outraggeously AMBITIOUS

Machine Learning Foundations for Product Managers

Course Overview

Duke
PRATT SCHOOL *of* ENGINEERING

Why Take This Course?

- Companies across every industry are using AI to make their products or services more **predictive**, **personalized** and **automated**
- AI is also creating the ability to solve previously unsolved problems
- Successfully bringing AI products to market requires a team effort
- Everyone needs to speak the same language and have the same fundamental understanding

AI Product Management Specialization

Course 1

Machine Learning Foundations for Product Managers

Course 2

Managing Machine Learning Projects

Course 3

Human Factors in AI

Course Learning Objectives

At the conclusion of this course, you should be able to:

1) Explain how machine learning works and the types of machine learning
2) Describe the challenges of modeling and strategies to overcome them
3) Identify the primary algorithms used for common ML tasks and their use cases
4) Explain deep learning and its strengths and challenges relative to other forms of machine learning
5) Implement best practices in evaluating and interpreting ML models

Course Outline

Module	Topic
1	What is machine learning?
2	The modeling process
3	Evaluating and interpreting models
4	Linear models for regression & classification
5	Tree models, ensembles, and clustering
6	Deep learning

outrageously
AMBITIOUS

Module 1: What is Machine Learning

Duke
PRATT SCHOOL *of* ENGINEERING

Module 1 Objectives:

At the conclusion of this week, you should be able to:

1) Describe what machine learning is and does
2) Explain why we should care about machine learning
3) Identify the common types of machine learning tasks
4) Define common ML terms to be able to understand articles and conversations about ML

outrageously
AMBITIOUS

Introduction to Machine Learning

Duke
PRATT SCHOOL of ENGINEERING

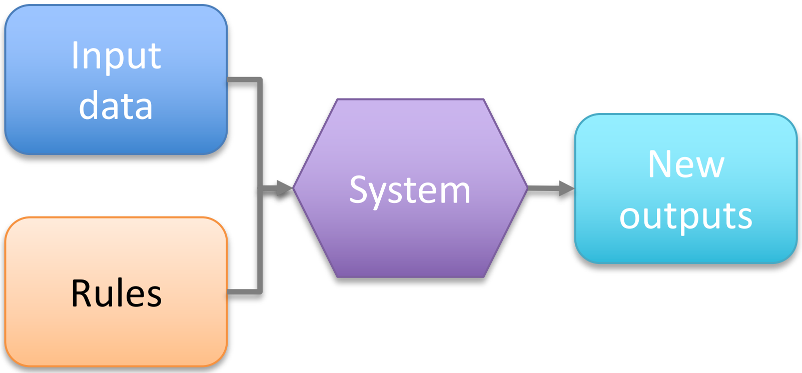
What is Machine Learning?

- "Field of study that gives computers the ability to learn **without being explicitly programmed**" – Arthur Samuel, IBM, 1959

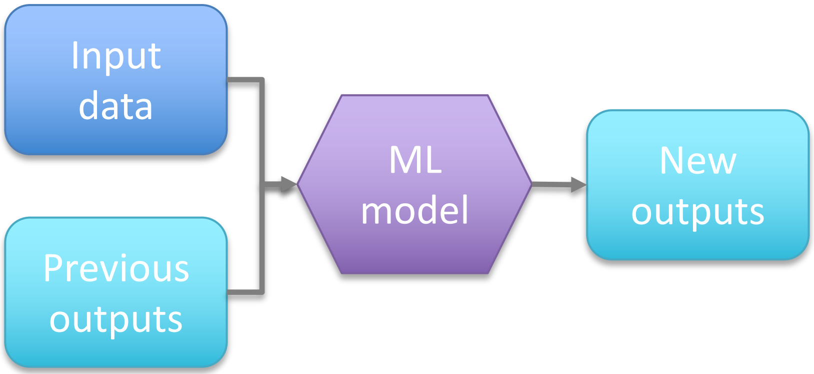
- Instead of providing a computer with exact instructions to solve a problem, we show it examples of the problem to solve and let it figure out how to solve it itself

ML vs. traditional software

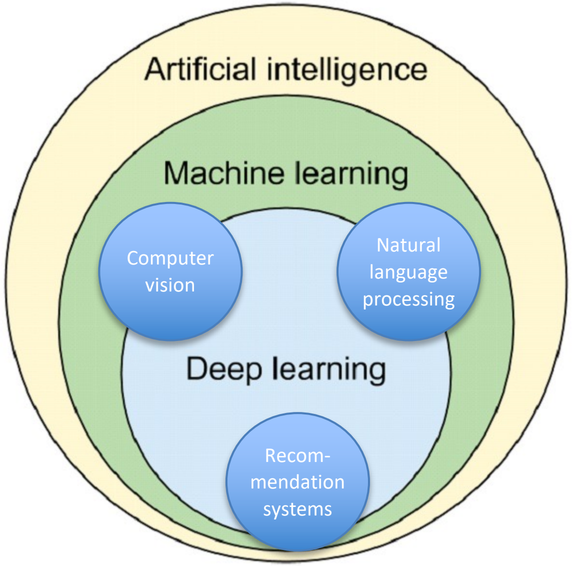
How traditional software generates predictions



How machine learning generates predictions



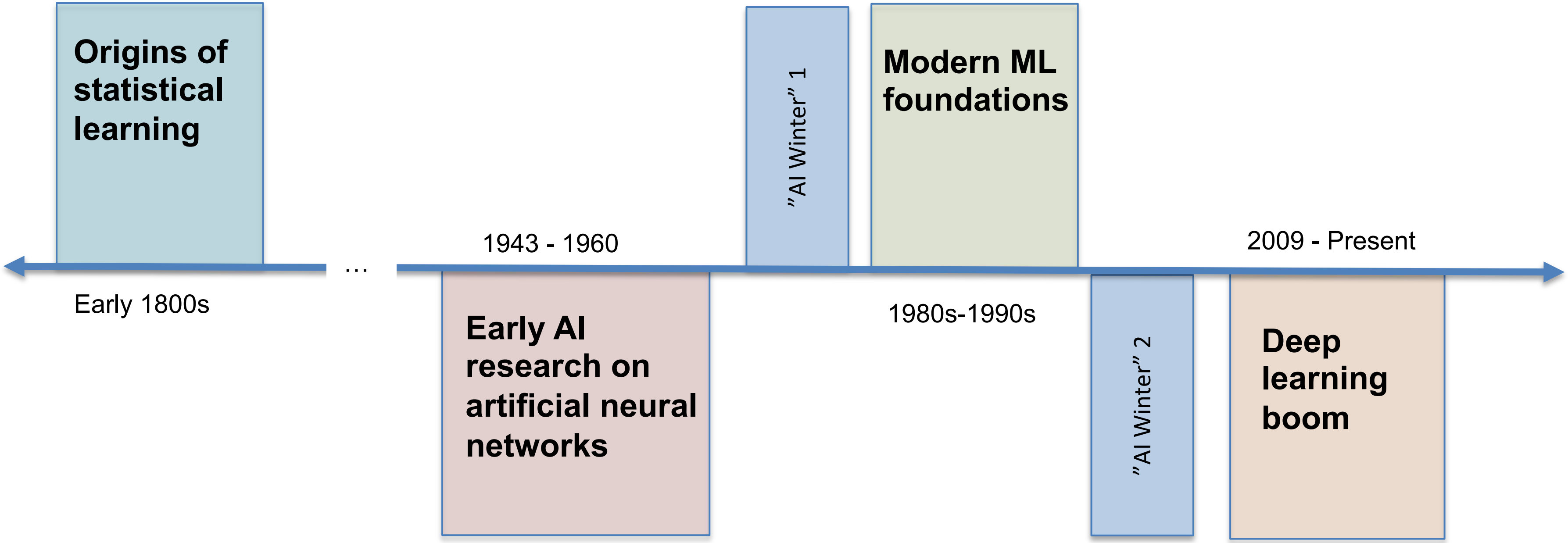
AI vs. Machine Learning



https://commons.wikimedia.org/wiki/File:Fig-X_All_ML_as_a_subfield_of_AI.jpg

- **Machine learning** is a set of methods & tools which help realize the goal of the field of **artificial intelligence**
- **Deep learning**, or the use of neural networks containing many layers, is a sub-field of machine learning
- **Computer vision, natural language processing, recommendation systems** etc. are sub-fields of AI which rely on machine learning methods

Brief History of AI/ML

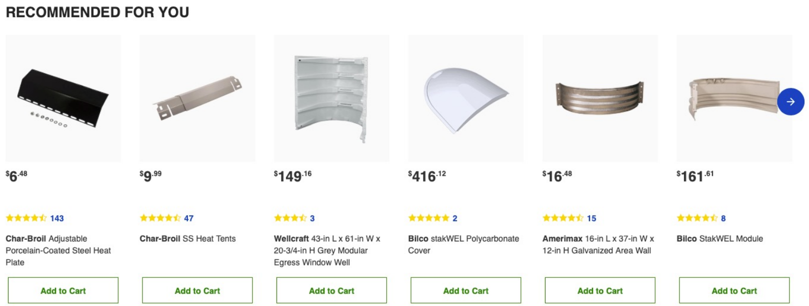


Machine Learning Today

- **Explosion in data**
  - Ubiquitous internet connectivity
  - Advances in sensor technology
  - Smart connected devices
- **Deep learning** has made what was impossible, possible
  - Massive increase in computational power – GPUs
  - Huge sets of labeled data for training
  - Algorithmic advances
- **Pervasiveness** of machine learning models in products and systems we interact with daily

Where Do We Find ML?

Product recommendations



Spam filters

Where Do We Find ML?

Mail routing via OCR

Credit card fraud detection

outrageously
AMBITIOUS

Data Terminology

Duke
PRATT SCHOOL of
ENGINEERING

Data Comes in Many Forms

“**Data** are characteristics or information, usually numerical, that are collected through observation.” [OECD Glossary of Statistical Terms]

Data Comes in Many Forms

Almost anything can be turned into numbers:

- Measurements
- Text
- Images
- Sound
- Video

Data may have different relationships:

- Spatial relationships
- Temporal relationships

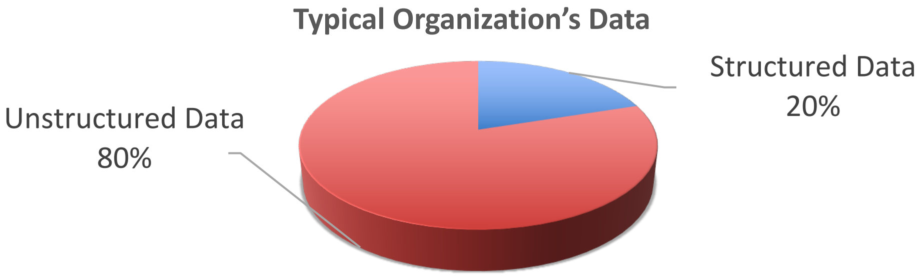
Structured vs. Unstructured Data

Structured data

- Set structure based on pre-defined fields for each record
- Often stored in relational databases
- Easy to enter, search and analyze
- Works well with common tools

Unstructured data

- Does not follow a defined format of fields
- Many types – images, videos, sounds, text
- Requires specialized tools to work with



Continuous vs Categorical Data

Continuous

- Numeric variable that has an infinite number of values between any two values
- E.g. length of a part, temperature, height, time

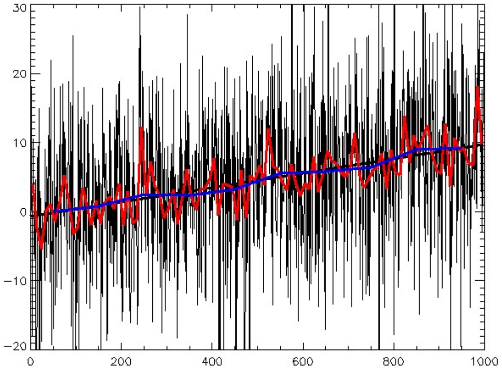
Categorical

- Finite number of categories / distinct groups
- May or may not have a logical order
- E.g. gender, student major, material type, color

Discrete

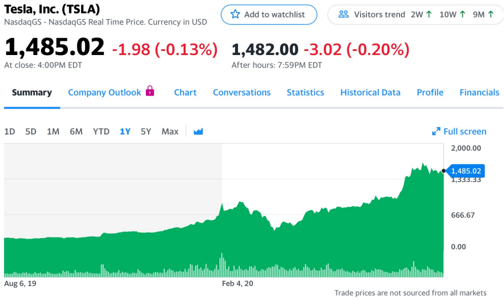
- Numeric variable that has a countable number of values between two values
- E.g. age, number of parts, year made
- Rule of thumb – if number of possible values small (e.g <10), treat as categorical

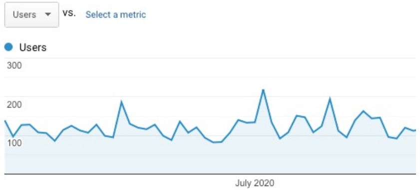
Time series data



Wikipedia

- Series of data points organized in time order
- Points are usually equally spaced by time
- Assumptions:
  - Time is considered one-way
  - Points close together in time are more related than points further apart





Terminology

Targets / Labels / Annotations / Response / Y Variable / Dependent Variable

Features / Factors / Predictors / X Variables / Independent Variables / Attributes / Dimensions

Observations / Instances / Examples / Feature Vectors

	Neighbor-hood	School district	Square footage	Number of bedrooms	Year built	Market sale price
House 1	Weycroft	Wake	3400	4	2010	$612,000
House 2	Horton Creek	Wake	4200	5	2008	$675,000
House 3	Cary Park	Chatham	3250	4	2012	$520,000
…	…	…	…	…	…	…

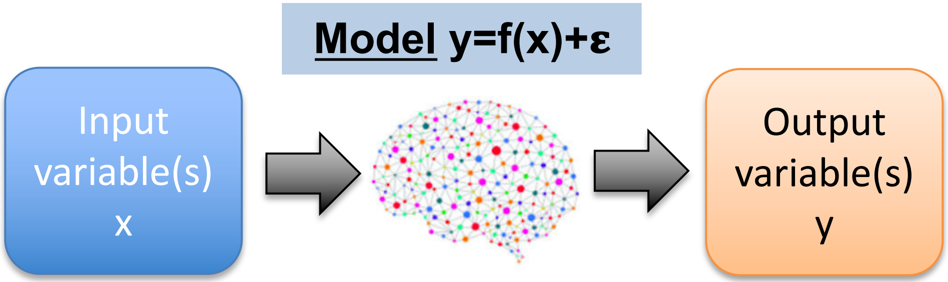
outrageously AMBITIOUS

What is a Model?

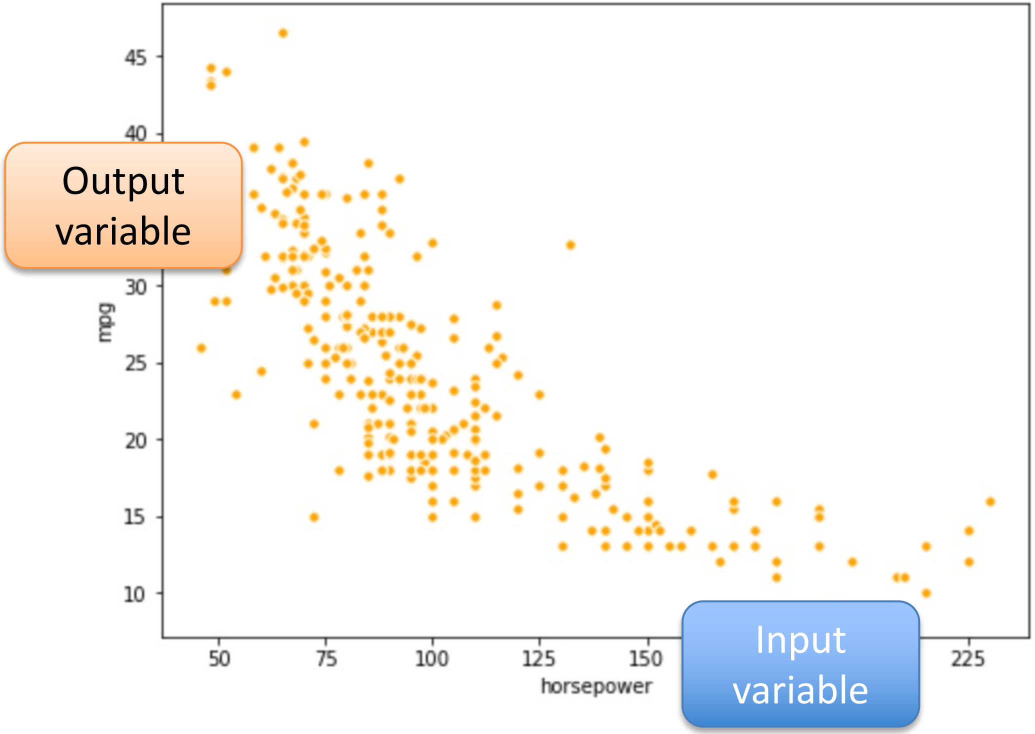
Duke
PRATT SCHOOL *of* ENGINEERING

What is a model?

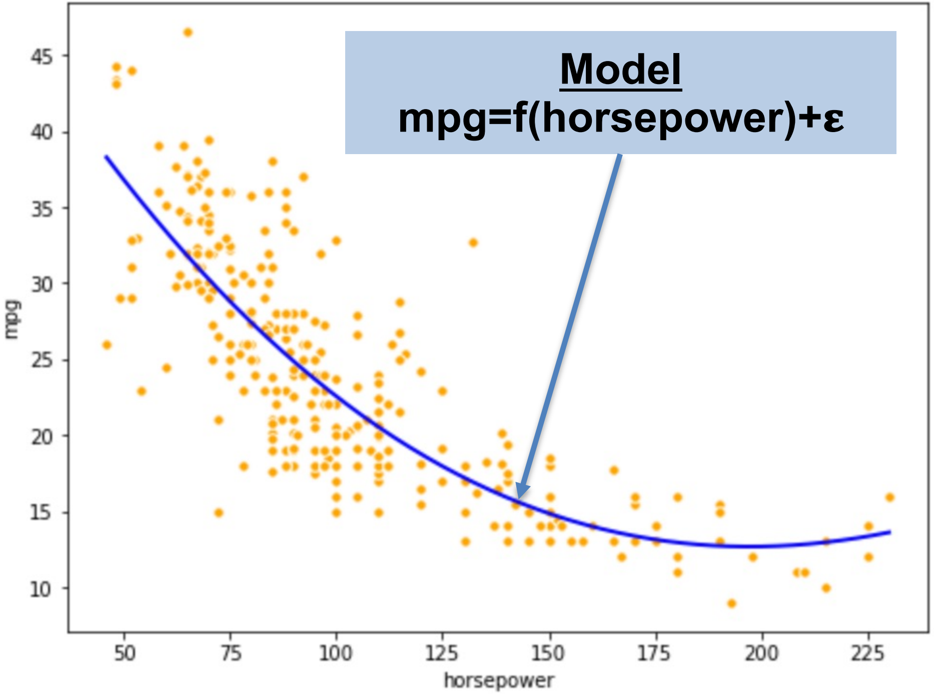
A **model** is an approximation of the relationship between two variables



What is a model?



What is a model?

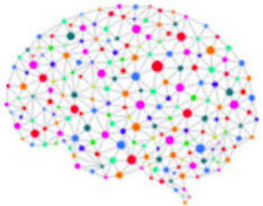


What is a model?

Observations of input data **(X)**

	Neighbor-hood	School district	Square footage	Number of bedrooms	Year built
House 1	Weycroft	Wake	3400	4	2010
House 2	Horton Creek	Wake	4200	5	2008
House 3	Cary Park	Chatham	3250	4	2012
…	…	…	…	…	…

Model y=f(X)+ε



Predictions of **target (y)**

Market sale price
$612,000
$675,000
$520,000
…

Building a model

To create a model we define four things:

1. **Features** to use
2. **Algorithm** – acts as a form/template for model
3. **Hyperparameter** values for algorithm
4. **Loss function** to optimize

We **train** our model using historical data:

- Algorithm & hyperparameters provide overall model form
- “Learn” values for the model which minimize loss function

outrageously
AMBITIOUS

Types of Machine Learning

Duke

PRATT SCHOOL *of*
ENGINEERING

Types of Machine Learning

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Objective	Prediction of a target variable	Organize data by inherent structure	Learn strategies via interaction
Learning Task(s)	Classification Regression	Clustering Anomaly detection	Achieve a goal
Target Data Required?	Yes	No	Yes, but delayed
Examples	• Identifying pneumonia from xray images • Predicting real estate prices	• Market segmentation • Identifying fraudulent activity	• AlphaZero • Autonomous vehicles

Supervised vs. Unsupervised Learning

Supervised learning

At least some past observations of the features (X_i) and targets (y_i) are known and used to build a model

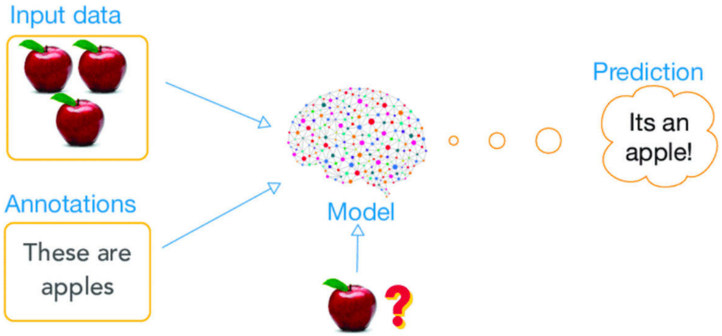


Image source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

Supervised vs. Unsupervised Learning

Unsupervised learning

We only have observations of the features (X_i). We need to use the observations to guess what the targets (y_i) would have been and build a model from there

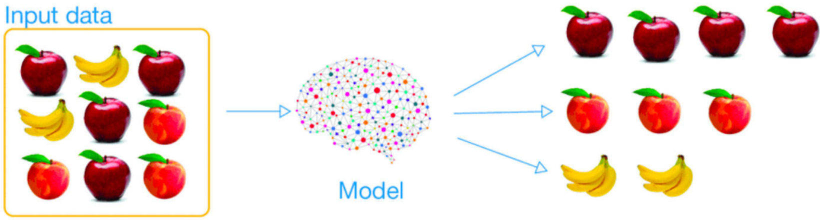
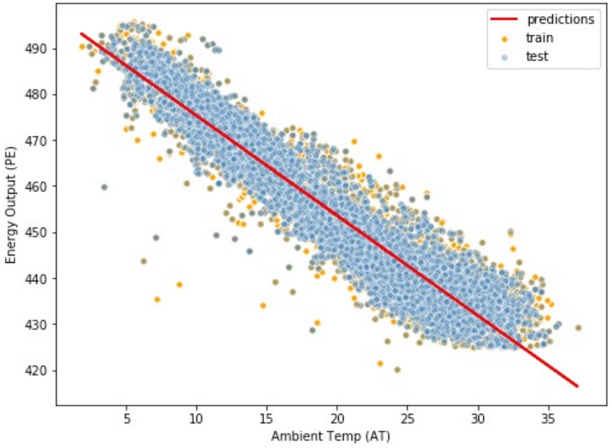


Image source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

Regression vs. Classification

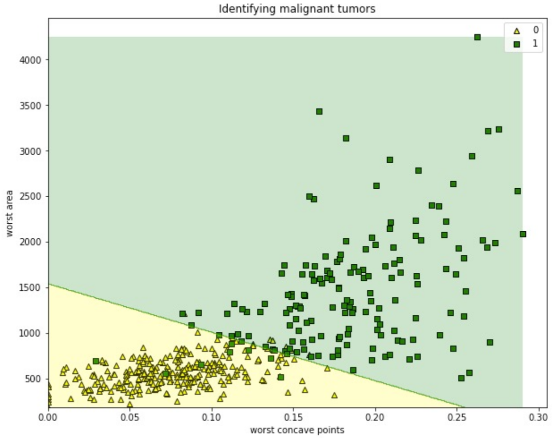
Regression

- Predict one or more **numerical** target variables
- E.g. home price, number of power outages, product demand



Classification

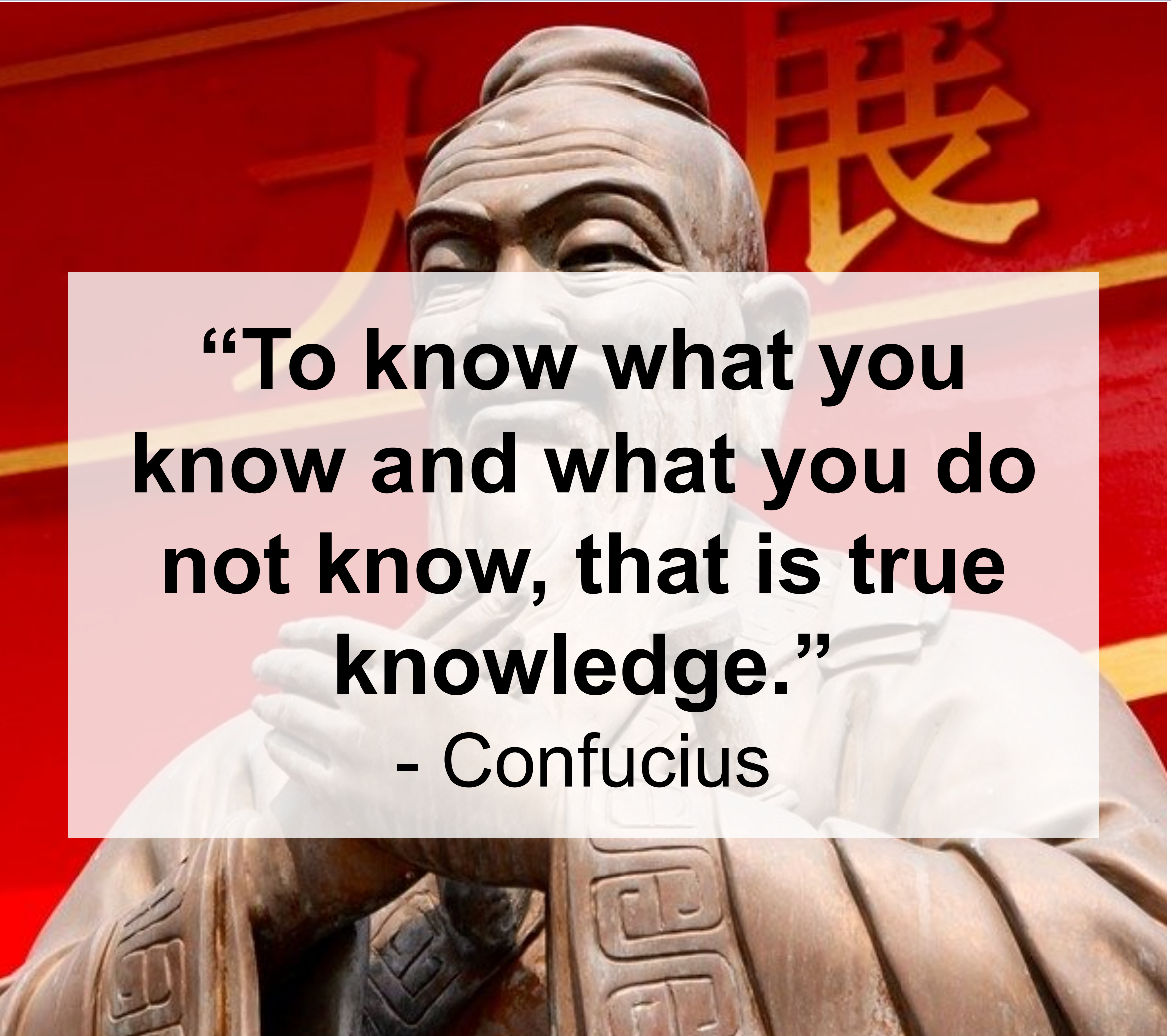
- Predicts a **class / category** – either binary or out of a set
- E.g. lung disease detection, identifying types of plants, sentiment analysis, detecting spam



What ML Can and Cannot Do Well

outrageously AMBITIOUS

Duke
PRATT SCHOOL of ENGINEERING



“To know what you know and what you do not know, that is true knowledge.”

- Confucius

What ML can do well*

- Automate straightforward tasks
- Make predictions by learning input-output relationships
- Personalize for individual users

* Given sufficient quantity and quality of data

What ML cannot do well

- Understand context
- Determine causation
- Explain “why” things happen
- Determine the impact of interventions / find solutions

outrageously
AMBITIOUS

Wrap-up

Duke
PRATT SCHOOL of ENGINEERING

Wrap Up

- ML enables computers to learn from experience without explicit instructions
  - **Data is the key!**
- Types: supervised vs. unsupervised vs. reinforcement learning
- Useful for automation, prediction and personalization
- NOT useful for explaining “why” or “how to fix”