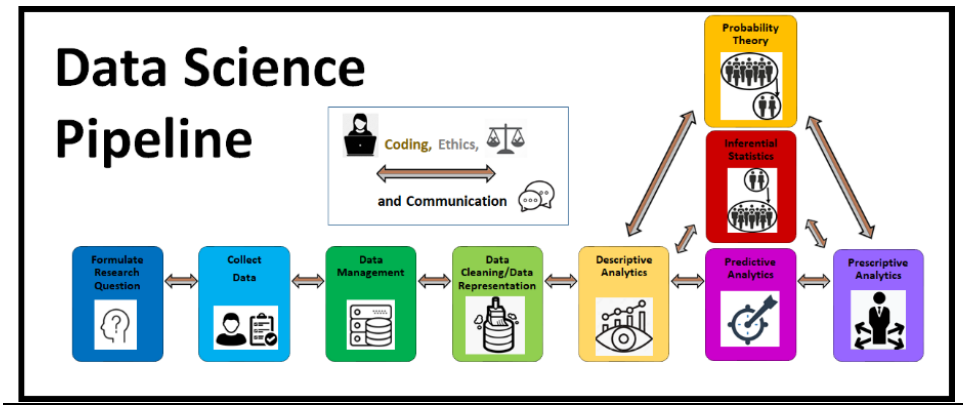


Class Introduction

***Case Study**: What is **unsupervised learning** and how does it fit into the "**full data science pipeline**"?*

Purpose of this Introduction:

- About you
- About me
- Where does machine learning fit into the "full data science pipeline"?
- What is machine learning
- Key assumption behind machine learning algorithms
- Supervised learning
- Unsupervised learning
- Clustering algorithms
- Dimensionality reduction algorithms
- Most common unsupervised learning algorithms
- Machine learning vs. statistics
- Class information
- Learning outcomes
- Lecture structure
- General course tips

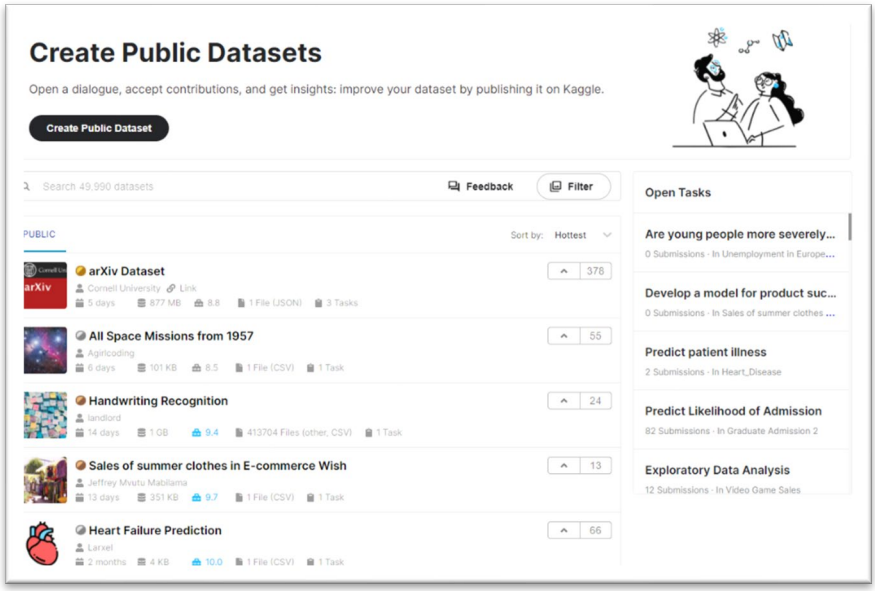
Additional Resources:

- ***Section 2.1 and 2.2*** *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York :Springer, 2013. https://www.statlearning.com/*

- *Matthew Stewart, The Actual Difference Between Statistics and Machine Learning,Towards Data Science. https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3#:~:text=%E2%80%9CThe%20major%20difference%20between%20machine,about%20the%20relationships%20between%20variables.%E2%80%9D*

ABOUT YOU!

What types of data sets would you like to **gain insights from, make predictions with,** and/or **use to help make better decisions?**

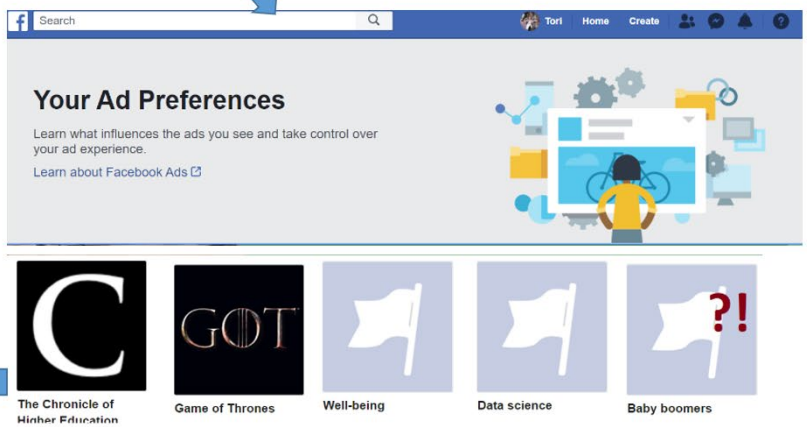


https://www.kaggle.com/datasets

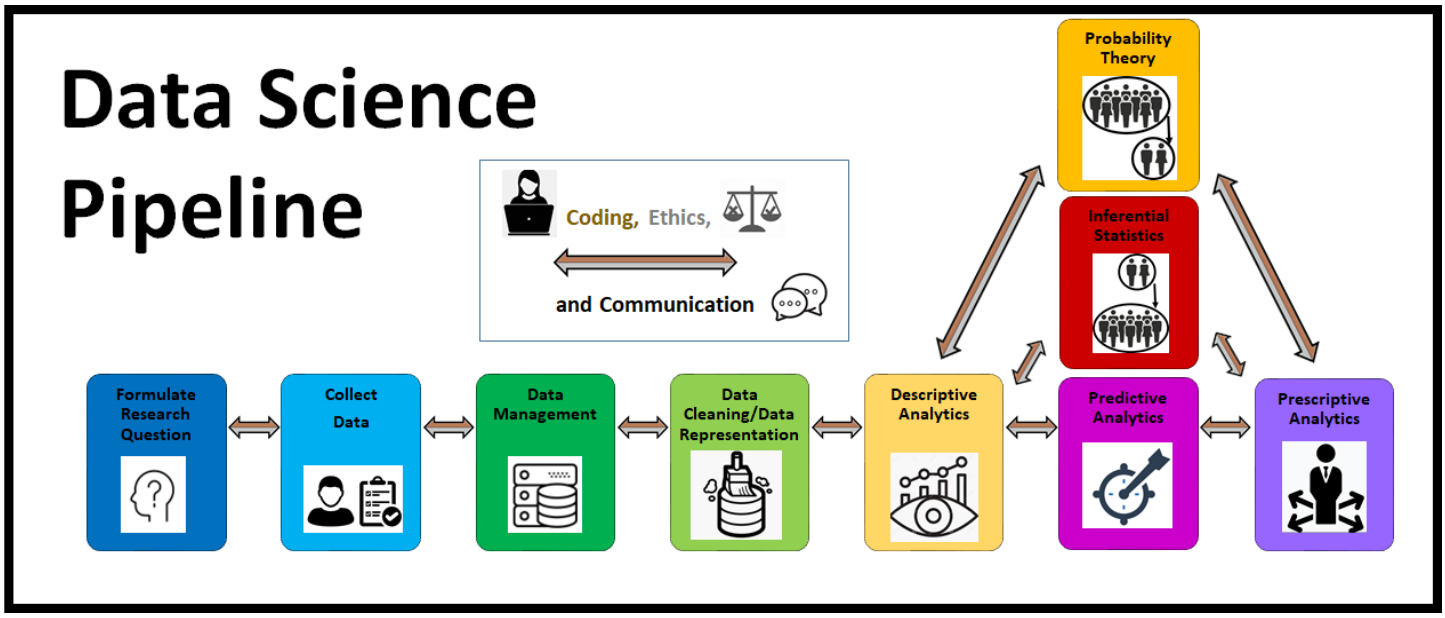
What places have you been able to find fun and interesting datasets from in the past?

About me

- Online Advertising
- TV Advertising
- Narcotics Detection
- Gene Expression Analysis
- Get Out the Vote Initiatives



WHERE DOES **MACHINE LEARNING** FIT IN THE "FULL DATA SCIENCE PIPELINE"?

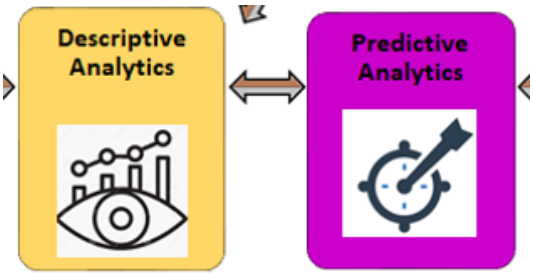


WHAT IS **MACHINE LEARNING**?

Area: Branch of computer science

Goal:

Use data to implement _____ models and _____ models



Two main kinds of machine learning algorithms

- Unsupervised Learning Algorithms: types of _____ *analytics algorithms*
- Supervised Learning Algorithms: types of _____ *analytics algorithms*

KEY ASSUMPTION BEHIND MACHINE LEARNING ALGORITHMS

In a machine learning analysis, there is assumed to be three elements.

1. **Set of Feature Vectors (or Explanatory Variable Values)**
   $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n} \in X$ belonging to a an **input space** X**.**

   Ex: $\boldsymbol{x_1} = (x_{1,hs}, x_{1,coll}, x_{1,unins})$

   Sample of U.S. Counties

	Feature Vectors		
	Highschool Graduation Rate	College Graduation Rate	Percent Uninsured
x1	0.8	0.7	0.8
x2	0.2	0.33	0.35
...	...	...	
xn	0.3	0.4	0.88

2. **Set of Target Values (or Labels/Response Variable Values)**
   $y_1, y_2, \ldots, y_n \in Y$ belonging to a an **output space** Y**.**

   Ex: $y_1 = 0.4$

   Sample of U.S. Counties

	Labels Poverty Rate
y1	0.4
y2	0.2
...	...
yn	0.3

3. **Fixed (Unknown) Mapping Function** $g: X \to Y$
   $Y = g(X) + \epsilon$
   - ϵ: random error term
   - ϵ is independent of *X*
   - ϵ has a mean of 0

   Ex: We might assume that this mapping function is linear.

$$g(x_{hs}, x_{coll}, x_{unins}) = b_0 + b_{hs}x_{hs} + b_{coll}x_{coll} + b_{unins}x_{unins} = \hat{y}$$

SUPERVISED LEARNING ALGORITHMS

In a **supervised learning algorithm**, it is assumed that the target values (or labels/response variable values) of the observations in your training dataset is _____.

- Input:

  Known Training Data: $X = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ \ldots, (\boldsymbol{x_n}, y_n)\}$

- Theoretical Goal: Out of all possible mapping functions $g: X \to Y$, choose the best function $\hat{g}$ such that $\hat{g}(X)$ is as close as possible to Y.

- Caveat:

- Output:

  Predicted Targets (Labels): $\{\hat{g}(\mathbf{x_1}), \hat{g}(\mathbf{x_2}), \ldots \hat{g}(\mathbf{x_n})\}$

How is this statement vague?

Example: *Ordinary least squares linear regression*

- Input:

  Training Data: $X = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ \ldots, (\boldsymbol{x_n}, y_n)\}$

	Feature Vectors		
	Highschool Graduation Rate	College Graduation Rate	Percent Uninsured
x1	0.8	0.7	0.8
x2	0.2	0.33	0.35
...	...	...	
xn	0.3	0.4	0.88

	Labels Poverty Rate
y1	0.4
y2	0.2
...	...
yn	0.3

- Goal: In ordinary least squares linear regression:
  - What types of functions $g: X \to Y$ are considered?
  - How is the best function $\hat{g}$ selected?
  - Can the function $\hat{g}(\boldsymbol{x})$ be easily known and interpreted in a linear regression?

- Output:

  Predicted Targets (Labels): $\{\hat{g}(\mathbf{x_1}), \hat{g}(\mathbf{x_2}), \ldots \hat{g}(\mathbf{x_n})\}$

What are some common supervised learning algorithms that you've heard of?

Main Goal in supervised learning algorithms

- Predict the _____ class labels (or target values/response variable values).

UNSUPERVISED LEARNING ALGORITHMS

In an **unsupervised learning algorithm**, it is assumed that the target values (or labels/response variable values) of the observations in your training dataset are _____.

- Input:

  Known Training Data: $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$

- Assumption:

  Unknown Implicit Training Data: $Y = \{y_1, y_2, \dots, y_n\}$

- Theoretical Goal: Out of all possible mapping functions $g: X \rightarrow Y$, choose the best function $\hat{g}$ such that $\hat{g}(X)$ is as close as possible to Y.

- Output:

  Predicted Targets (Labels): $\{\hat{g}(\mathbf{x_1}), \hat{g}(\mathbf{x_2}), \dots \hat{g}(\mathbf{x_n})\}$

What are some common *types* of unsupervised learning algorithms that you've heard of?

General Goal of Unsupervised Learning Algorithms: discover _____ in the feature vectors (or observations) in the dataset.

CLUSTERING ALGORITHMS

General Goal: Find 'grouping' relationships of feature vectors (or observations). These groups are called **clusters.** A set of clusters is called a _____.

Assumptions:

- Observations in the same clusters should be relatively _____ to each other.
- Observations in different clusters should be relatively _____ from each other.

What are some ways in which this definition is vague?

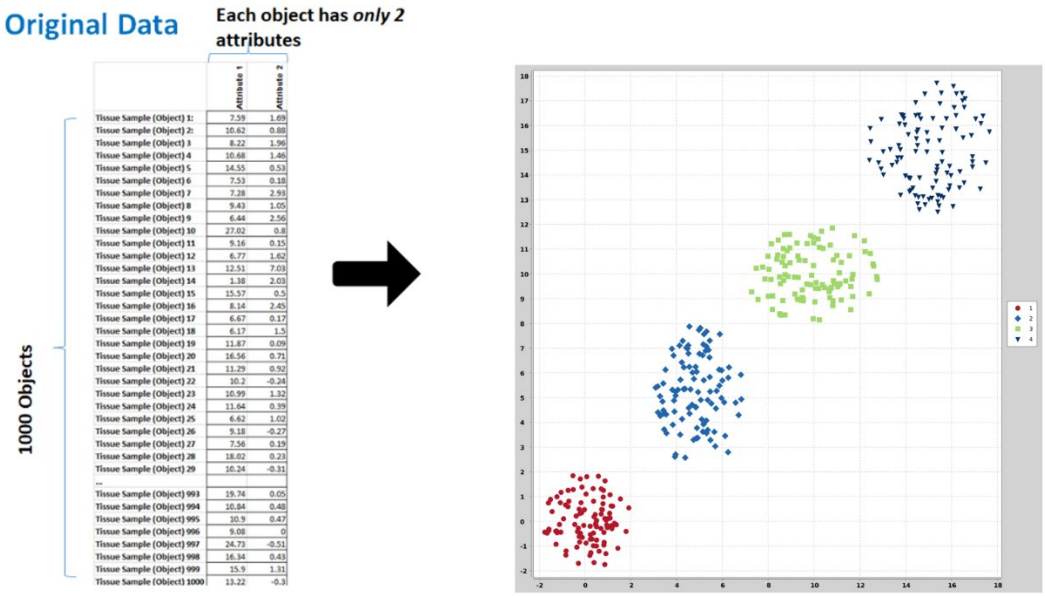
Different clustering algorithms will _____.

Ex: What's one way in which we can say two objects are _____ to one another?

***Example*:** Simplified version of k-means clustering

- Input:

  Known Training Data: $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$



- Assumption:

  Unknown Implicit Training Data: $Y = \{y_1, y_2, \dots, y_n\}$ (ie. the *actual* cluster labels for the *n* observations)

- Theoretical Goal: Out of all possible mapping functions $g: X \rightarrow Y$, choose the best function $\hat{g}$ such that $\hat{g}(X)$ is as close as possible to Y.

- Output:

  Cluster Labels

  Predicted Targets (Labels): $\{\hat{g}(\mathbf{x_1}), \hat{g}(\mathbf{x_2}), \ldots \hat{g}(\mathbf{x_n})\}$

DIMENSIONALITY REDUCTION ALGORITHMS

General Goal: Represent a high-dimensional dataset in a _____-dimensional space, while preserving _____ of the original datasets underlying structure.

Different dimensionality reduction algorithms will

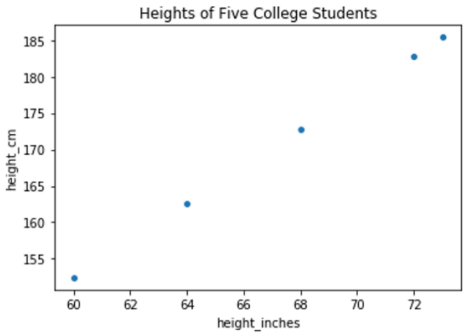
_____.

Example: *Principal Component Analysis*

- Input:

  Known Training Data: $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$

Original Data



	Feature Vectors		
	Height (in)	Height (cm)	
x1	60	152.4	
x2	64	162.56	
x3	68	172.72	
x4	72	182.88	
x5	73	185.42	
Variance	29.8	192.25768	Total Variance: 222.05768

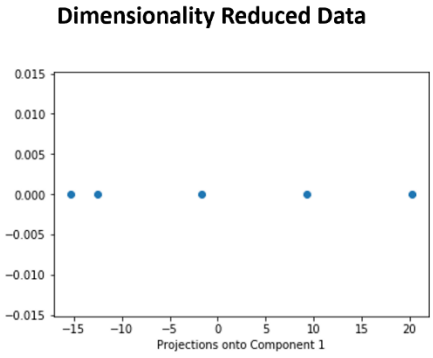
- Assumption:

  Unknown Implicit Training Data: $Y = \{y_1, y_2, \dots, y_n\}$ (ie. the *actual* response variable values that you would like to project the n observations onto)

- Theoretical Goal: Out of all possible mapping functions $g: X \to Y$, choose the best function $\hat{g}$ such that $\hat{g}(X)$ is as close as possible to Y.

- Output:

  Predicted Response Variable Values: $\{\hat{g}(\mathbf{x_1}), \hat{g}(\mathbf{x_2}), \ldots \hat{g}(\mathbf{x_n})\}$



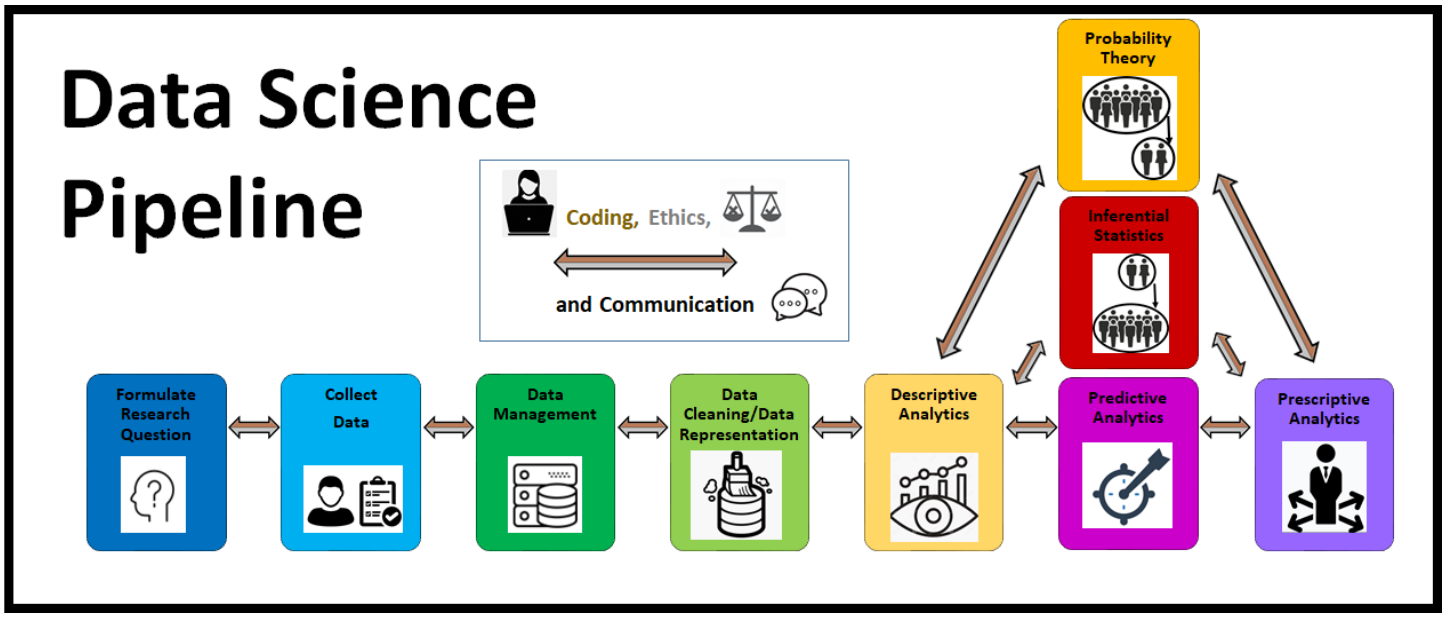
	Principal Component 1
y^1	20.200238
y^2	9.28119044
y^3	-1.6378571
y^4	-12.556905
y^5	-15.286667
Variance	222.05768

MACHINE LEARNING VS. STATISTICS

Models are built in both **machine learning** and traditional **inferential statistics.** But the goals for building these models differ.

- Goal of machine learning model:

- Goal of inferential statistics model:



CLASS INFORMATION

Main Course Website

Check out the course website http://courses.las.illinois.edu/spring2024/stat437 for the following.

1. Course schedule and *incomplete* lecture notes (to be filled out in the lecture).
2. Syllabus
3. Tech Guides
4. Course Staff Info

Canvas Page

On this STAT437 Canvas page, you can find the following (see the tabs on the left hand side).

1. Your grades
2. Assignment and Project Information
3. Lecture markups
    - We will markup the fill-in-the-blank unit lecture note materials during the lecture. You can find the incomplete notes here to follow along. The marked lecture notes will be posted to Canvas in this tab within 24 hours of the lecture.
4. Lecture videos
    - Videos of the lectures will be posted within 24 hours here.
5. Piazza: https://piazza.com/illinois/spring2024/stat437
    - If you have any content and/or non-personal course related questions, you can ask them here. The TAs and/or should be able to answer your question shortly.
        - If you have a personal question about the course, you can email Dr. Ellison at vellison@illinois.edu
6. Zoom Meeting Links
    - You can find the Zoom links for Tori's Office Hours and Peng's Office Hours here.
    - You can either attend office hours in person or over Zoom. It's up to you.

CLASS INFORMATION

General Goal:

Learn a series of tools (algorithms) that allow us to discover and describe hidden insights contained in high-dimensional unlabeled data.

Full Unsupervised Learning Analyses:

- Specifically given real-world data sets, students should be able to code a full unsupervised learning analysis in Python. This includes the following.
    - Be able to justify <u>when/if</u> it is useful to use a clustering algorithm or dimensionality reduction algorithm for a given dataset, research question, and research scenario.

    - Be able to justify <u>which</u> clustering and/or dimensionality reduction <u>algorithms</u> are most appropriate to use for a given dataset, research question, and research scenario.

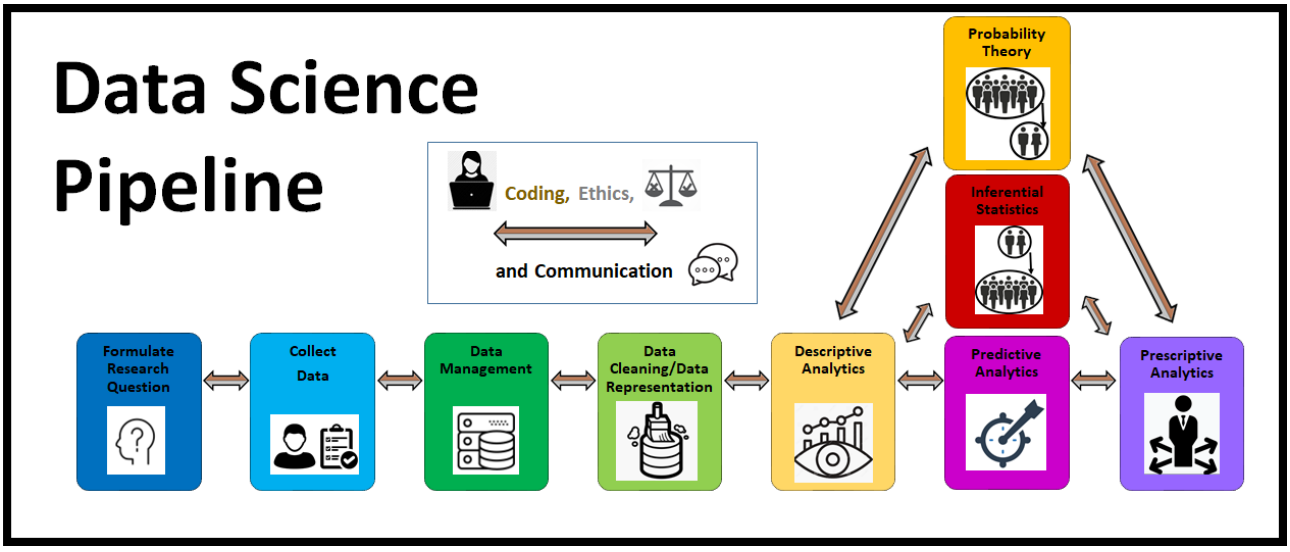
    - If the clustering and/or dimensionality reduction algorithm has different settings/parameters that can be utilized, be able to justify <u>which parameters to use</u>

- Be able to justify the evaluation metric(s)/methods that were used to: a.) select which algorithm/model/parameters to use as well b.) describe the nature of the results.

- Be able to interpret the results of the algorithms and effectively communicate as many hidden insights as possible about the dataset.

- Be able to understand how different aspects of data pre-processing might affect the results of the unsupervised learning algorithms.

- Be able to use these unsupervised learning insights to help make predictions as well as make good business decisions.



Develop Knowledgebase and Intuition about Unsupervised Learning Algorithms:

- In general, students should also know the following.
    - How each algorithm works and the output of each algorithm.
    - How each algorithm evaluation metric works.
    - Develop an intuition for what happens when we apply to these algorithms and evaluation metrics to 2-d datasets.
    - Students should know how to conduct at least one iteration of these algorithms and calculate these evaluation metrics by hand.
    - Students should know how to code these algorithms and evaluation metrics in Python.
    - Students should demonstrate best practices when effectively communicating and presenting data science results. (Ie: titles on graphs, label the axes etc.)

LECTURE STRUCTURE

Before Lecture

Will post lecture unit materials to course website.

Lecture Unit Materials

- Slides pdfs (*conceptual)*
- Jupyter notebook *(application)*
- Jupyter notebook pdf copy
- csv (usually)

General Course Tips

- Check your email regularly!
- Go to office hours
- Start working on your assignments early.
- Piazza Discussion board can be helpful!
- Ask questions if you get stuck.