Data Modelling Methods-II

CS771: Introduction to Machine Learning

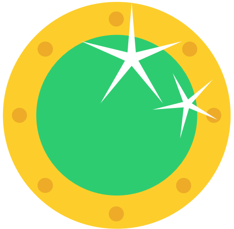
Purushottam Kar

Outline of today's discussion

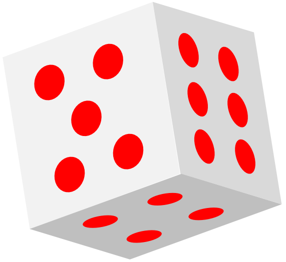
- Revise Naïve Bayes method
- Feature modelling methods for unlabelled data
- Gaussian mixture Models (GMMs)
- The alternating optimization approach to learning GMMs
- The k-means approach to learning GMMs

The Multinoulli Distribution

- Also known as categorical distribution
- Generalizes the Bernoulli distribution
  - Bernoulli models a coin with two outcomes (call them head and tail)
  - … using one “bias” parameter p (taken to be the probability of heads)



- Multinoulli models a K-sided dice
  - … using a K-dimensional vector π
  - π_k is taken to denote the probability of the k-th side turning up
  - $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$



- Conjugate prior for Bernoulli: Beta $\mathbb{P}\left[p;\alpha,\beta\right] = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$
- Conjugate prior for Multinoulli: Dirichlet $\mathbb{P}\left[\boldsymbol{\pi};\boldsymbol{\alpha}\right] = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{L} \boldsymbol{\pi}_i^{\boldsymbol{\alpha}_i - 1}$

App: Email Categorizer using Naïve Bayes

App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

HallPresi	Awesome talk at eSumm
HEC	**Urgent! Mess bill overdue**
HEC	**Urgent! Canteen bill over**

App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

HallPresi	Awesome talk at eSumm
HEC	**Urgent! Mess bill overdue**
HEC	**Urgent! Canteen bill over**

$\mathbf{x}$

talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

HallPresi	Awesome talk at eSumm
HEC	**Urgent! Mess bill overdue**
HEC	**Urgent! Canteen bill over**

$\mathbf{x}$

talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

$$y \in \{ \bullet , \bullet , \bullet \}$$

App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

HallPresi	Awesome talk at eSumm
HEC	**Urgent! Mess bill overdue**
HEC	**Urgent! Canteen bill over**

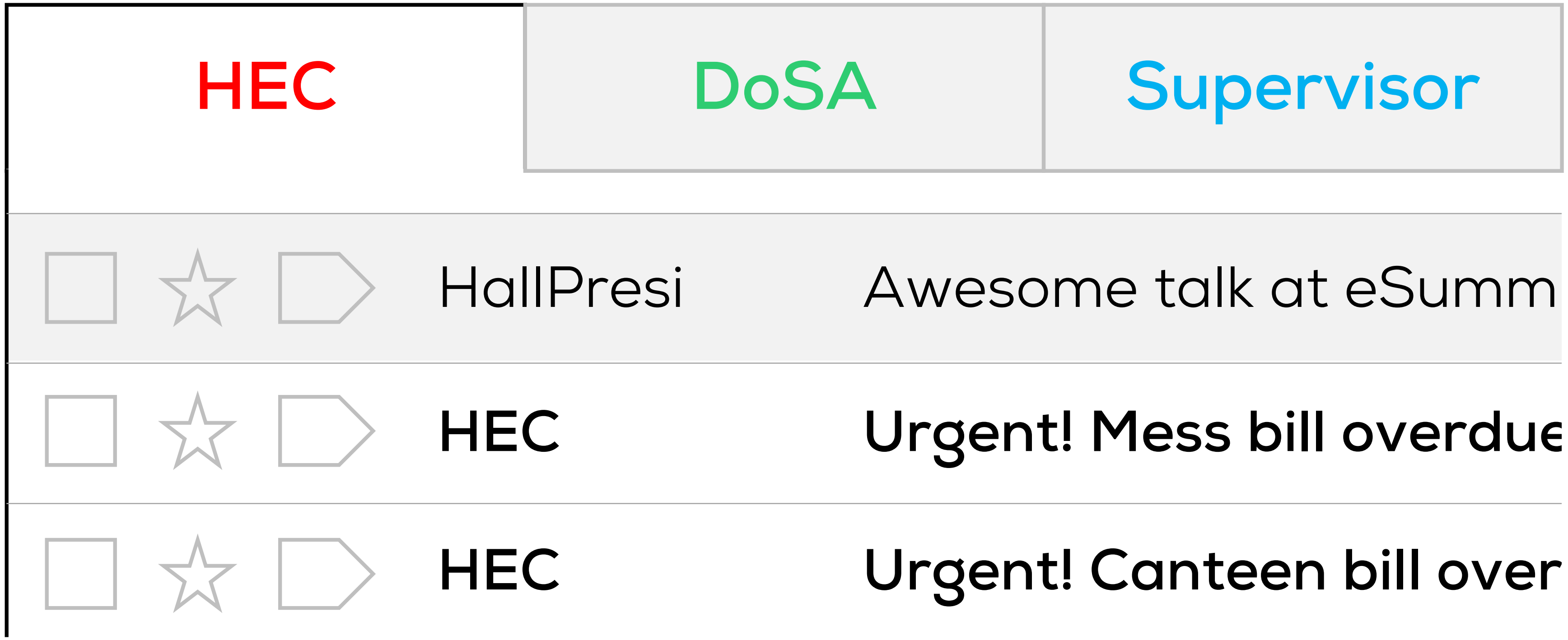
Bag of words feature can record just occurrence or count

$\mathbf{x}$

talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

$y \in \{ \bullet , \bullet , \bullet \}$

App: Email Categorizer using Naïve Bayes



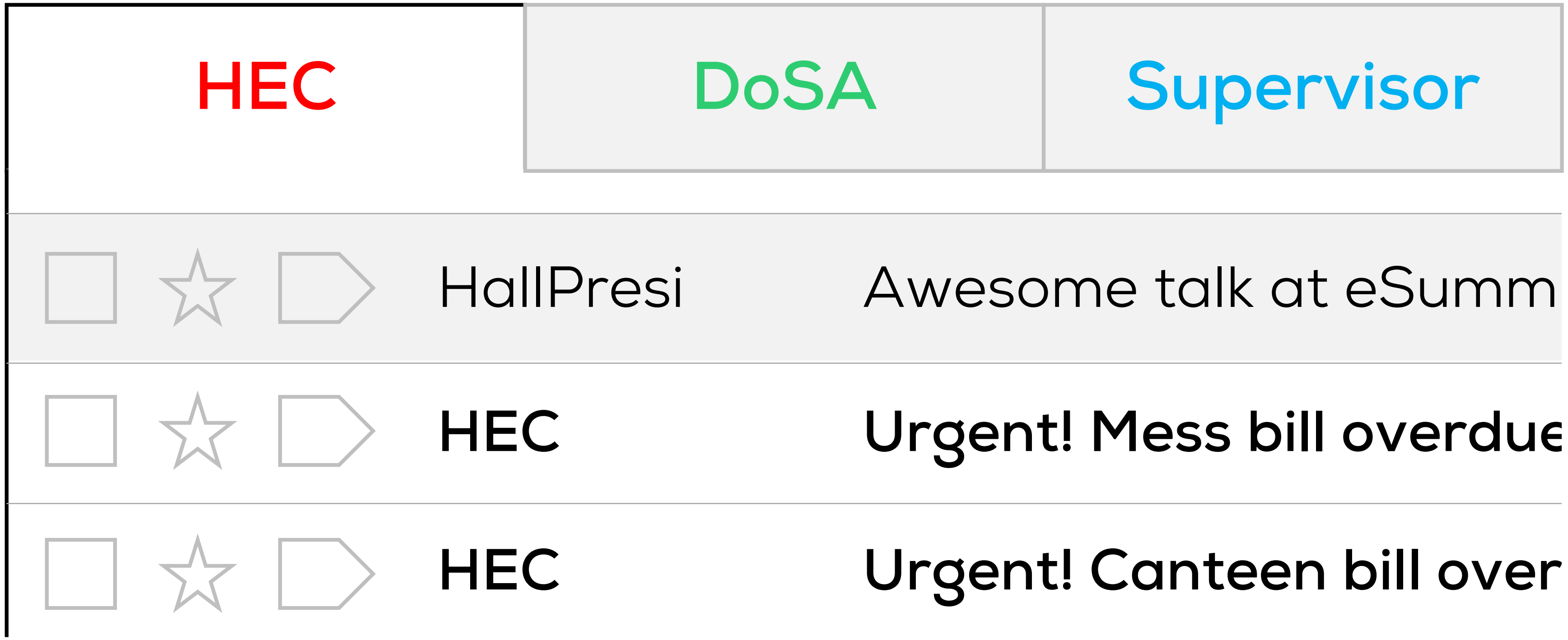
Bag of words feature can record just occurrence or count

$\mathbf{x}$

1	0	0	0	0	0	0	1	0	0	0	0	0
talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

$y \in \{\ ,\ ,\ \}$

App: Email Categorizer using Naïve Bayes

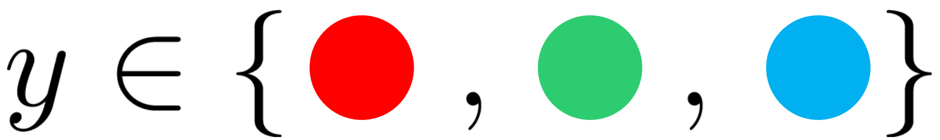


Choice of words crucial – stemming, throw away articles etc

Bag of words feature can record just occurrence or count

x

talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report
1	0	0	0	0	0	0	1	0	0	0	0	0



App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

HallPresi	Awesome talk at eSumm
HEC	**Urgent! Mess bill overdue**
HEC	**Urgent! Canteen bill over**

Commonly used in NLP

Choice of words crucial – stemming, throw away articles etc

Bag of words feature can record just occurrence or count

$\mathbf{x}$

1	0	0	0	0	0	0	1	0	0	0	0	0
talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

$$y \in \{\bullet, \bullet, \bullet\}$$

App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

☐ ☆	HallPresi	Awesome talk at eSumm
☐ ☆	**HEC**	**Urgent! Mess bill overdue**
☐ ☆	**HEC**	**Urgent! Canteen bill over**

Commonly used in NLP

Choice of words crucial – stemming, throw away articles etc

Bag of words feature can record just occurrence or count

$\mathbf{x}$

1	0	0	0	0	0	0	1	0	0	0	0	0
talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

Usually very high dimensional

$y \in \{ \bullet , \bullet , \bullet \}$

App: Email Categorizer using Naïve Bayes

HEC	DoSA	Supervisor

HallPresi	Awesome talk at eSumm
HEC	**Urgent! Mess bill overdue**
HEC	**Urgent! Canteen bill over**

Commonly used in NLP

Choice of words crucial – stemming, throw away articles etc

Bag of words feature can record just occurrence or count

$\mathbf{x}$

1	0	0	0	0	0	0	1	0	0	0	0	0
talk	meet	project	wake	meeting	wallet	keys	awesome	lost	trip	lost	mess	report

$y \in \{ \text{red}, \text{green}, \text{blue} \}$

Usually very high dimensional

Usually very very sparse

App: Email Categorizer using Naïve Bayes

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[\bullet] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[\bullet] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

$$\mathbb{P}[\text{awesome} = 1 \mid \bullet] = \frac{|\#\text{emails from sup. with "awesome"}|}{|\#\text{total emails from supervisor}|}$$

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[\bullet] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

$$\mathbb{P}[\text{awesome} = 1 \mid \bullet] = \frac{|\#\text{emails from sup. with "awesome"}|}{|\#\text{total emails from supervisor}|}$$

At test time ...

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[\bullet] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

$$\mathbb{P}[\text{awesome} = 1 \mid \bullet] = \frac{|\#\text{emails from sup. with "awesome"}|}{|\#\text{total emails from supervisor}|}$$

At test time …

$$\mathbb{P}[\mathbf{x}^t, \bullet] = \mathbb{P}[\bullet] \cdot \prod_{j=1}^{d} \mathbb{P}[\mathbf{x}_i^t \mid \bullet]$$

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[\bullet] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

$$\mathbb{P}[\text{awesome} = 1 \mid \bullet] = \frac{|\#\text{emails from sup. with "awesome"}|}{|\#\text{total emails from supervisor}|}$$

$\mathbb{P}[awesome = 0|\bullet]$
$= 1 - \mathbb{P}[awesome = 1|\bullet]$

At test time ...

$$\mathbb{P}[\mathbf{x}^t, \bullet] = \mathbb{P}[\bullet] \cdot \prod_{j=1}^{d} \mathbb{P}[\mathbf{x}_i^t \mid \bullet]$$

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[{\color{cyan}\bullet}] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

$$\mathbb{P}[\text{awesome} = 1 \mid {\color{cyan}\bullet}] = \frac{|\#\text{emails from sup. with "awesome"}|}{|\#\text{total emails from supervisor}|}$$

$$\mathbb{P}[awesome = 0|{\color{cyan}\bullet}]$$
$$= 1 - \mathbb{P}[awesome = 1|{\color{cyan}\bullet}]$$

At test time …

$$\mathbb{P}[\mathbf{x}^t, {\color{cyan}\bullet}] = \mathbb{P}[{\color{cyan}\bullet}] \cdot \prod_{j=1}^{d} \mathbb{P}[\mathbf{x}_i^t | {\color{cyan}\bullet}]$$

$$\hat{y}^t = \arg\max\{\mathbb{P}[\mathbf{x}^t, {\color{red}\bullet}], \mathbb{P}[\mathbf{x}^t, {\color{green}\bullet}], \mathbb{P}[\mathbf{x}^t, {\color{cyan}\bullet}]\}$$

App: Email Categorizer using Naïve Bayes

$$\mathbb{P}[\textcolor{cyan}{\bullet}] = \frac{|\#\text{emails from supervisor}|}{|\#\text{total emails}|}$$

$$\mathbb{P}[\text{awesome} = 1 \mid \textcolor{cyan}{\bullet}] = \frac{|\#\text{emails from sup. with "awesome"}|}{|\#\text{total emails from supervisor}|}$$

$$\mathbb{P}[awesome = 0|\textcolor{cyan}{\bullet}] = 1 - \mathbb{P}[awesome = 1|\textcolor{cyan}{\bullet}]$$

At test time …

$$\mathbb{P}[\mathbf{x}^t, \textcolor{cyan}{\bullet}] = \mathbb{P}[\textcolor{cyan}{\bullet}] \cdot \prod_{j=1}^{d} \mathbb{P}\left[\mathbf{x}_i^t \mid \textcolor{cyan}{\bullet}\right]$$

Will give the same result as

$$\arg\max\{\mathbb{P}[\textcolor{red}{\bullet} \mid \mathbf{x}^t], \mathbb{P}[\textcolor{green}{\bullet} \mid \mathbf{x}^t], \mathbb{P}[\textcolor{cyan}{\bullet} \mid \mathbf{x}^t]\}$$

$$\hat{y}^t = \arg\max\{\mathbb{P}[\mathbf{x}^t, \textcolor{red}{\bullet}], \mathbb{P}[\mathbf{x}^t, \textcolor{green}{\bullet}], \mathbb{P}[\mathbf{x}^t, \textcolor{cyan}{\bullet}]\}$$

App: Automatic Email Generator!

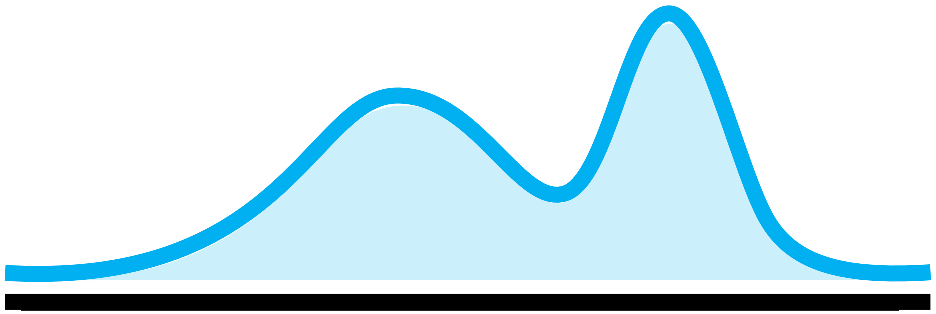
Class proportions

- Choose a category from {HEC, DoSA,Supervisor}
  - Toss a 3-sided "coin" aka categorical/multinoulii distribution using $\mathbb{P}[\bullet]$
  - Say we chose DoSA
- For each word in your dictionary of d words, toss a Bernoulli coin to decide whether to include that word in the mail or not
  - For $j \in [d]$, toss a coin that lands heads with probability $\mathbb{P}\left[x_j = 1 \mid \bullet\right]$
- Collect all words for which the toss landed heads
- Compose an email using only those words (and maybe a few articles, prepositions etc)
- Congratulations, you can now ask the dean to stop sending you emails – you will generate them yourself!

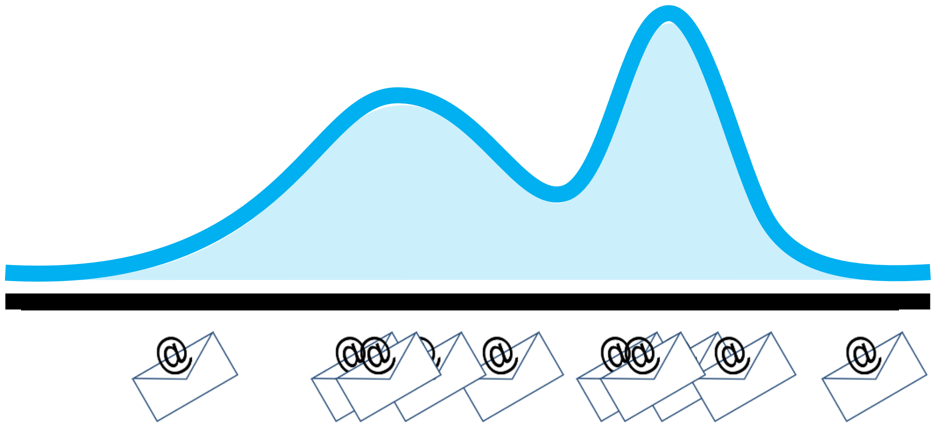
Already learnt from training data!

The generative story for unlabelled data??

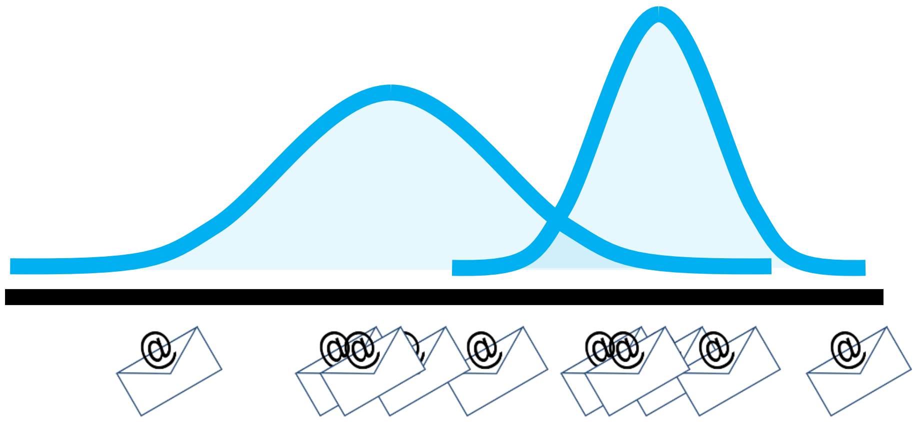
The generative story for unlabelled data??



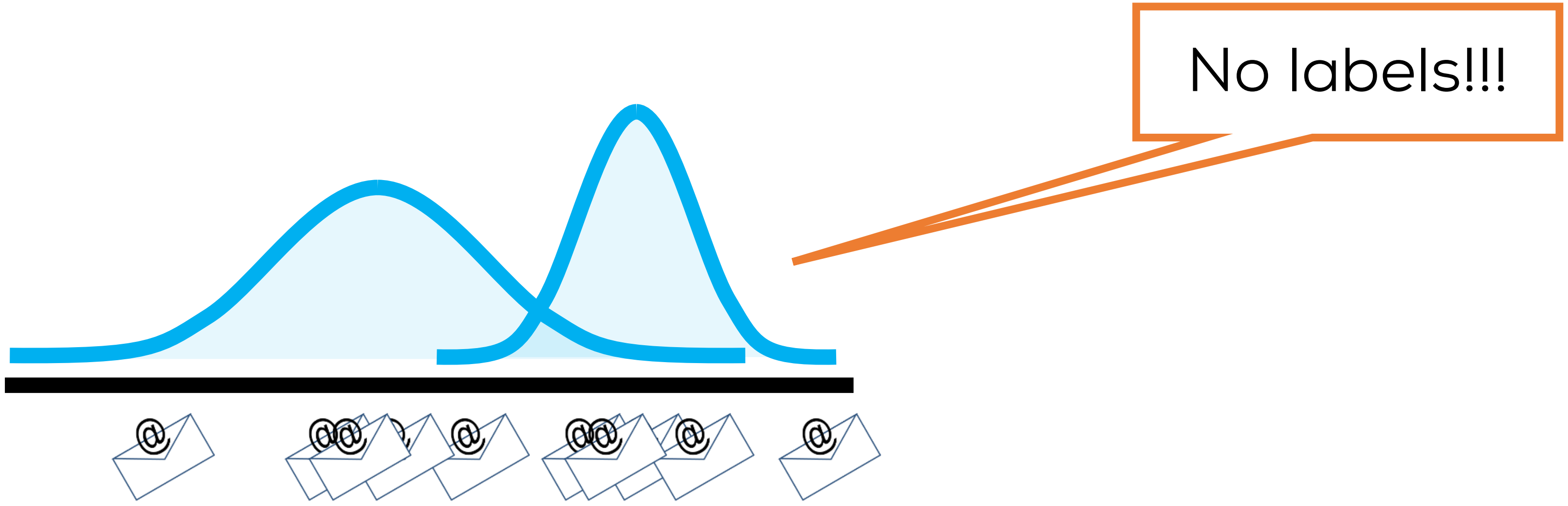
The generative story for unlabelled data??



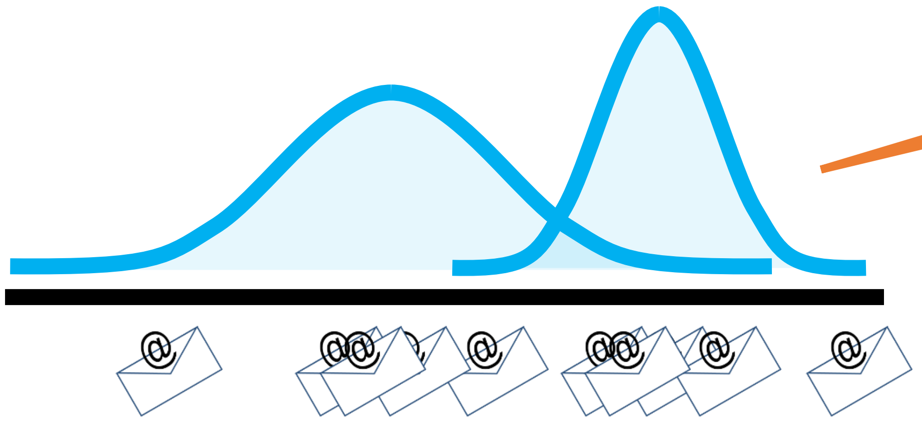
The generative story for unlabelled data??



The generative story for unlabelled data??



The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

How to grade a CS771 exam??

How to grade a CS771 exam??

gradescope

How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks

How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks

$\frac{x^2}{2} + b$

$x^2 + d$

$x^2/2$

$x^2/2$

$x^2 + C$

$\frac{x^2}{2} + c$

$x^2/2$

$x^2/2 + a$

$x^2 + \epsilon$

How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks

$\frac{x^2}{2} + a$

$\frac{x^2}{2} + b\frac{x^2}{2} + c$

$x^2/2$

$x^2/2 \quad x^2/2$

$x^2 + cx^2 + d$

$x^2 + \varepsilon$

How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks

$\frac{x^2}{2}$ $\quad x^2/2$ $\quad X^2/2$

$\frac{X^2}{2} + a$

$\frac{x^2}{2} + b$ $\quad \frac{x^2}{2} + c$

$x^2 + c$ $\quad X^2 + d$

$X^2 + \varepsilon$

How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks

$x^2/2$ $x^2/2$ $x^2/2$

4/5

$\frac{x^2}{2} + a$ $\frac{x^2}{2} + b$ $\frac{x^2}{2} + c$

5/5

$x^2 + c$ $x^2 + d$ $x^2 + e$

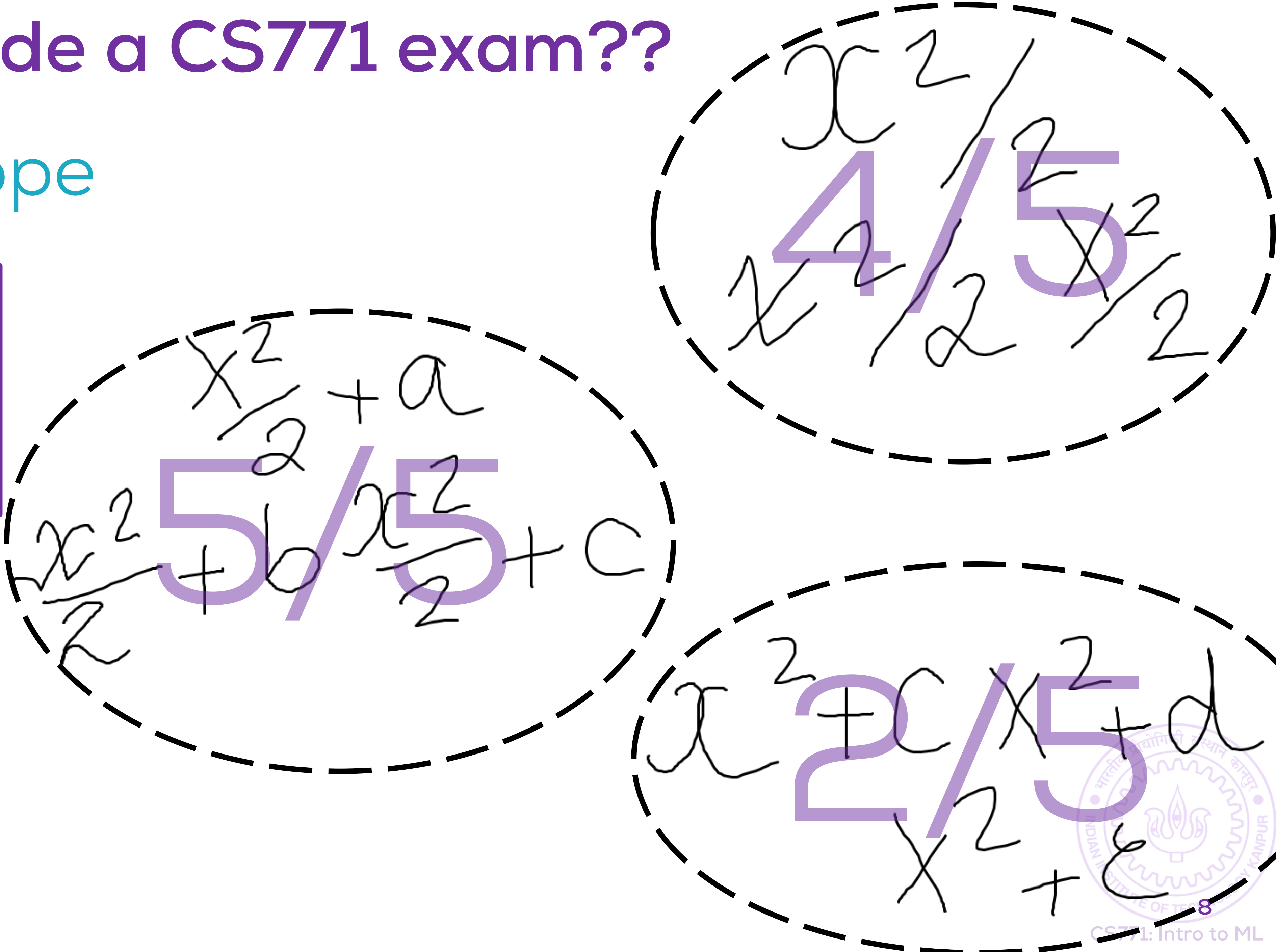
2/5

How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks



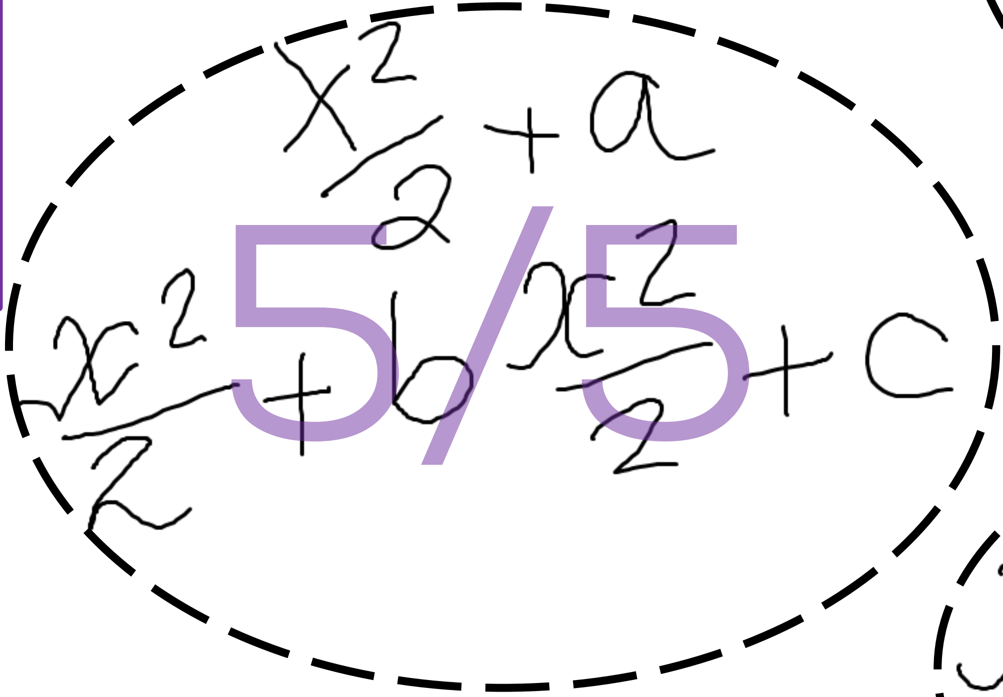
Clustering!

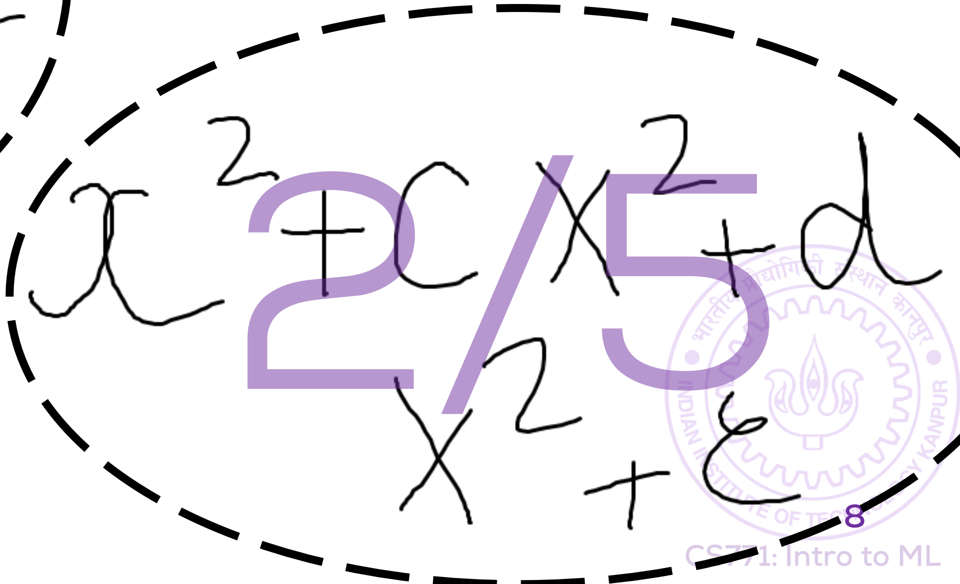
How to grade a CS771 exam??

gradescope

Q1. $\int x = ?$

5 marks

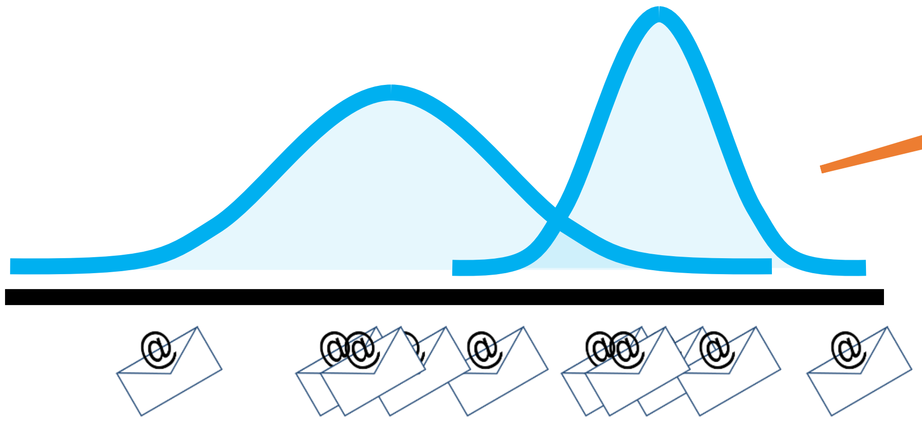




Clustering!

Like classification without labels ☺

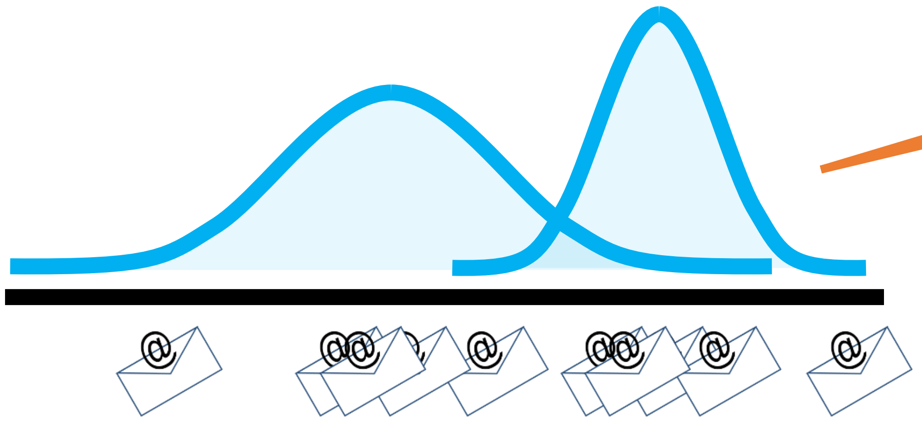
The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

The generative story for unlabelled data??

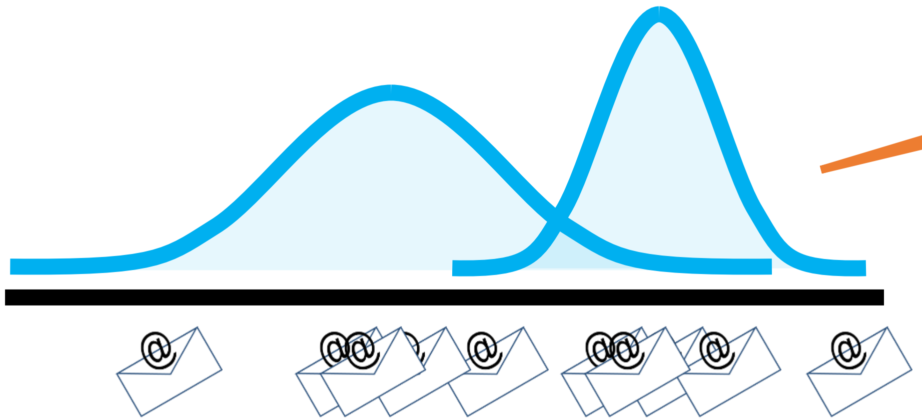


No labels!!!

Can we still recover the two Gaussian components in the mixture??

How do we know there are only two Gaussians?

The generative story for unlabelled data??



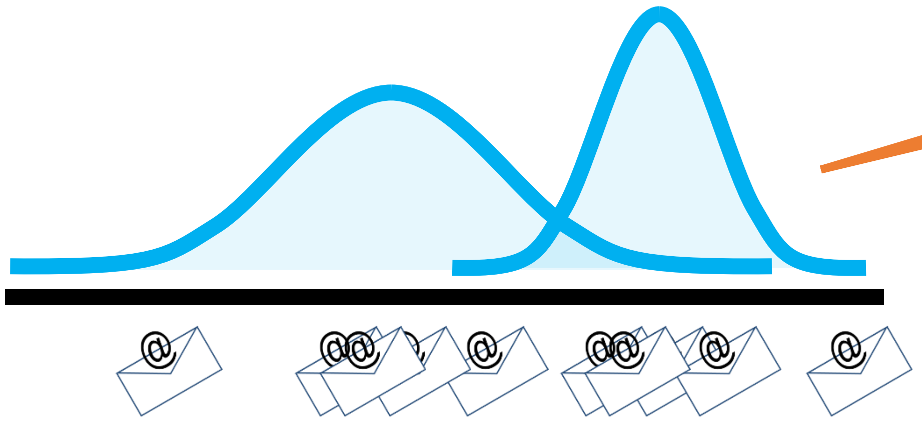
No labels!!!

Gaussians can be a hyper-parameter K you can tune

Can we still recover the two Gaussian components in the mixture??

How do we know there are only two Gaussians?

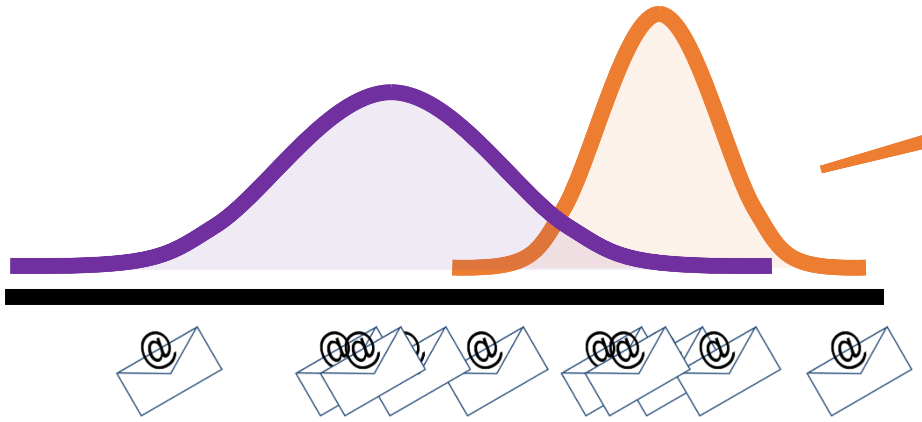
The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

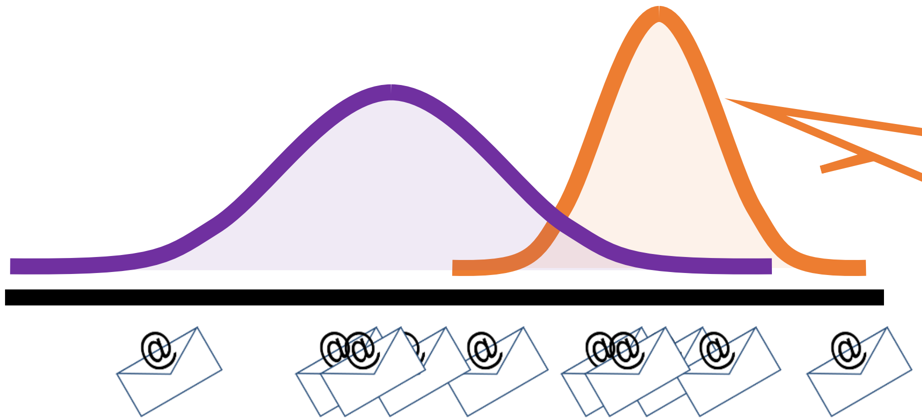
The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

The generative story for unlabelled data??

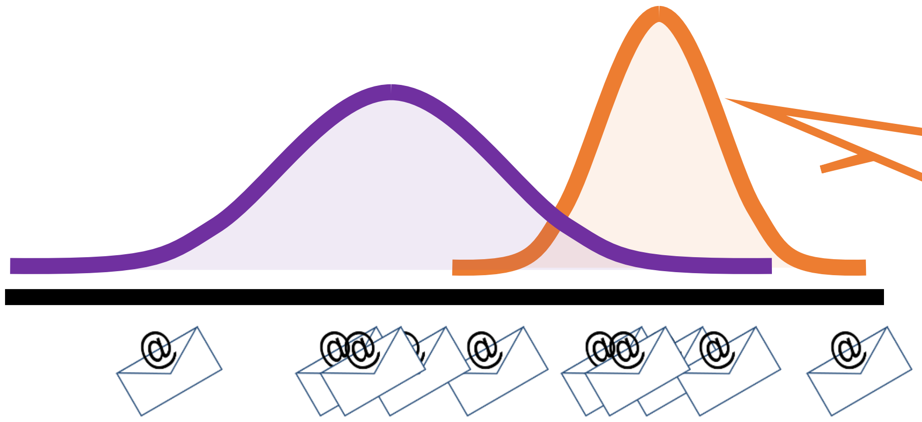


No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

The generative story for unlabelled data??



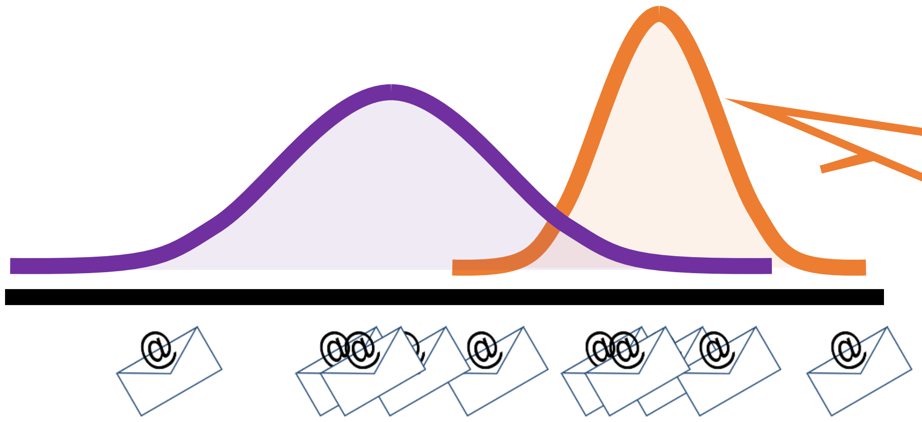
No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

The generative story for unlabelled data??



No labels!!!

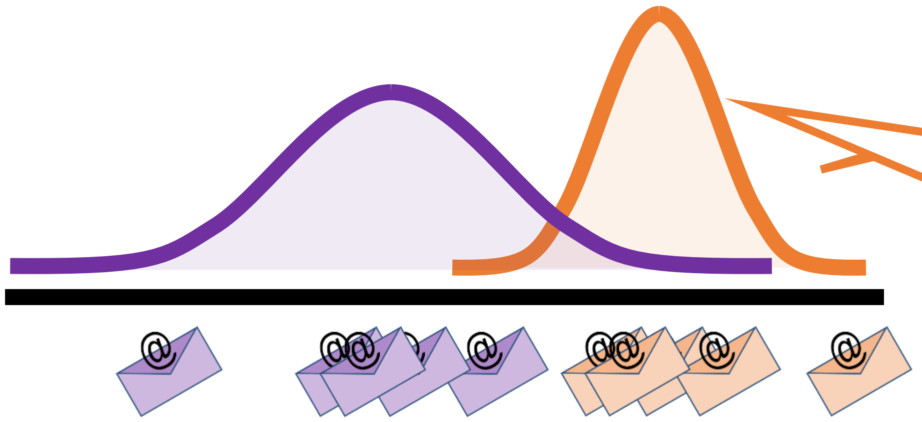
Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

The generative story for unlabelled data??



No labels!!!

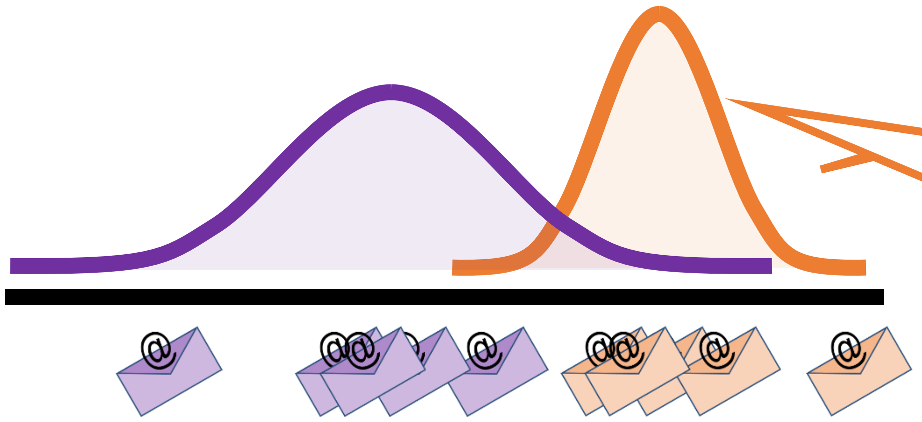
Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

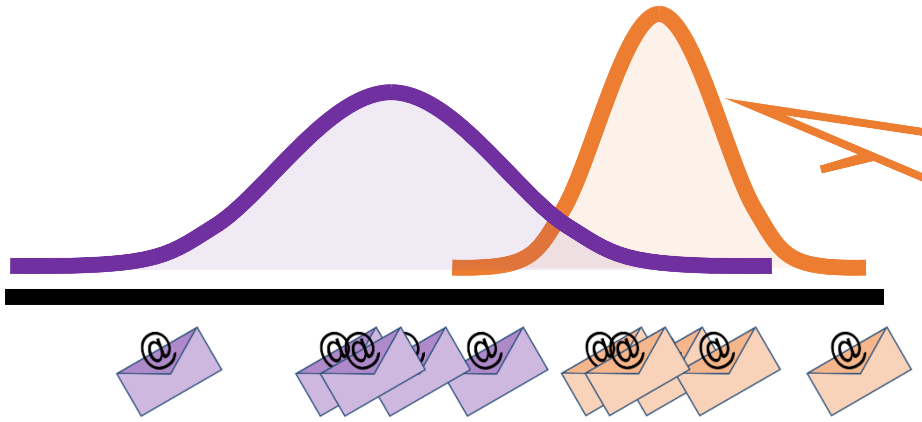
Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

Can estimate the class proportions, and means and variances of these 1D Gaussians using training data!

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

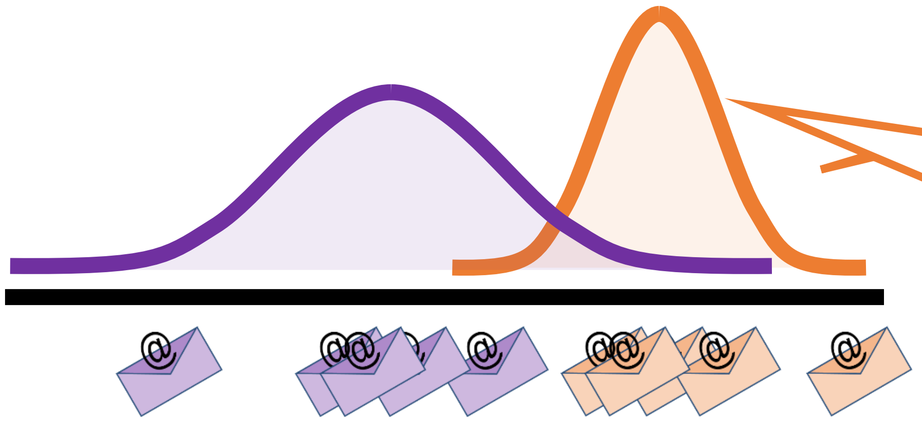
We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

Saw how to do this for Bernoulli distributions last time.

Can estimate the class proportions, and means and variances of these 1D Gaussians using training data!

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

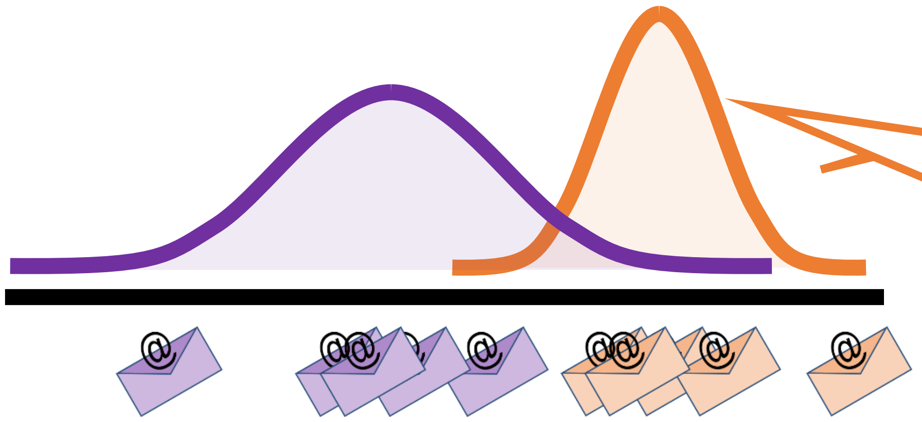
Saw how to do this for Bernoulli distributions last time.

Hmm … what if someone magically labelled the emails with the color

Can estimate the class proportions, and means and variances of these 1D Gaussians using training data!

Read **[DAU]** Sections 9.1-9.5

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Better take a look now ☺

Saw how to do this for Bernoulli distributions last time.

Hmm … what if someone magically labelled the emails with the color

Can estimate the class proportions, and means and variances of these 1D Gaussians using training data!

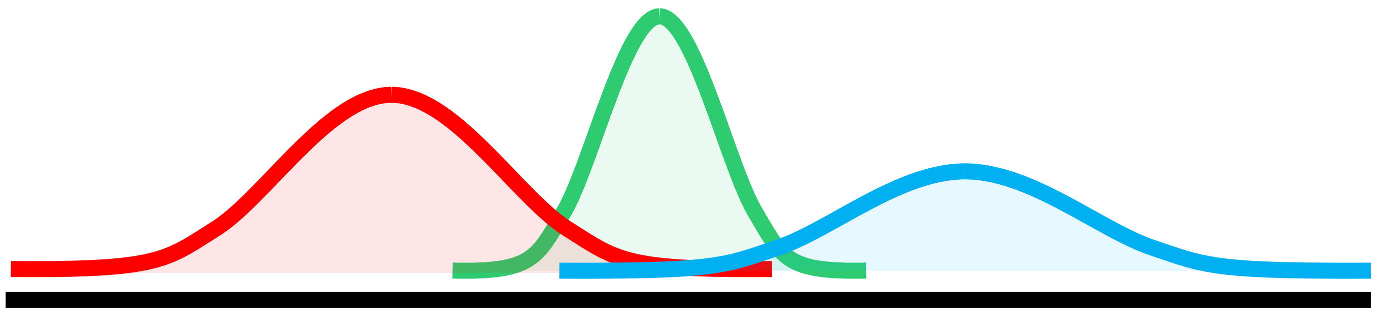
Read [**DAU**] Sections 9.1-9.5

$\langle$detour$\rangle$

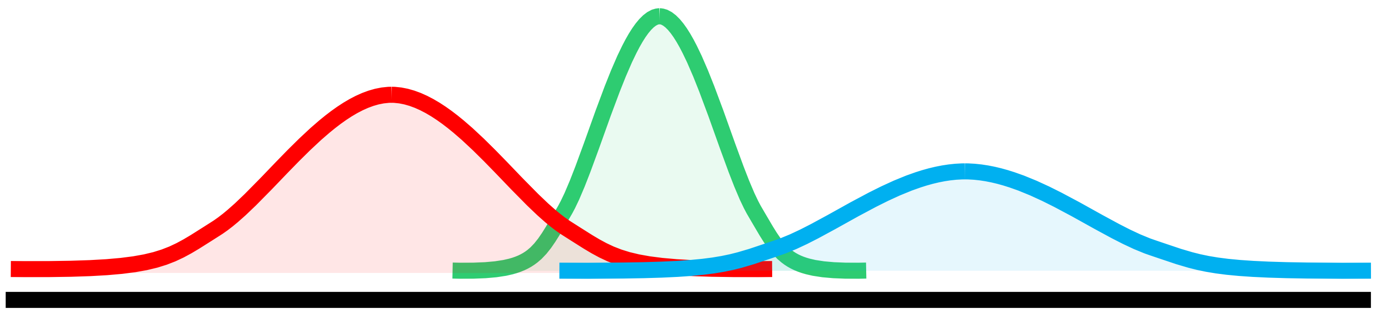
Learning a mixture of Gaussians in presence of labels

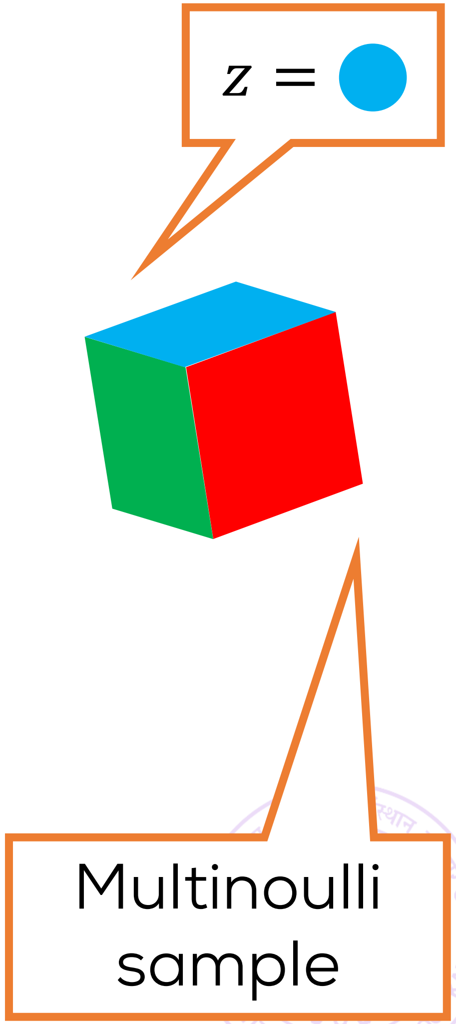
The generative story for labelled data

The generative story for labelled data

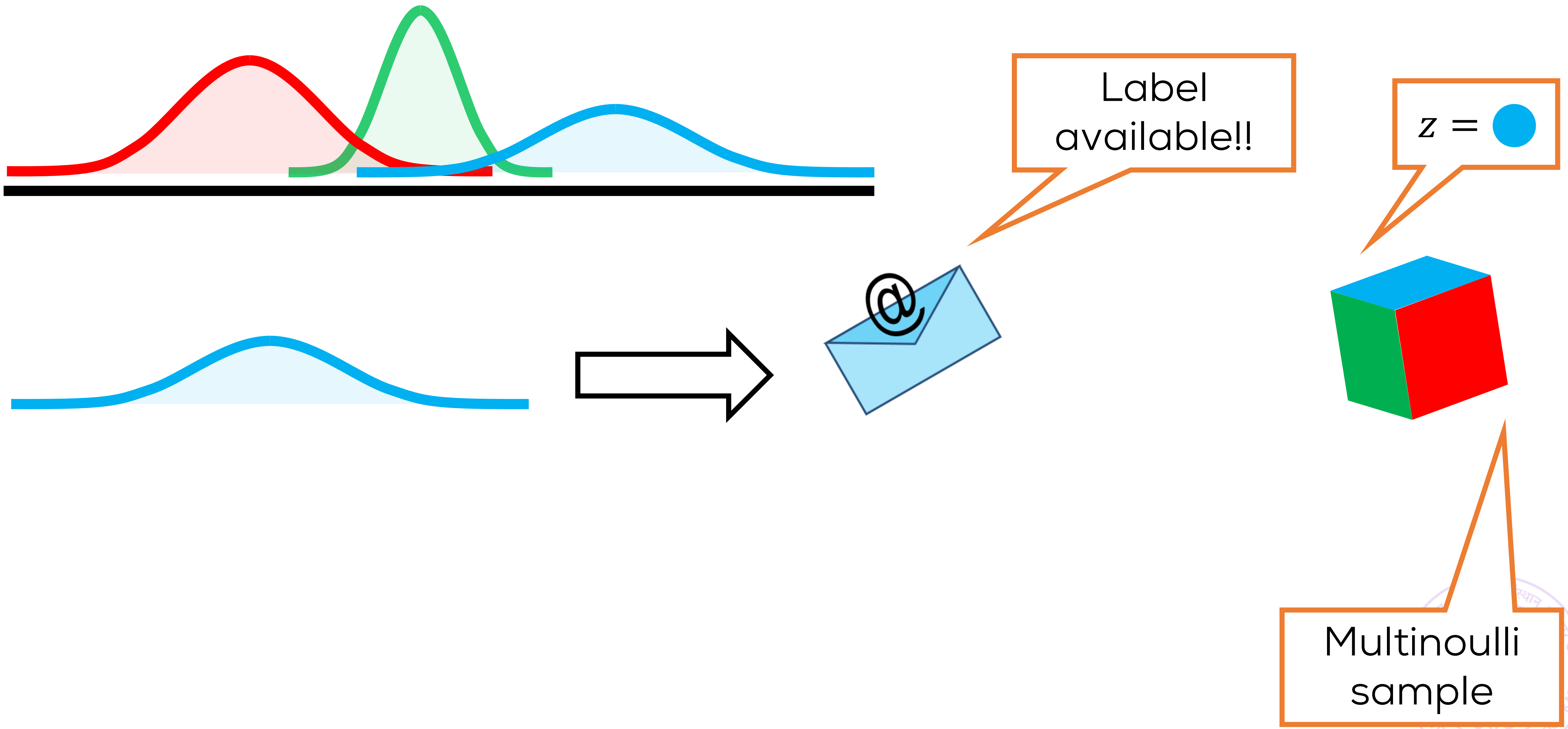


The generative story for labelled data

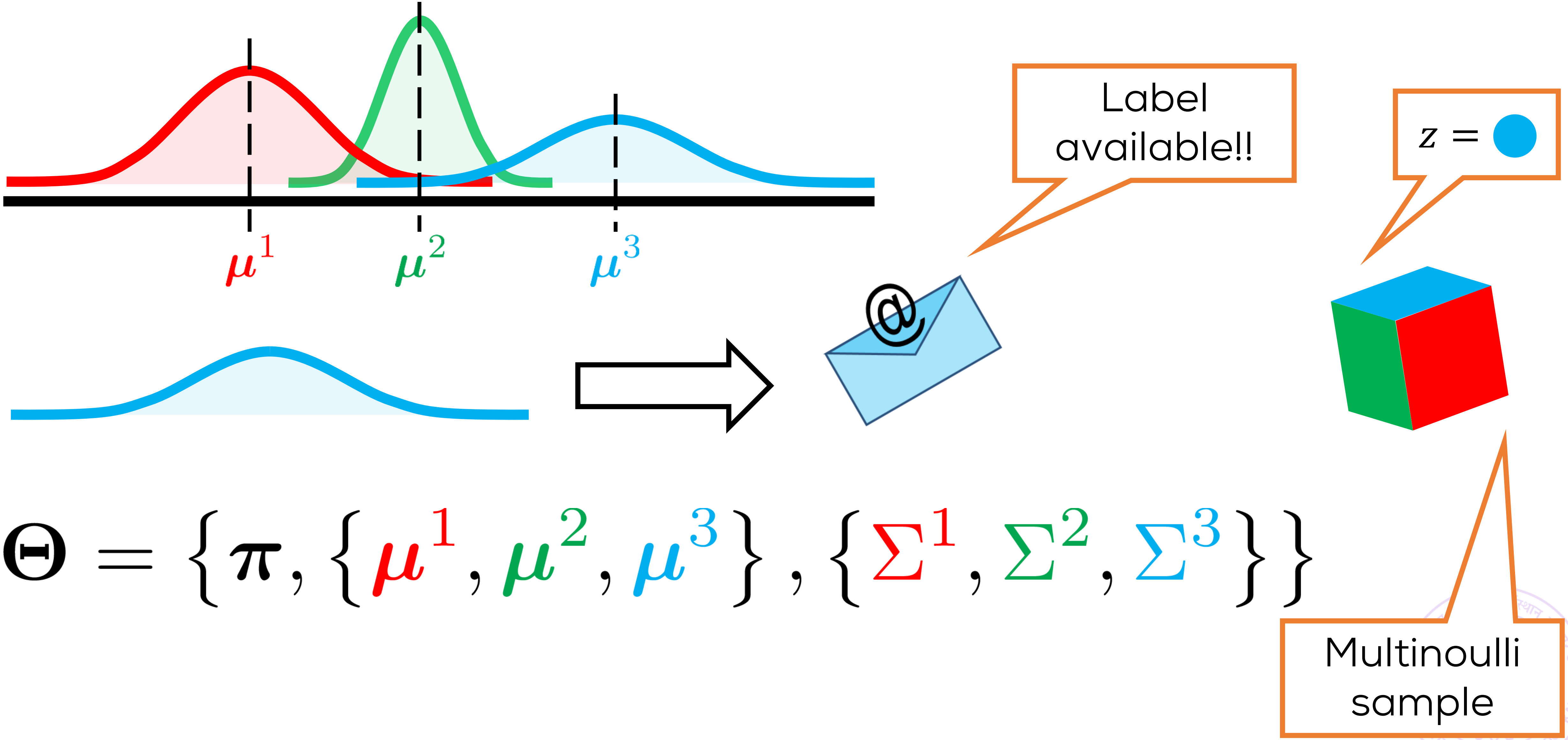




The generative story for labelled data

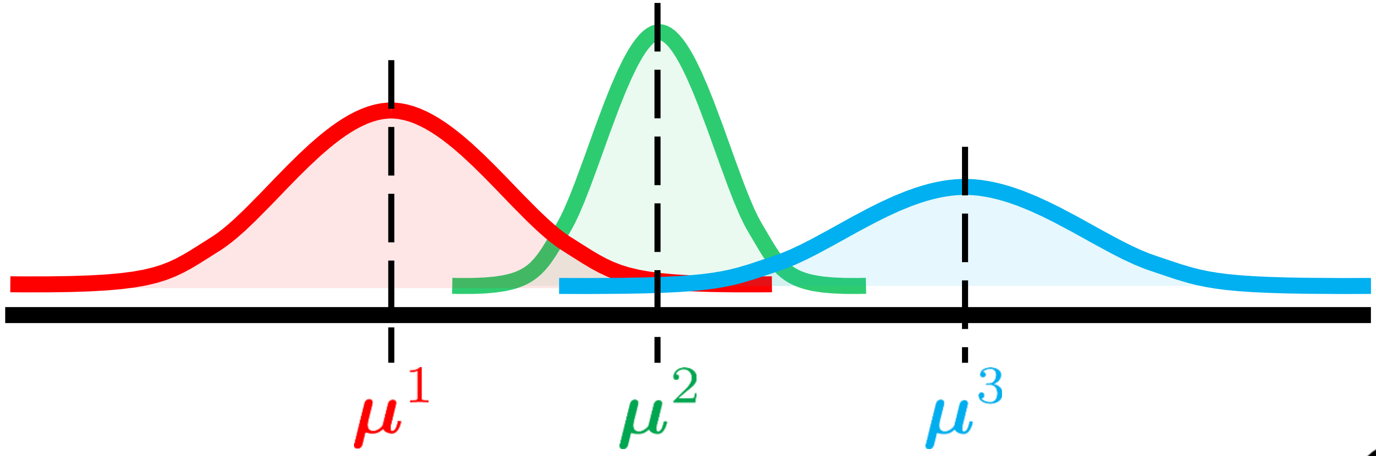


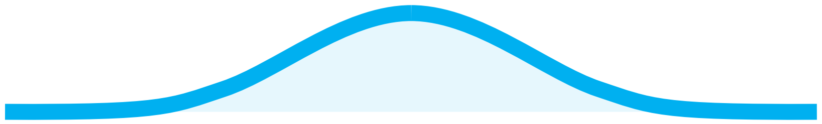
The generative story for labelled data

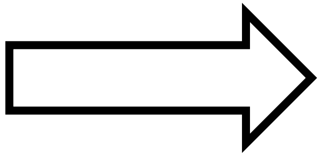


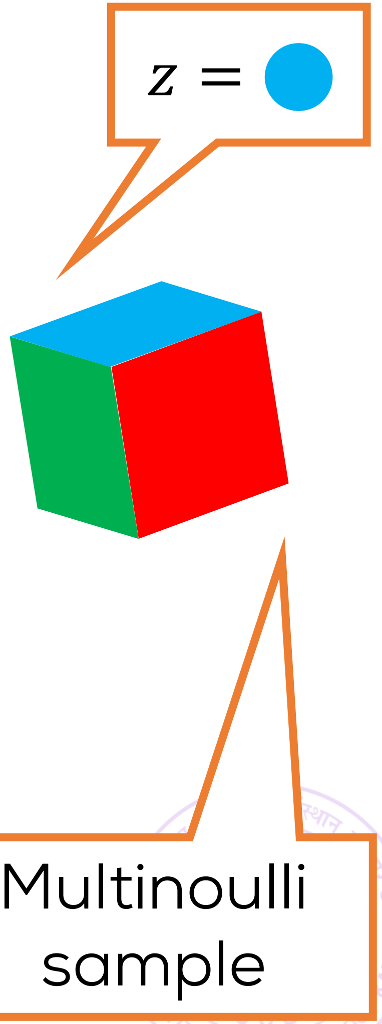
$$\Theta = \{\pi, \{\mu^1, \mu^2, \mu^3\}, \{\Sigma^1, \Sigma^2, \Sigma^3\}\}$$

The generative story for labelled data





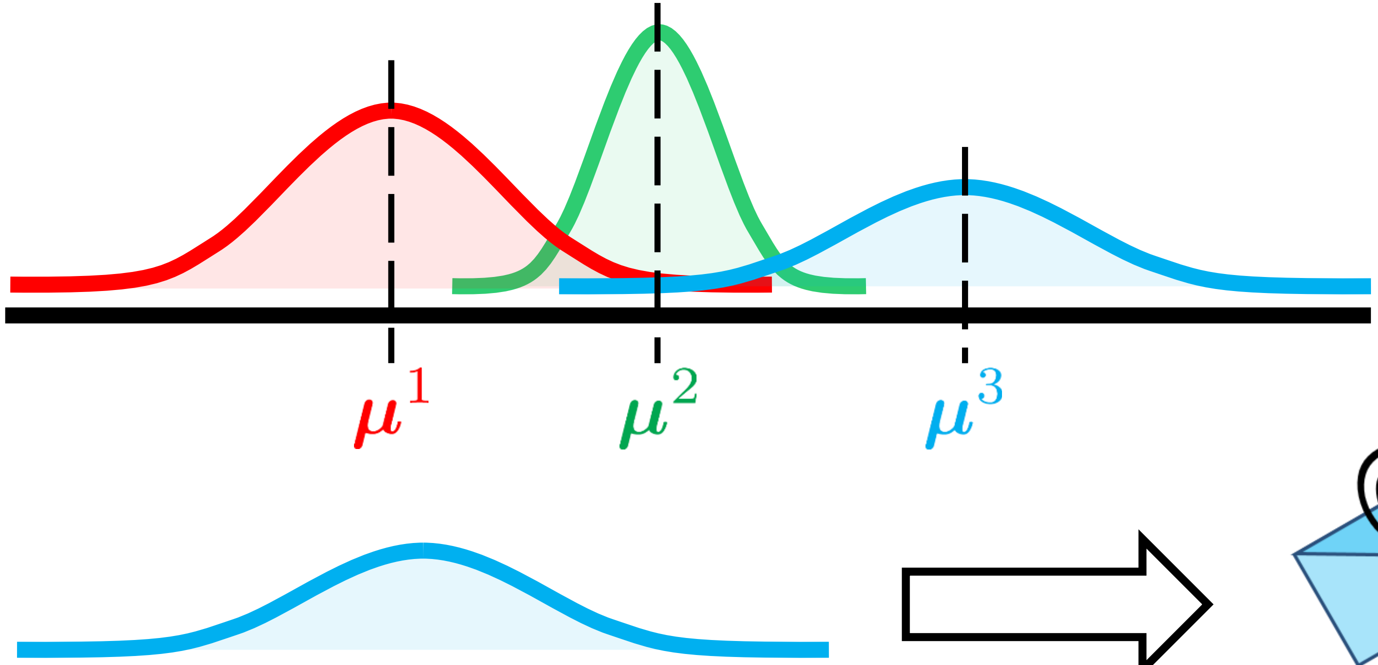


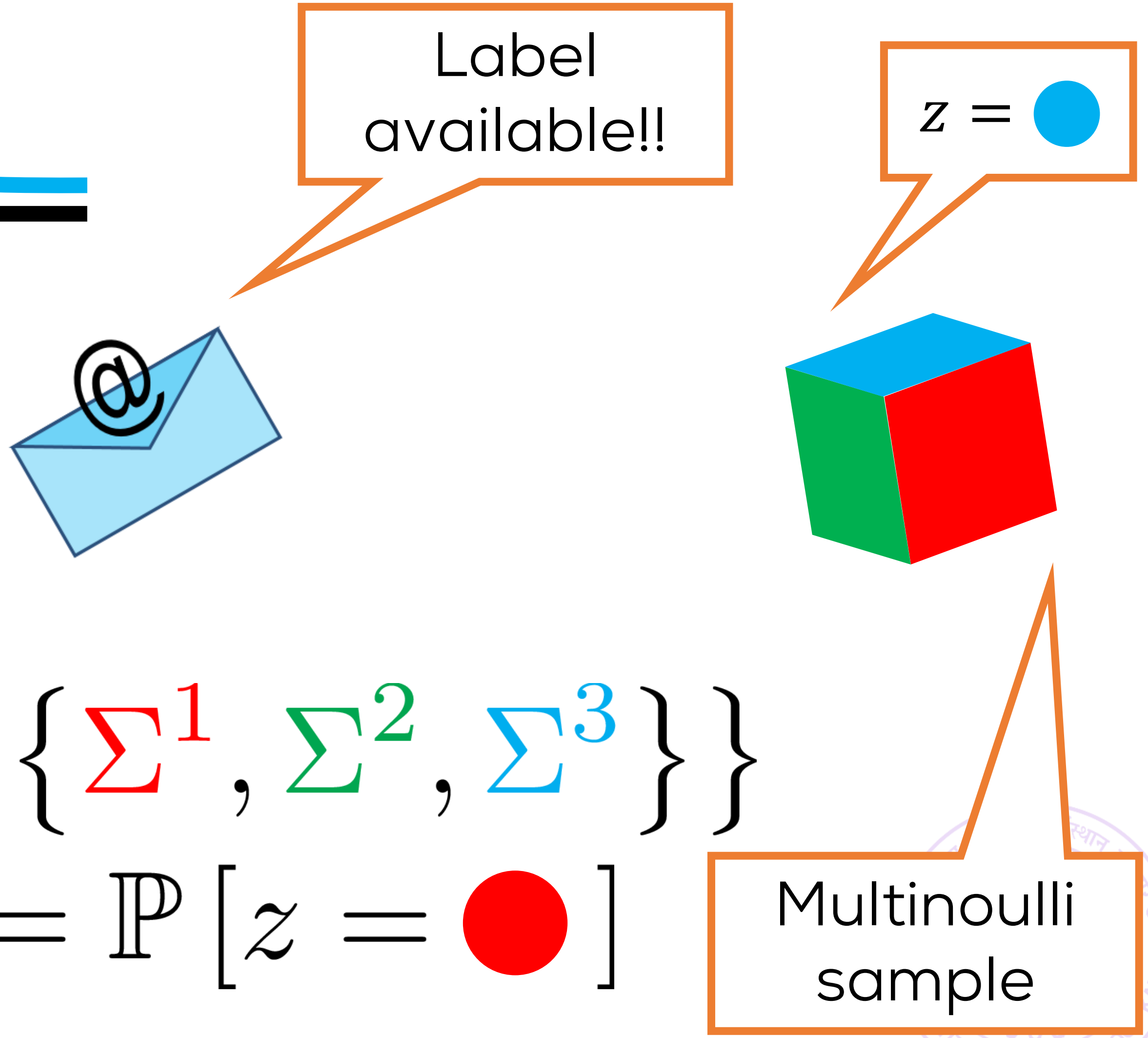


$$\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \{\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \boldsymbol{\mu}^3\}, \{\Sigma^1, \Sigma^2, \Sigma^3\}\}$$

$$\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3\}$$

The generative story for labelled data

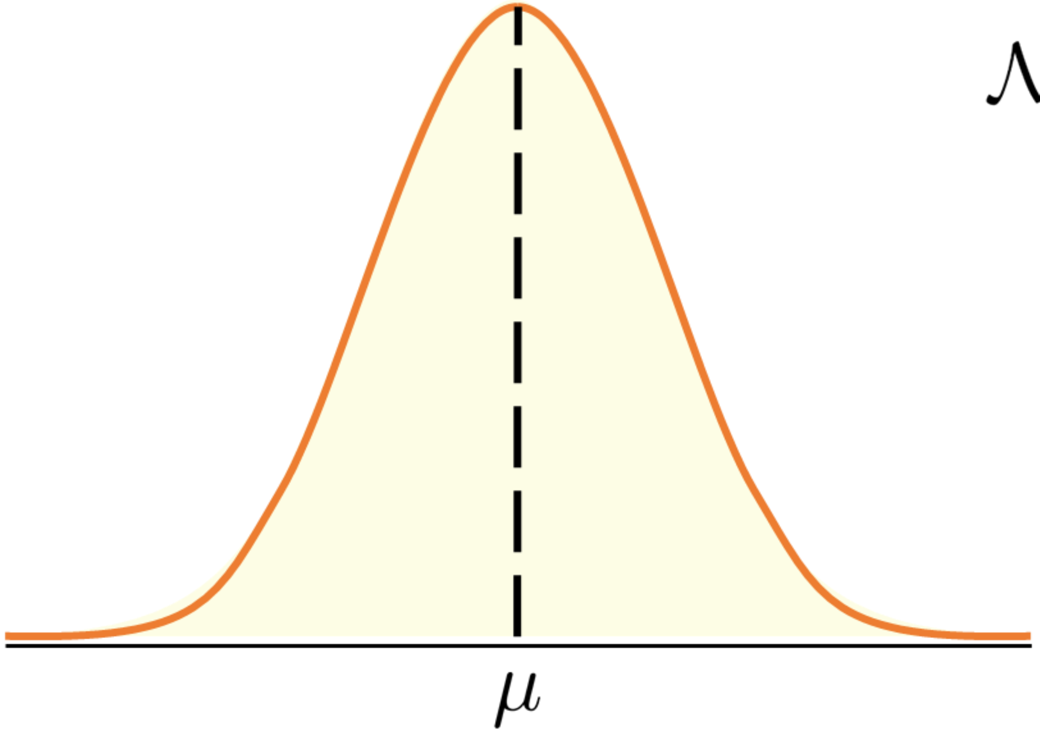




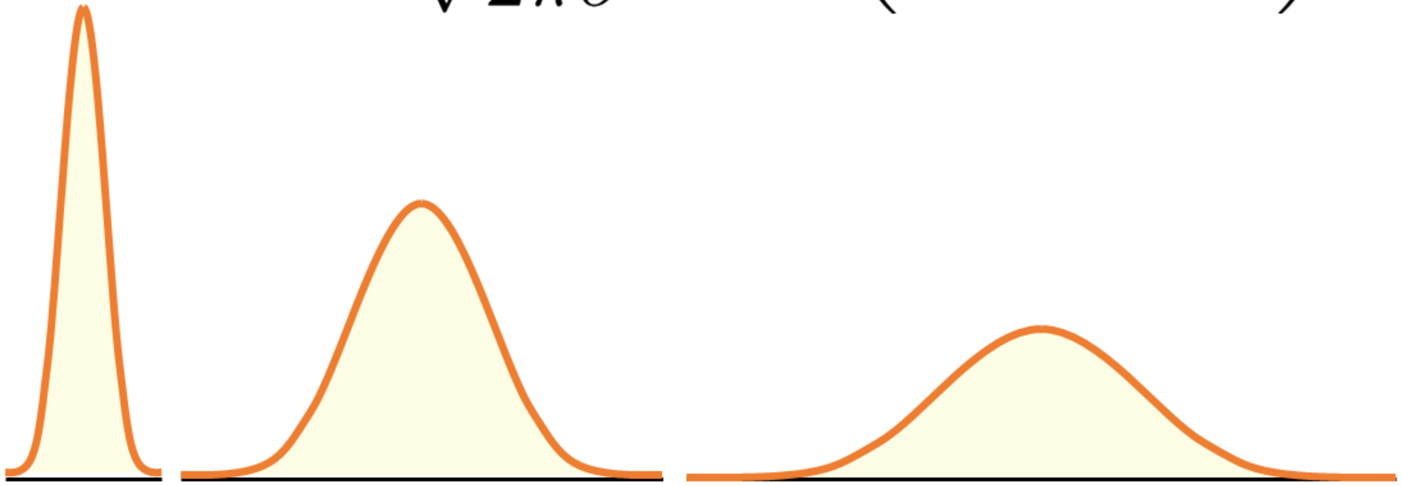
$$\Theta = \{\pi, \{\mu^1, \mu^2, \mu^3\}, \{\Sigma^1, \Sigma^2, \Sigma^3\}\}$$

$$\pi = \{\pi_1, \pi_2, \pi_3\} \quad \pi_1 = \mathbb{P}[z = \bullet]$$

The Gaussian Distribution



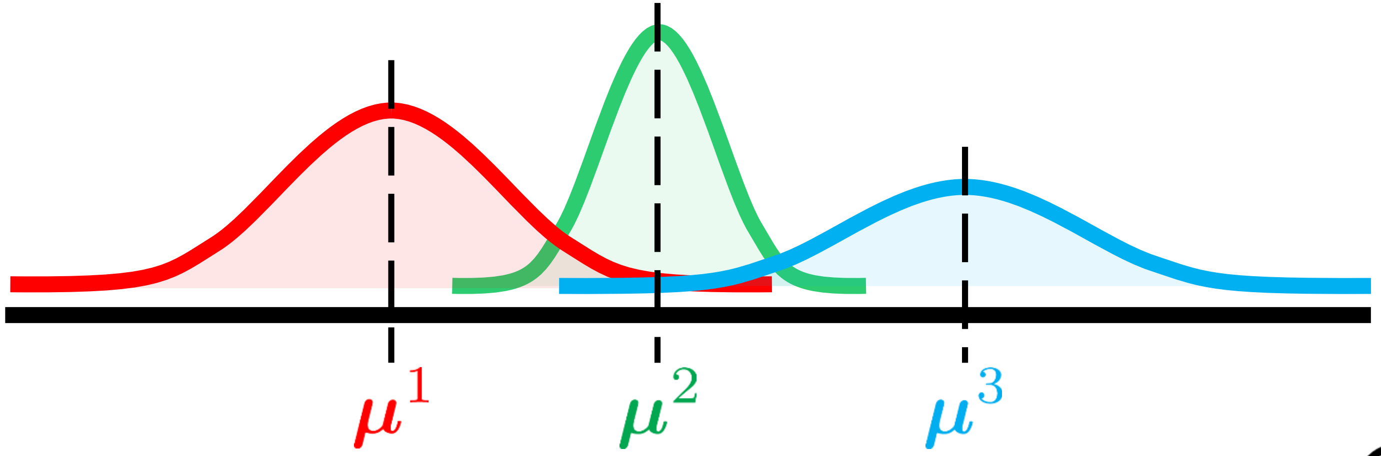
$$\mathcal{N}(t;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

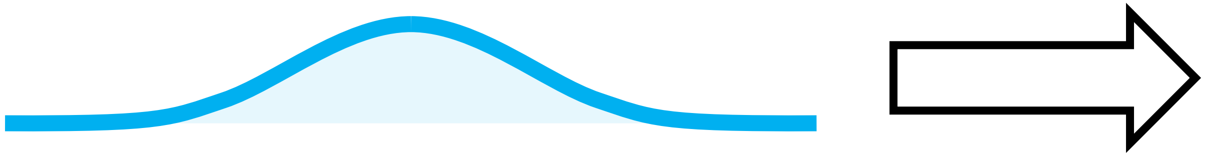


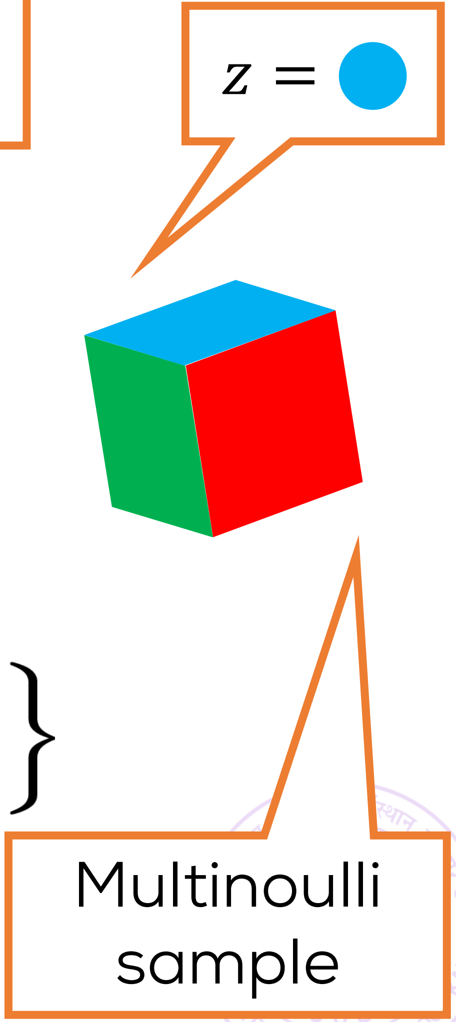
Multivariate Gaussian

$$\mathcal{N}(\mathbf{v};\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{v}-\boldsymbol{\mu})\right)$$

The generative story for labelled data



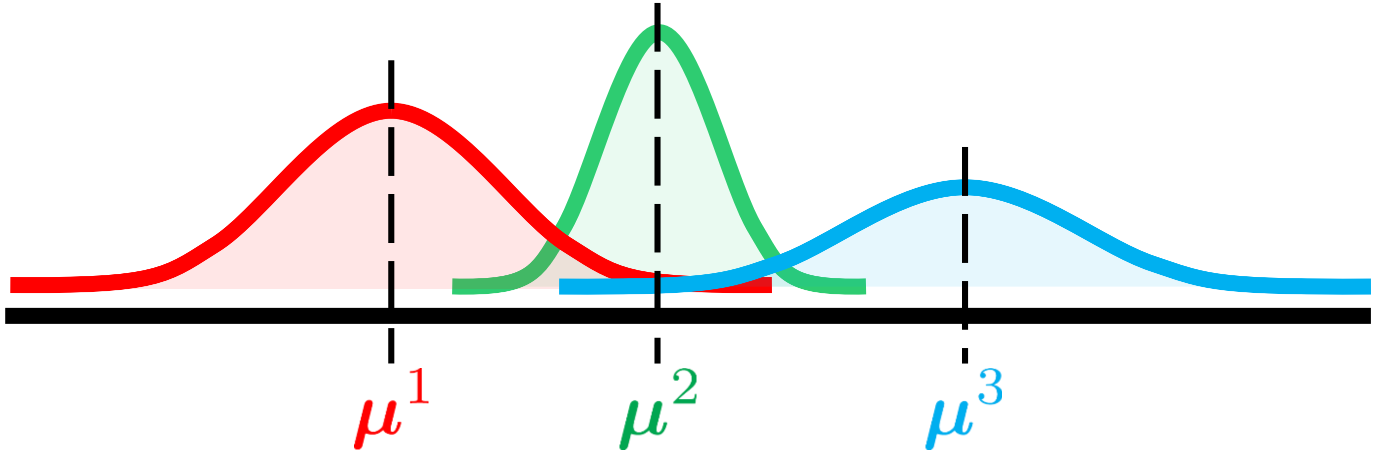




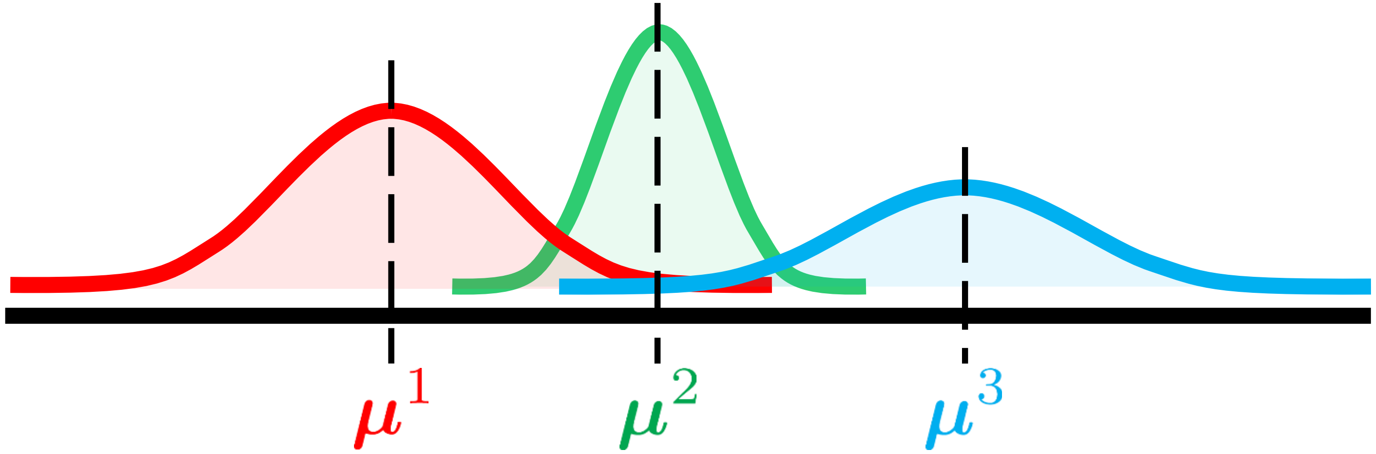
$$\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \{\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \boldsymbol{\mu}^3\}, \{\Sigma^1, \Sigma^2, \Sigma^3\}\}$$

$$\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3\} \quad \boldsymbol{\pi}_1 = \mathbb{P}[z = \bullet]$$

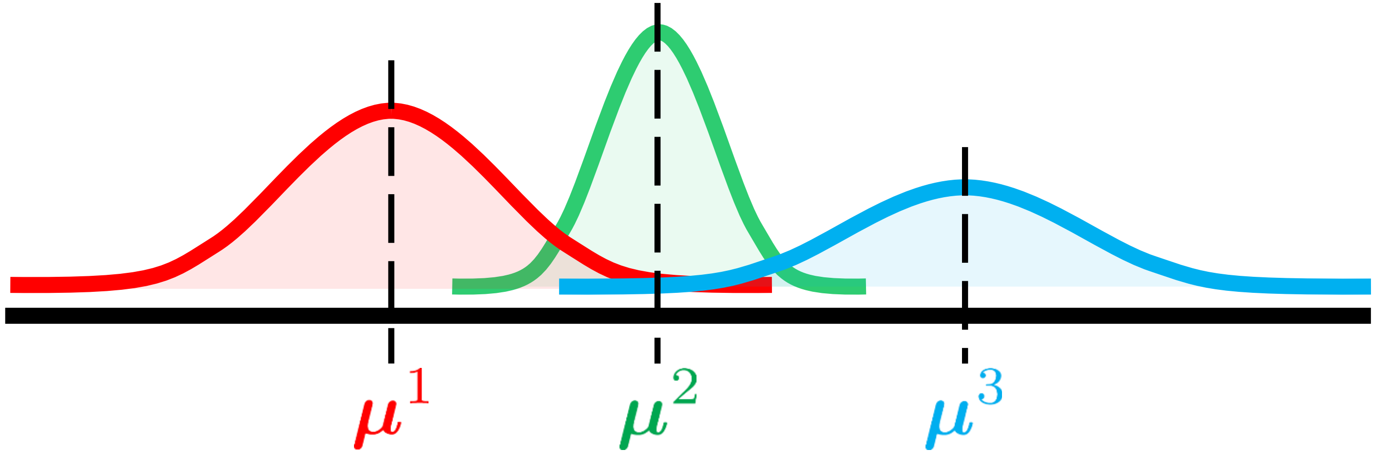
The generative story for labelled data



The generative story for labelled data

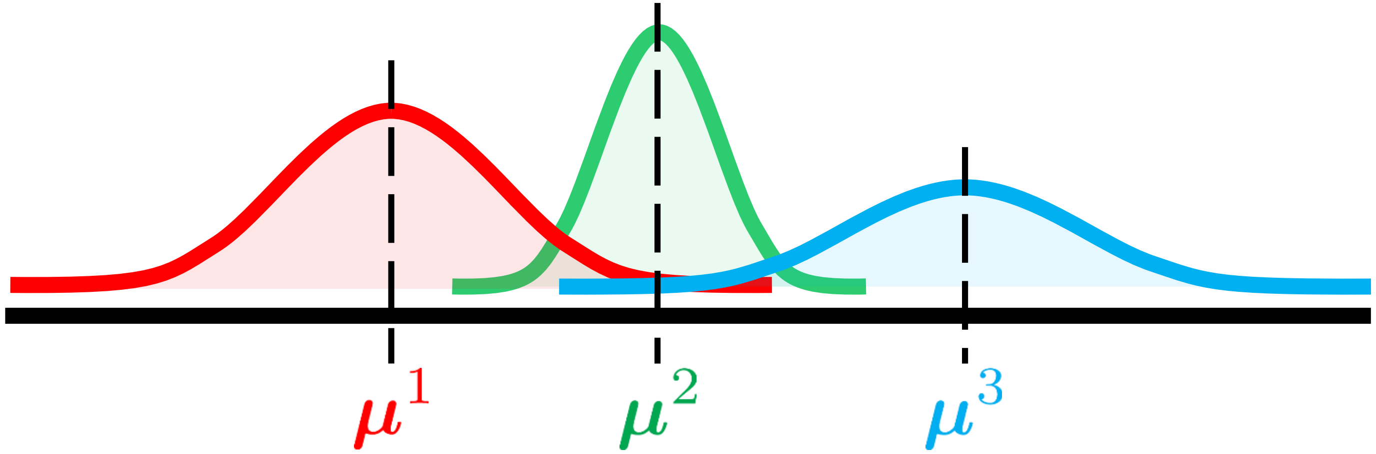


The generative story for labelled data



$$\mathbb{P}\left[\mathbf{x}^i, z^i \mid \mathbf{\Theta}\right] = \mathbb{P}\left[z^i \mid \mathbf{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i, \mathbf{\Theta}\right] = \boldsymbol{\pi}_{z^i} \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^{z^i}, \Sigma^{z^i})$$

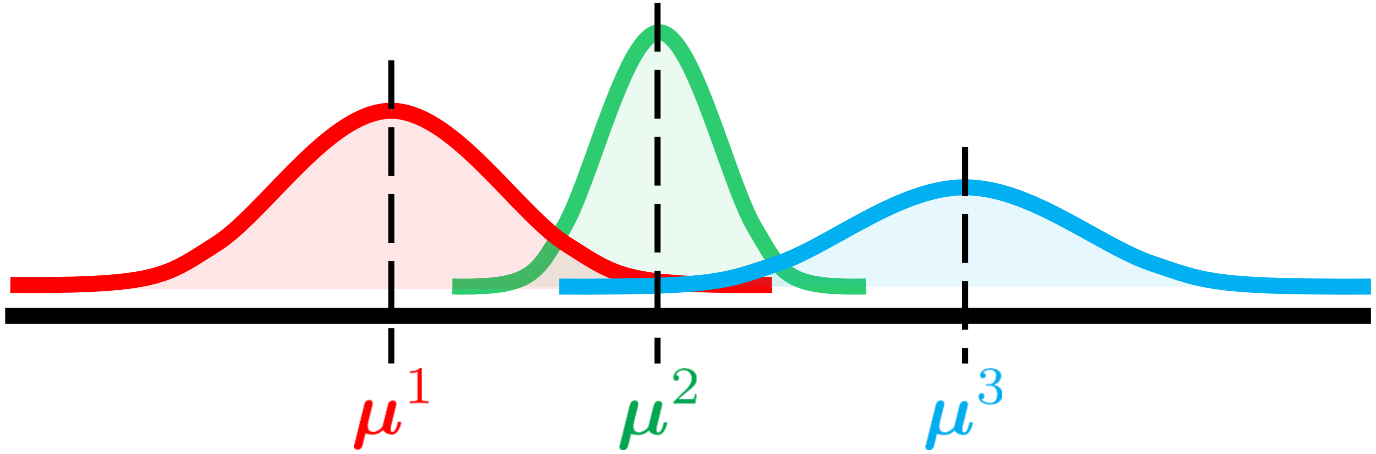
The generative story for labelled data



$$\mathbb{P}\left[\mathbf{x}^i, z^i \,|\, \boldsymbol{\Theta}\right] = \mathbb{P}\left[z^i \,|\, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \,|\, z^i, \boldsymbol{\Theta}\right] = \boldsymbol{\pi}_{z^i} \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^{z^i}, \Sigma^{z^i})$$

$$\mathbb{P}\left[X, \{z^i\} \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i, z^i \,|\, \boldsymbol{\Theta}\right]$$

The generative story for labelled data

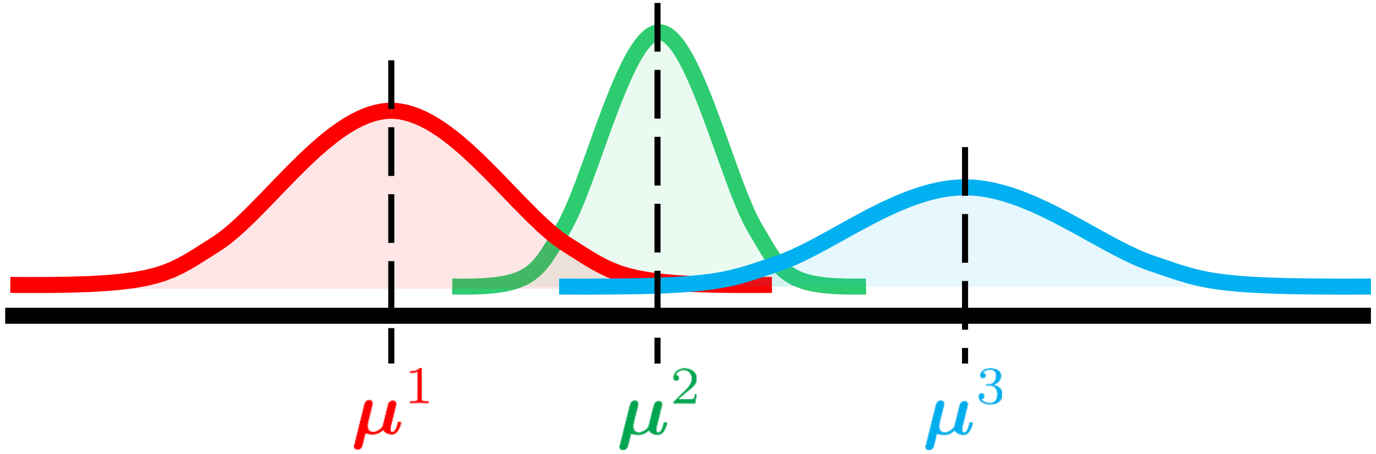


$$\mathbb{P}\left[\mathbf{x}^i, z^i \,|\, \boldsymbol{\Theta}\right] = \mathbb{P}\left[z^i \,|\, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \,|\, z^i, \boldsymbol{\Theta}\right] = \boldsymbol{\pi}_{z^i} \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^{z^i}, \Sigma^{z^i})$$

$$\mathbb{P}\left[X, \{z^i\} \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i, z^i \,|\, \boldsymbol{\Theta}\right]$$

$$\hat{\boldsymbol{\Theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \{z^i\} \,|\, \boldsymbol{\Theta}\right]$$

The generative story for labelled data



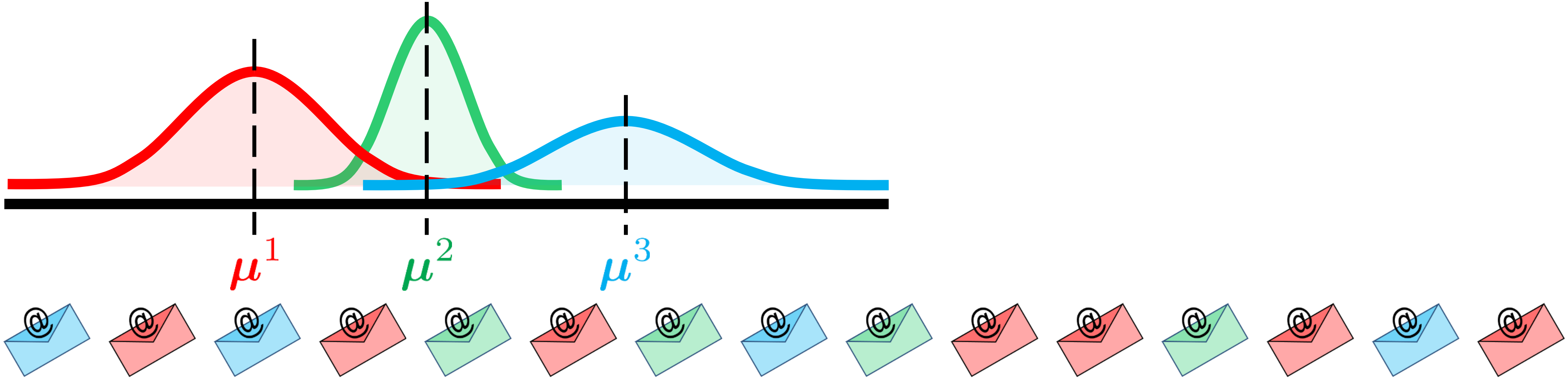
$$\mathbb{P}\left[\mathbf{x}^i, z^i \mid \mathbf{\Theta}\right] = \mathbb{P}\left[z^i \mid \mathbf{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i, \mathbf{\Theta}\right] = \boldsymbol{\pi}_{z^i} \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^{z^i}, \Sigma^{z^i})$$

$$\mathbb{P}\left[X, \{z^i\} \mid \mathbf{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i, z^i \mid \mathbf{\Theta}\right]$$

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^i\} \mid \mathbf{\Theta}\right]$$

Take log and apply 1st order optimality

The generative story for labelled data



$$\mathbb{P}\left[\mathbf{x}^i, z^i \mid \boldsymbol{\Theta}\right] = \mathbb{P}\left[z^i \mid \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i, \boldsymbol{\Theta}\right] = \boldsymbol{\pi}_{z^i} \cdot \mathcal{N}\left(\mathbf{x}^i \mid \boldsymbol{\mu}^{z^i}, \Sigma^{z^i}\right)$$

$$\mathbb{P}\left[X, \{z^i\} \mid \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i, z^i \mid \boldsymbol{\Theta}\right]$$

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \{z^i\} \mid \boldsymbol{\Theta}\right]$$

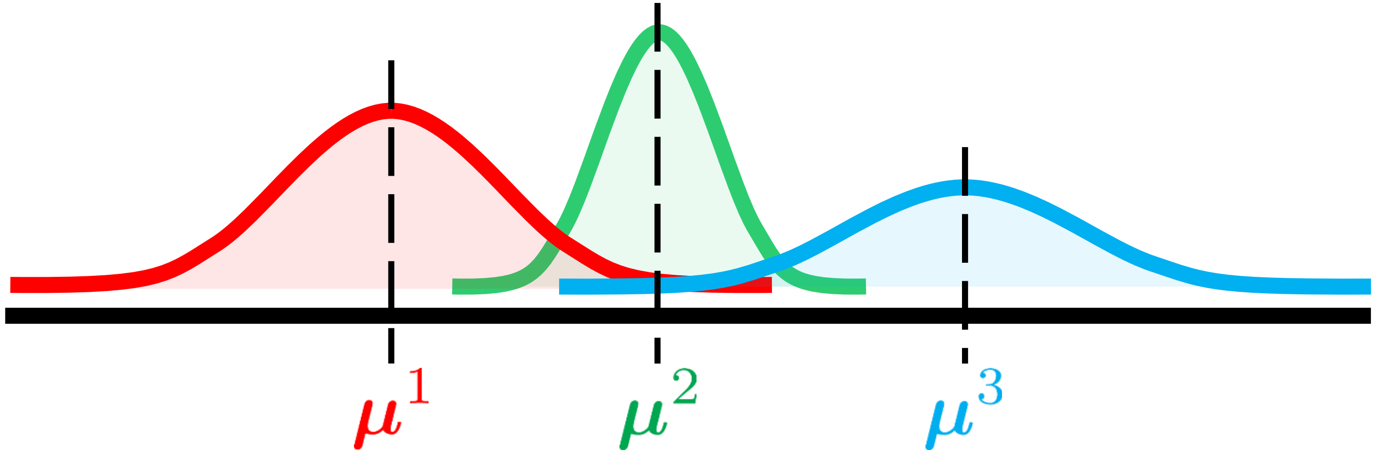
Take log and apply 1st order optimality

$$\boldsymbol{\pi}_1^{\text{MLE}} = \frac{\text{\# red emails}}{\text{\# total emails}} = \frac{n_r}{n}$$

$$\boldsymbol{\mu}_{\text{MLE}}^1 = \frac{1}{n_r} \sum_{i:z^i=\bullet} \mathbf{x}^i$$

$$\Sigma_{\text{MLE}}^1 = \frac{1}{n_r} \sum_{i:z^i=\bullet} \left(\mathbf{x}^i - \boldsymbol{\mu}_{MLE}^1\right)\left(\mathbf{x}^i - \boldsymbol{\mu}_{MLE}^1\right)^\top$$

The generative story for labelled data



$$\mathbb{P}\left[\mathbf{x}^i, z^i \mid \boldsymbol{\Theta}\right] = \mathbb{P}\left[z^i \mid \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i, \boldsymbol{\Theta}\right] = \boldsymbol{\pi}_{z^i} \cdot \mathcal{N}\left(\mathbf{x}^i \mid \boldsymbol{\mu}^{z^i}, \Sigma^{z^i}\right)$$

$$\mathbb{P}\left[X, \{z^i\} \mid \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i, z^i \mid \boldsymbol{\Theta}\right]$$

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \{z^i\} \mid \boldsymbol{\Theta}\right]$$

$$\boldsymbol{\pi}_1^{\text{MLE}} = \frac{\text{\# red emails}}{\text{\# total emails}} = \frac{n_r}{n}$$

$$\boldsymbol{\mu}^1_{\text{MLE}} = \frac{1}{n_r} \sum_{i : z^i = \bullet} \mathbf{x}^i$$

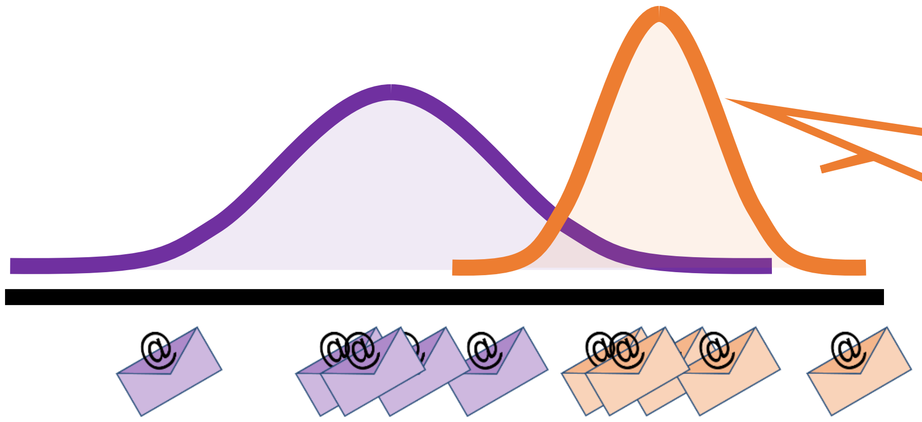
$$\Sigma^1_{\text{MLE}} = \frac{1}{n_r} \sum_{i : z^i = \bullet} \left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)\left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)^\top$$

Take log and apply 1st order optimality

Read [DAU] Sections 9.1-9.5

⟨/detour⟩

The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

We do not know which email is purple and which is orange – no labels!!

Better take a look now ☺

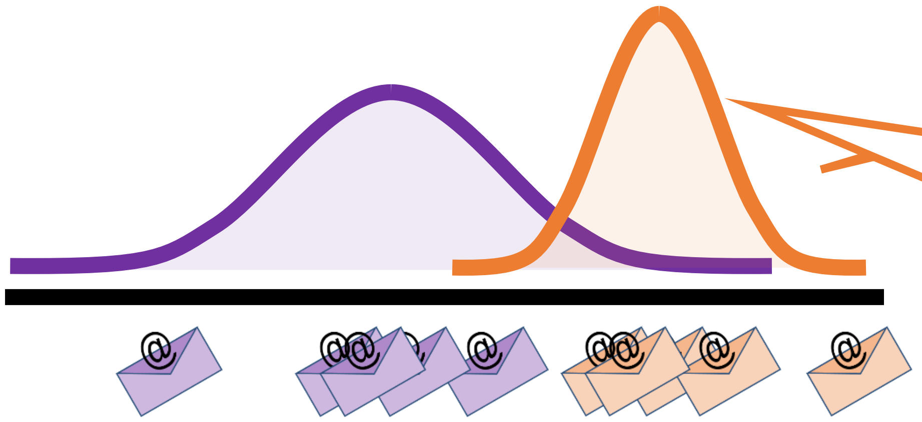
Saw how to do this for Bernoulli distributions last time.

Hmm … what if someone magically labelled the emails with the color

Can estimate the class proportions, and means and variances of these 1D Gaussians using training data!

Read **[DAU]** Sections 9.1-9.5

The generative story for unlabelled data??



No labels!!!

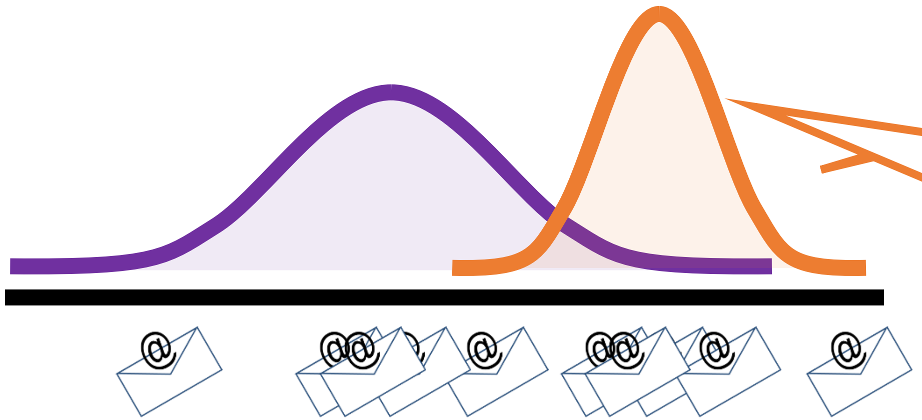
Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

The generative story for unlabelled data??



No labels!!!

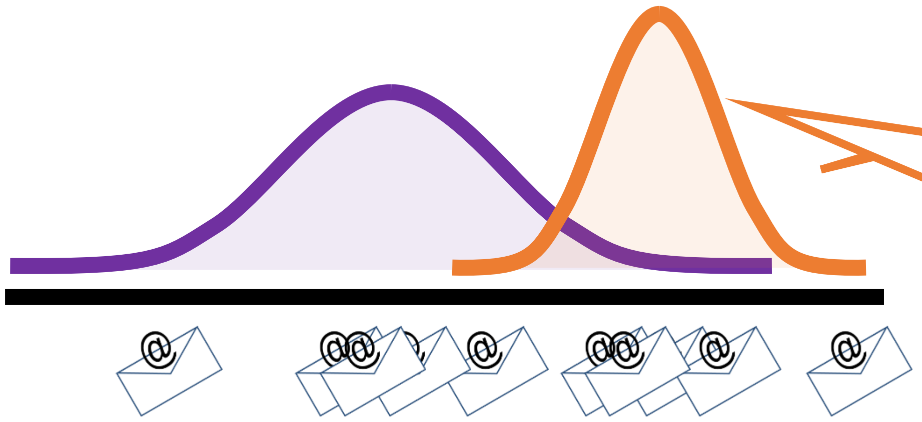
Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

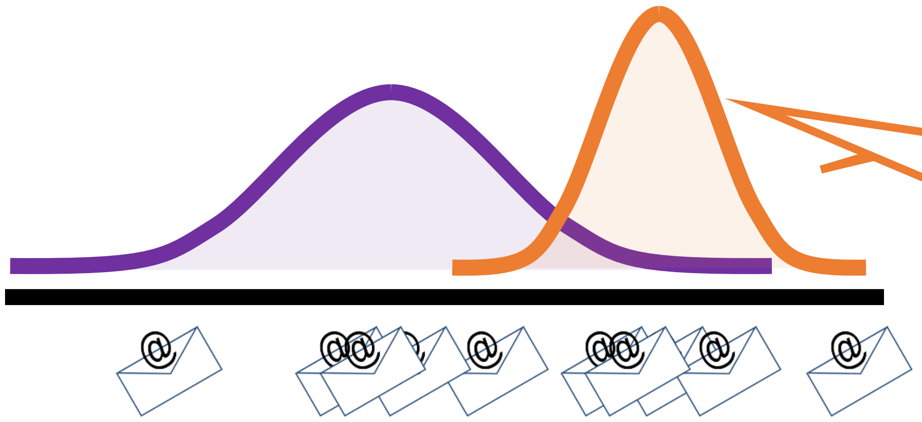
Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

How to get these magical labels??

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

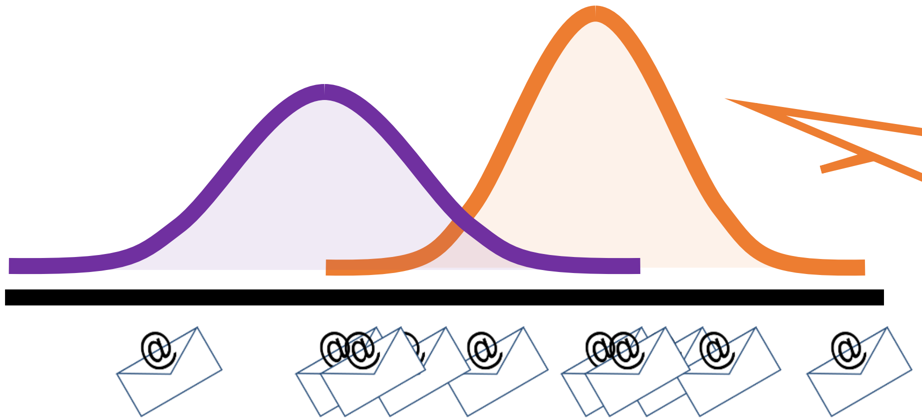
We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

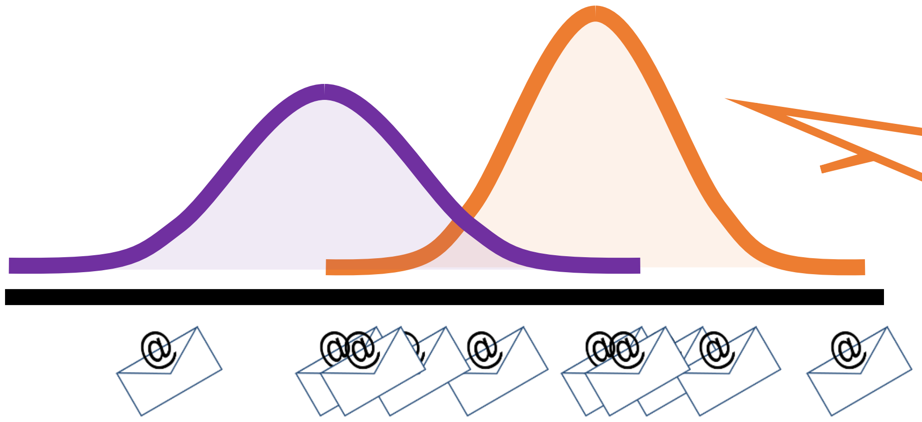
We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

How to get these magical labels??

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

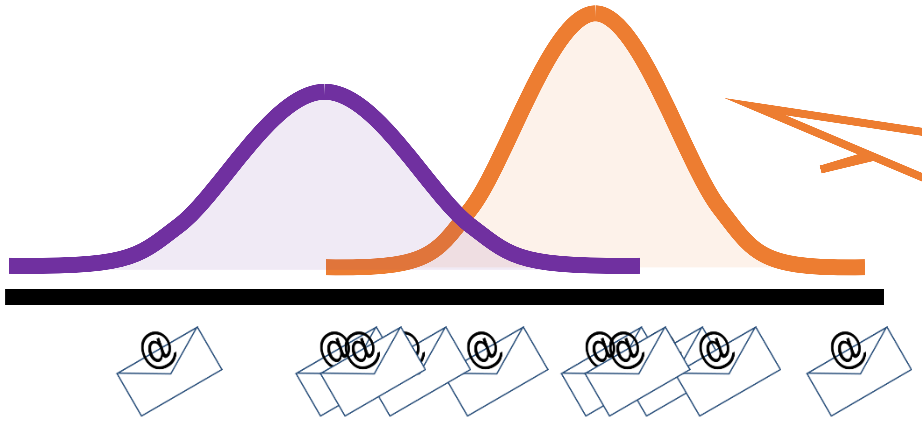
Hmm … what if someone magically labelled the emails with the color

Can use these to label emails!

How to get these magical labels??

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

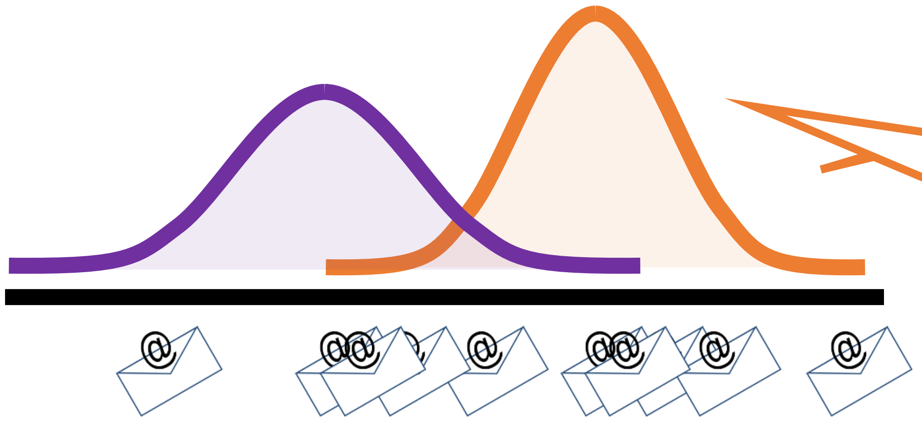
Can use these to label emails!

Hmm … what if someone magically labelled the emails with the color

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



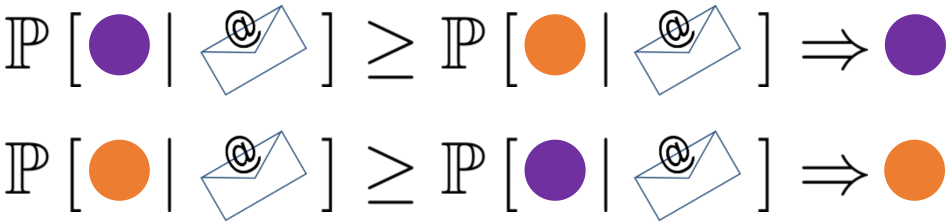
No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

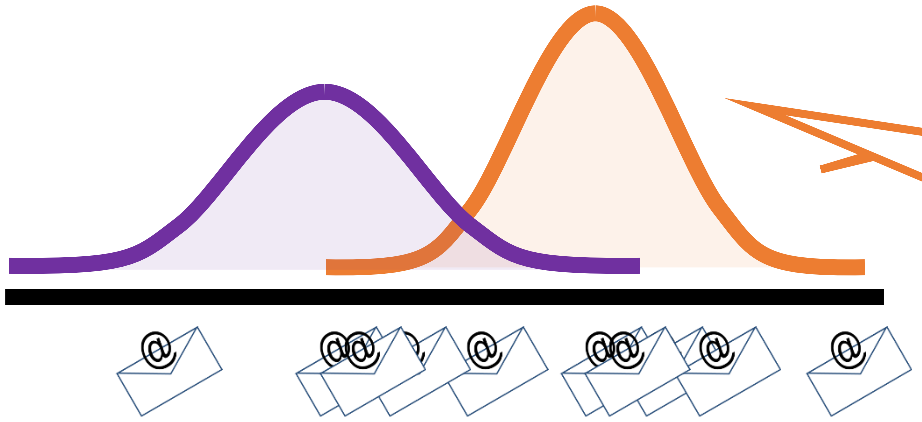


Can use these to label emails!

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



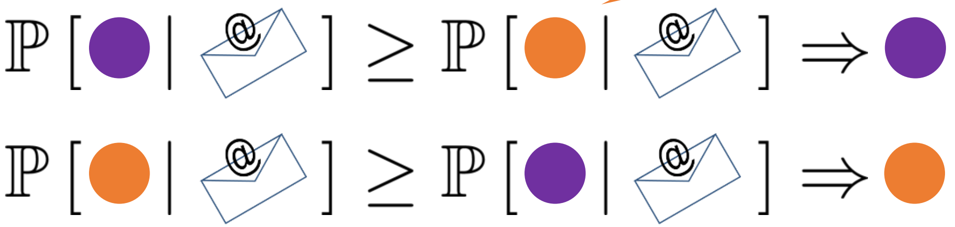
No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

We do not know which email is purple and which is orange – no labels!!

The dist. are wrong but trust them for now



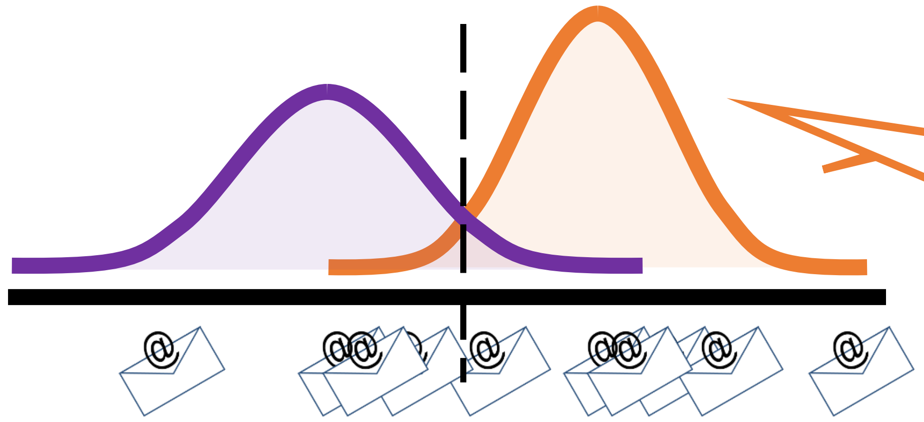
Can use these to label emails!

Hmm … what if someone magically labelled the emails with the color

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



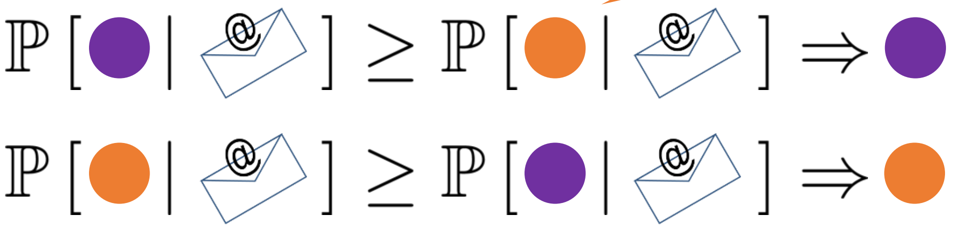
No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

The dist. are wrong but trust them for now



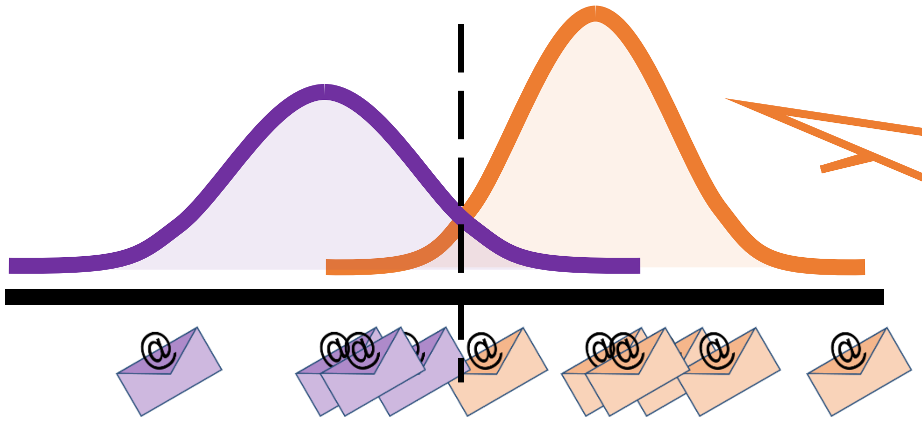
Can use these to label emails!

Hmm … what if someone magically labelled the emails with the color

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



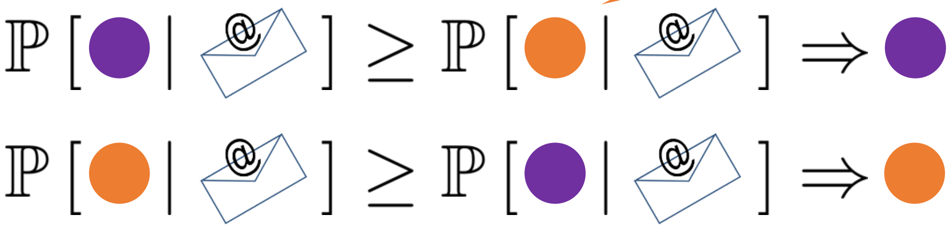
No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

We do not know which email is purple and which is orange – no labels!!

The dist. are wrong but trust them for now



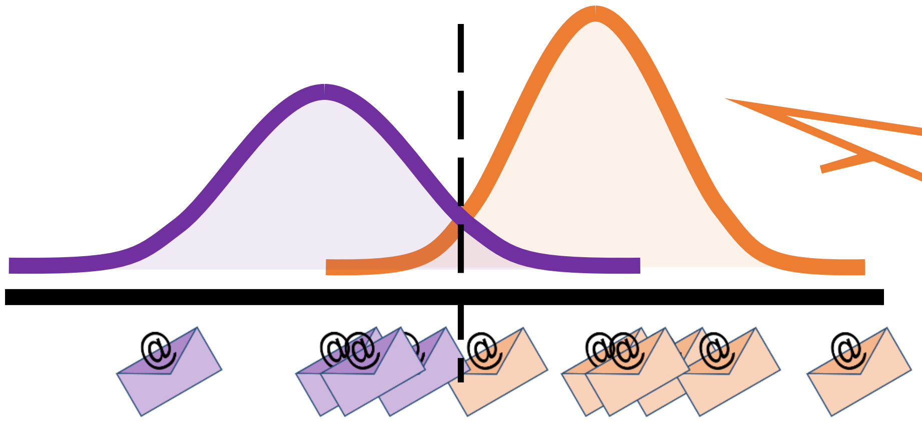
Hmm … what if someone magically labelled the emails with the color

Can use these to label emails!

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

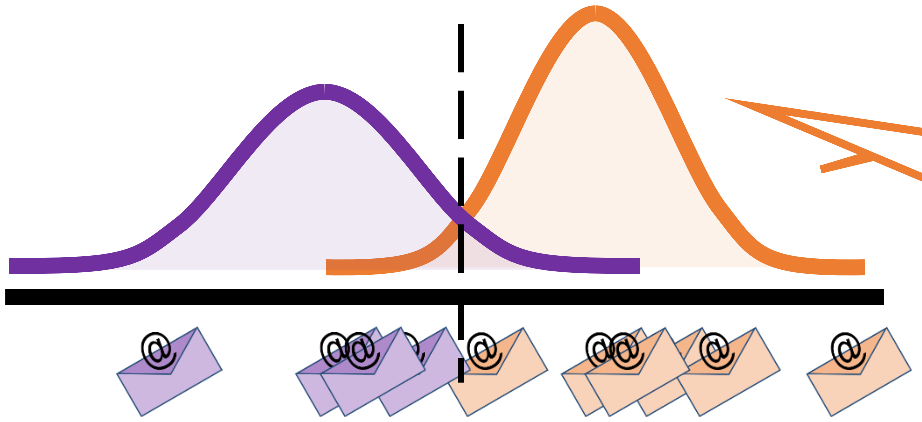
Hmm … what if someone magically labelled the emails with the color

Can use these to label emails!

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

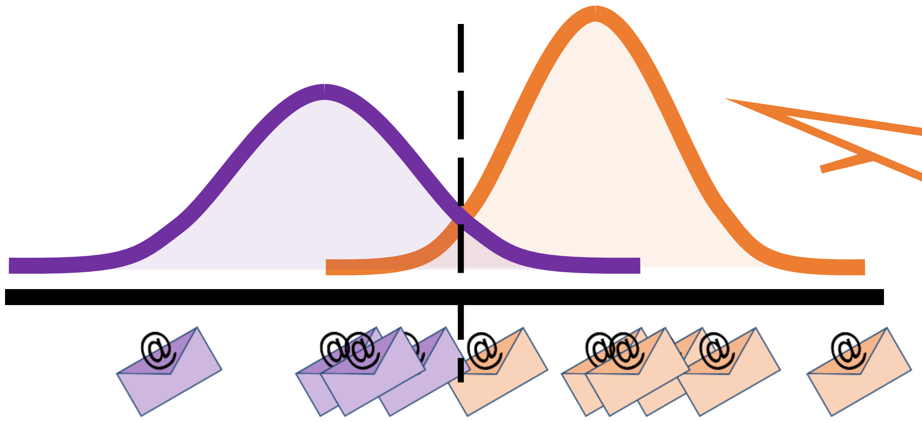
How to get these magical labels??

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

Can use these to label emails!

What if I use these “magical” labels to learn two Gaussians?

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Won't I just learn the wrong ones that generated the labels?

What if I use these "magical" labels to learn two Gaussians?

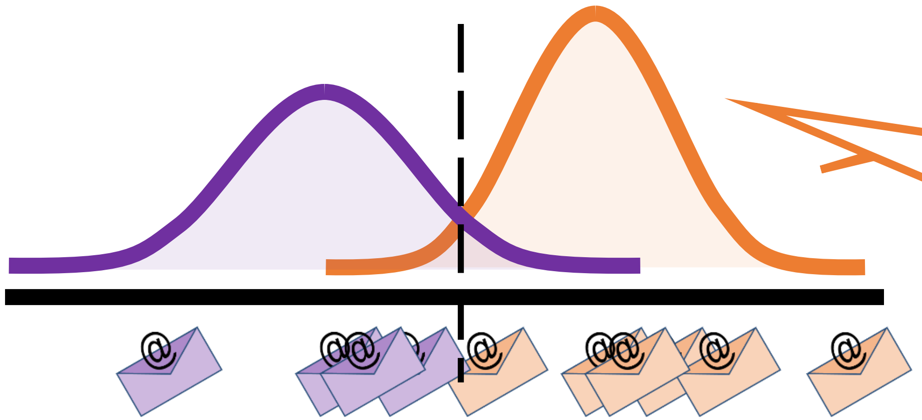
Can use these to label emails!

Hmm … what if someone magically labelled the emails with the color

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

How to get these magical labels??

The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

How to get these magical labels??

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

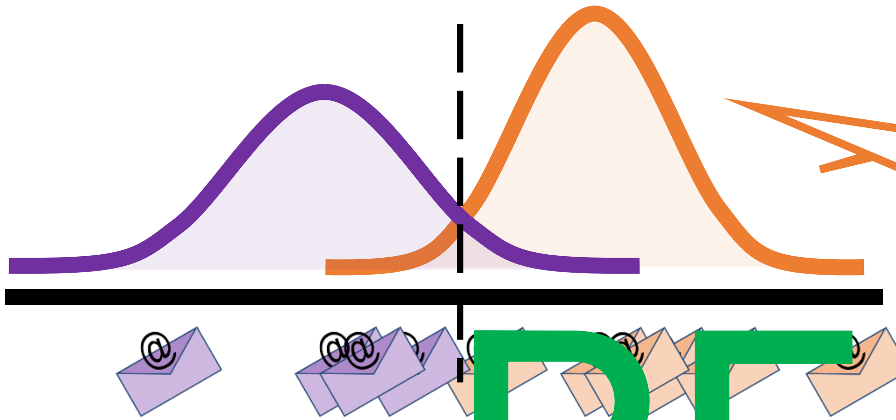
Can use these to label emails!

What if I use these “magical” labels to learn two Gaussians?

Won’t I just learn the wrong ones that generated the labels?

In practice: you will learn slightly “less wrong” ones ☺

The generative story for unlabelled data??



No labels!!!

Can we still recover the two Gaussian components in the mixture??

Give the Gaussians colors to help identifying them

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

How to get these magical labels??

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

Can use these to label emails!

What if I use these “magical” labels to learn two Gaussians?

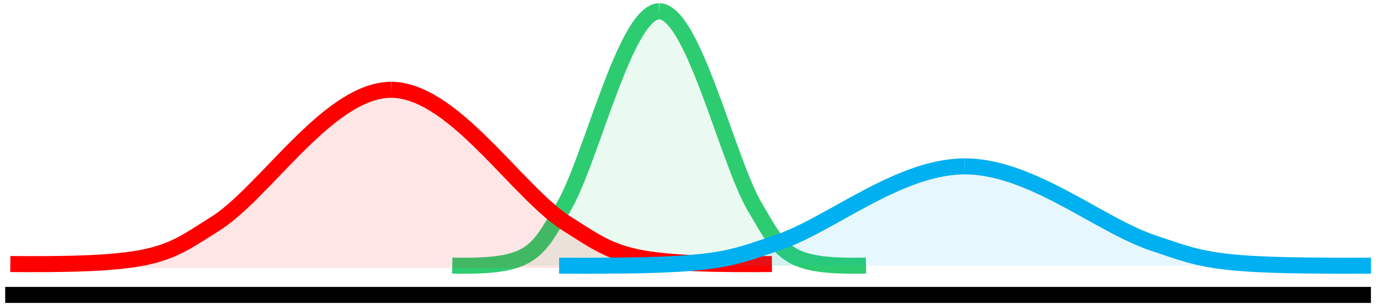
Won't I just learn the wrong ones that generated the labels?

In practice: you will learn slightly “less wrong” ones ☺

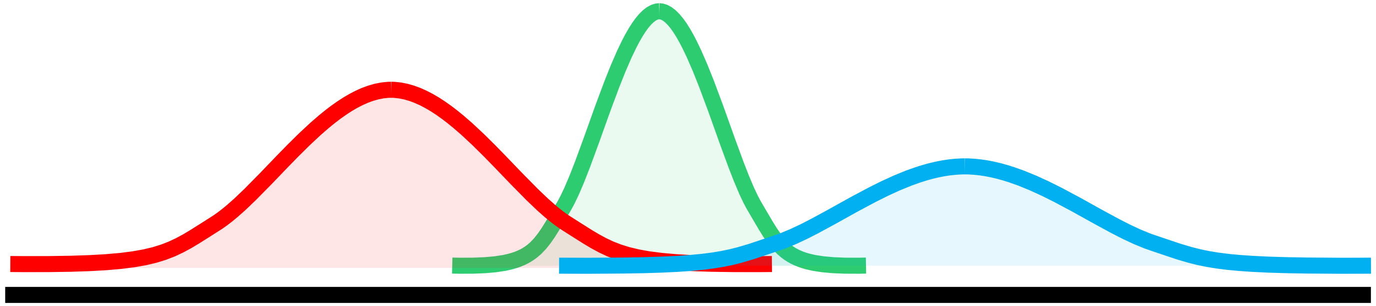
REPEAT!

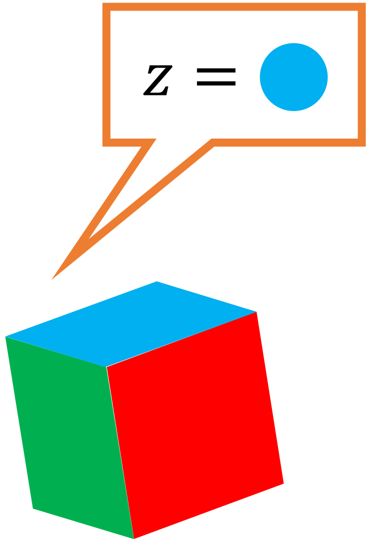
The generative story for unlabelled data

The generative story for unlabelled data

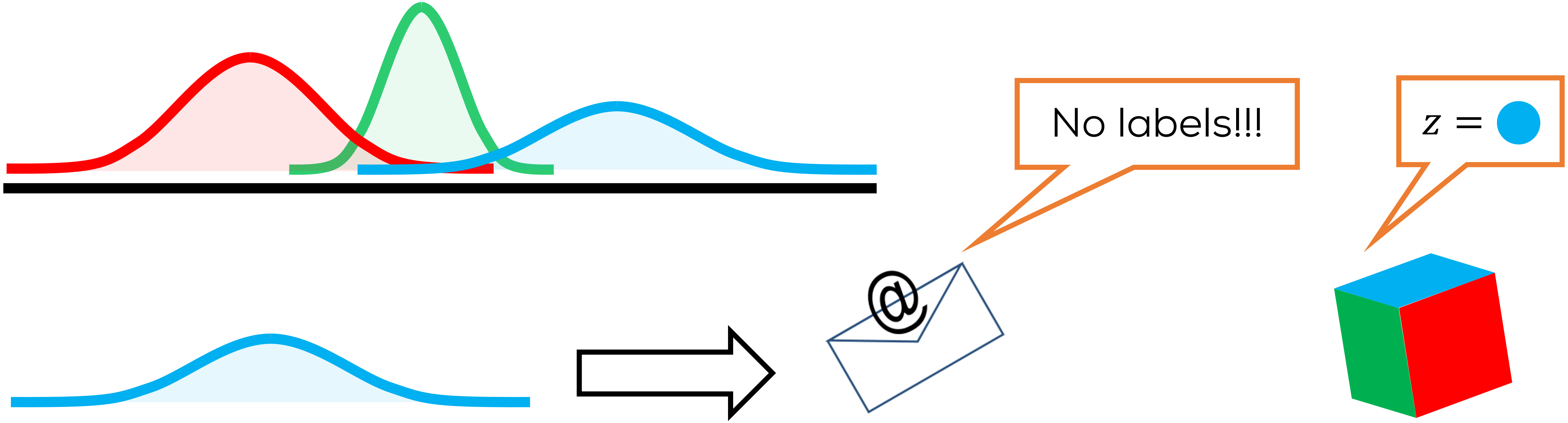


The generative story for unlabelled data

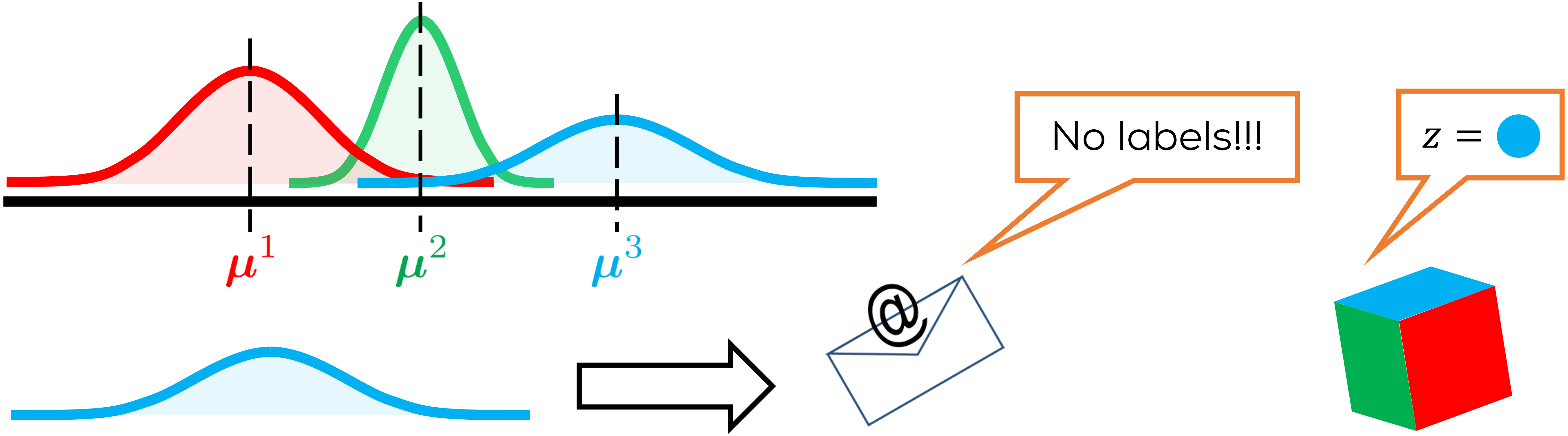




The generative story for unlabelled data

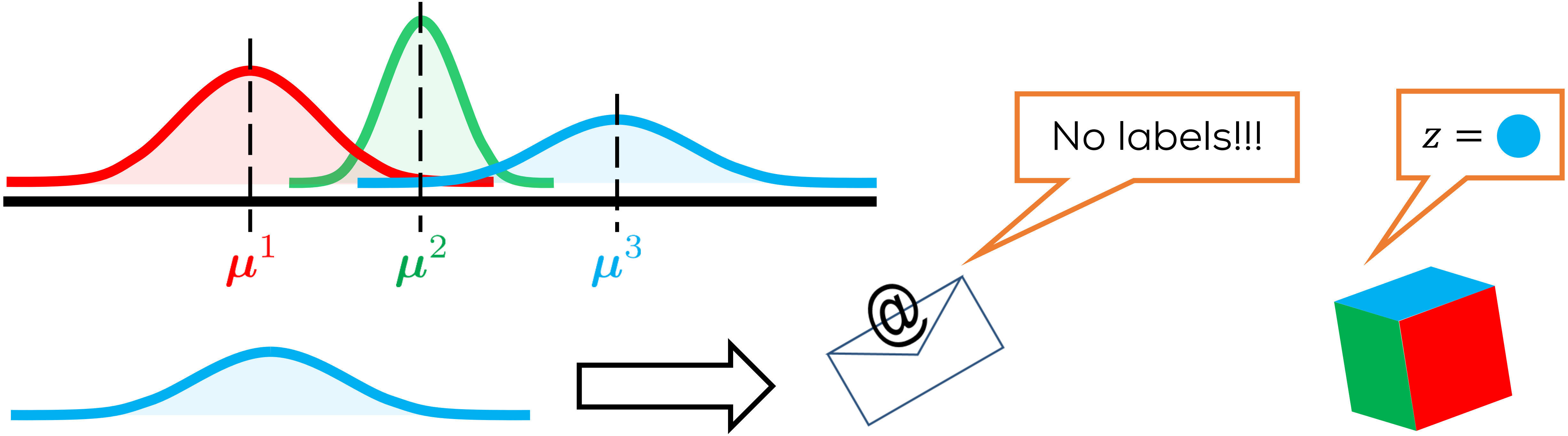


The generative story for unlabelled data



$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\pi}, \left\{ \boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \boldsymbol{\mu}^3 \right\}, \left\{ \Sigma^1, \Sigma^2, \Sigma^3 \right\} \right\}$$

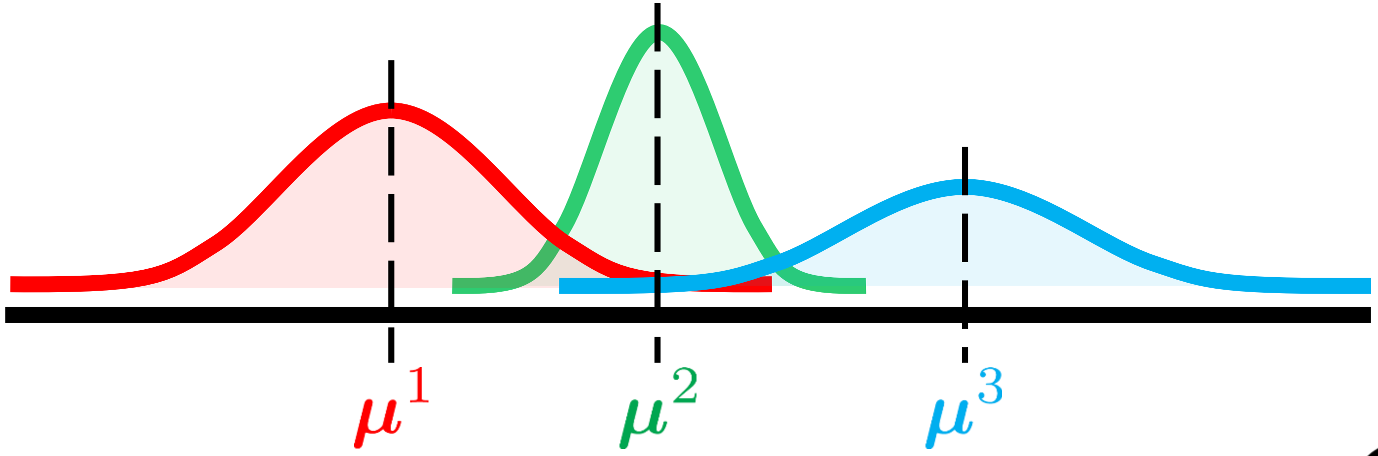
The generative story for unlabelled data

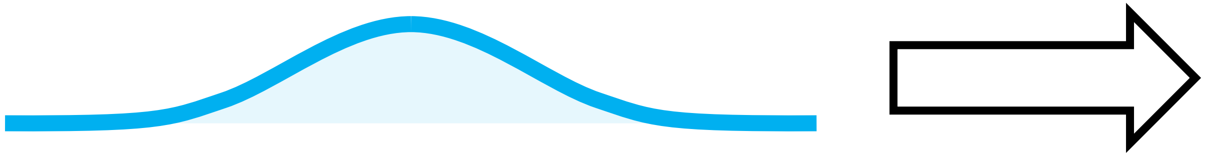


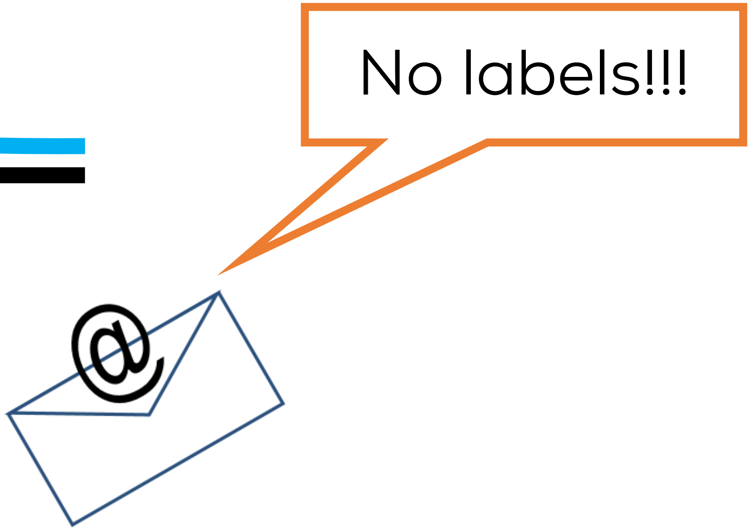
$$\Theta = \{\pi, \{\mu^1, \mu^2, \mu^3\}, \{\Sigma^1, \Sigma^2, \Sigma^3\}\}$$

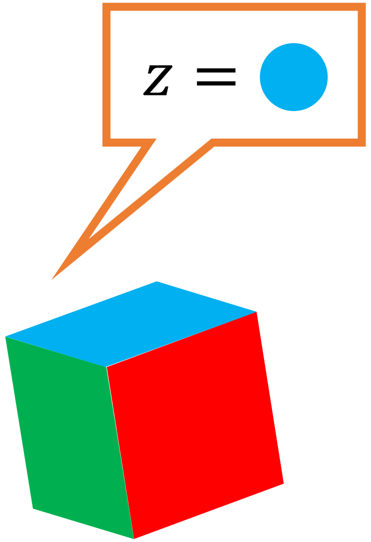
$$\pi = \{\pi_1, \pi_2, \pi_3\}$$

The generative story for unlabelled data







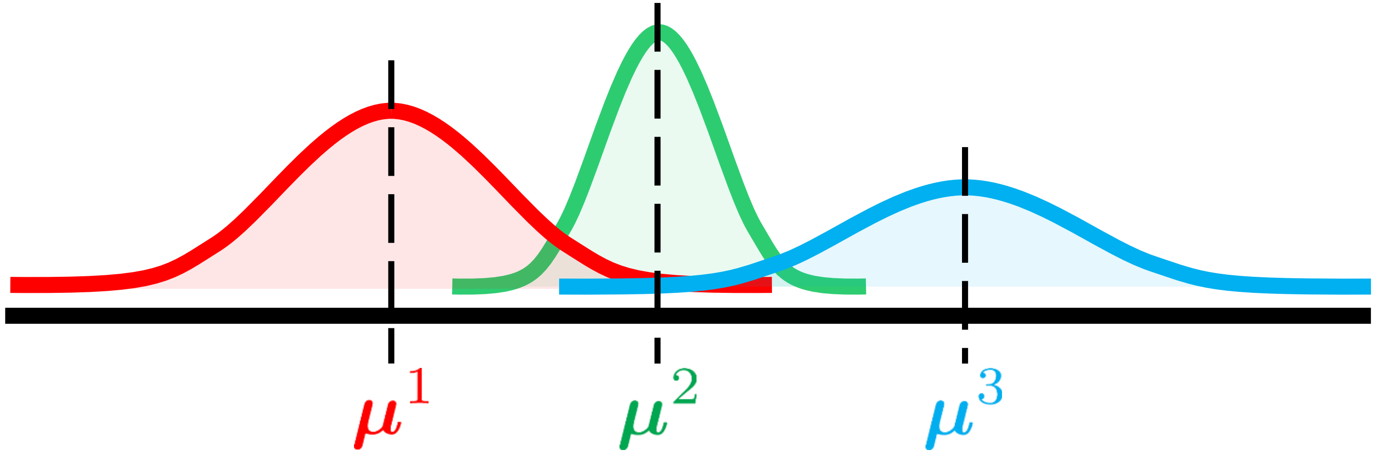


$$\Theta = \{\pi, \{\mu^1, \mu^2, \mu^3\}, \{\Sigma^1, \Sigma^2, \Sigma^3\}\}$$

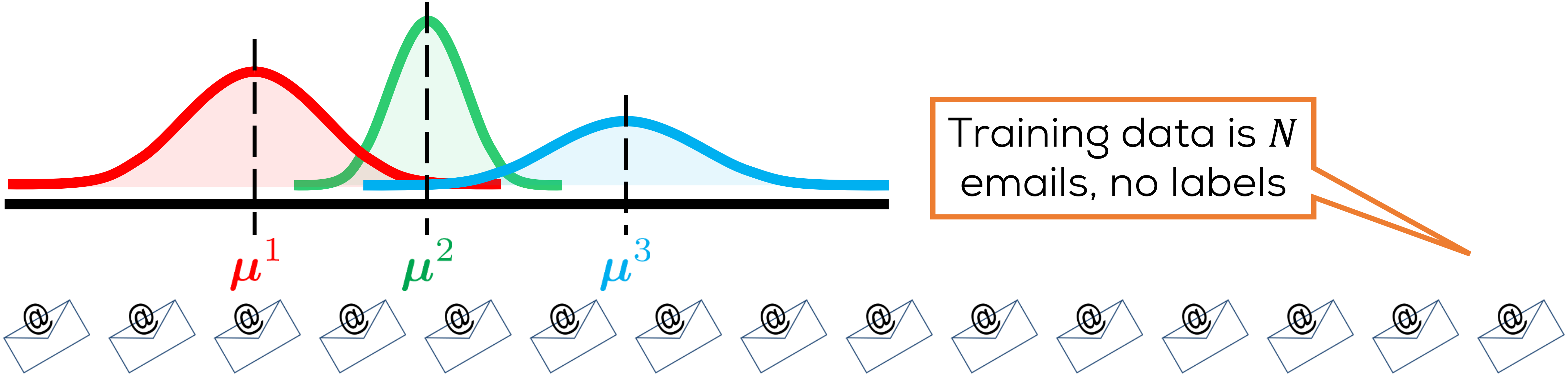
$$\pi = \{\pi_1, \pi_2, \pi_3\} \quad \pi_1 = \mathbb{P}[z = \bullet]$$

Just as before!

The generative story for unlabelled data

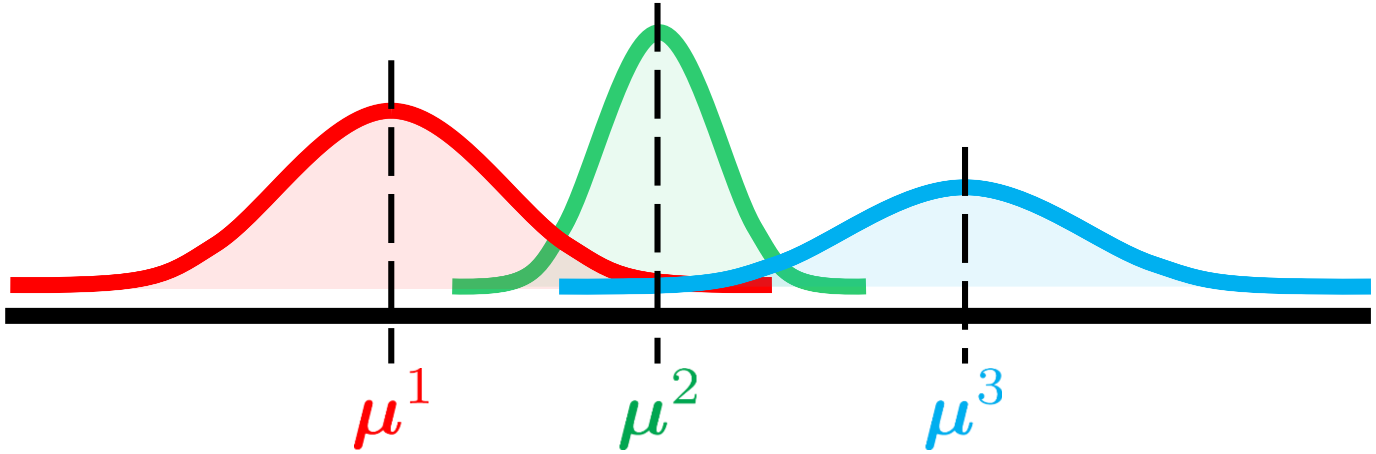


The generative story for unlabelled data

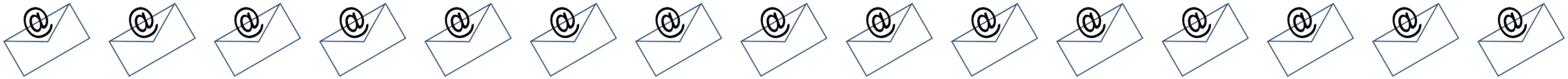


Training data is N emails, no labels

The generative story for unlabelled data

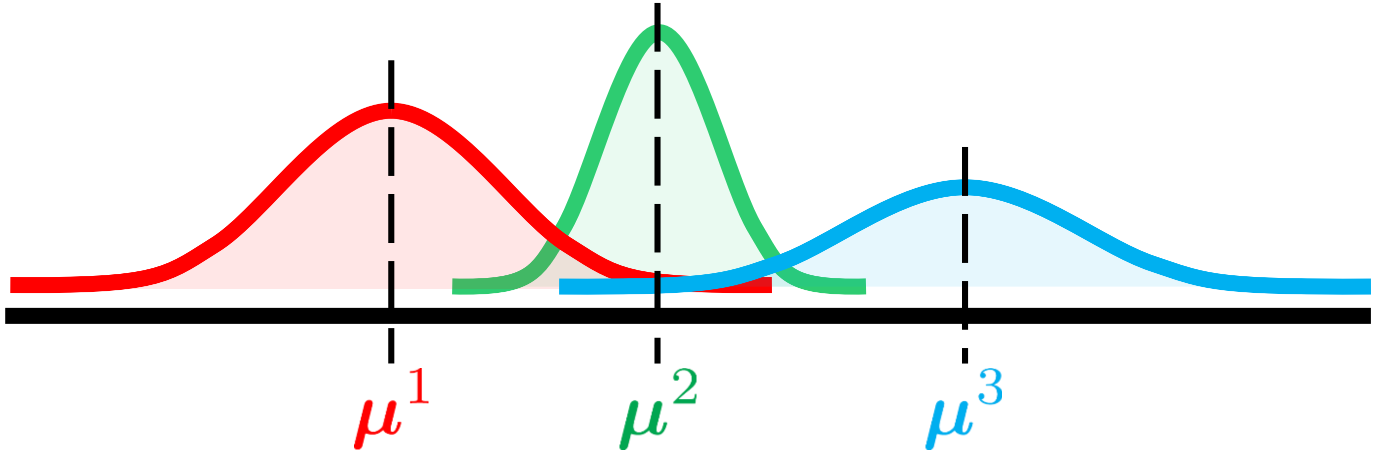


Training data is N emails, no labels

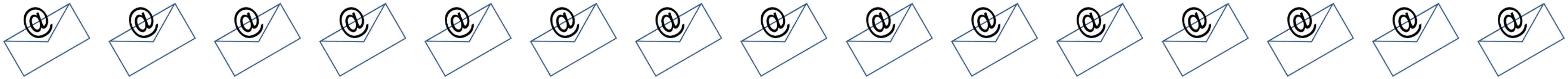


$$\mathbb{P}\left[\mathbf{x}^i \mid \mathbf{\Theta}\right]$$

The generative story for unlabelled data

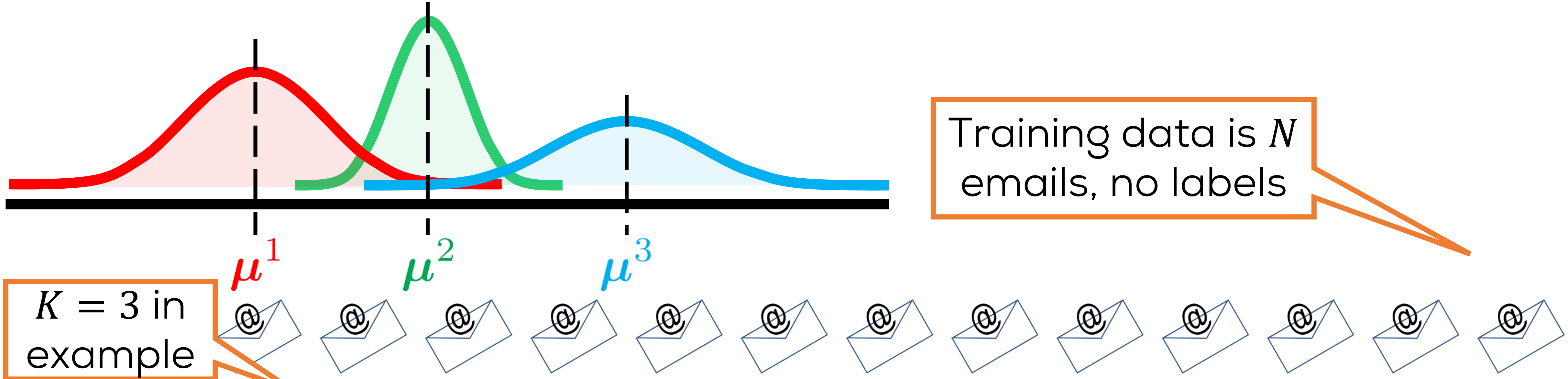


Training data is N emails, no labels



$$\mathbb{P}\left[\mathbf{x}^i \mid \mathbf{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \mathbf{\Theta}\right]$$

The generative story for unlabelled data

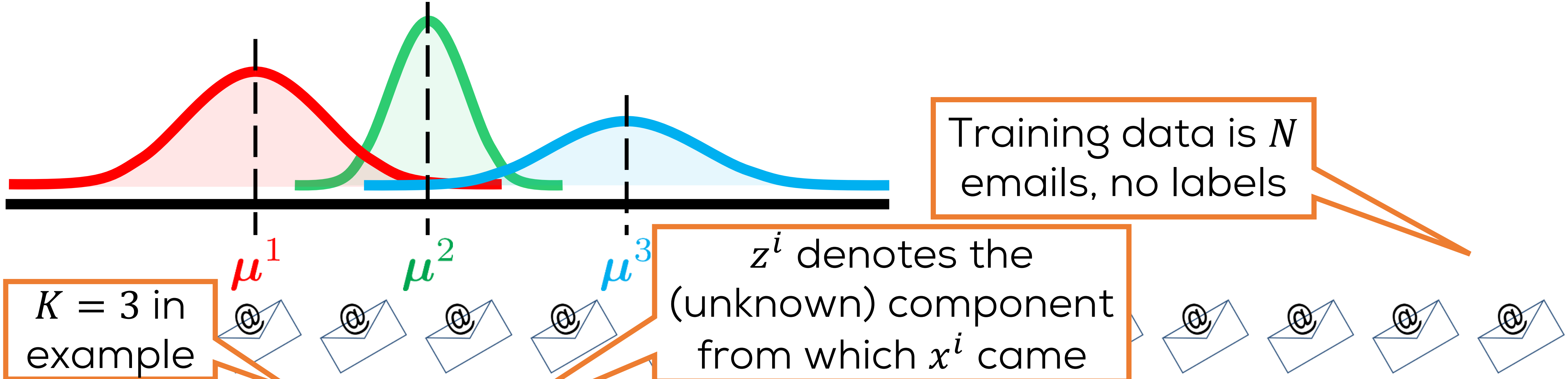


Training data is N emails, no labels

$K = 3$ in example

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \mathbf{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \,|\, \mathbf{\Theta}\right]$$

The generative story for unlabelled data



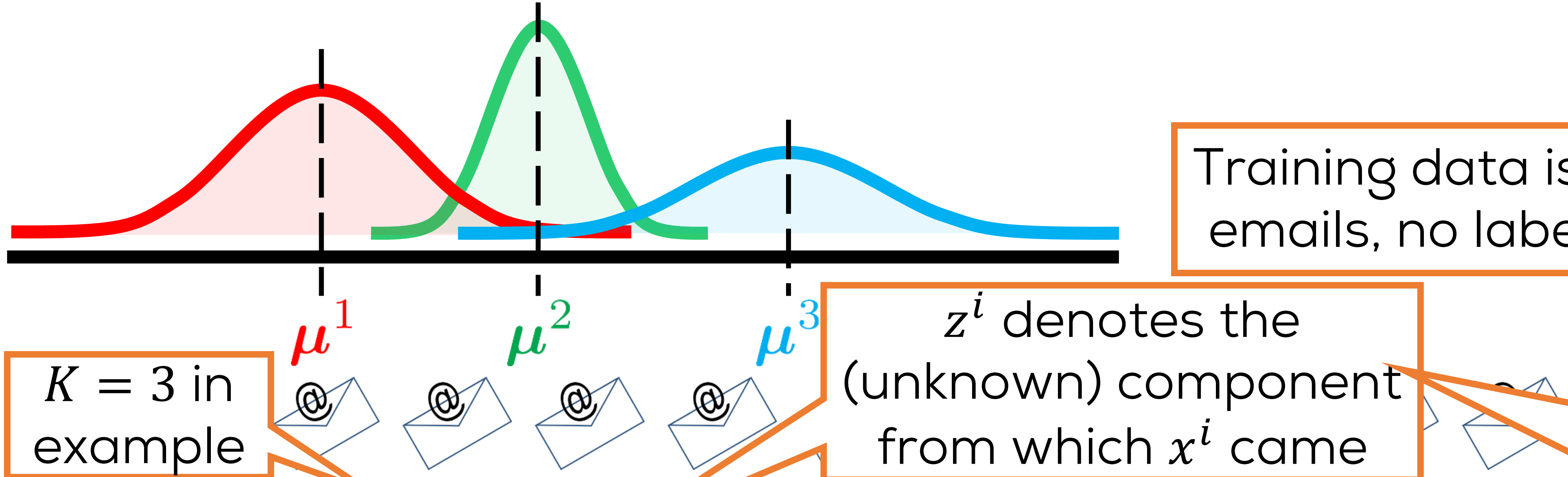
Training data is N emails, no labels

$K = 3$ in example

z^i denotes the (unknown) component from which x^i came

$$\mathbb{P}\left[\mathbf{x}^i \mid \mathbf{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \mathbf{\Theta}\right]$$

The generative story for unlabelled data



Training data is N emails, no labels

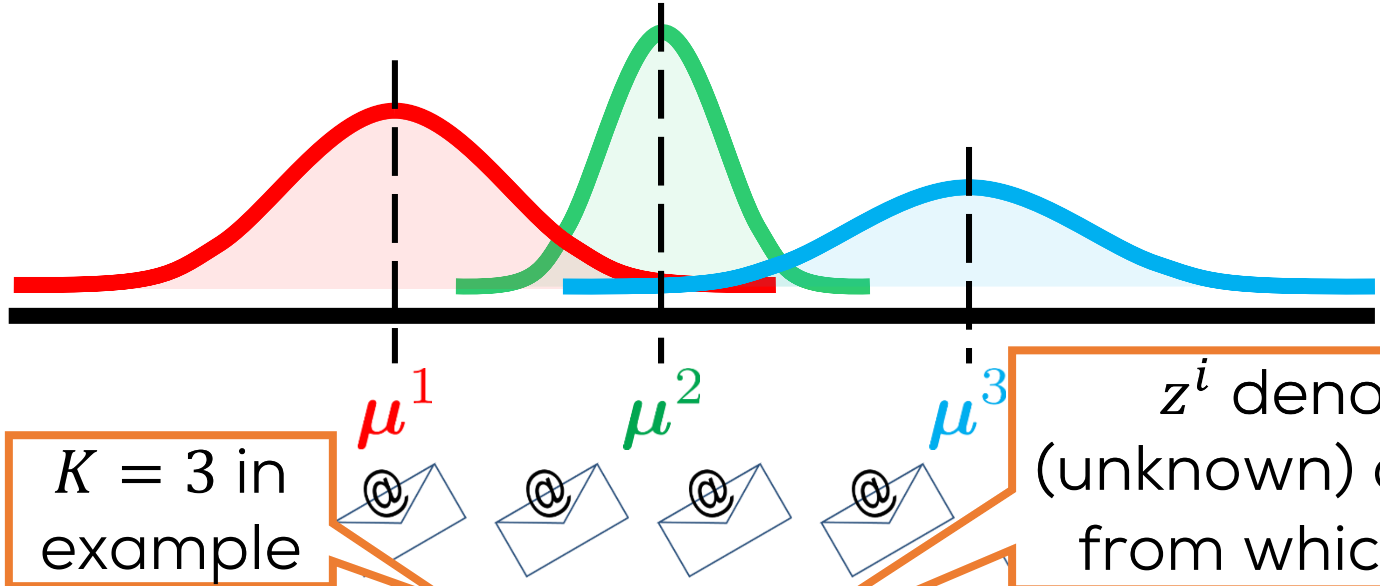
z^i denotes the (unknown) component from which x^i came

$K = 3$ in example

z^i not known from data. It is a *latent variable* (can take K values.

$$\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \boldsymbol{\Theta}\right]$$

The generative story for unlabelled data



Training data is N emails, no labels

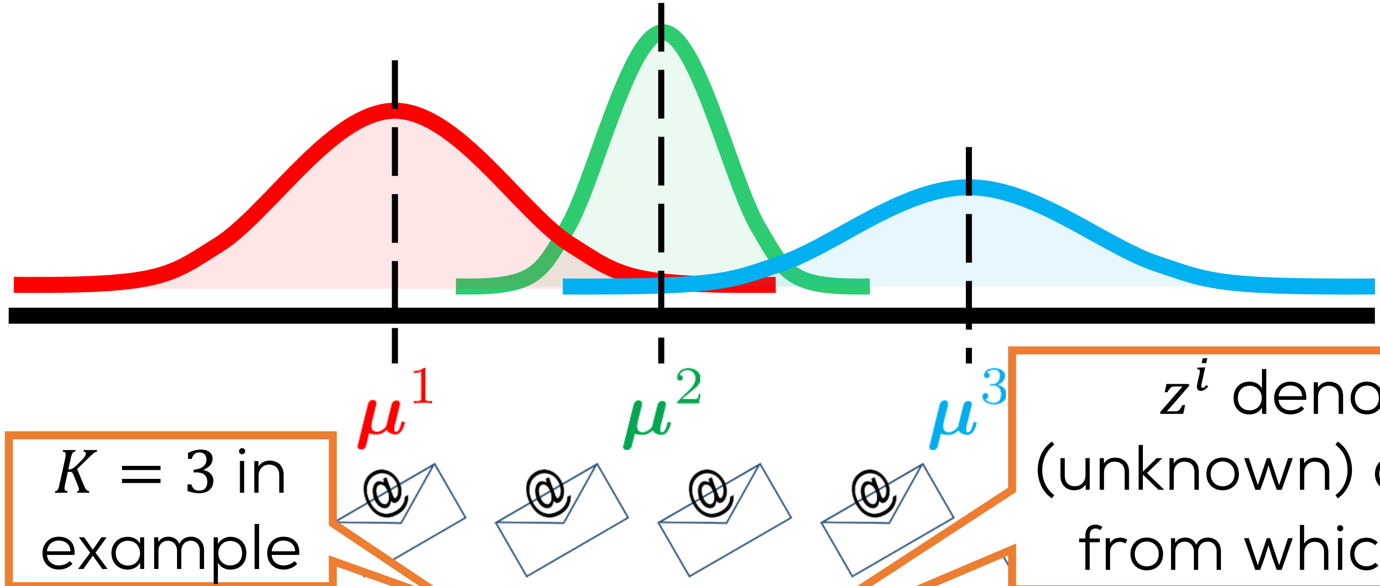
$K = 3$ in example

z^i denotes the (unknown) component from which x^i came

z^i not known from data. It is a *latent variable* (can take K values.

$$\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i = k, \boldsymbol{\Theta}\right]$$

The generative story for unlabelled data



Training data is N emails, no labels

$K = 3$ in example

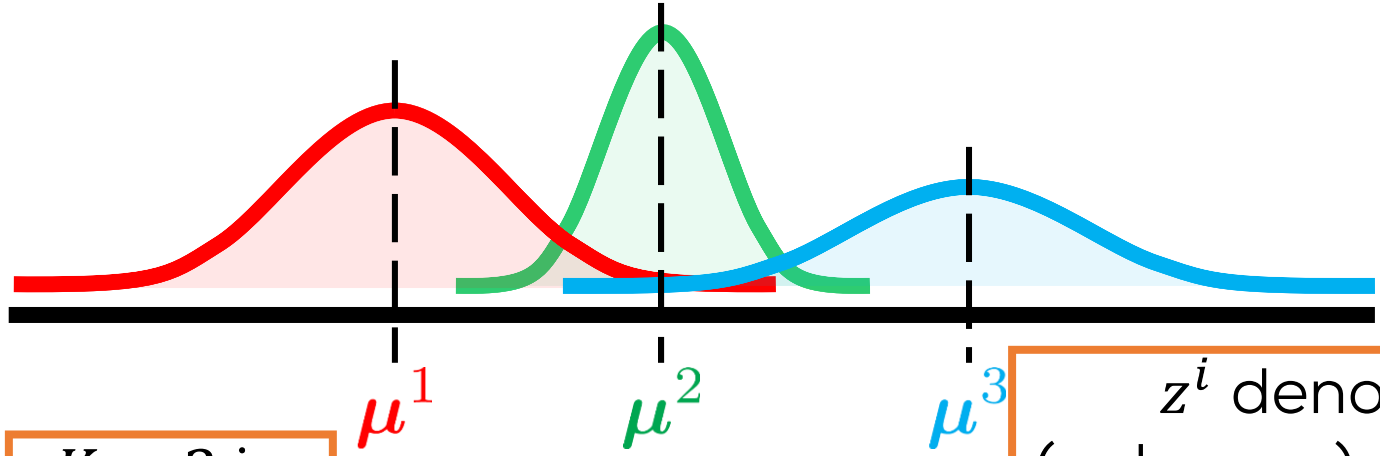
z^i denotes the (unknown) component from which x^i came

z^i not known from data. It is a *latent variable* (can take K values.

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \mathbf{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \,|\, \mathbf{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathbb{P}\left[\mathbf{x}^i \,|\, z^i = k, \mathbf{\Theta}\right]$$

$$= \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

The generative story for unlabelled data



Training data is N emails, no labels

$K = 3$ in example

z^i denotes the (unknown) component from which x^i came

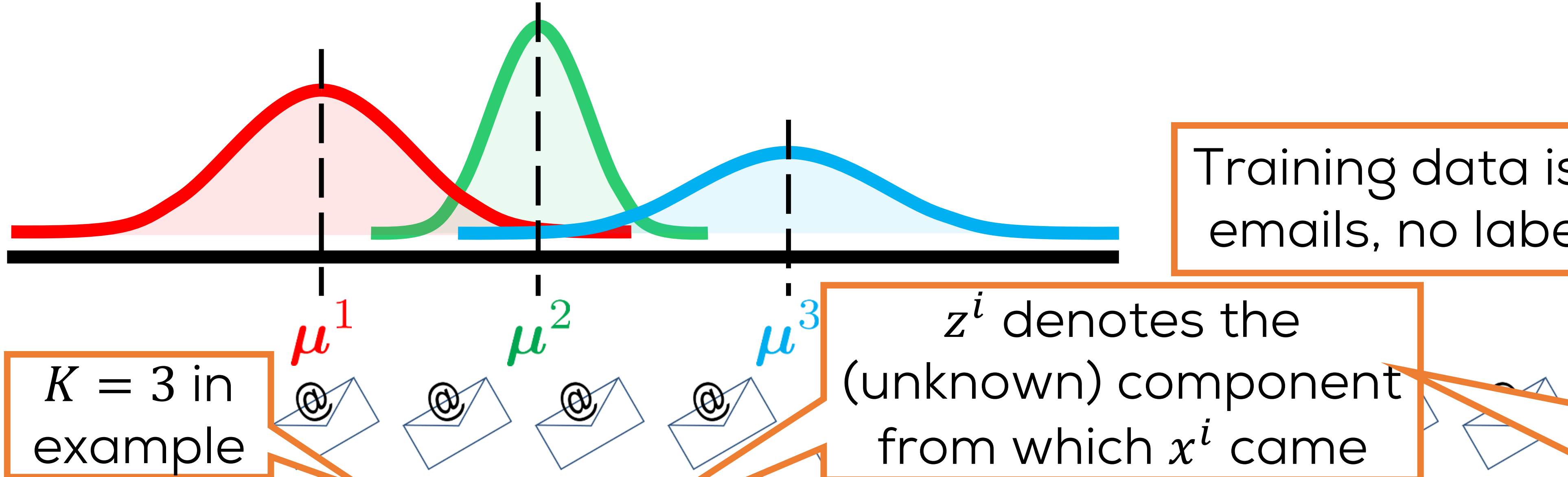
z^i not known from data. It is a *latent variable* (can take K values.

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathbb{P}\left[\mathbf{x}^i \,|\, z^i = k, \boldsymbol{\Theta}\right]$$

$$= \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

We assumed each component is a multidim. Gaussian

The generative story for unlabelled data



Training data is N emails, no labels

$K = 3$ in example

z^i denotes the (unknown) component from which x^i came

z^i not known from data. It is a *latent variable* (can take K values.

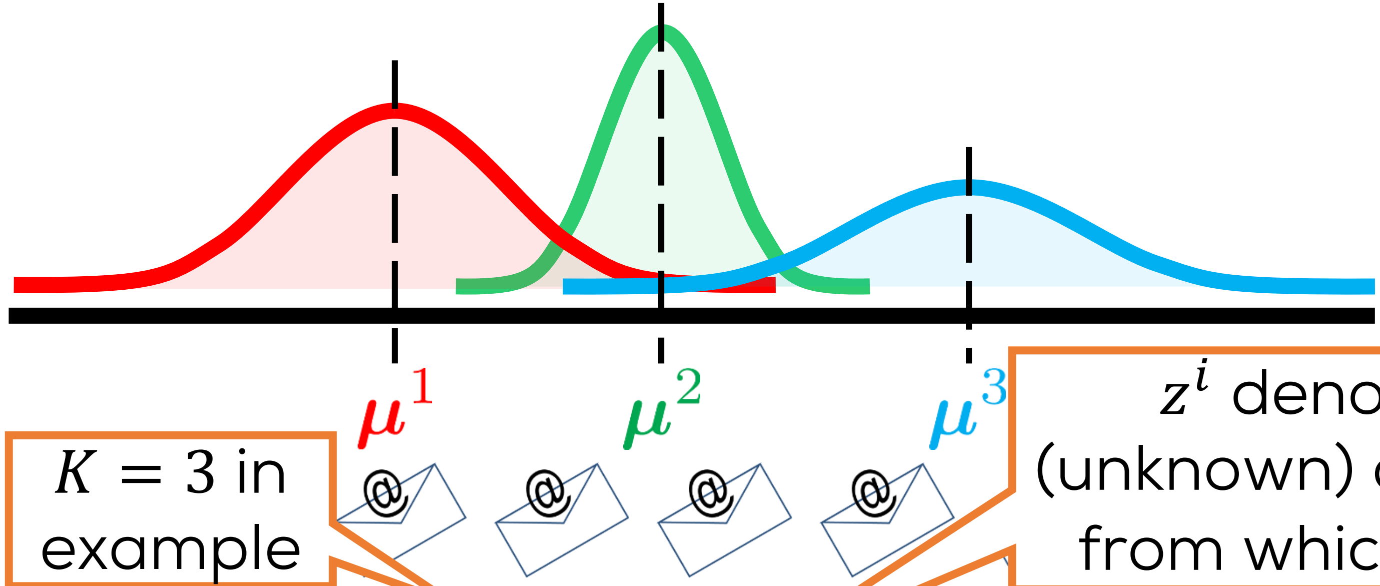
$$\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i = k, \boldsymbol{\Theta}\right]$$

$$= \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)$$

Gaussian Mixture Model (GMM) with K components

We assumed each component is a multidim. Gaussian

The generative story for unlabelled data



Training data is N emails, no labels

$K = 3$ in example

z^i denotes the (unknown) component from which x^i came

z^i not known from data. It is a *latent variable* (can take K values.

$$\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i = k, \boldsymbol{\Theta}\right]$$

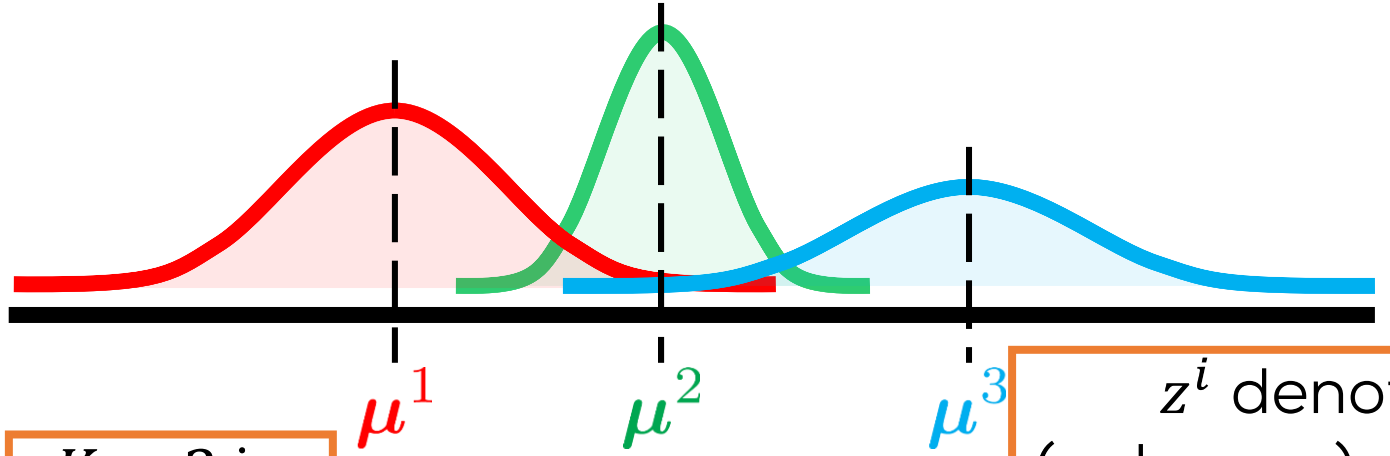
$$= \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)$$

Gaussian Mixture Model (GMM) with K components

We assumed each component is a multidim. Gaussian

$\mathbb{P}[z^i = k]$ prior prob. of $\mathbf{x}^i$ coming from k-th component

The generative story for unlabelled data



Training data is N emails, no labels

$K = 3$ in example

z^i denotes the (unknown) component from which x^i came

z^i not known from data. It is a *latent variable* (can take K values.

$$\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \mathbb{P}\left[\mathbf{x}^i, z^i = k \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathbb{P}\left[\mathbf{x}^i \mid z^i = k, \boldsymbol{\Theta}\right]$$

$$= \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)$$

Gaussian Mixture Model (GMM) with K components

Goal: incomplete data, learn $\boldsymbol{\mu}^k, \Sigma^k, \mathbb{P}[z = k]$

$\mathbb{P}[z^i = k]$ prior prob. of $\mathbf{x}^i$ coming from k-th component

We assumed each component is a multidim. Gaussian

The Likelihood Expression

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \mathbf{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right]$$

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \mid \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \mid \boldsymbol{\Theta}\right]$$

$$\begin{aligned}\hat{\boldsymbol{\Theta}}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X \mid \boldsymbol{\Theta}\right] \\ &= \arg\max_{\boldsymbol{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)\end{aligned}$$

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right]$$

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right]$$

$$= \arg\max_{\boldsymbol{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

Cannot apply first order Optimality to get solution

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right]$$

$$\begin{aligned}\hat{\boldsymbol{\Theta}}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right] \\ &= \arg\max_{\boldsymbol{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)\end{aligned}$$

Cannot apply first order Optimality to get solution

Horribly non-convex problem. Initialization matters!

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \mathbf{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \,|\, \mathbf{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,|\, \mathbf{\Theta}\right]$$

$$\hat{\mathbf{\Theta}}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \,|\, \mathbf{\Theta}\right]$$

$$= \arg\max_{\mathbf{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

Cannot apply first order Optimality to get solution

Horribly non-convex problem. Initialization matters!

Getting to global optimum may be NP-hard ☹

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right]$$

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right]$$

$$= \arg\max_{\boldsymbol{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

Things were so nice when we had the labels 😐

Cannot apply first order Optimality to get solution

Horribly non-convex problem. Initialization matters!

Getting to global optimum may be NP-hard ☹

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \mid \mathbf{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)$$

$$\mathbb{P}\left[X \mid \mathbf{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \mid \mathbf{\Theta}\right]$$

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

$$= \arg\max_{\mathbf{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^k, \Sigma^k)$$

Things were so nice when we had the labels 😐

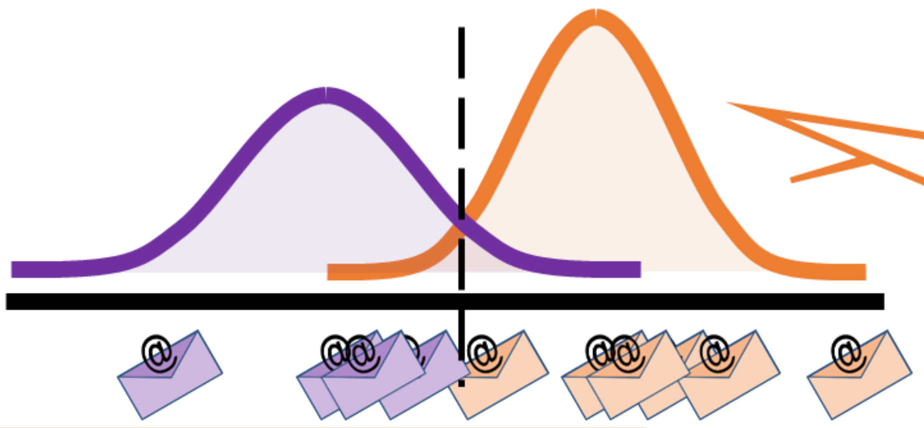
wait …

Cannot apply first order Optimality to get solution

Horribly non-convex problem. Initialization matters!

Getting to global optimum may be NP-hard ☹

The generative story for unlabelled data??



No labels!!!

Give the Gaussians colors to help identifying them

Can we still recover the two Gaussian components in the mixture??

We do not know which email is purple and which is orange – no labels!!

Hmm … what if someone magically labelled the emails with the color

How to get these magical labels??

Hmm … what if someone gave me a purple and an orange Gaussian, maybe slightly wrong ones

Can use these to label emails!

What if I use these “magical” labels to learn two Gaussians?

Won’t I just learn the wrong ones that generated the labels?

In practice: you will learn slightly “less wrong” ones ☺

The Likelihood Expression

$$\mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right] = \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

Things were so nice when we had the labels 😐

$$\mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}^i \,|\, \boldsymbol{\Theta}\right]$$

wait …

Cannot apply first order Optimality to get solution

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X \,|\, \boldsymbol{\Theta}\right]$$

$$= \arg\max_{\boldsymbol{\Theta}} \prod_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{\pi}_k \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^k, \Sigma^k)$$

Horribly non-convex problem. Initialization matters!

Getting to global optimum may be NP-hard ☹

A Ray of Hope

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

Can use a method like the one discussed in the "detour"!

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

Various ways of updating z^i

Can use a method like the one discussed in the "detour"!

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

Looks like block coordinate descent with $\mathbf{\Theta}, \{z^i\}$ being two blocks of "coordinates"

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

Various ways of updating z^i

Can use a method like the one discussed in the "detour"!

Hard Assignment

The K-means algorithm

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

Looks like block coordinate descent with $\Theta, \{z^i\}$ being two blocks of "coordinates"

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

Various ways of updating z^i

Can use a method like the one discussed in the "detour"!

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\mathrm{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \,|\, \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \,|\, \mathbf{\Theta}\right]$
4. Repeat until convergence

$$z^{i,t} = \arg\max_{k \in [K]} \mathbb{P}\left[k \,|\, \mathbf{x}^i, \mathbf{\Theta}^t\right]$$

A Ray of Hope

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X \mid \boldsymbol{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
3. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \boldsymbol{\Theta}\right]$
4. Repeat until convergence

$$z^{i,t} = \arg\max_{k \in [K]} \mathbb{P}\left[k \mid \mathbf{x}^i, \boldsymbol{\Theta}^t\right]$$

$$\boldsymbol{\Theta}^t = \left\{\boldsymbol{\pi}^t, \left\{\boldsymbol{\mu}^{1,t}, \boldsymbol{\mu}^{2,t}, \boldsymbol{\mu}^{3,t}\right\}, \left\{\Sigma^{1,t}, \Sigma^{2,t}, \Sigma^{3,t}\right\}\right\}$$

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

Bayes Rule!

$$z^{i,t} = \arg\max_{k \in [K]} \mathbb{P}\left[k \mid \mathbf{x}^i, \mathbf{\Theta}^t\right] = \arg\max_{k \in [K]} \mathbb{P}\left[k \mid \mathbf{\Theta}^t\right] \cdot \mathbb{P}\left[\mathbf{x}^i \mid k, \mathbf{\Theta}^t\right]$$

$$\mathbf{\Theta}^t = \left\{\boldsymbol{\pi}^t, \left\{\boldsymbol{\mu}^{1,t}, \boldsymbol{\mu}^{2,t}, \boldsymbol{\mu}^{3,t}\right\}, \left\{\Sigma^{1,t}, \Sigma^{2,t}, \Sigma^{3,t}\right\}\right\}$$

A Ray of Hope

$$\hat{\mathbf{\Theta}}_{\text{MLE}} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X \mid \mathbf{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\mathbf{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$
3. Update $\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \left\{z^{i,t}\right\} \mid \mathbf{\Theta}\right]$
4. Repeat until convergence

Bayes Rule!

$$z^{i,t} = \arg\max_{k \in [K]} \mathbb{P}\left[k \mid \mathbf{x}^i, \mathbf{\Theta}^t\right] = \arg\max_{k \in [K]} \boldsymbol{\pi}_k^t \cdot \mathbb{P}\left[\mathbf{x}^i \mid k, \mathbf{\Theta}^t\right]$$

$$\mathbf{\Theta}^t = \left\{\boldsymbol{\pi}^t, \left\{\boldsymbol{\mu}^{1,t}, \boldsymbol{\mu}^{2,t}, \boldsymbol{\mu}^{3,t}\right\}, \left\{\Sigma^{1,t}, \Sigma^{2,t}, \Sigma^{3,t}\right\}\right\}$$

A Ray of Hope

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \underset{\boldsymbol{\Theta}}{\arg\max}\ \mathbb{P}\left[X \mid \boldsymbol{\Theta}\right]$$

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
3. Update $\boldsymbol{\Theta}^{t+1} = \underset{\boldsymbol{\Theta}}{\arg\max}\ \mathbb{P}\left[X, \left\{z^{i,t}\right\} \mid \boldsymbol{\Theta}\right]$
4. Repeat until convergence

Bayes Rule!

$$z^{i,t} = \underset{k \in [K]}{\arg\max}\ \mathbb{P}\left[k \mid \mathbf{x}^i, \boldsymbol{\Theta}^t\right] = \underset{k \in [K]}{\arg\max}\ \boldsymbol{\pi}_k^t \cdot \mathcal{N}(\mathbf{x}^i \mid \boldsymbol{\mu}^{k,t}, \Sigma^{k,t})$$

$$\boldsymbol{\Theta}^t = \left\{\boldsymbol{\pi}^t, \left\{\boldsymbol{\mu}^{1,t}, \boldsymbol{\mu}^{2,t}, \boldsymbol{\mu}^{3,t}\right\}, \left\{\Sigma^{1,t}, \Sigma^{2,t}, \Sigma^{3,t}\right\}\right\}$$

Towards the K-Means Algorithm

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
    1. Let $z^{i,t} = \arg\max_k \boldsymbol{\pi}_k^t \cdot \mathcal{N}\left(\mathbf{x}^i \mid \boldsymbol{\mu}^{k,t}, \Sigma^{k,t}\right)$
3. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \left\{z^{i,t}\right\} \mid \boldsymbol{\Theta}\right]$
    1. Let $\boldsymbol{\pi}_k^{t+1} = \frac{n_k^t}{n}$, where $\mathrm{n}_k^t = \left|\left\{i: \, z^{i,t} = k\right\}\right|$
    2. Let $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i: z^{i,t} = k} \mathbf{x}^i$
    3. Let $\Sigma_k^{t+1} = \frac{1}{n_k^t} \sum_{i: z^{i,t} = k} \left(\mathbf{x}^i - \boldsymbol{\mu}^{k,t+1}\right)\left(\mathbf{x}^i - \boldsymbol{\mu}^{k,t+1}\right)^\top$
4. Repeat until convergence

A few simplifications

- Fix $\boldsymbol{\pi}_k^t = \frac{1}{K}$ for all iterations. Don't update it.
- Fix $\boldsymbol{\Sigma}^{k,t} = I$ for all iterations. Don't update it.

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   1. Let $z^{i,t} = \arg\max_k \mathcal{N}\left(\mathbf{x}^i \mid \boldsymbol{\mu}^{k,t}, I\right)$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

A few simplifications

- Fix $\boldsymbol{\pi}_k^t = \frac{1}{K}$ for all iterations. Don't update it.
- Fix $\boldsymbol{\Sigma}^{k,t} = I$ for all iterations. Don't update it.

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
    1. Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

A few simplifications

- Fix $\boldsymbol{\pi}_k^t = \frac{1}{K}$ for all iterations. Don't update it.
- Fix $\boldsymbol{\Sigma}^{k,t} = I$ for all iterations. Don't update it.

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
    1. Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

The K-Means Objective

The K-Means Objective

$$\hat{\mathbf{\Theta}}_{\mathrm{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^k\right\|_2^2$$

The K-Means Objective

$$\hat{\mathbf{\Theta}}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^k\right\|_2^2$$

For cluster k, we have cluster center $\boldsymbol{\mu}^k$

The K-Means Objective

$$\hat{\Theta}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

For cluster k,

$$\sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

is a measure of how much variance is in the cluster

For cluster k, we have cluster center $\boldsymbol{\mu}^k$

The K-Means Objective

$$\hat{\mathbf{\Theta}}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

For cluster k,

$$\sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

is a measure of how much variance is in the cluster

Want to minimize sum of variance across all clusters. Want tight clusters

For cluster k, we have cluster center $\boldsymbol{\mu}^k$

The K-Means Objective

$$\hat{\boldsymbol{\Theta}}_{\mathrm{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^k\right\|_2^2$$

The K-Means Objective

$$\hat{\mathbf{\Theta}}_{\mathrm{km}} = \arg\min_{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^k\right\|_2^2$$

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   1. Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

The K-Means Objective

$$\hat{\Theta}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\dots,K} \\ \{z^i\}_{i=1,\dots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

Alternates between updating $\{z^i\}$ and $\{\mu^k\}$

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\dots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   1. Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

The K-Means Objective

$$\hat{\boldsymbol{\Theta}}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^k\right\|_2^2$$

An FA approach to solving a data modelling task!

Alternates between updating $\{z^i\}$ and $\{\mu^k\}$

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
    1. Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

The K-Means Objective

NP-hard problem!

An FA approach to solving a data modelling task!

$$\hat{\Theta}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

Alternates between updating $\{z^i\}$ and $\{\mu^k\}$

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   1. Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

The K-Means Objective

NP-hard problem!

An FA approach to solving a data modelling task!

$$\hat{\Theta}_{\text{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

Alternates between updating $\{z^i\}$ and $\{\mu^k\}$

Very scalable but sensitive to initialization!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
    1. Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

The K-Means Objective

$$\hat{\Theta}_{\mathrm{km}} = \underset{\substack{\{\boldsymbol{\mu}^k\}_{k=1,\ldots,K} \\ \{z^i\}_{i=1,\ldots,n}}}{\arg\min} \sum_{k=1}^{K} \sum_{i:z^i=k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^k \right\|_2^2$$

NP-hard problem!

An FA approach to solving a data modelling task!

Alternates between updating $\{z^i\}$ and $\{\mu^k\}$

Very scalable but sensitive to initialization!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   1. Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

k-means++ initialization

1. Sample $i_1 \sim [n]$, let $\boldsymbol{\mu}^{1,0} = \mathbf{x}^{i_1}$
2. For k = 2,…K
   - Sample $i_k \propto$ min distance from $\{\boldsymbol{\mu}^{1,0}, \ldots, \boldsymbol{\mu}^{k-1,0}\}$
   - Let $\boldsymbol{\mu}^{k,0} = \mathbf{x}^{i_k}$

K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

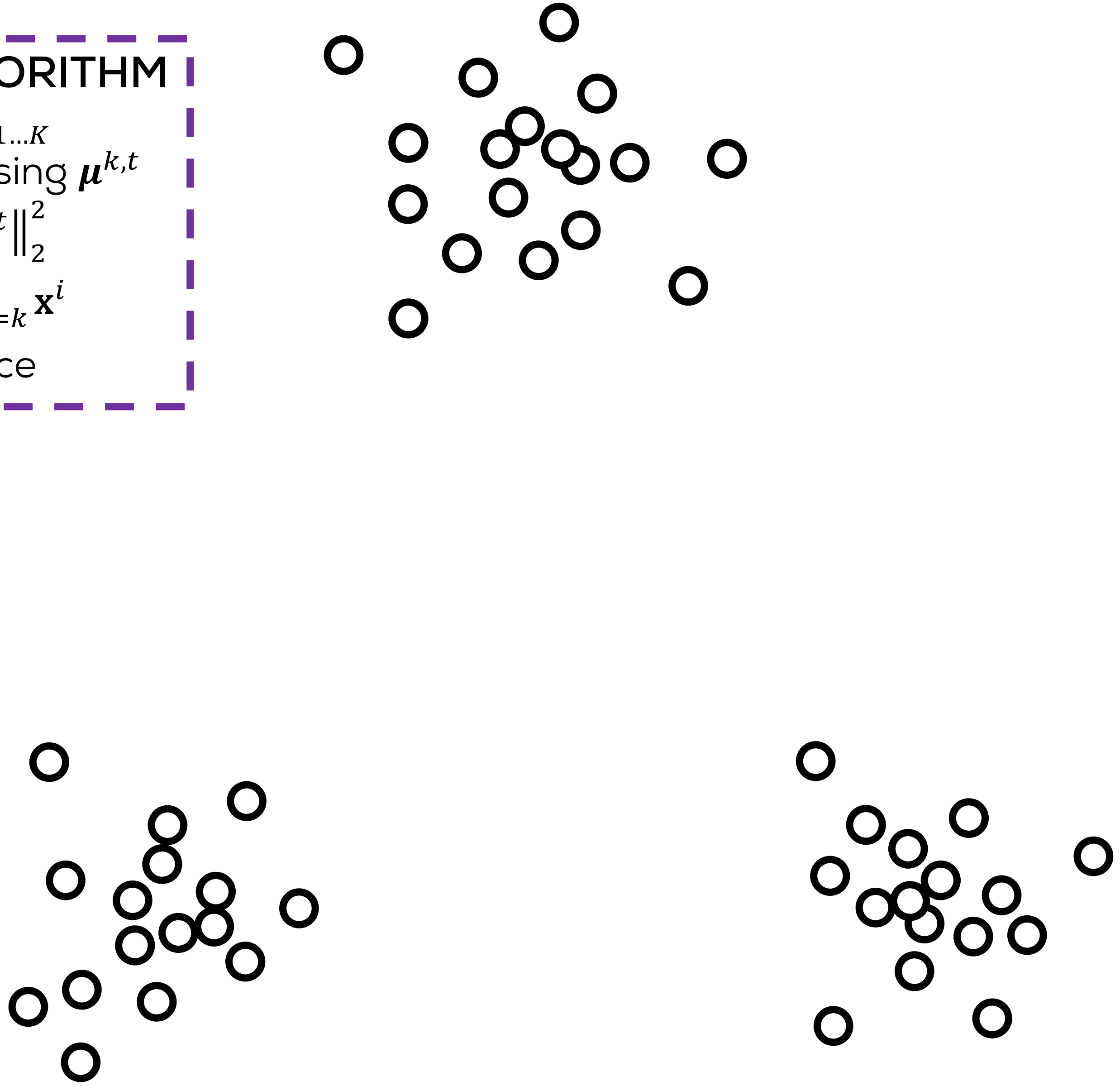
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

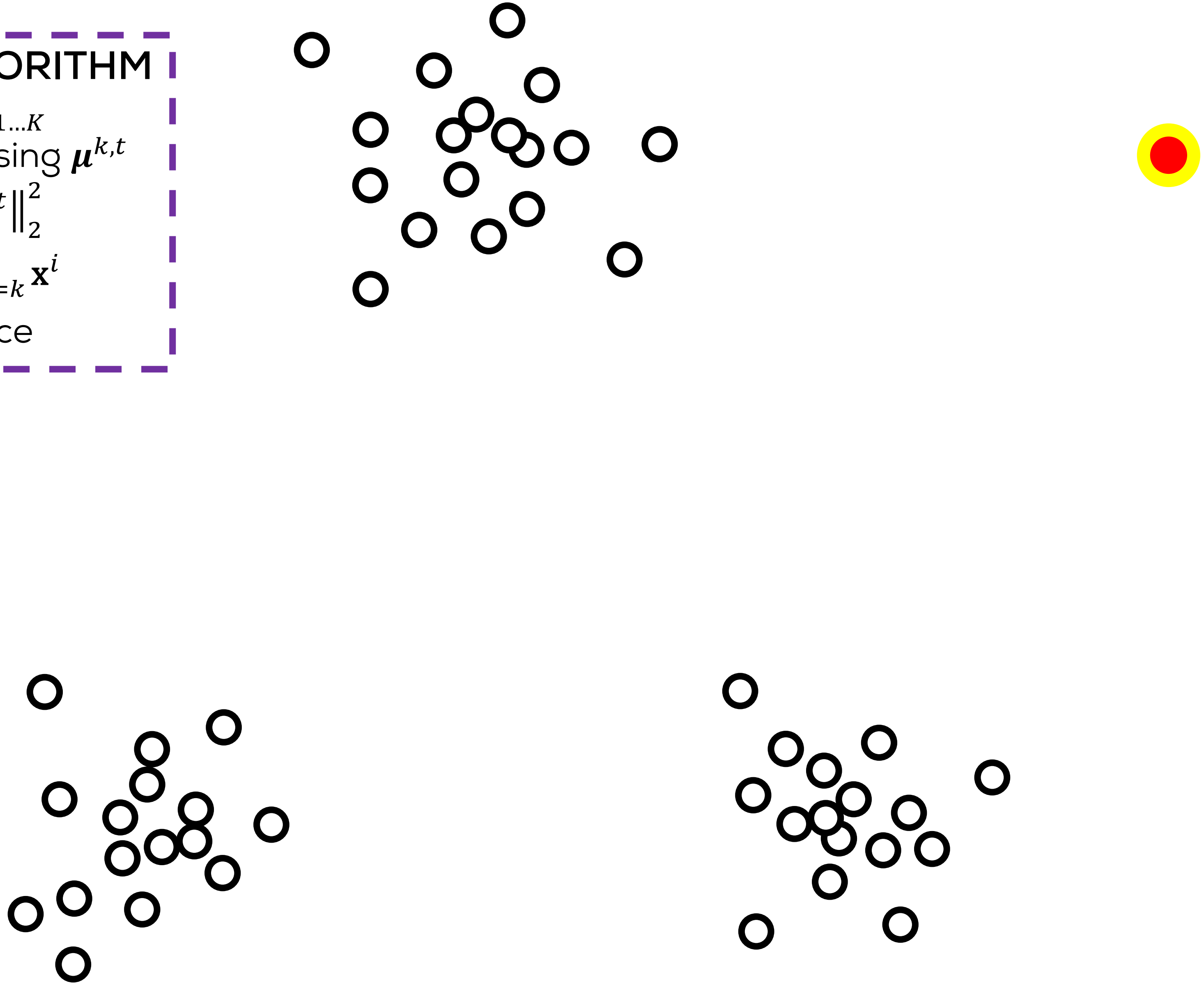
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

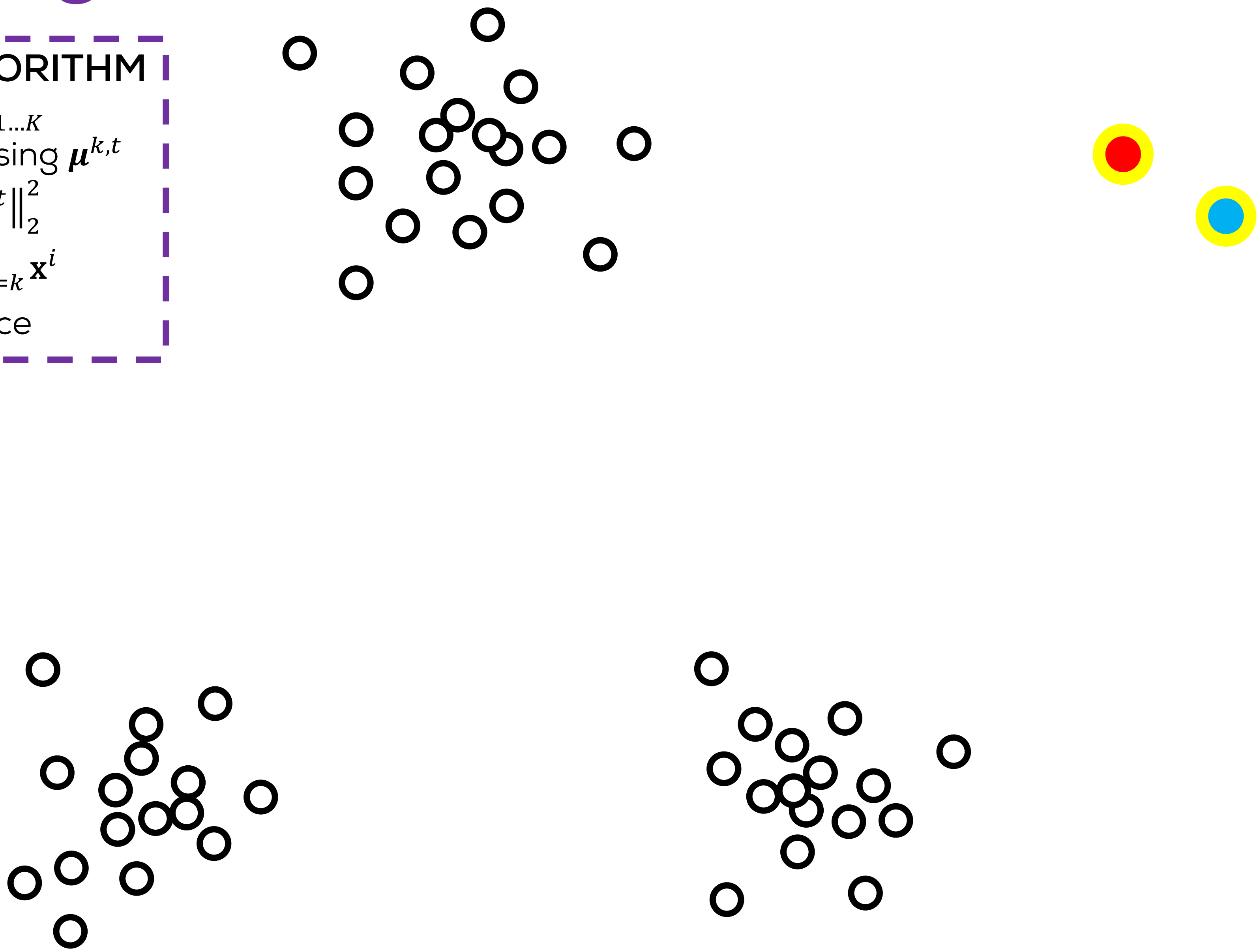


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

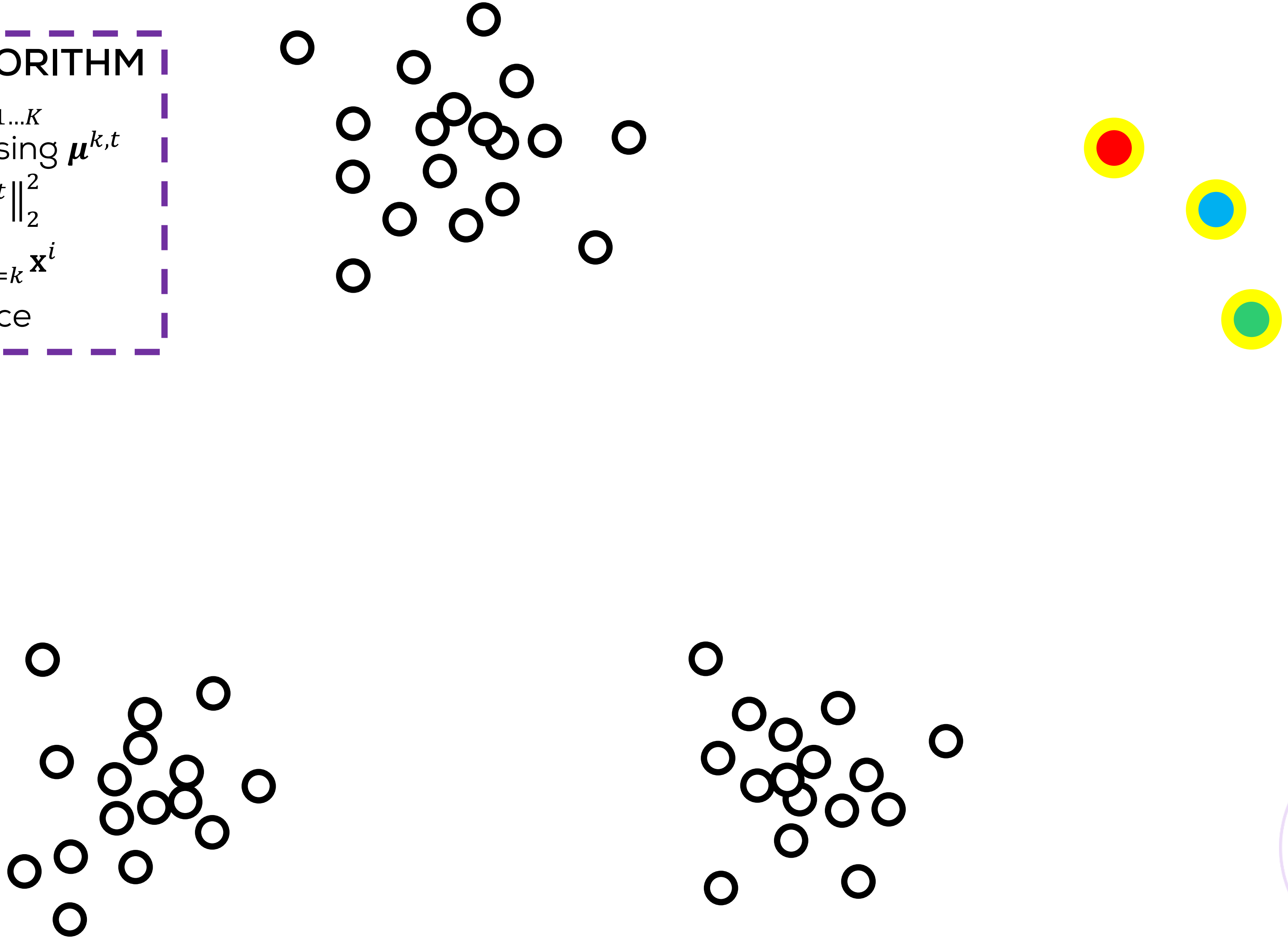


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i : z^{i,t} = k} \mathbf{x}^i$
4. Repeat until convergence

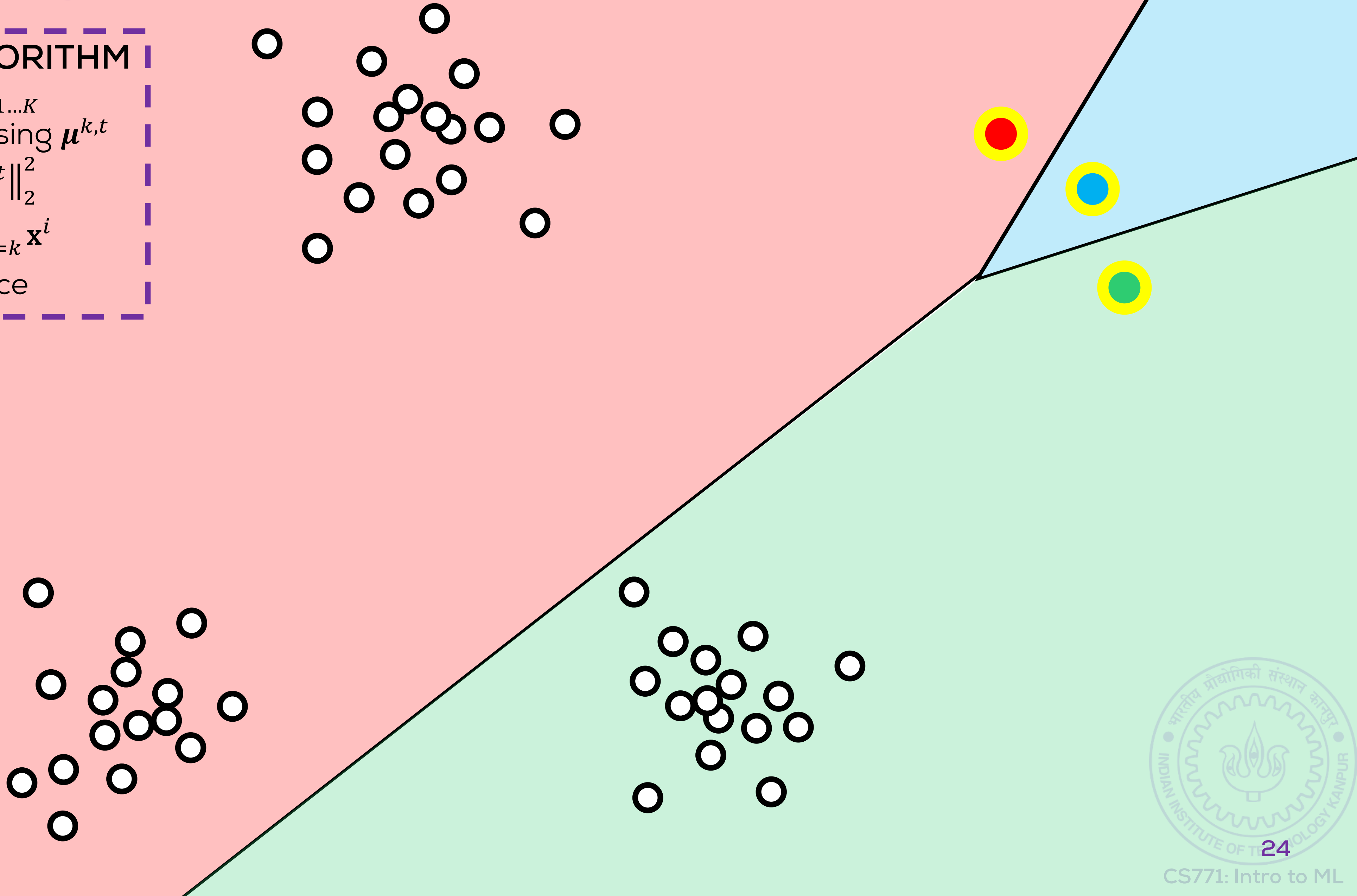


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

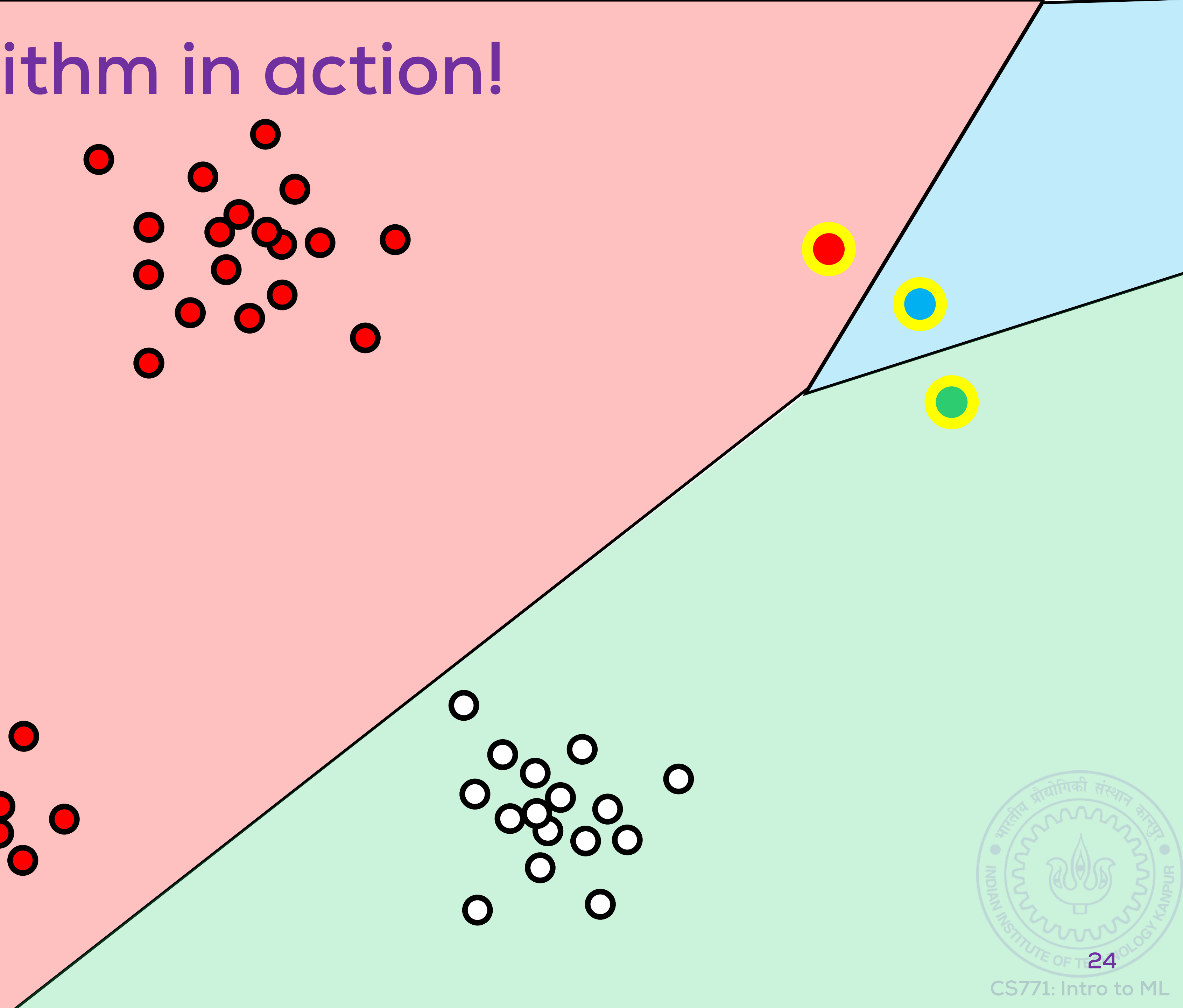


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

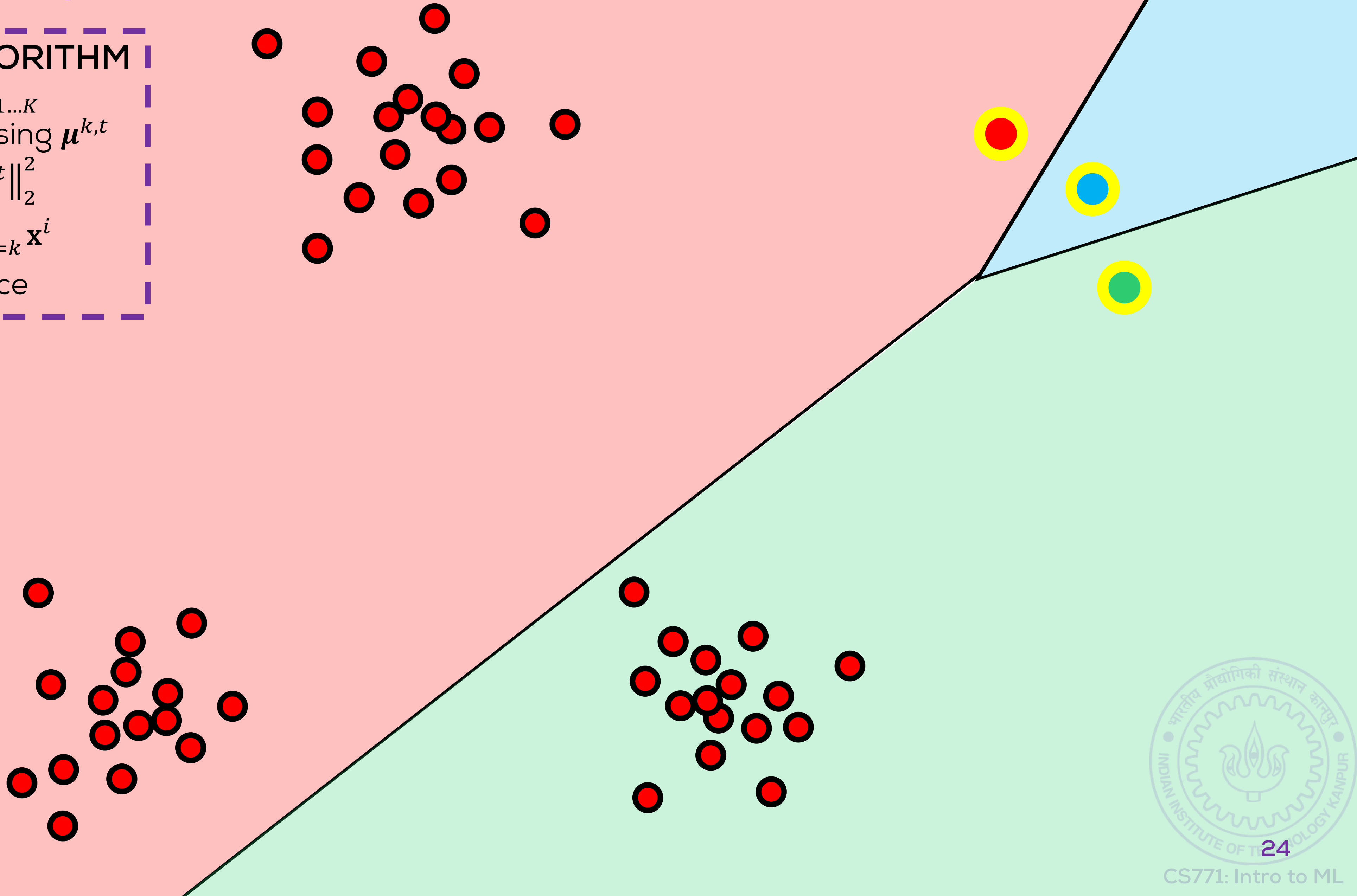


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

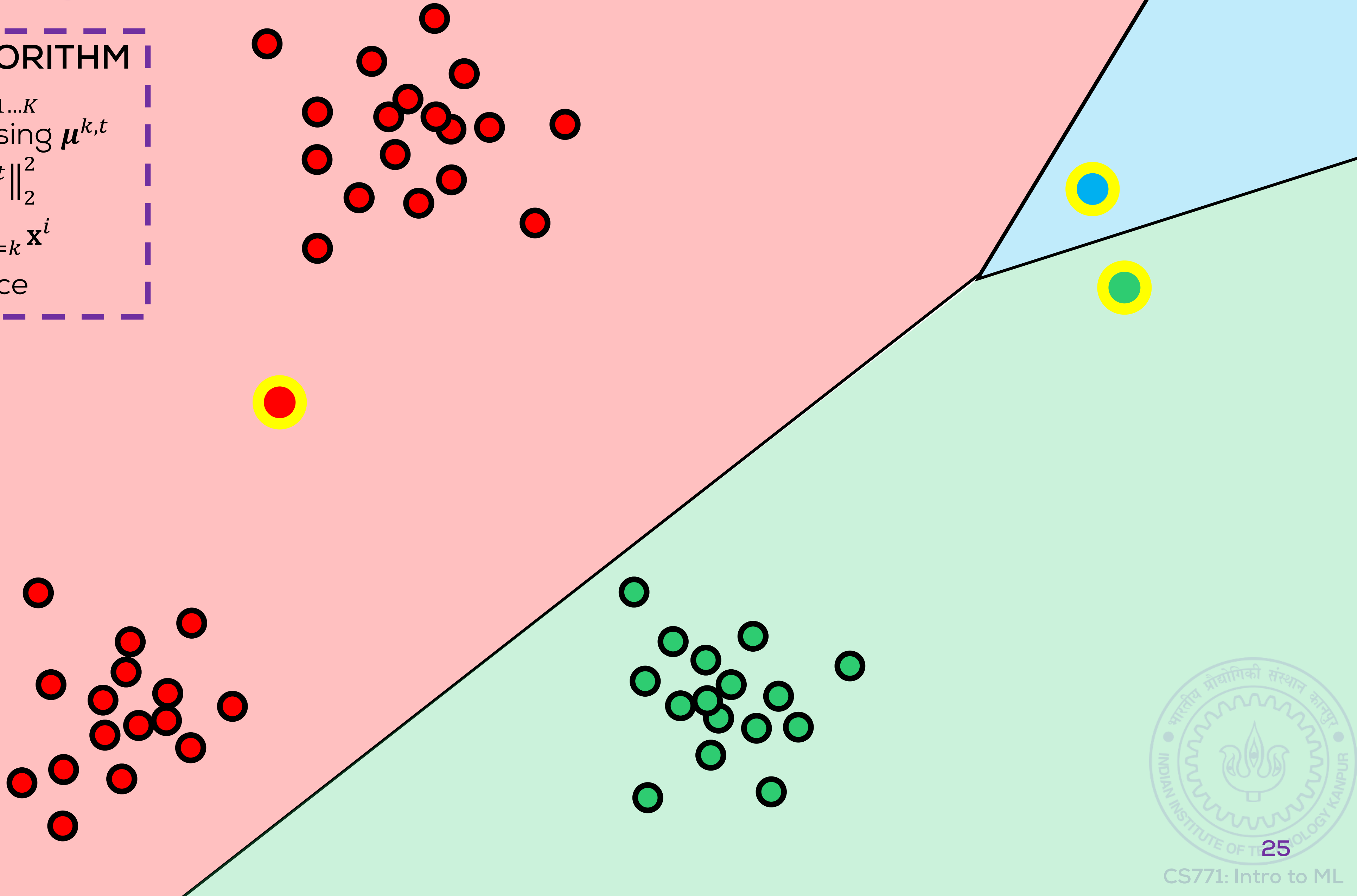


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

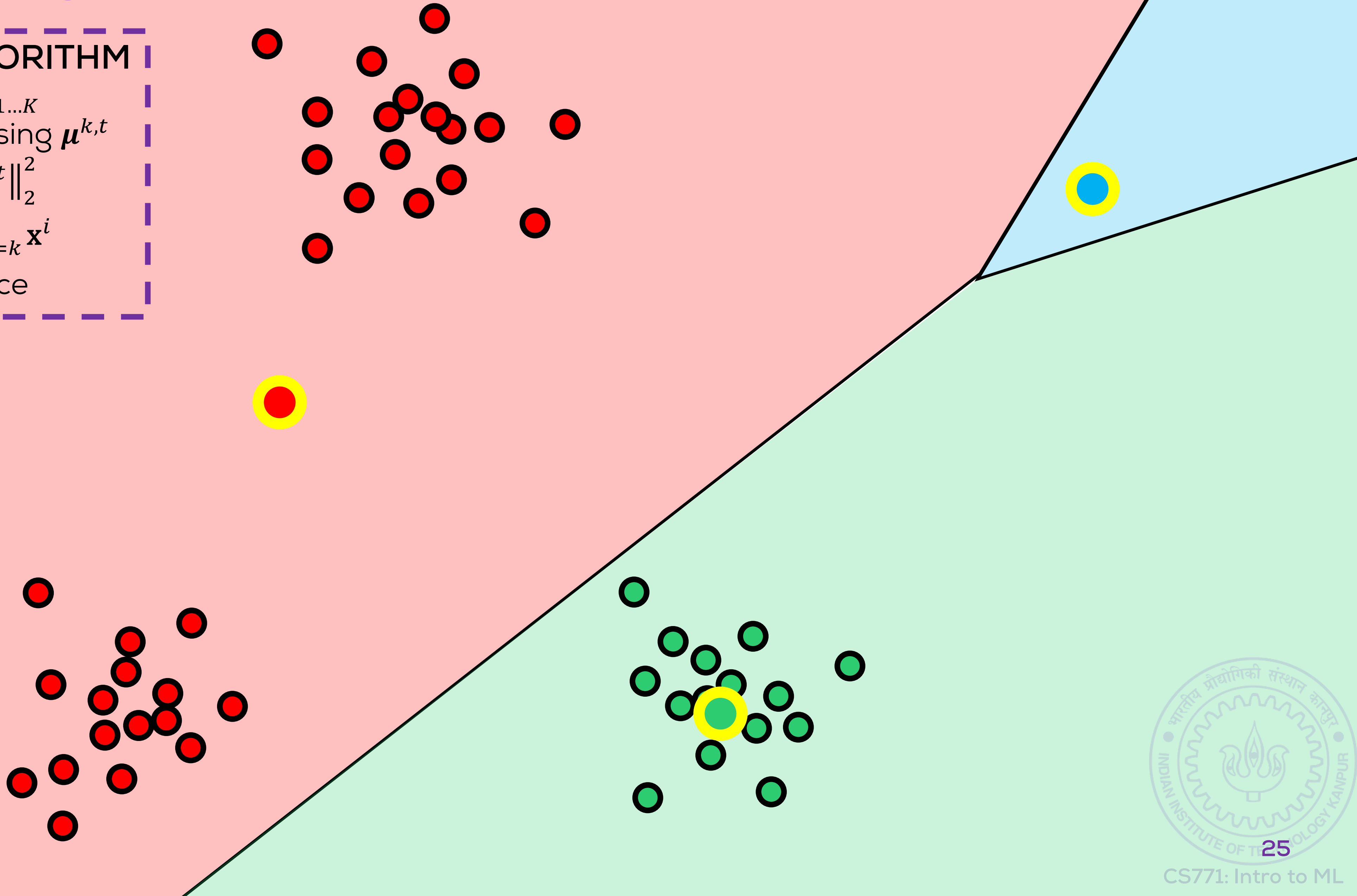


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

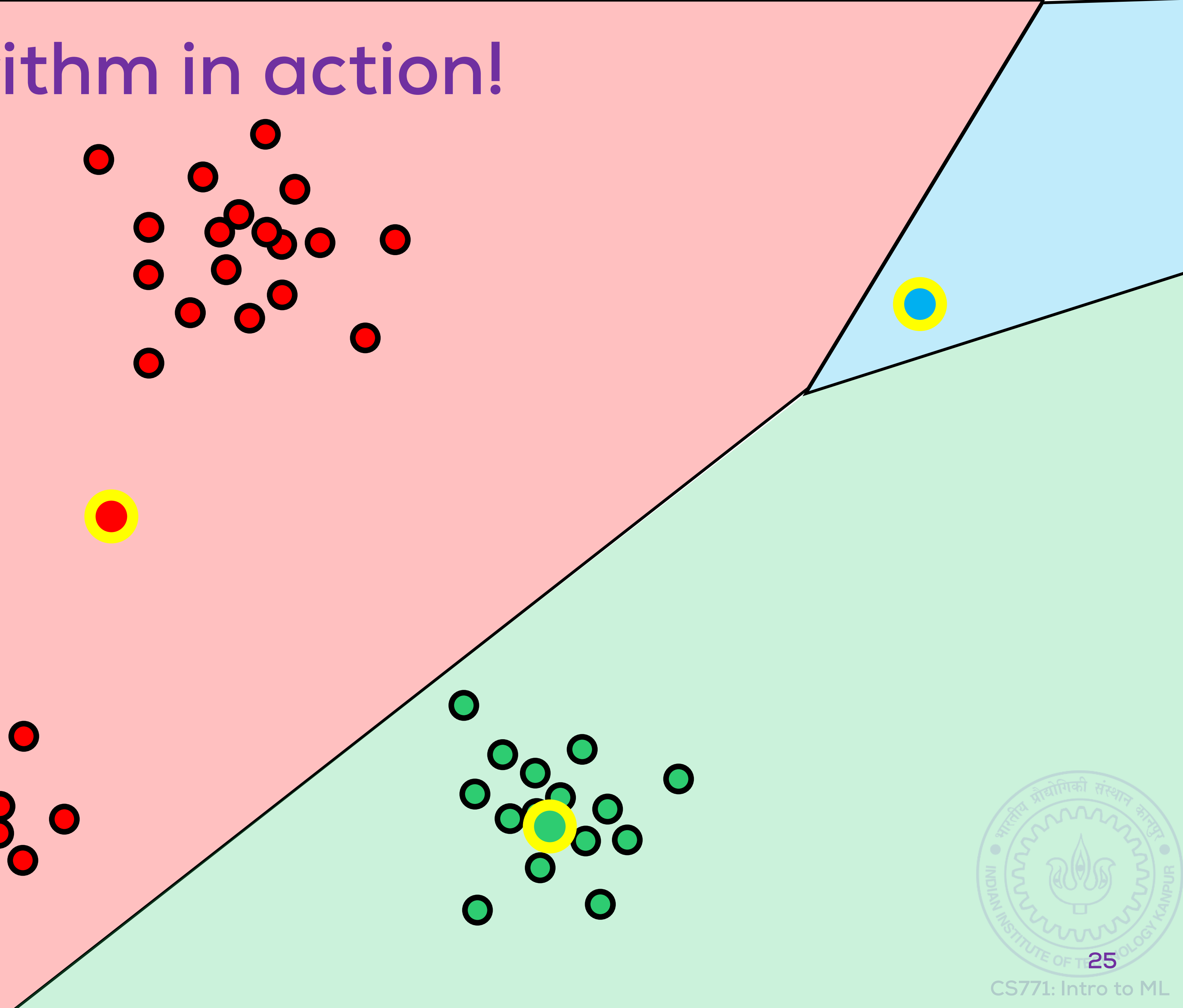


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

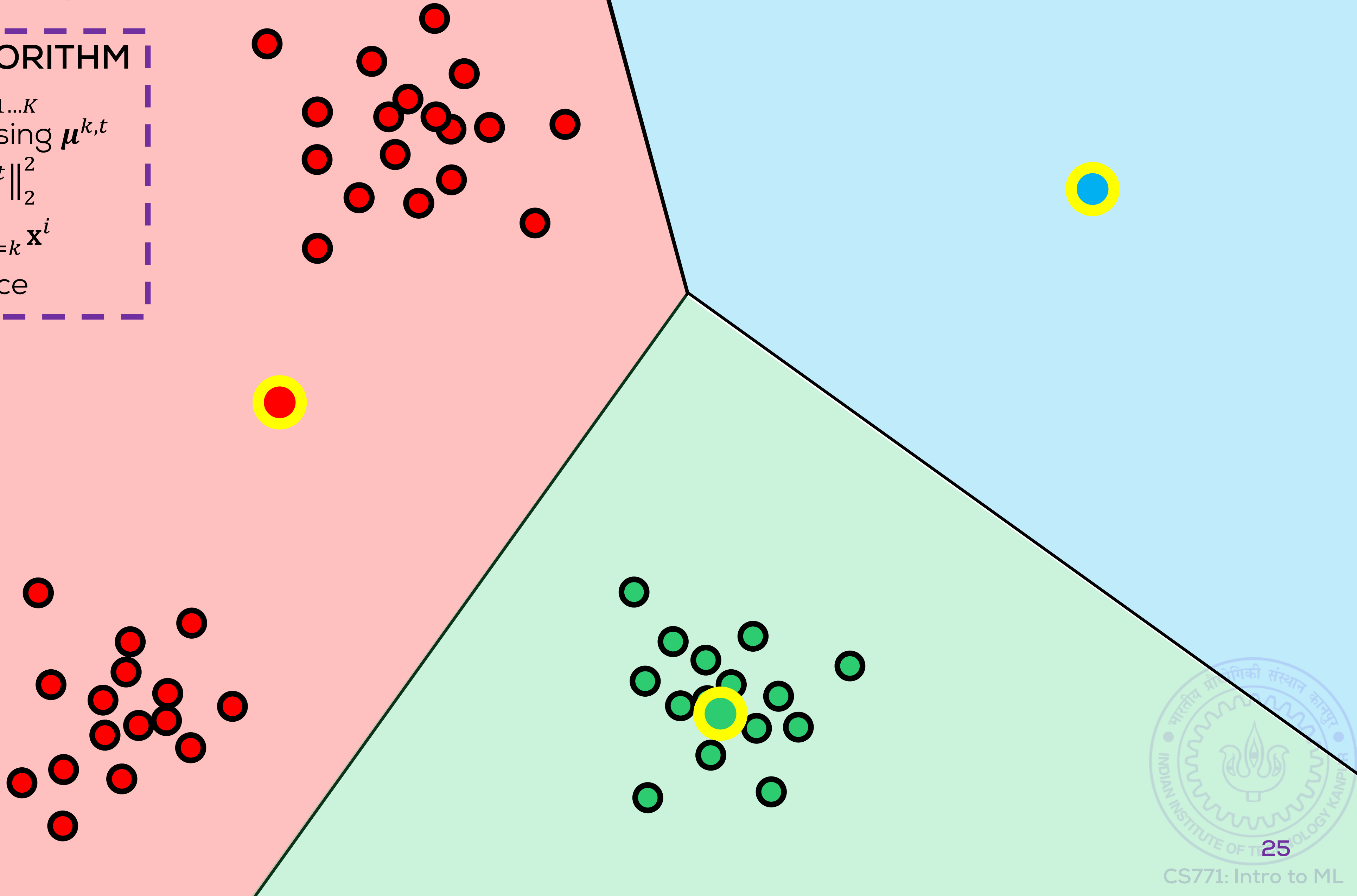


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

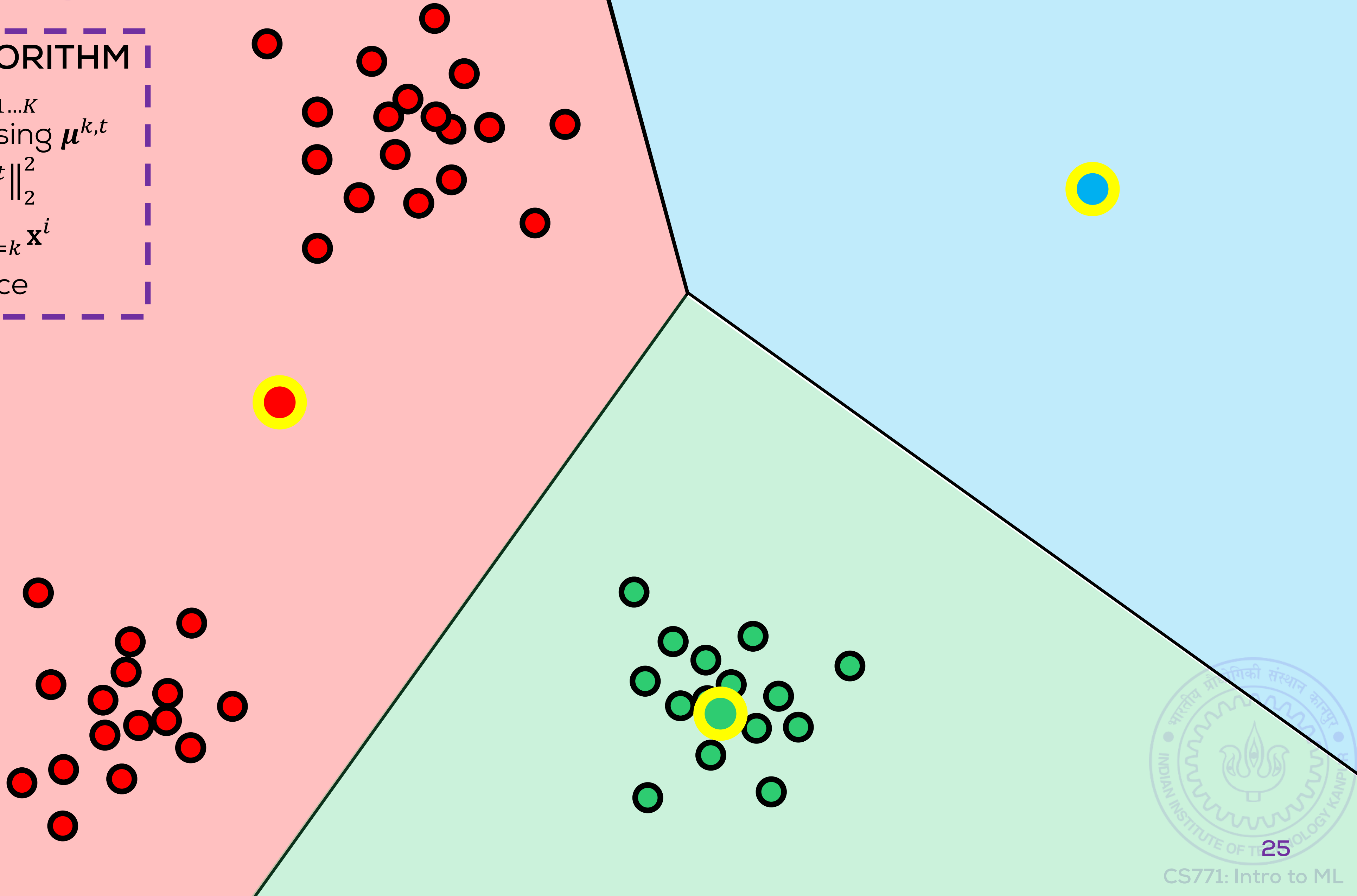
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   Let $z^{i,t} = \arg\min_{k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



Stuck!!!

K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

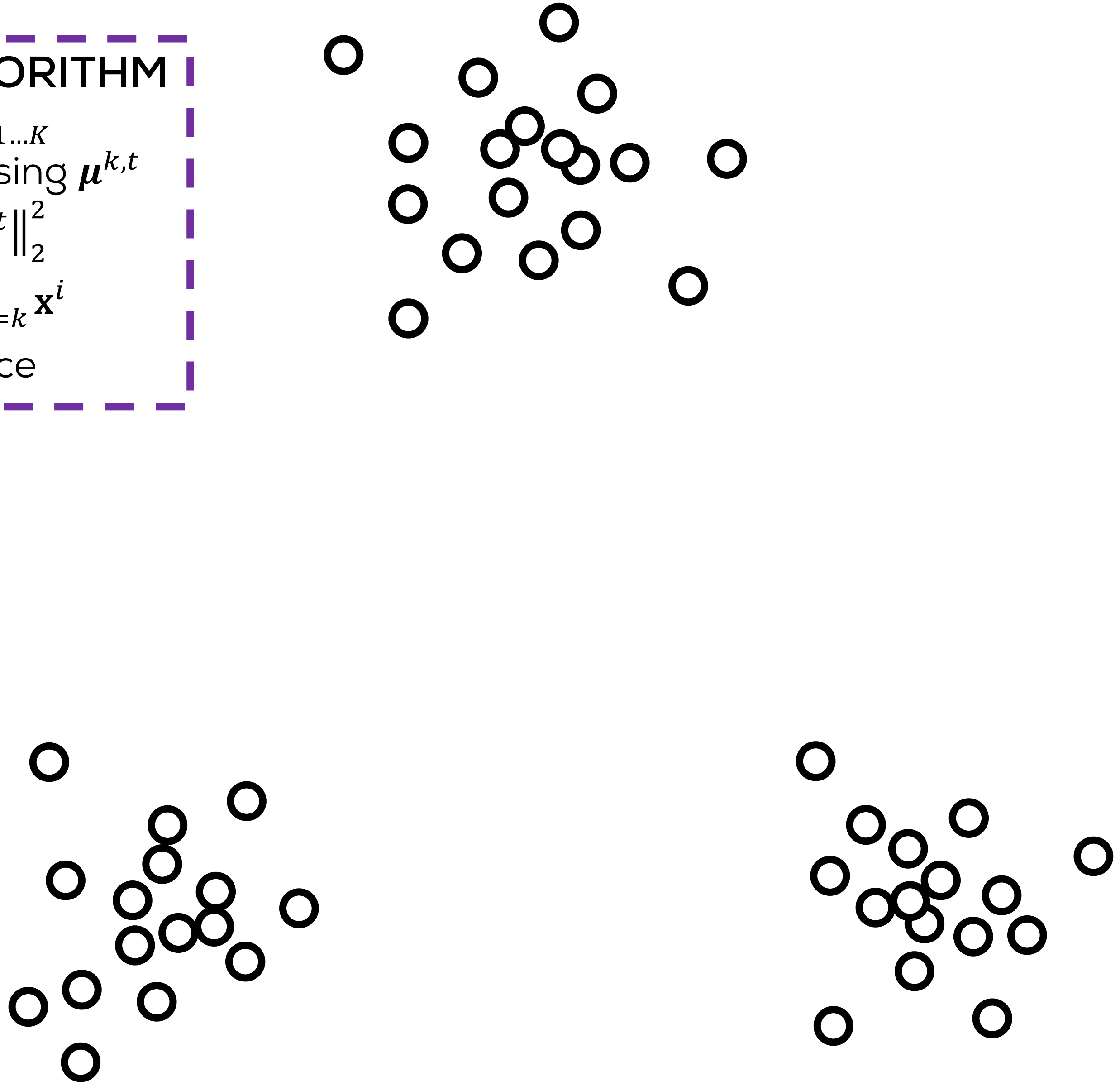
   Let $z^{i,t} = \arg\min_{k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_{k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

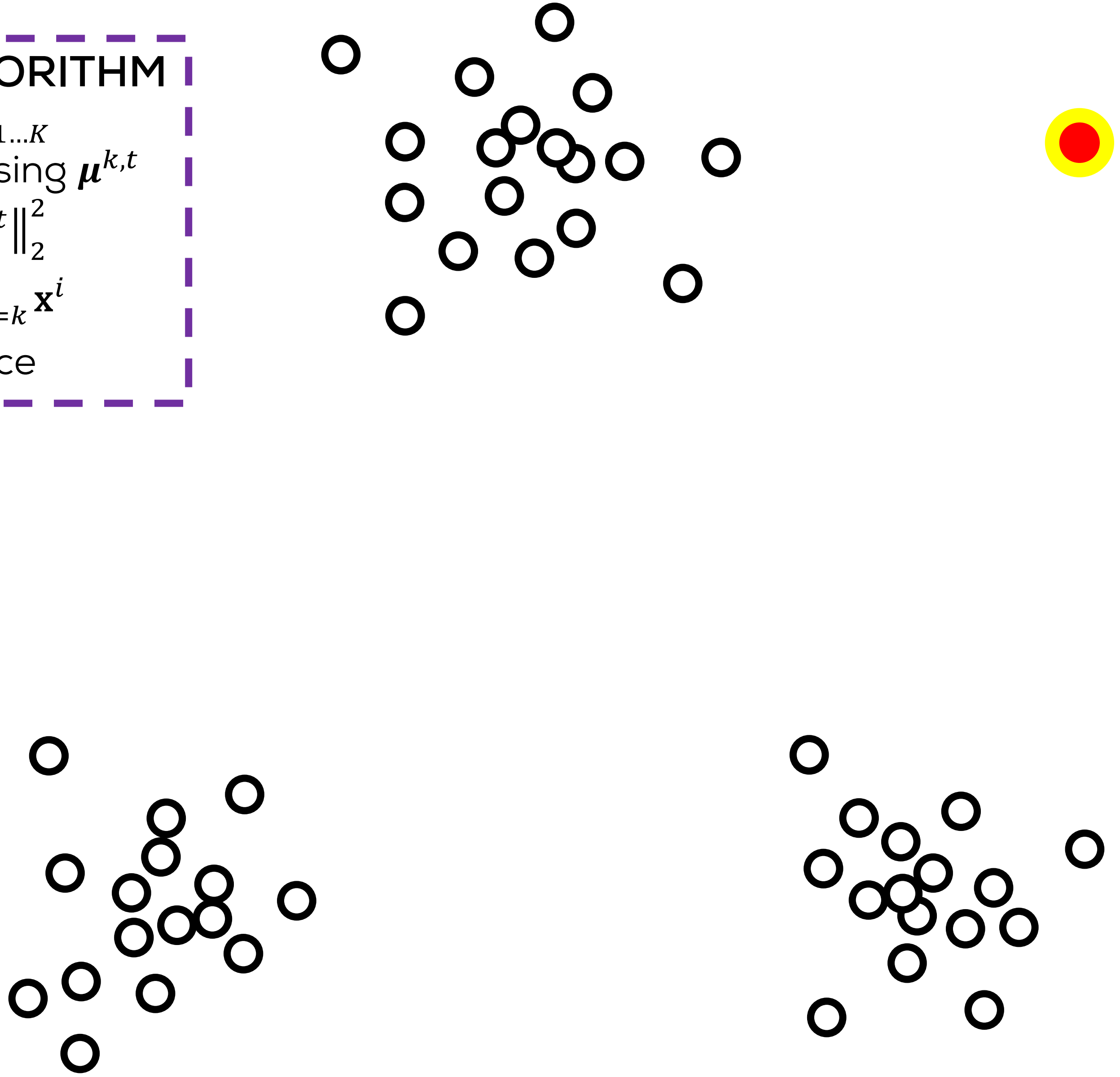


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i: z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

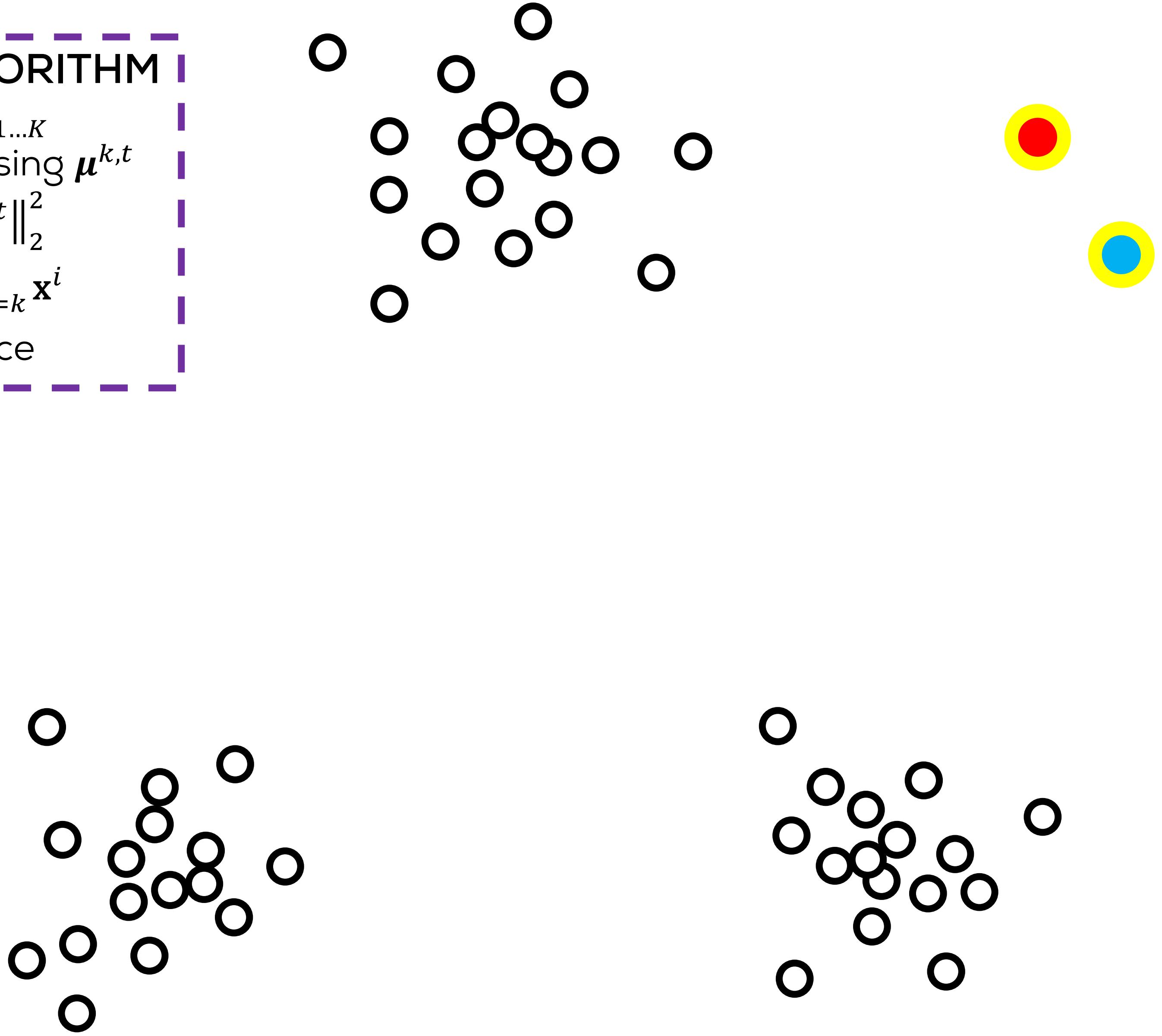


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

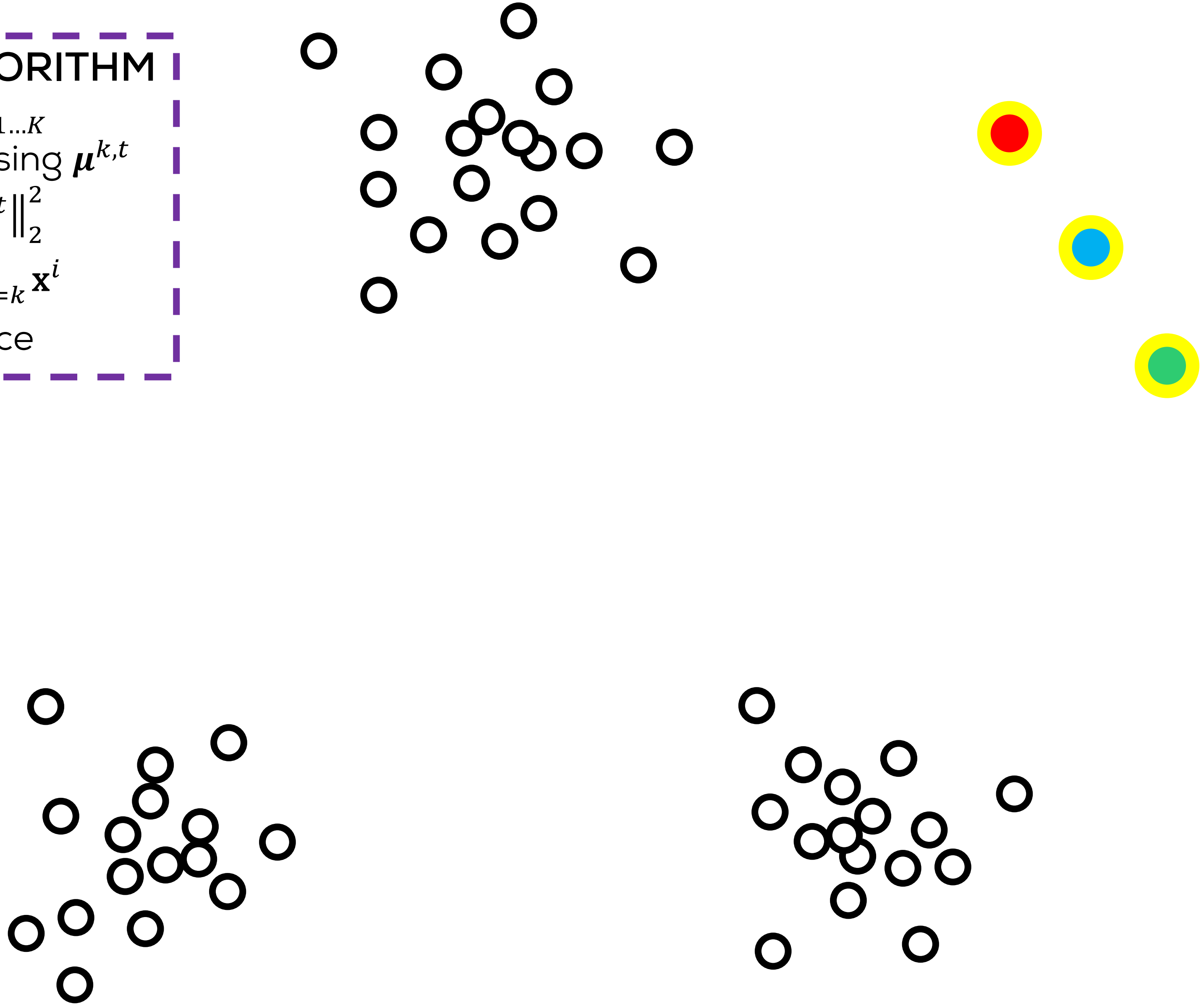


K-Means Algorithm in action!

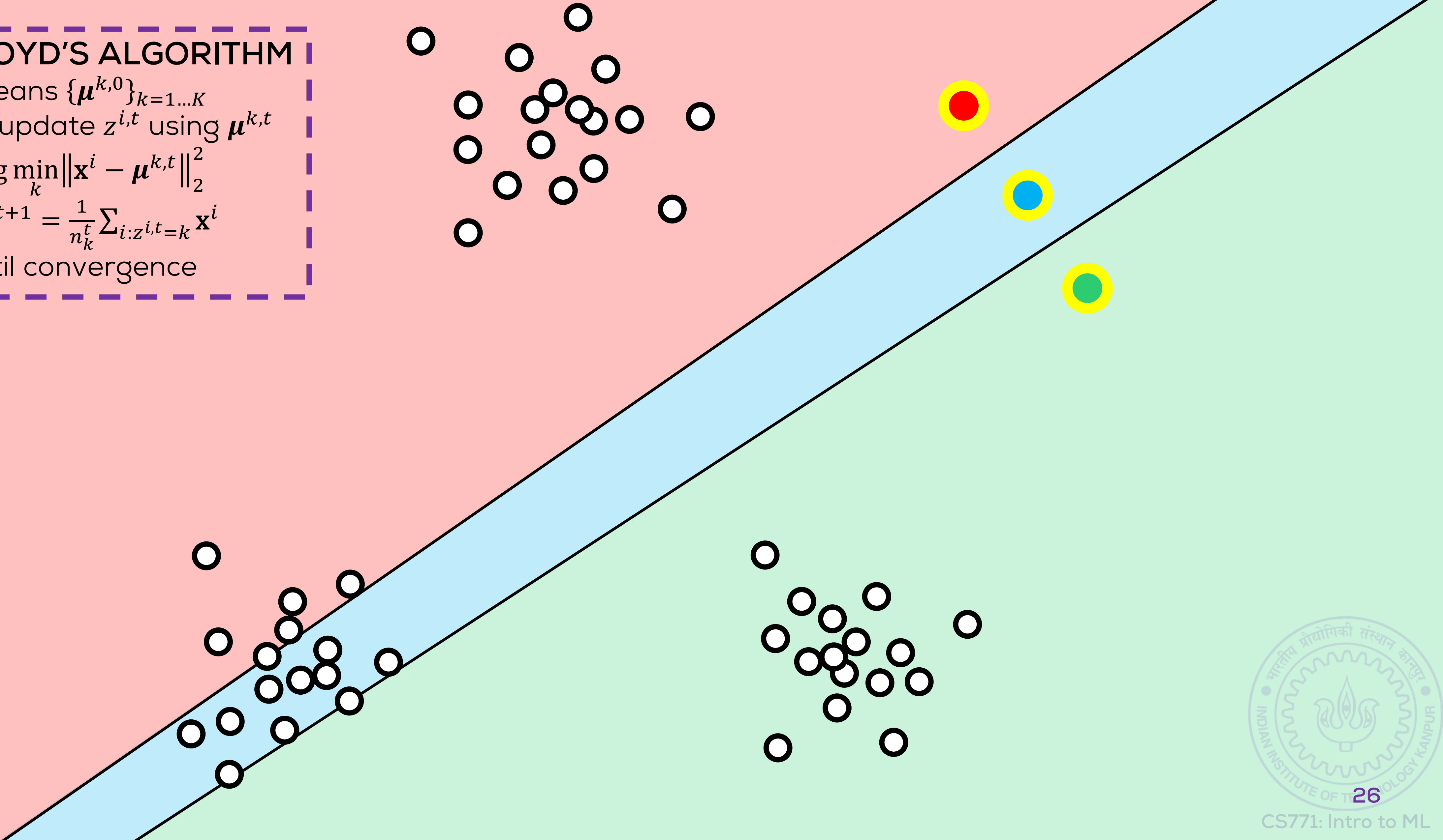
K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



K-Means Algorithm in action!



K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

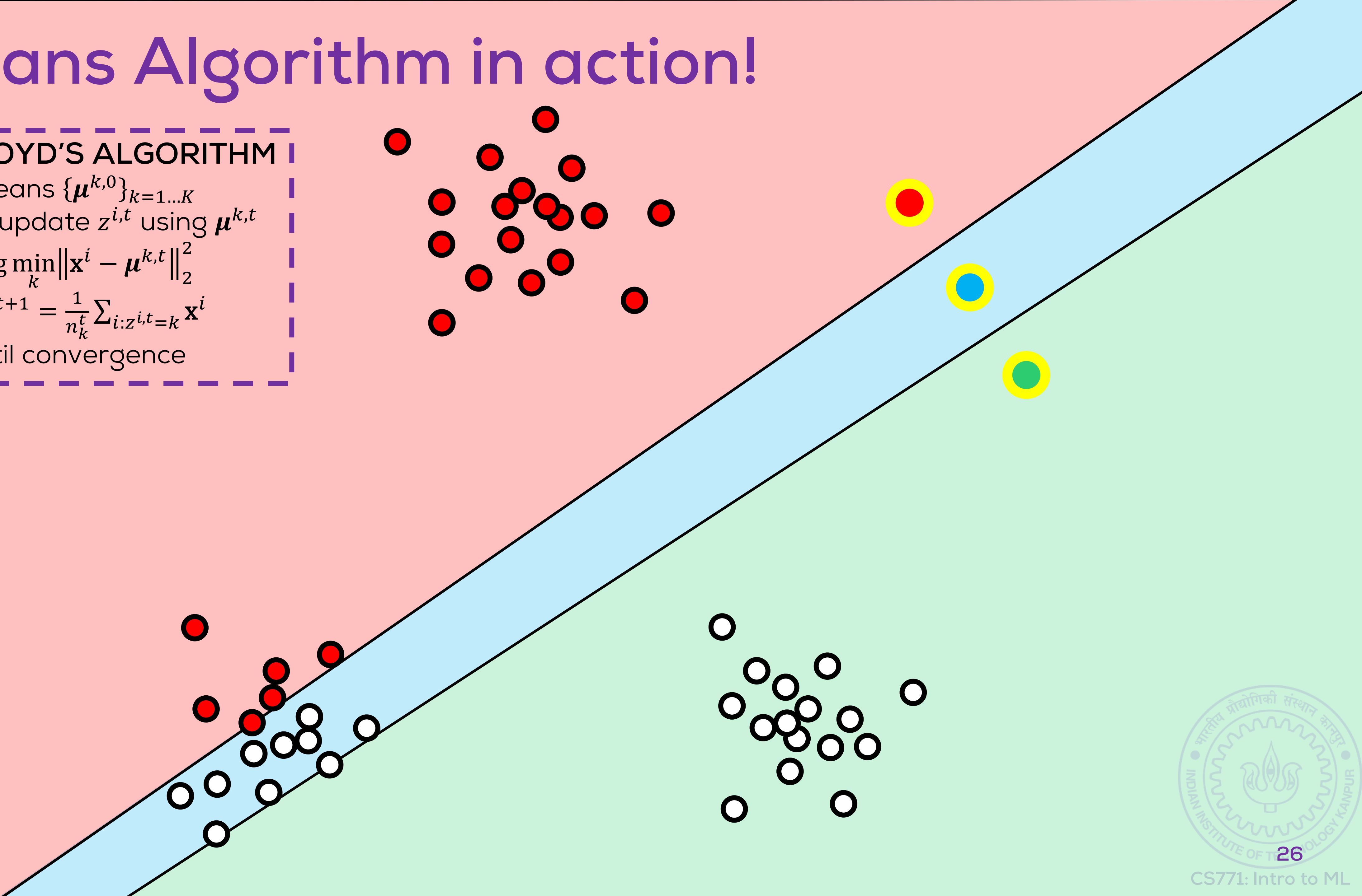
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

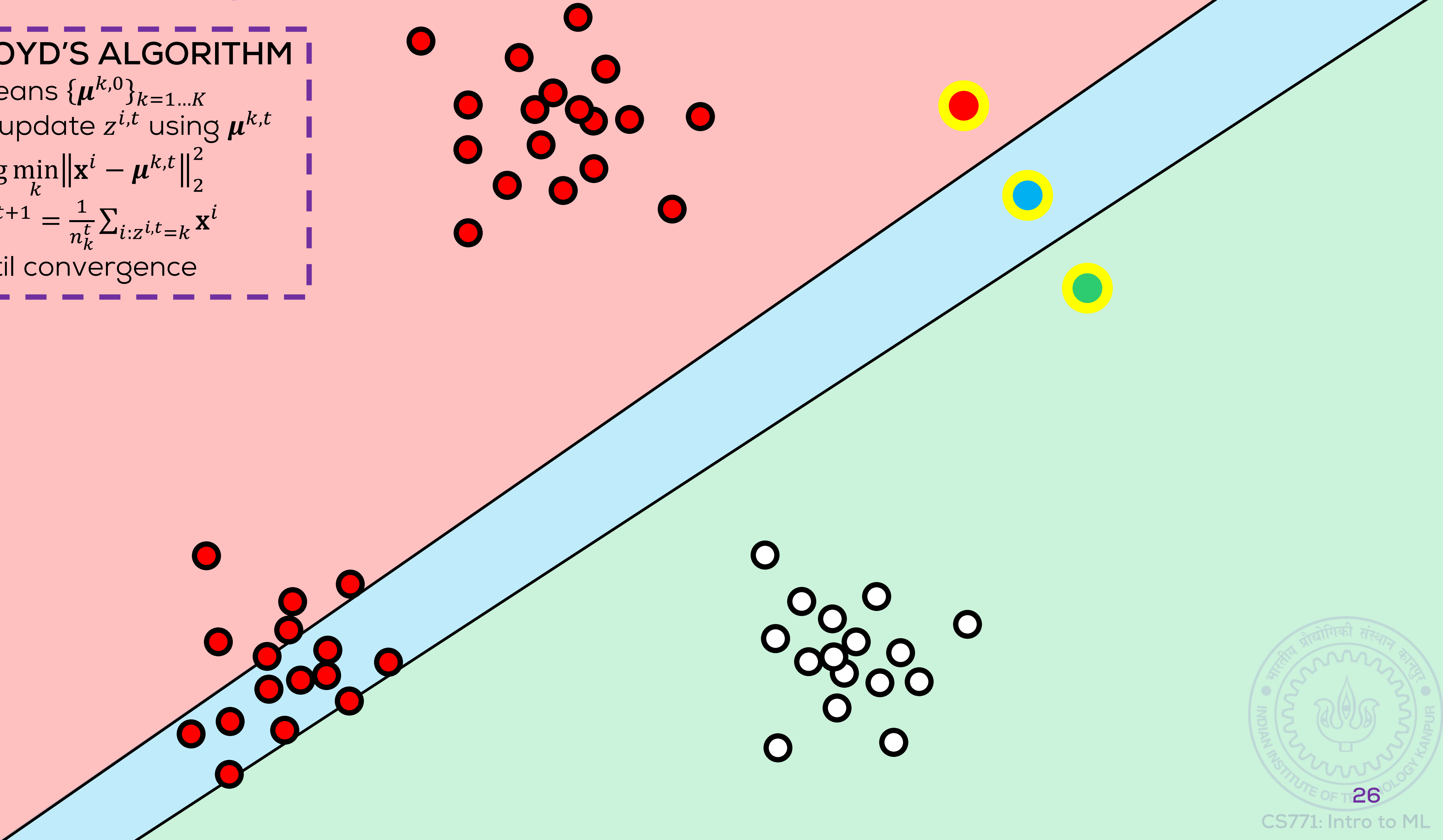


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

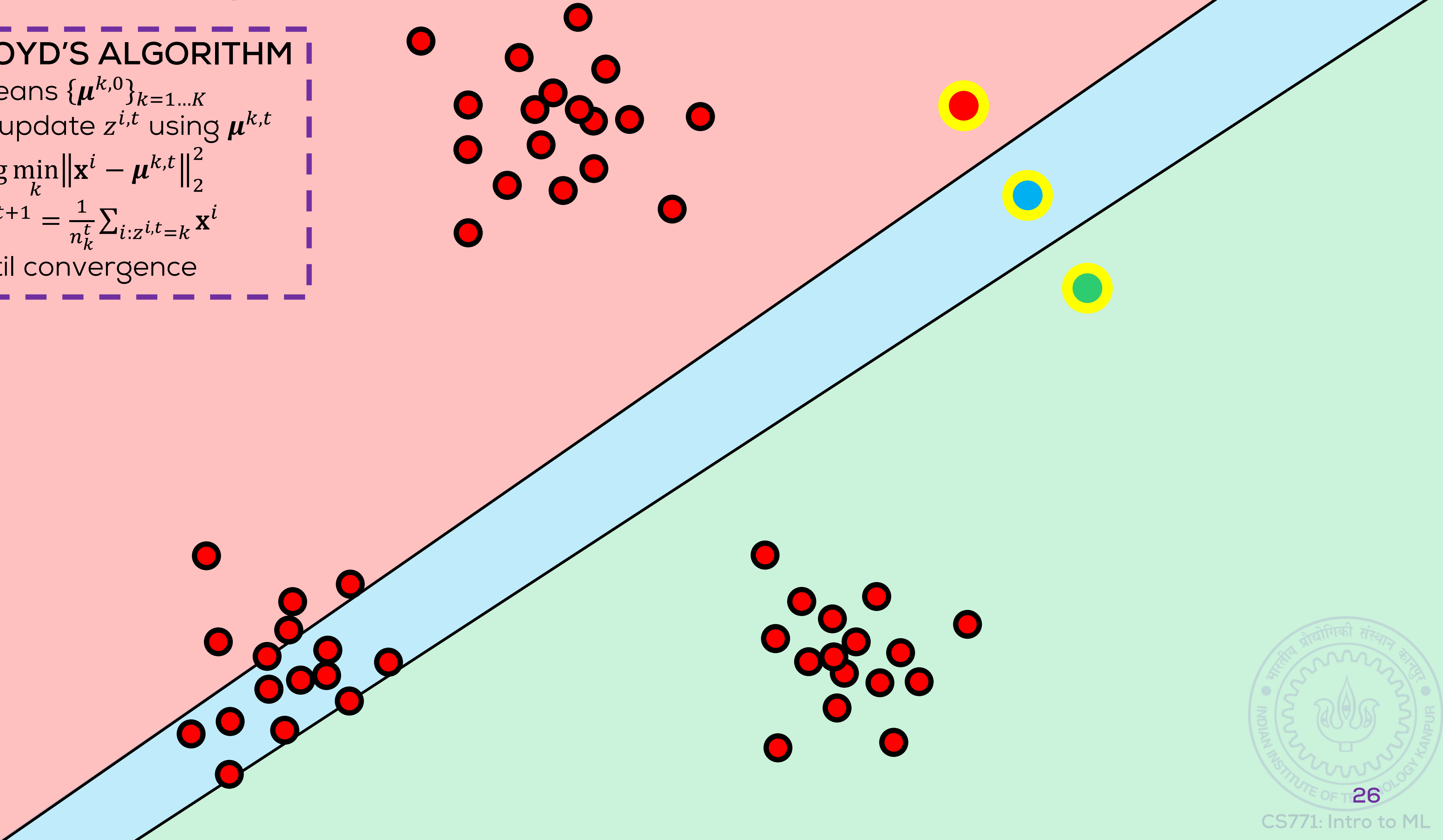


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_{k} \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

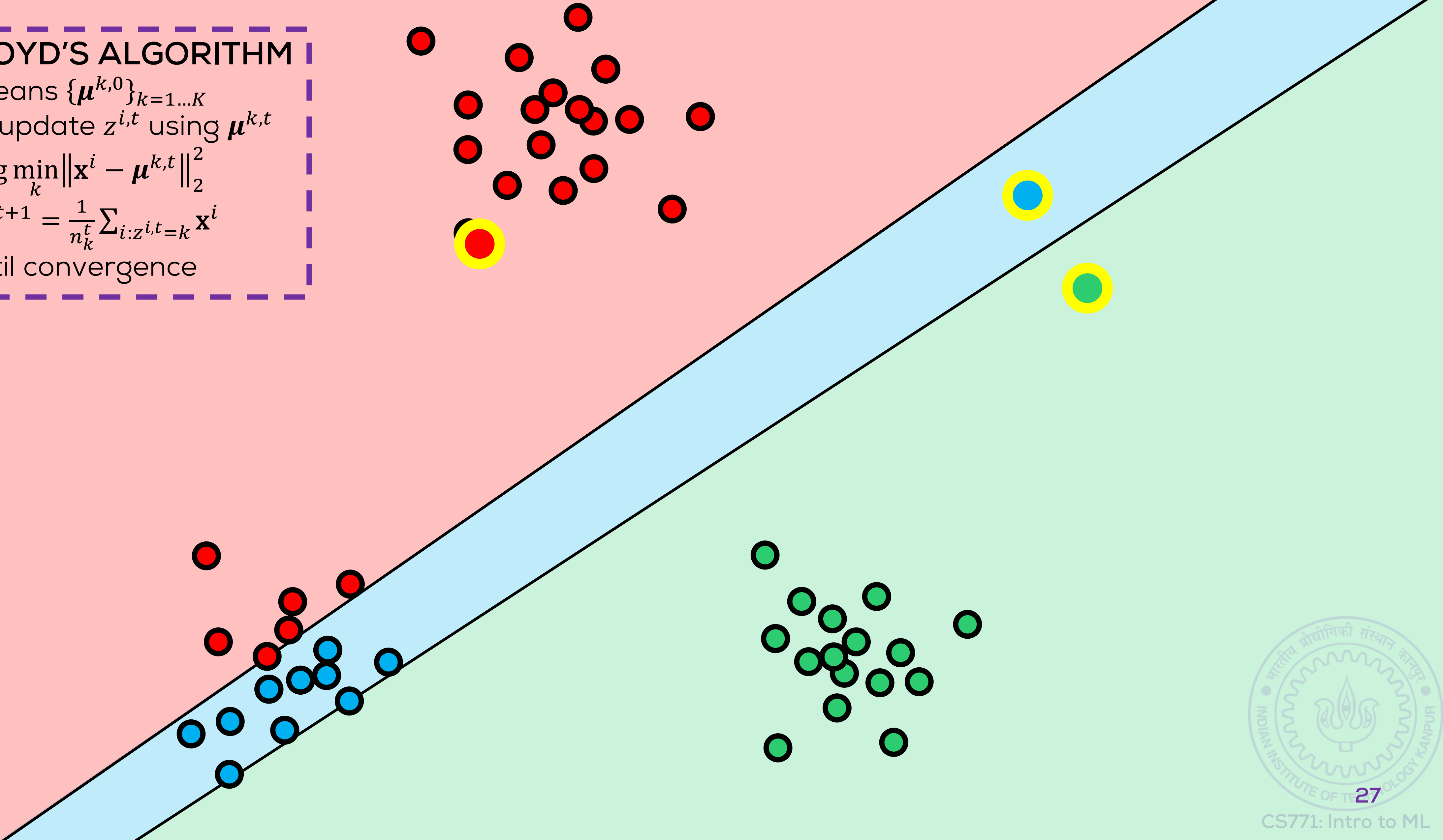


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

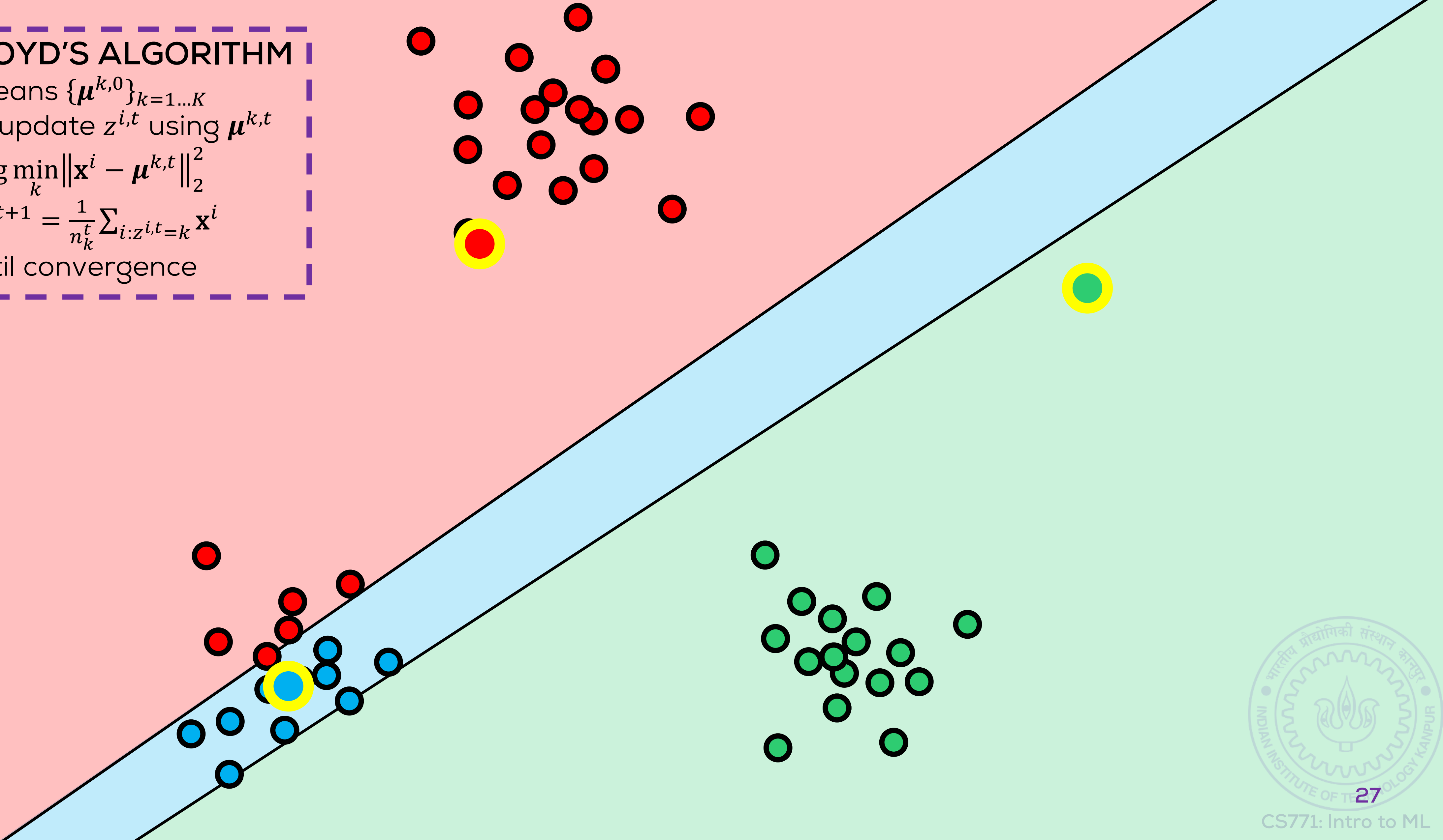


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_{k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

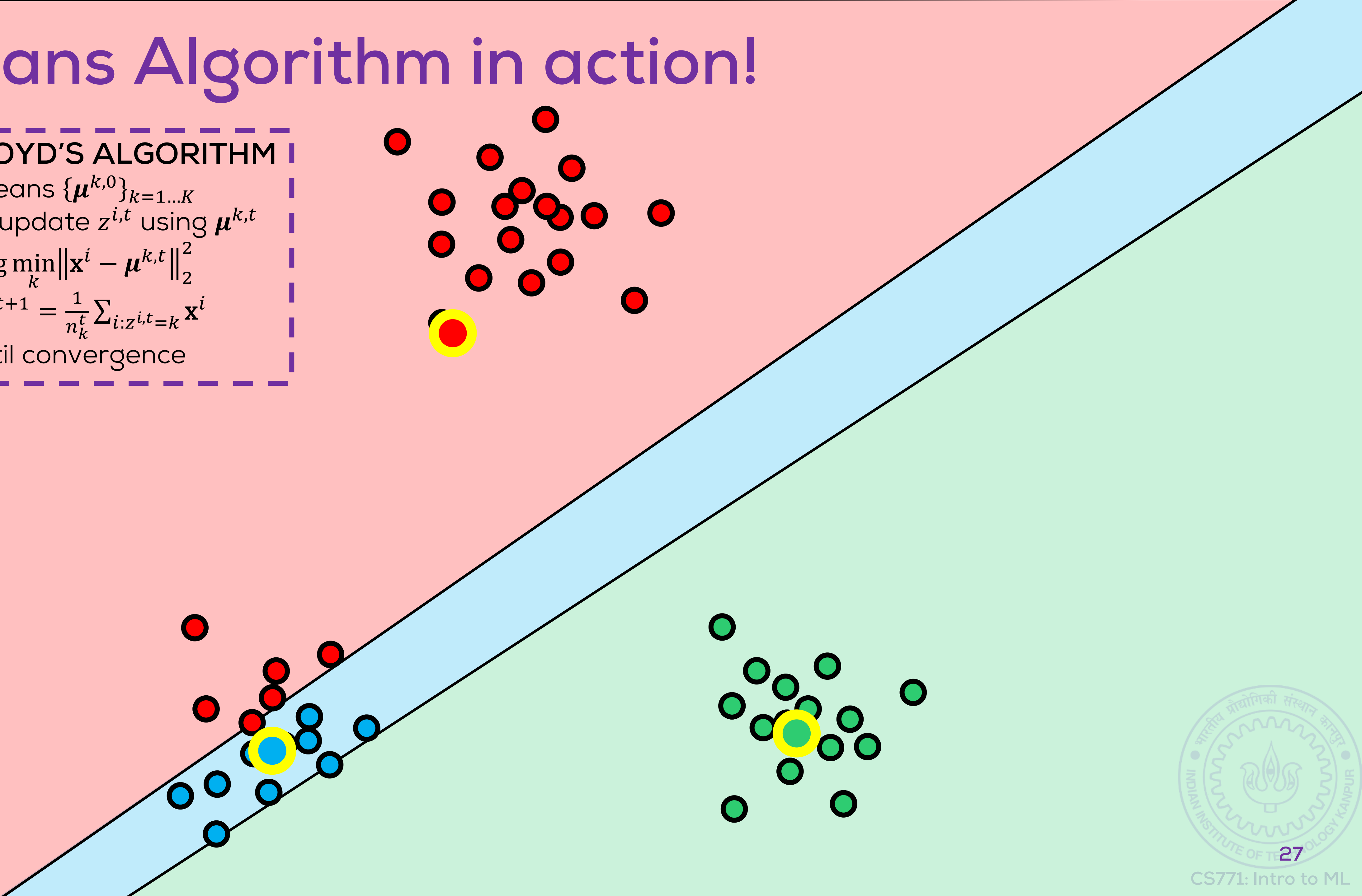


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

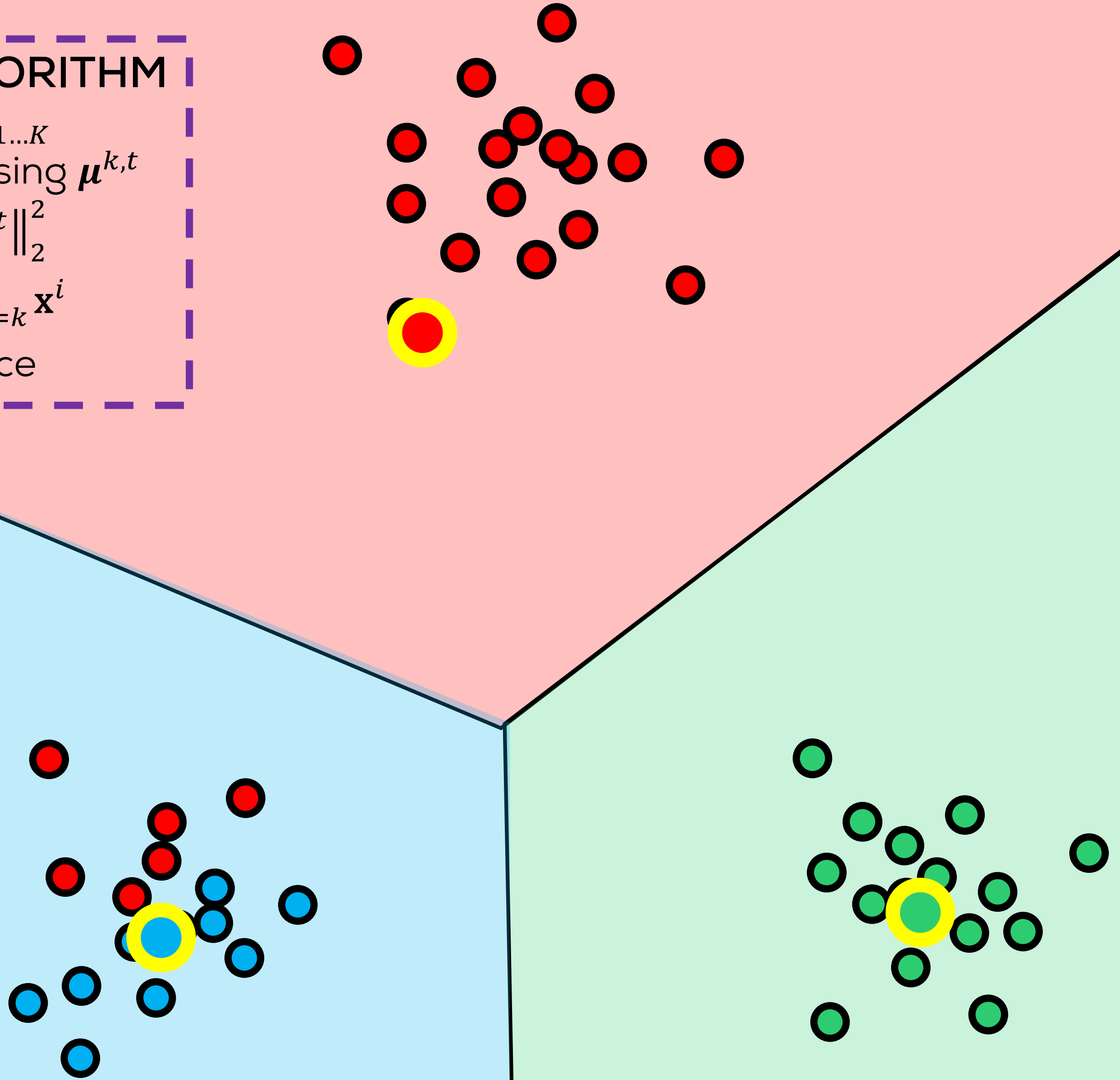


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

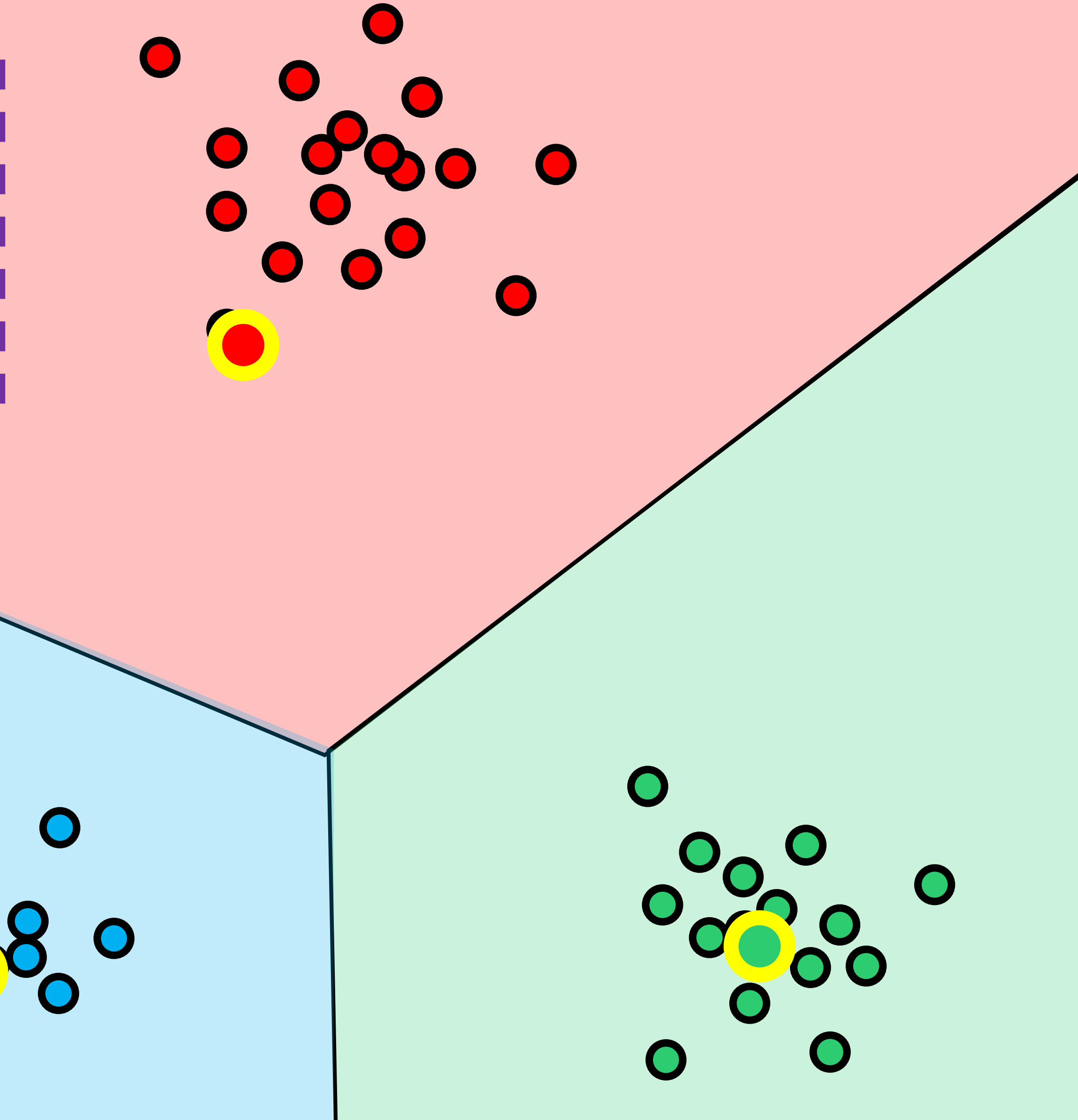


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

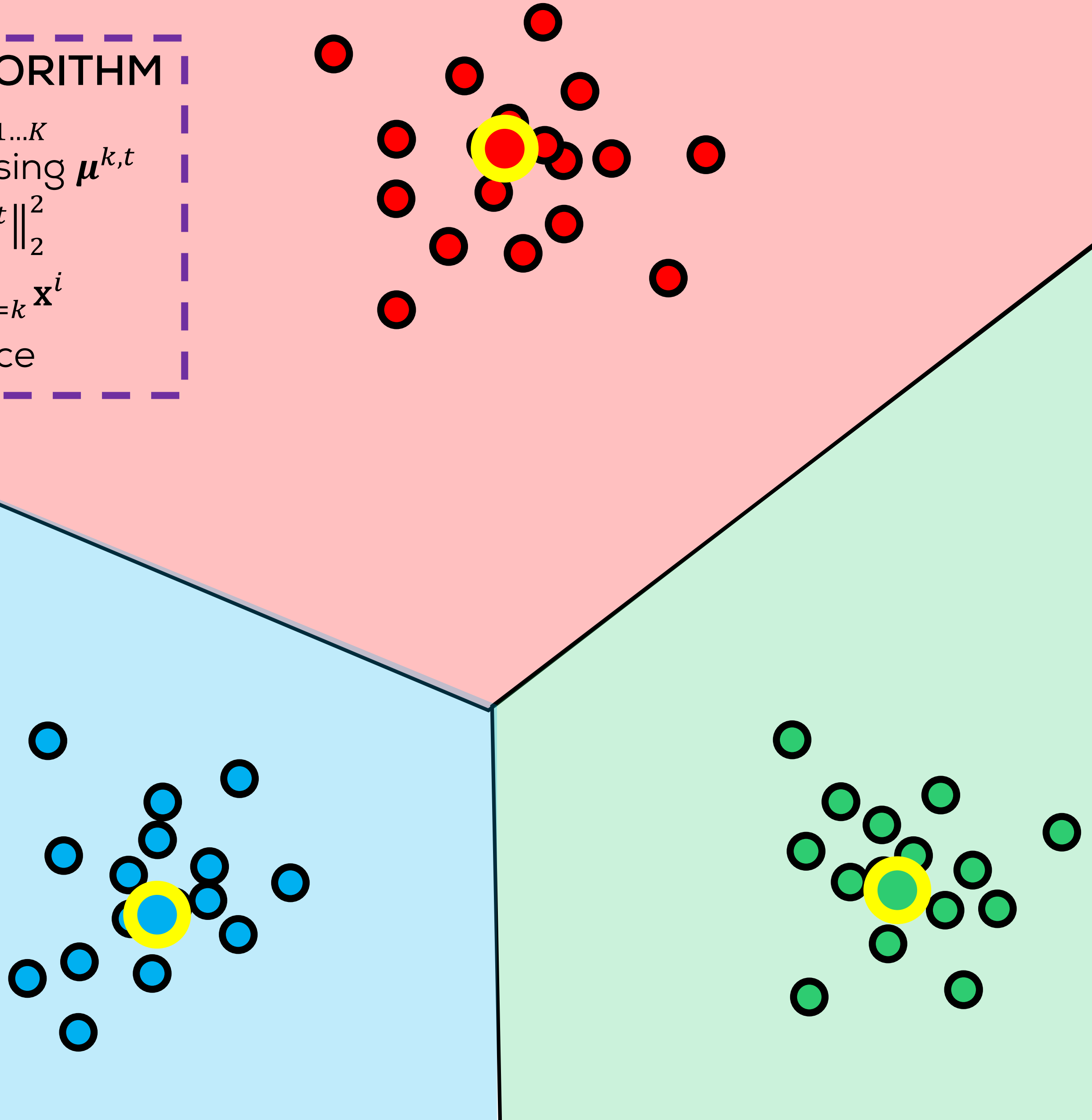
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



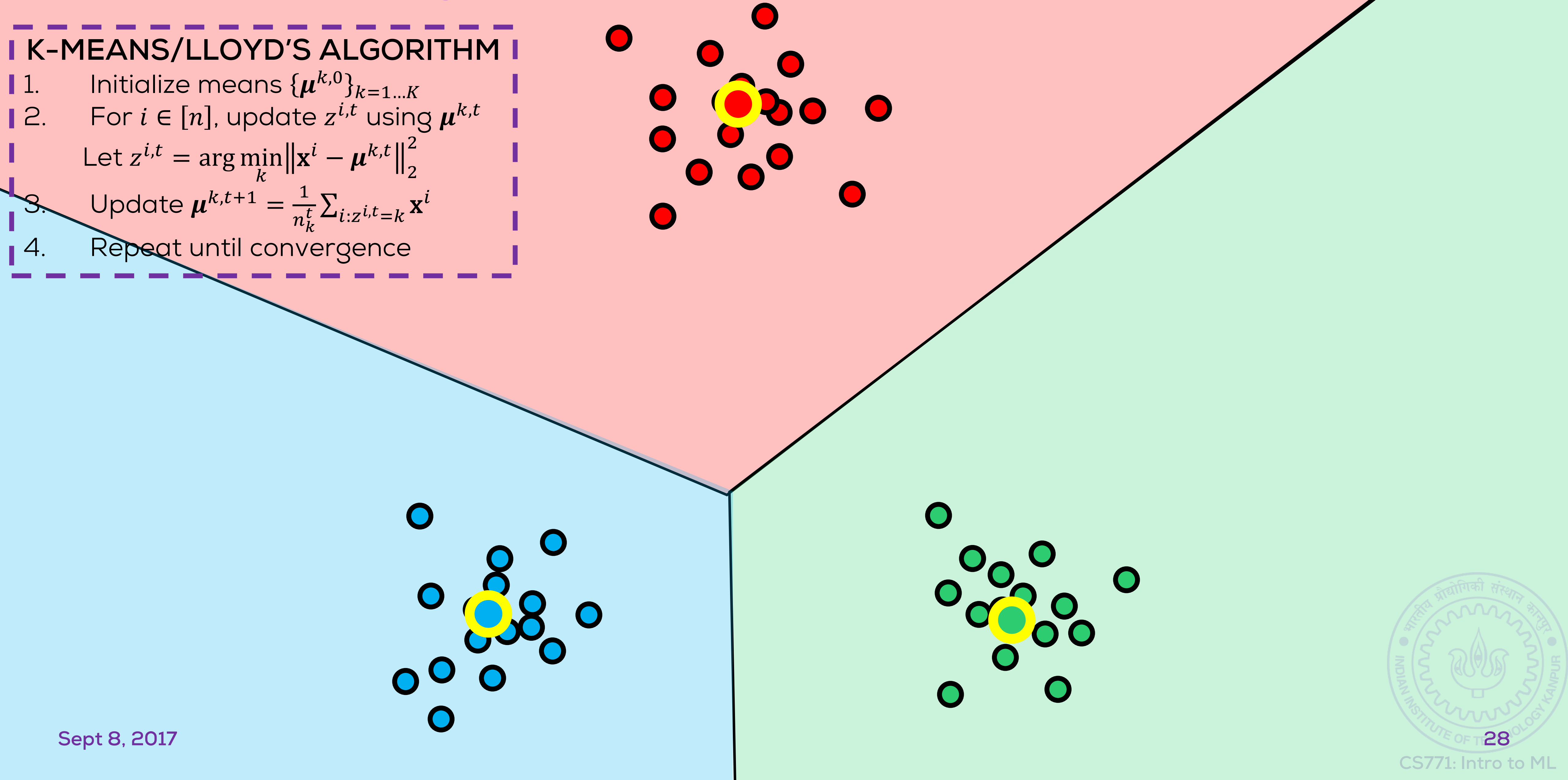
K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



K-Means Algorithm in action!



K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

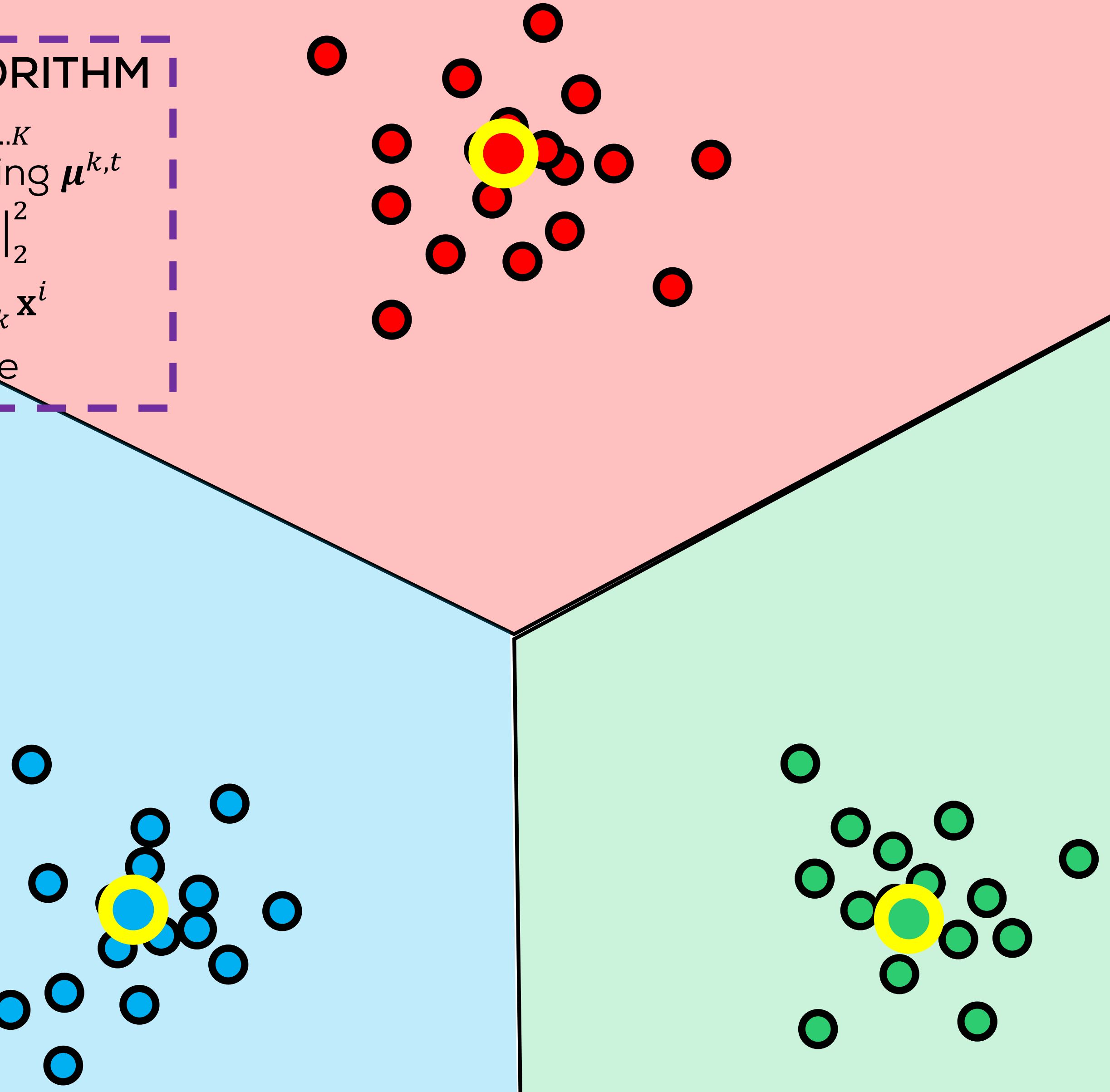
   Let $z^{i,t} = \arg\min_{k} \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

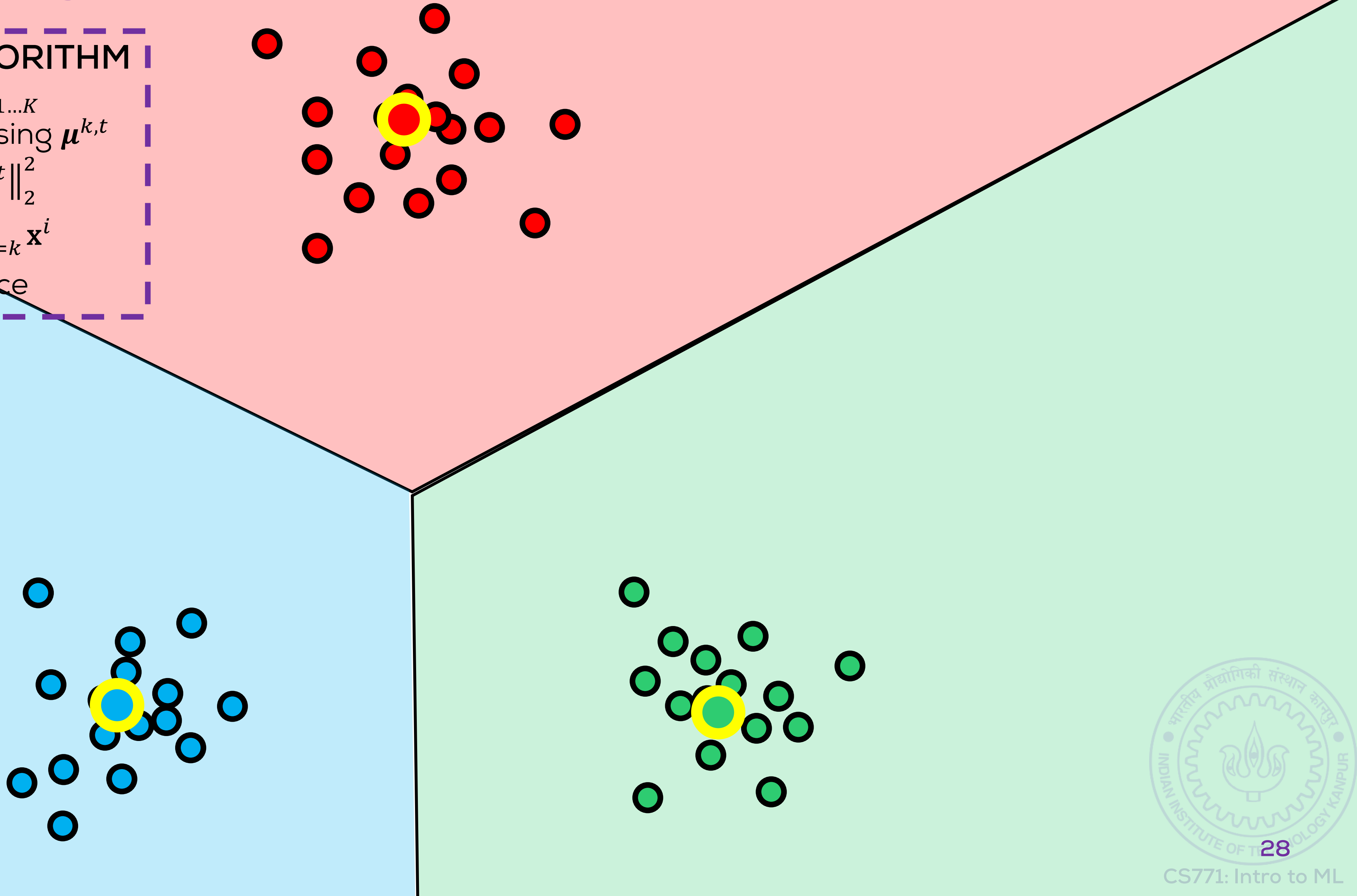


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



Stuck!!! … but at the global optimum ☺

Generating data from GMM learnt using Lloyd

- K clusters with means $\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, \ldots, \boldsymbol{\mu}^K$ learnt using k-means
- To generate a data point from this GMM
  1. Select a cluster $k \sim [K]$ uniformly at random
  2. Select a point from the Gaussian $\mathcal{N}(\boldsymbol{\mu}^k, I)$

K-means uses $\boldsymbol{\pi}_k = \frac{1}{K}$

K-means use $\boldsymbol{\Sigma}^k = I$

The K-means clustering algorithm

The K-means clustering algorithm

- Extremely popular

The K-means clustering algorithm

- Extremely popular
- Helps make sense of data

The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)

The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST

The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)

The K-means clustering algorithm

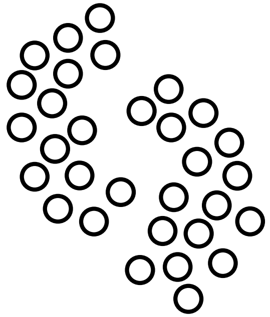
- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters

The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters

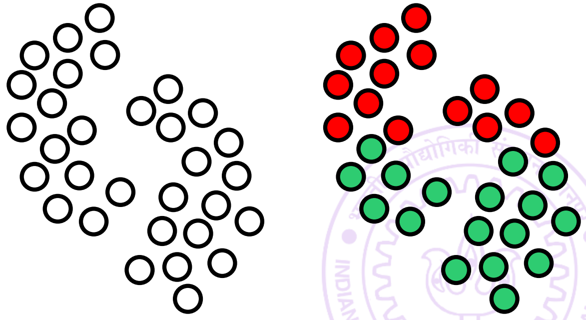
The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters



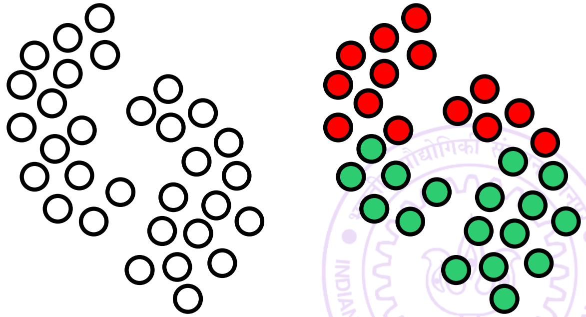
The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters



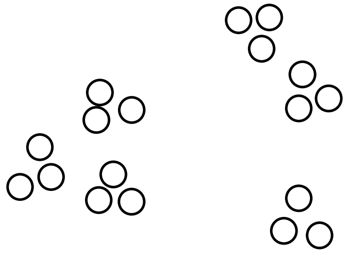
The K-means clustering algorithm

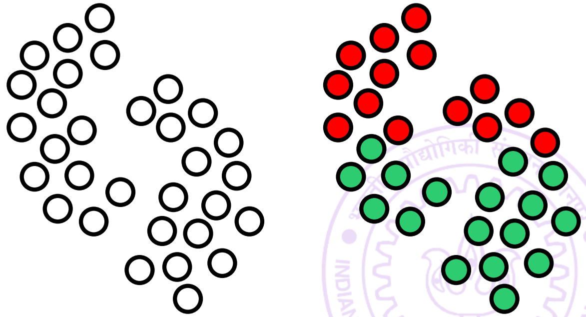
- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters
- Hierarchical clustering



The K-means clustering algorithm

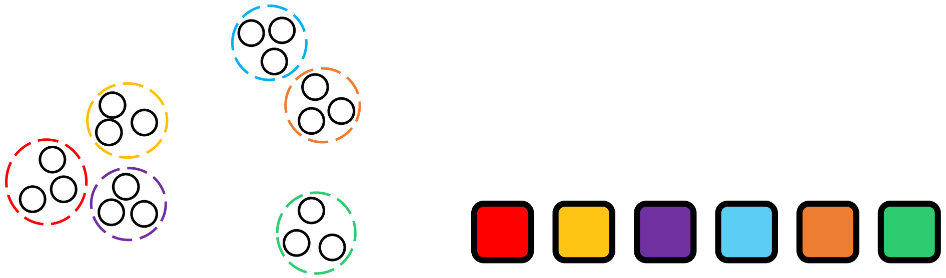
- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters
- Hierarchical clustering

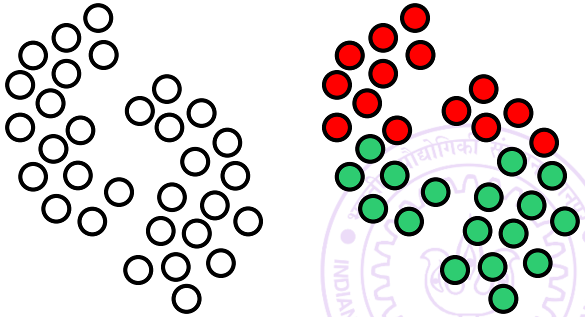




The K-means clustering algorithm

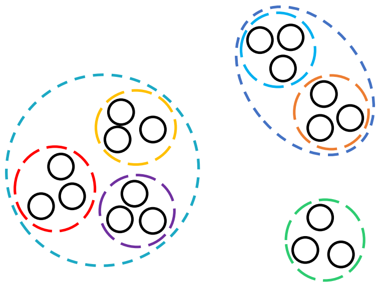
- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters
- Hierarchical clustering

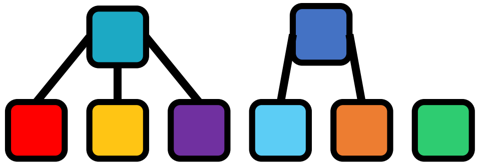


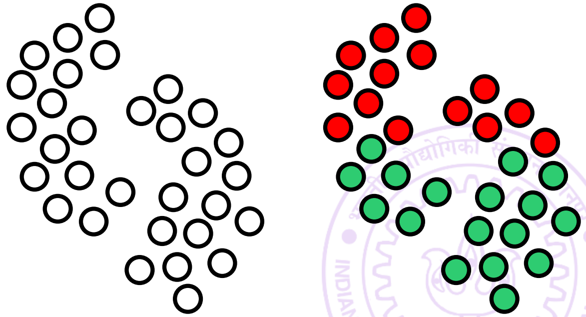


The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters
- Hierarchical clustering

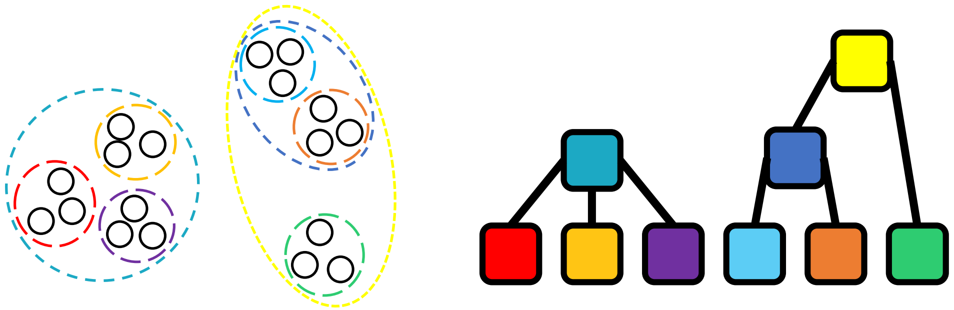


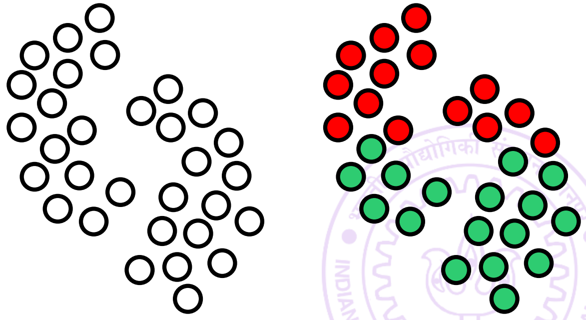




The K-means clustering algorithm

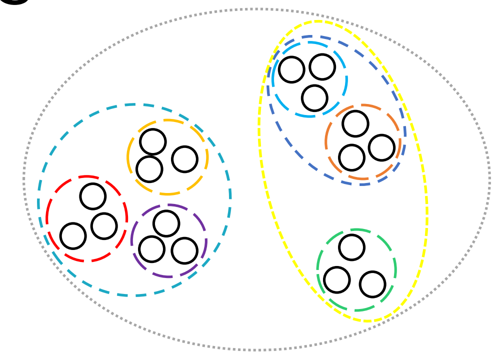
- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters
- Hierarchical clustering

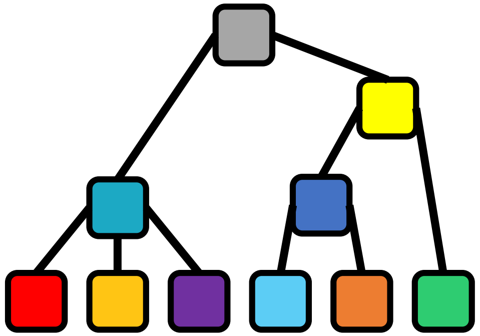


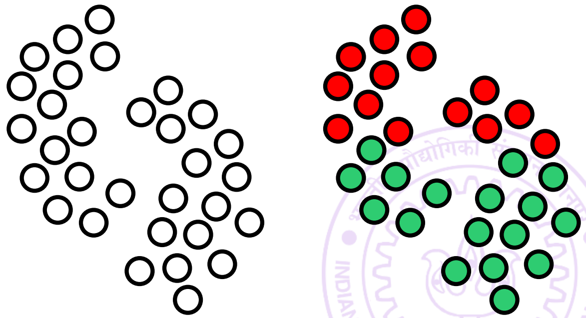


The K-means clustering algorithm

- Extremely popular
- Helps make sense of data
- Can speed up k-NN computation quite a bit (assignment ☺)
- Submission deadline: September 10, 2017, 2359 hrs, IST
- Works well if all clusters are balanced (comparable)
- Brittle in high dimensions and if many many small clusters
- Learns clusters that are ball shaped, does not do well on non-convex shaped clusters
- Hierarchical clustering







Please give your Feedback

http://tinyurl.com/ml17-18afb