**Quiz 3:** mean is 86% and average completion time 5min 18sec!

Image credit: Tenor (Queer Eye)

# #yas

# The week ahead

- Assignment 2 is out, due on Oct 5th 11:59pm (midnight)
- Fourth round of project seminars, available Thursday, Sep 17th
- Open office hours on Thursday, 7pm to 8pm
  - https://primetime.bluejeans.com/a2m/live-event/qfsqxjec
- Quiz 4, Friday, Sep 18th 6am until Sep 19th 11:59am (noon)
  - Gaussian mixture models, hierarchical clustering, density based clustering

# Coming up soon

- Assignment 2 Early bird special → 1 complete programming question by Wed, Sep 23rd
- Touch-point 1, survey for in-person version available tonight, deliverables due Sep 28th

CS4641B Machine Learning

# Lecture 08: Gaussian Mixture Model

Rodrigo Borela ▸ rborelav@gatech.edu

Some of the slides are based on slides from Jiawei Han Chao Zhang, Barnabás Póczos and Mahdi Roozbahani

# Outline

- Overview
- Gaussian Mixture Model
- The Expectation-Maximization Algorithm

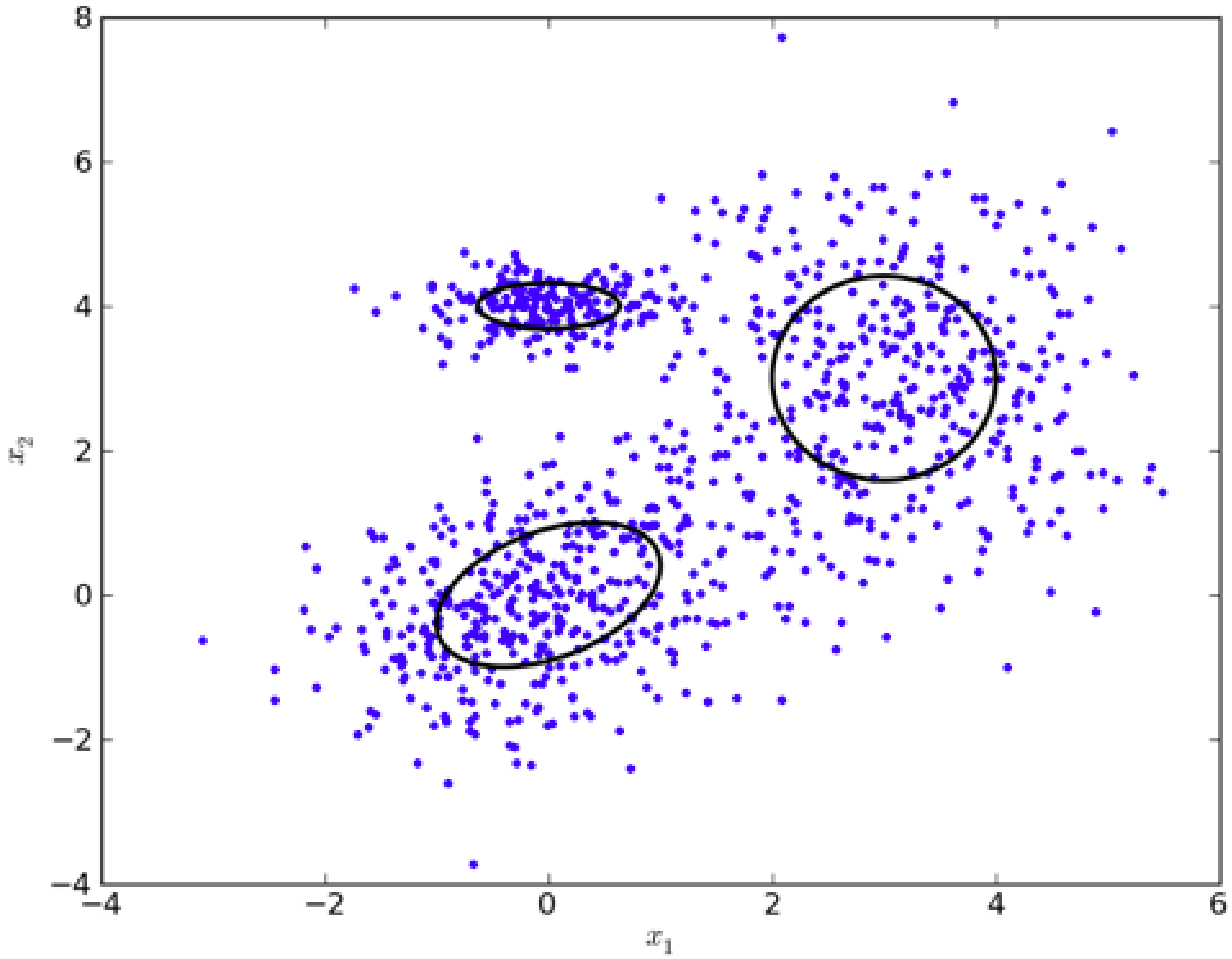
*Complementary reading: Bishop PRML – Chapter 9, Sections 9.2 through 9.3.3*

# Outline

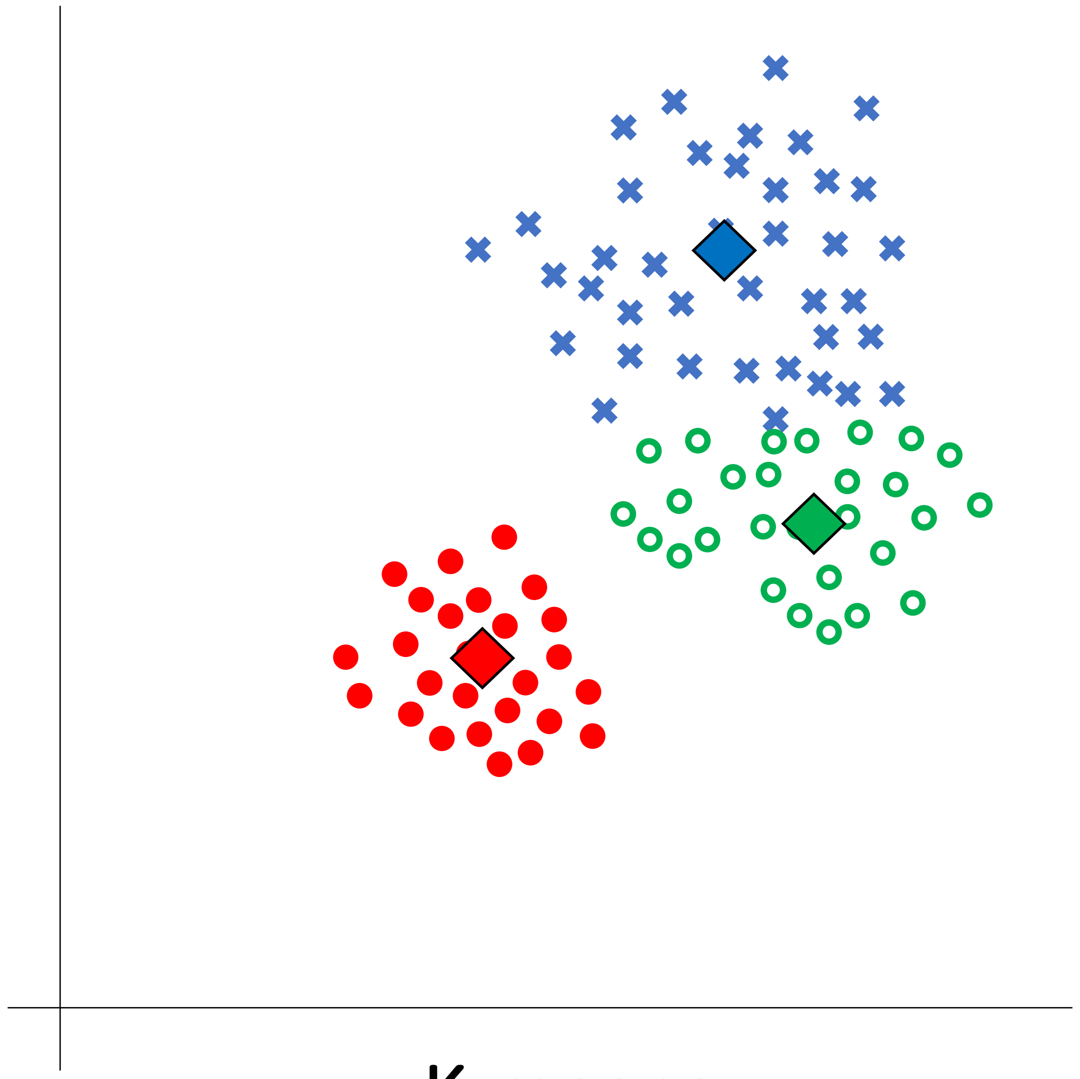
- Overview
- Gaussian Mixture Model
- The Expectation-Maximization Algorithm

# Hard clustering can be difficult

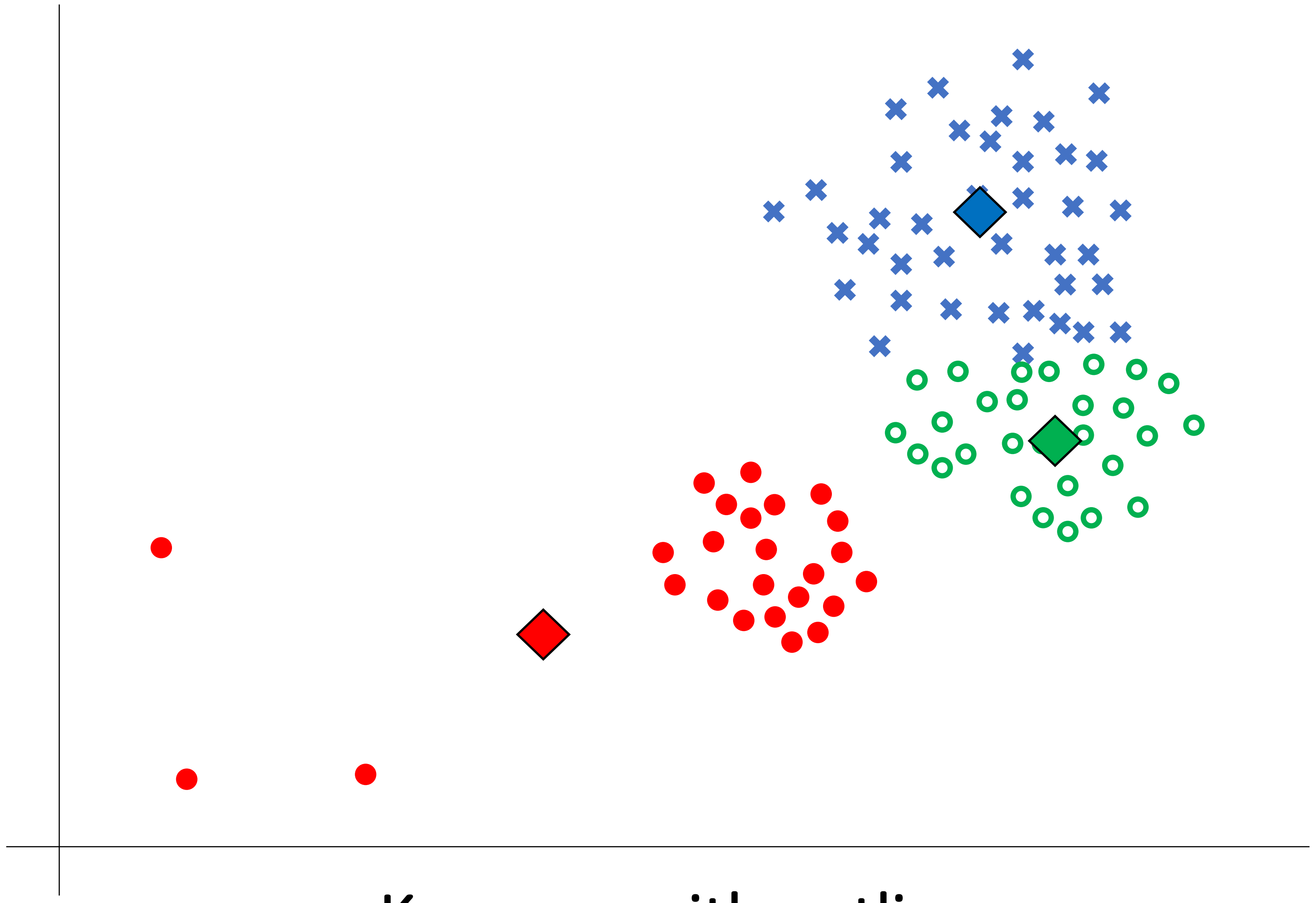
- Hard Clustering: K-Means, Hierarchical Clustering, DBSCAN



# How can we overcome some of the limitations of K-Means?

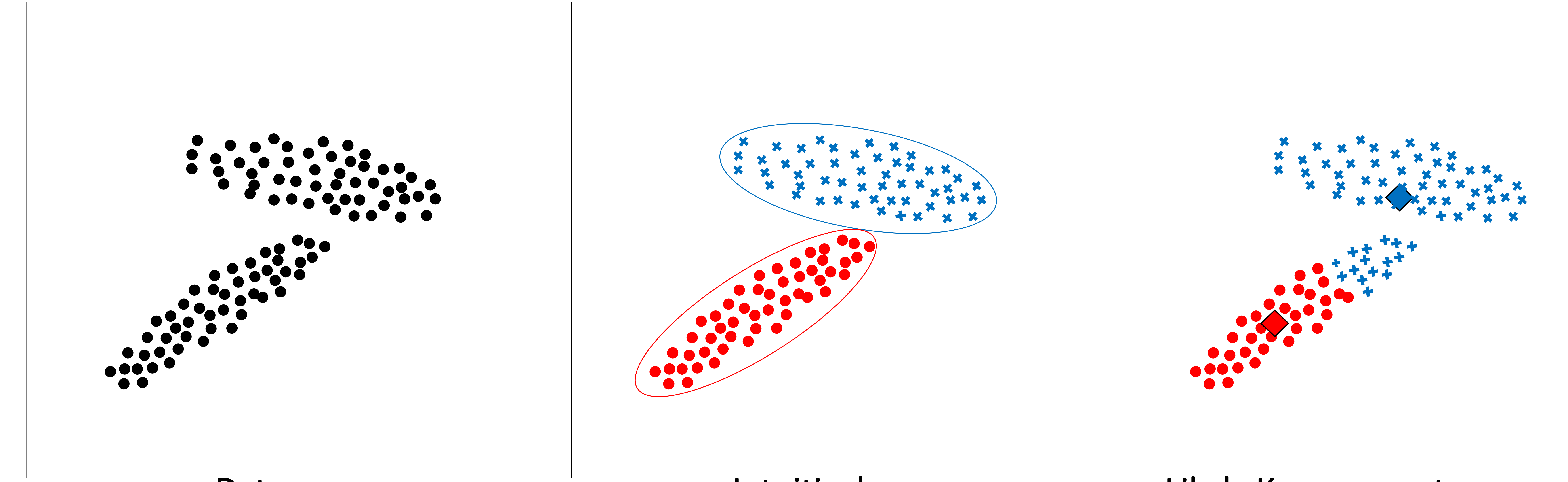


K-means



K-means with outliers

# How can we overcome some of the limitations of K-Means?



Data | Intuitively | Likely K-means outcome

# How can we overcome some of the limitations of K-Means (or hard clustering?)

- Hard cluster assignment

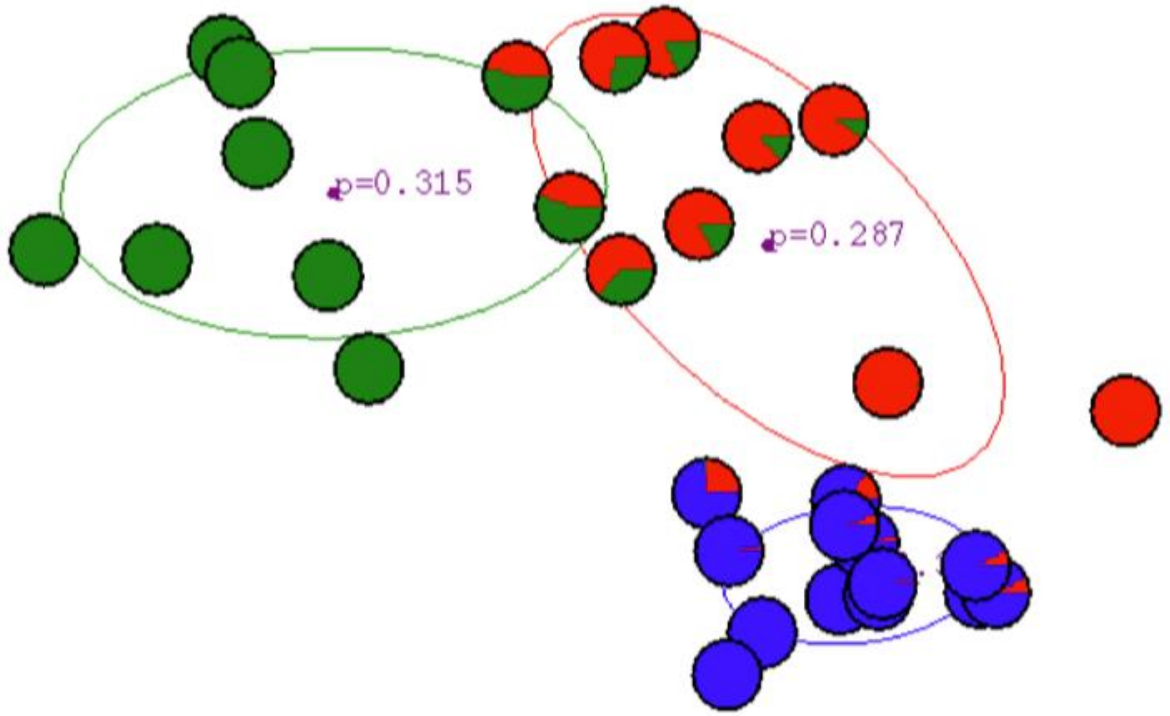
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

Cluster assignment: $\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NK} \end{bmatrix}_{N\times K}$ → $\boldsymbol{r}_n^T = [0 \quad 1 \quad \cdots \quad 0]$

# Towards soft clustering

- K-means
  - **Hard assignment:** each object belongs to only one cluster

- Mixture modeling
  - **Soft assignment:** probability that an object belongs to a cluster



# Outline

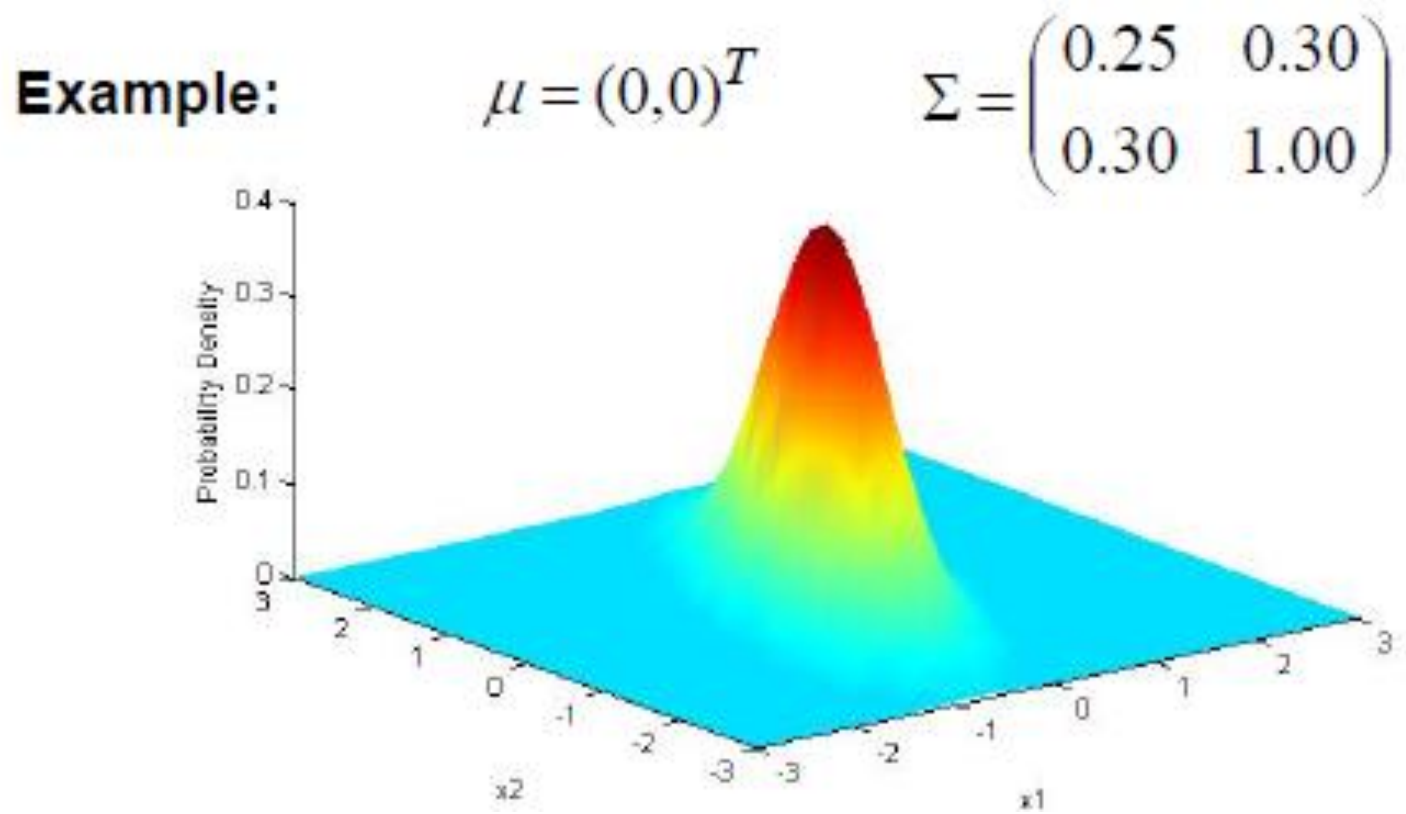
- Overview
- Gaussian Mixture Model
- The Expectation-Maximization Algorithm

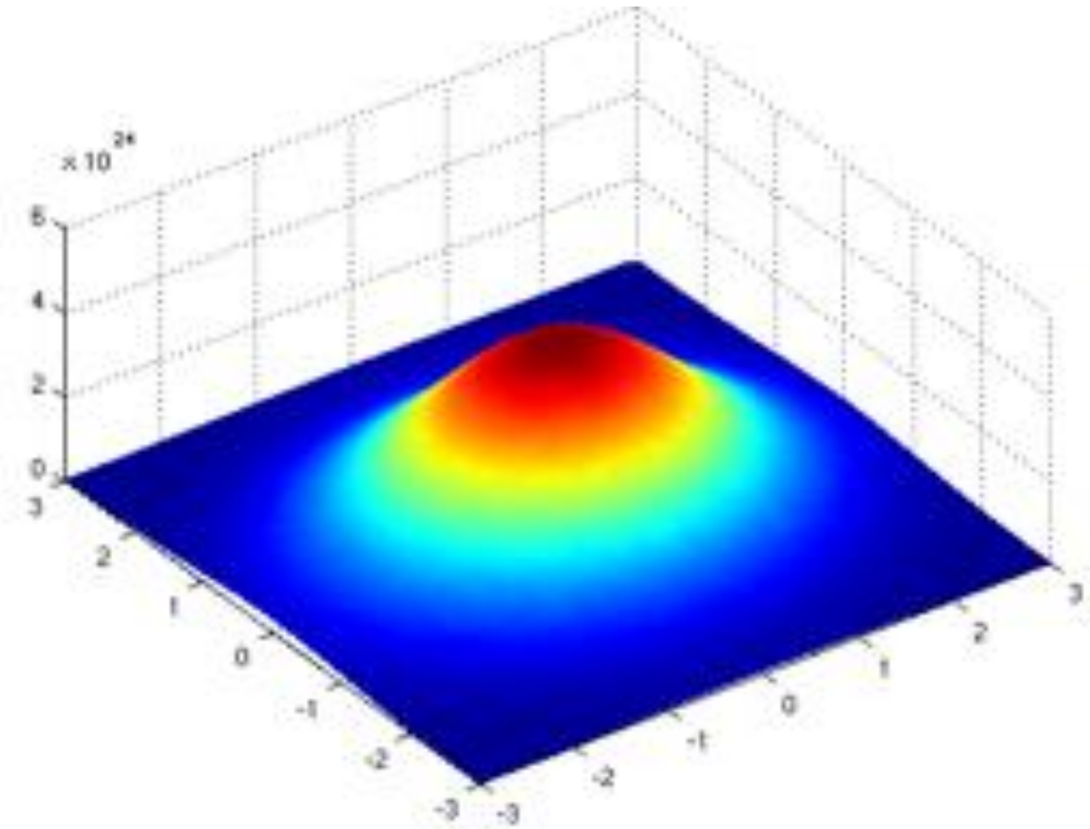
# What is a Gaussian?

- For $D$ dimensions the Gaussian distribution of a vector $\mathbf{x}^{\mathrm{T}} = [x_1, \dots, x_D]$ is defined by:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

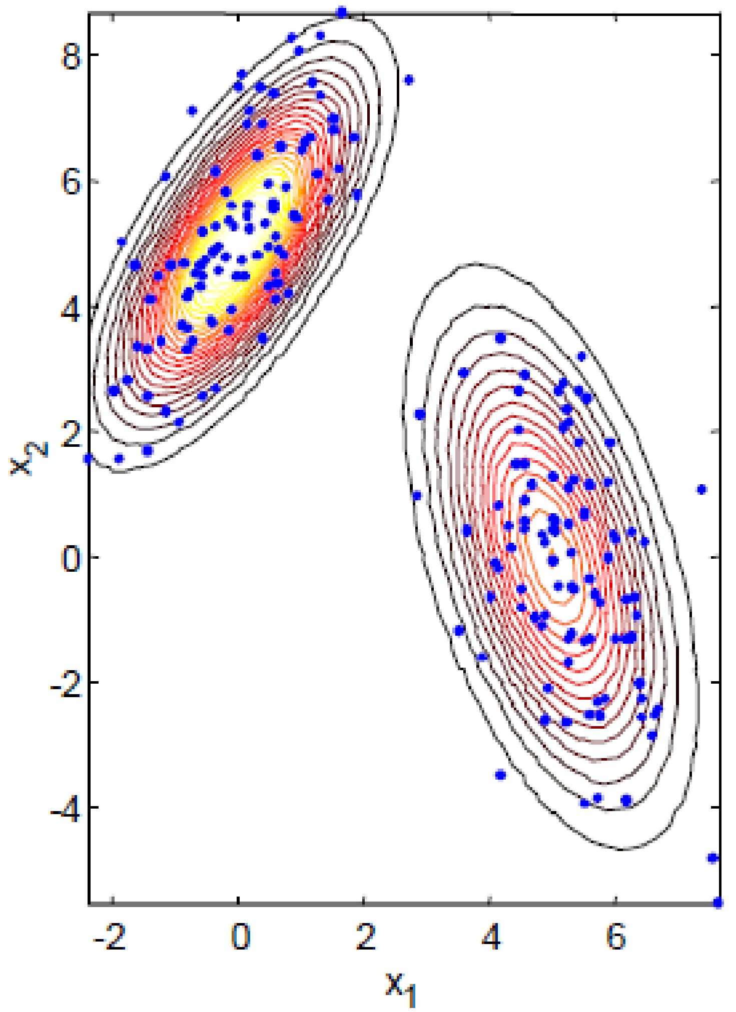
where $\boldsymbol{\mu}$ is the mean ($D$-dimensional vector) and $\boldsymbol{\Sigma}$ is the covariance matrix of the Gaussian ($D \times D$ matrix)

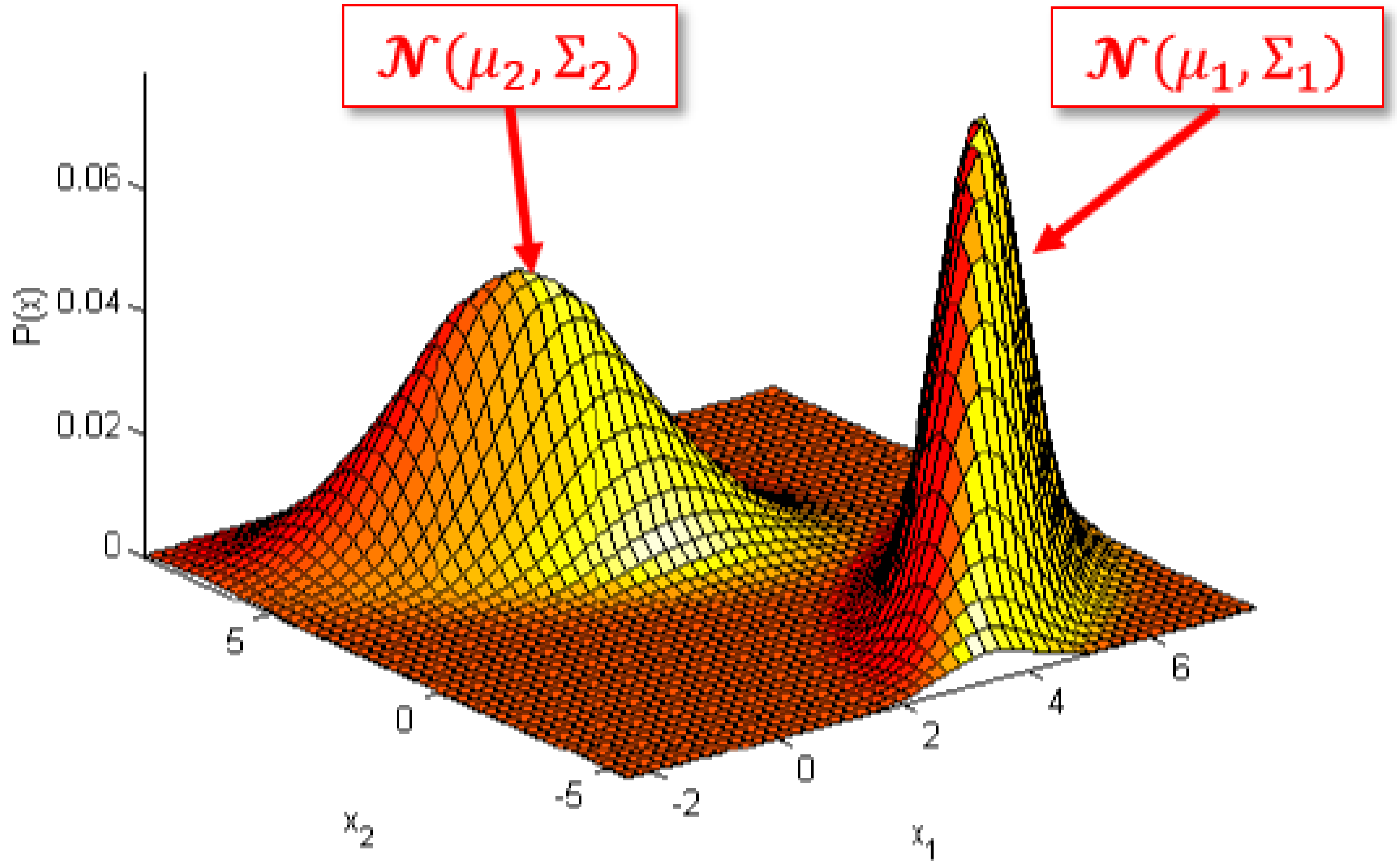


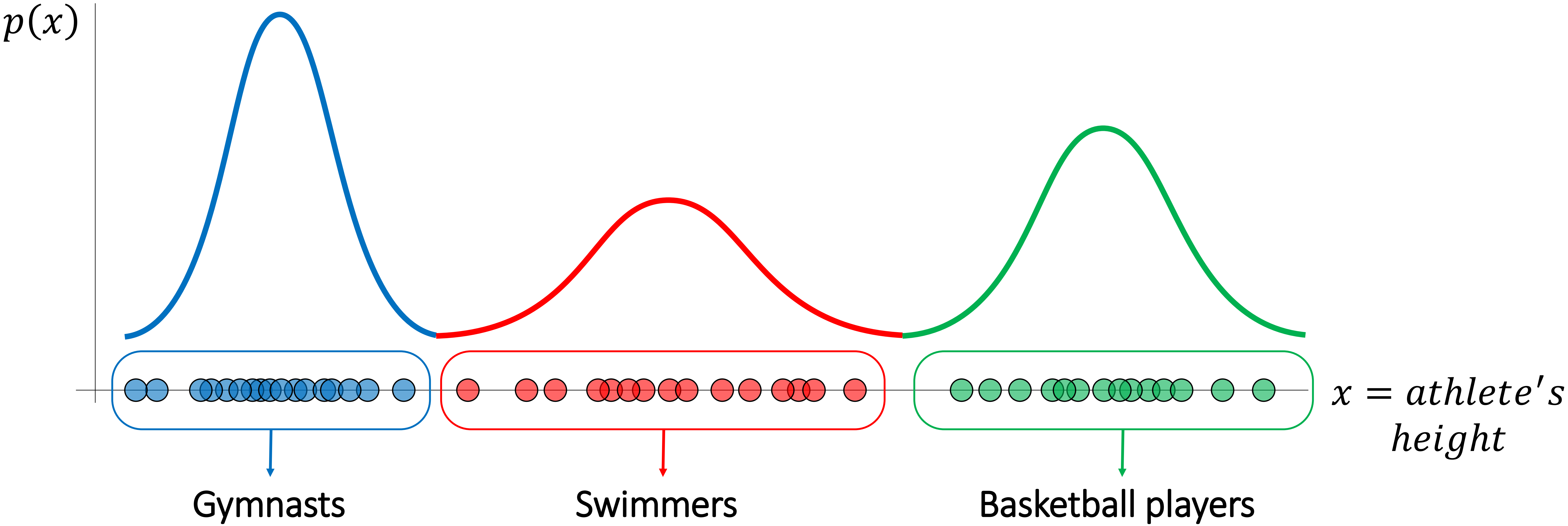
# What if our data is multimodal?

- What if we know the data consists of a few Gaussians
- What if we want to fit parametric models?

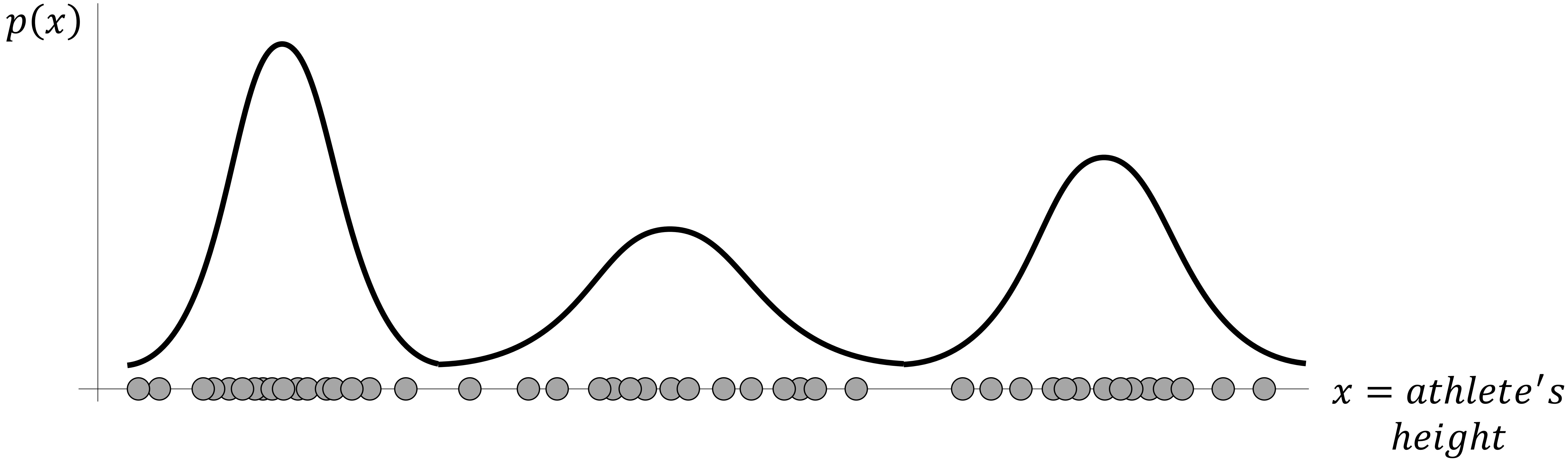


# What if our data is multimodal? Example



# What if our data is multimodal? Example



# Important observations

- Is summation of a bunch of Gaussians a Gaussian itself? Yes!
- $p(x)$ is a probability density function or it is also called a marginal distribution function.
- $p(x) =$ the density of selecting a data point from the probability density function which is created from a mixture model. Also, we know that the **area under a density function** is equal to 1.

# Mixture models

- Formally a Mixture Model is the weighted sum of a number of probability density functions where the weights are determined by a distribution:

$$p(x) = \pi_1 p_1(x) + \pi_2 p_2(x) + \cdots + \pi_K p_K(x) \rightarrow p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$$

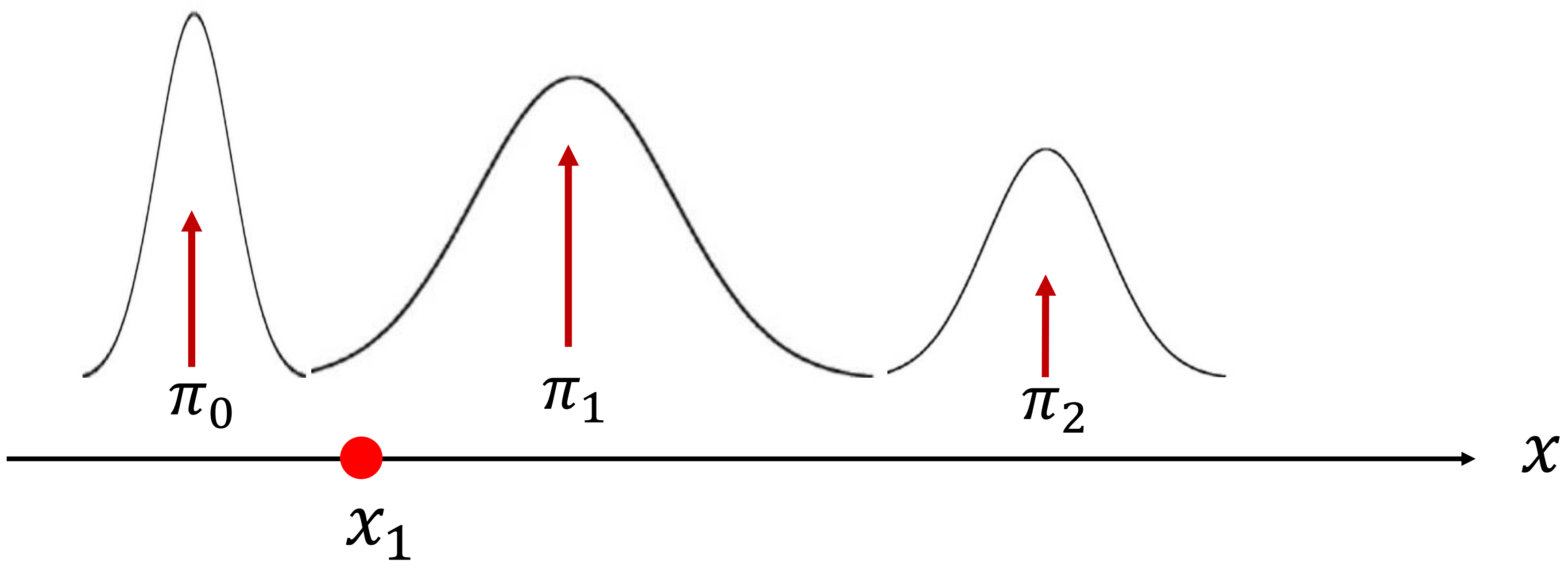
- Where $\sum_{k=1}^{K} \pi_k = 1$

$$\int p(x)dx = \int \{\pi_1 p_1(x)dx + \cdots + \pi_k p_k(x)\}dx = 1$$

$$\int p(x)dx = \pi_1 \int p_1(x)dx + \cdots + \pi_k \int p_k(x)dx = 1$$

$$\pi_1 \times 1 + \cdots + \pi_k \times 1 = 1$$

# Mixture models



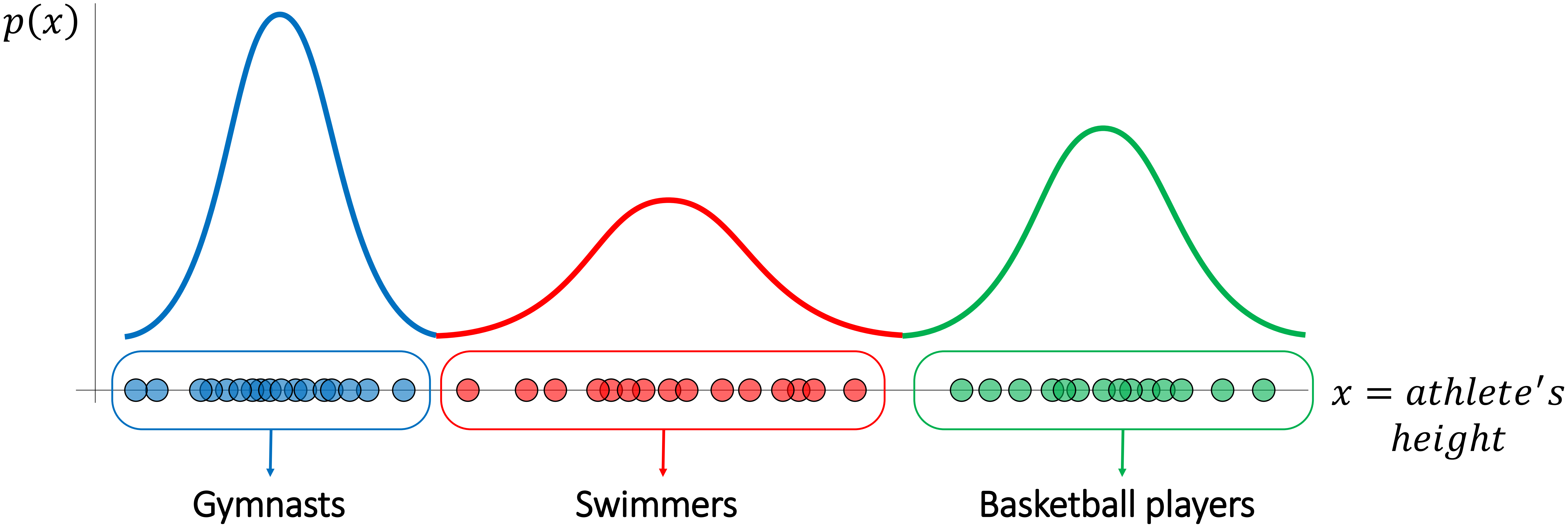

- What is the probability of a datapoint $x_1$ in each component?
- How many components we have here? 3
- How many probabilities? 3
- What is the sum value of the 3 probabilities for each datapoint? 1

# Latent variables

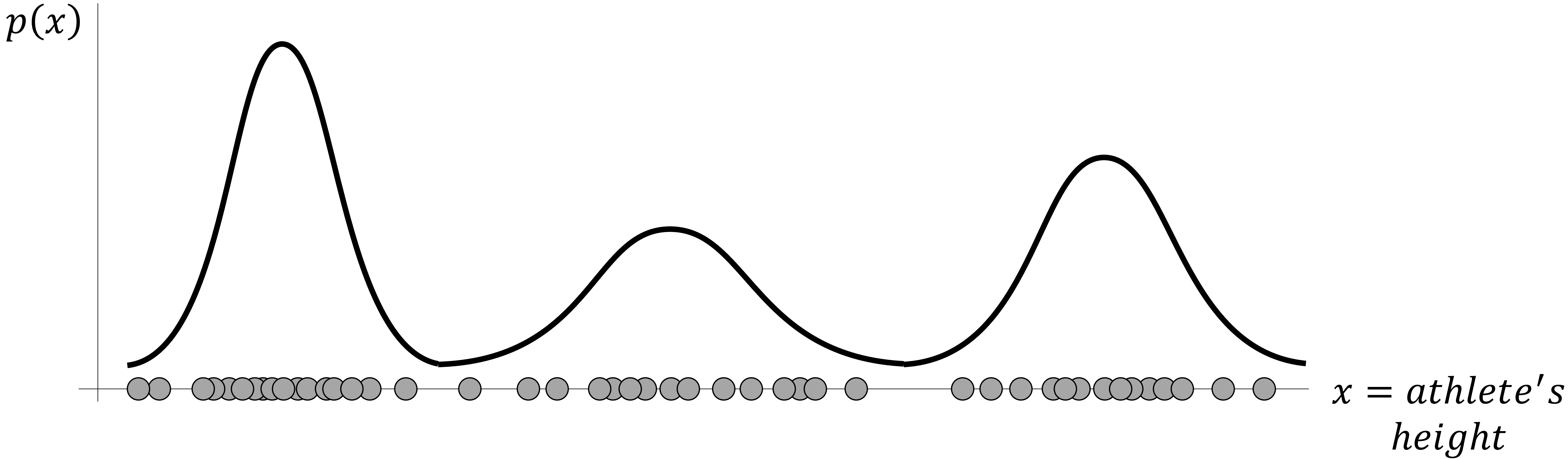
- A variable can be unobserved (latent) because:
  - It is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process.
    - e.g., speech recognition models, mixture models (soft clustering)…
  - it is a real-world object and/or phenomena, but difficult or impossible to measure
    - e.g., the temperature of a star, causes of a disease, evolutionary ancestors …
  - it is a real-world object and/or phenomena, but sometimes wasn’t measured, because of faulty sensors, etc.

- **Discrete latent variables** can be used to partition/cluster data into sub-groups.
- **Continuous latent variables** (factors) can be used for dimensionality reduction (factor analysis, etc).

# Latent variables



# Latent variables



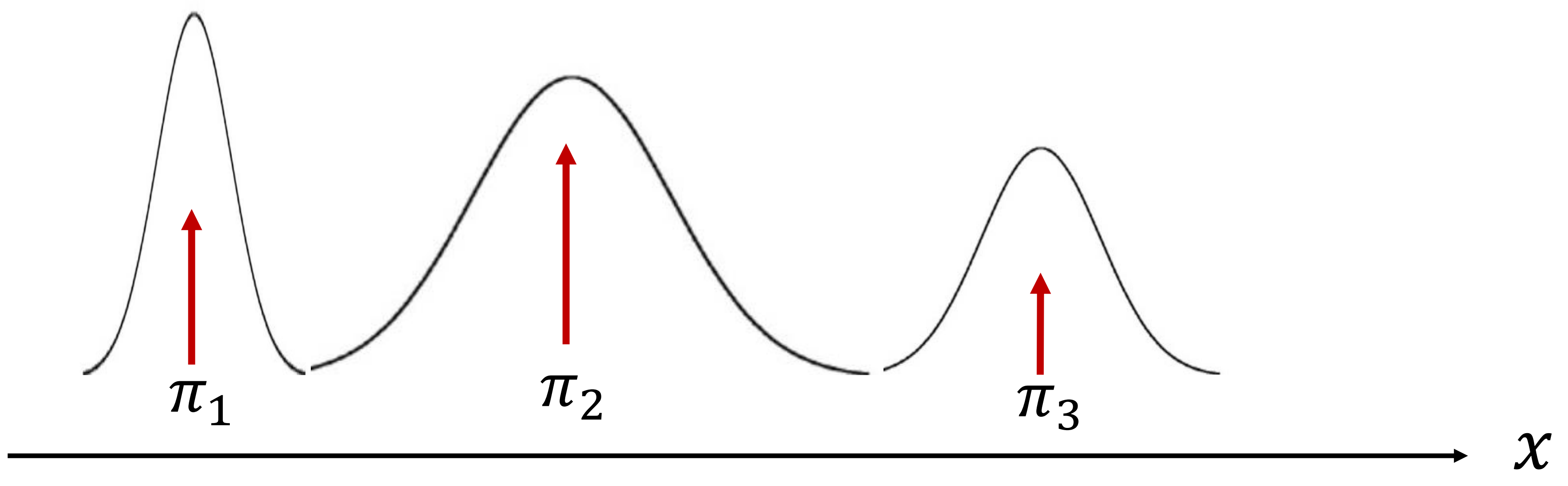

The latent variable becomes the Olympic sport from which we sampled the athlete’s heights

# Mixtures of Gaussians

- What is the probability of picking a mixture component (Gaussian model)= $p(z) = \pi_i$
- Picking data from that specific mixture component = $p(x|z)$
- $\mathbf{z}$ is latent, we observe $x$, but $\mathbf{z}$ is hidden

$p(x, \mathbf{z}) = p(x|\mathbf{z})p(\mathbf{z})$ → Generative model, joint distribution

$$p(x, \mathbf{z}) = \mathcal{N}(x|\mu_k, \sigma_k^2)\pi_k$$



# Latent variable representation

- A variable can be unobserved (latent) because:

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}, z_k) = \sum_k p(z_{nk})p(\mathbf{x}|z_{nk}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(z_k = 1) = \pi_k \rightarrow p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \rightarrow p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

- Why having the latent variable? The distribution that we can model using a mixture of Gaussian components is much more expressive than what we could have modeled using a single component.

# Inferring cluster membership

- We have representations of the joint $p(\mathbf{x}, z_k)$ and the marginal, $p(\mathbf{x})$
- The conditional of $p(z_k|\mathbf{x})$ can be derived using Bayes rule
- The responsibility that a mixture component takes for explaining an observation $x$.

$$\gamma(z_k) = p(z_k|\mathbf{x}) = \frac{p(z_k)p(\mathbf{x}|z_k)}{\sum_{j=1}^{K} p(z_j)p(x|z_j)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}})}$$

# How to calculate the probability of datapoints in the first component?

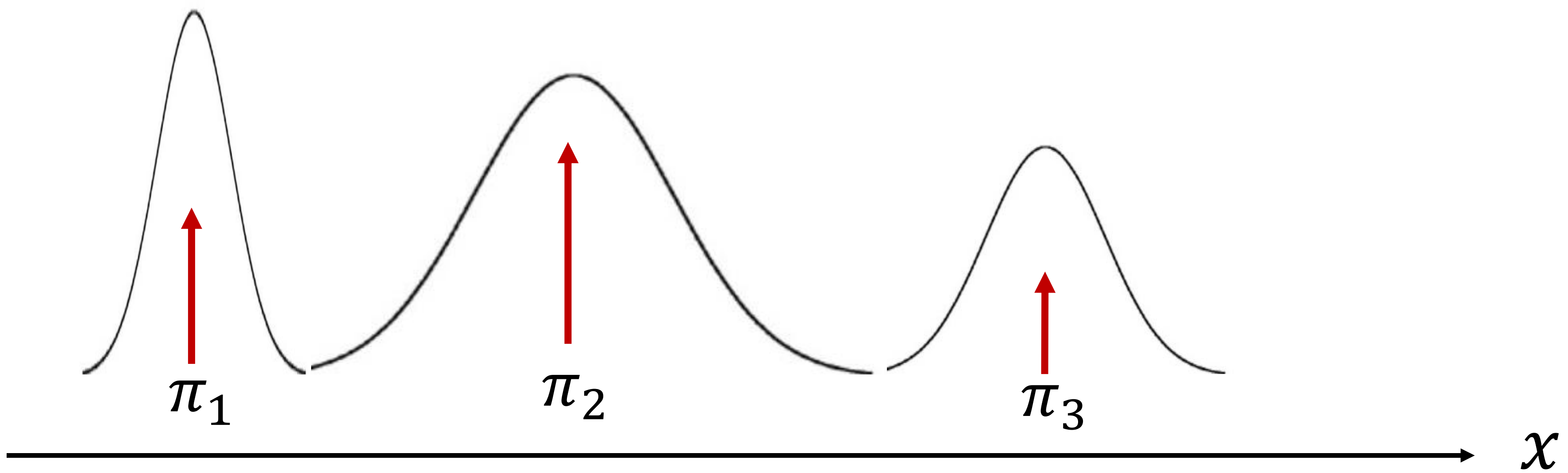
- Let's calculate the responsibility of the first component among the rest. Let's call that $\tau_0$

$$\gamma(z_1 = 1) = \frac{\mathcal{N}(x|\mu_1, \sigma_1^2)\pi_1}{\mathcal{N}(x|\mu_1, \sigma_1^2)\pi_1 + \mathcal{N}(x|\mu_2, \sigma_2^2)\pi_2 + \mathcal{N}(x|\mu_3, \sigma_3^2)\pi_3}$$

$$\gamma(z_1 = 1) = \frac{p(x|z_1)p(z_1)}{p(x|z_1)p(z_1) + p(x|z_2)p(z_2) + p(x|z_3)p(z_3)}$$

$$\gamma(z_1 = 1) = \frac{p(x, z_1)}{\sum_{k=1}^{k=3} p(x, z_k)} = \frac{p(x, z_1)}{p(x)} = p(z_1|x)$$

- Given a datapoint $x$, what is probability of that datapoint in component 1
- If I have 100 datapoints and 3 components, what is the size of $\gamma$?



# What are the GMM parameters?

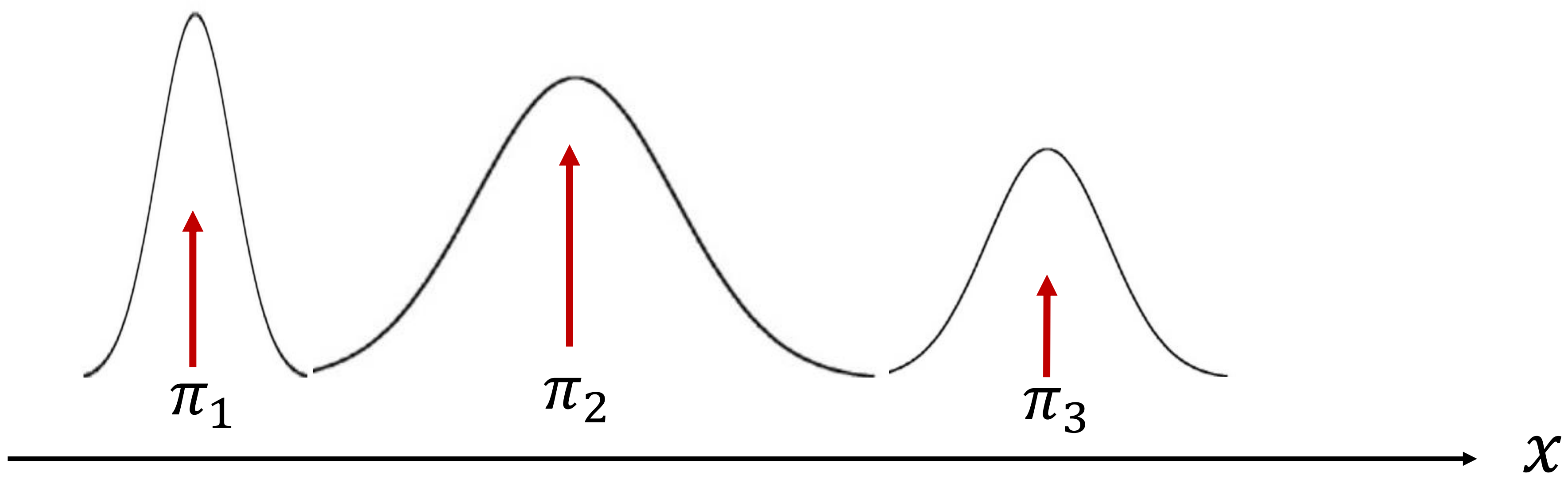
- Mean $\mu_k$, variance $\sigma_k^2$ and priors $\pi_k$ (1D Gaussian distribution)
- **Marginal probability distribution**

$$p(\mathrm{x}) = \sum_k p(x, z_k) = \sum_k p(x|z_k)p(z_k\ ) = \sum_k \mathcal{N}(x|\mu_k, \sigma_k^2)\pi_k$$

$p(z_k) = \pi_k$ Select a mixture component with probability $\pi_k$

$$p(x|z_k) = \mathcal{N}(x|\mu_k, \sigma_k^2)$$

- Sample from that component's Gaussian



# Well, we don't know $\pi_k, \mu_k, \Sigma_k$

- We can use maximum likelihood estimation (MLE) to solve the problem.

$$p(\mathbf{x}) = \sum\nolimits_k p(\mathbf{x}, z_k) = \sum\nolimits_k p(z_k)p(\mathbf{x}|z_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Let's identify a likelihood function, why?
- Because we use likelihood function to optimize the probabilistic model parameters!

$$\arg\max p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

# Maximum likelihood of a GMM

- Optimization of means

$$\ln p(x|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

$$\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)} \Sigma_k^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_k) = 0$$

$$\sum_{n=1}^{N} \gamma(z_{nk})\Sigma_{\mathrm{k}}^{-1}(x_k - \mu_k) = 0$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

# Maximum likelihood of a GMM

- Optimization of covariance

$$\ln p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

# Maximum likelihood of a GMM

- Optimization of mixing term

$$\ln p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n |\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}\left(\mathbf{x}_n |\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} + \lambda = 0$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

# Maximum likelihood of a GMM

- Defining $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{N_k}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{N_k}$$

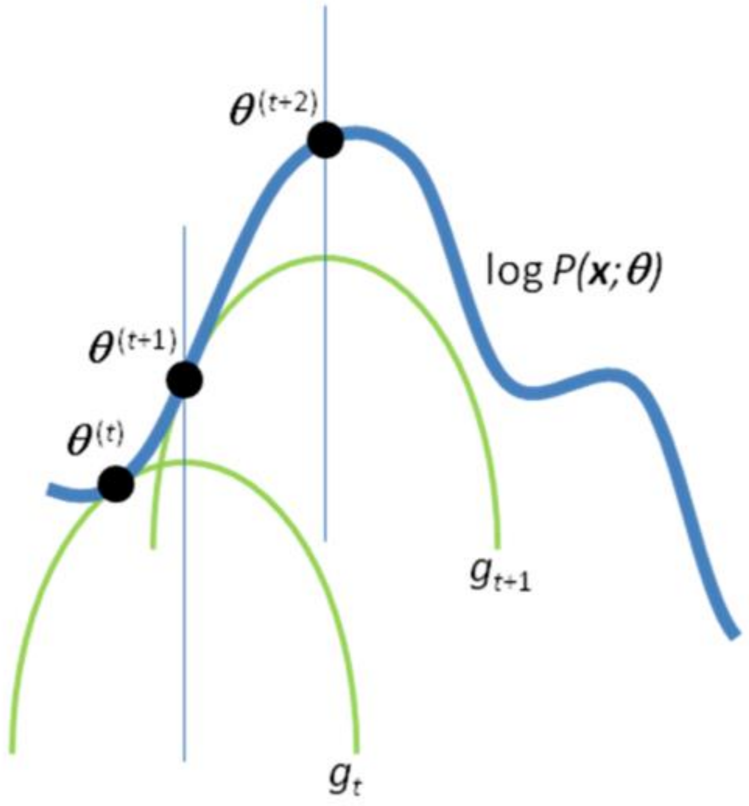
$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} = \frac{N_k}{N}$$

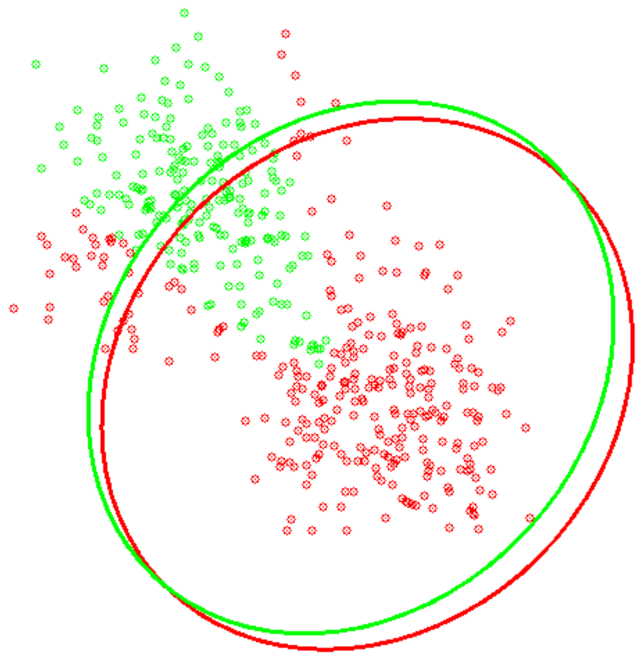
# Outline

- Overview
- Gaussian Mixture Model
- Expectation-Maximization Algorithm

# Expectation maximization

- Expectation Maximization (EM) is a general algorithm to deal with hidden variables.
- Two steps:
  - E-Step: Fill-in hidden values using inference
  - M-Step: Apply standard MLE method to estimate parameters
- EM always converges to a local minimum of the likelihood.



# EM for Gaussian Mixture Models

- Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters comprising the means and covariances of the components and the mixing coefficients.

- Initialize the means $\mu_k$, covariances $\Sigma_k$ and mixing coefficients $\boldsymbol{\pi}_k$ and evaluate the initial value of the log-likelihood.

- **E-step:** Evaluate the responsibilities using the current parameter values

$$\gamma(z_k) = p(z_k|\mathbf{x}) = \frac{p(z_k)p(\mathbf{x}|z_k)}{\sum_{j=1}^{K} p\left(z_j\right)p(\mathbf{x}|z_j)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_\mathrm{j})}$$

# EM for Gaussian Mixture Models

- **M-Step:** Re-estimate parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{N_k}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})^T}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})^T}{N_k}$$

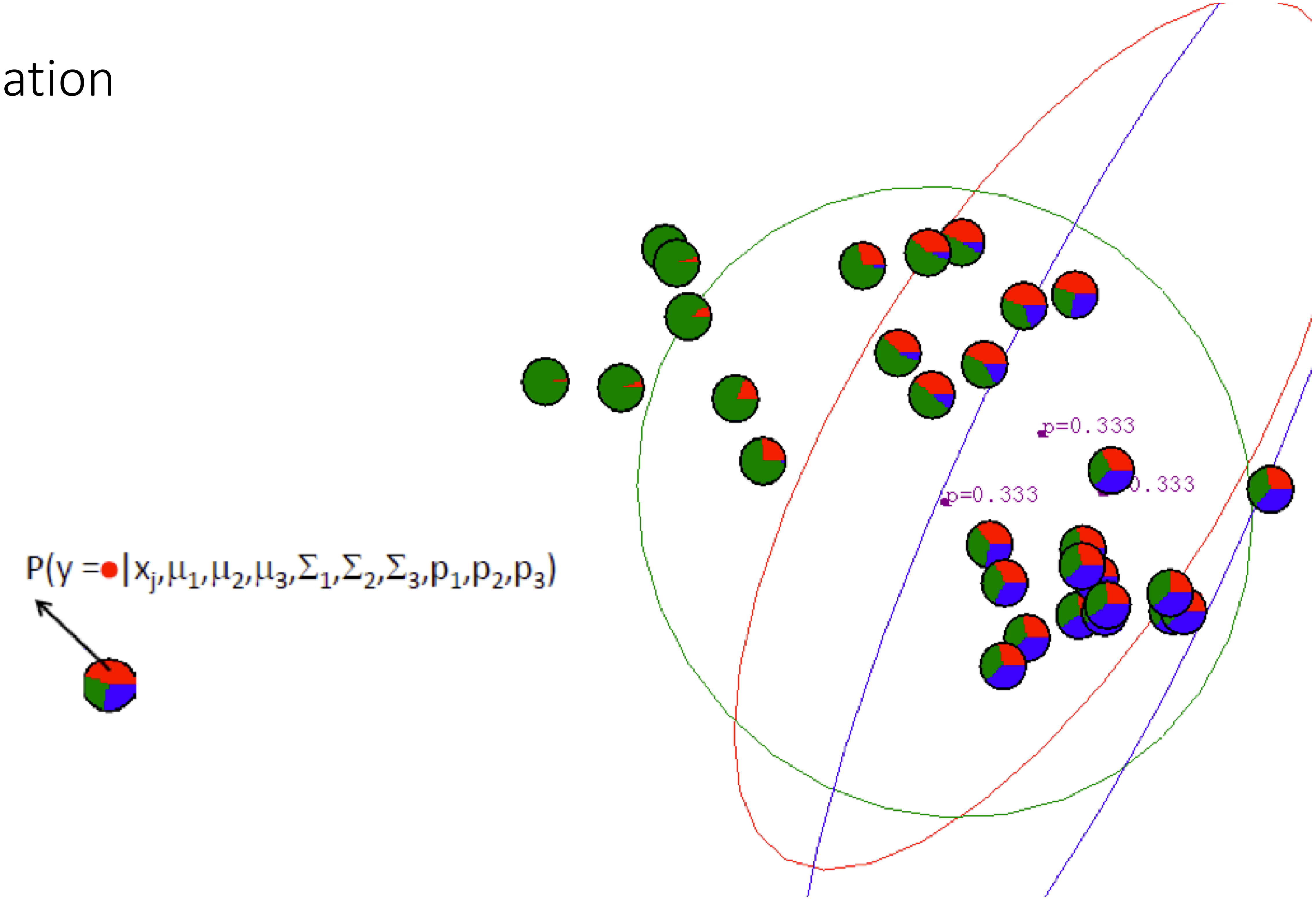
$$\pi_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} = \frac{N_k}{N}$$

# EM for Gaussian Mixture Models

Example

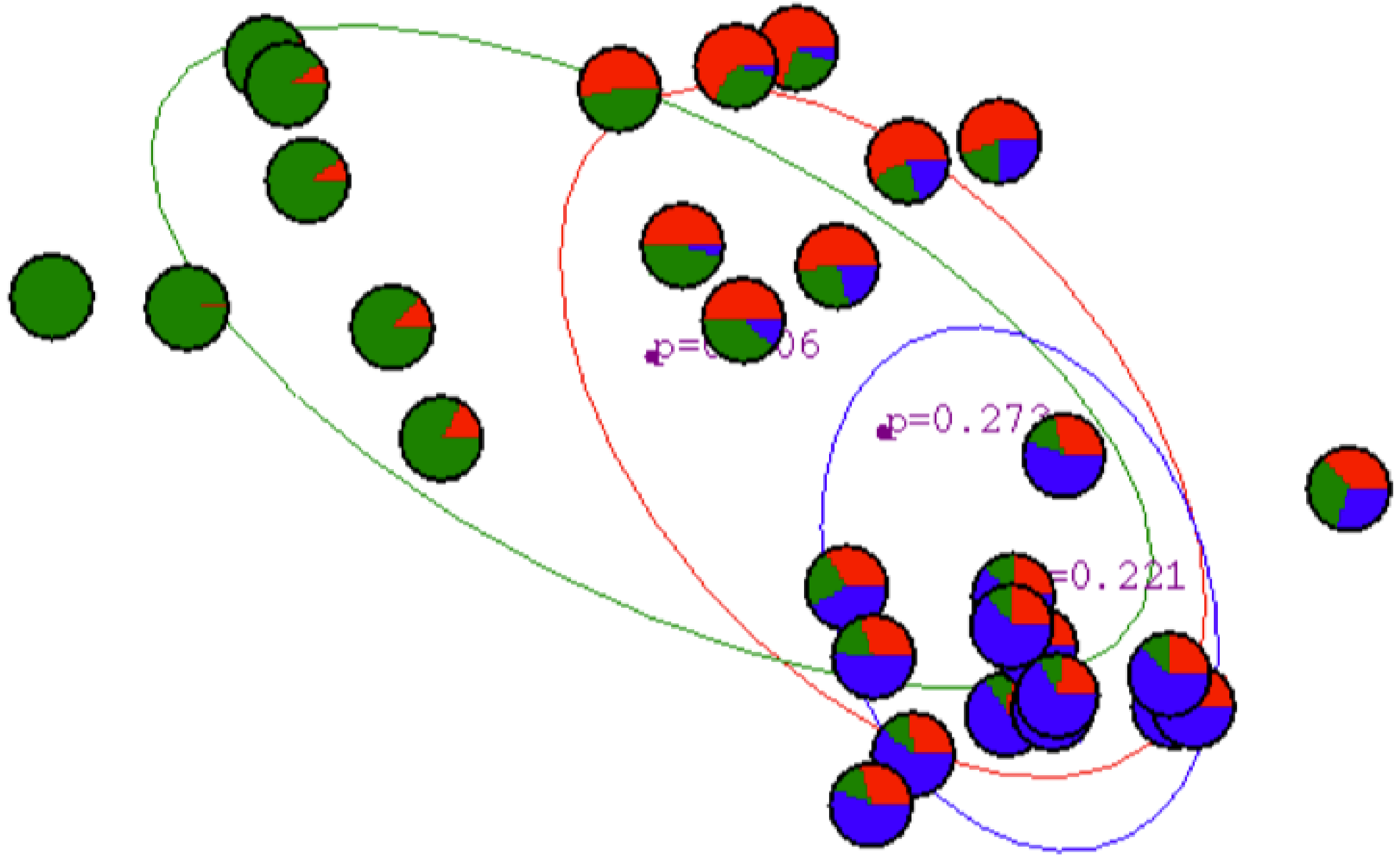
# EM for GMMs: Example

- Initialization



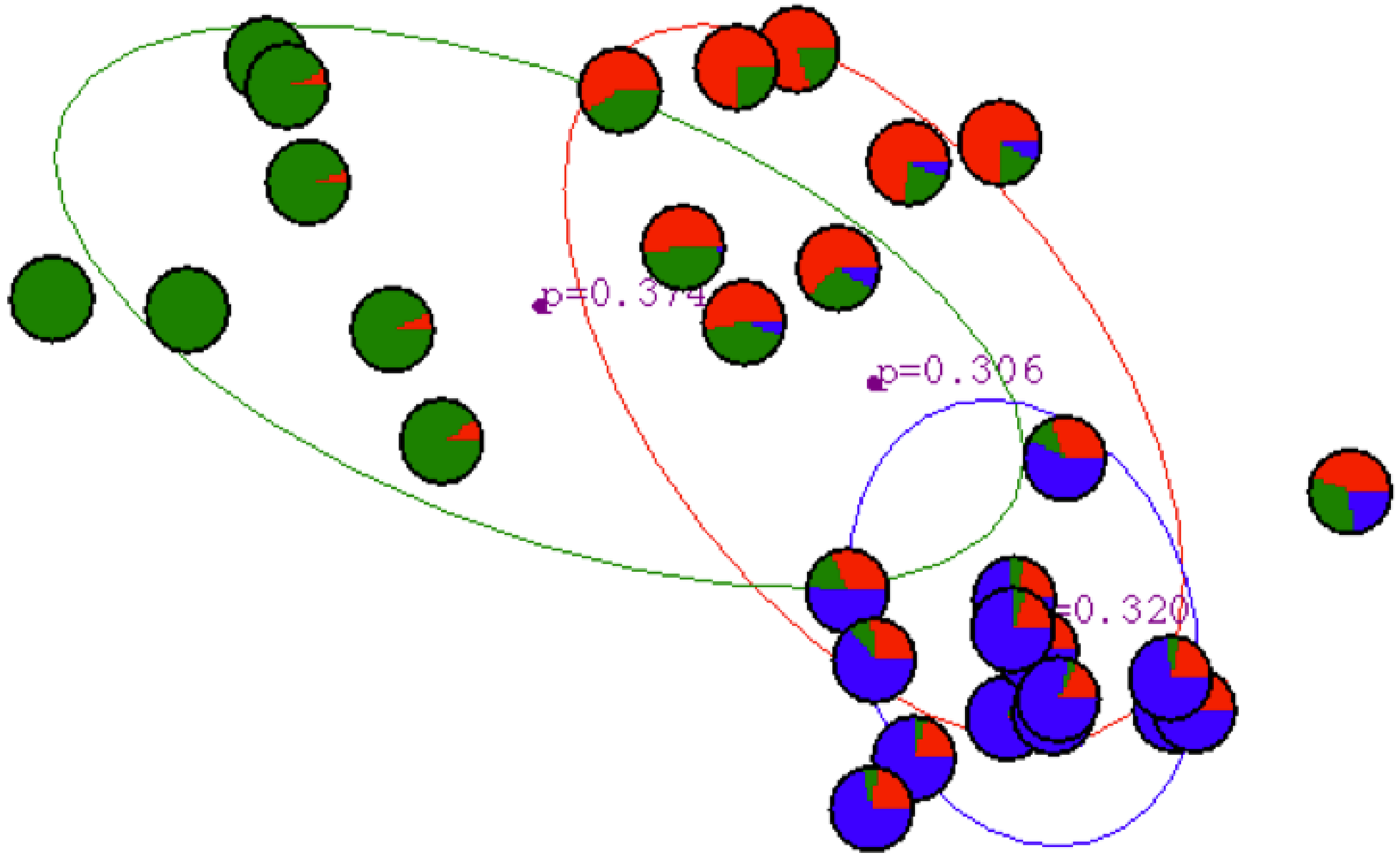
# EM for GMMs: Example

- After 1st iteration



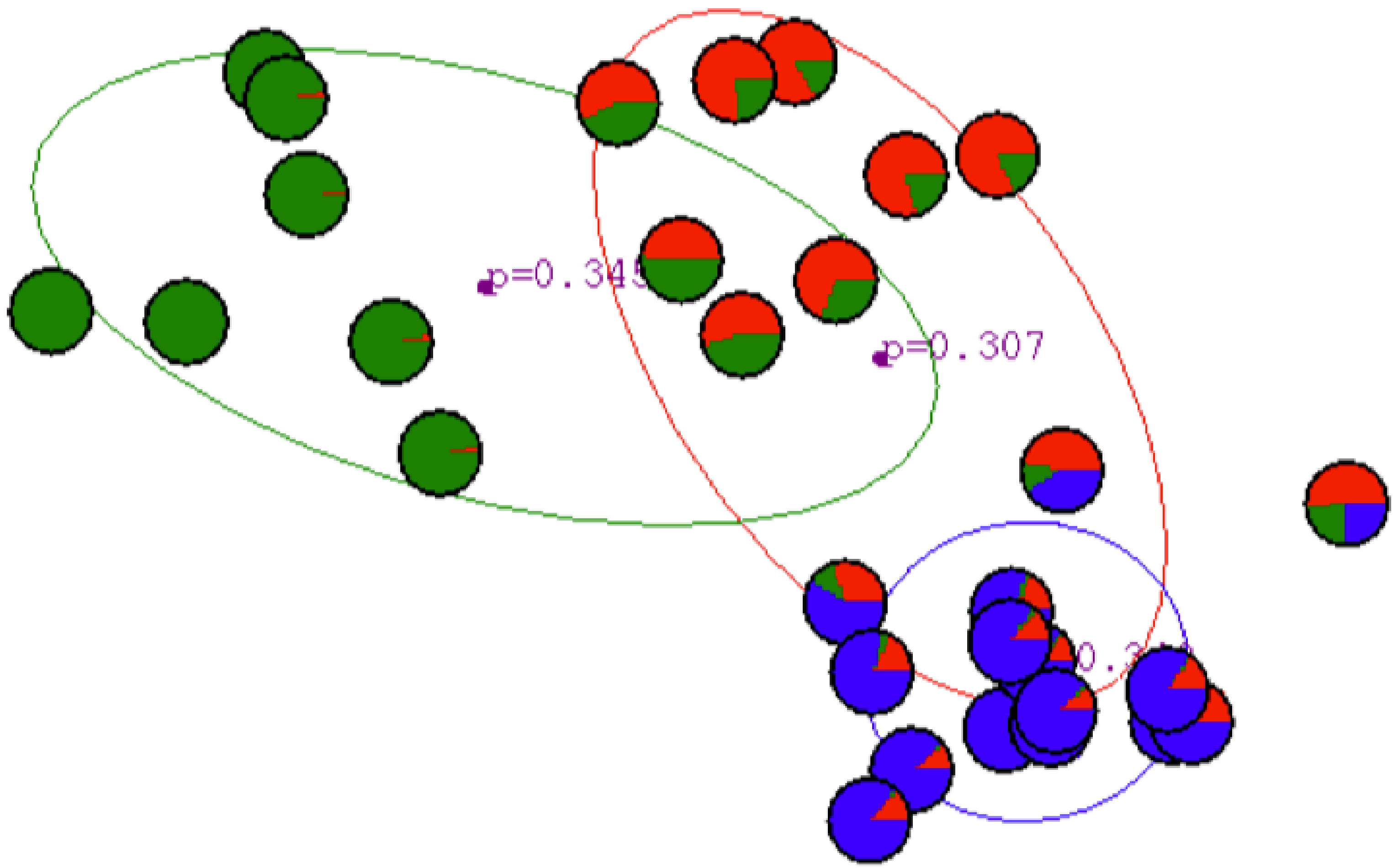
# EM for GMMs: Example

- After 2nd iteration



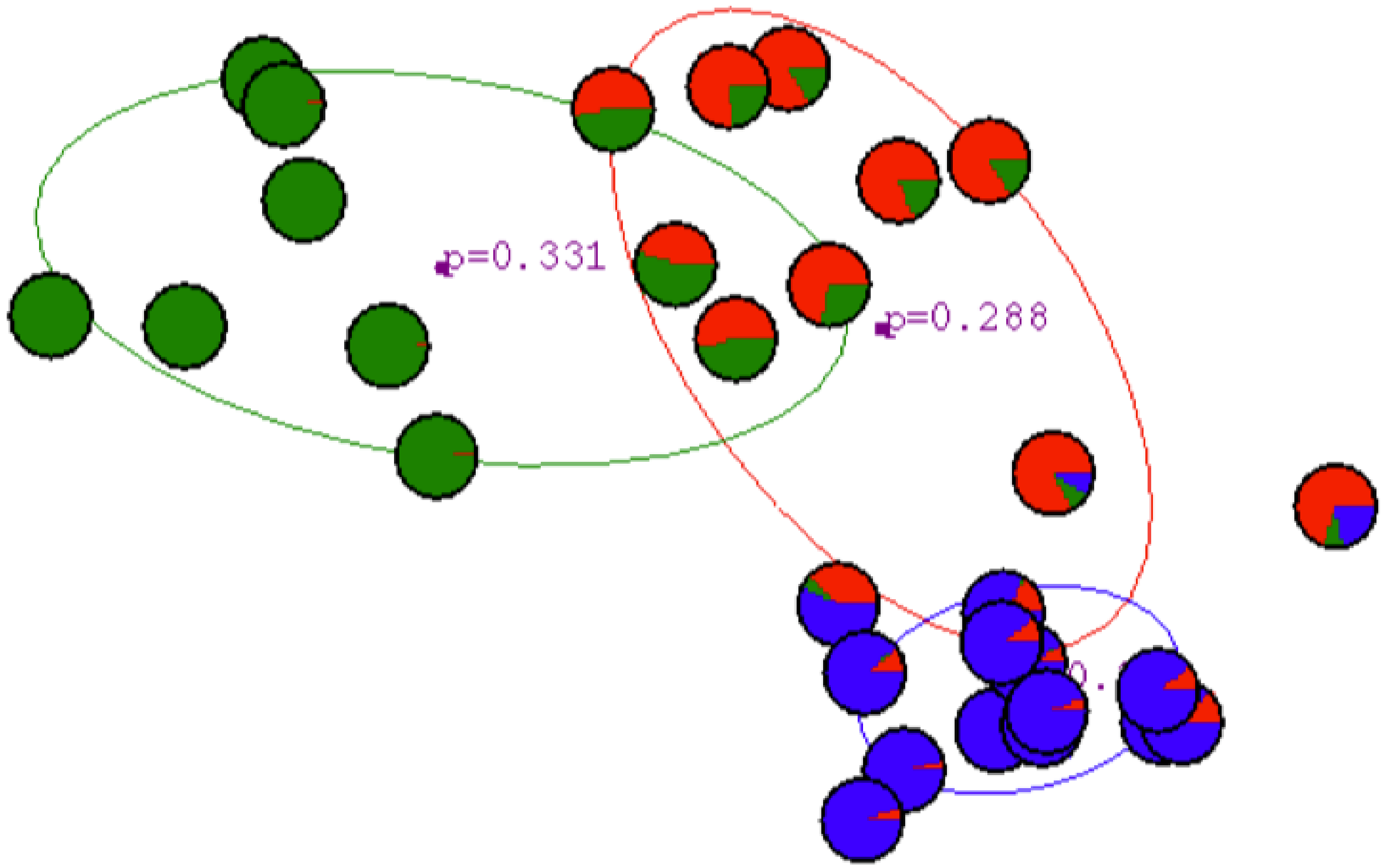
# EM for GMMs: Example

- After 3rd iteration



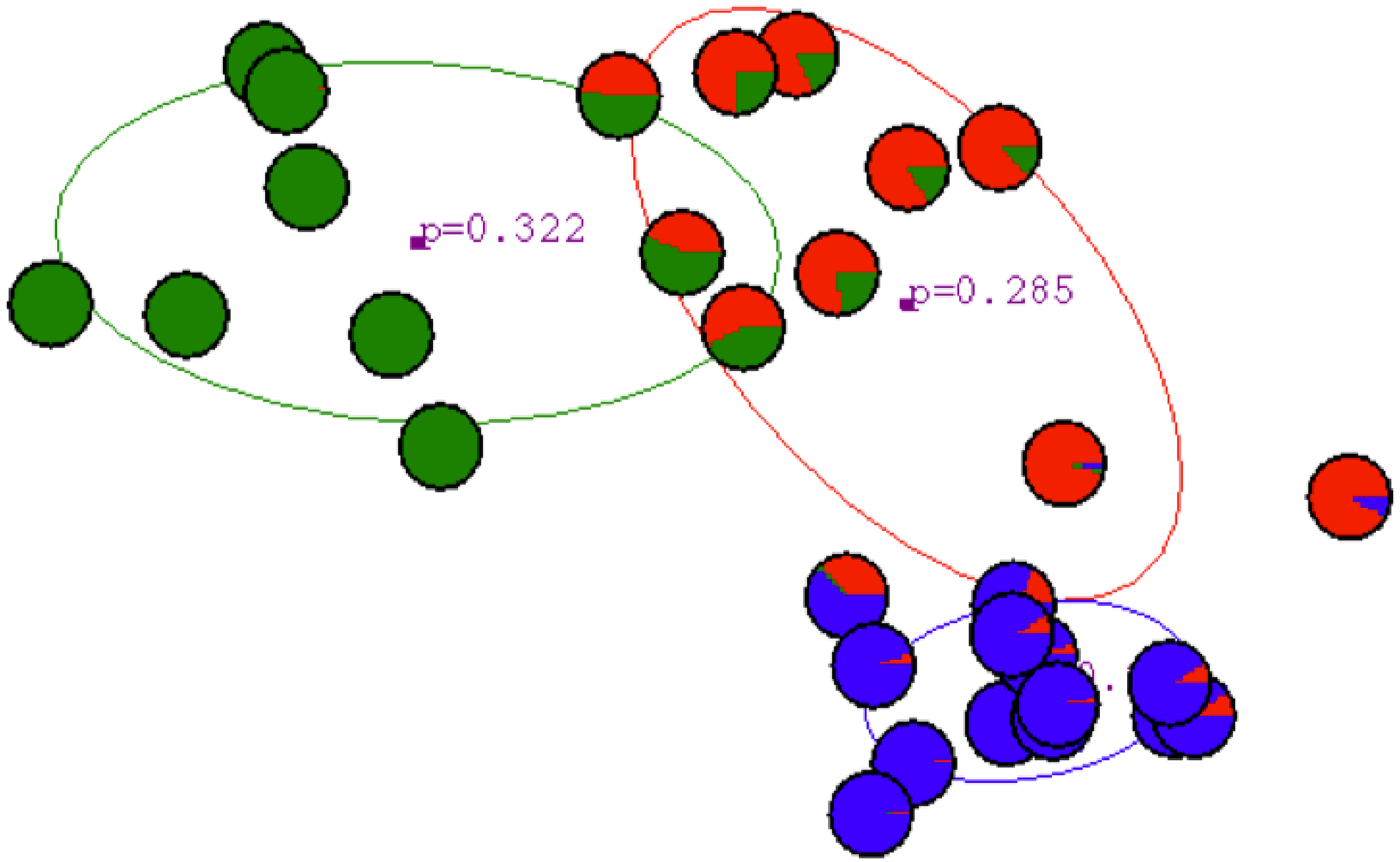
# EM for GMMs: Example

- After 4th iteration



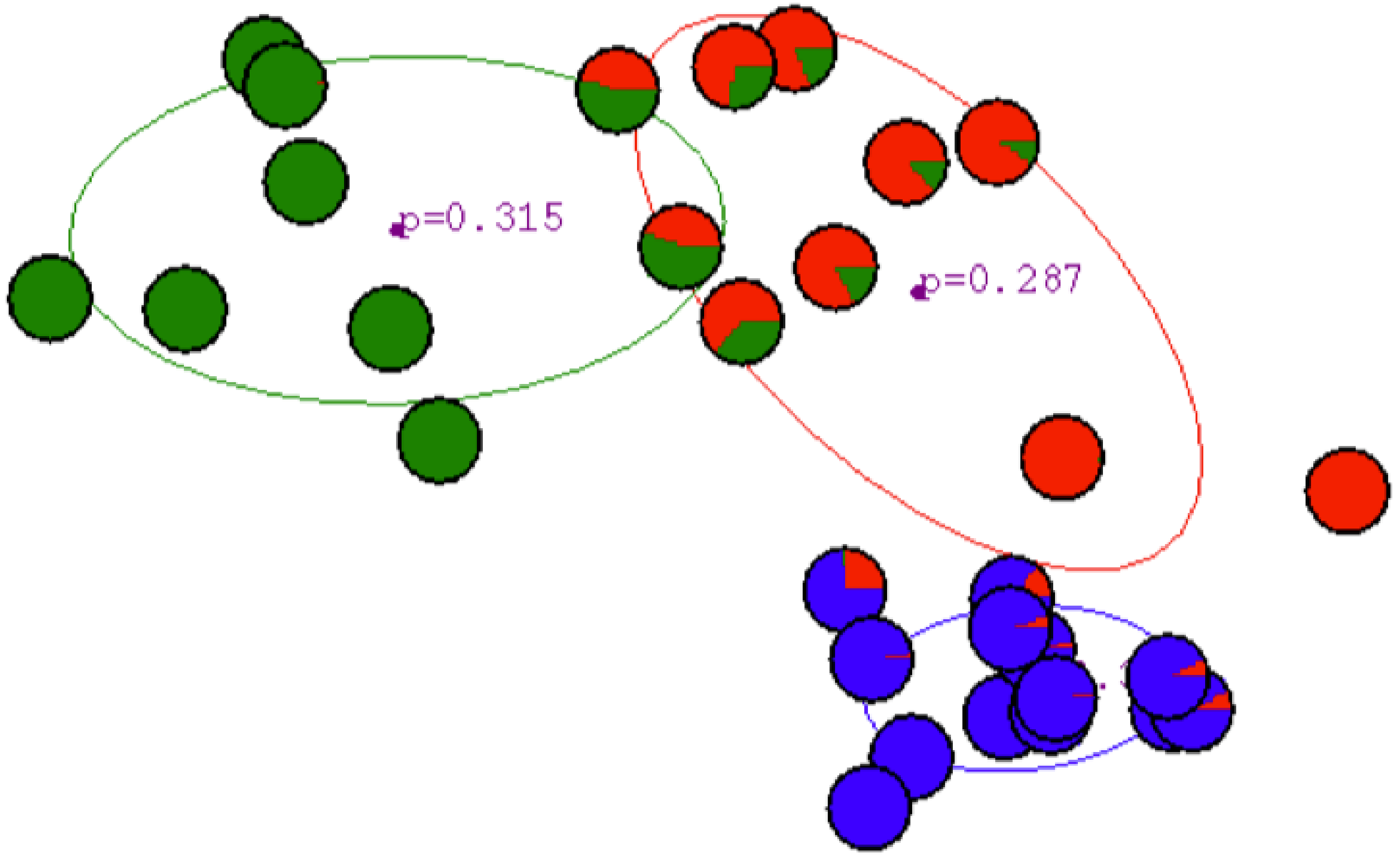
# EM for GMMs: Example

- After 5th iteration



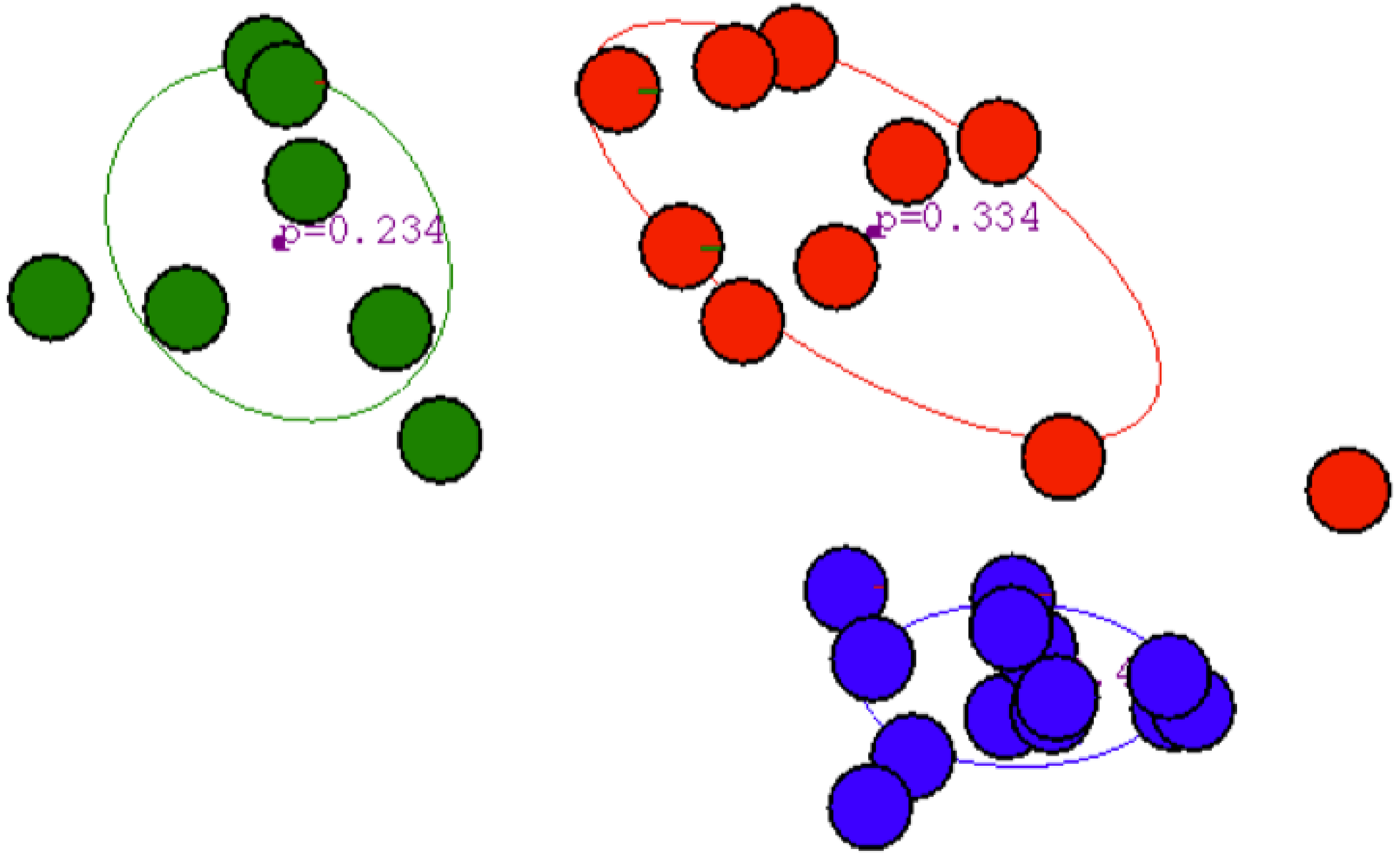
# EM for GMMs: Example

- After 6th iteration



# EM for GMMs: Example

- After 20th iteration



# Relationship to K-means

- K-means makes hard decisions.
  - Each data point gets assigned to a single cluster.
- GMM/EM makes soft decisions.
  - Each data point can yield a posterior $p(z|x)$
- K-means is a special case of EM

# General form of EM

- Givern a joint distribution over observed and latent variables: $p(x, z|\theta)$
- Want to maximize: $p(x|\theta)$

1. Initialize parameters: $\theta^{old}$
2. E-step: evaluate $p\left(z\middle|x, \theta^{old}\right)$
3. M-step: Re-estimate parameters (based on expectation of complete-data log likelihood

$$\theta^{new} = argmax_\theta \sum_z p\left(z|x, \theta^{old}\right) \ln p(x, z|\theta) = argmax_\theta \mathbb{E}[\ln p(x, z|\theta)]$$

4. Check for convergence of parameters or likelihood

# Jensen's inequality

$$
\begin{aligned}
l(\theta, x) &= \ln p(x|\theta) \\
&= \ln \sum_z p(x, z|\theta) \\
&= \ln \sum_z q(z|x) \frac{p(x, z|\theta)}{q(z|x)} \\
&\geq \sum_z q(z|x) \ln \frac{p(x, z|\theta)}{q(z|x)}
\end{aligned}
$$

Will lead to maximize this

↑

Maximizing this

$$
= \sum_z q(z|x) \ln \frac{p(x, z|\theta)}{q(z|x)} = \sum_z q(z|x) \ln p(x, z|\theta) - \sum_z q(z|x) \ln q(z|x) = \langle l_c(\theta, x, z) \rangle + H_q
$$

- The first term is the expected complete log likelihood and the second term, which does not depend on $\theta$, is the entropy.
- Thus, in the M-step, maximizing with respect to $\theta$ for fixed q we only need to consider the first term:

$$
\theta^{new} = argmax_\theta \langle l_c(\theta, x, z) \rangle_{q^{new}} = argmax_\theta \sum_z q(z|x) \ln p(x, z|\theta)
$$