

Introduction to Machine Learning – What is Machine Learning?

Dr. Ab Mosca (they/them)

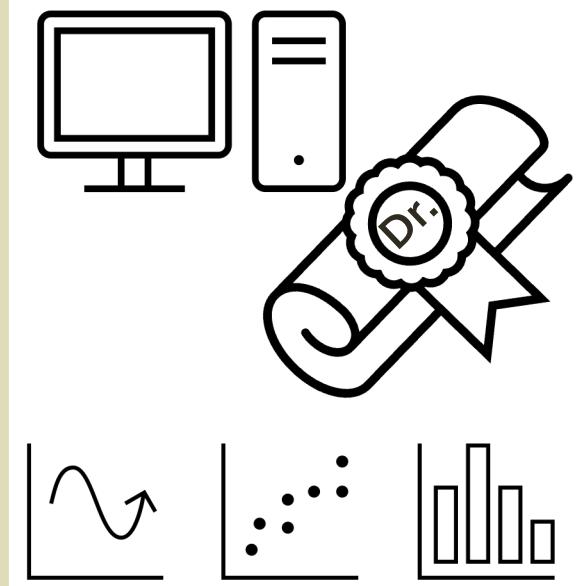
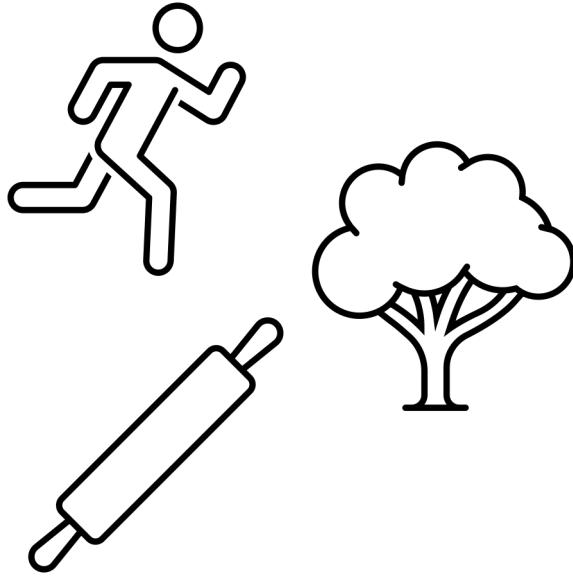
Plan for Today

- Who am I?
- Who are you?
- What will we do in this class?
- What is Machine Learning?

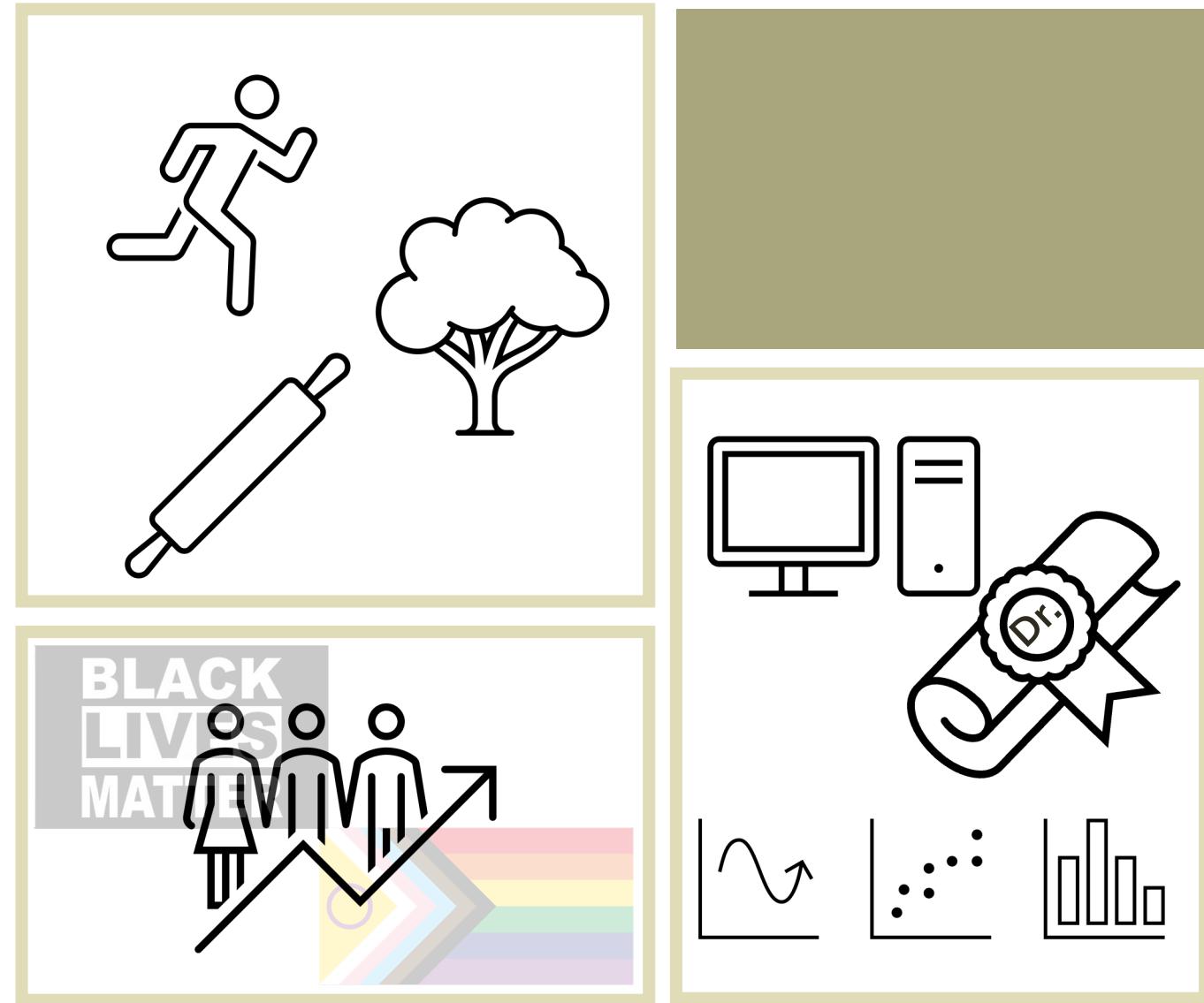
Who Am I?



Who Am I?



Who Am I?



Who Are You?

- Form groups of 3
- Introduce yourselves (name, pronouns)
- Share:
 - A highlight of your winter break
- Find 1 thing that your entire group has in common (favorite color? hometown? left-handed? Be creative!)
- After about 5 minutes we will go around, introduce ourselves, and share what each group has in common

Who Are You?

- Form **new groups** of 3 (move around!)
- Introduce yourselves (name, pronouns)
- Share:
 - Would you rather be able to touch a book and instantly read it OR remember everything you've ever read?
- After about 5 minutes we will go around, introduce ourselves, and share our would you rather answers

Who Are You?

- Form **new new groups** of 3 (move around!)
- Introduce yourselves (name, pronouns)
- Share:
 - Would you rather have your dishes wash themselves or your laundry wash itself?
- After about 5 minutes we will go around, introduce ourselves, and share our would you rather answers



What You Will Learn & Logistics

What is this class?



1. Understand what ML is (and isn't)



2. Learn some foundational methods / tools



3. Be able to choose methods that make sense

Important Info

- Course website (**write this down!**):
<https://amoscao1.github.io/CAIS380ML-S24/>
- Office Hours
 - Wilson Hall 325
 - Wednesday 09:30 - 11:00
 - Thursday 14:30 - 16:30
 - By Appointment

Important Info

- Textbook: *An Introduction to Statistical Learning*
 - See course website for instructions
- Assignments:
 - Turn in on Gradescope
(<https://help.gradescope.com/article/ccbpppziug-student-submit-work>)
- Due Dates: As listed on course schedule.
 - 24hr grace period; no late submissions
 - Lowest homework dropped
 - See syllabus for revise and resubmit policy

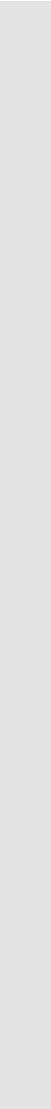
****Important Info****

Assignments

- Homework
 - Pair assignments
 - Graded on effort and correctness
- Quizzes
 - Individual assignments
 - Can re-take as many times as wanted before deadline
- In-class Activities
 - Graded on effort
- Final Project
 - Small group
 - Graded on creativity and correctness

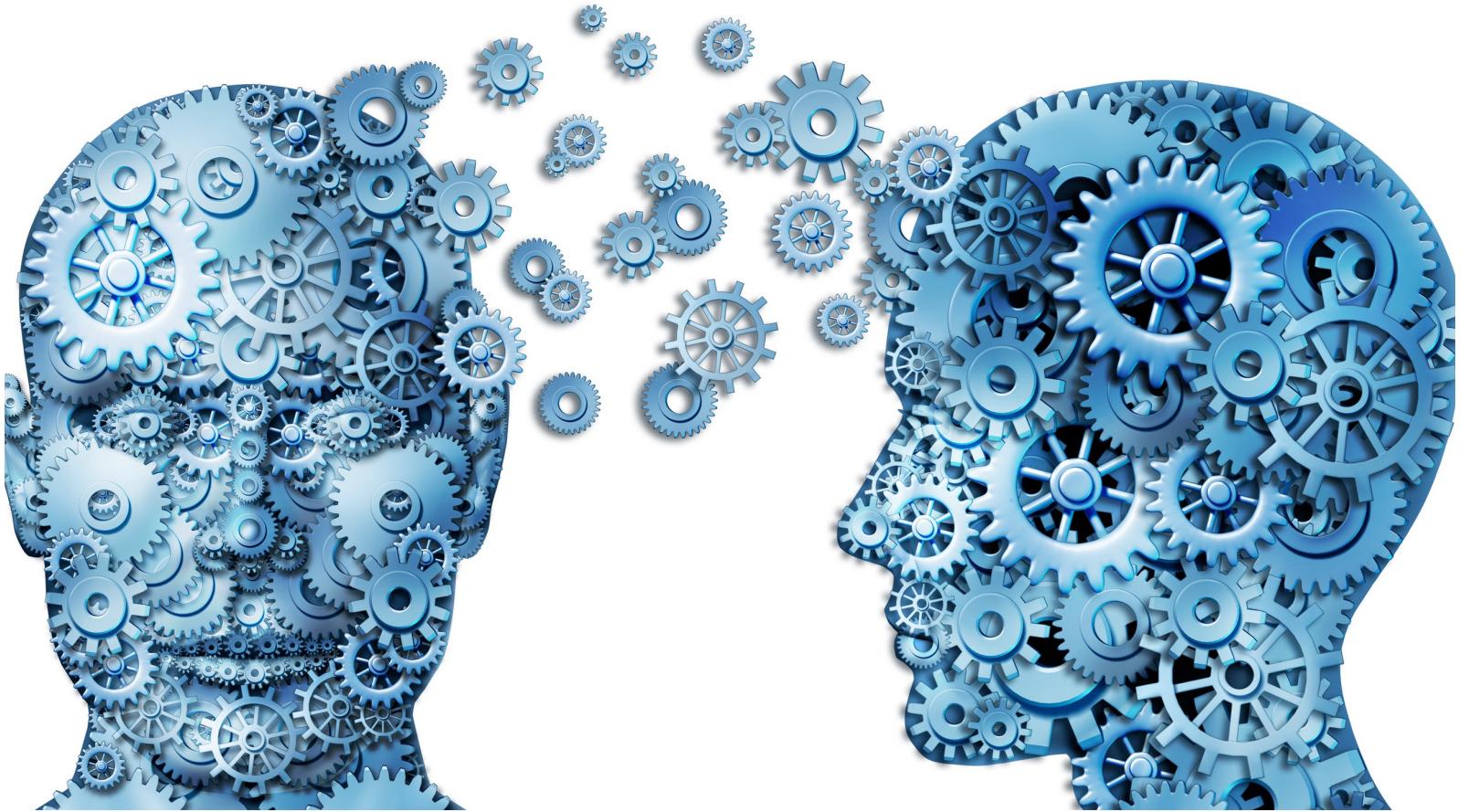
****Important Info****

- I'm here to help you succeed
- Please come to office hours or reach out if you need any additional support



Now the good stuff

What is
machine
learning?



What is
~~machine~~
learning?

learn·ing

/'lərnɪNG/ 

noun

the acquisition of knowledge or skills through experience, study, or by being taught.

"these children experienced difficulties in learning"

synonyms: study, studying, education, schooling, tuition, teaching, academic work; research

"a center of learning"

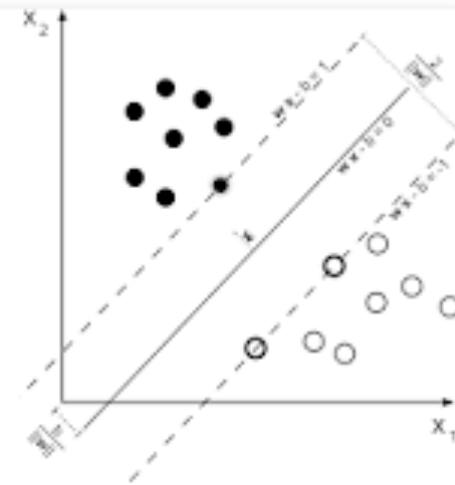


Translations, word origin, and more definitions

Machine learning: Wikipedia

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. **Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.**

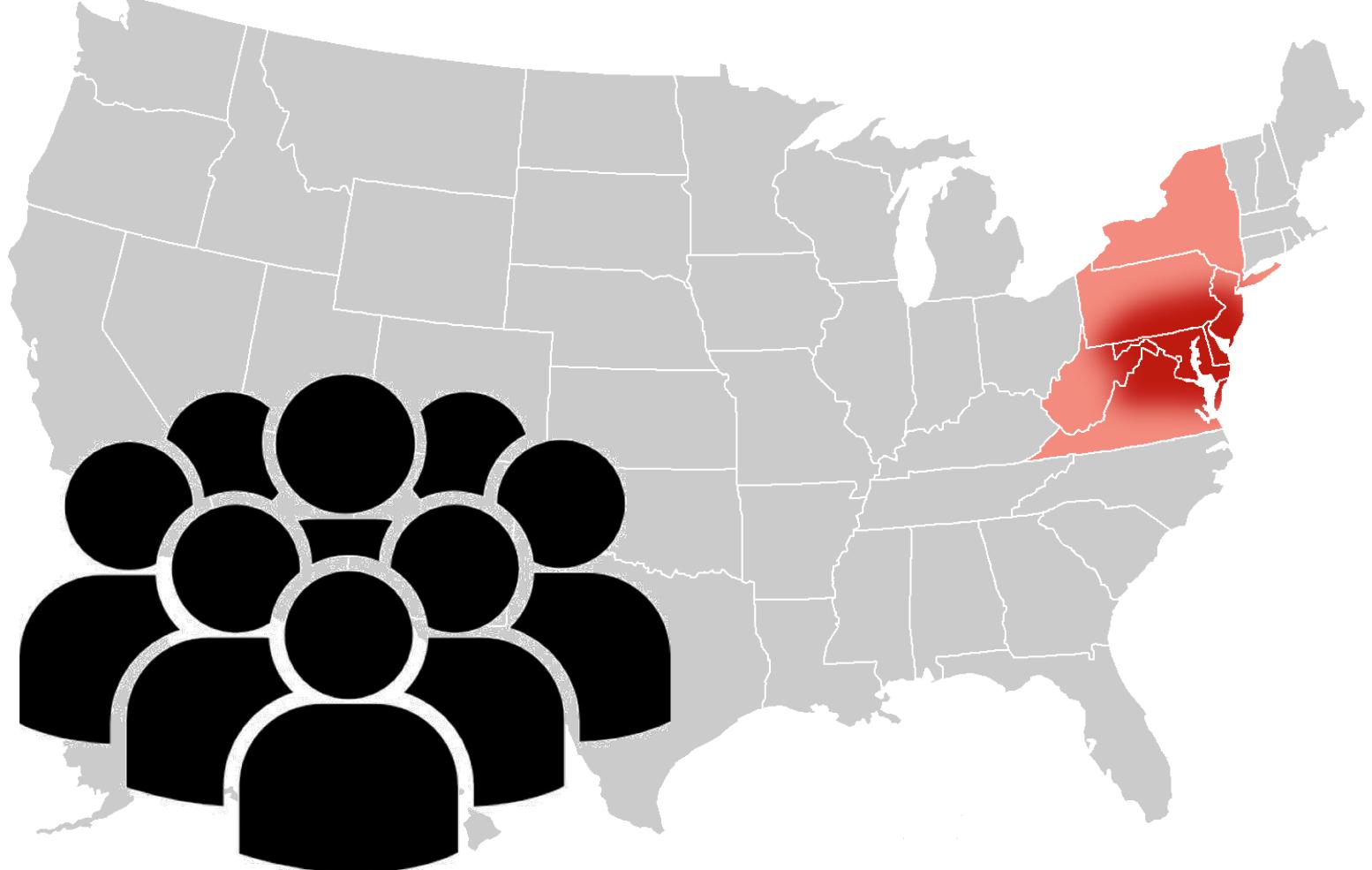
[Machine learning - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Machine_learning)
https://en.wikipedia.org/wiki/Machine_learning Wikipedia ▾



Machine learning: a working definition

- Machine learning is a set of **computational tools** for building **statistical models**
- These models can be used to:
 - **Group** similar data points together (*clustering*)
 - **Assign** new data points to the correct group (*classification*)
 - Identify the **relationships** between variables (*regression*)
 - Draw conclusions about the **population** (*density estimation*)
 - Figure out **which variables** are important (*dimension reduction*)

Example: men
& money in the
mid-Atlantic



Example: men & money in the mid-Atlantic

- `Wage` dataset available in the `ISLR` package
- **Sample:** 3000 male earners from the mid-Atlantic, surveyed between 2003 and 2009
- Dimensions:
 - Year each datapoint was collected
 - Age of respondent
 - Martial status
 - Race
 - Educational attainment
 - Job class
 - Health
 - Whether or not they have health insurance
 - Wage

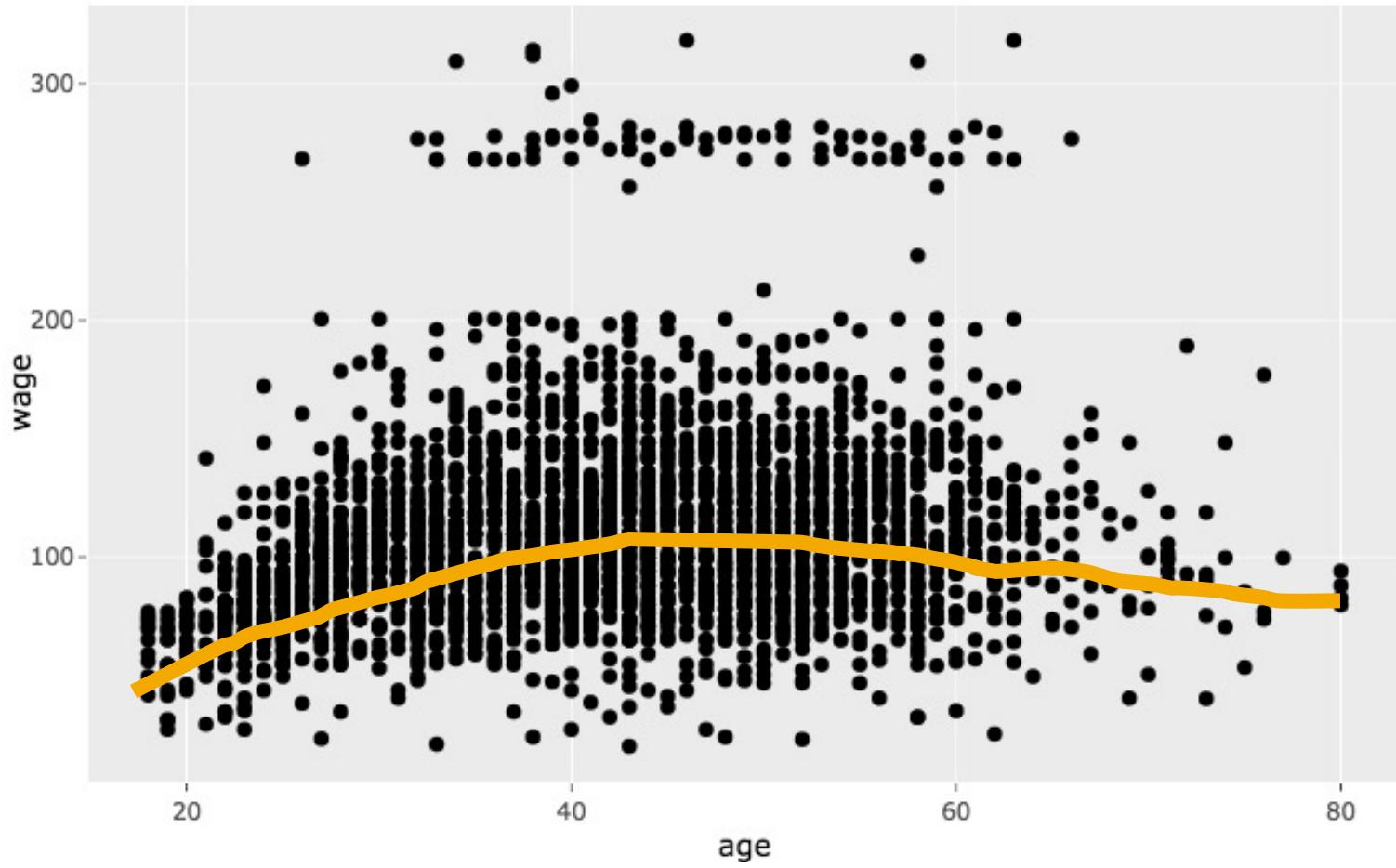
Example: men & money in the mid-Atlantic

- **Question:** what is the effect of an earner's `age`, `education`, and the `year` on his `wage`?
- Find some friends, then go explore the data at:
<https://amoscao1.github.io/CAIS380ML-S24/> under the Example tab

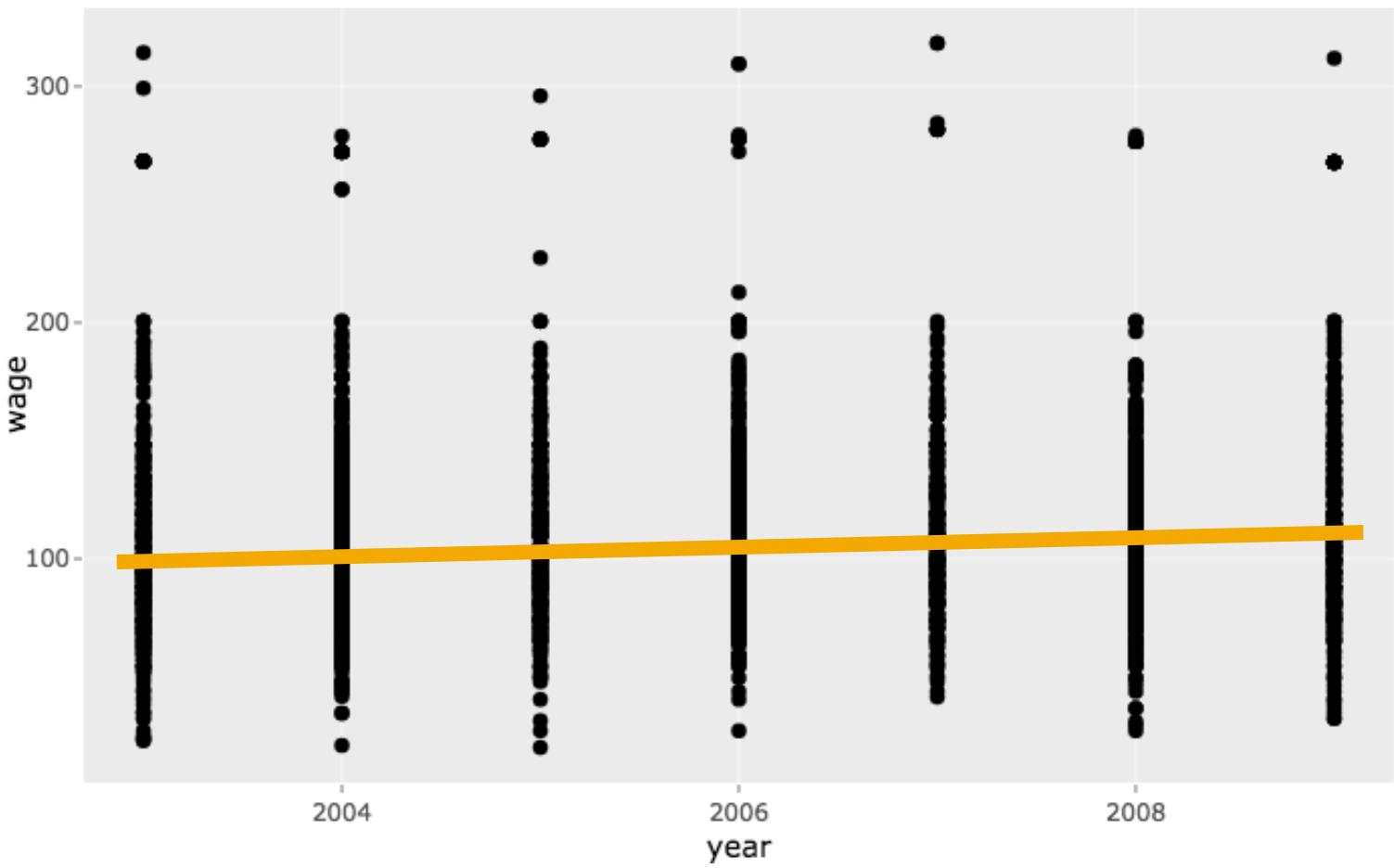
Example: men
& money in the
mid-Atlantic

<https://amoscao1.github.io/CAIS380ML-S24/examples/Wage.html>

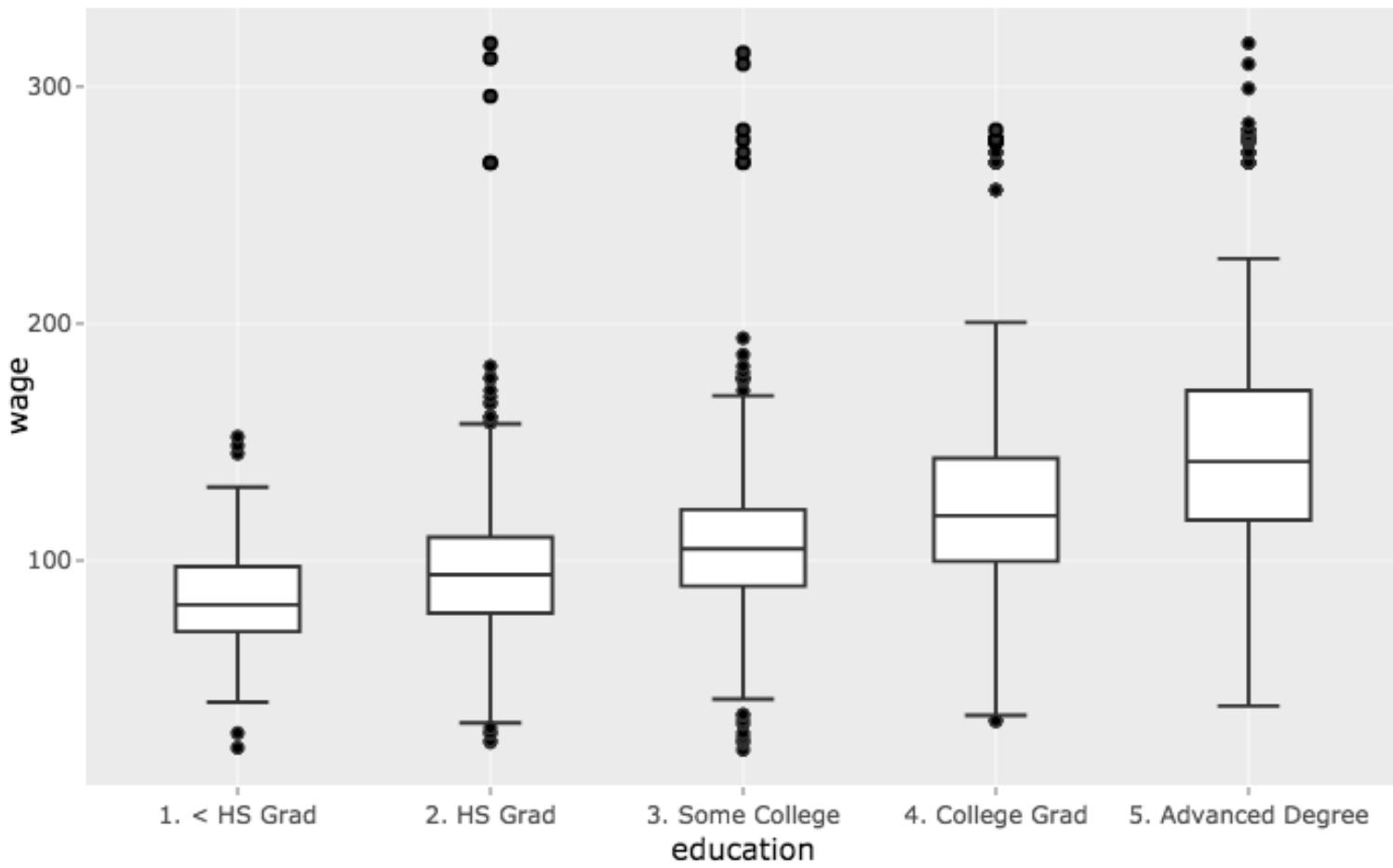
wage vs. age



wage vs.
year

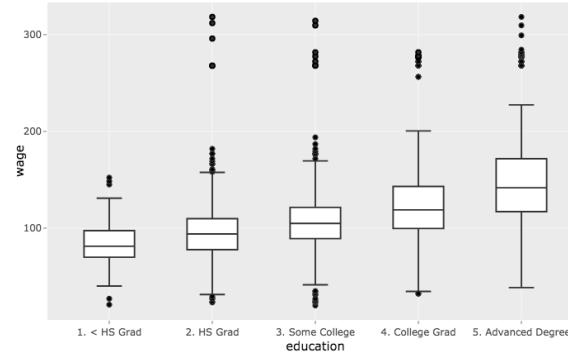


wage vs. education

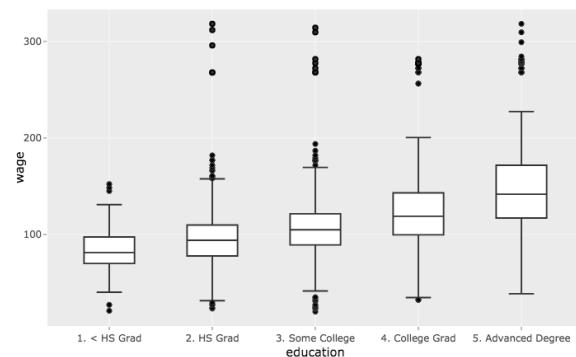
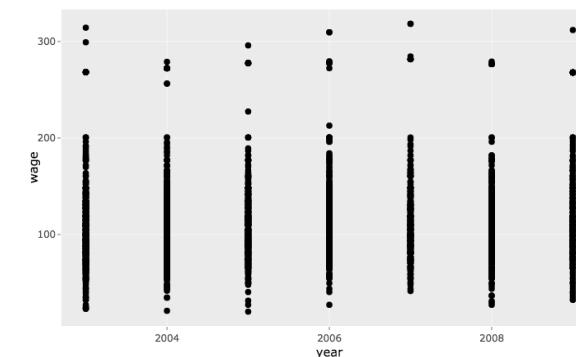
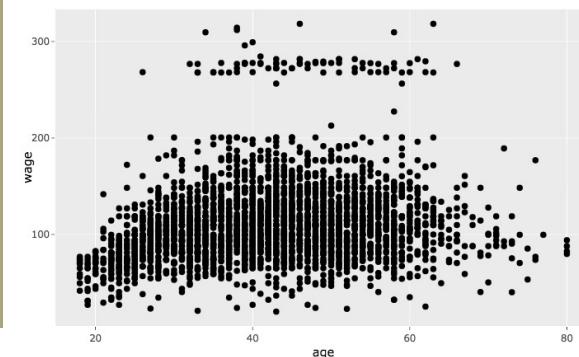


Example: men & money in the mid-Atlantic

- If we had to pick **just one**, we should probably use **education**



- In reality, the **best predictor** is probably a **combination** of all three



Supervised machine learning

- In this example, we used the value of **input variables** to predict the value of **output variables**
- Another way to think about this:



Supervised machine learning

- **Goal:** explain some observable phenomenon Y as a function of some set of predictors X :

$$Y = f(X) + \epsilon$$

- **Problem:** we don't know what the function actually looks like; we have to *estimate* it
- **Machine learning:** computational tools for estimating f

Unsupervised machine learning

- We sometimes have only **input variables**, but no clearly defined “response”
- Can’t check (“supervise”) our analysis: **unsupervised**
- Can’t fit a regression model (why?)
- What **can** we do?

Example:
personalized
marketing



Example:
personalized
marketing



Example: personalized marketing



Recommended for You

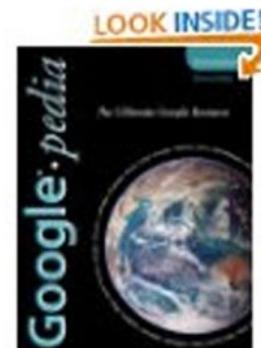
Amazon.com has new recommendations for you based on items you purchased or told us you own.



[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)



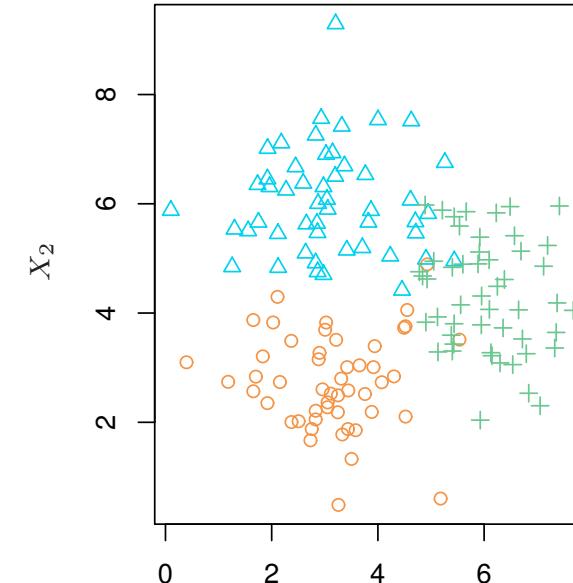
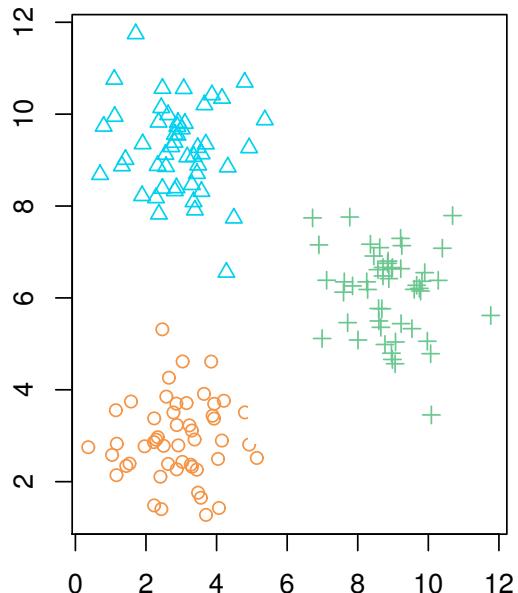
[Google Apps Administrator Guide: A Private-Label Web Workspace](#)



[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)

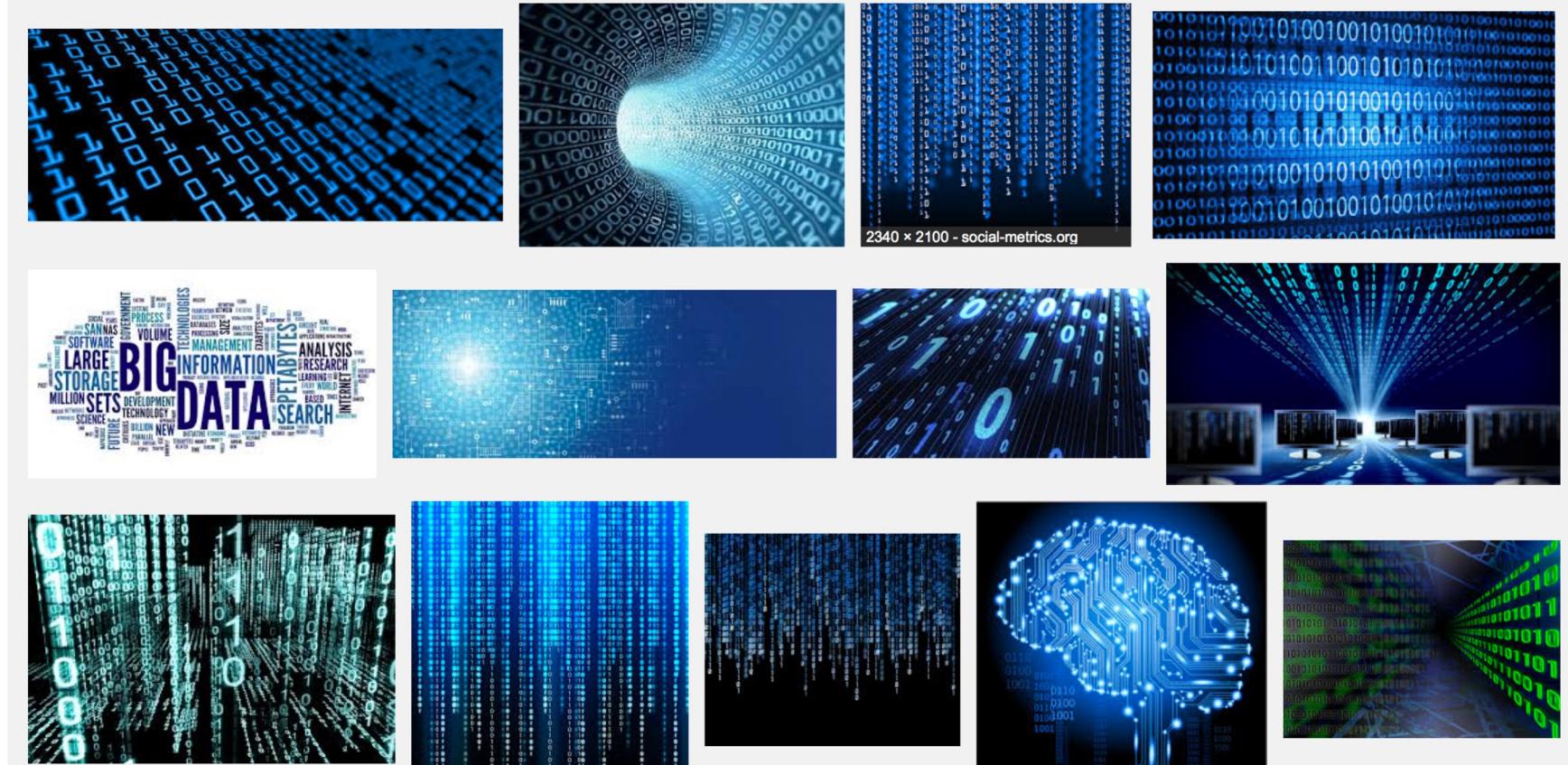
Unsupervised machine learning

- **Challenge:** identify whether the data separates into (relatively) distinct groups



- This kind of problem is called **cluster analysis** (Ch. 10)

Data science refresher: what is “data”?



Data: a definition

A dataset has some set of *variables* available for making predictions. For example:



*Tuition rates, enrollment numbers,
public vs. private, etc.*

Data: a definition

Each variable may be either *independent* or *dependent*:

- An *independent variable (iv)* is not controlled or affected by another variable (e.g., time in a time-series dataset)
- A *dependent variable (dv)* is affected by a variation in one or more associated independent variables (e.g., temperature in a region)

Data: a definition

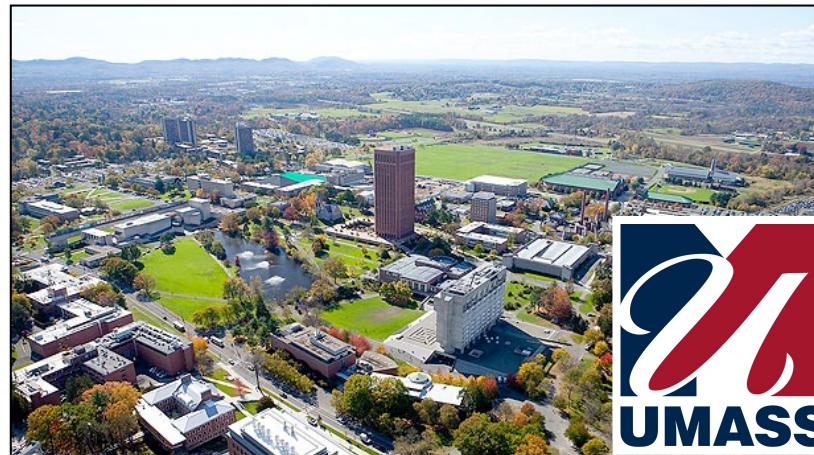
A dataset also contains a set of **observations** (also called *records*) over these variables. For example:



*tuition = \$26,978, enrollment = 4,239,
public, etc.*

Data: a definition

A dataset also contains a set of **observations** (also called *records*) over these variables. For example:



*tuition = \$32,434,
enrollment = 28,635,
public, etc.*

*tuition = \$78,438,
enrollment = 2,600
private, etc.*



One way to
think about
this:

| VARIABLES | | | | |
|--------------|--|--|--|--|
| OBSERVATIONS | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

Another way
to think about
this

```
class school_obs:  
    def __init__(tuition, enrollment,  
                 pub_or_priv):  
        self.tuition = tuition  
        self.enrollment = enrollment  
        self.pub_or_priv = pub_or_priv
```

OBSERVATIONS

VARIABLES

Basic data types

- Nominal
- Ordinal
- Scale / Quantitative
 - Ratio
 - Interval

An **unordered** set {...}
of non-numeric values

For example:

- Categorical (finite) data
{apple, orange, pear}
{red, green, blue}
- Arbitrary (infinite) data
{"12 Main St. Boston MA", "45 Wall St. New York NY", ...}
{"John Smith", "Jane Doe", ...}

Basic data types

- Nominal
- Ordinal
- Scale / Quantitative
 - Ratio
 - Interval

An **ordered set** <...>
(also known as a tuple)

For example:

- Numeric: <2, 4, 6, 8>
- Binary: <0, 1>
- Non-numeric:
<G, PG, PG-13, R>

Basic data types

- Nominal
- Ordinal
- Scale / Quantitative
 - Ratio
 - Interval

A numeric range

[...]

Ratios

- Distance from “absolute zero”
- Can be compared mathematically using division
- For example: height, weight

Intervals

- Ordered numeric elements that can be mathematically manipulated, but cannot be compared as ratios
- E.g.: date, current time

Converting between basic data types

- Q → O $[0, 100] \rightarrow \langle F, D, C, B, A \rangle$
- O → N $\langle F, D, C, B, A \rangle \rightarrow \{C, B, F, D, A\}$
- N → O (?)
 - $\{John, Mike, Bob\} \rightarrow \langle Bob, John, Mike \rangle$
 - $\{red, green, blue\} \rightarrow \langle blue, green, red \rangle$
- O → Q (?)
 - Hashing?
 - $Bob + John = ??$

Discussion: what do you notice?

Basic operations

- Nominal (N)
 - Equality: = and ≠
 - Frequency: how often does x appear?
- Ordinal (O)
 - Relation to other points: $>$, $<$, \geq , \leq
 - Distribution: inference on relative frequency
- Quantitative (Q)
 - Other mathematical operations: $(+, -, \times, /, \text{etc.})$
 - Descriptive statistics: *average, standard deviation, etc.*

(Hopefully) familiar statistical concepts

- We tend to refer to problems with a **quantitative** response as *regression* problems
- When the response is **qualitative** (i.e. nominal or ordinal), we're usually talking about a *classification* problem
- **Caveat:** the distinction isn't always that crisp. For example:
 - K-nearest neighbors (Ch. 2 and Ch. 4), which works with either
 - Logistic regression (Ch. 4), which estimates the probabilities of a qualitative response

What we'll cover in this class

- Ch. 2: Statistical Learning Overview (next class)
- Ch. 3: Linear Regression
- Ch. 4: Classification
- Ch. 5: Resampling Methods
- Ch. 6: Linear Model Selection
- Ch. 7: Beyond Linearity
- Ch. 8: Tree-Based Methods
- Ch. 9: Support Vector Machines
- Ch. 10: Unsupervised Learning

Backgrounds

- Who has experience coding?
- Who has experience with statistics?

Preparing for labs in R



You can install R Studio on your own
machine: rstudio.com

Preparing for
labs in python



You can install Anaconda here:
<https://www.anaconda.com/download>

I recommend developing in Spyder (which
comes with the anaconda download)

You'll need to know how to **install packages**

What I expect from you

- You like difficult problems and you're excited about "**figuring stuff out**"
- You have a solid foundation in **introductory statistics**
- You are proficient in **coding and debugging** (or are ready to work to get there)
- You're comfortable asking **questions**