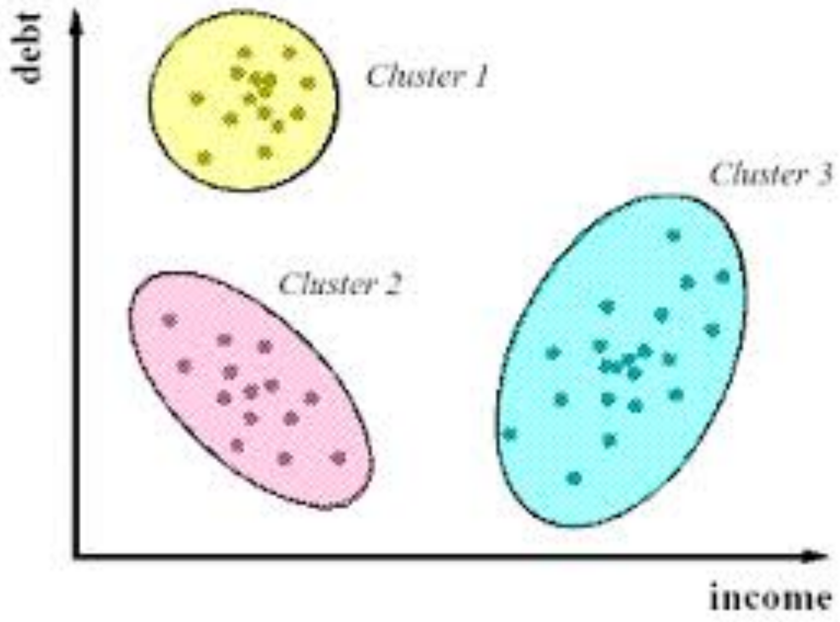
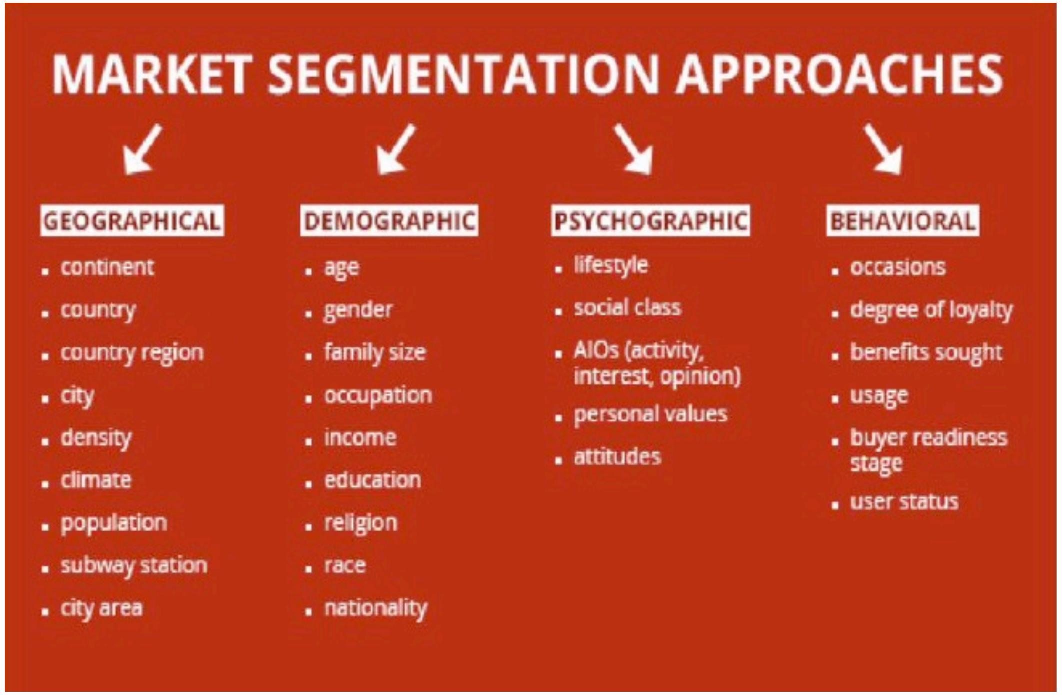
Clustering

- Given $N \times d$ data matrix: $\boldsymbol{X}$
  - Each row is one sample, x_n
- Problem: Group data into K clusters
- Mathematically:
  - Assign each sample to a cluster
  - Assign $\sigma_n \in \{1, \ldots, K\}$: Cluster label for each sample
- Want samples in same cluster to be "close"
  - $\|x_n - x_m\|$ is small when $\sigma_n = \sigma_m$



Clustering

- Clustering has many applications
  - Any time you want to segment data
  - Uncovering latent discrete variables
- Examples:
  - Segmenting sections of an image
  - Segmenting customers in market data

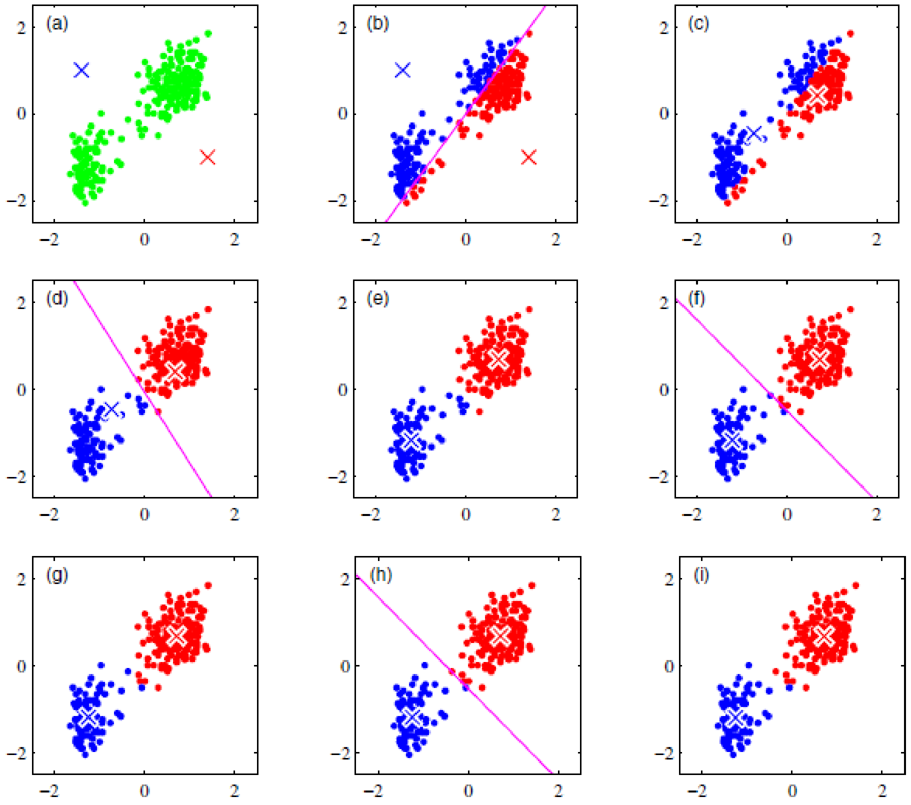


From: Market segmentation possibilities in the tourism market context of South Africa

K-means

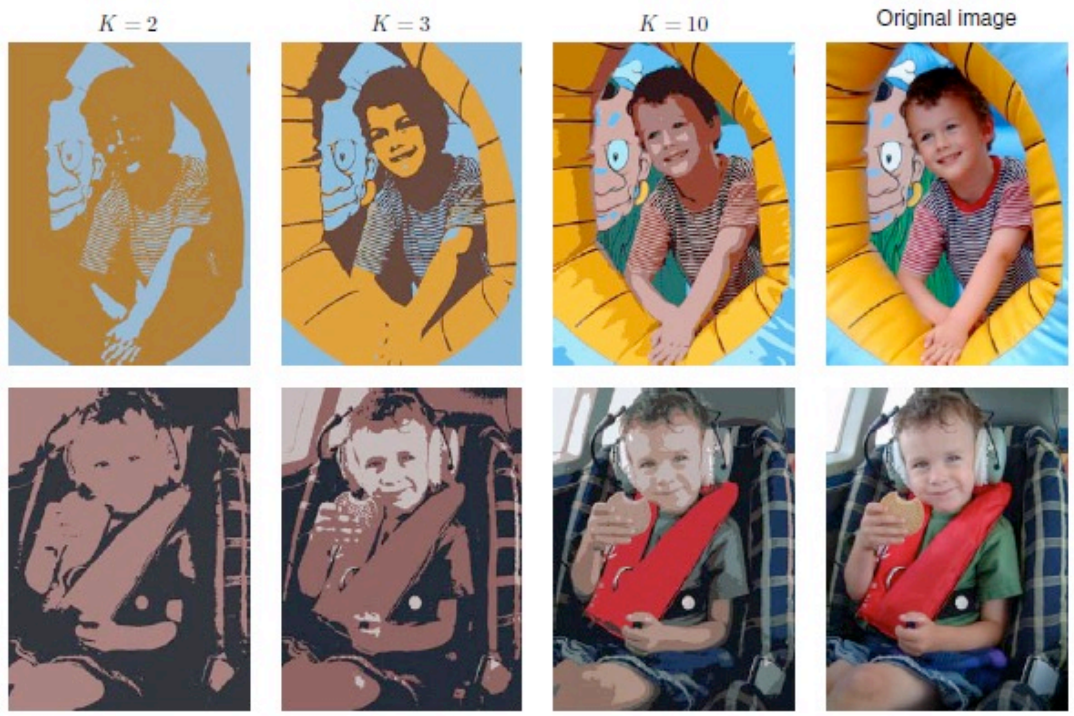
- A simple iterative algorithm to determine:
  - μ_i = mean of each cluster (hence, the name K-means)
  - $\sigma_n \in \{1, \ldots K\}$ = cluster that data point x_n belongs to
  - Minimize: $J = \sum_{i=1}^{K} \sum_{n \in C_i} ||x_n - \mu_i||^2$ (MSE of all samples in C_i from its center)
- Step 0: Start with guess at σ_n or μ_i
- Step 1: Update mean of each cluster: $\mu_i =$ average of x_n in C_i (centroid rule)
- Step 2: Update cluster membership: $\sigma_n = \arg\min_i ||x_n - \mu_i||^2$ (nearest neighbor rule)
- Return to step 1

K-Means illustrated



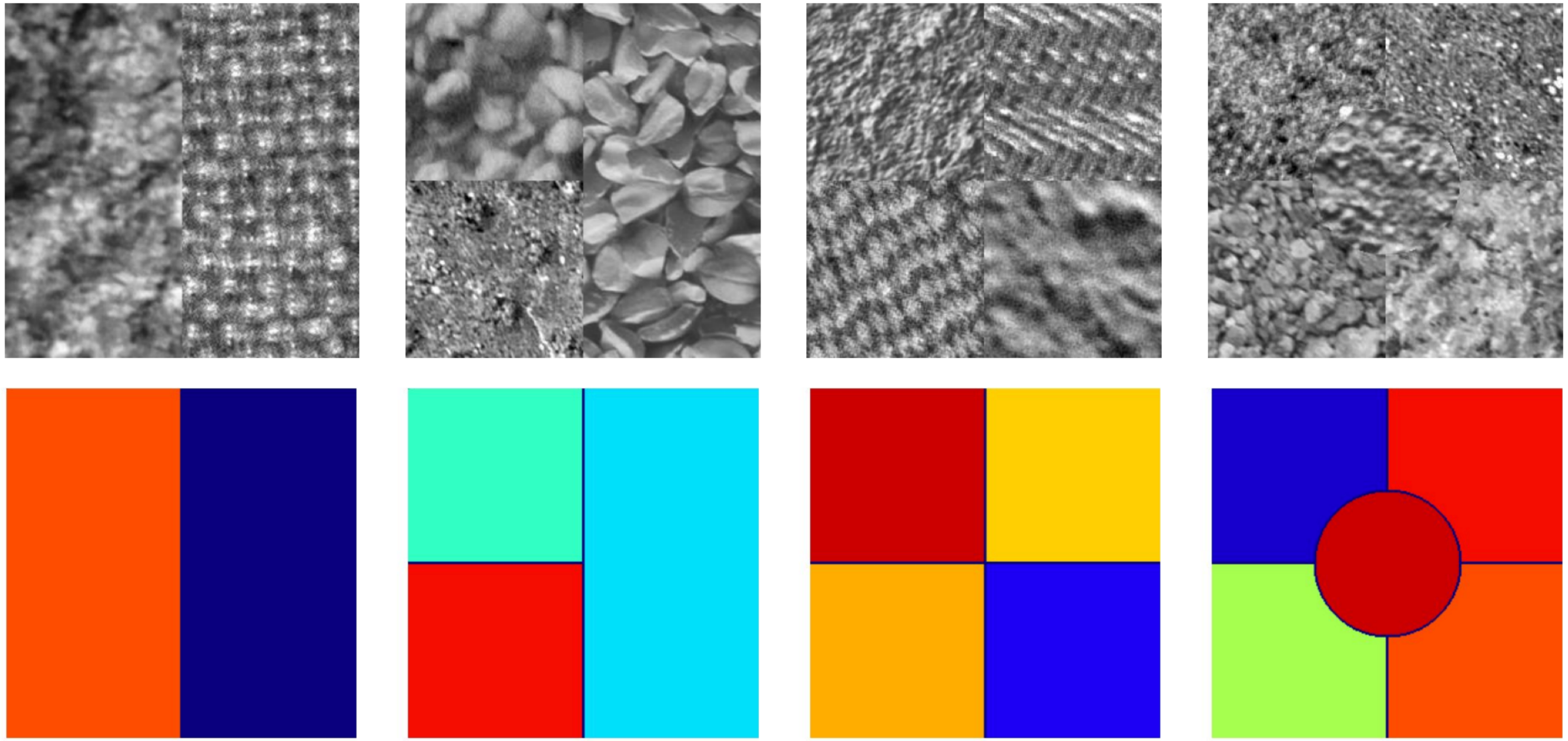
- From Bishop, Chapter 9.
- K-Means on "old faithful" data set

Image Segmentation Based on Color



- ❑Also from Bishop.
- ❑Use K-means on the RGB values (dimension = 3)

Image segmentation based on texture



Texture at each pixel is usually described by some statistics of the neighborhood surrounding the pixel.

Convergence

- Will always converge to a "local" minima of cost function

$$J = \sum_{i=1}^{K} \sum_{n=1}^{N} r_{ni} \|x_n - \mu_i\|^2$$

  - Subject to $r_{ni} = 0$ or 1 and $\sum_i r_{ni} = 1$

- K-means alternately decreases J
  - Proof on board

- But, can get stuck in a local minima
  - May need good selection of initial condition

Distance measures

- Distance measures
  - How to measure similarity between samples?
  - Above algorithms used squared distance $||x_n - x_m||$

- Many possibilities
  - What features to use?
  - Should you normalize entries?
  - What distance metric should you use?

Initialization

- Initialization:
  - Final limit of K-means depends on initial condition
  - May obtain poor clustering with bad initial condition
- Possible solutions:
  - K-means++: http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf
  - Provides good initial condition based on data
  - Multiple initial starts