



# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Septiembre, 2025



# Introducción <sub>(1)</sub>

- ¿Para qué sirve la Ciencia de Datos?
  - Para obtener información valiosa de los datos
- ¿Científico de datos?
  - Ayuda a convertir datos sin procesar en información.
  - Habilidades en: analítica, aprendizaje automático, minería de datos y estadística, experiencia en algoritmos y programación...
  - Habilidad más importante que debe poseer un científico de datos es:
    - Poseer la capacidad de explicar el significado de los datos de una manera que pueda ser fácilmente entendida por otros y a su vez que los datos cuenten historias y solucionen problemas basados en ellos.



# ¿Qué es la Ciencia de Datos? (1)

- Es un conjunto de principios, definición de problemas, algoritmos y procesos para extraer patrones no obvios y útiles en grandes volúmenes de datos. Muchos de estos principios se han desarrollado en campos relacionados con **Machine Learning** (ML) y **Minería de Datos** (MD). Estos términos a menudo se usan indistintamente...
  - La CD se centra en mejorar la **toma de decisiones** a través del **análisis de datos**, un aspecto muy destacable y central de esta disciplina. Sin embargo, aunque la CD se basa en otros campos, ésta tiene un alcance más amplio.
  - Por ejemplo, ML se centra en el diseño y evaluación de algoritmos para **extraer patrones** de datos y **clasificarlos**.
  - La MD en general se ocupa del **análisis de datos** que a menudo implica un énfasis en aplicaciones comerciales.
  - La CD toma estas consideraciones en cuenta, pero también asume otros desafíos como la **captura**, **limpieza** y **transformación** de datos no estructurados. Por ejemplo, desde las redes sociales, datos web, así como el uso de tecnologías de **big data** para almacenar grandes conjuntos de datos no estructurados.

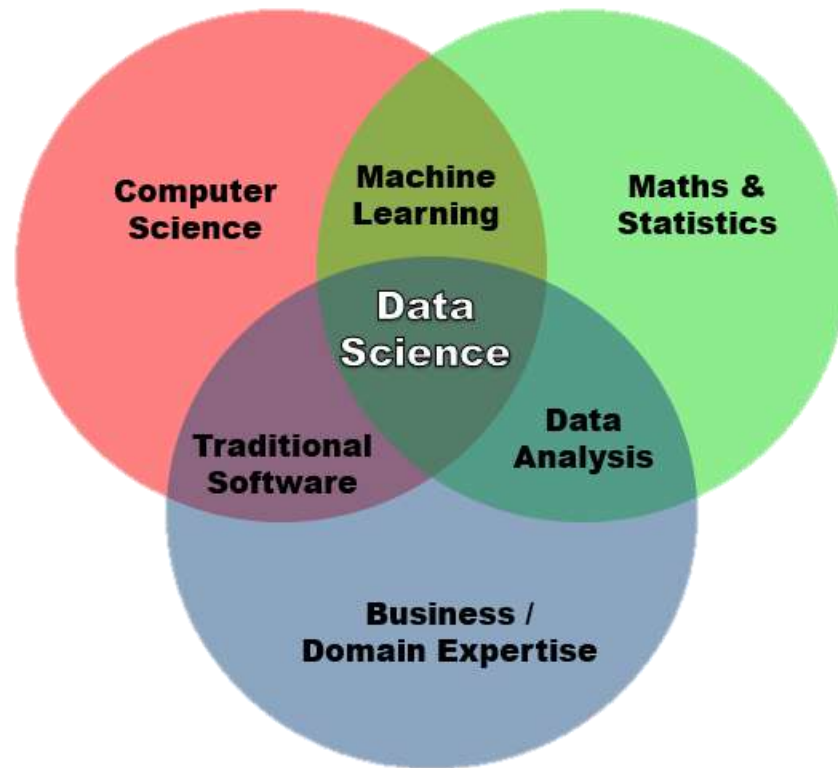


## ¿Qué es la Ciencia de Datos? (2)

- Usando la CD, podemos **extraer** diferentes tipos de **patrones**. Por ejemplo, aquellos que nos ayuden a identificar grupos de clientes siguiendo un comportamiento y gustos similares (**segmentación de clientes** en marketing) y en terminología de CD a esto se le denomina (**agrupación o clusters**).
- Alternativamente se puede extraer un patrón que identifique productos que son comprados juntos con frecuencia, y esto es un proceso de **MD**, conocido como **regla de asociación**, o tal vez requerimos extraer patrones que identifiquen eventos extraños como un reclamo de seguros, reclamos fraudulentos, y esto es un proceso conocido como **detección de anomalías y valores atípicos**.
- Finalmente es posible que necesitemos identificar patrones que nos ayuden a **clasificar** cosas... Entonces en general, la CD es una disciplina que nos permite **convertir datos sin procesar en entendimiento, comprensión y conocimiento**.



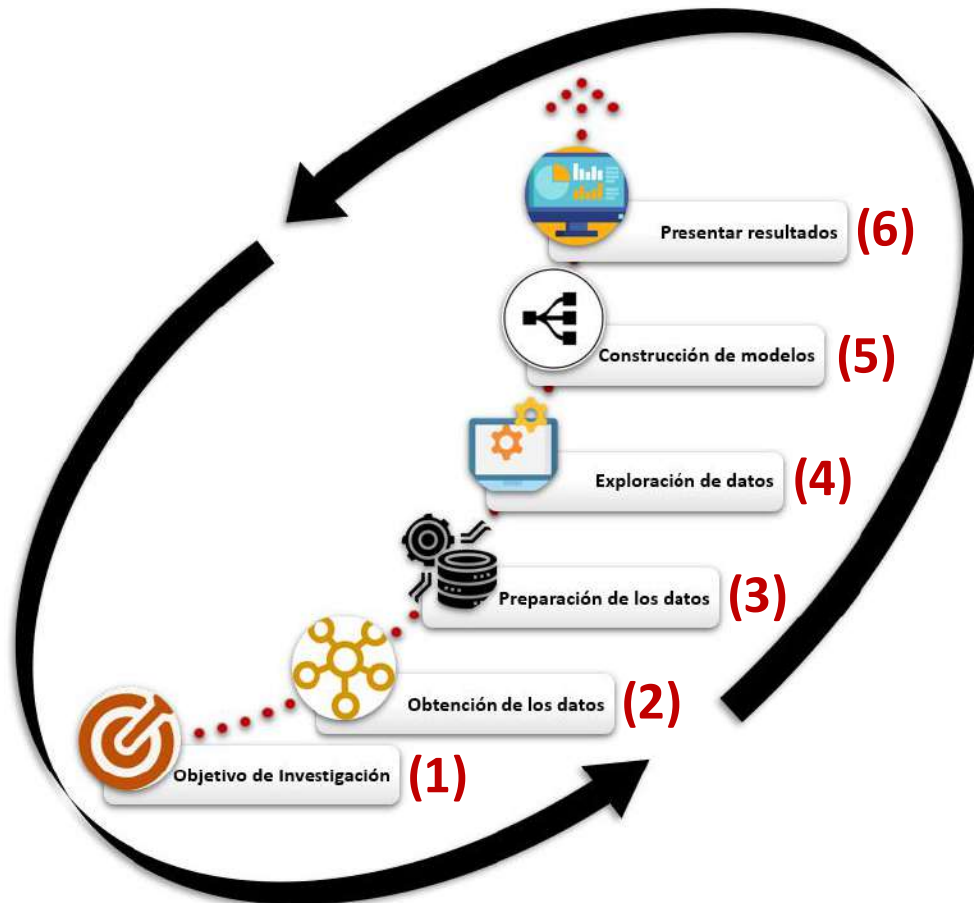
# ¿Qué es la Ciencia de Datos? (3)



- Podemos ver el diagrama de Venn, como la **unión** de varias disciplinas de varias ciencias del conocimiento, tales como:
  - Ciencia de la computación, matemáticas y estadística que entre ellas estas habilidades combinadas producen el ML.
  - Las habilidades del negocio con la ciencia de la computación generan la **programación** o **desarrollo de software**, utilizado en negocios tradicionalmente.
  - La combinación de estadística y matemáticas, con conocimientos del negocio, entonces formaliza el **análisis de datos**.
  - Por tanto, la CD es una **combinación de habilidades** en cada una de estas especialidades y ponerlas a disposición de la empresa para responder interrogantes del negocio.



# Proceso de la Ciencia de Datos <sup>(1)</sup>



- Todo proyecto de CD nace de una necesidad específica:
  - (1). Definición de hipótesis y descripción de qué queremos obtener como resultados.
  - (2). Cuáles son los posibles datos que puede arrojar esa respuesta o que se pueden procesar para obtener esas respuestas a esas interrogantes o comprobar o descartar esa hipótesis.
    - Datos de interés
      - Distintas fuentes de datos...
      - Datos de diversas disciplinas...



## Proceso de la Ciencia de Datos (2)

- (3). Hacer una **transformación**, una **estandarización**, un **modelado de datos**, en donde se necesite. Por ejemplo: estandarizar formatos de fechas, formatos de variables, combinar o calcular nuevas variables a partir de los datos, etc. Por tanto, es un proceso importante el estandarizar y luego explorar sobre ellos para saber ¿qué es lo que existe? ¿cómo están? ¿cuáles son sus patrones o posibles patrones que van a estar o qué vamos a identificar dentro de los datos?
  - Algo muy típico dentro de los datos, es que tenemos muchos valores que están perdidos o valores faltantes, errores de formato, de almacenamiento, entre otros. Estos detalles deben ser **limpiados** o **transformados** para contar con “**datos estandarizados**”, para posteriormente realizar una exploración de datos...





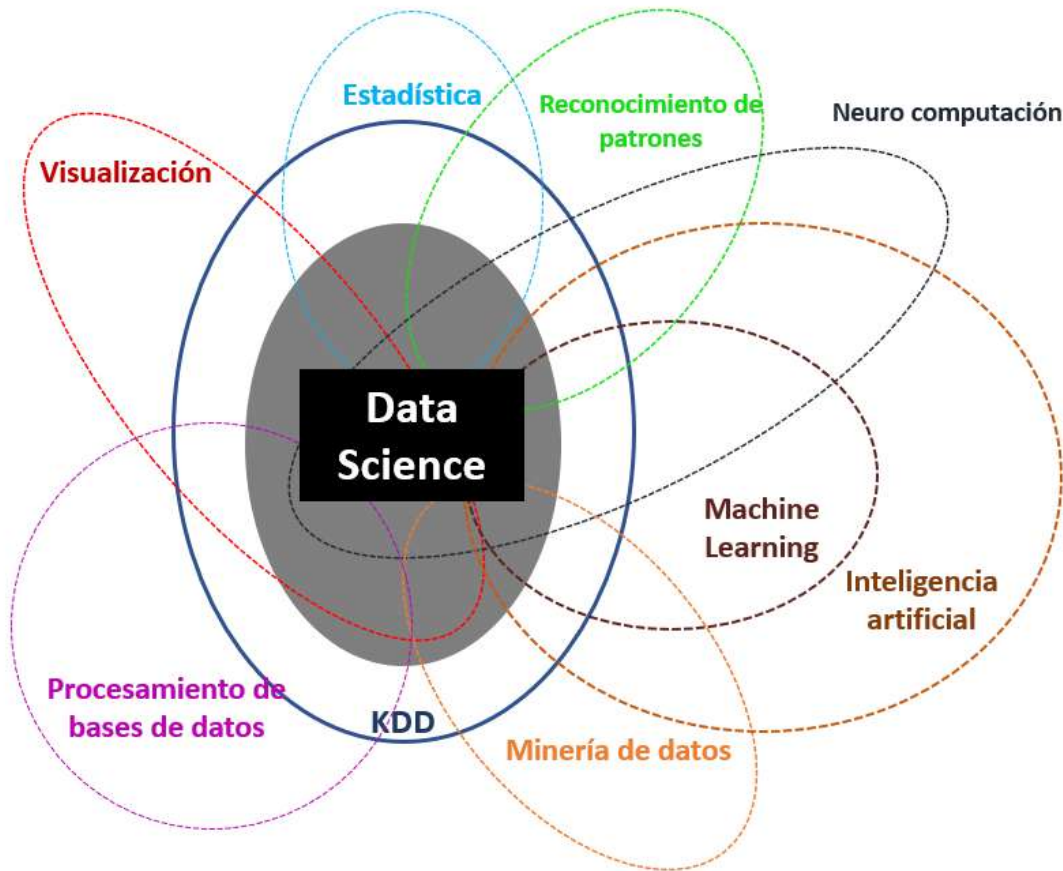
## Proceso de la Ciencia de Datos (3)

- (4). Esta tarea se da en distintas etapas como:
  - La exploración visual, realizando pruebas estadísticas dentro de esos datos, conocer la distribución, mínimos, máximos, promedios, conocer la densidad, entre otros.
- (5). Luego de ya explorados, se construye el **modelo de datos** que va a utilizarse para responder a diversas preguntas o definiciones planteadas en el objetivo, descartando aquellos datos que no se utilizarán para aplicar los algoritmos para desarrollar el modelo que va a procesar esos datos y los convertirá en información que se va a presentar...
- (6). Esta tarea consiste en resumir el producto final, conocer que **técnicas de visualización** pueden utilizarse, cómo **presentar esta información** con base en la audiencia a quién se dirige; ya sea gráficamente o textualmente.
  - En este sentido que sea **cognitivamente entendible** para las personas; es decir, elegir el modelo más adecuado que explique la información, darle valor al proceso que hubo detrás, mediante **gráficas, dashboards** que definan las conclusiones del objetivo de la investigación
    - Ejemplo: Our World in Data, Information is beautiful...





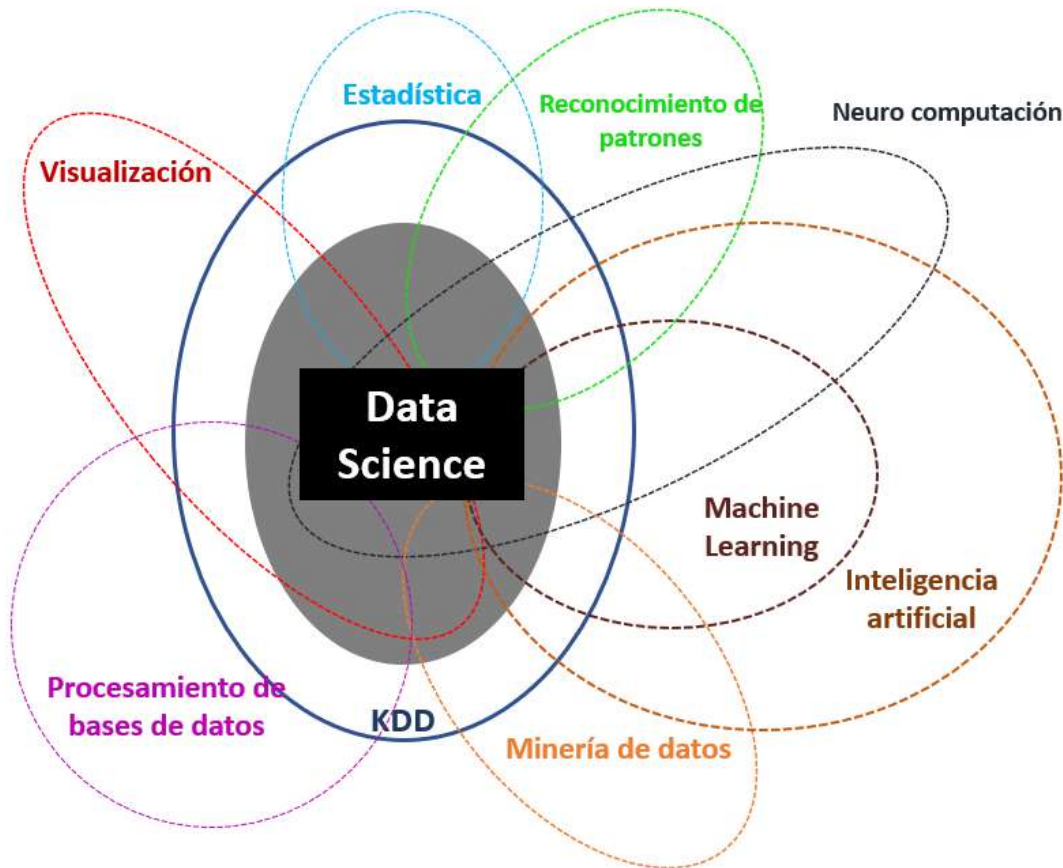
# Especialidades en la Ciencia de Datos <sup>(1)</sup>



- La CD engloba e integra muchas disciplinas para aportar en **Knowledge Discovery Database**:
  - **Estadística** para obtener un orden y análisis de un conjunto de datos y para obtener **explicaciones** y **predicciones** sobre fenómenos observados. Los métodos estadísticos permiten recolectar información para luego analizarla y extraer de ellos conclusiones relevantes. La estadística es la mejor representación de la CD y su principal objetivo es mejorar la comprensión de los hechos a partir de la información disponible, tiene como característica fundamental su **transversalidad**, por lo cual la hace aplicable a estudios de diversas disciplinas.



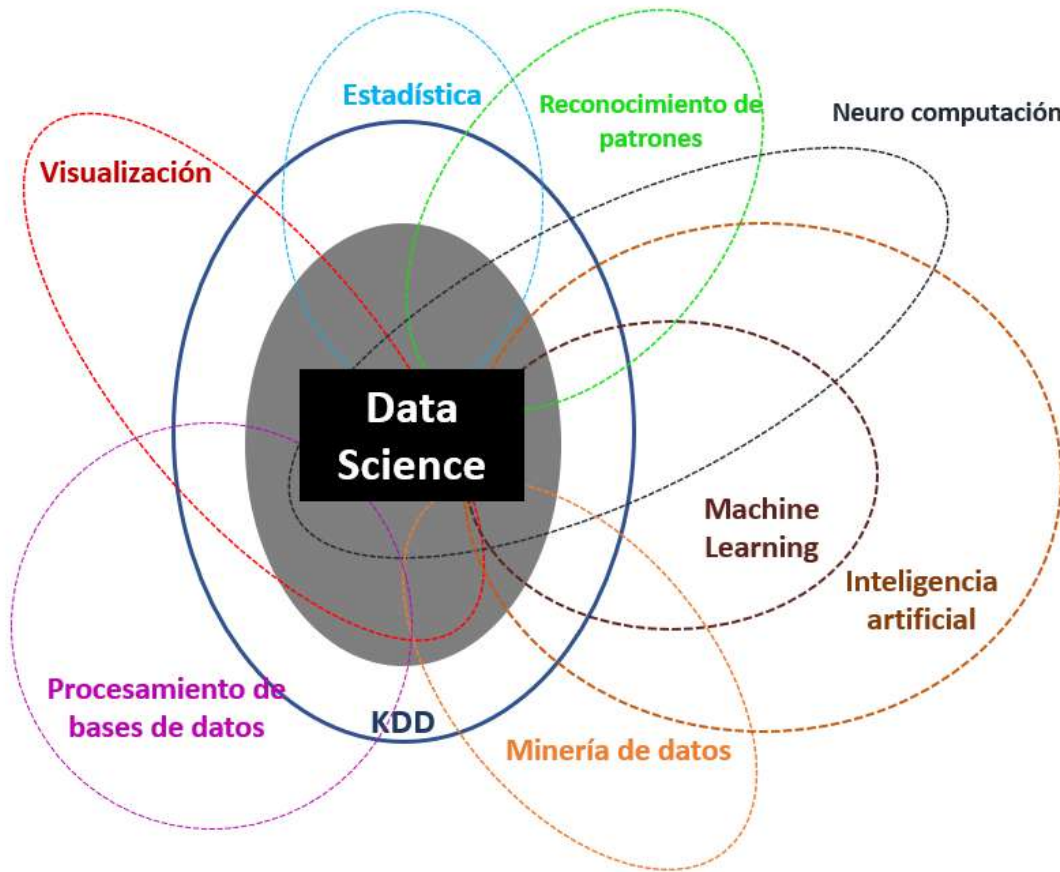
# Especialidades en la Ciencia de Datos (2)



- **Reconocimiento de patrones.** Es el procesamiento de información para solucionar diversos problemas, algunos de éstos resueltos por los humanos y otros son con máquinas, a través de métodos y algoritmos. La tarea fundamental del RP es la de **clasificar objetos** en un número específico de **categorías** o **clases**, dependiendo del dominio de aplicación. Estos objetos se denotan con el término genérico de **patrones**.
- **Neurocomputación.** Rama científica interdisciplinaria que enlaza diversos campos de la biofísica, neurociencia, ciencia cognitiva, ingeniería eléctrica ciencias de la computación y las matemáticas, y que en general persiguen **recrear en forma visual las redes neuronales y sus interacciones** en nuestro cerebro.
- **Machine Learning.** Es una disciplina dentro del ámbito de la IA, enfocada a crear sistemas que aprenden automáticamente. **“Aprender”** en este contexto significa **identificar patrones** complejos en millones de datos. La máquina que realmente **“aprende”** es un **algoritmo** que revisa los datos y es capaz de **predecir** comportamientos futuros. Automáticamente este contexto implica que estos sistemas se **mejoran** en forma **autónoma** con el tiempo, sin la **intervención humana**.



# Especialidades en la Ciencia de Datos (3)

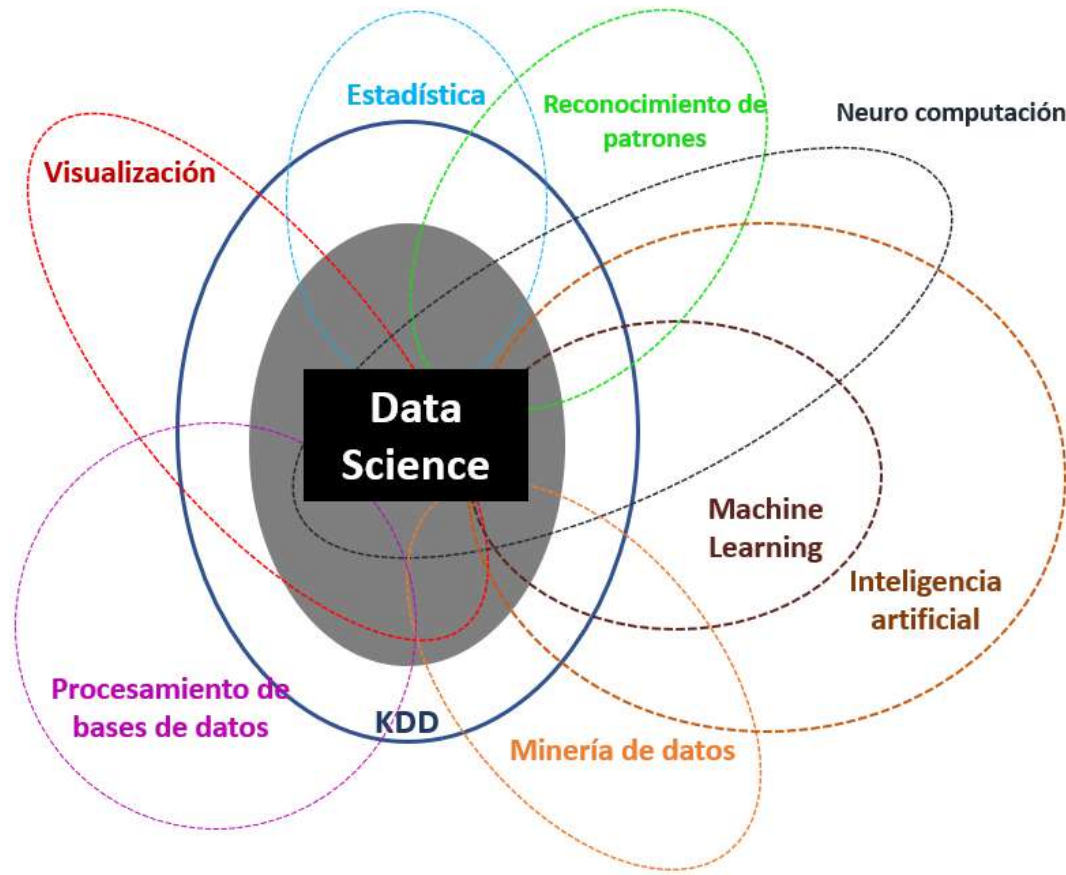


- **Inteligencia Artificial.** Es una disciplina que trata de crear sistemas capaces de **aprender** y **razonar** con un ser humano. Lo que se busca es **aprender** de la experiencia, **investigar** cómo resolver problemas ante unas condiciones dadas, **contrastar** información y **realizar** tareas lógicas.
- **Minería de Datos.** Es un enfoque que intenta **descubrir patrones** en grandes volúmenes de datos, utiliza los métodos de la IA, ML, estadística y SBD para **extraer información** de un conjunto de datos y **transformarla** en una **estructura comprensible** para su uso posterior. Dentro de la MD existe una rama en crecimiento la “minería de textos” (**text mining**), la cual refiere al proceso de **analizar** y **descubrir información nueva a partir de textos**, por medio de la identificación de patrones o correlación entre los términos, para encontrar información que no está explícita dentro del texto. Además, brinda la posibilidad de procesar textos, tomando como fuentes de datos páginas web, libros digitales, emails, reseñas, lista de artículos y productos, redes sociales, entre otros.





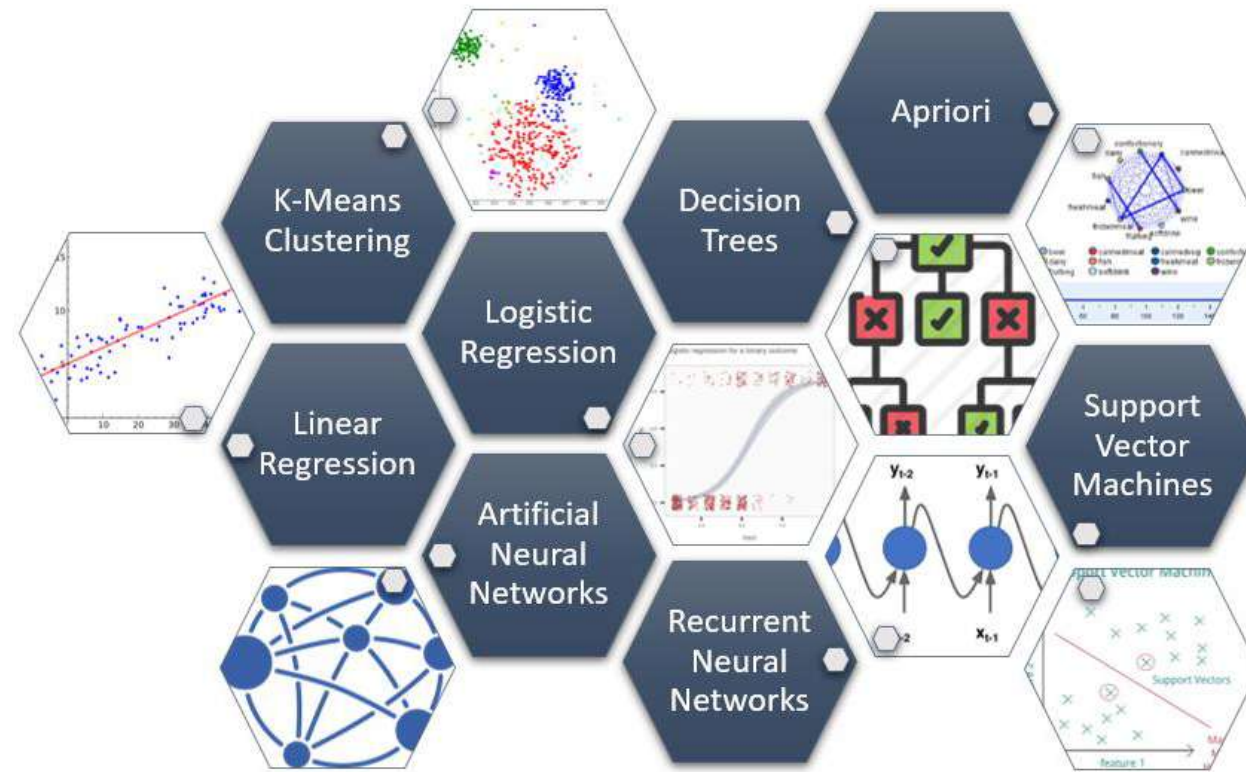
# Especialidades en la Ciencia de Datos (4)



- **Procesamiento de BD.** Trata sobre el **almacenamiento**, **manipulación** de elementos de datos para producir información significativa. Todos ellos apuntan sobre el tratamiento de datos y la CD de alguna forma maneja las **estructuras de datos** que se almacenan en una BD.
- **Visualización.** Un simple gráfico brinda más información a la mente del analista de datos que cualquier otro dispositivo. Por tanto, está comprobado científicamente que la información presentada en forma gráfica es más comprensible para el cerebro humano y que también un gráfico es la mejor representación de mucha información contenida en un gran conjunto de datos. La visualización es la **representación gráfica de información y datos** al usar elementos visuales tales como: cuadros, gráficos, mapas, etc. La visualización de datos proporciona una manera **accesible de ver y comprender tendencias, valores atípicos y patrones** en los datos. En el mundo del Big Data las herramientas y tecnologías de visualización de datos son esenciales para analizar grandes cantidades de información y tomar decisiones basadas en los datos.

Es importante resaltar que la integración y la relación entre todas estas disciplinas dentro de la CD, buscan la **generación de conocimiento**, y esa generación es para ayudar y facilitar la **toma de decisiones** en todos los ámbitos en los cuales se puede desarrollar cada una de estas disciplinas, en general la Ciencia de Datos.

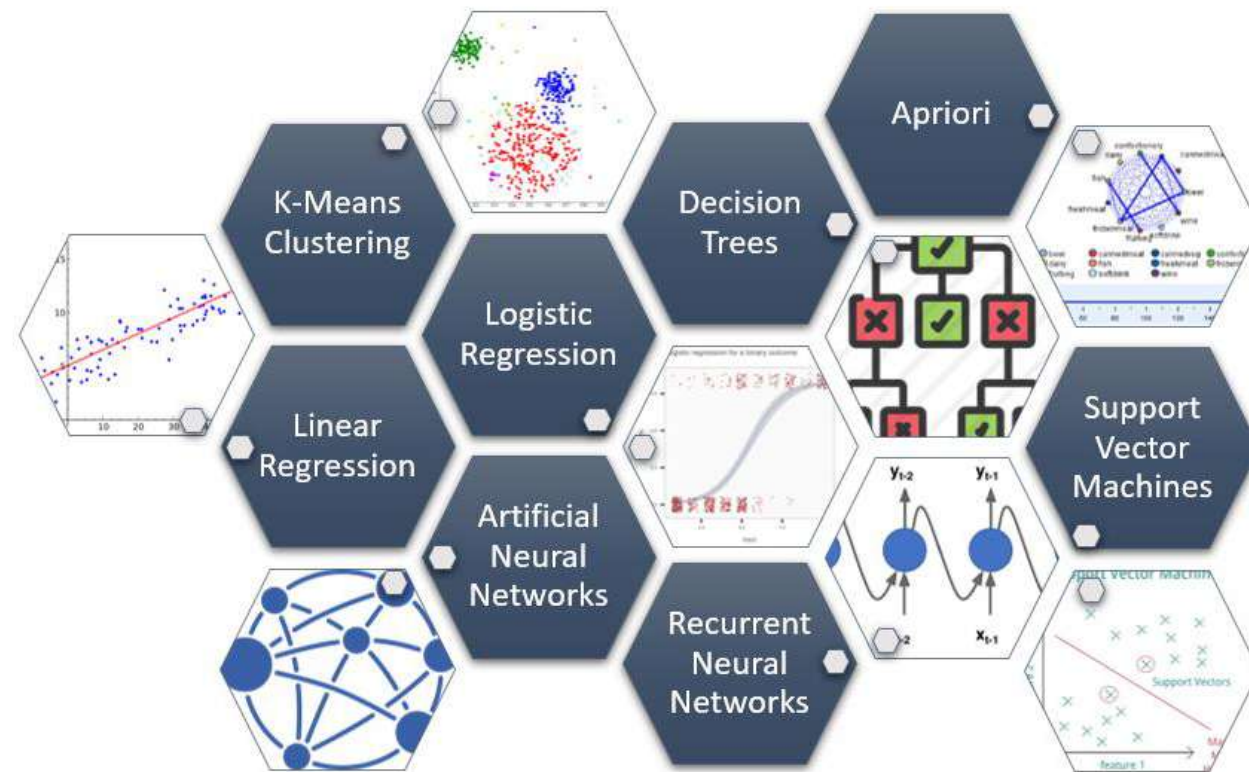
# Principales Algoritmos en la Ciencia de Datos <sub>(1)</sub>



- **Modelos de Regresión.** Son modelos matemáticos que buscan determinar la relación entre una variable dependiente ( $y$ ), con respecto a otras variables explicativas o independientes ( $x$ ). El modelo de regresión se suele usar en ciencias sociales con el fin de determinar si existe o no relación causal entre una variable dependiente ( $y$ ) y un conjunto de otras variables explicativas ( $x$ ). Asimismo, el modelo busca determinar cual será el **impacto sobre la variable dependiente ante un cambio de las variables explicativas o independientes**. Se busca tratar de **predecir** por ejemplo, en variables continuas o discretas, se puede predecir por ejemplo: el retiro no muy común de un cliente, el riesgo alto o bajo de morosidad, o tal vez obtener una cantidad de seguidores o la cantidad de clientes a futuro, haciendo un modelo de regresión lineal o regresiones logísticas.



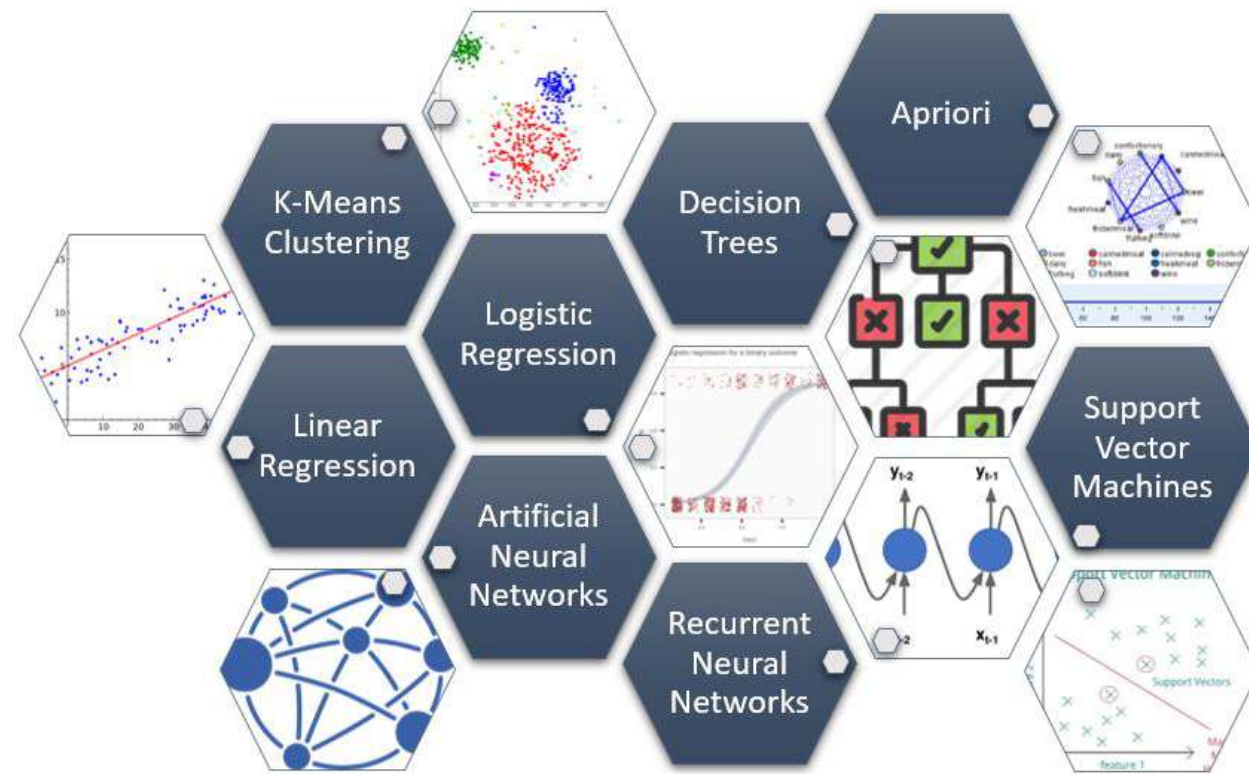
# Principales Algoritmos en la Ciencia de Datos (2)



- **Redes Neuronales Artificiales.** Es un modelo simplificado que emula el modo en que el cerebro humano procesa la información. Una RNA funciona simulando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de las neuronas humanas. Entonces, estas unidades de procesamiento se organizan en **capas**, principalmente una capa neuronal de entrada, una o varias capas ocultas y una capa de salida con la unidad o unidades que representan el campo o campos de destino. Estas unidades se conectan con fuerza de conexión variables o ponderaciones y los datos de entrada se presentan en la primera capa y los valores se propagan desde la neurona, hasta cada neurona de la siguiente capa. Al final se envía un resultado de la capa de salida, entonces una red aprende examinando los registros de manera individual, generando una **predicción** para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Entonces este proceso se repite varias veces y la red sigue mejorando su procesamiento hasta que haya alcanzado uno o varios criterios de parada. Su precisión y capacidad de procesamiento hacen que sean ampliamente usadas en problemas del mundo real, del mundo de negocios, y en ciencia de datos.



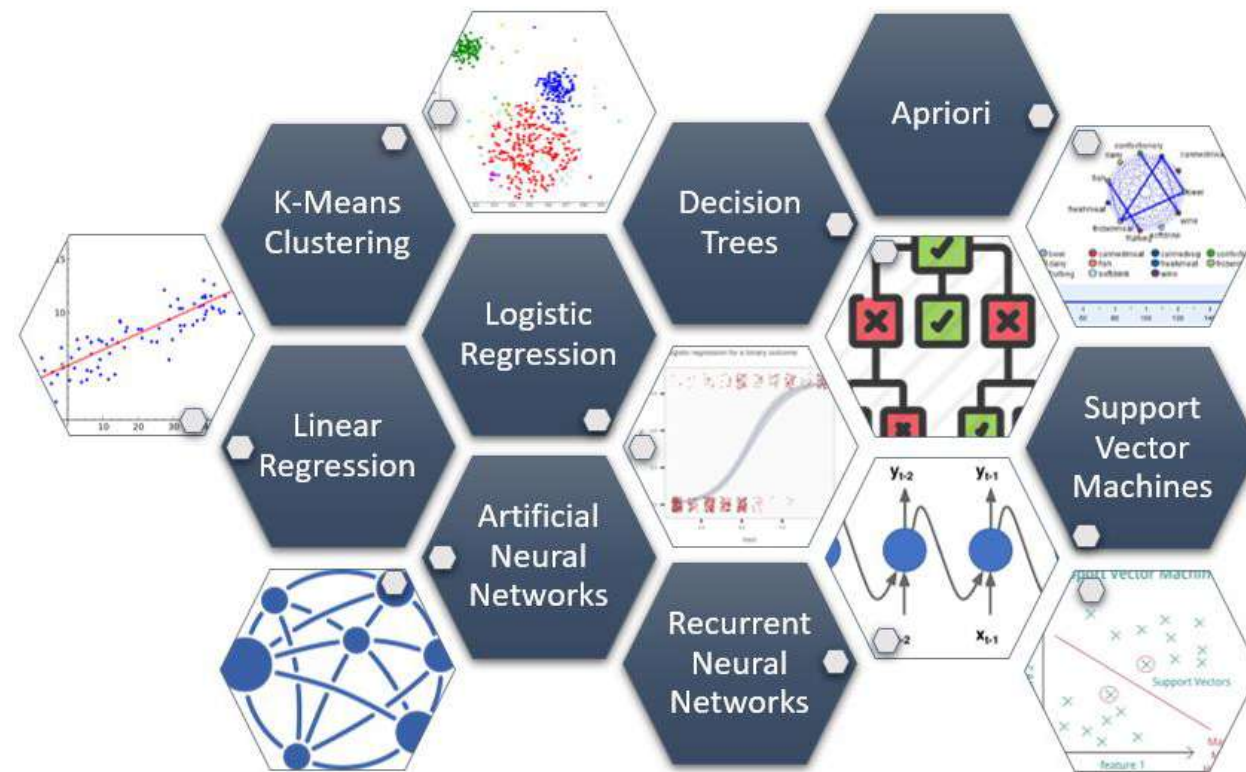
# Principales Algoritmos en la Ciencia de Datos (3)



- **Árboles de Decisión.** Son métodos analíticos que a través de una **representación esquemática** da las alternativas disponibles para facilitar y mejorar la toma de decisiones, especialmente cuando existen riesgos, costo beneficio y múltiples opciones. En este sentido, un AD va brindando los caminos a tomar según el escenario que se va presentando. Entonces, los AD son muy vistosos porque **muestran un esquema de cómo y cuál** es la mejor decisión a tomar cuando se presenta un escenario u otro.
- **Clustering.** El algoritmo de **clustering** más utilizado es **K-means** y la tarea principal que tiene un algoritmo de cluster es buscar agrupar un conjunto de objetos no etiquetados para lograr construir subconjuntos de datos conocidos como **clusters**. Entonces lo que se busca es por ejemplo, segmentar clientes con la finalidad de que se puedan ofrecer servicios personalizados o más adaptados según a la categoría de cliente que tiene una empresa, encontrar similitudes entre diversos perfiles de usuarios y en redes sociales y clasificarlos, entre otros.



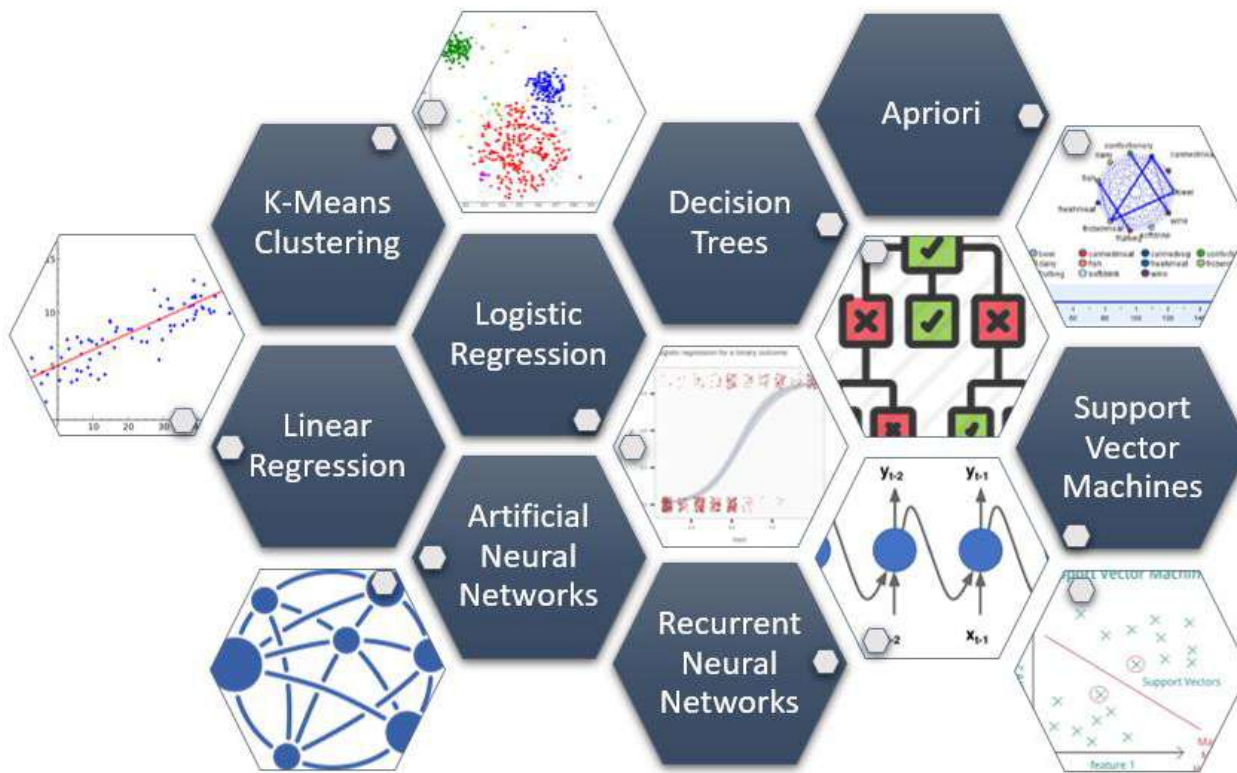
# Principales Algoritmos en la Ciencia de Datos (4)



- **Métodos apriori.** Son métodos que buscan una relación entre un algoritmo de minería de datos para buscar **relaciones** entre elementos y encontrar **reglas de asociación** de esos elementos y asocian la combinación de éstos en diferentes dominios de aplicación. Por tanto, se generan una gran cantidad de reglas que permiten ver la relación que tienen por ejemplo: las ventas en un retail, ventas de un productos, entre otros. Pueden responder a preguntas como: ¿qué probabilidad tiene que se de la venta de un producto? ¿cuántas veces se ha dado en el tiempo? Estos algoritmos son muy poderosos para **sistemas de recomendación**, que integran diversas aplicaciones orientadas a perfiles de usuario (i.e. Netflix, Amazon, Music, etc.).



# Principales Algoritmos en la Ciencia de Datos (5)



- **Máquinas de Soporte Vectorial.** Son métodos propiamente relacionados con problemas de **clasificación** y **regresión**, en donde se toma un conjunto de datos de entrada y se predice para cada una de las entradas, las dos clases de salida a las cuales pertenece, por lo que es un **clasificador no probabilístico lineal binario**, ya que solo escoge entre dos opciones. Entonces busca tener gran precisión y responder problemas más complejos que tal vez a través de modelos de regresión lineal o logística no pudieran responderse. Por tanto, estos son de los principales algoritmos que se pueden usar en la ciencia de datos, en donde el mismo algoritmo puede tener distintas aplicaciones dentro de un dominio de aplicación y ayuda no solamente a **predecir**, sino a **clasificar** entre otras muchas tareas.



# Principales Algoritmos en la Ciencia de Datos <sup>(6)</sup>

## Análisis de Regresión



**Análisis  
Descriptivo**

¿Qué ha  
pasado?



**Análisis  
Diagnóstico**

¿Por qué ha  
pasado?



**Análisis  
Predictivo**

¿Qué puede  
suceder?



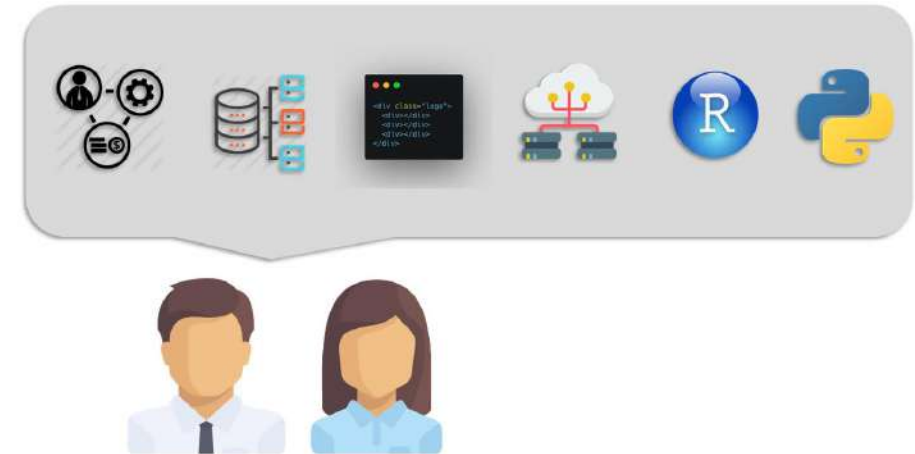
**Análisis  
Prescriptivo**

¿Qué debemos  
hacer?



# Científico de Datos (1)

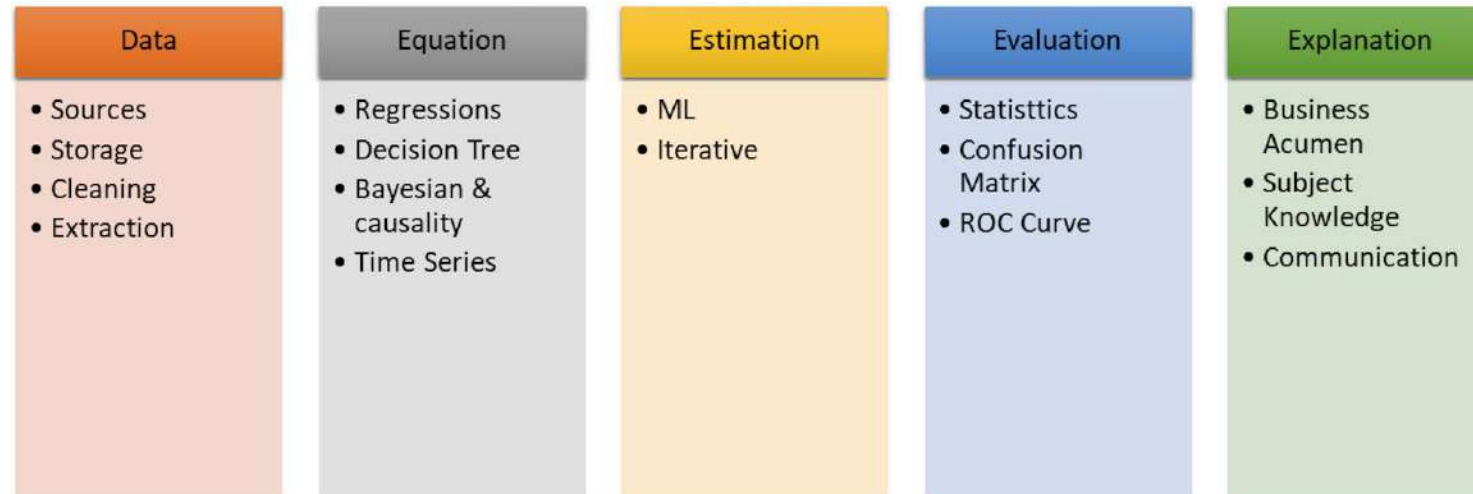
- Un **científico de datos** es un especialista en estadística que pone en práctica estos conocimientos a través de la programación de software para extraer el máximo valor de los datos, ya sea desde una fuente a múltiples fuentes de datos; en donde estos datos pueden ser:
  - Estructurados,
  - semi-estructurados o
  - no estructurados.
- Tareas principales de un científico de datos:
  - Analizar los datos del dominio de aplicación, identificar importancias y capturar aquellos que generan valor.
  - Limpieza de los datos, para darle una estructura analizable.
  - Evaluación de los modelos estadísticos para determinar la validez de los análisis.
  - Utilizar ML para construir mejores algoritmos predictivos.
  - Pruebas y mejora continua de la precisión de los modelos de ML.
  - Construir visualizaciones de datos para resumir la conclusión de un análisis avanzado.
  - Responder a requerimientos basados en las reglas del negocio con los datos.
  - Ayudar a mejorar la toma de decisiones.







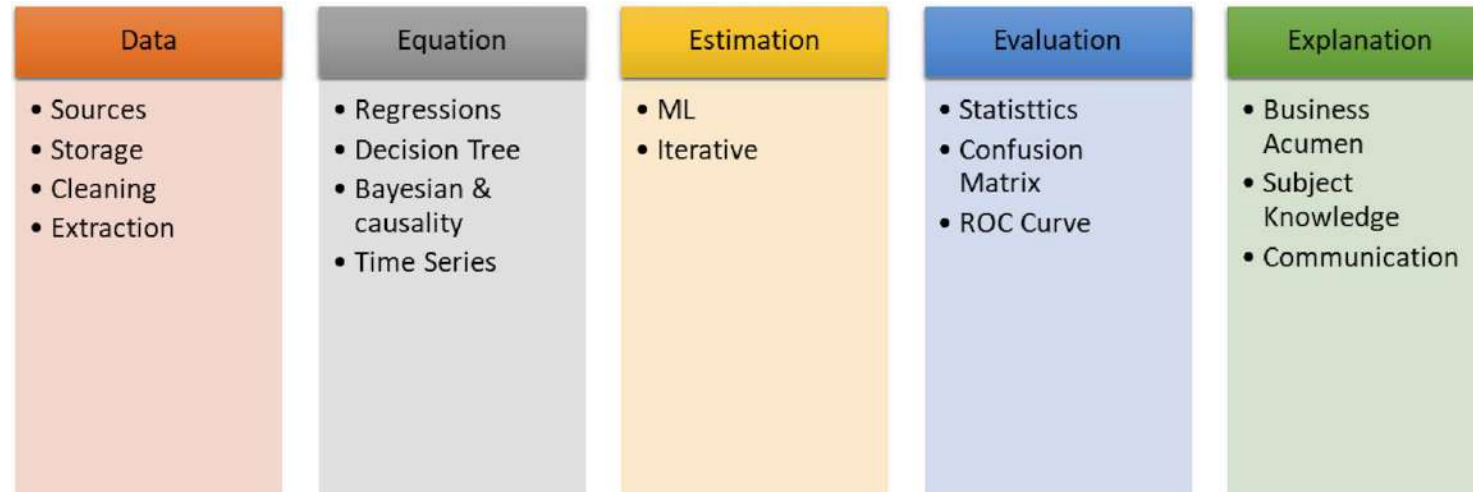
# Habilidades de un Científico de Datos (1)



- Habilidades en el manejo de:
  - **Fuentes de datos** y conocer la diversidad. Poder manejarse en distintas fuentes de datos **estructurados**, **semi** y **no estructurados**, así como en el conocimiento de los mismos datos y la **limpieza** de éstos.
  - **Transformación** de los datos para contar con una **estandarización** de los datos, una **limpieza**, **sustitución** de éstos y **extracción** de los mismos, ya sea de fuentes de datos internas o externas al dominio de aplicación, realizando estas tareas mediante APIs, conexiones web, conexiones con otros datos, etc.
  - **Ecuaciones** en los modelos de funciones específicas tales como regresión, árboles de decisión, probabilidad bayesiana, análisis de causalidad, series de tiempo, entre otros.
  - Entendimiento de los procesos de **estimación** basados en ML y procesos iterativos.



# Habilidades de un Científico de Datos (2)

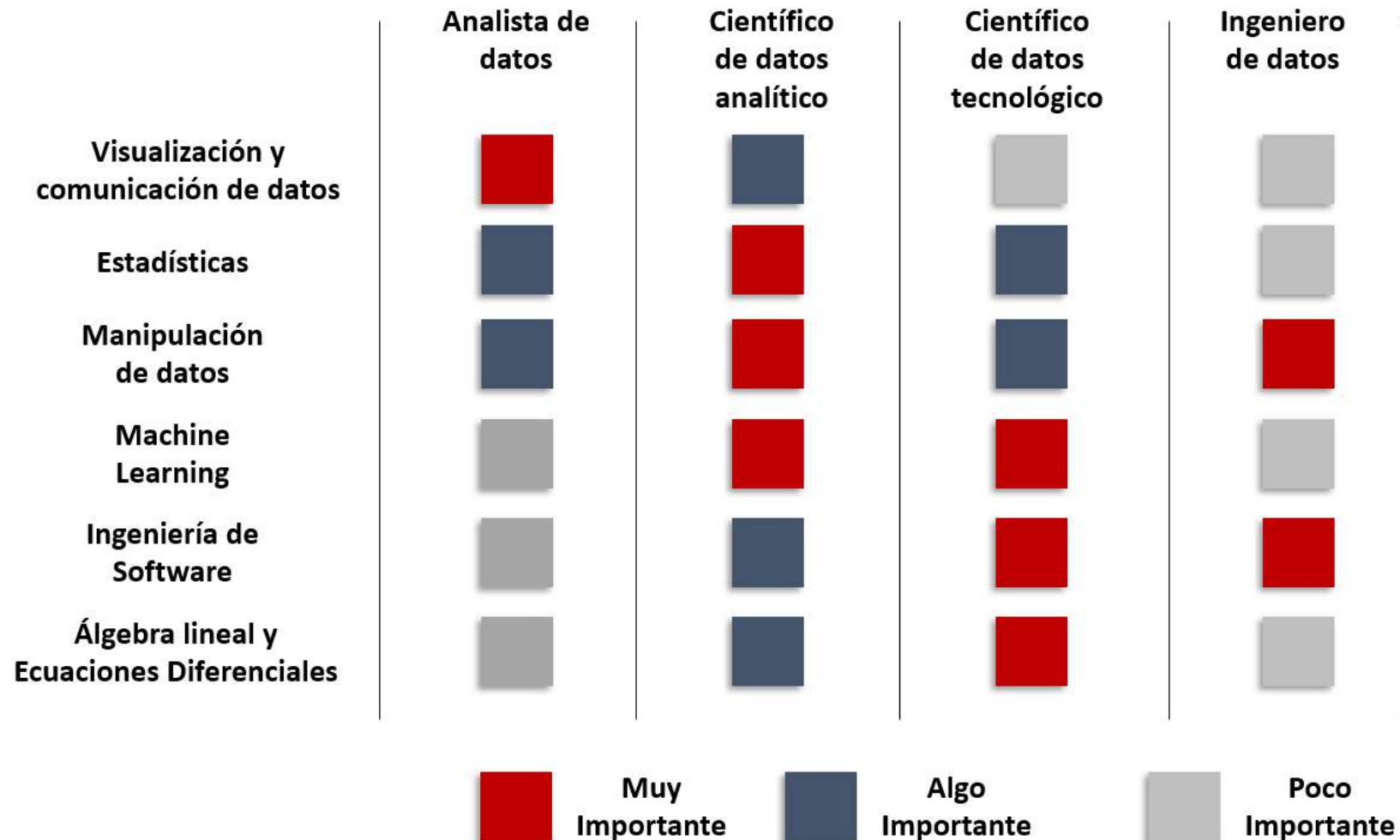


- Habilidades en el manejo de:
  - La **evaluación, procesamiento de resultados** para saber si es correcto o no la metodología aplicada. Por tanto, se requiere de una base estadística para usar herramientas como la matriz de confusión, la curva ROC para entender la precisión de los modelos que se obtienen a través del procesamiento de los datos.
  - La **explicación** de resultados ¿por qué es importante conocer el dominio, las reglas del negocio? Para conocer de lo que se habla, conocimiento técnico con relación a los cálculos que se realizan, saber comunicar, es decir, conocimiento del dominio para explicar claramente una interpretación de los resultados y cómo fueron éstos obtenidos.

En resumen, un científico de datos debe tener la habilidad y conocimiento de técnicas de modelado estadístico, conocimientos en matemáticas; particularmente álgebra lineal, desarrollo de software principalmente relacionado con lenguajes como R y Python, así como competencias en visualización de datos.



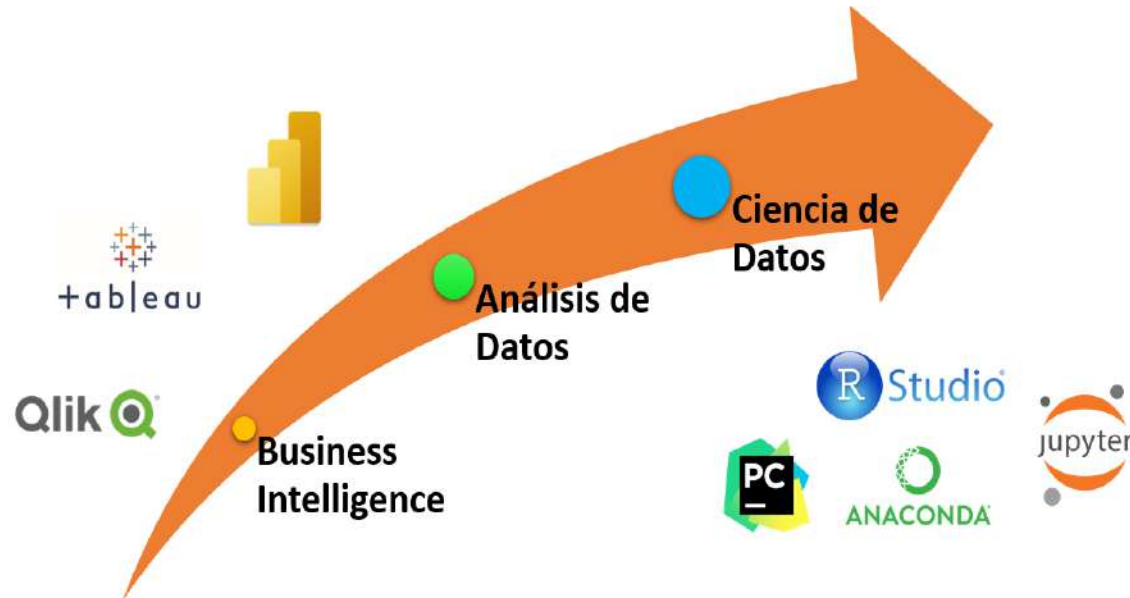
# Perfiles en un Científico de Datos (1)







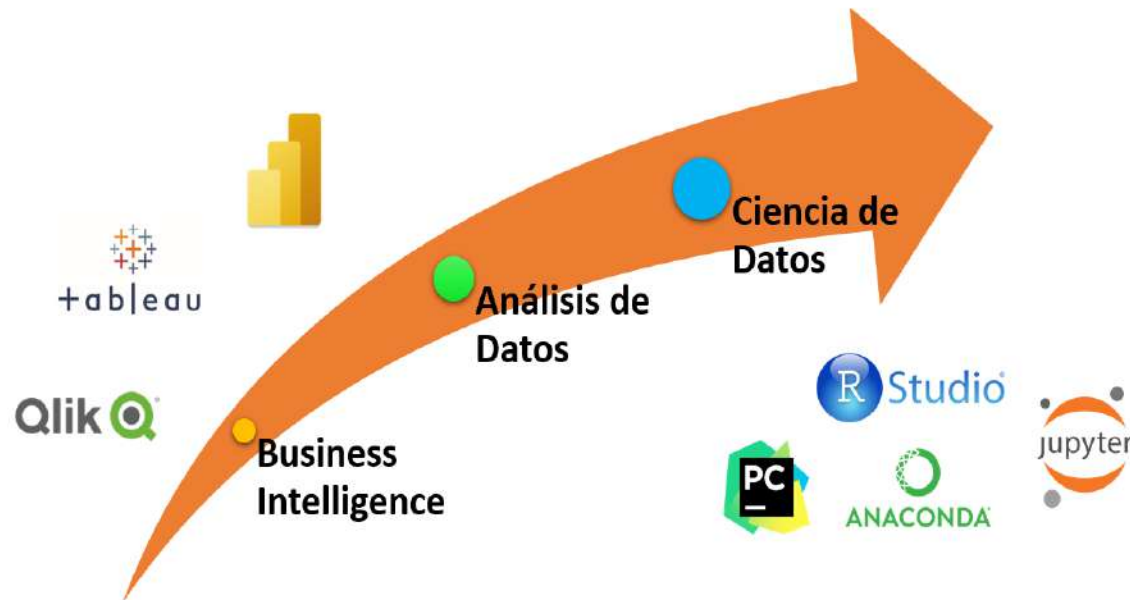
# Evolución del Análisis de Datos (1)



- La evolución del análisis de datos ha sufrido varios cambios con respecto al panorama de la analítica:
  - Muchas personas iniciaron en el manejo de datos a través de **Business Intelligence**, utilizando aplicaciones como: **Qlik**, **ClickSense**, **Tableau**, **Power BI**, iniciando con la interacción de los datos, creación de información gráfica sobre las principales áreas de un dominio de aplicación, procesos o tareas de análisis.
  - Se inició el trabajo con **indicadores** como KPIs, brindando información sobre el funcionamiento de algunas áreas o tareas específicas
  - **¿Cómo se ha hecho en el pasado para llegar al resultado actual?** Entonces, cuando ya se tiene una madurez dentro de la Inteligencia del Negocio y se ha generado una cultura de **“alfabetización”** de los datos, donde ya existen los mecanismos, búsquedas y usos de elementos para tomar decisiones, generación de gráficas, dashboards presentados en diversas plataformas, entre otros. Entonces pueden existir ya **modelos de datos generados a partir de BI**, conectados a una fuente de datos original que es un sistema transaccional del sitio, así como a fuentes de datos externas como redes sociales, sitios web, entre otros.



# Evolución del Análisis de Datos (2)



- La evolución del análisis de datos ha sufrido varios cambios con respecto al panorama de la analítica:
  - Ya cuando se tiene ese grado de **madurez de BI**, estamos enfocados en una comprensión del pasado o de una manera descriptiva de cómo se ha llegado a la situación actual. Al contar con este marco de trabajo se puede avanzar a la generación de pronósticos, la realización de segmentaciones, entre otros.
  - De esta manera se avanza formalmente al **análisis de datos**, incorporando algunos modelos de análisis, pruebas estadísticas que permitan evaluar la calidad de esos modelos, sus predicciones e incorporando dentro de las plataformas algoritmos que ayuden a realizar estos análisis.
  - Entre las principales plataformas se tienen **R** y **Python**, que generan nuevos resultados y ofrecen una nueva experiencia a usuarios finales, a través de dashboards, informes utilizando herramientas de estadística empresarial, etc.
  - Cuando se tiene implementado el marco de trabajo del análisis de datos, es necesario utilizar otras herramientas para crear ciencia de datos, con la intención que estos modelos de análisis sean reproducibles, más científicos en la **explicación** y tengan un sustento de **pruebas estadísticas** para satisfacer una hipótesis.
  - Entre las herramientas y las plataformas más comunes se tienen:
    - **R Studio** que es la interfaz para R por excelencia
    - Interfaces para **Python** como **Spyder**, **Anaconda**, **Jupyter Notebook**.
      - **Jupyter Notebook** es un entorno de desarrollo interactivo basado en la web para cuadernos, código y datos. Con una interfaz flexible que permite al usuario configurar y organizar flujos de trabajo de CD, computación científica, periodismo computacional y ML. Un diseño modular permite agregar extensiones para ampliar y enriquecer la funcionalidad, con lo cual se busca generar también estadística predictiva y empezar a realizar inferencias.



# R como lenguaje para Ciencia de Datos (1)

- R posee unas características especiales que lo hacen versátil para el manejo de elementos estadísticos, en concreto para operaciones con matrices y vectores, lo que facilita la manipulación de bases de datos. Por tanto, R permite manipular (seleccionar, recodificar y recuperar) datos rápidamente. De hecho existen algunos paquetes diseñados para ello como **plyr**, lo que hacen que este lenguaje de programación sea más hábil y eficiente en la preparación de los datos para su posterior análisis.
  - Fue un lenguaje diseñado específicamente para hacer análisis estadístico, es muy preciso y exacto para el análisis de datos.
  - Dispone de una gran cantidad de paquetes para la creación de gráficos, lo que aporta capacidades avanzadas en la visualización de los datos y los resultados del análisis. Incluye un paquete básico para funciones gráficas y se pueden agregar otros como **lattice** o **ggplot**.
  - Para Machine Learning, R tiene implementados una gran cantidad de algoritmos, como consecuencia de las diferentes líneas de investigación de grupos que dieron pie a su creación, debido precisamente al hecho de que R nació en el ámbito académico.
  - R tiene un enfoque orientado al análisis estadístico, lo que lo hace muy útil para la minería de datos. Es un multiparadigma orientado a objetos, vectorial y multiplataforma, tiene una gran cantidad de desarrolladores que lo mejoran y enriquecen.
  - R tiene una curva de aprendizaje más lenta a comparación de Phyton.





# R como lenguaje para Ciencia de Datos (2)

- Espacios colaborativos para el científico de datos
  - Ejemplo de manipulación de datos COVID en lenguaje R.
    - Se hacen las tareas de extracción de datos, de conexión de datos, transformación de esos datos y visualización de los datos a través de una gráfica. En este ejemplo no se aplicó ninguna prueba estadística para validar esos datos, ni se aplicó ninguna distribución sobre éstos. Simplemente los datos fueron conectados, transformados, limpiados y se creó un modelo simple, el cual contiene solo las variables de interés para posteriormente crear la visualización de esa información.
  - Ejemplo con datos de población mundial.





## Analítica de Datos, Minería de Datos y Descubrimiento de Conocimiento <sup>(1)</sup>

- **Minería de Datos**

- Su objetivo es **extraer** conocimiento de los datos. En este contexto el **conocimiento** se define como **patrones interesantes** que generalmente son válidos, novedosos, útiles y entendibles para el ser humano.
- El hecho de que los patrones extraídos sean interesantes o no, depende de la **aplicación en particular** y deben ser verificados por expertos. Con base en la retroalimentación, el **proceso de extracción** de conocimiento se refina de forma **interactiva** frecuentemente.

- **Analítica de Datos**

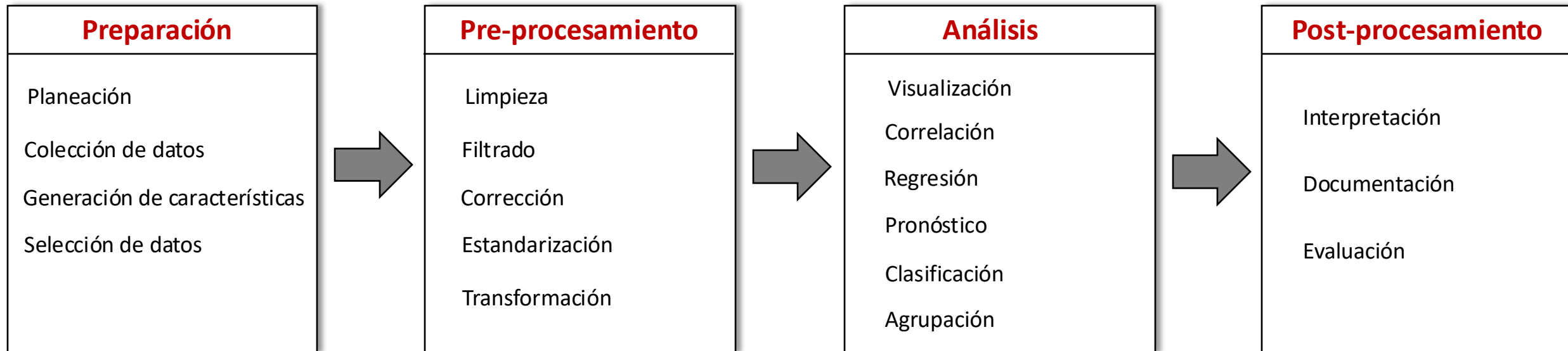
- Se define como la aplicación de sistemas de cómputo para **analizar** grandes conjuntos de datos que brinden un **soporte** a las decisiones.
- Es un **campo multidisciplinario** que ha adoptado aspectos de otras disciplinas tales como la estadística, el aprendizaje automático, el reconocimiento de patrones, la teoría de sistemas, investigación de operaciones o la Inteligencia Artificial.
- Los proyectos típicos de análisis de datos se pueden dividir en varias fases. Los datos se **evalúan** y **seleccionan**, **limpian** y **filtran**, **visualizan** y **analizan**; y los resultados del análisis finalmente se **interpretan** y **evalúan**.



## Analítica de Datos, Minería de Datos y Descubrimiento de Conocimiento (2)

### • Descubrimiento de Conocimiento en Bases de Datos

- El proceso de descubrimiento de conocimiento en bases de datos (KDD) está compuesto por 6 fases: **selección**, **preprocesamiento**, **transformación**, **extracción** de datos, **interpretación** y **evaluación**.
- Para simplificar el proceso, únicamente se consideran 4 fases: **preparación**, **pre-procesamiento**, **análisis** y **post-procesamiento**.





# Recursos

- Recursos de Python para #DataScience y #MachineLearning
  - Programación básica en Python:
    - [Listas, Tuplas, Diccionarios, Condicionales, Loops, etc.](#)
    - [Estructuras de Datos y Algoritmos](#)
    - [NumPy Arrays](#)
    - [Regular expressions \(Regex\)](#)
  - Manipulación de Datos
    - [Pandas](#)
    - [SQLAlchemy](#)
  - Visualización de Datos
    - [Matplotlib](#)
    - [Seaborn](#)
    - [Plotlib](#)
    - [Python Graph Library](#)
  - Machine Learning / Deep Learning
    - [Scikit-learn Tutorial](#)
    - [Deep Learning Tutorial](#)
    - [Kaggle Kernels](#)
    - [Otros cursos](#)





# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Septiembre, 2025



## Datos <sup>(1)</sup>

- **Conjunto de Datos de Iris**

- Para presentar los conceptos básicos del análisis de datos, consideraremos uno de los conjuntos de datos de referencia históricos más populares:
  - El **Conjunto de Datos Iris**
    - *E. Anderson. The Irises of the Gaspé Peninsula. Bull. of the American Iris Society, 59:2–5, 1935.*
  - Fue creado originalmente en 1935 por el botánico Edgar Anderson, quien examinó la **distribución geográfica de las flores de Iris** en la península de Gaspé en Quebec.
  - En 1936, Sir Ronald Aylmer Fisher usó el conjunto de datos Iris como ejemplo para el **análisis discriminante multivariable**.
    - *R. A. Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179–188, 1936.*
  - Posteriormente, el conjunto de datos de Iris se convirtió en uno de los conjuntos de datos de referencia más utilizados en **estadística** y **análisis de datos**.



## Datos <sub>(2)</sub>

- **Conjunto de Datos de Iris**

- El conjunto de datos de Iris comprende mediciones de 150 muestras de flores de Iris:
  - 50 de cada una de las tres especies Iris Setosa, Iris Virginica e Iris Versicolor.
  - Para cada una de las 150 flores, se midieron los valores de cuatro características numéricas elegidas por Anderson:
    - la longitud y la anchura de las hojas sépalo y pétalo en centímetros.
  - Se puede descargar de: <https://archive.ics.uci.edu/ml/datasets/Iris>
  - El conjunto de datos contiene **3 clases** de **50 instancias** cada una, donde cada clase se refiere a un tipo de planta de Iris.
- A continuación se presenta un fragmento del conjunto de datos Iris...



## Datos <sub>(3)</sub>

### • Conjunto de Datos de Iris

Setosa				Versicolor				Virginica			
Sepal		Petal		Sepal		Petal		Sepal		Petal	
Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

- En el análisis de datos llamamos a cada una de las 150 flores de iris como **objetos**, cada una de las 3 especies como **clases** y cada una de las 4 dimensiones como **características**.
- Algunas preguntas típicas para responder mediante el análisis de datos:
  - ¿Cuál de los datos podría contener errores o asignaciones de clases falsas?
  - ¿Cuál es el error causado por redondear los datos a un decimal?
  - ¿Cuál es la correlación entre la longitud y el ancho de los pétalos?
  - ¿Qué par de dimensiones están más correlacionadas?
  - Ninguna de las flores en el *dataset* tiene un ancho de sépalo de 1.8 cm. ¿Qué longitud de sépalo esperaríamos para una flor que tuviera 1.8 cm como ancho de sépalo?
  - ¿A qué especie pertenecería un Iris con un ancho de sépalo de 1.8 cm?
  - Las 3 especies contienen subespecies que pueden identificarse a partir de los datos?



## Tipos de datos estadísticos <sup>(1)</sup>

- **Datos categóricos**

- También conocidos como **datos cualitativos**, representan características como el género, idioma de las personas. También pueden tomar **valores numéricos**, por ejemplo: 1 – Mujeres, 0 – Hombres; sin que estos números tengan algún **significado** matemático. Los tipos de datos categóricos se clasifican en:
  - **Datos Nominales.** Se refieren a **valores** que representan **unidades discretas** y se usan para **etiquetar** variables que **no** tienen un valor **cuantitativo**. Estos datos **no** tienen un **orden**; es decir, que, aunque se cambiara el orden de sus valores, no cambia su significado.
  - **Datos Ordinales.** Estos representan **unidades discretas** y **ordenadas**. Por tanto, es casi lo mismo que los datos nominales, excepto por su orden que es importante. Las **escalas ordinales** se usan para **medir características no numéricas** como estado anímico, nivel de calidad, entre otros.

- **Datos numéricos**

- También se conocen como **datos cuantitativos** y se refieren a una **medida**. Se clasifican de la siguiente manera:



## Tipos de datos estadísticos (2)

- **Datos numéricos**

- **Datos discretos.** Son discretos cuando sus **valores** son **distintos** y **separados**; es decir, cuando los datos sólo pueden tomar ciertos valores. Este tipo de datos **no puede medirse**, pero se pueden **contar**. Básicamente representan información que se puede **clasificar**.
- **Datos continuos.** Representan **mediciones** y por lo tanto, sus **valores** no se pueden contar pero se pueden **medir**. A su vez, éstos se clasifican de la siguiente manera:
  - **Datos de intervalo.** Representan **unidades ordenadas** que tienen la misma **diferencia**. Por lo tanto, hablamos de datos de intervalo cuando tenemos una variable que contiene valores numéricos que están ordenados y donde conocemos las diferencias exactas entre los valores.
    - El problema con los datos de valores de intervalo es que podemos **sumar** y **restar**, pero no podemos **multiplicar**, **dividir** o **calcular razones**. Debido a que no existe un cero verdadero, no se pueden aplicar muchas estadísticas descriptivas e inferenciales.
  - **Datos de relación.** También son **unidades ordenadas** que tienen la misma **diferencia**. Los datos de relación son los mismos que los valores de intervalo, con la diferencia de que tienen un **cero absoluto**.



## Datos Nominales <sup>(1)</sup>

- Son datos “etiquetados” o “nombrados” que pueden dividirse en varios grupos que no se traslapan. En este caso, los datos no se miden ni se evalúan, sino que se **asignan** a varios **grupos**.
- Estos grupos son **únicos** y **no** tienen **elementos comunes**. El **orden** de los datos recopilados no puede establecerse utilizando datos nominales y, por lo tanto, si cambia el orden de los datos, su significado no se verá alterado.
- Presentan una similitud entre los diversos elementos, pero es posible que no se revelen los detalles relativos a esta similitud. Se trata simplemente de facilitar el **proceso de recolección** y **análisis de datos**.
- Como ejemplo: Si los datos binarios representan datos de “**dos valores**”, los datos nominales representan datos de “**múltiples valores**” y **no** pueden ser **cuantitativos**. Los datos nominales se consideran **discretos**.





## Datos Nominales <sup>(2)</sup>

- Los datos nominales nunca pueden ser **cuantificados**. Los datos nominales siempre estarán en forma de nomenclatura. Por ejemplo, una **encuesta para países asiáticos** puede incluir una pregunta como: *¿cuál es su origen étnico?* – con las siguientes opciones de respuesta.
  - Centroasiático
  - Indonesio
  - Asiático occidental
  - Japonés
- El análisis estadístico, lógico o numérico sobre estos datos **no es posible**, es decir, no se puede **sumar**, **restar** o **multiplicar** los datos recolectados o concluir que la variable 1 es mayor que la variable 2.



## Datos Nominales <sup>(3)</sup>

- Entre las principales características de los datos nominales se tienen:
  - **Ausencia de orden.** A diferencia de los datos ordinales, los datos nominales tampoco pueden ser **asignados** a un **orden definido**. El orden de las opciones de respuesta es irrelevante para las respuestas que proporciona, por ejemplo, un encuestado.
  - **Propiedad cualitativa.** Los datos recopilados siempre tendrán una propiedad **cualitativa**; es muy probable que las opciones de respuesta sean de esa naturaleza.
  - **No es posible calcular la media.** La media de los datos nominales no se puede establecer, incluso si los datos están ordenados alfabéticamente. En el ejemplo anterior, es imposible calcular el promedio de respuestas presentadas para las etnias **debido a la naturaleza cualitativa de las opciones**.
  - **Los datos son principalmente alfabéticos.** En la mayoría de los casos, los datos nominales son **alfabéticos** y **no numéricos**. Por ejemplo, en el caso mencionado los datos no numéricos también se pueden **categorizar** en varios grupos.



## Datos Nominales <sup>(4)</sup>

- Como ejemplos de datos nominales, se tienen los siguientes :
  - ¿Cuál es la raza de perros más querida?
    - Dálmata – 1
    - Doberman – 2
    - Labrador – 3
    - Pastor Alemán – 4
  - ¿A quién le gusta más viajar?
    - Hombres – 1
    - Mujeres – 2
  - ¿Qué tipo de casas prefieren los residentes de la ciudad de Nueva York?
    - Apartamentos – 1
    - Bungalows -2
    - Chalets – 3



## Datos Ordinales <sup>(1)</sup>

- Los datos ordinales son un tipo estadístico de **datos cuantitativos** en los que existen variables en **categorías ordenadas** que se producen de forma natural. La distancia entre dos categorías no se establece utilizando datos ordinales.
- En estadística, un grupo de números ordinales se representan usando una **escala ordinal**. La principal diferencia entre los datos nominales y ordinales es que los ordinales **tienen un orden de categorías** mientras que los nominales no.
- La **Escala Likert** es un ejemplo popular de datos ordinales.
  - Por ejemplo: para una pregunta como: *“Expresa la importancia que tiene el precio al comprar un producto”*, una escala Likert tendrá las siguientes opciones que están codificadas a 1, 2, 3, 4, 5 (números), en donde 1 es menor que 2, que es menor que 3, que es menor que 4, y que a su vez es menor que 5.

Muy importante	Importante	Neutral	Irrelevante	Muy irrelevante
1	2	3	4	5



## Datos Ordinales <sup>(2)</sup>

- Por lo tanto, los datos ordinales son un conjunto de **variables ordinales**, es decir, **variables con un orden particular** – “**bajo, medio, alto**”, pueden representarse como datos ordinales. Hay dos factores importantes a considerar para los datos ordinales.
  - Hay múltiples términos que representan el “**orden**” como “Alto, Superior, Máximo” o “Satisfecho, Insatisfecho, Extremadamente Insatisfecho”.
  - Pero la diferencia entre las variables **no es uniforme**.





## Niveles de medición <sup>(1)</sup>

- El primer paso en el análisis de datos es simplemente **entender** lo que éstos **significan**. Esto se facilita **clasificando** cada variable según su nivel de medición. El nivel de medición se refiere a la **relación entre los valores** que se asignan a los **atributos de una variable**.
- Una variable es cualquier cantidad que puede ser **medida** y cuyo **valor varía** a través de la **población o muestra**.
  - Por ejemplo, si consideramos una población de estudiantes, la nacionalidad del estudiante, género, calificaciones, etc., son todas las **variables definidas**, y su **valor** correspondiente **diferirá** para cada estudiante.
  - Si se desea calcular el salario promedio de los ciudadanos de un país, podemos salir y registrar el salario de todas y cada una de las personas para calcular el promedio o elegir un muestreo aleatorio de toda la población y calcular el salario promedio para esa muestra, y luego usar las pruebas estadísticas para obtener conclusiones para una población más amplia.
- Entonces, el tipo de prueba estadística que puede utilizarse para llegar a una conclusión sobre la población en general depende del **nivel de medición** de la variable considerada.
- El nivel de medición de una variable no es otra cosa que la **naturaleza matemática de una variable o cómo se mide una variable**.



## Niveles de medición <sup>(2)</sup>

- Tipos de niveles de medición
  - Los números se pueden agrupar en 4 tipos o niveles: **nominal**, **ordinal**, por **intervalos** y de **razón**.
  - **Nivel de medición nominal.** El nivel nominal es apenas una medida. Se refiere a la **cualidad** más que a la **cantidad**. Un nivel nominal de medición es simplemente una cuestión de diferenciar por nombre, por ejemplo, 0 = mujer, 1 = hombre. Aunque se usen los números 1 y 2, éstos no indican **cantidad**. En el mismo sentido, por ejemplo la categoría binaria de 0 y 1 utilizada para las computadoras es un nivel nominal de medición. Algunos ejemplos:
    - **PREFERENCIA DE COMIDA:** desayuno, comida, cena
    - **PREFERENCIA RELIGIOSA:** 1= Budista, 2= Musulmana, 3= Cristiana, 4= Judía, 5= Otra
    - **ORIENTACIÓN POLÍTICA:** Izquierda, Derecha, Independiente
    - Otros valores nominales son números de seguro social, códigos postales y números de teléfono.



## Niveles de medición <sup>(3)</sup>

- Tipos de niveles de medición
  - **Nivel de medición ordinal.** Este nivel se refiere al **orden en la medición**. Una escala ordinal indica la dirección, además de proporcionar información nominal. Bajo/Medio/Alto o Más Rápido/Más Lento son ejemplos de niveles ordinales de medición. Calificar una experiencia con un “9” en una escala de 1 a 10 nos indica que fue mejor que una experiencia calificada con un “6”. Muchas escalas o pruebas psicológicas utilizan la escala ordinal de medición. Algunos ejemplos:
    - **CLASIFICACIÓN:** 1er lugar, 2do lugar... último lugar
    - **NIVEL DE ACUERDO:** No, Tal vez, Sí
    - **CALIFICACIÓN CURSO:** 1, 2,..., 10



## Niveles de medición <sup>(4)</sup>

- Tipos de niveles de medición
  - **Nivel de medición de intervalo.** Proporcionan información sobre el **orden** y también poseen **intervalos iguales**. Del ejemplo anterior, si conociéramos que la distancia entre 1 y 2 es la misma que entre 7 y 8 en nuestra escala de calificación de 10 puntos, entonces tendríamos una **escala de intervalo**.
  - Un ejemplo de una escala de intervalo es la **temperatura**, medida en una escala Fahrenheit o Celsius. Un grado representa la **misma cantidad subyacente** de calor, independientemente de dónde ocurra en la escala.
  - Si lo medimos en unidades Fahrenheit, la diferencia entre una temperatura de 46 y 42 es la misma que la diferencia entre 72 y 68. Las escalas de medición de intervalos iguales pueden ser utilizadas para medir **opiniones** y **actitudes**. Algunos ejemplos:
    - **HORA DEL DÍA** en un reloj de 12 horas
      - Intervalo de tiempo de día – intervalos iguales; reloj analógico (12 horas), la diferencia entre la 1 y 2 pm es la misma que la diferencia entre las 11 y 12 am.



## Niveles de medición <sup>(5)</sup>

- Tipos de niveles de medición
  - **Nivel de medición de razón.** Además de poseer las cualidades de las escalas nominal, ordinal y de intervalo, una escala de razón tiene un **cero absoluto** (un punto donde no existe ninguna de las cualidades que se están midiendo).
  - Utilizar una escala de razón permite hacer **comparaciones** como: ser el doble de alto, o la mitad de alto de una persona. El tiempo de reacción (cuánto tiempo tarda en responder a una señal de algún tipo) utiliza una escala de medición de razón, el tiempo.
  - Aunque el tiempo de reacción de un individuo siempre es mayor que cero, **conceptualizamos** un punto cero en el tiempo y podemos afirmar que una respuesta de 24 milisegundos es dos veces más rápida que un tiempo de respuesta de 48 milisegundos. Algunos ejemplos:
    - **REGLA:** pulgadas o centímetros
    - **INGRESOS:** dinero ganado el año pasado
    - **AÑOS:** de experiencia laboral
    - **De RAZÓN:** el tiempo de 24 horas tiene un 0 absoluto (medianoche); 14 en punto está dos veces más lejos de la medianoche que las 7 en punto.





## Escala Nominal <sup>(1)</sup>

- Una escala nominal es una escala de medición en la cual los números sirven como “**etiquetas**”, solamente para **identificar** o **clasificar** un objeto. Una escala de medición nominal normalmente trata sólo con variables no numéricas (**cualitativas**).
- Por ejemplo, supongamos que se realiza esta pregunta: “¿Podrías seleccionar el grado de incomodidad de tu enfermedad?” 1-Leve; 2-Moderado; 3-Severo.
- Aquí los números simplemente son utilizados como etiquetas y no tienen ni un solo valor.
- La escala nominal posee solo la característica de descripción, y esto significa que posee **etiquetas únicas** que sirven para identificar o delegar valores de los objetos.
- Cuando la escala nominal se utiliza con fines de **identificación**, existe una correlación **uno a uno** entre un objeto y el valor asignado a él.
  - Por ejemplo, los números que están escritos en los autos de carrera simplemente están ahí para **identificar al conductor asociado con el automóvil**, la realidad es que estos números no tienen nada que ver con las características del automóvil.



## Escala Nominal <sup>(2)</sup>

- Pero cuando se utiliza la escala nominal para fines de **clasificación**, los números asignados al objeto sirven como **etiquetas** para **categorizar** y **organizar** objetos por clase.
  - Por ejemplo, en el caso de una escala de género, un individuo puede clasificarse como masculino o femenino. En este caso, todos los objetos de la categoría tienen el **mismo número**, por ejemplo, todos los hombres pueden ser número 1 y todas las mujeres pueden ser número 2. Tomar en cuenta que ese valor es puramente utilizado para fines de conteo.
  - Desde el punto de vista estadístico, la escala nominal es una de las escalas de medición más fáciles de comprender. Como se mencionó anteriormente, la escala nominal se asigna a objetos o elementos que **no son cuantitativos**, ni están orientados a un número.
    - Por ejemplo, supongamos que tenemos 5 colores, naranja, azul, rojo, negro y amarillo. Podríamos enumerar éstos en **cualquier orden** que nos guste, ya sea del 1 al 5 o del 5 al 1 en orden ascendente o descendente.
      - Aquí los números se **asignan a los colores** sólo con el propósito de **identificación**.



## Escala Nominal <sub>(3)</sub>

- Características de la escala nominal:
  - En una escala nominal, una variable se divide en **dos o más categorías**, por ejemplo, de acuerdo / en desacuerdo, si / no, etc. Es un mecanismo de medición en el que la respuesta a una pregunta en particular puede caer en cualquier categoría.
  - La escala nominal es de naturaleza **cualitativa**, lo que significa que los números se usan únicamente para **categorizar o identificar** objetos. Por ejemplo, en el fútbol, ¿has notado que los jugadores tienen un número en su camiseta? (cada uno tiene un número diferente). La realidad es que estos números no tienen nada que ver con la capacidad de los jugadores, sin embargo, pueden ayudar a identificar al jugador.
  - En una escala nominal, los números **no definen** las **características** relacionadas con el objeto, lo que significa que cada número se asigna a un objeto aleatorio o por decisión propia. El único aspecto permitido relacionado con los números en una escala nominal es que sirven para **“contar”**.
  - En una escala nominal es fácil generar respuestas utilizando **preguntas cerradas**, es por eso que se pueden recopilar muchas respuestas en un corto periodo de tiempo, lo que a su vez aumenta la confiabilidad de las respuestas.
    - Si volvemos al ejemplo de la clasificación de hombres y mujeres, 1 siendo hombres y 2 siendo mujeres, los números nos servirán para saber cuántos hombres (1) hay y cuántas mujeres (2) hay.
- Ejemplos:
  - ¿Cómo describirías tu comportamiento?
    - E – Extrovertido; I – Introverso; A – Ambas
  - Podrías seleccionar una opción que describa tu color de cabello
    - 0 – Negro; 1 – Café; 2 – Rojo; 3 – Amarillo; 4 – Otro
  - ¿Cuál es tu género? (Subtipo de escala nominal con solo dos categorías, conocido también por: **escala nominal dicotómica**)
    - H – Hombre; M – Mujer



## Escala Ordinal <sup>(1)</sup>

- La escala ordinal es uno de los niveles de medición que nos otorga la **clasificación** y el **orden** de los datos sin que realmente se establezca el grado de variación entre ellos.
- Los datos ordinales son básicamente datos estadísticos que tienen la misma naturalidad pero existe una diferencia entre ellos que es desconocida. Estos datos pueden ser **agrupados** o **clasificados**.
- Por lo tanto, se utiliza una escala ordinal como parámetro para comprender si las variables son mayores o menores. La tendencia central de la escala ordinal es la **mediana**.
- La **escala de Likert** es un ejemplo de porque la diferencia de intervalo entre las variables ordinales no se puede concluir. En esta escala de hecho, las opciones de respuesta suelen ser **polares**, como por ejemplo, algo como “totalmente satisfecho” o “totalmente insatisfecho”.
- La intensidad de la diferencia entre estas dos opciones **no puede ser relacionada a valores específicos**, ya que el valor de la diferencia entre totalmente satisfecho y totalmente insatisfecho es mucho mayor que la distancia entre satisfecho y neutral.
  - Supongamos que a una persona le encantan los automóviles Mercedes Benz, y se le aplica una encuesta que consta de una pregunta
  - ¿qué tan probable es que le recomiendes los automóviles de Mercedes Benz a tus amigos y familiares?
  - Supongamos que será muy fácil que este elija **“Extremadamente probable”** en lugar de **“probable”**. Pero qué pasa si fuera una persona **“neutral”**, a esta persona si le costaría tal vez un poco de trabajo elegir.
  - Es por eso que se utiliza una **escala ordinal cuando se debe deducir el orden de las opciones**, y no cuando se debe establecer una diferencia de intervalo.



## Escala Ordinal (2)

- Propiedades de la escala ordinal:
  - Además de **identificar** y **describir** la **magnitud**, la escala ordinal suele mostrar el **rango** relativo de variables.
  - Las propiedades del intervalo **no se conocen**.
  - Se **miden atributos no numéricos** como frecuencia, satisfacción, felicidad, etc.
  - Además de la información proporcionada por la escala nominal, la escala ordinal **identifica el rango de las variables**.
  - Utilizando esta escala, los encuestadores pueden **analizar el grado** de acuerdo o desacuerdo de los encuestados con respecto a una pregunta realizada.
  - Facilidad de comparación entre variables, ya que son extremadamente convenientes para **agrupar variables** después de que son **ordenadas**.
- Ejemplos:
  - Ranking de estudiantes de secundaria: 1ero, 3ero, 4to, 5to, etc. Un estudiante con un puntaje de 99/100 sería el primer rango, otro estudiante con puntaje de 98/100 sería el segundo, y así sucesivamente.
  - Encuestas de calificación en restaurantes: cuando se recibe una encuesta con una pregunta como: “¿Qué tan satisfecho está con la experiencia gastronómica?” En ésta las opciones de respuesta pueden ser algo como calificar del 0 al 10, siendo 10 extremadamente satisfecho y 0 extremadamente insatisfecho.
  - Escala de Likert: la escala de Likert es una variante de la escala ordinal que se utiliza para calcular niveles de satisfacción.





## Escala de Intervalo <sup>(1)</sup>

- La escala de intervalo se define como una escala de medición **cuantitativa** en la que se mide la **diferencia** entre dos variables. En otras palabras, las variables se miden en valores reales y no de forma relativa, donde la presencia de **cero es arbitraria**. Esto significa que la diferencia entre dos variables en una escala es una distancia real o igual.
  - Por ejemplo, la diferencia entre 40 grados centígrados y 50 grados centígrados es exactamente la misma que la diferencia entre 50 grados centígrados y 60 grados centígrados.
- El **“Intervalo”** equivale a la **distancia** entre dos variables. Otra manera fácil de recordar lo que es una escala de intervalo es considerando que ésta es la resta que se define entre dos variables. Esto es diferente a la escala de razón, donde la división se define entre dos variables.
- Los datos de la escala pueden ser:
  - Tipo **discretos**. Ejemplo: números tipo 8 grados, 4 años, 2 meses, etc.
  - Tipo **continuos**. Ejemplo: con números fraccionarios como 12.2 grados, 3.5 semanas o 4.2 kilómetros.



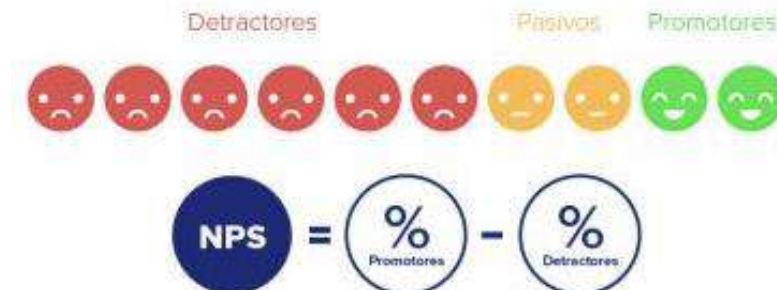
## Escala de Intervalo <sub>(2)</sub>

- Las características de esta escala son las siguientes:
  - La escala de intervalo es preferible a la escala nominal o la escala ordinal porque las dos últimas son **escalas cualitativas**. La escala es **cuantitativa** en el sentido de que se pueden cuantificar la diferencia entre dos valores.
  - Se pueden restar valores entre dos variables y esto te ayuda a **comprender** la **diferencia** entre dos variables.
  - Esta escala permite calcular la **media** de las variables.
  - Esta es una escala preferida en estadística porque permite **asignar** un **valor numérico** a cualquier evaluación arbitraria.



## Escala de Intervalo <sup>(3)</sup>

- Ejemplos:
- La escala de intervalo es el tipo de pregunta que se utiliza con mayor frecuencia en un estudio o investigación.
  - Para obtener cualquier tipo de respuesta, es indispensable que la pregunta solicitada requiera que los encuestados respondan en una **escala numérica** donde la **diferencia** entre los dos números sea la misma.
  - En la escala de intervalos, se requiere que la encuesta se diseñe de tal forma que la **dimensión** que se mida también se escale adecuadamente. Esto se puede lograr de manera numérica o verbal. Los tipos de preguntas para escalas de intervalo:
    - **Net Promoter Score (NPS).** En esta pregunta de intervalo, la pregunta se realiza usando una escala del 1 al 10 para responder. La pregunta NPS se basa en saber qué tan probable es que un cliente le recomiende tu negocio, producto o servicio a sus amigos, colegas y familiares.







## Escala de Razón <sup>(1)</sup>

- Los datos de escala de razón se definen como un tipo de datos **cuantitativos** que se caracterizan por un punto de **cero absoluto**, lo que significa que no hay ningún valor **numérico negativo**. Por ejemplo:
  - Cuatro personas son seleccionadas al azar y se les pregunta ¿cuánto dinero traen? Estos son los resultados: \$21, \$50, \$65 y \$300.
    - ¿Existe un orden para estos datos? **Sí**,  $\$21 < \$50 < \$65 < \$300$ .
    - ¿Las diferencias entre los valores de datos son significativas? **Sí**, la persona que tiene \$50 tiene \$29 más que la persona con \$21.
    - ¿Podemos calcular razones en función a estos datos? **Sí**, porque \$0 es la cantidad mínima absoluta de dinero que una persona podría traer con ella.
    - La persona con \$300 tienen 6 veces más que la persona con \$50.
  - Los datos de escala de razón tienen todas las **propiedades** de los datos de la escala de intervalo. Por ejemplo, los datos deben tener **valores numéricos**, la **distancia** entre los dos puntos es igual, etc. Sin embargo, a diferencia de los datos de intervalo donde el cero es arbitrario, en los datos de una escala de razón el **cero es absoluto**.





## Escala de Razón <sup>(2)</sup>

- Ejemplos de datos de escala de razón
  - La medición de **alturas**.
    - La **altura** puede medirse en centímetros, metros, pulgadas o pies. No es posible tener una altura negativa.
  - Si los comparamos con los datos de una escala de intervalo, por ejemplo, la **temperatura** puede ser de -10 grados, sin embargo, la **altura** no puede ser negativa como se mencionó anteriormente.
  - Los datos de escala de razón pueden ser **multiplicados** y **divididos**, ésta es una de las principales diferencias entre los datos de escala de razón y los datos de una escala de intervalo, los cuales solo pueden ser **sumados** y **restados**.
  - En los datos de escala de razón, la diferencia entre 1 y 2 es la misma que la diferencia entre 3 y 4, pero también aquí 4 es el doble que 2. Esta **comparación** es **imposible** en los datos de escala de intervalo.
- Análisis de datos de escala de razón
  - Los datos de escala de razón, junto con los otros 3 niveles de medición, son fundamentalmente un método de captura de **datos cuantitativos**. Lo que significa que se pueden aplicar todos los tipos de técnicas de análisis estadístico a los datos de razón.



## Escala de Razón <sup>(3)</sup>

- Características de la escala de razón:
  - **Punto de cero absoluto.** Una de las características distintivas de los datos de análisis de razón es el verdadero punto de **cero absoluto**, el cual hace que los datos sean **relevantes** y **significativos** de una manera que es correcto decir “un objeto es dos veces más largo que el otro” o 4 tiene el doble del valor que 2.
  - **Sin valor numérico negativo.** Los datos de escala de razón no tienen ningún valor **numérico negativo**. Por ejemplo, el peso no puede ser negativo, -20 Kgs no existe.
  - **Cálculo.** Los valores de datos de una escala de razón se pueden **sumar**, **restar**, **multiplicar** y **dividir**. Se puede realizar un análisis estadístico único para los datos de razón.
- Ejemplos:
  - ¿Cuál es tu peso en kilogramos?
    - Menos de 50 kgs
    - 51-60 kgs
    - 61-70 kgs
    - 71-80 kgs
    - 81-90 kgs
    - Más de 90 kgs



# Propiedades fundamentales de los datos <sup>(1)</sup>

## • Escalas de datos

- Las medidas numéricas pueden tener diferentes **significados semánticos**, incluso si están representadas por los mismos datos numéricos.
  - Dependiendo del significado semántico, diferentes tipos de operaciones matemáticas son apropiadas. Para el significado semántico de la medición numérica, se tienen cuatro escalas.

Escala	Operaciones		Estadística
Razón	.	/	Media generalizada
Intervalo	+	-	Media
Ordinal	>	<	Mediana
Nominal	=	≠	Moda



## Propiedades fundamentales de los datos (2)

### • Escalas de datos

- Para **datos nominales escalados**, solo son válidas las pruebas de igualdad o desigualdad.
  - Ejemplos de características nominales son **nombres de personas** o **códigos de objetos**. Los datos de una característica nominal se pueden representar por la **moda**.
    - La moda se define como el valor que se presenta con **mayor frecuencia**.
- Para **datos ordinales escalados** son válidas las operaciones “**mayor que**” y “**menor que**”.
- Para cada nivel de la escala también son válidas las operaciones y estadística de los niveles inferiores de la escala, por lo que para la **escala ordinal** tenemos la **igualdad**, la **desigualdad** y las **combinaciones** “**mayor o igual**” ( $\geq$ ) y “**menor o igual**” ( $\leq$ ).



## Propiedades fundamentales de los datos <sup>(3)</sup>

### • Escalas de datos

- La relación “**menor o igual**” ( $\leq$ ) define un **orden total**, tal que para cualquier  $x, y, z$  se tiene:
  - $(x \leq y) \wedge (y \leq x) \Rightarrow (x = y)$  (**antisimetría**),
  - $(x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z)$  (**transitividad**),
  - $(x \leq y) \vee (y \leq x)$  (**totalidad**).
- Ejemplos de **características ordinales** son las **calificaciones escolares**.
- Los datos de una **característica ordinal** se pueden representar mediante la **mediana**.
  - Que se define como el valor para el cual existen (casi) tantos valores más pequeños como más grandes.
  - La **mediana** no es válida para características ordinales, por lo que, por ejemplo, no tiene sentido decir que la calificación escolar promedio es C.
  - Para datos escalados por intervalos, la **suma** y la **resta** son válidas. Las entidades escaladas por intervalos tienen puntos **cero arbitrarios**.
    - Ejemplos: temperaturas en grados Celsius o Fahrenheit, por lo que no tiene sentido decir que 40 °C es el doble que 20 °C.



## Propiedades fundamentales de los datos (4)

### • Escalas de datos

- Los datos de una **característica de intervalo**, por ejemplo, dado un conjunto de valores  $X = \{x_1, \dots, x_n\}$ , se puede representar mediante la **media (aritmética)**.

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$$

- Para datos en **escala de proporción**, la **multiplicación** y la **división** son válidas.
  - Ejemplos de características en escala de proporción son las diferencias de tiempo como edades o temperaturas en la escala Kelvin. Los datos de una característica a escala de intervalo se pueden representar mediante la **media generalizada**.

$$m_{\alpha}(X) = \sqrt[\alpha]{\frac{1}{n} \sum_{k=1}^n x_k^{\alpha}}$$

- Con el parámetro  $\alpha \in \mathbb{R} \setminus \{0\}$ , que incluye los casos especiales mínimo ( $\alpha \rightarrow -\infty$ ), media armónica ( $\alpha = -1$ ), media geométrica ( $\alpha \rightarrow 0$ ), media aritmética ( $\alpha = 1$ ), media cuadrática ( $\alpha = 2$ ), y máxima ( $\alpha \rightarrow \infty$ ).





## Propiedades fundamentales de los datos (5)

### • Mediana

- Se define como el **número** o **valor central** que está justo en el medio de un conjunto de datos, ordenado de menor a mayor o de mayor a menor.
  - Cuando se tiene un conjunto de datos par, no existe un valor central, entonces es necesario calcular la **media aritmética** de los valores centrales del conjunto.

$$X = \{1, 2, 3, 4, 5\} \rightarrow M = 3$$

$$X = \{1, 2, 4, 5\}; \mu_A = \frac{2+4}{2} = 3 \rightarrow M = 3$$

### • Moda

- Se define como el número que está representado más veces dentro de un conjunto de datos; es decir, aquel valor que presenta una mayor **frecuencia** absoluta dentro de la muestra.
  - La moda se calcula tanto para variables **cuantitativas** como **cualitativas**.
  - **Moda Unimodal**: Cuando el máximo número de repeticiones se da para un solo número.

*Datos: [3, 5, 5, 6, 8];  $M_u = 5$ ;  $\rightarrow$  El valor de 5 se repite dos veces*



## Propiedades fundamentales de los datos (6)

### • Moda

- **Moda Bimodal:** Cuando el máximo número de repeticiones se da para dos números.

*Datos: [3, 5, 5, 6, 8, 8];  $M_B = 5 \text{ \& } 8$ ;  $\rightarrow$  Ambos se repiten dos veces*

- **Moda Multimodal:** Cuando el máximo número de repeticiones se da para tres o más valores.

*Datos: [3, 3, 5, 5, 6, 8, 8];  $M_B = 3, 5 \text{ \& } 8$ ;  $\rightarrow$  Tres valores con una repetición de dos veces*

### • Varianza

- Es una medida de **dispersión** que representa la **variabilidad** de una serie de datos respecto a su media.
- La **varianza**, junto con la **desviación estándar**, son medidas de **dispersión** de datos u observaciones.
  - La dispersión de estos datos indica la **variedad** que estos presentan, es decir, si todos los valores en un conjunto de datos son iguales, entonces no hay dispersión, pero en cambio, si todos son iguales entonces hay dispersión.



## Propiedades fundamentales de los datos (7)

- **Varianza**

- En resumen, esta **dispersión** puede ser grande o pequeña, dependiendo de qué tan cercanos sean los valores a la media. Procedimiento de obtención:
  - Calcular la media de conjunto muestra
  - Restar la media a cada número anterior y elevarlo al cuadrado
  - Calcular la media de las diferencias al cuadrado obtenidas en el punto anterior

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$$

- Donde:
  - $x_i$  es el número de observaciones de la variable o conjunto, tomando valores de 1 a  $n$ .
  - $n$  es el número de observaciones.
  - $\bar{X}$  es la media del conjunto o variable.



## Propiedades fundamentales de los datos (8)

- **Desviación estándar**

- Es una medida que **cuantifica** la cantidad de dispersión de las observaciones en un conjunto de datos. La baja desviación estándar es un indicador de la **cercanía** de las puntuaciones a la media aritmética y representa una desviación estándar alta. Las puntuaciones se dispersan en un rango de valores más alto.

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2}$$



## Propiedades fundamentales de los datos (9)

### • Diferencias entre la varianza y la desviación estándar

- La varianza es un valor numérico que describe la **variabilidad** de las observaciones desde su **media aritmética**. La desviación estándar es una medida de la **dispersión** de observaciones dentro de un conjunto de datos.
- La varianza no es más que un **promedio** de desviaciones al cuadrado. Por otro lado, la desviación estándar es la **desviación cuadrática media**.
- La varianza se denota por sigma-cuadrado ( $\sigma^2$ ) mientras que la desviación estándar se etiqueta como sigma ( $\sigma$ ).
- La **variación** se expresa en **unidades cuadradas** que generalmente son más grandes que los valores en el conjunto de datos dado. A diferencia de la **desviación estándar**, que se expresa en las **mismas unidades** que los valores en el conjunto de datos.
- La **varianza** mide **qué tan lejos están** los individuos en un grupo. Por el contrario, la **desviación estándar** mide la **cantidad de observaciones** de un conjunto de datos que **difiere** de su **media**.



## Propiedades fundamentales de los datos <sup>(10)</sup>

- **Escalas de datos**

- Las características del conjunto de datos de Iris están en escala de razón. Por tanto, podemos:
  - Estimar aproximadamente el área de la superficie del sépalo multiplicando la longitud del sépalo y el ancho del sépalo.
  - Calcular la moda, la mediana, la media, la varianza y la desviación estándar de cada una de las características del conjunto de datos para las 3 clases de flores.





## Proyecto 2. Hacer un programa en R o Python que permita llevar a cabo los siguientes cálculos sobre el conjunto de datos Iris

- Descargar del repositorio o URL indicada en la presentación, el conjunto de datos para llevar a cabo el análisis exploratorio que responda a lo siguiente:
- Calcular el área del sépalo y pétalo para cada especie de Iris.
- ¿Cuál de los datos podría contener errores o asignaciones de clases falsas?
- ¿Cuál es el error causado por redondear los datos a un decimal?
- ¿Cuál es la correlación entre la longitud y el ancho de los pétalos en cada flor?
- ¿Qué par de dimensiones están más correlacionadas?
- Ninguna de las flores en el *dataset* tiene un ancho de sépalo de 1.8 cm. Entonces, ¿qué longitud de sépalo se espera para una flor que tuviera 1.8 cm como ancho de sépalo?
- ¿A qué especie pertenecería un Iris con un ancho de sépalo de 1.8 cm?
- Pueden realizar para resolver estas preguntas el cálculo de máximos y mínimos de la anchura y longitud del sépalo y pétalo en cada flor.
- Calcular la media aritmética, media generalizada, mediana, moda, varianza, desviación estándar, así como los valores máximos y mínimos de las áreas de los sépalos y pétalos, de cada una de las 3 especies de flores Iris.
- Dar una interpretación de qué representan estas medidas a nivel macro (conjunto de datos) y a nivel micro (por cada especie de flor Iris).

### Consideraciones generales:

- Utilizar rutas dinámicas y permitir al usuario la entrada del archivo (CSV - conjunto de datos Iris)
- Adjuntar en la entrega un ZIP que contenga el conjunto de datos procesado, el código fuente para compilar y un archivo PDF en donde se presenten las respuestas a estas preguntas (pueden ser las tablas de resultados y sus gráficas correspondientes de interpretación con su correspondiente explicación).

**NOTA:** No pueden utilizar funciones predefinidas por el lenguaje, hay que implementar cada función (media, varianza, etc., es decir, programar las fórmulas para cada medida). Recuerden que pueden trabajar en equipos de 2 personas, pero ambos deben subir la asignación de la tarea aunque sea lo mismo. Identificar en el reporte quiénes son esas parejas.



# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Septiembre, 2022



## Representaciones de conjuntos y matrices <sup>(1)</sup>

- Para analizar datos nominales y ordinales, podemos definir **relaciones** entre pares de dichos datos, que pueden analizarse utilizando **métodos relacionales** específicos.

- Por tanto, denotamos datos de características numéricas como el **conjunto**:

$$X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$$

- Con  **$n$**  elementos, donde cada elemento es un **vector de características** de valor real con  $p$ -dimensiones, donde  **$n$**  y  **$p$**  son números enteros positivos. Para  **$p = 1$** , llamamos a  **$X$**  un **conjunto de datos escalares**. Como alternativa a la representación del conjunto, los datos de características numéricas también se representan a menudo como una **matriz**.

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(p)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix}$$

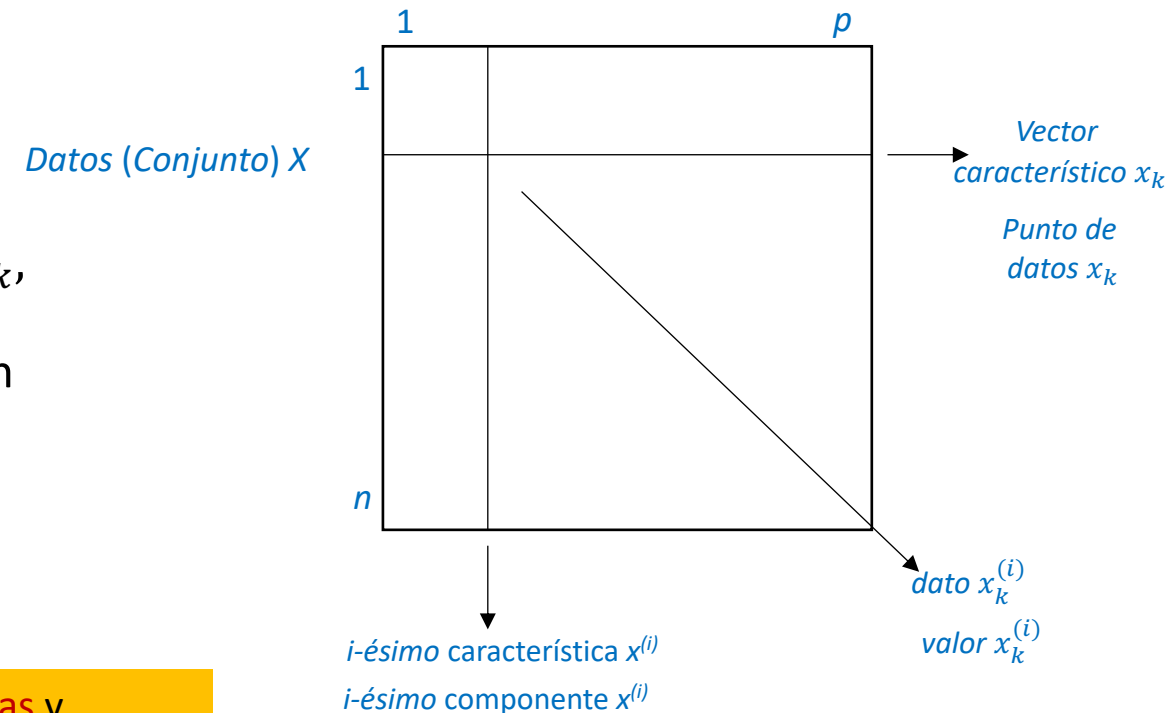
- Así, los vectores  **$x_1, \dots, x_n$**  son **vectores de fila**, aunque matemáticamente los conjuntos de datos y las matrices de datos se usan comúnmente como representaciones de datos equivalentes.



## Representaciones de conjuntos y matrices (2)

- En la siguiente figura se muestran los términos y notaciones comunes en la representación de una matriz de datos.
  - Cada **fila** de la matriz de datos corresponde a un **elemento** del conjunto de datos.
  - Se llama **vector de características** o **punto de datos**  $x_k$ ,  $k = 1, \dots, n$ .
  - Cada **columna** de la matriz de datos corresponde a un **componente** de todos los elementos del conjunto de datos.
  - Se denomina  $i$ -ésima característica o  $i$ -ésima componente  $x^{(i)}$ ,  $i = 1, \dots, p$

Las filas y columnas se identifican usando: **subíndices** para **filas** y **superíndices** entre paréntesis para **columnas**. Las notaciones alternativas en la literatura son:  $x(k, \cdot)$  y  $x(\cdot, i)$ . Un elemento único de la matriz se denomina **componente** de un elemento del conjunto de datos. Se llama **dato** o **valor**  $x_k^{(i)}$ ,  $k = 1, \dots, n$ ;  $i = 1, \dots, p$





## Representaciones de conjuntos y matrices (3)

- El conjunto de datos de Iris se puede escribir como una **matriz de datos** con 150 filas y 4 columnas,
  - donde cada **fila** representa un **objeto (flor)** y cada columna representa una **característica (dimensión)**.
  - La matriz de datos de Iris se puede obtener por **concatenación vertical** de las tres porciones que se muestran en la Tabla.
  - La información de **clase** (Setosa, Versicolor, Virginica) puede interpretarse como una **quinta característica**, en una **escala nominal**.

Setosa				Versicolor				Virginica			
Sepal		Petal		Sepal		Petal		Sepal		Petal	
Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8



## Relaciones <sub>(1)</sub>

- Considérese un conjunto de elementos (abstractos), sin referirse a **vectores de características numéricas**.

$$O = \{o_1, \dots, o_n\}$$

- Algunas veces no hay una **representación vectorial** disponible para los objetos  $o_k, k = 1, \dots, n$ ; por lo que los métodos convencionales de **análisis de datos basados en características** no son aplicables. En cambio, la relación de todos los pares de objetos a menudo se puede cuantificar y escribir como una matriz cuadrada.

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

- Cada valor de relación  $r_{ij}$ ,  $i, j = 1, \dots, n$ , puede referirse a un grado de **similitud**, **disimilitud**, **compatibilidad**, **incompatibilidad**, **proximidad** o **distancia** entre el par de objetos  $o_i$  y  $o_j$ .
- $R$  puede ser **simétrica**, entonces  $r_{ij} = r_{ji}$  para toda  $i, j = 1, \dots, n$ .  $R$  puede definirse manualmente o calcularse a partir de características.





## Relaciones <sub>(2)</sub>

- Si las características numéricas  $X$  están disponibles, entonces  $R$  puede calcularse a partir de  $X$  usando una función apropiada  $f: \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ .
  - Por ejemplo, un analista puede definir manualmente una **matriz relacional** para el conjunto de datos Iris, que compara visualmente y luego califica numéricamente alguna relación entre pares de flores, o  $R$  puede calcularse a partir de las longitudes y anchuras de sépalos y pétalos.
- A continuación se presentan dos clases importantes de relaciones, diferencias y similitudes.



## Medidas de disimilitud <sub>(1)</sub>

- Una función  $d$  se llama de **disimilitud** o **medida de distancia** si para toda  $x, y \in \mathbb{R}^p$ . Entonces tenemos los siguientes axiomas:
  - $d(x, y) = d(y, x)$
  - $d(x, y) = 0 \iff x = y$
  - $d(x, z) \leq d(x, y) + d(y, z)$
- De estos axiomas se tiene ahora que:
  - $d(x, y) \geq 0$
- Una clase de medidas de disimilitud se define usando una **norma**  $\|\cdot\|$  de  $x - y$ , entonces:
  - $d(x, y) = \|x - y\|$



## Medidas de disimilitud (2)

- Una función  $\|\cdot\|: \mathbb{R}^p \rightarrow \mathbb{R}^+$  es una **norma** sí y solo sí:
  - $\|x\| = 0 \iff x = (0,0, \dots, 0)$  (1)
  - $\|\alpha \cdot x\| = |\alpha| \cdot \|x\| \quad \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^p$  (2)
  - $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^p$  (3)
- Por ejemplo, la llamada **norma hiperbólica** de uso frecuente es:
  - $\|x\|_h = \prod_{i=1}^p x^{(i)}$
  - No es una norma según la definición anterior, ya que la condición (1) es violada por  $x = (0,1) \neq (0,0)$  con  $\|x\|_h = \|(0,1)\|_h = 0$ , o la condición (2) se viola por  $x = (1,1)$  y  $\alpha = 2$ , donde  $\|\alpha \cdot x\|_h = \|2 \cdot (1,1)\|_h = \|(2,2)\|_h = 4 \neq |\alpha| \cdot \|x\|_h = |2| \cdot \|(1,1)\|_h = 2$ .



## Medidas de disimilitud <sub>(3)</sub>

- Las clases de normas utilizadas con frecuencia son las:
  - Normas **matriciales**
  - Normas de **Lebesgue**
  - Norma de **Minkowski**
- La norma matricial se define como:  $\|x\|_A = \sqrt{xAx^T}$
- Con una matriz  $A \in \mathbb{R}^{n \times n}$ . Casos especiales importantes de la norma matricial son la **norma Euclídea**.

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- Para  $x \in \mathbb{R}^n$  se define su **norma Euclídea** como:

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \left( \sum_{j=1}^n x_j^2 \right)^{1/2}$$



## Medidas de disimilitud <sub>(4)</sub>

- Demostración de la **norma Euclídea**:

- Aplicando las propiedades se tiene que para cada  $x \in \mathbb{R}^n$  y  $\alpha \in \mathbb{R}$ , se cumple que:

*I.*  $\|x\| > 0$

$$\sqrt{\langle x, x \rangle} = \left( \sum_{j=1}^n x_j^2 \right)^{1/2} > 0$$

*II.*  $\|x\| = 0 \Leftrightarrow x = (0, 0, \dots, 0)$

$$\|x\| = 0 \Leftrightarrow \sqrt{\langle x, x \rangle} = 0 \Leftrightarrow \langle x, x \rangle = 0 \Leftrightarrow x = (0, 0, \dots, 0)$$

*III.*  $\|\alpha \cdot x\| = |\alpha| \cdot \|x\|$

$$\|\alpha \cdot x\| = \left( \sum_{j=1}^n \alpha x_j^2 \right)^{1/2} = \left( \sum_{j=1}^n (\alpha)^2 x_j^2 \right)^{1/2} = (\alpha^2)^{1/2} \left( \sum_{j=1}^n x_j^2 \right)^{1/2} = |\alpha| \cdot \|x\|$$



## Medidas de disimilitud <sup>(5)</sup>

- Demostración de la **norma Euclídea**:

*IV.*  $\|x + y\| \leq \|x\| + \|y\|$  para esta propiedad se tiene:

$$\begin{aligned}\|x + y\|^2 &\leq \langle x + y, x + y \rangle = \\ &= \langle x, x + y \rangle + \langle y, x + y \rangle = \\ &= \langle x + y, x \rangle + \langle x + y, y \rangle = \\ &= \langle x + x \rangle + \langle y + x \rangle + \langle x + y \rangle + \langle y + y \rangle = \\ &= \|x\|^2 + 2\langle x, y \rangle + \|y\|^2 \leq \\ &\leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = \\ &= (\|x\| + \|y\|)^2;\end{aligned}$$

Por lo tanto:

$$\|x + y\| \leq \|x\| + \|y\|$$





## Cálculo de la norma de una matriz, distancia entre matrices y producto interno entre matrices <sub>(1)</sub>

- Vamos a calcular la  $\| A \|$ . Para las matrices:  $A = \begin{bmatrix} -1 & 4 \\ 5 & -2 \end{bmatrix}$   $B = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}$

$$\| x \|_2 = \sqrt{\langle x, x \rangle} = \left( \sum_{j=1}^n x_j^2 \right)^{1/2}$$

$$\| x \|_1 = \sum_{j=1}^n |x_j|$$

$$\| x \|_\infty = \max |x_j|$$

$$\| A \|_2 = \langle A, A \rangle \quad \langle A, A \rangle = a_{11}a_{11} + a_{21}a_{21} + a_{12}a_{12} + a_{22}a_{22}$$

$$\langle A, A \rangle = a_{11}^2 + a_{21}^2 + a_{12}^2 + a_{22}^2$$

$$\langle A, A \rangle = a_{11}^2 + a_{21}^2 + a_{12}^2 + a_{22}^2$$

$$\| A \|_2 = (-1)^2 + 4^2 + 5^2 + (-2)^2$$

$$\| A \|_2 = 46$$

$$\sqrt{\| A \|^2} = \sqrt{46}$$

$$\| A \| = \sqrt{46}$$



## Cálculo de la norma de una matriz, distancia entre matrices y producto interno entre matrices <sub>(2)</sub>

- Ahora vamos a calcular la  $d(A, B)$  con las matrices:  $A = \begin{bmatrix} -1 & 4 \\ 5 & -2 \end{bmatrix}$   $B = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}$

$d(A, B) = \| A - B \|$  Primeramente calcular la operación  $A - B = ?$

$$A - B = \begin{bmatrix} -1 - 0 & 4 - (-1) \\ 5 - 2 & -2 - 1 \end{bmatrix} = \begin{bmatrix} -1 & 5 \\ 3 & -3 \end{bmatrix}$$

$$A - B = \begin{bmatrix} -1 & 5 \\ 3 & -3 \end{bmatrix}$$

Ahora se procede a calcular la norma de  $\| A - B \|^2$

$$\| A - B \|^2 = \langle A - B, A - B \rangle$$

$$\| A - B \|^2 = (-1)^2 + 5^2 + 3^2 + (-3)^2$$

$$\| A - B \|^2 = 44$$

$$\sqrt{\| A - B \|^2} = \sqrt{44}$$

$$\| A - B \| = \sqrt{44}$$

$$d(A, B) = \sqrt{44}$$

$$d(A, B) = 2\sqrt{11}$$



## Cálculo de la norma de una matriz, distancia entre matrices y producto interno entre matrices <sub>(2)</sub>

- Ahora vamos a calcular el producto interno de  $\langle A, B \rangle$  con las matrices:

$$\langle A, B \rangle = a_{11}b_{11} + a_{21}b_{21} + a_{12}b_{12} + a_{22}b_{22}$$

$$A = \begin{bmatrix} -1 & 4 \\ 5 & -2 \end{bmatrix} \quad B = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}$$

$$\langle A, B \rangle = -1(0) + 4(-1) + 5(2) - 2(1)$$

$$\langle A, B \rangle = 4$$



## Proyecto 2 <sub>(1)</sub>

**Proyecto 2. Programa en R que permita calcular la disimilitud con base en la Norma Euclídea sobre el conjunto de datos Iris, para las 3 clases de flores que están descritas en el conjunto de datos.**

- 1. Calcular el valor de disimilitud con la norma Euclídea entre las flores: Setosa, Versicolor y Virginica, tomando como base la longitud y anchura (área) del sépalo y pétalo de cada flor.**
- 2. Establecer el umbral de disimilitud entre las 3 clases de flores, con base en los valores de área de cada flor.**
- 3. Calcular la distancia entre los elementos de la clase Setosa-Versicolor; Setosa-Virginica; Versicolor-Virginica; Versicolor-Setosa; Virginica-Setosa; Virginica-Versicolor.**
- 4. Calcular el producto interno entre cada clase de flores Setosa-Versicolor; Setosa-Virginica; Versicolor-Virginica; Versicolor-Setosa; Virginica-Setosa; Virginica-Versicolor**
- 5. Calcular el producto interno entre: Setosa-Versicolor; Setosa-Virginica; Versicolor-Virginica; Versicolor-Setosa; Virginica-Setosa; Virginica-Versicolor**



# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz



## Distancia de Euclidiana <sup>(1)</sup>

- **Definición**

$$d_E = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

- Donde:

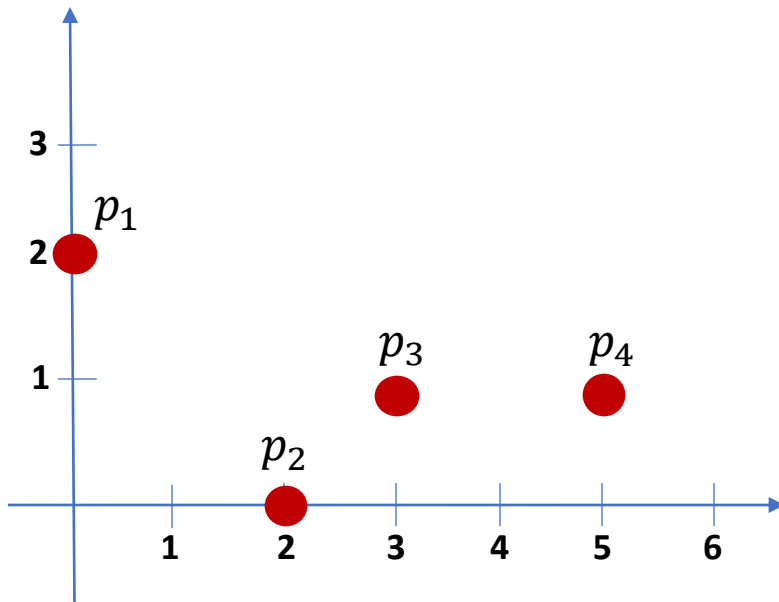
- $n$  es el número de dimensiones (atributos)
- $p_k$  y  $q_k$  son respectivamente los  $k$ -ésimos atributos (componentes) u objetos de datos  $p$  y  $q$ .
- $p_k$  y  $q_k$  son respectivamente los  $k$ -ésimos atributos (componentes) u objetos de datos  $p$  y  $q$ .

- Requiere de una estandarización, en caso de que las escalas difieran entre sí.





## Distancia de Euclidiana (2)



Punto	Altura	Peso
	$x$	$y$
$p_1$	0	2
$p_2$	2	0
$p_3$	3	1
$p_4$	5	1



	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	2.828	3.162	5.099
$p_2$	2.828	0	1.414	3.162
$p_3$	3.1622	1.414	0	2
$p_4$	5.099	3.162	2	0

**Matriz de Distancia**



## Distancia de Minkowski <sub>(1)</sub>

- **Definición**

- La distancia de Minkowski es una métrica en un espacio vectorial normado que puede considerarse como una generalización tanto de la distancia Euclidiana como de la distancia de Manhattan.
- La distancia del orden de Minkowski  $r$ , donde  $r$  es un número entero entre dos puntos.
- Por tanto, se tiene que  $P = (p_1, p_2, \dots, p_n)$  y  $Q = (q_1, q_2, \dots, q_n) \in \mathbb{R}^n$ , se define como:

$$Dist L_P = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

- Donde:
  - $r$  es un parámetro
  - $n$  es el número de dimensiones (atributos),
  - $p_k$  y  $q_k$  son respectivamente los  $k$ -ésimos atributos (componentes) u objetos de datos  $p$  y  $q$ .
- Para  $p \geq 1$ , la distancia de Minkowski es una métrica como resultado de la Desigualdad de Minkowski.
- Cuando  $p < 1$ , la distancia entre  $(0,0)$  y  $(1,1)$  es  $2^{\frac{1}{p}} > 2$ , pero el punto  $(0,1)$  está a la distancia de 1 de estos dos puntos.
- Esto viola la propiedad de desigualdad del triángulo, porque  $p < 1$  no es una métrica. Sin embargo se puede obtener una métrica para estos valores, simplemente eliminando el exponente  $\frac{1}{p}$



## Distancia de Minkowski <sub>(2)</sub>

- **Definición**

- Para  $r \geq 1$ , la distancia de Minkowski es una métrica como resultado de la Desigualdad de Minkowski.
- Cuando  $r < 1$ , la distancia entre  $(0,0)$  y  $(1,1)$  es  $2^{\frac{1}{r}} > 2$ , pero el punto  $(0,1)$  está a la distancia de 1 de estos dos puntos.
- Esto viola la propiedad de desigualdad del triángulo, porque  $r < 1$  no es una métrica. Sin embargo se puede obtener una métrica para estos valores, simplemente eliminando el exponente  $\frac{1}{r}$
- La distancia de Minkowski también se puede ver como un múltiplo de la potencia media de las diferencias por componentes de  $r$ .



## Distancia de Minkowski <sub>(3)</sub>

- **Definición**

- La familia de distancias de Minkowski  $Dist L_P = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$ 
  - Costo de evaluación  $O(n)$
  - Cumplen con las propiedades métricas
  - Cuando  $r = 1$ : Se conoce como City Block (Manhattan, [Taxicab](#), norma  $L_1$ )
    - Un ejemplo común de esto es la distancia de *Hamming*, que es solo la cantidad de bits que son diferentes entre dos vectores binarios.
  - Cuando  $r = 2$ : Distancia Euclidiana (norma  $L_2$ )
  - Cuando  $r = \infty$ : Distancia Chebyshev (chessboard, norma  $L_{max}$ , norma  $L_\infty$ , Distancia Máxima, Distancia Suprema)
    - Esta es la diferencia máxima entre cualquier componente de los vectores.
      - Por ejemplo:  $L_\infty$  de  $(1, 0, 2)$  y  $(6, 0, 3) = ??? = 5$
  - No confundir  $r$  con  $n$ , es decir, todas estas distancias están definidas para todos los números de dimensiones.

$$L_1(P, Q) = \sum_{k=1}^n |p_k - q_k|$$

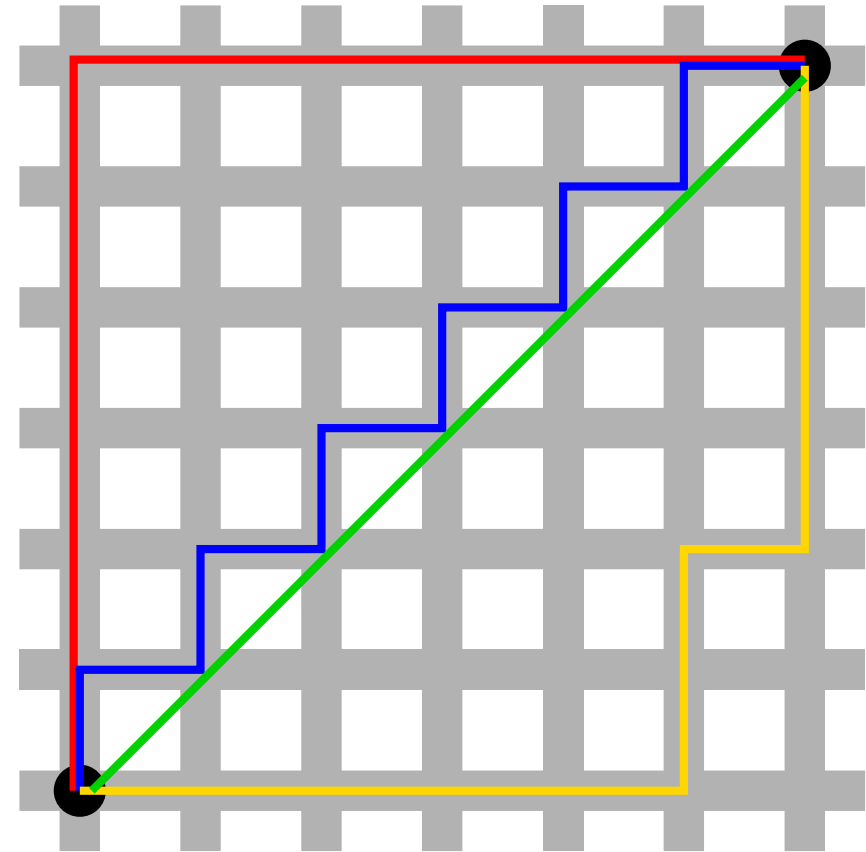
$$L_2(P, Q) = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

$$L_{max}(P, Q) = \max_{1 \leq k \leq n} \{|p_k - q_k|\}$$



# Distancia de Minkowski <sup>(4)</sup>

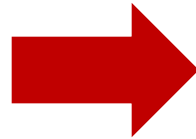
**Taxicab**





## Distancia de Minkowski <sup>(5)</sup>

Punto	$x$	$y$
$p_1$	0	2
$p_2$	2	0
$p_3$	3	1
$p_4$	5	1



### Matriz de Distancia

$L_1$	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	4	4	6
$p_2$	4	0	2	4
$p_3$	4	2	0	2
$p_4$	6	4	2	0

$L_2$	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	2.828	3.162	5.099
$p_2$	2.828	0	1.414	3.162
$p_3$	3.162	1.414	0	2
$p_4$	5.099	3.162	2	0

$L_\infty$	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	2	3	5
$p_2$	2	0	1	3
$p_3$	3	1	0	2
$p_4$	5	3	2	0



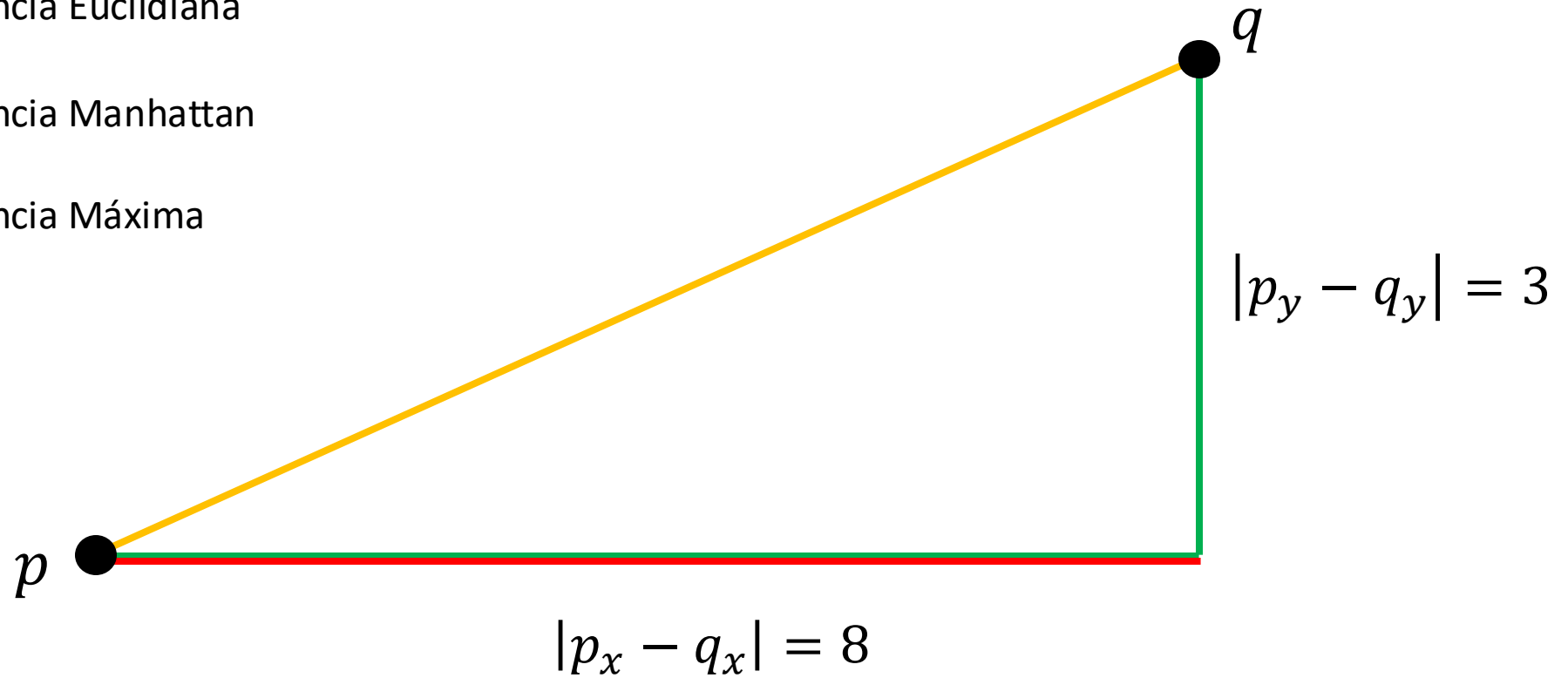


## Distancia de Minkowski <sup>(6)</sup>

 Distancia Euclidiana

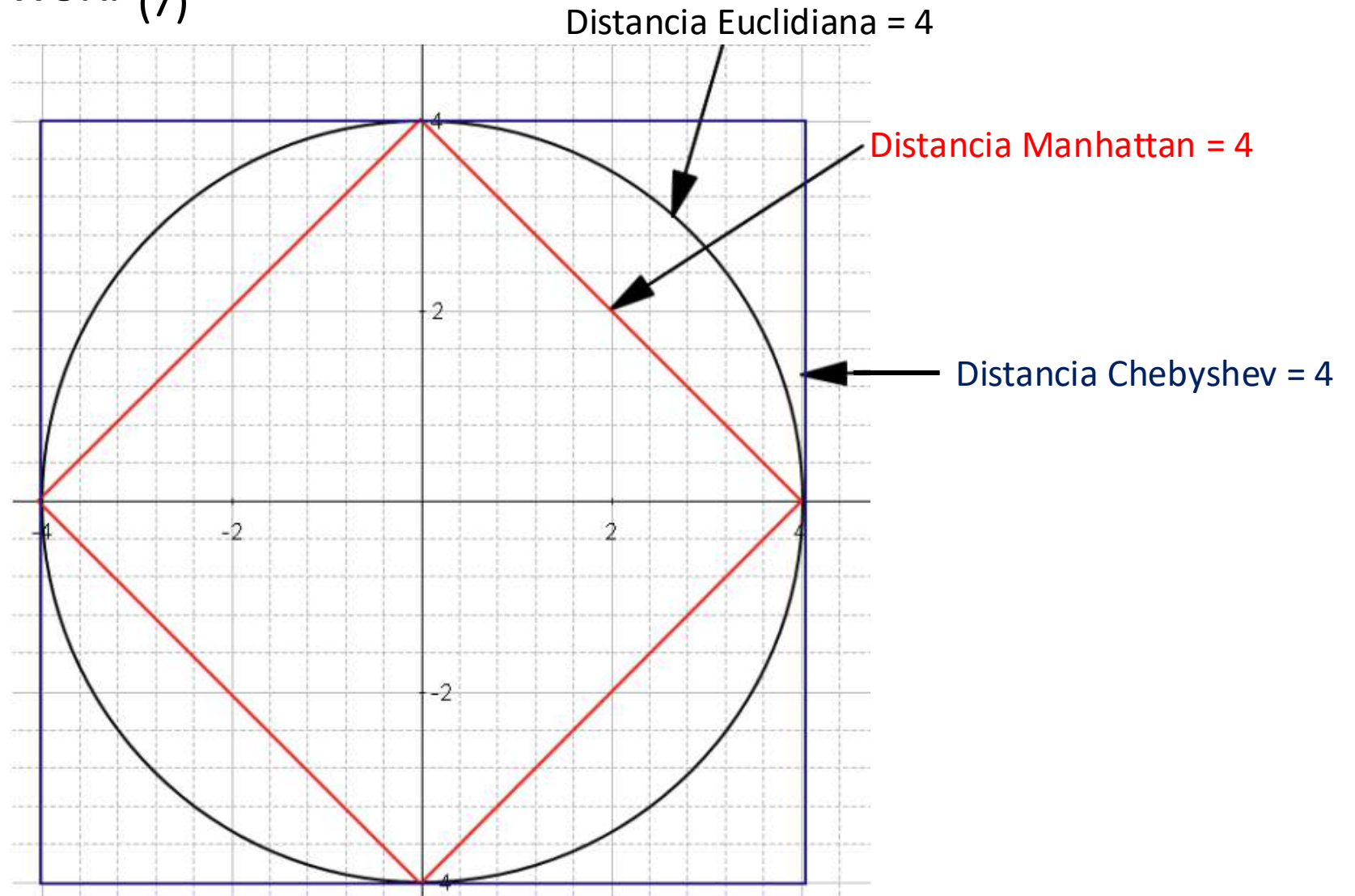
 Distancia Manhattan

 Distancia Máxima





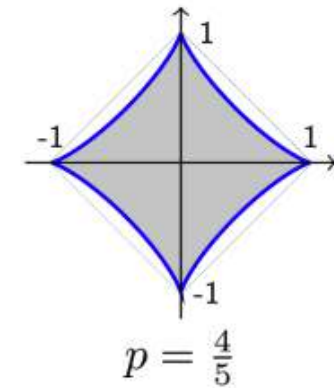
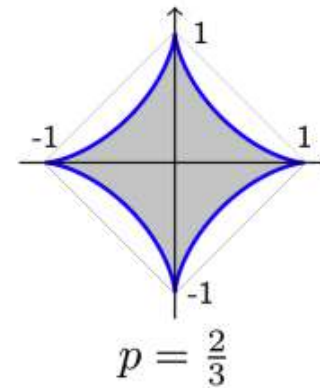
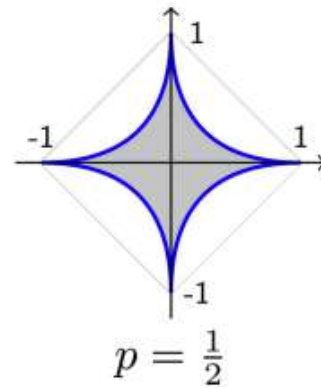
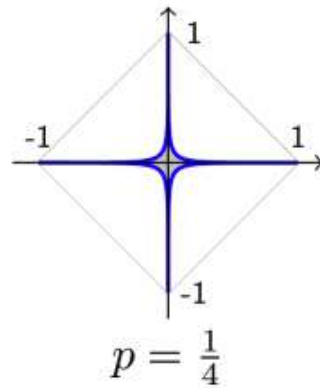
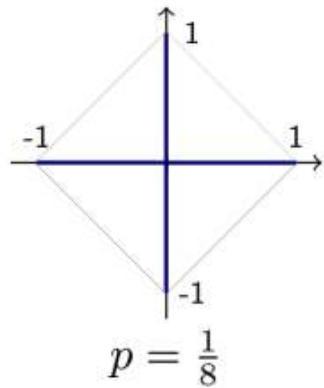
## Distancia de Minkowski <sup>(7)</sup>



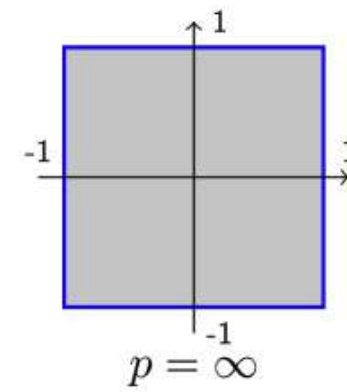
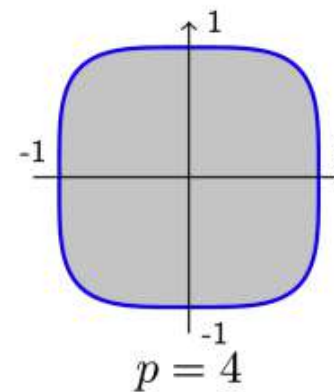
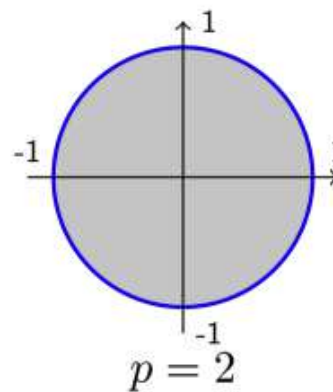
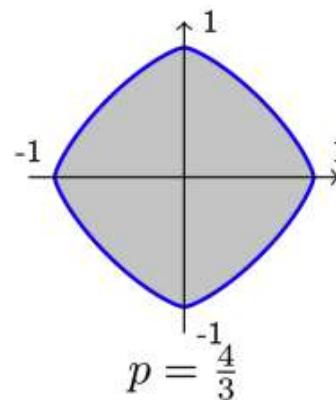
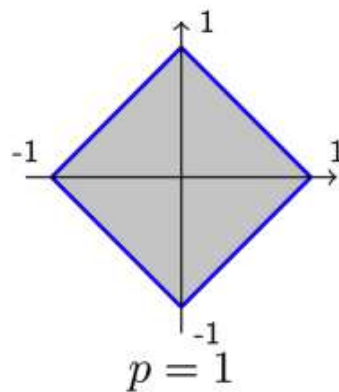


# Distancia de Minkowski (8)

$$C_p = \{(x, y) | (|x|^p + |y|^p)^{1/p} \leq 1\}$$



$p < 1$ : Conjuntos no convexos



$p \geq 1$ : Conjuntos convexos



# Distancia de Hamming <sup>(1)</sup>

- **Definición**

- La distancia de Hamming se define como:  $D_H(x, y) = \sum_{i=1}^p \rho(x^{(i)}, y^{(i)})$  y considerando la métrica discreta:

$$\rho(x, y) = \begin{cases} 0 & \text{si } x = y \\ 1 & \text{Otro caso} \end{cases}$$

- Entonces, la distancia de Hamming produce la cantidad de valores de características que no coinciden.
  - Para características binarias, la distancia de Hamming es igual a la distancia de Manhattan.
- Sin embargo, la distancia de Hamming no está asociada con una norma porque la condición  $\| \alpha \cdot x \| = |\alpha| \cdot \| x \| \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^p$  no se cumple.
- Las variantes de la distancia de Hamming usan funciones modificadas  $\rho$  para especificar similitudes entre características individuales.
  - Por ejemplo, si las características son páginas web (escala nominal), entonces  $\rho$  podría ser menor para pares de páginas con contenido similar y mayor para pares de páginas con contenido bastante diferente.



## Distancia de Hamming <sup>(2)</sup>

- **Definición**

- La distancia de Hamming se utiliza en procesamiento de señales y telecomunicaciones.
  - Contar el número de bits corruptos en la transmisión de un mensaje de una longitud determinada.
- Permite cuantificar la diferencia entre dos secuencias de símbolos.
- Es una distancia en el sentido matemático, con dos secuencias de símbolos de la misma longitud y asocia el número de posiciones donde difieren las dos secuencias.
- Para comparar secuencias de longitudes variables o cadenas de caracteres que pueden sufrir no solo sustituciones, sino también inserciones o borrados, se utiliza la **distancia de Levenshtein**.

- **Ejemplo:**

- $\alpha = (0\ 0\ 0\ 1\ 1\ 1\ 1)$ ;  $\beta = (1\ 1\ 0\ 1\ 0\ 1\ 1) \therefore d_H = 1 + 1 + 0 + 0 + 1 + 0 + 0 = 3$
- La distancia entre  $\alpha$  y  $\beta$  es igual a 3 porque **3 bits difieren**.
- La distancia de Hamming entre **(1 0 1 1 0 1)** y **(1 0 0 1 0 0 1)** es **2**
- La distancia de Hamming entre **(2 14 3 8 96)** y **(2 23 3 7 96)** es **3**
- La distancia de Hamming entre "**r a m e r**" y "**c a s e s**" es **3**



# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz



## Similitud y Disimilitud <sup>(1)</sup>

- **Similitud**

- Es una medida numérica que define qué tan parecidos son dos objetos de datos...
- Es más alta cuando los objetos son más parecidos.
- A menudo caen en el rango  $[0,1]$ .

- **Disimilitud**

- Es una medida numérica que define qué tan diferentes son dos objetos de datos...
  - Cuanto más bajo es el valor es cuando los objetos son más parecidos.
  - La disimilitud mínima suele ser 0.
  - El límite superior varía.
- En resumen, la **proximidad** se refiere a una similitud o disimilitud.





## Medidas de Similitud <sub>(1)</sub>

- En búsqueda y/o recuperación de información:
  - Una medida de similitud puede representar la similitud entre dos **documentos**, dos **consultas** o un documento y una consulta.
  - Es posible clasificar (**ranking**) los documentos recuperados en el orden de supuesta importancia.
  - Una medida de similitud es una función que calcula el **grado de similitud** entre un par de objetos (documentos).
  - Hay un gran número de medidas de similitud propuestas en la literatura,
    - Porque la **mejor** medida de similitud **no existe** (¡todavía!).



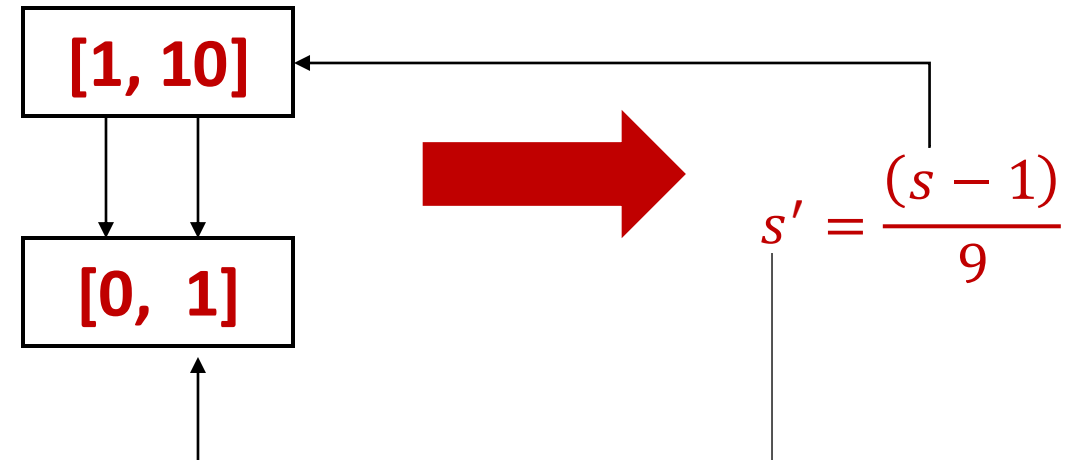
## Medidas de Similitud <sub>(2)</sub>

- Conversión de valores de similitud:

- Por ejemplo:
  - 1 muestra similitud incompatible
  - 10 muestra similitud absoluta

- En general podemos utilizar:

$$s' = \frac{(s - \min_s)}{(\max_s - \min_s)}$$





## Medidas de Similitud <sub>(3)</sub>

- Definición
  - Una función  $s$  se llama medida de **similitud** o proximidad si para toda  $x, y \in \mathbb{R}^p$ .
    - $s(x, y) = s(y, x)$
    - $s(x, y) \leq s(x, x)$
    - $s(x, y) \geq 0$
  - La función  $s$  se llama medida de **similitud normalizada** si:
    - $s(x, x) = 1$
- Cualquier medida de disimilitud  $d$  se puede usar para definir una medida de similitud correspondiente  $s$  y viceversa. Por ejemplo, usando una función positiva monótonicamente decreciente  $f(0) = 1$  tal como:

$$s(x, y) = \frac{1}{1 + d(x, y)}$$



## Modelo de Espacio Vectorial <sup>(1)</sup>

- Es un modelo algebraico utilizado para el **filtrado**, **recuperación**, **indexado** y cálculo de **relevancia** de **información**.
- Puede ser utilizado en datos de una manera formal mediante el uso de **vectores** (que pueden ser identificadores, o por ejemplo términos de búsqueda) en un **espacio lineal multidimensional**.
- Muchas de las tareas de recuperación de información como la **búsqueda**, **agrupamiento** o **categorización** de documentos tienen como primer objetivo procesar documentos en lenguaje natural.
- El problema que surge es que los algoritmos que pretenden resolver estas tareas necesitan representaciones internas **explícitas** de los **documentos**.



## Modelo de Espacio Vectorial <sup>(2)</sup>

- En el área de recuperación de información normalmente se usa una expresión vectorial, donde las **dimensiones** del **vector** representan términos, frases o conceptos que aparecen en el documento.
- En este aspecto la representación más adoptada es la conocida como **bolsa de palabras**: que es una colección de documentos compuesta por  **$n$**  documentos indexados y  **$m$**  términos representados por una matriz documento-término de  **$n \times m$** .
- Donde los  **$n$**  vectores renglón representan los  **$n$**  documentos; y el valor asignado a cada componente refleja la **importancia** o **frecuencia ponderada** que produce el término, frase o concepto  **$t_i$**  en la representación semántica del documento  **$j$** .

$$d_j(\omega_{1j}, \omega_{2j}, \dots, \omega_{mj})$$



## Modelo de Espacio Vectorial <sup>(3)</sup>

- Donde  $m$  es la cardinalidad del diccionario (una lista de términos únicos que aparecen en un conjunto de documentos) y  $0 \leq \omega_{ij} \leq 1$  representa la contribución del término  $t_i$  para la representación semántica del documento  $d_j$ .
- En esta representación vectorial de documentos el éxito o fracaso se basa en la ponderación o peso de los términos.
- Aunque existen muchos trabajos sobre técnicas de ponderación de términos, en realidad no hay un consenso sobre cuál método es el mejor.
- También hay que destacar que el espacio de renglones de la **matriz documento-término** determinan el contenido semántico de la colección de documentos. Sin embargo, una combinación lineal de dos vectores-documento no representa necesariamente un documento viable de la colección.
- Mediante el modelo espacio vectorial se pueden explotar las **relaciones geométricas** entre dos vectores documento (y términos) a fin de expresar las similitudes y diferencias entre términos.



## Modelo de Espacio Vectorial (4)

- Si bien el **rendimiento** de un sistema de recuperación de información depende en gran medida de las medidas de similitud entre documentos, la ponderación de términos desempeña un papel fundamental para que esa similitud entre documentos sea más confiable.
- Por ejemplo, mientras que una representación de documentos basada solo en las **frecuencias** o apariciones de términos no es capaz de representar adecuadamente el contenido semántico de los documentos, la representación de **términos ponderados** (aplicación de métodos de normalización a la matriz documento-término) hace frente a errores o incertidumbres asociadas a la representación simple de documentos.





## Modelo de Espacio Vectorial (5)

- Implementación del modelo
- Una colección de  $n$  documentos indexados por  $m$  términos puede ser representada por una matriz  $A$  de dimensión  $n \times m$ , donde cada elemento  $a_{ij}$  es usualmente definido por una frecuencia ponderada del término  $i$  en el documento  $j$  cuyo objetivo principal es mejorar el rendimiento en la recuperación de información.
- Entendiendo como rendimiento la **habilidad** de recuperar información **relevante** y descartar información **irrelevante**.

Matriz documento-término simple, donde cada columna representa un **término** en la colección, cada renglón un **documento** y cada celda o elemento de la matriz la **ocurrencia** en el documento



	Término 1	Término 2	Término 3
Documento 1	1	0	0
Documento 2	0	0	1
Documento 3	1	1	1
Documento 4	0	1	0

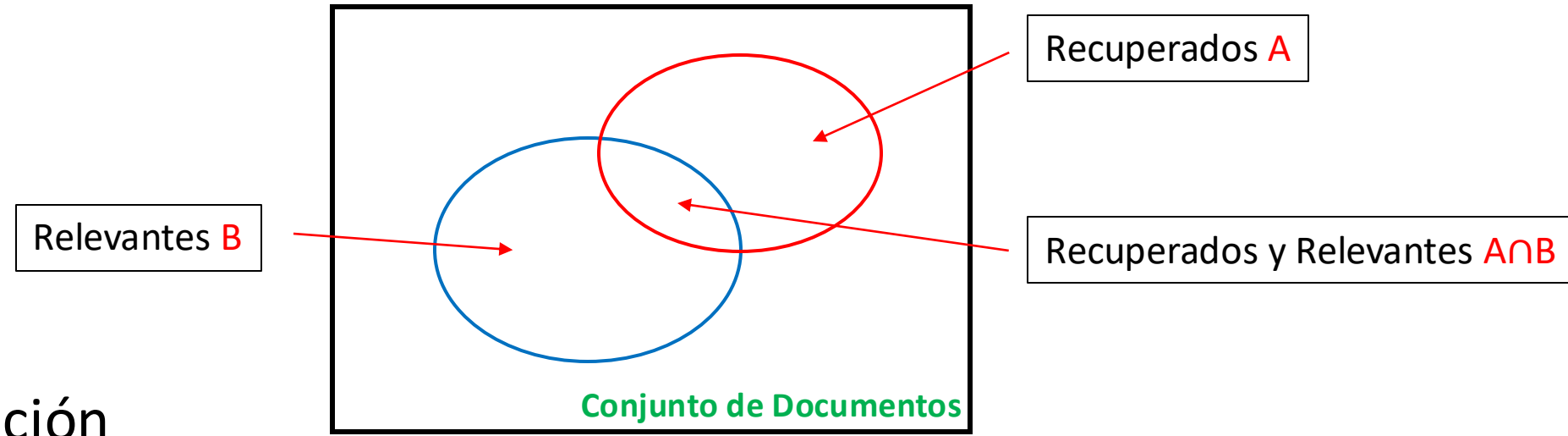


## Modelo de Espacio Vectorial <sup>(6)</sup>

- A partir de la matriz, se puede observar que el **Término 1** aparece en el **Documento 1 y 3**, pero no en los otros dos documentos. Se demuestra así que cada renglón de la matriz de 4×3 puede ser representado en un espacio de tres dimensiones.
- Entonces cada elemento  $a_{ij}$  de la matriz documento-término A queda definido como:  $a_{ij} = l_{ij} * g_i * d_j^{-1}$
- Donde:  $l_{ij}$  es el peso local del término  $i$  en el documento  $j$ , el cual mide la importancia de dicho término en el documento;  $g_i$  es el peso global del término  $i$  en la colección de documentos y  $d_j$  es el factor de normalización para el  $j$ -ésimo documento.
- **Peso local:** mide la importancia del término  $i$  en el documento  $j$  y solo depende de las frecuencias en el documento y no de otros documentos.
- **Peso global:** Son aquellos que toman información de la colección de documentos para obtener el peso de un término en un documento.



## Modelo de Espacio Vectorial (7)



- Medidas de evaluación

- $Precisión = \frac{\text{Documentos relevantes Devueltos}}{\text{Total de Documentos Devueltos}}$

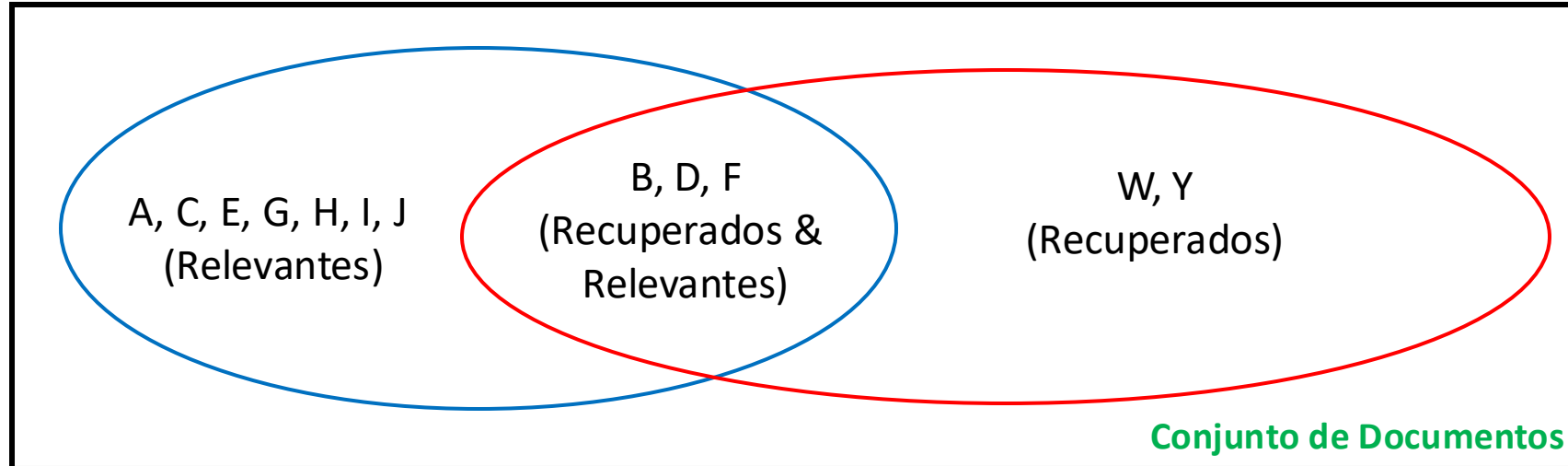
- $P(A, B) = \frac{|A \cap B|}{|A|}$

- $Recall = \frac{\text{Documentos relevantes Devueltos}}{\text{Total de Documentos Relevantes}}$

- $R(A, B) = \frac{|A \cap B|}{|B|}$



## Modelo de Espacio Vectorial (8)

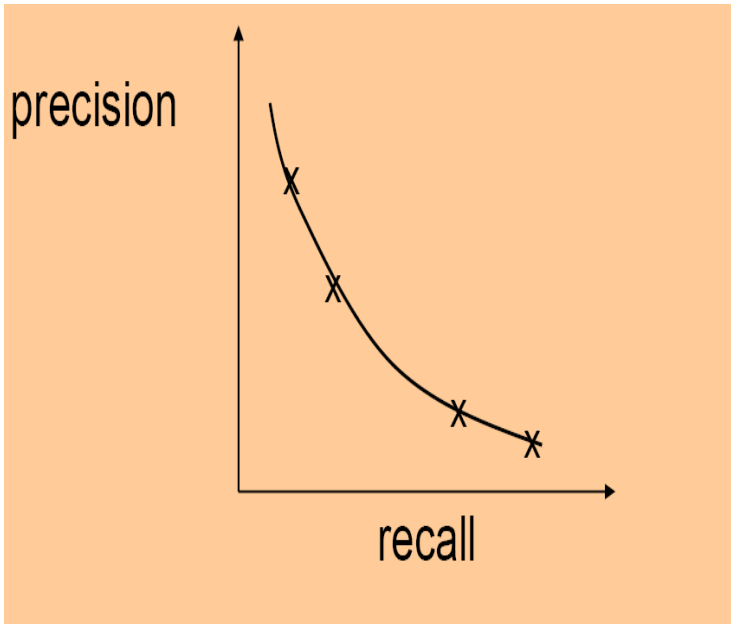


- Conjunto de documentos
  - $|A| = \{recuperados\} = \{B, D, F, W, Y\} = 5$
  - $|B| = \{relevantes\} = \{A, B, C, D, E, F, G, H, I, J\} = 10$
  - $|A \cap B| = \{recuperados\} \cap \{relevantes\} = \{B, D, F\} = 3$
  - $Precision = \frac{3}{5} = 60\%$
  - $Recall = \frac{3}{10} = 30\%$



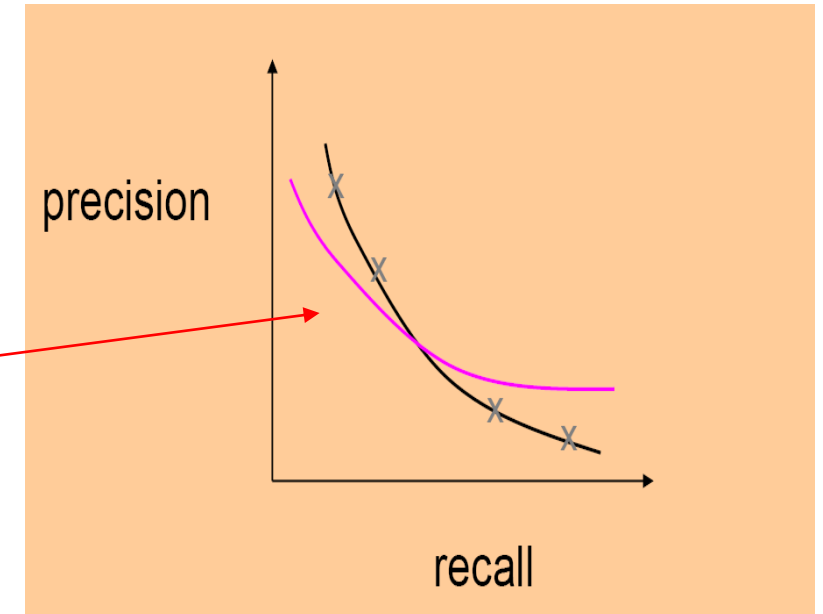
## Modelo de Espacio Vectorial (9)

- Siempre existe una compensación entre las medidas *precision* y *recall*.
- Por tanto, se mide el *precision* en diferentes niveles de *recall*.



Una consulta

Difícil determinar cuál de estos  
dos resultados hipotéticos es  
mejor



Dos consultas



## Definición de medidas de similitud <sup>(1)</sup>

- Consideremos primero las similitudes entre los vectores de características binarias.
  - Un par de vectores de características binarias pueden considerarse similares si muchos coinciden.
  - Esta **coincidencia** se puede representar mediante la operación producto, por lo que el **producto escalar** de los vectores de características es un candidato razonable para una medida de similitud.
  - También para características de valores reales no negativos  $x, y \in (\mathbb{R}^+)^p$ . Entonces, las medidas de similitud se pueden definir en función de productos escalares que se pueden normalizar de diferentes maneras.
    - Medida de Similitud del Coseno
    - Medida de Similitud Dice
    - Medida de Similitud Jaccard (Tanimoto)
    - Medida de Similitud de Sobreposición (Overlap)



## Definición de medidas de similitud <sup>(2)</sup>

- Estas expresiones no están definidas para vectores de características **cero** porque los **denominadores** son cero. Entonces, la similitud debe definirse explícitamente para este caso como cero.
- Por ejemplo, la similitud del coseno es **invariable** frente a la escala (positiva) de los vectores de características  $y$ , por lo tanto, considera la distribución relativa de las características, cumpliendo con:
  - $s(c \cdot x, y) = s(x, y)$
  - $s(x, c \cdot y) = s(x, y)$
  - Para toda  $x, y \in \mathbb{R}^p$  y  $c > 0$
- Por ejemplo, considerar dos recetas de pasteles:
  - **Receta 1.** 3 huevos,  $1 \frac{1}{2}$  tazas de azúcar,  $1 \frac{1}{2}$  tazas de harina y  $\frac{1}{2}$  taza de mantequilla.
  - **Receta 2.** 6 huevos, 3 tazas de azúcar, 3 tazas de harina y 1 taza de mantequilla.
- Obviamente, ambas recetas dan el mismo resultado, pero la segunda rinde el doble de pastel que la primera.
- Por tanto, siguiendo una expectativa intuitiva, la **similitud del coseno** entre las dos recetas es igual a **uno**.
- Entonces, podemos concluir que estas medidas cuantifican la similitud entre vectores de características (**filas de la matriz de datos**). Pero para cuantificar la similitud entre características (**columnas de la matriz de datos**), se utiliza la **correlación**.
- Si la matriz de datos es transpuesta (las filas y las columnas se intercambian), la correlación también se puede usar como una forma alternativa de cuantificar la similitud entre los vectores de características.





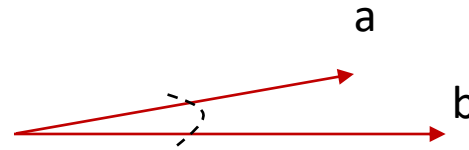
## Similitud del Coseno <sup>(1)</sup>

- En minería de datos, la similitud del coseno se refiere a la **distancia con dimensiones** que representan las **características** de los objetos de datos en un conjunto de datos.
- También conocida como **distancia coseno**, y se define como la medida de la **magnitud** de la diferencia entre dos individuos, usando el valor coseno del ángulo entre dos vectores en un espacio vectorial.
- En la similitud del coseno, los objetos de datos en un conjunto se procesan como **vectores**.
- Cuanto más **cercano** es el valor del coseno a **1**, más cerca está el ángulo a **0 grados**; es decir, los **dos vectores** son **más similares**, lo que se denomina como “similitud del coseno”.

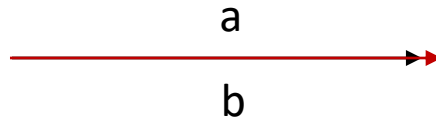


## Similitud del Coseno <sup>(2)</sup>

- Por ejemplo, en la figura se puede apreciar que el **ángulo** entre los dos vectores ***a*** y ***b*** es muy **pequeño**. Por tanto, se puede decir que el vector ***a*** y el vector ***b*** tienen una gran **similitud**.



- En casos extremos, cuando los **vectores *a* y *b*** coinciden completamente. Entonces ***a* y *b*** son iguales, lo que significa que los datos representados por los vectores ***a* y *b*** son completamente **similares o iguales**.



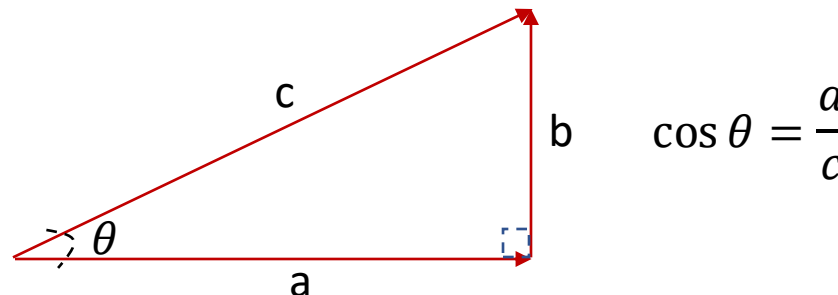


## Similitud del Coseno (3)

- Pero que pasa si el ángulo entre los vectores  $a$  y  $b$  es **grande**, o se ubica en la dirección opuesta, se puede decir que el vector  $a$  y el vector  $b$  tienen una **baja similitud**, o que el conjunto de datos representado por los vectores  $a$  y  $b$  no es básicamente similar.



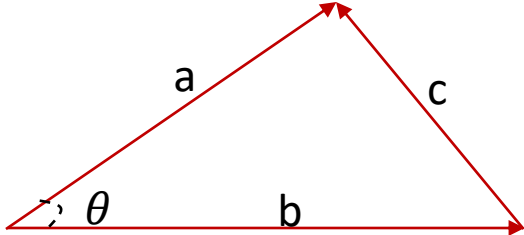
- Por tanto, la teoría de la similitud del coseno del espacio vectorial es un método para calcular la **similitud de los individuos**, a partir del siguiente análisis.
- Cuando se piensa en la fórmula del coseno, el método de cálculo más básico es  $\theta$ .





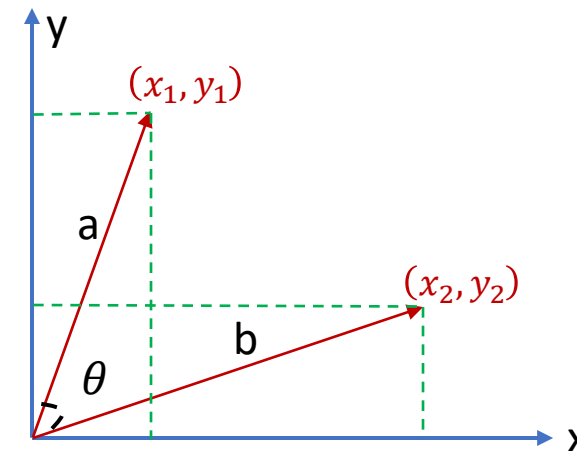
## Similitud del Coseno <sup>(4)</sup>

- Pero la fórmula anterior, es solo aplicable para **triángulos rectángulos**, y en los triángulos no rectángulos, la fórmula para calcular el coseno del ángulo **a** y **b** sería:


$$\cos \theta = \frac{a^2 + b^2 - c^2}{2ab}$$

- Por ejemplo, en el triángulo representado por el vector, suponiendo que el vector **a** es **(x<sub>1</sub>, y<sub>1</sub>)** y el vector **b** es **(x<sub>2</sub>, y<sub>2</sub>)**, el teorema del coseno se puede reescribir de la siguiente forma a partir de la figura.

$$\begin{aligned} \cos \theta &= \frac{a \cdot b}{||a|| \cdot ||b||} = \frac{(x_1, y_1) \cdot (x_2, y_2)}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}} \\ &= \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}} \end{aligned}$$





## Similitud del Coseno <sup>(5)</sup>

- Para el caso, si los vectores **a** y **b** no son bidimensionales, el método de cálculo del coseno anterior sigue siendo correcto. Suponiendo que **a** y **b** son dos vectores de  $n$ -dimensiones, **a** es **B** y **b** es **A**, entonces el coseno del ángulo entre **a** y **b** es igual a:

$$\cos \theta = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p (x_i)^2 \sum_{i=1}^p (y_i)^2}} = \frac{a \cdot b}{||a|| \cdot ||b||}$$

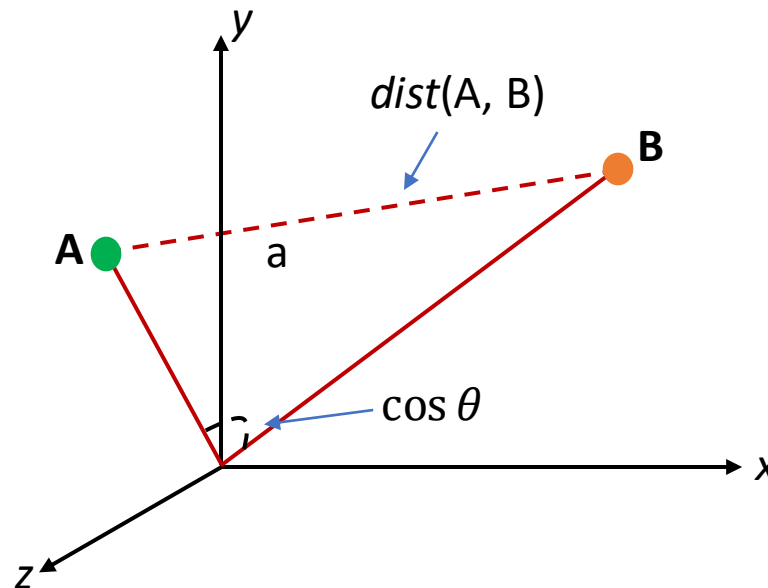
$$\text{sim}(x, y) = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p (x_i)^2 \sum_{i=1}^p (y_i)^2}}$$

- Por tanto, mientras más **cerca** esté el valor del coseno a **1**, más **cerca** estará el ángulo a **0 grados**, es decir, los **dos vectores** son más **similares** y el ángulo es igual a 0. Esto indica que los dos vectores son iguales, y se denomina **similitud del coseno**.



## Similitud del Coseno <sup>(6)</sup>

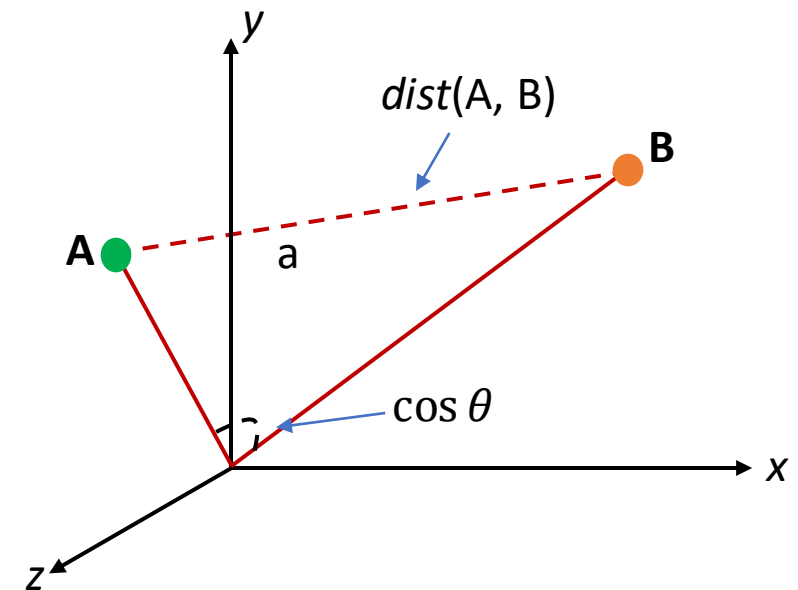
- Por otra parte, la distancia del coseno utiliza el valor del coseno del ángulo entre los dos vectores como una medida de la **diferencia** entre los dos **individuos**. En comparación con la distancia euclidiana, la distancia del coseno presta más atención a la **diferencia** en la **dirección** de los **vectores**.
- Utilizando un sistema con **coordenadas tridimensionales**, se puede observar la diferencia entre la distancia euclidiana y la del coseno.





## Similitud del Coseno <sup>(7)</sup>

- La distancia euclidiana mide la **distancia absoluta** de cada punto en el espacio y está directamente relacionada con las **coordenadas** de posición de cada punto.
- La distancia del coseno mide el **ángulo del vector espacial**, que se refleja más en la diferencia de la dirección, no de la ubicación, pero si mantiene la posición del punto A sin cambios y el punto B lejos del origen del eje de coordenadas en la dirección original.
- Entonces la distancia del coseno ( $\cos\theta$ ) permanece igual (**porque el ángulo no cambia**) y la distancia entre los puntos A y B si está cambiando.
- Esta es la diferencia entre la distancia euclidiana y la distancia del coseno.







## Similitud del Coseno <sup>(8)</sup>

- La distancia euclidiana y del coseno tienen diferentes **métodos de cálculo** y diferentes **características de medición**, por lo que son adecuadas para diferentes modelos de análisis de datos, por ejemplo:
  - La distancia **euclidiana** puede reflejar la **diferencia absoluta** de las características numéricas individuales, por lo que se usa más para aquellos análisis que requieren reflejar la diferencia del valor numérico de la dimensión, como el uso de indicadores de comportamiento del usuario para analizar la similitud o diferencia en el valor del usuario.
  - La distancia del **coseno** se usa más para **distinguir** la **diferencia** de la **dirección**, pero no es sensible al valor absoluto. Se utiliza para distinguir la **similitud** y la diferencia de interés por la calificación del contenido por parte del usuario. Al mismo tiempo, se corrigen los posibles estándares de medición entre usuarios.



## Similitud del Coseno <sup>(9)</sup>

- Ejemplo de cálculo de la similitud de coseno para encontrar similitud en texto:
    - **Texto 1.** Julie loves me more than Linda loves me
    - **Texto 2.** Jane likes me more than Julie loves me
  - ¿Cómo calcular la similitud entre los dos textos?
    - Queremos saber en qué medida se parecen estos textos, únicamente en términos de **recuento de palabras** (e ignorando el orden de las mismas). Comenzamos haciendo una lista de ambos textos.
- me Julie loves Linda than more likes Jane
- Se procede en obtener la **frecuencia** que aparece cada una de estas palabras en cada texto.

Palabra	Freq-T <sub>1</sub>	Freq-T <sub>2</sub>
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1



## Similitud del Coseno <sub>(10)</sub>

- Sin embargo, no nos interesan las palabras en sí; solo nos interesan esos **dos vectores verticales** de frecuencia.
  - Por ejemplo, hay dos casos de “**me**” en cada texto. Vamos a decidir lo **cerca** que están estos dos textos **entre sí** calculando una función de esos dos vectores, particularmente el coseno del ángulo entre ellos.
    - **a**: [2, 0, 1, 1, 0, 2, 1, 1]
    - **b**: [2, 1, 1, 0, 1, 1, 1, 1]
    - Aplicando la fórmula la **similitud del coseno**

$$sim(x, y) = \frac{\sum_{i=1}^p x_i y_i}{\sqrt{\sum_{i=1}^p (x_i)^2 \sum_{i=1}^p (y_i)^2}} = sim(x, y) = 0.822$$

- Estos vectores son de **8 dimensiones**. Una virtud de usar la similitud del coseno es que convierte una cuestión que va más allá de la capacidad humana para visualizarla, en una que puede serlo.
- En este caso se podría pensar esto, como un **ángulo de unos 35 grados**, que está a cierta “**distancia**” de 0 o de la concordancia perfecta.

Palabra	Freq-T <sub>1</sub>	Freq-T <sub>2</sub>
me	2	2
Jane	0	1
Julie	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1



## Proyecto 4 <sub>(1)</sub>

**Proyecto 5. Desarrollar un programa en R que permita calcular la similitud del coseno a partir del conjunto de datos: [Abstract COVID Papers.csv](#)**

### **Instrucciones:**

- 1. Analizar el conjunto de datos que contiene 3 atributos: “title”, “abstract” y “url”.**
- 2. Revisar el corpus de datos y los metadatos directamente de la página: <https://www.kaggle.com/datasets/anandhuh/covid-abstracts>**
- 3. Seleccionar 20 abstracts y aplicarles la similitud del coseno. Cada uno debe seleccionar diferentes títulos para que no sean iguales. El dataset está compuesto por 10,000 instancias.**
- 4. Posteriormente, *rankear* los abstracts por grado de similitud y visualizarlos por el título.**



# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

Diciembre, 2025



## Similitud de Dice <sub>(1)</sub>

- Esta medida de similitud también se le conoce como el **Coeficiente de Sørensen**. Es una métrica estadística que se utiliza para comparar la **similitud** de conjuntos de datos.
- La fórmula original de Sørensen está destinada a ser aplicada a datos con **presencia/ausencia** de valores y se define como sigue:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|} = D_S = \frac{2C}{A + B}$$

- Donde: **A** y **B** son el número de elementos muestra de un conjunto de datos respectivamente y **C** es el número de especies compartidas por los dos elementos muestra.
- **D<sub>S</sub>** es el coeficiente de similitud y varía entre 0 y 1. Esta expresión se extiende fácilmente a la **abundancia** en lugar de la presencia/ausencia de especies.
- La versión cuantitativa del índice de Sørensen también se conoce como el **índice binario de Czekanowski**.



## Similitud de Dice <sub>(2)</sub>

- El conjunto de operaciones se pueden expresar en términos de **operaciones vectoriales** sobre **vectores binarios**  $A$  y  $B$ , lo cual proporciona el mismo resultado en vectores binarios y también da una similitud más general sobre los vectores en términos generales.

$$D_S = \frac{2|A \cdot B|}{|A|^2 + |B|^2} = s(x, y) = \frac{2 \sum_{i=1}^p x_i y_i}{\sum_{i=1}^p (x_i)^2 + (y_i)^2}$$

- Para el caso de conjuntos por ejemplo,  $X$  e  $Y$  de **palabras clave** utilizadas en la recuperación de la información, el coeficiente puede ser definido como **dos veces** la información compartida (**intersección**) sobre la suma de las cardinalidades.
- Entonces el coeficiente se puede utilizar también como medida de similitud entre **cadenas**. Dadas dos secuencias  $x$  e  $y$ , se puede calcular el coeficiente como sigue:

$$D_S = \frac{2n_t}{n_x + n_y}$$



## Similitud de Dice <sub>(3)</sub>

$$D_S = \frac{2n_t}{n_x + n_y}$$

- Donde  $n_t$  es el número de **dígrafos** (n-gramas) (formado por dos caracteres consecutivos) en común a las dos cadenas,  $n_x$  es el número de dígrafos en la cadena x, y  $n_y$  es el número de dígrafos en la cadena y. Por ejemplo, para calcular la similitud entre:

night

nacht

- Se procede primeramente con el **cálculo** de los dígrafos de cada palabra:

ni	,	ig	,	gh	,	ht
na	,	ac	,	ch	,	ht

- Cada conjunto tiene **4 elementos** y su **intersección** se reduce a un elemento  $n_t$ . Con la fórmula dada, se obtiene la similitud.

$$D_S = \frac{2 \cdot 1}{(4 + 4)} = 0.25$$





## Similitud de Dice <sub>(4)</sub>

- De manera análoga **al índice de Jaccard**, la medida de distancia se obtiene al **restarle a 1 el valor de la similitud**.
- Dado un valor del **coeficiente de Dice (D)**, es posible calcular el **índice de Jaccard (J)** y viceversa, mediante las siguientes ecuaciones.

$$D = \frac{2J}{(1 + J)}; \quad J = \frac{D}{(2 - D)}$$

- En cierta forma, se puede observar que el coeficiente de Dice le da **mayor peso** a los elementos comunes entre ambos conjuntos, lo que se puede apreciar al comparar los resultados de calcular la similitud por ambos métodos.



## Similitud de Dice <sub>(5)</sub>

- La versión **cuantitativa** de la similitud de Dice o del coeficiente de Sørensen, se conoce como el **porcentaje de similitud**; en donde se considera como un índice que está basado en **datos de abundancia** (y no de presencia/ausencia).

$$I_{S\_cuant} = \frac{2p_n}{a_n + b_n}$$

- En este caso  $p_n$  representa la **sumatoria** de la **abundancia** más **baja** de cada una de las **especies compartidas** entre ambos sitios.  $a_n$  y  $b_n$  es el número total de individuos en el sitio A y B respectivamente.

T.Bil	G.Spi	C.Tri	C.Sca	Medidas	Tipo
1	21	11	16	49	A
1	8	3	0	12	B
1	8	3	0	12	$p_n$

$$I_{S\_cuant} = \frac{2 \cdot 12}{49 + 12} = 0.3934426$$



## Similitud de Dice <sub>(6)</sub>

- Conclusiones

- El coeficiente de Dice o Índice de Sørensen se utiliza como medida de **similitud** para **analizar** datos del dominio de la ecología, geografía, biología, medicina entre otros. Este índice expresa el **grado** en que dos muestras son **semejantes** por las especies presentes en ellas.
- En el área de recuperación de información encuentra uso en la **lexicografía infográfica**, donde interviene directamente en la medición de la puntuación de asociación léxica de **dos palabras**, así como el **análisis de términos**, considerando a éstas en un **espacio vectorial**.
- La razón de este uso es más empírica que teórica, se puede justificar teóricamente como la intersección de dos conjuntos difusos.
- En comparación con la distancia euclidiana, esta métrica se ajusta bien para **conjuntos de datos heterogéneos** y da **menos peso** a los casos desviados.



## Similitud de Jaccard <sup>(1)</sup>

- A diferencia de otras métricas, el **índice** o **coeficiente de similitud de Jaccard** opera sobre conjuntos, por lo que comúnmente se utiliza para comparar **sentencias** o **párrafos** completos como un **conjunto de palabras**.
- Sin embargo también puede ser utilizado para comparar palabras considerándolas como conjuntos de **letras** o **caracteres**. Cualquiera que sea el nivel de **tokenización** sobre el que se utilice, es interesante notar que la posición que ocupa el elemento no tiene relevancia y que elementos repetidos son considerados como uno solo dentro del conjunto.
- El coeficiente de similitud de Jaccard mide la **similitud** entre dos conjuntos de muestras. Se define como la **relación** entre el tamaño de la intersección de ambos conjuntos y el tamaño de la unión y es una medida de la similitud entre ambos; es decir, es la división entre el número de elementos en común que tienen los dos conjuntos sobre el número de elementos únicos que tiene la unión de ambos conjuntos.

$$sim_j(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



## Similitud de Jaccard <sup>(2)</sup>

- Por su parte, la **distancia de Jaccard** es el resultado de restarle a **1** el valor de la similitud.

$$d_J(A, B) = 1 - sim_J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

- El índice de similitud de Jaccard, compara **individuos** (instancias) de dos conjuntos para ver individuos **compartidos** y **distintos**.
- Es una medida de similitud para los dos conjuntos de datos, con un rango de **[0 a 1]**. Cuanto mayor sea el valor que se acerca a 1, más similares serán las dos poblaciones.
- El índice Jaccard es una estadística que sirve para comparar y medir qué tan similares son dos conjuntos **diferentes** entre sí. Aunque es fácil de interpretar, es susceptible de tamaños de muestra pequeños.



## Similitud de Jaccard <sup>(3)</sup>

- Puede dar resultados erróneos, especialmente con **muestras** más pequeñas o conjuntos de datos con **observaciones faltantes**.
- Para medir el grado de similitud entre un documento y/o una consulta debemos llevar o extender esta ecuación, que está expresada en función de conjuntos de términos, a una expresión en función de **vectores de términos**.
- Esta forma extendida del coeficiente de Jaccard también se conoce con el nombre de **coeficiente de Tanimoto**.

$$sim_j(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



## Similitud de Jaccard <sub>(4)</sub>

- Expresada en términos vectoriales:  $sim_J(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| + |\vec{q}| - \vec{d}_j \cdot \vec{q}}$

- Sustituyendo los términos en los conjuntos  $X$  e  $Y$ , se tiene que:

$$sim_J(x, y) = \frac{\sum_{i=1}^p x_i y_i}{\sum_{i=1}^p (x_i)^2 + \sum_{i=1}^p (y_i)^2 - \sum_{i=1}^p x_i y_i}$$

- En resumen, en ciencias de la computación se utiliza para medir **distancia** entre **vectores** definidos sobre un espacio vectorial booleano (componentes del vector 0 o 1).

$$sim_J(A, B) = \frac{|A \wedge B|}{|A \vee B|}$$

- Donde:

- $\wedge$  y  $\vee$  son respectivamente las operaciones  $\times$  (AND) y  $+$  (OR) de la lógica booleana y  $|A| = \sum a_i$



## Similitud de Jaccard <sup>(5)</sup>

- Hay dos alternativas principales para encontrar la métrica de similitud en textos:
  - El enfoque *basado en caracteres* se ocupa de los caracteres individuales presentes en el documento con la secuencia adecuada.
  - El enfoque *basado en términos* se ocupa de la palabra completa.
    - Las palabras a menudo se simplifican o *lematizan* antes de realizar la prueba según el proceso de limpieza de datos inicial utilizado para el propósito específico.
    - Por tanto, métrica de *similitud de Jaccard* es *adecuada* para este enfoque.
    - La similitud de Jaccard puede aplicarse a *vectores* de *documentos* que ya están en formato de *bolsa de palabras*.
    - La definición de la *distancia* es uno menos el tamaño de la intersección sobre el tamaño de la unión de los vectores (en este caso la distancia puede ser alta).
    - La similitud de Jaccard permite aceptar listas pares (es decir, *documentos*) como entradas. Cuando son los mismos vectores, el valor devuelto es **0**; lo que significa que la distancia es 0 y los dos *documentos* son *idénticos*.





## Similitud de Jaccard <sup>(6)</sup>

- En otras áreas como las ciencias médico-biológicas, la similitud o índice de Jaccard se utiliza para medir la **similitud** entre muestras o poblaciones, utilizando la siguiente fórmula:

$$I_j = \frac{c}{a + b - c}$$

- Donde:
  - a = Número de especies o individuos presentes en la muestra A.
  - b = Número de especies o individuos presentes en la muestra B.
  - c = Número de especies presentes en ambas muestras.
- En este sentido, el valor de **0**, indica que las muestras no presentan especies en común y tiende a **1** a medida que aumenta el número de especies compartidas.



## Similitud de Jaccard <sub>(7)</sub>

- Ejemplos:

- Con base en la fórmula:  $sim_j(A, B) = \frac{|A \cap B|}{|A \cup B|}$

- Si dos conjuntos de datos **comparten** exactamente los mismos miembros, su índice de **similitud Jaccard** será **1**. Por el contrario, si **no tienen miembros** en común, su similitud será **0**.
- Ejemplos de cómo calcular el índice de similitud de Jaccard para datasets diferentes.
  - Supongamos que se tienen los siguientes dos conjuntos de datos:
  - **A = [0, 1, 2, 5, 6, 8, 9]; B = [0, 2, 3, 4, 5, 7, 9]**
  - Para calcular la similitud de Jaccard entre ellos, primero encontramos el **número total** de observaciones en ambos conjuntos, luego **dividimos** por el número total de observaciones en cualquiera de los conjuntos:
  - **Número de observaciones en ambos:** {0, 2, 5, 9} = 4
  - **Número total de observaciones:** {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} = 10
  - $sim_j(A, B) = \frac{4}{10} = 0.4$



## Similitud de Jaccard <sub>(8)</sub>

- Ejemplos:
  - Supongamos que se tienen los siguientes dos conjuntos de datos:
  - **$C = [0, 1, 2, 3, 4, 5]$ ;  $D = [6, 7, 8, 9, 10]$**
  - Para calcular la similitud de Jaccard entre ellos, primero encontramos el **número total** de observaciones en ambos conjuntos, luego **dividimos** por el número total de observaciones en cualquiera de los conjuntos:
  - **Número de observaciones en ambos:**  $\{\} = 0$
  - **Número total de observaciones:**  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} = 11$
  - $sim_j(A, B) = \frac{0}{11} = 0$
  - El índice de similitud de Jaccard resulta ser 0. Esto significa que los dos conjuntos de datos **no comparten** miembros comunes.



## Similitud de Jaccard <sub>(9)</sub>

- Ejemplos:
  - También podemos usar el índice de similitud de Jaccard para conjuntos de datos que contienen **caracteres** en lugar de números. Por ejemplo, supongamos que tenemos los siguientes conjuntos:
  - $E = \text{'gato', 'perro', 'hipopótamo', 'mono'}$
  - $F = \text{'mono', 'rinoceronte', 'avestruz', 'salmón'}$
  - Para calcular la similitud de Jaccard entre ellos, primero encontramos el **número total** de observaciones en ambos conjuntos, luego **dividimos** por el número total de observaciones en cualquiera de los conjuntos:
    - **Número de observaciones en ambos:**  $\{\text{'mono'}\} = 1$
    - **Número total de observaciones:**  $\{\text{'gato', 'perro', 'hipopótamo', 'mono', 'rinoceronte', 'avestruz', 'salmón'}\} = 7$
    - $sim_j(A, B) = \frac{1}{7} = 0.142857$
  - El índice de similitud de Jaccard resulta ser 0.142857 . Dado que este número es bastante bajo, indica que los dos conjuntos son **bastante diferentes**.



## Similitud de Jaccard <sub>(10)</sub>

- Ejemplos:
  - La distancia de Jaccard mide la diferencia entre dos conjuntos de datos y se calcula como:  $d_J(A, B) = 1 - sim_J(A, B)$
  - Esta medida nos proporciona una idea de la diferencia entre dos conjuntos de datos o la diferencia entre ellos.
  - Por ejemplo, si dos conjuntos de datos tienen una similitud de Jaccard del 80%, entonces tendrían una distancia de Jaccard de  $1 - 0.8 = 0.2$  o equivalente al 20%.



## Proyecto 6<sub>(1)</sub>

### **Proyecto 6. Calcular los índices de Sørensen y Jaccard para la tabla tipo de suelo adjunta**

#### **Instrucciones:**

- 1. Calcular los índices mencionados para las especies que abundan en el tipo de suelo “Aluvial” y “Lacustre”.**
- 2. Considerar para el cálculo de estos índices la abundancia, además de que no deben considerar aquellas especies que se encuentren presentes en una extensión menor a 20 metros cuadrados en ambos tipos de suelo.**