Data Mining

Raimundo Sánchez, PhD

Facultad de Ingeniería y Ciencias

Universidad Adolfo Ibáñez

Raimundo Sánchez

- Ingeniero Industrial
- Doctor en Ingenieria de Sistemas Complejos
- Profesor de Data Science UAI
- Investigo deportes de resistencia
- Corredor de montañas

Data Mining

Uso de datos y métodos cuantitativos para mejorar la toma de decisiones

Análisis de datos en las empresas

Modelos estadísticos

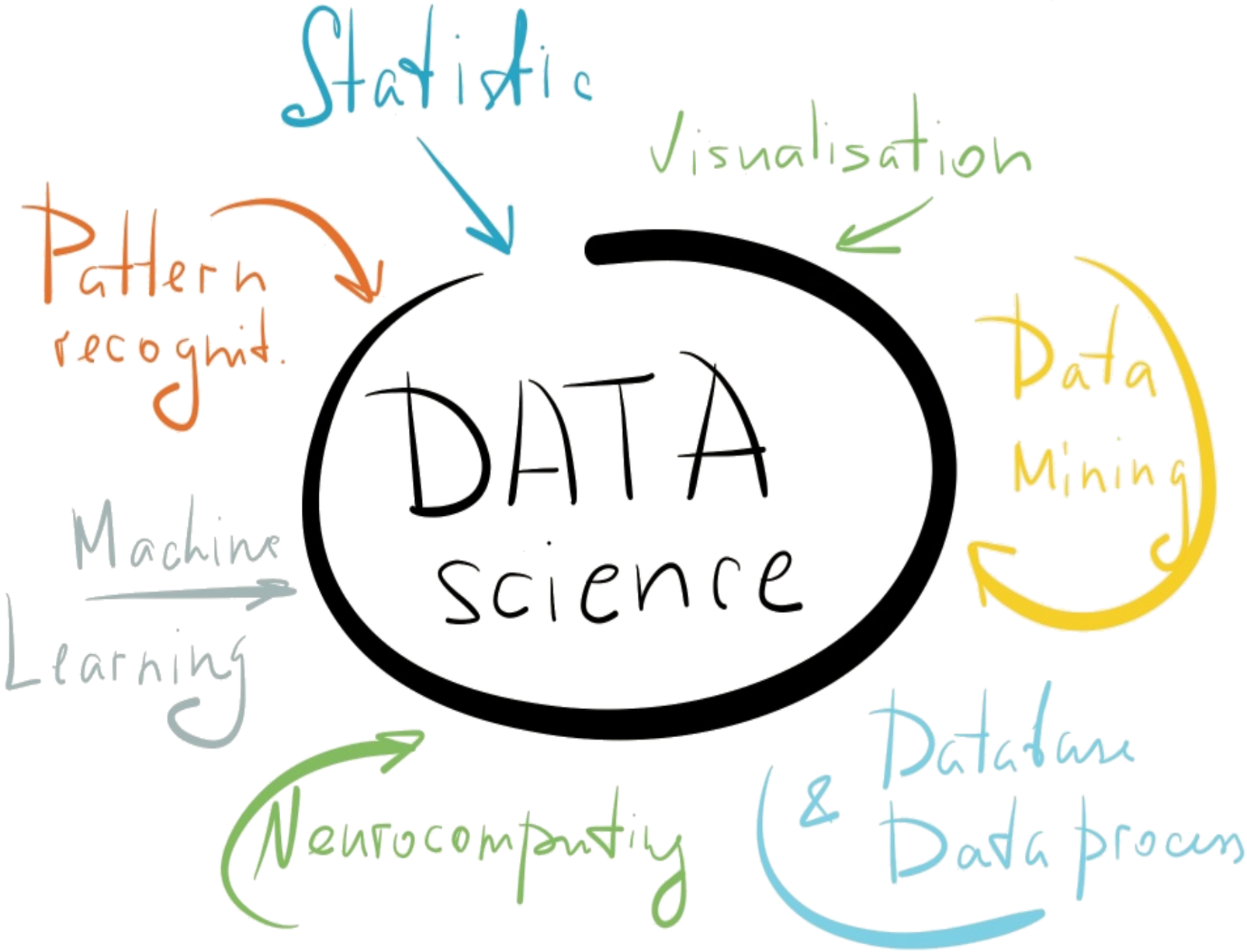
Intuición experta

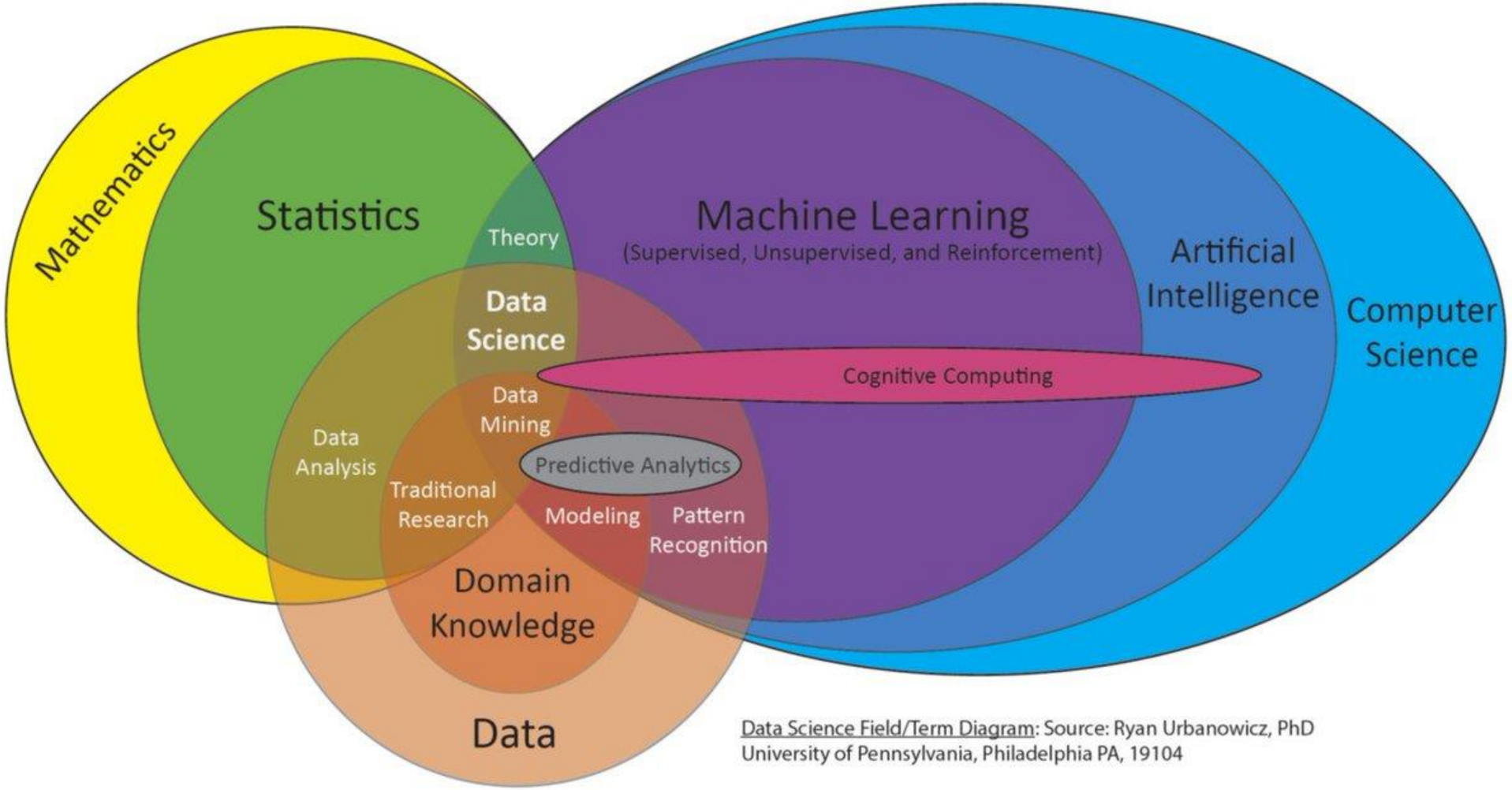
Sin data

Expertos + Data histórica

Data Science

¿Qué es Data Science?

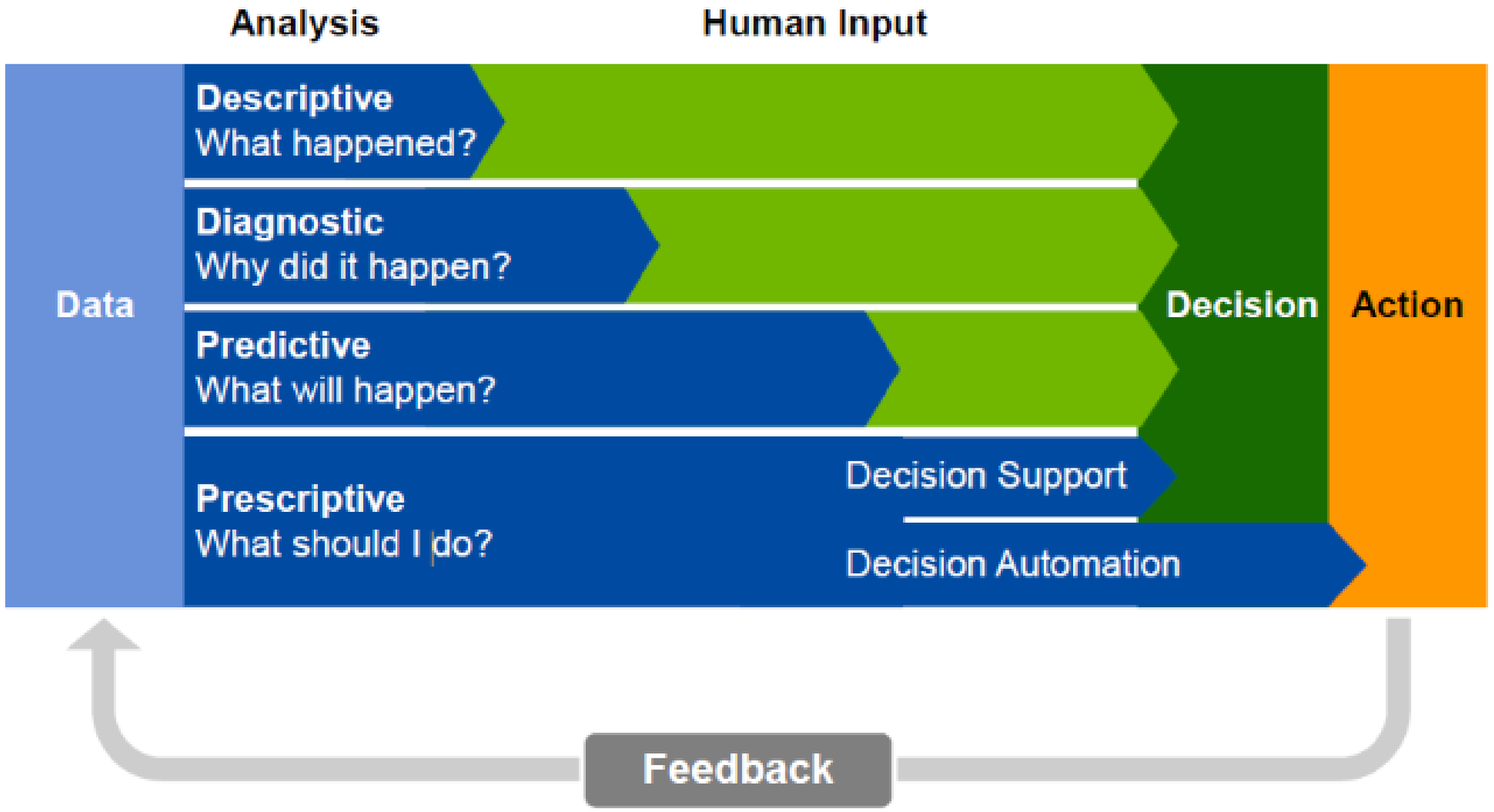




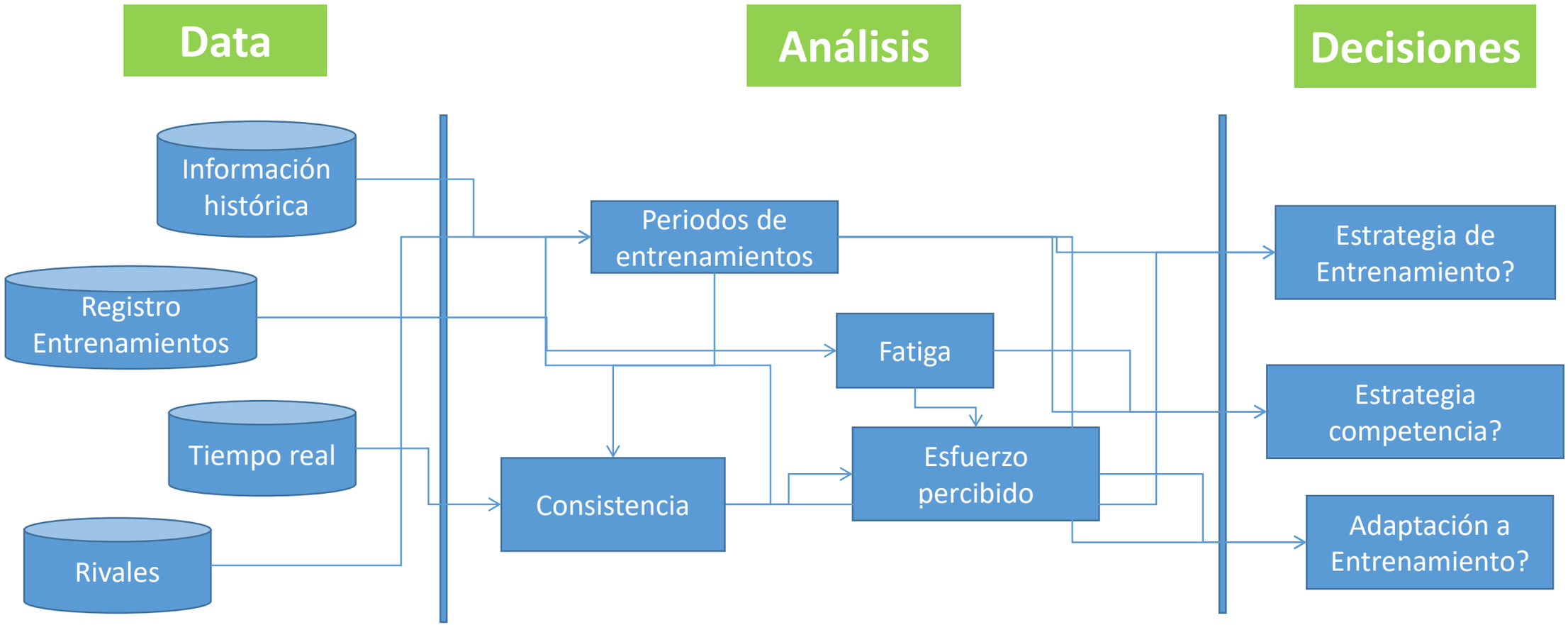
Machine learning: How can we build computer systems that automatically improve with experience? *(Mitchell 2006)*

Data mining: The process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data
(Fayyad, Piatetsky-Shapiro & Smith 1996)

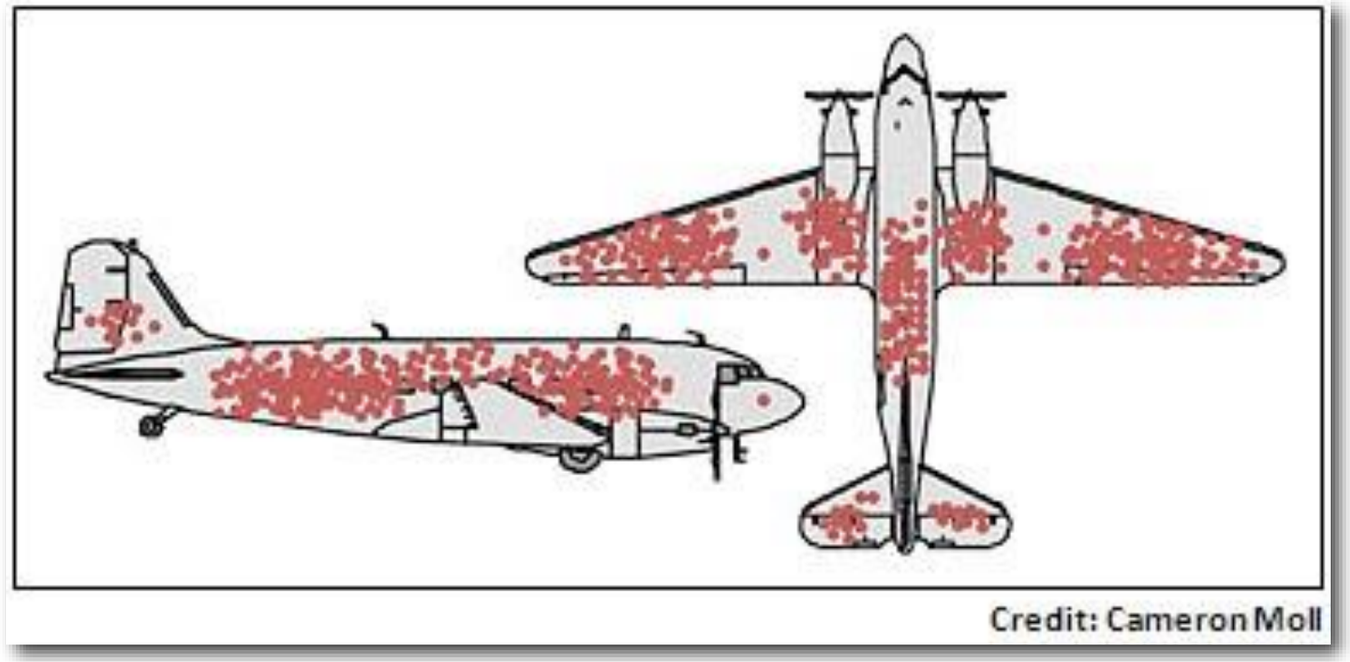
Niveles de decisión



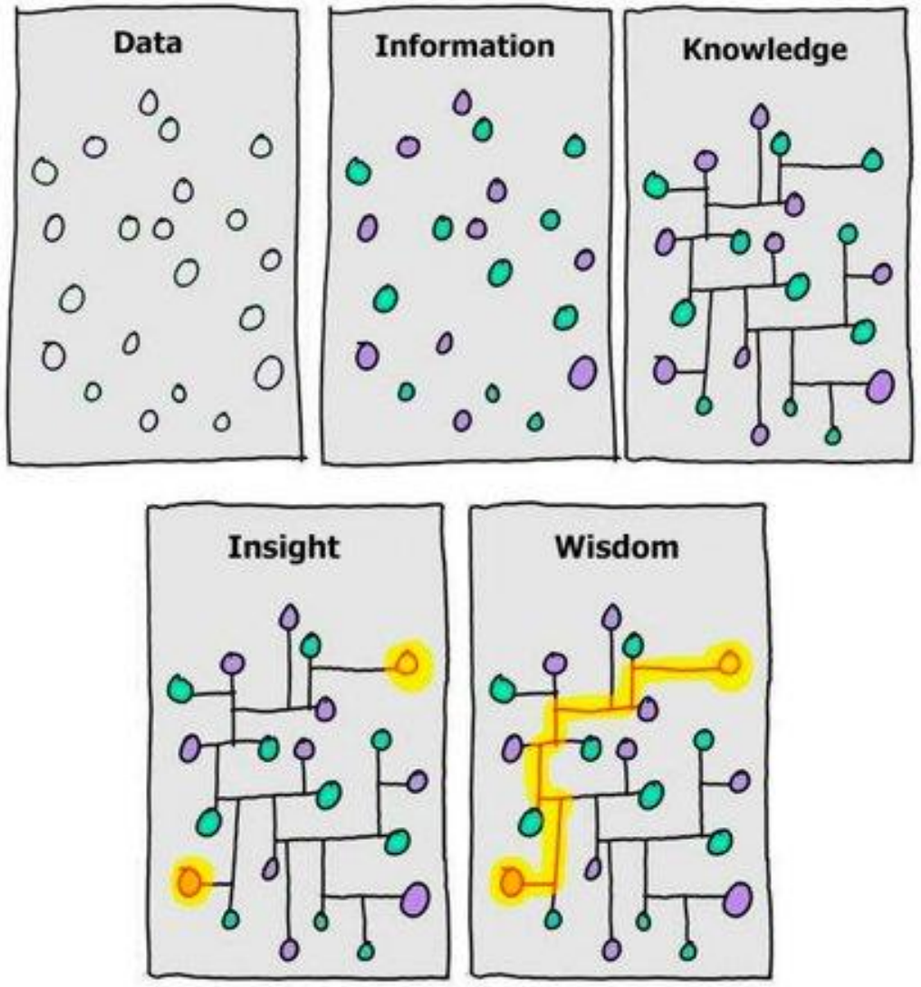
Flujo de información



Ejemplo



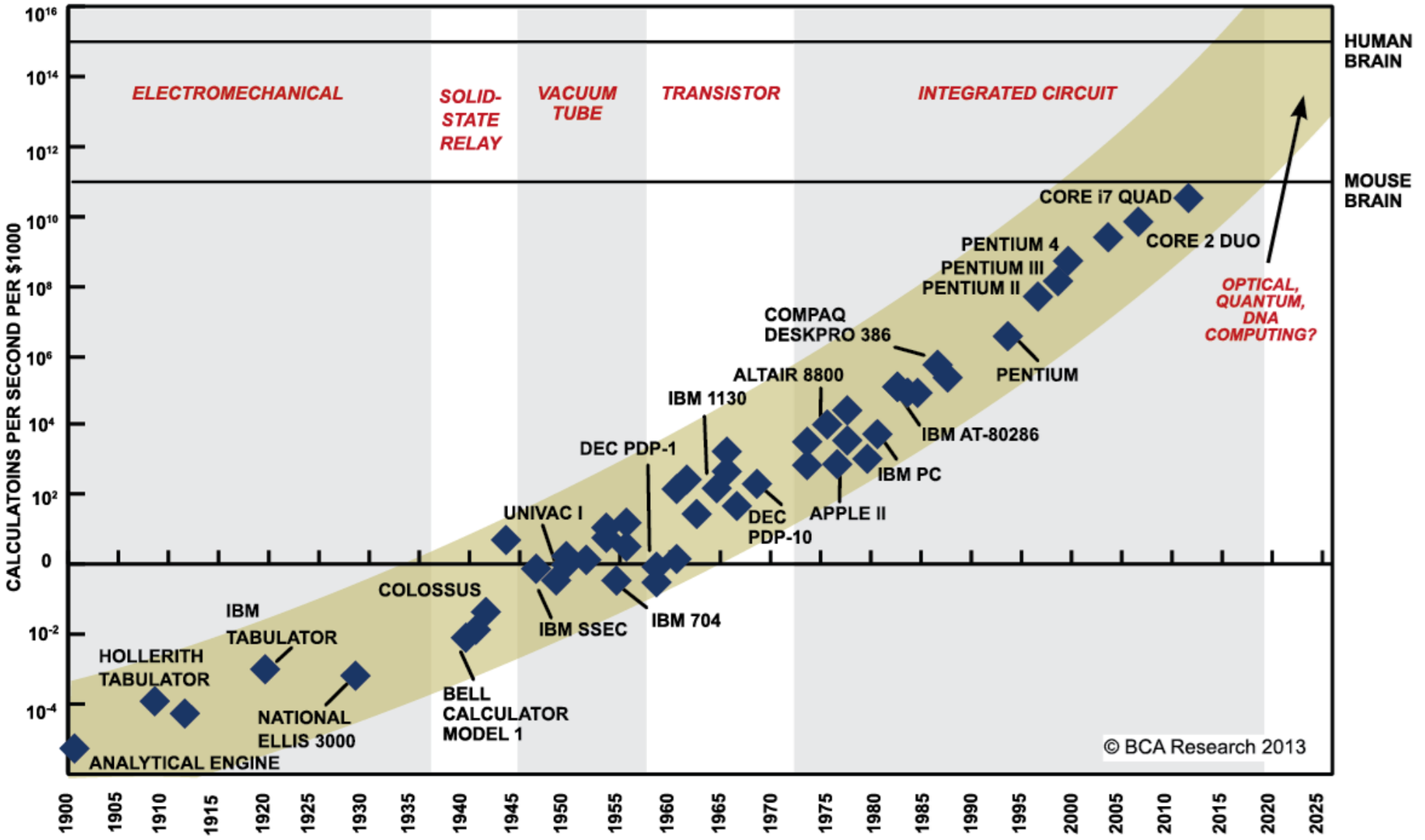
Durante la Segunda Guerra Mundial, se le pidió al estadístico Abraham Wald que ayudara a los británicos a decidir dónde añadir armadura a sus aviones


Data
Information
Knowledge
Insight
Wisdom

Los 3 Niveles de la IA

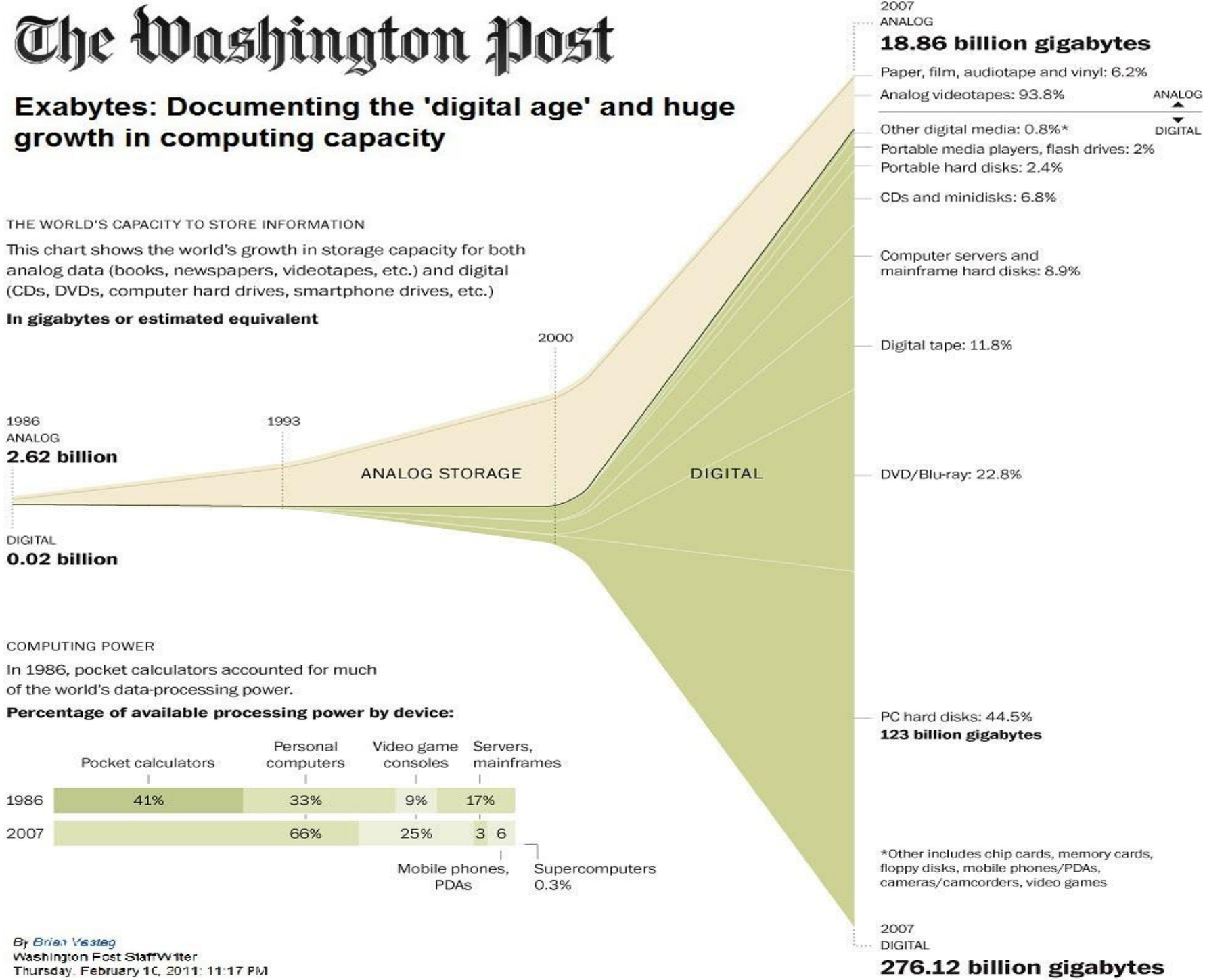
- Sistemas automáticos
- Redes Neuronales Artificiales
- Robótica Cognitiva

Singularidad cognitiva



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

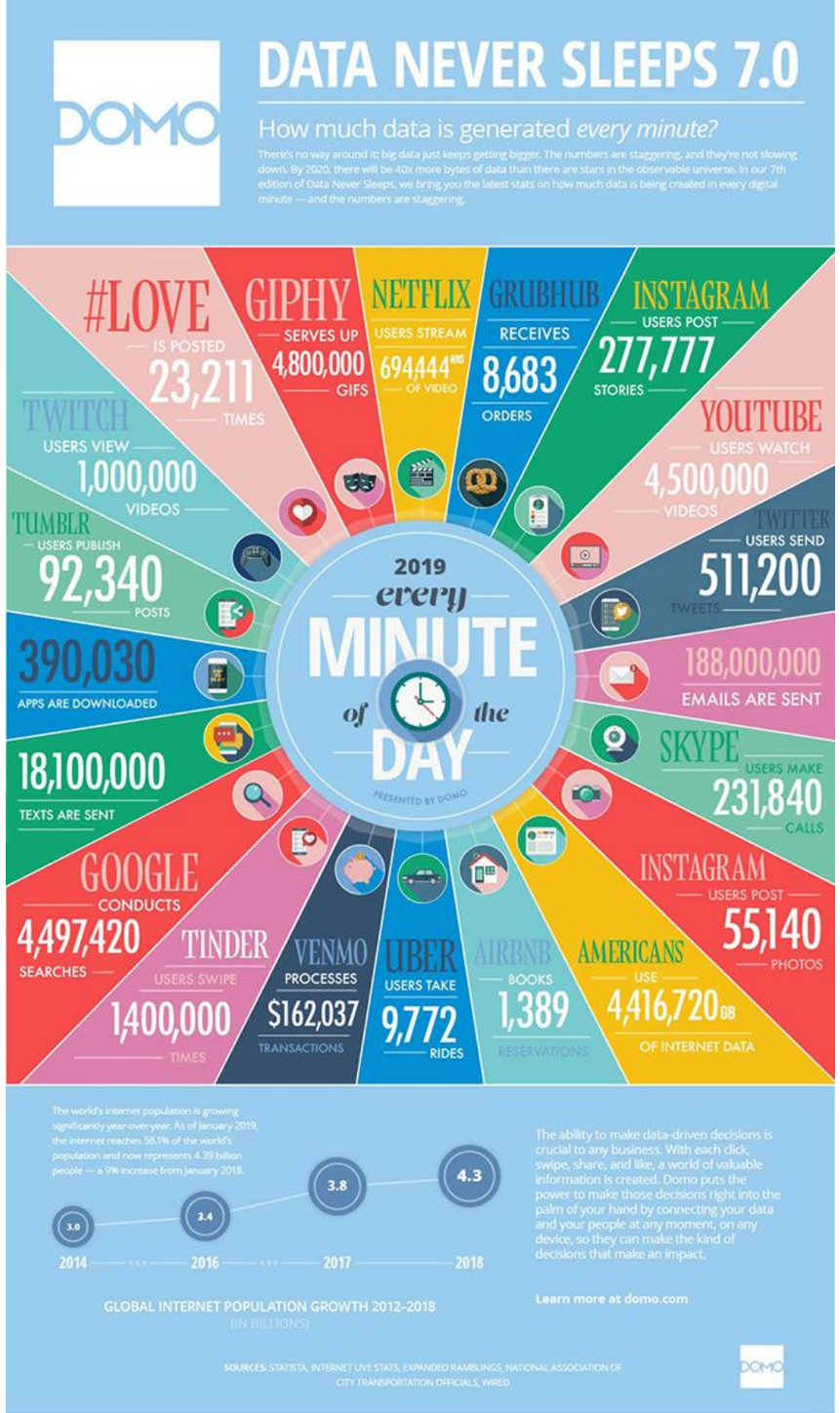
Singularidad digital

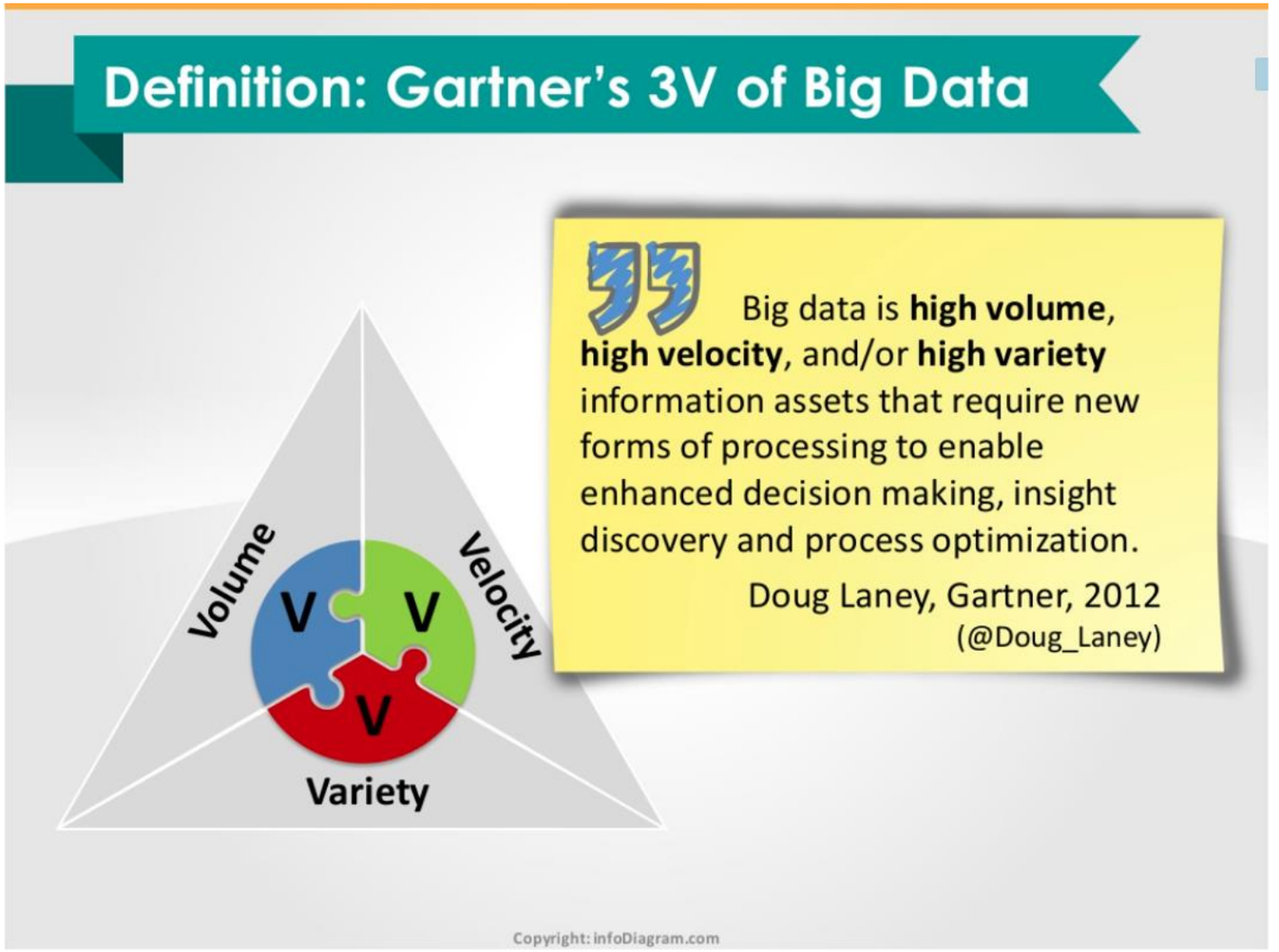


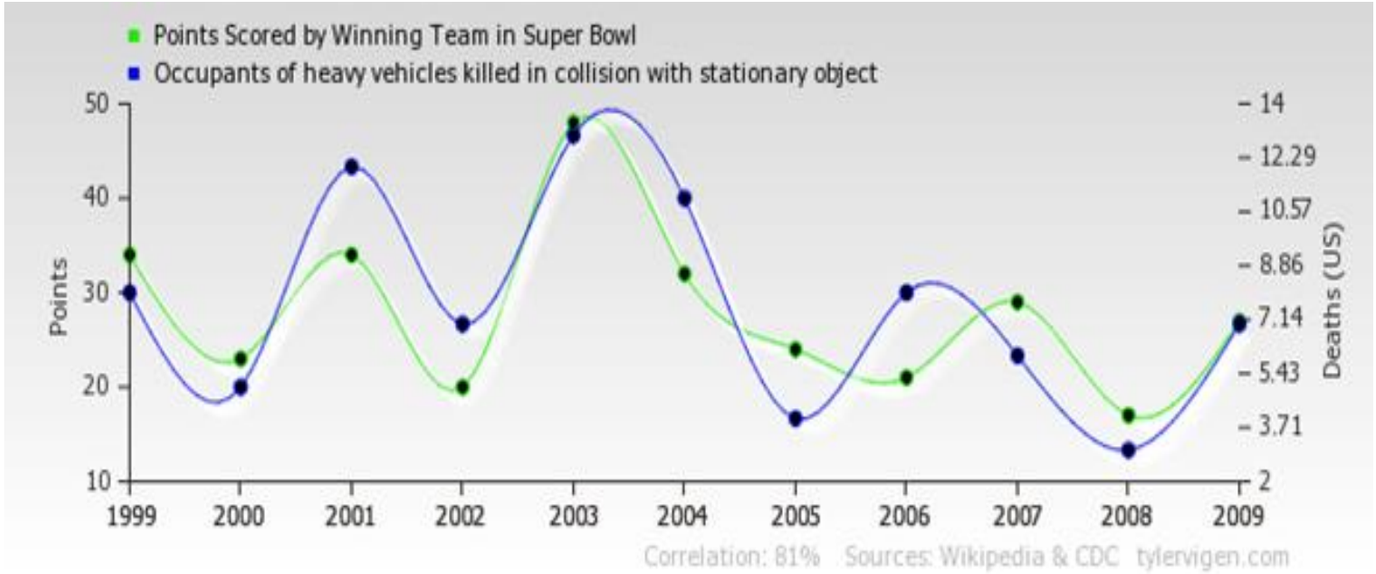
Big data

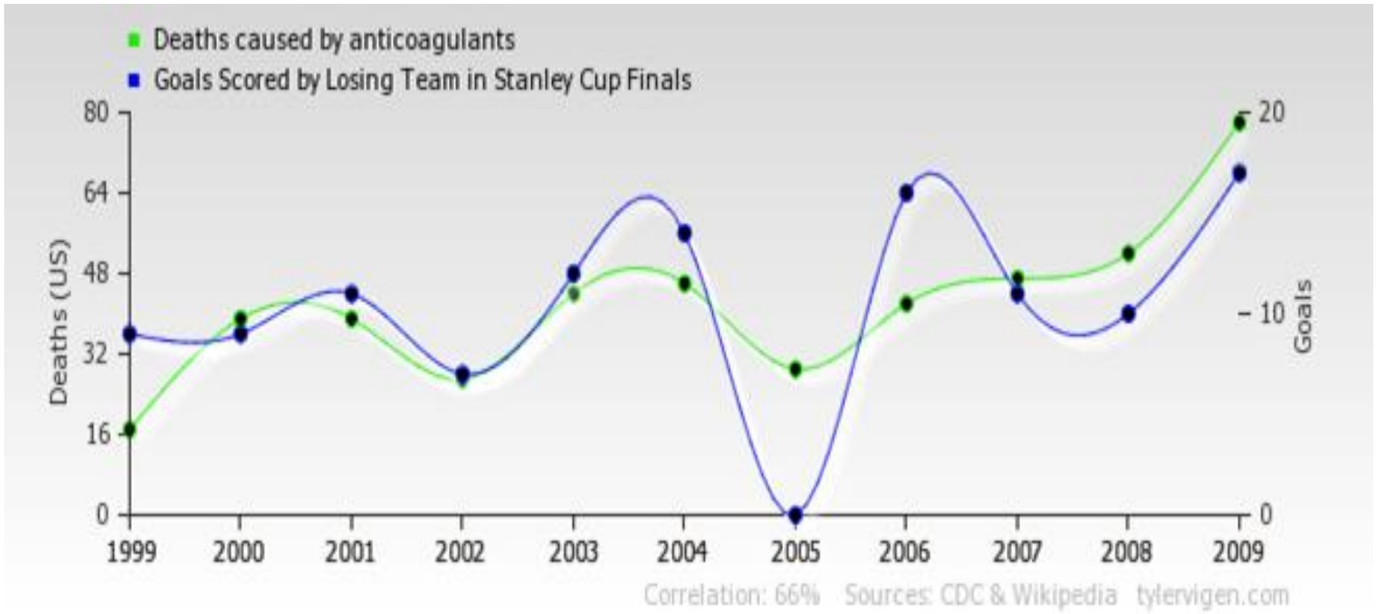
What is big data?

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data.**


DOMO
DATA NEVER SLEEPS 7.0
How much data is generated every minute?
There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute — and the numbers are staggering.
#LOVE IS POSTED 23,211 TIMES
GIPHY SERVES UP 4,800,000 GIFS
NETFLIX USERS STREAM 694,444 MINS OF VIDEO
GRUBHUB RECEIVES 8,683 ORDERS
INSTAGRAM USERS POST 277,777 STORIES
TWITCH USERS VIEW 1,000,000 VIDEOS
YOUTUBE USERS WATCH 4,500,000 VIDEOS
TUMBLR USERS PUBLISH 92,340 POSTS
TWITTER USERS SEND 511,200 TWEETS
2019 every MINUTE of the DAY
PRESENTED BY DOMO
390,030 APPS ARE DOWNLOADED
188,000,000 EMAILS ARE SENT
18,100,000 TEXTS ARE SENT
SKYPE USERS MAKE 231,840 CALLS
GOOGLE CONDUCTS 4,497,420 SEARCHES
INSTAGRAM USERS POST 55,140 PHOTOS
TINDER USERS SWIPE 1,400,000 TIMES
VENMO PROCESSES $162,037 TRANSACTIONS
UBER USERS TAKE 9,772 RIDES
AIRBNB BOOKS 1,389 RESERVATIONS
AMERICANS USE 4,416,720 GB OF INTERNET DATA
The world's internet population is growing significantly year-over-year. As of January 2019, the internet reaches 56.1% of the world's population and now represents 4.39 billion people — a 9% increase from January 2018.
3.0
3.4
3.8
4.3
2014
2016
2017
2018
GLOBAL INTERNET POPULATION GROWTH 2012-2018
(IN BILLIONS)
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.
Learn more at domo.com
SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED
DOMO


Definition: Gartner's 3V of Big Data
Volume
Velocity
Variety
V
V
V
Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.
Doug Laney, Gartner, 2012
(@Doug_Laney)
Copyright: infoDiagram.com





No basta con robots

Los humanos aún debemos hacer las preguntas!

Realidad del trabajo con datos

La creatividad humana no es automatizable y se volverá muy valiosa.

Objetivos del curso

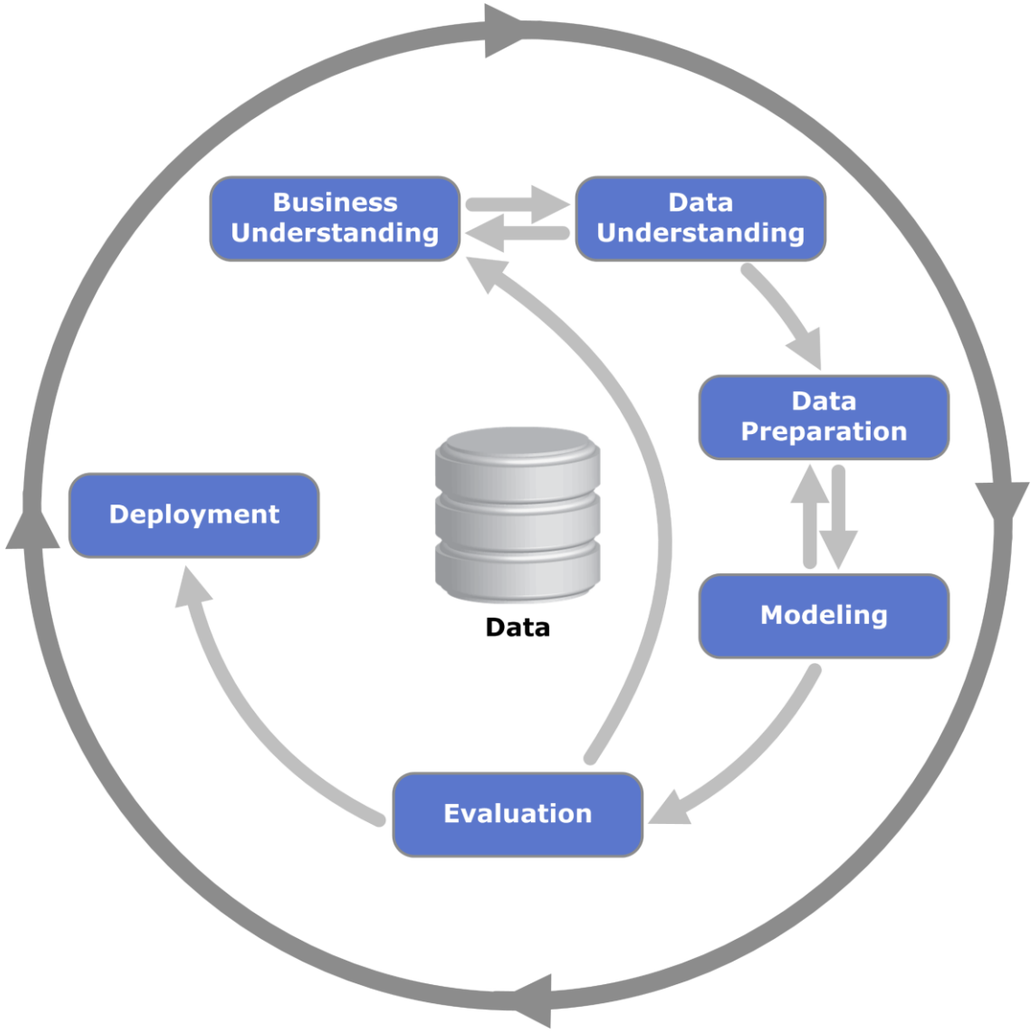
- Identificar elementos clave de la minería de datos
- Comprender cómo interactúan los elementos algorítmicos para afectar el rendimiento
- Comprender cómo elegir algoritmos para diferentes tareas de análisis
- Analizar los datos de una manera exploratoria y específica
- Implementar y aplicar algoritmos básicos para el aprendizaje supervisado y sin supervisión
- Evaluar con precisión el rendimiento de los algoritmos.

Calendario

02 mar 21	Introducción al Curso, R, RStudio, RMD y Github
09 mar 21	Data measurement, data types
16 mar 21	Exploratory Data analysis
23 mar 21	Outliers, Feature selection
30 mar 21	semana santa
06 abr 21	K-means, C-means (**ENTREGA P1**)
13 abr 21	Jerárquico, DBSCAN, GMM
20 abr 21	Evaluacion unsupervised
27 abr 21	dimensionality reduction
04 may 21	regresion lineal multiple
11 may 21	Regresión logística y perceptron multicapa (**ENTREGA P2**)
18 may 21	semana pausa
25 may 21	KNN, Naïve Bayes
01 jun 21	Evaluación, ROC, K-fold
08 jun 21	Decision Trees (Xgboost)
15 jun 21	redes neuronales profundas
22 jun 21	Modelos de Ensamblados, Random forest (**ENTREGA P3**)

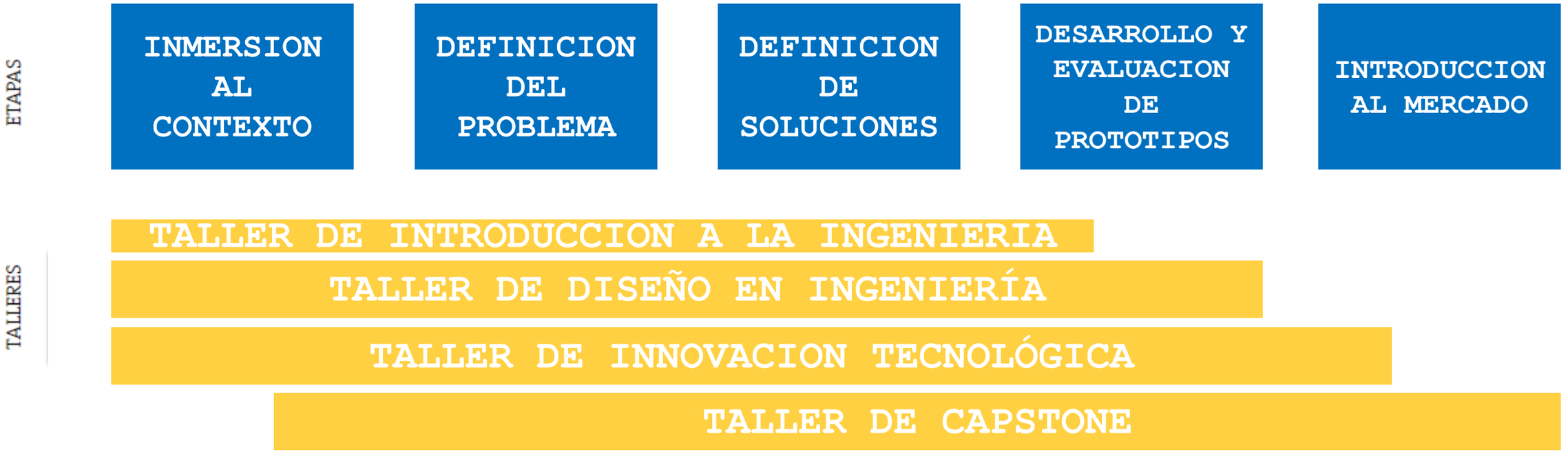
NF = 0.2 P1 + 0.3 P2 + 0.3 P3 + 0.2 Portafolio (Ayudantías)

CRISP DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Método Taller - UAI

Como resultado de un análisis de diversos métodos de resolución de problemas, se genero el Método Taller FIC, que consiste en 5 **etapas con objetivos acotados y asociados al desarrollo de la solución de ingeniería.** Al pasar por los talleres los alumnos repiten este proceso hasta internalizarlo, y son capaces de aplicarlo en cualquier contexto.



Software que utilizaremos este semestre

- R
- RStudio
- GitHub

EJEMPLO

- https://github.com/raimun2/curso_data_mining

Data Mining

Raimundo Sánchez, PhD

Facultad de Ingeniería y Ciencias

Universidad Adolfo Ibáñez