Lecture 4: Fundamentals of Machine Learning Pt. 1

INFO 1998: Introduction to Machine Learning

Introduction to Machine Learning and Tools

Presenter: Sri Kundurthy

srk247@cornell.edu

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 weeks (week of 03/17).
  - 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

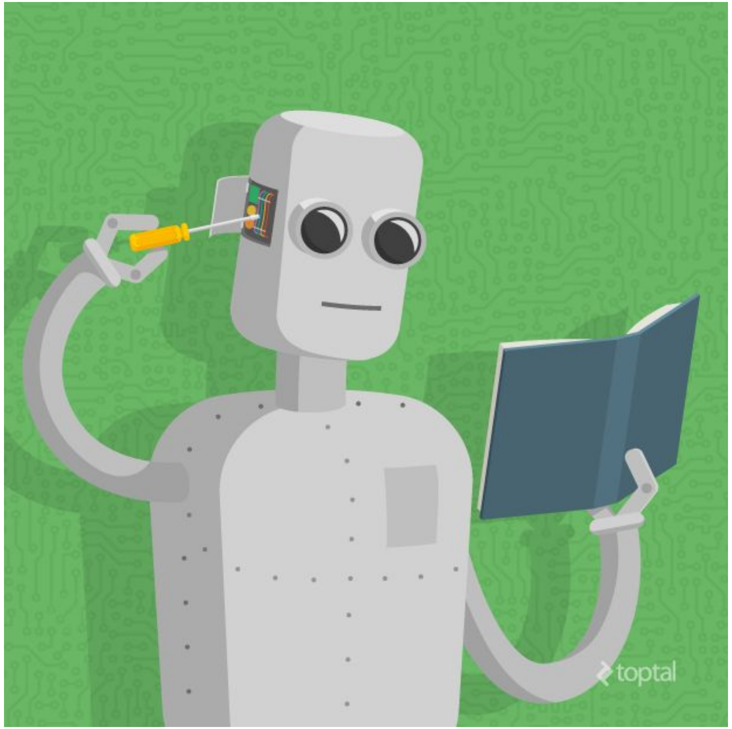
Today’s Goal: be able to write code to do some kind of ML (to some extent)

- **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

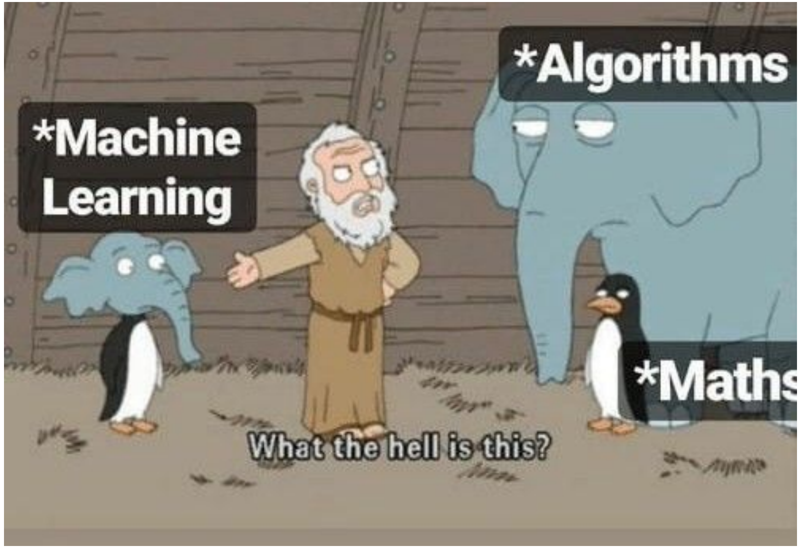
1. What are some things a Machine Learning Engineer does?
2. On a high level, how do you define “Machine Learning”?
3. What’s a Machine Learning Model?
4. What’s a *good* Machine Learning Model?

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 going to do:

Write as little code as possible!

- Use pandas to deal with data
- Use numpy to do math
- Use scikit-learn (“sklearn”) to make & analyze ML models

What we're gonna do:

Our main tasks:

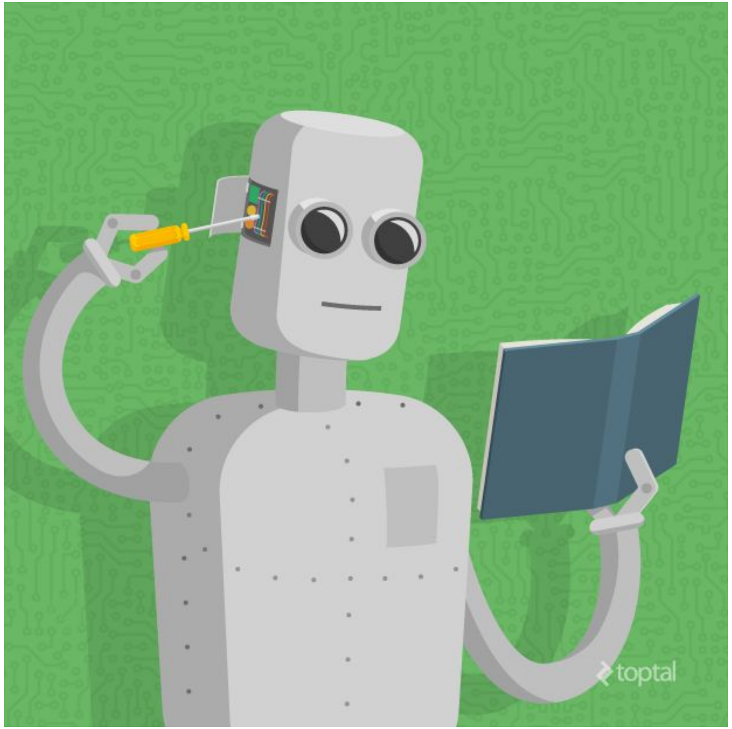
- **Formulate** a problem
- Find and **understand** data for that problem
- Choose a specific **algorithm class** to **solve** the problem
- Choose different parts of the data to **best** solve the problem
- Find which pandas, numpy, and scikit-learn functions do what we want
- Interpret the results and **fine-tune** our model

Quick analogy: studying

- Setup
  - Goal: Be able to solve the test problems *well*
  - 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 :)

*Note: this analogy describes supervised learning

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



Remark: We like to come up with *functions*

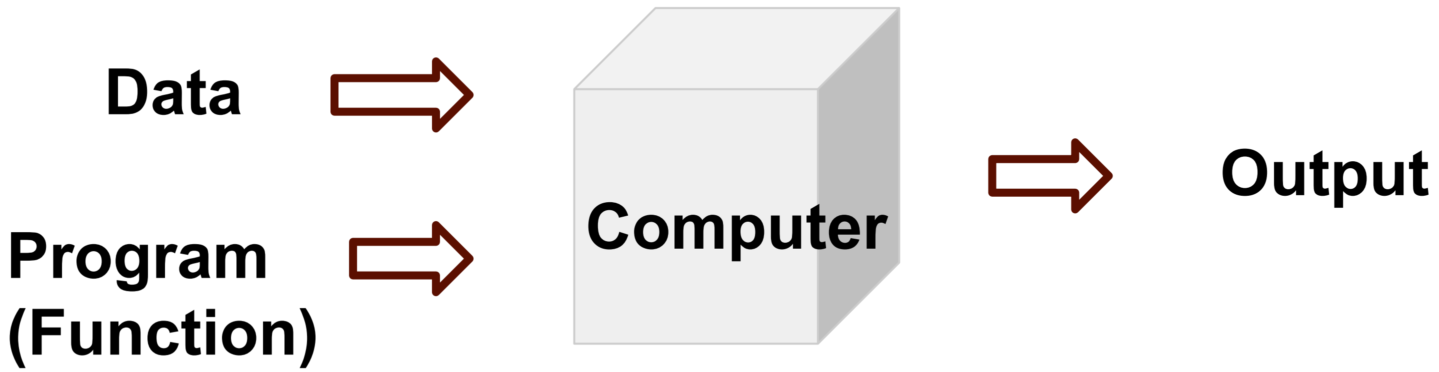
- Functions get us from *input data* to an *output*
  - In math, functions are how we relate information across different dimensions
  - Represent a sort of dependence
  - Show us how we can uncover an unknown value
- We can generalize beyond math!

Physics: A White-Box View of the World

- We have studied the world and crafted **models** *by ourselves* to represent it
- Our models are interpretable
- **White-box algorithms:** The inner workings of the algorithm are transparent

$$F=ma$$

Traditional Computer Science



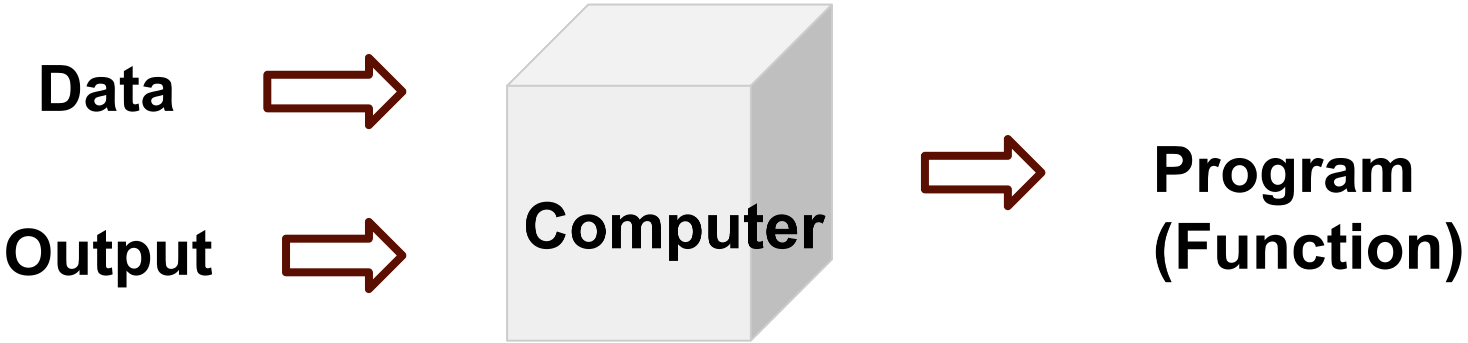
Complex Problems

- Challenge:
    - Write a mathematical equation to predict whether or not it is going to rain at 5:45 PM today.
- Several factors to consider
- Very complicated to come up with ourselves!

Black Box Models

- Perhaps we can derive a process to solve the problem
  - Determine what a function would roughly look like
  - Think of relevant inputs
  - Allow the function to *build itself*
- **Black Box Model:** Results may not be interpretable!

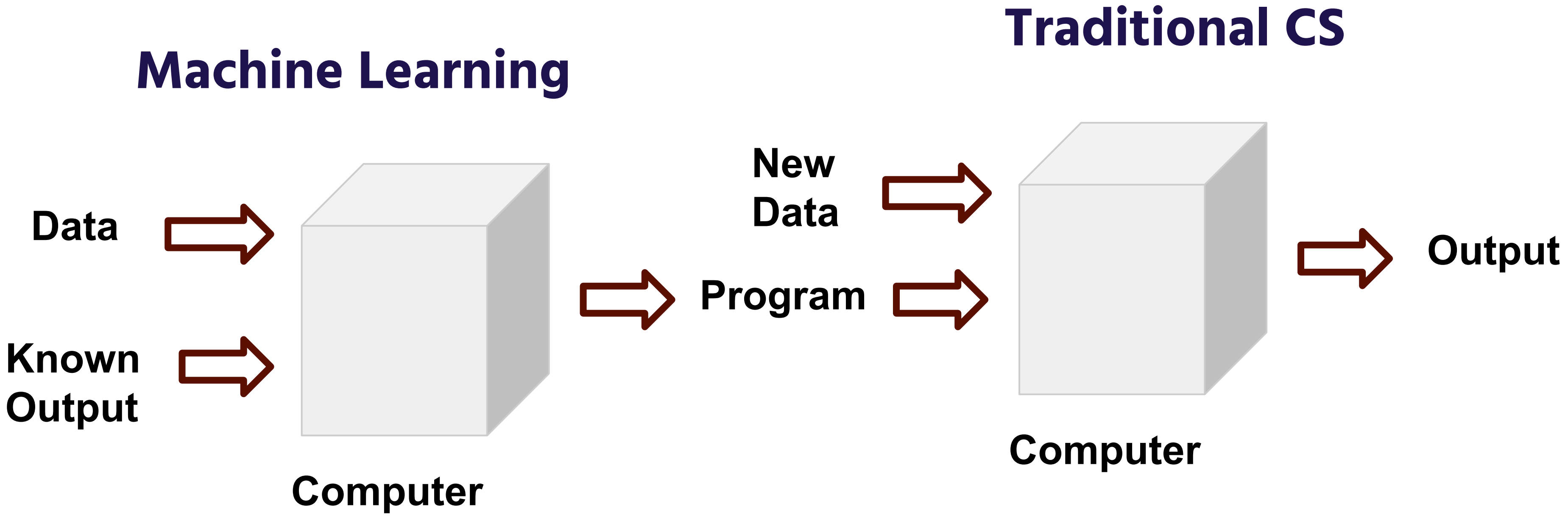
Machine Learning



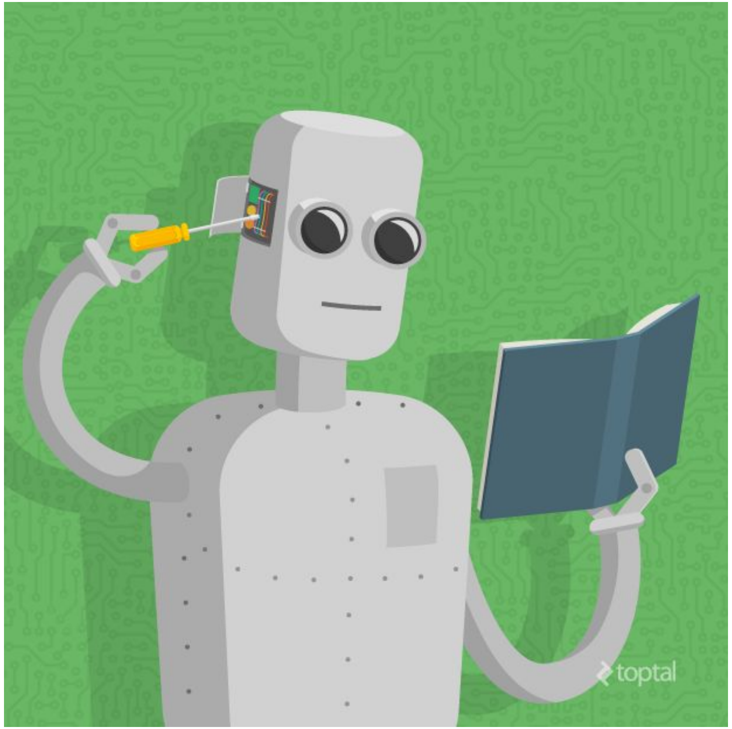
Some Definitions of ML

- Give computers the ability to learn without being explicitly programmed
- 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

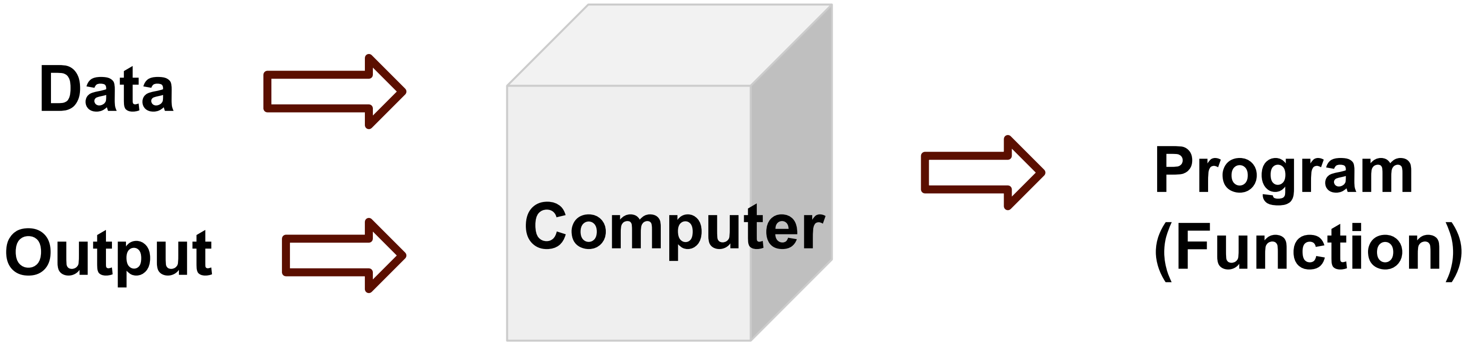
Using Machine Learning



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



ML Algorithm produces 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 parametrized function we can configure

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

A Sample Task

name	city	state	adm_rate	undergrads	cost	compl_4	median_hh_inc	median_earnings
Cornell University	Ithaca	NY	0.1507	14226	63596	0.8639	80346.48	73600
Washington University in St Louis	Saint Louis	MO	0.1674	7032	65887	0.8643	79298.58	66300
Lafayette College	Easton	PA	0.3025	2505	61905	0.8653	85923.51	67500
Johns Hopkins University	Baltimore	MD	0.1412	5862	63509	0.869	81539.46	69800
Vanderbilt University	Nashville	TN	0.1168	6857	62320	0.8697	76279.78	64500

What are some things we can do?

name	city	state	adm_rate	undergrads	cost	compl_4	median_hh_inc	median_earnings
Cornell University	Ithaca	NY	0.1507	14226	63596	0.8639	80346.48	73600
Washington University in St Louis	Saint Louis	MO	0.1674	7032	65887	0.8643	79298.58	66300
Lafayette College	Easton	PA	0.3025	2505	61905	0.8653	85923.51	67500
Johns Hopkins University	Baltimore	MD	0.1412	5862	63509	0.869	81539.46	69800
Vanderbilt University	Nashville	TN	0.1168	6857	62320	0.8697	76279.78	64500

Predict Median Graduate Earnings

name	city	state	adm_rate	undergrads	cost	compl_4	median_hh_inc	median_earnings
Cornell University	Ithaca	NY	0.1507	14226	63596	0.8639	80346.48	73600
Washington University in St Louis	Saint Louis	MO	0.1674	7032	65887	0.8643	79298.58	66300
Lafayette College	Easton	PA	0.3025	2505	61905	0.8653	85923.51	67500
Johns Hopkins University	Baltimore	MD	0.1412	5862	63509	0.869	81539.46	69800
Vanderbilt University	Nashville	TN	0.1168	6857	62320	0.8697	76279.78	64500

Pick Some Features

name	city	state	adm_rate	undergrads	cost	compl_4	median_hh_inc	median_earnings
Cornell University	Ithaca	NY	0.1507	14226	63596	0.8639	80346.48	73600
Washington University in St Louis	Saint Louis	MO	0.1674	7032	65887	0.8643	79298.58	66300
Lafayette College	Easton	PA	0.3025	2505	61905	0.8653	85923.51	67500
Johns Hopkins University	Baltimore	MD	0.1412	5862	63509	0.869	81539.46	69800
Vanderbilt University	Nashville	TN	0.1168	6857	62320	0.8697	76279.78	64500

Our Goal?

name	city	state	adm_rate	undergrads	cost	compl_4	median_hh_inc	median_earnings
Cornell University	Ithaca	NY	0.1507	14226	63596	0.8639	80346.48	73600
Washington University in St Louis	Saint Louis	MO	0.1674	7032	65887	0.8643	79298.58	66300
Lafayette College	Easton	PA	0.3025	2505	61905	0.8653	85923.51	67500
Johns Hopkins University	Baltimore	MD	0.1412	5862	63509	0.869	81539.46	69800
Vanderbilt University	Nashville	TN	0.1168	6857	62320	0.8697	76279.78	64500
Rutgers University	New Brunswick	NJ	0.5845	35102	29076	0.5838	82669.68	?
Case Western Reserve University	Cleveland	OH	0.3627	5039	59467	0.6311	69873.4	?

Machine Learning Algorithms

ML Algorithms

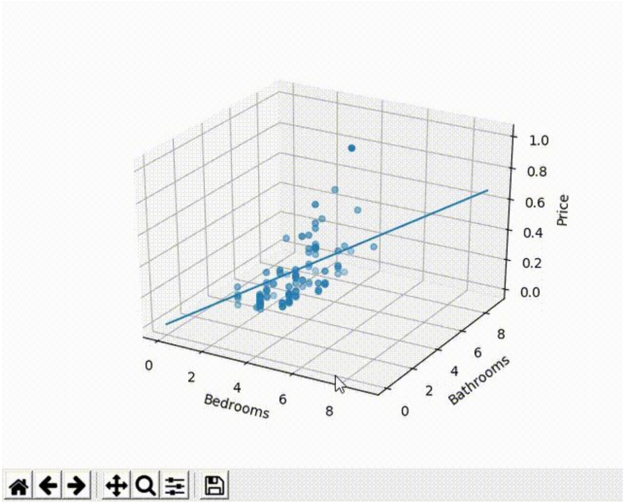
- We pick different kinds of algorithms to accomplish different tasks
- Classification
    - Group Data Into Distinct Classes
- Regression
    - Based on an input, provide a continuous-value output
- “All Models Make Assumptions”

Linear Regression

Linear Regression

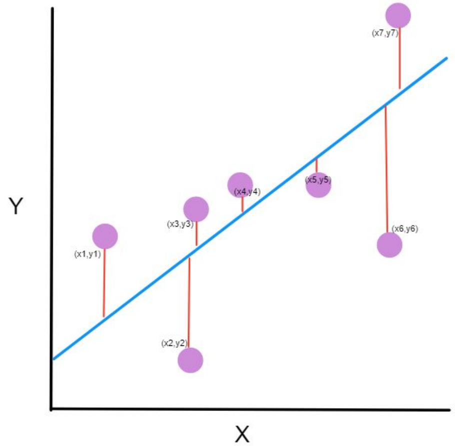
$$y = B_0 + B_1 x_1 + \dots + B_p x_p + \varepsilon$$

- x is an input; $x_1, x_2, \dots, x_p$ 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



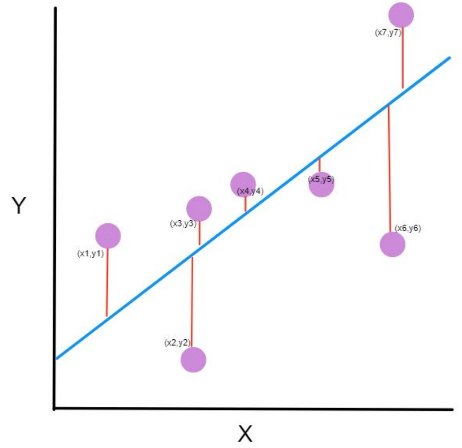
Linear Regression: Ordinary Least Squares

- 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



$$\mathrm{MSE} = \overset{\text{Mean}}{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}(\overset{\text{Error}}{\boxed{Y_i - \hat{Y}_i}})^{\overset{\text{Squared}}{\boxed{2}}}$$

"Training" a Model



- Dataset of n training points
- Datapoints: (X, Y_i) -> (input, output)
- Objective: Minimize MSE

1. Use the x values in our dataset to make a prediction
    a. Note: x is a vector

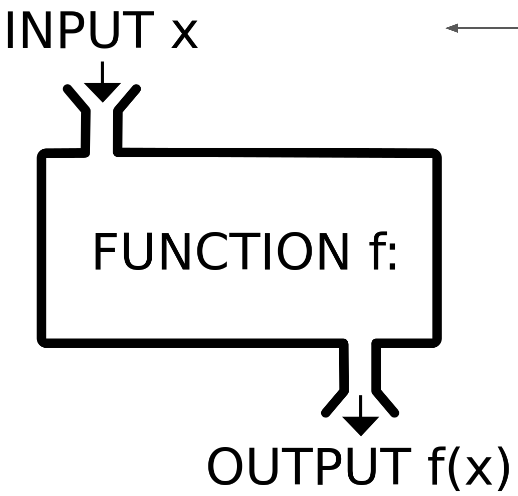
$$\hat{Y}_i = B_0 + B_1 x_1 + \ldots + B_p x_p + \varepsilon$$

2. Compare our prediction to the real Y_i
3. *Update B to get a better prediction*
    a. Special Algorithm: Gradient Descent
4. Repeat until MSE is as small as possible

$$\text{MSE} = \overset{\text{Mean}}{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}\left(\overset{\text{Error}}{\boxed{Y_i - \hat{Y}_i}}\right)^{\overset{\text{Squared}}{\boxed{2}}}$$

Linear Regression

Function



Weighted Sum

$$x_1 \dots x_p$$

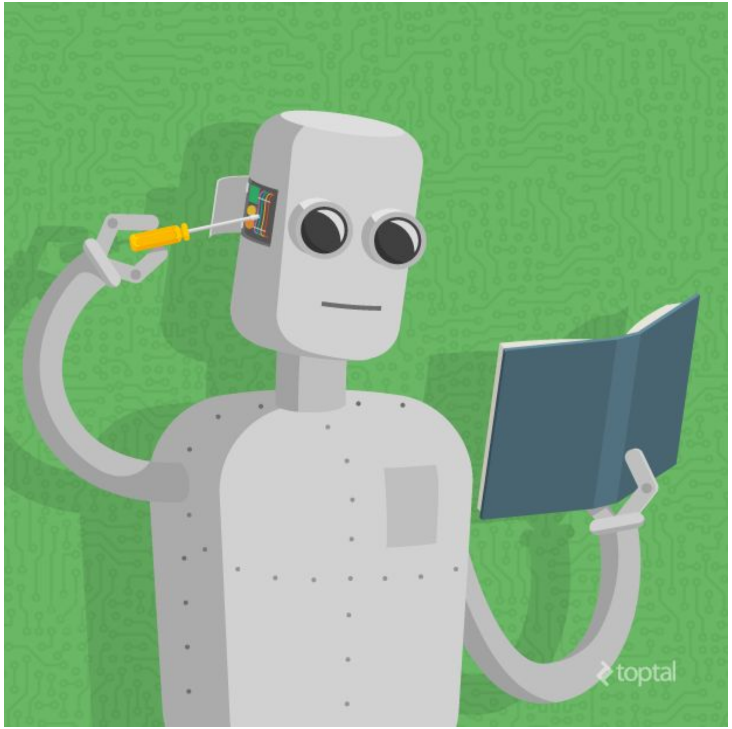
$$y = B_0 + B_1 x_1 + \dots + B_p x_p$$

$$y$$

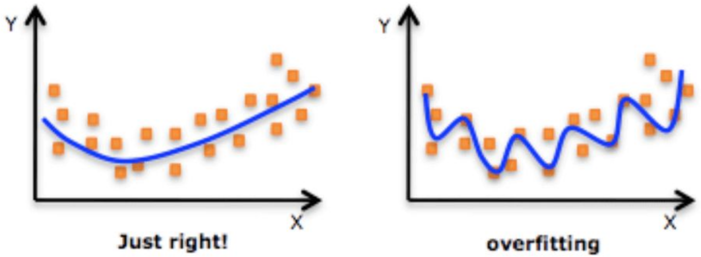
Assumptions of Linear Regression

We’re assuming output is linearly related to input features

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

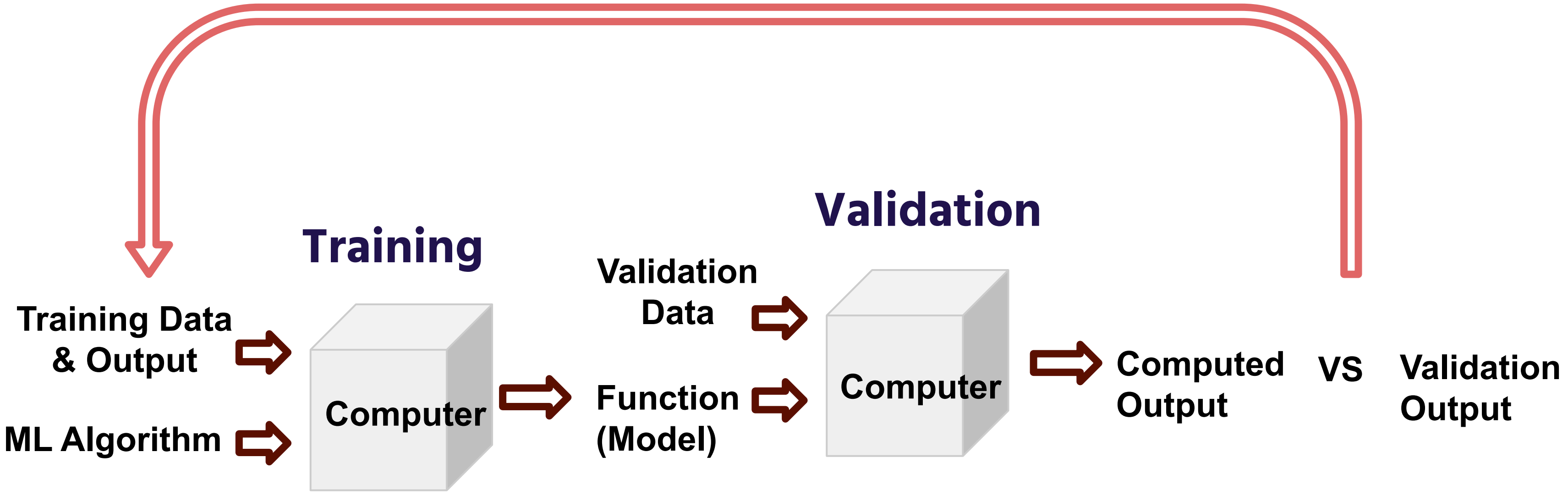
- We learned the specific mapping from **train input to train outputs…**
- But, we didn't learn the data's **general patterns** 🥲🥲🥲🥲🥲🥲

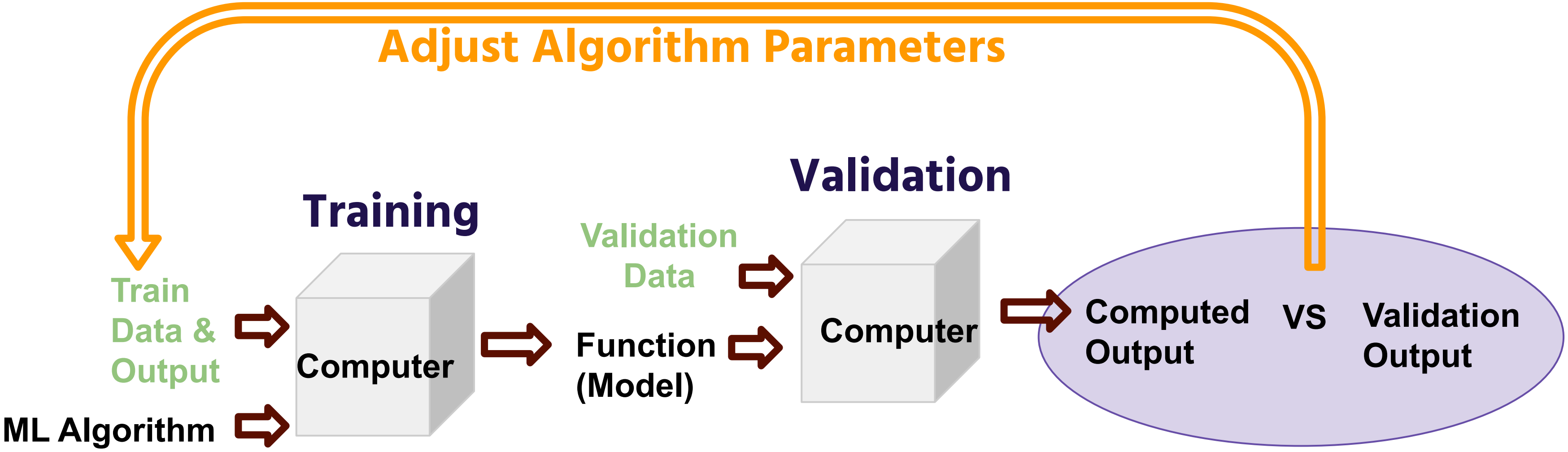
Solution: train on part of data, and check accuracy on a separate part of data (*validation* set)

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

Our ML Workflow





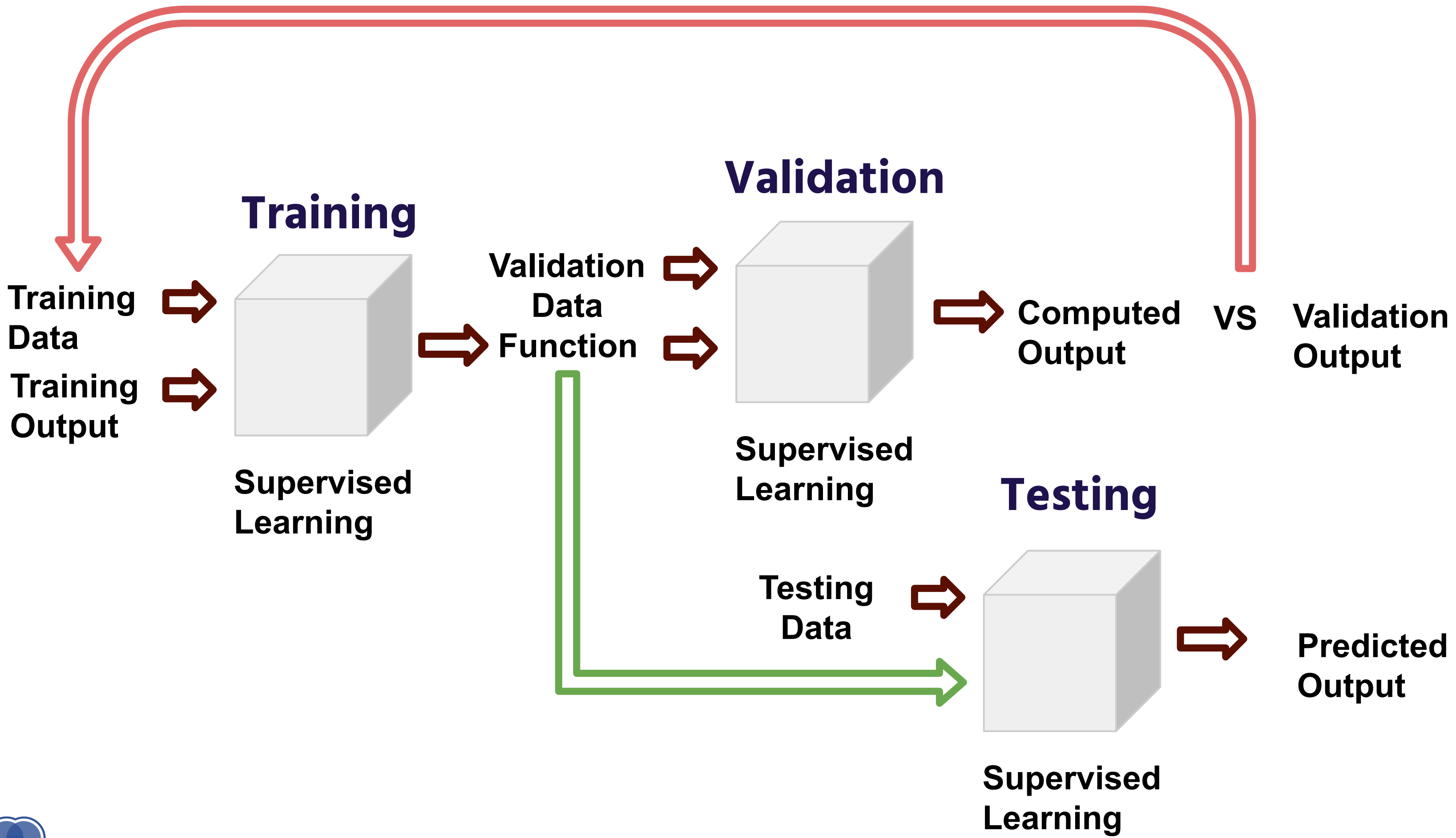
1. Select data
2. Assess model accuracy
3. Adjust Model

Pitfall of Validation: Overfitting

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

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

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


Training
Validation
Training Data
Training Output
Supervised Learning
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 3: Due tonight at 11:59pm EST

Assignment 4: Due at 11:59pm EST on **Wednesday, March 13th**

Next Lecture: Assessing Model Accuracy **+** Fundamentals of ML

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

Coming Up! Web Scraping Workshop 👀

CDS Education

We explore, learn, and educate big minds.