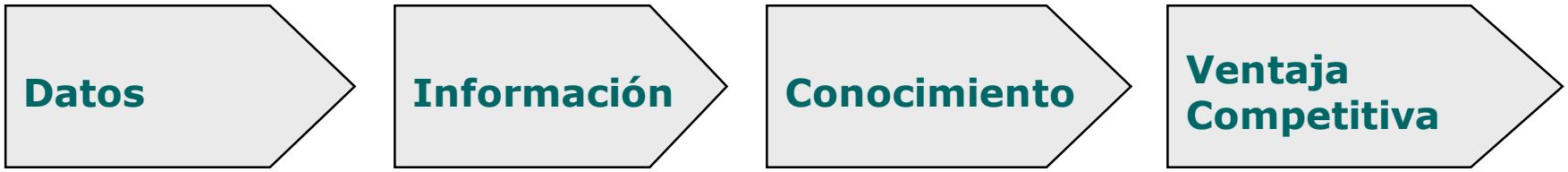
Inteligencia de Negocio

" Inteligencia de Negocio se refiere al proceso de convertir datos en conocimiento y conocimiento en acciones para crear la ventaja competitiva del negocio"

The Data Warehousing Institute



La analítica de negocios es el subconjunto de herramientas de inteligencia de negocio basado en la predicción, extracción de conocimiento y optimización

Inteligencia de Negocio

La analítica de negocios es el subconjunto de herramientas de inteligencia de negocio basado en la predicción, extracción de conocimiento y optimización.

Tema 2:
Minería de Datos.
Ciencia de Datos

INTELIGENCIA DE NEGOCIO

2019 - 2020

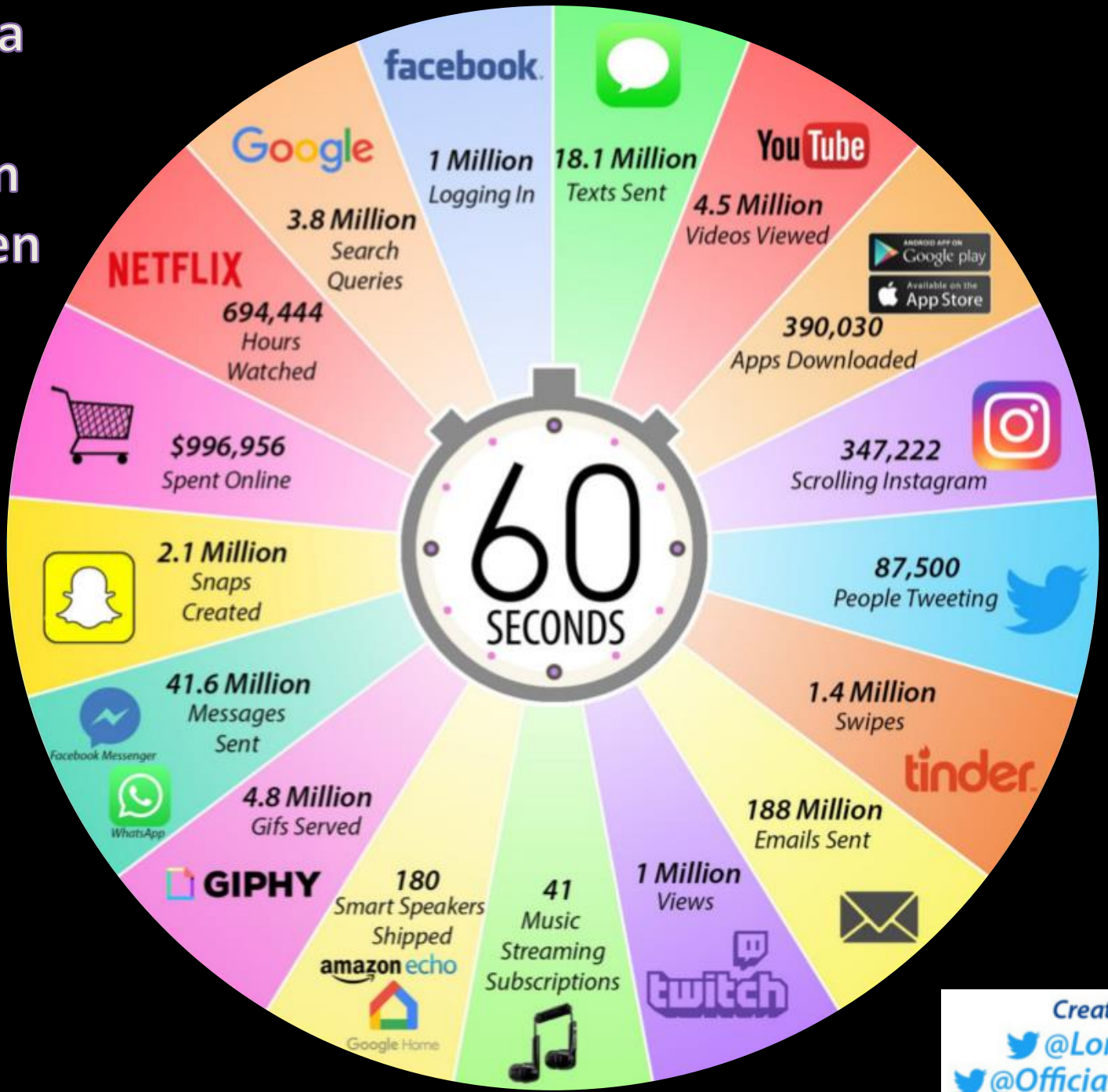
- **Tema 1. Introducción a la Inteligencia de Negocio**
- **Tema 2. Minería de Datos. Ciencia de Datos**
- **Tema 3. Modelos de Predicción: Clasificación, regresión y series temporales**
- **Tema 4. Preparación de Datos**
- **Tema 5. Modelos de Agrupamiento o Segmentación**
- **Tema 6. Modelos de Asociación**
- **Tema 7. Modelos Avanzados de Minería de Datos.**
- **Tema 8. Big Data**

Objetivos:

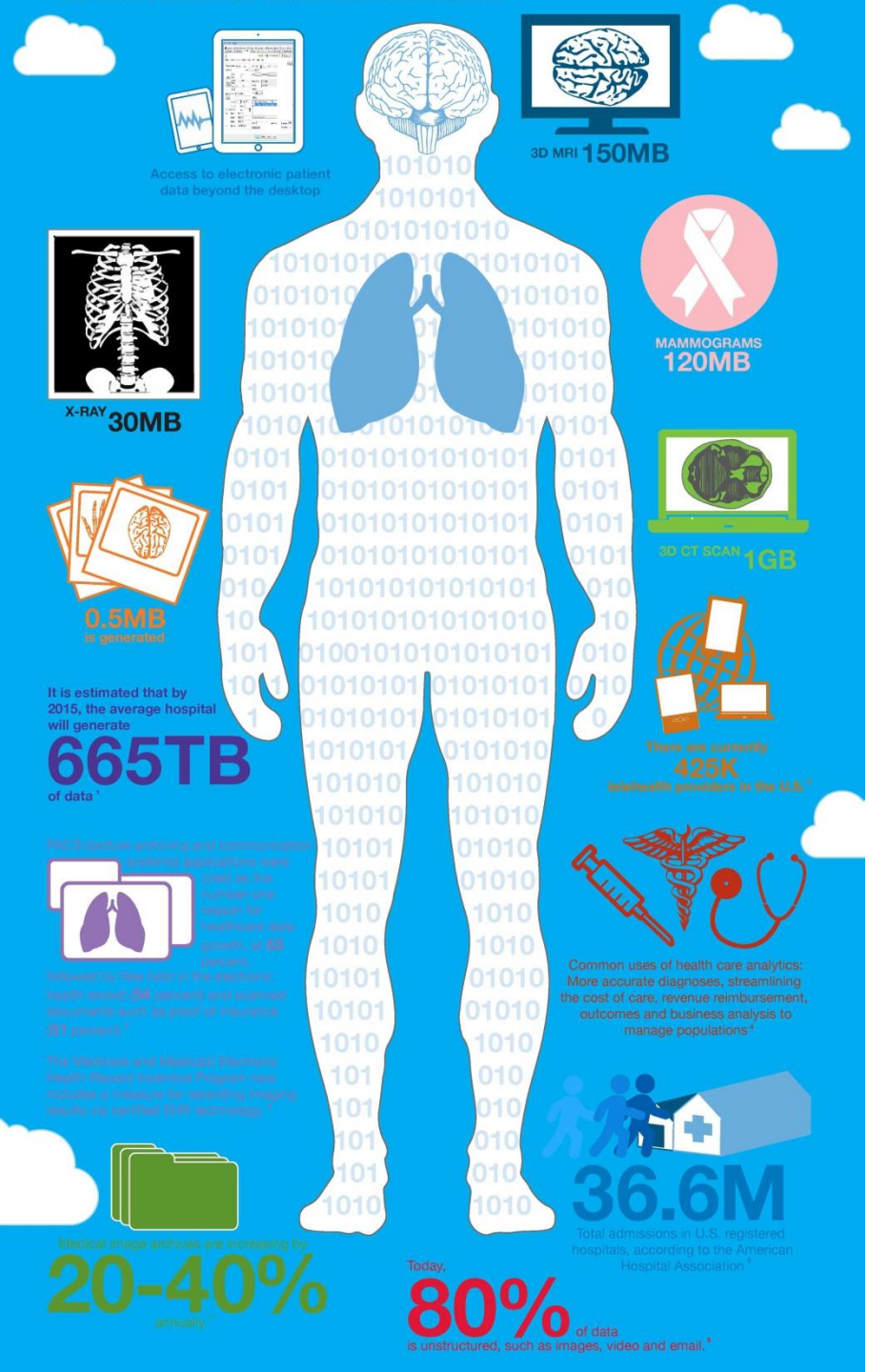
- Introducir los conceptos de Ciencia de Datos, Minería de Datos, Big Data
- Conocer las etapas del proceso de minería de datos
- Conocer los problemas clásicos de minería de datos

Cada **30 minutos**, la cantidad de datos creados crece **50 petabytes**, equivalente a todos los trabajos escritos en la **historia de la humanidad** en todas sus lenguas

¿Qué pasa en un minuto en internet en 2019?



Created By:
@LoriLewis
@OfficiallyChadd

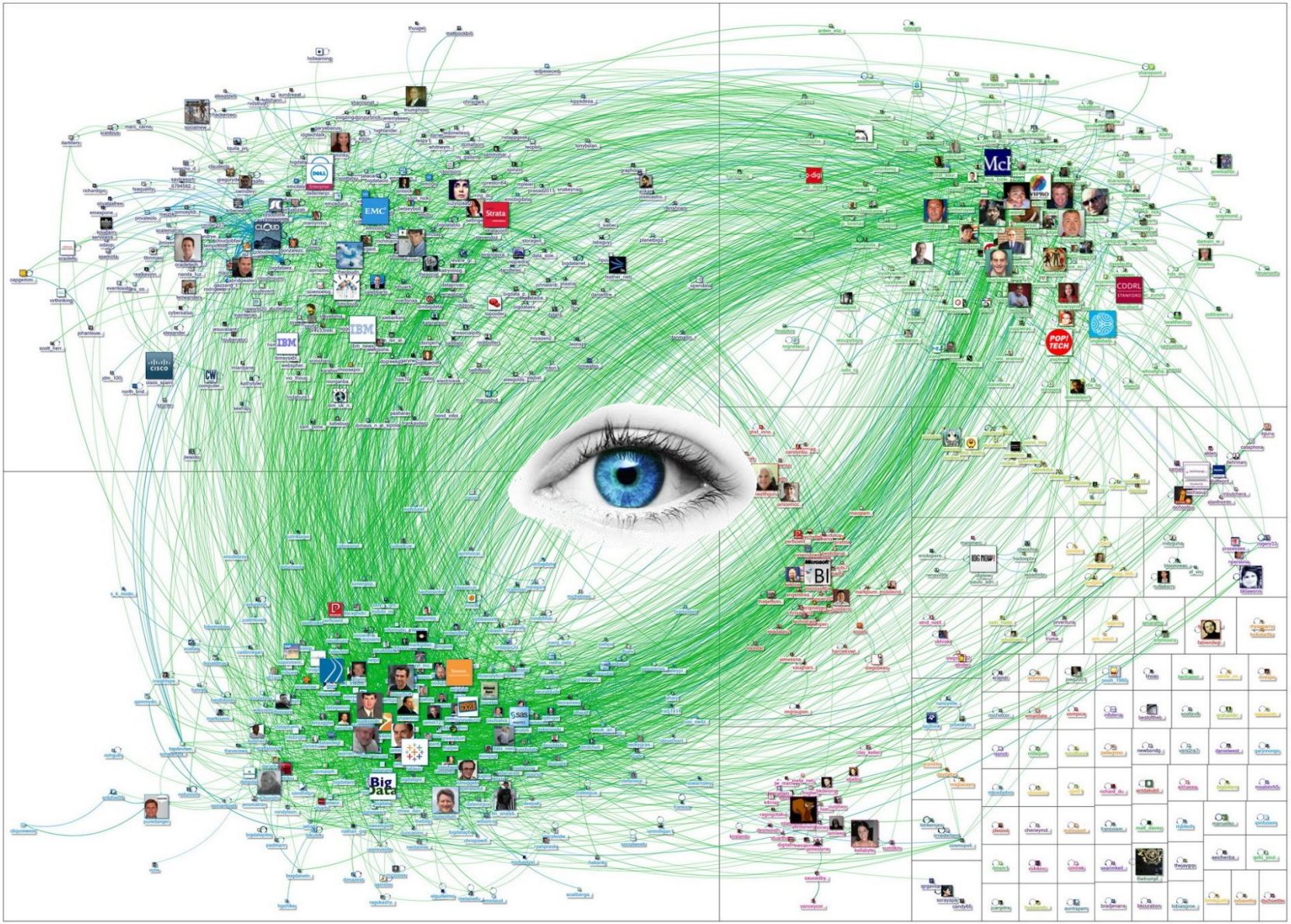


El cuerpo humano es una

FUENTE

ILIMITADA

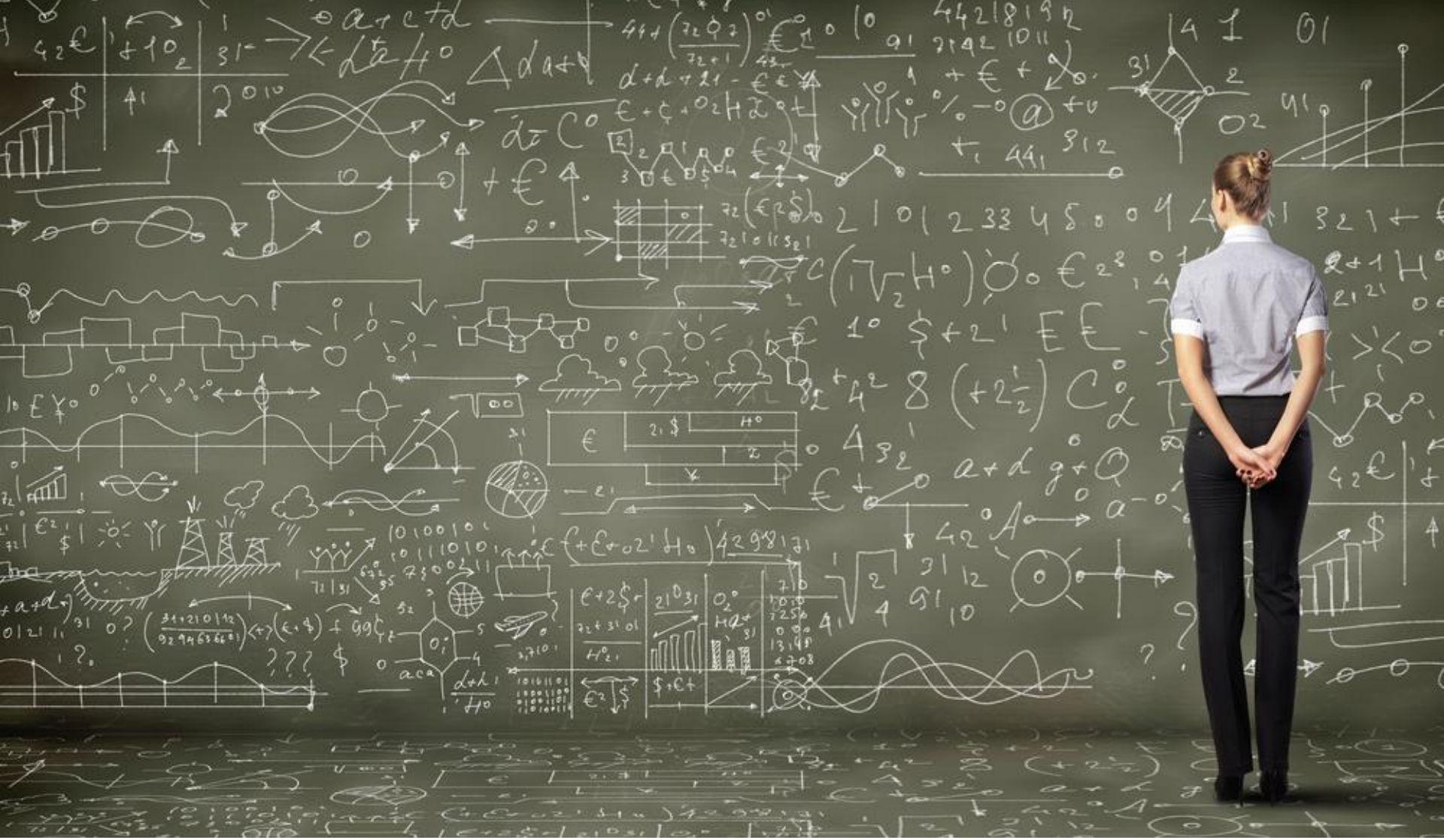
DE DATOS



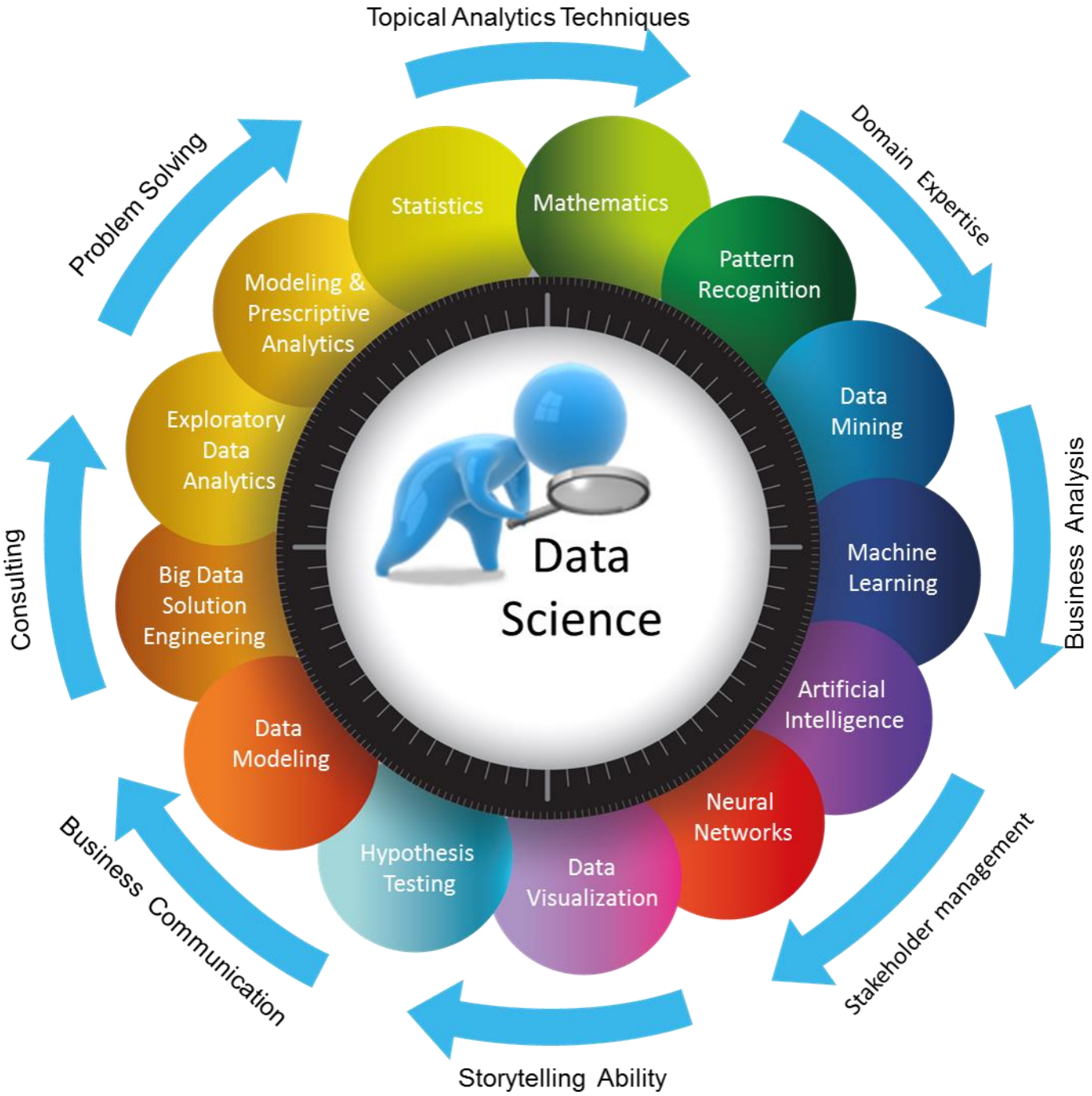
Discernir información relevante, **sintetizarla** y **extraer** conocimiento de ella es, cada vez, un aspecto más crítico en la sociedad en que vivimos


BIG
DATA
USE
ONE
SINCE
SOFTWARE
USING
TYPES
WORKING
GARTNER
MASSIVELY
BIOLOGICAL
PERFORMANCE
ALSO
RELATED
HUNDREDS
CREATED
DISK
RELATIONAL
GENOMICS
ZETTABYTES
COMPUTING
TOOLS
SET
SHARED
TIME
WITHIN
PROCESS
COMBAT
SIGNIFICANT
INCLUDE
SYSTEMS
DEFINITION
NETWORKS
BIOGEOCHEMICAL
SEARCH
RECONSIDER
OPPORTUNITIES
INFORMATION
CONNECTOMICS
RECORDS
COST
CONTINUES
TERABYTES
CITATION
COMPLEX
MPP
RESEARCH
DATABASES
STORAGE
INDEXING
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
SENSOR
ARCHIVES
AMOUNT
DESCRIBING
APPLIED
ELAPSED
CURRENT
THOUGHT
DESKTOP
CURRENTLY
FC
WORLD'S
TENS
CAPACITY
FORMS
PRESENTATIONS
PRACTITIONERS
NOW
ANALYTICS
ORGANIZATIONS
REQUIRING
BUSINESS
MOVING
UBIQUITOUS
SAN
RADIO-FREQUENCY
COMPLEXITY
SOLID
WIRELESS
TOLERABLE
PARALLEL
SIZE
NEEDED
QUALITIES
PETABYTES
BURIED
PROCESSING
INTERNET
LOGS
TECHNOLOGIES
USED
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
STORE
SOCIAL
LARGE
DEFINING
CASE
EVERY
LARGER

Se emplean grandes computadores para procesar esta información y extraer conocimiento útil (modelos)



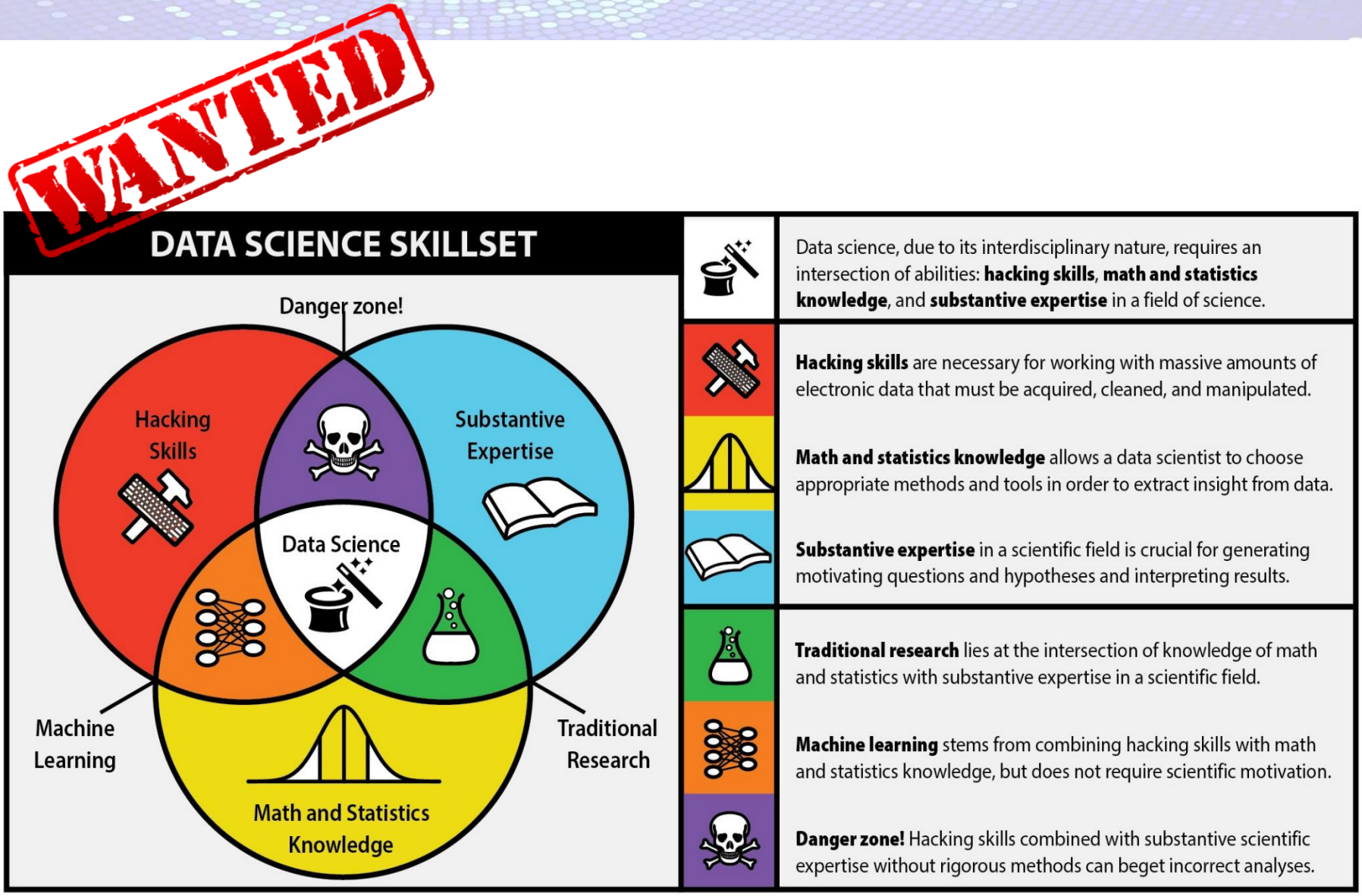
Los **modelos** representan el conocimiento extraído de los datos. Se pueden usar para **entender** una realidad compleja, **simular** condiciones, **detectar** fallos, **controlar** procesos, etc.


Topical Analytics Techniques
Domain Expertise
Business Analysis
Stakeholder management
Storytelling Ability
Business Communication
Consulting
Problem Solving
Statistics
Mathematics
Pattern Recognition
Data Mining
Machine Learning
Artificial Intelligence
Neural Networks
Data Visualization
Hypothesis Testing
Data Modeling
Big Data Solution Engineering
Exploratory Data Analytics
Modeling & Prescriptive Analytics
Data Science

Aplicaciones

- **Empresa/banca/seguros**: segmentación de clientes, **patrones de compra**, marketing digital, estudio de concesión de créditos, detección de **comportamiento fraudulento**, identificación de fuga de clientes…
- **Industria**: detección de productos defectuosos, identificación de **causas de fallos**, control del procesos, optimización adaptativa del rendimiento, mantenimiento predictivo, **eficiencia energética**…
- **Salud**: apoyo al **diagnóstico médico**, identificación de terapias para diferentes enfermedades, estudio de **factores de riesgo**, segmentación de pacientes en grupos afines, **gestión hospitalaria** y planificación temporal de salas, recomendación priorizada de fármacos, estudios en genética, selección de embriones en reproducción artificial…
- **Investigación científica**: medicina, biología, **astronomía**, geografía, **genética**, bioquímica, meteorología, ciencias sociales, ingenierías…

Científico de Datos



elEconomista.es
Tecnología
Jueves, 26 de Mayo de 2016 Actualizado a las 10:30

Telefónica nombra al 'hacker' Chema Alonso como nuevo responsable de Big Data e Innovación

EFE 26/05/2016 - 10:30 4 Comentarios

- *El experto informático se encargaba del departamento de ciberseguridad*

Más noticias sobre: TELEFÓNICA INTERNET INNOVACIÓN CÉSAR ALIERTA

Minería de Datos

Minería de Datos (*Data Mining*)

- La minería de datos (MD) es el proceso de **extracción de patrones** de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos

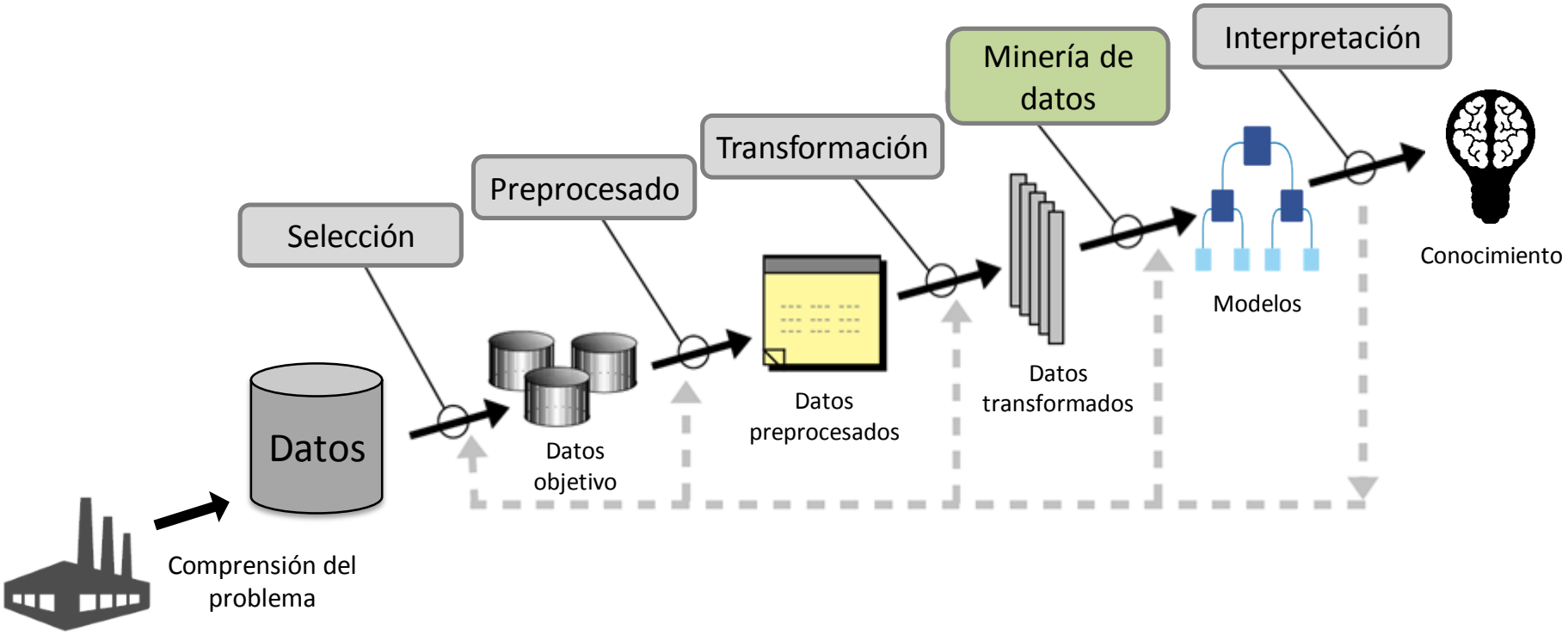
- También se conoce como:
  - Descubrimiento de conocimiento en bases de datos (KDD),
  - extracción del conocimiento,
  - análisis inteligente de datos/patrones,
  - …

Minería de Datos

- ¿Para qué se utiliza el 'conocimiento' obtenido?
  - hacer **predicciones** sobre nuevos datos
  - **explicar** los datos existentes
  - **resumir** una base de datos masiva para facilitar la toma de decisiones
  - **visualizar** datos altamente dimensionales, extrayendo estructura local simplificada…
- ¿A qué tipos de datos puede aplicarse las técnicas de Minería de Datos?
  - Bases de datos **relacionales, espaciales, temporales, documentales, multimedia**
  - World Wide **Web** (web mining)
  - Grandes volúmenes de datos: Big Data, Social Big Data

KDD

- KDD (*Knowledge Discovery from Databases*) es el **proceso completo** de extracción de conocimiento a partir de bases de datos
- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La Minería de Datos es sólo una etapa en el proceso de KDD

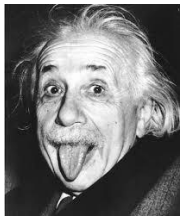


INTELIGENCIA ARTIFICAL

Algoritmo: conjunto de instrucciones bien definidas, ordenadas y finitas que permite realizar una actividad mediante pasos sucesivos

Aprendizaje Automático (*Machine Learning*)

Técnicas de Aprendizaje Automático (*Machine Learning*)



	Supervisado	No supervisado
Categórico	Clasificación	Reglas de asociación
Continuo	Regresión	Clustering

Aprendizaje Supervisado para Clasificación

etiqueta de clase

comestible

color, forma, tamaño…
en sombrero, tallo…

características / atributos

blanca, redonda, gruesa…

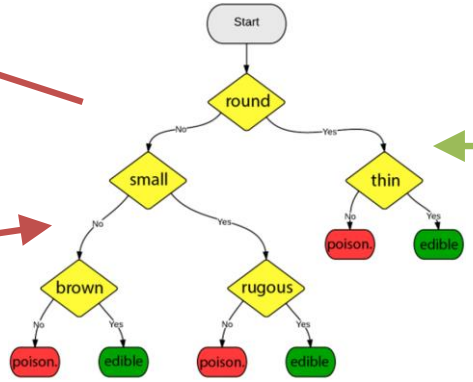
datos

caract. 1	comestible
caract. 2	comestible
caract. 3	venenosa
caract. 4	comestible
…	…

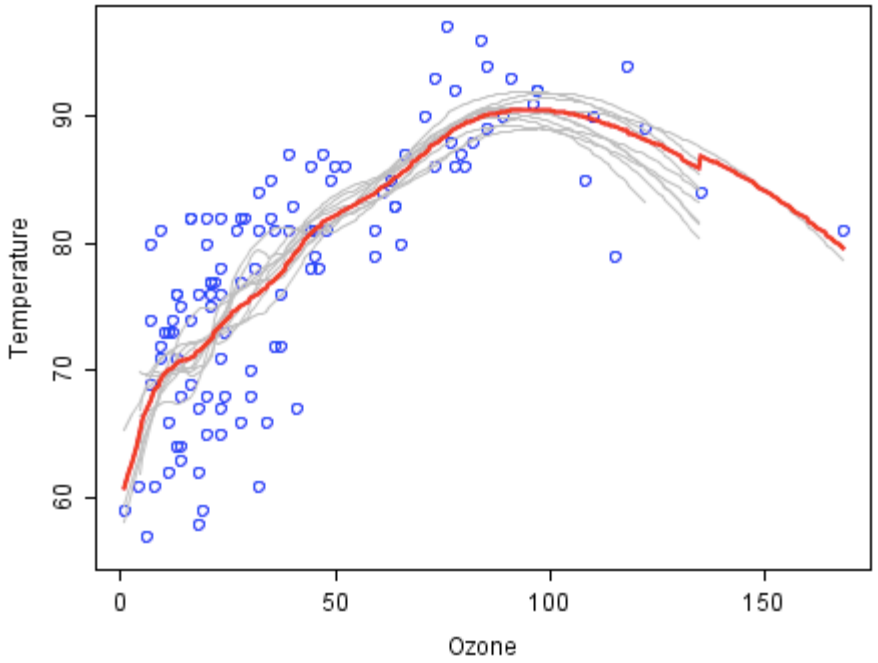
aprendizaje automático

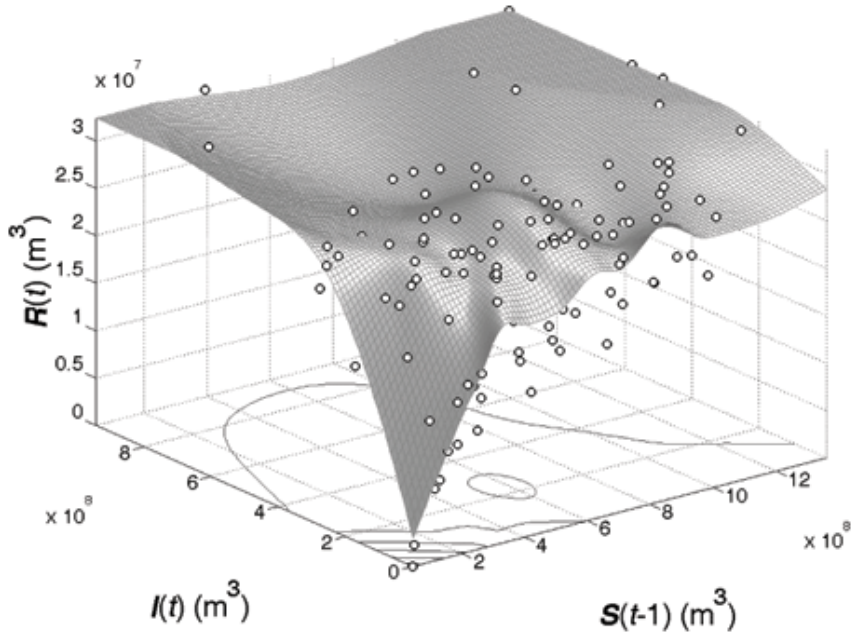
venenosa

?

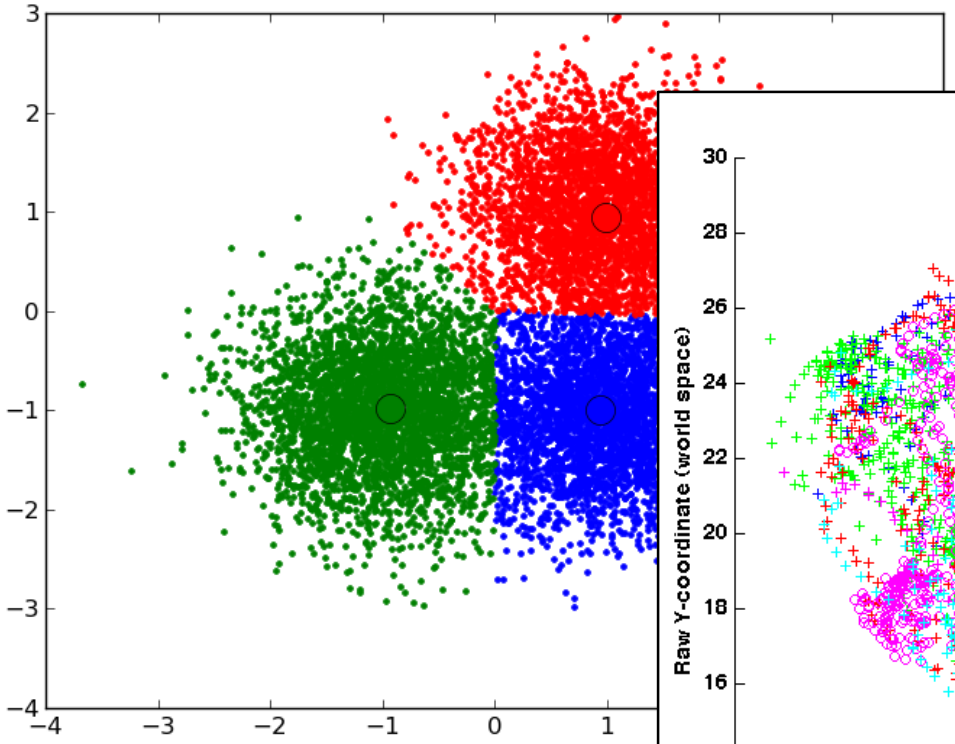


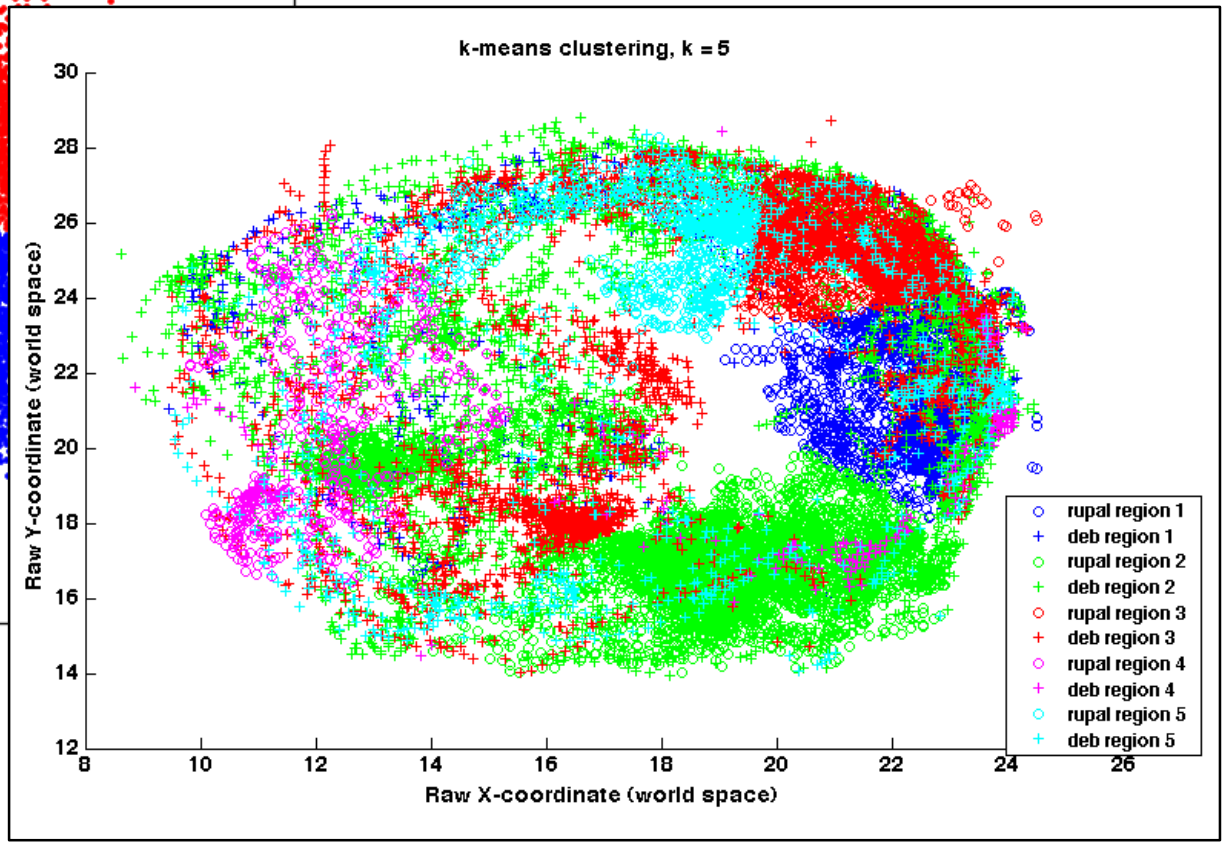
Aprendizaje Supervisado para Regresión





Aprendizaje no Supervisado para Segmentación (Agrupamiento o Clustering)



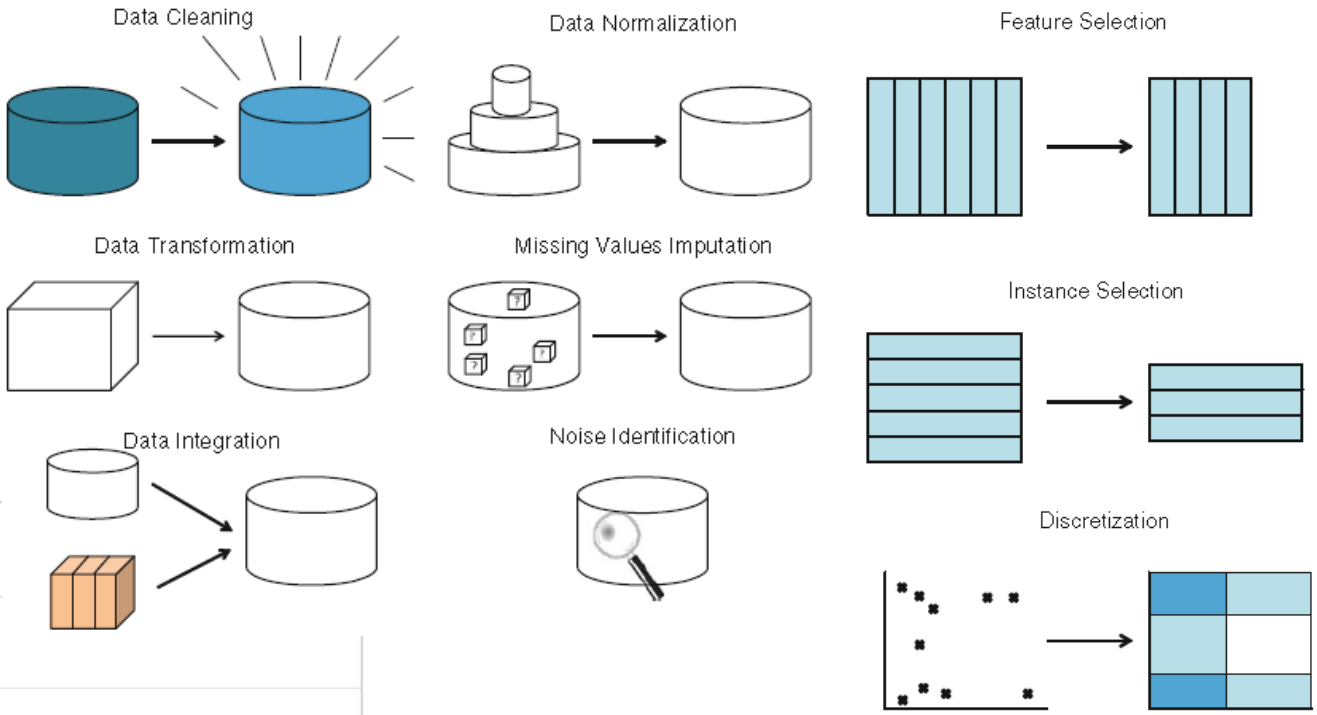


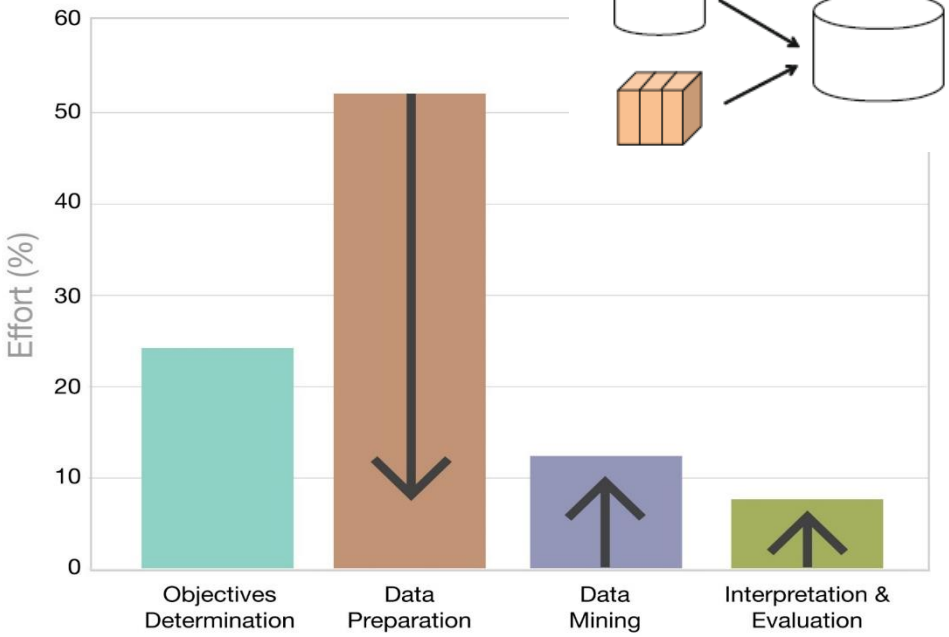
Aprendizaje no Supervisado para Asociación (Reglas de Asociación)

	Confianza
SI compra **cerveza**, ENTONCES compra **pañales**	50%
SI compra **pañales**, ENTONCES compra **cerveza**	67% ✓

Preprocesamiento: preparación y reducción de datos

Datos de calidad para decisiones de calidad



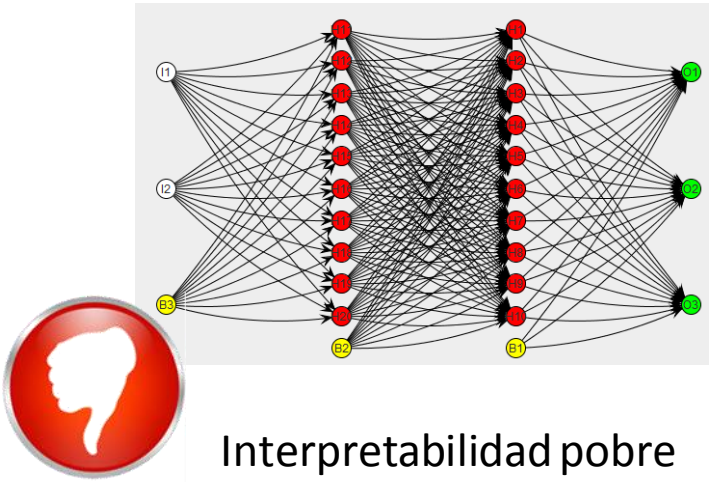


El esfuerzo de preprocesamiento suele suponer entre el 50% y el 80% del total

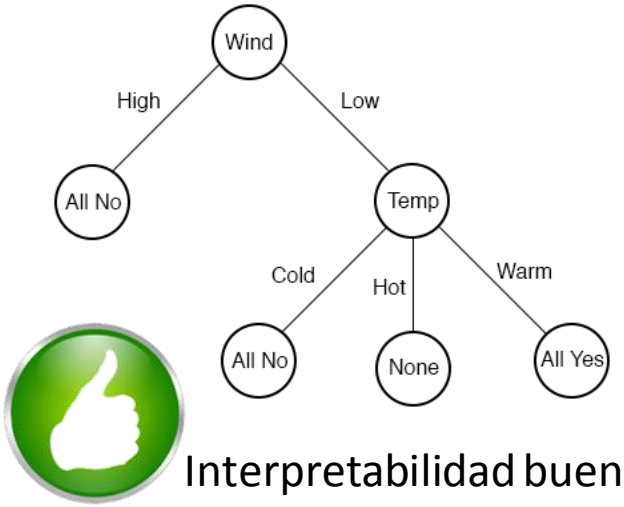
Modelizado

Interpretabilidad vs. Precisión

- Una de las utilidades del modelizado predictivo es comprender la lógica que lleva a realizar una predicción → para ello, obtener modelos legibles es crucial



Interpretabilidad pobre



Interpretabilidad buena

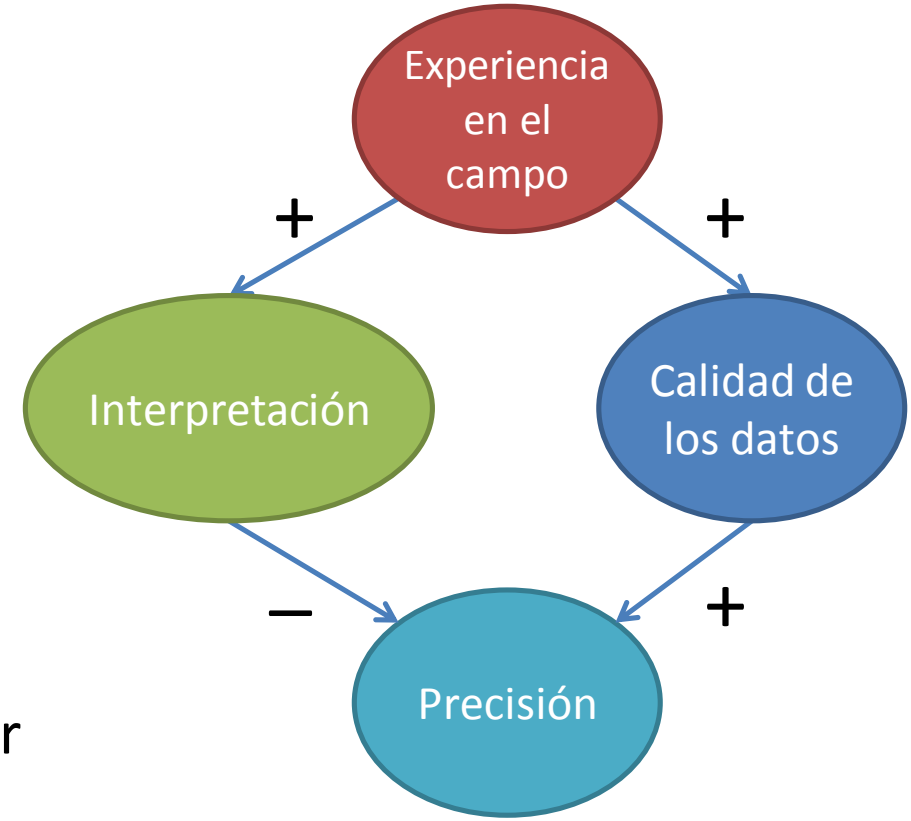
- Sin embargo, cuando se trata de aplicabilidad del modelo para justificar el retorno de la inversión (ROI), es esencial conseguir mayor precisión en la predicción

Interpretabilidad vs. Precisión

- Como es de esperar, la precisión y la interpretación no siempre van de la mano. En la práctica, se debe buscar un equilibrio y entran en juego otras consideraciones a la hora de seleccionar el mejor modelo

Interpretabilidad vs. Precisión

- A medida que aumenta la **interpretabilidad**, se reduce la exactitud del modelo
- A medida que aumenta la **calidad de los datos**, aumenta la precisión
- Cuanto mejor sea la **experiencia en el campo**, mejor será la calidad de los datos y la interpretación



https://jbsimha.wordpress.com/2011/06/24/model-performance-%E2%80%93-accuracy-vs-interpretability/

Big Data

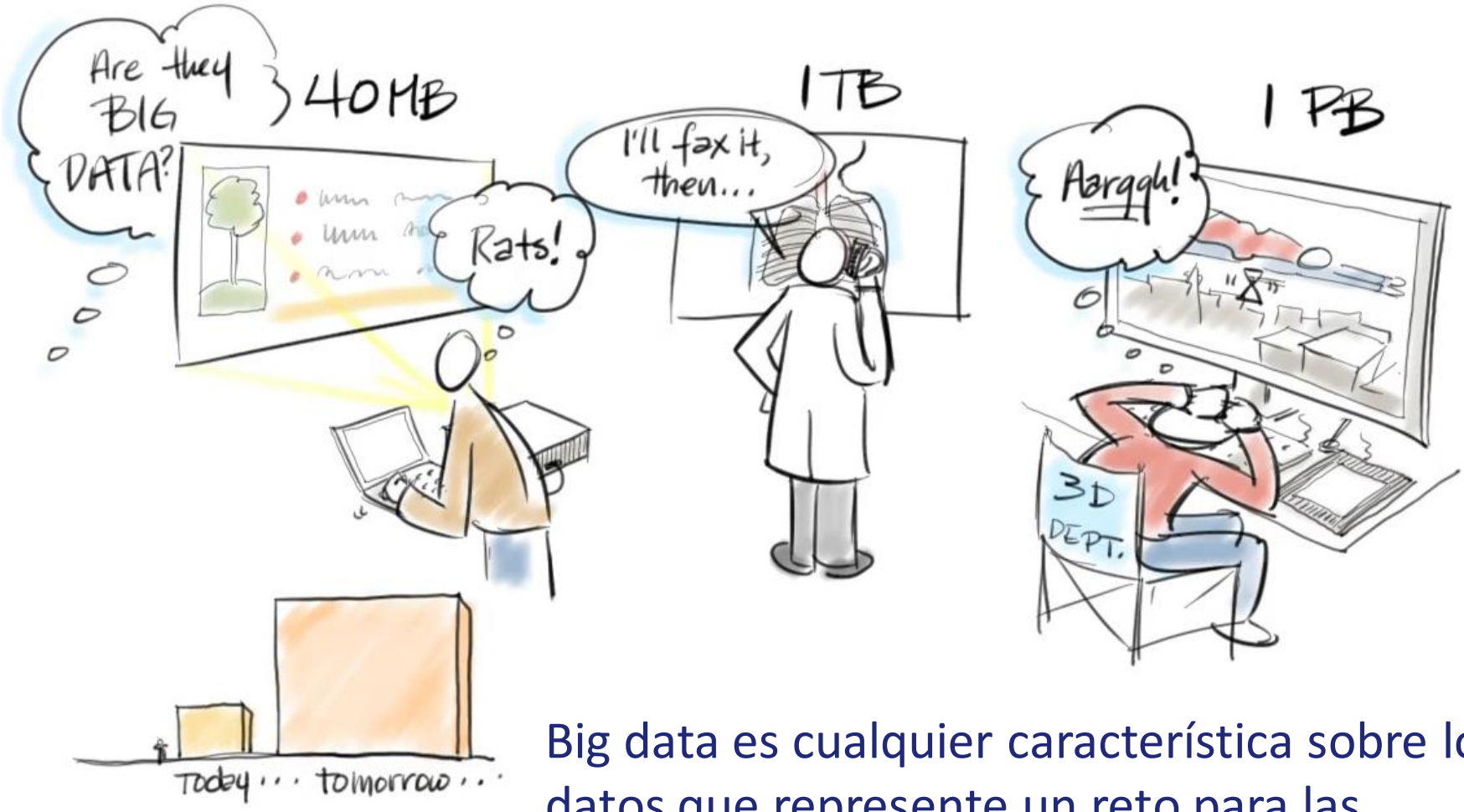
Dan Ariely

January 6, 2013 at 6:17pm ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Like | Comment | Share

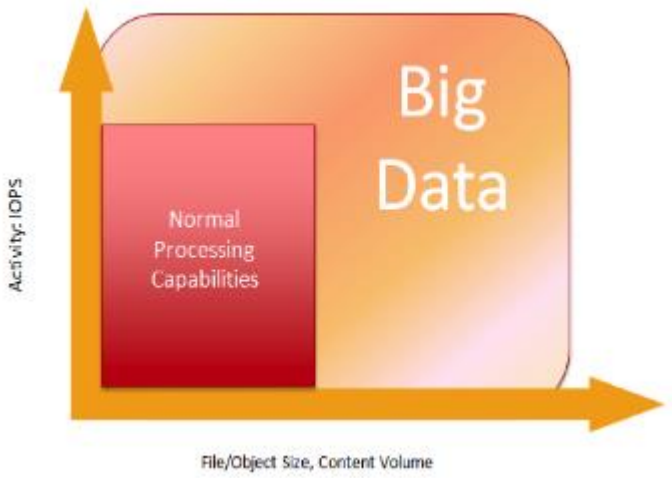
Big Data

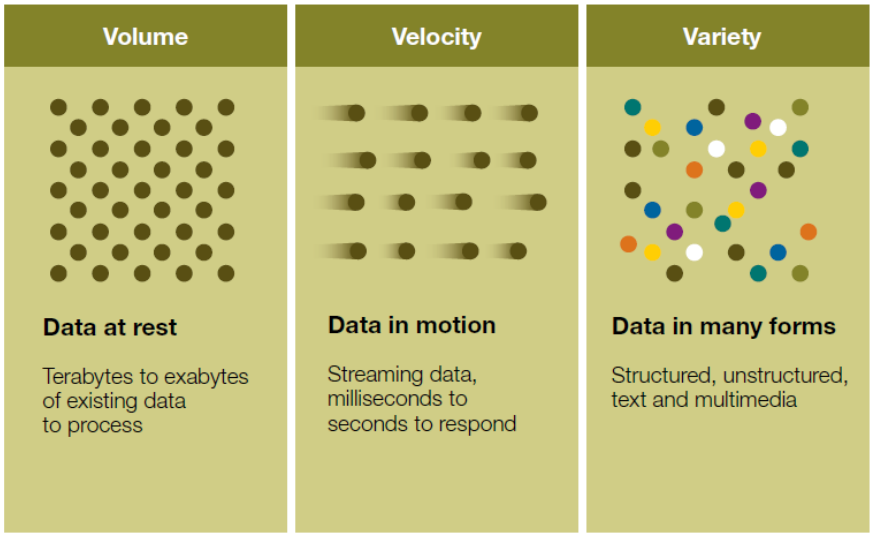


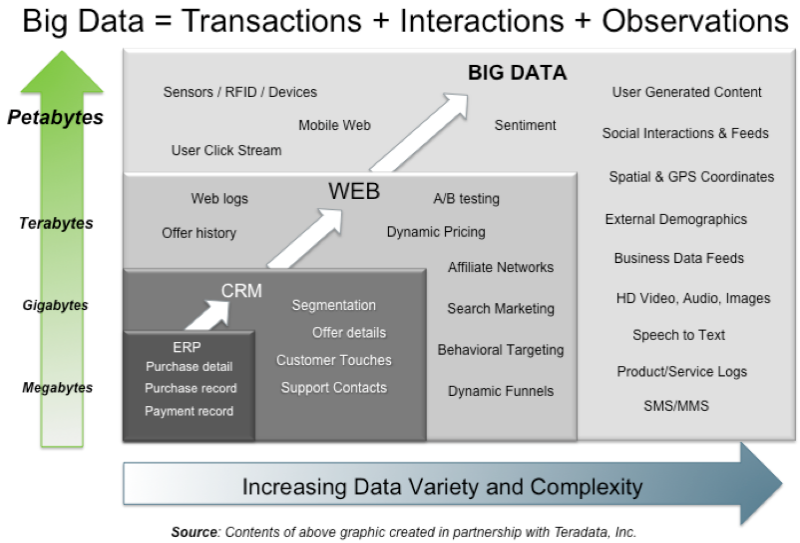
Big data es cualquier característica sobre los datos que represente un reto para las funcionalidades de un sistema

Big Data

- "*Big Data*" son datos cuyo volumen, diversidad y complejidad requieren nuevas arquitecturas, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos ...





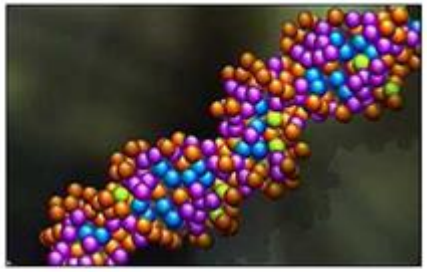


Source: *Contents of above graphic created in partnership with Teradata, Inc.*

Las tres 'V' del Big Data

Volumen

Ej. Genómica



- 25,000 genes in human genome
- 3 billion bases
- 3 Gigabytes of genetic data

Ej. Astronomía

- Astronomical sky surveys
- 120 Gigabytes/week
- 6.5 Terabytes/year

Ej. Transacciones de tarjetas de crédito

- 47.5 billion transactions in 2005 worldwide
- 115 Terabytes of data transmitted to VisaNet data processing center in 2004

Las tres 'V' del Big Data

Volumen

Velocidad

e PROMOTIONS

Ej. E-Promociones: Basadas en la posición actual e historial de compra ➔ envío de promociones en el momento de comercios cercanos a la posición

Las tres ‘V’ del Big Data

Ej. Huella digital de pasajeros

Volumen

Interior encarga un megacerebro capaz de localizar terroristas entre los pasajeros

¿Un proyecto de ciencia ficción? El Ministerio del Interior cree que no. Que es posible tener un megacerebro electrónico capaz de localizar —a través de cálculos estadísticos y de cruzar ingentes cantidades de datos en milisengundos—...

"identificación automática del perfil demográfico y sociológico del pasajero"

Velocidad

Variedad

http://article.wn.com/view/2015/04/04/Interior_encarga_un_megacerebro_capaz_de_localizar_terrorist_3/

Application Delivery Strategies — META Group

Date: 6 February 2001 — File: 949

Author: Doug Laney

3D Data Management: Controlling Data Volume, Velocity, and Variety. Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.

META Trend: During 2001/02, leading enterprises will increasingly use a centralized data warehouse to define a common business vocabulary that improves internal and external collaboration. Through 2003/04, data quality and integration woes will be tempered by data profiling technologies (for generating metadata, consolidated schemas, and integration logic) and information logistics agents. By 2005/06, data, document, and knowledge management will coalesce, driven by schema-agnostic indexing strategies and portal maturity.

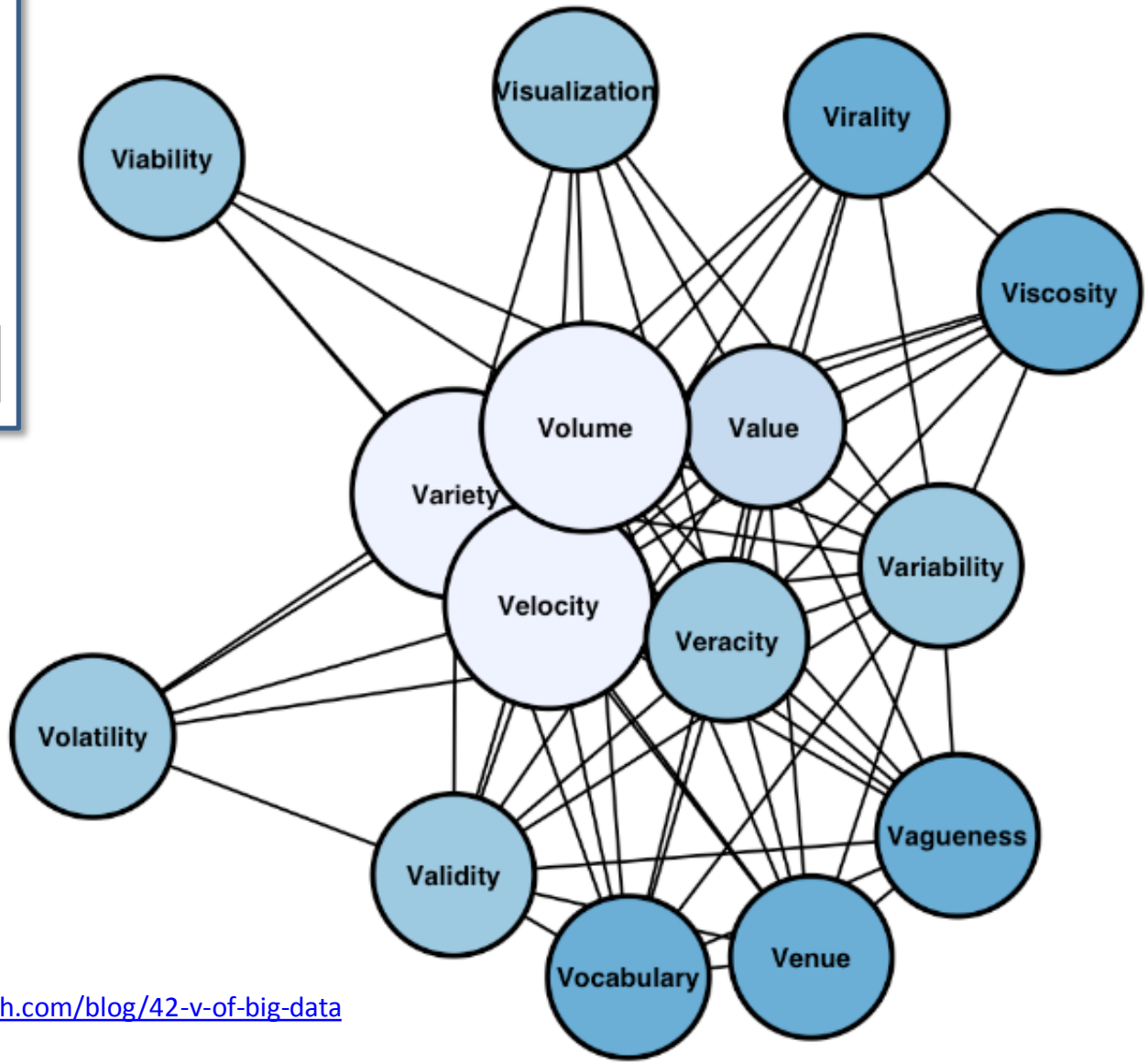
Business Impact

Attention to data management, particularly in a climate of e-commerce, and greater need for collaboration can enable enterprises to achieve greater returns on their information assets.

Bottom Line

In 2001/02, IT organizations must look beyond traditional direct brute-force physical approaches to data management. Through 2003/04, practices for resolving e-commerce accelerated data volume, velocity, and variety issues will become more formalized/diverse. Increasingly, these techniques involve tradeoffs and architectural solutions that involve/impact application portfolios and business strategy decisions.

http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf



First Occurrence

- 2001
- 2012
- 2013
- 2014

Fuente: https://www.elderresearch.com/blog/42-v-of-big-data

Gran Presencia en los Medios

EL MUNDO

España Opinión Internacional Eco

Inicio › Economía › INnovadores

El 27% de los ya son rentab

· Un estudio de Capgemini e Ir emprendiendo las empresas aún quedan muchos retos po

INNOVADORES

21/06/2016 11:37

Actualmente, más de la mitad (concretamente, el 55%) tiener el Big Data, ya sea de forma tra departamento individual. De el de Big Data son actualmente re se encuentra en posición de eq **datos de un estudio de Capger cómo un 12% restante de los p 12% considera que aún es prer**

EL PAÍS

'Big data', la nueva materi

La privacidad gana importancia en un mundo

Uno de los debates del foro. /SANTI BURGOS

que se quede embarazada?", espetó. No Un patrón de compra y búsqueda en la we

La historia, relatada por Charles Duhigg e (Urano, 2012), hace énfasis en uno de los mundo digital: el manejo de datos a gran ¿oportunidades sin límites?, organizado p Data Pop Alliance. Lo que desconocía el p desarrollado un sistema, que predice, cor embarazo de sus clientes en base a su his que pedir disculpas a la compañía.

elEconomista.es

Jueves, 21 de Enero de 2016 Actualizado a las 10:48

Big Data, un universo a tu servicio

ENLACES RELACIONADOS

Las 10 tendencias tecnológicas para el 2016: del Big Data a la realidad virtual (20/01)

El Big Data ha llegado a nuestras vidas para quedarse. Y ¿qué es eso del Big Data? Pues grosso modo y con poco tecnicismo, es la forma de aprovechar y sacar partido de forma inteligente a los exagerados volúmenes de información que se producen a diario.

Cada vez son más los organismos públicos, y privados, que lo están implantando para mejorar el servicio a la ciudadanía. En este post quiero referirme solo a dos de ellos para ver los progresos que es capaz de conseguir en nuestro día a día.

El primero de ellos viene de la mano del Cuerpo Nacional de Policía y de la Universidad de Granada, con el desarrollo de un sistema informático basado en algoritmos para predecir cuántos delitos y de qué tipo se van a producir en el próximo turno policial. Se trata de una aplicación de métodos científicos que combina métodos de policía predictiva con un modelo matemático de patrullaje. La implantación del sistema permitiría una organización de patrullas y turnos policiales mucho más eficiente, con lo que se evitaría un gran número de víctimas de delitos, además del ahorro económico.

El segundo, con un tema muy habitual en estos días de invierno, la temida gripe. ¿Quién de nosotros no ha realizado alguna vez una búsqueda en Google para saber cuáles son

Gran Presencia en los Medios

elEconomista.es

Martes, 21 de Junio de 2016 Actualizado a las 12:12

¿Qué se pierden los partidos por no utilizar el Big Data en sus campañas?

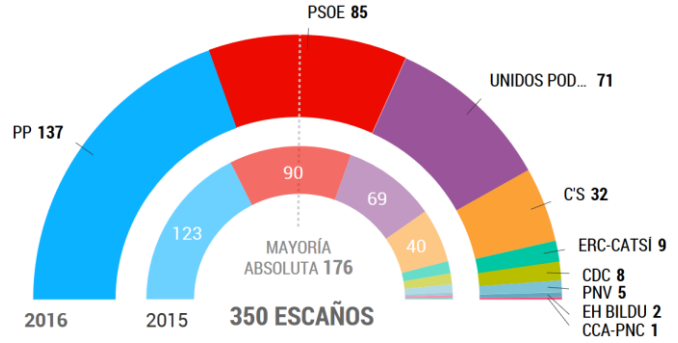
ENLACES RELACIONADOS

Big Data, un universo a tu servicio (21/01)

Nos enfrentamos a una jornada electoral caracterizada por un gran número de ciudadanos indecisos, en concreto, 1.357.997 de electores serán los responsables de decidir el resultado en las últimas 24 horas. El Big Data puede ayudar a los partidos políticos a rescatar adeptos de última hora y cambiar su intención de voto. La carrera ha comenzado.

Para Tinámica, compañía española especializada en soluciones tecnológicas del entorno Business Intelligence y Big Data, la clave está en conocer qué es exactamente lo que les haría apostar por uno u otro partido y, para ello, las técnicas de Big Data permiten analizar las demandas y elaborar la mejor fórmula y mensajes para acertar.

En paralelo a ganar indecisos, la compañía también ha identificado una serie de prácticas que los partidos políticos españoles están o deberían estar realizando para aprovechar esta tecnología y sus ventajas. De este modo, el Big Data está permitiendo o permitiría en esta campaña:



Impacto Económico

La demanda de profesionales formados en Ciencia de Datos y *Big Data* es enorme

España necesitará para 2015 más de 60.000 profesionales con formación en Ciencia de Datos y *Big Data*

Se estima que la conversión de datos en información útil generará un mercado de 132.000 millones de dólares en 2015 y que se crearán más de 4.4 millones de empleos

Impacto Económico

Las posiciones y las competencias más demandadas en el mercado laboral español

POSICIONES MÁS DEMANDADAS En 2015	Variación sobre 2014
1 'Big data' (tecnología)	14 ▲
2 Gestor de cuentas (comercial)	-1 ▼
3 Ingeniero industrial (ingeniería)	5 ▲
4 Gestor jefe de cuentas clave de la empresa (comercial)	9 ▲
5 Comerciales para nuevos mercados (comercial)	-2 ▼
6 Ingeniero informático (ingeniería)	24 ▲
7 Comercial digital (comercial)	14 ▲
8 Técnico comercial (comercial)	4 ▲
9 I+D (tecnología)	-7 ▼
10 Delegados de venta (comercial)	4 ▲

POSICIONES DIFÍCILES DE CUBRIR A dos o tres años. Frecuencia relativa	
1 'Big data' (tecnología)	5,12%
2 I+D (tecnología)	5,12%
3 Comerciales para nuevos mercados (comercial)	4,33%
4 Gestor de cuentas (comercial)	3,94%
5 Comercial digital (comercial)	3,15%
6 Comerciales de exportación (comercial)	3,15%
7 Ingeniero informático (ingeniería)	3,15%
8 Operadores de mantenimiento (operarios cualificados)	3,15%
9 Responsable de estrategia digital (marketing)	2,76%
10 Líder de proyecto (tecnología)	2,76%

Fuente: Informe Epyce 2015

ALEJANDRO MERAVIGLIA / CINCO DÍAS

CincoDías

MANUEL G. PASCUAL - MADRID 27-04-2016 08:41

PROFESIONES DE FUTURO

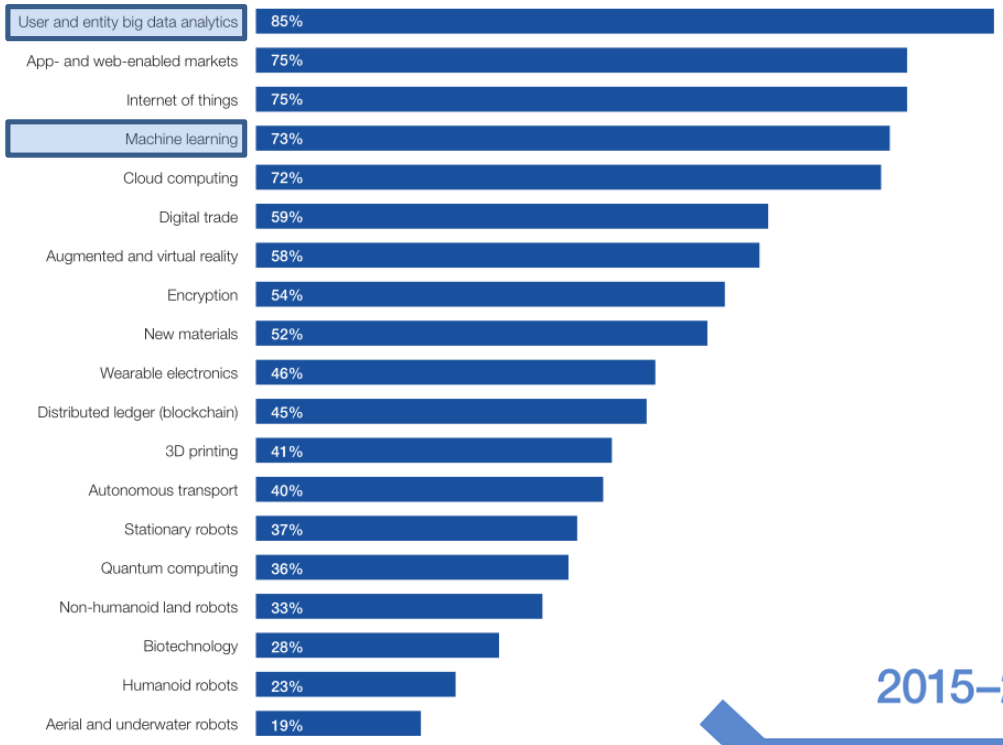
El 'big data' asalta el mercado laboral

Los directores de recursos humanos apuestan fuerte por las profesiones técnicas

COMPETENCIAS MÁS DEMANDADAS En 2015	Variación sobre 2014
1 Compromiso	0 ■
2 Resilencia (inteligencia emocional)	5 ▲
3 Visión estratégica y orientación estratégica	3 ▲
4 Liderazgo	-2 ▼
5 Idiomas y habilidades lingüísticas	-2 ▼
6 Gestión del cambio y adaptabilidad	-2 ▼
7 Iniciativa y proactividad	1 ▲
8 Innovación	2 ▲
9 Flexibilidad	-4 ▼
10 Orientación a resultados	-1 ▲

Impacto Económico

Figure 2: Technologies by proportion of companies likely to adopt them by 2022 (projected)



WORLD ECONOMIC FORUM

The Future of Jobs

Emerging in-demand roles: Among the range of established roles that are set to experience increasing demand in the period up to 2022 are **Data Analysts and Scientists,** Software and Applications Developers, and Ecommerce and Social Media Specialists, roles that are significantly based on and enhanced by the use of technology. […] Moreover, our analysis finds extensive evidence of accelerating demand for a variety of wholly new specialist roles related to understanding and leveraging the latest emerging technologies: **AI and Machine Learning Specialists**, **Big Data Specialists**, Process Automation Experts, Information Security Analysts, User Experience and Human-Machine Interaction Designers, Robotics Engineers, and Blockchain Specialists.

2015–2017

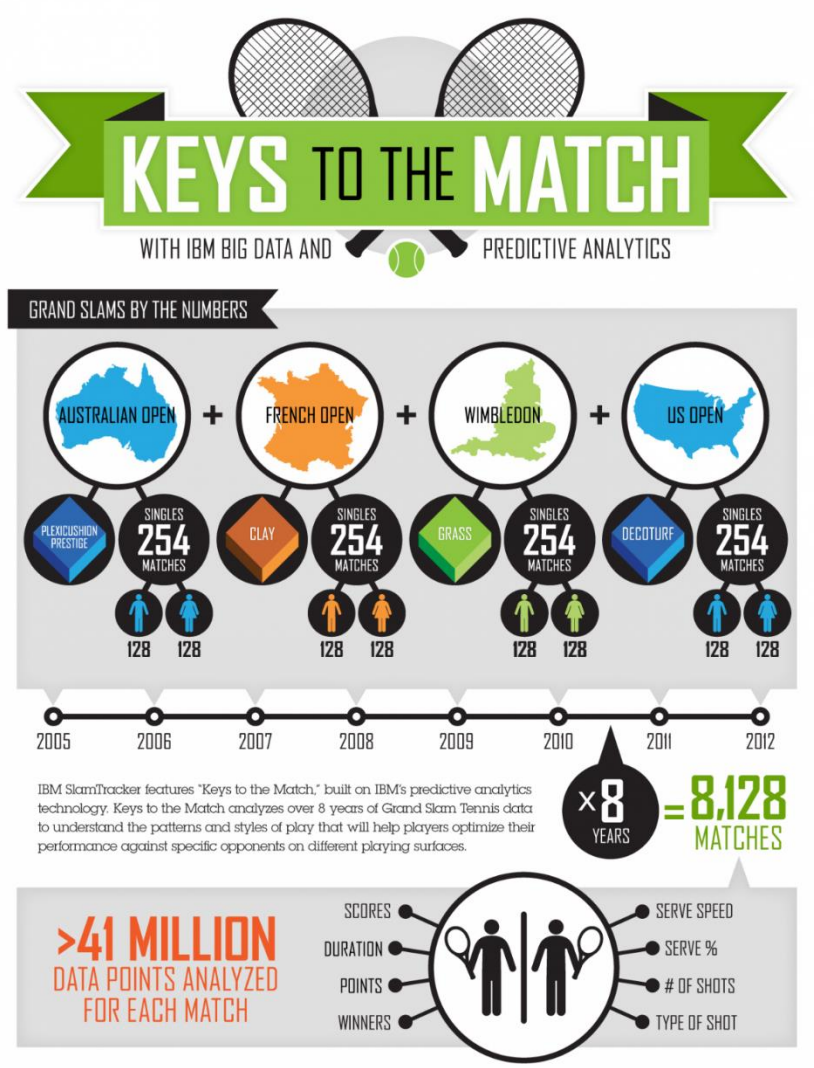
- New energy supplies and technologies
- The Internet of Things
- Advanced manufacturing and 3D printing
- Longevity and ageing societies
- New consumer concerns about ethical and privacy issues
- Women's rising aspirations and economic power

2018–2020

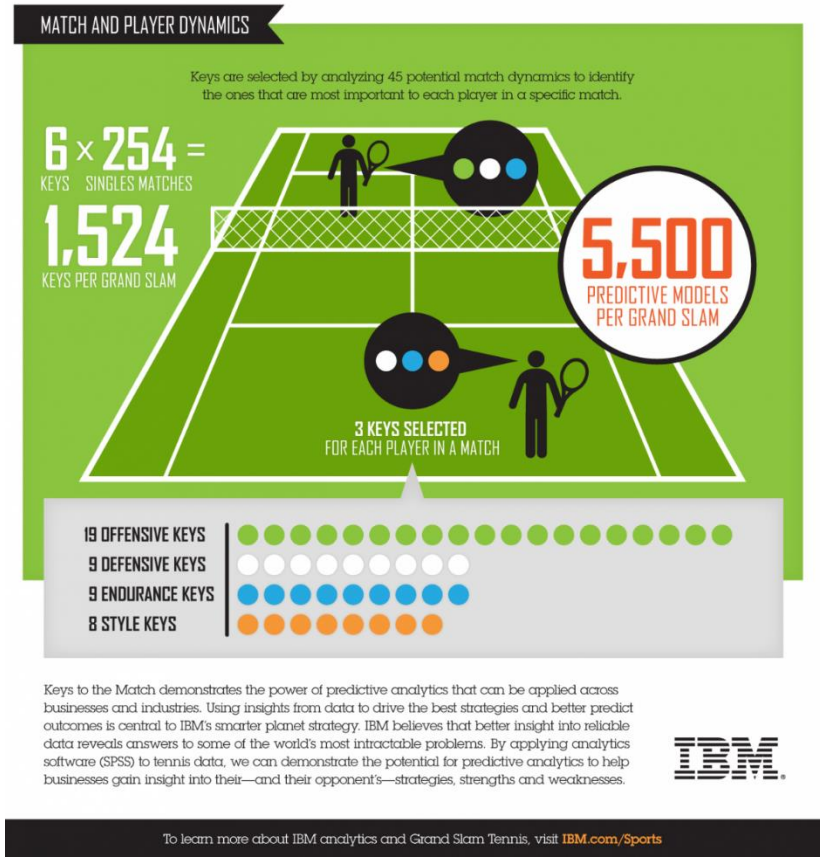
- Advanced robotics and autonomous transport
- Artificial intelligence and machine learning
- Advanced materials, biotechnology and genomics

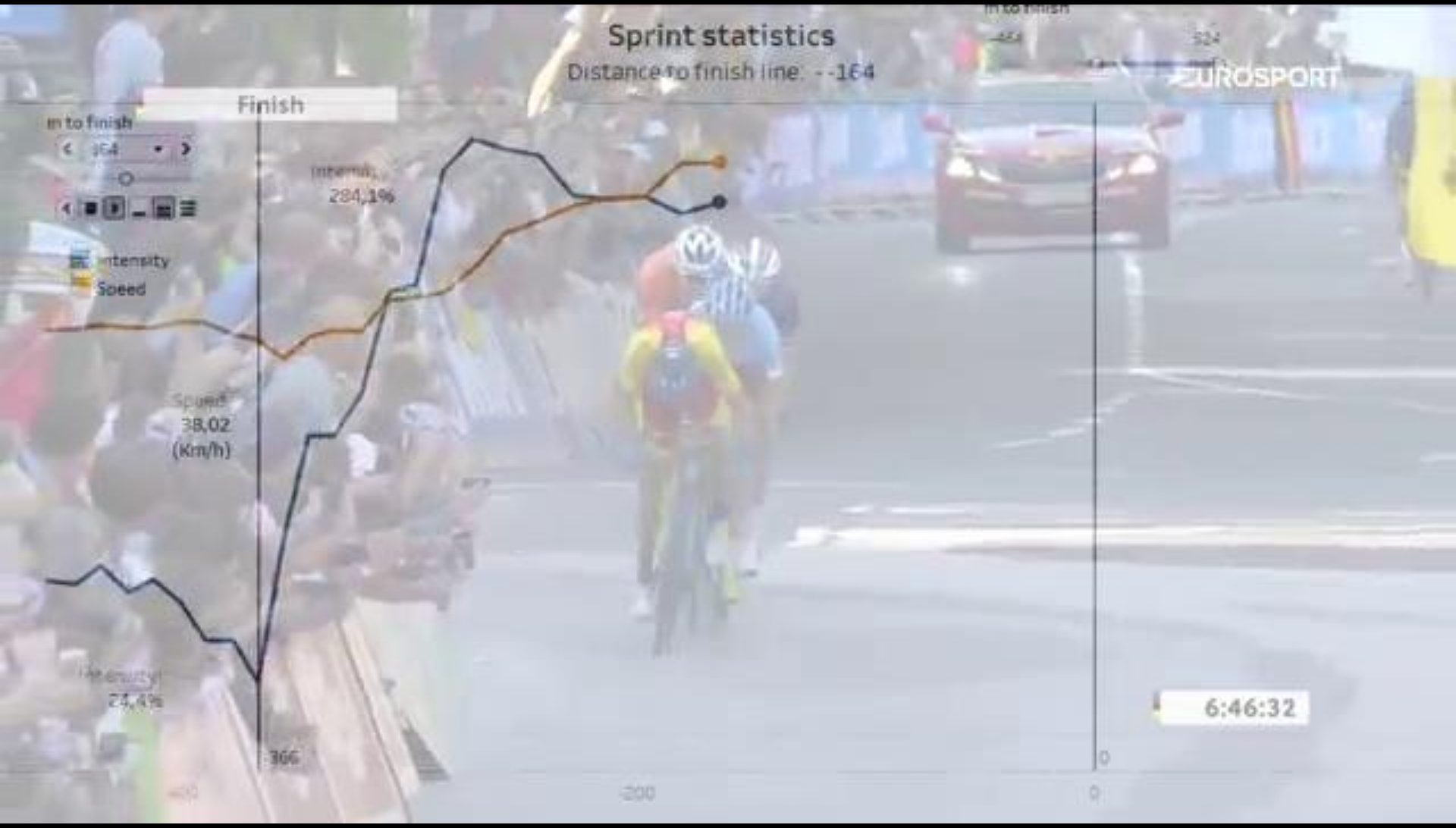
Casos de Uso

IBM SlamTracker



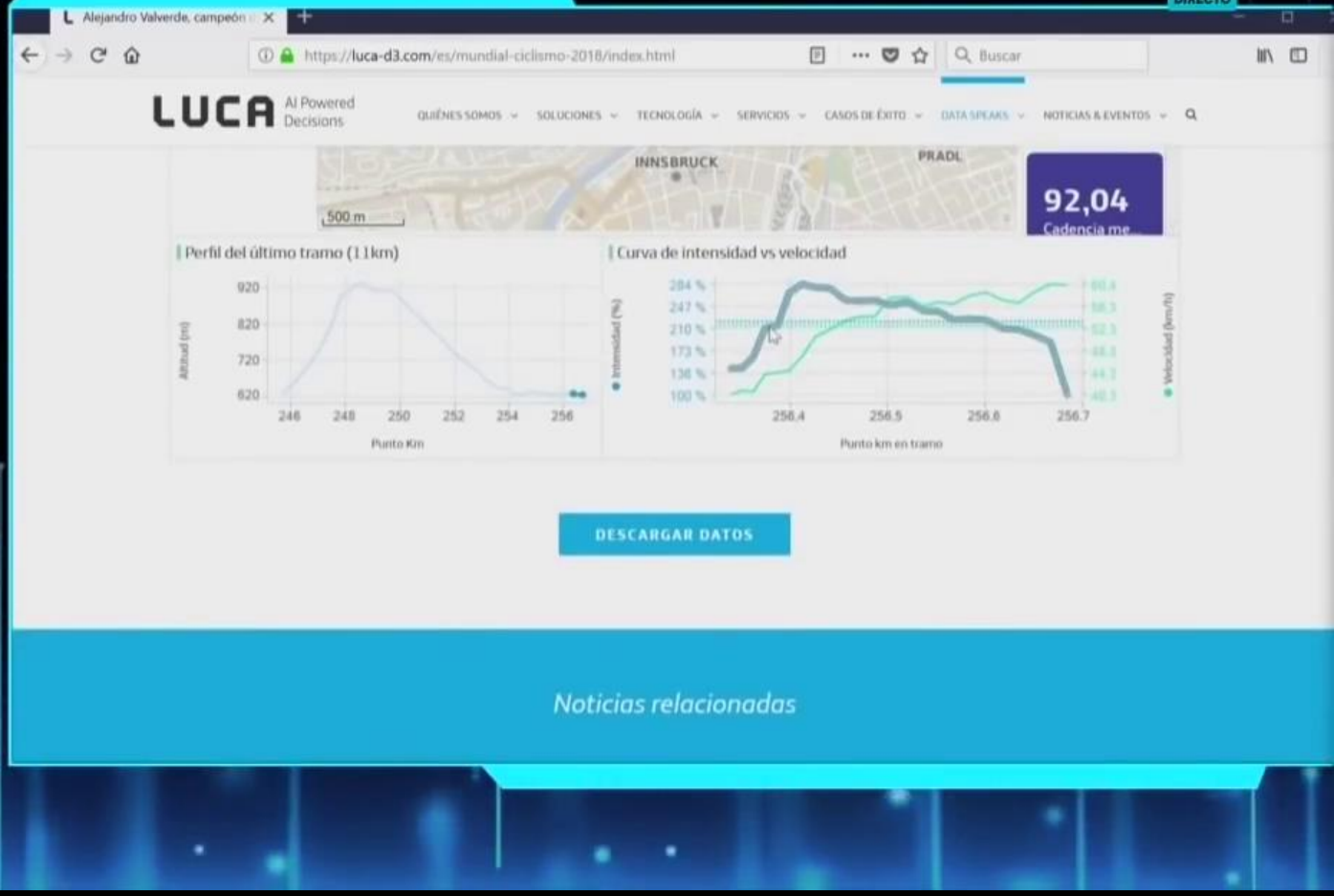
http://IBM.com/Sports





https://www.youtube.com/watch?v=ApJIH7dqhUo

#MovistarTeam2019



https://luca-d3.com/es/mundial-ciclismo-2018/

Google Flu Trends

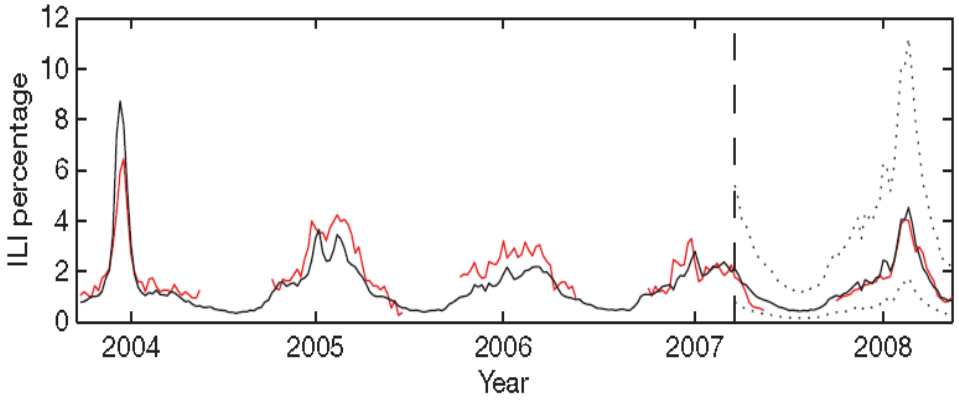
"Google puede predecir la propagación de la gripe analizando lo que la gente busca en internet"

Detecting influenza epidemics using search engine query data
Nature 475 (2009) 1012-1014

nature
International journal of science

- +30.000M de búsquedas
- 50M de términos de búsqueda más utilizados semanalmente
- Encontraron una combinación de 45 términos de búsqueda que combinados con un modelo matemático presentaba una correlación fuerte entre su predicción y las cifras oficiales de la enfermedad

- Podían decir, como los centros de control y prevención de enfermedades, a dónde se había propagado la gripe pero casi en **tiempo real**, no una o dos semanas después



Google Flu Trends

- En 2013 sobreestimó los niveles de gripe (dos veces CDC)
  - La sobreestimación puede deberse a la amplia cobertura mediática de la gripe que puede modificar comportamientos de búsqueda

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer[1,2,*], Ryan Kennedy[1,3,4], Gary King[3], Alessandro Vespignani[5,6,3]

Science 14 Mar 2014:
Vol. 343, Issue 6176, pp. 1203-1205
DOI: 10.1126/science.1248506

Science

- En 2014, Google cierra este proyecto (https://www.google.org/flutrends/about/)

- El uso de Wikipedia parece más estable y fiable

Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time

David J. McIver , John S. Brownstein, Harvard Medical School, Boston Children's Hospital, Plos Computational Biology, 2014.

Análisis de asociaciones

TARGET

Consumidora de Target ficticia: Jenny Ward, 23 años, vive en Atlanta. En marzo compró **loción** de manteca de cacao, un **bolso** grande para doblar, **suplementos** de zinc y magnesio y una **alfombra** azul brillante. Existe, digamos, un **87%** de probabilidad de que esté **embarazada** y que su fecha de parto sea en agosto

Modelo de predicción de clientes embarazadas por medio de sus patrones de compra

Descubrimiento: Cremas sin perfume al tercer mes. Con dos docenas de productos, predicción de fecha parto

Acción: Envío de cupones para cada fase del embarazo

El poder de los hábitos (2012) - Charles Duhigg

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did

Identificación de personas a partir de 4 compras con tarjetas de crédito

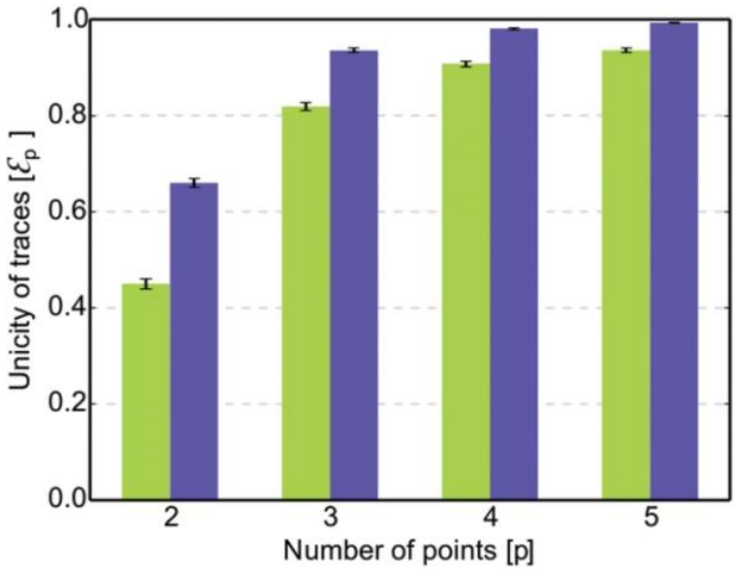
Unique in the shopping mall: On the reidentifiability of credit card metadata

Y.-A. de Montjoye, L. Radaelli, V. Kumar Singh, A. "Sandy" Pentland
Science 30 Jan 2015:
Vol. 347, Issue 6221, pp. 536-539
DOI: 10.1126/science.1256297

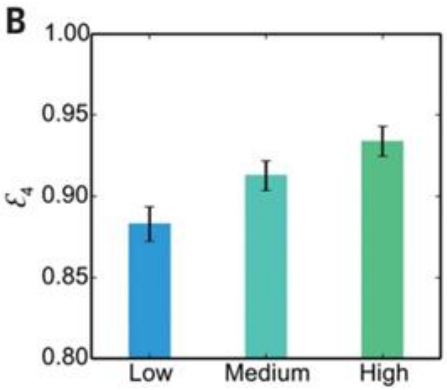
Science AAAS

http://www.sciencemag.org/content/347/6221/536

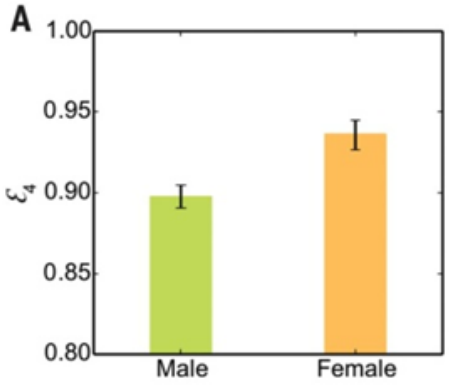
Los conjuntos de datos a gran escala del comportamiento humano tienen el potencial de transformar fundamentalmente la manera en que luchamos contra las enfermedades, diseñamos ciudades o realizamos investigaciones. Los meta-datos, sin embargo, contienen información sensible. **Comprender la privacidad de estos conjuntos de datos es clave** para su amplio uso y, en última instancia, su impacto.



Identificación por número de compras



Identificación por poder adquisitivo



Identificación por género

Herramientas

Herramientas, Lenguajes, Kaggle

- Software libre para Ciencia de Datos: https://datafloq.com/big-data-open-source-tools/os-home/

Weka

- Fue el primer software libre de aprendizaje automático en Java
- Desarrollado por The University of Waikato, Nueva Zelanda. http://www.cs.waikato.ac.nz/ml/weka/

KEEL

- Software en Java. Se caracteriza por tener una buena batería de algoritmos de preprocesado de datos y algoritmos de aprendizaje basado en computación evolutiva
- Desarrollado en la Universidad de Granada. http://www.keel.es

KNIME

http://www.knime.com/

- **KNIME** (o Konstanz Information Miner) es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. KNIME está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en Java
- Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania
- Se integra con Weka, R, Python, Hadoop, Spark, TensorFlow…
- En 2017, KNIME agregó versiones en la nube de su plataforma para AWS y Microsoft Azure, características mejoradas de calidad de datos y capacidades ampliadas de *deep learning*

Open for Innovation ®
KNIME

Gartner 2018 Magic Quadrant for Data Science and Machine Learning Platforms

scikit-learn (Phyton)

- http://scikit-learn.org/

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

CRAN en R

CRAN: The Comprehensive R Archive Network

https://cran.r-project.org/web/views/MachineLearning.html

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Contributed Packages

Available Packages

Currently, the CRAN package repository features 13346 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.

CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 37 views are available.

Package Check Results

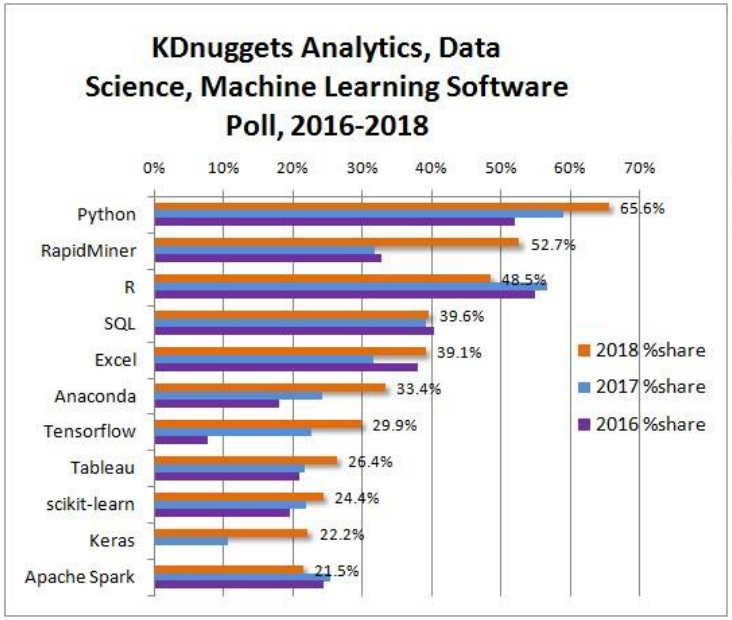
All packages are tested regularly on machines running Debian GNU/Linux, Fedora, OS X, Solaris and Windows.

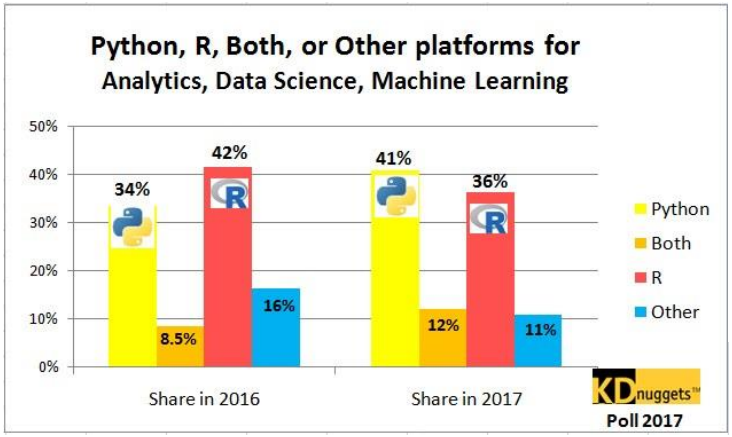
The results are summarized in the check summary (some timings are also available). Additional details for Windows checking and building can be found in the Windows check summary.

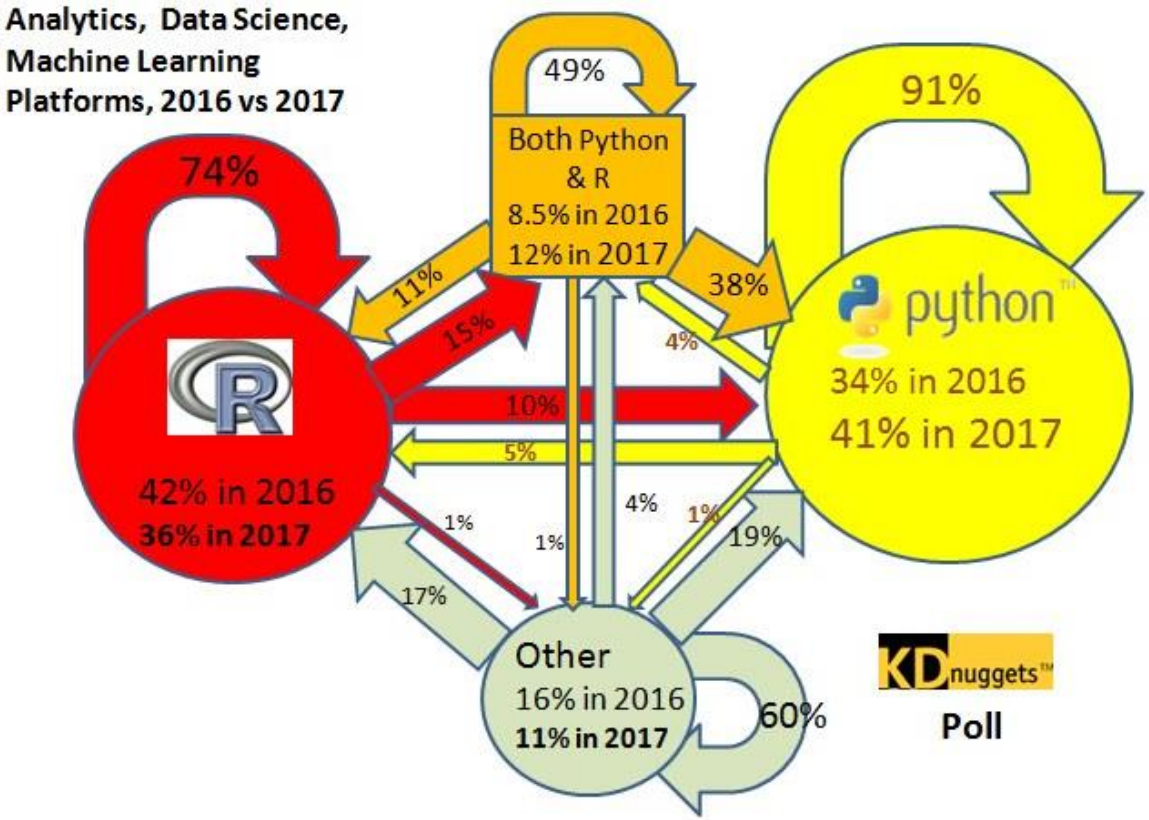
Writing Your Own Packages

The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.

Encuestas sobre el uso de herramientas de ciencia de datos







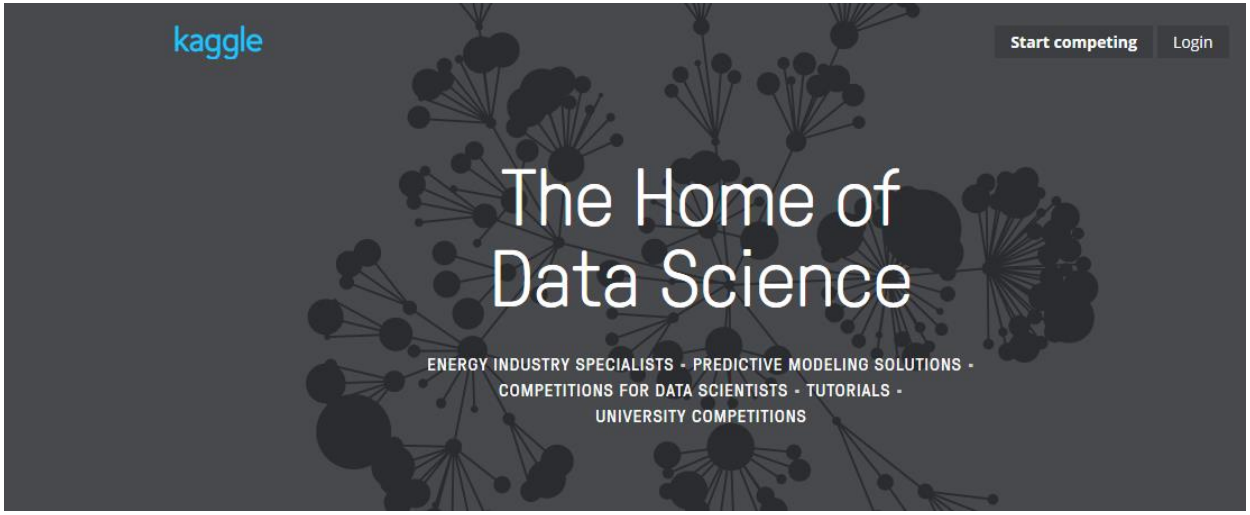
http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

Kaggle: The Home of Data Science

http://www.kaggle.com/

- Es un portal web que ofrece competiciones, tutoriales, actividades académicas…



Competiciones activas en Kaggle

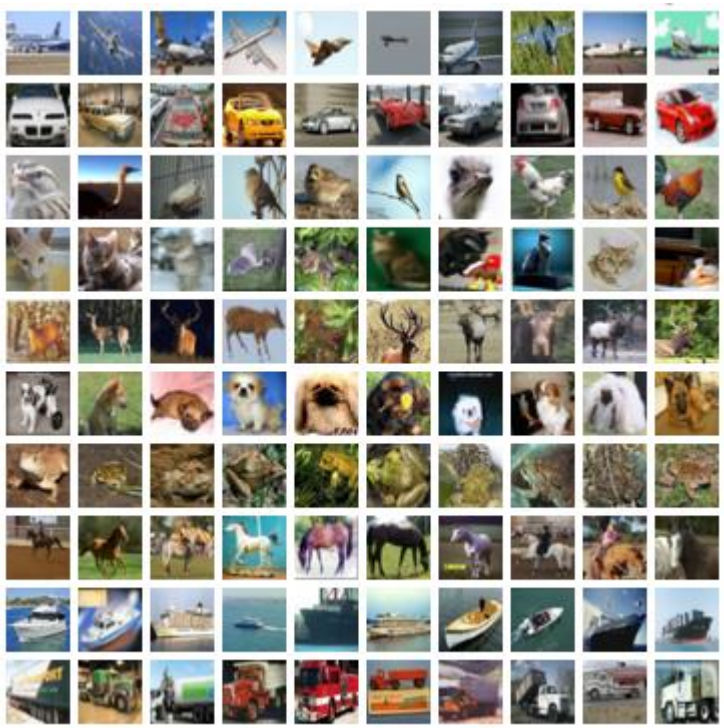
Two Sigma: Using News to Predict Stock Movements Use news analytics to predict stock price performance Featured · Kernels Competition · 2 months to go · news agen...	$100,000 2,927 teams	**House Prices: Advanced Regression Techniques** Predict sales prices and practice feature engineering, RFs, and gra... Getting Started · Ongoing · tabular data, regression	Knowledge 4,816 teams
Jigsaw Unintended Bias in Toxicity Classification Detect toxicity across a diverse range of conversations Featured · Kernels Competition · a month to go · biases, nlp, ...	$65,000 2,303 teams	**ImageNet Object Localization Challenge** Identify the objects in images Research · 11 years to go · object detection, image data	Knowledge 38 teams
LANL Earthquake Prediction Can you predict upcoming laboratory earthquakes? Research · 7 days to go · signal processing, physics, earth scie...	$50,000 4,337 teams	**Predict Future Sales** Final project for "How to win a data science competition" Courser... Playground · 7 months to go	Kudos 3,217 teams
Google Landmark Recognition 2019 Label famous (and not-so-famous) landmarks in images Research · 7 days to go	$25,000 274 teams	**iMaterialist (Fashion) 2019 at FGVC6** Fine-grained segmentation task for fashion and apparel Research · 14 days to go	Kudos 157 teams
Google Landmark Retrieval 2019 Given an image, can you find all of the same landmarks in a datas... Research · 7 days to go	$25,000 141 teams	**iNaturalist 2019 at FGVC6** Fine-grained classification spanning a thousand species Research · 14 days to go	Kudos 195 teams
Data Science for Good: City of Los Angeles Help the City of Los Angeles to structure and analyze its job descr... Analytics · 25 days to go · employment, text data, image data,...	$15,000	**iWildCam 2019 - FGVC6** Categorize animals in the wild Playground · 11 days to go · multiclass classification, image da...	Kudos 287 teams
Instant Gratification A synchronous Kernels-only competition Featured · Kernels Competition · 24 days to go · tabular data,...	$5,000 696 teams	**iMet Collection 2019 - FGVC6** Recognize artwork attributes from The Metropolitan Museum of ... Research · Kernels Competition · 8 days to go · image data, v...	Kudos 472 teams
Freesound Audio Tagging 2019 Automatically recognize sounds and apply tags of varying natures Research · Kernels Competition · 14 days to go · sound techn...	$5,000 739 teams	**Northeastern SMILE Lab - Recognizing Faces in the W...** Can you determine if two individuals are related? Playground · 2 months to go · relationships, image data	Knowledge 181 teams
Digit Recognizer Learn computer vision fundamentals with the famous MNIST data Getting Started · Ongoing · object identification, multiclass cl...	Knowledge 3,049 teams	**Aerial Cactus Identification** Determine whether an image contains a columnar cactus Playground · Kernels Competition · a month to go · image da...	Knowledge 669 teams
Titanic: Machine Learning from Disaster Start here! Predict survival on the Titanic and get familiar with M... Getting Started · Ongoing · tabular data, binary classification,...	Knowledge 11,436 teams	**TMDB Box Office Prediction** Can you predict a movie's worldwide box office revenue? Playground · 3 days to go · film, tabular data	Knowledge 1,317 teams

Ejemplo de competición

CIFAR-10 - Object Recognition in Images

Identify the subject of 60,000 labeled images

55 days
190 teams
Knowledge



60,000 32x32 color images containing one of 10 object classes, with 6000 images per class.

Dashboard

Leaderboard - CIFAR-10 - Object Recognition in Images

This leaderboard is calculated on all of the test data.

See someone using multiple

#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	Raillou	0.92610	7	Wed, 16 Apr 2014 14:45:45
2	–	DeepCNet	0.91820	1	Mon, 20 Jan 2014 12:12:53
3	↑3	git.io/nagadomi-cifar10	0.91730	7	Thu, 21 Aug 2014 21:40:00
4	↓1	Dmitriy Anisimov	0.90670	2	Thu, 17 Apr 2014 09:18:53
5	↓1	fastml.com/dropconnect	0.90660	4	Sun, 22 Dec 2013 15:08:58
6	↓1	Yan	0.90530	13	Fri, 29 Nov 2013 05:33:29

Comunidad Kaggle

Es una muy buena oportunidad para practicar en la resolución de problemas reales y la adquisición de habilidades en Ciencia de Datos

Rank	Tier	User		Medals			Points
1		Giba	joined 6 years ago	42	30	23	183,755
2		bestfitting	joined 2 years ago	14	4	0	183,346
3		Μαριος Μιχαηλιδης KazAnova	joined 5 years ago	32	30	27	145,039
4		Little Boat	joined 5 years ago	16	18	6	109,007
5		raddar	joined 3 years ago	10	11	4	107,060
6		Komaki	joined 6 years ago	5	7	3	102,525
7		Eureka	joined 5 years ago	20	20	5	100,895
8		ZFTurbo	joined 3 years ago	11	18	10	97,621
9		alijs	joined 2 years ago	4	11	6	92,373
10		Kohei	joined 8 years ago	15	31	19	91,599

124 Grandmasters

1,034 Masters

3,721 Experts

44,376 Contributors

44,036 Novices

DrivenData

DRIVENDATA

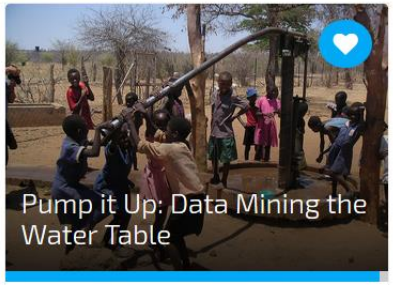
https://www.drivendata.org/

1 MONTH LEFT

Using environmental data collected by U.S. Federal Government agencies, can you predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru?

LuShaofeng
CURRENT LEADER

COMPETE →

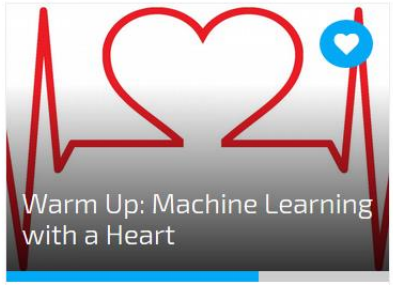


1 MONTH LEFT

Can you predict which water pumps are faulty to promote access to clean, potable water across Tanzania? This is an intermediate-level practice competition.

smw012
CURRENT LEADER

COMPETE →



5 MONTHS LEFT

Can you predict the presence or absence of heart disease in patients given basic medical information? This is the smallest, least complex dataset on DrivenData, and a great place to dive into the world of data science competitions.

unknow
CURRENT LEADER

COMPETE →

7 MONTHS, 1 WEEK LEFT

Can you predict the level of damage to buildings caused by the 2015 Gorkha earthquake in Nepal based on aspects of building location and construction?

Gillesvdw
CURRENT LEADER

COMPETE →

10 MONTHS LEFT

The UN's Millennium Development Goals provide the big-picture perspective on international development. Using indicators aggregated and collected by the World Bank, try to predict progress towards select MDGs.

Ganesh221B
CURRENT LEADER

COMPETE →

10 MONTHS LEFT

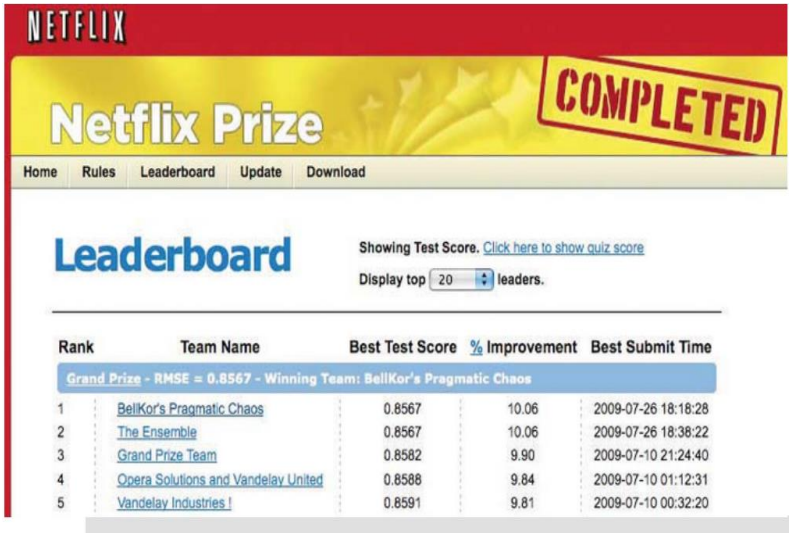
We're rebooting our first prized competition for fun and education! Tag school budgets automatically to help districts get a better grasp of their spending and how to improve the impact of their scarce resources.

NUDT_DINGZH...
CURRENT LEADERS

COMPETE →

Conclusiones

No todo en la vida es *big data*



- El sistema de recomendación de Netflix le proporciona 75% de las ventas, así que en 2009 ofreció un premio de 1 millón de dólares a quien mejorara en un 10% su sistema. Luego comprendieron que al final era una mejora marginal
- Así que usaron **Thick Data** (datos gruesos) para mejorar la experiencia. Contrataron al etnógrafo Grant McCraken, que entró en los salones de usuarios de TV en EE.UU. y Canadá para conocer su nuevos hábitos
- Se observó que los usuarios demandaban ver series completas del tirón en lugar de esperar al siguiente episodio cada semana

https://media.netflix.com/en/press-releases/netflix-declares-binge-watching-is-the-new-normal-migration-1

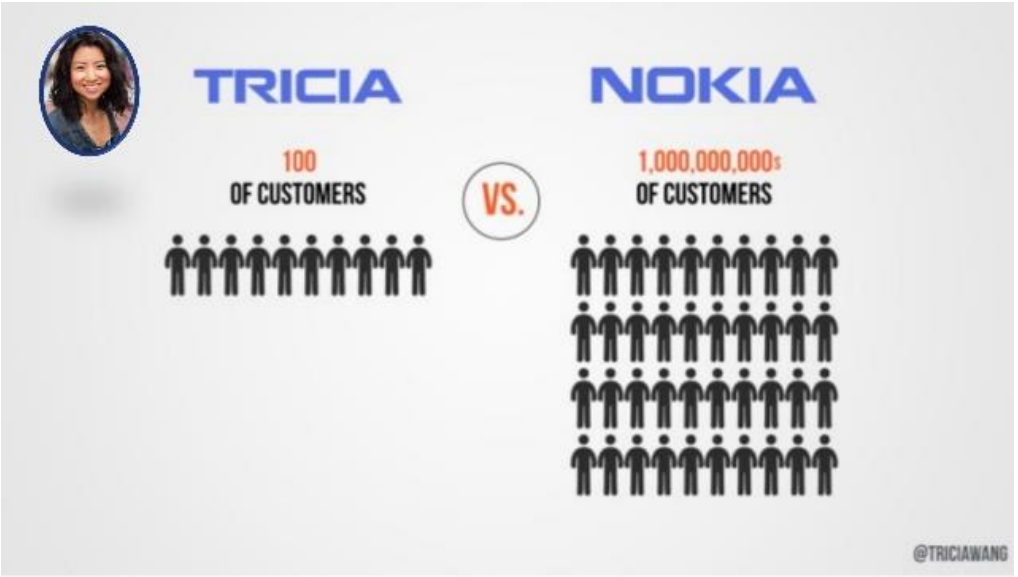
Thick Data (datos gruesos)

- Para formar una imagen completa, tanto el ***big data*** como el ***thick data*** son críticos porque producen diferentes tipos de información a diferentes escalas y profundidades

- No todo es observable cuantitativamente, también hay factores cualitativos que no recoge el *big data*

- Nokia ignoró este aspecto e hizo caso a sus datos que indicaban que era mejor orientar sus productos a la exclusividad en lugar de hacer el móvil accesible a todos como concluía su etnógrafa **Tricia Wang**

- Ya sabemos cómo acabó Nokia…



https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data

Errores comunes

- Evitar aprender de cosas que no son ciertas
  - Patrones que no representan ninguna regla subyacente
  - Datos que no reflejan lo relevante
  - Datos con un nivel de detalle erróneo

- Evitar aprender cosas ciertas, pero inútiles
  - Aprender información ya conocida
  - Aprender cosas que no se pueden utilizar
- Evitar usar un algoritmo/método para todos los problemas
- Evitar sesgos que surgen en distintas fases del proceso: recolección de datos, preparación de datos, modelizado, evaluación e implementación

Tipos de sesgo en aprendizaje automático

- El **sesgo de muestreo** puede producir modelos entrenados en datos que no son totalmente representativos de casos futuros
- El **sesgo de rendimiento** puede exagerar las percepciones de poder predictivo, generalización y homogeneidad de rendimiento en todos los segmentos de datos
- El **sesgo de confirmación** puede hacer que la información sea buscada, interpretada, enfatizada y recordada de una manera que confirma los preconceptos
- El **sesgo de anclaje** puede llevar a un exceso de confianza en la primera información examinada

https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

¿Cómo evitar el sesgo?

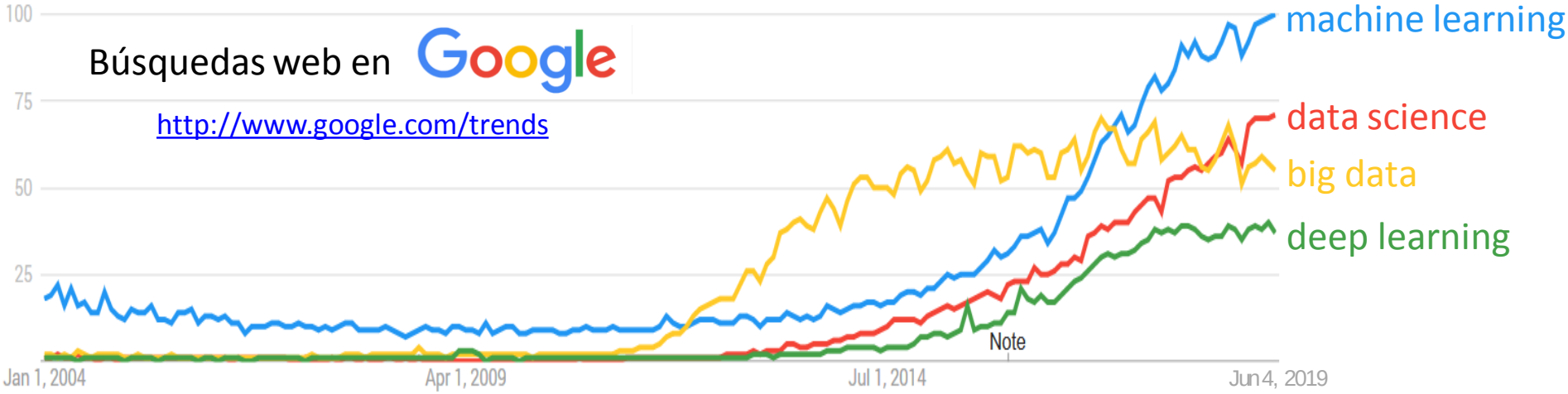
1. Unir científicos de datos con **expertos en ciencias sociales** para introducir una perspectiva humanista a los datos y a la solución a implementar
2. Anotar los datos **con precaución**. Hacer que los responsables de anotar los datos que van a alimentar un modelo de aprendizaje automático sean conscientes de sus posibles sesgos
3. Medir la **equidad**. Utilizando métricas para medir la imparcialidad que ayuden a corregir el sesgo
4. No concentrarse sólo en la representatividad. Una vez que se hayan establecido las medidas de equidad, encontrar un **equilibrio** entre la representatividad de los casos futuros y la **infrarrepresentación** de las minorías
5. Tener en cuenta la eliminación del sesgo y, si es necesario, **evitar** la categorización basada principalmente en **variables sociodemográficas**

https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

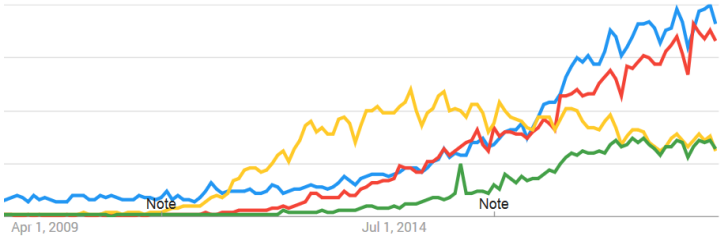
Evolución de la popularidad de ciencia de datos y otros conceptos afines

Búsquedas web en Google

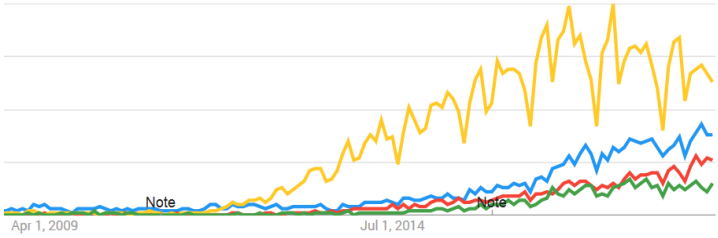
http://www.google.com/trends



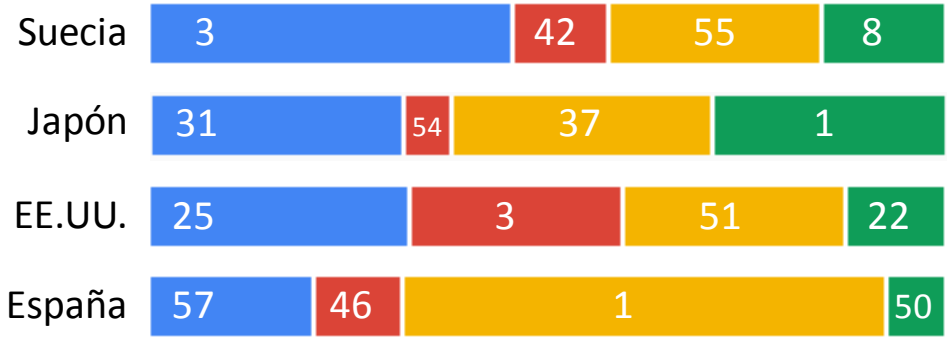
EE.UU.



España



Distribución relativa. Ranking entre **57** países



Primeros pasos en aprendizaje automático

- **Nivel principiante:**
  A visual introduction to machine learning
  http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

- **Si entendiste el tutorial anterior, continua con este otro:**
  Intro to Machine Learning
  https://www.kaggle.com/learn/intro-to-machine-learning

- **¿Estás pillando la idea, más o menos? Genial, intenta este video para convertirte en experto:**
  Complete Data Science Course | What is Data Science? | Data Science for Beginners | Edureka
  https://www.youtube.com/watch?v=aGu0fbkHhek

- **¿Te has perdido en el segundo tutorial? No te agobies, olvida el vídeo anterior y mejor mira esto:**
  What is Data Science? | Introduction to Data Science | Data Science for Beginners | Simplilearn
  https://www.youtube.com/watch?v=KxryzSO1Fjs

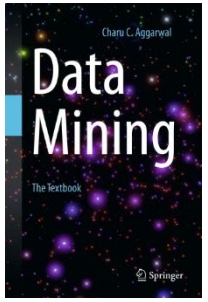
- **¿Sabías que aprendizaje automático puede ser sexista y racista? ¿Por qué? Lee esto:**
  A Tutorial on Fairness in Machine Learning
  https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

- **Mis vídeos ☺:**
  http://decsai.ugr.es/~casillas/videos.html

Bibliografía

- Libros buenos e interesantes:
  - http://www.datasciencecentral.com/profiles/blogs/15-books-every-data-scientist-should-read
- Libros gratis:
  - http://www.wzchen.com/data-science-books/
- Compendio completo sobre Data Mining:

  C.C. Aggarwal, Data Mining, Springer, 2015



- Sobre divulgación en Big Data:

  V. Mayer-Schönberger y K. Cukier,
  Big Data. La Revolución de los Datos Masivos,
  Noema, 2015

La cara oculta del *big data*

- **Byung-Chul Han**,
  La Sociedad de la Transparencia. Herder (2013)
- **Safiya Umoja Noble**,
  Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press (2018)
- **Cathy O'Neil**,
  Armas de destrucción matemática. Capitán Swing Libros S.L. (2018)
- **Scott Galloway**,
  The Four: The Hidden DNA of Amazon, Apple, Facebook, and Google. Portfolio Penguin (2017)
- **Yuval Noah Harari**,
  Homo Deus. Debate (2018)

http://decsai.ugr.es/~casillas/talks.html