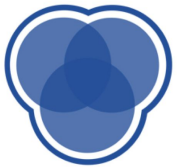
INFO 1998: Introduction to Machine Learning

(Please download the Lecture 4 Demo on the Course Website!)



CDS Education

We explore, learn, and educate big minds.

Lecture 4: Fundamentals of Machine Learning Pt. 1

INFO 1998: Introduction to Machine Learning

Introduction to Machine Learning and Tools

CDS Education

We explore, learn, and educate big minds.

Project

- Start thinking about what datasets and questions you want to explore
- If you still need a partner, post on Ed or stay after class
- Rubric can be found on our website under “Final Project”

Project Check-in

- We'll have a check-in in about 2-3 weeks.
  - Expecting hypothesis/question/problem to solve
  - Chosen dataset
  - Some progress on data cleaning/data visualization

- Come to OH if you need help or if there's a problem

What We'll Cover

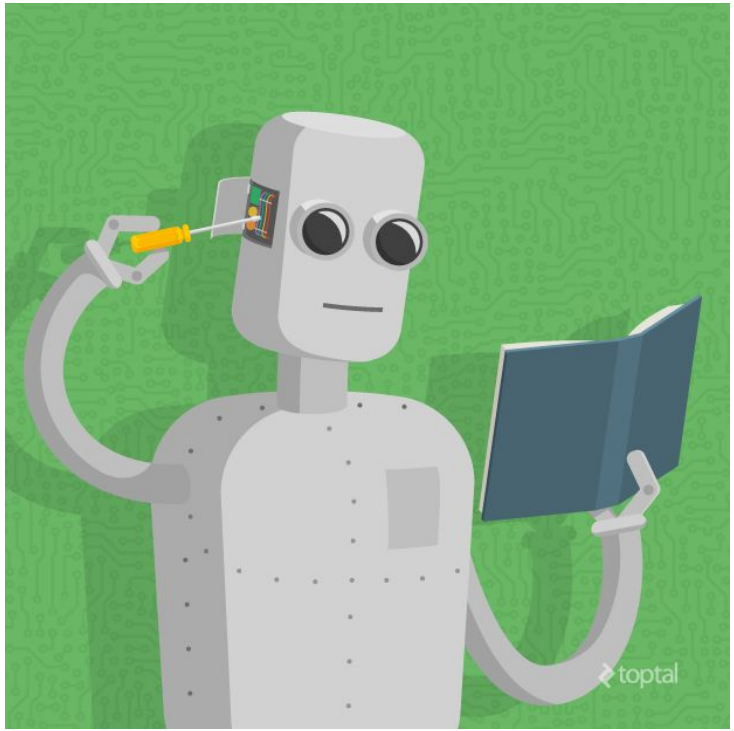
Today's Goal: be able to write code to do some kind of ML (to some extent)

- **Define Machine Learning:** or like, 5 definitions
- **Start learning the language of ML:** There's a lot of terminology 😢
- **Try Linear Regression (via ScikitLearn)**: Our first ML algorithm!
- **Introduce our Workflow**: An outline for developing an ML model
- **Discuss Some Important Considerations**: What should we be thinking about as we're MLing?

Agenda

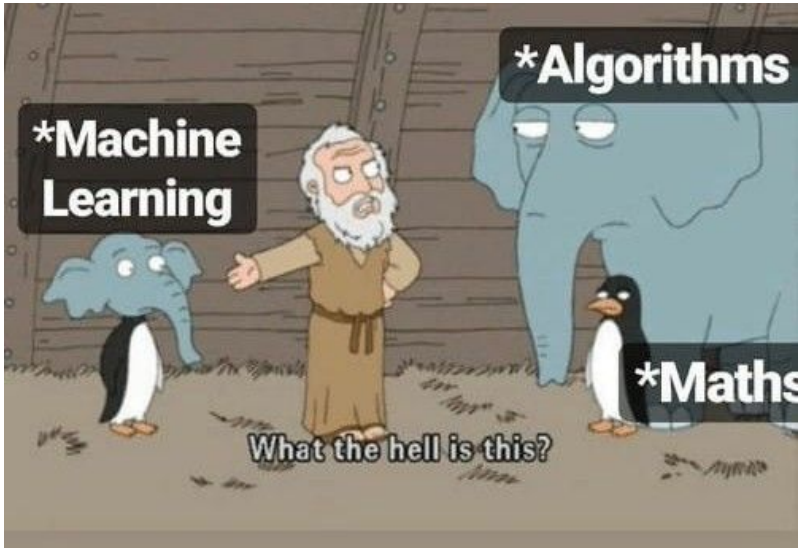
1. What does a Machine Learning Engineer do?
2. On a high level, how do you define “Machine Learning”?
3. What’s a Machine Learning Model?
4. What’s a *good* Machine Learning Model?

What's Machine Learning? Part 1: what does an ML engineer do



Machine Learning can involve:

- Preprocessing data
- Splitting and selecting pieces of data
- Doing mathematical analysis on the data
- Deciding what data structures are needed to efficiently implement algorithms
- Implementing accuracy metrics
- ...and a lot more



How do *we* do machine learning?

What we’re gonna do:

Write as little code as possible!

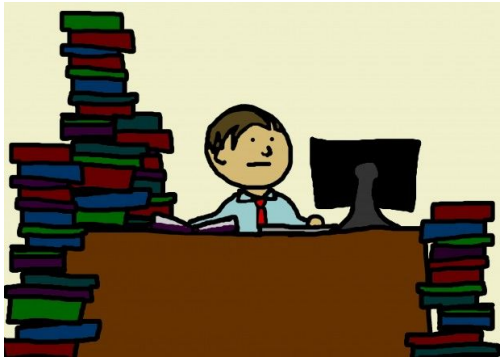
- Use pandas to deal with data
- Use numpy to do math
- Use scikit-learn (“sklearn”) to make ML models (and other useful stuff)

What we’re gonna do:

Our main tasks:

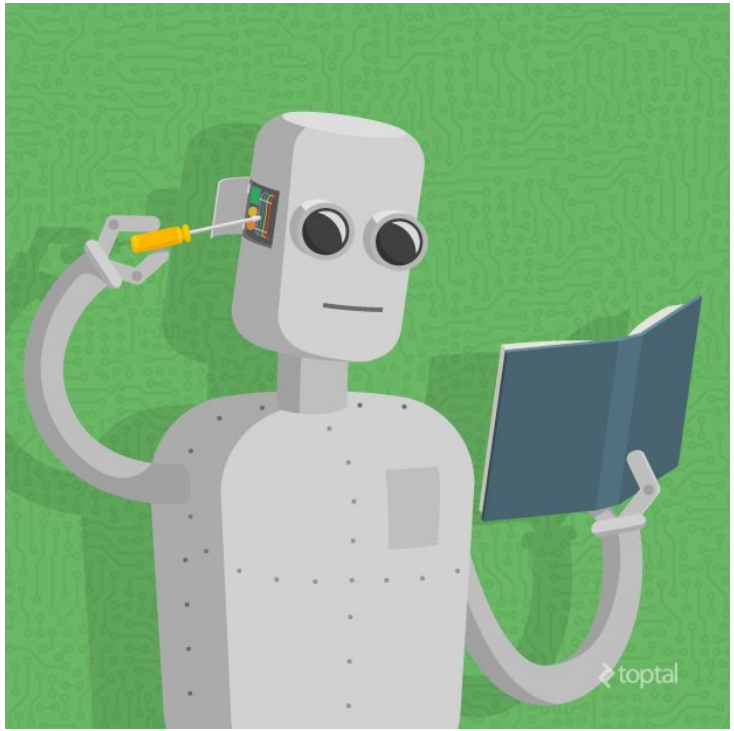
- Choose an algorithm
- Choose how to use different parts of the data
- Find which pandas, numpy, and scikit-learn functions do what we want
- Interpret the results and fine-tune our model

Quick analogy: studying



- Setup
  - Goal: Be able to solve the test problems
  - Resources: Practice problems + answers
- Method
  - You study those practice problems and answers. Given a problem, how do you get the answer?
- Result:
  - On the real test, the problems aren't the exact same as the practice problems. But they're similar!
  - Since you learned generally how to solve the practice problems, you can solve the similar test problems too :)

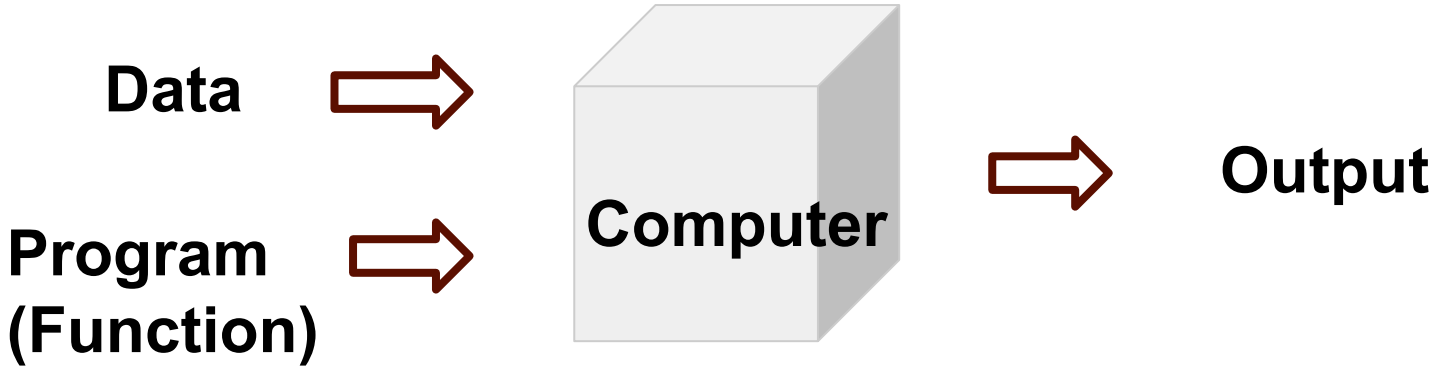
What's Machine Learning? Part 2: like seriously what is it



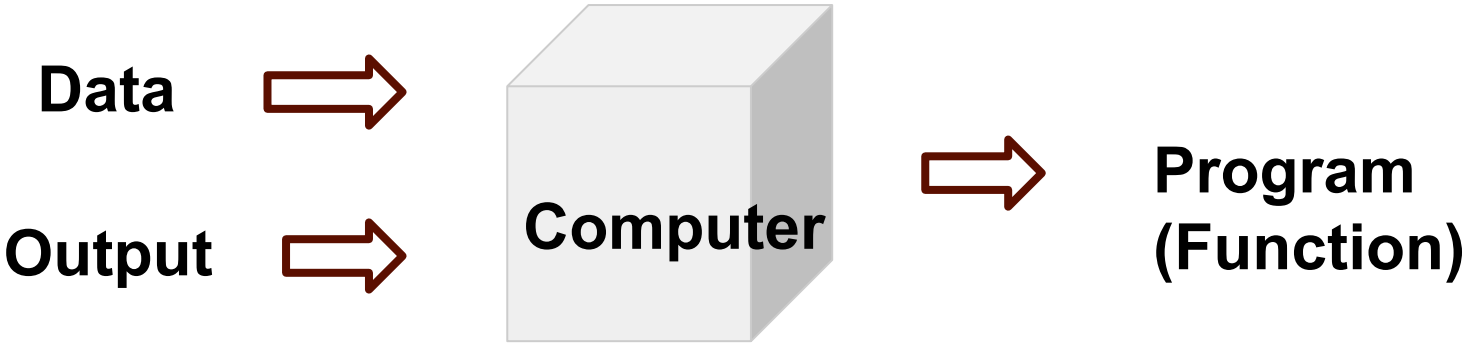
Some Definitions of ML

- Give computers the ability to learn without being explicitly programmed
  (^ that one's a pretty sucky definition)
- Build a useful mathematical model, based on sample data, to make inferences
- Take in data and make predictions or decisions
- Help your computer learn patterns

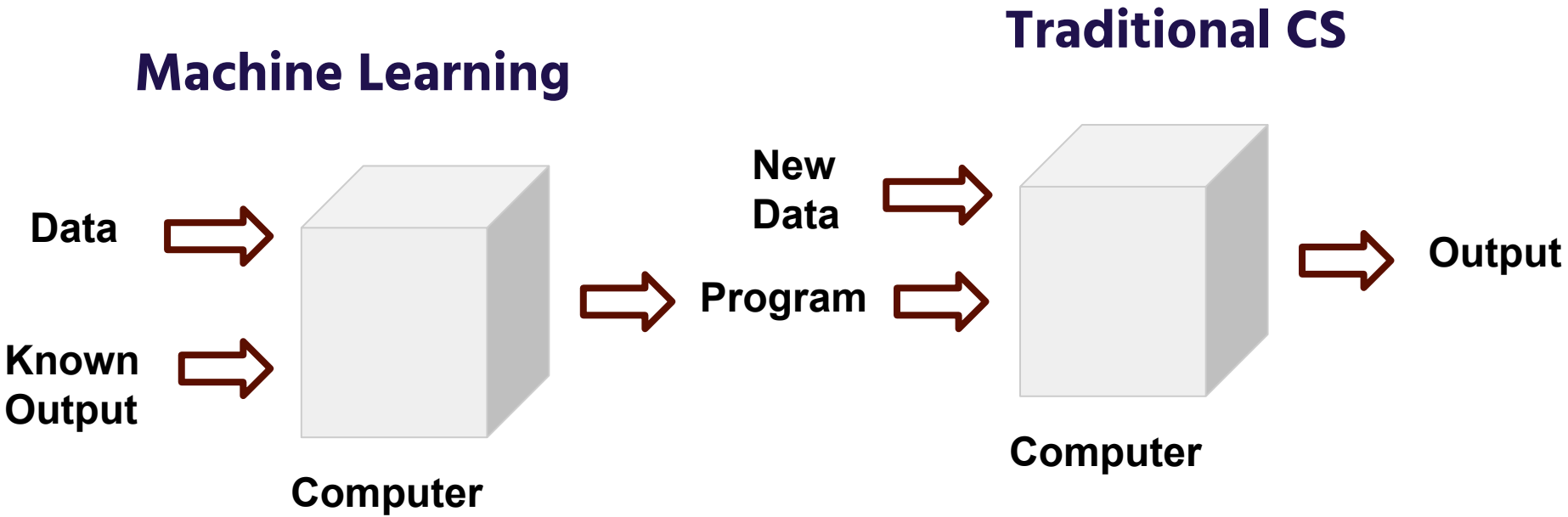
Traditional Computer Science



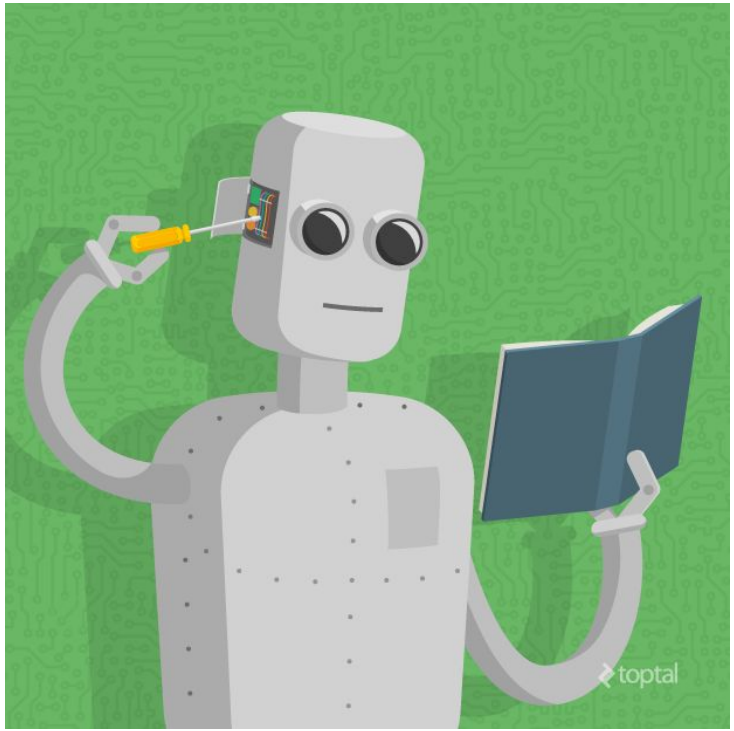
Machine Learning



Using Machine Learning



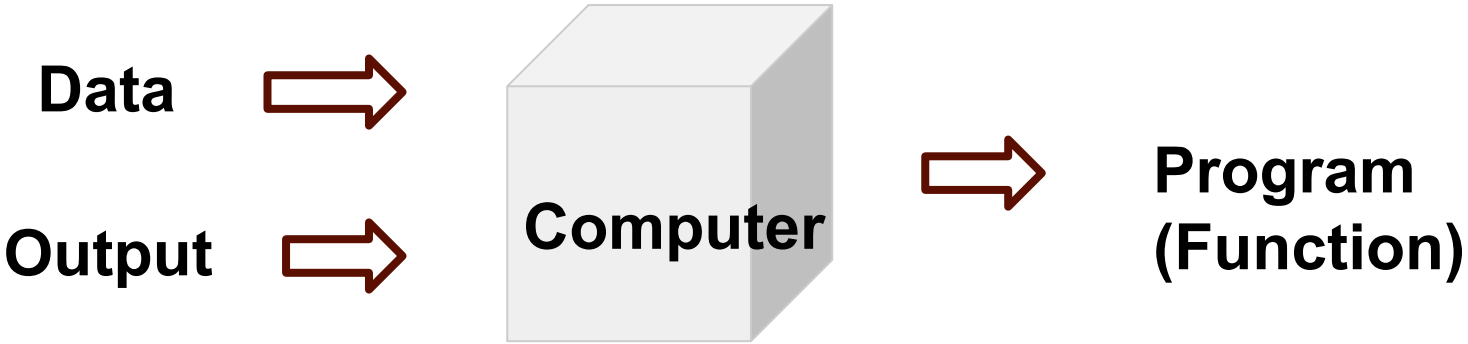
What's Machine Learning? Part 3: what's a model?



What's a model?

1. The output of a machine learning algorithm
2. A procedure to produce some outputs when given some inputs
3. A relationship between inputs and outputs
4. A guess at how inputs and outputs are related
5. A set of assumptions we're imposing on the dataset
6. A configurable thing (hyperparameters)

ML Algorithm produces a Model



Review: Dataset Structure

- rows are data points
  - aka samples
- columns are features
  - a sample is made of lots of features, including the goal

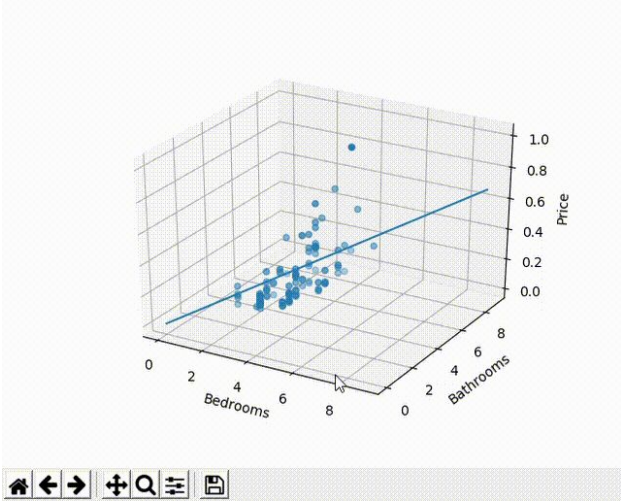
	Name	Age	Major
0	Ann	19	Computer Science
1	Chris	20	Sociology
2	Dylan	19	Computer Science
3	Camilo	NaN	NaN
4	Tanmay	NaN	NaN

Linear Regression

Linear Regression

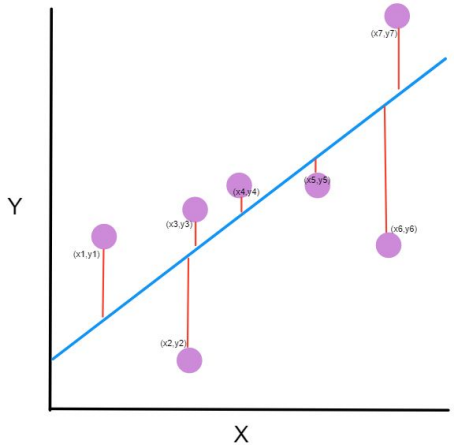
$$y = B_0 + B_1 x_1 + \ldots + B_p x_p + \varepsilon$$

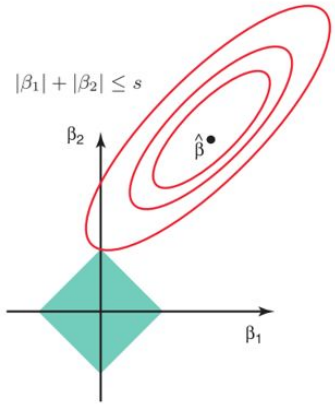
- x is an input; $x_1, x_2, \ldots, x_p$ are the features of x
- y is an output (usually a single value)
- B's are "weights"
  - A linear regression equation is defined by its B's
  - This linear regression equation is the "program" produced by ML
- Given a set of x's and y's, the program finds a set of B's that (almost) satisfy the equation above for all x's and y's
  - Then, you can plug in the feature values of a new x and to predict its y



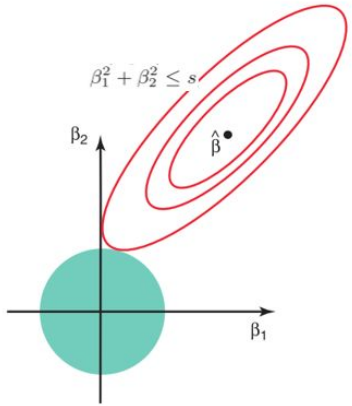
Linear Regression: Ordinary Least Squares

- There are different types of linear regression algorithms
- We are using *ordinary least squares*
- This calculates the weight vector B by minimizing the **mean-squared error** of the predicted y-values
- There are other types of linear regression such as ridge regression, which use different *loss functions* to calculate the weights





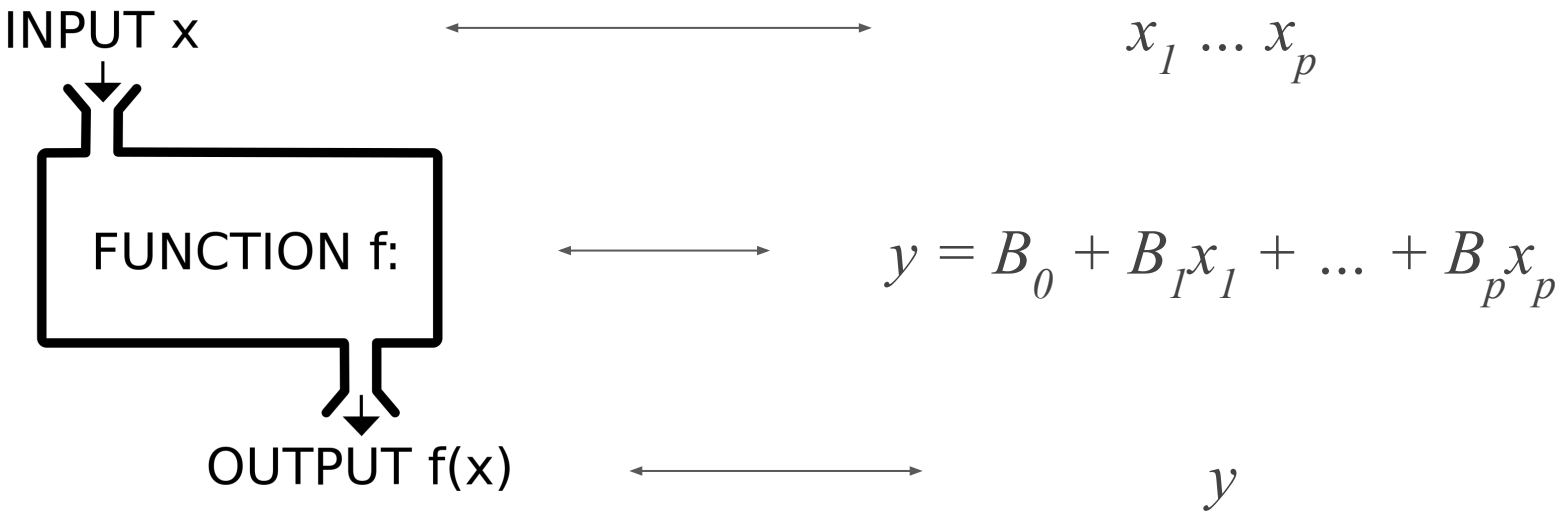
Lasso Regression



Ridge Regression

Linear Regression

Function **Weighted Sum**



$y = B_0 + B_1 x_1 + \ldots$ is a model

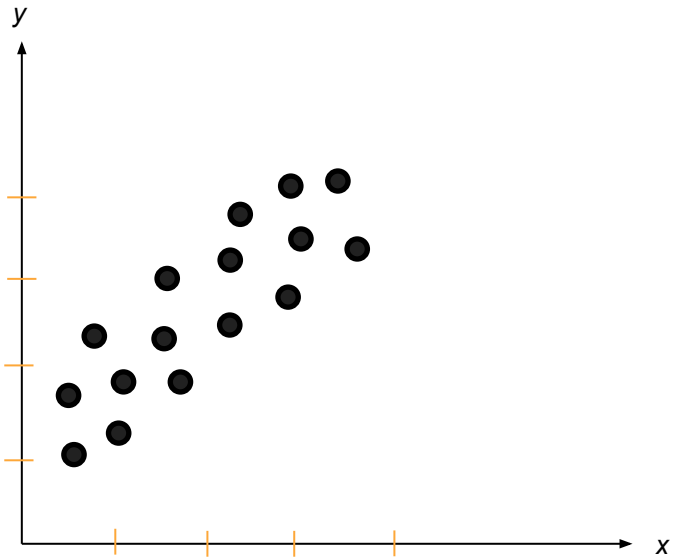
- A relationship between inputs and outputs

  $y = B_0 + B_1 x_1 + \ldots$ relates inputs to outputs
- A guess at how inputs and outputs are related

  but $y = B_0 + B_1 x_1 + \ldots$ is just a guess/estimate; it's not exactly true
- A set of assumptions we're imposing on the dataset

  We're assuming output is linearly related to input features
- A configurable thing (hyperparameters)

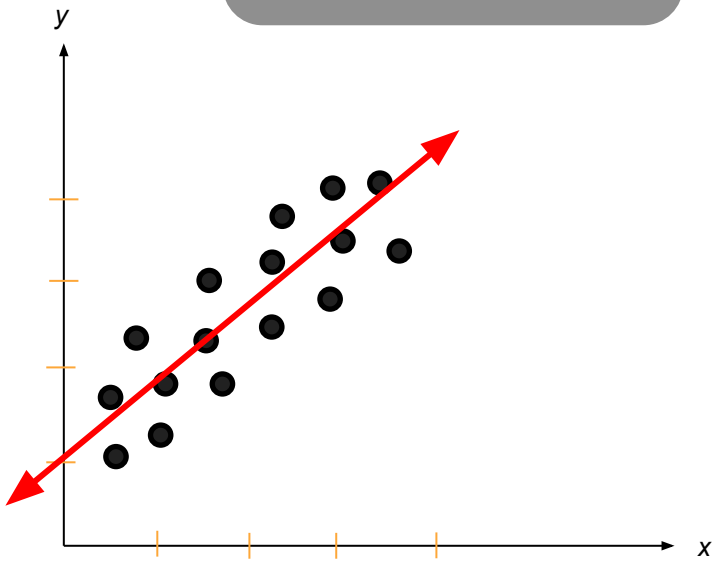
  Sorry, we don't cover very much linear regression configuration here 😅



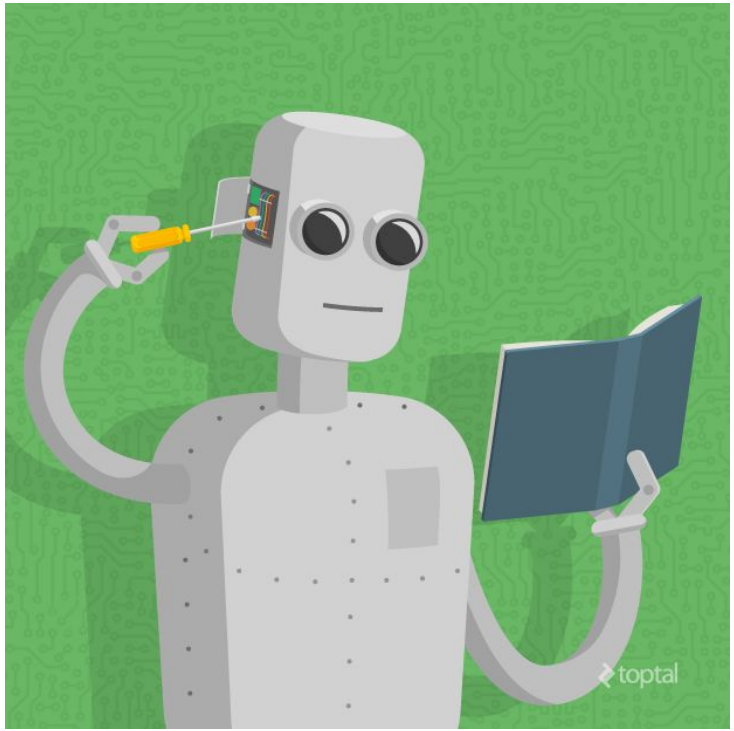
$y = B_0 + B_1 x_1 + \ldots$ is a model

Use algorithm to “learn” parameters that give us this line of best fit

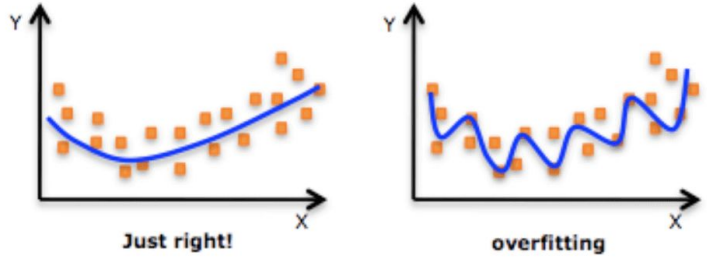
- **A relationship between inputs and outputs**
  $y = B_0 + B_1 x_1 + \ldots$ relates inputs to outputs
- **A guess at how inputs and outputs are related**
  but $y = B_0 + B_1 x_1 + \ldots$ is just a guess/estimate; it’s not exactly true
- **A set of assumptions we’re imposing on the dataset**
  We’re assuming output is linearly related to input features
- **A configurable thing (hyperparameters)**
  Sorry, we don’t cover very much linear regression configuration here 😅



What's Machine Learning? Part 4: What makes a *good* model?



Pitfall of Training: Overfitting



Model is accurate for **train** data

≠

Model can accurately predict **new** data

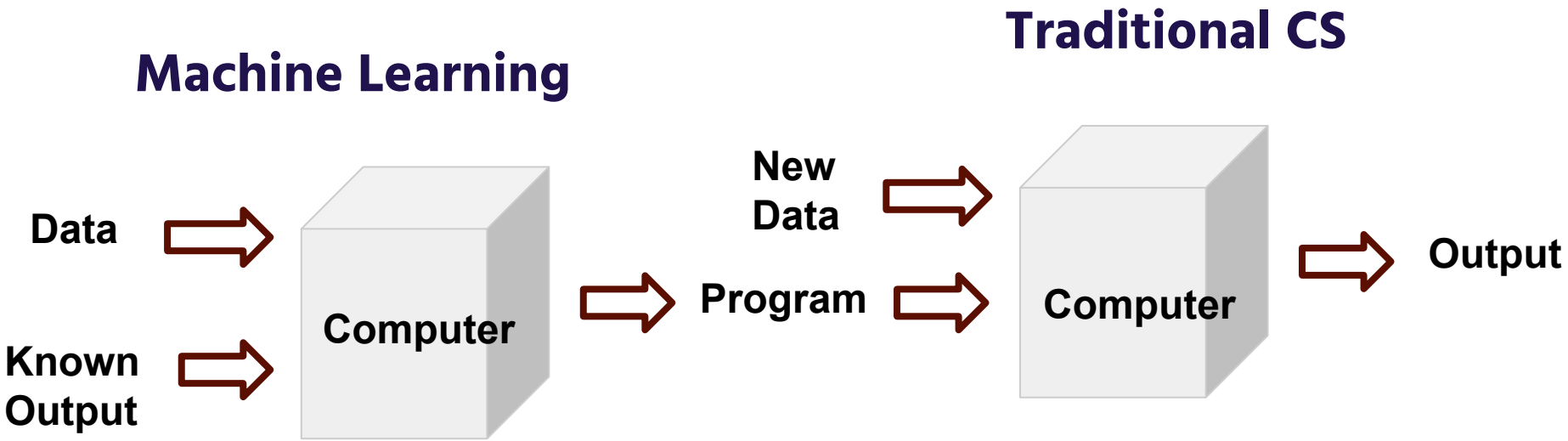
- We learned the specific mapping from **train input to train outputs…**
- But, we didn't learn the data's **general patterns** 🥲🥲🥲🥲🥲🥲

Solution: train on part of data, and check accuracy on a separate part of data (*validation* set)

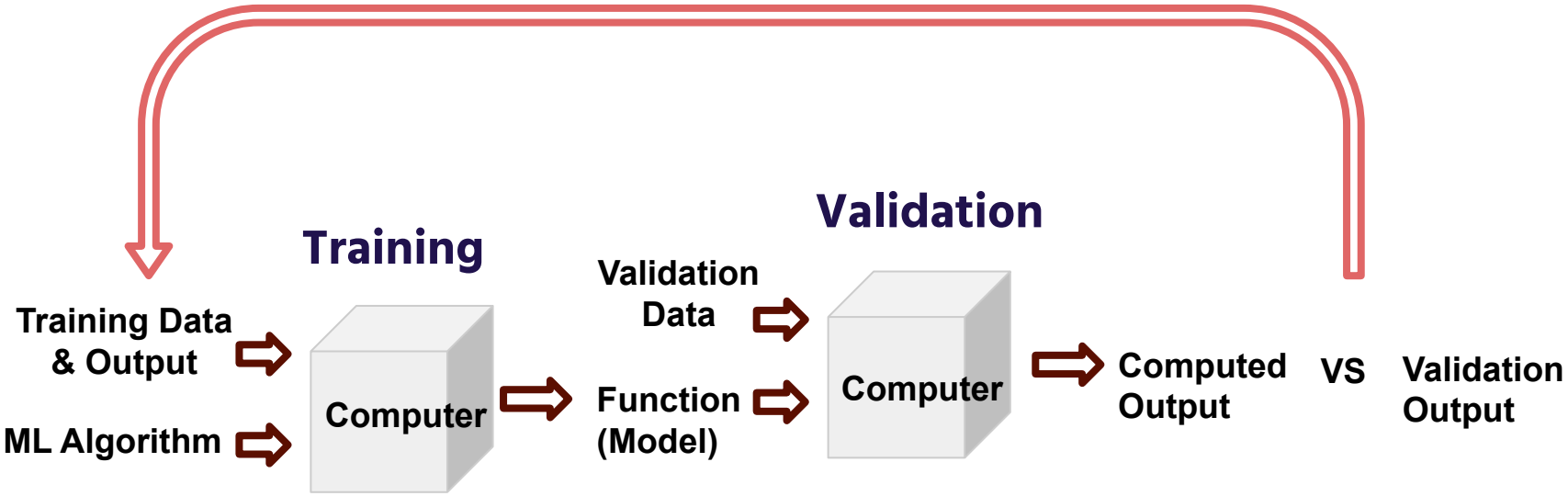
Terminology: Training and Validating

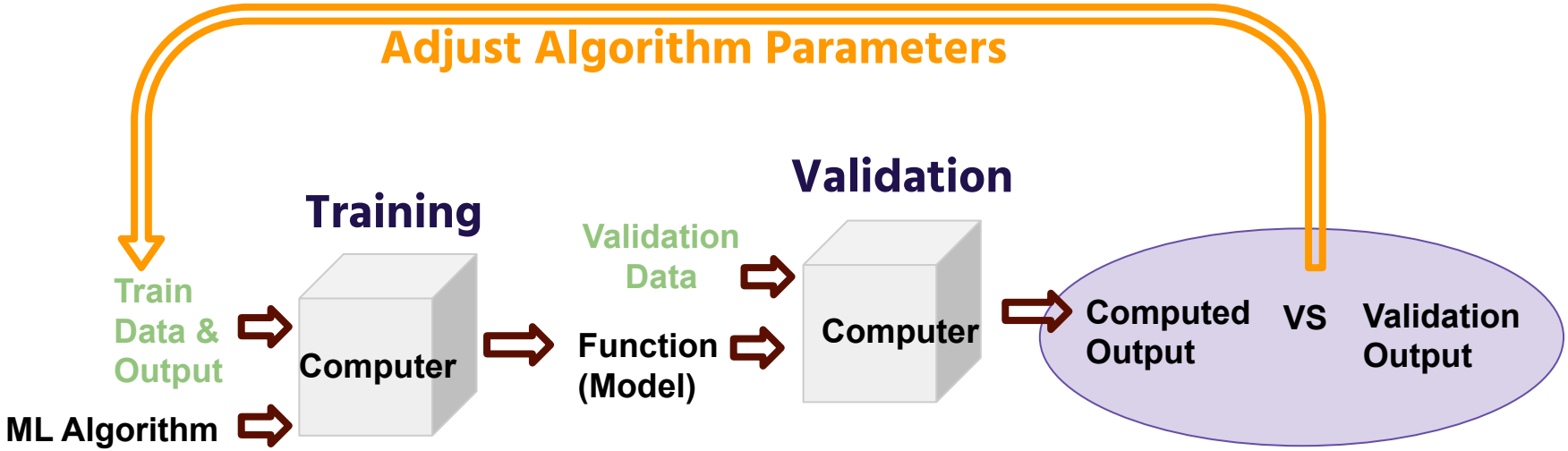
- Split data into two sets
- Train model on one, validate on the other
- “Model training” = learn a relationship/program
    - e.g. give the linear regression data so it can define the B’s
- “Model validation” = see if the learned relationship is accurate on other data

Machine Learning



Our ML Workflow





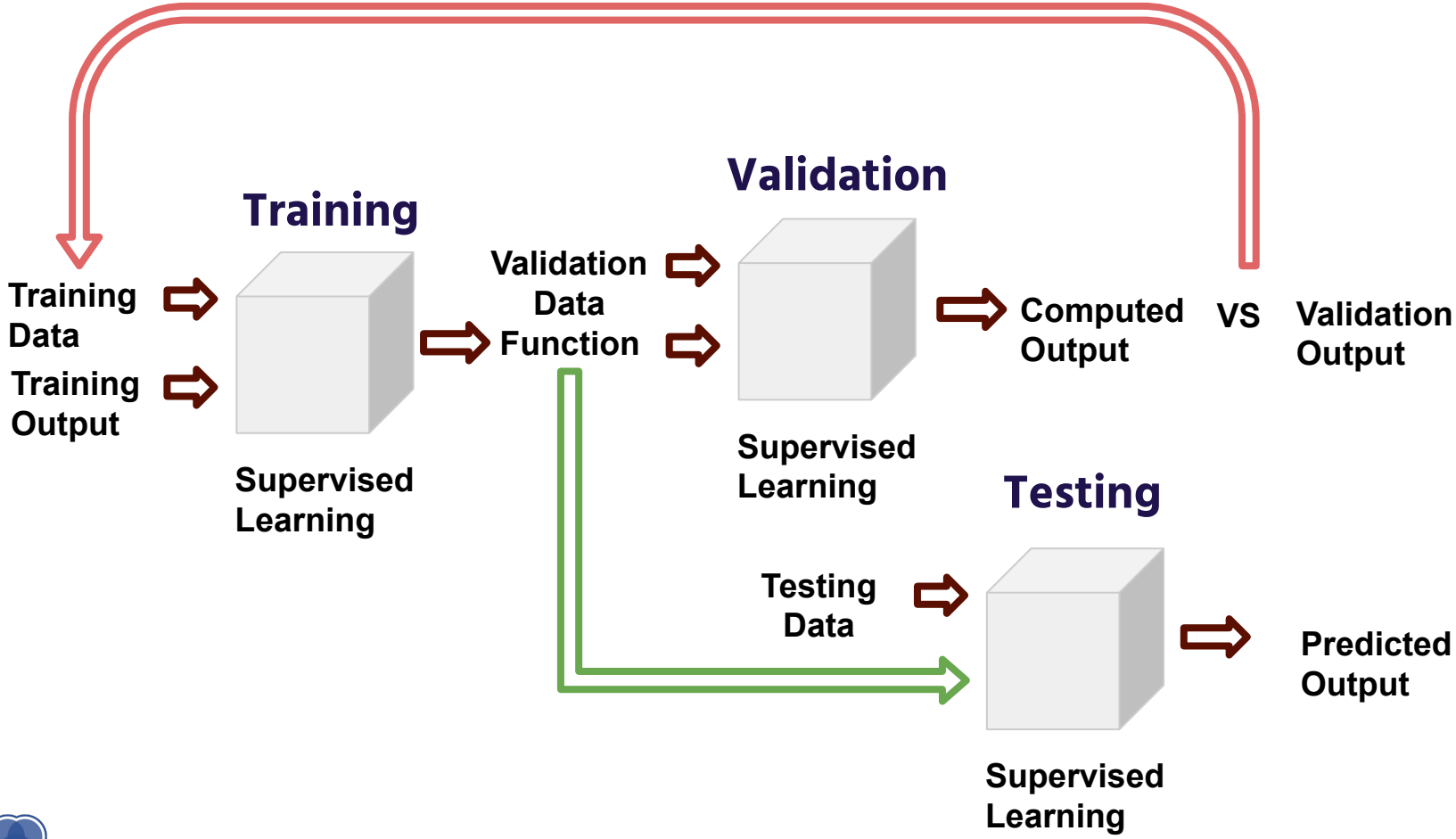
1. Select data
2. Assess model accuracy
3. Adjust Model

Pitfall of Validation: Overfitting

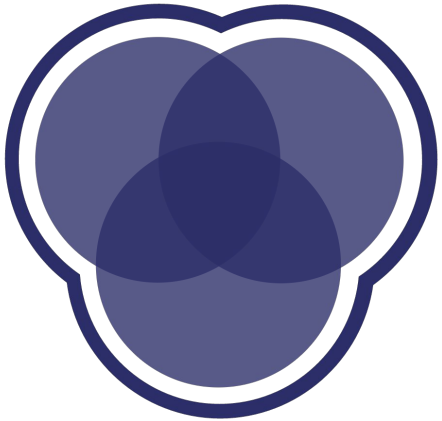
Predicting well on validation set ≠ Predicting well on new data

- We used the validation set to make our adjustments.
  - ⇒ Our model is **biased** to the validation set. 🥲🥲

Solution: keep a separate, rarely-used ***test*** set


Training
Training Data
Training Output
Supervised Learning
Validation
Validation Data Function
Supervised Learning
Computed Output
VS
Validation Output
Testing
Testing Data
Supervised Learning
Predicted Output

Demo



Model Goals

When training a model we want our models to:

- Capture the trends of the training data
- Generalize well to other samples of the population
- Be moderately interpretable

The first two are especially difficult to do simultaneously!
The more sensitive the model, the less generalizable and vice versa.

Putting things into perspective

- Linear Regression alone is weak, but it can be very strong when combined with feature selection and feature engineering.
- Linear Regression is just one algorithm — we'll cover many more! 🤠
- The "model" produced by an algorithm is not always a simple equation like in linear regression.
- Validation is *really* important.
  - Overfitting is a huge problem!
  - We'll delve deeper in the next few lectures...

Coming Up

Assignment 4: Due at 11:59pm EST on **October 13th**

Next Lecture: Assessing Model Accuracy + Fundamentals of ML

(a.k.a. *What's Machine Learning? Part* ∞)

TBD: Web Scraping Workshop 👀

CDS Education

We explore, learn, and educate big minds.