UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

SMILe
Soft Management of Internet and Learning

Análisis de Sentimientos y otros retos del aprovechamiento inteligente de los datos masivos

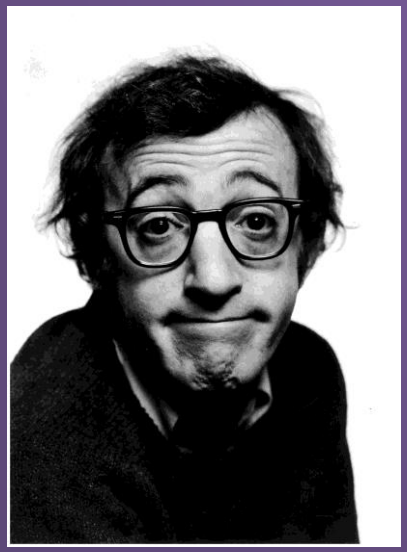
José A. Olivas
Joseangel.olivas@uclm.es

UCM, mayo, 2018

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

SMILe
Soft Management of Internet and Learning

Desmontando a Harry, a Google y otros mitos de la era digital…



UCM, mayo, 2018

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

SMILe

Soft Management of Internet and Learning

El aprendiz de Data Scientist…

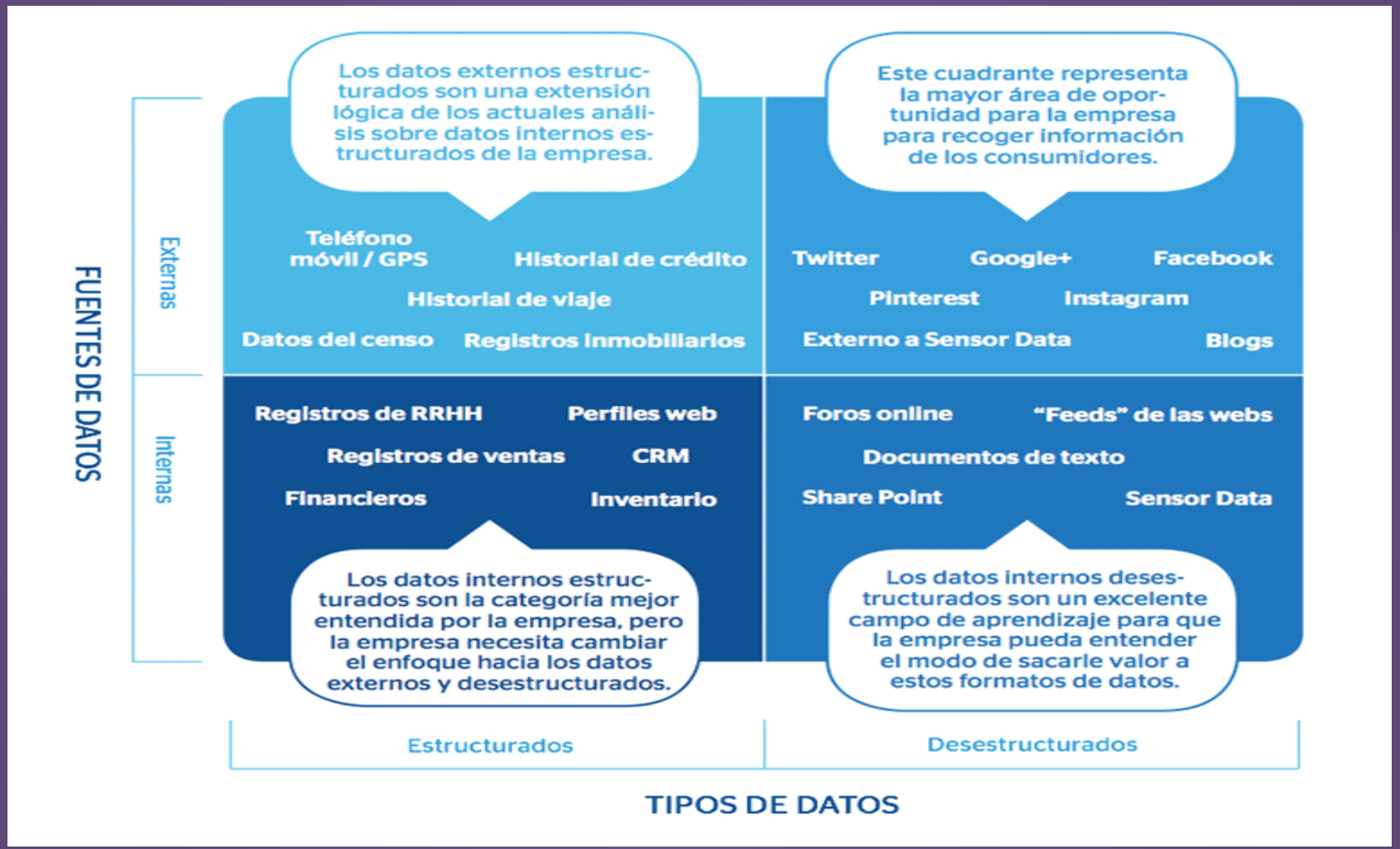
UCM, mayo, 2018

Comencemos por el final…

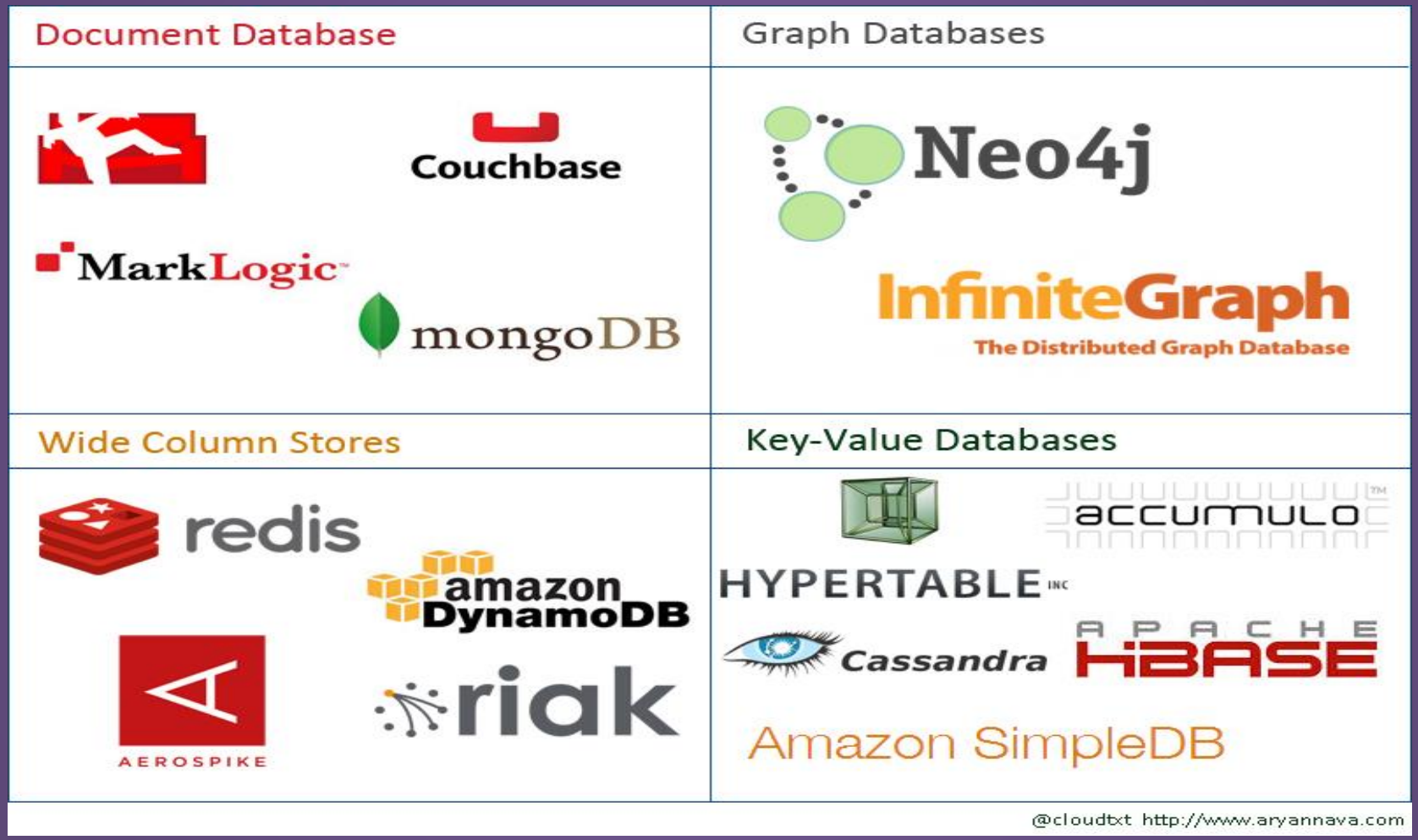
…los datos masivos

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

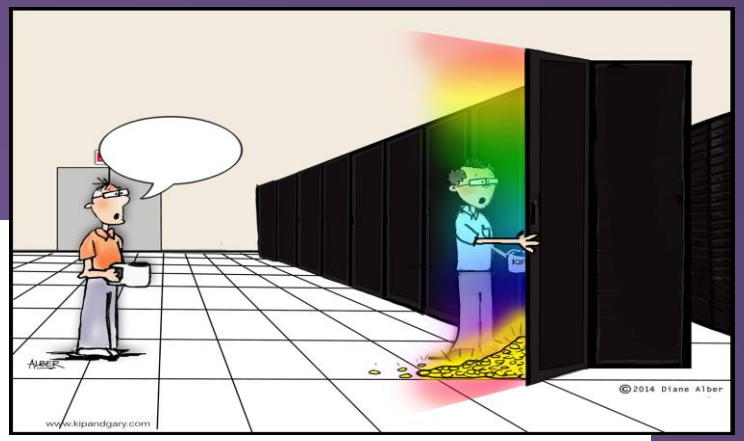
Tipos de datos (I)



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.
- IoT, Smart Cities…
- **Redes sociales**
- Etc, Etc…

¿ 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.
- IoT, Smart Cities…
- **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…

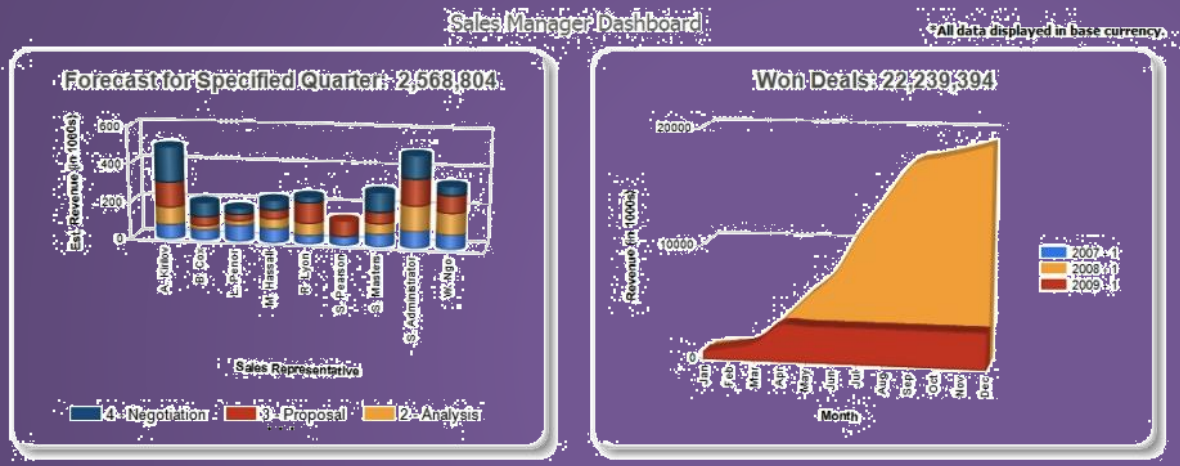
Demasiado restringido:

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

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

- Veamos las posibles salidas…

Outputs…

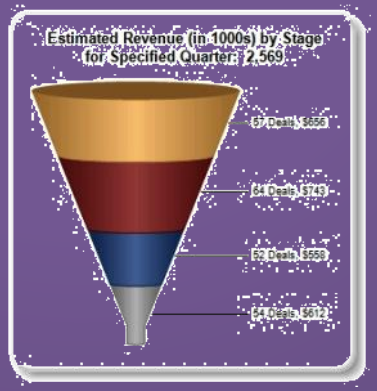


Top 10 Key Deals

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.	2/17/2010	$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	$337.78	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


Past
Future
Data Analytics
Descriptive Analytics
Diagnostic Analytics
Predictive Analytics
Prescriptive Analytics
Business Intelligence
Big Data Studio
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)

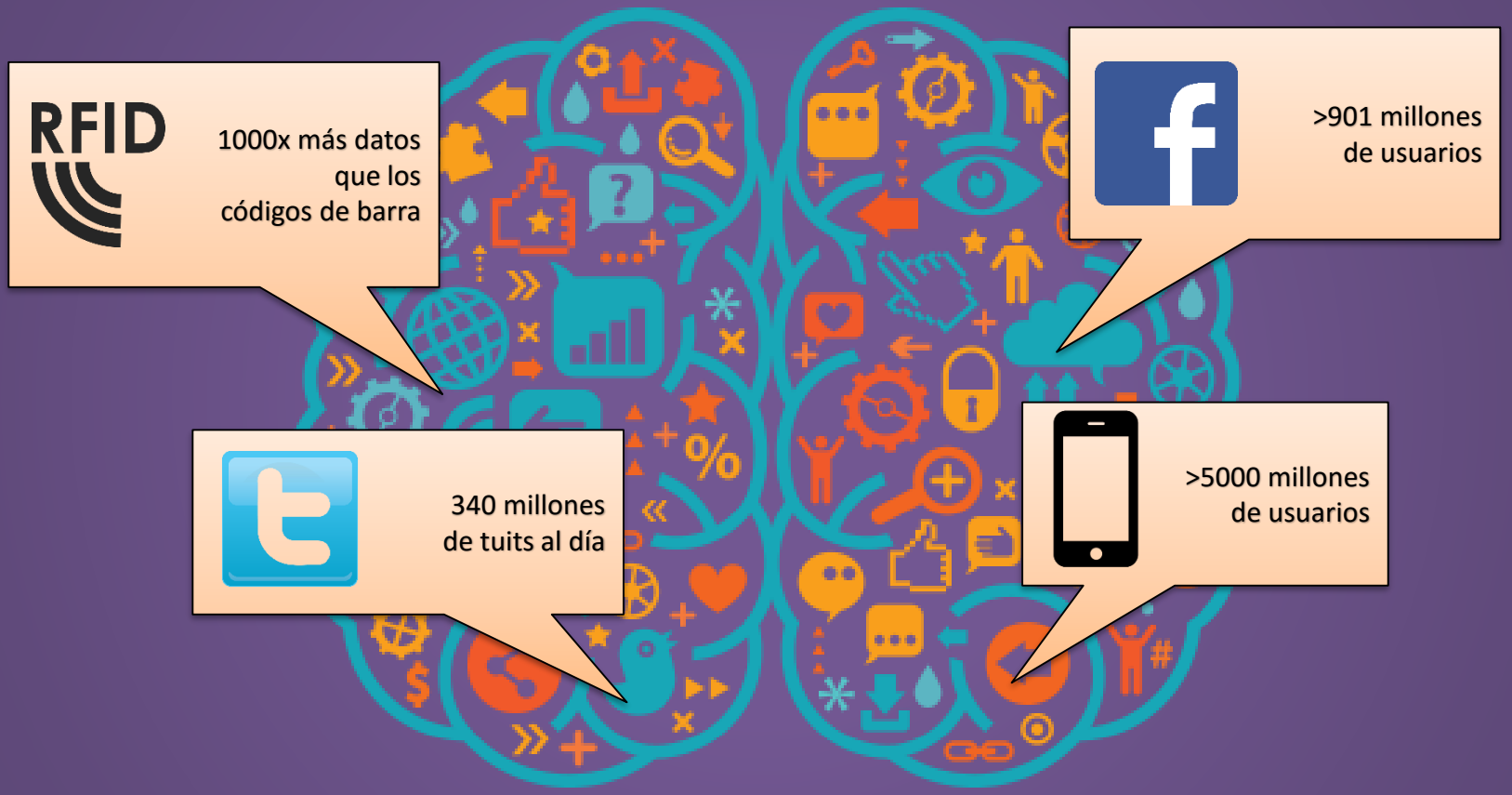
Big Data

- 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)

¡El nudo Gordiano de Google!


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

¿Cómo se percibe el Big Data?

10% Grandes volúmenes de datos

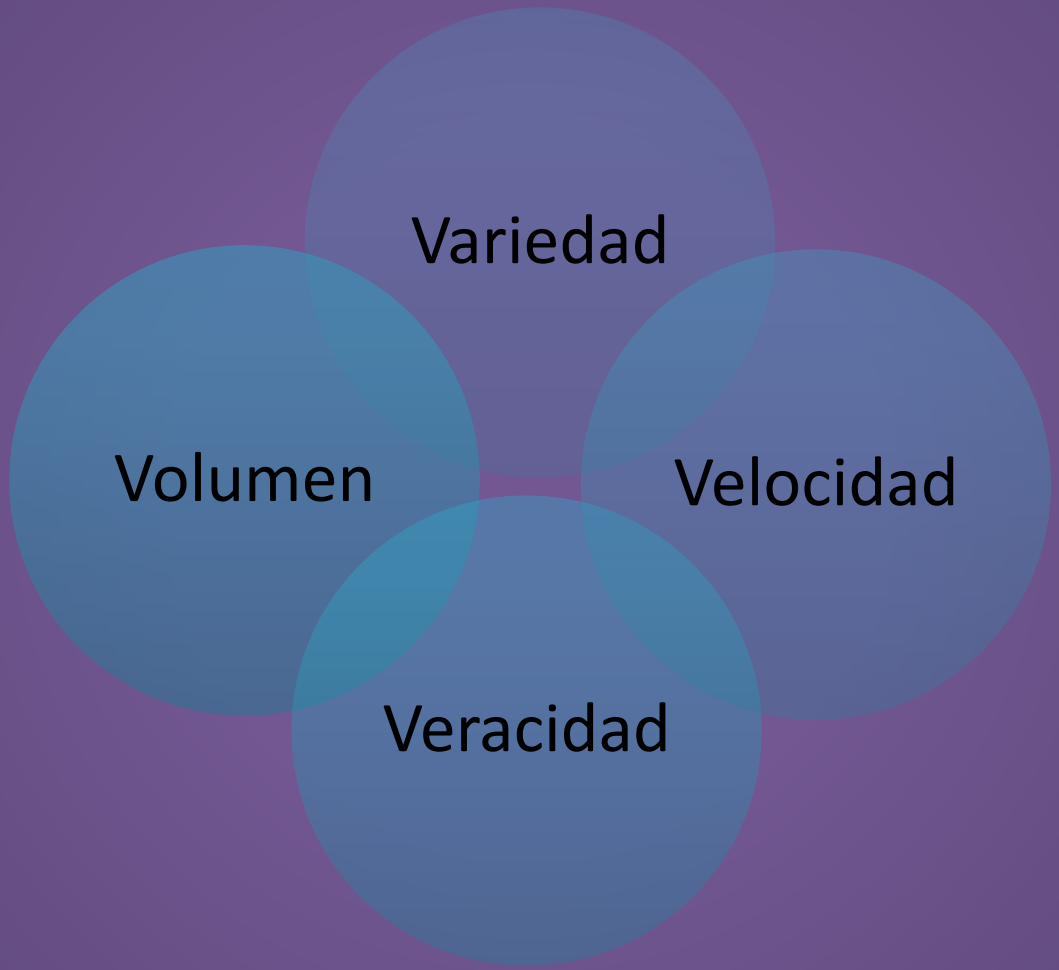
7% Datos de medios sociales

15% Información en tiempo real

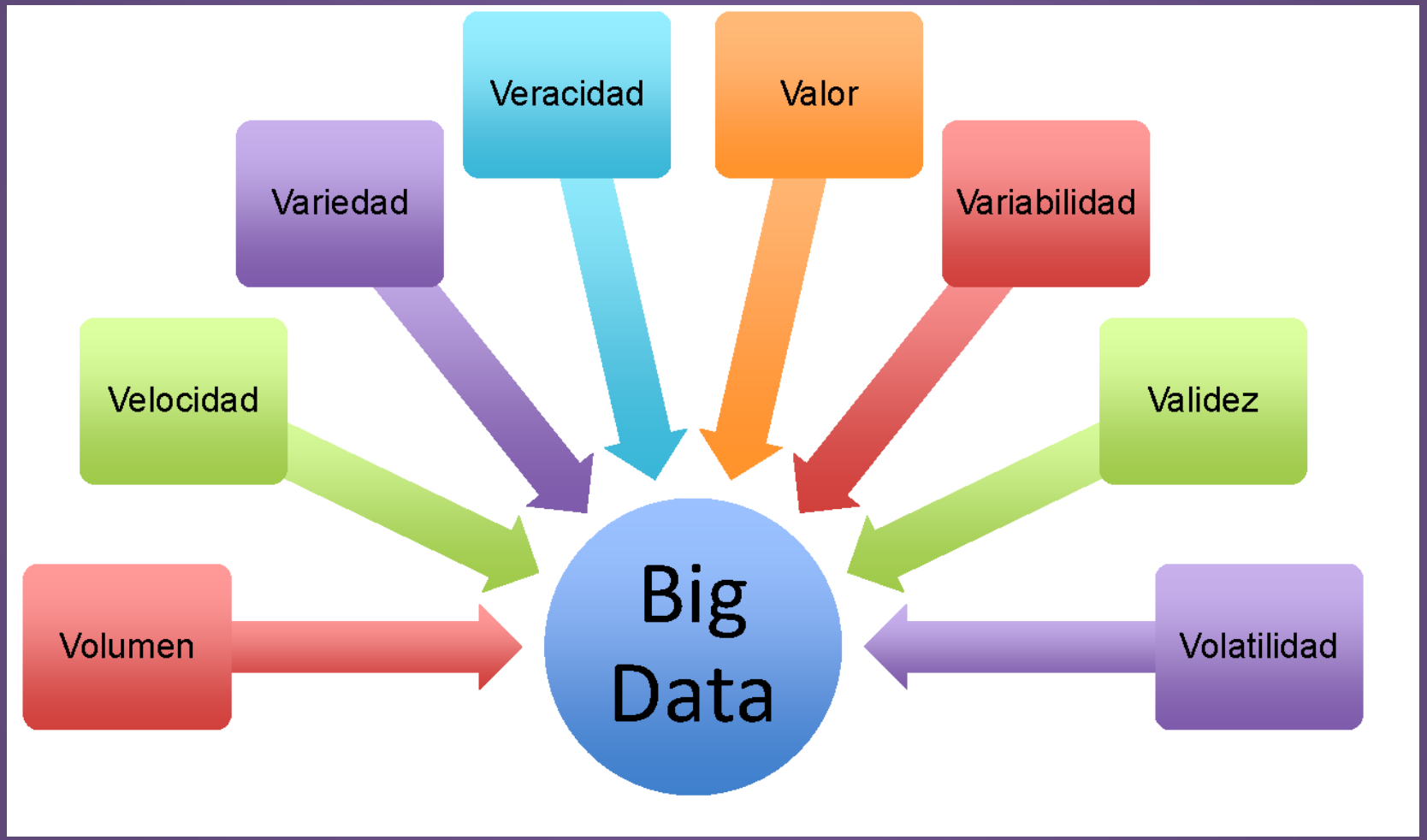
16% Nuevos tipos de datos y análisis

18% Un abanico mayor de información

8% Una nueva palabra de moda


Variedad
Volumen
Velocidad
Veracidad

¿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!!

- 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?

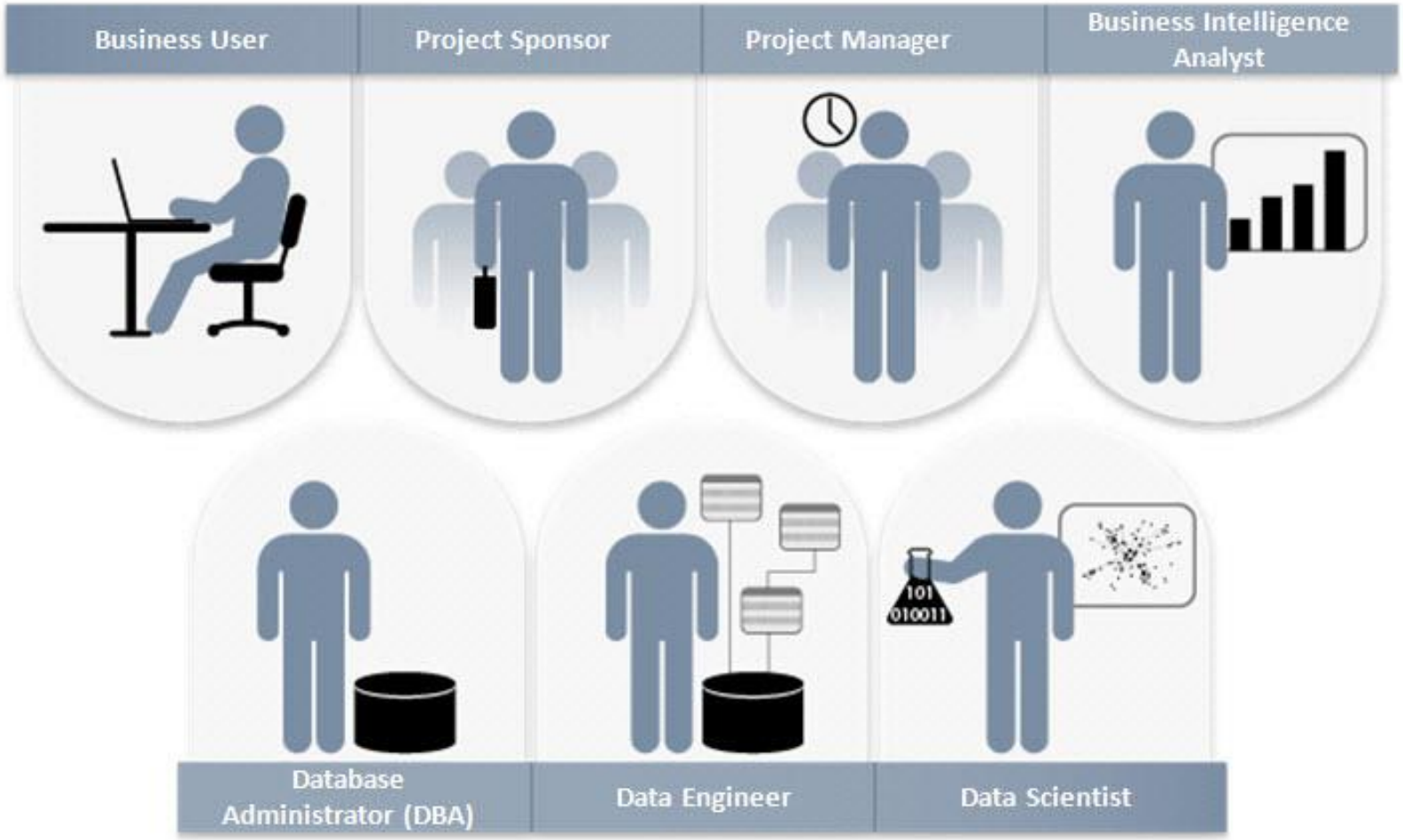
¡¡Explosión en la cantidad de datos!!

¿Habitualmente qué hacemos con todos estos datos?

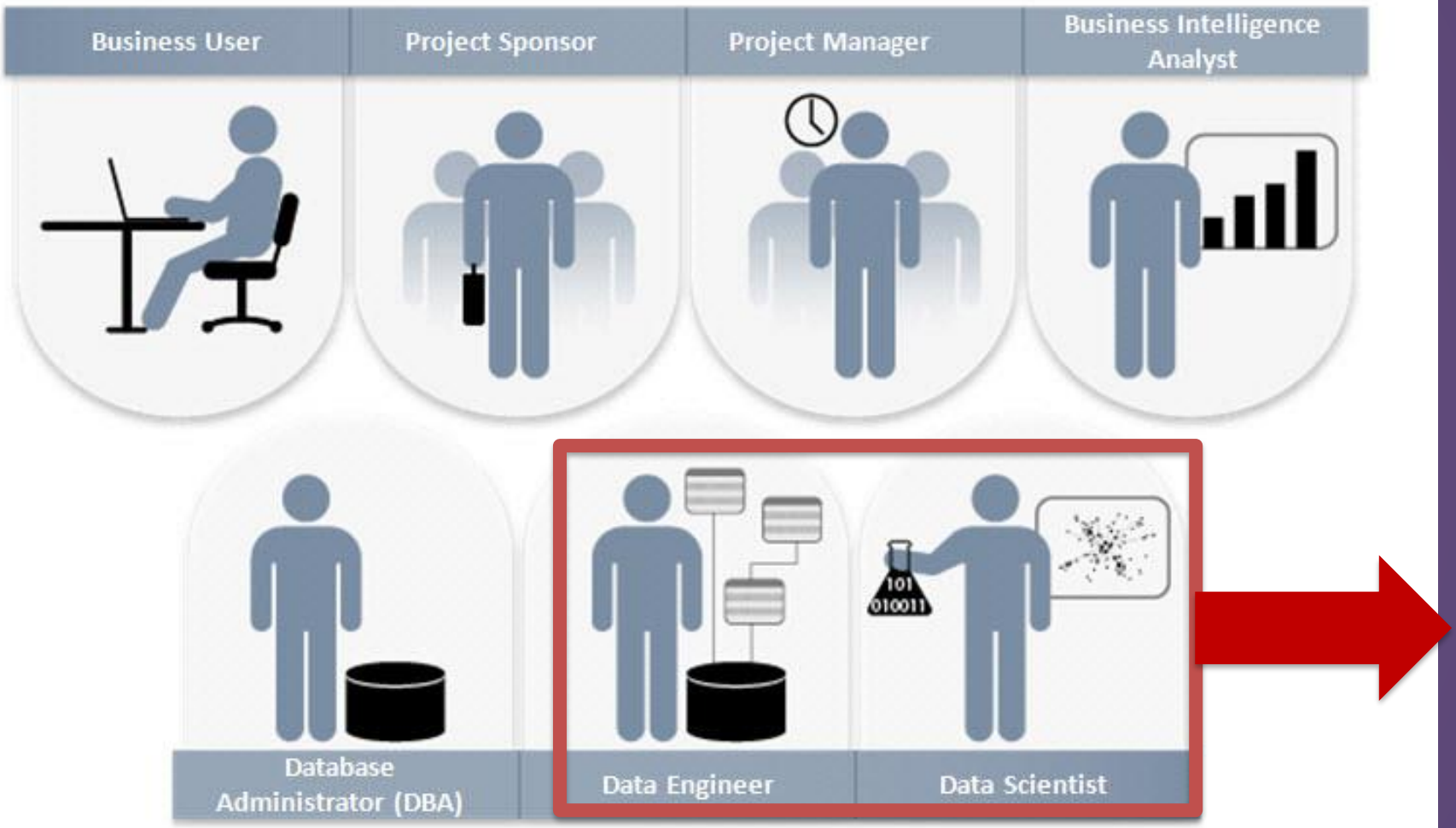
¡IGNORARLOS!

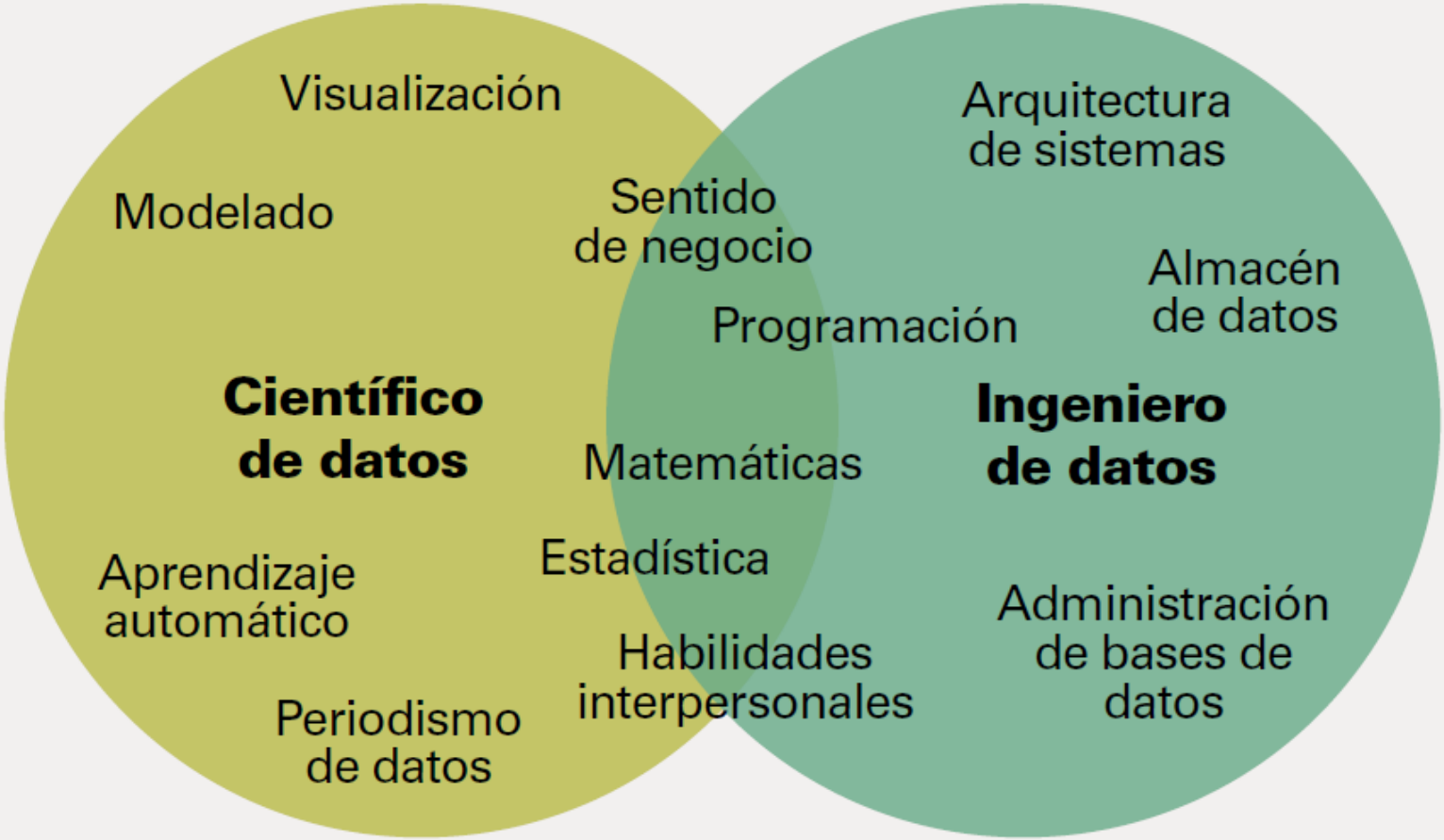
…el aprovechamiento inteligente

Key Roles for a Successful Analytic Project



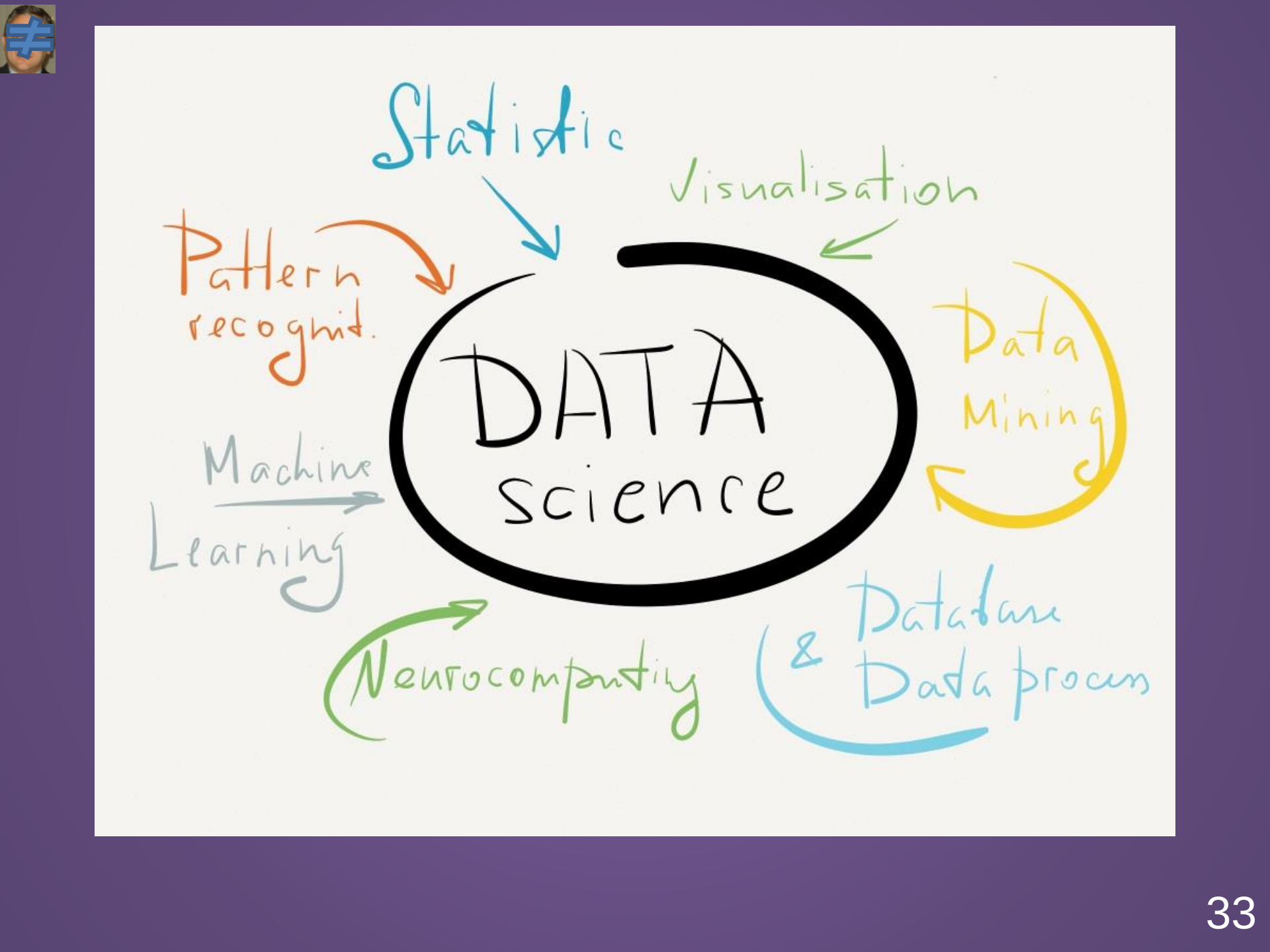
Key Roles for a Successful Analytic Project





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


Make way!
I'm a data scientist!


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

Métodos basados en la estadística

- Técnicas de **Regresión** y correlación
  - **Lineal**
  - **Múltiple**
  - **Logística**
  - **CART** (**C**lassification **A**nd **R**egression **T**rees, Leo Breiman)
- Técnicas de **extrapolación** de funciones.
- Técnicas de **aproximación** y **ajuste** de funciones.
- Técnicas de **agrupamiento** basadas en medidas estadísticas (**clustering**).

- Etc.

Métodos de Machine Learning (Aprendizaje automático)

- **Aprendizaje por Analogía.**
  - (Transformacional, derivacional, Razonamiento basado en Casos…)
- **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:* **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:* **EJEMPLOS**

- **Paradigma Inductivo**. Árboles de decisión:
  - ID3,
  - CART,
  - C4.5,
  - See5,
  - **Random Forest** (de moda en Big Data), Leo Breiman 2001,
  - Etc.

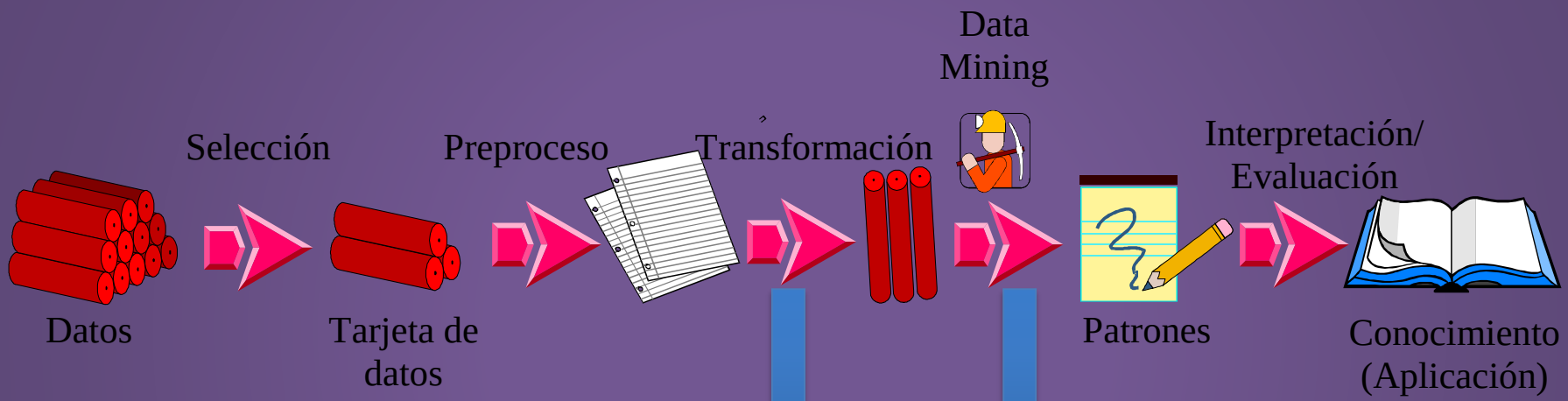
Técnicas de *Clasificación:* **EJEMPLOS**

- **Paradigma Conexionista**. Redes Neuronales Artificiales:
  - Perceptrón Multicapa (con backpropagation),
  - Convolucionales,
  - Neocognitrones,
  - Redes de Hopfield,
  - Redes recurrentes,
  - Adaline,
  - **Deep Learning** (de moda en Big Data),
  - Etc.

Técnicas de *Clasificación:* **EJEMPLOS**

- **Modelos estadísticos y probabilistas**.
  - Redes Bayesianas,
  - Naive-Bayes,
  - Máquinas de Soporte Vectorial (SVM),
  - Metaheurísticas,
  - Etc.

La importancia del KDD



Consisten habitualmente en convertir un proceso de ‘Clustering’ en uno de ‘clasificación’.

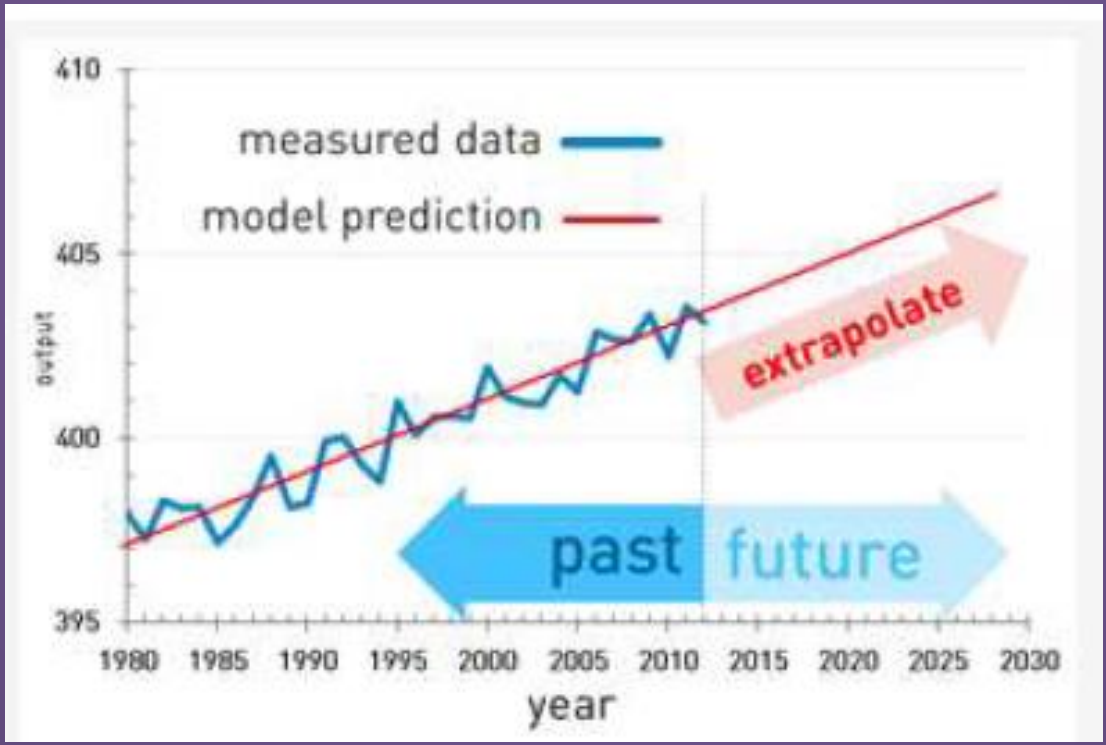
Adecuación de los métodos a los problemas.

Análisis Predictivo

- **Extrapolación de funciones** (Tendencia para el futuro, pero no hay capacidad de pronóstico –hechos/cambios puntuales-).
- **Correlaciones entre variables** (Demasiado evidente, no suele funcionar de forma muy fina).
- Encontrar **‘patrones’** en los datos que puedan ser aplicados a situaciones futuras (KDD y Data Mining).
- Métodos de **CLUSTERING** Y **CLASIFICACIÓN**.

Análisis Predictivo

- **Extrapolación de funciones** (Por ejemplo Estimaciones o Líneas de Tendencia).



Análisis Prescriptivo

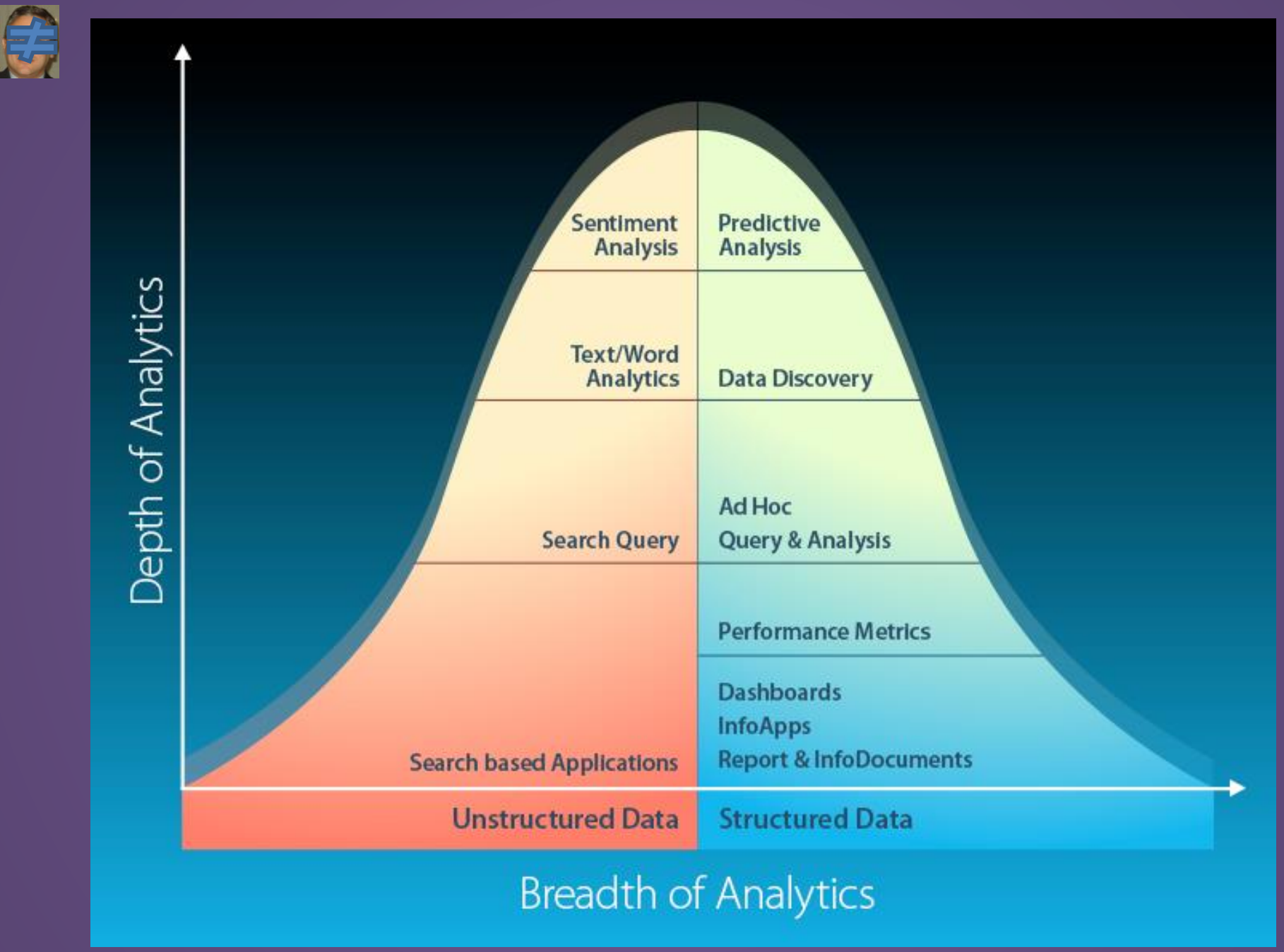
- El análisis predictivo se centra en **un escenario** futuro.
- El prescriptivo se centra en **múltiples alternativas**.
- Por lo tanto, un modelo prescriptivo puede ser considerado como **una combinación de modelos predictivos** (uno por cada posible escenario), que **se ejecutan en paralelo**.
- El objetivo es encontrar la mejor opción posible: **OPTIMIZACIÓN**.

Análisis Prescriptivo

- Técnicas:
  - Técnicas de Investigación Operativa,
  - Algoritmos Genéticos,
  - Técnicas estocásticas,
  - Metaheurísticas,
  - Etc.

Análisis Prescriptivo

…análisis de sentimientos y otros retos



http://www.informationbuilders.com/

Acceso y la búsqueda de información digital.

- **Los buscadores son eficientes, pero no eficaces.**
- **Ejemplos: sinonimia, veracidad (reputación), variedades diatópicas, operadores, tendencias…**

"Búsqueda eficaz de información en la Web" (EDULP, 2011)

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  “asíncrono” vs. “síncrono”

  Reflexión/preparación vs. Inmediatez/visceralidad

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  “asíncrono” vs. “síncrono”

  Reflexión/preparación vs. Inmediatez/visceralidad
- ¡Dimensión Humana!

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  “asíncrono” vs. “síncrono”

  Reflexión/preparación vs. Inmediatez/visceralidad

- ¡Dimensión Humana! CONTEXTO

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  “asíncrono” vs. “síncrono”

  Reflexión/preparación vs. Inmediatez/visceralidad

- ¡Dimensión Humana! CONTEXTO

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  “asíncrono” vs. “síncrono”

  Reflexión/preparación vs. Inmediatez/visceralidad
- ¡Dimensión Humana! CONTEXTO
- El gran reto: PLN

Internet y las redes sociales.

- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  "asíncrono" vs. "síncrono"

  Reflexión/preparación vs. Inmediatez/visceralidad
- ¡Dimensión Humana! CONTEXTO
- El gran reto: PLN
- Análisis de Sentimientos.

"Sentiment analysis: A review and comparative analysis of web services" (Information Sciences, 2015)

Internet y las redes sociales.

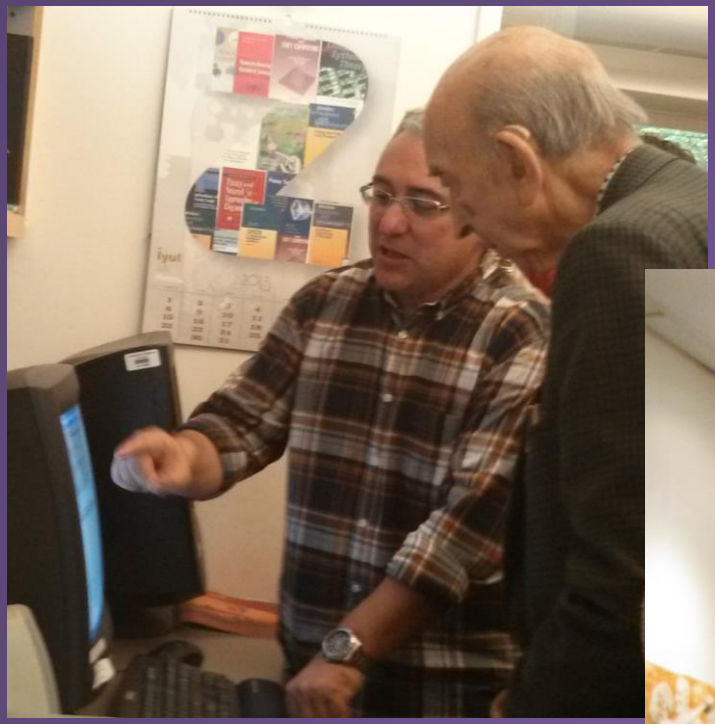
- El nuevo reto del análisis inteligente en Internet y las redes sociales:

  "asíncrono" vs. "síncrono"

  Reflexión/preparación vs. Inmediatez/visceralidad
- ¡Dimensión Humana! CONTEXTO
- El gran reto: PLN
- Análisis de Sentimientos. ¿SIRI?

"Sentiment analysis: A review and comparative analysis of web services" (Information Sciences, 2015)

www.prometeusgs.com



Lógica Borrosa

Lotfi A. Zadeh

1921-2017

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

SMILe
Soft Management of Internet and Learning

Análisis de Sentimientos y otros retos del aprovechamiento inteligente de los datos masivos

¡¡MUCHAS GRACIAS!!

José A. Olivas
Joseangel.olivas@uclm.es

UCM, mayo, 2018

¡Hasta siempre Lotfi!