Lecture 19
Hierarchical Clustering

Dennis Sun
Stanford University
DATASCI / STATS 112

February 27, 2023

1. Review
2. Hierarchical Clustering: Conceptual
3. Hierarchical Clustering: Coding
4. About Exams, Project, and Grading

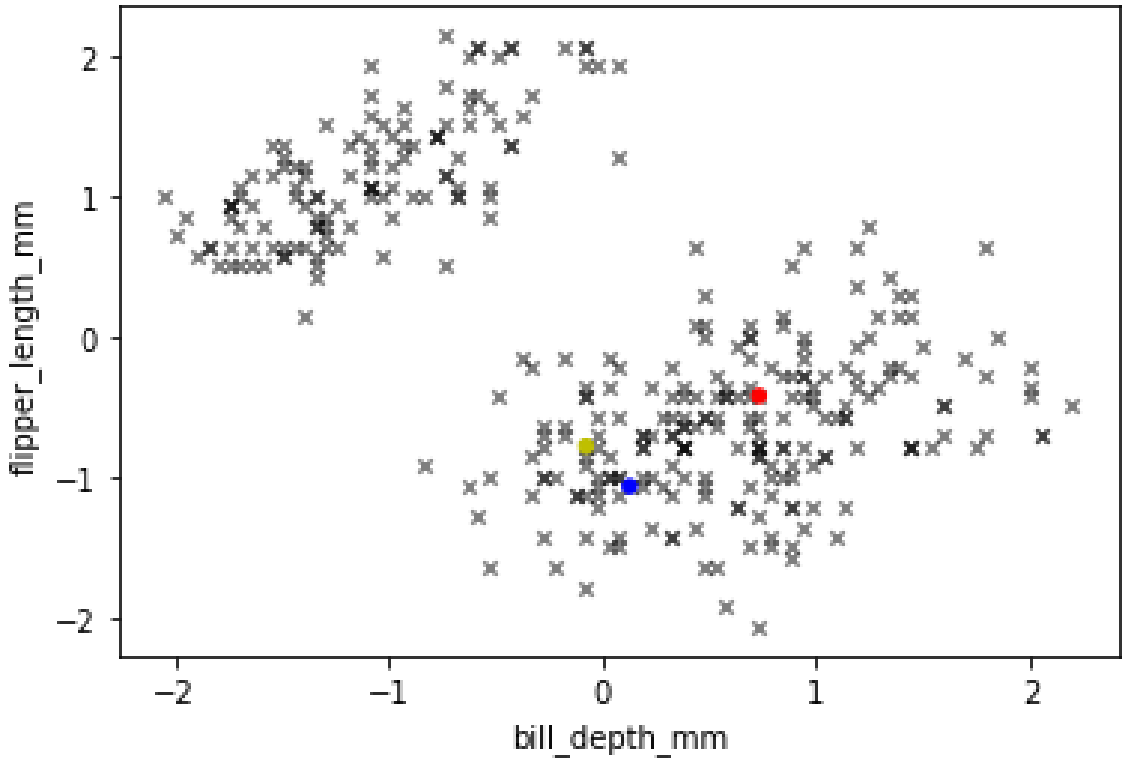
1. Review
2. Hierarchical Clustering: Conceptual
3. Hierarchical Clustering: Coding
4. About Exams, Project, and Grading

K-Means Clustering

Last time, we learned about **unsupervised learning**.

One important type of unsupervised learning is **clustering**.

One algorithm for finding clusters is **k-means**, which finds the centroids of the clusters.



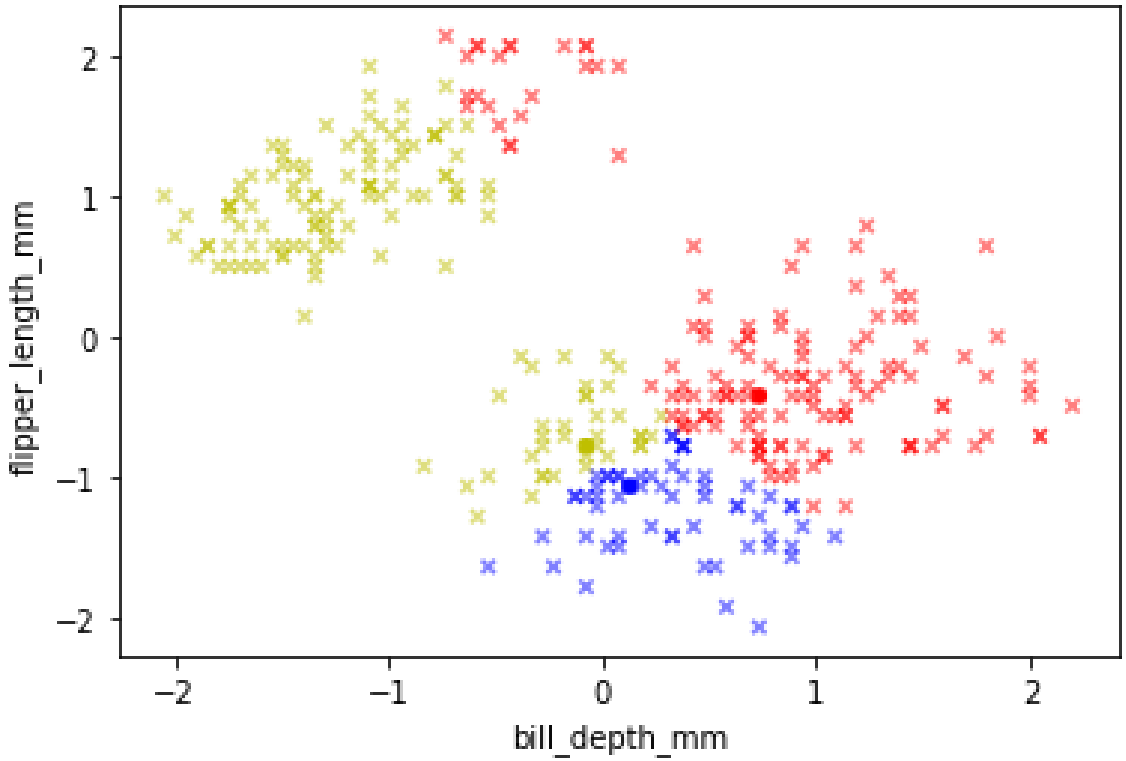
Initialize centroids at random.

K-Means Clustering

Last time, we learned about **unsupervised learning**.

One important type of unsupervised learning is **clustering**.

One algorithm for finding clusters is **k-means**, which finds the centroids of the clusters.



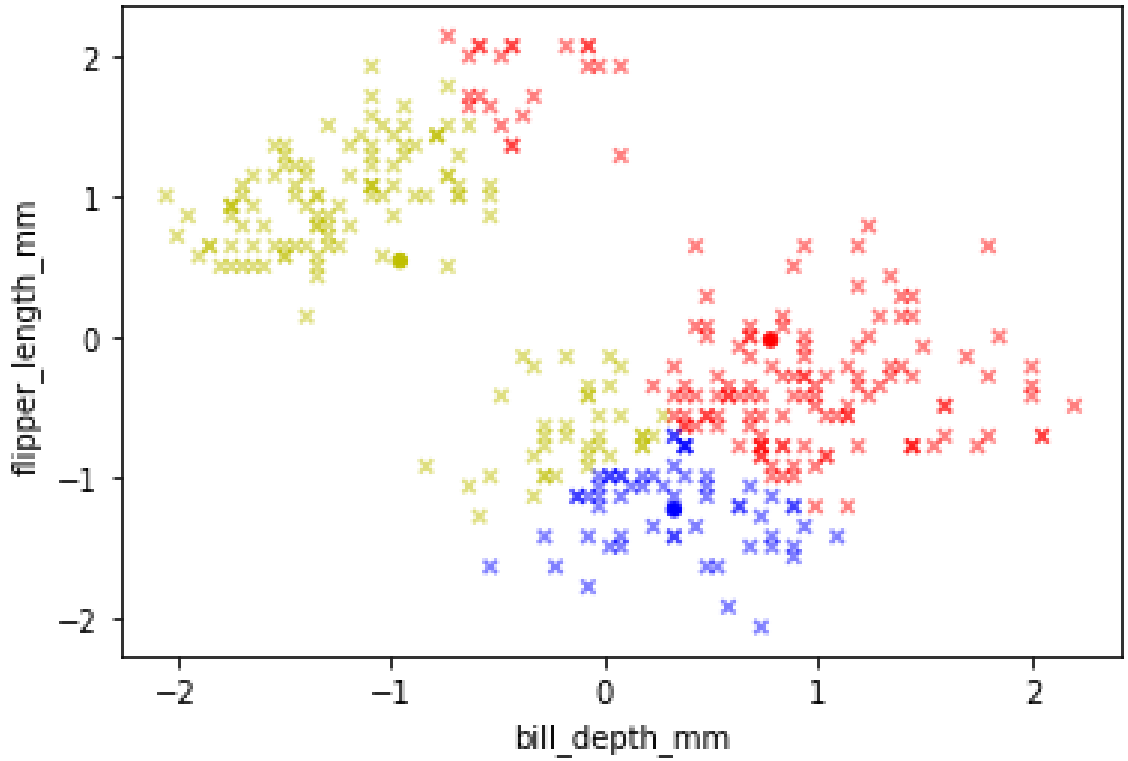
Assign observations to the cluster of the nearest centroid.

K-Means Clustering

Last time, we learned about **unsupervised learning**.

One important type of unsupervised learning is **clustering**.

One algorithm for finding clusters is **k-means**, which finds the centroids of the clusters.



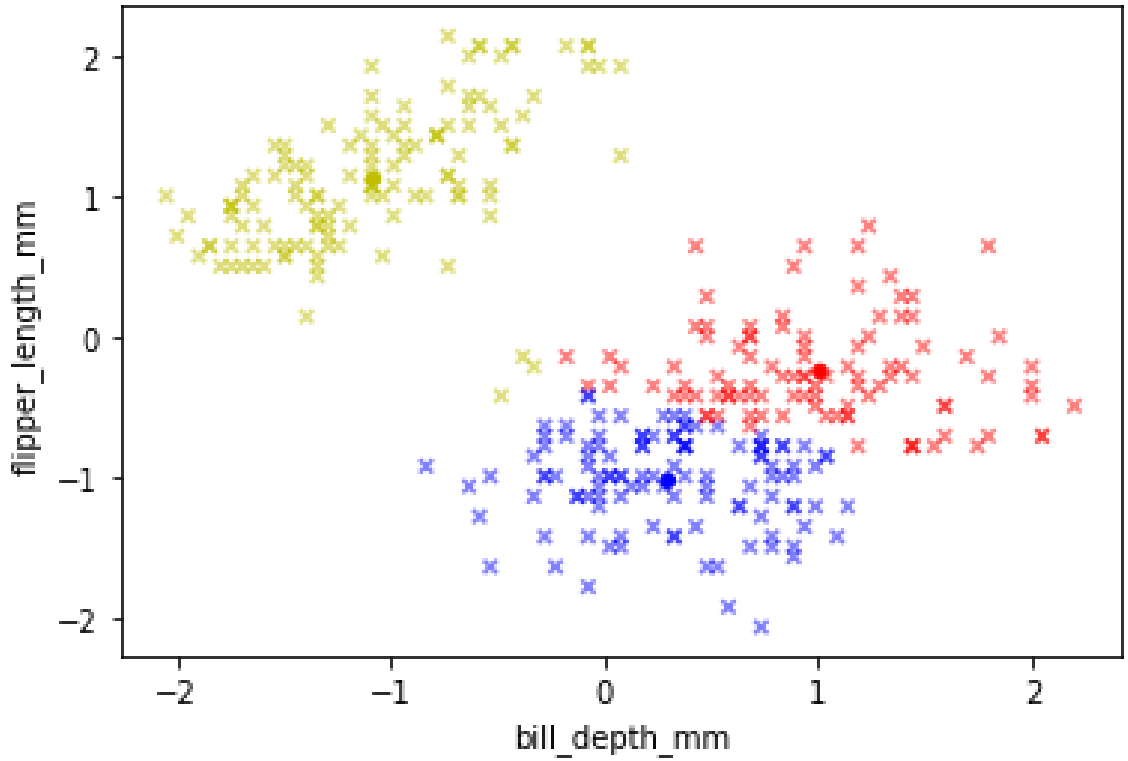
Recalculate centroids based on cluster assignments.

K-Means Clustering

Last time, we learned about **unsupervised learning**.

One important type of unsupervised learning is **clustering**.

One algorithm for finding clusters is k**-means**, which finds the centroids of the clusters.



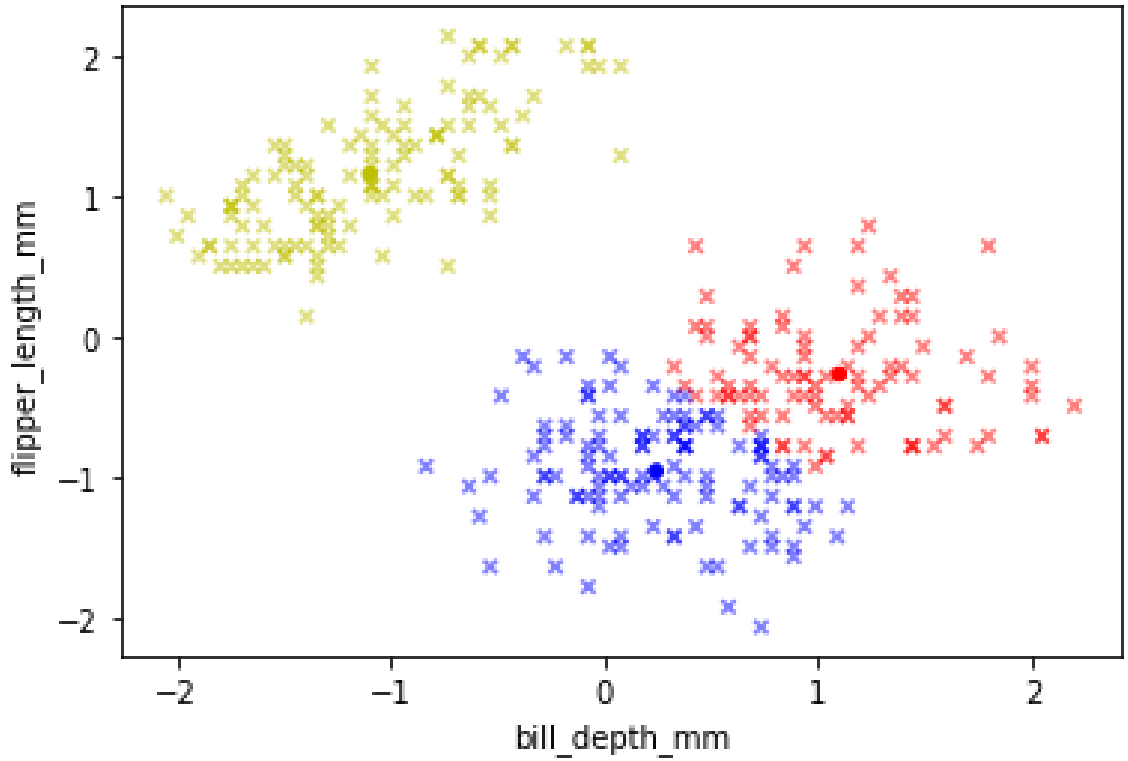
Repeat this process...

K-Means Clustering

Last time, we learned about **unsupervised learning**.

One important type of unsupervised learning is **clustering**.

One algorithm for finding clusters is **k-means**, which finds the centroids of the clusters.



...until the cluster assignments stop changing.

1 Review

2 Hierarchical Clustering: Conceptual

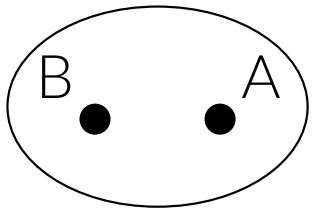
3 Hierarchical Clustering: Coding

4 About Exams, Project, and Grading

Hierarchical Clustering

Hierarchical clustering is a clustering algorithm based on distances between observations (not distances from centroids).

Suppose we have this (toy) data set, consisting of 5 observations.

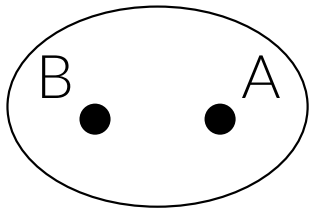


First, we merge the two closest points into a cluster.

Hierarchical Clustering

Hierarchical clustering is a clustering algorithm based on distances between observations (not distances from centroids).

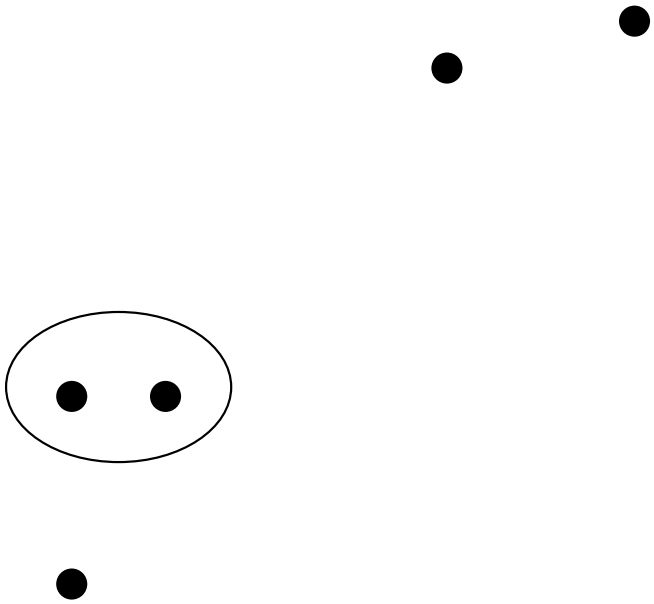
Suppose we have this (toy) data set, consisting of 5 observations.



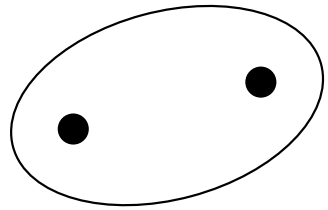
Now, we want to merge the next closest into a cluster.

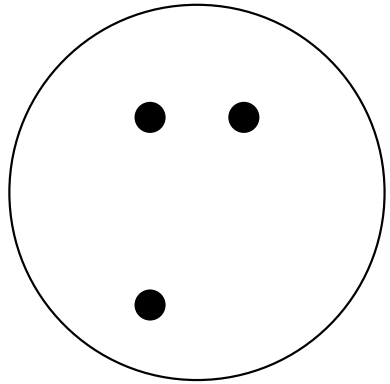
But there's a problem!

How do we measure distance between a cluster and a point?



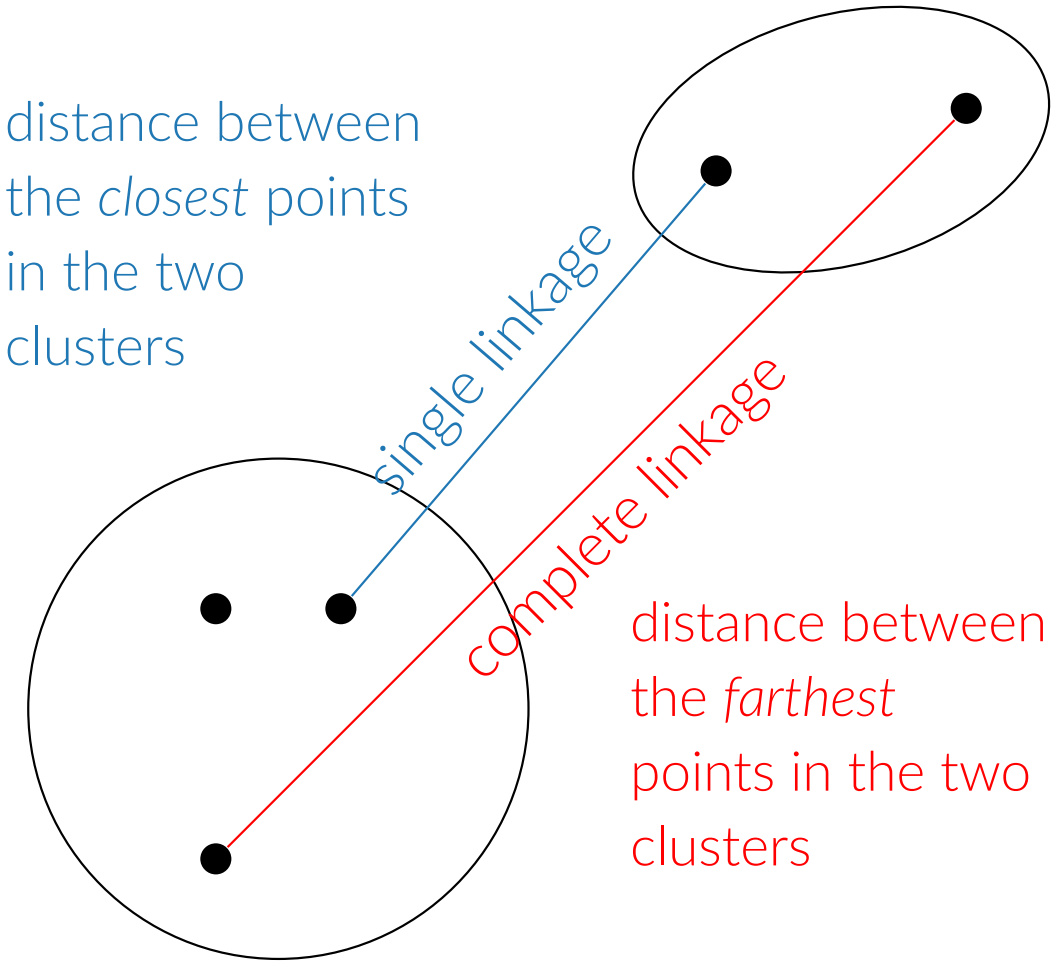
More generally, how do we measure distance between two clusters?





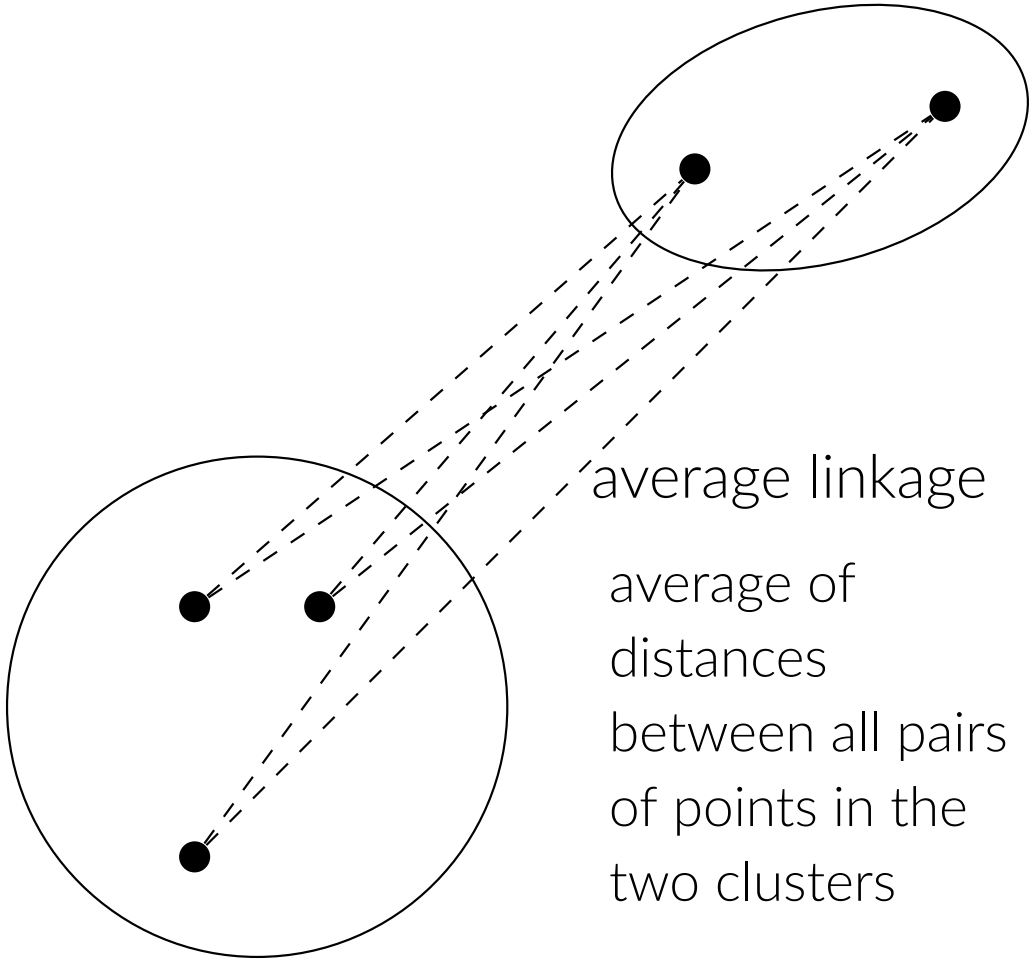
Measuring Distances between Clusters

The choice of how to measure distances between clusters is called the **linkage**.



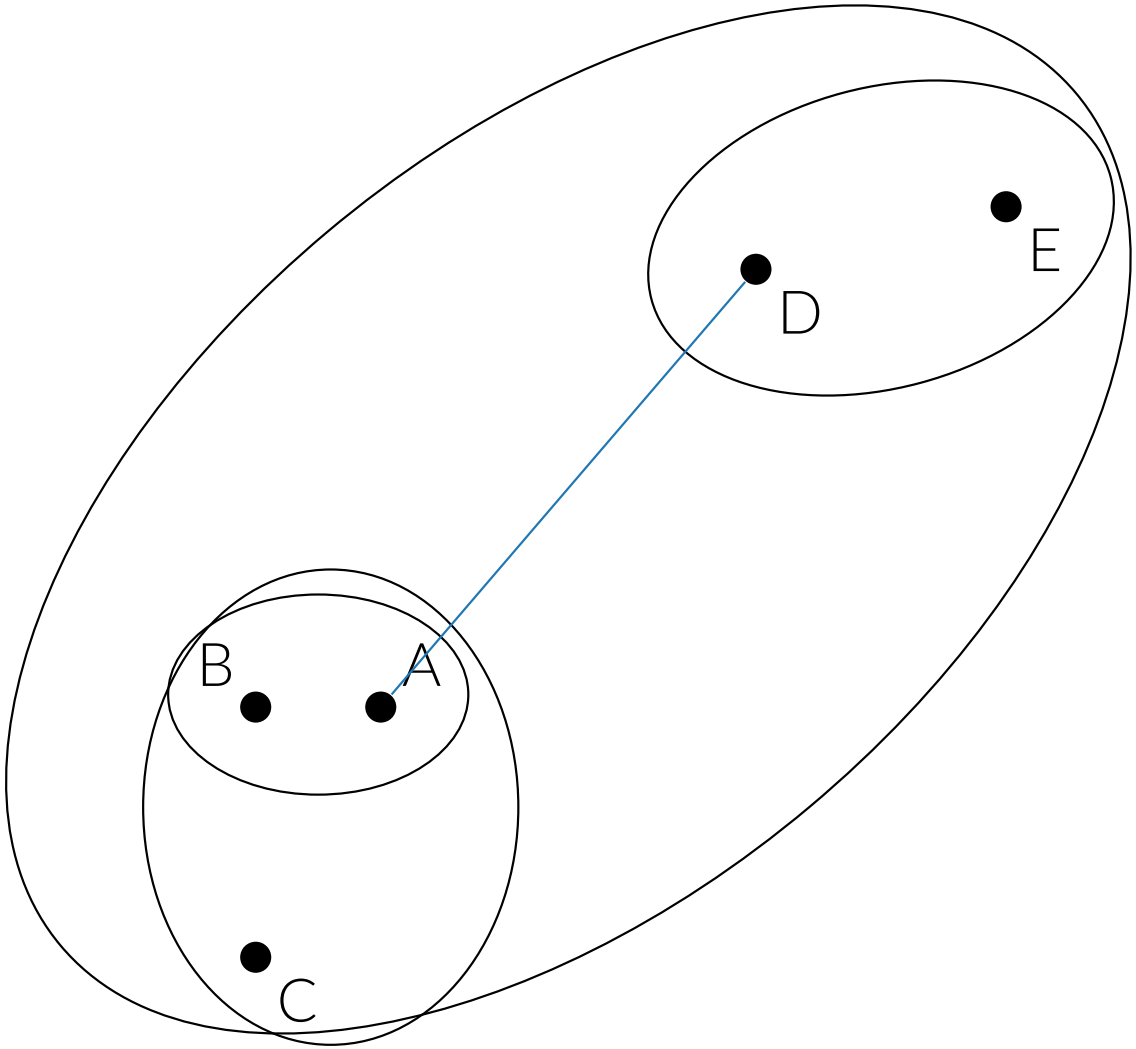
Measuring Distances between Clusters

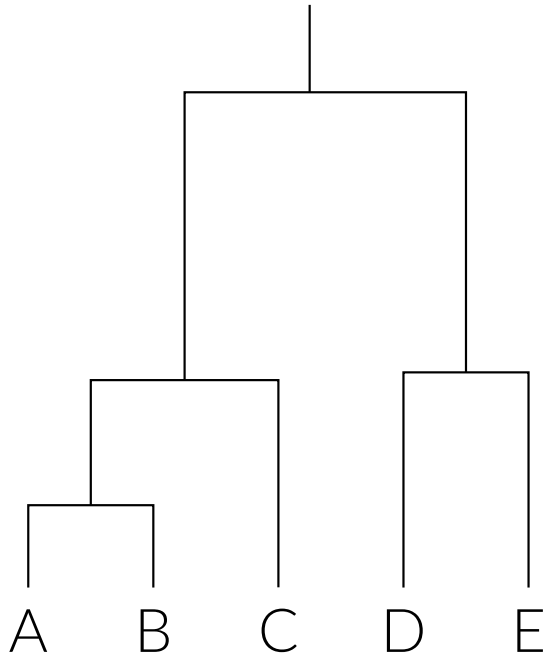
The choice of how to measure distances between clusters is called the **linkage**.



Hierarchical Clustering

Let's use hierarchical clustering on the data set using *single linkage*.

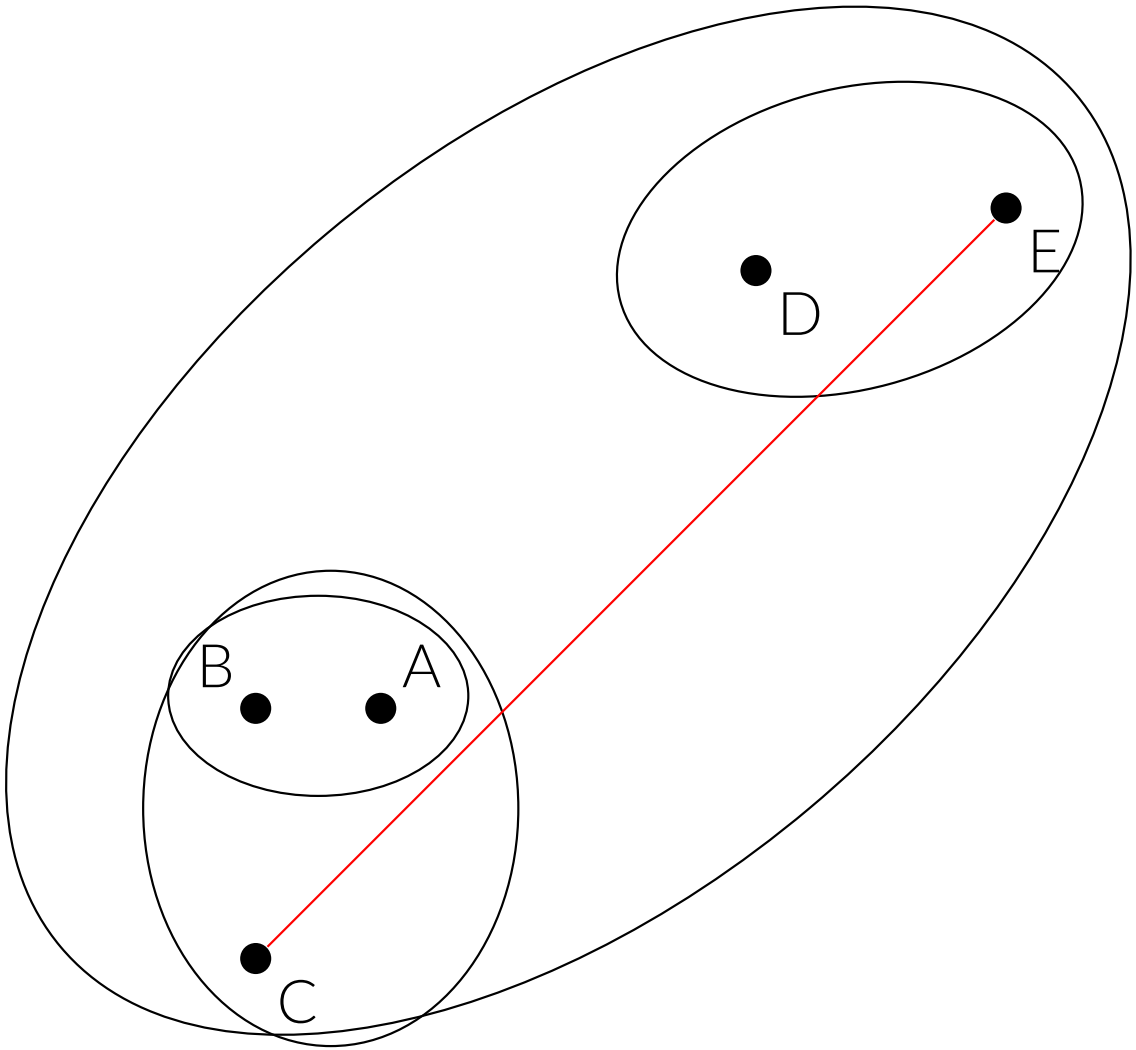


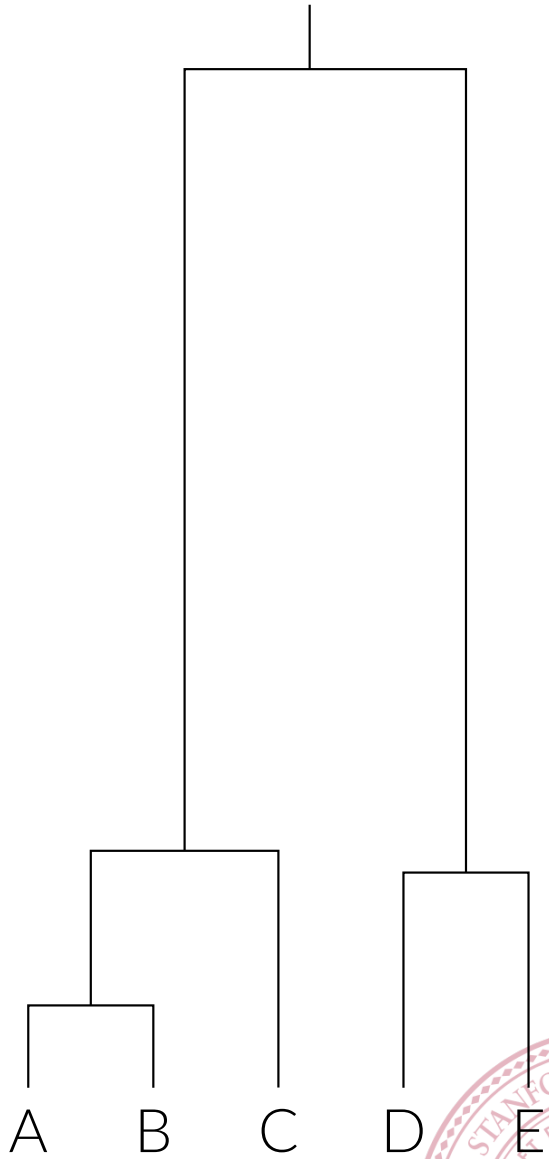


The results can be represented using a **dendrogram**.

Hierarchical Clustering

What if we instead use *complete linkage*?



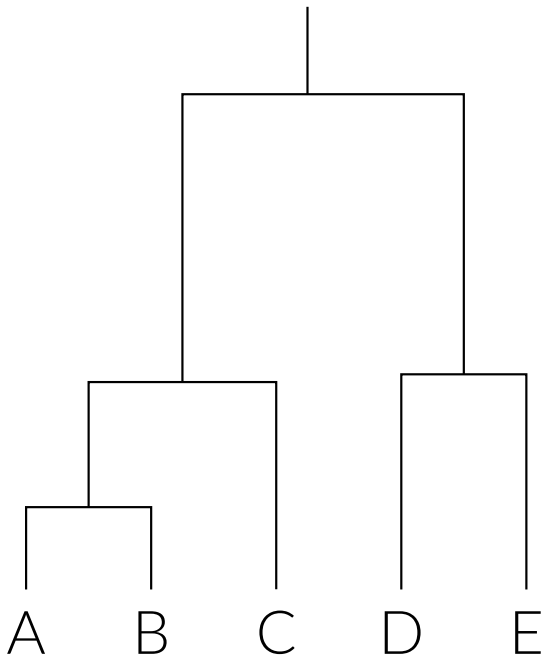


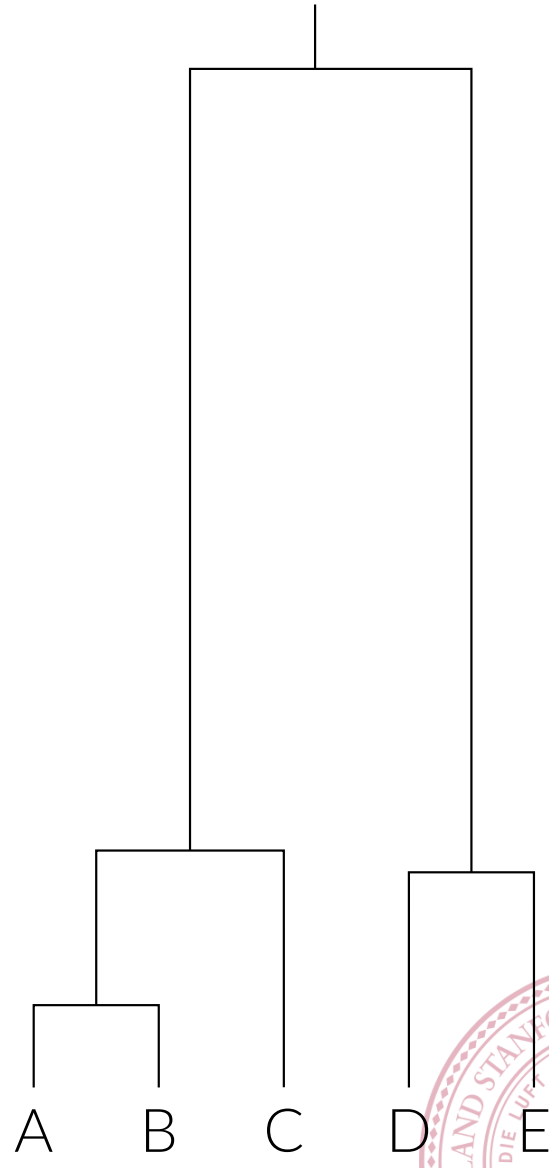
Comparing Linkages

single linkage

Notice that the clustering is different depending on the linkage you choose!

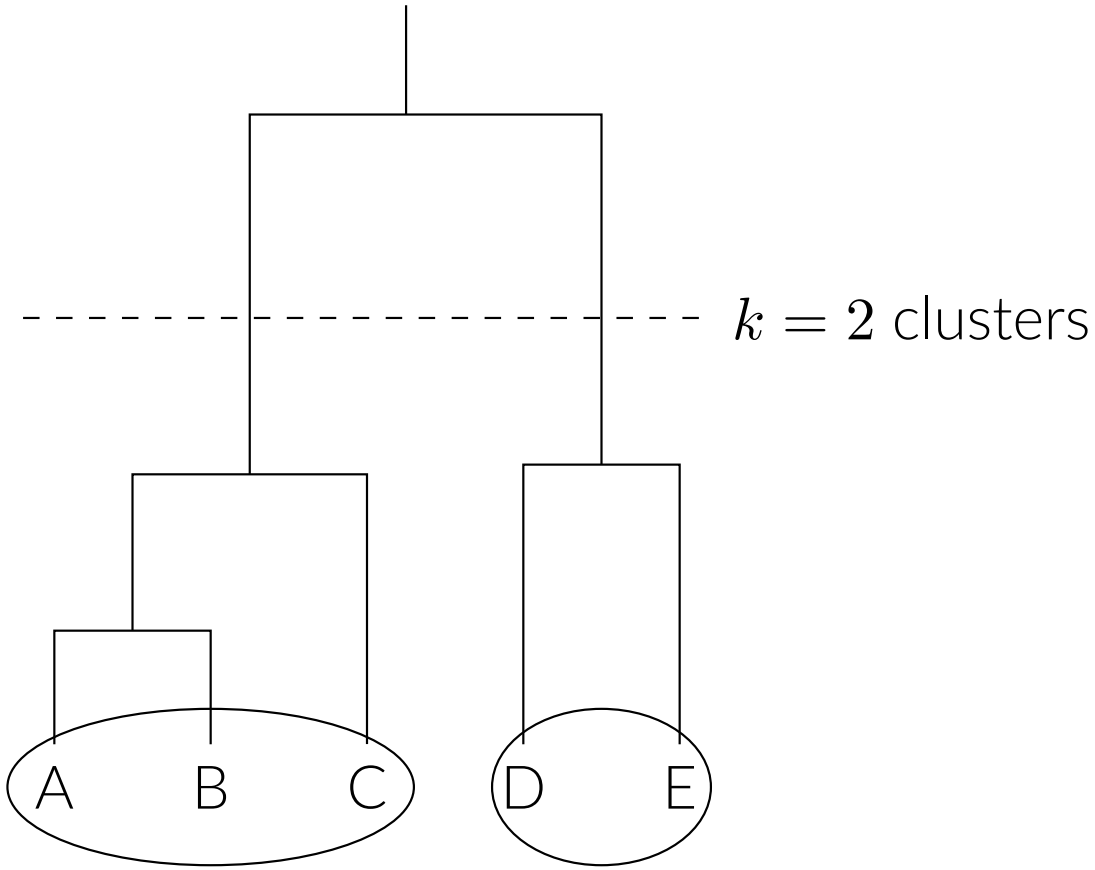
complete linkage





Finding k Clusters

To get k clusters, we "cut" the dendrogram at some height.



1 Review

2 Hierarchical Clustering: Conceptual

3 Hierarchical Clustering: Coding

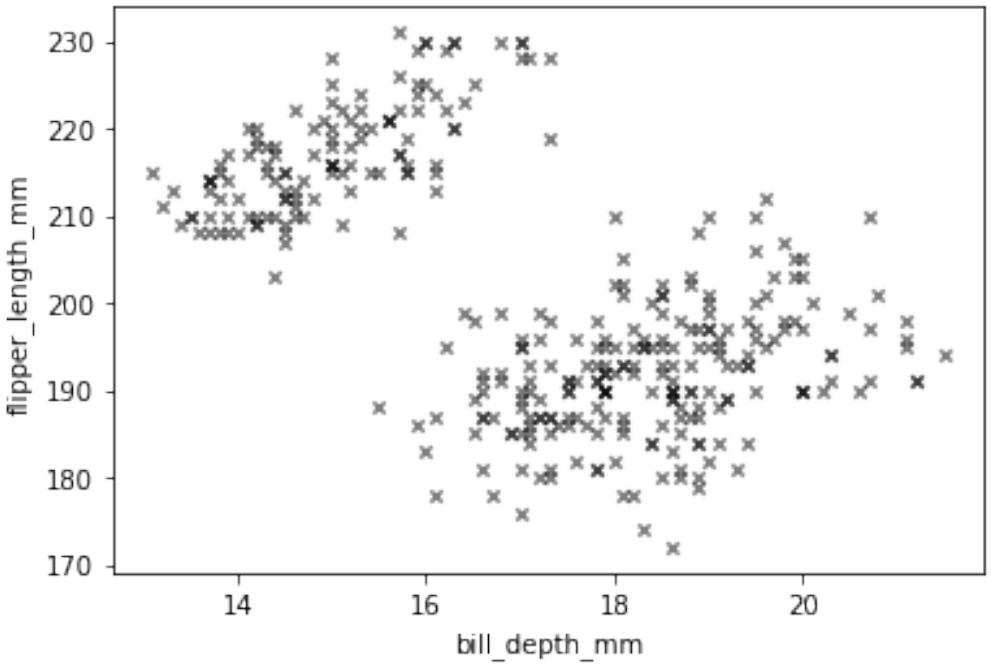
4 About Exams, Project, and Grading

Hierarchical Clustering in Scikit-Learn

```
import pandas as pd

data_dir = "https://dlsun.github.io/stats112/data/"
df_penguins = pd.read_csv(data_dir + "penguins.csv")

X_train = df_penguins[["bill_depth_mm", "flipper_length_mm"]].dropna()
X_train.plot.scatter(x="bill_depth_mm", y="flipper_length_mm",
                     c="black", marker="x", alpha=.5)
```



Hierarchical Clustering in Scikit-Learn

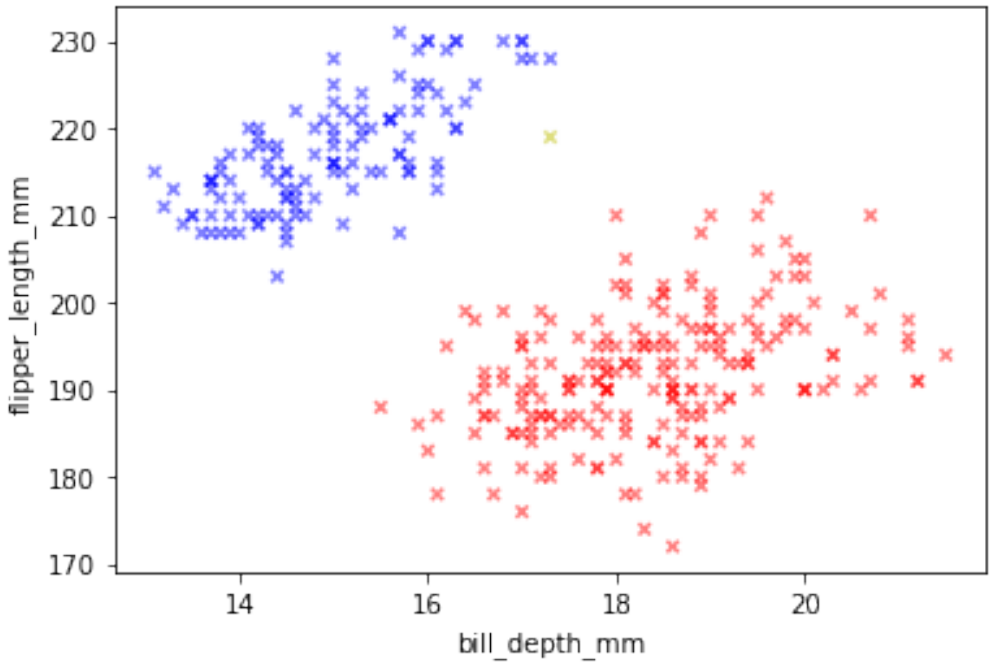
```
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import AgglomerativeClustering

model = AgglomerativeClustering(n_clusters=3, linkage="single")
pipeline = make_pipeline(
    StandardScaler(),
    model)
pipeline.fit(X_train)
clusters = model.labels_
clusters
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

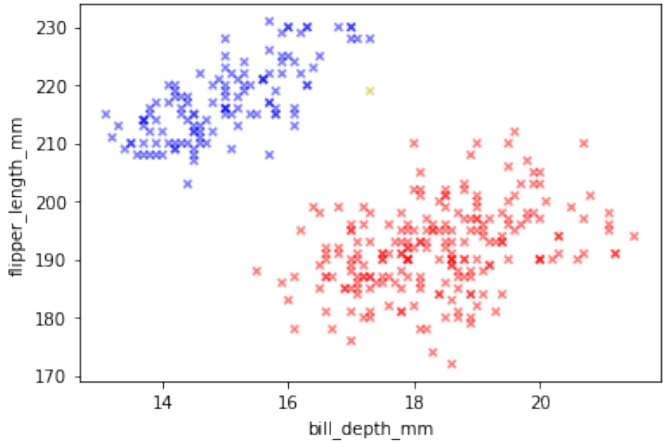
Hierarchical Clustering in Scikit-Learn

```
clusters = pd.Series(clusters).map({
    0: "r",
    1: "b",
    2: "y"
})
X_train.plot.scatter(x="bill_depth_mm", y="flipper_length_mm",
                     c=clusters, marker="x", alpha=.5)
```

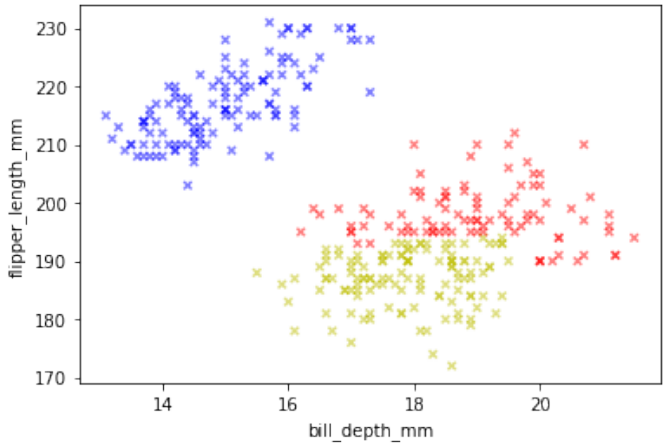


Comparison of Linkages

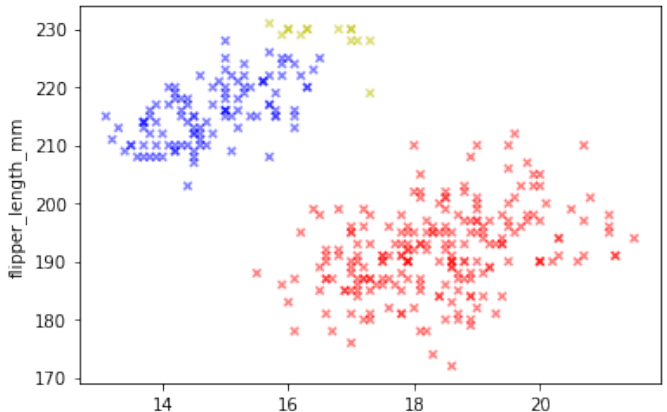
```
AgglomerativeClustering(
    n_clusters=3,
    linkage="single")
```




```
AgglomerativeClustering(
    n_clusters=3,
    linkage="complete")
```




```
AgglomerativeClustering(
    n_clusters=3,
    linkage="average")
```



Evaluating Clustering Models

There are many choices to make in clustering:

- k-means vs. hierarchical
- metric
- number of clusters k
- linkage (for hierarchical clustering)

How do we know which choice is best?

- There is no analog of cross-validation for clustering.
- The determination of whether a clustering is "good" is subjective.

The Colab for section tomorrow will give you more experience with evaluating clustering models (and understanding the different algorithms).

1 Review

2 Hierarchical Clustering: Conceptual

3 Hierarchical Clustering: Coding

4 About Exams, Project, and Grading

About Grading

- I don't decide letter grades until the end of the quarter.
- But among students who
  - complete the assignments,
  - earn full participation in section,
  - and submit an acceptable final project,

  at least half will earn an A-range grade.
- On the flip side, I don't hesitate to give Ds and Fs to students who don't do homework or submit a sloppy final project.

Homework Reminders

- Don't forget to do the Colab ("Clustering in Practice and in Theory") for section tomorrow.
- Assignment 5 due Tuesday, Assignment 6 due Friday.
- **No extensions** because we need to post solutions (so that you can study for your exam).
- They are short, so start early and come to office hours.

Exam 2 Reminders

- Exam 2 is in class next Monday. Same policy as last time (1 page of handwritten notes allowed).
- The exam covers material up to today.
- I have posted a practice exam. Solutions will be posted later in the week.
- We have also posted solutions to all the assignments and will post solutions to Assignments 5 and 6 before the exam.

Project Reminders

- Sign up for a final project presentation here: [link to form].
- The final project files are due on Canvas on Wednesday 3/22 at 11:59 PM.
- Look at the rubric and example projects I've posted.
- If you haven't started collecting data yet, it's getting very late!