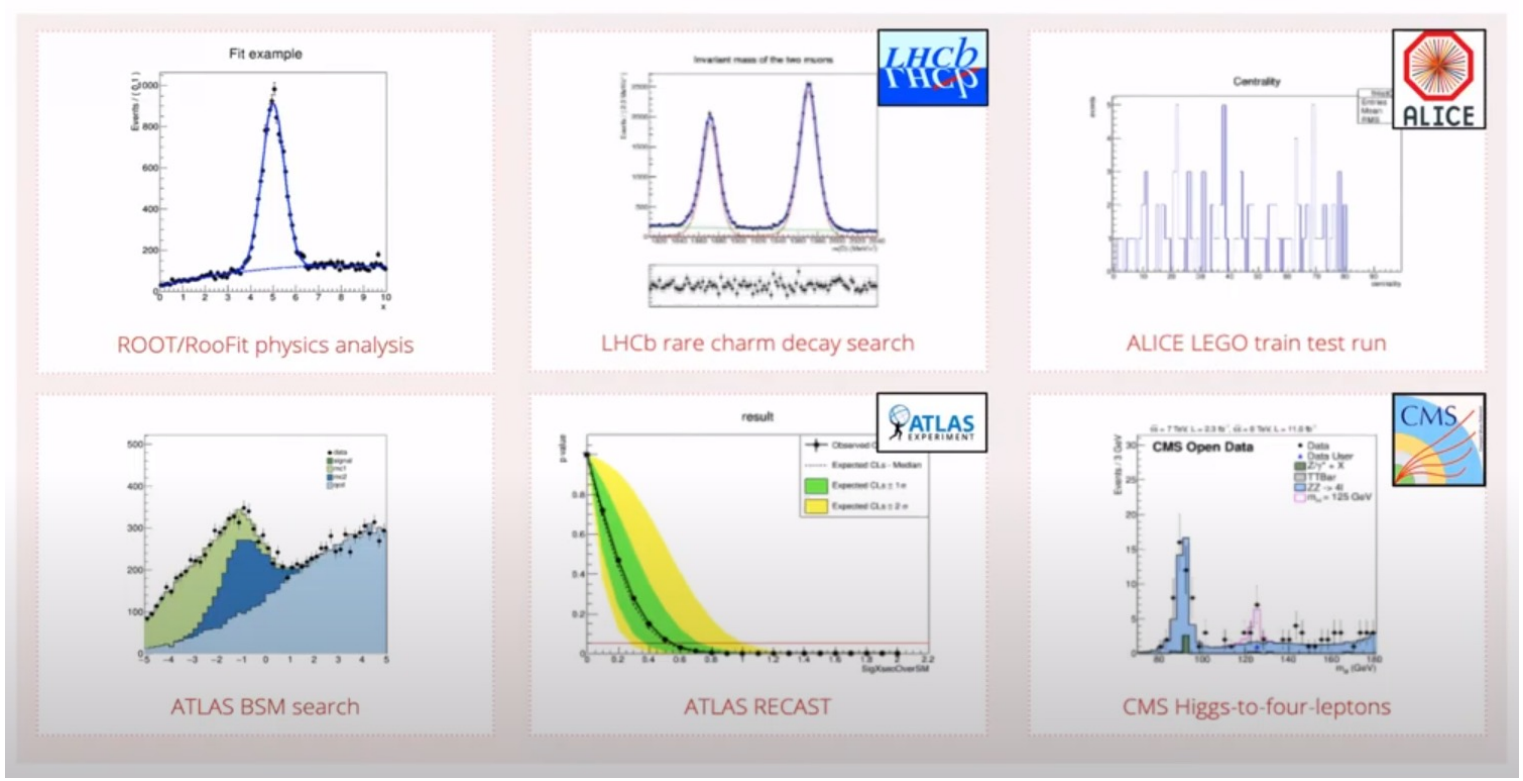
Containerise once, reuse elsewhere. Cloud-native.

## Free

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

CERN

# Examples from HEP…




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

# CERN technologies

*Services and tools*

zenodo

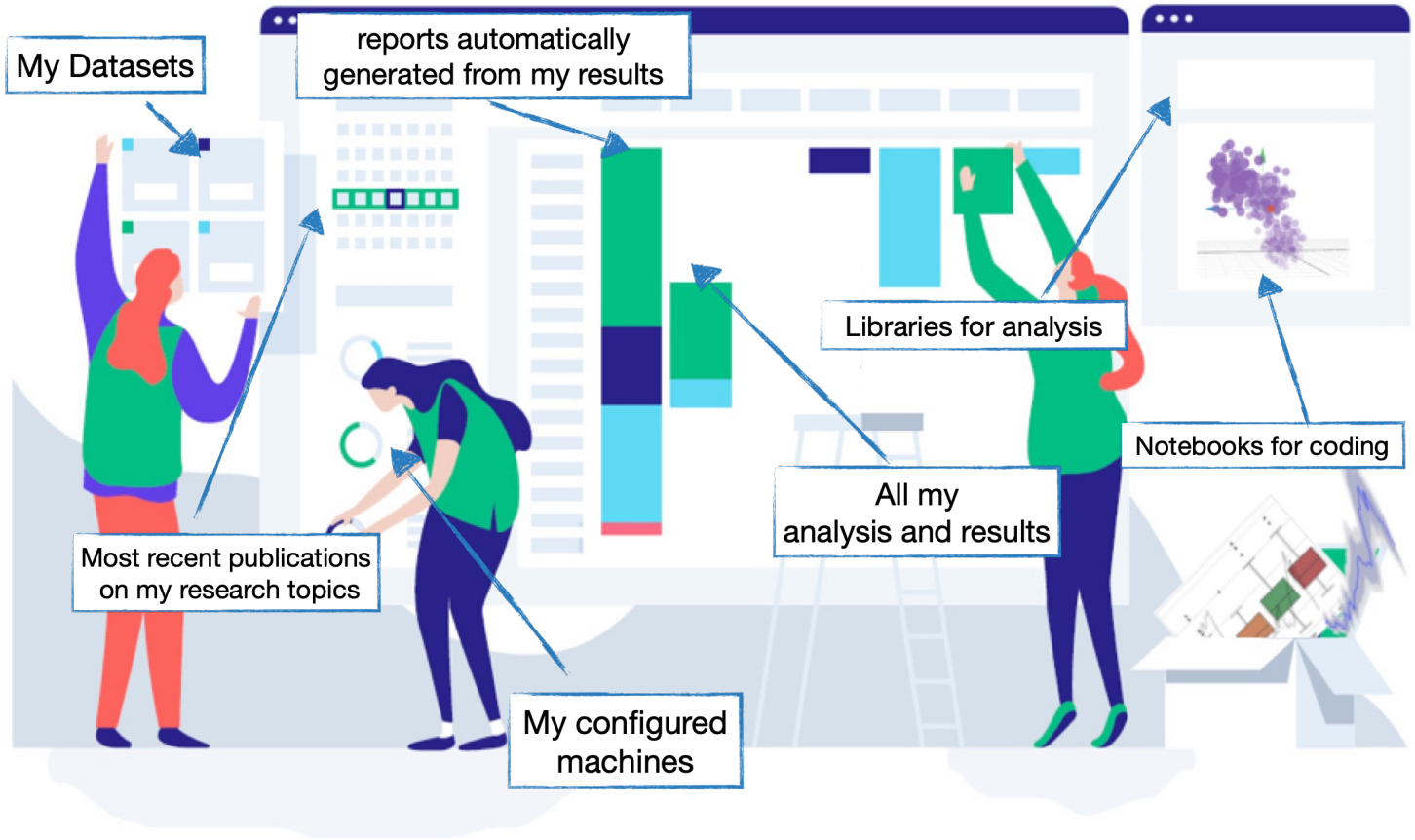
reana

SWAN

CERNBox

# CS4OD Project

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




Share

Publish

Reproduce

# 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

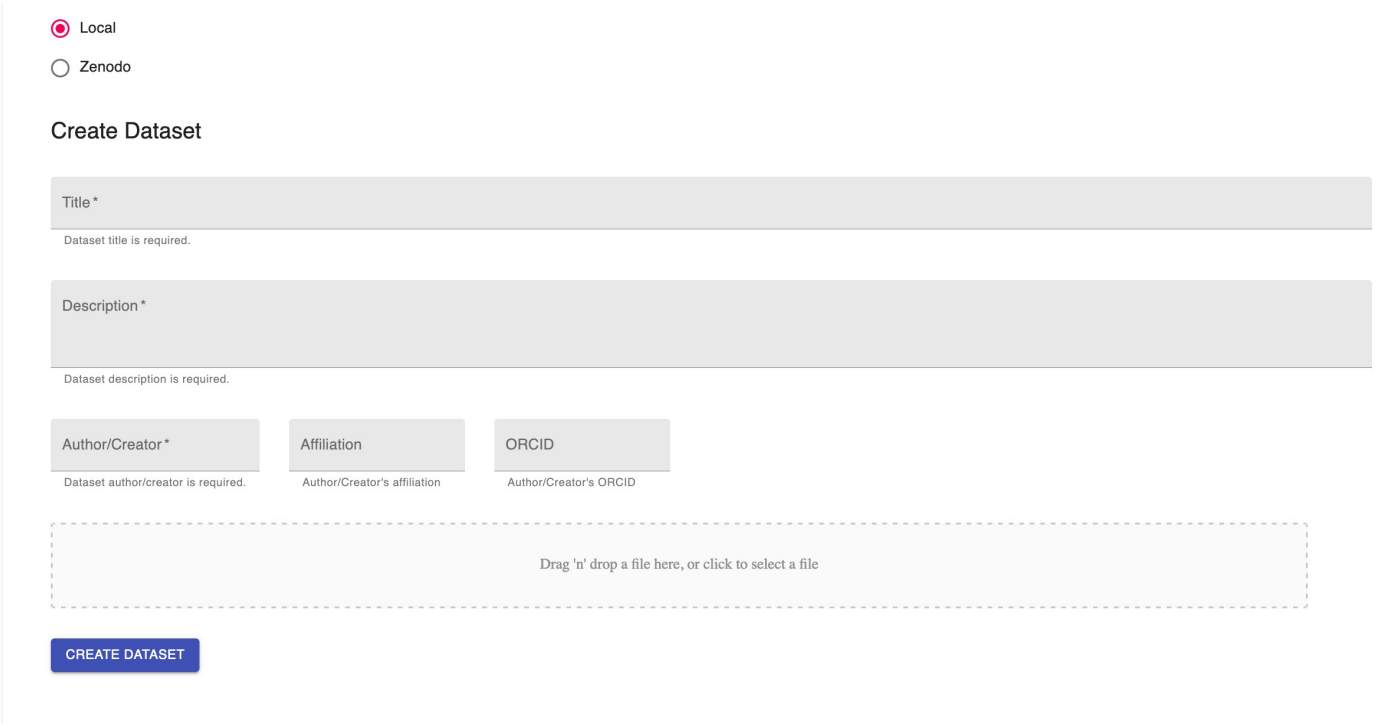
Sustainable Development Goals

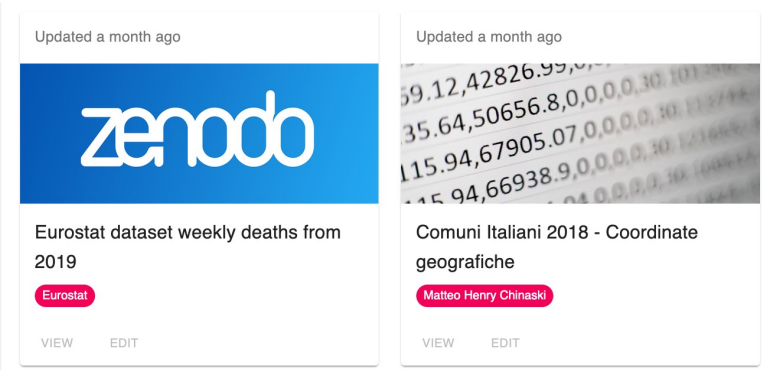
Zenodo Communities

Search for Zenodo communities

CREATE PROJECT

# Upload Datasets and Notebooks



# Create Preview, Visualization and Analysis

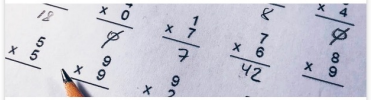
### Preview and Concatenation

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

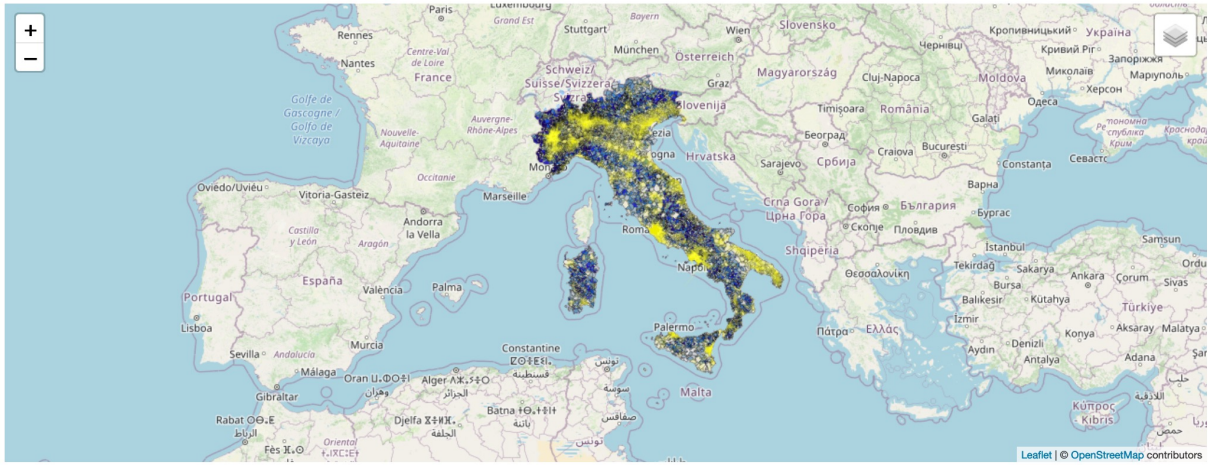
### Charts

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

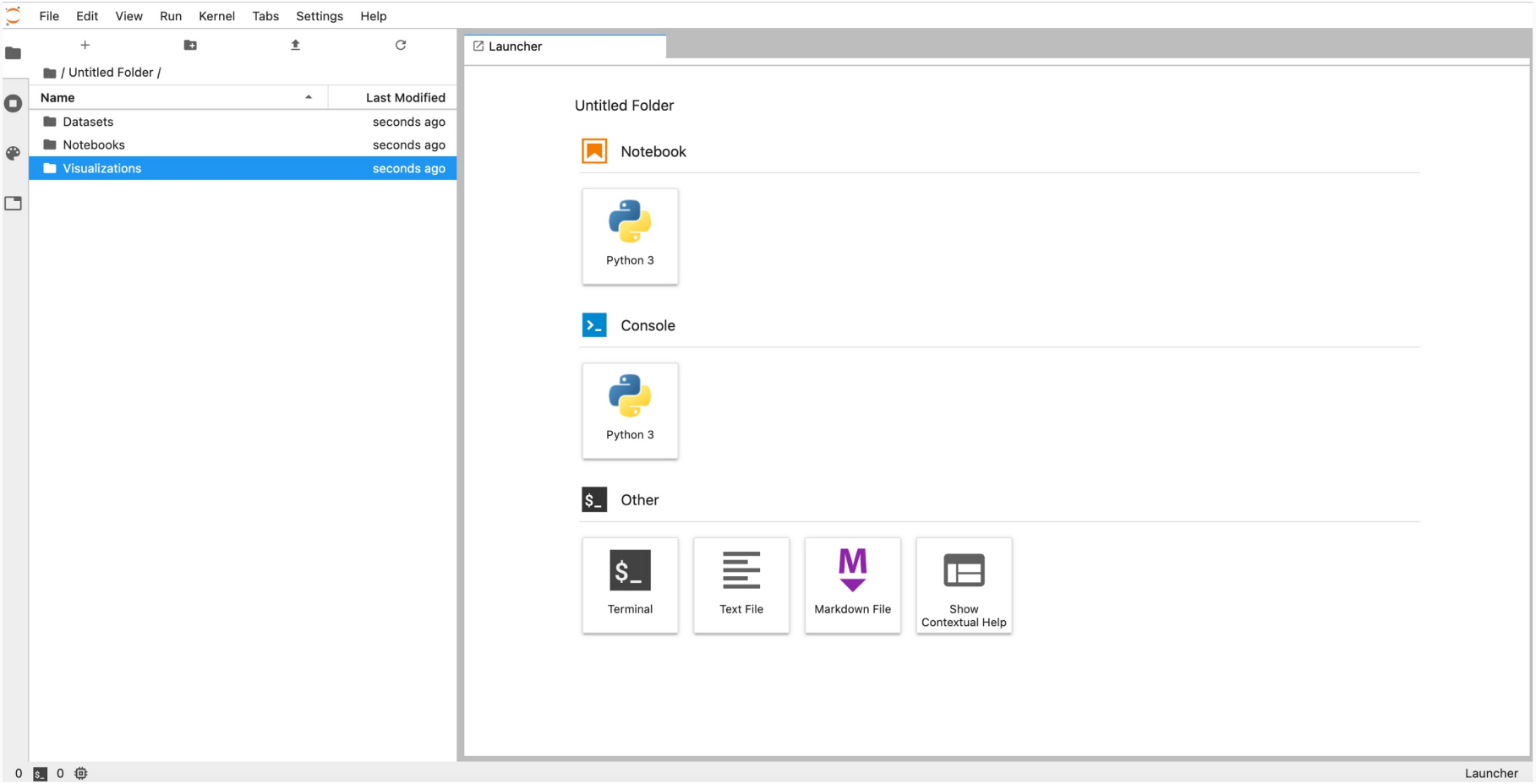
### Analysis



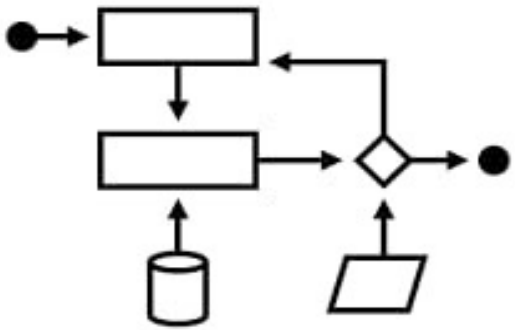
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.



# Your Workspace



# Make your analysis reproducible



```yaml
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*

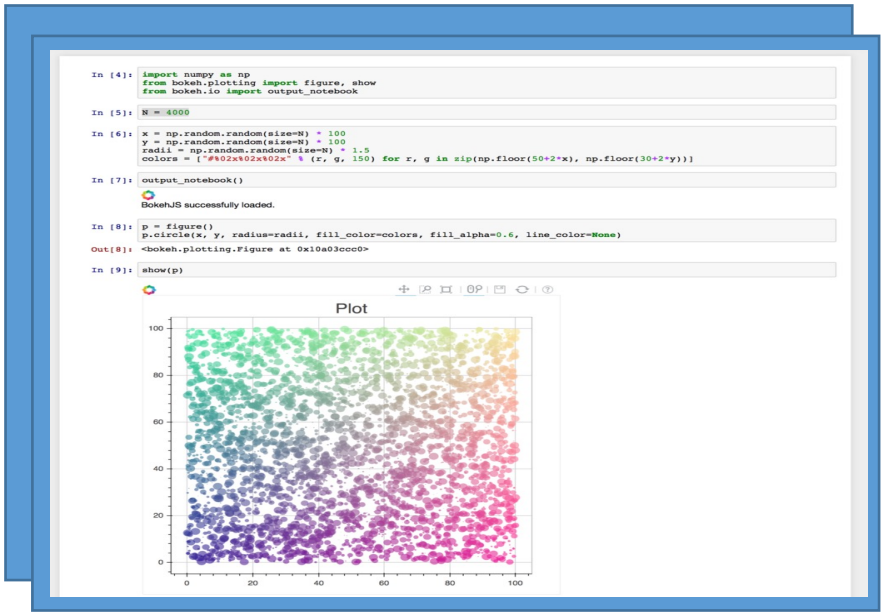
*Homogeneisation of data*

*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 prectices

- **Implementation additional tools and pipelines for target users:**
  - Data Scientists and Engeneers
  - 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?*