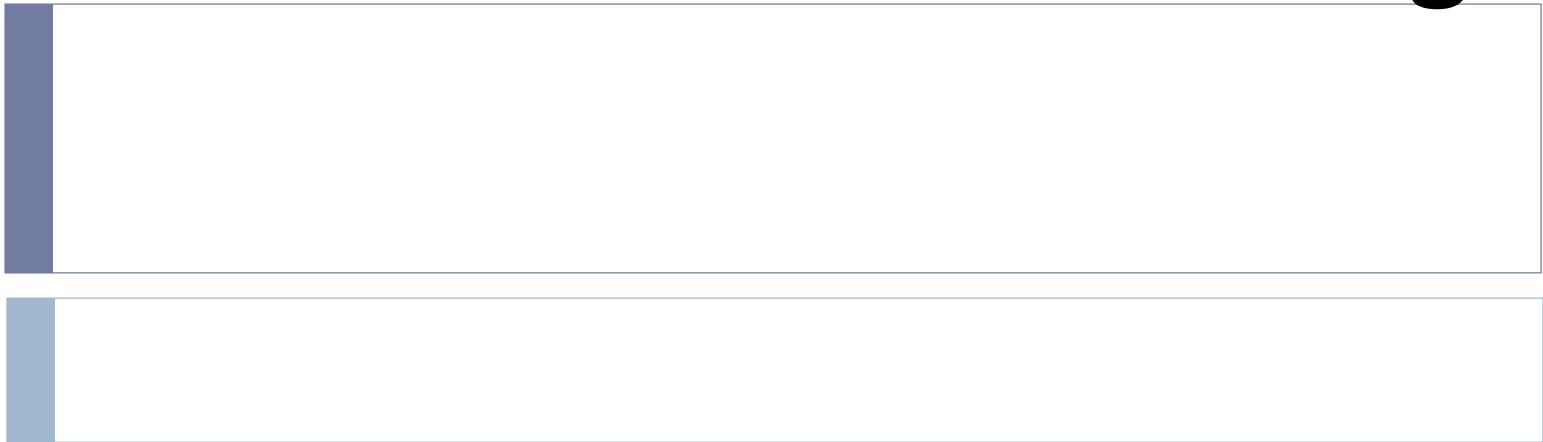
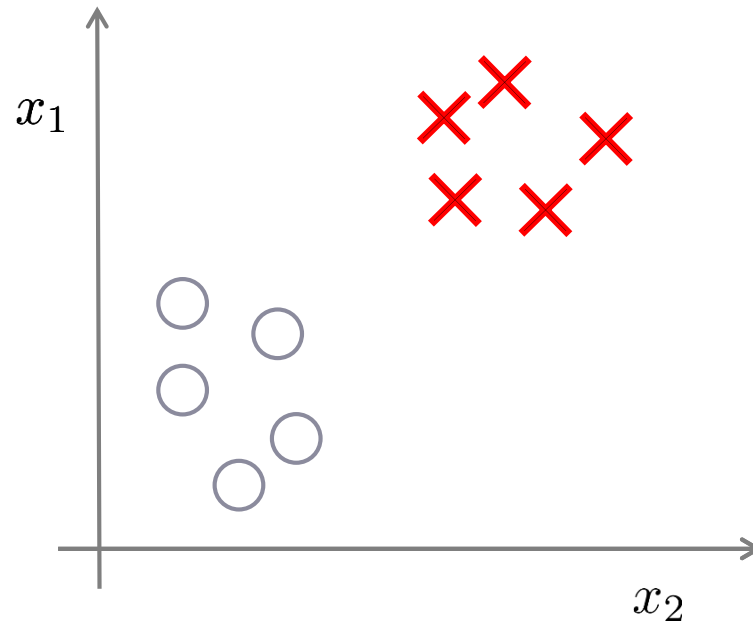


Segmentation by Clustering



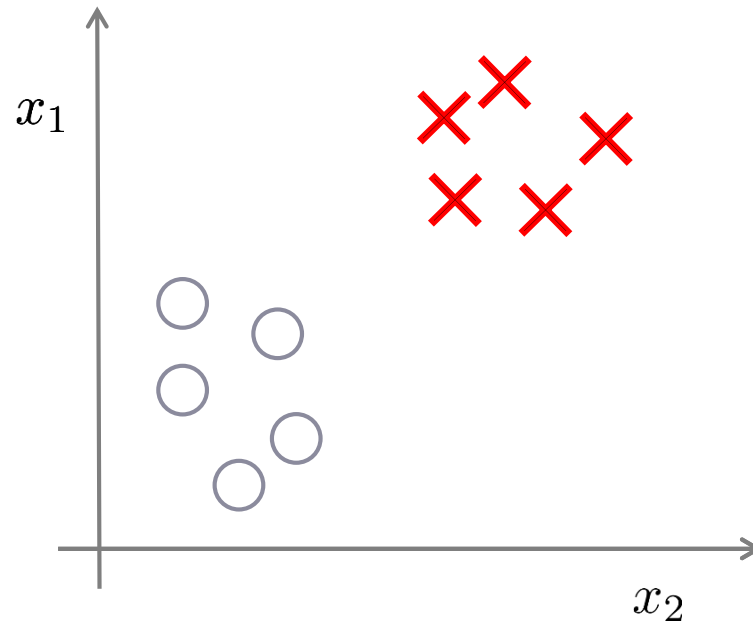
Supervised learning



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



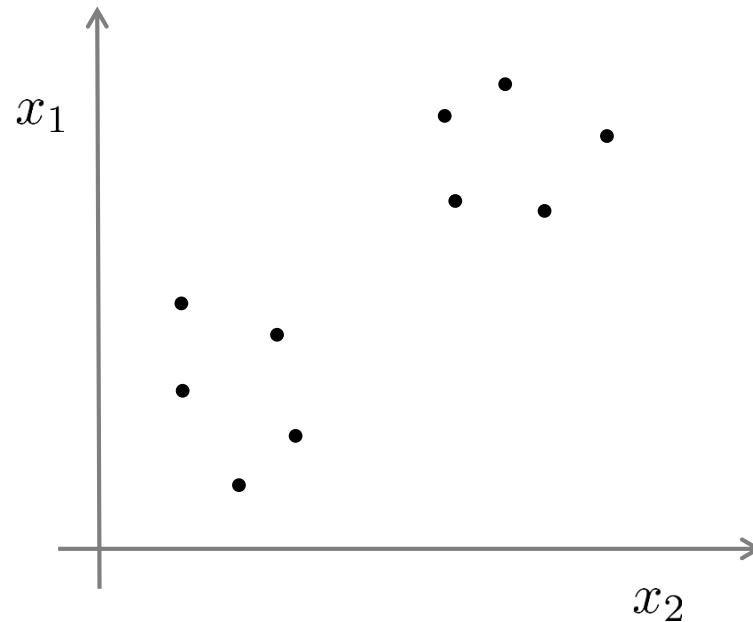
Supervised learning



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



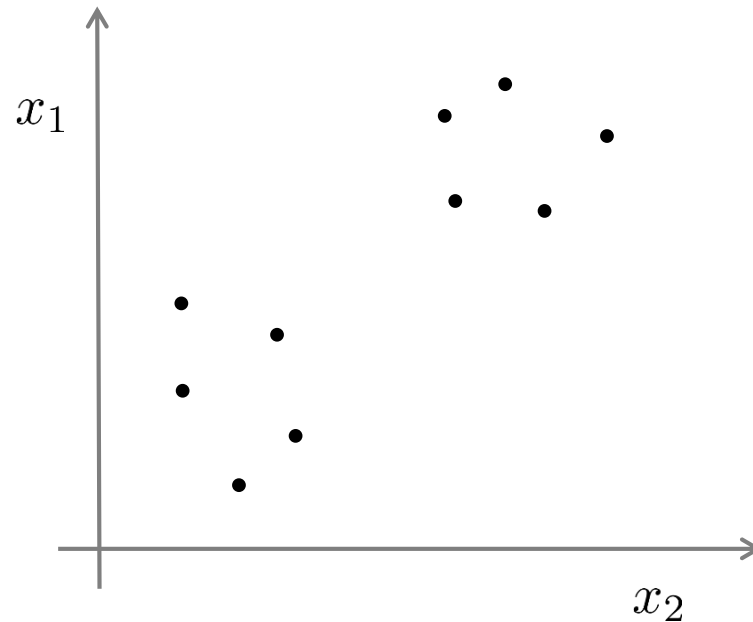
Unsupervised learning



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



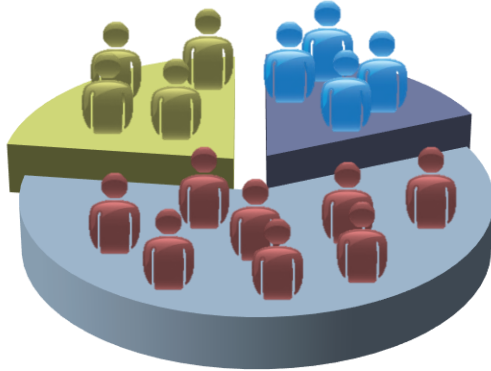
Unsupervised learning



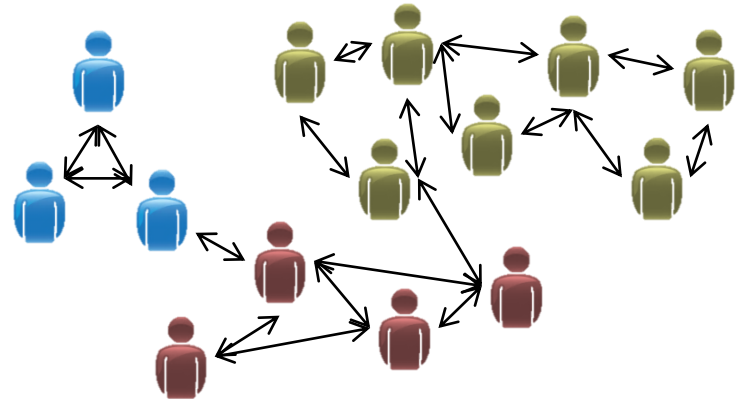
Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$



Applications of clustering



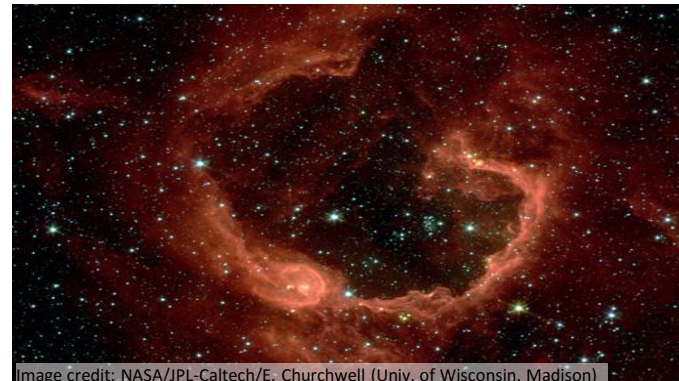
Market segmentation



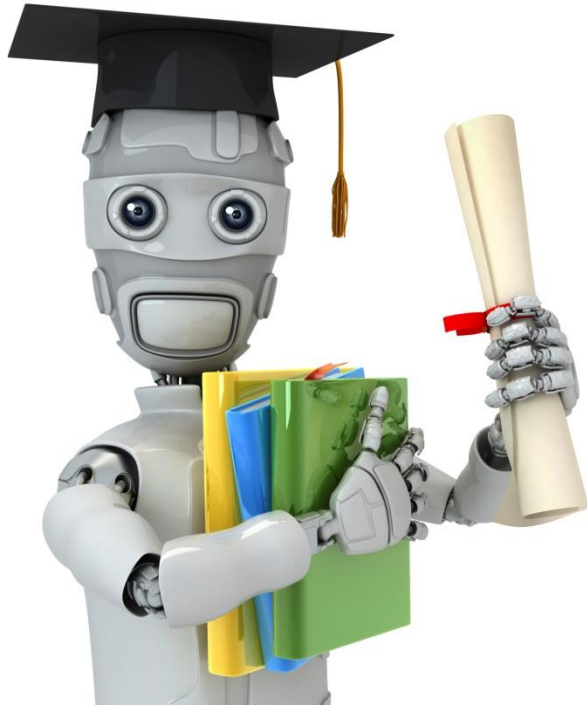
Social network analysis



Organize computing clusters



Astronomical data analysis



Clustering

Application



Google News: automatic clustering gives an effective news presentation metaphor

The screenshot shows the Google News homepage with a browser window. The address bar shows <http://news.google.com/>. The page is organized into sections based on automatic clustering:

- World »** (edit icon)
 - Pirates Demand \$25 Million Ransom for Hijacked Tanker (Update1)**
Bloomberg - 36 minutes ago
By Caroline Alexander and Hamsa Omar Nov. 20 (Bloomberg) -- Somali pirates are demanding \$25 million in ransom to release an oil-laden Saudi supertanker seized off the East African coast, and called on the ship's owners to pay up "soon."
[Somali pirates demand \\$25M for Saudi ship](#) United Press International
[African Union says Somali politicians fuel piracy](#) Washington Post
[BBC News](#) - [guardian.co.uk](#) - [Aljazeera.net](#) - [RIA Novosti](#)
[all 4,015 news articles »](#)
 - Pakistan protests over US missile strikes**
Reuters - 2 hours ago
By Simon Cameron-Moore ISLAMABAD (Reuters) - Pakistan summoned US ambassador Anne Patterson on Thursday to protest over missile strikes launched by pilotless drone aircraft against militant targets in Pakistan.
[Pakistan protests US drone attacks, Taliban warns of reprisals](#) AFP
[Pakistan warns US over missile strike](#) CNN International
[Telegraph.co.uk](#) - [China Daily](#) - [Xinhua](#) - [PRESS TV](#)
[all 560 news articles »](#)
 - Nighttime attack on Thai antigovernment protesters wounds at least 20**
Christian Science Monitor - 30 minutes ago
The government denied attacking demonstrators, who have called for the ouster of the prime minister. By Huma Yusuf One person has been killed and 23 others wounded in a grenade attack Thursday against antigovernment protesters occupying the Thai prime ...
[Blast Kills 1, Wounds 23 at Thai Prime Minister's Office](#) Washington Post
[Anti-government protestor in Thailand dies in grenade attack](#) International Herald Tribune
[Xinhua](#) - [United Press International](#) - [The Associated Press](#) - [AsiaOne](#)
[all 688 news articles »](#)
- U.S. »** (edit icon)
 - Top Court in California Will Review Proposition 8**
New York Times - 1 hour ago
By JESSE MCKINLEY SAN FRANCISCO - Responding to pleas for legal clarity from those on both sides of the issue, the California Supreme Court said Wednesday that it would take up the case of whether a voter-approved ban on same-sex unions was ...
[California Supreme Court to decide fate of Prop. 8 same-sex ...](#)
[San Jose Mercury News](#)
[Prop. 8 gay marriage ban goes to Supreme Court](#) Los Angeles Times
[The Miami Herald](#) - [San Diego Union Tribune](#) - [Indiana Daily Student](#) - [San Francisco Chronicle](#)
[all 1,241 news articles »](#)
 - Drop That Cigarette, Today Is The Great American Smokeout**
dBTechno - 1 hour ago
Washington (dbTechno) - Today marks the annual Great American Smokeout hosted by the American Cancer Society, and is trying to get people all across the US to drop their cigarettes for just one day.
[Great American Smokeout: Time to kick the habit](#) Capital Times
[National Smoke Out Day is Thursday: be a quitter](#) Las Cruces Sun-News
[MPNnow.com](#) - [eMaxHealth.com](#) - [Times Tribune of Corbin](#) - [ABC15.com \(KNXV-TV\)](#)
[all 338 news articles »](#)
 - Perino: Bush would sign jobless benefits extension**
The Associated Press - 47 minutes ago
WASHINGTON (AP) - With weekly jobless claims benefits at a 16-year high, the White House said Thursday that President George W. Bush would quickly sign legislation pending in Congress to provide further unemployment benefits.
[Bush would sign measure to extend jobless benefits](#) Houston Chronicle
[Jobless claims show need for benefits extension: White House](#) AFP
[Washington Times](#) - [Wall Street Journal Blogs](#) - [WOI](#) - [Tampabay.com](#)
[all 599 news articles »](#)

At the bottom of the page, there are links to [Show more stories](#) and [Show fewer stories](#) for each section.

The URL at the bottom of the browser window is <http://www.google.com/hostednews/ap/article/ALeqM5hGjNbXI6O23C8QzqZMY0pGPAik-AD94INLTG1>

聚类分析无处不在

- ▶ 谁经常光顾商店，谁买什么东西，买多少？
 - ▶ 按忠诚卡记录的光临次数、光临时间、性别、年龄、职业、购物种类、金额等变量分类
 - ▶ 这样商店可以....
 - ▶ 识别顾客购买模式（如喜欢一大早来买酸奶和鲜肉，习惯周末时一次性大采购）
 - ▶ 刻画不同的客户群的特征



聚类分析无处不在

- ▶ 谁是银行信用卡的黄金客户？
 - ▶ 利用储蓄额、刷卡消费金额、诚信度等变量对客户分类，找出“黄金客户”！
 - ▶ 这样银行可以.....
 - ▶ 制定更吸引的服务，留住客户！比如：
 - ▶ 一定额度和期限的免息透支服务！
 - ▶ 百盛的贵宾打折卡！
 - ▶ 在他或她生日的时候送上一个小蛋糕！



聚类的应用领域

▶ 经济领域：

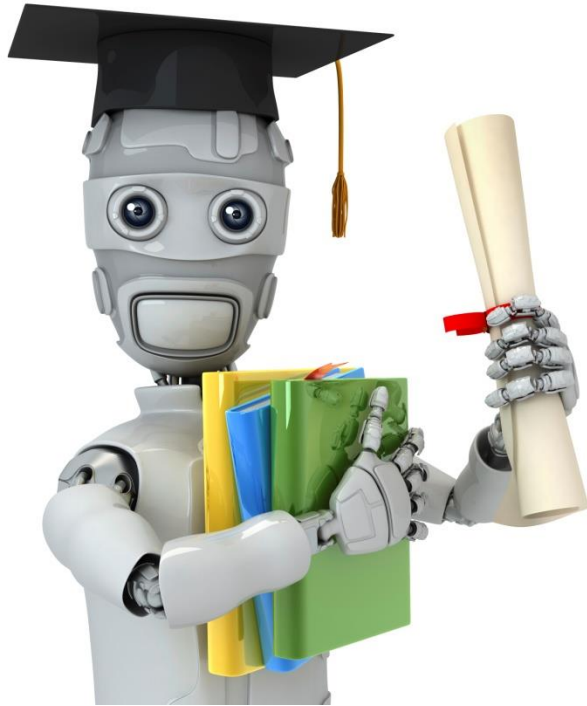
- ▶ 帮助市场分析人员从客户数据库中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。
- ▶ 谁喜欢打国际长途，在什么时间，打到那里？
- ▶ 对住宅区进行聚类，确定自动提款机ATM的安放位置
- ▶ 股票市场板块分析，找出最具活力的板块龙头股
- ▶ 企业信用等级分类

▶ 生物学领域

- ▶ 推导植物和动物的分类；
- ▶ 对基因分类，获得对种群的认识

▶ 数据挖掘领域

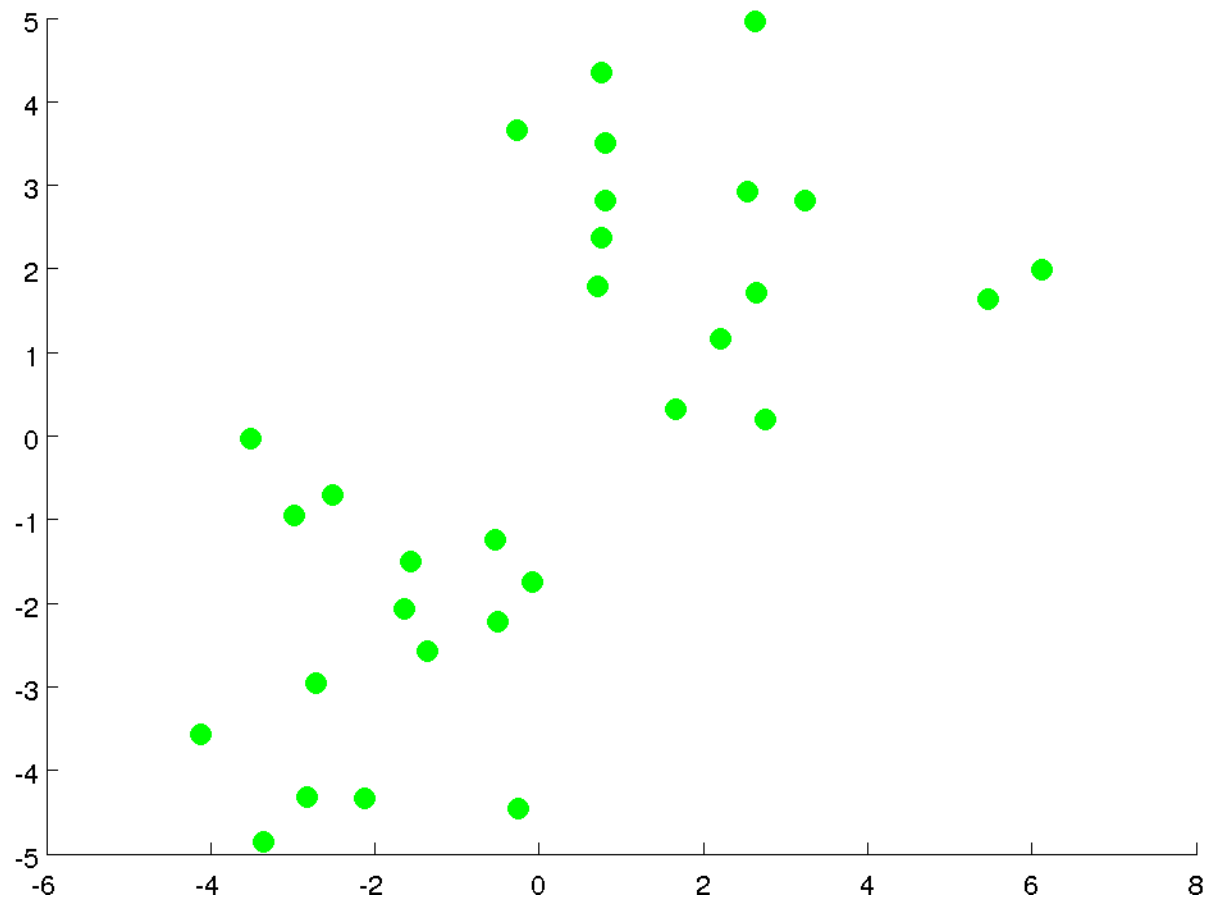
- ▶ 作为其他数学算法的预处理步骤，获得数据分布状况，集中对特定的类做进一步的研究

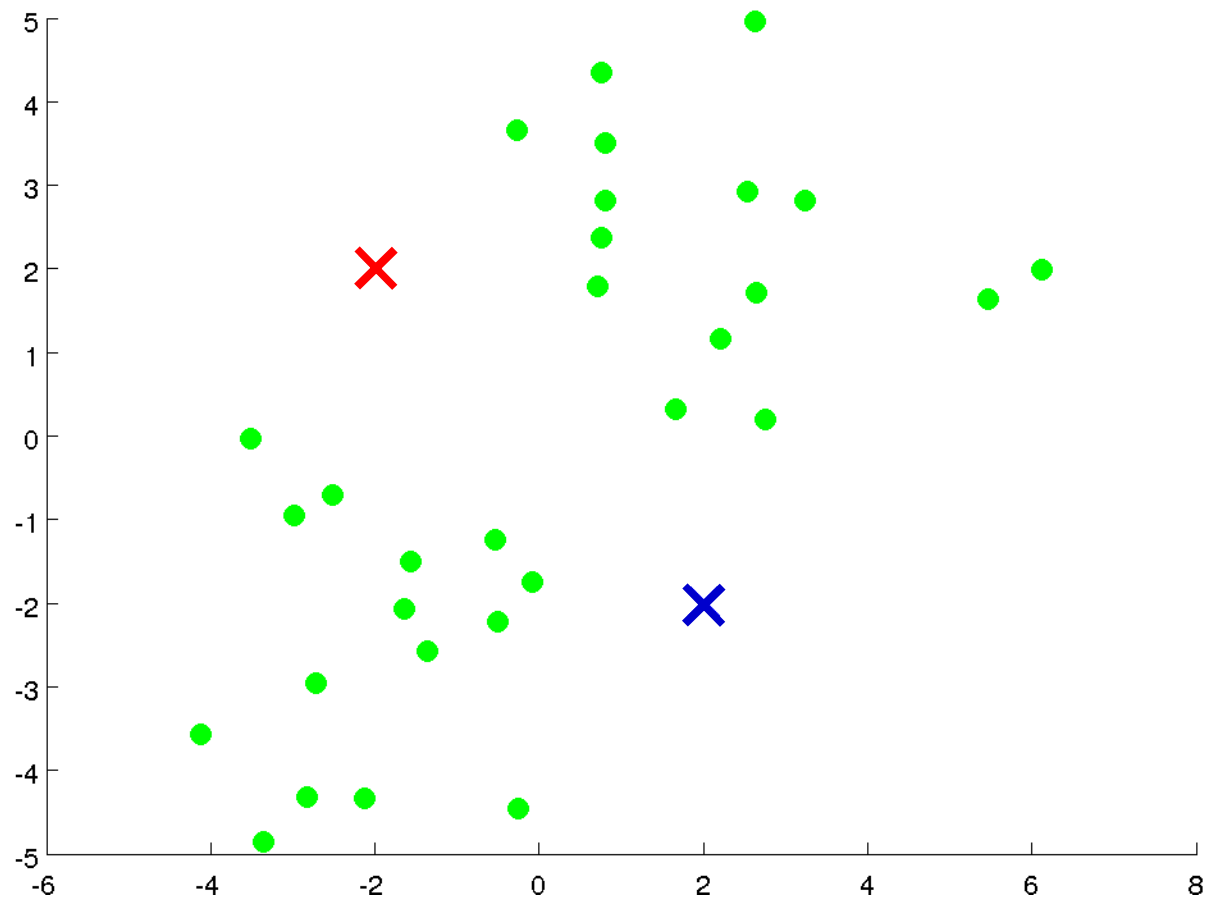


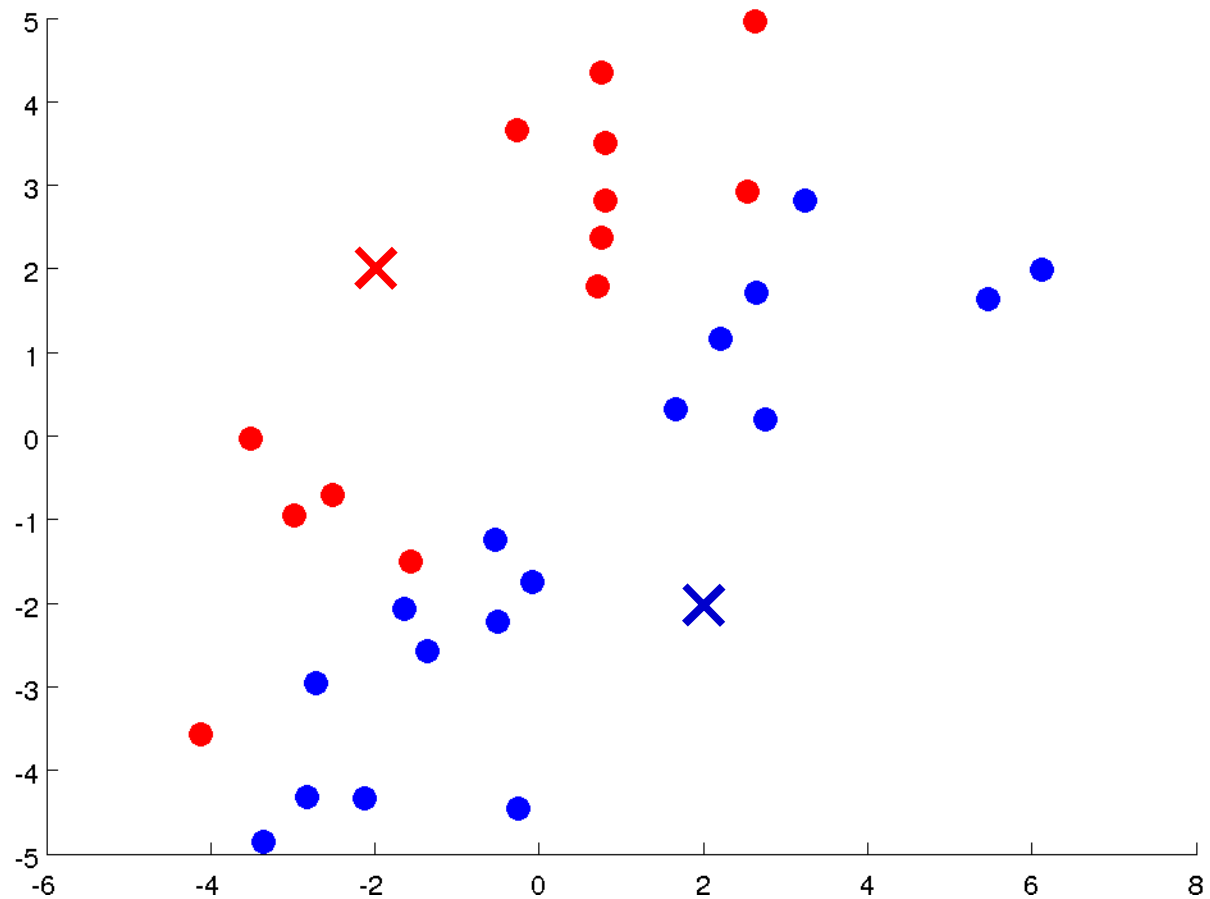
Clustering

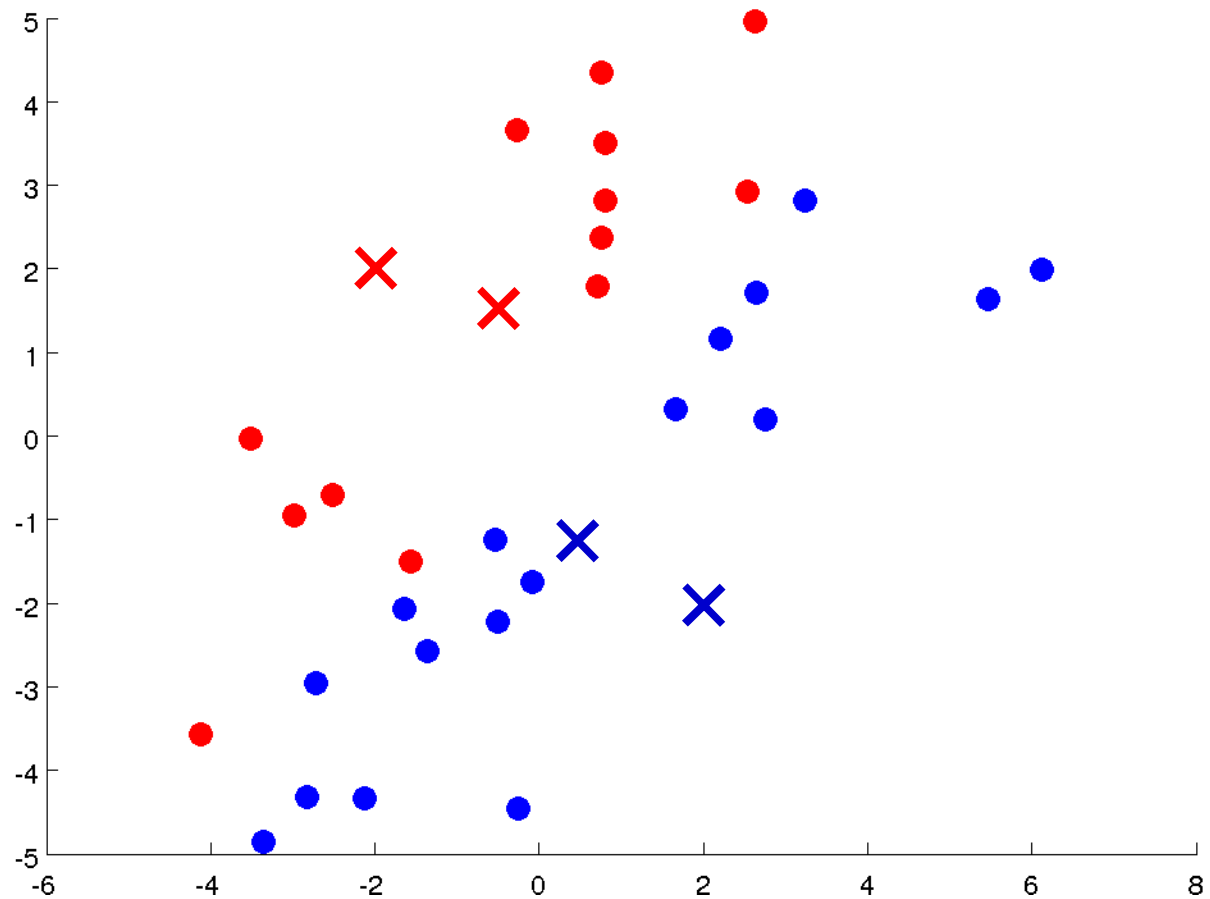
K-means algorithm

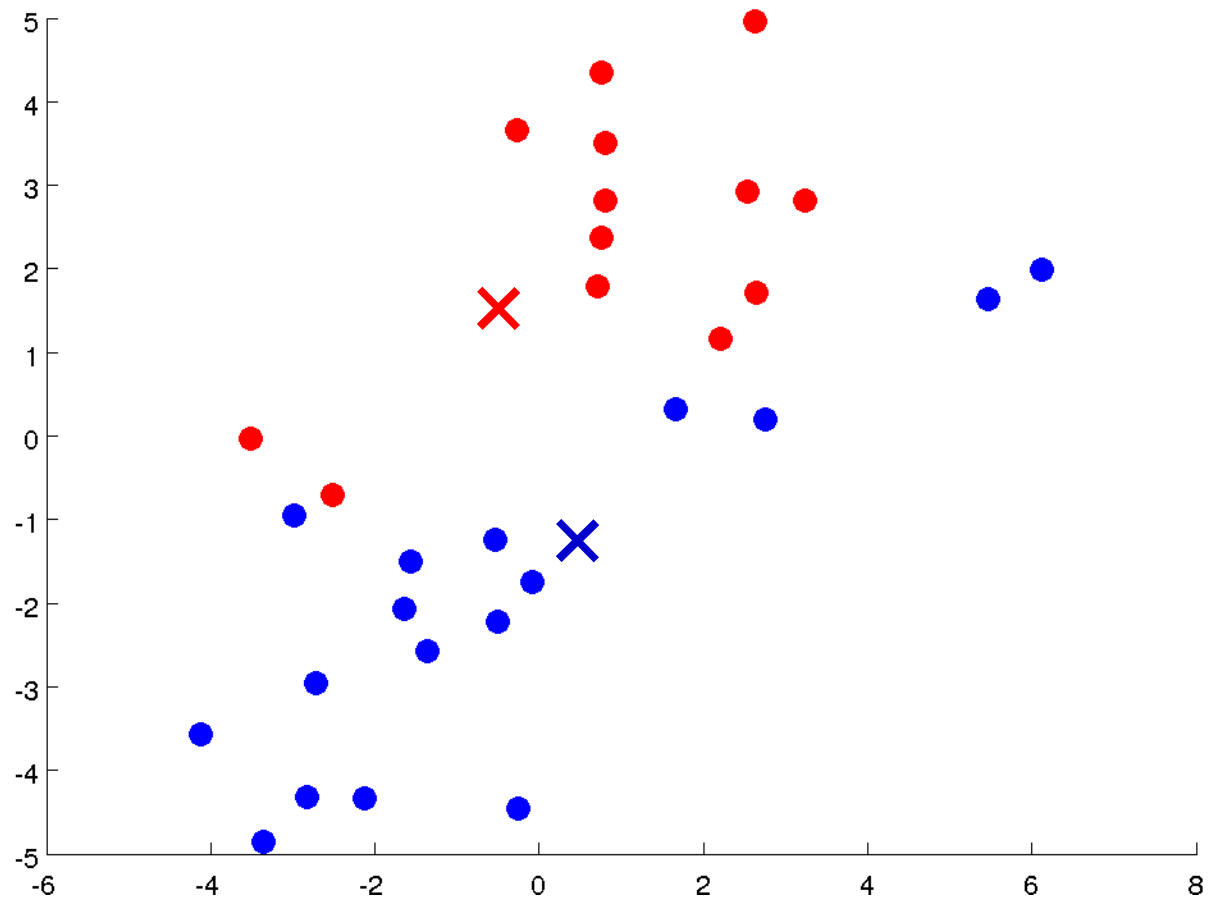


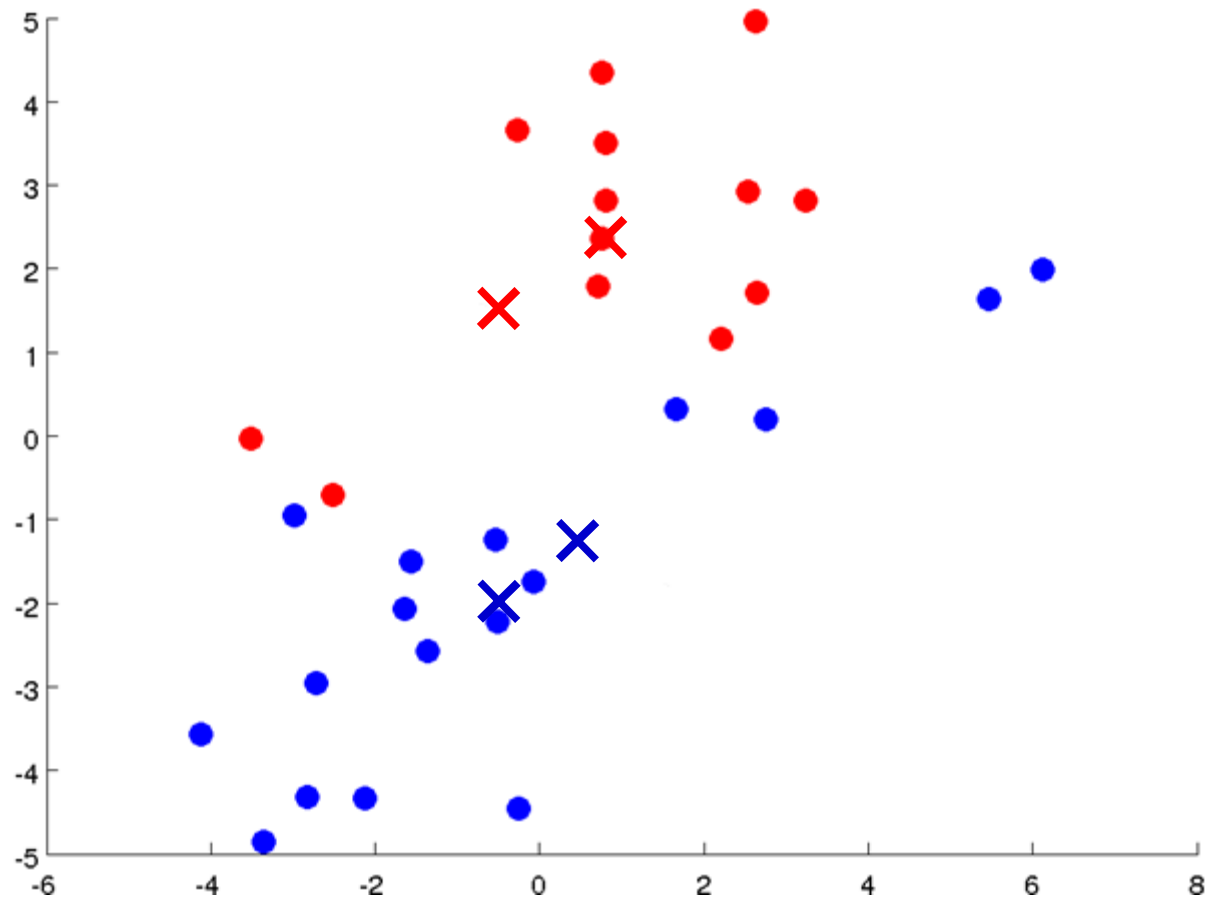


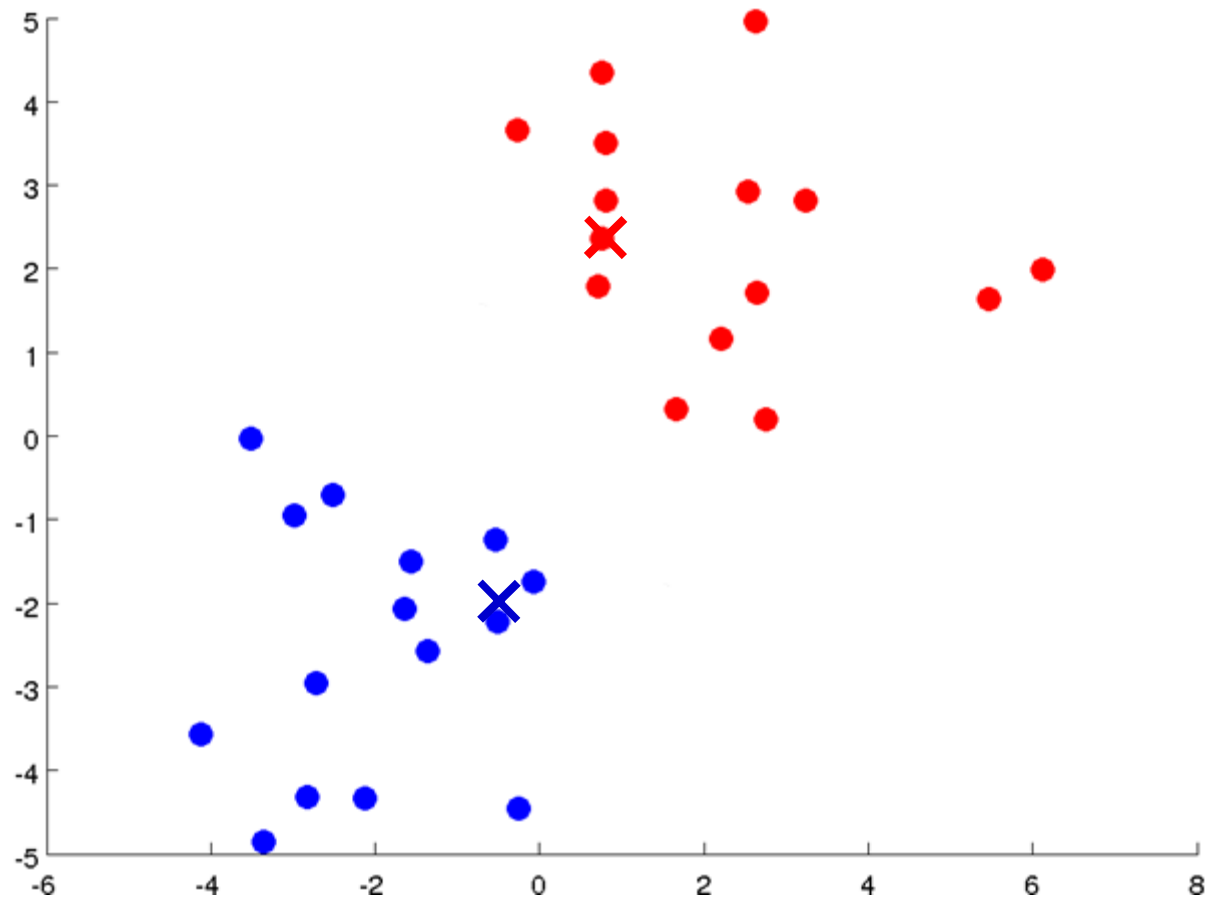


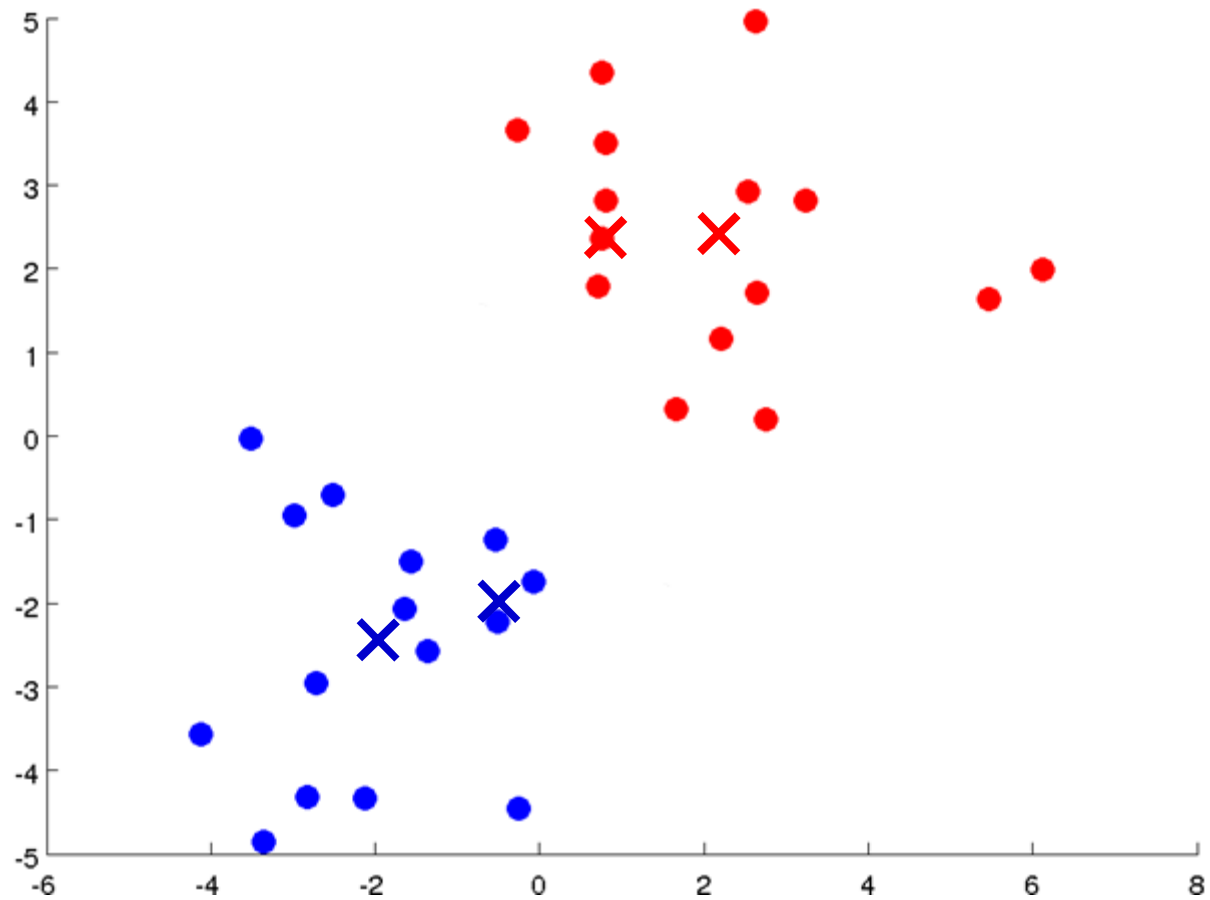


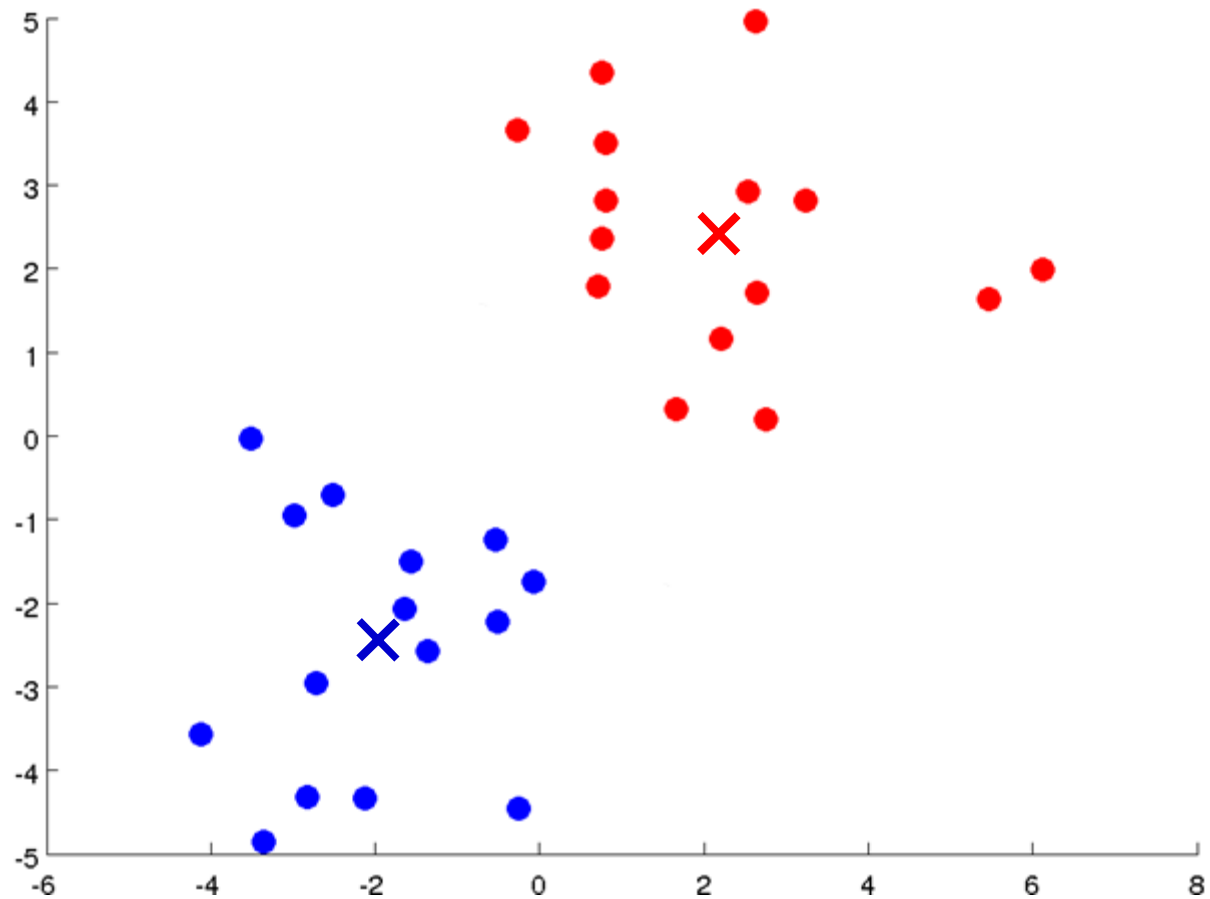












K-means algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

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



K-means algorithm

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

Repeat {

 for $i = 1$ to m

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

 for $k = 1$ to K

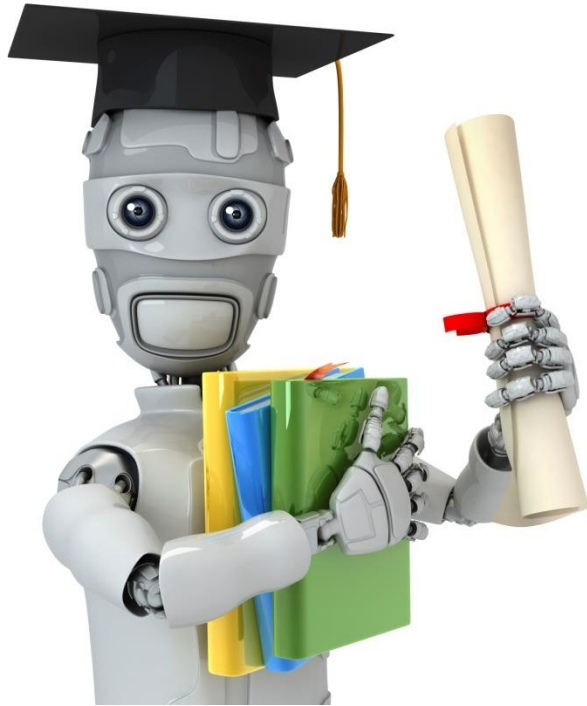
$\mu_k :=$ average (mean) of points assigned to cluster k

}



K-means for non-separated clusters





Clustering Optimization objective



K-means optimization objective

$c^{(i)}$ = index of cluster $(1, 2, \dots, K)$ 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:

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

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

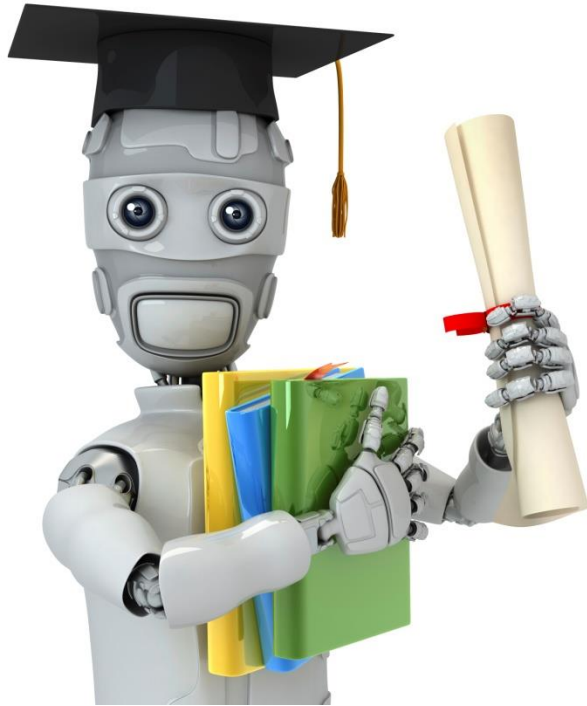


K-means algorithm

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

Repeat {
 for $i = 1$ to m
 $c^{(i)} :=$ index (from 1 to K) of cluster centroid
 closest to $x^{(i)}$
 for $k = 1$ to K
 $\mu_k :=$ average (mean) of points assigned to cluster k
}





Clustering

Random initialization



K-means algorithm

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

Repeat {
 for $i = 1$ to m
 $c^{(i)} :=$ index (from 1 to K) of cluster centroid
 closest to $x^{(i)}$
 for $k = 1$ to K
 $\mu_k :=$ average (mean) of points assigned to cluster k
}

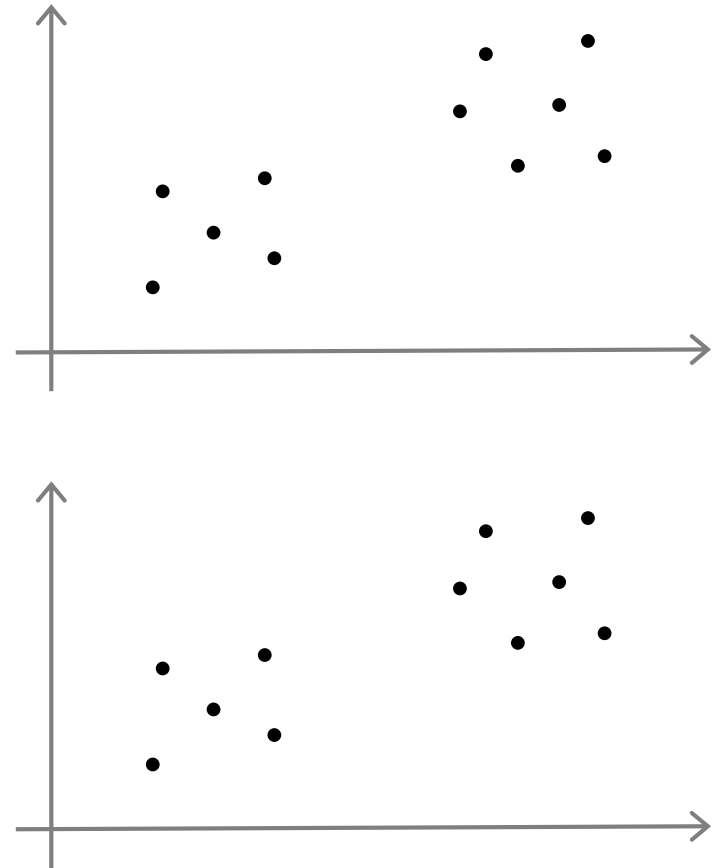


Random initialization

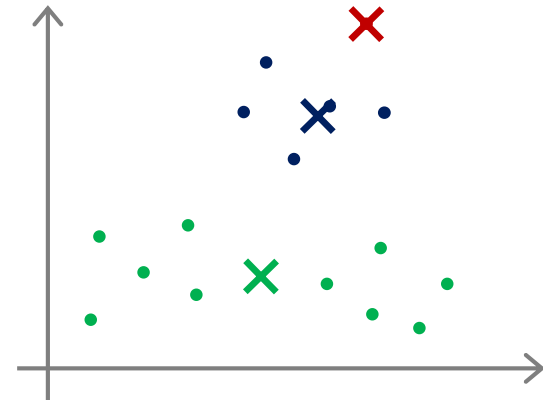
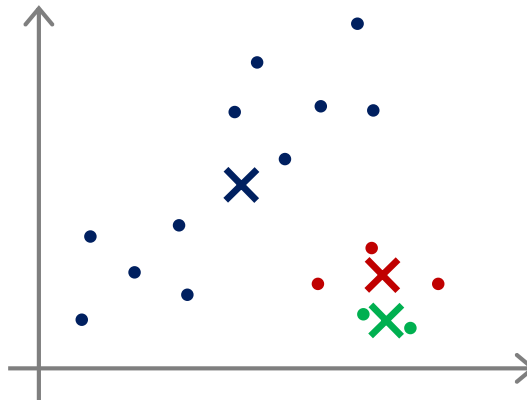
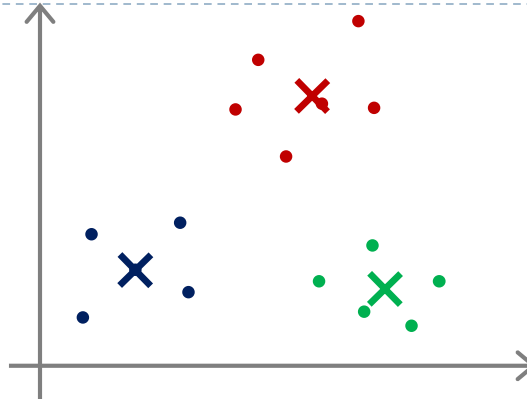
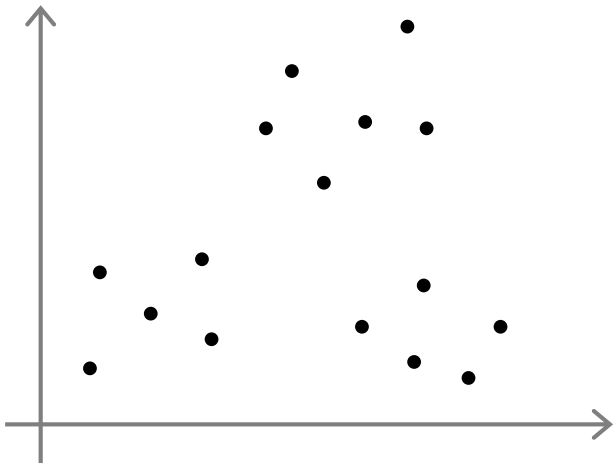
Should have $K < m$

Randomly pick K training examples.

Set μ_1, \dots, μ_K equal to these K examples.



Local optima



Random initialization

For $i = 1$ to 100 {

 Randomly initialize K-means.

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

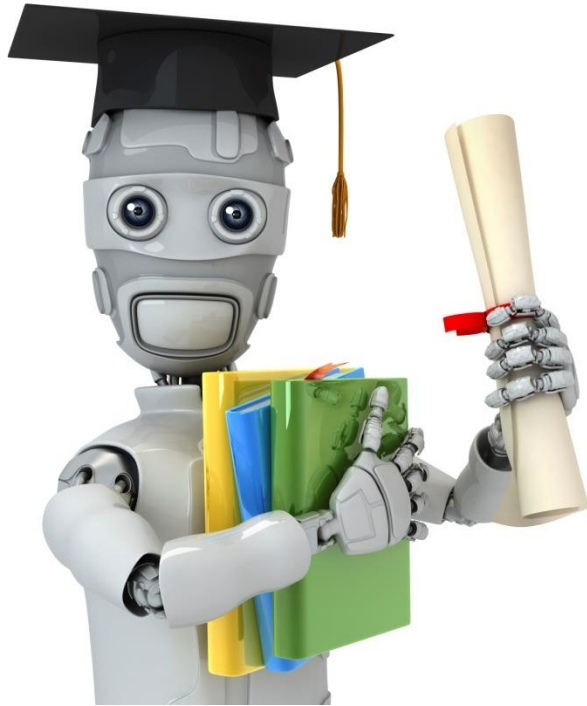
 Compute cost function (distortion)

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

}

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



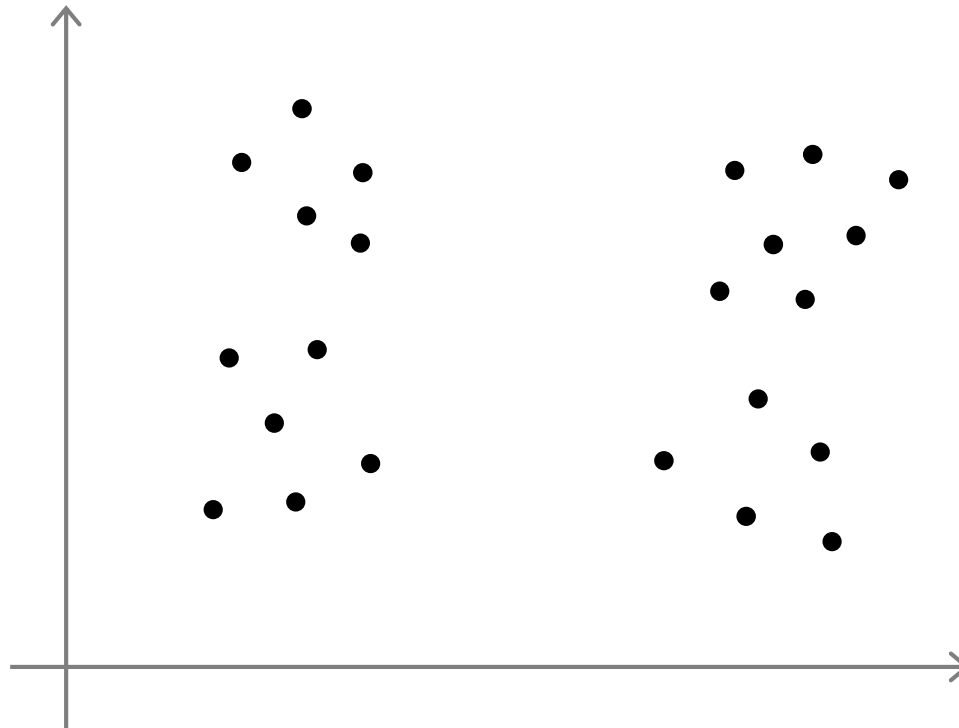


Clustering

Choosing the
number of clusters

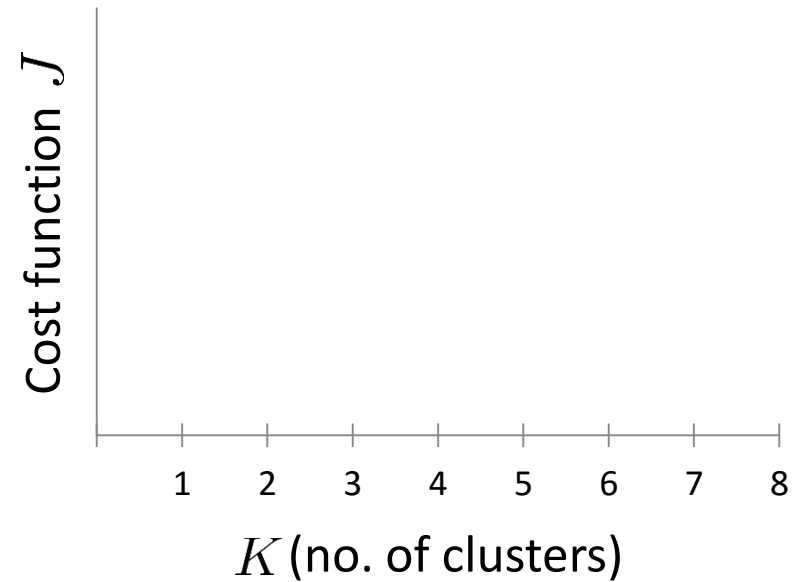
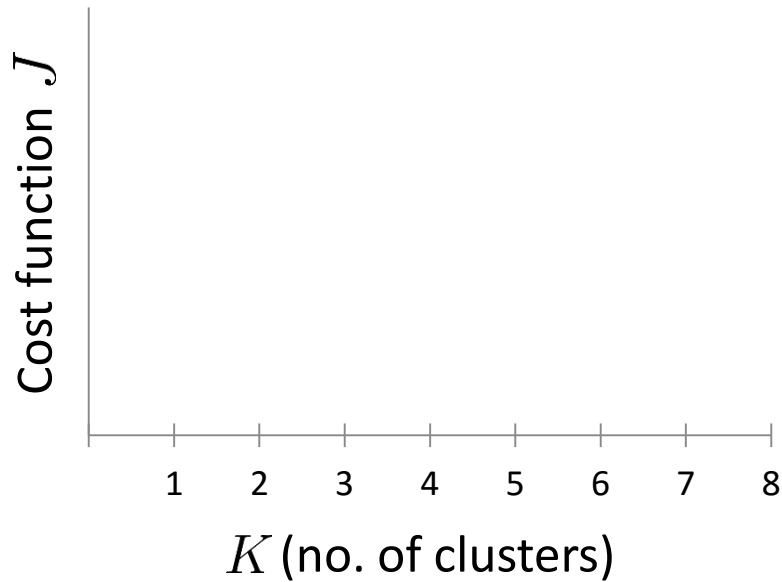


What is the right value of K?



Choosing the value of K

Elbow method:



Choosing the value of K

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

E.g.

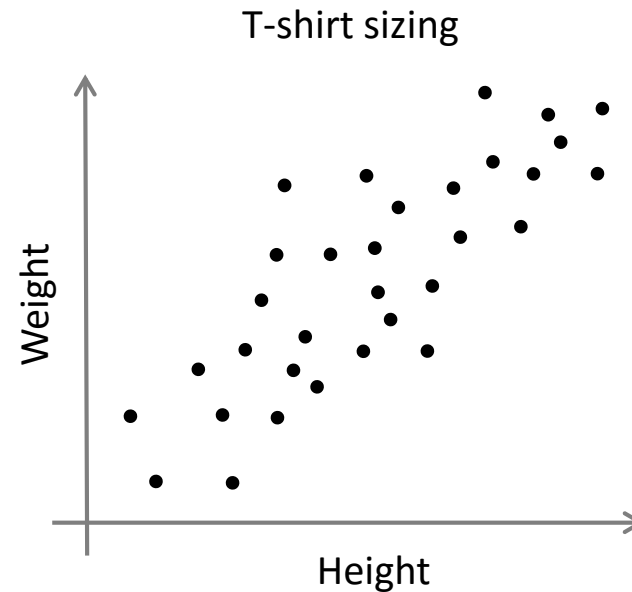
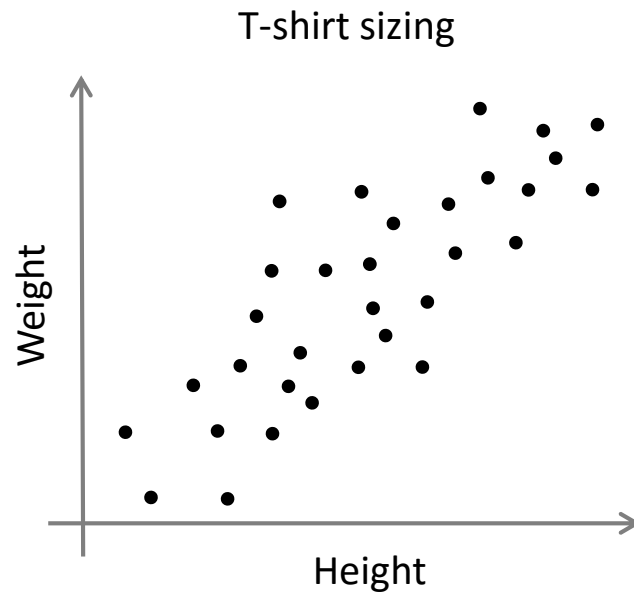


Image Segmentation by K-means



Original



K = 2



K = 8



K = 11



K = 14



K = 15

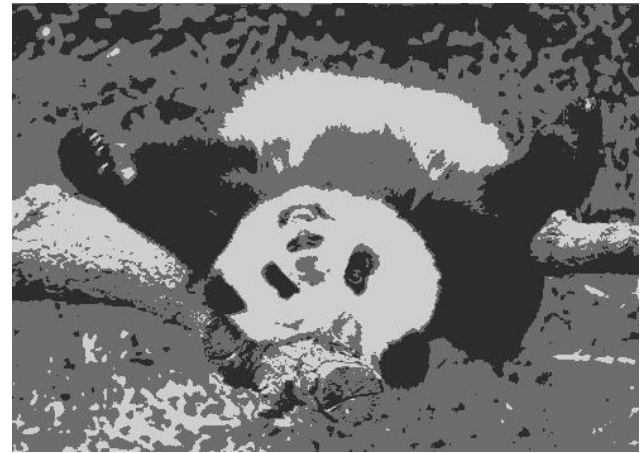
Segmentation as Clustering



Original image



2 clusters



3 clusters

Feature Space

- ▶ Depending on what we choose as the *feature space*, we can group pixels in different ways.
- ▶ Grouping pixels based on **intensity** similarity

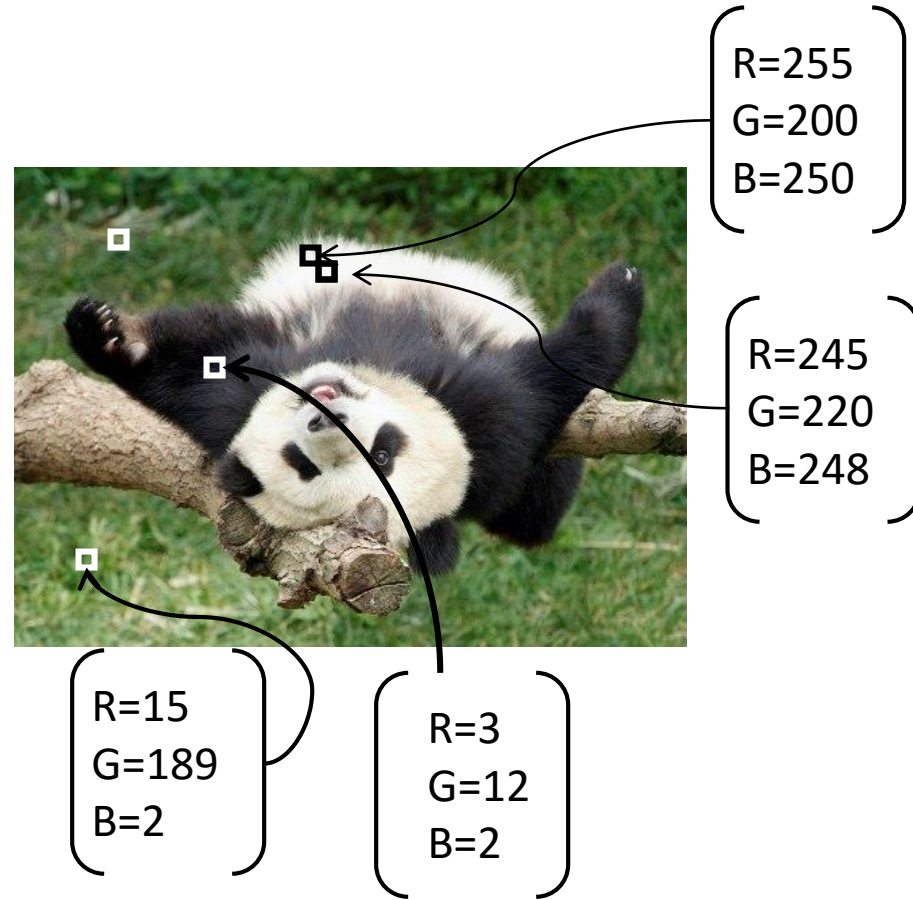
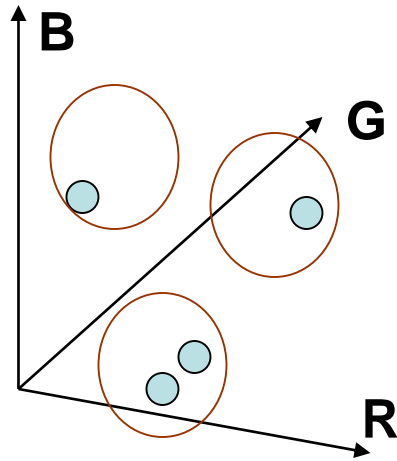


- ▶ Feature space: intensity value (1D)

Slide credit: Kristen Grauman

Feature Space

- ▶ Depending on what we choose as the *feature space*, we can group pixels in different ways.
- ▶ Grouping pixels based on **color** similarity

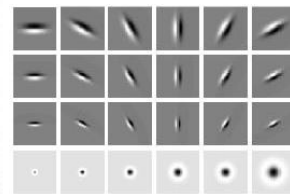
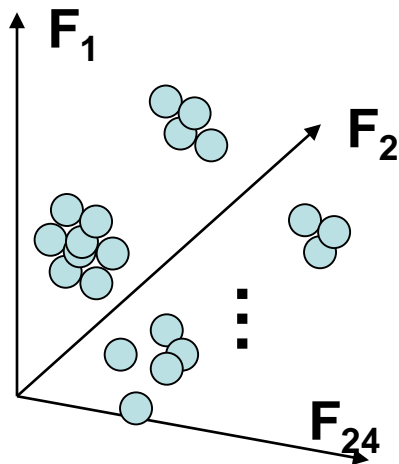


- ▶ Feature space: color value (3D)

Slide credit: Kristen Grauman

Feature Space

- ▶ Depending on what we choose as the *feature space*, we can group pixels in different ways.
- ▶ Grouping pixels based on **texture** similarity



Filter bank of
24 filters

- ▶ Feature space: filter bank responses (e.g., 24D)

K-means Clustering—特点

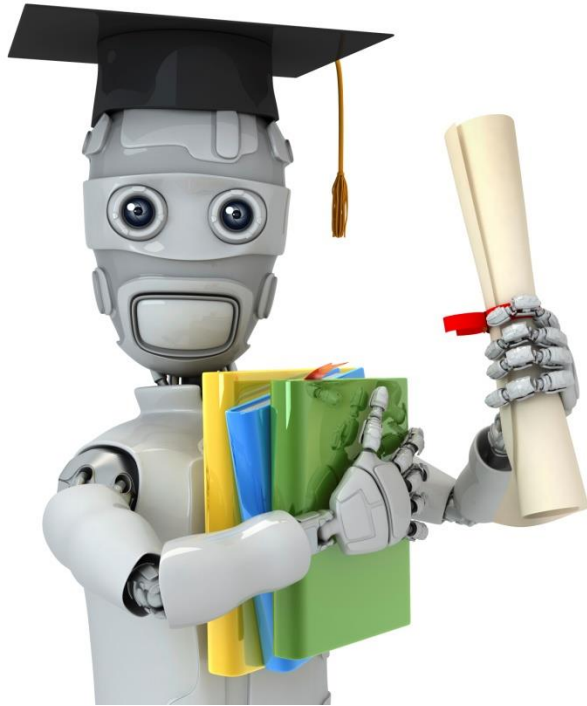
▶ 优点:

- ▶ 当类密集, 且类与类之间区别明显 (比如球型聚集) 时, 聚类效果很好;
- ▶ 强的一致性
- ▶ 算法的复杂度是 $O(Nmt)$ (t 为迭代次数), 对处理大数据集是高效的。

▶ 缺点:

- ▶ 结果与初始质心有关;
- ▶ 必须预先给出聚类的类别数 m ;
- ▶ 对“噪声”和孤立点数据敏感, 少量的这些数据对平均值产生较大的影响;
- ▶ 不适合发现非凸面形状的聚类





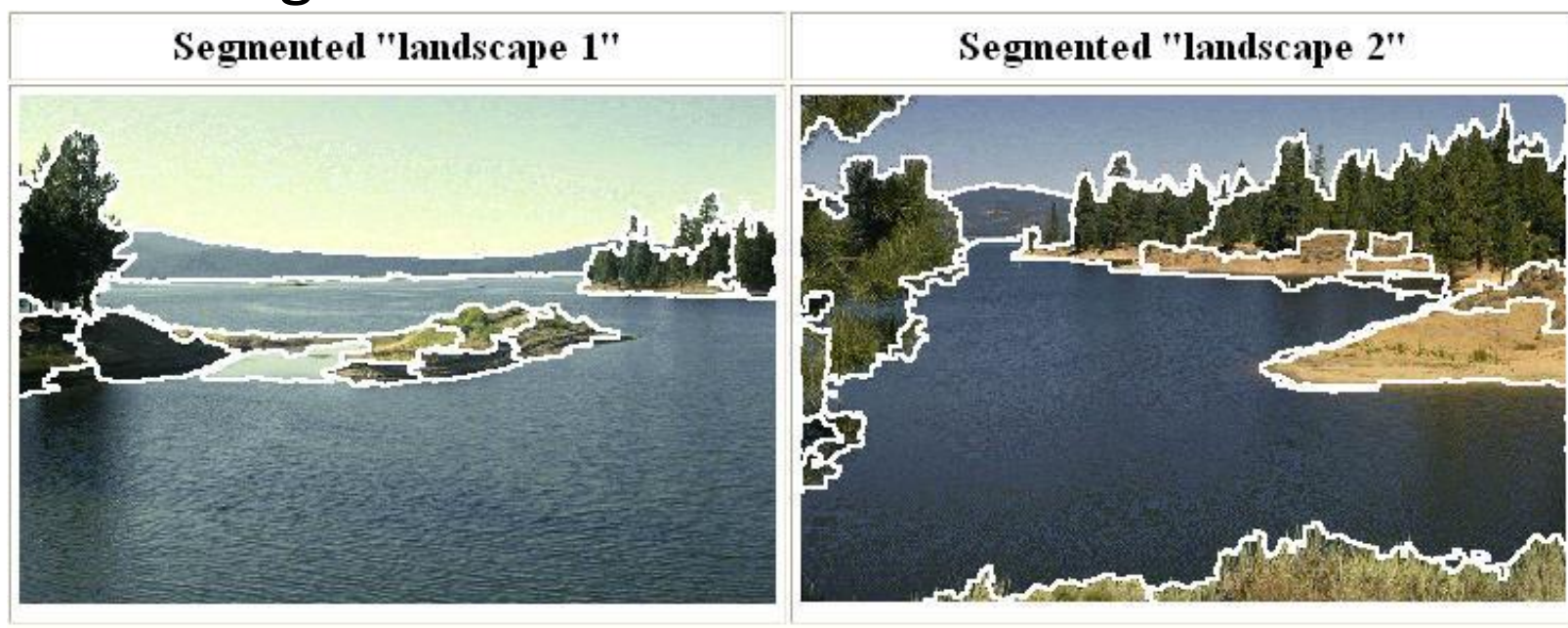
Clustering

Mean-shift



Mean-Shift Segmentation

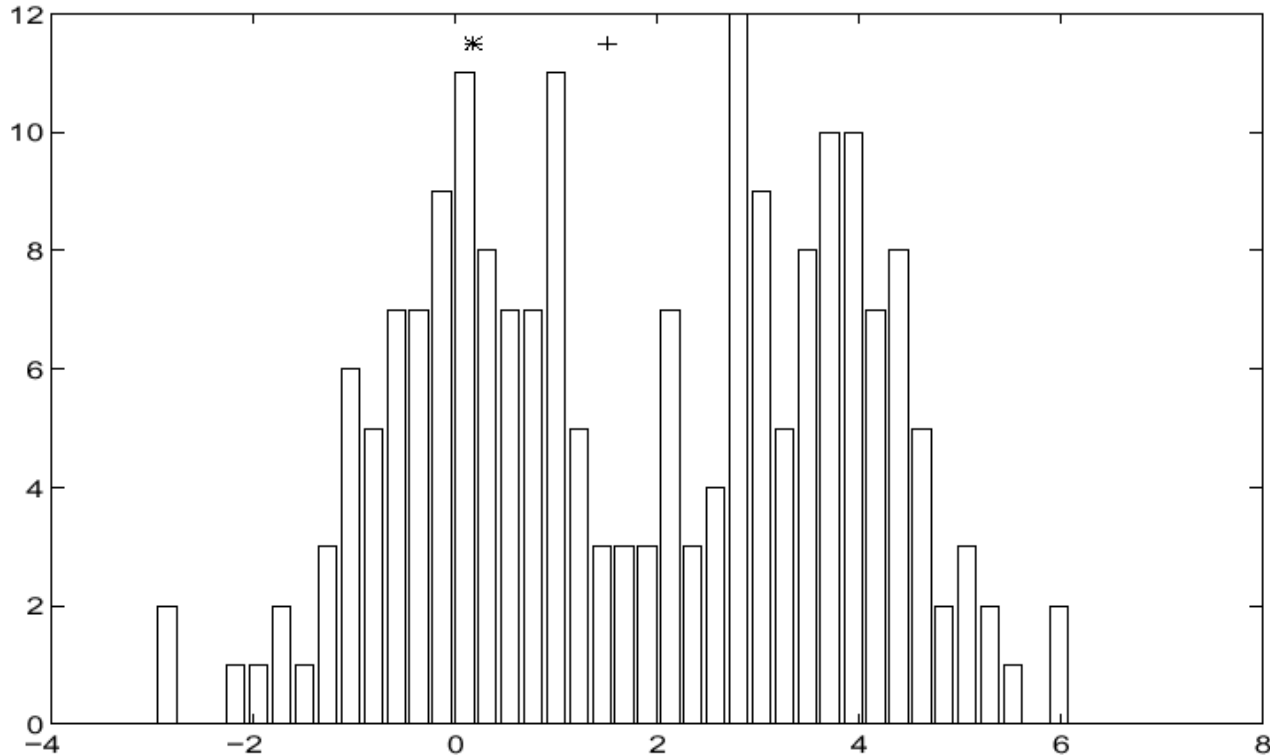
- ▶ An advanced and versatile technique for clustering-based segmentation



<http://www.caip.rutgers.edu/~comanici/MSPAMI/msPamiResults.html>

D. Comaniciu and P. Meer, [Mean Shift: A Robust Approach toward Feature Space Analysis](#), PAMI 2002.

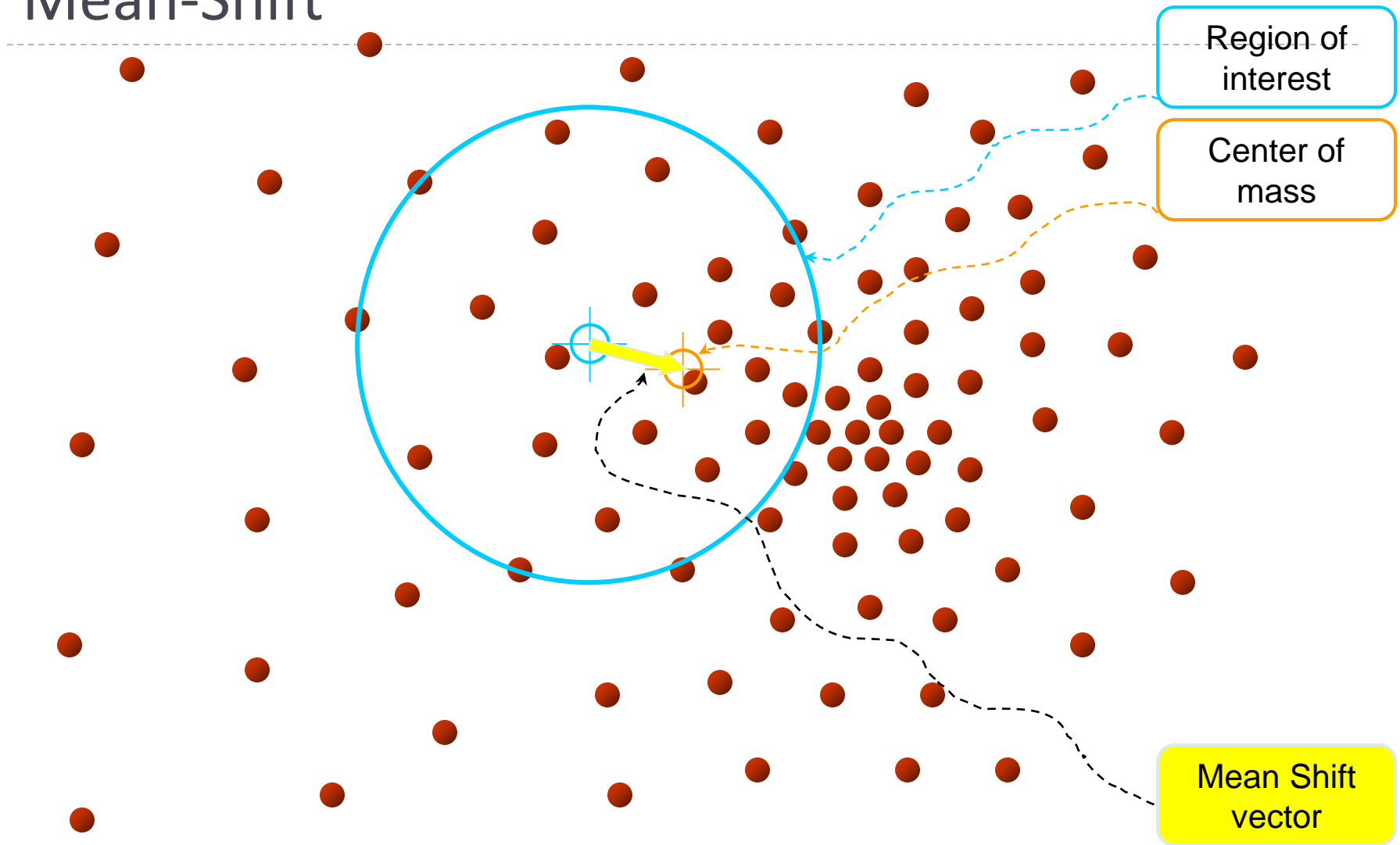
Mean Shift Algorithm



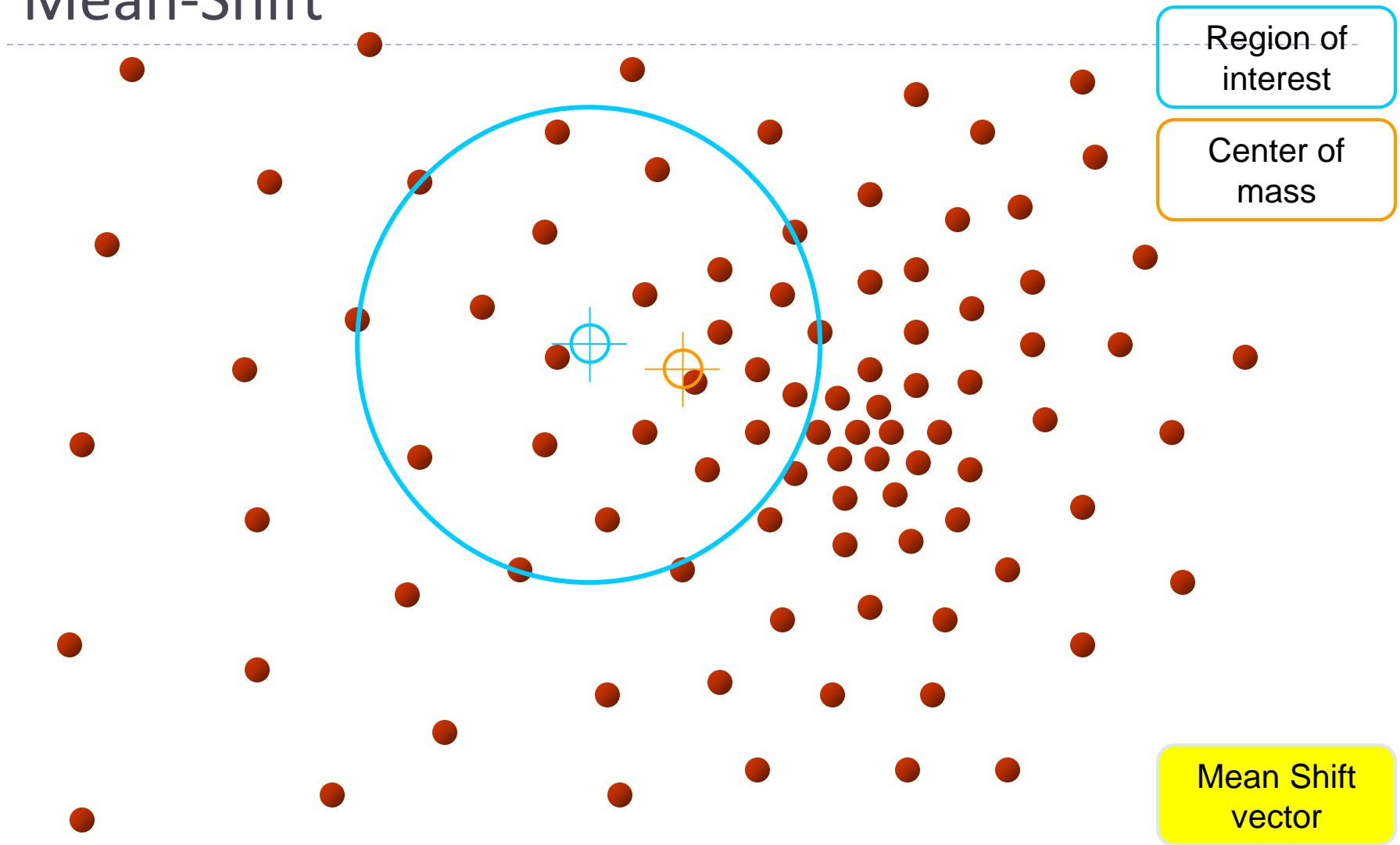
► Iterative Mode Search

1. Initialize random seed, and window W
2. Calculate center of gravity (the “mean” $\sum_{x \in W} xH(x)$)
3. Shift the search window to the mean
4. Repeat Step 2 until convergence

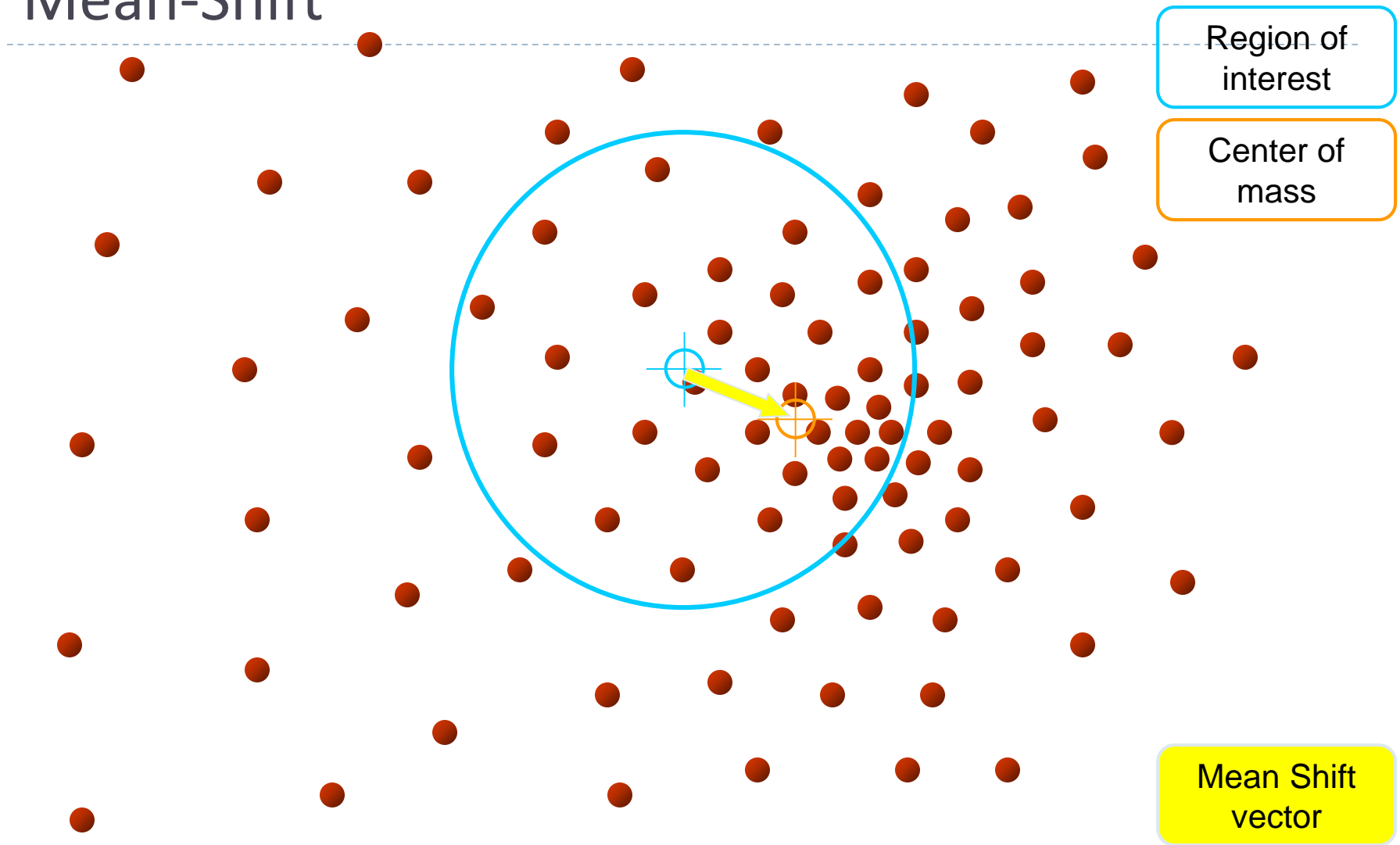
Mean-Shift



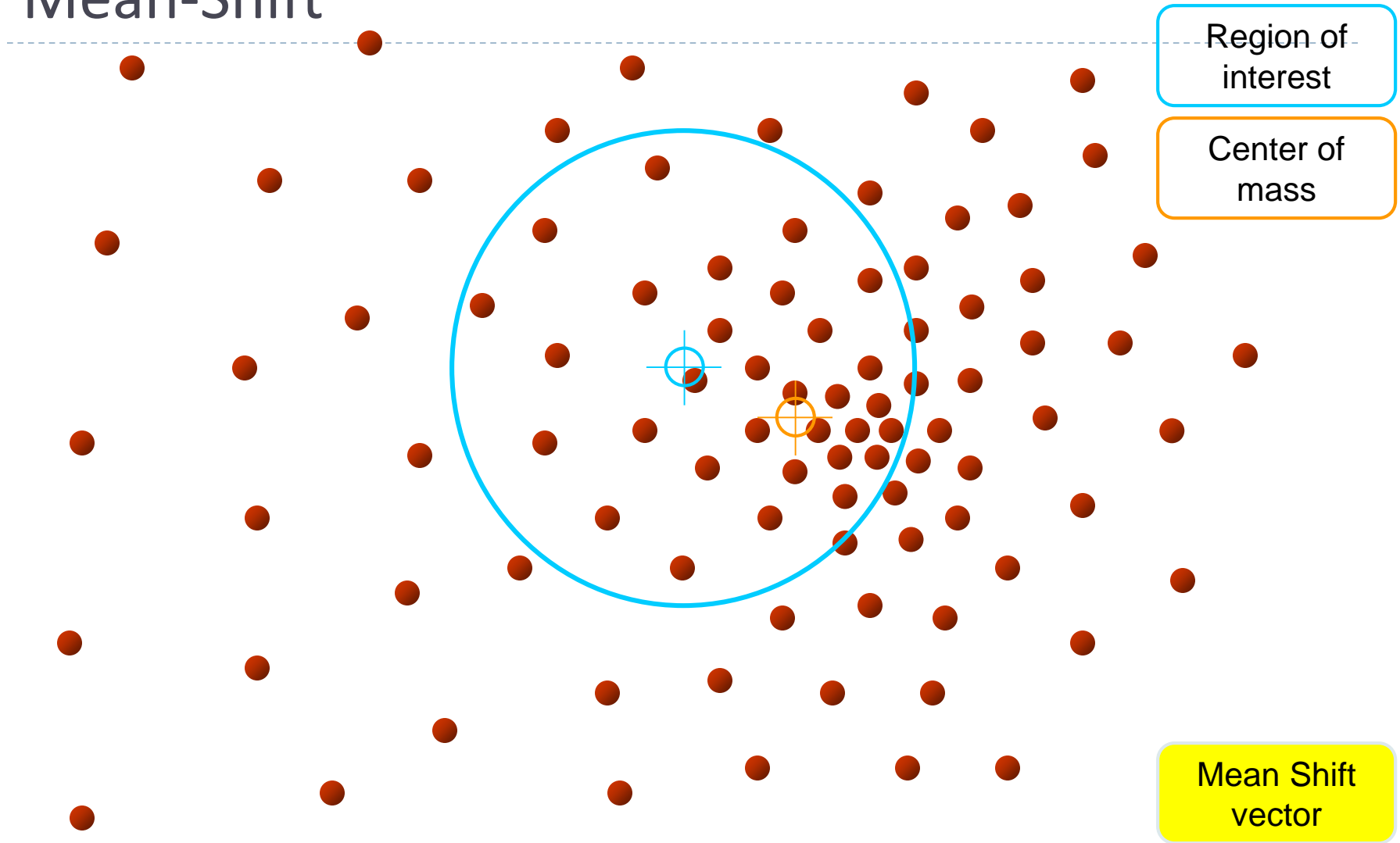
Mean-Shift



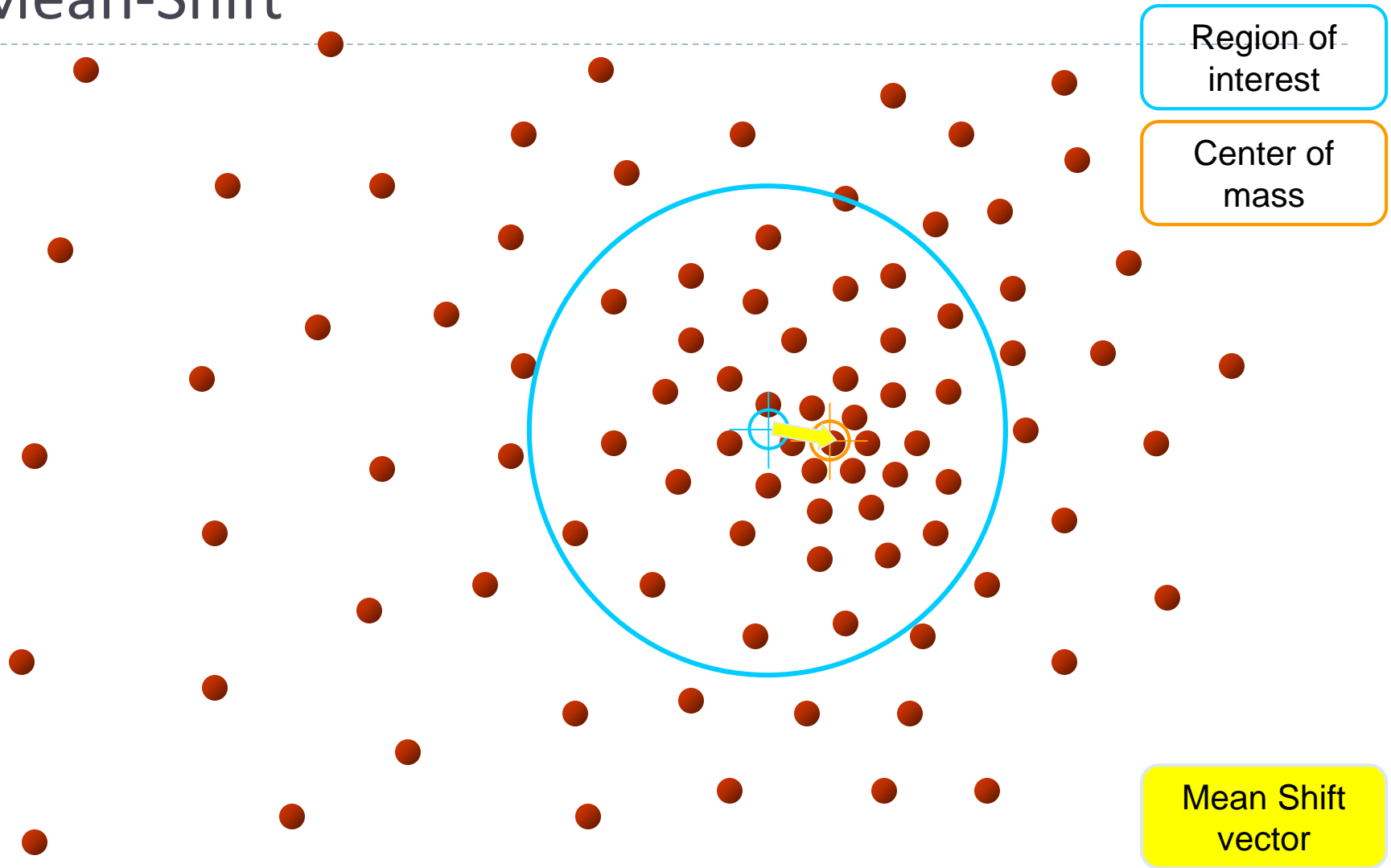
Mean-Shift



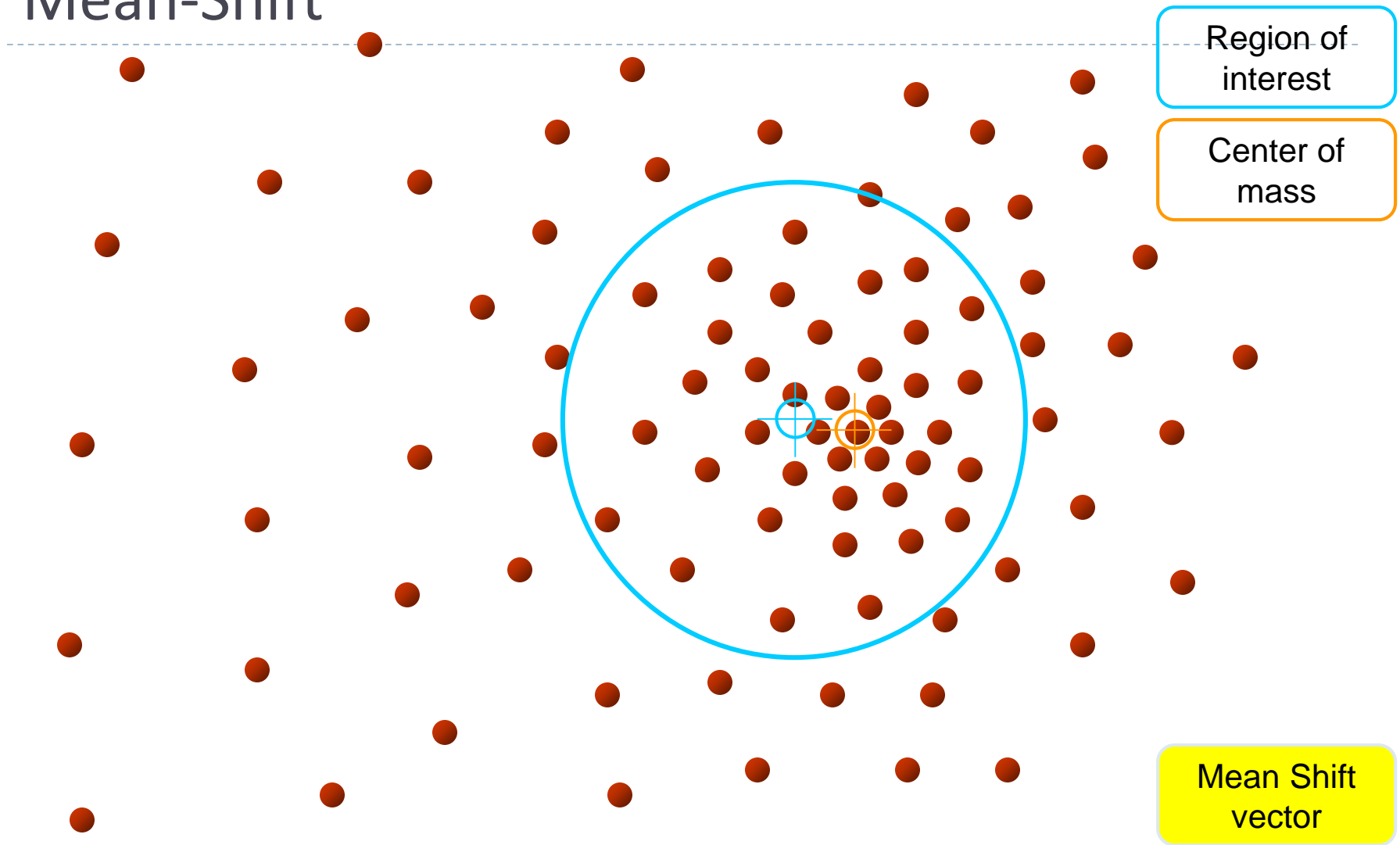
Mean-Shift



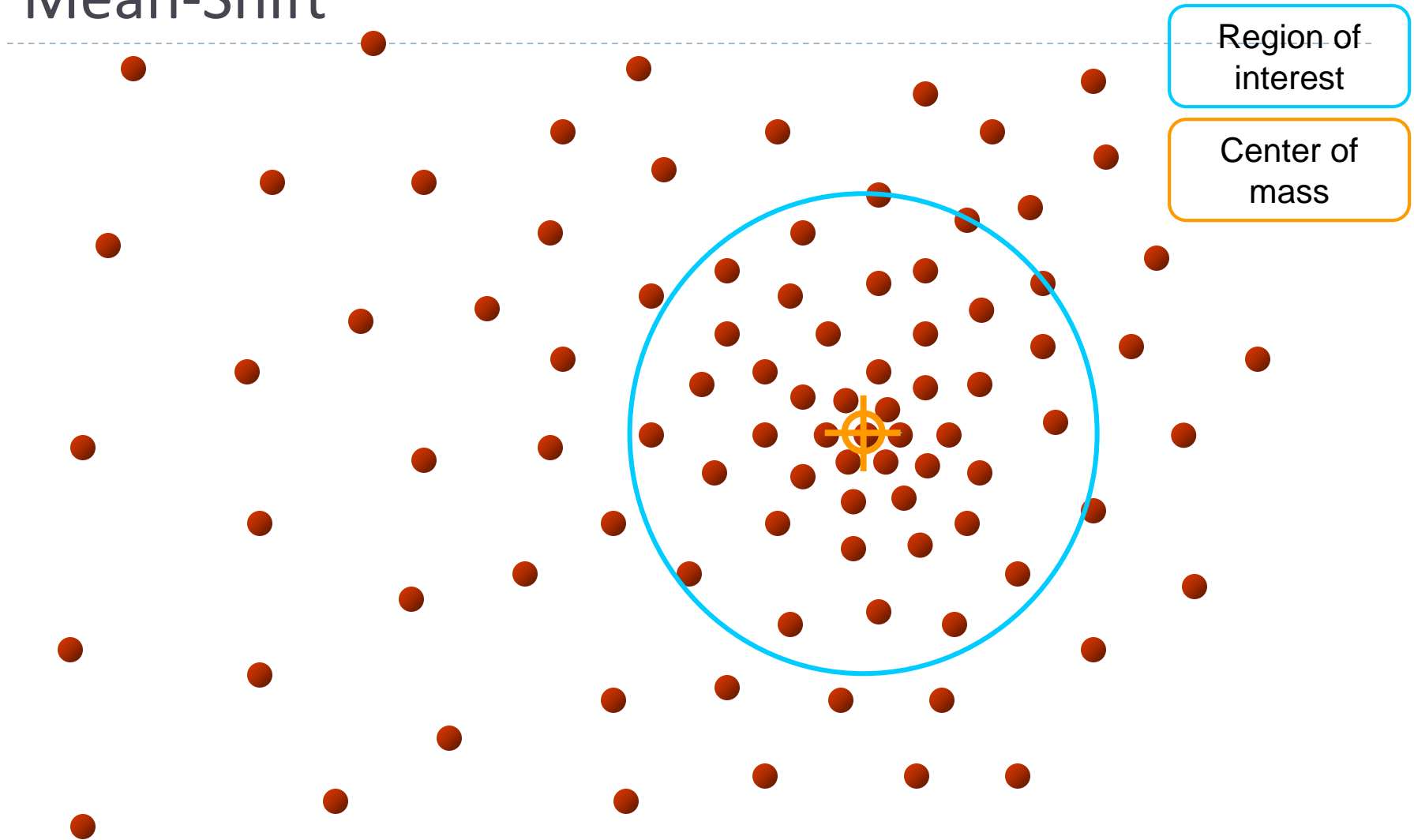
Mean-Shift



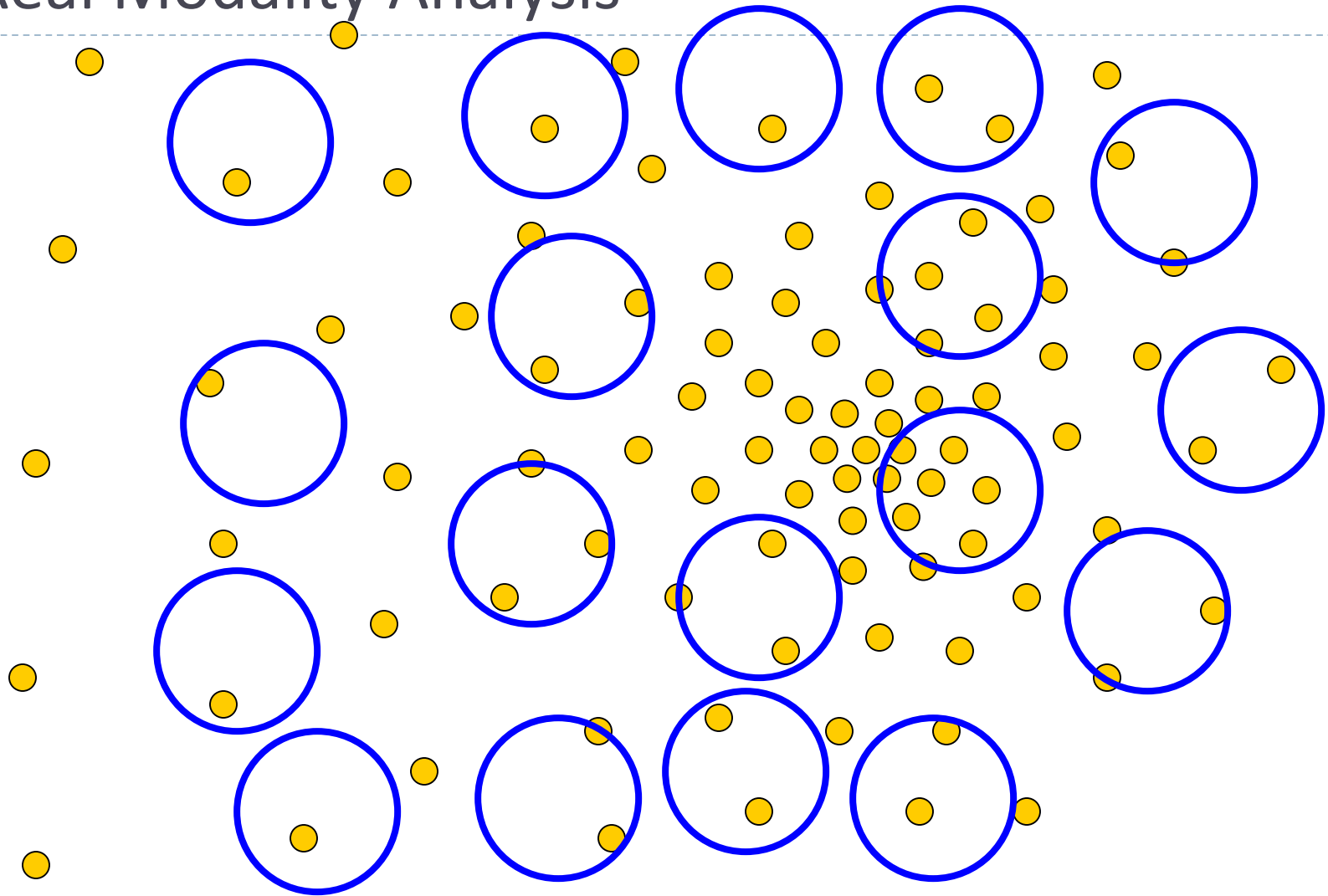
Mean-Shift



Mean-Shift



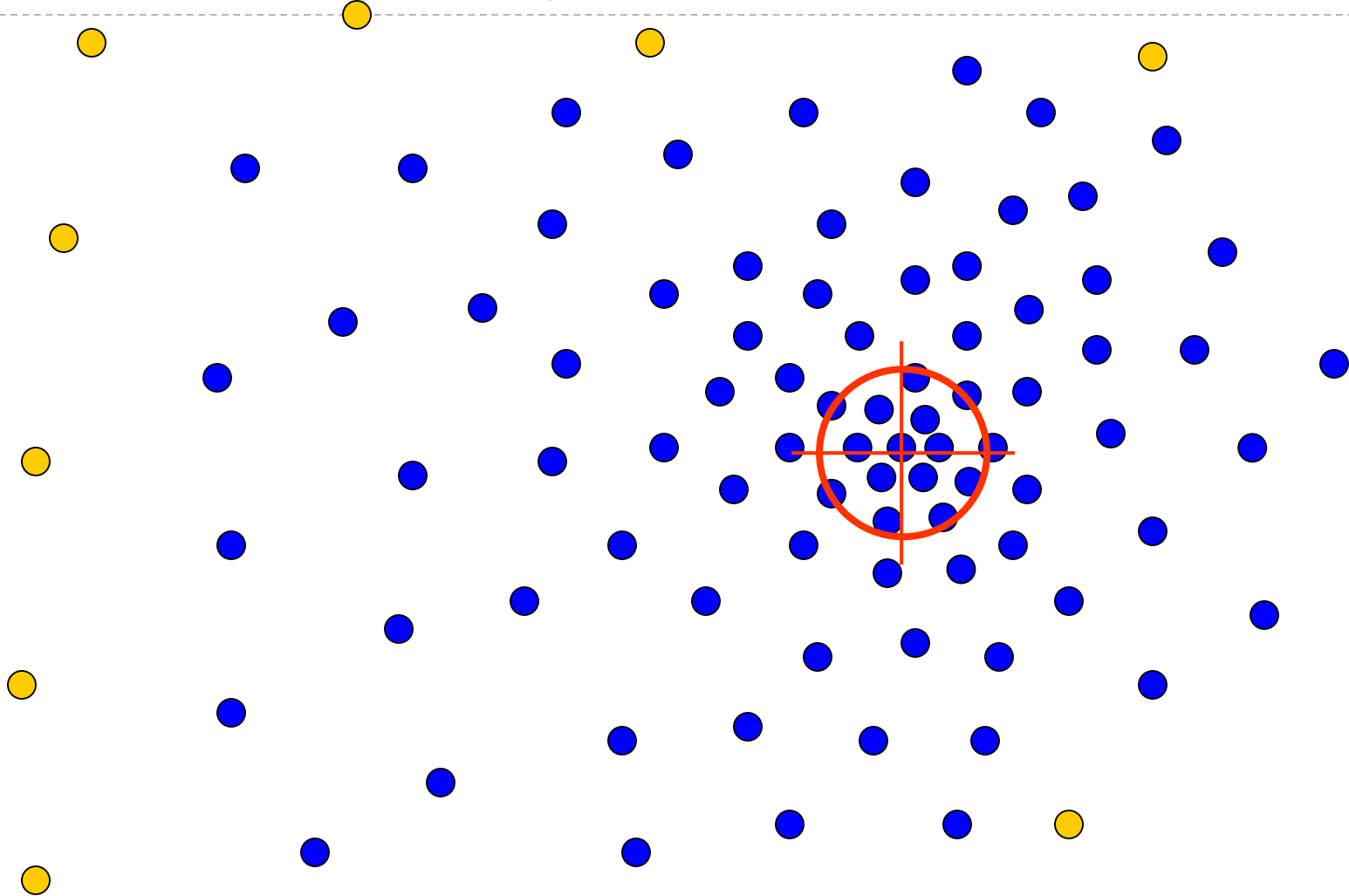
Real Modality Analysis



Tessellate the space with windows

Run the procedure in parallel

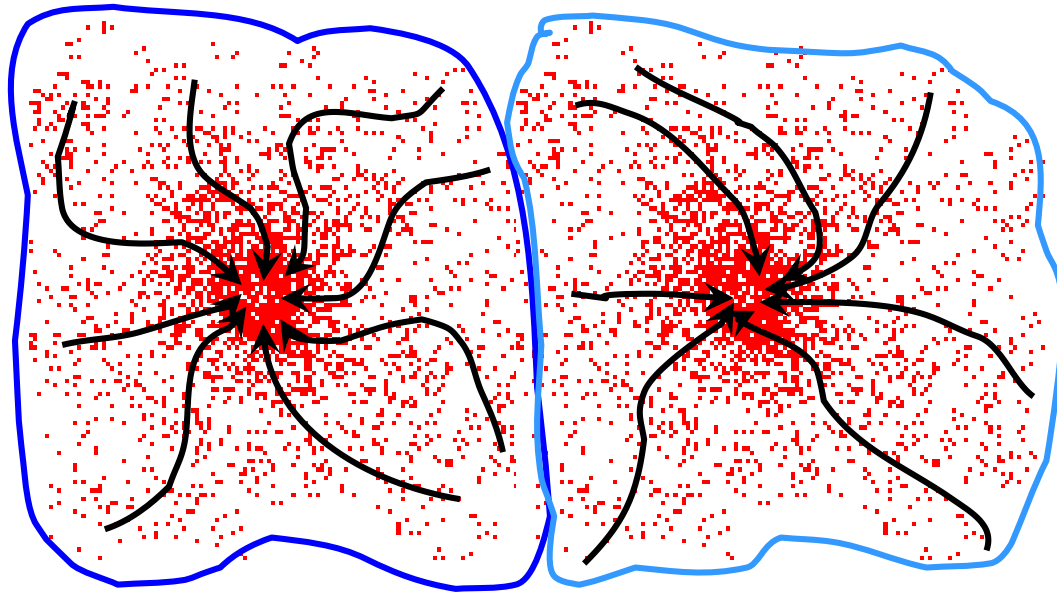
Real Modality Analysis



The **blue** data points were traversed by the windows towards the mode.

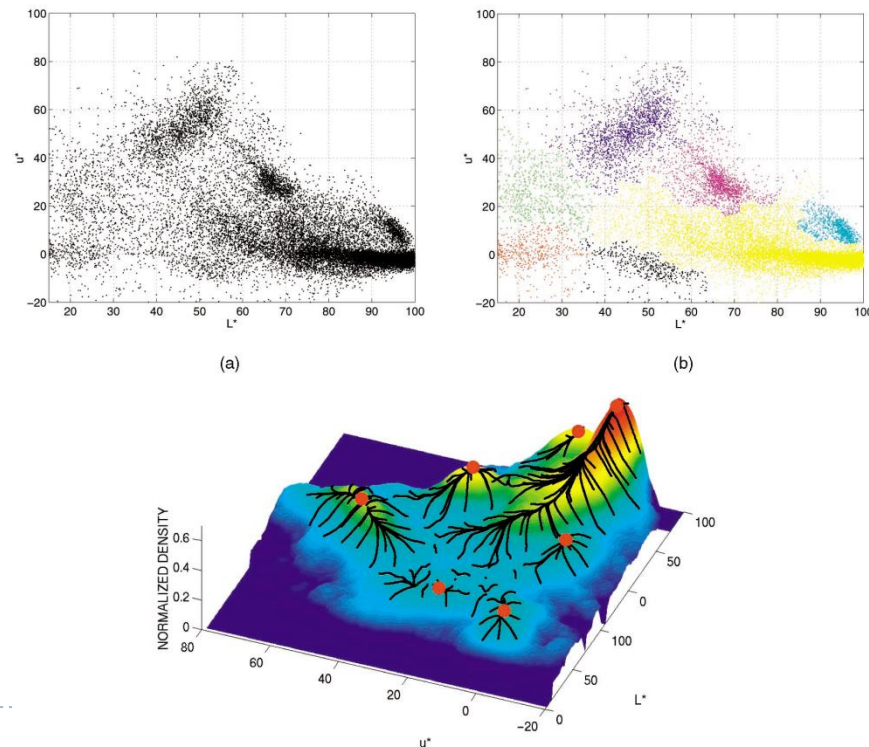
Mean-Shift Clustering

- ▶ Cluster: all data points in the attraction basin of a mode
- ▶ Attraction basin: the region for which all trajectories lead to the same mode

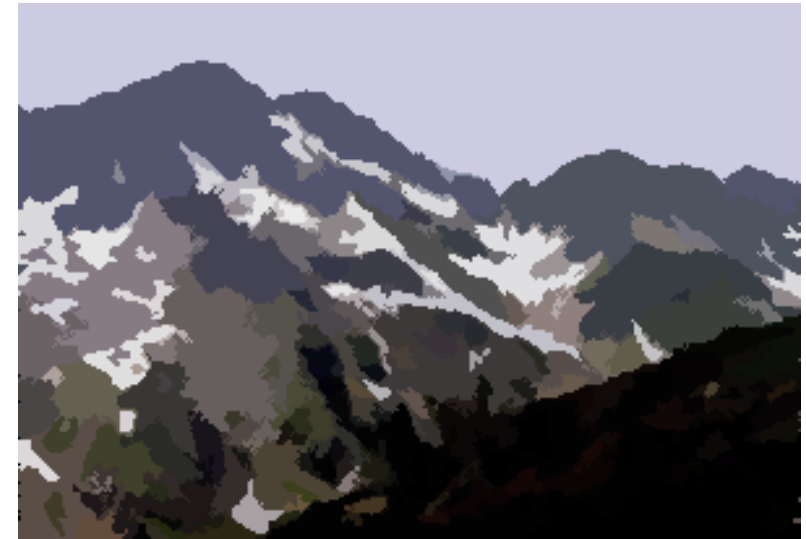


Mean-Shift Clustering/Segmentation

- ▶ Find features (color, gradients, texture, etc)
- ▶ Initialize windows at individual pixel locations
- ▶ Perform mean shift for each window until convergence
- ▶ Merge windows that end up near the same “peak” or mode



Mean-Shift Segmentation Results



<http://www.caip.rutgers.edu/~comanici/MSPAMI/msPamiResults.html>

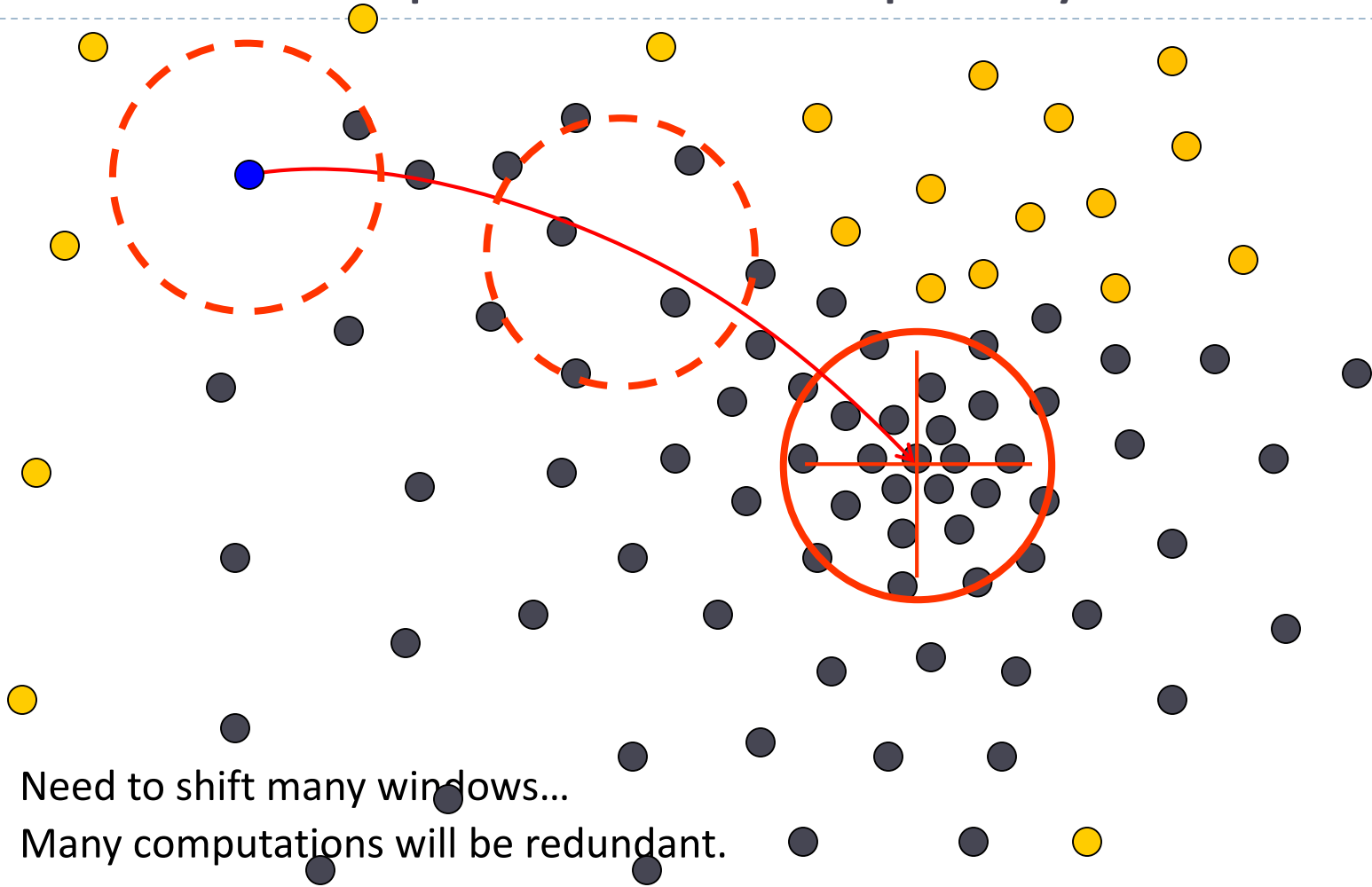
More Results



More Results

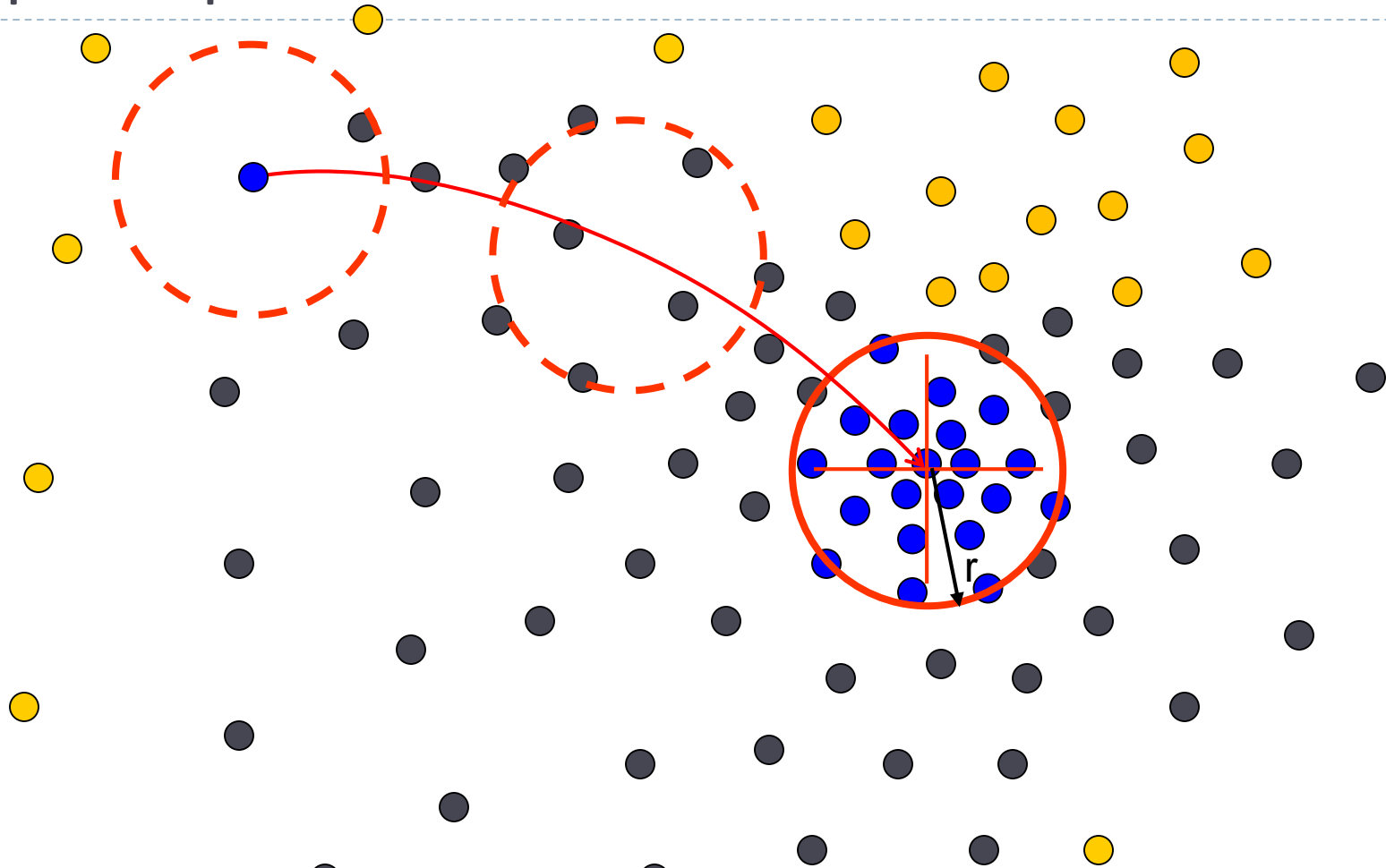


Problem: Computational Complexity



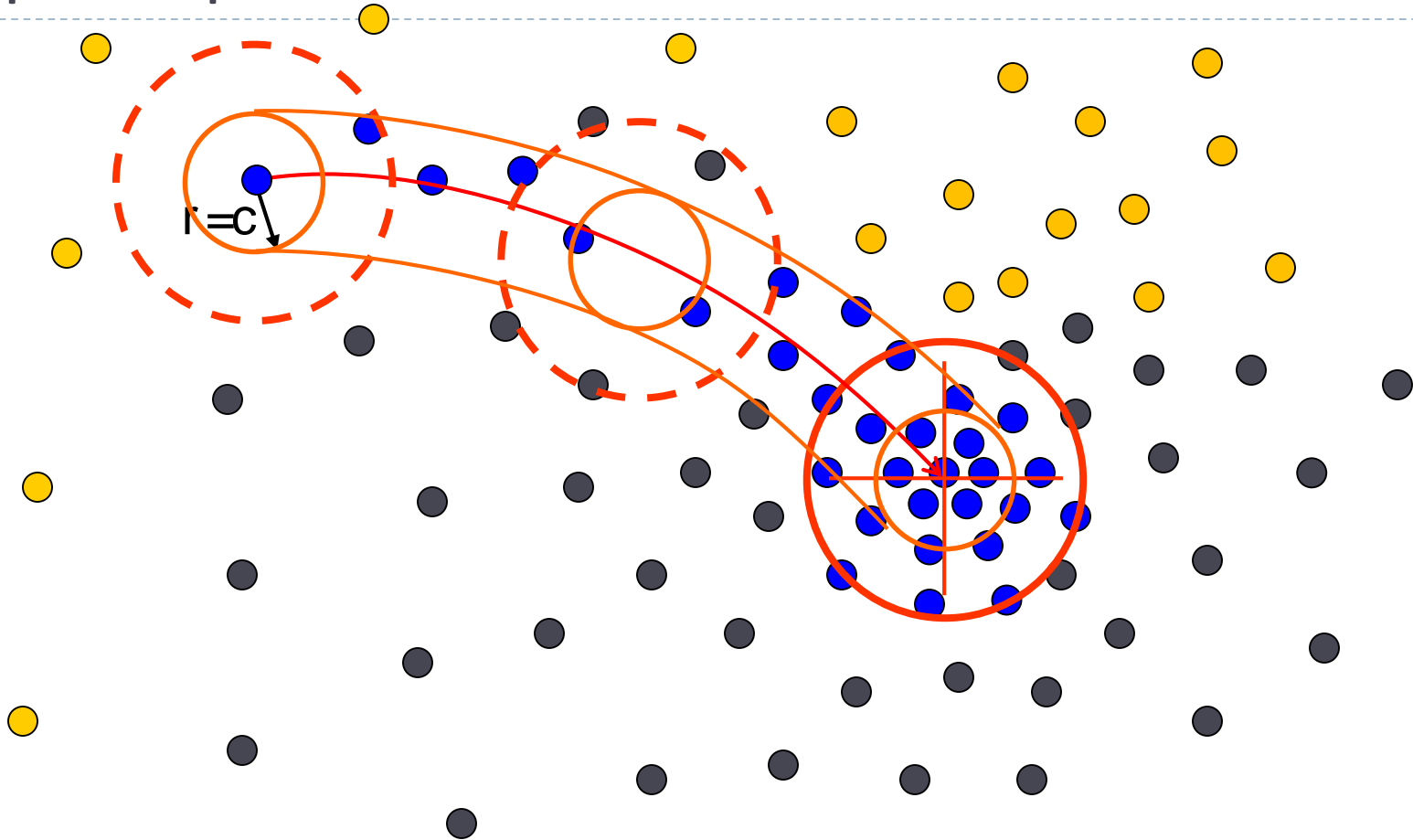
- ▶ Need to shift many windows...
- ▶ Many computations will be redundant.

Speedups: Basin of Attraction



1. Assign all points within radius r of end point to the mode.

Speedups



2. Assign all points within radius r/c of the search path to the mode -> reduce the number of data points to search.

Technical Details

Given n data points $\mathbf{x}_i \in \mathbb{R}^d$, the multivariate kernel density estimate using a radially symmetric kernel¹ (e.g., Epanechnikov and Gaussian kernels), $K(\mathbf{x})$, is given by,

$$\hat{f}_K = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \quad (1)$$

where h (termed the *bandwidth* parameter) defines the radius of kernel. The radially symmetric kernel is defined as,

$$K(\mathbf{x}) = c_k k(\|\mathbf{x}\|^2), \quad (2)$$

where c_k represents a normalization constant.

Technical Details

$$\nabla \hat{f}(\mathbf{x}) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^n g \left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \right]}_{\text{term 1}} \underbrace{\left[\frac{\sum_{i=1}^n \mathbf{x}_i g \left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^n g \left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \right]}_{\text{term 2}}, \quad (3)$$

where $g(x) = -k'(x)$ denotes the derivative of the selected kernel profile.

- Term1: this is proportional to the density estimate at \mathbf{x} (similar to equation 1 from the previous slide).
- Term2: this is the mean-shift vector that points towards the direction of maximum density.

Technical Details

Finally, the mean shift procedure from a given point \mathbf{x}_t is:

1. Compute the mean shift vector \mathbf{m} :

$$\left[\frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \right]$$

2. Translate the density window:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{m}(\mathbf{x}_i^t).$$

3. Iterate steps 1 and 2 until convergence.

$$\nabla f(\mathbf{x}_i) = 0.$$

Summary Mean-Shift

▶ Pros

- ▶ General, application-independent tool
- ▶ Model-free, does not assume any prior shape (spherical, elliptical, etc.) on data clusters
- ▶ Just a single parameter (window size h)
 - ▶ h has a physical meaning (unlike k-means)
- ▶ Finds variable number of modes
- ▶ Robust to outliers

▶ Cons

- ▶ Output depends on window size
- ▶ Window size (bandwidth) selection is not trivial
- ▶ Computationally (relatively) expensive ($\sim 2s/\text{image}$)
- ▶ Does not scale well with dimension of feature space