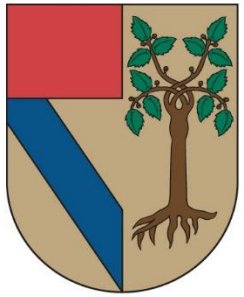


UNIVERSIDAD
PANAMERICANA

Machine Learning

(https://leonpalafox.github.io/mlclase/)

Leon F. Palafox PhD

Accuracy

- Mide el Bias
- Es una de las métricas mas usadas
- De todas las clasificaciones, cuales fueron correctas?

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision

- De todas tus detecciones positivas, cuales estan bien.
- Te da una buena sensación de que tan Bueno es el clasificador detectando positivos.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall

- Nos ayuda con sets desbalanceados
- Detecta los casos positivos

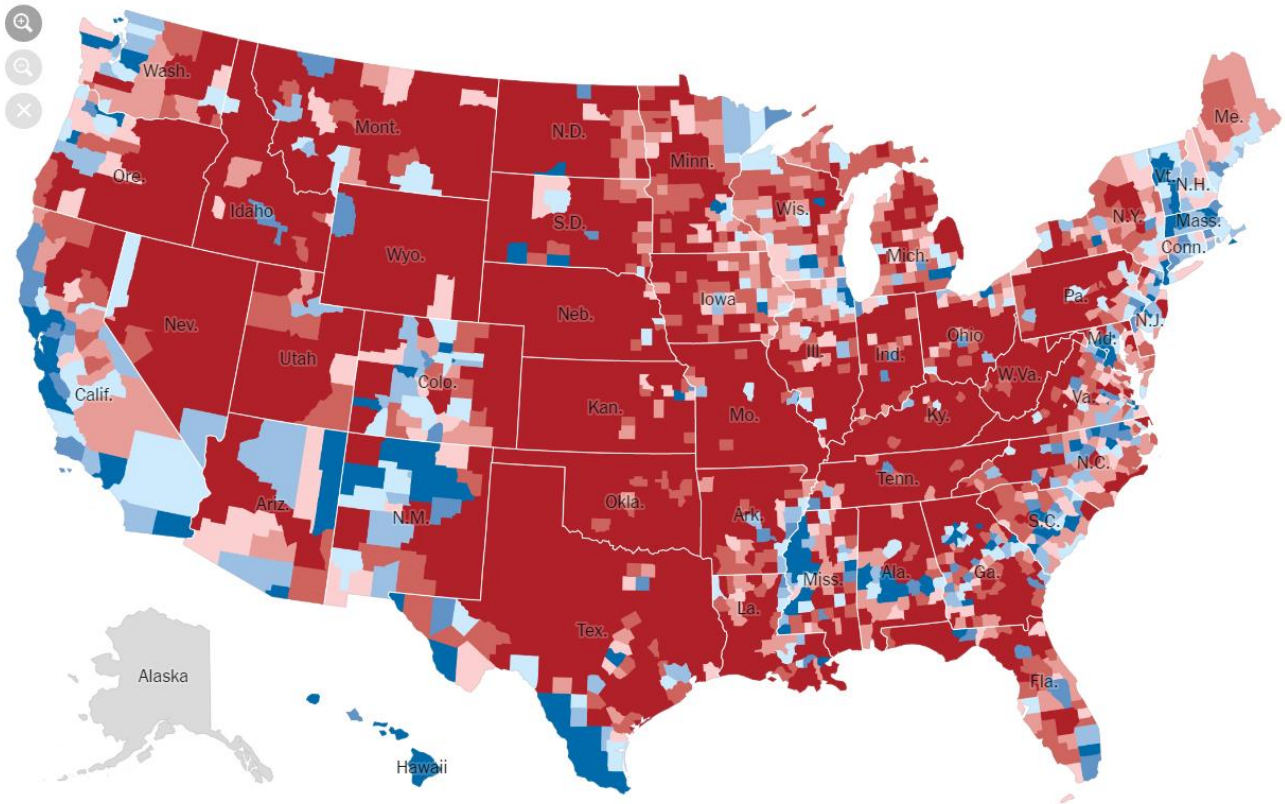
$$\text{Recall} = \frac{tp}{tp + fn}$$

F1-Score

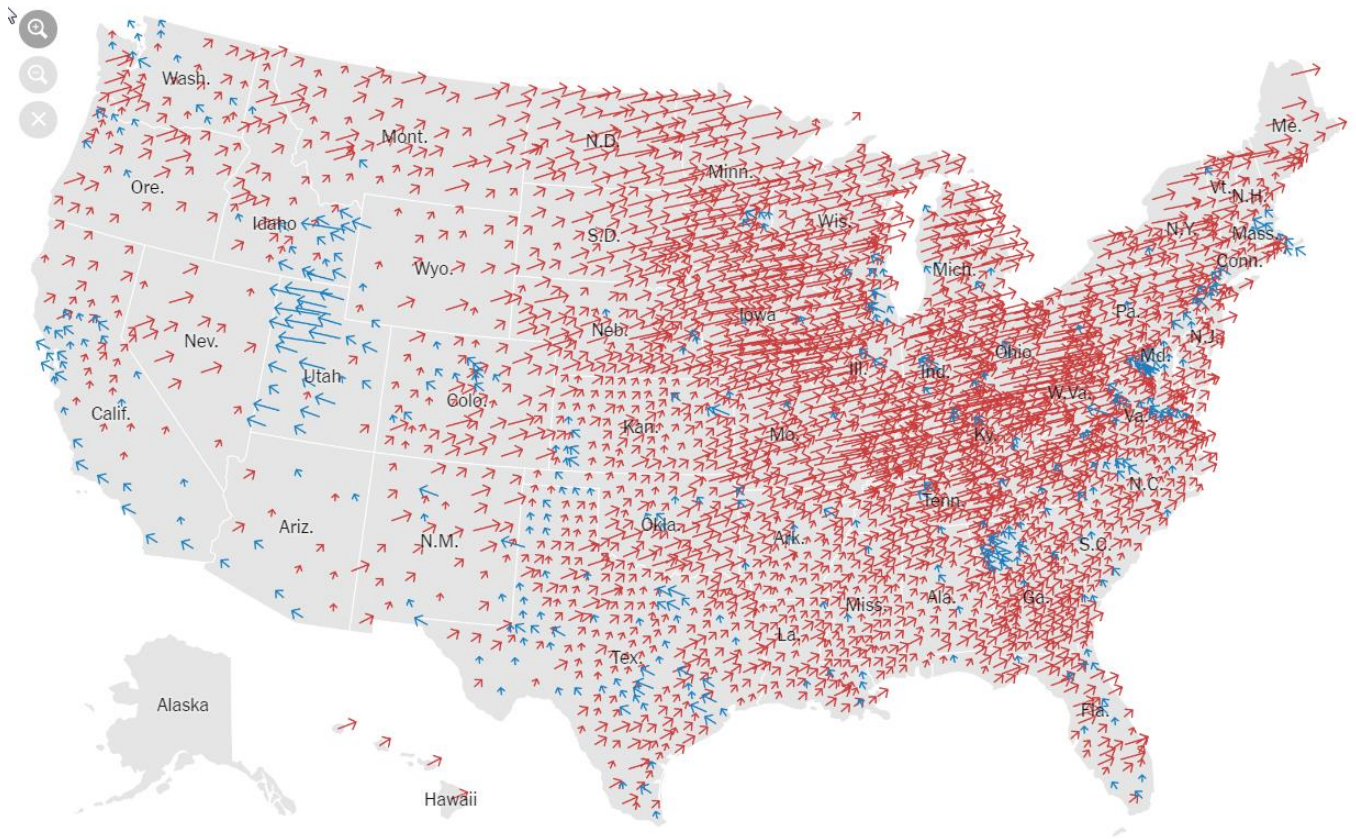
- Es una mezcla de precision y recall
- Es un solo número que nos indica que tan Bueno es el clasificador.

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

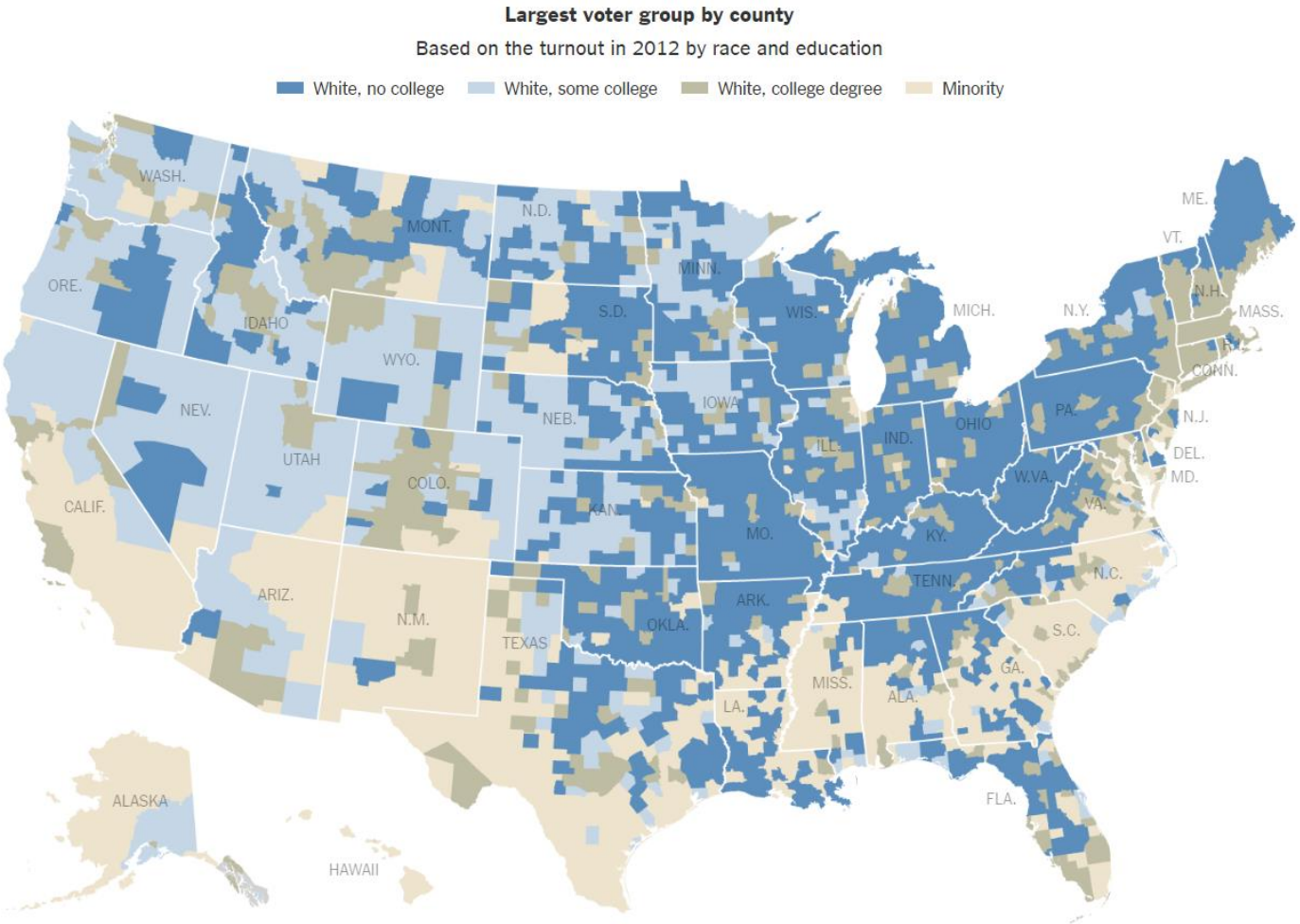
Group Activity



Group Activity



Group Activity



Manifolds

- http://cs.stanford.edu/people/karpathy/tsnejs/

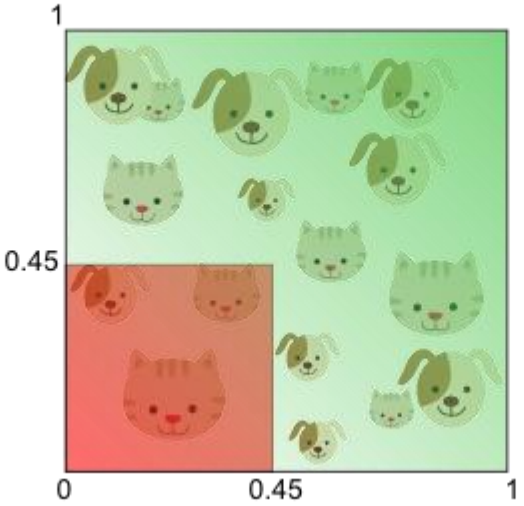
- En un manifold, los datos estan cerca los unos de los otros
- La cercania se basa en los “features”
  - Pixels
  - Palabras
  - Data

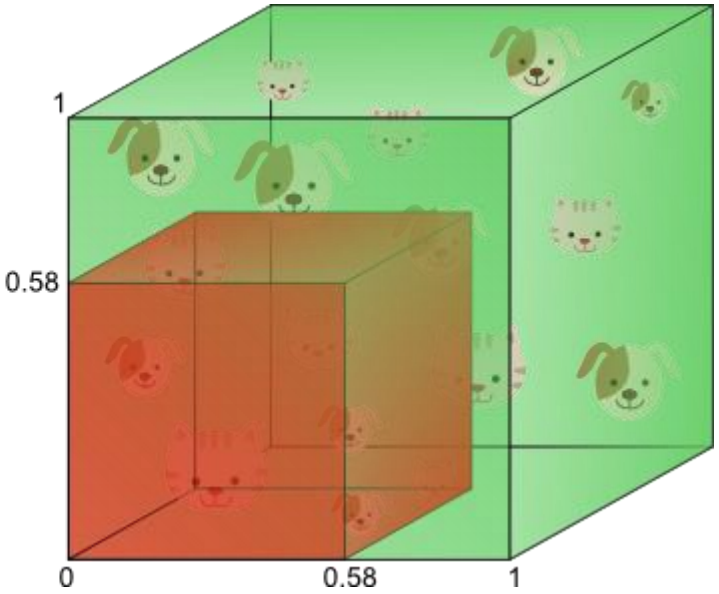
Maldición de la dimensionalidad

- Conforme tenemos más dimensiones, resulta mas dificil hacer clusters
  - También aplica a clasificadores
- Afecta la distancia euclideana
  - Si tu algoritmo usa esta distancia, ten mucho cuidado.

Maldición de la dimensionalidad

Imagina que quieres un clasificador que abarque el 20% de la población de perros y gatos

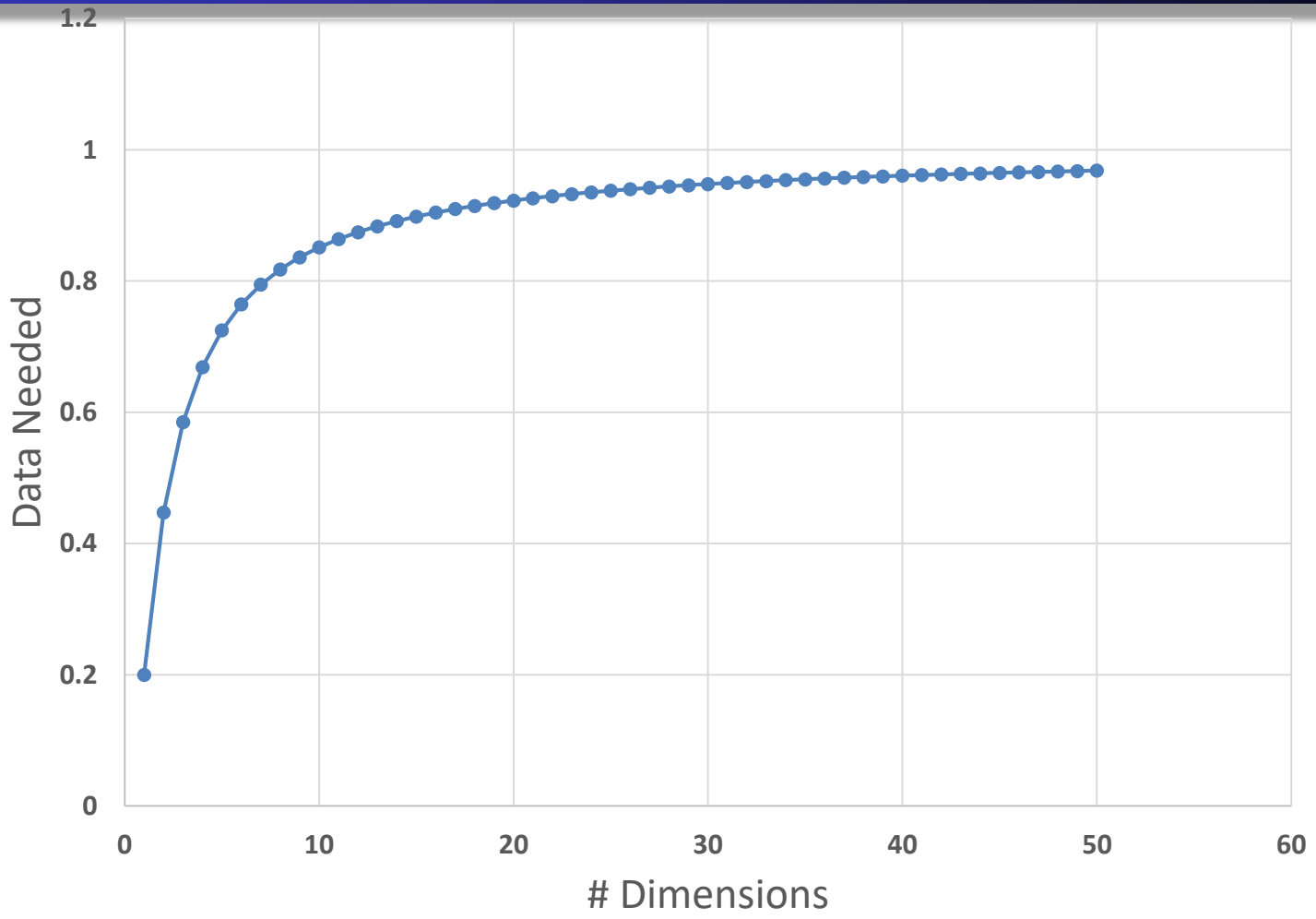




0.45^2 = 0.2

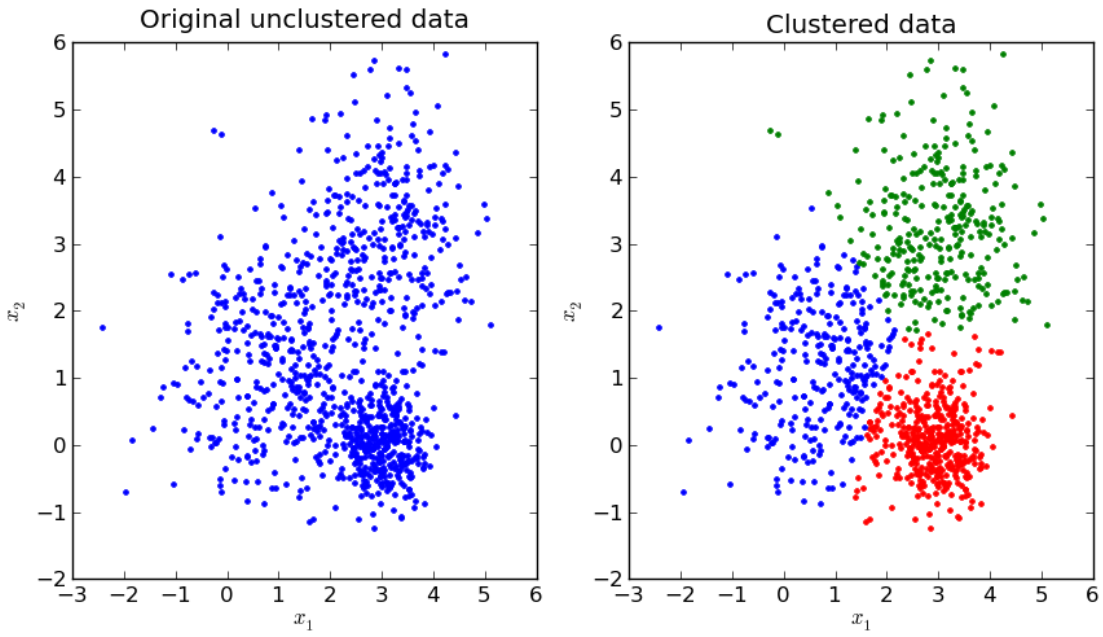
0.58^3 = 0.2

Curse of dimensionality

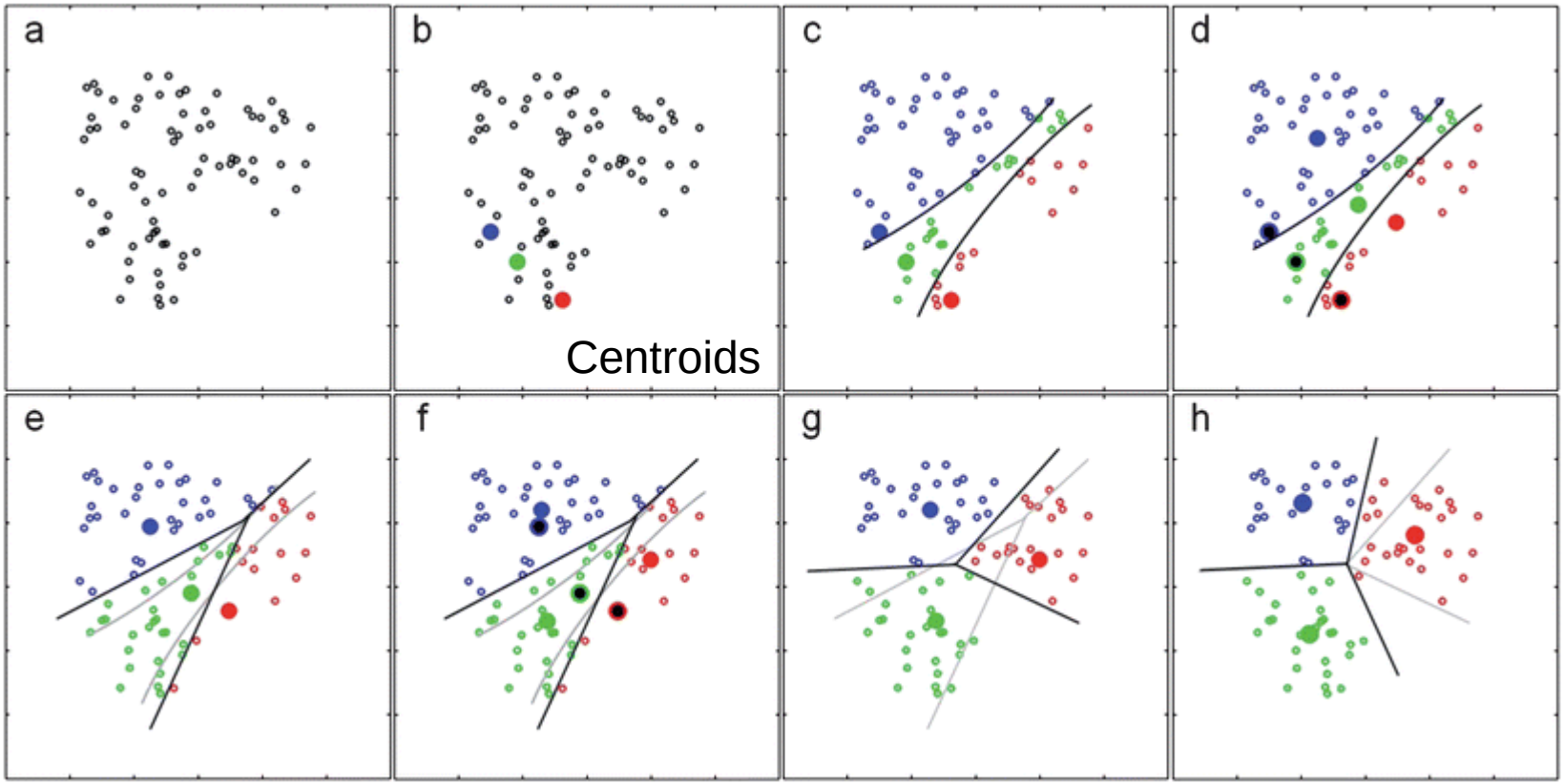


K-Means

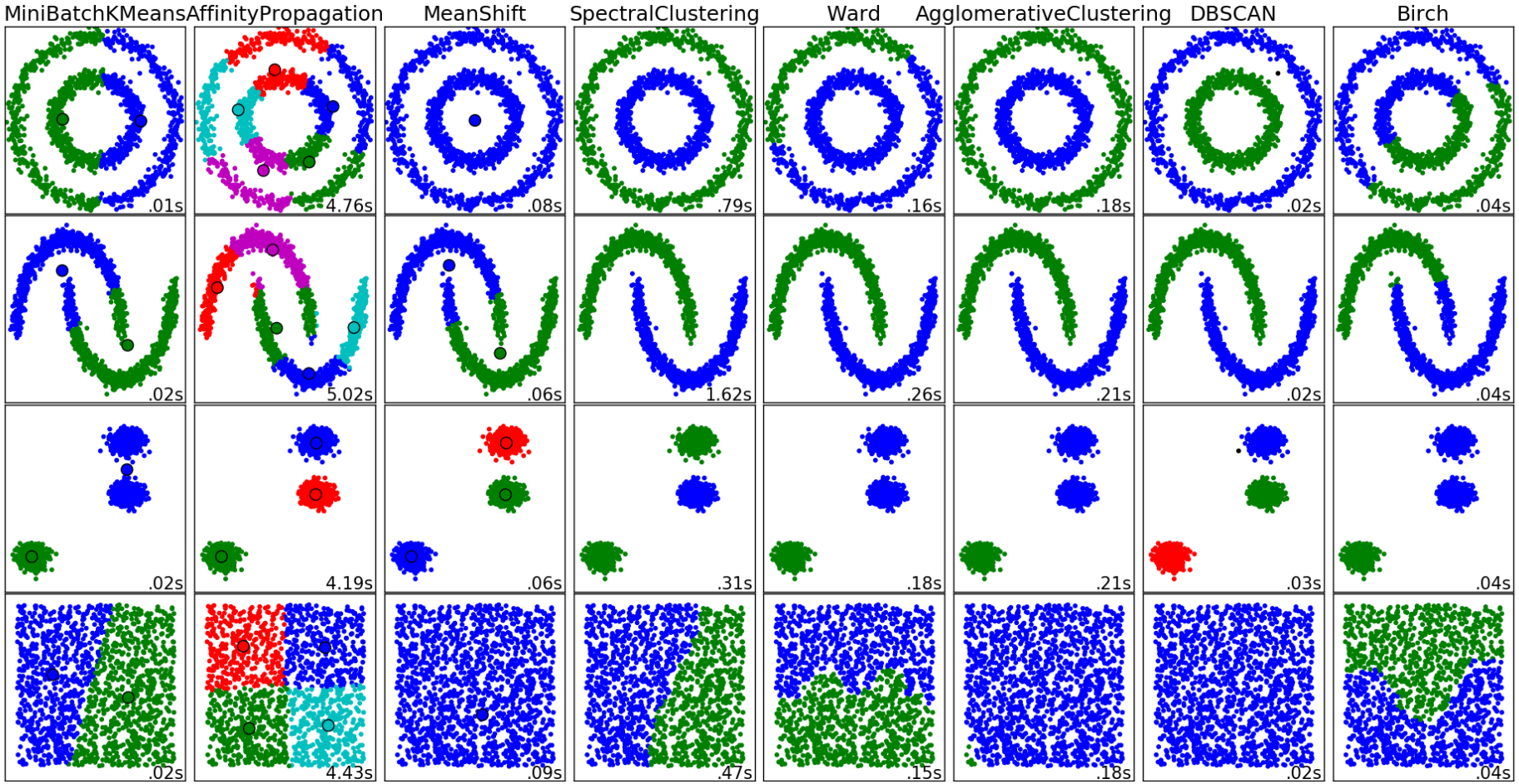
- http://kluster.j38.net/



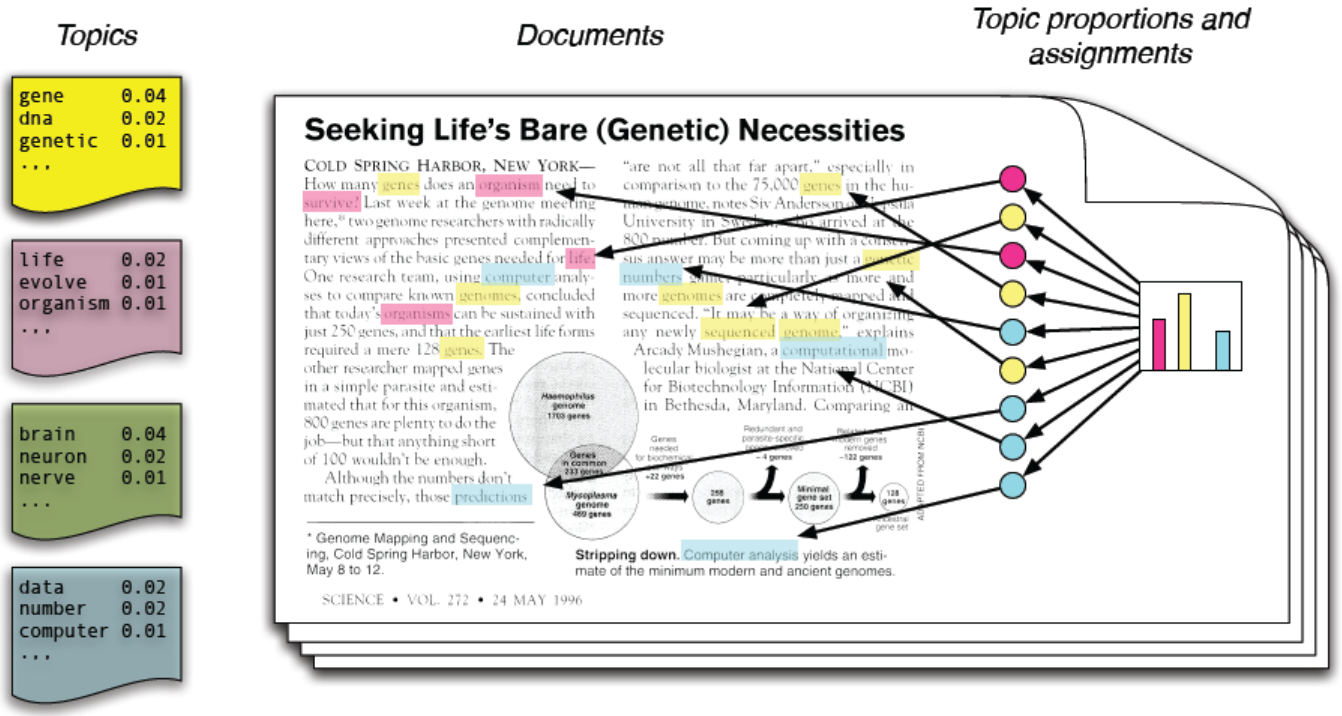
K-Means



Other Clustering Techniques



Latent Dirichlet Allocation (LDA)



https://www.lpl.arizona.edu/~leonp/lplpapers2014/lpl2014lda5Topics.html
https://www.lpl.arizona.edu/~leonp/HiRISE/HiRISELDA5Topics.html

Como evaluamos K-Means

- Sin etiquetas:
  - Davies-Boudin Index

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

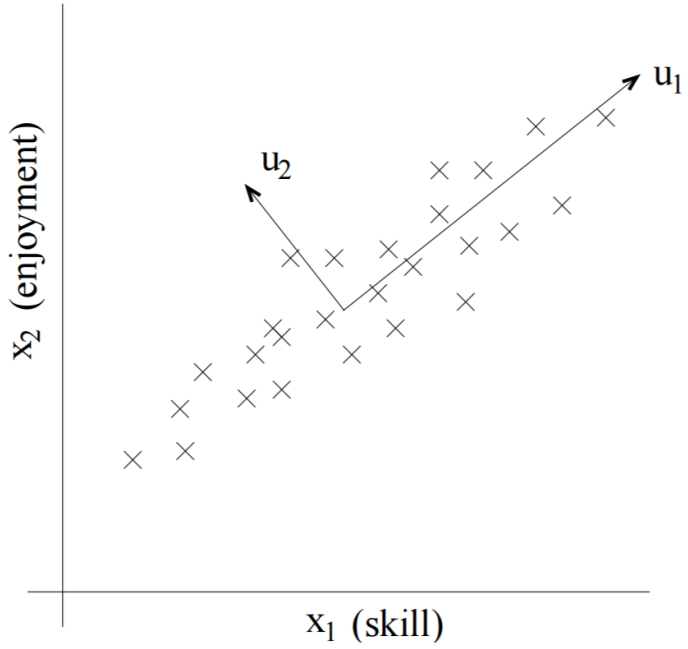
  - σ es la distancia promedio de los elementos del cluster al centroide
  - d(c,c) es la distancia entre centroides.
  - Da scores bajos para baja intra-distancia y alta inter-distancia.

Escenario

- Estamos tratando de detectar si alguien es buen conductor o mal conductor.
- Hacemos una encuesta:
  - Medimos distintas variables:
    - Skill (basado en metricas)
    - Diversion (basado en tiempo)
    - Precisión
    - Edad, genero, etc

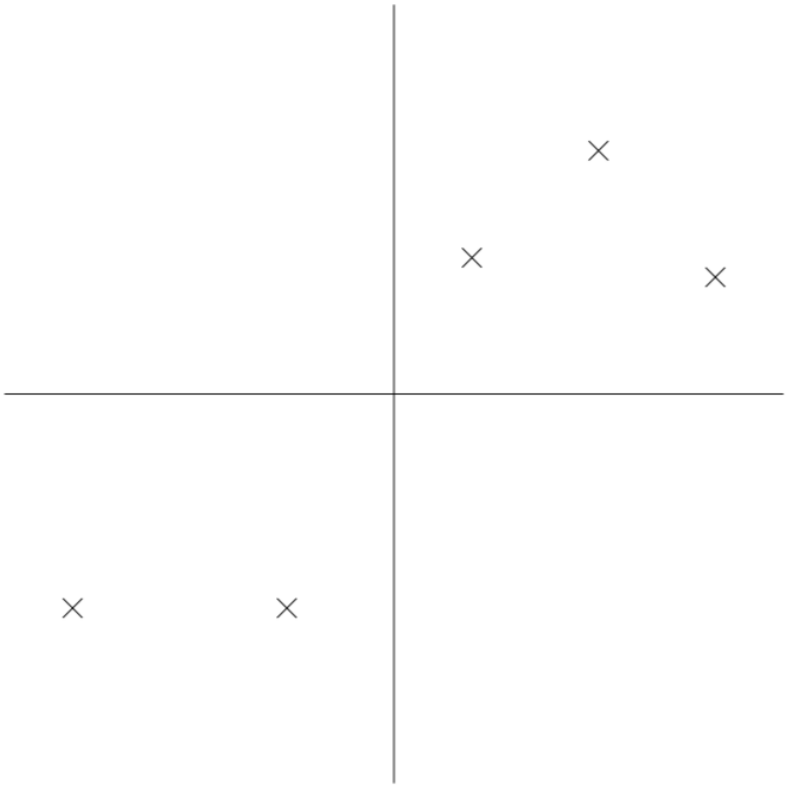
Hay variables que sobran

- Creamos una nueva variable
  - Esta variable es el “karma” de cada usuario
  - La nueva variable captura la varianza



PCA

- Buscar la dirección donde los datos varían mas



Options, options!