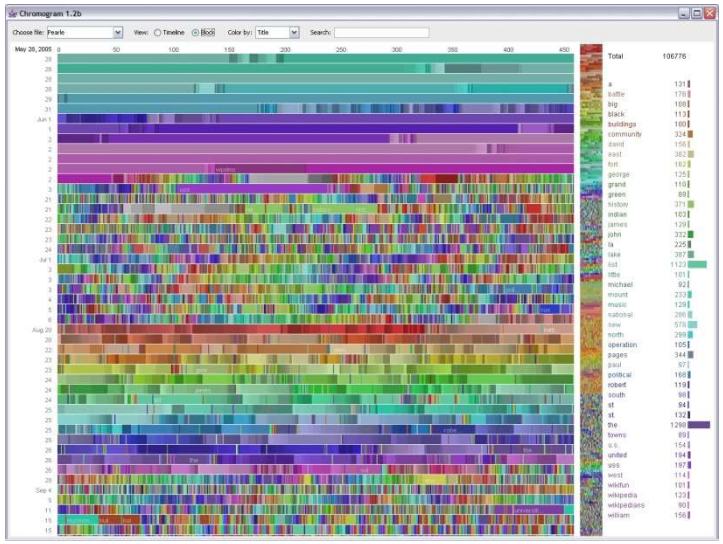
Una visión de BIG DATA

- Big Data según 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."



- Un poco de historia y evolución:

*"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**."*

Concepto de BIG DATA: Introducción

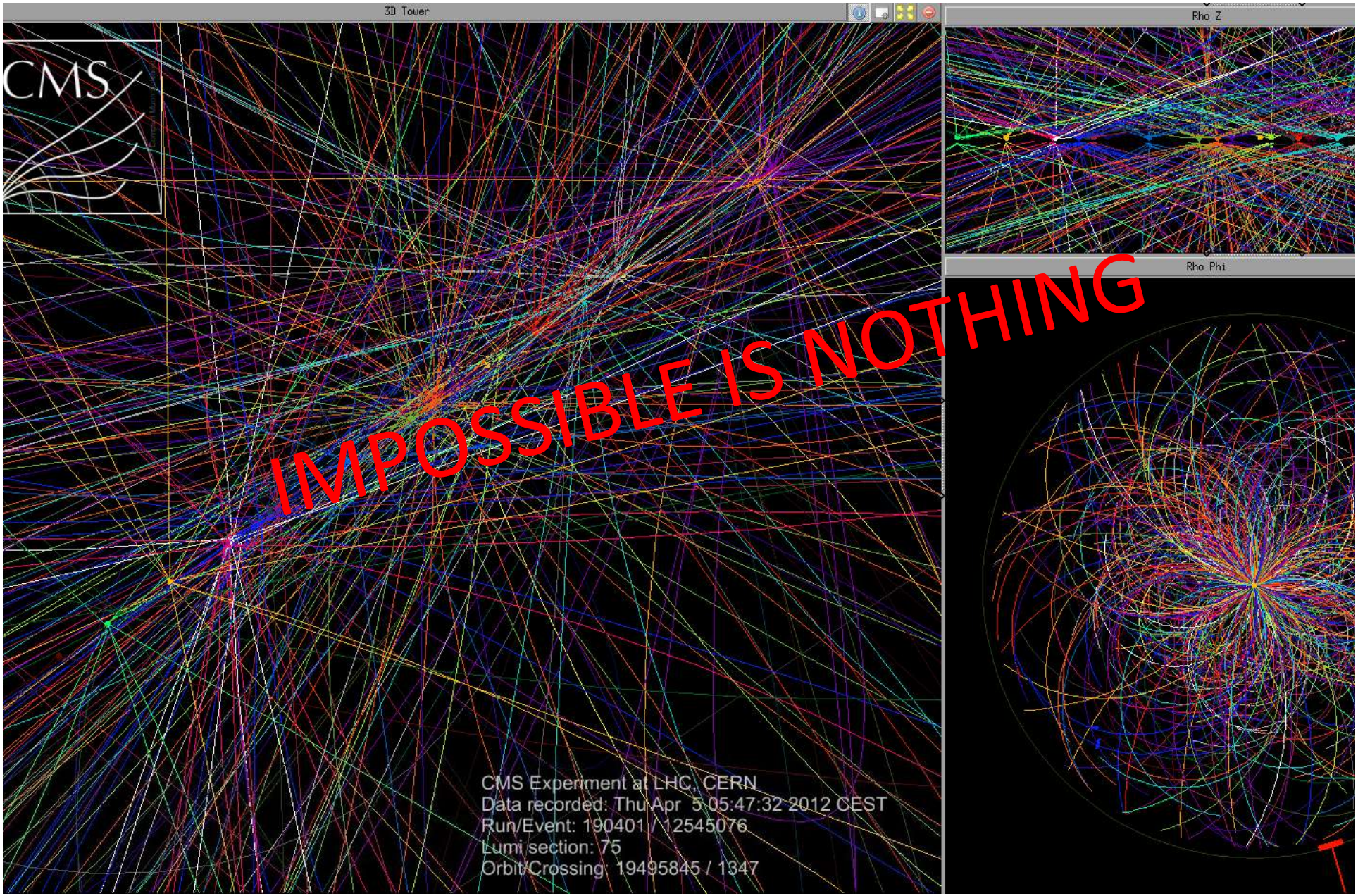
- En estos últimos años, los ámbitos **empresarial**, académico, investigador y de la **administración** han estado haciendo frente a la avalancha de datos, con la ayuda de un nuevo término, el Big Data.
- ¿Cómo podemos definir Big Data?
  - *"es el término que describe grandes volúmenes de datos (de terabytes pasamos a **zetabytes**) que se generan a gran velocidad (pasamos de datos en lotes/archivos a datos en "**streaming**"), con una posible componente de complejidad y variabilidad en el formato de esos datos (pasamos de datos estructurados a datos semi-estructurados o **no estructurados**) y que requieren de técnicas y tecnologías específicas para su captura, almacenamiento, distribución, gestión, y **análisis** de la información*".
- Otras "definiciones":
  - *"Se considera Big Data cuando el volumen de los datos se convierte en sí mismo parte del problema a solventar" (O'Reilly Radar).*
  - *"Las tecnologías de Big Data describen un nuevo conjunto de tecnologías y arquitecturas, diseñadas para **extraer valor** y beneficio de grandes volúmenes de datos con una amplia variedad en su naturaleza, mediante procesos que permitan capturar, descubrir y analizar información a alta velocidad y con un **coste reducido**". (EMC/IDC)*

Una visión de BIG DATA: Introducción

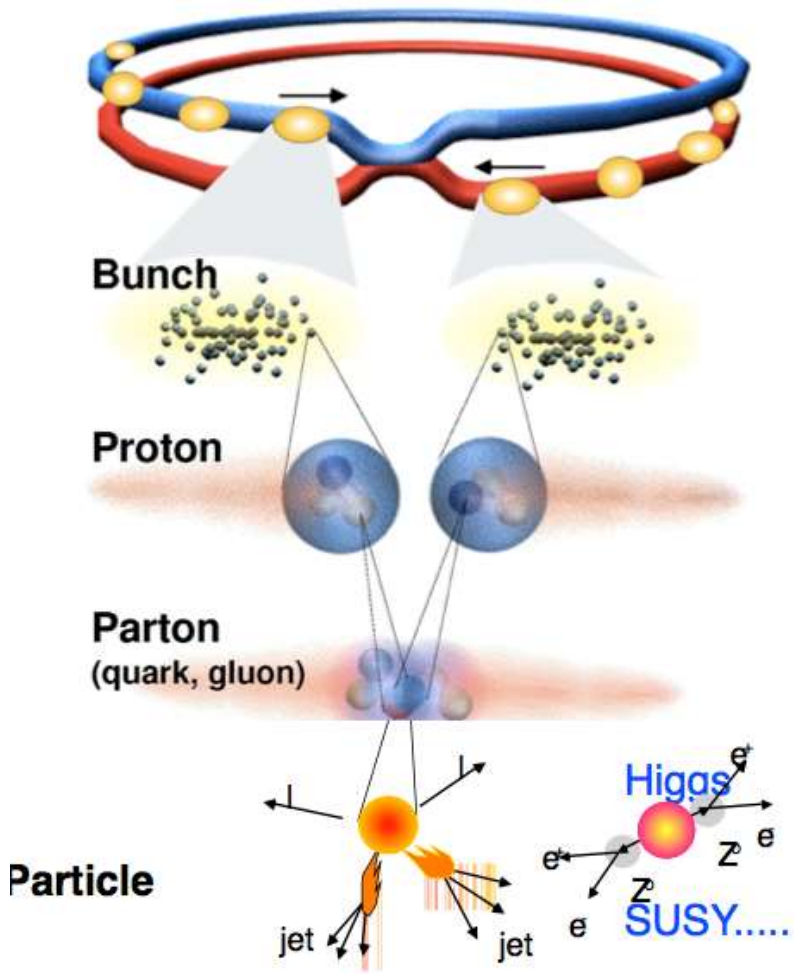
- **Big Data no es una tecnología en sí misma,** si no más bien un planteamiento de trabajo para la obtención de valor y beneficios de los grandes volúmenes de datos que se están generando hoy en día.
- Se deben contemplar aspectos como los siguientes:
  - Cómo capturar, gestionar y explotar todos estos datos.
  - Cómo asegurar estos datos y sus derivados, así como su validez y fiabilidad.
  - Cómo disponer la compartición de estos datos y sus derivados en la organización para la obtener mejoras y beneficios.
  - Cómo comunicar estos datos y sus derivados (técnicas de visualización, herramientas, y formatos) para facilitar la toma de decisión y posteriores análisis.

- ¡DEBEMOS CONSTRUIR UNA "VISIÓN" PROPIA DE BIG DATA!
  - Ejemplo: en investigación la tecnología GRID nos ha permitido resolver el reto del procesado de datos de LHC, que "era" un problema Big Data.
  - Para construir esta "visión" **necesitamos conocer la tecnología disponible**
  - Estar "al tanto" de los desarrollos tecnológicos es un reto en sí:
    - Evolución muy rápida de técnicas y capacidades
    - Dificultad de separar interés real y el interés profesional/comercial

Capacidades Técnicas



Una extraordinaria maquina muy compleja…

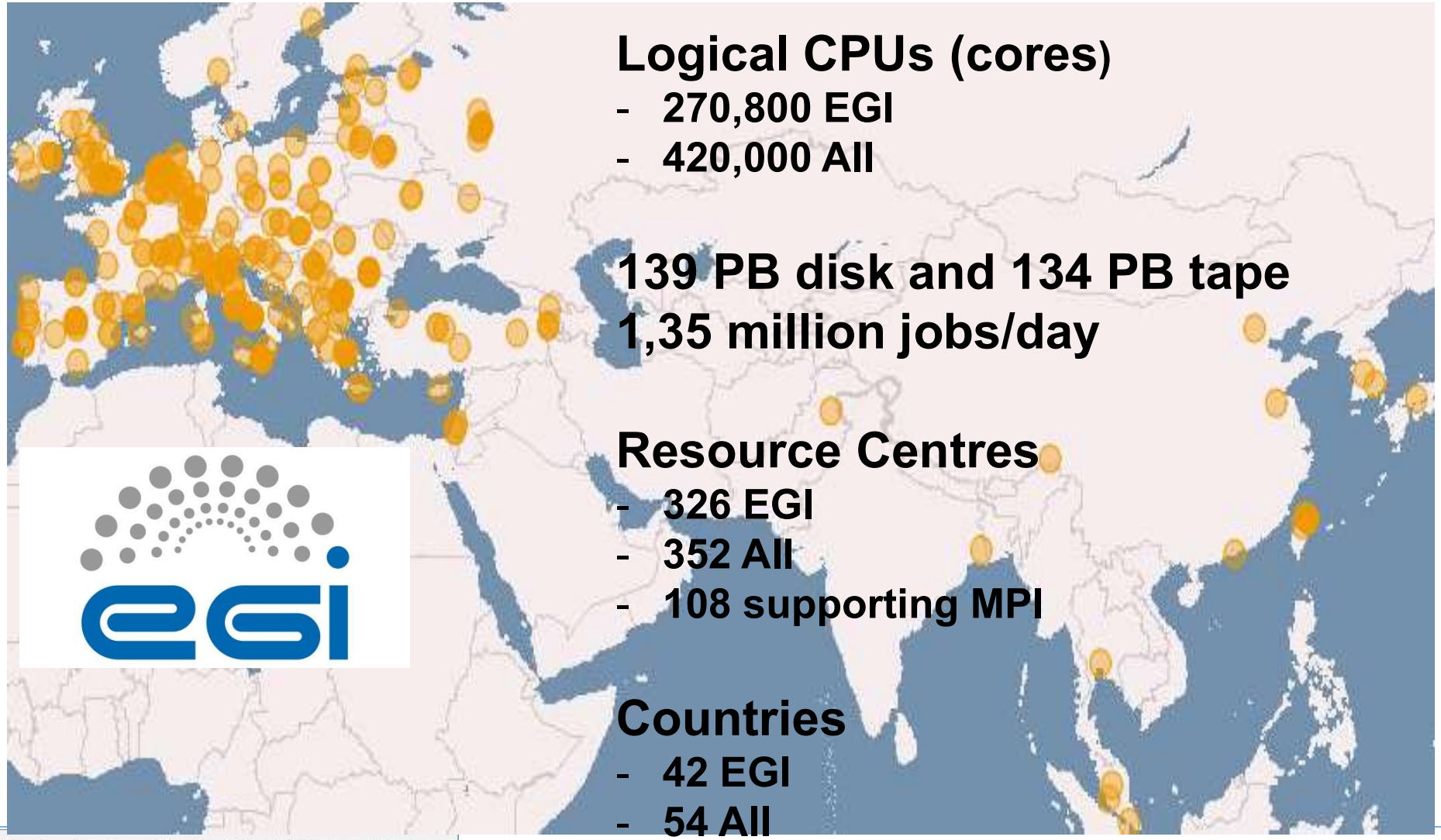


Proton - Proton	2808 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	$10^{34} cm^{-2} s^{-1}$
Crossing rate	40 MHz
Collision rate ≈	10^{7}-10^{9}

New physics rate ≈ .00001 Hz

Event selection:
1 in 10,000,000,000,000

European Grid Infrastructure (2012)



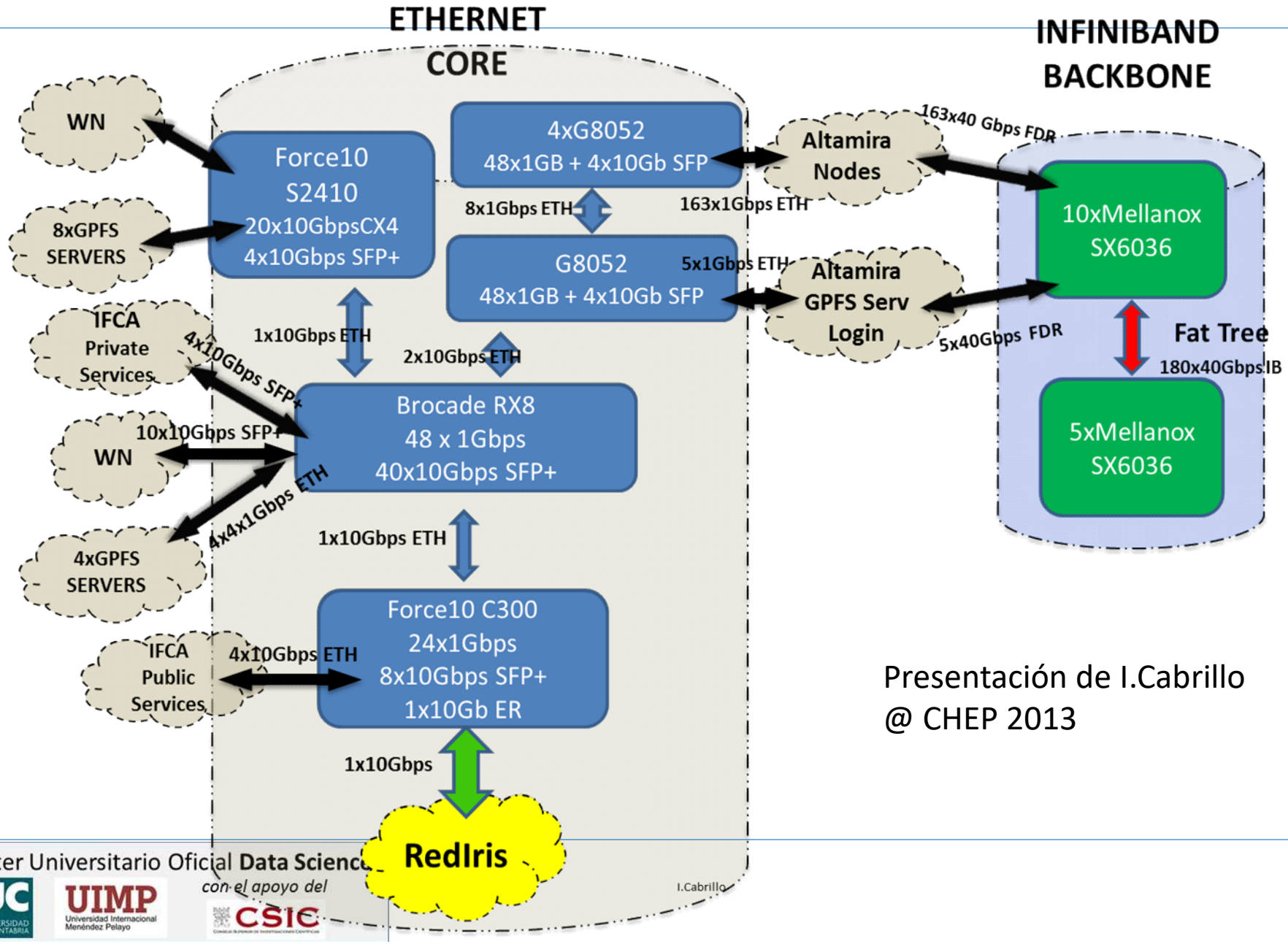
Logical CPUs (cores)
- 270,800 EGI
- 420,000 All

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

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

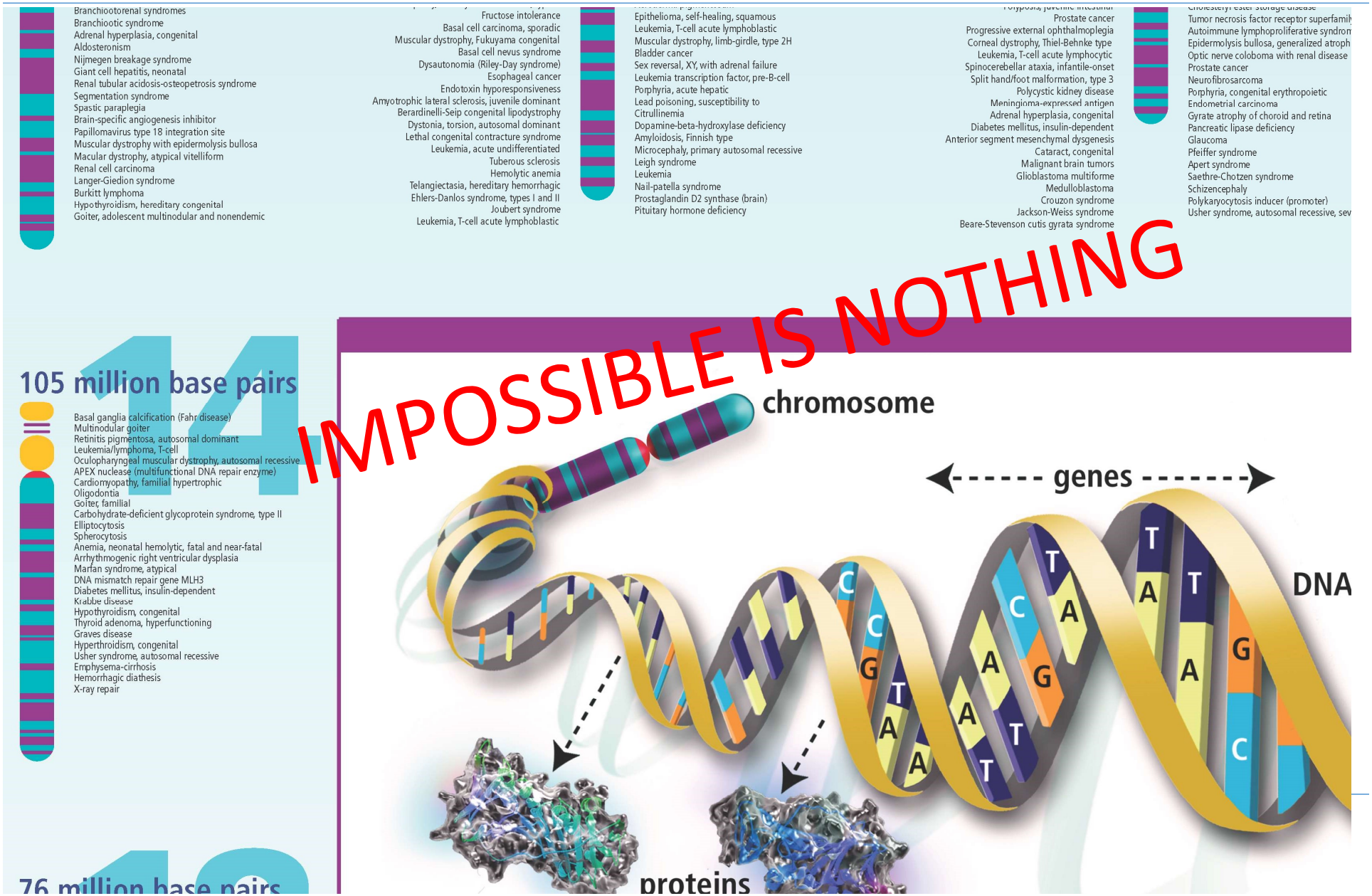
Countries
- 42 EGI
- 54 All

ALTAMIRA INTEGRATION: Network View



Presentación de I.Cabrillo
@ CHEP 2013

Capacidades Técnicas



The $1,000 genome is a term that predicts a new era of predictive and personalized medicine during which the cost of FULL genome sequencing an individual or patient drops to roughly USD$1,000

Privacy of individuals undergoing genetic testing must be protected under all circumstances


ict 2013
Create
Connect
Grow
6-8 November
Vilnius, Lithuania
How will Big Data yield a new economy?
Ecosystem
investment funds
public & private R&D
public launching customers
data reuse
enabling infrastructure
Application domains
industrial
business decision support
services
cross-sector?
@ICT2013
Health
Value chain mechanisms
new business functions
collaborative models
ownership
Hurdles
job destruction!
corporate psychology
privacy!
data ownership
fast-pace development
Technologies
networks
HPC & cloud
analysis & visual.
storage
Semantic web
Drivers
open data
market making
business cases, examples
sandboxes

rdles

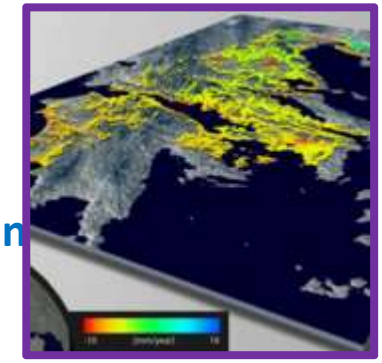
corporate psychology

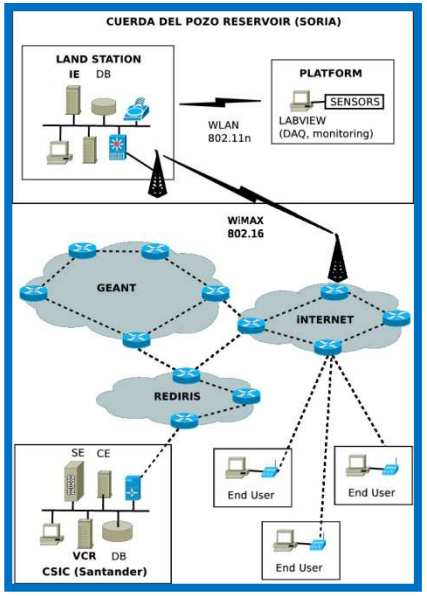
privacy!

- legal
- technological

Capacidades Técnicas: Adquisición y Transmisión de Datos

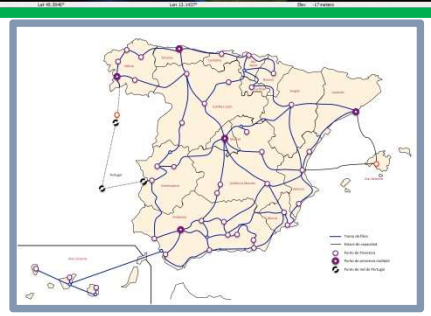
- Fuentes de Información (Sensores):
  - Instrumentación
    - Redes de sensores
    - Cámaras
    - Satélites
  - Uso personal
    - Smartphones
    - Automóviles
    - Instrumentación person
  - **Mensajería en la red**
  - **OPEN DATA**
- Ejemplos Globales:
  - **ESA's GMES (Observing the Earth)**
  - SmartCities
- Integración: **Estándares: Sensor Web Enablemen**





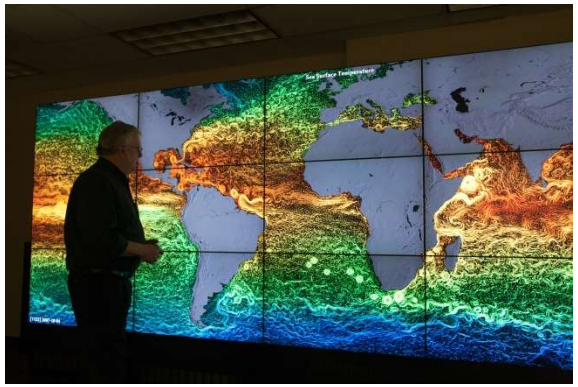
Capacidades Técnicas: Infraestructura

- Centros de procesado de datos
  - Supercomputadores top500: hasta ~100 Petaflops
    - España: Red Española de Supercomputación (RES)
  - GRID: hasta 700.000 cores (WLCG), >200 Petabytes
    - España: IberGrid
  - Componentes:
    - Almacenamiento: HADOOP, GPFS, Lustre…
    - Clusters: Redes Infiniband
- Redes de comunicación
  - España: RedIris-Nova (fibra oscura, n x 10Gb/s)
- Cloud
  - Amazon , IBM, Google, MS Azure

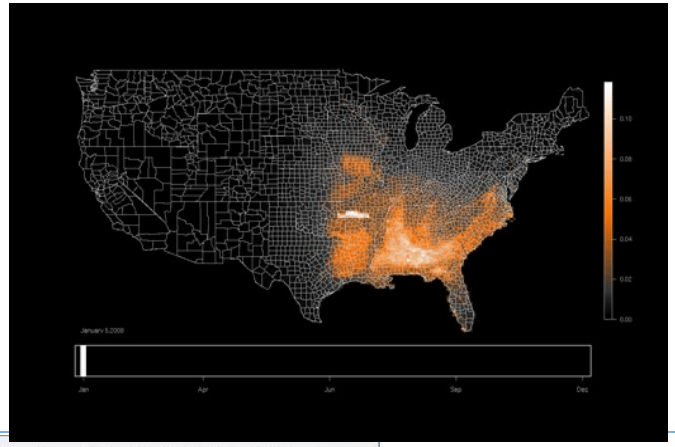


Presentación

- **“Hardware”: Dashboards, Walls, Visores personales 3D**



- **Software/Gráficos:**

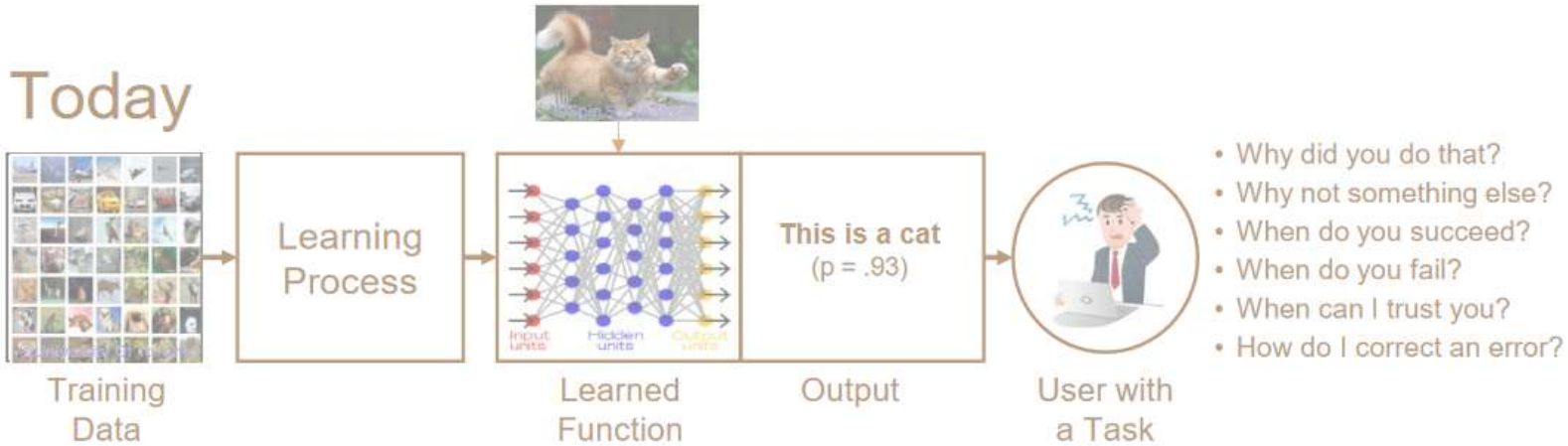


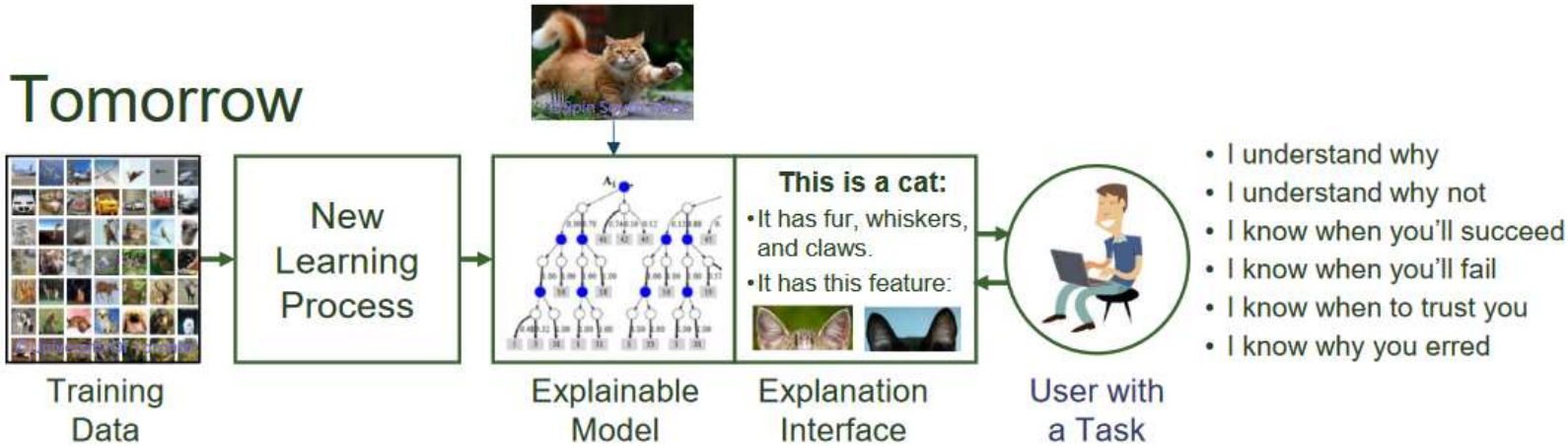
My favorite digital images of last years:

-my digital fingerprints plus an updated photo
-an accurate 3D profile of my face (TLS)
-an accurate 3D profile of my body (AMW)
-my country, my city, my street
-my car parked in front of my home
-my photo crossing a finish line, with my timing

Joining my favorite digital info of last years:
-my “private” gmail
-my searches in Internet
-my “private” wifi information (keys included)

Explainable AI – What Are We Trying To Do?





Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

9

BREAK

- Mañana veremos un ejemplo sencillo de red neuronal.

Ejercicio 1

- Analiza cómo problema de Open Science el descubrimiento del bosón de Higgs (2012)
  - Publicado en abierto
  - ¿Datos en abierto?
  - ¿Transparencia? ¿Reproducibilidad?

•*"Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". Physics Letters B.* ***716****: 1. 2012. Bibcode:2012PhLB..716....1A. arXiv:1207.7214 doi:10.1016/j.physletb.2012.08.020.*

•*"Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". Physics Letters B.* ***716****: 30. 2012. Bibcode:2012PhLB..716...30C. arXiv:1207.7235 doi:10.1016/j.physletb.2012.08.021.*

Ejercicio 1

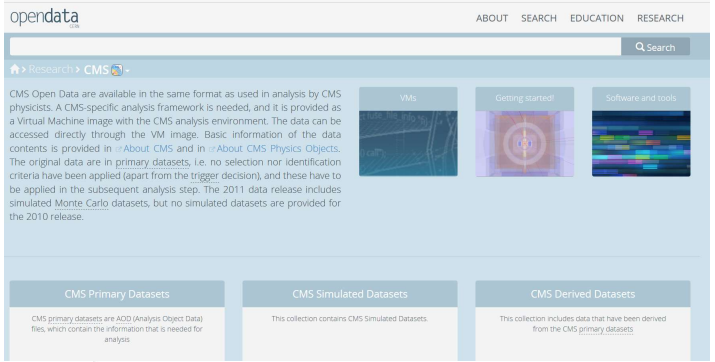
- Analiza cómo problema de Open Science el descubrimiento del bosón de Higgs (2012)
  - **Publicado en abierto**
  - ¿Datos en abierto? **http://opendata.cern.ch/research/CMS**
  - ¿Transparencia? ¿Reproducibilidad?

First open-access data from large collider confirm subatomic particle patterns

September 30, 2017 by Jennifer Chu

1.8K Like G+ Tweet 1 reddit Favorites Email Print PDF

The Compact Muon Solenoid is a general-purpose detector at the Large Hadron Collider. Credit: CERN



Ejercicio 1

- Revisa el estado cómo problema de Open Science de la detección de ondas gravitacionales (Nobel 2017)
  - Publicado en abierto
  - ¿Datos en abierto?
  - ¿Transparencia?

https://es.wikipedia.org/wiki/Deteccion_de_ondas_gravitacionales

[116] LIGO Open Science Center (LOSC) https://losc.ligo.org/events/GW150914/

Ejercicio 2

- Identifica el interés y los problemas de un enfoque “Open Science” para el modelado del embalse de Cuerda del Pozo
  - Datos externos para el modelo

    AEMET, IGN, CHD,
  - Publicación de datos propios

    Servidor DBMS doriie
  - Modelo en código abierto (open source):

    Delft-3D, DELTARES Modelo de negocio Open?
  - Publicaciones del área en abierto

    WoS:
  - Validaciones en la literatura

  - Modelo de explotación de los resultados

    ?

¿Donde estamos en Open Science/ Open Access?

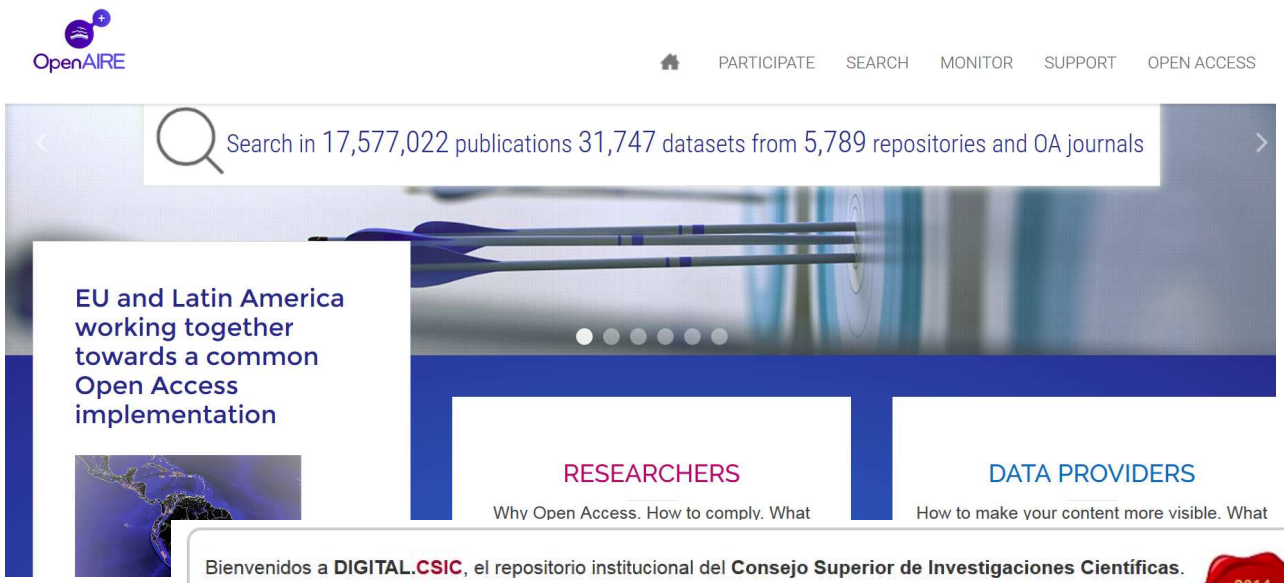
Open Science and Open Access

*Open Access, as defined in the Berlin Declaration,[1] means unrestricted, online access to peer-reviewed, scholarly research papers for reading and productive **re-use**, not impeded by any financial, organisational, legal or technical barriers. Ideally, the only restriction on use is an obligation to attribute the work to the author.*

OpenAire; los expertos de las Bibliotecas…

Launched at the 'Berlin 12' conference in December 2015, the OA2020 initiative aims to accelerate the transition by transforming subscription-based scientific journals to OA business models. OA2020 is based on a financial analysis published by the Max Planck Digital Library. According to the analysis, there should be enough money in the system to allow for a transition to OA at potentially neutral cost.
The OA2020 initiative is outlined in an Expression of Interest statement that was endorsed by 53 parties at the end of June 2016, including SE Mos (SNSF, CSIC, NWO, MPG, Leibniz Association, FCT, DFG and FWF).

OpenAire, Digital CSIC



Master Universit
UC UNIVERSIDAD DE CANTABRIA
UIM Universidad Internacional Menéndez Pelayo

¿Donde estamos en Open Science/ Open Access?

Open Science and Research Data Management (RDM)

Given the diversity in Europe, a common vision, strategy and funding practice is not easy to accomplish. The increasing shift to an Open Science approach offers a good starting point for the layout of a layered, component-based RDI with complementary RDM support functions at various levels: international/ national/local and mono/inter/multi-disciplinary, offering various types of RDI services (computing, storage, network, data, research support, training and education).

Text Data Mining

- ***Science Europe TDM Workshop***

Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research

Lucie Guibault, professor of information law at the University of Amsterdam, pointed out that TDM is not directly mentioned in the Directive on Copyright in the Information Society (Infosoc Directive). Nevertheless, the technology is hindered by the directive because TDM regularly involves making copies of the works to be mined. This infringes the reproduction right that is broadly protected by Article 2 of the Infosoc Directive. Copying protected works either needs an appropriate license or an exception within the law.

Professor Guibault focused particularly on Article 5 of the Infosoc Directive, which lists exceptions to the reproduction right and the right of communication to the public. Article 5(1) of the Infosoc Directive allows for broad "transient and incidental reproductions" of copyrighted works. However, transient copies for TDM purposes are only allowed if they are an integral and essential part of a technological process whose sole purpose is to enable (a) a transmission in a network between third parties by an intermediary, or (b) a lawful use. However, a use can be 'lawful' if it is authorised by either the rights owner or by law. Since there is no specific provision in the Directive authorising TDM, this means that the article does not provide a guarantee of the right to carry out TDM without the consent of rights holders.

¿Estrategia para llegar a una Ciencia en Abierto?

CSIC Vision 2020: "A Digital Knowledge Strategy to support Open Science"

"Open Science is a broad term, covering the many exciting developments in how science is becoming more open, accessible, efficient, democratic, and transparent. This Open Science revolution is being driven by new, digital tools for scientific collaboration, experiments and analysis and which make scientific knowledge more easily accessible by professionals and the general public, anywhere, at any time..."

The experience from different research projects and activities lead by CSIC (Spanish National Research Council), indicates that supporting this concept of Open Science is a key first step towards a new way of discovering, sharing and preserving knowledge.

Three key components must be considered under this approach:

A) ***Open Access to research publications****, enabling direct access, without any kind of restriction, registration or subscription.*

B) ***Enhanced Research Data Management****, covering the full data cycle, from planning, acquisition and curation to publication, integration in analysis and preservation.*

C) ***Advanced e-Infrastructures*** *enabling the process of large datasets, the mining of scientific databases and literature, as well as the distributed collaboration among researchers at all levels, including the contribution from citizen science.*

¿Estrategia para llegar a una Ciencia en Abierto?

- **Under the vision proposed for 2020:**

-Researchers are able to directly explore, access and use research data and publications of different areas when preparing new interdisciplinary studies, employing a well defined framework. They will be able to integrate and analyze the data, using the required computing infrastructure, and also to store and publish the new results including a description of the analysis under a semantic framework so they can be further shared and preserved.

-Relevant data and analysis results published will be further explored, re-used and referenced by the research community, and proper recognition to their quality and impact attributed to the authors

-Adequate technical and financial support is provided to these Open Science pillars, including the formation of new specialists and the dissemination of the techniques and results.

-Citizens are engaged in the support of science, being able to directly explore new results and contribute, when possible, in different ways, from data provision to crowd sourced tasks.

-The research initiatives launched to target the integration of the semantic framework in the Open Science context provide successful examples of interdisciplinary achievements.

¿Estrategia para llegar a una Ciencia en Abierto?

A) Open Access (OA)

1) Research publications are one of the main results of research process. Both Research Performing and Research Funding institutions share the vision of increasing the impact and reducing the costs of research publications by moving to a system of Open Access

- How to assure that research publications are either published in an Open Access journal or deposited, as soon as possible, in a repository?
- It is crucial to support any valid approach to achieve Open Access goals (green-gold), recognising repositories as a key strategic infrastructure.
- **The hybrid publication model as currently defined and implemented by publishers, is not a working and viable pathway to Open Access. The "double dipping" must be prevented and publishers cost transparency improved.**

2) Open Access is not only about the right of access, but also about the use and re-use information, subject to proper attributions.

- The final goal is to shift to a research publication system in which free access to research publications is guaranteed. This involves a move towards Open Access, replacing the present subscriptions system with other publications models redirecting and reorganising the current resources accordingly.

¿Estrategia para llegar a una Ciencia en Abierto?

B) Enhanced Research Data Management (RDM)

1) There is a clear need for a (common EU) policy for RDM activities

- The framework must establish what structure must be used to assure an effective organization of RDM activities, which responsibilities in the data cycle should be defined and how to assure that curation activities are close to the required expertise.

2) How to convince all actors (RFO, RPO, research teams) of the importance of RDM activities and open data reuse and exploitation?

- By defining institutional policies, and enforcing them, for example new indicators for assessment exercises.
-Data licensing issues should be carefully considered to guarantee proper attribution (following for instance open source software licensing experience)

3) Definition of the scope of the RDM activities, and in particular long term preservation, for the datasets collected/produced in a given project

-In that sense it should be established how to define the interest of the datasets, and associated software and recipes; how to promote that researchers use correctly data embargo and also that open data and metadata formats employed are useful for reuse, and finally how to balance the investment required with this interest on reuse.

¿Estrategia para llegar a una Ciencia en Abierto?

C) e-Infrastructures (e-INFRA)

1) e-Infrastructures must be offered and accessed in a unified way as services supporting Open Science

-Assuring a coordination of the different Data, HPC and Distributed Computing/Cloud Computing resources

-Providing Single Sign On mechanisms and tools for management of Virtual Organizations.

2) Virtual Research Environments must be productive for researchers

-Enabling new capacities/capabilities (access to new algorithms, to new resources, simplifying the deployment of new applications)

-Providing a transparent way to share data, analysis and discussions.

¿Estrategia para llegar a una Ciencia en Abierto?

Additionally, two clear findings are transversal to these pillars:

INTEGRATION: How to support RDM and OA activities using services on top of e-infrastructures

-Guaranteeing the closeness to institution/experts and guarantee national involvement

-Exploiting an adequate scale factor

FUNDING: How to assure a baseline funding for RDM, OA and e-INFRA

-Considering a formal overhead (3-10%) for any project, depending on the weight of these activities

-By transforming a punctual funding into a long-term budget within the institution

EU to the rescue!

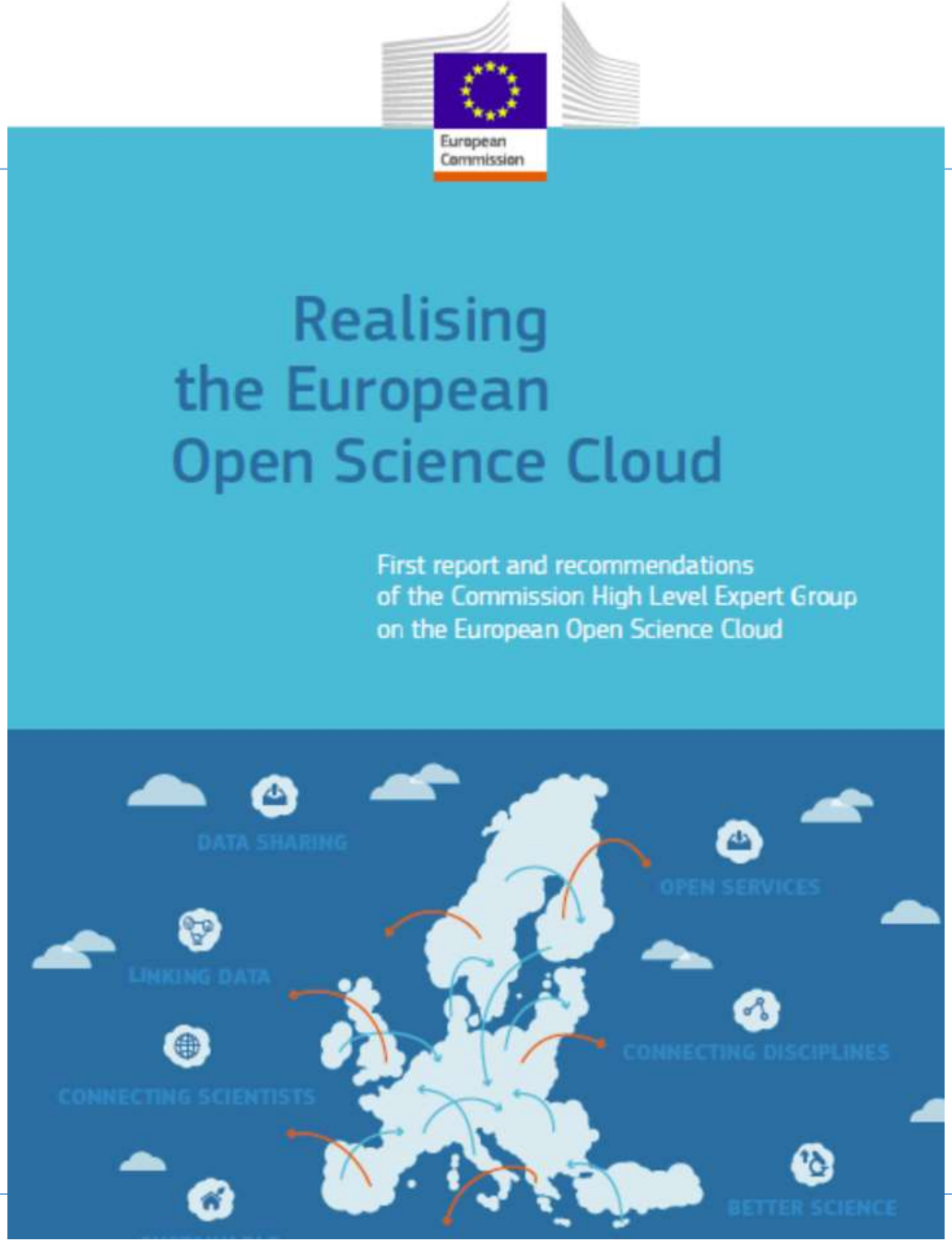
- ***European Open Science Cloud***

New H2020 calls

~250M€ in 2017-2020

Complex!

Research + e-Infrastructures



“Collateral” impacts

- ***Research Information and Open Science!***

Research Performing Organisations (RPOs) and Research Funding Organisations (RFOs) collect and use data about their own activities from various and heterogeneous sources. This kind of data – data about research activities rather than research data generated by researchers – is stored in research information systems.

RPOs and RFOs use research information systems for a variety of different purposes, such as monitoring and evaluating research activities and outputs, allocating funding, supporting decision making on their policies and strategies, tracking researchers’ careers, and describing their systemic role to policy-makers, stakeholders and the public.

As a result, decision makers and research organisation managers alike increasingly depend on indicators, reports and studies that draw data from research information systems.

“Collateral” impacts

- ***Open Data and Open Access open a lot of possibilities…if you have resources***
- ***Example: Deep Learning applied to image recognition (ex: how to become a world expert on plants identification!)***
- ***Is it true that Google and US-GS store whole (all time) Sentinel ESA data (and EU is not able to do it? Ask Javier)***
- ***Who can track all drugs and clinical essays in the market?***

Tarea a completar

(deadline miércoles 20h)

Name of the team
Contact person
E-mail
Organisation (university, company, etc.)
Background (education, work, etc.)
Language of communication

¿¿¿ EXAMINAR DATASETS Y ENCONTRAR IDEAS
o
PENSAR EN IDEAS Y BUSCAR DATASETS ???

Which EU dataset do you want to use (URL)?

Short description of the idea: **which question/challenge will you address?**

Short description of the idea: **how will you present the data?**

Short description of the idea: **how would you further develop this idea?**

JRC: Datos Bio/Medio Ambiente
Ejemplo: Impacto de los asentamientos urbanos en las zonas Natura 2000
-cambio democrático? Compromiso con la naturaleza, evaluación especies invasoras
-puestos de trabajo? Ideas en torno a zonas Natura 2000
COPERNICUS: Satelites
Ejemplo: cambio en las zonas Natura2000
GBIF:
Ejemplo: registros con geolocalización en zonas Natura 2000