



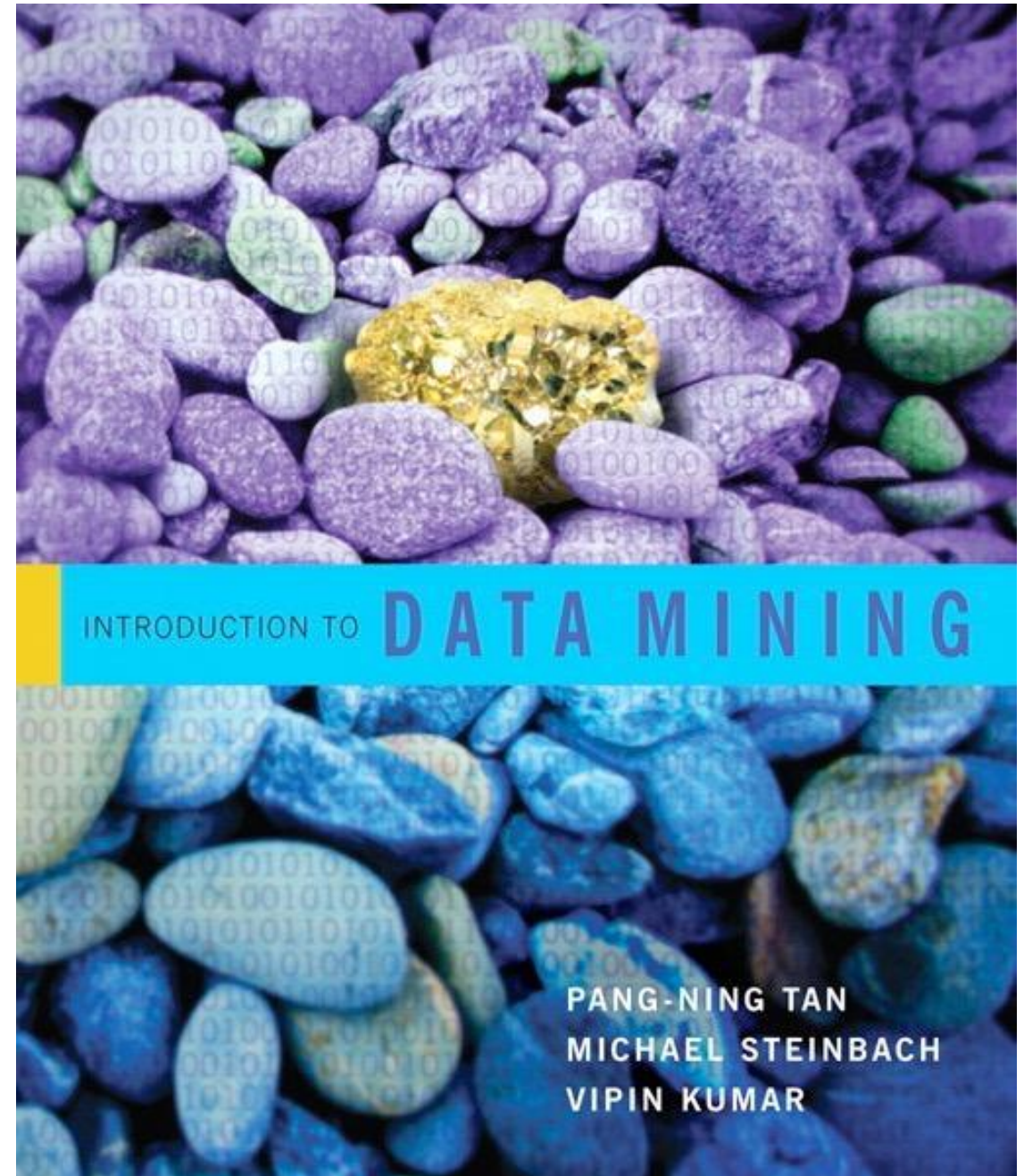
# Introducción a la Minería de Datos

Profesores: Felipe Bravo y  
Hernán Sarmiento

Basado en las slides de Bárbara Poblete

# Libro del Curso

- Introduction to Data Mining
- Autores: Pang-Ning Tan, Michael Steinbach, Vipin Kumar

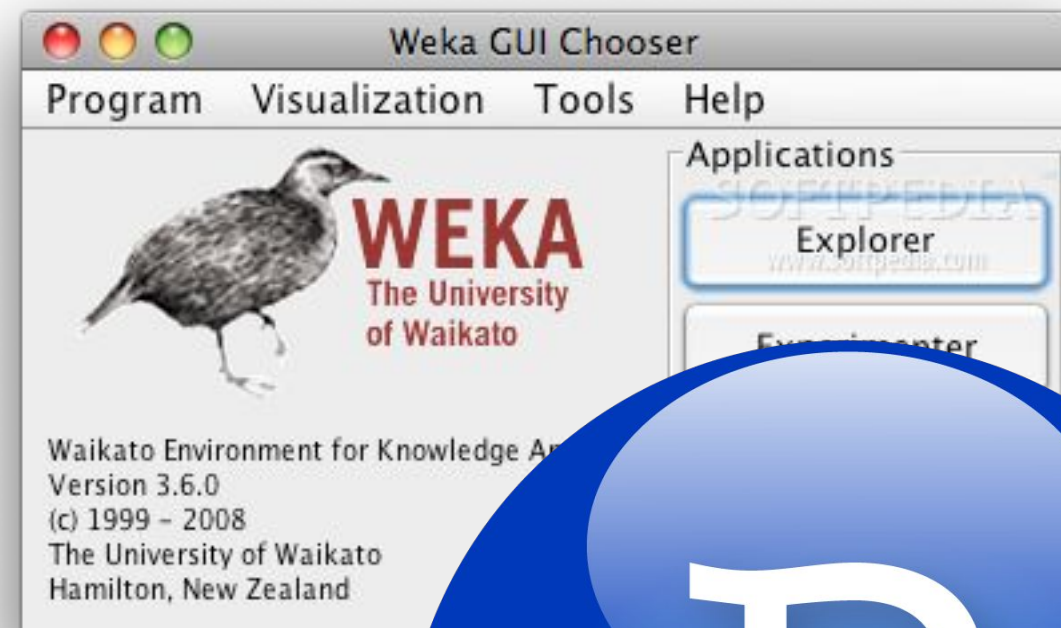


# Herramientas del curso

- ~~WEKA~~

- R (R Studio)

- Python



python™



# Objetivos del curso

- **Curso introductorio**
- Aprender a aplicar el proceso de DM a datos reales
- Conocer, seleccionar y utilizar las técnicas básicas de DM
- Aprender a interpretar los resultados de estos procesos
- Proveer la base para adquirir conocimiento más avanzado

# ¿Qué significa **Minar**?

Según la RAE:

*“Hacer grandes diligencias para conseguir algo”*



## EL PROCESO DE PRODUCCIÓN PASO A PASO

### 1 EXPLORACIÓN

Consiste en ubicar zonas donde exista la presencia de minerales cuya explotación sea económicamente rentable. Se recogen muestras (rocas) del suelo para conocer los elementos y minerales que las conforman.

**PERFORACIÓN.** Si los análisis de las muestras dan resultados positivos se procede con la perforación: se sacan muestras de diferentes profundidades (testigos) para determinar tipo, cantidad, profundidad y otras características del mineral.

### 2 MINADO

Consiste en la extracción y transporte del material que contiene oro y plata desde el tajo hasta las pilas de lixiviación (PAD).

Se perfora el terreno para colocar los explosivos y fragmentar el suelo para el carguo.

### 3 CARGUO Y ACARREO

Camiones gigantes llevan el mineral extraído del tajo a la pila de lixiviación (PAD) acondicionada previamente.

### 4 LIXIVIACIÓN

El mineral descargado en las pilas de lixiviación es regado con solución cianurada para recuperar el oro y la plata. La solución rica (cargada con oro y plata) es conducida hacia las pozas de procesos a través de tuberías colectoras.

### PLANTA DE TRATAMIENTO DE AGUAS ÁCIDAS

La solución cianurada se vierte en el PAD a través de un sistema de riego por goteo. Se utiliza, en promedio, una solución con 50 partes por millón; es decir, 50 g de cianuro por cada 1.000 litros de agua.

**TRATAMIENTO DE AGUAS DE EXCESO.** Antes de ser devuelta al medio ambiente, el agua pasa por una planta de tratamiento de aguas de procesos donde se le aplica tratamiento de Osmosis Inversa.

### 6 REFINERÍA (Fundición)

El oro obtenido en el proceso Merrill Crowe es sometido a operaciones de secado en hornos de retortas a 650°C. Finalmente, el producto obtenido pasa por un proceso de fundición en horno de arco eléctrico a 1.200°C para obtener el doré, que es el producto final.

# EL PROCESO DEL ORO DE PRINCIPIO A FIN

UNA VEZ DESCUBIERTA LA ZONA MINERALIZADA, EL ÁREA DE GEOLOGÍA DE YANACOCCHA REALIZA ESTUDIOS MÁS DETALLADOS DE LA ZONA QUE PERMITEN IDENTIFICAR CANTIDADES PRECISAS DE MINERAL. EN 1990 SE LLEVARON A CABO LOS ESTUDIOS DE FACTIBILIDAD PARA INICIAR LOS TRABAJOS EN UNA PLANTA PILOTO PARA LA LIXIVIACIÓN EN PILAS. CON EL INICIO DE LAS OPERACIONES DE CARACHUGO, EL 7 DE AGOSTO DE 1993, LA EMPRESA PRODUJO SU PRIMERA BARRA DORÉ. PARA EXPLOTAR Y OBTENER EL ORO UTILIZA EL MÉTODO DE MINERÍA A TAJO ABIERTO O A CIELO ABIERTO.

2006

Yanacocha fue reconocida como la primera productora de oro a nivel mundial durante ese año.

Más de

12 millones

de onzas de oro se produjeron entre el 2003 y el 2007.

Casi US\$ 400 millones de inversión ambiental desde el inicio de las operaciones de Yanacocha.

### DEPÓSITOS DE DESMONTE.

Son las estructuras donde se acumula el material extraído del tajo con bajo porcentaje de metal y que no sirve para lixiviar.

Todos los camiones y las palas están controlados a través de un sistema computarizado que permite conocer por satélite su ubicación exacta en todo momento.

Pozos de control de aguas subterráneas.

**Geomembrana.** Es una cubierta plástica de alta resistencia, ubicada en la base del PAD, que impide el contacto de los químicos con el suelo, cuidando la calidad del agua.

La mayor parte del agua utilizada en el proceso es reincorporada al circuito cerrado donde se vuelve a reponer el contenido de cianuro en la solución.

3,3 millones

de onzas ha sido el récord absoluto de producción durante un año. Esta cifra se consiguió en el 2005.

### 5 PROCESO MERRILL CROWE

Un proceso al que es sometida la solución rica en oro y plata. Primero es filtrada y limpiada; luego se elimina el oxígeno para finalmente añadir polvo de zinc para precipitar el metal y hacerlo sólido.

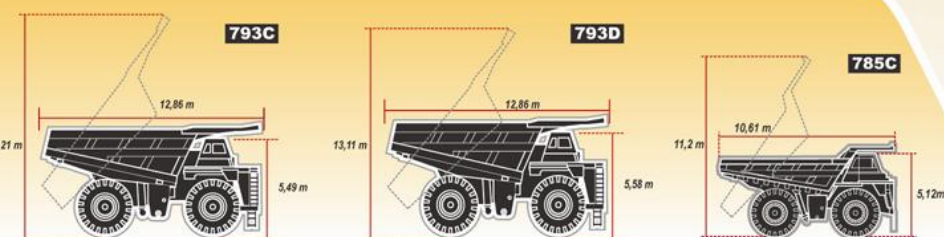
En esta etapa se procede a rehabilitar las áreas donde se realizó la actividad minera. Tiene por objetivo devolver a las áreas trabajadas condiciones similares o mejores a las que tenían antes de iniciar las operaciones. El paisaje, los cursos de agua o vegetación, por ejemplo, se reincorporan al entorno original.

## UN VISTAZO A LA MAQUINARIA

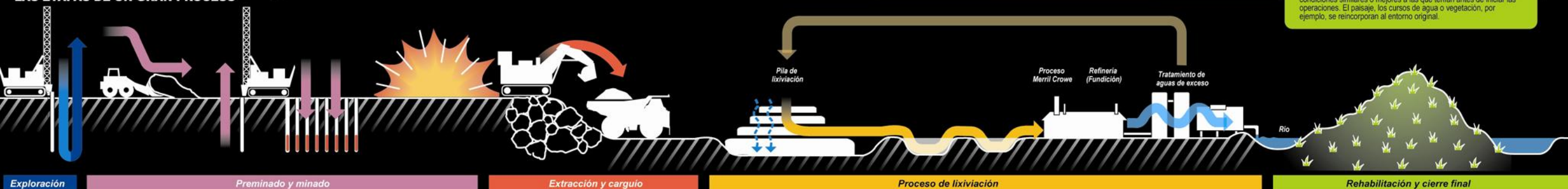
793C

793D

785C



## LAS ETAPAS DE UN GRAN PROCESO



Exploración

Preminado y minado

Extracción y carguo

Proceso de lixiviación

Rehabilitación y cierre final



# ¿Qué es la Minería de Datos?

- Descubrir automáticamente información útil en grandes repositorios de datos

# ¿Qué es la Minería de Datos?

- Descubrir **automáticamente** información útil en grandes repositorios de datos



# ¿Qué es la Minería de Datos?

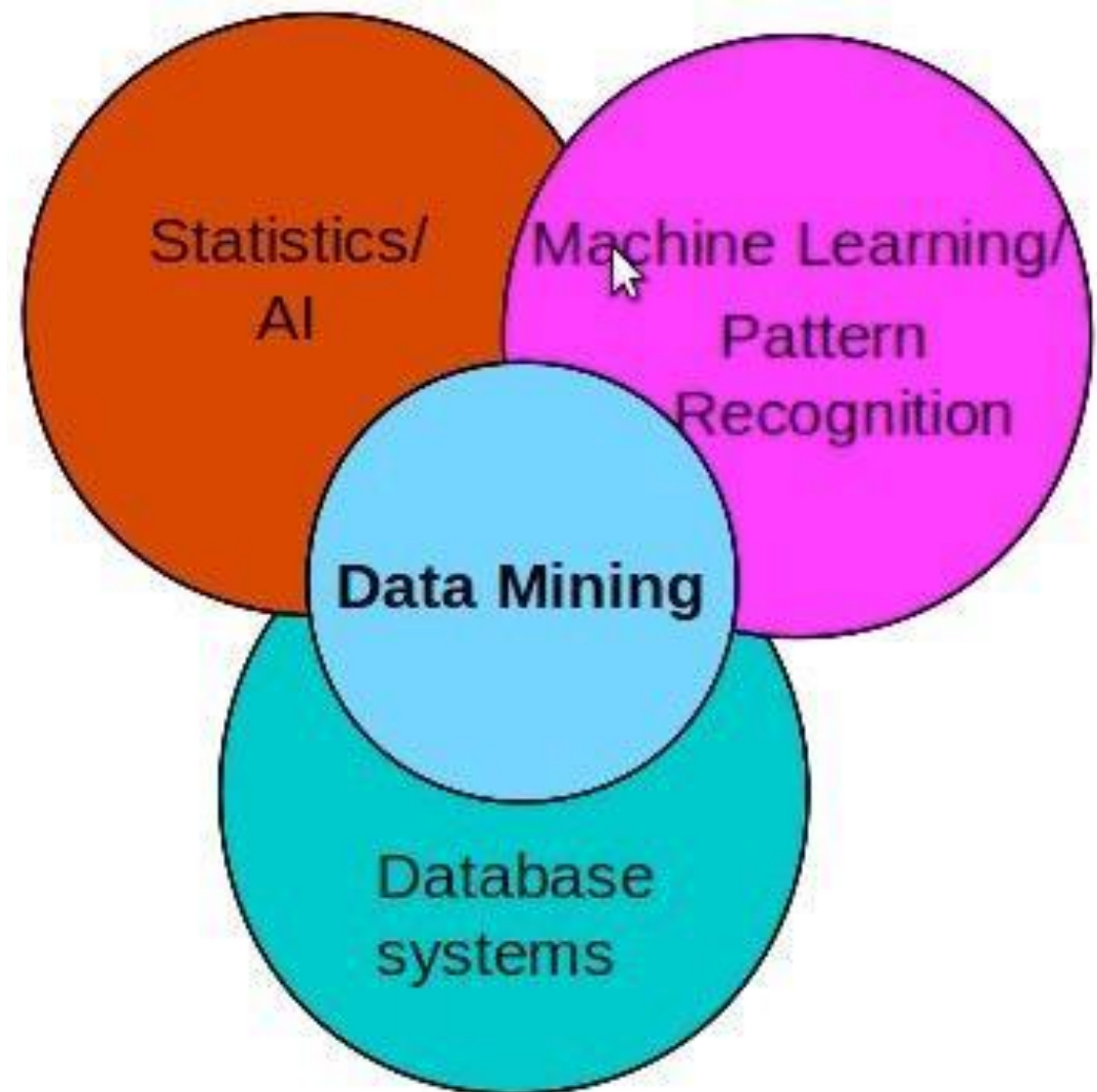
- Descubrir automáticamente información **útil** en grandes repositorios de datos

# ¿Qué es la Minería de Datos?

- Descubrir automáticamente información útil en **grandes repositorios** de datos

# Orígenes de la MD

- Une ideas de ML/AI, reconocimiento de patrones, estadística y BD
- Enfoques tradicionales fallan con datos masivos (alta dim., datos heterogéneos y distribuidos)





# entre Data Science, Machine Learning e

- Están de moda, pero no son lo mismo, ni son intercambiables

## Inteligencia Artificial?

- **Data Science** es el nombre reciente para algo mucho más antiguo: **Data Mining (90's)**
- Definición (sobre) simplista:
  - **Data mining** genera **entendimiento**.
  - **Machine learning** genera **predicciones**.
  - **Artificial intelligence** genera **acciones**.

# entre Data Science, Machine Learning e

- **Artificial Intelligence:** auto reconoce una señal de STOP y toma la acción de frenar.
- **Machine Learning:** auto reconoce señales de STOP usando cámaras y **predice** en base a un entrenamiento cuando debe parar.
- **Data Mining:** auto transita por las calles y nos damos cuenta que su rendimiento no es el esperado. Luego, **entendemos** que esto se debe a varios factores externos.



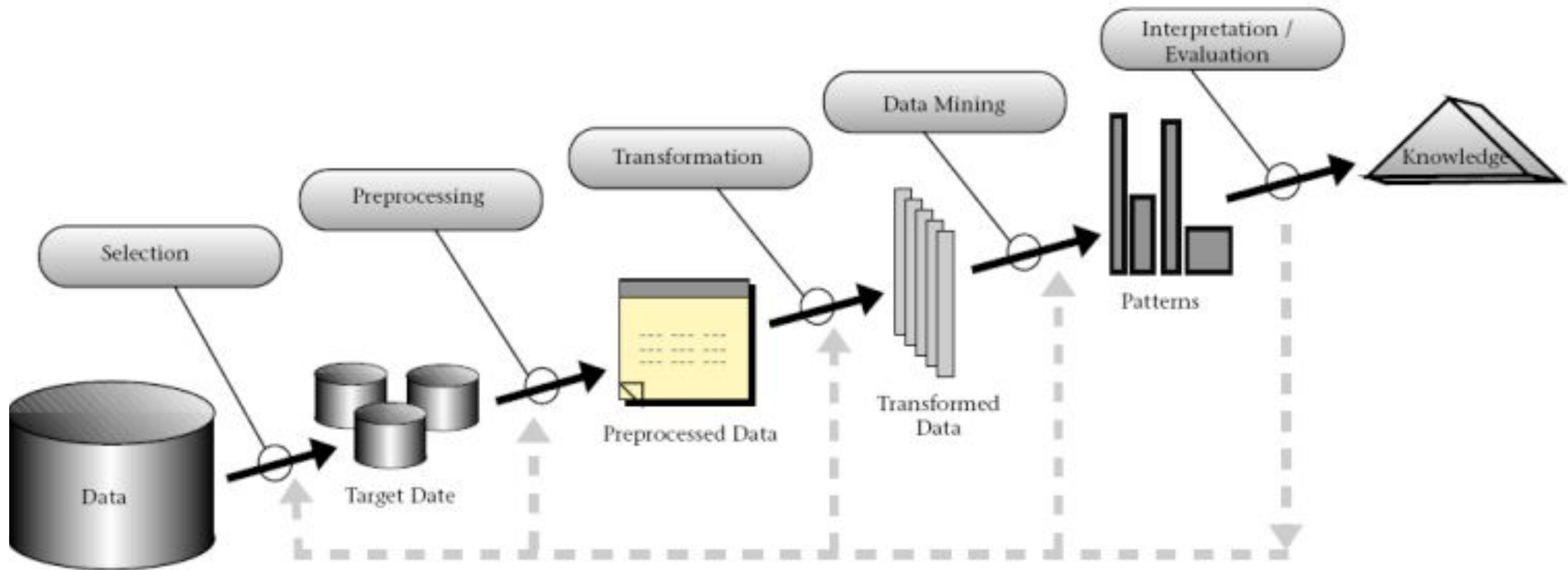
# ¿Cuál es la diferencia entre Data Science, Machine Learning e Inteligencia Artificial?

- Las definiciones tampoco sirven para describir el trabajo de alguien:
- "*Yo soy Data Scientist*" no depende de lo que uno haga, sino que de experiencia que se tenga y enfoque principal del trabajo que se hace.
- El hecho que alguien escriba no lo convierte en escritor.



# ¿Por qué es importante entender estas diferencias?

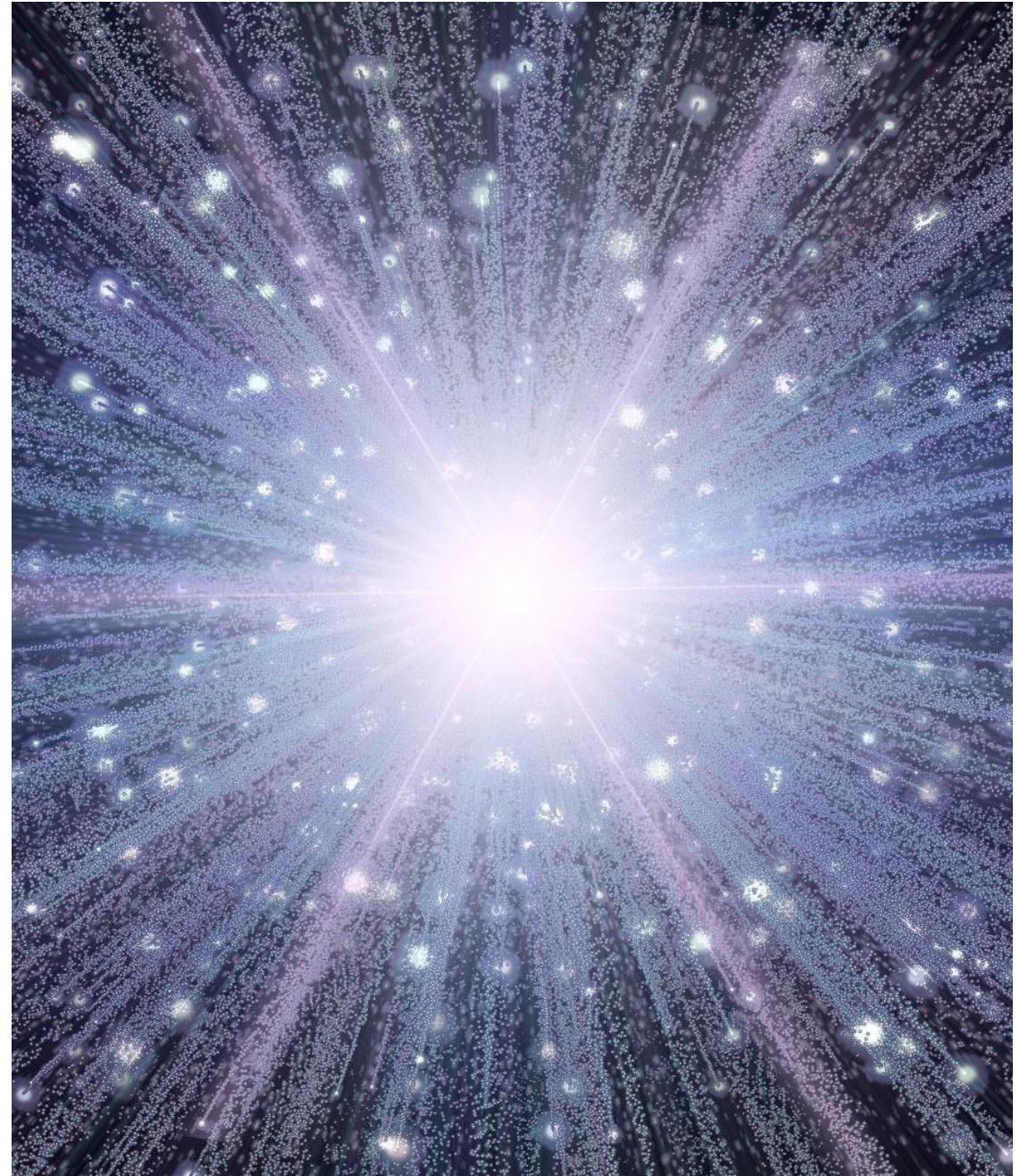
- Porque este **no es un curso de Machine Learning**, es un **curso de Minería de Datos**.
- **ML: Estudio, diseño y desarrollo de algoritmos** que permiten a los computadores aprender sin ser explícitamente programados (Arthur Samuel). Técnicas genéricas, aplicables a varios dominios.
- **Minería de Datos:** El enfoque está en **extraer conocimiento**, o patrones previamente desconocidos, a partir de (grandes) volúmenes de datos (en su mayoría no estructurados). Para esto se pueden utilizar técnicas de ML, entre otras. Requiere conocimiento de los datos mismos y su dominio.



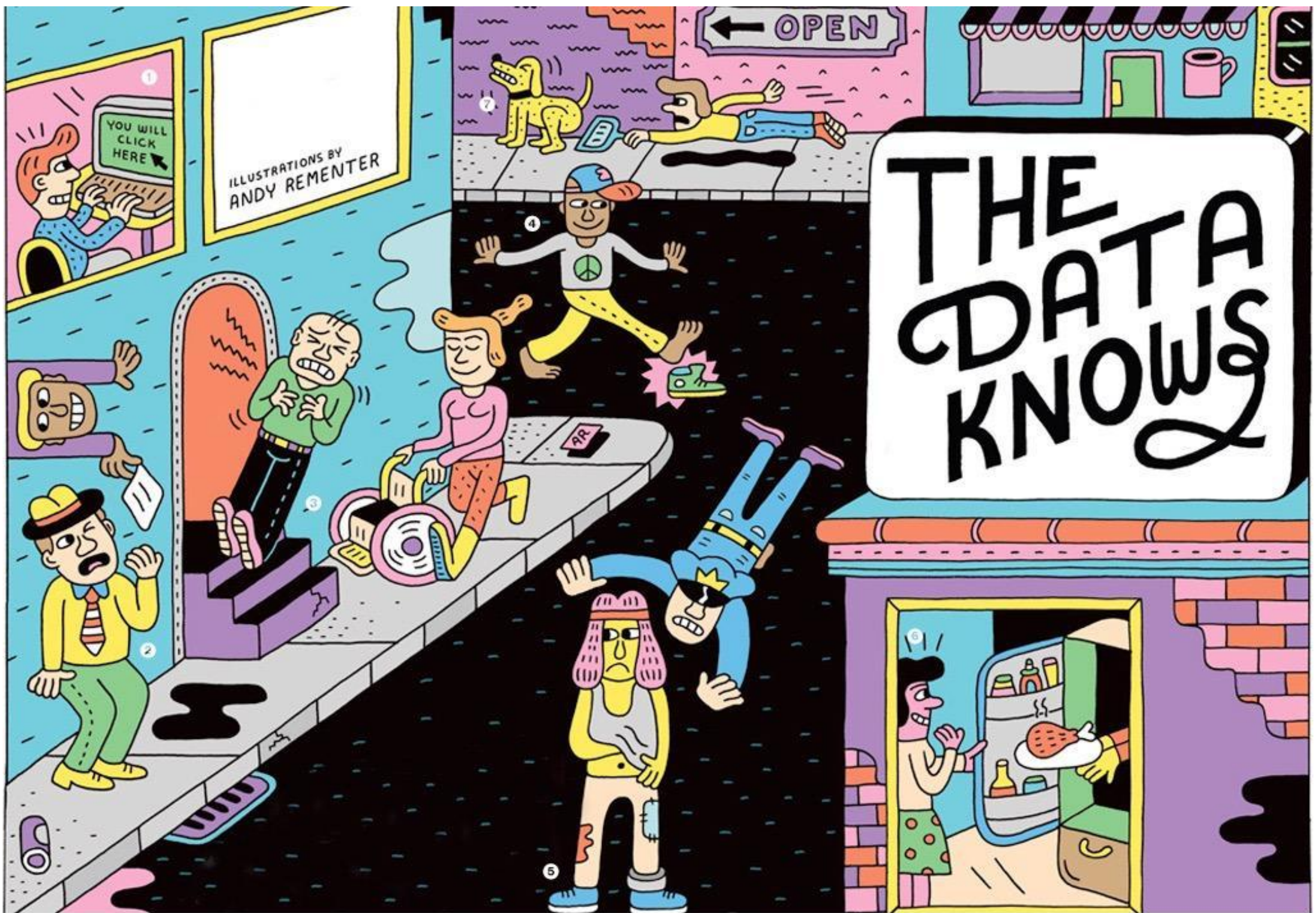
## Knowledge Discovery in Databases (KDD)

# BIG BANG

- 2006 Hadoop
- Análisis de datos masivos al alcance de todos (cientos de start-ups)









# ¿Por qué hacer minería de datos?

- Aspecto comercial
- Aspecto científico



# ¿Por qué hacer minería de datos?

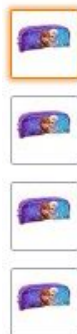
¿Motivación Comercial?

- Recolección de MUCHOS datos comerciales:
- Datos Web, e-commerce
- Compras en tiendas
- Transacciones en Bancos/ Tarjetas de Crédito









## 1 X Disney Frozen Pencil Case

by Innovative Designs, LLC

★★★★★ 4 customer reviews

Price: **\$5.30**

**In Stock.**

This item ships to **Santiago, Chile**. Want it **Friday, March 11**? Order within **9 hrs 49 mins** and choose **Amazon Global Priority Shipping** at checkout. [Learn more](#)

Sold by **JACOB'S** and Fulfilled by **Amazon**. Gift-wrap available.

Package Quantity: **1**

Style Name: **Purple**

- 1 Disney Frozen Pencil Case

15 new from **\$1.50**



Roll over image to zoom in

### Frequently Bought Together



Total price: **\$22.55**

Add both to Cart

Add both to List

- ✓ **This item:** 1 X Disney Frozen Pencil Case **\$5.30**
- ✓ Thermos 12 Ounce Funtainer Bottle, Frozen Purple **\$17.25**

### Customers Who Bought This Item Also Bought

Page



Disney Frozen Light Blue Stationery Set Pack with Case (13 Pcs)

★★★★★ 39

\$7.40 ✓Prime



Disney Frozen Rolling 16" Backpack and Lunch Bag - Lunchbox 2pc

★★★★★ 12

\$49.95 ✓Prime



Disney Frozen 1 Subject Wide Ruled Notebook - (Colors/Graphics Vary)

★★★★★ 14

\$4.67



Disney Frozen Elsa and Anna Kids Stationery Set (17 Pcs)

★★★★★ 19

\$8.95 ✓Prime



American Greetings Frozen Party Accessories, Pencils, 12 Count

★★★★★ 89

\$5.26 ✓Prime



Disney Frozen Hot Pink Elsa Anna and Olaf Stationery Set Pack with Case (13 Pcs)

★★★★★ 34



Thermos 12 Ounce Funtainer Bottle, Frozen Purple

★★★★★ 4

\$17.25 ✓Prime



## TV Thrillers &amp; Mysteries



## Romantic Movies



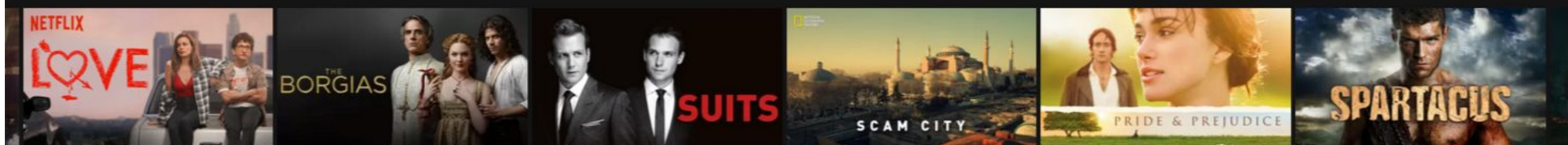
## Continue Watching for Barbara



## Watch It Again



## Top Picks for Barbara





Top Picks for Barbara



House, M.D.



★★★★★ 2014 TV-14 9 Seasons  
An awkward forensic anthropologist. An arrogant FBI agent. Together, they find justice in the dead.



★★★★★ 2010 TV-14 3 Seasons  
His deception detection is second to none. But his social skills? Well, they could use a little work.



★★★★★ 2015 18 11 Seasons  
Neither their patients' problems nor their own relationships are black-and-white. It's all shades of grey.



★★★★★ 2013 14 1 Season  
Elite FBI profilers play minds games to catch serial killers. Getting into murderers' heads can also get into yours.



★★★★★ 2015 TV-14 3 Seasons  
The legendary detective needs a doctor to keep him clean -- and maybe help round up a few murderers.

OVERVIEW

EPISODES

MORE LIKE THIS

DETAILS

Because you watched Cooked

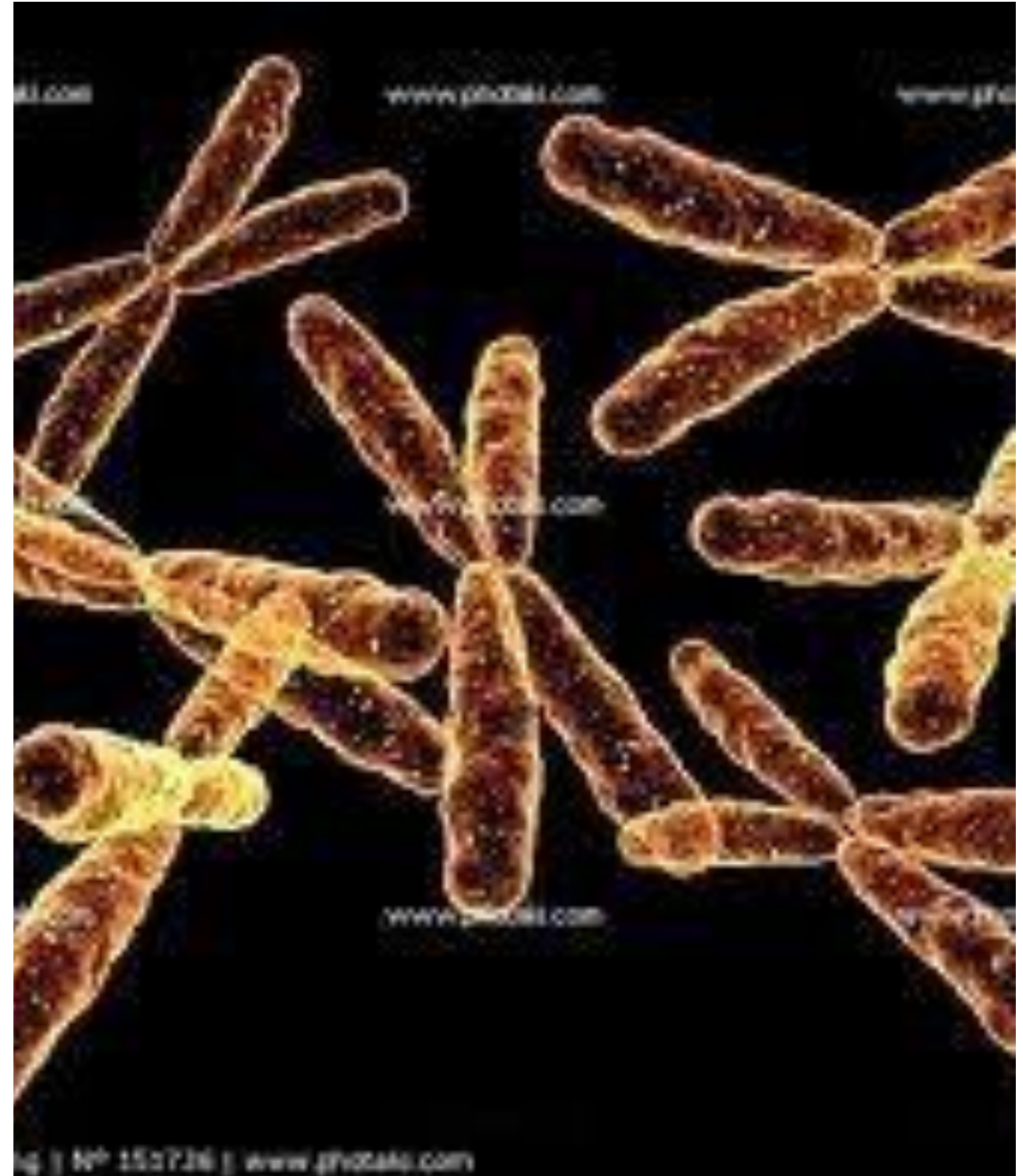




# ¿Por qué hacer minería de datos?

## ¿Motivación Científica?

- Datos (observaciones) recolectadas a gran velocidad (GB/hr, Tb/día)
- Telescopios, Satélites, Requerimientos Web, ADN, etc ([Google Flu Trends](#))



[Google.org home](#)

## Flu Trends

Select country ▼

[Home](#)

[How does this work?](#)

[FAQ](#)

### Flu activity



## Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)



[Download world flu activity data](#)



# Métodos utilizados en DM

- **Métodos predictivos:** Usar variables para predecir variables desconocidas o valores futuros de otras variables
- **Métodos descriptivos:** Encontrar patrones interpretables por humanos que permitan describir los datos



# Métodos utilizados en DM

- **Clasificación (Predictivo)**
- **Clustering (Descriptivo)**
- **Descubrimiento de Reglas de Asociación (Descriptivo)**
- **Descubrimiento de Patrones Secuenciales (Descriptivo)**
- **Regresión (Predictivo)**
- **Detección de Desviación (Predictivo)**



# Clasificación

- Set de Entrenamiento (atributos incluyendo clase)
- Busca modelar en atributo clase
- Objetivo: asignar la clase más correcta a records nuevos
- Set de Evaluación

*categorical*  
*categorical*  
*continuous*  
*class*

| <i>Tid</i> | Refund | Marital Status | Taxable Income | Cheat |
|------------|--------|----------------|----------------|-------|
| 1          | Yes    | Single         | 125K           | No    |
| 2          | No     | Married        | 100K           | No    |
| 3          | No     | Single         | 70K            | No    |
| 4          | Yes    | Married        | 120K           | No    |
| 5          | No     | Divorced       | 95K            | Yes   |
| 6          | No     | Married        | 60K            | No    |
| 7          | Yes    | Divorced       | 220K           | No    |
| 8          | No     | Single         | 85K            | Yes   |
| 9          | No     | Married        | 75K            | No    |
| 10         | No     | Single         | 90K            | Yes   |

| Refund | Marital Status | Taxable Income | Cheat |
|--------|----------------|----------------|-------|
| No     | Single         | 75K            | ?     |
| Yes    | Married        | 50K            | ?     |
| No     | Married        | 150K           | ?     |
| Yes    | Divorced       | 90K            | ?     |
| No     | Single         | 40K            | ?     |
| No     | Married        | 80K            | ?     |



# Clasificación: Aplicación 1

- Marketing directo
- Meta: Reducir costos de publicidad apuntando directamente a potenciales compradores.
- ¿Cómo?

# Clasificación:

## Aplicación 2

- Detección de Fraude
- Meta: Predecir transacciones fraudulentas en el uso de tarjetas de crédito
- ¿Cómo?



# Clasificación: Aplicación 3

- Fidelidad de Clientes
- Meta: Predecir si es posible perder a un cliente a la competencia
- ¿Cómo?

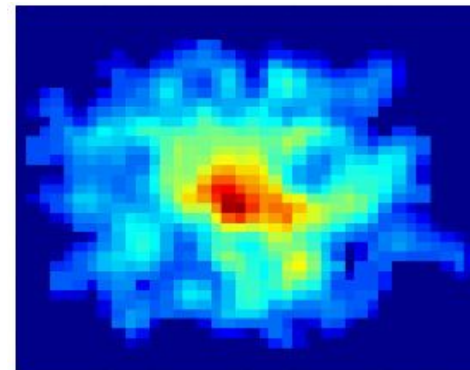
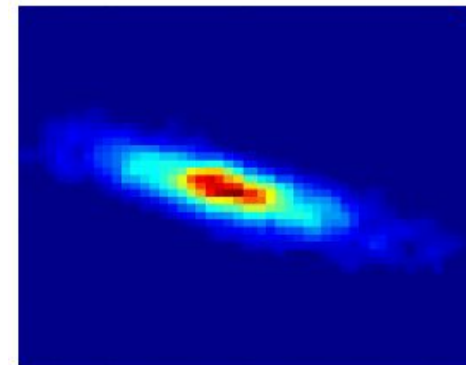
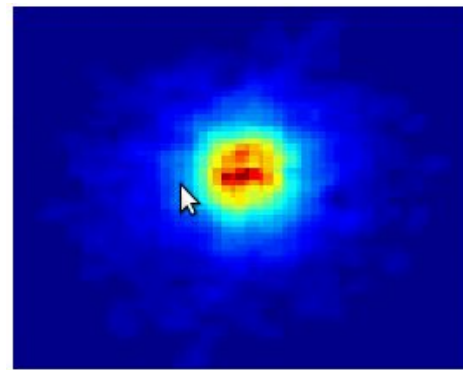
# Clasificación:

## Aplicación 4

- Catalogación de exploración del espacio
- Meta: Predecir la clase (estrella o galaxia) de objetos en el espacio, en especial de objetos poco visibles, basándose en exploración de telescopios (3000 imágenes de 23.040x23.040 píxeles por imagen, del observatorio Palomar)
- ¿Cómo?

# Clasificando galaxias

- Tamaño de los datos:
  - 72M de estrellas, 20M de galaxias
  - Catalogo de objetos: 9GB
  - BD de imágenes: 150GB
- Clase:
  - Etapas de formación (temprana, intermedia, tardía)
- Atributos:
  - Características de las imágenes
  - Características de las ondas de luz, etc



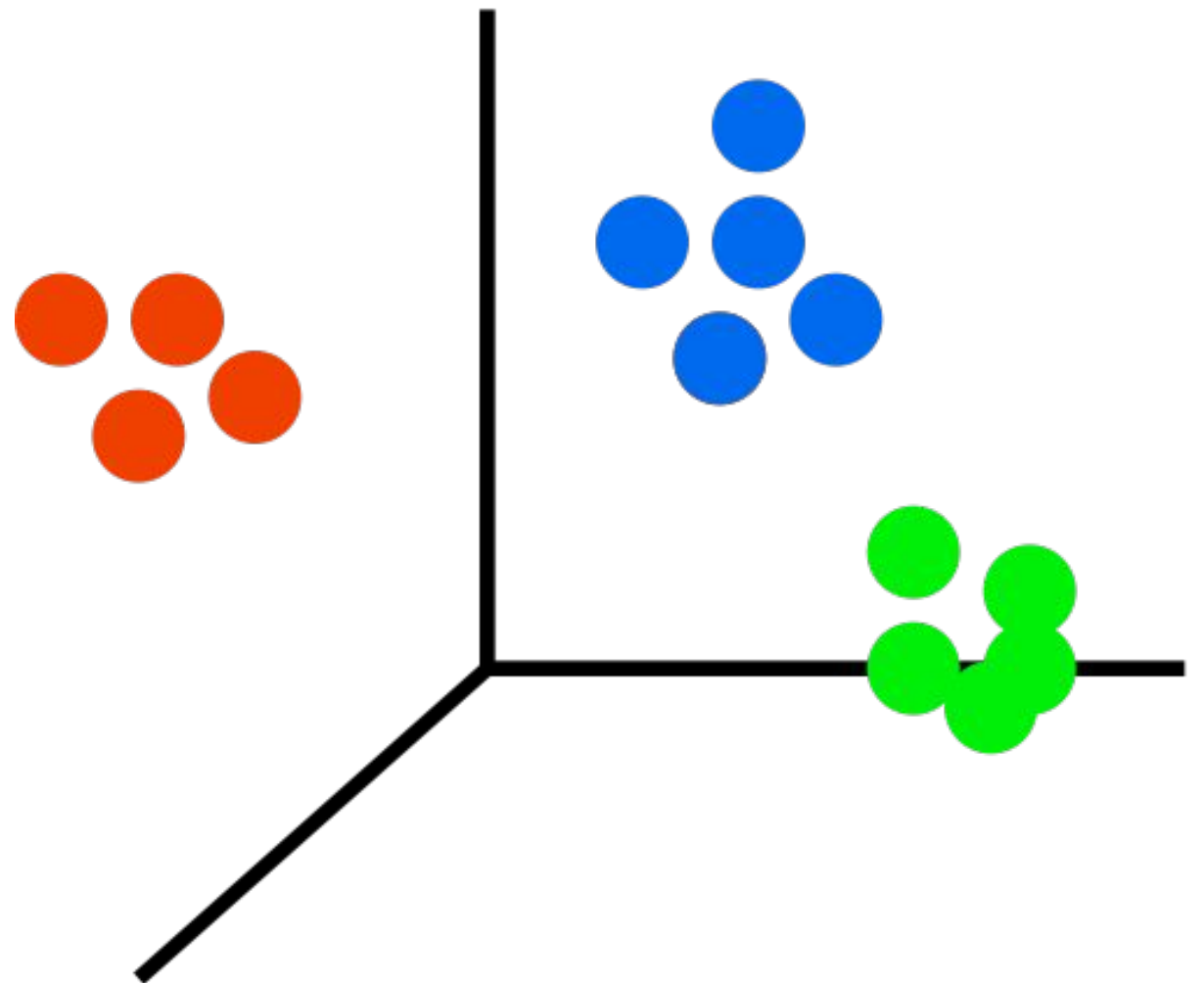


# CLUSTERING

- Conjunto de puntos (datos), cada uno con un set de atributos y una medida de similitud
- Encontrar conjuntos tales que:
  - Puntos en un *cluster* sean más similares entre sí
  - Puntos en conjuntos diferentes sean menos similares entre sí

# Visualización de clustering

- Clustering 3D basado basado en distancia Euclidiana
- Distancia intra-cluster es minimizada
- Distancia inter-cluster es maximizada



# Clustering

## Aplicación 1

- Segmentación de mercado
- Meta: Subdividir un mercado en subconjuntos de clientes en donde cualquier conjunto es un potencial objetivo de marketing (ej: Netflix, Amazon)
- ¿Cómo?

# Clustering

## Aplicación 2

- Clustering de documentos
  - Meta: Encontrar grupos de documentos que son similares entre sí, basándose en las palabras más importantes que contienen. (Directorios, Wikipedia)
  - ¿Cómo?



# Ejemplo

- Clustering de puntos: 3204 artículos del L.A. Times
- Medida de similitud: cuántas palabras tienen en común estos documentos (después de filtrar algunas palabras).

| <b>Category</b>      | <b>Total Articles</b> | <b>Correctly Placed</b> |
|----------------------|-----------------------|-------------------------|
| <b>Financial</b>     | 555                   | 364                     |
| <b>Foreign</b>       | 341                   | 260                     |
| <b>National</b>      | 273                   | 36                      |
| <b>Metro</b>         | 943                   | 746                     |
| <b>Sports</b>        | 738                   | 573                     |
| <b>Entertainment</b> | 354                   | 278                     |

# Reglas de Asociación

- Dado un conjunto de records, cada uno contiene un número de elementos de una colección determinada
- Objetivo: Producir reglas de dependencia que predecirán la ocurrencia de un elemento (ítem) basándose en ocurrencias de otros ítems.

# Reglas de Asociación

TID

Items

1

Pan, Coca-cola, Pañales, Leche

2

Cerveza, Pan

3

Cerveza, Coca-cola, Pañales, Leche

4

Cerveza, Pan, Pañales, Leche

5

Coca-cola, Pañales, Leche

# Reglas de Asociación

## Aplicación 1

- Promoción de Marketing y Ventas
  - Sea la regla encontrada del tipo

$\{\text{Queso}, \dots\} \rightarrow \{\text{PapasFritas}\}$



# Reglas de Asociación

## Aplicación 2

- Manejo de góndolas en los supermercados
- Meta: Identificar elementos que son comprados juntos por un número suficientemente grande de personas
- ¿Cómo?

# Reglas de Asociación

## Aplicación 3

- Manejo de inventario
- Meta: Una empresa de reparaciones a domicilio quiere anticipar la naturaleza de las reparaciones de sus consumidores, para mantener sus vehículos equipados y reducir nro. de viajes.
- ¿Cómo?

# Patrones secuenciales

- Dado un set de objetos asociados a una línea de tiempo de eventos, encontrar los elementos que tengan fuertes dependencias secuenciales entre ellos
- Se forman reglas descubriendo patrones y luego se aplican restricciones de tiempo

# Regresión

- Predecir el valor de una variable continua, en base a valores de otras variables, asumiendo modelo de dependencia lineal o no-lineal.
- Estadística y redes neuronales



# Detección de desviación/anomalía

- Detectar desviaciones significativas de los valores normales

# Desafíos de DM

- Escalabilidad
- Dimensionalidad
- Datos complejos y heterogéneos
- Calidad de los datos
- Distribución de los datos y propiedad
- Privacidad
- Streaming

# Próxima Clase

- Leer reglas del curso: ver que no haya problemas con los requisitos de asistencia, se entiende que Ud. puede cumplirlos si sigue en el curso.
- Bonus track ver el video de [Hans Rosling](#).