MACHINE LEARNING & BIG DATA

Conceptos básicos, Estadística, Exploración y Data Quality

Instructor: José Nelson Zepeda

San Salvador, octubre 2018

Fundamentos Machine Learning

Conceptos Básicos

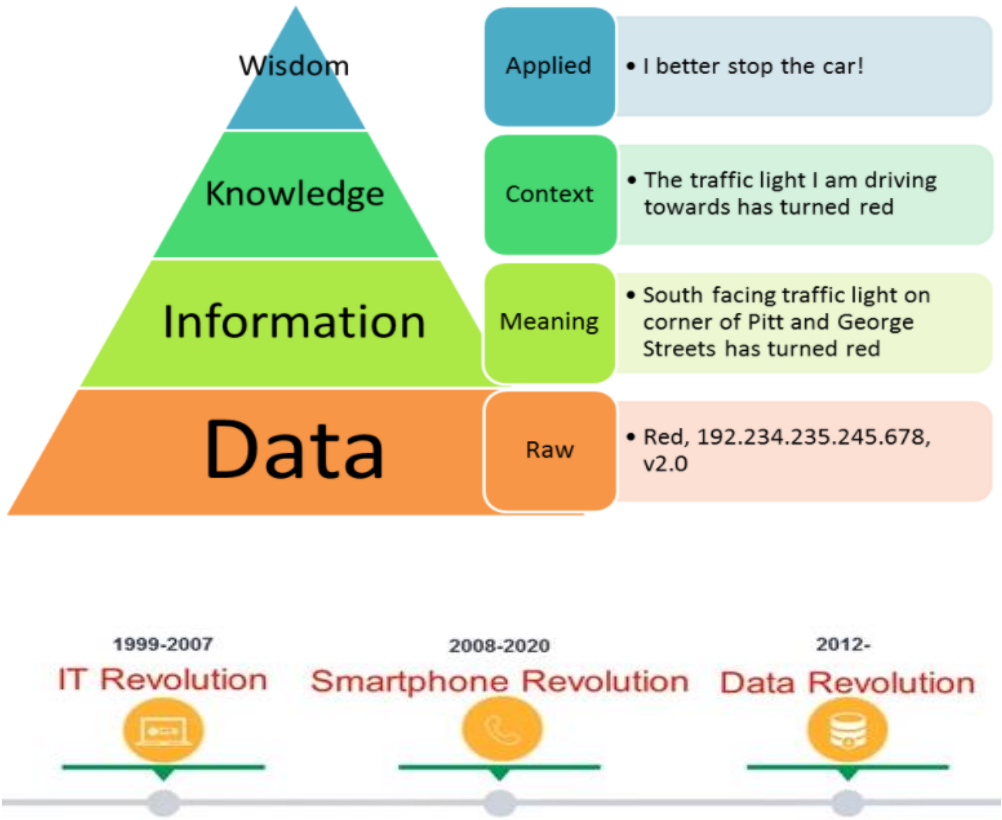
Estadística Básica

Análisis Exploratorio

Data Quality

Conceptos Básicos

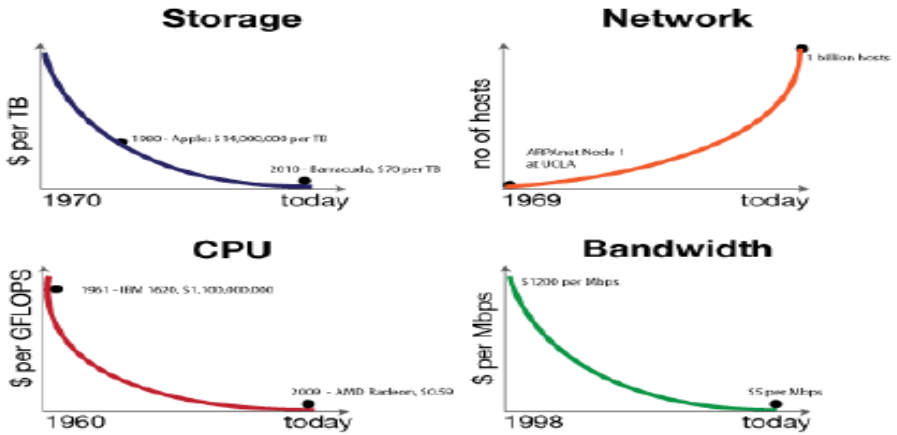
¿Qué es Data?



Data is the **seed** from which information, knowledge and wisdom sprouts and blossoms.

Data is the **key** to answer the right question

Data is a **set of values** of qualitative or quantitative variables.

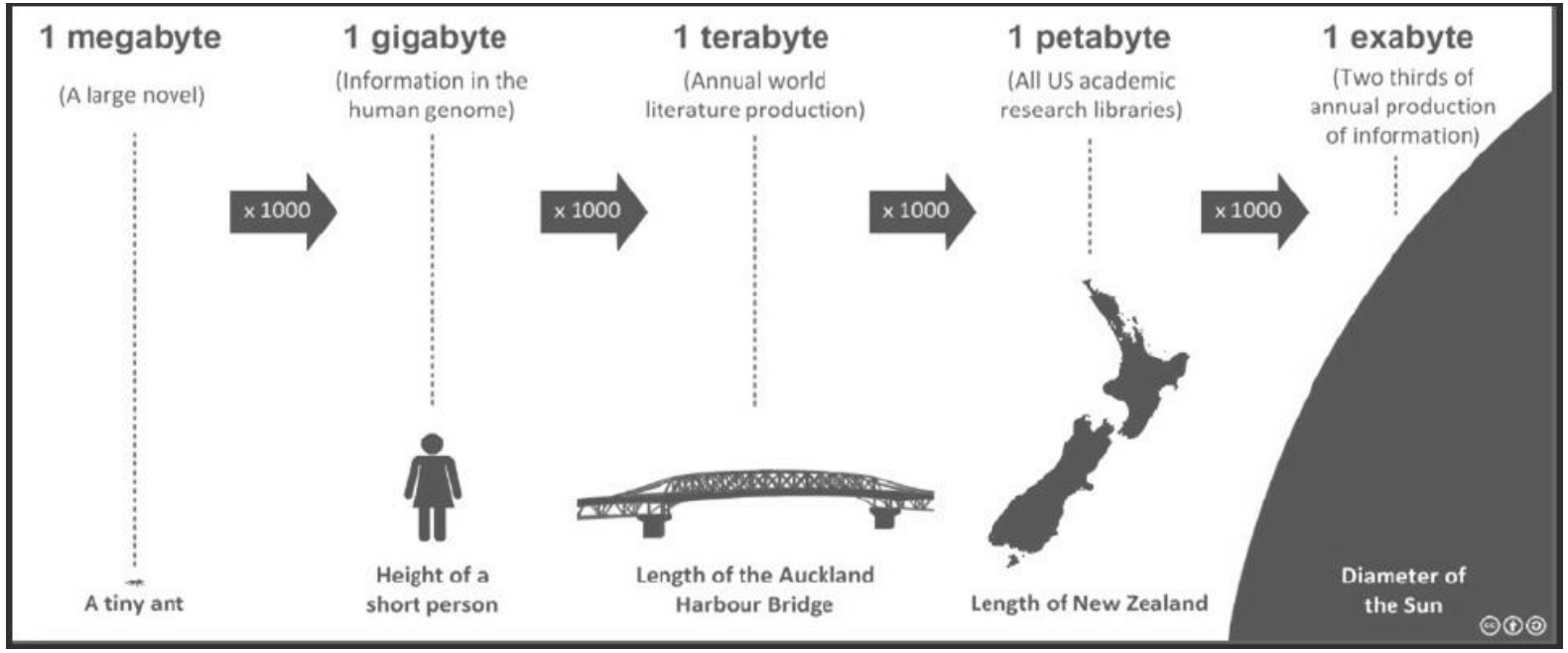


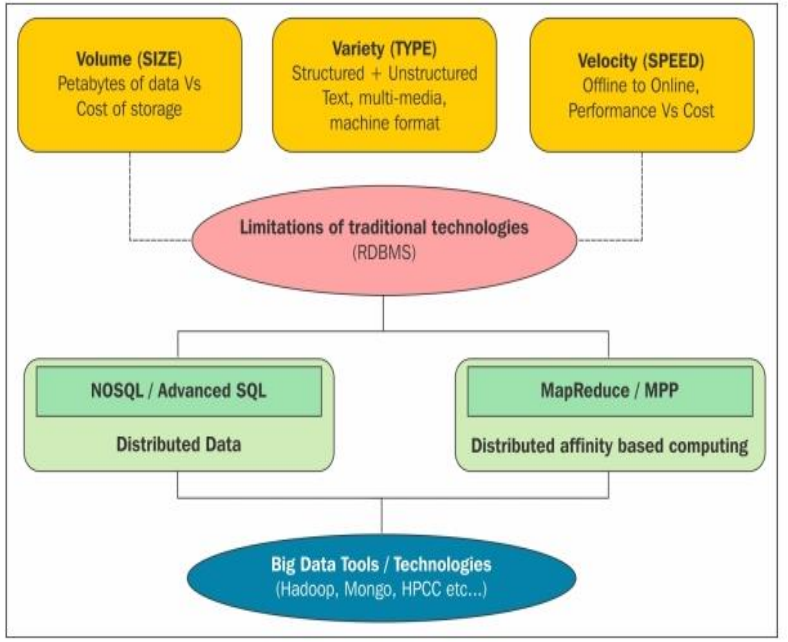
https://www.youtube.com/watch?v=jbkSRLYSojo&t=2s

¿Qué es Big Data?

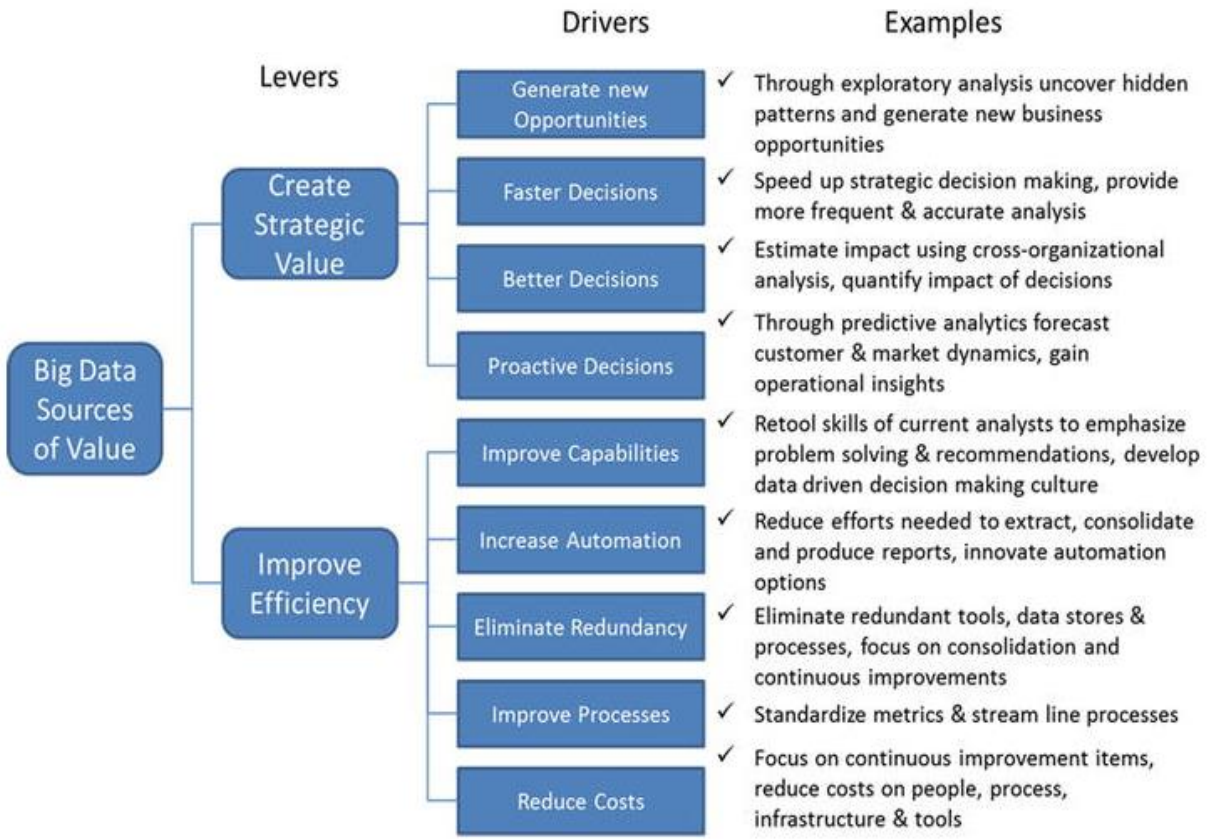
"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..."

Dan Ariely, Duke University





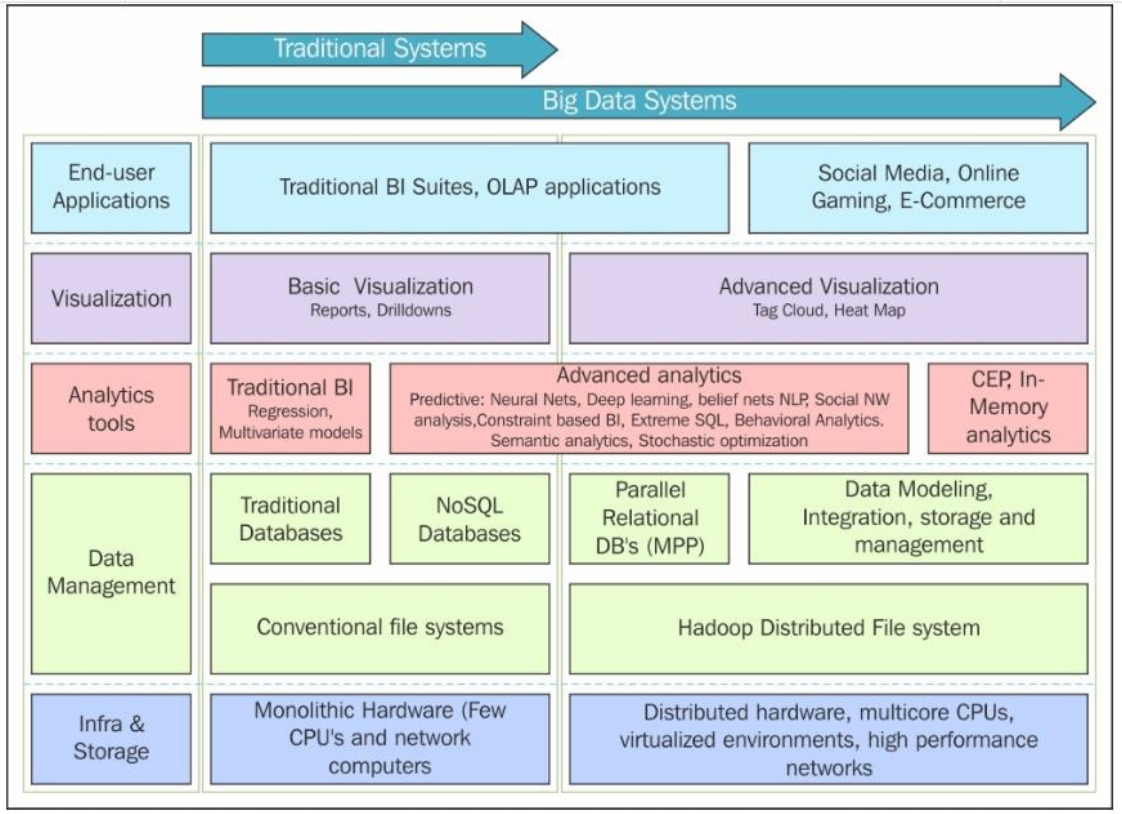
Casos de Uso de Big Data



Retail		Manufacturing	
✓ Customer Relationship Management ✓ Store Location & Layout	✓ Fraud Detection & Prevention ✓ Supply-Chain optimization ✓ Dynamic Pricing	✓ Product Research ✓ Engineering Analysis ✓ Predictive Maintenance	✓ Process & Quality Metrics ✓ Distribution Optimization
Financial Services		**Media & Telecommunications**	
✓ Algorithmic Trading ✓ Risk Analysis	✓ Fraud Detection ✓ Portfolio Analysis	✓ Network Optimization ✓ Customer Scoring	✓ Churn Prevention ✓ Fraud Prevention
Advertising & Public Relations		**Energy**	
✓ Demand Signaling ✓ Targeted Advertising	✓ Sentiment Analysis ✓ Customer Acquisition	✓ Smart Grid ✓ Exploration	✓ Operational Modeling ✓ Power-Line Sensors
Government		**Healthcare & Life Sciences**	
✓ Market Governance ✓ Weapon Systems & Counter Terrorism	✓ Econometrics ✓ Health Informatics	✓ Pharmacogenomics ✓ Bioinformatics	✓ Pharmaceutical Research ✓ Clinical Outcomes Research

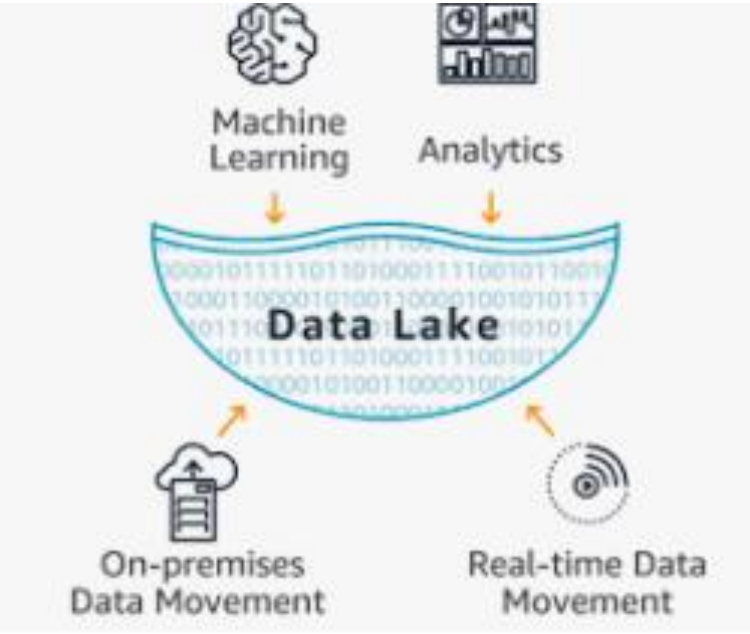
Diferencias entre DWH/BI y Big Data

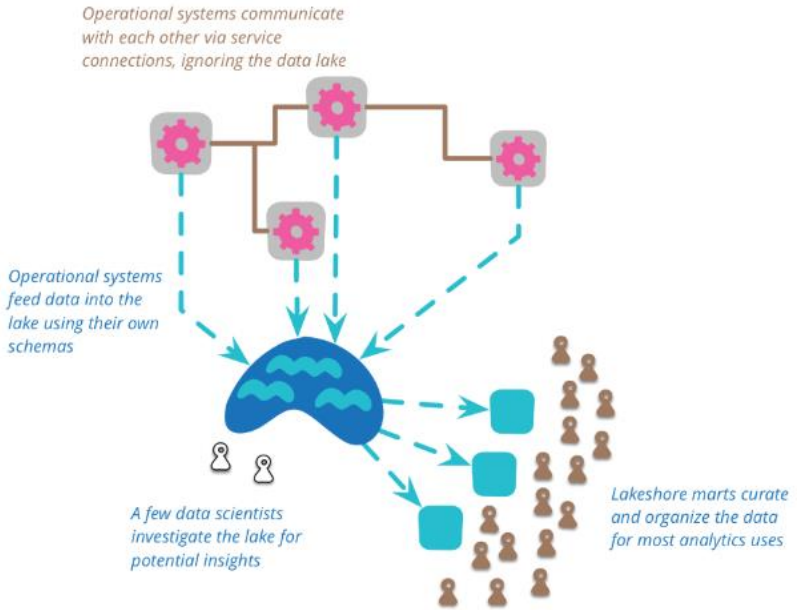
- La escalabilidad del almacenamiento y el poder de procesamiento son diferentes.
- En el enfoque tradicional, la data proviene de sistemas relacionales y estructurados, en la nueva era del Big Data la data puede provenir de todo tipo de fuentes incluyendo las no estructuradas.
- La velocidad de procesamiento de los sistemas tradicionales es menor.
- La complejidad de los algoritmos que se pueden aplicar sobre la data.
- El enfoque tradicional ofrece reporteria y cubos con drill-downs, el nuevo enfoque es mucho más visual incluyendo mapas de calor, graficas de N dimensiones, etc. El Story teller es una realidad y una necesidad.



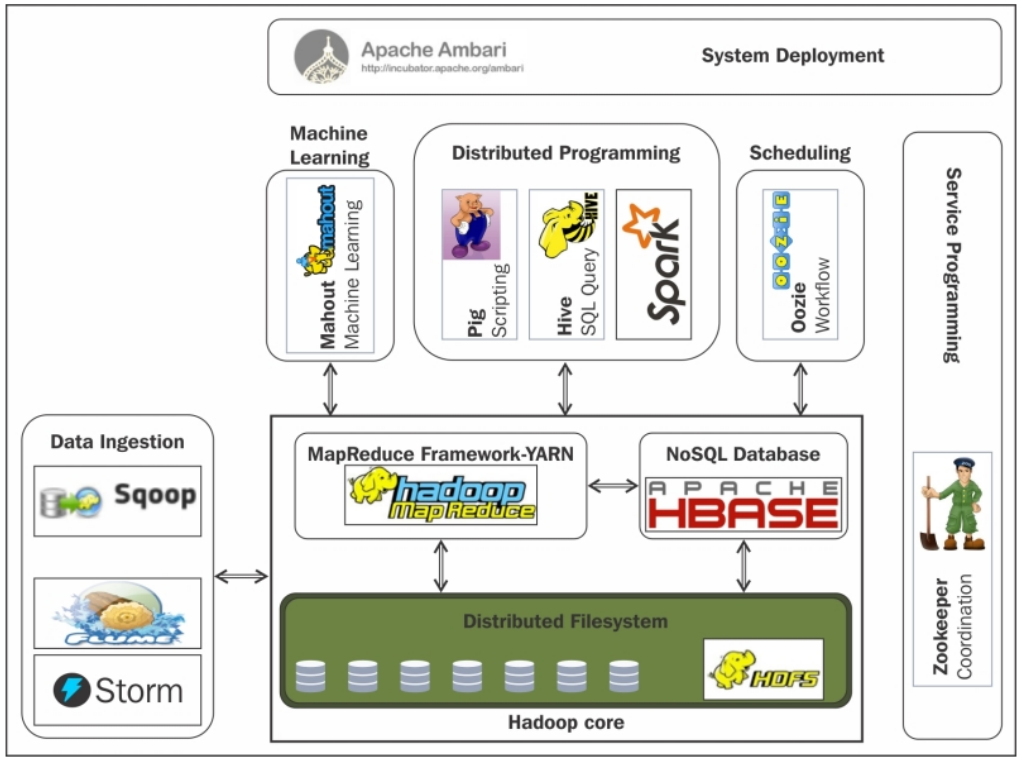
Que es un Datalake

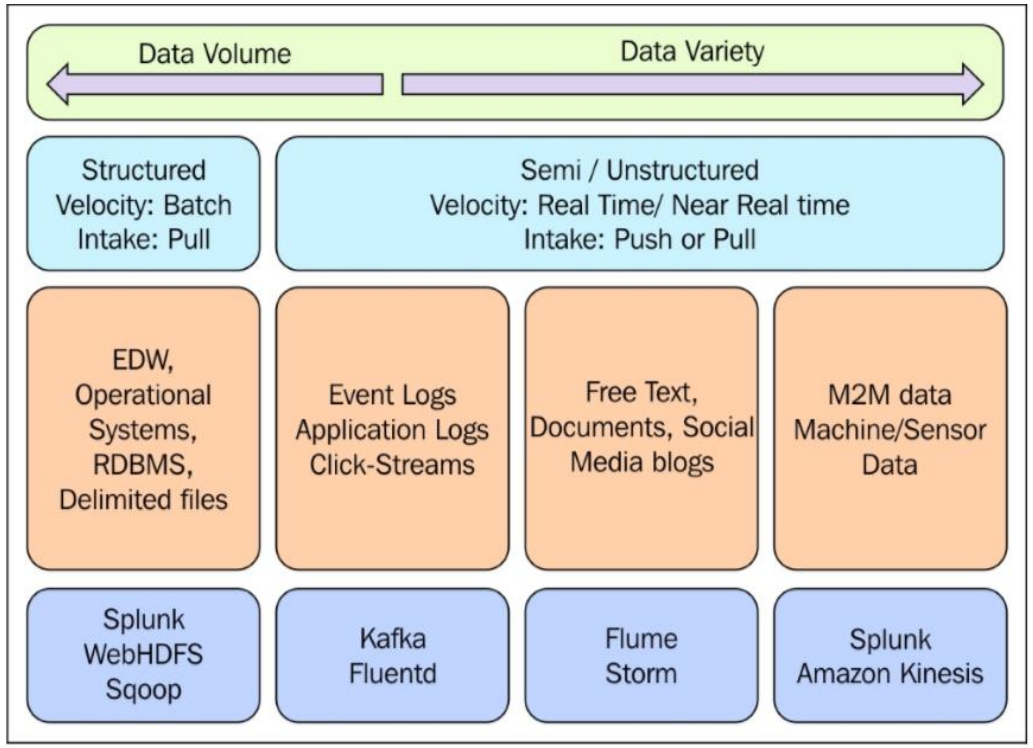
Un Data Lake, es un repositorio que almacena una gran cantidad de datos estructurados, semi-estructurados y no estructurados en su formato natural, es decir todo está almacenado de forma plana y los datos se van procesando/preparando según sea necesario. Debe ser reconocido como un punto de integración de la data para propósitos de análisis, no como un puente o colaboración entre los sistemas operacionales





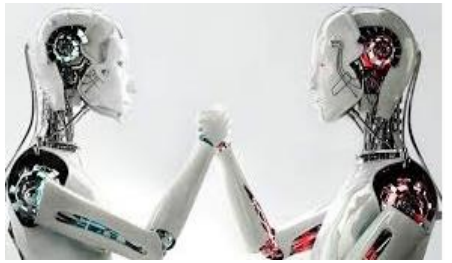
Ecosistema en el Datalake (Hadoop)





Que no es Machine Learning

Supongamos que tienes un problema de Machine Learning que debes resolver, sin embargo, no conoces que es Machine Learning. Empezaremos por decirte lo que no es:



No es una investigación sobre las capacidades de un algoritmo.

No es el desarrollo de un algoritmo o de alguna teoria.

No es una investigación esoterica de algun tipo de aprendizaje.

No es la construcción de un agente de inteligencia artificial

No es la construcción de un circuito que emita señales

https://machinelearningmastery.com/think-machine-learning/
http://blogs.teradata.com/data-points/todays-machine-learning-ai-not/

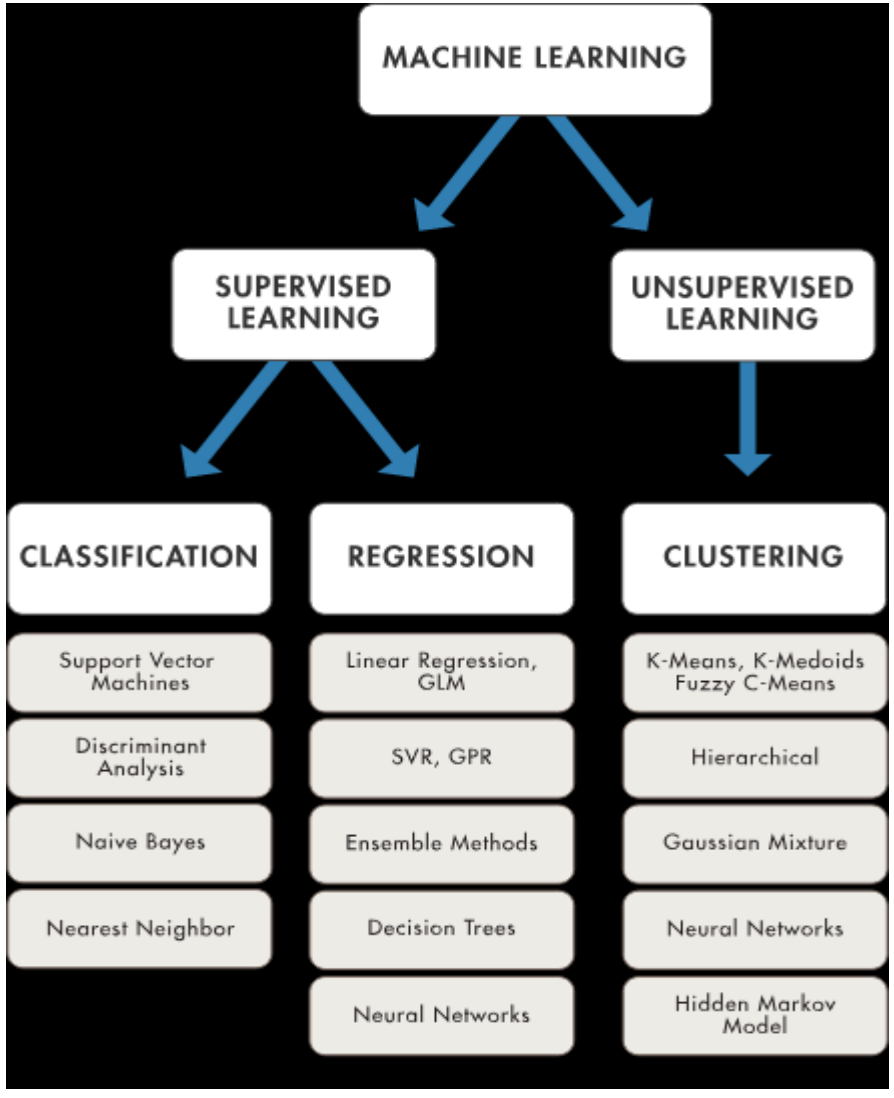
Machine Learning

Machine Learning son un conjunto de métodos/algoritmos diseñados para encontrar patrones y tendencias en los datos. Se encuentra en la intersección entre las matemáticas y estadística con la ingeniería de software y ciencias de la computación.

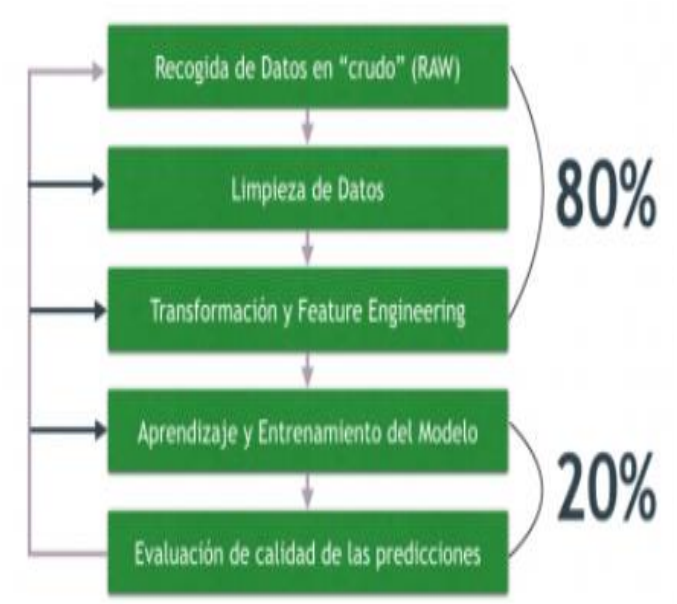
Familias de técnicas de ML

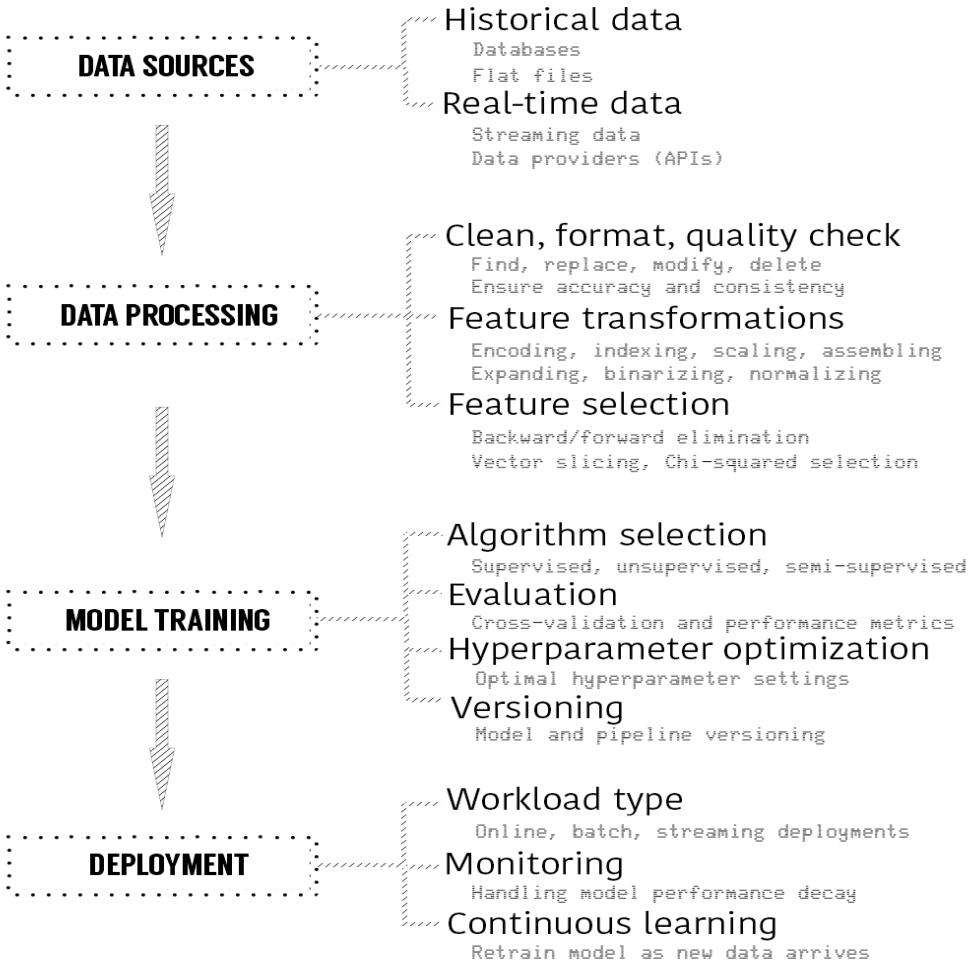
1. Aprendizaje Supervisado: En este proceso de aprendizaje la variable de salida está bien definida (variable objetivo), es decir estas técnicas nos son útiles cuando nos interesa hacer predicciones sobre una variable objetivo.

2. Aprendizaje No Supervisado: Este proceso de aprendizaje no implica tener una variable objetivo bien identificada, su objetivo no es hacer predicciones.



Ciclo Vida Machine Learning

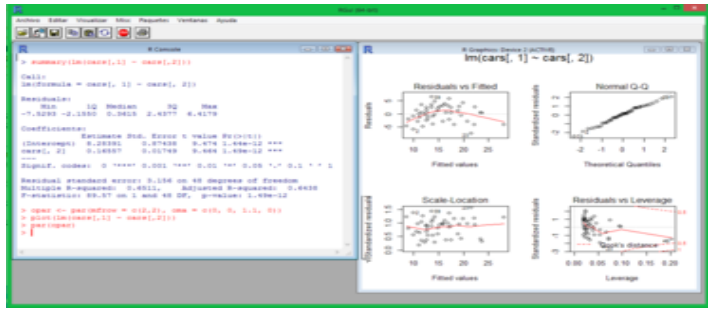




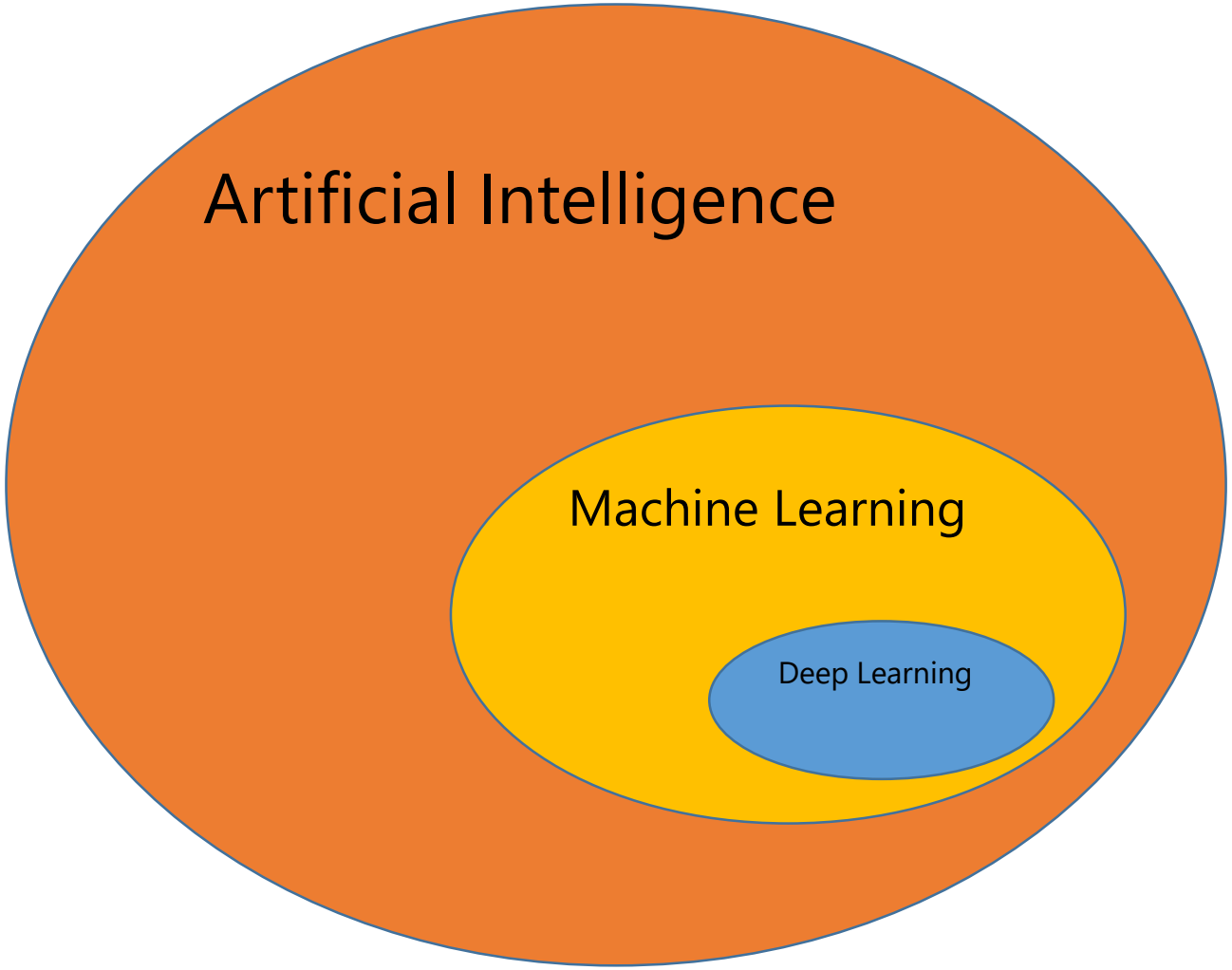
Herramientas para Machine Learning

Las herramientas para soportar las actividades de ML son una gran cantidad, entre las más populares destacan:

- Lenguaje R
- Python
- Weka
- Knime
- RapidMiner
- Azure ML Studio
- TensorFlow
- BigML
- SkyTree
- IBM Watson
- MLIB Spark
- Julia
- Jupyter



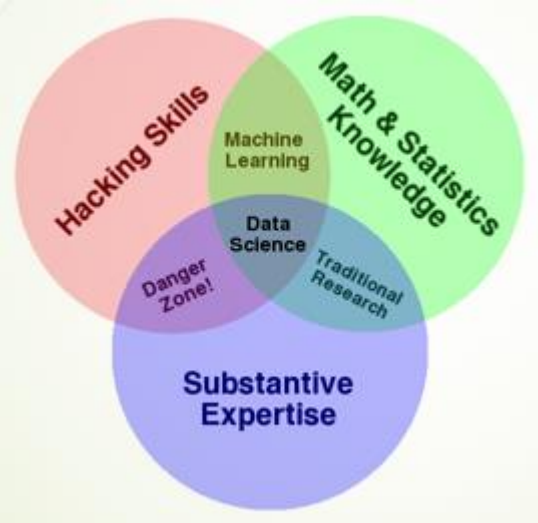
AI vs ML vs DL

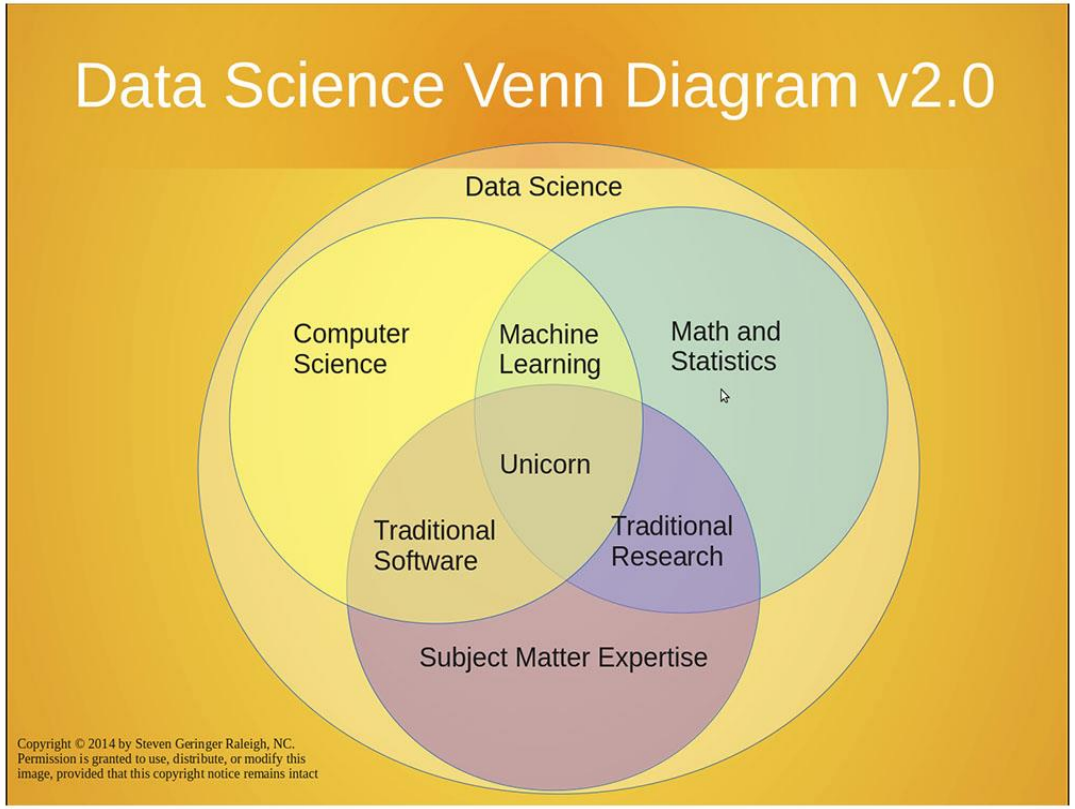


*Machine Learning is a **current application of AI** based around the idea that we should really just be able to give machines access to data and let them learn for themselves*

Deep Learning — A Technique for Implementing Machine Learning

El Científico de Datos





Estadística

Estadística: Concepto

Estadística

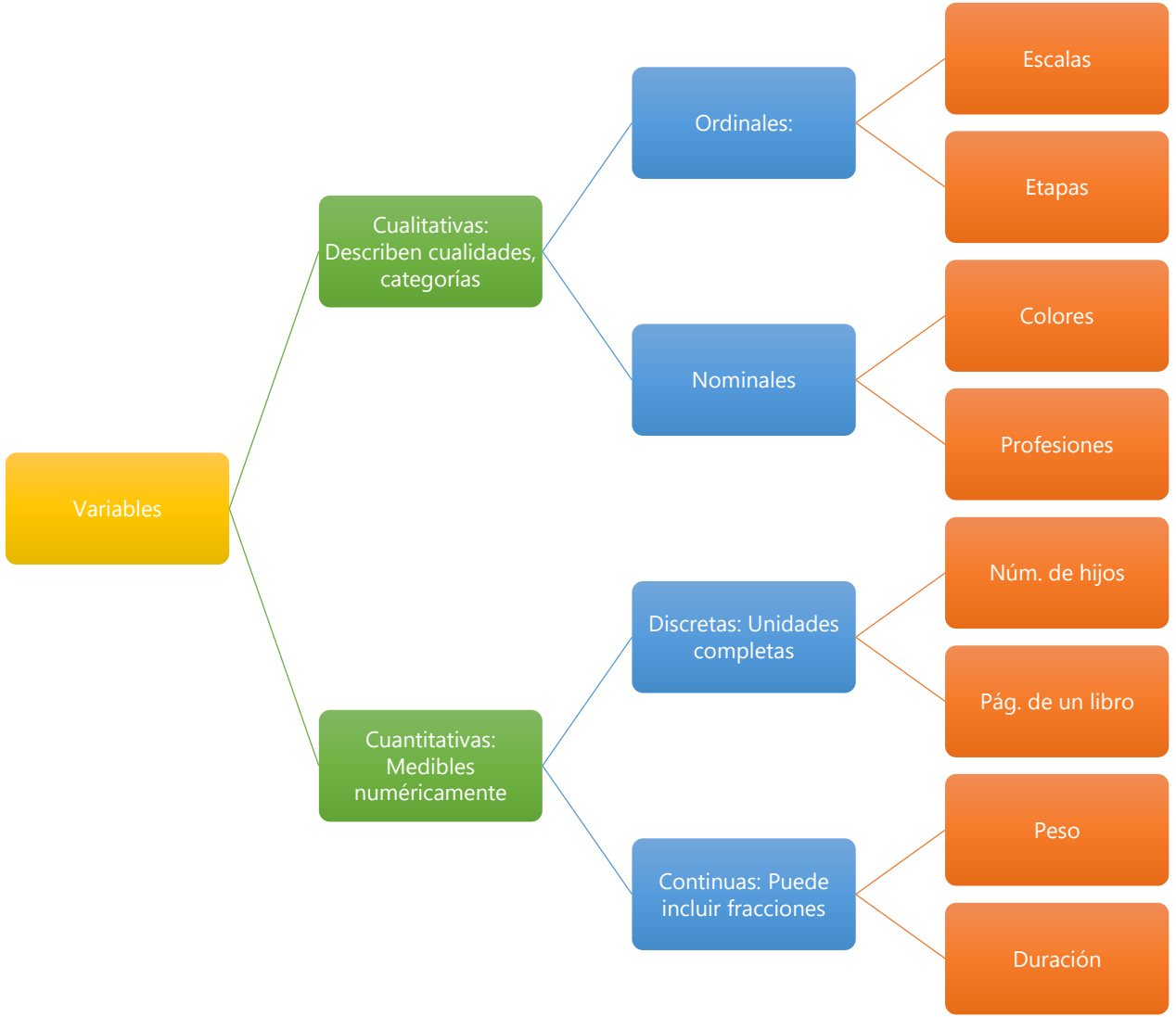
La estadística es la parte de las matemáticas que se encarga del estudio de una determinada característica de una población, recogiendo los datos, organizándolos en tablas, representándolos gráficamente y analizándolos para sacar conclusiones.

Existen dos tipos de estadística:

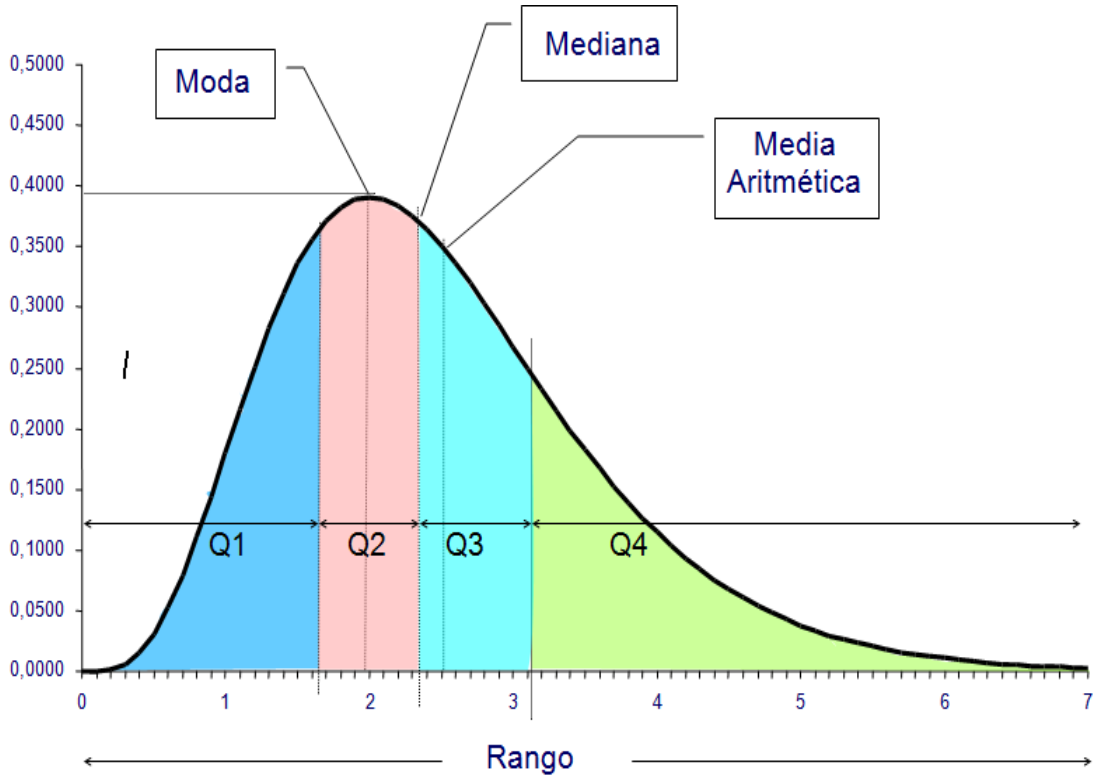
Estadística Descriptiva: Realiza estudios sobre los datos, para resumir la información de la forma más sencilla y presentable posible obteniendo así los parámetros que distinguen las características de un conjunto de observaciones, es decir, trata del recuento, ordenación, clasificación y presentación de los datos.

Estadística Inferencial: Realiza el estudio sobre un subconjunto de la población llamado muestra y, posteriormente, extiende/infiere los resultados obtenidos a toda la población. En otras palabras, la estadística inferencial utiliza los resultados de la estadística descriptiva y se apoya en el cálculo de probabilidades para la obtención de conclusiones sobre una población a partir de los resultados obtenidos de una muestra.

Estadística: Tipos de Variables

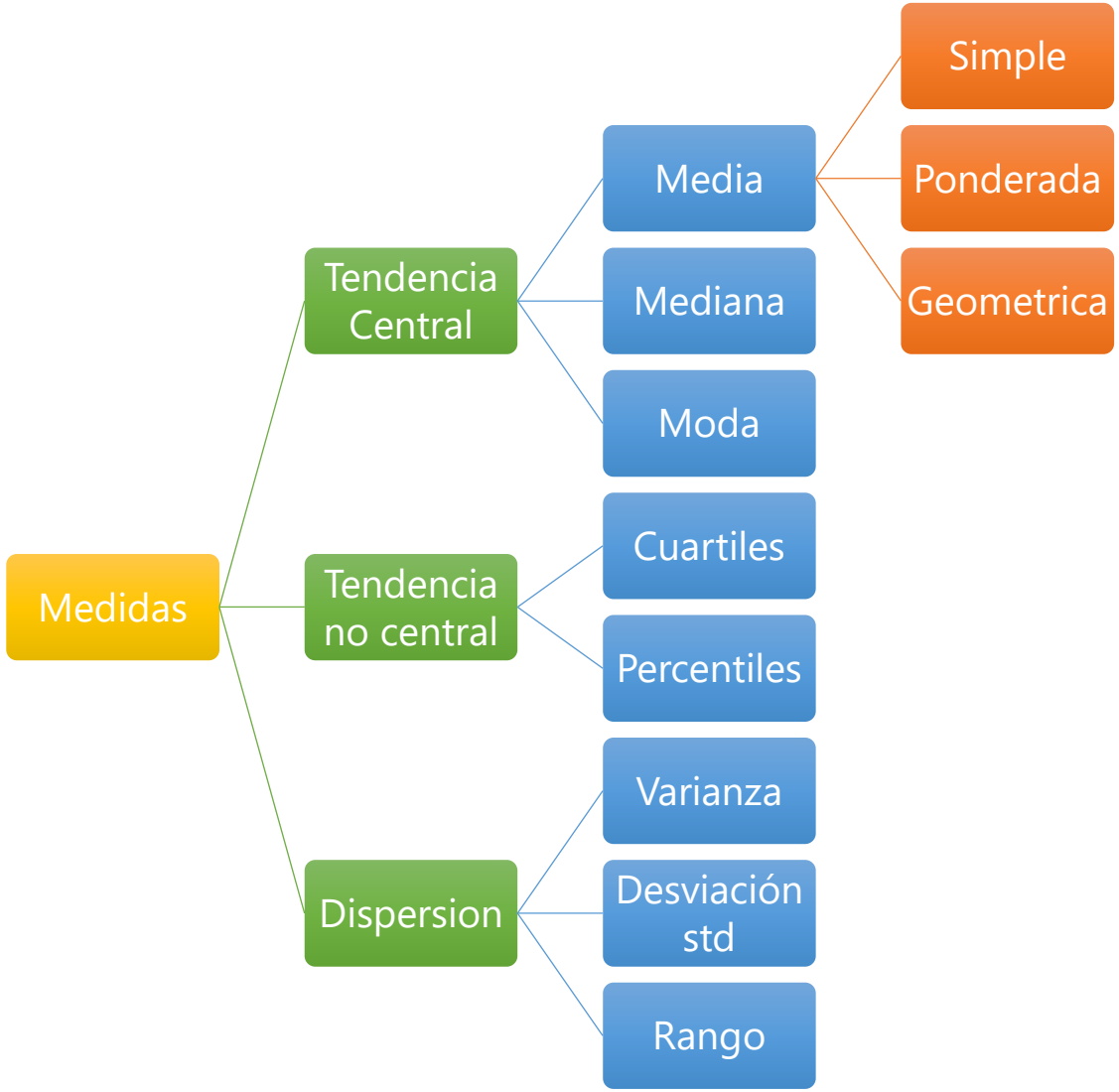


Estadística: Medidas de Tendencia Central



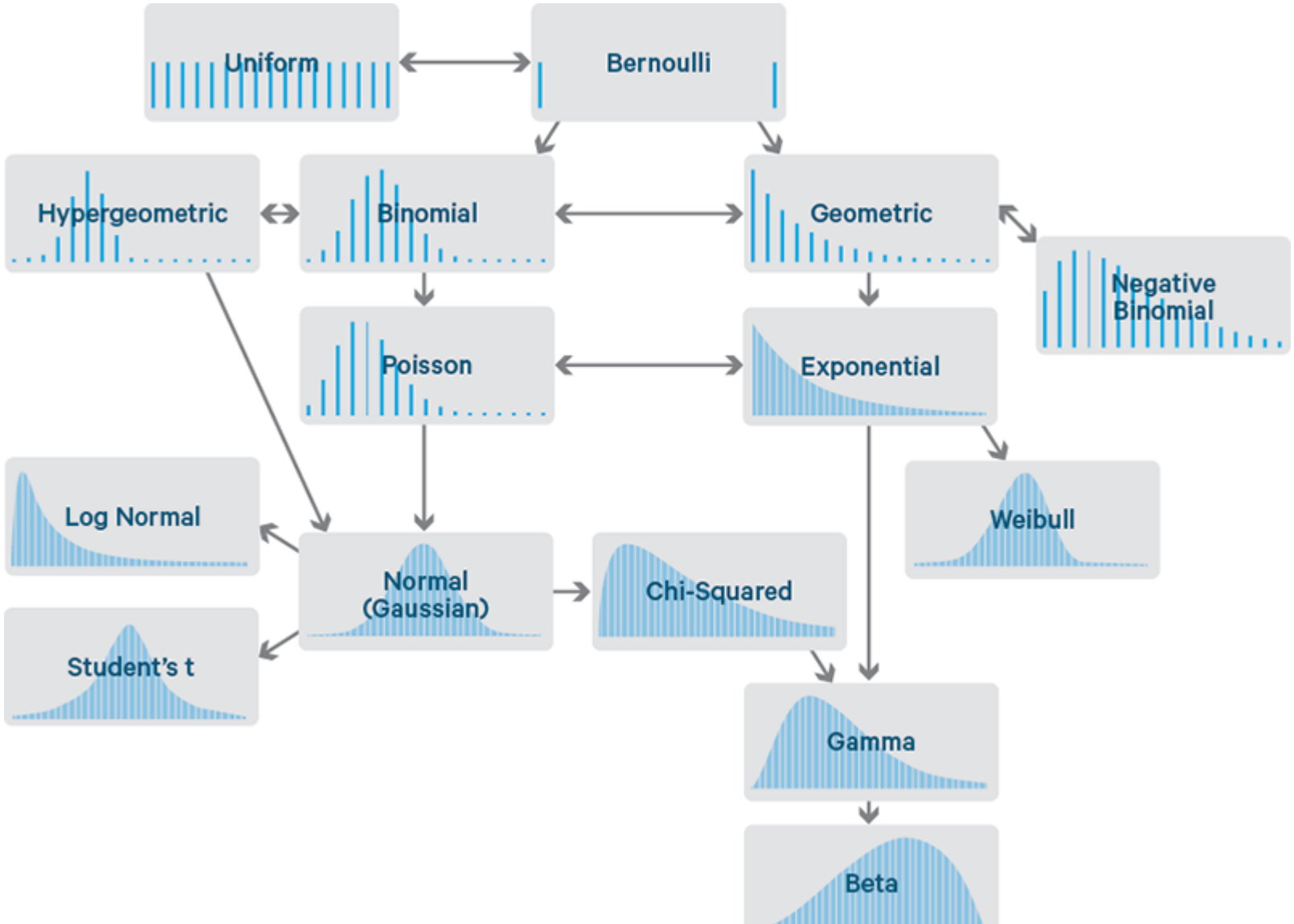
Fuente: Allende H y Ahumada S, ILI-280

Estadística: Familias de Medidas



¿Qué es una Distribución de Probabilidad?

En teoría de la probabilidad y estadística, la **distribución de probabilidad** de una variable aleatoria es una función que asigna a cada suceso definido sobre
la variable la probabilidad de que dicho suceso ocurra. La distribución de probabilidad está definida sobre el conjunto de todos los sucesos y cada uno de los sucesos es el rango de valores de la variable aleatoria.



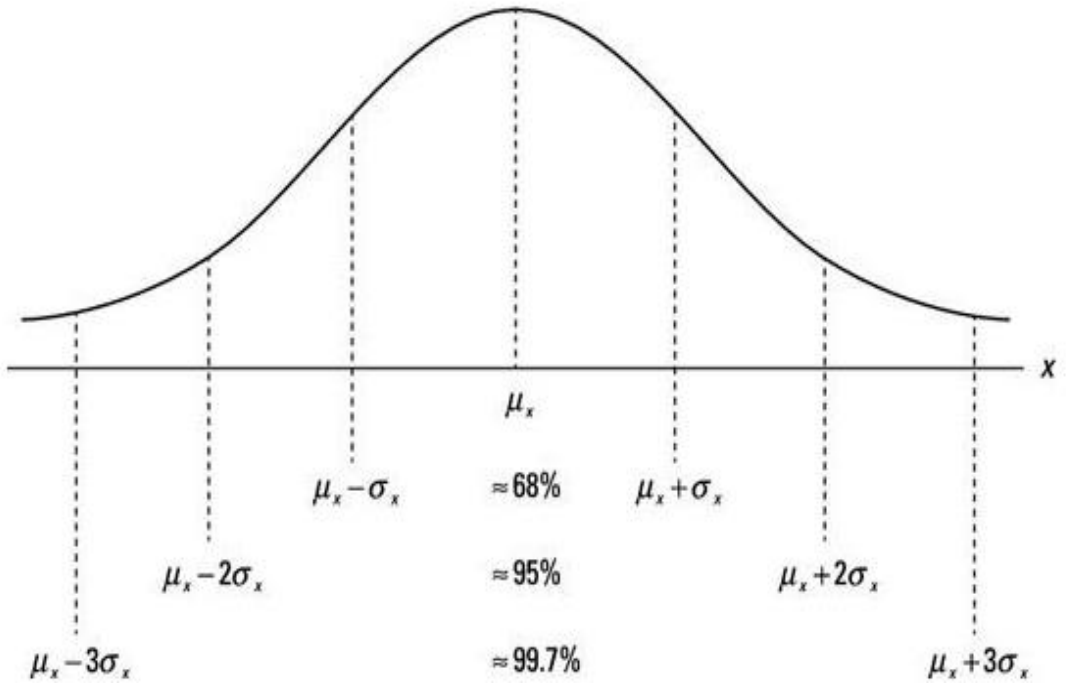
La distribución Normal

La distribución normal es la más importante de todas las distribuciones de probabilidad. Es una distribución de **variable continua** cuyo rango es del menos infinito al más infinito. La popularidad se debe a tres razones principales:

-La **gran cantidad de fenómenos reales** que se pueden modelizar con esta distribución.

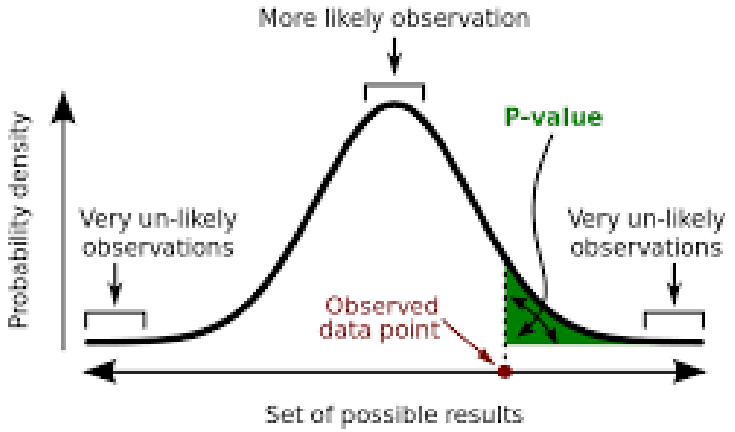
-Muchas de las distribuciones de uso frecuente **tienden a aproximarse a la distribución normal bajo** ciertas condiciones.

-En virtud del teorema central del límite, todas aquellas variables que puedan considerarse causadas por **un gran número de pequeños efectos tienden a distribuirse con una distribución normal**.

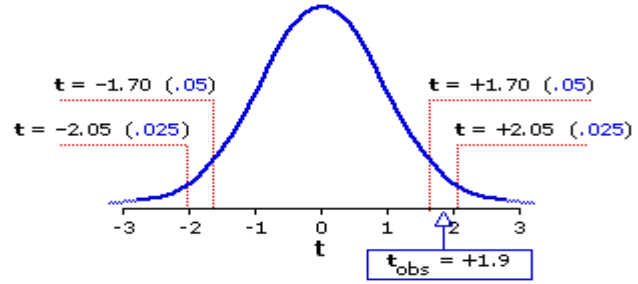


Otros Coeficientes

Valor P: El valor P es una medida de la fuerza de la evidencia en sus datos en contra de la hipótesis nula. Por lo general mientras más pequeño sea el valor P, más fuerte será la evidencia para rechazar la hipótesis nula. Tradicionalmente el valor P se compara con valores menores que **0.05 o 0.01**, dependiendo del campo de estudio.



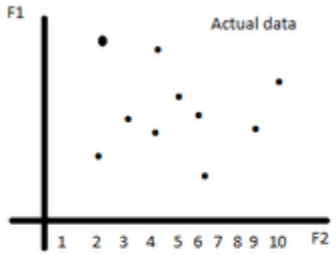
Valor T: Un valor t es el resultado de una prueba estadística. El valor se encuentra en la distribución t de Student que es apropiado para los grados de libertad. La ubicación especifica la probabilidad de obtener el valor t por casualidad. Si la probabilidad es menor que el nivel de significación, el resultado se juzga que es estadísticamente significativo. es aceptable si es **mayor que +2 y menor que -2.**

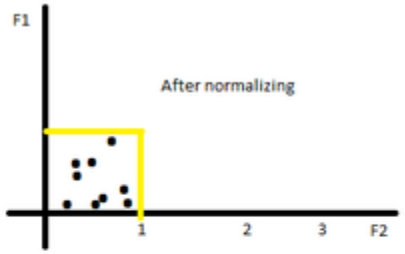


Level of Significance for a Directional Test					
	.05	.025	.01	.005	.0005
Level of Significance for a Non-Directional Test					
	---	.05	.02	.01	.001
df = 28	1.70	2.05	2.47	2.76	3.67

Estandarización

La estandarización de datos es importante ya que en la mayoría de casos nos encontraremos que dentro de un mismo set de datos, **los atributos tienen diferente naturaleza**, origen y forma de medición, en otras palabras, si los datos no son estandarizados estos no serían comparables





- Escalamiento por base decimal: Se basa en la transformación
  - $X' = X/(10 \wedge h)$, h es el parámetro que determina la intensidad del escalamiento que se aplicara, el valor transformado estará en el rango [-1,1]
- Mínimos y máximos: esta transformación se basa en el mínimo y máximo del set de datos en análisis y su salida siempre se espera en el rango [-1,1]
- Índice Z: Esta transformación se basa en el uso de la media y la desviación estándar de la variable a analizar
  - $X' = \frac{X-\mu}{\sigma}$, si la distribución es normal o cercana a esta, esta transformación devolverá valores en el rango [-3,3]

Análisis Exploratorio

Análisis Exploratorio de Datos

Independientemente de la complejidad de los datos disponibles y del procedimiento estadístico que se tenga intención de utilizar, una exploración minuciosa de los datos previa al inicio de cualquier análisis posee importantes ventajas que un analista no puede pasar por alto.

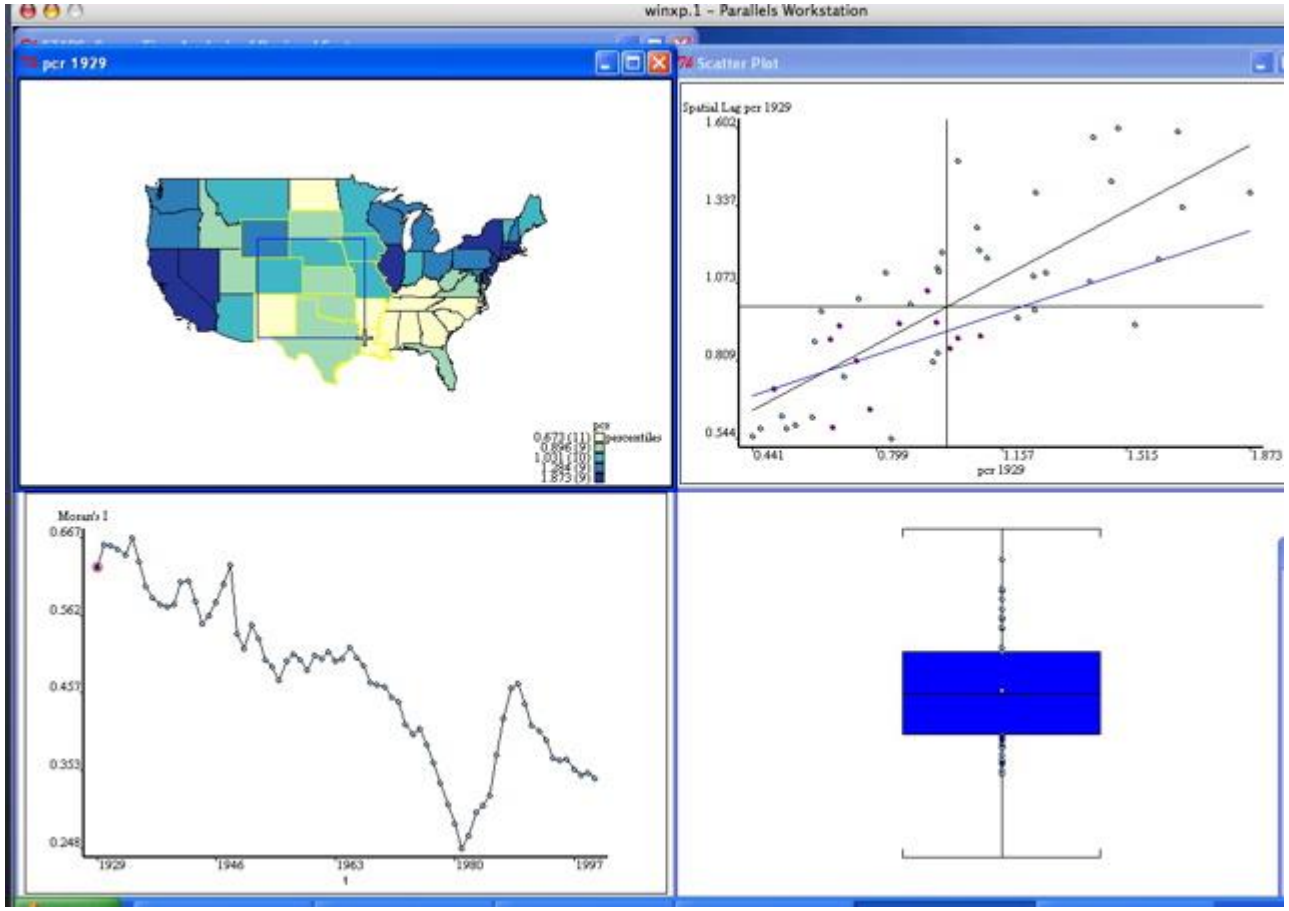
Una exploración minuciosa de los datos permite identificar entre otras cosas:

Posibles errores (datos mal introducidos, respuestas mal codificadas, etc.)

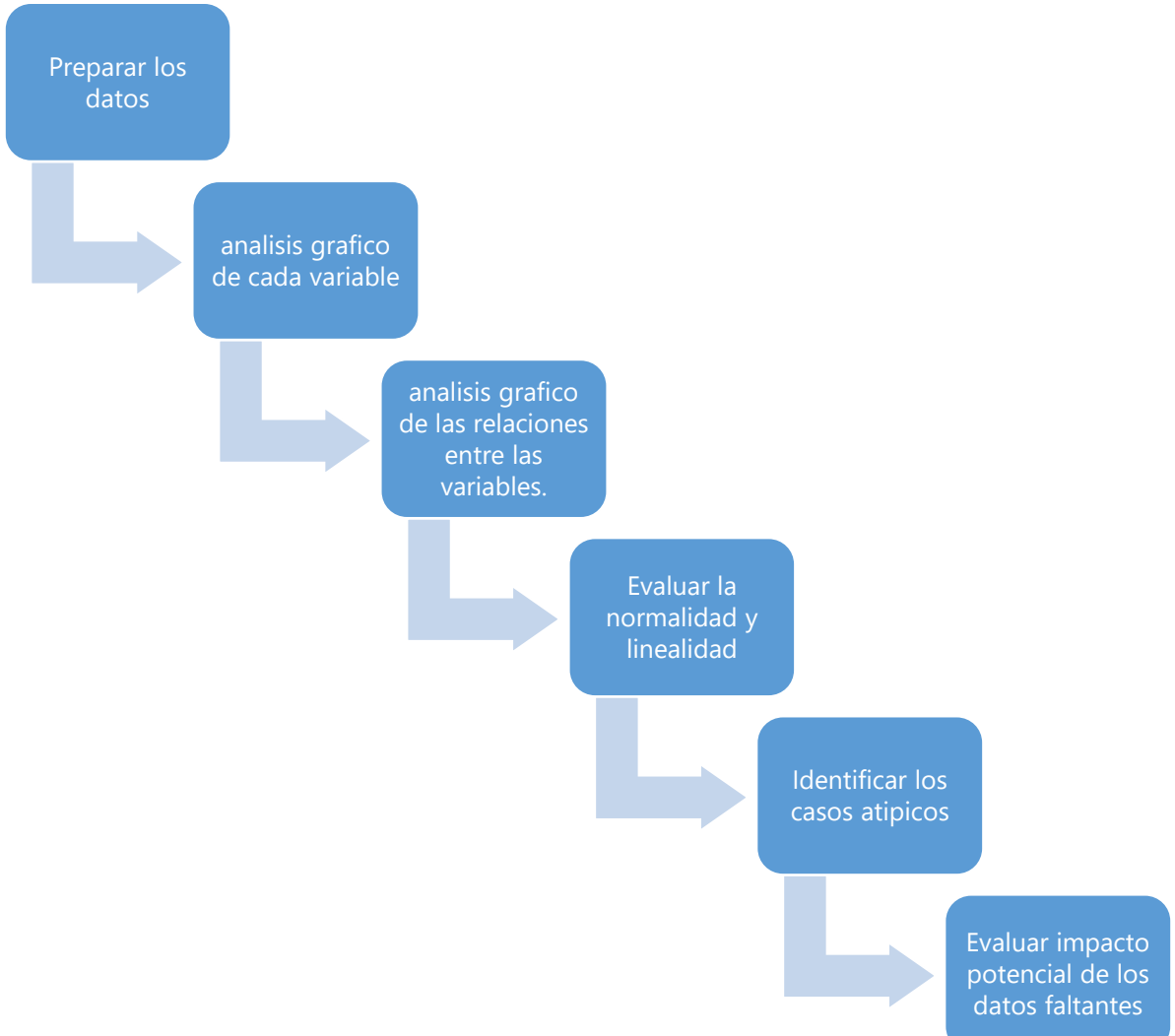
Valores extremos (valores que se alejan demasiado del centro)

Pautas extrañas en los datos (valores que se repiten demasiado o que no aparecen nunca, etc.)

Variabilidad no esperada



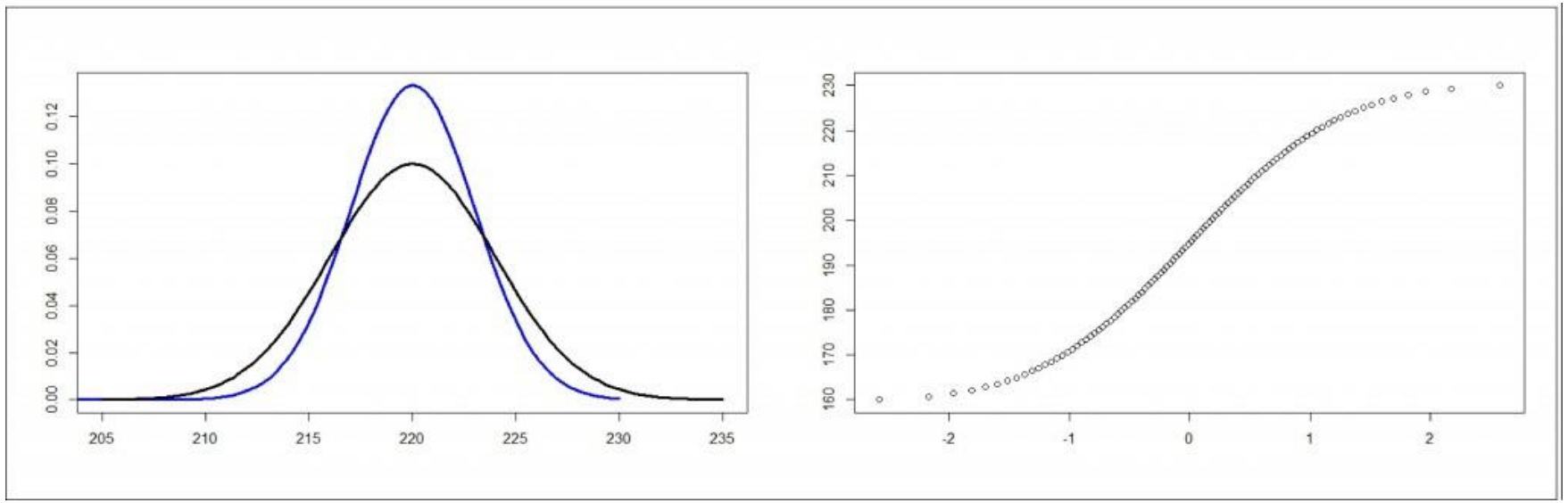
AED

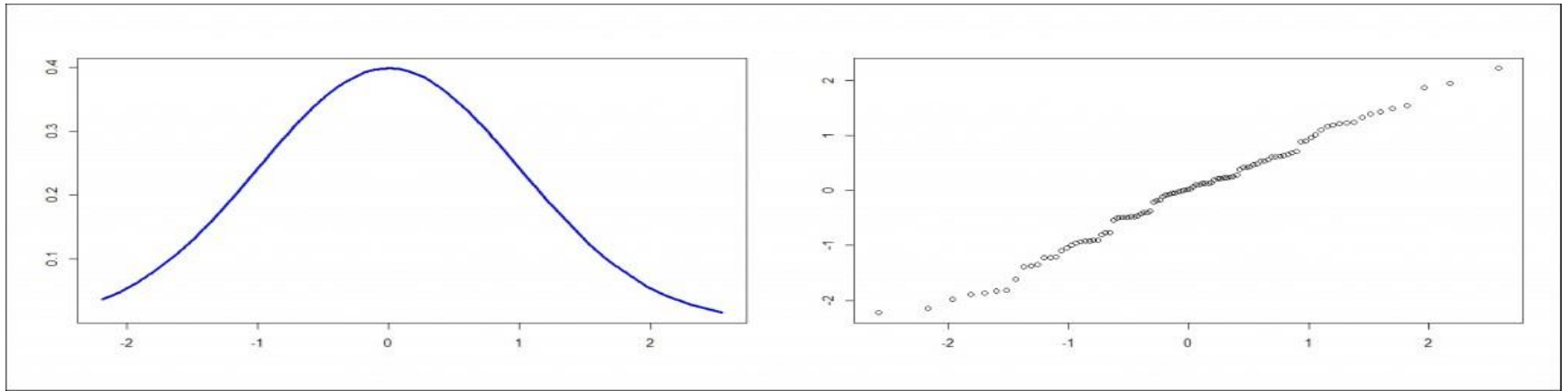

Preparar los datos
analisis grafico de cada variable
analisis grafico de las relaciones entre las variables.
Evaluar la normalidad y linealidad
Identificar los casos atipicos
Evaluar impacto potencial de los datos faltantes

AED

Escala de Medida	Tipo de Gráfica	Medidas Tendencia Central	Medidas de Dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango Intercuartilico
Intervalo	Histograma	Media	Desviación
Razón		Media Geométrica	Coefic. De Variación.

QQ Plot

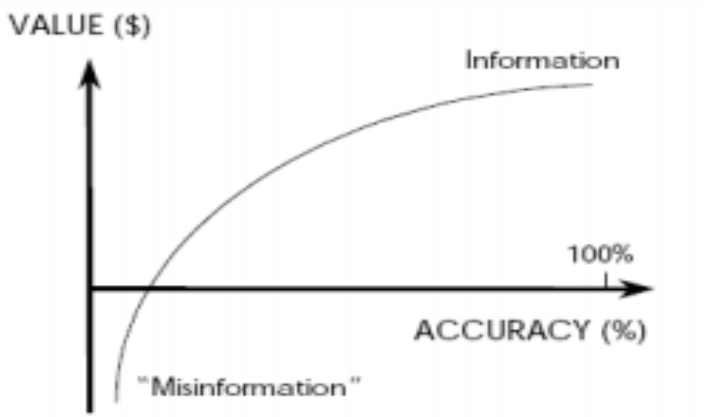

0.00
0.02
0.04
0.06
0.08
0.10
0.12
205
210
215
220
225
230
235
160
170
180
190
200
210
220
230
-2
-1
0
1
2

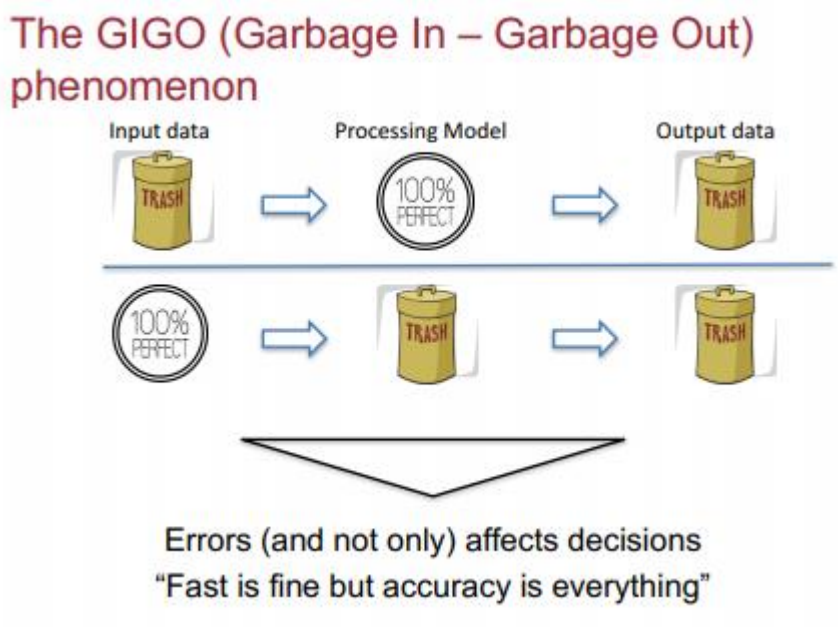

0.1
0.2
0.3
0.4
-2
-1
0
1
2
-2
-1
0
1
2
-2
-1
0
1
2

Data Quality

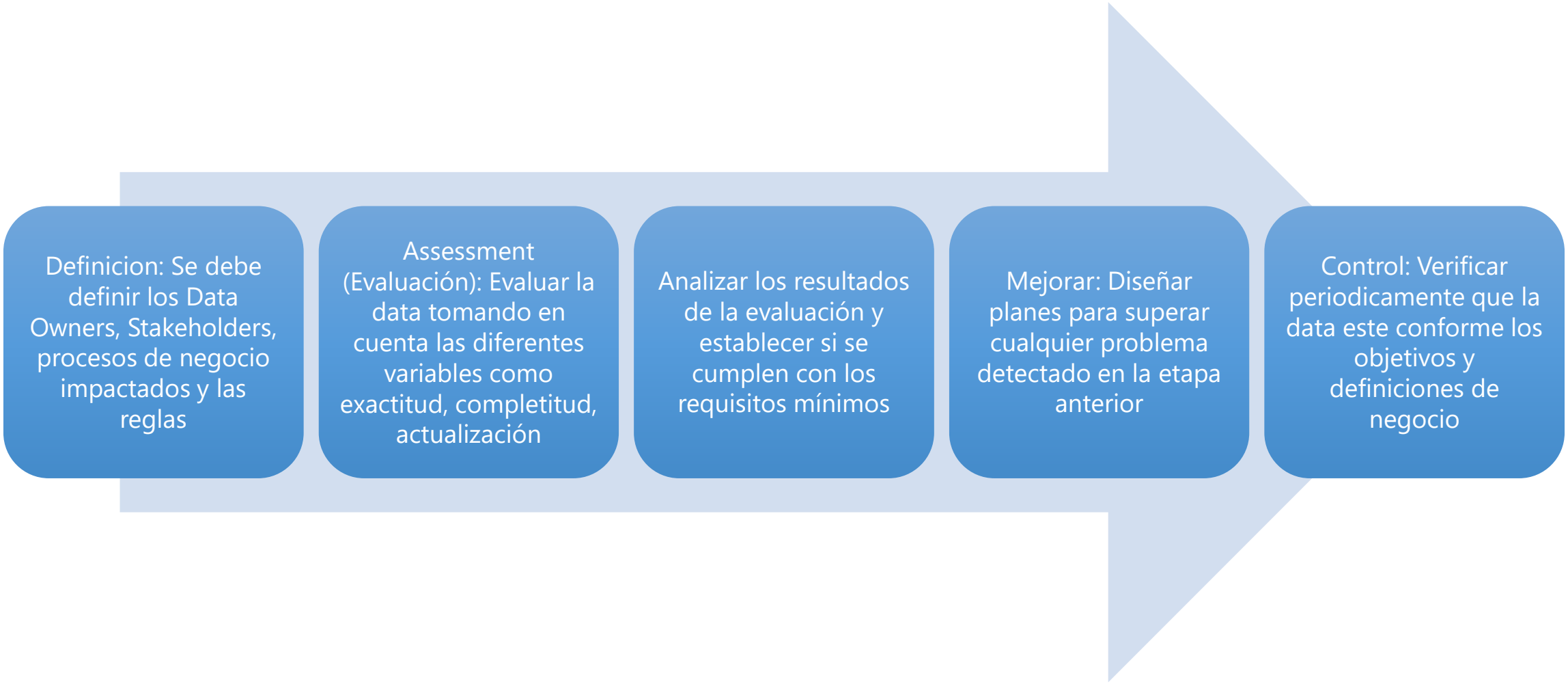
Data Quality: Definición

Se puede definir como el conjunto de técnicas/metodologías para mantener la información de las organizaciones, completa, precisa, consistente, actualizada, única y valida





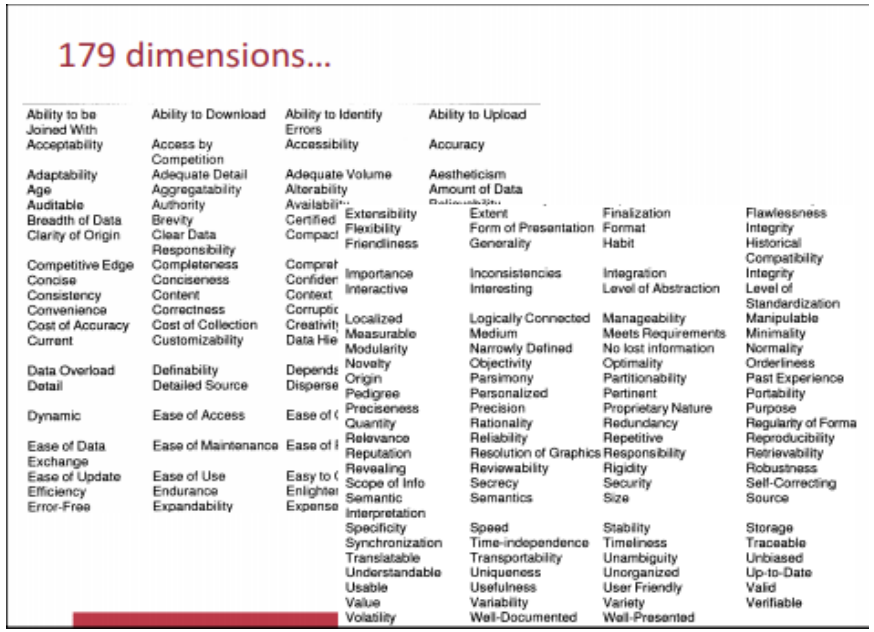
Data Quality: Proceso



Data Quality: Dimensiones

Completeness	All the required values are electronically recorded
Standards-based	Data conforms to industry standards
Consistency	Data values aligned across systems
Accuracy	Data values are right, at the right time
Time-stamped	Validity timeframe of data is clear

*Source: GCI/CapGemini Report: "Internal Data Alignment", May 2004



Tipos de Errores

Q & A

Bibliografía

Big Data Analytics: Turning Big Data into Big Money

by Frank J. Ohlhorst, November 2012

Hadoop Essentials

by Swizec Teller, April 2015

Scalable Big Data Architecture: A Practitioner's Guide to Choosing Relevant Big Data Architecture

by Bahaaldine Azarmi,2016

Regression Analysis by Example, 4th Edition

by Ali S. Hadi; Samprit Chatterjee,2006

Basic Statistics for Trainers

by Jean Houston Shore,2006

A Framework for Analysis of Data Quality Research

by Richard Y. Wang,1995

An Introduction to Data Cleaning with R

by Edwin de Jonge & Mark van der Loo,2013