GENERAL ASSEMBLY

CLUSTERING

Adam Jones, PhD

Data Scientist @ Critical Juncture

LEARNING OBJECTIVES

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

UNSUPERVISED LEARNING

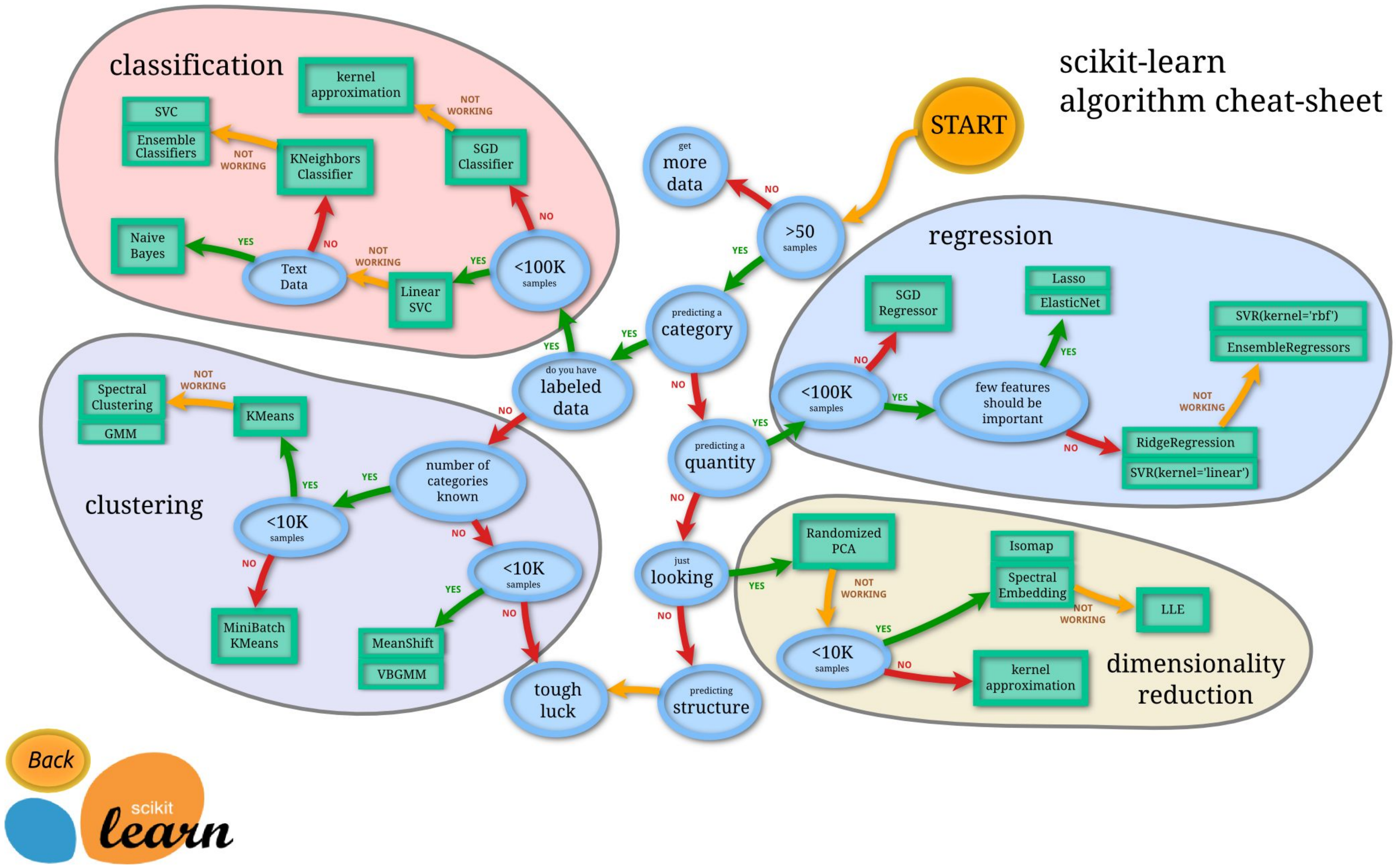
UNSUPERVISED LEARNING

- 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



scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
<10K samples
MeanShift
VBGMM
regression
<100K samples
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
dimensionality reduction
Randomized PCA
<10K samples
Isomap
Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING
Back
scikit
learn

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. How is unsupervised learning different from classification?

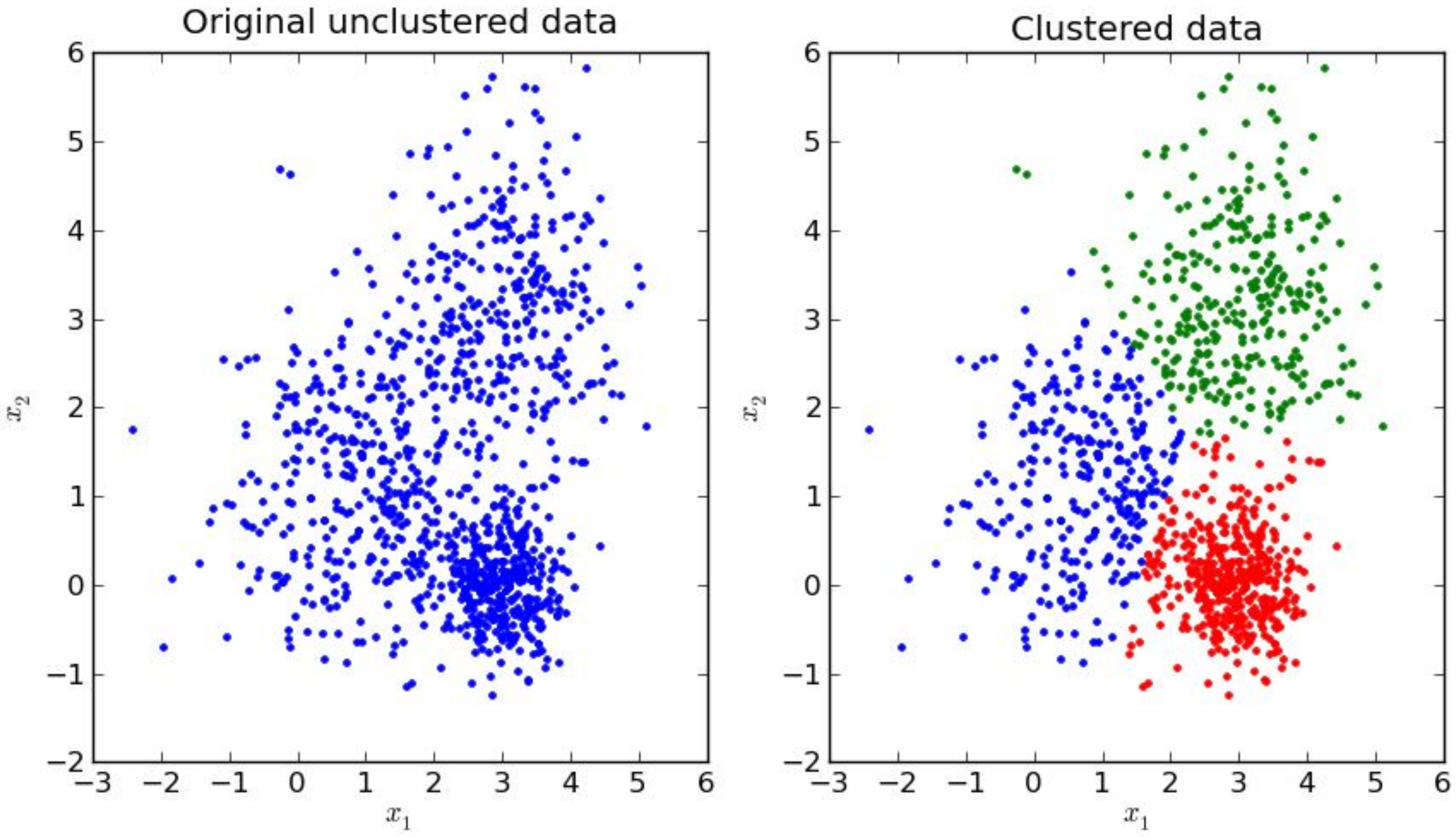
DELIVERABLE

Answers to the above questions

INTRODUCTION

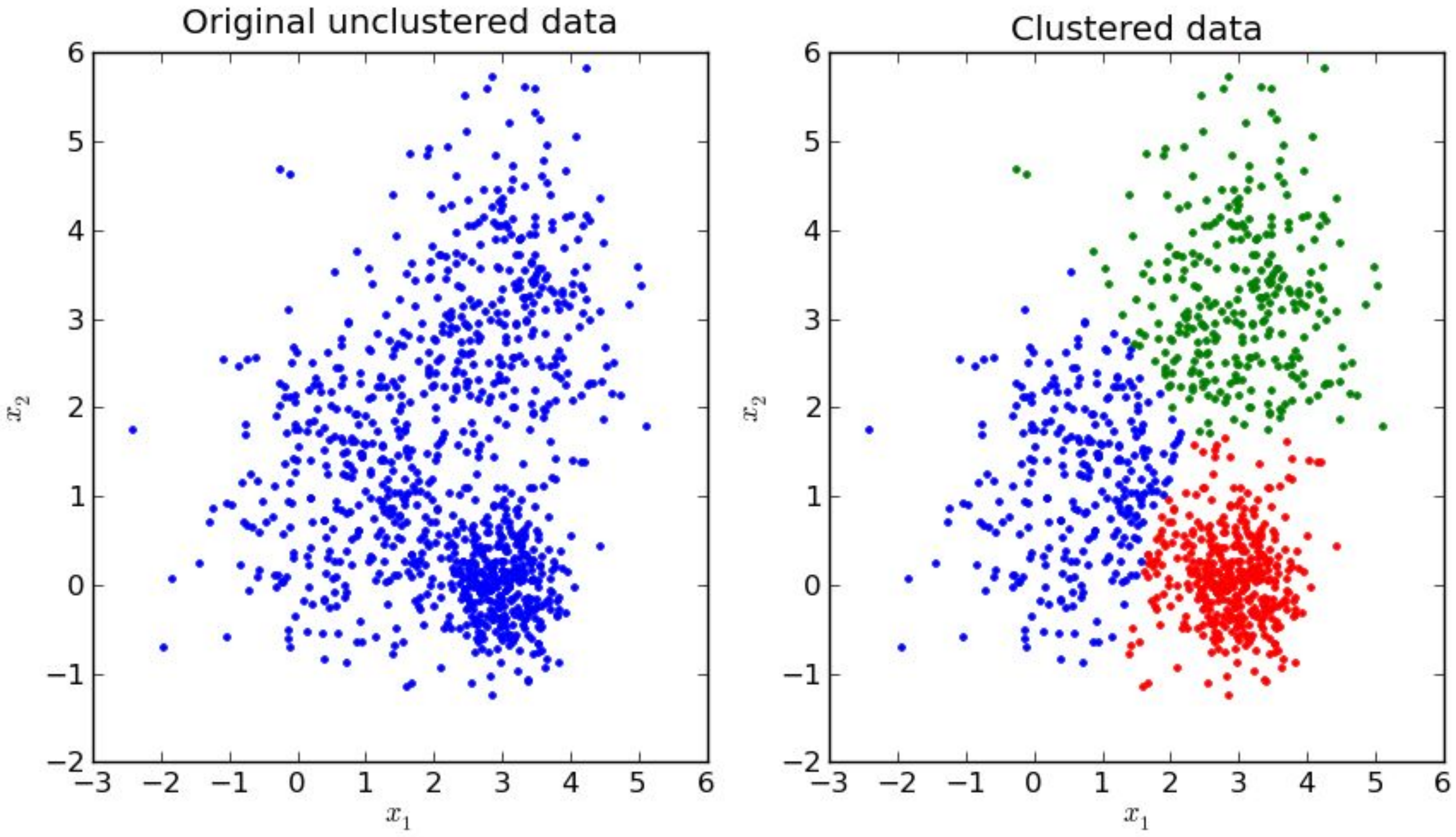
CLUSTERING

CLUSTERING: Centroids



Source: http://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue

CLUSTERING: Centroids



Source: http://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

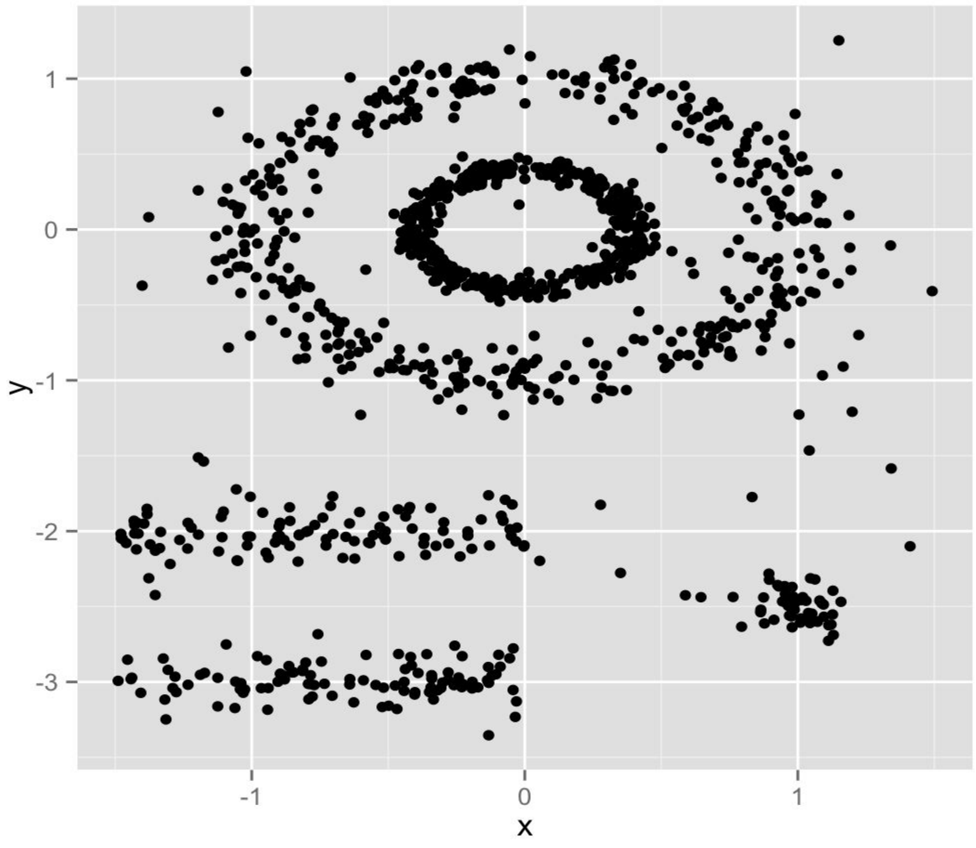
ANSWER THE FOLLOWING QUESTIONS

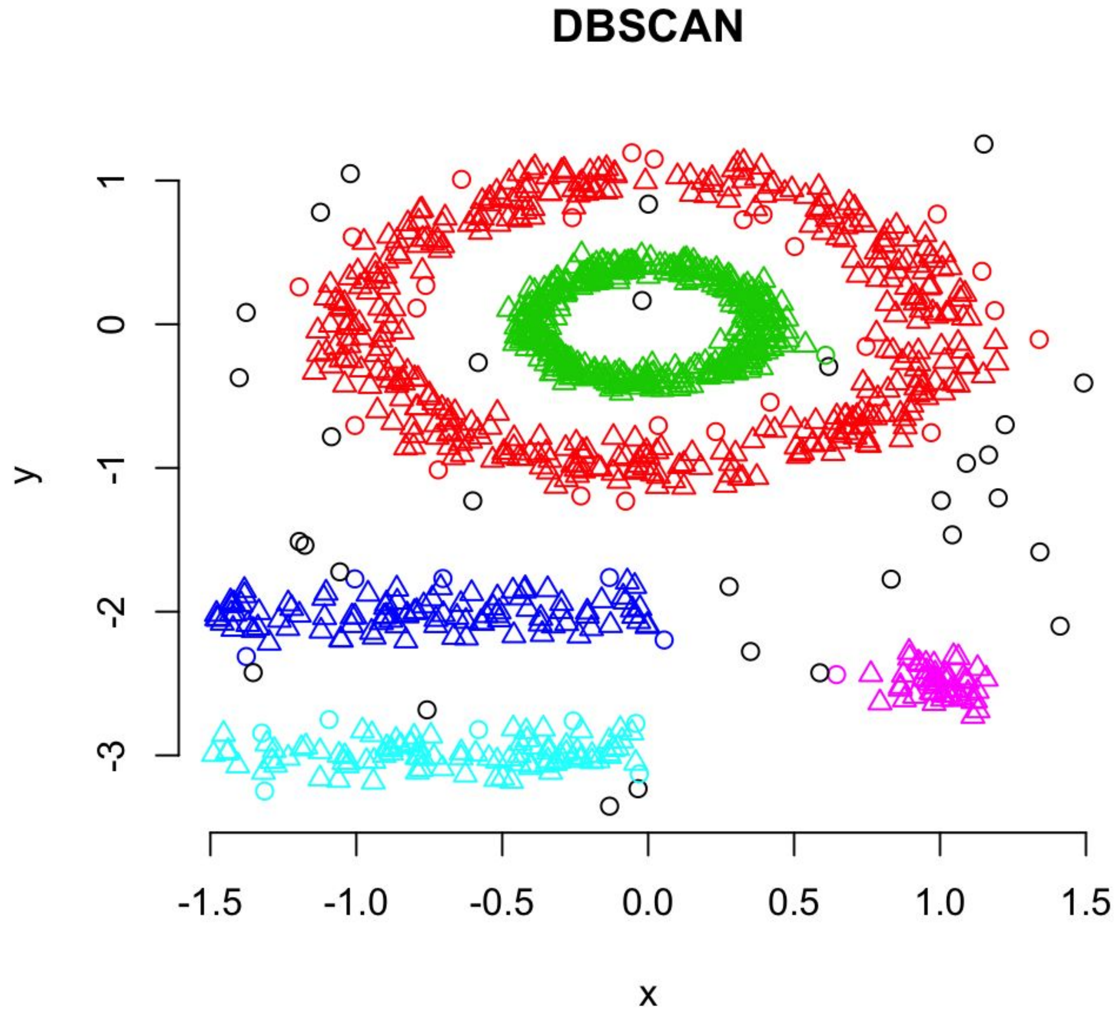
1. Why might data often appear in centered clusters?

DELIVERABLE

Answers to the above questions

CLUSTERING: Density-Based





Source: http://www.sthda.com/english/wiki/dbscan-density-based-clustering-for-discovering-clusters-in-large-datasets-with-noise-unsupervised-machine-learning

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

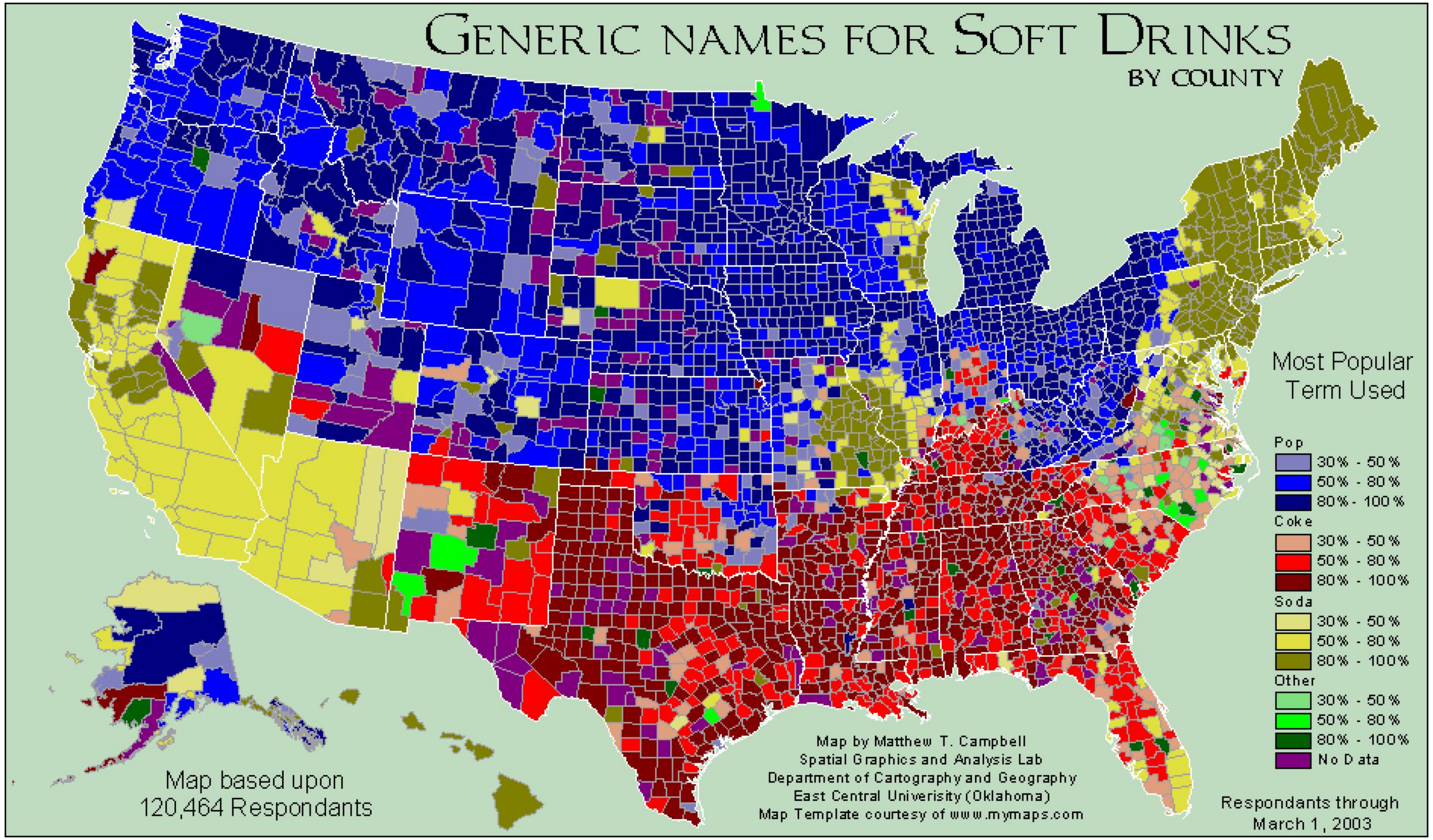
ANSWER THE FOLLOWING QUESTIONS

1. Why might data often appear in density-based clusters?

DELIVERABLE

Answers to the above questions

ACTIVITY: KNOWLEDGE CHECK

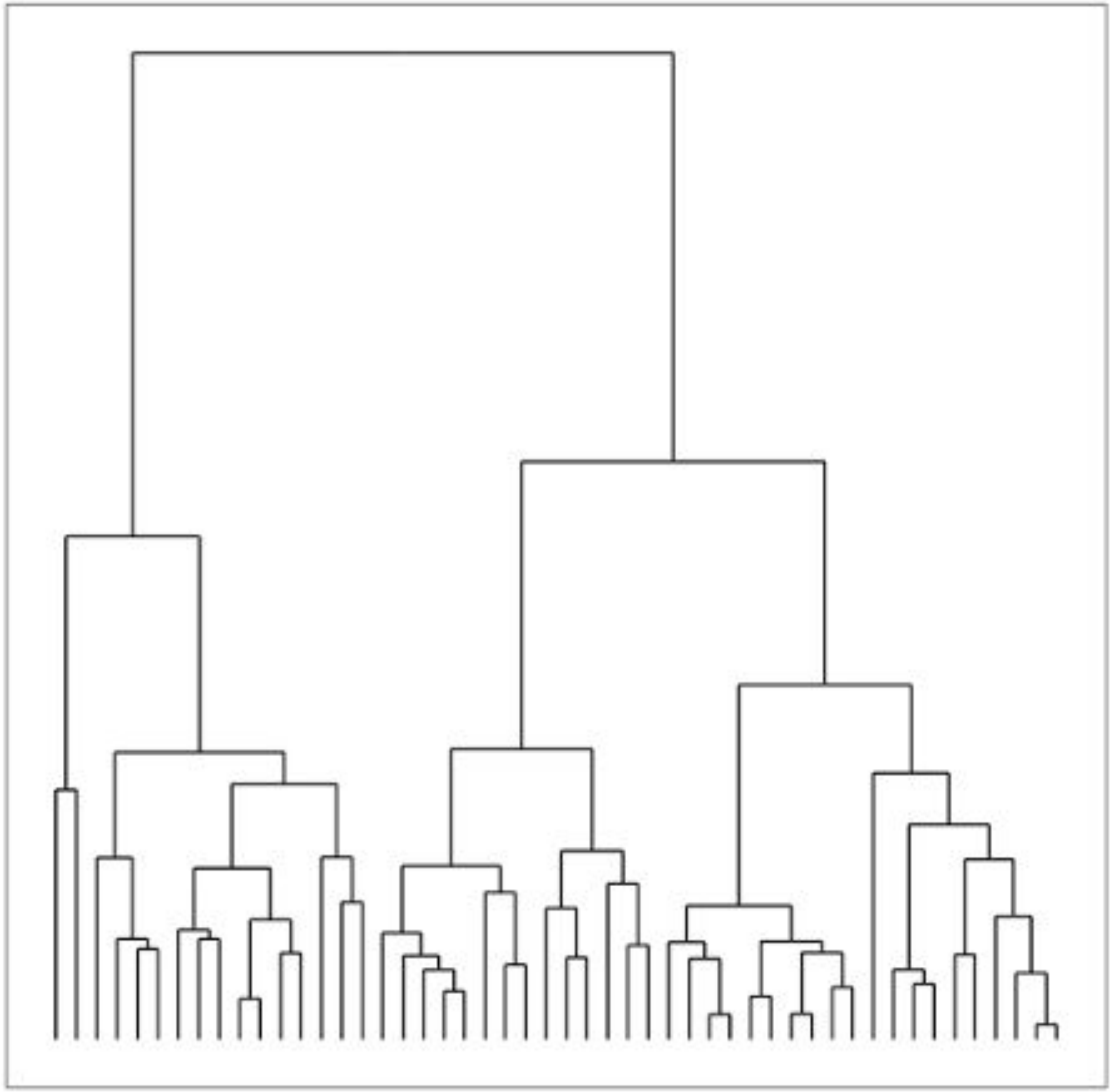


See also: http://www4.ncsu.edu/~jakatz2/files/dialectposter.png

CLUSTERING: Hierarchical

- Goal: Build hierarchies that form clusters

- Based on **classification trees**
  - (see *next* lesson)



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

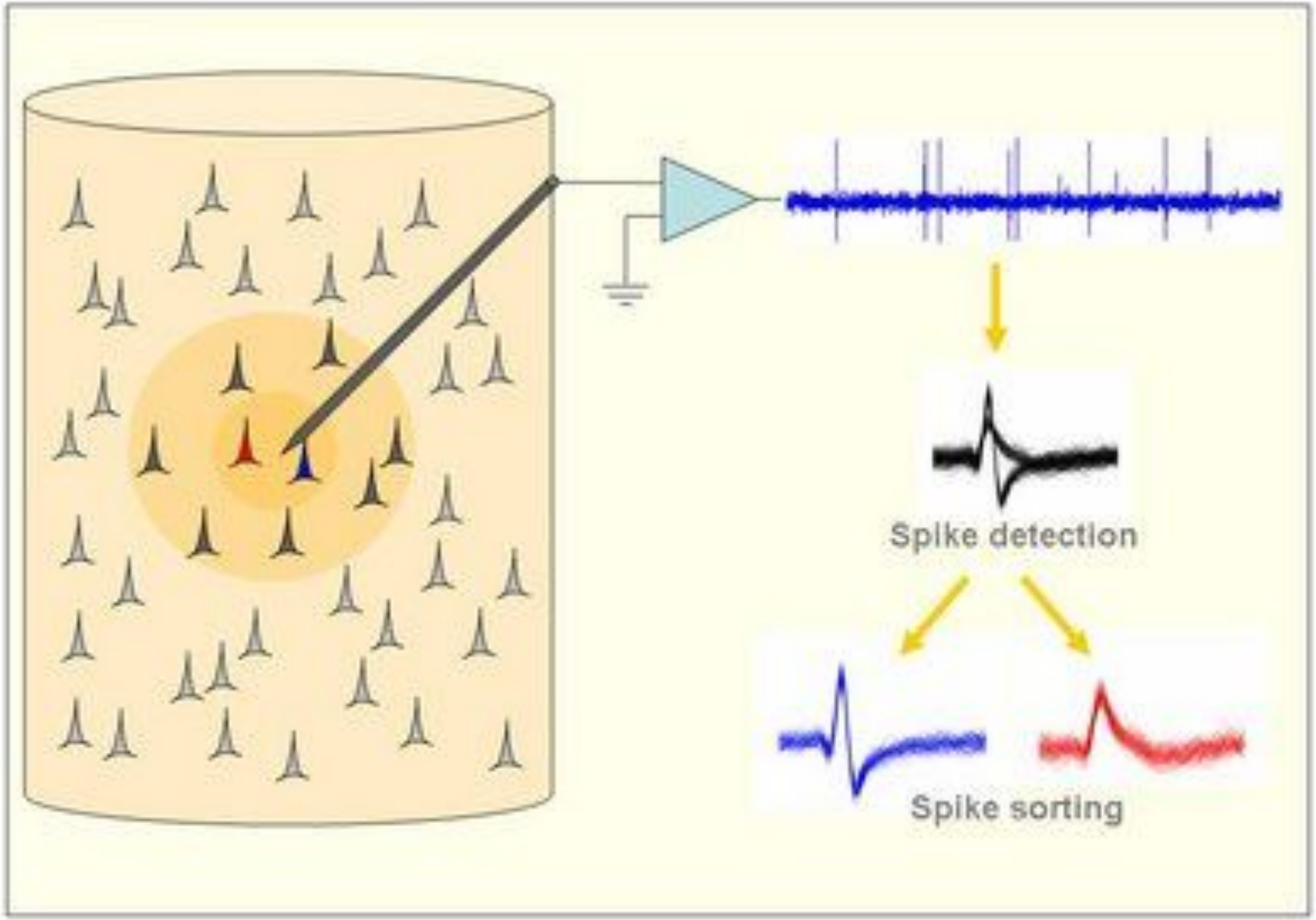
1. Can you think of a real-world clustering application?

DELIVERABLE

Answers to the above questions

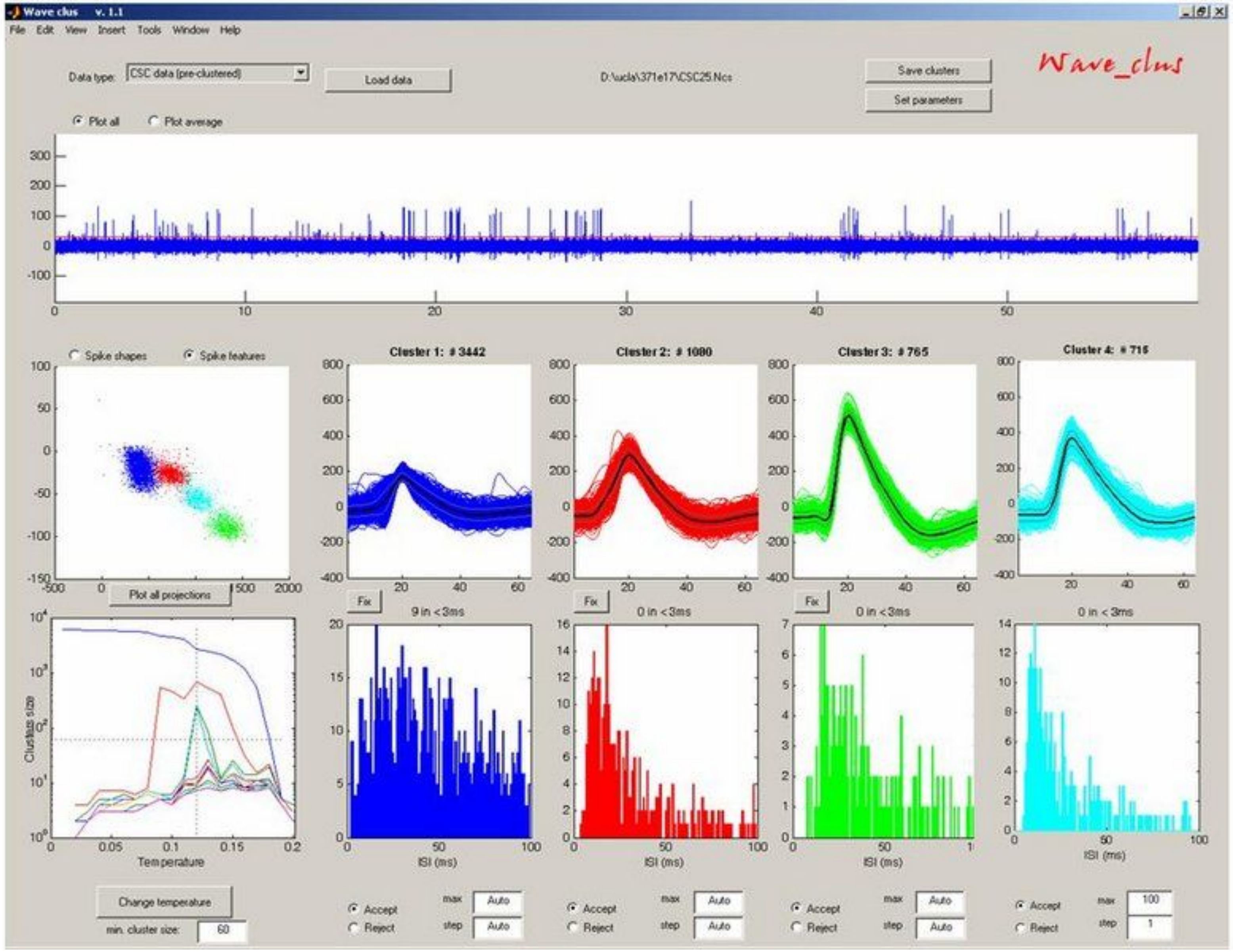
ACTIVITY: KNOWLEDGE CHECK

EXERCISE



ACTIVITY: KNOWLEDGE CHECK

EXERCISE



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWERS

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

K-MEANS: CENTROID 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
- Awesome Demo

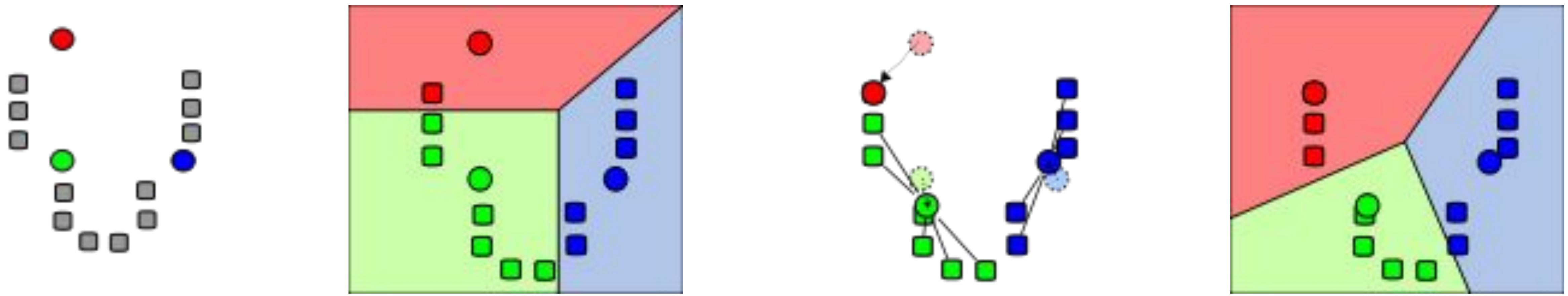
K-MEANS CLUSTERING

- This is a computationally difficult problem to solve so we rely on heuristics

- The "standard" heuristic is called "Lloyd's Algorithm":
  - Start with k initial mean values
  - Data points are then split up into a Voronoi diagram
    - Each point is assigned to the "closest" mean
  - Calculate new means based on centroids of points in the cluster
  - Repeat until clusters do not change

K-MEANS CLUSTERING

- Start with initial k mean values
- Data points are then split up into a Voronoi diagram
- Calculate new means based on centroids



K-MEANS CLUSTERING

- k-Means seeks to minimize the sum of squares about the means
- Precisely, find k subsets S_1, … S_k of the data with means mu_1, …, mu_k that minimizes:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

K-MEANS CLUSTERING

- **k-Means** assumptions:
  - k is the correct number of clusters
  - the variance is the same for each variable
  - clusters are roughly the same size

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What is **cluster analysis**?
2. How do we assign meaning to the clusters we find?
- Do clusters always have meaning?

DELIVERABLE

Answers to the above questions

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 for the Netflix Prize

DBSCAN: DENSITY BASED CLUSTERING

DBSCAN CLUSTERING

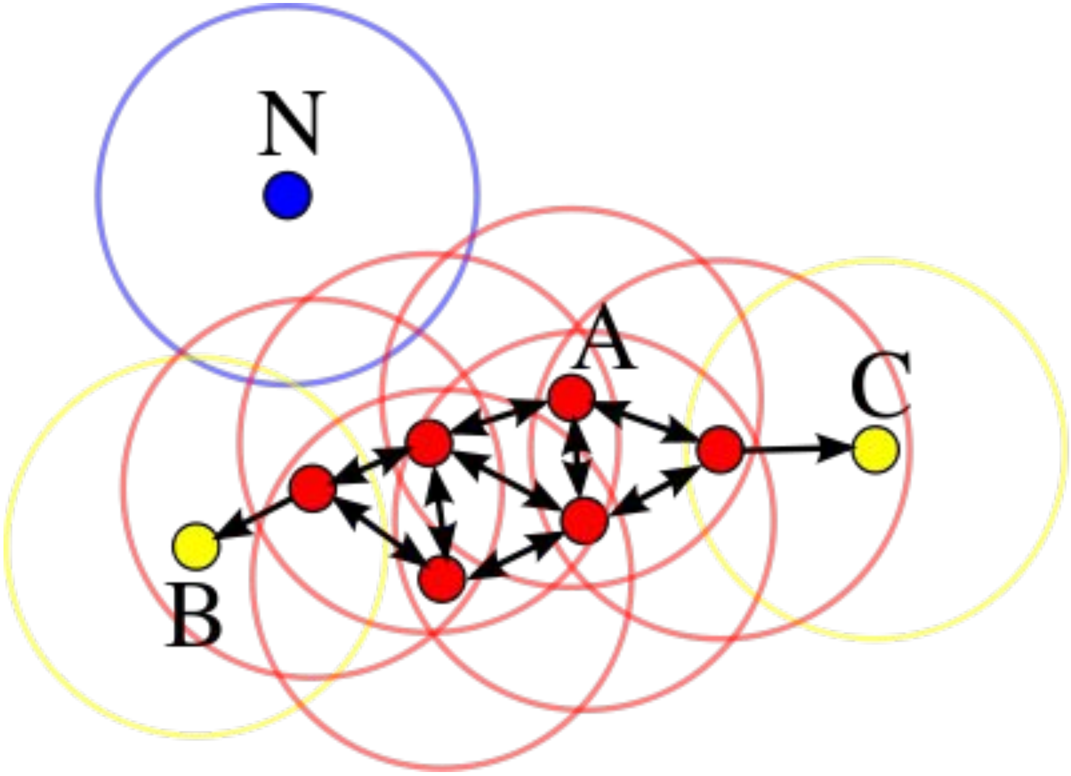
- DBSCAN: Density-based spatial clustering of applications with noise

- Main idea: Group together closely-packed points by identifying
  - Core points
  - Reachable points
  - Outliers (not reachable)

- Two parameters:
  - ε (eps)
  - min_samples

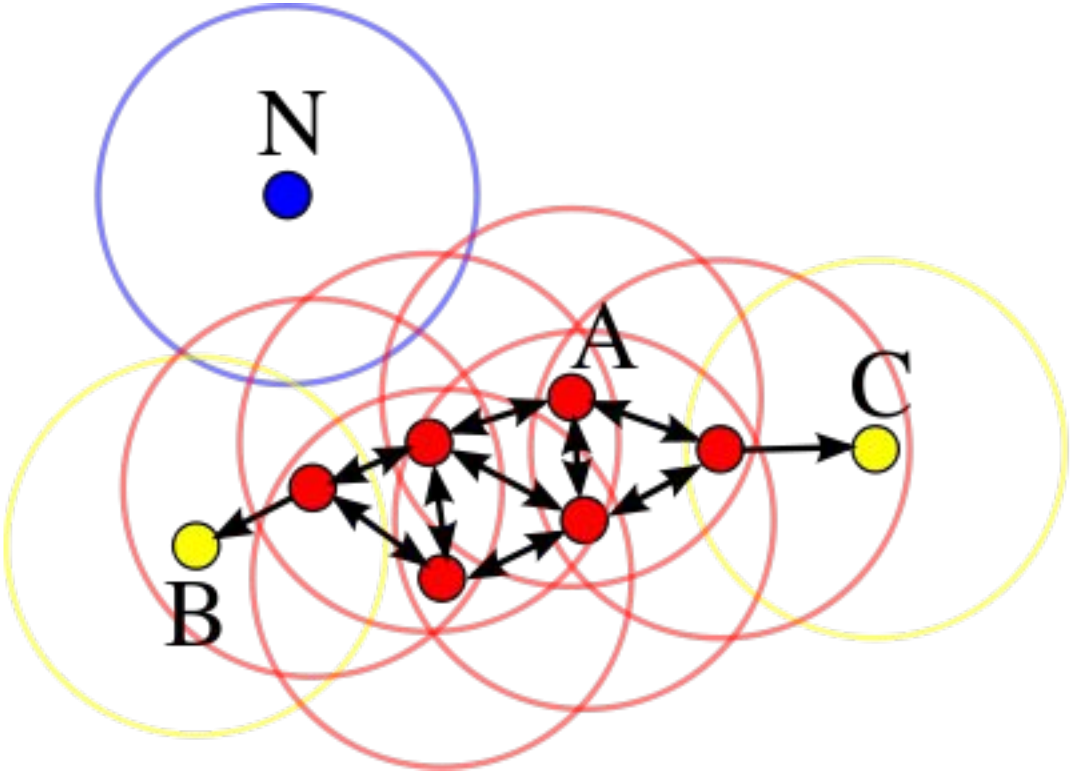
DBSCAN CLUSTERING

- 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

- Each cluster is a collection of connected points reachable by ε (or less)



CLUSTERING: Density-Based

Advantages:

- Finds non linearly separable (arbitrarily-shaped) clusters
- No need to assume a fixed number of clusters
- Robust to outliers

CLUSTERING: Density-Based

Disadvantages:

- Sensitivity to Euclidean distance measure problems
    - “Curse of dimensionality”
- Doesn’t work well when clusters are of varying densities

- Awesome Demo

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. How does DBSCAN differ from k-means?

DELIVERABLE

Answers to the above questions

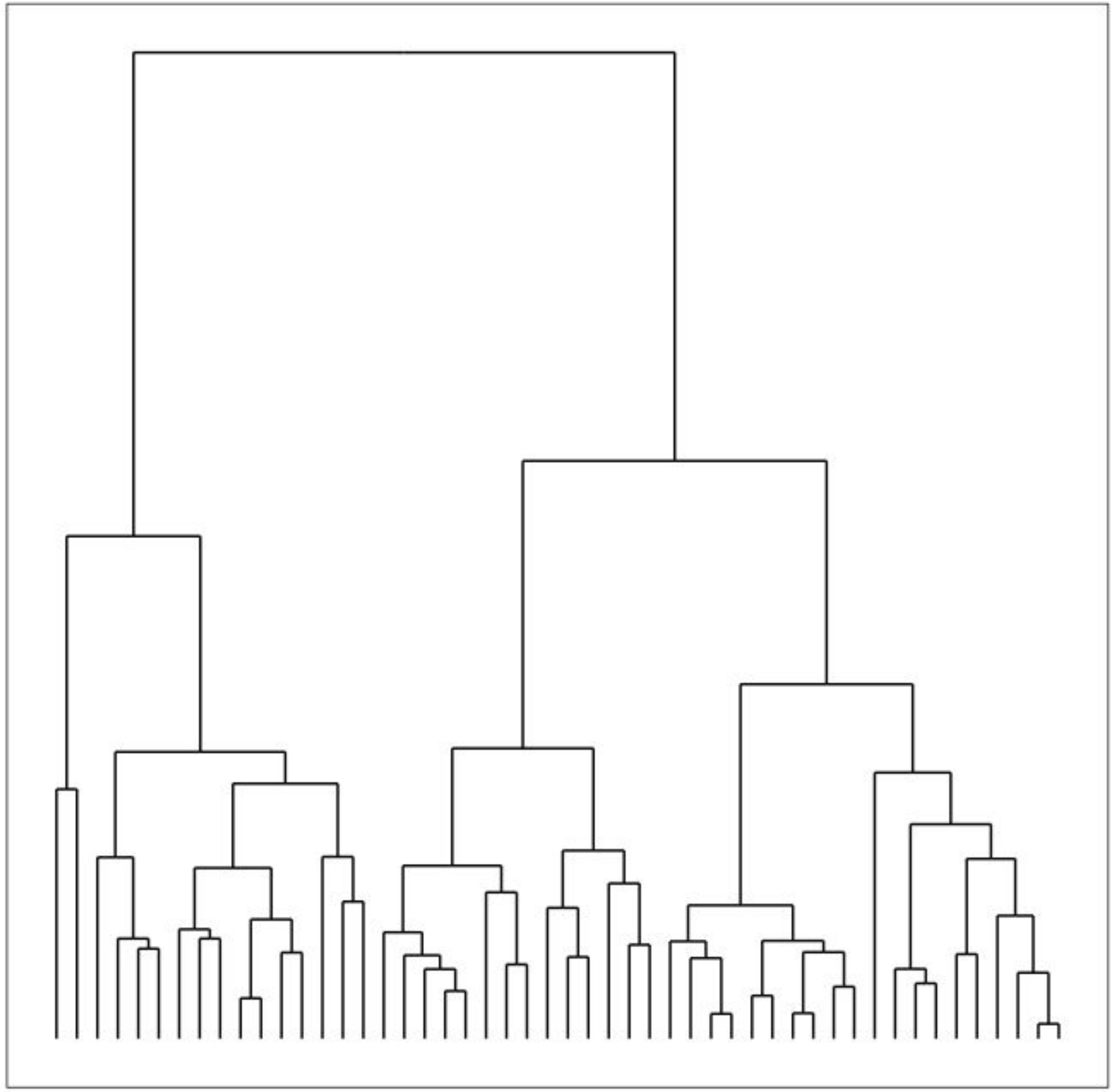
HIERARCHICAL CLUSTERING

CLUSTERING: Hierarchical

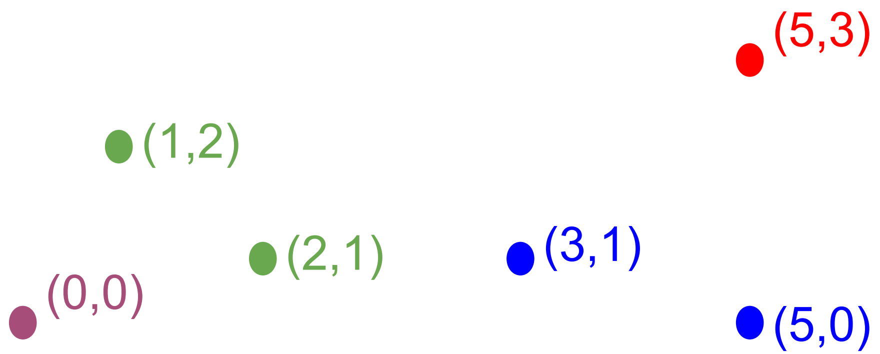
- Build hierarchies of clusters
  - Based on classification trees (next lesson)

<u>Benefits:</u>

- No fixed number of clusters
- Dendrogram displays multiple granularities of clustering
- Multiple distance metric options
- Captures non-spherical clusters

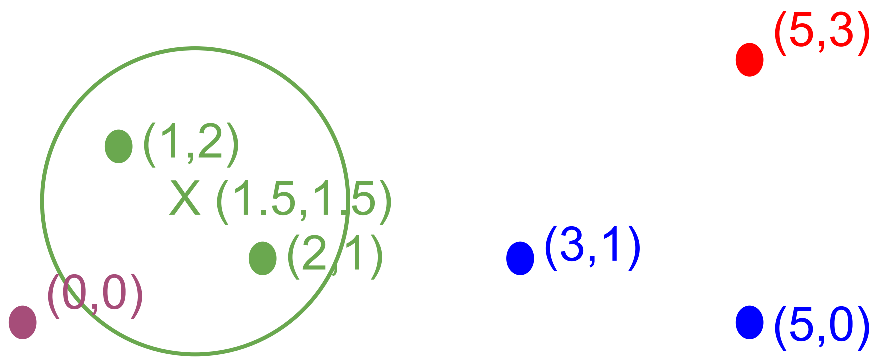


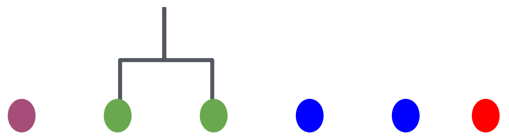
HIERARCHICAL CLUSTERING



Dendrogram

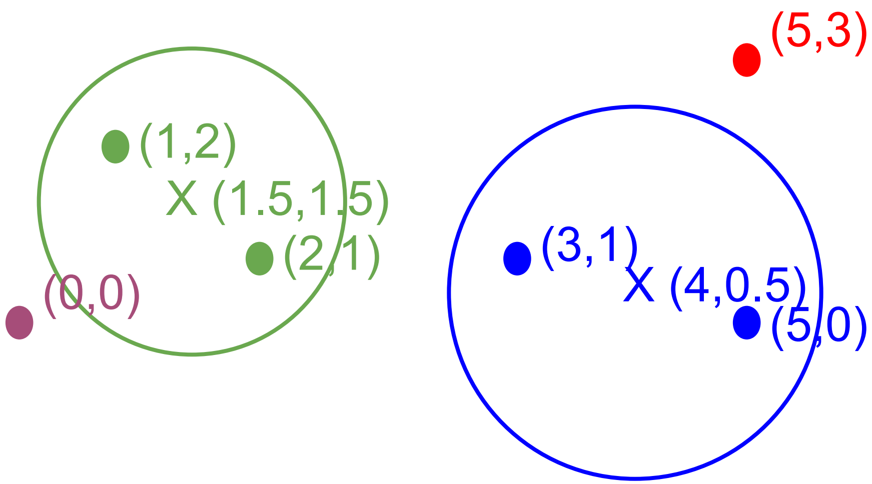
HIERARCHICAL CLUSTERING

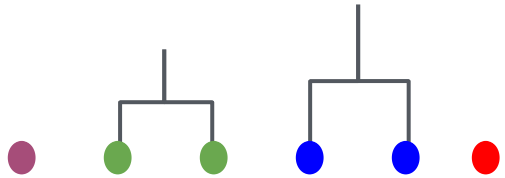




Dendrogram

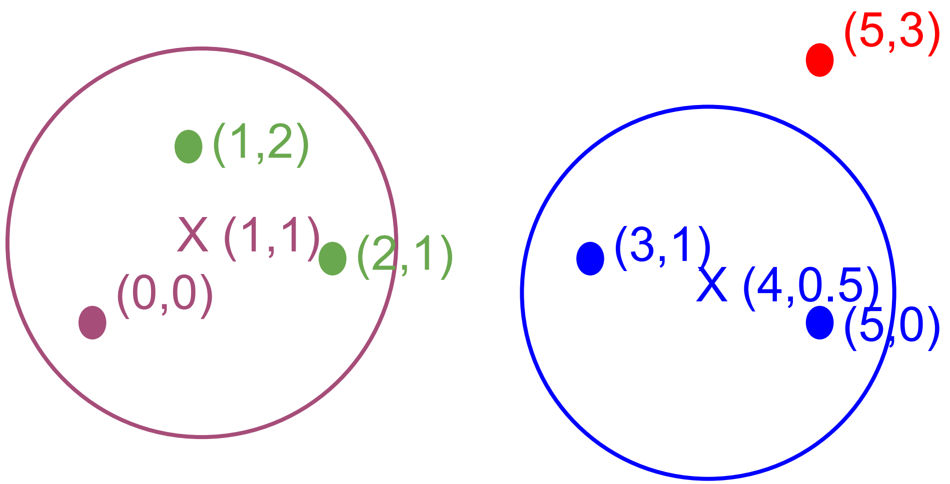
HIERARCHICAL CLUSTERING

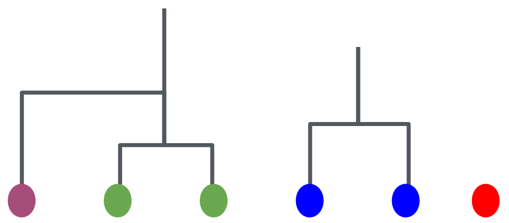




Dendrogram

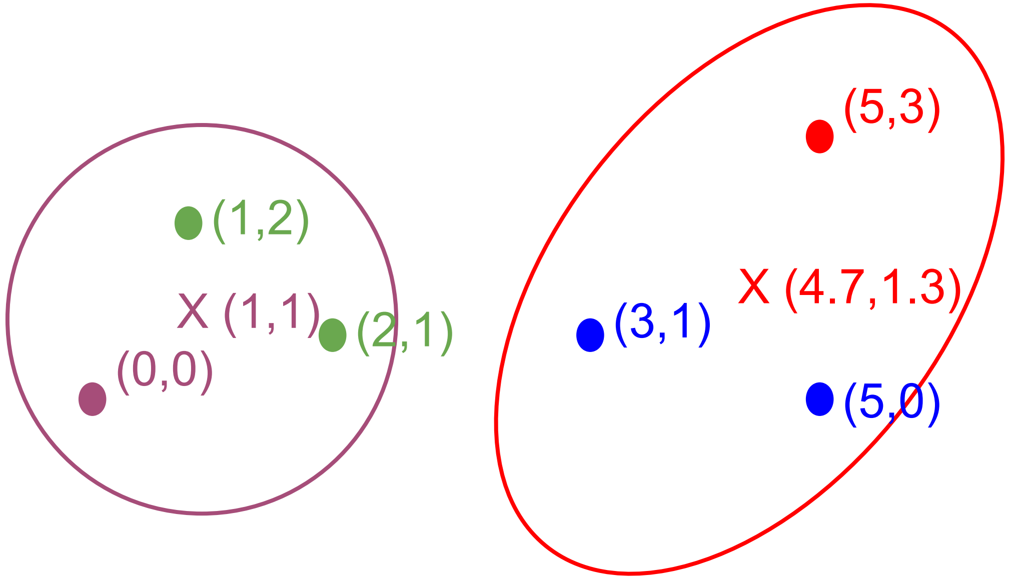
HIERARCHICAL CLUSTERING

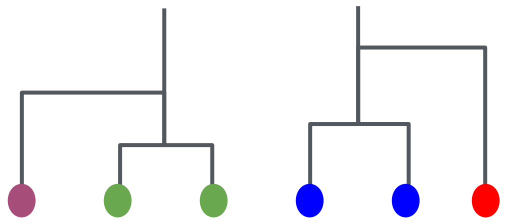




Dendrogram

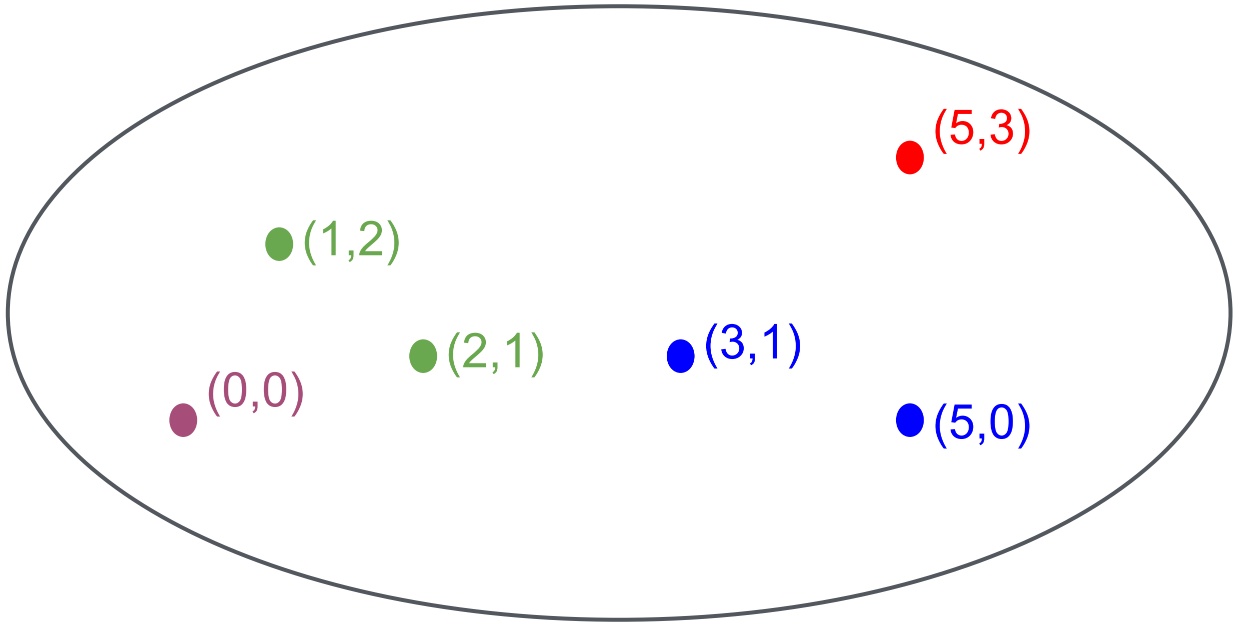
HIERARCHICAL CLUSTERING

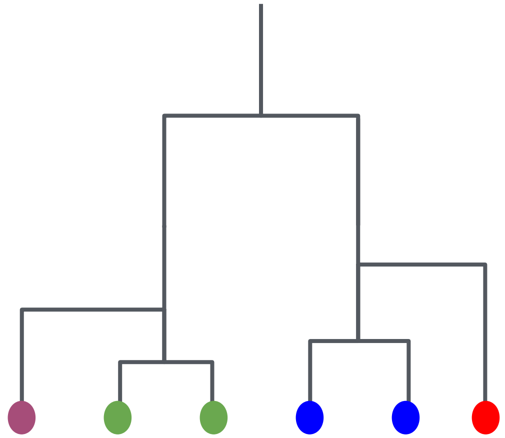




Dendrogram

HIERARCHICAL CLUSTERING





Dendrogram

HIERARCHICAL CLUSTERING

We’ll discuss hierarchical models more once we cover decision trees

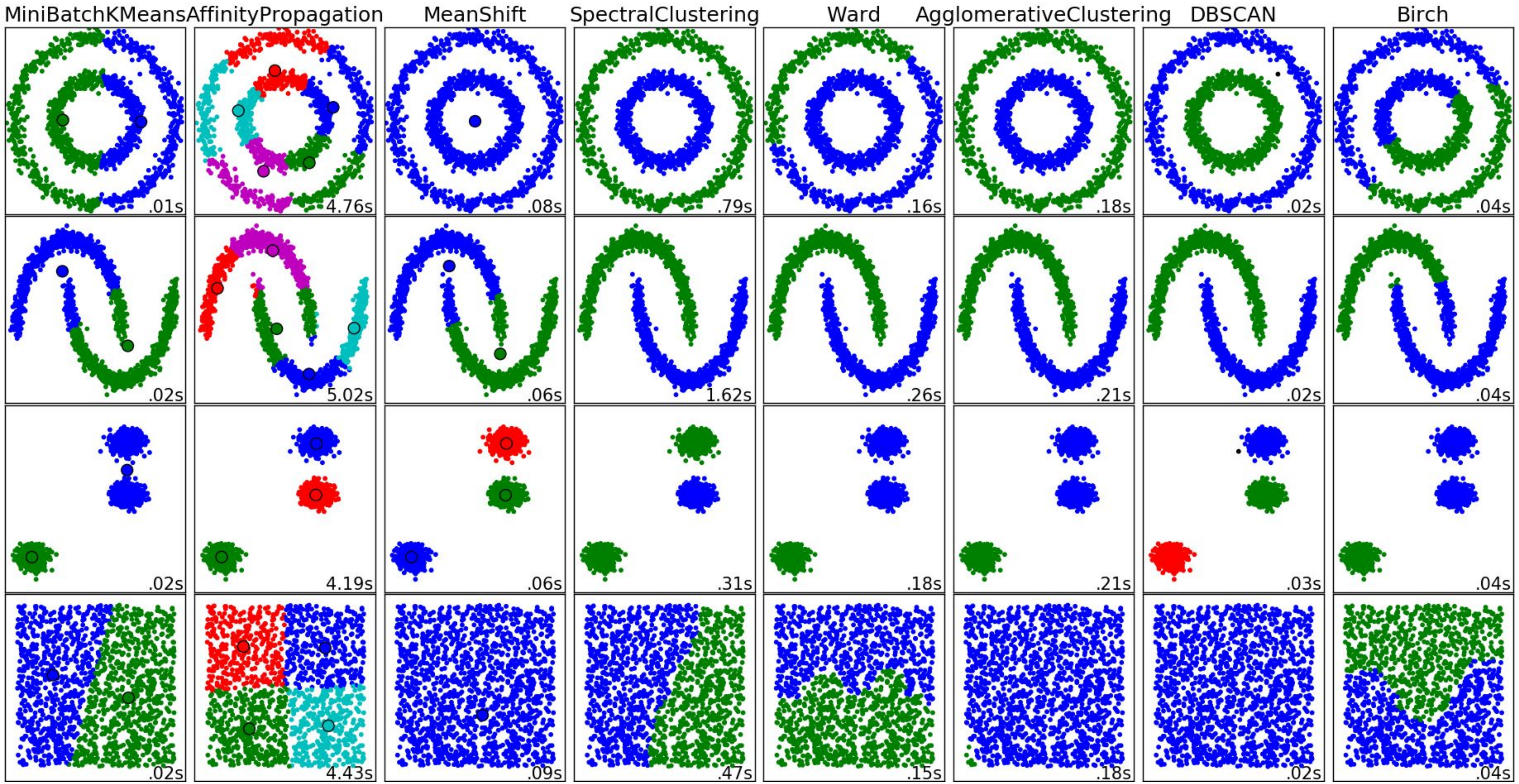
- For now we can fit with sklearn

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

We’ll try it out in the starter-code

CLUSTERING OVERVIEW

- There are [many](#) [clustering algorithms](#)



GUIDED PRACTICE

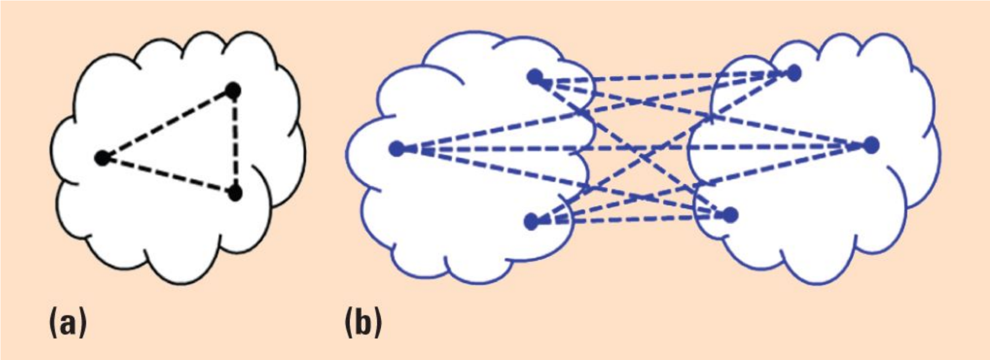
CLUSTERING METRICS

CLUSTERING METRICS

- As usual we need a metric to evaluate *model fit*

CLUSTERING METRICS

- 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

```
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')
```

CLUSTERING METRICS

- There are also a number of other metrics based on:
    - Mutual Information
    - Homogeneity
    - Adjusted Rand Index

INDEPENDENT PRACTICE

CLUSTERING, CLASSIFICATION, AND REGRESSION

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

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 (i.e. unsupervised prediction)

- 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

BEFORE NEXT CLASS

UPCOMING

- **Unit Project 4** and **Project Proposal*** - Due Thurs!

  * just send me the info for your Final Project selection via ***Slack***

    - i.e. which project you decided to go with

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET