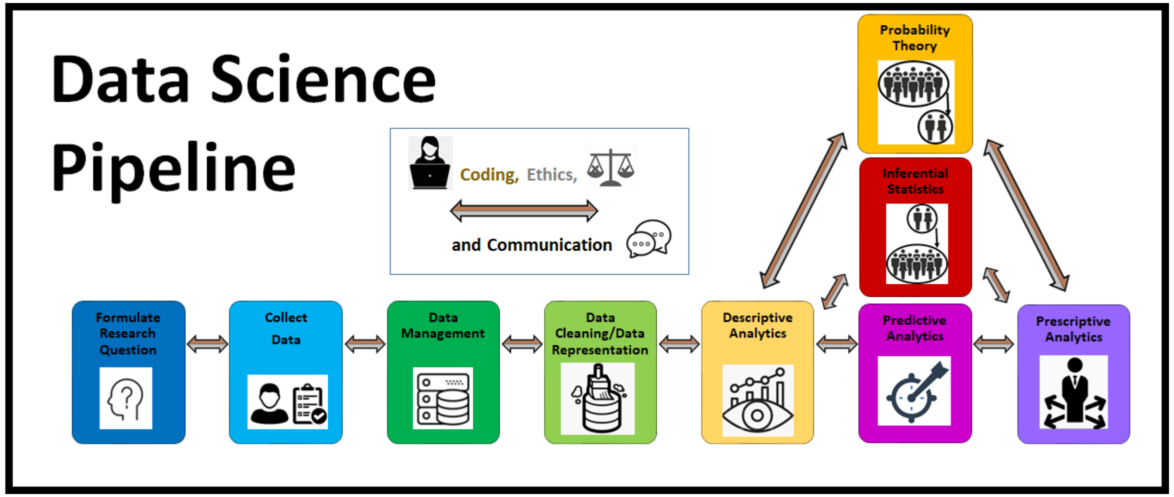
Introduction to the STAT430 Unsupervised Learning Course

August 25, 2020



Data Science Pipeline

Coding, Ethics, and Communication

Formulate Research Question

Collect Data

Data Management

Data Cleaning/Data Representation

Descriptive Analytics

Probability Theory

Inferential Statistics

Predictive Analytics

Prescriptive Analytics

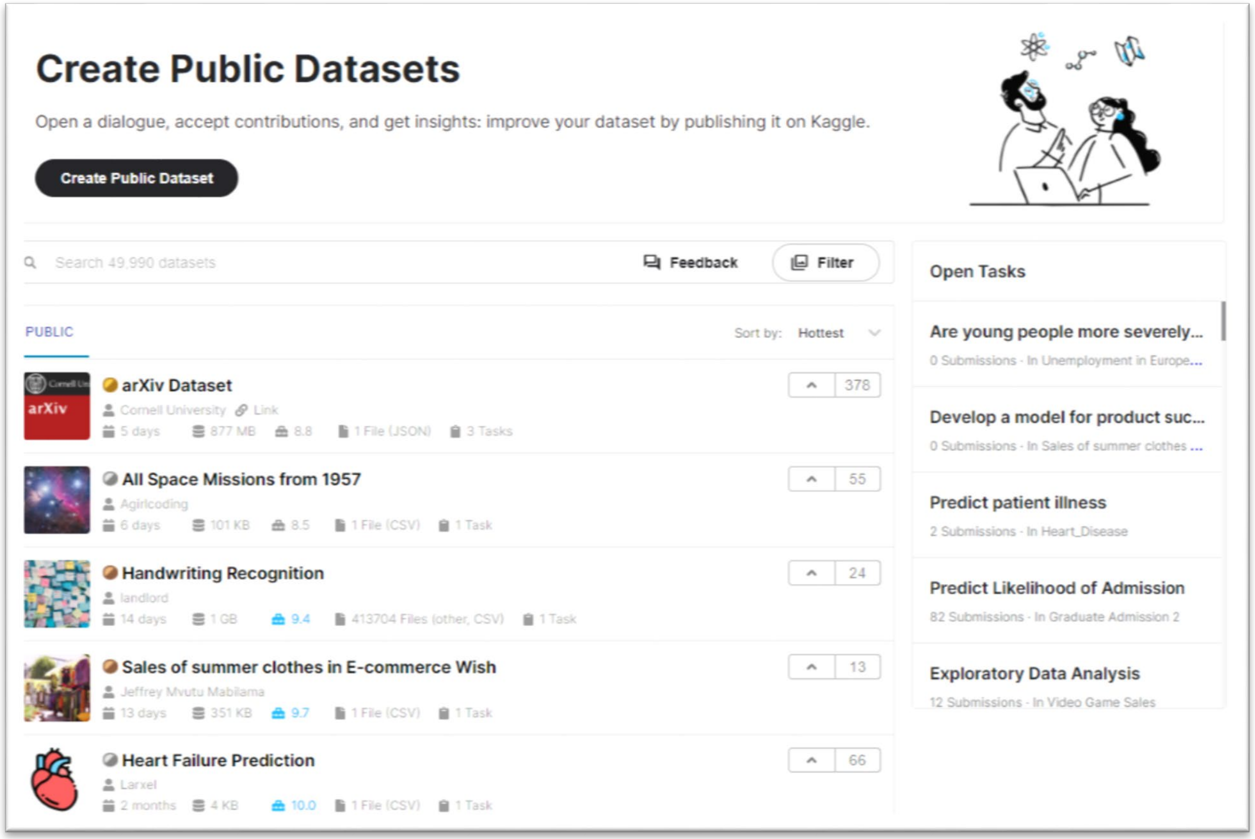
Introduction to this Course

- About you
- About me
- What is machine learning?
- Supervised vs. unsupervised learning
- Most common unsupervised learning algorithms

- Class Information
- Learning content tips/course goals
- Lecture tips
- General course tips

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?**



Create Public Datasets

Open a dialogue, accept contributions, and get insights: improve your dataset by publishing it on Kaggle.

arXiv Dataset — 378
Cornell University · Link
5 days · 877 MB · 8.8 · 1 File (JSON) · 3 Tasks

All Space Missions from 1957 — 55
Agirlcoding
6 days · 101 KB · 8.5 · 1 File (CSV) · 1 Task

Handwriting Recognition — 24
landlord
14 days · 1 GB · 9.4 · 413704 Files (other, CSV) · 1 Task

Sales of summer clothes in E-commerce Wish — 13
Jeffrey Mvutu Mabilama
13 days · 351 KB · 9.7 · 1 File (CSV) · 1 Task

Heart Failure Prediction — 66
Larxel
2 months · 4 KB · 10.0 · 1 File (CSV) · 1 Task

Open Tasks

Are young people more severely...
0 Submissions · In Unemployment in Europe...

Develop a model for product suc...
0 Submissions · In Sales of summer clothes ...

Predict patient illness
2 Submissions · In Heart_Disease

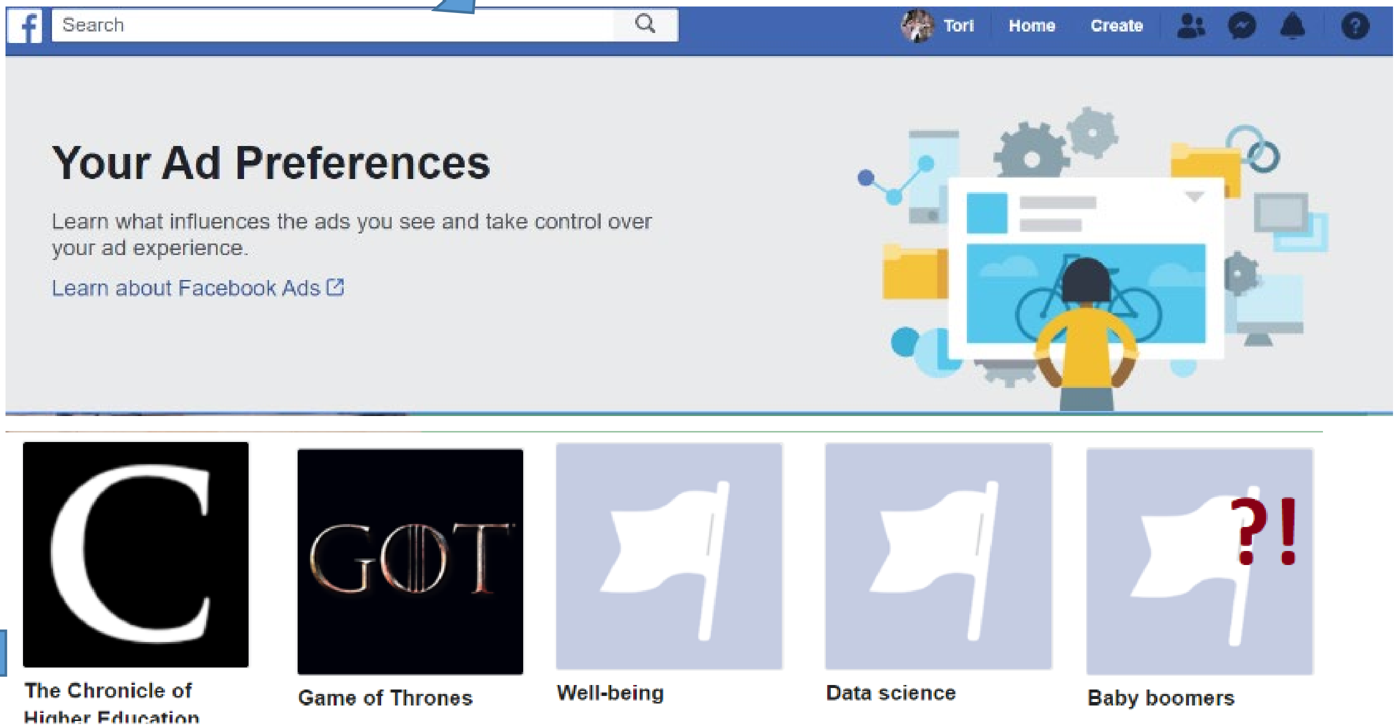
Predict Likelihood of Admission
82 Submissions · In Graduate Admission 2

Exploratory Data Analysis
12 Submissions · In Video Game Sales

https://www.kaggle.com/datasets

About Me

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



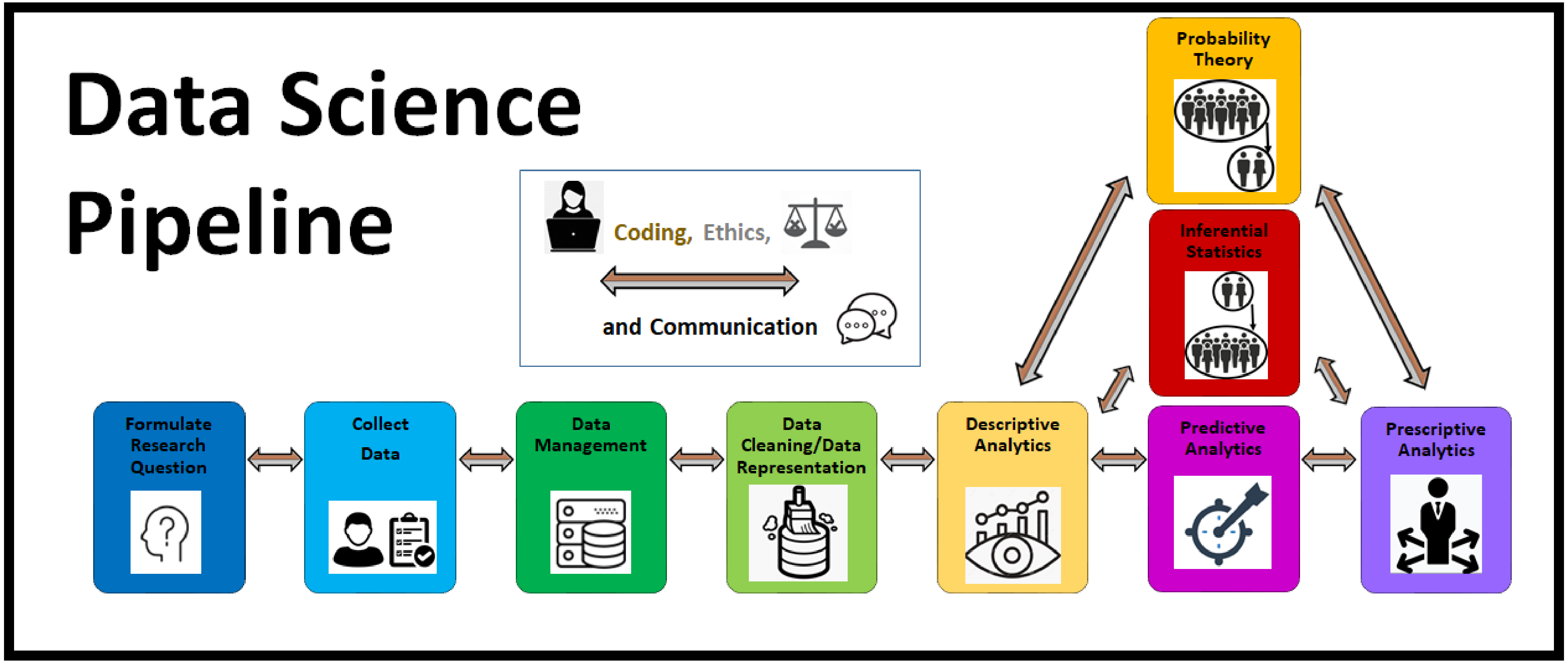
Your Ad Preferences

Learn what influences the ads you see and take control over your ad experience.

Learn about Facebook Ads

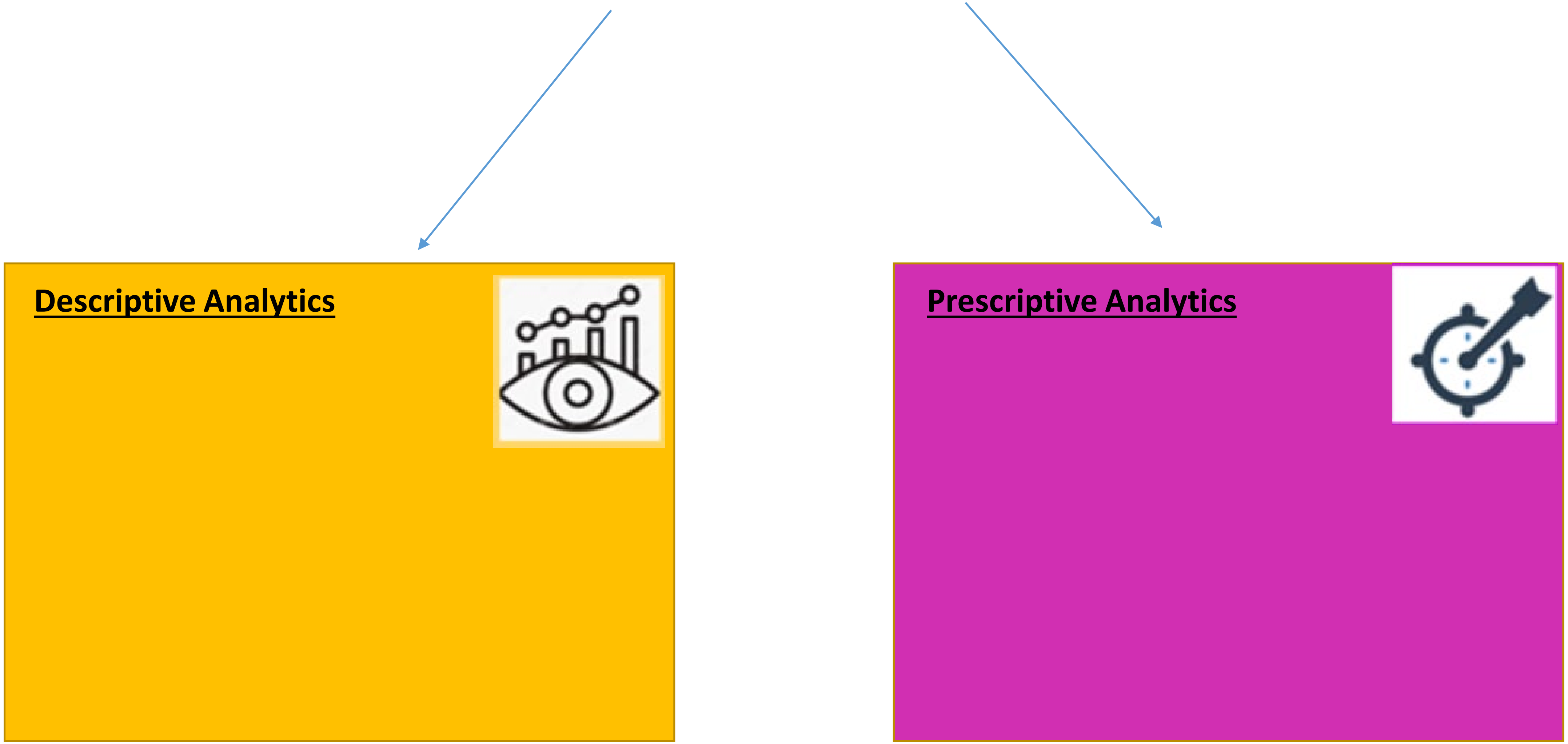
- The Chronicle of Higher Education
- Game of Thrones
- Well-being
- Data science
- Baby boomers

What is machine learning?



What is machine learning?

- **Area:** Branch of computer science
- **Goal**:
  - Use data to implement descriptive models and predictive models



Descriptive Analytics

Prescriptive Analytics

Supervised vs. Unsupervised Learning Algorithms

- **Two main kinds:**
  - Supervised Learning Algorithms: types of *predictive analytics algorithms*
  - Unsupervised Learning Algorithms: types of *descriptive analytics algorithms*

- **Supervised Learning Algorithms:**

  Labels

  - Input:
    - **Training Data**: $X = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \dots, (\boldsymbol{x_n}, y_n)\}$

  Feature Vectors

  - Output:
    - Function $g: X \to Y$
      - $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n} \in X$ $(\boldsymbol{Input\ Space})$
      - $y_1, y_2, \dots, y_n \in Y$ $(\boldsymbol{Output\ Space})$

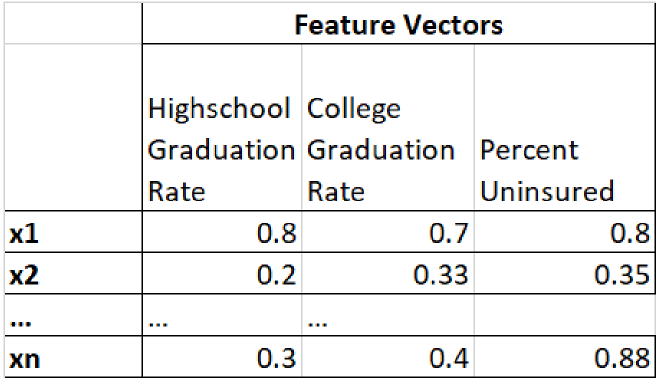
Supervised vs. Unsupervised Learning Algorithms

- **Supervised Learning Algorithms:**

Labels

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

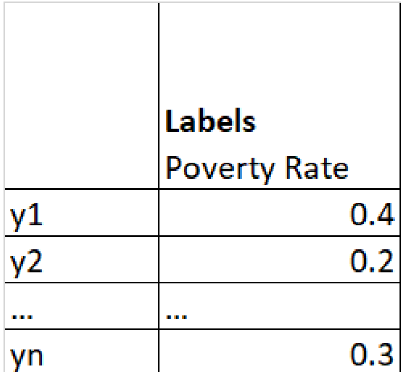
Feature Vectors



	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

Example:
- Linear regression

$$\hat{y} = b_0 + b_{hs}x_{hs} + b_{coll}x_{coll} + b_{unins}x_{unins}$$



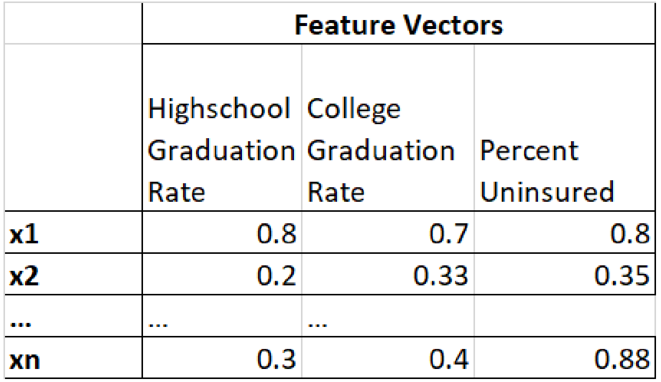
	Labels Poverty Rate
y1	0.4
y2	0.2
...	...
yn	0.3

  - Output:
    - Function $g: X \rightarrow Y$
      - $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n} \in X$ $(\boldsymbol{Input\ Space})$
      - $y_1, y_2, \ldots, y_n \in Y$ $(\boldsymbol{Output\ Space})$

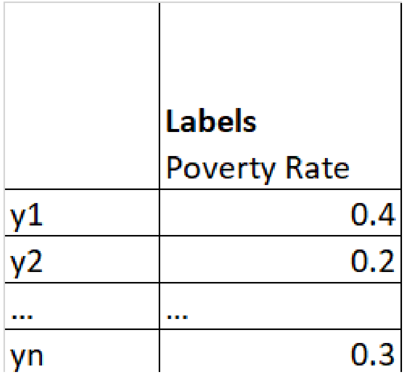
Supervised vs. Unsupervised Learning Algorithms

- **Supervised Learning Algorithms:**

Training Data



	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

Common Supervised Learning Algorithms:

- Linear regression
- Logistic regression
- Naïve Bayes
- Decision Trees/Random Forests
- Linear Discriminant Analysis
- K-nearest neighbors algorithms
- Support vector machines
- Neural networks

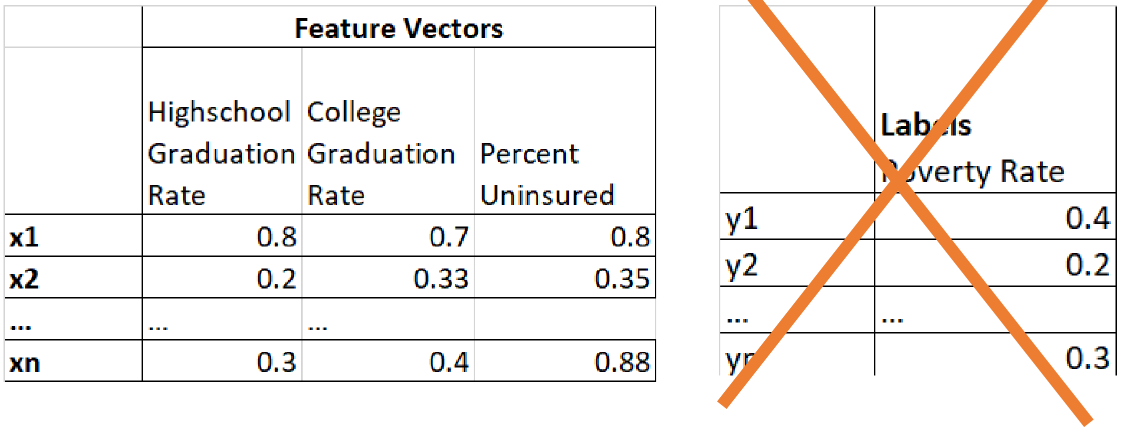
Key point:

- The **training data** in supervised learning algorithms always has **labels.**
- General goal is to **predict the labels.**

Supervised vs. Unsupervised Learning Algorithms

- **Unsupervised Learning Algorithms:**

Training Data



	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

Common Supervised Learning Algorithms:

- Clustering algorithms
- Dimensionality reduction algorithms

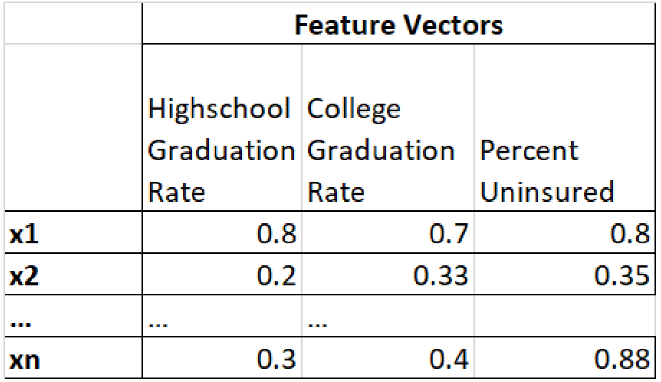
Key point:

- The **training data** in UNsupervised learning algorithms **doesn't have labels.**
- General goal **discover hidden patterns** in the feature vectors.

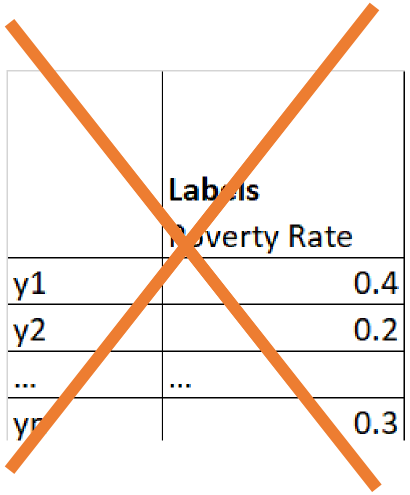
Supervised vs. Unsupervised Learning Algorithms

- **Unsupervised Learning Algorithms:**

Training Data



	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

Common Supervised Learning Algorithms:

- Clustering algorithms
- Dimensionality reduction algorithms

Key point:

- The **training data** in UNsupervised learning algorithms **doesn't have labels.**
- General goal **discover hidden patterns** in the feature vectors.

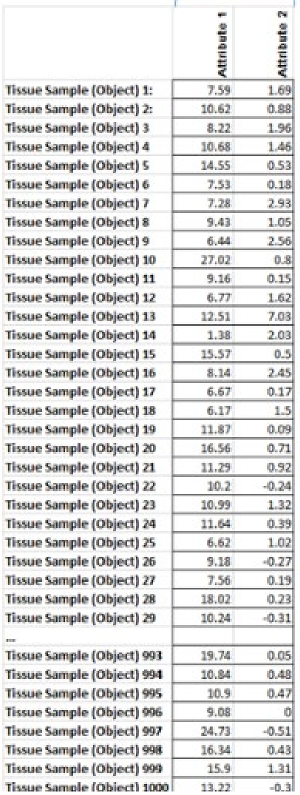
Types of Unsupervised Learning Algorithms

- **Common Types of Unsupervised Learning Algorithms:**
    - **Clustering algorithms**
        - Goal: Hidden ‘grouping’ relationships in the data.

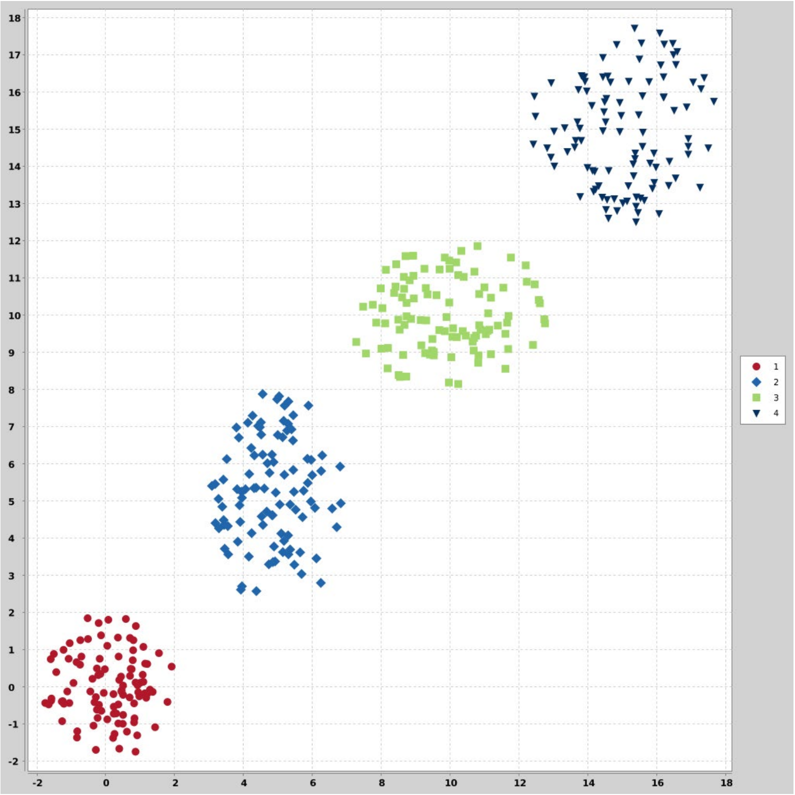
Original Data

Each object has *only 2* attributes

1000 Objects



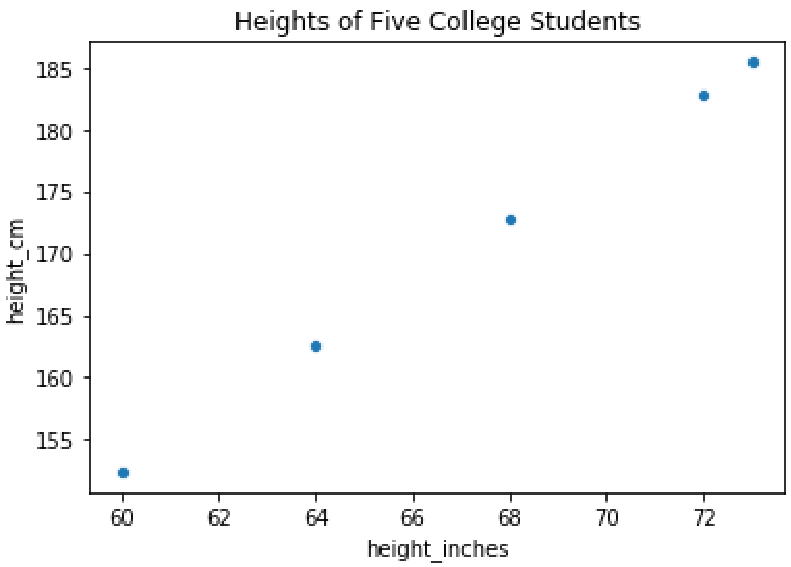
	Attribute 1	Attribute 2
Tissue Sample (Object) 1:	7.59	1.69
Tissue Sample (Object) 2:	10.62	0.88
Tissue Sample (Object) 3	8.22	1.96
Tissue Sample (Object) 4	10.68	1.46
Tissue Sample (Object) 5	14.55	0.53
Tissue Sample (Object) 6	7.53	0.18
Tissue Sample (Object) 7	7.28	2.93
Tissue Sample (Object) 8	9.43	1.05
Tissue Sample (Object) 9	6.44	2.56
Tissue Sample (Object) 10	27.02	0.8
Tissue Sample (Object) 11	9.16	0.15
Tissue Sample (Object) 12	6.77	1.62
Tissue Sample (Object) 13	12.51	7.03
Tissue Sample (Object) 14	1.38	2.03
Tissue Sample (Object) 15	15.57	0.5
Tissue Sample (Object) 16	8.14	2.45
Tissue Sample (Object) 17	6.67	0.17
Tissue Sample (Object) 18	6.17	1.5
Tissue Sample (Object) 19	11.87	0.09
Tissue Sample (Object) 20	16.56	0.71
Tissue Sample (Object) 21	11.29	0.92
Tissue Sample (Object) 22	10.2	-0.24
Tissue Sample (Object) 23	10.99	1.32
Tissue Sample (Object) 24	11.64	0.39
Tissue Sample (Object) 25	6.62	1.02
Tissue Sample (Object) 26	9.18	-0.27
Tissue Sample (Object) 27	7.56	0.19
Tissue Sample (Object) 28	18.02	0.23
Tissue Sample (Object) 29	10.24	-0.31
...		
Tissue Sample (Object) 993	19.74	0.05
Tissue Sample (Object) 994	10.84	0.48
Tissue Sample (Object) 995	10.9	0.47
Tissue Sample (Object) 996	9.08	0
Tissue Sample (Object) 997	24.73	-0.51
Tissue Sample (Object) 998	16.34	0.43
Tissue Sample (Object) 999	15.9	1.31
Tissue Sample (Object) 1000	13.22	-0.3



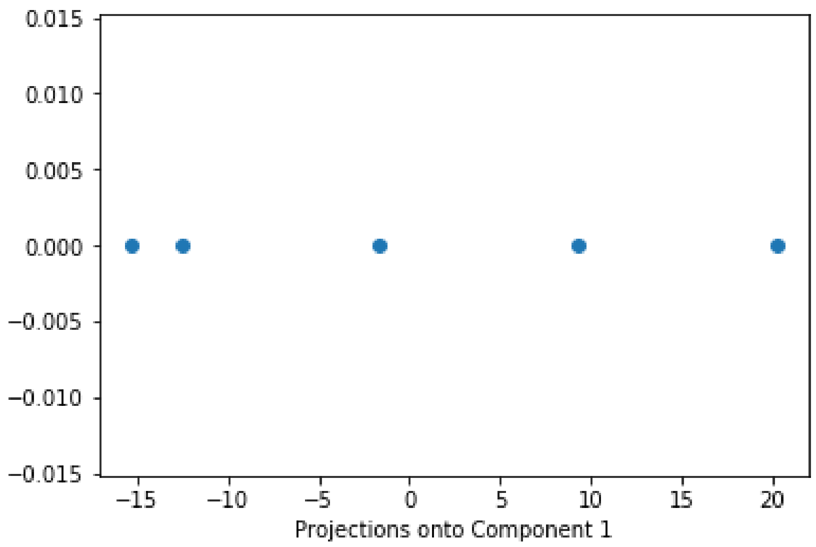
Types of Unsupervised Learning Algorithms

- **Common Types of Unsupervised Learning Algorithms:**
  - **Dimensionality Reduction Algorithms**
    - Goal: Represent a high-dimensional datasets in a lower-dimensional space, while preserving *certain aspects* of the original datasets underlying structure.

Original Data



Dimensionality Reduced Data



Structure: Total Variance of Height (inches) and Height (cm) = 29.8

=

Structure: Total variability of dimensionality reduced data =29.8

Course Website and Syllabus

Course Website: http://courses.las.illinois.edu/fall2020/stat430

- Schedule
- Syllabus
- Course information
- Assignments
- Python Resource Help Pages

Compass Page: https://compass2g.illinois.edu/

- Zoom links for:
  - Lectures
  - My office hours + TA office hours
- Videos Posted of the lecture
- Grades

Piazza: https://piazza.com/illinois/fall2020/stat430

- Content and non-personal course related questions.

Learning Content Tips

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 **when/if 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 **which** clustering and/or dimensionality reduction **algorithms 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 **which parameters to use**
  - 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.**

Learning Content Tips

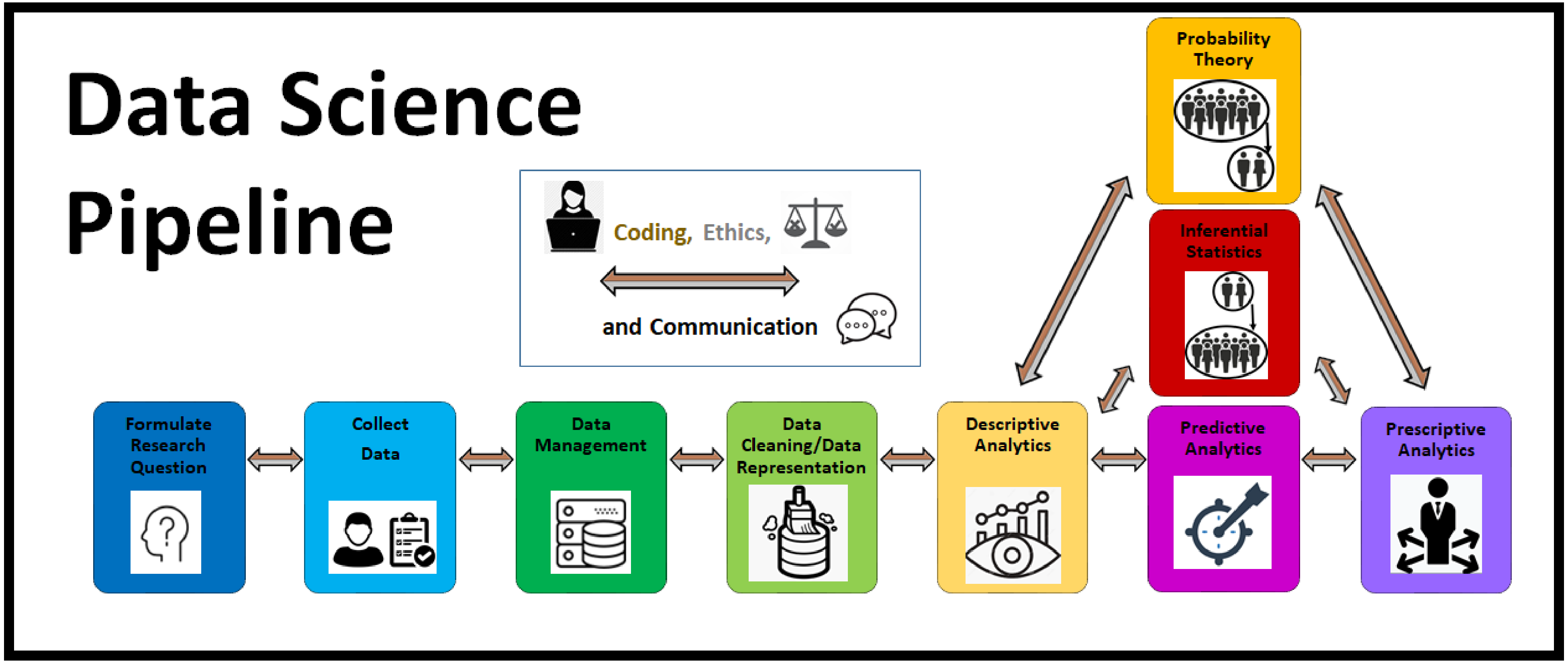
General Goal:

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

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.)

Learning Content Tips



Lecture Tips - Synchronous

- **Synchronous:** strongly encouraged if you are able to, but not required!

- **Each class download these (posted by 8am CST before class)**
  - Python Notebooks
  - Pdf

- **Note-taking Ideas**
  - Printing the pdf, hand written notes
  - Onenote *(or other similar notetaking software)*
  - Make notes in your Python Notebook

- **Following Along with Code**
  - Download .ipynb before class and try to follow along *(not all class notes will be in .ipynbs, but all code will be in the pdf)*

- **Engaging during Lecture**
  - Zoom chatroom
  - Private chatroom messages to TA.

- **Breakout Rooms**
  - Ask classmates for help in the breakout rooms.
  - Ask me/CAs for help in the breakout rooms.

Lecture Tips - Asynchronous

- **Expectation: Watch videos within 24 hours of posting**
  - Try watching with classmates
  - Try watching during office hours/lab to ask question.
  - Ask questions on Piazza.

General Course Tips

- Check your email regularly!
- Go to office hours
  - **Tori: Fridays 9:30-10:30am CST**
  - **Rong: Wednesdays 5:30-7:30pm CST**
  - More coming soon (after survey)
- Start working on your assignments early.
- Piazza can be helpful!
- Ask questions if you get stuck.

- New Idea: After class, write around 4 sentences describing what you just learned.