INFO 1998: Introduction to Machine Learning

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.

Midsemester Project

- Midsemester Project Released (Due 10/30)
- Can work in groups of 1-3
- Find a dataset, clean it, and make some visualizations
- At the end of lecture today, we will host a team finding session.

What We'll Cover

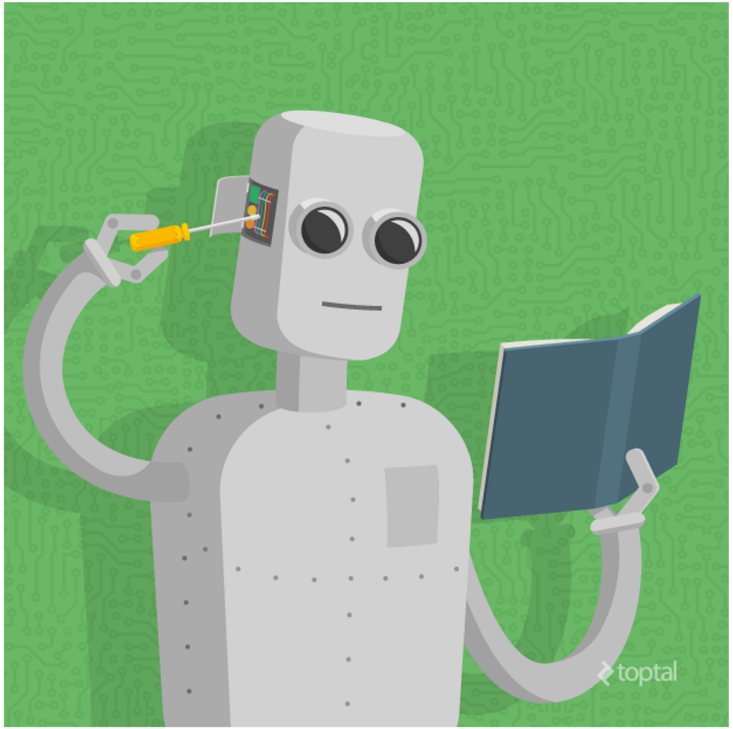
Today's Goal: be able to write code to do some kind of ML (to some extent)

- **Learn how to use ScikitLearn:** It's intimidating at first but you'll catch on quickly
- **Define Machine Learning:** or like, 5 definitions
- **Start learning the language of ML:** There's a lot of terminology :(
- **Try Linear Regression**: 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
- Manipulating those data structures as the algorithm indicates
- Optimizing for hardware infrastructure
- 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:

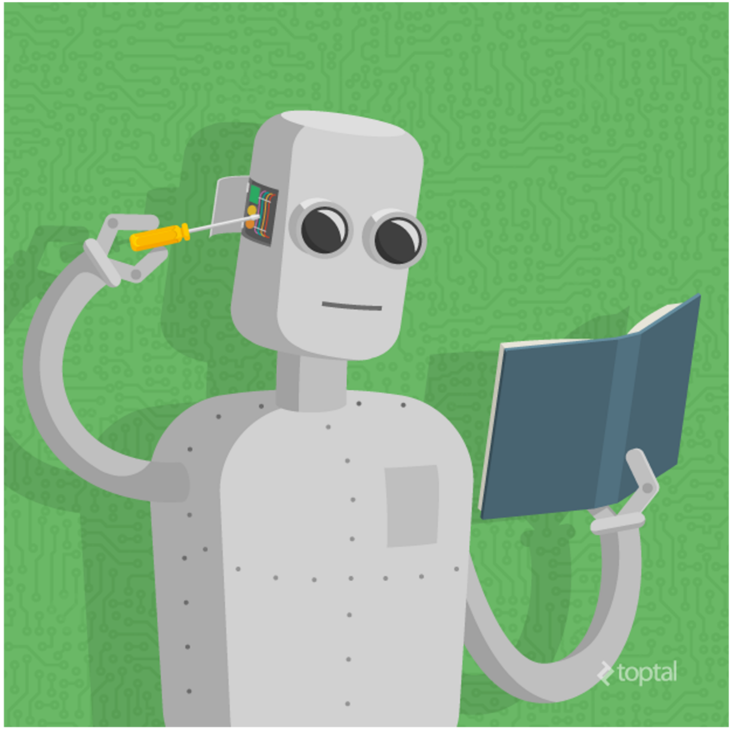
- Choose an algorithm
- Choose how to use different parts of the data
- Find which pandas, numpy, and scikit-learn functions do what we want
- Interpret the results and fine-tune our model

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 :)

Demo

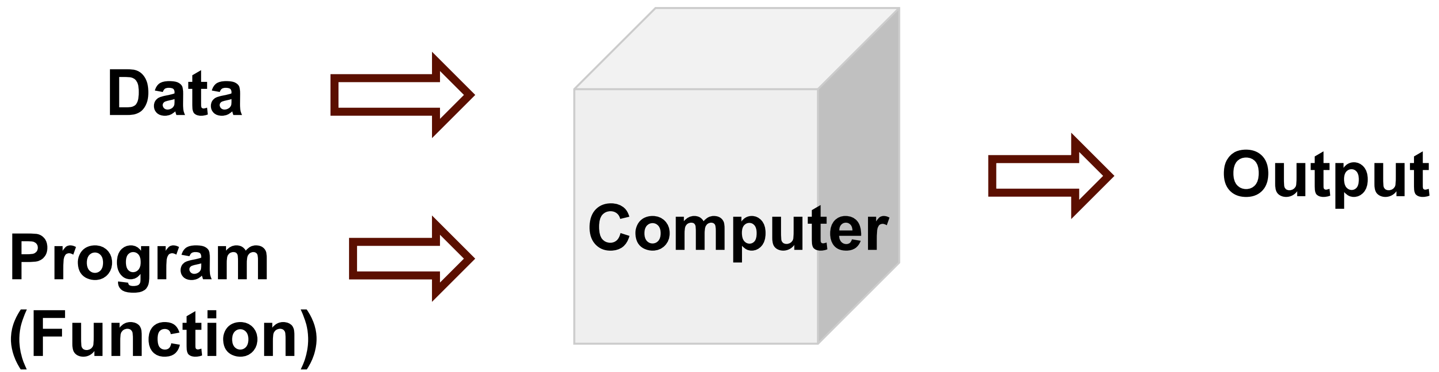
What's Machine Learning? Part 2: like seriously what is it



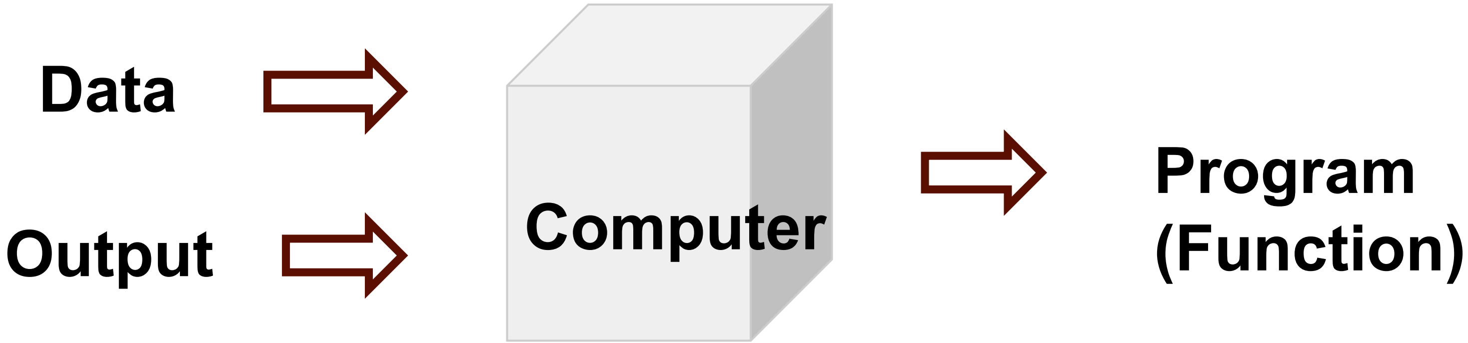
Some Definitions of ML

- 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

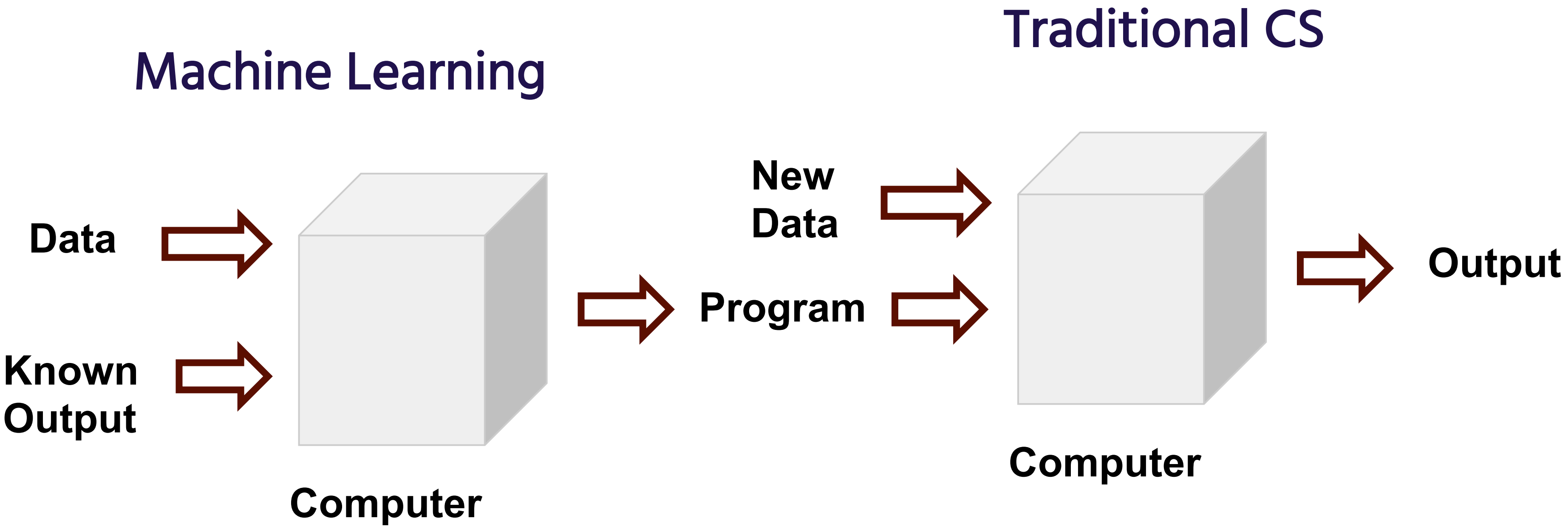
Traditional Computer Science



Machine Learning

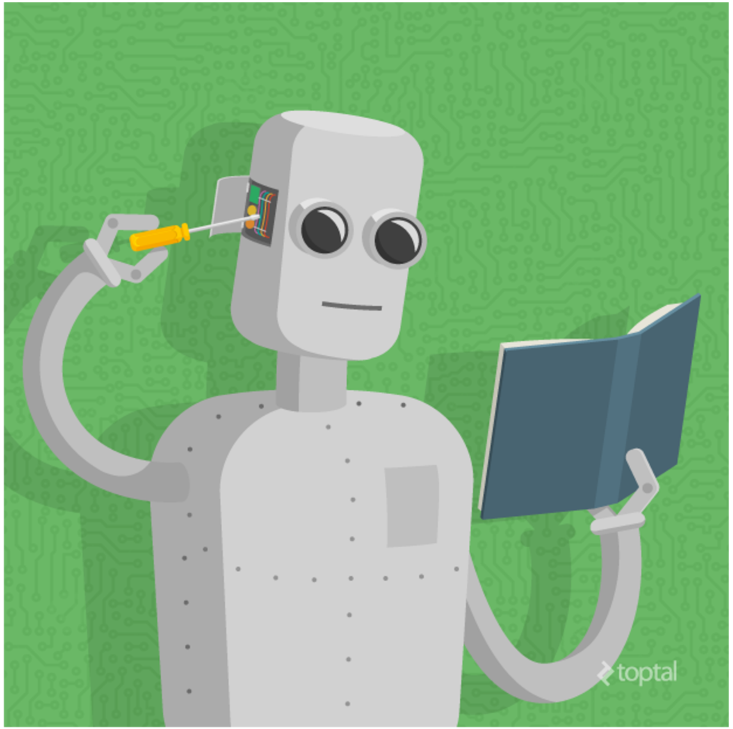


Using Machine Learning



Demo

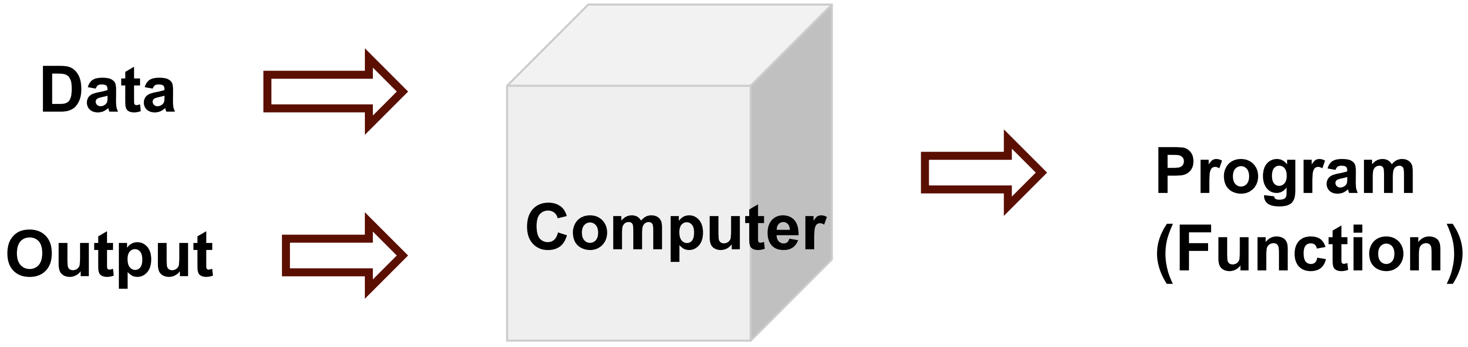
What's Machine Learning? Part 3: what's a model?



What’s a model?

- The output of a machine learning algorithm
- A procedure to produce some outputs when given some inputs
- A relationship between inputs and outputs
- A guess at how inputs and outputs are related
- A set of assumptions we’re imposing on the dataset
- 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

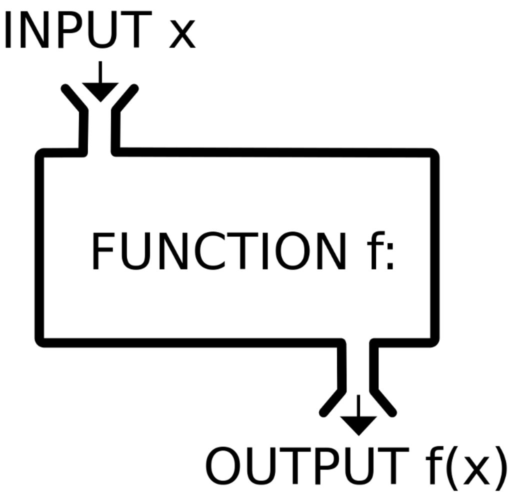
Linear Regression

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

- 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 make 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

Function



Weighted Sum

$$x_1 \ldots x_p$$

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

$$y$$

$y = B_0 + B_1x_1 + \ldots$ is a model

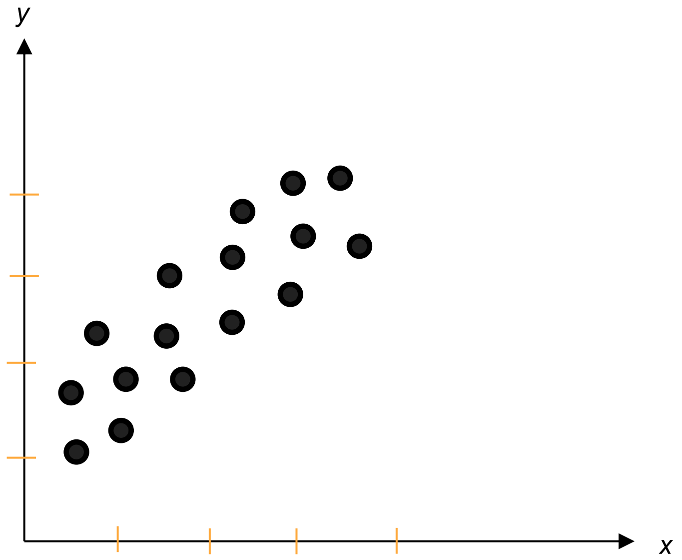
- A relationship between inputs and outputs

  $y = B_0 + B_1x_1 + \ldots$ relates inputs to outputs
- A guess at how inputs and outputs are related

  but $y = B_0 + B_1x_1 + \ldots$ 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 linear with input features and input features are ordered
- A configurable thing (hyperparameters)

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



$y = B_0 + B_1x_1 + \ldots$ is a model

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

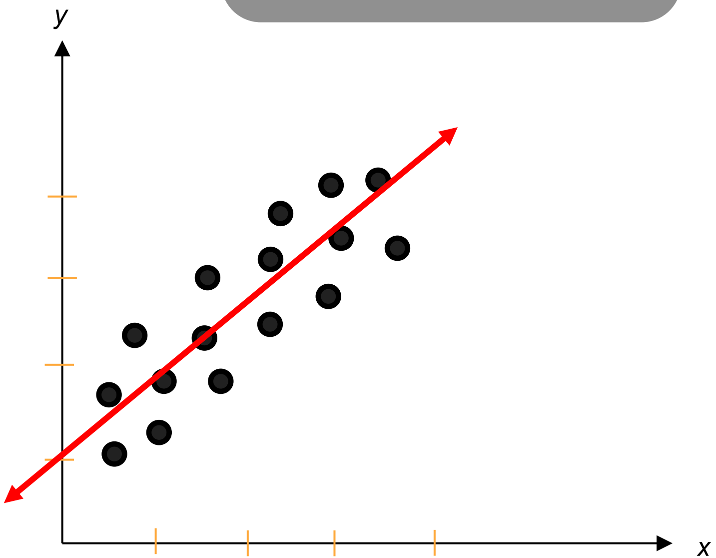
- A relationship between inputs and outputs

  $y = B_0 + B_1x_1 + \ldots$ relates inputs to outputs
- A guess at how inputs and outputs are related

  but $y = B_0 + B_1x_1 + \ldots$ is just a guess/estimate; it’s not exactly true
- A set of assumptions we’re imposing on the dataset

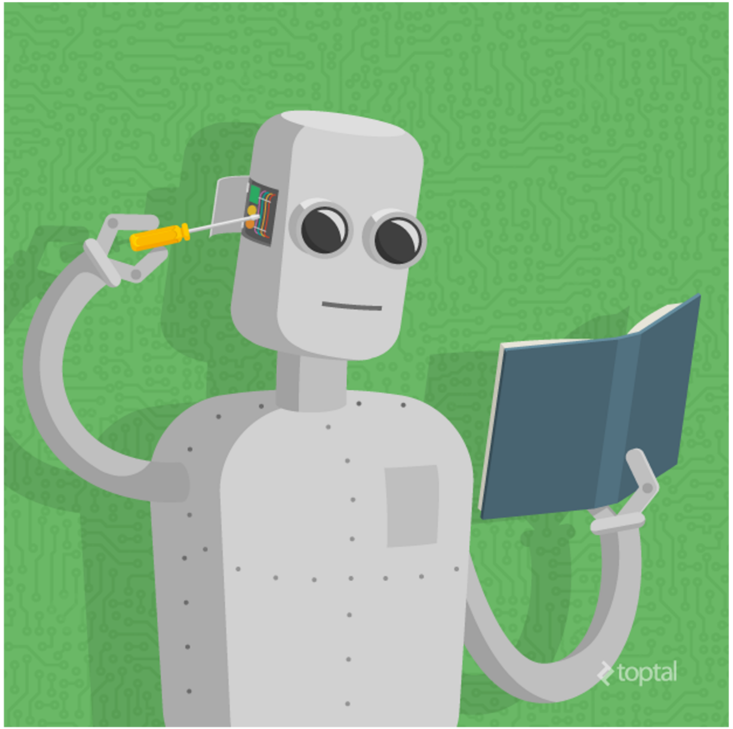
  We’re assuming output is linear with input features and input features are ordered
- A configurable thing (hyperparameters)

  Sorry, we don’t cover very much linear regression configuration here :(



Demo

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

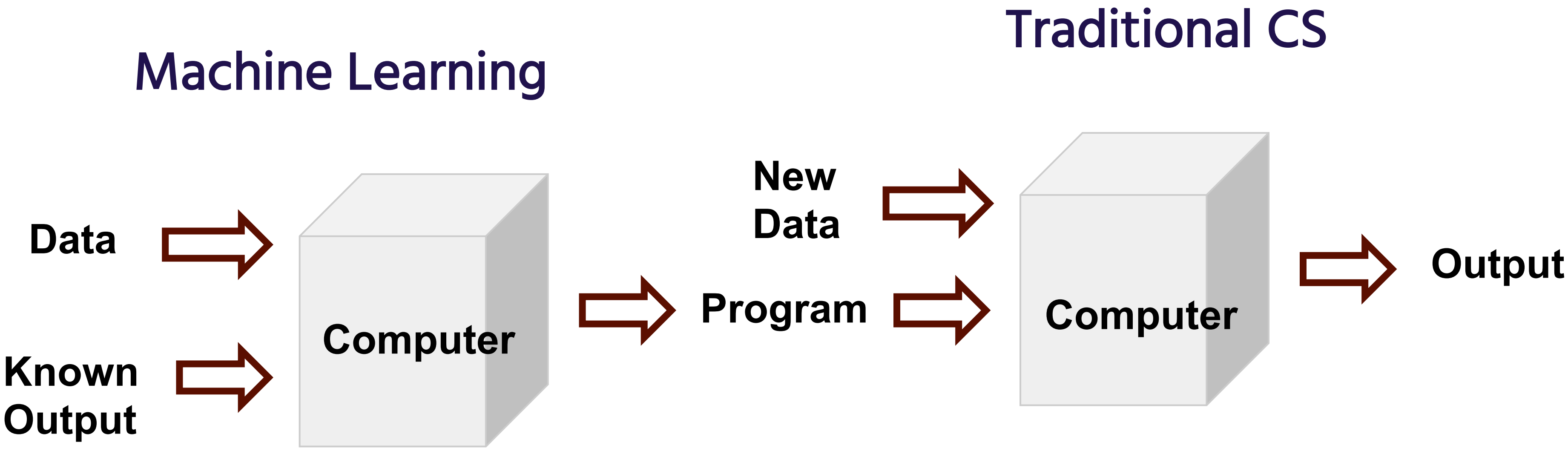
- We didn't learn the data's general patterns :(
- We learned the specific mapping from **train input to train outputs** :((((

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

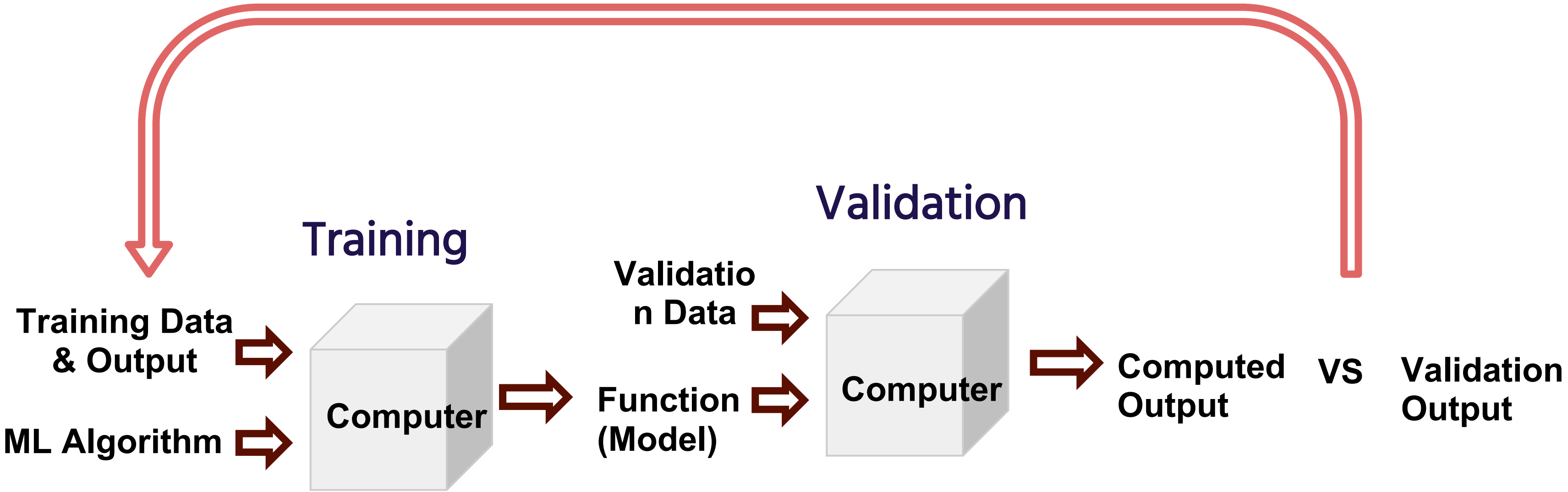
Terminology: Training and Validating

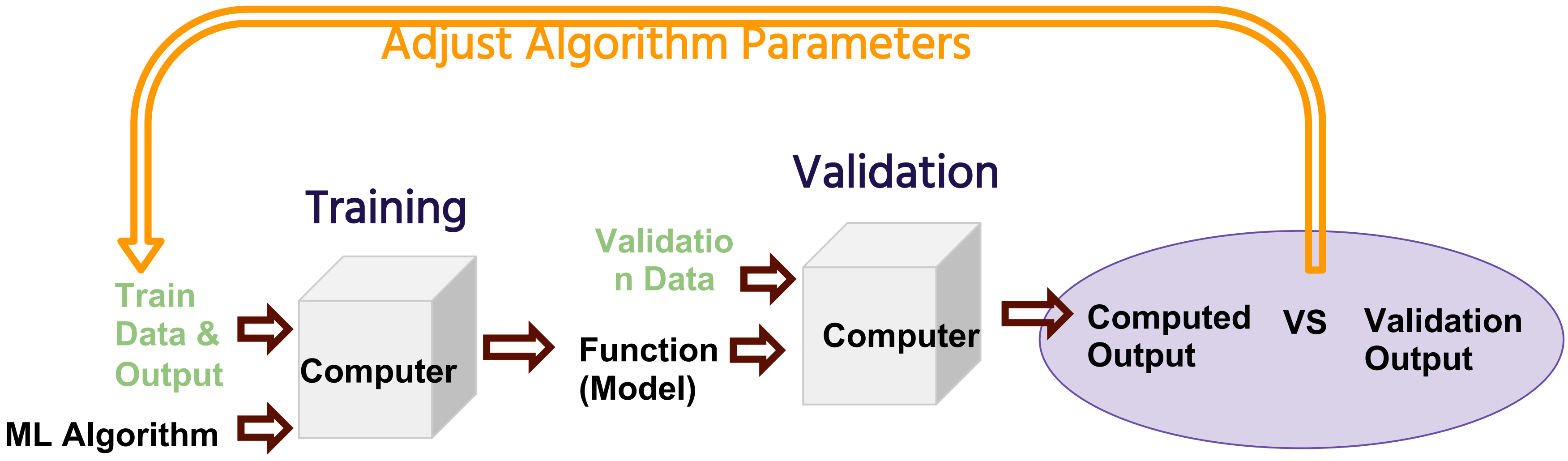
- Split data into two sets
- Train model on one, validate on the other
- “Model training” = learn a relationship/program
  - e.g. give the linear regression data so it can define the B’s
- “Model validation” = see if the learned relationship is accurate on other data

Machine Learning



Our ML Workflow





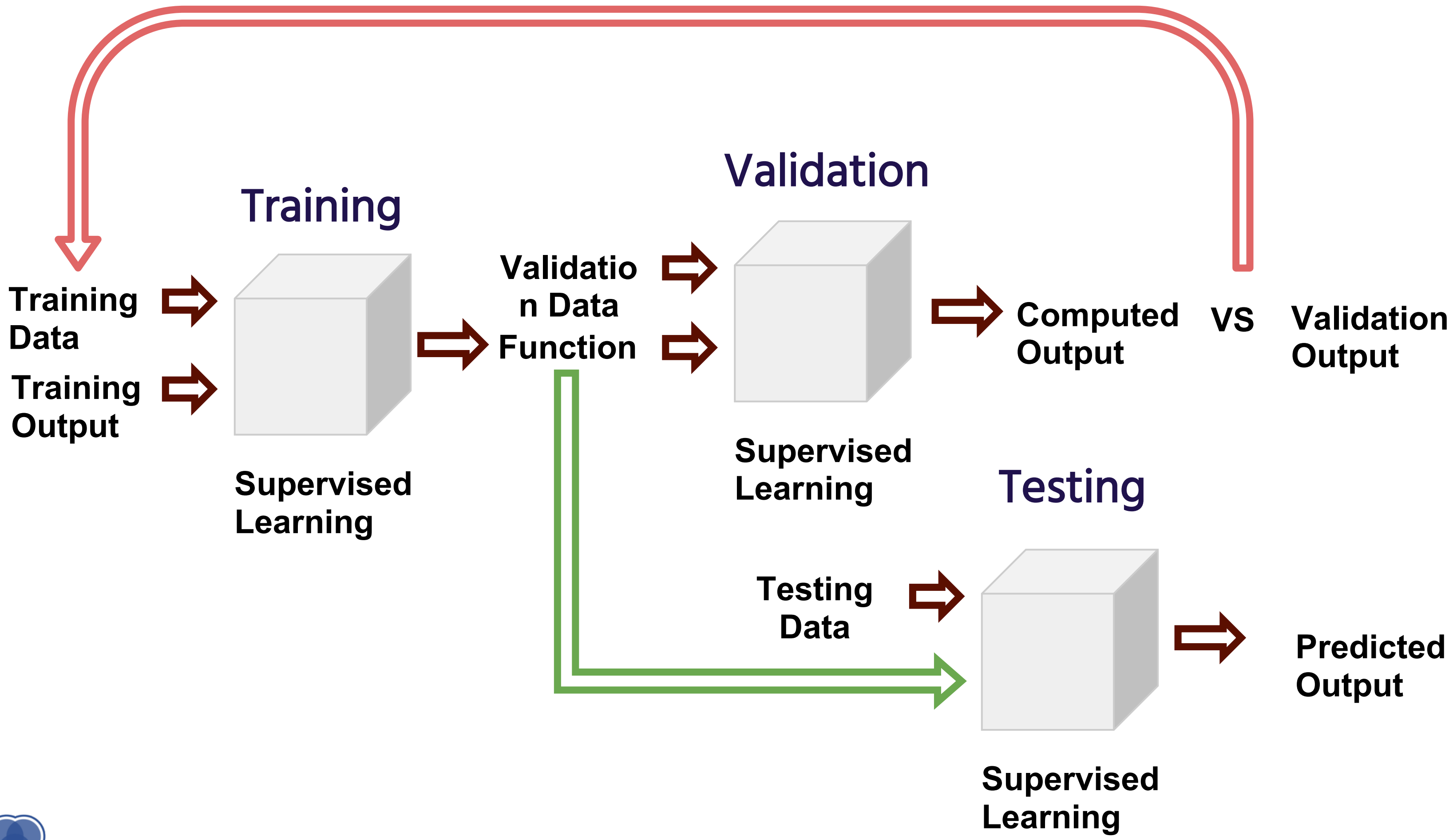
1. Select data
2. Assess model accuracy
3. Adjust Model

Pitfall of validation: Overfitting

Predicting well on validation set $\neq$ Predicting well on new data

- We used the validation set to make our adjustments. This means our model is **biased** to the validation set ☹

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


Training
Validation
Training Data
Training Output
Supervised Learning
Validatio n 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 5:30pm EST on October 28^{th}, 2020

Next Lecture: Assessing Model Accuracy + Fundamentals of ML

(a.k.a. *What's Machine Learning? Part 99999999*)

Midterm Project: Due at 11:59pm EST on October 30^{th}, 2020

CDS Education

We explore, learn, and educate big minds.