#CSIC

Challenges related to data in scientific research

Particle physics: from fundamental science to society

Tribute to Gaspar Barreira

September 11th 2019 @Lisbon

Jesús Marco, CSIC

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

In one slide…

- DATA takes a meaning only in a context
- For physicists, this context seemed obvious
- In informatics, this context seemed obvious
- But what is, or how to, describe the context?
  - Experiment, Theory
  - Software, Metadata
- DUALITY DATA-SOFTWARE IS NOT ENOUGH TO GRASP KNOWLEDGE
- Workflows and the data lifecycle
- Open data: facilitator (join,confront,trace,reproduce,secure)
- Enters ontology…path started towards AI via machine learning

#CSIC

An historical (subjective but very short) view…

- DELPHI Collaboration at LEP (CERN):

-Main “problem” was data structure (HYDRA/ZEBRA banks) to accommodate “trees”

-other typical “old times” precision problems with data

-Preservation? Open? What for? (but see DPHEP effort)

- CMS Collaboration at LHC (CERN):

-Welcome to OO (data->objects, data&methods)

-context (alias software) more and more complex

-LARGE VOLUMES: distributed computing: **GRID/WLCG**

-Open Data finally agreed, but who can do it?

-Reproducibility???

-What to preserve? (unique, but everything potentially interesting…)

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

An historical (subjective but very short) view…

- **APPARENTLY THERE IS LIFE OUTSIDE HEP!!!**

-CONFRONT OTHER COMMUNITIES (Why? We don't need them! Or maybe we do…)

-Common e-infrastructure: from GRID to CLOUD

-What about common methodologies also for data?

NOT THE SAME PRIORITIES

Drawback: Machine learning methods raising outside HEP

Drawback: Late for whole-cycle-workflows

Drawback: Late for cloud and to attract young tech

- **DATA AND CLOUD EMERGING AS CENTRAL TOPIC**

INDIGO project (coordinated by INFN, relevant participation of LIP and CSIC):

DMPs, DATA driven requirements

support to DATA LIFE CYCLE (from acquisition to preservation)

EVOLUTION TOWARDS DEEP LEARNING (AND OTHER TECHNIQUES REQUIRING DATA PLUS HPC INFRASTRUCTURE): DEEP HYBRID –DATACLOUD

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

Looking forward

-BIG DATA hype is officially dead but…

-AI (Artificial Intelligence) hype is back , and ACADEMIA is VERY LATE

Powerful hardware + sophisticated methods (like deep learning and beyond)

DATA AVAILABILITY IS CRITICAL although mainly internet, for commercial use, maybe next IoT

New methods may reduce this dependency on data

-LARGE OPPORTUNITIES OPEN (YET) FOR SCIENCE

Interdisciplinary studies (ex. from clima to planet impact)

Privacy data (medicine but also daily activities) regulation

Science should exploit them for the benefit of the whole society

-NEW TECHNIQUES MAY HAVE A VERY LARGE IMPACT

Quantum techniques for transmission and on decryption and on processing (not data)

And still we have to learn how we store data in our brain…

and this is really A CHALLENGE (cf. HBP, and clinical approaches)

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

But we know the path…

COLLABORATION

To share our data

To share our infrastructures

To share our ideas

To share our teams!

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

Photo credits: Victor Castelo (CSIC)

CSIC
Consejo Superior de Investigaciones Científicas