Applied Data Science
fall 2017
Session 8: Clustering

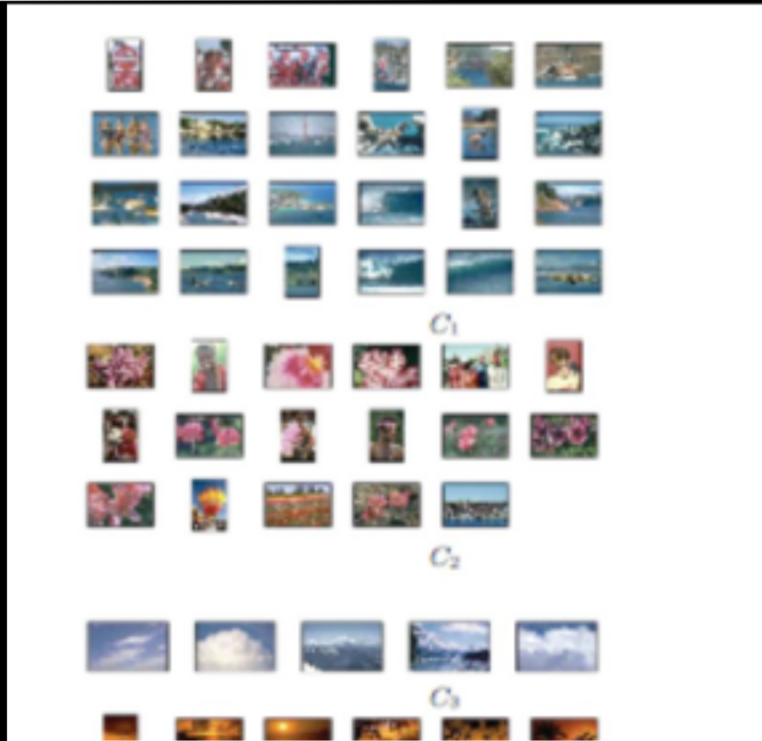
Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, Maxim Temnogorod

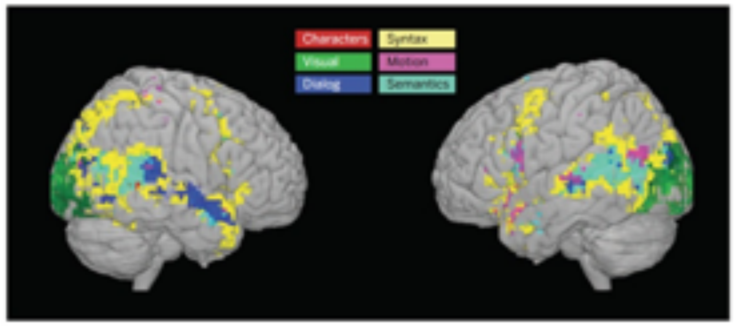
So what is the clustering?

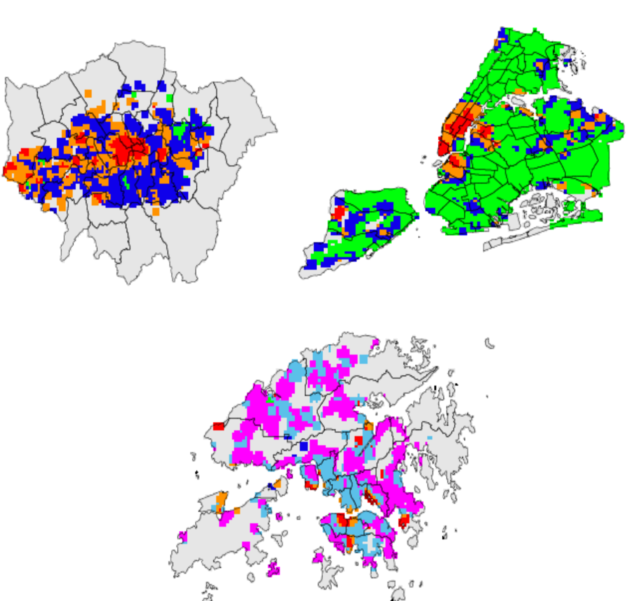
group objects of similar characteristics such that within-group similarity is higher compared to between-group similarity

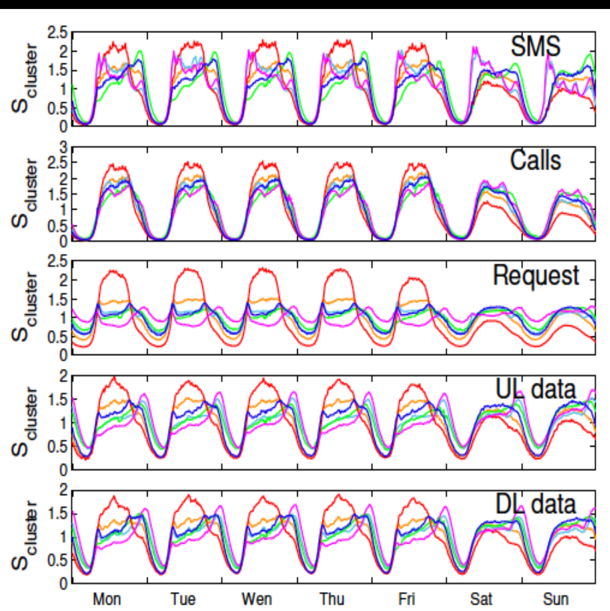
Clustering - applications

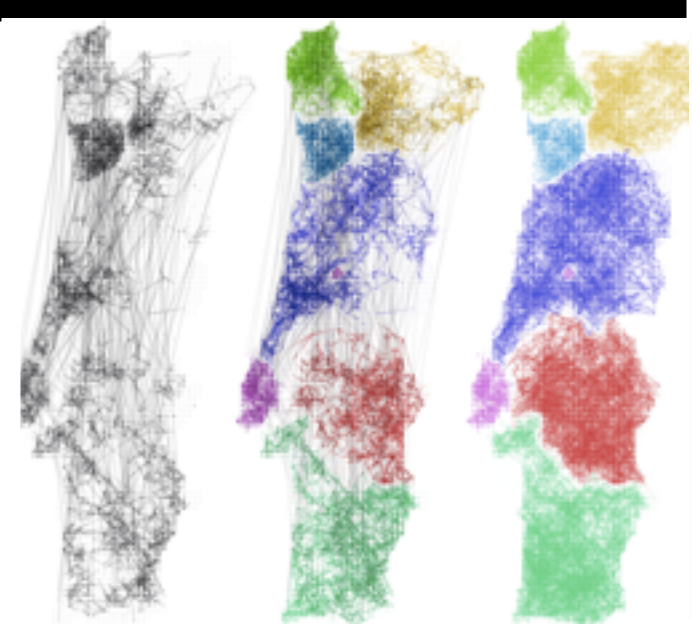
- similar consumers into market segments
- topic detection - groups of similar messages
- groups of similar text documents
- groups of connected individuals: community detection in social networks
- similar areas - neighborhood typology
- land use classification
- connected areas - regions
- similar noise samples - noise/speech recognition
- image compression - cluster pixels by RGB value
- remove duplicate or near-duplicate records
- criminal hotspots
- brain activity patterns

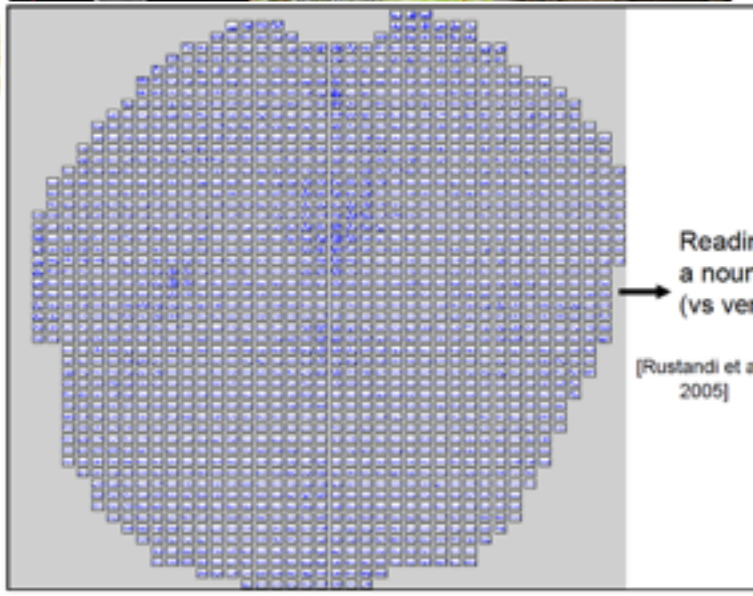




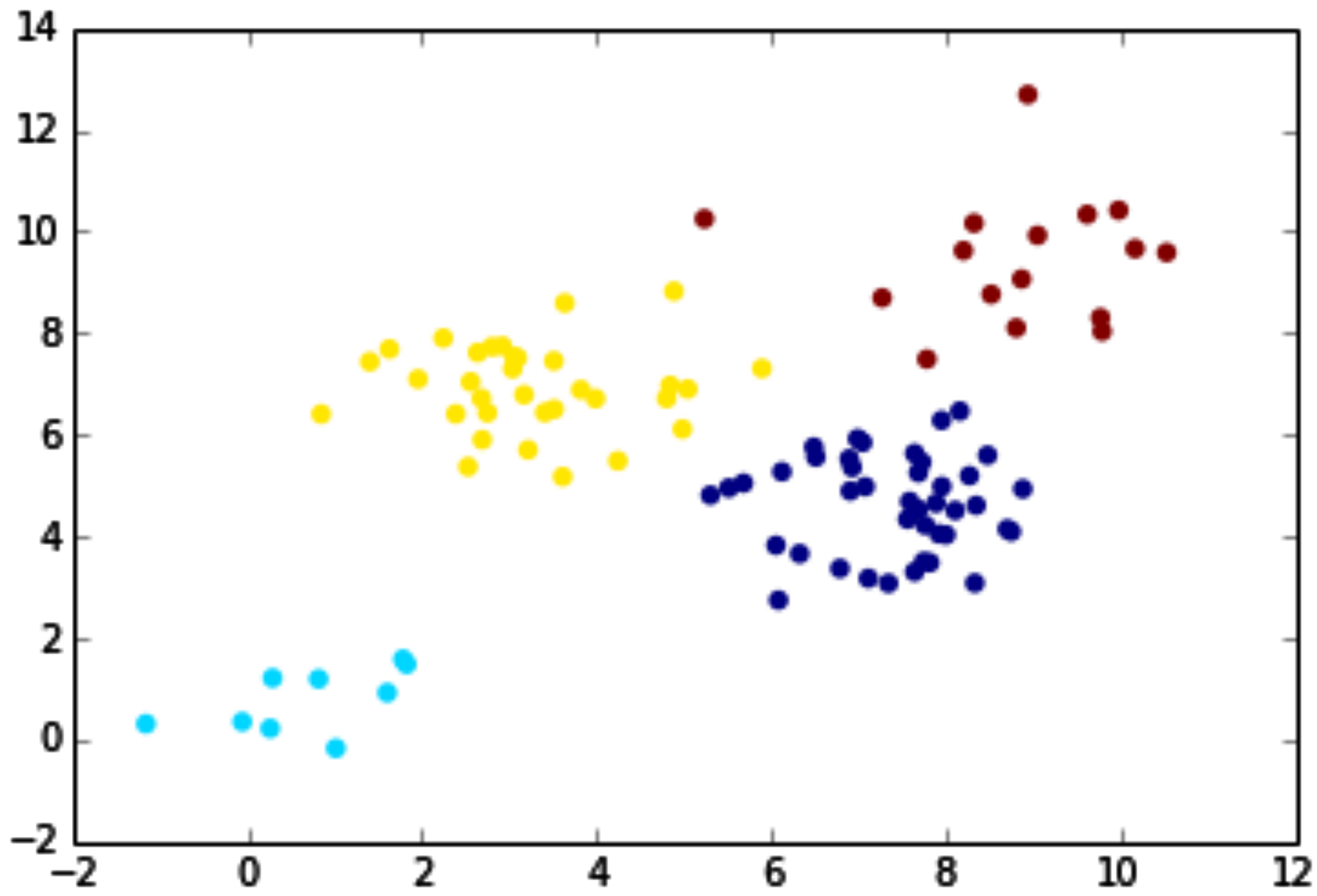








Clustering



Clustering

Given the data points

$$X = \{x_i, i = 1..N\} = \{x_i^j, i = 1..N, j = 1..n\}$$

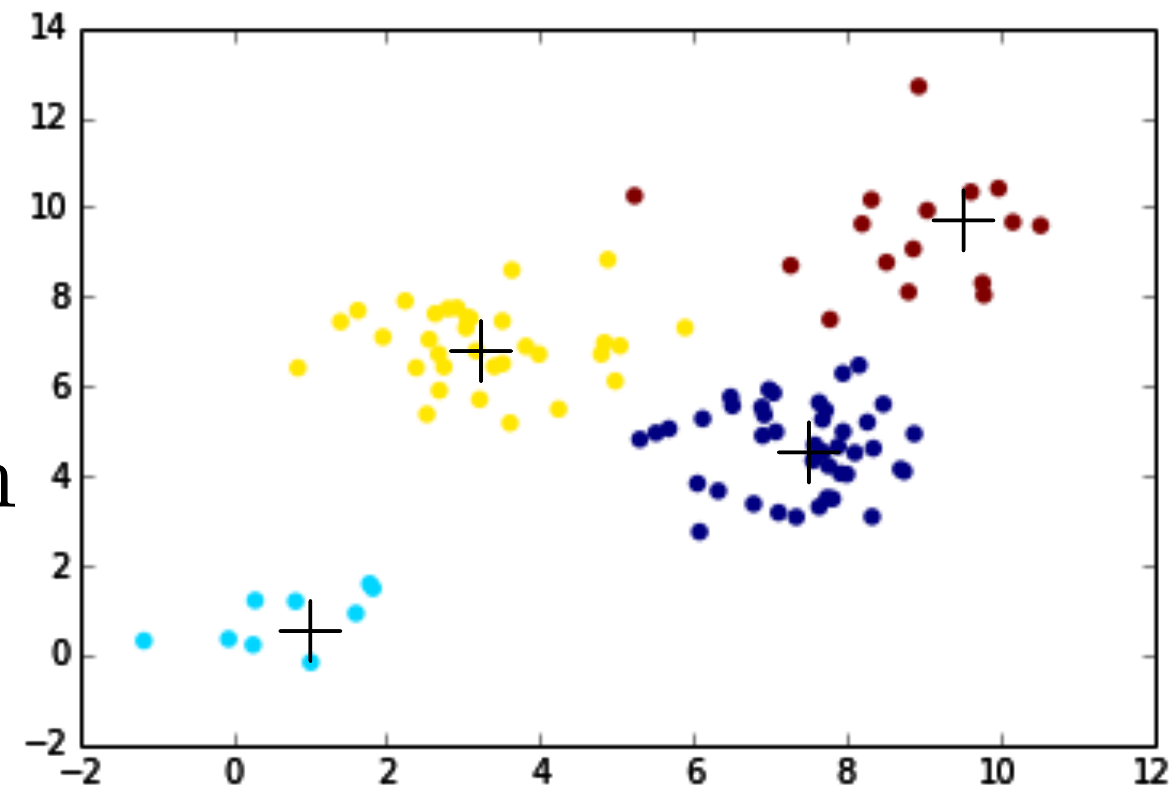
Assign cluster numbers

$$c_i = 1, 2, ..., M$$

Centroids

Where to put a point x^*, y^*

$$\sum_i \left[(x_i - x^*)^2 + (y_i - y^*)^2\right] \to \min$$



$$x^* = \sum_i x_i/N \quad y^* = \sum_i y_i/N$$

$$\mu = \sum_i x_i/N$$

Clusters define centroids and vice versa

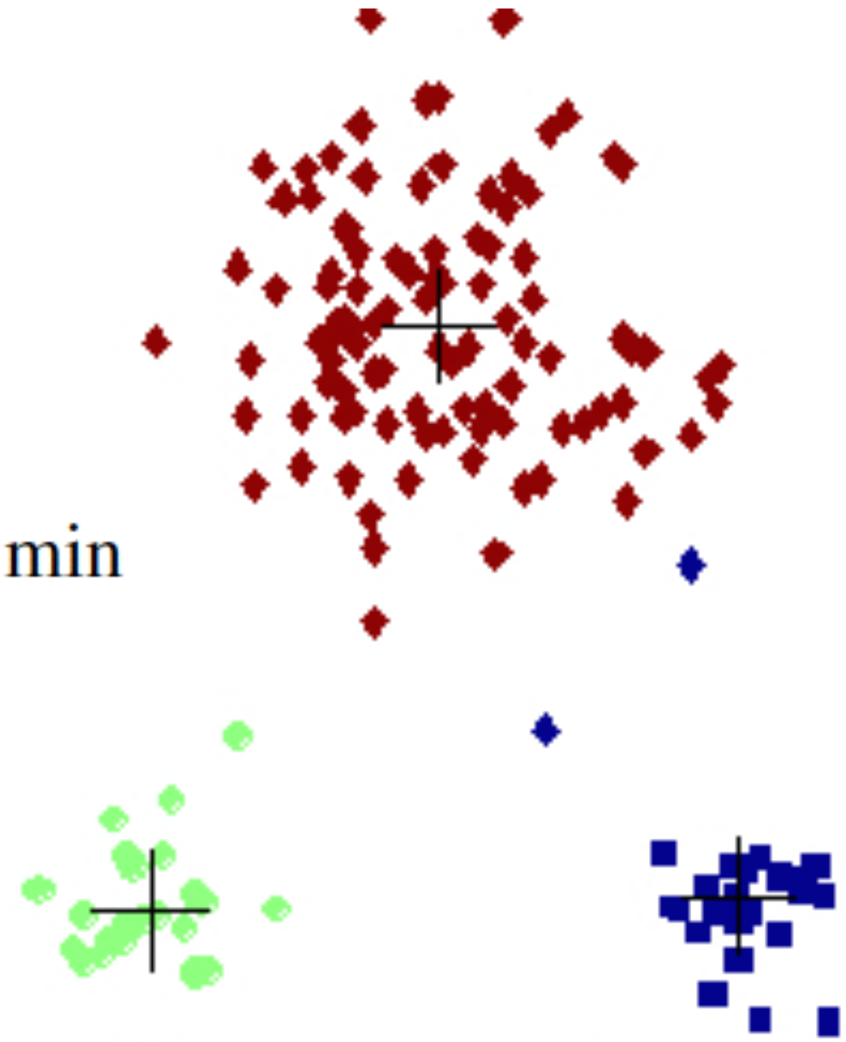
K-means clustering

Look for $c_i = 1, 2, ..., M$

For each cluster let

$$\mu_c = \sum_{i,c(i)=c} x_i / m(c)$$

$$SD = \sum_i \|x_i - \mu_{c_i}\|^2 = \sum_{i,j} \left(x_i^j - \mu_{c_i}^j\right)^2 \to \min$$

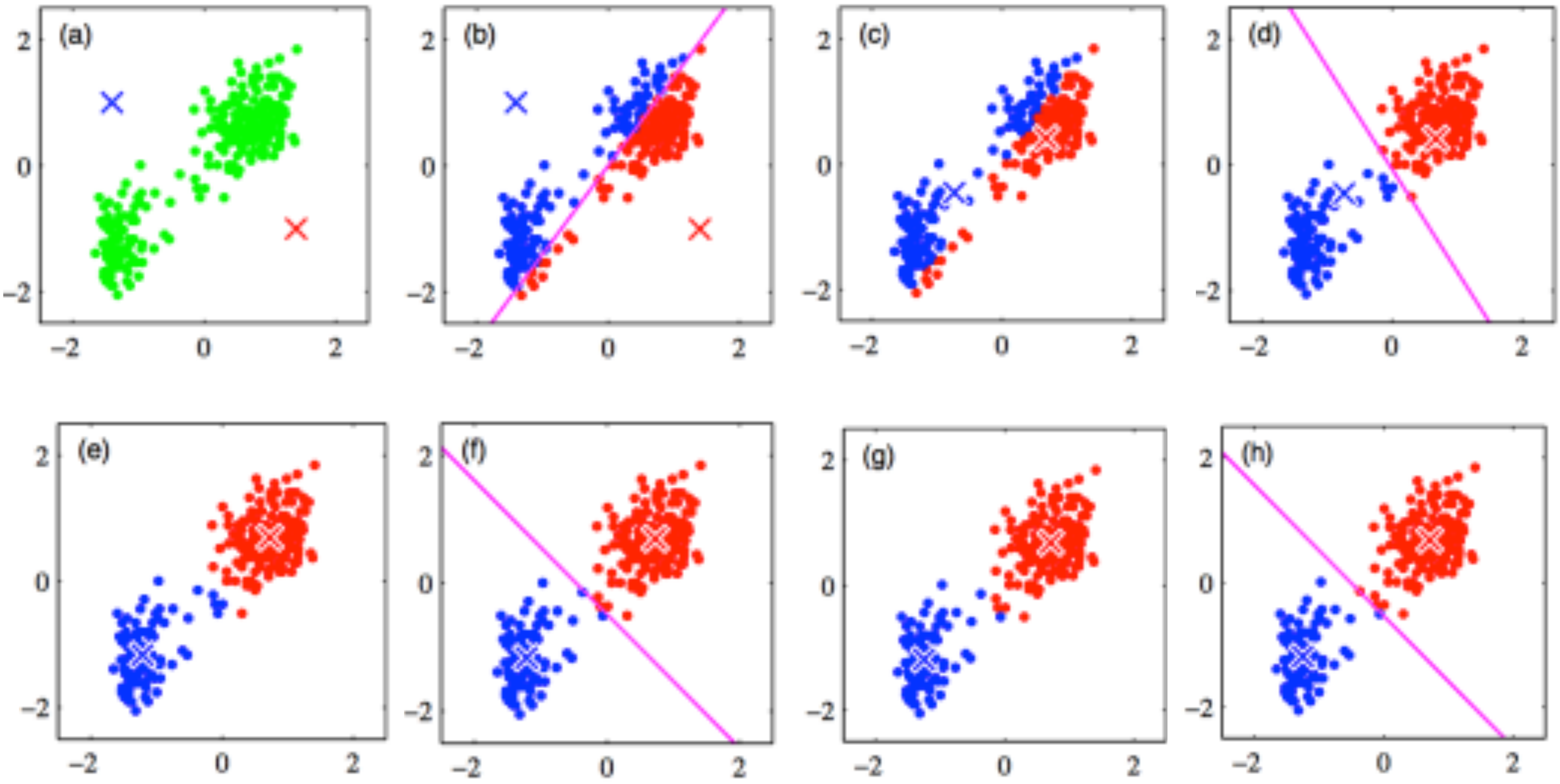


How small is the variance with respect to knowing the clusters/ cluster centroids vs the original variance of the dataset?

K-means clustering

A. Start with random cluster centroids
B. Attach each point to the closest centroid creating clusters
C. Re-compute the centroids
D. If centroids have shifted repeat from B, otherwise – stop

K-means clustering



Issues with k-means

- Stability - sensitive to initial cluster centers choices
- Which distance metrics is the right one?
- Choosing the correct number of clusters
- Real clusters may not be spherical (or similar size)

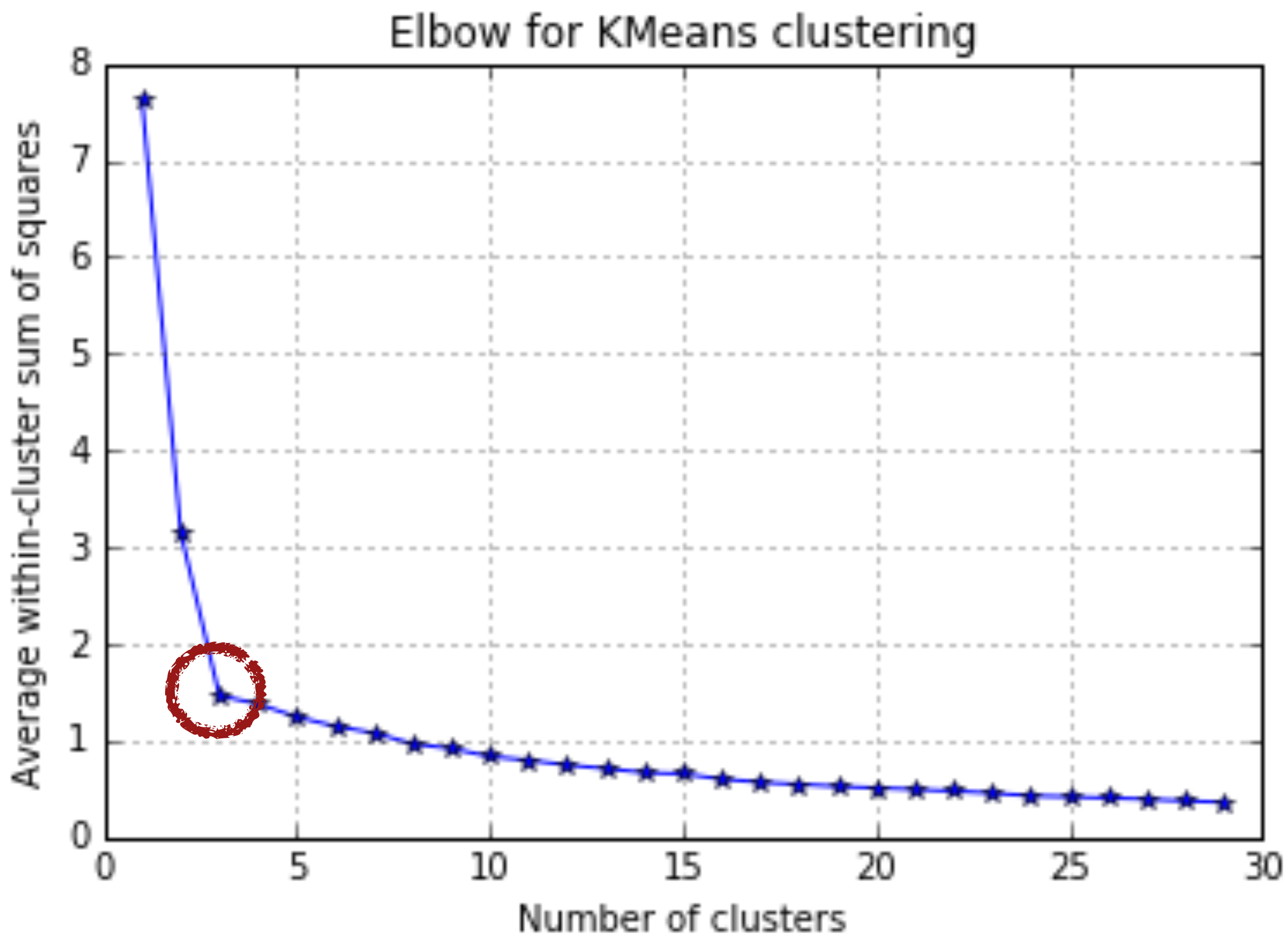
Alternative distances

$$SD = \sum_i \|x_i - \mu_{c_i}\| = \sum_i \sqrt{\sum_j \left(x_i^j - \mu_{c_i}^j\right)^2} \to \min$$

$$\mu_c \in \{x_i : c_i = c\}$$

Selecting the number of clusters: Elbow method

$$SD = \sum_i \|x_i - \mu_{c_i}\|^2 = \sum_{i,j} \left(x_i^j - \mu_{c_i}^j\right)^2$$ vs number of clusters

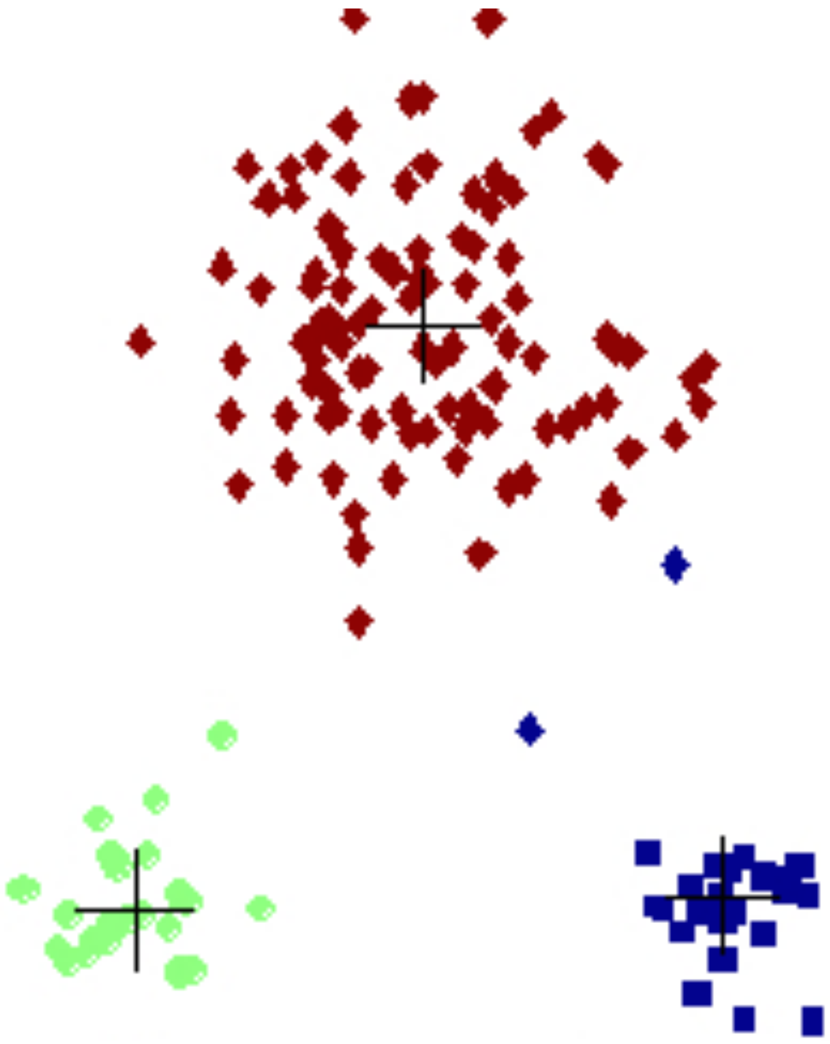


Selecting the number of clusters - Silhouette

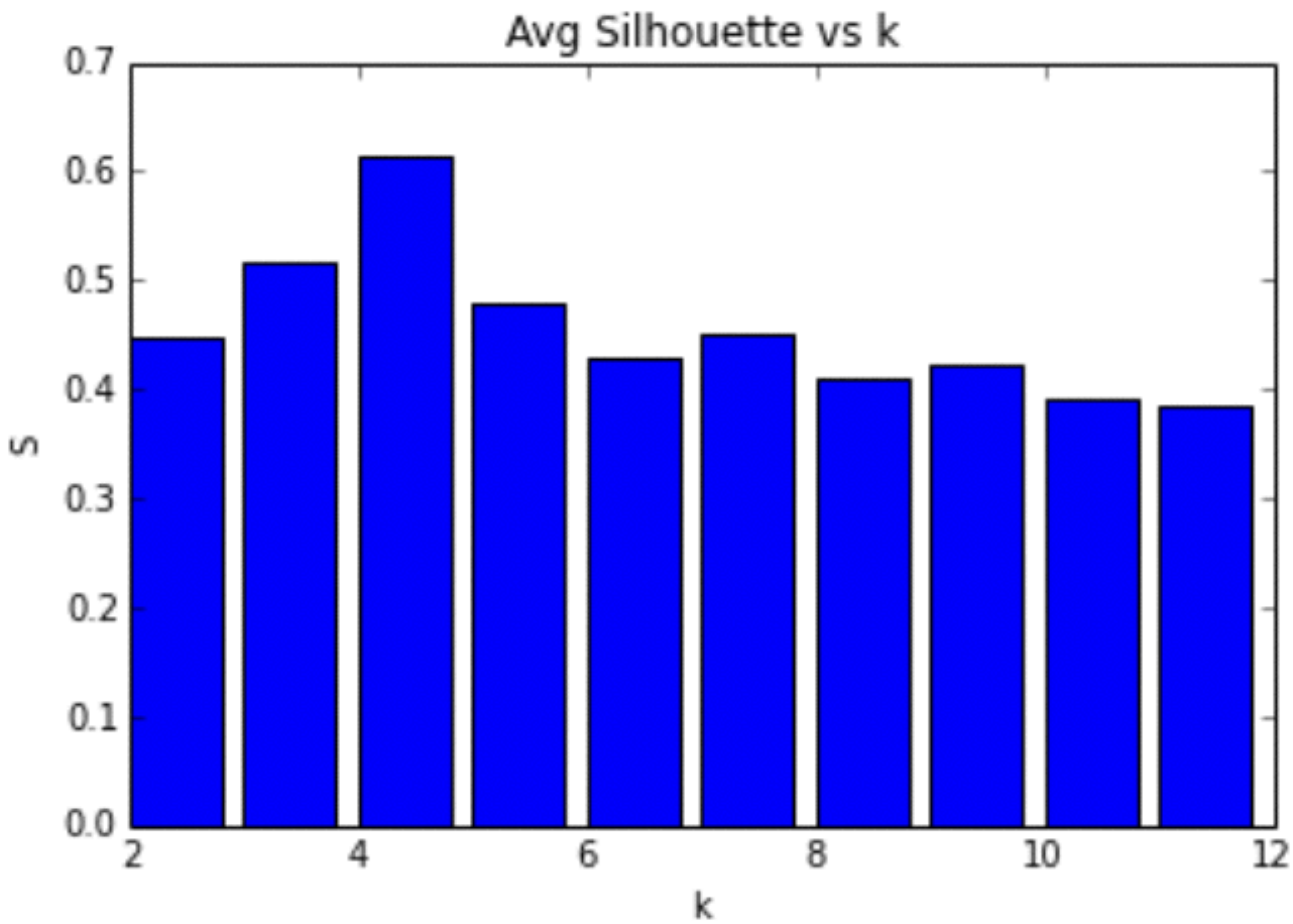
$$s(i) = \frac{\min_{k \neq c_i} \|x_i - \mu_{c_k}\| - \|x_i - \mu_{c_i}\|}{\max\{\|x_i - \mu_{c_i}\|, \min_{k \neq c_i} \|x_i - \mu_{c_k}\|\}}$$

$$S = \frac{\sum_i s(i)}{N}$$

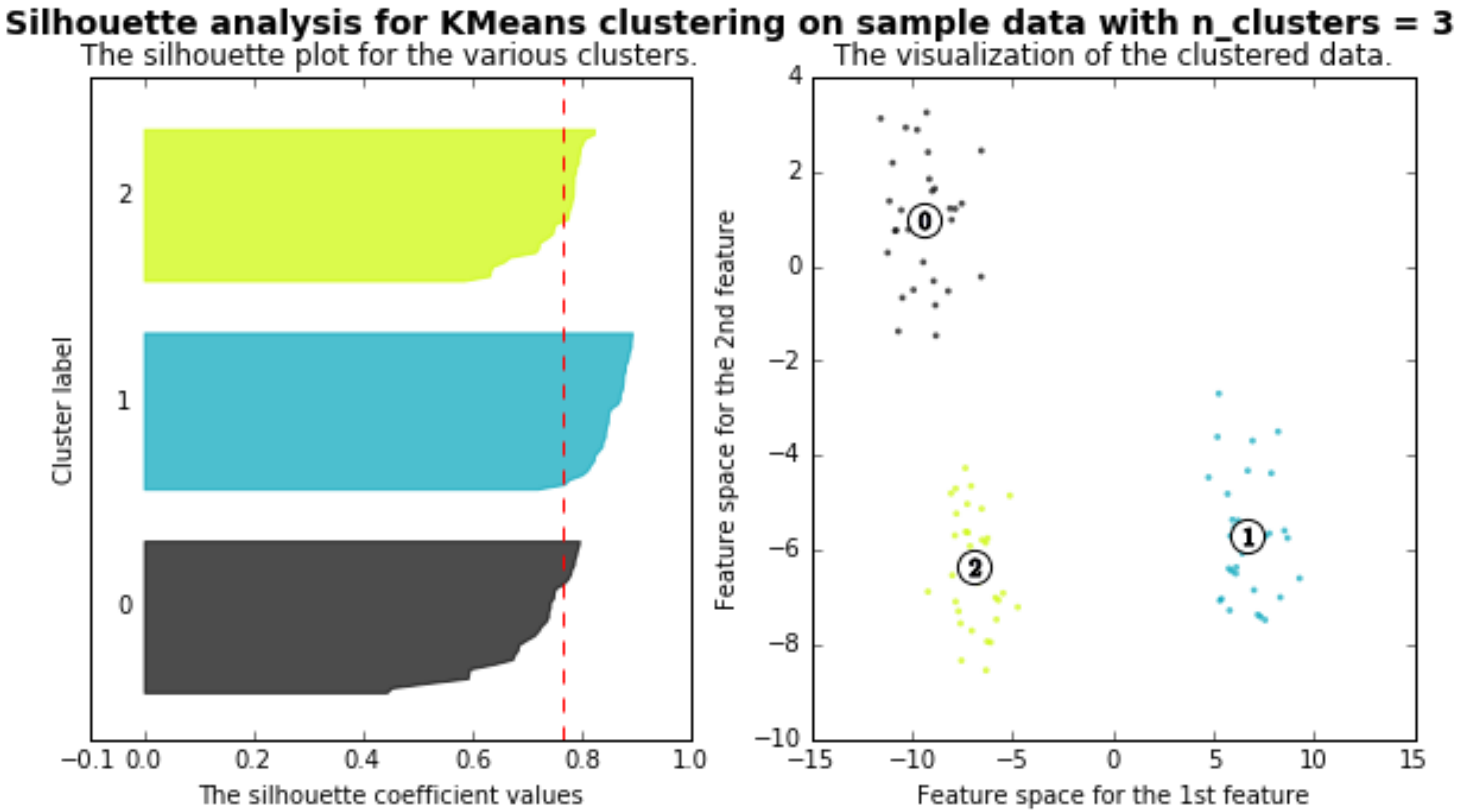
$$-1 \leq S \leq 1$$



Selecting number of clusters



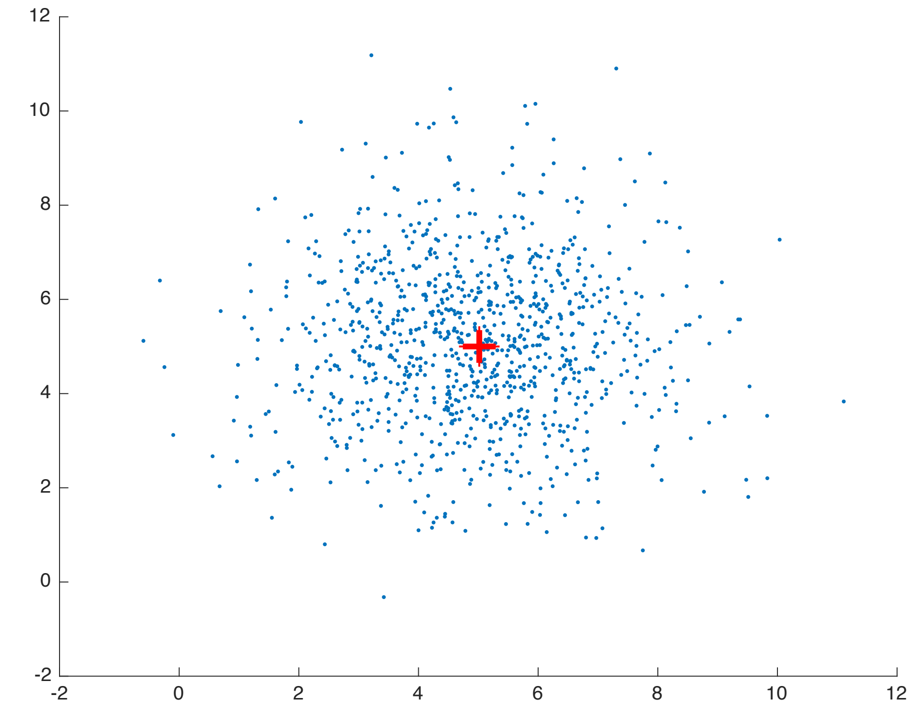
Silhouette analysis



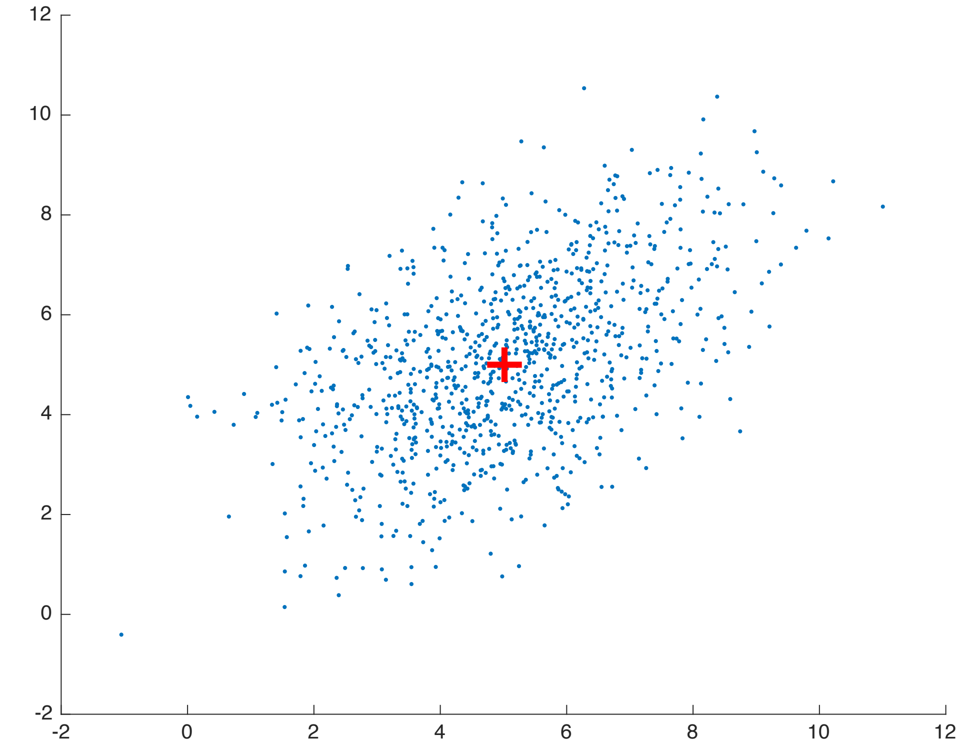
Probabilistic approach

Assume data is produced by random variables:

$$x_i^j \sim \mathcal{N}(\mu, \sigma^2)$$



$$x_i \sim \mathcal{N}(\mu, \Sigma)$$



Probabilistic approach

Assume data is produced by several random variables:

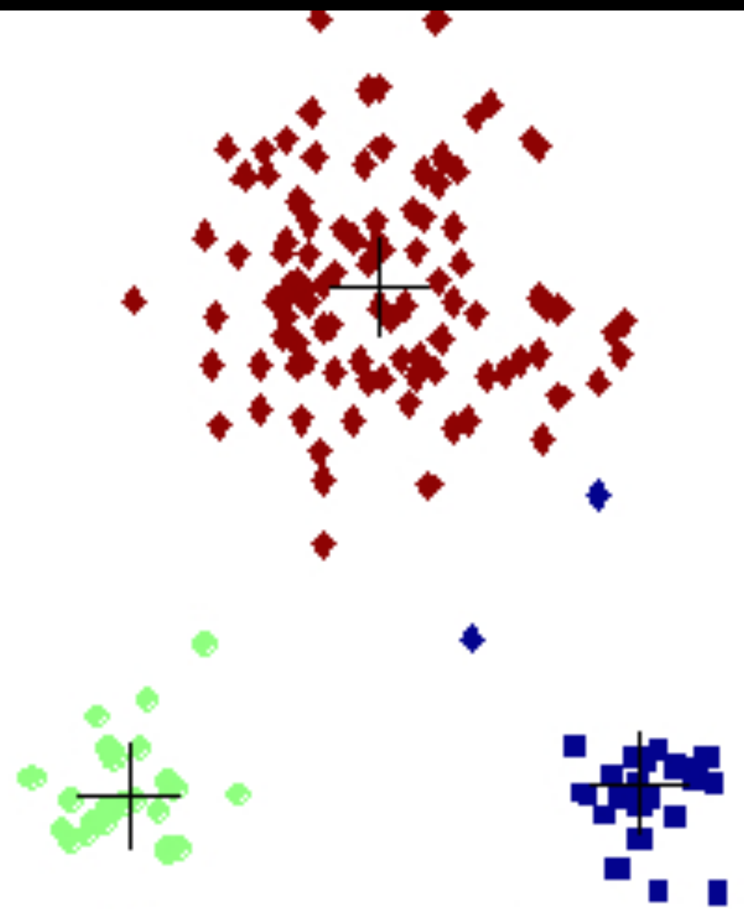
$$x_i^j \sim \mathcal{N}(\mu_{c_i}^j, \sigma^2)$$

Physical characteristics of adult male dogs based on their breed

$$p(x_i^j|c_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i^j-\mu_{c_i}^j)^2}{\sigma^2}}$$

$$\prod_{i,j} p(x_i^j|c_i) \to \max$$

k-means derivation



$$p(x_i^j|c_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i^j - \mu_{c_i}^j)^2}{\sigma^2}}$$

$$\prod_{i,j} p(x_i^j|c_i) \to \max$$

$$\sum_{i,j} ln(p(x_i^j|c_i)) \to \max$$

$$-Nn \cdot ln\sigma - \sum_{i,j} \frac{(x_i^j - \mu_{c_i}^j)^2}{\sigma^2} \to \max$$

$$SD = \sum_{i,j} (x_i^j - \mu_{c_i}^j)^2 \to \min$$

Mixture model

What if we admit uncertainty of the clustering, i.e. multiple distribution contribute to a single data point with certain weights

mixed/uncertain breeds

$$c_i = \mathrm{argmax}_k \pi_k(i)$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \sum_{k=1}^{K} \pi_k = 1.$$

Maximum Likelihood: $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$