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
The Big Data Challenge

Data size

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

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

Anna Ferrari - a.ferrari@cern.ch
Information vs Knowledge

Data overload

- **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** 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 fields

- **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 multi-disciplinary expertises exchange*
Swan for Data Management

CERN technologies, softwares, tools, infrastructures

UI/Core
Analysis platforms
Compute
Software
Storage
Infrastructure
Zenodo as Data repository

CERN technologies for data size and heterogeneity

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

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

Most recent publications on my research topics

My configured machines

Libraries for analysis

Notebooks for coding

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

Researcher:

Use of libraries,
Computation of analysis
Platform layers

High level layer

General user:

Interested on trends and phenomana, can interact with the graphs and analysis.
Use Case: Excess of Mortality

Collaboration with Bocconi University of Milan for implementing a **platform for multidisciplinary data.**

Piloting on Italian data to estimate and represent graphically:

- All-cause death rates
- Excess mortality
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?