Unsupervised Learning

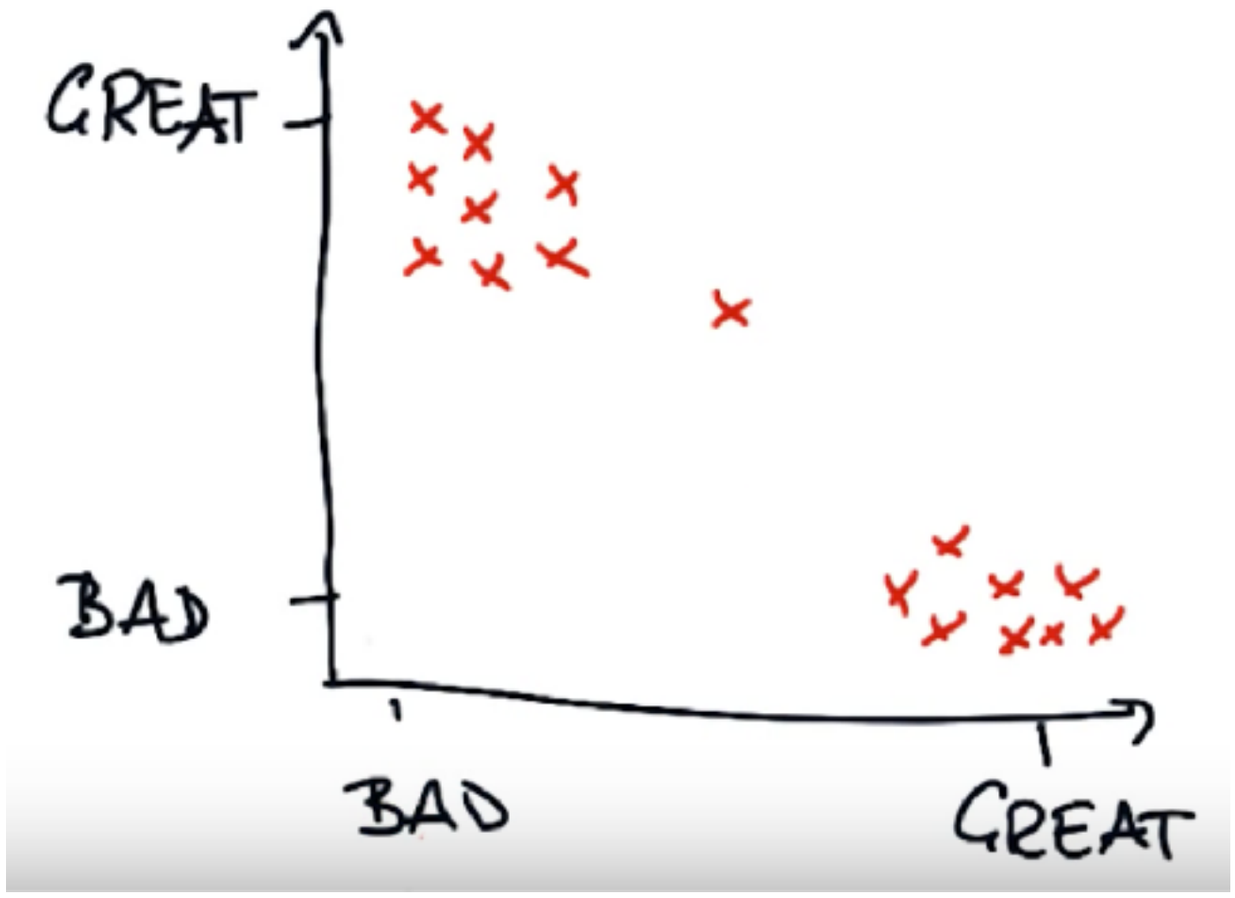
How to deal with unlabelled data?

Outline

- Clustering
- Dimensionality reduction

Example

Student A



Student B

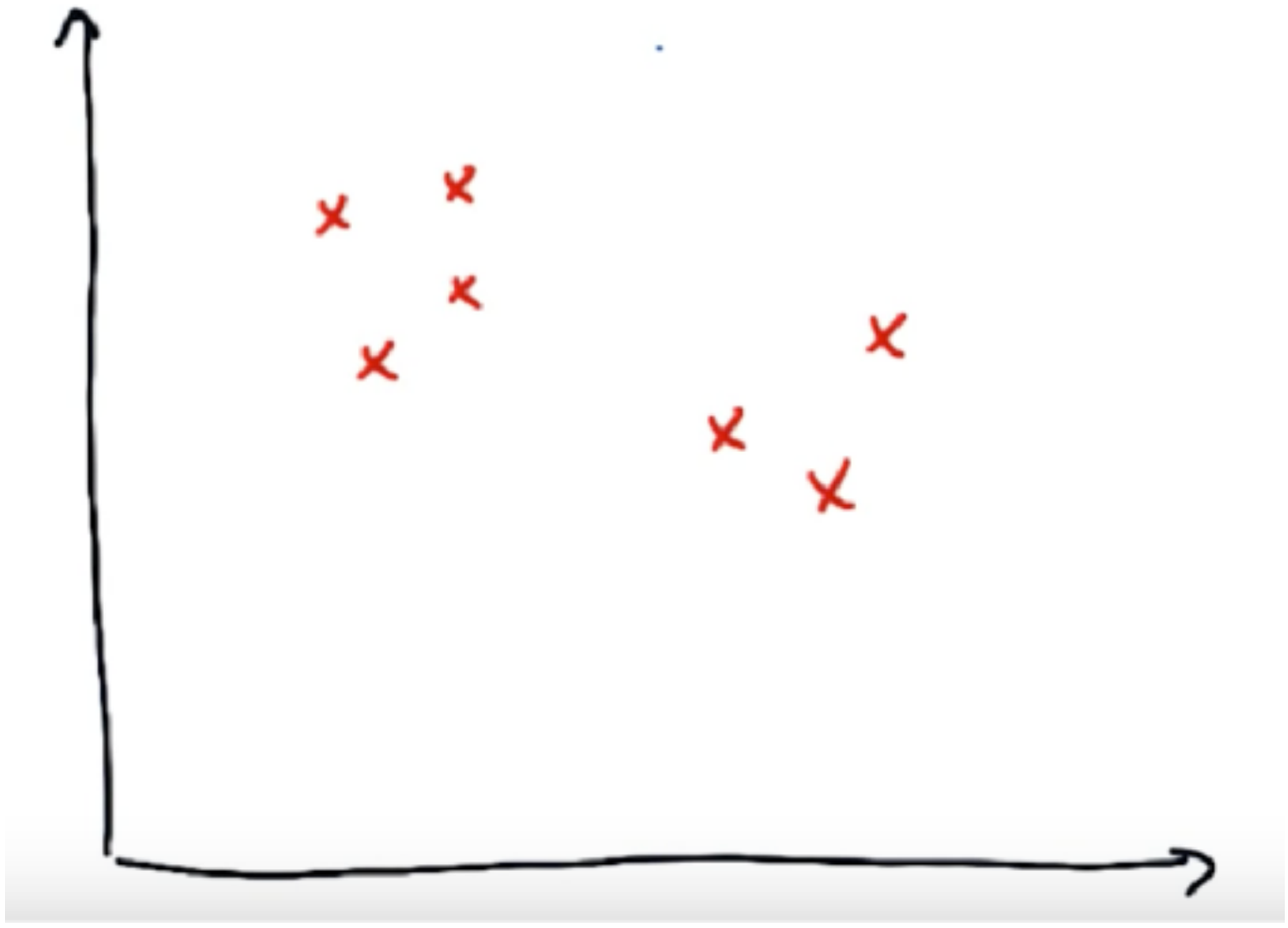
Example

Student A

Student B

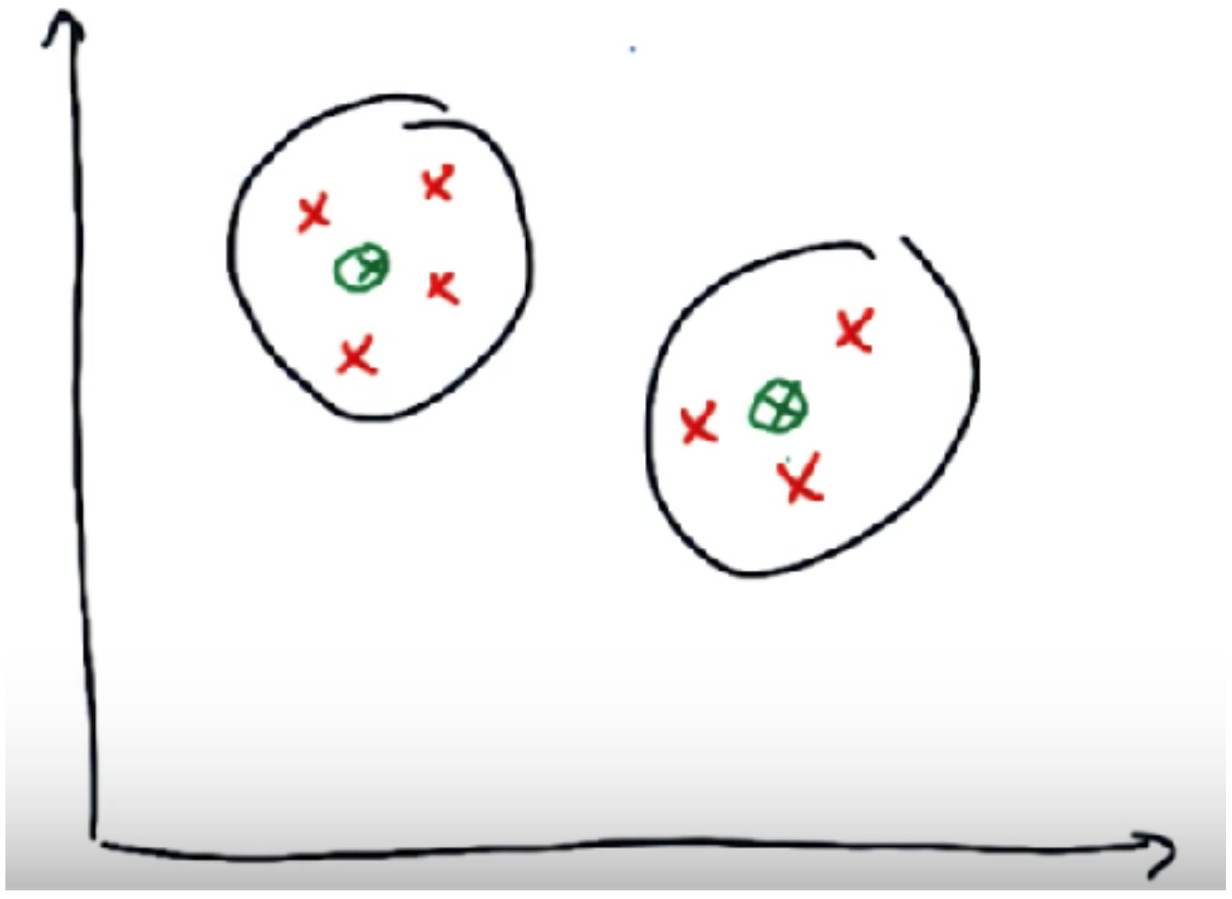
K-Means Algorithm

- How many clusters do you see?



K-Means Algorithm

- How many clusters do you see?



K-Means Algorithm

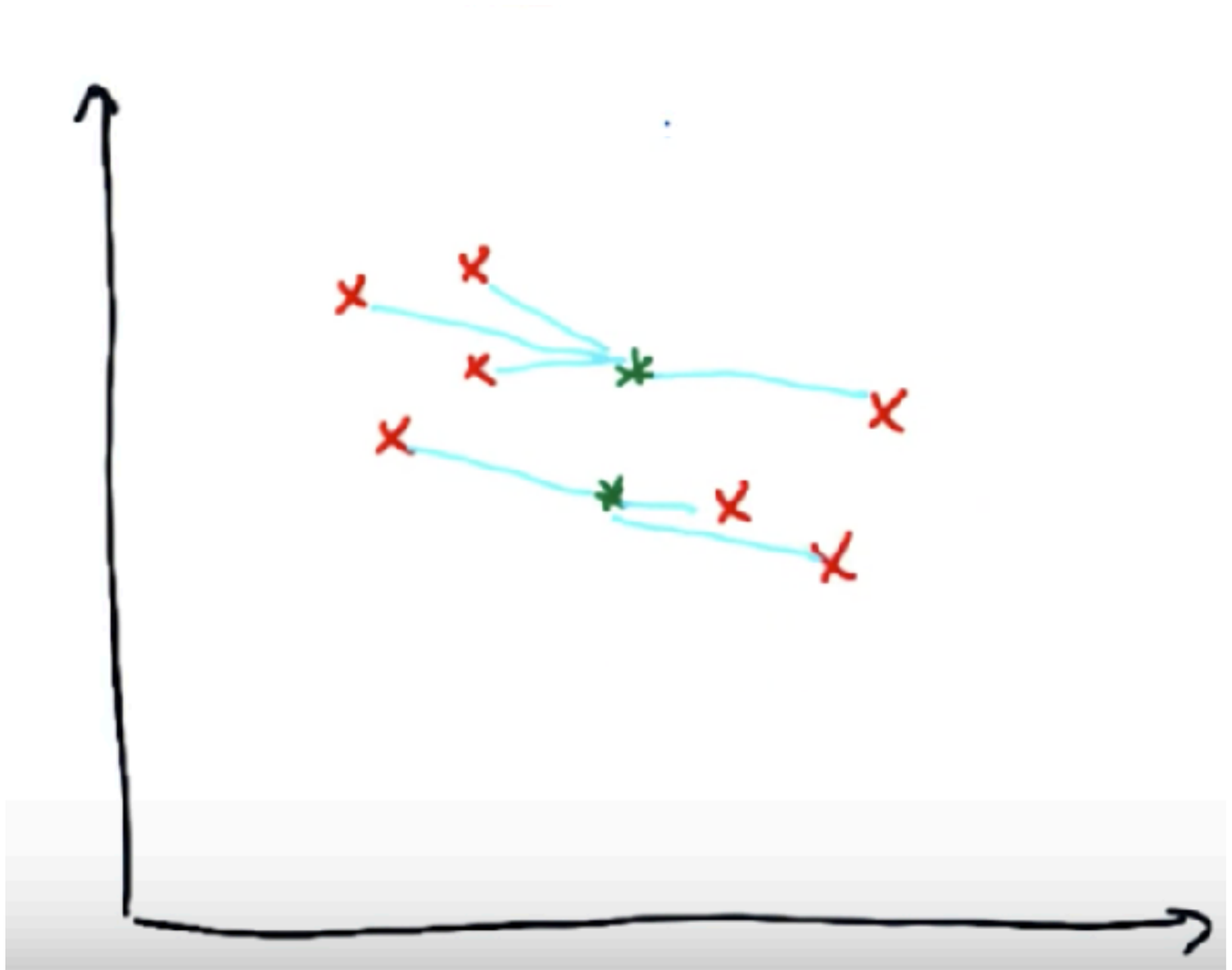
- Back and forth between
  - Assign
  - Move

K-Means Algorithm

- Back and forth between
  - Assign
  - Move

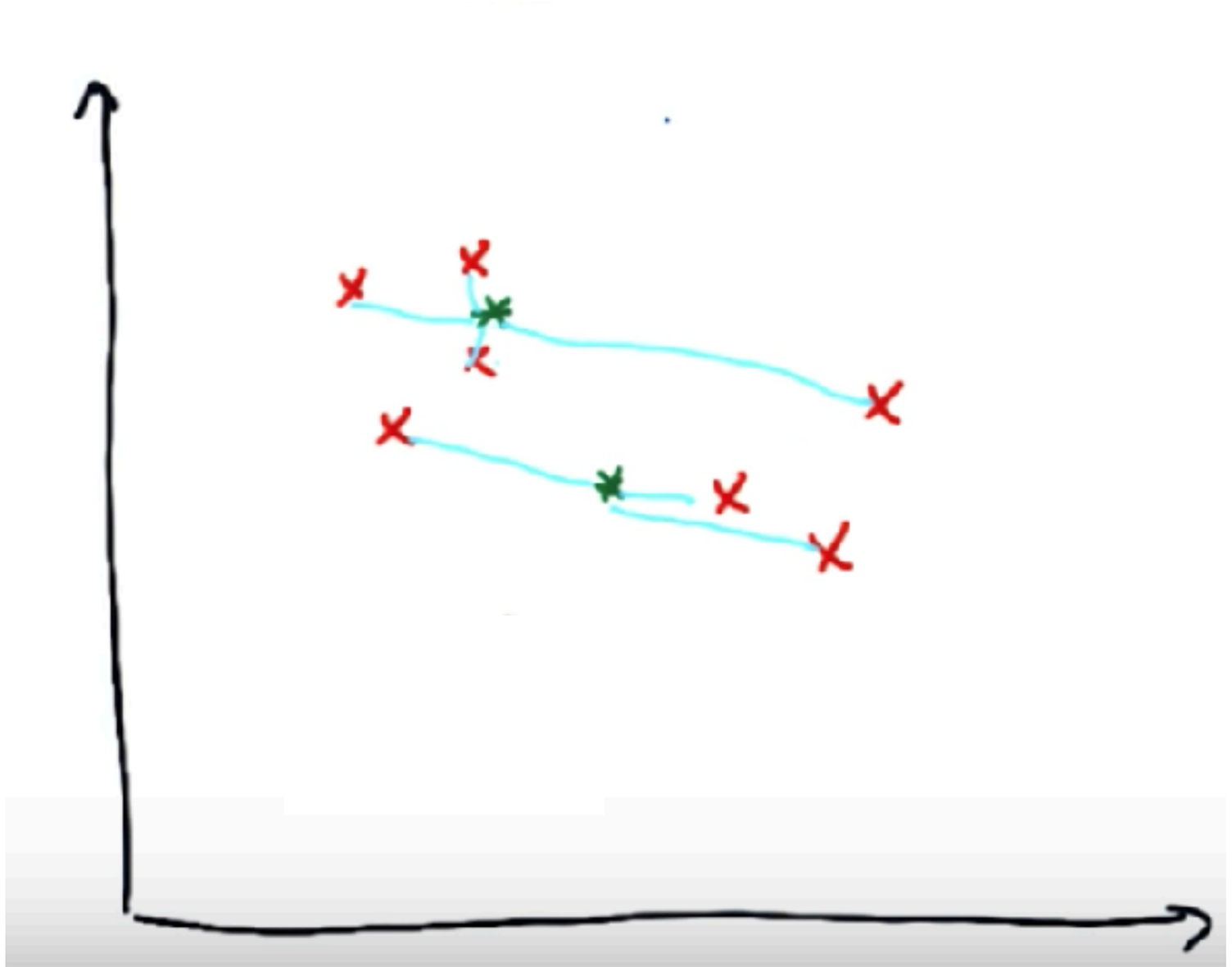
K-Means Algorithm

- Back and forth between
  - Assign
  - Move



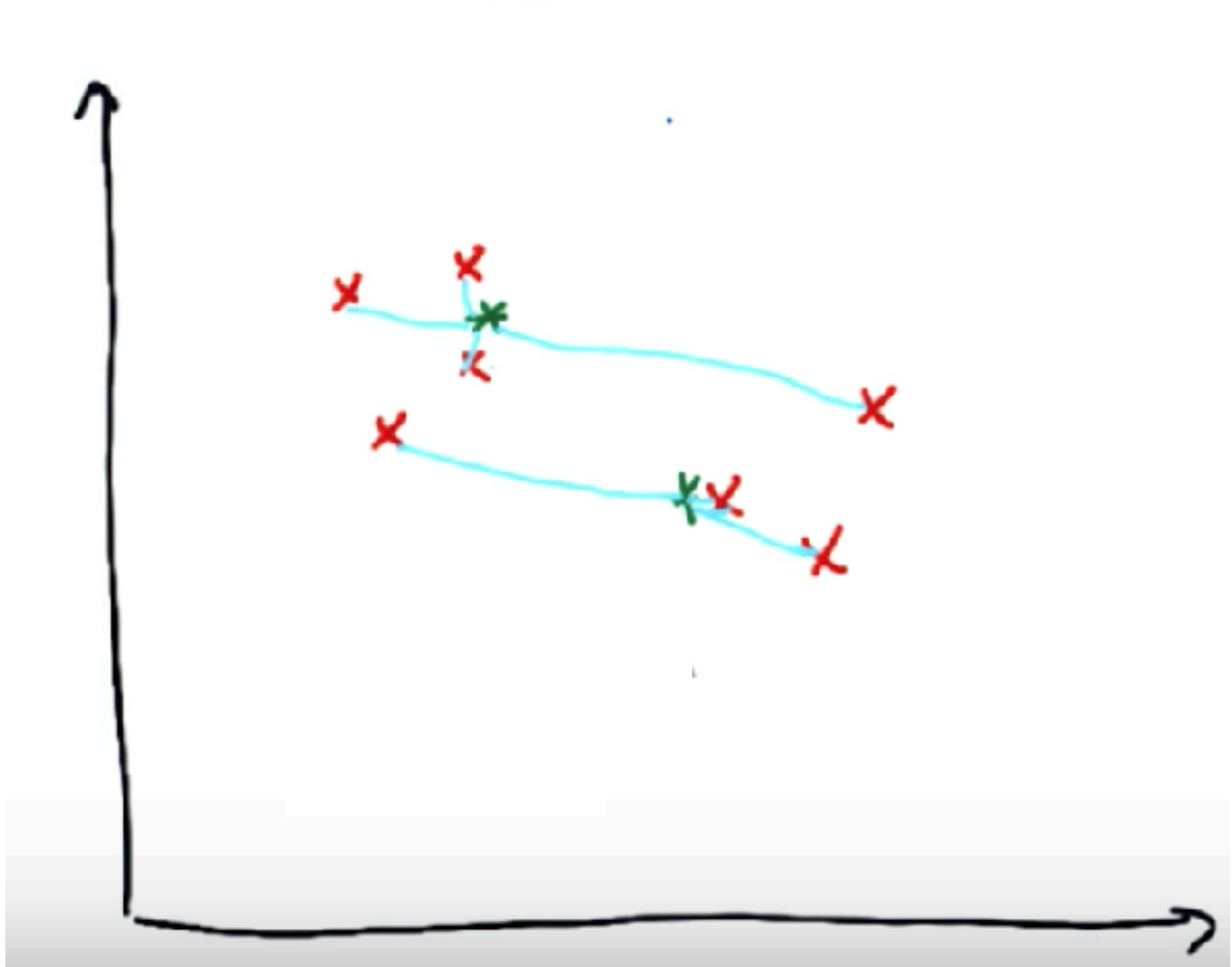
K-Means Algorithm

- Back and forth between
  - Assign
  - Move



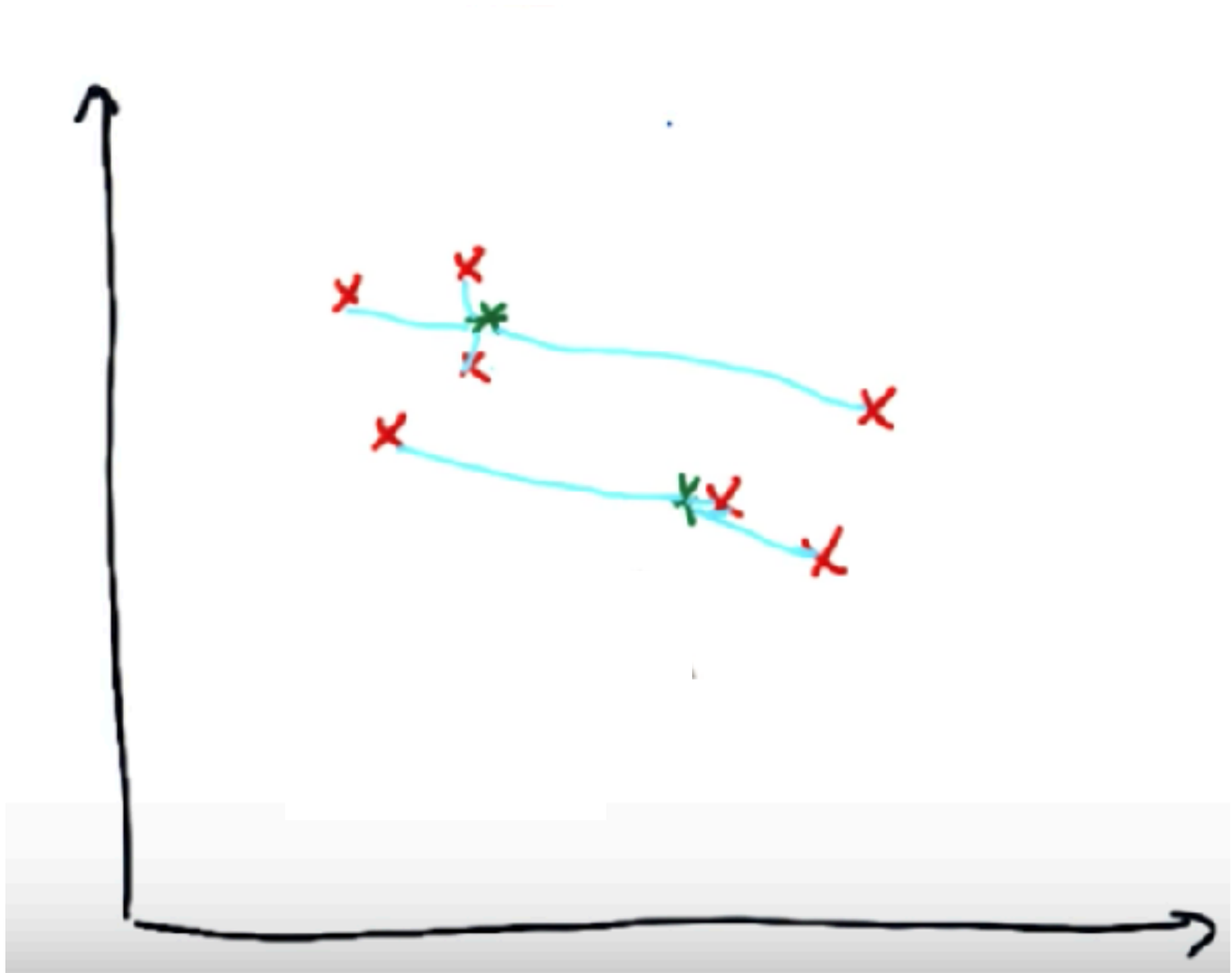
K-Means Algorithm

- Back and forth between
  - Assign
  - Move



K-Means Clustering

- Back and forth between
  - Assign
  - Move



K-Means Clustering Visualization

- Naftali Harris's nice visualisation tool

Scikit-learnk Library

- Useful Python library found here

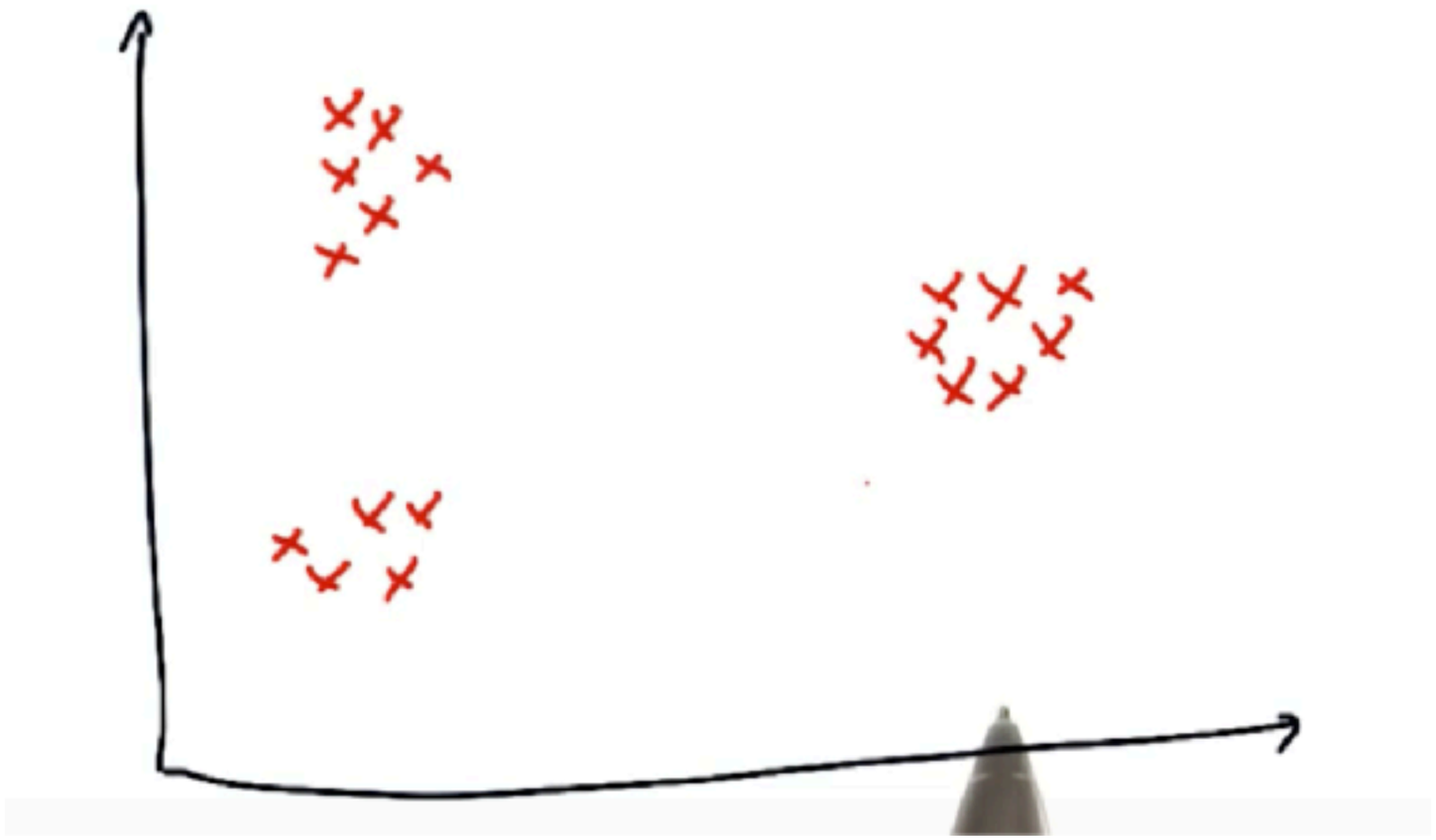
Limitation of K-Means Clustering

- Depending on initialization
- Sometimes stuck in local minima.

Limitation of K-Means Clustering

- Will output of k-means clustering for any fixed training set always be same?

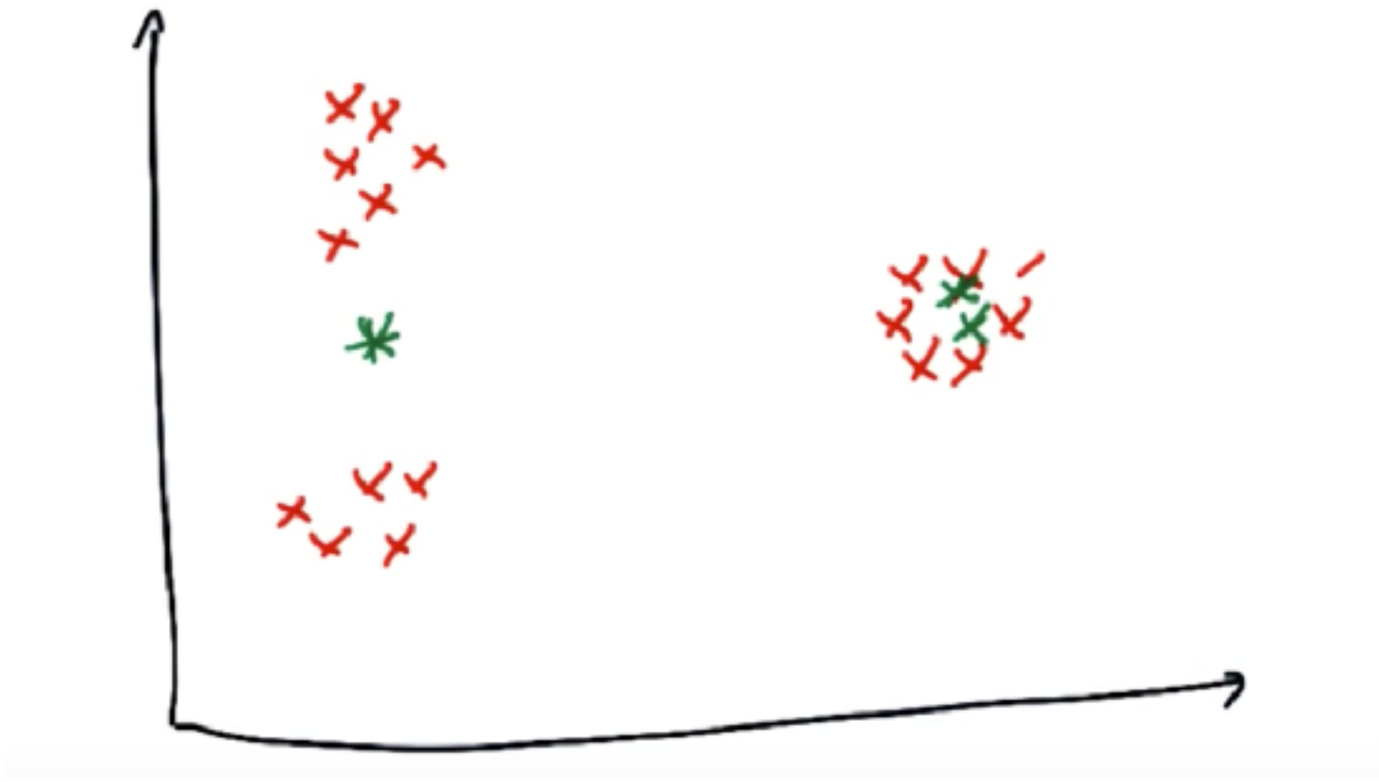
Limitation of K-Means Clustering



Limitation of K-Means Clustering

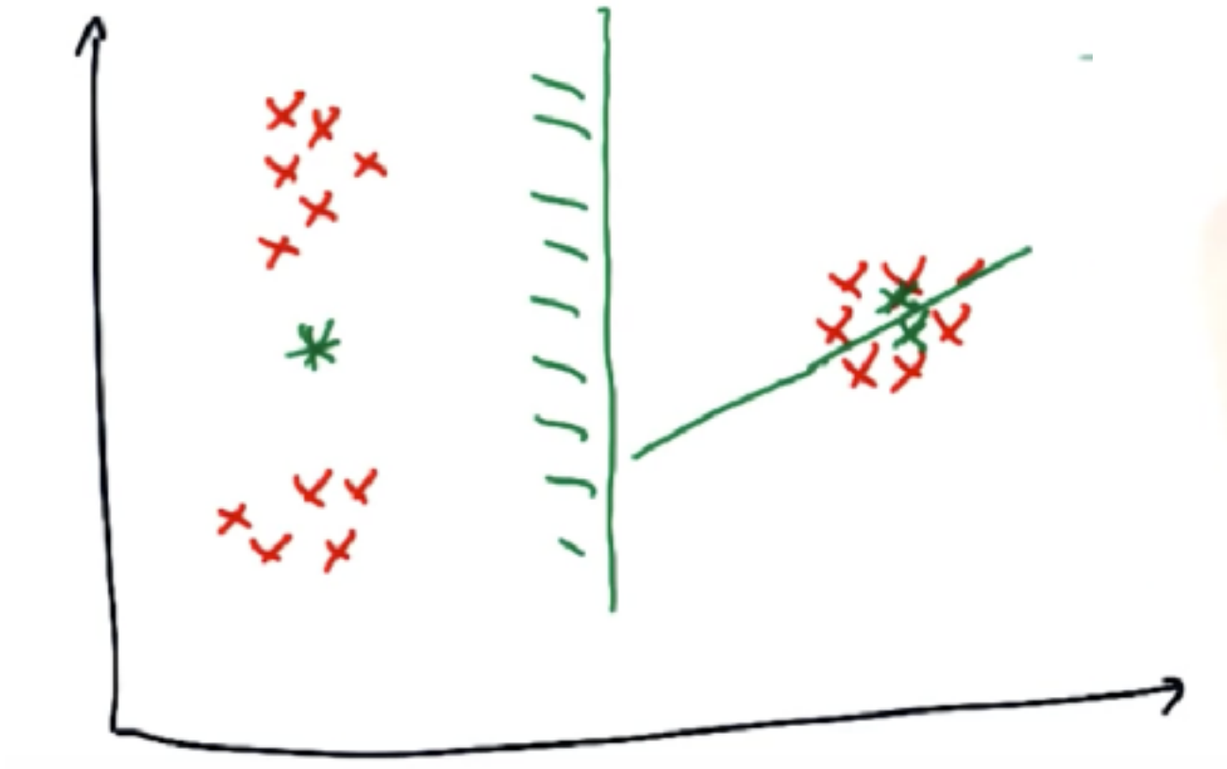
- Do you think k-means clustering could produce following result?

Limitation of K-Means Clustering



- Do you think k-means clustering could produce following result?

Limitation of K-Means Clustering



- Do you think k-means clustering could produce following result?

K-Means Clustering Demo

- Apply k-means clustering to Enron financial data

K-Means Clustering Demo

1. Run starter code k_means_cluster.py that reads email and financial Enron dataset and gets us ready for clustering

2. Start performing k-means based on just two features

3. Take look at code, determine which features code uses for clustering

4. Perform k-means clustering on data with 2 clusters as parameter. Store your cluster predictions to list called red, so that call to Draw() function at bottom works properly. Are the clusters what you expected?

K-Means Clustering Demo

5. Add third feature_list “total_payments” Rerun clustering using 3 input features instead of 2. Compare plot with clusterings to earlier one obtained with 2 input features. Do any points switch clusters? How many?