CS5228 LECTURE 2: CLUSTERING

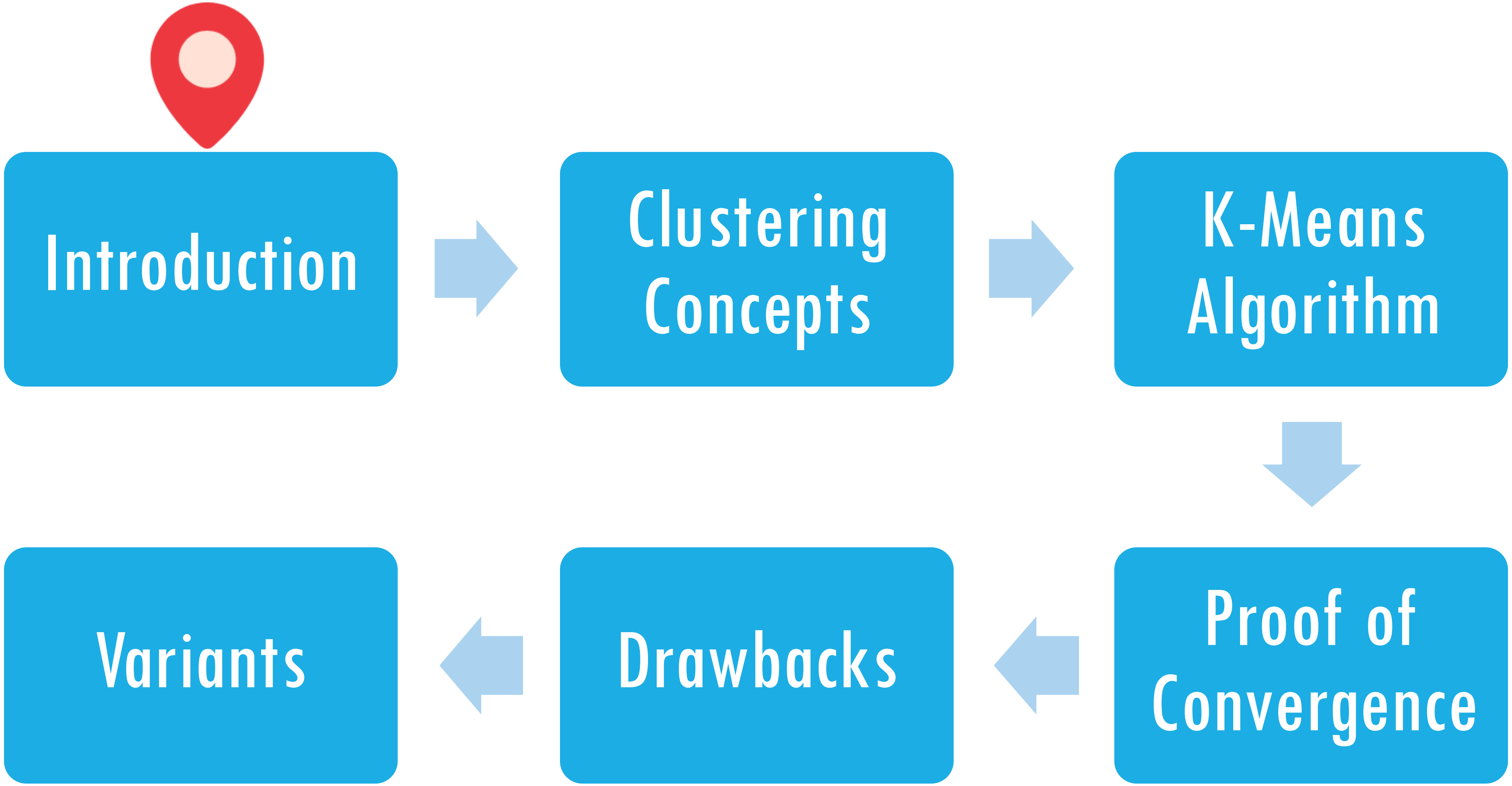
Bryan Hooi

School of Computing

National University of Singapore

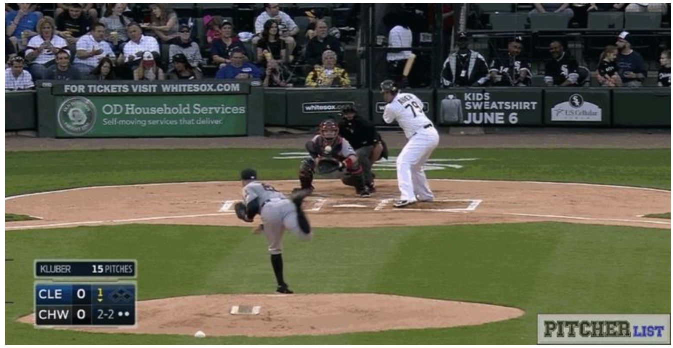
Slide credits: Wang Wei, Ng See Kiong, Wynne Hsu

OUTLINE


Introduction
Clustering Concepts
K-Means Algorithm
Proof of Convergence
Drawbacks
Variants


at&t


mariners.com
T-Mobile
IWAKUMA 33 PITCHES
TB 0
SEA 1 0-2
PITCHER GIFS


FOR TICKETS VISIT WHITESOX.COM
OD Household Services
KIDS SWEATSHIRT JUNE 6
KLUBER 15 PITCHES
CLE 0
CHW 0 2-2
PITCHER LIST


KIA 4 5 삼성
삼성카드 삼성증권
SPOTV
대구은행
치킨더홈


FOR TICKETS VISIT WHITESOX.COM
U.S. Cellular
CROSSTOWN CUP
MAY 7-8
TICKETS AVAILABLE
SCHERZER 24 PITCHES
DET 0
CHW 0 0-2
PITCHER GIFS


SFGIANTS.COM/VOT
at&t


FOR TICKETS VISIT WHITESOX.COM
OD Household Services
KIDS SWEATSHIRT JUNE 6
KLUBER 15 PITCHES
CLE 0
CHW 0 2-2
PITCHER LIST

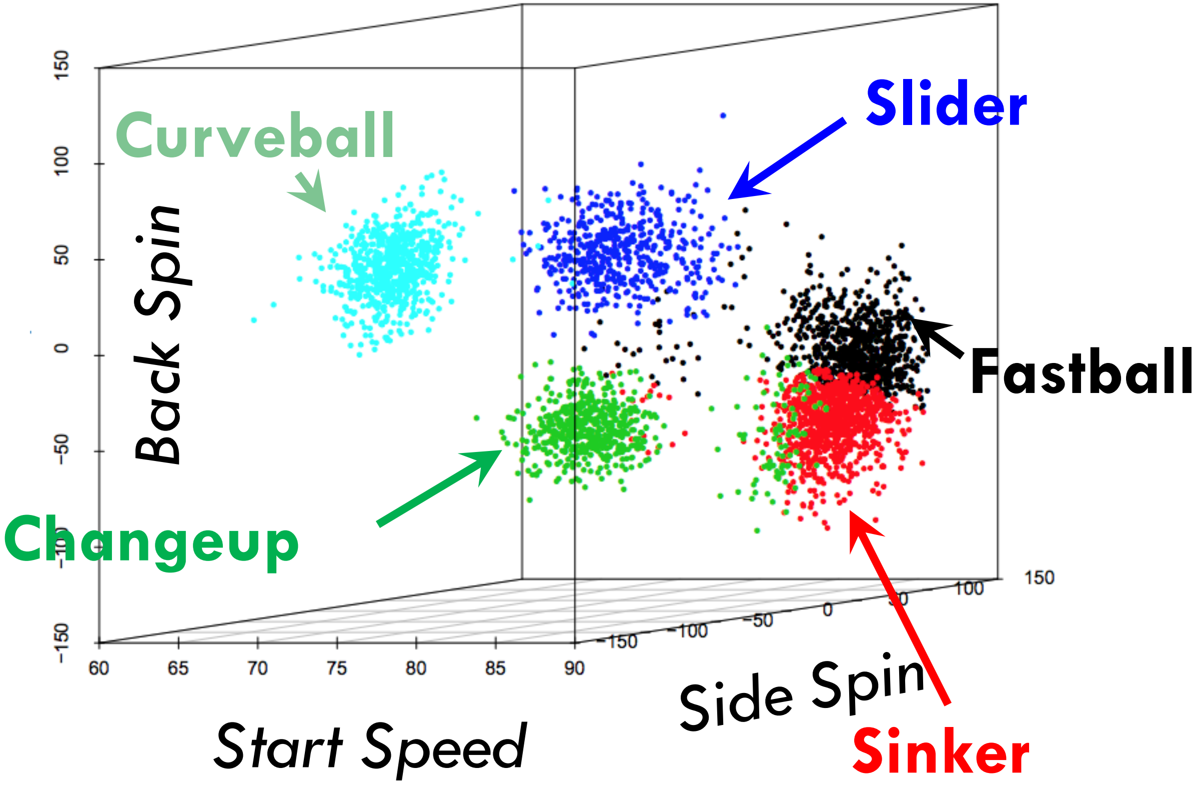

KIA 4 5 삼성
삼성카드 삼성증권
대구은행
치킨더홈
SPOTV

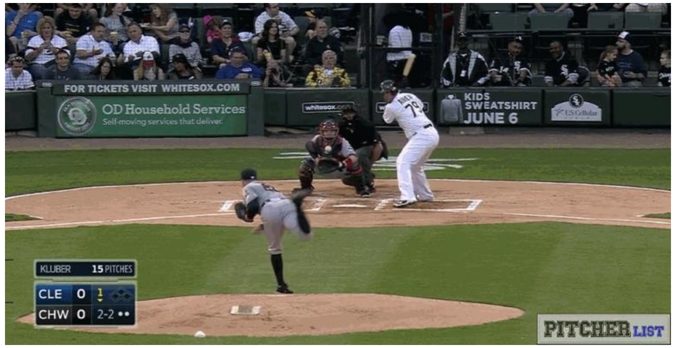

T-Mobile
IWAKUMA 33 PITCHES
TB 0
SEA 1 0-2
PITCHER GIFS


FOR TICKETS VISIT WHITESOX.COM
U.S. Cellular
CROSSTOWN CUP MAY 7-8 TICKETS AVAILABLE
SCHERZER 24 PITCHES
DET 0
CHW 0 0-2
PITCHER GIFS


SFGIANTS.COM/VO
at&t
FRA
TEMI


T · ·Mobile·
·Mobile·
mariners.com
IWAKUMA 33 PITCHES
TB 0
SEA 1 0-2
PITCHER GIFS


Curveball
Slider
Fastball
Changeup
Sinker
Back Spin
Start Speed
Side Spin
150
100
50
0
-50
-100
-150
60
65
70
75
80
85
90
-150
-100
-50
0
50
100
150


FOR TICKETS VISIT WHITESOX.COM
OD Household Services
KIDS SWEATSHIRT JUNE 6
KLUBER 15 PITCHES
CLE 0
CHW 0 2-2
PITCHER LIST

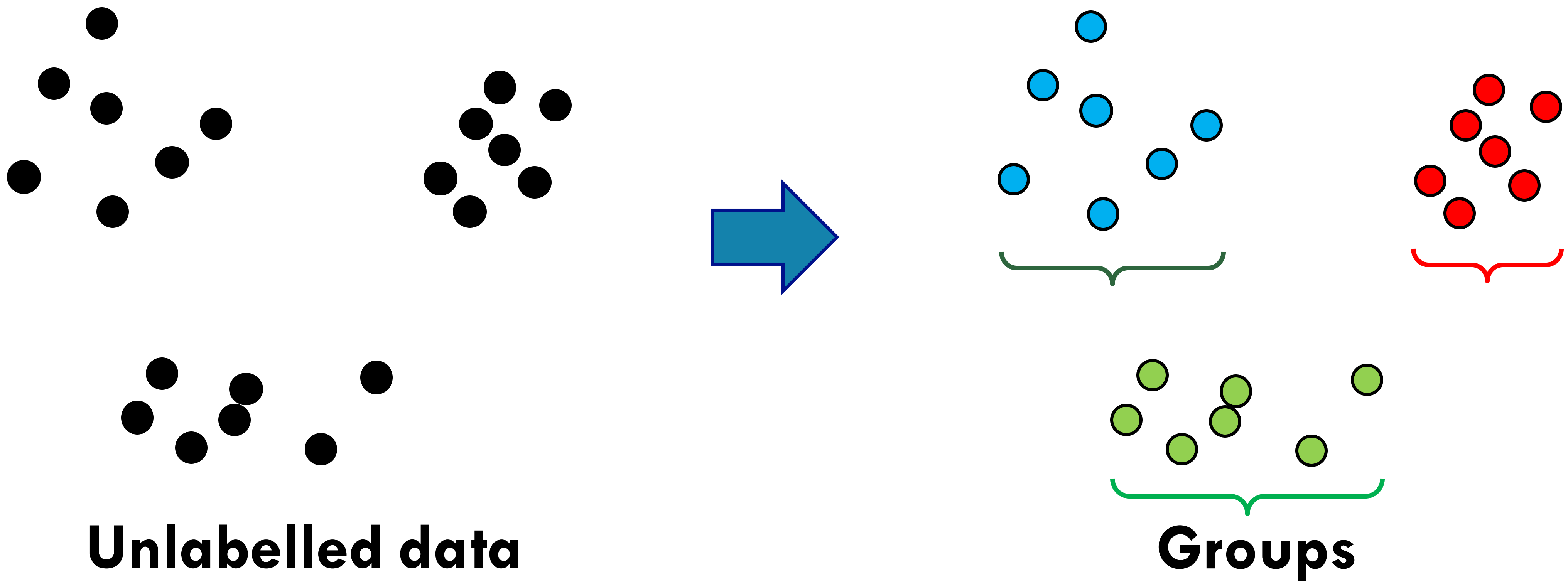

KIA 4 5 삼성
SPOTV
대구은행
치킨더홈
땅땅치킨


FOR TICKETS VISIT WHITESOX.COM
U.S. Cellular
CROSSTOWN CUP MAY 7-8 TICKETS AVAILABLE
SCHERZER 24 PITCHES
DET 0
CHW 0 0-2
PITCHER GIFS

GOAL OF CLUSTERING

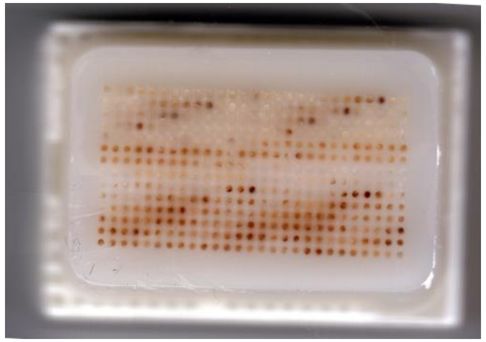
Clustering separates **unlabelled** data into **groups** of similar points.

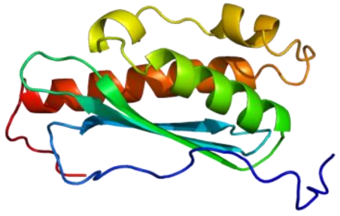
Clusters should have high **intra-cluster similarity**, and low **inter-cluster similarity**.



APPLICATIONS OF CLUSTERING

Many applications:





Microbiology: find groups of related genes (or proteins etc.)

Users you may be interested in

Recommendation & Social Networks: find groups of similar users

Google News

Trump, North Korea's Kim to hold second summit in late February
Channel NewsAsia · today

Trump to hold second summit with Kim Jong Un in February
The Straits Times · today

View more

Google

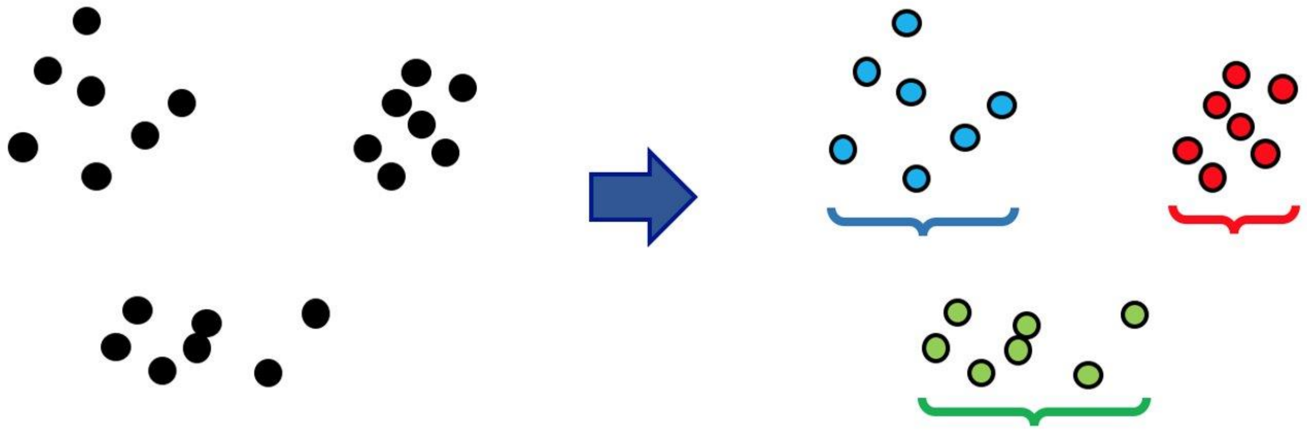
Introduction to K-means Clustering - DataScience.com
https://www.datascience.com/blog/k-means-clustering
Dec 6, 2016 - Learn data science with data scientist Dr. Andrea Trevino's step-by-step tutorial on the K-**means** clustering unsupervised machine learning ...

K Means
stanford.edu/~cpiech/cs221/handouts/kmeans.html
K-Means is one of the most popular "clustering" algorithms. **K-means** stores centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.

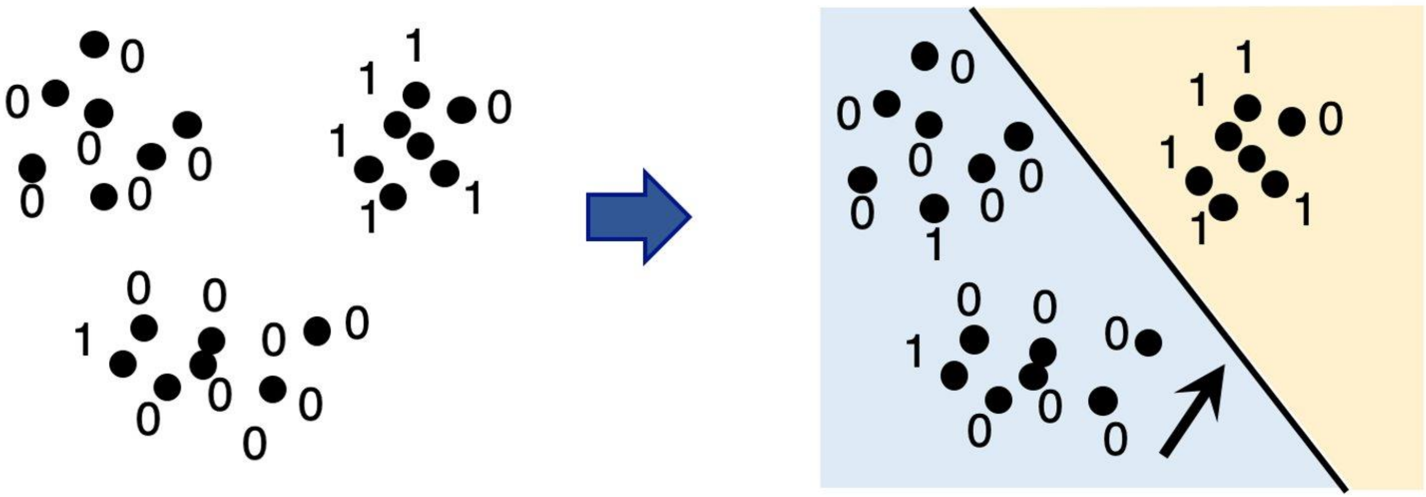
Search & Information Retrieval: grouping similar search (or news etc.) results

CLUSTERING VS. CLASSIFICATION

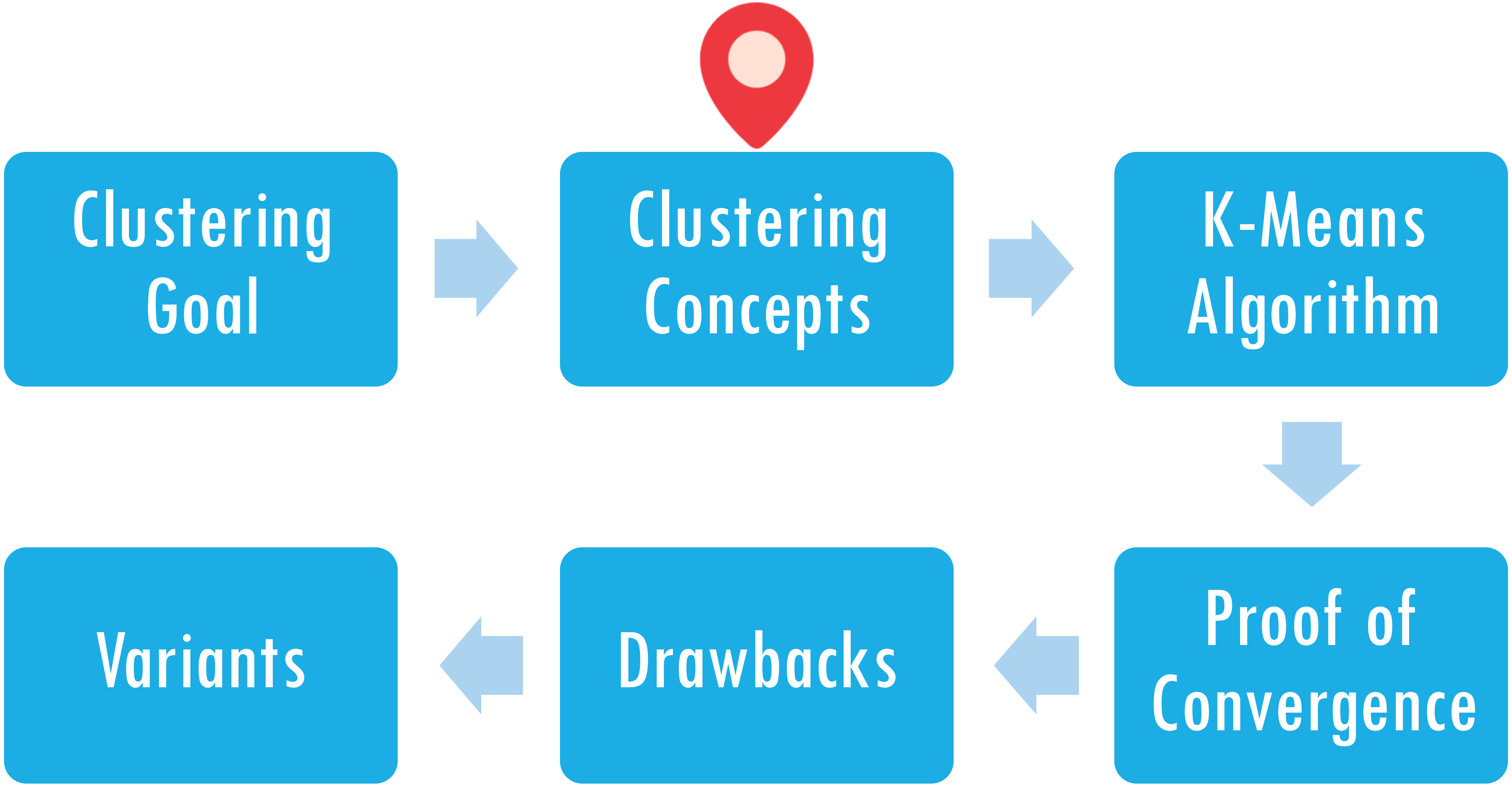
Clustering



Classification



OUTLINE



Clustering Goal
Clustering Concepts
K-Means Algorithm
Proof of Convergence
Drawbacks
Variants

WHAT DOES SIMILARITY MEAN?

(These are quite similar at the **pixel level**, but not **semantically**)

(In terms of their "meaning")

DEFINITION OF A DISTANCE METRIC

Given a set S, a **distance metric** is a **nonnegative** function $d : S \times S \to \mathbb{R}^{\geq 0}$ satisfying the properties:

Equivalent to

Nonnegative real numbers

- Uniqueness: $d(a, b) = 0 \Leftrightarrow a = b$

  (We don't want there to be objects that we cannot tell apart)

- Symmetry: $d(a, b) = d(b, a)$

  (If Alice is like Bob, then Bob is like Alice)

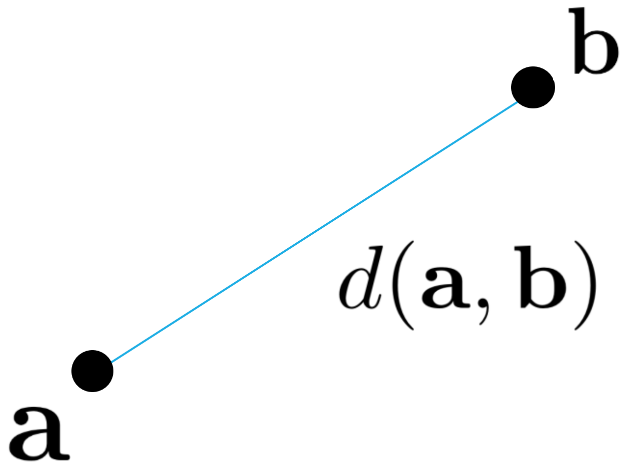
- Triangle Inequality: $d(a, b) \leq d(a, c) + d(c, b)$

  (Otherwise, Alice could be very like Carol, and Carol very like Bob, but Alice very unlike Bob)

COMMON DISTANCE / SIMILARITY METRICS

- **Euclidean distance**

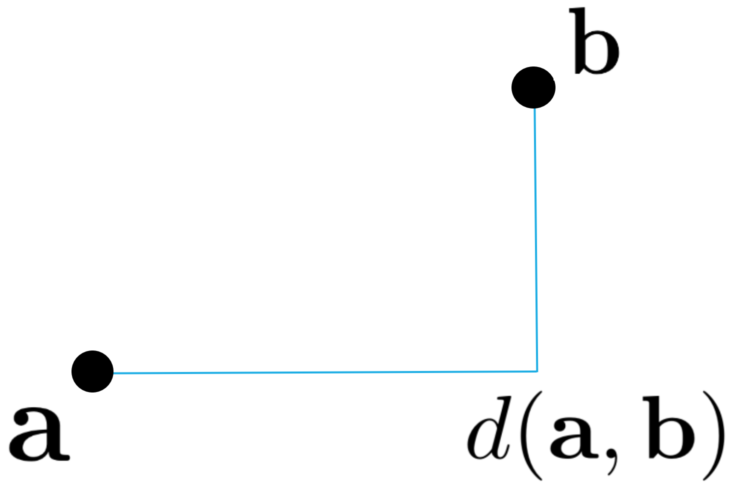
$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2)}$$



COMMON DISTANCE / SIMILARITY METRICS

- **Manhattan distance**

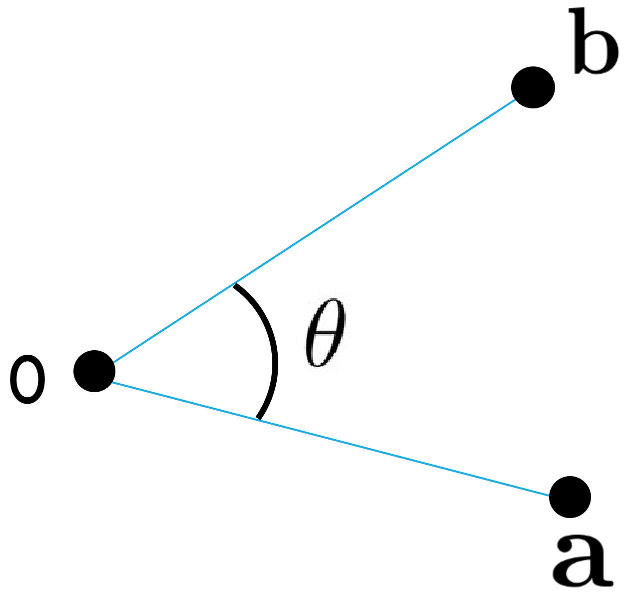
$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_1 = \sum_{i=1}^{p} |a_i - b_i|$$



COMMON DISTANCE / SIMILARITY METRICS

- **Cosine distance**

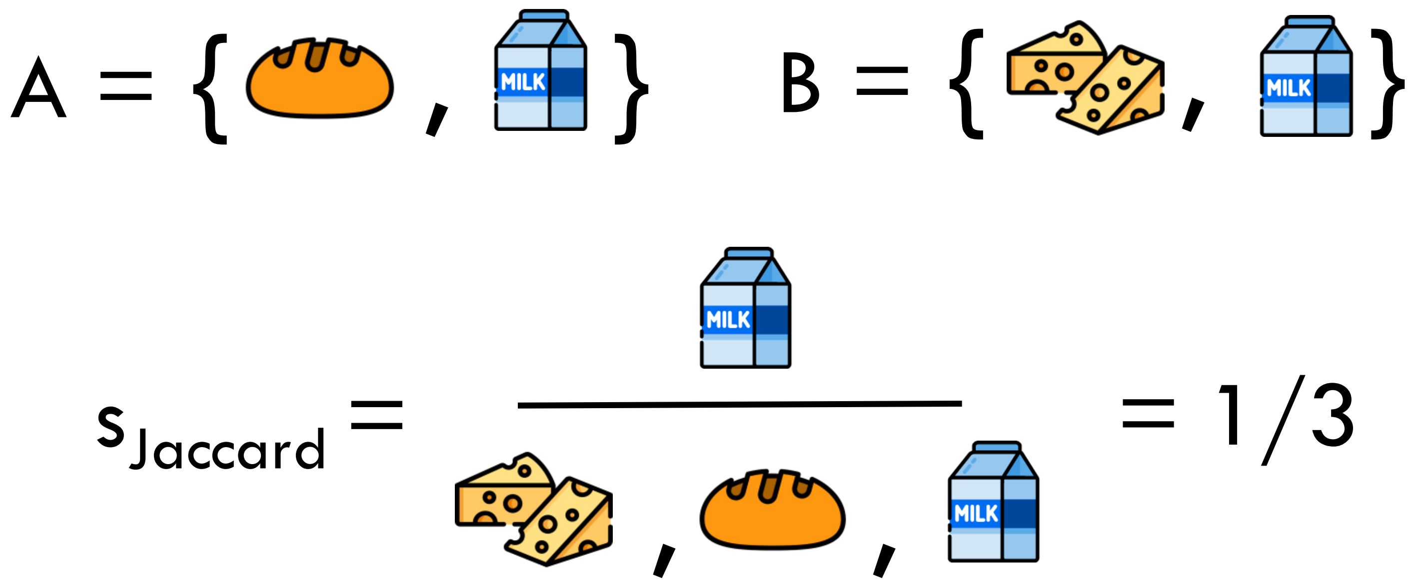
$$d(\mathbf{a}, \mathbf{b}) = \cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|a\| \cdot \|b\|}$$



COMMON DISTANCE / SIMILARITY METRICS

- **Jaccard Similarity** (between **sets** A and B)

$$s_{\text{Jaccard}}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

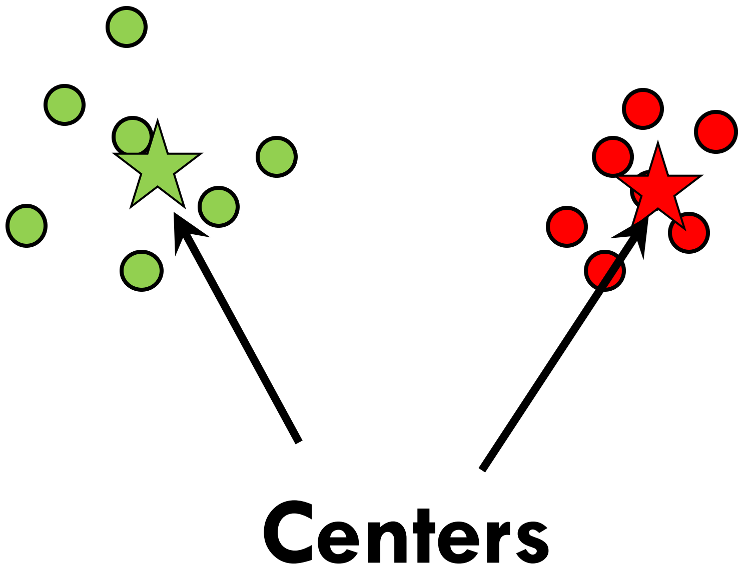


- **Jaccard Distance**

$$d_{\text{Jaccard}}(A, B) = 1 - s_{\text{Jaccard}}(A, B)$$

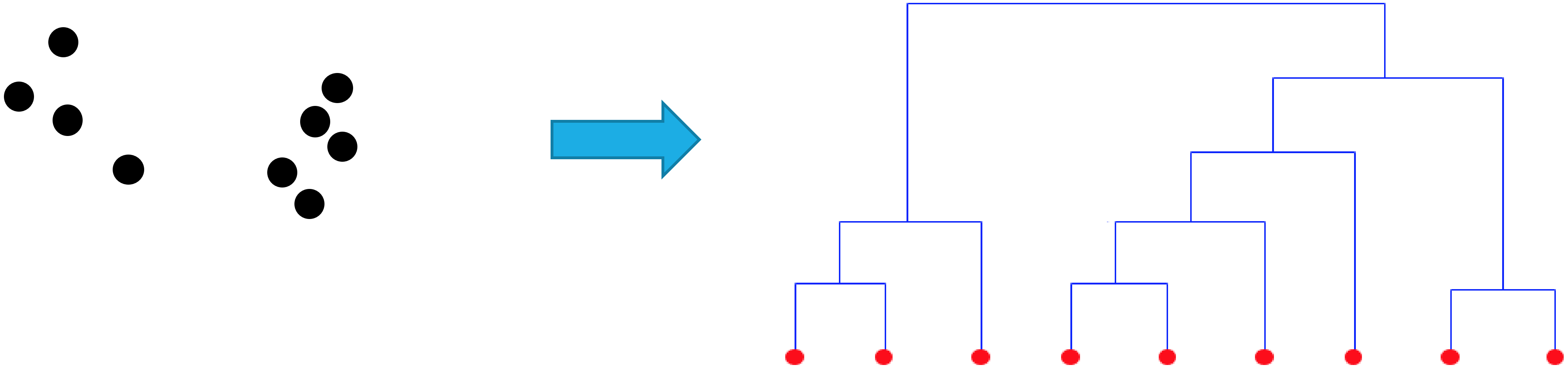
OVERVIEW OF CLUSTERING APPROACHES

- **Center-based:** each cluster is characterized by its center



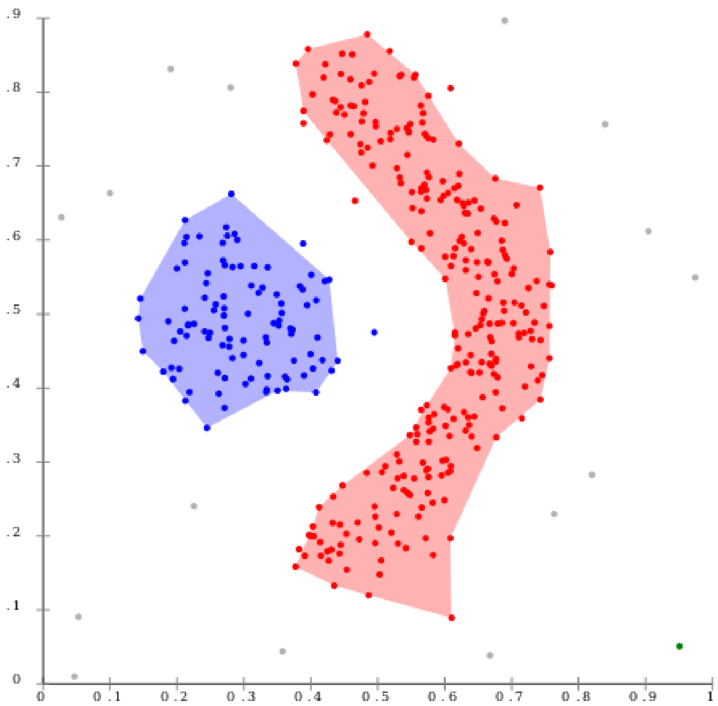
OVERVIEW OF CLUSTERING APPROACHES

- **Hierarchical:** points are organized according to a hierarchy (or tree structure)



OVERVIEW OF CLUSTERING APPROACHES

- **Density-based:** clusters are high-density regions surrounded by low-density regions



OUTLINE

1. Clustering Goal
2. Clustering Concepts
3. K-Means Algorithm
4. Proof of Convergence
5. Drawbacks
6. Variants

K-MEANS ALGORITHM: STEPS

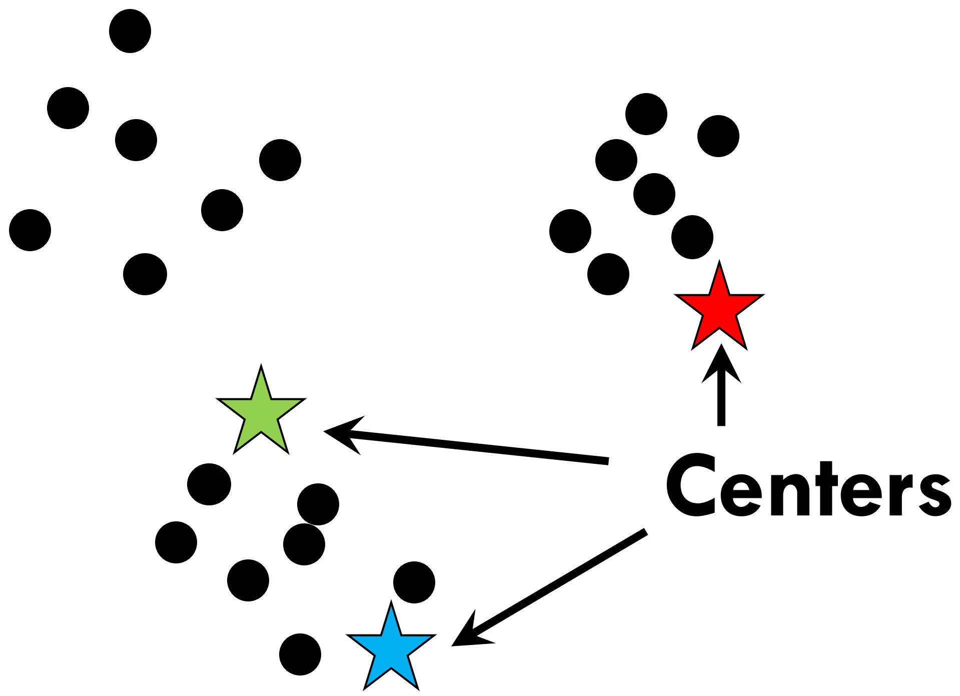
1. Initialization: Pick K random points as centers

2. Repeat:

a) **Assignment:** assign each point to nearest cluster

b) **Update:** move each cluster center to average of its assigned points

Stop if no assignments change



K-MEANS ALGORITHM: STEPS

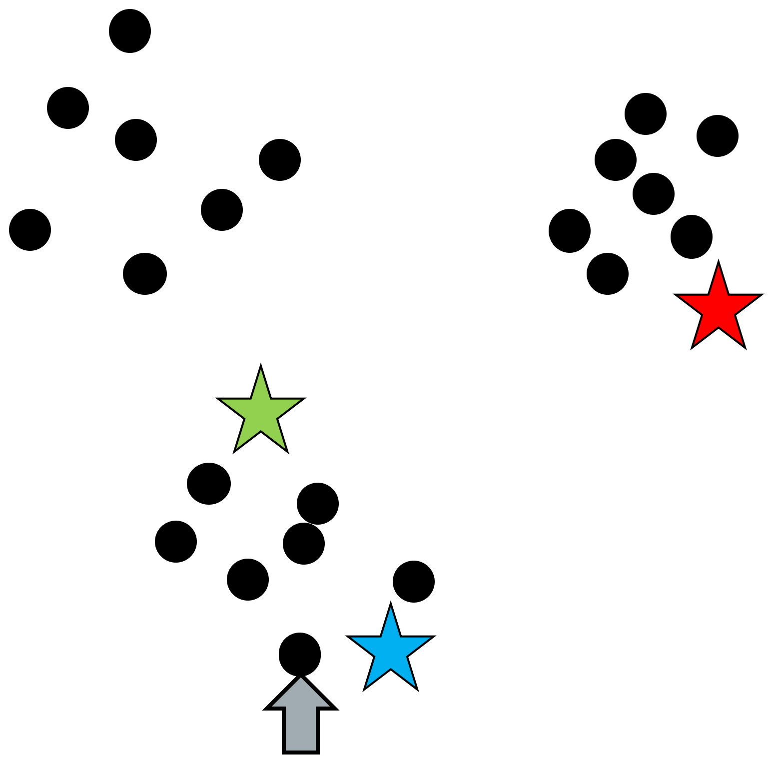
1. Initialization: Pick K random points as centers

2. Repeat:

a) Assignment: assign each point to nearest cluster

b) Update: move each cluster center to average of its assigned points

Stop if no assignments change



K-MEANS ALGORITHM: STEPS

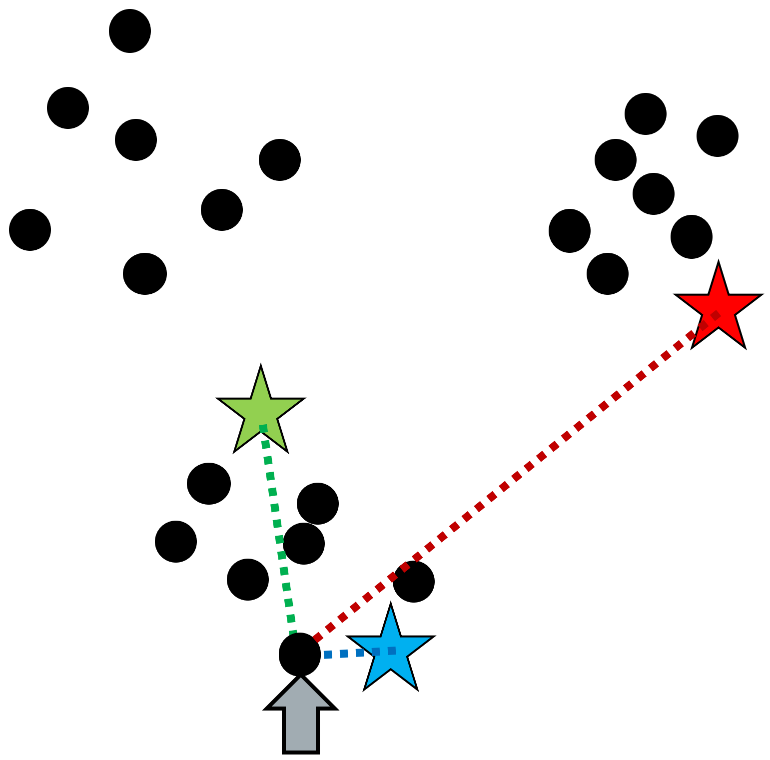
1. Initialization: Pick K random points as centers

2. Repeat:

a) Assignment: assign each point to nearest cluster

b) Update: move each cluster center to average of its assigned points

Stop if no assignments change



K-MEANS ALGORITHM: STEPS

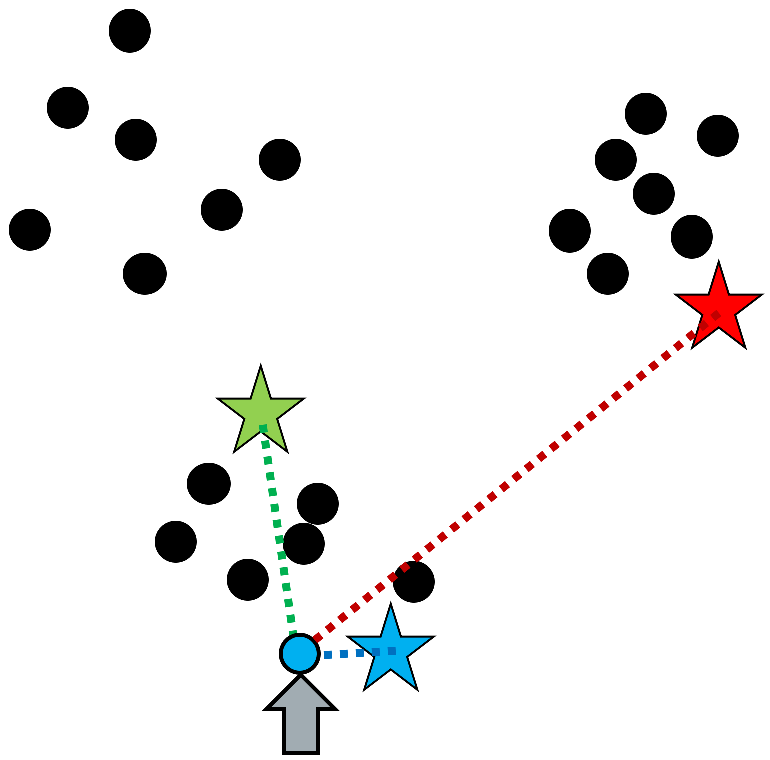
<u>1. Initialization</u>: Pick K random points as centers

2. Repeat:

a) <u>Assignment:</u> assign each point to nearest cluster

b) <u>Update:</u> move each cluster center to average of its assigned points

<u>Stop</u> if no assignments change



K-MEANS ALGORITHM: STEPS

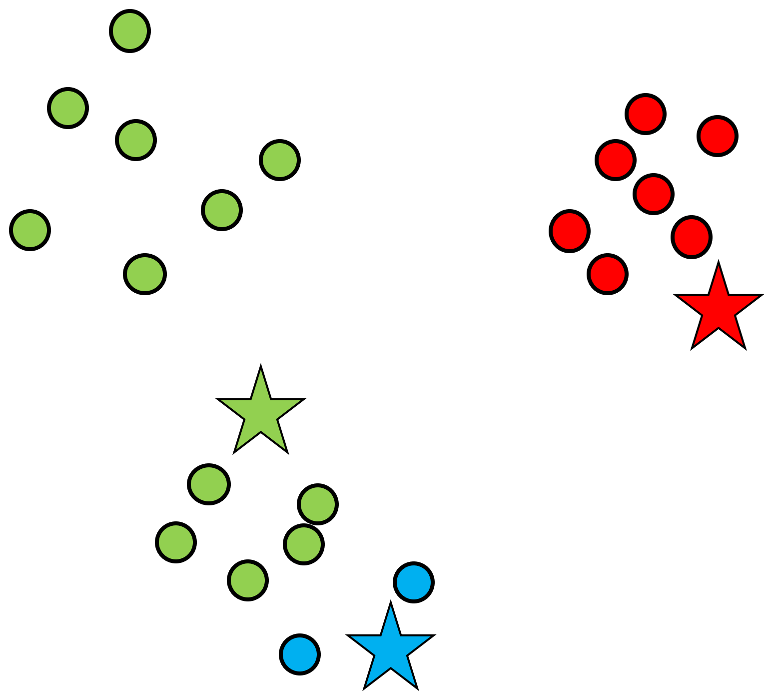
<u>1. Initialization</u>: Pick K random points as centers

2. Repeat:

a) <u>Assignment:</u> assign each point to nearest cluster

b) <u>Update:</u> move each cluster center to average of its assigned points

<u>Stop</u> if no assignments change



K-MEANS ALGORITHM: STEPS

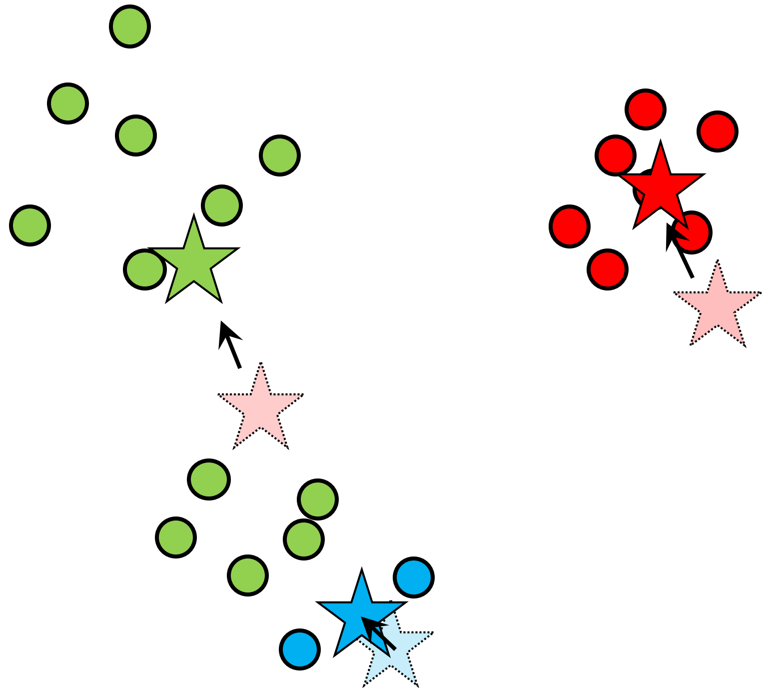
1. Initialization: Pick K random points as centers

2. Repeat:

a) **Assignment:** assign each point to nearest cluster

b) **Update:** move each cluster center to average of its assigned points

Stop if no assignments change



K-MEANS ALGORITHM: STEPS

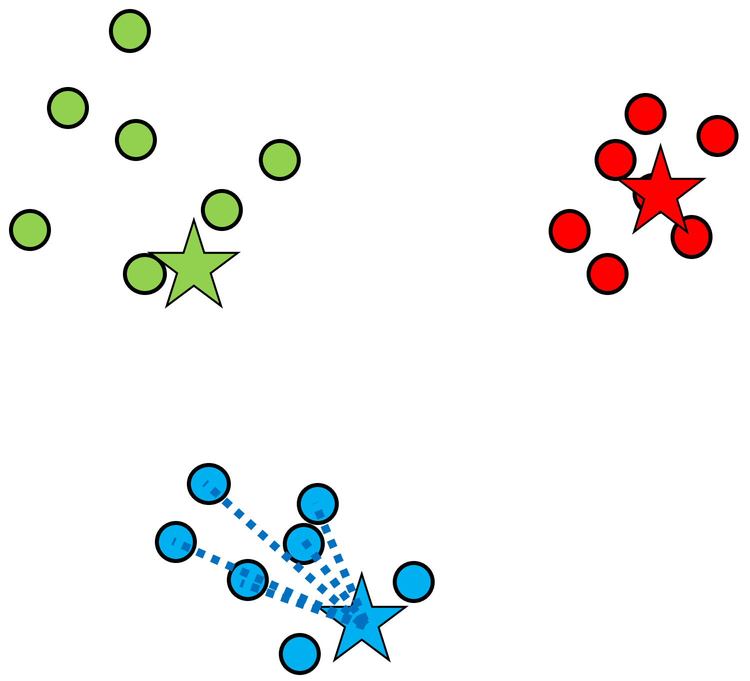
1. Initialization: Pick K random points as centers

2. Repeat:

a) **Assignment:** assign each point to nearest cluster

b) **Update:** move each cluster center to average of its assigned points

Stop if no assignments change



K-MEANS ALGORITHM: STEPS

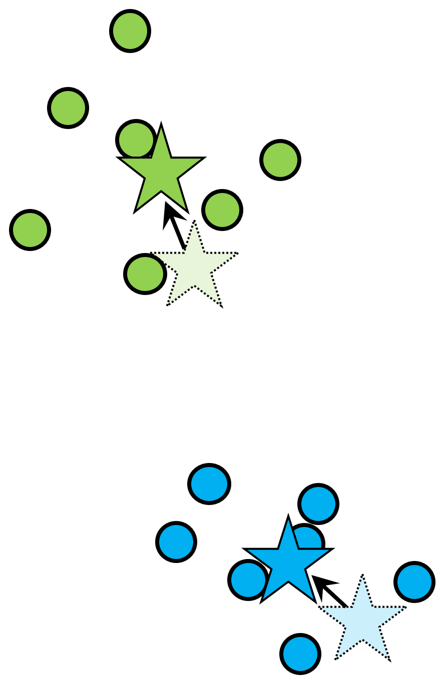
1. Initialization: Pick K random points as centers

2. Repeat:

a) **Assignment:** assign each point to nearest cluster

b) **Update:** move each cluster center to average of its assigned points

Stop if no assignments change



K-MEANS ALGORITHM: STEPS

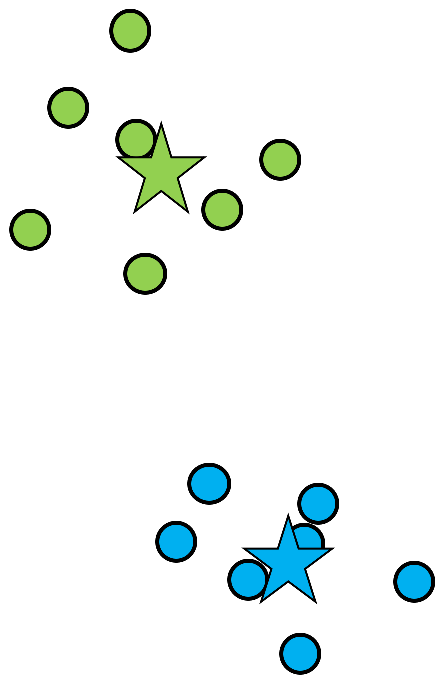
1. Initialization: Pick K random points as centers

2. Repeat:

a) **Assignment:** assign each point to nearest cluster

b) **Update:** move each cluster center to average of its assigned points

Stop if no assignments change



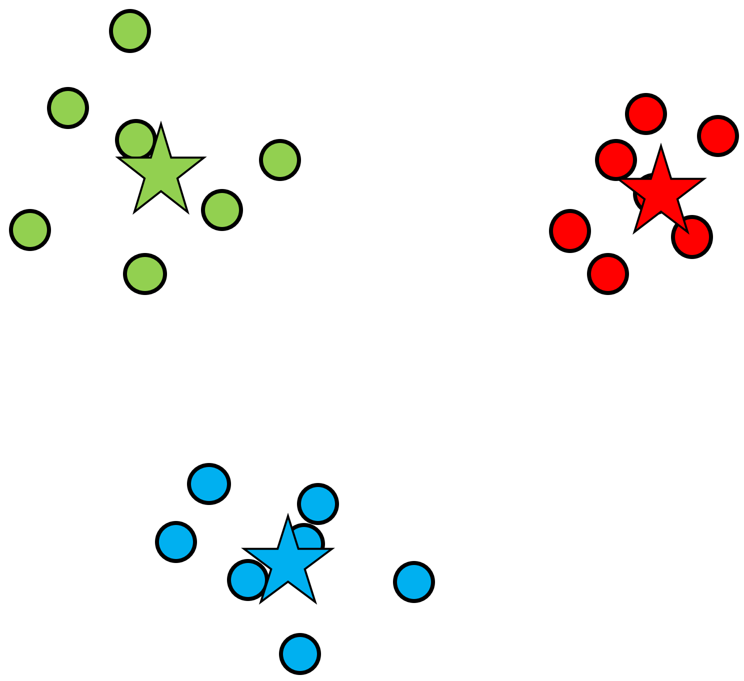
K-MEANS ALGORITHM: STEPS

1. Initialization: Pick K random points as centers

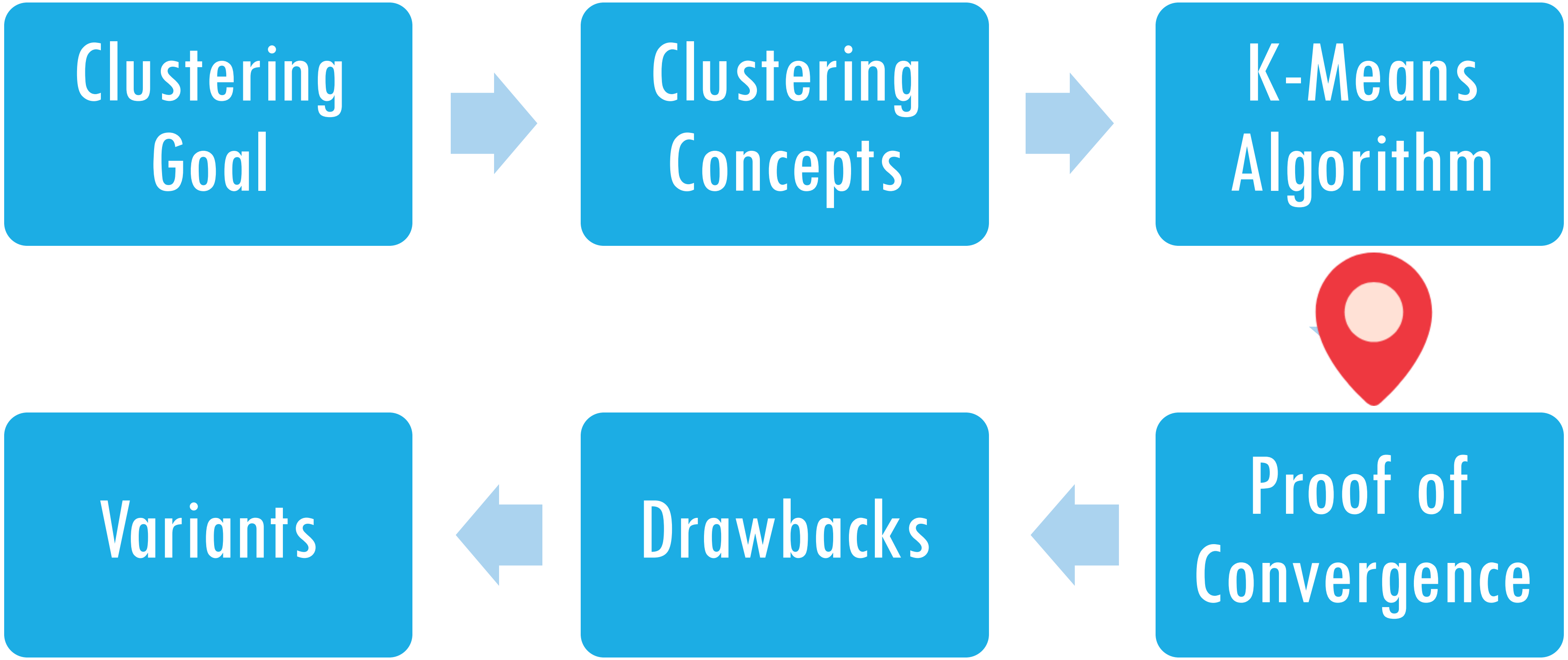
2. Repeat:

- a) **Assignment:** assign each point to nearest cluster
- b) **Update:** move each cluster center to average of its assigned points

Stop if no assignments change

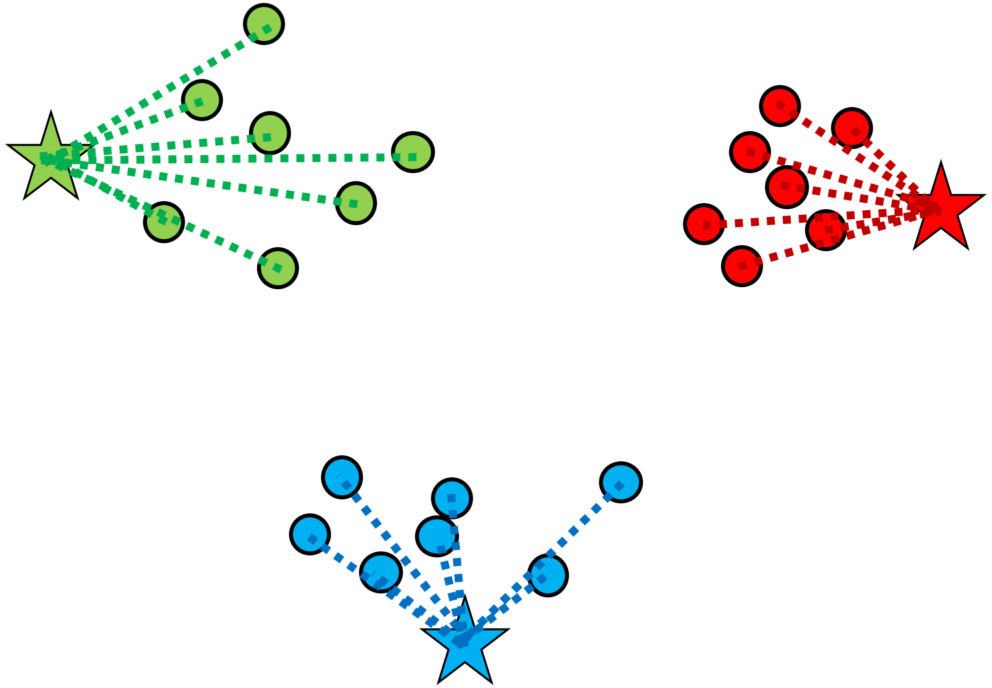


OUTLINE


Clustering Goal
Clustering Concepts
K-Means Algorithm
Proof of Convergence
Drawbacks
Variants

OPTIMIZATION OBJECTIVE

Within-Cluster Sum of Squares (WCSS): sum of squared distances between each point and its cluster center



ith cluster center

$$\mathrm{WCSS} = \sum_{i=1}^{K} \underbrace{\sum_{x \in C_i}}_{\text{Set of points in ith cluster}} \underbrace{\|x - c_i\|_2^2}_{\text{Squared distance between point and center}}$$

Sum over clusters

K-MEANS AS ALTERNATING MINIMIZATION

1. Initialization: Pick K random points as centers

2. Repeat:

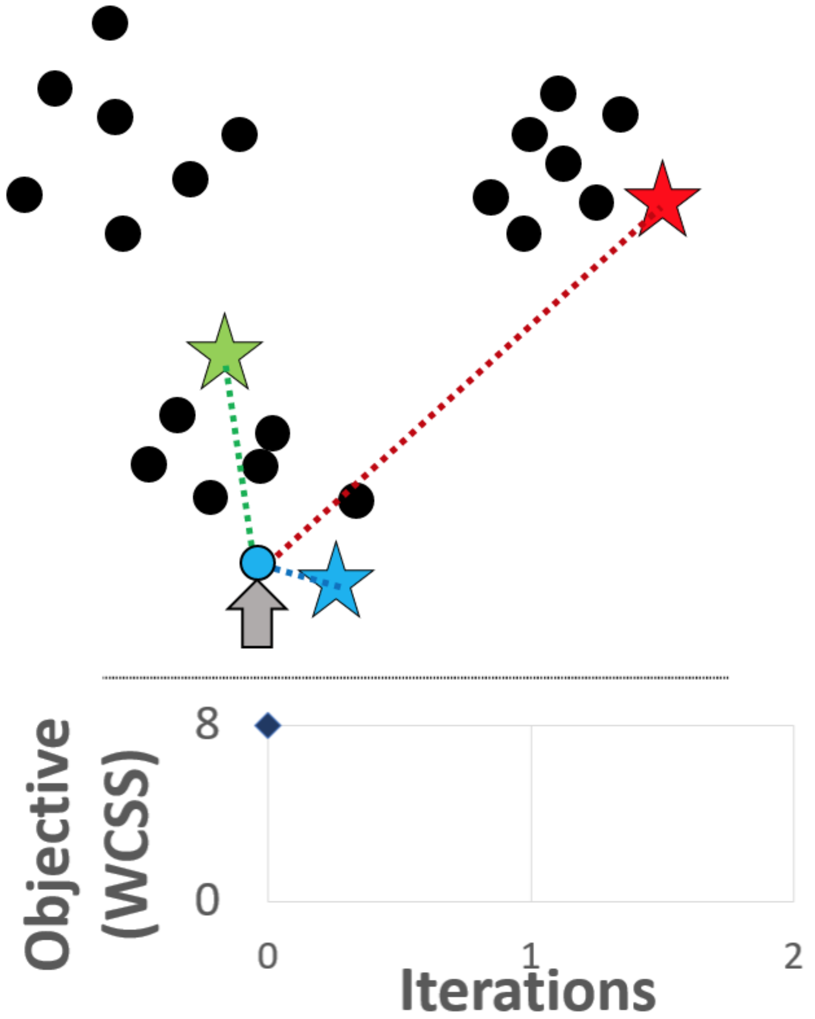
a) Assignment: assign each point to nearest cluster

$$\underset{C_1,\cdots,C_K}{\text{minimize}} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2$$

b) Update: move each cluster center to average of its assigned points

$$\underset{c_1,\cdots,c_K}{\text{minimize}} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2$$

Stop if no assignments change



K-MEANS AS ALTERNATING MINIMIZATION

1. Initialization: Pick K random points as centers

2. Repeat:

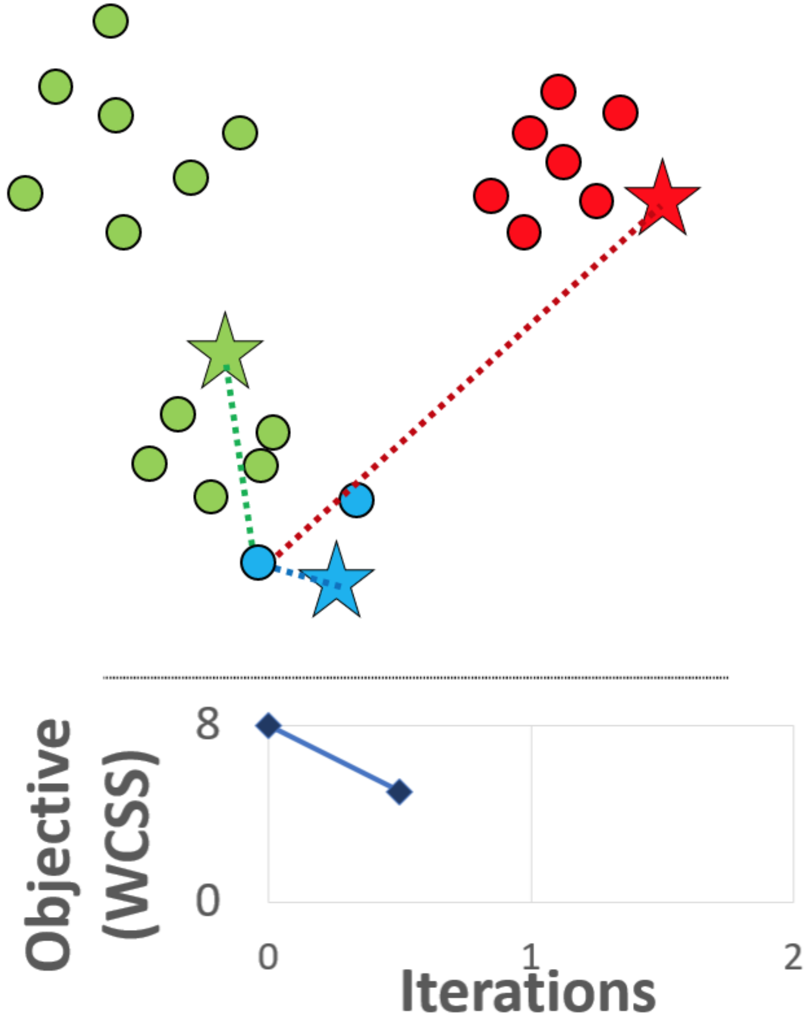
a) **Assignment:** assign each point to nearest cluster

$$\underset{C_1,\cdots,C_K}{\text{minimize}} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2$$

b) **Update:** move each cluster center to average of its assigned points

$$\underset{c_1,\cdots,c_K}{\text{minimize}} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2$$

Stop if no assignments change



K-MEANS AS ALTERNATING MINIMIZATION

1. **Initialization**: Pick K random points as centers

2. Repeat:

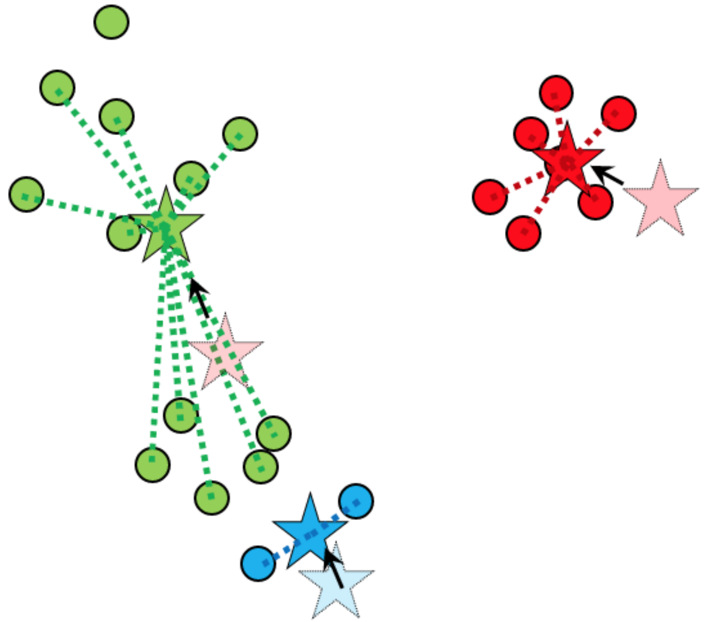
a) **Assignment:** assign each point to nearest cluster

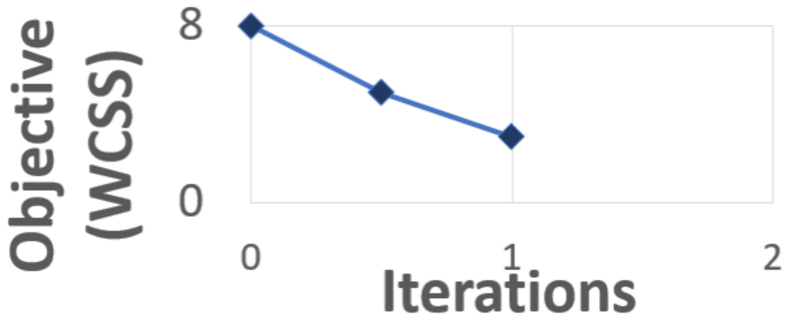
$$\underset{C_1,\cdots,C_K}{\text{minimize}} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2$$

b) Update: move each cluster center to average of its assigned points

$$\underset{c_1,\cdots,c_K}{\text{minimize}} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|_2^2$$

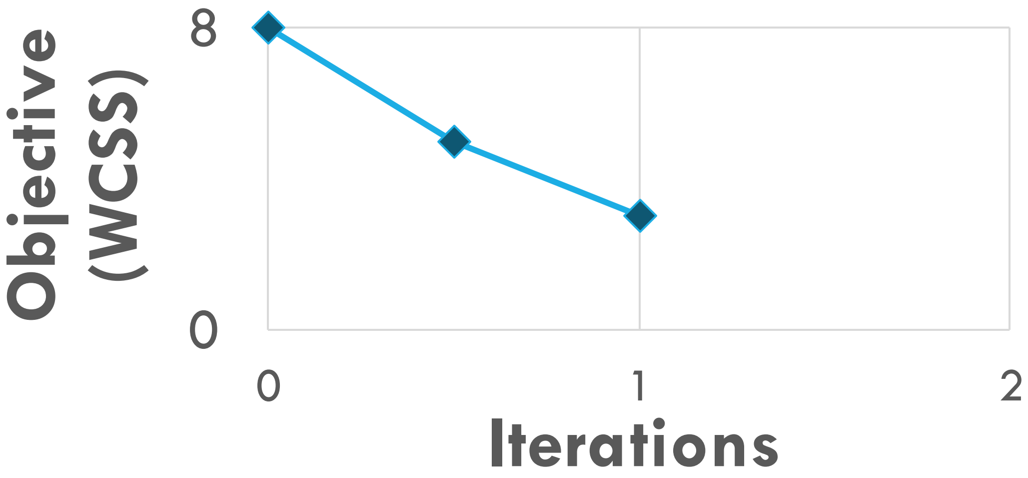
Stop if no assignments change



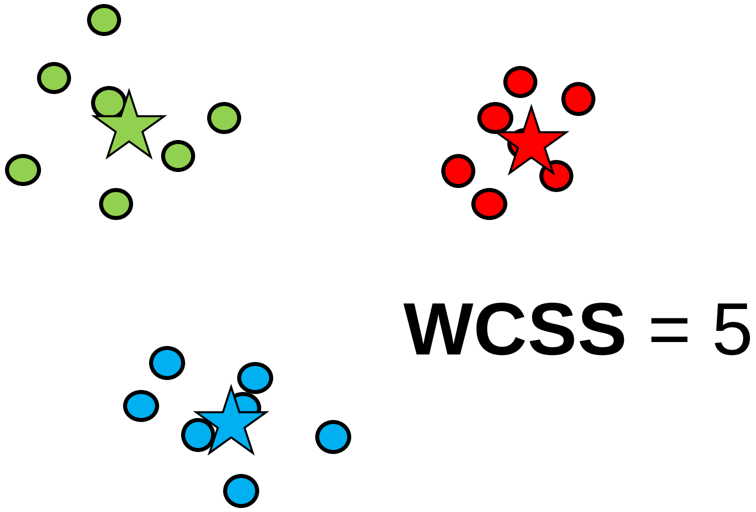


PROOF OF CONVERGENCE

1. The WCSS objective is strictly decreasing

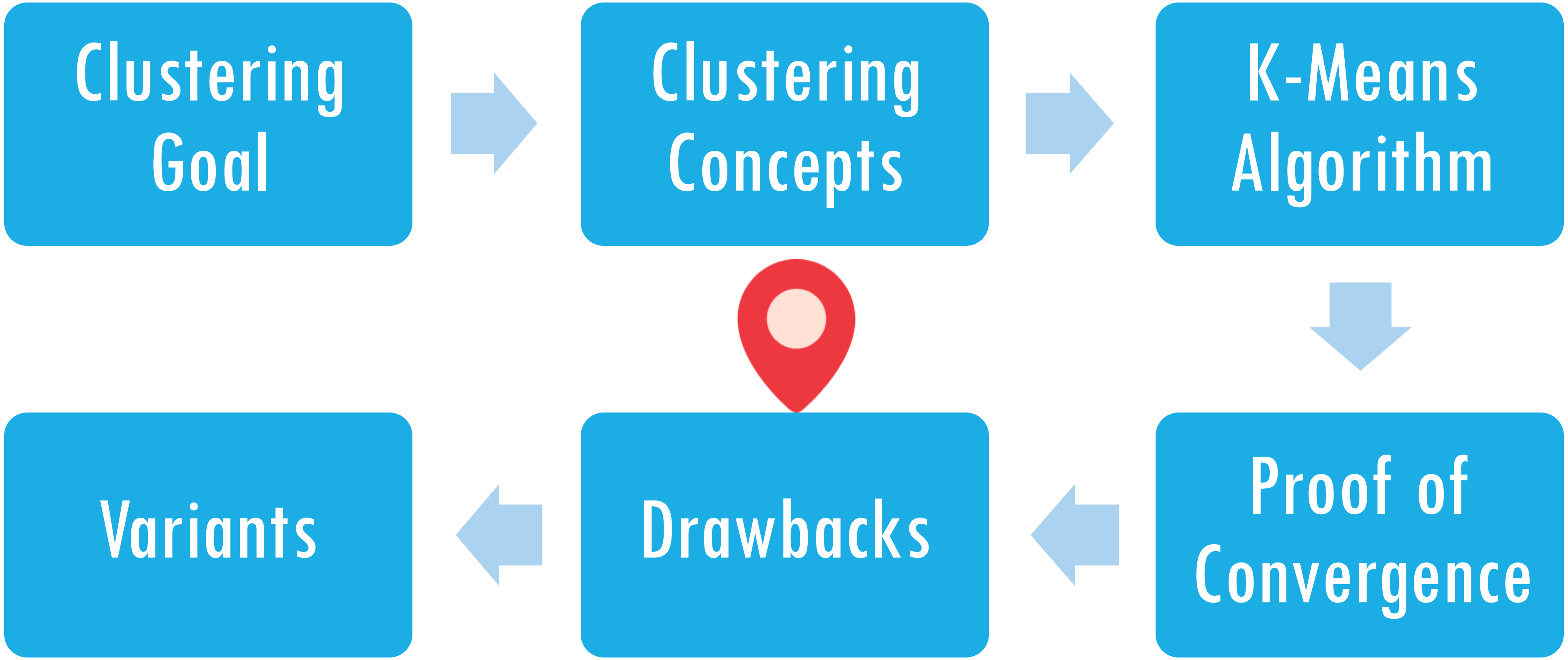


2. There are a finite number of possible clusterings



The algorithm must eventually stop!

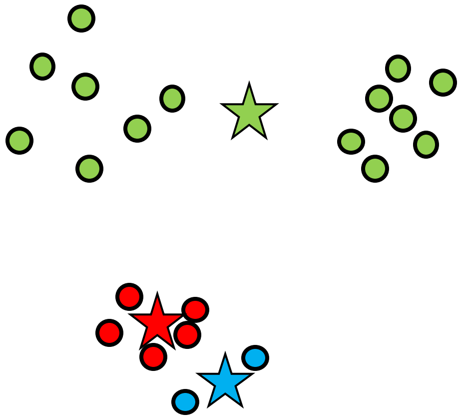
OUTLINE


Clustering Goal
Clustering Concepts
K-Means Algorithm
Proof of Convergence
Drawbacks
Variants

LOCAL, NOT GLOBAL OPTIMUM

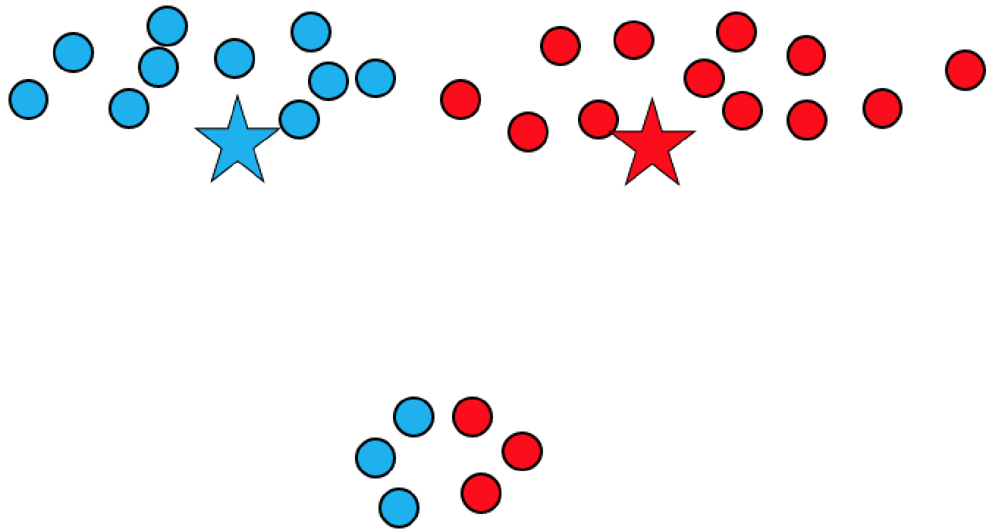
- The algorithm only returns a **local**, not a **global** optimum!
    - (Finding the global optimum is NP-hard)
- Initialization is important

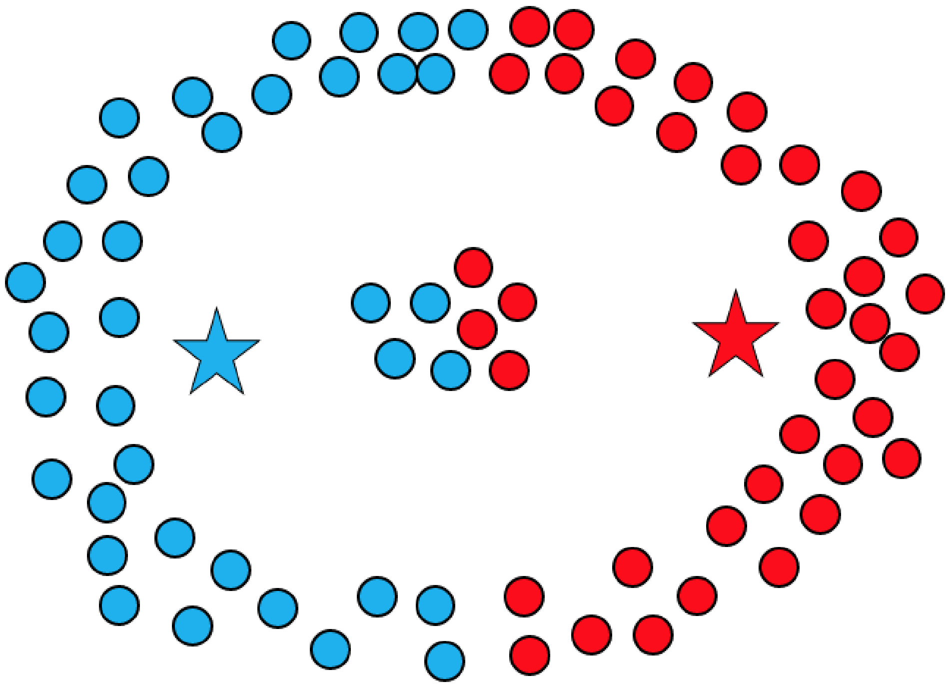
Example of a local minima



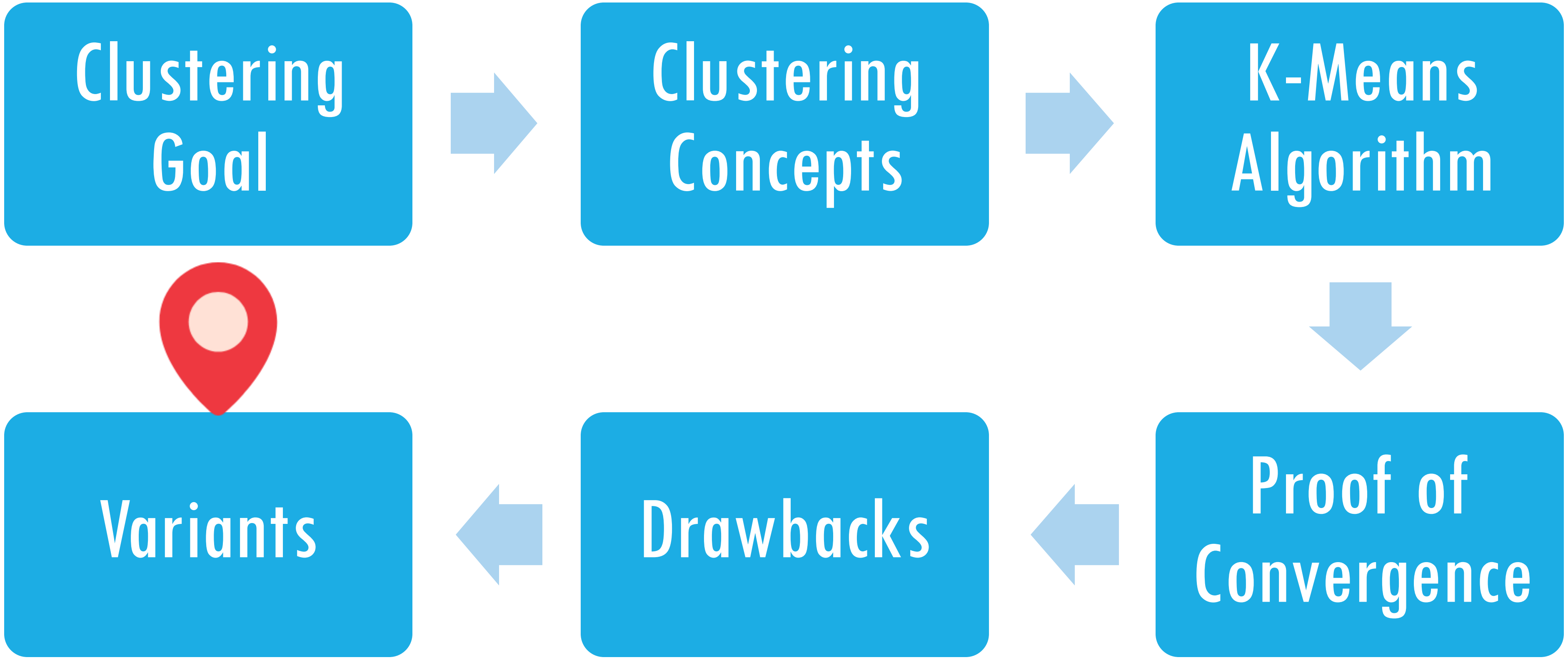
NON-SPHERE-LIKE CLUSTERS

- Optimization objective results in roughly sphere shaped clusters



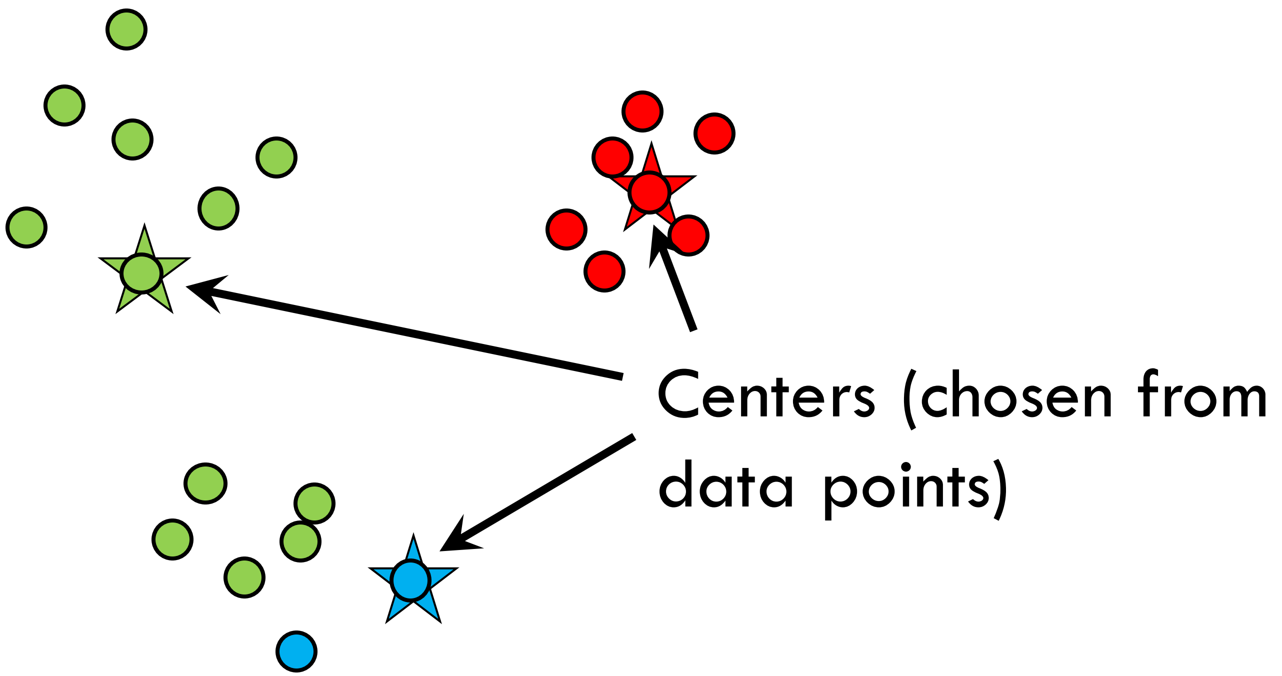


OUTLINE


Clustering Goal
Clustering Concepts
K-Means Algorithm
Proof of Convergence
Drawbacks
Variants

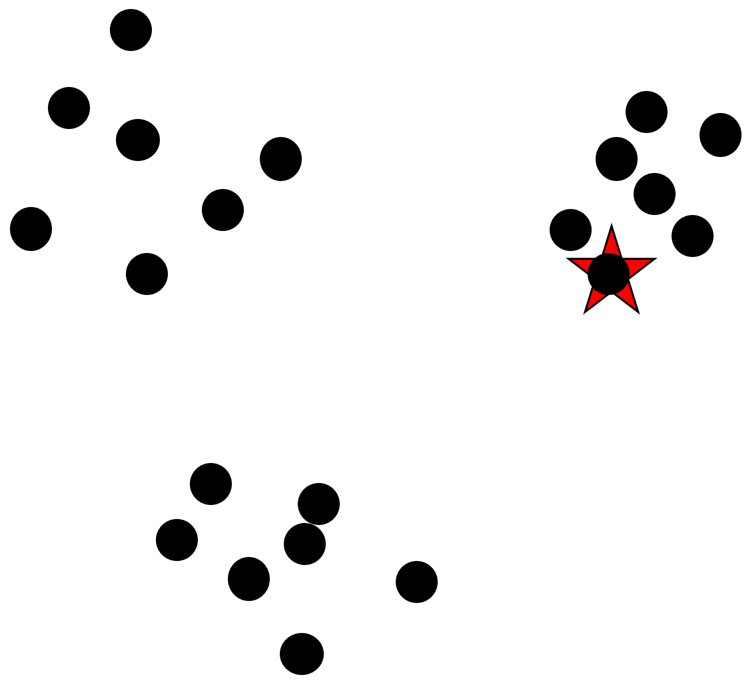
1. K-MEDOIDS ALGORITHM

- **K-Medoids:** like K-Means, but centers are chosen from data points
- Useful when:
  - We want data points as cluster representatives
  - Complex data types – we can only measure distances between data points



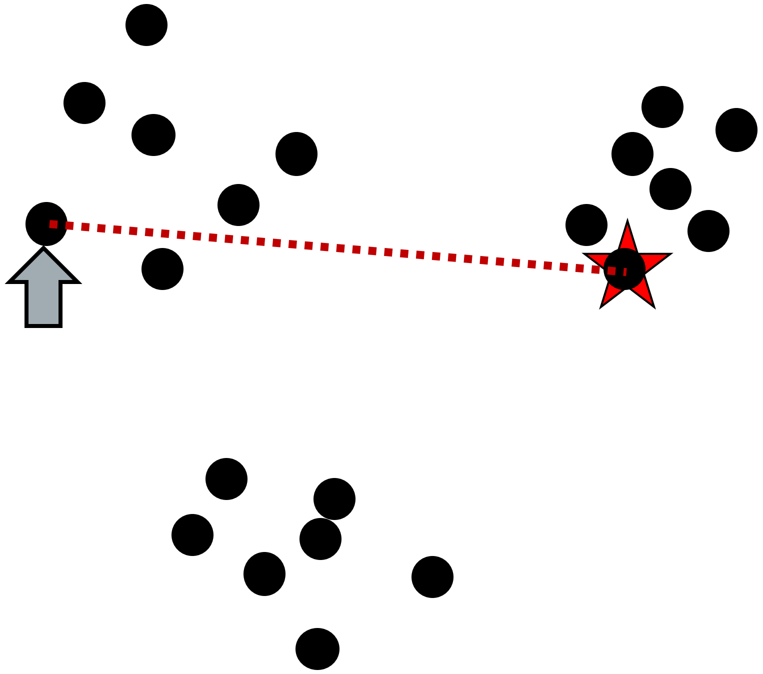
2. K-MEANS++ ALGORITHM

- **K-Means++:** only changes the initialization step
- **"Spread out centers":**
  - First center is a uniformly random point



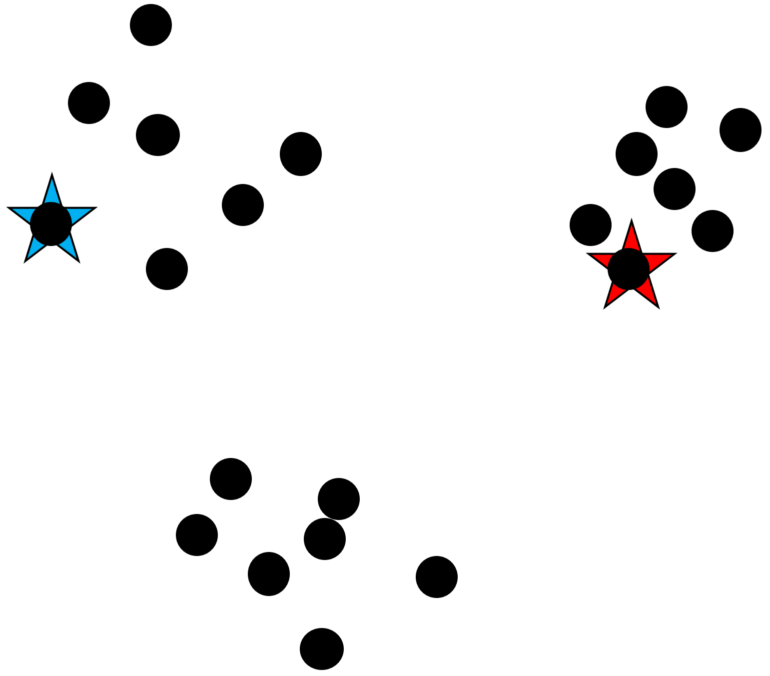
2. K-MEANS++ ALGORITHM

- **K-Means++:** only changes the initialization step
- **"Spread out centers":**
  - First center is a uniformly random point
  - Next centers: each point chosen with probability proportional to square of distance to its closest center



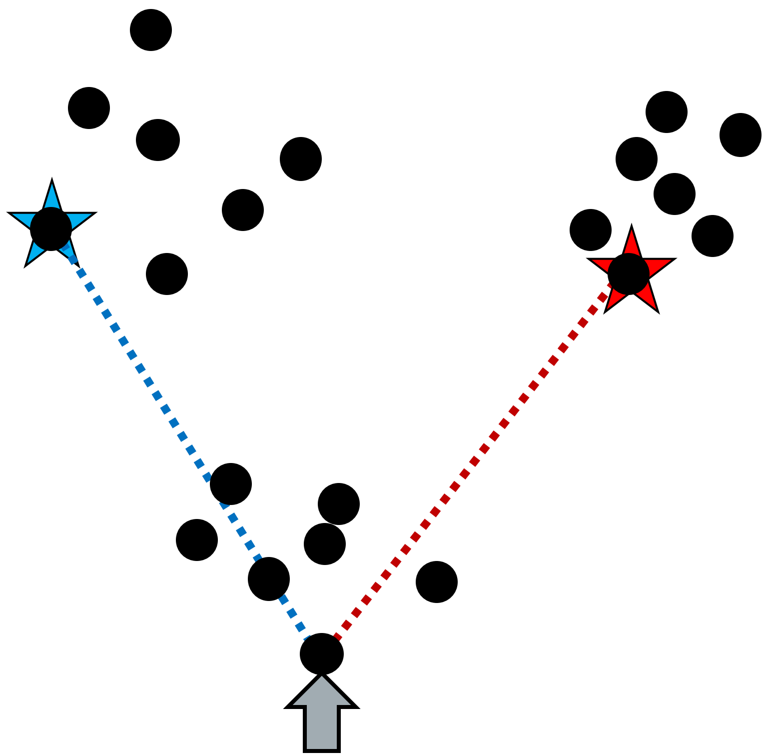
2. K-MEANS++ ALGORITHM

- **K-Means++:** only changes the initialization step
- **“Spread out centers”:**
  - First center is a uniformly random point
  - Next centers: each point chosen with probability proportional to square of distance to its closest center



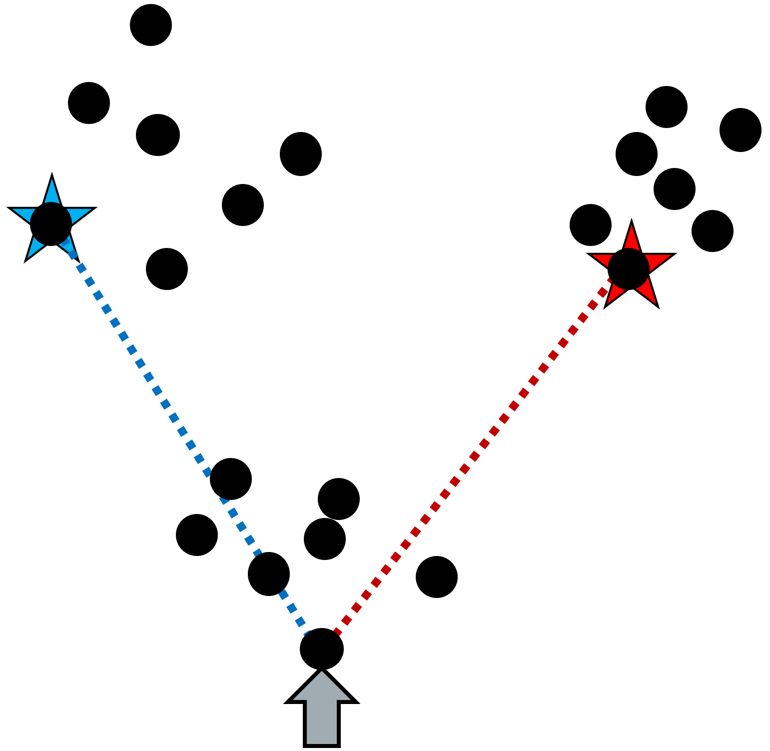
2. K-MEANS++ ALGORITHM

- **K-Means++:** only changes the initialization step
- **"Spread out centers":**
  - First center is a uniformly random point
  - Next centers: each point chosen with probability proportional to square of distance to its closest center



2. K-MEANS++ ALGORITHM

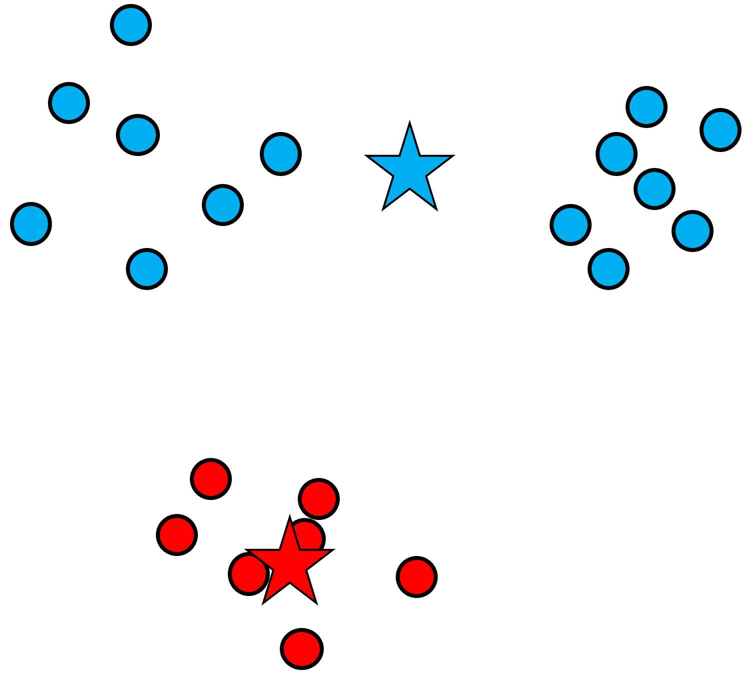
- **K-Means++:** only changes the initialization step
- Better practical performance
- Theoretical guarantee: O(log k) approximation ratio in expectation



3. X-MEANS ALGORITHM

- **Automatic way to choose K**

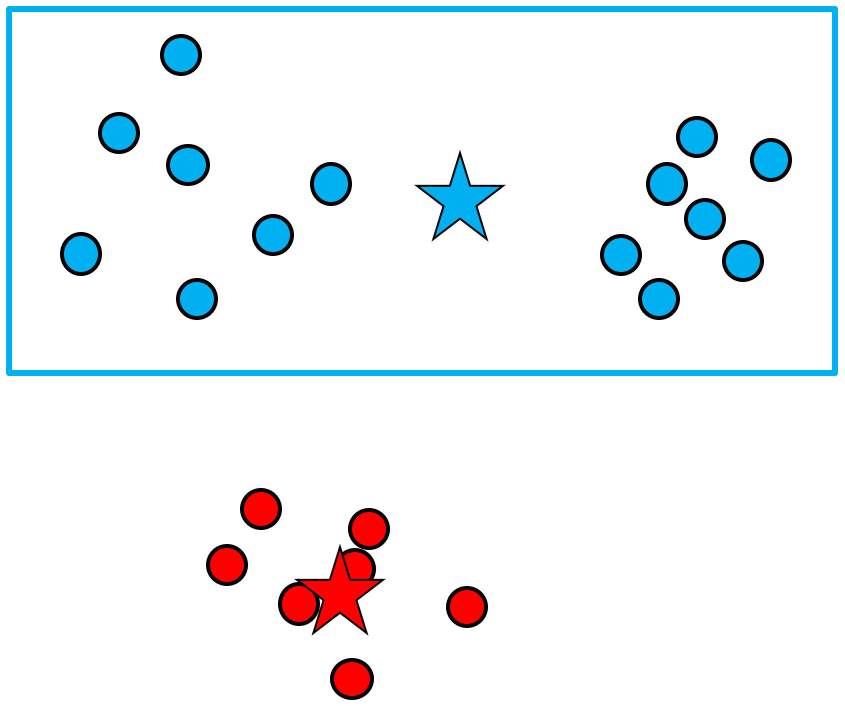
1. Run usual K-Means with K=2



3. X-MEANS ALGORITHM

- **Automatic way to choose K**

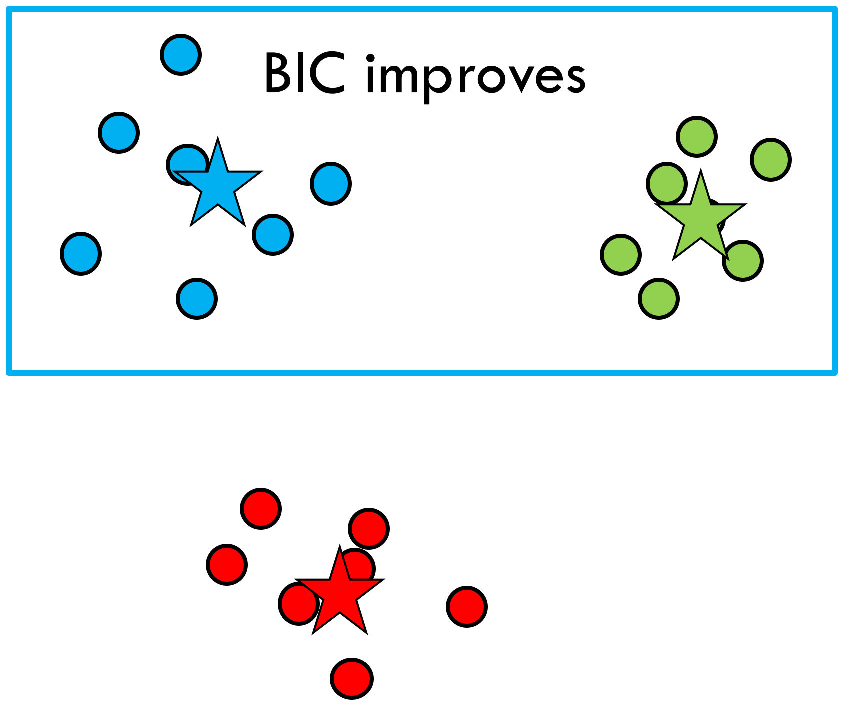
1. Run usual K-Means with K=2
2. Attempt to split each cluster by running K-Means with K=2 only within that cluster



3. X-MEANS ALGORITHM

- **Automatic way to choose K**

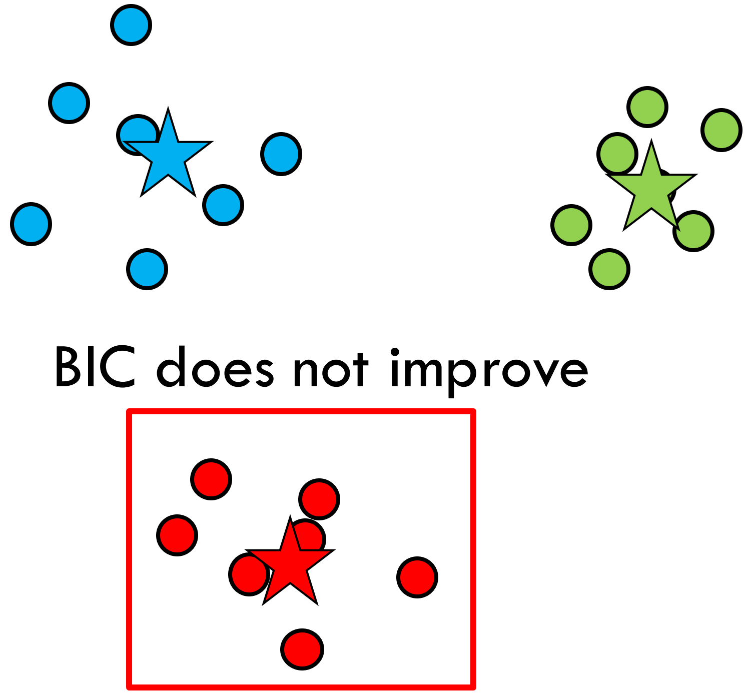
1. Run usual K-Means with K=2
2. Attempt to split each cluster by running K-Means with K=2 only within that cluster
   - Use “Bayesian Information Criterion” (BIC) to decide whether to split



3. X-MEANS ALGORITHM

- **Automatic way to choose K**

1. Run usual K-Means with K=2
2. Attempt to split each cluster by running K-Means with K=2 only within that cluster
   - Use "Bayesian Information Criterion" (BIC) to decide whether to split



New activity

Search...

Default

4 activities

Logout