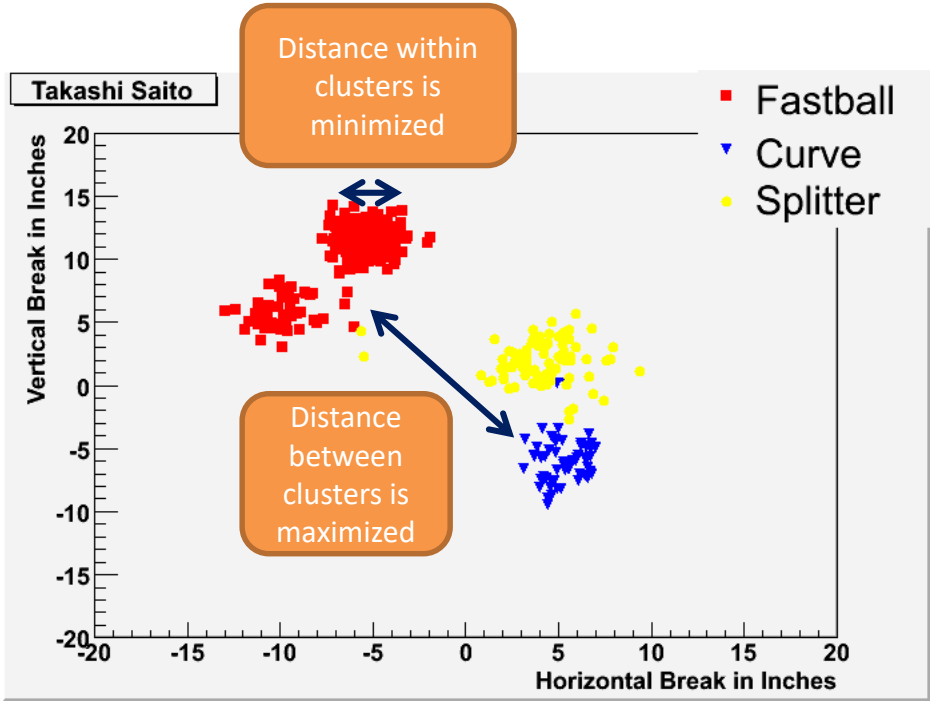
# MIS2502:
# Data Analytics
# *Clustering and Segmentation*

**Alvin Zuyin Zheng**
**zheng**@temple.edu
http://community.mis.temple.edu/zuyinzheng/

# What is Cluster Analysis?

Grouping data so that elements in a group will be

- Similar (or related) to one another
- Different (or unrelated) from elements in other groups




http://www.baseball.bornbybits.com/blog/uploaded_images/Takashi_Saito-703616.gif

# Applications

## Understanding data

- Group related documents for browsing
- Create groups of similar customers
- Discover which stocks have similar price fluctuations
- Group similar plants into species

## Summarizing data

- Reduce the size of large data sets
- Data in similar groups can be combined into a single data point

# Applications

## Marketing (Market Segmentation)

- Discover distinct customer groups for targeted promotions

## Industry analysis

- Finding groups of similar firms based on profitability, growth rate, market size, products, etc.

## Political forecasting

- Group neighborhoods by demographics, lifestyles and political view

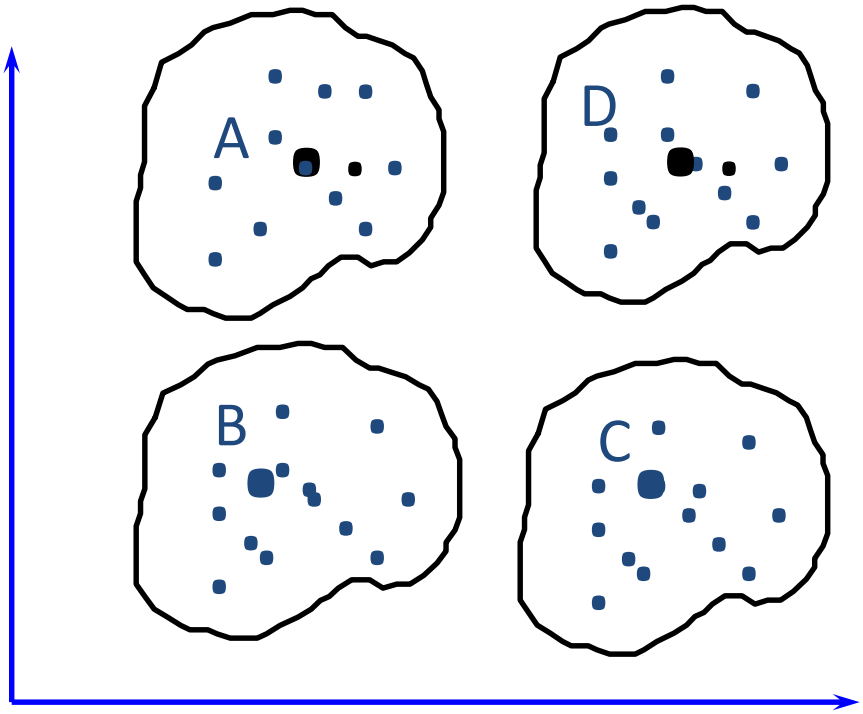
## Biology

- Group similar plants into species

## Finance

- Define groups of similar stocks based on financial characteristics, and select stocks from different groups to create balanced portfolios

# Market Segmentation (for Carpet Fibers)




**Strength**
(Importance)

**Water Resistance**
(Importance)

Typical clients:

- A: schools
- B: light commercial
- C: indoor/outdoor carpeting
- D: health clubs

https://www.stat.auckland.ac.nz/~balemi/Segmentation%20&%20Targeting.ppt

# What cluster analysis is NOT

Classification (like Decision Trees)

People simply place items into categories

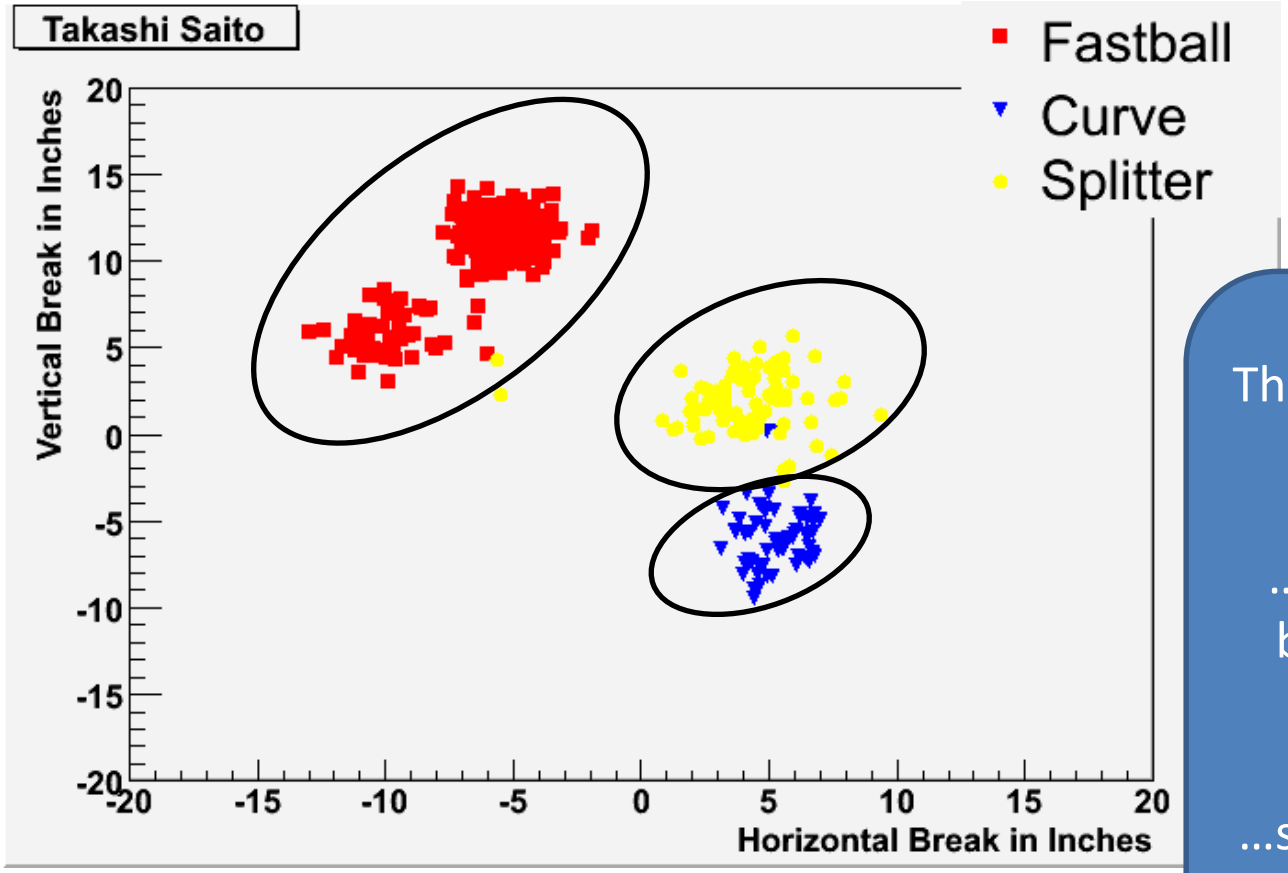
Simple categorization by attributes

Dividing students into groups by last name

The clusters must be learned from the data, not from external specifications.

Creating the "buckets" beforehand is categorization, but not clustering.

# (Partitional) Clustering




Three distinct groups emerge, but…

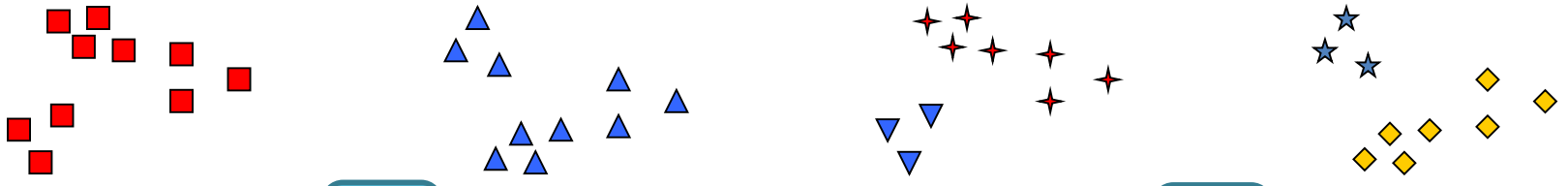
…some curveballs behave more like splitters.

…some splitters look more like fastballs.

# Clusters can be ambiguous

**How many clusters?**

6



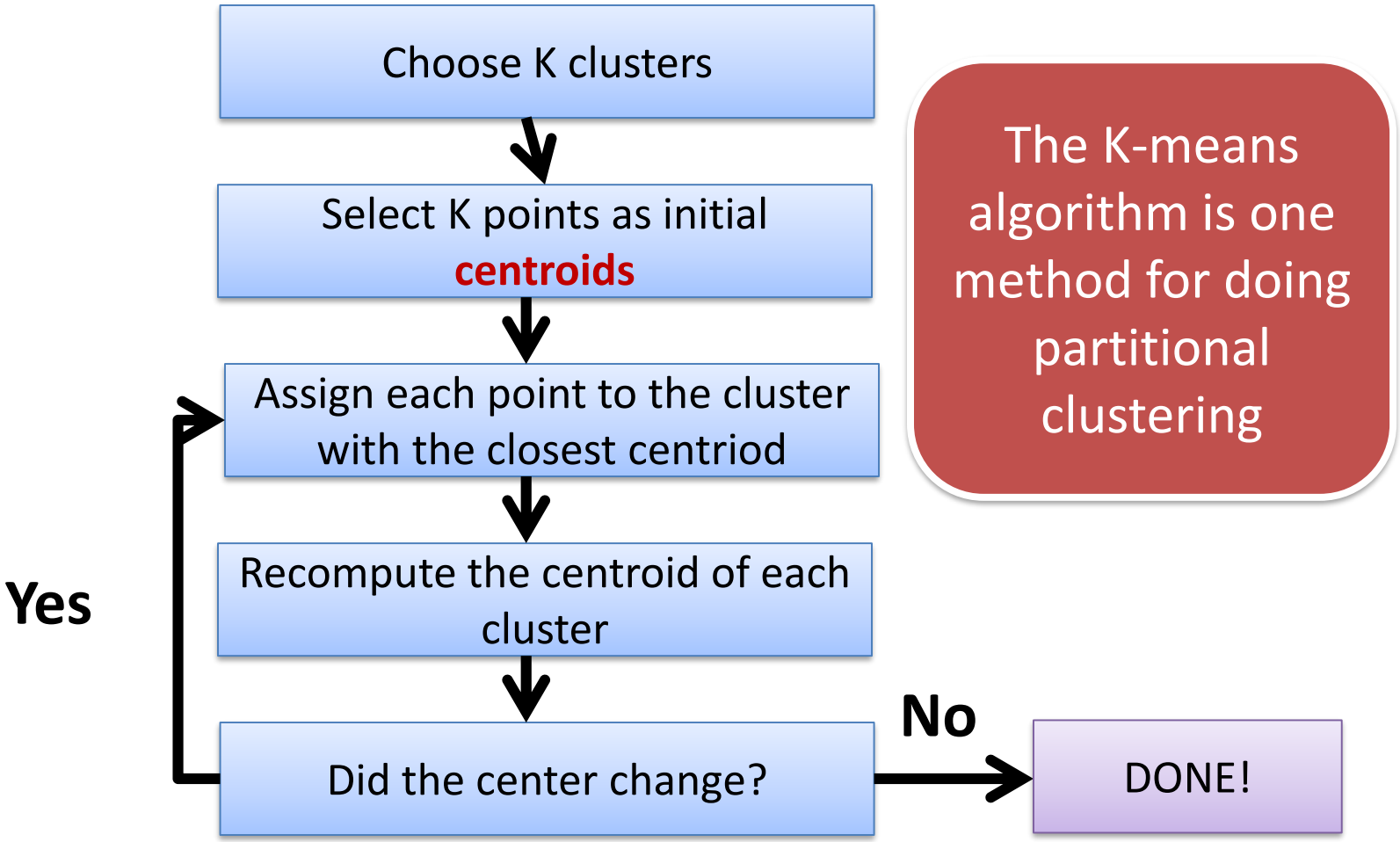
2

4

***The difference is the threshold you set.***
***How distinct must a cluster be to be it's own cluster?***

adapted from Tan, Steinbach, and Kumar. Introduction to Data Mining (2004)

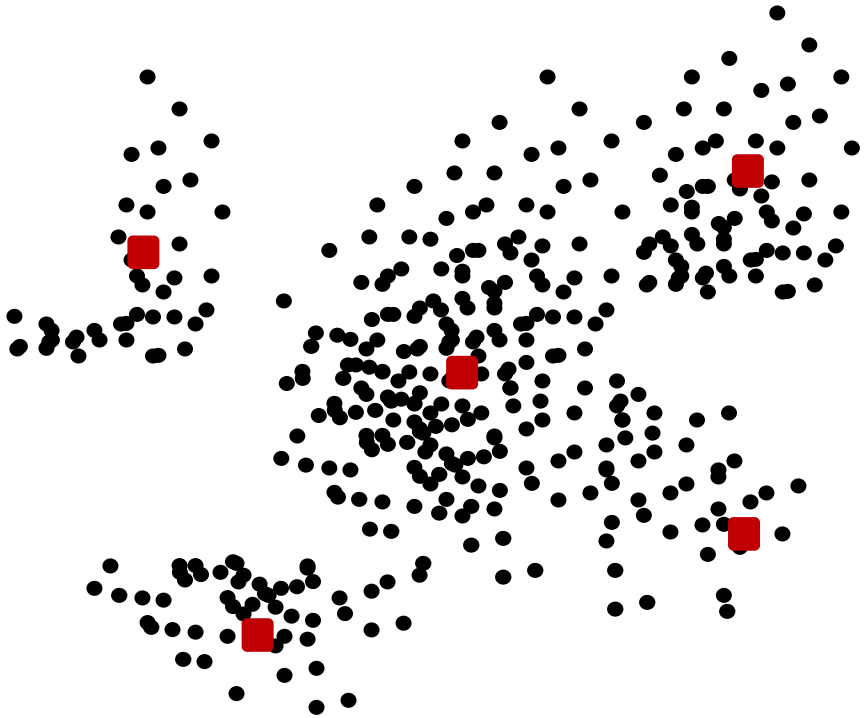
# K-means (partitional)



# K-Means Demonstration

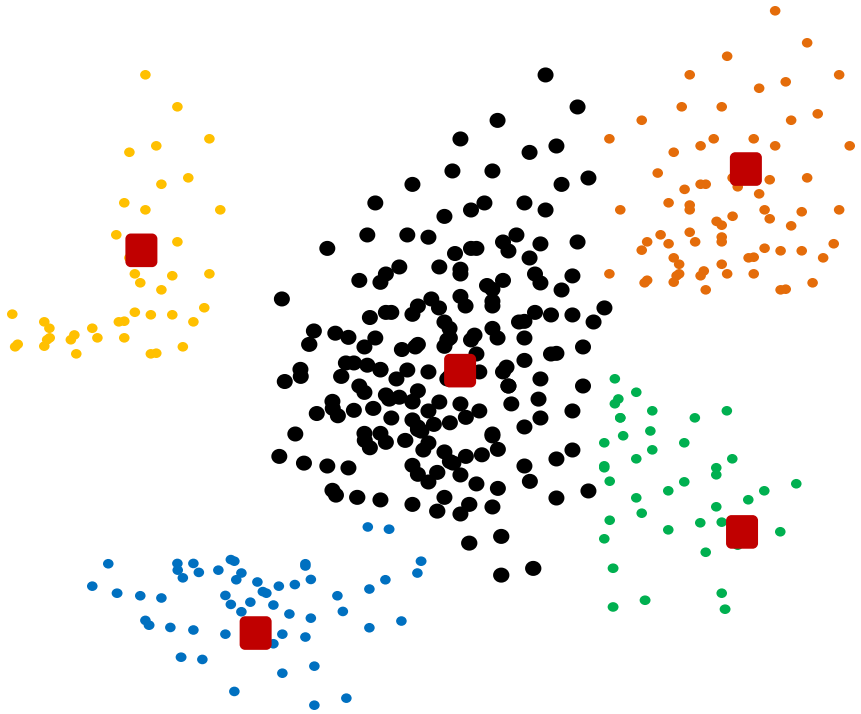
Here is the initial data set

# K-Means Demonstration



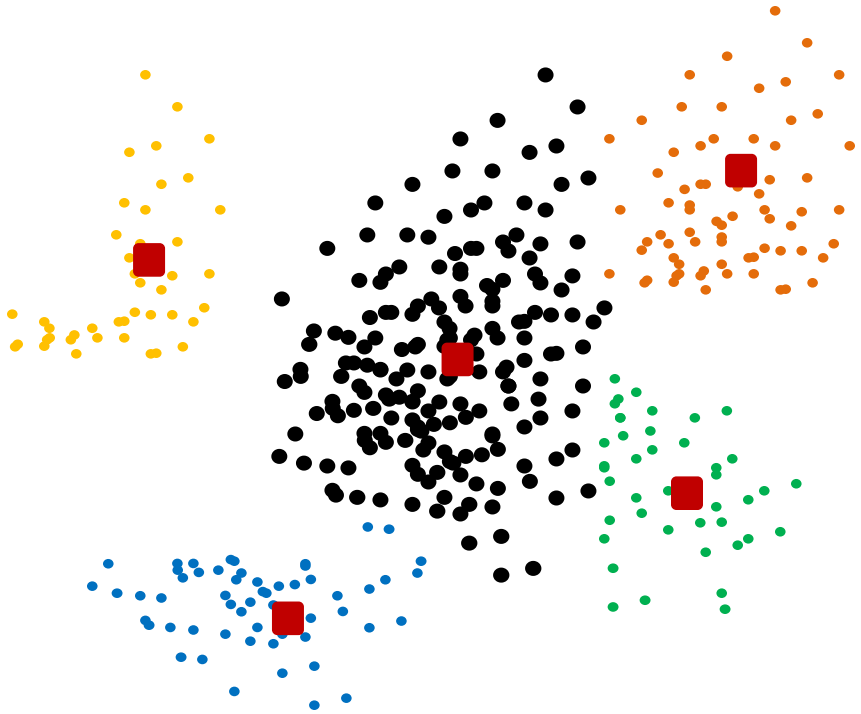
Choose K points as initial centroids

# K-Means Demonstration



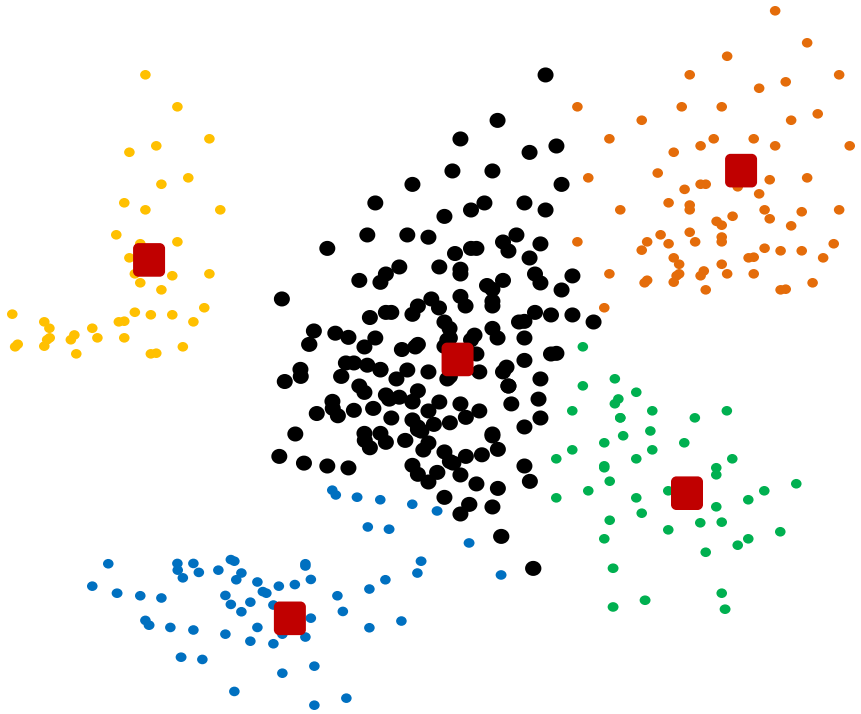
Assign data points according to distance

# K-Means Demonstration



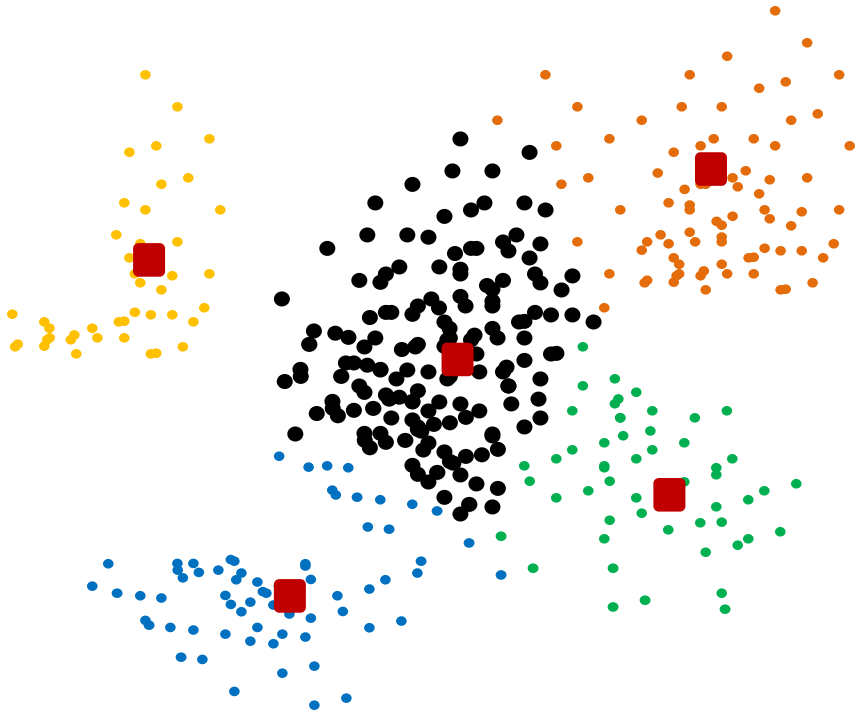
Recalculate the centroids

# K-Means Demonstration



And re-assign the points

# K-Means Demonstration



And keep doing that until you settle on a final set of clusters

# Another Interaction Demonstration

http://cs.joensuu.fi/sipu/clustering/animator/

# Choosing the initial centroids

There's no single, best way to choose initial centroids

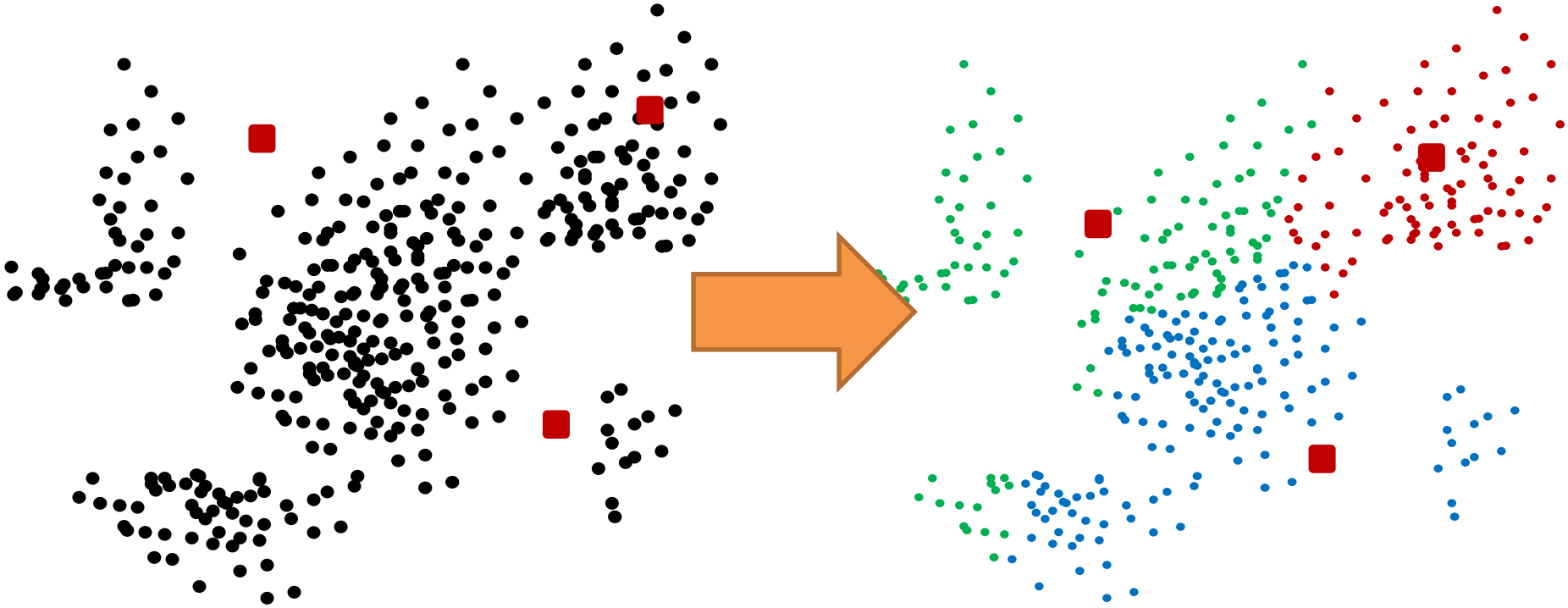
## It matters

- Choosing the right number
- Choosing the right initial location

## Bad choices create bad groupings

- They won't make sense within the context of the problem
- Unrelated data points will be included in the same group

# Example of Poor Initialization



This may “work” mathematically but the clusters don’t make much sense.

# Pre-processing: Normalize (Standardize) the data

- Reduces dispersion and influence of outliers
- Adjusts for differences in scale (income in dollars versus age in years)

- Normalizing data means
  - Subtracting with the average and dividing by the standard deviation
  - This puts data into a “standard” distribution
    - New average = 0
    - New standard deviation = 1

# Normalization: A Numeric Example

- If we have 10 students' age values:
  - 19, 22, 23, 20, 20, 21, 21, 22, 18, 20
  - Average: 20.6; standard deviation: 1.51
  - Normalizing 1st student's age: (19-20.6)/1.506= -1.06
  - Similarly, normalizing 2nd student's age: (22-20.6)/1.506= 0.93

- The normalized age values would be
  - -1.06, 0.93, 1.59, -0.40, -0.40, 0.27, 0.27, 0.93, -1.73, -0.40

# Normalization: A Numeric Example

If we have 10 students' age values:

| Age |
|---|
| 19 |
| 22 |
| 23 |
| 20 |
| 20 |
| 21 |
| 21 |
| 22 |
| 18 |
| 20 |

Average: 20.6
Standard deviation: 1.51

The normalized age values would be:

| Age |
|---|
| -1.06 |
| 0.93 |
| 1.59 |
| -0.40 |
| -0.40 |
| 0.27 |
| 0.27 |
| 0.93 |
| -1.73 |
| -0.40 |

New average: 0
New standard deviation: 1

- Normalizing 1st student's age: (19-20.6)/1.51= -1.06
- Normalizing 2nd student's age: (22-20.6)/1.51= 0.93

# Pre-processing: Remove outliers

- Also reduces dispersion that can skew the cluster centroids
- They don’t represent the population anyway

# Evaluating K-means Clusters

- On the previous slides, we did it visually, but there is a mathematical test

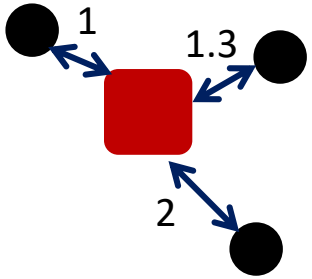
- Sum-of-Squares Error (SSE)
  - The distance to the nearest cluster center
  - How close does each point get to the center?

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

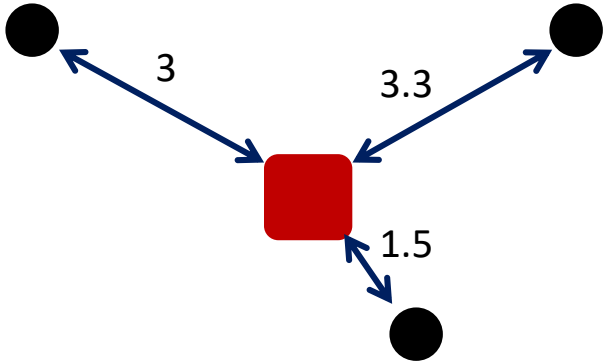
  - This just means
    - For each cluster $i$, compute distance from a point ($x$) to the cluster center ($m_i$)
    - Square that distance (so sign isn't an issue)
    - Add them all together
  - Hence SSE is always non-negative (>=0)

# Example: SSE within a Cluster (Within-cluster SSE)

**Cluster 1**




**Cluster 2**




$SSE_1 = 1^2 + 1.3^2 + 2^2 = 1 + 1.69 + 4 = 6.69$

$SSE_2 = 3^2 + 3.3^2 + 1.5^2 = 9 + 10.89 + 2.25 = 22.14$

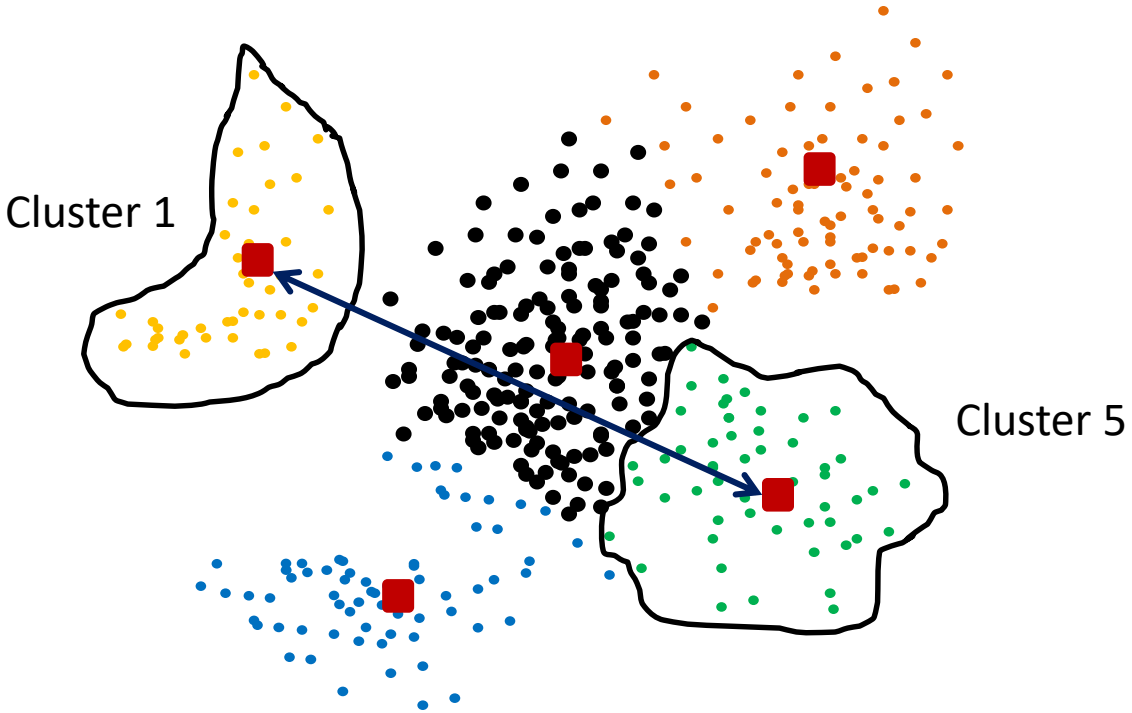
**Considerations**

- Lower individual cluster SSE = a better cluster
- Lower total SSE = a better set of clusters
- **More clusters will reduce SSE**

Reducing SSE within a cluster increases **cohesion** **(we want that)**

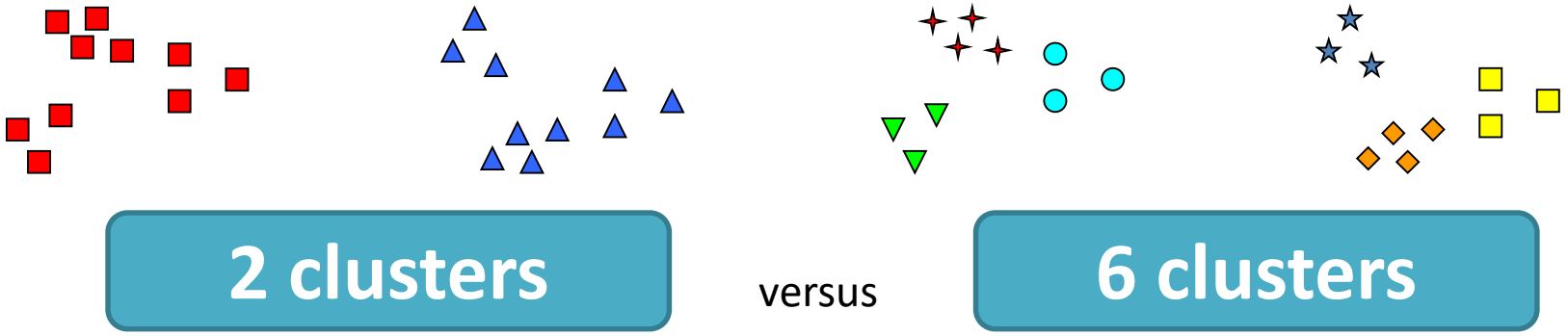
# SSE between clusters (Between-cluster SSE)

- Most common: distance between centroids
- Also can use SSE
  - Look at distance between cluster 1's points and other centroids
  - You'd want to *maximize* SSE ***between*** clusters




Increasing SSE across clusters increases **separation** **(we want that)**

# Trade-off: Cohesion versus Separation




- More clusters → higher cohesion (good)
  (lower within cluster SSE)
- More clusters → lower separation (bad)
  (lower between cluster SSE)

# Choosing the number of clusters (K)

- Can be determined by external reasons
- In many cases, there's no single answer…

- **But here's what we can do:**
  - Choose solutions with the fewest possible clusters.
  - But also make sure the clusters are describing distinct groups (separation).
  - Make sure that the range of values on each variable within a cluster is not too large to be useful (cohesion).

# Figuring out if our clusters are good

- "Good" means
  - Meaningful
  - Useful
  - Provides insight
- How to interpret the clusters?
  - Obtain summary statistics
  - Label the clusters
- This is somewhat subjective and depends upon the expectations of the analyst

# The pitfalls

- Poor clusters reveal incorrect associations
- Poor clusters reveal inconclusive associations
- There might be room for improvement and we can’t tell

# Limitations of K-Means Clustering

K-Means gives unreliable results when

- Clusters vary widely in size
- Clusters vary widely in density
- Clusters are not in rounded shapes
- The data set has a lot of outliers

The clusters may **never** make sense.
In that case, the data may just not be well-suited for clustering!

# The Keys to Successful Clustering

- We want high **cohesion** within clusters (minimize differences)
  - Low within cluster SSE
- And high **separation** between clusters (maximize differences)
  - High between cluster SSE
- Choose the right number of clusters
- Choose the right initial centroids
- No easy way to do this
- Trial-and-error, knowledge of the problem, and looking at the output

In R, **cohesion** is measured by **within cluster sum of squares error**…

…and **separation** measured by **between cluster sum of squares error**

# Classification versus Clustering

| Differences | Classification | Clustering |
|---|---|---|
| Prior Knowledge of classes (categories) | Yes | No |
| Goal | Classify new cases into known classes | Suggest groups based on patterns in data |
| Algorithm | Decision Trees | K-means |
| Type of model | Predictive<br>(to predict future data) | Descriptive<br>(to understand/explore data) |