Clustering: Models and Algorithms

Shikui Tu
Shanghai Jiao Tong University
2021-03-01

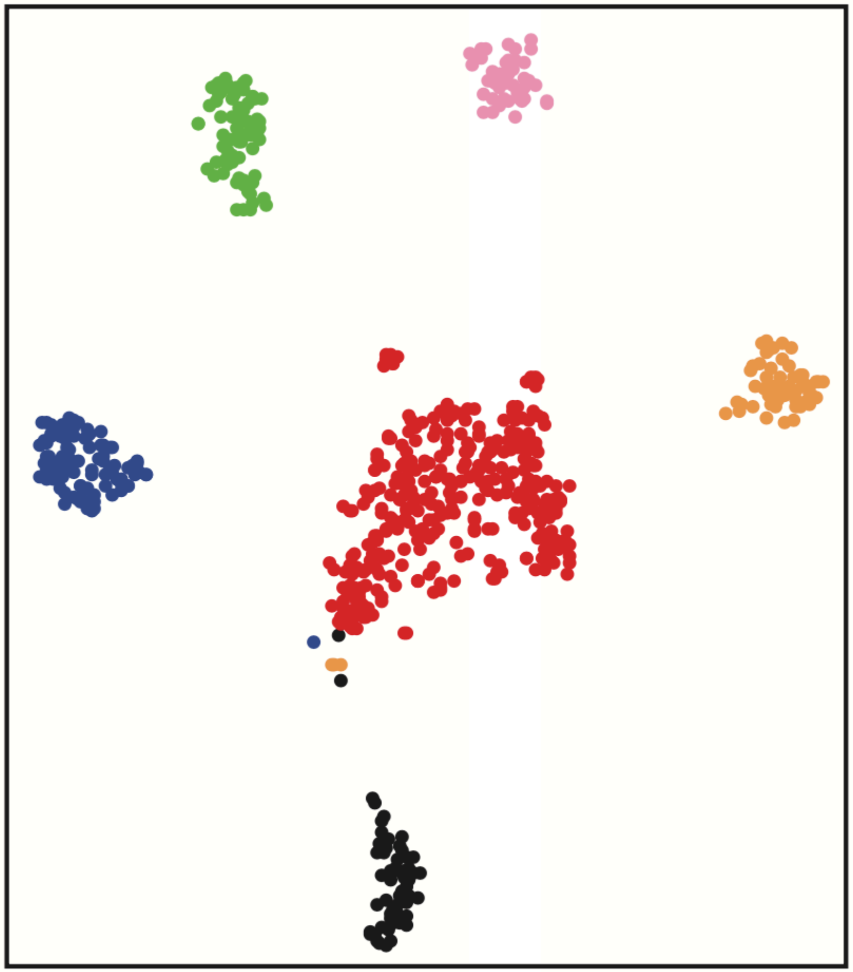
Outline

- Clustering
  - K-mean clustering, hierarchical clustering
- Adaptive learning (online learning)
  - CL, FSCL, RPCL
- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood

What is clustering?

例子：不同类型的癌细胞会各自聚在一起

物以类聚



Six malignant tumors (melanoma)

8 APRIL 2016 • VOL 352 ISSUE 6282, SCIENCE

Clustering analysis of COVID-19 virus

Lancet, January 29, 2020 https://doi.org/10.1016/S0140-6736(20)30251-8

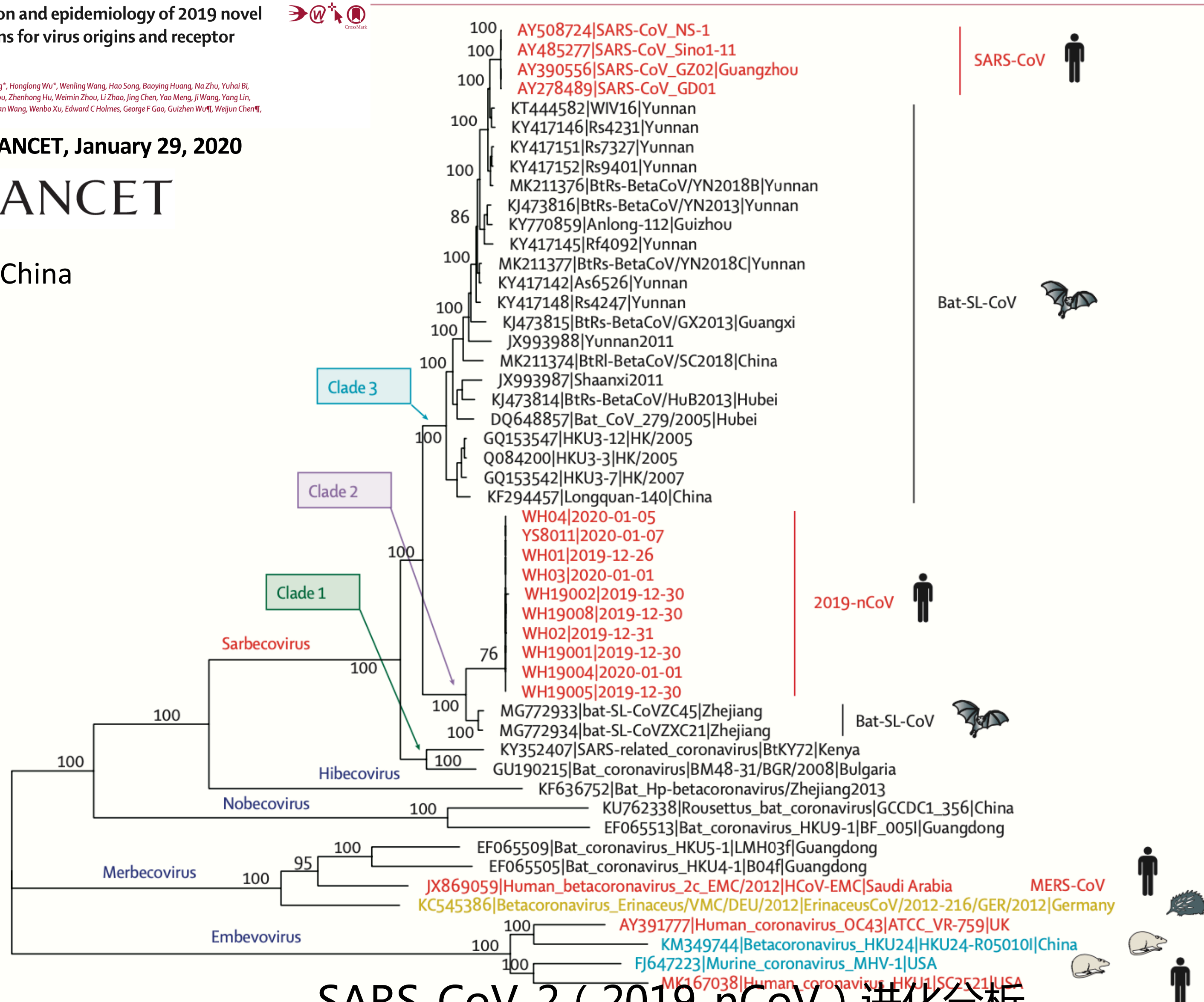
Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding

Roujian Lu*, Xiang Zhao*, Juan Li*, Peihua Niu*, Bo Yang*, Honglong Wu*, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J Liu, Dayan Wang, Wenbo Xu, Edward C Holmes, George F Gao, Guizhen Wu¶, Weijun Chen¶, Weifeng Shi¶, Wenjie Tan¶

THE LANCET, January 29, 2020

THE LANCET

CDC China

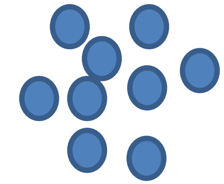


SARS-CoV-2（2019-nCoV）进化分析

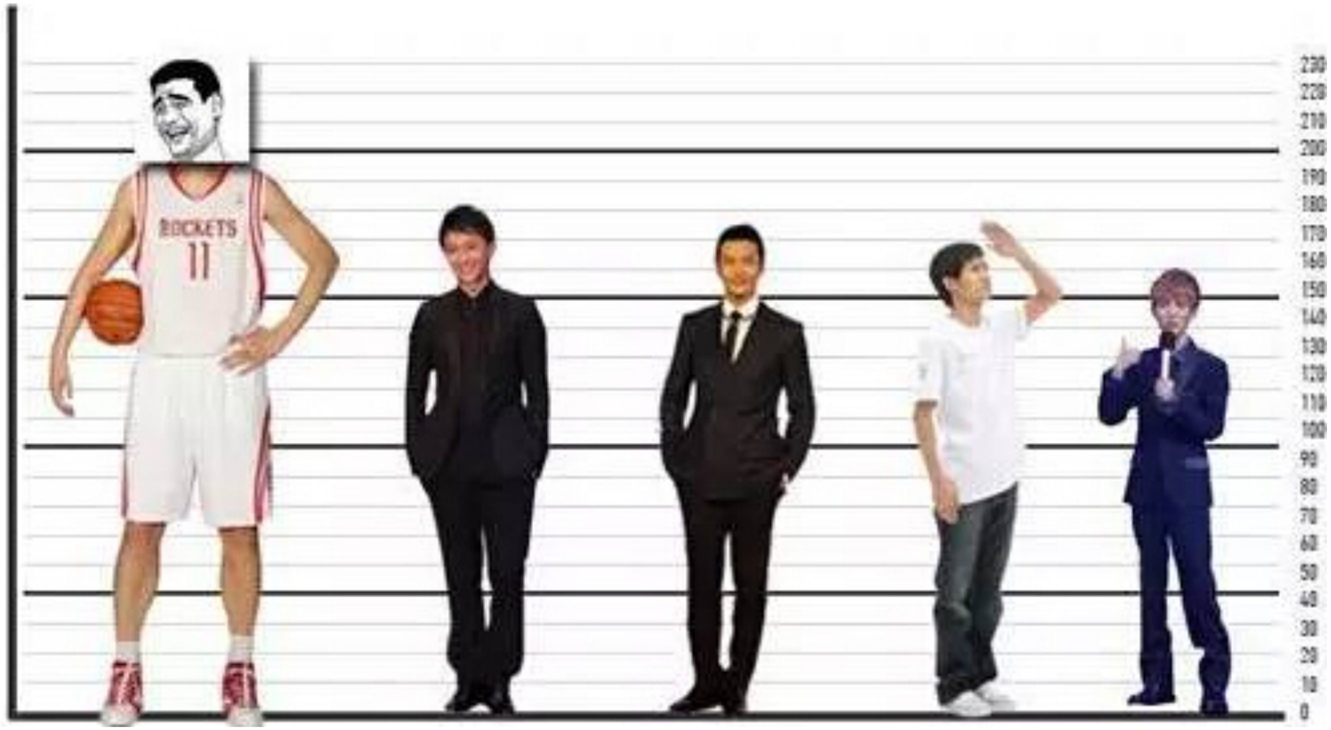
Figure 3: **Phylogenetic analysis of full-length genomes of 2019-nCoV and representative viruses of the genus Betacoronavirus**

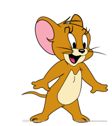
2019-nCoV=2019 novel coronavirus. MERS-CoV=Middle East respiratory syndrome coronavirus. SARS-CoV=severe acute respiratory syndrome coronavirus.

How to represent a cluster



- 例如：将每个人的身高记下来





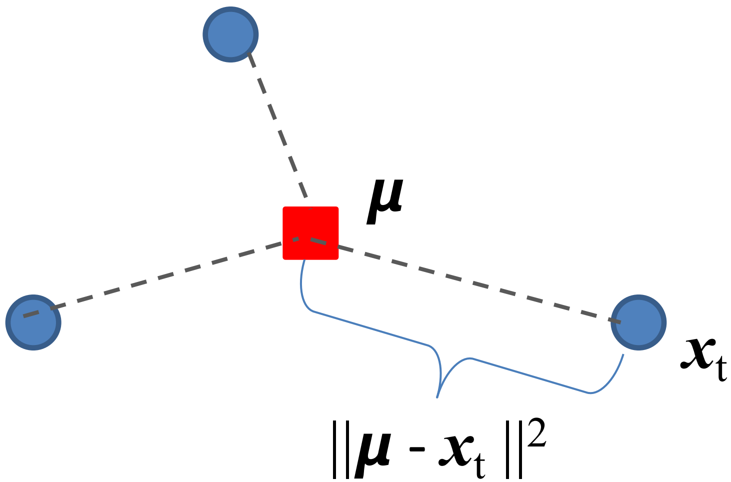
但是，如果只能记一个身高数值...

平均值

总误差最小

How to define error?

Square distance:

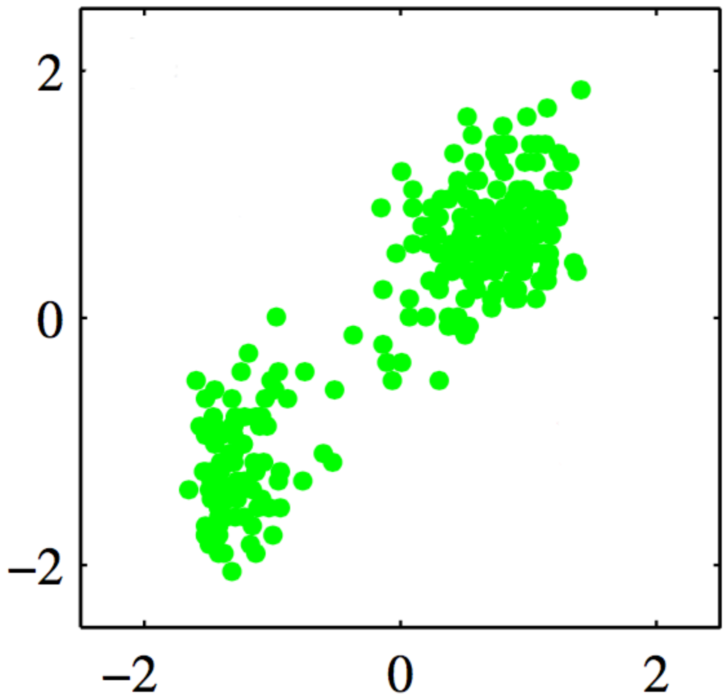


$$\|\boldsymbol{\mu} - \boldsymbol{x}_1\|^2 + \|\boldsymbol{\mu} - \boldsymbol{x}_2\|^2 + \|\boldsymbol{\mu} - \boldsymbol{x}_3\|^2$$

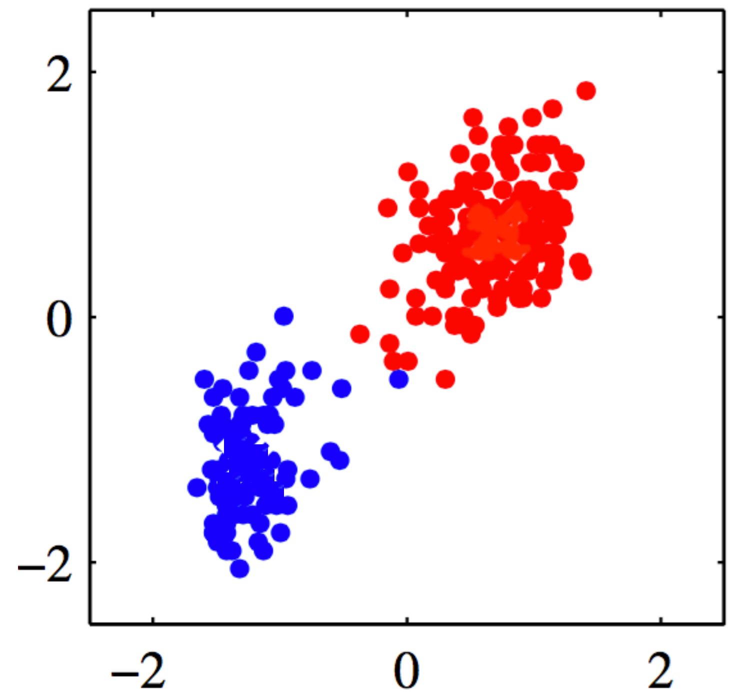
可以证明：当$\boldsymbol{\mu}$是所有数据点的均值时，平方距离和最小

Clustering the data

We have the following data:



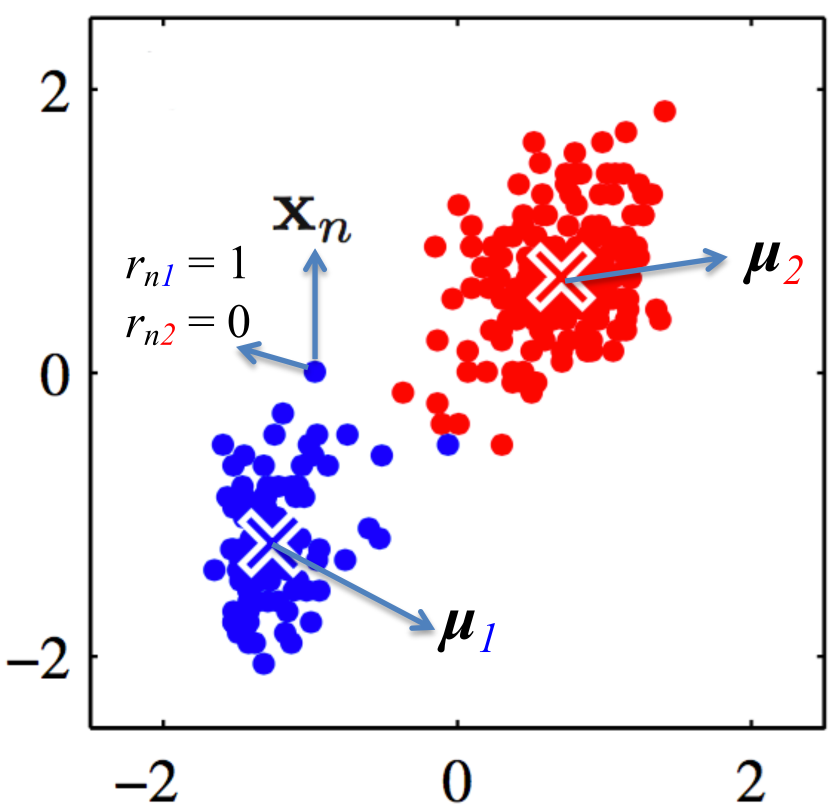
We want to cluster the data into two clusters (red and blue)



How?

Minimize the sum of square distances J

$$\text{minimize} \quad J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$



$r_{nk} = 1$ if and only if data point $\mathbf{x}_n$ is assigned to cluster k; otherwise $r_{nk} = 0$.

$k = 1, 2$; K = 2 clusters

$n = 1, \ldots, N$;
N: the total number of points.

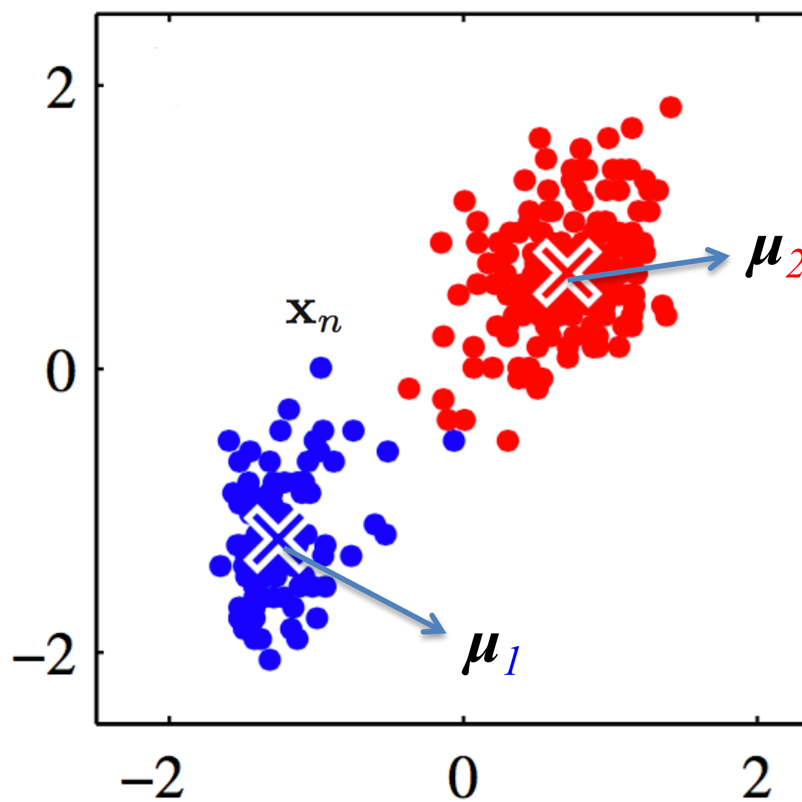
We need to calculate $\{ r_{nk} \}$ and $\{ \boldsymbol{\mu}_k \}$.

If we know r_{n1} , r_{n2} for all n=1,…,N

Since the points have been assigned to cluster 1 or cluster 2, we calculate

$\boldsymbol{\mu}_1$ = mean of the points in cluster 1

$\boldsymbol{\mu}_2$ = mean of the points in cluster 2



Or formally

$$\boxed{\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}}$$

We call it the **M Step.**

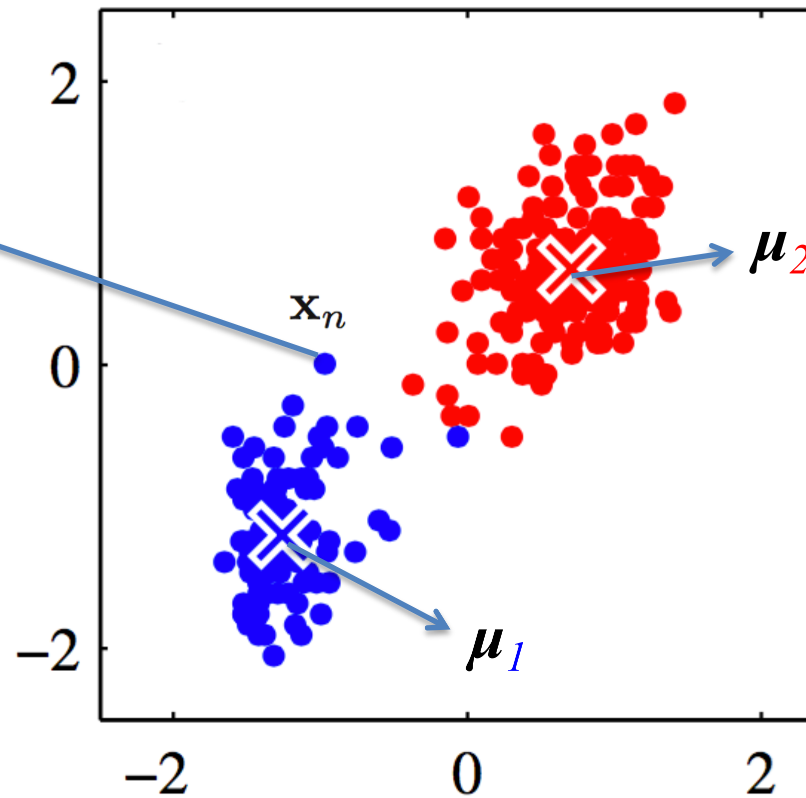
If we know $\boldsymbol{\mu}_{1}, \boldsymbol{\mu}_{2}$

We should assign point $\mathbf{x}_n$ to cluster 1, because

$$\| \mathbf{x}_n - \boldsymbol{\mu}_1 \|^2 < \| \mathbf{x}_n - \boldsymbol{\mu}_2 \|^2$$

Then, $r_{n1} = 1$

$r_{n2} = 0$

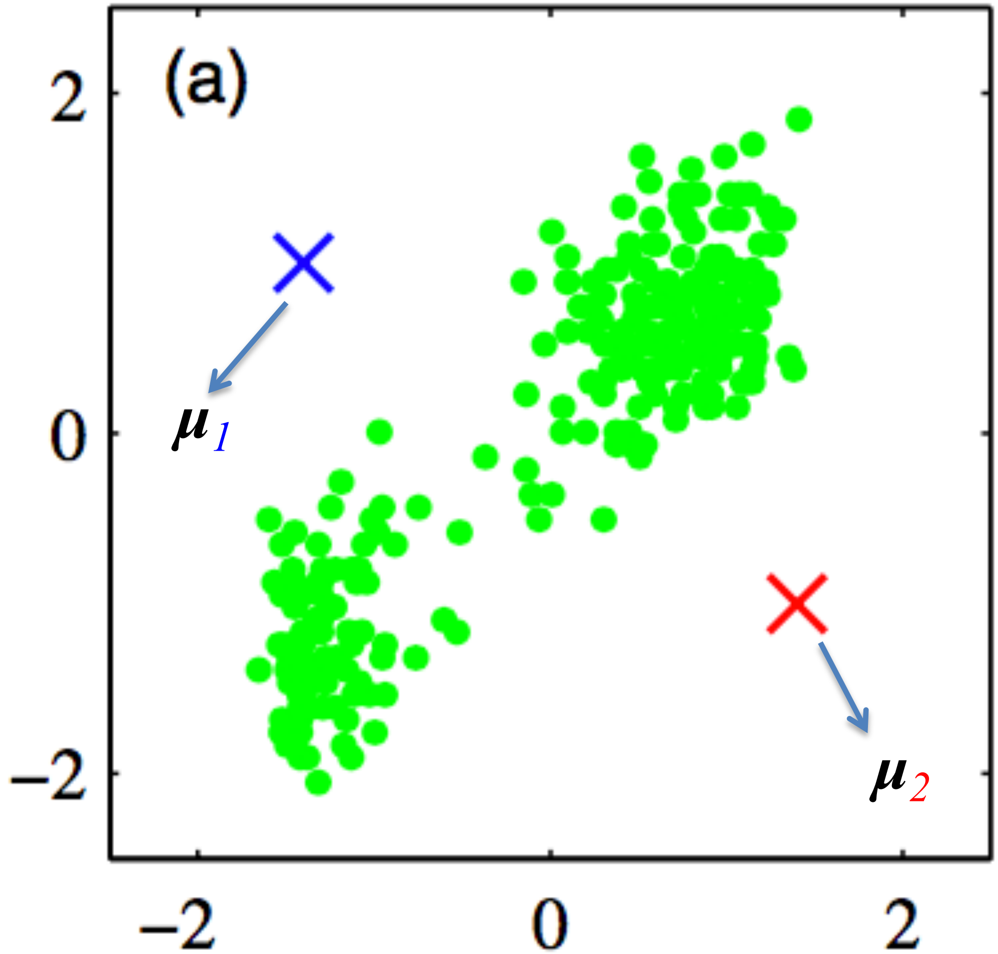


Or formally

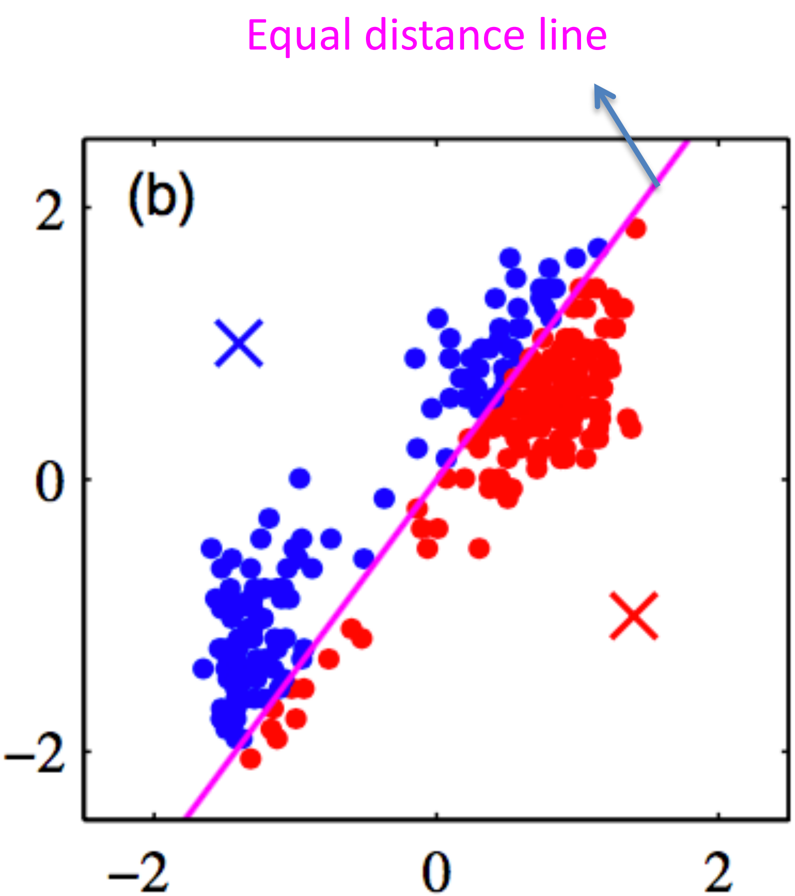
$$\boxed{r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}}$$

We call it the **E Step**

Initialization



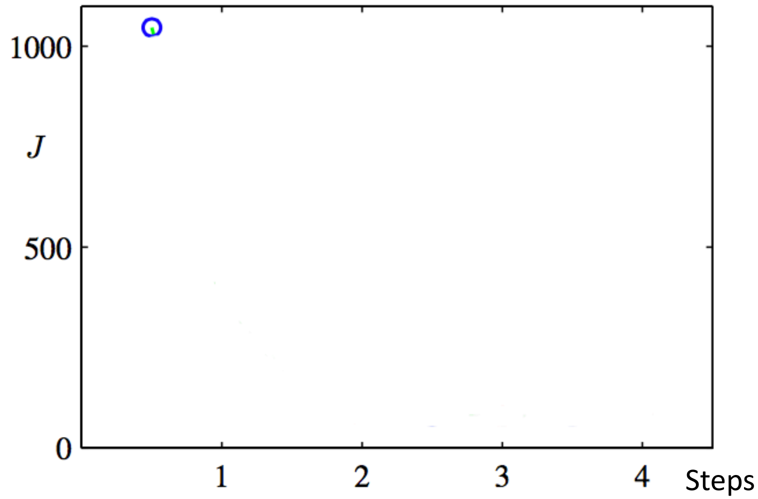
Given $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, calculate r_{n1}, r_{n2} for all n=1,…,N



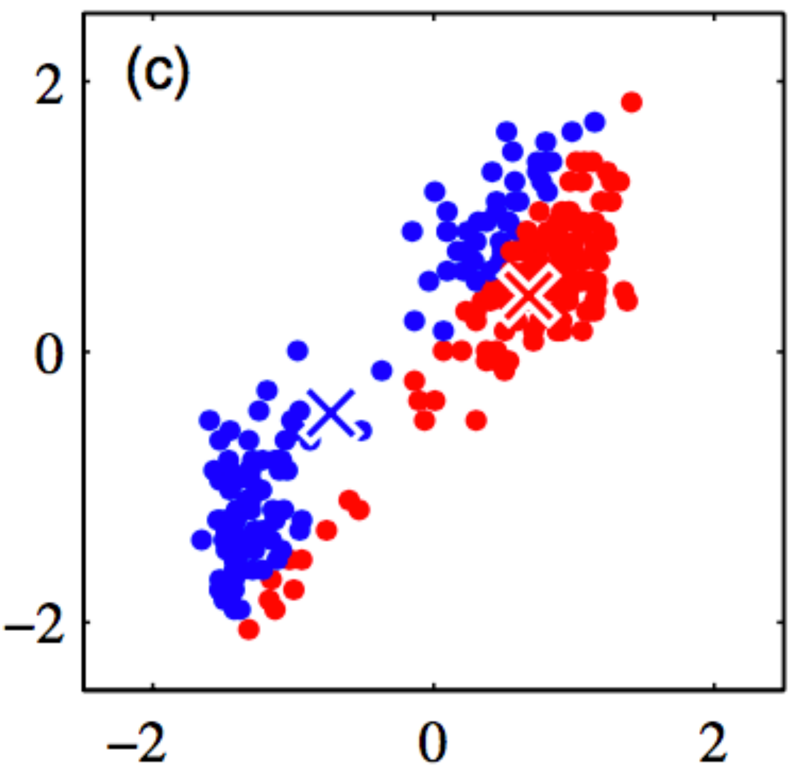
E Step

Assign the points to the nearest cluster:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$



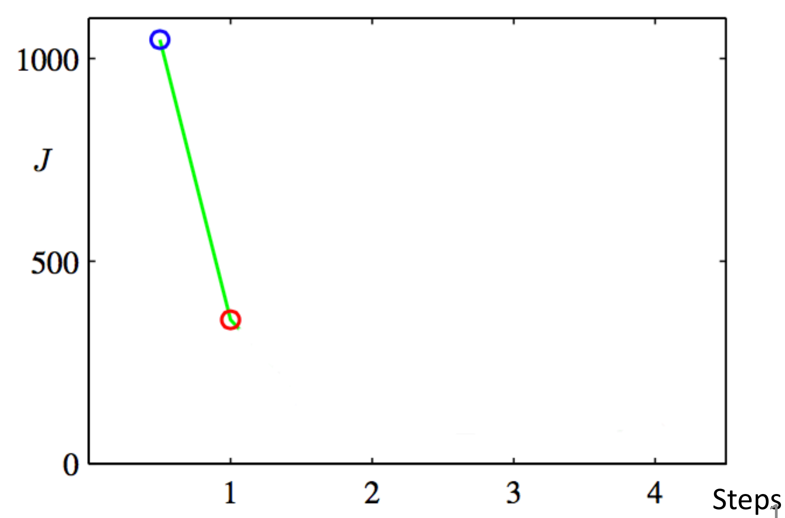
Given r_{n1} , r_{n2} , calculate $\boldsymbol{\mu}_{1,}$ $\boldsymbol{\mu}_2$



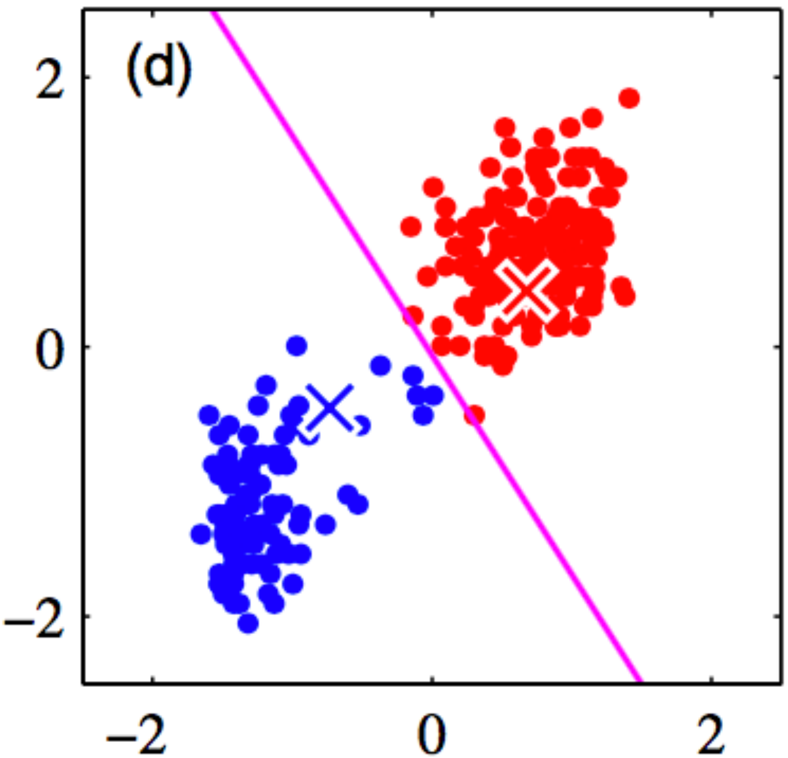
M Step

Calculate the means of the points in each cluster:

$$\boxed{\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}}$$



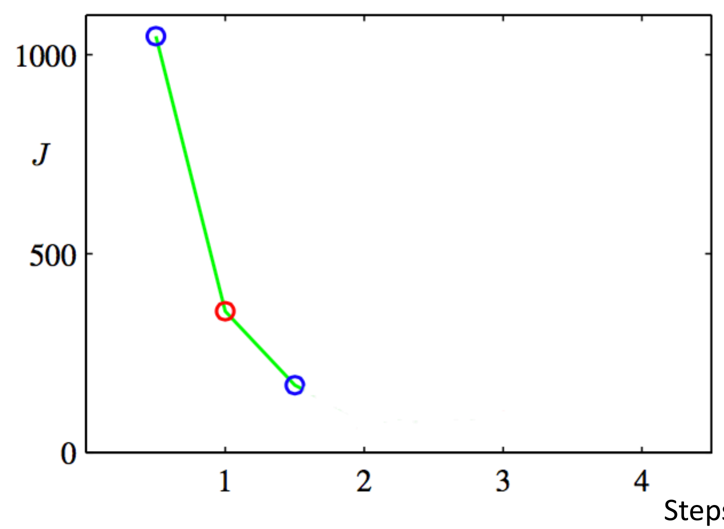
Given $\boldsymbol{\mu}_{1}, \boldsymbol{\mu}_{2}$, calculate r_{n1}, r_{n2} for all $n=1,\ldots,N$



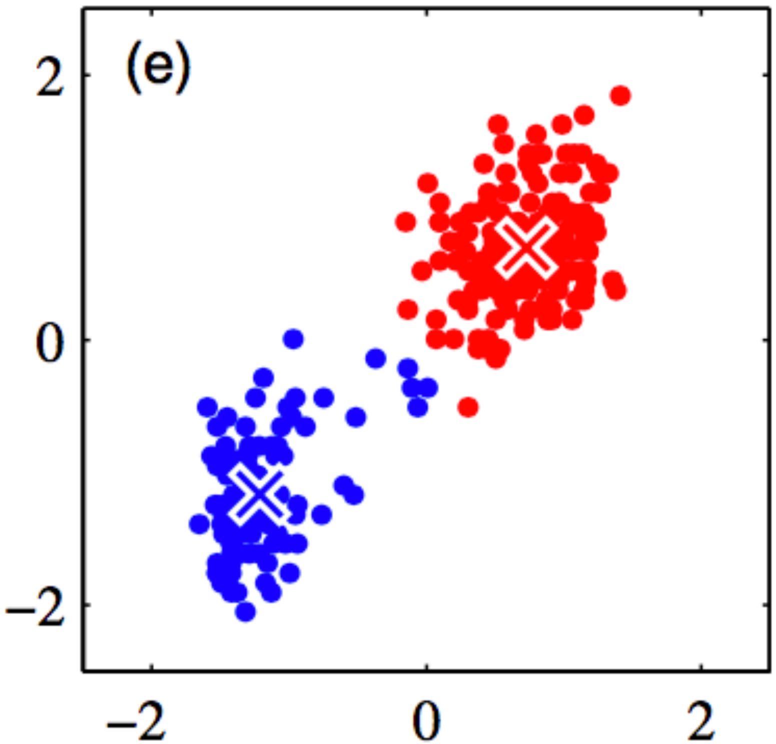
E Step

Assign the points to the nearest cluster:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$



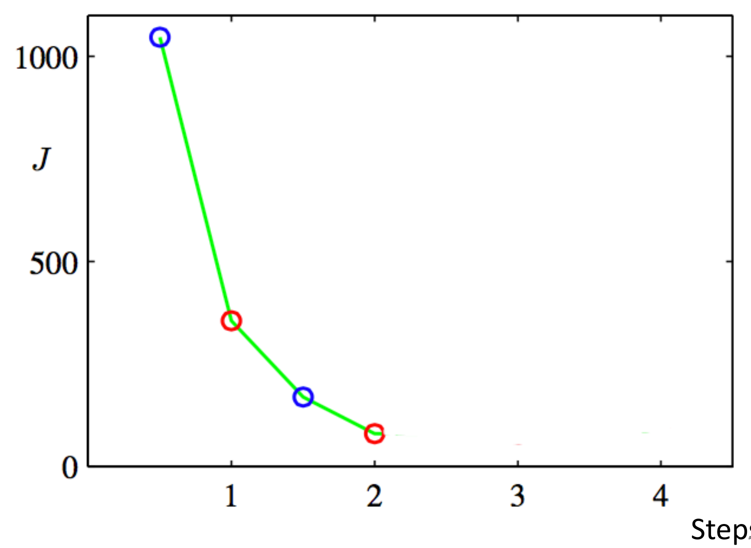
Given r_{n1}, r_{n2}, calculate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$

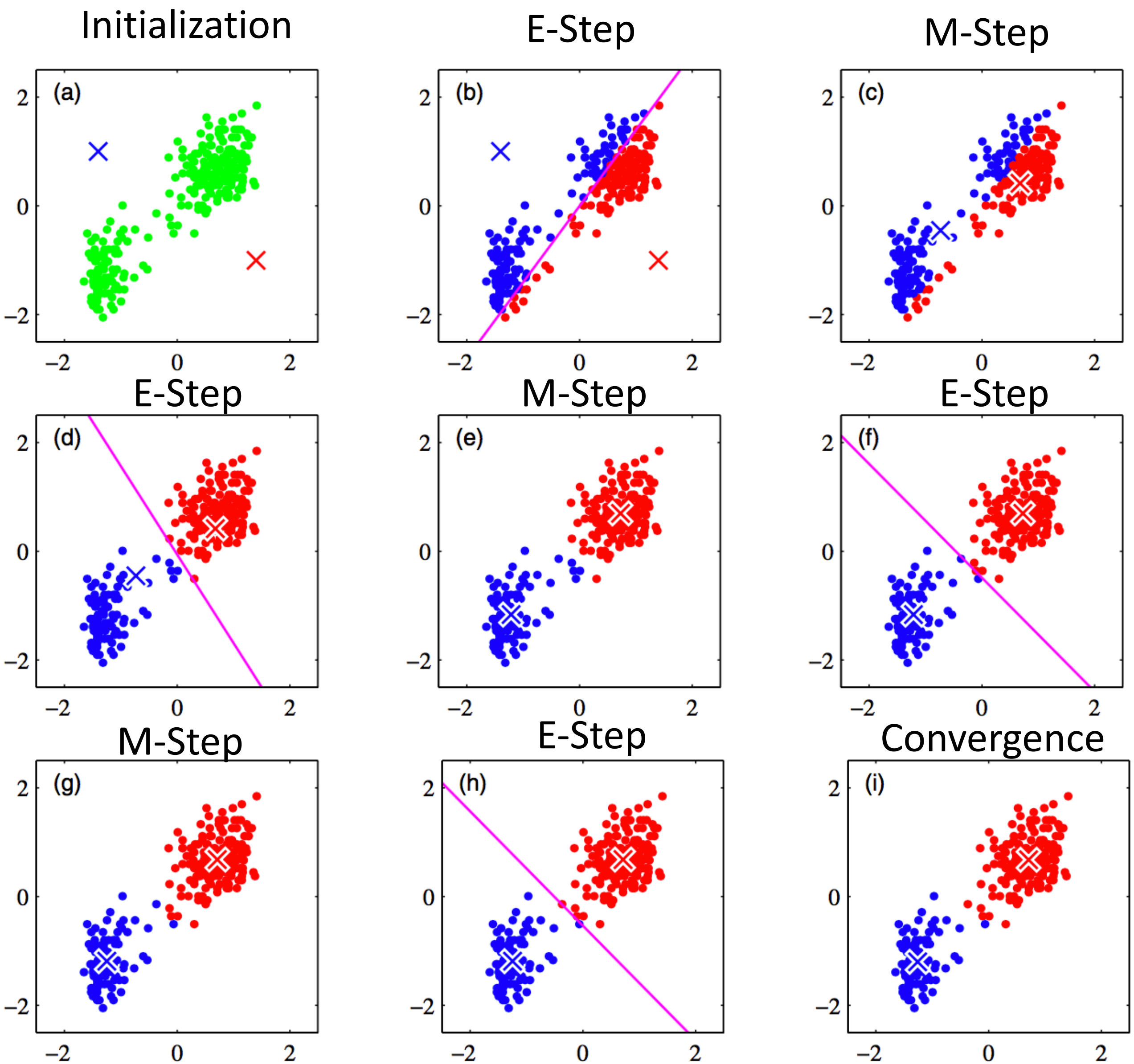


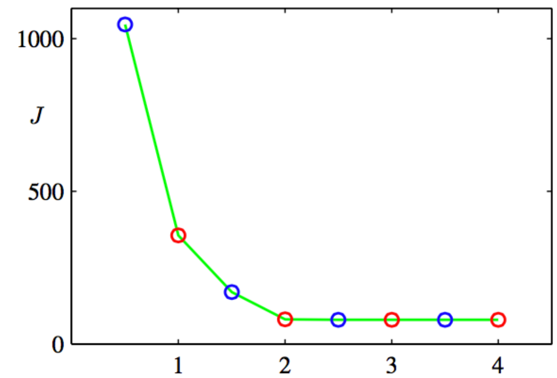
M Step

Calculate the means of the points in each cluster:

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$







If J does not change, or $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ do not change, then the algorithm converges.

K均值法小结

- 初始化均值点$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$
- 迭代如下
  - 把每个数据点按照就近原则分配给相应的$\boldsymbol{\mu}_i$
  - 把$\boldsymbol{\mu}_i$更新为所分配的数据点的均值
- 迭代停止，如果聚类分配不变

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$
Repeat
    For all $\boldsymbol{x}^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

    For all $\boldsymbol{m}_i, i = 1, \ldots, k$

$$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$

Until $\boldsymbol{m}_i$ converge

Basic ingredients

- Model or structure
- Objective function
- Algorithm
- Convergence

Questions

- How many possible assignments for K-mean clustering? K^N

- Can K-mean algorithm always converge? Why?

- Possible limitations of K-mean clustering?

Outline

- Clustering
  - K-mean clustering, **hierarchical clustering**
- Adaptive learning (online learning)
  - CL, FSCL, RPCL
- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood

Hierarchical Clustering

- Agglomerative clustering
- A very simple procedure:
  - Assign each data point into its own group
  - Repeat: look for the two **closest** groups and merge them into one group
  - Stop when all the data points are merged into a single cluster

Adapted from Blei, D. *Hierarchial Cluster* [PwerPoint slides]. www.cs.princeton.edu/courses/archive/spr08/cos424/slides/clustering-2.pdf

Hierarchical Clustering

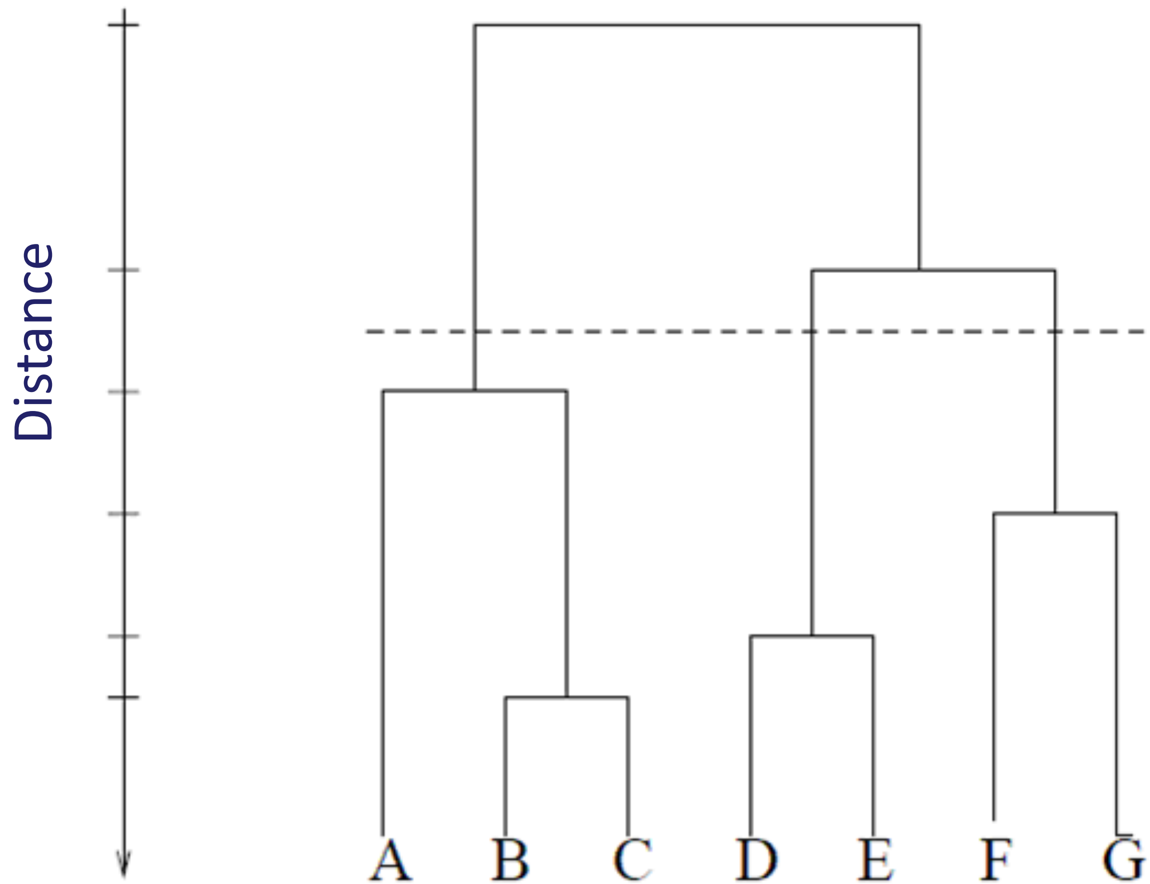
- *k*-means clustering requires
  - *k*
  - Positions of initial centers
  - A distance measure between points (*e.g.* Euclidean distance)
- Hierarchical clustering requires a measure of distance between *groups* of data points

Adapted from Blei, D. *Hierarchial Cluster* [PwerPoint slides]. www.cs.princeton.edu/courses/archive/spr08/cos424/slides/clustering-2.pdf

Distance Measure

- Distance between data points *a* and *b*:
  - $d(a, b)$
- Group A and B
  - Single-linkage
$$d(A, B) = \min_{a \in A, b \in B} d(a, b)$$
  - Complete-linkage
$$d(A, B) = \max_{a \in A, b \in B} d(a, b)$$
  - Average-linkage
$$d(A, B) = \frac{\sum_{a \in A, b \in B} d(a, b)}{|A| \cdot |B|}$$

Dendrogram



Jain, A. K., Murty, M. N., Flynn, P. J. (1999) "Data Clustering: A Review". ACM Computing Surveys (CSUR), 31(3), p.264-323, 1999.

基于层次聚类的进化关系分析

THE LANCET, January 29, 2020

CDC China

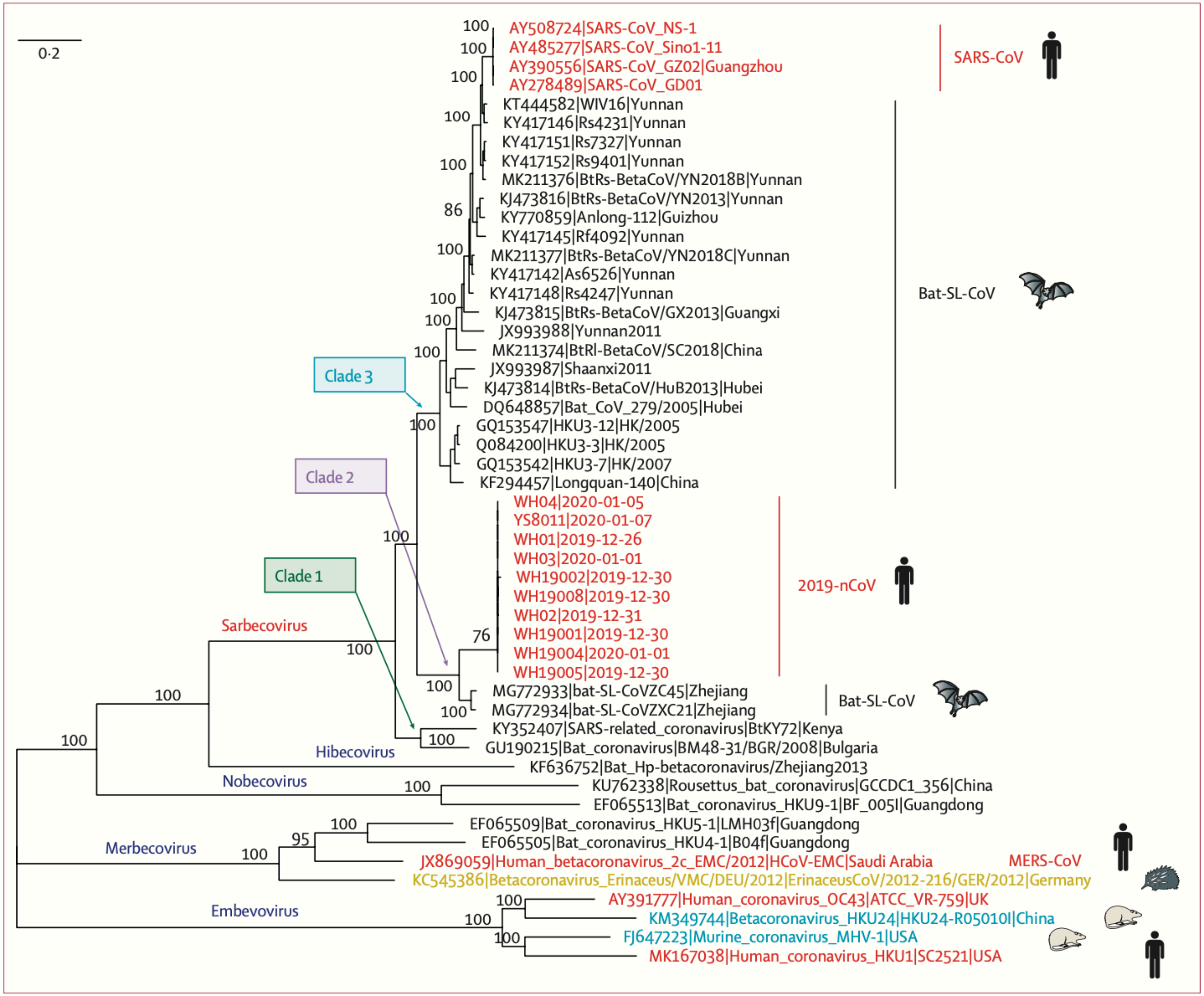


Figure 3: Phylogenetic analysis of full-length genomes of 2019-nCoV and representative viruses of the genus Betacoronavirus

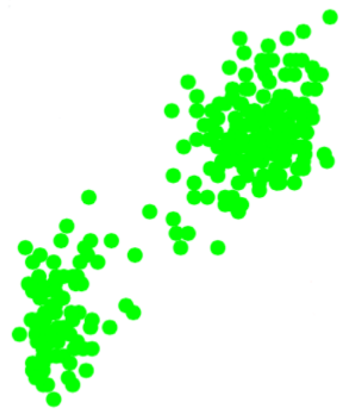
2019-nCoV=2019 novel coronavirus. MERS-CoV=Middle East respiratory syndrome coronavirus. SARS-CoV=severe acute respiratory syndrome coronavirus.

Outline

- Clustering
  - K-mean clustering, hierarchical clustering
- Adaptive learning (online learning)
  - CL, FSCL, RPCL
- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood

From batch to adaptive

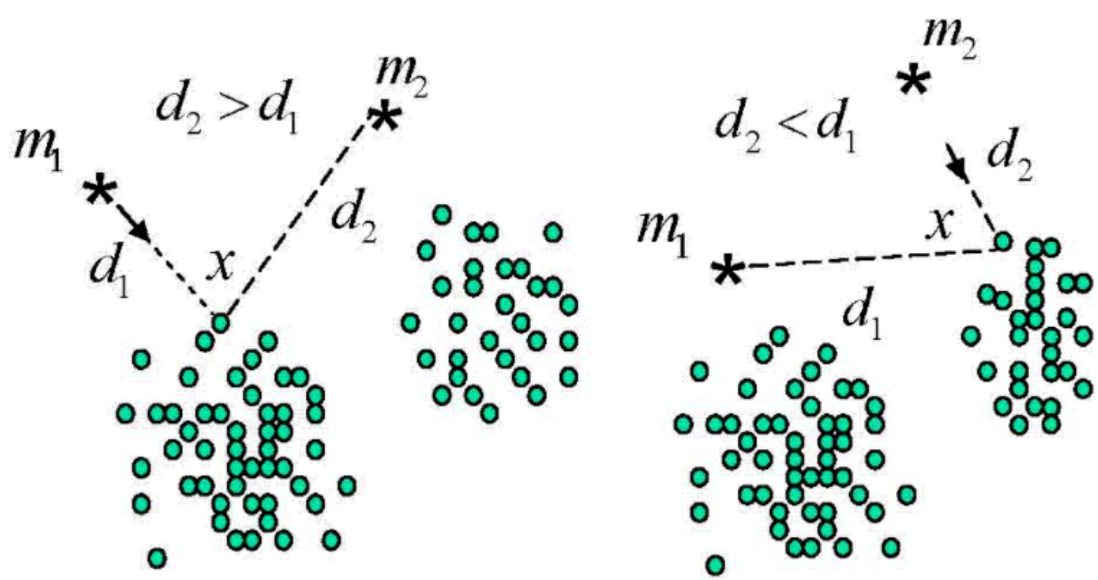
- Given a batch of data points



- Data points come one by one:

Competitive learning

- Data points come one by one:



(a) m_1 is the winner (b) m_2 is the winner

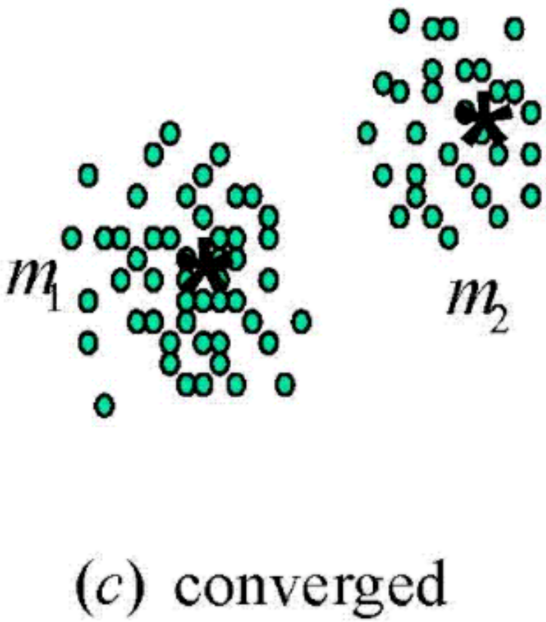
$$\varepsilon_t(\theta_j) = \|x_t - m_j\|^2$$

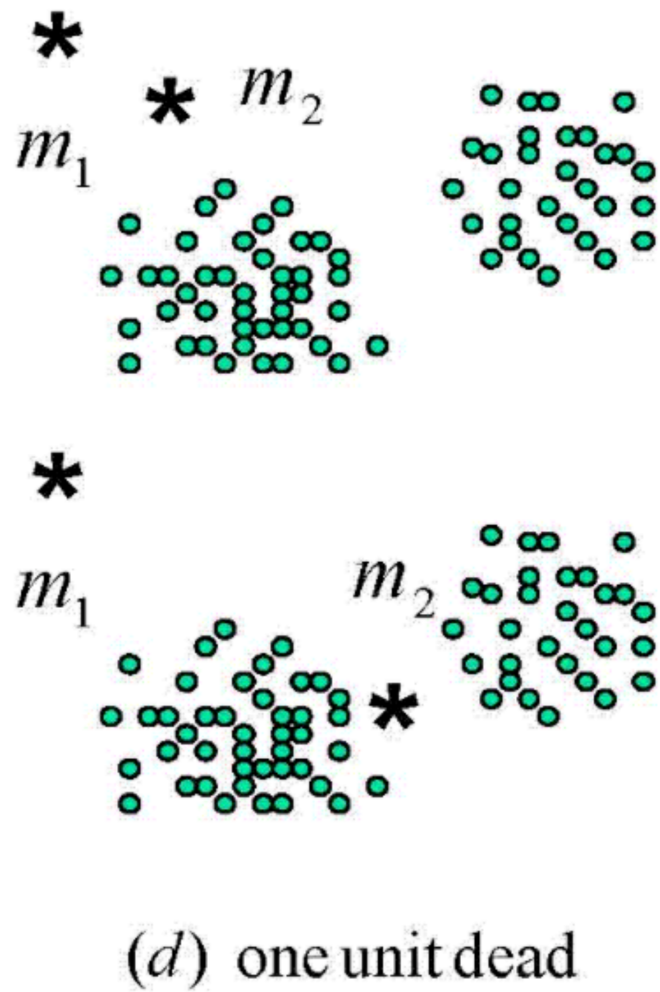
$$p_{j,t} = \begin{cases} 1, & \text{if} \quad j = c, \\ 0, & \text{otherwise;} \end{cases}$$

$$c = arg\ min_j \varepsilon_t(\theta_j).$$

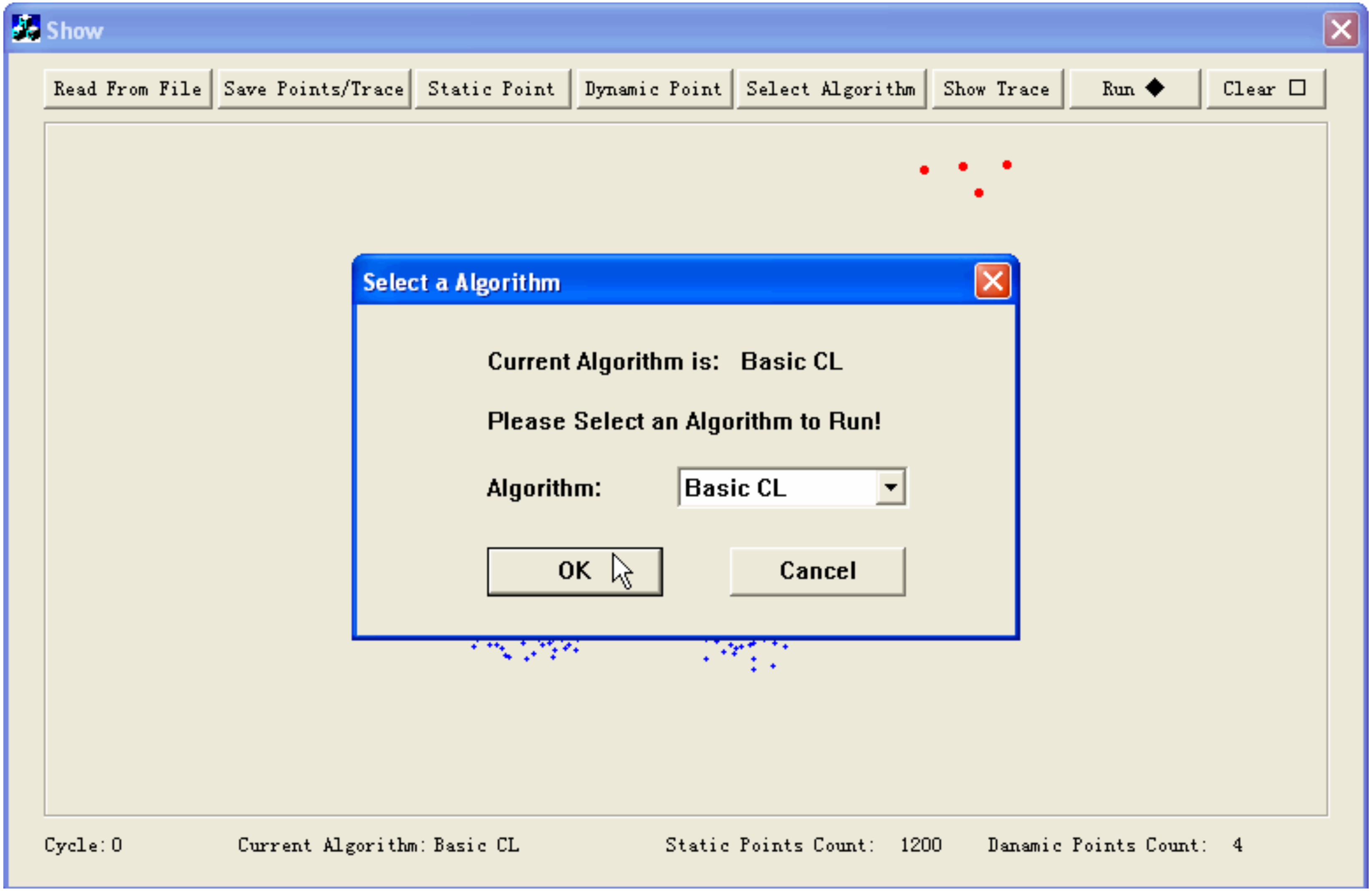
$$m_j^{new} = m_j^{old} + \eta p_{j,t}(x_t - m_j^{old}).$$

When starting with "bad initializations"





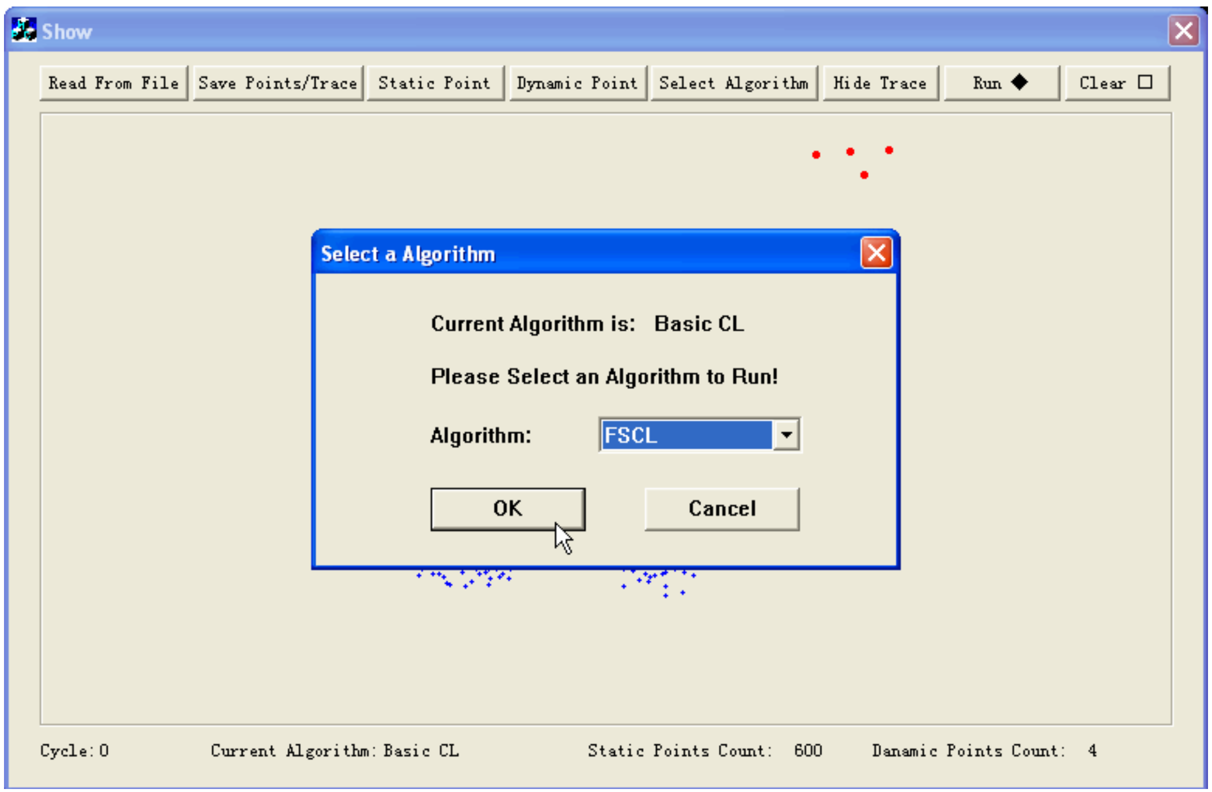
A four-cluster case



frequency sensitive competitive learning (FSCL) [Ahalt et al., 1990]

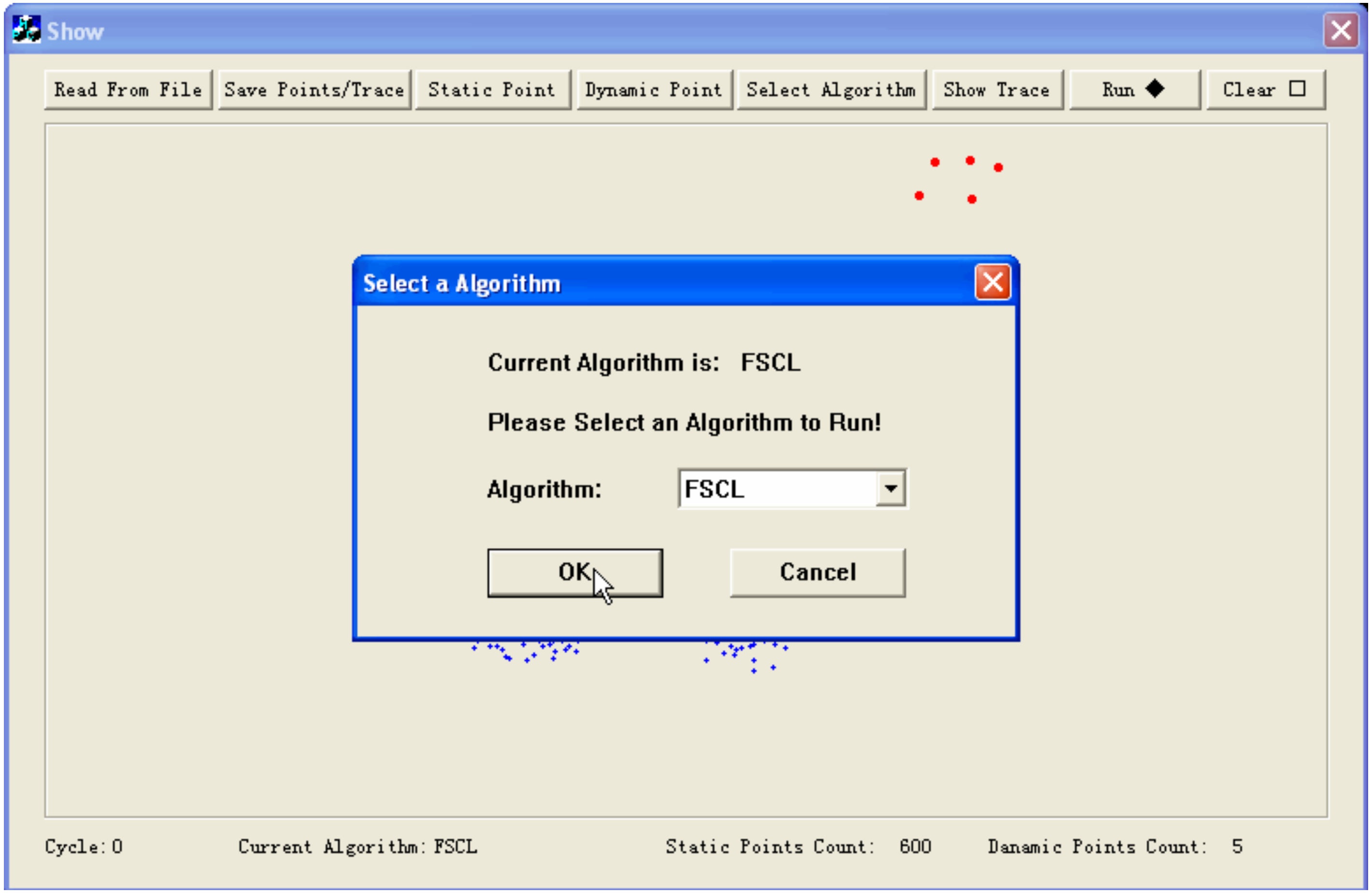
The idea is to penalize the frequent winners:

$$\varepsilon_t(\theta_j) = \alpha_j ||x_t - m_j||^2$$



FSCL is not good when there are extra centers

When k is pre-assigned to 5. the frequency sensitive mechanism also brings the extra one into data to disturb the correct locations of others

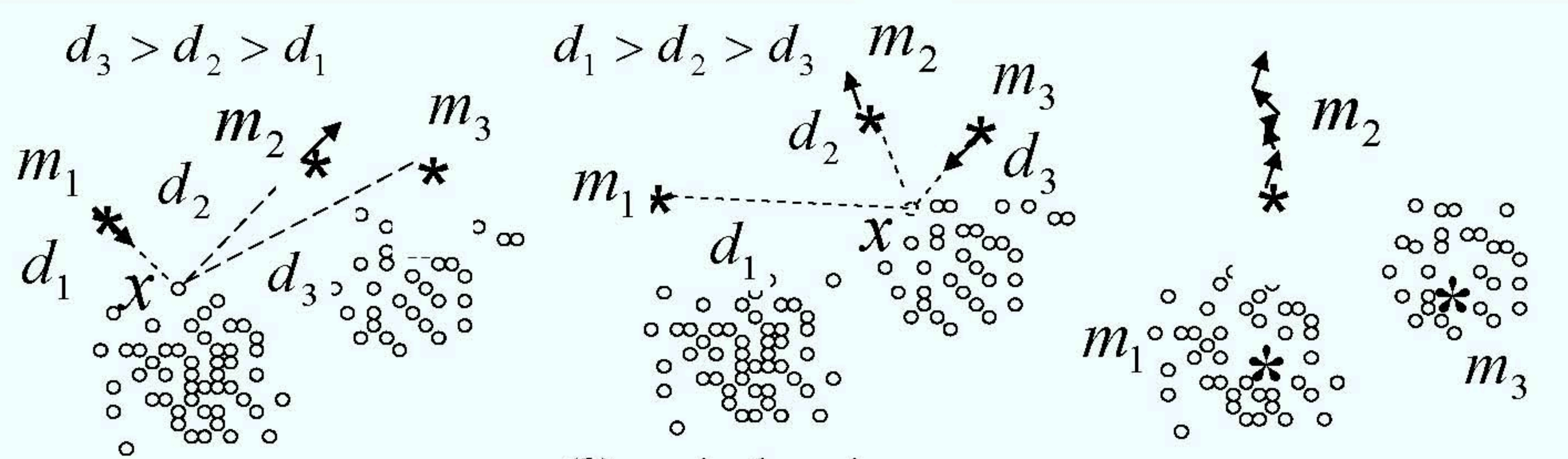


Rival penalized competitive learning (RPCL) (Xu, Krzyzak, & Oja, 1992 , 1993)

The RPCL differs from FSCL by implementing $p_{j,t}$ as follows:

$$p_{j,t} = \begin{cases} 1, & \text{if } j = c, \\ -\gamma, & \text{if } j = r, \\ 0, & \text{otherwise,} \end{cases} \quad \begin{cases} c = arg\ min_j \varepsilon_t(\theta_j), \\ r = arg\ min_{j \neq c} \varepsilon_t(\theta_j), \end{cases}$$

where γ approximately takes a number between 0.05 and 0.1 for controlling the penalizing strength.

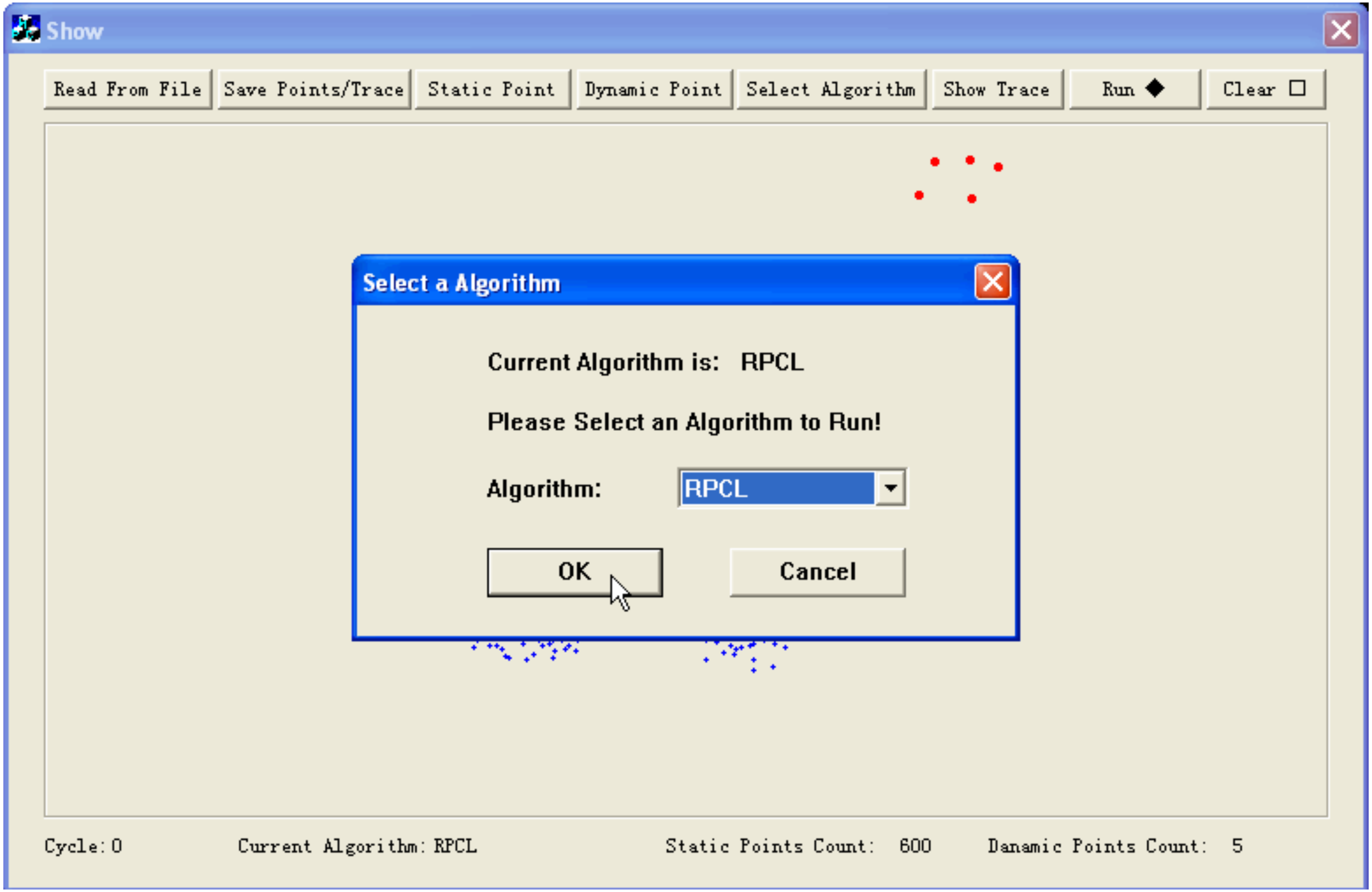


(a) m_1 is the winner
m_2 is the rival

(b) m_3 is the winner
m_2 is the rival

(c) m_1 and m_3 are converged
m_2 is driven far away

Rival penalized mechanism makes extra agents driven far away.



Questions

- Are competitive learning (CL) and K-mean equivalent?

- Could you come up with new algorithms to tackle the “bad initialization” problem of competitive learning (or K-mean)?

- Can you design a K-mean version of RPCL?

Thank you!

Matrix derivatives

$$\left[\frac{\partial \mathbf{x}}{\partial y}\right]_i = \frac{\partial x_i}{\partial y} \qquad \left[\frac{\partial x}{\partial \mathbf{y}}\right]_i = \frac{\partial x}{\partial y_i} \qquad \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right]_{ij} = \frac{\partial x_i}{\partial y_j}$$

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \tag{69}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \tag{70}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T \tag{71}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \tag{72}$$

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \det(\mathbf{X})(\mathbf{X}^{-1})^T \tag{49}$$

$$\frac{\partial \mathbf{Y}^{-1}}{\partial x} = -\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\mathbf{Y}^{-1} \tag{59}$$

http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf