UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

Soft Management of Internet and Learning

Algunos retos para la Inteligencia Artificial en el siglo XXI

Dpto. de Tecnologías y Sistemas de Información
Universidad de Castilla-La Mancha
http://smile.esi.uclm.es

José A. Olivas
Joseangel.olivas@uclm.es

La Plata, CACIC 2017

Idea principal / objetivo

- Presentar un panorama de diversos **retos** vinculados con la computación actual, con una **posición crítica**, asociados a elementos tecnológicos **cotidianos**.
- Centrarse en aquellos retos que tienen que ver con la **gestión** y el **aprovechamiento inteligente** de la ingente cantidad de datos que se generan a partir de diversas fuentes y en diversos ámbitos de la **sociedad digital actual**.
- Mostrar una visión panorámica de las **tecnologías** y **técnicas** emergentes para el **manejo masivo** "**inteligente**" de datos e información.
- Describir algunos **ejemplos** de interés en el tema.

¡¡ Aproximación **Crítica** !!

Contenido

El nuevo reto de la Inteligencia Artificial en:

- **La gestión y extracción de conocimiento de grandes volúmenes de datos (Big Data y KDD).**
- **El acceso y la búsqueda de información en las grandes bases de datos digitales.**
- **Internet y las redes sociales.**

BIG DATA: Origen: los datos

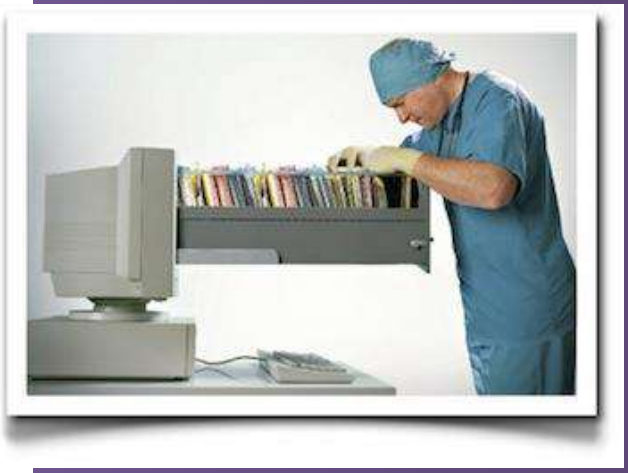
Contenido

- Mayer-Schönberger, V.; Cukier, K.: Big data. La revolución de los datos masivos. Turner 2013.
- Piatetsky-Shapiro, G.; Frawley, W.: Knowledge Discovery in Databases. AAAI/MIT Press, Cambridge MA, 1991.
- Siegel E.: Analítica predictiva. Predecir el futuro utilizando Big Data. Anaya Multimedia-Anaya Interactiva, 2013.
- D. Agrawal, S. Das and A. E. Abbadi, “Big Data and Cloud Computing: Current State and Future Opportunities” ETDB 2011, Uppsala, Sweden.
- D. Agrawal, S. Das and A. E. Abbadi, “Big Data and Cloud Computing: New Wine or Just New Bottles?” VLDB 2010, Vol. 3, No. 2.

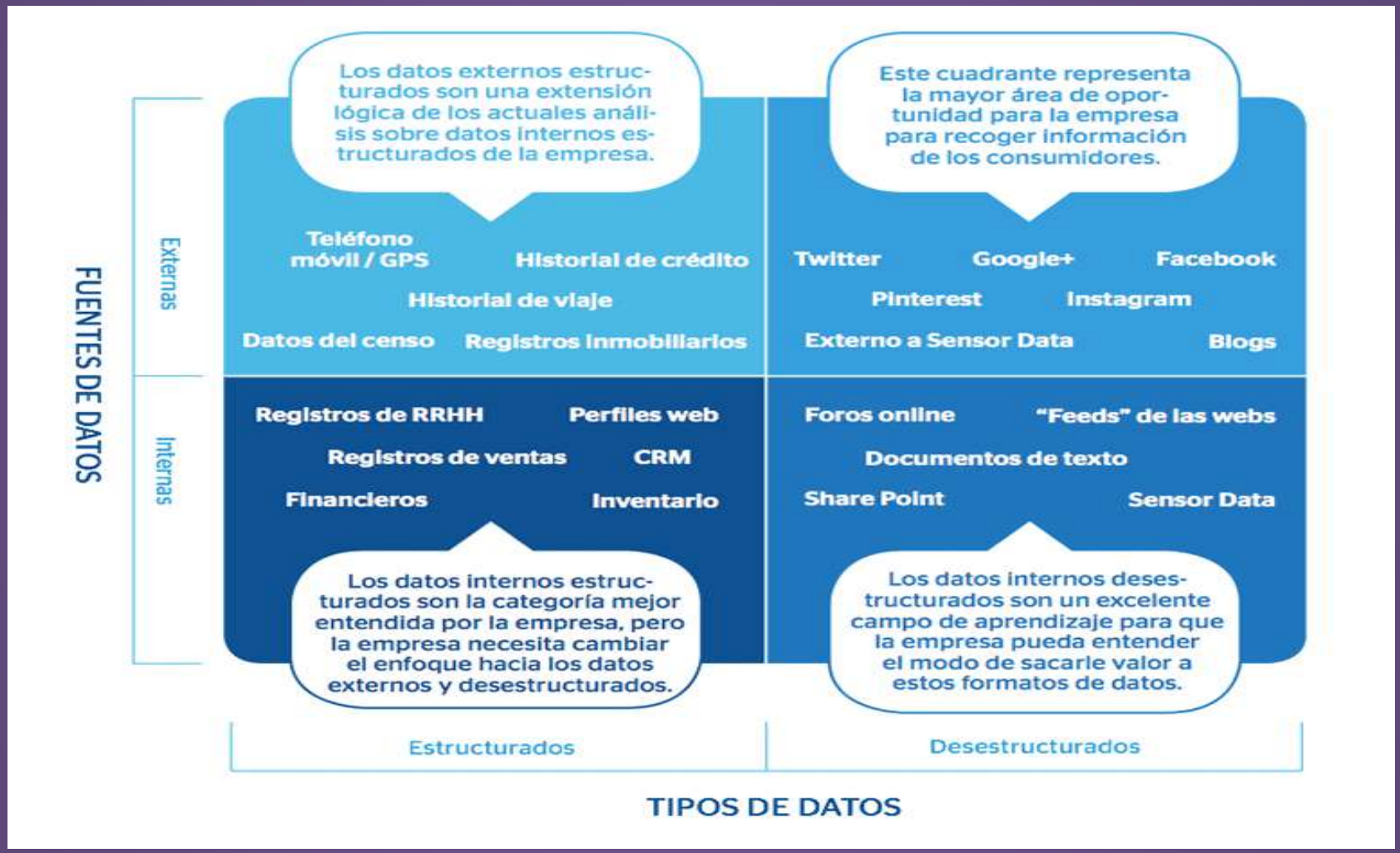
Big Data

- **Aproximación ingenua y crítica.**
- **Definición abierta de big data.**

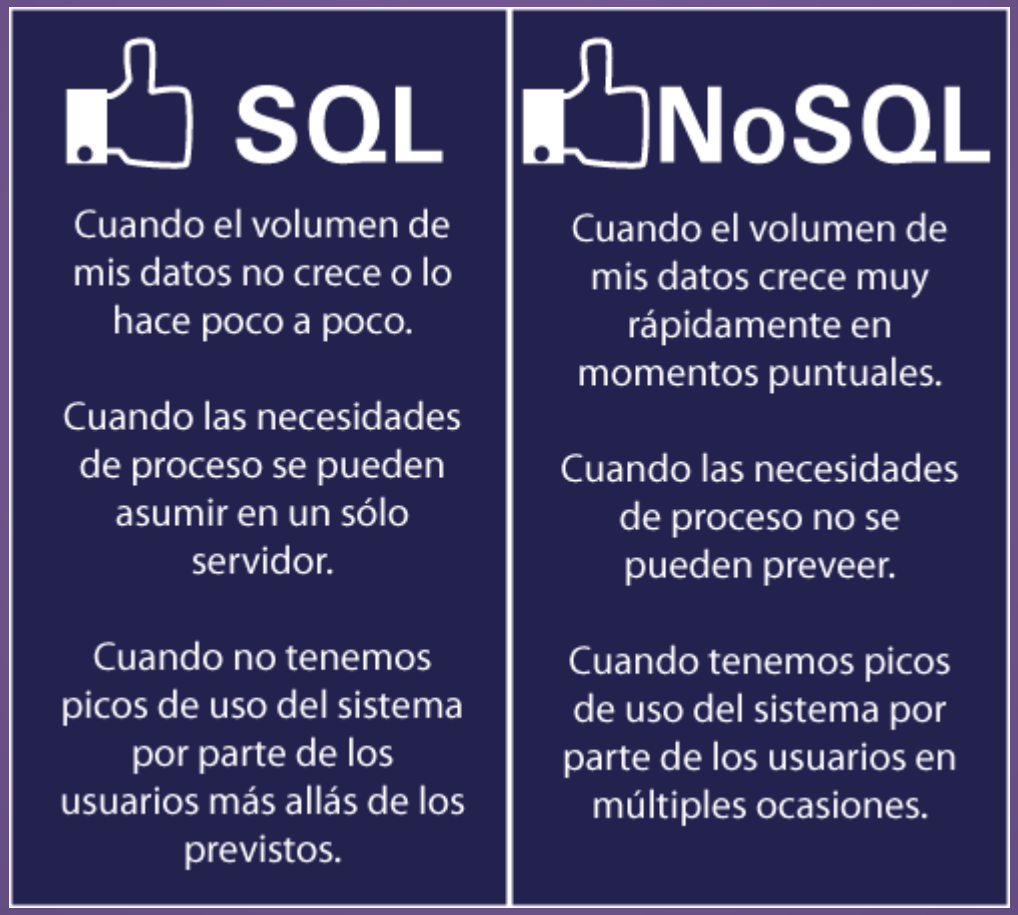
¿Qué son los datos?: DATOS / INFORMACIÓN /CONOCIMIENTO



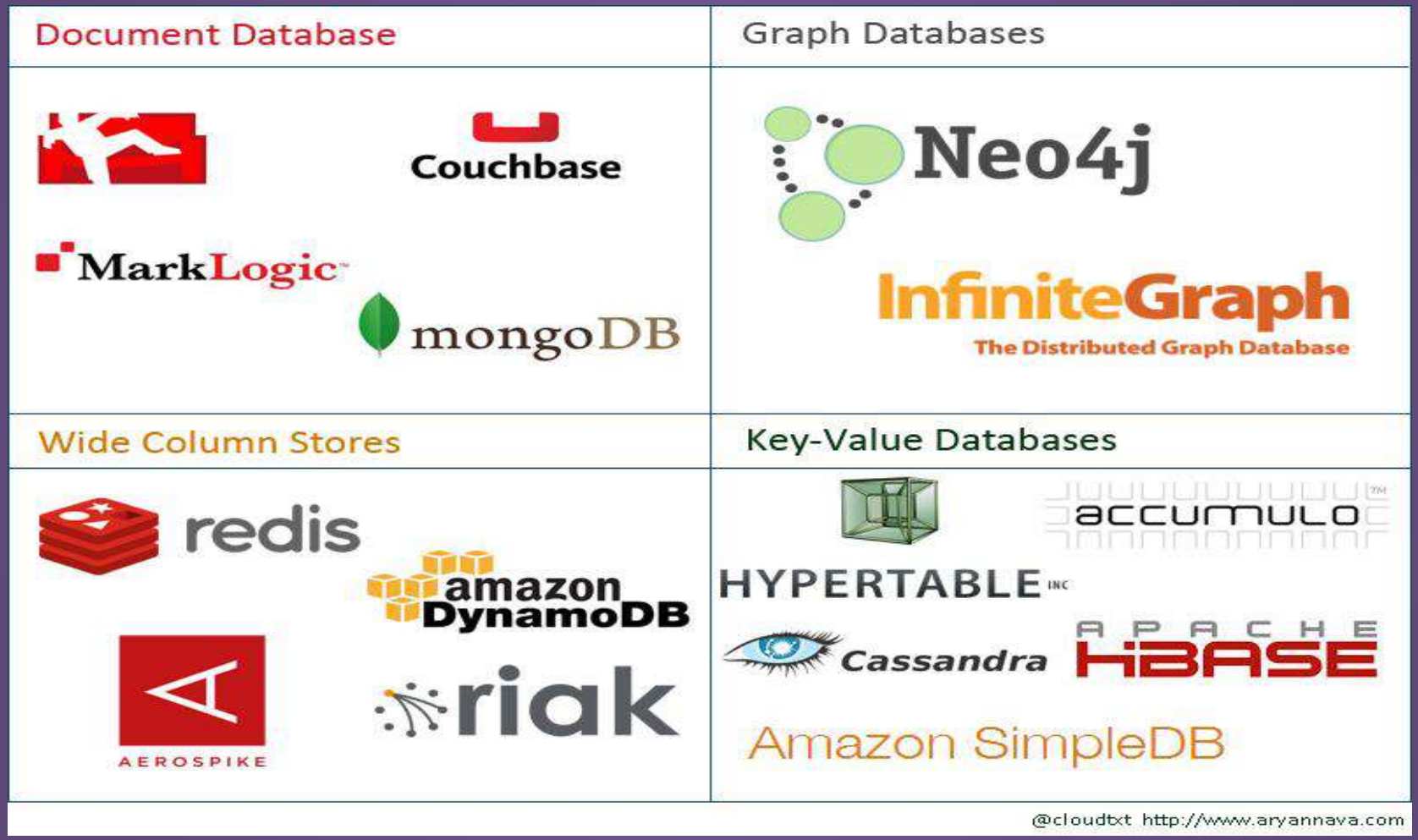
Tipos de datos (I)



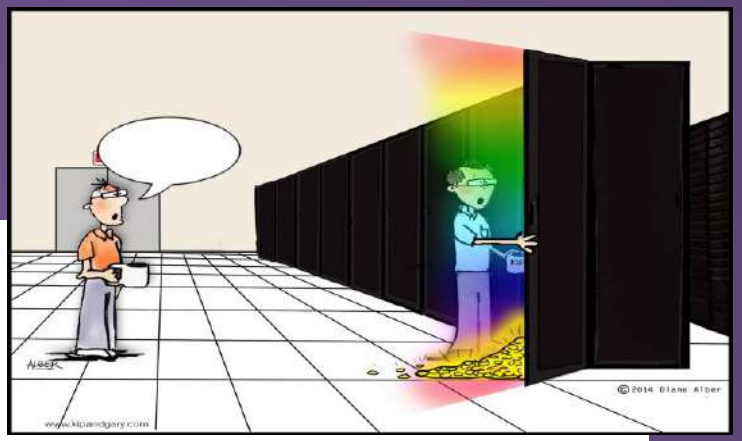
Tipos de datos (II)



Tipos de datos (III)



Origen: los datos



¿ Dónde residen los datos ?

What is a data lake?

A repository for large quantities and varieties of data, both structured and unstructured.

Data generalists/ programmers can tap the stream data for real-time analytics.

The lake can serve as a staging area for the data warehouse, the location of more carefully "treated" data for reporting and analysis in batch mode.

The data lake accepts input from various sources and can preserve both the original data fidelity and the lineage of data transformations. Data

¿ Cómo se consiguen ?

- **Sistemas Transaccionales** (operadores que recogen las peticiones a través de Call-Centers)
- **Transacciones** que se generan en las **Webs** (ficheros weblogs)
- Los **sensores** permiten capturar las magnitudes físicas o químicas y convertirlas en datos, por ejemplo temperatura, luz, distancia, aceleración, inclinación, desplazamiento, presión, fuerza, humedad, sonido, movimiento o el pH.
- **Redes sociales**
- Etc, Etc…

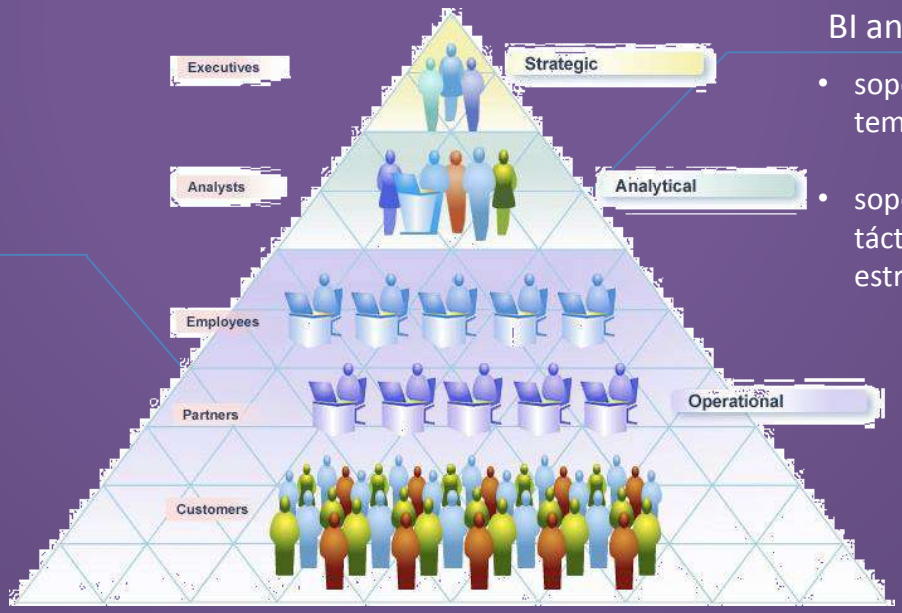
El *Business Intelligence* (BI)

definición de business intelligence (BI)

La capacidad de transformar datos en información para ayudar a gestionar una empresa es el dominio de la **inteligencia empresarial de negocios (BI)**, que consiste en los procesos, aplicaciones y prácticas que apoyen la toma de decisiones ejecutivas

BI operacional

- soporta funciones al nivel operacional
- capacidad en tiempo real o cerca de real-time
- comprende y cubre los procesos.



BI analítico y estratégico

- soporta a los executivos y en temas estratégicos
- soporta a los gestores en niveles tácticos que contribuyen a la estrategia

Crítica…

Demasiado restringido:

- …transformar datos en información…
- …apoyen la toma de decisiones…

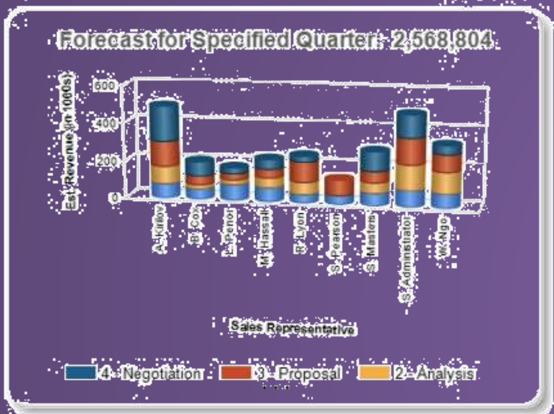
¡¡ Hay muchas otras cosas que se pueden hacer !!

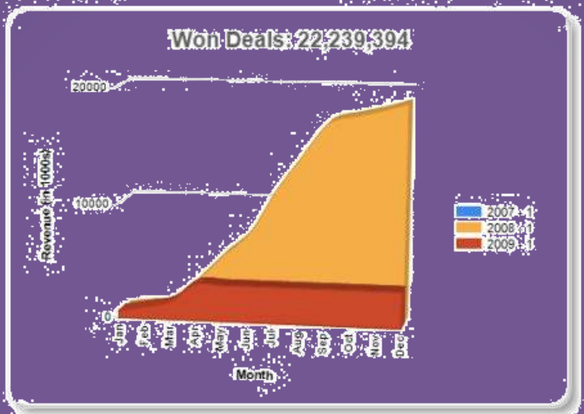
- Veamos las posibles salidas…

Outputs…

Sales Manager Dashboard

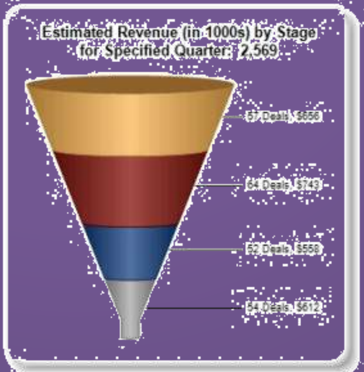
All data displayed in base currency.





Top 10 Key Deals			
Account	Est. Close Date	Est. Revenue (in 1000s)	Recent Activity
Wide World Importers	9/2/2008	$725.00	0
Trey Research	8/4/2010	$435.70	0
Fabrikam Inc.	2/21/2010	$434.30	0
Tailspin Toys	8/25/2010	$431.50	0
Consolidated Messenger	7/16/2010	$395.00	0
City Power & Light	5/23/2009	$339.69	0
Contoso Parts	3/23/2009	$338.51	0
Fabrikam Inc.	4/10/2009	$338.12	0
City Power & Light	7/9/2009	$337.73	0
Madrona Solutions Group	8/9/2009	$336.94	0

Top 10 Sales Leaders in 2009		
Sales Representative	Actual Revenue (in 1000s)	Win Rate
Anton Kirilov	$4,518.5821	47 %
System Administrator	$3,640.8966	46 %
Simon Pearson	$2,760.7613	43 %
Mark Hassall	$1,867.9274	45 %
Brian Cox	$1,798.2422	46 %
William Ngo	$1,515.8373	47 %
Robert Lyon	$1,454.1014	46 %
Lori Penor	$1,372.6267	45 %
Steve Masters	$1,359.9723	44 %
John Chen	$799.0690	41 %



Crítica…

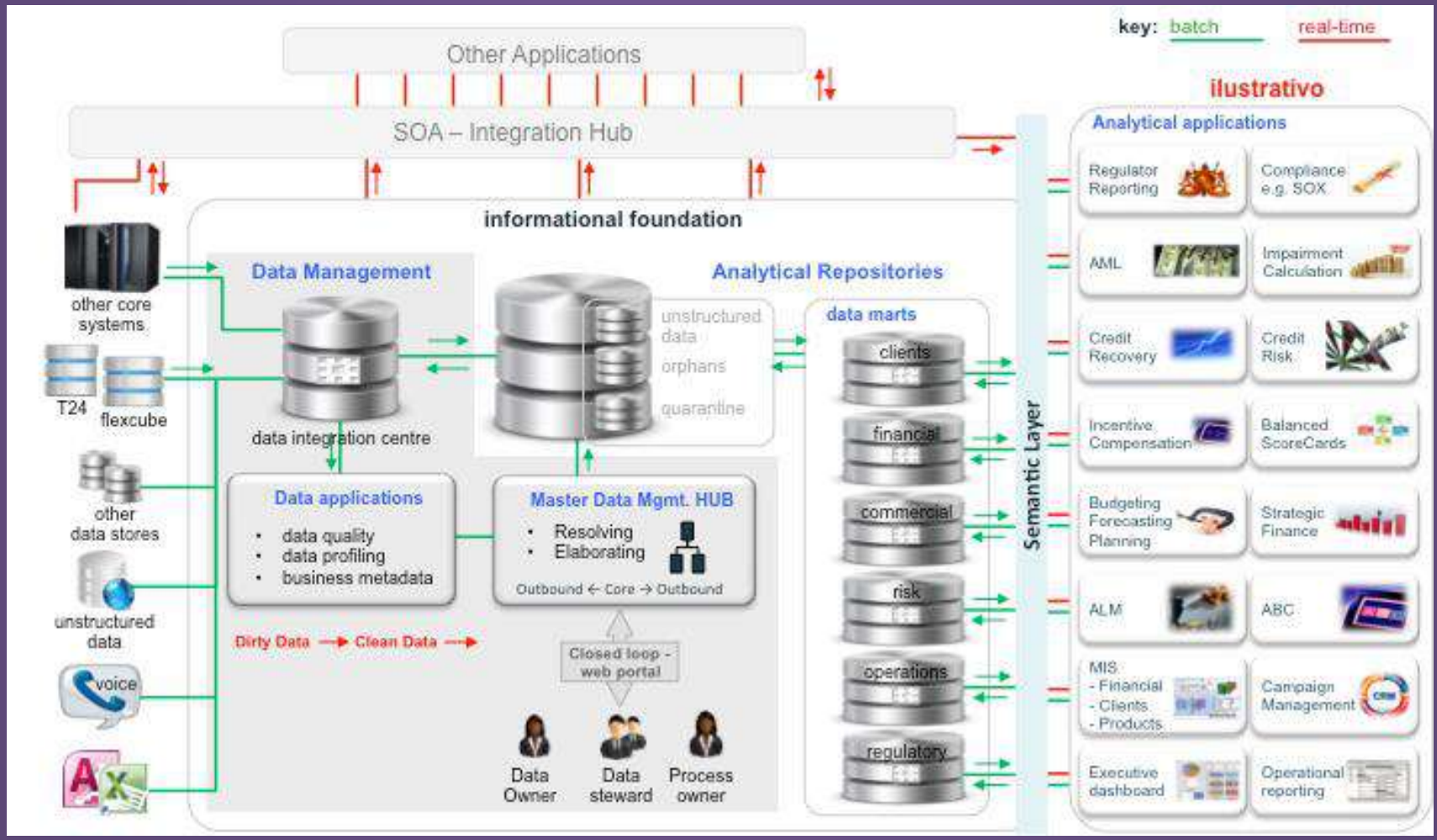
De nuevo demasiado restringido:

- Esto es sólo visualización
- Conocimiento…
  - Sistemas de Ayuda a la Decisión (DSS).
  - Sistemas Recomendadores (Recommender Systems).
  - Análisis de series temporales (Predicción vs Pronóstico).
- Segmentación.

¡¡ Patrones !!

- Las salidas condicionan todo el proceso.
- No se debe ir “a ciegas” hacia delante

Arquitectura de Referencia



Ejemplos de Soluciones de BI

Information Management	Reporting	Advanced Analytics	EPM (2)
Calidad de Datos	Cuadros de Mando	Segmentación	Presupuestación y Planificación
Gobierno del Dato	Vizualización	Next Best Offer	Consolidación Financiera
ETL(1)	Agile BI	Mantenimiento Preventivo	Rentabilidad y Costeo
EDW y Data Marts	Mobile	Modelos de Riesgo	Balanced Scorecard

(1) Extraer, Transformar y Cargar
(2) Enterprise Project Management

Tendencias

Internet, dispositivos y redes sociales

Algunas cosas a tener en cuenta…

- ✓ La imprecisión de los datos
- ✓ Privacidad y Seguridad datos
- ✓ Demanda de más capacidad de supercomputación
- ✓ Soporte a la toma de decisiones basado en datos

  (Big Data es clave en la toma de decisiones)
- ✓ Formación a gran escala

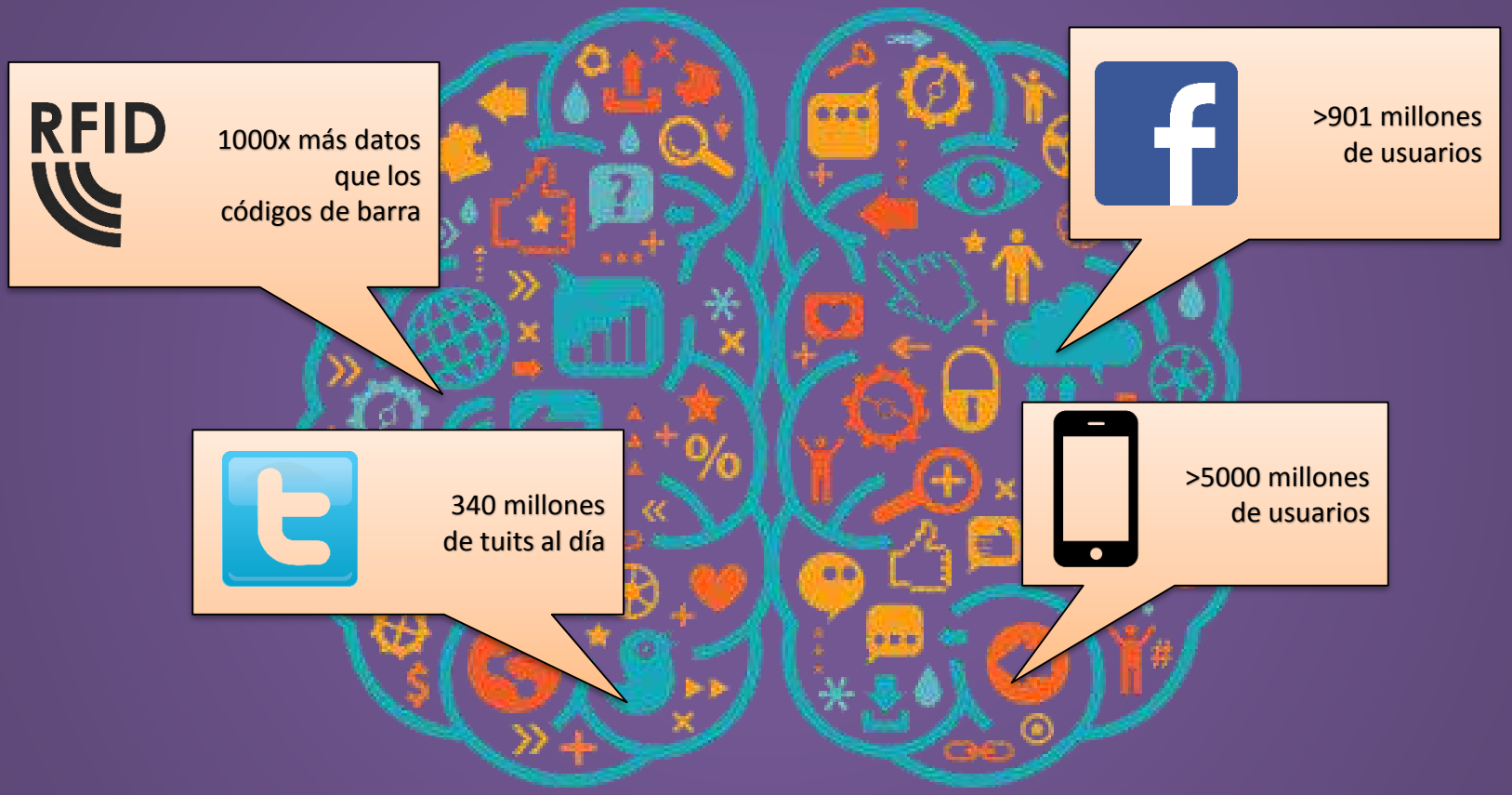

Past
Future
Data Analytics
Descriptive Analytics
Diagnostic Analytics
Predictive Analytics
Prescriptive Analytics
Business Intelligence
BIG DATA

Big Data

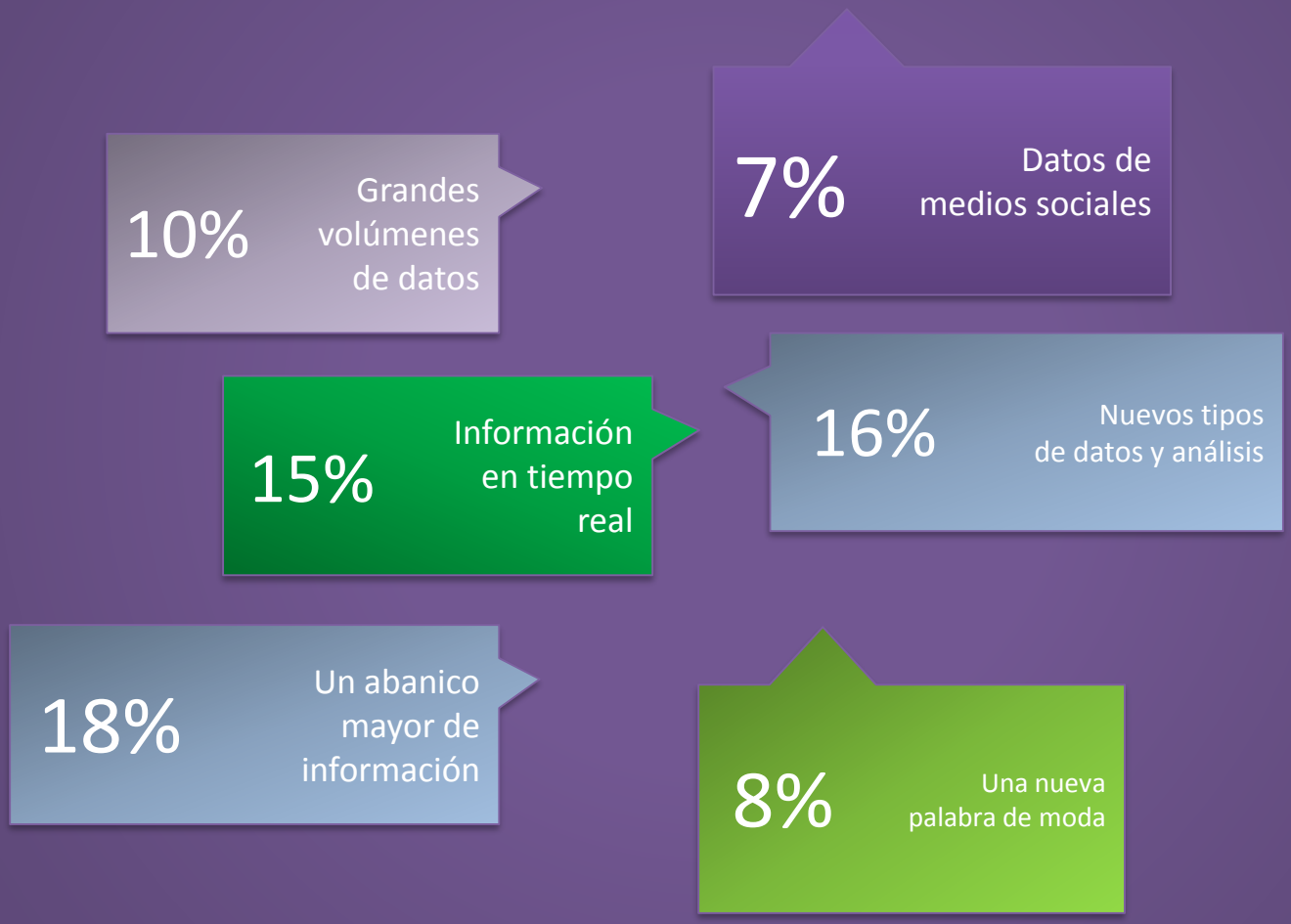
- Aproximación ingenua y crítica.
- Definición abierta de Big Data.

"**Big Data**" es en el sector de tecnologías de la información y la comunicación una referencia a los sistemas que manipulan grandes conjuntos de datos. Las dificultades más habituales en estos casos se centran en la captura, el almacenamiento, búsqueda, compartición, análisis y visualización.
(Wikipedia)

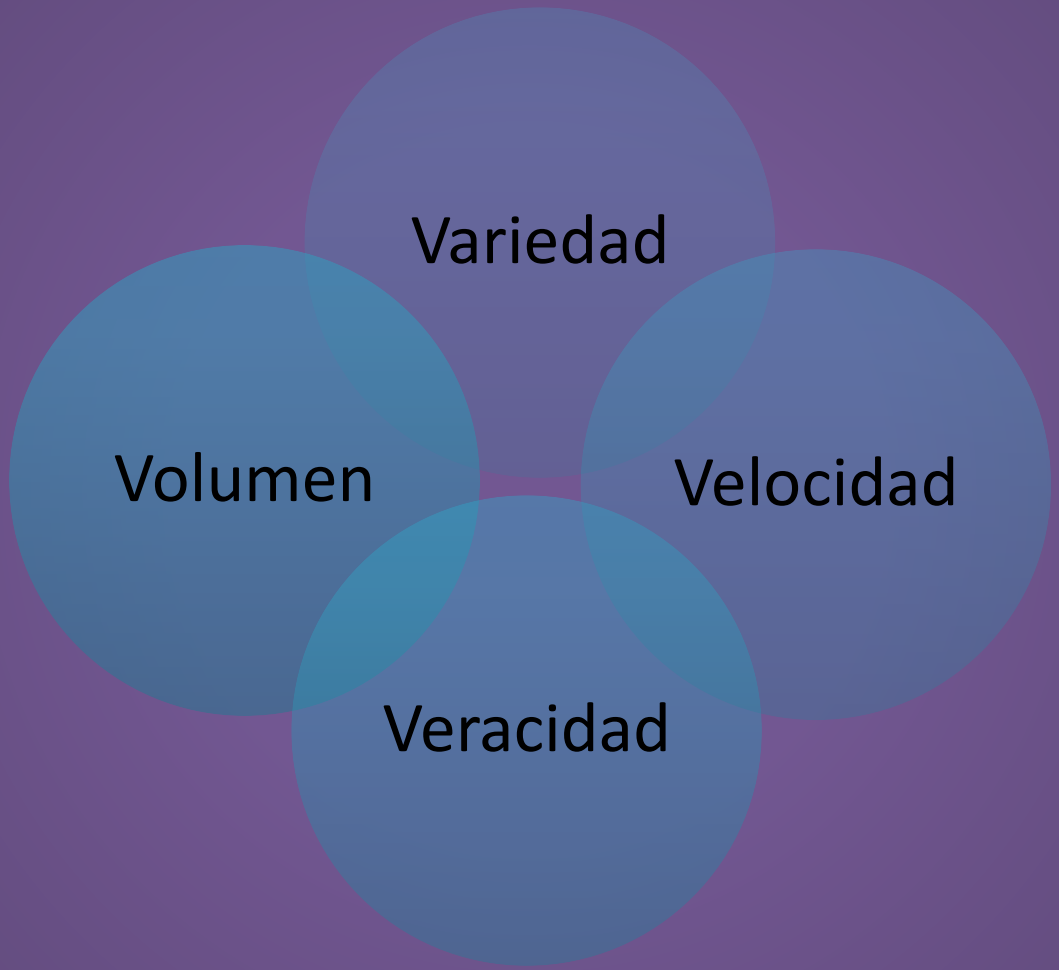
"Big data" es un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable. Los tamaños del "big data" se hallan constantemente en aumento.
(Wikipedia)


RFID
1000x más datos que los códigos de barra
340 millones de tuits al día
>901 millones de usuarios
>5000 millones de usuarios

¿Cómo se percibe el Big Data?



Las cuatro V's del Big Data



Volumen

El crecimiento de los datos

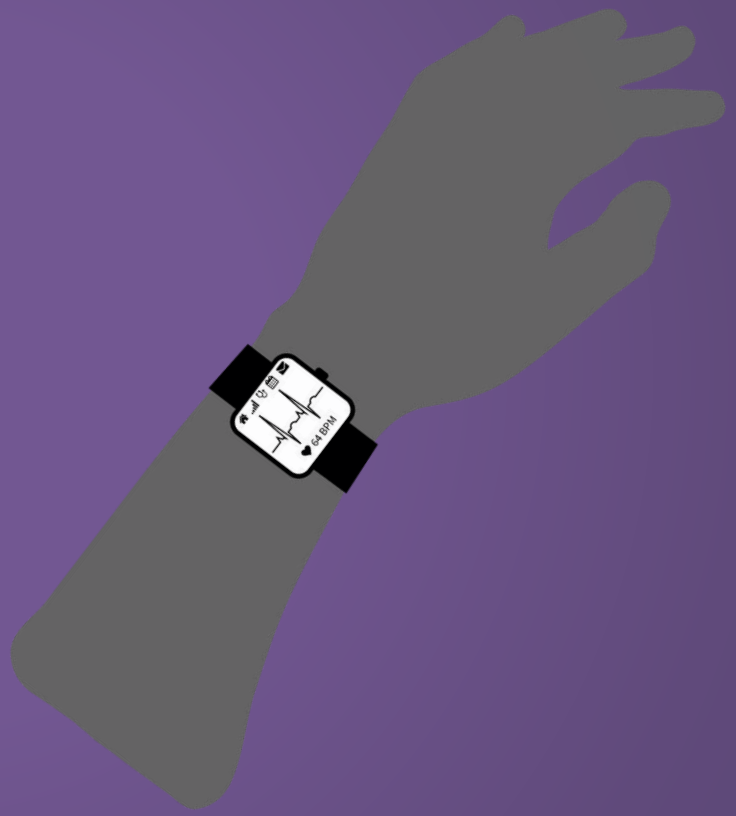
- 40 Zettabytes* de datos serán creados en 2020.
- Se estima que 2.5 quintillones de bytes de datos son creados cada día.
- 6 billones de personas tendrán smartphone.
- La mayoría de las compañías en USA, tendrán 100 Terabytes de datos almacenados.

*1 ZB= 10^3 EB = 10^6 PB = 10^9 TB = 10^{12} GB = 10^{15} MB = 10^{18} kB = 10^{21} bytes.

Variedad

Diferentes tipos de datos

- A partir del 2011, el tamaño total de datos relacionados con salud, fue estimado en 150 Exabytes.
- Para el año 2015, se prevé que haya 420 millones de wearables y monitores de salud Wireless.
- Más de cuatro millones de vídeos serán vistos en Youtube cada mes.
- 400 millones de tweets serán enviados cada día por usuarios activos.



Velocidad

Análisis del flujo de datos

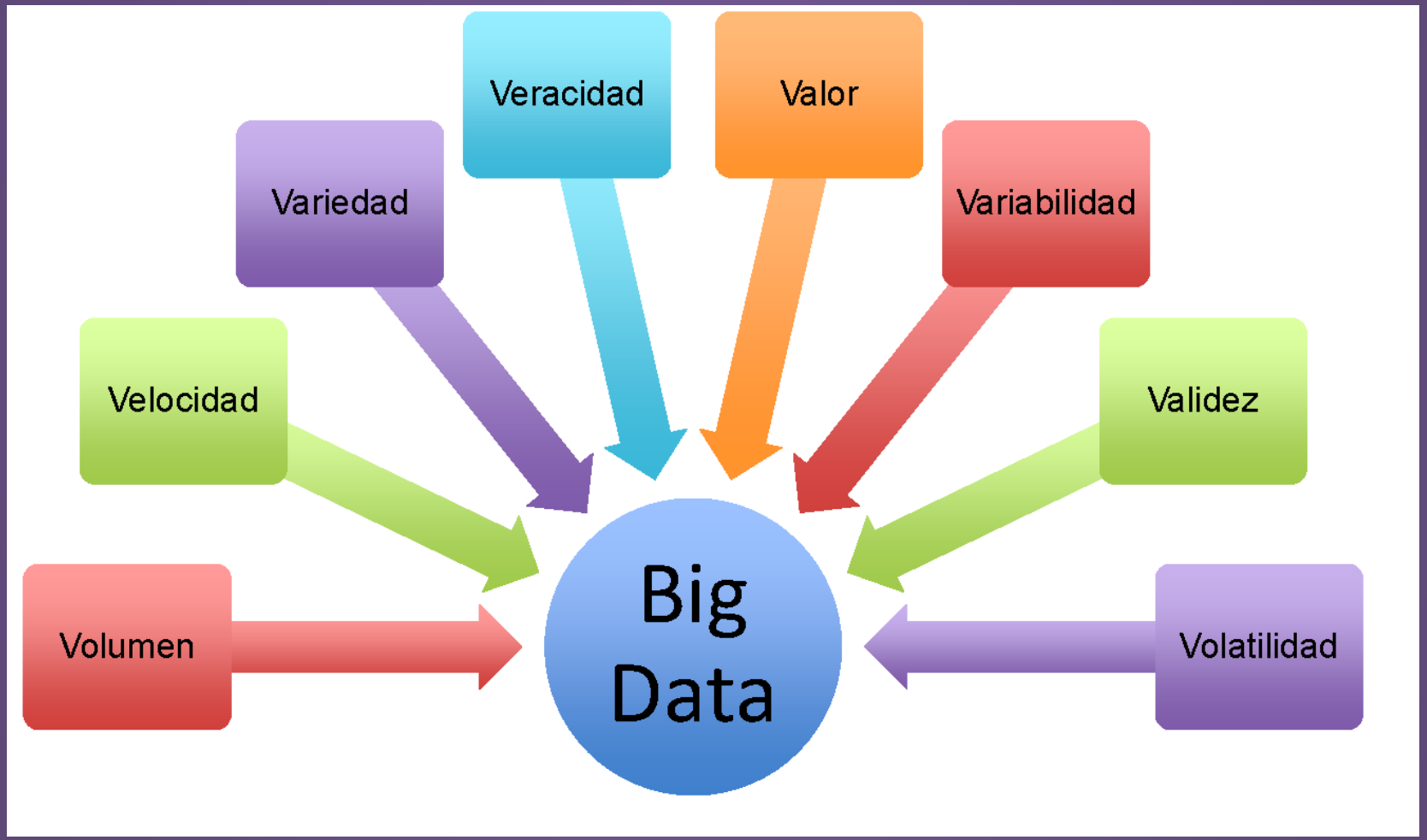
- Los coches modernos tendrán cerca de 100 sensores que monitorizarán partes del vehículo.
- Para 2016, está previsto que haya 18,9 billones de conexiones a la red. Casi 2,5 conexiones por persona en la tierra.

Veracidad

Incertidumbre de los datos

- 1 de cada 3 directores de negocio no confía en la veracidad de la información a la hora de tomar decisiones.
- La pobre calidad de los datos genera un coste en USA de 3,1 trillones de $ al año.
- El 27% de los encuentados no estuvieron seguros de cuantos de sus datos son inexactos.

¿O las 8 V's?...


Veracidad
Valor
Variedad
Variabilidad
Velocidad
Validez
Volumen
Big Data
Volatilidad

Definición

- Datos…
- Información…
- ¿Conocimiento?
  - Abstracción-Patrones…
  - Dimensión humana…
  - La Web…
  - Google…

¡¡Explosión en la cantidad de datos!!

¡EL MUNDO DIRIGIDO POR DATOS!

- Science
  - Data bases from astronomy, genomics, environmental data, transportation data, …
- Humanities and Social Sciences
  - Scanned books, historical documents, social interactions data, new technology like GPS …
- Business & Commerce
  - Corporate sales, stock market transactions, census, airline traffic, …
- Entertainment
  - Internet images, Hollywood movies, MP3 files, …
- Medicine
  - MRI & CT scans, patient health records, …

¡¡Explosión en la cantidad de datos!!

- A380:
  - Más de 1 billón de líneas de código.
  - Cada motor genera 10 Tb cada media hora.
  - Mas de 640 Tb de información por vuelo.
- Twitter genera más de 15 Tb de datos al día.
- Las principales bolsas generan más de 1 Tb al día.
- La capacidad de almacenamiento de ha doblado cada 3 años desde los 80s.

¡¡Explosión en la cantidad de datos!!

- Historias Clínicas Electrónicas:

9.000.000.000 documentos sólo en España…

¡¡Problemas graves al gestionarlos!!

- A380 de Quantas (32-2009) ¡SATURACIÓN!
- A330 de Air France (447-2010) ¡INCONSISTENCIA!
- B777 Malayo (370-2014) ¡INCERTIDUMBRE!
- Twitter, ¡ANÁLISIS DE SENTIMIENTOS! PLN.
- No se usan las Historias Clínicas Electrónicas.

¡¡Explosión en la cantidad de datos!!

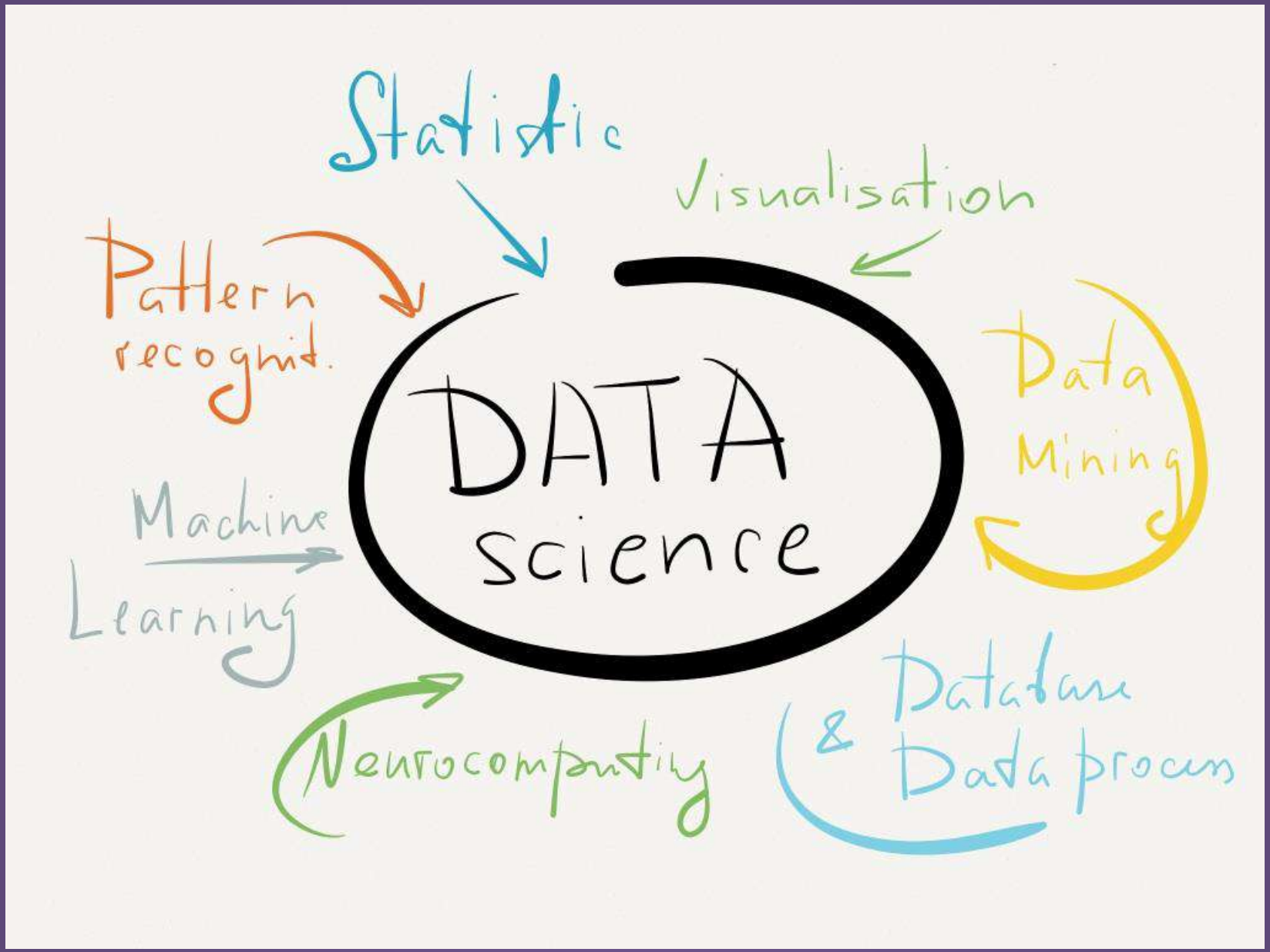
¿Habitualmente qué hacemos con todos estos datos?

¡¡Explosión en la cantidad de datos!!

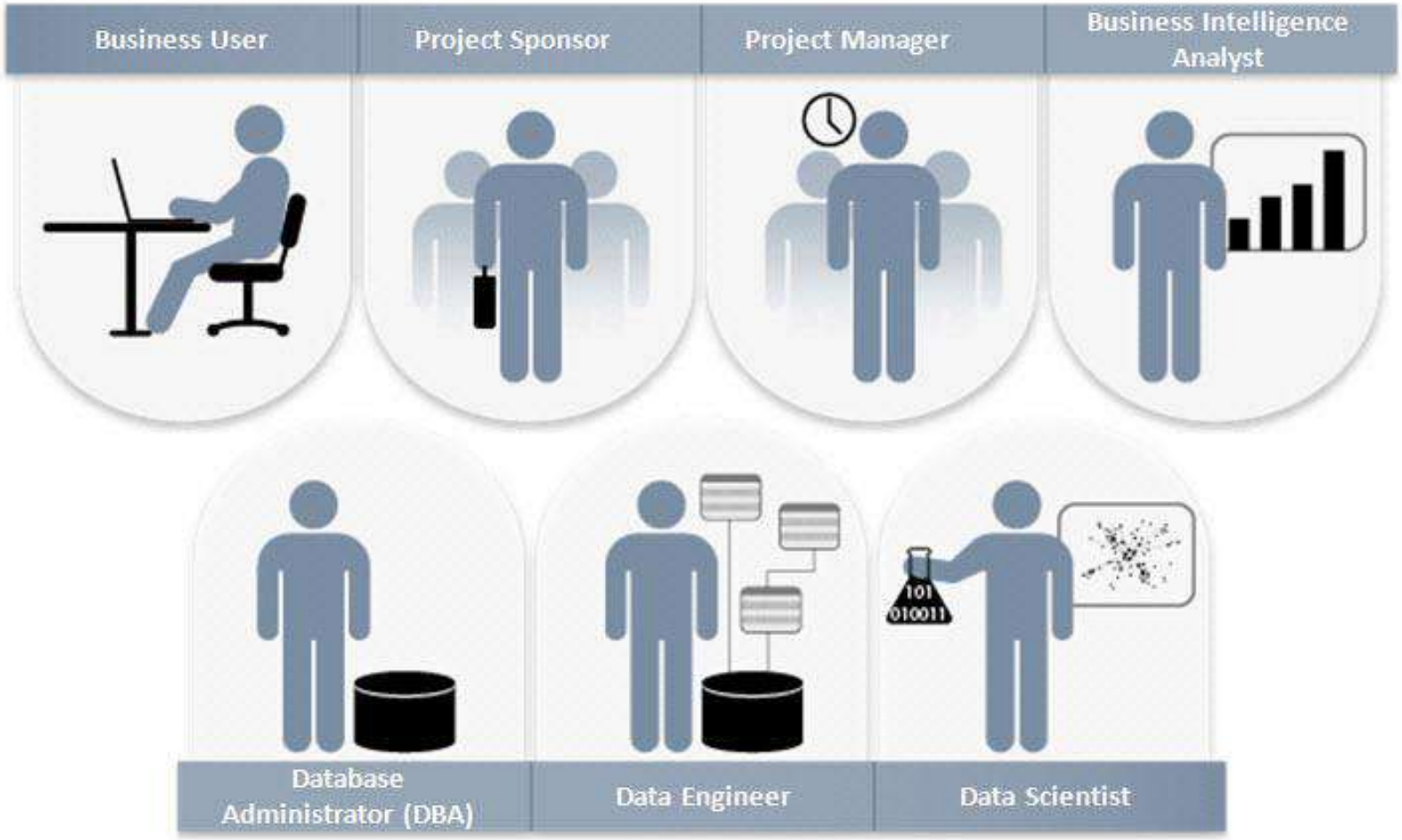
¿Habitualmente qué hacemos con todos estos datos?

¡IGNORARLOS!

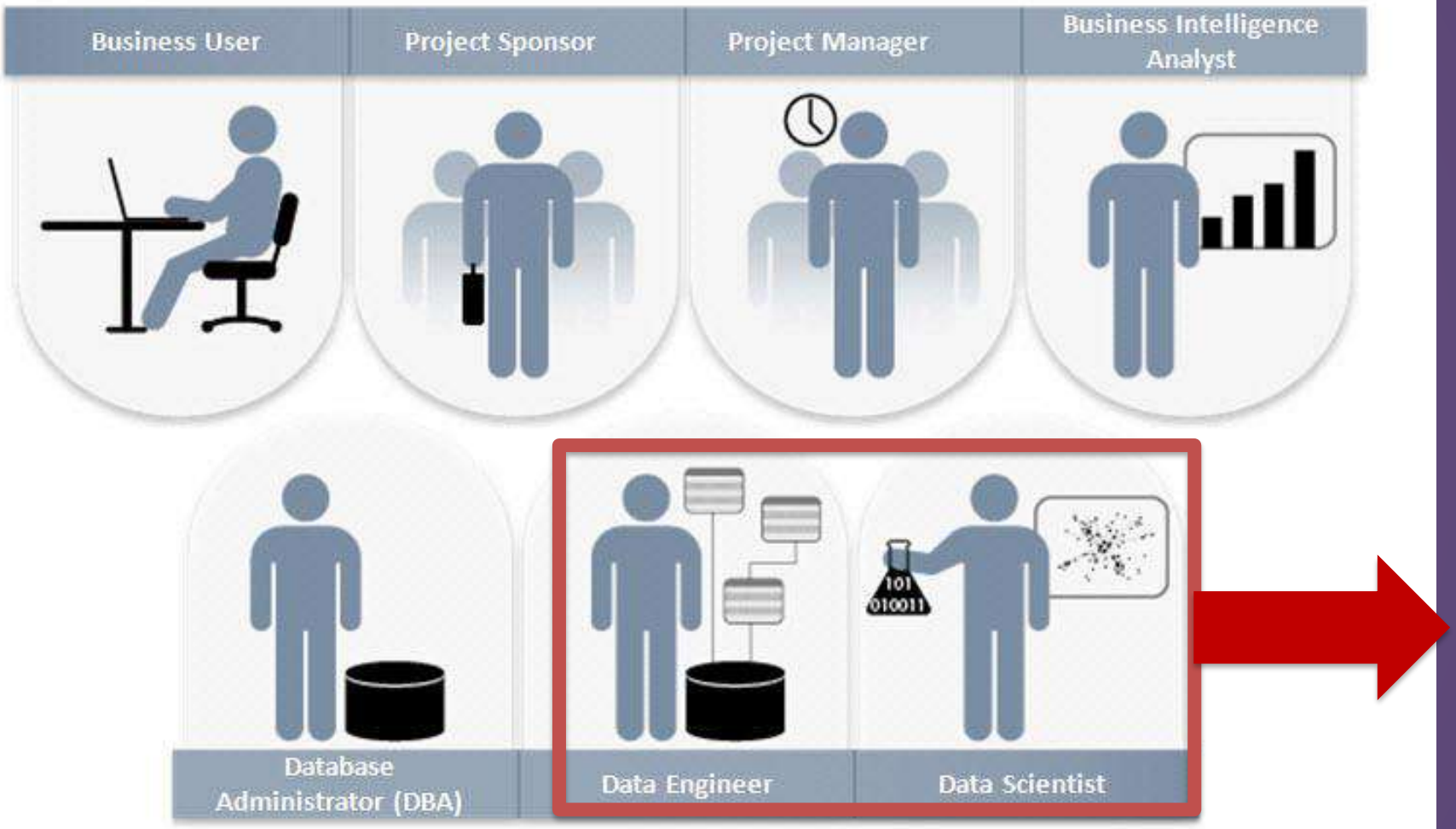
Análisis de datos

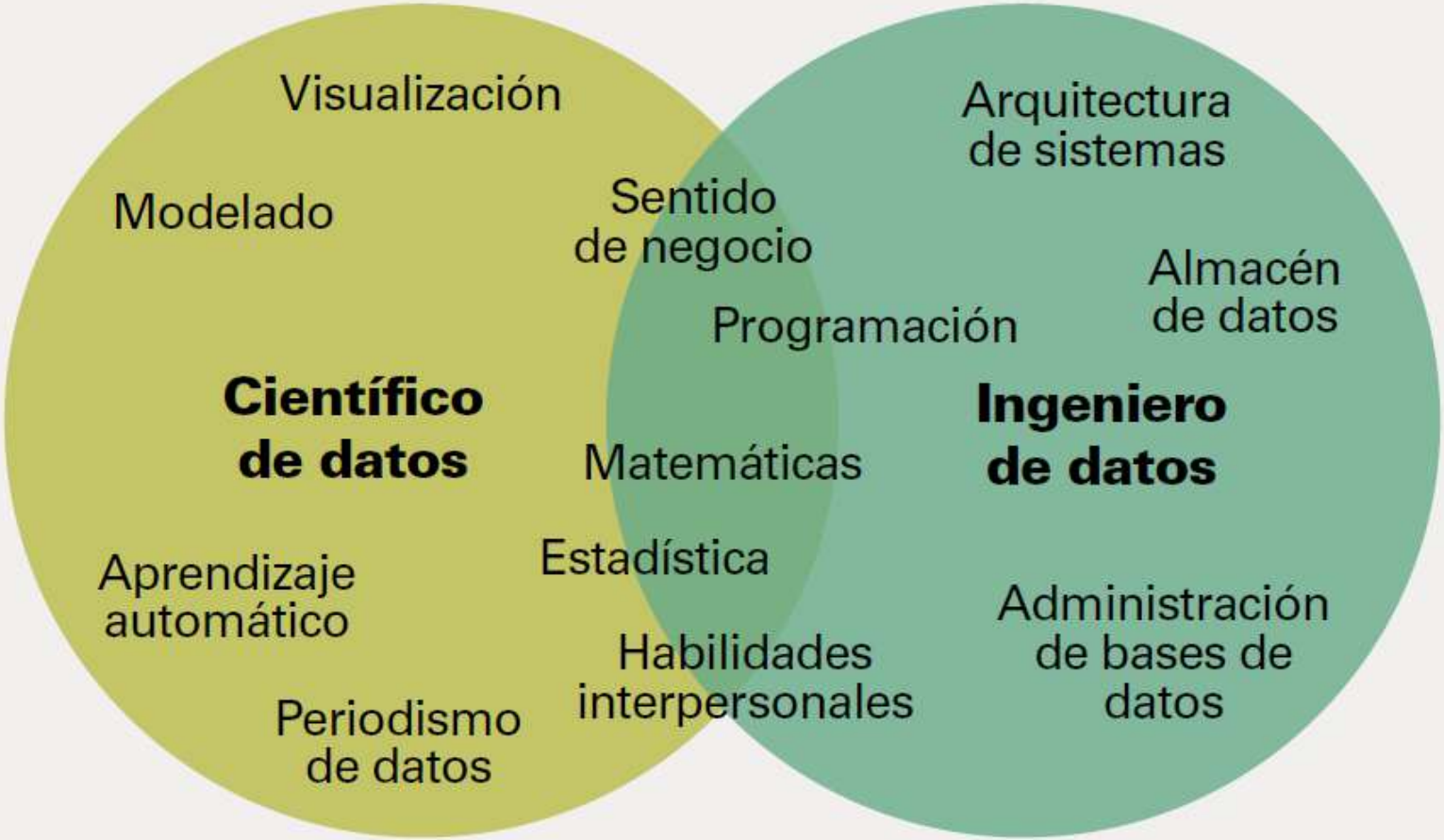

Statistic
Visualisation
Pattern recognit.
DATA science
Data Mining
Machine Learning
Neurocomputing
Database & Data procss

Key Roles for a Successful Analytic Project



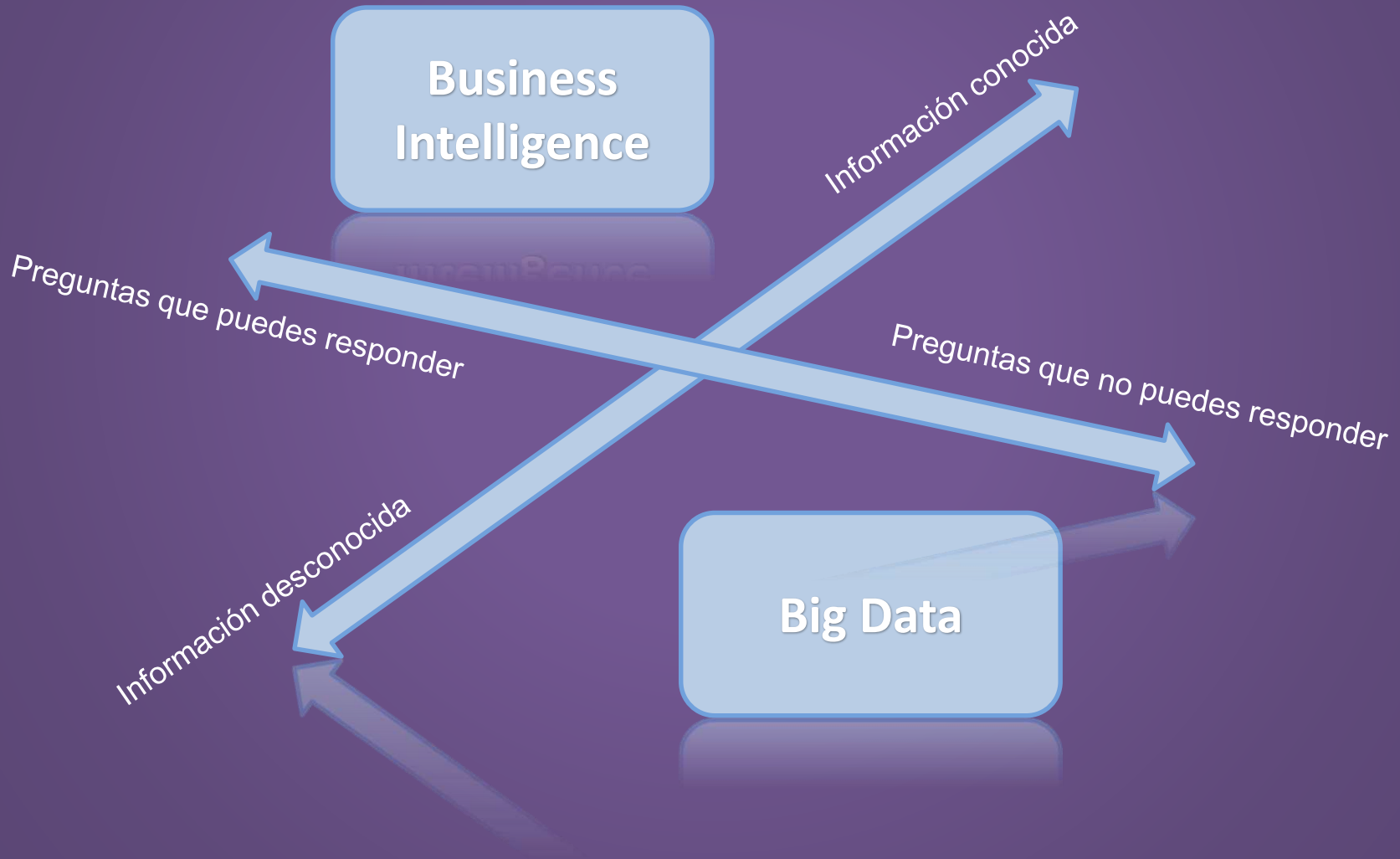
Key Roles for a Successful Analytic Project



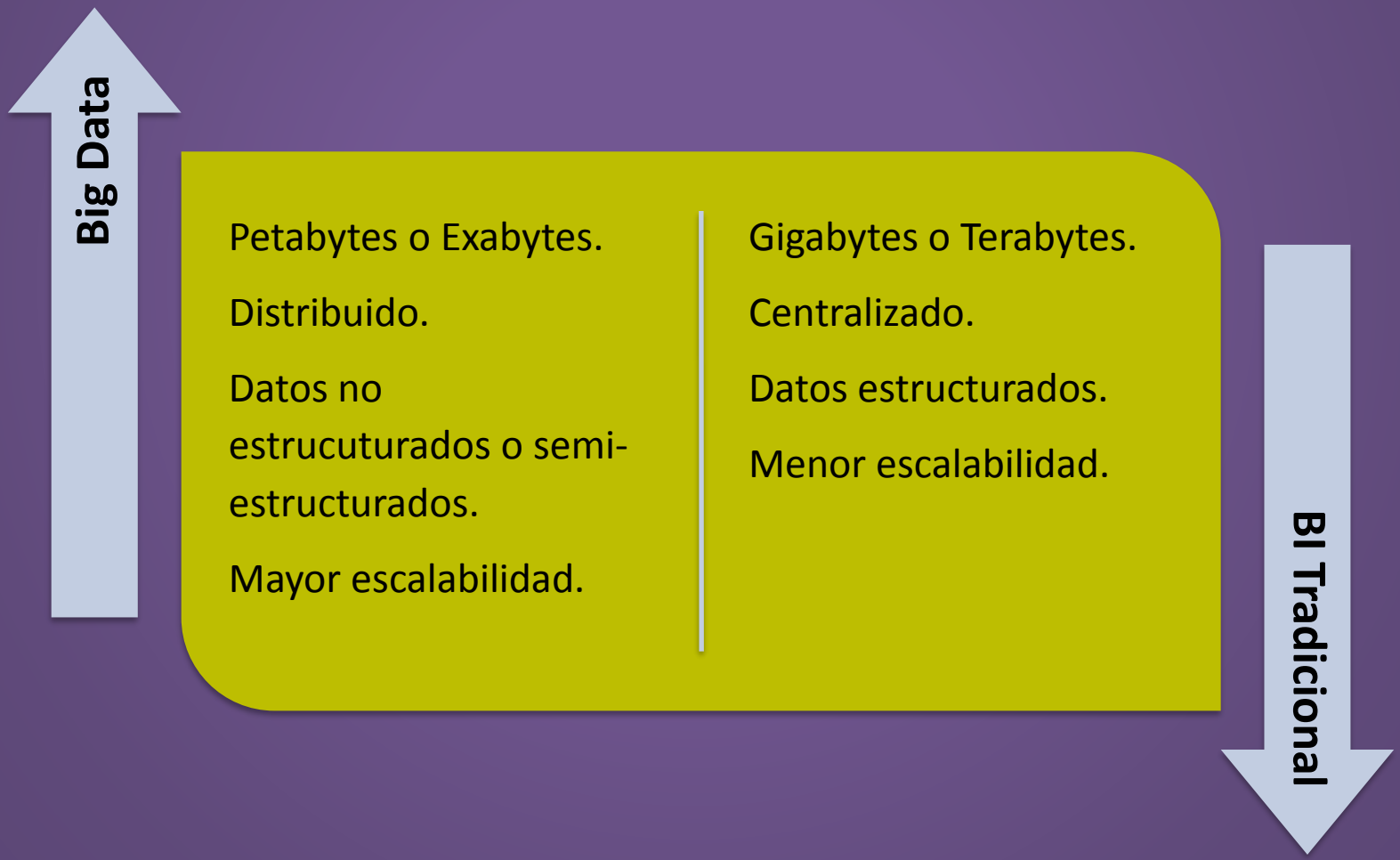


Fuente: Universitat Oberta de Catalunya. Máster en *Business Intelligence* y Big Data (2016)


Make way!
I'm a data scientist!


Business Intelligence
Información conocida
Preguntas que puedes responder
Preguntas que no puedes responder
Información desconocida
Big Data

Diferencias entre BI y Big Data



Métodos basados en la estadística para BA.

Técnicas de Regresión y correlación

- Formalización de una relación significativa entre dos o más variables para calcular pronósticos a partir del conocimiento de los valores en un individuo concreto.
  - **Lineal** (aproximación de la dependencia entre una variable dependiente y variables independientes)
  - **Múltiple** (para predecir el valor de una variable dependiente a partir de variables independientes)
  - **Logística** (para predecir variables categóricas)
  - **CART** (**C**lassification **A**nd **R**egression **T**rees, Leo Breiman)
  - Etc.

Otras Técnicas

- Técnicas de **extrapolación** de funciones.
- Técnicas de **aproximación** y **ajuste** de funciones.
- Técnicas de **agrupamiento** basadas en medidas estadísticas (**clustering**).
- Etc.

Muchas se pueden englobar tanto en Técnicas estadísticas como de Machine Learning…

Métodos basados en Inteligencia Artificial (Machine Learning) para BA.

Machine Learning (Aprendizaje automático)

- Rama de la **Inteligencia Artificial** en la que se diseñan mecanismos para dotar a los sistemas computacionales de capacidad de aprendizaje.
- Aprendizaje en el sentido de la capacidad de **descubrir regularidades (patrones)** en datos o situaciones anteriores y aplicarlos a nuevos problemas o situaciones análogas.

Principales paradigmas en Machine Learning

- **Paradigma Analógico** (Aprendizaje por analogía).
  - Pretende encontrar una solución a un problema que se presenta ahora **usando el mismo procedimiento** usado en la resolución de uno similar que se presentó en otra ocasión anterior.
  - Si dos problemas son similares en algún aspecto de su formulación entonces pueden serlo también en sus soluciones. Nuevos problemas pueden ser abordados reduciéndolos a problemas análogos resueltos.

Principales paradigmas en Machine Learning

- **Paradigma Analógico** (Ejemplos).
  - Analogía por transformación.
  - Analogía por derivación.
  - Razonamiento basado en casos.
  - Etc.

Principales paradigmas en Machine Learning

- **Paradigma Inductivo**.
  - Árboles de decisión, algoritmos de inducción pura…
- **Paradigma Conexionista**.
  - Redes Neuronales Artificiales…
- **Paradigma Evolutivo**.
  - Algoritmos Genéticos, otros métodos de optimización, colonias de insectos, descenso estocástico del gradiente…
- **Modelos gráficos probabilistas**.
  - Bayesianos, cadenas de Markov, Filtros de Kalman, redes de creencia, Máquinas de Soporte Vectorial (SVM), Metaheurísticas…

Técnicas de *Clustering* (Aprendizaje no supervisado)

- **Agrupar los elementos** de una colección en subconjuntos (clases, categorías, *clusters*), nítidos o borrosos, **en base a su similitud**. Es **no supervisado** porque las clases o categorías no se conocen a priori, las determinaran las propias similitudes entre los elementos.
- Por lo tanto, se centran en la "**medida de similitud**" entre elementos, de la que puede haber infinidad de variantes: **estadísticas**, distancias **euclídeas**, distancias **vectoriales** (coseno), distancias **borrosas**, etc…

Técnicas de *Clustering:* **EJEMPLOS**

- **Clustering Jerárquico.**
- **Paradigma Conexionista**.
  - Redes Neuronales Artificiales: **SOM** (**S**elf **O**rganized **M**aps, Mapas de Kohonen). Toolbox de Matlab SOM.
  - Etc.

Técnicas de *Clustering:* **EJEMPLOS**

- **Modelos estadísticos y probabilistas**.
  - K-means, c-means,
  - K-nearest neighbours (KNN),
  - Mean shift (ventanas circulares con un centroide),
  - Dirichlet process (estocásticos basados en distribuciones de probabilidad). LDA (Latent Dirichlet Allocation),
  - Modelos Gaussianos,
  - Etc.

Técnicas de *Clustering:* **EJEMPLOS**

- **Extensiones basadas en Lógica Borrosa**.
  - Fuzzy K-means,
  - Fuzzy c-means,
  - Isodata,
  - Etc.

Técnicas de Clasificación (Aprendizaje supervisado)

- Asignar una clase a un nuevo elemento en base a un conjunto de categorías previamente establecidas (**supervisado**), por ejemplo, evaluar los síntomas de un nuevo paciente y decir que tiene gripe (clase previamente establecida).
- Se basan en un **entrenamiento** en base a ejemplos con la **solución conocida** (supervisado) para crear **modelos** que permitan **clasificar nuevos casos** análogos.