CCSS Workshop Series

Unsupervised Learning Workshop

ANGEL HSING-CHI HWANG

HH695@CORNELL.EDU

CORNELL CENTER for SOCIAL SCIENCES

CCSS Research Support Code of Conduct

The Cornell Center for Social Sciences provides a welcoming environment for everyone embracing all backgrounds or identities. All instructors and attendees agree to abide by our community norms. We encourage the following behaviors in our workshops:

- Respect differing viewpoints and ideas
- Share your own perspectives and ask any questions
- Accept constructive criticism
- Use welcoming and inclusive language
- Show courtesy and respect for all instructors and attendees

If you believe that an instructor or attendee has violated the code of conduct, please report the violation to CCSS-ResearchSupport@cornell.edu. We take all reported incidents seriously.

Land Acknowledgement

Cornell University is located on the traditional homelands of the Gayogo̱hó:nǫ' (the Cayuga Nation). The Gayogo̱hó:nǫ' are members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic and contemporary presence on this land. The Confederacy precedes the establishment of Cornell University, New York state, and the United States of America. We acknowledge the painful history of Gayogo̱hó:nǫ' dispossession, and honor the ongoing connection of Gayogo̱hó:nǫ' people, past and present, to these lands and waters.

Here are additional links for more on the history of Cornell's violent, colonial formation, the movement to return native lands, and about the AIISP program at Cornell.

Consider donating to the Gayogo̱hó:nǫ' sovereignty initiative here.

CORNELL CENTER for SOCIAL SCIENCES

Today's Agenda

LECTURE

- Supervised vs. Unsupervised Learning
- Intro to Clustering
- K-Means Essentials
- Common Challenges

HANDS-ON WORKSHOP

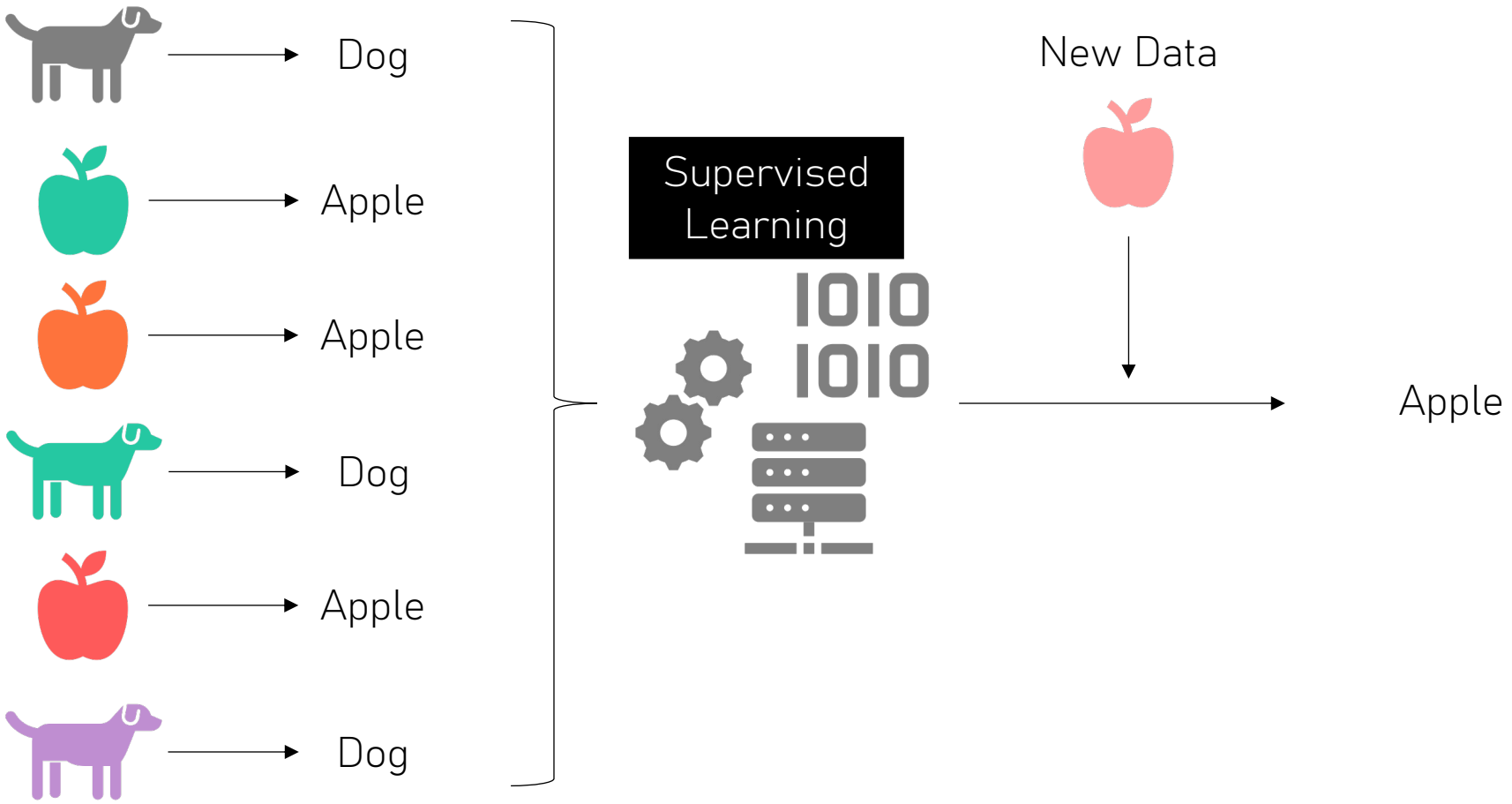
- Mini Case Study: Clustering houses in Boston

Supervised Learning vs. Unsupervised Learning

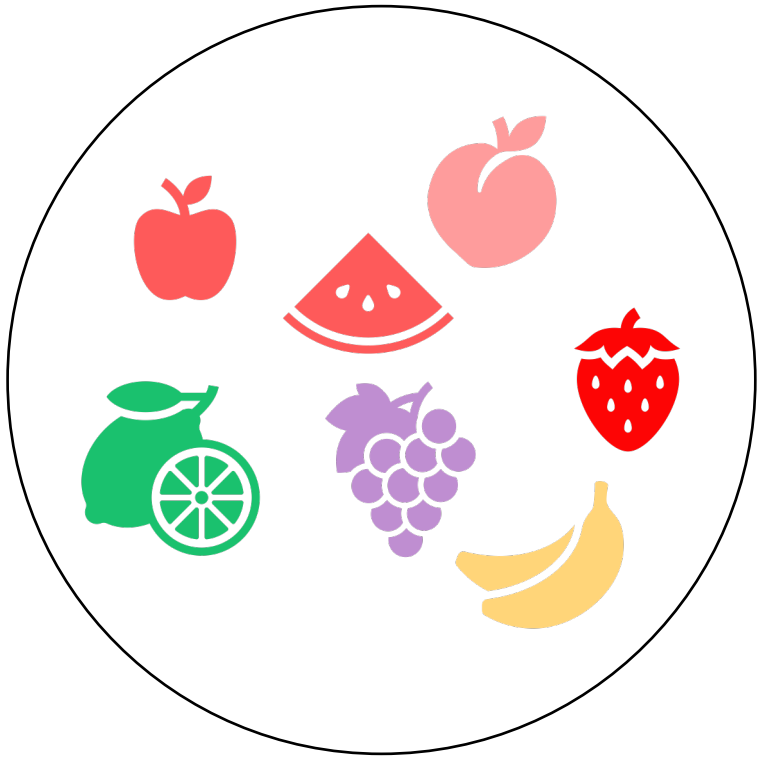
Learning from labeled data

Learning from unlabeled data

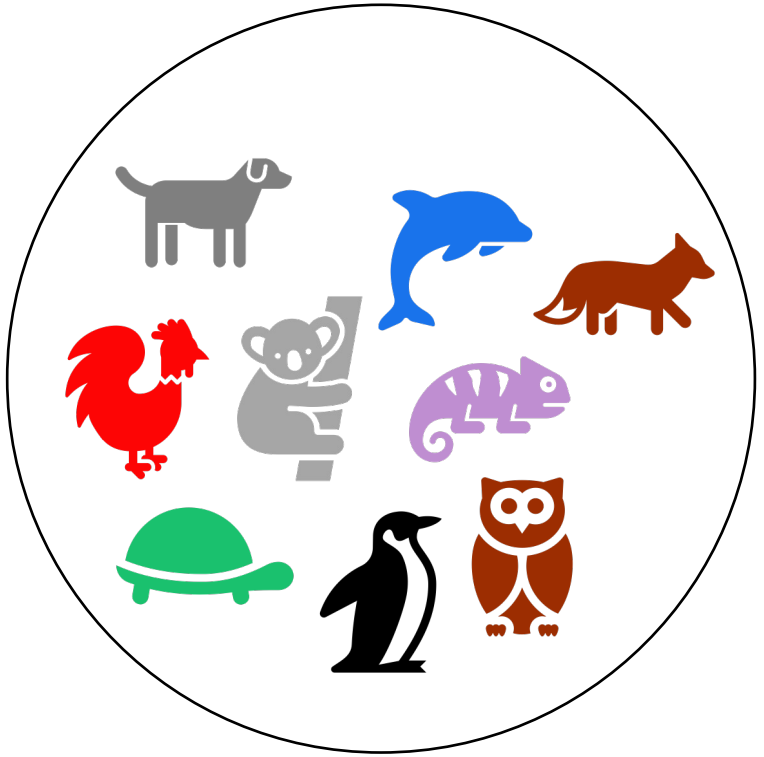
Supervised Learning vs. Unsupervised Learning



Supervised Learning vs. Unsupervised Learning



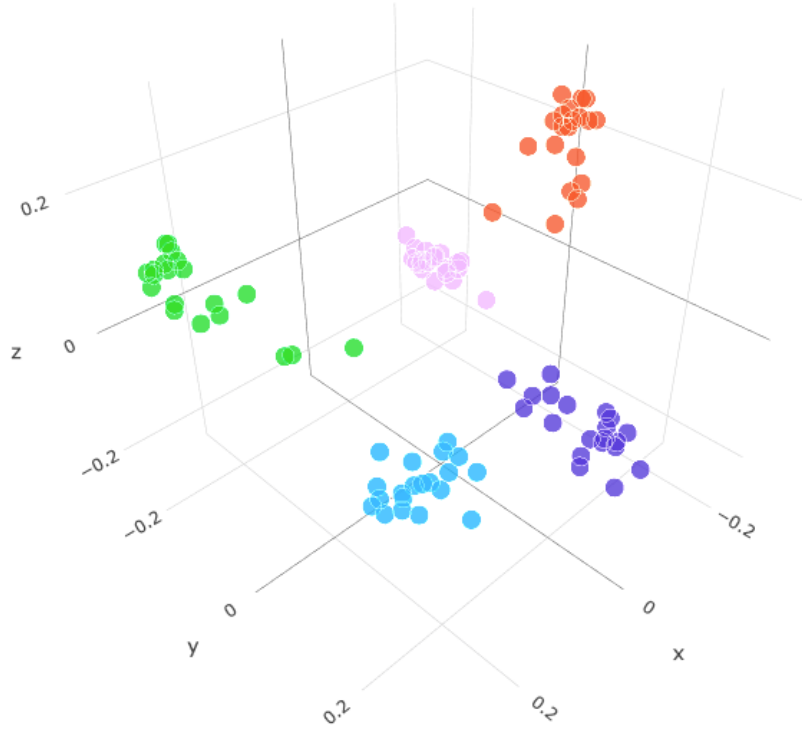
Fruit



Animal

Intro to Clustering

- Data in the real world are often unlabeled
- The goal of unsupervised learning is to find patterns and/or structure in such data
- Clustering is the most commonly used unsupervised learning technique
- Clustering identifies similar "groups" (i.e., clusters) in unlabeled data

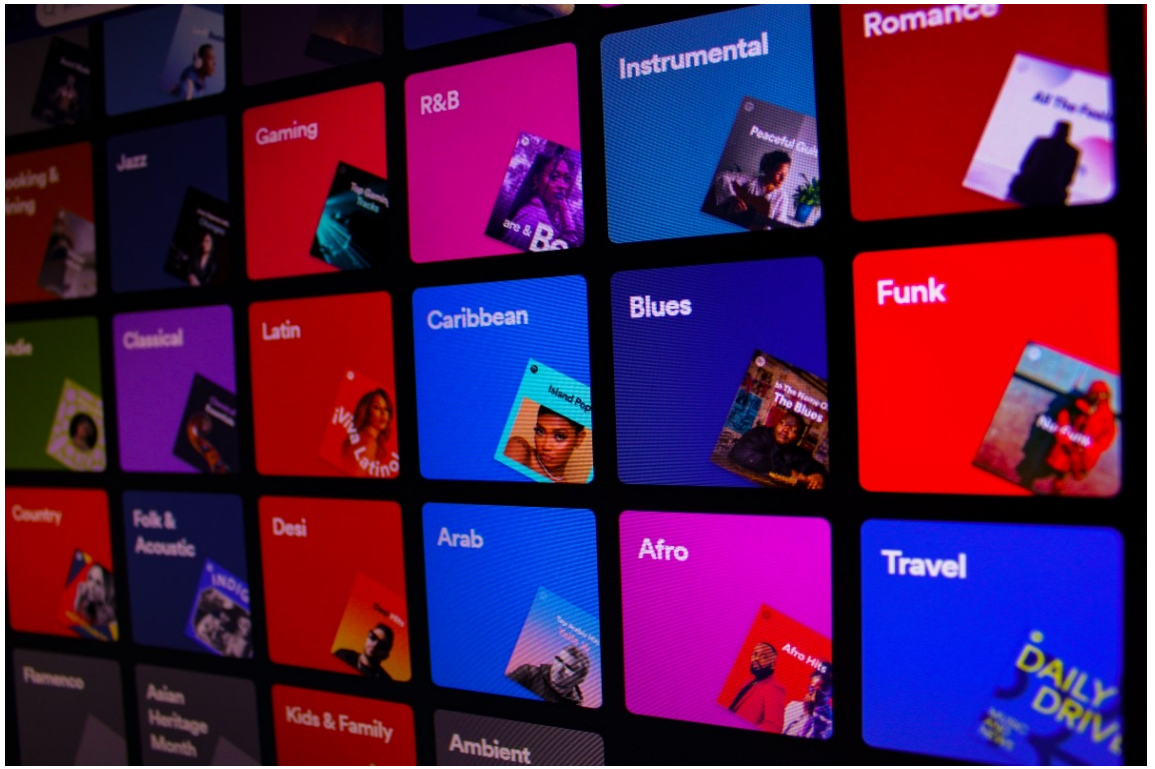


Credit: https://openai.com/blog/introducing-text-and-code-embeddings/

Intro to Clustering

Some applications of clustering…

- Recommendation systems
- Search results
- Market segmentation



Credit: https://medium.com/intro-to-artificial-intelligence/semantic-segmentation-udaitys-self-driving-car-engineer-nanodegree-c01eb6eaf9d

Data Structure

```
      Var. 1   Var. 2   Var. 3   Var. 4
[[  5.1,  3.5,  1.4,  0.2 ... ]
 [  4.9,  3. ,  1.4,  0.2 ... ]
 [  4.7,  3.2,  1.3,  0.2 ... ]
 [  4.6,  3.1,  1.5,  0.2 ... ]
    . . .
 [  5.9,  3. ,  5.1,  1.8 ... ]]
```

Data Structure

Var. 1	Var. 2	Var. 3	Var. 4		Cluster
[[5.1,	3.5,	1.4,	0.2	...]	?
[4.9,	3. ,	1.4,	0.2	...]	?
[4.7,	3.2,	1.3,	0.2	...]	?
[4.6,	3.1,	1.5,	0.2	...]	?
. . .					
[5.9,	3. ,	5.1,	1.8	...]]	?

K-Means Clustering

The goal of clustering is to separate data so that data similar to one another are in the same group, while data different from one another are in different groups. So, two questions arise:

How many groups do we choose?

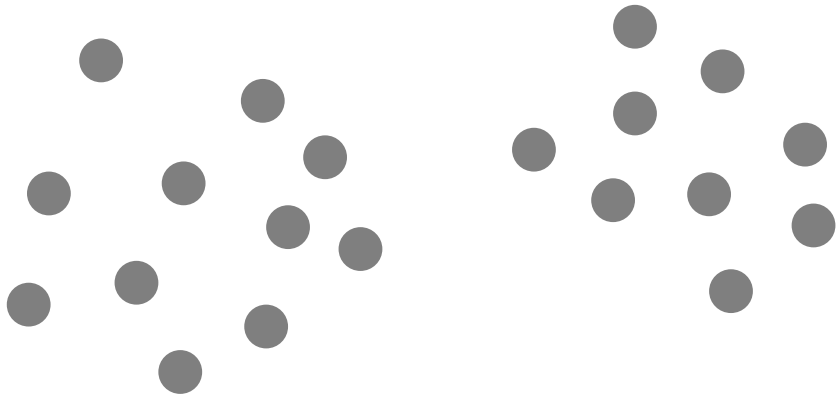
How do we define similarity?

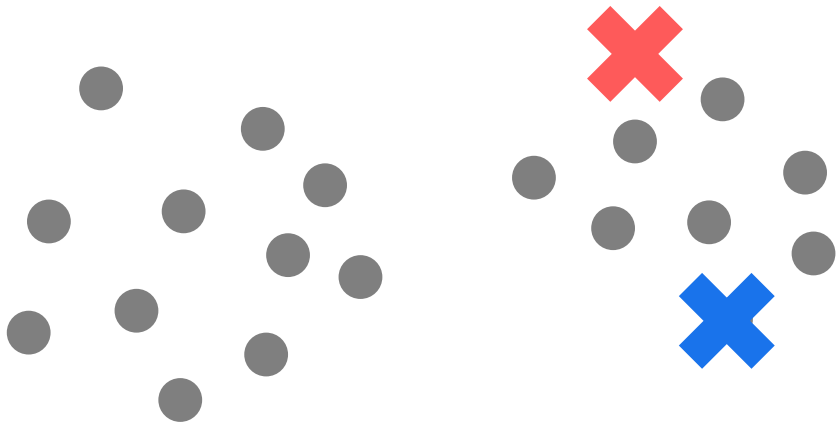
K-Means Clustering

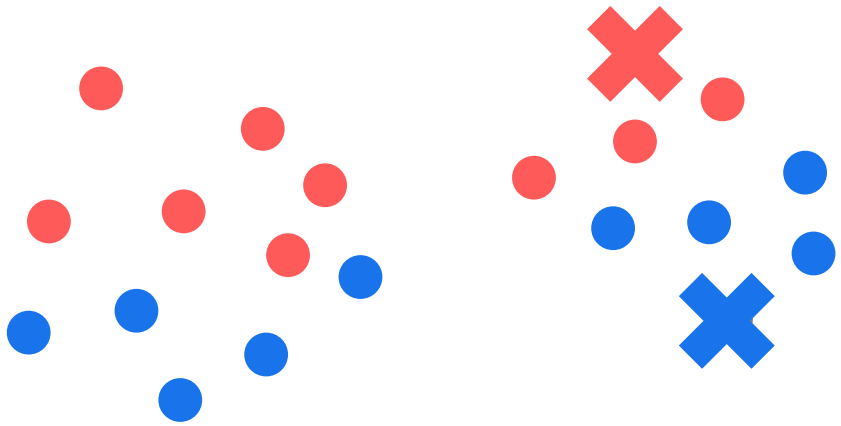
The goal of clustering is to separate data so that data similar to one another are in the same group, while data different from one another are in different groups. So, two questions arise:

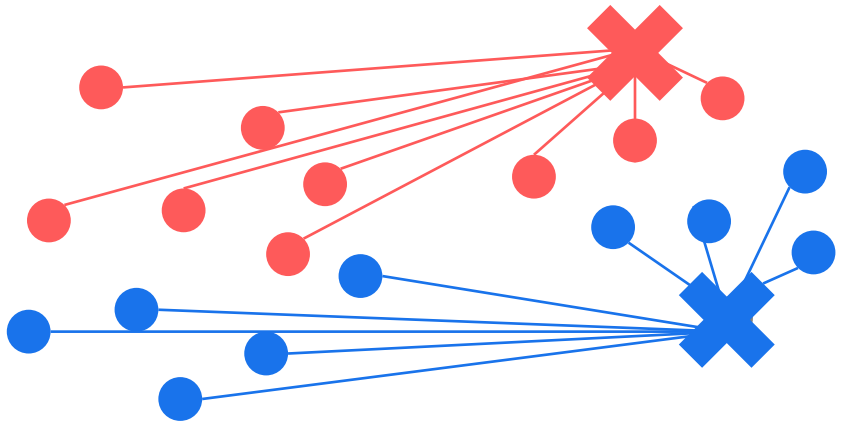
How many groups do we choose? → The "K" refers to the number of clusters (groups) we expect to find in a dataset

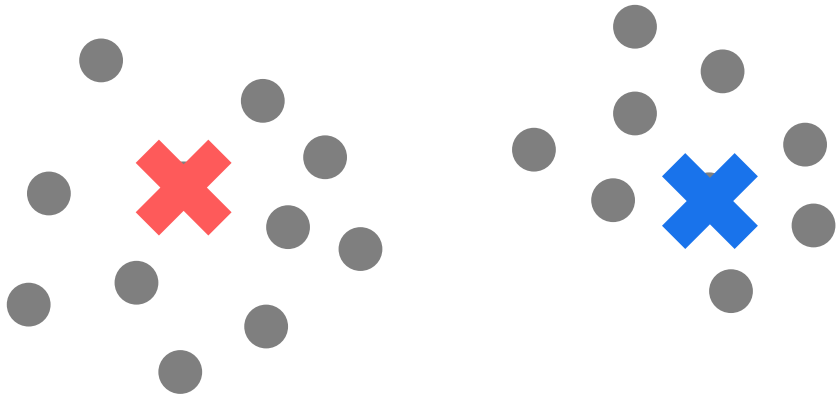
How do we define similarity? → The "Means" refers to the average distance of data to each cluster center, also known as the *centroid*, which we are trying to minimize.

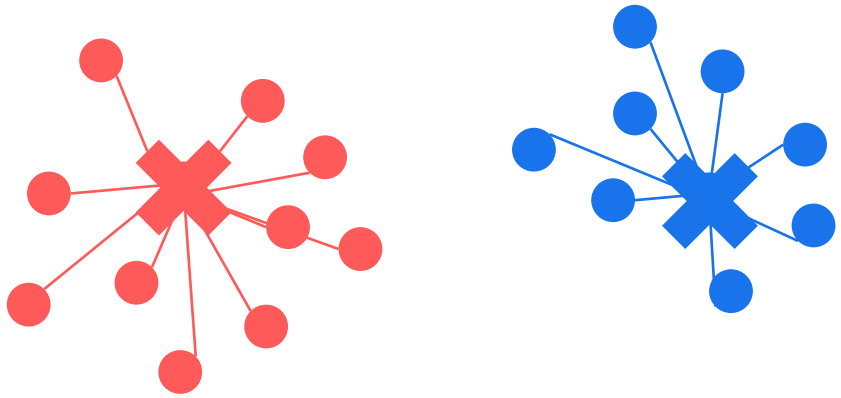












Implement K-Means Clustering

K-Means Clustering is conducted through an iterative process:

1. Place k centroids for the initial clusters

```
model = KMeans(n_clusters = k)
```

2. Assign data samples to the nearest centroid

3. Update centroids based on the above-assigned data samples

```
model.fit(X)
```

4. Assign each data point to a cluster

```
model.predict(X)
```

Let's work on a mini case study in Google Colab

A few unresolved questions …

How do I determine the # of clusters for k-means?

- Data exploration & visualization
- Theory & domain knowledge
- Other practical reasons (e.g., budget)
- …

Do I really need all these features for clustering?

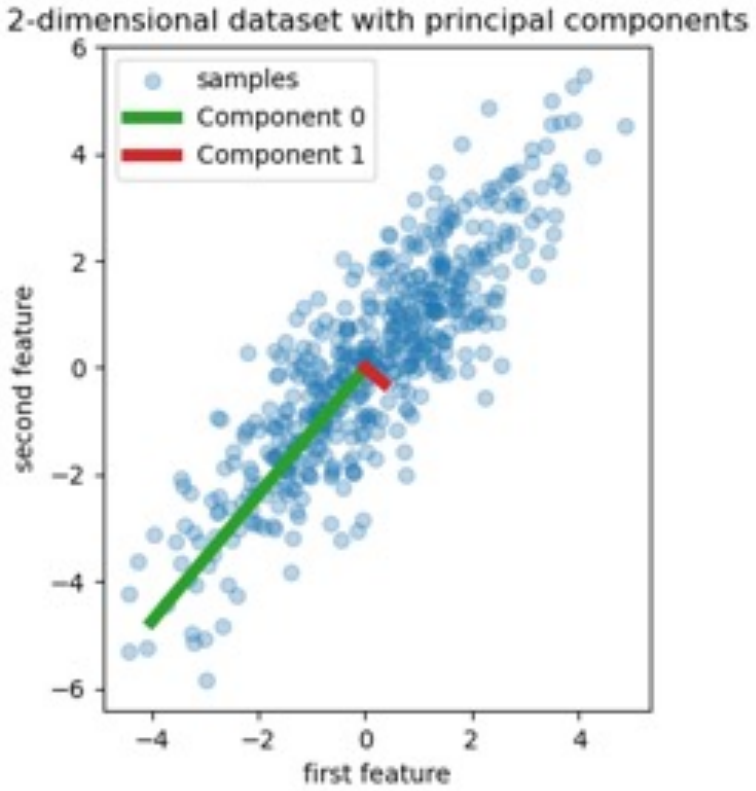
- Dimension scaling/reduction
  - Understand your data in a simpler way
  - Not all features are equally important
  - Multiple features may provide repetitive information
- Principal component analysis (PCA)

Do my clustering results make sense?

- Examine clustering results
- Visualization after clustering

Principal Component Analysis (PCA)

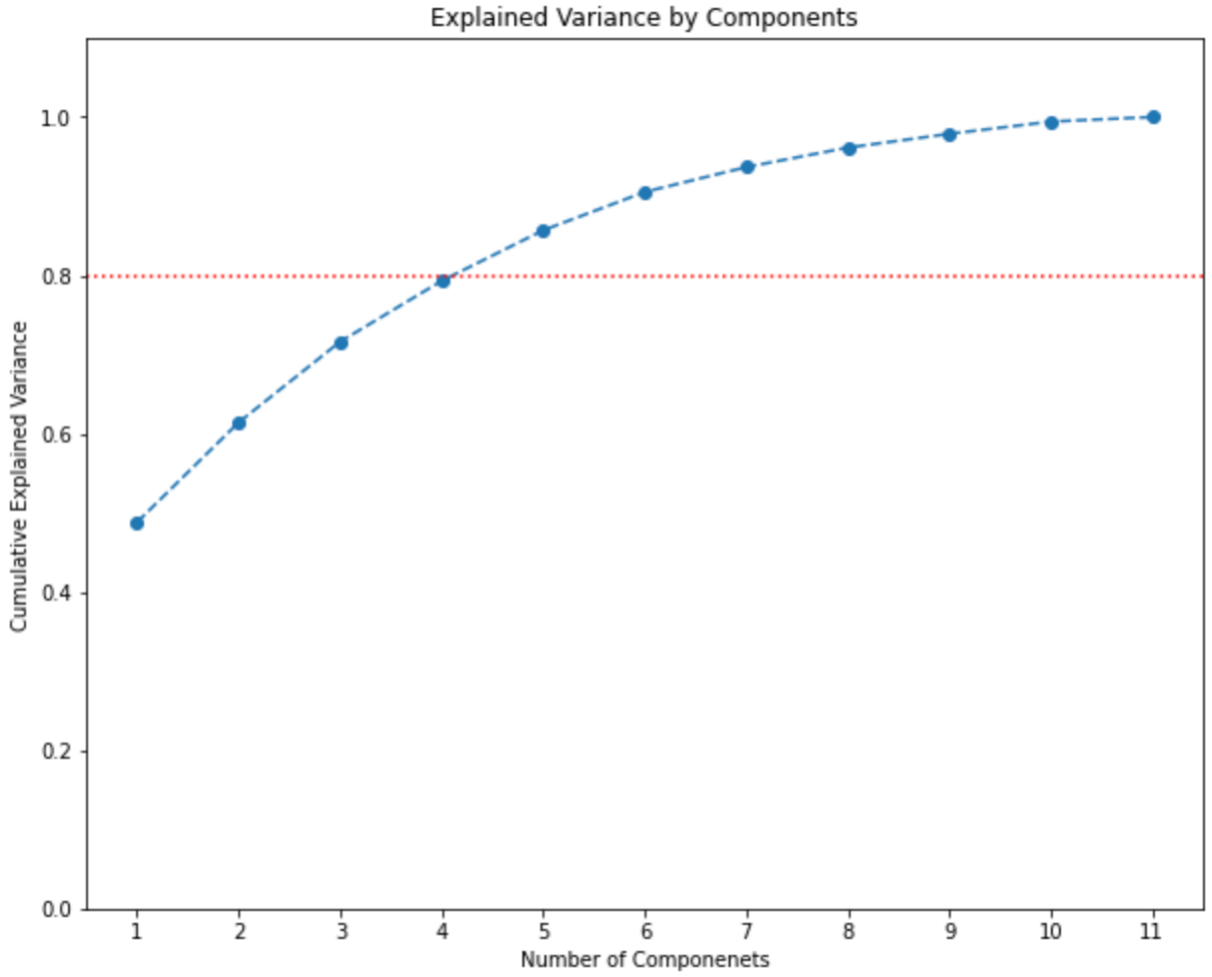
- Not all of the features in your dataset will be equally helpful for clustering
- PCA is a technique for dimension reduction (i.e., to reduce your dataset from higher to lower dimensions)
- Each "component" is a linear combination of your features



PCA using scikit-learn

Principal Component Analysis (PCA)

- To choose the number of component we need, we want to make sure we are explaining sufficient variances in the data when we scale down its dimensions.



A few unresolved questions …

How do I determine the # of clusters for k-means?

- Data exploration & visualization
- Theory & domain knowledge
- Other practical reasons (e.g., budget)
- …

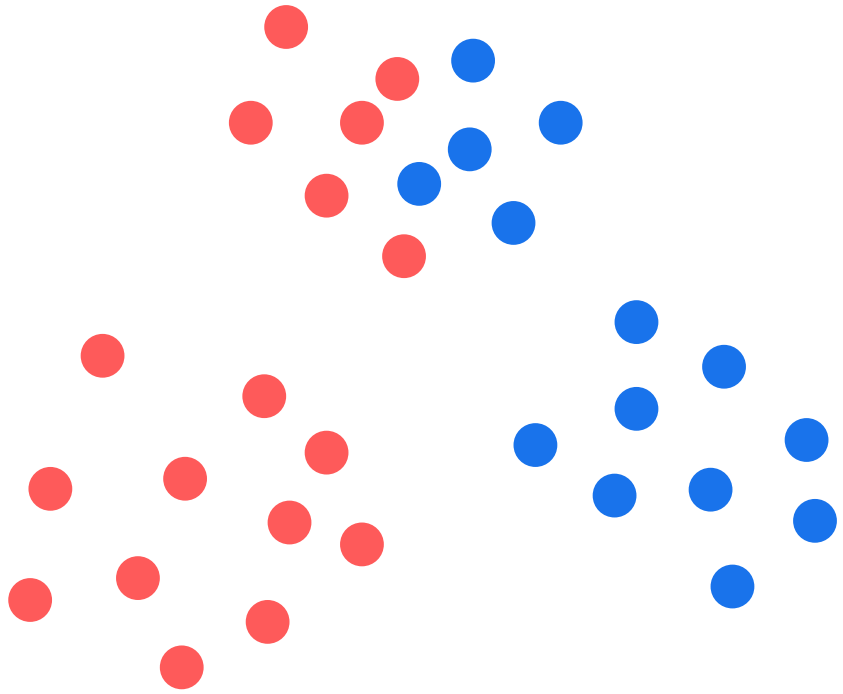
Do I really need all these features for clustering?

- Dimension scaling/reduction
- Principal component analysis (PCA)

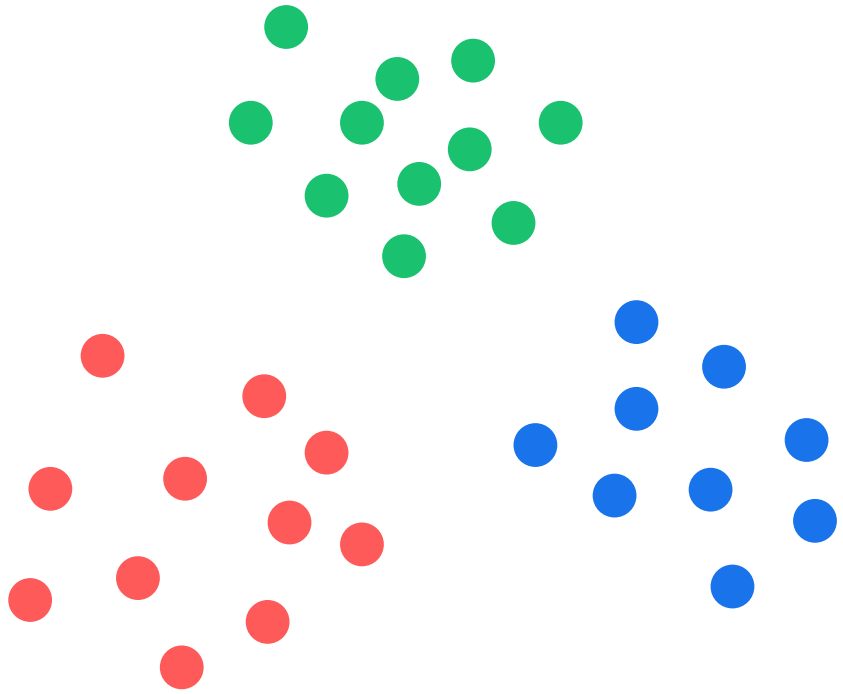
Do my clustering results make sense?

- Examine clustering results
- Visualization after clustering

Examine & Visualize Clustering Results



Examine & Visualize Clustering Results



Summary: The Full Process of Unsupervised Learning

Visualize before k-means	Explore the data – Does it even make sense to cluster?
Dimension reduction using PCA	Do we need all the variables to help us cluster the data?
Determine the # of clusters	How many clusters should we group the data into?
Fit the k-means model	How do we implement k-means clustering?
Examine the clustering results	What are some unique characteristics of each cluster?
(Handle new data)	When we get new data, which cluster does each instance belong to?

Beyond this current workshop…

- Unsupervised learning is more than just K-Means
  Check out unsupervised learning on Scikit-Learn
- In research, exploration is often just the first step!
- Stay tuned – CCSS is rolling out an interactive Machine Learning for Social Sciences Handbook in May

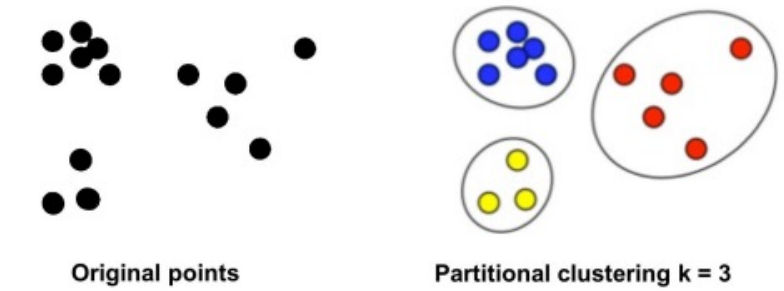


FIGURE 1. **Example of clustering based on partitional methods.**

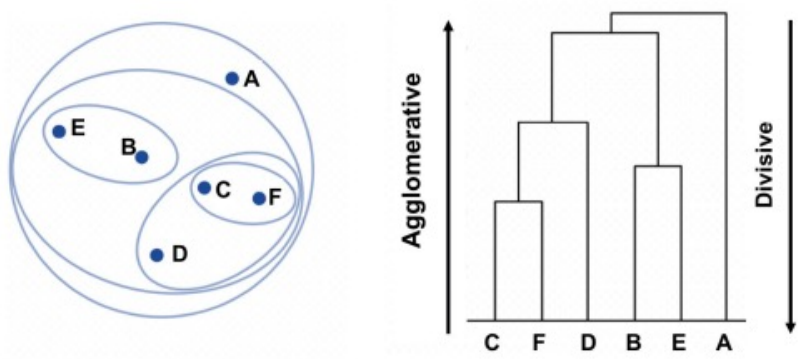
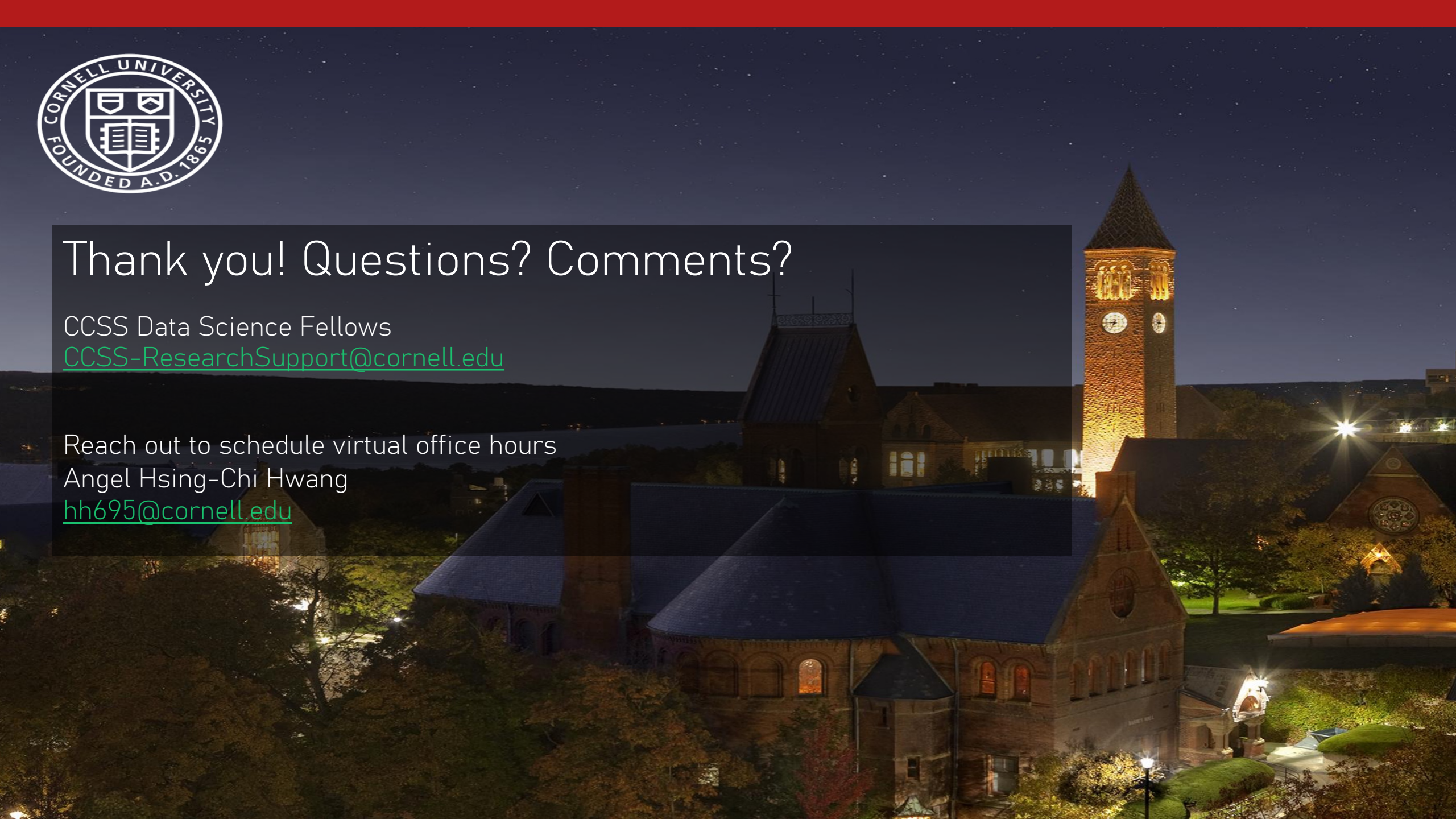


FIGURE 2. **Example of clustering based on hierarchical methods.**

Luna-Romera et al., 2019



Thank you! Questions? Comments?

CCSS Data Science Fellows
CCSS-ResearchSupport@cornell.edu

Reach out to schedule virtual office hours
Angel Hsing-Chi Hwang
hh695@cornell.edu