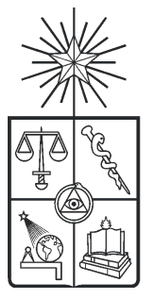


dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Curso DM
“Datos II”

Bárbara Poblete

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?

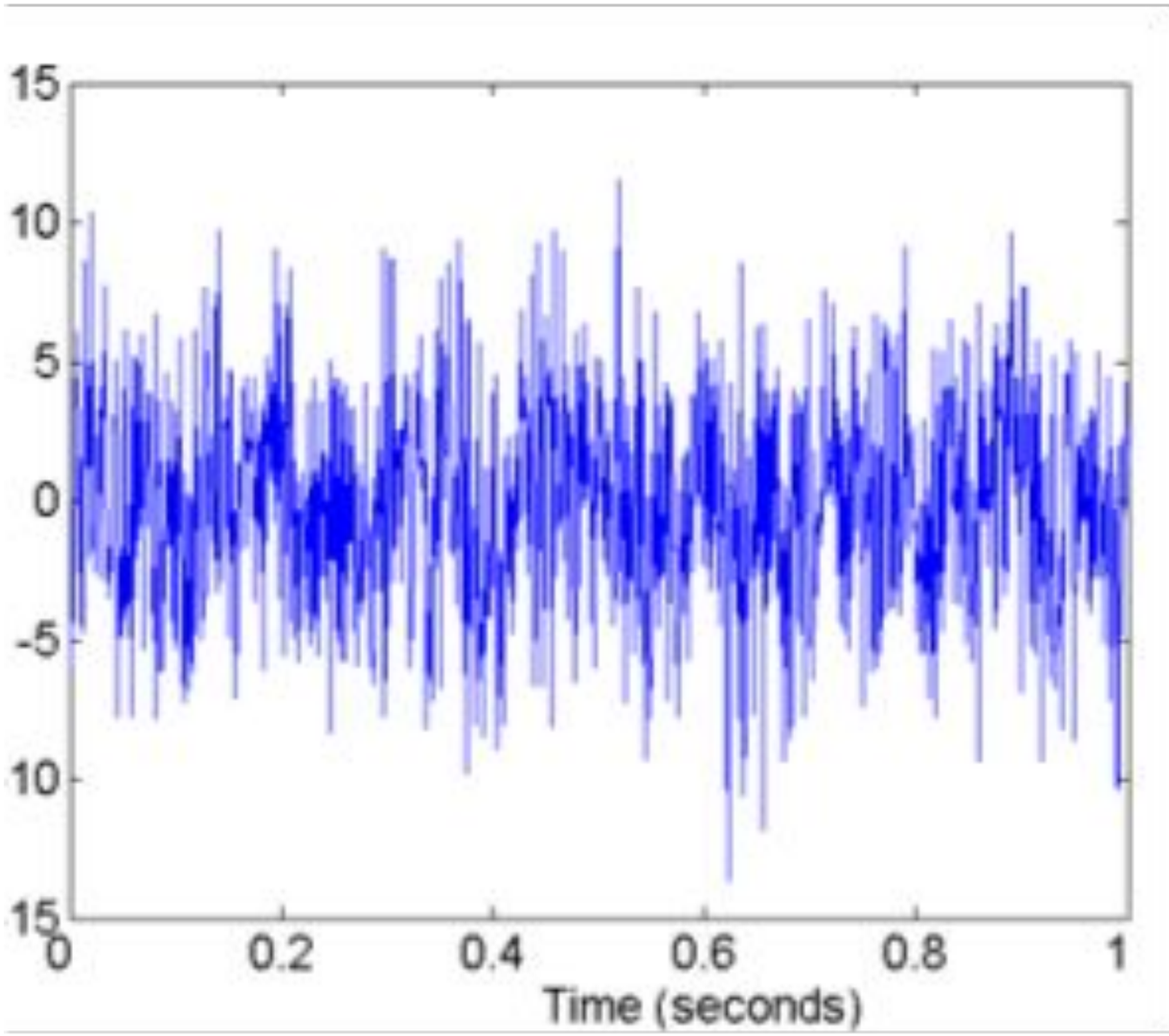
Tipos de errores:

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

¿Qué es el ruido?

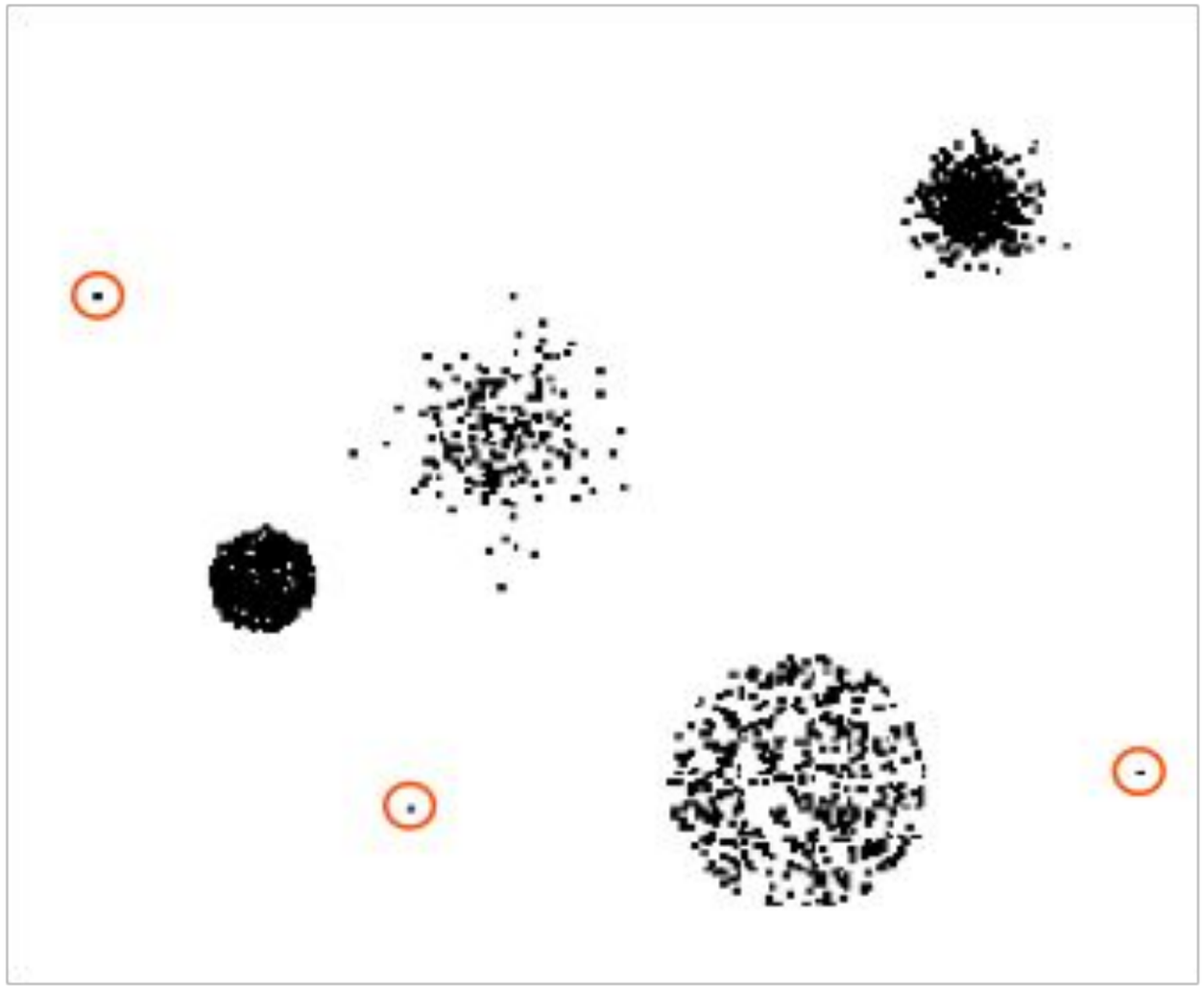
- Componente aleatoria en la medición (distorsión de voz en un teléfono malo)

- Datos espaciales, temporales



Outliers

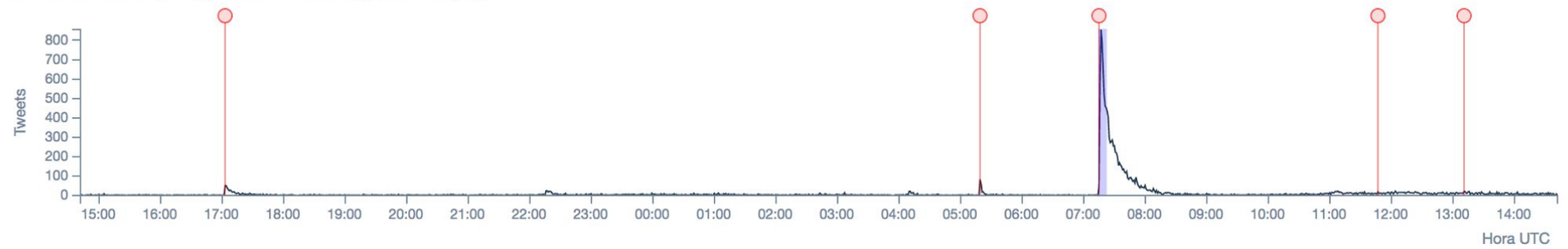
- Objetos con características considerablemente diferentes a la mayoría



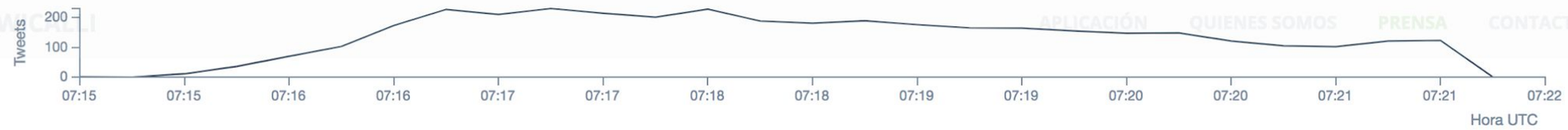
Frecuencia de tweets

REANUDAR 08/02/2017

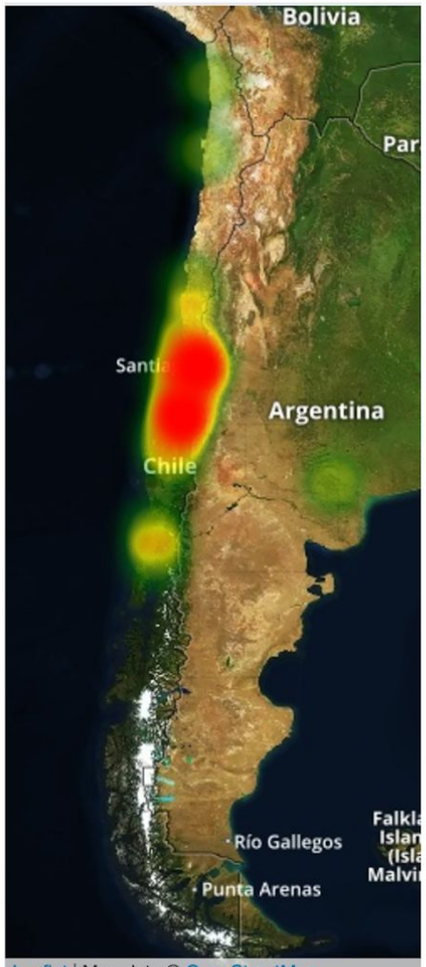
Últimas 24 horas (01 Agosto 2017 - 02 Agosto 2017)



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

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

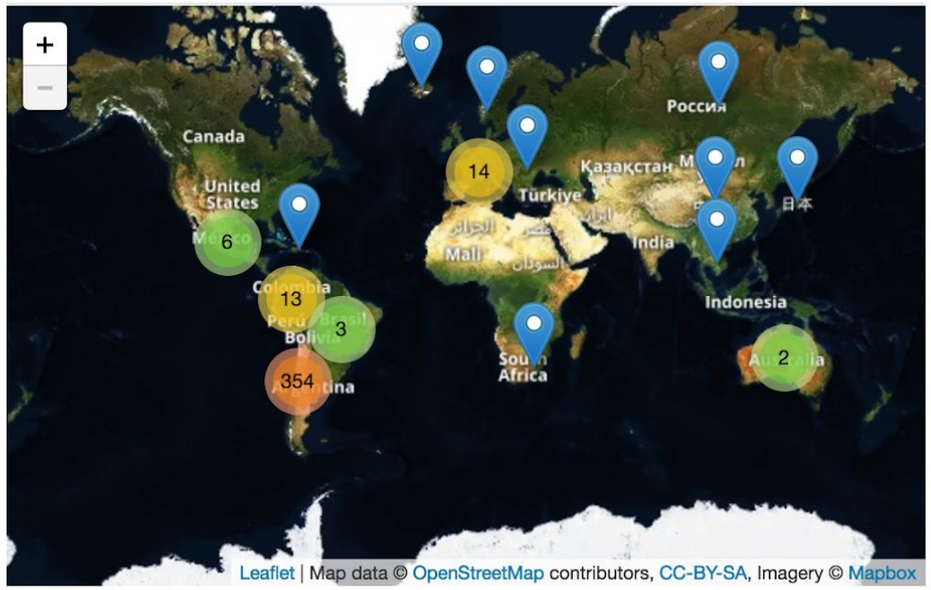
Andriu @andres_munoza *07:21:59 02/08/17*
Ojala nunca pasen un temblor volados ctm, estoy mal

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

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

Tweets con geolocalización

Filtrar marcadores



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

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

Pre-procesamiento de datos

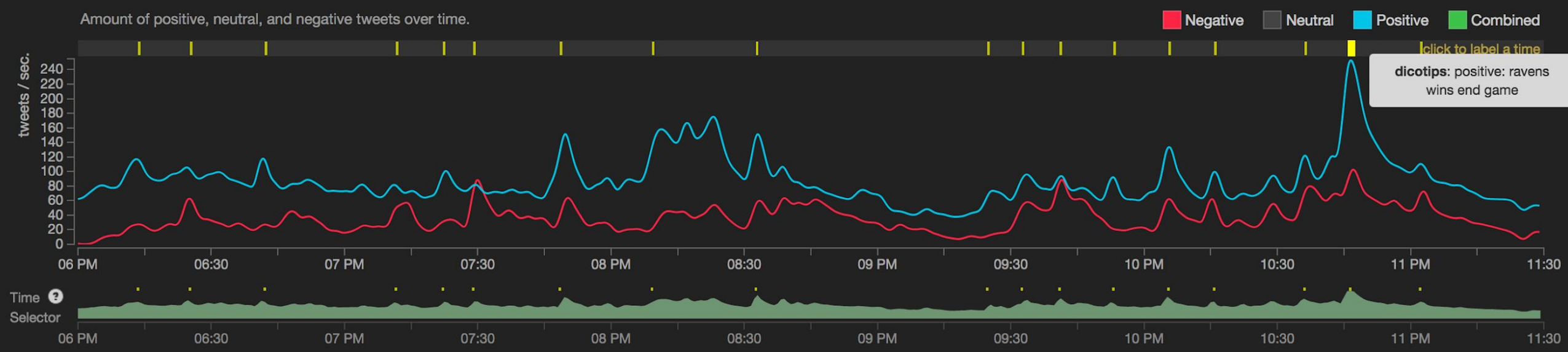
- Creación de atributos
- Selección de un subconjunto de atributos
- Agregación
- Normalización
- Muestreo
- Reducción de dimensionalidad
- Discretización y binarización

Agregación 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



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

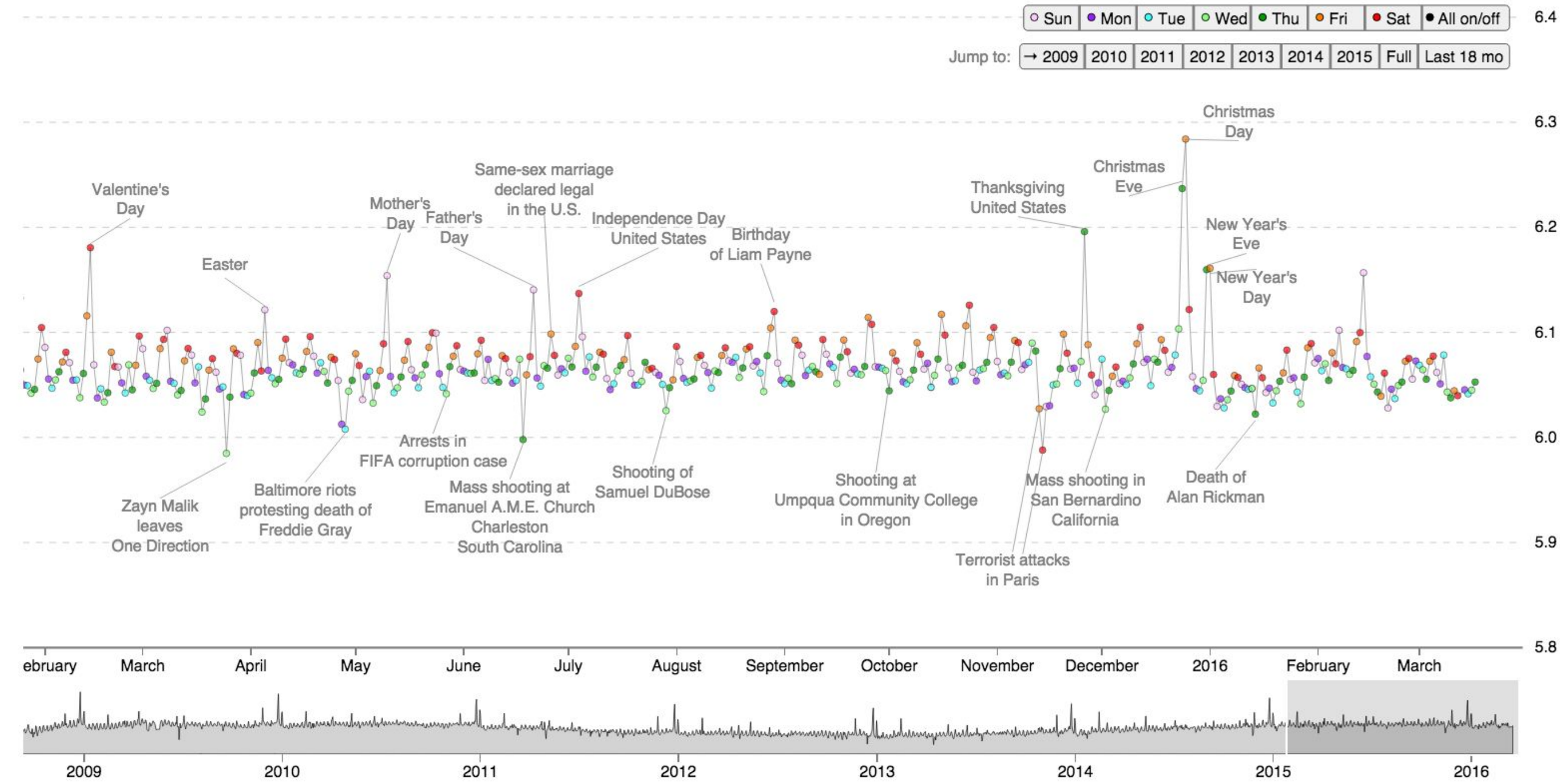
> #RT The reason Why Ray Lewis is Whooping the 49ers is b/c he Joined Body By Vi #superbowl #TOTL #lightsout #bodybyvi http://t.co/HfuELYIT

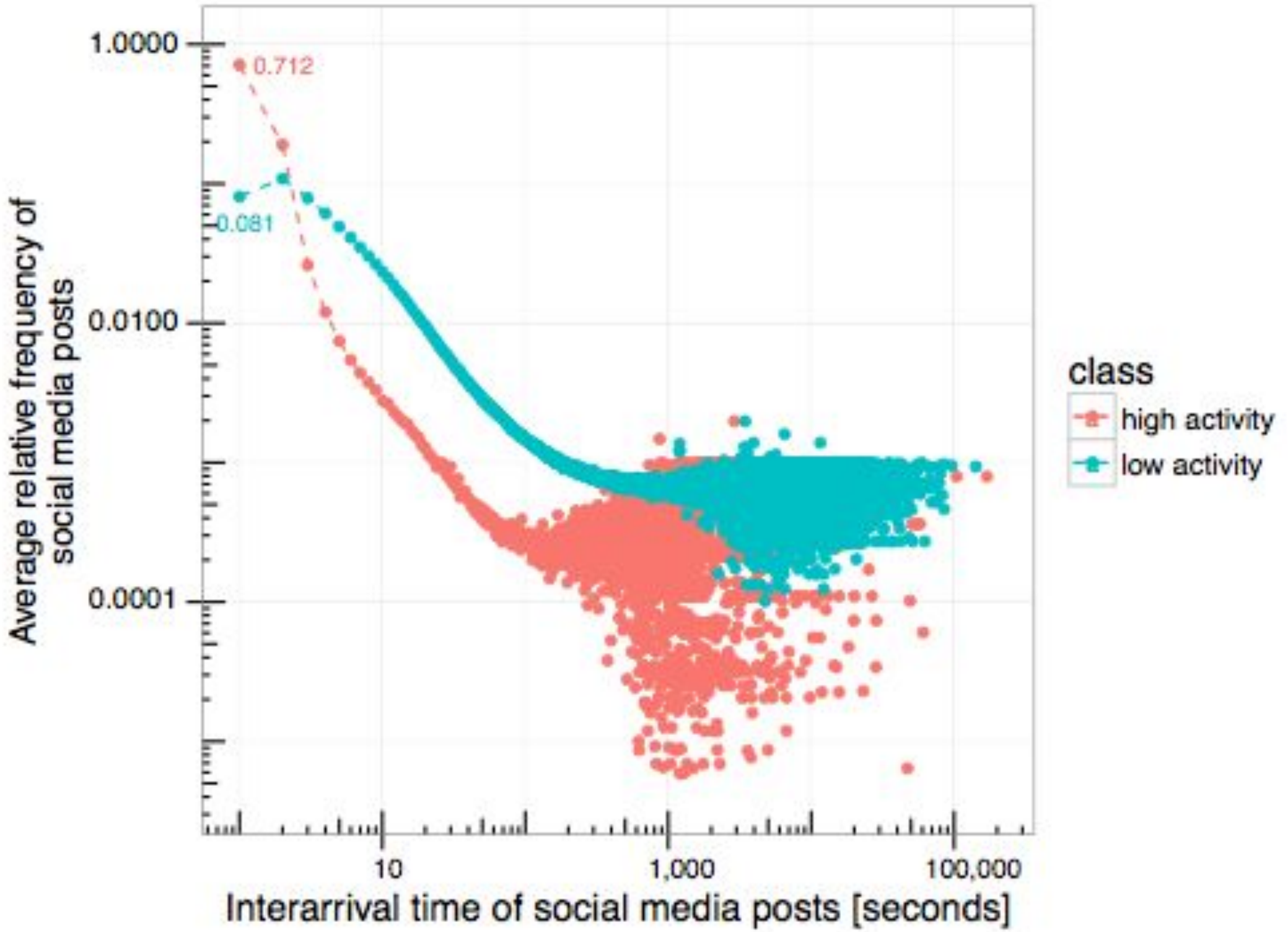
@TheIlluminati — 6992 retweets

> Yes, we turned the lights off at the #Superbowl

@piersmorgan — 5453 retweets

Average Happiness for Twitter




1.0000
0.0100
0.0001
0.712
0.081
Average relative frequency of social media posts
10
1,000
100,000
Interarrival time of social media posts [seconds]
class
high activity
low activity

Agregación

Muestreo

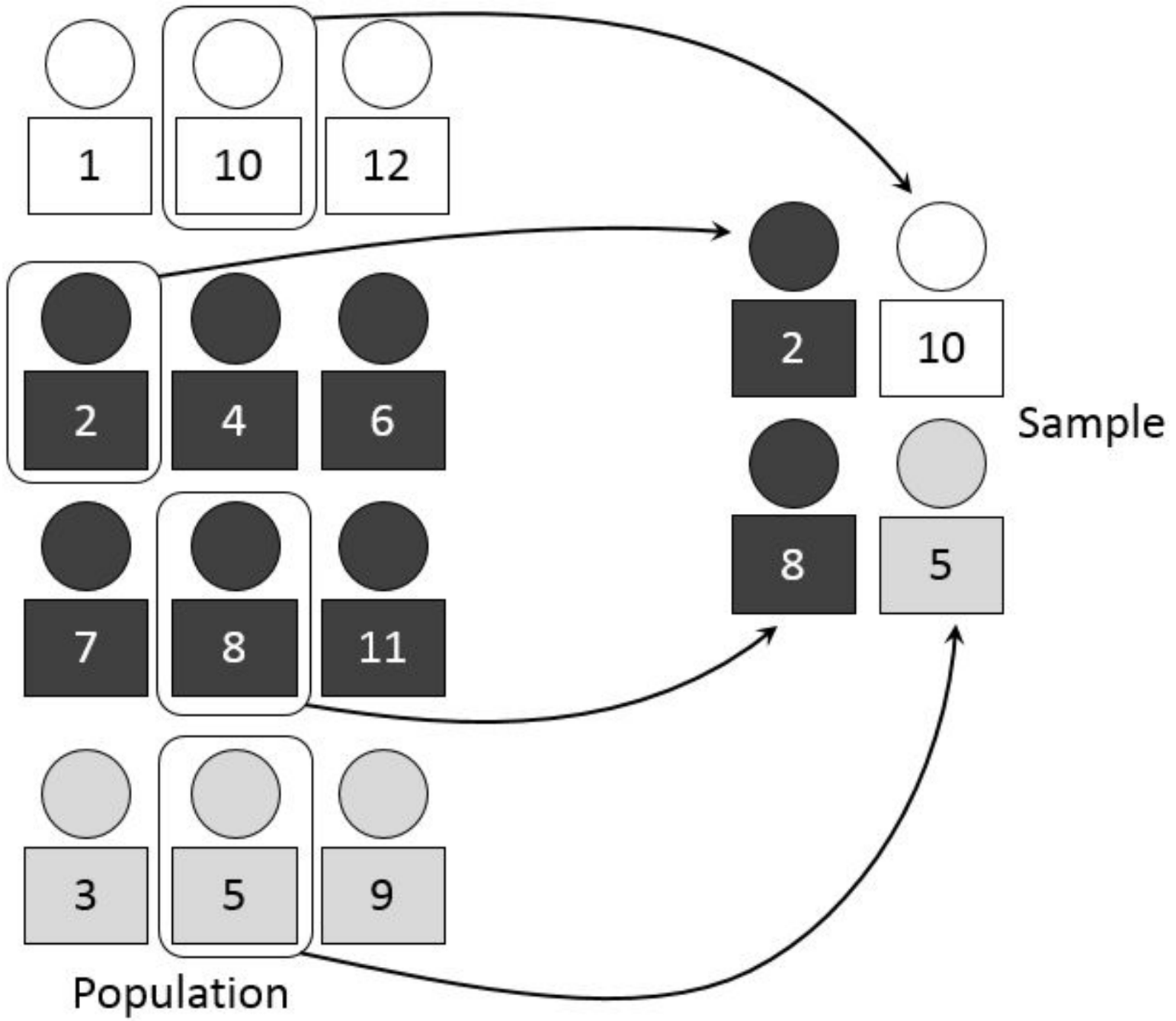
- Principal técnica de selección de datos (investigación preliminar o final)
- Usado en Estadística y DM
- ¿Cuándo es efectivo?

Tipos de Muestreo

- Muestreo aleatorio
- ¿Ventajas?
- ¿Desventajas?

Tipos de Muestreo

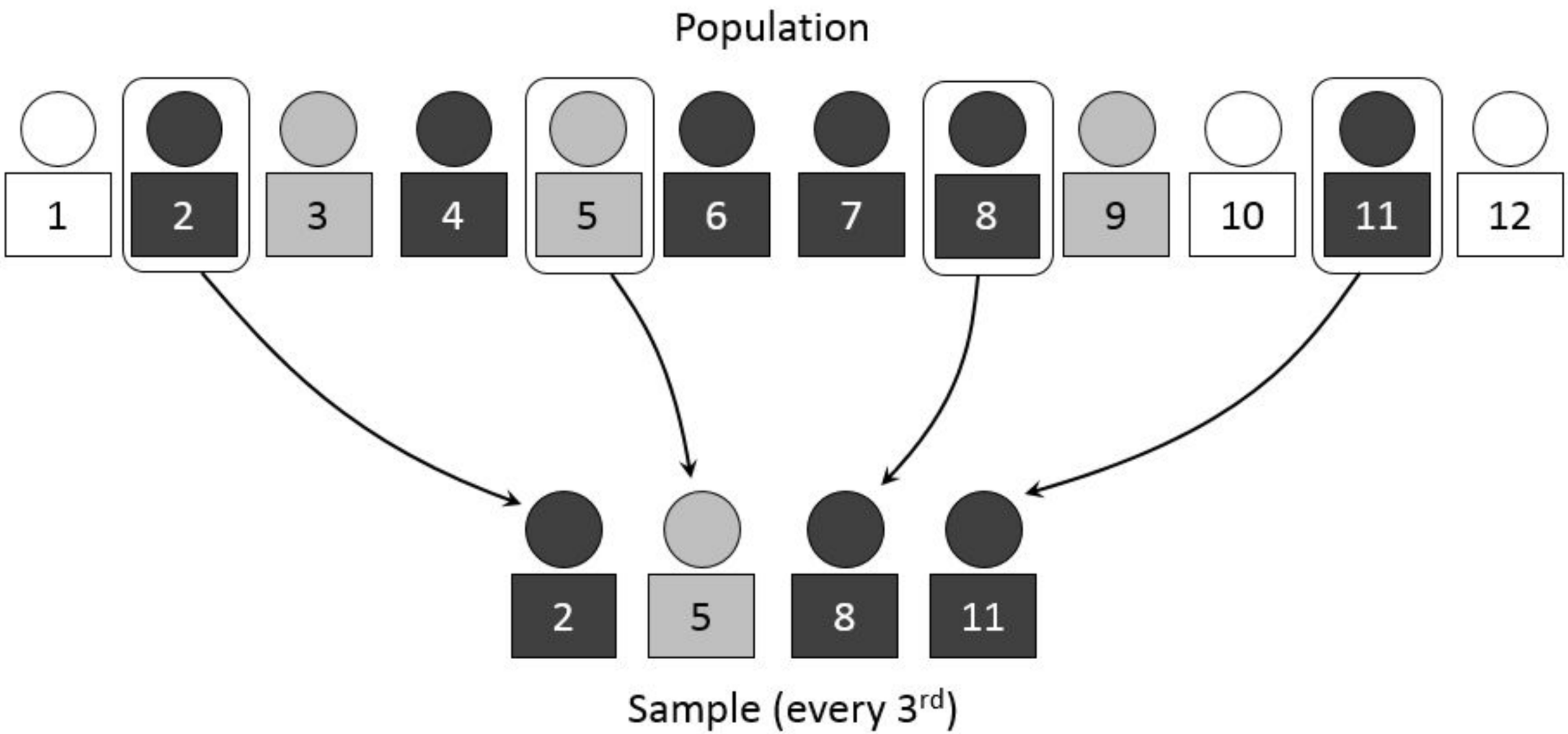
- Muestreo estratificado
- ¿Ventajas?
- ¿Desventajas?

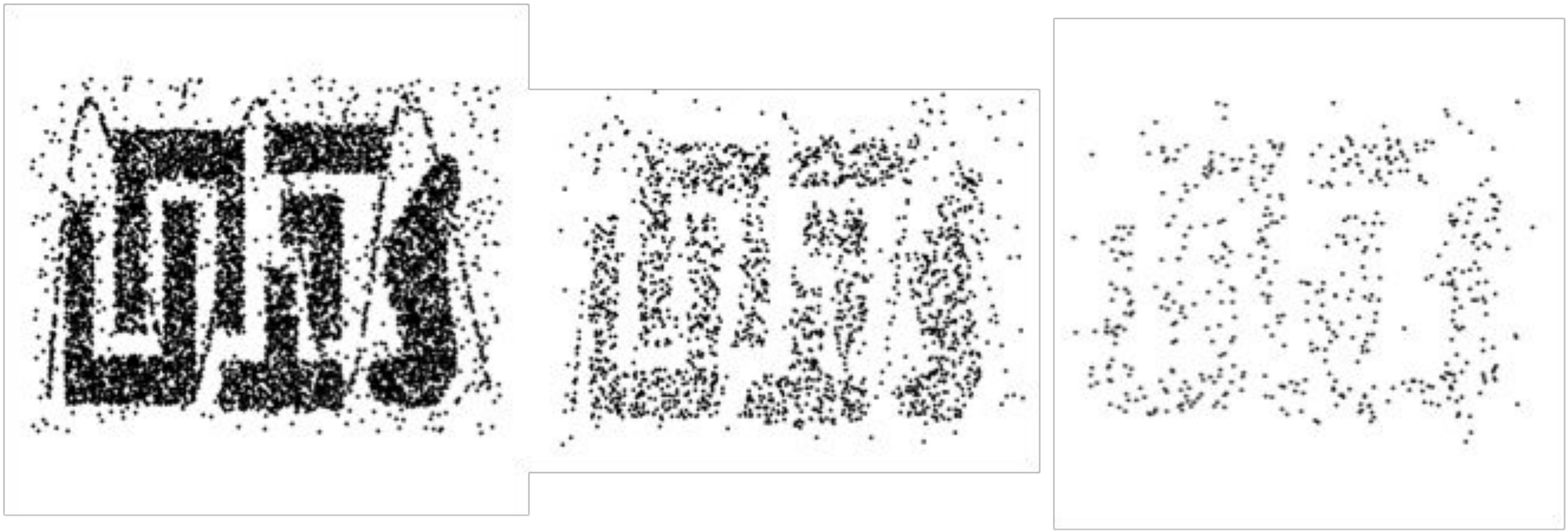


Tipos de Muestreo

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

Systematic random sample

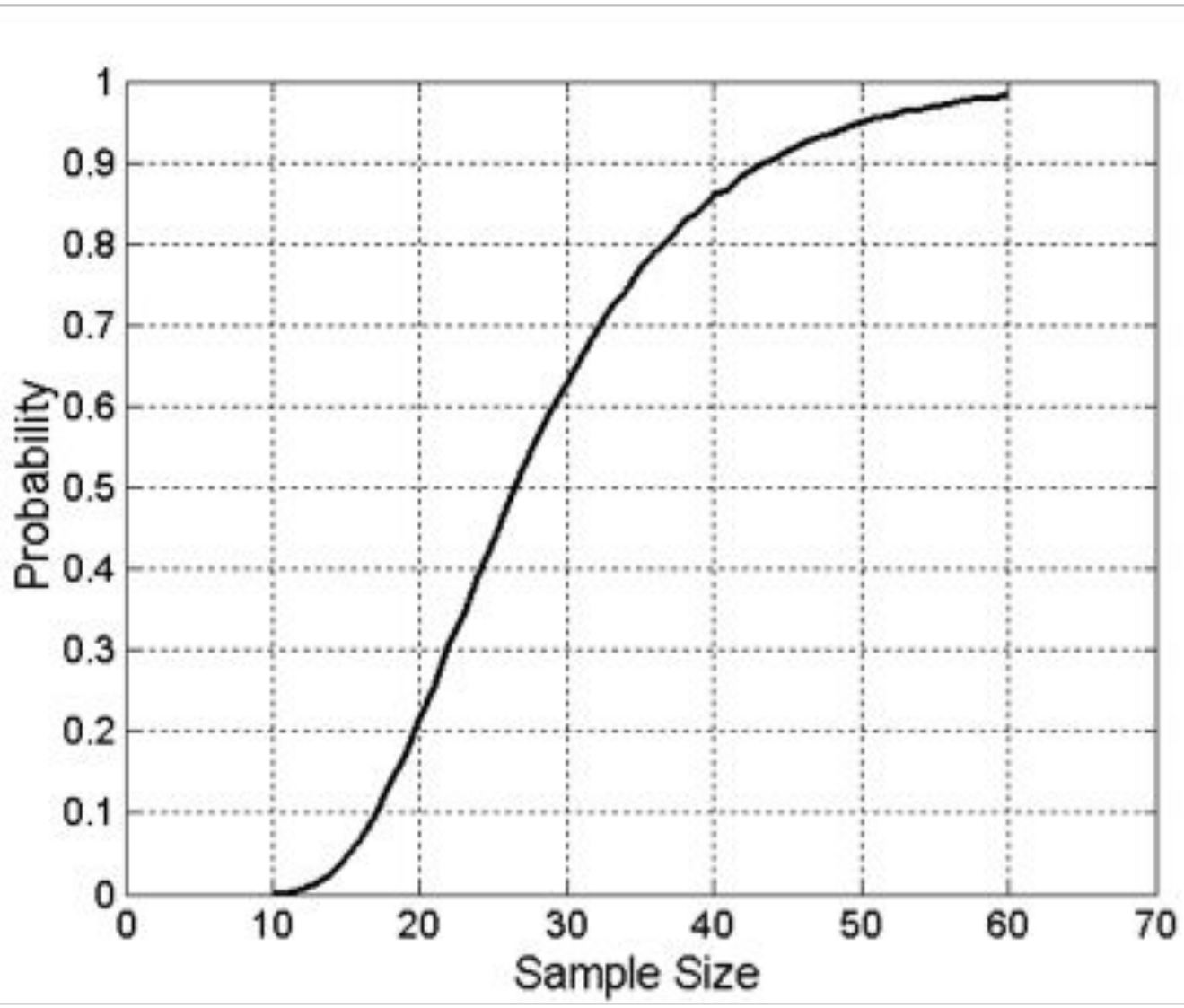




8000 points

2000 Points

500 Points

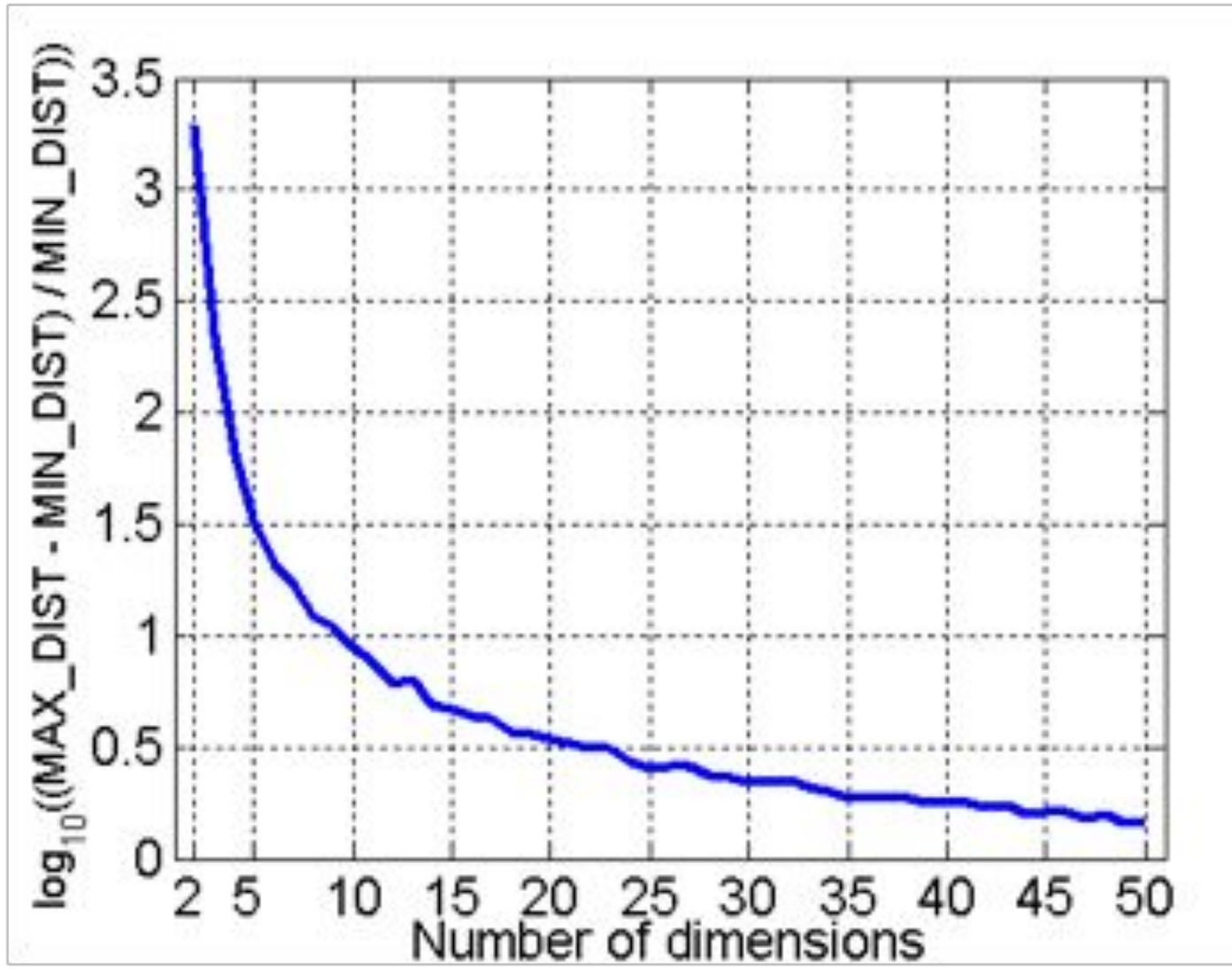


- ¿Cómo obtener al menos un objeto de cada uno de los 10 grupos?

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

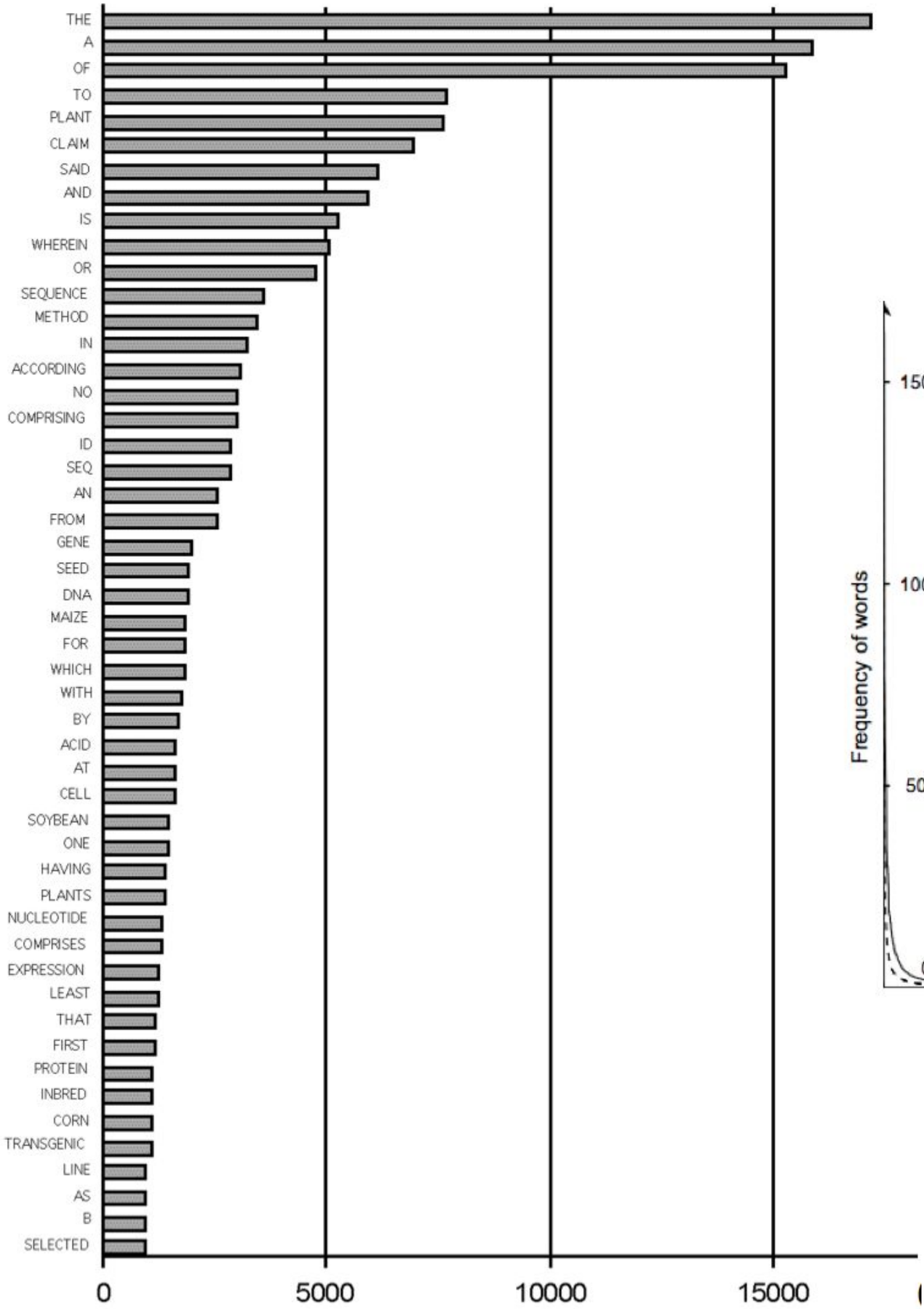
- ¿Propósito?

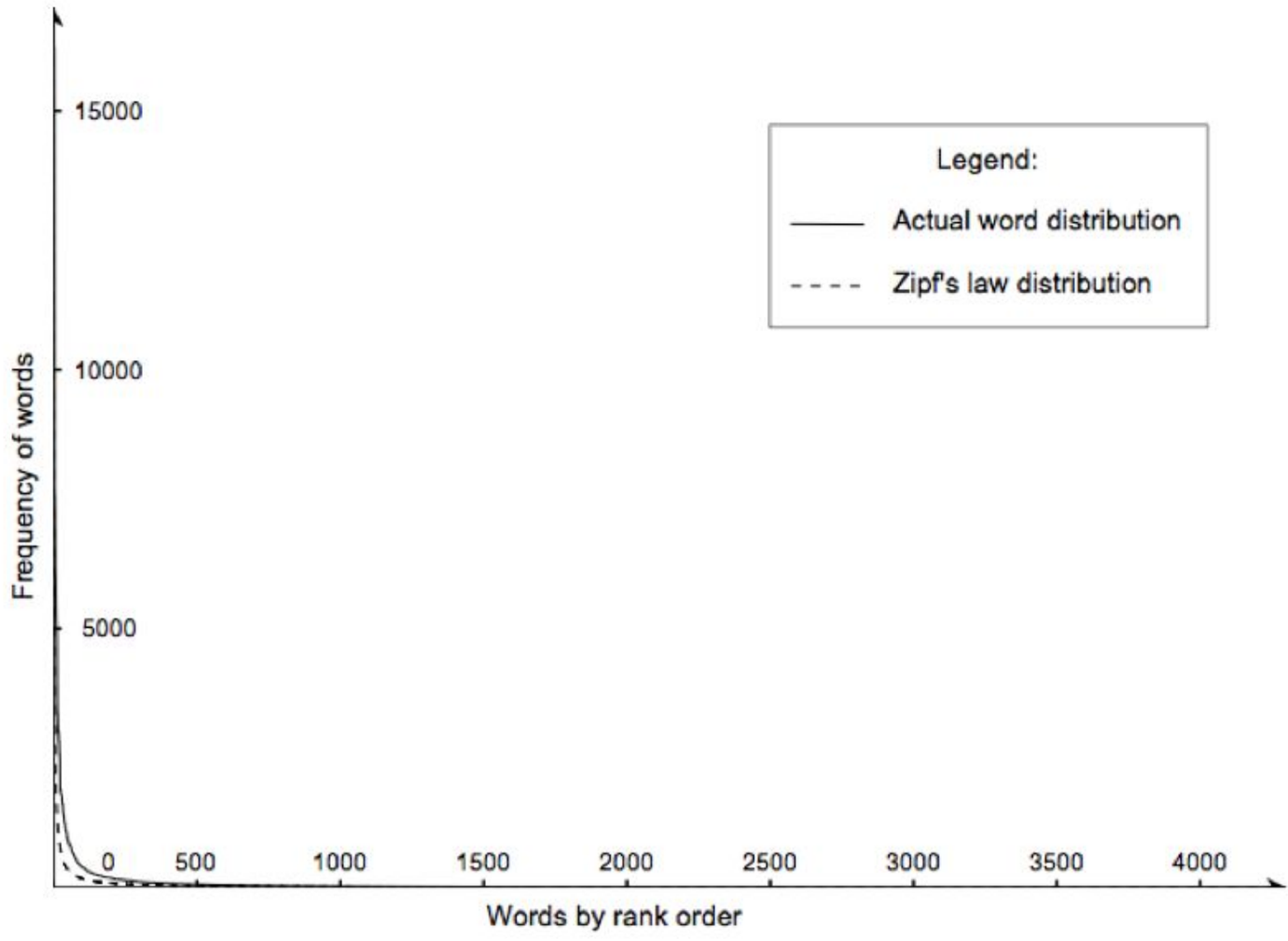
- Evitar curse of dimensionality

- Reducir costos asociados a aplicar el algoritmos (tiempo, memoria)

- Mejor visualización de los datos

- Ayuda a quitar atributos irrelevantes o ruidosos


THE
A
OF
TO
PLANT
CLAIM
SAID
AND
IS
WHEREIN
OR
SEQUENCE
METHOD
IN
ACCORDING
NO
COMPRISING
ID
SEQ
AN
FROM
GENE
SEED
DNA
MAIZE
FOR
WHICH
WITH
BY
ACID
AT
CELL
SOYBEAN
ONE
HAVING
PLANTS
NUCLEOTIDE
COMPRISES
EXPRESSION
LEAST
THAT
FIRST
PROTEIN
INBRED
CORN
TRANSGENIC
LINE
AS
B
SELECTED
0
5000
10000
15000


Frequency of words
15000
10000
5000
0
500
1000
1500
2000
2500
3000
3500
4000
Words by rank order

Legend:
Actual word distribution
Zipf's law distribution

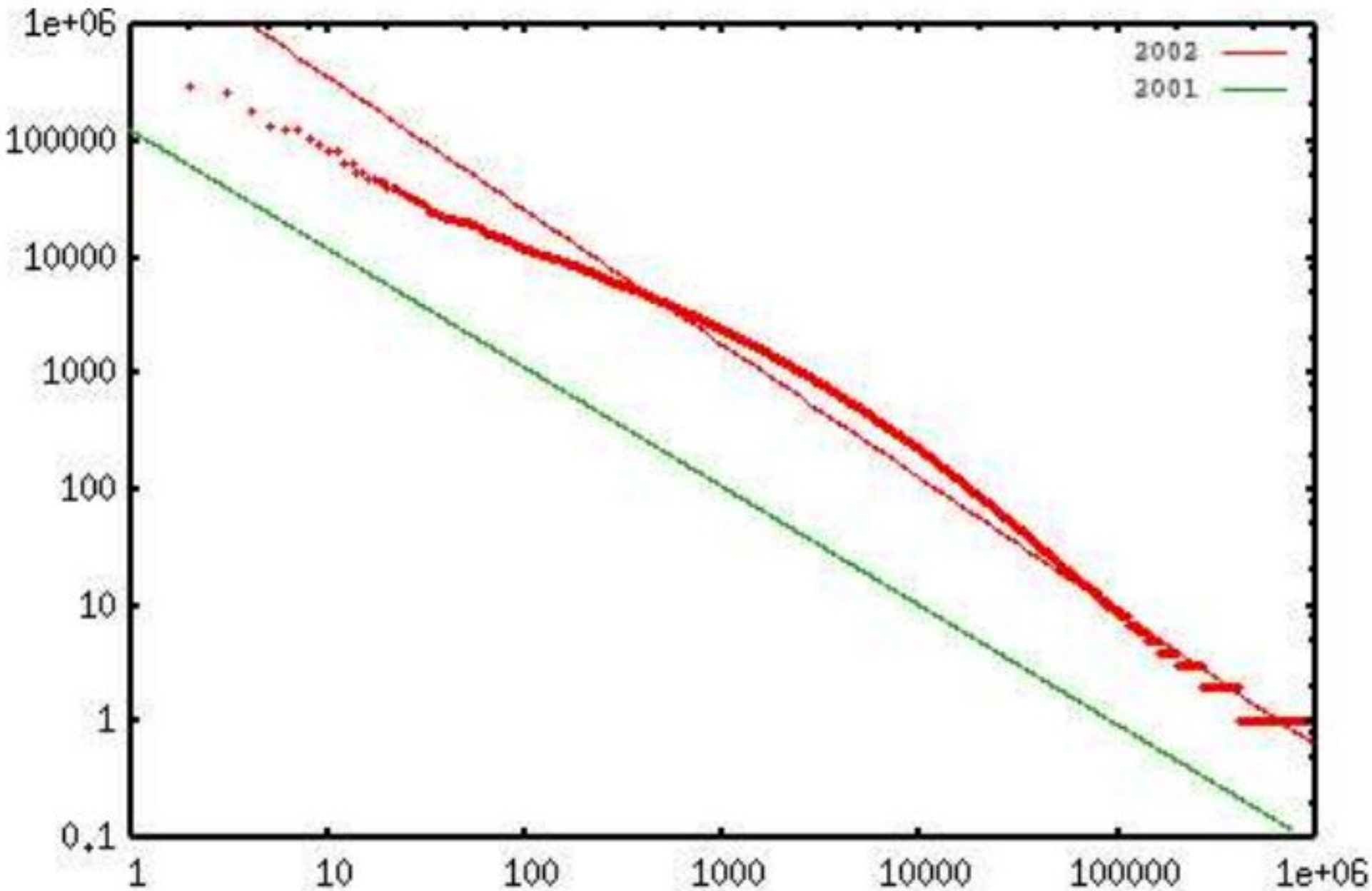


Figura 8: Frecuencia de consulta vs. Palabras.

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%	

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

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

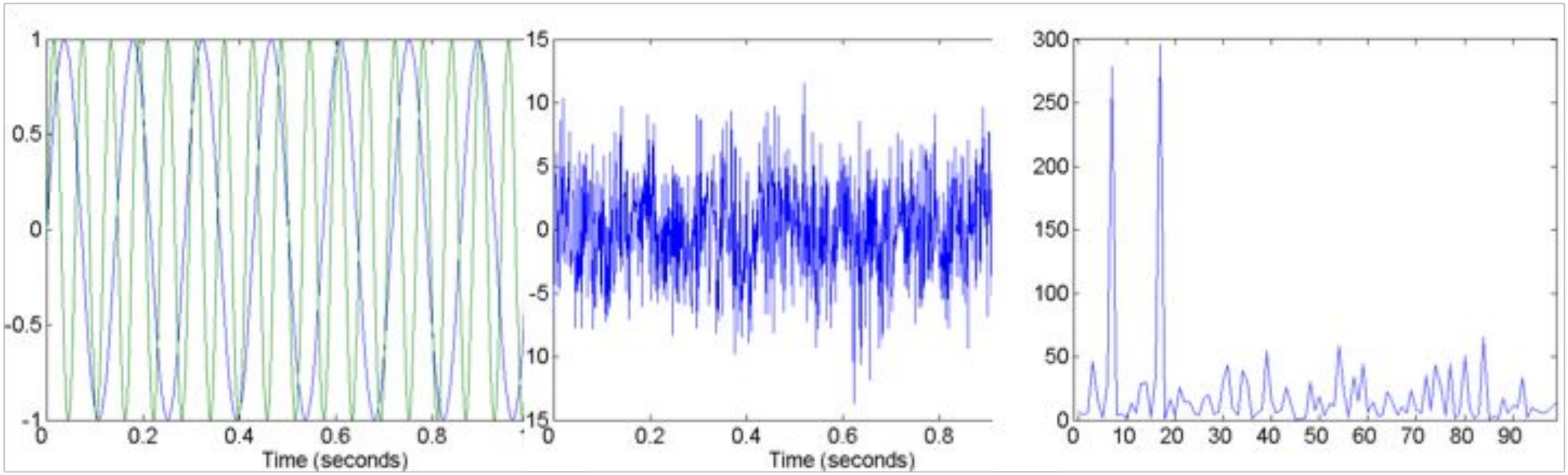
Dar pesos a los atributos

- 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

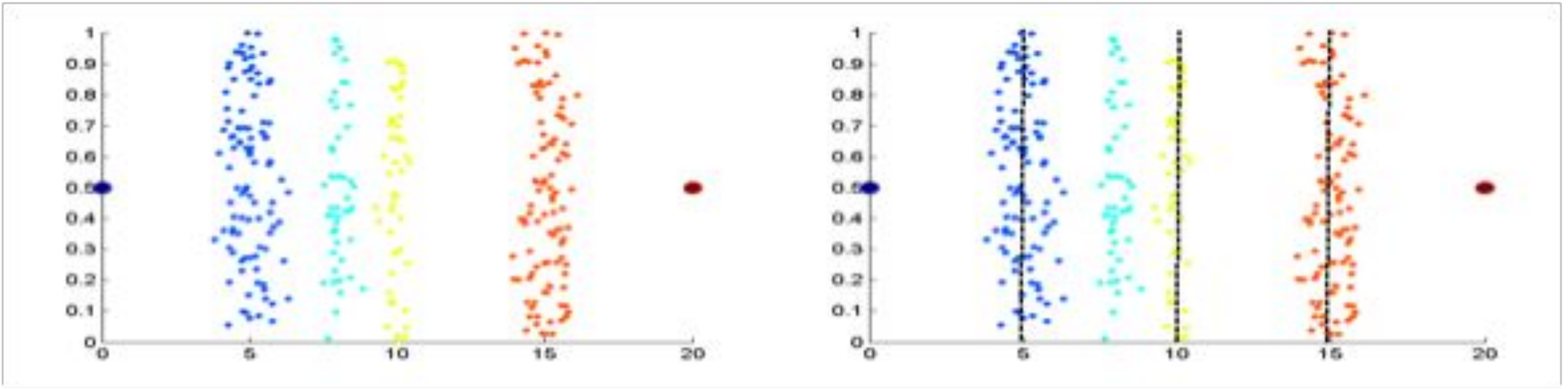
Frequency

Discretizar

- Decidir cuántas categorías tendremos

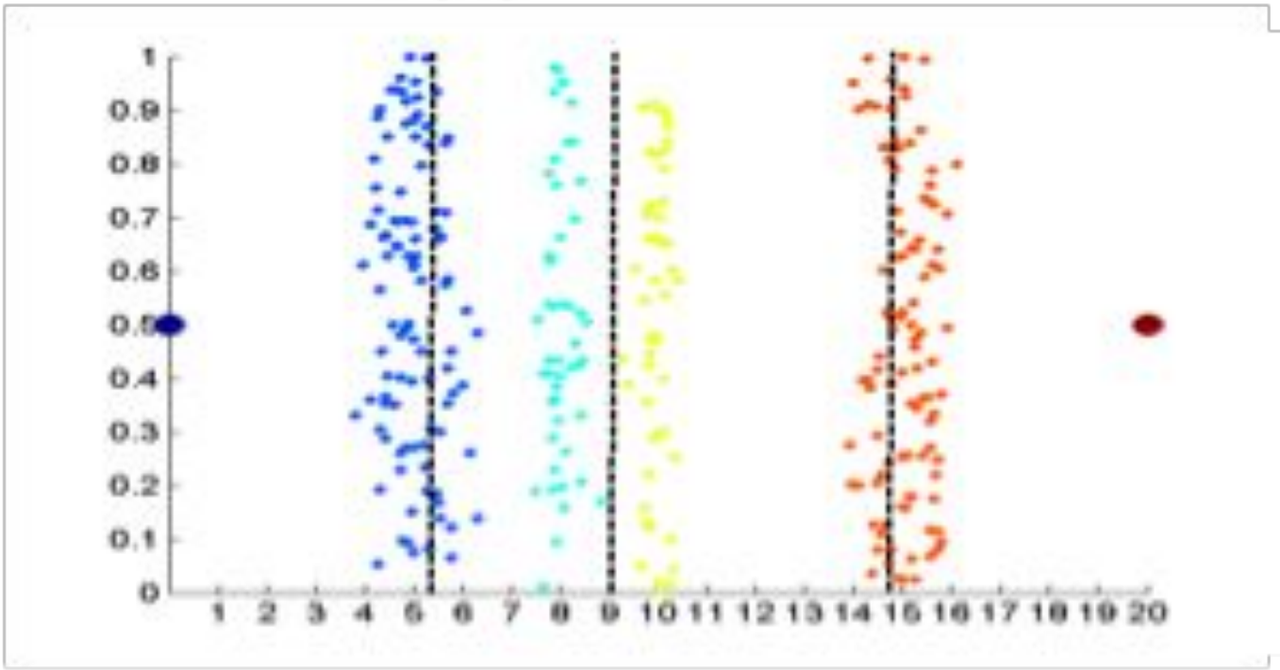
- Supervisado (con clases, considerando entropía y pureza)

- y no-supervisado (mismo intervalo, misma cantidad)

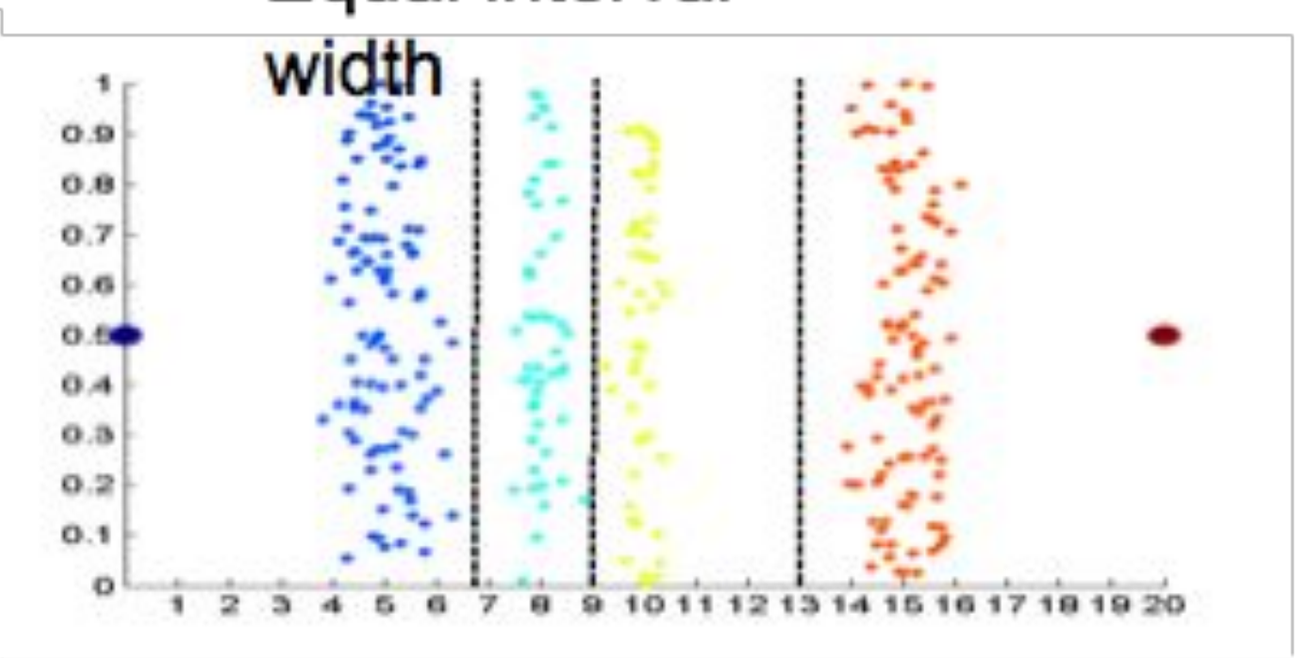


Data

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 Normalizacion

Próxima Clase

- Lab Exploración (2 sesiones: 1.1 y 1.2)

- Hito I (conformar grupos, elegir tema):

  - Exploración

  - (Pre-procesamiento)

  - Objetivos/Hipótesis iniciales (¿Por qué?)

  - primeros análisis

Proyectos de años anteriores...

- Análisis transporte en Santiago
- PIB por regiones
- Películas: Predecir ranking de las películas
- Música: Encontrar conjuntos de artistas similares para recomendación
- Recolección y análisis de datos de las elecciones presidenciales en Twitter
- Análisis de foro de u-cursos
- Análisis de datos de galaxias
- Proyecto Redes

dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

/ DCCUCHILE