MINERIA DE DATOS

Introducción

Elizabeth León Guzmán

Research Group on Data Mining – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

2021

Agenda

Agenda

Conceptos

¿Qué es un dato?

Población

- Conjunto de estudio cuyos elementos tienen propiedades en común
- Conjunto bien definido (es posible identificar cuales elementos pertenecen al conjunto y cuales no)
  Ejemplo: Personas que viven en mi barrio

Características de la población

Se pueden medir. En estadística se definen Variables (características medibles)
Ejemplo: edad

Conceptos

¿Qué es un dato?

Población

Personas que viven en mi barrio

Características de la población (variable)

Ejemplo: edad

edad = 25, 15, 6, 75, 15, ...

Conceptos

¿Qué es un dato?

Variable

$edad(x) = ?$

- La variable tiene rango
- La variable aleatoria es una función
- La posibilidad de que se tome un x y x tome un valor y, se le llama EVENTO

Conceptos

¿Qué es un dato?

- Hecho individual acerca de algo de interés para alguién.
- Representación simbólica de una variable numérica o categórica

Ejemplos

Temperatura: 17, 28, 15
Ciudad: Bogotá, Cartagena
Fecha Julio 20 2015, Julio 20 2016

Conceptos

Generación de Datos

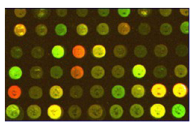
Source: Tan et al. (2005)

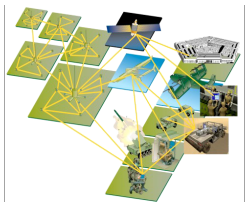
Comerciales

Web (e-commerce, e-learning)
Supermercados (compras)
Bancos (transacciones con tarjetas, web

Conceptos

Generación de Datos





Source: Tan et al. (2005)

Científicos

Satélites
Telescopios
Microarrays (información genética)
Simulaciones

Conceptos

¿Qué es información?

- Datos estructurados y relacionados. Almacenados generalmente en Bases de Datos
- Consultas para obtener información

ciudad	temperatura	fecha
Bogotá	15	July 20 2015
Cartagena	28	July 20 2015
Bogotá	17	July 20 2016

Información

Información

Algo peor que no tener información disponible,

es

¡tener mucha información y no saber qué hacer con ella!

Conocimiento

Si se entiende lo que significan los datos, Usarlos para la toma de decisiones

- Datos definen "Oportunidad"
- Analizar datos para encontrar **CONOCIMIENTO** $\rightarrow$ Toma de decisiones

Conocimiento

Data Mining - Minería de Datos

- Datos crecen exponencialmente.
- Inmensas Bases de datos contienen datos, que todavía no han sido "explotados" (¡Valiosa información!)
- Ej: Datos en Internet. Los usuarios esperan información más sofisticada

"Descubrir información oculta"

Minería de Datos

NO ES

- Buscar un número telefónico en un directorio
- Buscar en Google
- Generar histogramas de salarios por grupos de edades diferentes

Minería de Datos

ES

- Encontrar grupos de personas con similares hobbies
- ¿Hay mas probabilidad de desarrollar cáncer si se vive cerca de una planta eléctrica?

Minería de Datos - Ejemplo

NO ES

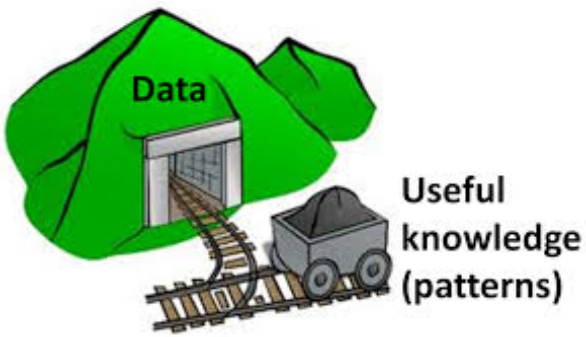
- Encontrar las personas que aplicaron a crédito con apellido Martínez
- Identificar los clientes que compraron mas de $1,000,000 en el último mes
- Encontrar todos los clientes que han comprado leche

Minería de Datos - Ejemplo

ES

- Encontrar las personas que aplicaron a crédito con **poco riesgo** de pago del crédito
- Identificar clientes con **tendencias similares** de compra
- Encontrar todos los artículos que son comprados **frecuentemente** con leche

Minería de Datos



Definición Fayyad est al 1996

Proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles de los datos

Definición

Proceso de aplicar metodologías basadas en computador para descubrir conocimiento de los datos

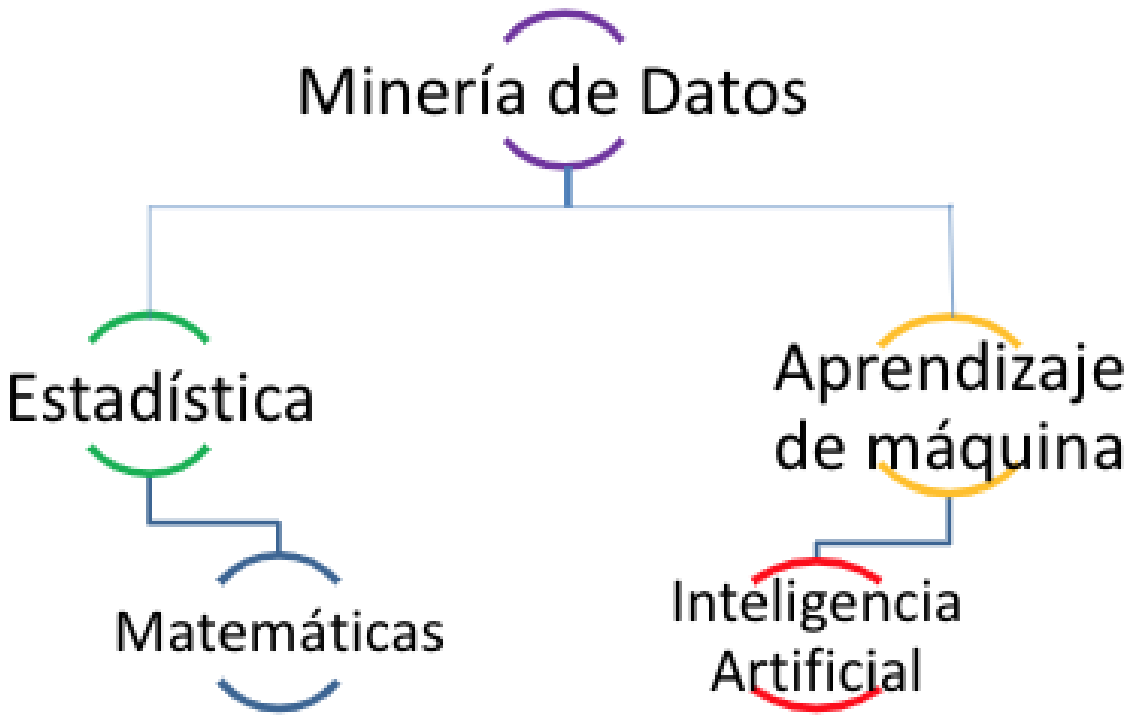
Origen Minería de Datos

Teoría de Control

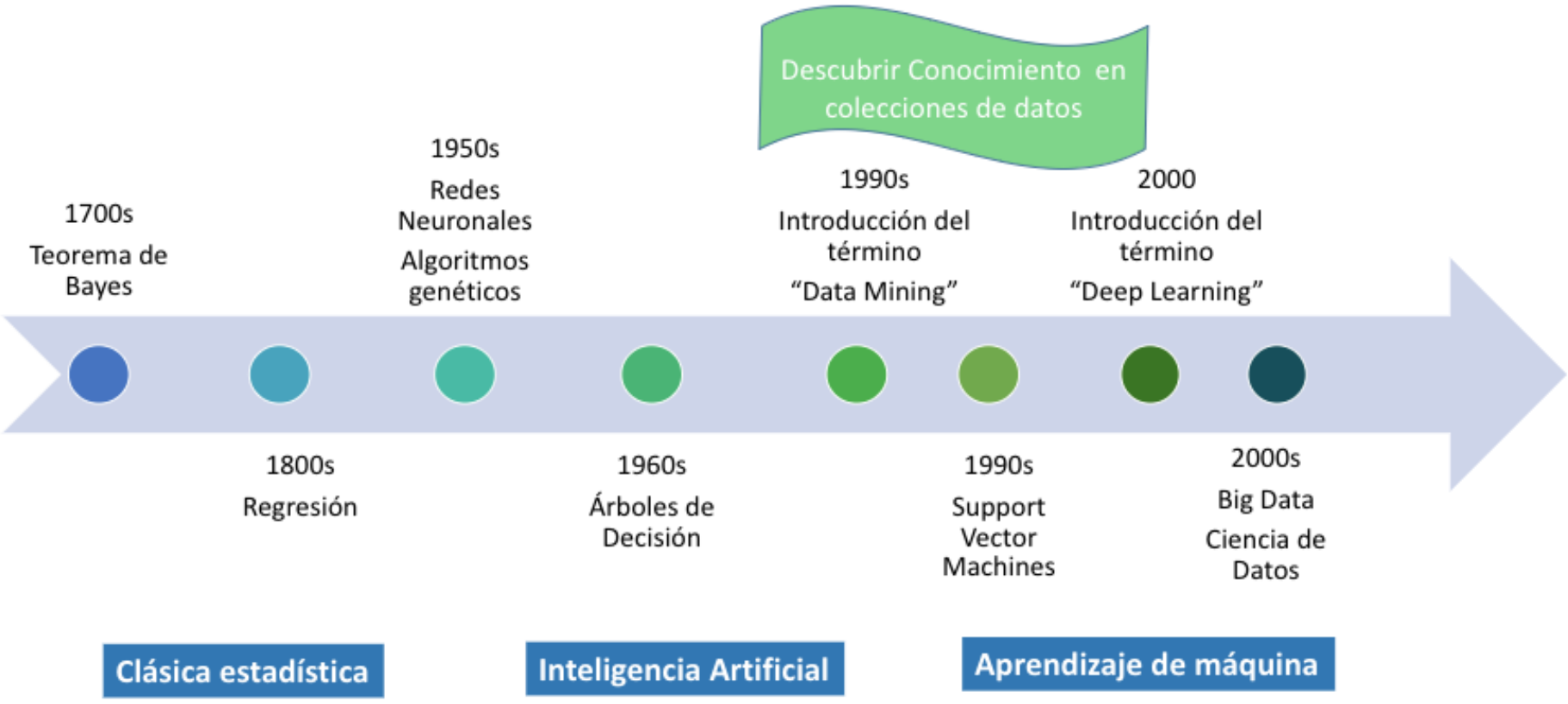
- Determinar modelo matemático de un sistema desconocido observando sus datos de entrada y salida:
  - Predecir el comportamiento del sistema
  - Explicar la interacción y relación entre las variables del sistema.
- Sistemas complejos para formalizar matemáticamente
- Crecimiento de los computadores ha generado grandes cantidades de datos. Datos son usados para generar modelos estimando las relaciones entre variables.

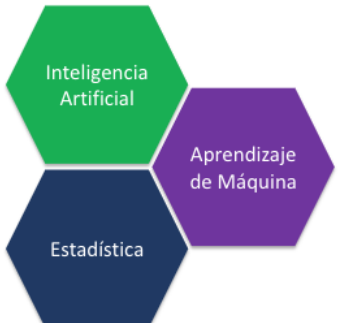
Origen Minería de Datos

- Raíces en "Data Analysis"
- Origen en varias disciplinas: estadística y aprendizaje de máquina.



Historia de la Minería de Datos





Conferencias

Auge en los años 90

SIAM International Conference on Data Mining

ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING

ECML PKDD - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases

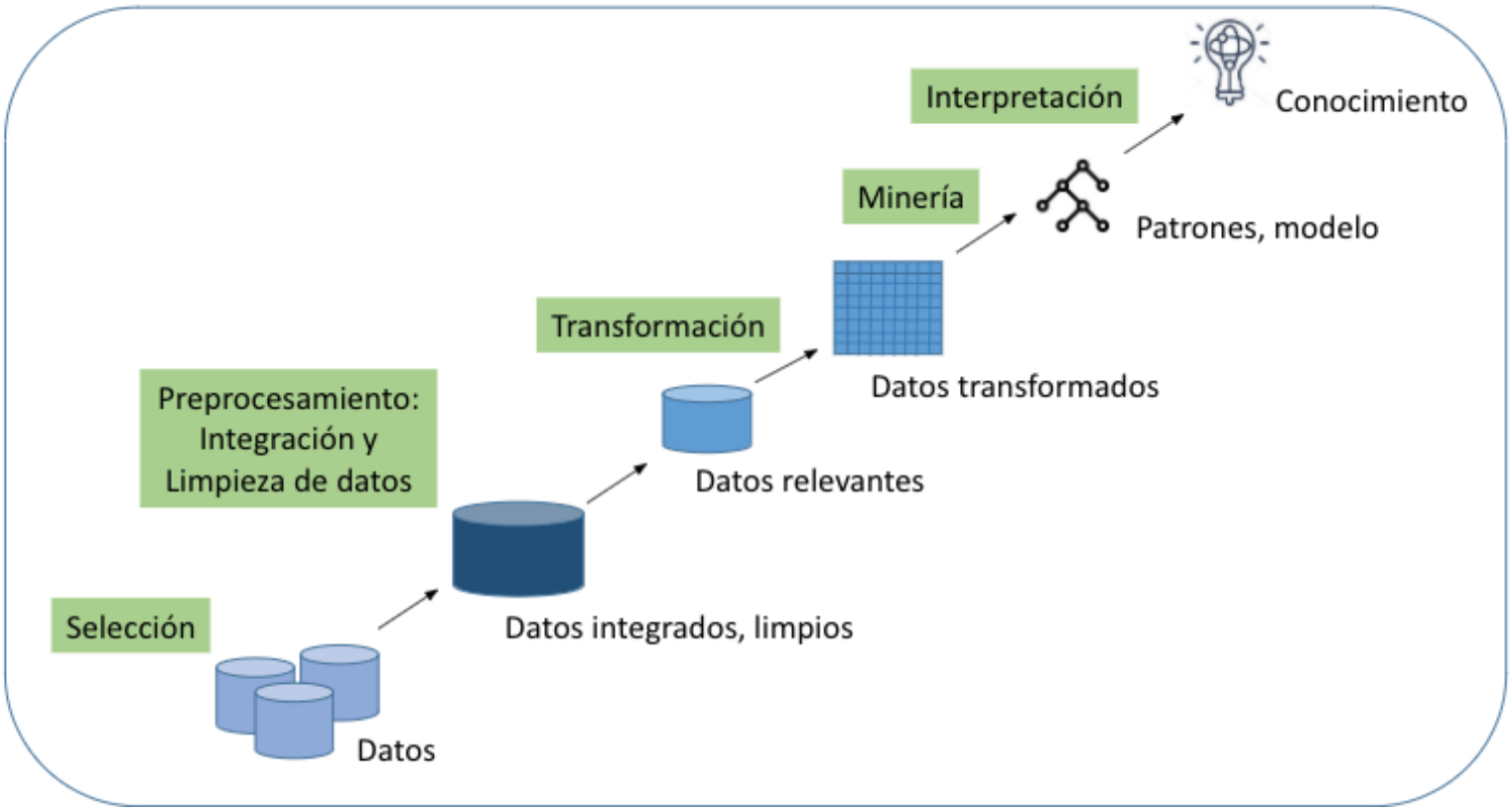
IEEE ICDM IEEE International Conference on Data Mining

PAKDD : Pacific-Asia Conference on Knowledge Discovery and Data Mining

Proceso KDD

- **Knowledge Discovery in Databases (KDD):** Descubrir información y conocimiento útil de grandes repositorios de datos (patrones, asociaciones, etc.)
- **Minería de Datos:** métodos inteligentes para extraer conocimiento "cavando por oro"

Proceso KDD



Source: Modificado de Dunham (2002)

Proceso KDD

- **Selección:** Obtener datos de varias fuentes
- **Preprocesamiento:** limpiar los datos
- **Transformación:** Convertir a formato común. Transformar a nevo formato
- **Minería:**Obtener resultados esperados
- **Interpretación/evaluación:**presentar resultados a usuario final de manera entendible

Ejemplo KDD - Análisis Web Log

u_ex150803.log - Notepad

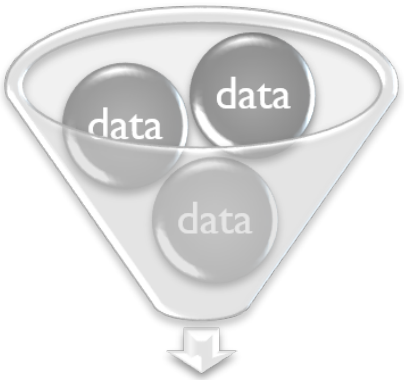
File Edit Format View Help

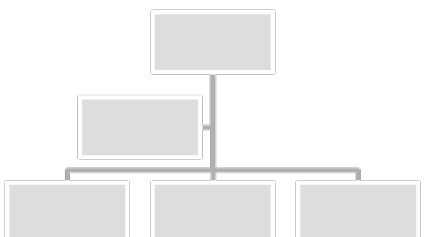
```
date time s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-
Agent) sc-status sc-substatus sc-win32-status time-taken
2015-08-03 12:40:57 209.133.7.95 GET /course-eligibility.asp - 80 - 115.118.114.159
Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu
+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0 1234
2015-08-03 12:40:58 209.133.7.95 GET /css/font-awesome.min.css - 80 -
115.118.114.159 Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like
+Gecko)+Ubuntu+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0
578
2015-08-03 12:40:58 209.133.7.95 GET /images/ftrlogo.png - 80 - 115.118.114.159
Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu
+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0 312
2015-08-03 12:40:58 209.133.7.95 GET /css/styles.css - 80 - 115.118.114.159
Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu
+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0 609
2015-08-03 12:40:58 209.133.7.95 GET /js/modernizr.custom.86080.js - 80 -
115.118.114.159 Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like
+Gecko)+Ubuntu+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0
281
2015-08-03 12:40:58 209.133.7.95 GET /css/bootstrap.min.css - 80 - 115.118.114.159
Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu
+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0 1171
2015-08-03 12:40:58 209.133.7.95 GET /js/bootstrap.min.js - 80 - 115.118.114.159
Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu
+Chromium/37.0.2062.120+Chrome/37.0.2062.120+Safari/537.36 200 0 0 593
```

Ejemplo KDD - Análisis Web Log

- **Selección:** seleccionar datos del log (fechas y lugares)
- **Preprocesamiento:**
  - Remover id de urls
  - Remover errores del log
- **Transformación:** Armar sesiones
- **Minería:** Identificar patrones en la estructura de los datos
- **Interpretación/evaluación:**
  - Visualizar las secuencias de navegación más frecuentes
  - Personalización, Recomendación

Minería de Datos





Data driven model

- Exploratory data analysis
- Data driven discovery
- Deductive learning

Construir modelo a partir de los datos

Aprendiendo de los Datos

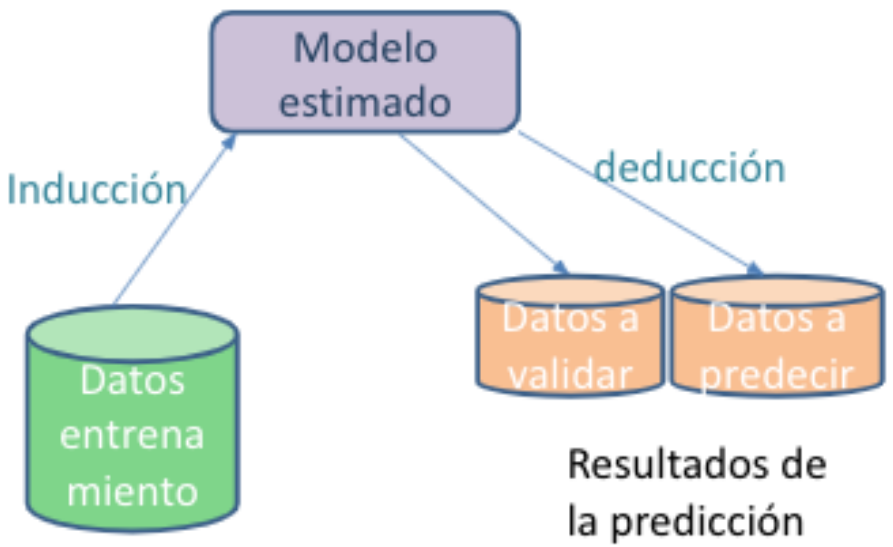
"las personas aprenden a través de la interacción de datos con el ambiente"

Aprendiendo de los Datos

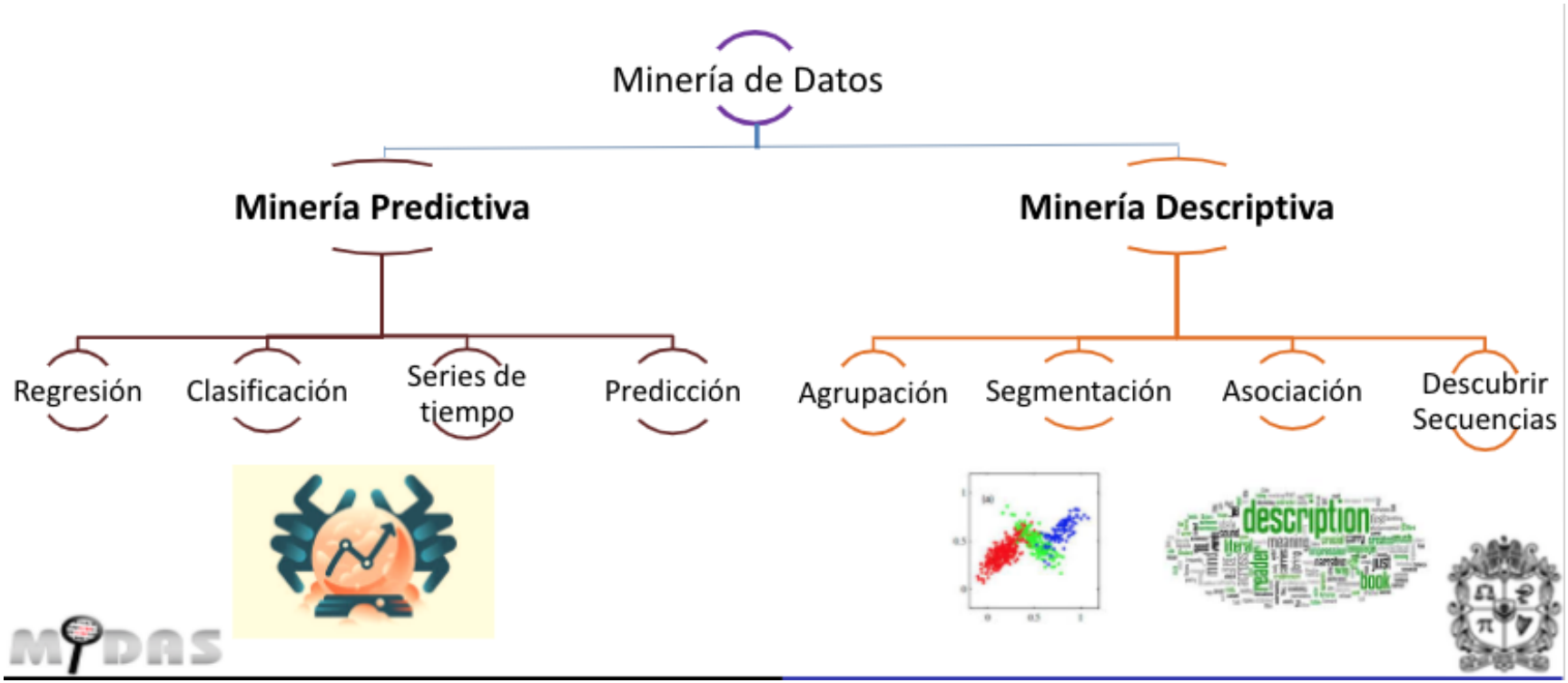
Aprendizaje Predictivo:

Aprendizaje por ejemplos de datos en dos pasos:

1. **Inducción**: Procesar de casos particulares (datos de entrenamiento) a un modelo general
2. **Deducción**: Dar valores de salidas de casos particulares teniendo en cuenta un modelo general

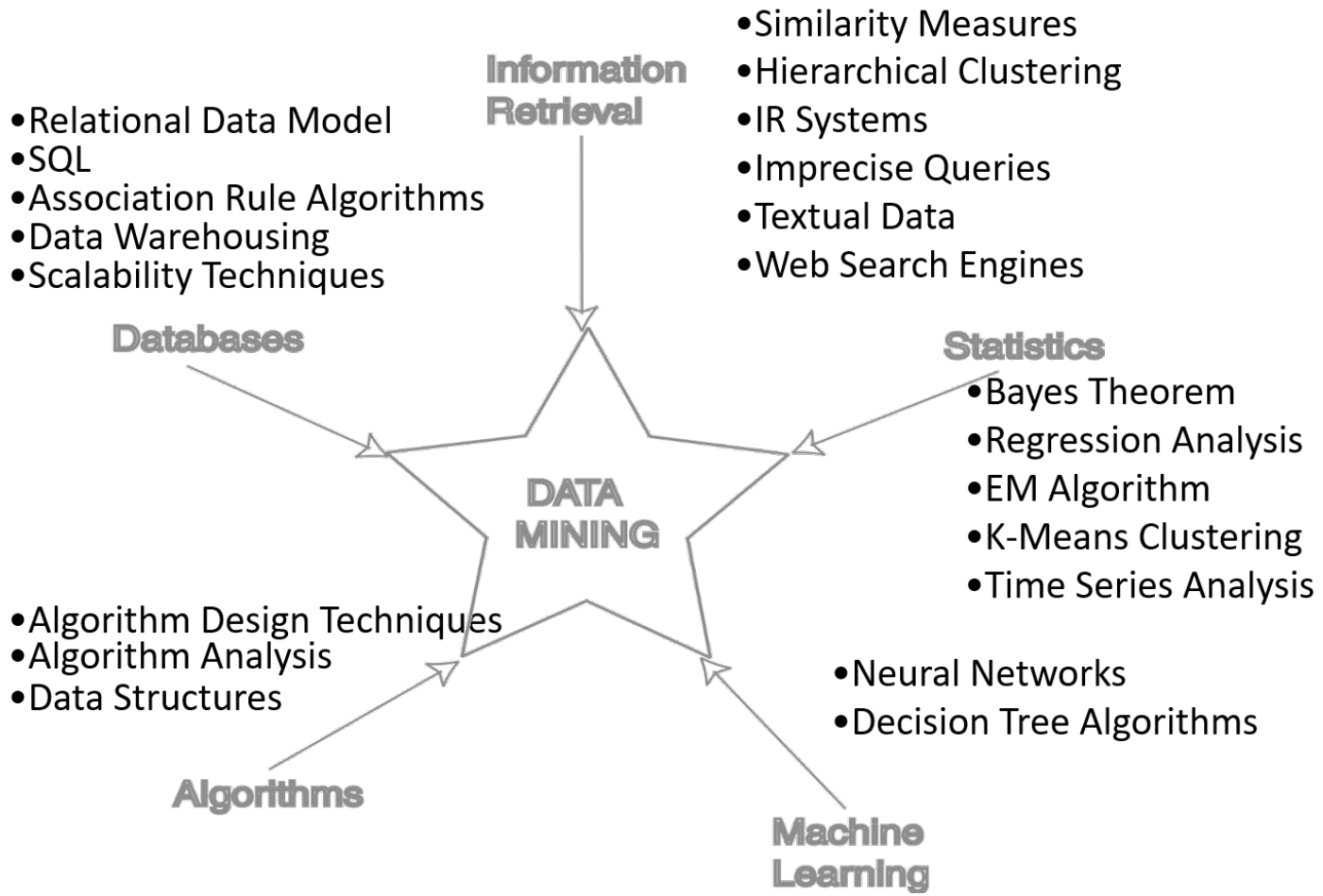


Modelos y Tareas de Minería de Datos



Source: Modificada de Dunham (2002)

Tecnología, Conocimiento para Minería de Datos



Source: Dunham (2002)

Industria 4.0

[https://www.laprensagrafica.com/__export/1508349926661/sites/prensagrafica/img/2017/10/18/innovacixn_empresarial.jpg]

Ciencia de Datos

Continuación de algunos campos de análisis de datos como la estadística, minería de datos, aprendizaje automático y analítica predictiva.

"Un concepto para unificar estadísticas, análisis de datos, aprendizaje automático y sus métodos relacionados para comprender y analizar los fenómenos reales"

Big data

- Recopilar, procesar, extraer de diversas fuentes de datos
- Limpiar, detección ruido, calidad (valores perdidos, etc)
- Preprocesar (reducción dimensionalidad, sampling, etc)
- Modelar (minería, aprendizaje maquinal, estadística)
- Visualizar

En la actualidad

Retos que trae Big Data para Minería de Datos: Volumen, variedad, velocidad, etc.

- Modelos dinámicos, stream (flujos de datos)
- Modelos escalables
- Modelos con fusión de datos (video, texto, imágenes, etc.)
- Modelos con rta en tiempo real
- Modelos usando aprendizaje profundo (Deep learning)

Referencias I

Dunham, M. H. (2002). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining, (first edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.