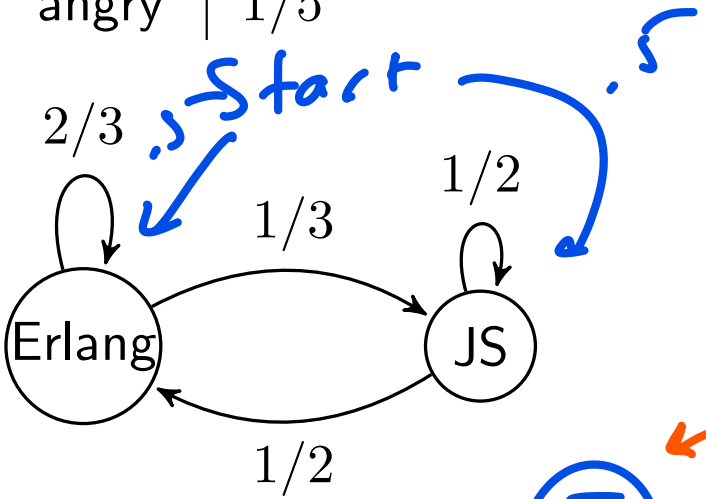
Erlang/JavaScript example

$\delta_2(J)$ (most likely prob that uses JS day 2)

E	$p(E\|X = \text{Erlang})$
happy	$4/5$
angry	$1/5$

E	$p(E\|X = \text{JavaScript})$
happy	$1/4$
angry	$3/4$



prev = JS : $\frac{1}{8} \cdot \frac{1}{2} \cdot \frac{1}{4}$

prev = E : $\frac{2}{5} \cdot \frac{1}{3} \cdot \frac{1}{4}$

day 2

used to comp. $\delta_2(E)$

$\delta_1(E) = \frac{4}{5} \cdot \frac{1}{2}$

$\delta_2(E)$:

$\frac{2}{5} \cdot \frac{2}{3} \cdot \frac{4}{5}$

or $\frac{1}{8} \cdot \frac{1}{2} \cdot \frac{4}{5}$

Start

E

E

J

J

$\delta_1(J) = \frac{1}{4} \cdot \frac{1}{2}$

ICS Summer Academy Session II
Topic 8: Clustering

Michael Shindler

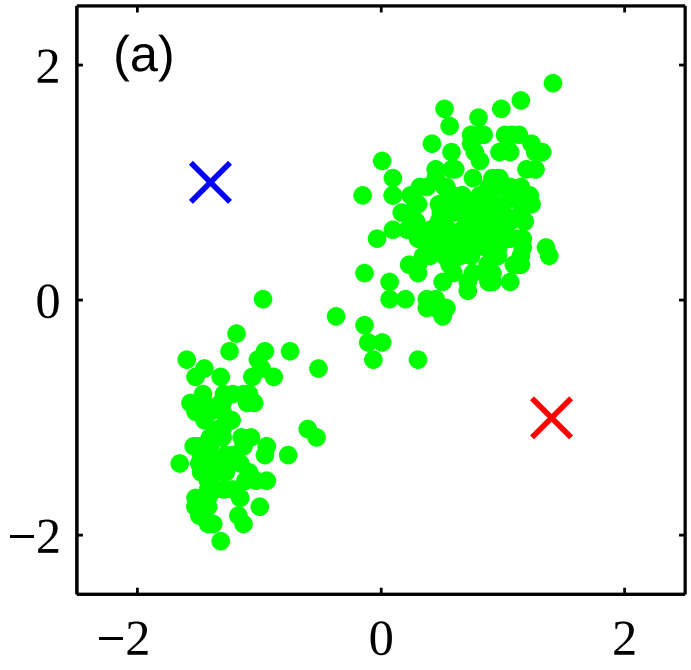
Clustering

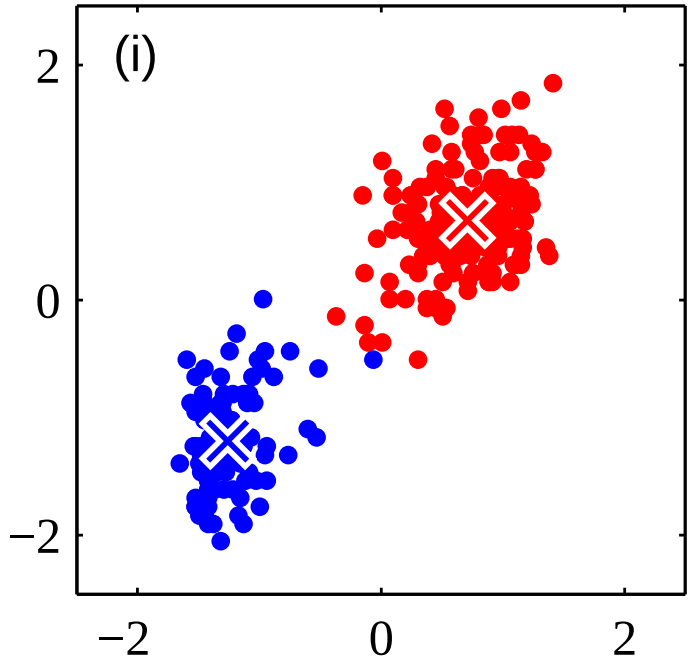
unsupervised

Setup Given $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$ and K, we want to output

- $\{\boldsymbol{\mu}_k\}_{k=1}^K$: prototypes of clusters
- $A(\boldsymbol{x}_n) \in \{1, 2, \ldots, K\}$: the cluster membership

Example Cluster data into two clusters.





Why Clustering?

- Biology
- Information Retrieval
- Psychology and Medicine
- Business
- Compression
- Nearest Neighbors

Application: who wrote which Federalist papers?

- Federalist Papers: essays written anonymously
- True authors a subject of much speculation
- 1963: Frederick Mosteller and David L. Wallace: *Inference in an Authorship Problem*
- 2018: a student applies modern statistical methods (k-means and TFIDF)
- Used three features:
  - lexical similarity in sentence structure
  - lexical similarity in punctuation
  - syntactic similarity.
- Two clusters, removing the papers of John Jay.
- Prediction similar to Mosteller and Wallace

Problem: k-means clustering

Select k vectors μ

Intuition Data points assigned to cluster k should be close to $\boldsymbol{\mu}_k$

Distortion measure (clustering objective function, cost function)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|_2^2$$

where $r_{nk} \in \{0, 1\}$ is an indicator variable

Alternate Distortions : Total Cohesion

$$\sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} \text{cosine}(\boldsymbol{x}, \boldsymbol{c}_k)$$

Basic k-means Algorithm (Lloyd's)

$r_{ij} = 1$ iff point i assigned to $\vec{\mu}_j$ (j^{th} center)

Minimize distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|_2^2$$

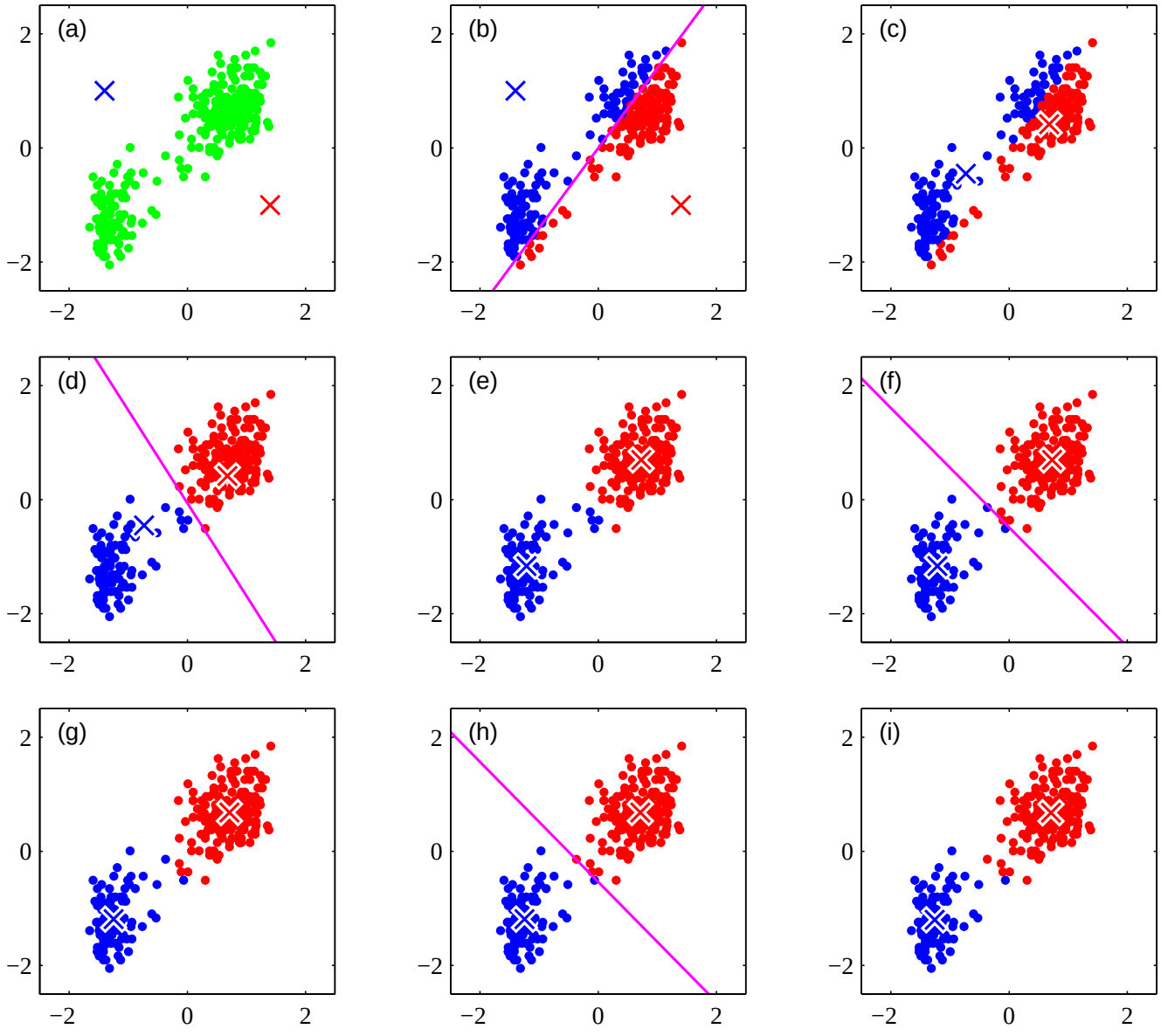
- **Step 0** Initialize $\{\boldsymbol{\mu}_k\}$ to some values
- **Step 1** Optimize $\{r_{nk}\}$ values, keeping $\{\boldsymbol{\mu}_k\}$ fixed.

i.e., assign all points to nearest center.

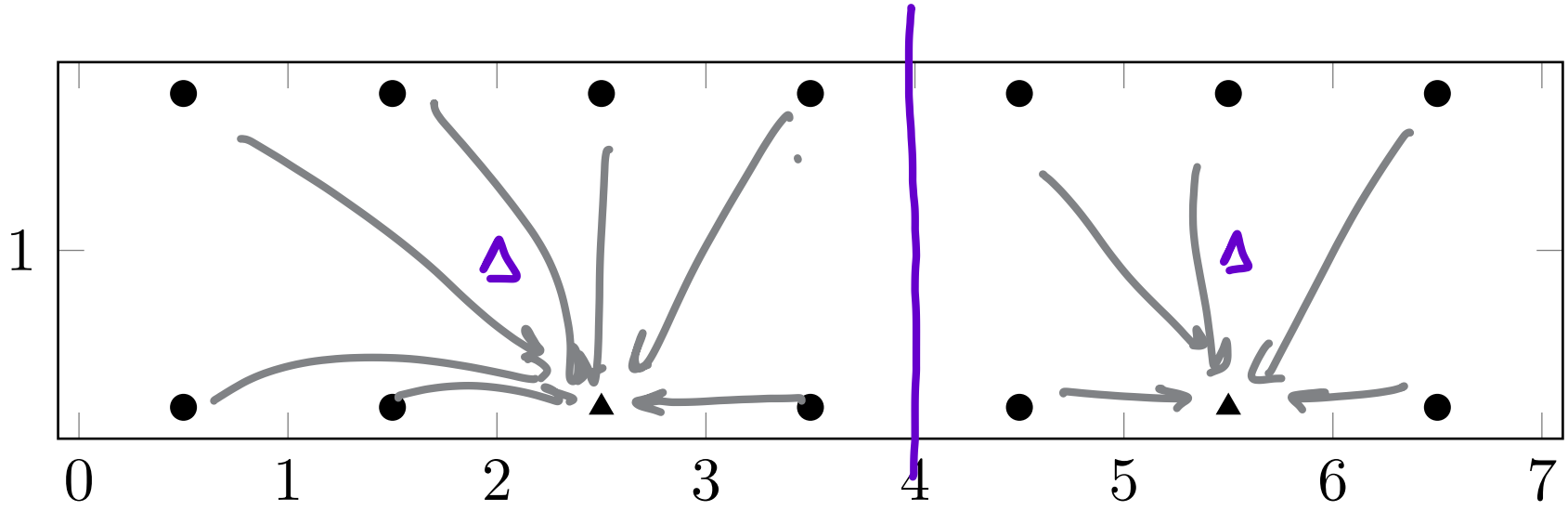
- **Step 2** Optimize $\{\boldsymbol{\mu}_k\}$ values, keeping $\{r_{nk}\}$ fixed.

move each $\vec{\mu}_k$ to center of mass of assigned points: $\vec{\mu}_k \leftarrow \frac{\text{sum vecs w/ } r_{nk}=1}{\sum r_{nn}}$

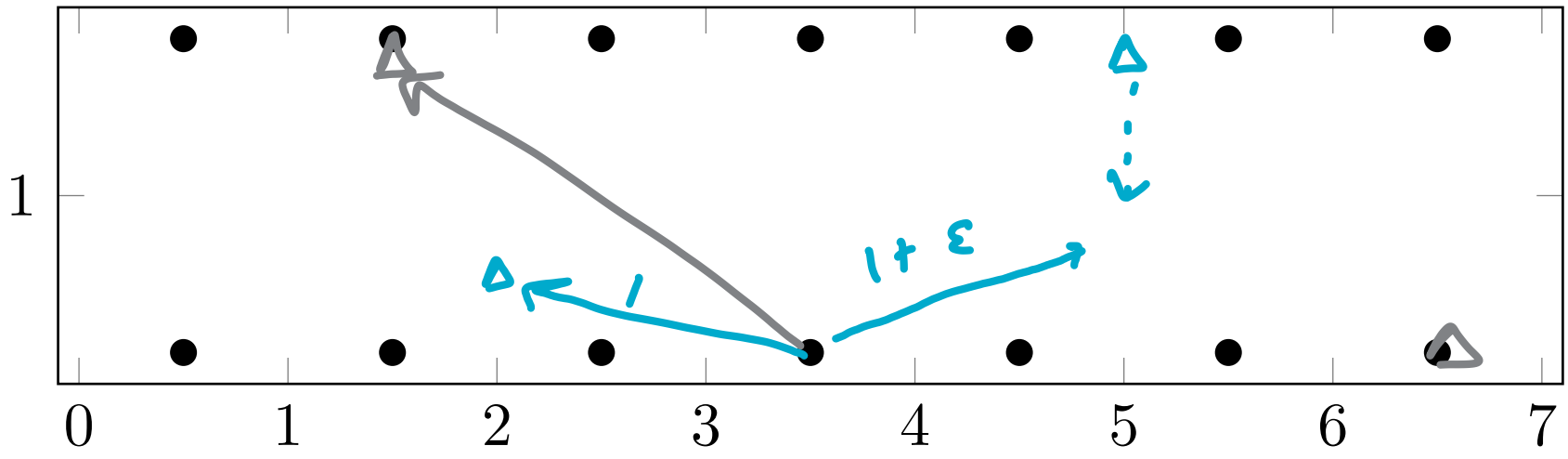
Example of running K-means algorithm



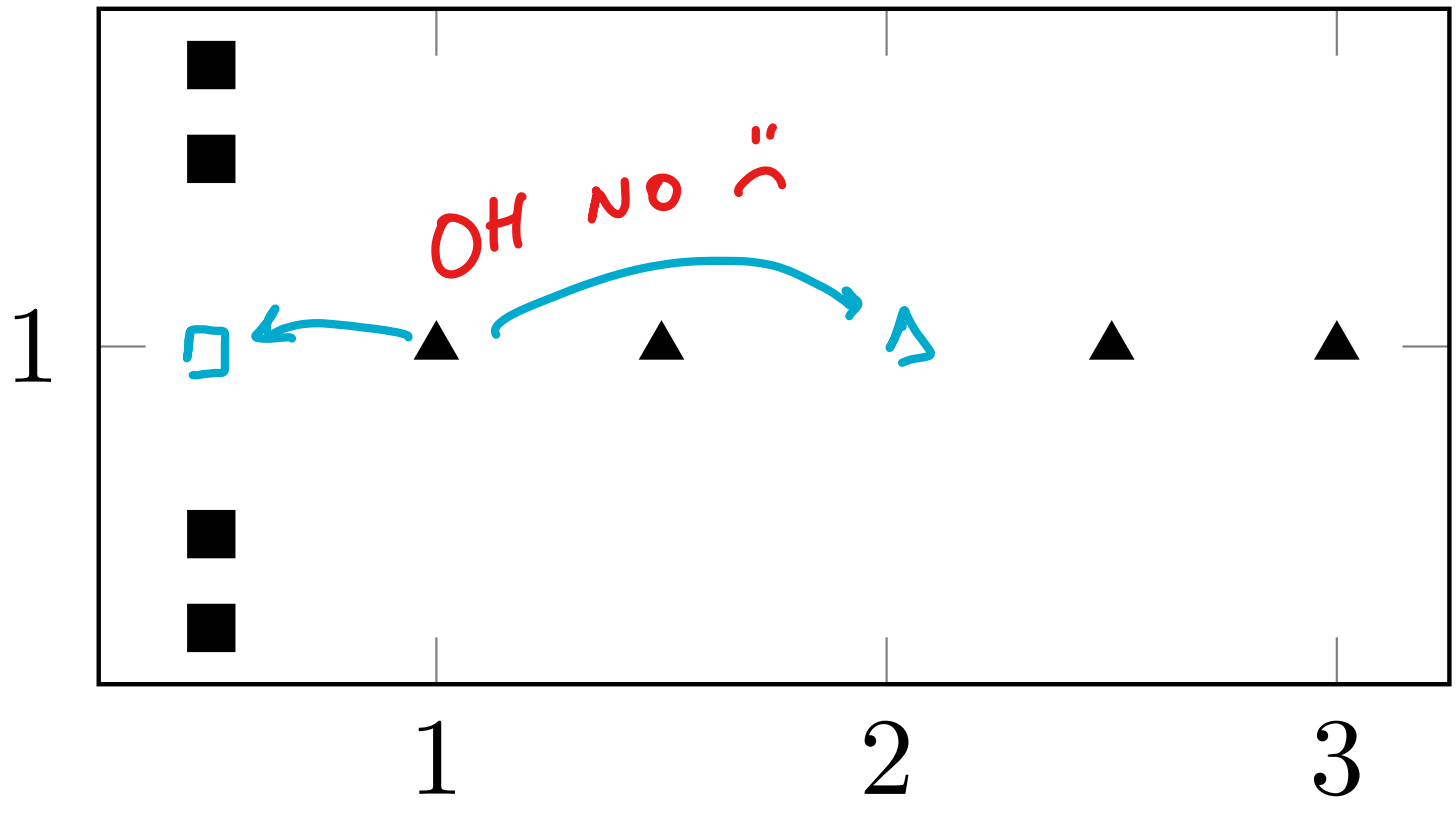
Question 3



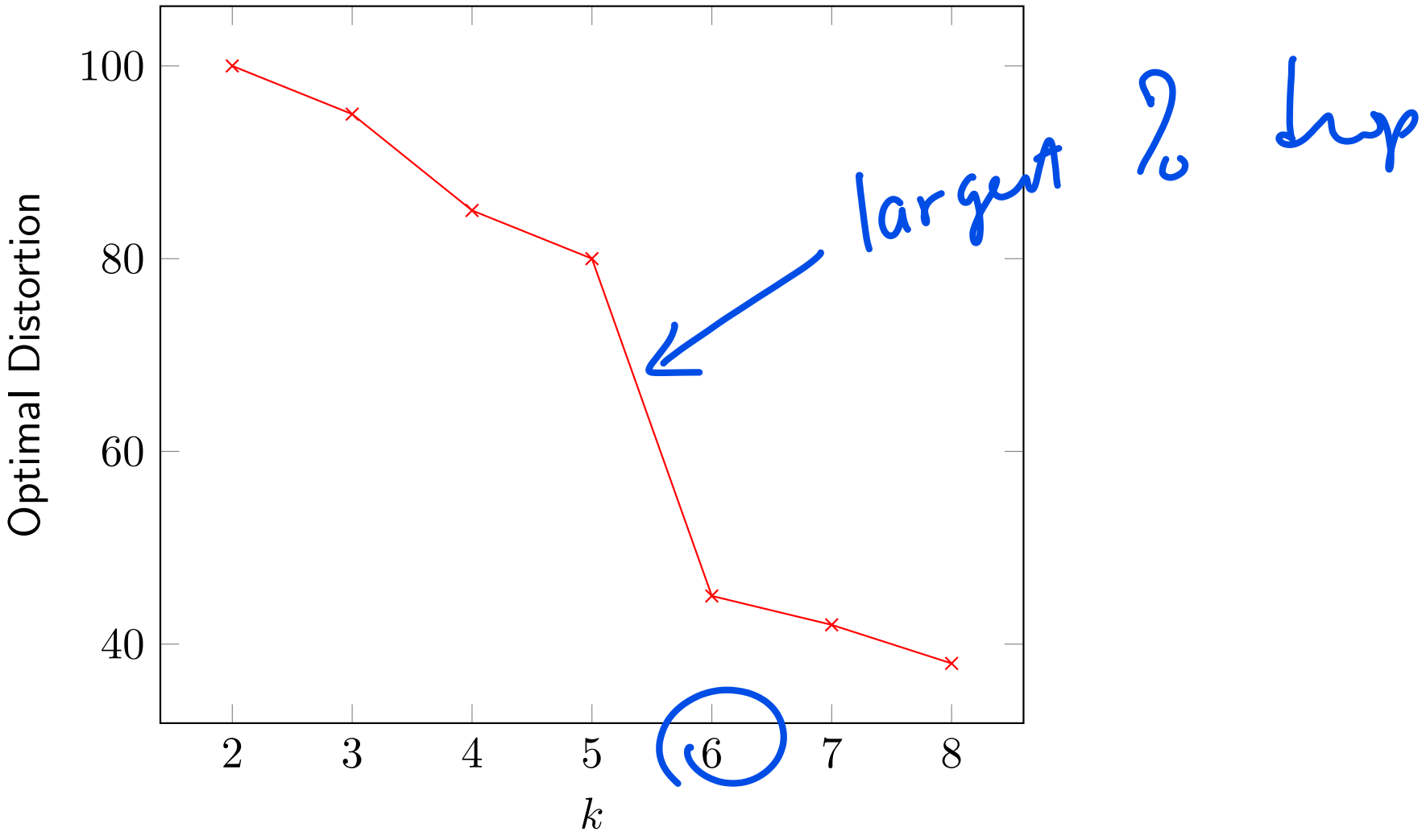
Question 4: Create the same clustering



Question 5: Did I run Lloyd's?



How should we choose k?



What do you suppose the "natural" cluster count for this data is?