

CLUSTERING

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

What is Data Science	Lesson 1
Research Design and Pandas	Lesson 2
Statistics Fundamentals I	Lesson 3
Statistics Fundamentals II	Lesson 4
Flexible Class Session	Lesson 5

UNIT 2: FOUNDATIONS OF DATA MODELING

Introduction to Regression	Lesson 6
Evaluating Model Fit	Lesson 7
Introduction to Classification	Lesson 8
Introduction to Logistic Regression	Lesson 9
Communicating Logistic Regression Results	Lesson 10
Flexible Class Session	Lesson 11

UNIT 3: DATA SCIENCE IN THE REAL WORLD

Decision Trees and Random Forests	Lesson 12
Natural Language Processing	Lesson 13
Dimensionality Reduction	Lesson 14
Time Series Data I	Lesson 15
Time Series Data II	Lesson 16
Database Technologies	Lesson 17
Where to Go Next	Lesson 18
Flexible Class Session	Lesson 19
Final Project Presentations	Lesson 20



LAST CLASS

WHAT DID WE LEARN?

- Explain the trade-offs between the precision and recall of a model while articulating the cost of false positives vs. false negatives
- Describe the difference between visualization for presentations vs. exploratory data analysis
- Practice, practice, and practice!



You got all objectives?



Not all of them...

Let's form groups of 1's and 2's ...

LAST CLASS

ANNOUNCEMENTS

- ❖ You will need to return your parking garage passes.
- ❖ Mid-class survey
- ❖ Unit Project 3

There is a lot to learn in this class and the pace is pretty rapid. I gave this feedback to Abbas and he made some changes which really helped me and I believe others in the class. We were still able to cover what we needed and I was able to understand the lessons deeper.

I don't have a coding back ground and the prework given (codecademy) does not focus on data science python. For me a Python Data Science class would have been a must to get the most out of course

I think the time is really strict for this class to the instructor has limited time on each subjects and questions he can answer?

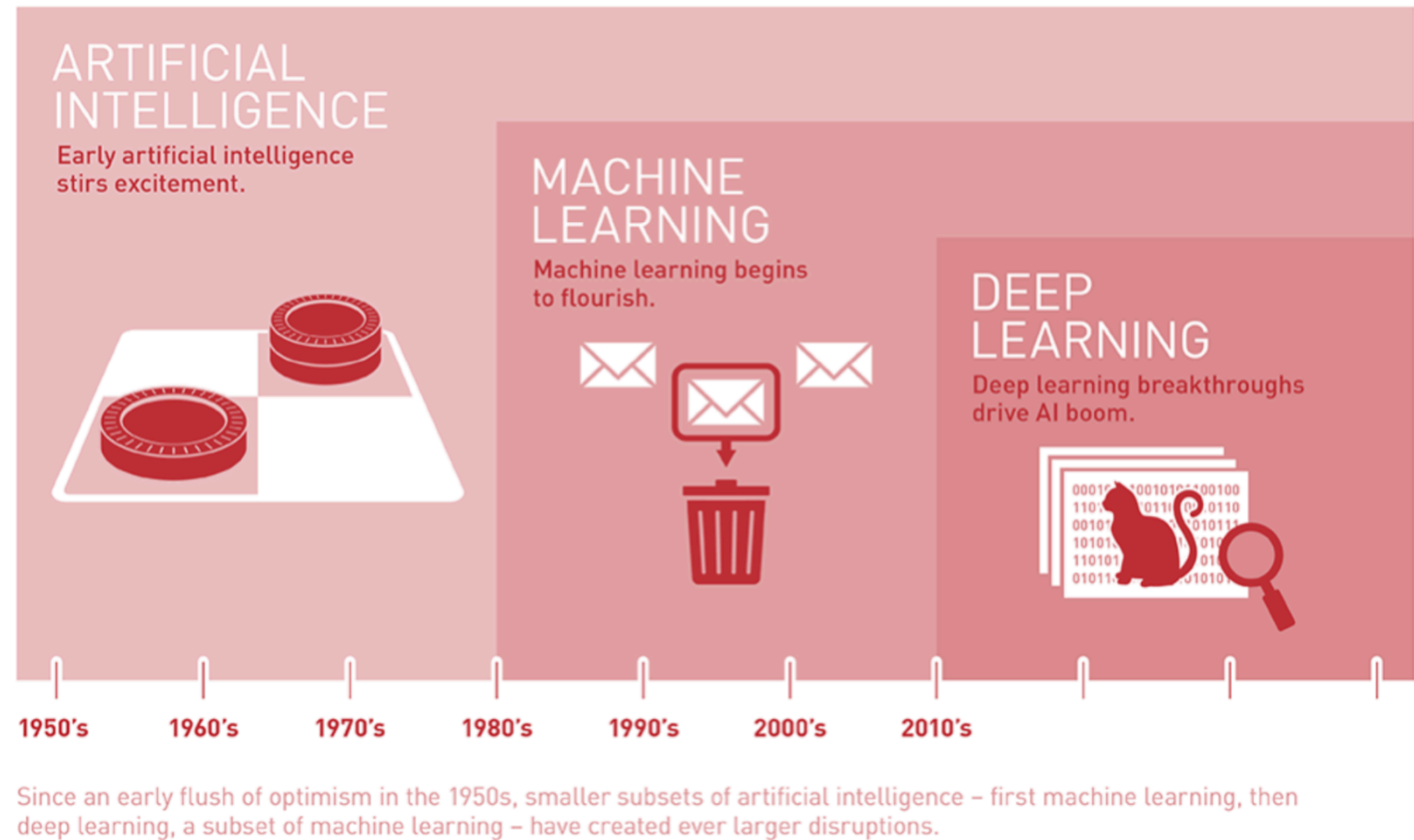
it is covering all the crucial topics and if you put the time in, you can come away almost prepared to do professional data science - or at least find an entry level job in the field.

Its hard to teach a class with such a wide variety of experience. I would recommend instituting a prerequisite test or class (python 101) either before or concurrently with this one. Teaching the math behind data science and teaching the code to implement it are two separate subjects.

WHAT IS MACHINE LEARNING

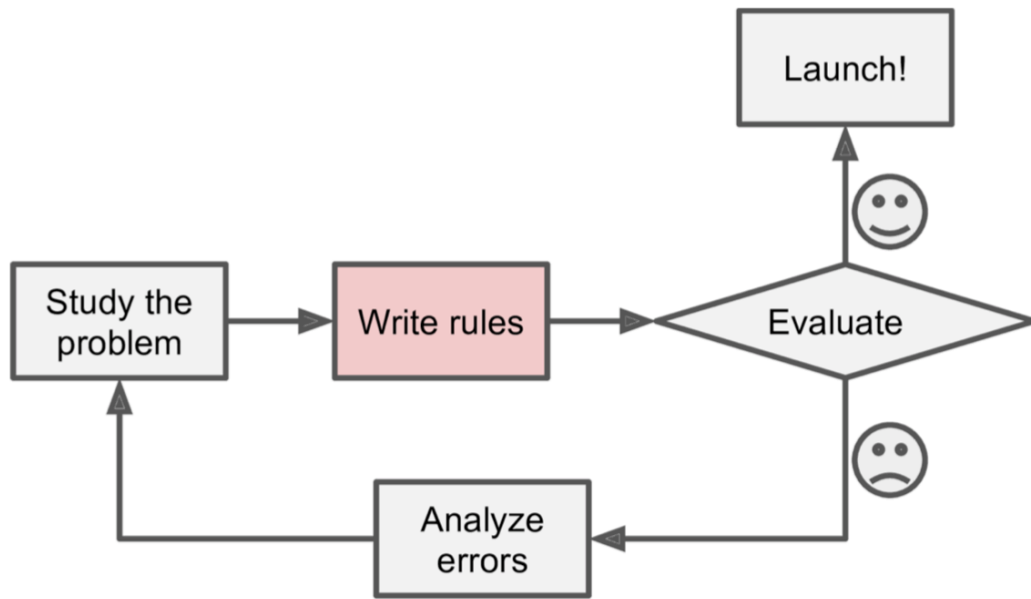
[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959



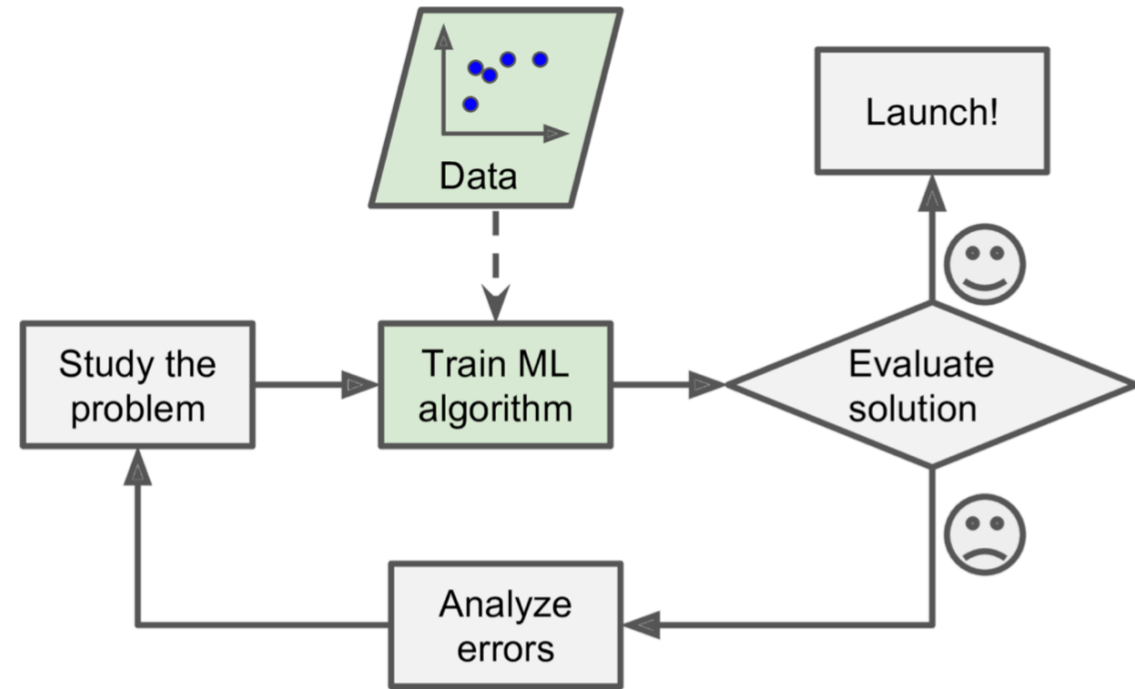
Machine Learning: is the science (and art) of programming computers so they can *learn from data*.

WHY TO USE MACHINE LEARNING?



Traditional approach

Vs.



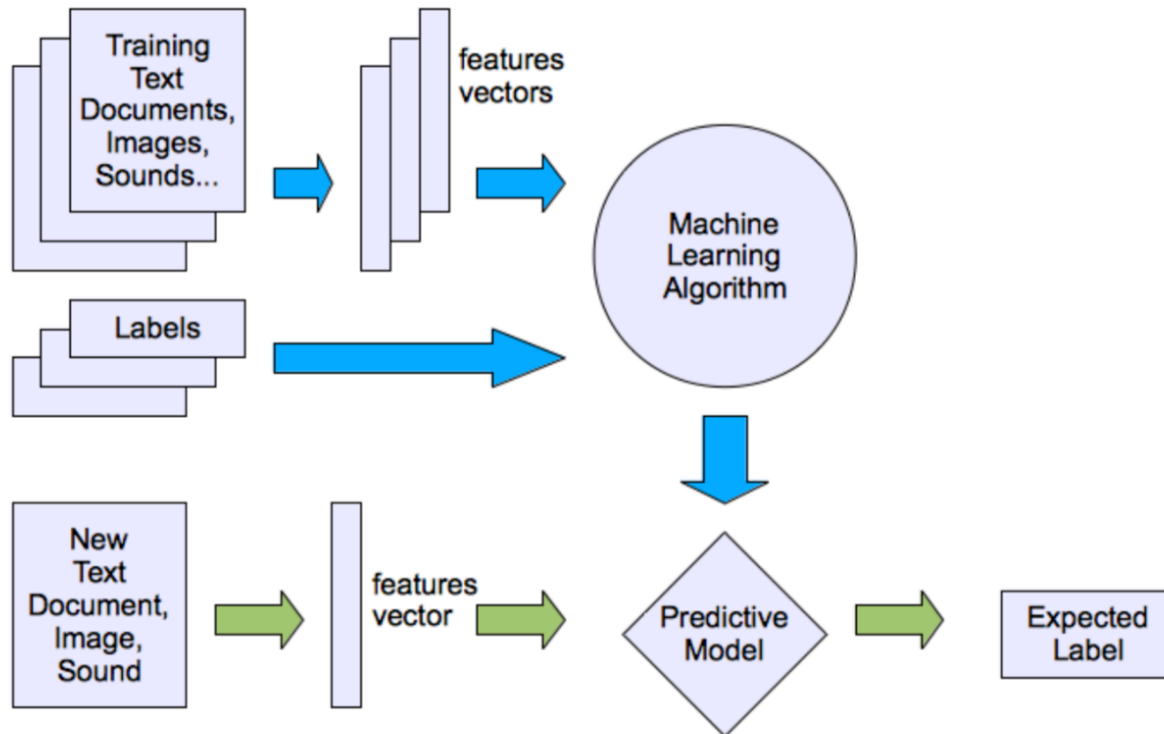
Machine Learning approach

Did you know?

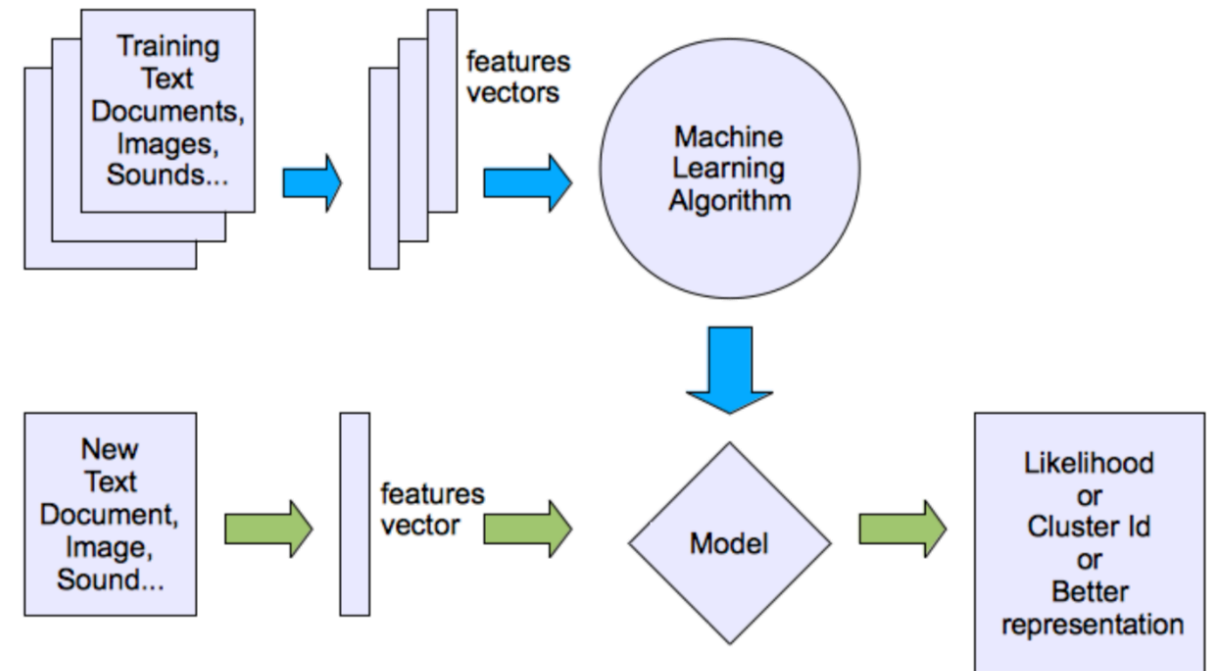
Machine learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years.

TYPES OF MACHINE LEARNING

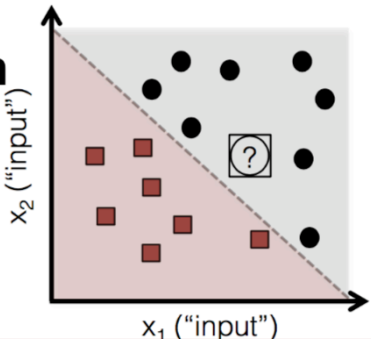
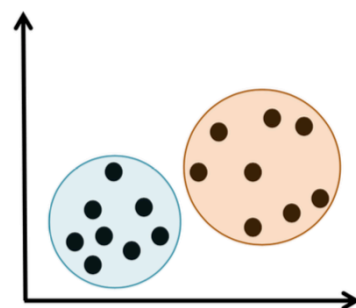
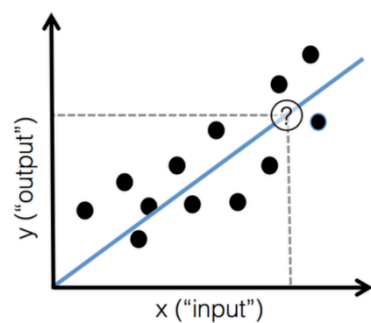
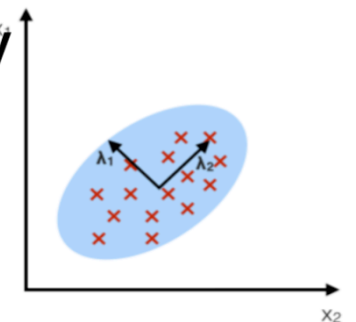
Supervised



Unsupervised



TYPES OF MACHINE LEARNING

	Supervised Working with Labeled Data	Unsupervised Working with Unlabeled Data
Discrete Countable Data	Classification 	Clustering 
Continuous Infinite Data	Regression 	Dimensionality Reduction 

COMMUNICATING RESULTS

LEARNING OBJECTIVES

- Supervised vs unsupervised algorithms
- Understand and apply k-means clustering
- Density-based clustering: DBSCAN
- Silhouette Metric

OPENING

UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

- So far all the algorithms we have used are *supervised*: each observation (row of data) came with one or more *labels*, either *categorical variables* (classes) or *measurements* (regression)
- **Unsupervised learning** has a different goal: **feature discovery**
- **Clustering** is a common and fundamental example of unsupervised learning
- **Clustering** algorithms try to find meaningful groups within data

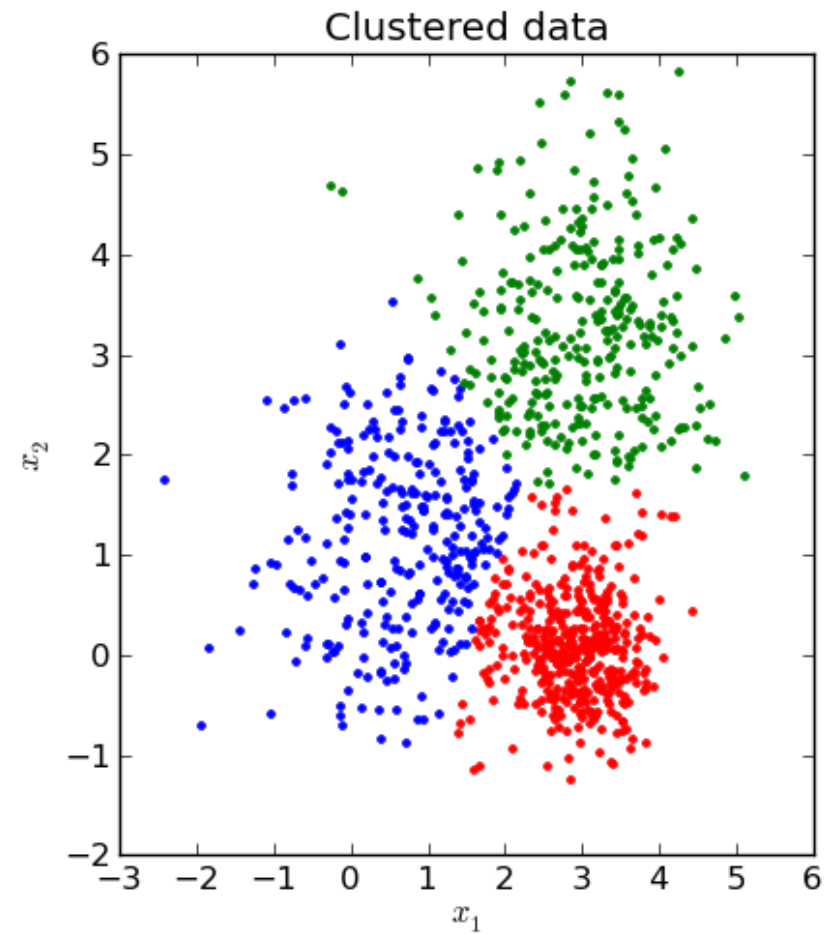
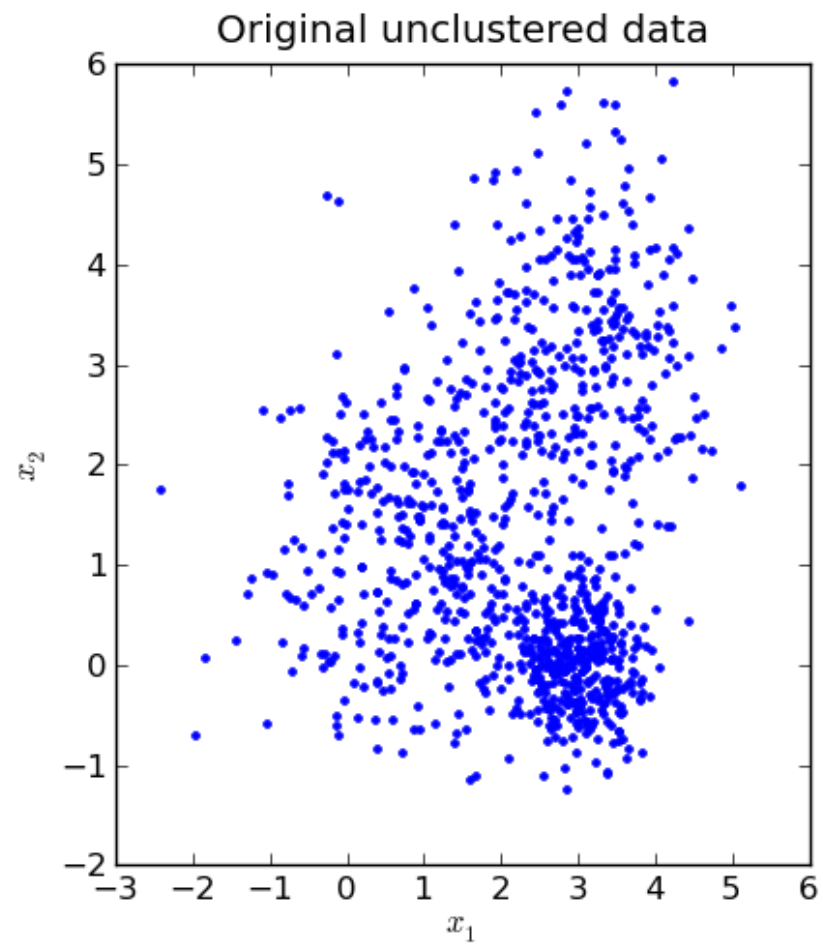
CLUSTERING

CLUSTERING

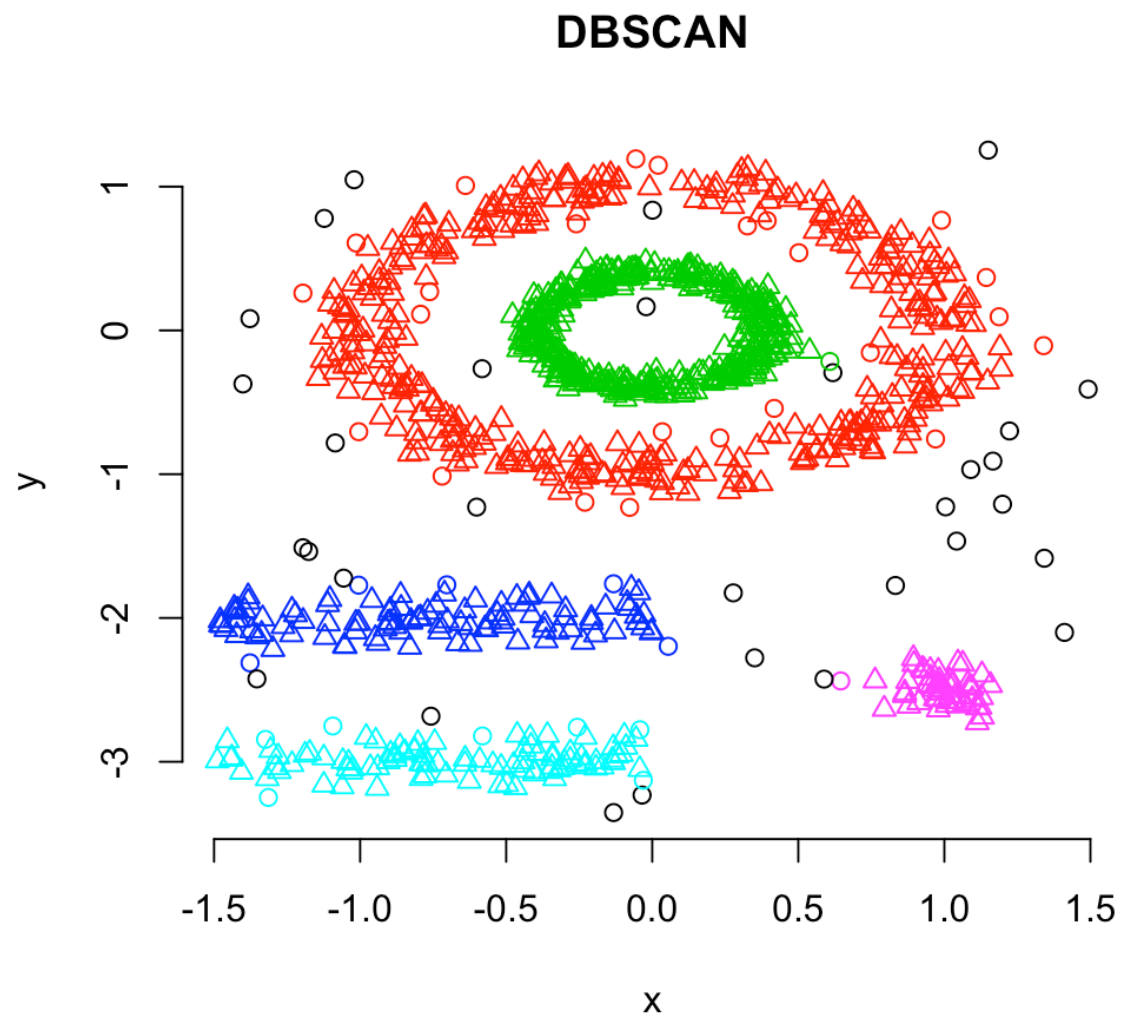
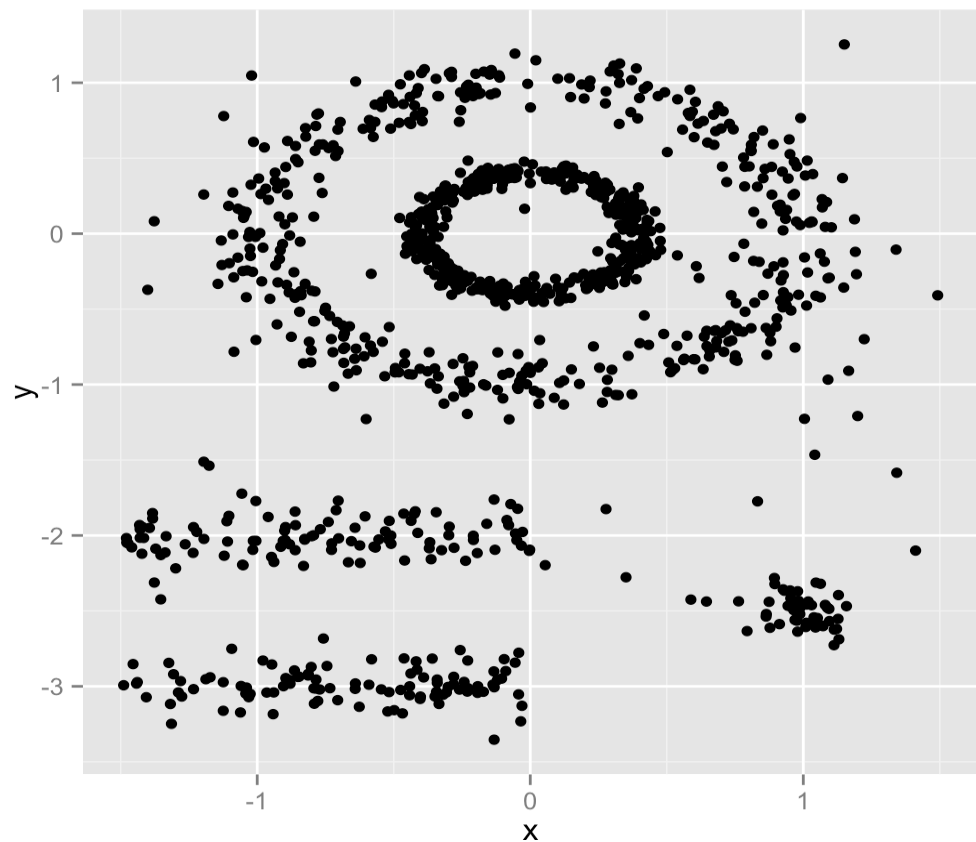
CLUSTERING

- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.
- In simple words, the aim is to segregate groups with similar traits and assign them into clusters.
- There are 3 main clustering concepts:
 - ✓ Centroids
 - ✓ Density-Based
 - ✓ Hierarchical

CLUSTERING: Centroids

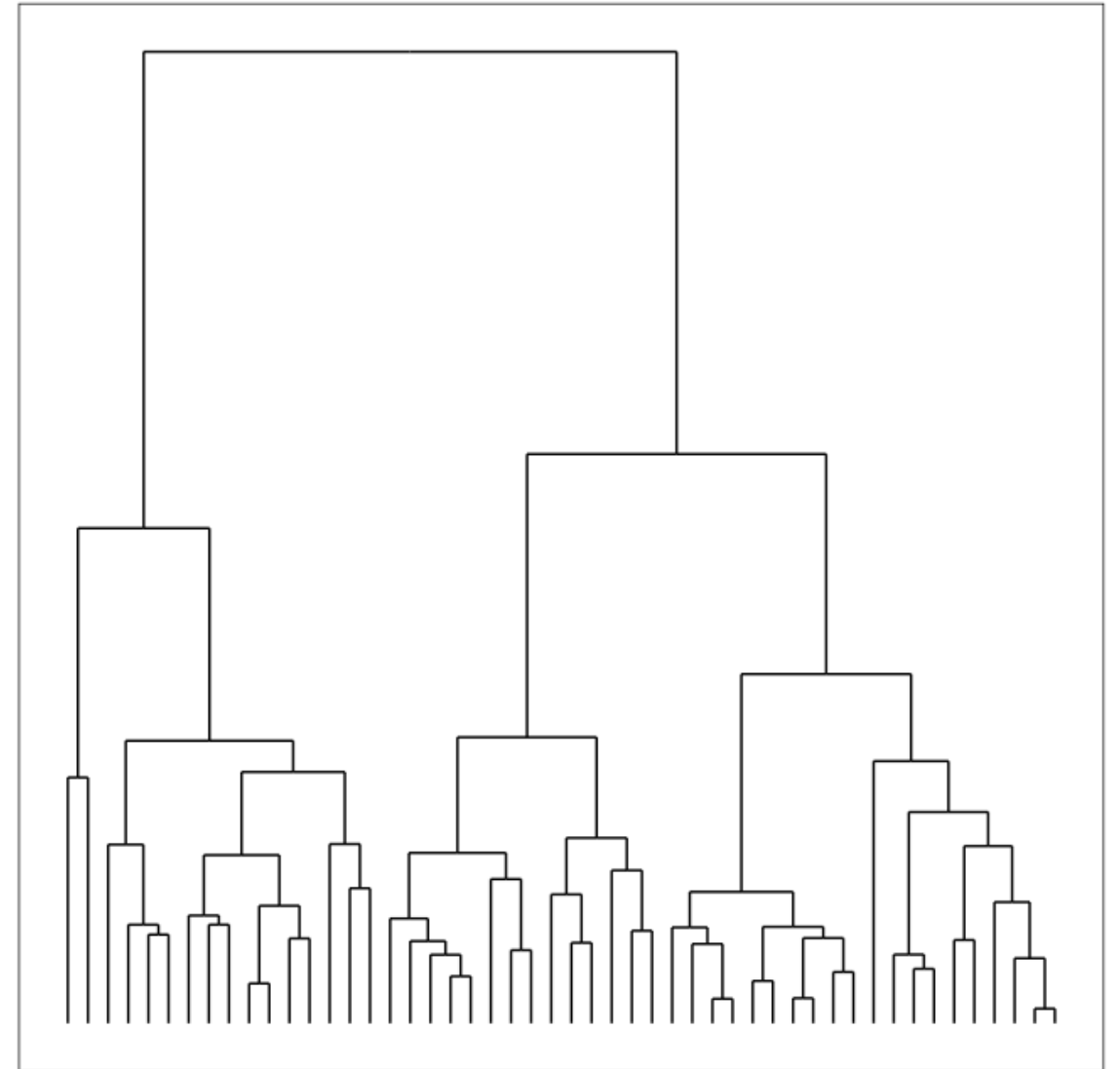


CLUSTERING: Density-Based



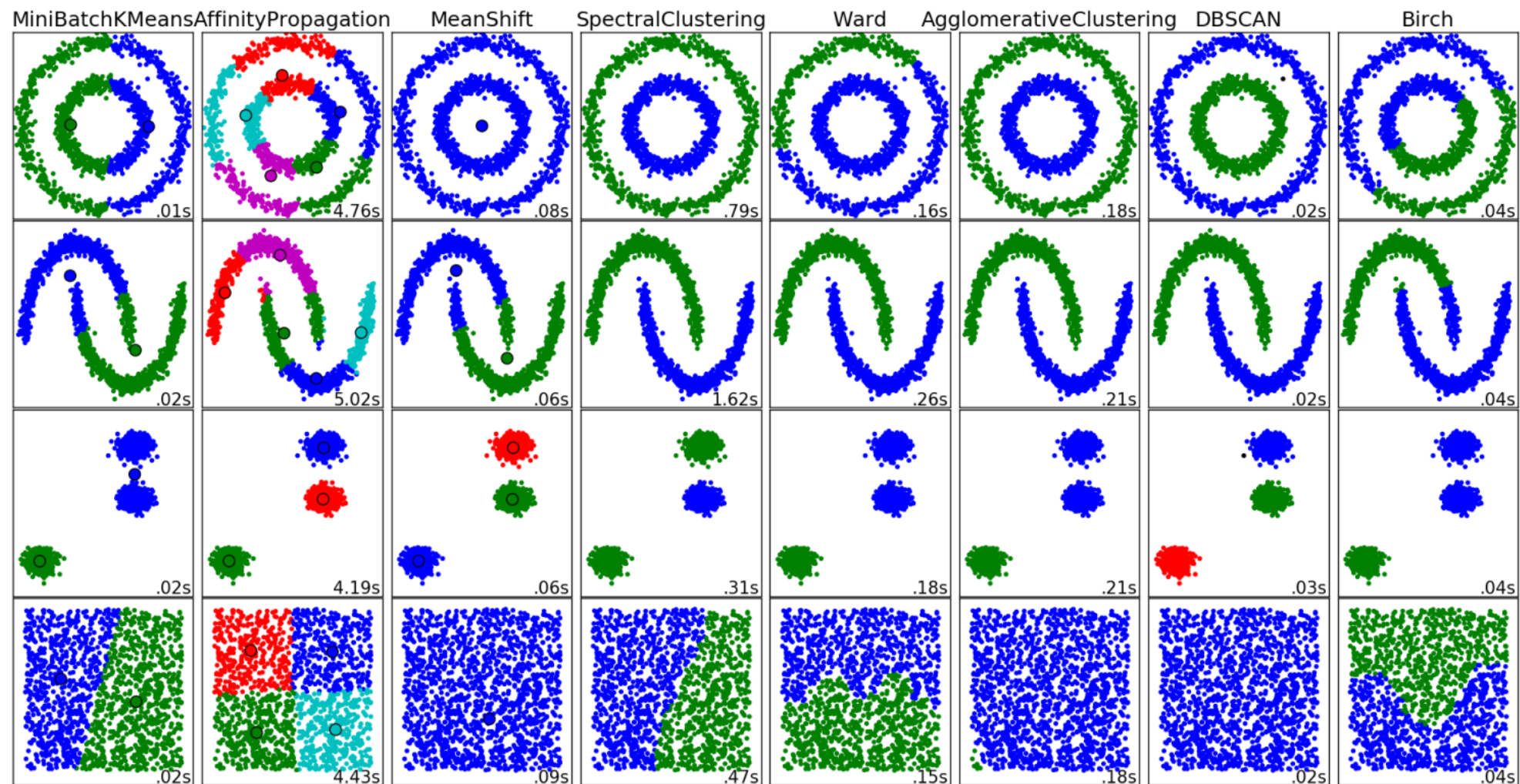
CLUSTERING: Hierarchical

- Build hierarchies that form clusters
- Based on classification trees (next lesson)



CLUSTERING

- There are [many clustering algorithms](#)



ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS



EXERCISE

1. Can you think of a real-world clustering application?

DELIVERABLE

Answers to the above questions

ACTIVITY: KNOWLEDGE CHECK

ANSWERS



EXERCISE

1. Recommendation Systems e.g. Netflix genres
2. Medical Imaging: differentiate tissues
3. Identifying market segments
4. Discover communities in social networks
5. Lots of applications for genomic sequences (homologous sequences, genotypes)
6. Earthquake epicenters
7. Fraud detection

CLUSTERING

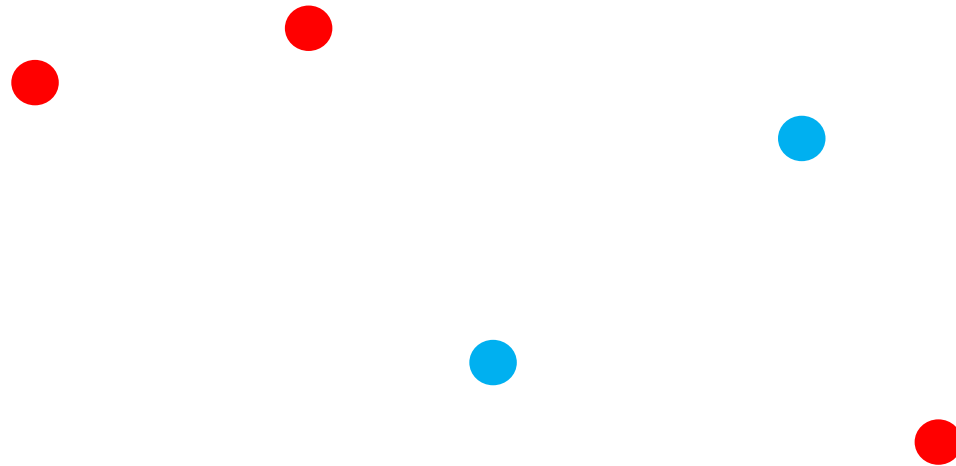
K-MEANS: CENTRIOD CLUSTERING

K-MEANS CLUSTERING

- K-Means clustering is a popular centroid-based clustering algorithm
- Basic idea: find k clusters in the data centrally located around various mean points

K-MEANS CLUSTERING: 5 Steps

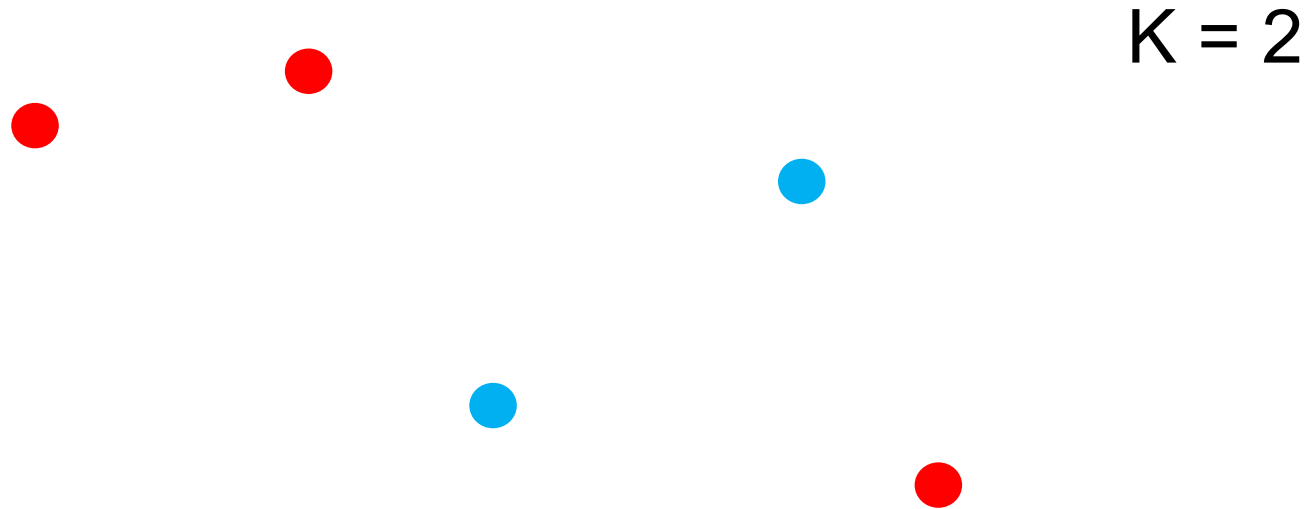
1. Specify the desired number of clusters K : Let us choose $k=2$ for these 5 data points in 2-D space.



$K = 2$

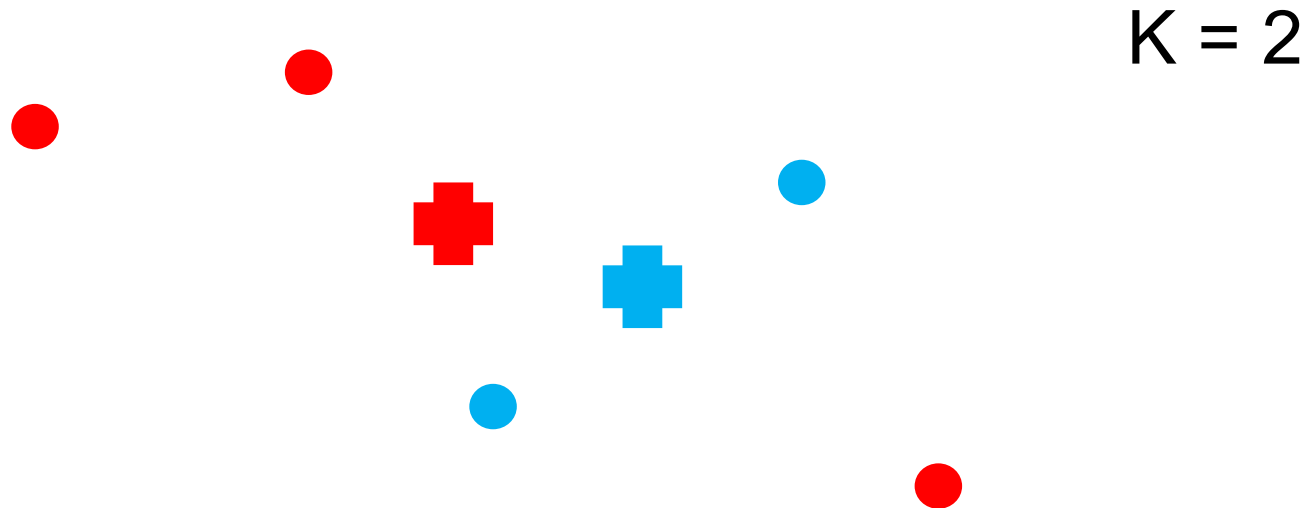
K-MEANS CLUSTERING: 5 Steps

2. Randomly assign each data point to a cluster : Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using blue color.



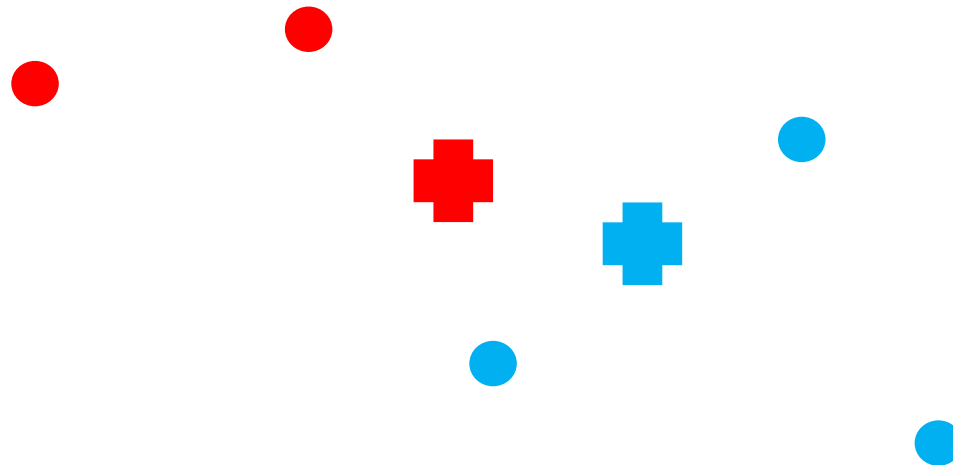
K-MEANS CLUSTERING: 5 Steps

3. Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in blue cluster using blue cross.



K-MEANS CLUSTERING: 5 Steps

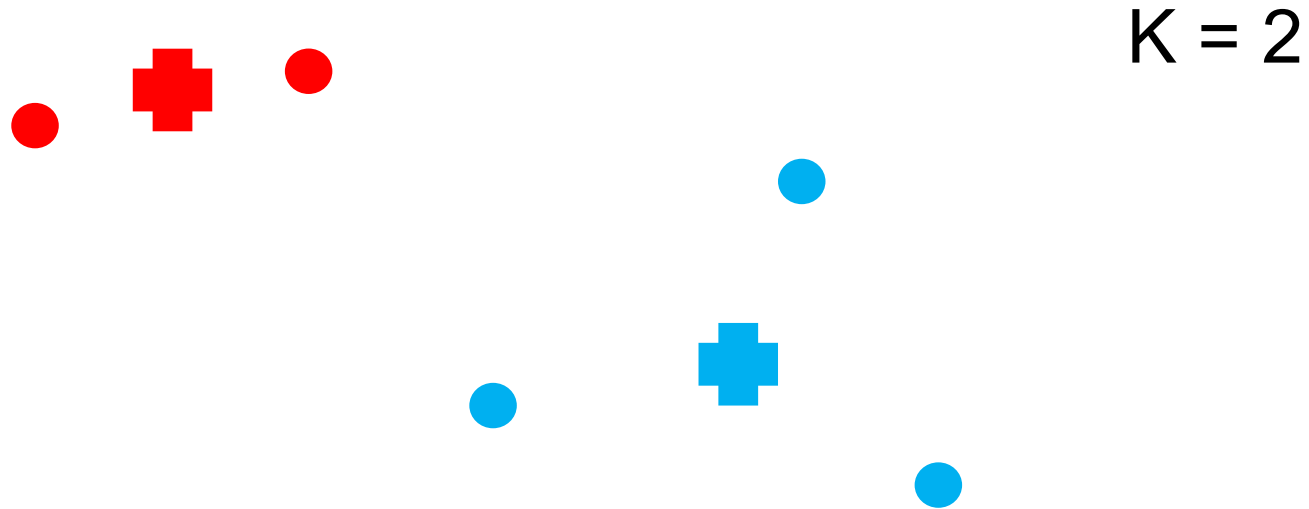
4. Re-assign each point to the closest cluster centroid : Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of blue cluster. Thus, we assign that data point into blue cluster



$K = 2$

K-MEANS CLUSTERING: 5 Steps

5. Re-compute cluster centroids : Now, re-computing the centroids for both the clusters.



Repeat until no improvements are possible. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm.

K-MEANS CLUSTERING

- `from sklearn.cluster import KMeans`
- `est = KMeans(n_clusters=3)`
- `est.fit(X)`
- `labels = est.labels_`

Let's try it out!

Go to the Lab Starter!

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS



EXERCISE

1. How do we assign meaning to the clusters we find?
2. Do clusters always have meaning?

DELIVERABLE

Answers to the above questions

K-MEANS CLUSTERING

- Assumptions are important! k-Means assumes:
 - k is the correct number of clusters
 - the data is isotropically distributed (circular/spherical distribution)
 - the variance is the same for each variable
 - clusters are roughly the same size

K-MEANS CLUSTERING

- Netflix prize: Predict how users will rate a movie
 - How might you do this with clustering?
 - Cluster similar users together and take the average rating for a given movie by users in the cluster (which have rated the movie)
 - Use the average as the prediction for users that have not yet rated the movie
- In other words, fit a model to users in a cluster for each cluster and make predictions per cluster

K-MEANS CLUSTERING

- K-MEANS advantages:
 - Simple
 - Fast for low dimensional data
 - It can find pure sub clusters if large number of clusters is specified
- K-MEANS disadvantages:
 - Will not identify outliers
 - Restricted to data which has the notion of a center
 - Clustered are assumed to be discrete: no overlap is allowed

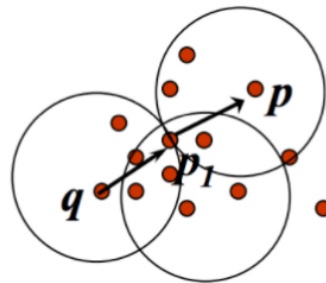
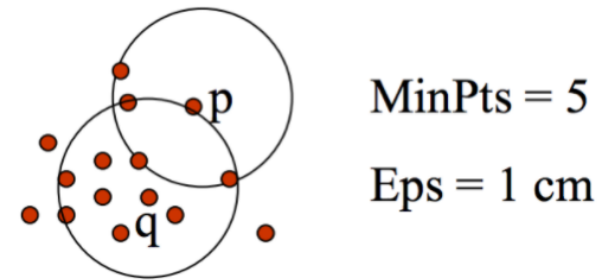
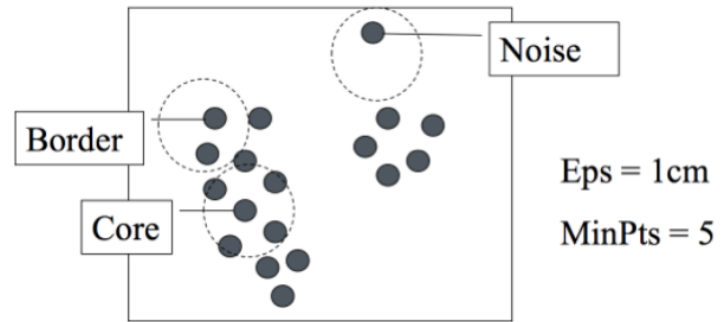
CLUSTERING

DBSCAN: DENSITY BASED CLUSTERING

DBSCAN CLUSTERING

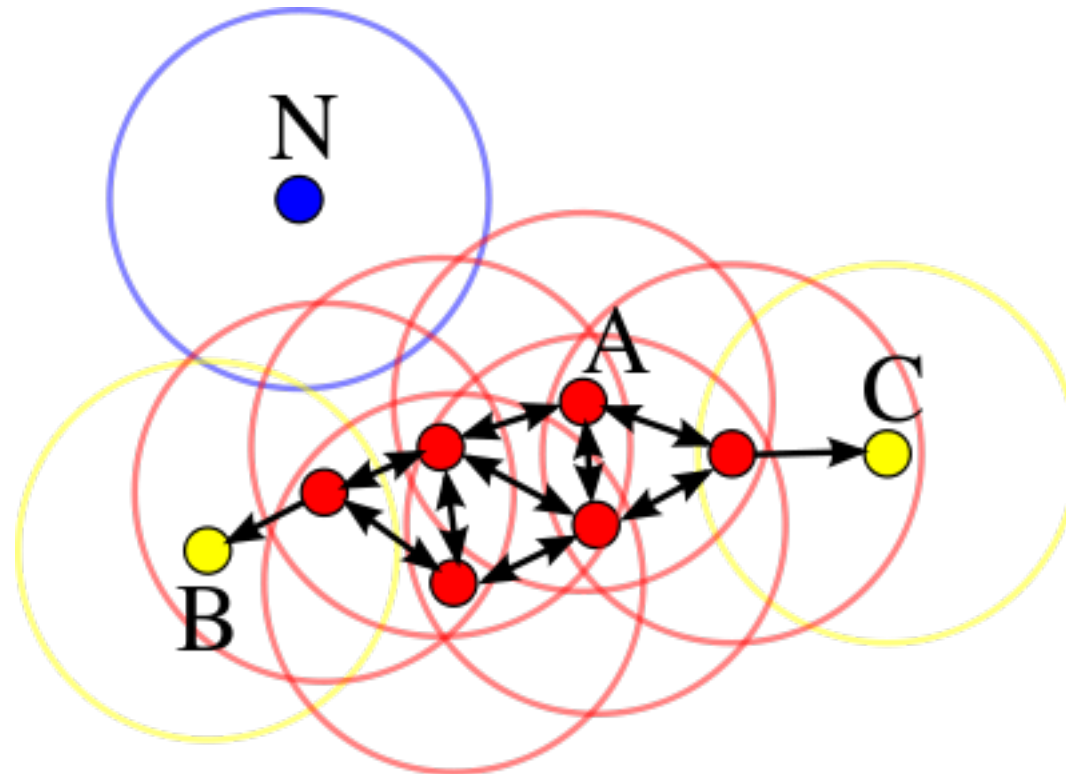
- DBSCAN: Density-based spatial clustering of applications with noise (1996)
- Main idea: Group together closely-packed points by identifying
 - Core points
 - Reachable points
 - Outliers (not reachable)
- Two parameters:
 - min_samples
 - eps

DBSCAN CLUSTERING



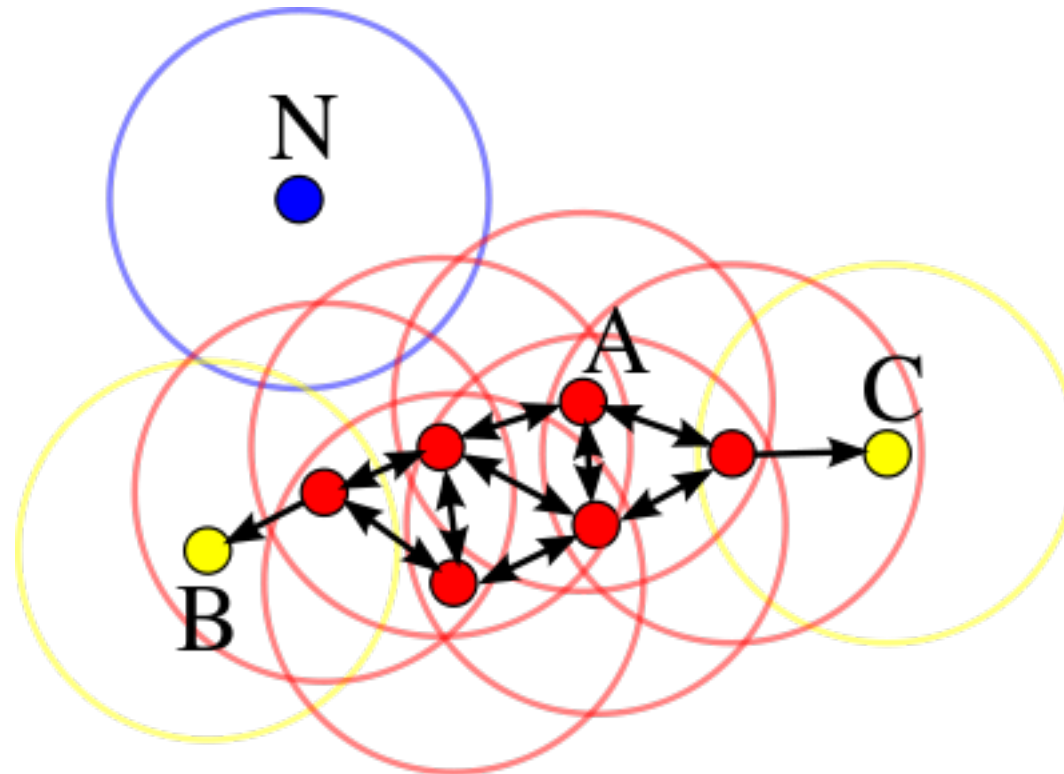
DBSCAN CLUSTERING

- Core points: at least **min_samples** points within **eps** of the core point
 - Such points are *directly reachable* from the core point
- Reachable: point q is reachable from p if there is a path of core points from p to q
- Outlier: not reachable



DBSCAN CLUSTERING

- A cluster is a collection of connected core and reachable points



DBSCAN CLUSTERING

The Algorithm

1. Randomly choose a point p .
2. Retrieve all points density-reachable from p w.r.t. Eps and $MinPts$.
3. If p is a core point, a cluster is formed.
4. If p is a border point, no points are density-reachable from p , then visit the next point.
5. Repeat the process until all the data points have been processed.

DBSCAN CLUSTERING

- `from sklearn.cluster import DBSCAN`
- `est = DBSCAN(eps=0.5, min_samples=10)`
- `est.fit(X)`
- `labels = est.labels_`

Let's try it out!

Go to the Lab Starter!

DBSCAN CLUSTERING

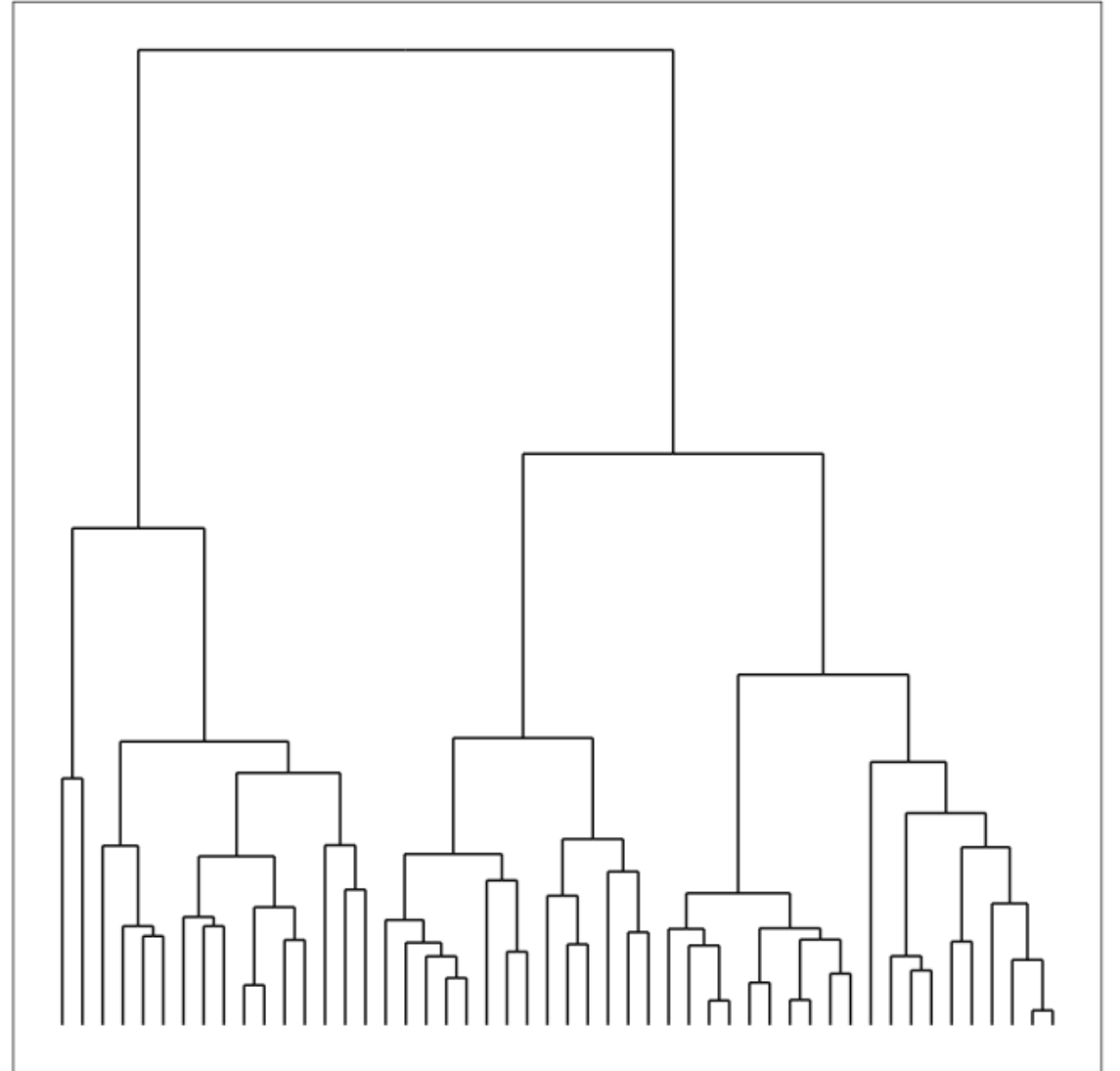
- DBSCAN advantages:
 - Can find arbitrarily-shaped clusters
 - Don't have to specify number of clusters
 - Robust to outliers
- DBSCAN disadvantages:
 - Doesn't work well when clusters are of varying densities
 - hard to chose parameters that work for all clusters
 - Can be hard to chose correct parameters regardless

CLUSTERING

HIERARCHICAL CLUSTERING

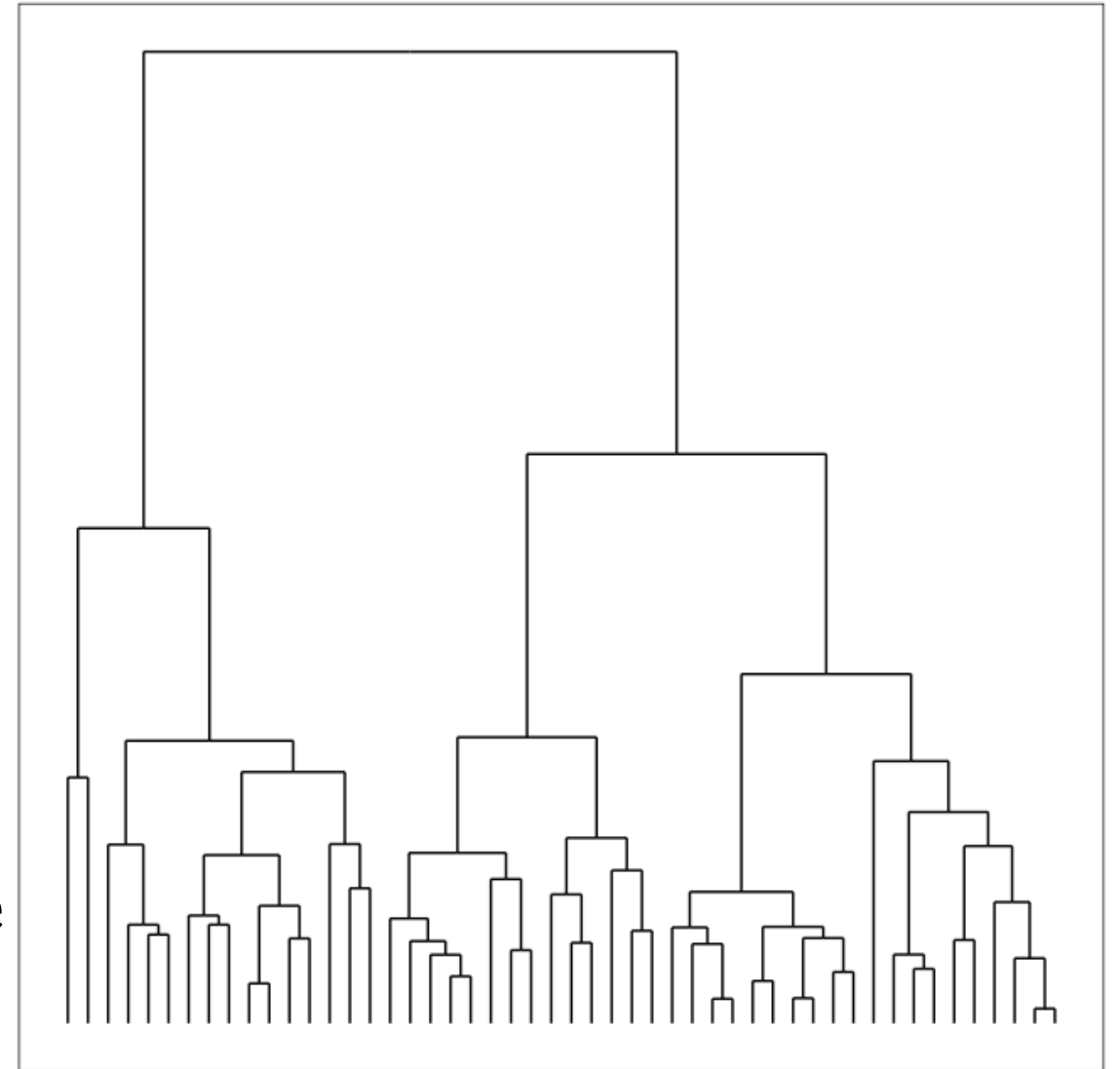
CLUSTERING: Hierarchical

- Build hierarchies that form clusters
- Based on classification trees (next lesson)



CLUSTERING: Hierarchical (Additional Tips: not required)

- At the bottom, we start with 50 data points, each assigned to separate clusters.
- Two closest clusters are then merged till we have just one cluster at the top.
- The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space.
- The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram.



HIERARCHICAL CLUSTERING

We'll discuss the details once we cover decision trees. For now we can black box the model and fit with sklearn

- `from sklearn.cluster import AgglomerativeClustering`
- `est = AgglomerativeClustering(n_clusters=4)`
- `est.fit(X)`
- `labels = est.labels_`

Let's try it out!

Go to the Lab Starter!

HIERARCHICAL CLUSTERING

- HIERARCHICAL advantages:
 - Easy to implement
 - Outputs a hierarchy to decide on the number of clusters
- HIERARCHICAL disadvantages:
 - Not suitable for large datasets
 - The order of the data has an impact on final results
 - Very sensitive to outliers

CLUSTERING

CLUSTERING METRICS

CLUSTERING METRICS

- As usual we need a metric to evaluate model fit
- For clustering we use a metric called the [Silhouette Coefficient](#)
 - **a** is the mean distance between a sample and all other points in the cluster
 - **b** is the mean distance between a sample and all other points in the *nearest* cluster

- The Silhouette Coefficient is:

$$\frac{b - a}{\max(a, b)}$$

- Ranges between 1 and -1
- Average over all points to judge the cluster algorithm

CLUSTERING METRICS (Additional Tips: not required)

Unsupervised method

Silhouette Coefficient: Evaluate how well the **compactness** and the **separation** of the clusters are. (Note that the notation below is consistent with the above content.) Using *Silhouette Coefficient*, we can choose an optimal value for number of clusters.

$a(x_i)$ denotes the **mean intra-cluster distance**. Evaluate the compactness of the cluster to which x_i belongs. (The smaller the more compact)

$$a(x_i) = \frac{\sum_{x_k \in C_j, k \neq i} D(x_i, x_k)}{|C_j| - 1}$$

For the data point x_i , calculate its average distance to all the other data points in its cluster. (Minusing one in denominator part is to leave out the current data point x_i)

$b(x_i)$ denotes the **mean nearest-cluster distance**. Evaluate how x_i is separated from other clusters. (The larger the more separated)

$$b(x_i) = \min_{C_j: 1 \leq j \leq k, x_i \notin C_j} \left\{ \frac{\sum_{x_k \in C_j} D(x_i, x_k)}{|C_j|} \right\}$$

For the data point x_i and all the other clusters not containing x_i , calculate its average distance to all the other data points in the given clusters. Find the minimum distance value with respect to the given clusters.

Finally, *Silhouette Coefficient*: $s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}$, $-1 \leq s(x_i) \leq 1$. Want $a(x_i) < b(x_i)$ and $a(x_i) \rightarrow 0$ so as to $s(x_i) \rightarrow 1$.

CLUSTERING METRICS

- `from sklearn import metrics`
- `from sklearn.cluster import KMeans`
- `kmeans_model = KMeans(n_clusters=3, random_state=1).fit(X)`
- `labels = kmeans_model.labels_`
- `metrics.silhouette_score(X, labels, metric='euclidean')`

Let's check the starter lab!

CLUSTERING METRICS

- There are a number of [other metrics](#) based on:
 - Mutual Information
 - Homogeneity
 - Adjusted Rand Index (when you know the labels on the training data)

PUTTING IT TOGETHER

**CLUSTERING,
CLASSIFICATION,
AND REGRESSION**

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS



EXERCISE

1. How might we combine clustering and classification?

DELIVERABLE

Answers to the above questions

CLUSTERING, CLASSIFICATION, AND REGRESSION

- We can use clustering to discover new features and then use those features for either classification or regression
- For classification, we could use e.g. k-NN to classify new points into the discovered clusters
- For regression, we could use a dummy variable for the clusters as a variable in our regression

ACTIVITY: CLUSTERING + CLASSIFICATION

EXERCISE



EXERCISE

1. Using the starter code, perform a k-means clustering on the flight delay data
2. Use the clustering to create a classifier

DELIVERABLE

A completed notebook

CONCLUSION

TOPIC REVIEW

REVIEW AND NEXT STEPS

- Clustering is used to discover features, e.g. segment users or assign labels (such as species)
- Clustering may be the goal (user marketing) or a step in a data science pipeline

COURSE

**BEFORE NEXT
CLASS**

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

What is Data Science	Lesson 1
Research Design and Pandas	Lesson 2
Statistics Fundamentals I	Lesson 3
Statistics Fundamentals II	Lesson 4
Flexible Class Session	Lesson 5

UNIT 2: FOUNDATIONS OF DATA MODELING

Introduction to Regression	Lesson 6
Evaluating Model Fit	Lesson 7
Introduction to Classification	Lesson 8
Introduction to Logistic Regression	Lesson 9
Communicating Logistic Regression Results	Lesson 10
Flexible Class Session	Lesson 11

UNIT 3: DATA SCIENCE IN THE REAL WORLD

› Decision Trees and Random Forests	Lesson 12
› Natural Language Processing	Lesson 13
› Dimensionality Reduction	Lesson 14
› Time Series Data I	Lesson 15
› Time Series Data II	Lesson 16
› Database Technologies	Lesson 17
› Where to Go Next	Lesson 18
› Flexible Class Session	Lesson 19
› Final Project Presentations	Lesson 20



BEFORE NEXT CLASS

UPCOMING

- Project: Final Project 2 due on Tuesday

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET