Data Modelling Methods-III

CS771: Introduction to Machine Learning

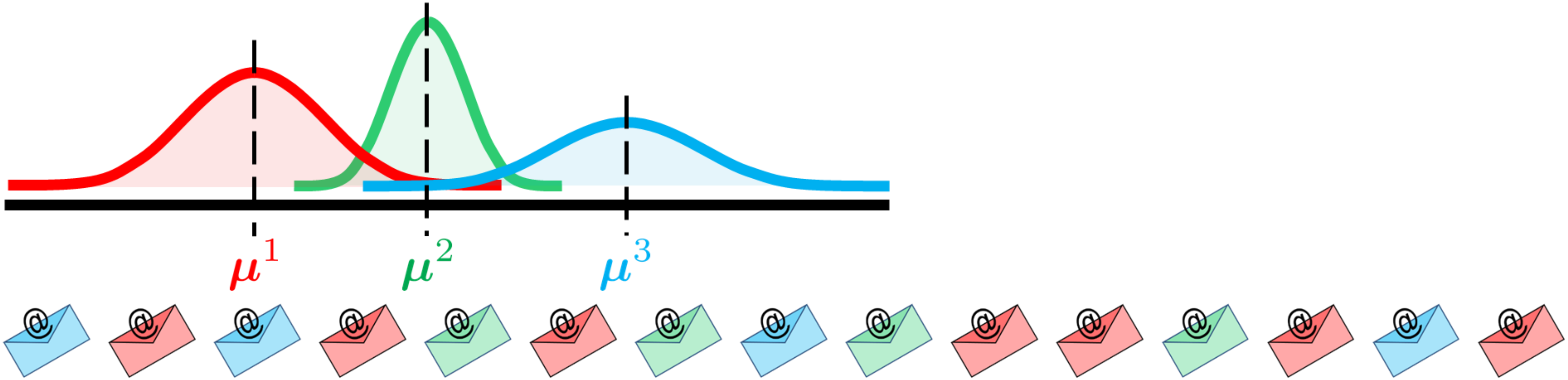
Purushottam Kar

Mid Semester Examination

- September 21st, 2017 (Thursday) 1300-1500 hrs
- Venue L18, 19, L20 (all OROS)
- Syllabus: till whatever we cover today
- Open notes (handwritten only)
- No printed/photocopied material
- No laptops, i-pads, mobile phones (switched off)
- Please bring a notepad with you for rough work
- Please bring a pencil/eraser with you – we will not provide these
- Answers will have to be written on the question paper itself

Recap

The generative story for labelled data



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

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

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

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

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

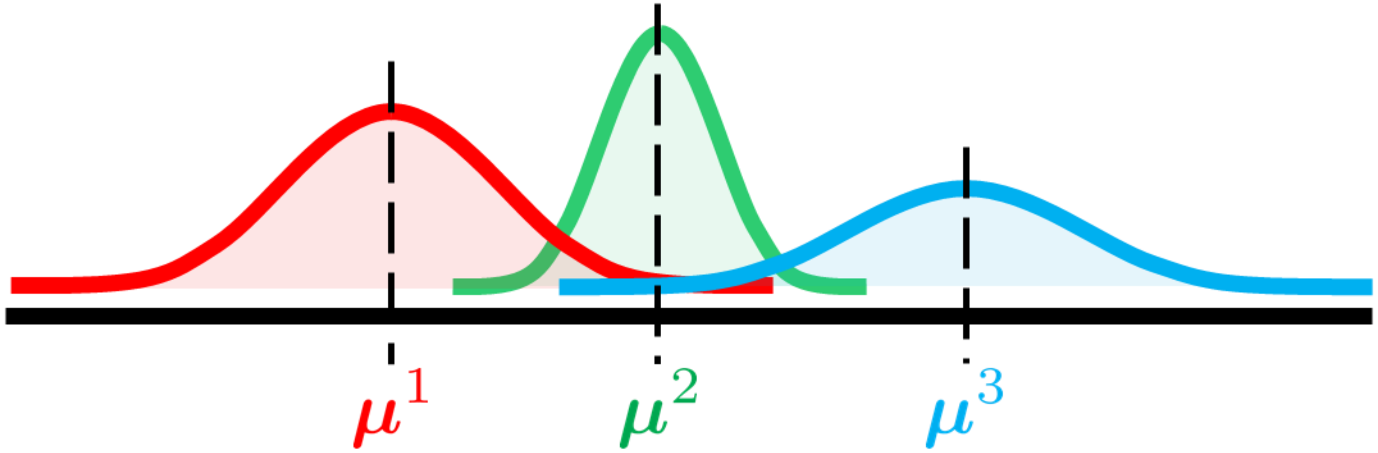
$$\Sigma^1_{\text{MLE}} = \frac{1}{n_r} \sum_{i:z^i=\bullet} (\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE})(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE})^{\mathsf{T}}$$

Take log and apply 1st order optimality

Read [DAU] Sections 9.1-9.5

Recap

The generative story for unlabelled data



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

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

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

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

Gaussian Mixture Model (GMM) with K components

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

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

Recap

A Ray of Hope

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

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

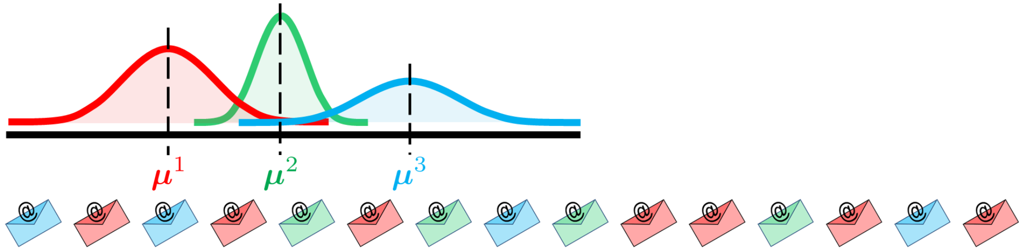
ALTERNATING OPTIMIZATION

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

Recap

The generative story for labelled data



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

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

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

Take log and apply 1st order optimality

Read [DAU] Sections 9.1-9.5

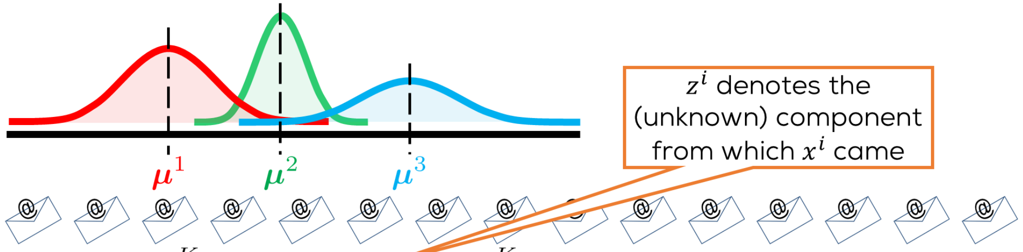
$\boldsymbol{\pi}_1^{\text{MLE}} = \frac{\#\text{ red emails}}{\#\text{ total emails}} = \frac{n_r}{n}$

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

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

12

The generative story for unlabelled data



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

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

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

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

Gaussian Mixture Model (GMM) with K components

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

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

Sept

4

A Ray of Hope

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

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

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

Various ways of updating z^i

Sept 8, 2017

18

Hard Assignment

The K-means algorithm

A Ray of Hope

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

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

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

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

ALTERNATING OPTIMIZATION

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

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

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

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

ALTERNATING OPTIMIZATION

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

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

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

Bayes Rule!

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

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

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 $\mathrm{n}_k^t = \left|\left\{i\colon z^{i,t} = k\right\}\right|$
   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

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

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

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

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

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

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++ 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/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 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 $\boldsymbol{\mu}^{k,t}$

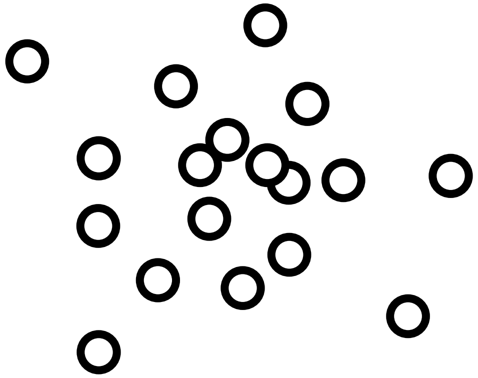
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k}\mathbf{x}^i$
4. Repeat until convergence

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 $\boldsymbol{\mu}^{k,t}$

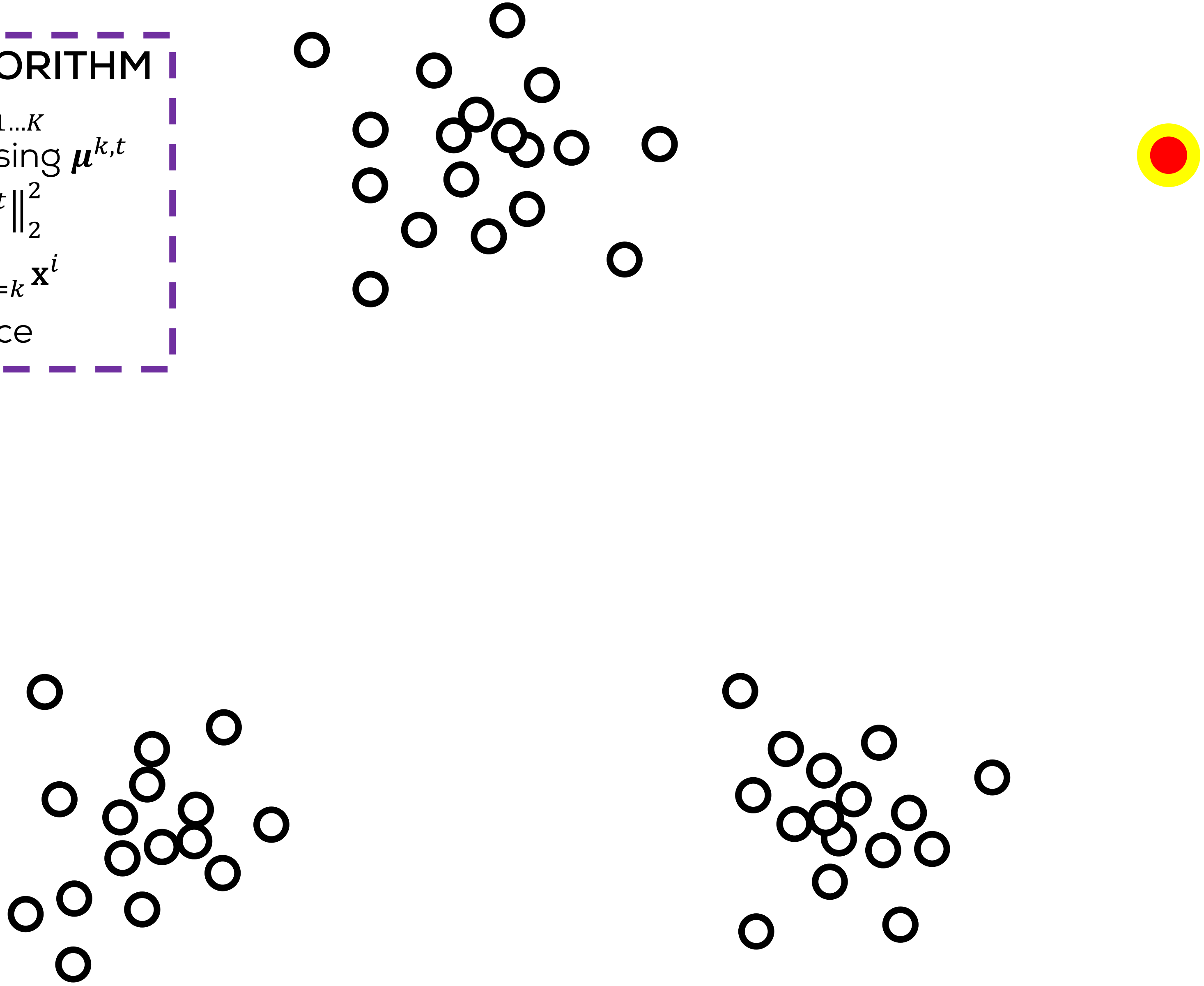
   Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i : z^{i,t} = k} \mathbf{x}^i$
4. Repeat until convergence



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

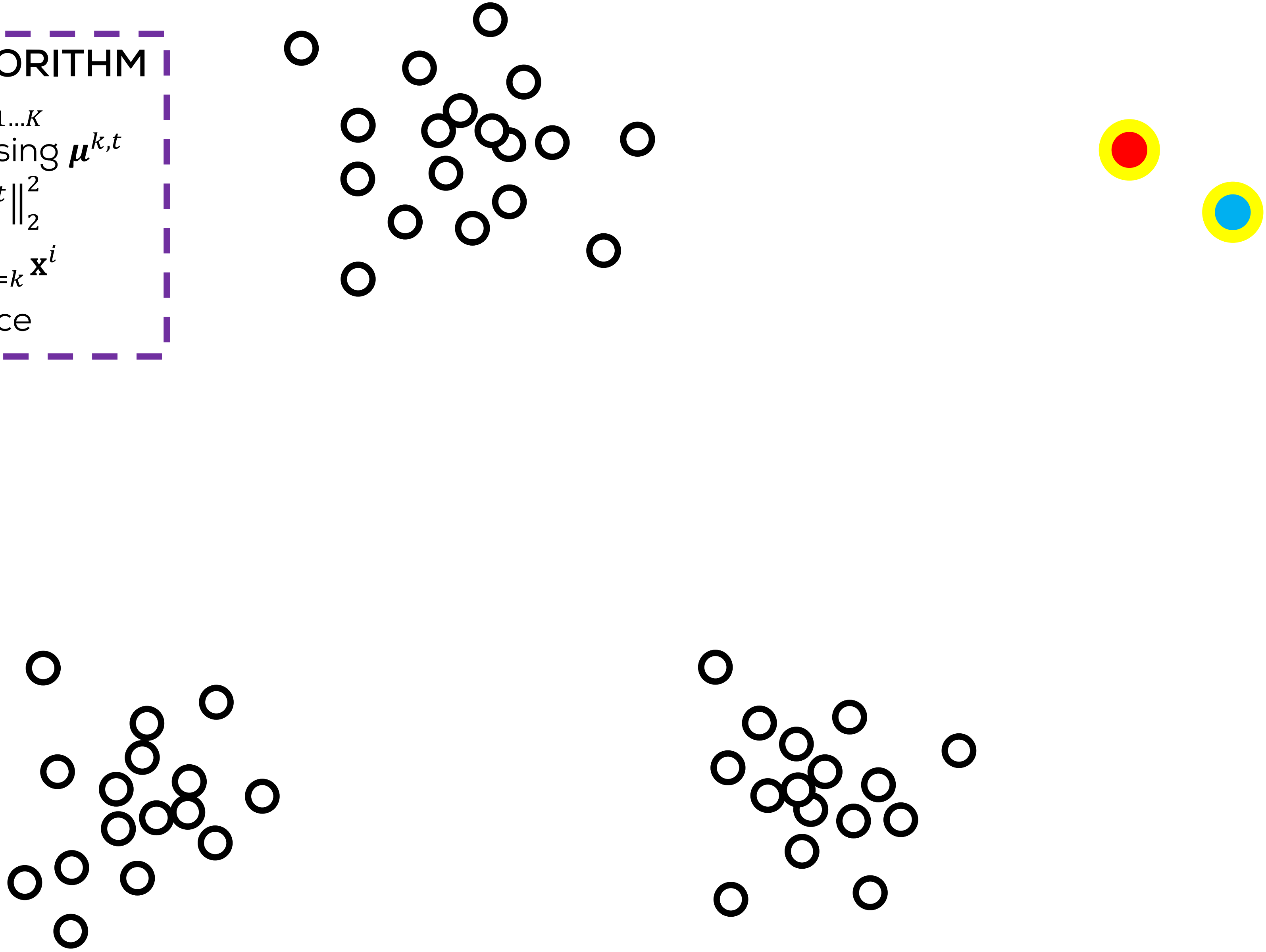


K-Means Algorithm in action!

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

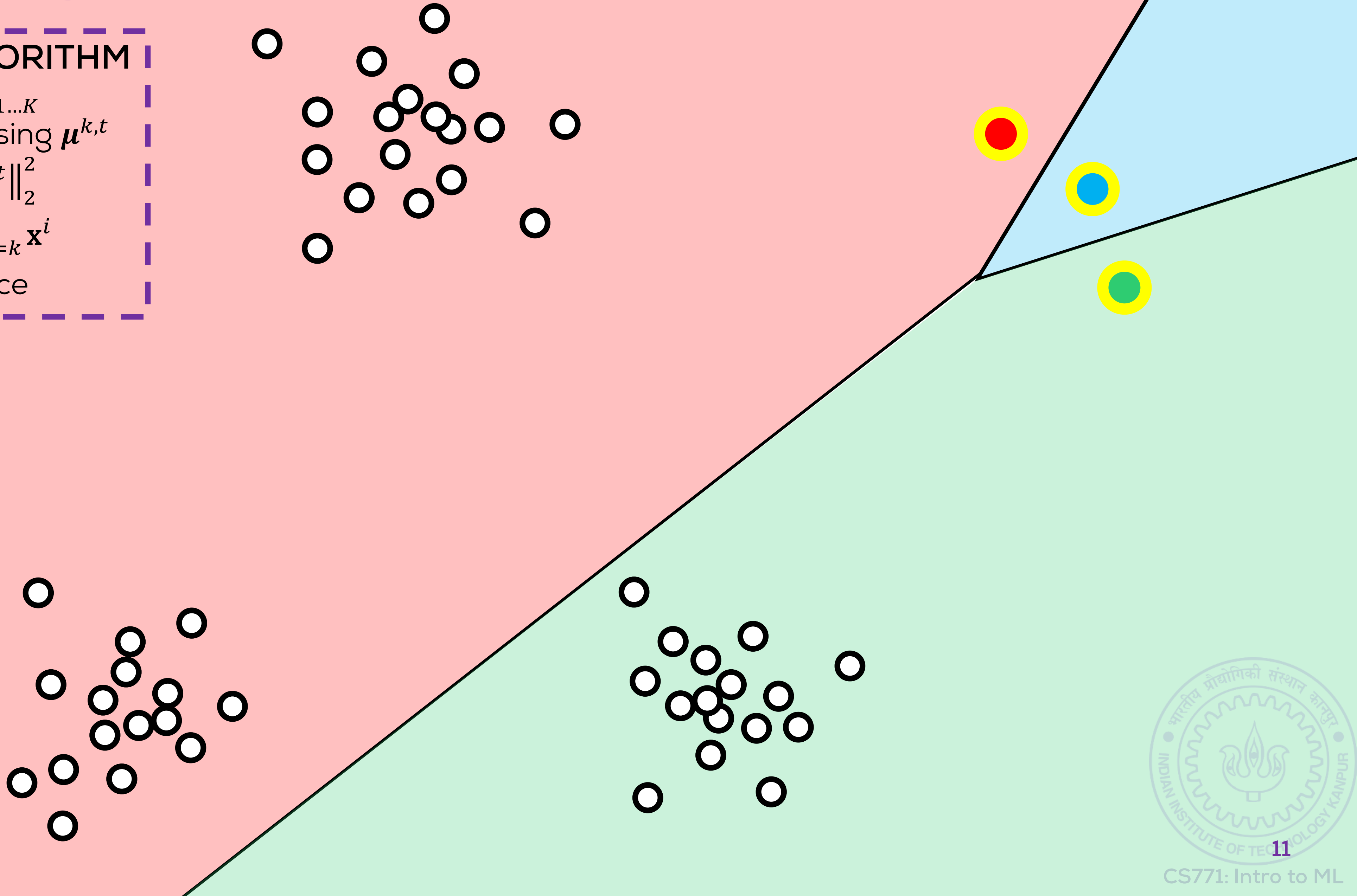
1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1\ldots K}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\mu}^{k,t}$

   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

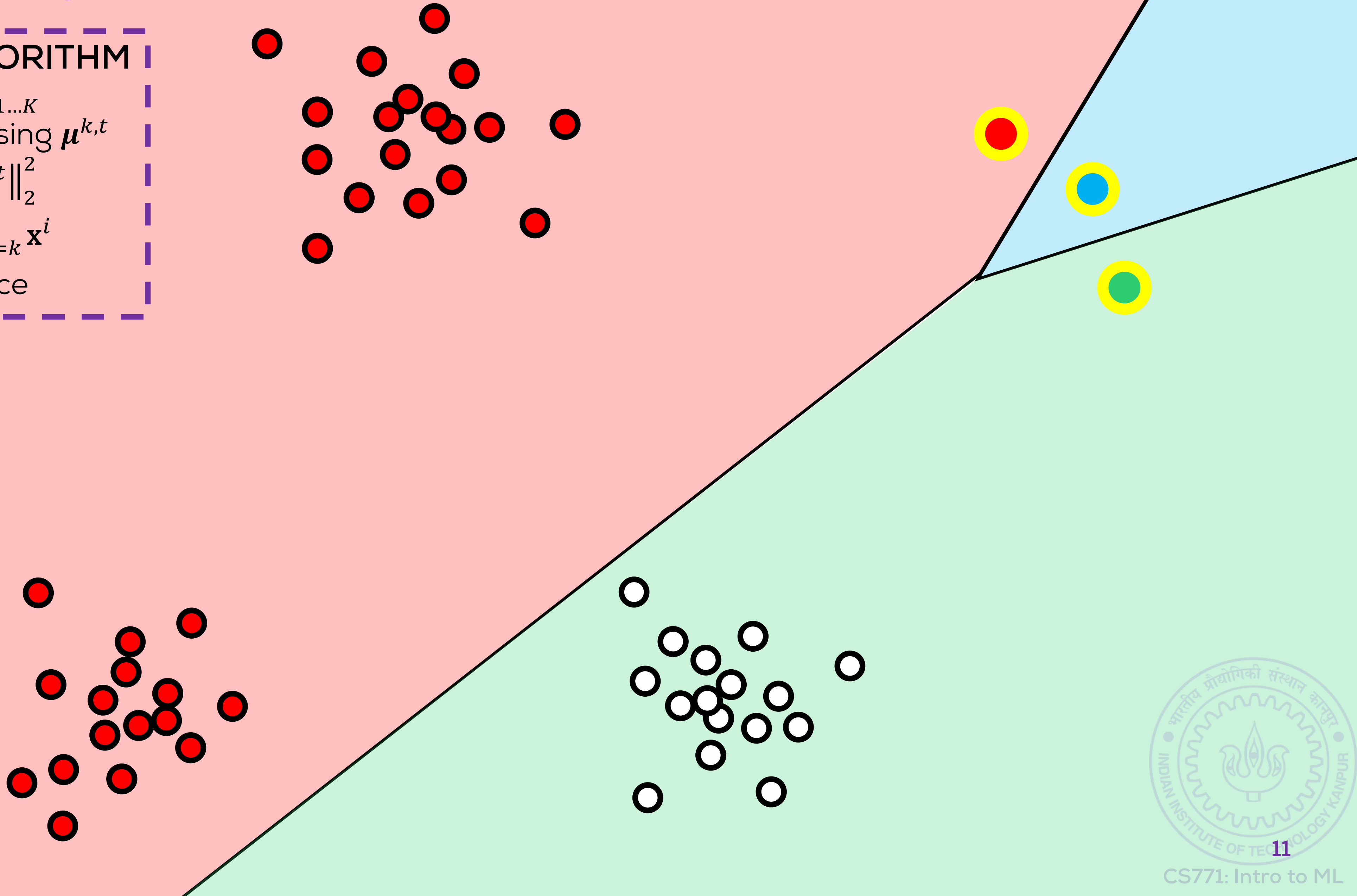


K-Means Algorithm in action!

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

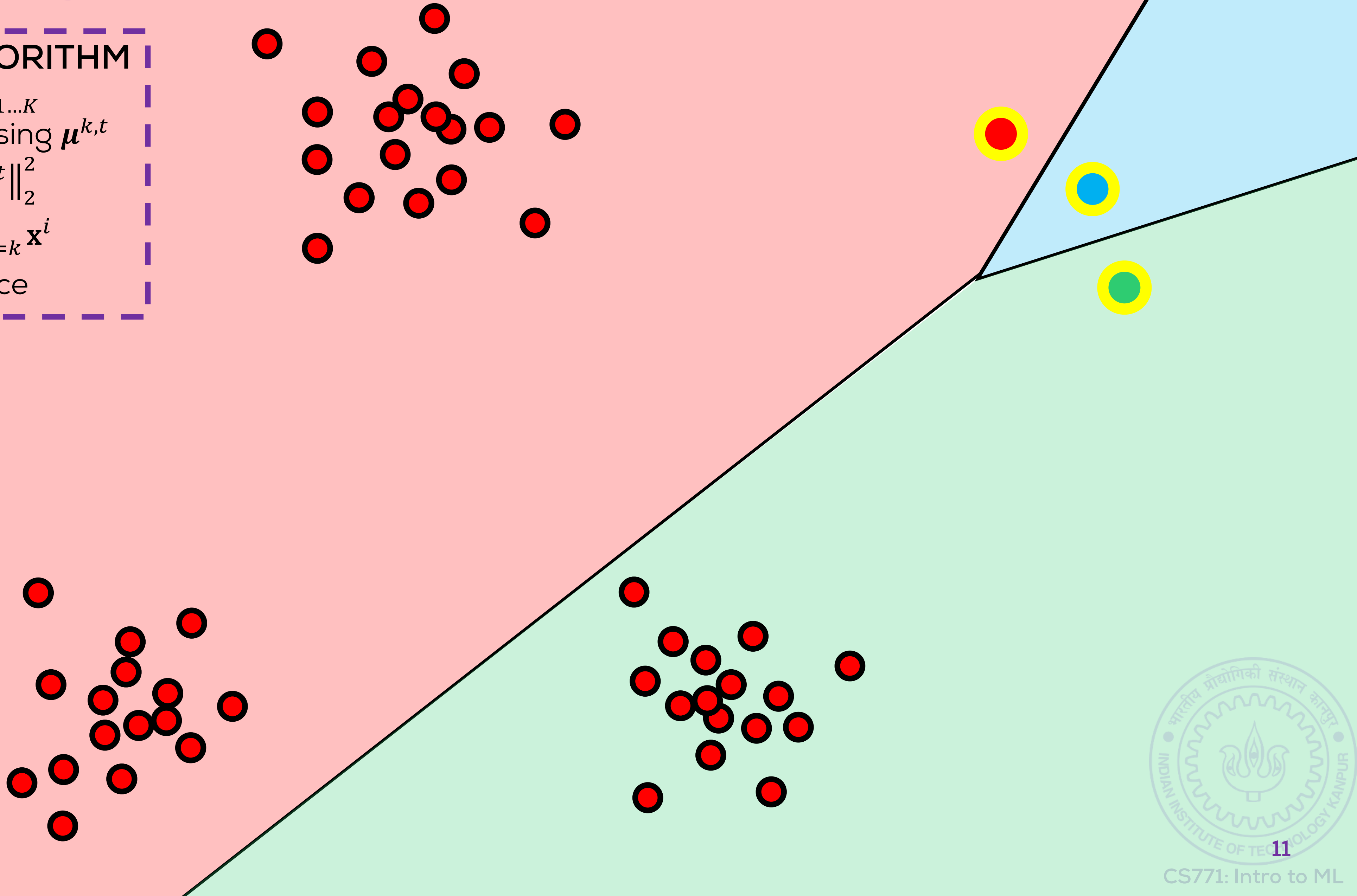


K-Means Algorithm in action!

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

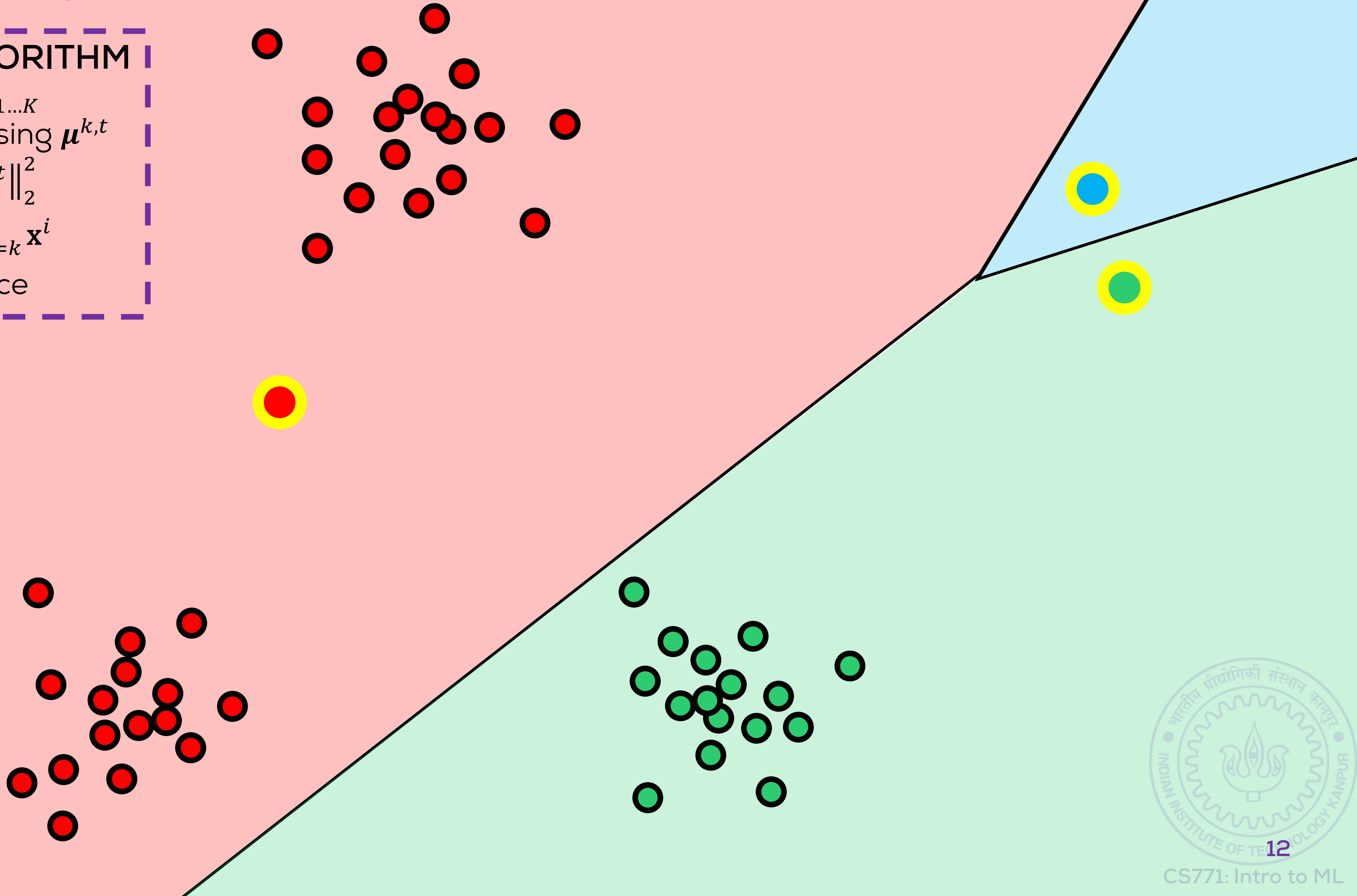


K-Means Algorithm in action!

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

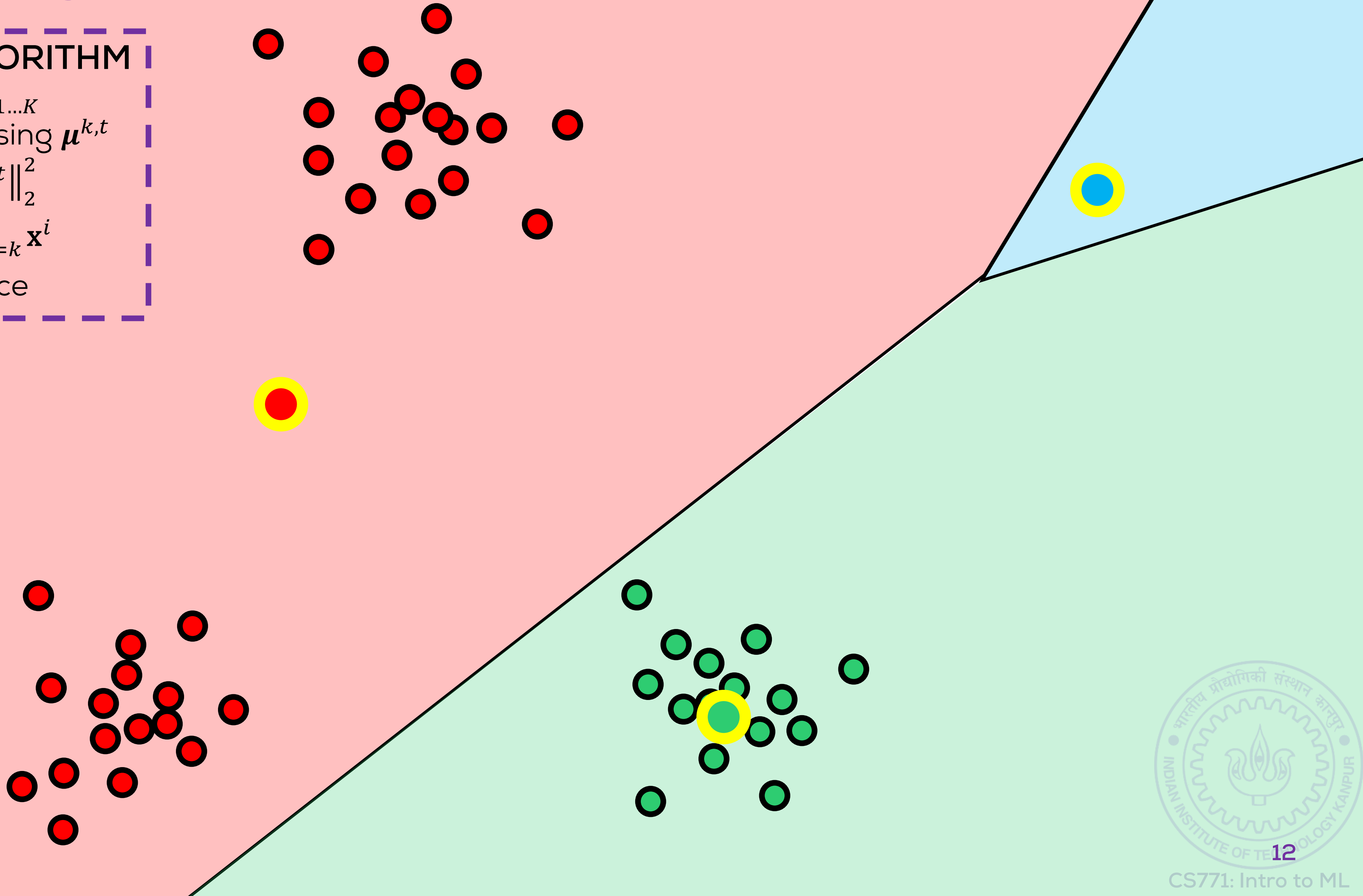


K-Means Algorithm in action!

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 $\boldsymbol{\mu}^{k,t}$

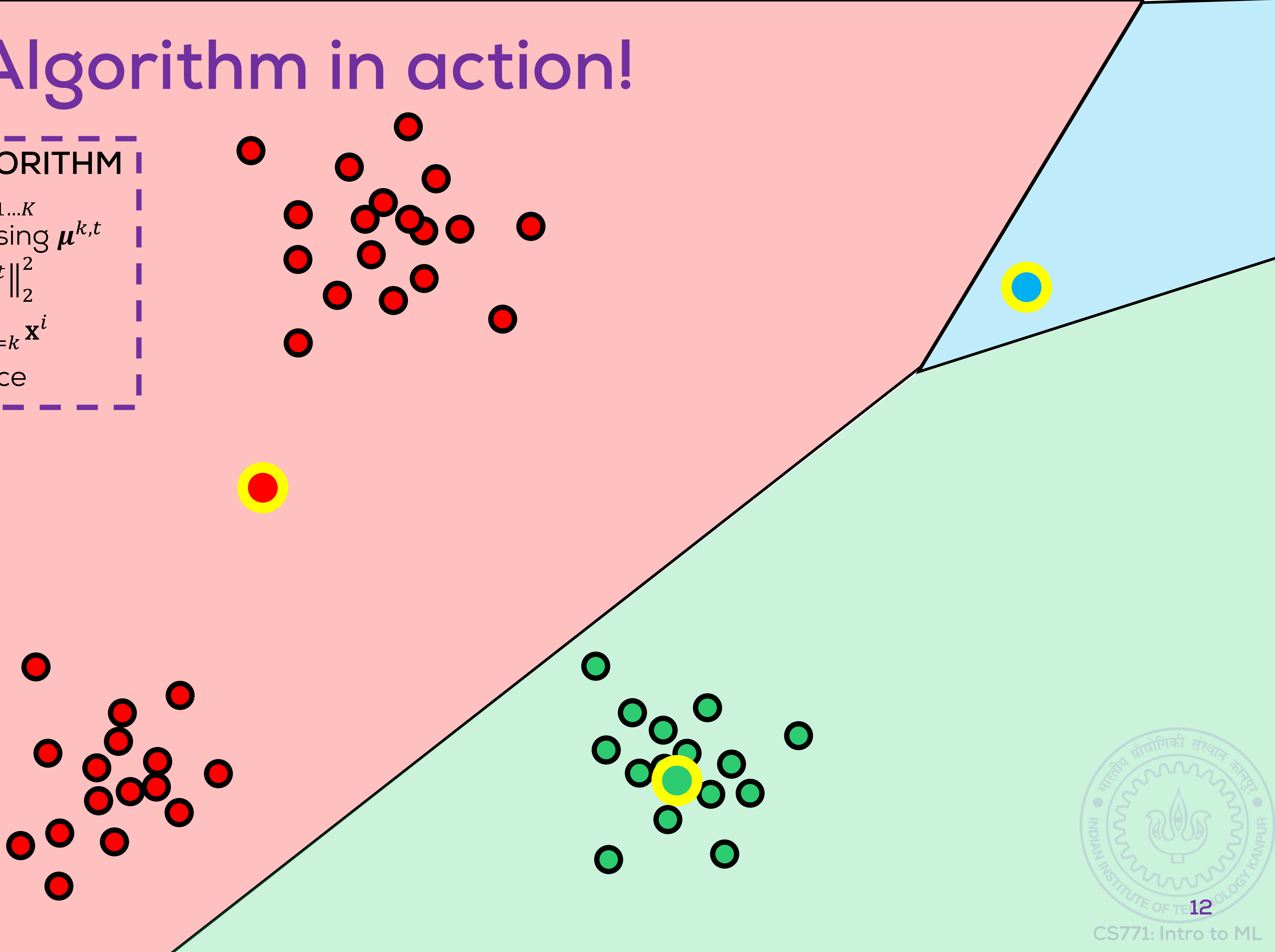
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence



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

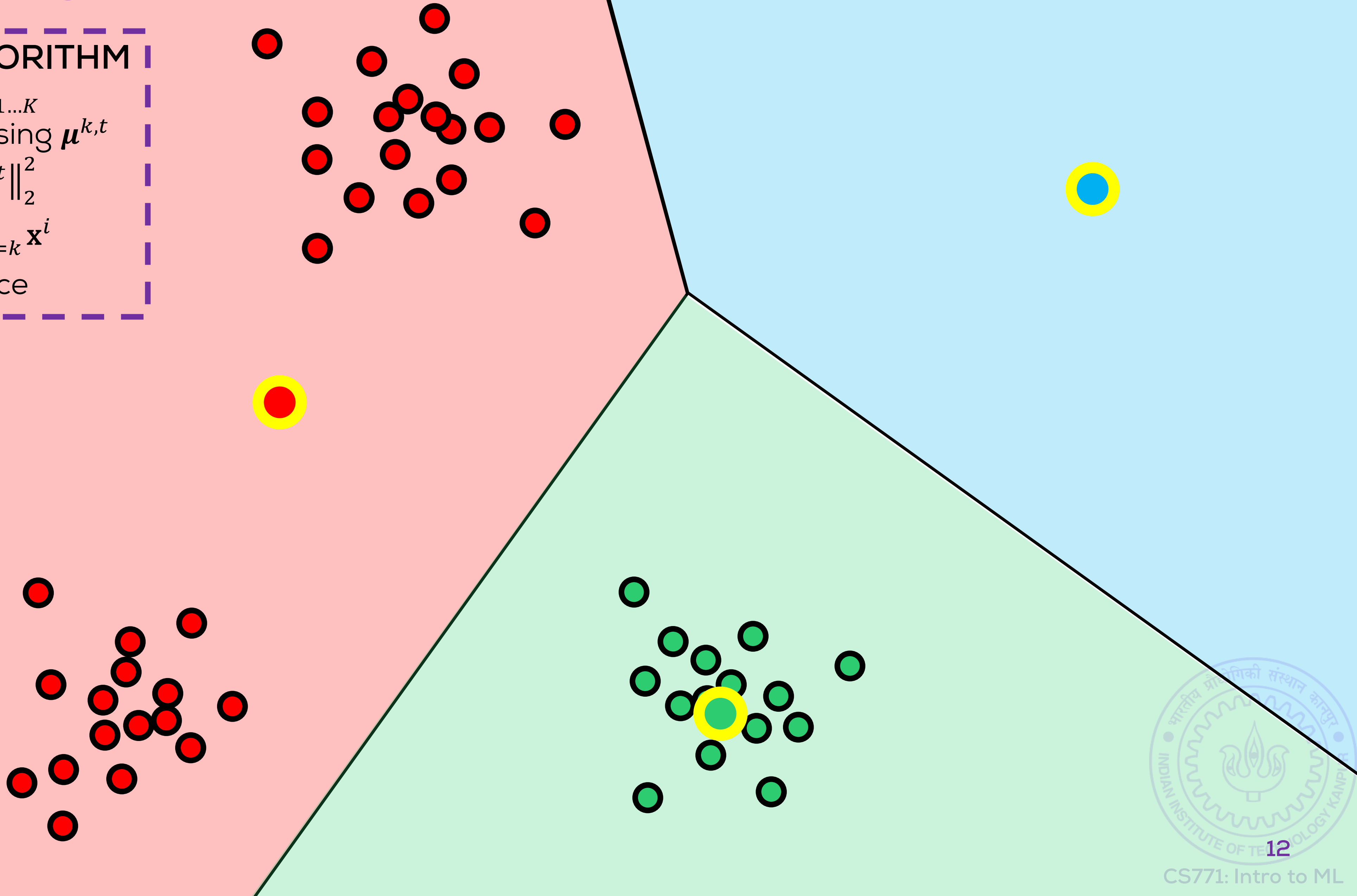


K-Means Algorithm in action!

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

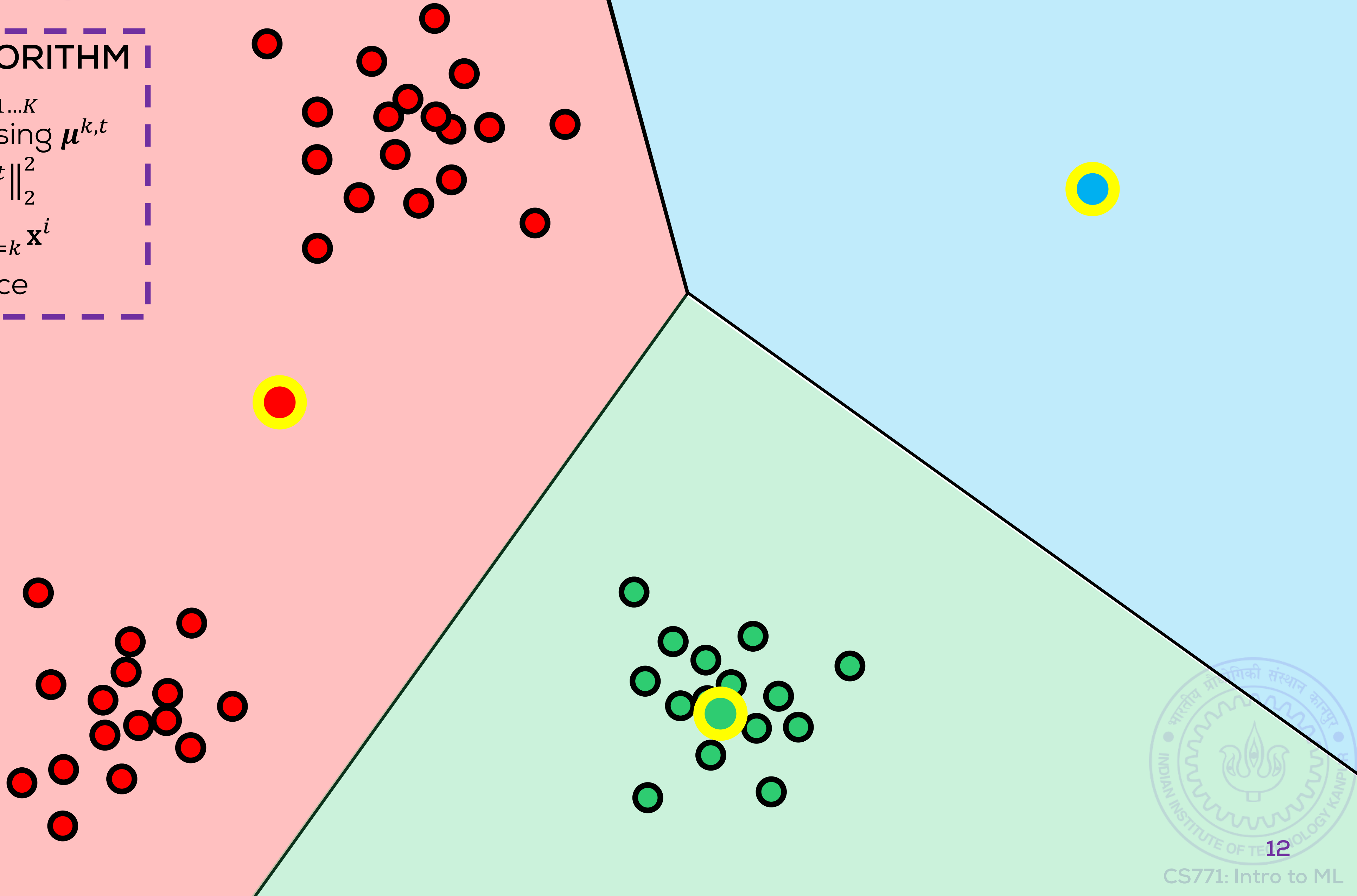


K-Means Algorithm in action!

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 $\boldsymbol{\mu}^{k,t}$

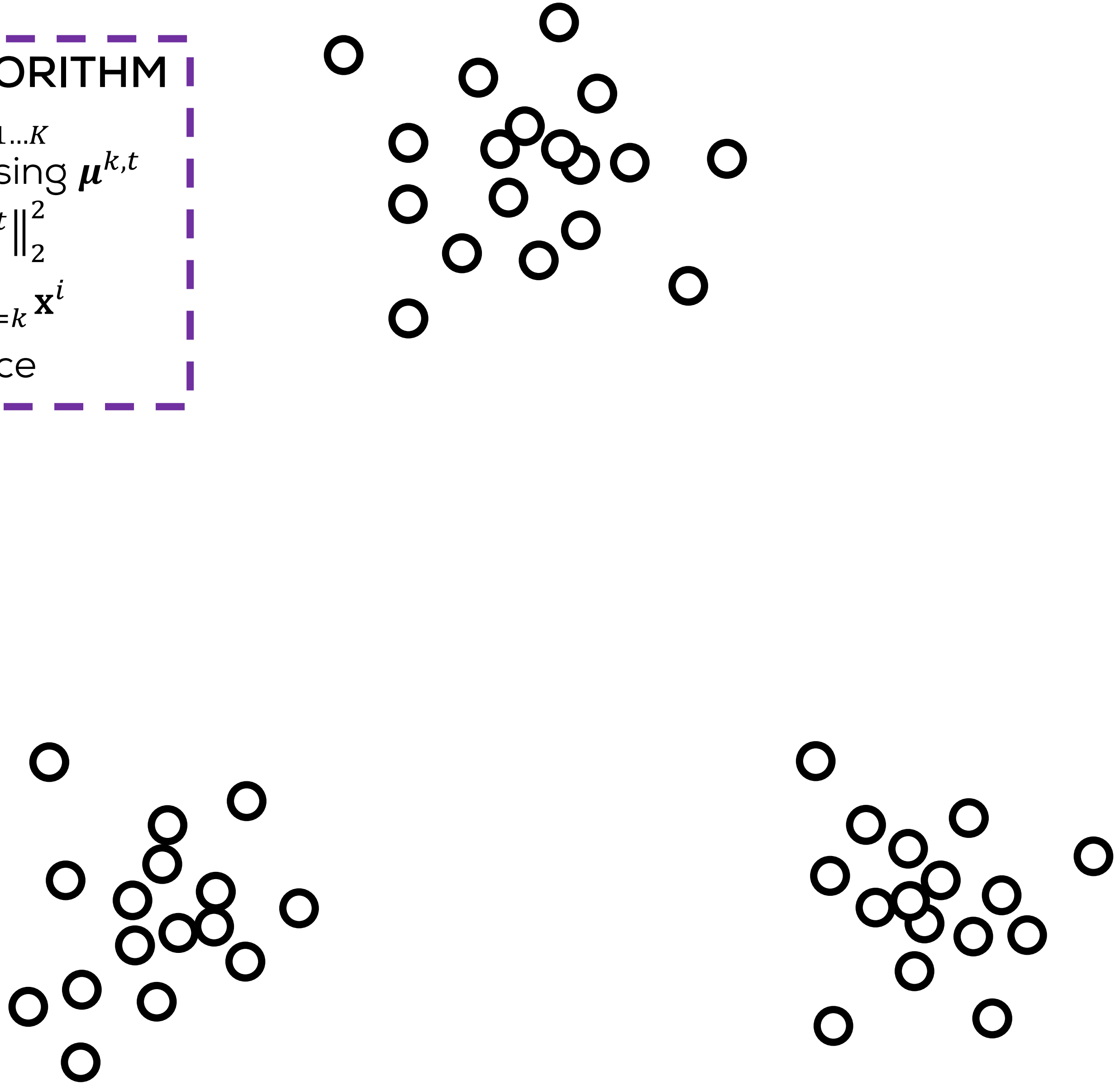
   Let $z^{i,t} = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
4. Repeat until convergence

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

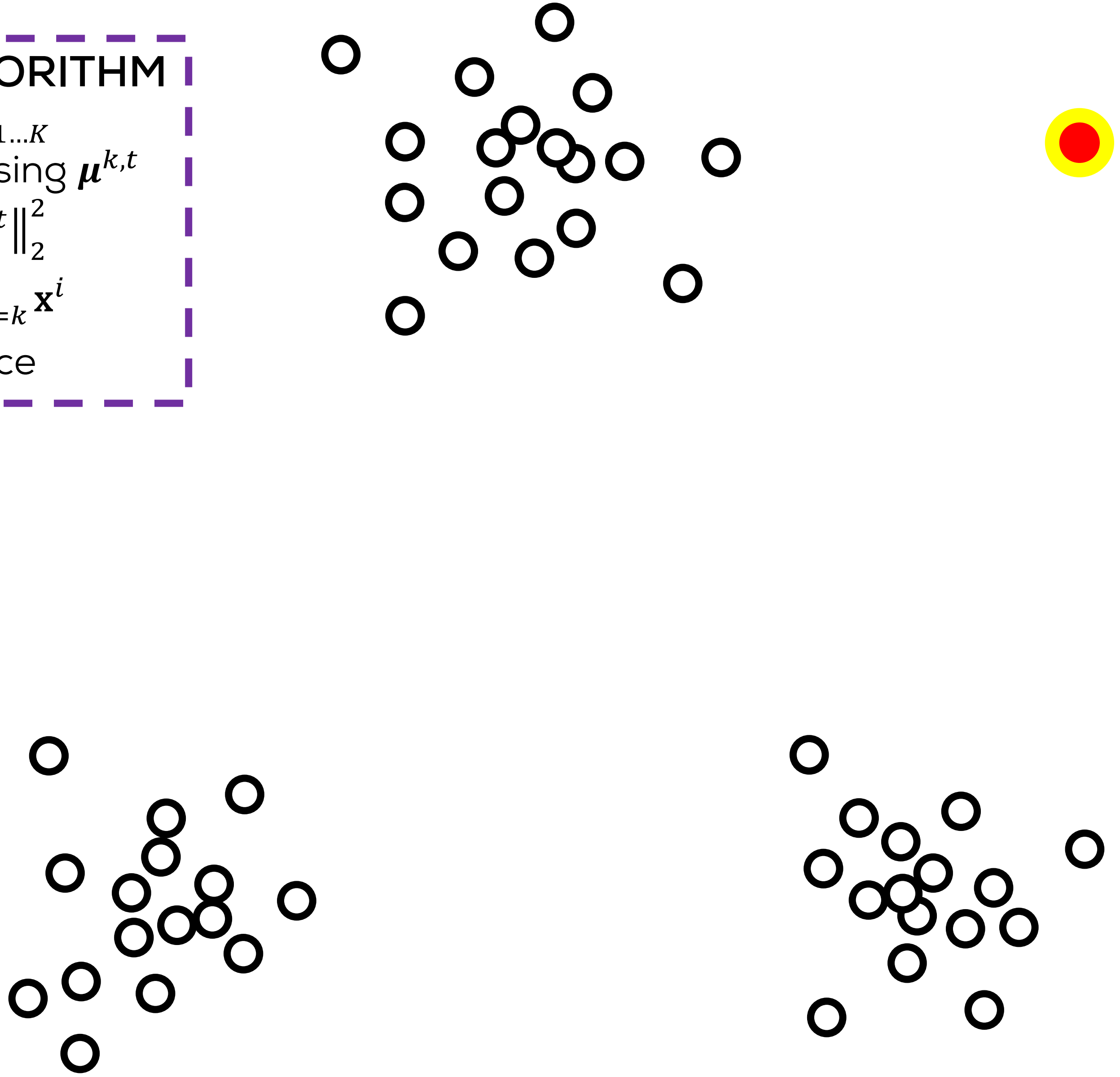


K-Means Algorithm in action!

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

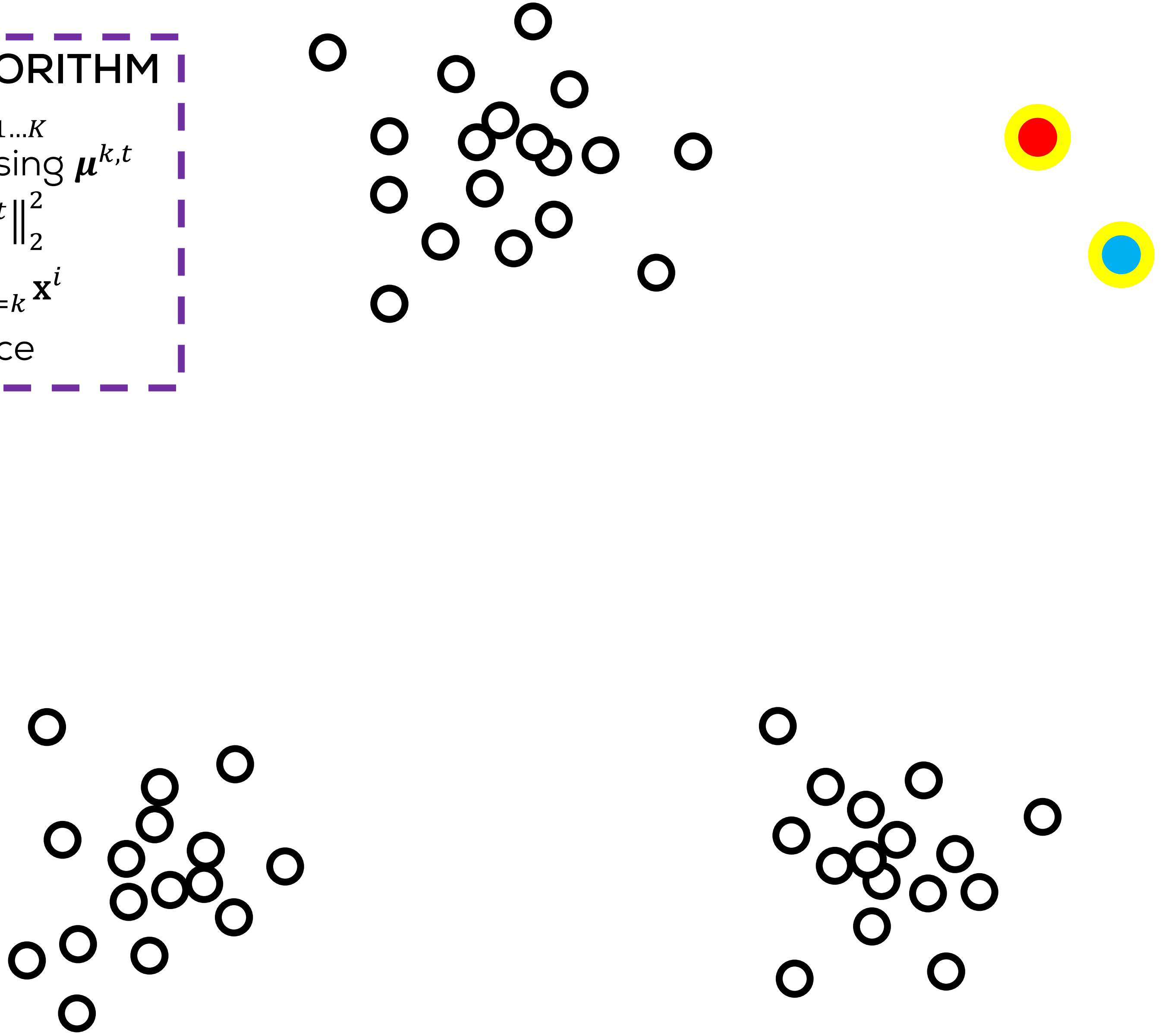


K-Means Algorithm in action!

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

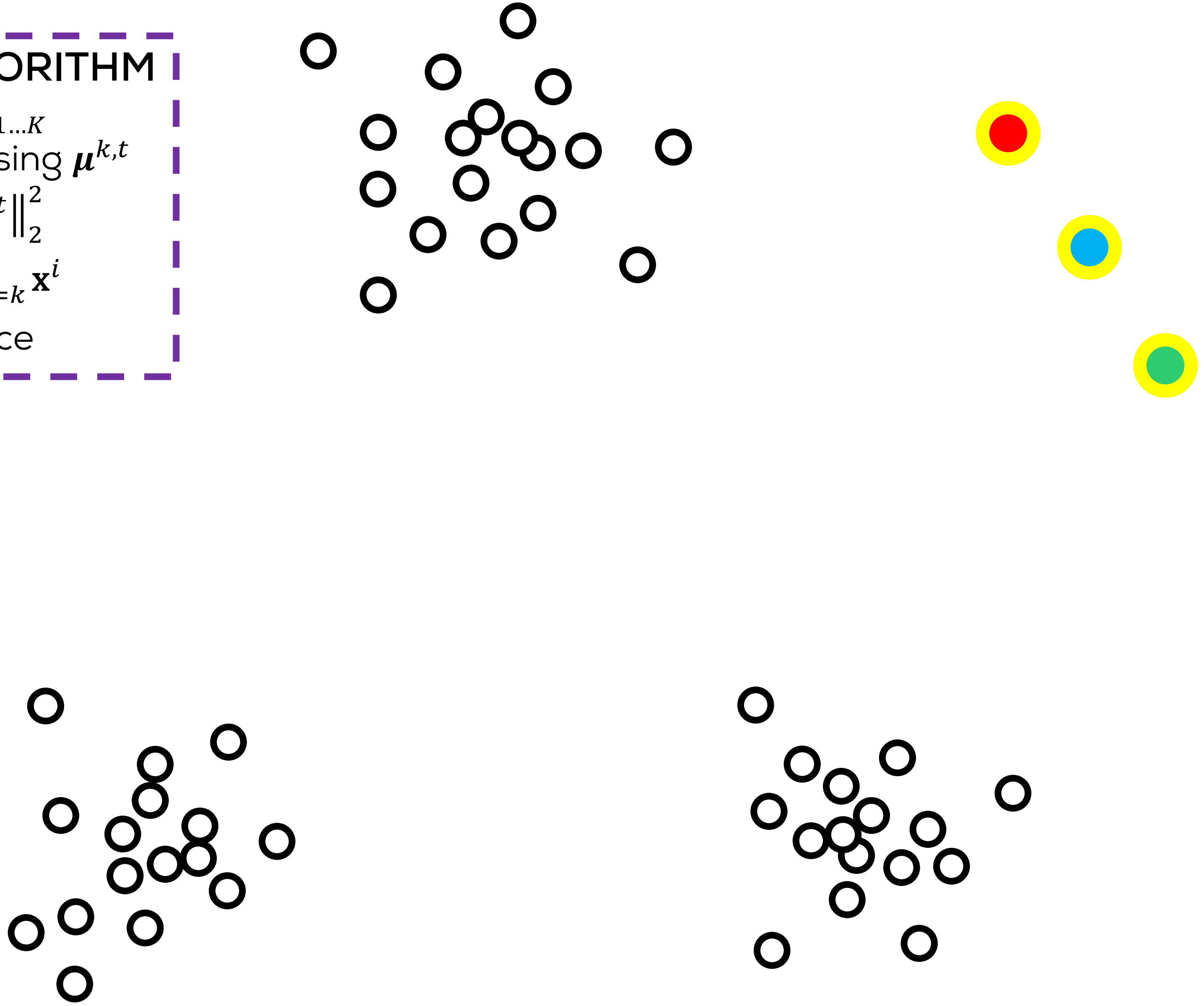


K-Means Algorithm in action!

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

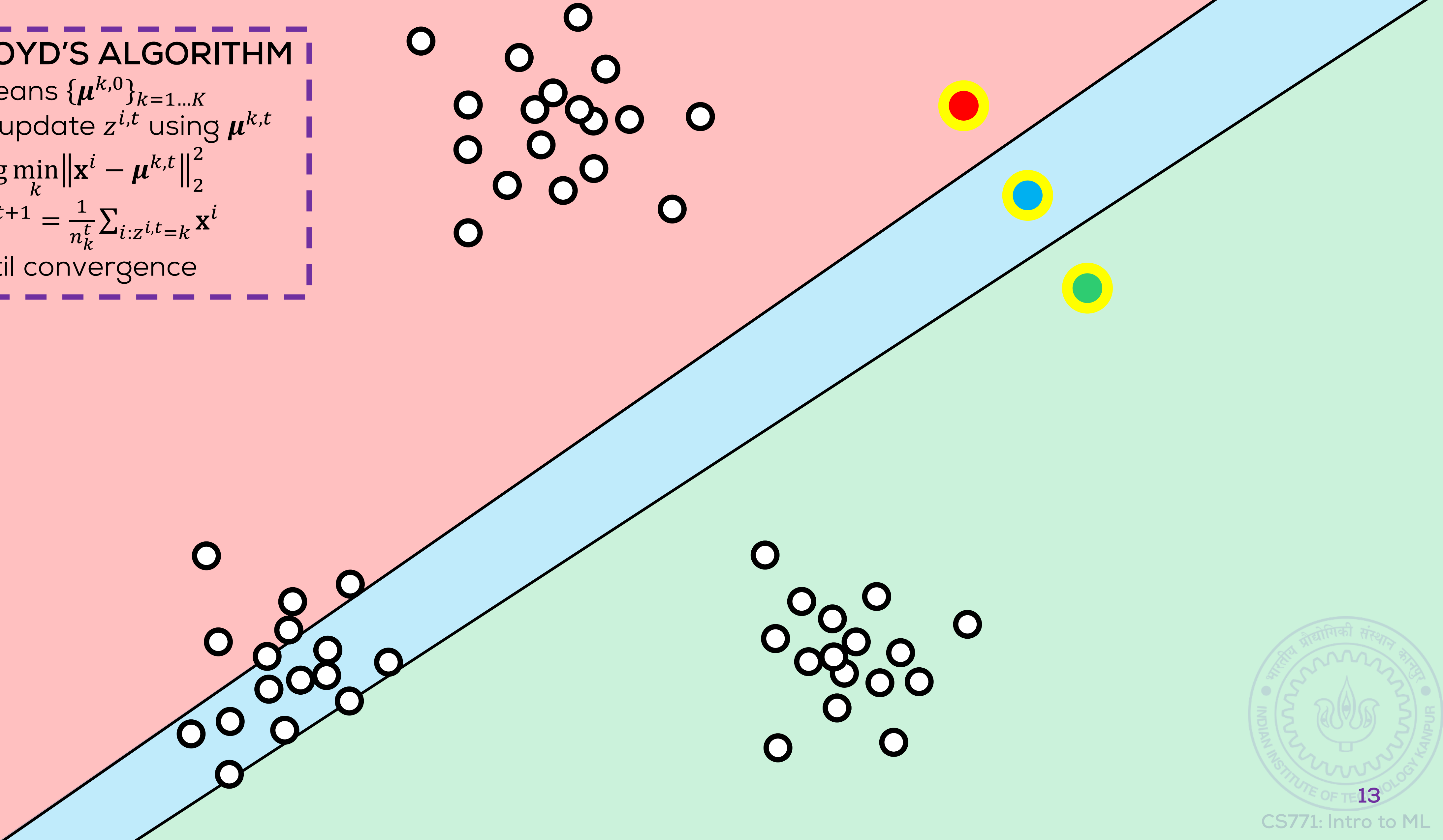


K-Means Algorithm in action!

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

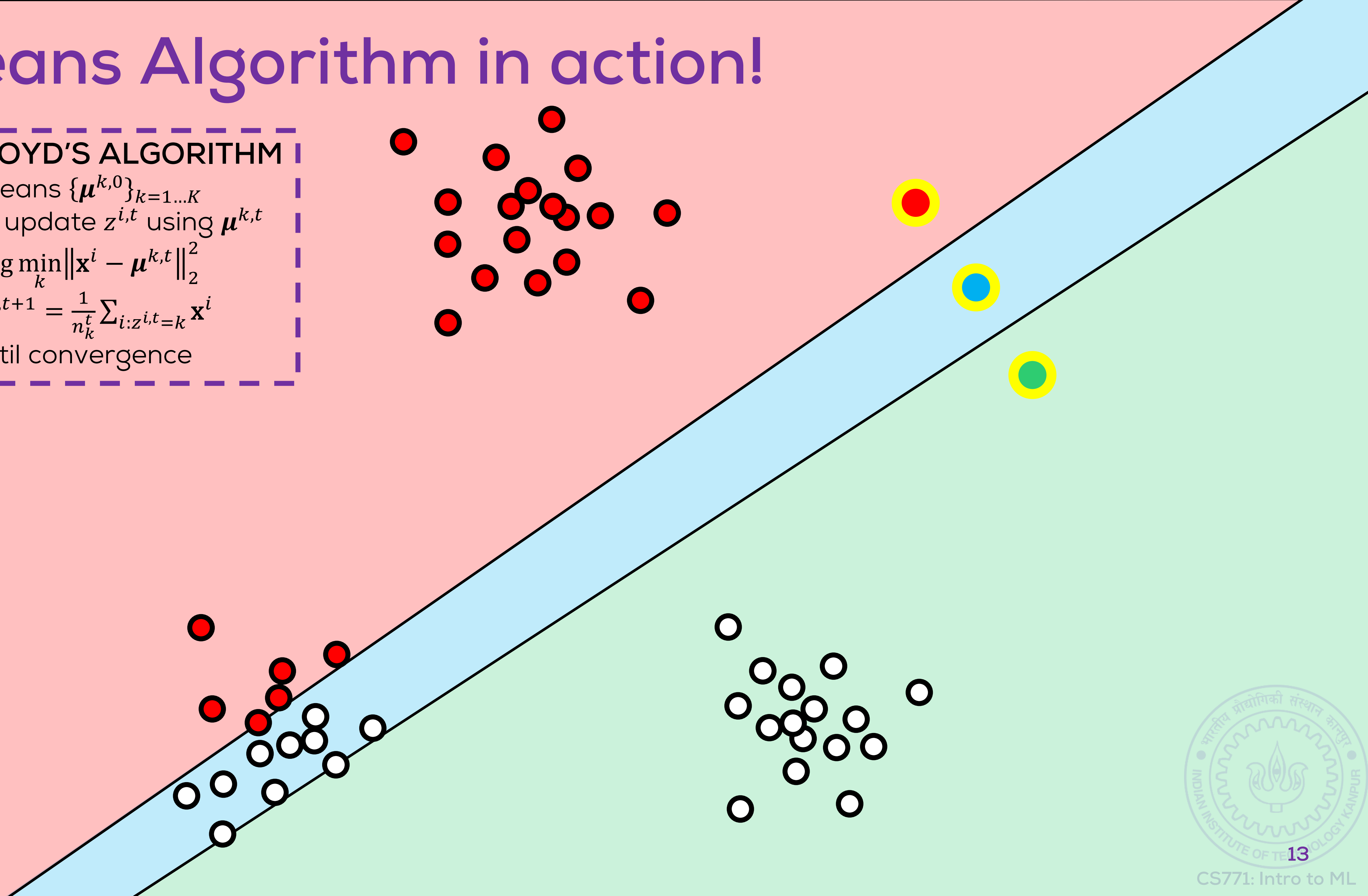


K-Means Algorithm in action!

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

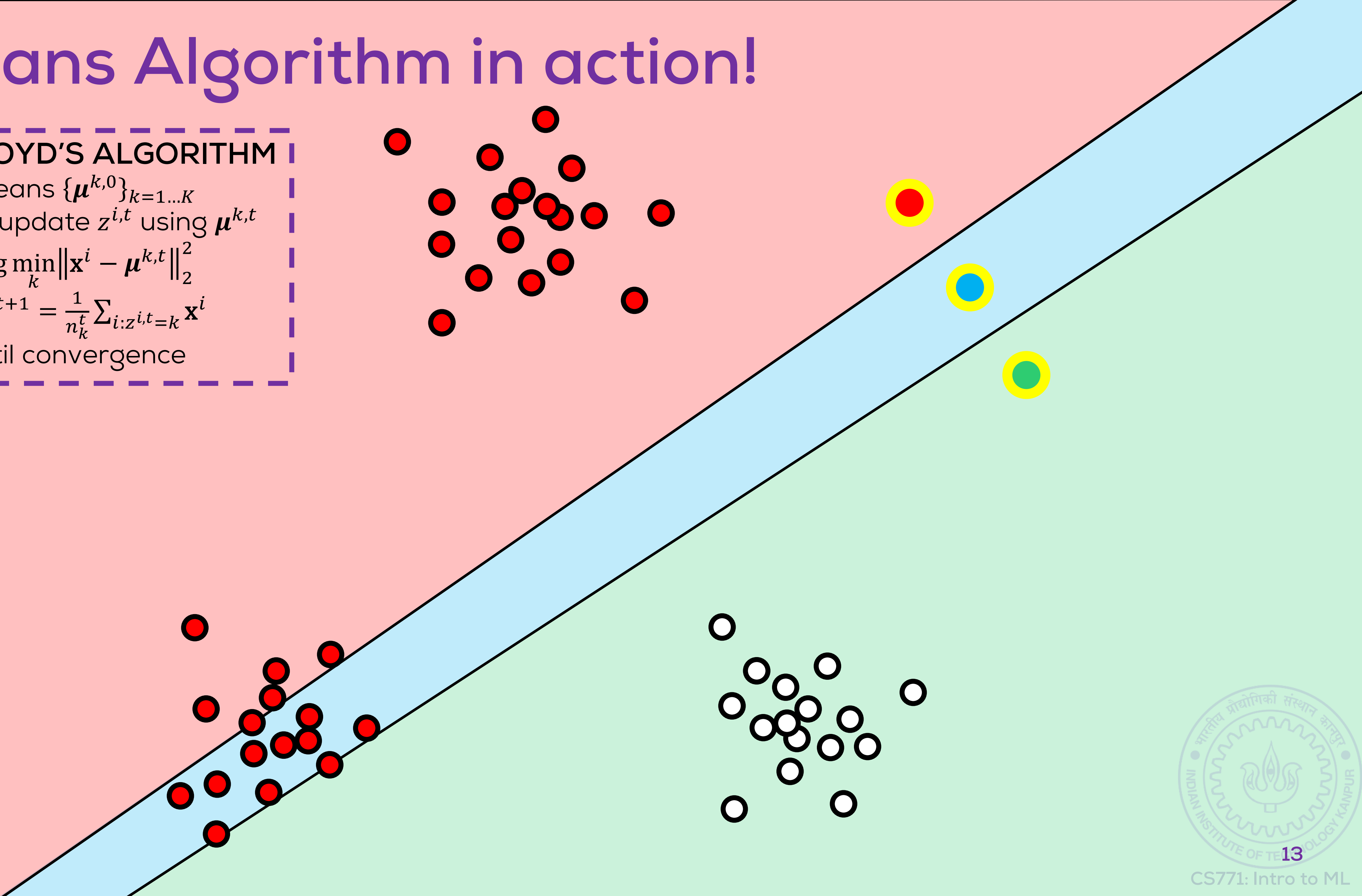


K-Means Algorithm in action!

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

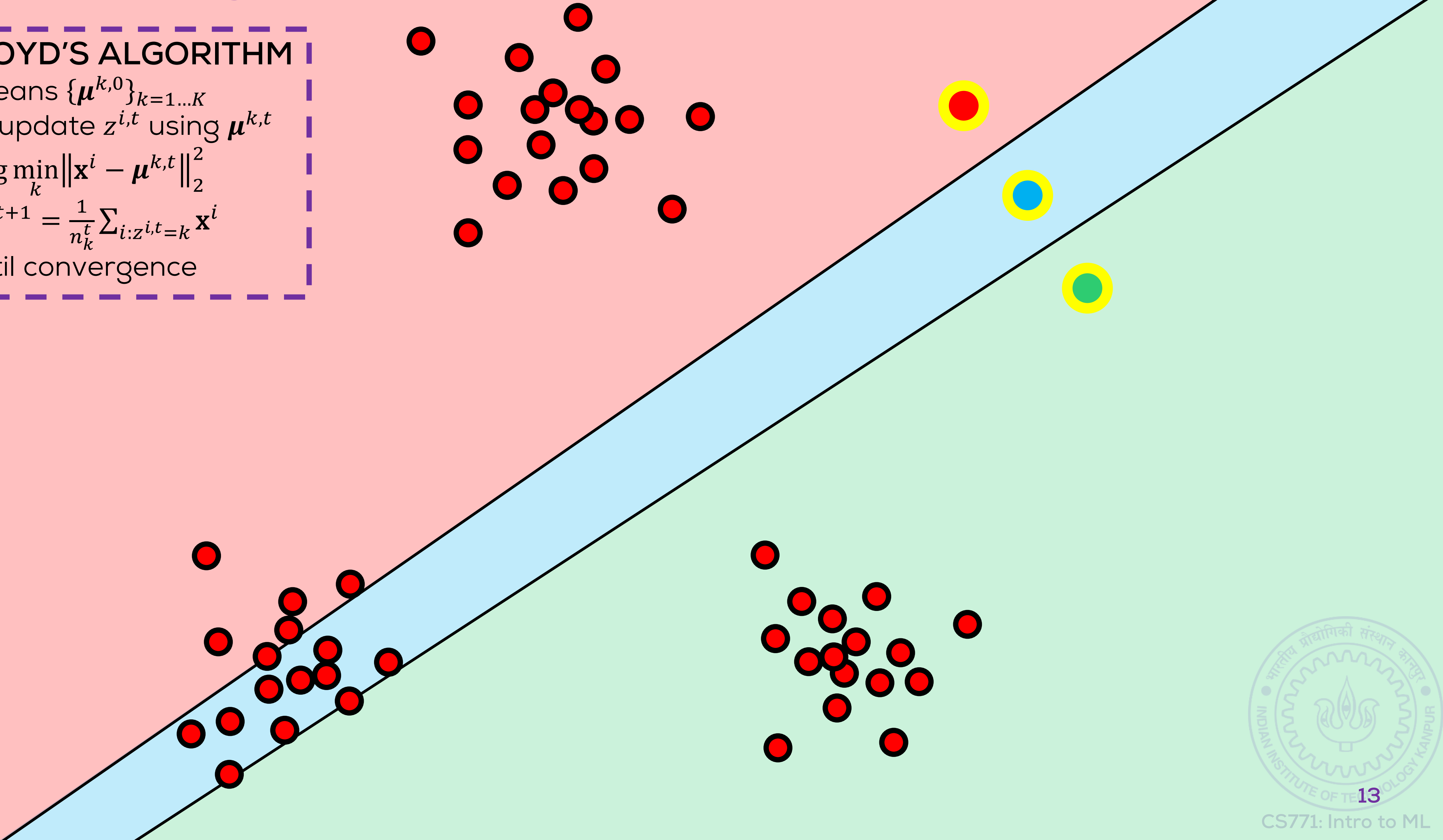


K-Means Algorithm in action!

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

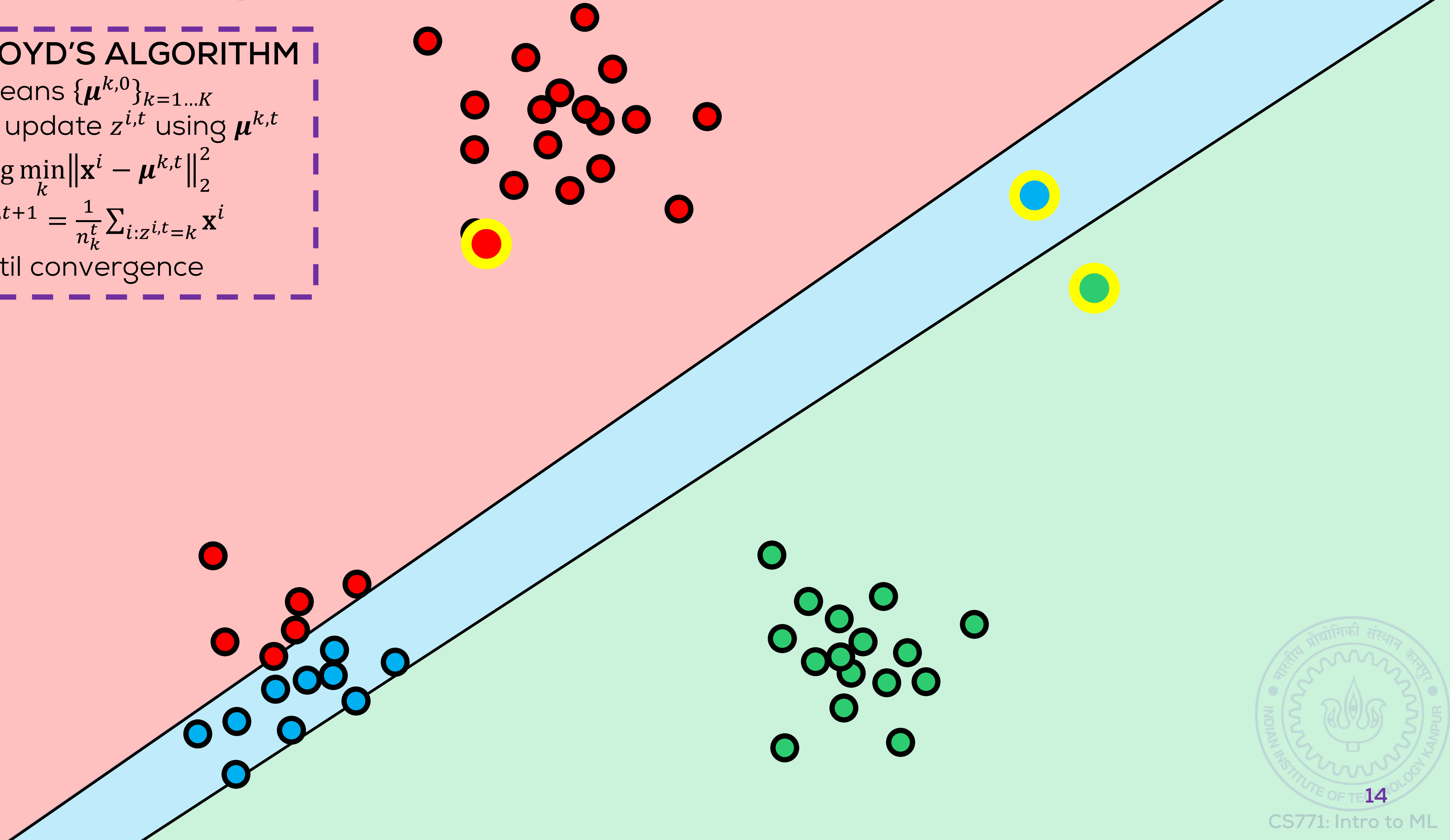


K-Means Algorithm in action!

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

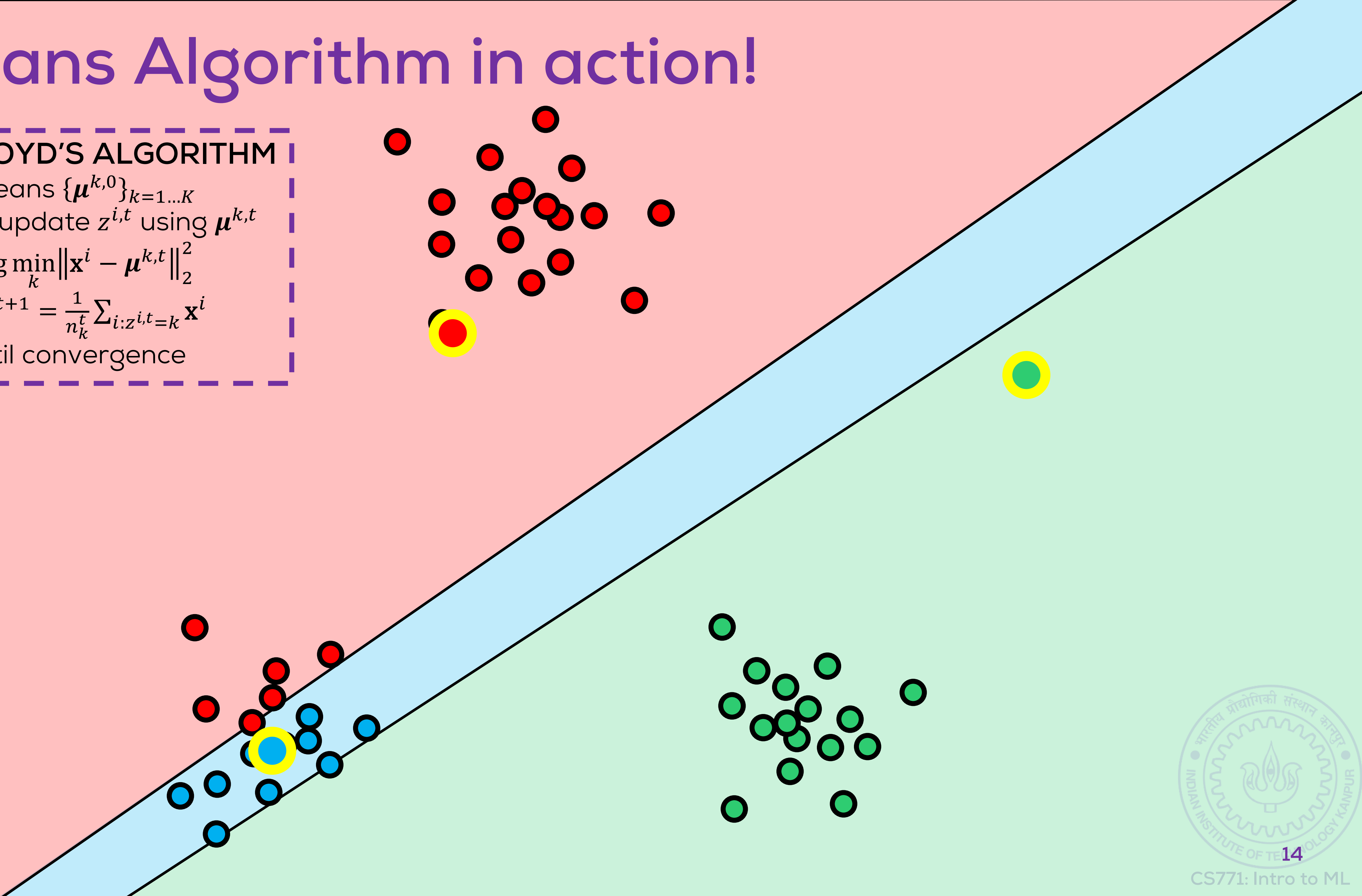


K-Means Algorithm in action!

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

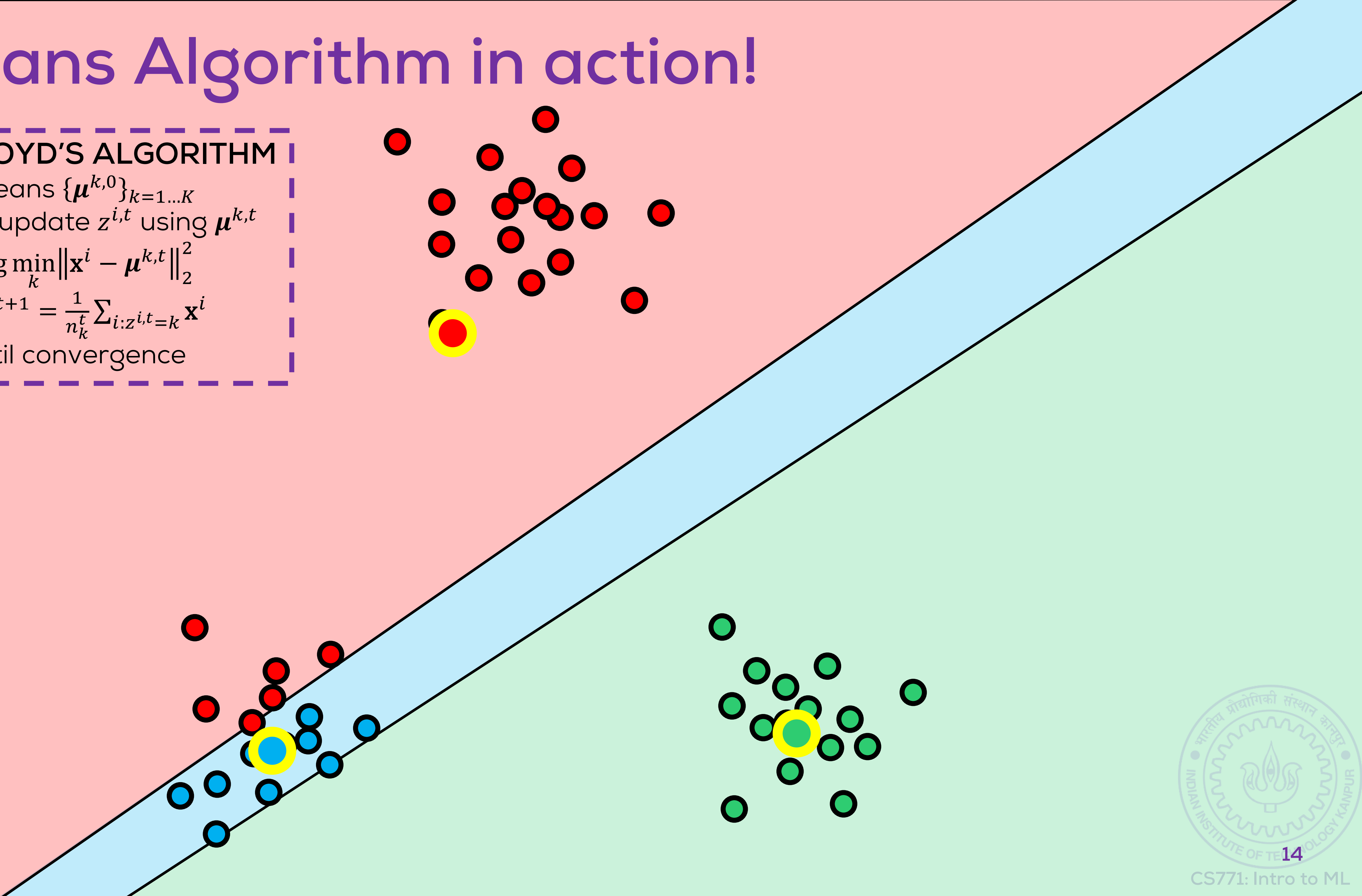


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1...K}$
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

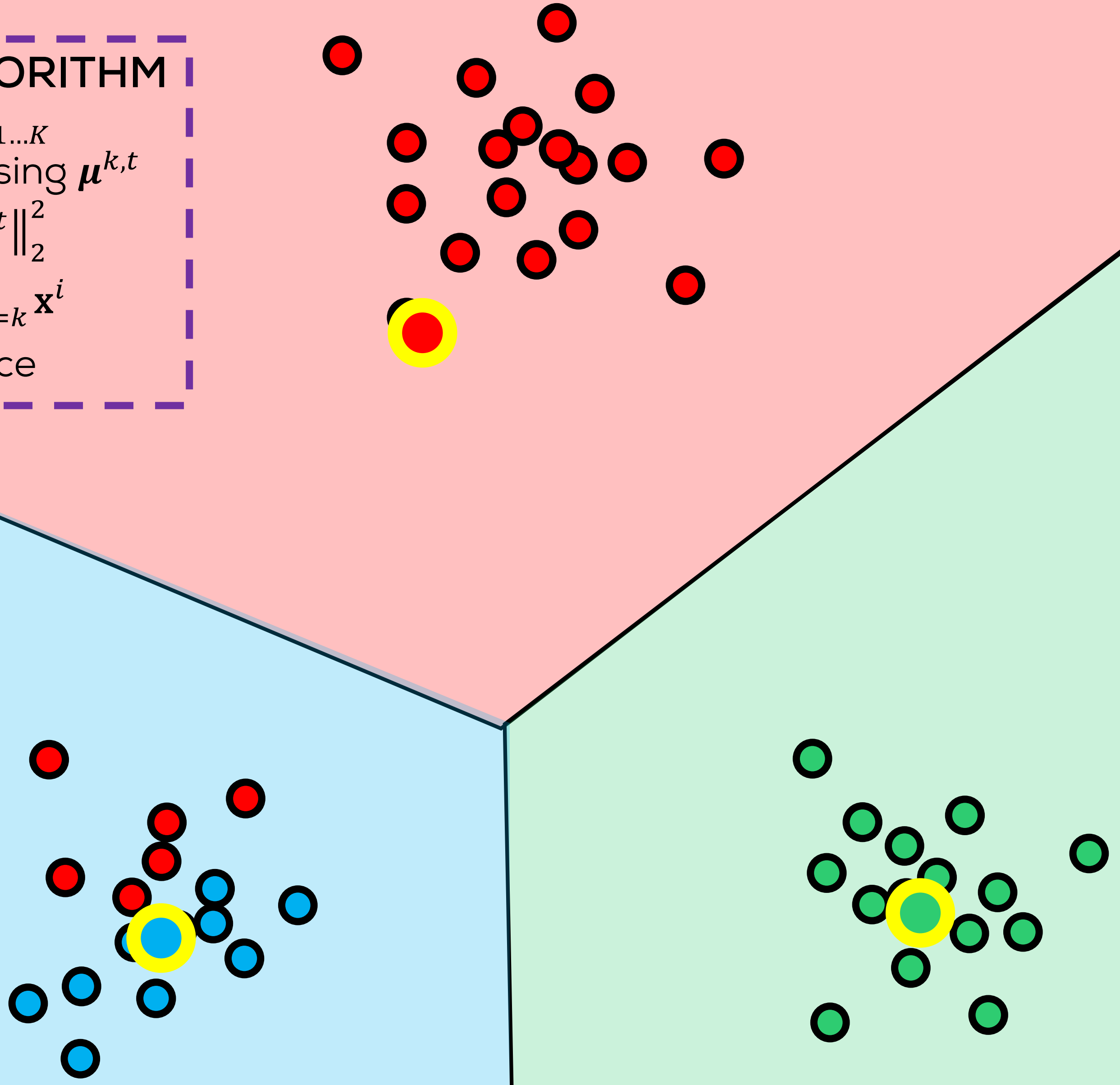


K-Means Algorithm in action!

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

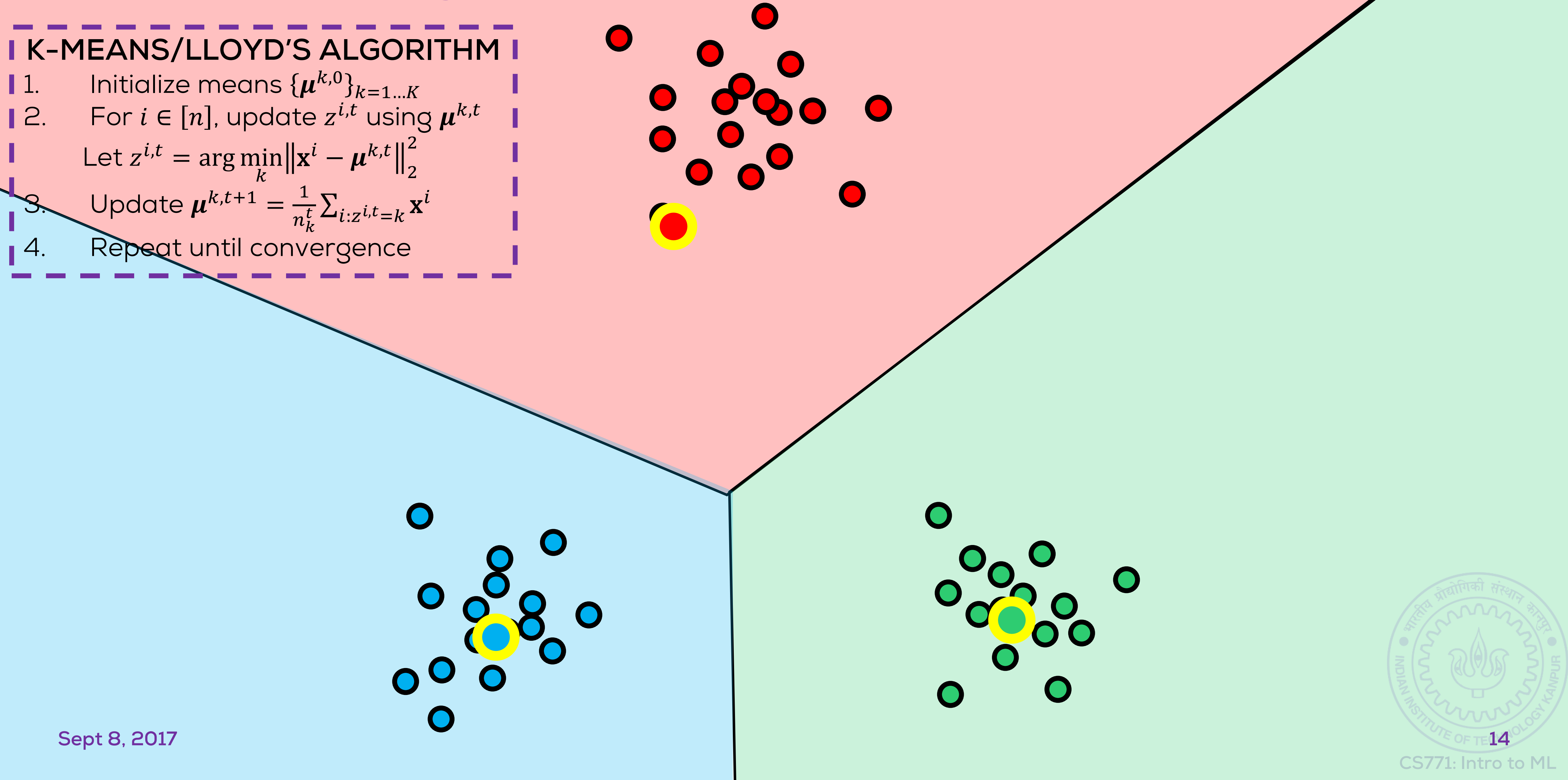


K-Means Algorithm in action!

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

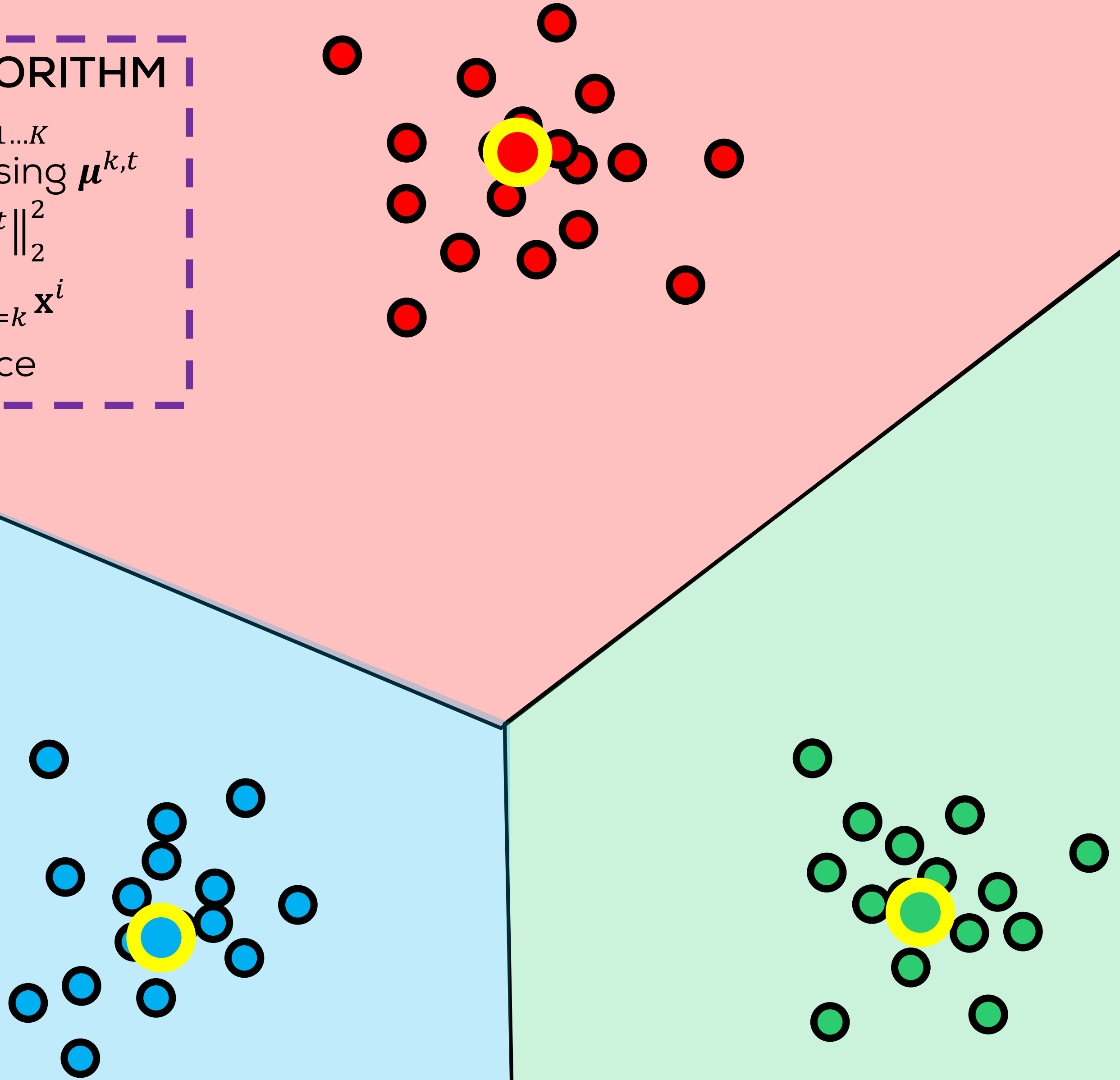


K-Means Algorithm in action!

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

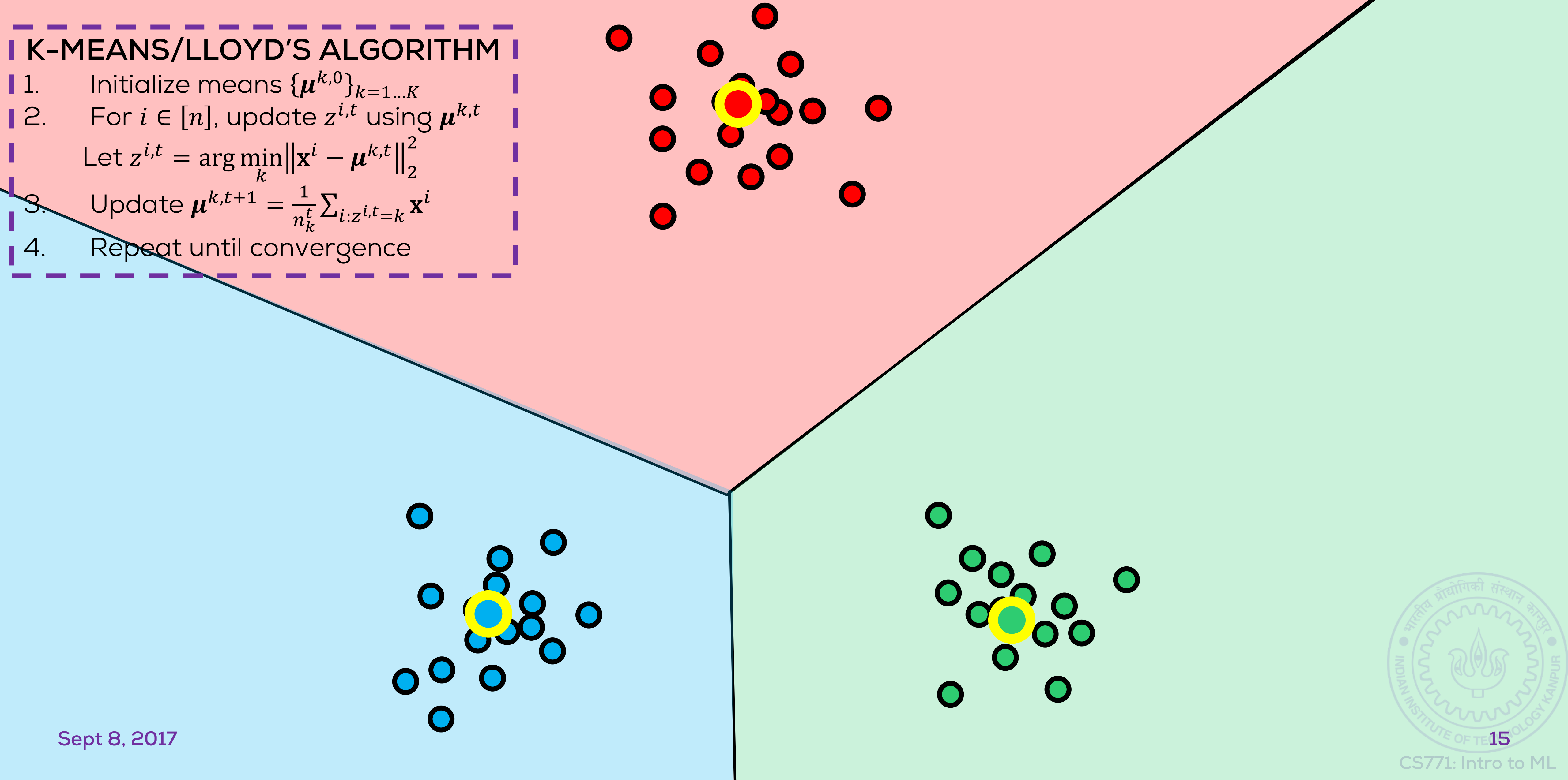


K-Means Algorithm in action!

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

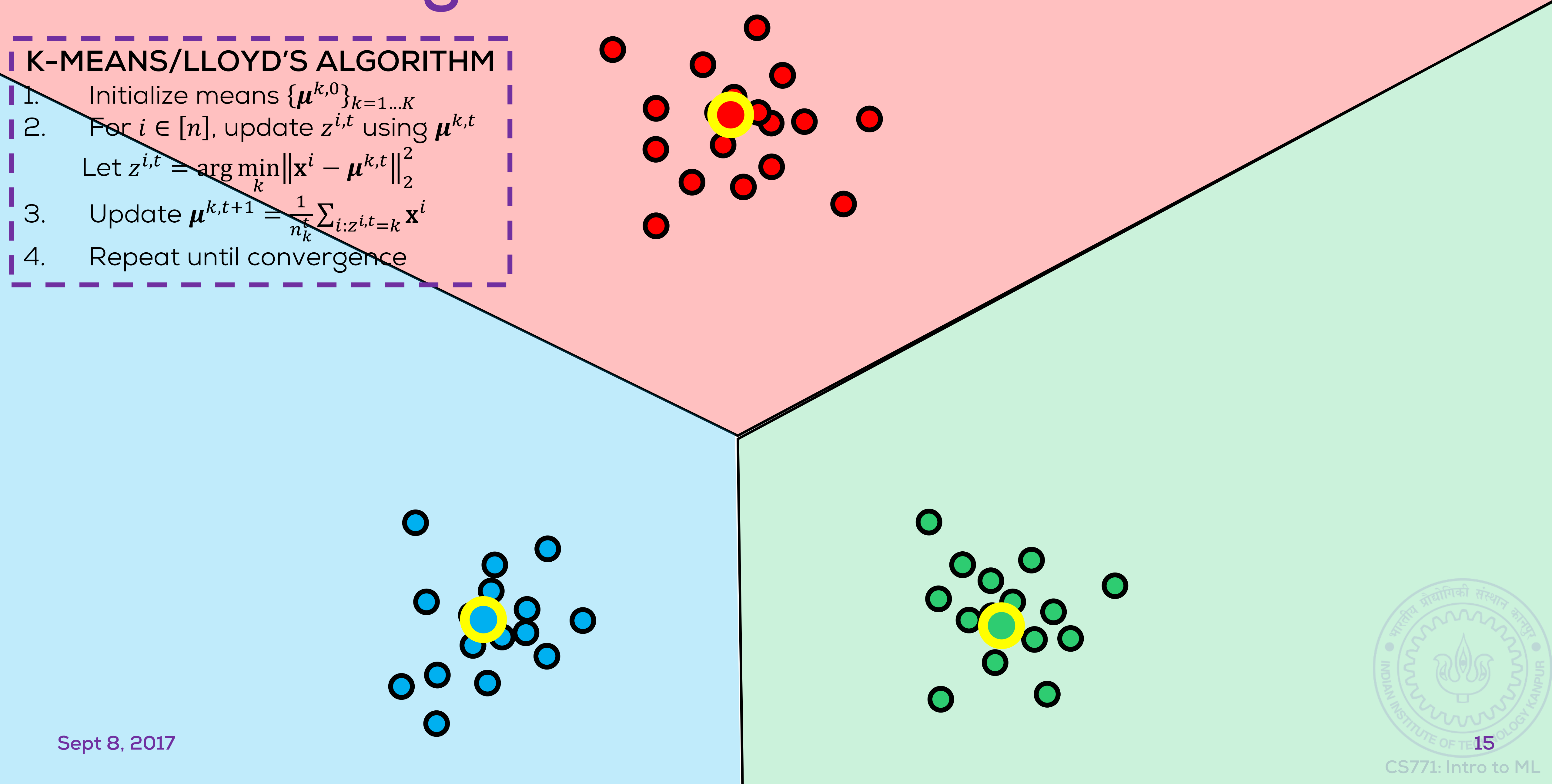


K-Means Algorithm in action!

K-MEANS/LLOYD'S ALGORITHM

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

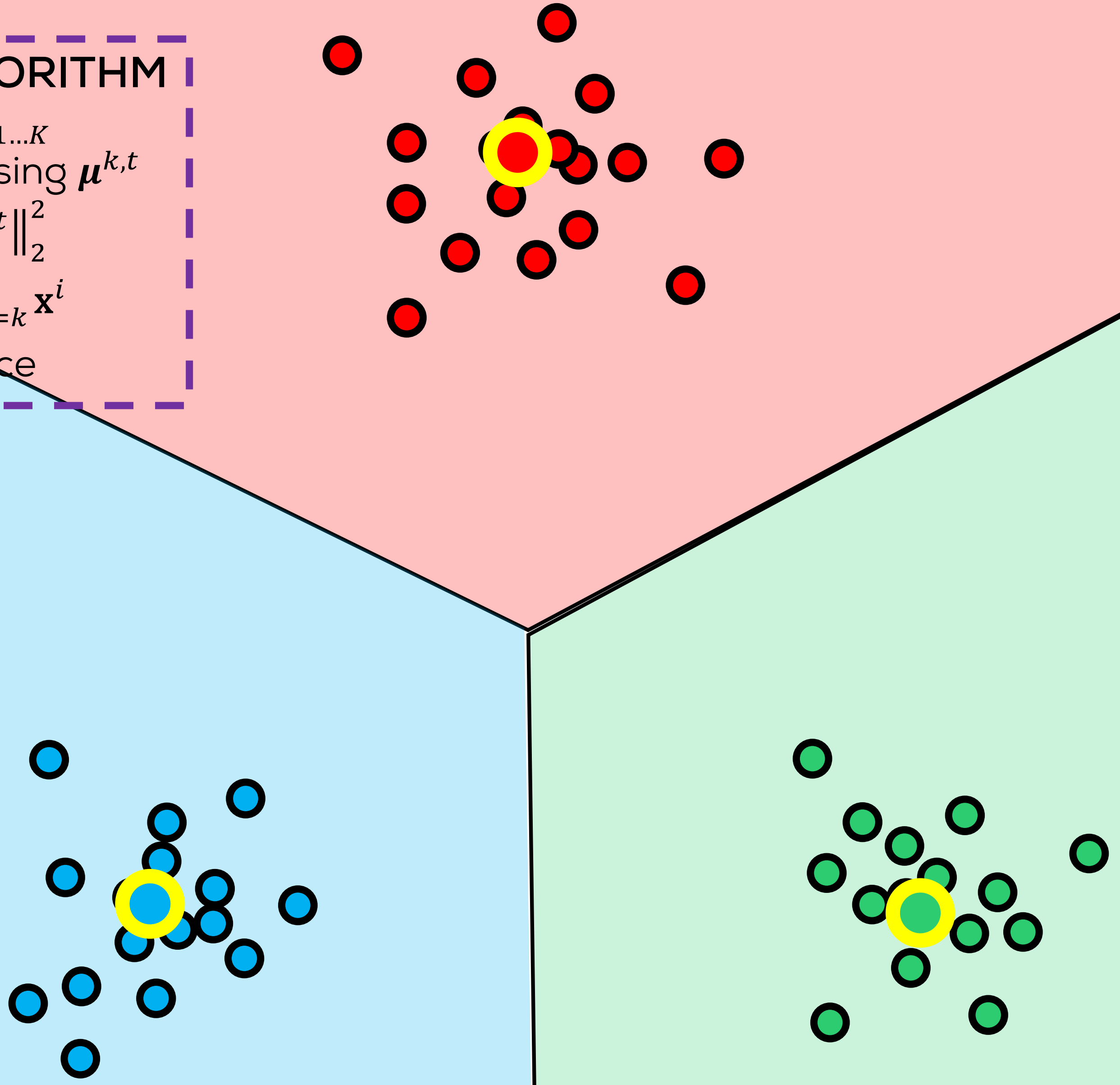
   Let $z^{i,t} = \arg\min_k \left\| \mathbf{x}^i - \boldsymbol{\mu}^{k,t} \right\|_2^2$
3. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t} \sum_{i : z^{i,t} = k} \mathbf{x}^i$
4. Repeat until convergence



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

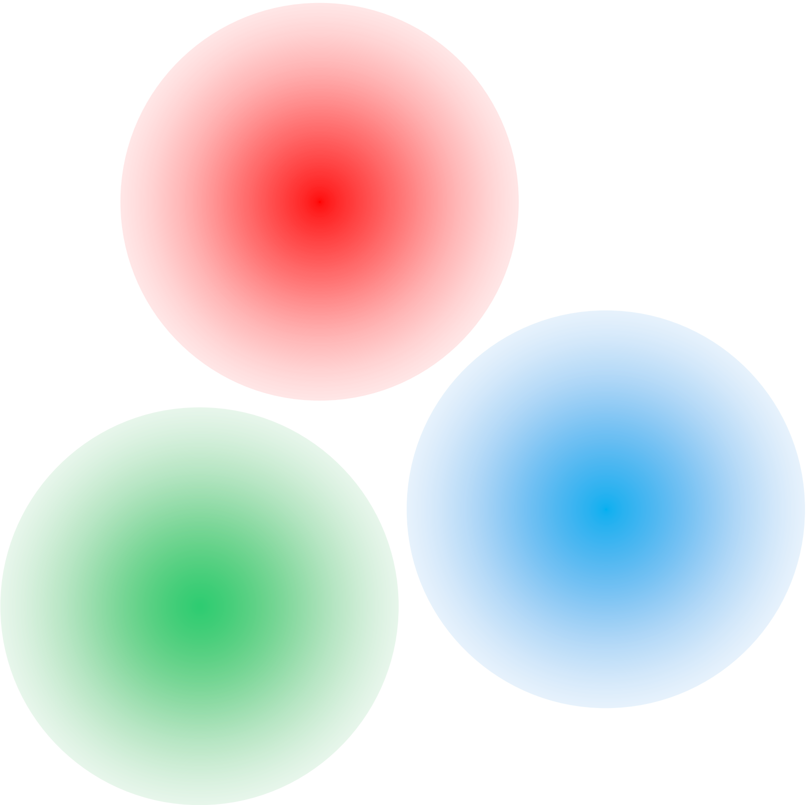


Stuck!!! … but at the global optimum ☺

The Magic is not Always very Useful!

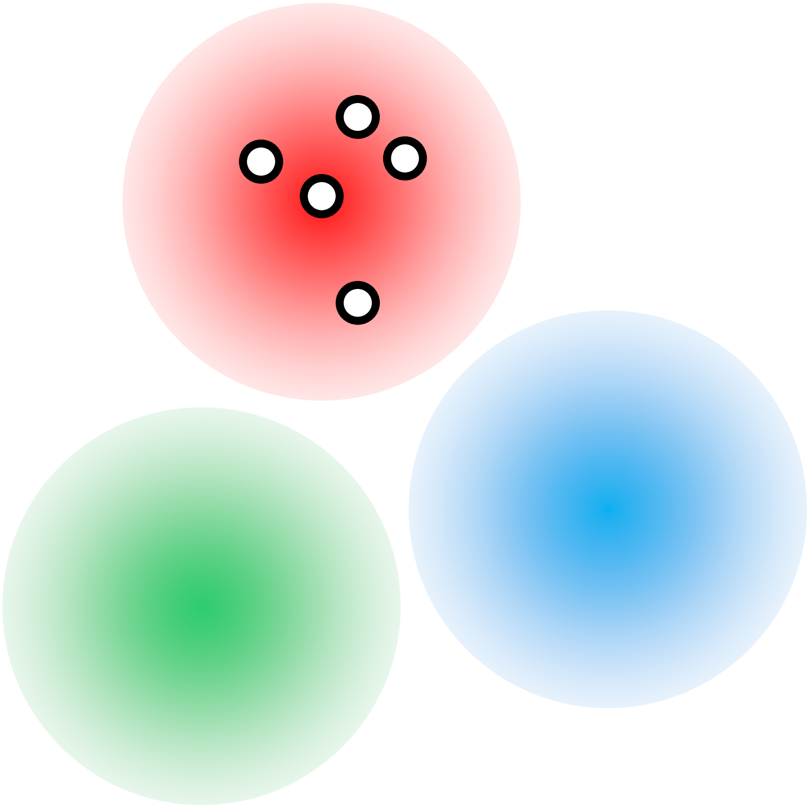
lotr.wikia.com

The Magic is not Always very Useful!



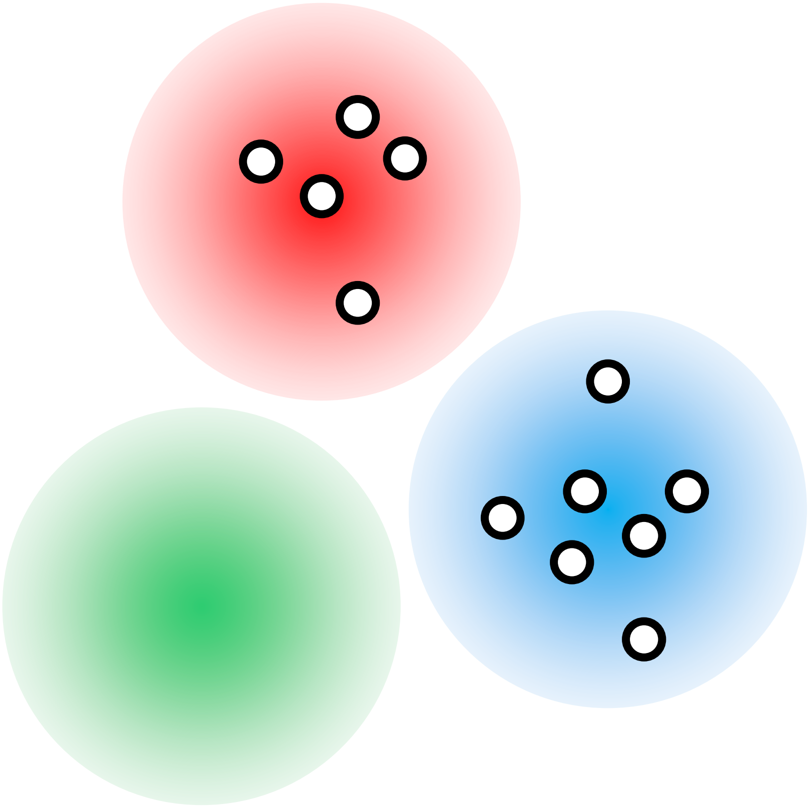
lotr.wikia.com

The Magic is not Always very Useful!



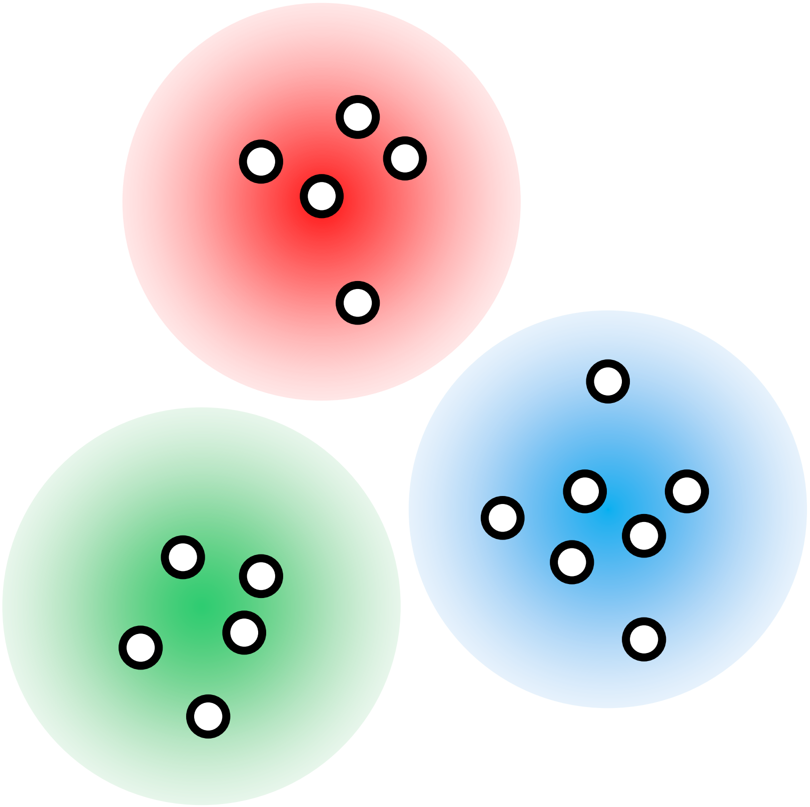
lotr.wikia.com

The Magic is not Always very Useful!

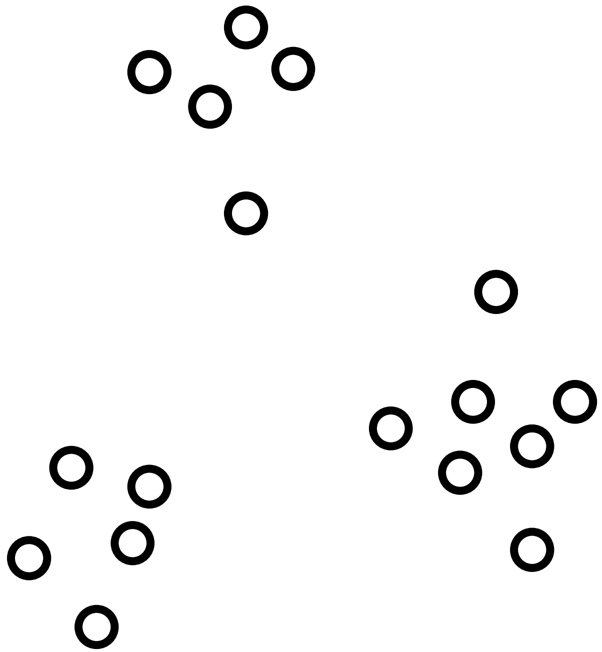


lotr.wikia.com

The Magic is not Always very Useful!

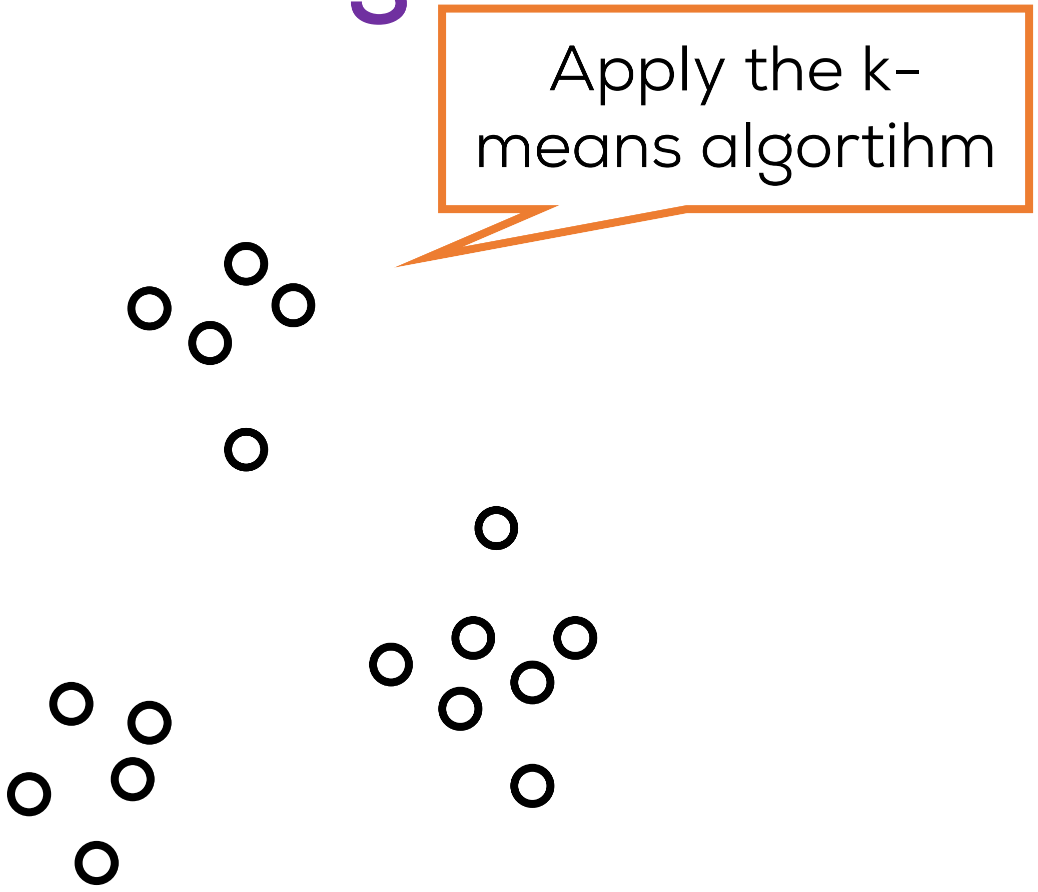


The Magic is not Always very Useful!

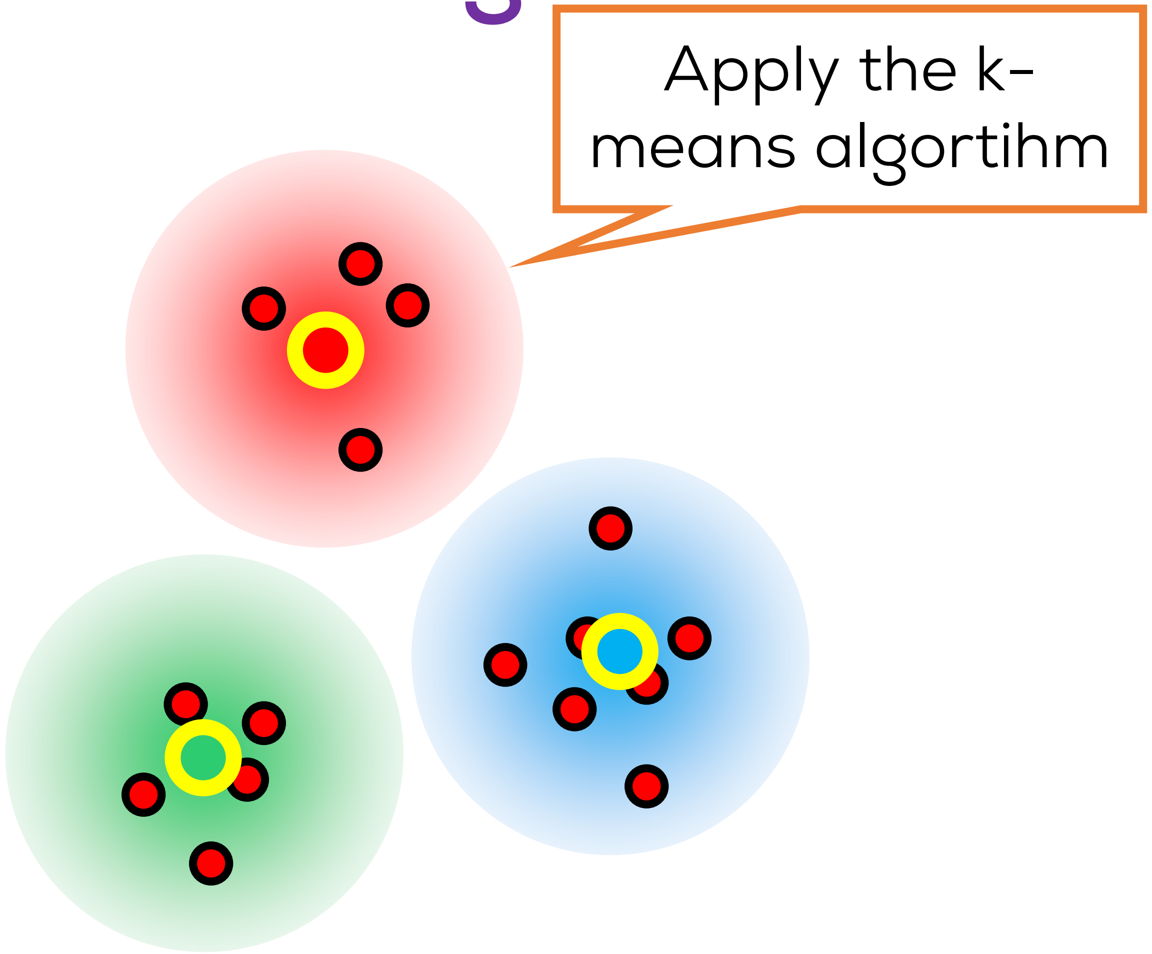


lotr.wikia.com

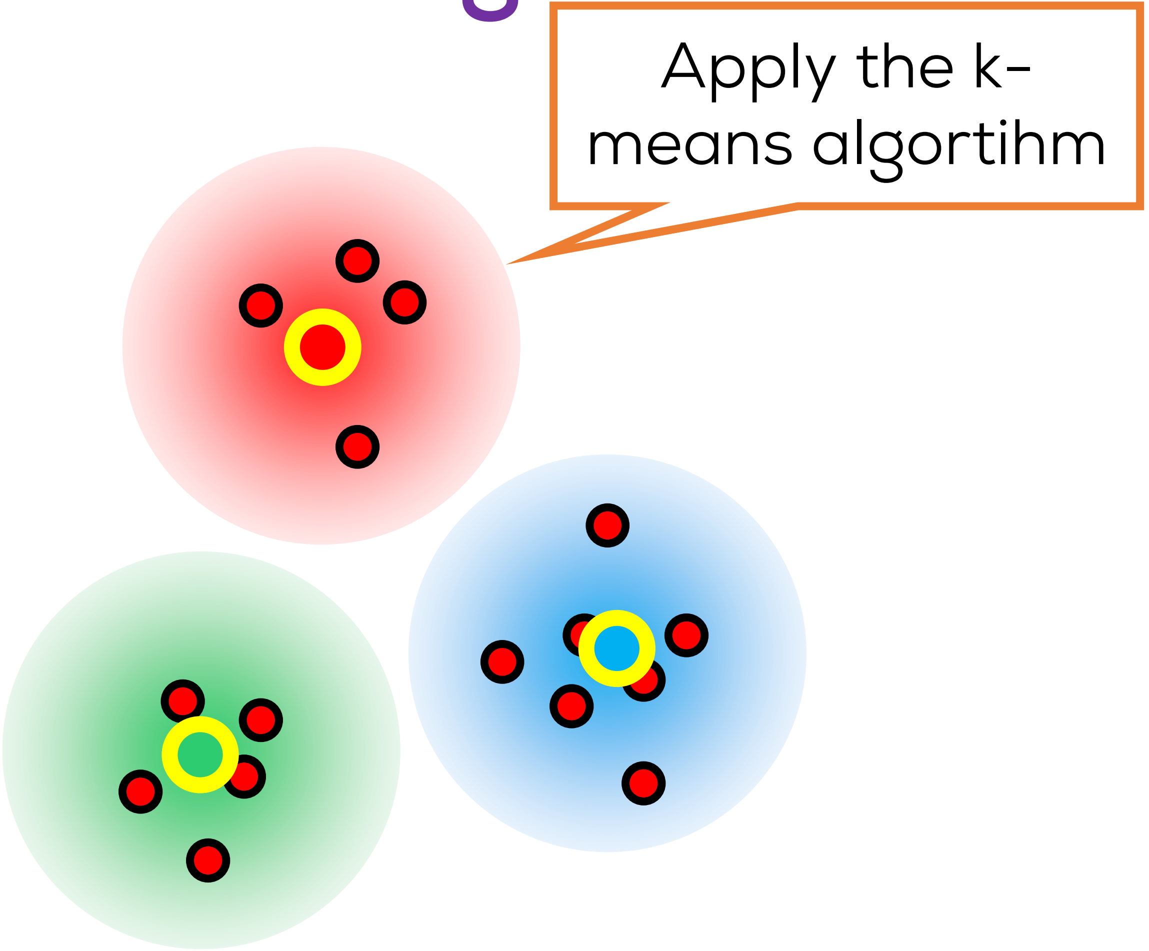
The Magic is not Always very Useful!

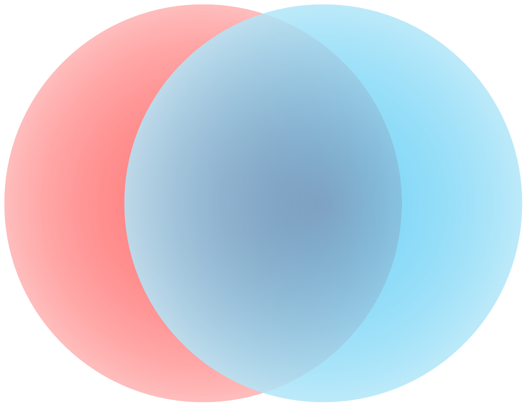


The Magic is not Always very Useful!



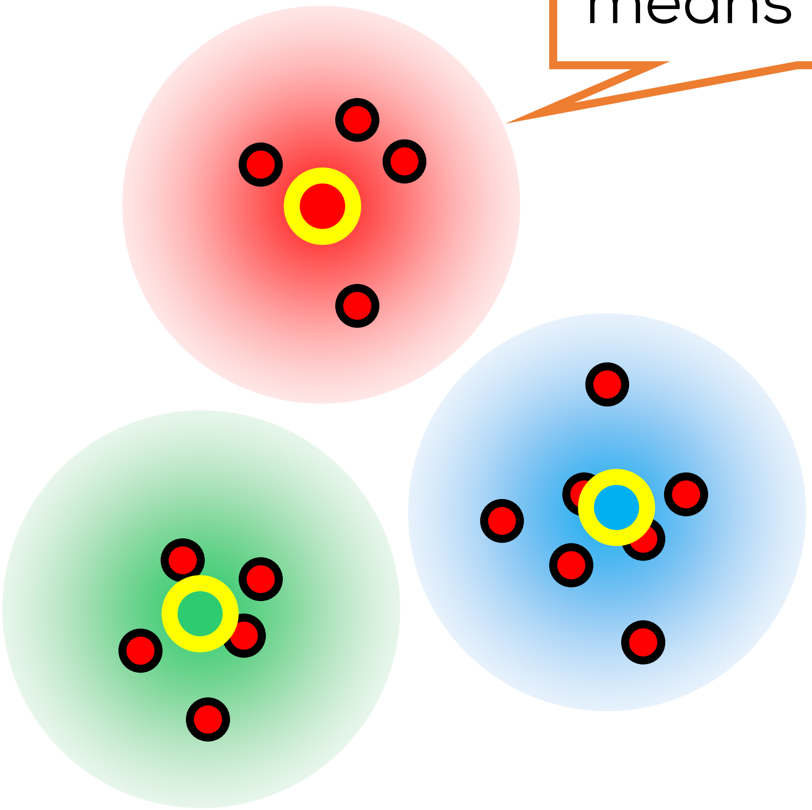
The Magic is not Always very Useful!

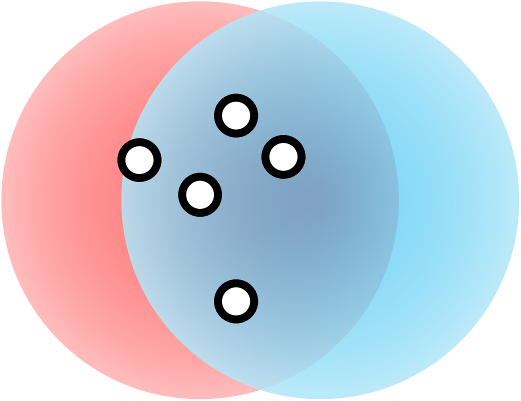




The Magic is not Always very Useful!

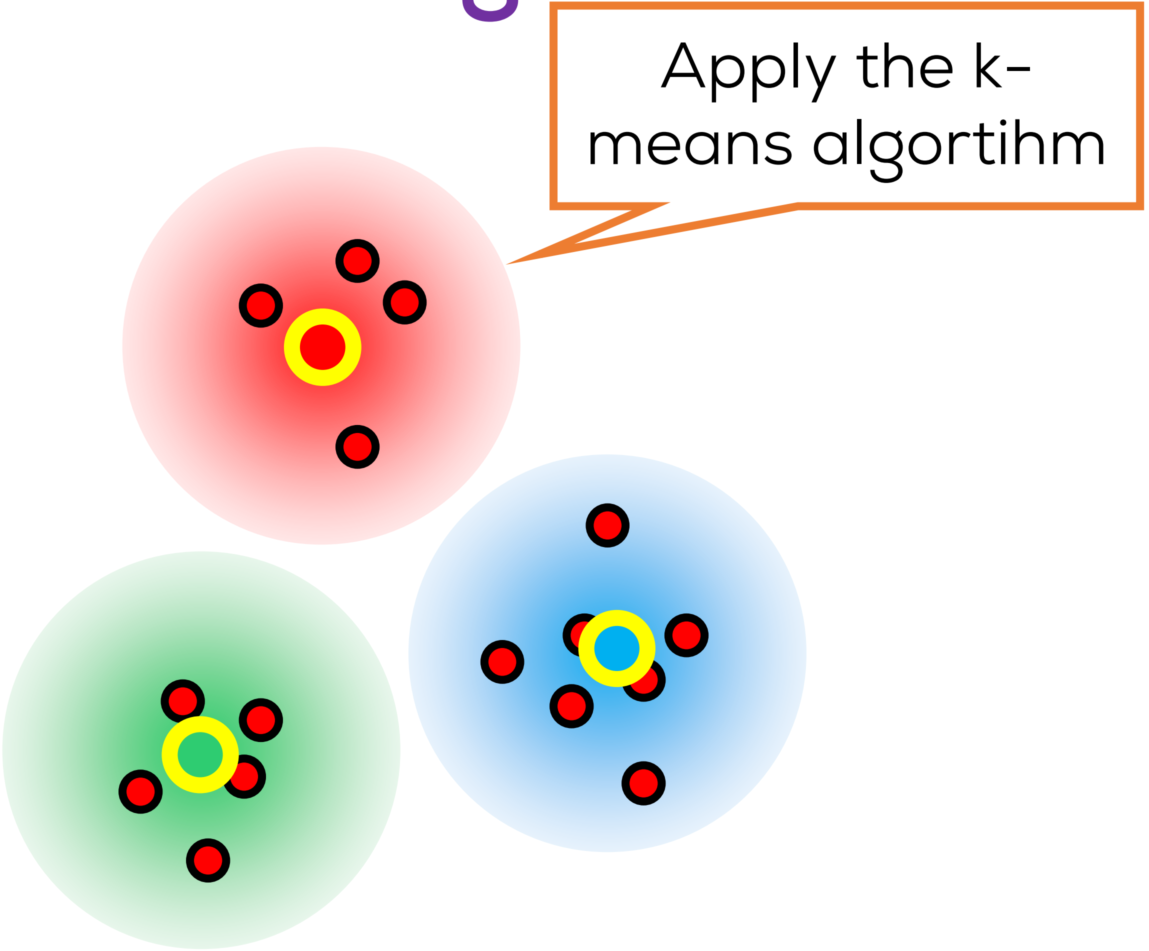
Apply the k-means algortihm

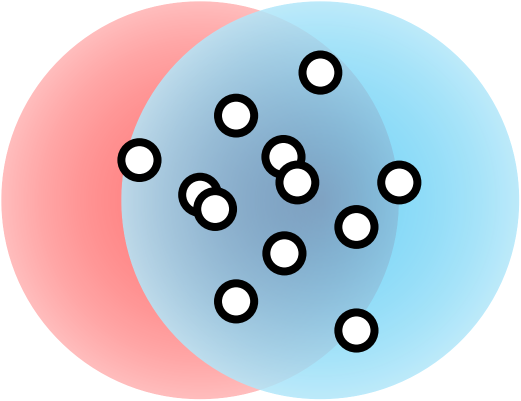




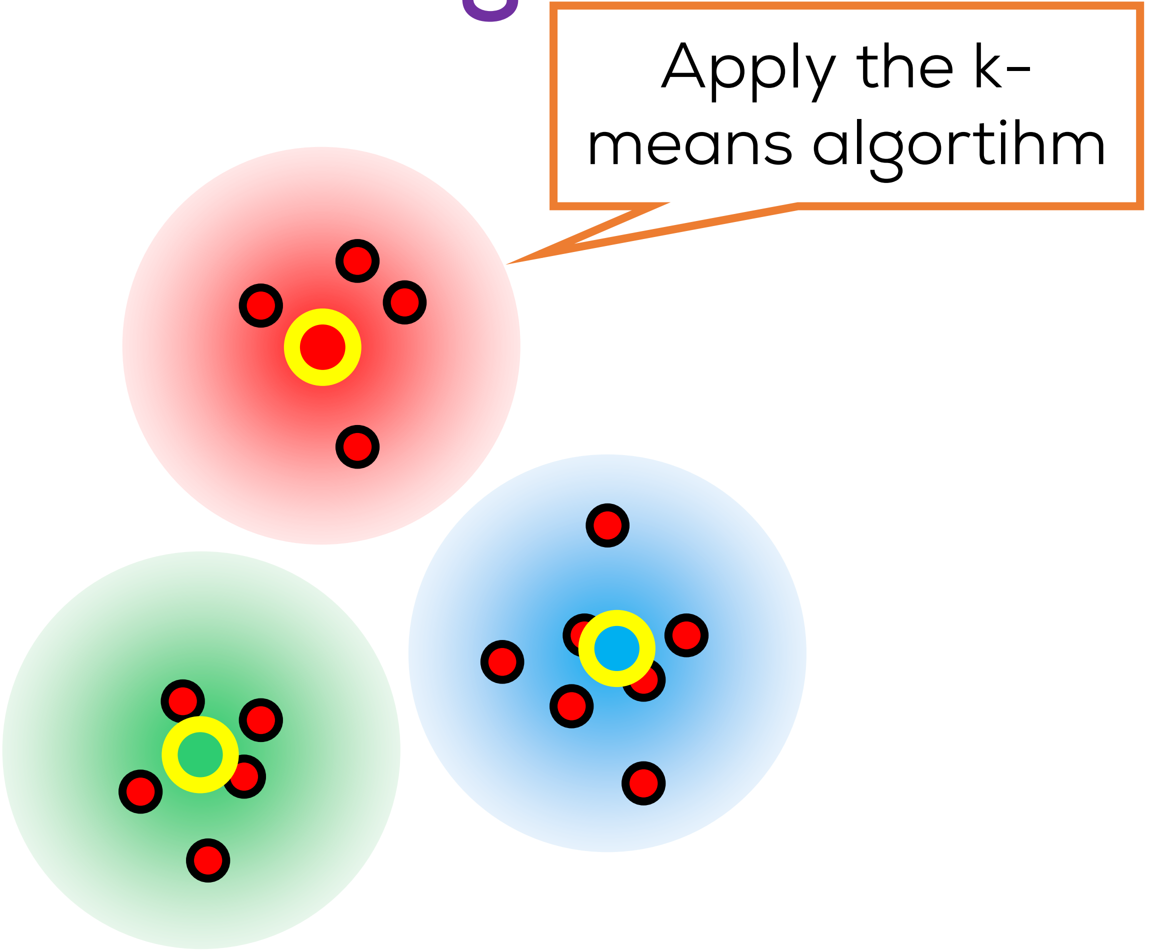
lotr.wikia.com

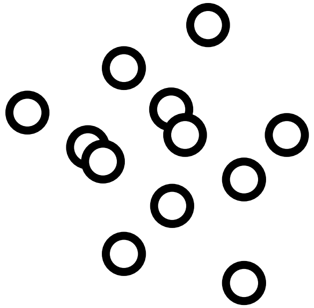
The Magic is not Always very Useful!



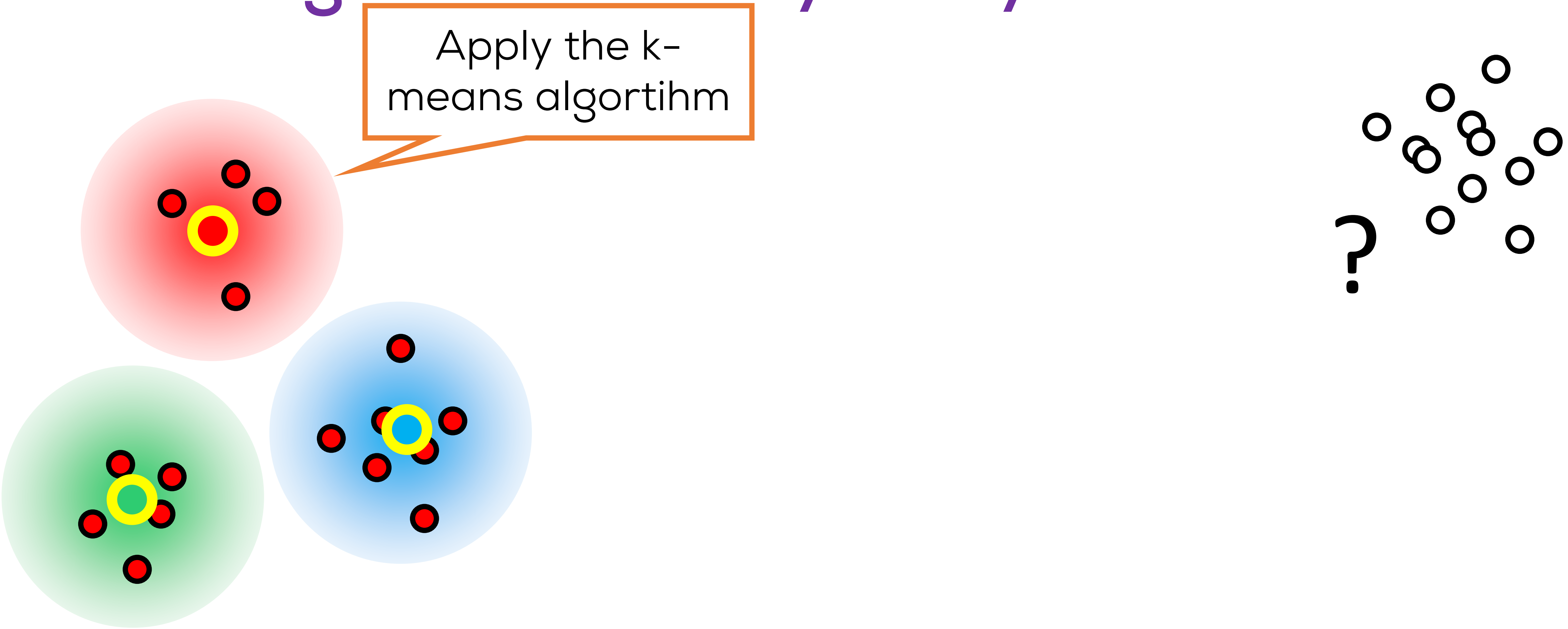


The Magic is not Always very Useful!



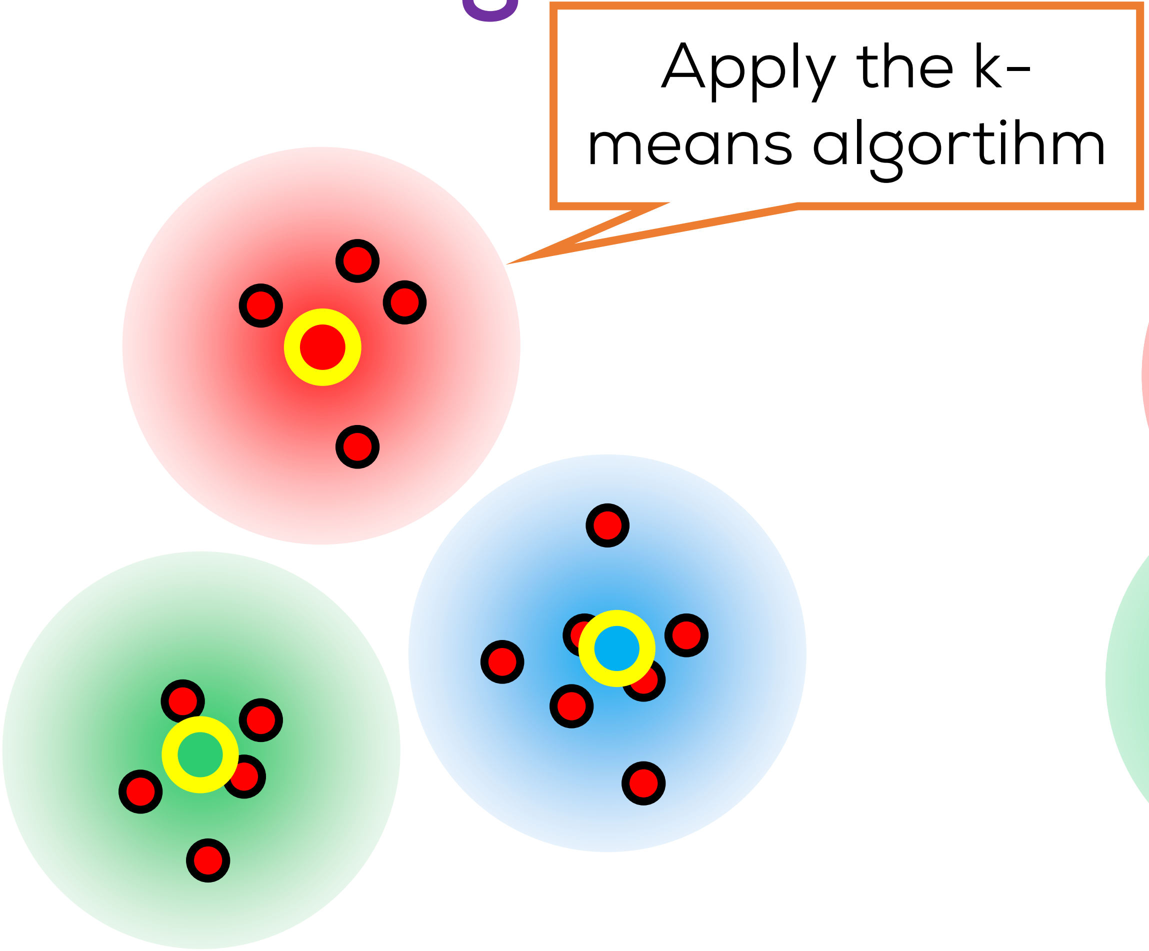


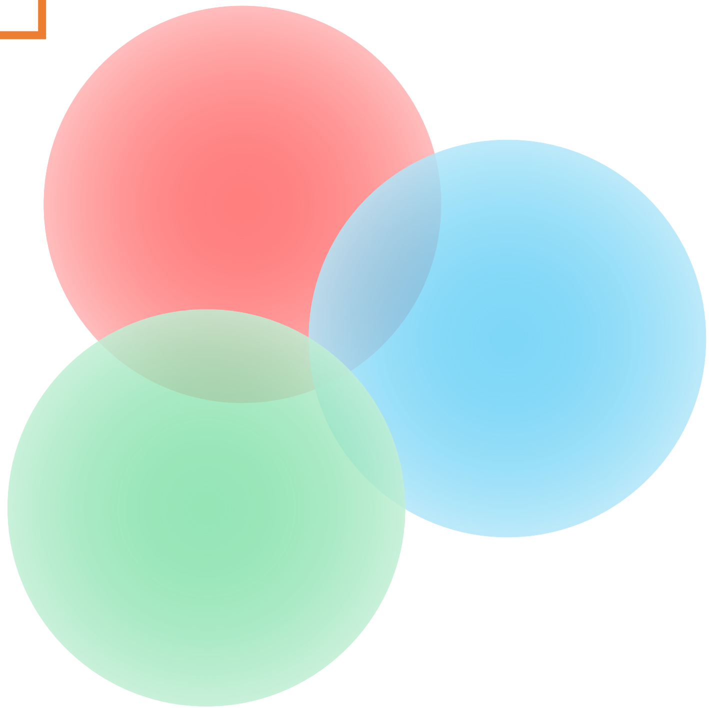
The Magic is not Always very Useful!

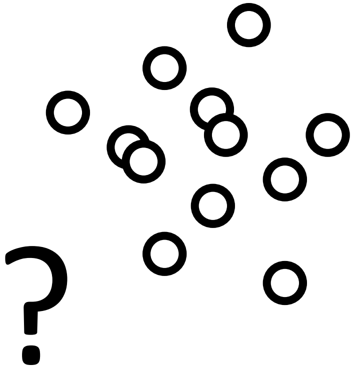


lotr.wikia.com

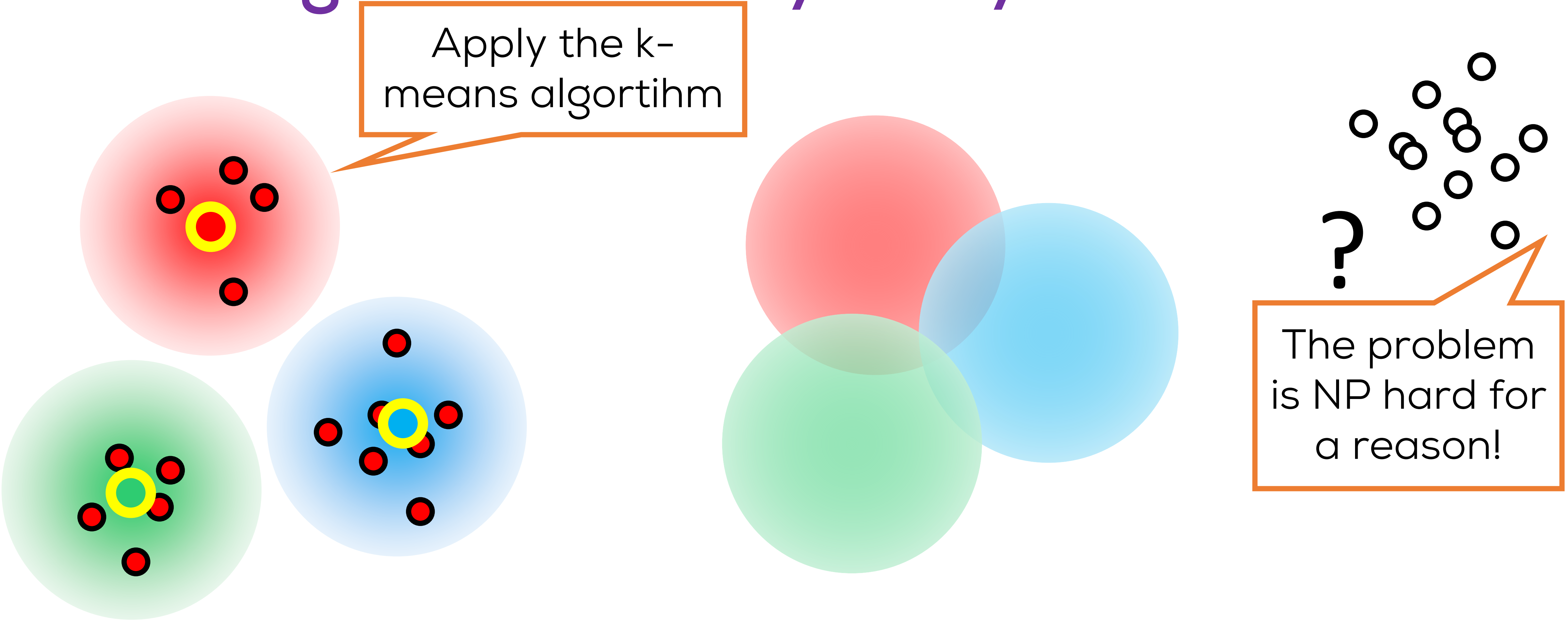
The Magic is not Always very Useful!







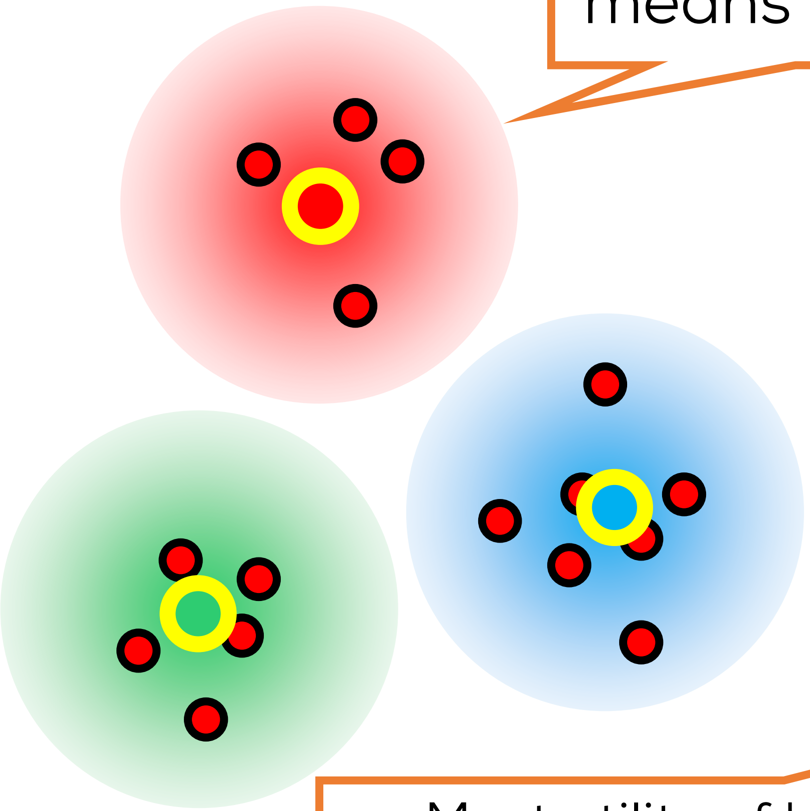
The Magic is not Always very Useful!



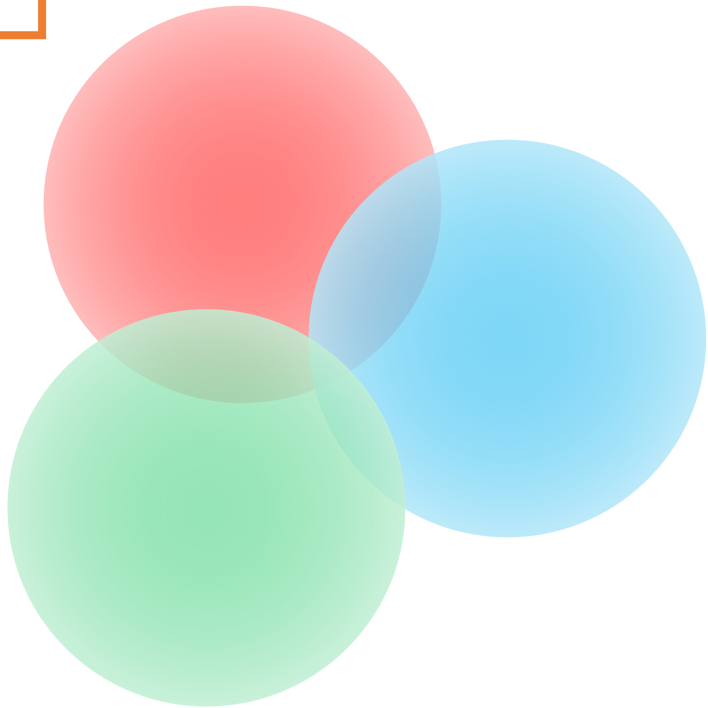
lotr.wikia.com

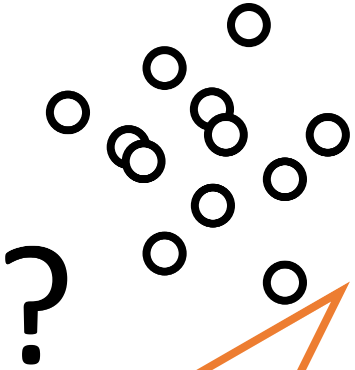
The Magic is not Always very Useful!

Apply the k-means algortihm



Most utility of k-means comes in problems that have some apparent stucture





The problem is NP hard for a reason!

Soft Assignment

The EM algorithm

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

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

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

A Ray of Hope

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

ALTERNATING OPTIMIZATION

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

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

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

A Ray of Hope

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

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

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

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

A Ray of Hope

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

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

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

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

A Ray of Hope

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

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

Bayes Rule!

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

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

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 $\boldsymbol{\Theta}^t$

$^{+1} = \underset{\boldsymbol{\Theta}}{\arg\max}\, \mathbb{P}\left[X, \{z^{i,t}\} \mid \boldsymbol{\Theta}\right]$

il convergence

May be throwing away a lot of information!

Bayes Rule!

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

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

A Ray of Hope

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

E.g. $\mathbb{P}[\,{\color{red}\bullet} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = 0.5$,
$\mathbb{P}[\,{\color{cyan}\bullet} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = 0.4$,
$\mathbb{P}[\,{\color{green}\bullet} \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = 0.1$,

ERNATING OPTIMIZATION

alize $\boldsymbol{\Theta}^0$

2. For $i \in [n]$ update $z^{i,t}$ using $\boldsymbol{\Theta}^t$

$^{+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \boldsymbol{\Theta}\right]$

til convergence

May be throwing away a lot of information!

Bayes Rule!

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

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

A Ray of Hope

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

E.g. $\mathbb{P}[\,\text{red}\,|\mathbf{x}^i, \mathbf{\Theta}^t] = 0.5$,
$\mathbb{P}[\,\text{blue}\,|\mathbf{x}^i, \mathbf{\Theta}^t] = 0.4$,
$\mathbb{P}[\,\text{green}\,|\mathbf{x}^i, \mathbf{\Theta}^t] = 0.1$,

ERNA … ATION

…alize …

Can we use $\mathbb{P}[k|\mathbf{x}^i, \mathbf{\Theta}^t]$ as weights instead?

2. For $i \in [n]$, update $z^{i,t}$ using $\mathbf{\Theta}^t$

… $^{+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$

May be throwing away a lot of information!

…til convergence

Bayes Rule!

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

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

A Ray of Hope

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

E.g. $\mathbb{P}[\,\text{●}|\mathbf{x}^i, \mathbf{\Theta}^t] = 0.5$,
$\mathbb{P}[\,\text{●}|\mathbf{x}^i, \mathbf{\Theta}^t] = 0.4$,
$\mathbb{P}[\,\text{●}|\mathbf{x}^i, \mathbf{\Theta}^t] = 0.1$,

ERNA ... ZATION

...alize ...

2. For $i \in [n]$ update $z^{i,t}$ using $\mathbf{\Theta}^t$

... $^{+1} = \arg\max_{\mathbf{\Theta}} \mathbb{P}\left[X, \{z^{i,t}\} \mid \mathbf{\Theta}\right]$

... til convergence

Can we use $\mathbb{P}[k|\mathbf{x}^i, \mathbf{\Theta}^t]$ as weights instead?

Assign point $\mathbf{x}^i$ to cluster k with weight $\propto \mathbb{P}[k|\mathbf{x}^i, \mathbf{\Theta}^t]$

May be throwing away a lot of information!

Bayes Rule!

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

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

A Ray of Hope

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

Assign point $\mathbf{x}^i$ to cluster k with weight $\propto \mathbb{P}[k|\mathbf{x}^i, \Theta^t]$

E.g. $\mathbb{P}[\,\bullet|\mathbf{x}^i, \Theta^t] = 0.5,$
$\mathbb{P}[\,\bullet|\mathbf{x}^i, \Theta^t] = 0.4,$
$\mathbb{P}[\,\bullet|\mathbf{x}^i, \Theta^t] = 0.1,$

Can we use $\mathbb{P}[k|\mathbf{x}^i, \Theta^t]$ as weights instead?

Has a "Bayesian" feel to it – use all available posterior information

...ERNA... ...AT...

...alize ...

2. For $i \in [n]$, update $z^{i,t}$ using Θ^t

... $^{+1} = \arg\max_{\Theta} \mathbb{P}\left[X, \{z^{i,t}\} \mid \Theta\right]$

May be throwing away a lot of information!

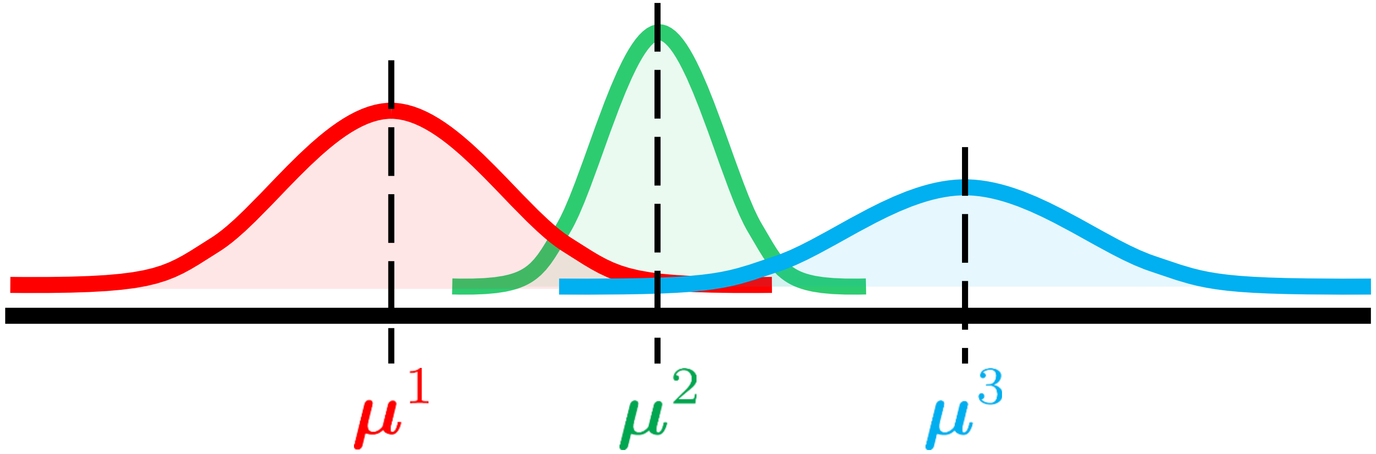
...peat until convergence

Bayes Rule!

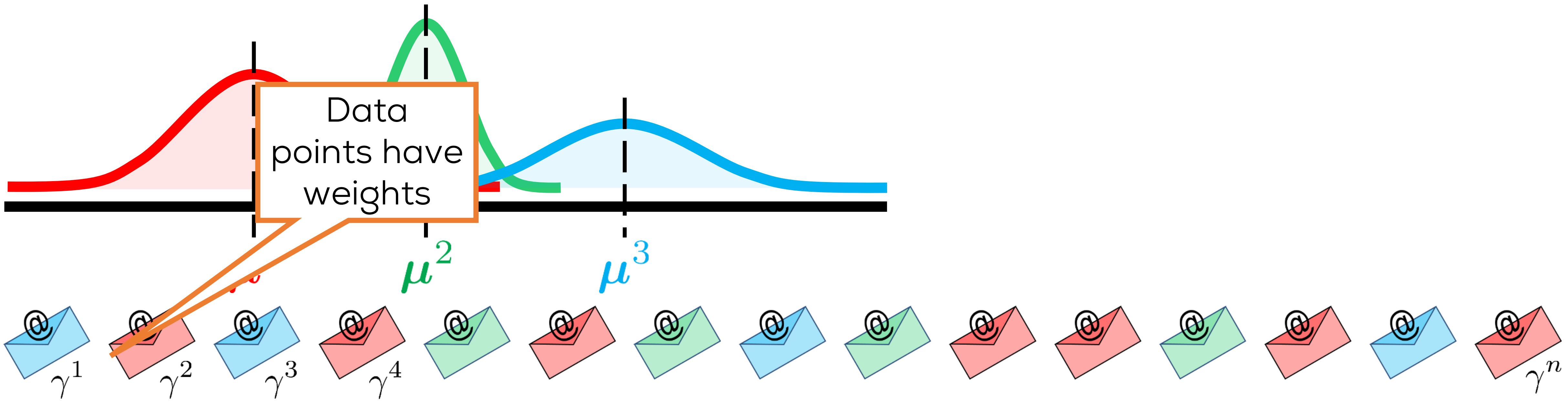
$$z^{i,t} = \arg\max_{k \in [K]} \mathbb{P}\left[k \mid \mathbf{x}^i, \Theta^t\right] = \arg\max_{k \in [K]} \quad \pi_k^t \quad \cdot \quad \mathcal{N}(\mathbf{x}^i \mid \mu^{k,t}, \Sigma^{k,t})$$

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

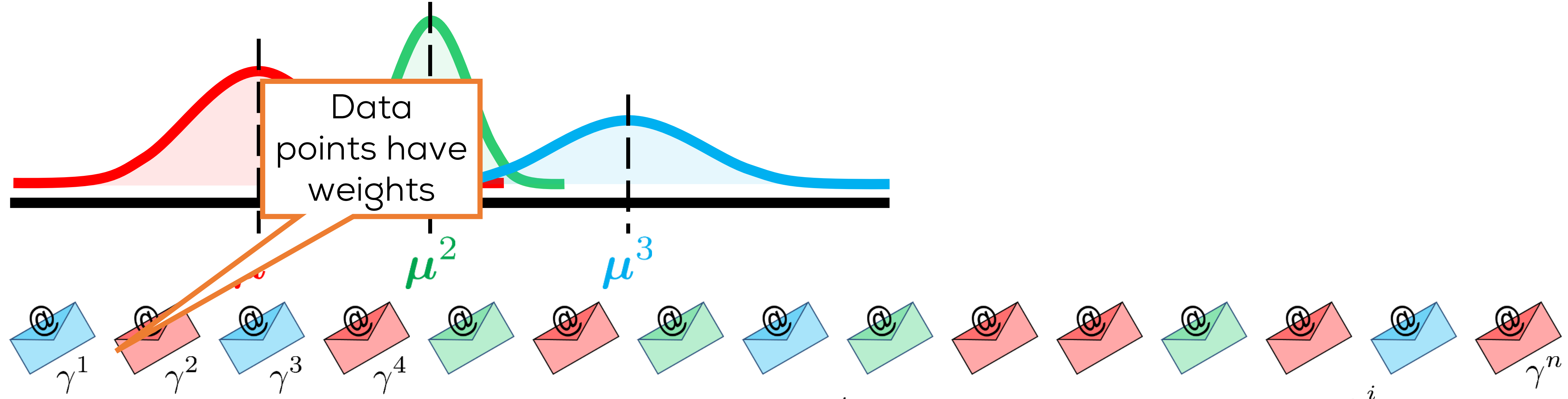
Weighted MLE



Weighted MLE

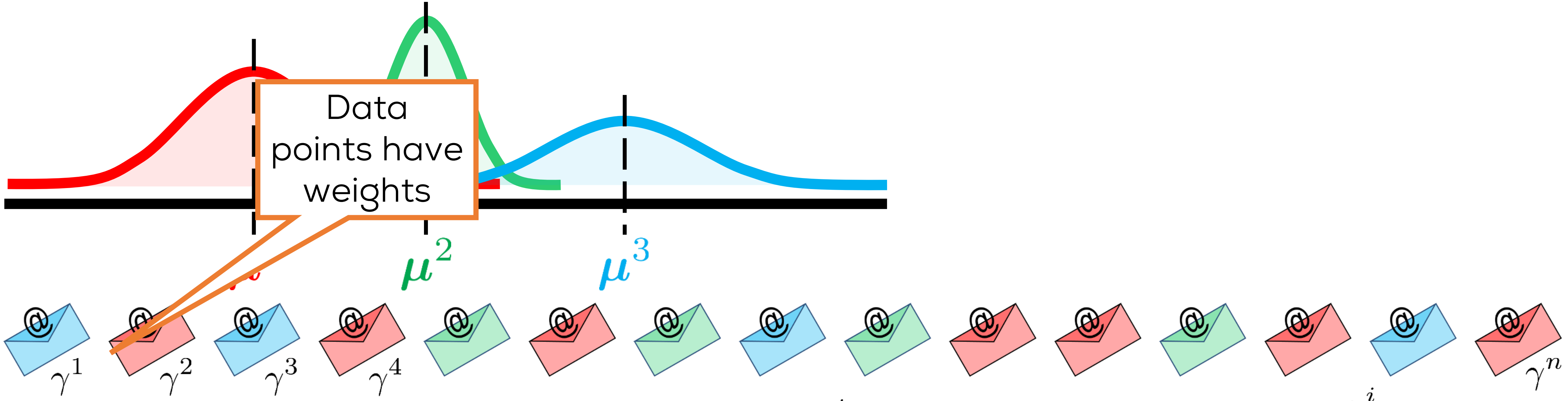


Weighted MLE



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

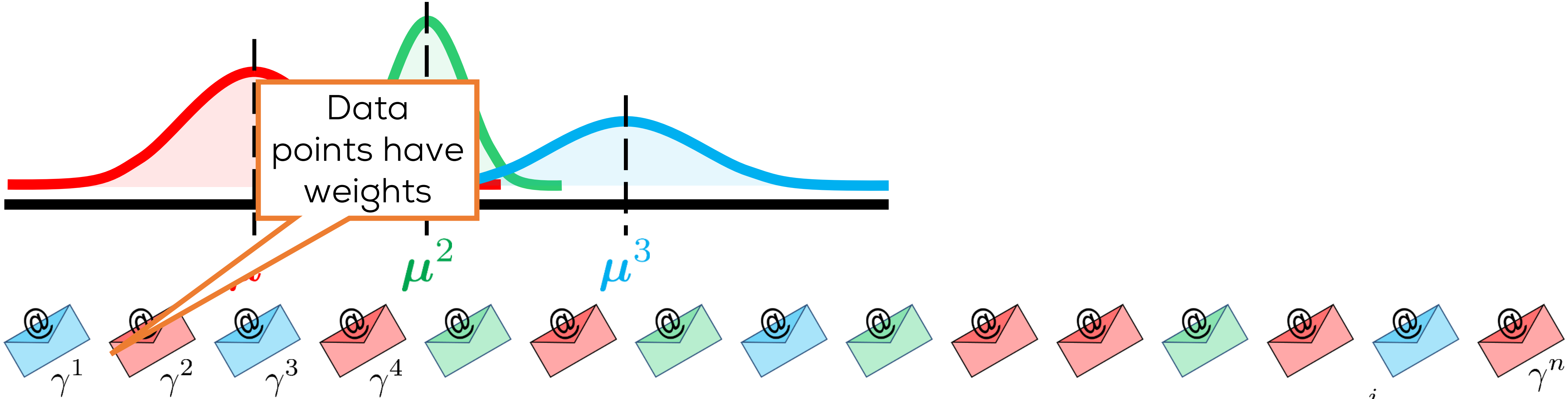
Weighted MLE



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

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

Weighted MLE

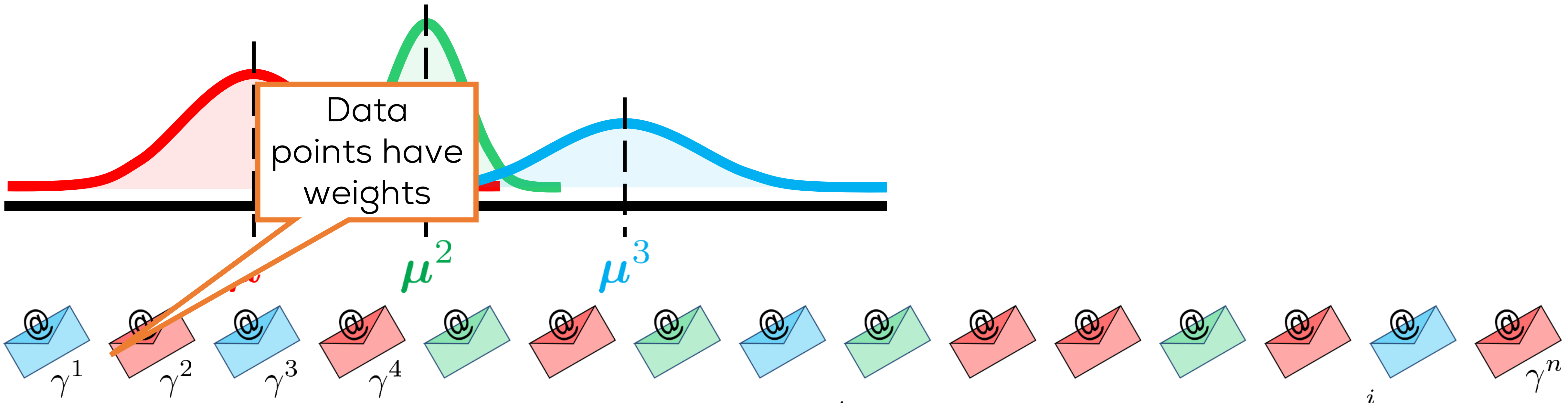


$$\mathbb{P}\left[\mathbf{x}^i, z^i, \gamma^i \,|\, \boldsymbol{\Theta}\right] = \left(\mathbb{P}\left[z^i \,|\, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \,|\, z^i, \boldsymbol{\Theta}\right]\right)^{\gamma^i} = \left(\boldsymbol{\pi}_{z^i} \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^{z^i},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

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

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

Weighted MLE



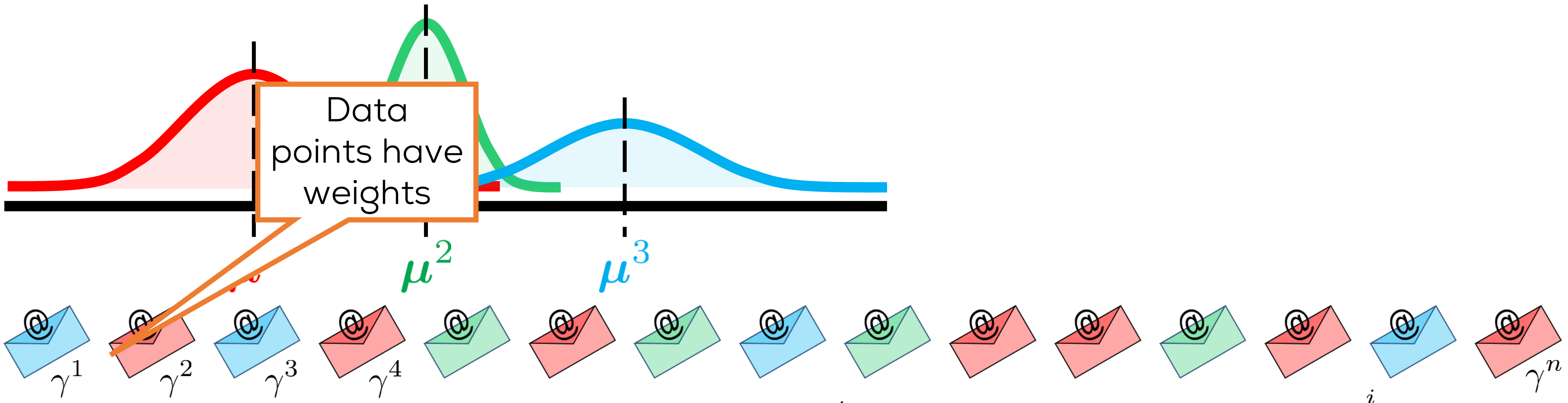
$$\mathbb{P}\left[\mathbf{x}^i, z^i, \gamma^i \,|\, \boldsymbol{\Theta}\right] = \left(\mathbb{P}\left[z^i \,|\, \boldsymbol{\Theta}\right] \cdot \mathbb{P}\left[\mathbf{x}^i \,|\, z^i, \boldsymbol{\Theta}\right]\right)^{\gamma^i} = \left(\boldsymbol{\pi}_{z^i} \cdot \mathcal{N}(\mathbf{x}^i \,|\, \boldsymbol{\mu}^{z^i},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

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

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

Take log and apply 1st order optimality

Weighted MLE



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

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

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

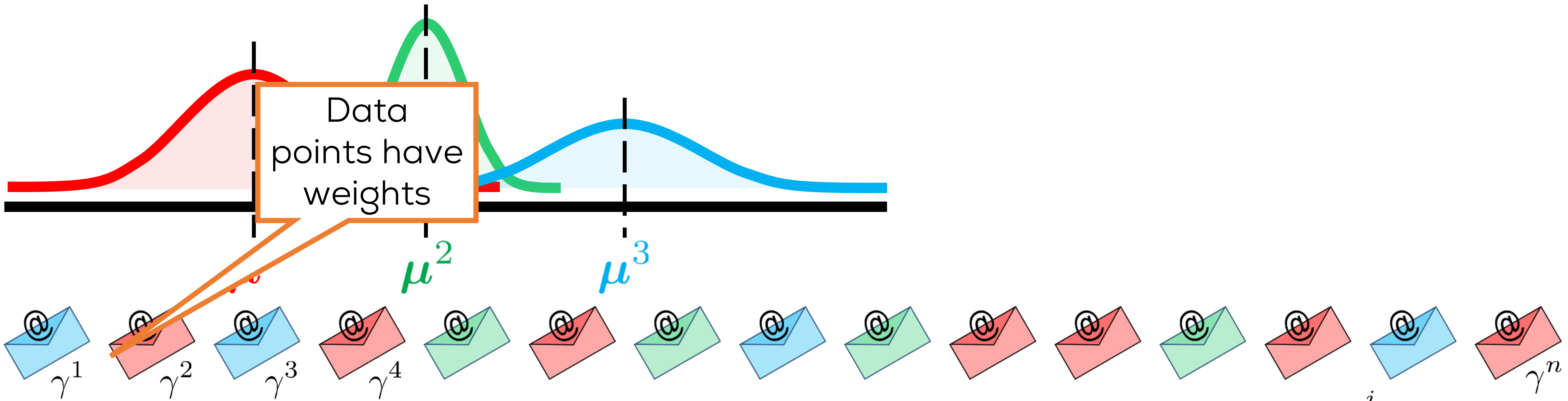
Take log and apply 1st order optimality

$$\boldsymbol{\pi}_1^{\mathrm{MLE}} = \frac{\text{\# eff. red emails}}{\text{\# eff. total emails}} = \frac{\tilde{n}_r}{n} = \frac{\sum_{i:z^i=\bullet} \gamma^i}{\sum_{j=1}^{n} \gamma^j}$$

$$\boldsymbol{\mu}_{\mathrm{MLE}}^1 = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot \mathbf{x}^i$$

$$\Sigma_{\mathrm{MLE}}^1 = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot \left(\mathbf{x}^i - \boldsymbol{\mu}_{MLE}^1\right)\left(\mathbf{x}^i - \boldsymbol{\mu}_{MLE}^1\right)^{\top}$$

Weighted MLE



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

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

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

Take log and apply 1st order optimality

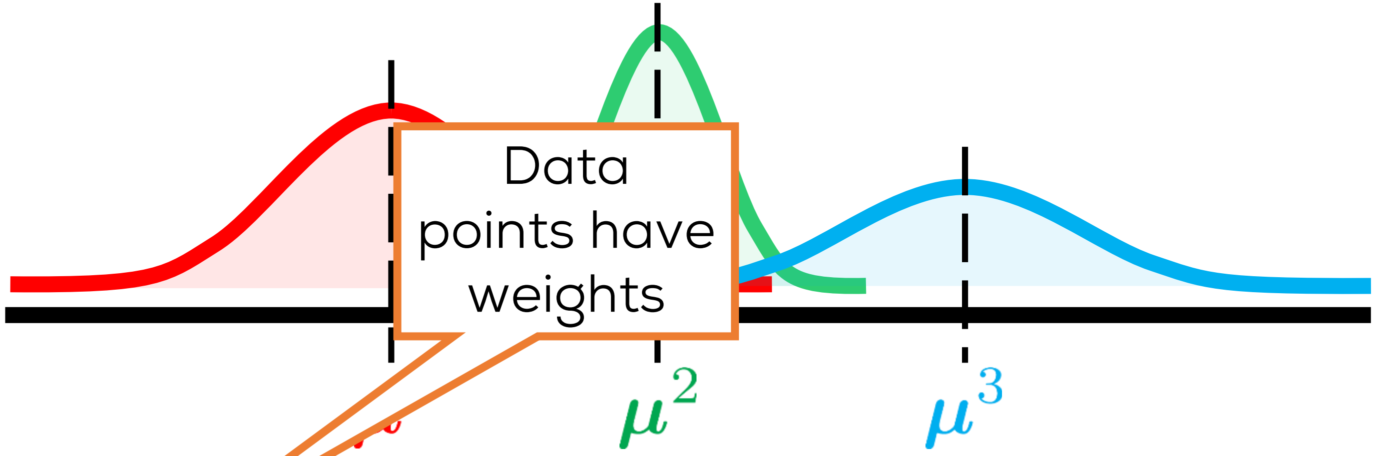
$$\boldsymbol{\pi}_1^{\mathrm{MLE}} = \frac{\#\text{ eff. red emails}}{\#\text{ eff. total emails}} = \frac{\tilde{n}_r}{n} = \frac{\sum_{i:z^i={\color{red}\bullet}} \gamma^i}{\sum_{j=1}^{n} \gamma^j}$$

$$\boldsymbol{\mu}^1_{\mathrm{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i={\color{red}\bullet}} \gamma^i \cdot \mathbf{x}^i$$

$$\Sigma^1_{\mathrm{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i={\color{red}\bullet}} \gamma^i \cdot \left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)\left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)^\top$$

Exercise!

Weighted MLE



Reduces to normal MLE if $\gamma^i = 1$ for all i

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

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

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

$$\boldsymbol{\pi}_1^{\text{MLE}} = \frac{\text{\# eff. red emails}}{\text{\# eff. total emails}} = \frac{\tilde{n}_r}{n} = \frac{\sum_{i:z^i=\bullet} \gamma^i}{\sum_{j=1}^{n} \gamma^j}$$

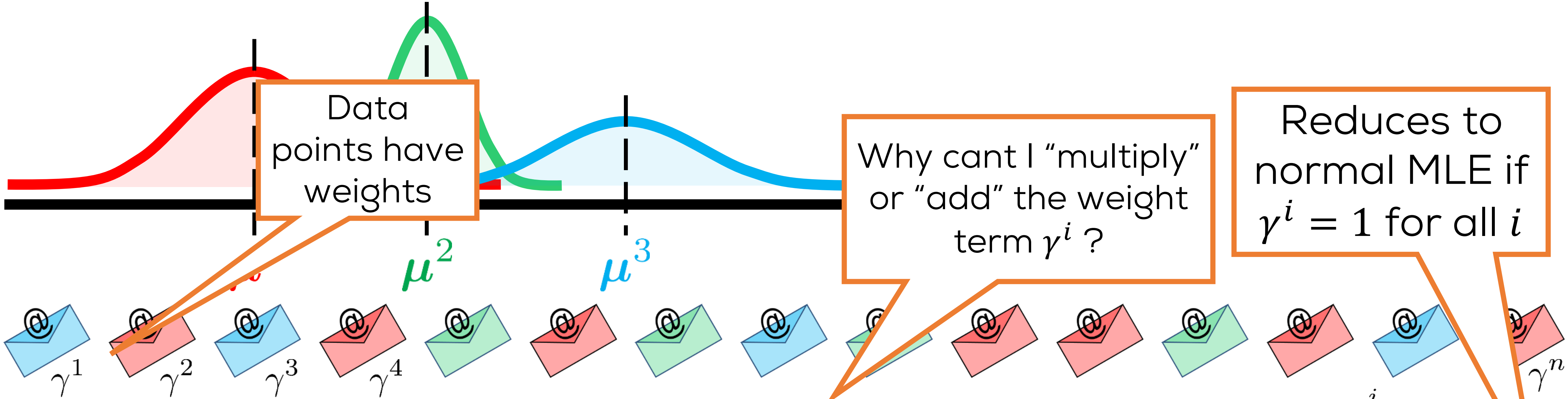
$$\boldsymbol{\mu}_{\text{MLE}}^1 = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot \mathbf{x}^i$$

$$\Sigma_{\text{MLE}}^1 = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot (\mathbf{x}^i - \boldsymbol{\mu}_{MLE}^1)(\mathbf{x}^i - \boldsymbol{\mu}_{MLE}^1)^{\mathsf{T}}$$

Take log and apply 1st order optimality

Exercise!

Weighted MLE



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

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

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

Take log and apply 1st order optimality

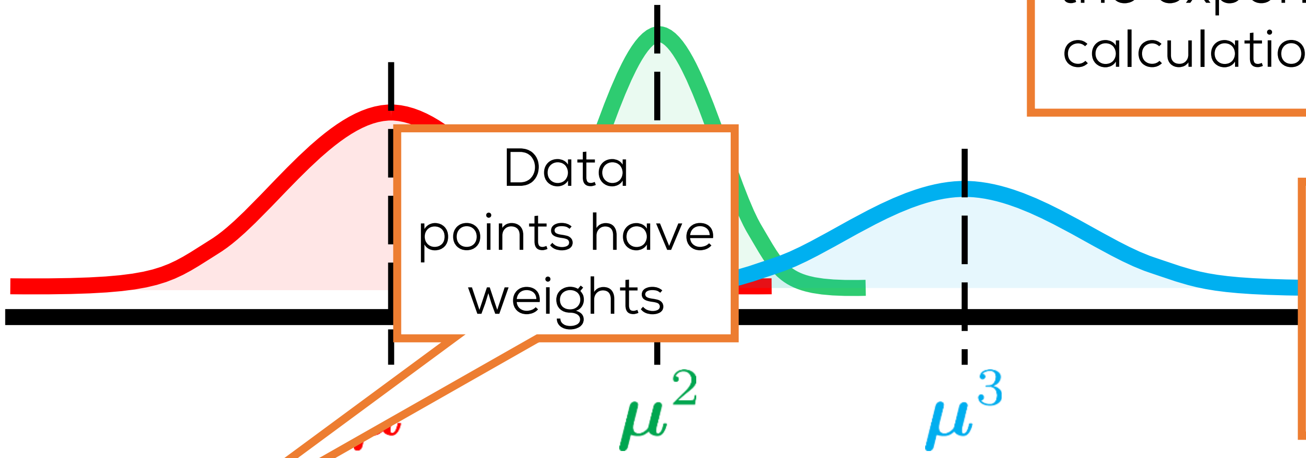
$$\boldsymbol{\pi}_1^{\text{MLE}} = \frac{\text{\# eff. red emails}}{\text{\# eff. total emails}} = \frac{\tilde{n}_r}{n} = \frac{\sum_{i:z^i=\bullet} \gamma^i}{\sum_{j=1}^{n} \gamma^j}$$

$$\boldsymbol{\mu}^1_{\text{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot \mathbf{x}^i$$

$$\Sigma^1_{\text{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot \left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)\left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)^{\mathsf{T}}$$

Exercise!

Weighted MLE



Data points have weights

Incorporating γ^i into the exponent makes calculations simpler

Why cant I "multiply" or "add" the weight term γ^i ?

Reduces to normal MLE if $\gamma^i = 1$ for all i

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

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

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

Take log and apply 1st order optimality

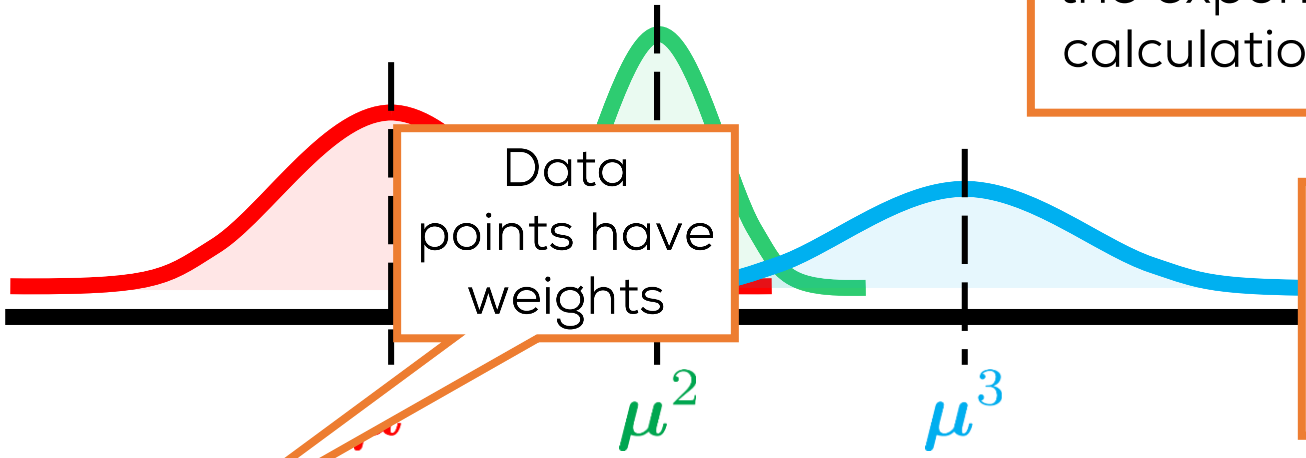
$$\boldsymbol{\pi}_1^{\text{MLE}} = \frac{\text{\# eff. red emails}}{\text{\# eff. total emails}} = \frac{\tilde{n}_r}{n} = \frac{\sum_{i:z^i=\bullet} \gamma^i}{\sum_{j=1}^{n} \gamma^j}$$

$$\boldsymbol{\mu}^1_{\text{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i=\bullet} \gamma^i \cdot \mathbf{x}^i$$

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

Exercise!

Weighted MLE



Data points have weights

Incorporating γ^i into the exponent makes calculations simpler

Also makes more sense (when we study the EM algo)

Why cant I "multiply" or "add" the weight term γ^i ?

Reduces to normal MLE if $\gamma^i = 1$ for all i

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

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

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

Take log and apply 1st order optimality

Exercise!

$$\boldsymbol{\pi}_1^{\mathrm{MLE}} = \frac{\text{\# eff. red emails}}{\text{\# eff. total emails}} = \frac{\tilde{n}_r}{n} = \frac{\sum_{i:z^i = \bullet} \gamma^i}{\sum_{j=1}^{n} \gamma^j}$$

$$\boldsymbol{\mu}^1_{\mathrm{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i = \bullet} \gamma^i \cdot \mathbf{x}^i$$

$$\Sigma^1_{\mathrm{MLE}} = \frac{1}{\tilde{n}_r} \sum_{i:z^i = \bullet} \gamma^i \cdot \left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)\left(\mathbf{x}^i - \boldsymbol{\mu}^1_{MLE}\right)^\top$$

Hard Alternating Minimization

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 $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[X, \left\{z^{i,t}\right\} \mid \boldsymbol{\Theta}\right]$
4. Repeat until convergence

Soft Alternating Minimization

ALTERNATING OPTIMIZATION

1. For $i \in [n]$, create k copies of the data point
   1. Let $\mathbf{x}^i \to \left\{\mathbf{x}^{\{i,1\}}, \mathbf{x}^{\{i,2\}}, \dots, \mathbf{x}^{\{i,k\}}\right\}$
   2. Assign the k-th copy label k i.e. $z^{\{i,k\}} = k$
2. Initialize $\boldsymbol{\Theta}^0$
3. Update weights $\gamma^{i,k,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $\gamma^{i,k,t} = \mathbb{P}[k \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = \dfrac{\boldsymbol{\pi}_k^t \cdot \mathcal{N}\left(\mathbf{X}^i \mid \boldsymbol{\mu}^{k,t}, \Sigma^{k,t}\right)}{\sum_j \boldsymbol{\pi}_j^t \cdot \mathcal{N}\left(\mathbf{X}^i \mid \boldsymbol{\mu}^{j,t}, \Sigma^{j,t}\right)}$
4. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[\left\{x^{\{i,k\}}\right\}, \left\{z^{\{i,k\}}\right\}, \left\{\gamma^{\{i,k,t\}}\right\} \mid \boldsymbol{\Theta}\right]$
5. Repeat until convergence

Soft Alternating Minimization

ALTERNATING OPTIMIZATION

1. For $i \in [n]$, create k copies of the data point
   1. Let $\mathbf{x}^i \to \left\{\mathbf{x}^{\{i,1\}}, \mathbf{x}^{\{i,2\}}, \ldots, \mathbf{x}^{\{i,k\}}\right\}$
   2. Assign the k-th copy label k i.e. $z^{\{i,k\}} = k$
2. Initialize $\boldsymbol{\Theta}^0$
3. Update weights $\gamma^{i,k,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $\gamma^{i,k,t} = \mathbb{P}[k \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = \dfrac{\boldsymbol{\pi}_k^t \cdot \mathcal{N}\left(\mathbf{X}^i \mid \boldsymbol{\mu}^{k,t}, \Sigma^{k,t}\right)}{\sum_j \boldsymbol{\pi}_j^t \cdot \mathcal{N}\left(\mathbf{X}^i \mid \boldsymbol{\mu}^{j,t}, \Sigma^{j,t}\right)}$
4. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[\left\{x^{\{i,k\}}\right\}, \left\{z^{\{i,k\}}\right\}, \left\{\gamma^{\{i,k,t\}}\right\} \mid \boldsymbol{\Theta}\right]$
5. Repeat until convergence

$\sum_j \gamma^{\{i,j,t\}} = 1$ for all i and all t

Soft Alternating Minimization

ALTERNATING OPTIMIZATION

1. For $i \in [n]$, create k copies of the data point
   1. Let $\mathbf{x}^i \to \left\{\mathbf{x}^{\{i,1\}}, \mathbf{x}^{\{i,2\}}, \ldots, \mathbf{x}^{\{i,k\}}\right\}$
   2. Assign the k-th copy label k i.e. $z^{\{i,k\}} = k$
2. Initialize $\boldsymbol{\Theta}^0$
3. Update weights $\gamma^{i,k,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $\gamma^{i,k,t} = \mathbb{P}[k \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = \dfrac{\boldsymbol{\pi}_k^t \cdot \mathcal{N}(\mathbf{X}^i \mid \boldsymbol{\mu}^{k,t}, \Sigma^{k,t})}{\sum_j \boldsymbol{\pi}_j^t \cdot \mathcal{N}(\mathbf{X}^i \mid \boldsymbol{\mu}^{j,t}, \Sigma^{j,t})}$
4. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[\left\{x^{\{i,k\}}\right\}, \left\{z^{\{i,k\}}\right\}, \left\{\gamma^{\{i,k,t\}}\right\} \mid \boldsymbol{\Theta}\right]$
5. Repeat until convergence

Distribute a unit weight across clusters

$\sum_j \gamma^{\{i,j,t\}} = 1$ for all i and all t

Soft Alternating Minimization

Hard-AM: all weight was on a single cluster

Distribute a unit weight across clusters

$\sum_j \gamma^{\{i,j,t\}} = 1$ for all i and all t

ALTERNATING OPTIMIZATION

1. For $i \in [n]$, create k copies of the data poin
   1. Let $\mathbf{x}^i \rightarrow \left\{\mathbf{x}^{\{i,1\}}, \mathbf{x}^{\{i,2\}}, \ldots, \mathbf{x}^{\{i,k\}}\right\}$
   2. Assign the k-th copy label k i.e. $z^{\{i,k\}} = k$
2. Initialize $\boldsymbol{\Theta}^0$
3. Update weights $\gamma^{i,k,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $\gamma^{i,k,t} = \mathbb{P}[k \mid \mathbf{x}^i, \boldsymbol{\Theta}^t] = \dfrac{\boldsymbol{\pi}_k^t \cdot \mathcal{N}(\mathbf{X}^i \mid \boldsymbol{\mu}^{k,t}, \Sigma^{k,t})}{\sum_j \boldsymbol{\pi}_j^t \cdot \mathcal{N}(\mathbf{X}^i \mid \boldsymbol{\mu}^{j,t}, \Sigma^{j,t})}$
4. Update $\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathbb{P}\left[\left\{x^{\{i,k\}}\right\}, \left\{z^{\{i,k\}}\right\}, \left\{\gamma^{\{i,k,t\}}\right\} \mid \boldsymbol{\Theta}\right]$
5. Repeat until convergence

Soft k-means Algorithm

- 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 ALGO

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For all i, 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. Let $n_k^t = \left|i: z^{\{i,t\}} = k\right|$
4. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
5. Repeat until convergence

SOFT K-MEANS ALGO

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For all i, update $\gamma^{i,k,t}$ using $\boldsymbol{\mu}^{k,t}$

   Let $\gamma^{i,k,t} = \frac{\exp\left(-\frac{\left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2}{2}\right)}{\sum_j \exp\left(-\frac{\left\|\mathbf{x}^i - \boldsymbol{\mu}^{j,t}\right\|_2^2}{2}\right)}$
3. Let $\tilde{n}_k^t = \sum_i \gamma^{\{i,k,t\}}$
4. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{\tilde{n}_k^t}\sum_i \gamma^{\{i,k,t\}} \cdot \mathbf{x}^i$
5. Repeat until convergence

Soft k-means Algorithm

- 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 ALGO

1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For all i, 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. Let $n_k^t = \left|i : z^{\{i,t\}} = k\right|$
4. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{n_k^t}\sum_{i:z^{i,t}=k} \mathbf{x}^i$
5. Repeat until convergence

SOFT K-MEANS ALGO

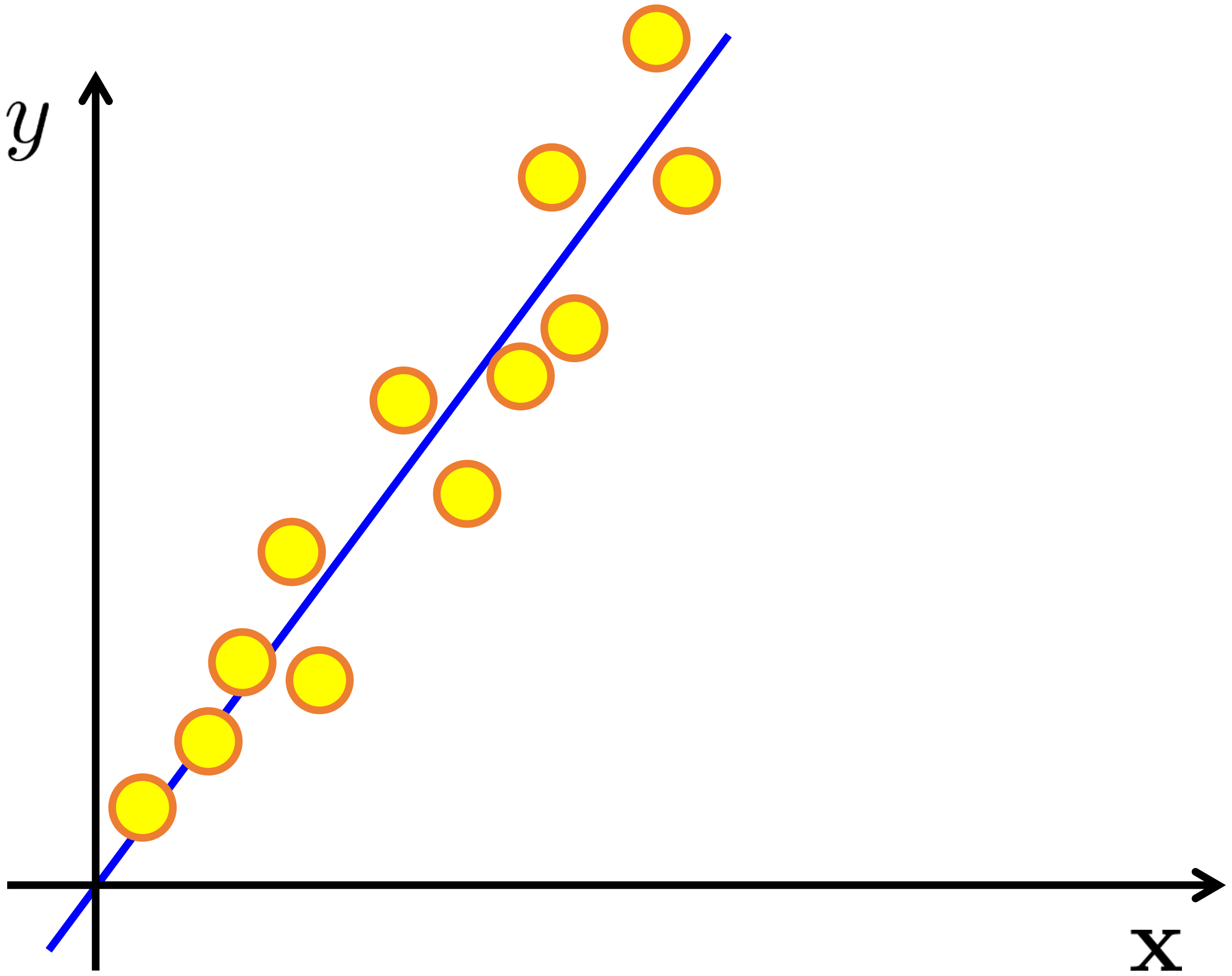
1. Initialize means $\{\boldsymbol{\mu}^{k,0}\}_{k=1,\ldots K}$
2. For all i, update $\gamma^{i,k,t}$ using $\boldsymbol{\mu}^{k,t}$
   Let $\gamma^{i,k,t} = \dfrac{\exp\left(-\frac{\left\|\mathbf{x}^i - \boldsymbol{\mu}^{k,t}\right\|_2^2}{2}\right)}{\sum_j \exp\left(-\frac{\left\|\mathbf{x}^i - \boldsymbol{\mu}^{j,t}\right\|_2^2}{2}\right)}$
3. Let $\tilde{n}_k^t = \sum_i \gamma^{\{i,k,t\}}$
4. Update $\boldsymbol{\mu}^{k,t+1} = \frac{1}{\tilde{n}_k^t}\sum_i \gamma^{\{i,k,t\}} \cdot \mathbf{x}^i$
5. Repeat until convergence

Mixed Regression

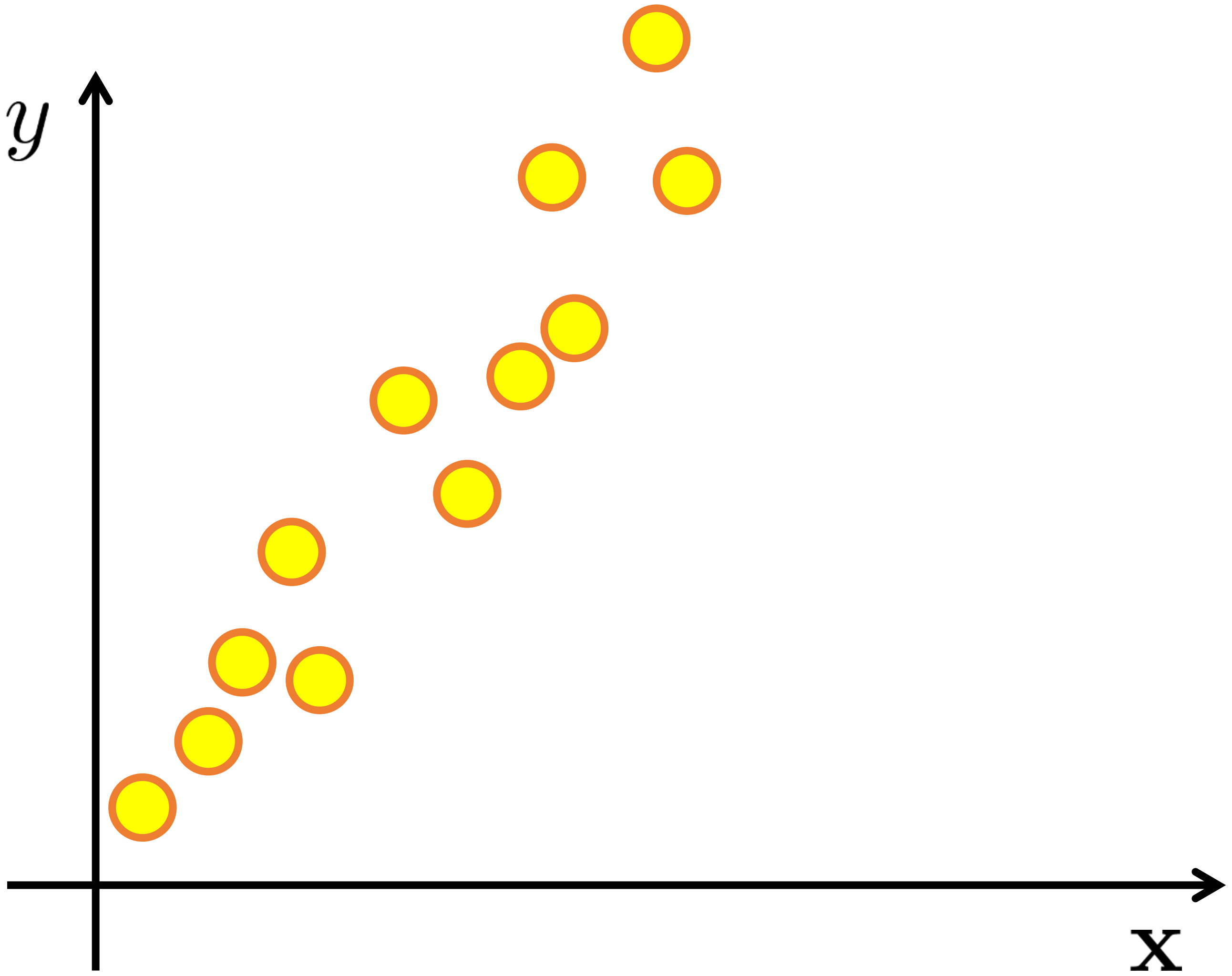
Mixed Regression

Mixed Regression

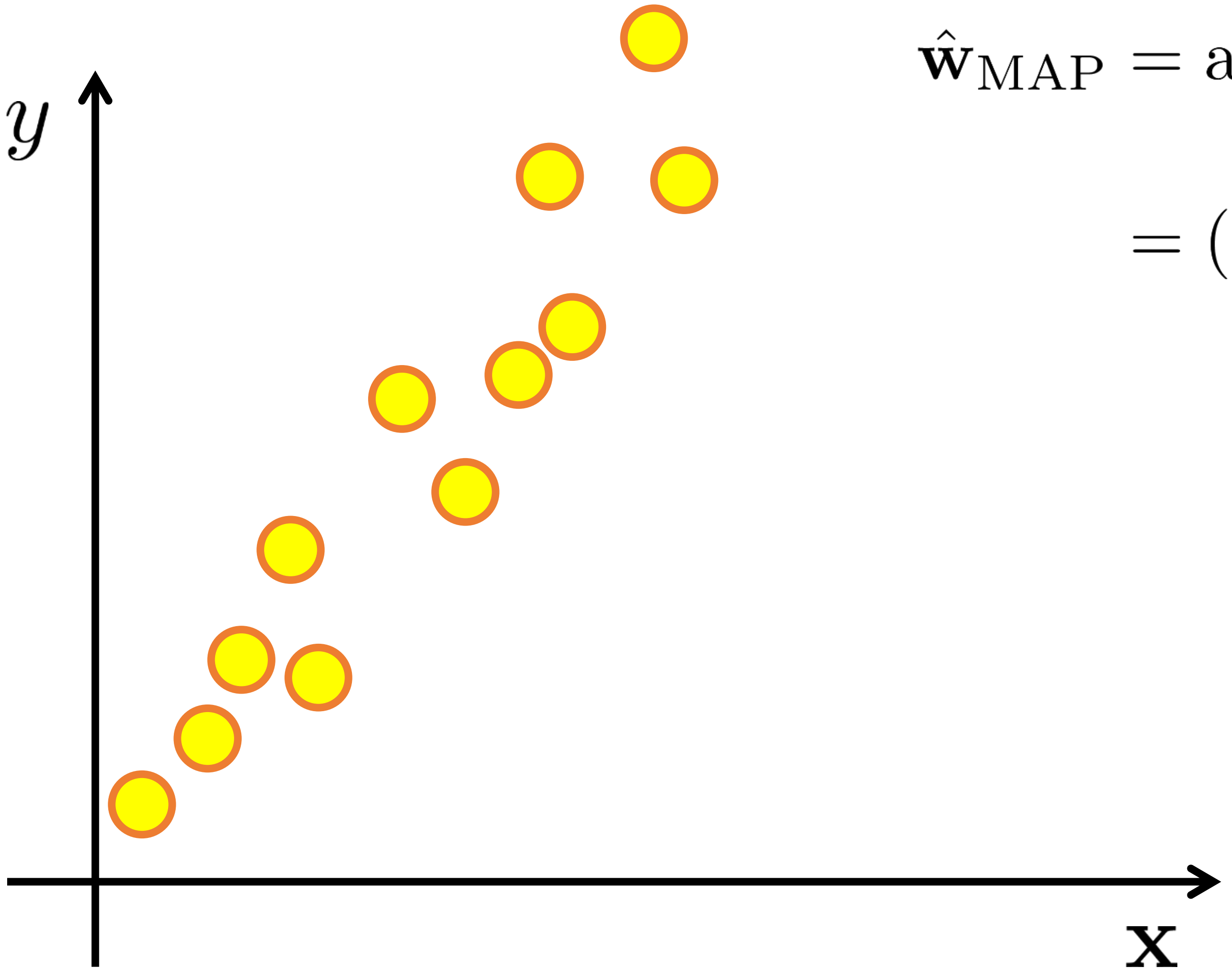
Mixed Regression



Mixed Regression



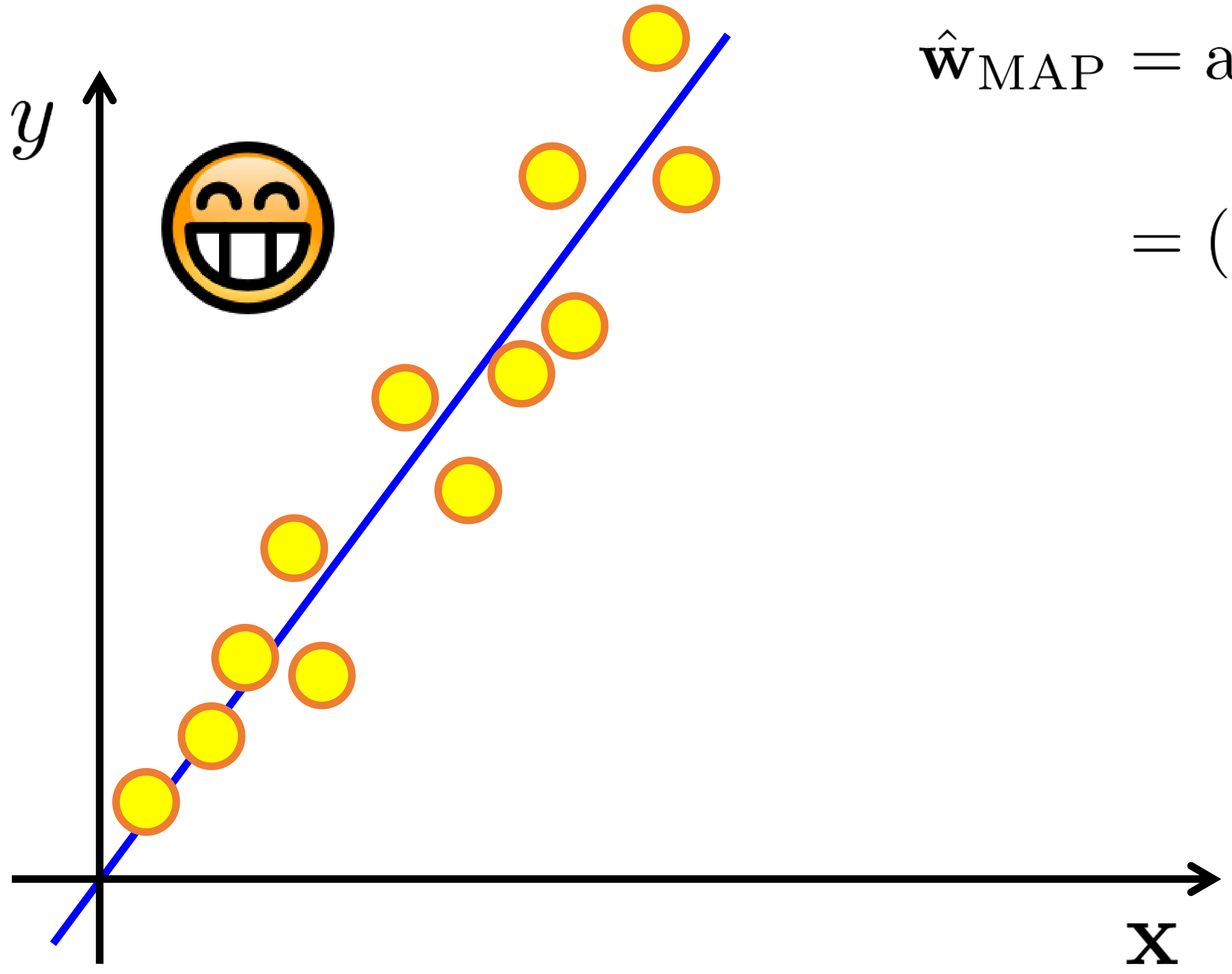
Mixed Regression



$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$$

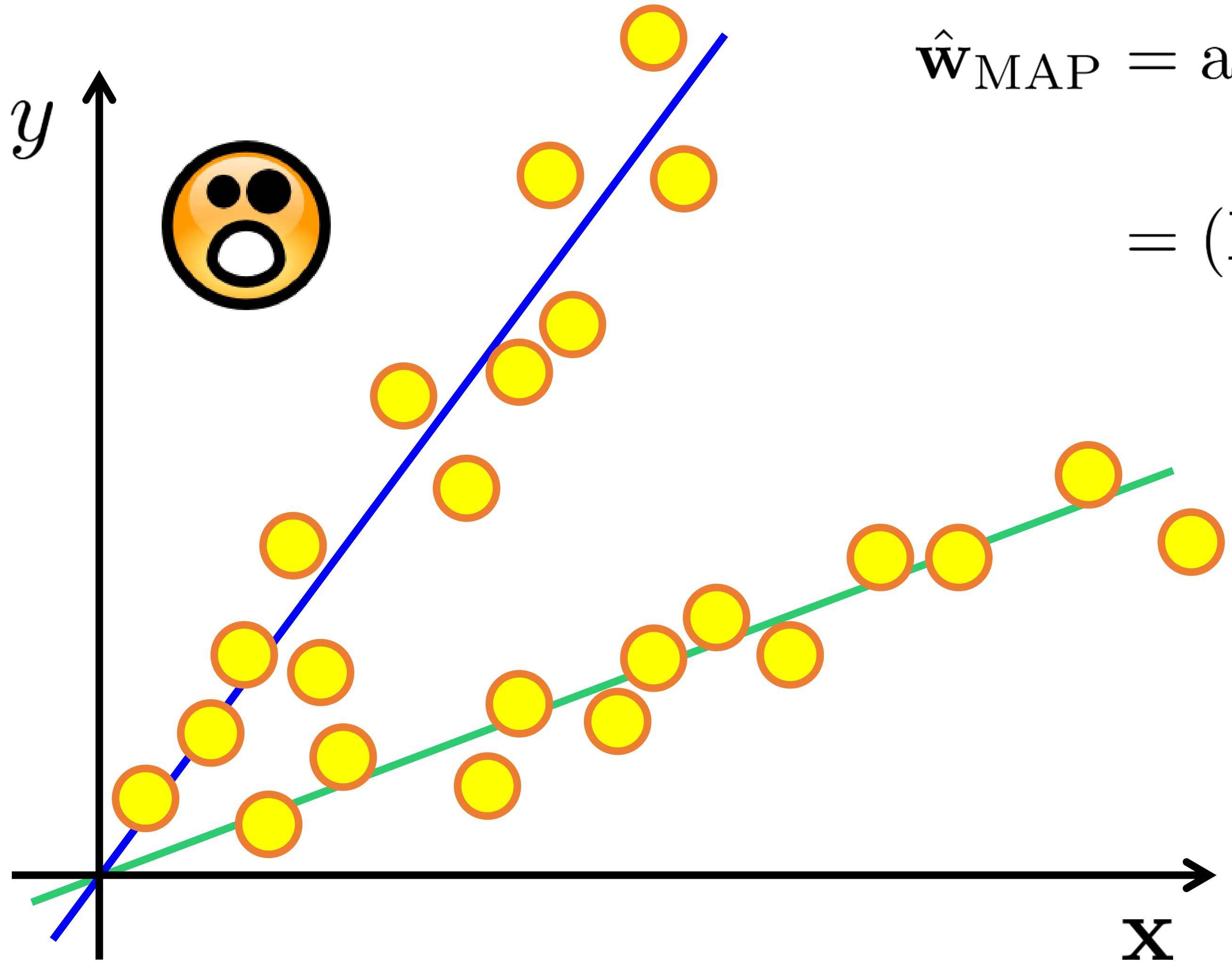
Mixed Regression



$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

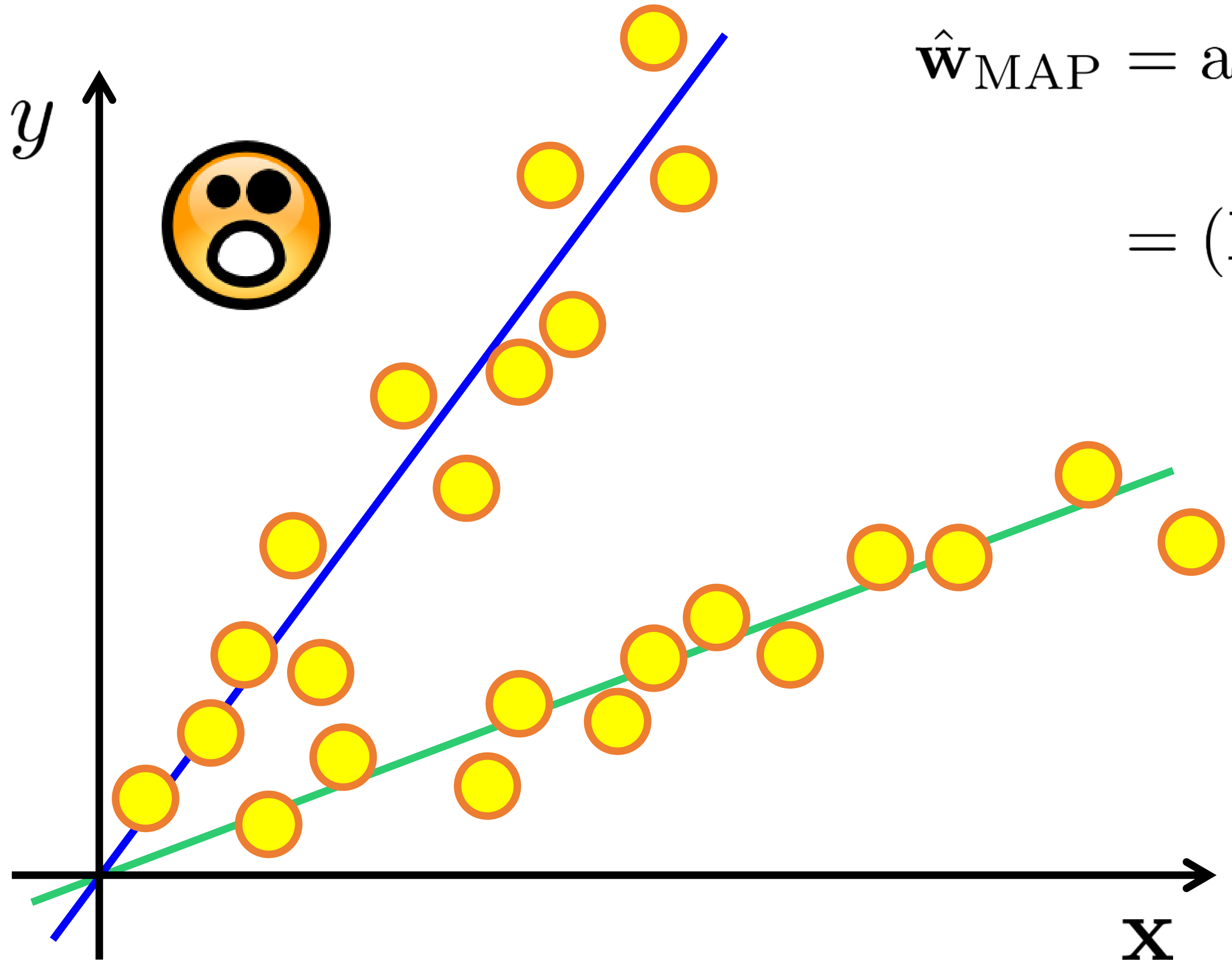
$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$$

Mixed Regression



$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$
$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1} \mathbf{X}\mathbf{y}$$

Mixed Regression

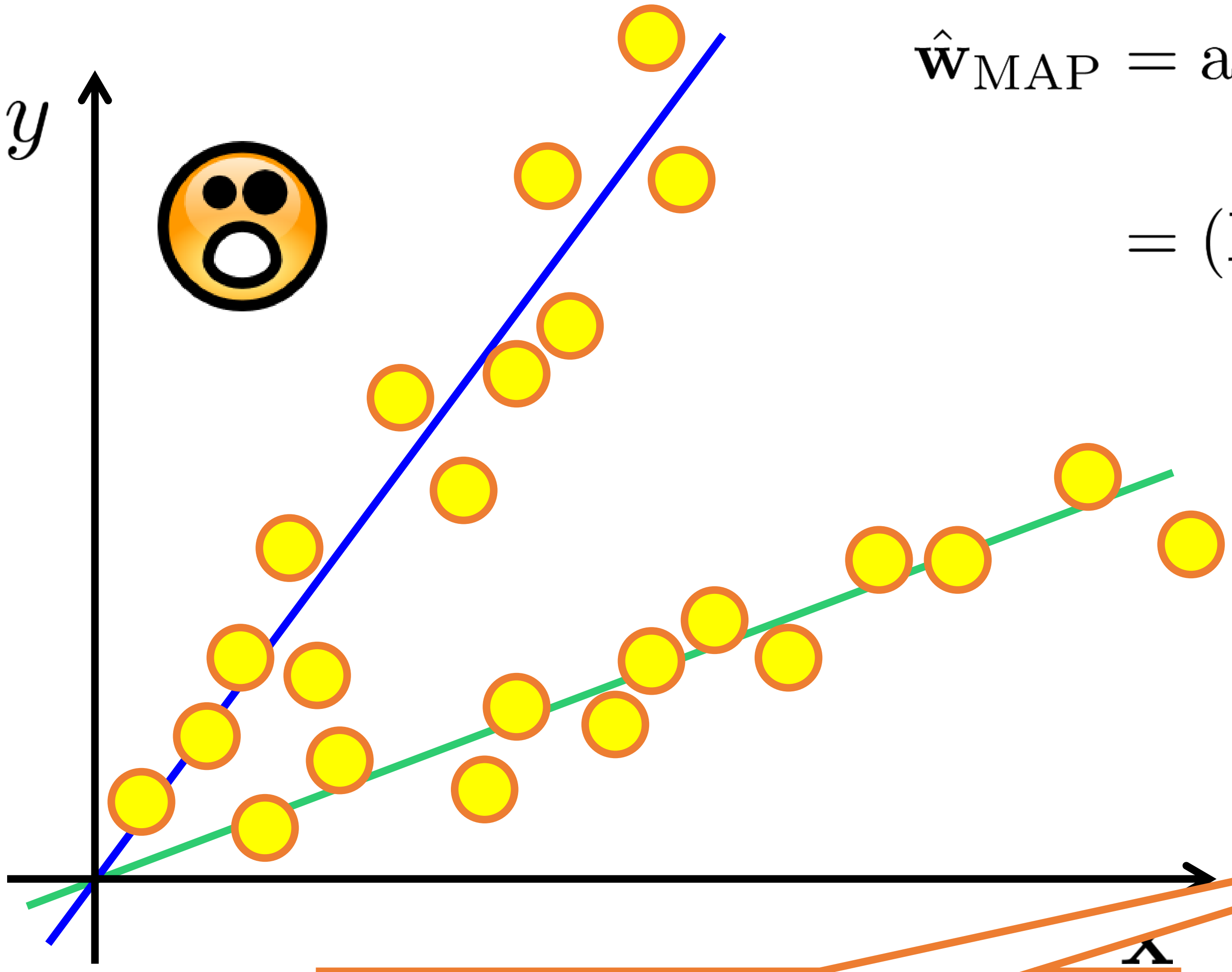


$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$
$$= \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y}$$

Regression with a classification touch ☺

E.g. recommendation systems, a different system might work for young and old but we don't know age of users

Mixed Regression



$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$
$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$$

Regression with a classification touch ☺

E.g. recommendation systems, a different system might work for young and old but we don't know age of users

Mixed models, Mixture of Experts

Mixed Regression

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, z^i, \mathbf{w}\right] = \mathcal{N}(\left\langle \mathbf{w}^{z^i}, \mathbf{x}^i \right\rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \left\langle \mathbf{w}^{z^i}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2}\right)$$

- Assume for sake of simplicity that both models are equally sampled $\mathbb{P}[z = 0] = \mathbb{P}[z = 1] = 0.5$
- Assume for sake of simplicity that Gaussian noise $\sigma_1 = \sigma_2 = 1$
- Only unknowns are the two linear models $\mathbf{w}^1, \mathbf{w}^2$

Mixed Regression

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, z^i, \mathbf{w}\right] = \mathcal{N}(\left\langle \mathbf{w}^{z^i}, \mathbf{x}^i \right\rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \left\langle \mathbf{w}^{z^i}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2} \right)$$

$z^i \in \{0,1\}$ indicates which line generated the data

- Assume for sake of simplicity that both models are equally sampled $\mathbb{P}[z = 0] = \mathbb{P}[z = 1] = 0.5$
- Assume for sake of simplicity that Gaussian noise $\sigma_1 = \sigma_2 = 1$
- Only unknowns are the two linear models $\mathbf{w}^1, \mathbf{w}^2$

Mixed Regression

$$\mathbb{P}\left[y \,|\, \mathbf{x}^i, z^i, \mathbf{w}\right] = \mathcal{N}(\left\langle \mathbf{w}^{z^i}, \mathbf{x}^i \right\rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \left\langle \mathbf{w}^{z^i}, \mathbf{x}^i \right\rangle)^2}{2\sigma^2}\right)$$

$z^i \in \{0,1\}$ indicates which line generated the data

z^i is a latent variable!

- Assume for sake of simplicity that both models are equally sampled $\mathbb{P}[z = 0] = \mathbb{P}[z = 1] = 0.5$
- Assume for sake of simplicity that Gaussian noise $\sigma_1 = \sigma_2 = 1$
- Only unknowns are the two linear models $\mathbf{w}^1, \mathbf{w}^2$

Hard Mixed Regression

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $z^{i,t} = \arg\max_{k \in \{0,1\}} \mathcal{N}\left(y^i \mid \mathbf{x}^i, \mathbf{w}^{k,t}\right)$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_{i : z^{\{i,t\}} = k} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

Hard Mixed Regression

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $z^{i,t} = \arg\max_{k\in\{0,1\}} \mathcal{N}\left(y^i \mid \mathbf{x}^i, \mathbf{w}^{k,t}\right)$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_{i:z^{\{i,t\}}=k}\left(y^i - \langle \mathbf{w}, \mathbf{x}^i\rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

Exercise: verify these updates

Hard Mixed Regression

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
    1. Let $z^{i,t} = \arg\min_{k \in \{0,1\}} \left|y^i - \left\langle \mathbf{w}^{t,k}, \mathbf{x}^i \right\rangle\right|$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_{i:z^{\{i,t\}}=k} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

Exercise: verify these updates

Hard Mixed Regression

Assign to the "closest" line!

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $z^{i,t} = \arg\min_{k \in \{0,1\}} \left| y^i - \langle \mathbf{w}^{t,k}, \mathbf{x}^i \rangle \right|$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_{i:z^{\{i,t\}}=k} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\}$
5. Repeat until convergence

Exercise: verify these updates

Hard Mixed Regression

ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $z^{i,t}$ using $\boldsymbol{\Theta}^t$
   1. Let $z^{i,t} = \arg\min_{k \in \{0,1\}} \left| y^i - \left\langle \mathbf{w}^{t,k}, \mathbf{x}^i \right\rangle \right|$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_{i: z^{\{i,t\}} = k} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

Assign to the “closest” line!

In k-means, we assigned to the closest mean

Exercise: verify these updates

Weighted Regression

Weighted Regression

$$\hat{\mathbf{w}}_{\mathrm{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$
$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$$

Weighted Regression

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \gamma_i \cdot \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$= (M + \lambda \cdot I)^{-1}\mathbf{b}$$

$$M = \sum_{i=1}^{n} \gamma_i \mathbf{x}^i (\mathbf{x}^i)^\top$$

$$\mathbf{b} = \sum_{i=1}^{n} \gamma_i y_i \cdot \mathbf{x}^i$$

Weighted Regression

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$= (\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$$

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} \gamma_i \cdot \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$= (M + \lambda \cdot I)^{-1}\mathbf{b}$$

$$M = \sum_{i=1}^{n} \gamma_i \mathbf{x}^i (\mathbf{x}^i)^\top$$

$$\mathbf{b} = \sum_{i=1}^{n} \gamma_i y_i \cdot \mathbf{x}^i$$

Exercise

Soft Mixed Regression

SOFT ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $\gamma^{\{i,k,t\}}$ using $\boldsymbol{\Theta}^t$
    1. Let $\mathrm{c}^{\{i,k,t\}} = \exp\left(-\frac{\left(y^i - \langle \mathbf{w}^{\{t,k\}}, \mathbf{x}^i \rangle\right)^2}{2}\right)$
    2. Let $\gamma^{\{i,k,t\}} = \frac{\mathrm{c}^{\{i,k,t\}}}{\mathrm{c}^{\{i,0,t\}} + \mathrm{c}^{\{i,1,t\}}}$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_i \gamma^{\{i,k,t\}} \cdot \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

Soft Mixed Regression

SOFT ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $\gamma^{\{i,k,t\}}$ using $\boldsymbol{\Theta}^t$
   1. Let $\mathrm{c}^{\{i,k,t\}} = \exp\left(-\frac{\left(y^i - \langle \mathbf{w}^{\{t,k\}}, \mathbf{x}^i \rangle\right)^2}{2}\right)$
   2. Let $\gamma^{\{i,k,t\}} = \frac{\mathrm{c}^{\{i,k,t\}}}{\mathrm{c}^{\{i,0,t\}} + \mathrm{c}^{\{i,1,t\}}}$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_i \gamma^{\{i,k,t\}} \cdot \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

Exercise: derive these updates

Soft Mixed Regression

SOFT ALTERNATING OPTIMIZATION

1. Initialize $\boldsymbol{\Theta}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\right\}$
2. For $i \in [n]$, update $\gamma^{\{i,k,t\}}$ using $\boldsymbol{\Theta}^t$
    1. Let $c^{\{i,k,t\}} = \exp\left(-\frac{\left(y^i - \left\langle \mathbf{w}^{\{t,k\}}, \mathbf{x}^i \right\rangle\right)^2}{2}\right)$
    2. Let $\gamma^{\{i,k,t\}} = \frac{c^{\{i,k,t\}}}{c^{\{i,0,t\}} + c^{\{i,1,t\}}}$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_i \gamma^{\{i,k,t\}} \cdot \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\right\}$
5. Repeat until convergence

$\propto \mathbb{P}\left[k \mid y^i, \mathbf{x}^i, \boldsymbol{\Theta}^t\right]$

Exercise: derive these updates

Soft Mixed Regression

SOFT ALTERNATING OPTIMIZATION

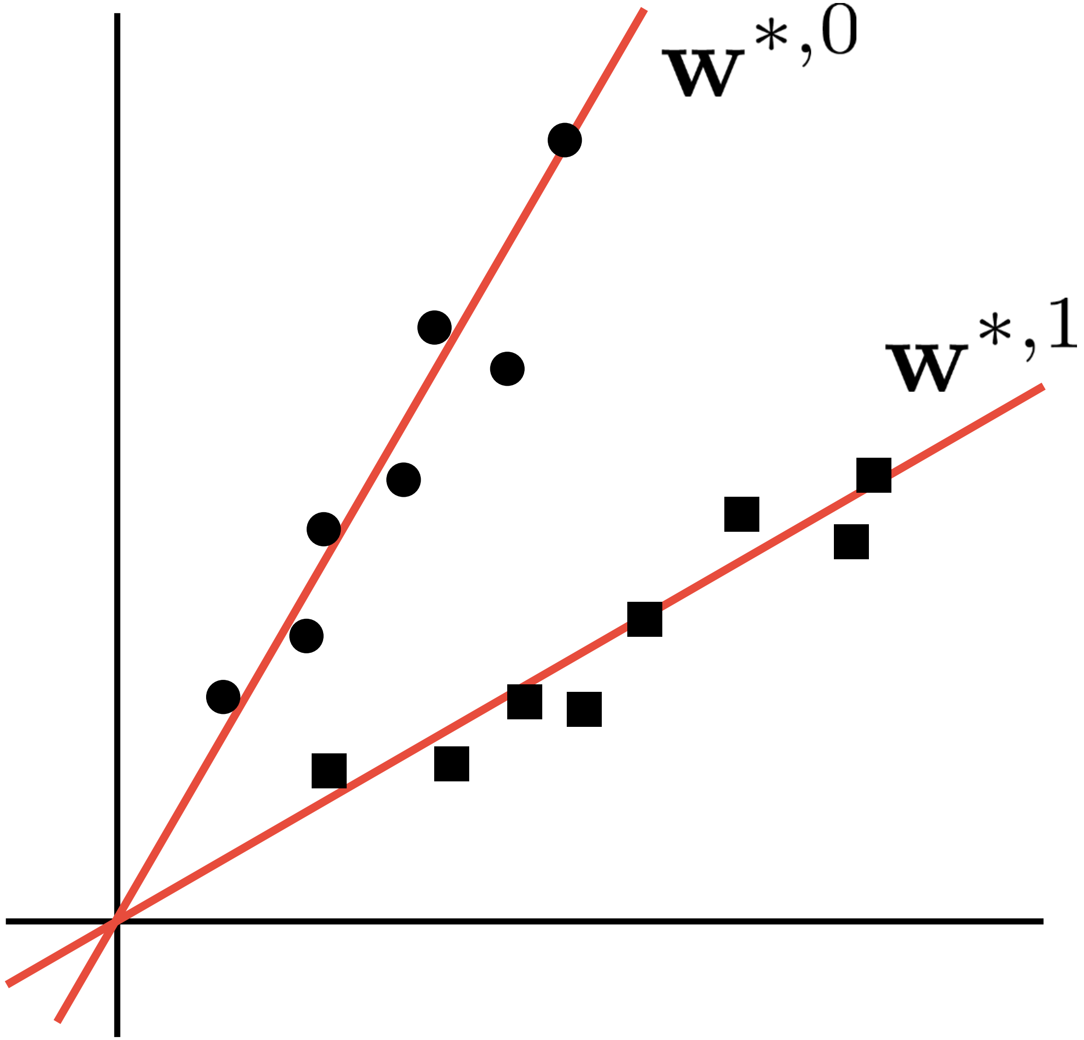
1. Initialize $\boldsymbol{\Theta}^0 = \{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}\}$
2. For $i \in [n]$, update $\gamma^{\{i,k,t\}}$ using $\boldsymbol{\Theta}^t$
    1. Let $c^{\{i,k,t\}} = \exp\left(-\frac{\left(y^i - \langle \mathbf{w}^{\{t,k\}}, \mathbf{x}^i \rangle\right)^2}{2}\right)$
    2. Let $\gamma^{\{i,k,t\}} = \frac{c^{\{i,k,t\}}}{c^{\{i,0,t\}} + c^{\{i,1,t\}}}$
3. Update $\mathbf{w}^{\{t+1,k\}} = \arg\min_{\mathbf{w}} \sum_i \gamma^{\{i,k,t\}} \cdot \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \frac{1}{2}\|\mathbf{w}\|_2^2$
4. Set $\boldsymbol{\Theta}^{t+1} = \{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}\}$
5. Repeat until convergence

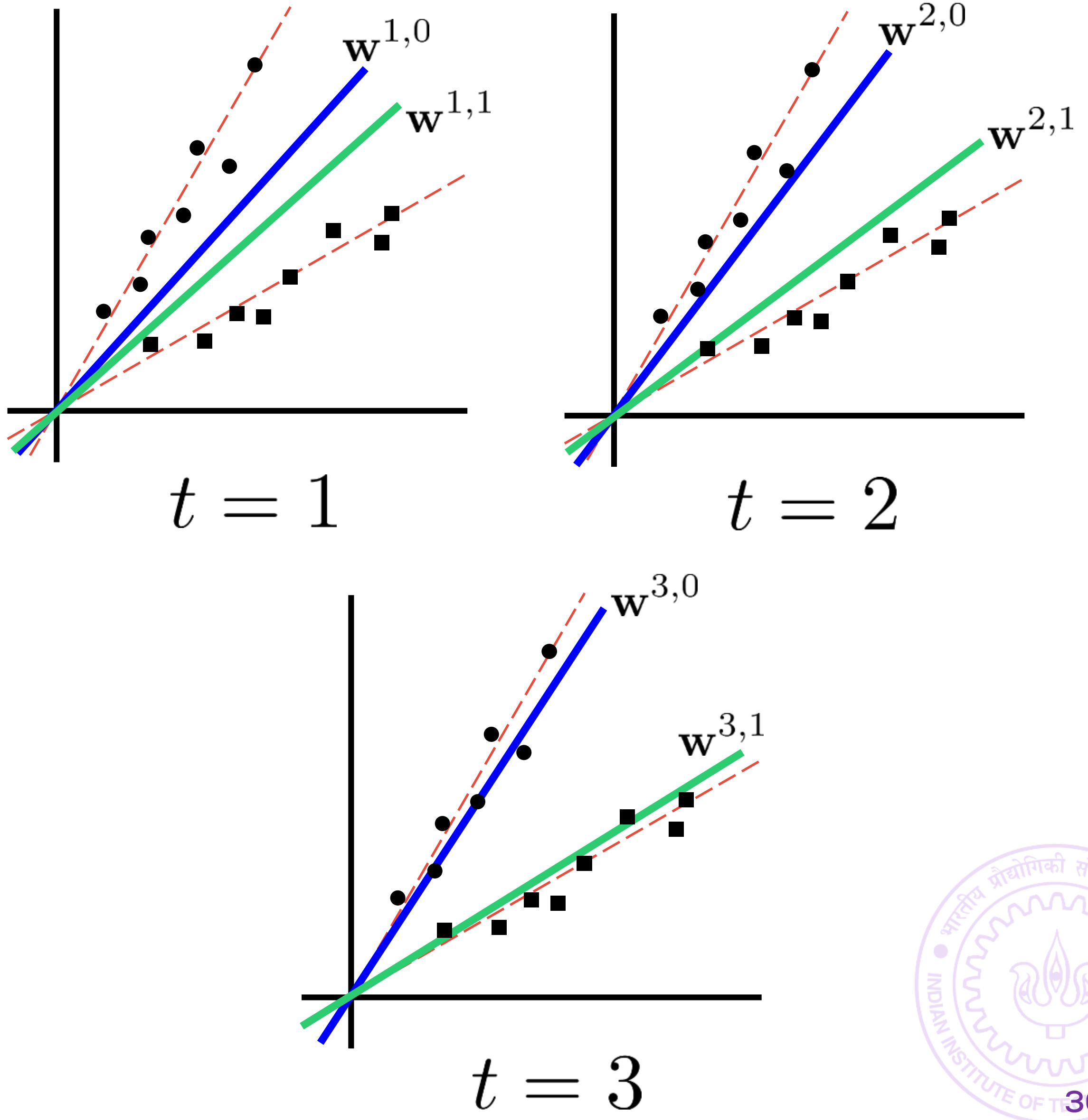
$\propto \mathbb{P}[k \mid y^i, \mathbf{x}^i, \boldsymbol{\Theta}^t]$

$= \mathbb{P}[k \mid y^i, \mathbf{x}^i, \boldsymbol{\Theta}^t]$

Exercise: derive these updates

Soft MR in action





The EM Algorithm

- Generalizes the notion of "soft" updates
- Very powerful algorithm
- Related to alternating minimization
- Will study this next time!

Please give your Feedback

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