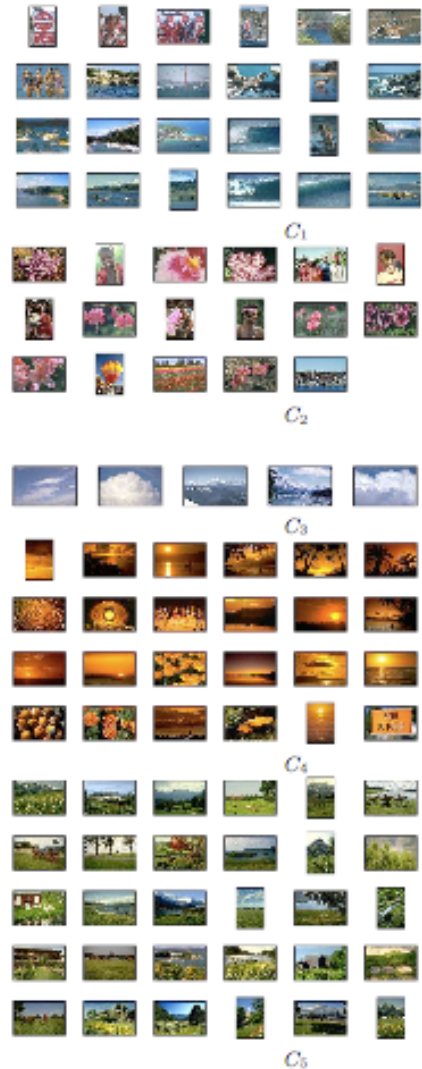
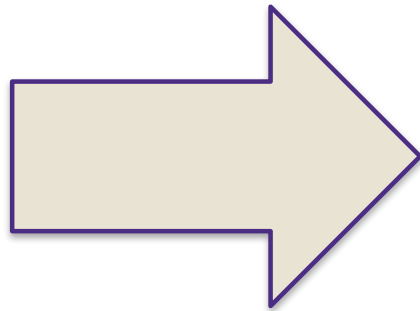


Unsupervised Learning.  $\{X_i\}_{i=1}^n$   $\left\{ \begin{array}{l} \text{dimensionality reduction : PCA} \\ \text{clustering} \end{array} \right.$

# Clustering with $k$ -means


---

# Clustering images



[Goldberger et al.]

# Clustering web search results



web news images wikipedia blogs jobs more »

race

Search

advanced preferences

clusters sources sites

remix

All Results (238)

Car (28)

Race cars (7)

Photos, Races Scheduled (5)

Game (4)

Track (3)

Nascar (2)

Equipment And Safety (2)

Other Topics (7)

Photos (22)

Game (14)

Definition (13)

Team (18)

Human (8)

Classification Of Human (2)

Statement, Evolved (2)

Other Topics (4)

Weekend (8)

Ethnicity And Race (7)

Race for the Cure (8)

Race Information (8)

more | all clusters

find in clusters:

Find

Cluster Human contains 8 documents.

Search Results

1. [Race \(classification of human beings\) - Wikipedia, the free ...](#)

The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **races**, vary by culture and over time, and are often controversial for scientific as well as social and political reasons. History · Modern debates · Political and ...  
[en.wikipedia.org/wiki/Race\\_\(classification\\_of\\_human\\_beings\)](#) - [cache] - Live, Ask

2. [Race - Wikipedia, the free encyclopedia](#)

General. **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of human beings) **Race** and ethnicity in the United States Census, official definitions of "**race**" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games  
[en.wikipedia.org/wiki/Race](#) - [cache] - Live, Ask

3. [Publications | Human Rights Watch](#)

The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...  
[www.hrw.org/backgrounder/usa/race](#) - [cache] - Ask

4. [Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...](#)

Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...** From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...  
[www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861](#) - [cache] - Live

5. [AAPA Statement on Biological Aspects of Race](#)

AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study **human** evolution and variation, ...  
[www.physanth.org/positions/race.html](#) - [cache] - Ask

6. [race: Definition from Answers.com](#)

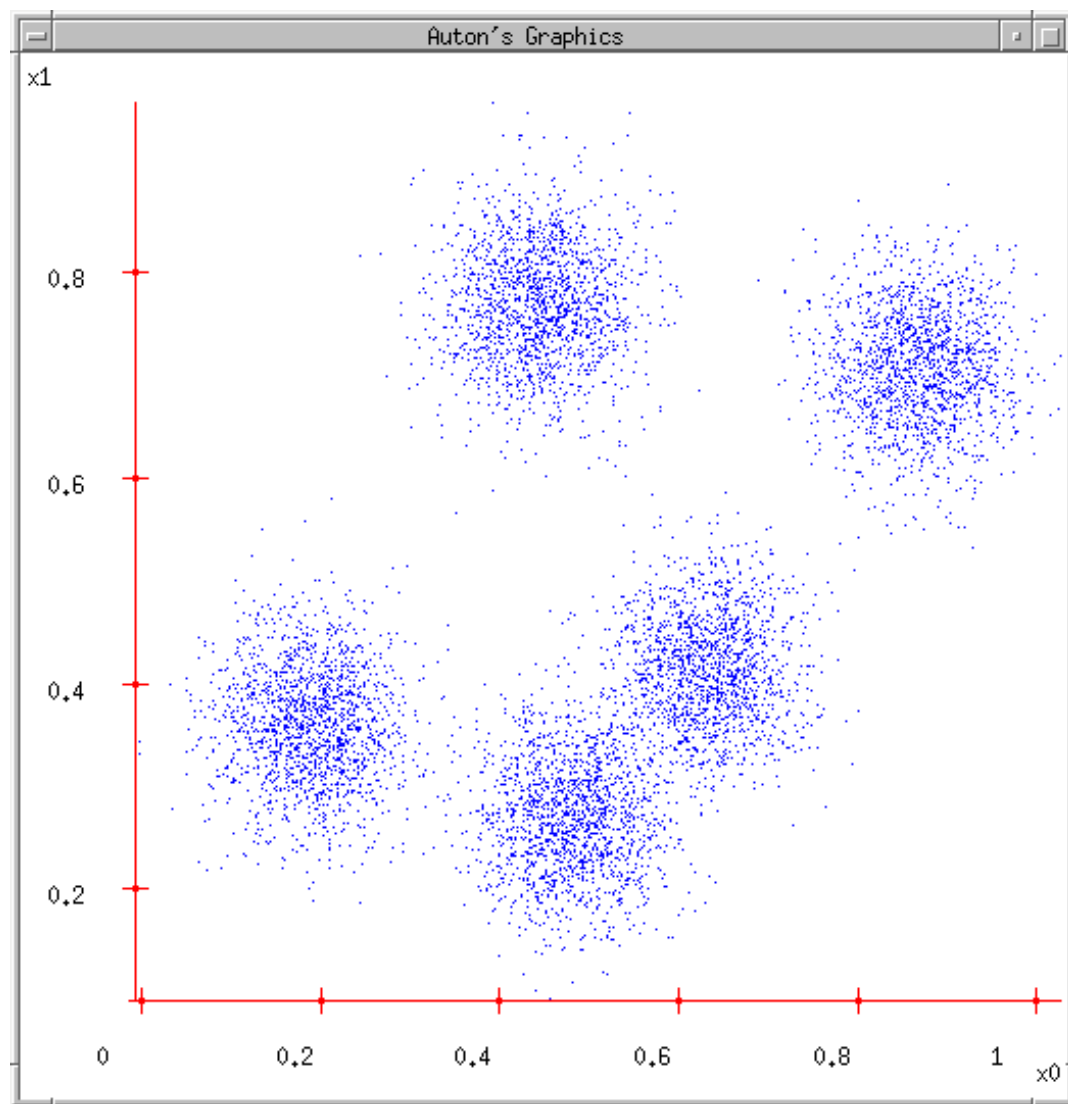
**race** n. A local geographic or global **human** population distinguished as a more or less distinct group by genetically transmitted physical  
[www.answers.com/topic/race-1](#) - [cache] - Live

7. [Dopefish.com](#)

Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the **human race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.  
[www.dopefish.com](#) - [cache] - Open Directory

# Some Data

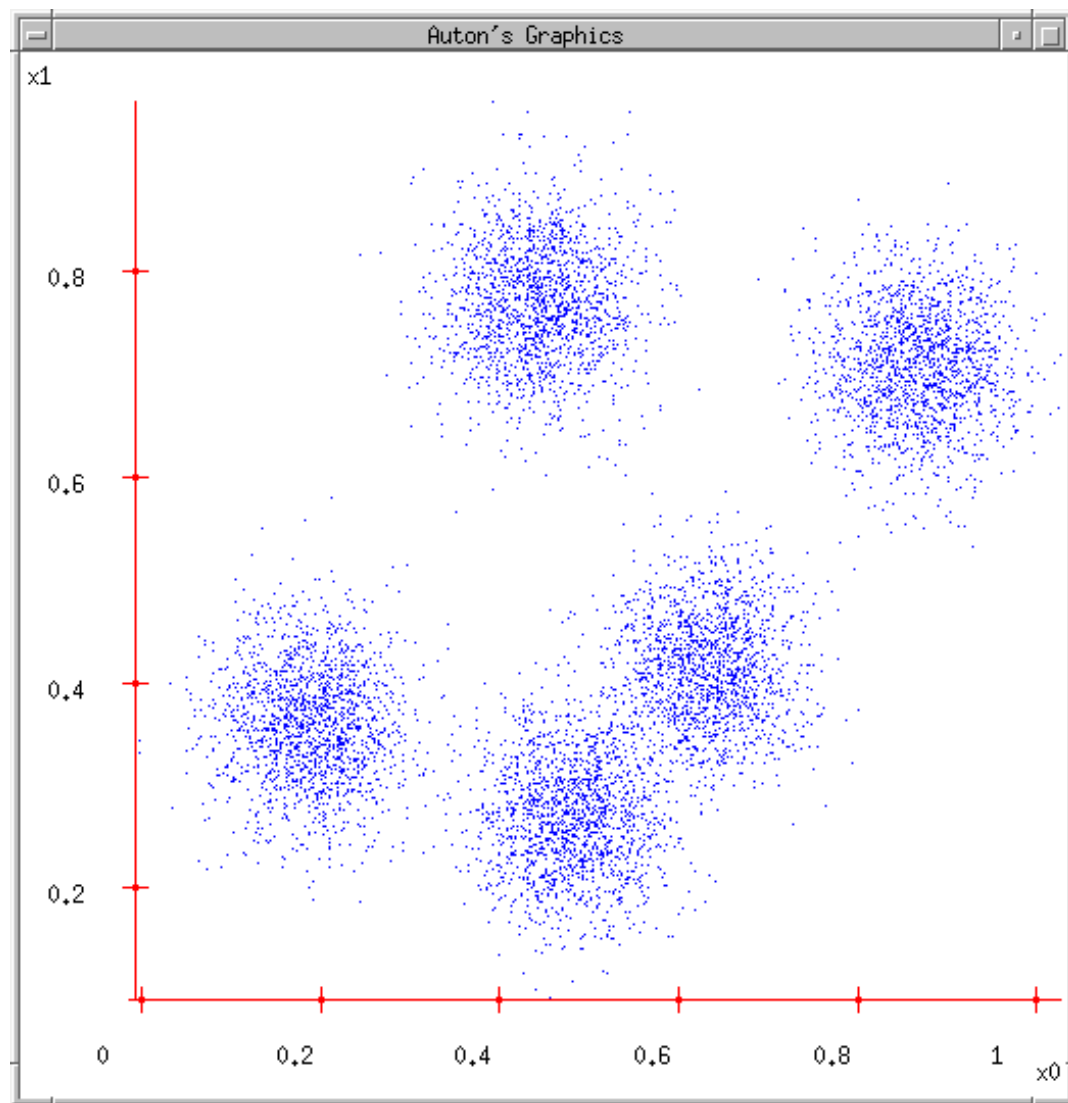
---



# *k*-means

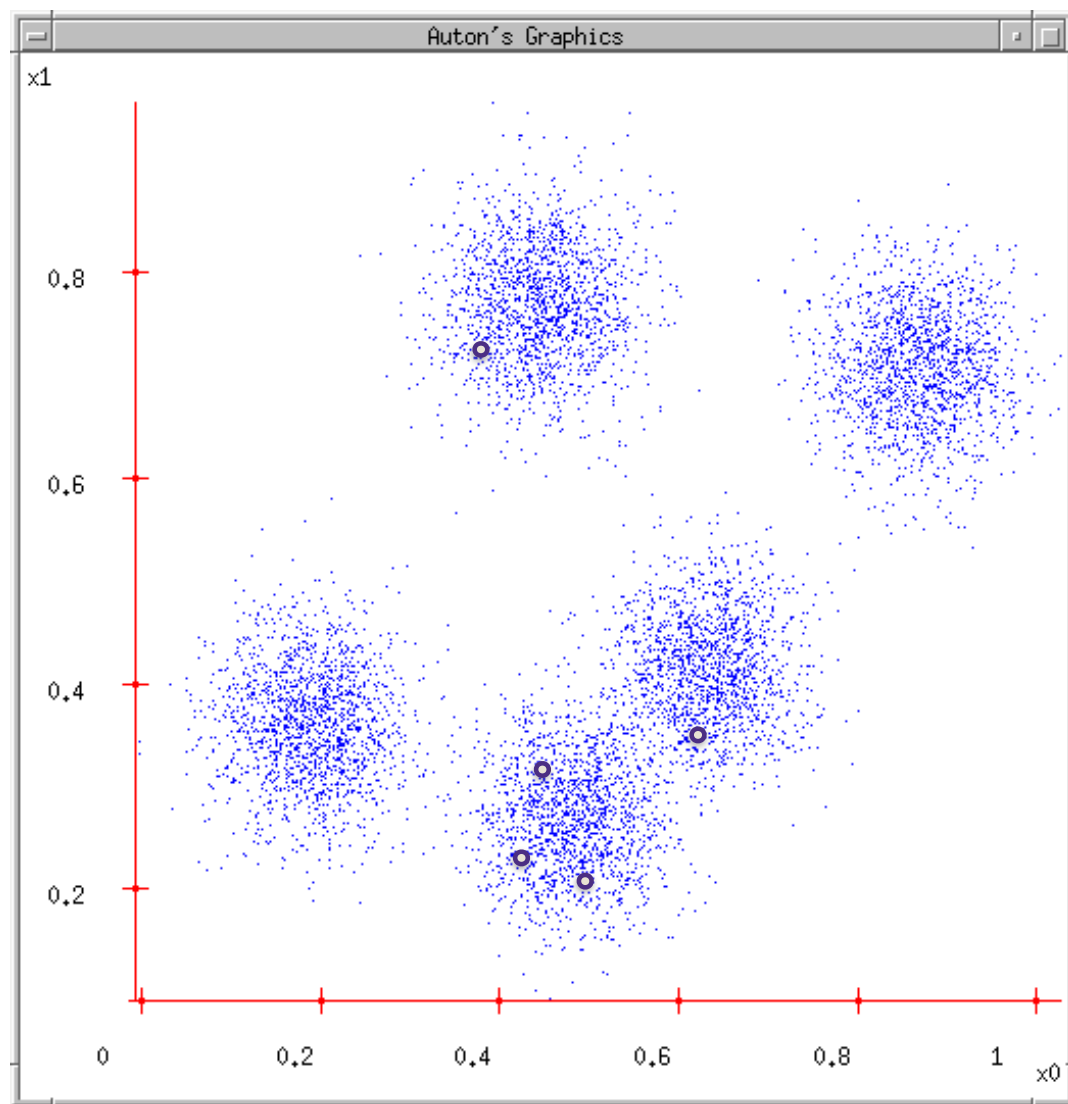
---

1. Ask user how many clusters they'd like. (e.g.  $k=5$ )



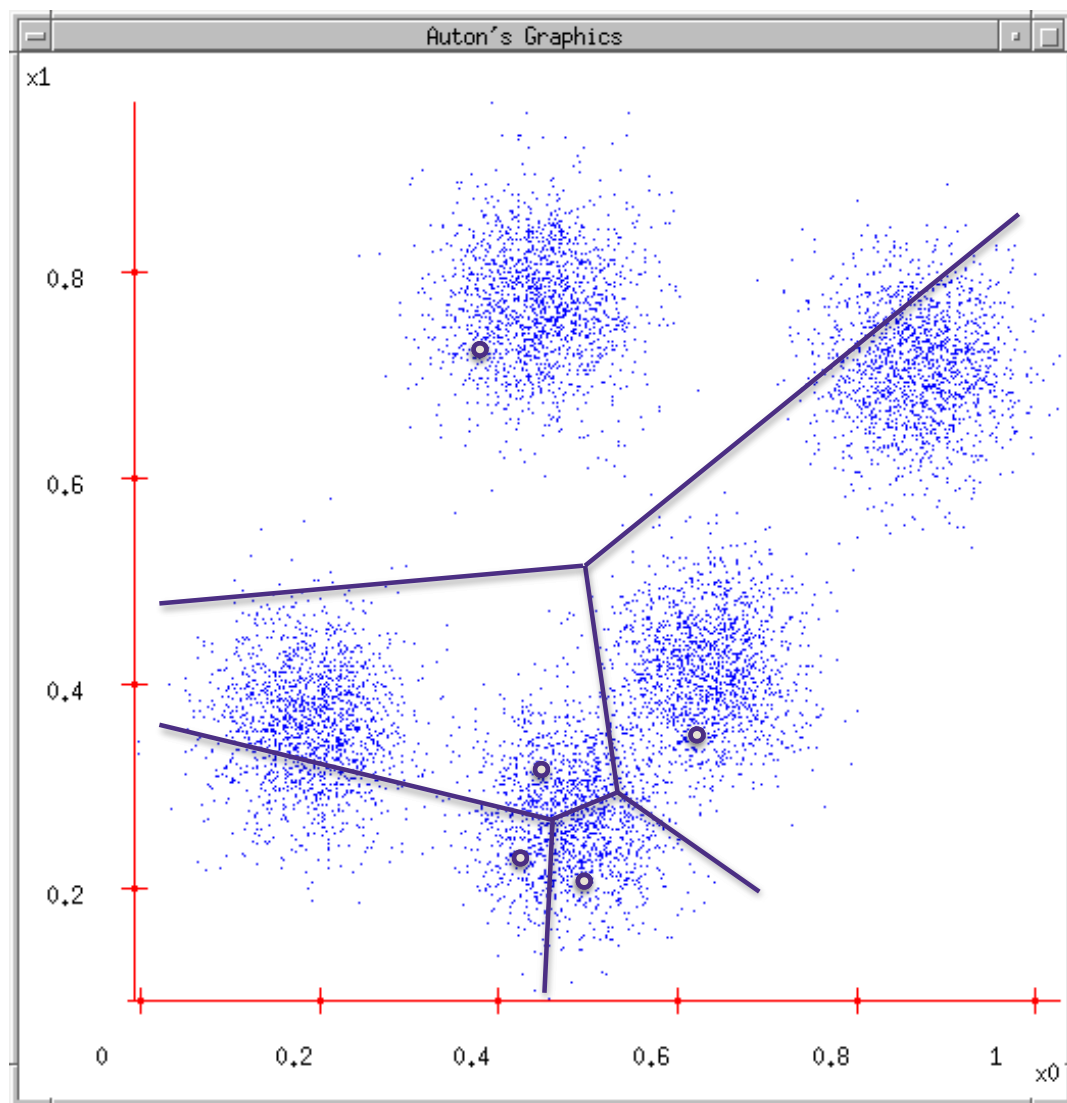
# *k*-means

1. Ask user how many clusters they'd like. (e.g.  $k=5$ )
2. Initialize: Randomly guess  $k$  cluster Center locations



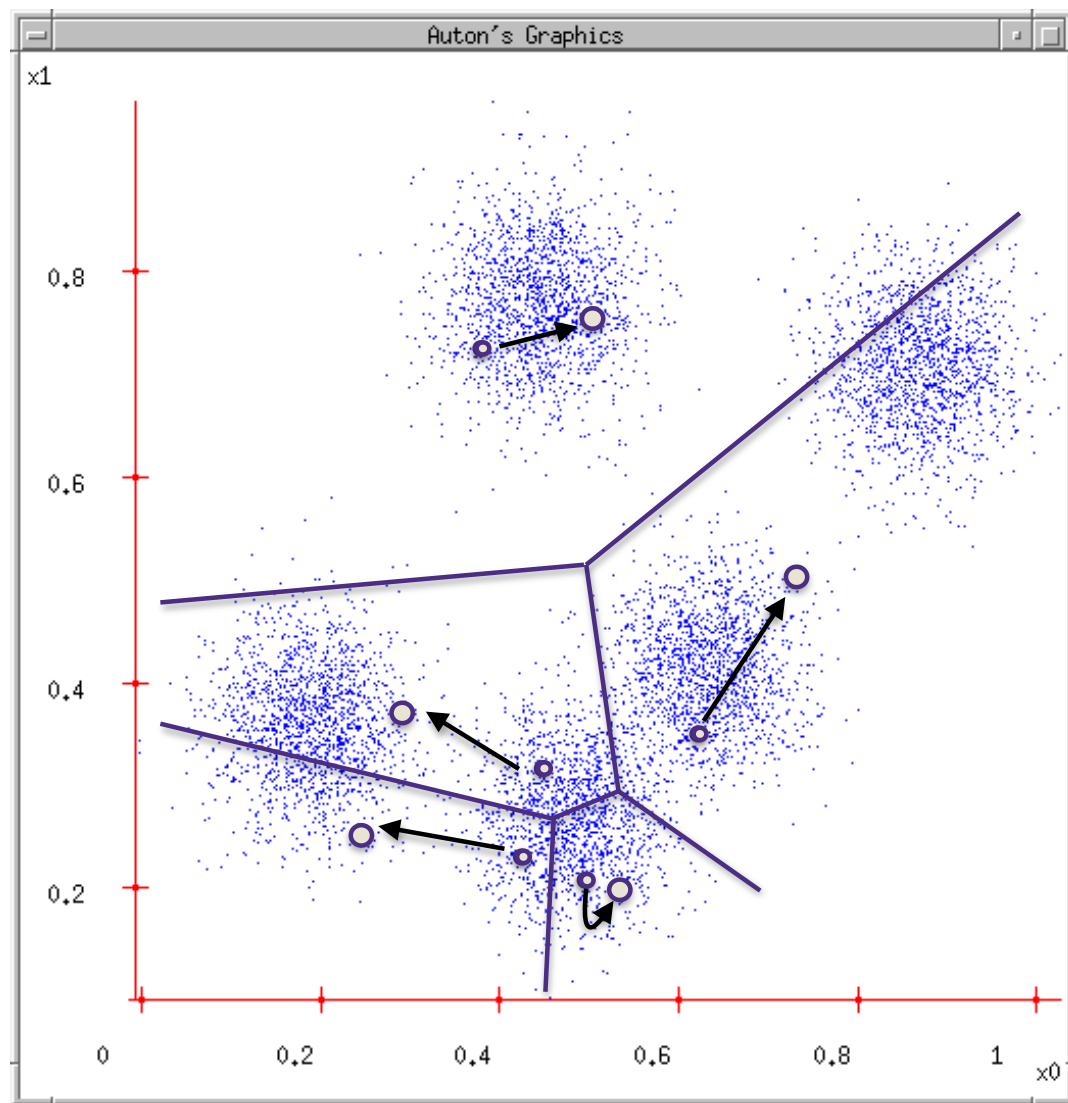
# $k$ -means

1. Ask user how many clusters they'd like. (e.g.  $k=5$ )
2. Initialize: Randomly guess  $k$  cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



# *k*-means

1. Ask user how many clusters they'd like. (e.g.  $k=5$ )
2. Initialize: Randomly guess  $k$  cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the **centroid** of the points it owns

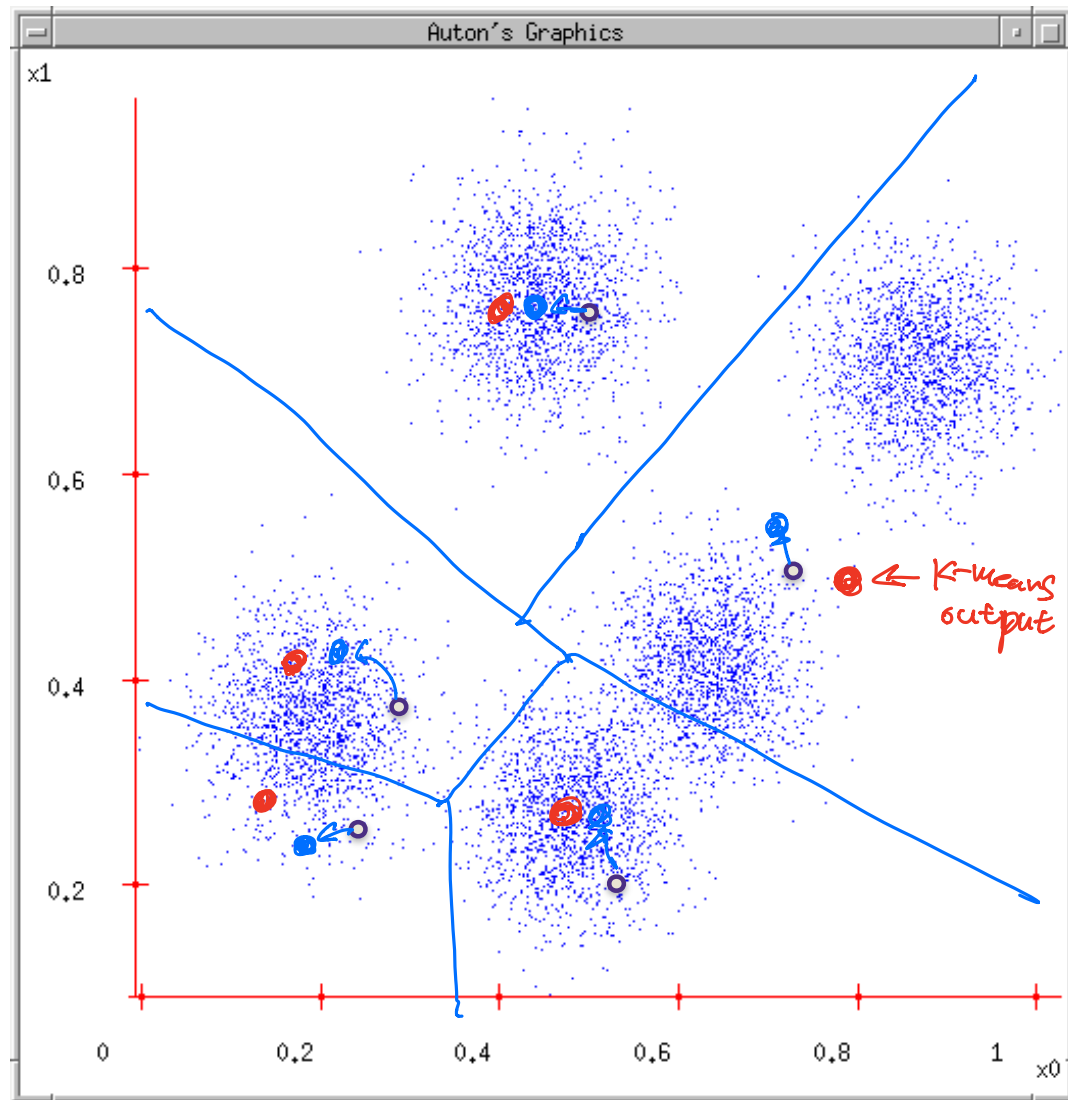




# k-means

1. Ask user how many clusters they'd like. (e.g.  $k=5$ )
2. Initialize: Randomly guess  $k$  cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the **centroid** of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!

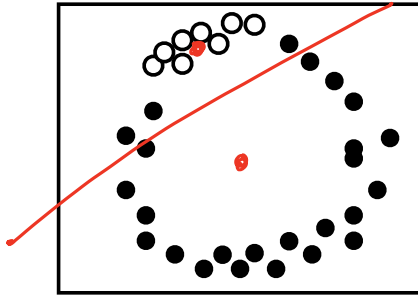
Average of all  
data points in that  
region



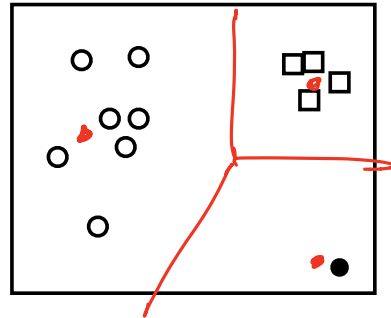
# Which one is a snapshot of a converged $k$ -means

Can this be  
output of terminated  $k$ -means

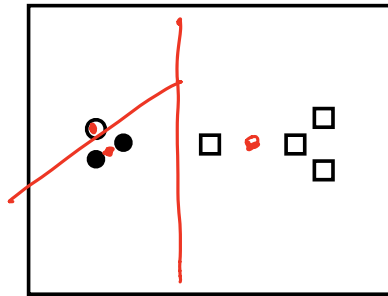
Example (a) No



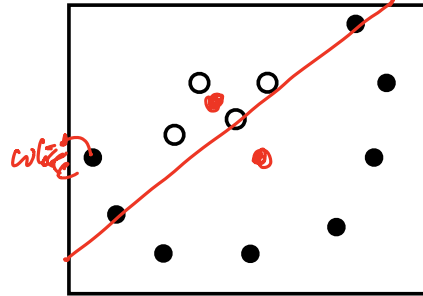
Example (b) (0)



Example (c) Yes



Example (d) ? No.



# k-means

> **Initialize**  $k$  centers (as random data points)

–  $\mu^{(0)} = (\mu_1^{(0)}, \dots, \mu_k^{(0)})$

> **Classify:** assign each point  $j \in \{1, \dots, n\}$  to nearest center:

– For each  $j \in \{1, \dots, n\}$ ,  $C^{(t)}(j) \leftarrow \arg \min_{i \in \{1, \dots, k\}} \|\mu_i^{(t)} - x_j\|_2^2$

$\uparrow$   
 $\{1, \dots, k\}$

> **Recenter:**  $\mu_i$  becomes centroid of its point:

– For each  $i \in \{1, \dots, k\}$ ,  $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C^{(t)}(j)=i} \|\mu - x_j\|_2^2$

– Equivalent to  $\mu_i \leftarrow$  average of its points!

distance

$\mu^* = \frac{1}{|C^{(t)}(i)|} \sum_{j: C^{(t)}(j)=i} x_j$

$\uparrow$  size of cluster

no tie true for other distance.

# *k*-means

> **Initialize**  $k$  centers (as random data points)

- $\mu^{(0)} = (\mu_1^{(0)}, \dots, \mu_k^{(0)})$

> **Classify**: assign each point  $j \in \{1, \dots, n\}$  to nearest center:

- For each  $j \in \{1, \dots, n\}$ ,  $C^{(t)}(j) \leftarrow \arg \min_{i \in \{1, \dots, k\}} \|\mu_i^{(t)} - x_j\|_2^2$

> **Recenter**:  $\mu_i$  becomes centroid of its point:

- For each  $i \in \{1, \dots, k\}$ ,  $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C^{(t)}(j)=i} \|\mu - x_j\|_2^2$

- Equivalent to  $\mu_i \leftarrow$  average of its points!

What does  $k$ -means do? Coordinate descent on

$$F(\underbrace{\mu_1, \dots, \mu_k}_{\mu}, \underbrace{C(1), \dots, C(n)}_C) = \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

# Does $k$ -means converge??

>  $k$ -means is trying to minimize the following objective

$$\min_{\mu \in \mathbb{R}^d, C \subseteq \{1, \dots, n\}} F(\mu, C) = \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

> Via coordinate descent:

> Fix  $\mu$ , optimize  $C$

$$\min_{C(j) \in \{1, \dots, n\}} \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

→ decomposed into  $n$  separate problems.

$$x_j \quad \min_{C(j) \in \{1, \dots, k\}} \|\mu_{C(j)} - x_j\|_2^2$$

# Does $k$ -means converge??

>  $k$ -means is trying to minimize the following objective

$$F(\mu, C) = \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

> Via coordinate descent:

> Fix  $\mu$ , optimize  $C$

$$\min_C \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

by solving  $n$  separate problems:

$$\min_{C(j)} \|\mu_{C(j)} - x_j\|_2^2$$

whose solution is

$$C(j) \leftarrow \arg \min_{i \in \{1, \dots, k\}} \|\mu_{C(j)} - x_j\|_2^2$$

# Does $k$ -means converge??

- >  $k$ -means is trying to minimize the following objective

$$F(\mu, C) = \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

- > Via coordinate descent:

- > Fix  $C$ , optimize  $\mu$

$$\min_{\mu} \sum_{j=1}^n \|\mu_{C(j)} - x_j\|_2^2$$

by solving  $k$  separate problems

$$\min_{\mu_i} \sum_{j:C(j)=i} \|\mu_i - x_j\|_2^2$$

whose solution is

$$\mu_i \leftarrow \frac{1}{|\{j : C(j) = i\}|} \sum_{j:C(j)=i} x_j$$

# Does $k$ -means converge??

---

- there is only a finite set of values that  $\{C(j)\}_{j=1}^n$  can take ( $k^n$  is large but finite)
- so there is only finite,  $k^n$  at most, values for  $\mu$  also
- each time we update them, we will never increase the objective

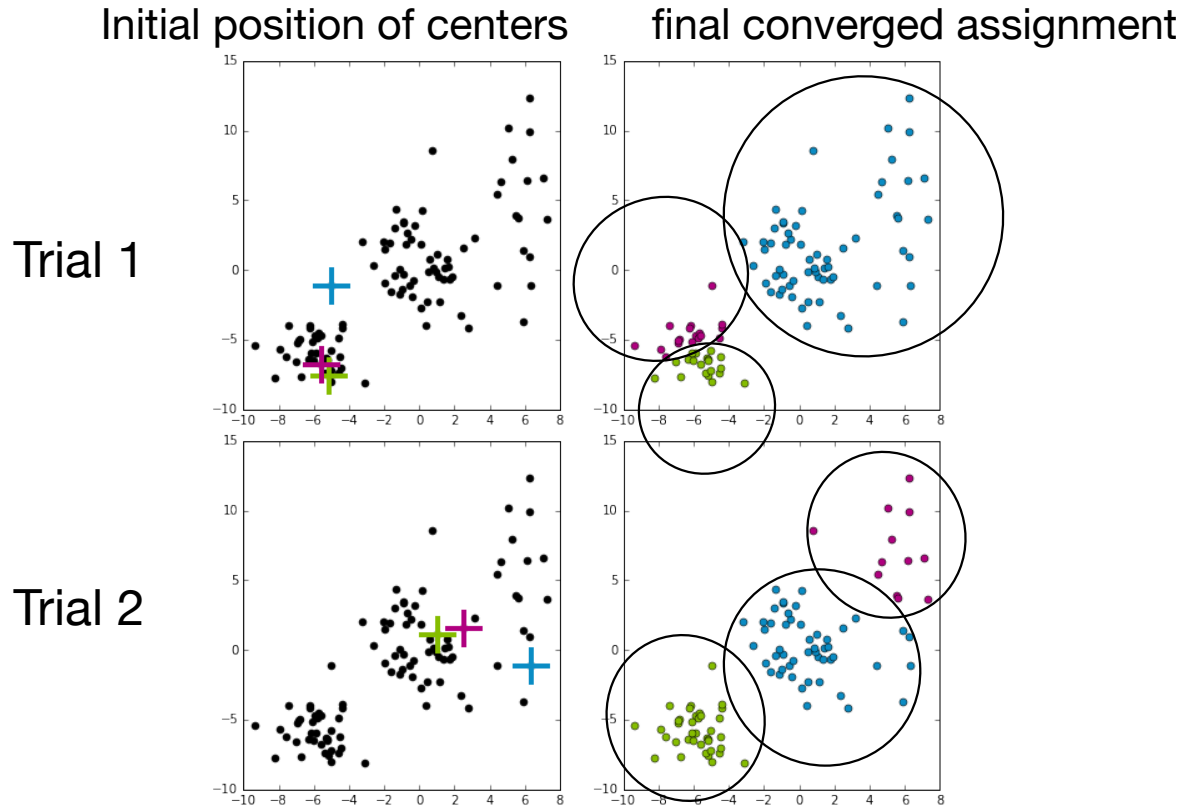
function 
$$\sum_{i=1}^k \sum_{j:C(j)=i} \|x_j - \mu_i\|_2^2$$

- the objective is lower bounded by zero
- after at most  $k^n$  steps, the algorithm must converge (as the assignments  $\{C(j)\}_{j=1}^n$  cannot return to previous assignments in the course of  $k$ -means iterations)



# Downside of $k$ -means

- the final solution depends on the initialization  
(as it is a coordinate descent on a non-convex problem)



# *k*-means ++: a smart initialization

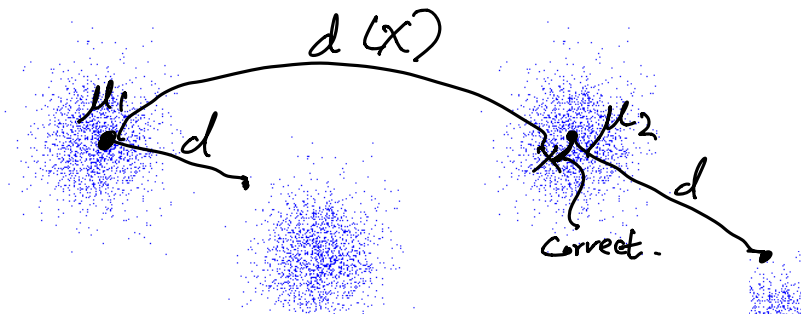
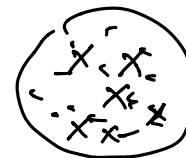
## Smart initialization:

1. Choose first cluster center uniformly at random from data points
2. Repeat  $k - 1$  times
  3. For each data point  $x_j$ , compute distance  $d_j$  to the nearest cluster center  $\approx \|\mu_1 - x_j\|_2$
  4. Choose new cluster center from amongst data points, with probability of  $x_j$  being chosen proportional to  $(d_j)^2$

$$P(\mu_2 = x_j) \propto \frac{d_j^2}{\sum_j d_j^2}$$

- apply standard K-means after the initialization
- *k*-means++ achieves *k*-means error at most a factor of  $\log k$  worse than the optimal [Arther, Vassilvitskii, 2007]

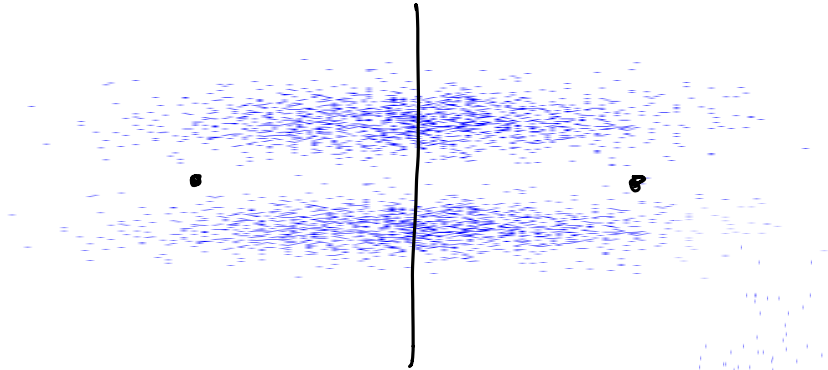
$$k=4$$



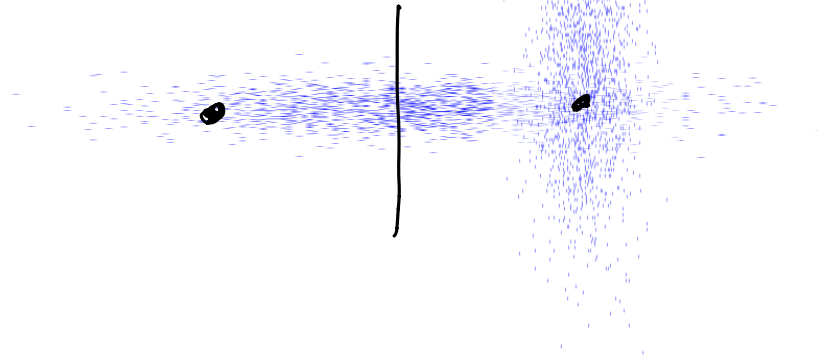
# Downside of $k$ -means

---

- Cluster shapes can be different

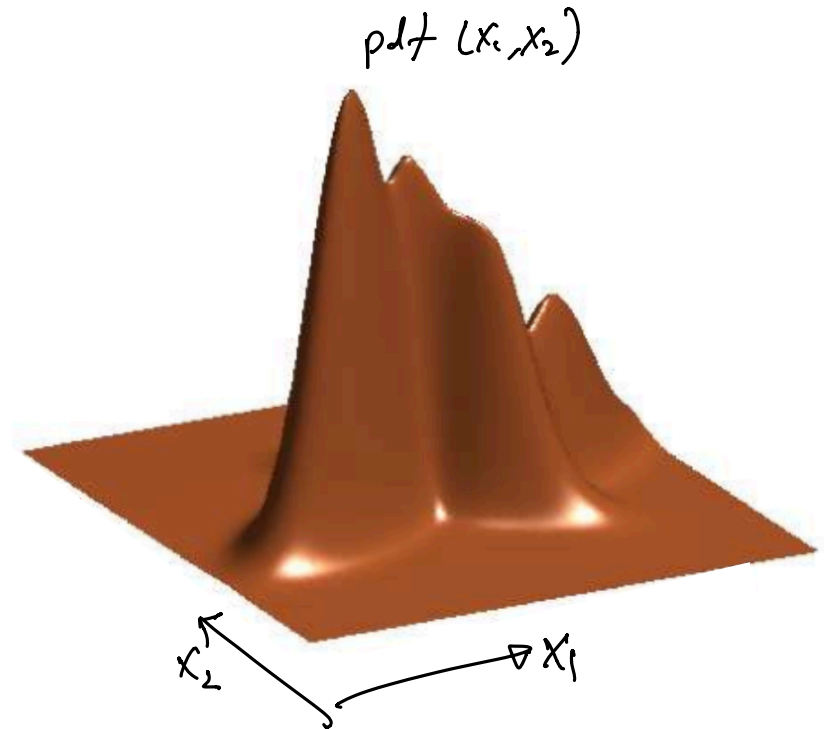
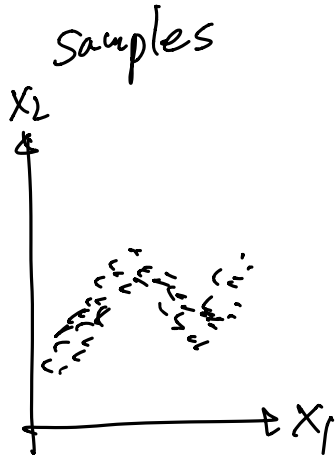


- Or clusters can have overlaps



# Solution: density estimation

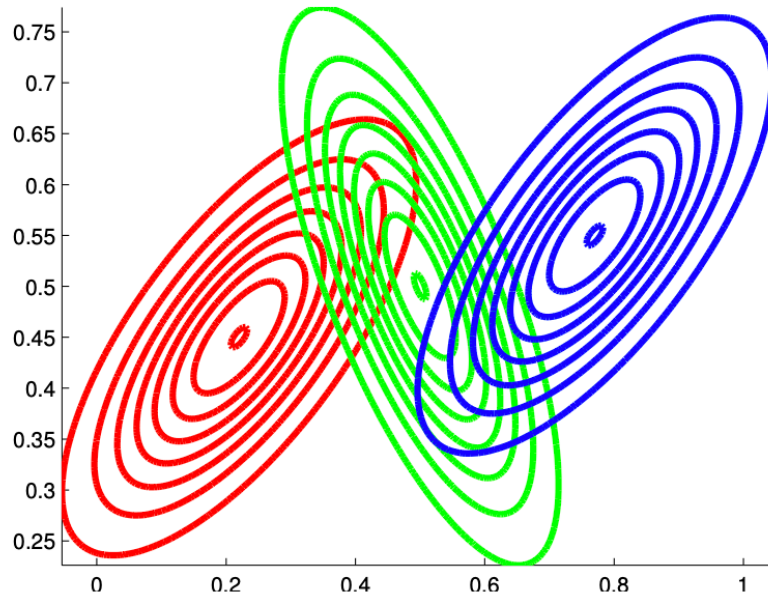
- > Estimate probability density function from  $n$  i.i.d. samples  $x_1, x_2, \dots, x_n$



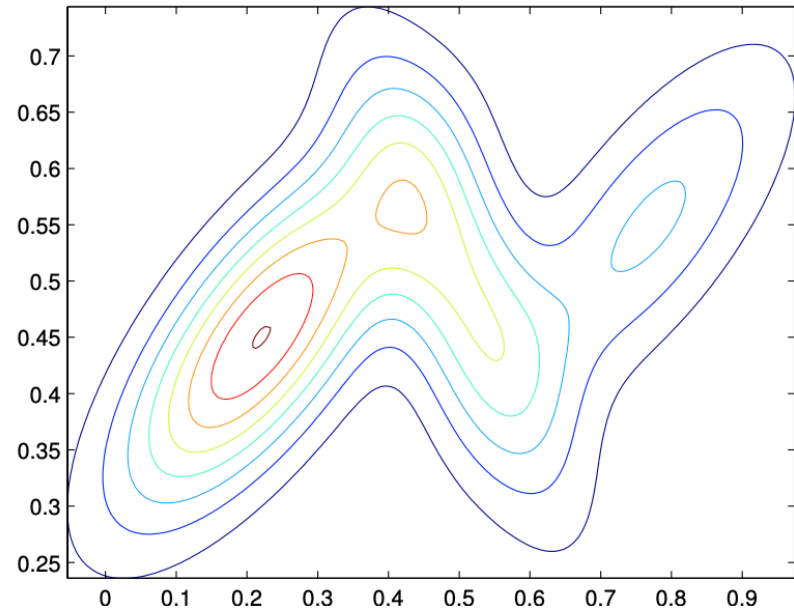
# Density as mixture of Gaussians

- > Approximate unknown density with a mixture of Gaussians

*Mixture of 3 Gaussians*



*Contour Plot of Joint Density*



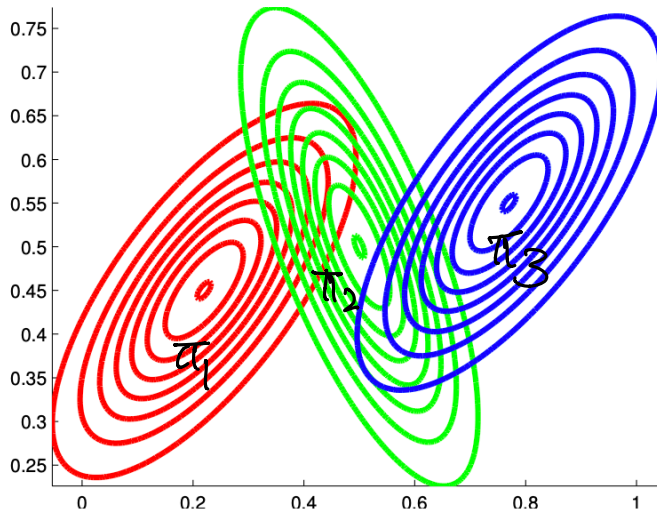
# Mixture of Gaussians

$$P(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

> Approximate unknown density with a mixture of Gaussians

$$P(x_j; \pi, \mu, \Sigma) = \sum_{(\pi_1, \pi_2, \pi_3)} \underbrace{P(x_j; \mu_1, \Sigma_1)}_{\pi_1} + \underbrace{P(x_j; \mu_2, \Sigma_2)}_{\pi_2} + \underbrace{P(x_j; \mu_3, \Sigma_3)}_{\pi_3}$$

Mixture of 3 Gaussians



$$\sum_{i=1}^k \pi_i = 1$$

$$\pi_i \geq 0$$

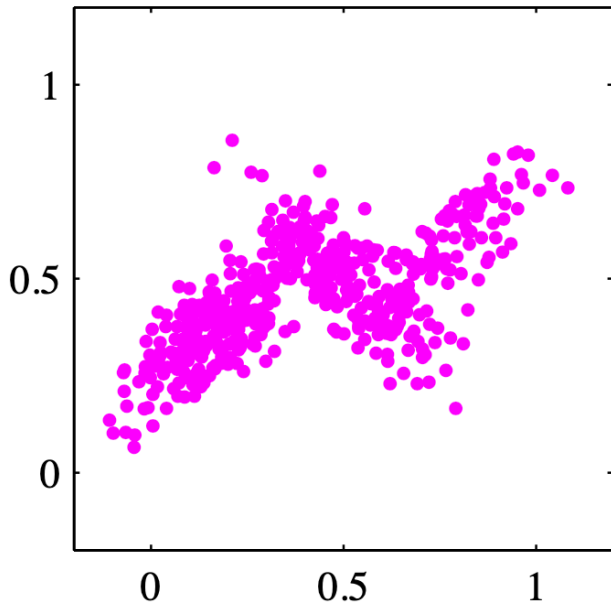
# Maximum likelihood solves clustering

$$\max_{\pi, \mu, \Sigma} \sum_{j=1}^n \log P(x_j; (\pi, \mu, \Sigma))$$

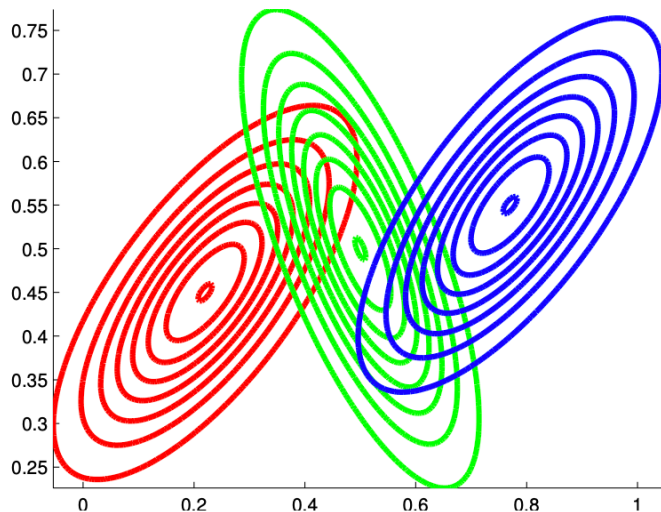
$\longrightarrow \pi^*, \mu^*, \Sigma^*$   
Algorithm

E-M (Expectation Maximization)

Our actual observations



Mixture of 3 Gaussians



# Maximum likelihood solves clustering

$$P(\underline{z}_j, x_j; \pi, \mu, \Sigma)$$

$\pi, \mu, \Sigma$  = parameters

$$\underline{z}_j \sim \pi$$

$$x_j \sim N(\mu_{\underline{z}_j}, \Sigma_{\underline{z}_j})$$

To assign clusters, we define latent cluster indicator  $z_j \in \{1, \dots, k\}$

Suppose for just now that we have  $z_j$  (true cluster indicator),  
 $z_j \in \{1, 2, 3\}$

then we have  $P(x_j; z_j, \pi, \mu, \Sigma) =$

$$\pi_{\underline{z}_j} \cdot P(x_j; \mu_{\underline{z}_j}, \Sigma_{\underline{z}_j})$$

$$\underline{z}_j = 1$$

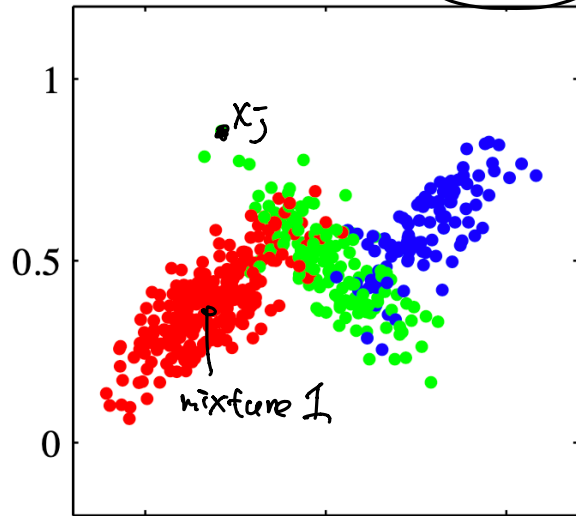
$$= \pi_1 \cdot P(x_j; \mu_1, \Sigma_1)$$

$$= P(x_j, \underline{z}_j = 1; \pi, \mu, \Sigma)$$

We can now infer the clusters  
for each sample using this formula

$$P(\underline{x}_j | \underline{z}_j = 1; \pi, \mu, \Sigma)$$

$$= P(x_j; \mu_1, \Sigma_1)$$



Complete data labeled  
by true cluster assignments



# Maximum likelihood solves clustering

But in practice we do not know  $z_j$ 's

but we can now infer the clusters, by computing the posterior probability on  $z_j$ 's

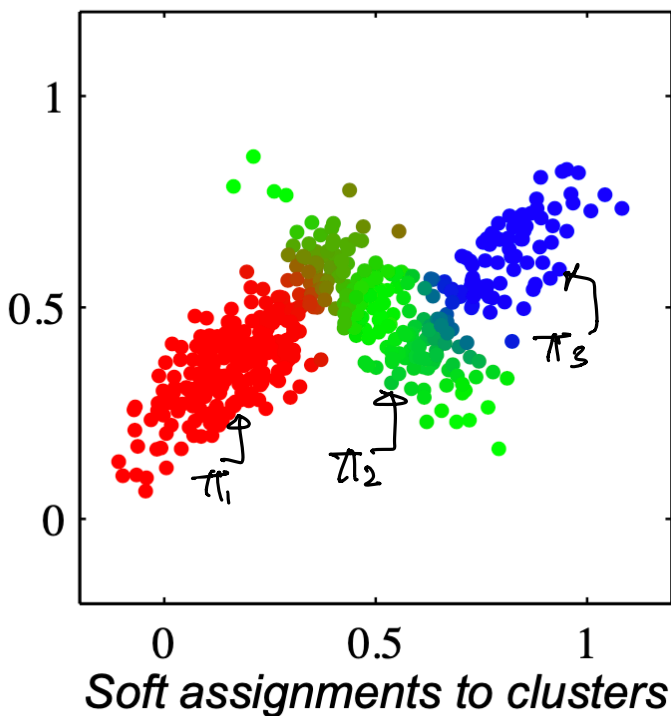
- ① pick a cluster with prob.  $(\pi_1, \pi_2, \pi_3) \rightarrow z_j$
- ② sample  $x_j$  from Gaussian from  $z_j$ -cluster.

$$r_{ji} = P(\text{sample } j \text{ belongs to cluster } i)$$

$$= P(z_j = i | x_j; \pi, \mu, \Sigma)$$

$$= \frac{\pi_i P(x_j; \mu_i, \Sigma_i)}{\sum_{i'} \pi_{i'} P(x_j; \mu_{i'}, \Sigma_{i'})}$$

find  $\boxed{\pi^*, \mu^*, \Sigma^*}$  depends on  
integration

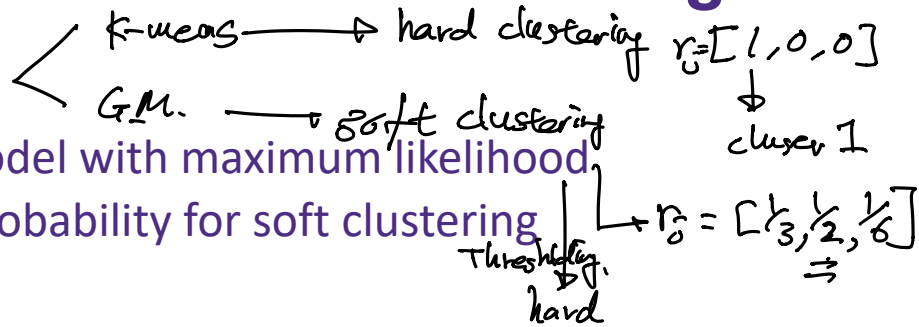


# Recap: Mixture of Gaussians for clustering

Given a set of samples

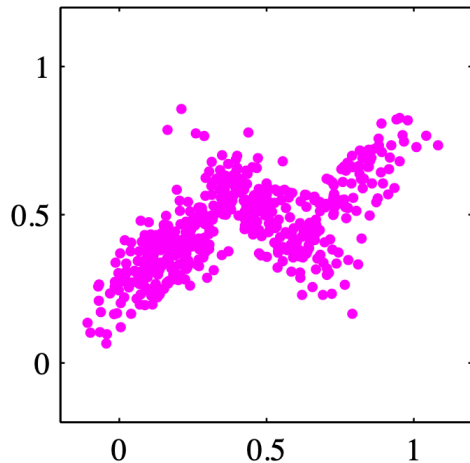
1. Fit a mixture of Gaussian model with maximum likelihood

2. Use posterior assignment probability for soft clustering

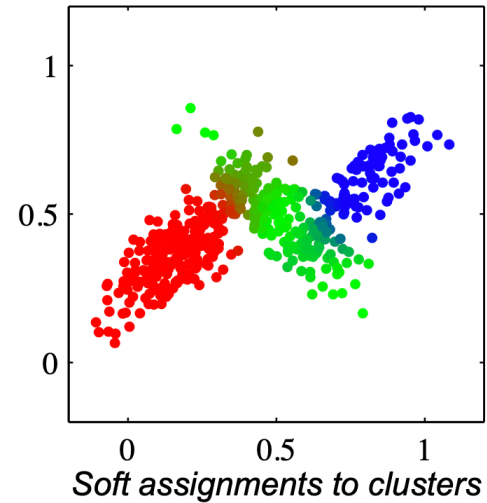
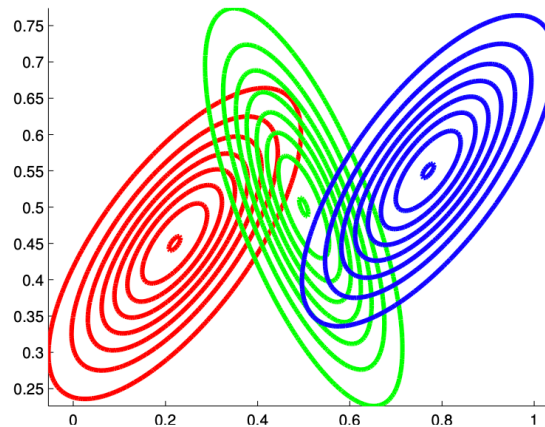


this can handle overlapping clusters, and clusters of various (oval) shapes and not just circles

Our actual observations



Mixture of 3 Gaussians



# Questions?

---