COMPUTAEX

JUNTA DE EXTREMADURA
Consejería de Economía e Infraestructuras
Secretaría General de
Ciencia, Tecnología e Innovación

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa

CÉNITS

SOC INFO Big Data

# EXPERIENCIAS PRÁCTICAS DE BIG DATA EN AAPP (5)

Asociación de la Prensa, Madrid, 15 Noviembre de 2016

# BIG DATA Y SUPERCOMPUTACIÓN: PAREJA *DE FACTO*

1.- ENFOQUE
2.- RETOS Y TENDENCIAS
3.- EXPERIENCIAS

joseluis.gonzalez@cenits.es

# 1.- ENFOQUE

- La fusión de la supercomputación y la analítica de datos permite conducir descubrimientos científicos y técnicos
- Big Data se refiere a datos que no es fácil capturar, gestionar y analizar con herramientas tradicionales debido a limitaciones (Vs y Ps)
- Supercomputación resuelve grandes retos en ciencia, ingeniería y analítica.

# 1.- ENFOQUE

# MÉTODOS ANALÍTICOS DE BIG DATA

Buscar

La aguja en el pajar

Descubrir

La aguja entre agujas

# 1.- ENFOQUE

## MANTRAS SOCIOLÓGICOS:

- Internet
- Cloud computing
- Smartphones
- Redes sociales
- Big Data
- Open Data
- M2M
- Smart Cities
- Internet de las cosas
- Internet del Futuro
- Industria 4.0

# 1.- ENFOQUE

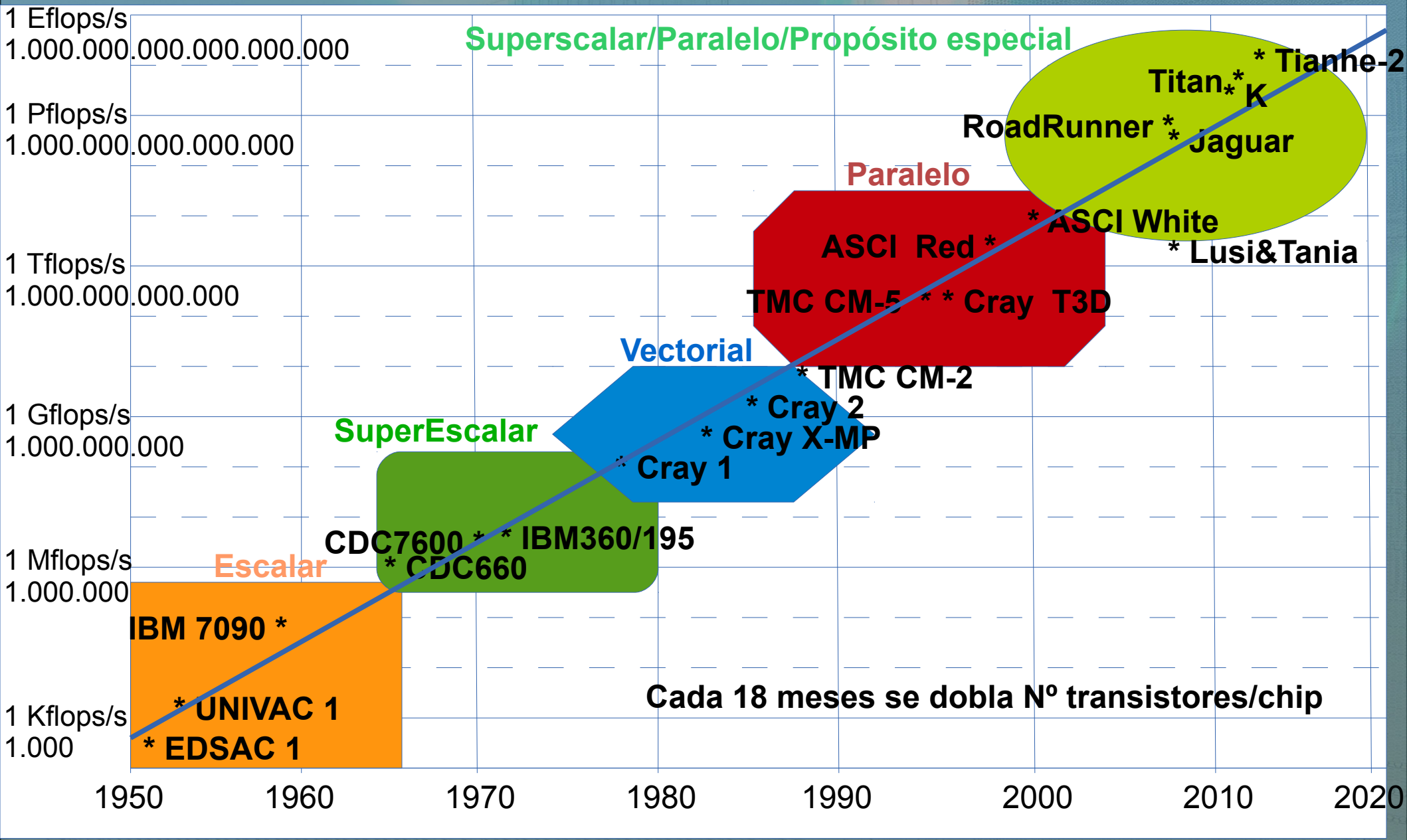
# VISIÓN DEL MUNDO COMPUTACIONAL:

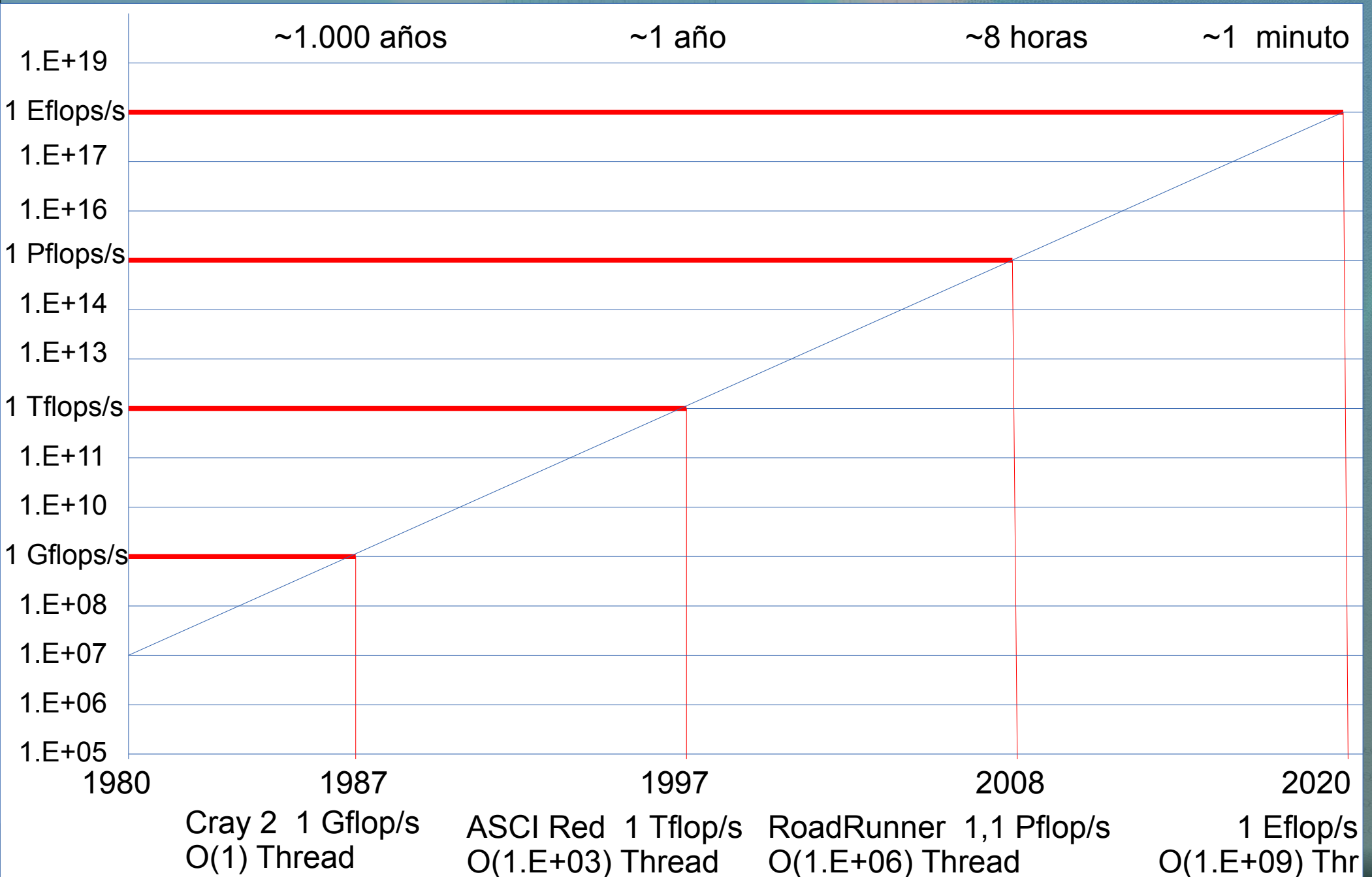
- CLUSTER COMPUTING
- SMART/GRID COMPUTING
- CLUSTER DE GPUS
- HIGH PERFORMANCE COMPUTING
- HIGH THROUGHPUT PERFORMANCE
- CLOUD COMPUTING
- HIGH PERFORMANCE CLOUD COMPUTING
- SMART COMPUTING
- BIG/OPEN DATA
- GREEN COMPUTING
- COMPUTACIÓN/COMUNICACIÓN CUÁNTICA

# 1.- ENFOQUE

# LOS CIUDADANOS DEMANDAMOS...

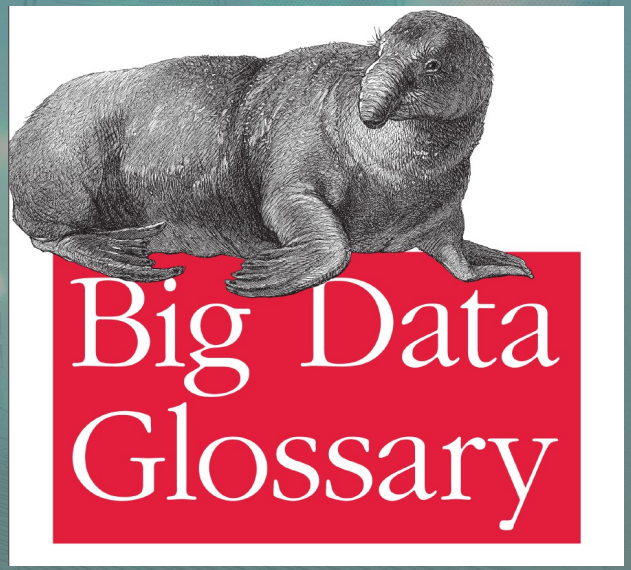
- Eficiencia
- Sostenibilidad
- Gestión óptima de recursos
- Servicios y nuevas oportunidades
- Calidad de vida


1 Eflops/s
1.000.000.000.000.000.000
1 Pflops/s
1.000.000.000.000.000
1 Tflops/s
1.000.000.000.000
1 Gflops/s
1.000.000.000
1 Mflops/s
1.000.000
1 Kflops/s
1.000
1950
1960
1970
1980
1990
2000
2010
2020
Superscalar/Paralelo/Propósito especial
Paralelo
Vectorial
SuperEscalar
Escalar
* Tianhe-2
Titan * * K
RoadRunner * * Jaguar
* ASCI White
* Lusi&Tania
ASCI Red *
TMC CM-5 * * Cray T3D
* TMC CM-2
* Cray 2
* Cray X-MP
* Cray 1
CDC7600 * * IBM360/195
* CDC660
IBM 7090 *
* UNIVAC 1
* EDSAC 1
Cada 18 meses se dobla Nº transistores/chip


~1.000 años
~1 año
~8 horas
~1 minuto
1.E+19
1 Eflops/s
1.E+17
1.E+16
1 Pflops/s
1.E+14
1.E+13
1 Tflops/s
1.E+11
1.E+10
1 Gflops/s
1.E+08
1.E+07
1.E+06
1.E+05
1980
1987
1997
2008
2020
Cray 2 1 Gflop/s
O(1) Thread
ASCI Red 1 Tflop/s
O(1.E+03) Thread
RoadRunner 1,1 Pflop/s
O(1.E+06) Thread
1 Eflop/s
O(1.E+09) Thr

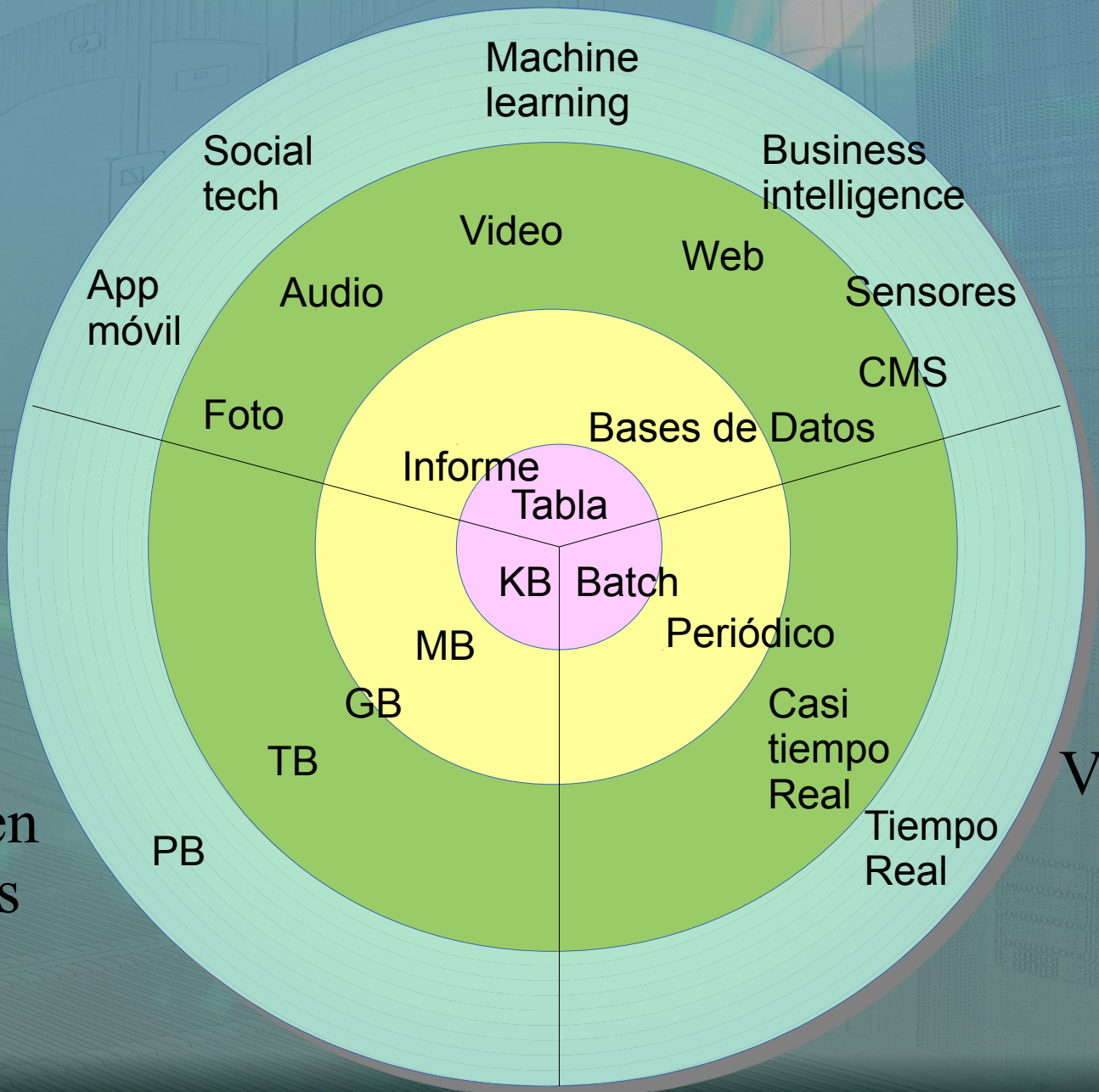
# 1.- ENFOQUE

- En el tiempo que antes se necesitaba para evaluar una hipótesis ahora se avalúan 1.000 hipotesis, lo que ahumenta considerablemente nuestra garantía y velocidad de éxitos




DATOS
• Volumen
• Velocidad
• Variedad
• Volatilidad
USOS
• Personalización
• Predicción
• Prevención
Big & Quick Data
CONOCIMIENTO
• Veracidad
• Valor
• Visualización
• Validez
• Viabilidad

# Variedad de datos




Volumen de datos

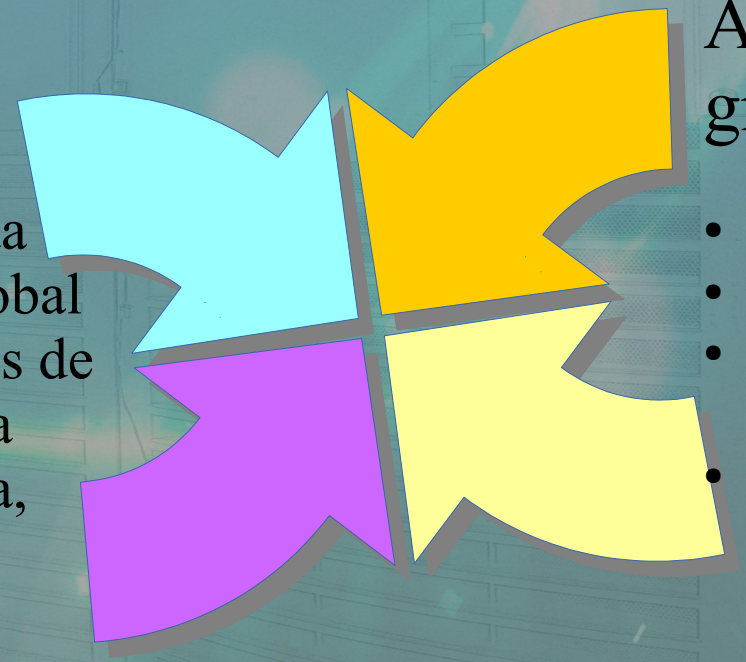
Velocidad de datos

# 1.- ENFOQUE

## ARQUITECTURAS DE SISTEMAS DIFERENTES....

### Supercomputación

- Computación escalable
- Elevado ancho de banda
- Baja latencia, mem. global
- Minimizar movimientos de datos carga en memoria
- Mueve datos para carga, *check-point* o almacén



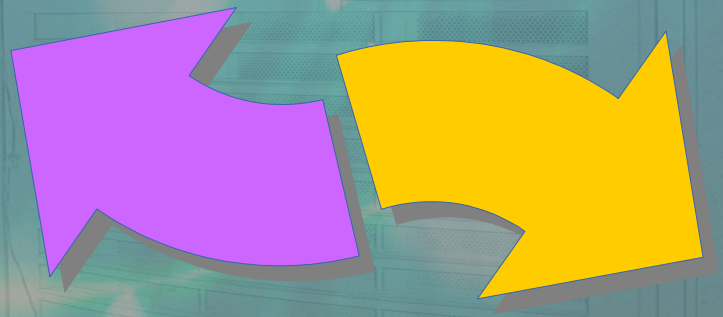
### Analítica de datos gran escala

- Computación distribuida
- Divide-y-vencerás en SOA
- Maximiza movimientos de datos: *Scan-Sort-Stream* todo a la vez
- Bajo coste procesador-memoria-interconexión y almacenamiento

## SOPORTAN APLICACIONES DIFERENTES....

# 1.- ENFOQUE
# ¿PUEDE CLOUD HACERLO?....

Supercomputación



Cloud

# 1.- ENFOQUE

# ¿PUEDE CLOUD HACERLO?....

- IaaS hace posible Big Data
- Principio de elasticidad de Cloud computing: escalabilidad de máquinas y recursos
- Esta escalabilidad: máquinas virtuales pueden instalarse en sistemas diseñados y desarrollados para procesamiento paralelo
- Tecnologia Big Data inmersa en Cloud computing
- Cassandra: base de datos estándar especialmente diseñada para para ser integrada en clusters gestionados en Cloud

# 1.- ENFOQUE
# ¿PUEDE CLOUD HACERLO?....

- Hadoop es un framework (conjunto de metodologías y herramientas asociadas a un lenguaje de programación) software diseñado para el procesamiento masivo paralelo (que corre en una plataforma masivamente paralela.
- Tal vez ha pasado el tiempo en que la computación paralela estaba sólo al alcance de una pequeña comunidad de científicos especializados, desarrolladores y expertos.
- En todo caso, las infraestructuras han de ser diseñadas, programadas y gestionadas en centros especializados.

# 1.- ENFOQUE

# PAREJA DE FACTO: FUSION O CONVERGENCIA

Captación datos (Cloud)

Simulación


Analítica

# 2.- RETOS Y TENDENCIAS

## CLUSTER COMPUTING

- Convergencia de necesidades y recursos:
  - Existencia de micros potentes y económicos.
  - Disponibilidad de redes de alta velocidad.
  - Implementado software de cómputo distribuido de alto rendimiento.
  - Necesidad de aplicaciones con requerimientos de potencia de cómputo.

- Clusteres para aplicaciones comerciales: google, wikipedia, flickr, YouTube, facebook, etc.

- Clusteres científicos: Beowulf, Now, Terascale ó Cluster X, RES, Thunder, ASCI Q, LUSITANIA, etc.

# 2.- RETOS Y TENDENCIAS

## CLUSTER COMPUTING

- Servicios aportados por un cluster:
  - Alta disponibilidad.
  - Alto rendimiento.
  - Alta eficiencia.
  - Balanceo de carga.
  - Escalabilidad
- Componentes del cluster:
  - Nodos.
  - Electrónica de red.
  - Protocolos de comunicaciones.
  - SOs y middleware.
  - Almacenamiento y periferia.
  - Servicios y aplicaciones
  - Entorno programación paralela.

# 2.- RETOS Y TENDENCIAS
# LUSITANIA

# 2.- RETOS Y TENDENCIAS
# LUSITANIA

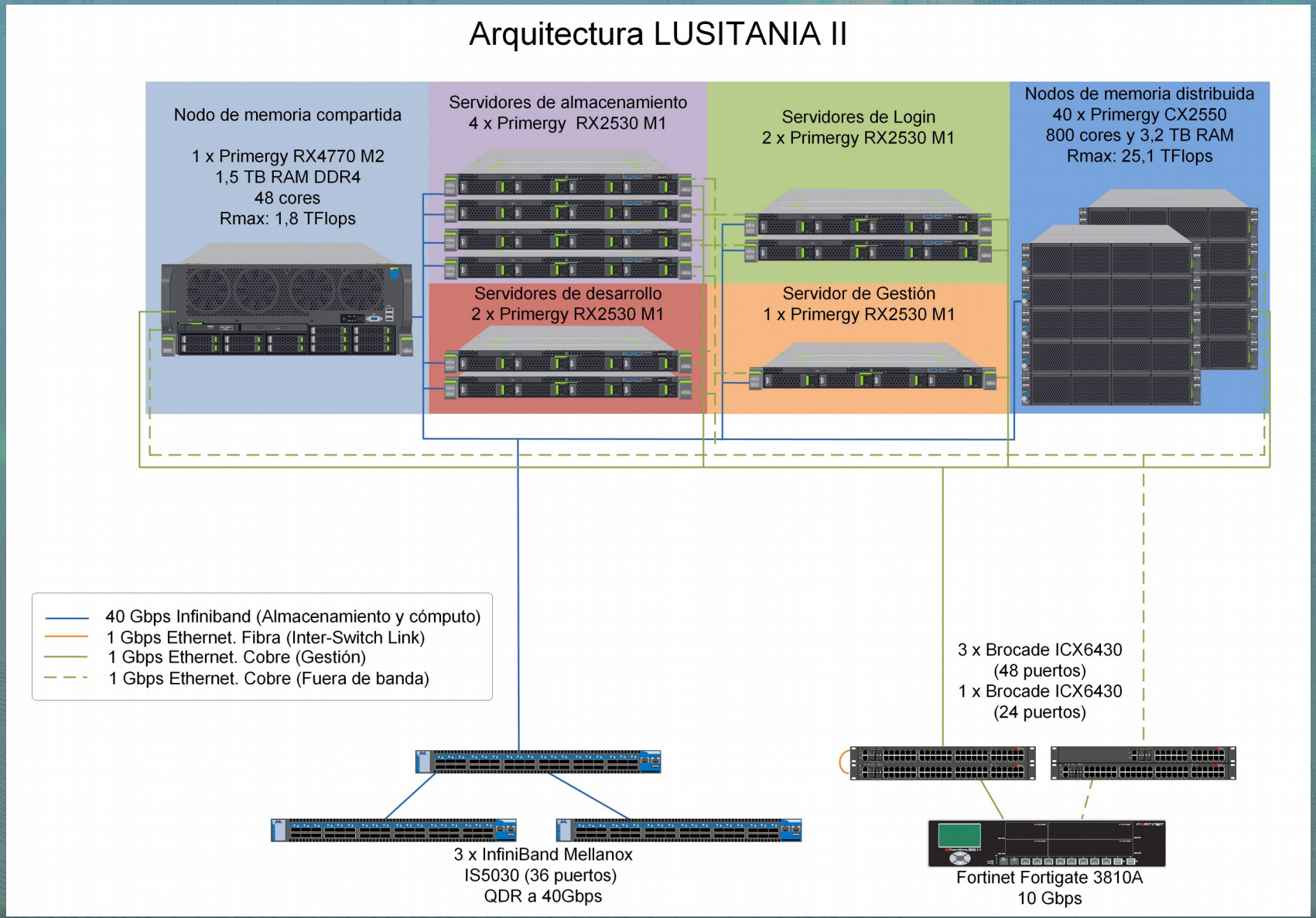
## 2 HP integrity Superdomes sx2000

- 2 x (64 procesadores/128 cores):
  - Total 128 procesadores/256 cores
  - 2 x 0,8192 = 1,63 Teraflops pico.
- Itanium®2 Dual Core Montvale @ 1.6 GHz, 18 MB cache
- 768 GB de memoria principal
- 2x 1TB memoria en una imagen:
  - Total 2 TB memoria.
- 2x 40 x 146 GB SAS Disks = 11,68 TB de scratch
- SuSe Linux SLES 10
- Particiones:
  - Hasta 16 particiones físicas
  - Hasta 64 particiones virtuales PRM, WLM, IVM en HP-UX,gWLM multiSO

# LUSITANIA II

# LUSITANIA II

## Arquitectura LUSITANIA II



# 2.- RETOS Y TENDENCIAS

DCs juegan un papel muy importante pues grandes CPDs aportan las siguientes ventajas:

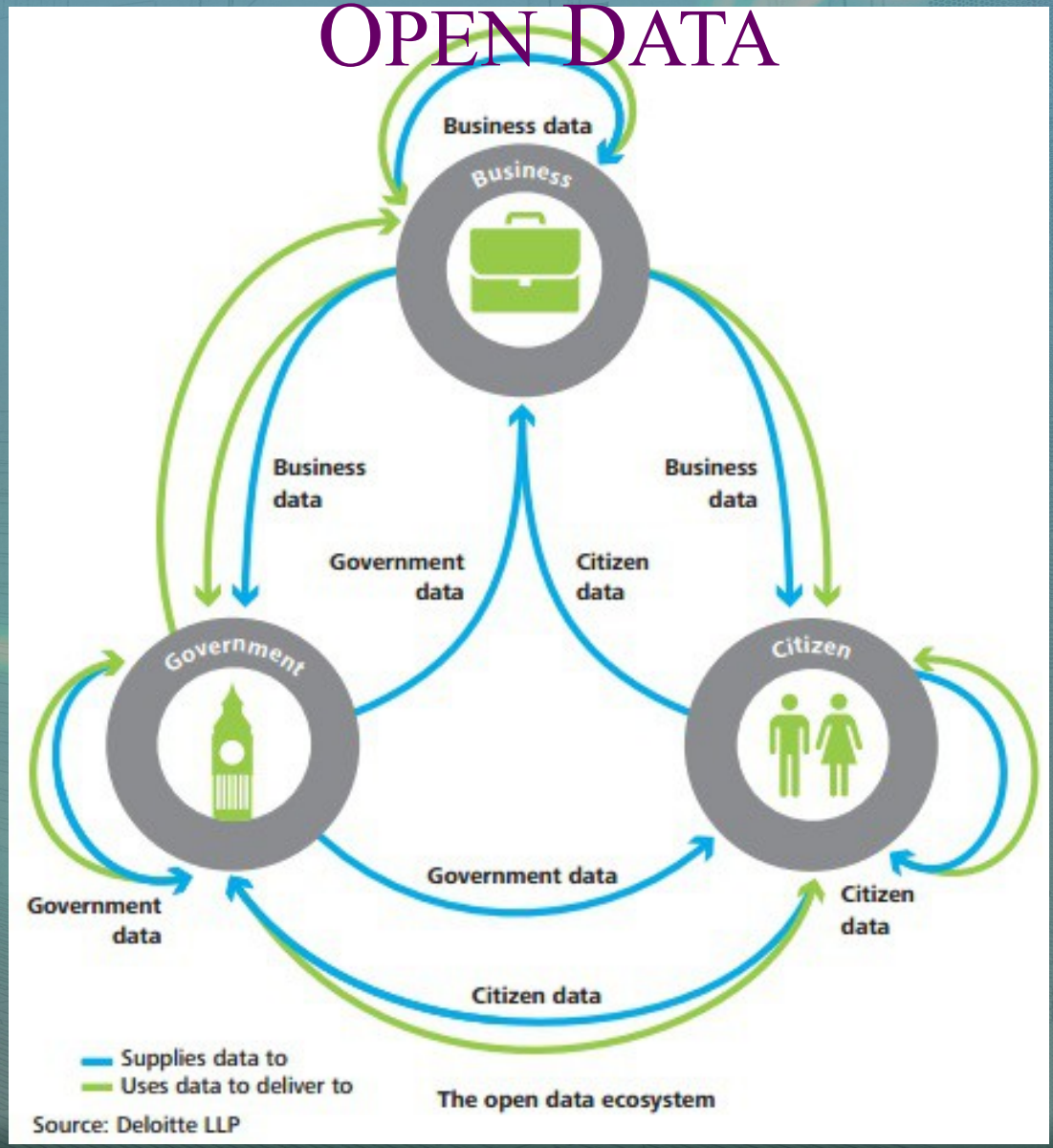
- Alta disponibilidad.
- Gran capacidad de:
  - almacenamiento,
  - procesamiento y
  - acceso a la información.
- Seguridad y fiabilidad de la información.
- Eficiencia energética.

# 2.- RETOS Y TENDENCIAS
# BIG DATA

- *Big Data* significa capacidad de manejar y gestionar grandes volúmenes de información a gran velocidad.
- Conjunto de procesos, tecnologías y modelos de negocio basados en la captación, análisis y explotación de cantidades masivas de datos
- Diariamente se generan 2,5 trillones de bytes.
- 90% de los datos actuales generados los dos últimos años
- Big data importante en la toma de decisiones y mejora la competitividad de las empresas

# 2.- RETOS Y TENDENCIAS

## OPEN DATA



# 2.- RETOS Y TENDENCIAS

## IoT

- *Conexión a Internet de cualquier cosa en cualquier momento y lugar*

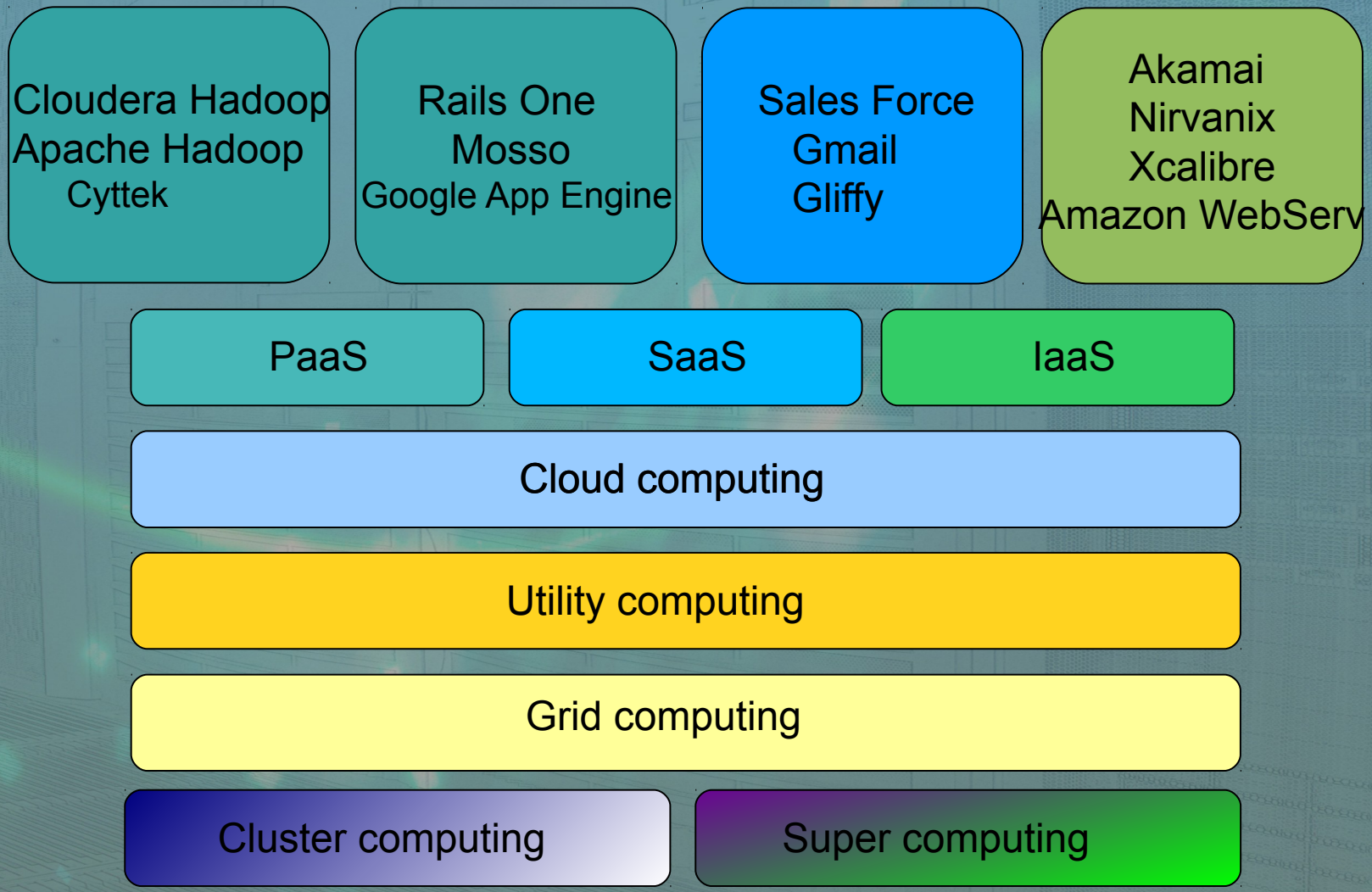
- Cualquier objeto puede ser “inteligente”:

  - Sensores
  - Redes fijas e inalámbricas
  - Dispositivos

- Actualmente 60 millones de dispositivos conectados en UE

- En 2020 previsión de ¿200 millones de objetos? conectados

# 2.- RETOS Y TENDENCIAS



# 3. EXPERIENCIAS

- La extracción de información (Big Data+HPC),
- de diferentes fuentes o sensores (Internet de las Cosas),
- dotada de una capa de inteligencia (Smart),
- y sustentada en una infraestructura ubícua (Cloud Computing) y
- con acceso abierto para todos los ciudadanos (Open Data),
- puede desarrollar una ciudad o región inteligente (Smart city/region) trabajando en red para
- afrontar numerosos proyectos innovadores.

# 3. EXPERIENCIAS
# TIPOS DE PROYECTOS BIG DATA

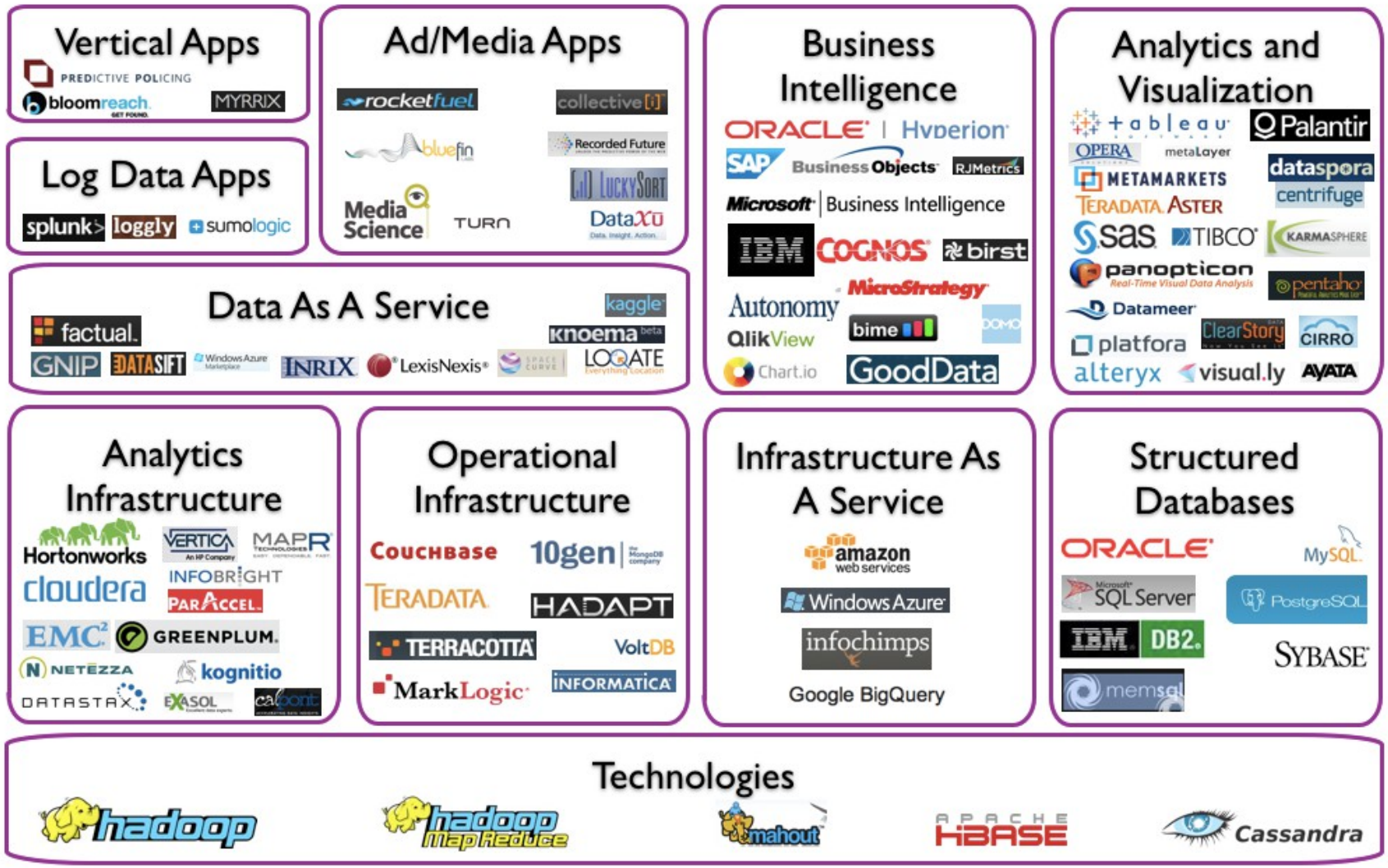
- Procesamiento de datos en tiempo real
- Procesamiento datos “almacenados”
- Diferentes: enfoques, arquitecturas técnicas, herramientas y datos

# 3. EXPERIENCIAS

## COMPONENTES DE PROYECTOS BIG DATA:

- Tecnologías BD: hardware y software
- Metodología específica
- Aspectos legales relacionados con la manipulacion de los datos y usos previstos
- Componente social:circulacion y uso de datos personales

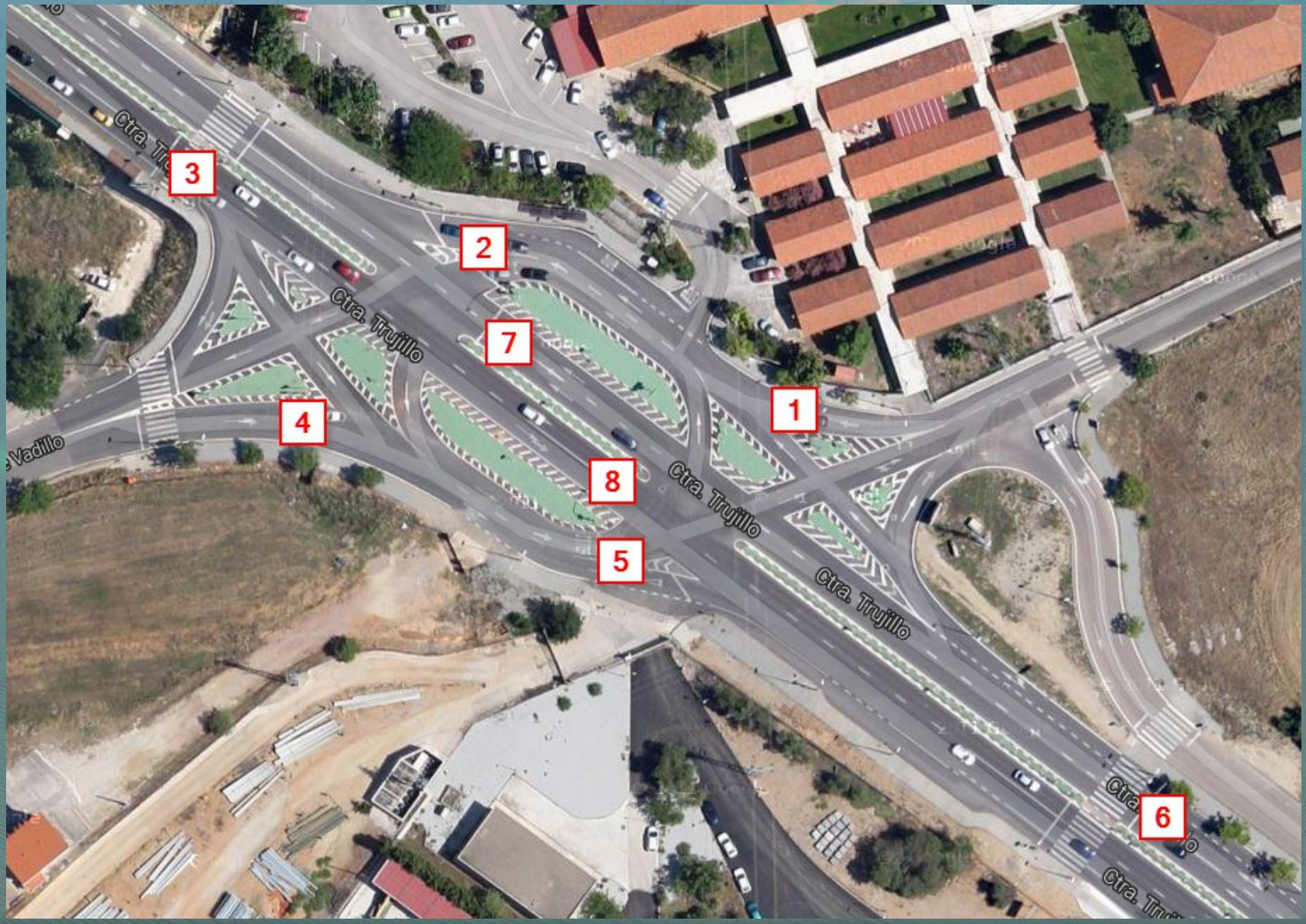
# Big Data Landscape



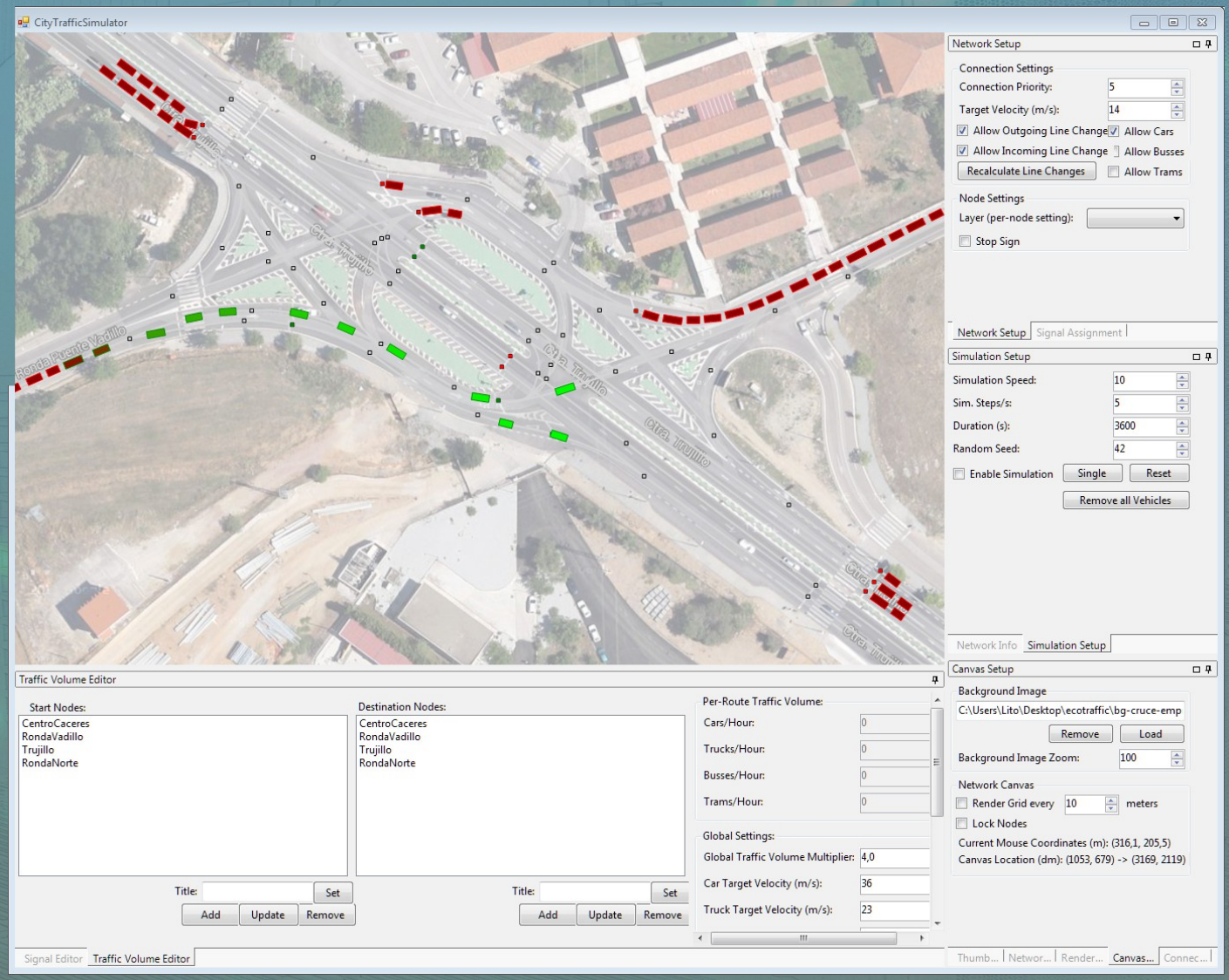


dave@vcdave.com
blogs.forbes.com/davefeinleib

# 3. EXPERIENCIAS: ECO-TRAFIC



# 3. EXPERIENCIAS: ECO-TRAFIC

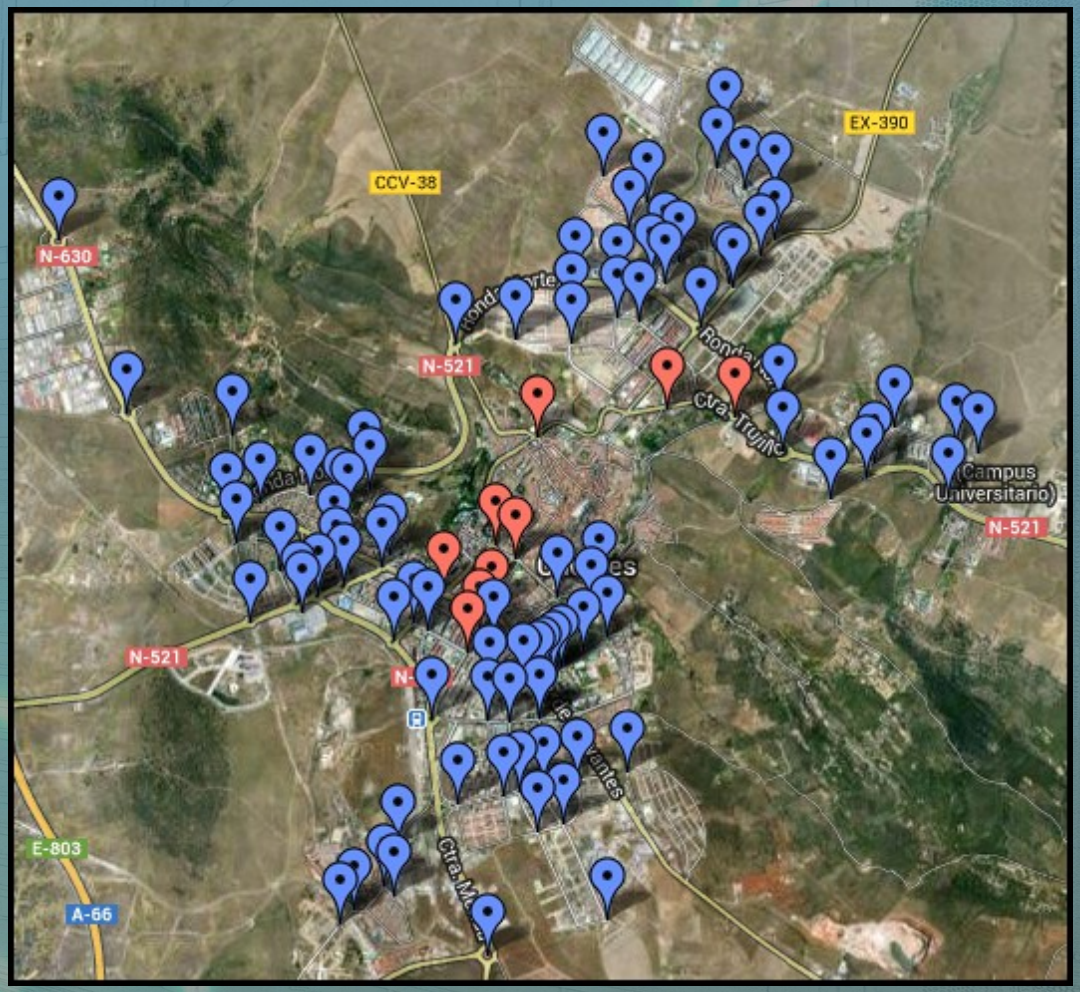


# 3. EXPERIENCIAS: ECO-TRAFIC

- Ahorros y eficiencias:
  - 26.000 vehículos diarios
  - 108 litros de combustible diarios.
  - 270 Kg. de CO2 diarios.
  - 79.083 horas de ahorro anual.
  - 3 horas de ahorro anual por vehículo.

- Predictibilidad de flujos de tráfico con Big Data

# 3. EXPERIENCIAS: ECO-TRAFIC




*http://www.cenits.es/noticias/31012014-computaex-presenta-resultados-eco-traffic-modelado-trafico-smart-eco-region*

# 3. EXPERIENCIAS: CONSUMAR

- Objetivos:
  - Lograr eficiencia energética, tanto en el ámbito doméstico como empresarial.
  - Aplicación de técnicas de Big Data al análisis de los datos de consumo energético.
  - Búsqueda de la sostenibilidad económica y ecológica a través de las nuevas tecnologías.
  - Un producto software que actúe como prospector de las diversas tarifas eléctricas en cada momento.

# 3. EXPERIENCIAS: CONSUMAR

- Diseñado un conjunto de herramientas que realice las siguientes labores:
  - Recogida de datos fijos.
  - Visualización de las diferentes tarifas eléctricas.
  - Plataforma de satisfacción del servicio eléctrico.
  - Herramientas de análisis de datos.

# 3. EXPERIENCIAS: CONSUMAR


Arquitectura proyecto ConSumar

# 3. EXPERIENCIAS: CONSUMAR

- Base de datos para la gestión del sistema de información, basada en documentos y esquema libre:
  - Documentos bastante numerosos.
  - De diferente formato.
  - Aúna consultas rápidas (C++) con almacenamiento heterogéneo (BSON).
  - Base de datos única para datos climáticos, costes energéticos y plataforma de satisfacción de usuarios.


Plataforma de almacenamiento de ConSumar

# 3. EXPERIENCIAS: CONSUMAR

- Sistema de recogida diaria de datos:
  - Aplicaciones desarrolladas en Python.
  - Los datos recogidos diariamente alimentan mongoDB.
  - Desarrollo de *parsers* para transformación de XML, CVS y Excel a JSON.
  - Captación diaria:
    - Costes energéticos a las 23h.
    - Predicción climática para 3 días a las 5:55h.
    - Clima diario a las 23:55h.


Captación de datos ConSumar

# 3. EXPERIENCIAS: CONSUMAR

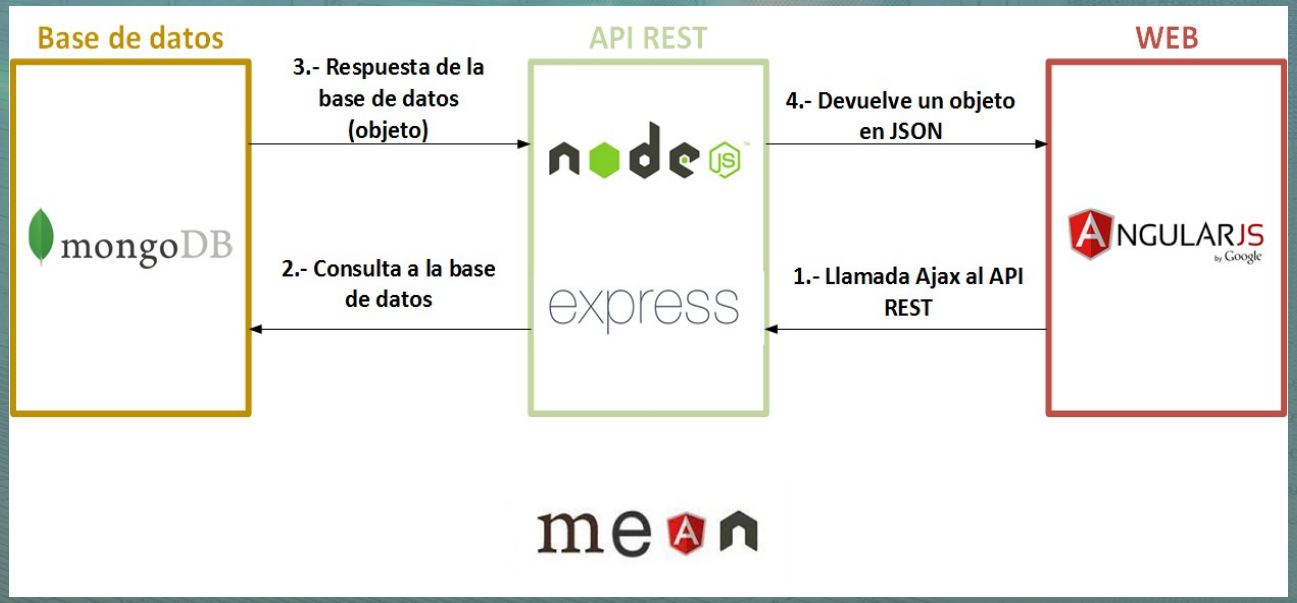
Herramienta de visualización de tarifas y precios diarios:

- Basada en la pila de aplicaciones MEAN (MongoDB, Express, AngularJS y Node.js) en JavaScript
- Permite crear aplicaciones distribuidas usando JavaScript en el cliente, servidor y en la capa de almacenamiento.

# 3. EXPERIENCIAS: CONSUMAR

Herramienta de visualización de tarifas y precios diarios:

- AngularJS: Framework de JavaScript libre para desarrollo de aplicaciones web en el cliente.
- Node.js: intérprete JavaScript en el servidor.
- Express: framework en el servidor para desarrollo de aplicaciones web con Node.js



# 3. EXPERIENCIAS: CONSUMAR

## Tarifa elegida para el día 26-02-2016

**Tarifa Diurna - 2.0A**

| Hora | Precio kwh |
|---|---|
| 0 - 1 | 0.089 € |
| 1 - 2 | 0.085 € |
| 2 - 3 | 0.084 € |
| 3 - 4 | 0.083 € |
| 4 - 5 | 0.083 € |
| 5 - 6 | 0.086 € |
| 6 - 7 | 0.091 € |
| 7 - 8 | 0.098 € |
| 8 - 9 | 0.101 € |
| 9 - 10 | 0.095 € |
| 10 - 11 | 0.093 € |
| 11 - 12 | 0.087 € |
| 12 - 13 | 0.086 € |
| 13 - 14 | 0.085 € |
| 14 - 15 | 0.084 € |
| 15 - 16 | 0.082 € |
| 16 - 17 | 0.081 € |
| 17 - 18 | 0.082 € |
| 18 - 19 | 0.082 € |
| 19 - 20 | 0.096 € |
| 20 - 21 | 0.097 € |

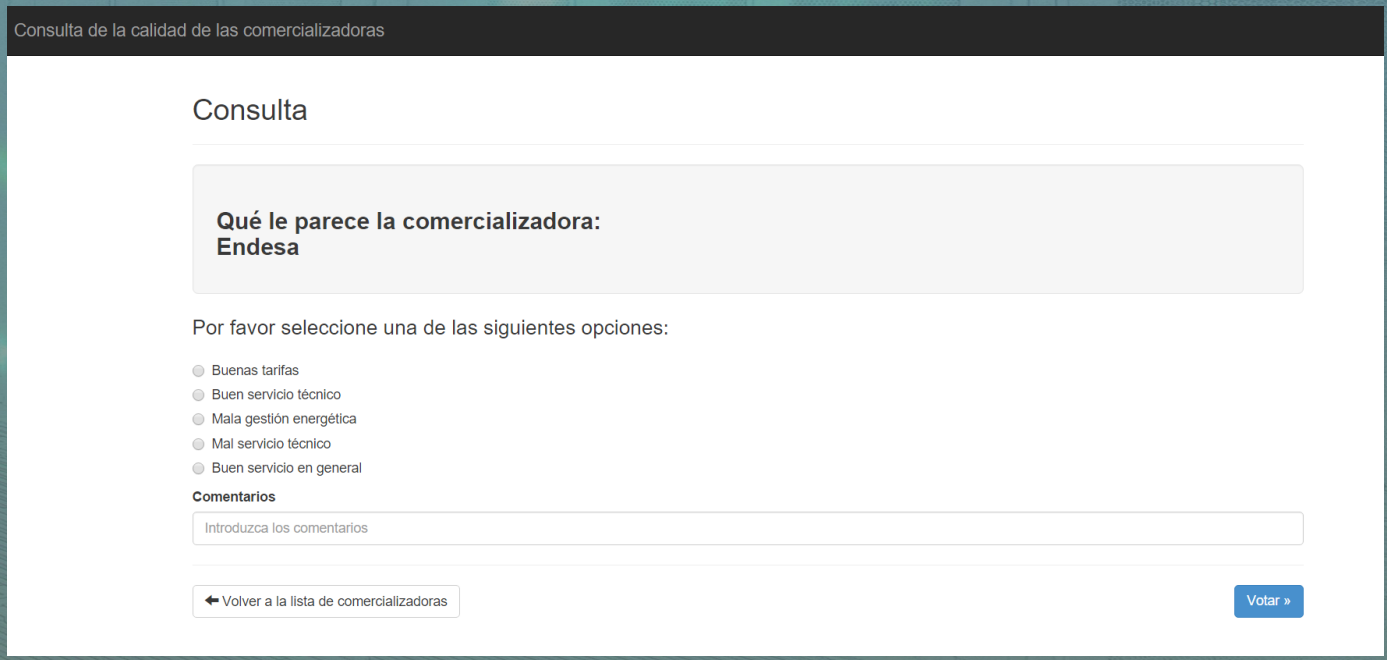
Tarifa General | Tarifa Nocturna | Tarifa Vehículo Eléctrico

Consulta de tarifas de producción de energía

# 3. EXPERIENCIAS: CONSUMAR

Plataforma de satisfacción del servicio eléctrico:

- Vierte datos a la BD creada con mongoDB.
- Aplicación web también desarrollada en MEAN.
- Almacena en sistema Big Data: compañías eléctricas, encuesta, voto, votantes y opiniones usuarios. Fundamental para analítica de datos.

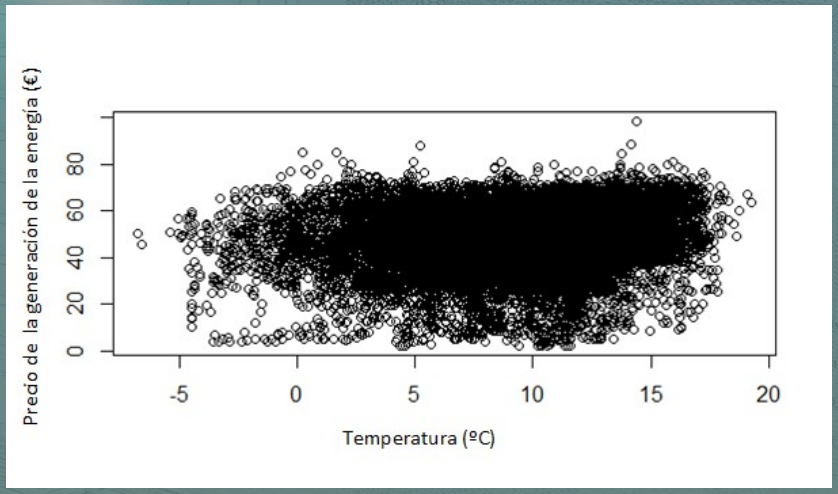



Plataforma satisfacción de ConSumar

# 3. EXPERIENCIAS: CONSUMAR

Analítica de datos:

- Herramienta de analítica para estudio de costes.
- Varias aplicaciones desarrolladas en R bajo óptica Machine Learning.
- Obtención de conjunto de valores, tratados para obtener patrones, realizar predicciones y tomar decisiones sobre la forma de ahorros en el consumo energético
- Nube de puntos usada para la analítica de datos y toma de decisiones.

# 3. EXPERIENCIAS: CONSUMAR

Análisis y predicción de datos aplicando regresión lineal

- Análisis de la dispersión de los datos y estudio de agrupación de valores.
- Dispersión muy importante: variación del coste de producción de energía depende de numerosos factores (renovables, mes, hora, etc) para calcular el precio.
- Modelo de regresión lineal múltiple: una variable dependiente a la que le influyen un conjunto de variables explicativas

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + u$$

# 3. EXPERIENCIAS: CONSUMAR

- Modelado Coste de Producción *CPh* de la energía:

$$CPh = (Pmh + SAh + Och)$$

- *Pmh*: Precio medio horario obtenido del mercado diario en hora h:

$$Pmh = PMDh * EMDh + \sum(PMIh_n * EMIh_n)/EMDh + \sum EMIh_n$$

- *Coste de Servicios de Ajuste* del sistema asociados al suministro :

$$SAh = PMASh + CDSVh$$

- *Otros Costes* asociados al suministro (financiaciones, etc.):

$$OCh = CCOMh + CCOSh + CAPh + INTh$$

# 3. EXPERIENCIAS: CONSUMAR

Análisis y predicción de datos aplicando regresión lineal

- Análisis de precios diarios de generación de energía muestran diferencias dependiendo de la hora y del mes (factores climáticos y transporte energía por la red eléctrica)
- Predicción: información de lluvia (mm/h); temperatura (º C); velocidad viento (Km/h). Modelo de regresión lineal múltiple:

```
CPhconLluvia+Temperatura+VelocidadViento
```

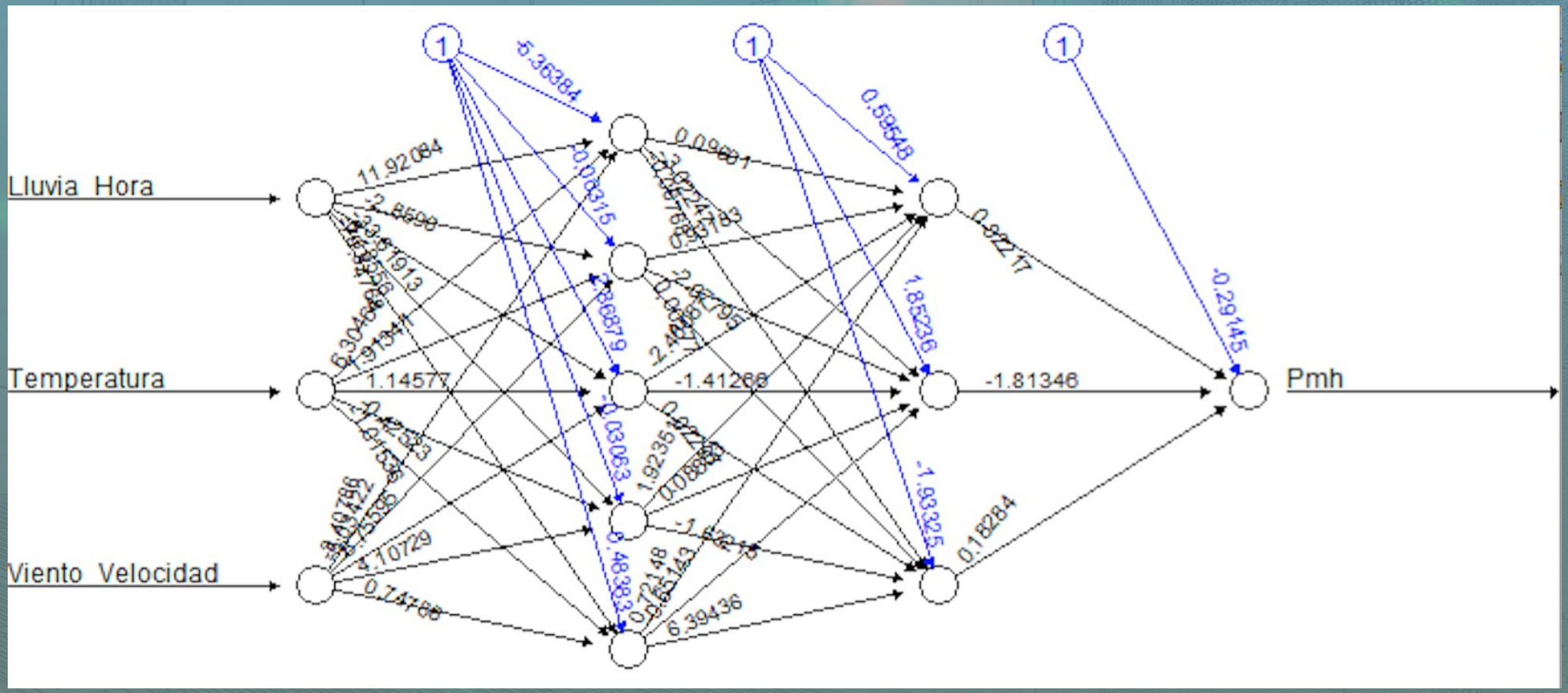
# 3. EXPERIENCIAS: CONSUMAR

- Consulta de predicciones de producción energética:
- Herramienta MEAN para mostrar predicciones de precios de energía usando un modelo de datos que use todas las variables que intervienen


| Hora | Precio kwh |
|---|---|
| 0 - 1 | 0.089 € |
| 1 - 2 | 0.085 € |
| 2 - 3 | 0.084 € |
| 3 - 4 | 0.083 € |
| 4 - 5 | 0.083 € |
| 5 - 6 | 0.086 € |
| 6 - 7 | 0.091 € |
| 7 - 8 | 0.098 € |
| 8 - 9 | 0.101 € |
| 9 - 10 | 0.095 € |
| 10 - 11 | 0.093 € |
| 11 - 12 | 0.087 € |
| 12 - 13 | 0.086 € |
| 13 - 14 | 0.085 € |
| 14 - 15 | 0.084 € |

# 3. EXPERIENCIAS: CONSUMAR

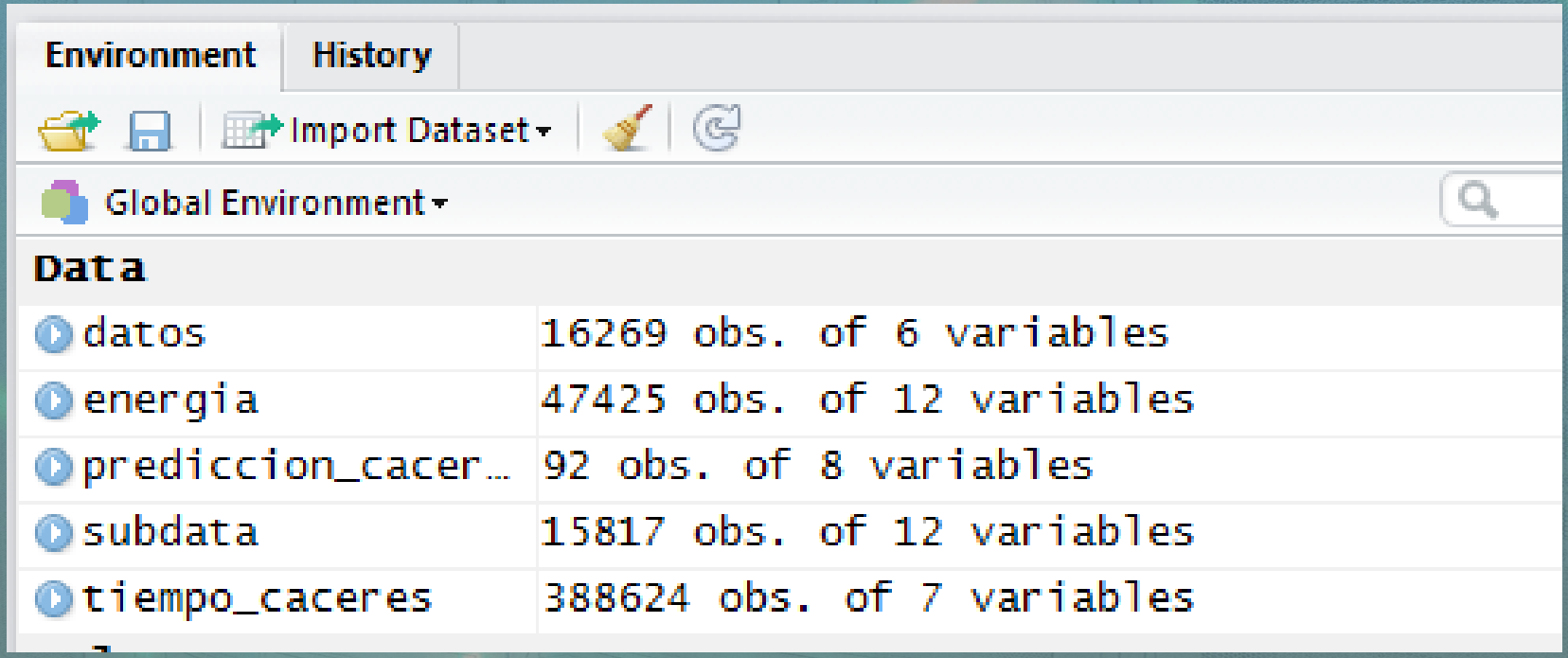
- Red de capa simple para análisis y de capa múltiple para predicción de datos
- Cálculo de errores cuadráticos medios de la predicción




Red neuronal para predicción de precio medio horario

# 3. EXPERIENCIAS: CONSUMAR

*http://www.cenits.es/noticias/11042016-computaex-presenta-resultados-proyecto-consumar-sobre-aplicacion-tecnicas-big-data*




- Fujitsu Server PRIMERGY RX350 S8: 2 procesadores Intel Xeon E5 2620v2 (2,10GHz/6 cores/15MB); 256 GB de memoria RAM y dos discos duros SAS de 300GB.
- 150 horas de cómputo en procesamiento secuencial

# 3. EXPERIENCIAS: HERITAGEN

- Ultrasecuenciación genética y supercomputación para la unificación del patrimonio genético.
- Aplicación al estudio de enfermedades hereditarias.
- 1 de cada 200 nacimientos puede estar afectado por las 6000 enfermedades monogénicas conocidas.
- Procesar datos open y secuencias genéticas garantizando:
  - Almacenamiento
  - Seguridad
  - Disponibilidad
  - Velocidad

*http://www.cenits.es/noticias/20052015-computaex-presenta-continuacion-su-proyecto-ultrasecuenciacion-genetica-estirpex-2*

COMPUTAEX

JUNTA DE EXTREMADURA
Consejería de Economía e Infraestructuras
Secretaría General de
Ciencia, Tecnología e Innovación

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa

CÉNITS

SOC INFO Big Data

# MUY AGRADECIDO POR SU ATENCIÓN

Asociación de la Prensa, Madrid, 15 Noviembre de 2016
joseluis.gonzalez@cenits.es