Ciencia de datos

Mauricio Sepúlveda

Ingeniero civil en informática USACH
Doctor en Ingeniería mención Automática
Académico USS

Temario

- Publicaciones 2024
- Ciencia de datos e IA

Publicaciones recientes

INDIAN JOURNAL OF INDUSTRIAL AND APPLIED MATHEMATICS
Vol. 15, No. 1&2, January–December 2024, pp. 1–17

IndianJournals.com

DOI:

Models for Predicting the Monetary Value and Number of Transactions in a Neobank

Mauricio Sepúlveda Cárdenas[1*], Bárbara Valenzuela Klagges[2], Danilo Gómez Correa[3#], Sebastián Manríquez Robles[4$] and Roberto Acevedo[5~]**

[1]Universidad San Sebastian, Bellavista N° 7, Recoleta, Santiago, Chile
[2]Universidad Mayor, Manuel Montt N° 367, Providencia, Santiago, Chile
[3]Universidad del Desarrollo, Avenida Sanhueza N° 1750, Concepción, Chile
[4]Universidad San Sebastián, Bellavista 7
(*Corresponding author) E-mail: *mauricio.sepulveda@uss.cl, **barbara.valenzuela@umayor.cl, #d.gomez@udd.cl, $smanriquezr@gmail.com, ~roberto.acevedo.llanos@gmail.com

ORCID:[1]0000-0002-5522-3194; [2]0000-0002-7584-8183; [3]0000-0002-8735-7832; [5]0000-0001-6847-0285

INDIAN JOURNAL OF INDUSTRIAL AND APPLIED MATHEMATICS
Vol. 15, No. 1&2, January–December 2024, pp. 1–15

IndianJournals.com

DOI:

Automatic Generation of Recursive Algorithms for List Manipulation

Mauricio Sepúlveda Cárdenas[1*], Victor Parada Daza[2], Juan Parra Galvez[1] and Roberto Acevedo[3]

[1]Universidad San Sebastian, Bellavista N° 7, Recoleta, Santiago, Chile
[2]Universidad de Santiago de Chile, Santiago, Chile
[3]Universidad San Sebastián, Bellavista 7
(*Corresponding author) Email id: *mauricio.sepulveda@uss.cl, [2]victor.parada@usach.cl, [1]juan.parra@uss.cl, [3]roberto.acevedo.llanos@gmail.com

ORCID: 0000-0002-5522-3194[1]; 0000-0002-8649-5694[2]; 0000-0001-9244-0019[3]; 0000-0001-6847-0285[4]

Generation of domestic tourism travel time series using Big Data from mobile phone data.

Mauricio Sepúlveda Cárdenas, Jaime Miranda Fierro, Mauricio Hidalgo Barrientos

Abstract

Detecting and measuring domestic tourism is a complex task because it is hidden among many types of population movements. In addition, traditional measurement methods suffer from high resource consumption and significant time windows. This paper describes the rules and process for obtaining time series of domestic tourism flows for overnight trips from passive mobile phone data (MPD) from Chile. The generated data are monthly for the period from January 2015 to March 2019 (pre COVID). To validate the data, the time series for the region with the highest domestic tourism flow in the country (Valparaíso) is analyzed. For this purpose, data from Google Trends and data on domestic tourist arrivals in protected tourist centers are used. With both, important similarities and causal correlations are obtained. To obtain information about the time series, forecasting models are built using Linear Regression with Seasonal Component (MLR) and SARIMA models. The latter is widely used in tourism forecasting. From the SARIMA model, we obtain a forecast that is less than or equal to 6.0 %ADD and information indicating that lags 1 and 12 are important for both models. In conclusion, the time series constructed show important correlations and coincidences with data collected from other sources and the forecasting models show a behavior consistent with what is expected for tourism, which allows us to consider the process promising for obtaining reliable time series of domestic tourism with cost and time savings and increased frequency.

Key words: Time series, mobile phone data, domestic tourism, Forecasting, Overnight trips

Modeling the density of chlorinated brines with nonlinear multivariate regressions.

Mauricio Sepúlveda[1*], Thierry Bertrand De Saint Pierre Sarrut[1], Andrés Soto-Bubert[1], Rashmi Bhardway[2], Roberto Acevedo[1]

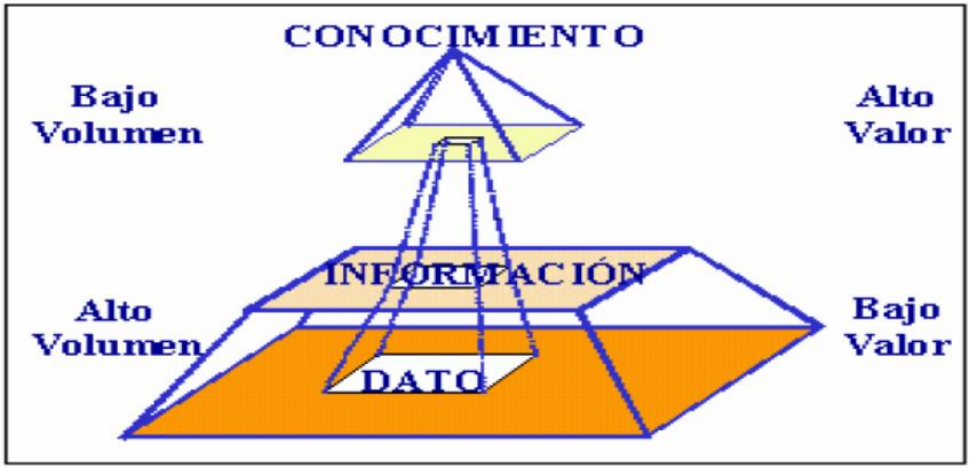
Abstract

There is still no conclusive or definitive model for calculating brine density. While there are soft computing models that consider pressure, temperature, molarity, and brine type, they are black boxes that do not allow us to observe the relationships of attributes or generalize results in new brine. Some techniques enable modeling "interpretable" regressions for multivariate and nonlinear data. These include Symbolic Regression, M5P trees, and the MARS method. To proceed further on and to work out a generic regression for each technique, this work uses data from the density of the brine NaCl, KI, KCl, $MgCl_2$, $SrCl_2$, and $CaCl_2$ that have already been published. The results show that all obtained models have a % AAD lower than 0.72 in test data. Although this result is less accurate than published ones, it is only slightly offset by the models' simplicity and their ability to be used in new untrained brine, such as (Na_2SO_4, $NaHCO_3$, and $AlCl_3$). The residual of the generated regressions is studied, concluding that the models still must incorporate new attributes. The regression models confirm a non-linear relationship between the data attributes. A constant is observed in them, which is similar between the models. A temperature variable shows a relationship with a significant tendency towards linearity and inverse with respect to density, which differs from that indicated in several publications. Similarly, molarity shows a linear and positive behaviour with a small influence in magnitude. It is also observed that the models include the salt molar weight attribute, which interacts strongly with the pressure and temperature attributes and is the one that allows obtaining regressions that provide good accuracy even with new brines. It is concluded that it is possible to generate general regressions for single component brines and to obtain information on the behaviour of the variables from these models. This could serve as a basis for future rigorous formulations of single component brines.

$$\rho = MW * \left(0.28P + \ln(P + 0.7)\right) + 0.33m - 0.7T + \sqrt{T} + \frac{P\sqrt{T} - P}{MW - 76.77} + 1187.37 \qquad (7.0)$$

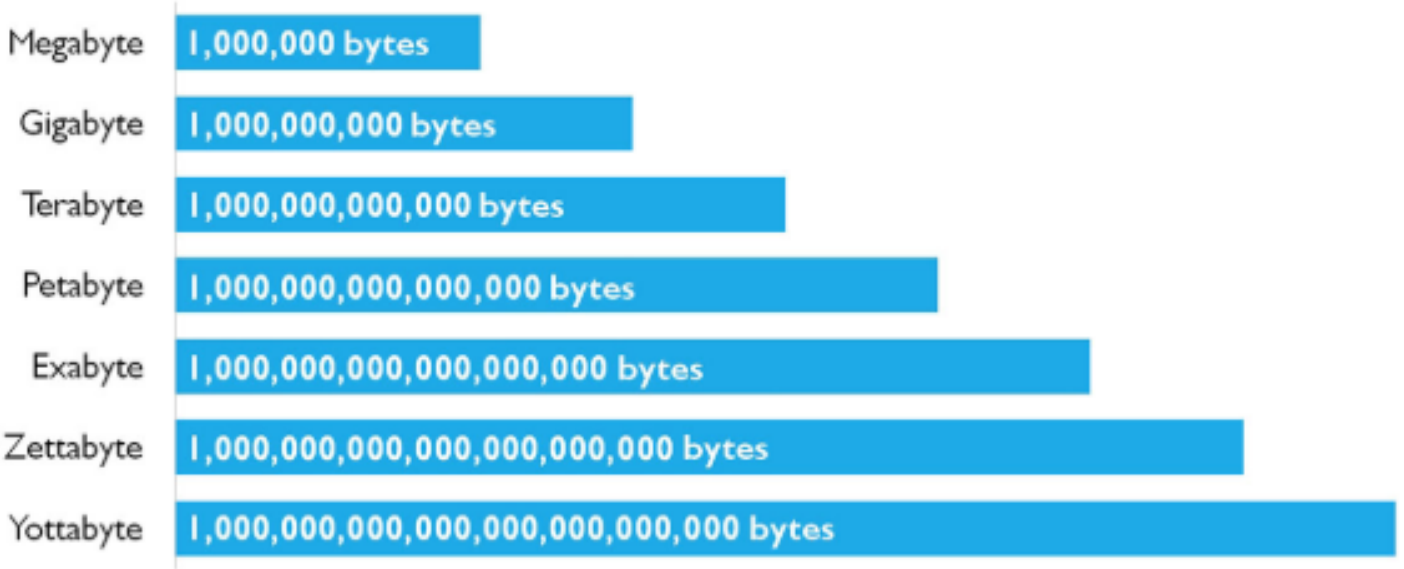
Introducción

- Datos
- Información
- Conocimiento



Introducción

Medida de crecimiento de los datos



↑ Oportunidad de encontrar conocimiento.

Introducción

✓ Definición del Data Science

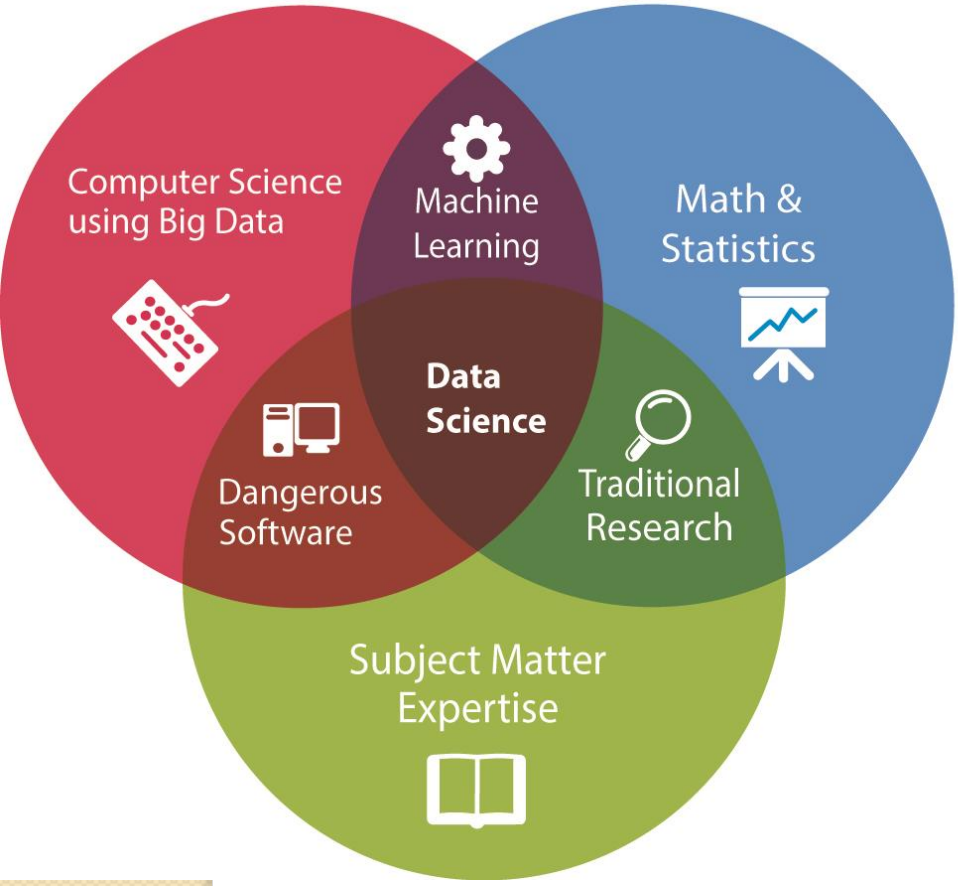
- La ciencia de datos es un campo multidisciplinario que se enfoca en encontrar información y conocimiento a partir de grandes conjuntos de datos estructurados y otros sin procesar.

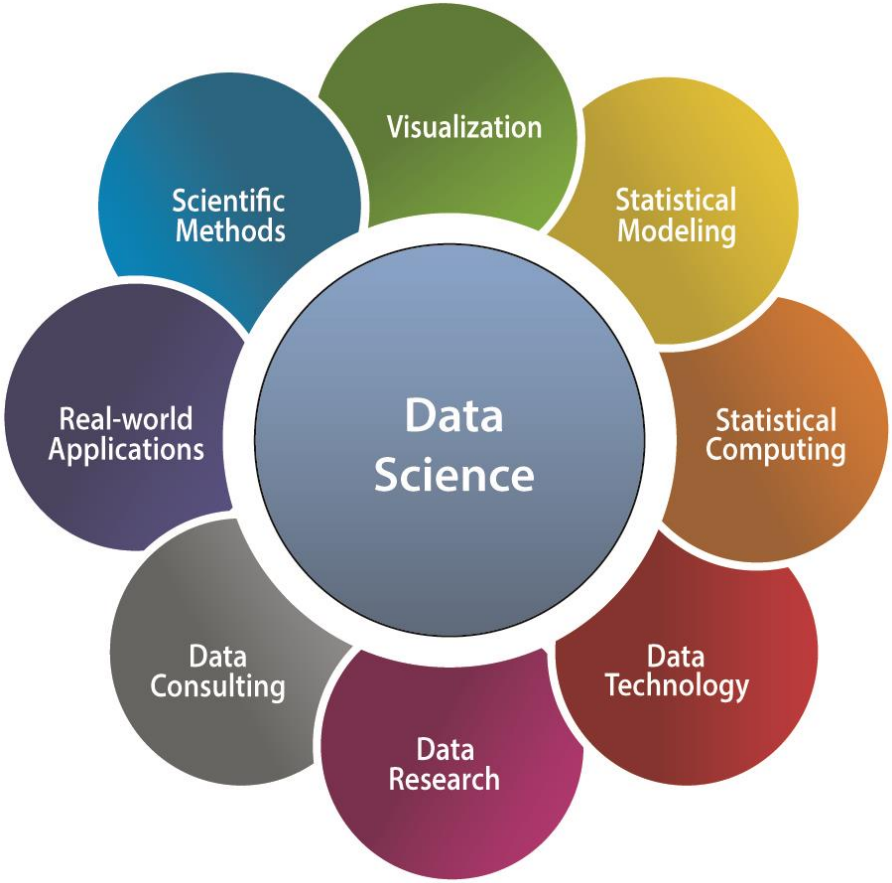
- El campo se centra principalmente en descubrir respuestas a las cosas que no sabemos.

- Los expertos en ciencia de datos usan varias técnicas diferentes para obtener respuestas, incorporando el método científico, las ciencias de la computación, análisis predictivo, estadísticas y aprendizaje automático para analizar conjuntos de datos masivos en un esfuerzo por establecer soluciones a problemas que aún no se han pensado.

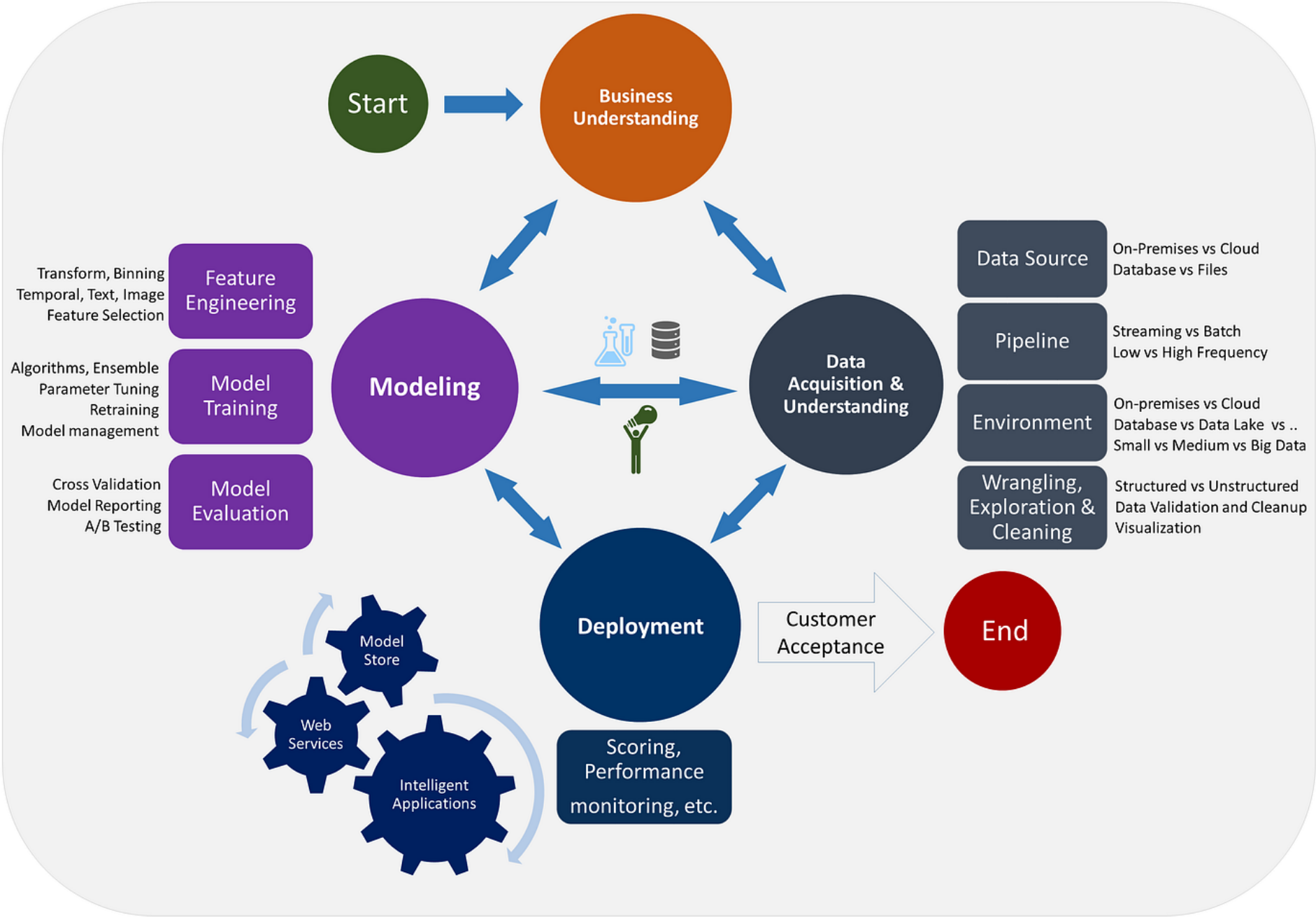
Introducción

- ✓ Contexto del Data Science





Data Science Lifecycle



Introducción

✓ Contexto del Data Science

Business intelligence

Hace referencia a un conjunto de productos y servicios para acceder a los datos, analizarlos y convertirlos en información.

La misión consiste en mejorar el proceso de toma de decisiones en los negocios basándose en los datos corporativos.

Big data

Big Data es un conjunto de tecnologías que permiten la recopilación, almacenamiento, gestión, análisis y visualización, potencialmente en condiciones de tiempo real, de grandes conjuntos de datos con características heterogéneas.


Ciencia de Datos
Ecuador
El tamaño de la base de datos
La capacidad de mi computadora

Introducción

✓ Contexto del Data Science

Cloud Computing

El Cloud Computing es un entorno o plataforma bajo la que almacenamos los datos y dónde ejecutamos las aplicaciones y software especializado para procesar y acceder a estos datos. Ejemplos: Google, Amazon, Azure

IA en Data Science

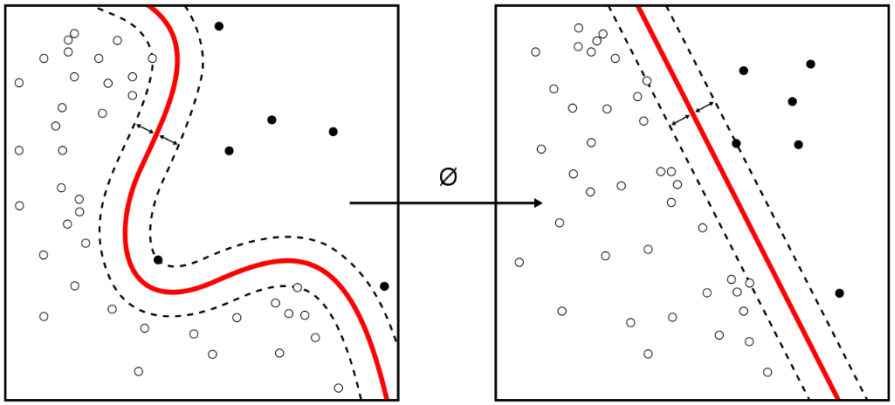
La IA es una suma de algoritmos informáticos complejos que imitan la inteligencia del ser humano. Muchos de estos algoritmos se usan en Ciencia de datos, tales como el Machine Learning, las redes neuronales profundas (Deep Learning), El aprendizaje por refuerzo.
Incluso los modelos de IA requieren data science para estudiar la representatividad de la muestra de datos y para evaluar los modelos.

Introducción

Principales modelos de Data Science

- Clustering/agrupación.
- Clasificación.
- Estimación.
- Predicción.
- Descubrimiento de patrones (reglas de asociación).
- Minería de texto

Introducción



Agrupamiento

Ejemplo:

- Agrupar a los clientes según indicadores de Frecuencia de compra, Monto, cuotas impagas, frecuencia de atrasos, etc. en segmentos de comportamientos homogéneos.
- Resultado: Clientes Buenos, Clientes Medios, Clientes Malos.
- Se reconocen los cluster ya que el 78% de la facturación se concentra en el Cluster considerado de Clientes Buenos.
- Los Clientes Buenos son casados, con hijos, trabajadores autónomos con ingreso superior a $3000 pesos.

Introducción

Clasificación y Estimación

Ejemplo:
Clasificar un "**nuevo**" cliente de acuerdo a su perfil sociodemográfico como un cliente:

- Bueno.
- Medio.
- Malo.

Esto se logra basado en la historia de cómo se han clasificado anteriormente a los clientes.

Otro ejemplo: Estimar el consumo de un determinado rubro de artículos de un grupo de clientes en el próximo trimestre. Se puede usar un modelo de regresión.

Introducción

Predicción

Predecir el abandono de un cliente:

- Para una compañía de telefonía celular.
- Para una AFP.
- Para una tarjeta de crédito.
- Para una universidad

Otro tipo de predicciones se logra gracias a que generalmente, existe una serie de tiempo con datos de como se ha portado la variable en el tiempo.

Introducción

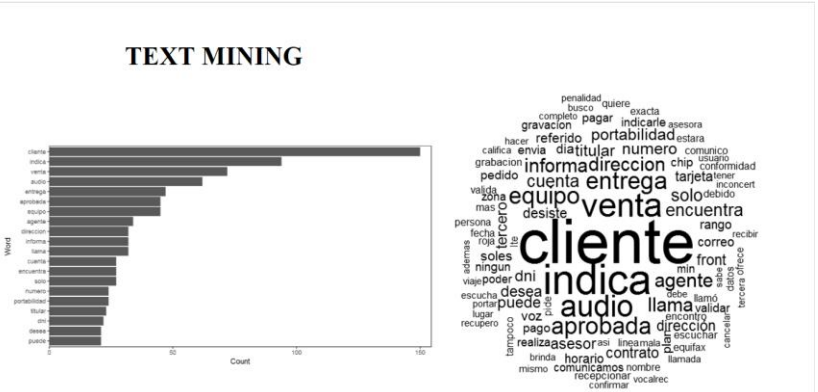
Asociación

Encontrar las reglas que determinan la interrelación entre productos para clientes online de un banco.

Por ejemplo:

"Cuando un cliente se activa en Caja de Ahorros, el siguiente producto donde se activa es Préstamos Personales. Este patrón ocurre el 65 % de los casos."

Introducción



Minería de texto

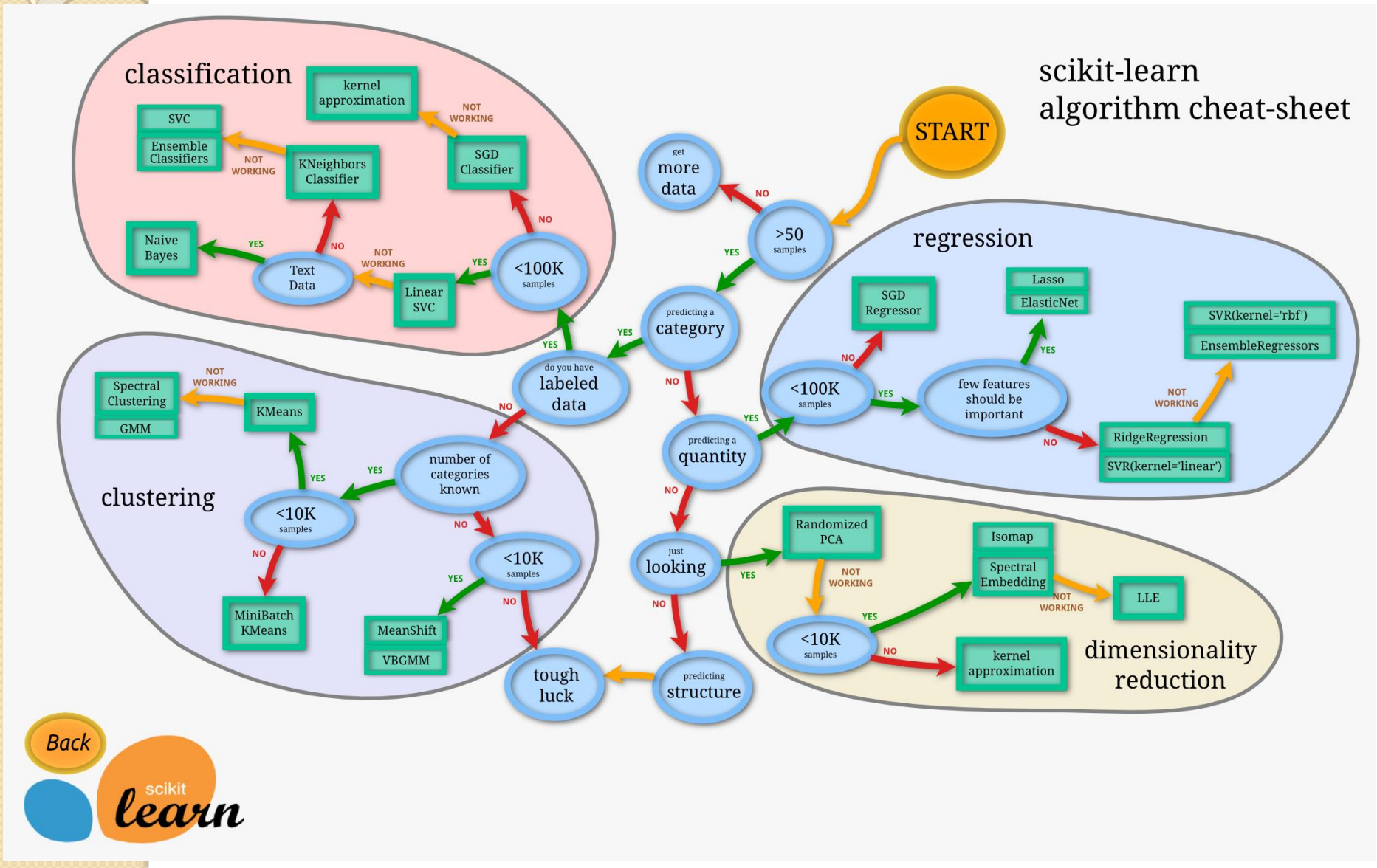
En base al texto de fichas médicas de urgencia, reconocer a pacientes con ideación suicida.

Uso de análisis de sentimientos para texto de comentarios de clientes en redes sociales.

Agrupación de documentos con textos similares o reconocimiento del autor.

Uso de software de redes generativas como chatGPT.

Ejemplo de técnicas en Data Science



Introducción

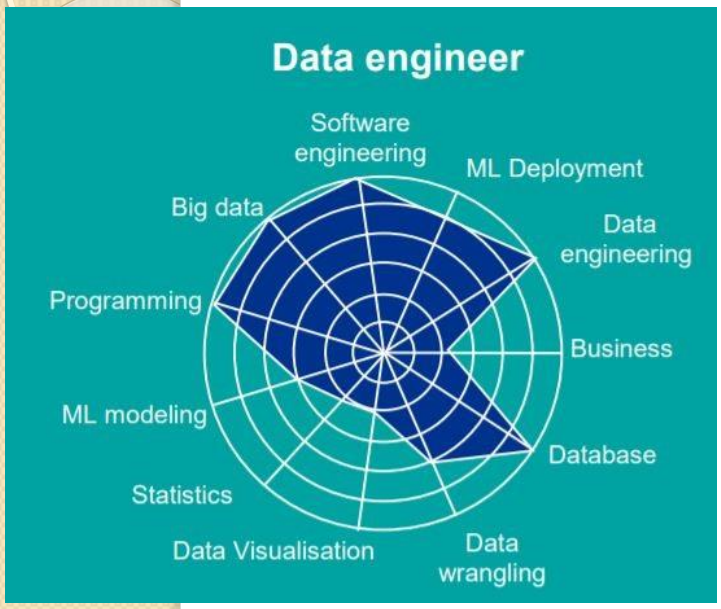
Limitaciones

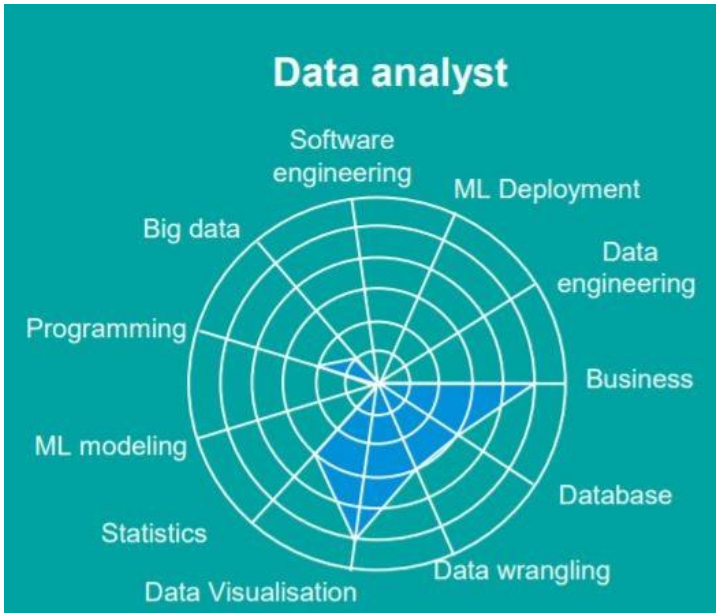
La ciencia de datos tiene el potencial de proporcionar ideas valiosas y respuestas a través del análisis de datos.

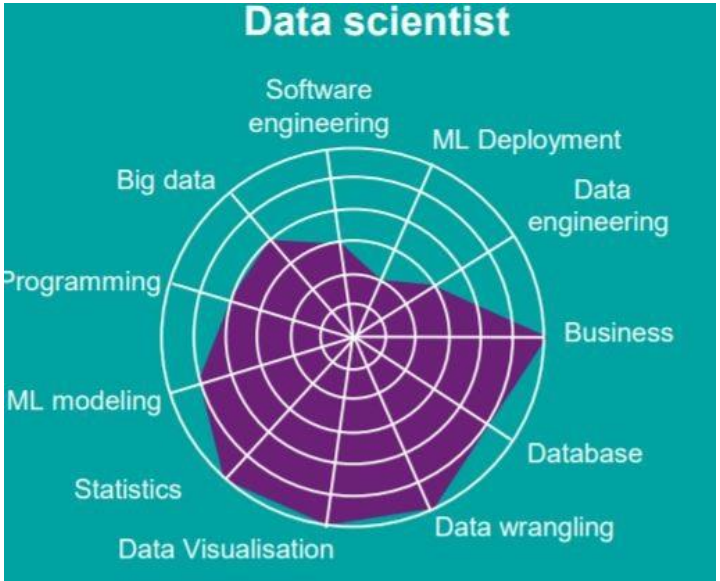
- Permite descubrir **patrones** ocultos, identificar **tendencias**, realizar **predicciones** y tomar **decisiones** informadas.

- Sin embargo, también tiene limitaciones. No puede proporcionar respuestas **definitivas** o predecir el **futuro** con certeza absoluta. Además, la ciencia de datos puede verse afectada por **sesgos** y limitaciones inherentes a los datos utilizados, como la falta de **representatividad** o la presencia de datos **faltantes**.

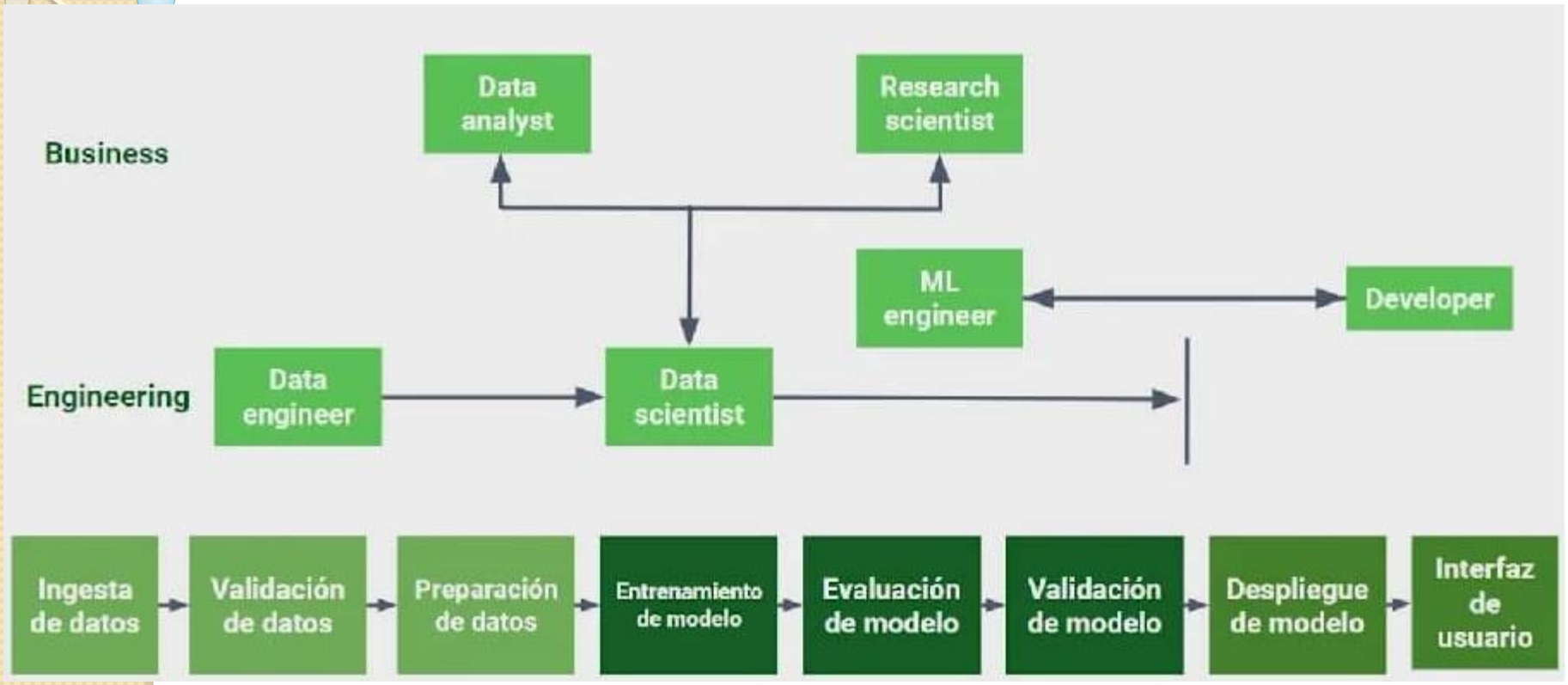
Perfiles o roles en data science







Como funcionan los equipos



Recopilación y Preparación de Datos

Recopilación y Preparación de Datos

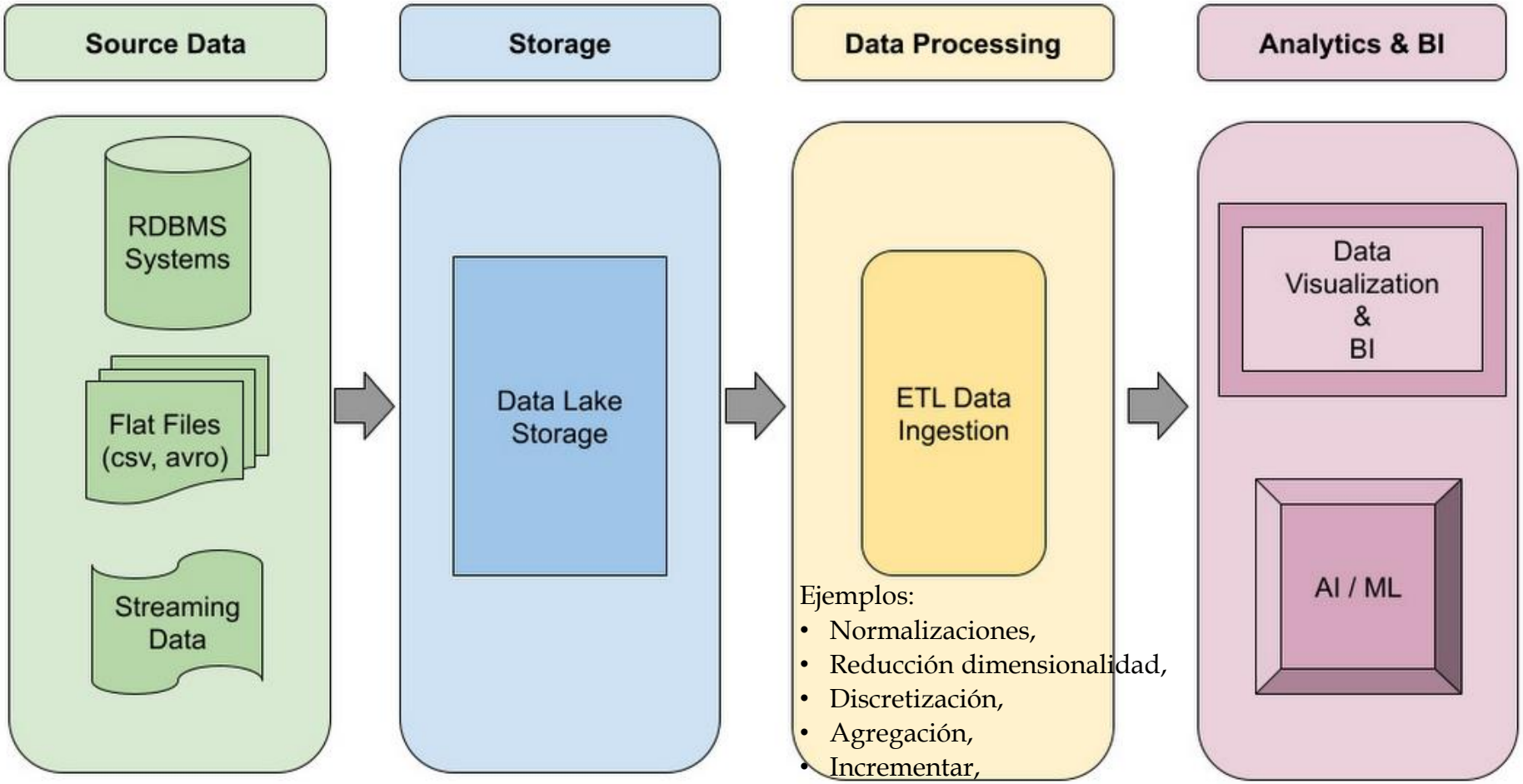
Recopilación y Preparación de Datos

Fuentes y tipos de datos

Recursos de entrada de información	Tipo de información Generada
Correo electrónico, SMS, mensaje instantáneo, YouTube, WhatsApp, Web	Video textual, gráfico y audio
Instrumentos médicos electrónicos, científicos datos experimentales y observacionales	Numérico (es decir, temperatura, presión, etc.) e imágenes de diagnóstico (es decir, tomografía computarizada, ECG, etc.)
Sensores ambientales	Numérico, textual, gráfico, audio-video
Transacciones financieras	Textual y Numérico
Base de datos tradicional y Datawarehouse	Numérico y textual
Satélite	Numérico y gráfico

Recopilación y Preparación de Datos

ETL (Extracción, Transformación, Carga)



Ejemplos:

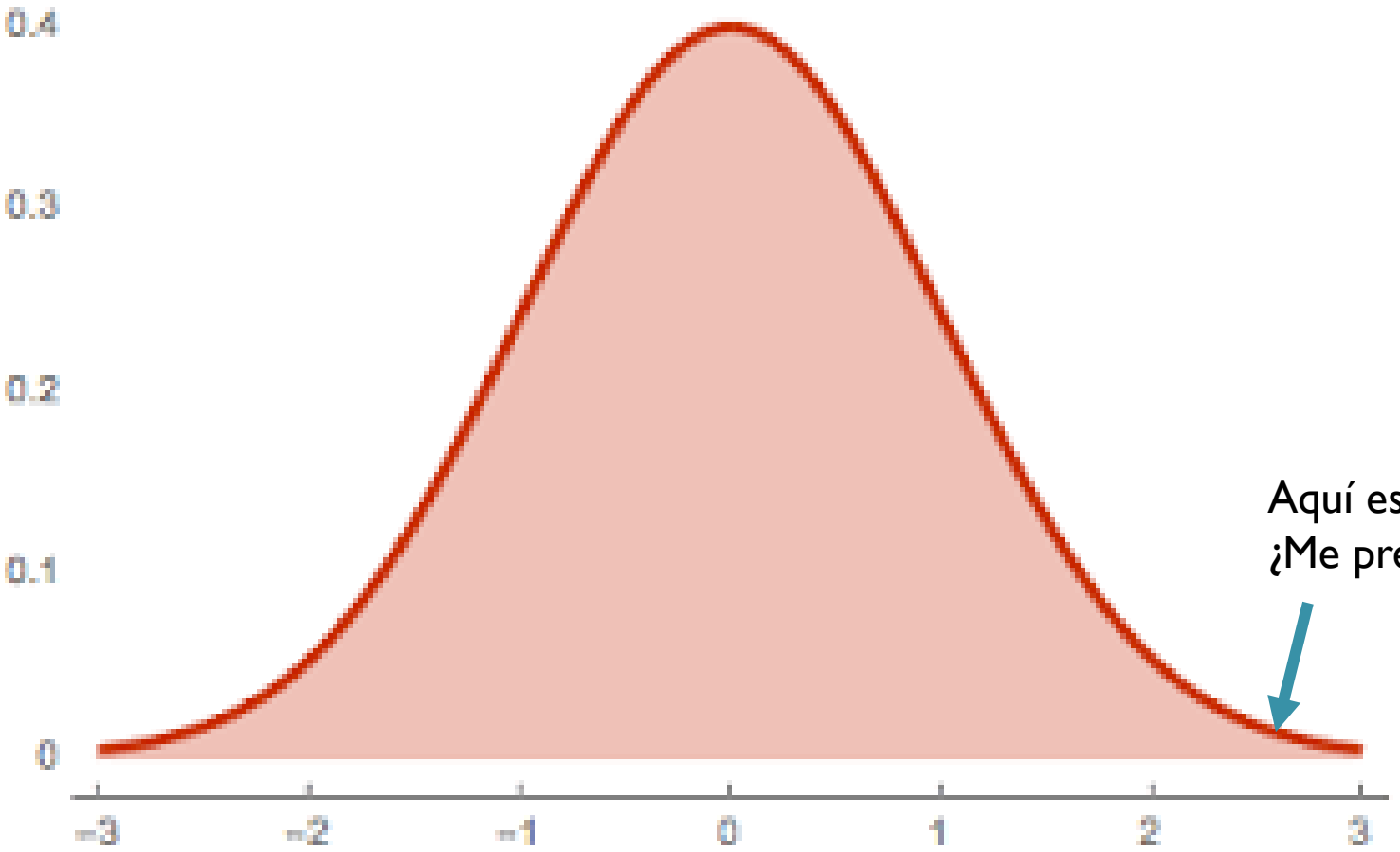
- Normalizaciones,
- Reducción dimensionalidad,
- Discretización,
- Agregación,
- Incrementar,
- Recodificar, etc.

Recopilación y Preparación de Datos

TechBI analytics

TODOS QUIEREN CONVERTIRSE EN DATA SCIENTIST, PERO NADIE QUIERE LIMPIAR DATOS

Análisis de Datos y Estadísticas




SEE
YO SOY TU
PADRE
DE: ESTADÍSTICA Y COMPUTACIÓN
A: MACHINE LEARNING Y CIENCIA DE DATOS

Estadística descriptiva e inferencial

Estadística: ciencia que se ocupa de recoger, clasificar, representar y resumir los datos de muestras, y de hacer inferencias (extraer conclusiones) acerca de las poblaciones de las que éstas proceden.

1. **Estadística descriptiva:** parte de la estadística que se ocupa de recoger, clasificar, representar y resumir los datos de las muestras.
2. **Estadística inferencial:** parte de la estadística que se ocupa de llegar a conclusiones (inferencias) acerca de las poblaciones a partir de los datos de las muestras extraídas de ellas.

Estadística descriptiva e inferencial

ESTADÍSTICA

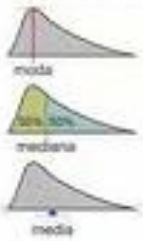
DESCRIPTIVA

Cualitativos

- Proporciones

Cuantitativos

- Tendencia central



- Dispersión

σ^2

σ

IC

INFERENCIAL

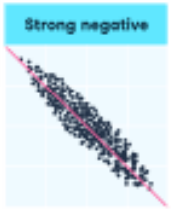
- Descripciones
- Predicciones
- Comparaciones
- Generalizaciones

Conclusiones

- Estimaciones
- Probabilidad
- Hipótesis

Cualitativos → Chi Cuadrada, Probabilidad exacta de Fisher

Correlaciones



Estadística – Hoja resumen

	Continuous, parametric	Continuous, non-parametric	Categorical	Survival
Describe one group	Mean, SD, box plot, frequency distribution, stem and leaf plot	Median, IQR, box plot, frequency distribution, stem and leaf plot	Proportions, frequency table	Kaplan Meier survival curve
Compare one group to a hypothetical value	CI for mean, one sample t-test	CI for median, Wilcoxon signed rank sum test	CI for proportion, Chi square test	
Compare two unpaired groups	CI for difference in means, Two sample t-test	CI for difference in medians, Mann-Whitney test (Wilcoxon rank sum test)	CI for difference in proportions Chi square test Fisher's Exact test Odds ratio, Risk Ratio	Log-rank test or Mantel Haenszel
Compare three of more groups	One-way ANOVA	Kruskal-Wallis test	Cochran Mantel-Haenszel, Chi Square test	Cox proportional hazards (PH) regression
Compare two paired groups	CI for paired differences, Paired t-test	Wilcoxon signed rank sum test	McNemar's test	Conditional PH regression
Examine association between two variables	Pearson correlation Scatterplot	Spearman correlation	Chi-square test	
Predict value from another variable	Linear regression	Non-parametric regression, Linear regression on transformed variable	Logistic regression	Cox PH regression

Análisis de Datos y Estadísticas

¿Qué es la visualización de datos?

La visualización de datos es el proceso de utilizar elementos visuales como gráficos o mapas para representar datos. De esta manera, se trasladan datos complejos, de alto volumen o numéricos a una representación visual más fácil de procesar.

Análisis de Datos y Estadísticas

Diferentes tipos de técnicas de visualización

Visualización temporal de datos
Gráfico de líneas, una tabla de líneas o una línea de tiempo.

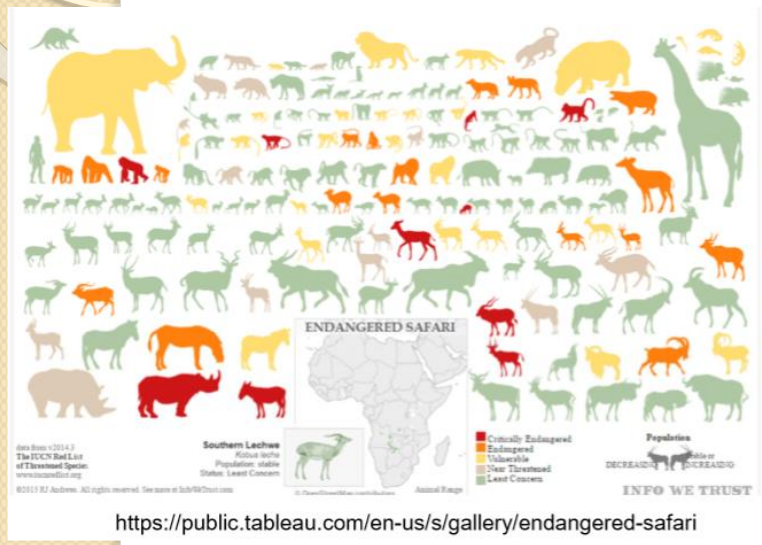
Visualización jerárquica de datos
Árboles de datos para mostrar clústeres de información.

Visualización de datos de la red: Representar la compleja relación entre diferentes tipos de datos que están o no correlacionados.
Gráficos de dispersión
Gráficos de burbujas
Nubes de palabras
Visualización de datos multidimensionales

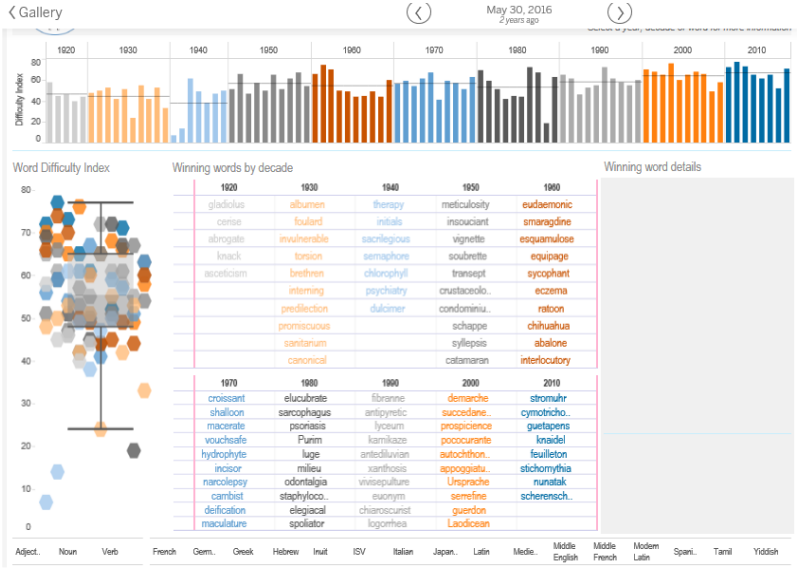
Visualización de datos geoespaciales
Mapas de calor, los mapas de densidad o los cartogramas.

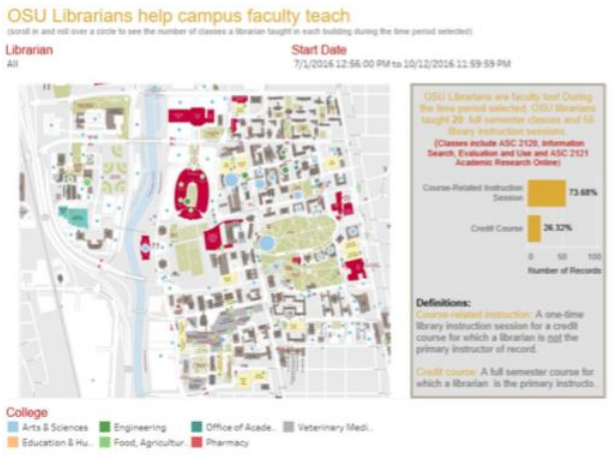
Análisis de Datos y Estadísticas

Visualización de resultados



https://public.tableau.com/en-us/s/gallery/endangered-safari





http://go.osu.edu/REACH-Dashboard

Machine Learning

Machine Learning

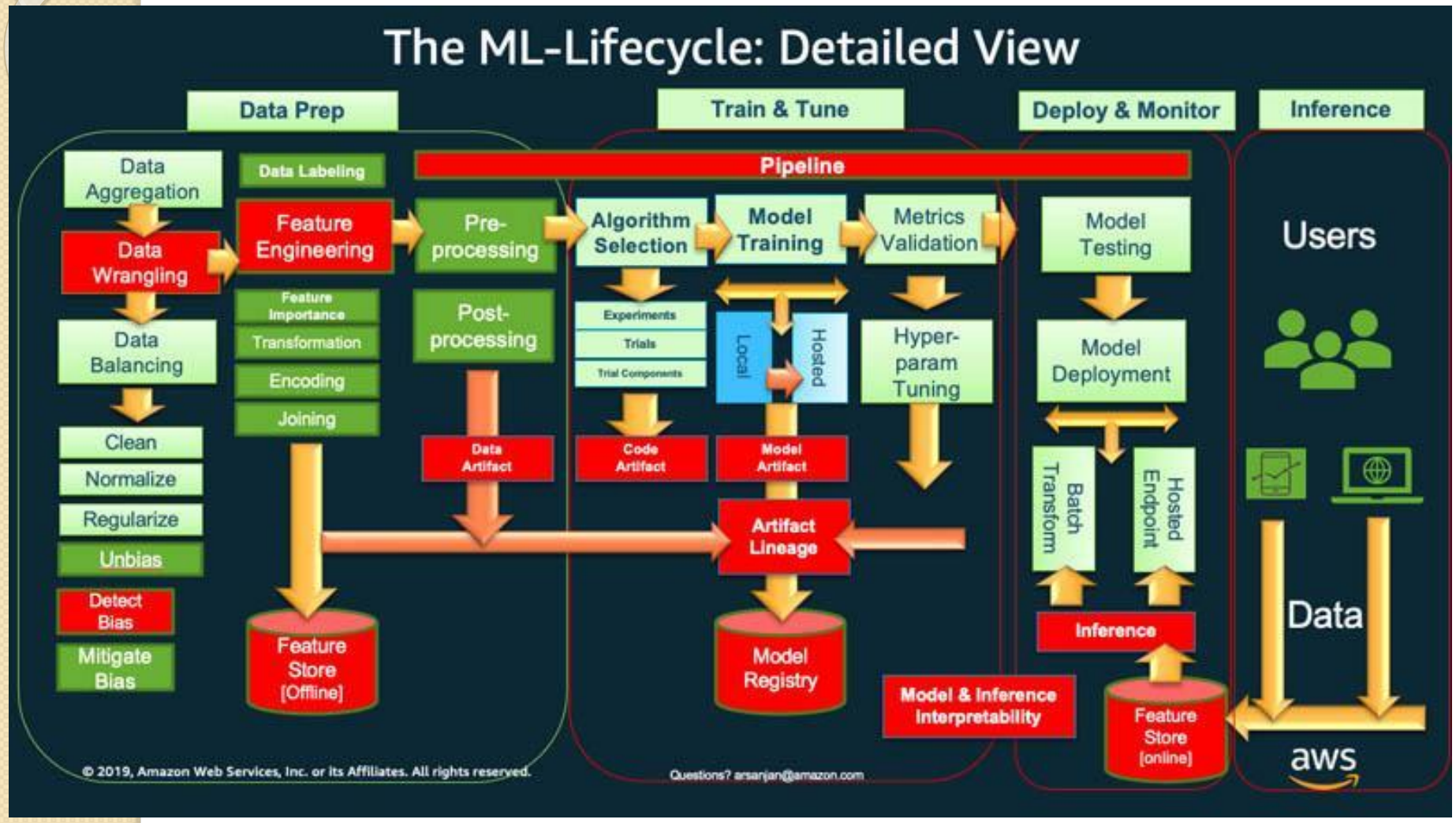
¿Qué es el machine Learning?

Traditional Programming

Machine Learning

Requiere entrenamiento!

Machine Learning

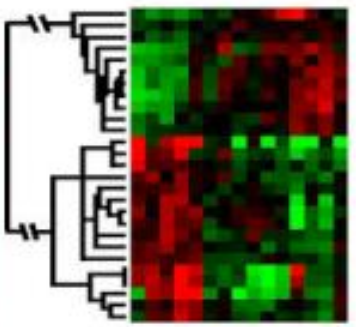


Machine Learning

¿Cuándo usar el machine Learning?

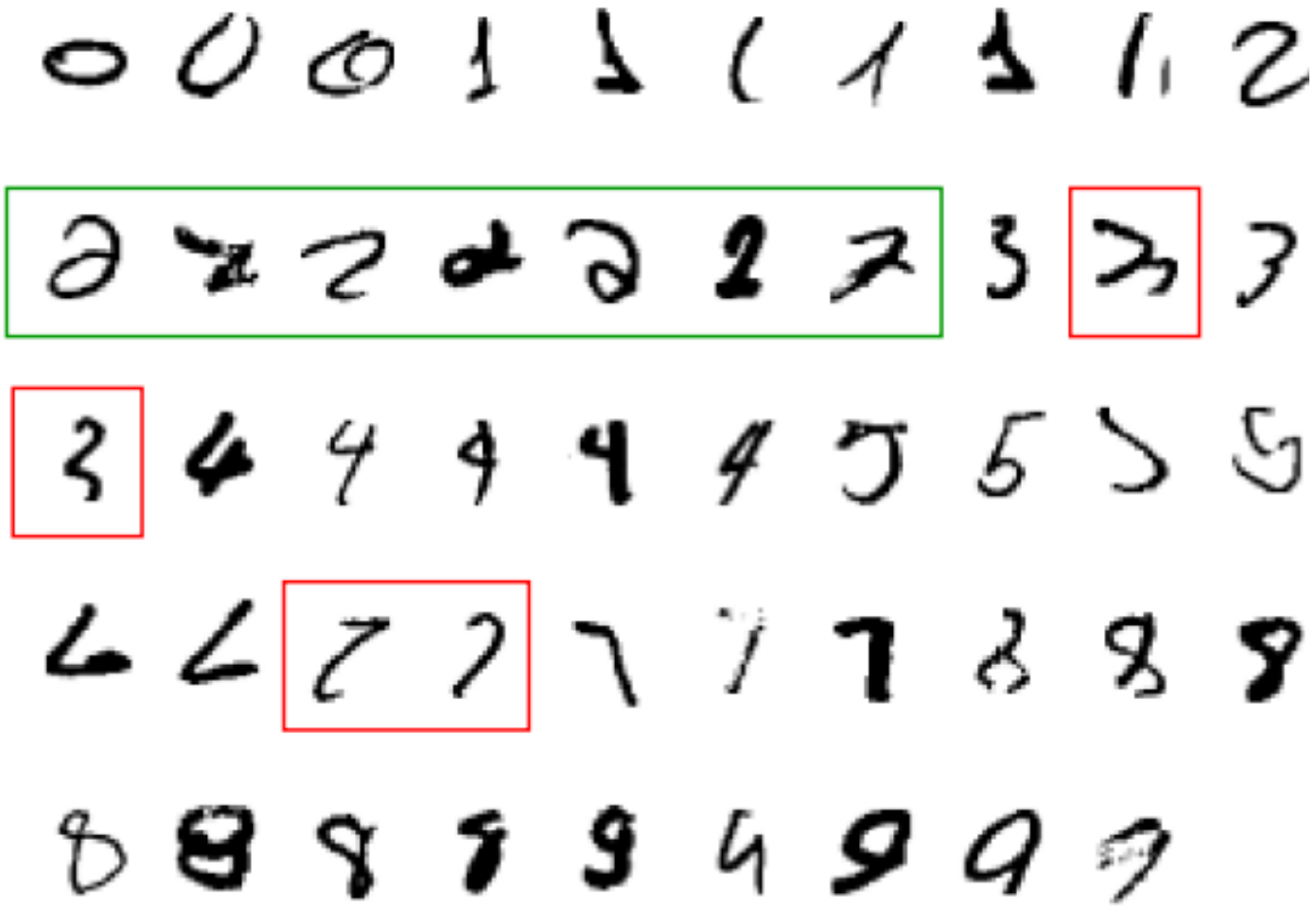
- La experiencia humana no existe (navegar en Marte)
- Los humanos no pueden explicar su experiencia (reconocimiento de voz)
- Los modelos deben ser personalizados (medicina personalizada)
- Los modelos se basan en enormes cantidades de datos y estos están disponibles (genómica)



Machine Learning

Ejemplo, determinar si un número es 2.



Machine Learning
Tipos de aprendizaje

- Aprendizaje supervisado (inductivo)

Dado: datos de entrenamiento + resultados deseados

- Aprendizaje sin supervisión

Dado: datos de entrenamiento (sin resultados deseados)

- Aprendizaje semisupervisado

Dado: datos de entrenamiento + algunos resultados deseados

- Aprendizaje reforzado

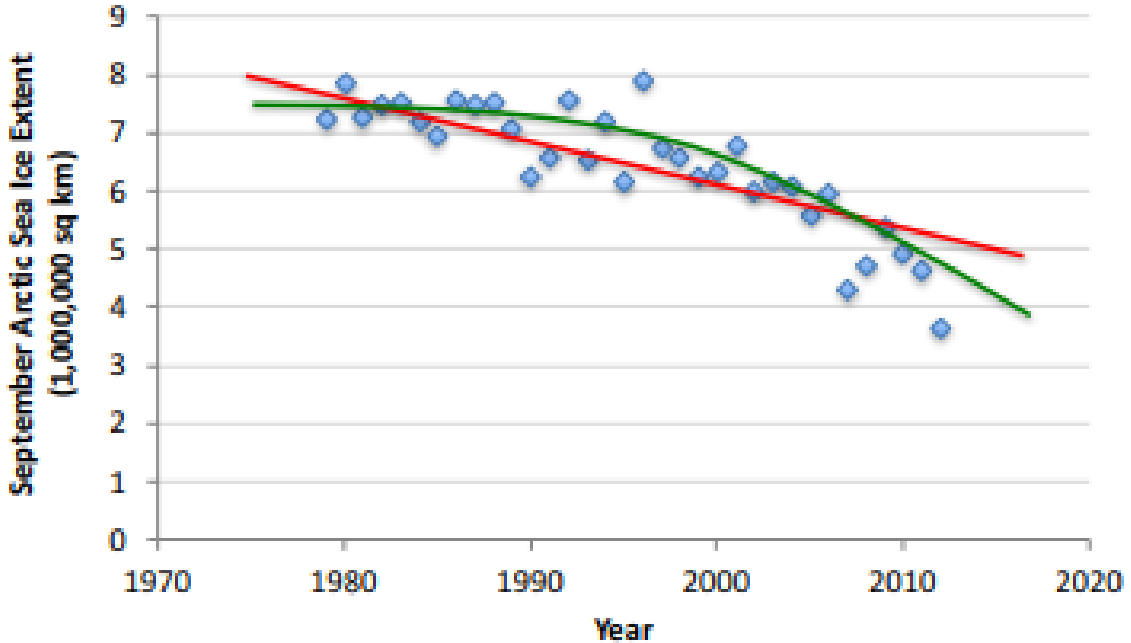
Recompensas por secuencia de acciones.

Machine Learning

Ejemplo, aprendizaje supervisado

Regression

- Given (x_1, y_1), (x_2, y_2), …, (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression

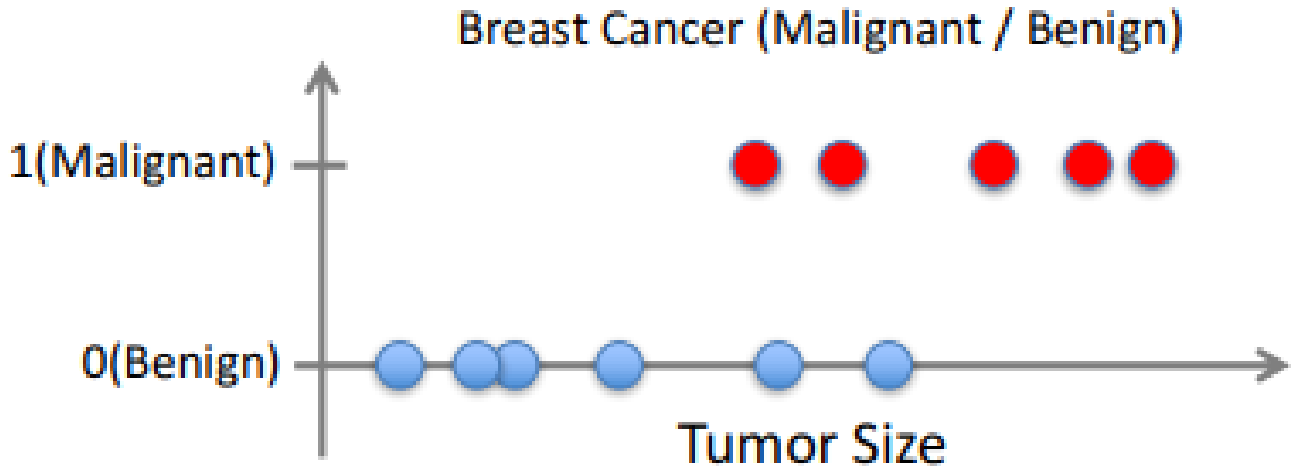


Machine Learning

Ejemplo, aprendizaje supervisado

Classification

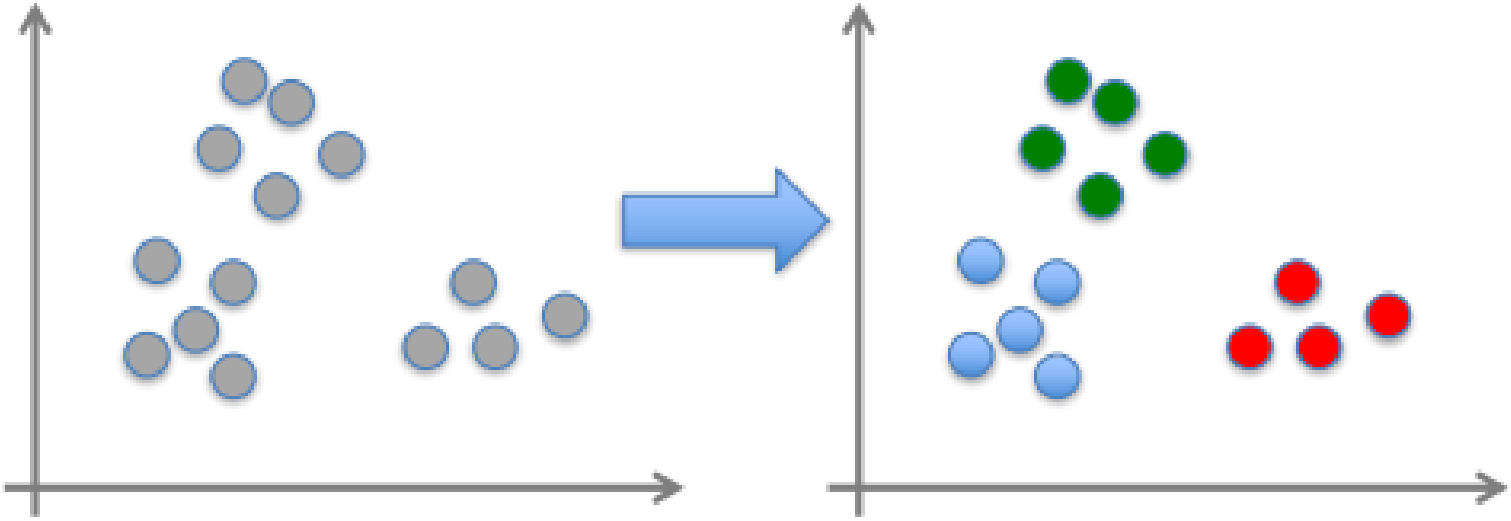
- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification



Machine Learning

Ejemplo, aprendizaje no supervisado

- Given $x_1, x_2, \ldots, x_n$ (without labels)
- Output hidden structure behind the x's
  - E.g., clustering

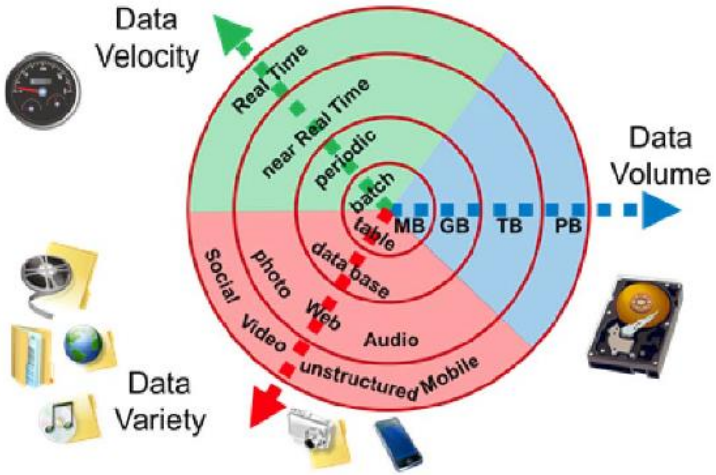


Big data

BIG
DATA
HIGH
TRENDS
INSIGHT
LOGS
TYPES
THINGS
ANY
BLANKET
COLLECTION
DEPENDENCIES
STREAM
ENABLE
TRIGGER
AREAS
WIRELESS
OUT
DENSITY
ENHANCED
TOOLS
TERM
TRANSFER
ANALYSIS
INCLUDE
CAPABILITIES
DISCOVERY
HUNDREDS
MAKING
RESEARCH
SIMULATIONS
NETWORKS
TIME
ENTERPRISE
BUSINESS
SOFTWARE
DATA
SHARING
CAPTURE
PROCESSING
TAKE
SCIENTISTS
STORAGE
DECISION
SPEED
PACKAGES
INTERNET
REPORT
PROCESS
GENOMICS
BIG
CURATE
OPTIMIZATION
TERABYTES
CAPTURE
INFORMATION
SENSOR
MANAGEMENT
PROCESSING
AND
VELOCITY
VARIETY
CHALLENGES
ON-HAND
PETABYTES
AMOUNT
IN
STATISTICS
DOMAIN
DIFFICULT
3V
REQUIRE
RELATIONAL
METEOROLOGY
SIMULATION
THAT
PROCESS
DATABASE
FOR
RANGE
VELOCITY
GROWS
EXPLORE
COMPLEX
ASSETS
VOLUME
THAT
CAMERAS
APPLICATIONS
GARTNER
AND
BYTES
IS A
MANAGE
PROCESS
SOURCES
DOUG
LANEY
NEW
BUSINESS INTELLIGENCE
GOVERNMENT
SEARCH
FLOW
SYSTEMS
SIZE

Big data

LAS "V" DEL BIG DATA



VOLUMEN

- Se refiere al gran volumen de información que se maneja.
- Los datos se acumulan con un crecimiento exponencial, requiriendo ampliar continuamente el almacenamiento de datos.
- Cuando se habla de bases de datos masivas se refiere a magnitudes del orden de petabytes o exabytes.

VELOCIDAD

- Es la enorme velocidad en la generación, recogida y proceso de la información.
- Hay que ser capaz de almacenar y procesar en tiempo real millones de datos generados por segundo por fuentes de información tales como sensores, cámaras de videos, redes sociales, blogs, páginas webs,…

Big data

LAS "V" DEL BIG DATA

VARIEDAD

- Necesidad de agregar información procedente de una amplia variedad de fuentes de información independientes: redes sociales, sensores, máquinas o personas individuales
- En general son datos desestructurados, así como gráficos, texto, sonido o imágenes.
- Estos datos no pueden gestionarse fácilmente con bases de datos relacionales y las herramientas de inteligencia de negocio Tradicionales
- También hace relación a datos con gran número de variables.

Big data

- LAS "V" DEL BIG DATA

VALOR

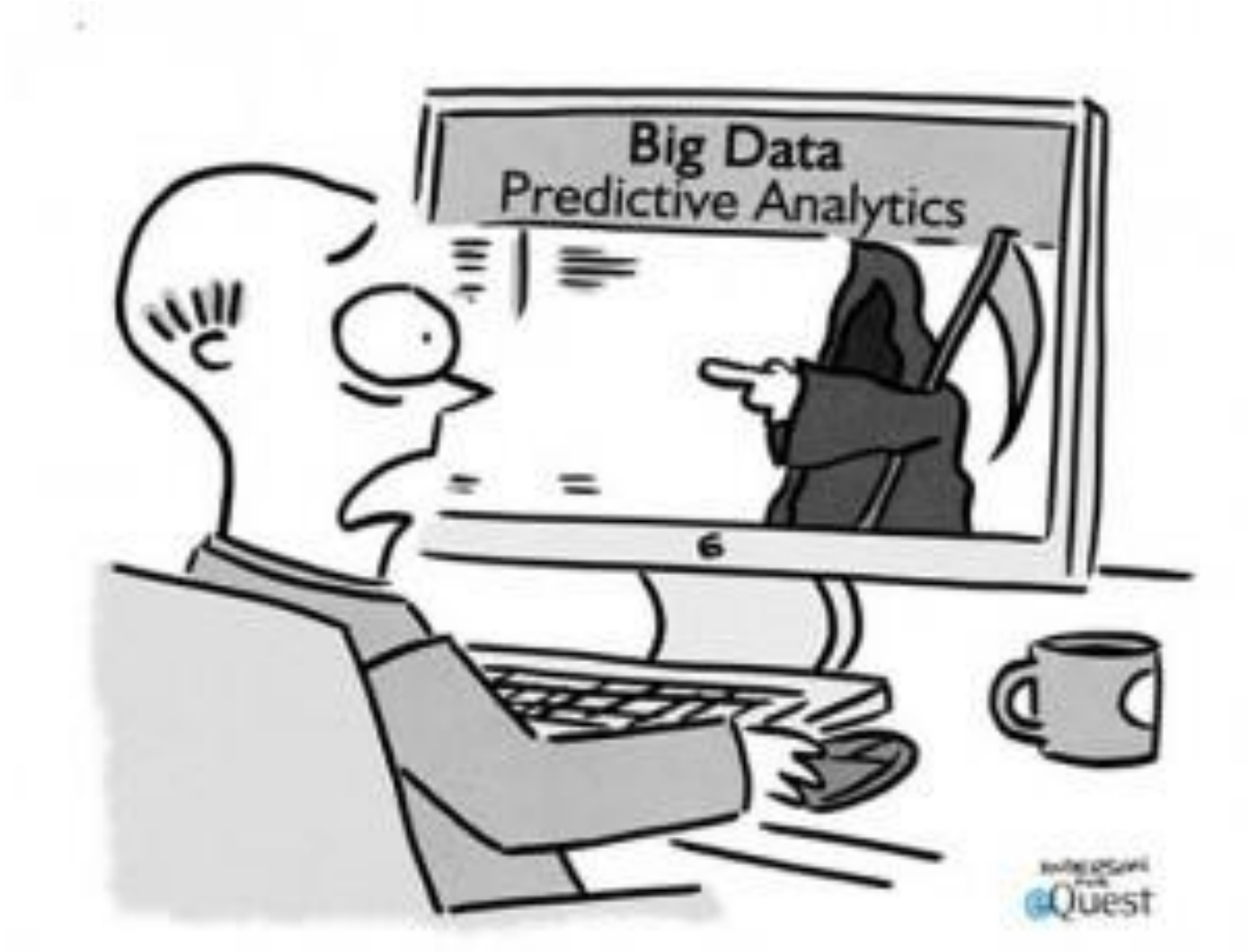
• Es la creación de una ventaja competitiva al identificar y procesar los datos claves, permitiendo así:

  • Monetizar los datos.
  • Obtener nuevos clientes.
  • Generar fidelidad.
  • Reducir costes.
  • Mejorar la imagen de marca.

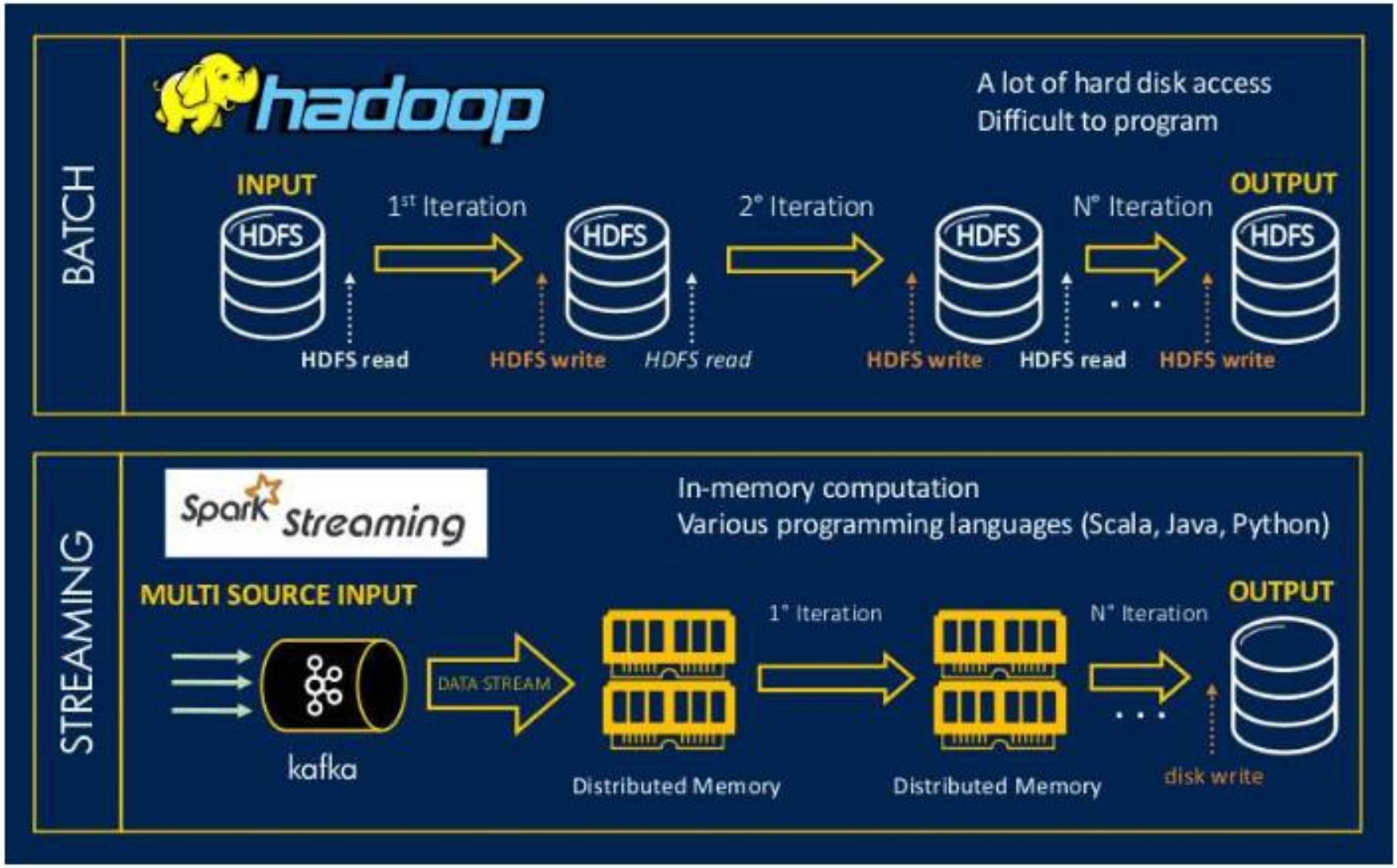
VERACIDAD

• Se debe analizar inteligentemente un gran volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar nuestra toma de decisiones.

Big data

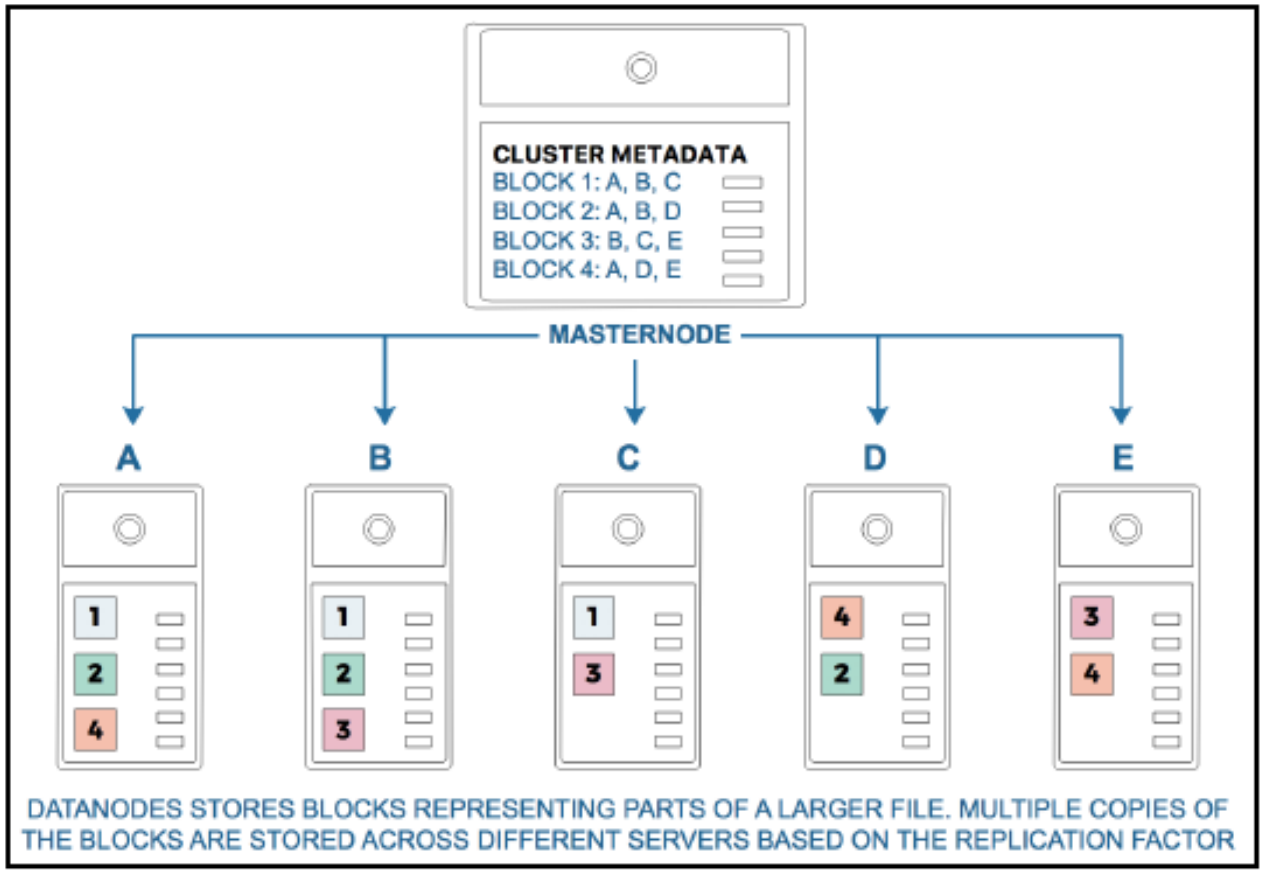


Big data
Arquitectura



¿Como lo hace Hadoop?

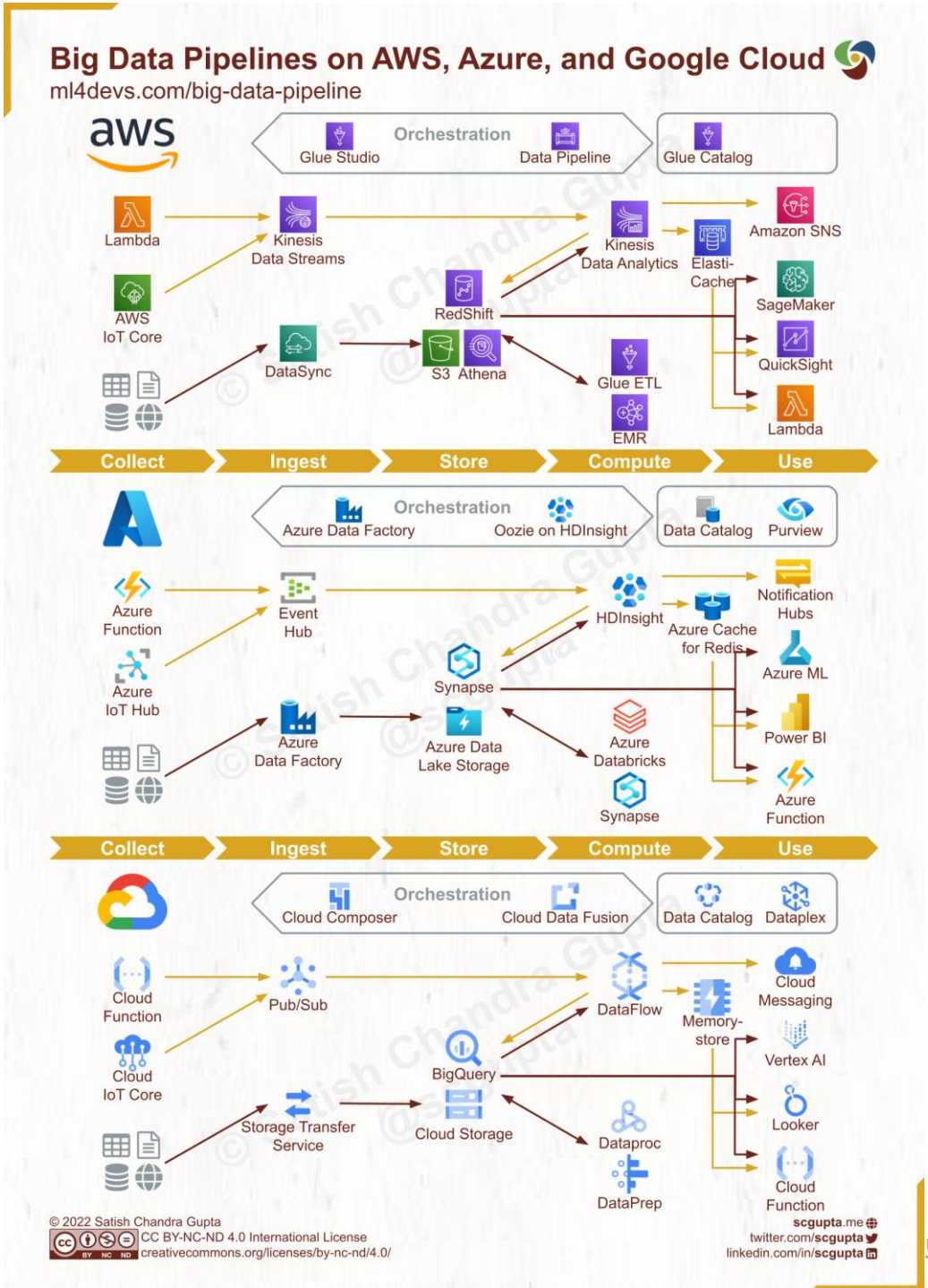
- Data storage process in HDFS



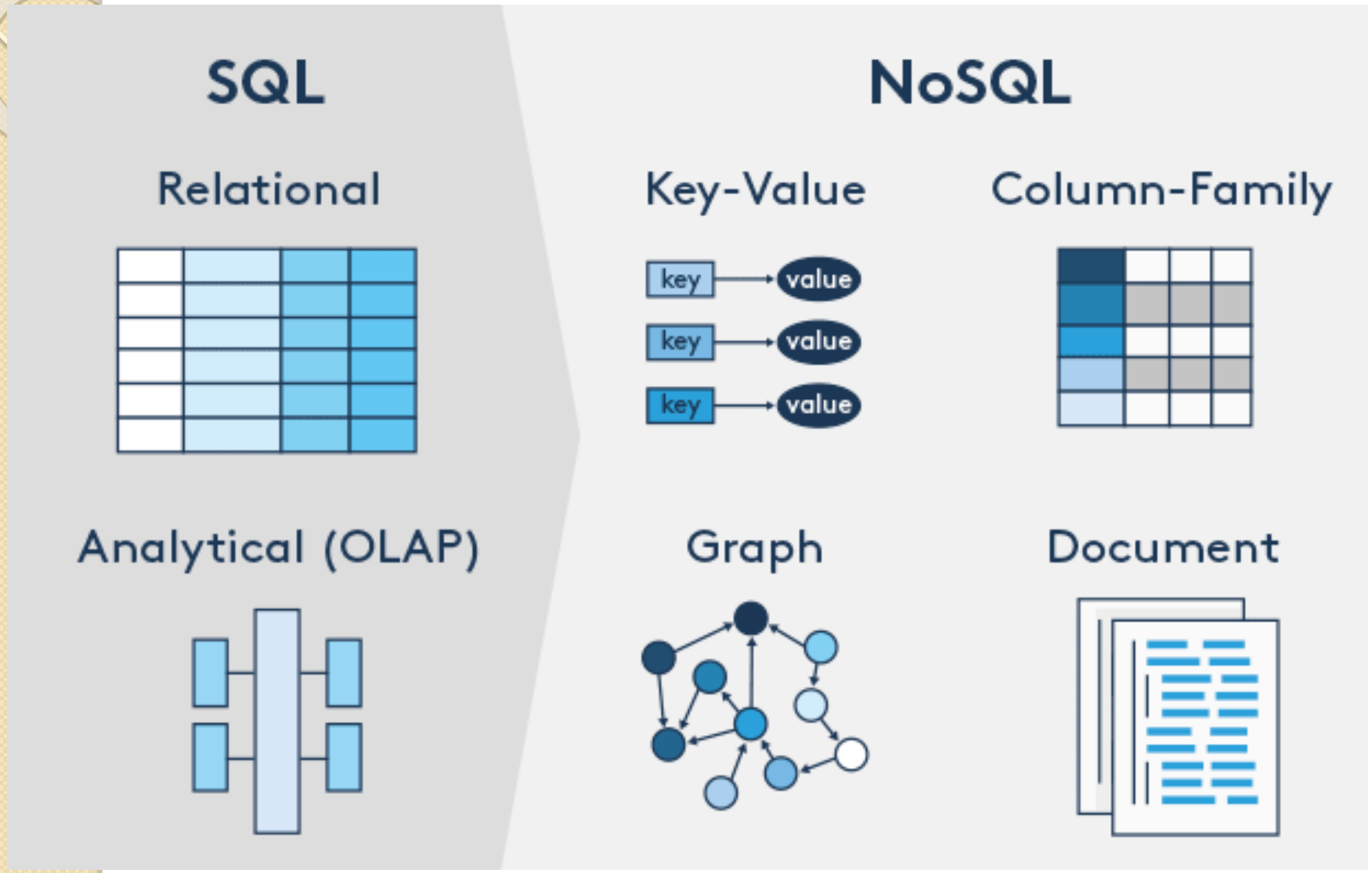
Datos son divididos en bloques de 128 MB
Se copia a un Datanode y este lo replica a otro y este a otro.

Big data

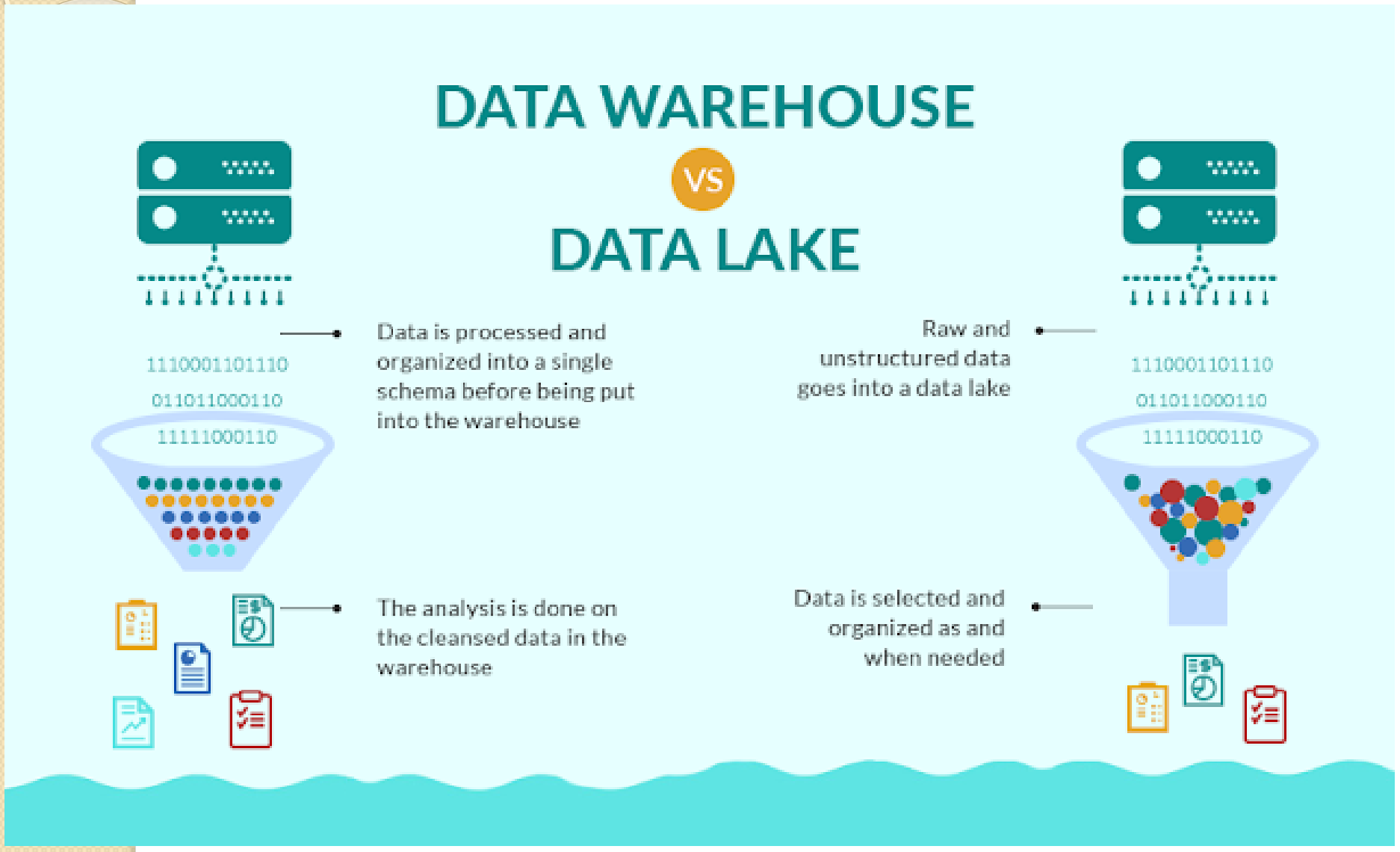
- Arquitectura
- En la nube



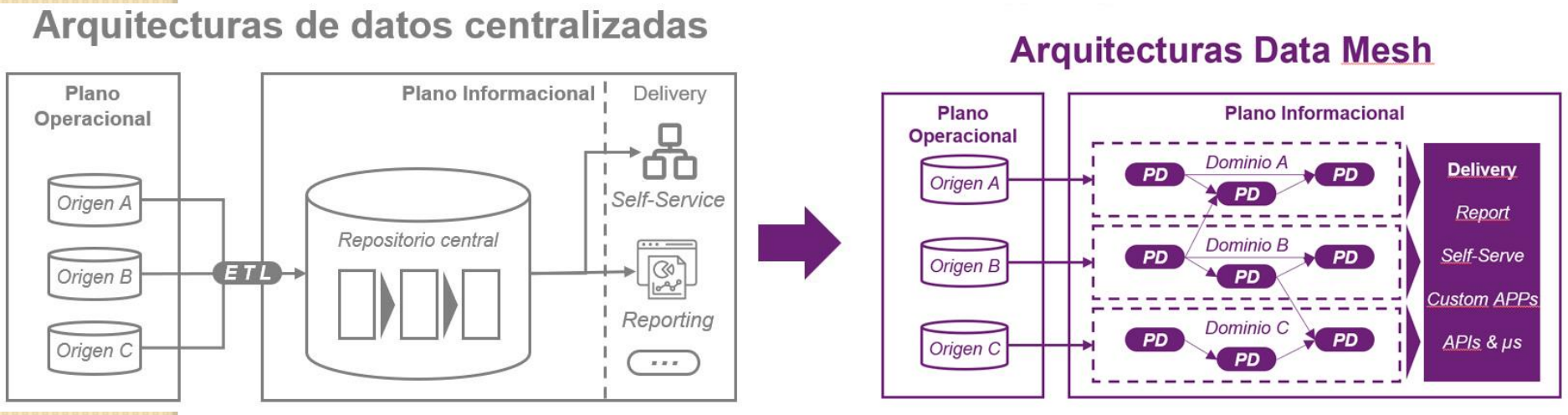
Big data


SQL
NoSQL
Relational
Key-Value
Column-Family
key
value
key
value
key
value
Analytical (OLAP)
Graph
Document

Big data


DATA WAREHOUSE
VS
DATA LAKE
Data is processed and organized into a single schema before being put into the warehouse
1110001101110
011011000110
11111000110
The analysis is done on the cleansed data in the warehouse
Raw and unstructured data goes into a data lake
1110001101110
011011000110
11111000110
Data is selected and organized as and when needed

Big data


Arquitecturas de datos centralizadas
Plano Operacional
Origen A
Origen B
Origen C
ETL
Plano Informacional
Repositorio central
Delivery
Self-Service
Reporting
...
Arquitecturas Data Mesh
Plano Operacional
Origen A
Origen B
Origen C
Plano Informacional
Dominio A
Dominio B
Dominio C
PD
Delivery
Report
Self-Serve
Custom APPs
APIs & μs

Ética y privacidad

Ética y privacidad

Retos de la Big Data: Privacidad vs. Accesibilidad.

- De quien es la información: del negocio o del cliente/ciudadano?
- ¿Pueden los gobiernos usarla en forma discrecional para los temas de estado (seguridad, lucha contra el crimen, etc.)?

La privacidad de la información es la relación entre la recopilación y la difusión de:

- Datos personales y empresariales
- Tecnología
- La expectativa pública de privacidad.
- Problemas legales y políticos que los rodean.

Ética e Información Privada: ¿Que es Ética?

La ética, o filosofía moral, es la rama de la filosofía que estudia la conducta humana, lo correcto y lo incorrecto, lo bueno y lo malo, la moral, el buen vivir, la virtud, la felicidad y el deber.

Las posibles Soluciones

Leyes y regulaciones en Chile.

Ley 19.628 sobre protección de la vida privada o protección de datos de carácter personal, última versión - 09-MAY-2023

- Título i, de la utilización de datos personales.
- Título ii, de los derechos de los titulares de datos.
- Título iii, de la utilización de datos personales relativos a obligaciones de carácter económico, financiero, bancario o comercial.
- Título iv, del tratamiento de datos por los organismos públicos.

Leyes y regulaciones en Chile.

- Ley 20285 sobre acceso a la información pública.

Regula el principio de transparencia de la función pública, el derecho de acceso a la información de los órganos de la Administración del Estado, los procedimientos para el ejercicio del derecho y para su amparo, y las excepciones a la publicidad de la información.

Ley 21459 establece normas sobre delitos informáticos.

Ley Marco de Ciberseguridad e Infraestructura Crítica.

Crea la Agencia Nacional de Ciberseguridad (ANCI).

También el CSIRT Nacional y el CSIRT de Defensa.

Aplicaciones del Data Science

Ejemplos

AGÈNCIA VALENCIANA DE SALUT

Hospital Universitario de San Juan de Alicante

Objetivo: Reducción de costes al permitir un óptimo consumo de recursos.

Problema planteado: Las pruebas preparatorias que se llevan a cabo antes de cualquier intervención suelen ser excesivas, invasivas, caras y generan listas de espera para la intervención.

Solución: Analizando los datos del histórico de operaciones y aplicando técnicas de Minería de Datos se descubren aquellos casos en que dichas pruebas son prescindibles. El sistema de calidad proporciona información detallada del resultado de la operación, de forma que aquellos casos en los que no se han realizado las pruebas y sí hubiesen sido necesarias permiten al sistema seguir aprendiendo y mejorar la identificación de los patrones adecuados.

Ejemplos

Portal B2B Neumáticos Soledad

Objetivo: Aumentar las ventas a través del portal.

Problema planteado: Cómo modificar el portal de compra online que usan los talleres asociados para aumentar las ventas por este canal.

Solución: Extraer patrones de comportamiento de los usuarios sobre el motor de búsquedas del portal, analizando aquellas búsquedas que terminan en pedido y las que no.

Ejemplos

Goldcar

Objetivo: Reducción de gastos anticipándose a problemas derivados de la demanda.

Problema planteado: Las reservas a través de su portal online que no terminan en alquiler generan grandes gastos. Como no se requiere pago previo para realizar una reserva, muchos usuarios no se presentar a recoger el coche reservado. Esto genera grandes gastos por los coches que quedan esperando a esos usuarios que nunca llegan.

Solución: Por medio del análisis de los datos de las reservan se identifican perfiles concretos que terminan en casos de reservas canceladas o clientes no presentados, en función de la la procedencia, la temporada, antelación y otros factores clave.

Banca: Identificación de personas con las compras de tarjetas de crédito

http://elpais.com/elpais/2015/01/29/ciencia/1422520042_066660.html

PRIVACIDAD EN INTERNET »

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT

Science

AAAS

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

Yves-Alexandre de Montjoye,[1]* Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex "Sandy" Pentland[1]

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that women are more reidentifiable than men in credit card metadata.

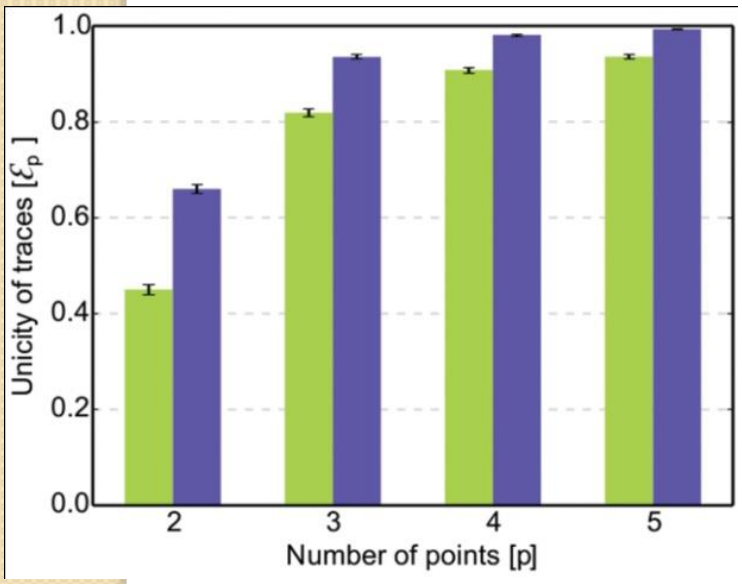
http://www.sciencemag.org/content/347/6221/536

Banca: Identificación de personas con las compras de tarjetas de crédito

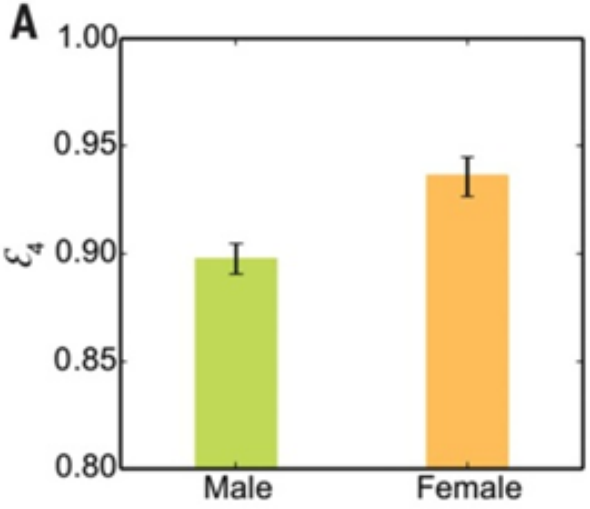
PRIVACIDAD EN INTERNET »

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

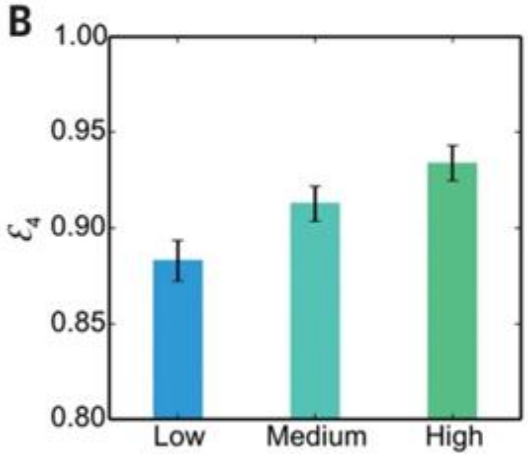
- Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT



Identificación por el número de compras

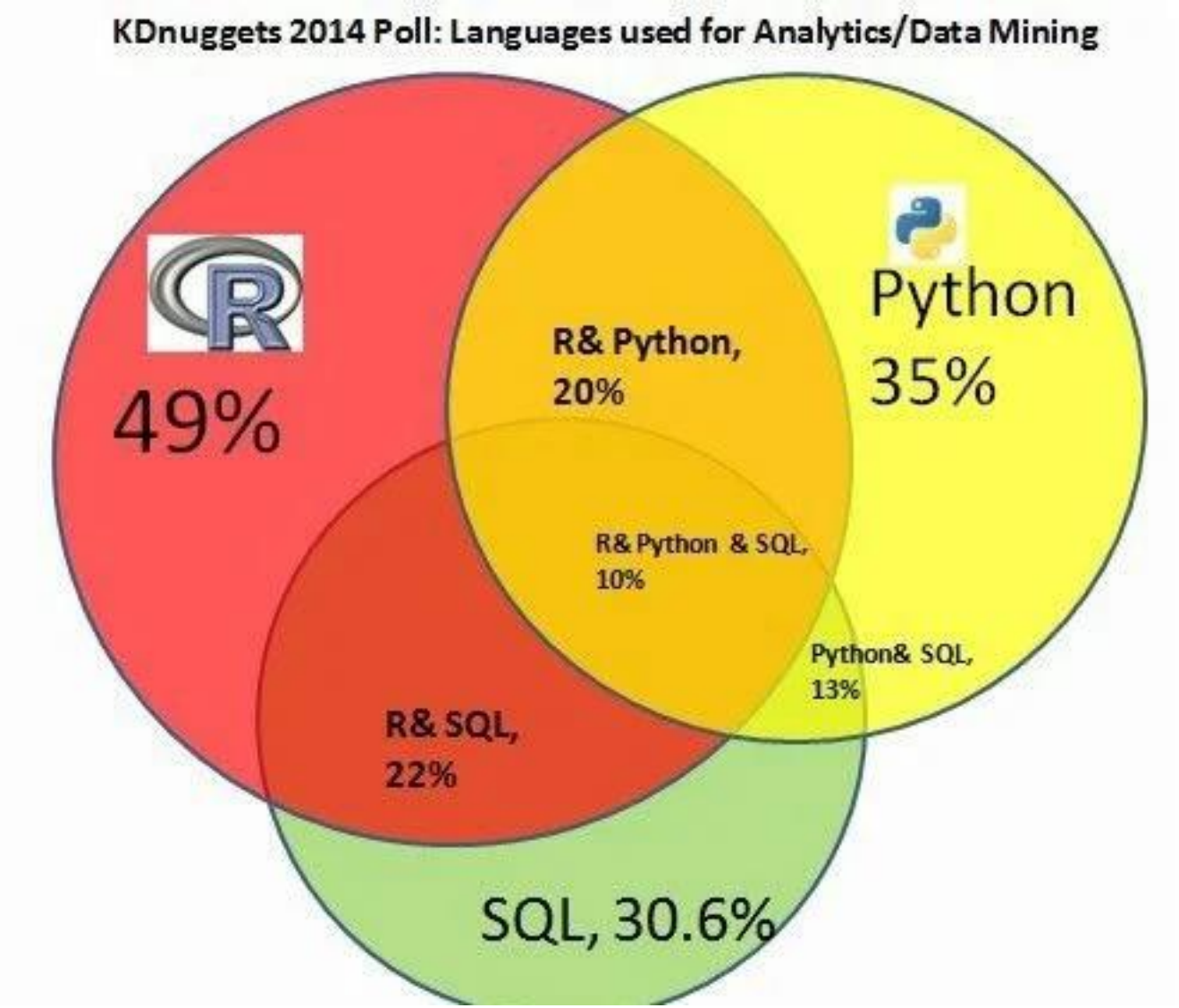


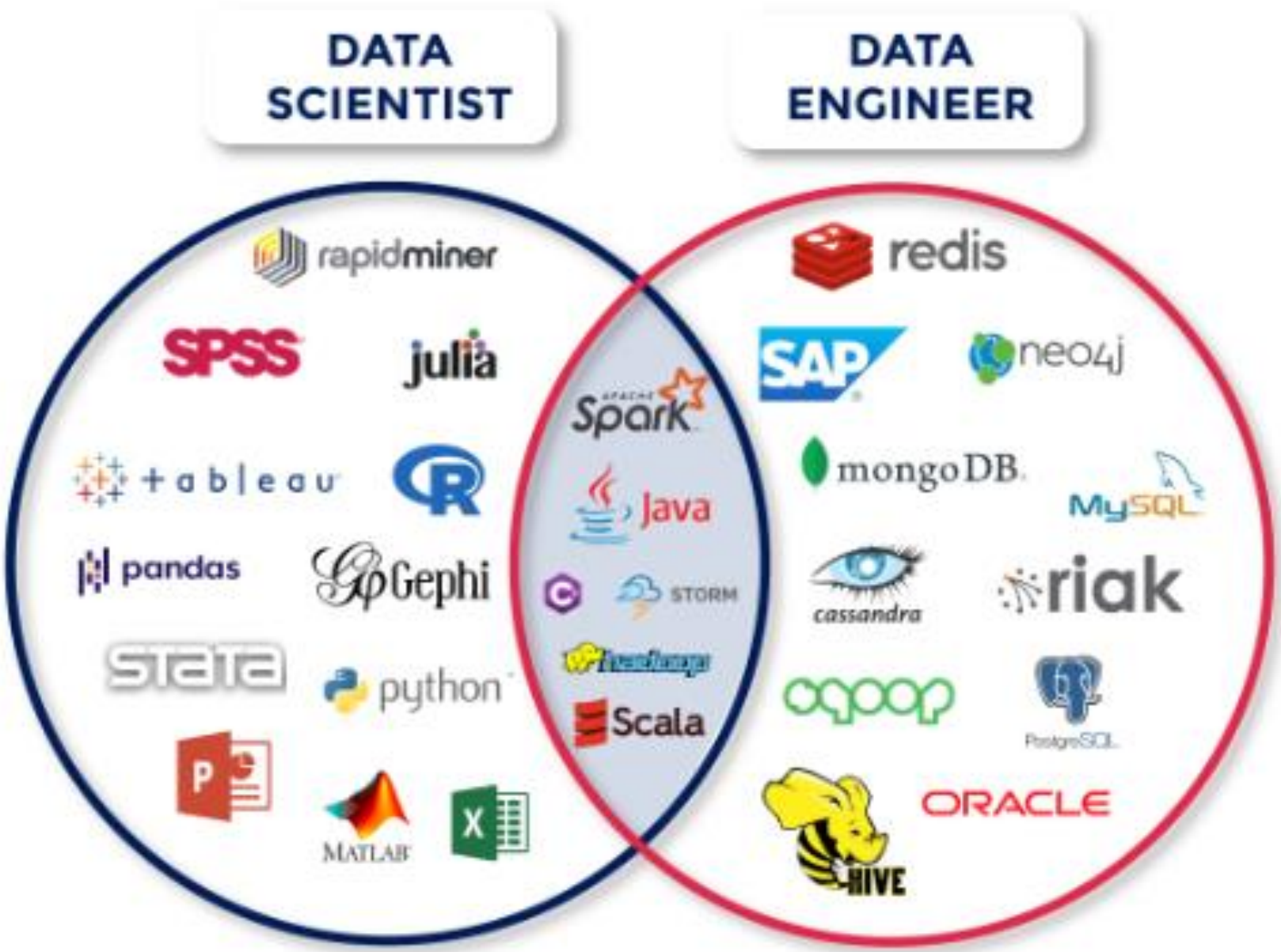
Identificación por el género



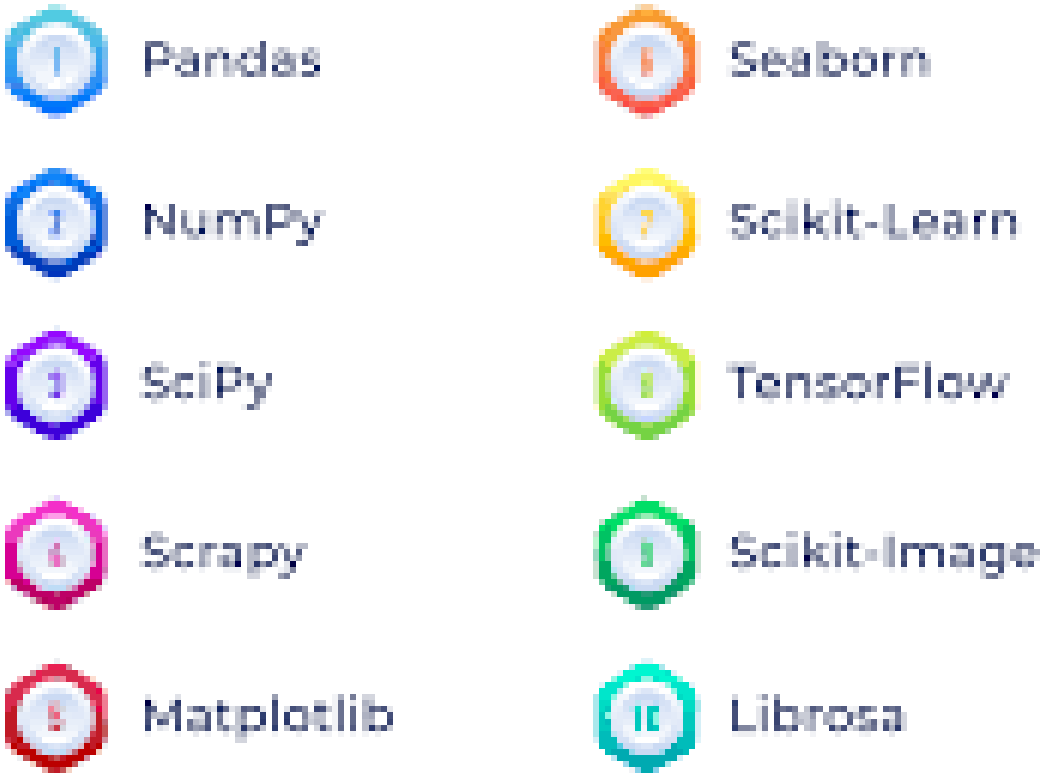
Identificación por el poder adquisitivo

Herramientas y Lenguajes de Programación



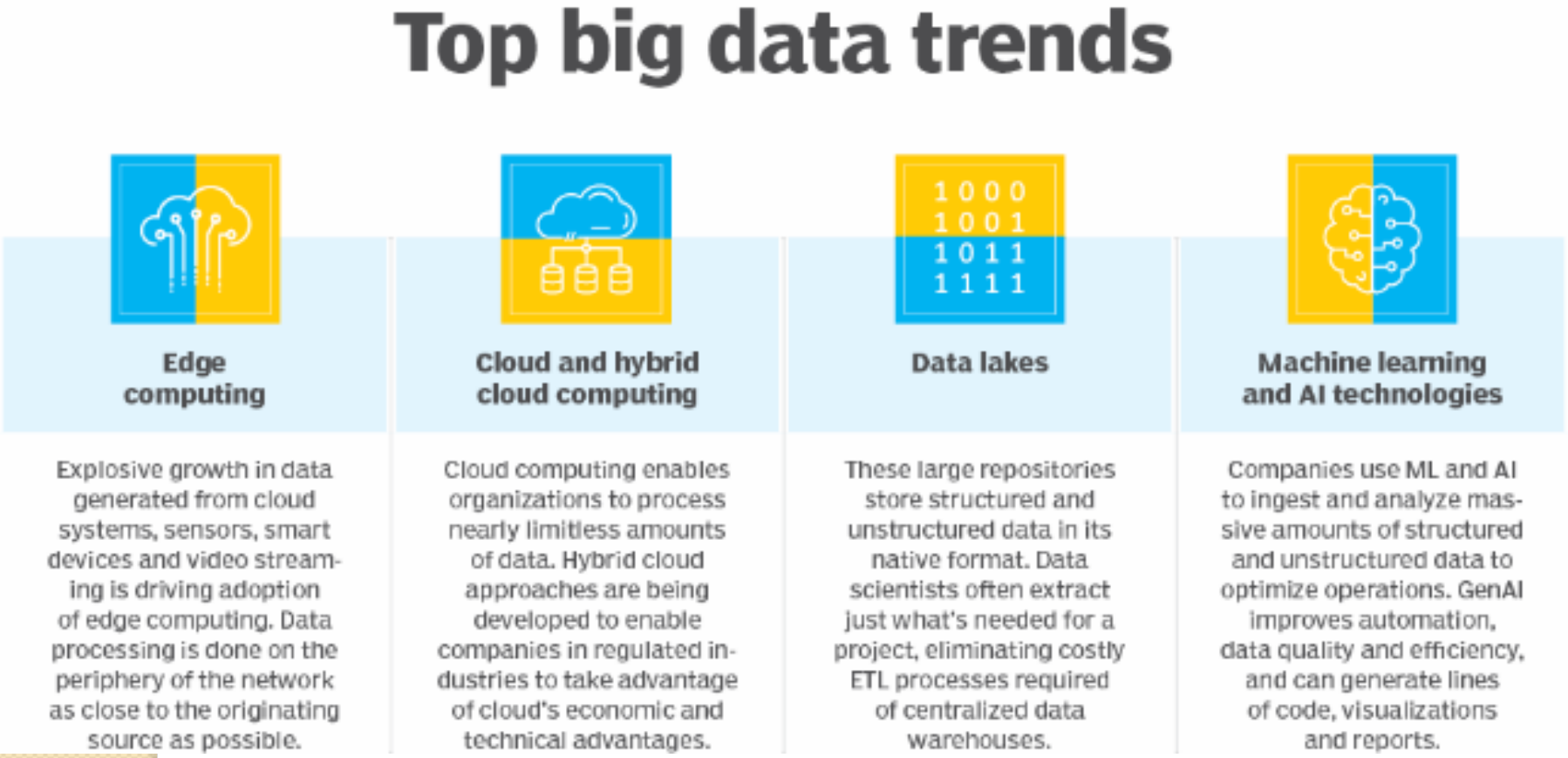

DATA SCIENTIST
DATA ENGINEER
rapidminer
SPSS
julia
+ableau
pandas
Gephi
STATA
python
MATLAB
Spark
Java
STORM
Scala
redis
SAP
neo4j
mongoDB.
MySQL
cassandra
riak
oqoop
PostgreSQL
ORACLE
HIVE

Principales bibliotecas de Python



Data Science

Desafíos y Tendencias



Desafíos y Tendencias

Machine learning and AI systems

- Recognition systems for image, video and text data.
- Automated classification of data.
- Natural language processing (NLP) capabilities for chatbots and voice and text analysis.
- Autonomous business process automation.
- Personalization and recommendation features in websites and services.
- Analytics systems that can find optimal solutions to business problems among a sea of data.

Fin