



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

SEDE VIÑA DEL MAR

Ciencia de Datos

Unidad 1: Proceso de Minería de Datos,
Tipos de Aprendizaje y Ciencia de Datos

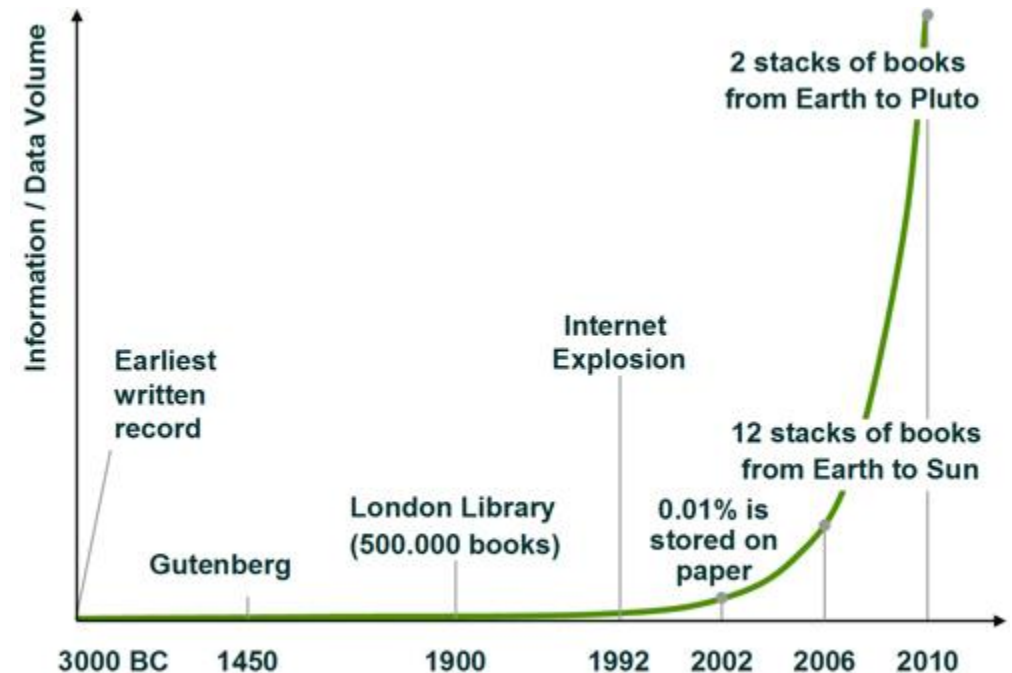
Profesor: Gabriel Jara

gabriel.jara@usm.cl

Segundo Semestre 2024

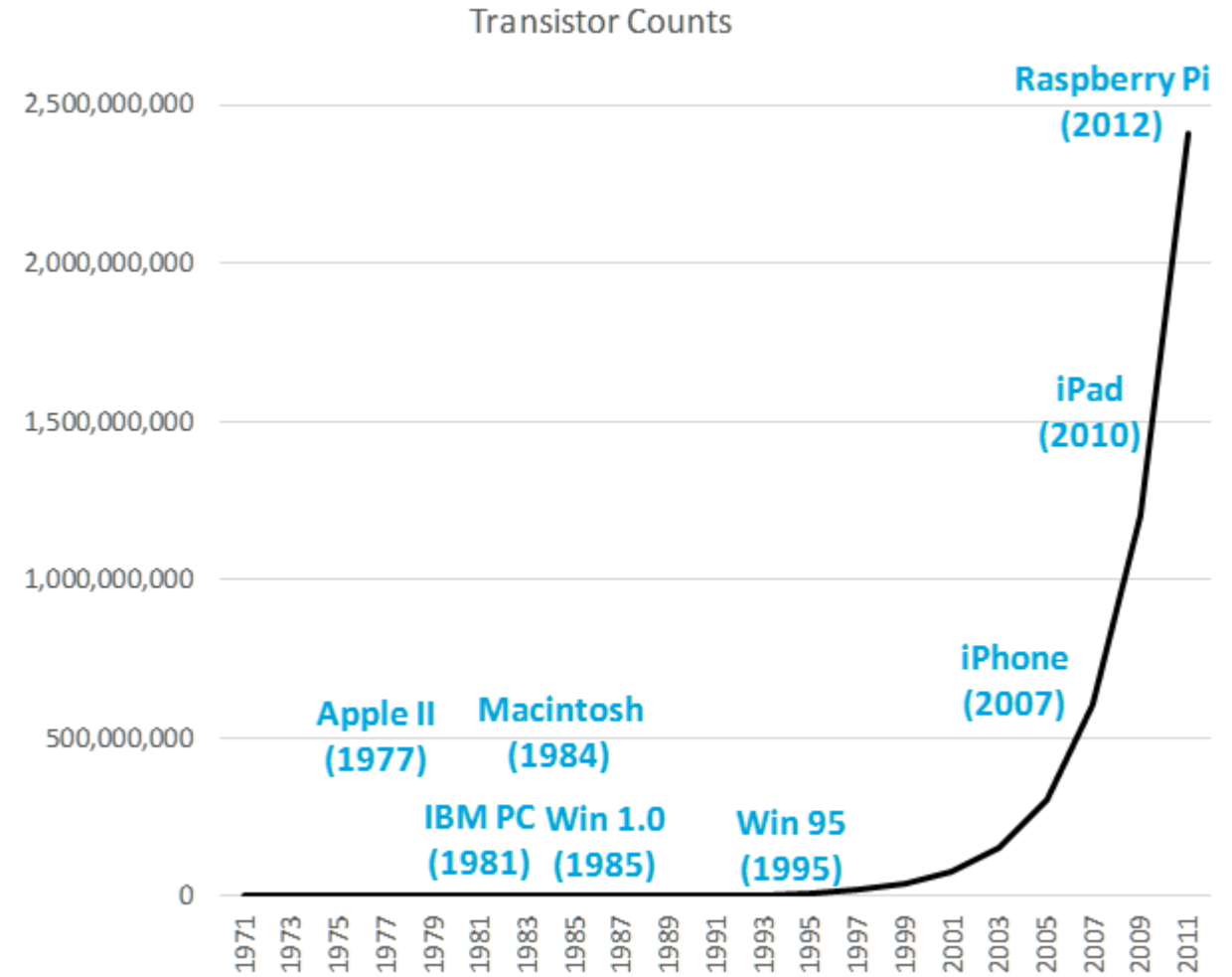
El problema de la Información

A lo largo de la historia, la información que producimos y acumulamos por generaciones ha crecido en forma exponencial. Los sistemas de información surgen en respuesta directa a esta realidad.



Ley de Moore

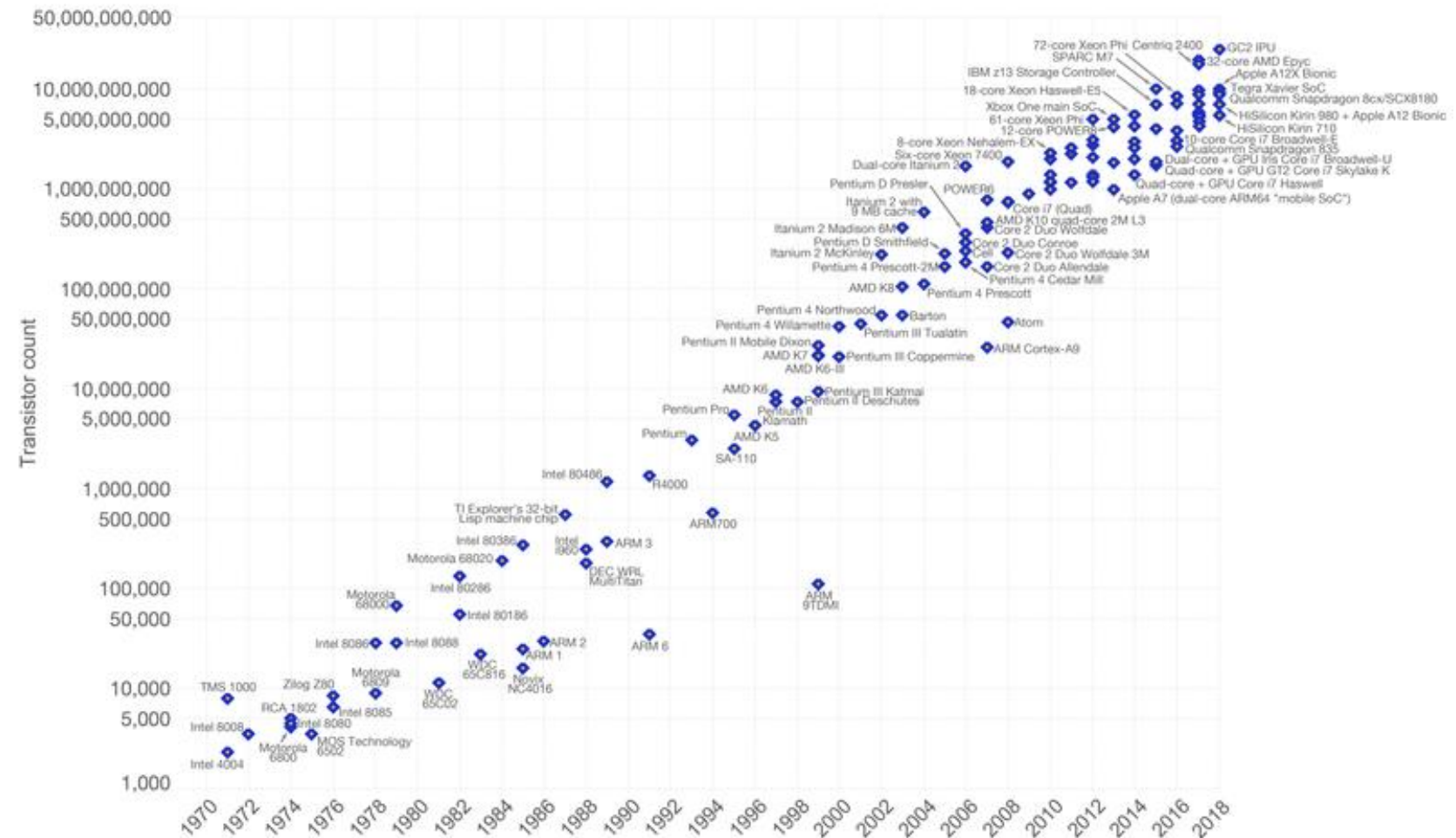
La masificación de las herramientas computacionales siempre ha estado limitada por la capacidad de cómputo y memoria. Sin embargo, ya en los años 70 se identificó que la capacidad de los sistemas computacionales crece también a un patrón exponencial.



La evidencia
acumulada a la
fecha no
contradice esta
observación.

La capacidad de
cómputo del
hardware crece a
ritmo
exponencial.

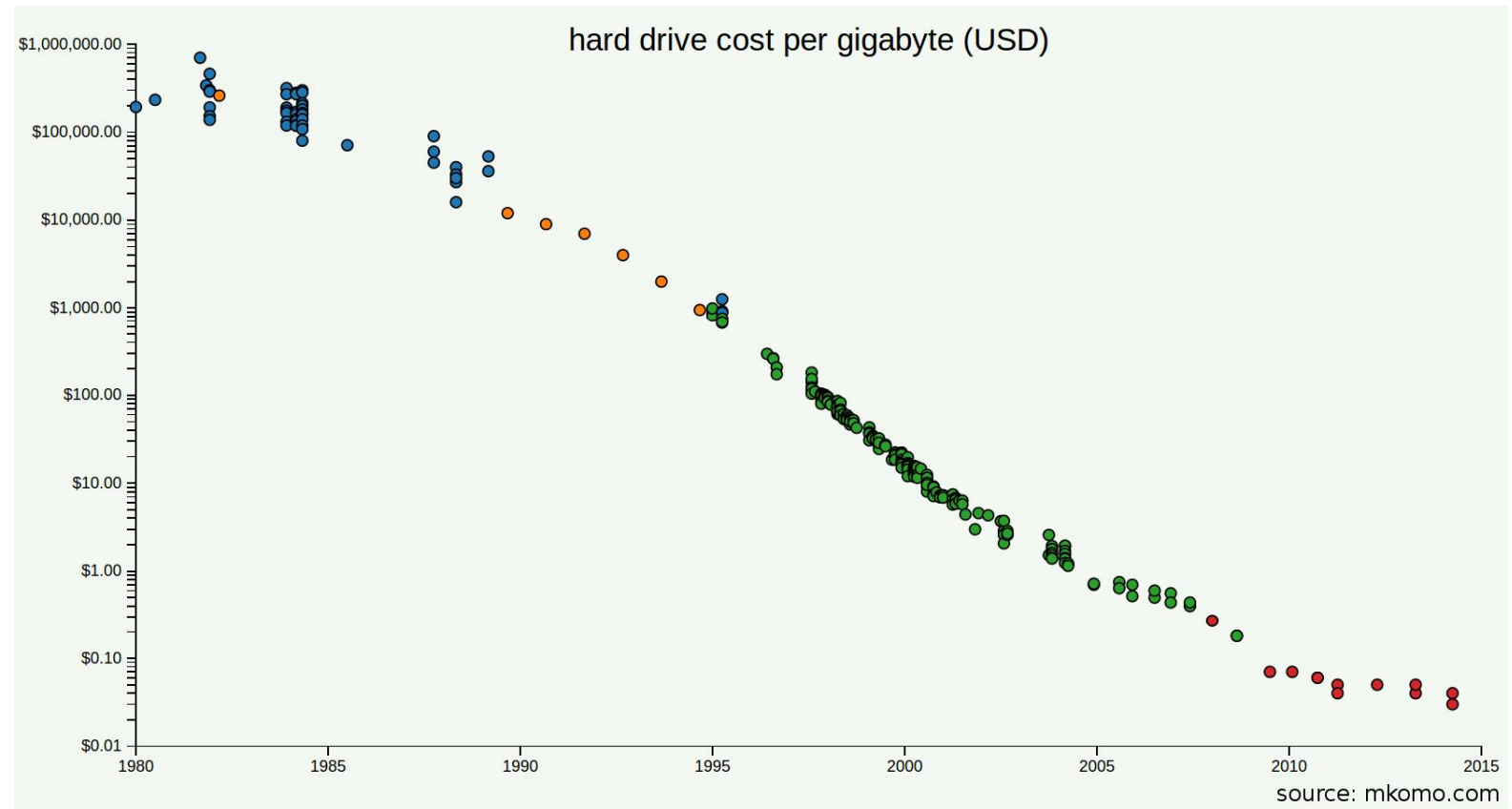
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.



Licensed under [CC-BY-SA](#) by the author Max Roser.

Ley de Moore

El costo de almacenar datos decrece a ritmo exponencial.



El problema de la Información

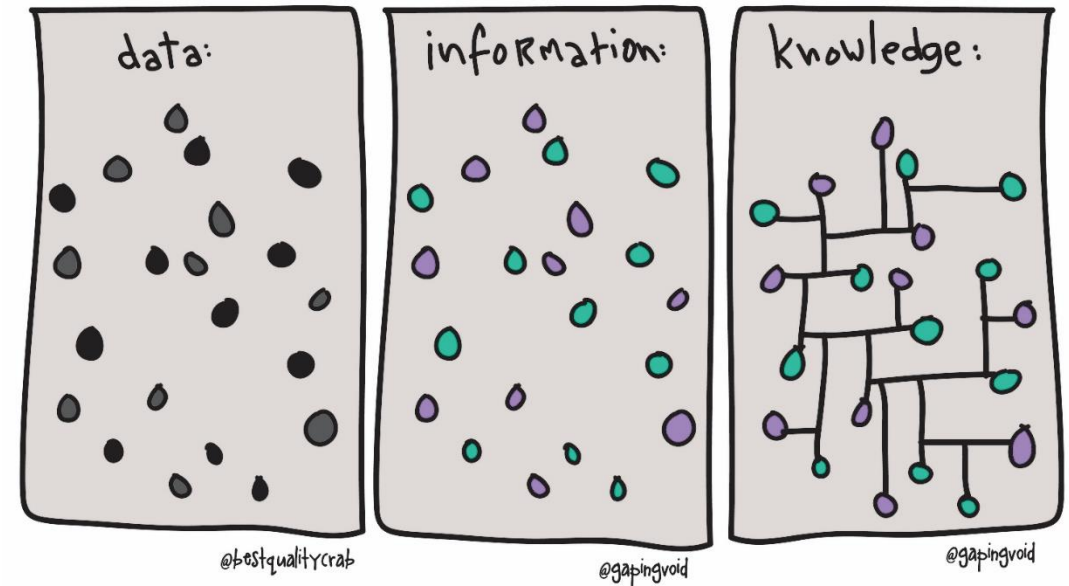
La capacidad de las personas, sin embargo, NO crece a ritmo exponencial. Personas y organizaciones, en todo el mundo, enfrentan la dificultad de administrar y aprovechar volúmenes crecientes de información.



¿Pero qué es la Información?

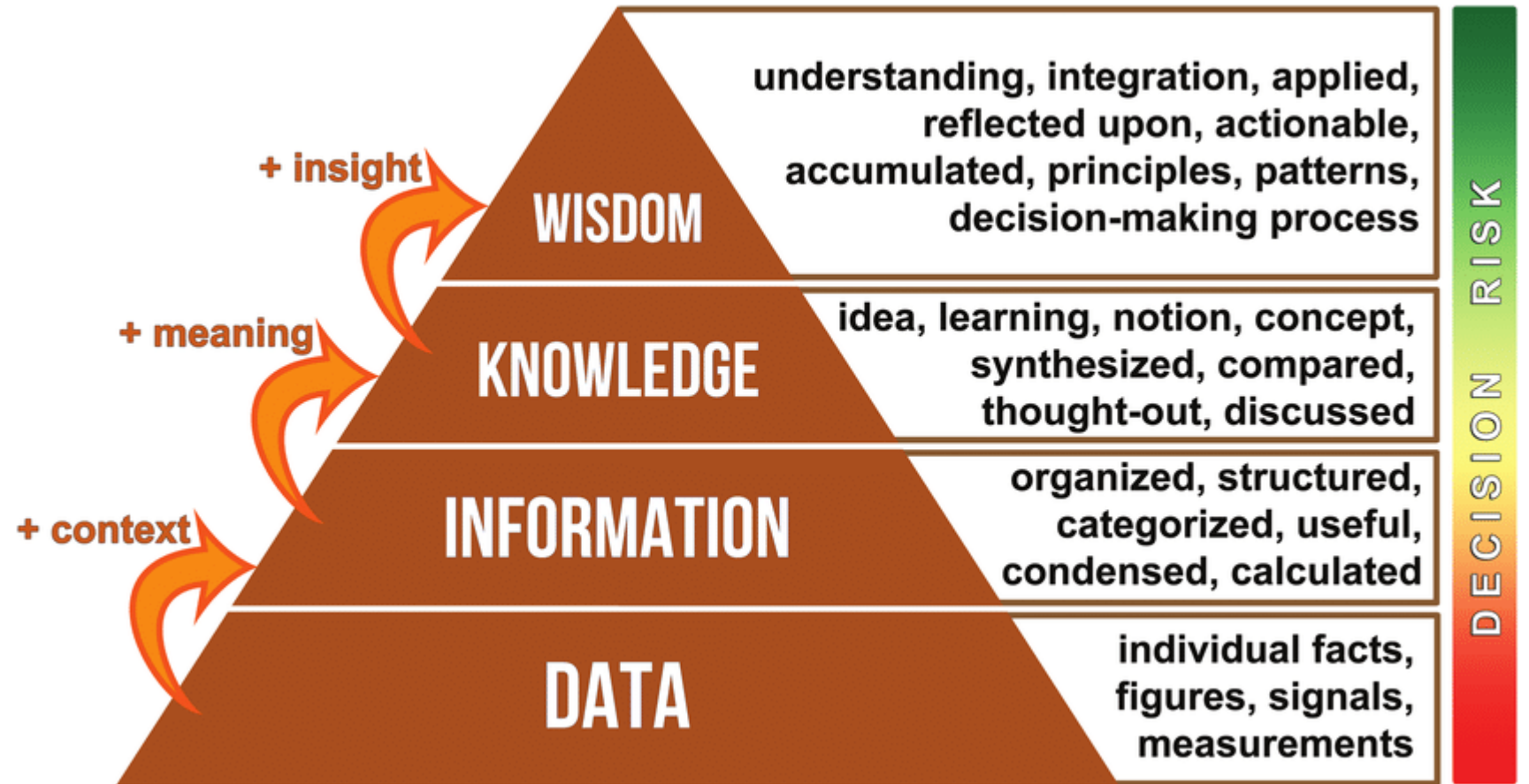
La **información** se obtiene de la organización y contextualización de **datos**, siendo estos últimos: observaciones directas tomadas de la realidad.

La información permite el **conocimiento**, que resulta de interpretación, relación y comprensión de la información.



Datos → Información → Conocimiento → Sabiduría

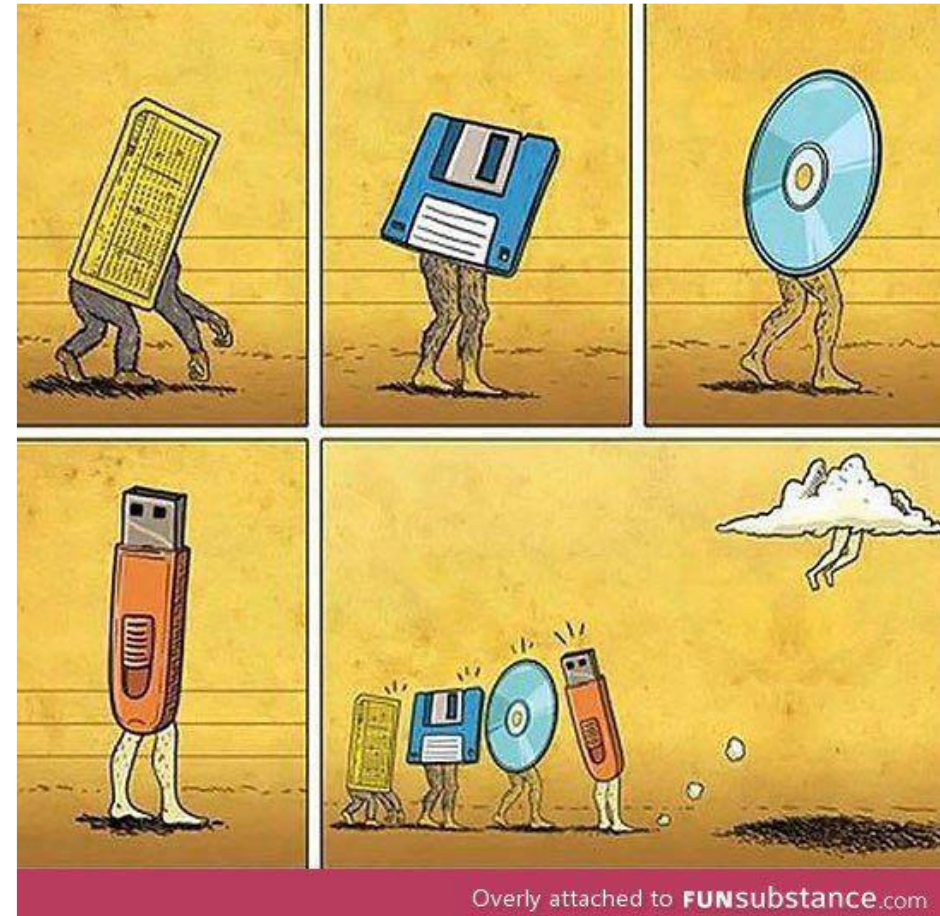
La toma de decisión basada en datos constituye, en última instancia, una fuente de *sabiduría* para las organizaciones.



La base son los datos

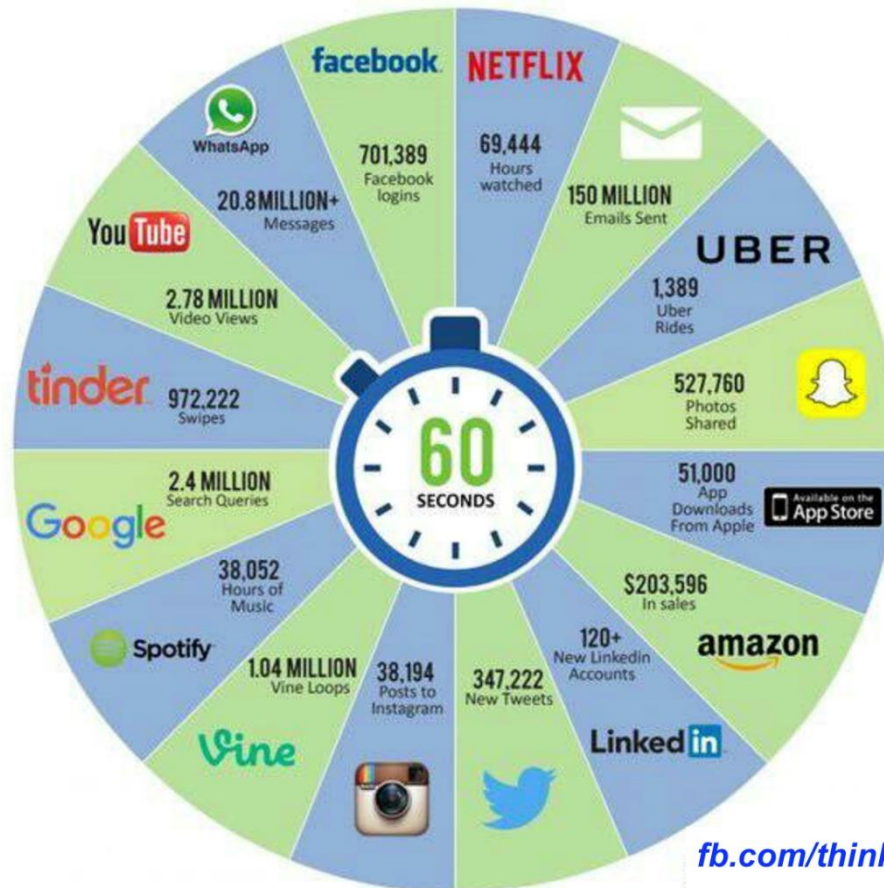
Actualmente nuestros datos se encuentran alojados y respaldados mayoritariamente y crecientemente en la *nube*, donde virtualmente no encontramos límites para seguir acumulando.

¿Qué nuevos desafíos emergen a partir de este cambio?



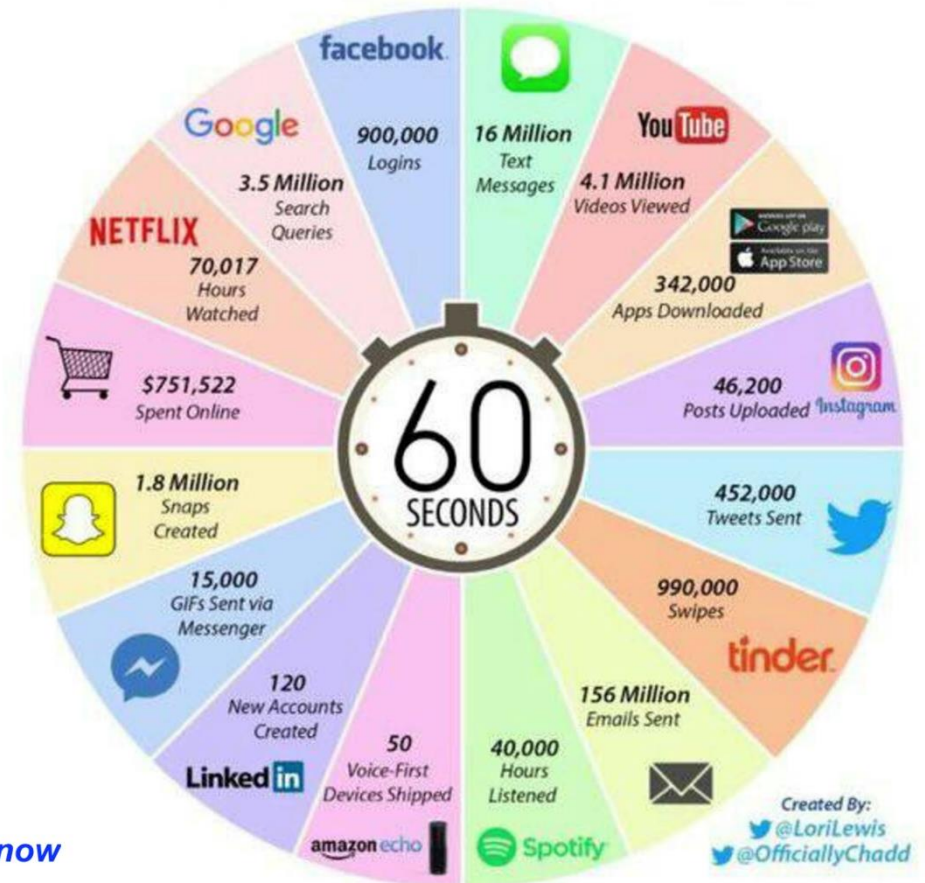
Big Data

2016 What happens in an INTERNET MINUTE?



[fb.com/thinkwittynow](https://www.facebook.com/thinkwittynow)

2017 This Is What Happens In An Internet Minute



Created By:
[@LoriLewis](#)
[@OfficiallyChadd](#)

Big Data

2018 *This Is What Happens In An Internet Minute*



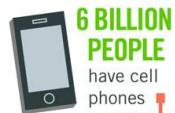
2019 *This Is What Happens In An Internet Minute*



40 ZETTABYTES

[43 TRILLION GIGABYTES]

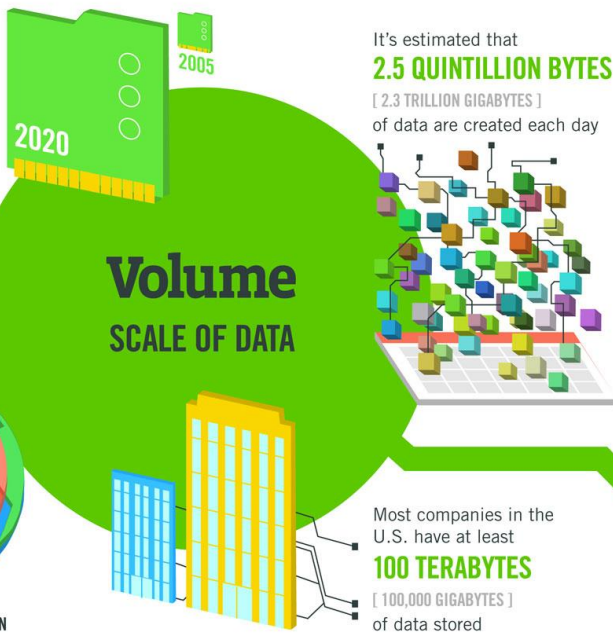
of data will be created by 2020, an increase of 300 times from 2005



6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT

are shared on Facebook every month



Variety
DIFFERENT FORMS OF DATA

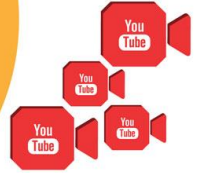
By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS



4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



400 MILLION TWEETS

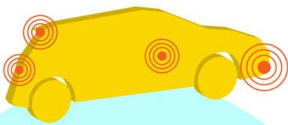
are sent per day by about 200 million monthly active users



The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to **100 SENSORS**

that monitor items such as fuel level and tire pressure

Velocity
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

— almost 2.5 connections per person on earth



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



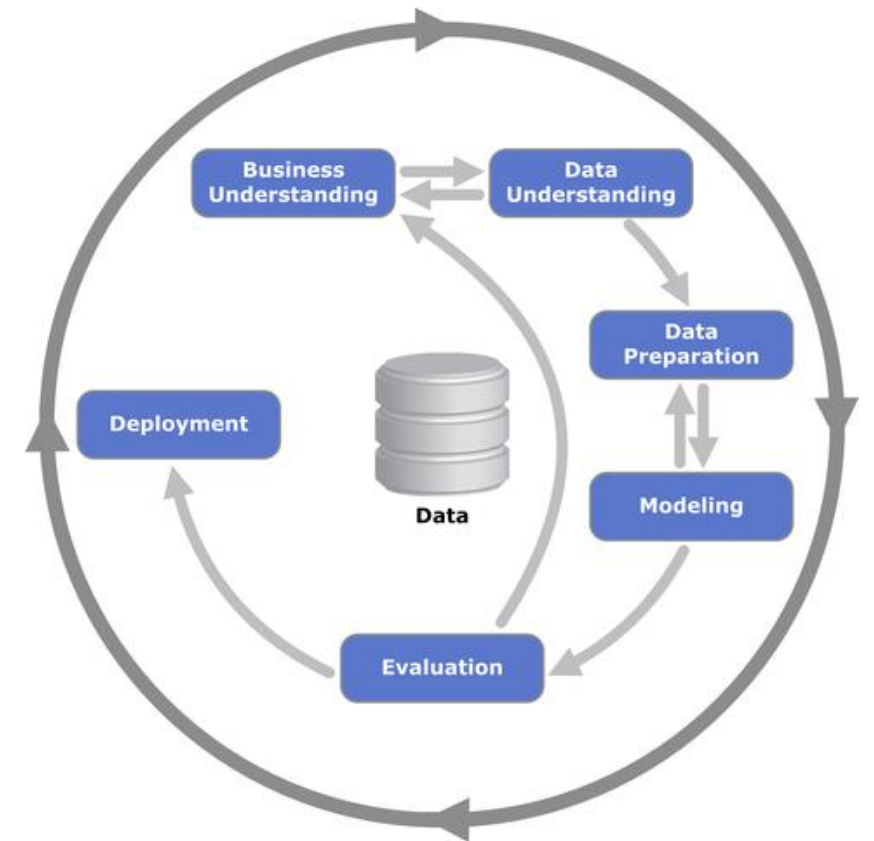
in one survey were unsure of how much of their data was inaccurate

Veracity
UNCERTAINTY OF DATA

Data Mining

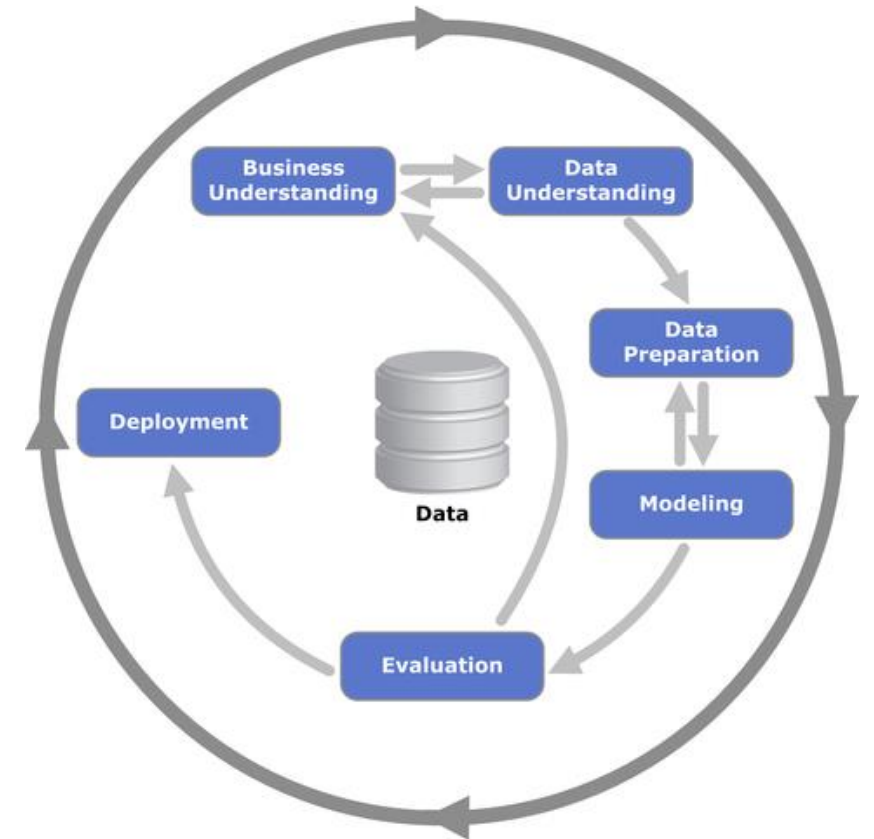
Durante los años 80 surge el concepto **Minería de Datos (Data Mining)**, consistente en la práctica de buscar extraer información a partir de los datos. Se distinguió de las estadísticas por no ser guiado por hipótesis, sino directamente por los datos, algo mal visto por la comunidad estadística hasta años recientes.

Cross-industry Standard Process for Data Mining (CRISP-DM) es un estándar propuesto a partir de la experiencia levantada en la industria (1997), y según estudios recientes, sigue siendo uno de los modelos de Minería de Datos más comúnmente utilizados.



Data Mining

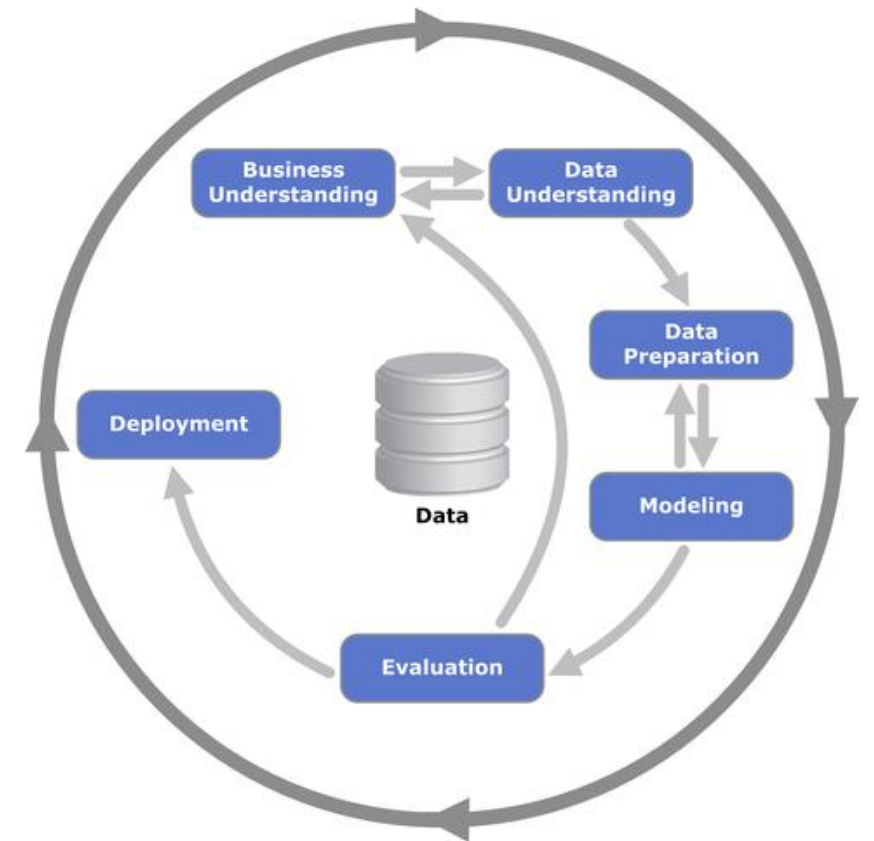
- Comprensión de negocio.
- Comprensión de los datos.
- Preparación de los datos.
- Modelamiento
- Evaluación
- Despliegue



Data Mining

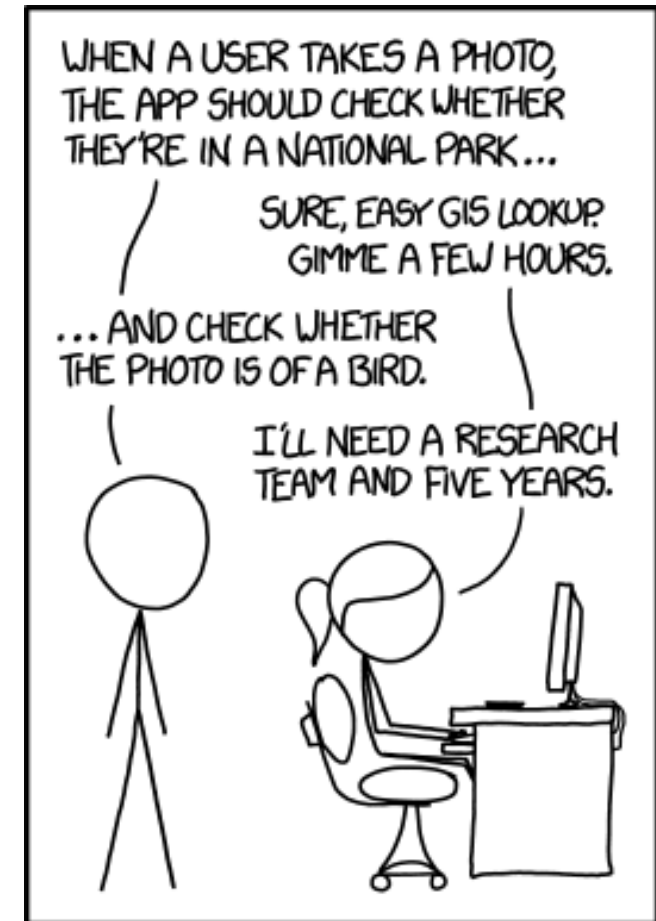
Las fases propuestas en CRISP-DM corresponden a las más frecuentemente observadas en la industria. Estas actividades no se ejecutan en forma secuencial estricta, al contrario, la metodología asume que habrá distintos posibles caminos, identificando los que tienen mayor probabilidad.

Por ejemplo, la evaluación de resultados obtenidos en la etapa de modelamiento puede derivar en su despliegue o derivar de vuelta a la etapa de entendimiento del negocio.



Machine Learning

La fase de modelamiento de un proceso de Minería de Datos puede ser, en realidad, una composición de distintos tipos de análisis y actividades. Los algoritmos y modelos más avanzados que disponemos, cuyo desarrollo y uso están en pleno auge, están clasificados en general en lo que se denomina **Machine Learning**. Acá encontraremos los modelos predictivos de avanzada. Es también el motor de la Inteligencia Artificial.



IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.

Machine Learning

Nota al margen: Hace pocos años este chiste tenía sentido. El avance del campo Ciencia de Datos ha hecho que lo que el requerimiento expuesto por el cliente ya no sea *virtualmente imposible*, y hasta resulta relativamente fácil de implementar.

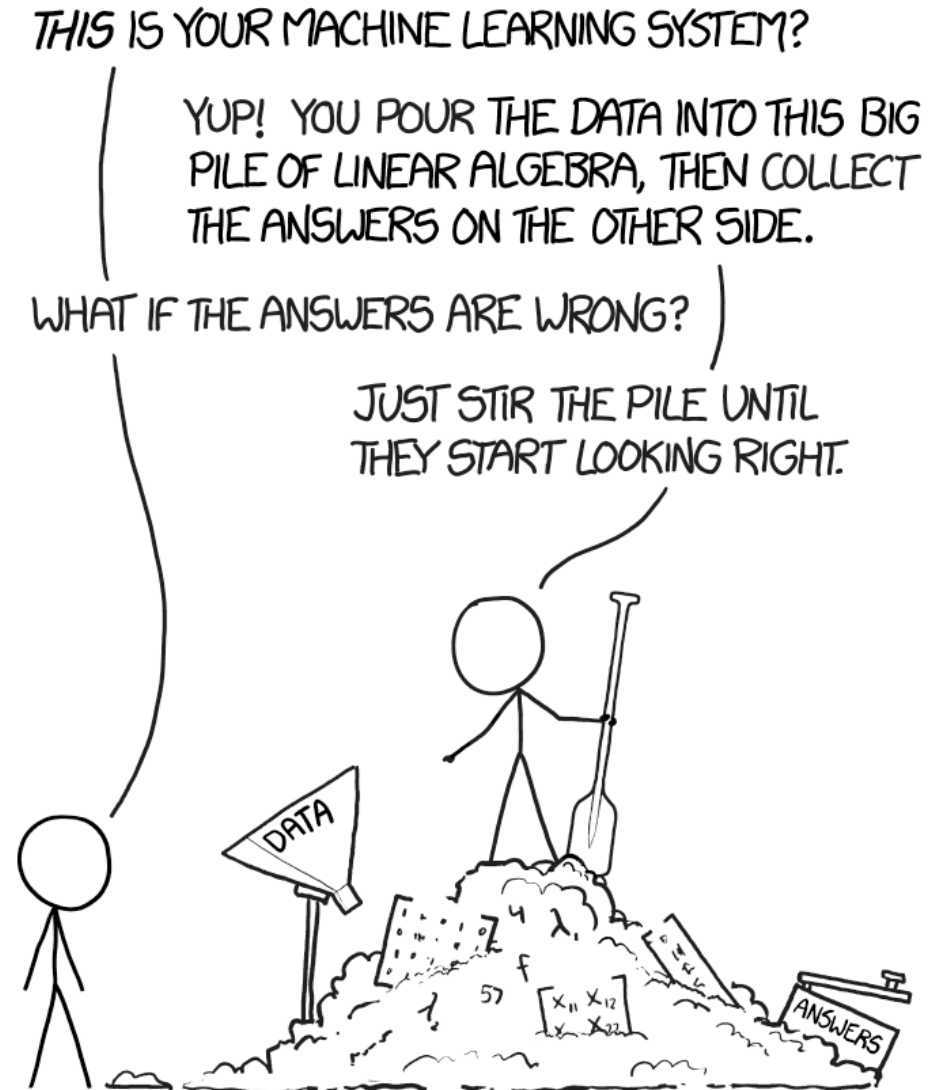


IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.

Machine Learning

Literalmente, son máquinas diseñadas para aprender a partir de los datos, para extraer información valiosa en la resolución de un problema, en forma automática.

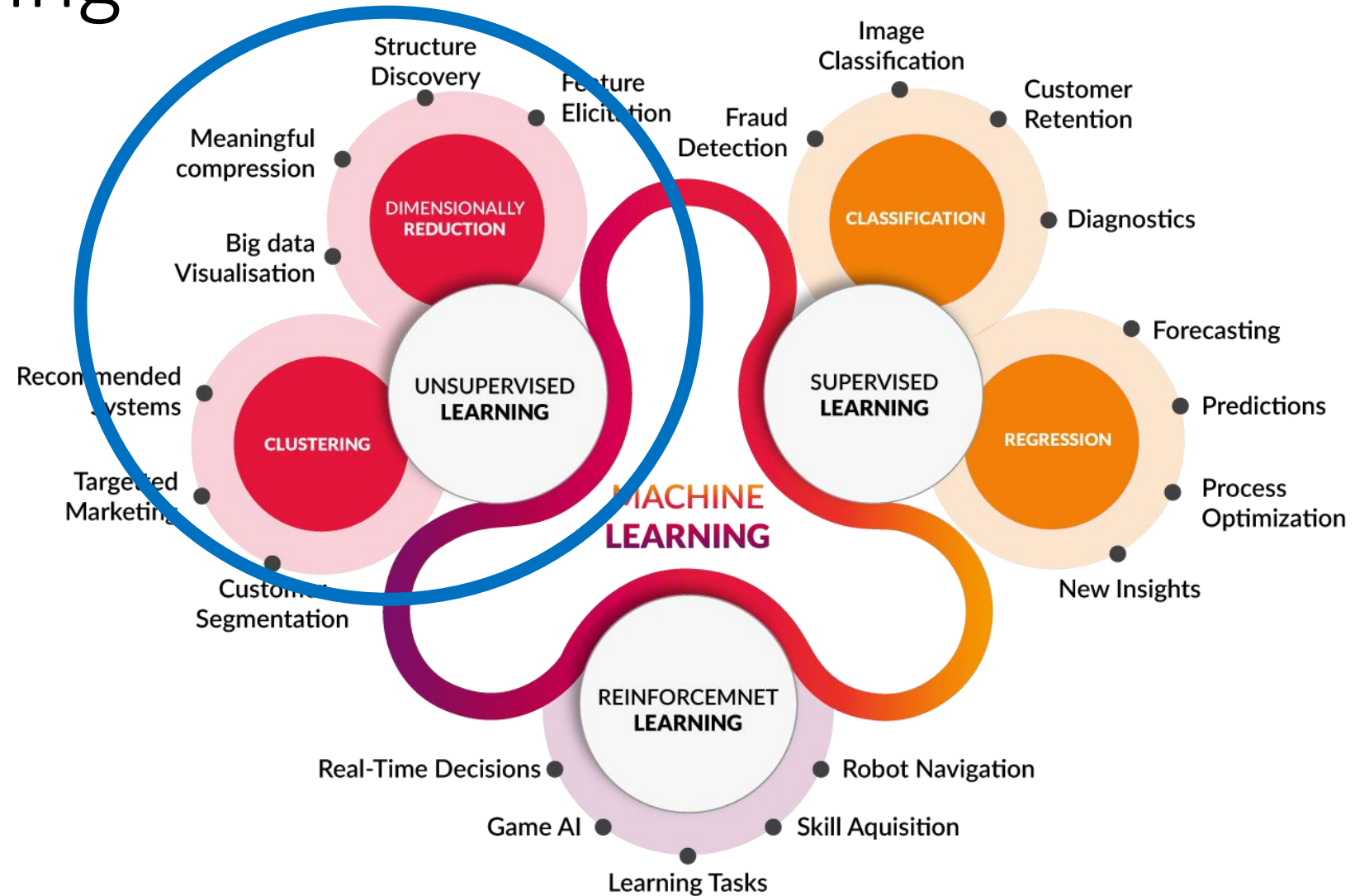
Consiste en implementar modelo matemático y entrenarlos a partir de datos obtenidos de observaciones de la realidad.



Machine Learning

La clasificación clásica de los algoritmos de ML distingue: Supervisado de No-Supervisado.

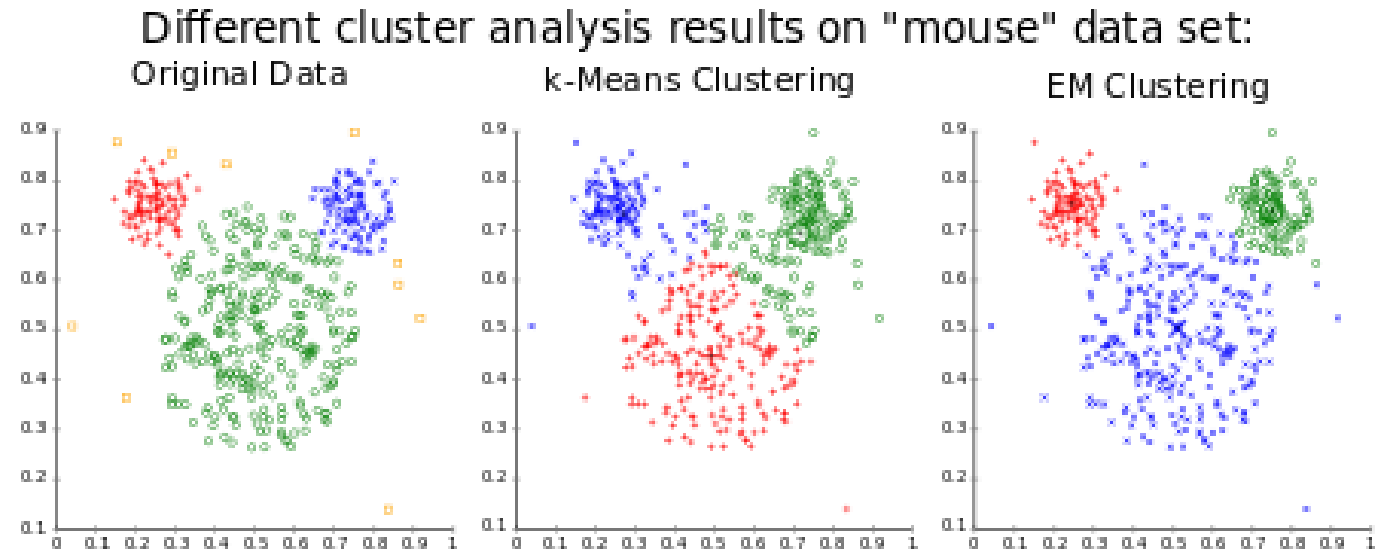
Los algoritmos No supervisados permiten levantar patrones y tendencias a partir de los datos.



Aprendizaje No Supervisado

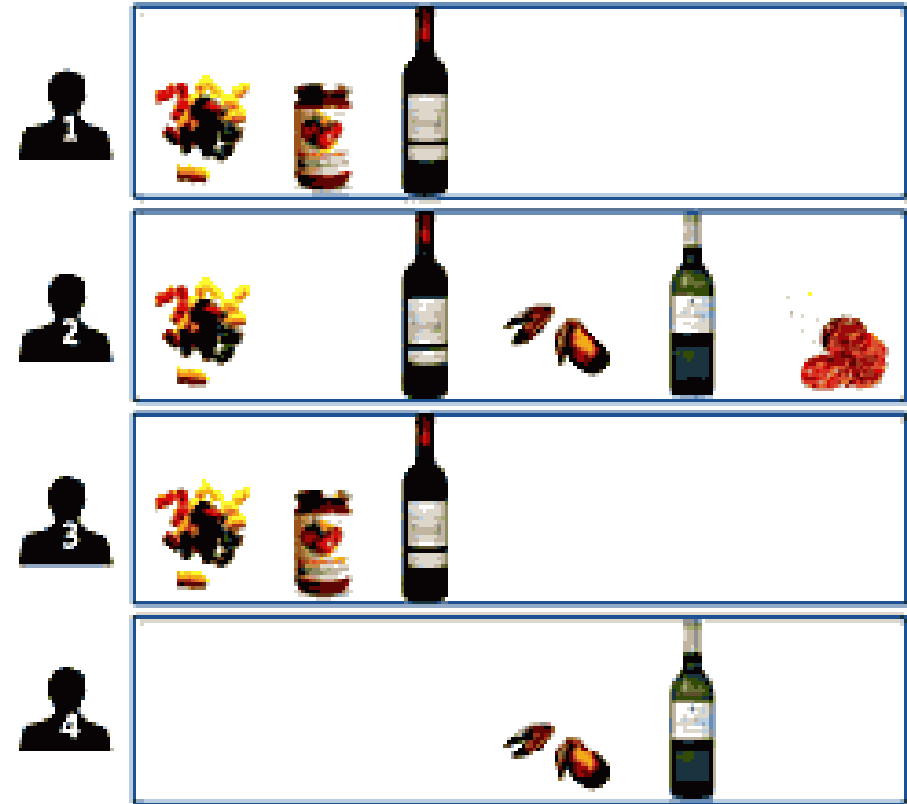
Clustering:

Consiste en agrupar las observaciones de acuerdo a similitud. Existe varias familias de modelos de este tipo, según que se entiende por similitud, a partir de lo cual se define alguna función de distancia.



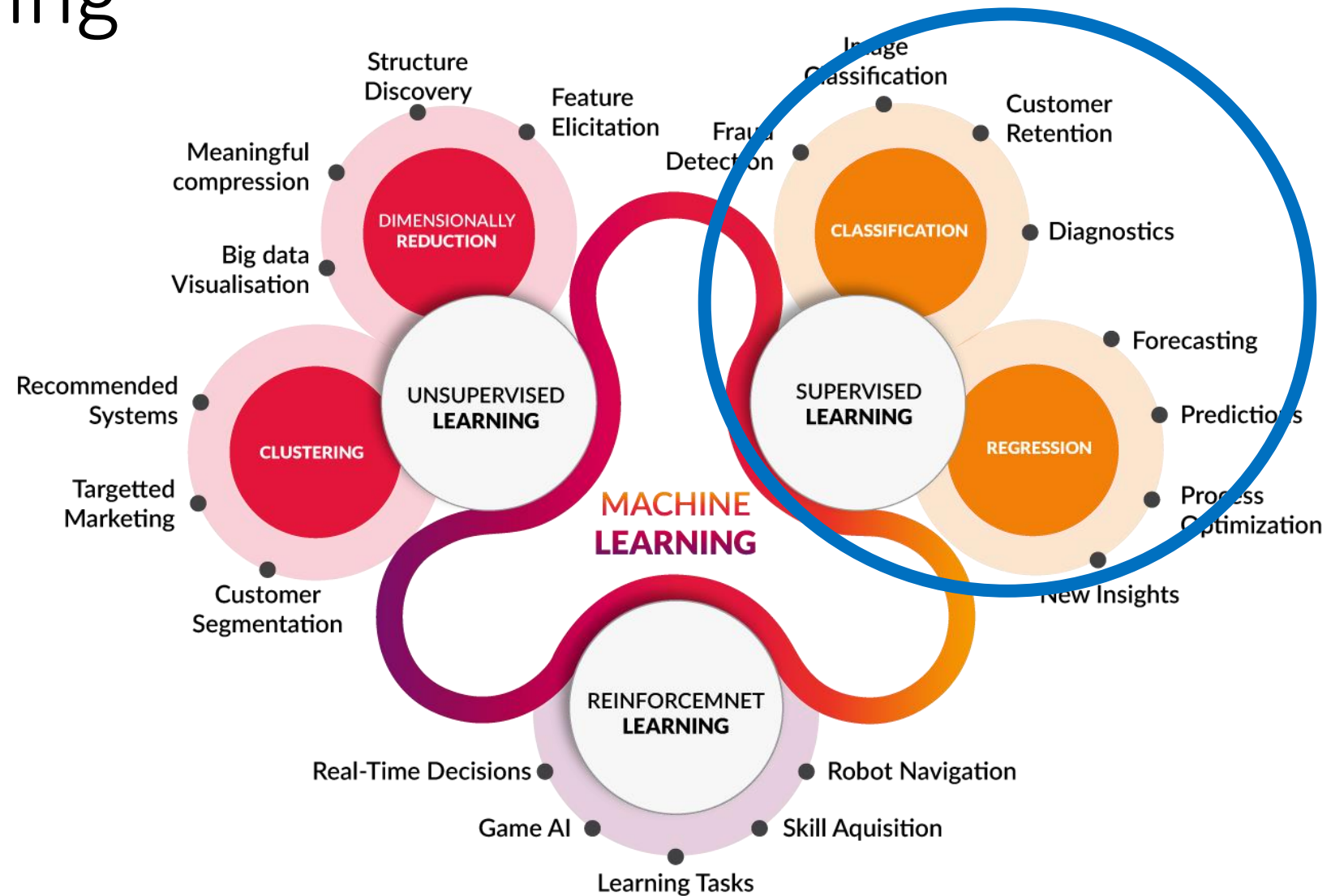
Aprendizaje No Supervisado

Reglas de Asociación:
Método para descubrir relaciones de interés en grandes sets de datos. Se basan en identificar reglas fuertes que maximicen alguna métrica de interés (confianza, soporte, elevación y convicción)



Machine Learning

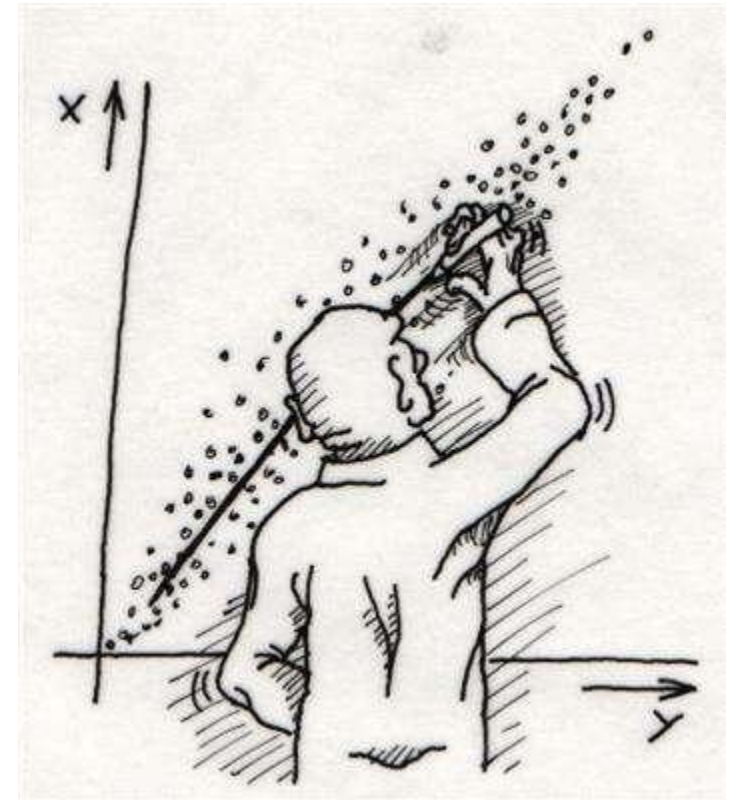
El aprendizaje supervisado se entrena en base a observaciones previamente etiquetadas.



Aprendizaje Supervisado

Se distinguen dos grandes clases:
Regresión y Clasificación.

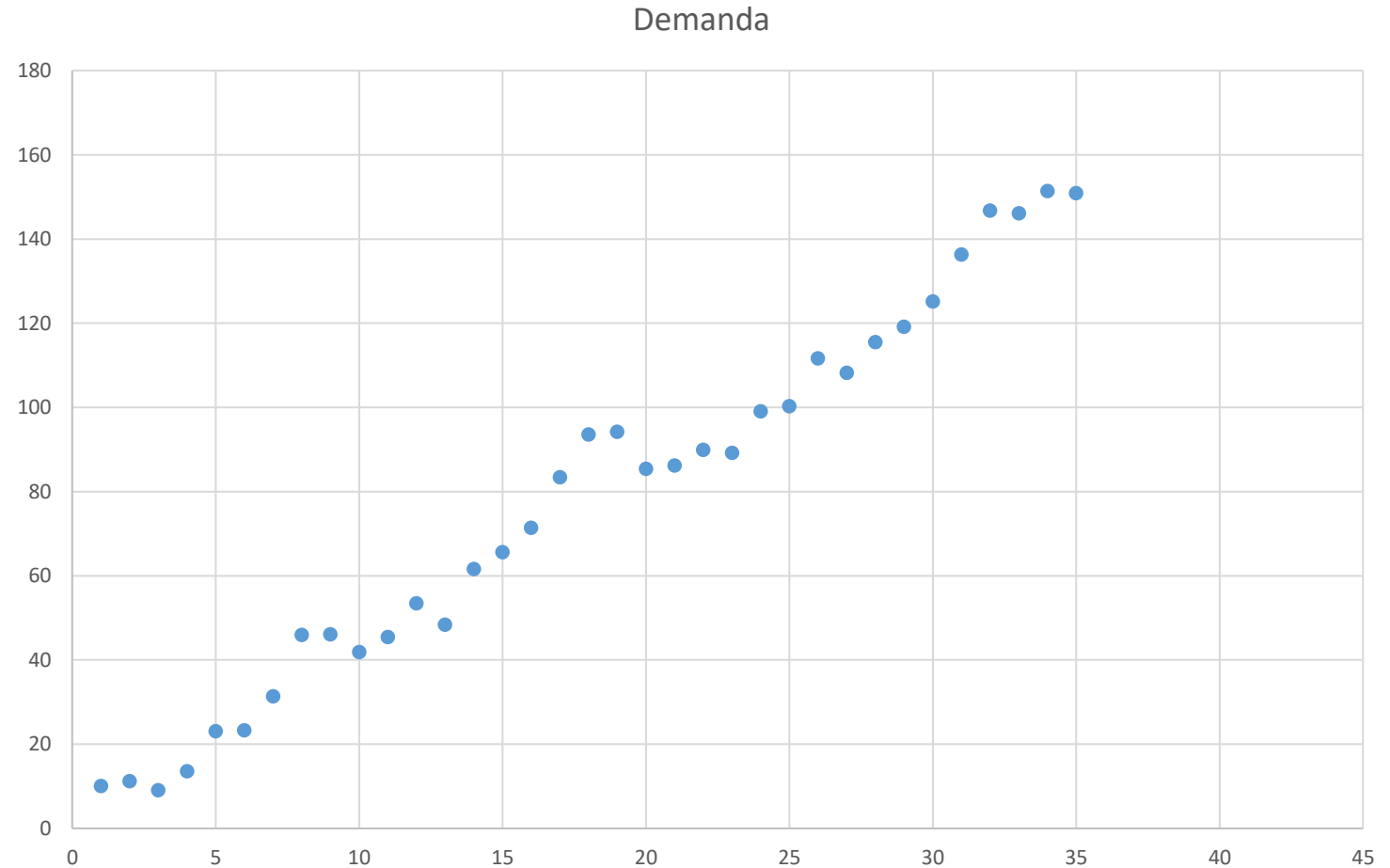
Regresión : Directamente estima la esperanza condicional de una variable dependiente, a partir del valor de una o varias variables independientes. Involucra asumir la forma de dicha relación, lo más común es lineal.



Paréntesis Estadístico

$$\hat{y} = \beta_0 + \beta_1 x$$

Regresión Lineal:
Consiste en construir un modelo lineal que estima el valor de una variable dependiente y , a partir de una variable independiente x .

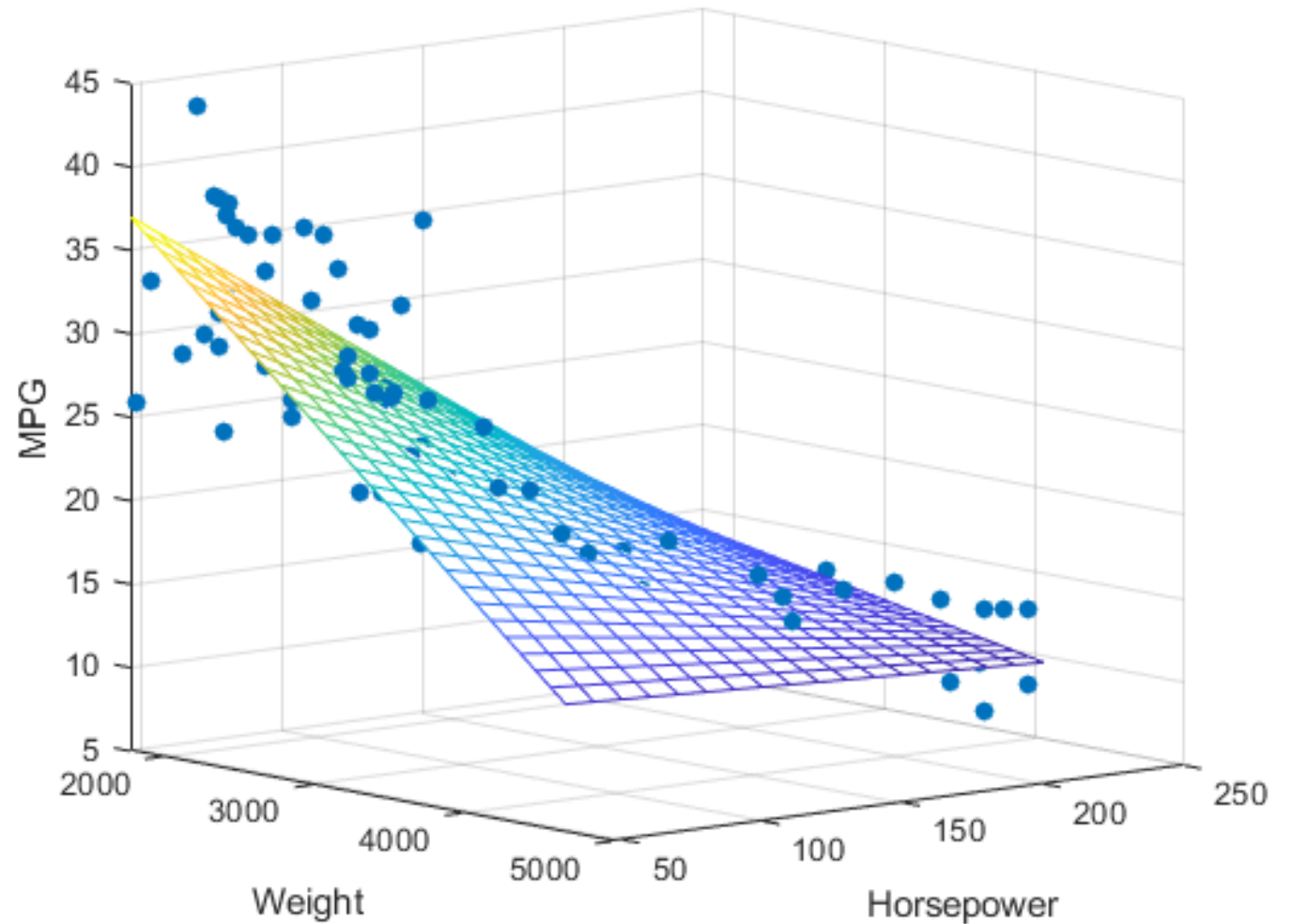


Paréntesis Estadístico

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

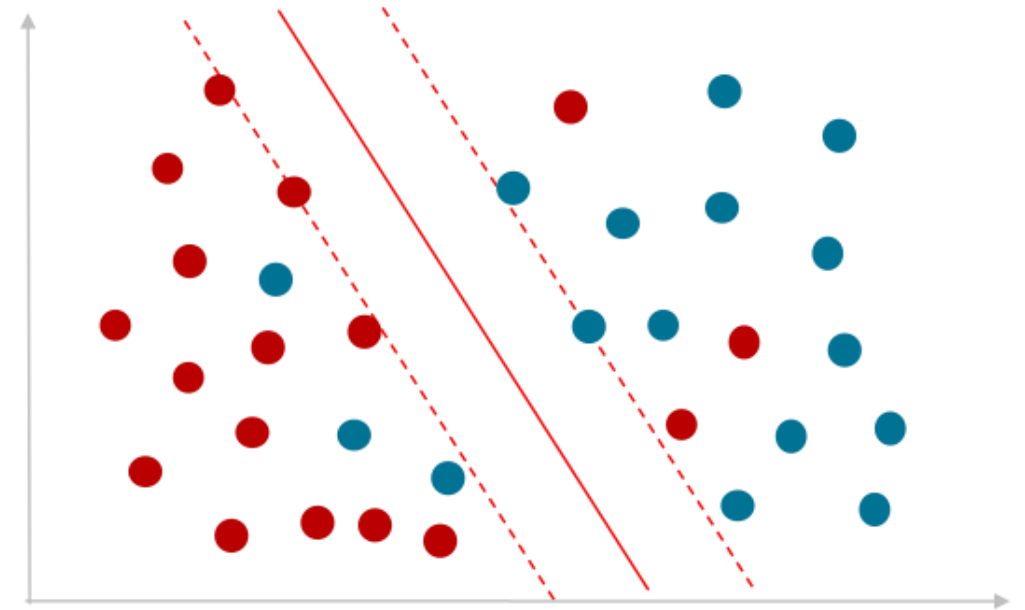
Regresión Lineal:

Puede ser una
variable
independiente, o
bien muchas
variables
independientes.



Aprendizaje Supervisado

Clasificación : Consiste en determinar a qué categoría pertenecen las observaciones, a partir de sus atributos. Es uno de los campos más versátiles y desarrollados de *Machine Learning*, existe una muy amplia variedad de modelos diseñados para clasificación. Al ser supervisado, es necesario contar con una muestra etiquetada para entrenar.

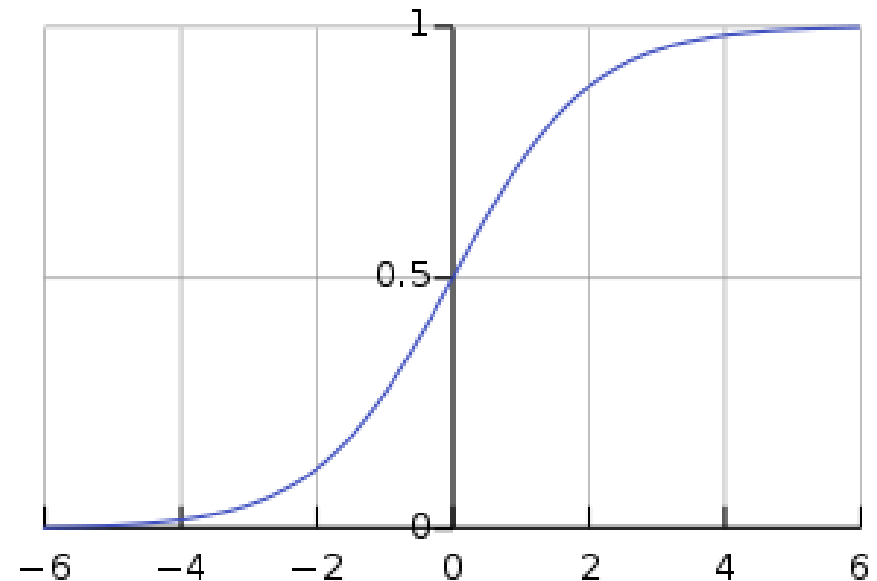


Paréntesis Estadístico

Regresión Logística:

Una primera aproximación al problema de clasificación, de mecánica similar a la regresión lineal, pero adoptando la función logística.

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

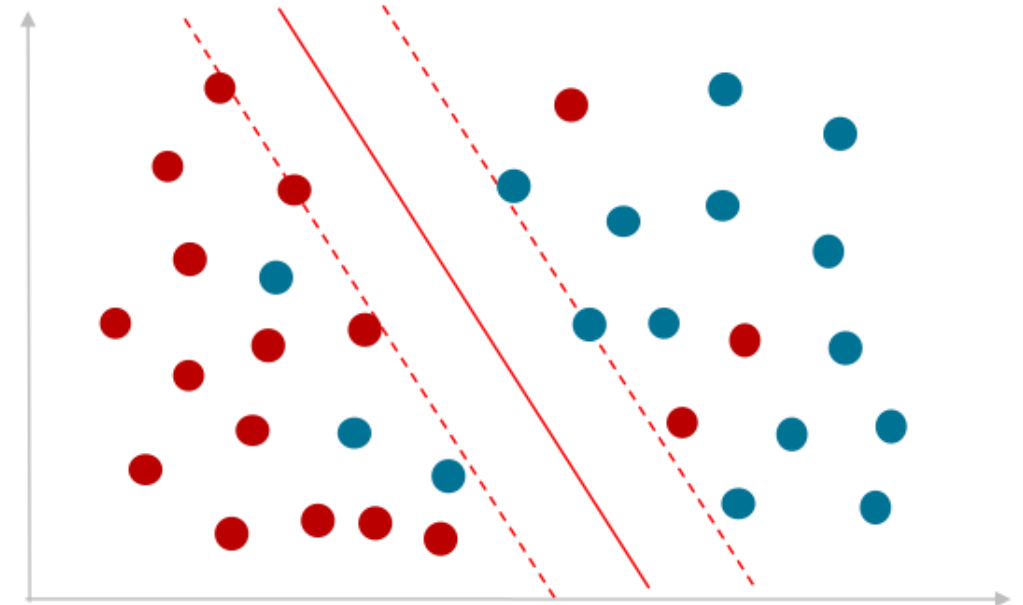


Paréntesis Estadístico

Regresión Logística:

El resultado es una frontera lineal que divide las observaciones, en dos grupos, tratando de separar ambas categorías lo mejor posible.

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

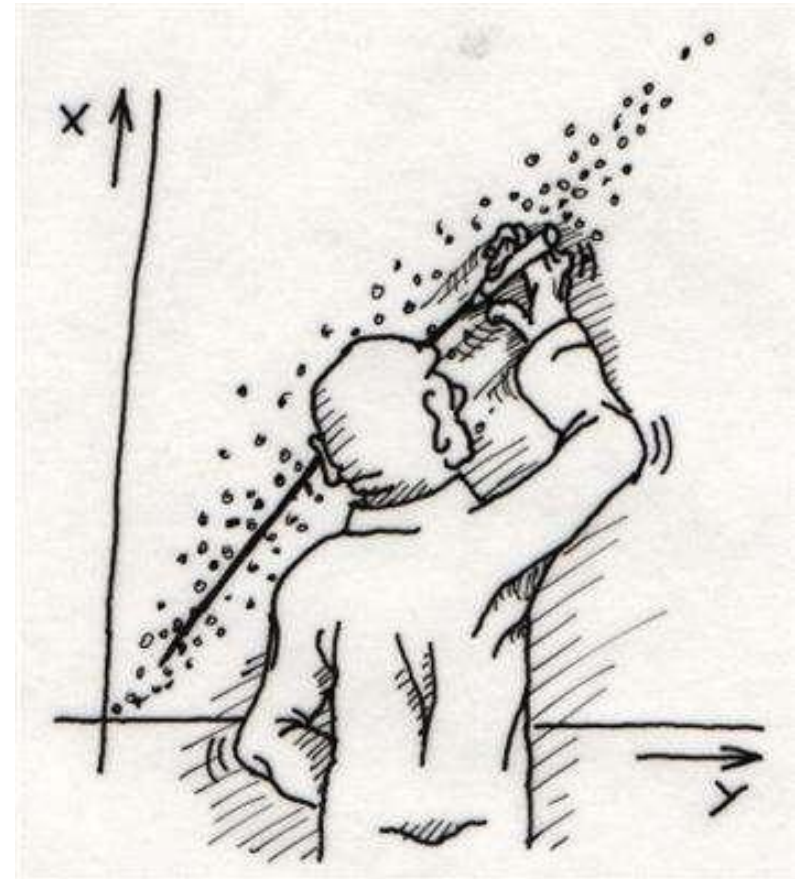


Paréntesis Estadístico

¿Cómo logran estos modelos predecir la variable dependiente y a partir de las variables independientes x ?

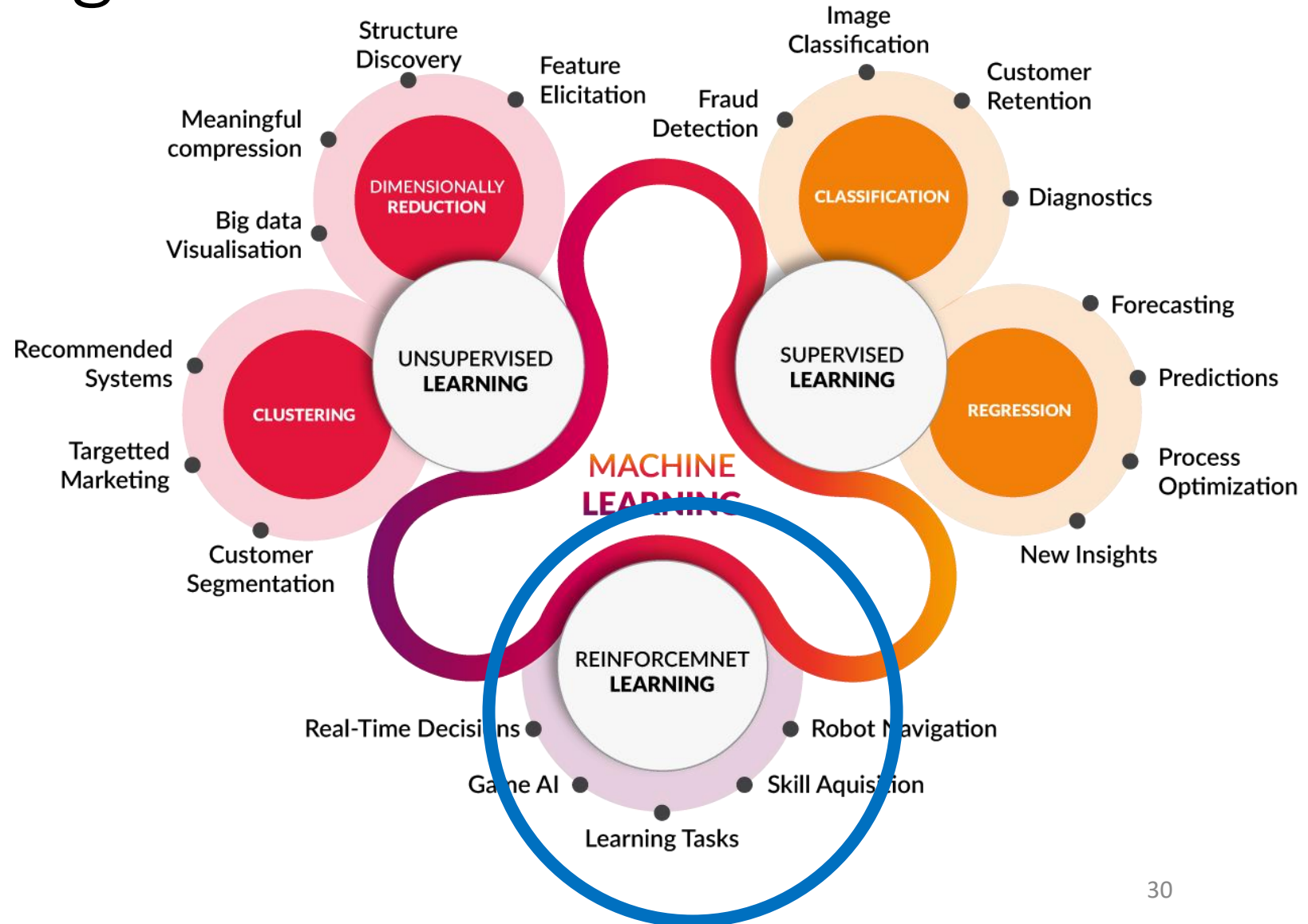
El proceso de alimentar el modelo es usar observaciones conocidas de (x, y) , para definir el valor de los parámetros β

$$\hat{y} = \beta_0 + \beta_1 x_1$$



Machine Learning

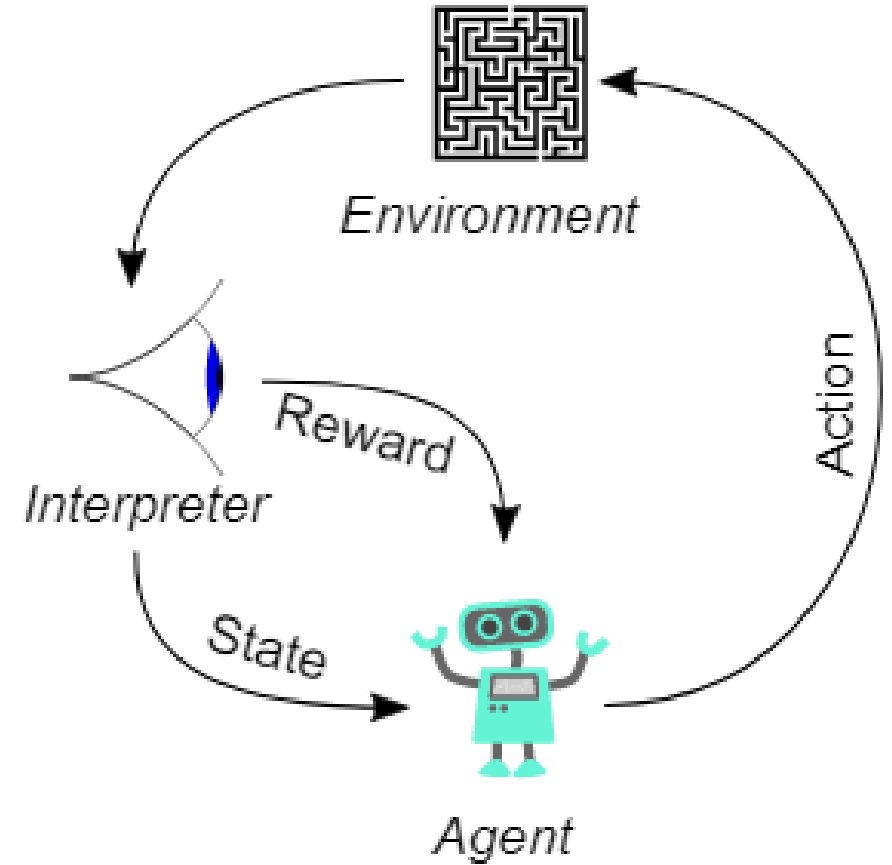
El aprendizaje reforzado se obtiene por algoritmos que realizan acciones en un medio ambiente, evalúan sus resultados y ajustan su comportamiento para mejorar.



Aprendizaje Reforzado

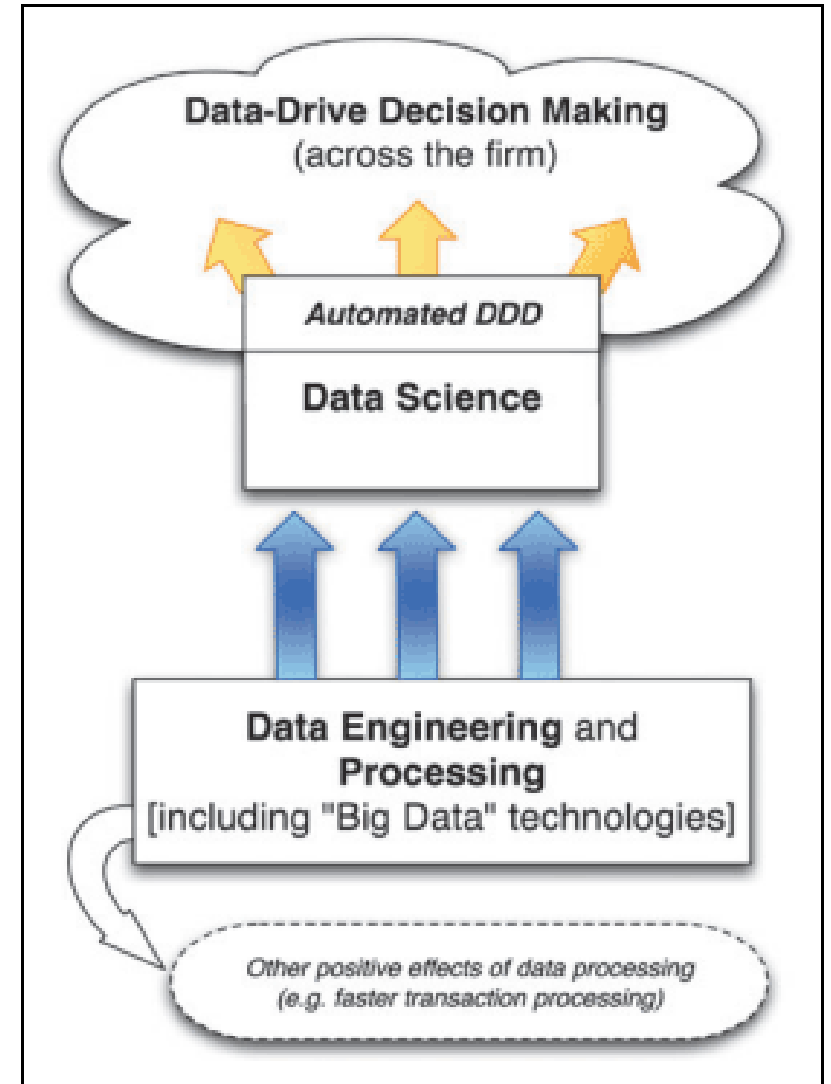
Emerger como una tercera categoría de aprendizaje. Se caracteriza por modelos que aprenden a tomar decisiones en el contexto de tareas cuyo éxito (o fracaso) usan como retroalimentación.

Es una especie de aprendizaje supervisado, pero sin supervisión externa. Es el aprendizaje que más relacionamos con ***Inteligencia Artificial***.



Data-Driven Decision Making

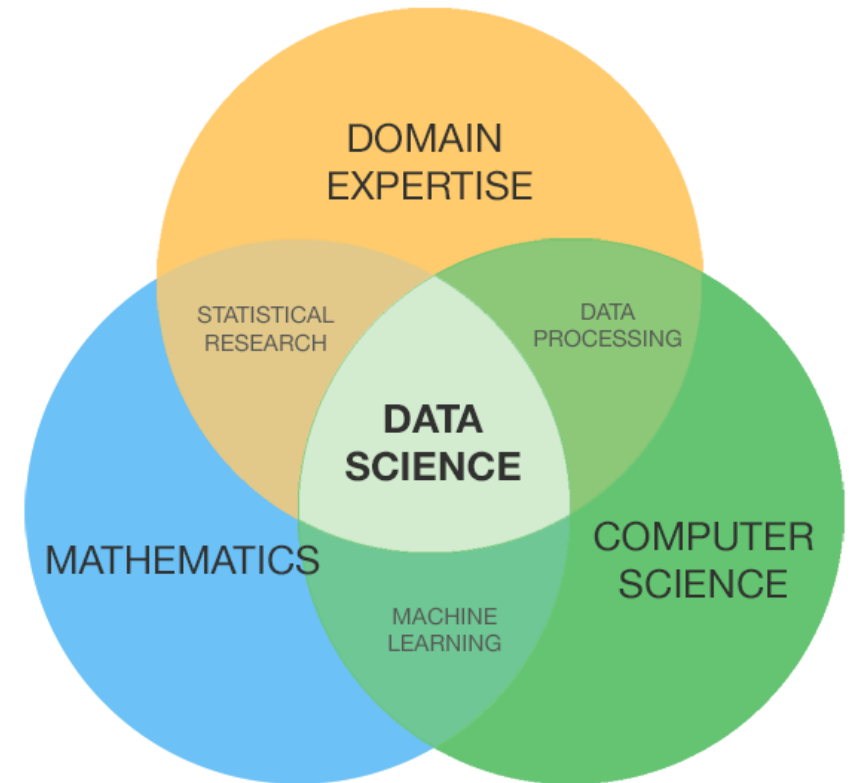
Las organizaciones crecientemente están basando su proceso de decisión en datos levantados de la realidad de sus procesos, entorno competitivo y macro-entorno. El motor analítico, que hace posible la toma de decisión, corresponde a todas las tecnologías que en conjunto constituyen lo que se conoce como **Ciencia de Datos** (*Data Science*).



Data Science

La Ciencia de Datos es un campo multidisciplinario, que a través de diversos métodos, procesos y algoritmos busca la extracción de conocimiento a partir de datos estructurados y no estructurados.

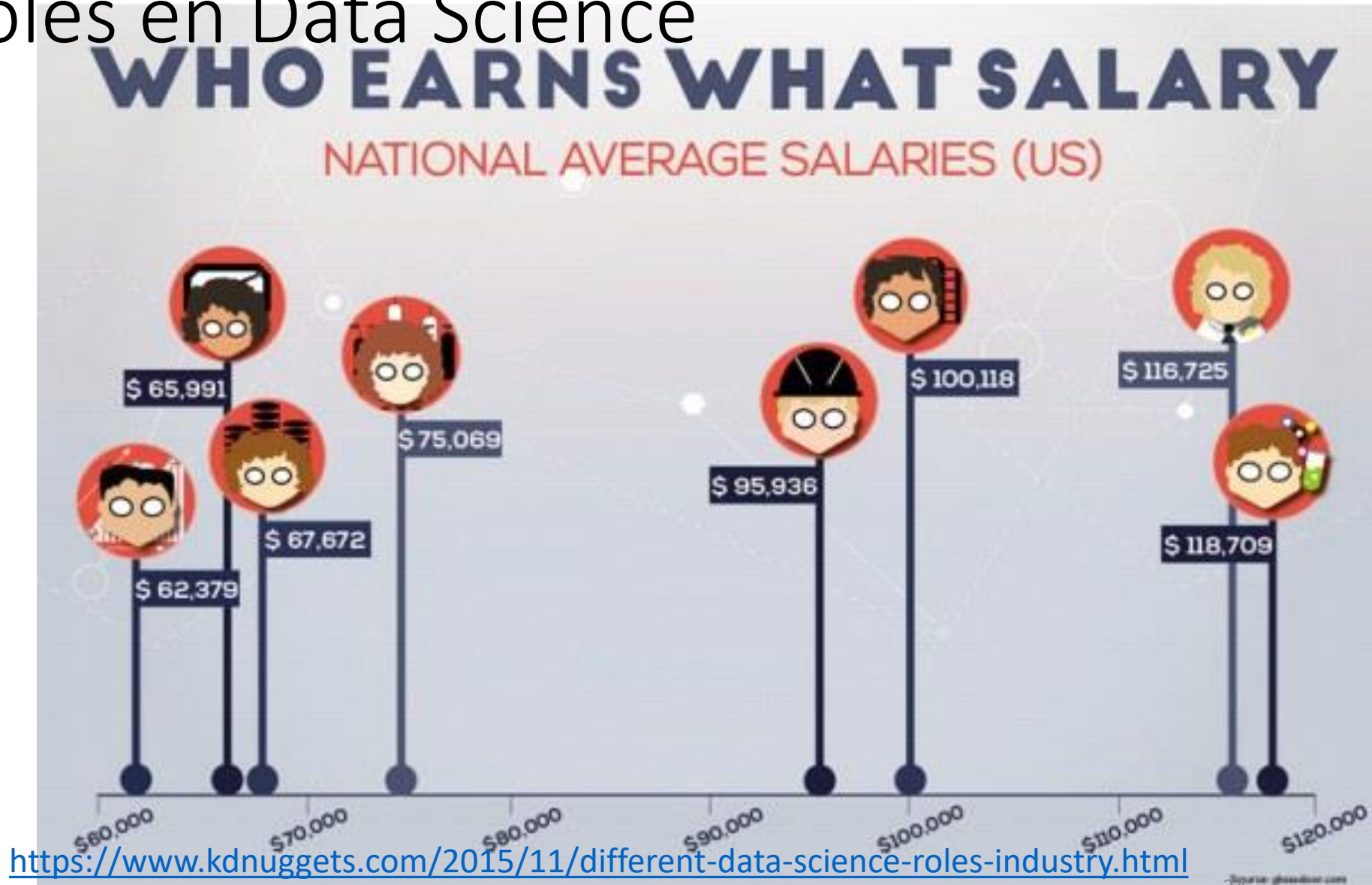
Unifica conceptos previos como Minería de Datos, Análisis de Datos, Análisis Estadístico, Machine Learning.



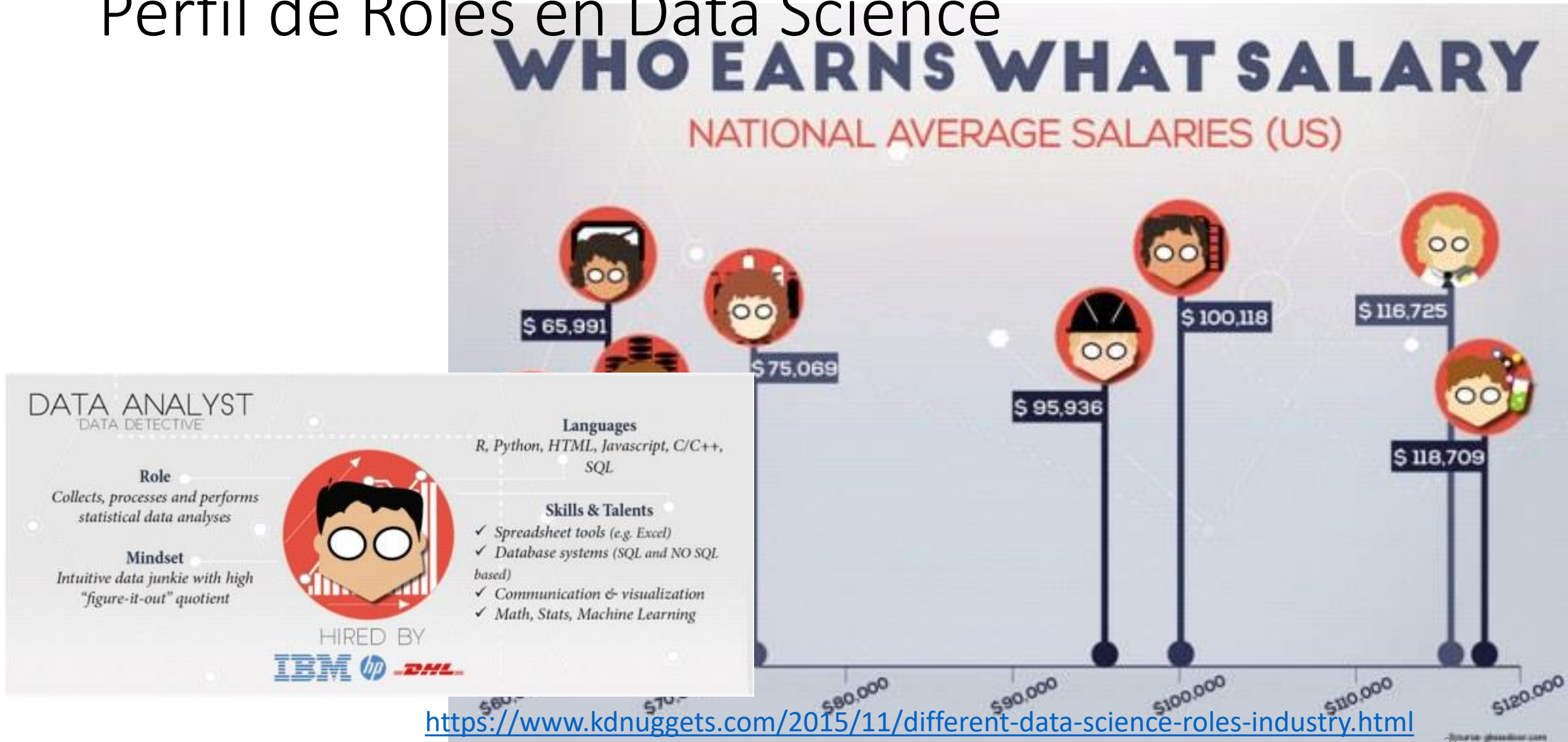
*Source: Palmer, Shelly. Data Science for the C-Suite.
New York: Digital Living Press, 2015. Print.*

Perfil de Roles en Data Science

Las ciencias de datos se han convertido en un dominio más dentro del ámbito de la ingeniería en informática (y similares).

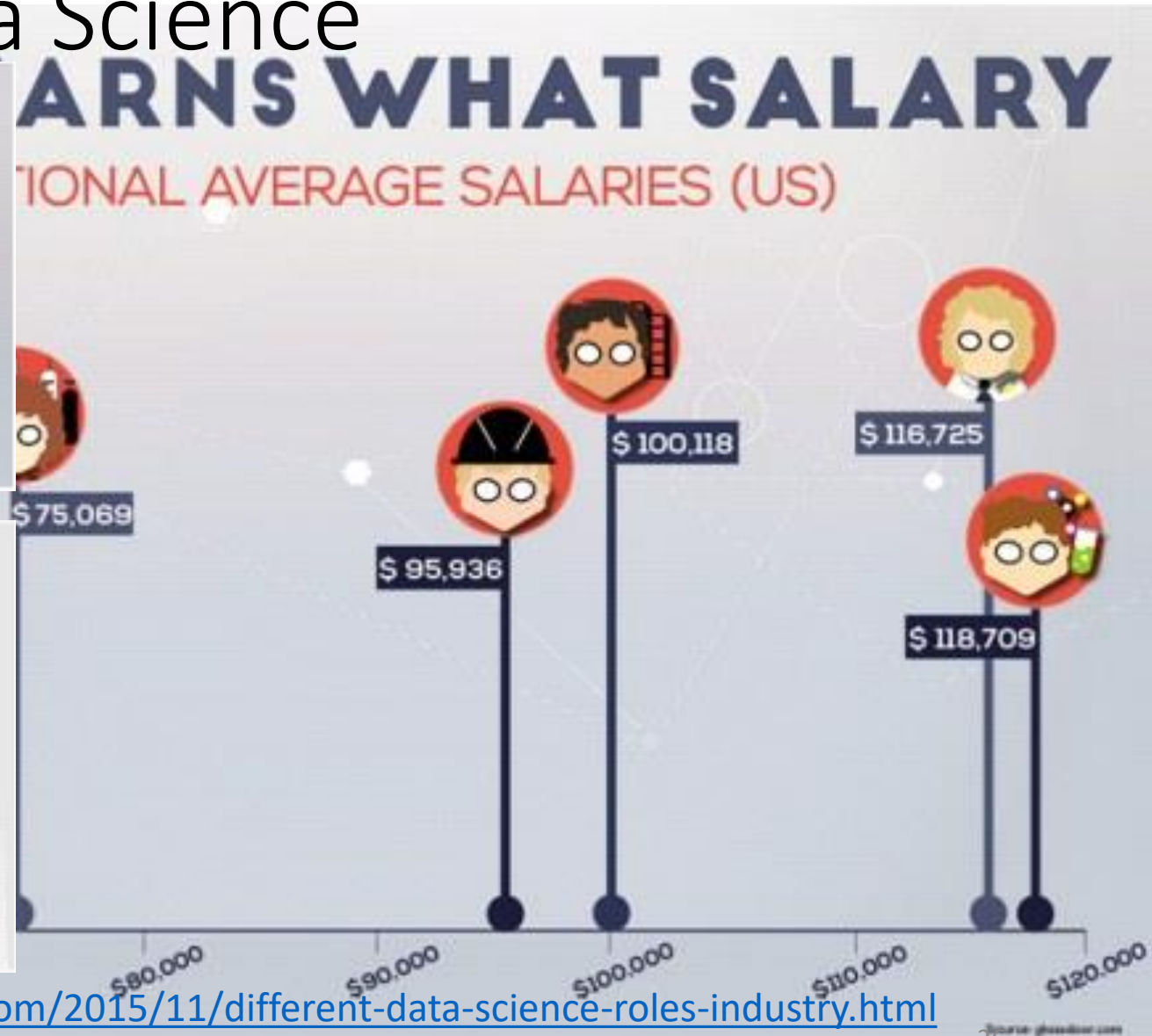
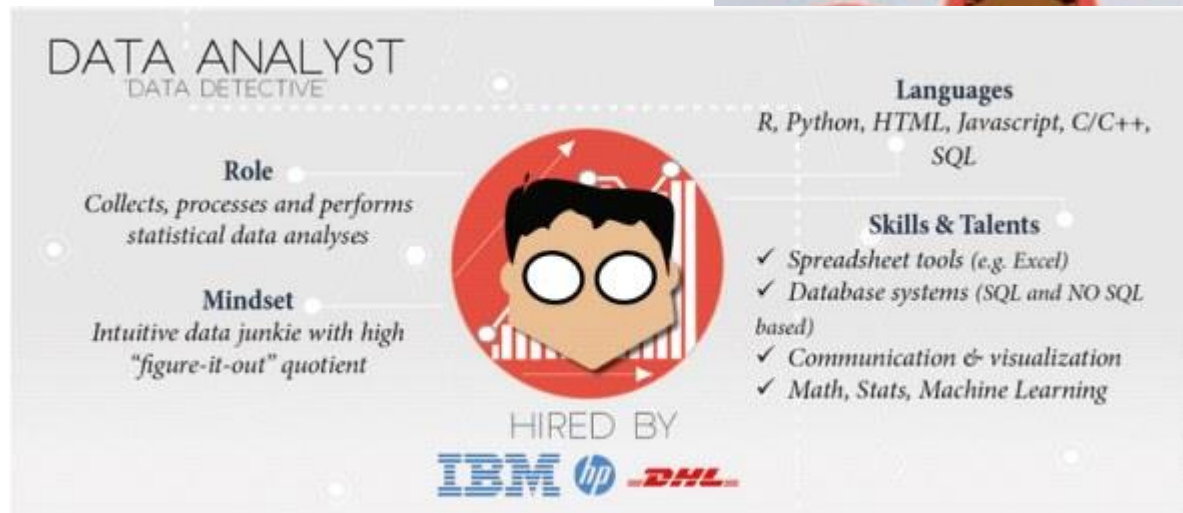


Perfil de Roles en Data Science



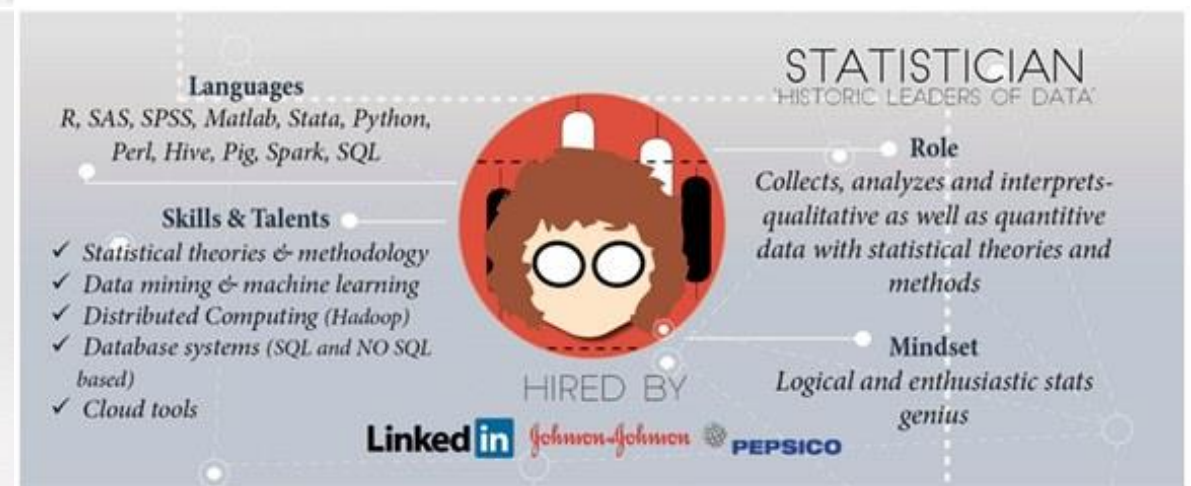
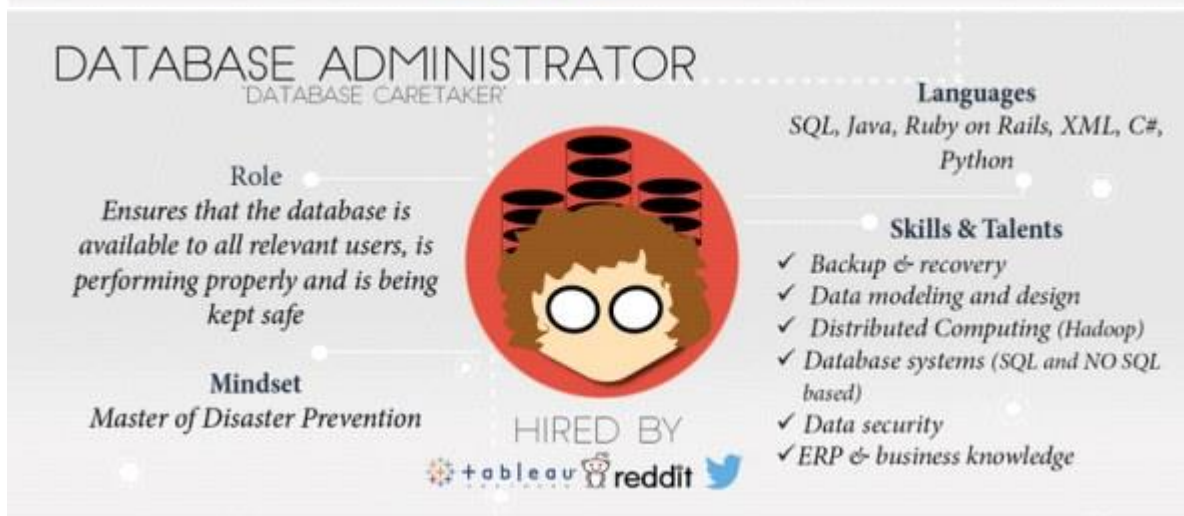
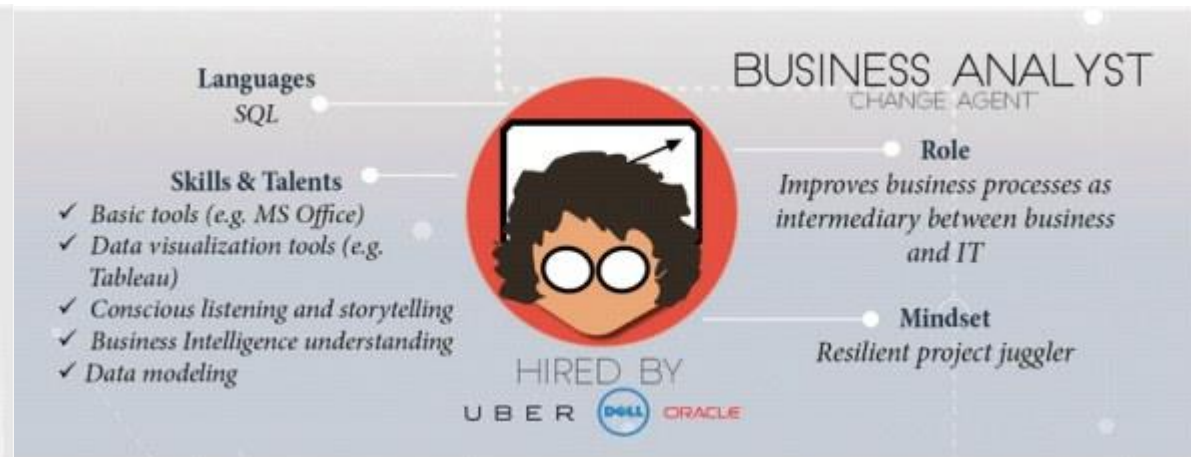
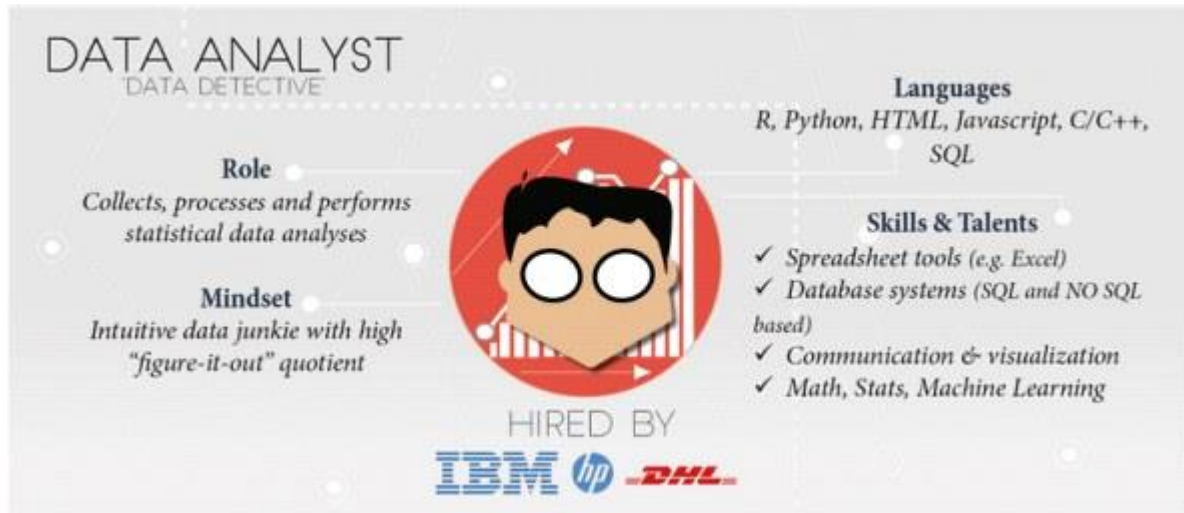
<https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html>

Perfil de Roles en Data Science



<https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html>

Perfil de Roles en Data Science



Perfil de Roles en Data Science

DATA ENGINEER
SOFTWARE ENGINEERS BY TRADE

Role
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset
All-purpose everyman



HIRED BY
Spotify f a

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl


Skills & Talents

- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

DATA ARCHITECT
THE CONTEMPORARY DATA MODELLER

Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:
Inquiring ninja with a love for data architecture design patterns



HIRED BY
VISA Coca-Cola logitech

Languages
SQL, XML, Hive, Pig, Spark


Skills & Talents

- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development

DATA AND ANALYTICS MANAGER
DATA SCIENCE TEAM LEADER

Role
Manages a team of analysts and data scientists

Mindset
Data Wizards' Cheerleader



HIRED BY
coursera slack MOTOROLA SOLUTIONS

Languages
SQL, R, SAS, Python, Matlab, Java

Skills & Talents

- ✓ Database systems (SQL and NO SQL based)
- ✓ Leadership & project management
- ✓ Interpersonal communication
- ✓ Data mining & predictive modeling

DATA SCIENTIST
"AS RARE AS UNICORNS"

Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard



HIRED BY
Google Microsoft Adobe

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

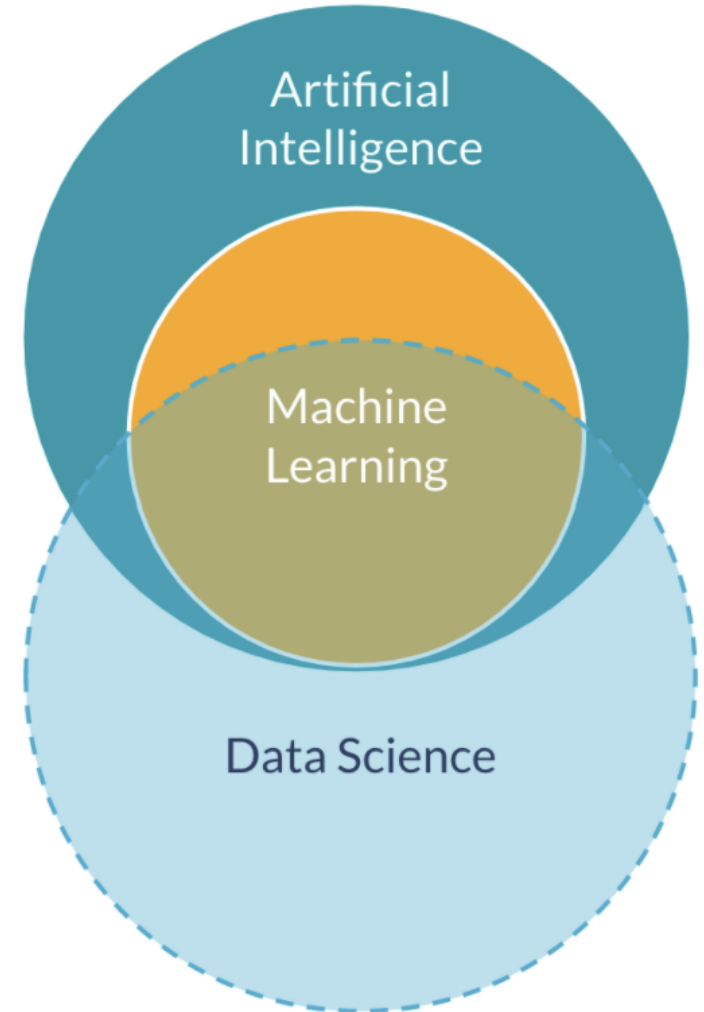
Skills & Talents

- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

Ciencia de Datos \neq Inteligencia Artificial

El éxito que vive la Inteligencia Artificial ha llevado a incrementar la confusión entre términos como: *Data Science*, *Machine Learning* y *Artificial Intelligence*. Conceptos aún más específicos como *Deep Learning* (Aprendizaje Profundo) o *Big Data* se entremezclan con los anteriores.

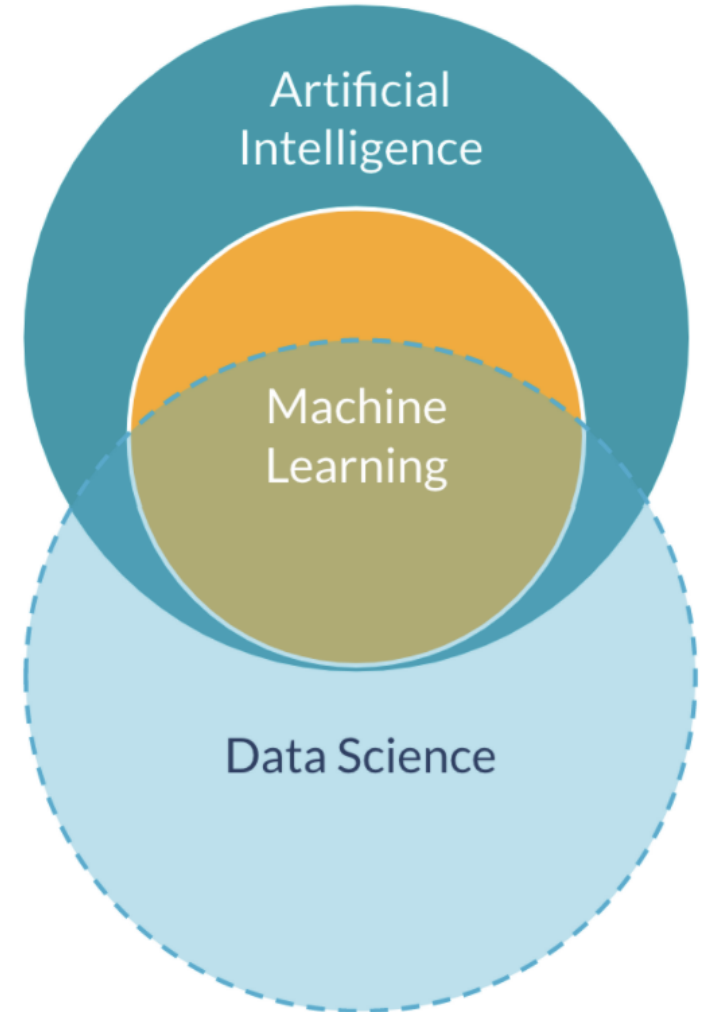
El significado exacto de cada uno de estos términos aún no se asienta del todo.



Ciencia de Datos \neq Inteligencia Artificial

La Ciencia de Datos es un campo multidisciplinario y amplio, cuyo propósito es la extracción de conocimiento a partir de datos.

La ciencia de datos aprovecha el aprendizaje automático para generar modelos predictivos y encontrar patrones en los datos. *ML* es una herramienta en el contexto de *DS*, permite cumplir con los objetivos de investigación usando técnicas eficaces, simples de desarrollar y flexibles.

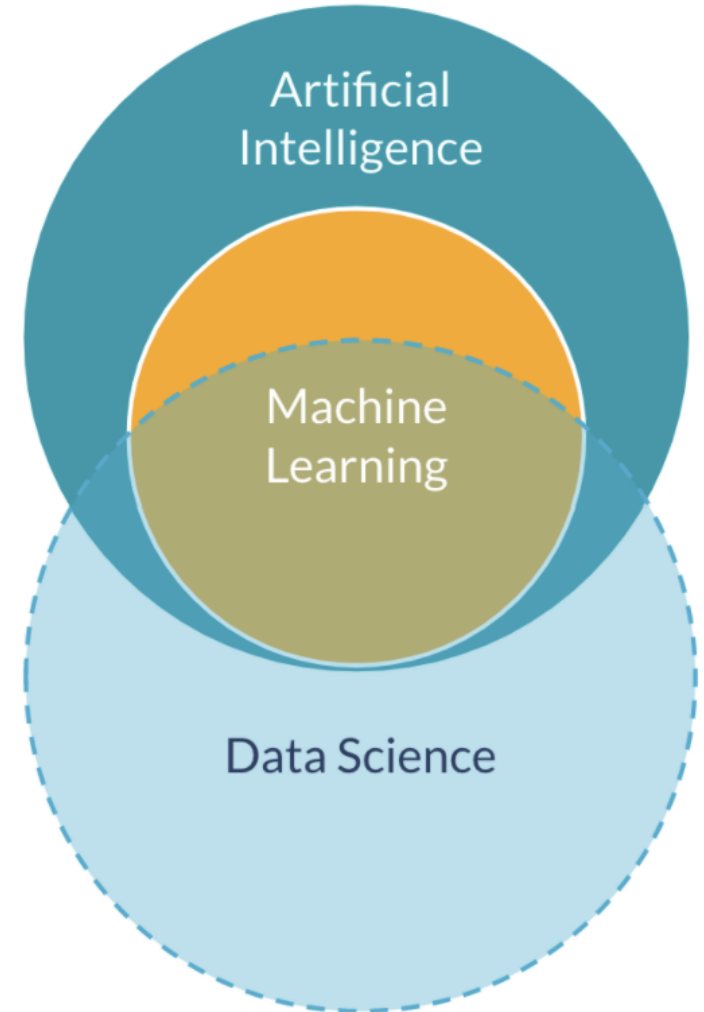


Ciencia de Datos \neq Inteligencia Artificial

La Inteligencia Artificial es la simulación de la inteligencia biológica por medios computacionales. Provee capacidades humanas en máquinas artificiales.

El aprendizaje automático es el principal mecanismo por el cual se ha logrado construir sistemas que califican como inteligencia artificial, pero no es el único. Por lo mismo, *ML* no es sinónimo de *IA*, pero la mayoría de las *IA* se obtuvieron a través de *ML*.

¡PERO ES LA CIENCIA DE DATOS LA QUE LO HA HECHO POSIBLE!





UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

SEDE VIÑA DEL MAR

Ciencia de Datos

Profesor: Gabriel Jara

gabriel.jara@usm.cl

Segundo Semestre 2024