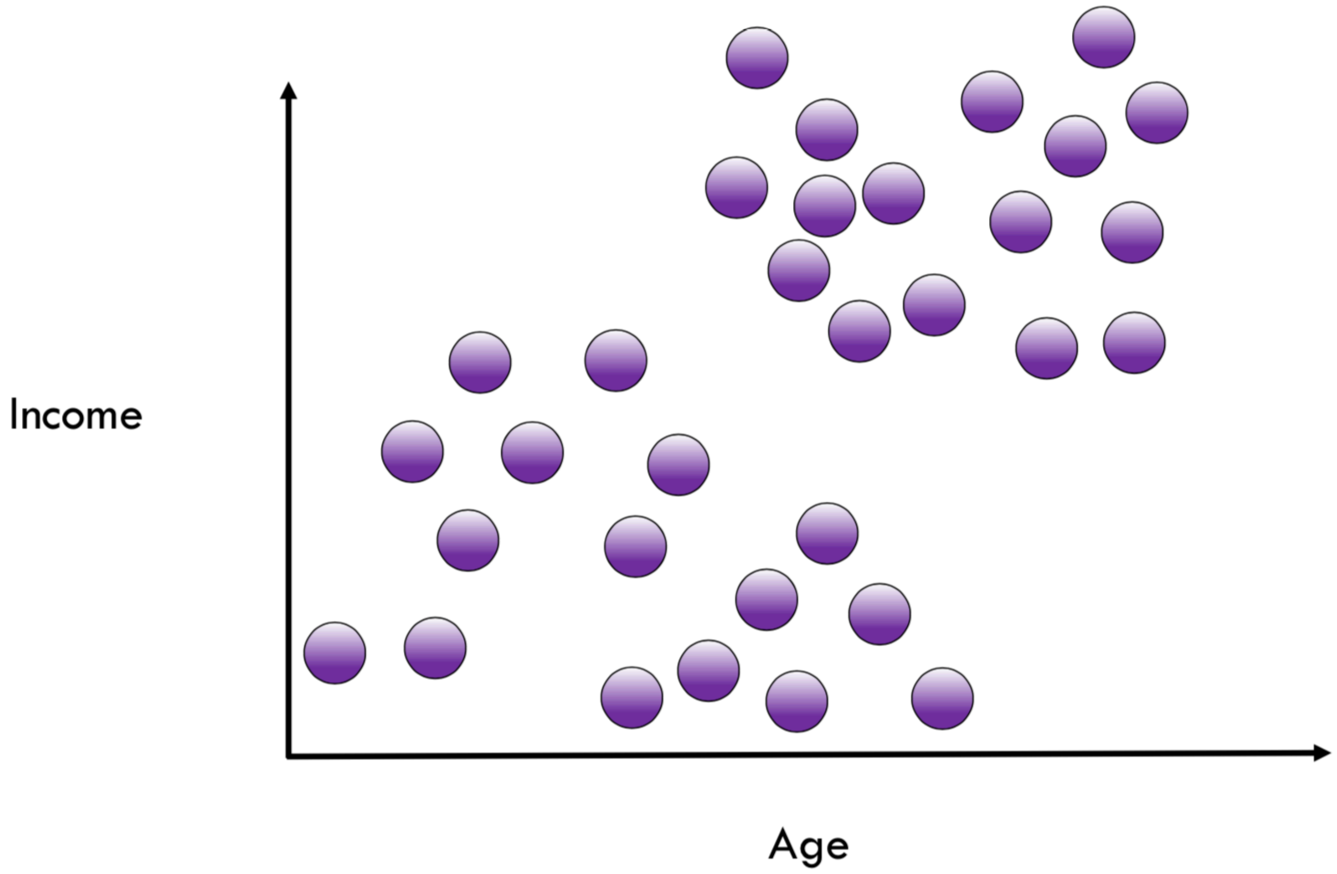
Gaussian Mixture Model

Kookjin Lee

(kookjin.Lee@asu.edu)

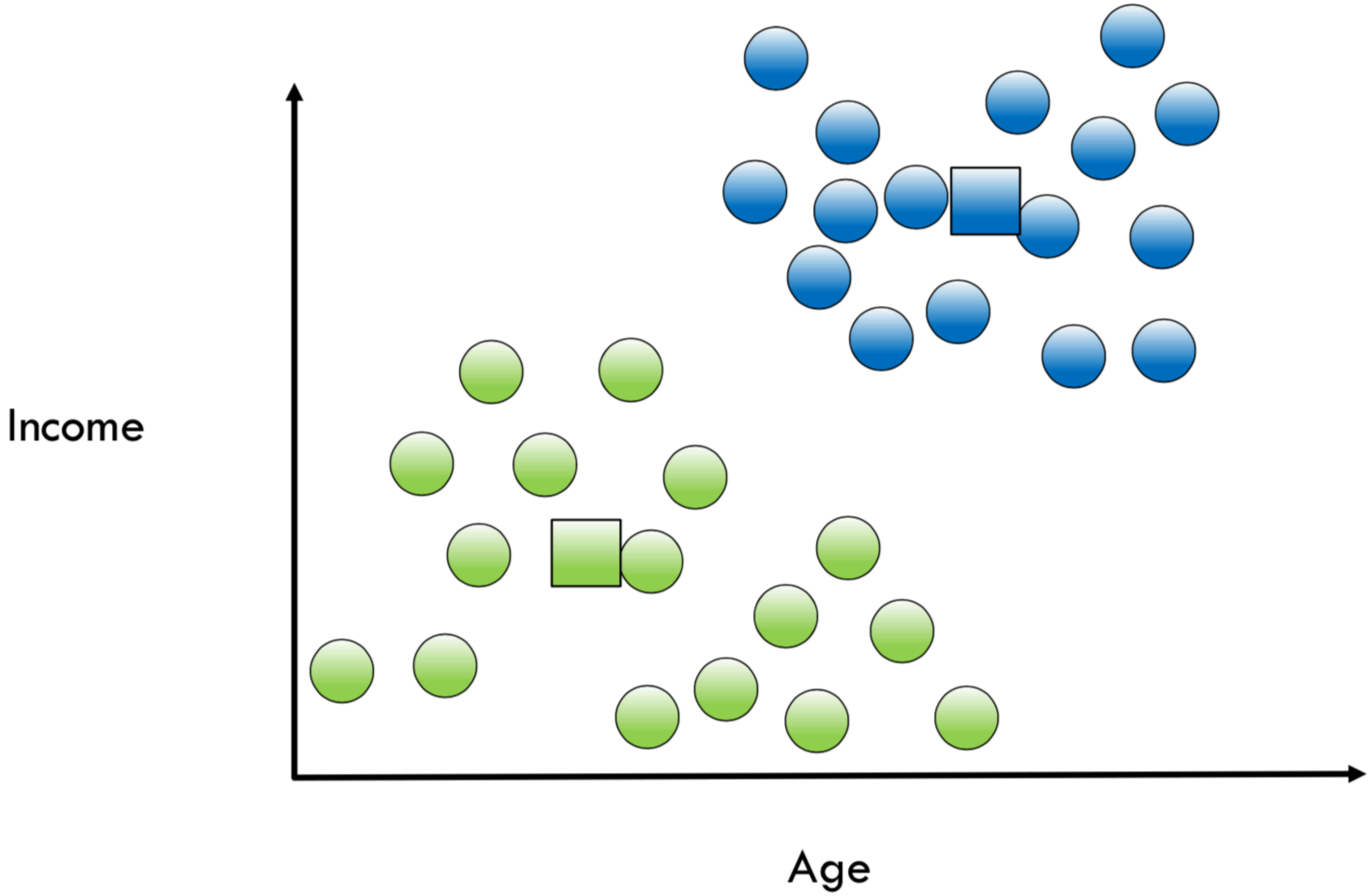
K-means clustering

- K=2



K-means clustering

- K=2



Probabilistic ways of assigning centers

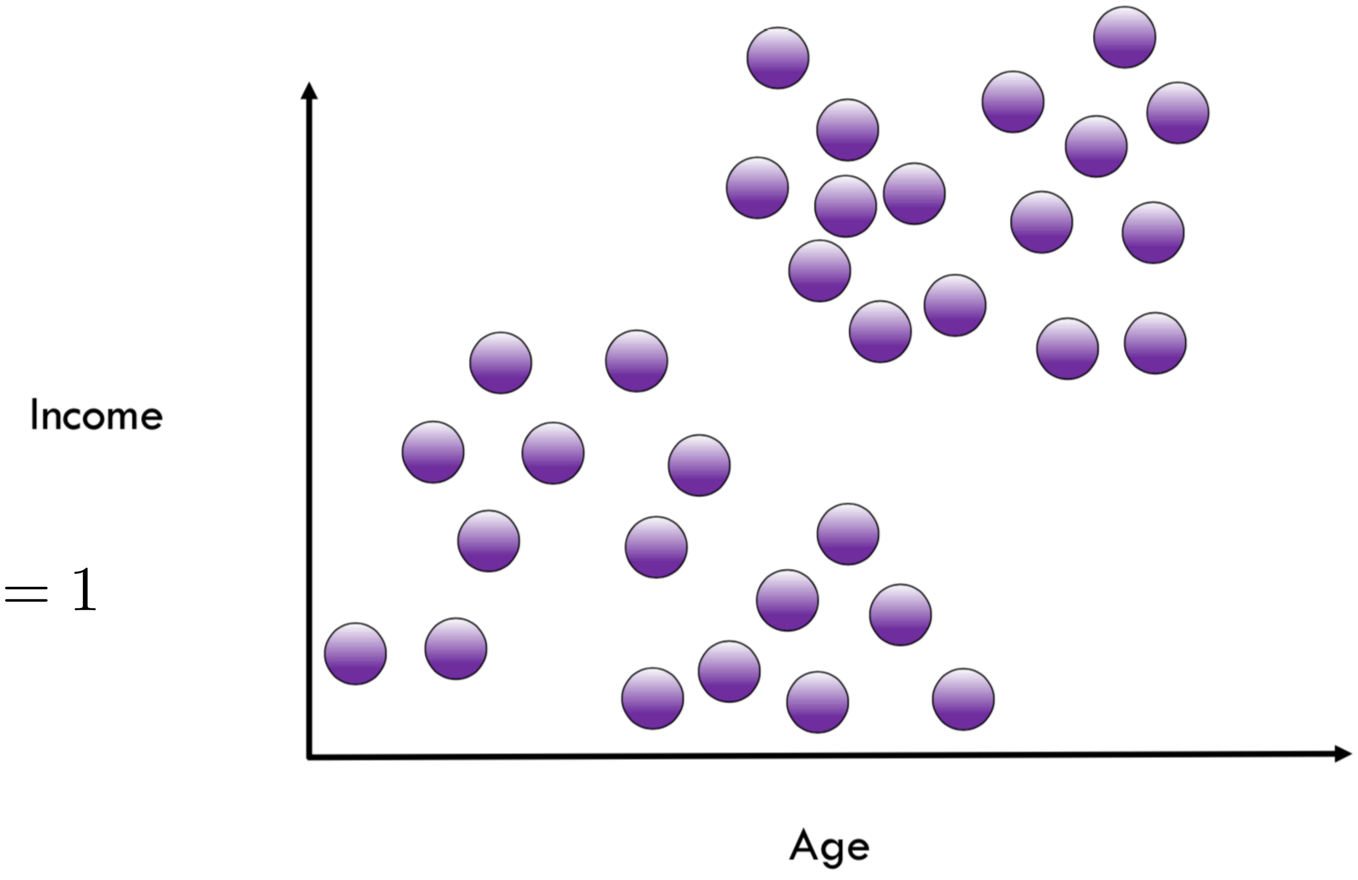
- K clusters
  - Assign the probability
  - Introducing a latent (hidden) variable $z^{(i)}$

$$z^{(i)} \in \{1, \ldots, K\}$$

$$z^{(i)} \sim \text{Multinomial}(\phi)$$

$$\phi = (\phi_1, \ldots, \phi_K), \quad \sum \phi_i = 1$$

$$p(z^{(i)} = j)$$



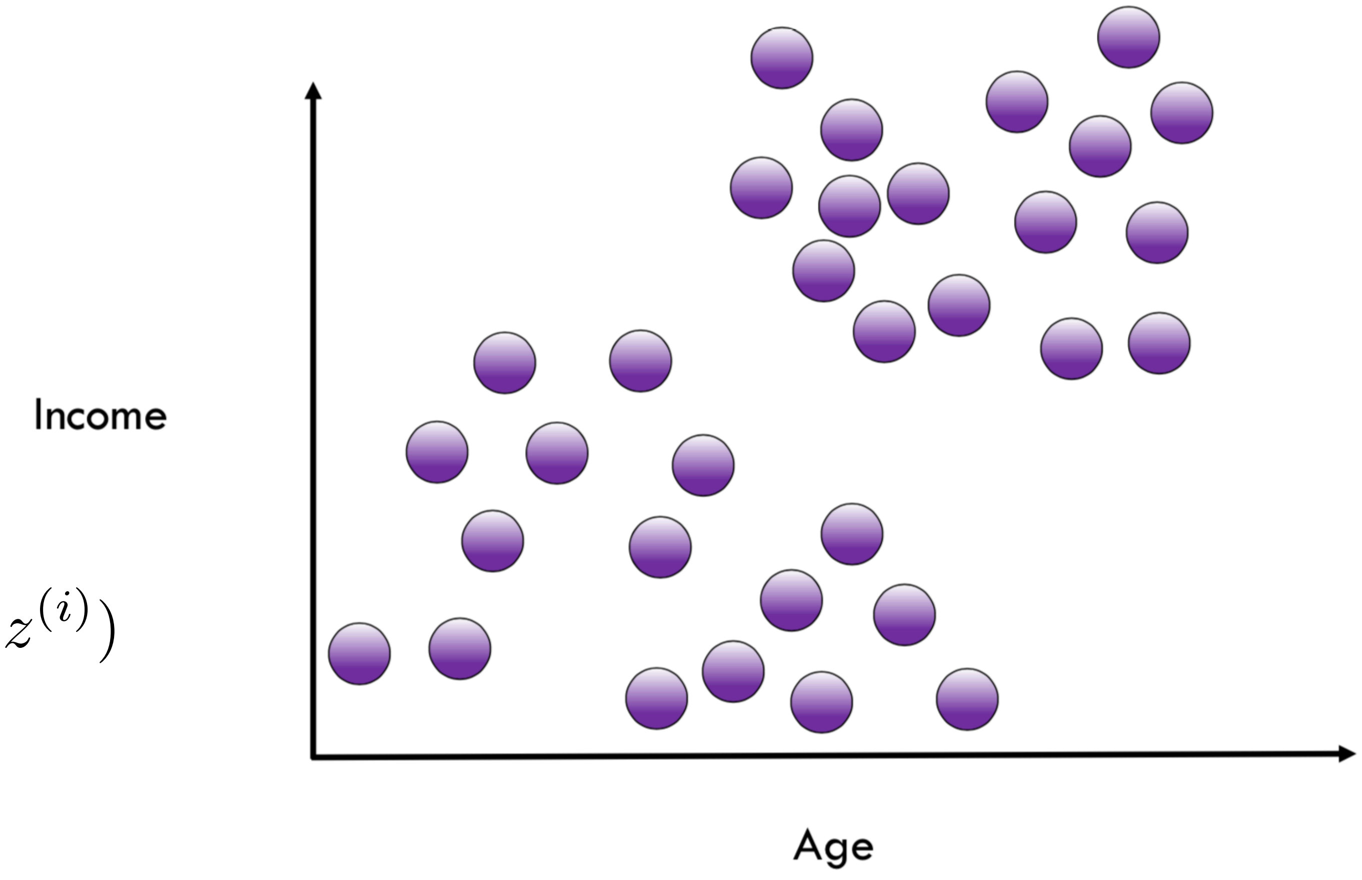
Probabilistic ways of assigning centers

- K clusters
  - Assign the probability
  - Introducing a latent (hidden) variable $z^{(i)}$
  - Conditional probability

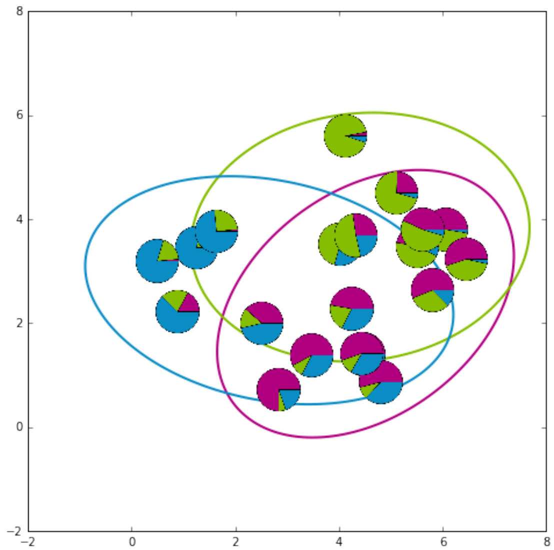
    $$x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

  - All together

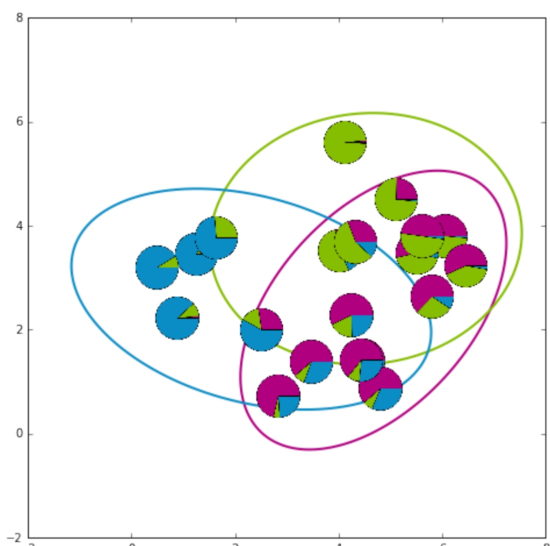
    $$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$$



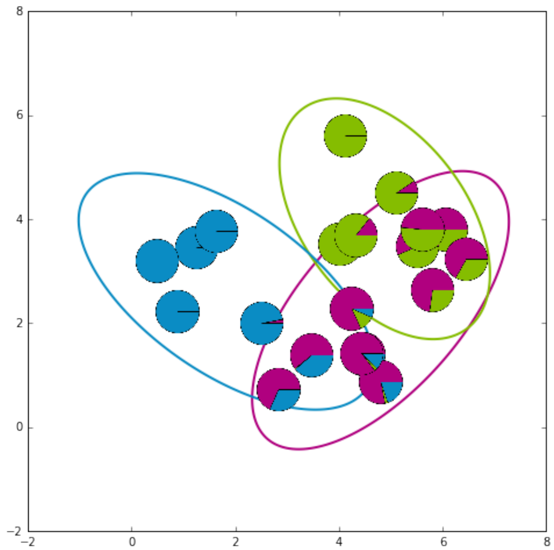
EM for mixtures of Gaussians in pictures – initialization



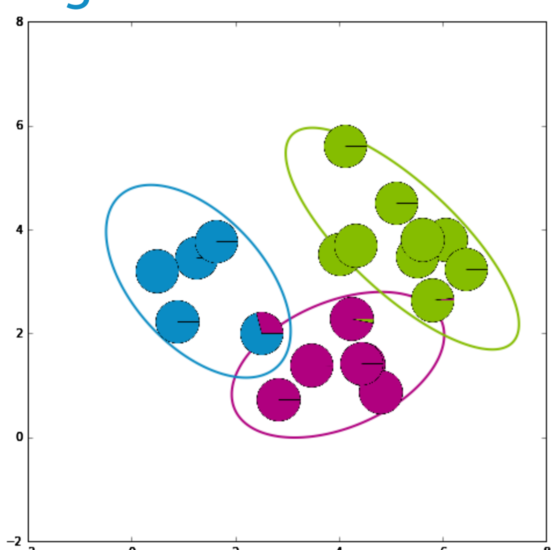
EM for mixtures of Gaussians in pictures – after 1st iteration



EM for mixtures of Gaussians in pictures – after 2nd iteration



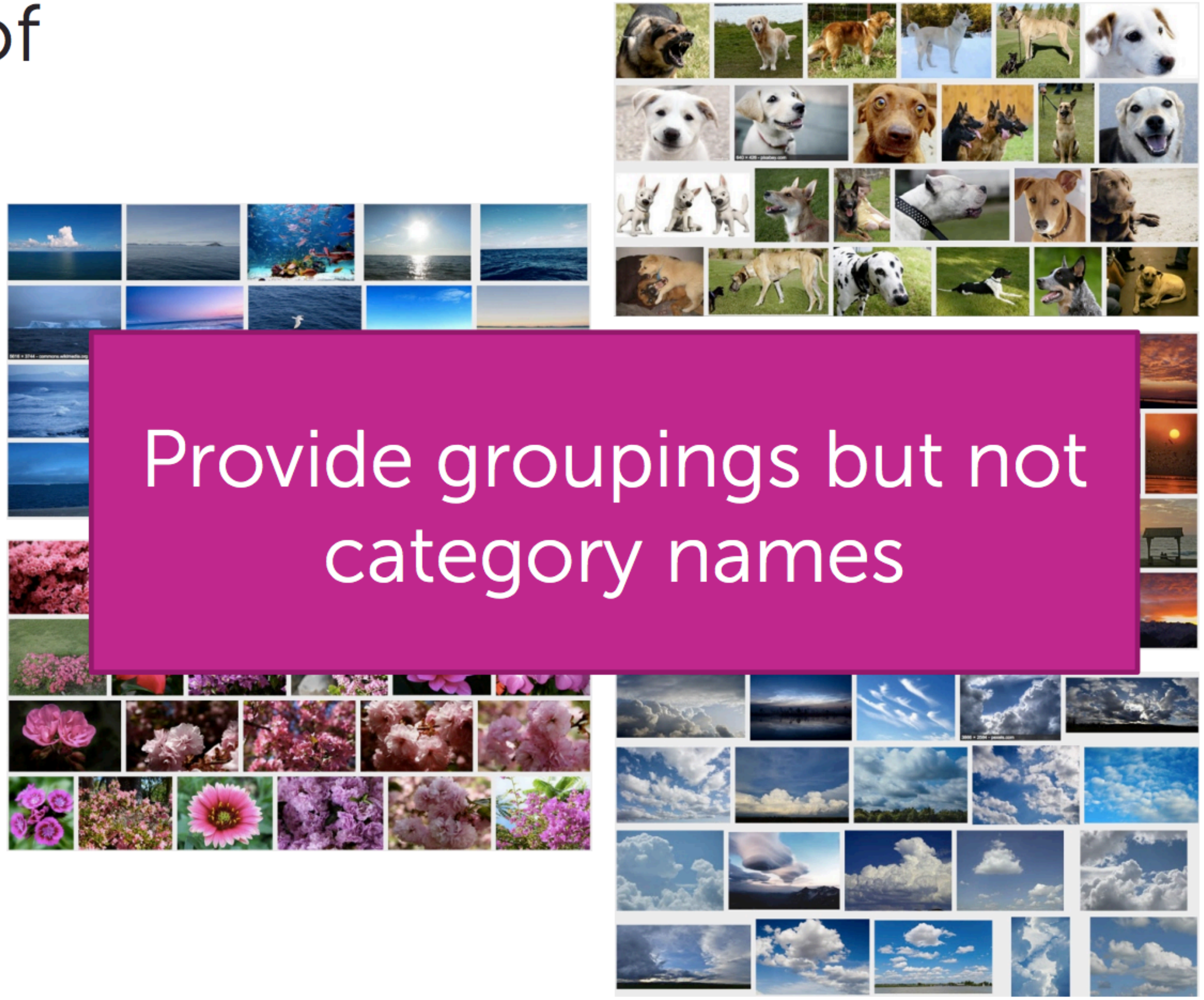
EM for mixtures of Gaussians in pictures – converged solution



Examples

Discover groups of similar images

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...

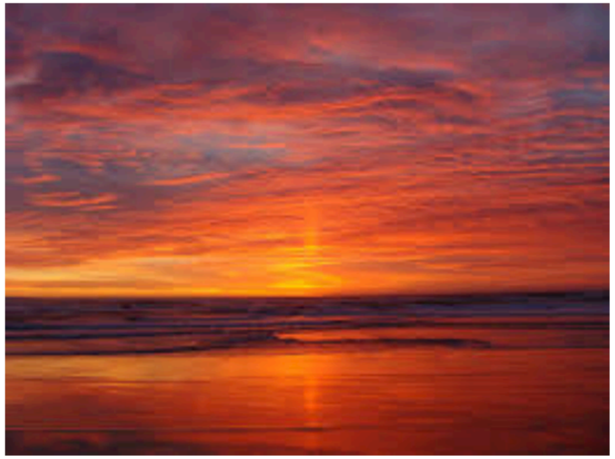


Provide groupings but not category names

Simple image representation

Consider average red, green, blue pixel intensities

[R = 0.05, G = 0.7, B = 0.9]

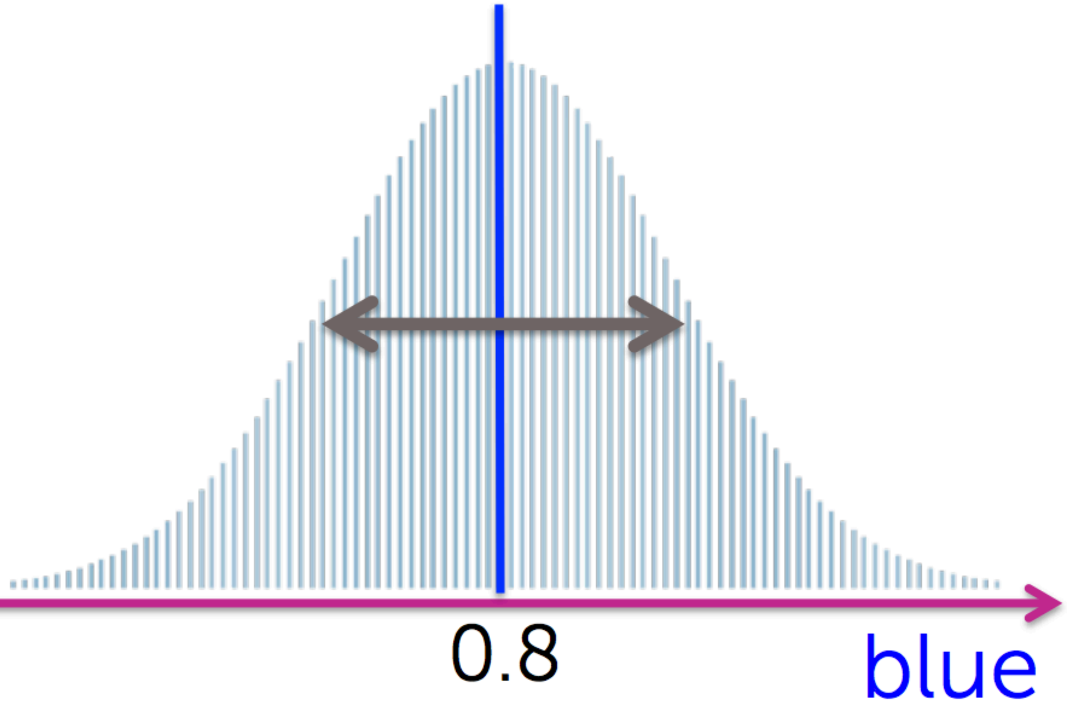


[R = 0.85, G = 0.05, B = 0.35]

[R = 0.02, G = 0.95, B = 0.4]

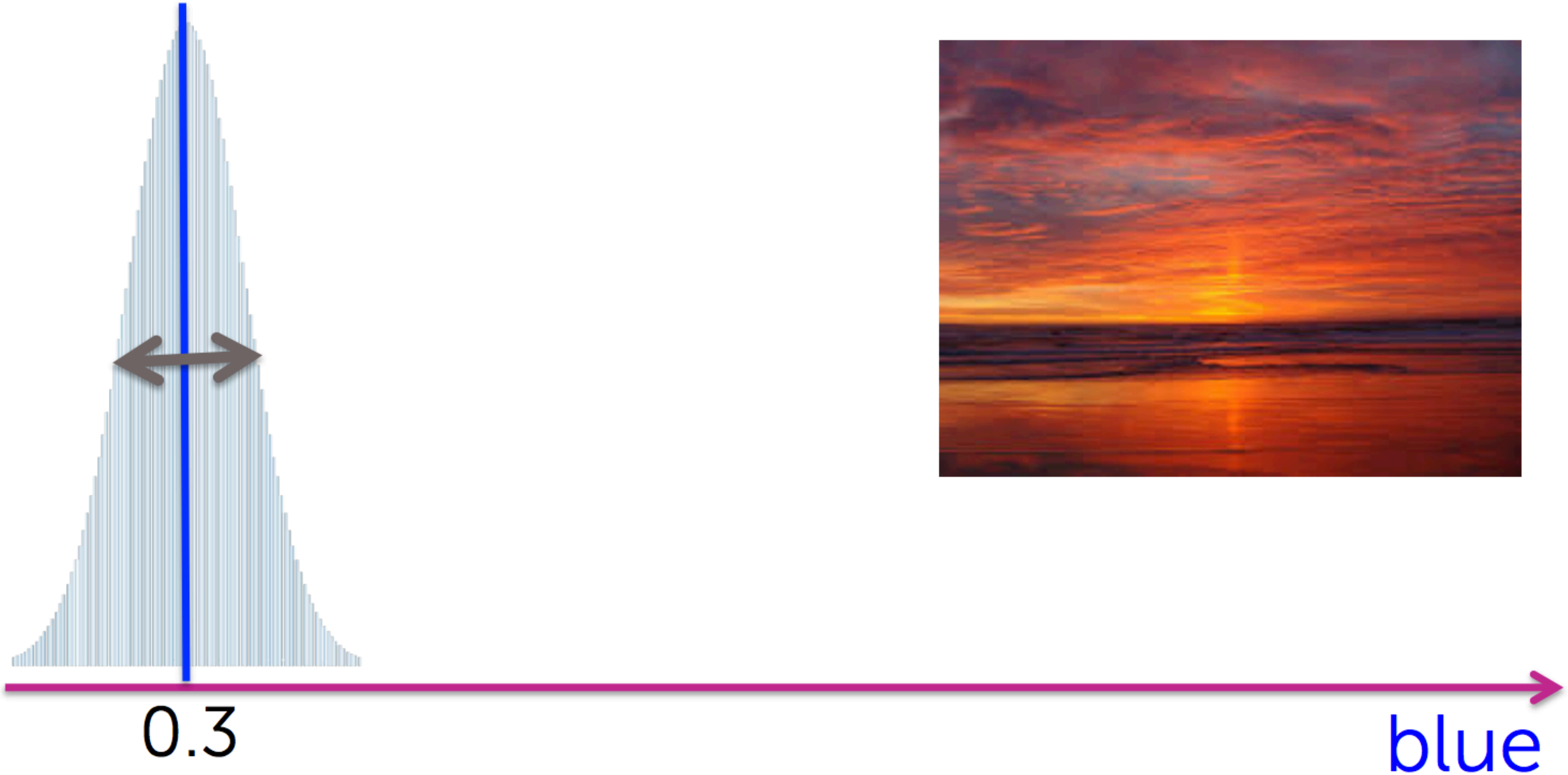
Distribution over all cloud images

Let's look at just the blue dimension



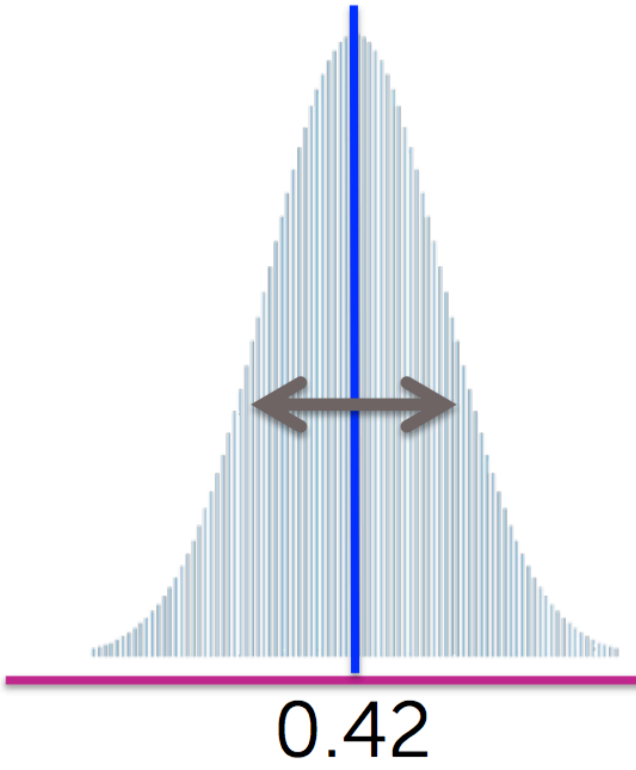
Distribution over all sunset images

Let’s look at just the blue dimension

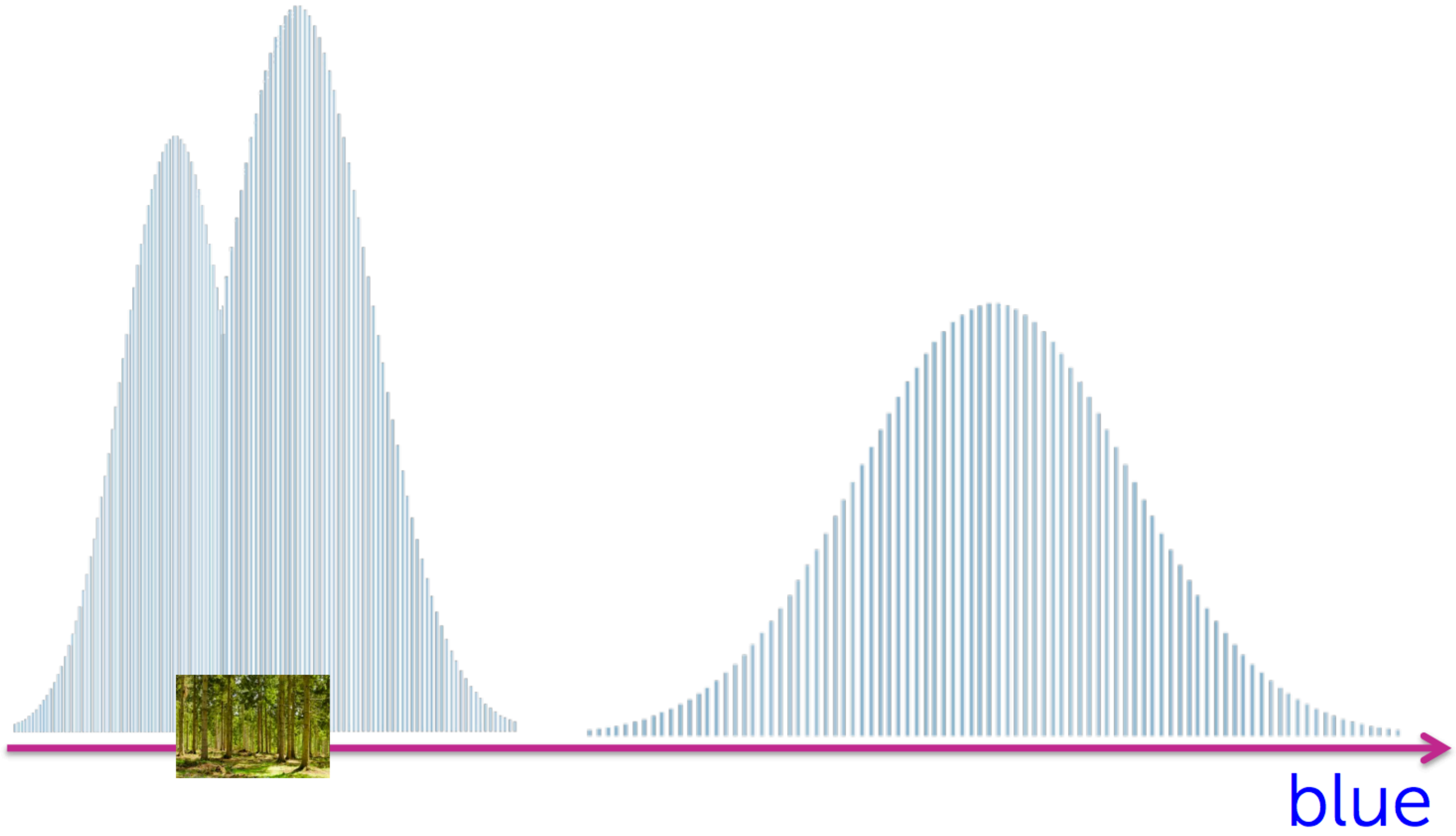


Distribution over all forest images

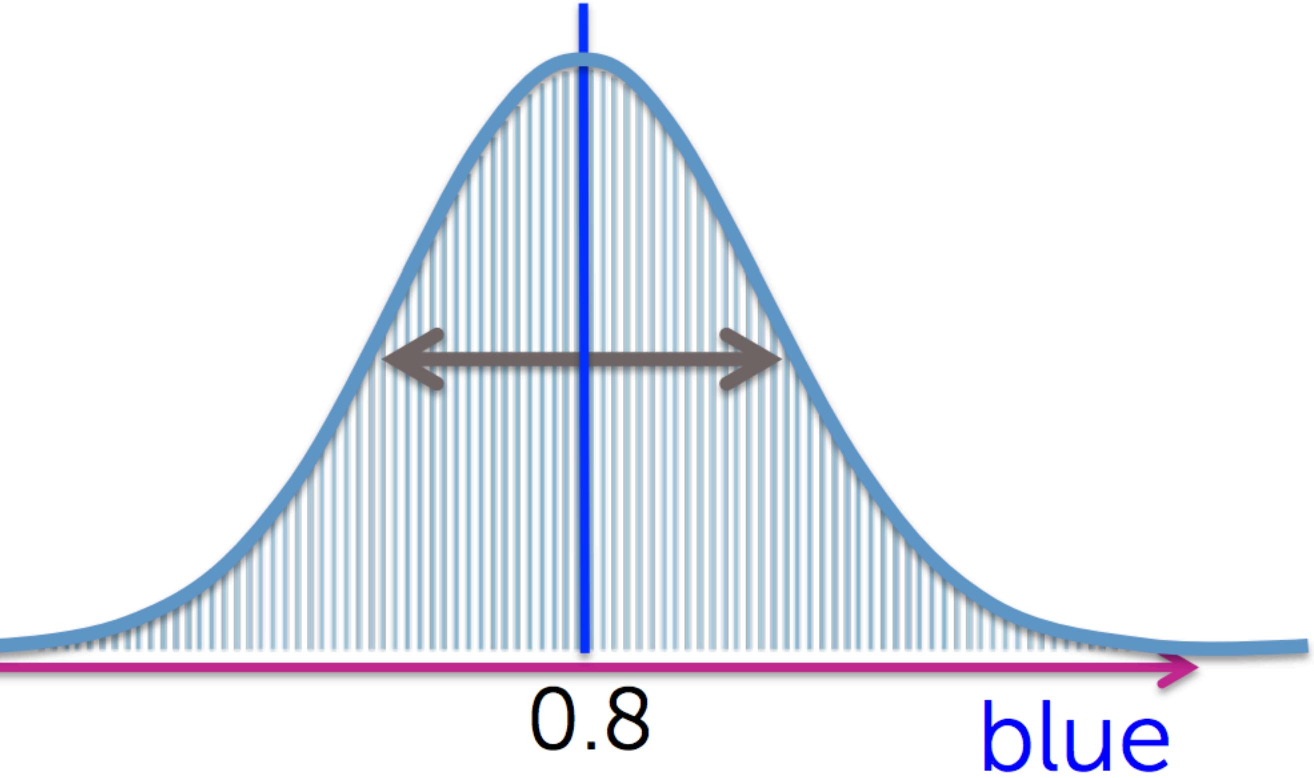
Let's look at just the blue dimension



Distribution over all images

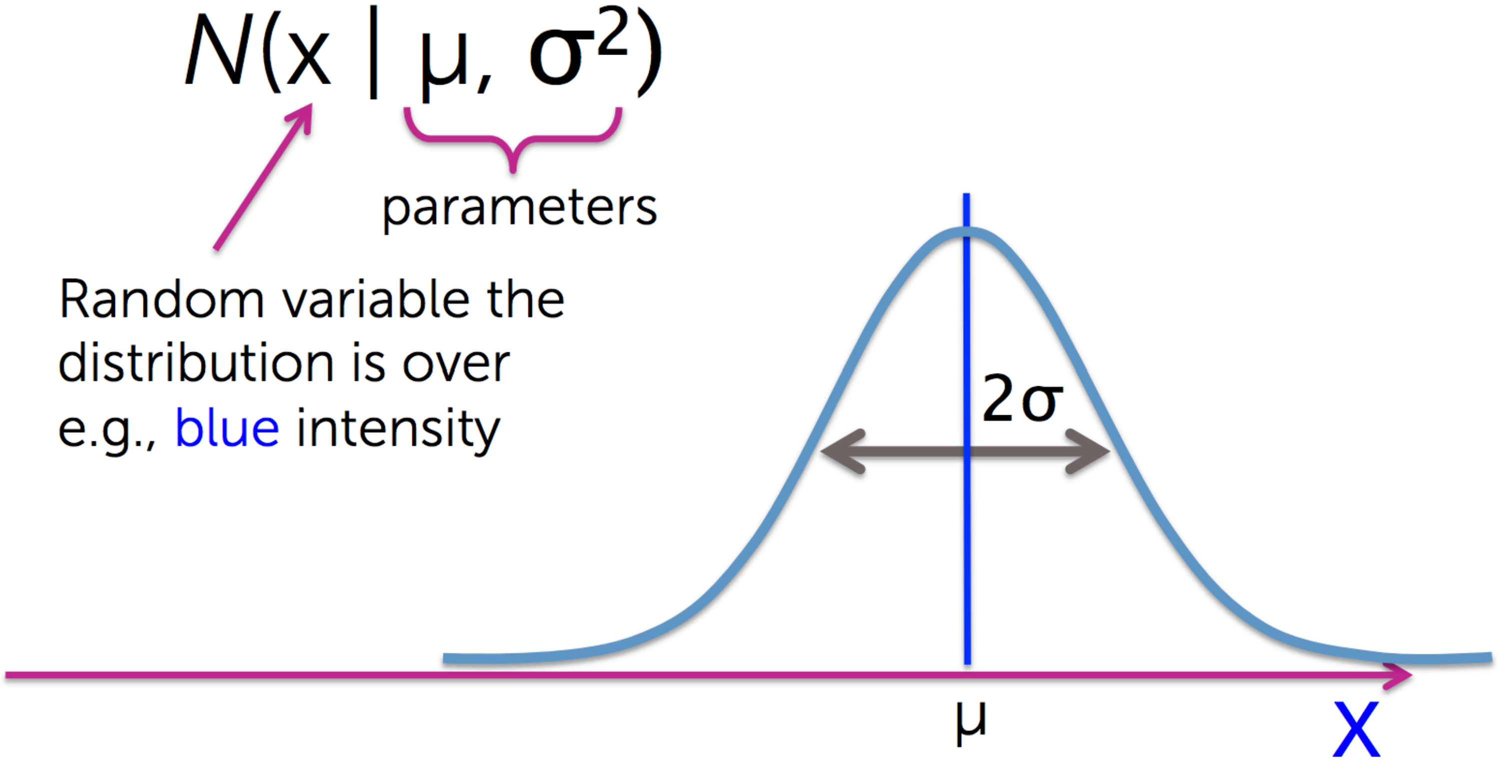


Assume Gaussian distribution



Revisiting Gaussian

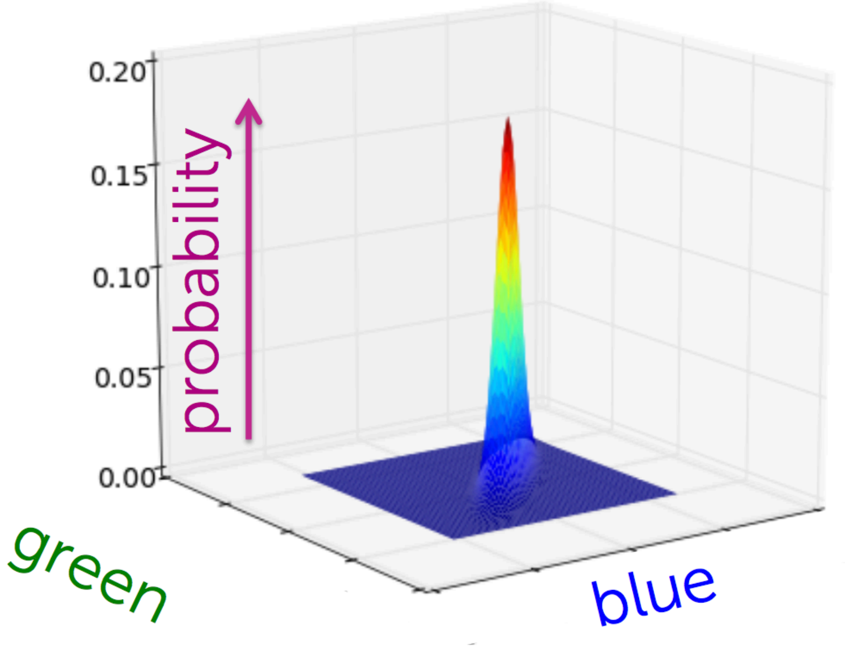
- 1D gaussian



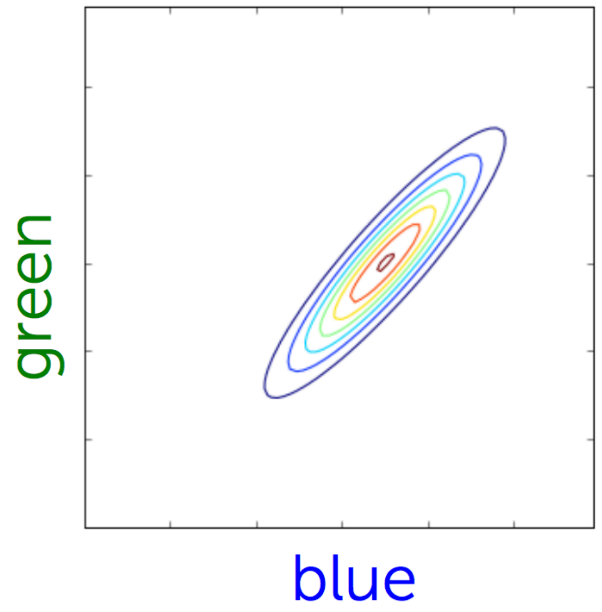
Revisiting Gaussian

- 2D Gaussian

3D mesh plot



Contour plot



Revisiting Gaussian

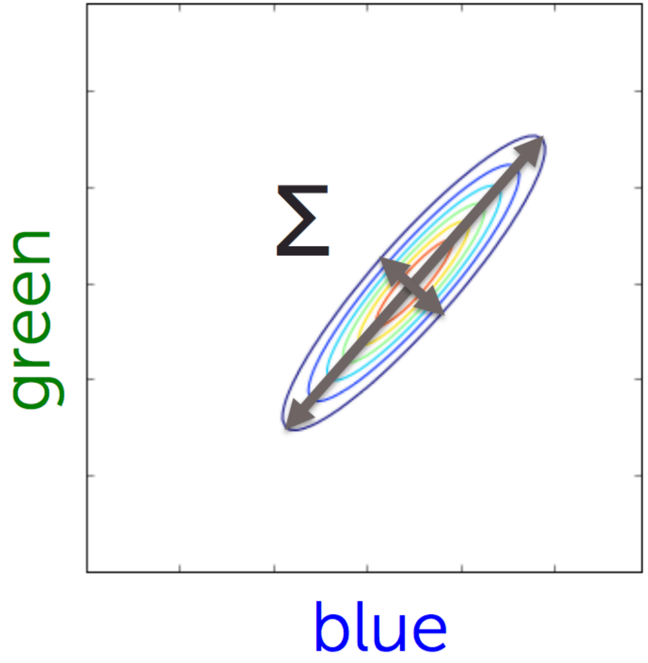
- 2D Gaussian

Fully specified by **mean** μ and **covariance** Σ

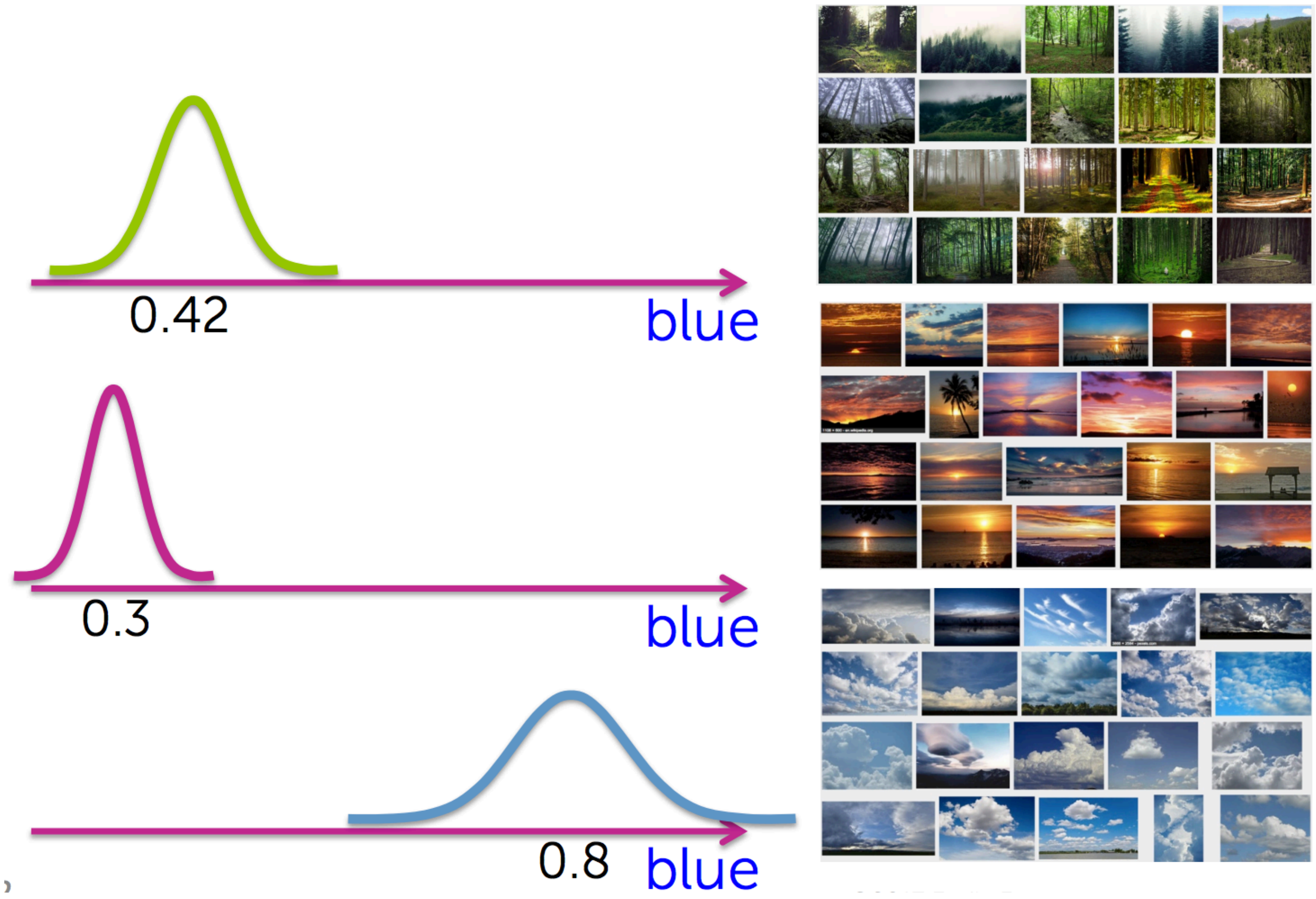
$$\boldsymbol{\mu} = [\mu_{blue}, \mu_{green}]$$

$$\Sigma = \begin{pmatrix} \sigma_{blue}^2 & \sigma_{blue,green} \\ \sigma_{green,blue} & \sigma_{green}^2 \end{pmatrix}$$

covariance determines
orientation + spread

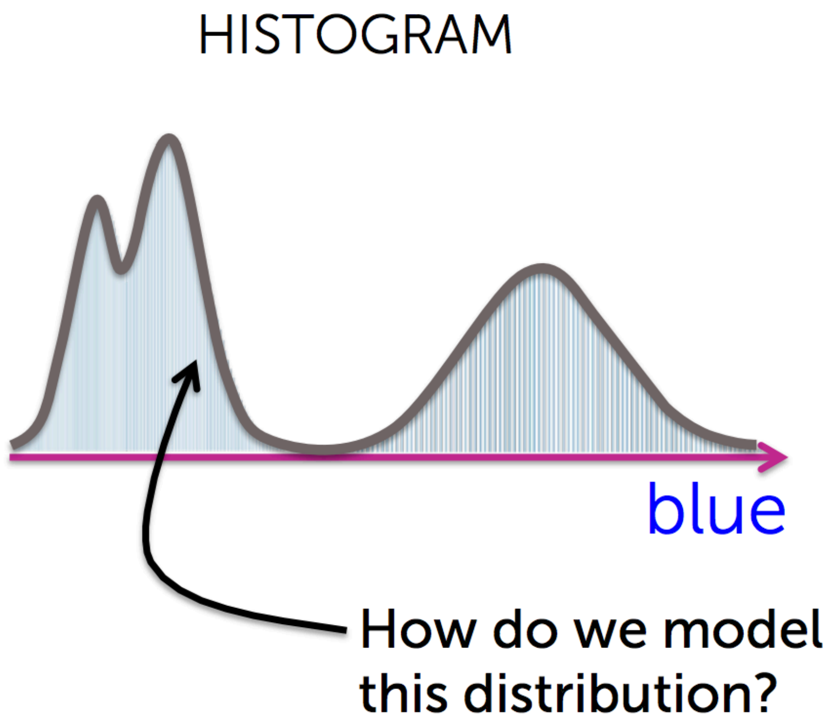


Gaussian per each cluster/group



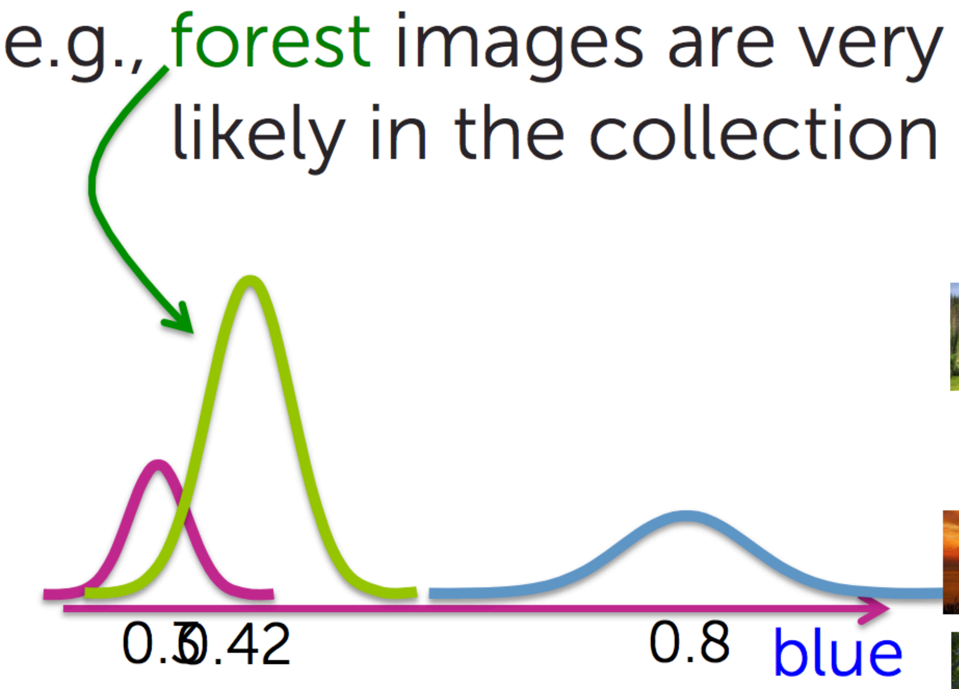
In reality

- No labels



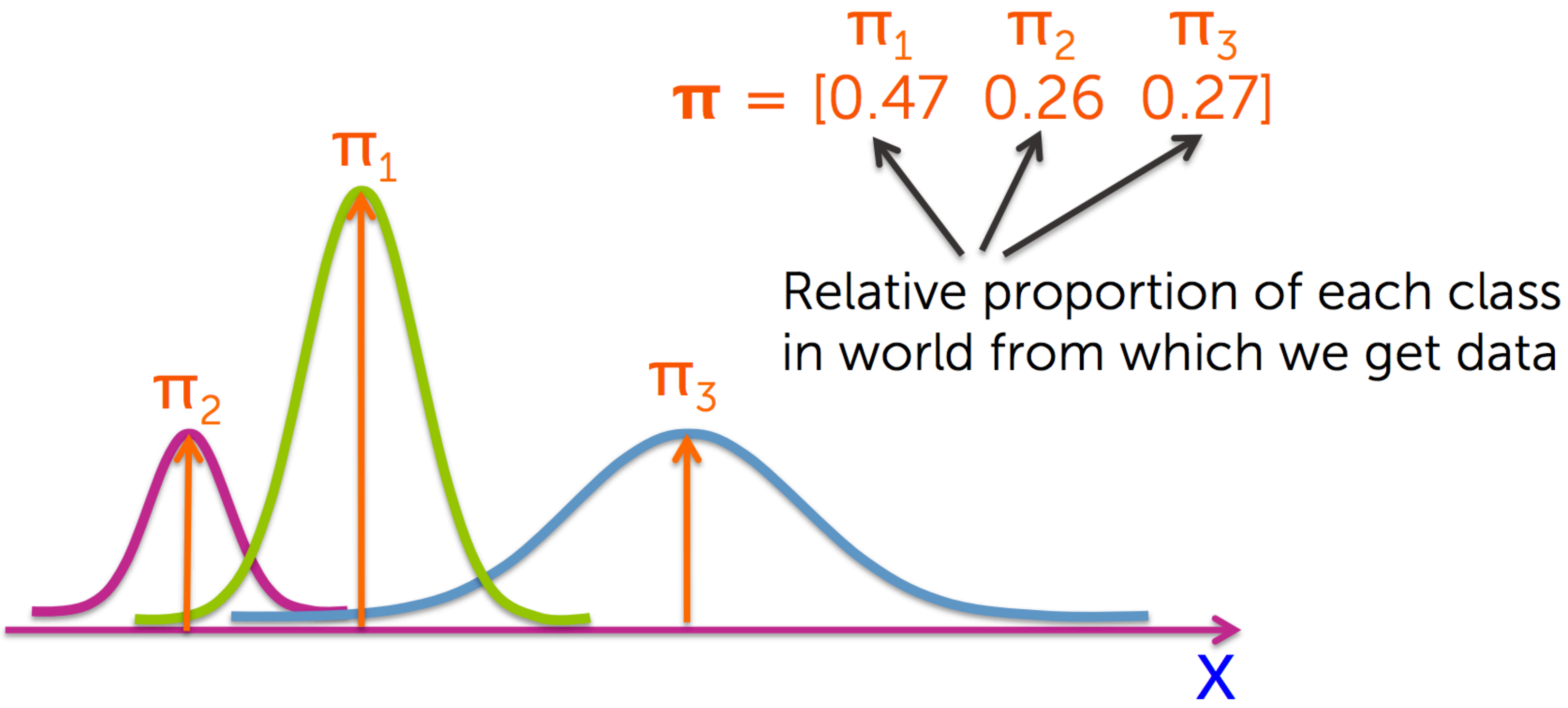
In reality

- No labels



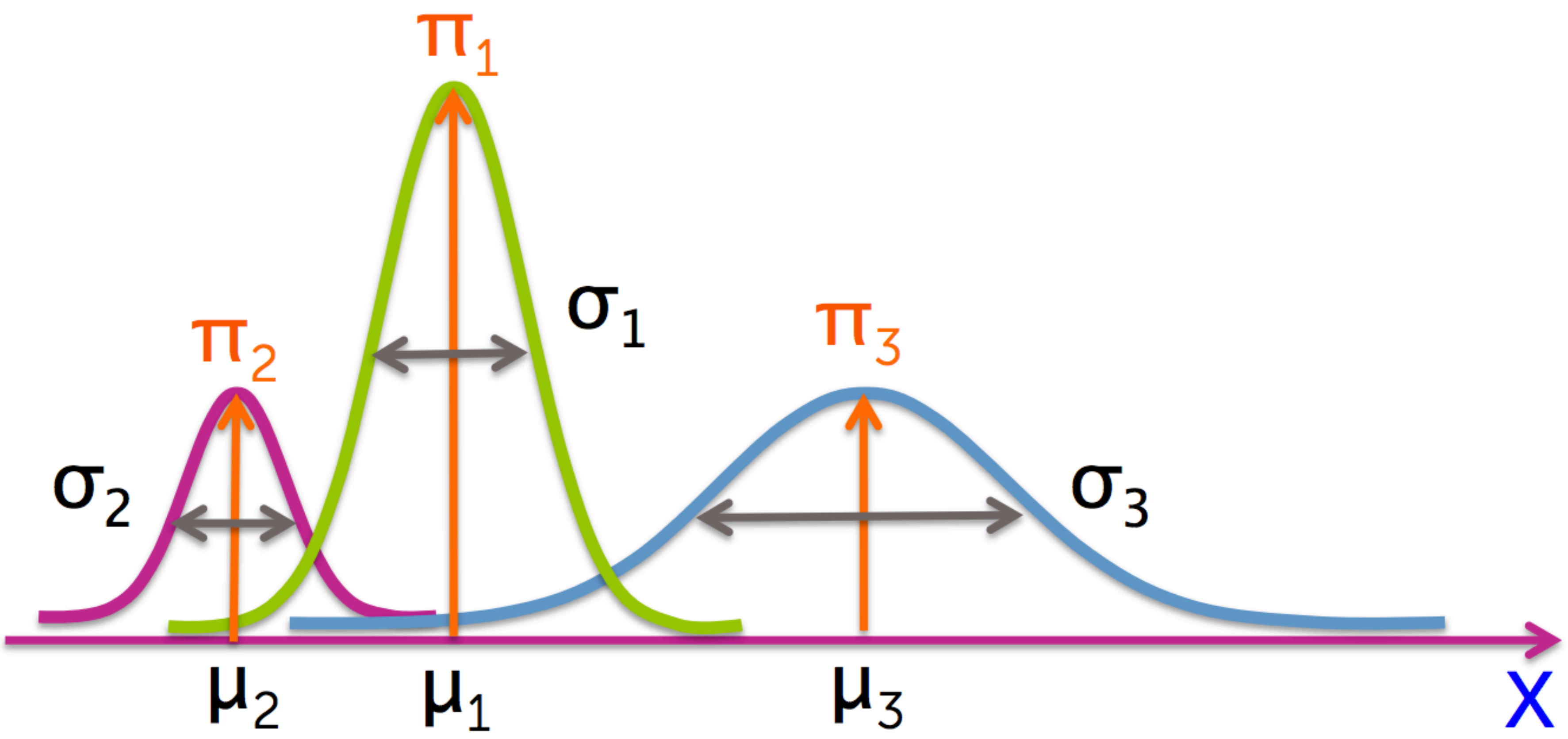
Mixture of Gaussian

- Weighting each Gaussian



Mixture of Gaussian

- Characterize each component



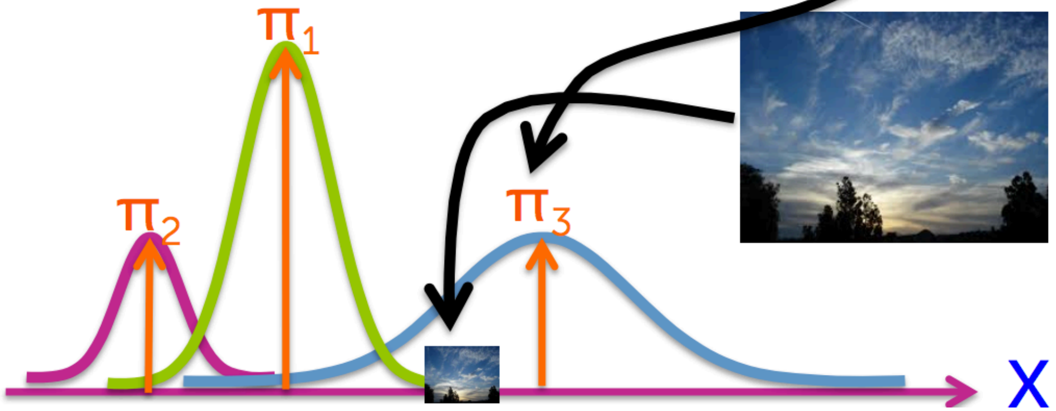
Mixture of Gaussian

Without observing the image content, what's the probability it's from cluster k? (e.g., prob. of seeing "clouds" image)

$$p(z_i = k) = \pi_k$$

Given observation **x**$_i$ is from cluster k, what's the likelihood of seeing **x**$_i$?
(e.g., just look at distribution for "clouds")

$$p(x_i \mid z_i = k, \mu_k, \Sigma_k) = N(x_i | \mu_k, \Sigma_k)$$



Probabilistic ways of assigning centers

- Again, we need to find optimal values of parameters
  - How? Maximum (log) likelihood

$$\ell(\phi, \mu, \Sigma) = \sum_i \log p(x^{(i)}; \phi, \mu, \Sigma)$$

$$\ell(\phi, \mu, \Sigma) = \sum_i \log \sum_{z^{(i)}=1}^{K} p(x^{(i)} | z^{(i)}; \phi, \mu, \Sigma) p(z^{(i)}; \phi)$$

Assumptions on the values of z

- Assume: we know what the value of $z^{(i)}$ is

$$\ell(\phi, \mu, \Sigma) = \sum_i \log p(x^{(i)} | z^{(i)}; \phi, \mu, \Sigma) p(z^{(i)}; \phi)$$

$$\ell(\phi, \mu, \Sigma) = \sum_i \Big(\log p(x^{(i)} | z^{(i)}; \phi, \mu, \Sigma) + \log p(z^{(i)}; \phi) \Big)$$

Recall GDA

- Optimal values of parameters?

$$\phi_j = \frac{\sum_{i=1}^{n_{\text{train}}} \mathbb{1}[z^{(i)} = j]}{n_{\text{train}}}$$

$$\mu_j = \frac{\sum_{i=1}^{n_{\text{train}}} \mathbb{1}[z^{(i)} = j] x^{(i)}}{\sum_{i=1}^{n_{\text{train}}} \mathbb{1}[z^{(i)} = j]}$$

$$\Sigma_j = \frac{\sum_{i=1}^{n_{\text{train}}} \mathbb{1}[z^{(i)} = j](x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^\mathsf{T}}{\sum_{i=1}^{n_{\text{train}}} \mathbb{1}[z^{(i)} = j]}$$

Expectation-Maximization algorithm

- EM algorithm

$$\ell(\phi, \mu, \Sigma) = \sum_i \log \sum_{z^{(i)}=1}^{K} p(x^{(i)}|z^{(i)}; \phi, \mu, \Sigma) p(z^{(i)}; \phi)$$

$$\ell(\phi, \mu, \Sigma) = \sum_i \left(\log p(x^{(i)}|z^{(i)}; \phi, \mu, \Sigma) + \log p(z^{(i)}; \phi) \right)$$

  - Expectation: guess the values of $z^{(i)}$
  - Maximization: update the model parameter based on the guess

EM algorithm

- Algorithm
  - E-step:

$$w_j^{(i)} = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

  - M-step:

$$\phi_j = \frac{1}{n} \sum_i w_j^{(i)}$$

$$\mu_j = \frac{\sum_i w_j^{(i)} x^{(i)}}{\sum_i w_j^{(i)}} \qquad \Sigma_j = \frac{\sum_i w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^{\mathsf{T}}}{\sum_i w_j^{(i)}}$$

Maximum likelihood estimation from soft assignments

Just like in boosting with weighted observations…

R	G	B	r_{i1}	r_{i2}	r_{i3}
$\mathbf{x}_1[1]$	$\mathbf{x}_1[2]$	$\mathbf{x}_1[3]$	0.30	0.18	0.52
$\mathbf{x}_2[1]$	$\mathbf{x}_2[2]$	$\mathbf{x}_2[3]$	0.01	0.26	0.73
$\mathbf{x}_3[1]$	$\mathbf{x}_3[2]$	$\mathbf{x}_3[3]$	0.002	0.008	0.99
$\mathbf{x}_4[1]$	$\mathbf{x}_4[2]$	$\mathbf{x}_4[3]$	0.75	0.10	0.15
$\mathbf{x}_5[1]$	$\mathbf{x}_5[2]$	$\mathbf{x}_5[3]$	0.05	0.93	0.02
$\mathbf{x}_6[1]$	$\mathbf{x}_6[2]$	$\mathbf{x}_6[3]$	0.13	0.86	0.01

52% chance this obs is in cluster 3

Total weight in cluster: (effective # of obs)	**1.242**	**2.8**	**2.42**

Maximum likelihood estimation from soft assignments

R	G	B	r_{i1}	r_{i2}	r_{i3}
$\mathbf{x}_1[1]$	$\mathbf{x}_1[2]$	$\mathbf{x}_1[3]$	0.30	0.18	0.52
$\mathbf{x}_2[1]$	$\mathbf{x}_2[2]$	$\mathbf{x}_2[3]$	0.01	0.26	0.73
$\mathbf{x}_3[1]$	$\mathbf{x}_3[2]$	$\mathbf{x}_3[3]$	0.002	0.008	0.99
$\mathbf{x}_4[1]$	$\mathbf{x}_4[2]$	$\mathbf{x}_4[3]$	0.75	0.10	0.15
$\mathbf{x}_5[1]$	$\mathbf{x}_5[2]$	$\mathbf{x}_5[3]$	0.05	0.93	0.02
$\mathbf{x}_6[1]$	$\mathbf{x}_6[2]$	$\mathbf{x}_6[3]$	0.13	0.86	0.01

MLE of cluster proportions $\hat{\pi}_k$

r_{i1}	r_{i2}	r_{i3}
0.30	0.18	0.52
0.01	0.26	0.73
0.002	0.008	0.99
0.75	0.10	0.15
0.05	0.93	0.02
0.13	0.86	0.01

Total weight in cluster:	**1.242**	**2.8**	**2.42**

Total weight in dataset: 6

datapoints N

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

$$N_k^{\text{soft}} = \sum_{i=1}^{N} r_{ik}$$

Total weight in cluster k = effective # obs

Estimate cluster proportions from relative weights

Defaults to hard assignment case when r_{ij} in {0,1}

Hard assignments have:

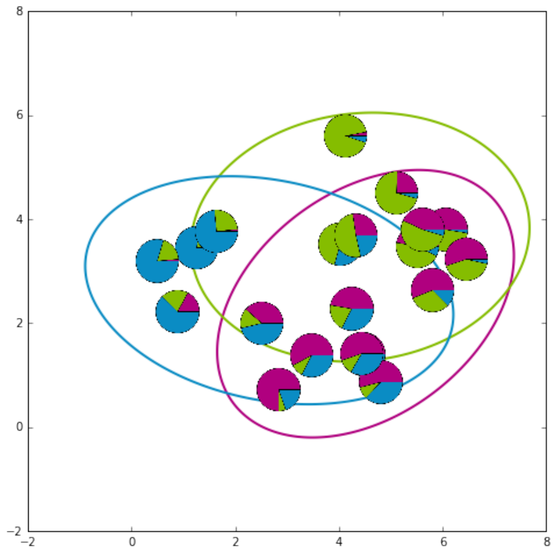
$$r_{ik} = \begin{cases} 1 & i \text{ in } k \\ 0 & \text{otherwise} \end{cases}$$

R	G	B	r_{i1}	r_{i2}	r_{i3}
$\mathbf{x}_1[1]$	$\mathbf{x}_1[2]$	$\mathbf{x}_1[3]$	0	0	1
$\mathbf{x}_2[1]$	$\mathbf{x}_2[2]$	$\mathbf{x}_2[3]$	0	0	1
$\mathbf{x}_3[1]$	$\mathbf{x}_3[2]$	$\mathbf{x}_3[3]$	0	0	1
$\mathbf{x}_4[1]$	$\mathbf{x}_4[2]$	$\mathbf{x}_4[3]$	1	0	0
$\mathbf{x}_5[1]$	$\mathbf{x}_5[2]$	$\mathbf{x}_5[3]$	0	1	0
$\mathbf{x}_6[1]$	$\mathbf{x}_6[2]$	$\mathbf{x}_6[3]$	0	1	0

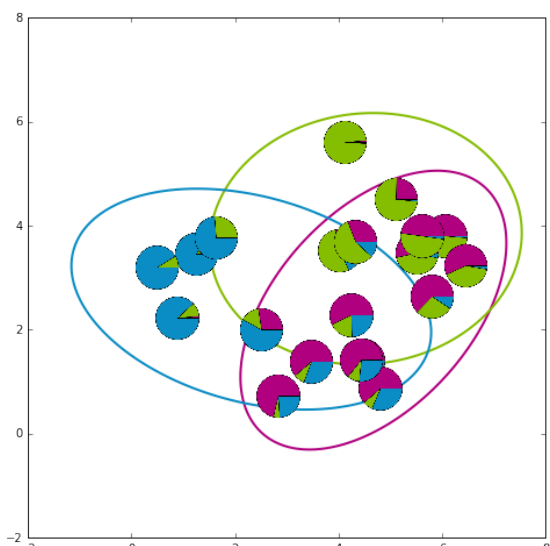
Total weight in cluster:	**1**	**2**	**3**

One-hot encoding of cluster assignment

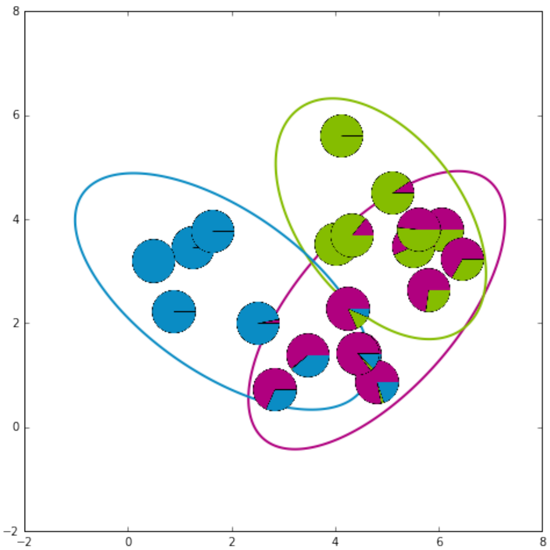
EM for mixtures of Gaussians in pictures – initialization



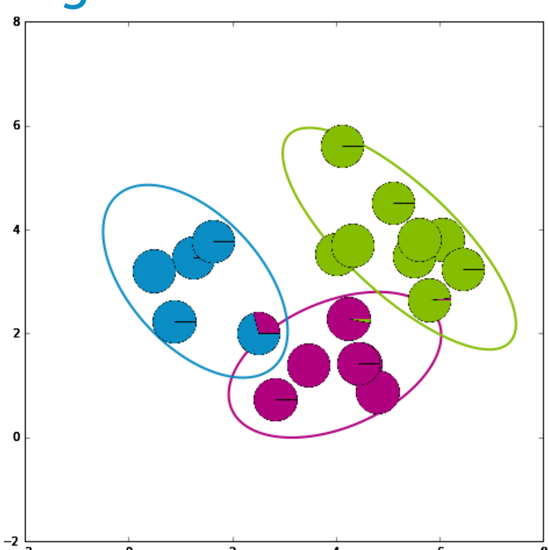
EM for mixtures of Gaussians in pictures – after 1st iteration



EM for mixtures of Gaussians in pictures – after 2nd iteration



EM for mixtures of Gaussians in pictures – converged solution



How do we compute E-step?

- Bayes' rule!

$$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_l p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$