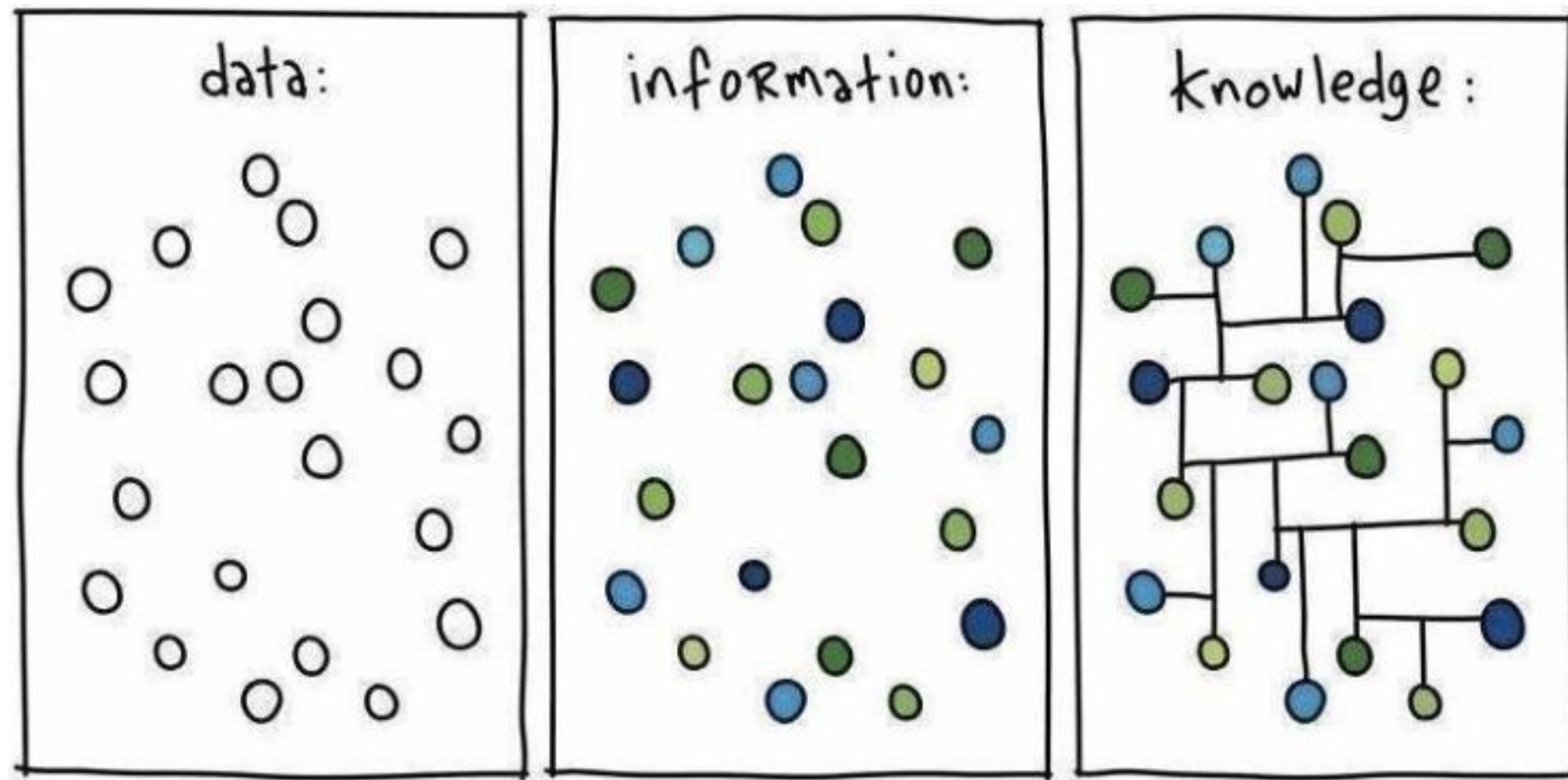


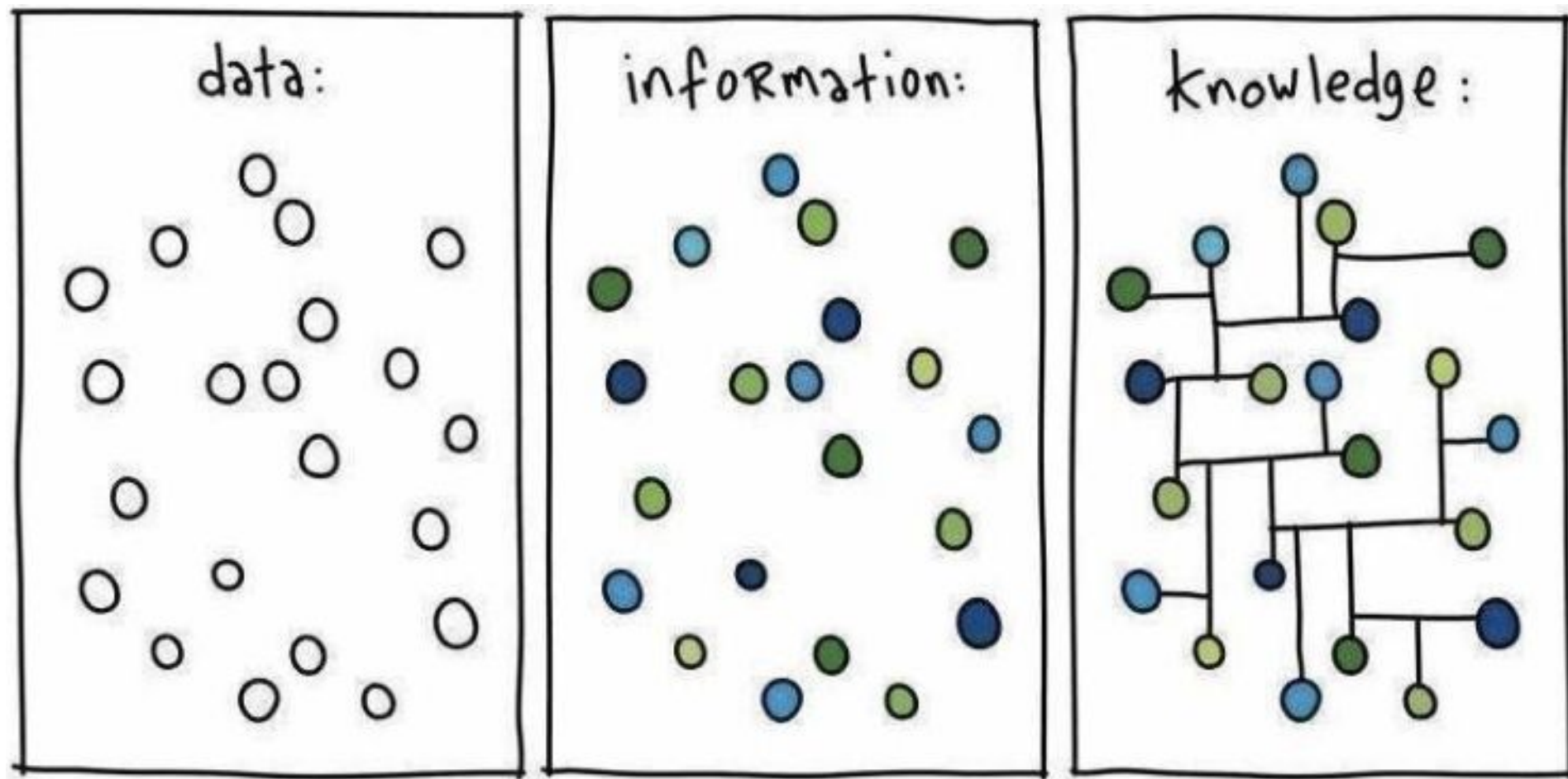


Datos

¿Qué entendemos por dato?



¿Qué entendemos por dato?



Juan Pérez obtuvo un 5.8 en la segunda prueba de historia

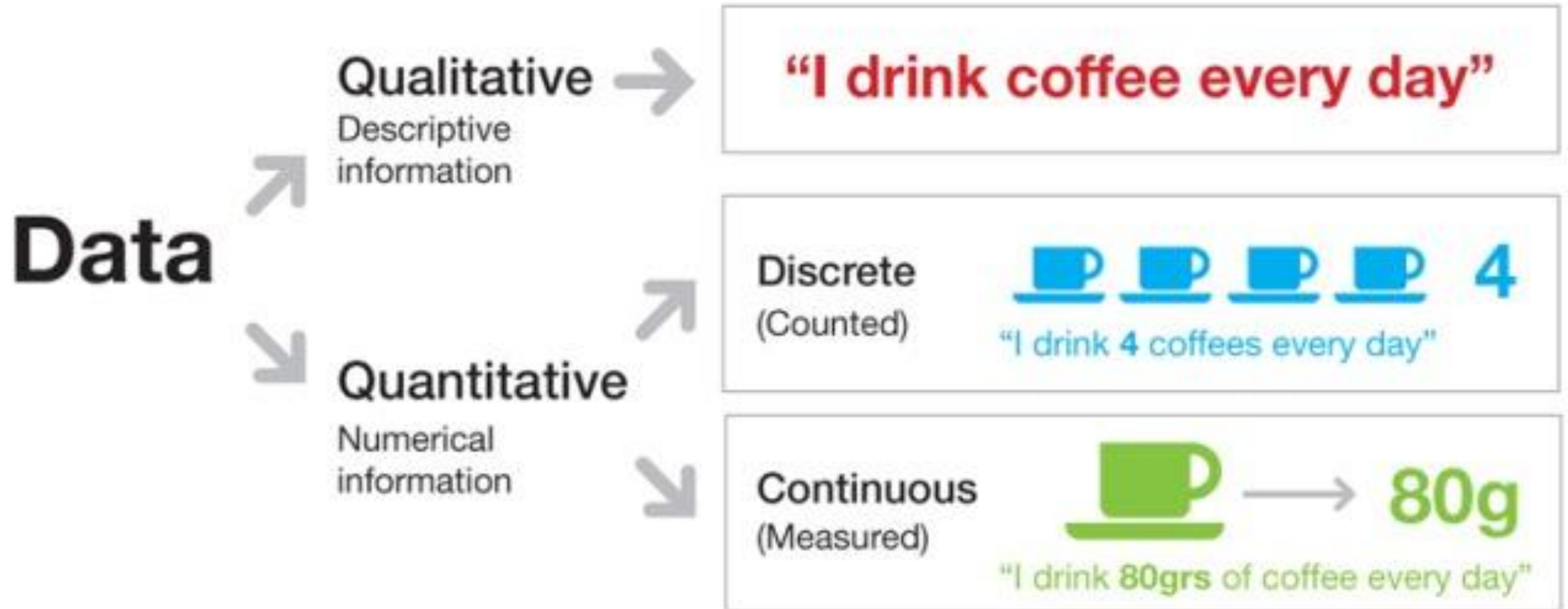
Juan Pérez obtuvo mejor nota que el 75% del resto del curso

Juan Pérez es un buen alumno

¿Qué entendemos por dato?

- Colección de hechos tales como números, palabras, mediciones, o solo descripción de cosas.
- Utilizado generalmente para **análisis**.

¿Cómo describimos un dato?



Attributes

<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

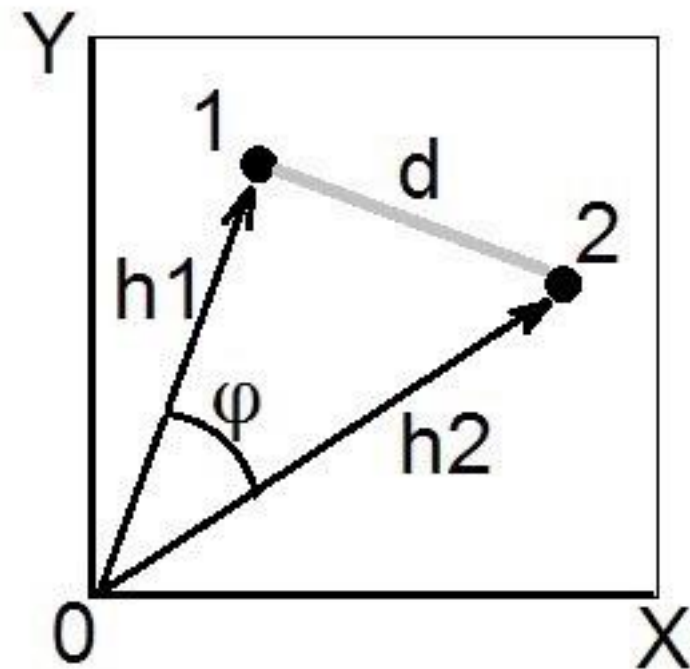
Objects

Estructura básica

Dataset, records, atributos

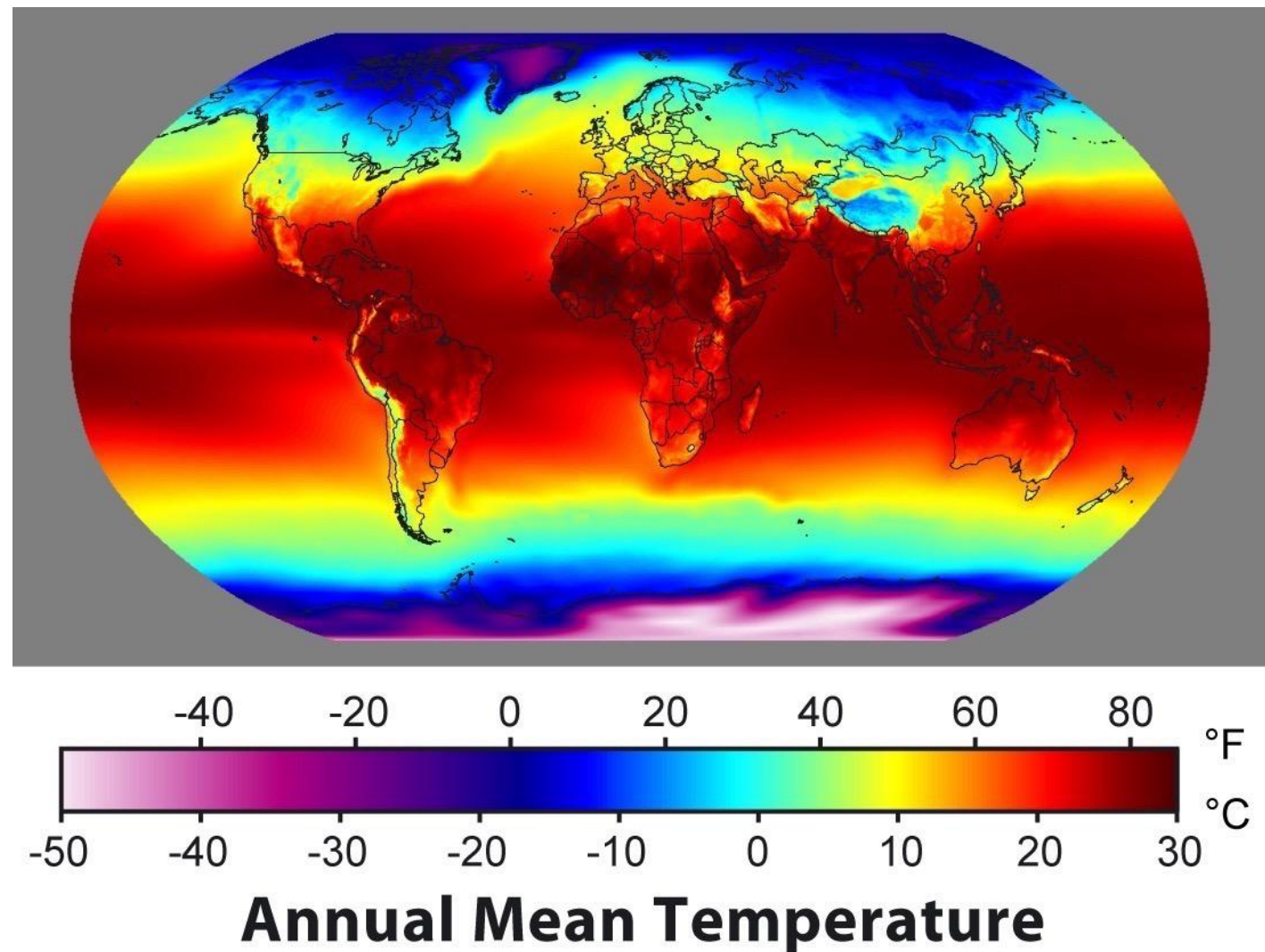
Diferentes enfoques

- Analizar en base a las relaciones entre datos
- Extraer relaciones y luego trabajar con estos valores, no con los datos mismos



Atributos y métricas

- **Atributo:** propiedad que puede variar de un objeto a otro
- **Escala de medición o métrica:** es necesario definirla de forma exacta, para poder comparar.



TIPOS DE ATRIBUTOS

Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

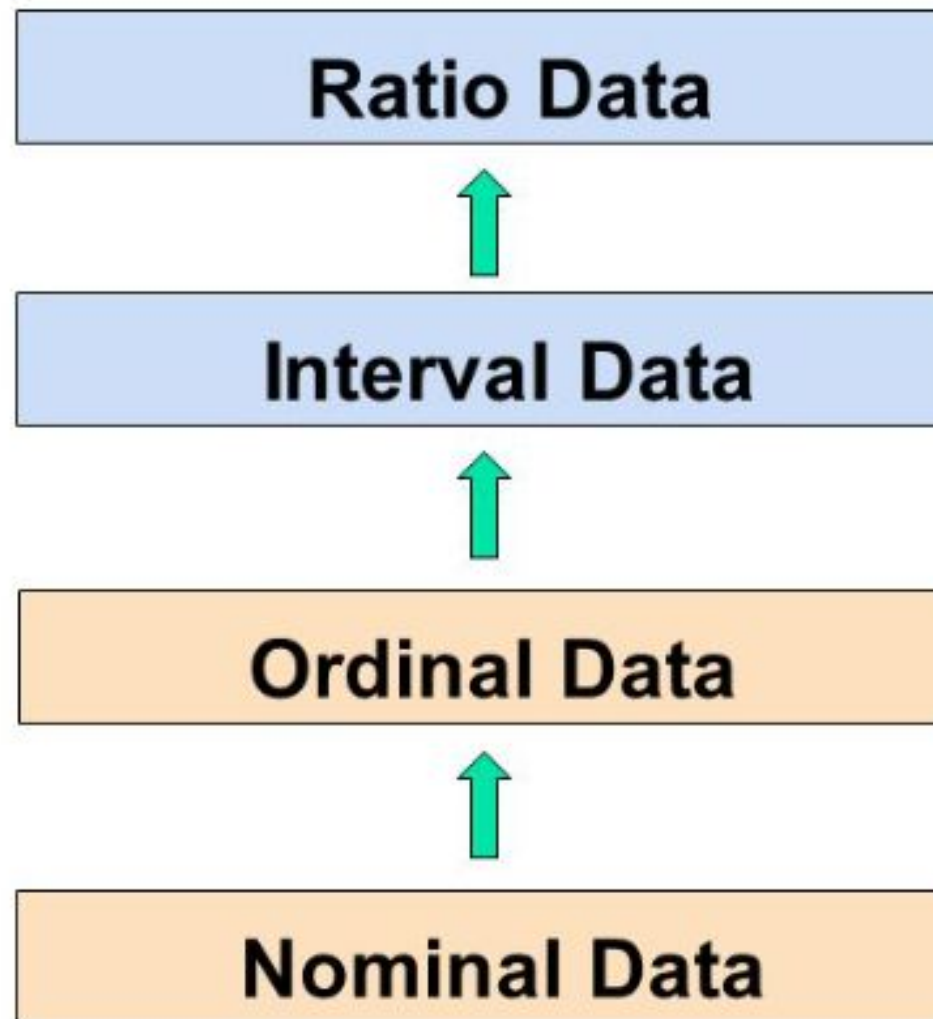
Ordered Categories (rankings, order, or scaling)

Ordinal Data

Qualitative Data

Categories (no ordering or direction)

Nominal Data



TIPOS DE ATRIBUTOS

- **NOMINAL** (IDs, color de ojos, categorías de bacterias)
- **ORDINAL** (rankings, notas, altura en {alto, mediano, bajo})
- **INTERVALO** (Fechas, temperaturas °C o °F)
- **RAZÓN** (Temperatura Kelvin, largo, hora)

TIPOS DE ATRIBUTOS

(Operaciones y propiedades)

- **NOMINAL** (IDs, color de ojos, categorías de bacterias)
Distinción $==$, $!=$ (solo diferencia de nombre)
- **ORDINAL** (rankings, notas, altura en {alto, mediano, bajo})
Orden $<$, $<=$, $>$, $>=$ (permite ordenar los datos)
- **INTERVALO** (Fechas, temperaturas $^{\circ}\text{C}$ o $^{\circ}\text{F}$)
Adición $+$ $-$, grado de diferencia entre medidas pero no la razón entre ellos (el cero es arbitrario).
- **RAZÓN** (Temperatura Kelvin, largo, peso)
Multiplicación $*$ y $/$

- **CUALITATIVOS (Categóricos)**

Nominal y Ordinal, aunque sean numéricos deben ser tratados como símbolos

- **CUANTITATIVOS (Numéricos)**

Intervalo y Razón deben ser tratados como números y tienen propiedades numéricas (continuos o discretos)

Por qué la temperatura en Kelvin es Razón y Celsius es Intervalo?

Operaciones o propiedades

1. **Distinción** $=$ y \neq
2. **Orden** $<$, \leq , $>$ y \geq
3. **Adición** $+$ y $-$
4. **Multiplicación** $*$ y $/$

- NOMINAL (1)
- ORDINAL (1,2)
- INTERVALO (1,2,3)
- RAZÓN (1,2,3,4)

- **Atributos DISCRETOS:**

- Atributos categóricos.
- A menudo representados como enteros
- binarios (finitos)
- Ej: códigos postales o conjunto de palabras en docs.

- **Atributos CONTINUOS:**

- Números reales o valores flotantes
- Ej: temperatura, altura, peso (infinitos)

Formatos de datos

Datos estructurados

Datos que han sido ***formateados y modelados*** para un “fácil” acceso.

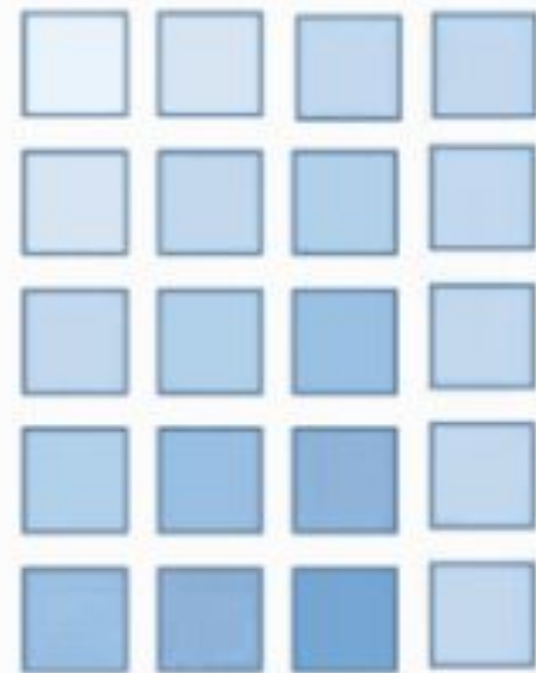
Usualmente se organizan en **tablas**.

Poseen una estructura fija mediante **columnas**.

Un nuevo valor se representa como una **fila**.

Ejemplo: excel, .csv o tablas relacionales (SQL)

Structured Data





Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7
Tony	Felino	DPC	0.5	Macho	2
Lorenzo	Canino	Mestizo	3	Macho	1.5

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7
Tony	Felino	DPC	0.5	Macho	2
Lorenzo	Canino	Mestizo	3	Macho	1.5

¿Qué **problemas** podríamos tener con **datos** en este formato?

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7
Tony	Felino	DPC	0.5	Macho	2
Lorenzo	Canino	Mestizo	3	Macho	1.5

Agregar nuevas columnas:

- Chip
- Celo
- Castración
- Esterilización
- Crías

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso	Celo	Alergias
Doris	Canino	Schnauzer	3	Hembra	8	NO	SI
Clotilde	Canino	Mestizo	14	Hembra	7	SI	NO
Tony	Felino	Ángora	0.5	Macho	2	??	??
Lorenzo	Canino	Mestizo	3	Macho	1.5	??	SI

Agregar nuevas columnas:

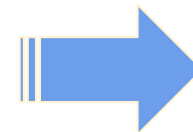
- Chip
- Celos
- Castración
- Esterilización
- Alergias

Datos estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso	Celo	Alergias
Doris	Canino	Schnauzer	3	Hembra	8	NO	SI
Clotilde	Canino	Mestizo	14	Hembra	7	SI	NO
Tony	Felino	Ángora	0.5	Macho	2	??	??
Lorenzo	Canino	Mestizo	3	Macho	1.5	??	SI

Agregar nuevas columnas:

- Chip
- Celos
- Castración
- Esterilización
- Alergias



Muchos valores
VACÍOS

Formatos de datos

Datos semi-estructurados

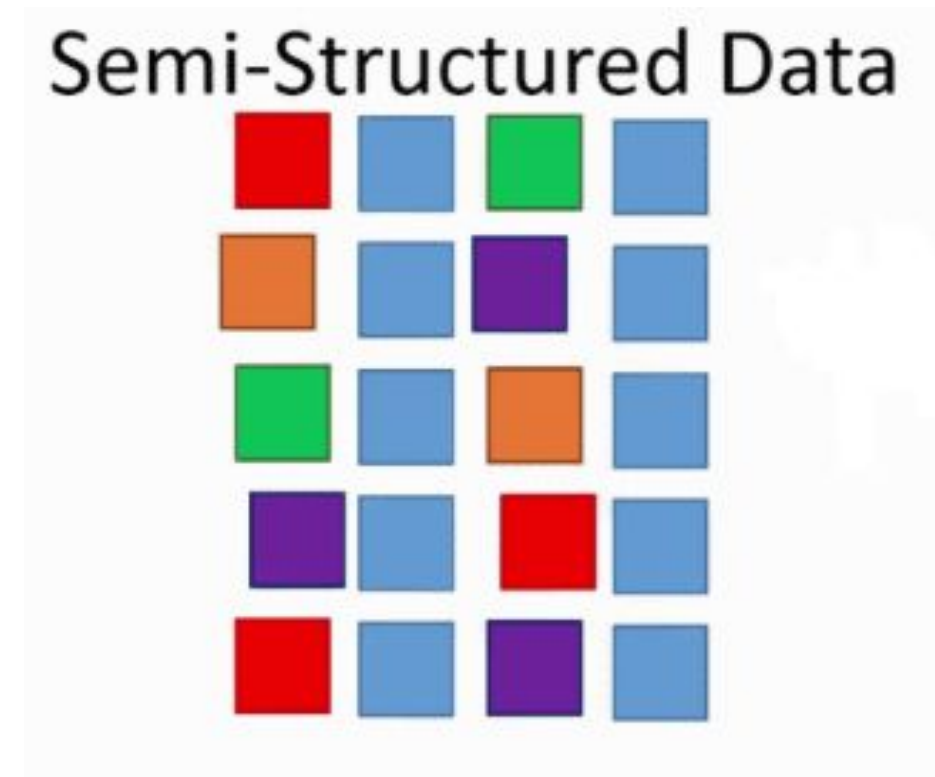
Datos que sigue una estructura más **dinámica** y **flexible**.

Usualmente se organizan de manera jerárquica.

No necesita modificar la estructura completa al agregar un nuevo atributo.

En lugar de utilizar columnas, se usan “**tags**” para representar atributos

Ejemplos: XML, json, e-mails, algunas web



Datos semi-estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7



XML

```
<mascota>
  <nombre>Doris</nombre>
  <especie>Canino</especie>
  <raza>Schnauzer</raza>
  <edad>3</edad>
  <color>Sal y pimienta</color>
  <peso>8</peso>
  <celo>SI </celo>
</mascota>
```

json

```
{
  "nombre": "Doris",
  "especie": "Canino",
  "raza": "Schnauzer",
  "edad": "3",
  "color": "Sal y pimienta",
  "peso": "8"
}
```

Formatos de datos

Datos NO estructurados

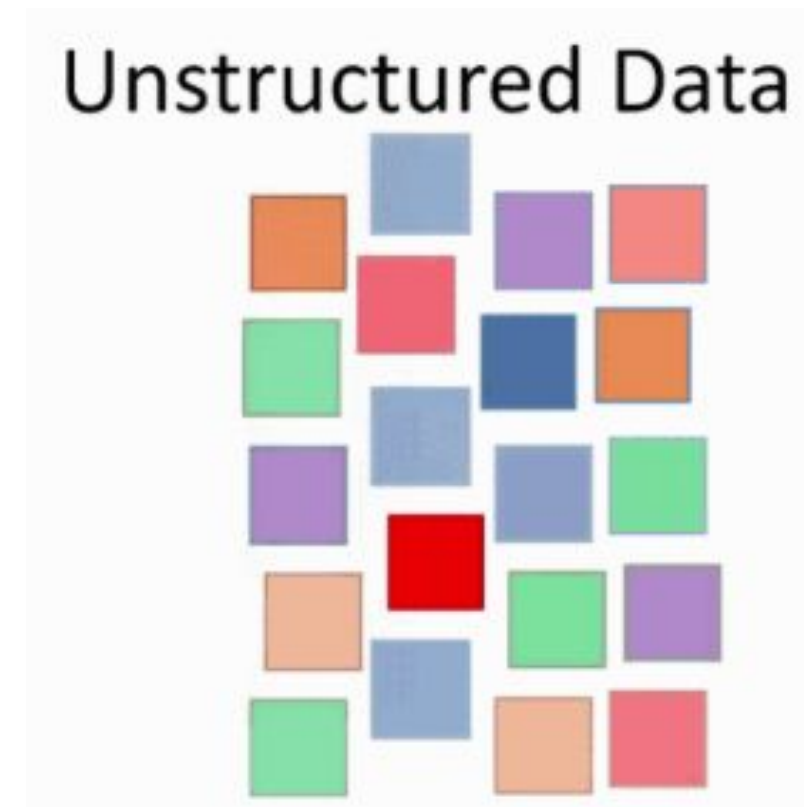
Datos **no** poseen estructura alguna **previamente** definida.

Es posible encontrar **patrones** de cómo están organizados los datos.

Sin embargo, involucra mucho trabajo de pre-procesamiento de datos.

80% de los datos están en este formato.

Ejemplos: documentos de texto, imágenes, video, audio, etc.



Datos NO estructurados

Nombre mascota	Especie	Raza	Edad	Sexo	Peso
Doris	Canino	Schnauzer	3	Hembra	8
Clotilde	Canino	Mestizo	14	Hembra	7



Se ha ingresado el paciente **Doris**, de especie **canino**. Nació aproximadamente el 2017, por lo que tiene **3 años**. Su manto es color **sal y pimienta** y patas blancas. Este es su primer control, pesando **8kg** aproximadamente.



NO Estructurados

Semi Estructurados

Estructurados



(-) formato

(+) formato



(+) cantidad

(-) cantidad



(+) pre-procesamiento

(-) pre-procesamiento

Datos tipo records

<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

Datos tipo records

	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

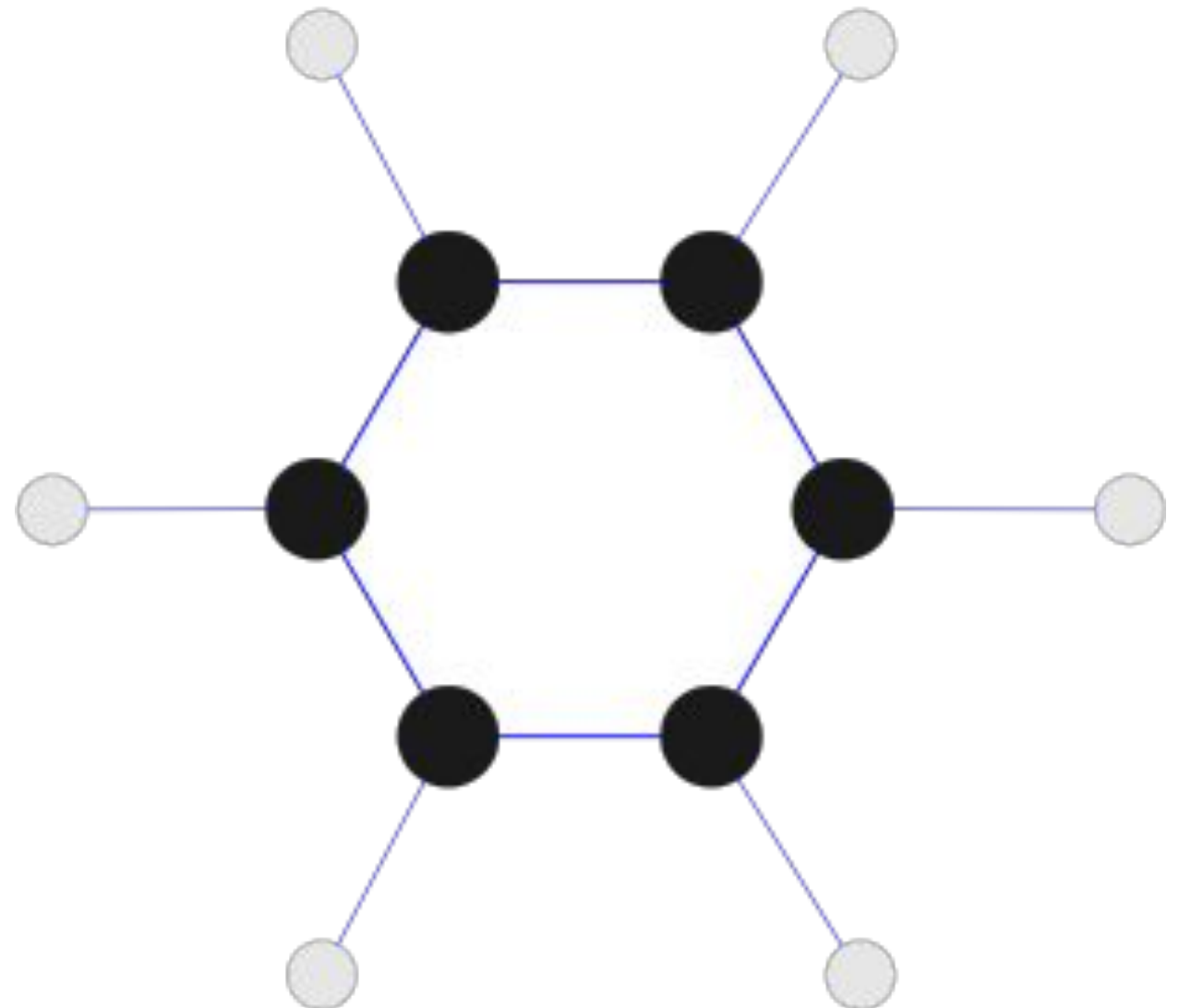
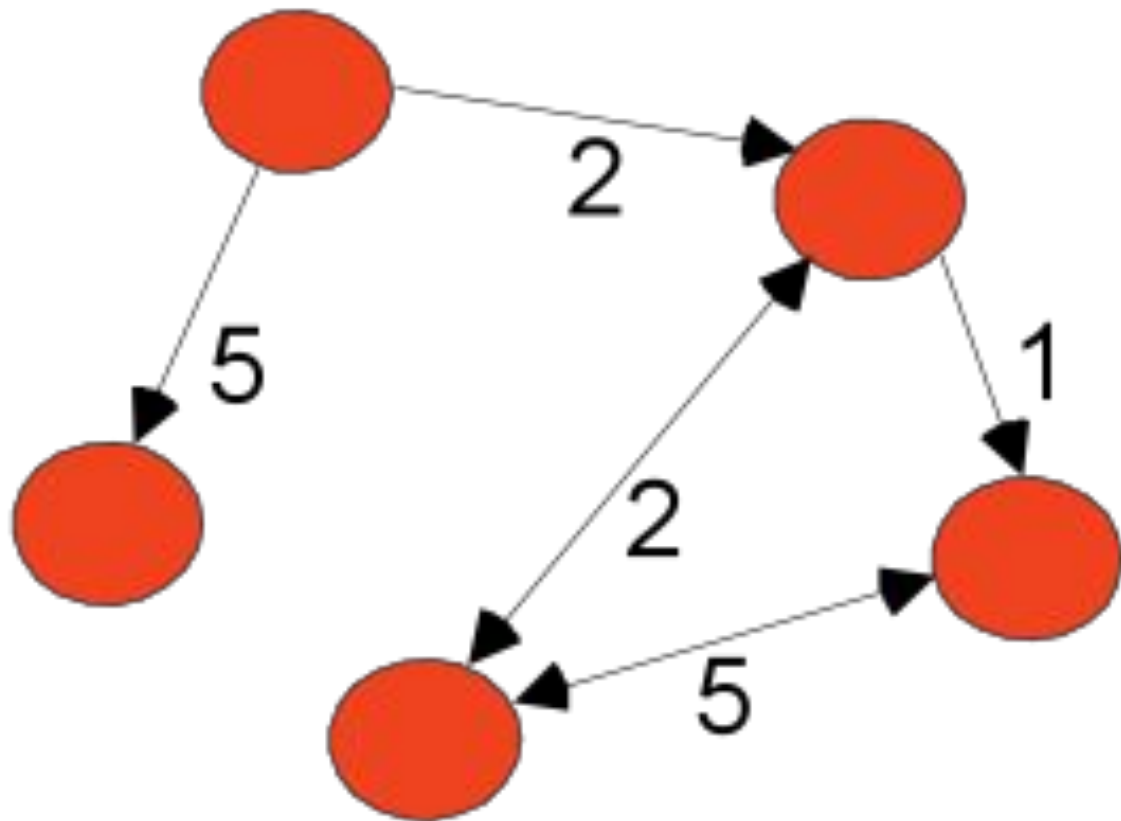
Datos tipo records

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Datos tipo records

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 tipo grafos



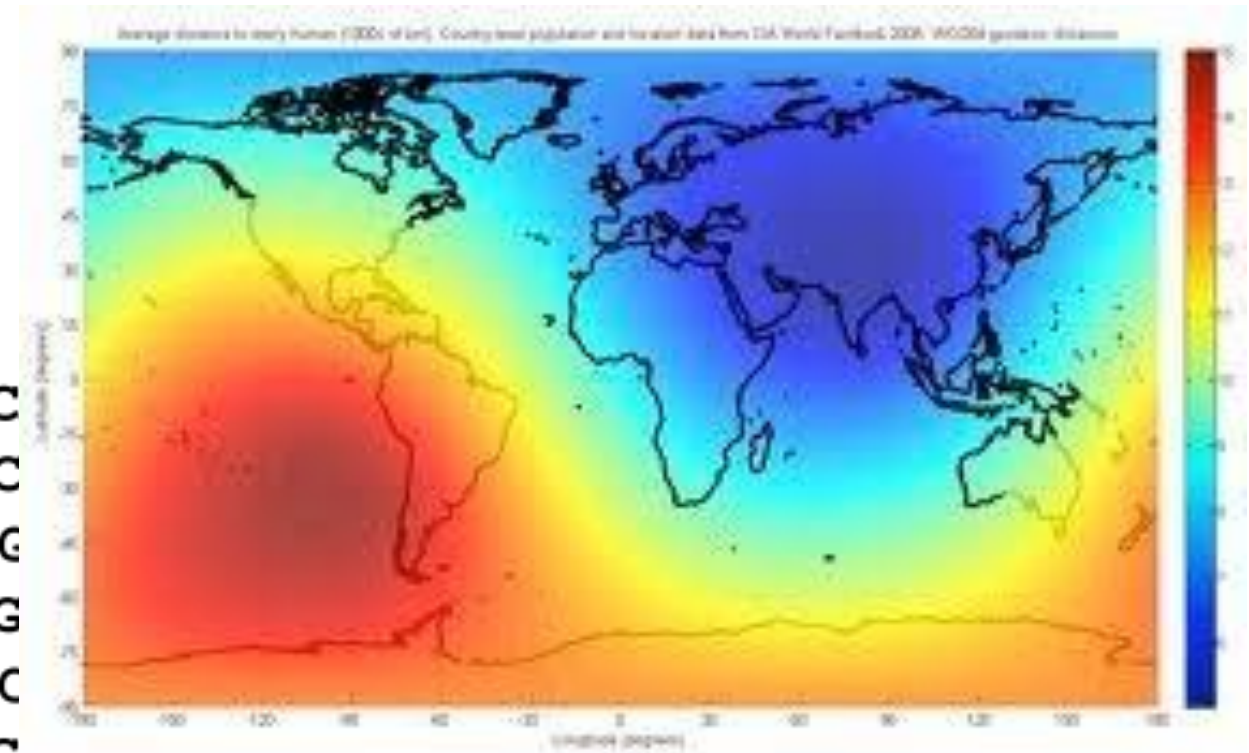
Datos ordenados

Items/Events

(A B) (D) (C E)
 (B D) (C) (E)
 (C D) (B) (A E)

An element of
the sequence

GGTTC CGCCTTC
 CGCAGGGCCCGC
 GAGAAGGGCCCG
 GGGGGAGGCGGG
 CCAACCGAGTCC
 CCCTCTGCTCGG
 GCTCATTAGGCGGCAGCGGACAG
 GCCAAGTAGAACACGCGAAGCGC
 TGGGCTGCCTGCTGCGACCAGGG



- 1) **datos secuenciales** (transacciones con tiempo asociado: libros embarazo -> pañales)
- 2) **secuencias datos ordenados pero sin tiempo** (ADN, secuencias de genes, etc.)
- 3) **Datos ordenados en el espacio**
- 4) **Datos ordenados en el tiempo:** series de tiempo (fluctuaciones de la bolsa)
- 5) **Autocorrelación temporal:** objetos cercanos en el tiempo son parecidos (mediciones temp en 2 minutos continuos)
- 6) **Autocorrelación espacial:** obj. cercanos en el espacio son parecidos (i.e., ley del metal)

Características Generales (sets de Datos)

- **Dimensionalidad:** nro. de atributos, maldición de la dimensionalidad (curse of dimensionality) tiene que ver con problemas al trabajar con muchas dimensiones (preprocesamiento: reducción de dimensionalidad)
- **Dispersión:** mayoría de las dimensiones son 0 para los datos, puede tener ventajas, como no necesitar almacenar los valores 0, sólo los 1s. (Ej. Grafo de la Web y sus enlaces).
- **Resolución** (ej. variaciones de presión atmosférica en horas es notoria, pero en meses no se detecta).

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

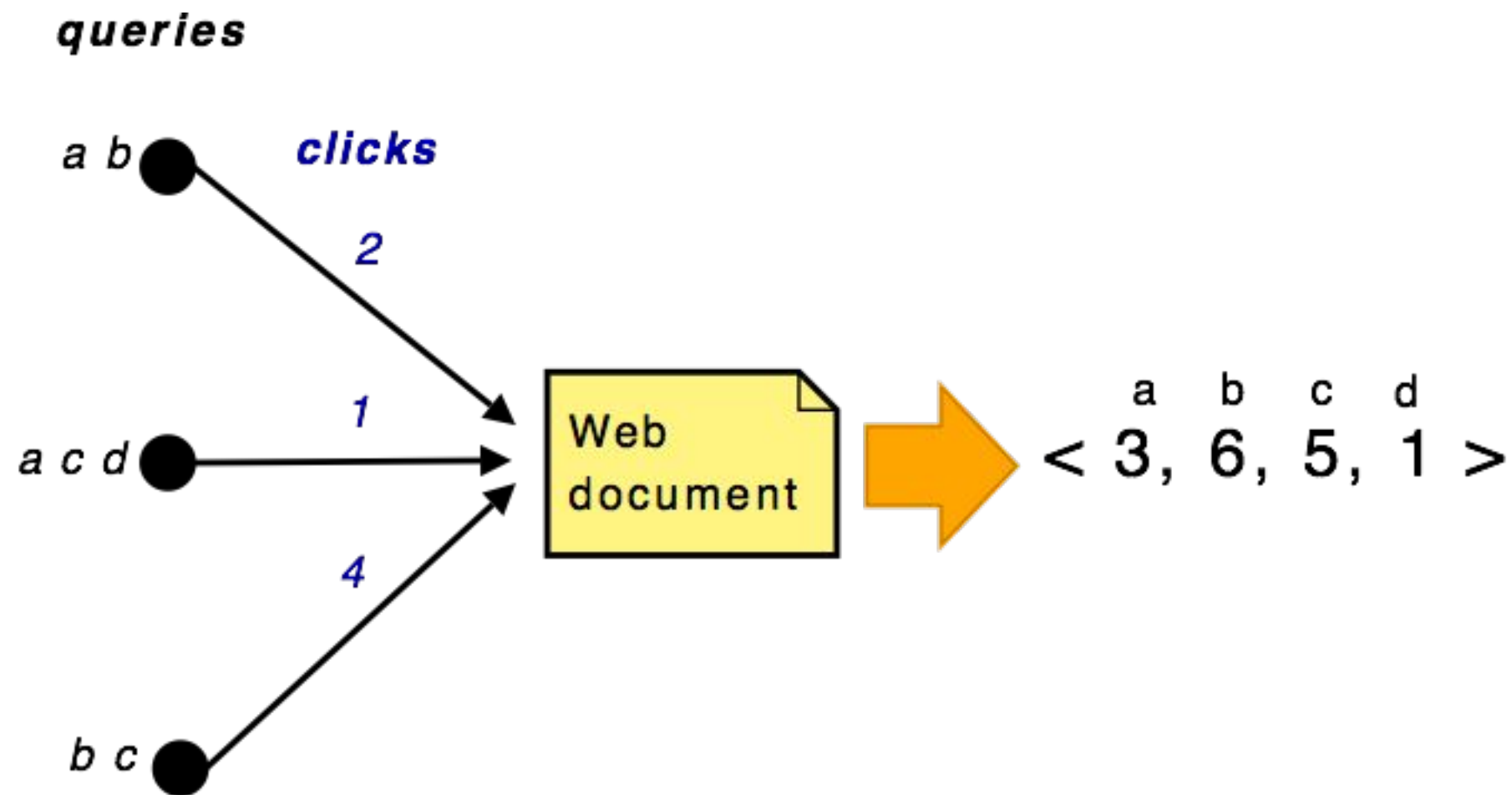
Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to



Reducción de dimensionalidad
(DOCUMENTO WEB)

Calidad de los datos

- No poseen la calidad deseada a priori, los algoritmos de DM se enfocan en:
 1. Detección y corrección de problemas de calidad
 2. Usar algoritmos que toleren datos de poca calidad
- i.e., limpieza de datos

Calidad de los datos

Time spent in a life of a Data Scientist

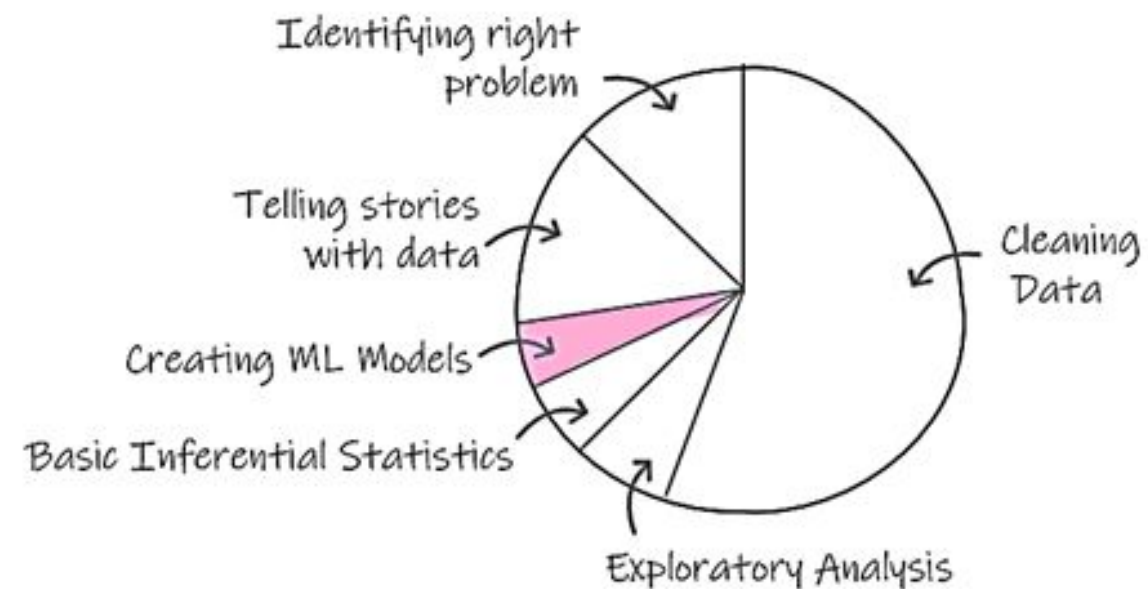
@datavizzdom

Gulrez

Perception



Reality



¿Por qué se producen errores?

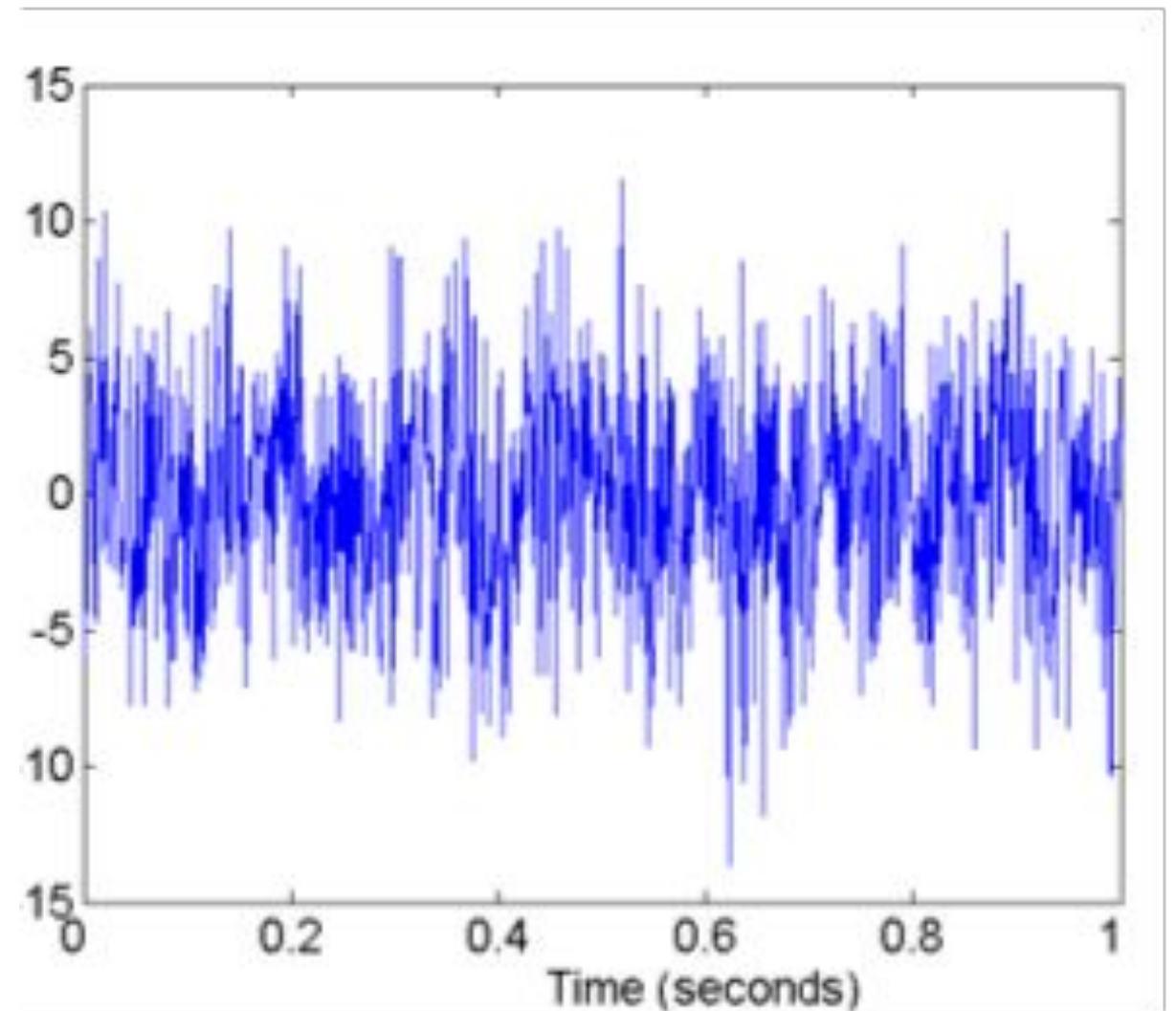
Tipos de errores:

- Ruido y outliers
- Valores faltantes
- Datos duplicados



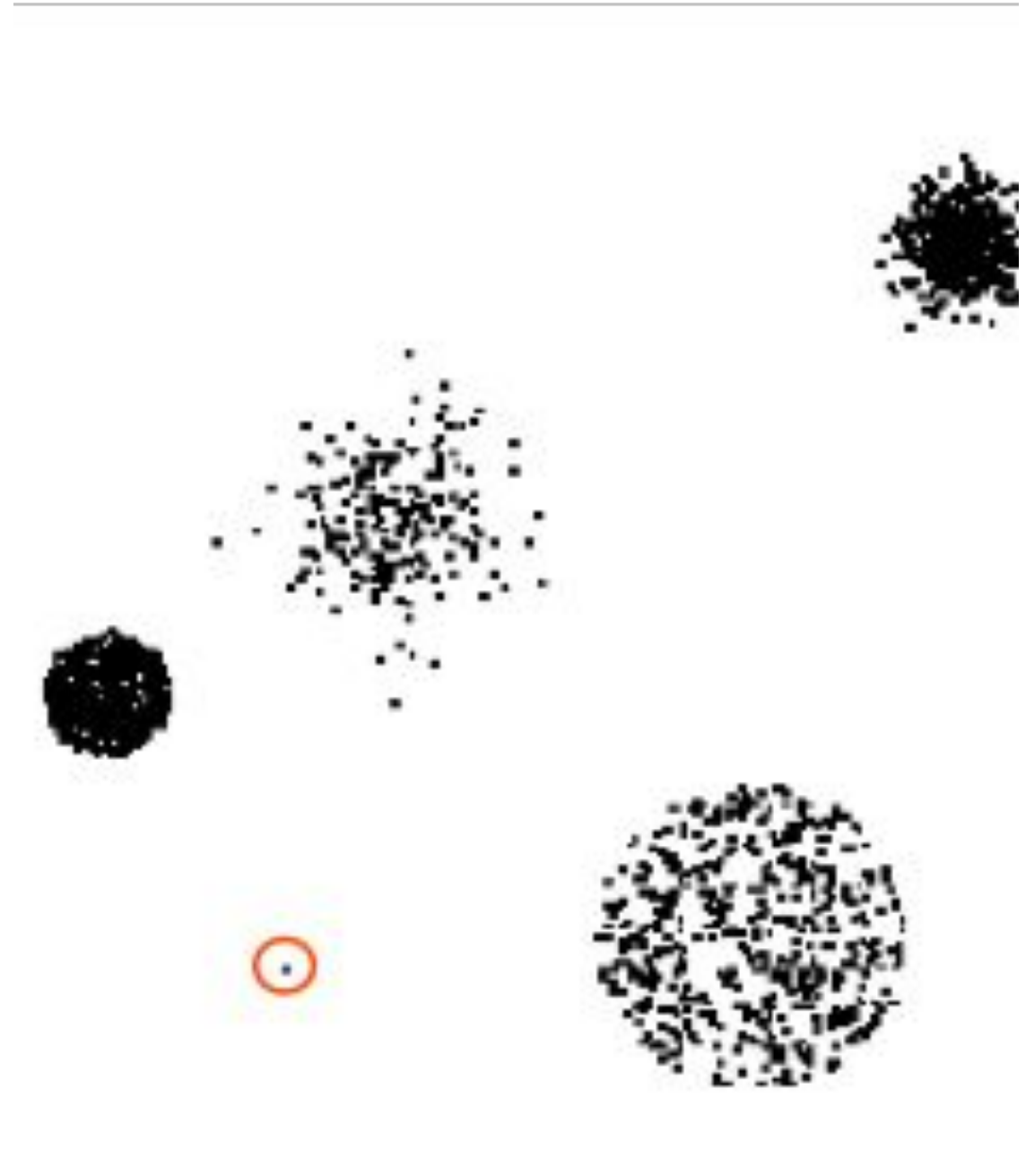
¿Qué es el ruido?

- Componente aleatoria en la medición (distorsión de voz en un teléfono malo)
- Datos espaciales, temporales



Outliers

- Objetos con características considerablemente diferentes a la mayoría



Valores faltantes

- ¿Motivos?
- Información no recolectada (e.j: no quieren dar edad y/o peso)
- Atributos no aplicables a todos (e.j: impuesto en niños)

Valores faltantes...

- ¿Cómo los manejo?
- Eliminando el objeto
- Estimando (interpolando) valores
- Ignorar



**15.800.000 = Población
efectivamente
censada**



16.600.000 = Población estimada
incluye
Población efectivamente censada
+
Moradores ausentes
(670.772 Hab.)

censo 2012

Valores inconsistentes

- Datos mal ingresados

Datos duplicados

- El dataset incluye datos duplicados o cuasi-duplicados
- Gran problema al juntar datos de fuentes múltiples
- e.j: RT (casos deseados, no deseados)



Segunda Parte

Calidad de los datos

- No poseen la calidad deseada a priori, los algoritmos de DM se enfocan en:
 1. Detección y corrección de problemas de calidad
 2. Usar algoritmos que toleren datos de poca calidad
- i.e., limpieza de datos

¿Por qué se producen errores?

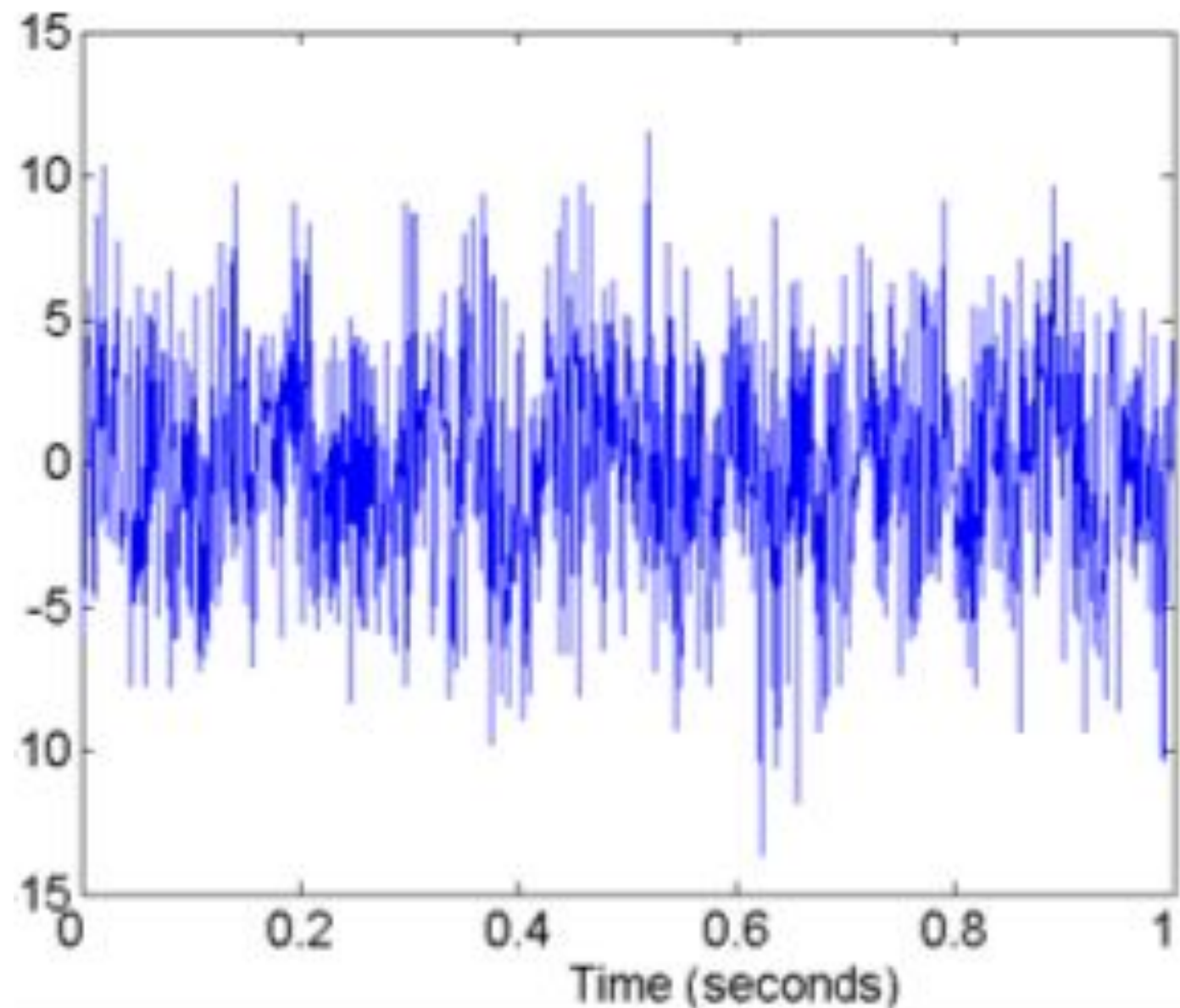
Tipos de errores:

- Ruido y outliers
- Valores faltantes
- Datos duplicados
- Sesgo (Bias)



¿Qué es el ruido?

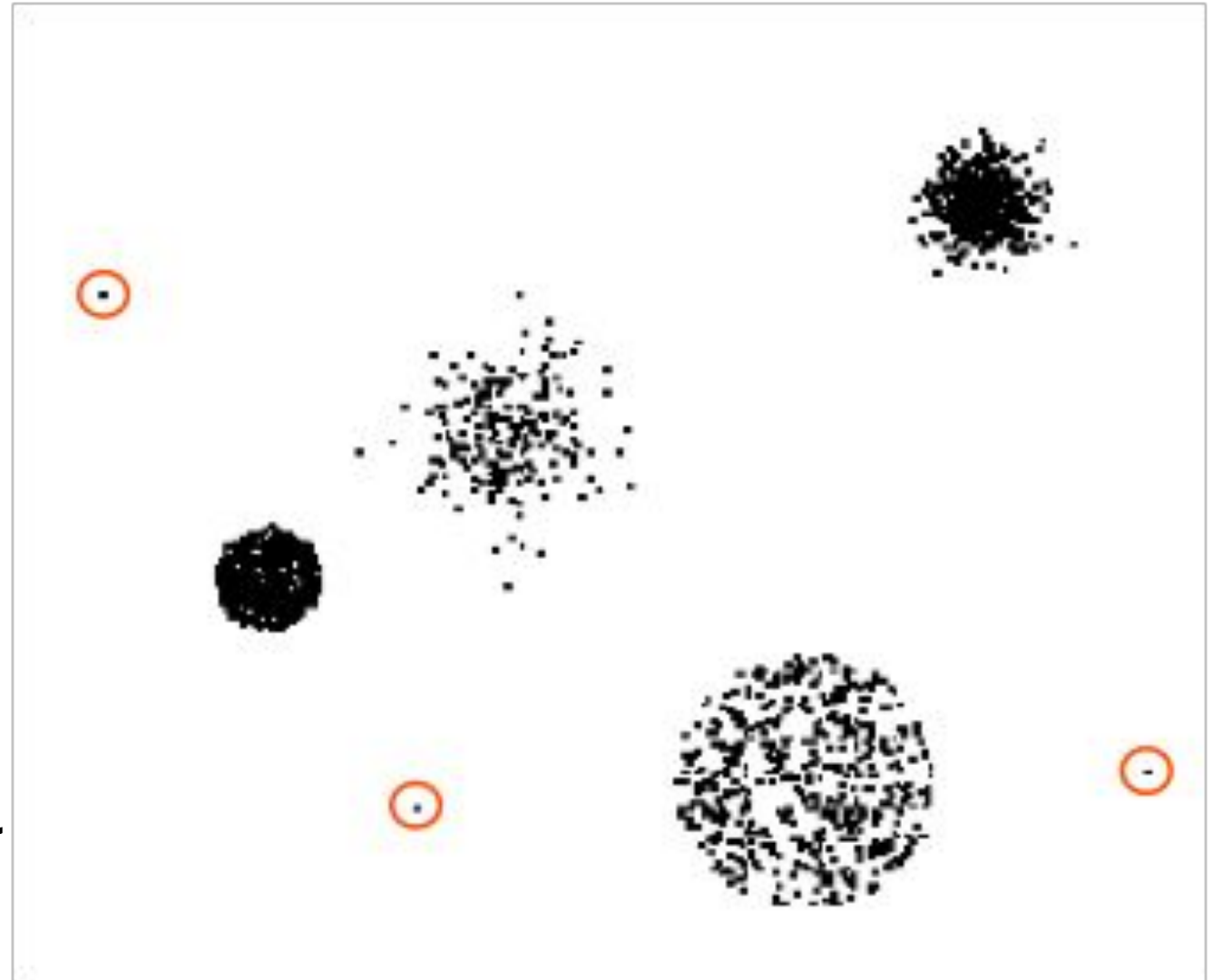
- Componente aleatoria en la medición (distorsión de voz en un teléfono malo)
- Datos espaciales, temporales





Outliers

- Objetos con características considerablemente diferentes a la mayoría



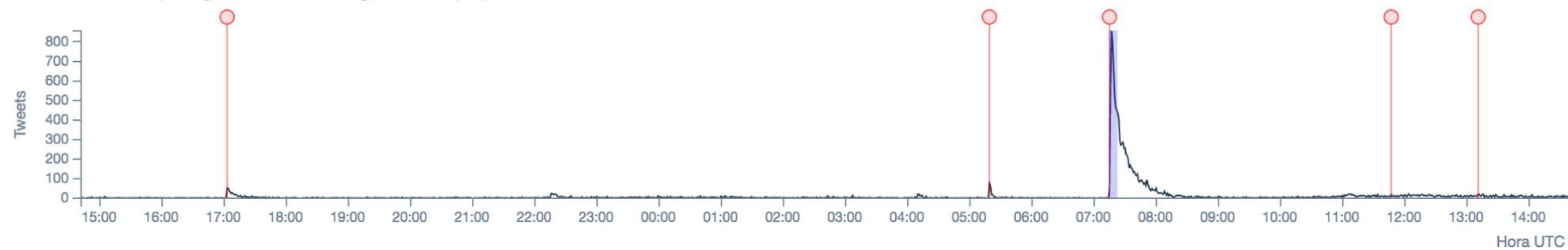
Frecuencia de tweets

REANUDAR

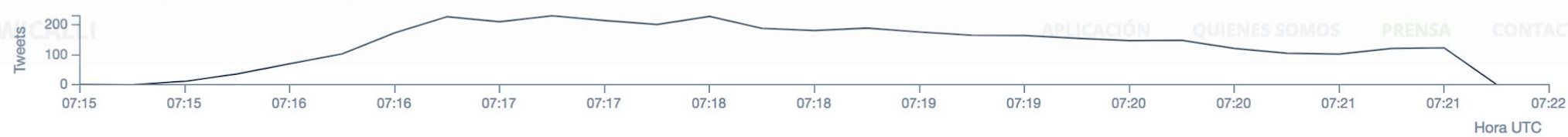
08/02/2017



Últimas 24 horas (01 Agosto 2017 - 02 Agosto 2017)



Ampliación del rango entre las 07:15 y las 07:22 horas



Mapa de Calor



Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Tweets

Ordenar

Lady Lolo 07:21:59 02/08/17
@LadyLoloA
Un terrible temblor... ojala no sigan mas!!
Sobretudo por los que viven en malas
condiciones.



Xavier M.B. 07:21:59 02/08/17
@Xavier_Ever
Solo me despierte para escribir que
temblo asique ahora seguire durmiendo
#Temblor



Andriu 07:21:59 02/08/17
@andres_munoza
Ojala nunca pasen un temblor volados
ctm, estoy mal



Claudia Escalera Ch. 07:21:59 02/08/17
@Claudita_34
Fuerte el temblor! Largo y ruidoso
#canal24horas



sachi 07:21:59 02/08/17
@nn_sachiyo
jesus christ there was a temblor and our
windows where this || close to break apart



Tweets con geolocalización

Filtrar marcadores



Información

La información aquí presentada no es de carácter oficial. Para obtener información acerca de sismos en Chile por favor dirigirse a la página del [Centro Sismológico Nacional de la Universidad de Chile](#).

Centro de Investigación
de la Web Semántica

- # Pre-procesamiento de datos
- Creación de atributos
 - Selección de un subconjunto de atributos
 - Agregación
 - Normalización
 - Muestreo
 - Reducción de dimensionalidad
 - Discretización y binarización

Aggregación de datos

- Combinar 2 o más atributos (o objetos) en un único atributo (o objeto)
- ¿Propósito?
 - Reducción de datos
 - Cambio de escala
 - Datos más estables



Agave

Super Bowl 47

2/3/2013 6:30pm - 2/3/2013 10:00pm EST

7.9 million tweets

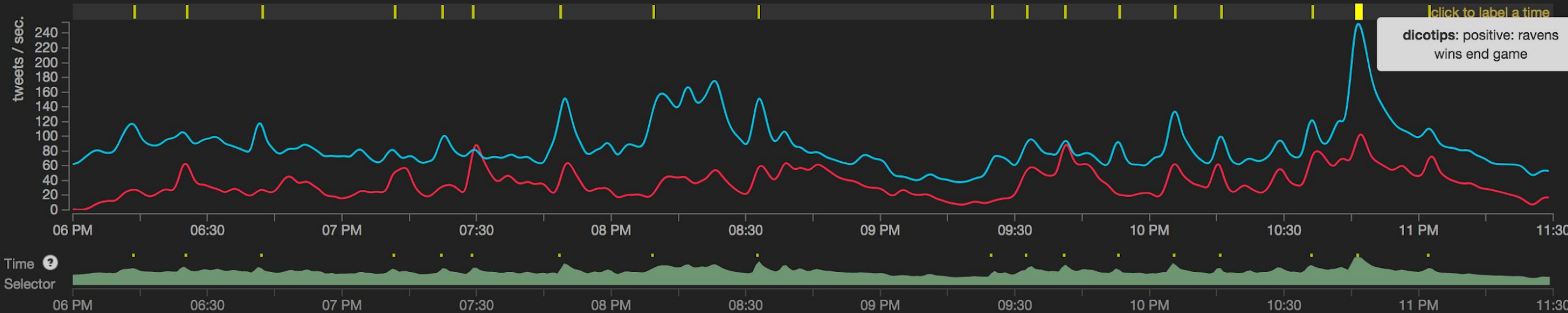
3.8 million authors

Welcome, dicotips!

Sign out

Amount of positive, neutral, and negative tweets over time.

Negative Neutral Positive Combined



Search tweet text



@author



Show RTs



all

Tweets

Top 100 most retweeted Tweets

@AlfredoFlores

16835 retweets

Justin is currently in recovery from an apparent collapse after Beyonce's halftime performance. He wanted me to tell you all he is fine.

@CalebU85

7684 retweets

#RT The reason Why Ray Lewis is Whooping the 49ers is b/c he Joined Body By Vi #superbowl #TOTL #lightsout #bodybyvi <http://t.co/HfuELYIT>

@TheIlluminati

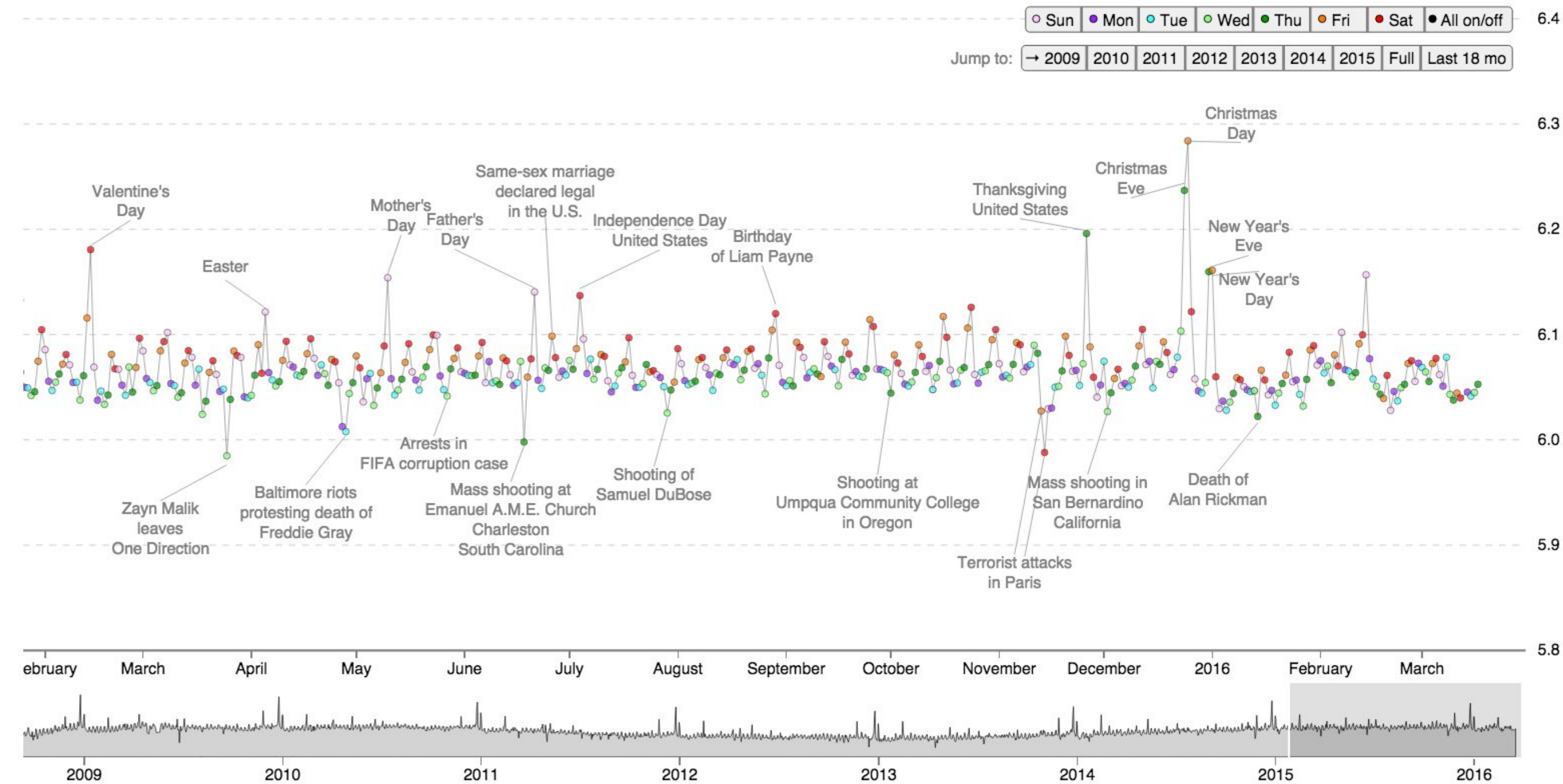
6992 retweets

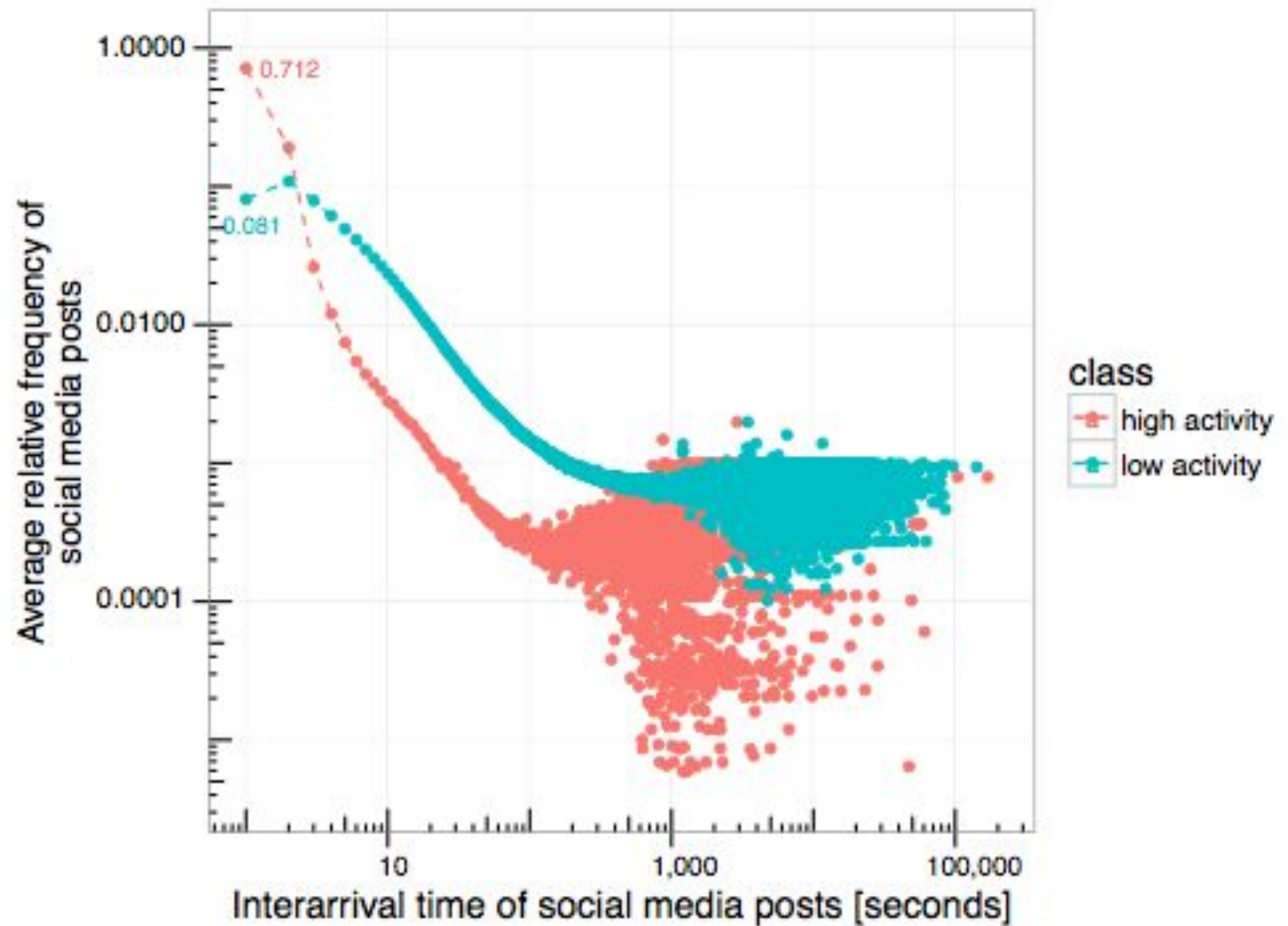
Yes, we turned the lights off at the #Superbowl

@piersmorgan

5453 retweets

Average Happiness for Twitter





Agregación

Muestreo

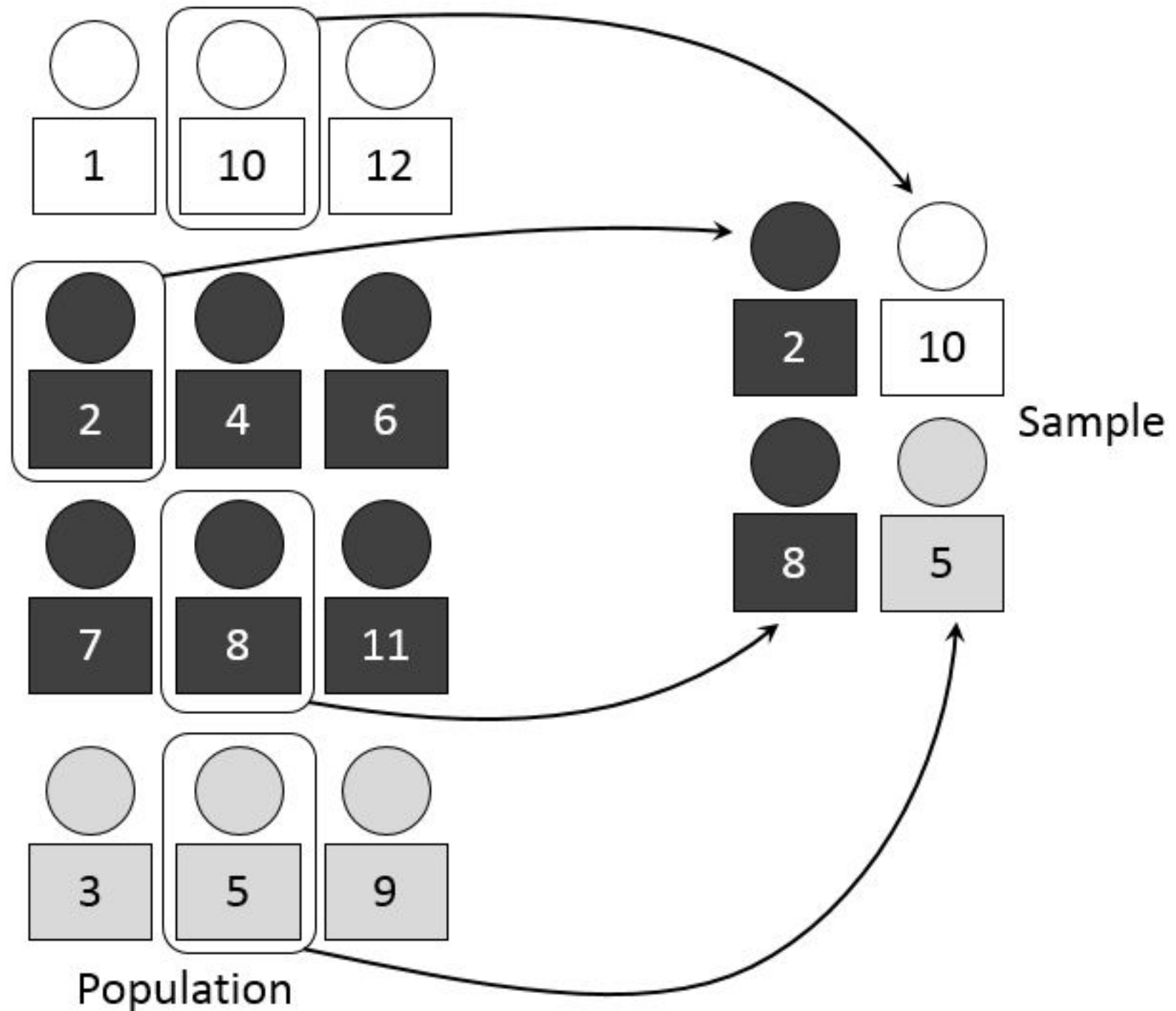
- Principal técnica de selección de datos (investigación preliminar o final)
- Usado en Estadística y DM
- ¿Cuándo es efectivo?

Tipos de Muestreo

- Muestreo aleatorio
- ¿Ventajas?
- ¿Desventajas?

Tipos de Muestreo

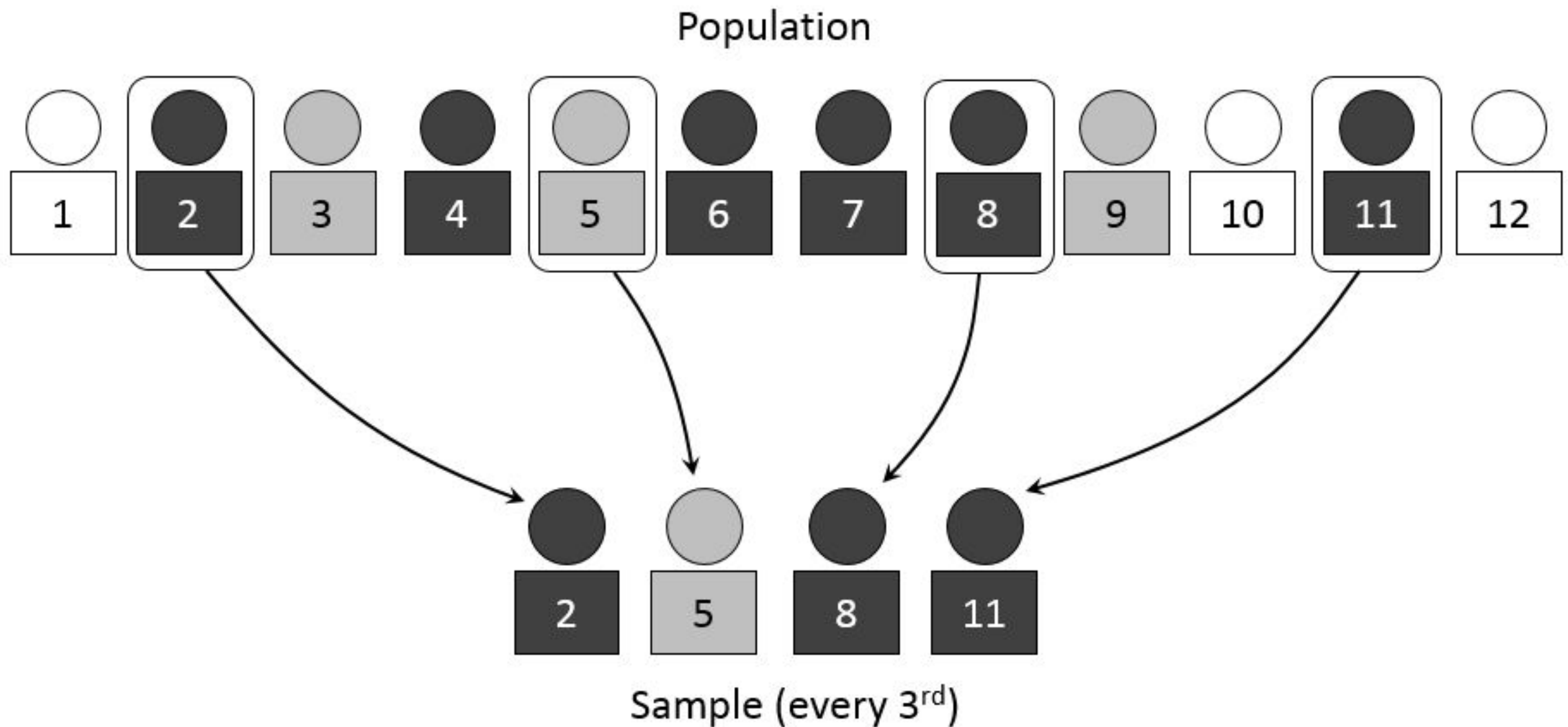
- Muestreo estratificado
- ¿Ventajas?
- ¿Desventajas?



Tipos de Muestreo

- Muestreo aleatorio
- Muestreo estratificado
- Muchos otros tipos de muestreo

Systematic random sample





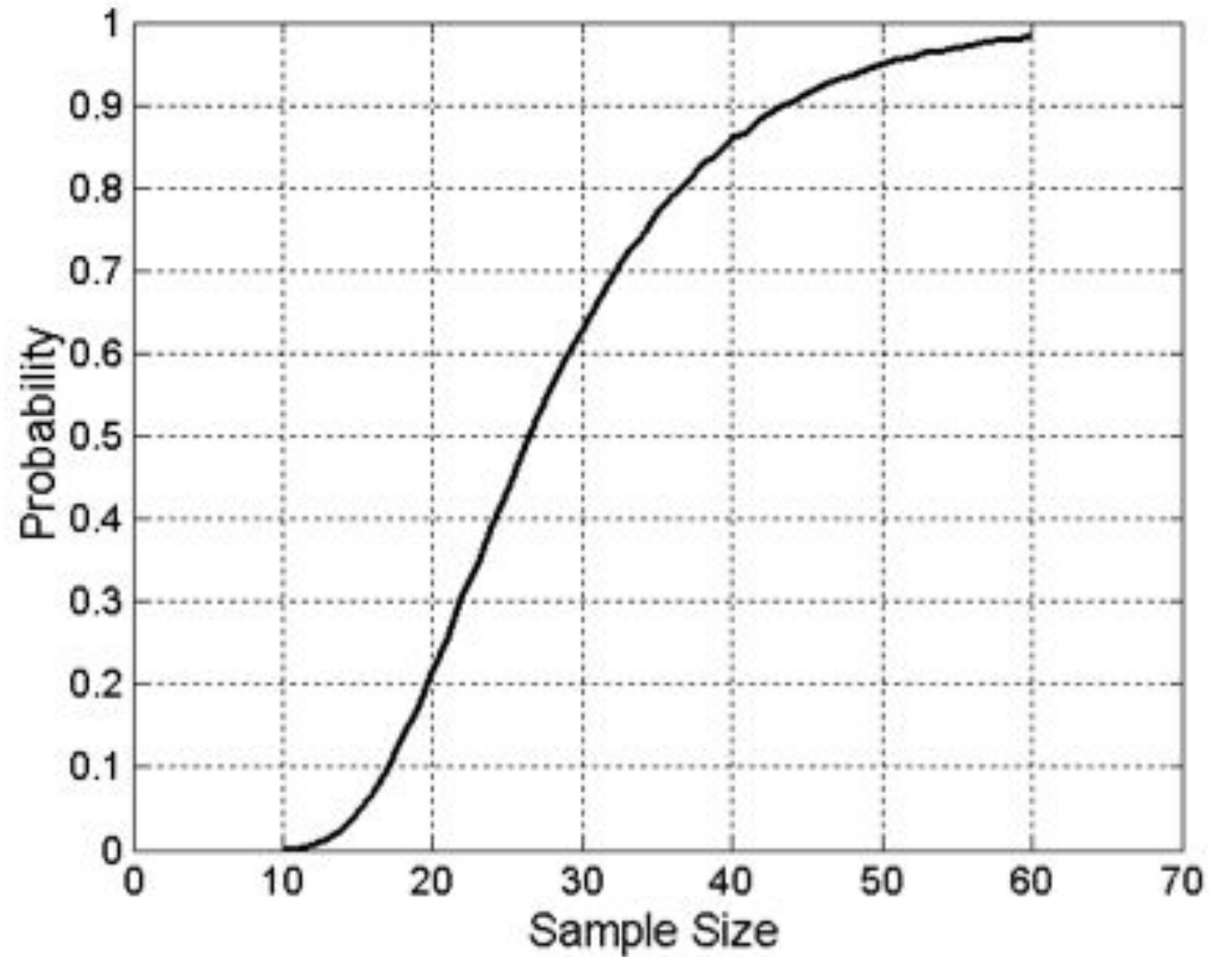
8000 points



2000 Points



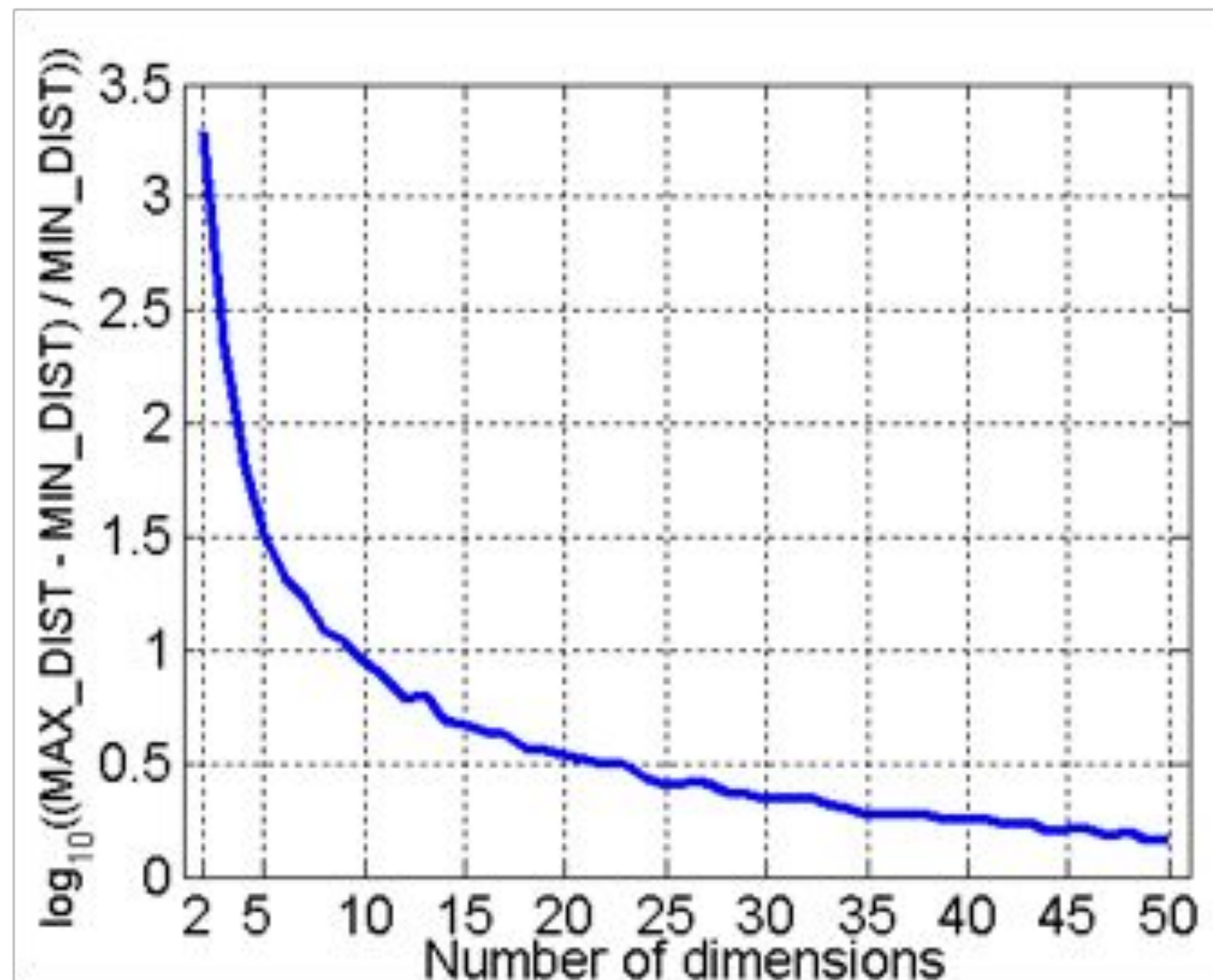
500 Points



- ¿Cómo obtener al menos un objeto de cada uno de los 10 grupos?

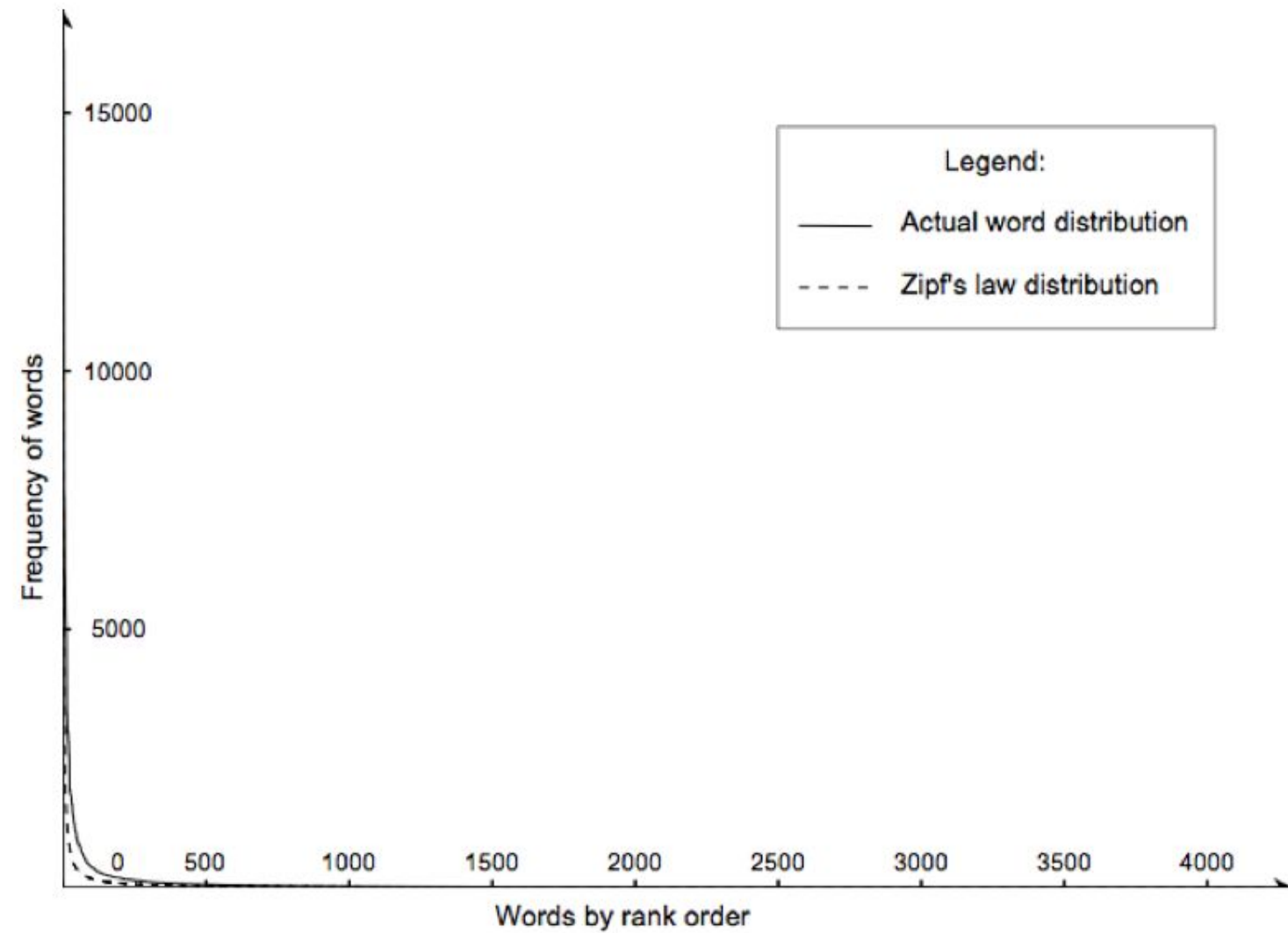
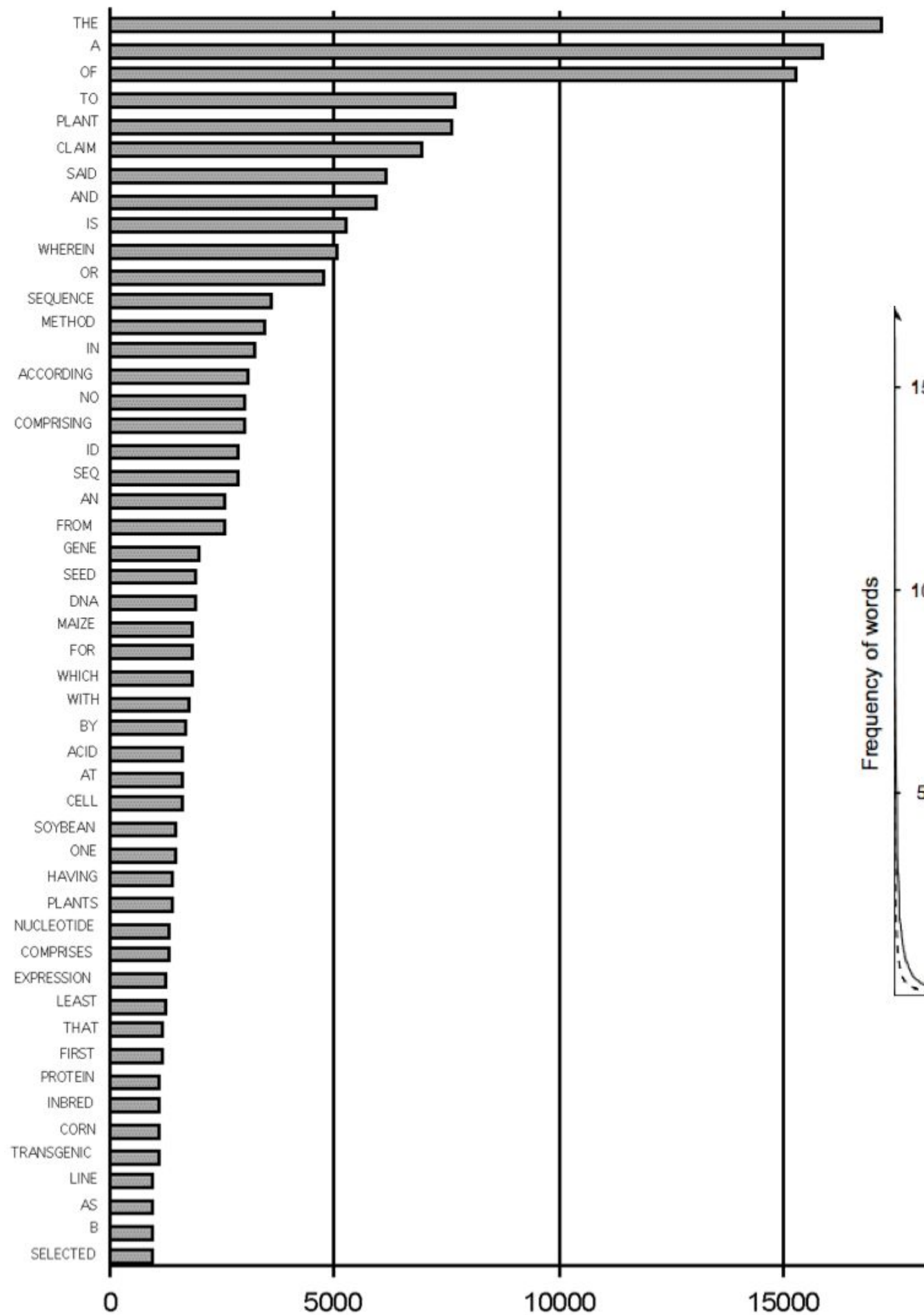
CURSE OF DIMENSIONALITY

- Al aumentar dimensionalidad, los datos se vuelven más dispersos en el espacio
- Pierden significado las medidas, i.e. densidad y distancia entre puntos (clustering y detección de outliers)



Reducción de Dimensionalidad

- ¿Propósito?
- Evitar curse of dimensionality
- Reducir costos asociados a aplicar el algoritmos (tiempo, memoria)
- Mejor visualización de los datos
- Ayuda a quitar atributos irrelevantes o ruidosos



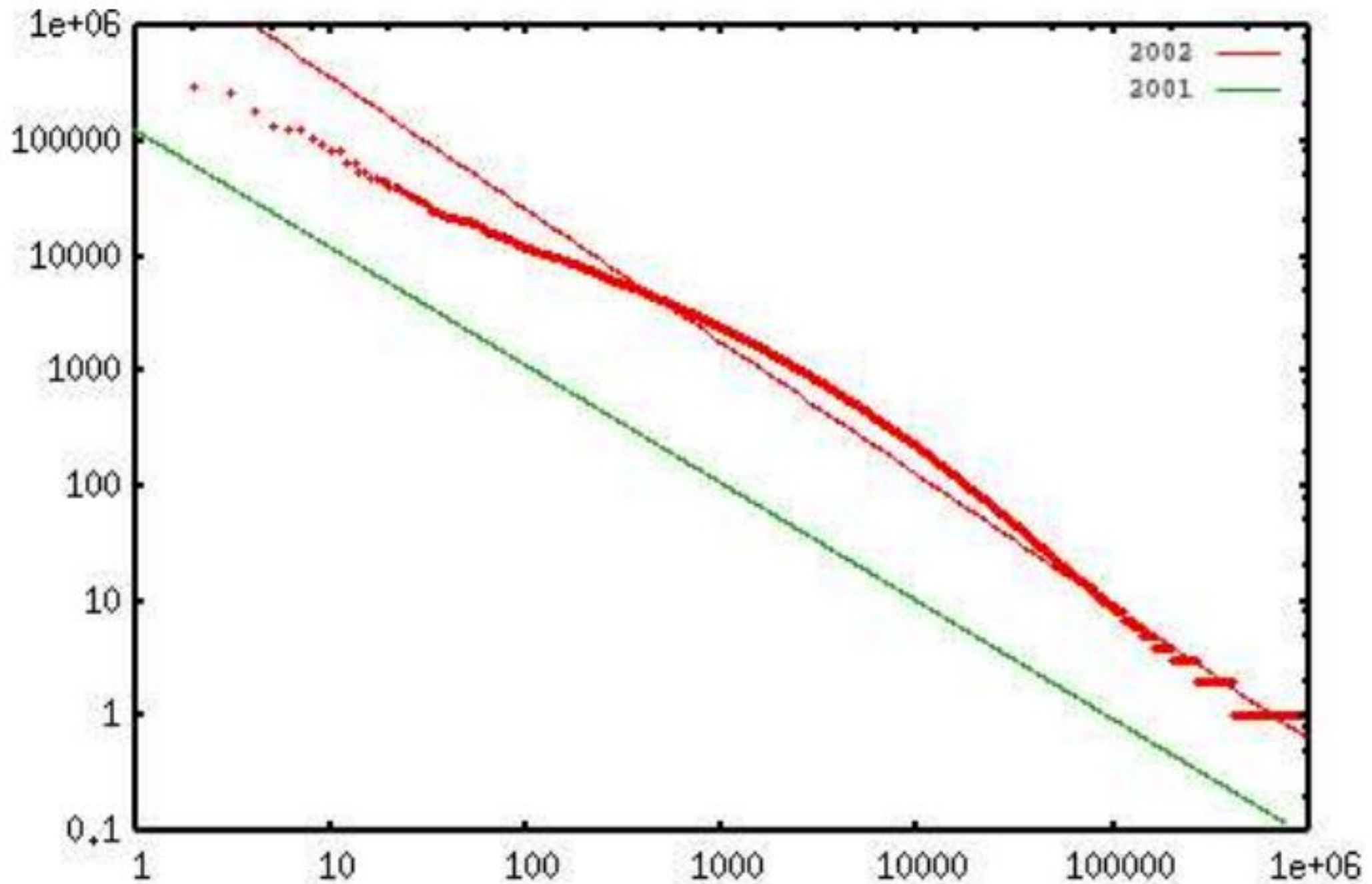


Figura 8: Frecuencia de consulta vs. Palabras.

Agregación

Reducción de Dimensionalidad y Selección de atributos

- Sirven para lo mismo:
 - Selección de atributos
 - Reducción de dimensionalidad

Selección de atributos

- Fuerza bruta: trial and error muchas veces
- Missing Values Ratio
- Low Variance Filter
- High Correlation Filter
- <http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html>

Reducción de Dimensionalidad

- Random Forest/Ensemble Trees
- Backwards/Forward Feature Elimination/Construction (pocas columnas)
- Técnicas de álgebra lineal: PCA, SVD, ISOMAP (nuevo espacio, pierde interpretabilidad)

Dimensionality Reduction	Reduction Rate	Accuracy on validation set	Best Threshold	AuC	Notes
Baseline	0%	73%	-	81%	Baseline models are using all input features
Missing Values Ratio	71%	76%	0.4	82%	-
Low Variance Filter	73%	82%	0.03	82%	Only for numerical columns
High Correlation Filter	74%	79%	0.2	82%	No correlation available between numerical and nominal columns
PCA	62%	74%	-	72%	Only for numerical columns
Random Forrest / Ensemble Trees	86%	76%	-	82%	-
Backward Feature Elimination + missing values ratio	99%	94%	-	78%	Backward Feature Elimination and Forward Feature Construction are prohibitively slow on high dimensional data sets. It becomes practical to use them, only if following other dimensionality reduction techniques, like here the one based on the number of missing values.
Forward Feature Construction + missing values ratio	91%	83%	-	63%	

PCA

- Principal Component Analysis
- Para atributos continuos
- Busca un nuevo set de atributos (componentes principales) que
 1. Son combinaciones lineales de los atributos originales
 2. Son ortogonales (perpendiculares) entre sí
 3. Capturan la máxima variación de los datos

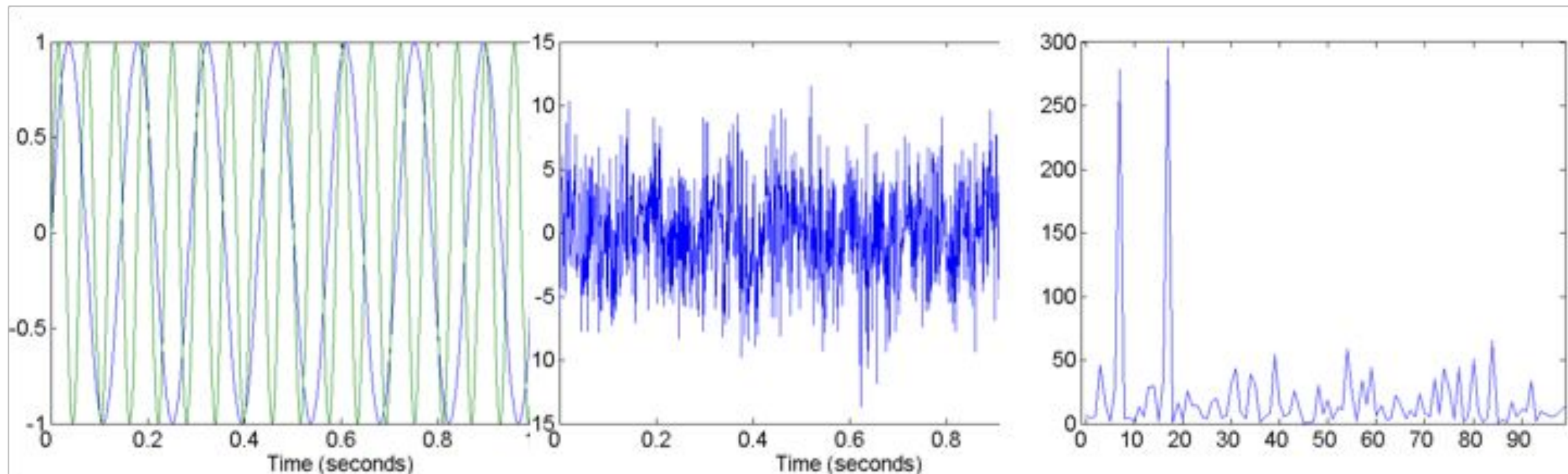
Dar pesos a los atributos

- Se asigna peso a los atributos según su importancia
- SVM lo hace automáticamente
- Normalización

Crear atributos

- Extraer atributos
- Mapear atributos a un nuevo espacio
- Construir atributos

Mapear a un nuevo espacio



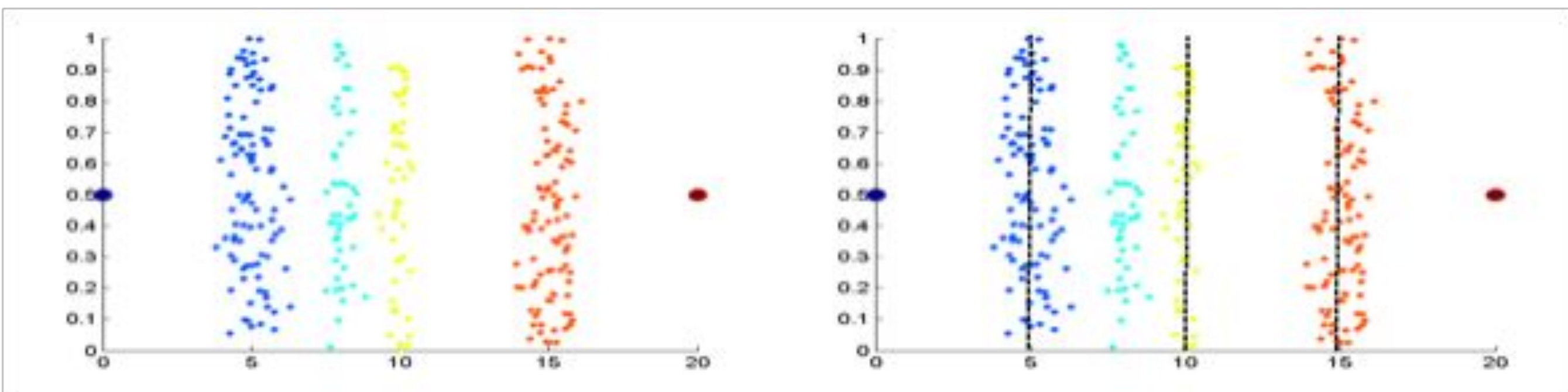
Two Sine Waves

Two Sine Waves +
Noise

Frequency

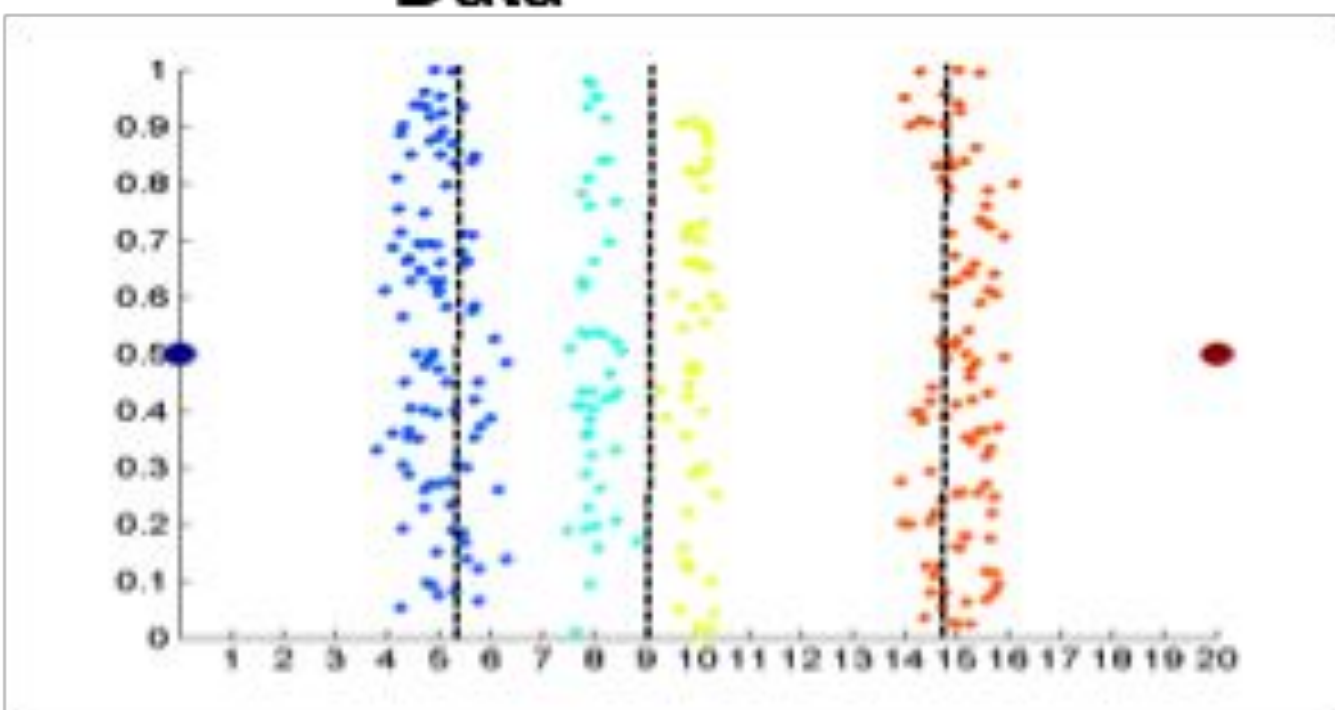
Discretizar

- Decidir cuántas categorías tendremos
- Supervisado (con clases, considerando entropía y pureza)
- y no-supervisado (mismo intervalo, misma cantidad)

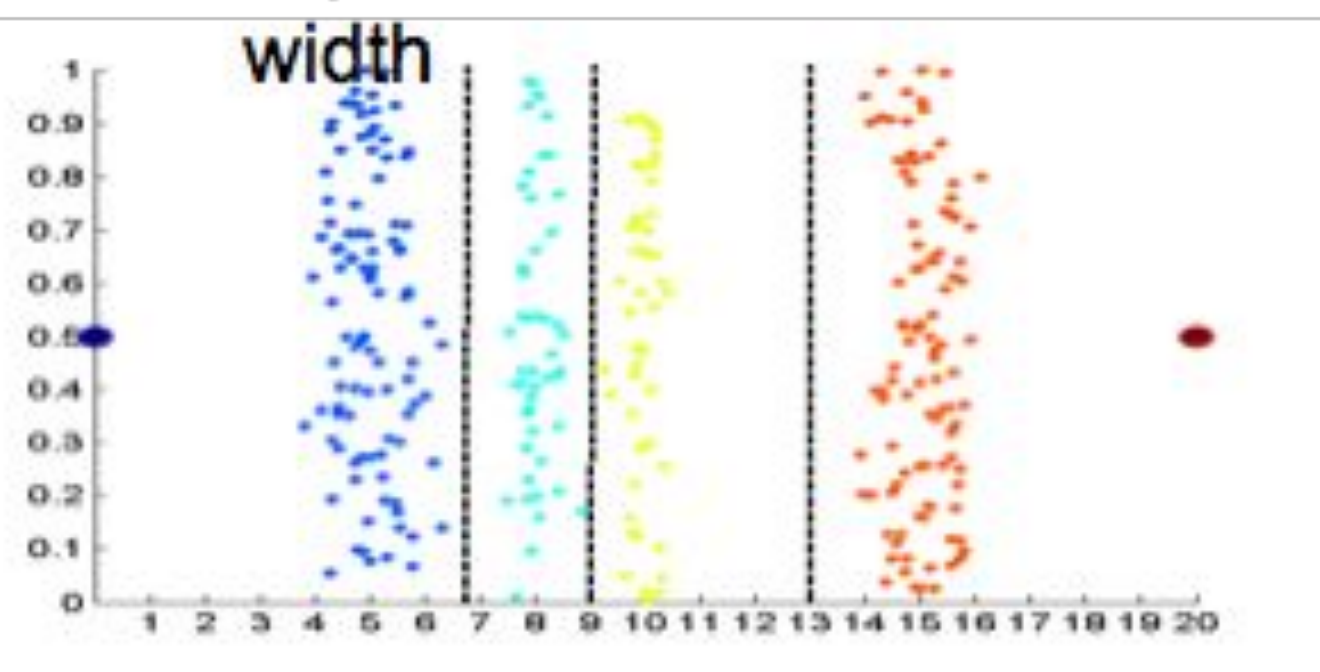


Data

Equal interval
width



Equal frequency



K-means

Transformación de atributos

- Una función que mapea el set de valores a otro set de datos.
- Funciones simples x^k , $\log(x)$, e^x , $|x|$
- Estandarización y Normalización



dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

f @ in  / DCCUCHILE