BIG DATA & FUNDAMENTAL RESEARCH CHALLENGES & PERSPECTIVES

Jesús Marco de Lucas
CSIC (National Research Council, Spain)

ON INTELLIGENT SIGNAL PROCESSING FOR FRONTIER PHYSICS & INDUSTRY
WELCOME TO INFIERI 2019
MAY 12TH-26TH, WUHAN, CHINA
Outlook

- The lecture will introduce different fundamental research challenges where Big Data techniques are key, how they are being addressed, and the new ideas being explored at different levels:
  - Infrastructure
  - cloud-based platforms
  - data science solutions
- A general framework developed in the framework of EU initiatives will be presented and used as a guide to understand the roles of the various stakeholders. It will be translated into specific examples in different areas (physics, biodiversity, earth observation).
- An approach to the implementation of the full data life cycle in an open science framework will be introduced, describing the importance of data “fairness” for reuse, and the need for a cloud platform supporting this activity.
- Finally, the specific application of deep learning techniques to big data to derive new ideas in fundamental research will be discussed.
Seven Messages Today

• Most Research Challenges in 21\textsuperscript{st} century imply Big Data
• Big Data requires Efficient Infrastructure + Effective Software
• 3Vs (VOLUME, VELOCITY, VARIETY) are addressed by technology
• DATA implies MANAGEMENT (DMP): PLAN full DATA life cycle (6S)
• SERVICE ORIENTED ARCHITECTURE (SOA): consider CLOUD framework
• DESIGN TOGETHER DATA+PROCESSING to GET VALUE
• 4V = VALUE requires “intelligence”: DATA SCIENCE (+ “true” IA next?)
Spanish National Research Council

120 RESEARCH CENTERS

12,000 papers / 900 PhD / 120 patents annually, in all scientific areas

<table>
<thead>
<tr>
<th>Category</th>
<th>Women</th>
<th>Men</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientists</td>
<td>3,531</td>
<td>2,211</td>
<td>5,742</td>
</tr>
<tr>
<td>PhD Students</td>
<td>1,242</td>
<td>643</td>
<td>1,885</td>
</tr>
<tr>
<td>Technicians</td>
<td>5,047</td>
<td>2,320</td>
<td>7,367</td>
</tr>
<tr>
<td>Admin</td>
<td>1,265</td>
<td>784</td>
<td>2,049</td>
</tr>
<tr>
<td>Total General</td>
<td>11,085</td>
<td>5,611</td>
<td></td>
</tr>
</tbody>
</table>

*Includes PhD students and technicians.
Research: **21st century challenges?**

- Healthy food for everyone
- **Global Change:**
  - Protection of the environment (water, soil, atmosphere, bio)
  - Clean and renewable energy
  - Mobility and cities
- **Circular Economy and new materials**
- **Cure of chronic and infectious diseases**
- **Exploration and Colonization of Space**
- **Quantum and cloud computing + Connectivity**
- **Artificial Intelligence and Robotics**
- +
  - **New models of elementary particles and cosmology**
  - **Synthetic cells and origin of life**
  - **Aging and mind**
Es sorprendente que el #LHC, que es la máquina más grande y compleja construida por la humanidad, no tenga un fin militar ni económico, sino puramente romántico: conocer la realidad en su nivel más profundo. (A. Casas) bit.ly/2ldZHj4
¿How does our brain work? ¿How does it age?

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180400
Cuenta oficial del Consejo Superior de Investigaciones Científicas (CSIC).
Hablamos de ciencia y del trabajo de los investigadores.
Algae Bloom: a Big Data Open Science Problem

1-D models are not enough, we need to understand the evolution in 3D
Guiding Exercise

- AN ANTHROPOCENE PROBLEM!
- DESIGN THE ALGAE BLOOM PROBLEM AS A BIG DATA CHALLENGE:
  - Define the problem!
  - Understand the “DATA” REQUIRED
  - Design the “DATA PROCESSING”
  - MODELLING NEEDS?
  - WHAT VALUE DO YOU GET ???
- ARE WE USING AN AGILE APPROACH?
- IS THIS AN INTERDISCIPLINARY PROBLEM??
BIG DATA: introduction

Big Data according to Wikipedia:

“Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.”

History and evolution:

“In a 2001 research report, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources) (3V).

In 2012, Gartner updated its definition as follows: "Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."
BIG DATA Concept

BIG DATA HYPE IS NOW OFFICIALLY DEAD, but...

What was the definition for Big Data?

“it is the term that describes large volumes of data (from terabytes we go to zetabytes) that are generated at high speed (we go from batch/file data to streaming data), with a possible component of complexity and variability in the format of those data (we go from structured data to semi-structured or unstructured data) and that require specific techniques and technologies for their capture, storage, distribution, management, and analysis of the information.

Other “definitions”:

• “We consider Big Data when the volume of data becomes part of the problem to be solved" (O'Reilly Radar).

• "Big Data's technologies describe a new set of technologies and architectures, designed to extract value and benefit from large volumes of data with a wide variety in nature, through processes that allow information to be captured, discovered and analyzed at high speed and at low cost. (EMC/IDC)
BIG DATA: our vision

- Big Data is not a technology in itself, but rather a working approach to obtaining value and benefits from the large volumes of data that are being generated today. Aspects such as the following should be considered:
  - How to capture, manage and exploit all this data.
  - How to secure this data and its derivatives, as well as its validity and reliability.
  - How to arrange the sharing of this data and its derivatives in the organization to obtain improvements and benefits.
  - How to communicate this data and its derivatives (visualization techniques, tools, and formats) to facilitate decision making and subsequent analysis.

WE MUST BUILD OUR OWN "VISION" OF BIG DATA!

- Example: GRID technology has allowed us to solve the challenge of LHC data processing, which "was" a Big Data problem.
- To build this "vision" we need to know the available technology.
- Being "aware" of technological developments is a challenge in itself:
  - very rapid evolution of techniques and capabilities
  - Difficulty separating real interest and professional/commercial interest
Technical Requirements to solve a challenge
A very complex “instrument”

**Proton - Proton**
- 2808 bunch/beam
- Protons/bunch: $10^{11}$
- Beam energy: 7 TeV ($7 \times 10^{12} \text{ eV}$)
- Luminosity: $10^{34} \text{ cm}^{-2} \text{s}^{-1}$

**Crossing rate**: 40 MHz

**Collision rate**: $10^7 - 10^9$

**New physics rate**: $0.0001 \text{ Hz}$

**Event selection**: 1 in $10,000,000,000,000,000$
European Grid Infrastructure & WLCG

Logical CPUs (cores)
- 270,800 EGI
- 420,000 All

139 PB disk and 134 PB tape
1,35 million jobs/day

Resource Centres
- 326 EGI
- 352 All
- 108 supporting MPI

Countries
- 42 EGI
- 54 All
ASSEMBLING A BIG DATA SOLUTION

DESIGN OF A SOLUTION:

• Basic: DATA MANAGEMENT PLAN
• Basic: INFRASTRUCTURE RESOURCES
• Basic: PROCESSING ALGORITHMS
• Advanced: PARALLELIZATION/DISTRIBUTION/GRAPHS
  • Data structures
  • Processing frameworks (HADOOP...)
• Advanced: DATA SCIENCE SOLUTIONS
  • STATISTICAL METHODS
  • DEEP LEARNING
@CSIC

DATA SOURCES

• Sensors (information sources):
  • Instrumentation
    • Sensor networks
    • Cameras
    • Satellites
  • Personal devices
    • Smartphones
    • Cars
    • Other personal devices (DNA chips?)
  • Internet messaging
  • OPEN DATA
• Global examples:
  • Copernicus@ESA (Observing the Earth)
  • SmartCities
• Integration: Sensor Web Enablement/ Web Services
INDIGO Data Life Cycle ("6S")

**Stage 1: Plan:** DMP
**Stage 2: Collect:** process of getting data
**Stage 3: Curate:** actions performed over the data.
**Stage 4: Analyse:** also called “Process”, given the data an added.
**Stage 5: Ingest (\& Publish):** including other steps like “Access”, “Use” or “Re-use”, in this stage, data is normally associated to metadata, has a persistent identifier and is published in an accessible repository or catalogue, under a format that makes it useful for further re-use.
**Stage 6: Preserve:** "store" both data and analysis for long-term. Licenses and methods need to be taken into account.
INDIGO Data Ingestion: The Arbor metaphor

“Data Ingestion as the process that ends with the data being ready for sharing/(re-)use, following the usual community requirements”
“Data Ingestion as the process that ends with the data being ready for sharing/(re-)use, following the usual community requirements”

FAIR
+ Reproducibility
+ Security/Legal
“Data Ingestion as the process that ends with the data being ready for sharing/(re-)use, following the usual community requirements”
# INDIGO Data Integrity Test

<table>
<thead>
<tr>
<th>Stage</th>
<th>How?</th>
<th>INDIGO-DataCloud Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1: Plan</strong></td>
<td>Check DMP Existence</td>
<td>Manual</td>
</tr>
<tr>
<td></td>
<td>Next gen: Machine Actionable DMPs</td>
<td>Automatic linking (not implemented)</td>
</tr>
<tr>
<td><strong>Stage 2: Collect</strong></td>
<td>DataSet existence</td>
<td>EML – OneData</td>
</tr>
<tr>
<td></td>
<td>DataSet Integrity (checksum)</td>
<td></td>
</tr>
<tr>
<td><strong>Stage 3: Curate</strong></td>
<td>Qc/Qa description OK</td>
<td>EML – OneData</td>
</tr>
<tr>
<td></td>
<td>Curating, Quality Software (optional)</td>
<td></td>
</tr>
<tr>
<td><strong>Stage 4: Analyze</strong></td>
<td>Parameters description OK</td>
<td>EML – OneData</td>
</tr>
<tr>
<td></td>
<td>Processable Check</td>
<td></td>
</tr>
<tr>
<td><strong>Stage 5: Ingest</strong></td>
<td>Previous stages OK</td>
<td>EML – OneData</td>
</tr>
<tr>
<td></td>
<td>Have PID/DOI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Open Protocol (OAI-PMH)</td>
<td>Compatible with OneData</td>
</tr>
<tr>
<td><strong>Stage 6: Preserve</strong></td>
<td>License Definition</td>
<td>EML – OneData</td>
</tr>
<tr>
<td></td>
<td>Preservation details</td>
<td>QoS - OneData</td>
</tr>
</tbody>
</table>
Persistent IDs for Data: Globally unique, resolvable, persistent IDs for research data make the data more findable and accessible, enable citation, and permit linking to data from within publications and other kinds of research presentations. Digital Object Identifiers (DOIs) offer a common example of a persistent ID.

Machine-readable DMPs: When written effectively, DMPs clarify how researchers will effectively disseminate and share research results, data, and associated materials. However, DMPs can also contain complex and/or ambiguous terms that produce uncertainty about the benefits of data management activities. Such ambiguity can produce situations where the DMP does not adequately explain what data will be created or where the data will be deposited. For this reason, NSF encourages the use of DMP tools, such as EZDMP7 or the DMPTool8, to create machine-readable DMPs. The DMP specifies how data will be produced, prepared, curated, and stored. A machine-readable document allows a computer program to interpret the DMP, such as to prepare a data repository for an eventual deposit of a large or complicated dataset.
VISUALIZATION

• “Hardware”: Dashboards, Walls, 3D glasses, etc.

• Software/Graphics:
Seven Messages Today

• Most Research Challenges in 21\textsuperscript{st} century imply Big Data
• Big Data requires Efficient Infrastructure + Effective Software
• 3Vs (VOLUME, VELOCITY, VARIETY) are addressed by technology
• \textbf{DATA implies MANAGEMENT (DMP): PLAN full DATA life cycle (6S)}
• SERVICE ORIENTED ARCHITECTURE (SOA): consider CLOUD framework
• DESIGN TOGETHER DATA+PROCESSING to GET VALUE
• 4V = VALUE requires “intelligence”: DATA SCIENCE (+ “true” IA next?)
Infraestructure Components

• Data Centers
  • Supercomputers top500: up to 140 Petaflops
  • GRID: up to 800,000 cores (WLCG), >200 Petabytes
• Equipment:
  • Storage + File system (HADOOP, GPFS, Lustre...)
  • Clusters: Infiniband network
• Network
  • Dark fiber (then n x 100Gb/s)
• Cloud
  • Amazon, Google, IBM,
  • EUROPEAN OPEN SCIENCE CLOUD (EOSC)
Supercomputers *(from top500)*

#1 Summit, IBM-built system, Oak Ridge National Laboratory, USA, **143.5 Pflop/s** on the HPL benchmark, 4,356 nodes, each one 2xPower9 CPUs (22 cores) + six NVIDIA Tesla V100 GPUs with 80 streaming multiprocessors (SM). Connected with Mellanox dual-rail EDR InfiniBand network.

#2 Sierra, a system at the Lawrence Livermore National Laboratory, USA, 94.6 Pflop/s. 4,320 nodes with 2xPower9 CPUs + four NVIDIA Tesla V100 GPUs.

#3 Sunway TaihuLight, a system developed by China’s National Research Center of Parallel Computer Engineering & Technology (NRCPC) installed at the National Supercomputing Center in Wuxi, 93 Pflop/s.

#4 Tianhe-2A (Milky Way-2A) National Supercomputer Center in Guangzho, China. 61 Pflop/s upgraded earlier this year by replacing the Xeon PHI accelerators with the new proprietary Matrix-2000 chips.

#5 Piz Daint, Cray XC50 system installed at the Swiss National Supercomputing Centre (CSCS) in Lugano.

…

#446  **AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , NEC, National Institute of Advanced Industrial Science and Technology, Japan
Big Data & CLOUD

- **Cloud Computing**
  - A paradigm to offer computing services via Internet.
  - Technically based on virtualization
  - Mature technology, embraced by the industry and, more and more, by the science

**¿Why Big Data & Cloud Computing?**

*e-IRG white paper* 15 may 2013:

“These two major challenges, Big Data and cloud computing, are not totally independent: not only because Big Data may require a huge computing power but also because Big Data could represent the killer application for clouds”.

CLOUD Service Layered Architecture

PROVIDERS

ENABLERS

SOLVERS

Software

“External” Data

“Internal” Data

SITE

Network

Storage

Servers

SITE ...

USER APPLICATIONS

COMPOSITION

e-INFRASTRUCTURE

RESOURCES

Software as a Service (SaaS)

Platform as a Service (PaaS)

Infrastructure as a Service (IaaS)
@CSIC

USER APPLICATIONS

INDIGO SW

Solver

Enabler

Providers

Infrastructure as a Service (IaaS)

Platform as a Service (PaaS)

Software as a Service (SaaS)

e-INFRAsTRUcTURE - FedCloud

Delft3D Modeling Service

Future Gateway

D3D

OneData

Orchestrator

Repositories

D3D Docker

OneClient

Docker

1CPU, 1024 MB mem

User

Base Model/Input

Model Output

RESOURCES

Bari INDIGO Testbed

SITE

Servers

Storage

Network

SITE...

Repositories

1:N
Complete Example: AlgaeBloom

- IAM Integrated
- Last version of Interface for N simulations (range, N steps)
- Last version of Orchestrator
- Last version of OneData – Tested with Big Models
- Parallel applications: Geoserver
Results and impact

- Eutrophication impacts directly in water quality and human health.
- Tools for eutrophication management optimization.
- Reproduce the process in other reservoirs with the same problem.
- Understand how eutrophication happens.
Seven Messages Today

• Most Research Challenges in 21st century imply Big Data
• Big Data requires Efficient Infrastructure + Effective Software
• 3Vs (VOLUME, VELOCITY, VARIETY) are addressed by technology
• DATA implies MANAGEMENT (DMP): PLAN full DATA life cycle (6S)
• SERVICE ORIENTED ARCHITECTURE (SOA): consider CLOUD framework
• DESIGN TOGETHER DATA+PROCESSING to GET VALUE
• 4V = VALUE requires “intelligence”: DATA SCIENCE (+ “true” IA next?)
On Knowledge, Science and Data

- Knowledge = Culture based in Stories
- Science without numbers?
- Galileo: modern physics requires measurements
- French revolution: Nature and Science
- Industrial Revolution
- Technological revolution

- Anthropocene

- Eventually, we are only data? then, DATA SCIENCE
DEEP Learning as a Service

DEEP Open Catalog and marketplace

- Collection of ready-to-use modules (for inference, training, retraining, etc.)
  - Machine learning, deep learning, big data analytics tools + TOSCA templates
  - GitHub: https://github.com/deephdc
  - Docker Hub: https://hub.docker.com/u/deephdc
- Based on DEEPaaS API component
  - Expose underlying model functionality with a common API
  - Based on OpenAPI specifications
  - Minimal modifications to user applications.
- Final goal: execute the same module on any infrastructure:
  - HPC, laptop, server, Kubernetes, Mesos, FaaS, etc.
DEEP Learning as a Service

DEEPaaS API

- Offering models as a service implies management of APIs and web applications.
  - Scientists need to know what a (REST) API is
    - GET, POST, PUT, etc.
  - Lack of API consistency → hard for external developers to consume them.
- Our solution: provide users with a generic API (OpenAPI) component where their application can be plugged with no modifications.
Deep Learning in the CLOUD

- High-level GUIs for the composition of application architectures to be deployed on the Cloud.
- Mechanisms to encapsulate and execute Machine Learning models across platforms.
- A catalog of models aimed at end-users.
Use Case Goals

- **Objectives:** Integrate different and heterogeneous data sources: satellite data, real-time monitoring system based on sensors, observations, and meteorological data to feed the hydrological and water quality models, thus automating modeling and prediction of water quality.

- **XDC Services Requirements:**
  - XDC IAM
  - Onedata:
    - Onedata Attachment
    - Onedata Discovery
  - PaaS Orchestrator
DATA SCIENCE MASTER

OUR TIME IS CHARACTERIZED BY THE AVALANCHE OF DATA OF ALL KINDS, AND IN ALL FIELDS: EXTRA KNOWLEDGE OF THE DATA IS THE NEW ENGINE OF SCIENTIFIC, TECHNOLOGICAL, AND ECONOMIC DEVELOPMENT.

ARE YOU A PHYSICIST, A MATHEMATICIAN, AN ENGINEER, AN ECONOMIST ...?
Final remark on the Anthropocene

• All our knowledge is based in our experience (even if indirect)
  • PERCEPTION
  • NEURAL PROCESSING
  • ACTION
• We are designing and building bio-inspired computers and robots
  • EXTENDING OUR PERCEPTION
  • ASSISTING OUR ACTIONS
  • MORE AND MORE POWERFUL COMPARED TO OUR BRAIN...
• Sooner or later we will have a AI system more powerful than us
• Two questions:
  • Will science be done by machines?
  • Will we be smart enough to provide education (ethics included?)

BIG DATA HYPE IS DEAD, BUT DATA SCIENCE + AI IS NOT!