GENERAL ASSEMBLY

CLUSTERING

Mason Gallo, Data Scientist

AGENDA

- Unsupervised Learning
- Intro to Clustering
- Uses for Clustering
- How Clustering works
- Implement Clustering

OBJECTIVES

- Understand how you can use Clustering to solve problems
- Intuition for how to evaluate Clustering problems
- Python implementation

MOTIVATING EXAMPLE: CLUSTERING PEOPLE

CENSUS DATA

age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

YOU WILL CREATE SEGMENTS TO BETTER UNDERSTAND THE POPULATION

UNSUPERVISED LEARNING AND CLUSTERING

CLUSTER ANALYSIS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

CLUSTER ANALYSIS

supervised	*making predictions*
unsupervised	*discovering patterns*

Q: What is a cluster?

CLUSTER ANALYSIS

Q: What is a cluster?

A: A group of **similar** *data points.*

The concept of similarity is central to the definition of a cluster, and therefore to cluster analysis.

Examples: distance between points, number of common words, etc.

CLUSTER ANALYSIS

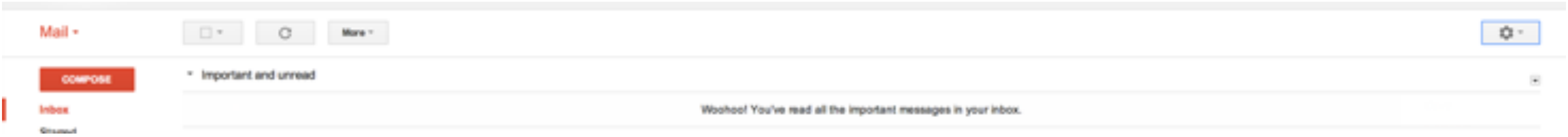
Q: What is the purpose of cluster analysis?

CLUSTER ANALYSIS

Q: What is the purpose of cluster analysis?

A: To enhance our understanding of a dataset by dividing the data into groups.

CLUSTER ANALYSIS



Priority Inbox: Unsupervised Learning

Group mails into groups and decide which group represents important mails

Q: How do you solve a clustering problem?

CLUSTER ANALYSIS

Q: How do you solve a clustering problem?

A: Think of a cluster as a "potential class"; then the solution to a clustering problem is to programmatically determine these classes.

Q: How do you solve a clustering problem?

A: Think of a cluster as a "potential class"; then the solution to a clustering problem is to programmatically determine these classes.

THINK: WE'RE ALLOCATING THE ROWS INTO CLASSES / GROUPS

DRAWING ON THE BOARD

K-MEANS CLUSTERING

K-MEANS CLUSTERING

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Q: What is k-means clustering?

K-MEANS CLUSTERING

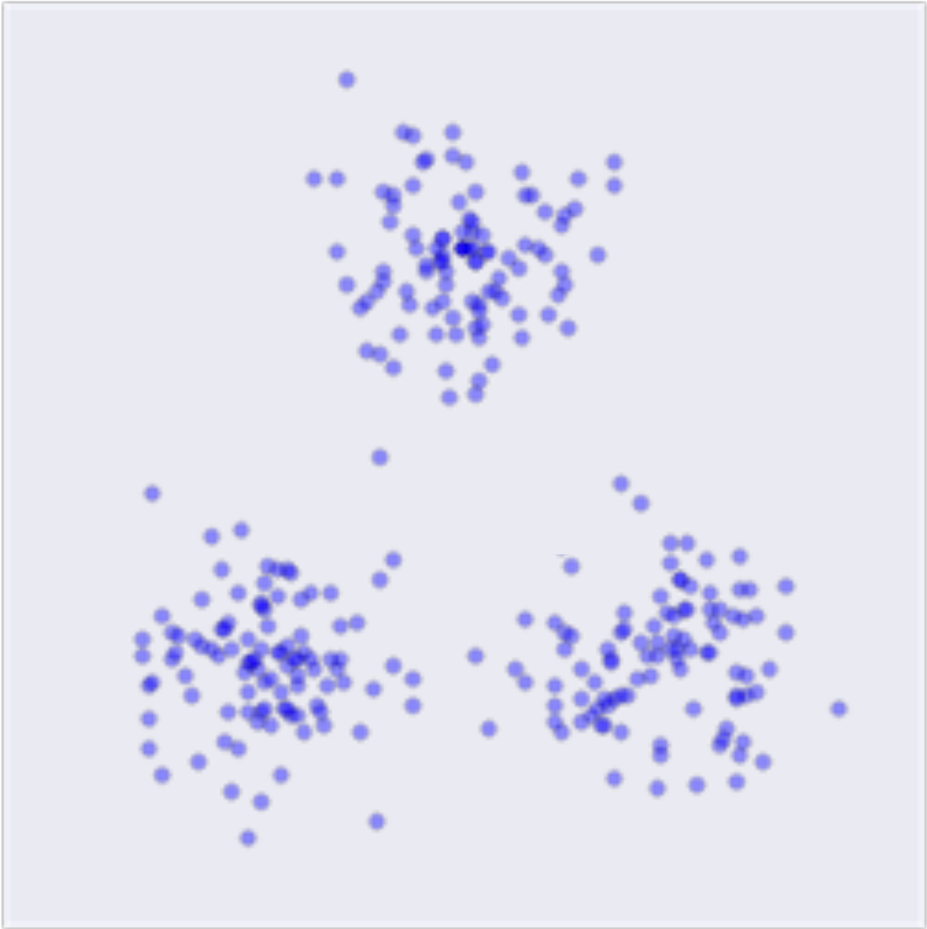
Q: What is k-means clustering?

A: A **greedy** *learner that* **partitions** *a data set into k clusters.*

greedy – *captures local structure (depends on initial conditions)*

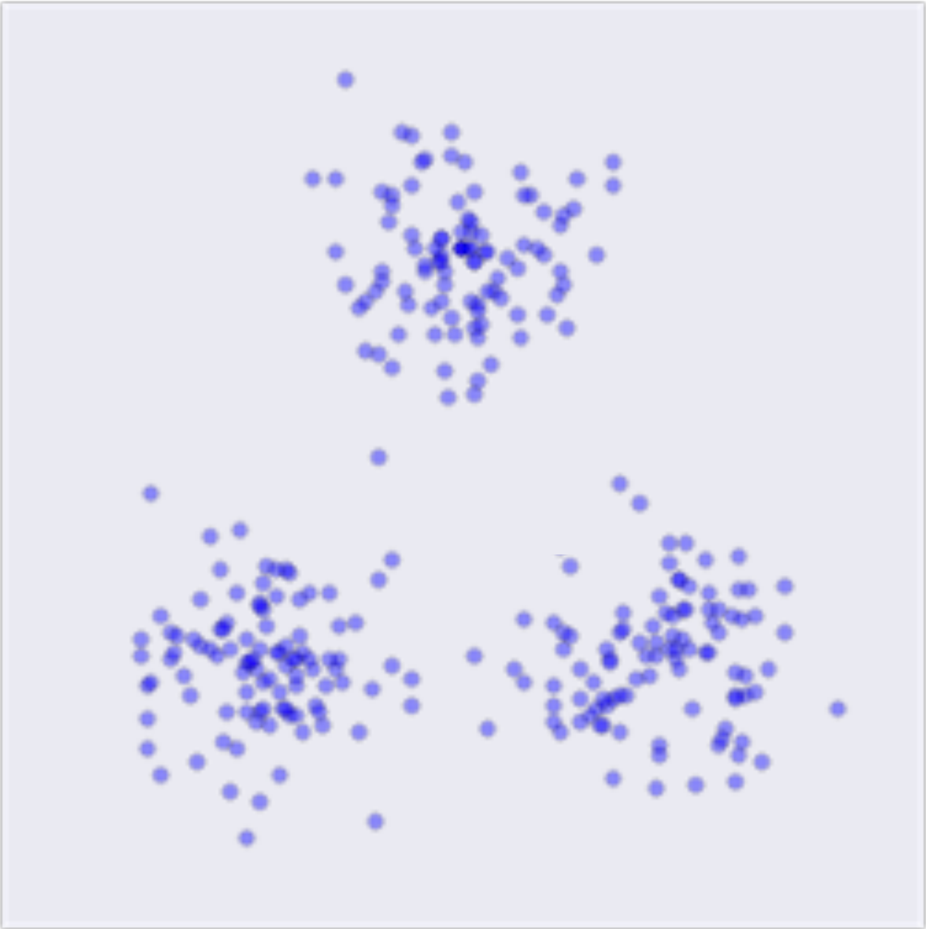
partition – *each point belongs to exactly one cluster*

K-MEANS CLUSTERING



Suppose we are given some unsupervised data (i.e., no class labels)

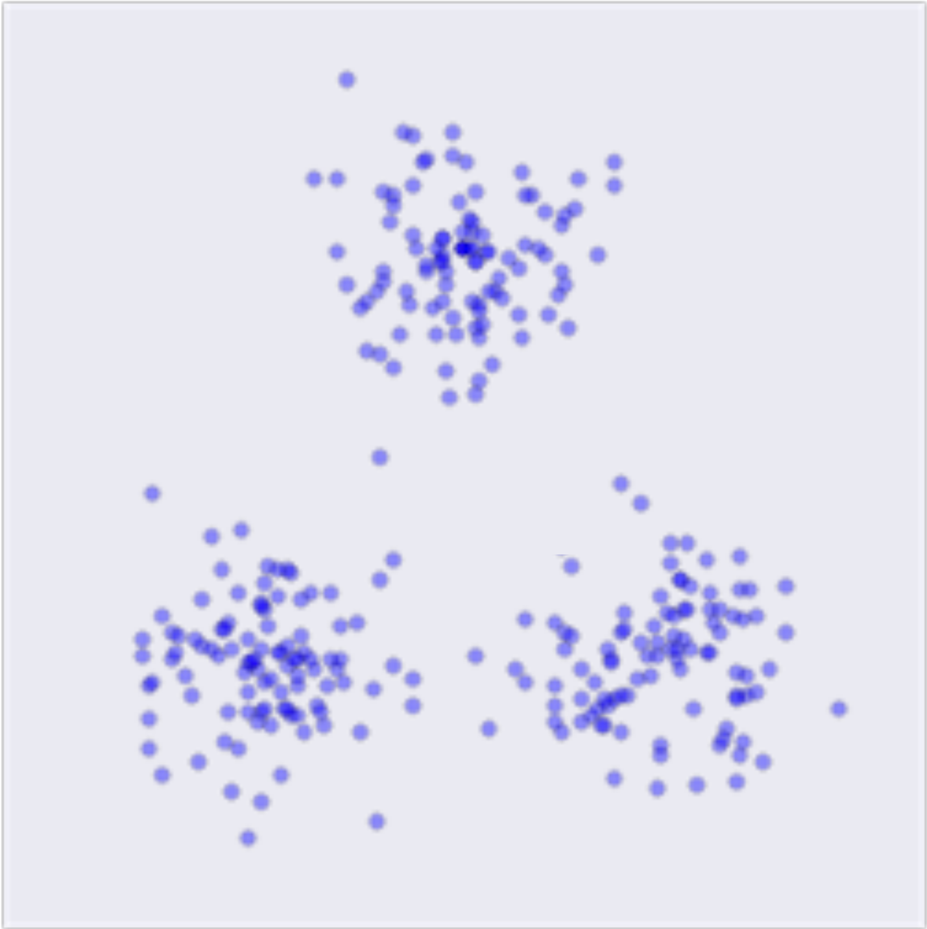
K-MEANS CLUSTERING



Suppose we are given some unsupervised data (i.e., no class labels)

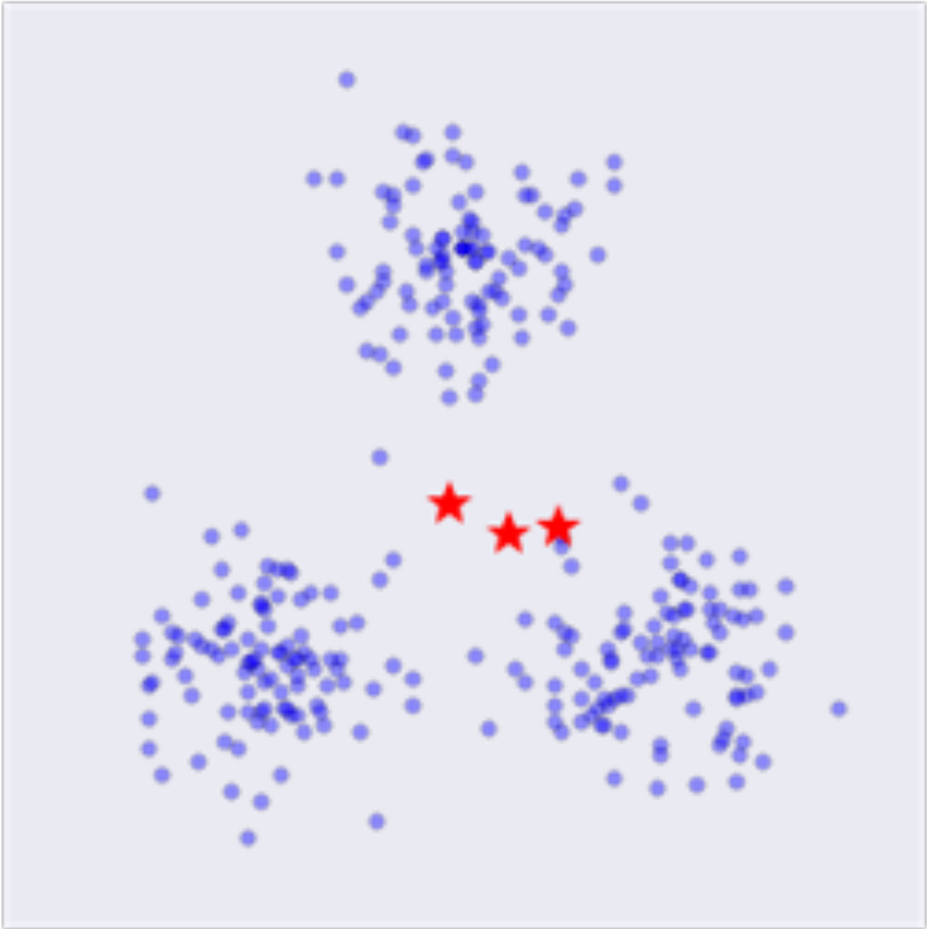
We could like to infer class labels from the data, i.e., cluster the data into similar groups

K-MEANS CLUSTERING



Steps of k-means algorithm

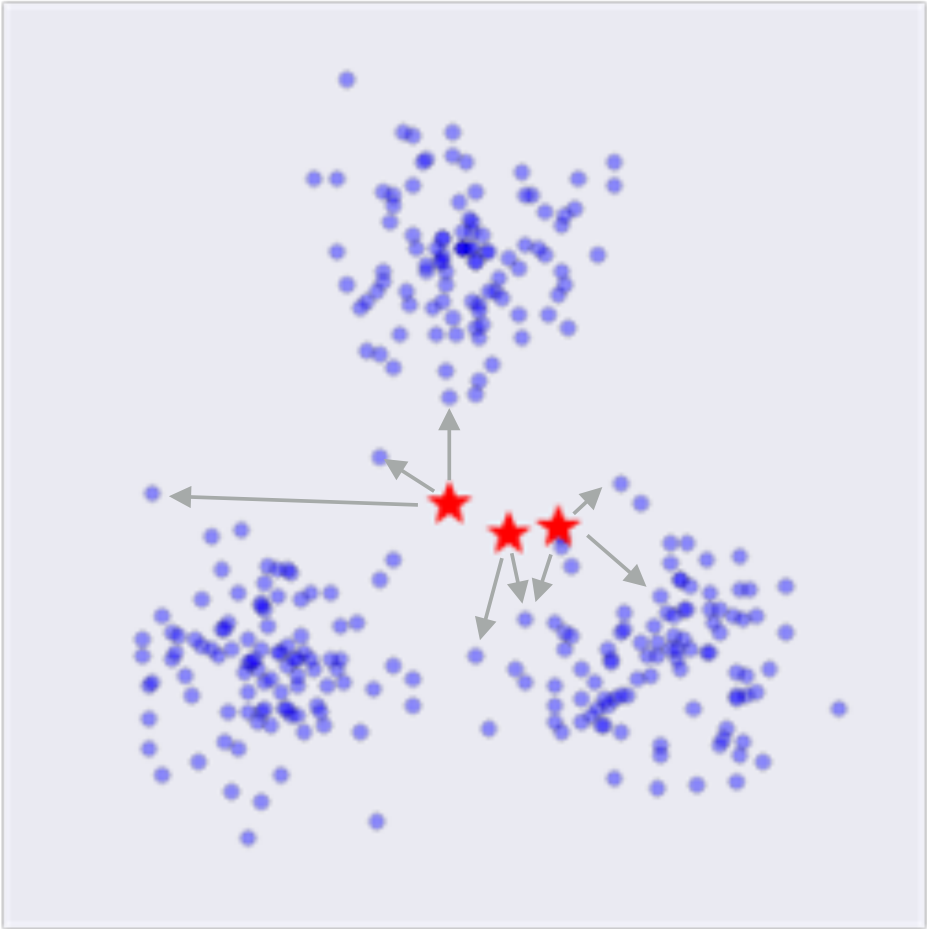
K-MEANS CLUSTERING



Steps of k-means algorithm

Start with k cluster centers chosen at random

K-MEANS CLUSTERING

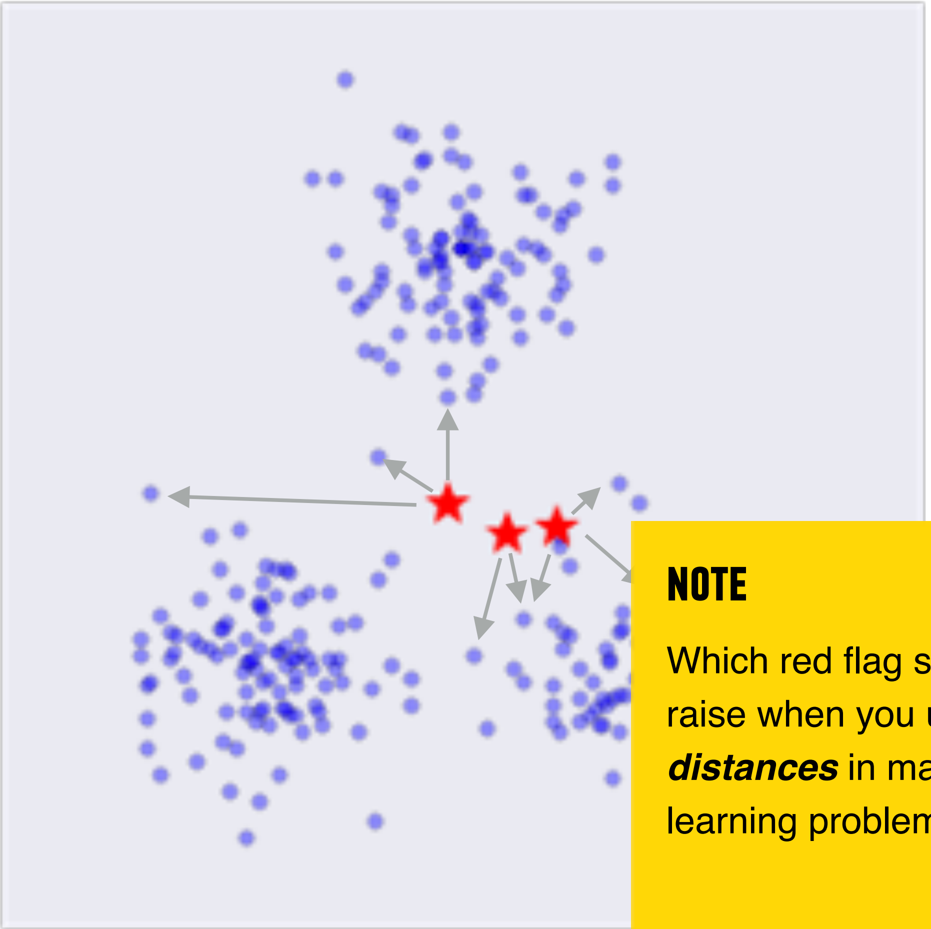


Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*

K-MEANS CLUSTERING



Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*

NOTE

Which red flag should raise when you use ***distances*** in machine learning problems?

K-MEANS CLUSTERING

NOTE

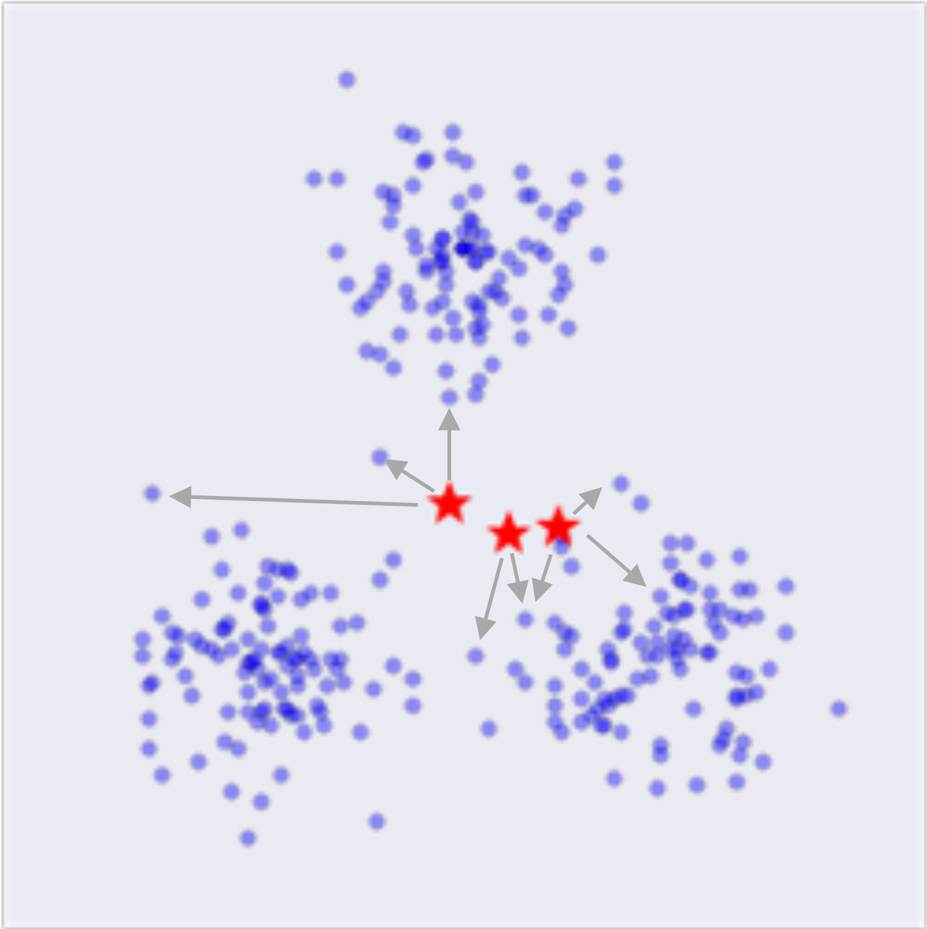
The k-means algorithm require ***feature scaling***, as we saw earlier with kNN

Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*

K-MEANS CLUSTERING

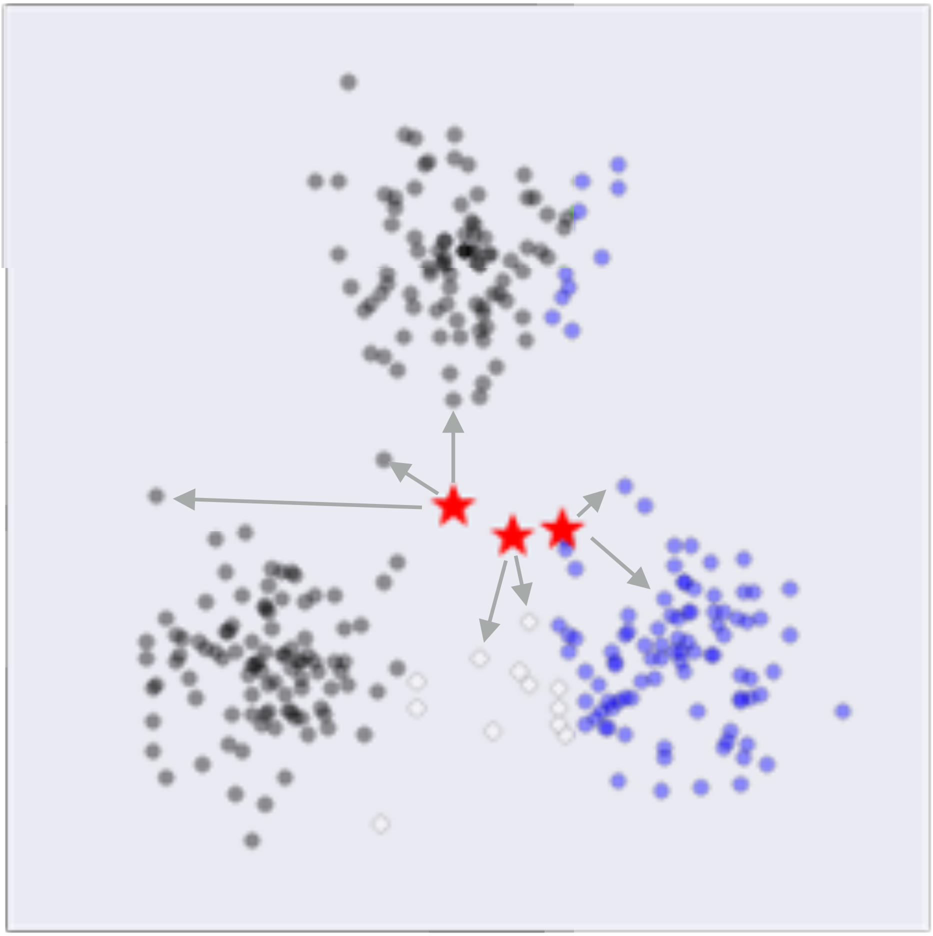


Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*

K-MEANS CLUSTERING

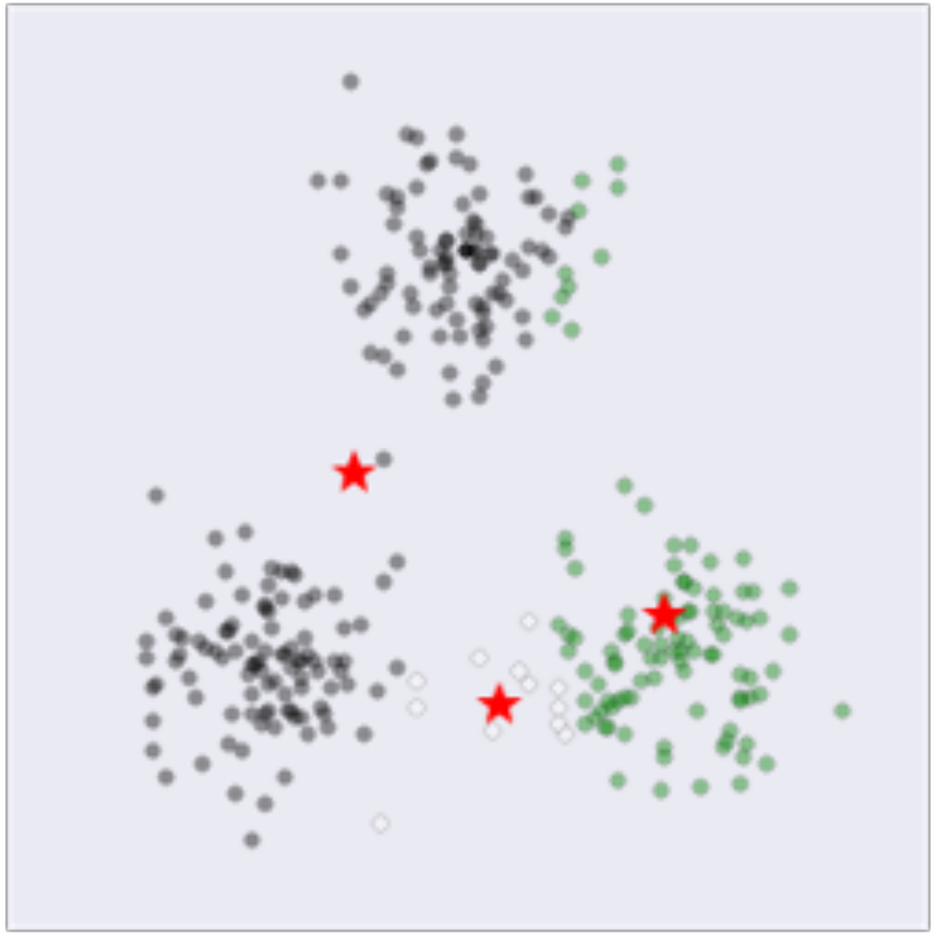


Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*
2. *Label data according to their closest cluster*

K-MEANS CLUSTERING

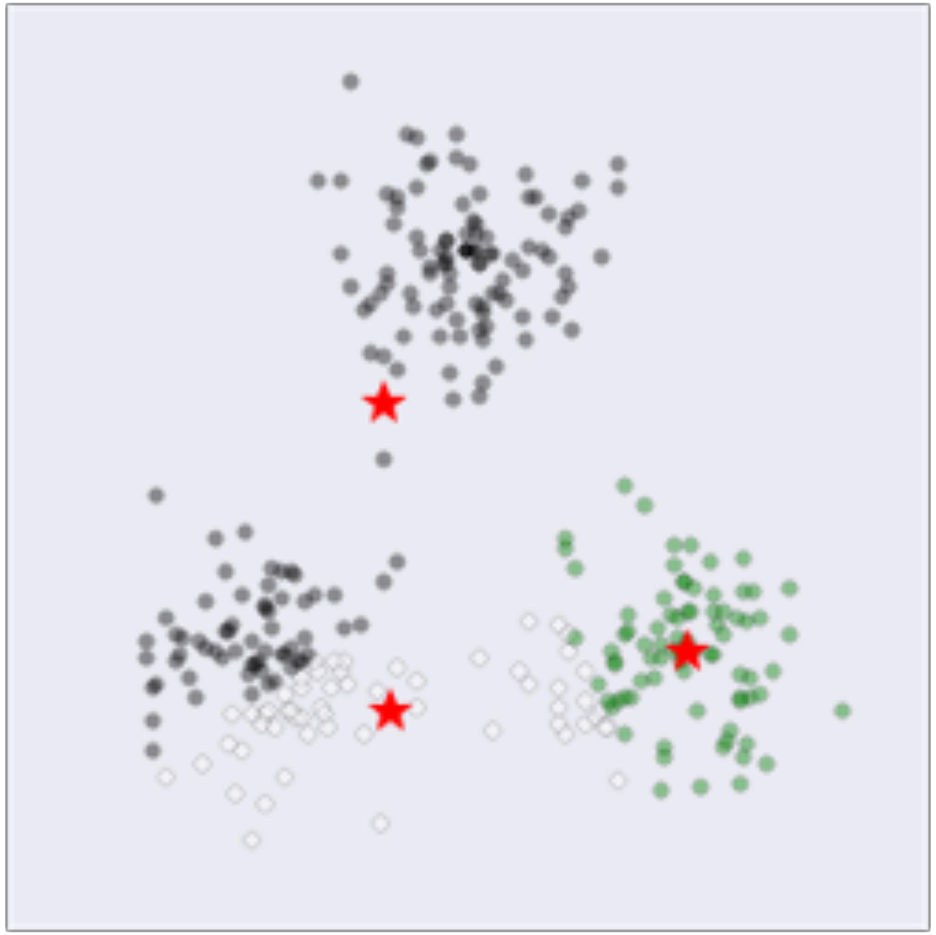


Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*
2. *Label data according to their closest cluster*
3. *Recompute cluster centers*

K-MEANS CLUSTERING



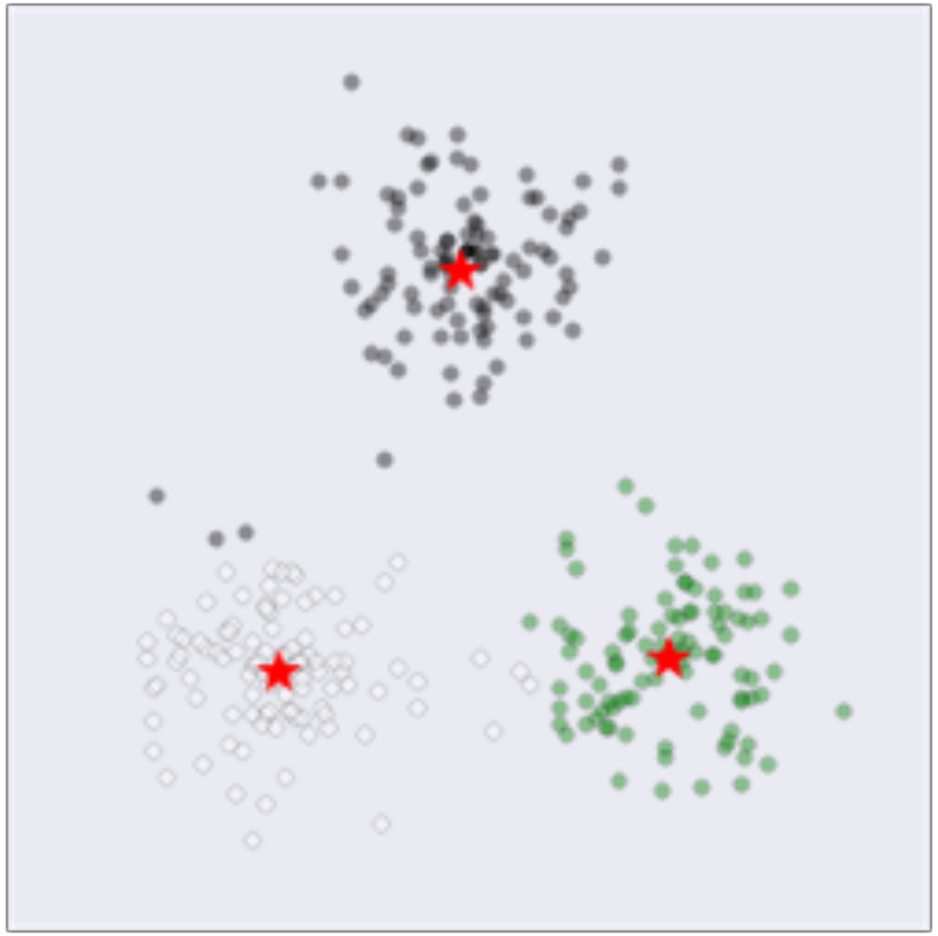
Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*
2. *Label data according to their closest cluster*
3. *Recompute cluster centers*

Repeat 1-3 until labels don't change
(or some maximum iteration has been reached)

K-MEANS CLUSTERING



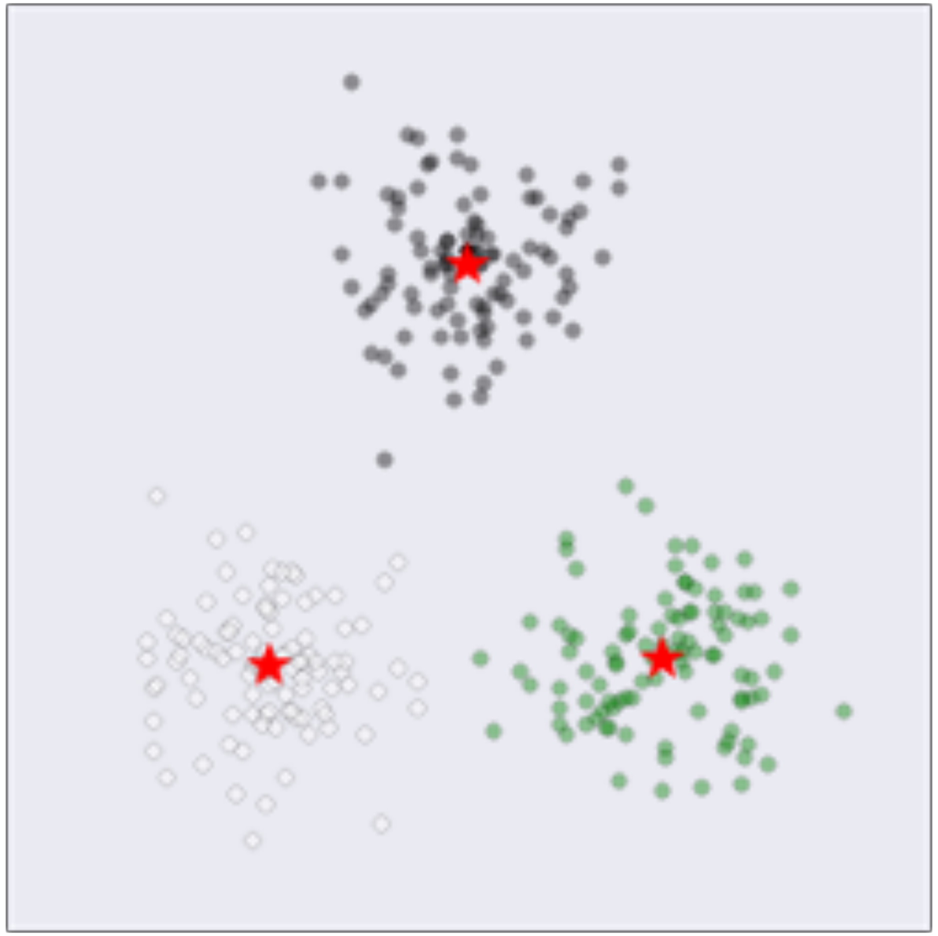
Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*
2. *Label data according to their closest cluster*
3. *Recompute cluster centers*

Repeat 1-3 until labels don't change
(or some maximum iteration has been reached)

K-MEANS CLUSTERING



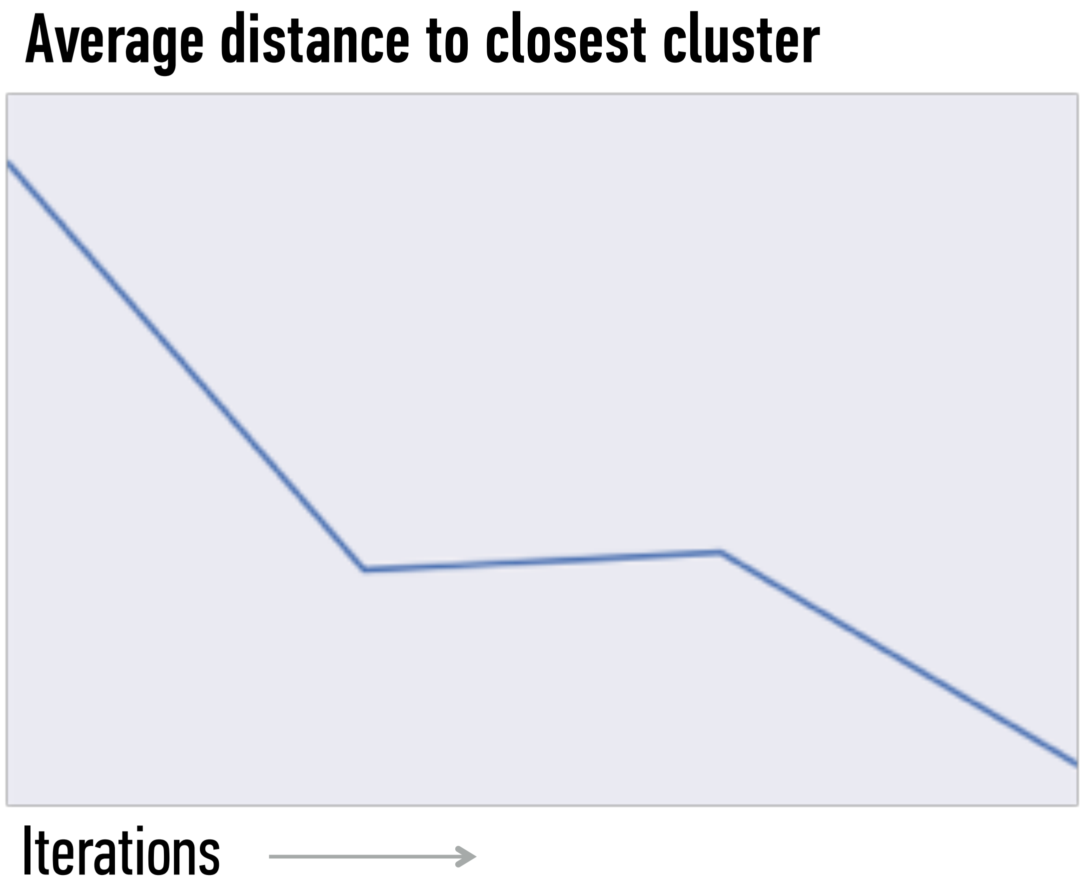
Steps of k-means algorithm

Start with k cluster centers chosen at random

1. *Compute distances from each point to centers*
2. *Label data according to their closest cluster*
3. *Recompute cluster centers*

Repeat 1-3 until labels don't change
(or some maximum iteration has been reached)

COST FUNCTION



At each step, we compute the **average distance** *to the closest cluster center as its 'cost'*

GETTING THE “BEST” CLUSTERS

ASSESSING ML PERFORMANCE

supervised	*test out your predictions*
unsupervised	*...*

ASSESSING ML PERFORMANCE

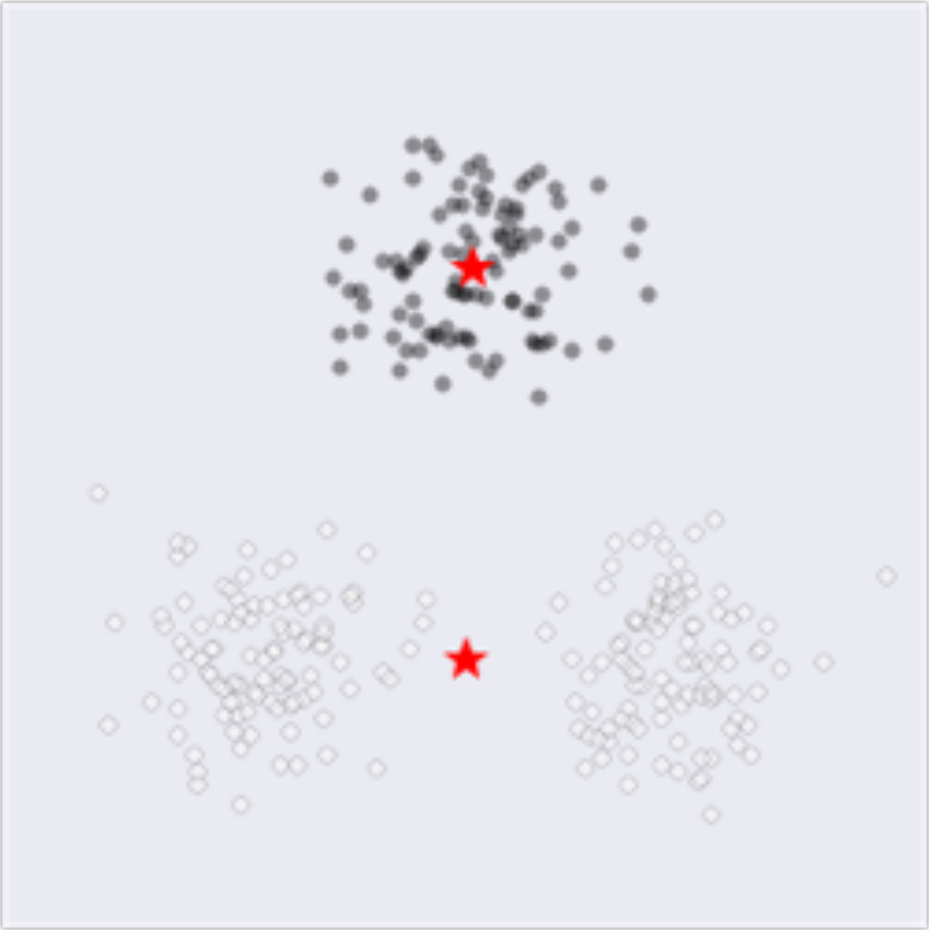
supervised	*test out your predictions*
unsupervised	*can't really*

CLUSTER VALIDATION

How do we choose k?

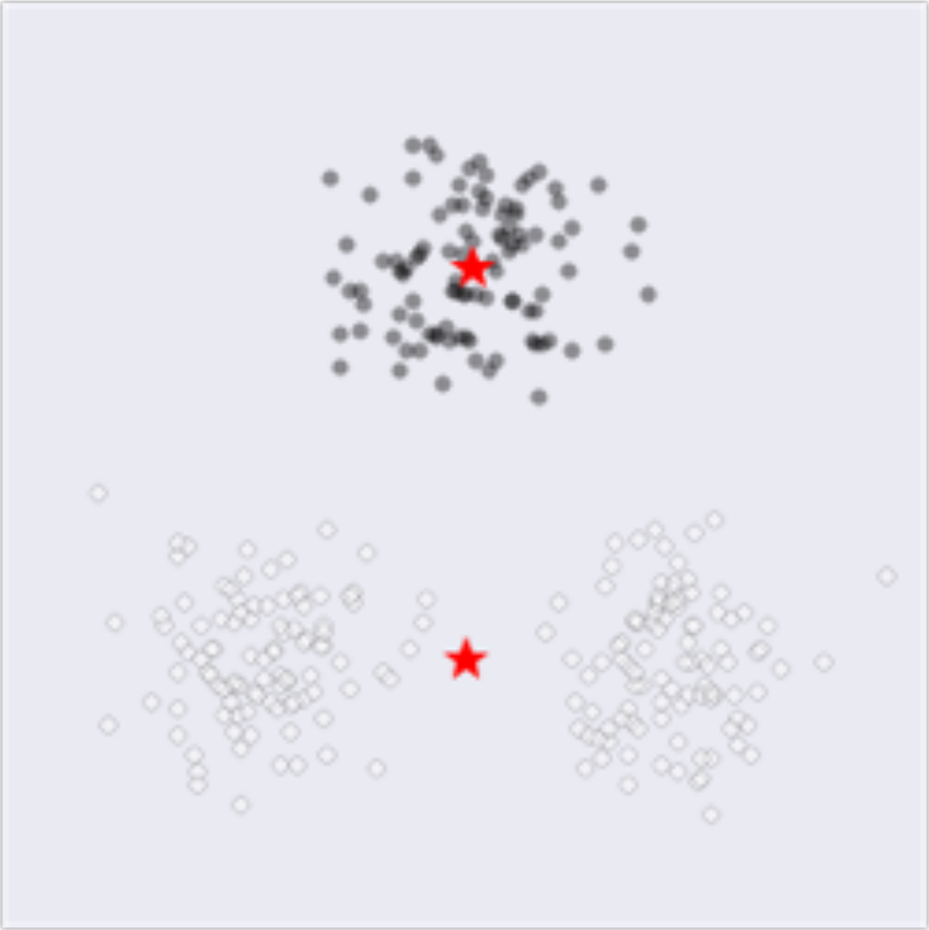
CLUSTER VALIDATION

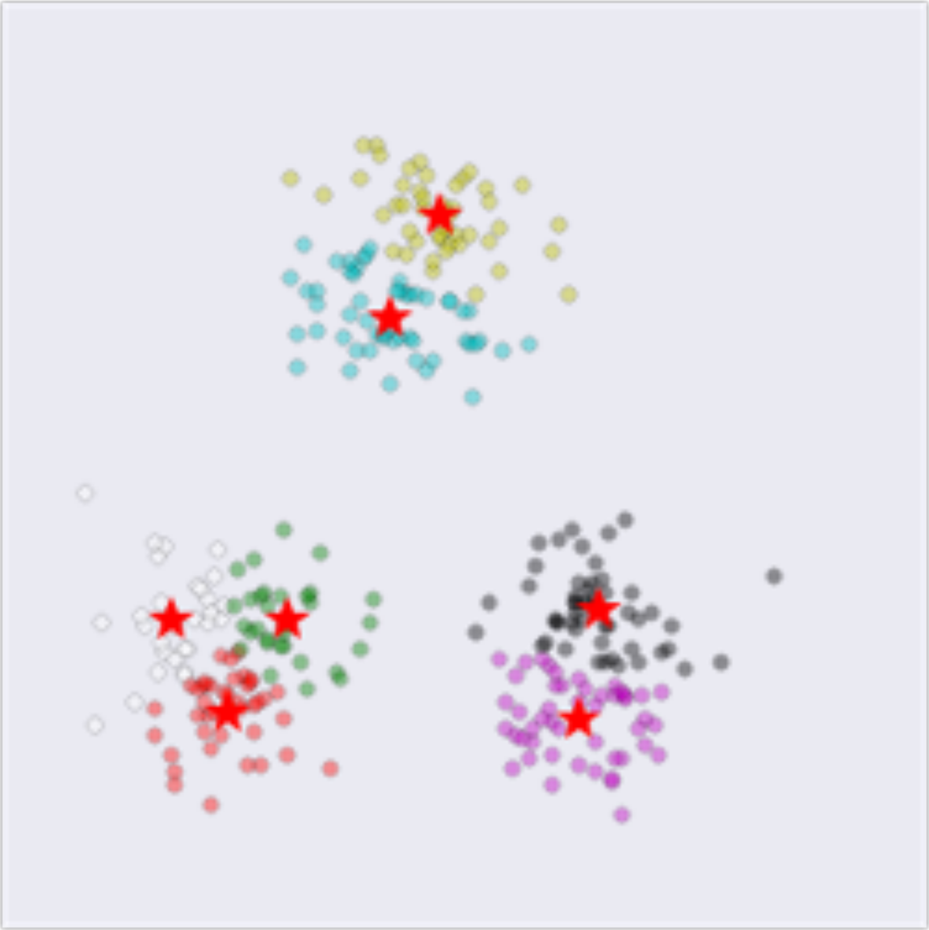
How do we choose k?



CLUSTER VALIDATION

How do we choose k?





CLUSTER VALIDATION

In general, k-means will converge to a solution and return a partition of k clusters, even if no natural clusters exist in the data.

We will look at two validation metrics useful for partitional clustering, **cohesion** *and* **separation***.*

CLUSTER VALIDATION

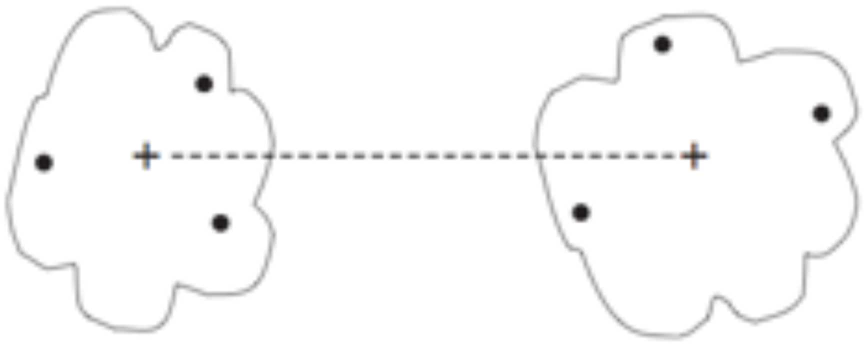
Cohesion *measures clustering effectiveness within a cluster.*

CLUSTER VALIDATION

Cohesion *measures clustering effectiveness within a cluster.*

$$\hat{C}(C_i) = \sum_{x \in C_i} d(x, c_i)$$

Separation *measures clustering effectiveness between clusters.*



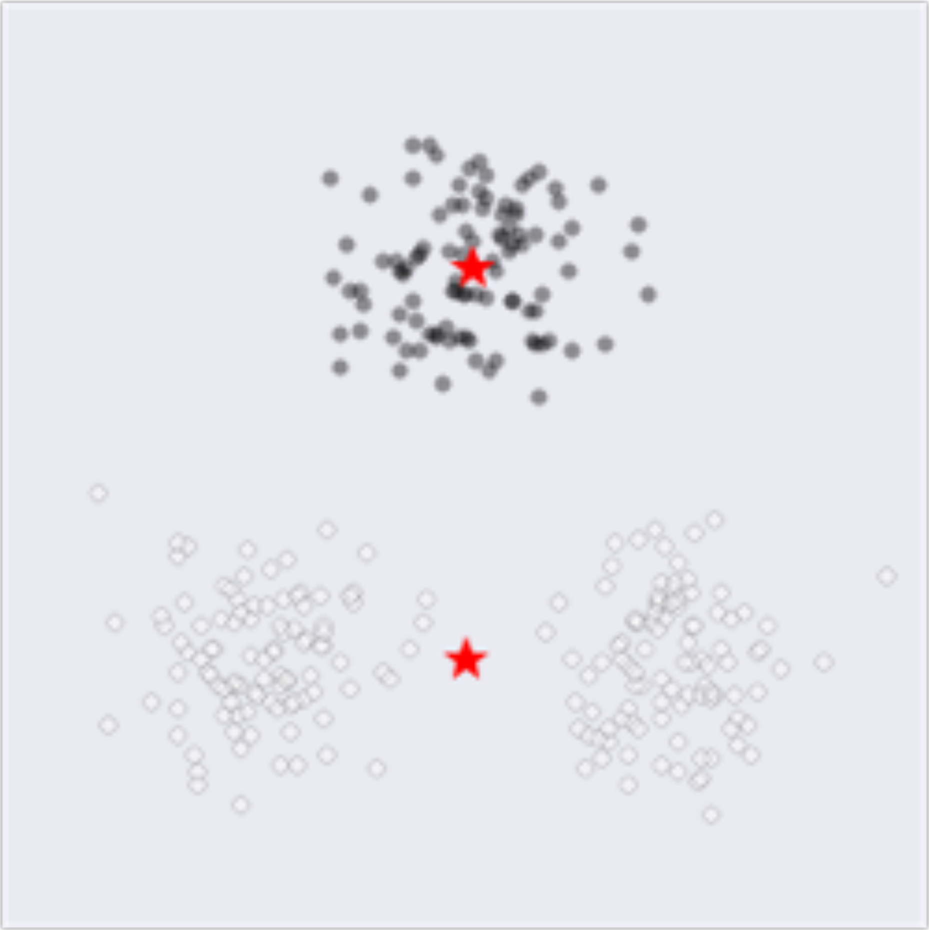
$$\hat{S}(C_i, C_j) = d(c_i, c_j)$$

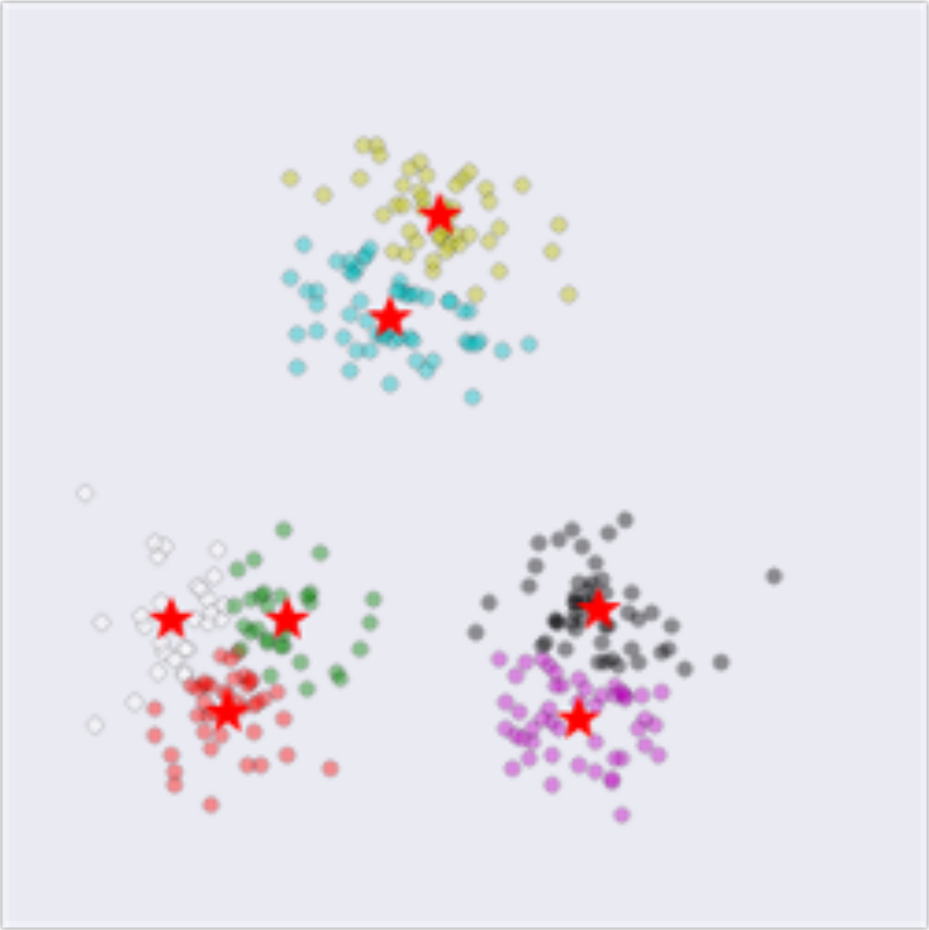
CLUSTER VALIDATION

How do we choose k?

CLUSTER VALIDATION

How do we choose k?

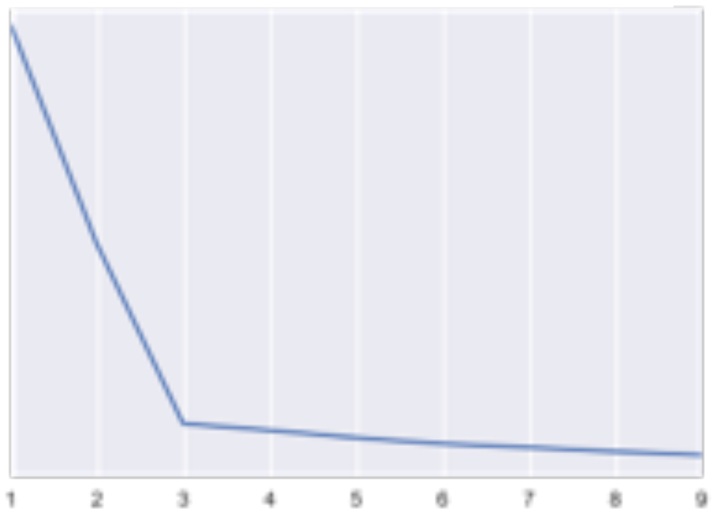




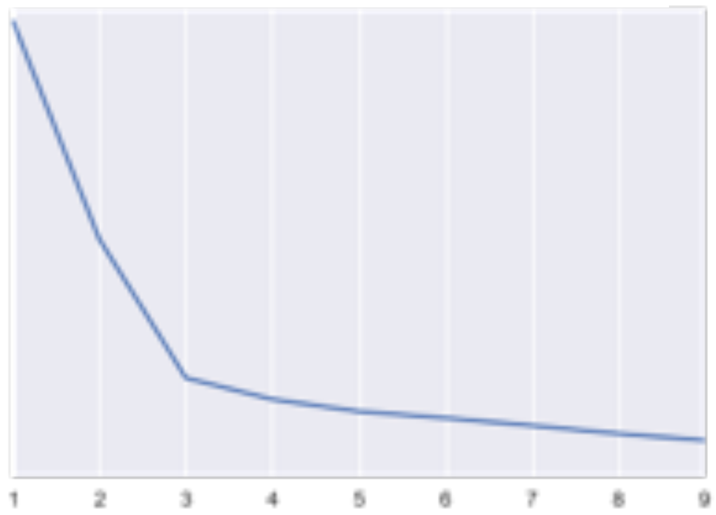
CLUSTER VALIDATION

How do we choose k?

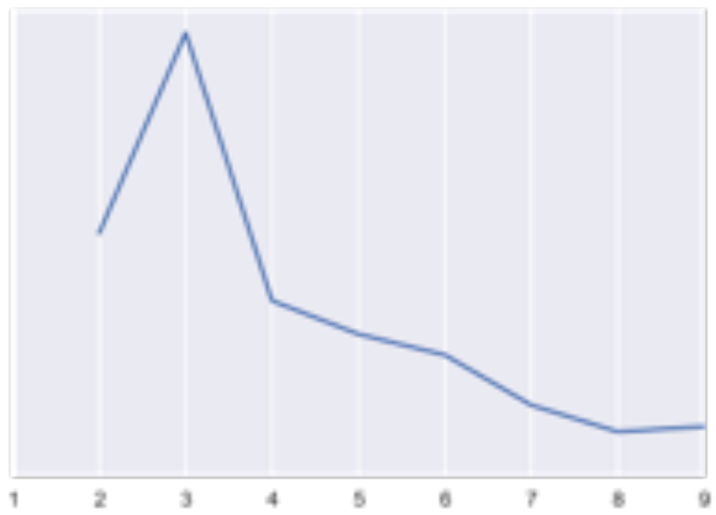
Average distance to closest cluster



Average cohesion within clusters



Average separation between clusters



CLUSTER VALIDATION

How do we choose k?

Average distance to closest cluster

Average cohesion within clusters

Average separation between clusters

Look for the **largest kink** *in the cost curve (this is called the* **elbow method***)*

Or look for the **largest separation** *between clusters*

CLUSTER VALIDATION

In practice, you'd choose k with a certain application in mind

CLUSTER VALIDATION

In practice, you'd choose k with a certain application in mind

For example, you'd like to manufacture three sizes of clothing: small, medium or large

REAL WORLD EXAMPLE FROM MY WORK

BUSINESS PROBLEM

Major retailer rebranding to fight decreasing sales and increased competition

BUSINESS PROBLEM

Major retailer rebranding to fight decreasing sales and increased competition

To rebrand, they want to understand their customers and how to reach them

THE DATA

Retailer tracks all credit card sales nationwide

Purchase behavior and demographic information

HOW WE DID IT

K-Means Clustering to find meaningful customer segments

WHAT THEY DO WITH THE SEGMENTS

Figure out the customer segments that are most profitable

Spend money marketing to the most profitable customer segments

WHAT THEY DO WITH THE SEGMENTS

Figure out the customer segments that are most profitable

Spend money marketing to the most profitable customer segments

STOP WASTING MONEY ON CUSTOMERS THAT HAVE LOW LIFETIME VALUES

K-MEANS CLUSTERING FAQ

VISUALIZING K-MEANS CLUSTERING

http://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-MEANS CLUSTERING

Questions:

- What if no clusters exist?
- How to choose K?
- How to choose the initial centroid positions?
- When to stop the algorithm?
- When might it produce poor results?

K-MEANS CLUSTERING

What if no clusters exist?

- It will still find clusters!
  - Visualization: I’ll Choose, Uniform Points

K-MEANS CLUSTERING

How to choose K?

- It will find the number of clusters specified
    - Visualization: I'll Choose, Gaussian Mixture, K=2/3/4
- Try different values for K and pick the "best"

K-MEANS CLUSTERING

How to choose the initial centroid positions?

- Randomly
  - Doesn’t tend to work well
  - Visualization: http://asa.1gb.ru/kmeans/1.html
- Farthest point
  - Visualization: Farthest Point, Packed Circles, K=7
- K-means++
  - Similar to farthest point, but adds some randomness
  - Used by default in scikit-learn
- In all cases: Run it several times and pick the best result to avoid local minima

K-MEANS CLUSTERING

When to stop the algorithm?

- Tends to converge quickly
- Set stopping criteria:
    - Maximum number of iterations
    - Once centroids move less than a threshold amount
    - Once fewer points than a threshold amount change clusters
    - Visualization: Randomly, Pimpled Smiley, K=6

K-MEANS CLUSTERING

When might it produce poor results?

- Data with varying shapes
  - Visualization: I’ll Choose, Smiley Face, K=4
  - Visualization: I’ll Choose, Density Bars, K=1
- Data with different scales
- Still the most popular clustering algorithm, used for a wide range of applications

MASON, I REALLY LOVE CLUSTERING SO SHOW ME MORE

http://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods

LET’S CODE!