Minería de Datos
IIC2433

Datos

Vicente Domínguez

Genéricamente

- Inteligencia de Máquina: **Aprender de los Datos**
- Revisemos de modo conceptual un ejemplo: Construir para un banco un sistema que automáticamente apruebe o niegue crédito

Rechazar o
Aprobar credito??

Applicant information:

age	23 years
gender	male
annual salary	$30,000
years in residence	1 year
years in job	1 year
current debt	$15,000
...	...

Formalización del Problema de Aprendizaje

- Encontrar la fórmula de aprobación *g()* que se aproxime lo más posible a la fórmula ideal *f()*

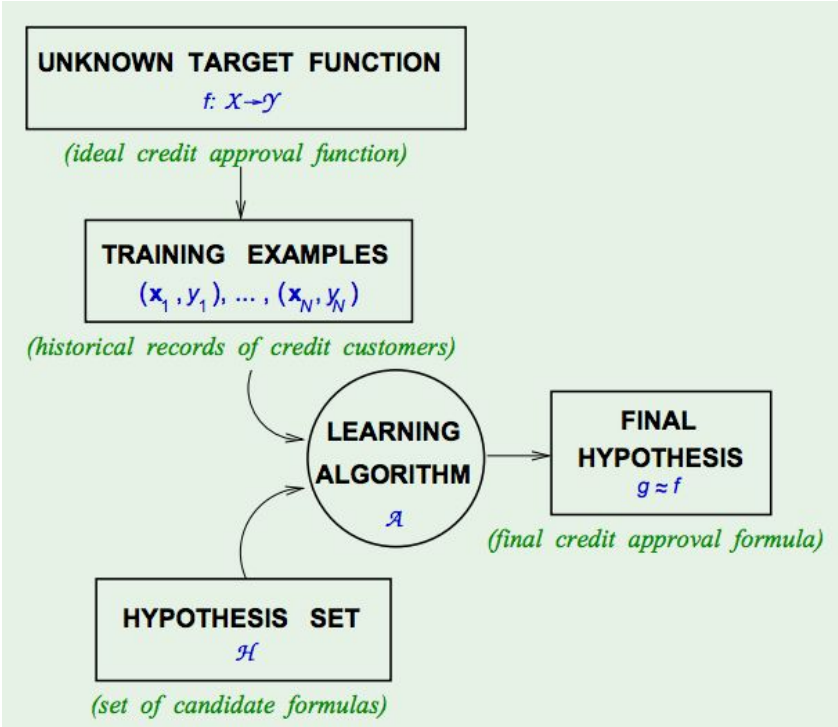
- Input: $\mathbf{x}$ (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ (*historical records*)

$\downarrow \quad \downarrow \quad \downarrow$

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (*formula to be used*)

Formalización del Problema de Aprendizaje

- El conjunto de hipótesis H



¿Qué son los datos?

En esta clase, aprenderás

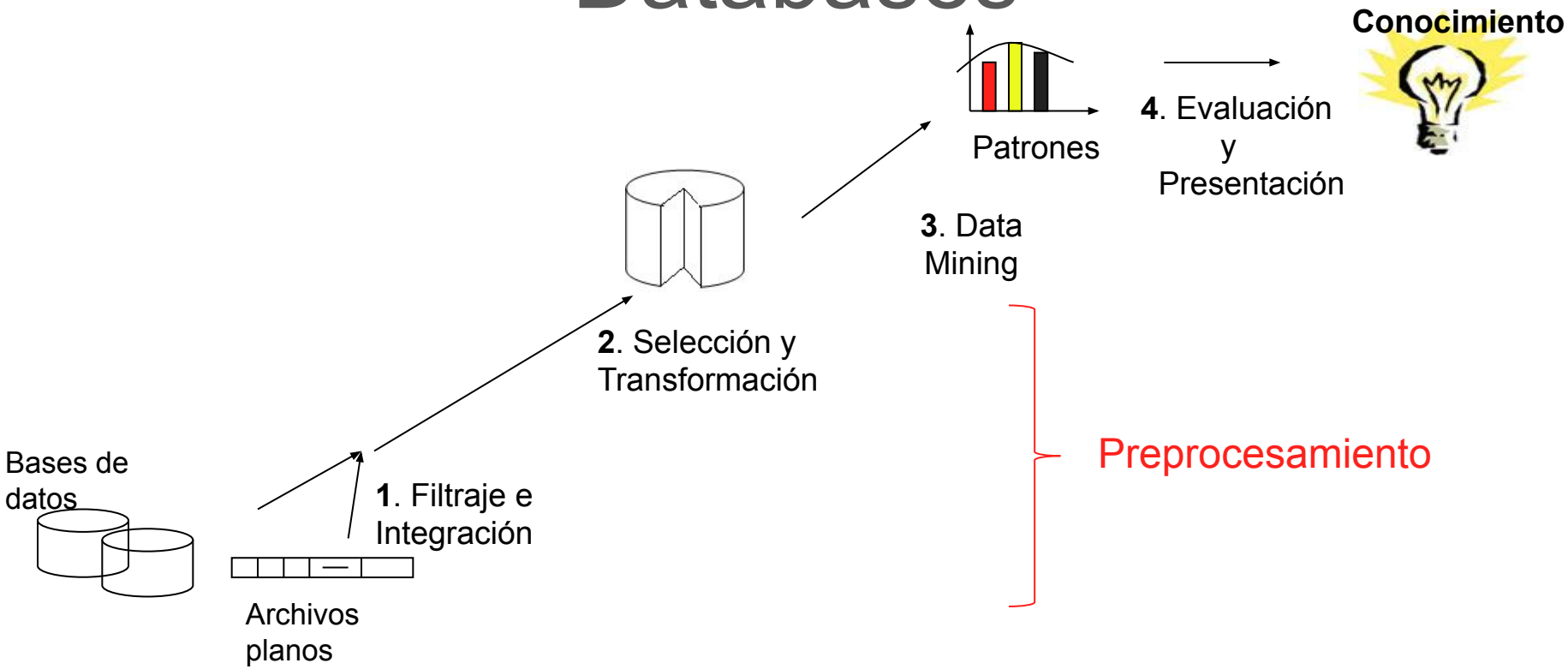
- Qué tipos de datos existen y cómo suelen almacenarse
- Métricas y visualizaciones para exploración de datos
- Las siguientes slides están reproducidas o parcialmente modificadas de:

http://www-users.cs.umn.edu/~kumar/dmbook/index.php

de los autores del libro:

Introduction to Data Mining (2005) de Tan, Steinbach y Kumar.

Knowledge Discovery in Databases



¿Qué son los datos?

- Colecciones de objetos y sus atributos
- Un atributo es una propiedad o característica de un objeto
  - Los **atributos** se conocen también como variables, campos, características, o *features*
- Un conjunto de atributos define un objeto
  - Los **objetos** se conocen también como puntos, casos, entidades o instancias.

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

Valores de atributos

- *“Attribute values are numbers or symbols assigned to an attribute”*
- Los **valores** de un **atributo** son números o símbolos asignados a un atributo
- Distinción entre un atributo y su valor
  - A un mismo atributo se le pueden asignar diferentes valores
    - Ejemplo: Altura medida en metros o pies
  - A diferentes atributos se les puede asignar el mismo conjunto de valores
    - Example: Valores para ID y edad son números enteros
    - Pero sus propiedades son diferentes

Ejemplo en base de datos

¿Cuáles son atributos y cuales valores?

	confID	year	subjectivity	polarity
1	AAAI	2009	0.0000000	0.00000000
2	AAAI	2010	0.3358810	0.17477655
3	AAAI	2011	0.3700711	0.13407506
4	AAAI	2012	0.3468204	0.18819547
5	AAAI	2013	0.3342377	0.11930844
6	CHI	2009	0.3535978	0.14754842
7	CHI	2010	0.3625253	0.15583003
8	CHI	2011	0.3466291	0.15390548
9	CHI	2012	0.3715251	0.16388508
10	CHI	2013	0.3441474	0.14595761
11	CIKM	2009	0.3307190	0.15835698

Tipos de atributos

- Hay cuatro tipos de atributos
  - Nominal
    - Examples: ID, color de ojos, código postal
  - Ordinal
    - Examples: rankings (ej., sabor de papas fritas del 1 al 10), notas, altura in {tall, medium, short}
  - Intervalo
    - Examples: calendar dates, temperaturas en Celsius o Fahrenheit.
  - Razón
    - Examples: temperatura en Kelvin, largo, tiempo, cuenta

Ejemplo en bases de datos

¿Qué tipos de atributos se observan aquí?

	confID	year	subjectivity	polarity
1	AAAI	2009	0.0000000	0.00000000
2	AAAI	2010	0.3358810	0.17477655
3	AAAI	2011	0.3700711	0.13407506
4	AAAI	2012	0.3468204	0.18819547
5	AAAI	2013	0.3342377	0.11930844
6	CHI	2009	0.3535978	0.14754842
7	CHI	2010	0.3625253	0.15583003
8	CHI	2011	0.3466291	0.15390548
9	CHI	2012	0.3715251	0.16388508
10	CHI	2013	0.3441474	0.14595761
11	CIKM	2009	0.3307190	0.15835698

Properties of Attribute Values

- El tipo de atributo depende de qué propiedades posee:
  - Distinción: = ≠
  - Orden: < >
  - Adición: + -
  - Multiplicación: * /
  - **Nominal:** distinción
  - **Ordinal:** distinción y orden
  - **Intervalo:** distinción, orden y adición
  - **Razón:** las cuatro propiedades

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Atributos discretos y continuos

- Discreto:
  - Tiene un conjunto de valores finito o infinito numerable
  - Ejemplos: códigos postales, conjunto de palabras en un documento
  - Representados usualmente con **números enteros**
  - Note: binary attributes are a special case of discrete attributes
- Continuos:
  - Números reales como valores de atributo
  - Ejemplos: temperatura, altura, peso
  - En la práctica los valores reales solo pueden ser medidos y representados usando un número finito de dígitos.
  - Representados usualmente con **variables de punto flotante**

Tipos de *datasets*

- Registros
  - Matriz de datos
  - Datos de documentos
  - Datos de transacciones
- Grafos
  - WWW
  - Estructura molecular
- Ordenados
  - Datos espaciales
  - Datos temporales
  - Datos secuenciales
  - Datos de secuencias genéticas

Datos de registros

- Colección de registros con una cantidad fija de atributos

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

Matriz de datos

- Si los datos tienen una cantidad fija de atributos (sin valores faltantes), entonces se puede representar como una matriz de n x m.
- Generalmente una matriz de datos tiene solo datos de tipo “razón”

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Datos de documentos

- Cada documento pasa a ser un **vector de términos**
- Cada término es una componente del vector
  - El valor de la componente indica cuántas veces aparece ese término en el documento

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

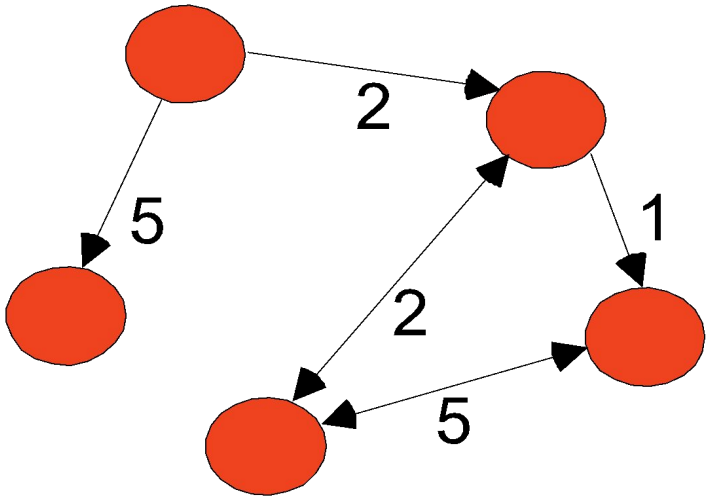
Datos de transacciones

- Un caso especial de registros, donde
  - Cada registro es una transacción que involucra ciertos ítems
  - Por ejemplo en un supermercado.
    - Una **transacción** es lo que compró una persona en una visita
    - Un **ítem** es un producto del supermercado

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Datos de grafos

- Ejemplos: Grafo genérico y links html




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Datos ordenados

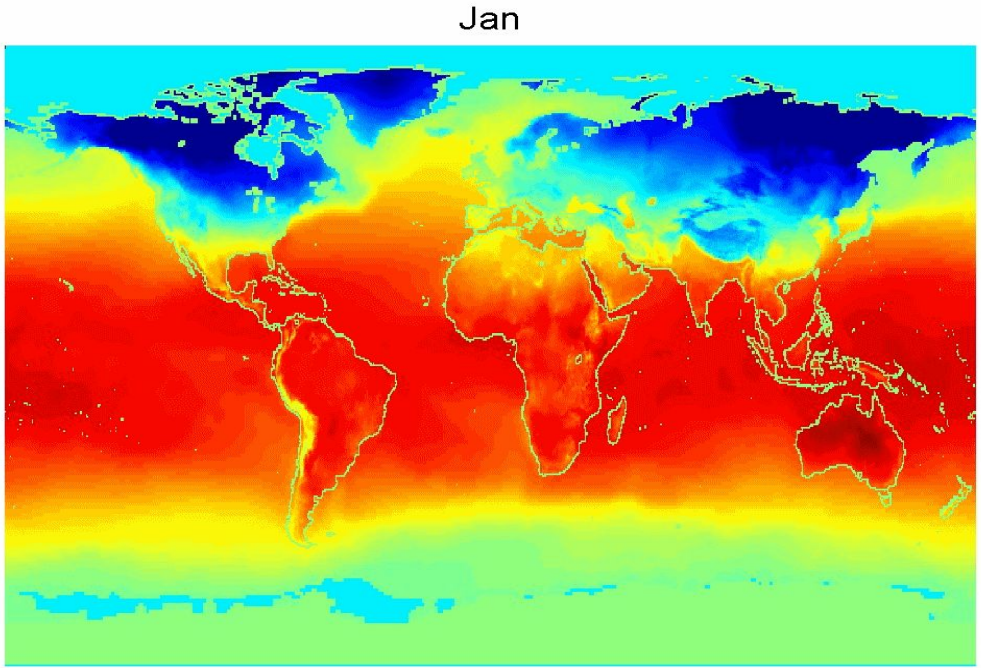
- Secuencia de genes

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Datos ordenados

- Datos espacio-temporales

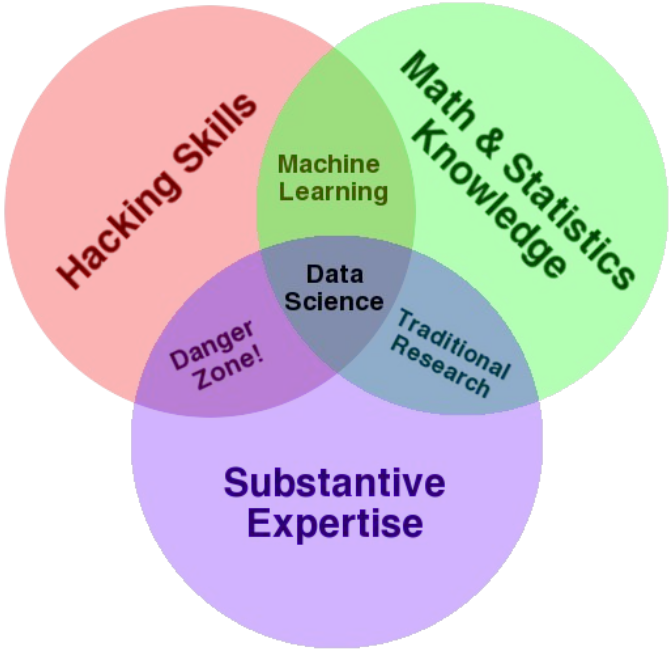
Temperatura
promedio mensual



¿Qué veremos esta clase?

- Visualización de datos
- Pre-procesamiento de datos

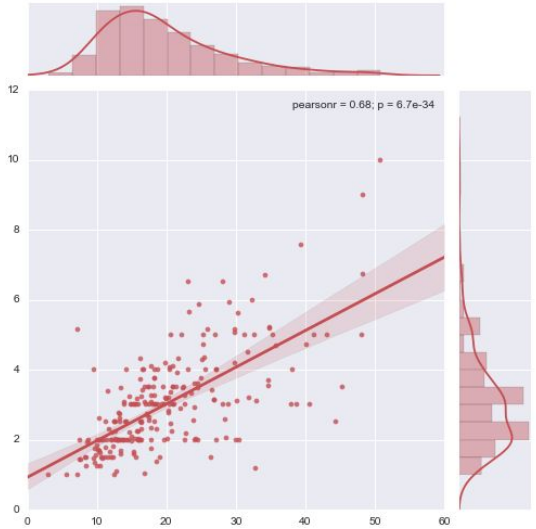
Ciencia de Datos *(Data Science)*

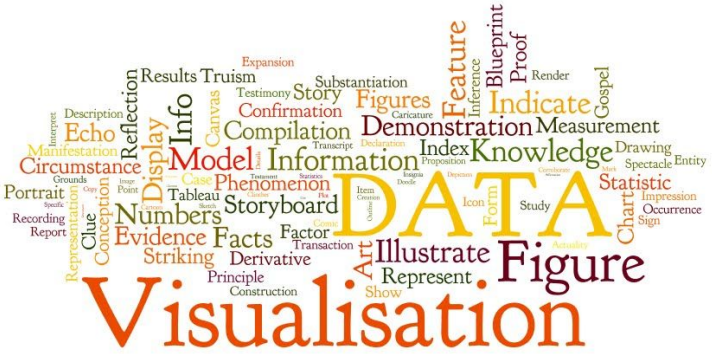


Fuente: Drew Conway

Visualización de datos

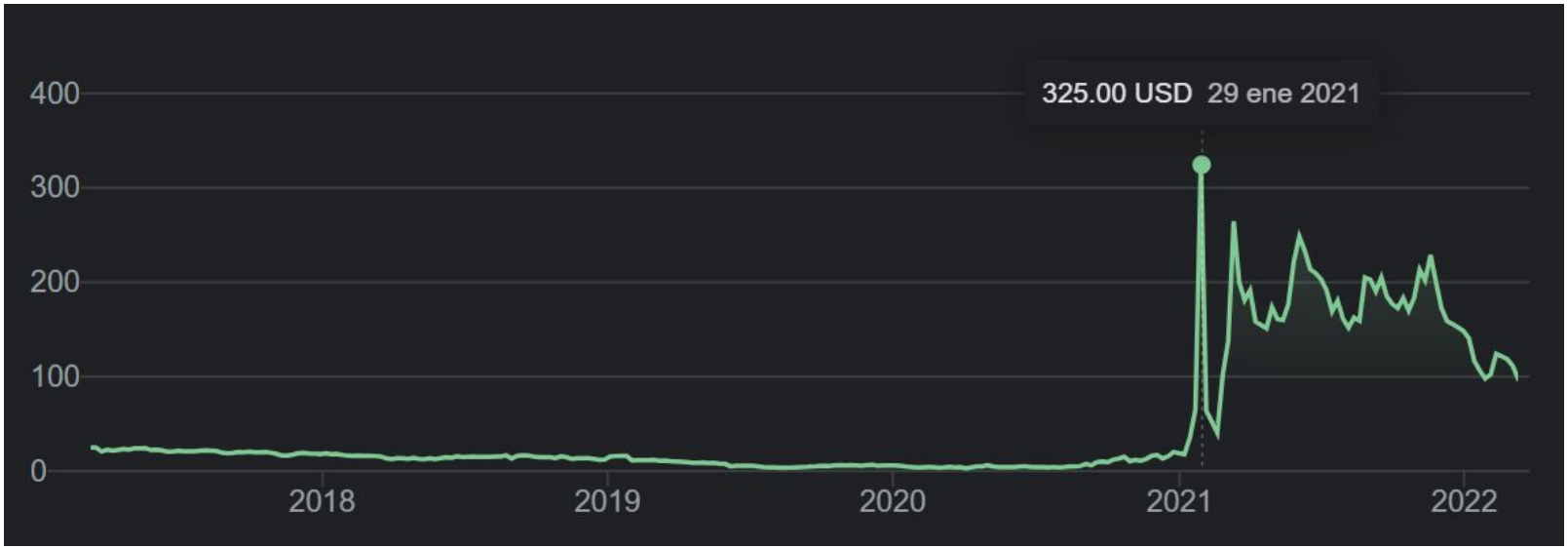
¿Qué es la visualización de datos?





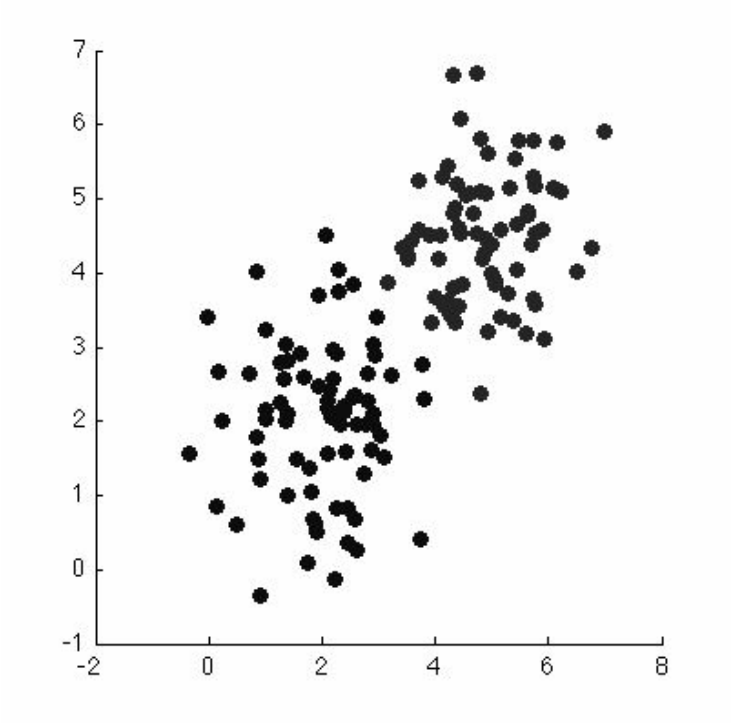
Visualización de datos

Comunicar



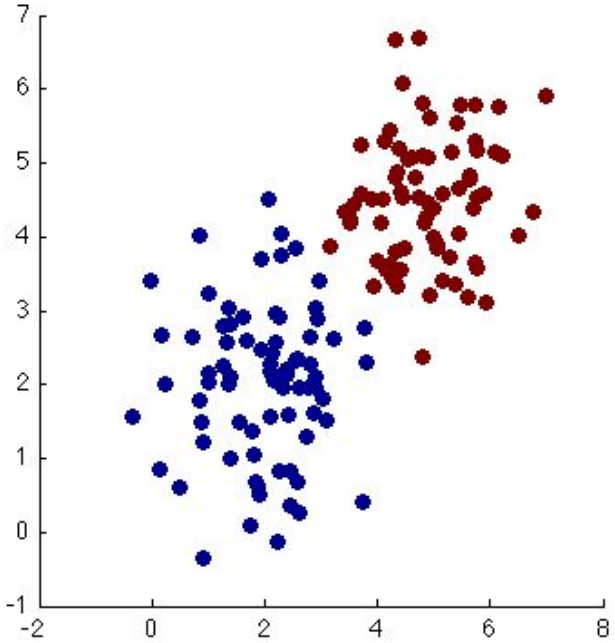
Visualización de datos

Explorar



Visualización de datos

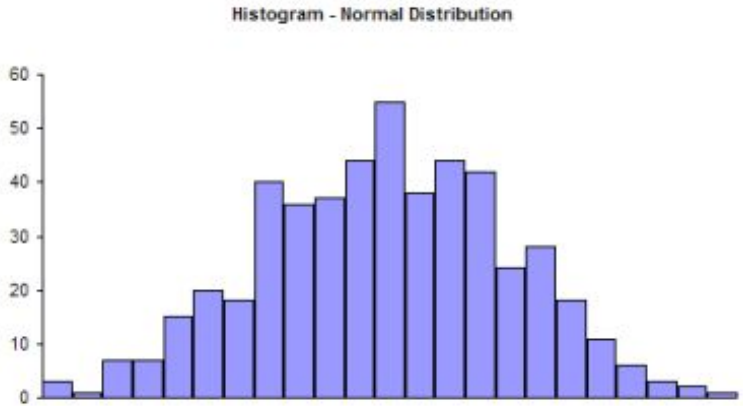
Explorar



Visualización de datos

Explorar

- Buscar patrones, tendencias, irregularidades, estructura



Histograma de valores

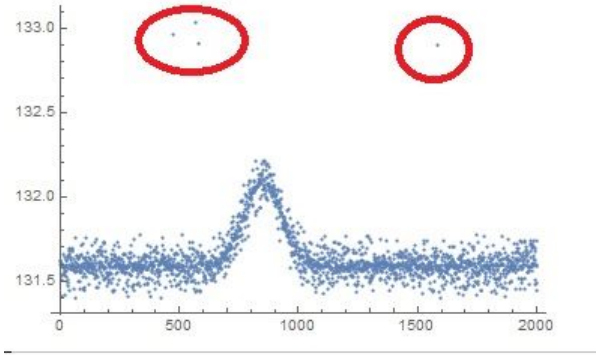


Gráfico de puntos

Pre-procesamiento de datos

Las siguientes slides están basadas en las slides del profesor Karim Pichara.

Pre-procesamiento de datos

Limpieza de datos

- Datos faltantes
- Datos erróneos
- Datos inconsistentes

Pre-procesamiento de datos

Datos faltantes

- Muchas veces un atributo viene vacío, y eso afecta el proceso de análisis:

- Soluciones posibles (todas tienen pros y contras):
  - Ignorar la tupla
  - Llenar los datos manualmente
  - Usar una cte. Global para llenar los valores: *Ej: “desconocido”, “-∞”, etc.*
  - Usar la media del atributo
  - Usar la media por clases
  - Usar el valor más probable (según herramienta de inferencia)

Pre-procesamiento de datos

Datos faltantes

- No siempre un dato faltante es un error. Ej, persona no tiene licencia de conducir, no usa tarjeta de crédito, etc.
- En esos casos es importante tener valores definidos como “no se aplica”, etc.

Integración de datos

La información en la mayoría de los casos debe ser integrada desde múltiples fuentes de datos.

Algunos problemas típicos:

- Identificación de la entidad
- Redundancia
- Detección y resolución de conflictos entre valores

Identificación de la entidad

La misma entidad tiene distintos nombres en diferentes fuentes de datos, ej, ***customer_id***, ***cust_number***.

Para esto se utiliza Metadata donde se almacena información sobre las entidades en cada fuente de datos, ej: nombre, significado, tipo de datos, rango, valores nulos, etc.

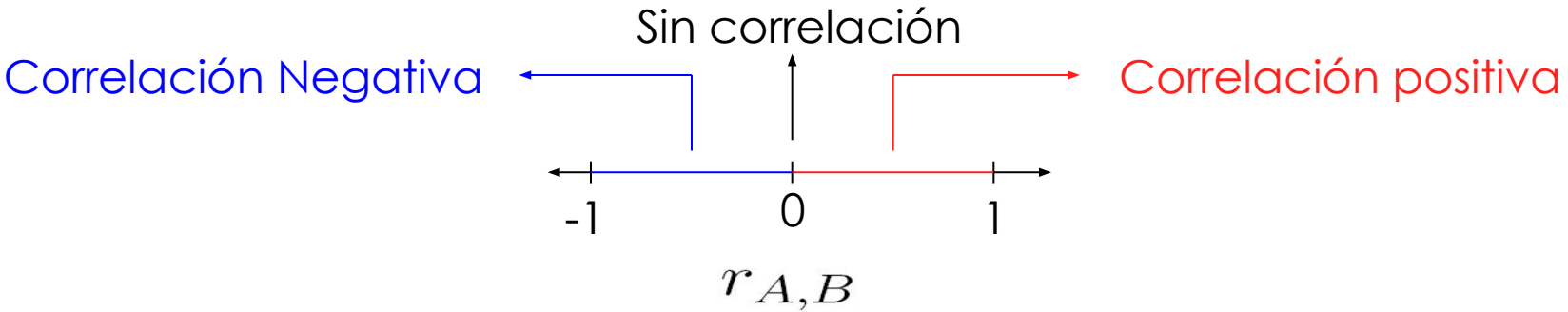
Redundancia

Un atributo es redundante si puede ser derivado de otro. Errores en la identificación de la entidad suelen llevar a situaciones de redundancia

Puede ser detectada realizando un análisis de correlación:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} \qquad -1 \leq r_{A,B} \leq 1$$

Redundancia



Meme del día

Más adelante habilitaremos un medio para recibir sus aportes para ser el meme del día.

Requisito: debe estar relacionado al contenido del curso.

Premio: ser el meme del día.