Clustering
Lecture 14

David Sontag

New York University

Slides adapted from Luke Zettlemoyer, Vibhav Gogate,
Carlos Guestrin, Andrew Moore, Dan Klein
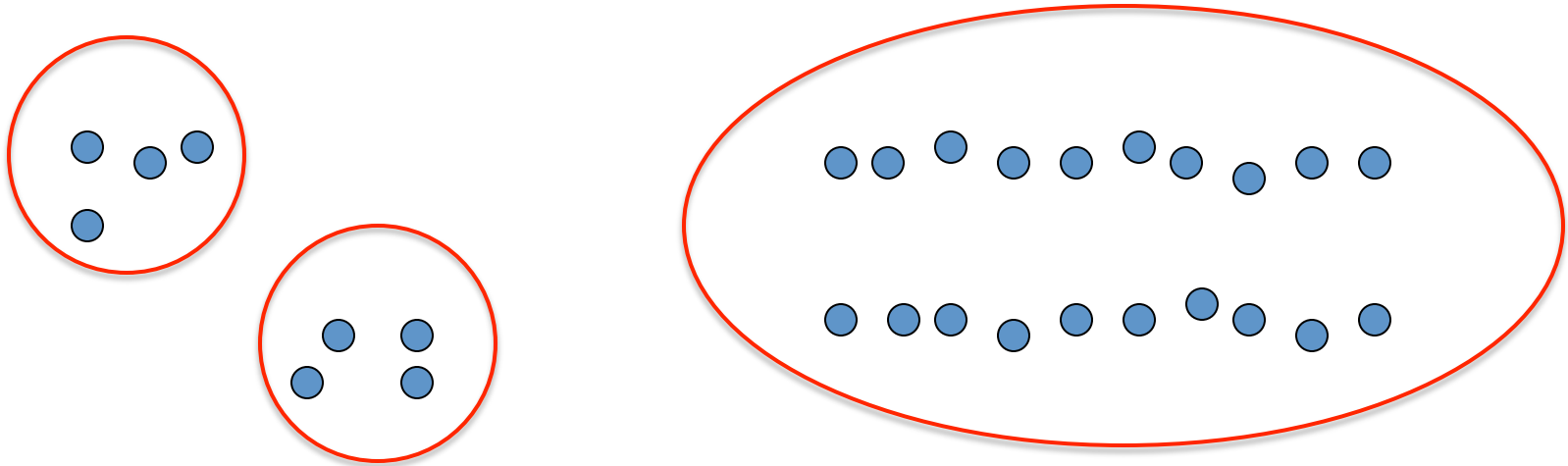
Clustering

Clustering:

- Unsupervised learning

- Requires data, but no labels

- Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images

- Useful when don't know what you're looking for
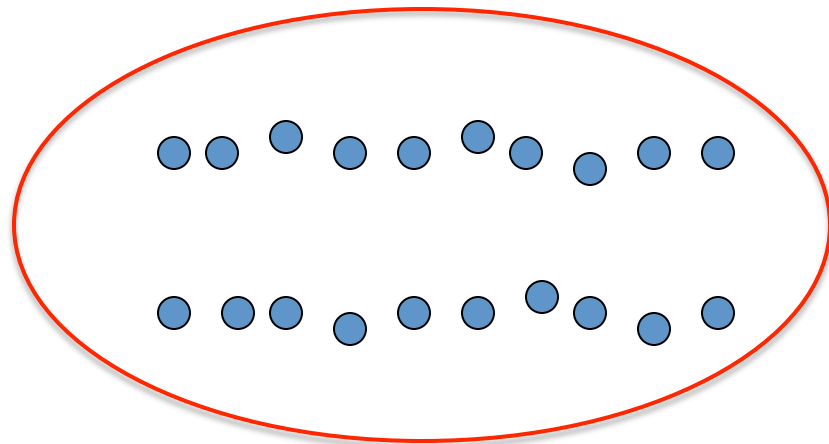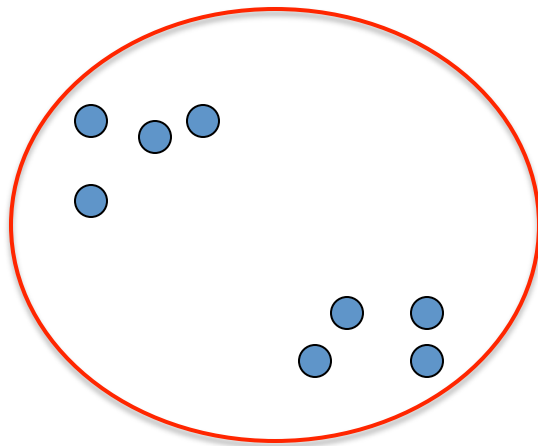
- But: can get gibberish

Clustering

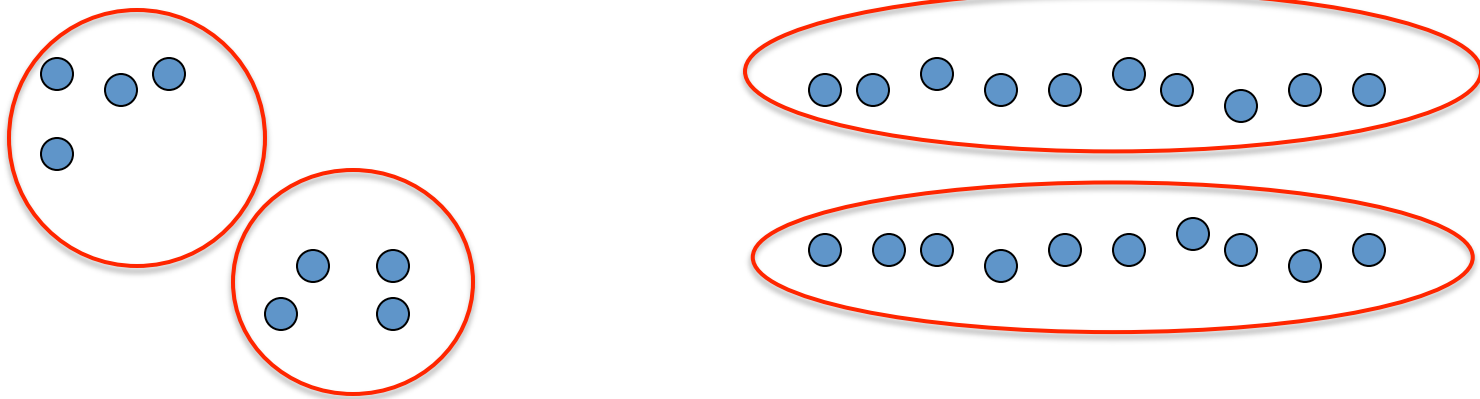- Basic idea: group together similar instances
- Example: 2D point patterns

Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns

Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns

- What could "similar" mean?
  - One option: small Euclidean distance (squared)

$$\mathrm{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

  - Clustering results are crucially dependent on the measure of similarity (or distance) between "points" to be clustered
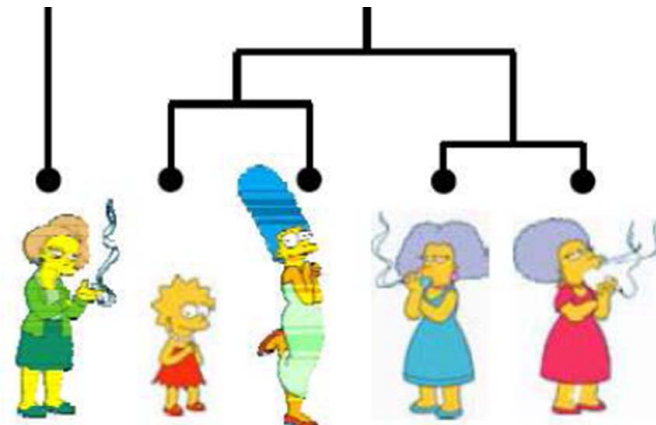
Clustering algorithms

- Partition algorithms (Flat)
  - K-means
  - Mixture of Gaussian
  - Spectral Clustering
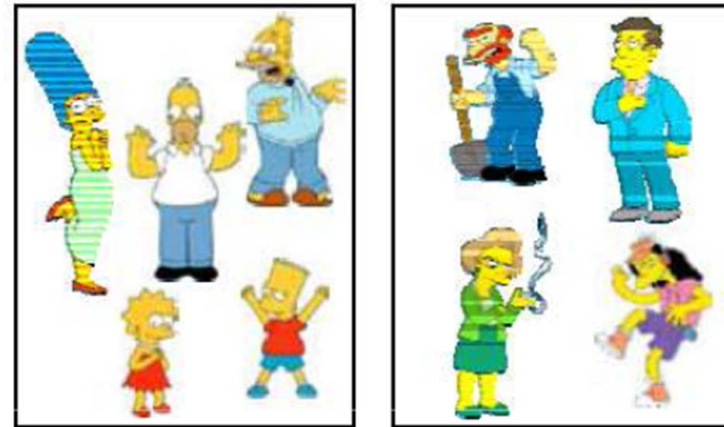
- Hierarchical algorithms
  - Bottom up – agglomerative
  - Top down – divisive
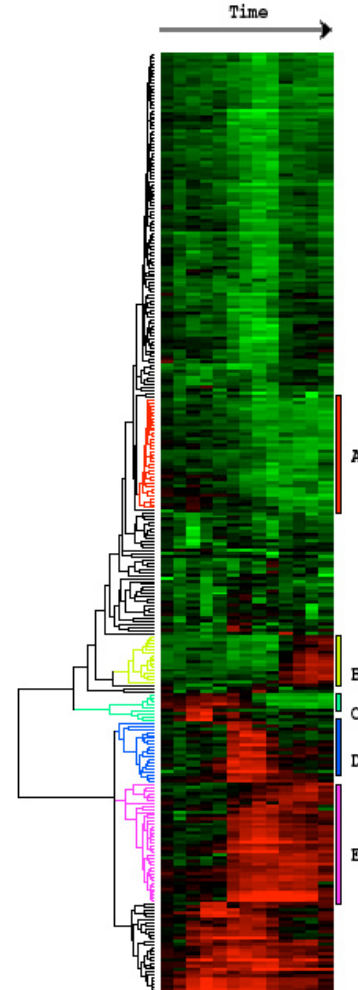
Clustering examples

Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions

Clustering examples

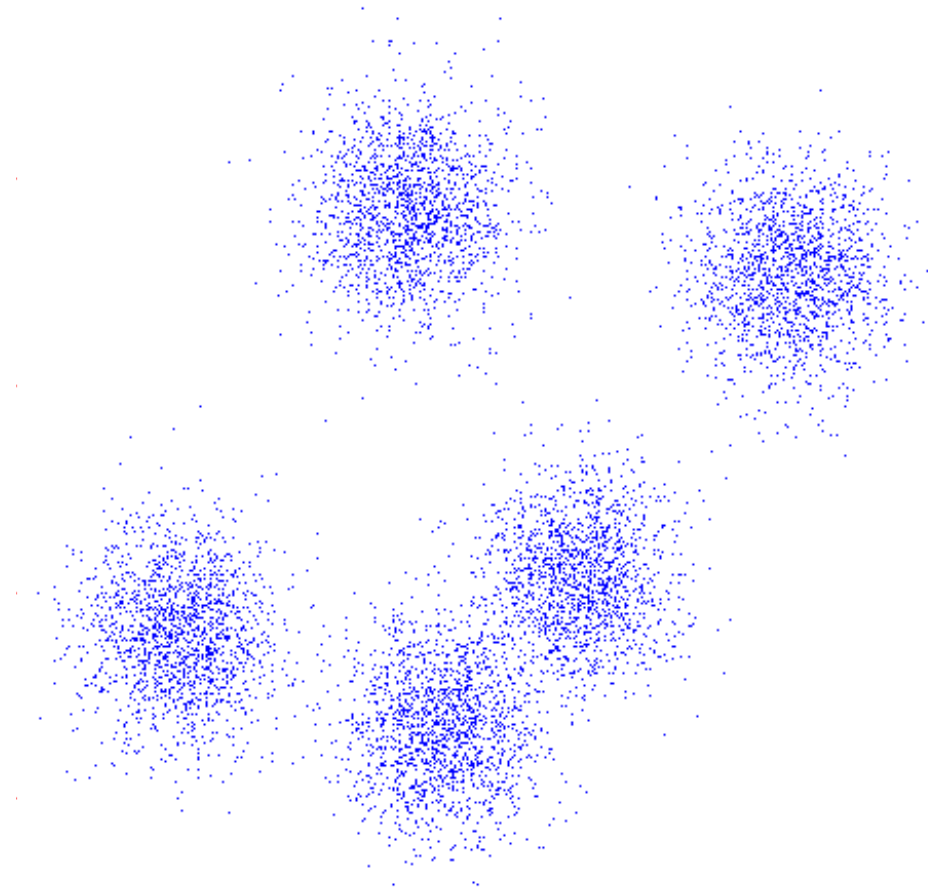**Clustering gene expression data**

Time

A
B
C
D
E

Eisen et al, PNAS 1998

K-Means

- An iterative clustering algorithm

  - Initialize: Pick K random points as cluster centers

  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points

  - Stop when no points' assignments change

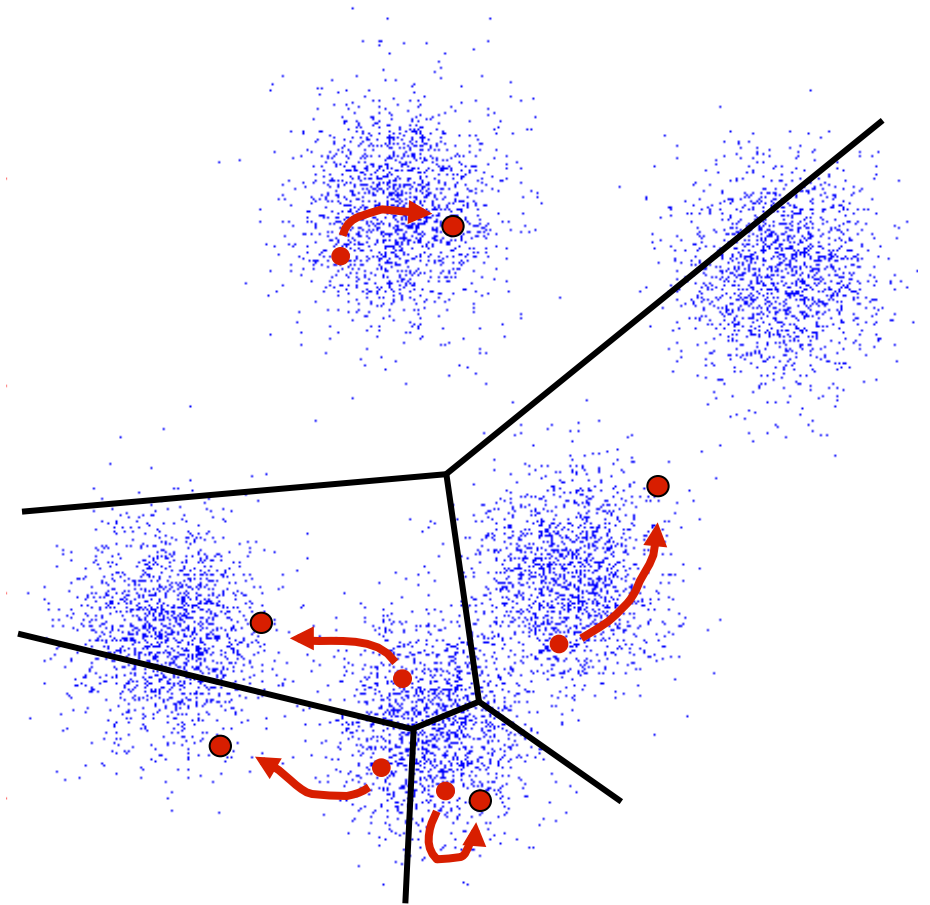# K-Means

- An iterative clustering algorithm

    - Initialize: Pick K random points as cluster centers

    - Alternate:
        1. Assign data points to closest cluster center
        2. Change the cluster center to the average of its assigned points
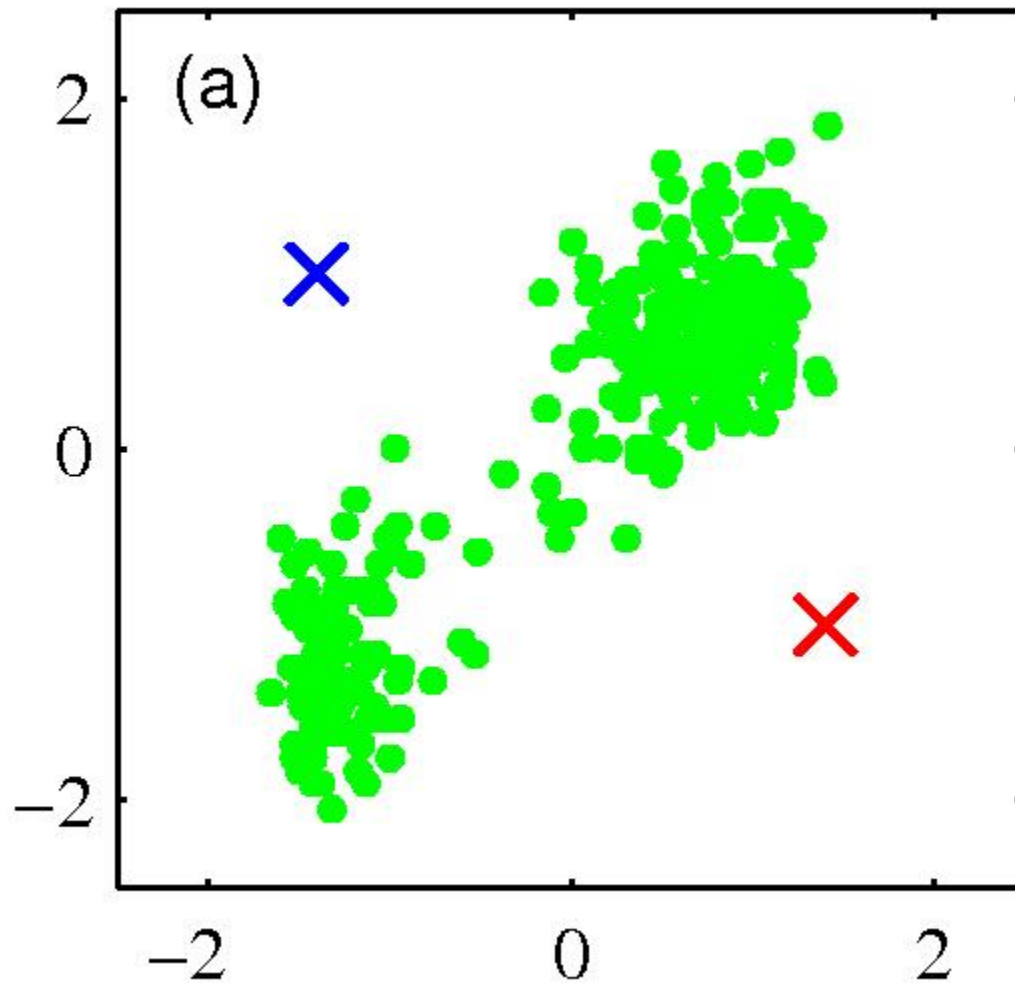
    - Stop when no points' assignments change
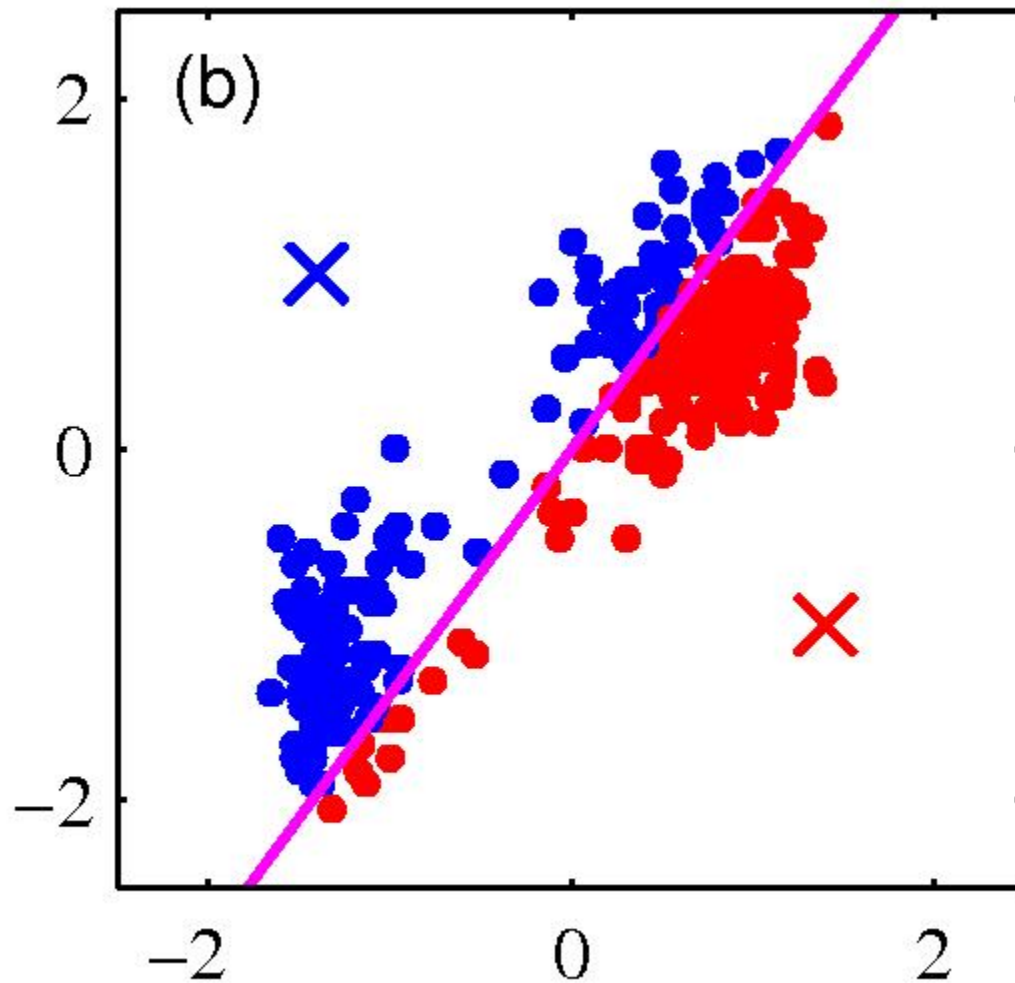
K-means clustering: Example

- Pick K random points as cluster centers (means)
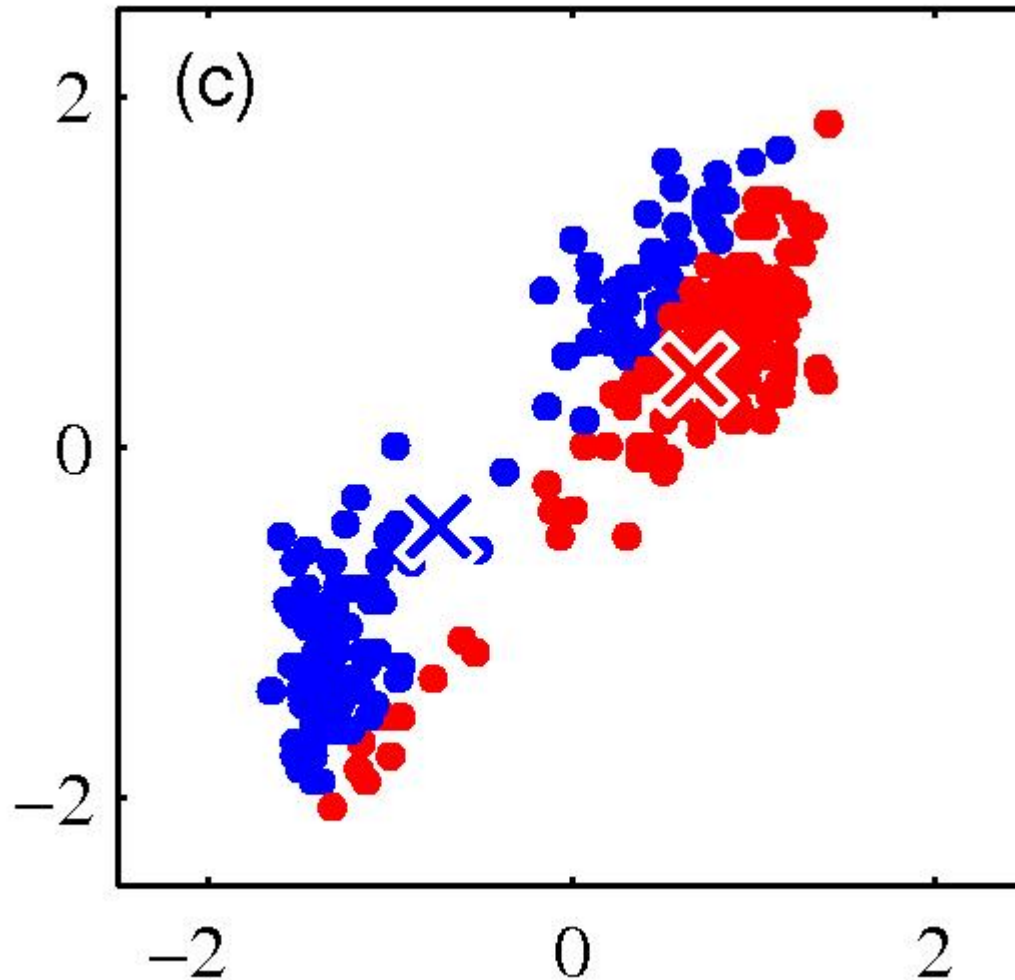
Shown here for $K=2$

K-means clustering: Example

Iterative Step 1

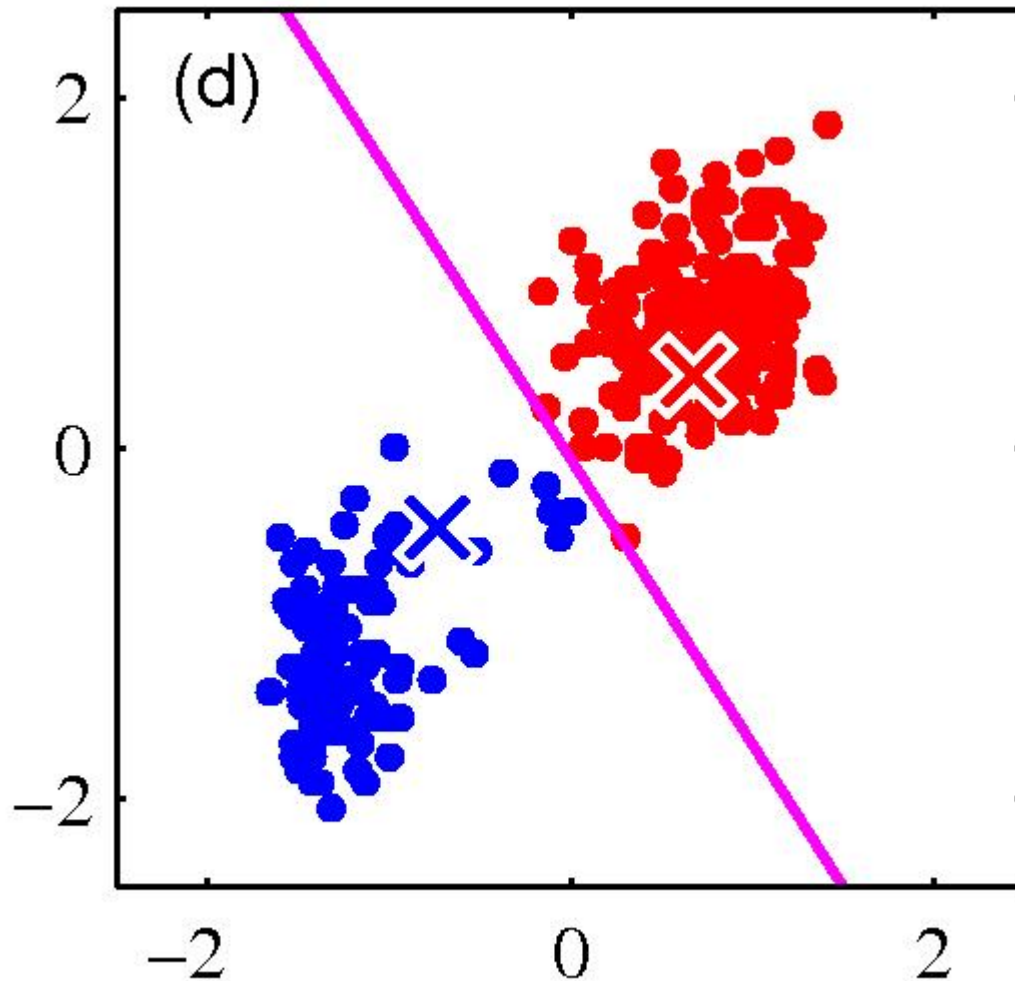- Assign data points to closest cluster center

K-means clustering: Example
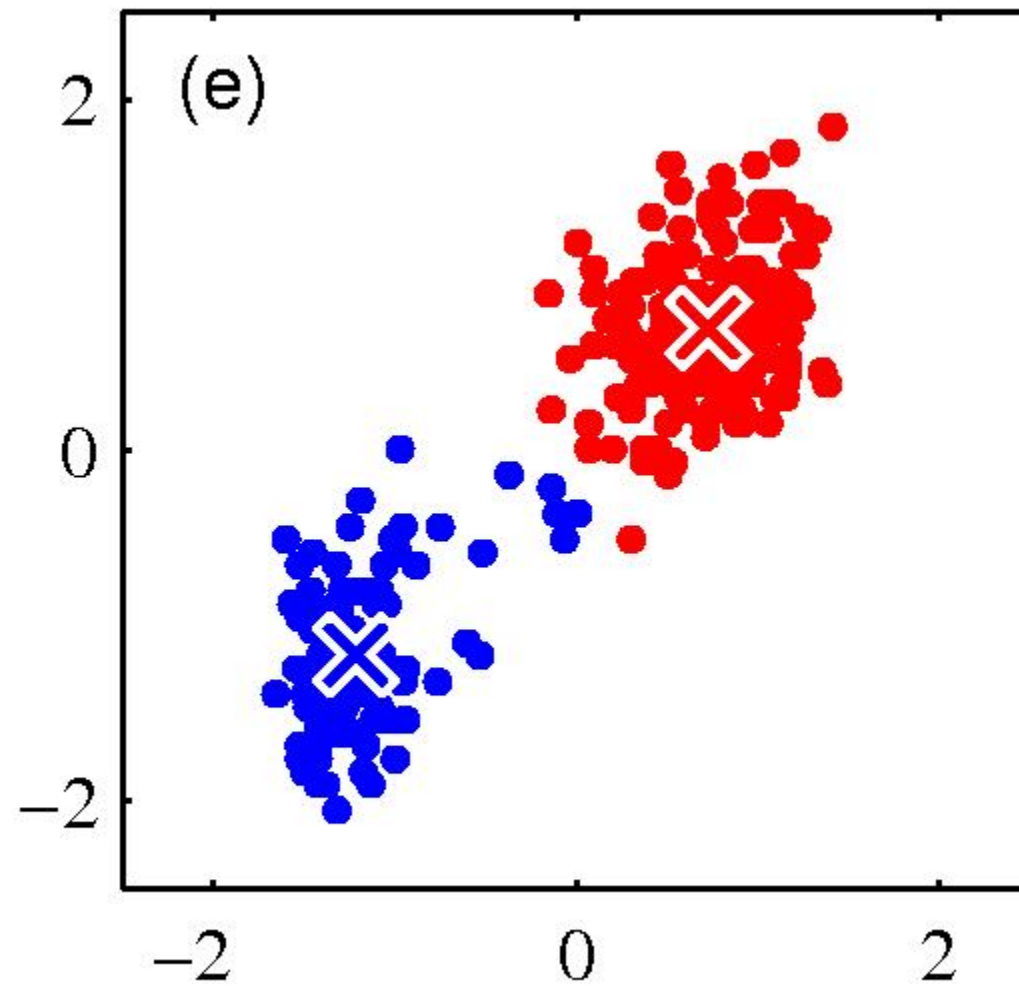

Iterative Step 2

- Change the cluster center to the average of the assigned points

K-means clustering: Example

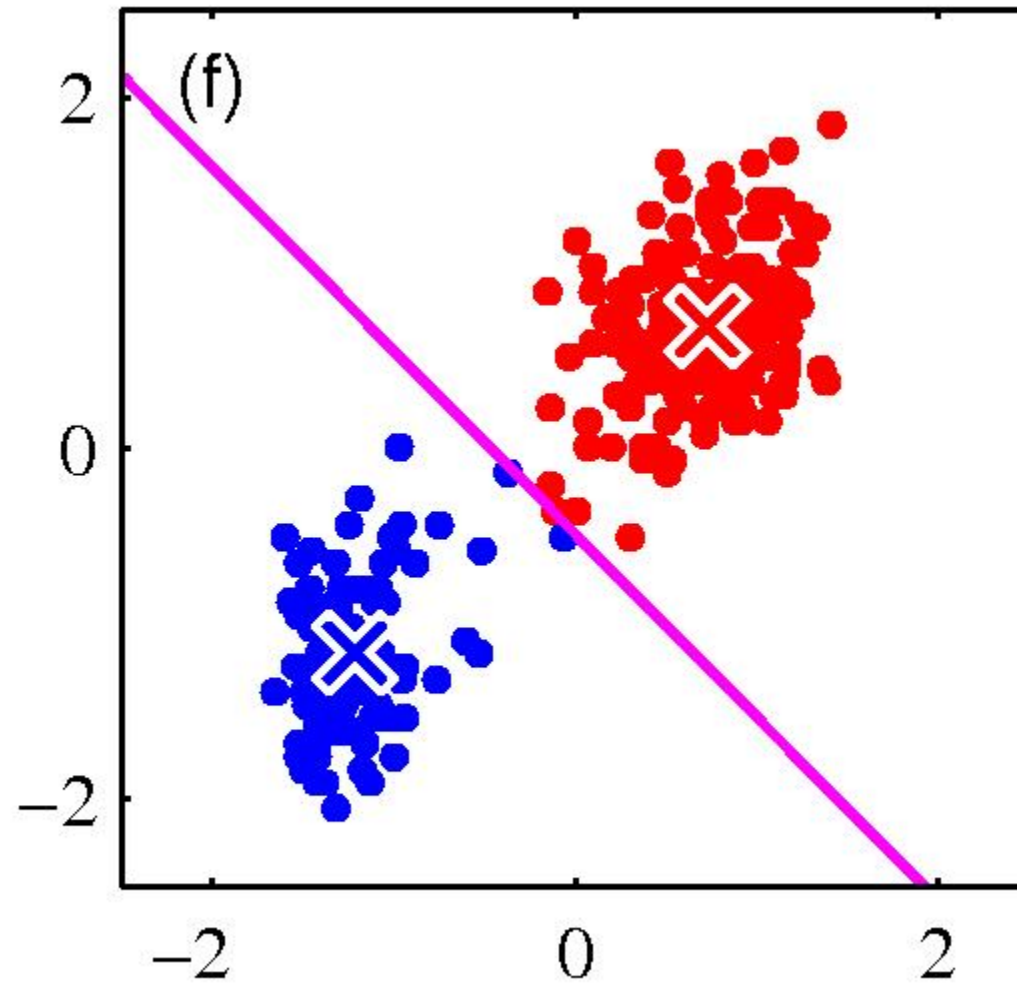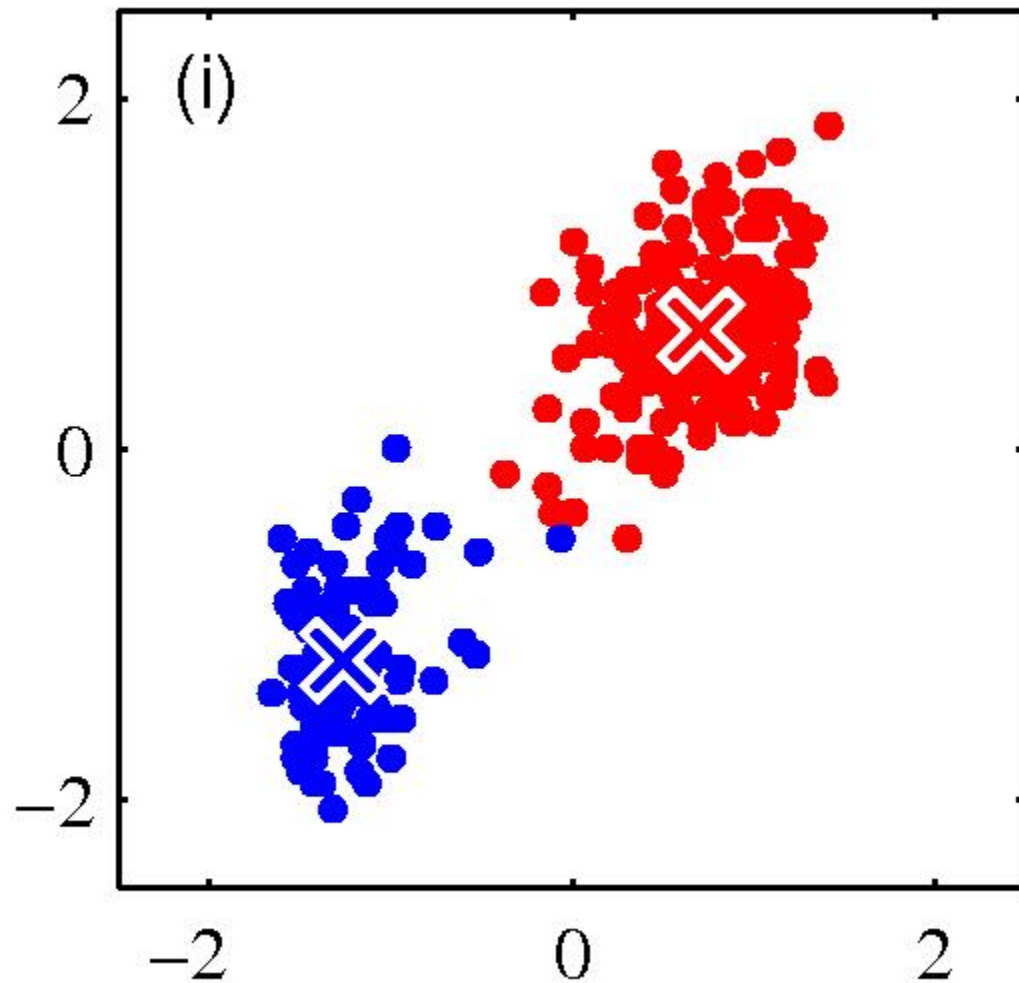- Repeat until convergence

K-means clustering: Example

K-means clustering: Example

K-means clustering: Example

Properties of K-means **algorithm**

- Guaranteed to converge in a finite number of iterations

- Running time per iteration:
    1. Assign data points to closest cluster center

       O(KN) time

    2. Change the cluster center to the average of its assigned points

       O(N)

What properties should a distance measure have?

- Symmetric
  - D(A,B)=D(B,A)
  - Otherwise, we can say A looks like B but B does not look like A
- Positivity, and self-similarity
  - D(A,B)$\geq$0, and D(A,B)=0 iff A=B
  - Otherwise there will different objects that we cannot tell apart
- Triangle inequality
  - D(A,B)+D(B,C) $\geq$ D(A,C)
  - Otherwise one can say "A is like B, B is like C, but A is not like C at all"

[Slide from Alan Fern]

Kmeans Convergence

Objective

$$\min_{\mu}\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

1. Fix μ, optimize C:

$$\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 = \min_{c} \sum_{i}^{n} \left|x_i - \mu_{x_i}\right|^2$$

2. Fix C, optimize μ:

$$\min_{\mu} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

– Take partial derivative of μ_i and set to zero, we have

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Step 2 of kmeans

Kmeans takes an alternating optimization approach, each step is guaranteed to decrease the objective – thus guaranteed to converge

[Slide from Alan Fern]

Example: K-Means for Segmentation

K=2

Original

Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Example: K-Means for Segmentation

K=2   K=3   Original
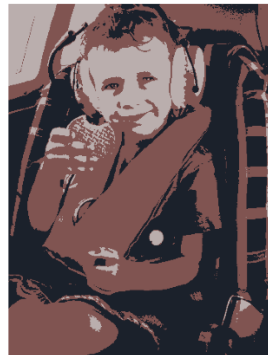
Example: K-Means for Segmentation

K=2 K=3 K=10 Original

4% 8% 17%

Example: Vector quantization

FIGURE 14.9. *Sir Ronald A. Fisher (1890 − 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel*
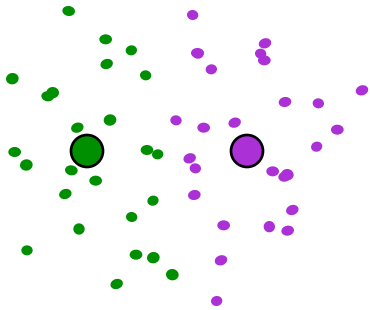
[Figure from Hastie *et al.* book]

Initialization

- K-means **algorithm** is a heuristic
  - Requires initial means
  - It does matter what you pick!

  - What can go wrong?

  - Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics
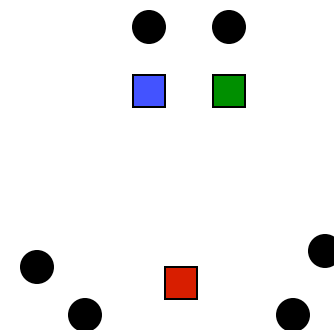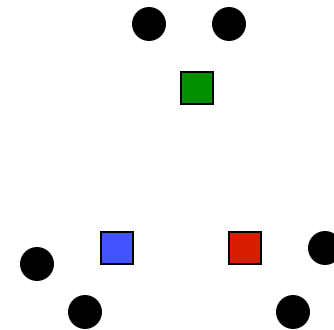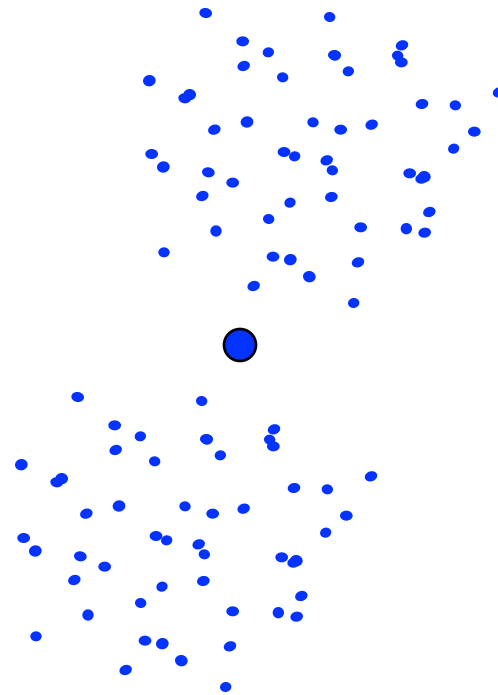
K-Means Getting Stuck

A local optimum:

Would be better to have
one cluster here

… and two clusters here

Extra

- K-means Applets:
    - http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

    - http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster/