

# Clustering

## Lecture 14

David Sontag  
New York University

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

# Clustering

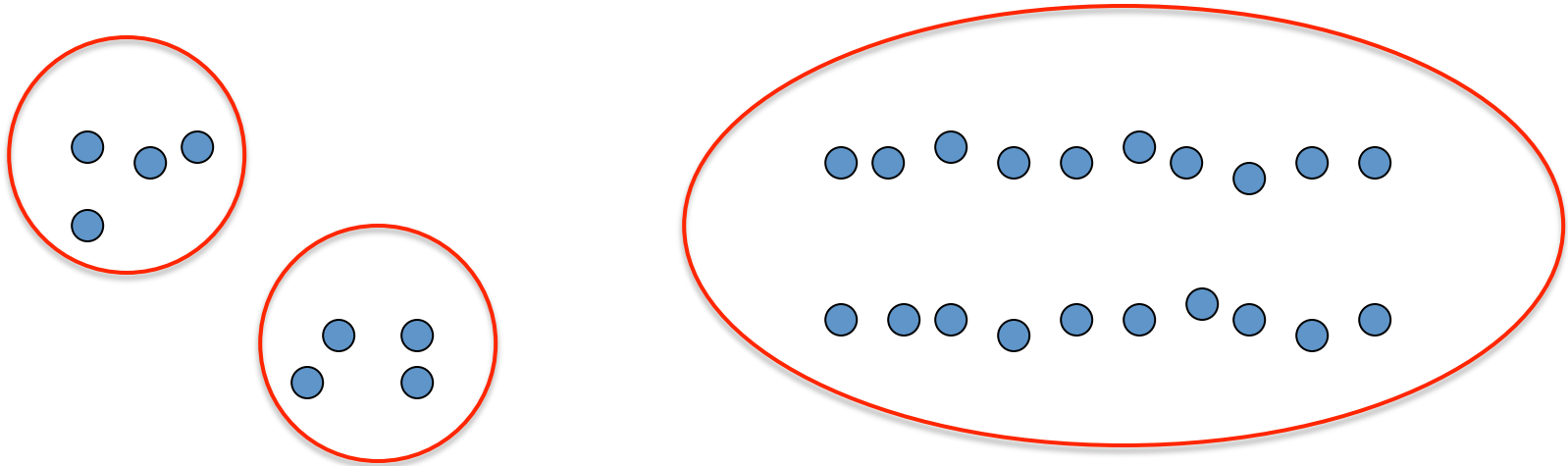
## Clustering:

- 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



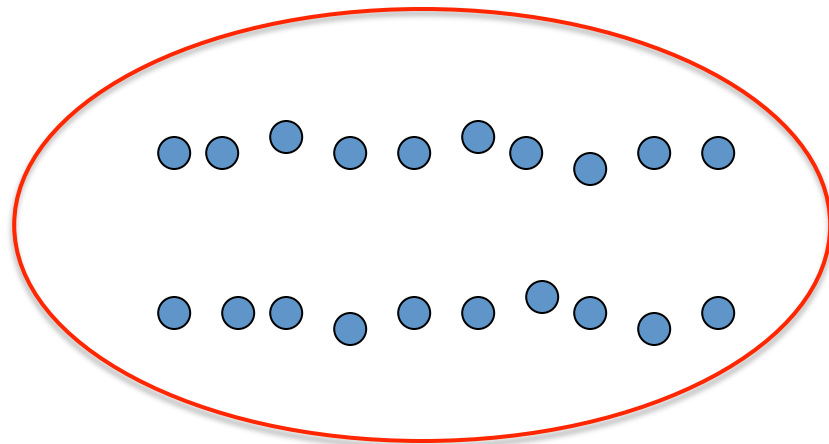
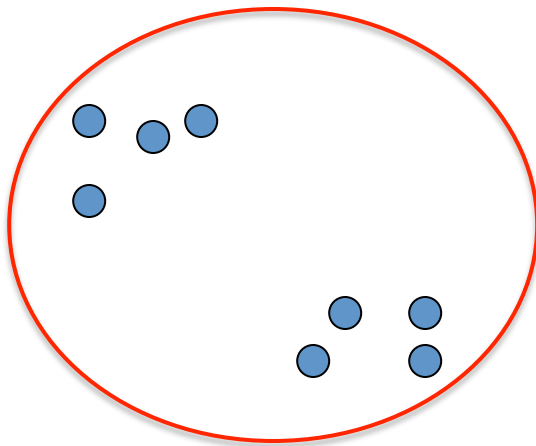
# Clustering

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



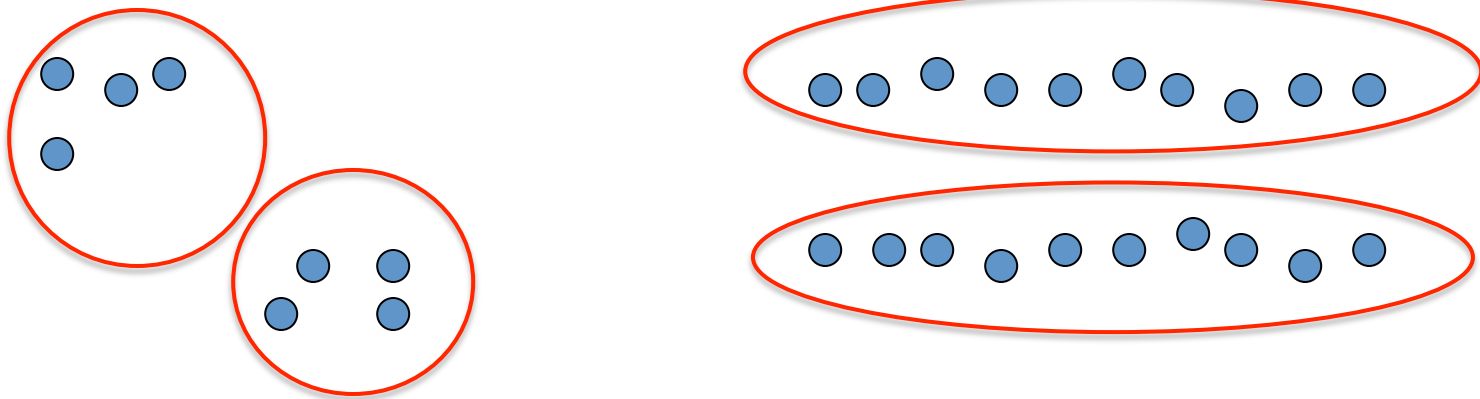
# Clustering

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



# Clustering

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

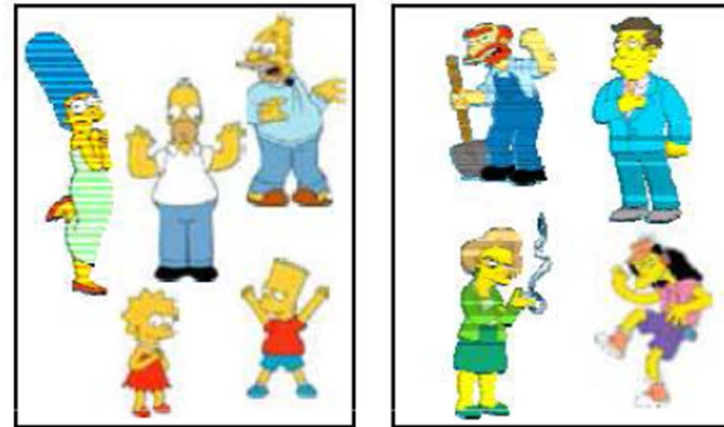


- **What could “similar” mean?**
  - One option: small Euclidean distance (squared)
$$\text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$
  - Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered

# Clustering algorithms

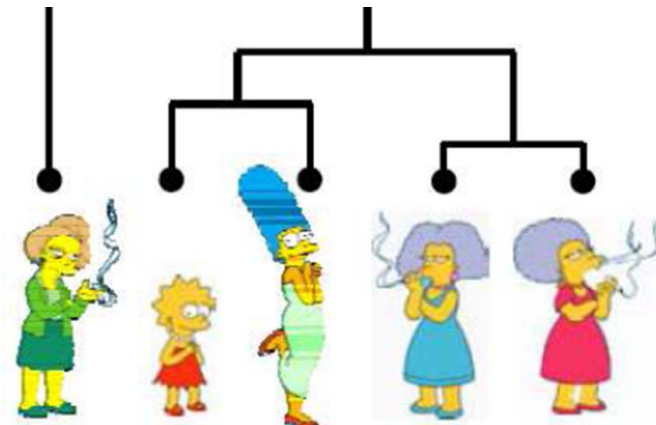
- Partition algorithms (Flat)

- K-means
- Mixture of Gaussian
- Spectral Clustering



- Hierarchical algorithms

- Bottom up – agglomerative
- Top down – divisive



# Clustering examples

## Image segmentation

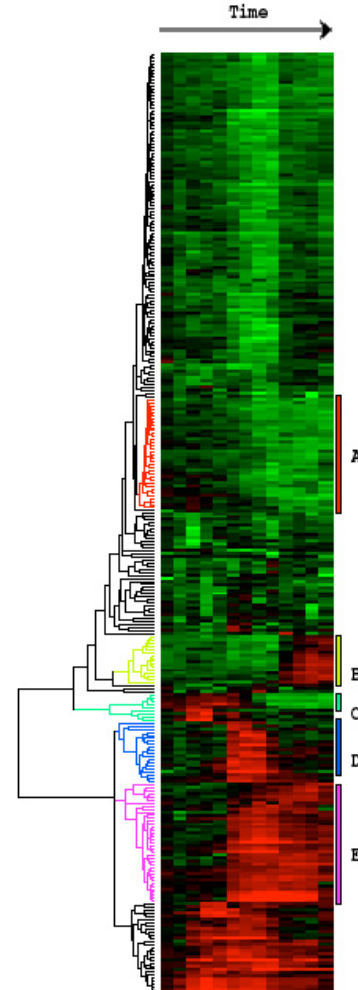
Goal: Break up the image into meaningful or perceptually similar regions



[Slide from James Hayes]

# Clustering examples

## Clustering gene expression data



Eisen et al, PNAS 1998



# Clustering examples

## Cluster news articles

Google

News

U.S. edition Classic

**Top Stories**

Boston Red Sox  
Apple Inc.  
Angela Merkel  
Nokia Lumia  
Bashar al-Assad  
Republican Party  
Facebook  
Pets  
Katy Perry  
Bushfires in Australia  
New York, New York

Recommended

U.S.  
World  
Sci/Tech  
Business  
More Top Stories  
Health  
Spotlight  
Elections  
Entertainment  
Sports  
Technology  
Science

**Top Stories**

**Teen suspect saw movie moments after allegedly killing beloved Massachusetts ...**  
Fox News - 8 minutes ago  
The 14-year-old student who authorities say murdered a beloved math teacher at a Massachusetts high school admitted to police that he slashed her throat with a box cutter, a source told MyFoxBoston.  
Colleen Ritzer, slain Danvers High School teacher, remembered as passionate ... CBS News  
14-Year-Old Charged in Brutal Murder of Massachusetts Teacher New York Magazine  
Highly Cited: 14-year-old student held without bail in slaying of Danvers High teacher Boston.com  
Opinion: Hesham: Heartbroken friends say Colleen was born to teach Boston Herald  
In Depth: Student, 14, arraigned in murder of Mass. teacher USA TODAY  
Wikipedia: Danvers, Massachusetts  
[See realtime coverage »](#)

**Obamacare contractors tell their stories at congressional hearing**  
CNN - 40 minutes ago  
Washington (CNN) -- [Breaking news update at 10:09 a.m.]. [URGENT - Congress-Obamacare-Testing]. (CNN) -- A contractor on the problem-plagued government website for President Barack Obama's signature health care reforms said Thursday his ...  
Hearing on health care website today to focus on blame WXXIA-TV  
Contractors Point Fingers Over Health-Law Website AllThingsD  
[See realtime coverage »](#)

**EU leaders meet amid concern about US spying claims**  
CNN - 1 hour ago  
(CNN) -- European Union leaders are meeting Thursday in Brussels for a summit that may be overshadowed by anger about allegations that the United States has been spying on its European allies.  
Germany summons US ambassador over spying claims USA TODAY  
Germany Summons US Envoy Over Alleged NSA Spying ABC News  
Highly Cited: Readout of the President's Phone Call with Chancellor Merkel of Germany Whitehouse.gov (press release)  
From Germany: Press Review: Outrage over NSA eavesdropping Deutsche Welle  
Opinion: The Handyüberwachung Disaster New York Times  
In Depth: US ambassador to Germany summoned in Merkel mobile row BBC News  
[See realtime coverage »](#)

**US jobless claims miss forecasts, trade deficit widens slightly**  
Reuters - 59 minutes ago  
WASHINGTON | Thu Oct 24, 2013 9:19am EDT. WASHINGTON (Reuters) - The number of Americans filing new claims for unemployment benefits fell less than expected last week, but a lingering backlog of applications in California makes it difficult to get a ...  
Weekly Jobless Claims Fall to 350,000 Fox Business  
How States Fared on Unemployment Benefit Claims ABC News  
In Depth: More Americans Than Forecast Filed Jobless Claims Businessweek  
[See realtime coverage »](#)

**Kennedy cousin gets new trial in 1975 killing of neighbor; victim's mother ...**

ABC News

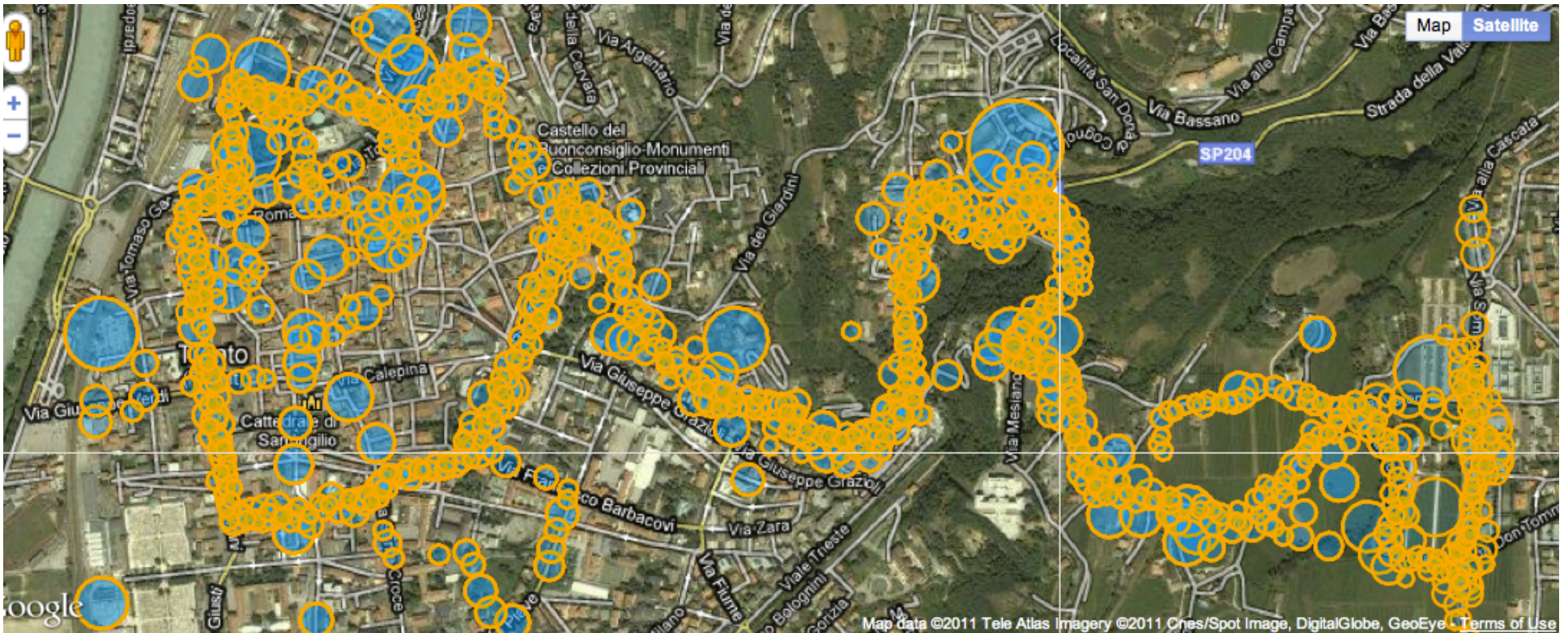
Wall Street Journal

National Post

The Olympian

# Clustering examples

Cluster people by space and time

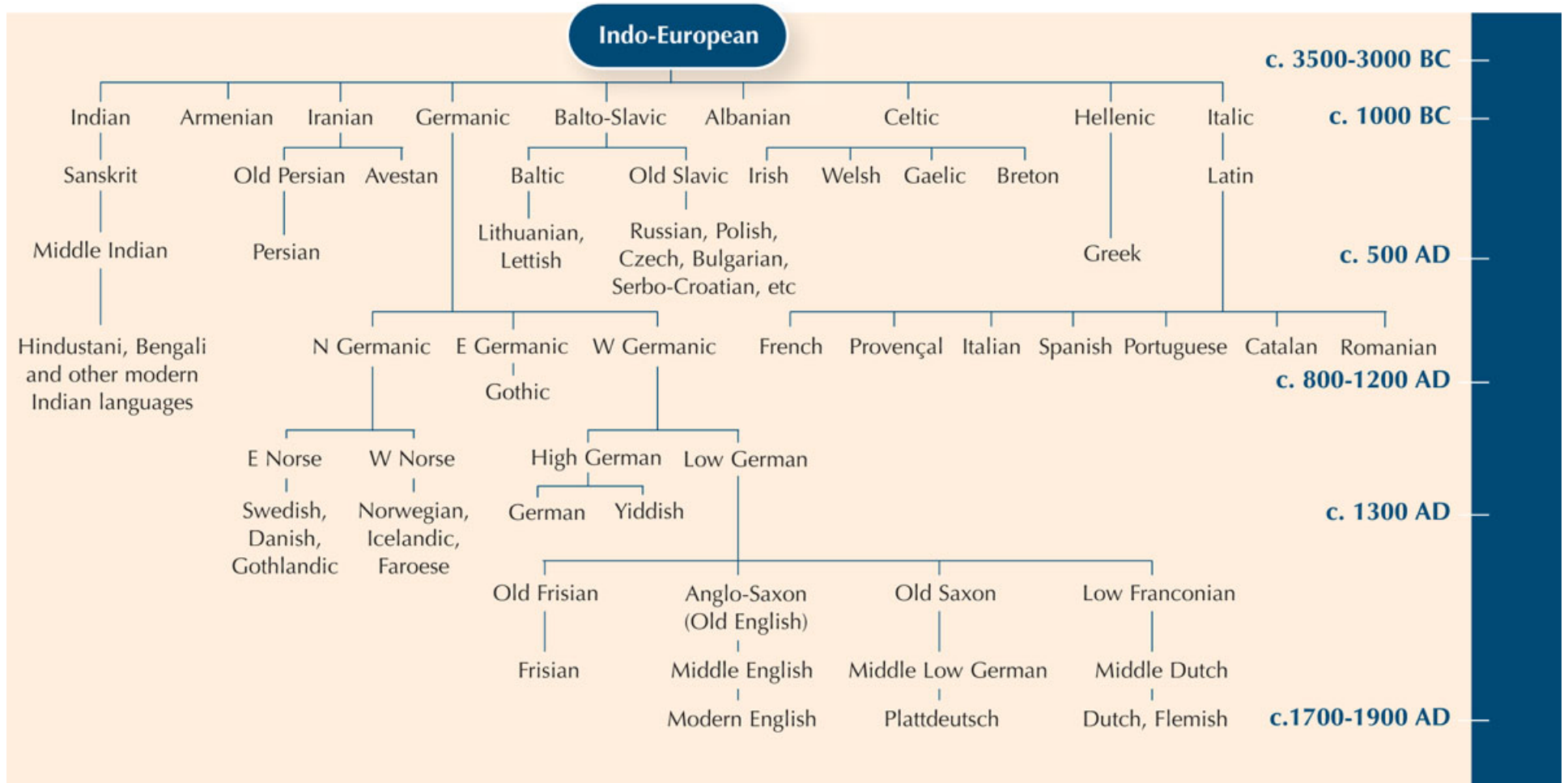


[Image from Pilho Kim]



# Clustering examples

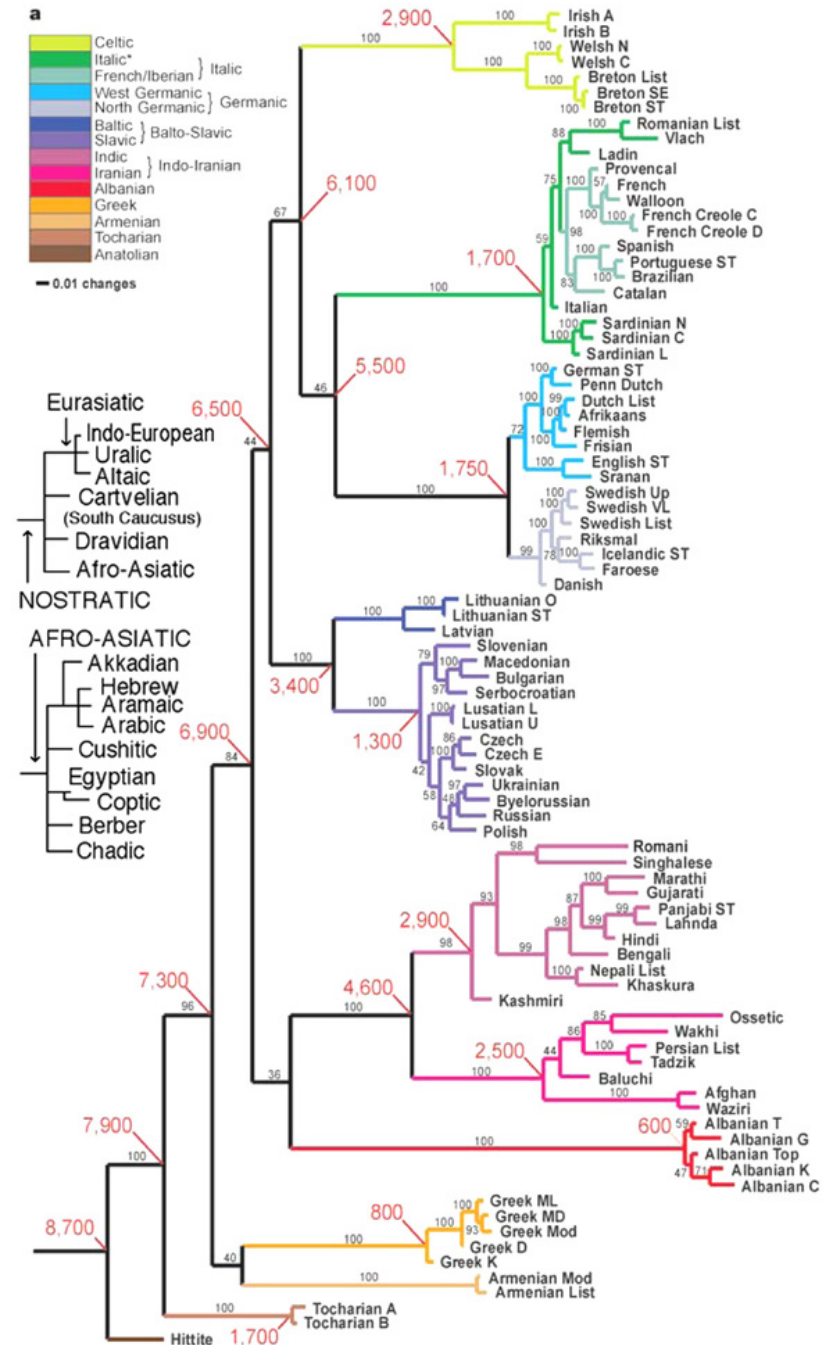
## Clustering languages



[Image from scienceinschool.org]

# Clustering examples

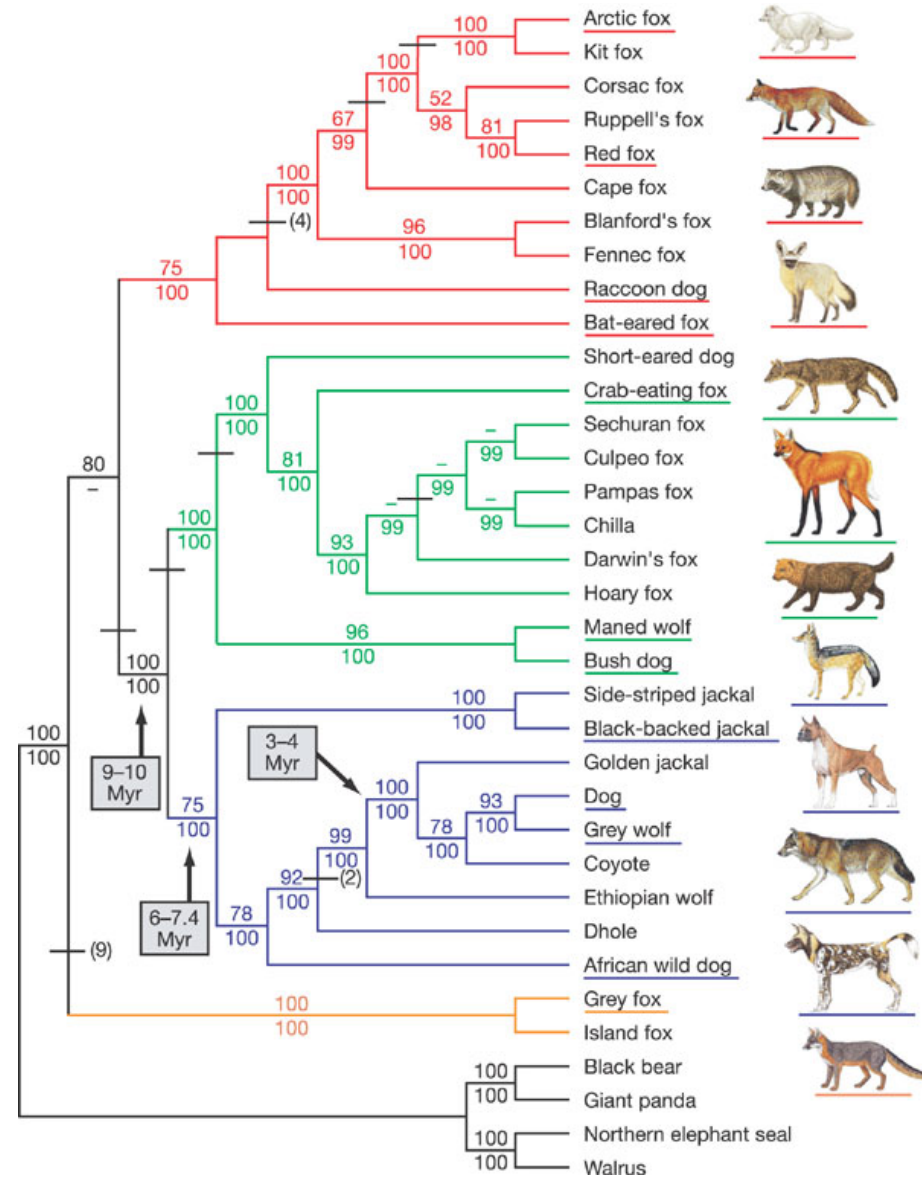
## Clustering languages



[Image from dhushara.com]

# Clustering examples

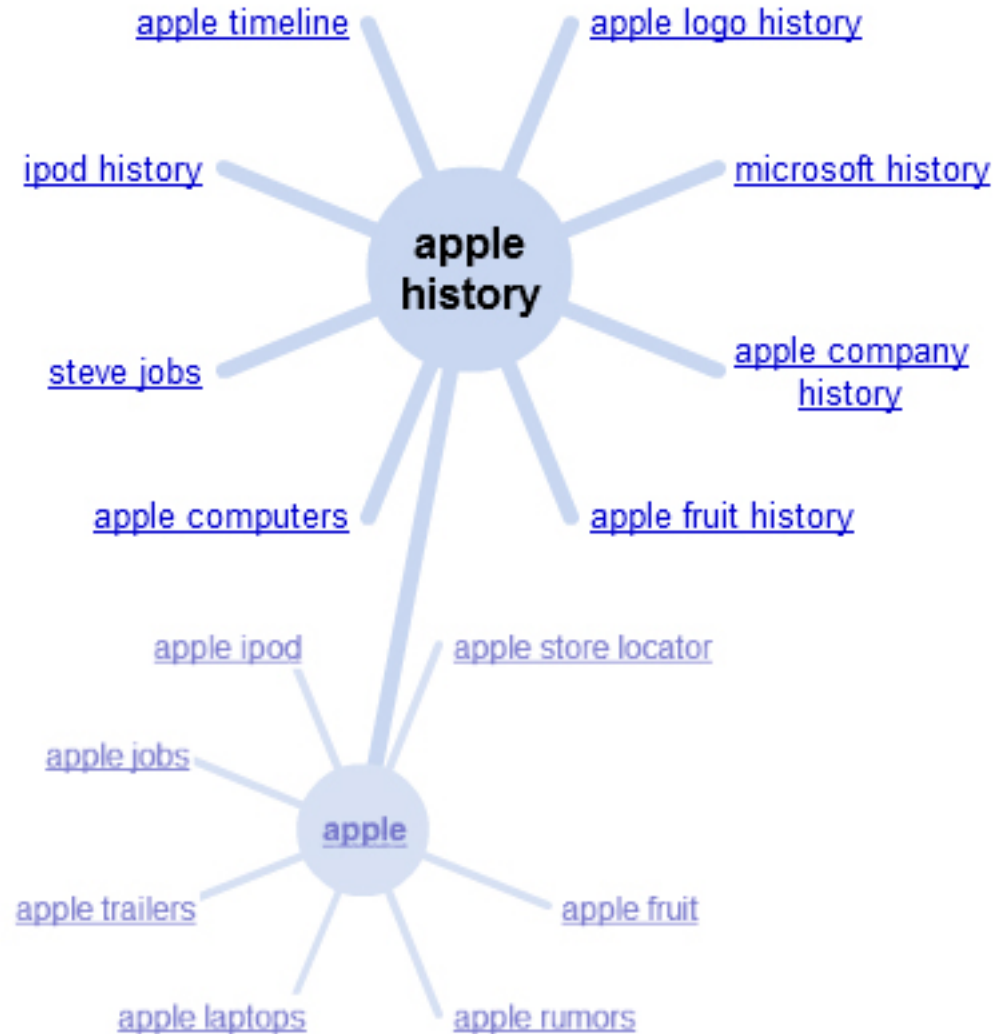
## Clustering species ("phylogeny")



[Lindblad-Toh et al., Nature 2005]

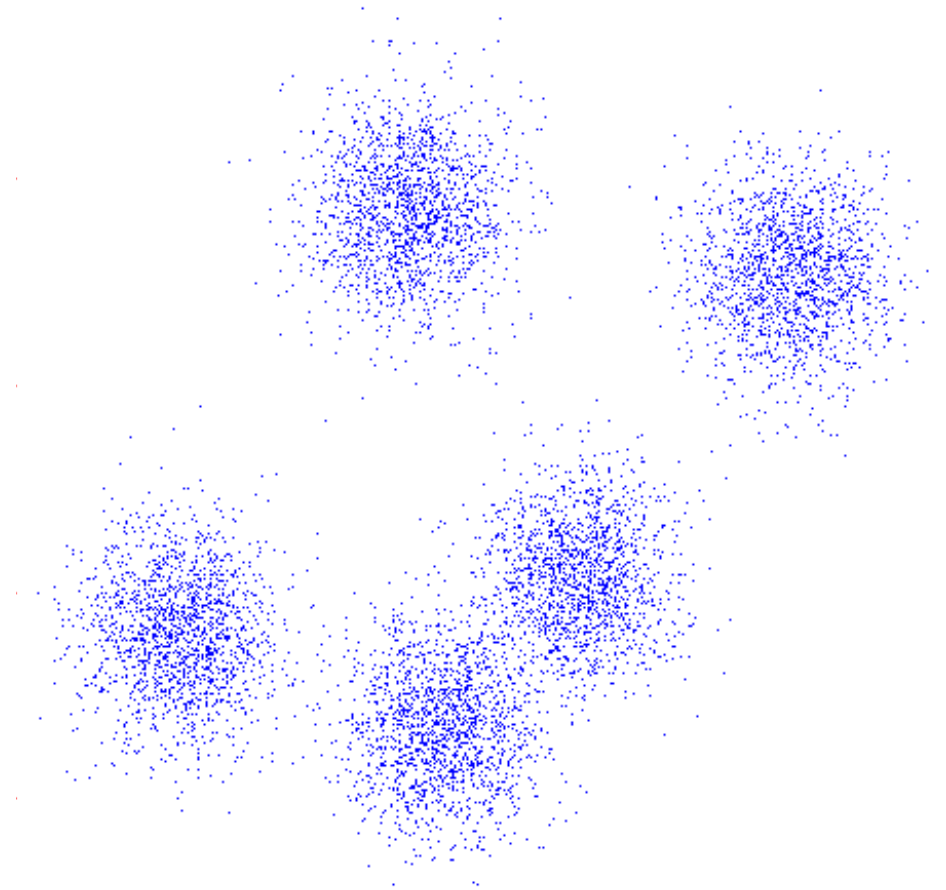
# Clustering examples

## Clustering search queries



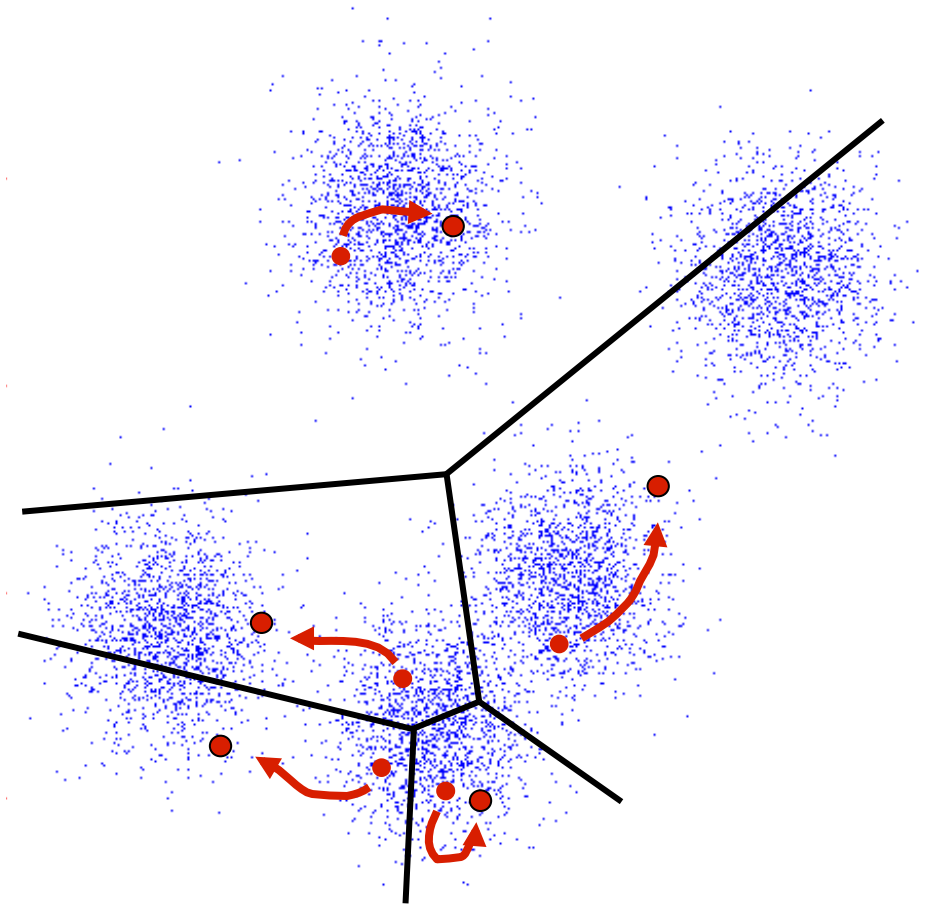
# K-Means

- 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



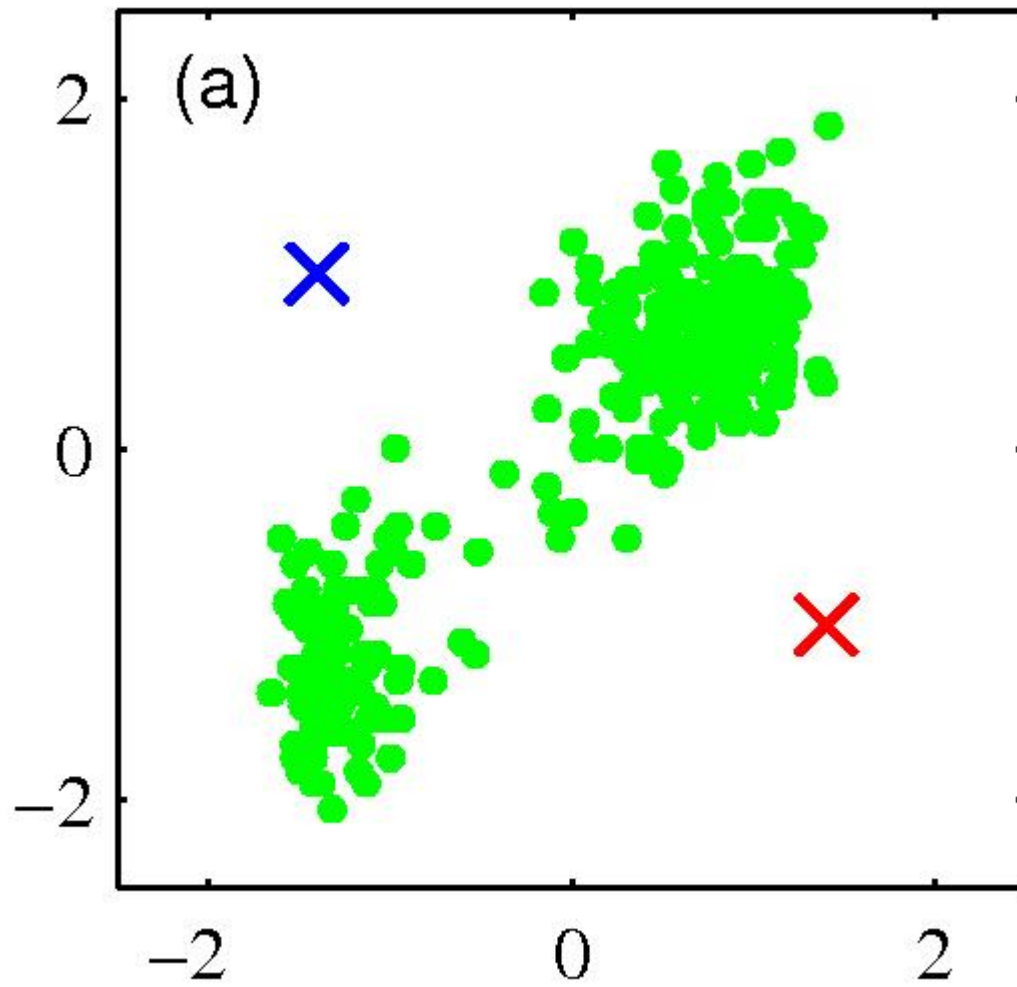
# K-Means

- 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





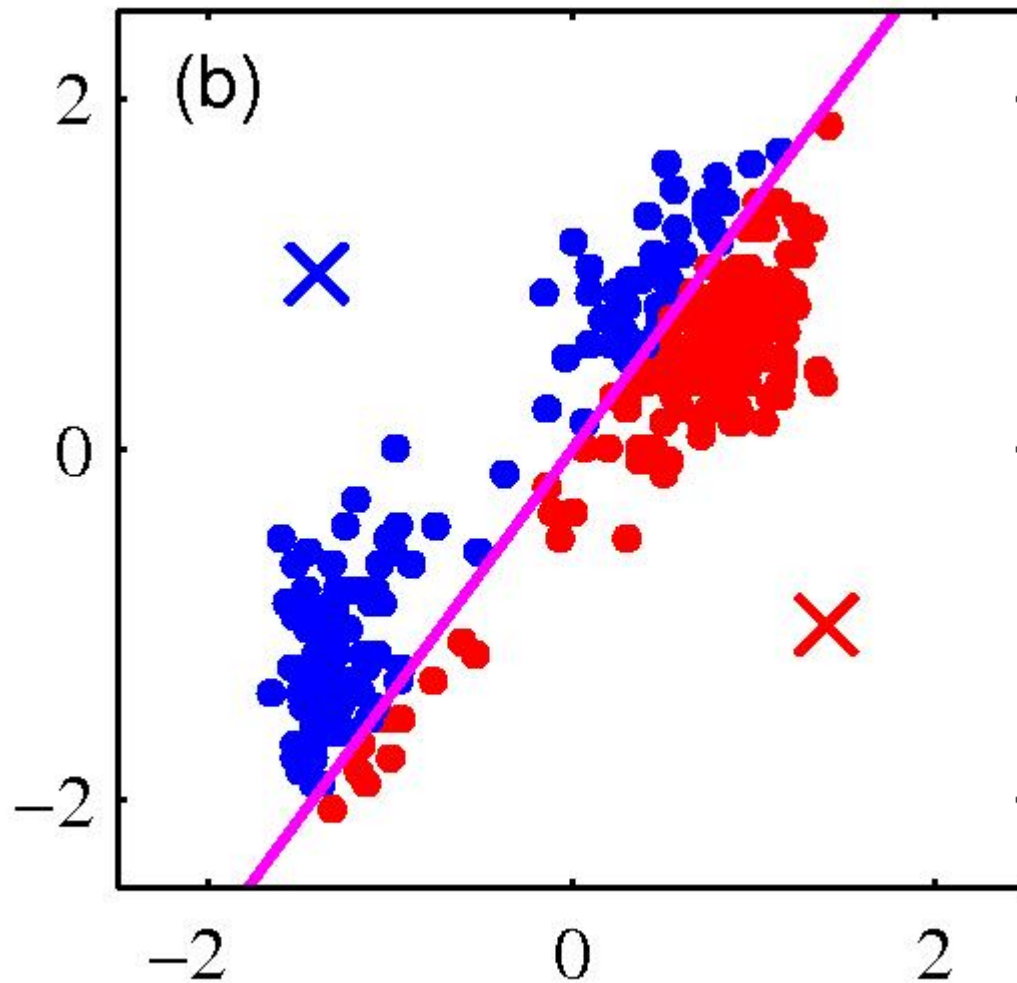
# K-means clustering: Example



- Pick  $K$  random points as cluster centers (means)

Shown here for  $K=2$

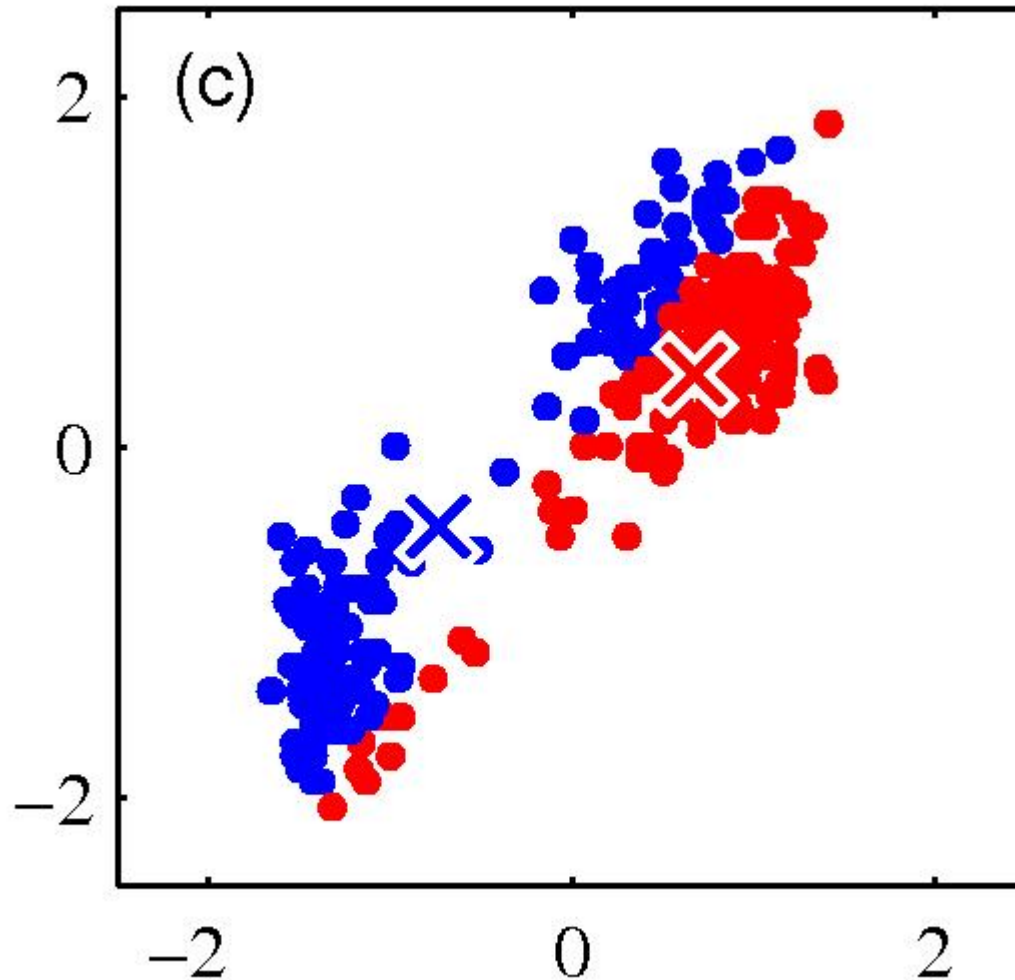
# K-means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

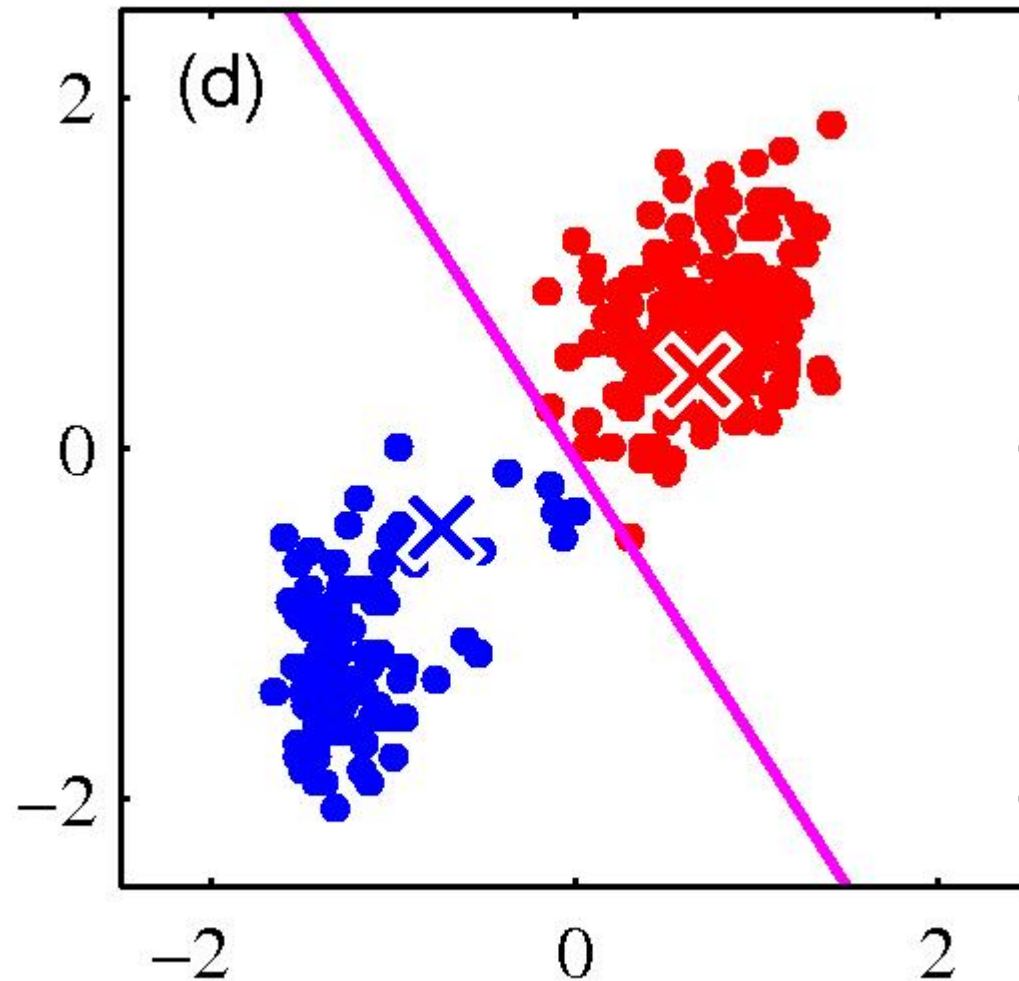
# K-means clustering: Example



## Iterative Step 2

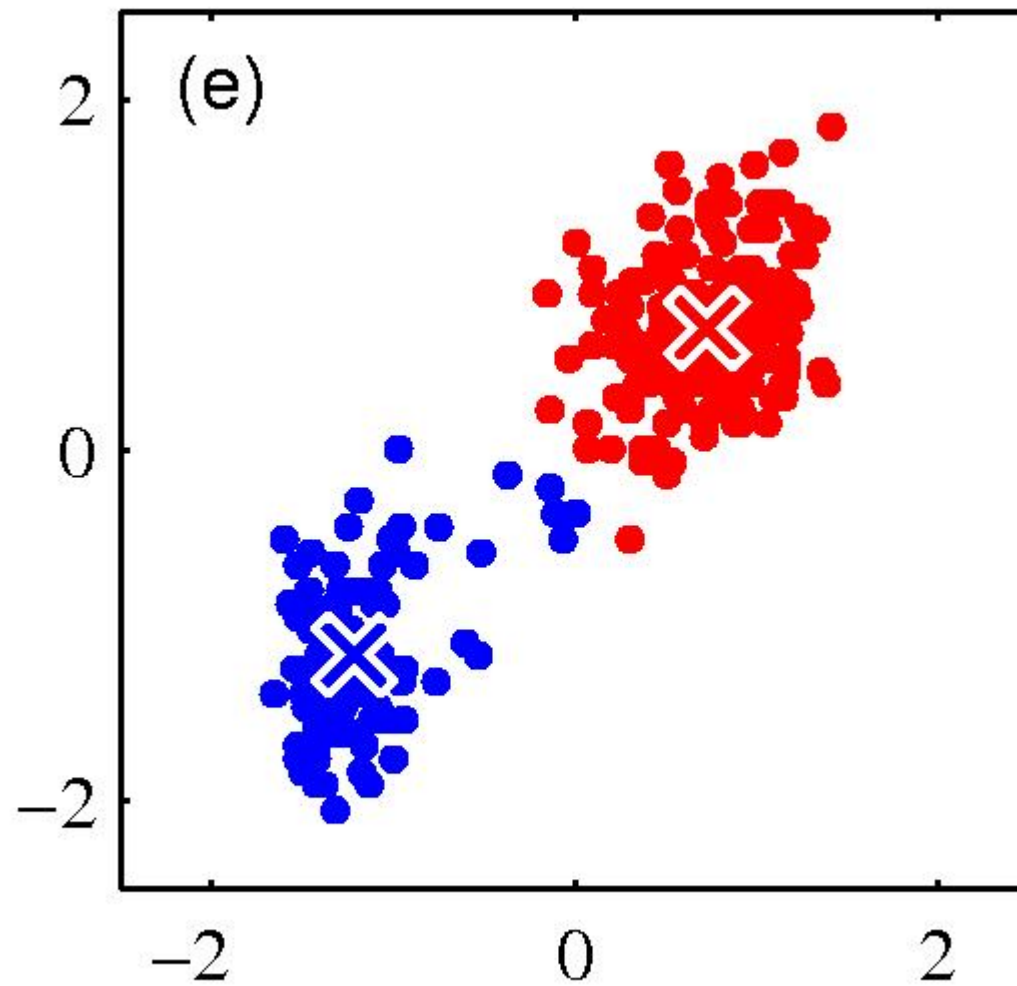
- Change the cluster center to the average of the assigned points

# K-means clustering: Example

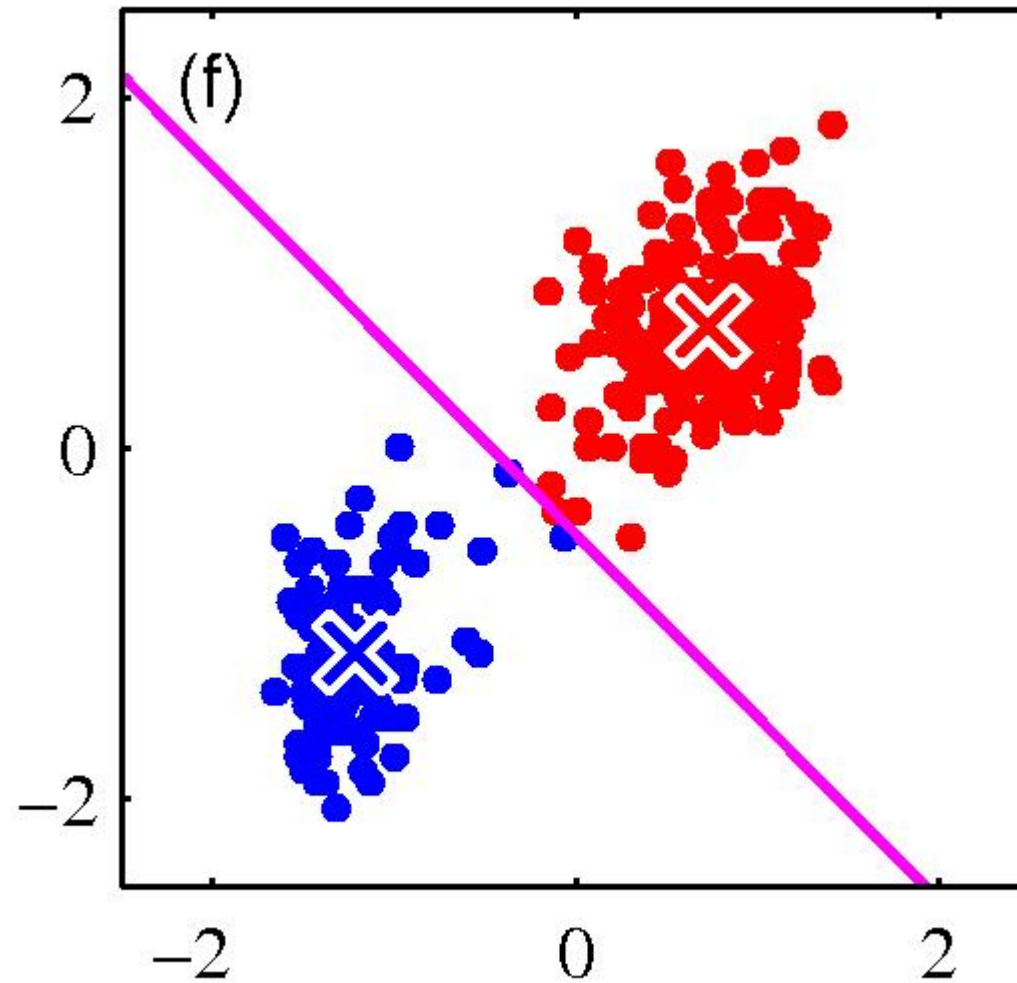


- Repeat until convergence

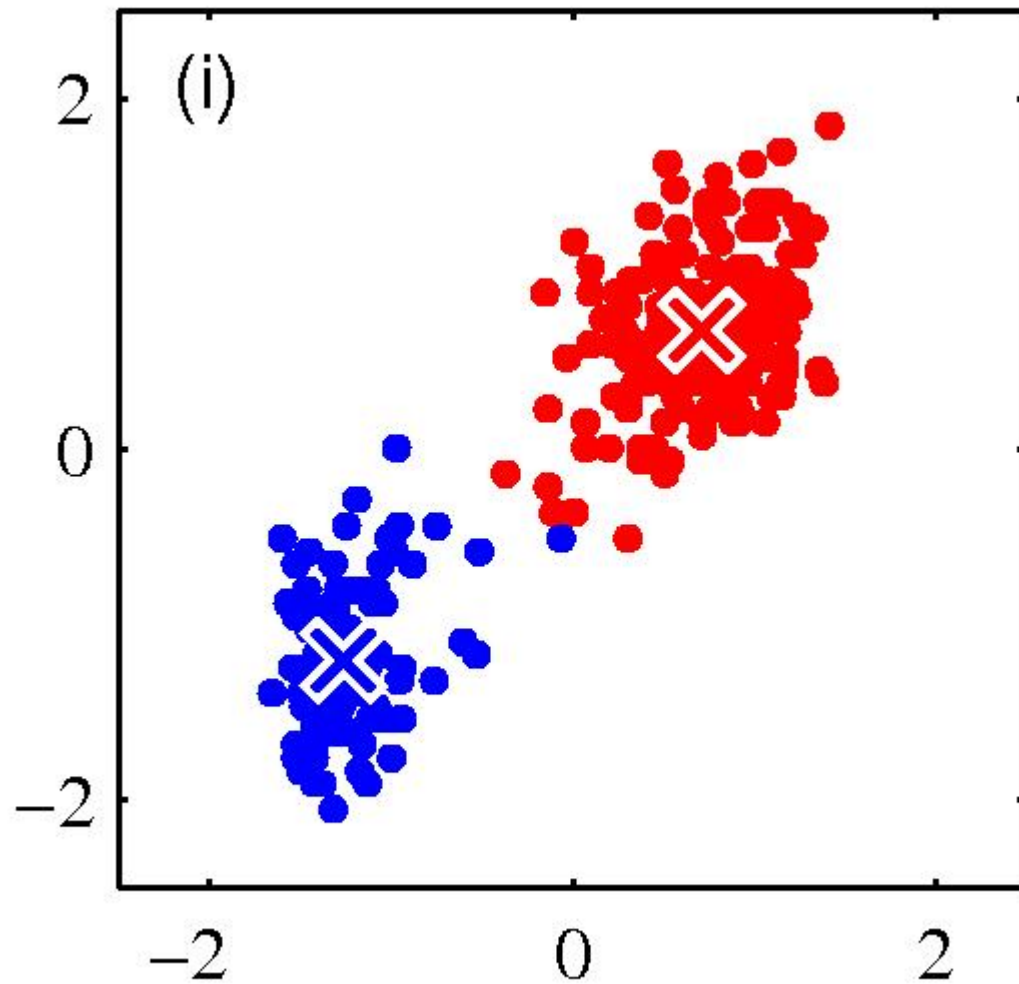
# K-means clustering: Example



# 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  
 $O(N)$



# Kmeans Convergence

## Objective

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

1. Fix  $\mu$ , optimize  $C$ :

$$\min_C \sum_{i=1}^k \sum_{x \in C_i} |x - \mu_i|^2 = \min_c \sum_i^n |x_i - \mu_{x_i}|^2$$

**Step 1 of kmeans**

2. Fix  $C$ , optimize  $\mu$ :

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

- Take partial derivative of  $\mu_i$  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

# Example: K-Means for Segmentation

K=2



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

Original



# Example: K-Means for Segmentation

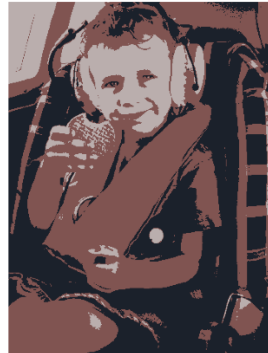
K=2



K=3



Original



# Example: K-Means for Segmentation

K=2



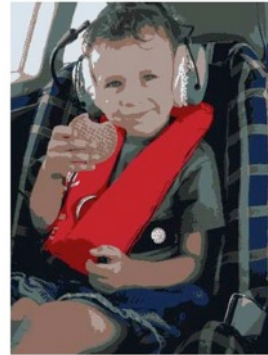
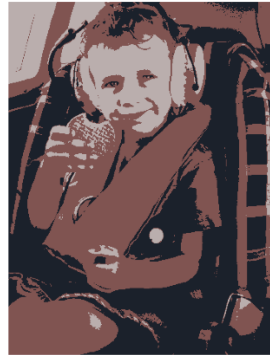
K=3



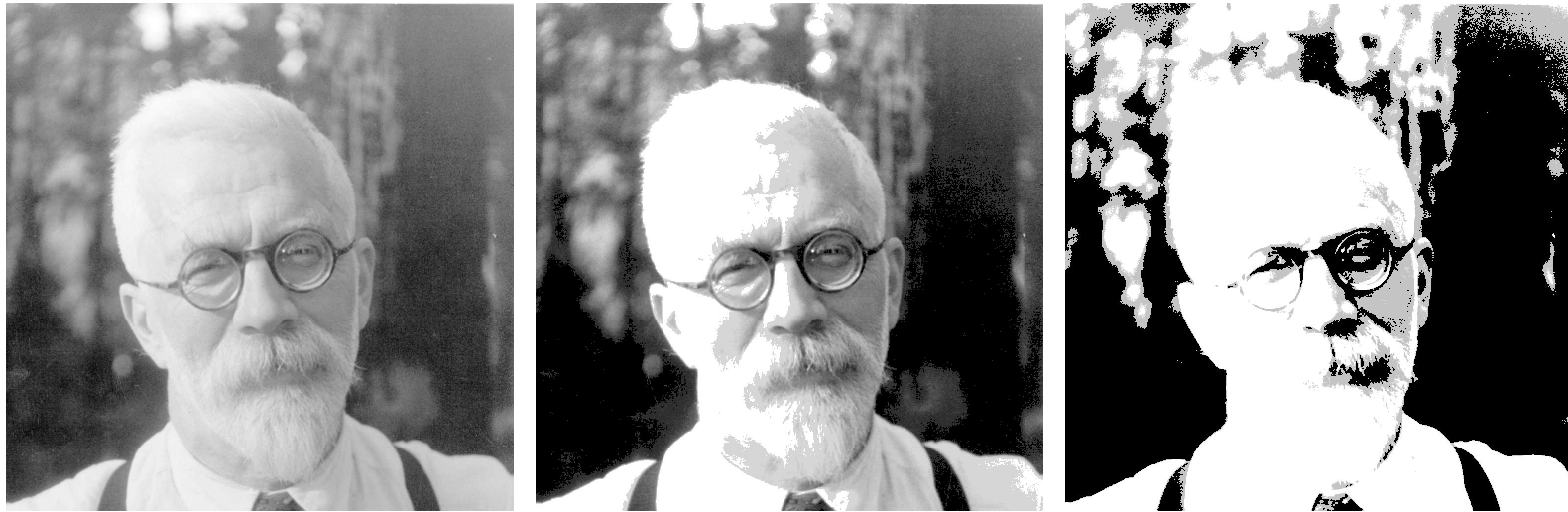
K=10



Original



# Example: Vector quantization

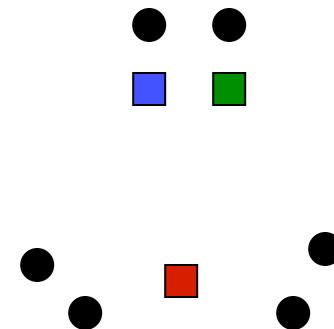
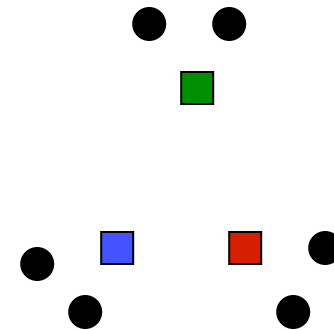


**FIGURE 14.9.** *Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a  $1024 \times 1024$  grayscale image at 8 bits per pixel. The center image is the result of  $2 \times 2$  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]

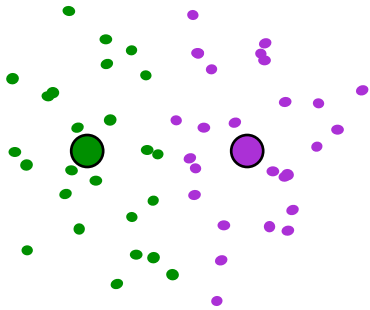
# Initialization

- 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

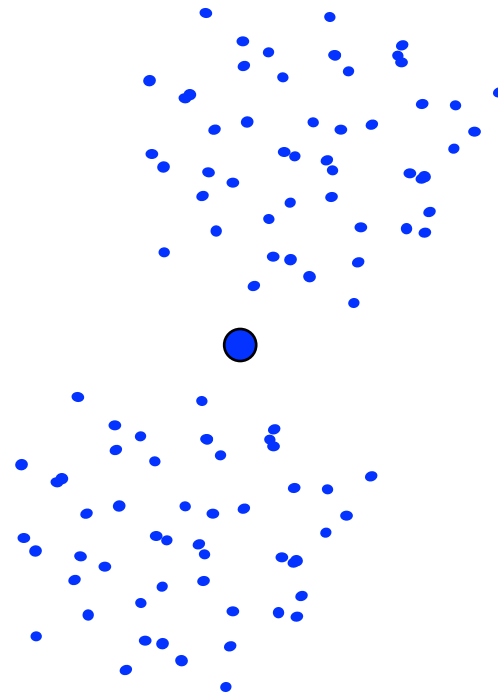


# K-Means Getting Stuck

A local optimum:

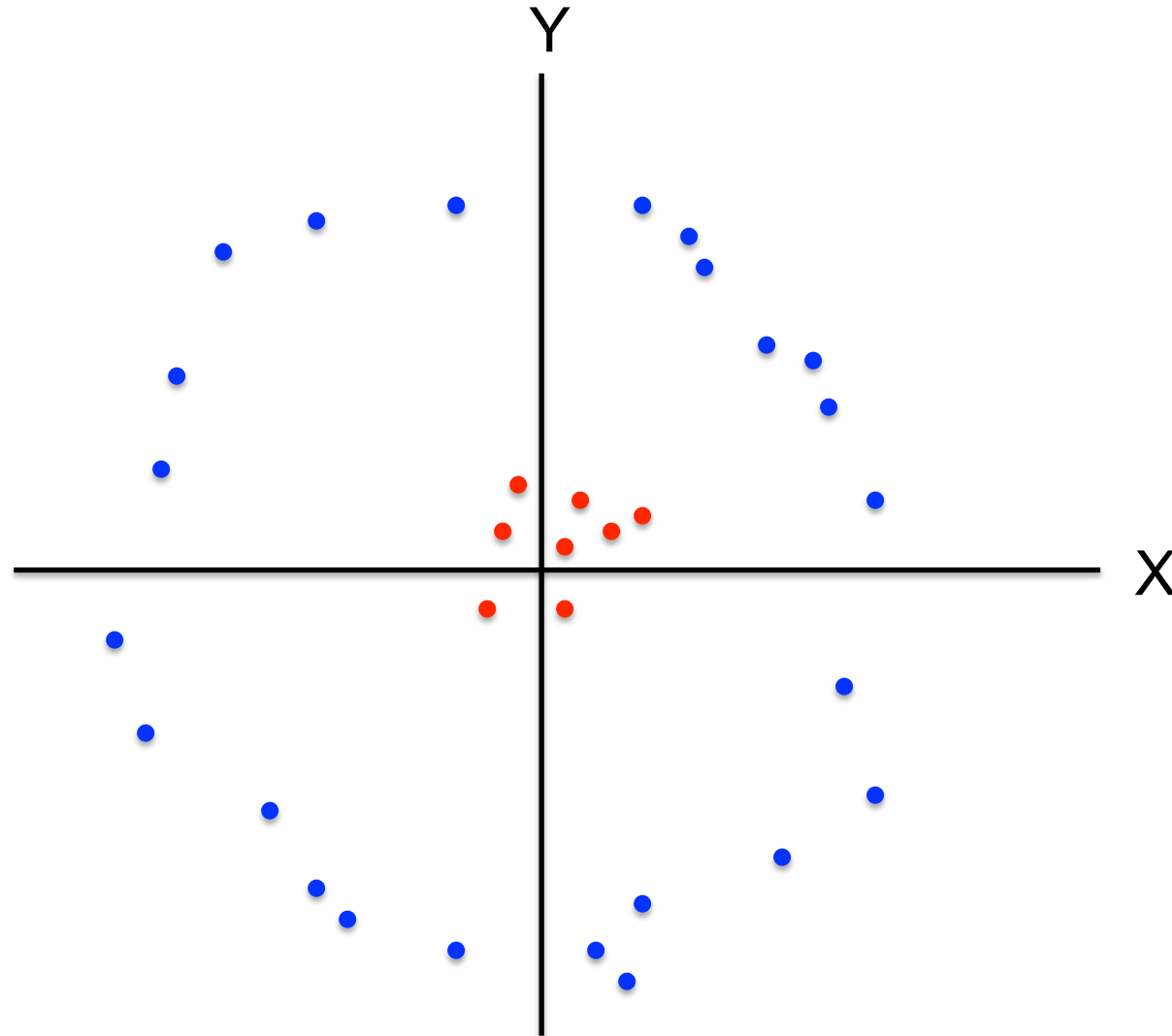


Would be better to have  
one cluster here



... and two clusters here

## K-means not able to properly cluster





Changing the features (distance function)  
can help

