



Curso DM

“Datos II”

Primavera 2023

Basado en las slides de Bárbara Poblete

Calidad de los Datos

- Datos no poseen la calidad deseada a priori.
- Los algoritmos de DM se enfocan en:
 1. Limpieza de Datos: Detección y corrección de problemas de calidad
 2. Usar algoritmos que toleren datos de poca calidad

¿Por qué se producen errores?

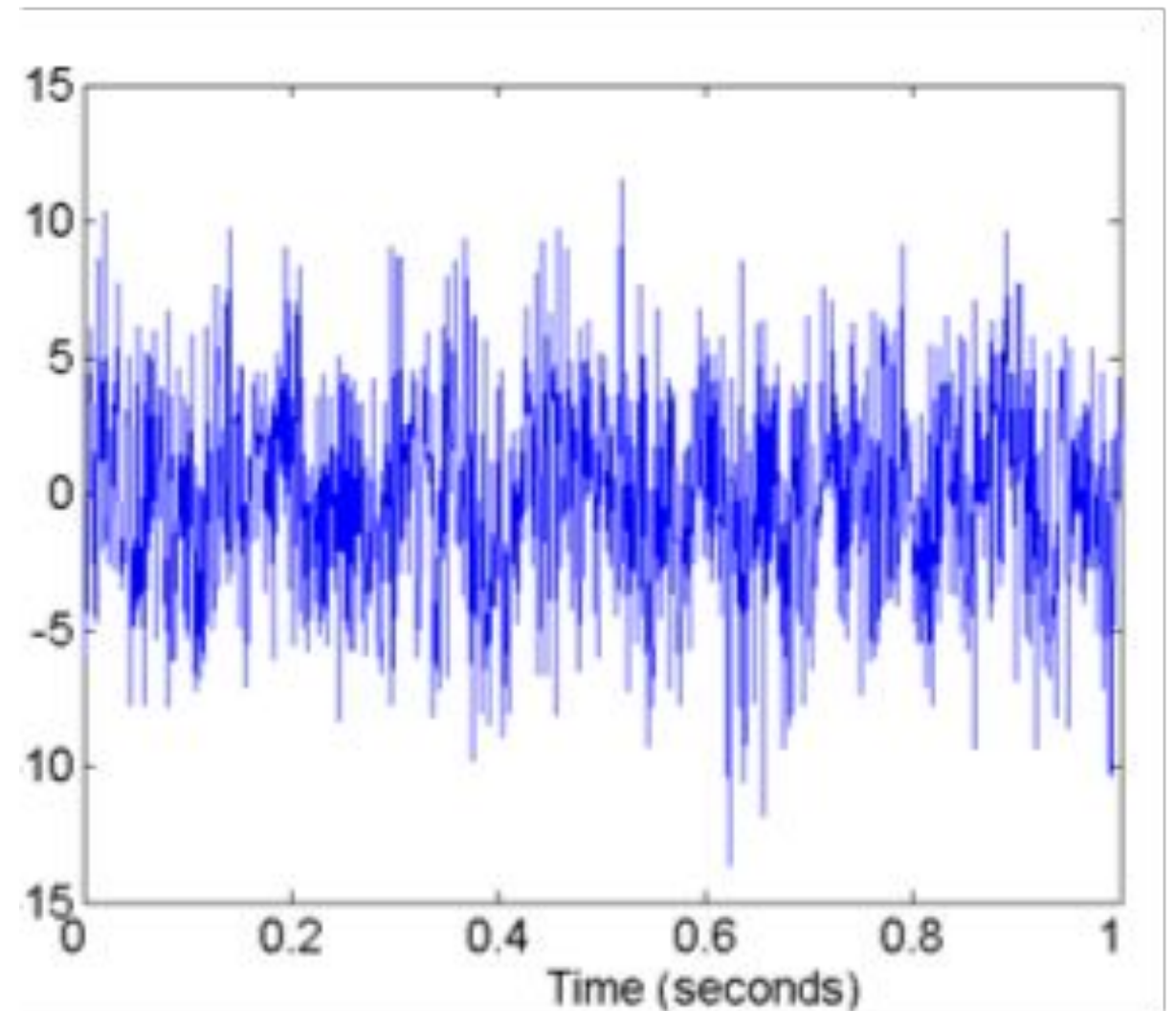
- Ruido y outliers
- Valores faltantes
- Datos duplicados
- Sesgo (Bias)



Artist
All rights obtainable from
iStock.com

¿Qué es el ruido?

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

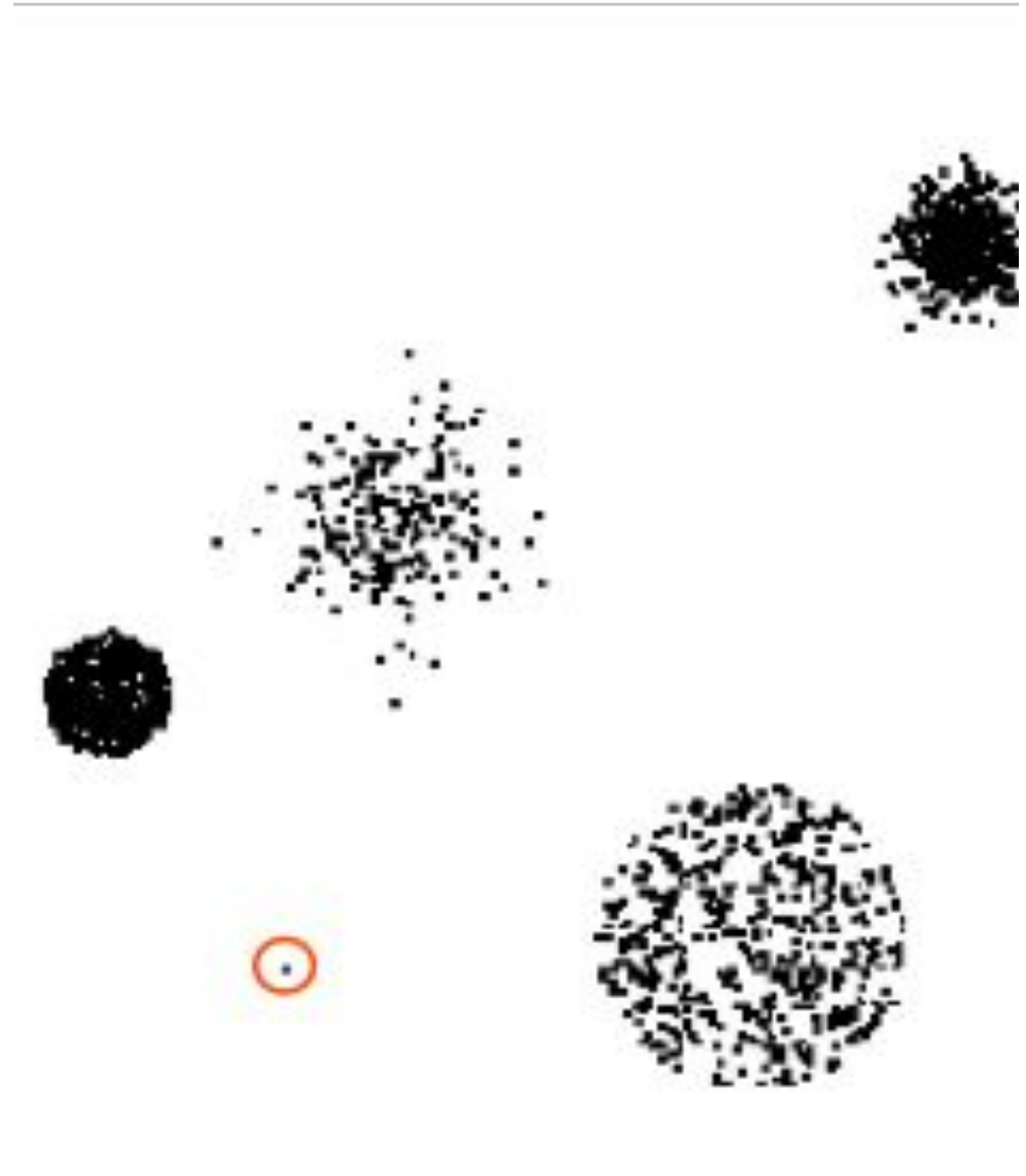


¿Qué es el ruido?



Outliers

- Objetos con características considerablemente diferentes a la mayoría



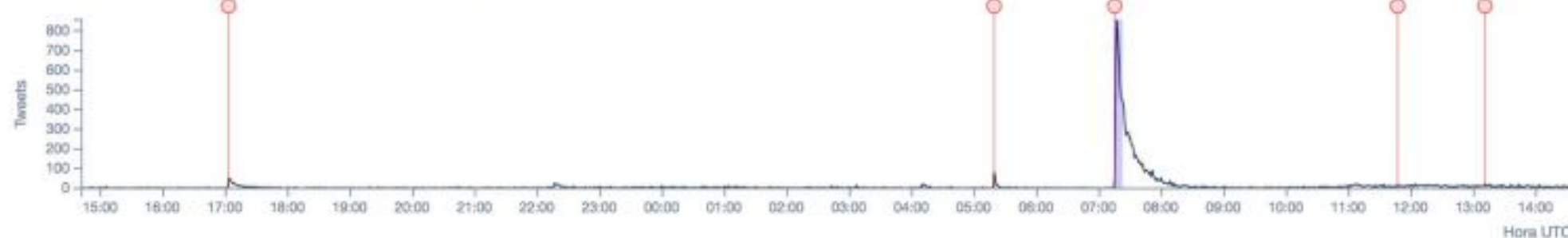
Frecuencia de tweets

REANUDAR

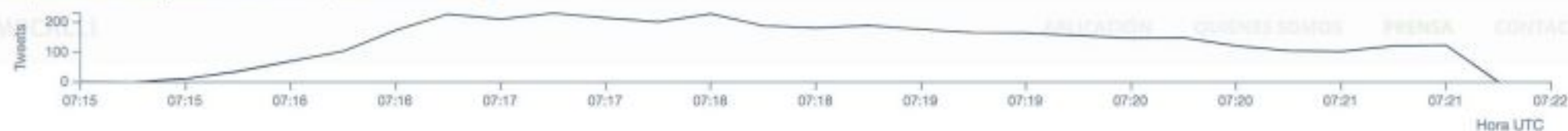
08/02/2017



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
@LadyLoloA 02/08/17
Un terrible temblor... ojala no sigan mas!!
Sobretudo por los que viven en malas
condiciones.

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

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

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

sachi☆ 07:21:59
@nn_sachiyo 02/08/17
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](#).



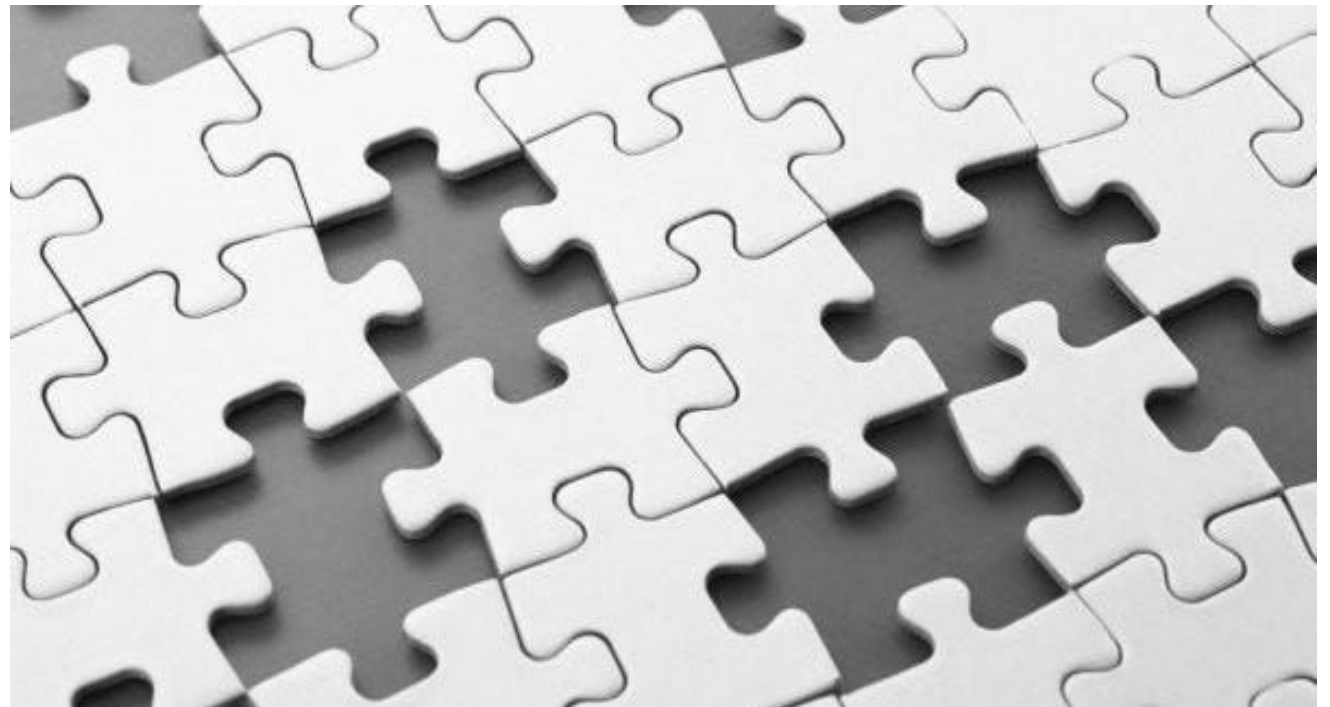
Centro de Investigación
de la Web Semántica



CSN
CENTRO SISMOLÓGICO NACIONAL
UNIVERSIDAD DE CHILE

Valores faltantes

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



```
sum(is.na(dataframe$column)) #nulos de una columna  
sum(!complete.cases(dataframe)) #nulos en el dataset
```


Datos duplicados

- Puede ocurrir al juntar datos de fuentes múltiples



```
sum(duplicated(dataframe))
```

Preprocesamiento de Datos

- Agregación
- Muestreo
- Selección de un subconjunto de atributos
- Reducción de dimensionalidad
- Normalización
- Creación de atributos
- Discretización y binarización

Preprocesamiento de Datos

- **Agregación**

- Muestreo
- Selección de un subconjunto de atributos
- Reducción de dimensionalidad
- Normalización
- Creación de atributos
- Discretización y binarización

Aggregación de Datos

Combinar 2 o más atributos (u objetos) en un único atributo (u 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

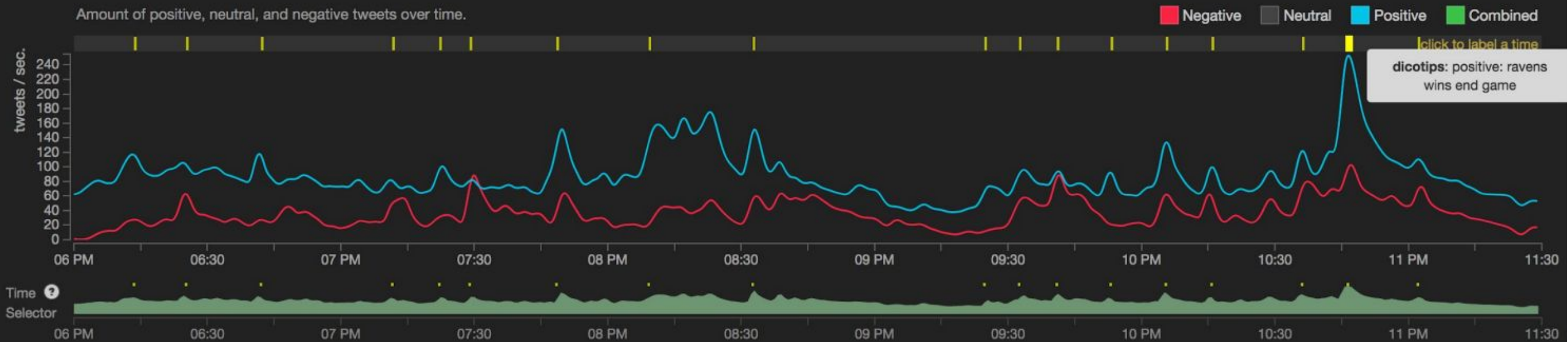
7.9 million tweets

3.8 million authors

Welcome, dicotips!

Sign out

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



Search tweet text



@author



Show RTs



all



Tweets

Top 100 most retweeted Tweets

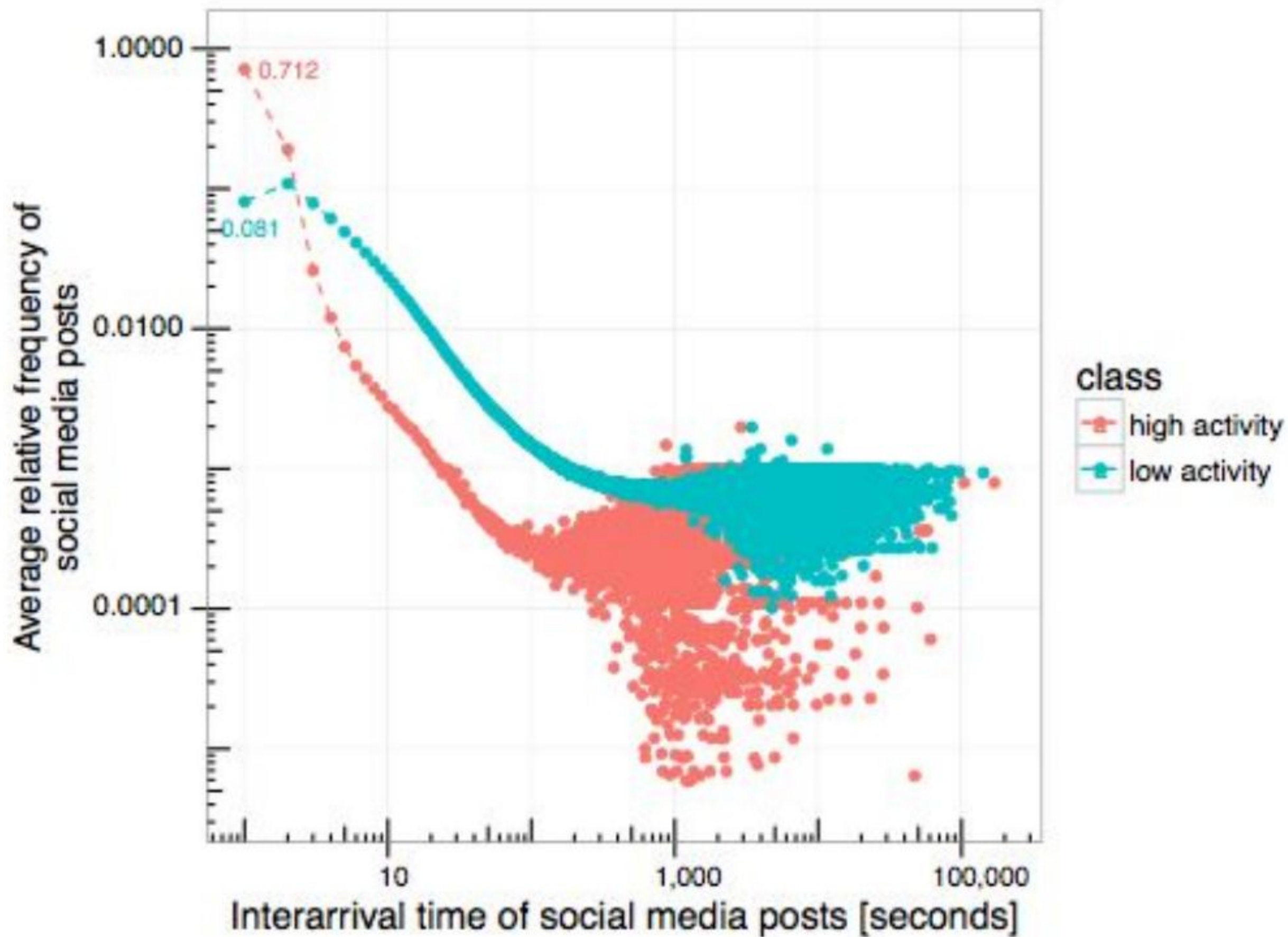
- @AlfredoFlores** 16835 retweets
Justin is currently in recovery from an apparent collapse after Beyonce's halftime performance. He wanted me to tell you all he is fine.
- @CalebU85** 7684 retweets
#RT The reason Why Ray Lewis is Whooping the 49ers is b/c he Joined Body By Vi #superbowl #TOTL #lightsout #bodybyvi <http://t.co/HfuELYIT>
- @TheIlluminati** 6992 retweets
Yes, we turned the lights off at the #Superbowl
- @piersmorgan** 5453 retweets

Average Happiness for Twitter

All Tweets in English.



https://hedonometer.org/timeseries/en_all/



Preprocesamiento de Datos

- Agregación
- **Muestreo**
- Selección de un subconjunto de atributos
- Reducción de dimensionalidad
- Normalización
- Creación de atributos
- Discretización y binarización

Muestreo

Principal técnica de selección de datos (investigación preliminar o final)

- Usado en Estadística y Minería de Datos
- ¿Cuándo es efectivo?

Muestreo Aleatorio



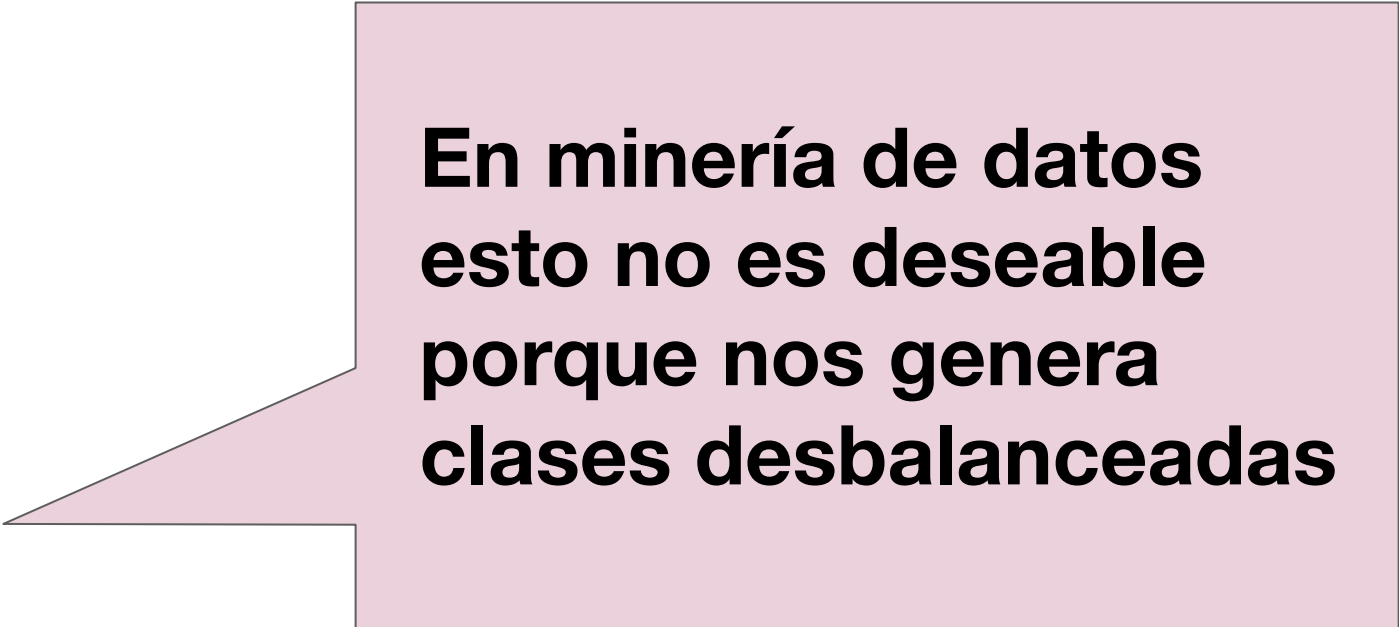
Muestreo Aleatorio

¿Ventajas?

Reduce la probabilidad de introducir sesgos

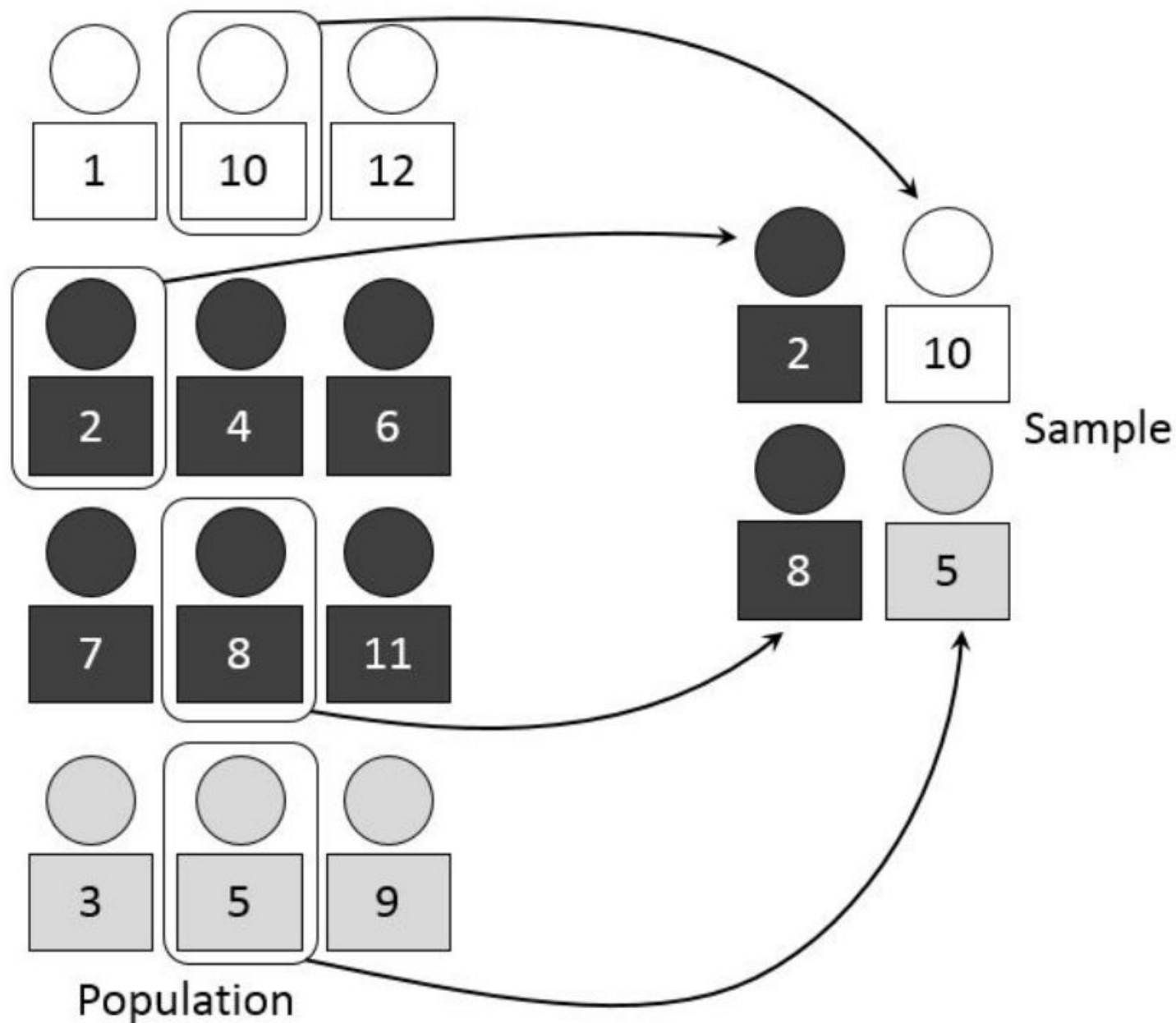
¿Desventajas?

Puede resultar en muestras poco representativas de la población



En minería de datos esto no es deseable porque nos genera clases desbalanceadas

Muestreo Estratificado

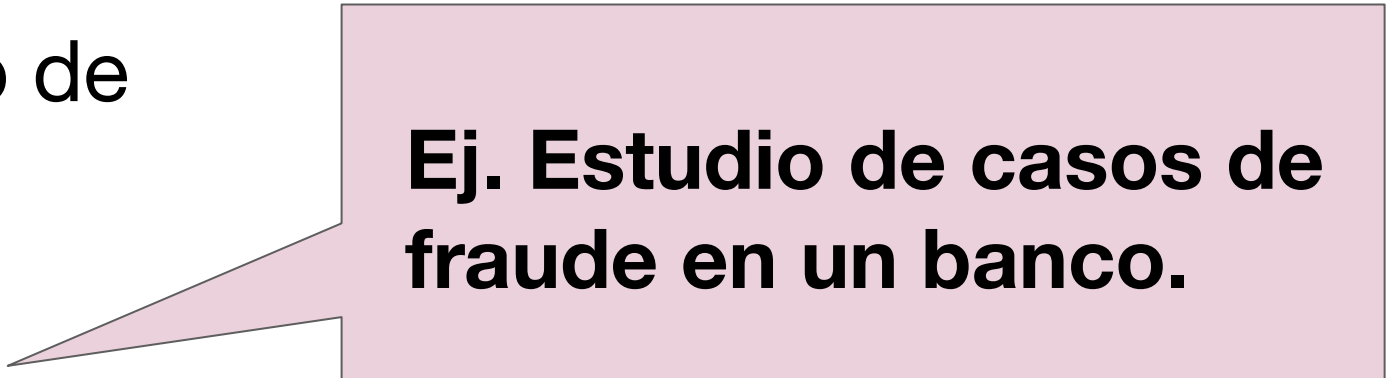


- Debo conocer cuáles son las clases o subpoblación
- Para cada una tomamos una muestra aleatoria

Muestreo Estratificado

En estadística

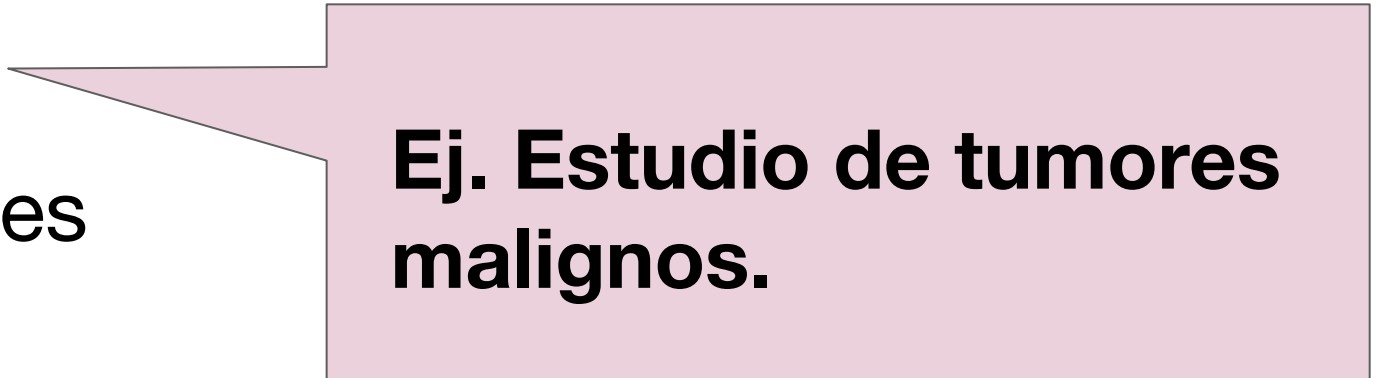
Buscamos muestras
proporcionales al tamaño de
la población real



**Ej. Estudio de casos de
fraude en un banco.**

En minería de datos

Puedo necesitar
sobrerrepresentar las clases
que son más pequeñas



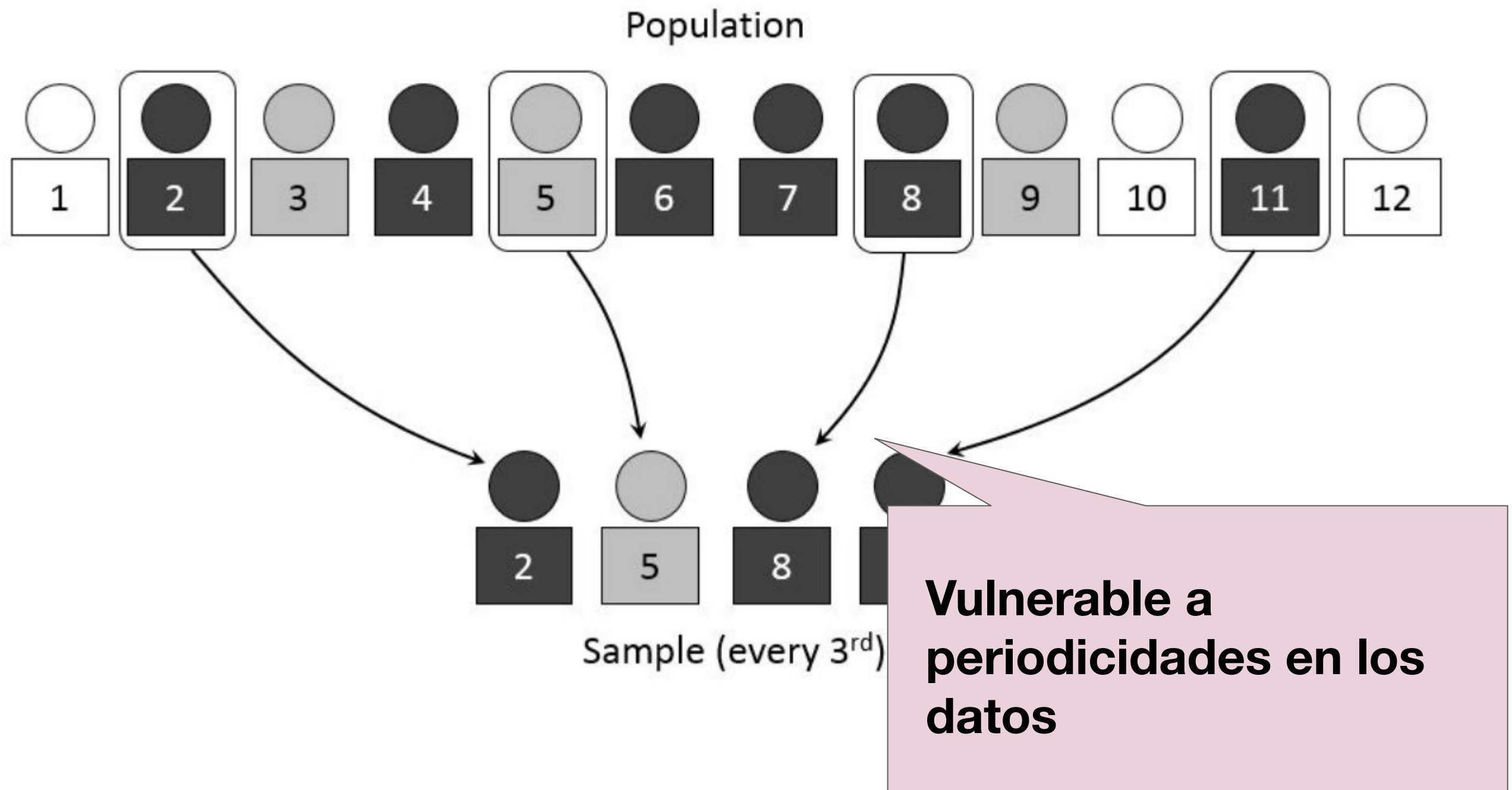
**Ej. Estudio de tumores
malignos.**

Oversampling y Subsampling

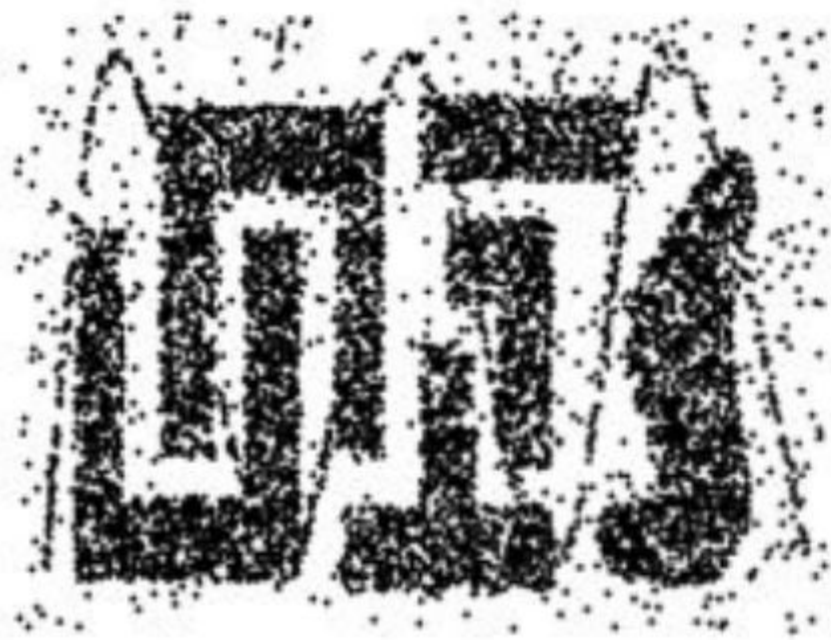
Oversampling: Sobrerrepresentación de una población o clase que es más pequeña.

Subsampling: Disminuir la población o clase que es de mayor tamaño

Systematic random sample



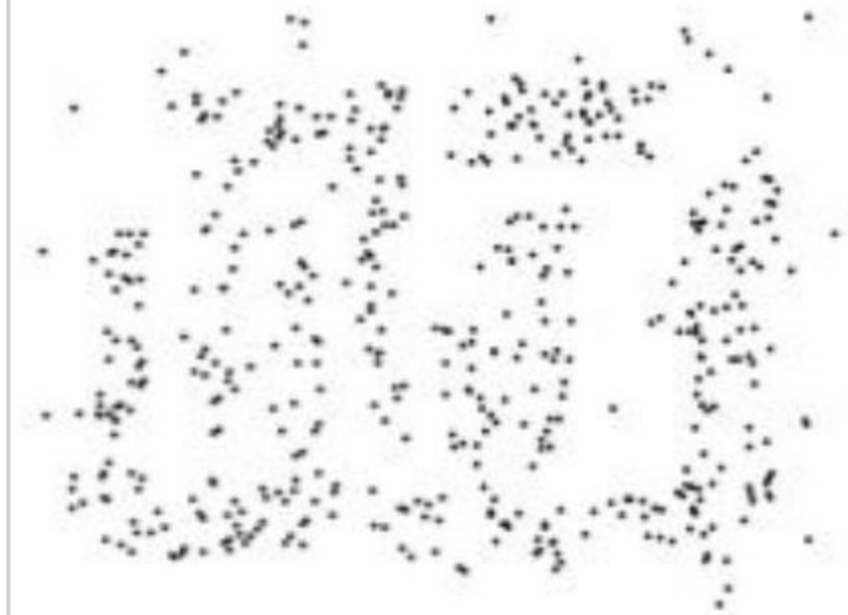
Tamaño de la muestra



8000 points



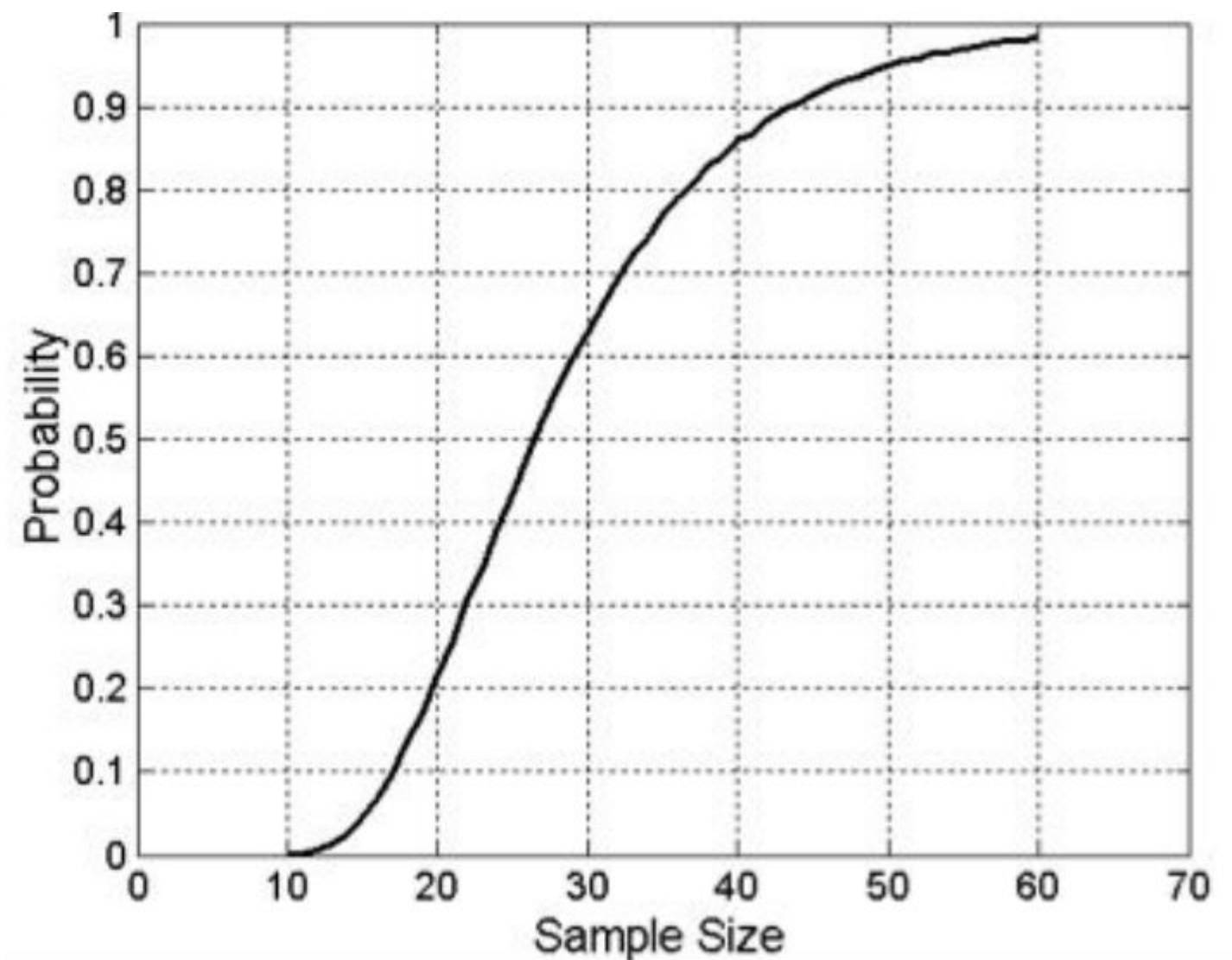
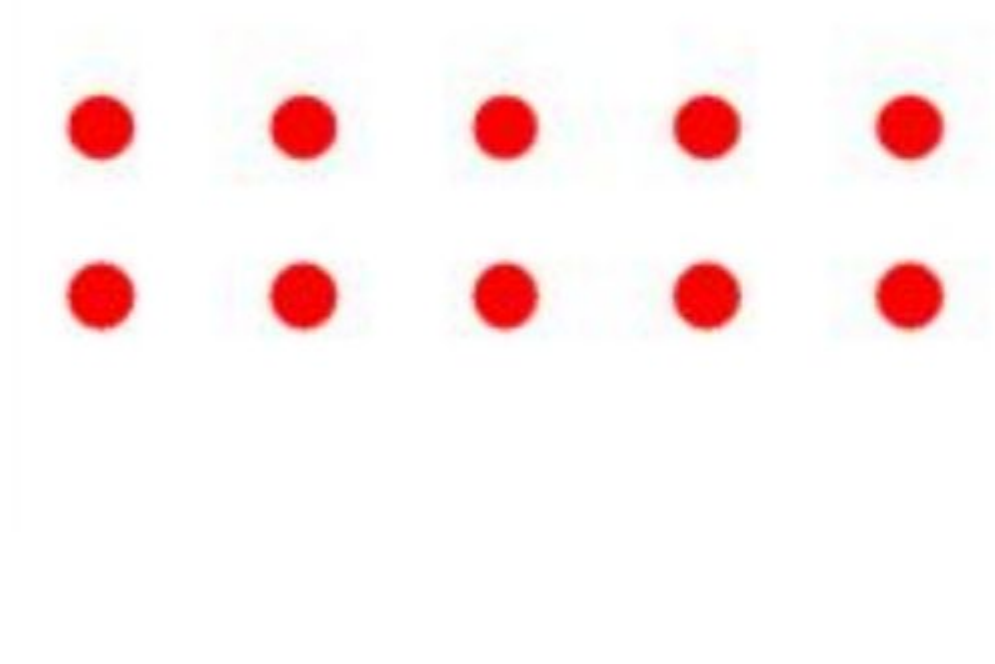
2000 Points



500 Points

En pocos datos se pierden los patrones

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



Preprocesamiento de Datos

- Agregación
- Muestreo
- **Selección de atributos**
- **Reducción de dimensionalidad**
- Normalización
- Creación de atributos
- Discretización y binarización

Maldición de la Dimensionalidad

Se habla de dataset de alta dimensionalidad cuando tenemos más dimensiones que observaciones

- Al aumentar la dimensionalidad, los datos se vuelven más dispersos en el espacio.
- Pierden significado las medidas, i.e. densidad y distancia entre puntos



Reducción de Dimensionalidad y Selección de Atributos

Propósito

- Evitar maldición de la dimensionalidad
- Reducir costos asociados a aplicar algoritmos (tiempo, memoria)
- Mejor visualización de los datos
- Ayuda a quitar atributos irrelevantes o ruidosos

Selección de Atributos

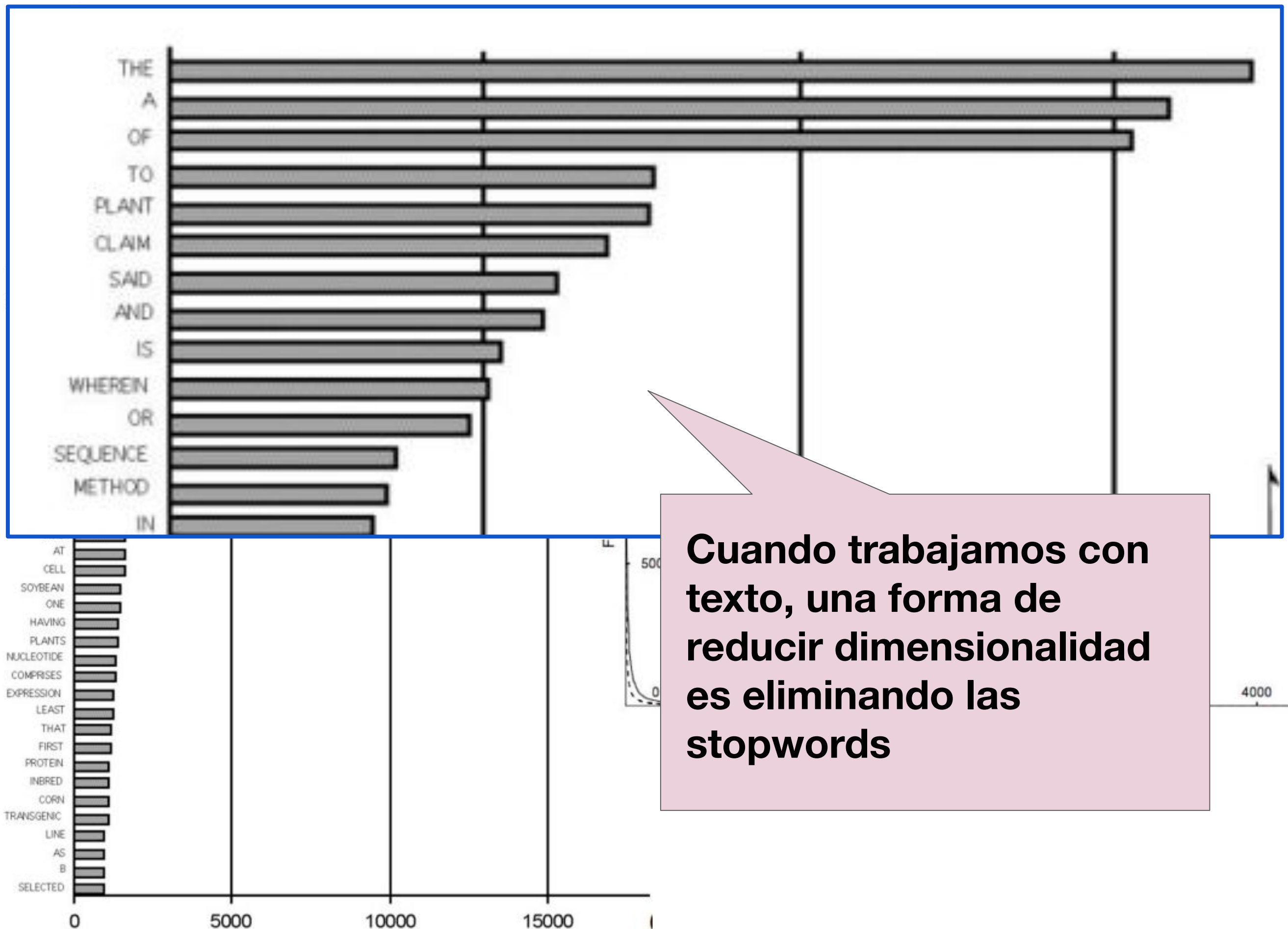
Elegimos atributos

- Missing Values Ratio
- Low Variance Filter
- High Correlation Filter

Si una columna posee muchos valores nulos podría ser candidata para eliminarla

Si una columna tiene varianza casi cero, quiere decir que no aporta mucha info.

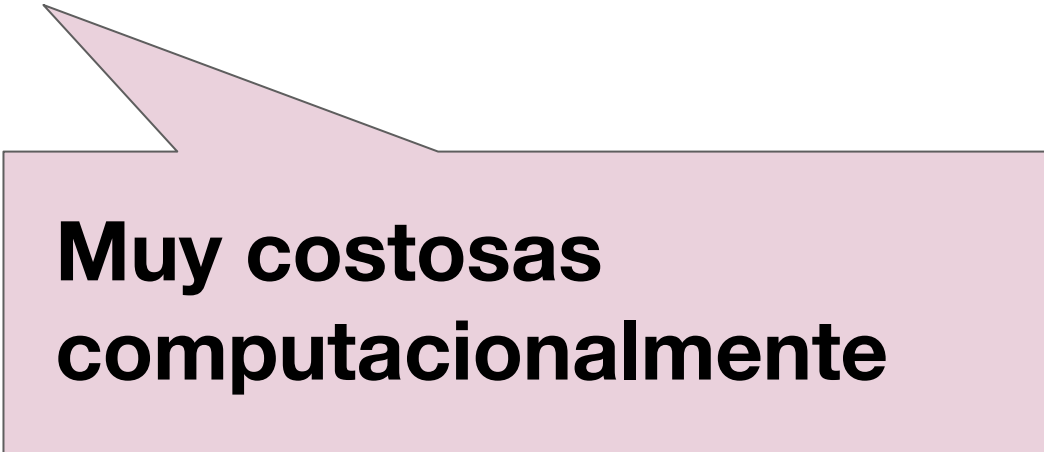
Si dos columnas están muy correlacionadas, me puedo quedar con una solamente



Selección de Atributos

Técnicas automáticas que seleccionan atributos

- Random Forest / Ensemble Trees
- Backwards/Forward Feature Elimination/Construction

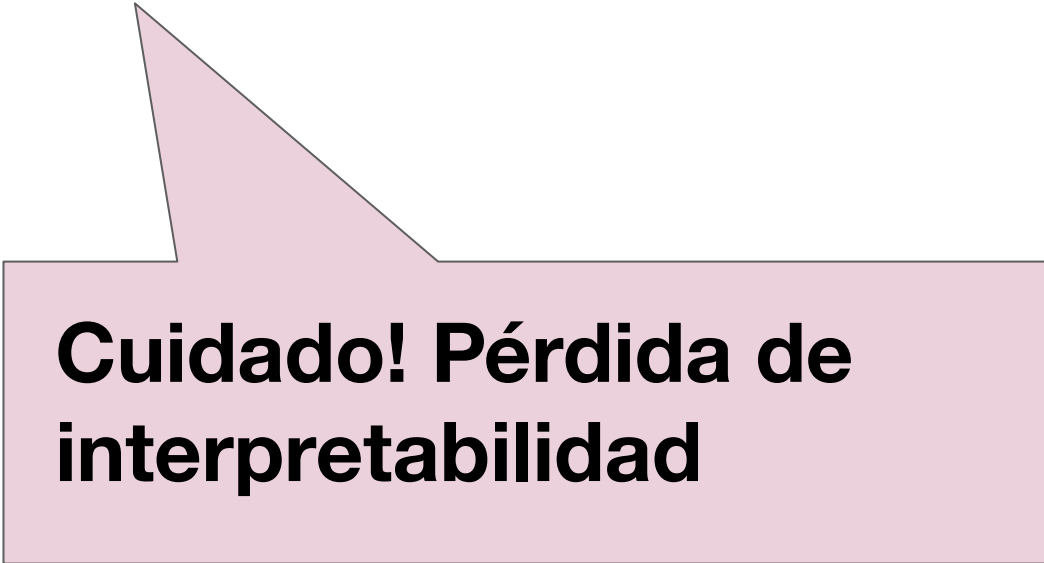


**Muy costosas
computacionalmente**

Reducción de Dimensionalidad

Técnicas de Álgebra lineal para transformar el espacio dimensional a uno de menor tamaño

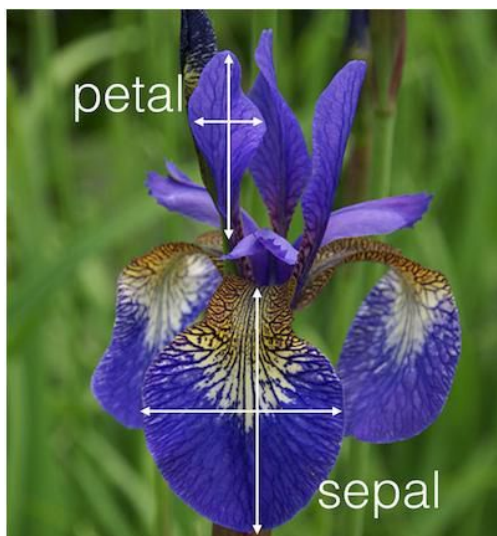
- PCA (Análisis de componentes principales)
- LDA (Análisis discriminante lineal)
- SVD
- Isomap



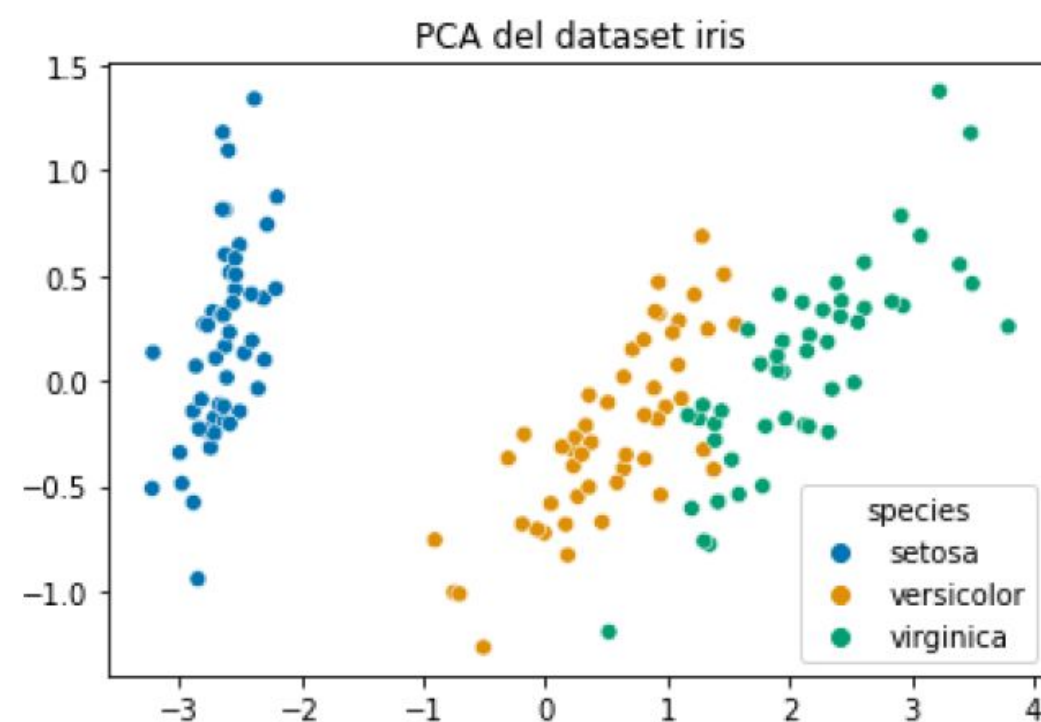
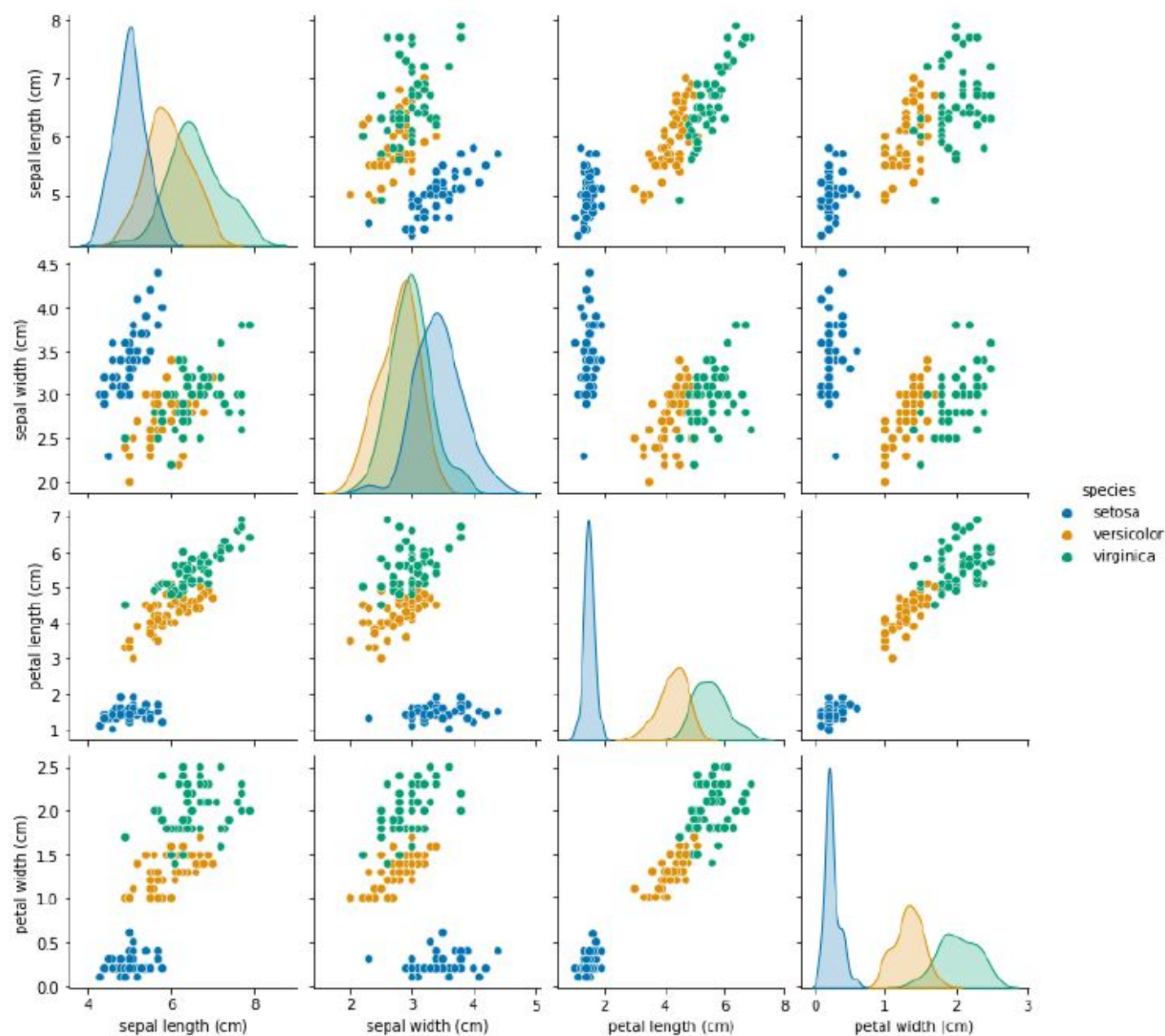
Cuidado! Pérdida de interpretabilidad

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%	

<http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html>



Reducir la dimensionalidad también nos sirve para visualizar los datos!



Preprocesamiento de Datos

- Agregación
- Muestreo
- Selección de atributos
- Reducción de dimensionalidad
- **Normalización**
- Creación de atributos
- Discretización y binarización

Dar pesos a los atributos

Se asigna peso a los atributos según su importancia

- SVM (Support Vector Machine) lo hace automáticamente
- Normalización

Normalización

- Escalamos nuestros atributos para que estén en el mismo rango
- Nos ayuda a que atributos con mayor escala no tengan una importancia desmedida
- Lo más común es escalarlos a un rango entre 0 y 1

Normalización:

variable con rango 0-1

$$= \frac{x - \min(x)}{\max(x) - \min(x)}$$

Estandarización:

variable con media 0 y desviación estándar 1

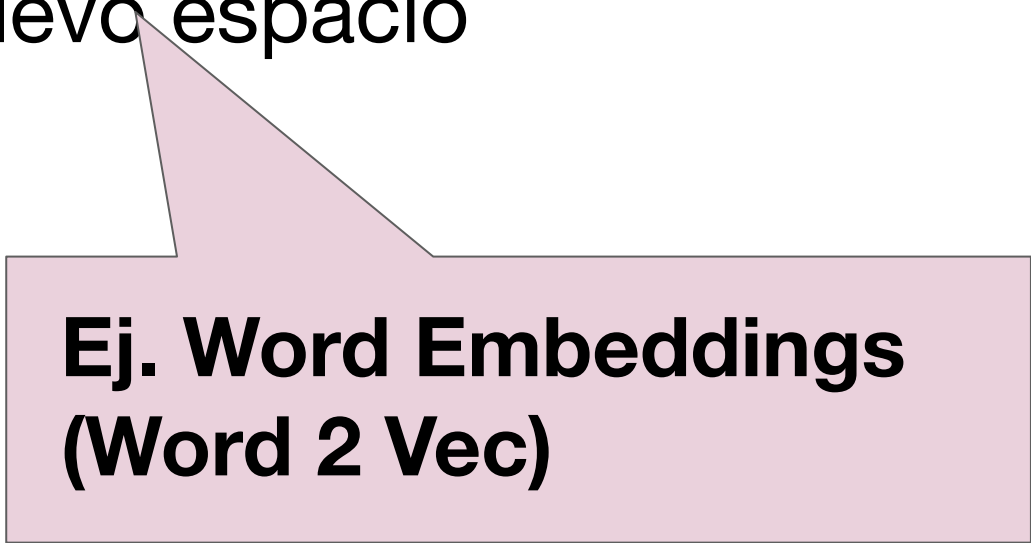
$$x' = (x - \bar{x}) / s_x$$

Preprocesamiento de Datos

- Agregación
- Muestreo
- Selección de atributos
- Reducción de dimensionalidad
- Pesos y Normalización
- **Creación de atributos**
- Discretización y binarización

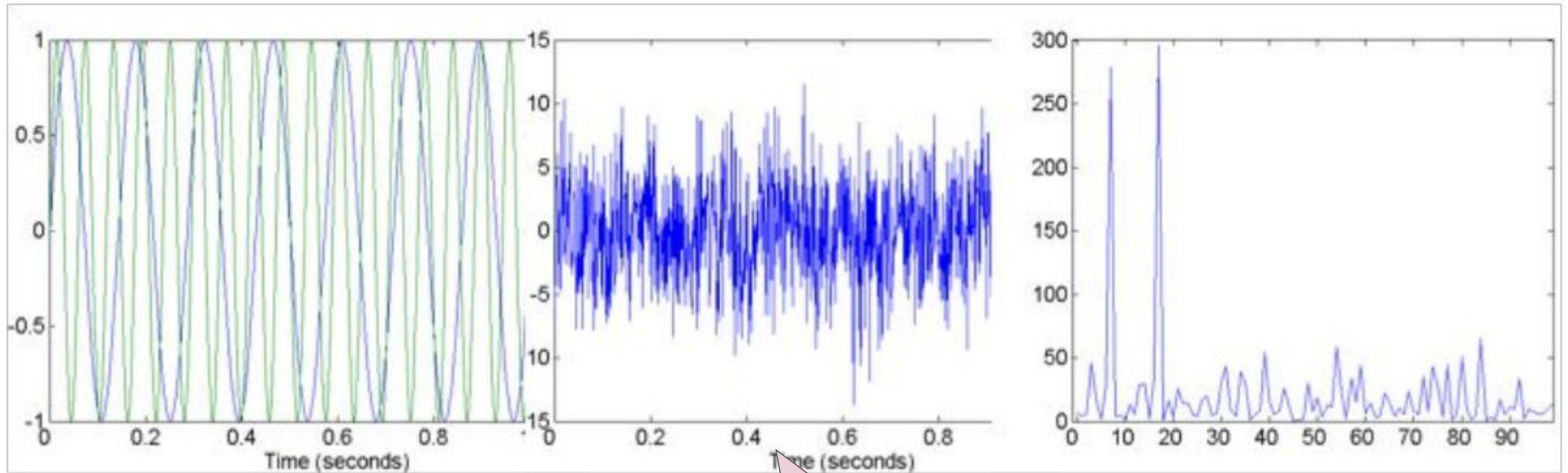
Crear atributos

- Aplicamos una función a uno o más atributos para crear uno nuevo.
- Mapeamos atributos a un nuevo espacio



**Ej. Word Embeddings
(Word 2 Vec)**

Mapear a un nuevo espacio



Two Sine Wave

Two Sine Wave +
Noise

Frequency

La transformada de Fourier nos permite llevar la señal al dominio de la frecuencia y detectar ruido

Preprocesamiento de Datos

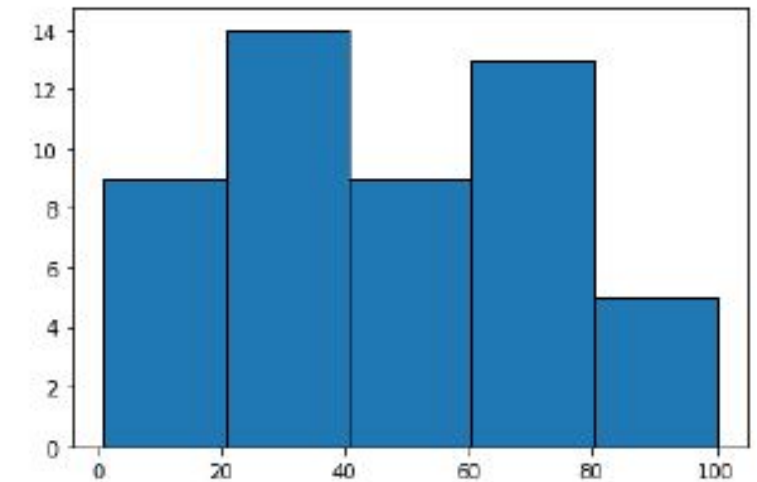
- Agregación
- Muestreo
- Selección de atributos
- Reducción de dimensionalidad
- Pesos y Normalización
- Creación de atributos
- **Discretización y binarización**

Discretizar y Binarizar

- Proceso de convertir un atributo continuo en un atributo categórico
- Decidir cuántas categorías tendremos
- Supervisado (yo elijo cuáles clases)
- No-supervisado (ej. división usando intervalos fijos, usando clustering, etc.)

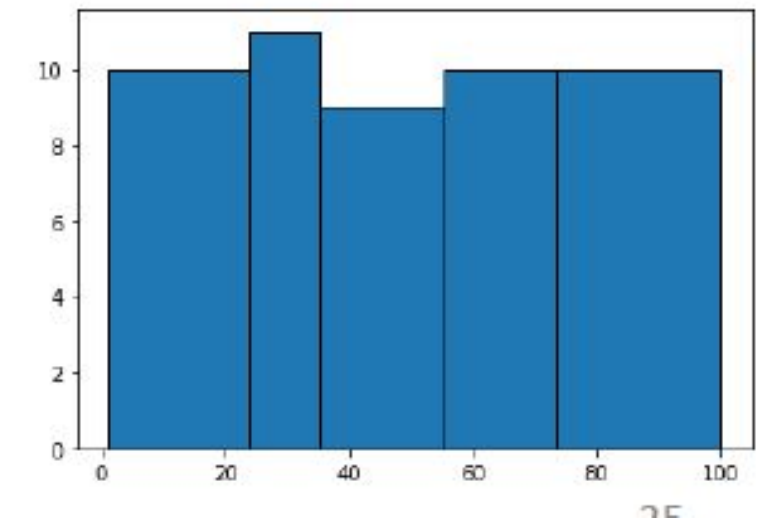
Equal width
Intervalos del mismo ancho

(0.901, 20.8]	9
(20.8, 40.6]	14
(40.6, 60.4]	9
(60.4, 80.2]	13
(80.2, 100.0]	5



Equal depth
Intervalos con casi el mismo
número de instancias

(0.999, 23.8]	10
(23.8, 35.0]	11
(35.0, 55.2]	9
(55.2, 73.6]	10
(73.6, 100.0]	10



**o usando clustering,
por ej K-means**

Próxima Clase

- Cómo lidiar con datos faltantes
- Profundización en Selección de Atributos y Reducción de dimensionalidad
- Ejemplos prácticos

Repaso en Casa!! Revisar Tutorial I

- Familiarizarse con funciones de manipulación de vectores y matrices.
- Función **aggregate**: Sirve para agrupar objetos de acuerdo a una función y atributo específico (como group by en SQL)
- Funciones para Filtrar y Ordenar una matriz en base a columnas específicas.
- Función **melt**: Sirve para transformar la matriz
- Gráficos usando plot y/o ggplot



dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

f @ in  / DCCUCHILE