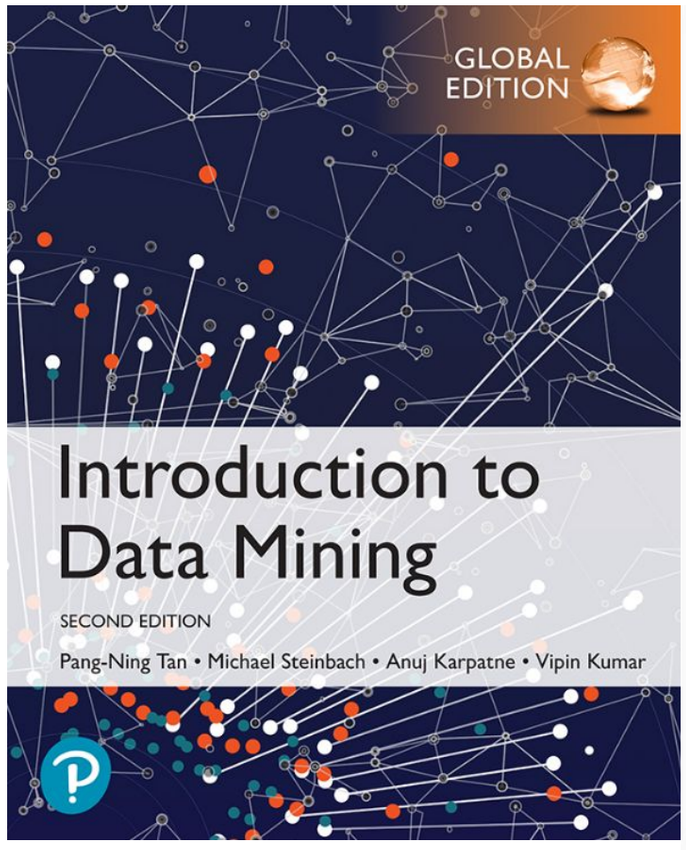
dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

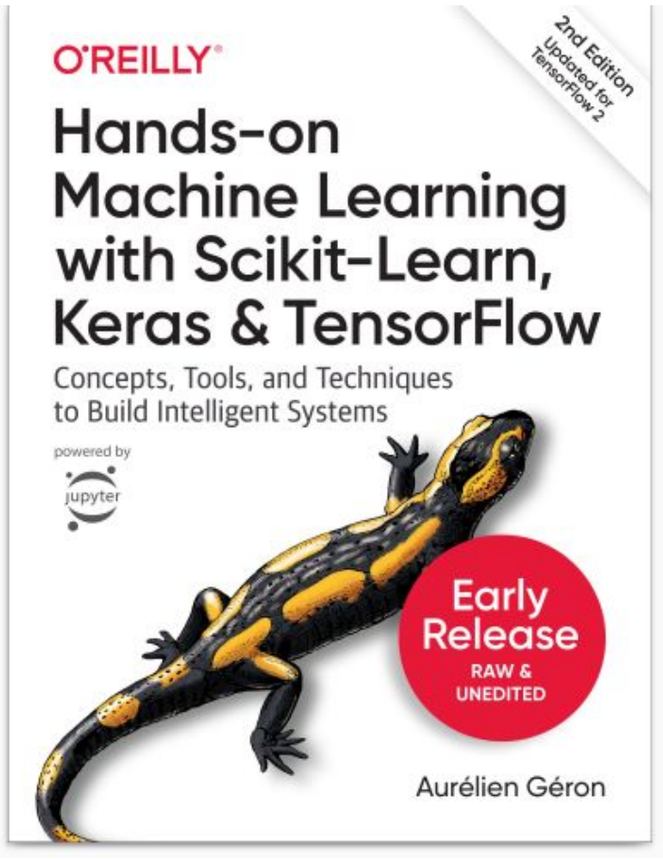
Introducción a la Minería de Datos

Libros del Curso

1. Introduction to Data Mining Segunda Edición
   - Autores: Pang-Ning Tan, Michael Steinbach, Vipin Kumar



2. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow
   - Autores: Aurélien Géron



Herramientas del Curso

Python es considerado uno de los lenguajes de programación más populares que provee bibliotecas potentes para Data Science y ML. https://www.python.org/downloads/

Scikit-learn es una biblioteca de aprendizaje automático de código abierto que permite el aprendizaje supervisado y no supervisado. https://scikit-learn.org/stable/index.html

Objetivos del curso

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

Contexto

- Estamos experimentando rápidos avances en tecnología de recolección y almacenamiento de datos (la Web, observatorios).
- A ese explosivo crecimiento de datos se le llama “big data”.
- Muchas áreas (industria, ciencia, salud, gobierno) necesitan tomar acciones en base a los datos.
- Desafíos: volumen, variedad (imágenes, videos, texto), velocidad de datos.
- Necesidad de herramientas automatizadas para extraer información útil.

¿Qué significa **Minar**?

Según la RAE:

“Hacer grandes diligencias para conseguir algo”

EL PROCESO DEL ORO DE PRINCIPIO A FIN

UNA VEZ DESCUBIERTA LA ZONA MINERALIZADA, EL ÁREA DE GEOLOGÍA DE YANACOCHA 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.

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 carguío.

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.

3 CARGUÍO Y ACARREO

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

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.

PAD. Es la estructura donde se acumula el mineral extraído del cerro para ser lixiviado y así recuperar el oro existente. También se le denomina "pila de lixiviación".

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.

TOP SOIL. Es la capa de tierra orgánica que permite el crecimiento de vegetación en la superficie. Esta capa se retira antes de iniciar el trabajo de explotación en sí y se almacena en áreas especiales para ser utilizada posteriormente en los trabajos de restauración del terreno.

Una vez culminada la etapa de explotación, el PAD es recubierto con el TOP SOIL extraído.

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.

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.

PLANTA DE TRATAMIENTO DE AGUAS ÁCIDAS

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 Ósmosis Inversa.

Planta de Ósmosis Inversa.

Tanque de floculación y precipitación.

Tanque de destrucción de cianuro.

Tanque de clarificación.

Pozas de procesos.

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.

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.

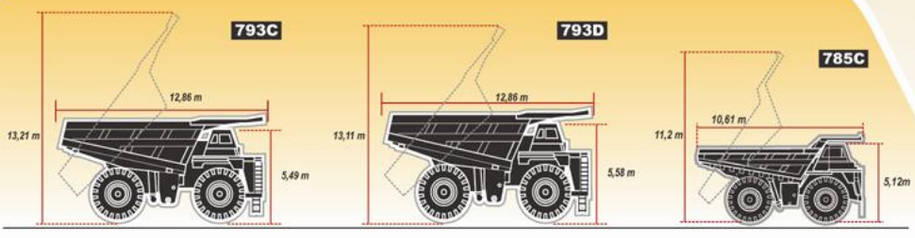
6 REFINERÍA (Fundición)

El oro obtenido en el proceso Merrill Crowe es sometido a operaciones de secado en hornos de retortas a 650ºC. Finalmente, el producto obtenido pasa por un proceso de fundición en horno de arco eléctrico a 1 200ºC para obtener el doré, que es el producto final.

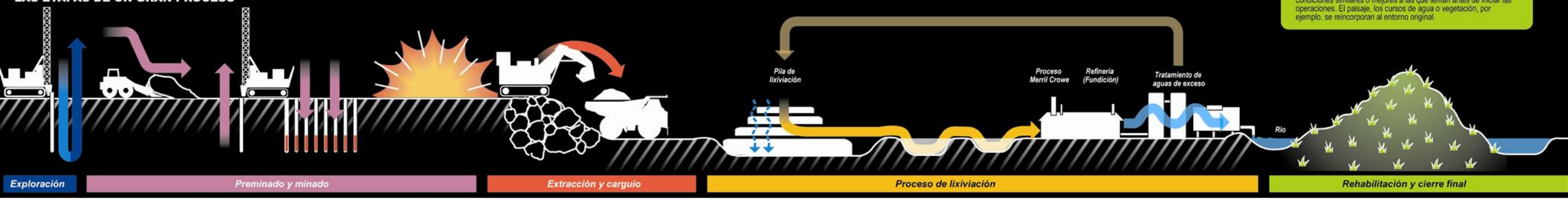
Reservorio San José

3,3 millones
de onzas ha sido el récord absoluto de producción durante un año. Esta cifra se consiguió en el 2005.

UN VISTAZO A LA MAQUINARIA



LAS ETAPAS DE UN GRAN PROCESO



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.

¿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

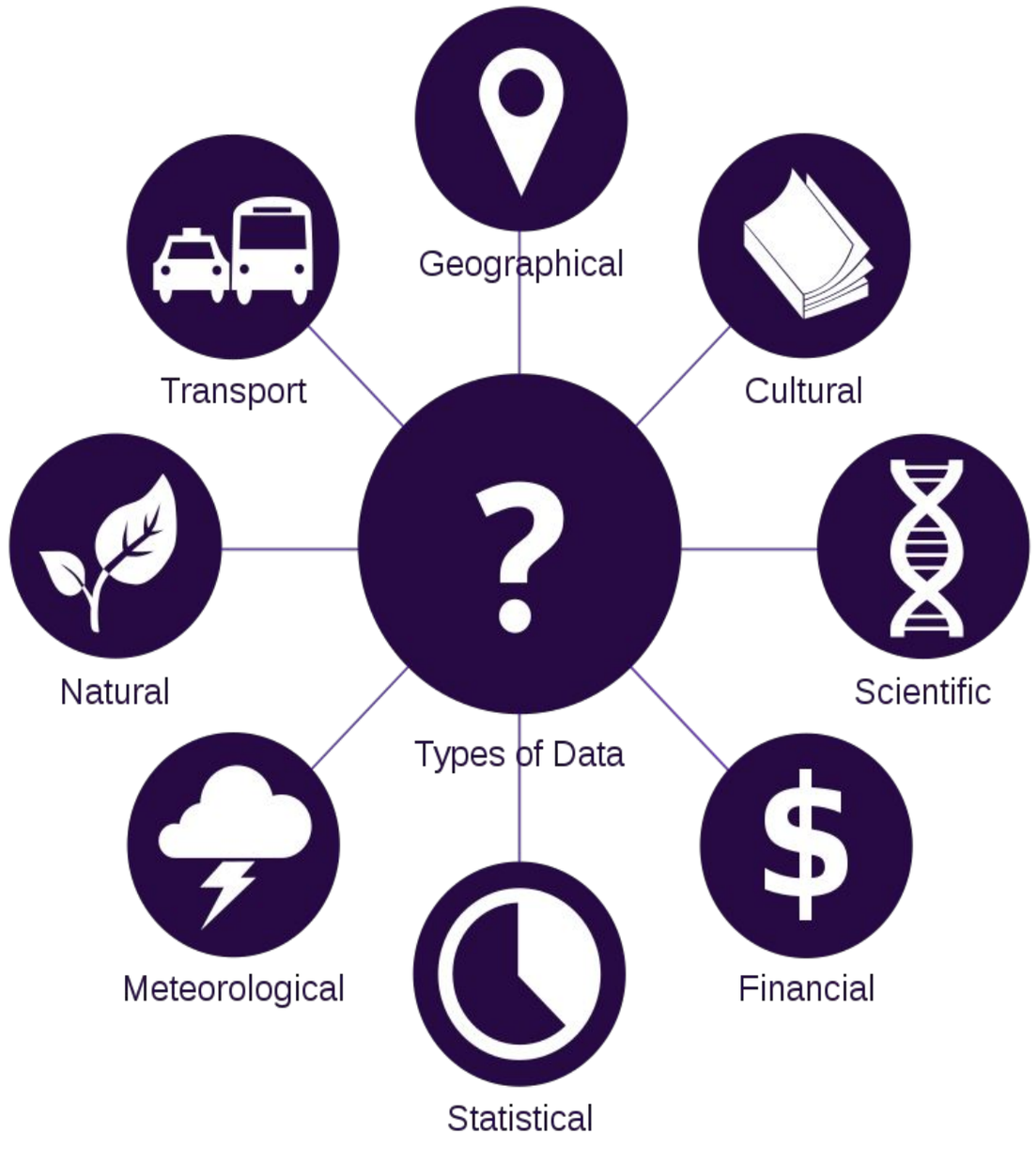
Datos

- Hasta ahora hemos dicho que ML consiste en aprender a partir de **datos**.
- Pero, ¿qué son los datos?
  - Según la OECD: Los datos son **características o información**, generalmente numéricos, que se recogen mediante la observación.

  - Según el Australian Bureau of Statistics: los datos son un conjunto de valores de variables **cualitativas** o **cuantitativas** sobre una o más personas u objetos.

Datos

- Existen muchos tipos de datos complejos como las secuencias de ADN, el texto, las imágenes, el video, el audio, datos espacio-temporales, etc.
- En este curso nos enfocaremos en **datos tabulares**.
- En cursos más avanzados aprenderán a trabajar con datos más complejos: procesamiento de **lenguaje natural**, procesamiento de **imágenes,** etc.



Fuente: Wikipedia

Datos Tabulares

- Un dataset tabular es una colección de **objetos** con sus **atributos**

- Un **atributo** es una propiedad o característica de un objeto
  - Ejemplos: color de ojos, temperatura, etc.
  - Un atributo también se conoce como: variable, campo, característica, dimensión.
- Un **objeto** es una colección de atributos
  - Un objeto también se puede llamar: registro, ejemplo, punto, instancia, observación.
  - Todos los objetos de un dataset tienen la misma cantidad de atributos.

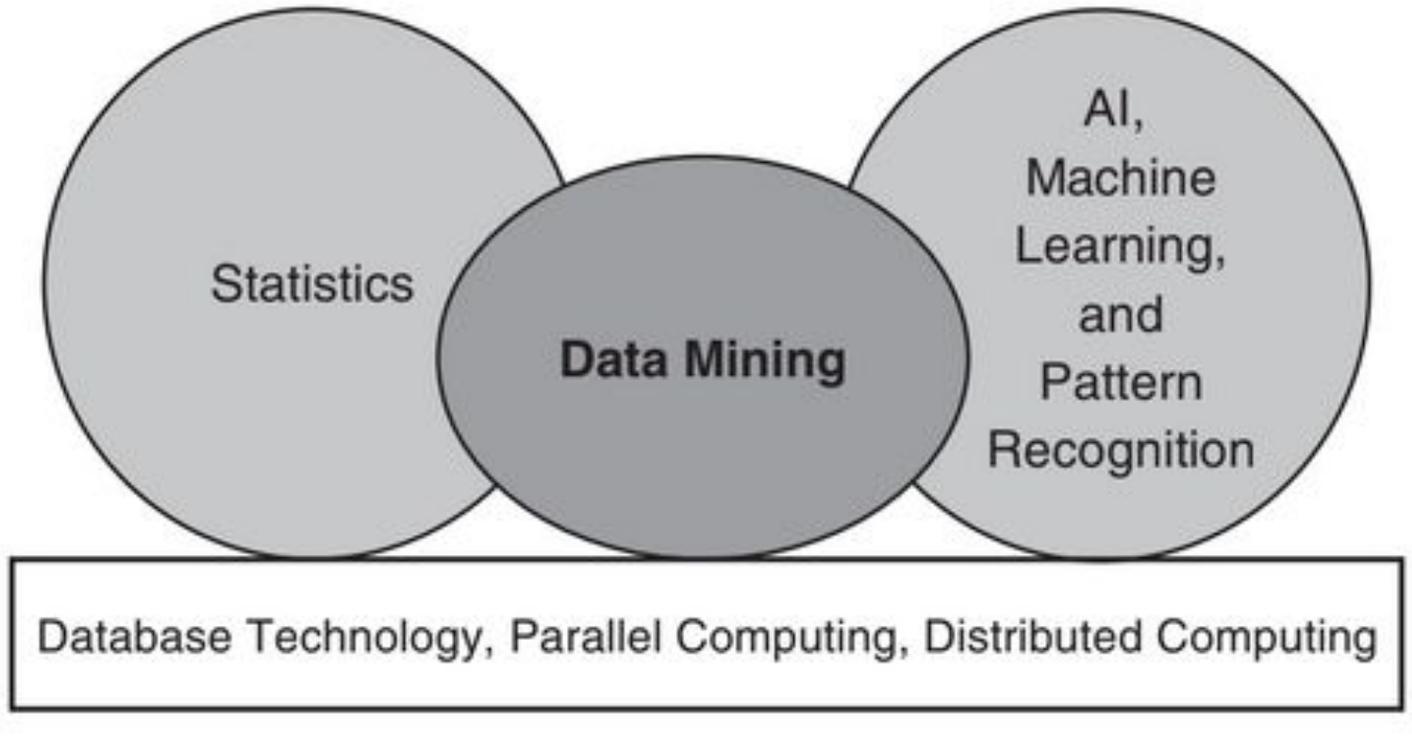
Atributos

Objetos

Tid	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

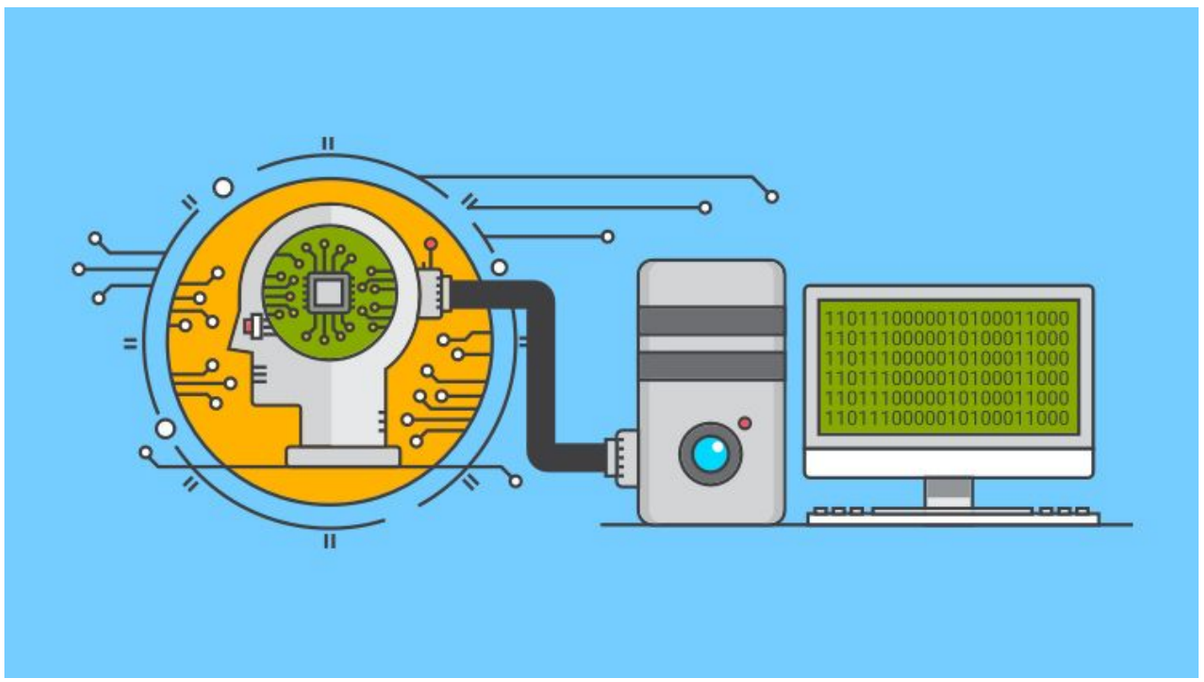
Orígenes de la MD

- La minería de datos y la ciencia de datos surgen como disciplinas que combinan ideas de varias otras disciplinas.
- Inteligencia Artificial (AI), Aprendizaje Automático (AI), reconocimiento de patrones, inferencia estadística y Bases de datos.
- La MD ofrece una "metodología" para aplicar técnicas de todas estas disciplinas para poder extraer conocimiento de grandes bases de datos.



Aprendizaje Automático (definiciones)

- Machine Learning es la ciencia (y el arte) de programar un computador para que pueda **aprender** a partir de los **datos**.
- Machine Learning es el campo de estudio que da a los computadores la capacidad de **aprender** sin ser programados explícitamente. Arthur Samuel, 1959.
- Se dice que un programa computacional aprende de la experiencia **E** con respecto a alguna tarea **T** y alguna medida de rendimiento **P**, si su rendimiento en **T**, medido por **P**, mejora con la experiencia **E**. Tom Mitchell, 1997.
- Machine Learning y su variante más popular “Deep Learning” son las áreas de mayor impacto en la **Inteligencia Artificial**.



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

- **Data Science** es el nombre reciente (2013) a la Minería de Datos **Data Mining (90's)**

- Una definición (sobre) simplista para distinguir a es:
  - **Data mining** genera **entendimiento**.
  - **Machine learning** genera **predicciones**.
  - **Artificial intelligence** genera **acciones**.

http://varianceexplained.org/r/ds-ml-ai/

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

- **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.

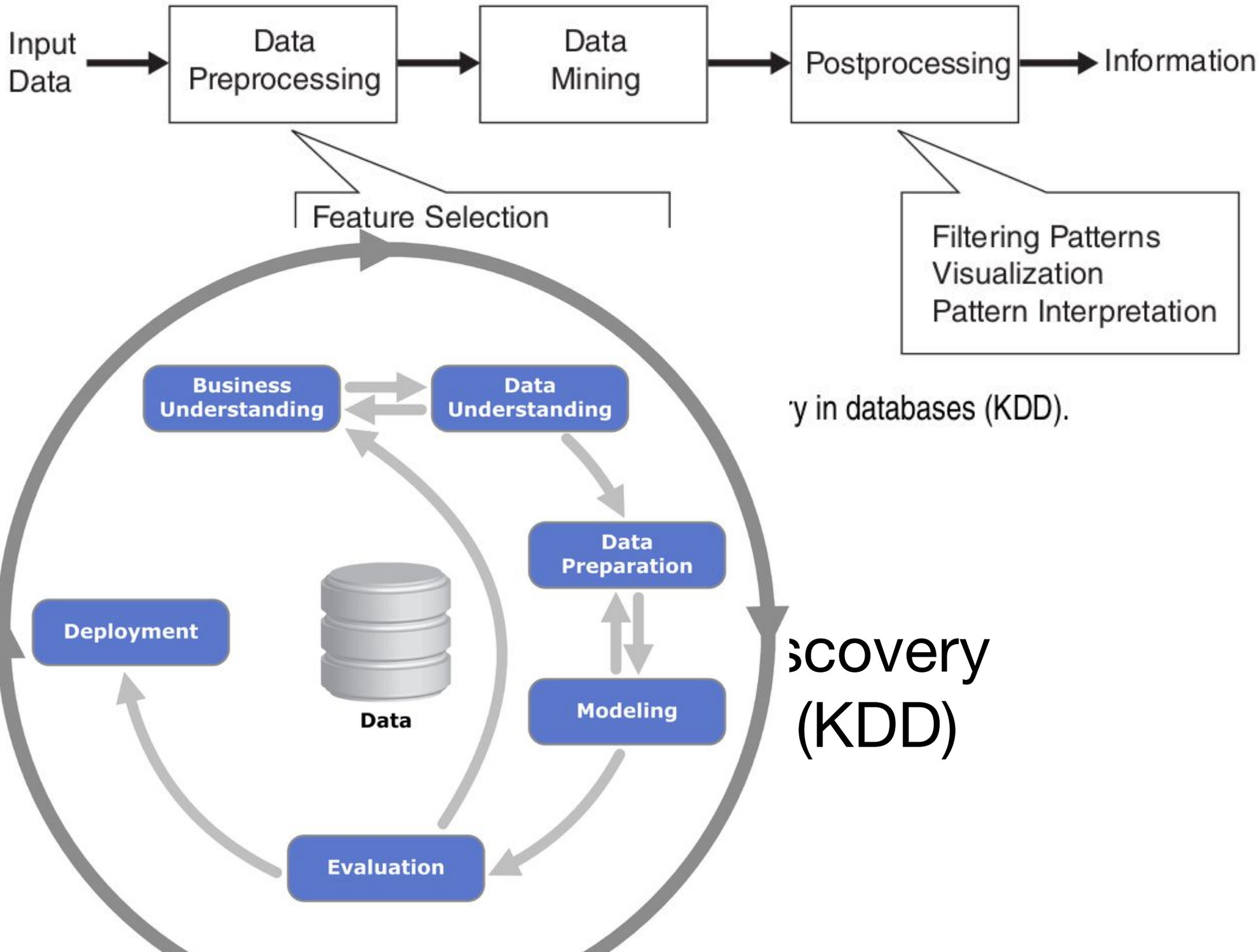
- El hecho que alguien escriba no lo convierte en escritor.

¿Por qué es importante entender estas diferencias?

- Porque este curso se centra en introducir la metodología de **de Minería de Datos**, enseñando los conceptos principales de Machine Learning.

- **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.

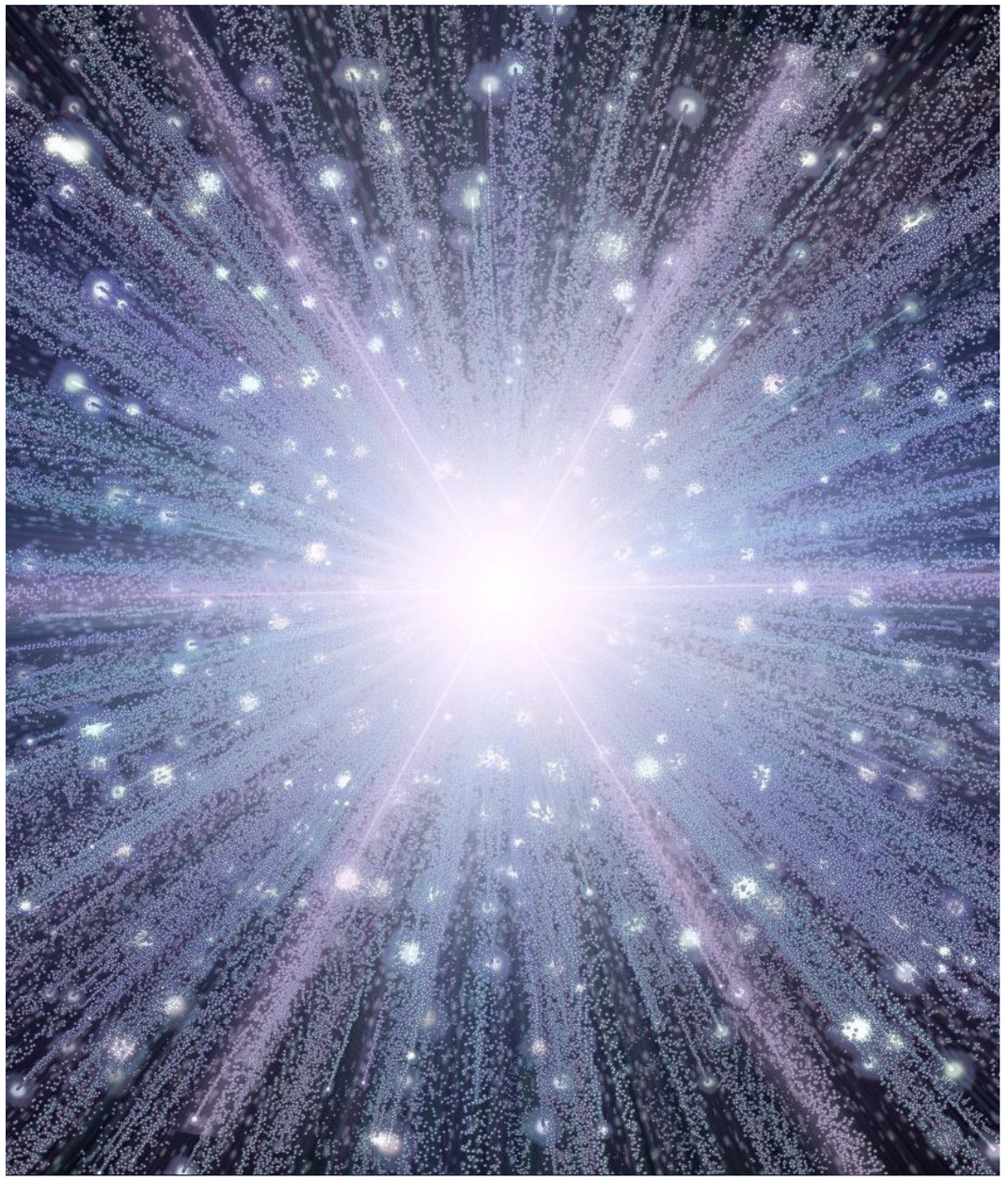


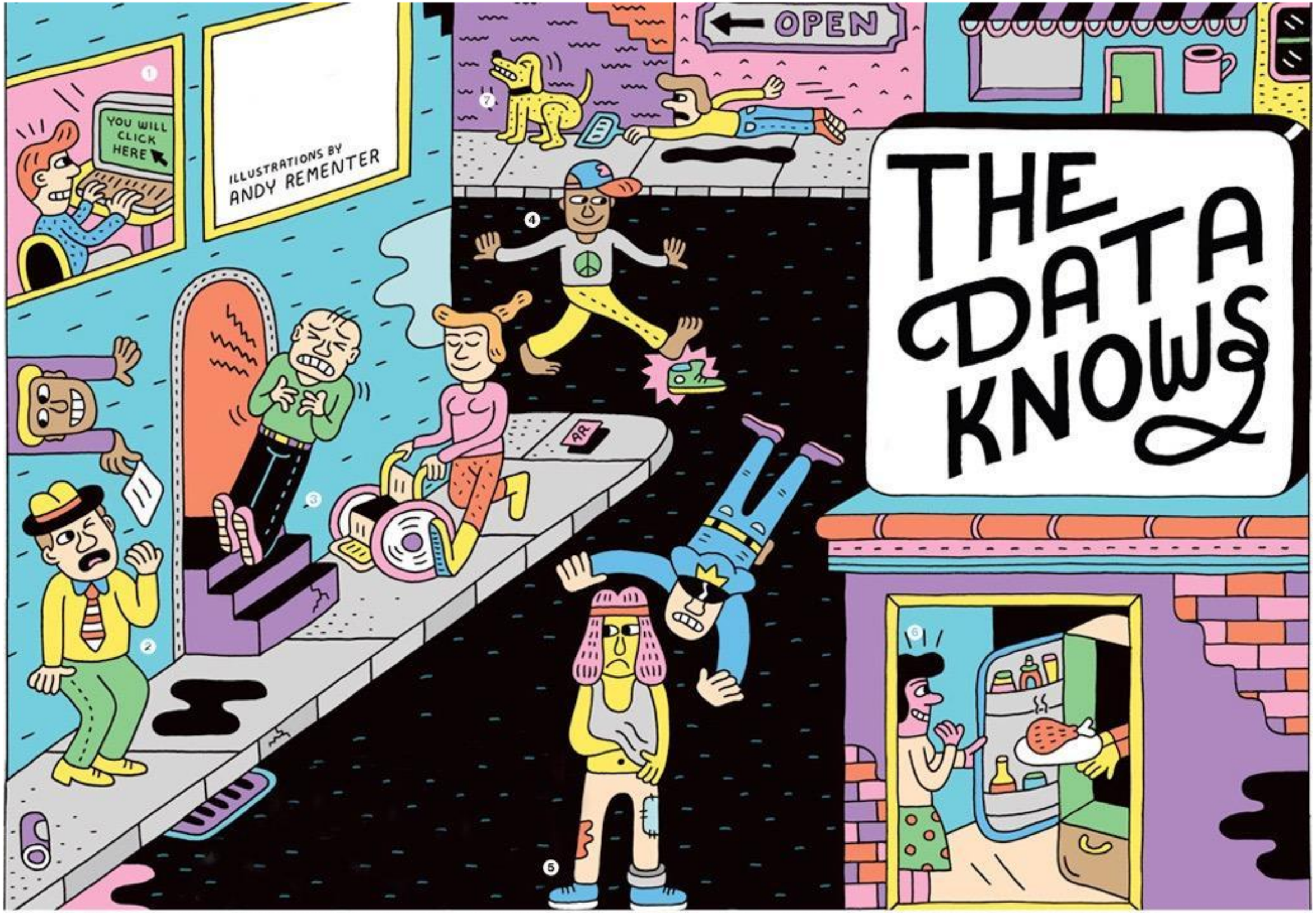
y in databases (KDD).

scovery
(KDD)

BIG BANG

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





❶ They know what you're going to buy ❷ They know when you're going to sneeze ❸ They know when you're about to have a heart attack
❹ They know the shoes you like ❺ They know when you're being bad ❻ They know what you want for dinner ❼ Sorry, can't help you there

¿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

Barbara's Amazon

ON ORDER
0 items

AMAZON PRIME
Try Prime
View benefits

AUDIBLE AUDIOBOOKS
Try Audible
Get 2 free audiobooks

CUSTOMER SINCE
2008

Recommended for you, Barbara

Literature & Fiction
100 ITEMS

Science Fiction & Fantasy Books
41 ITEMS

Prime Video – Unlimited Streaming for Prime Members
27 ITEMS

Mystery, Thriller & Suspense Books
55 ITEMS

Personal Care Products
81 ITEMS

Recommended Based On *Sketching User Experiences: Getting the Design...*
16 ITEMS

Office & School Supplies
20 ITEMS

Cell Phones & Accessories
10 ITEMS

Roll over image to zoom in

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 **AmazonGl**
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

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

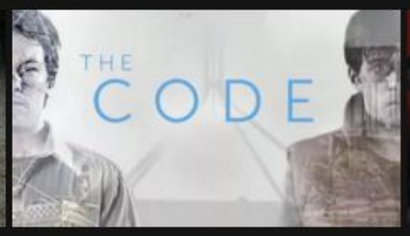
American Greetings Frozen Party Accessories, Pencils, 12 Count
89
$5.26 Prime

Disney Frozen Hot Pink Elsa Anna and Olaf Stationery Set Pack with Case (13 Pcs)
34

Thermos 12 Ou
Funtainer Bottl
Purple
$17.25 Prime

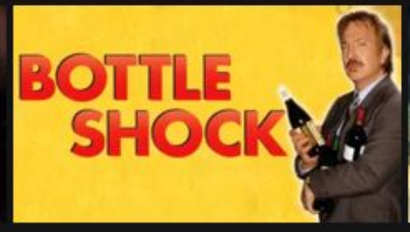
NETFLIX Browse Kids Search 1 Barbara

TV Thrillers & Mysteries



Romantic Movies

Continue Watching for Barbara



Watch It Again

Top Picks for Barbara

NETFLIX Browse Kids Search 1 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.

★★★★½ 2010 TV-14 3 Seasons

His deception detection is second to none. But his social skills? Well, they could use a little work.

★★★★½ 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.

★★★★½ 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

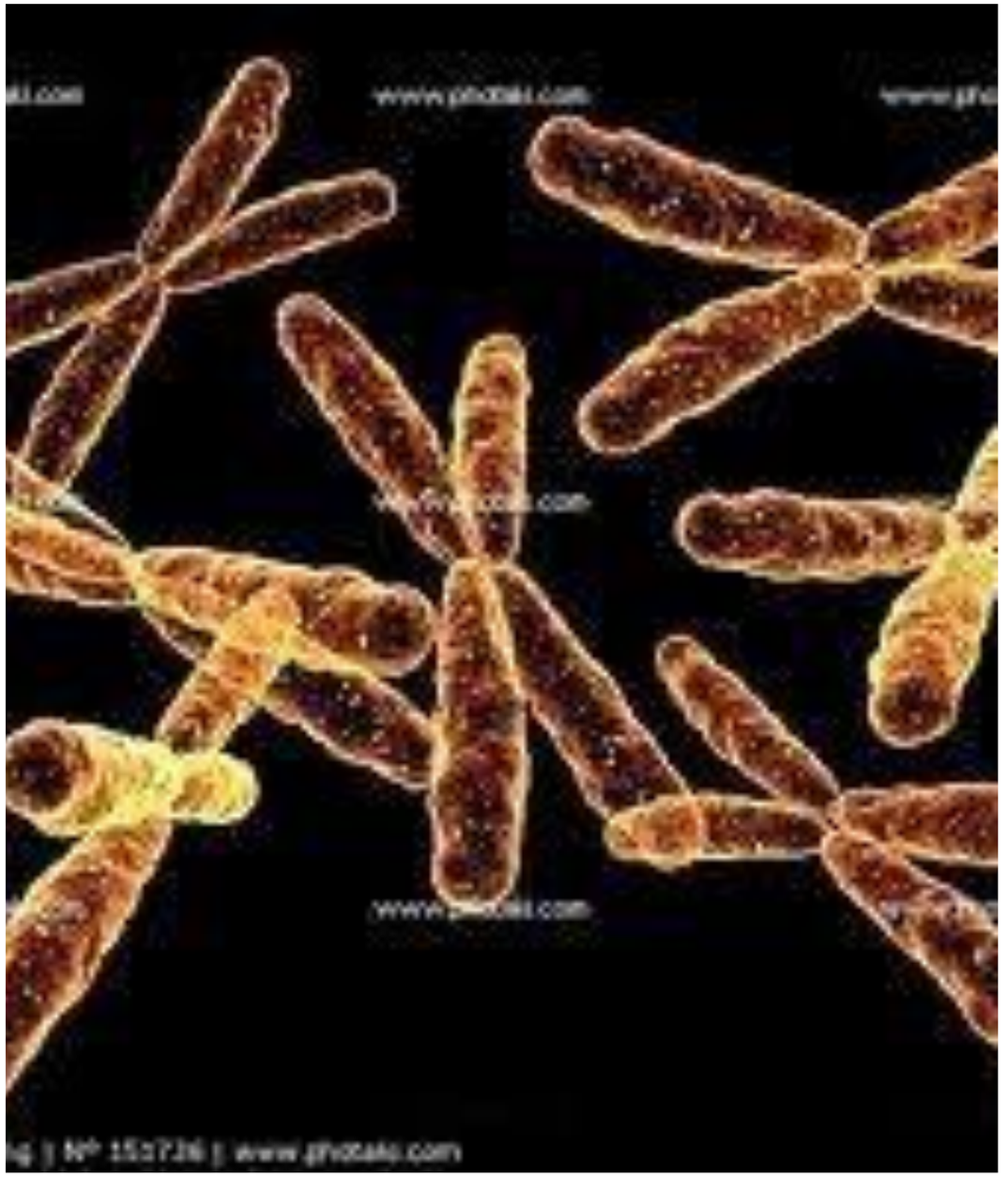
Because you watched Cooked

¿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 Flu Trends

Google.org home

Flu Trends

Select country

Home

How does this work?

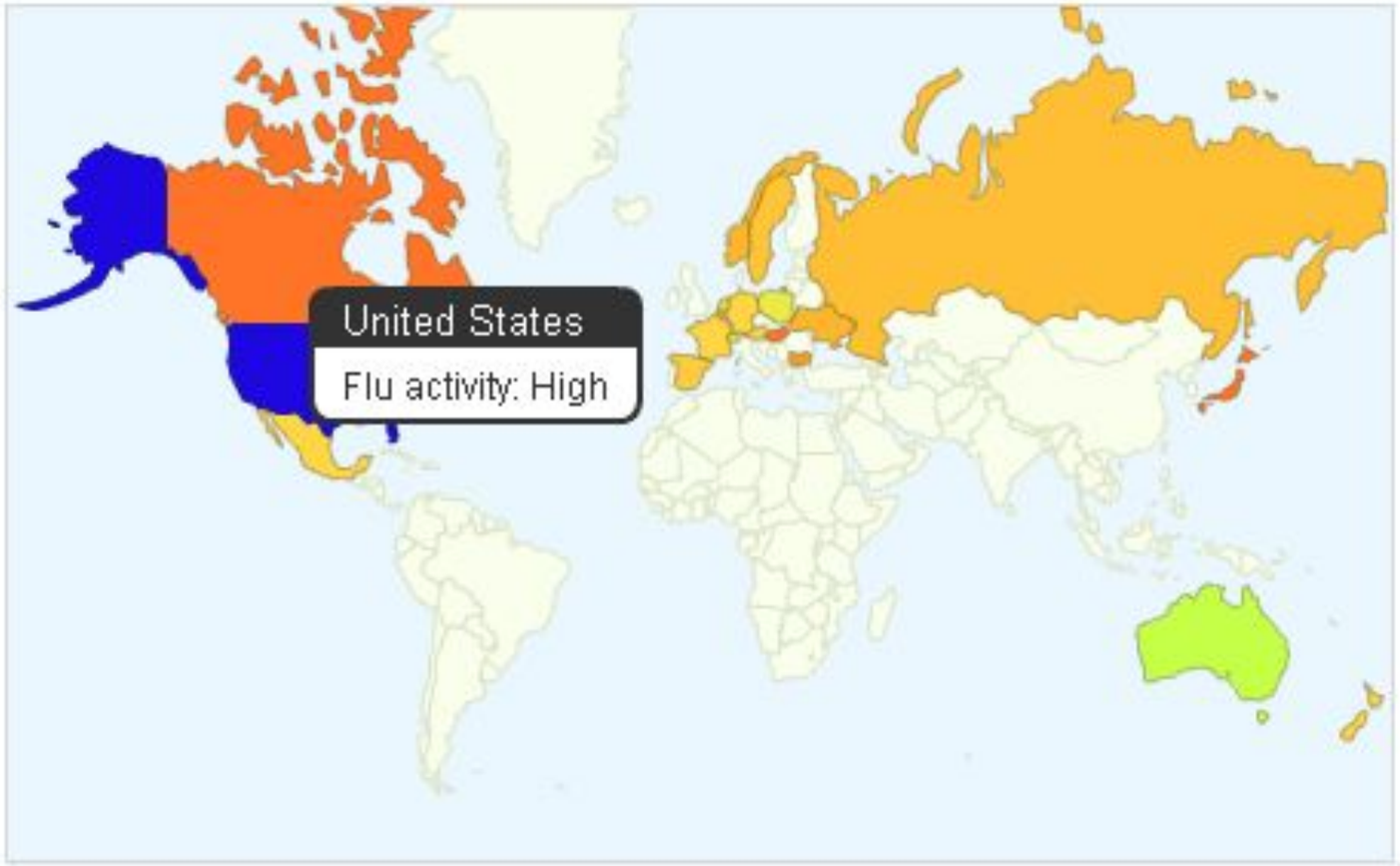
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

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)**

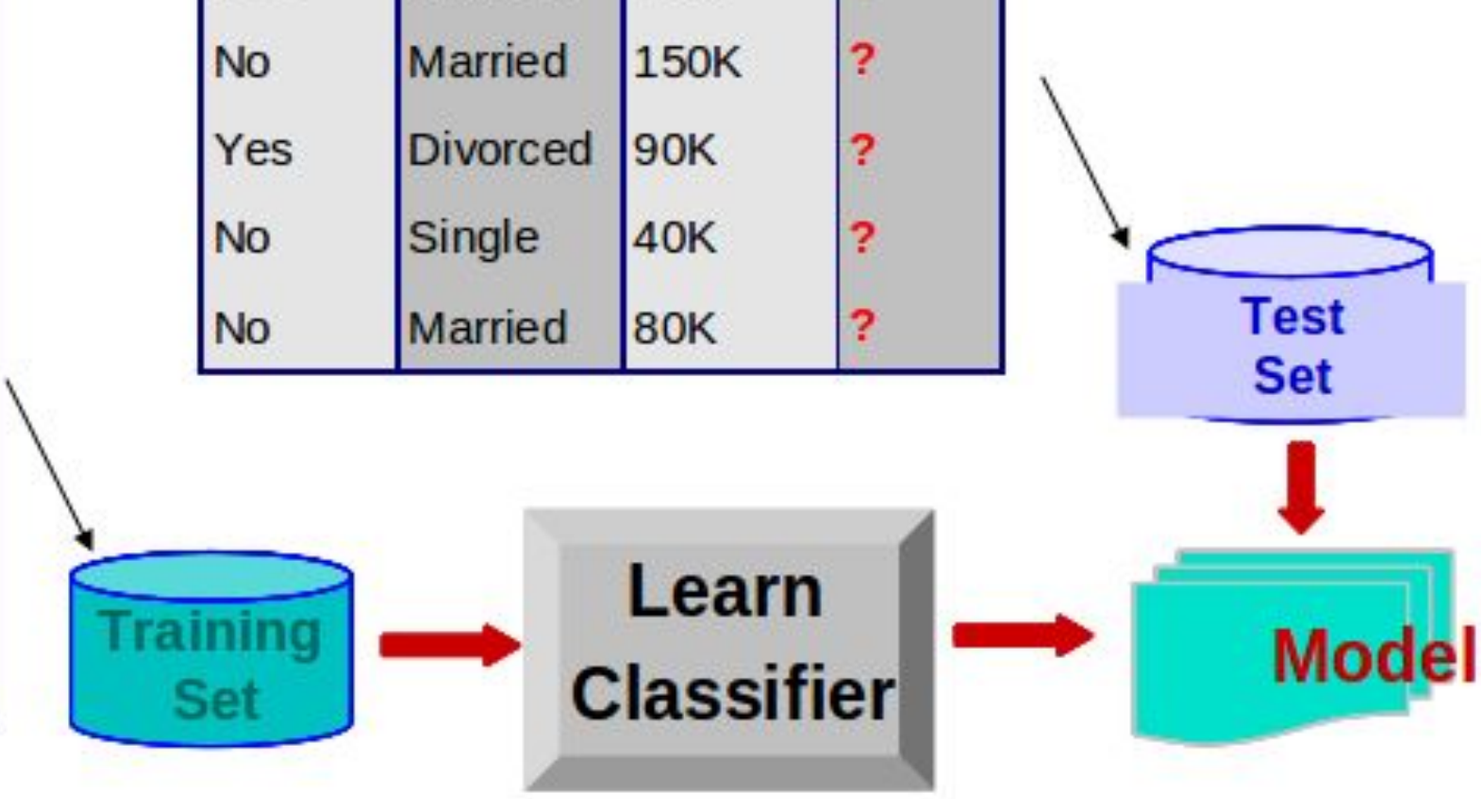
Clasificación

- 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

Tid	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?

Clustering

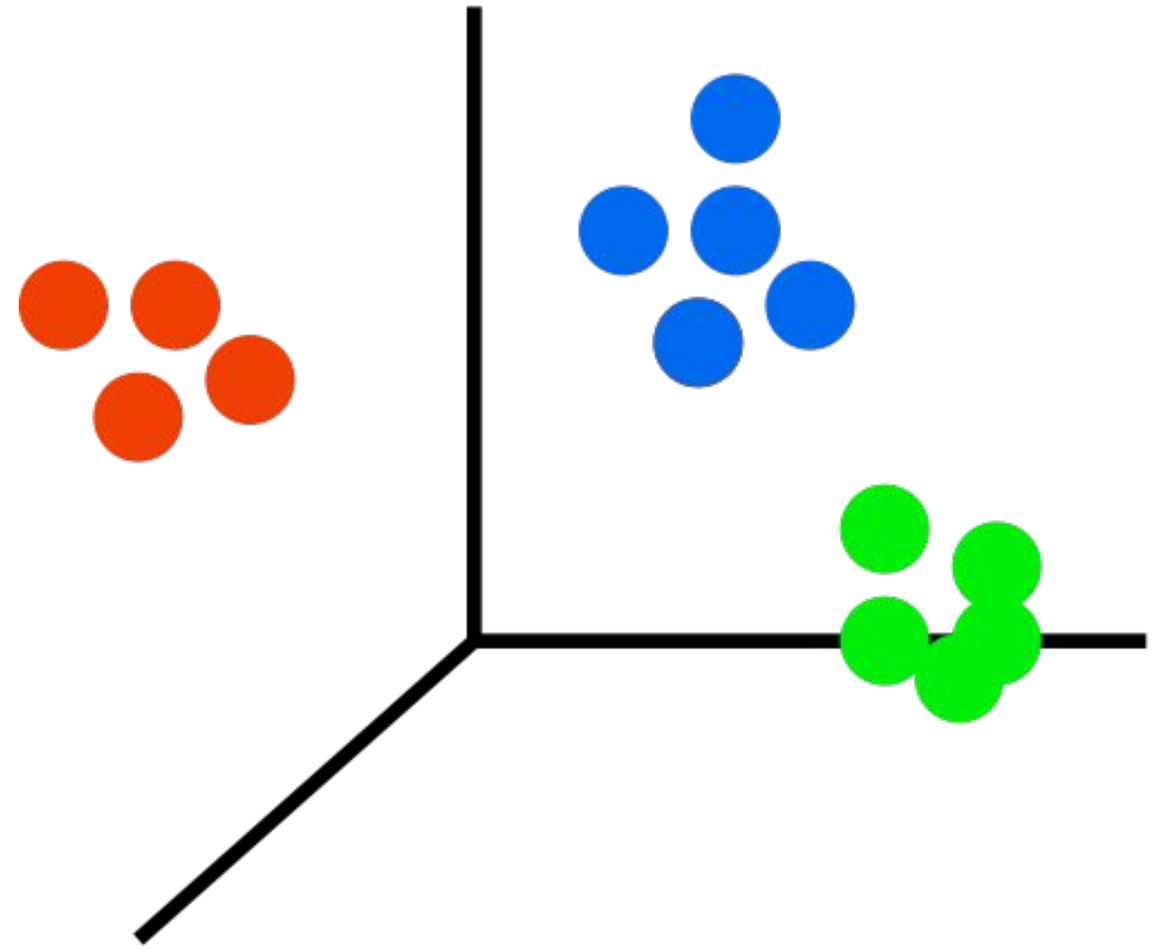
- Conjunto de puntos (datos), cada uno con un set de atributos y una medida de similitud
- Encontrar conjuntos tales que:
  - Puntos en un *cluster* sean más similares entre sí
  - Puntos en conjuntos diferentes sean menos similares entre sí

categorical categorical continuous class

Tid	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

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).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	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

{Queso, ...} $\longrightarrow$ {Papas Fritas}

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

Contenido Recomendado

- Hans Rosling: The best stats you've ever seen

dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

/ DCCUCHILE