Clustering

Jia-Bin Huang
Virginia Tech

Administrative

- HW 2 grades out.
- HW 3 March 27.
- Started grading Midterm. Results expected to be out by March 27
- Final project question?
- Feedback please! https://goo.gl/forms/A6NSq5yMpkuWEMR43

Where are we? (Supervised Learning)

- K-Nearest Neighbor
- Linear Regression
- Naive Bayes
- Logistic Regression
- Regularization
- Support Vector Machines
- Neural Networks
- Diagnosing ML systems

What’s next? (Unsupervised Learning)

- Clustering
- Expectation Maximization (EM) and Gaussian Mixture Models (GMM)
- Dimensionality reduction
- Anomaly Detection
- Recommender systems

What's next? (Advanced topics)

- Supervised Learning
  - Ensemble methods
  - Structured prediction
  - Semi-supervised learning
- Generative Models
- Probabilistic Graphical Models
- Sequence Prediction Models
- Reinforcement Learning
  - Markov Decision Process, Q-learning, Policy gradient

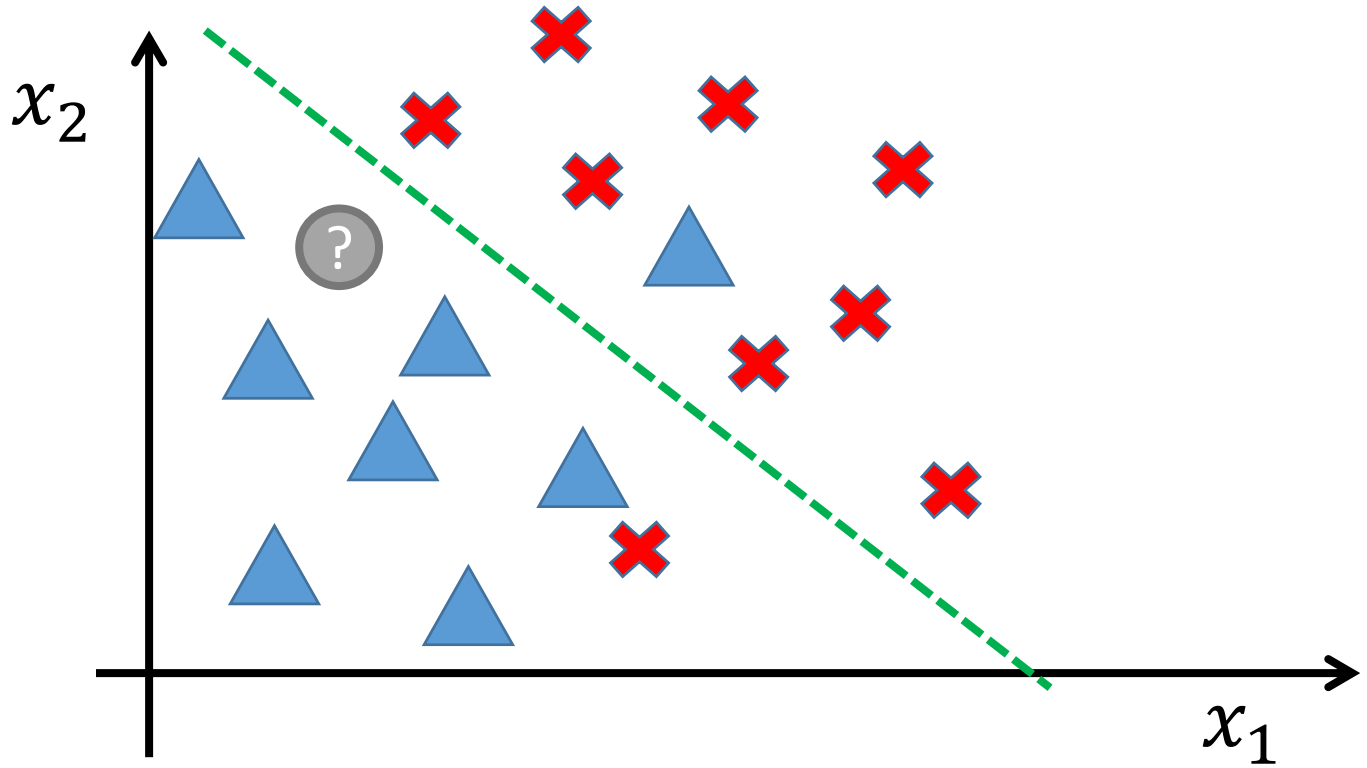
Clustering

- Intro to unsupervised learning
- K-means algorithm
- Optimization objective
- Initialization and the number of clusters
- Hierarchical clustering

Clustering

- **Intro to unsupervised learning**
- K-means algorithm
- Optimization objective
- Initialization and the number of clusters
- Hierarchical clustering

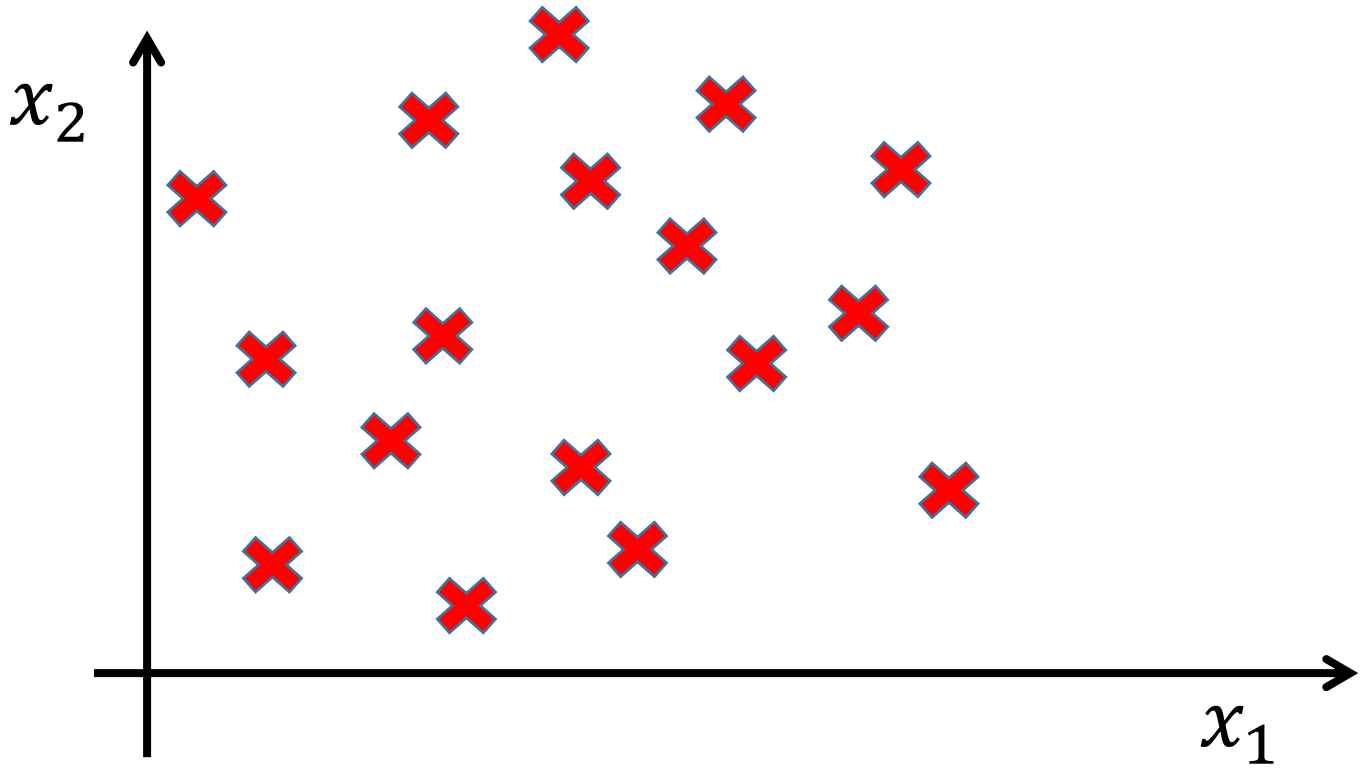
Supervised learning



- Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

Slide credit: Andrew Ng

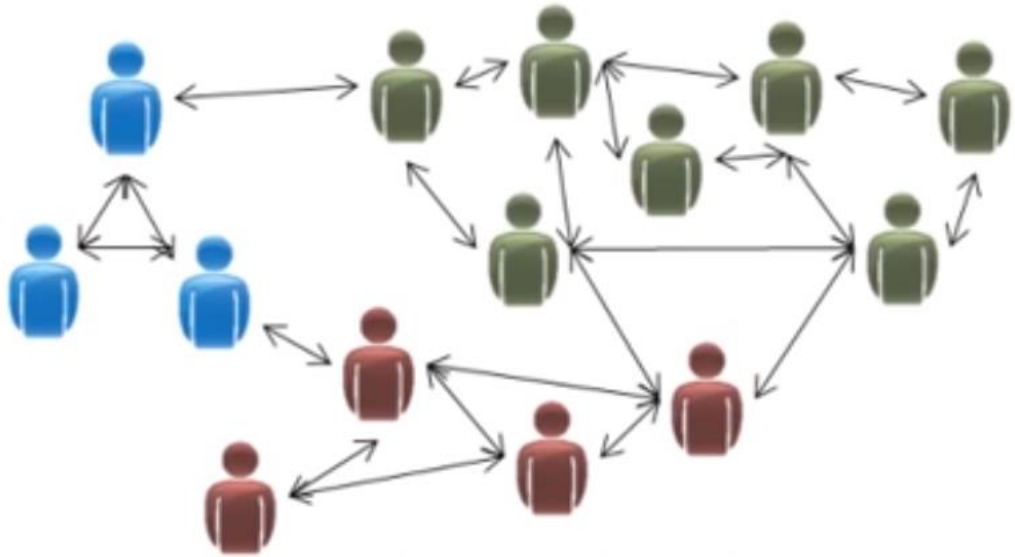
Unsupervised learning



- Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$

Slide credit: Andrew Ng

Organize computing clusters



Social network analysis

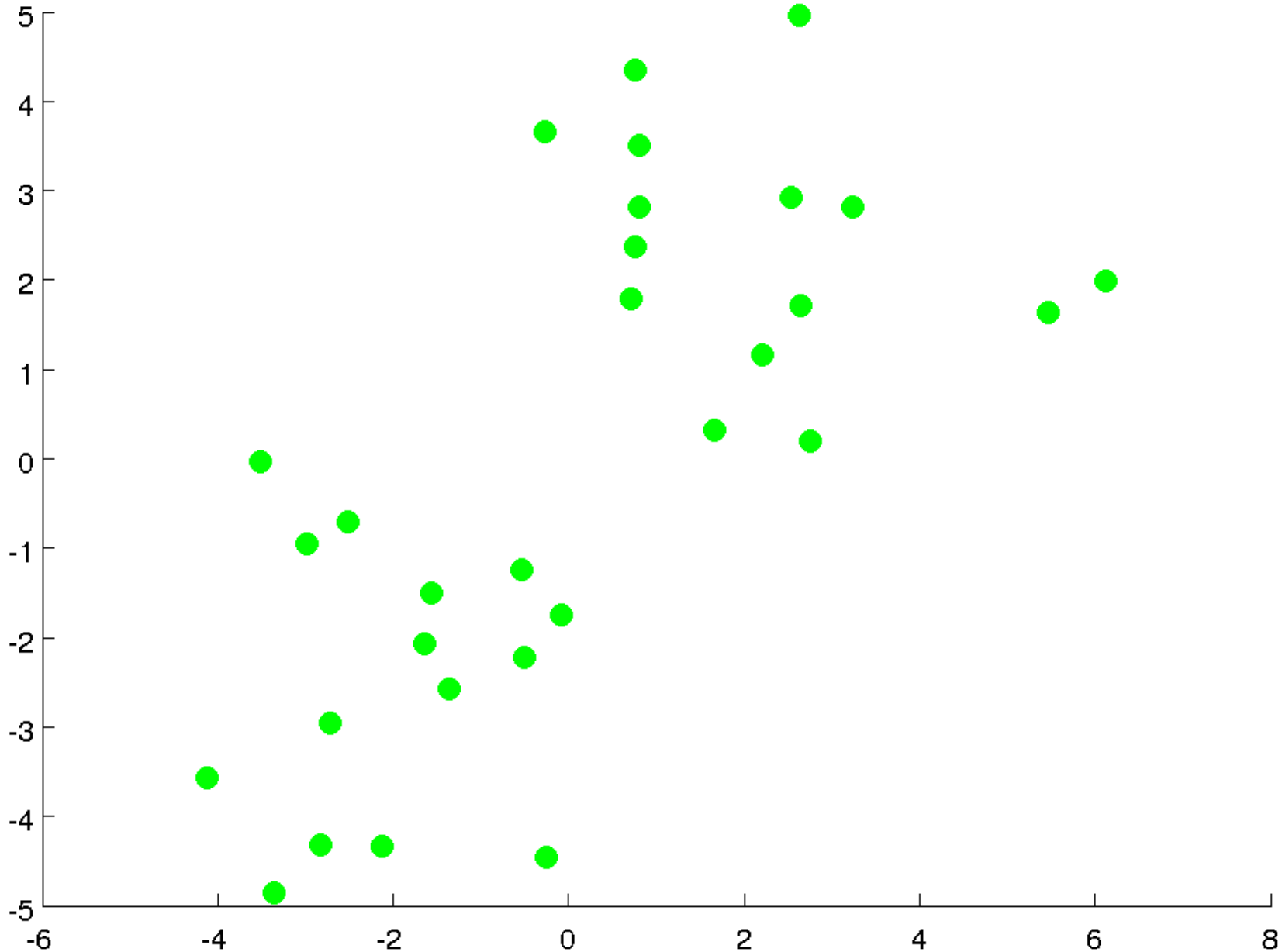
Market segmentation

Astronomical data analysis

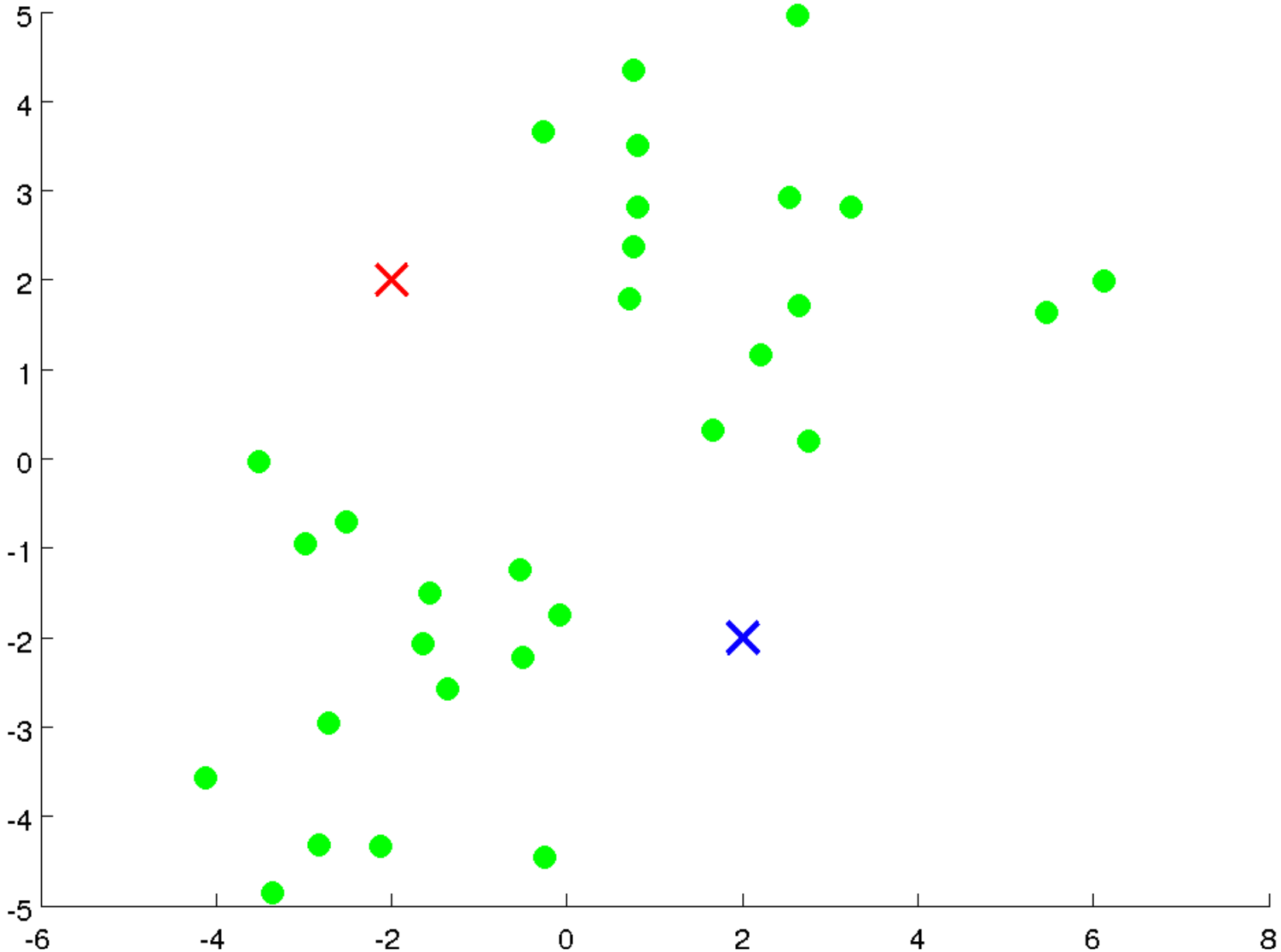
Slide credit: Andrew Ng

Clustering

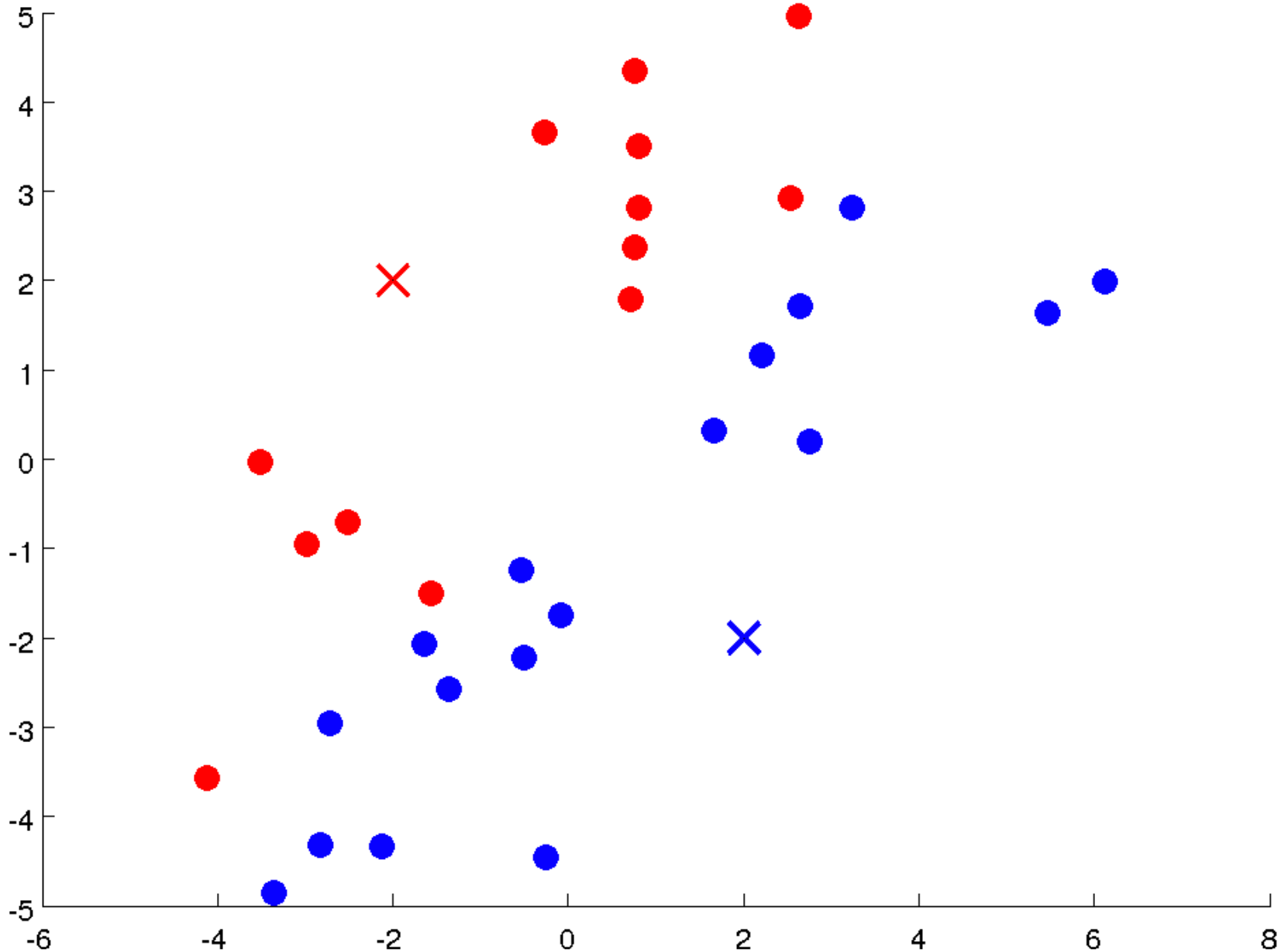
- Intro to unsupervised learning
- **K-means algorithm**
- Optimization objective
- Initialization and the number of clusters
- Hierarchical clustering



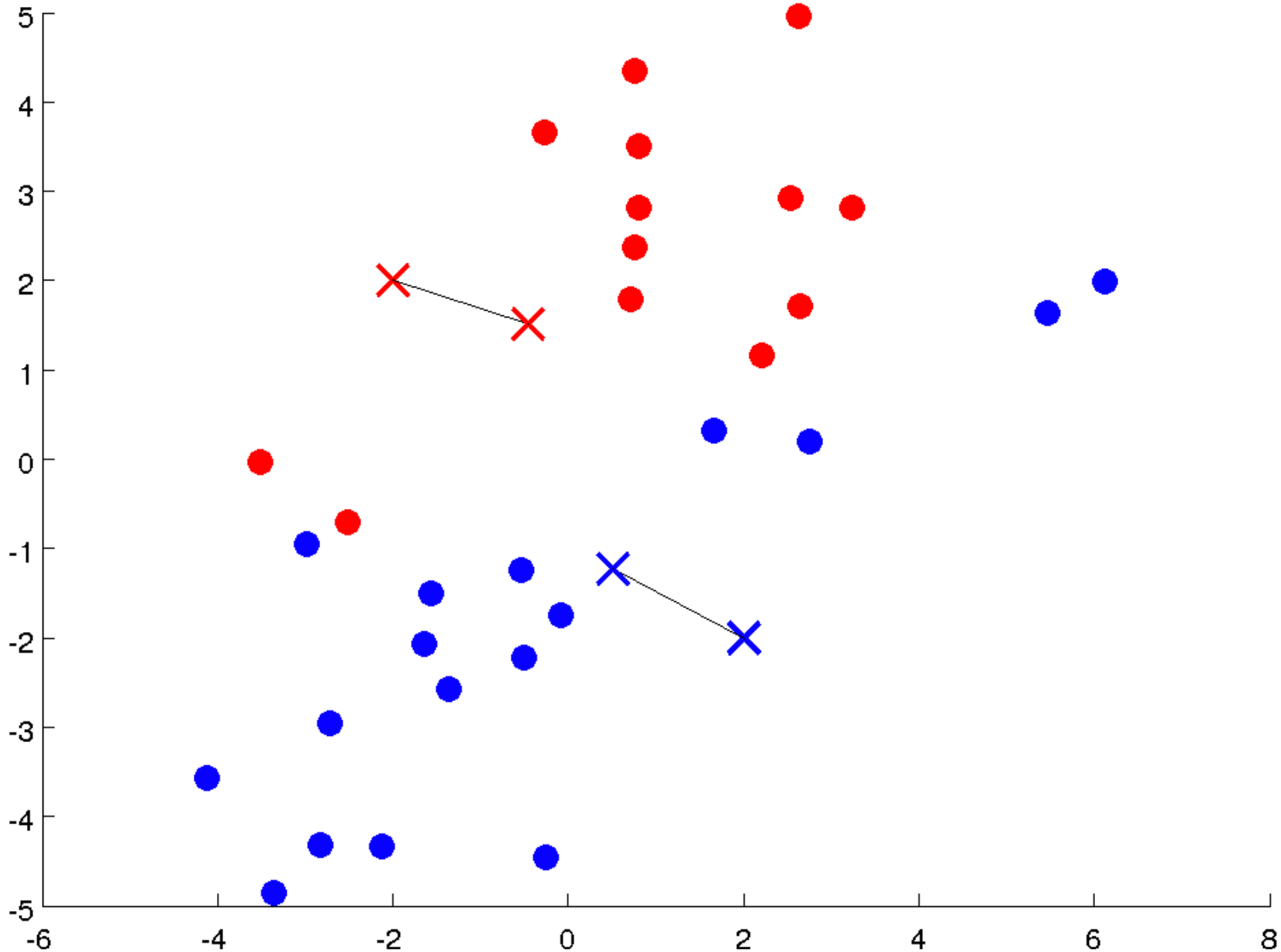
Slide credit: Andrew Ng



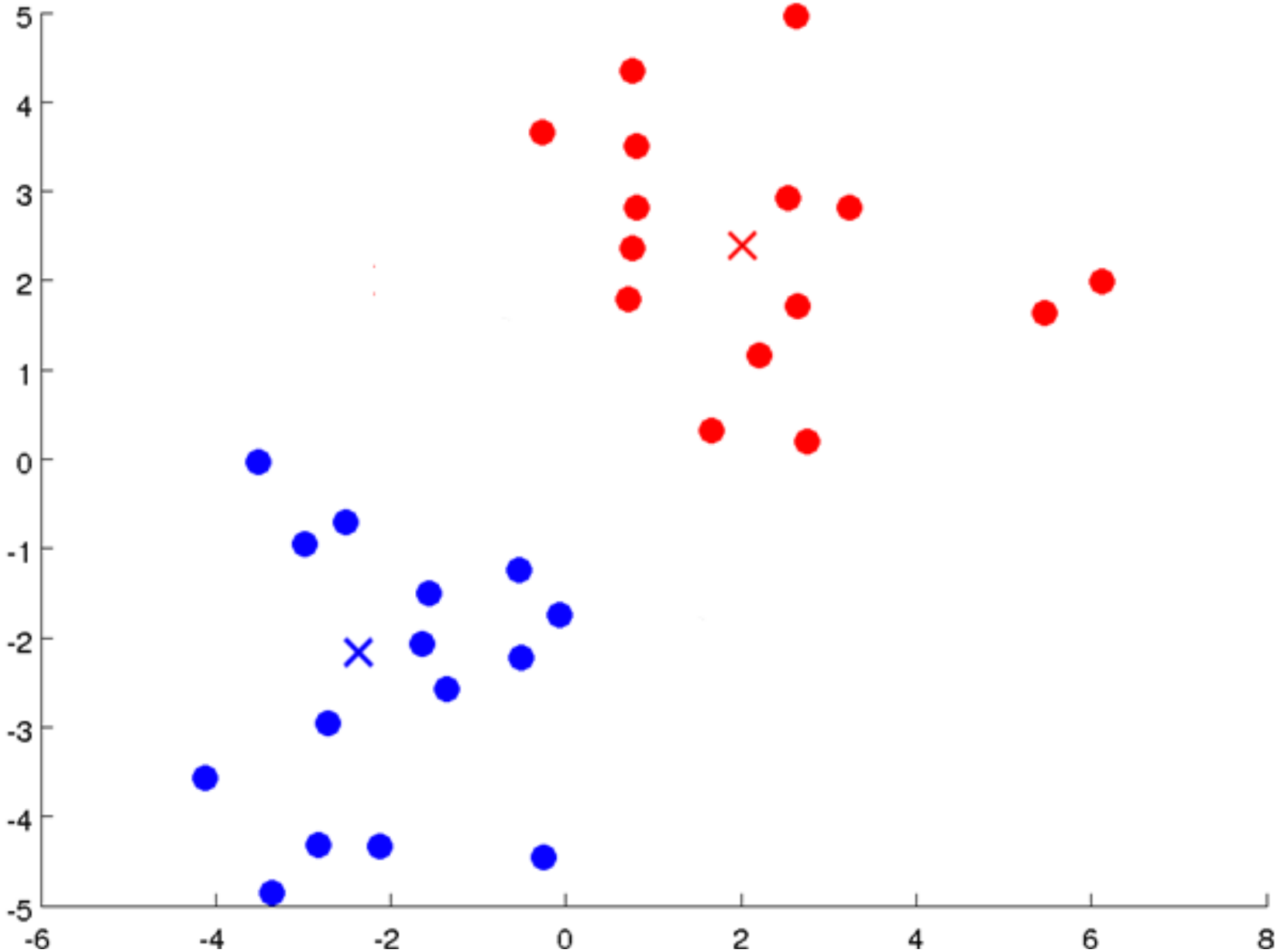
Slide credit: Andrew Ng



Slide credit: Andrew Ng



Slide credit: Andrew Ng



Slide credit: Andrew Ng

K-means algorithm

- Input:
  - K (number of clusters)
  - Training set $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$

- $x^{(i)} \in \mathbb{R}^n$ (note: drop $x_0 = 1$ convention)

Slide credit: Andrew Ng

K-means algorithm

- Randomly initialize K cluster centroids $\mu_1, \mu_2, \cdots, \mu_K \in \mathbb{R}^n$

Repeat{

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

Cluster assignment step

for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

Centroid update step

}

Slide credit: Andrew Ng

Clustering

- Intro to unsupervised learning
- K-means algorithm
- **Optimization objective**
- Initialization and the number of clusters
- Hierarchical clustering

K-means optimization objective

- $c^{(i)}$ = Index of cluster (1, 2, … K) to which example $x^{(i)}$ is currently assigned
- μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$)
- $\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned
- Optimization objective:

$$J\left(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K\right) = \frac{1}{m}\sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

$$\min_{\substack{c^{(1)},\cdots,c^{(m)} \\ \mu_1,\cdots,\mu_K}} J\left(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K\right)$$

Example:

$x^{(i)} = 5$

$c^{(i)} = 5$

$\mu_{c^{(i)}} = \mu_5$

Slide credit: Andrew Ng

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \cdots, \mu_K \in \mathbb{R}^n$

Repeat{

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

Cluster assignment step

$$J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K) = \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

Centroid update step

$$J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K) = \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

}

Slide credit: Andrew Ng

Objective based clustering

- Input:
  - K (number of clusters)
  - Training set $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$
  - Distance/dissimilarity measure $\mathrm{d}(x^{(i)}, x^{(j)})$

- Goal: output a partition of the data
  - K-means: Euclidean distance
  - K-median: Absolute distance
  - K-center: minimize the maximum radius

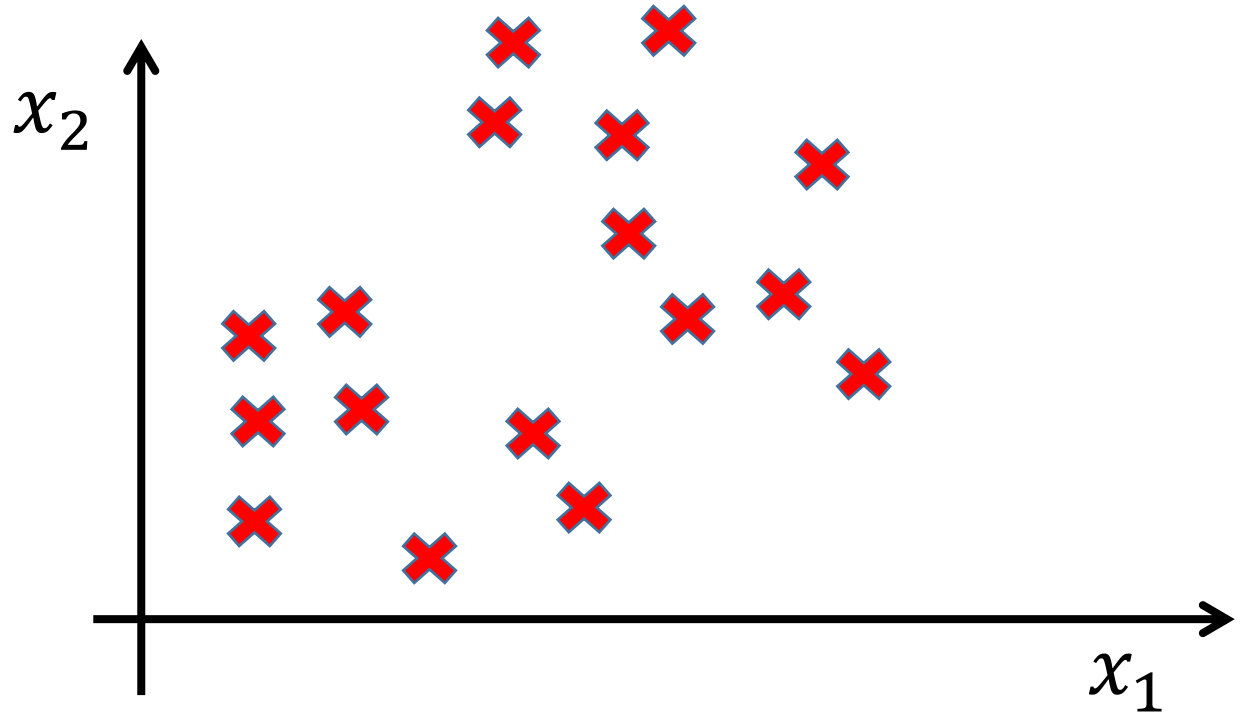
Clustering

- Intro to unsupervised learning
- K-means algorithm
- Optimization objective
- **Initialization and the number of clusters**
- Hierarchical clustering

Random initialization

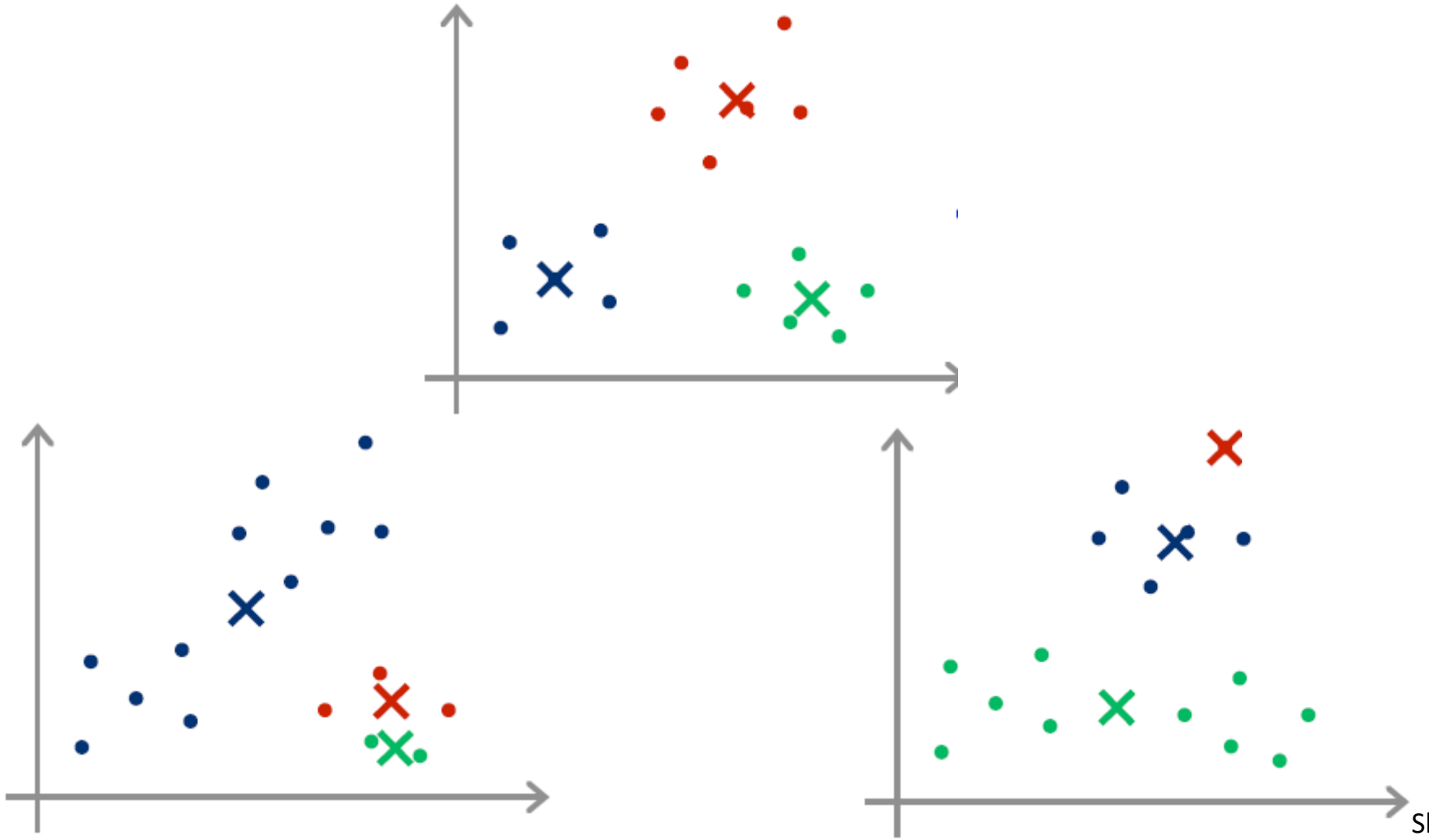
- Randomly pick K training examples

- Set $\mu_1, \mu_2, \cdots, \mu_K$ equal to those K examples



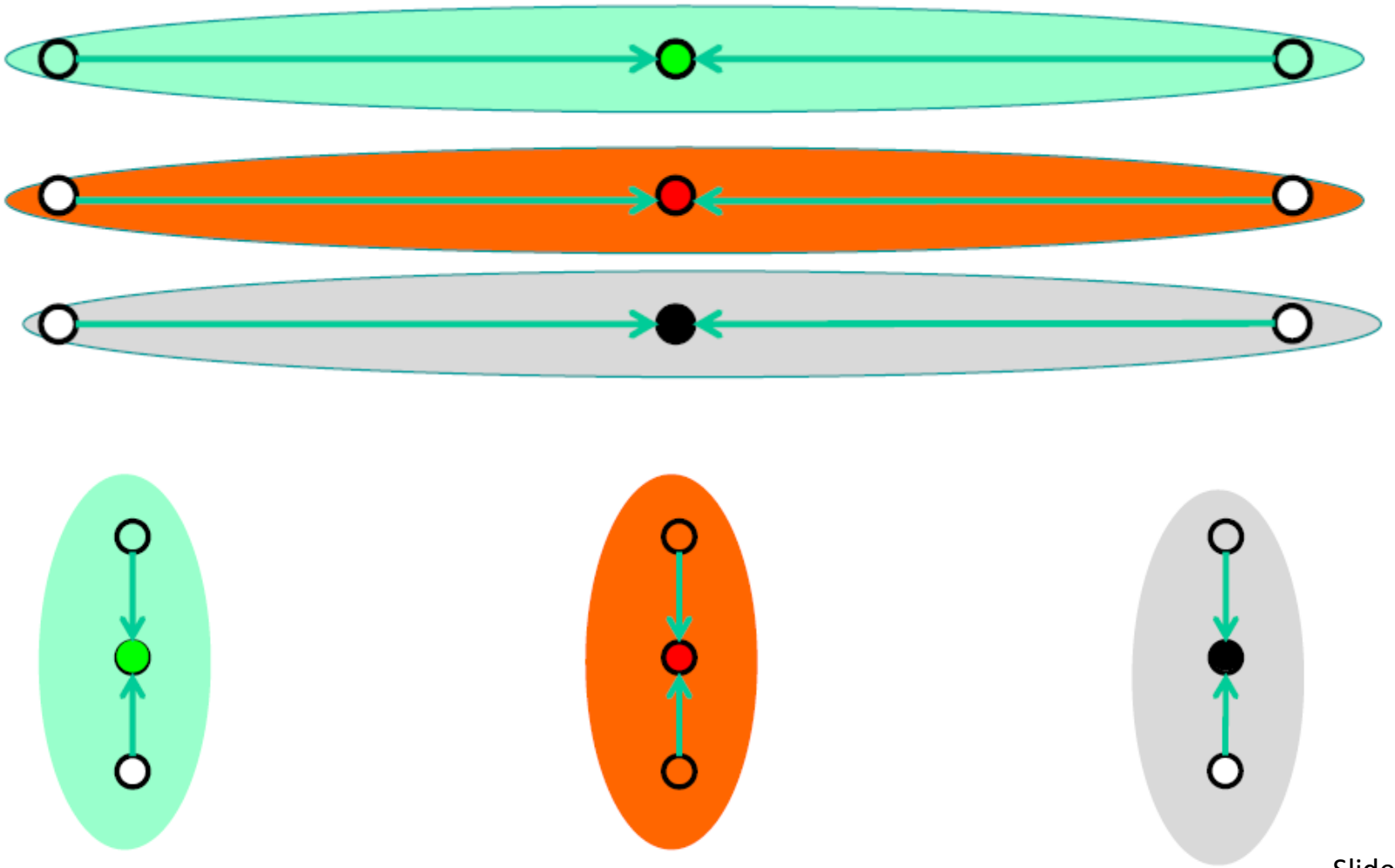
Slide credit: Andrew Ng

Local optima



Slide credit: Andrew Ng

Local optima



Slide credit: Maria-Florina Balcan

1) Multiple random initialization

For i = 1 to 100 {

Randomly initialize K-means

Run K-means. Get $c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K$

Compute the cost function (distortion)

$J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K)$

}

Pick clustering that gave the lowest cost $J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K)$

Slide credit: Andrew Ng

2) Furthest Point Heuristic

Choose μ_1 arbitrarily (or at random)

For j = 2 to K

Pick μ_j among data points $x^{(1)}, x^{(2)}, \cdots, x^{(m)}$ that is farthest from previously chosen $\mu_1, \mu_2, \cdots, \mu_{j-1}$

Slide credit: Maria-Florina Balcan

3) K-means++ initialization

- Interpolate between random and furthest point initialization
- Idea: Let $\mathrm{D}(x)$ be the distance between a point x and its nearest center. Chose the next center proportional to $\mathrm{D}^2(x)$.

Choose μ_1 arbitrarily (or at random)

For j = 2 to K

Pick μ_j among data points $x^{(1)}, x^{(2)}, \cdots, x^{(m)}$ according to the distribution

$$\Pr\left(\mu_j = x^{(i)}\right) \propto \min_{j'<j} \left\| x^{(i)} - \mu_{j'} \right\|^2$$

[Arthur and Vassilvitskii 2007]

3) K-means++ initialization

- Interpolate between random and furthest point initialization
- Idea: Let $\mathrm{D}(x)$ be the distance between a point x and its nearest center. Chose the next center proportional to $\mathrm{D}^{\alpha}(x)$.

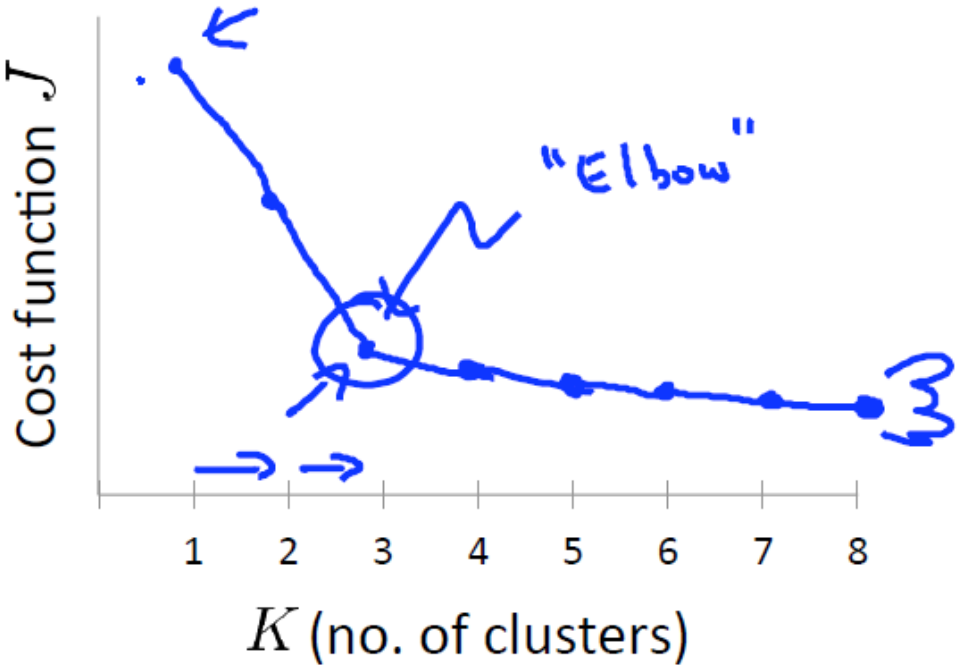
- $\alpha = 0$: random sampling
- $\alpha = \infty$: furthest point
- $\alpha = 2$: K-means++

scikit-learn uses K-means++ as default

[Arthur and Vassilvitskii 2007]

How to choose K?

- 1) Elbow method

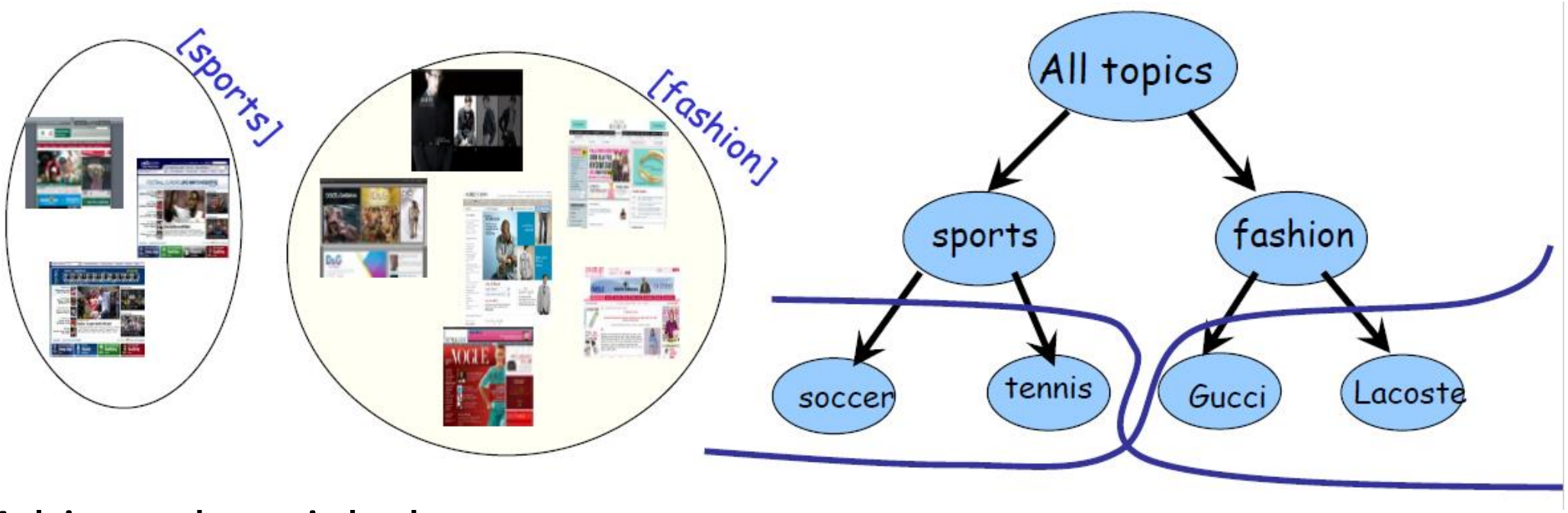


- 2) Try multiple K and use performance from downstream task for selection

Clustering

- Intro to unsupervised learning
- K-means algorithm
- Optimization objective
- Initialization and the number of clusters
- **Hierarchical clustering**

Hierarchical Clustering



- A hierarchy might be more nature
- Different users might care about different levels of granularity or even prunings.

Slide credit: Maria-Florina Balcan

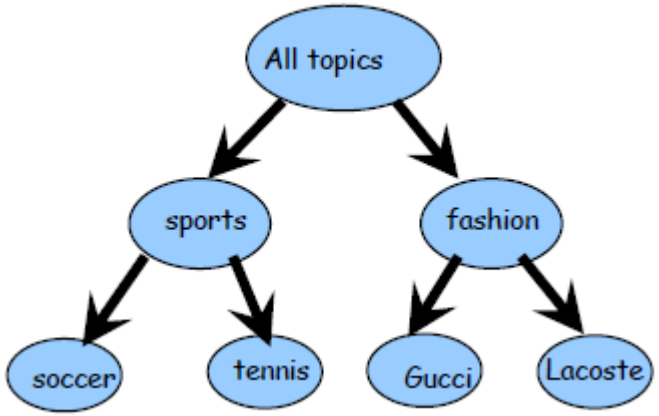
Hierarchical Clustering

- **Top-down (divisive)**
  - Partition data into 2-groups (e.g., 2-means)
  - Recursively cluster each group

- **Bottom-up (agglomerative)**
  - Start with every point in its own cluster.
  - Repeatedly merge the “closest” two clusters
  - Different definitions of “closest” give different algorithms.

Slide credit: Maria-Florina Balcan

Bottom-up (agglomerative)



- Have a distance measure on pairs of objects.
- $d(x, y)$: Distance between x and y

- Single linkage: $\mathrm{dist}(\mathrm{A}, \mathrm{B}) = \min_{\mathrm{x} \in A, x' \in B} \mathrm{d}(\mathrm{x}, \mathrm{x}')$

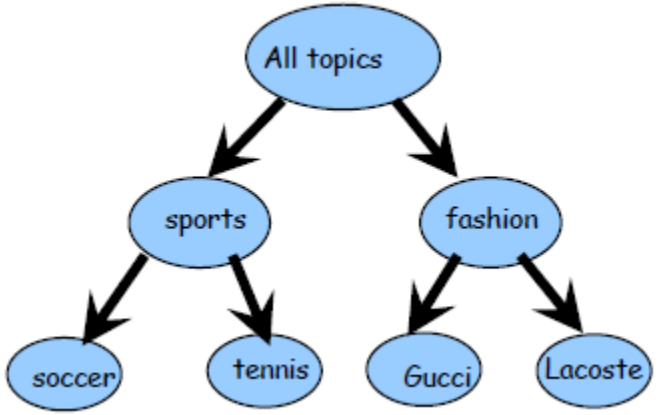
- Complete linkage: $\mathrm{dist}(\mathrm{A}, \mathrm{B}) = \max_{\mathrm{x} \in A, x' \in B} \mathrm{d}(\mathrm{x}, \mathrm{x}')$

- Average linkage: $\mathrm{dist}(\mathrm{A}, \mathrm{B}) = \underset{\mathrm{x} \in A, x' \in B}{\mathrm{average}}\, \mathrm{d}(\mathrm{x}, \mathrm{x}')$

- Ward's method $\mathrm{dist}(\mathrm{A}, \mathrm{B}) = \frac{|A|\,|B|}{|A|+|B|} \|\mathrm{mean}(A) - \mathrm{mean}(B)\|^2$

Slide credit: Maria-Florina Balcan

Bottom-up (agglomerative)



- **Single linkage**: $\text{dist}(\text{A}, \text{B}) = \min_{\text{x}\in A, x'\in B} \text{d}(\text{x}, \text{x}')$
  - At any time, distance between any two points in a connected components < r.

- **Complete linkage**: $\text{dist}(\text{A}, \text{B}) = \max_{\text{x}\in A, x'\in B} \text{d}(\text{x}, \text{x}')$
  - Keep max diameter as small as possible at any level

- **Ward's method** $\text{dist}(\text{A}, \text{B}) = \frac{|A|\,|B|}{|A|+|B|} \|\text{mean}(A) - \text{mean}(B)\|^2$
  - Merge the two clusters such that the increase in k-means cost is as small as possible.
  - Works well in practice

Slide credit: Maria-Florina Balcan

Things to remember

- Intro to unsupervised learning
- K-means algorithm
- Optimization objective
- Initialization and the number of clusters
- Hierarchical clustering