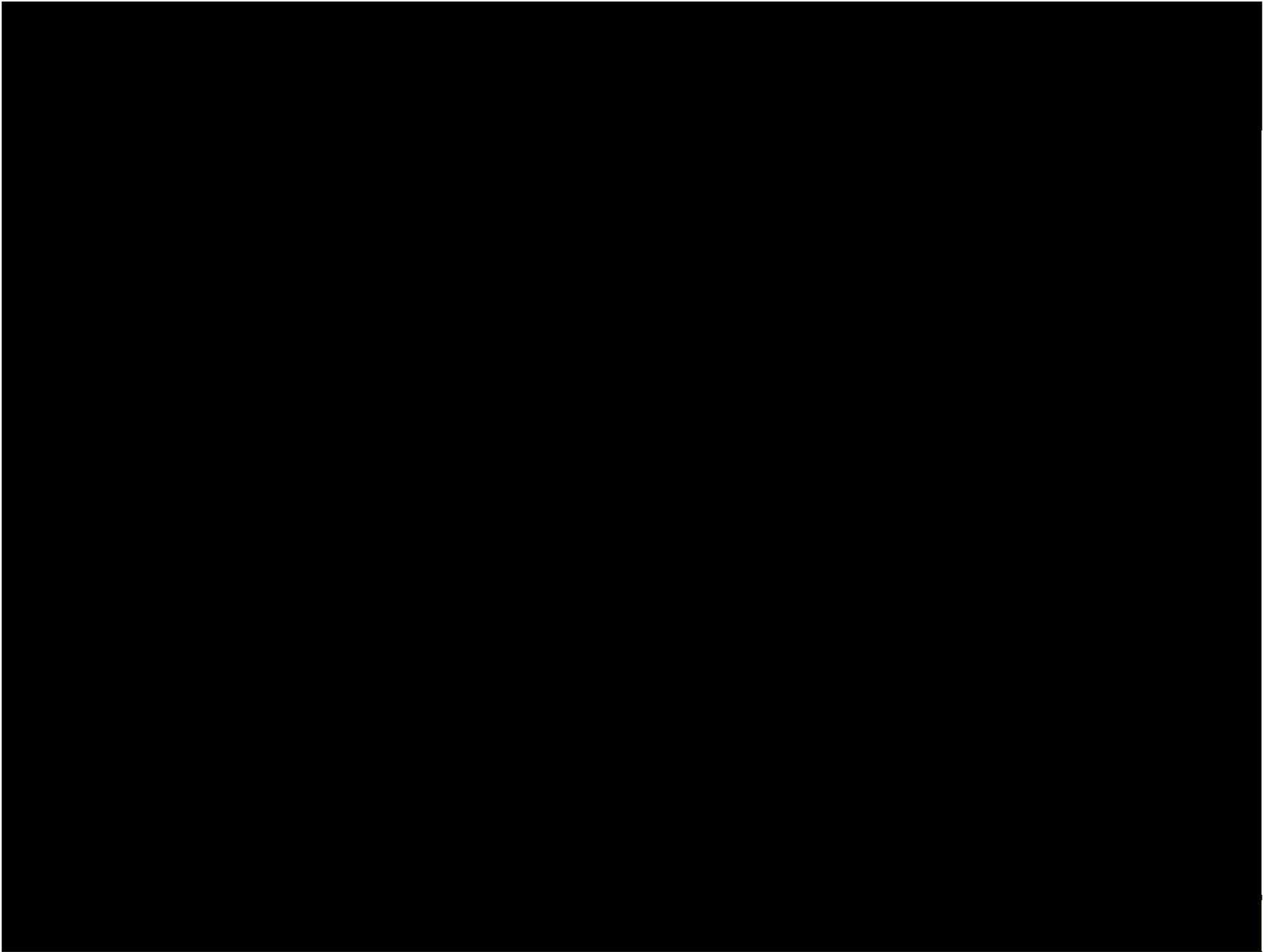


# ***BIG DATA Y PRIVACIDAD***



Jesús Marco de Lucas, Instituto de Física de Cantabria, CSIC-UC



These slides were prepared in 2013, after I suffered a

## "brain phase change"

Why?

-Finished a short course on genetics (MIT)

-Did an intro course named "Justice" (Harvard)

-Visited several US labs

-EGI-TF, CHEP 2013, LHC computing, e-IRG, ICT2013 Vilnius

*And after I read a book in Stansted Airport, I started to prepare these slides*

-Presented in a meeting on Defense (2013)

*All opinions/facts presented in these slides are on my own*

Two years later... the war is lost

Privacy is no longer possible

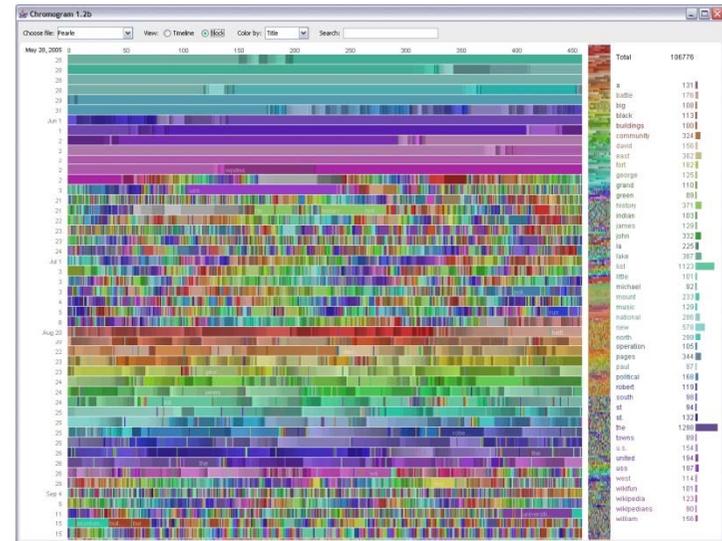
- Participated in COPERNICUS meeting (2014)
- CITIZENFOUR (Snowden) (2014)
- Discussion with experts at Edinburgh (2015)
- Today I am travelling to e-IRG, Luxembourg



# 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

GOBIERNO DE ESPAÑA  
MINISTERIO DE DEFENSA  
SECRETARÍA DE ESTADO DE DEFENSA  
DIRECCIÓN GENERAL DE ARMAMENTO Y MATERIAL

Portal de Tecnología e Innovación del Ministerio de Defensa

MAPA WEB CONTACTO

BUSCAR

Presentación Estrategia de Tecnología e Innovación Contenidos del Portal Contacto y Participación

Inicio ETID > Contenidos del Portal > Referencia de interés

## Referencias de Interés

28/11/2012

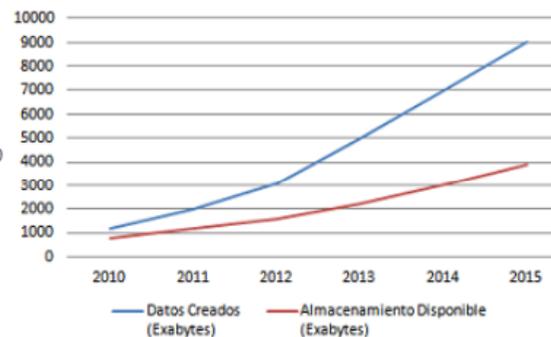
### Concepto de BIG DATA y su aplicación a defensa y seguridad

#### Ámbito: OTROS

En la película "Cortocircuito", de 1986, el robot Número 5 expresaba su necesidad de conocimientos con la frase: ¡datos, necesito más datos!. Hoy en día alimentar esa necesidad no sería ningún problema, solamente en 2010 se generaron más de 1200 exabytes de información (un exabyte equivale a 1000 millones de gigabytes). Enmarcando el ejemplo en el ámbito de la defensa, baste decir que durante 2009 los UAVs de EE.UU. remitieron a sus mandos de control el equivalente a 24 años de grabación.

Sin embargo, hoy en día "Número 5" tendría problemas para:

- Almacenar la información que recibiera.
- Procesar y analizar esa información en forma y en tiempo. El 95 por ciento de esos 1200 exabytes creados en 2010, eran datos no estructurados. Este volumen de datos, conjuntamente con la falta de estructura, supone un desafío técnico para los sistemas de bases de datos relacionales existentes hoy en día.



#### Eventos

May 2013

May 2013						
«		today			»	
M	T	W	T	F	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

#### Próximos eventos

**Seminario sobre Aplicaciones de Big Data en los entornos de Defensa y Seguridad**

Del 30/05/2013 al 30/05/2013

El CESEDEN/IEEE en colaboración con la Fundación Círculo e Isdefe organizan un seminario sobre "Las tecnologías de Big Data y sus aplicaciones a los entornos de Defensa y Seguridad". El evento tendrá lugar el próximo día 30 de mayo en las instalaciones de Isdefe a las 9:30 de la mañana. [Mas información](#)

# 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)*

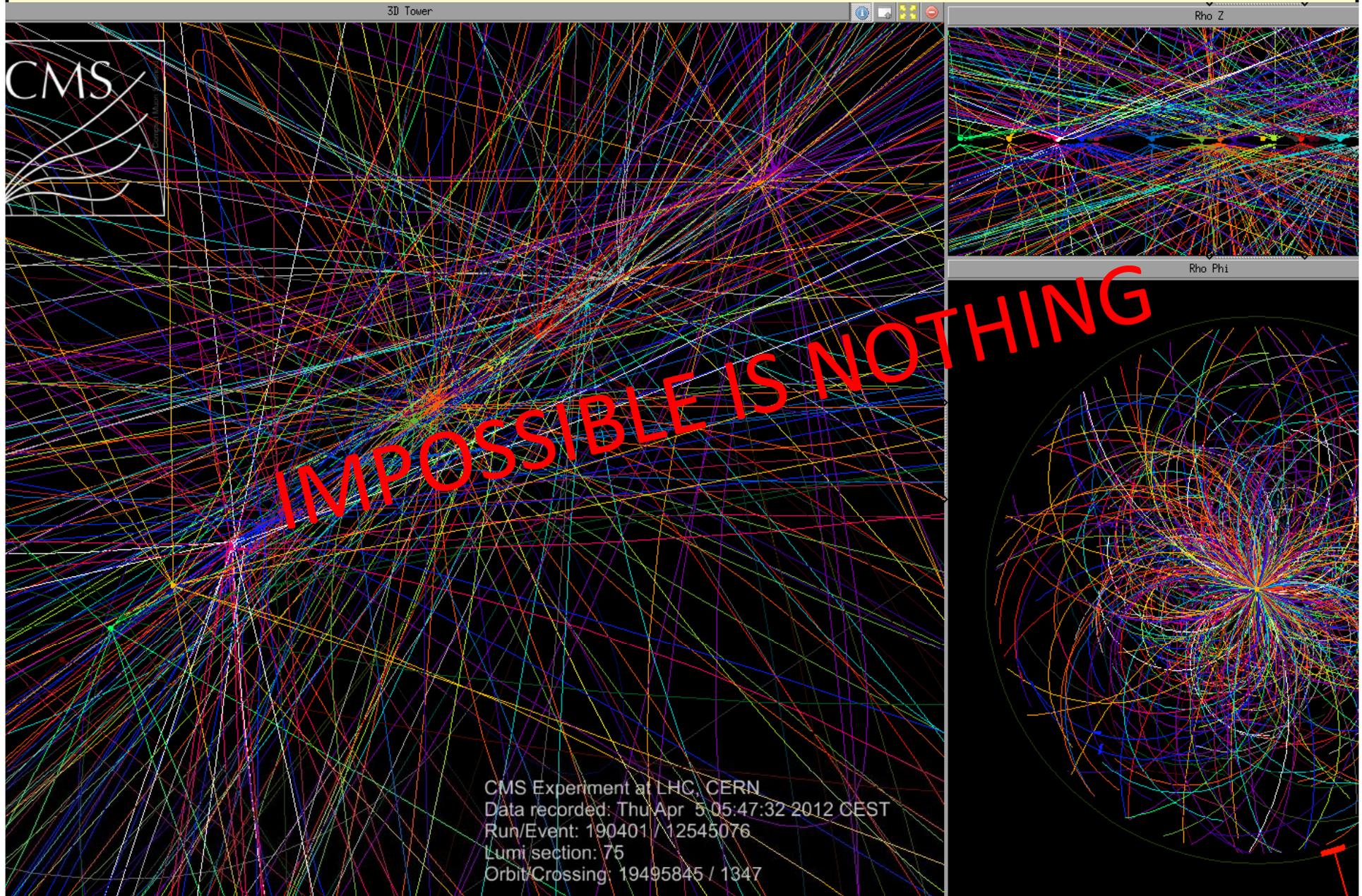
# *Una visión de BIG DATA: Introducción*

- 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.

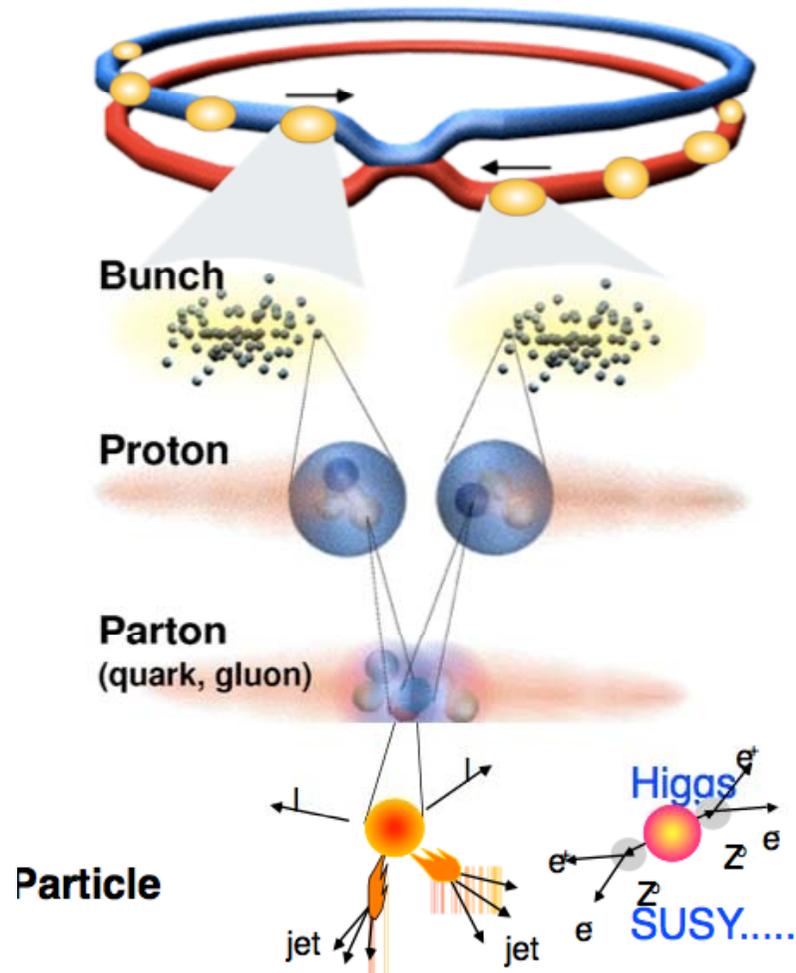
## ● ¡DEBEMOS CONSTRUIR UNA “VISIÓN” PROPIA DE BIG DATA!

- 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 ( $7 \times 10^{12}$  eV)  
**Luminosity**  $10^{34} \text{cm}^{-2} \text{s}^{-1}$

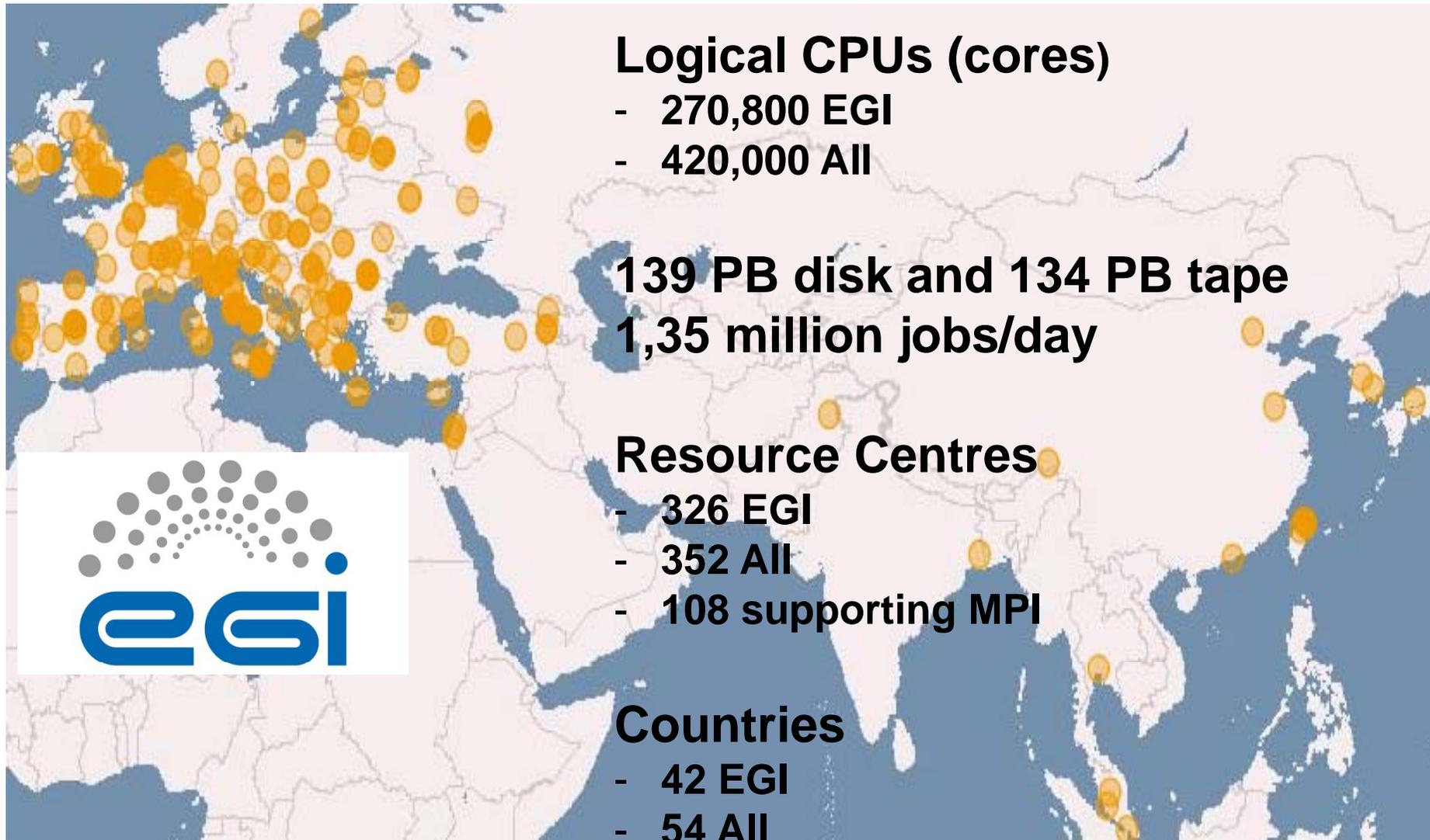
**Crossing rate** 40 MHz

**Collision rate  $\approx$**   $10^7 - 10^9$

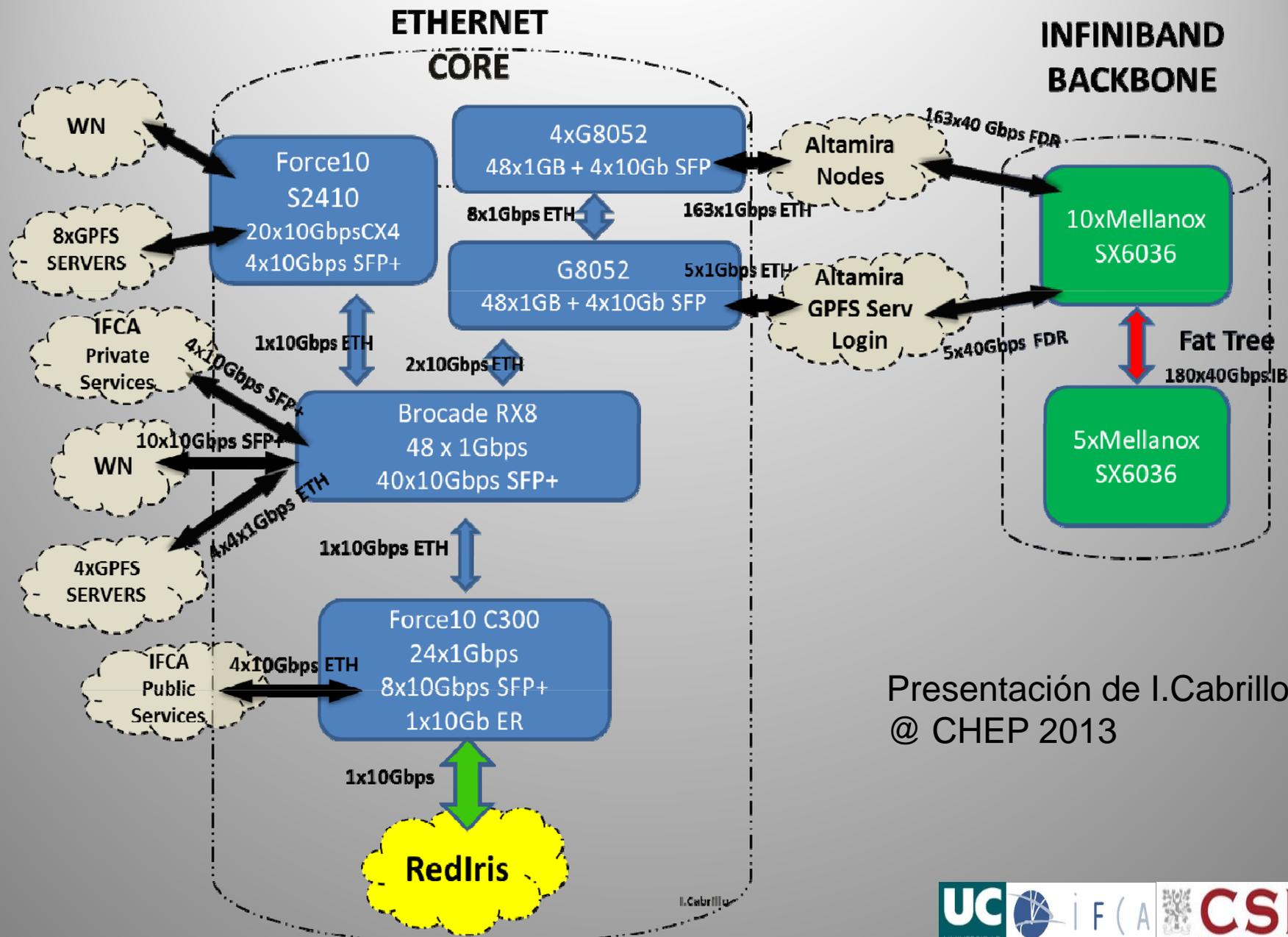
**New physics rate  $\approx$**  .00001 Hz

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

# *European Grid Infrastructure*



# ALTAMIRA INTEGRATION: Network View



Presentación de I.Cabrillo  
@ CHEP 2013

# Capacidades Técnicas



Branchiootrenal syndromes  
Branchiootic syndrome  
Adrenal hyperplasia, congenital  
Aldosteronism  
Nijmegen breakage syndrome  
Giant cell hepatitis, neonatal  
Renal tubular acidosis-osteopetrosis syndrome  
Segmentation syndrome  
Spastic paraplegia  
Brain-specific angiogenesis inhibitor  
Papillomavirus type 18 integration site  
Muscular dystrophy with epidermolysis bullosa  
Macular dystrophy, atypical vitelliform  
Renal cell carcinoma  
Langer-Giedion syndrome  
Burkitt lymphoma  
Hypothyroidism, hereditary congenital  
Goiter, adolescent multinodular and nonendemic

Fructose intolerance  
Basal cell carcinoma, sporadic  
Muscular dystrophy, Fukuyama congenital  
Basal cell nevus syndrome  
Dysautonomia (Riley-Day syndrome)  
Esophageal cancer  
Endotoxin hyporesponsiveness  
Amyotrophic lateral sclerosis, juvenile dominant  
Berardinelli-Seip congenital lipodystrophy  
Dystonia, torsion, autosomal dominant  
Lethal congenital contracture syndrome  
Leukemia, acute undifferentiated  
Tuberous sclerosis  
Hemolytic anemia  
Telangiectasia, hereditary hemorrhagic  
Ehlers-Danlos syndrome, types I and II  
Joubert syndrome  
Leukemia, T-cell acute lymphoblastic



Epithelioma, self-healing, squamous  
Leukemia, T-cell acute lymphoblastic  
Muscular dystrophy, limb-girdle, type 2H  
Bladder cancer  
Sex reversal, XY, with adrenal failure  
Leukemia transcription factor, pre-B-cell  
Porphyria, acute hepatic  
Lead poisoning, susceptibility to  
Citrullinemia  
Dopamine-beta-hydroxylase deficiency  
Amyloidosis, Finnish type  
Microcephaly, primary autosomal recessive  
Leigh syndrome  
Leukemia  
Nail-patella syndrome  
Prostaglandin D2 synthase (brain)  
Pituitary hormone deficiency

Polycystic juvenile osteoarthritis  
Prostate cancer  
Progressive external ophthalmoplegia  
Corneal dystrophy, Thiel-Behnke type  
Leukemia, T-cell acute lymphocytic  
Spinocerebellar ataxia, infantile-onset  
Split hand/foot malformation, type 3  
Polycystic kidney disease  
Meningioma-expressed antigen  
Adrenal hyperplasia, congenital  
Diabetes mellitus, insulin-dependent  
Anterior segment mesenchymal dysgenesis  
Cataract, congenital  
Malignant brain tumors  
Glioblastoma multiforme  
Medulloblastoma  
Crozon syndrome  
Jackson-Weiss syndrome  
Beare-Stevenson cutis gyrate syndrome



Cholesteryl ester storage disease  
Tumor necrosis factor receptor superfamily  
Autoimmune lymphoproliferative syndrome  
Epidermolysis bullosa, generalized atroph  
Optic nerve coloboma with renal disease  
Prostate cancer  
Neurofibrosarcoma  
Porphyria, congenital erythropoietic  
Endometrial carcinoma  
Gyrate atrophy of choroid and retina  
Pancreatic lipase deficiency  
Glaucoma  
Pfeiffer syndrome  
Apert syndrome  
Saethre-Chatzen syndrome  
Schizencephaly  
Polykaryocytosis inducer (promoter)  
Usher syndrome, autosomal recessive, sev

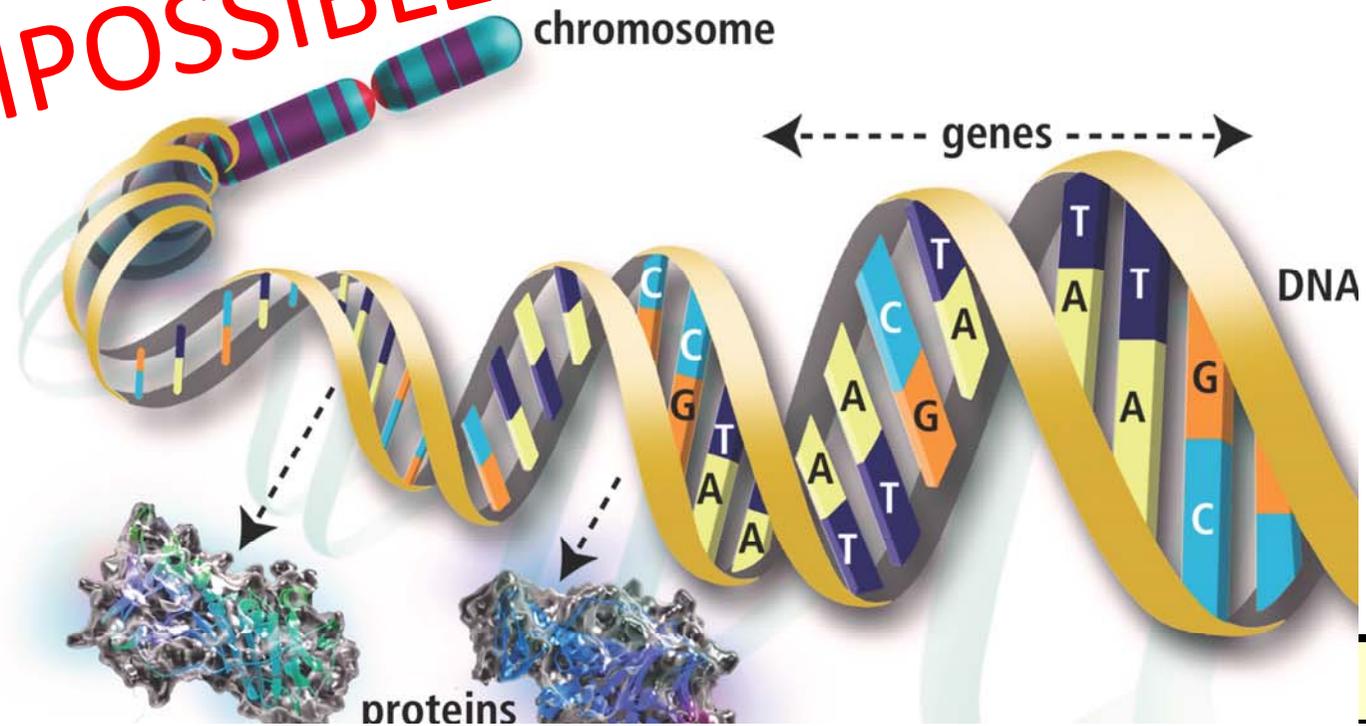
**IMPOSSIBLE IS NOTHING**

105 million base pairs



Basal ganglia calcification (Fahr disease)  
Multinodular goiter  
Retinitis pigmentosa, autosomal dominant  
Leukemia/lymphoma, T-cell  
Oculopharyngeal muscular dystrophy, autosomal recessive  
APEX nuclease (multifunctional DNA repair enzyme)  
Cardiomyopathy, familial hypertrophic  
Oligodontia  
Goiter, familial  
Carbohydrate-deficient glycoprotein syndrome, type II  
Elliptocytosis  
Spherocytosis  
Anemia, neonatal hemolytic, fatal and near-fatal  
Arrhythmic right ventricular dysplasia  
Marfan syndrome, atypical  
DNA mismatch repair gene MLH3  
Diabetes mellitus, insulin-dependent  
Krabbe disease  
Hypothyroidism, congenital  
Thyroid adenoma, hyperfunctioning  
Graves disease  
Hyperthyroidism, congenital  
Usher syndrome, autosomal recessive  
Emphysema-cirrhosis  
Hemorrhagic diathesis  
X-ray repair

76 million base pairs



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



corporate  
psychology

rdles

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 personal (DNA chips)

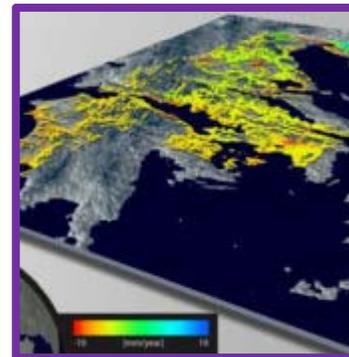


### Mensajería en la red

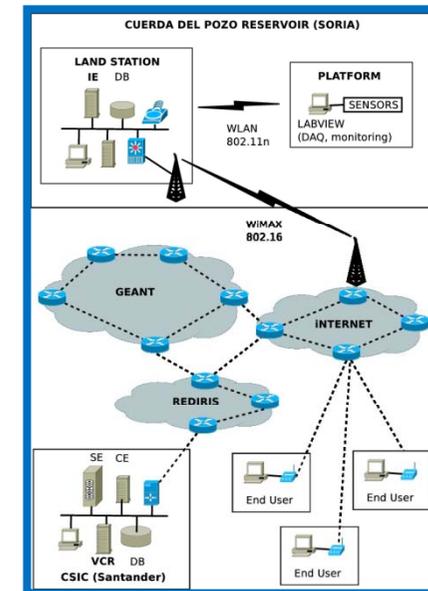
### OPEN DATA

## Ejemplos Globales:

- **ESA's GMES (Observing the Earth)**
- SmartCities



## Integración: Estándares: Sensor Web Enablement/ Web Services



# Capacidades Técnicas: Infraestructura

## Centros de procesamiento de datos

- **Supercomputadores** top500: hasta 50 Petaflops
  - España: Red Española de Supercomputación (RES)
- **GRID**: hasta 600.000 cores (EGI.eu), >200 Petabytes
  - España: IberGrid
- Componentes:
  - Almacenamiento: HADOOP, GPFS, Lustre...
  - Clusters: Redes Infiniband



## Redes de comunicación

- España: **RedIris-Nova** (fibra oscura, n x 10Gb/s)

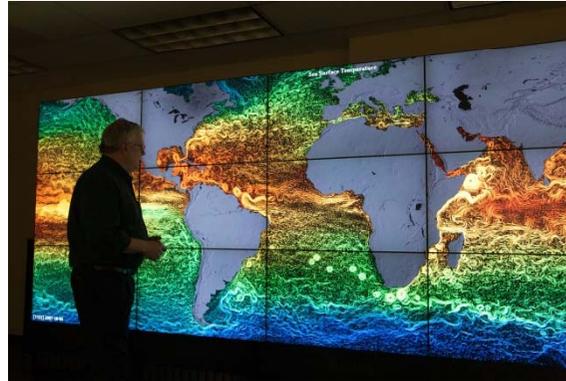
## Cloud

- Amazon , IBM, BT, Verizon, ..., Arsys, INDRA,



# Presentación

## ➤ "Hardware": Dashboards, Walls, Visores personales 3D



## ➤ Software/Gráficos:

