Introduction to Artificial Intelligence

Professor Kartik Hosanagar

Wharton
UNIVERSITY of PENNSYLVANIA

Agenda

- **Artificial Intelligence (AI) Overview**
  – What AI is, types of AI, history of AI and the expert systems approach
- **Machine Learning (ML) Overview**
  – How ML differs from AI and the three types of ML (supervised, unsupervised, and reinforcement learning)
- **Detailed View of ML**
  – ML at 30,000 feet and factors that influence accuracy in ML
- **Specific ML Methods: A Deep Dive**
  – Logistic regression, decision trees & random forests, and neural networks

Wharton
UNIVERSITY *of* PENNSYLVANIA

Artificial Intelligence Overview

- What is Artificial Intelligence
- Types of AI
- History of AI
- Expert Systems approach to AI

What is Artificial Intelligence (AI)?

- Definition of AI
  - “the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.” – Merriam Webster
- AI requires knowledge of language, visual perception, reasoning, and learning.

At its core, AI is about getting computers to do things that require human intelligence.

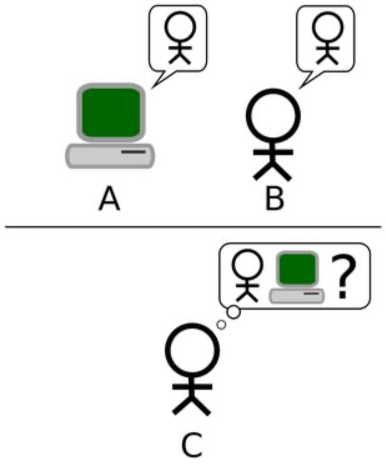
Wharton
UNIVERSITY of PENNSYLVANIA

Intelligence on a Spectrum

- AI can refer to many different types of intelligence:
  - **Artificial Narrow Intelligence (weak AI):**
    - AI that is very good at one specific task
      - A chess algorithm that can beat any human
      - Amazon's book recommendations
  - **Artificial General Intelligence (strong AI):**
    - A computer program that could do all intelligent things a human could do, just as quickly and easily
      - Artificial Neural Networks are the closest thing to reaching this level
  - **Artificial Super Intelligence:**
    - A computer program that can rapidly improve itself and do all things a human could do at a significant increase in speed and competency

History of AI

- Origins
  - "My contention is that machines can be constructed which will simulate the behavior of the human mind very closely."
    - Alan Turing, Cryptography
  - Can machines think?
    - Proposed the Turing test for machine intelligence (1950): can machines do well in imitation games?



- Soon after, a workshop called the Dartmouth Summer Research Project on Artificial Intelligence was held, which historians believe is what coined the term AI.
- "Calling it AI made it extremely ambitious, and it inspired many people to enter the field, which has been responsible for a lot of the progress." (Pedro Domingos)

Content/quote: "A Human's Guide to Machine Intelligence"

AI in the Press

- Much of what the general public knows about AI is about AI playing popular games, which has been widely covered in the press.

The New York Times

Computer Wins on 'Jeopardy!': Trivial, It's Not

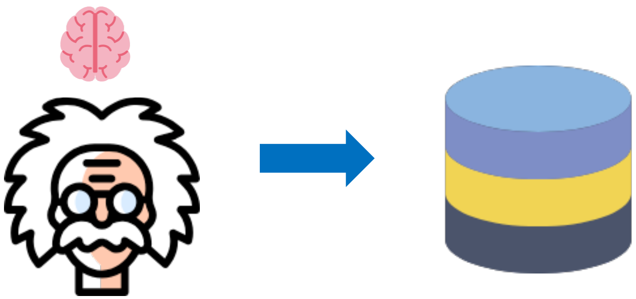
By JOHN MARKOFF FEB. 16, 2011

Progression of AI in Games

- **IBM's Deep Blue**
  – May 11, 1997, IBM's Chess Playing computer beat #1 in the World Garry Kasparov
  – No ML in system: Ability to analyze > 200,000 moves per second
- **IBM's Watson**
  – 2011, IBM Watson beats Ken Jennings and Brad Rutter (two of Jeopardy's best)
  – Using Natural Language Processing (NLP) & Question Answer (QA) information retrieval
- **Google + DeepMind's AlphaGO**
  – Wins 4:1 v World Champion Lee Sedol in GO
  – Techniques: Machine Learning (Neural Nets & Reinforcement Learning)

How to Build AI: Expert Systems Approach

- Expert systems are an older approach to AI that involves transferring knowledge from experts to a knowledge base.



- Limitations of expert systems: tacit knowledge
  - Doctors arrive at diagnoses in seconds based off pattern matching, not rules.
  - Drivers cannot articulate all the rules they use to drive cars
  - Drivers can’t articulate all their rules for driving, but videos can capture their actual driving behavior.
- Machine learning (ML) is a newer approach to AI that addresses the limitations of expert systems.

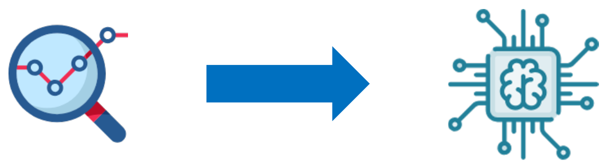
Wharton
UNIVERSITY of PENNSYLVANIA

Machine Learning Overview

- Overview
- Types of Machine Learning
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning

What is Machine Learning?

- ML is a subfield of AI
  - ML methods are characterized by their ability to learn from data without being explicitly programmed.
  - ML is most concerned with making predictions.



PREDICTION TASKS

Structured Data: “Is a transaction fraudulent?”

Text: “Is an email spam?”

Images: Image recognition in driverless cars

Audio: Speech recognition

APPLICATIONS OF ML

Healthcare: Automated medical diagnosis

HR: Which applicants are best suited for a job

Tech: Voice interfaces, Autonomous cars, personalization

Finance: Investing

Wharton
UNIVERSITY of PENNSYLVANIA

Three Types of Machine Learning

1. Supervised Learning

Develop predictive model based on input & output data

- Classification
- Regression

2. Unsupervised Learning

Group & interpret observations based only on input data

- Anomaly Detection
- Clustering

3. Reinforcement Learning

Acquire new data by taking actions & receiving ad hoc feedback

- Bandit Algorithms
- Q-learning

Wharton
University *of* Pennsylvania

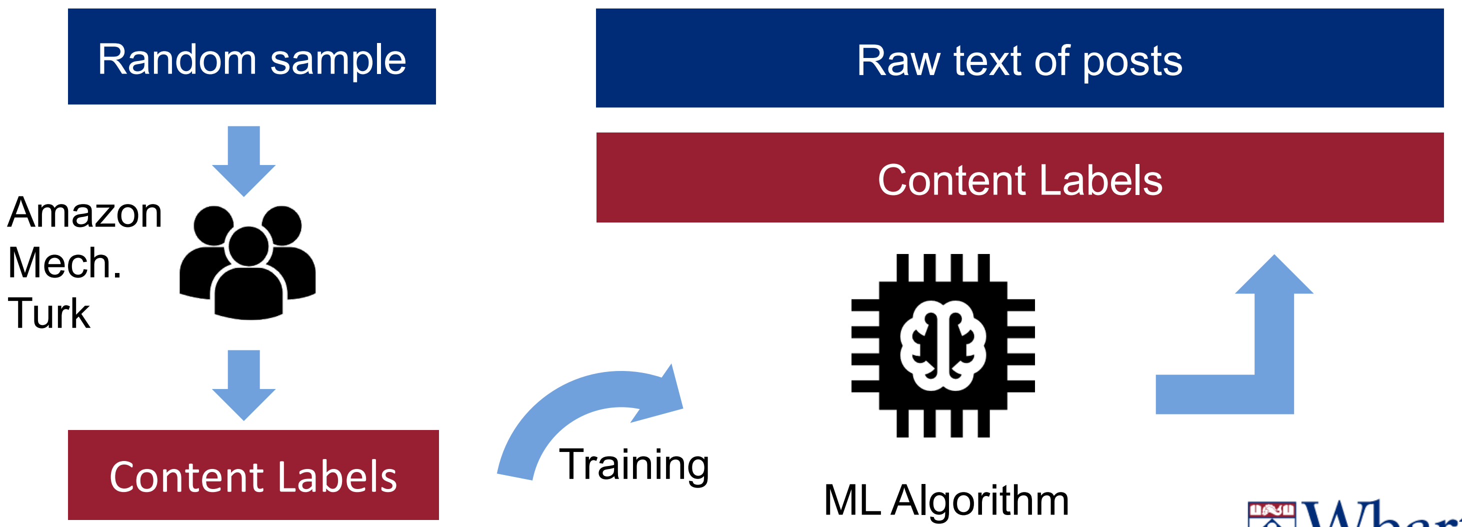
Supervised Learning

- Goal is to build a model that predicts an output (called a *target*) given input data (features/covariates). The model is given a training dataset with labeled inputs and outputs.
  - Used when there is a clearly labeled “correct” answer in the data
- Why would you want to do this?
  - Allows you to use historical data (where you already know what happened) to predict future outcomes
    - Example, forecasting sales based on month, weather, etc

Wharton
UNIVERSITY *of* PENNSYLVANIA

Supervised Learning Example

- Concrete example: Tagging text content on Facebook
  - 106,316 Facebook posts; wanted to identify what types of posts were associated with highest engagement
  - Manually tagging content is expensive so we want to automate it
  - Solution: Have a random sample tagged manually on Amazon Mechanical Turk & use as a training dataset for supervised ML



Wharton
University of Pennsylvania

NLP Algorithm Performance

	Accuracy	Precision	Recall
REMFACT	0.998	0.998	0.998
EMOTION	0.996	0.992	0.999
HUMOR	0.999	0.999	1
PHILANTHROPIC	0.999	0.999	1
FRIENDLIKELY	0.997	0.996	0.998
SMALLTALK	0.858	0.884	0.803
DEAL	0.996	0.999	0.994
PRICECOMPARE	0.999	0.999	1
TARGETING	0.999	0.998	1
PRODAVAILABILITY	0.999	0.998	1
PRODLOCATION	0.970	0.999	0.901

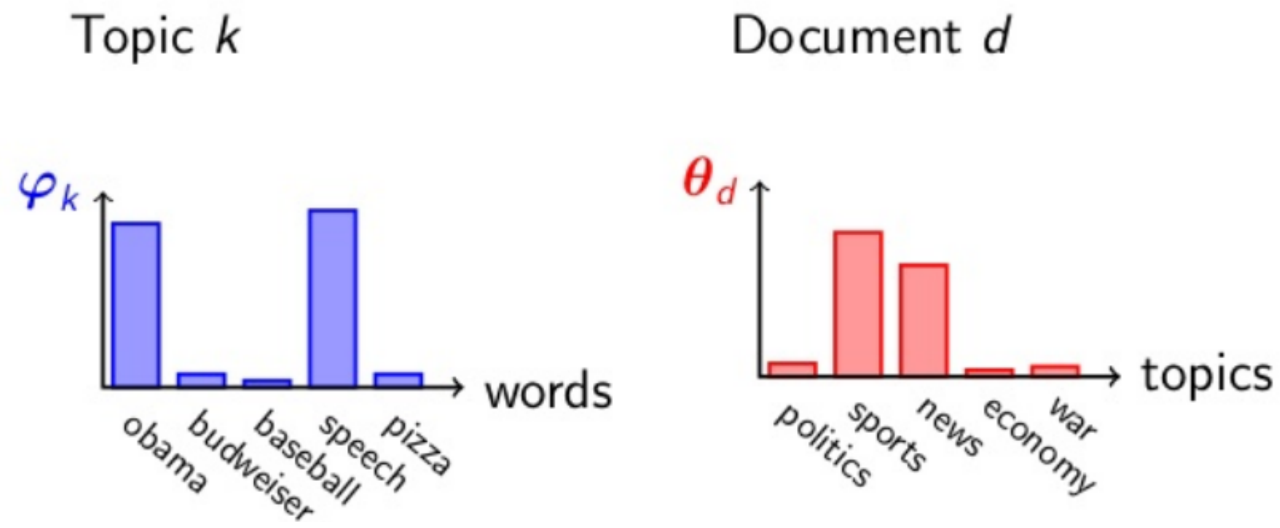
Algorithm described in detail at:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2290802

Unsupervised Learning

- No fixed set of outputs provided; must be learned from inputs
- Goal is to cluster and identify important features
- Why would you want to do this?
  - Data is too large to label manually or it is expensive to label data and create a training set
  - Absolute precision isn't needed
  - Automated data exploration
- Example: topic analysis from web page content
  - Latent Dirichlet Allocation (LDA)

Unsupervised Learning

- Latent Dirichlet Allocation (LDA)
  - Takes as input a set of documents
  - Identifies both common topics across set of documents, and
  - Labels documents with corresponding topics



Unsupervised Learning

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

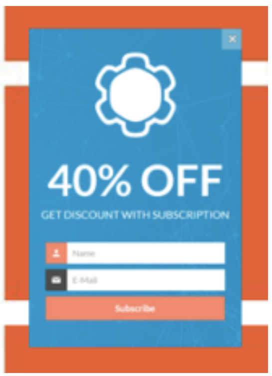
Reinforcement Learning

- Idea: Let algorithms learn by testing various actions/strategies and deciding which one works best
  - Very powerful method for simulation and robotics- based applications
    - At the hear of AlphaGo and other state-of-the-art gaming AI
- In many applications experimentation isn't free!
  - You often want to balance both exploration and exploitation
  - Multi-armed bandit algorithms

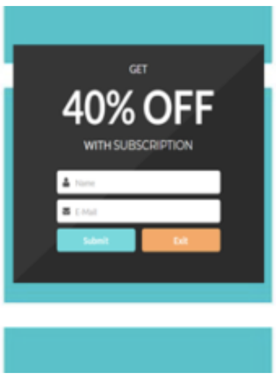
Wharton
UNIVERSITY of PENNSYLVANIA

Reinforcement Learning Example: Bandits

- Suppose you have two ad copies and you don't know which will attract more clicks (and therefore visitors to your website).
- Traditional A/B testing involves showing ad A 50% of the time & ad B 50% of the time, and then assessing which ad performed better.

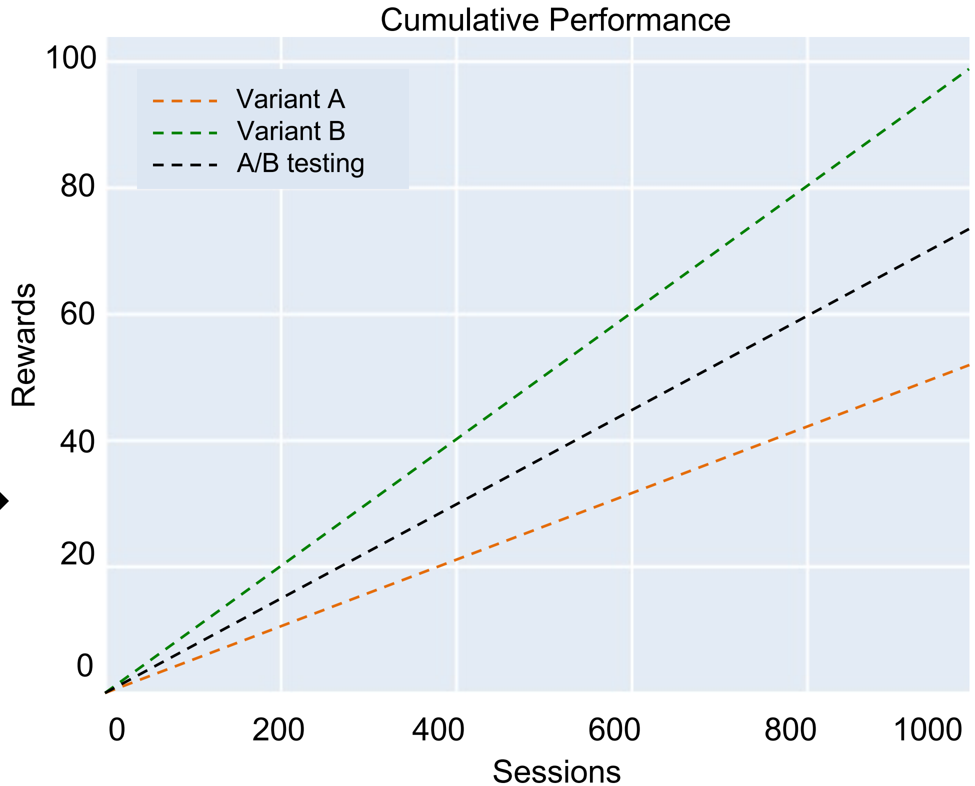


Ad A



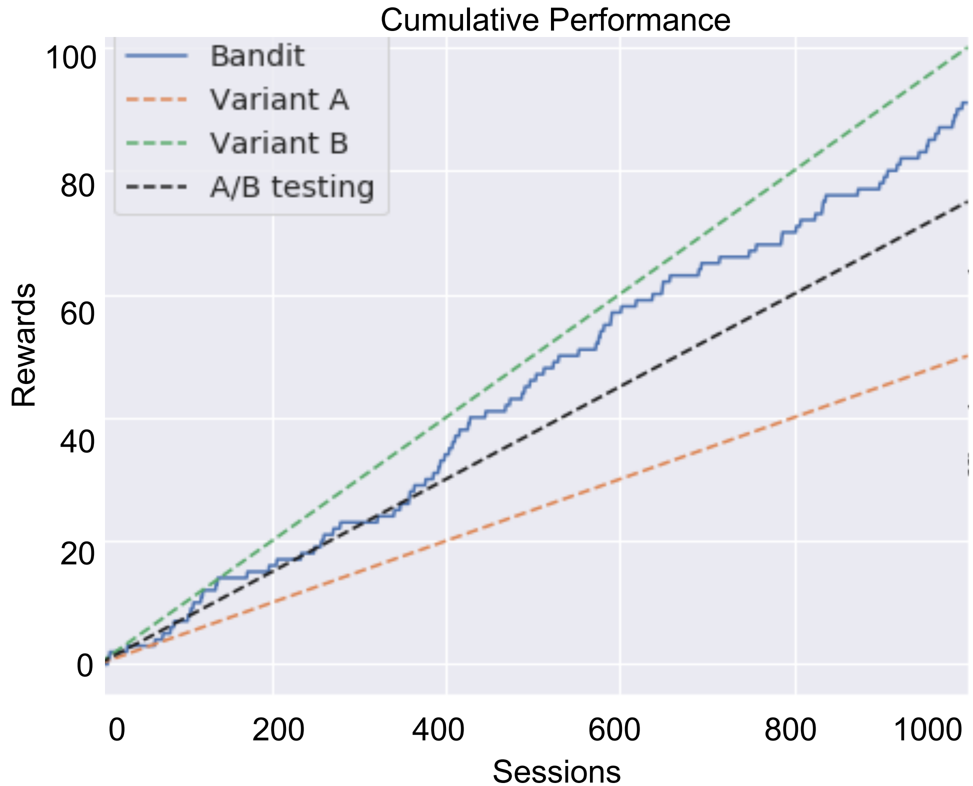
Ad B

B performs better than A. The A/B test line is midway between the A & B lines as it is 50% A and 50% B.



Reinforcement Learning Example: Bandits

- Machine learning can improve upon A/B testing through Bandit algorithms.
- Bandit algorithms update beliefs based upon performance.
  - They spend more time on best performers early on while still learning and improving over time.
    - The bandit begins by showing 50% A & 50% B, but slowly starts allocating more & more traffic to the higher-performing ad as it learns & confirms which one is better.



A Detailed View of ML

- ML at 30,000 feet
- What influences accuracy in ML

ML at 30,000 Feet

- Focus on supervised learning:
  - At its core, supervised machine learning is about using “a set of variables” to predict “an outcome”
  - You do this all the time:
    - Dark clouds + strong winds → rain
    - What someone is wearing, how they interact with you → whether you’ll be friends
    - Resume (school, experience, skills) → good employee
  - Applications in business are abundant:
    - Will someone buy your product?
    - Will someone click your ad?
    - Will this machine fail in the next day?

Wharton
UNIVERSITY of PENNSYLVANIA

ML at 30,000 Feet

- Focus on supervised learning:
  - At its core, supervised machine learning is about using "a set of variables" to predict "an outcome"

UserID	PgsViewed	ZipCode	OS
1	3	19104	iOS
2	10	19141	Windows
…	…	…	…

→

Purchase
Yes
No
…

- Names for input data:
  - Predictors, features, "data", variables, characteristics, covariates

ML at 30,000 Feet

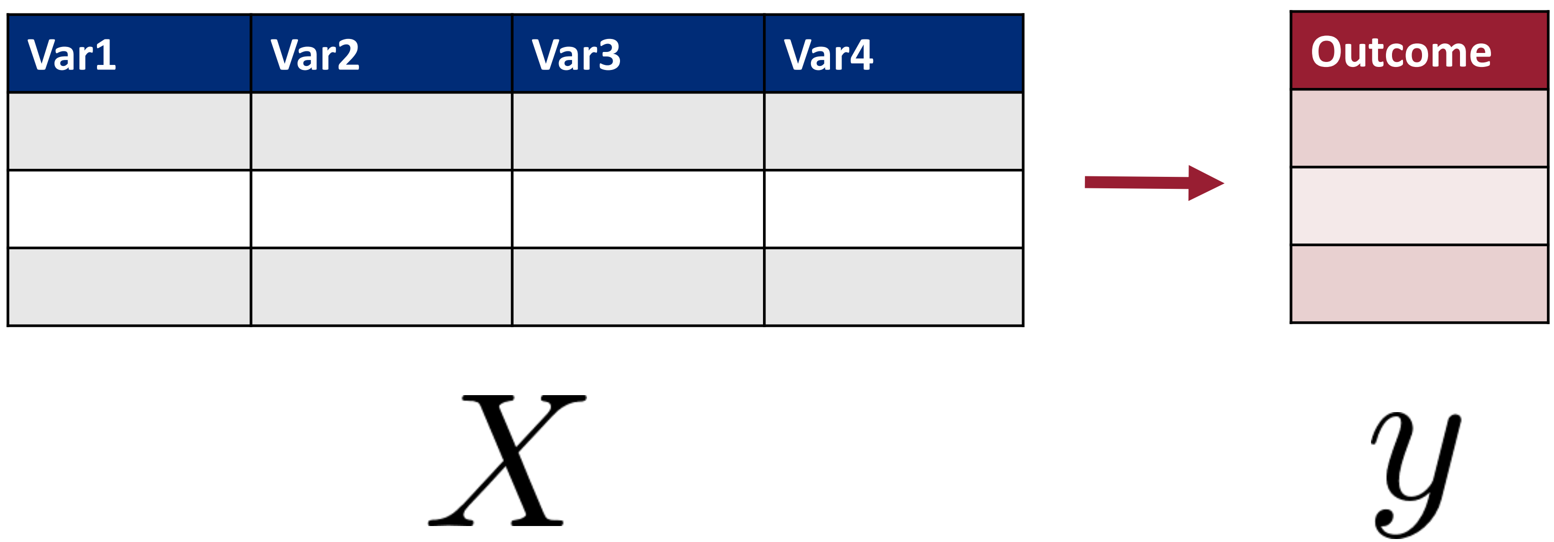
- Focus on supervised learning:
  - At its core, supervised machine learning is about using “a set of variables” to predict “an outcome”

Var1	Var2	Var3	Var4

Outcome

ML at 30,000 Feet

- Focus on supervised learning:
  - At its core, supervised machine learning is about using “a set of variables” to predict “an outcome”



ML at 30,000 Feet

- Focus on supervised learning:
  - At its core, supervised machine learning is about using “a set of variables” to predict “an outcome”

Var1	Var2	Var3	Var4

Outcome

$$f(X) = y$$

ML at 30,000 Feet

- Focus on supervised learning:
  - At its core, supervised machine learning is about using “a set of variables” to predict “an outcome”

Var1	Var2	Var3	Var4

→

Outcome

$$f(X) = y$$

All of supervised ML comes down to approximating this function with high fidelity

What Influences Accuracy in ML?

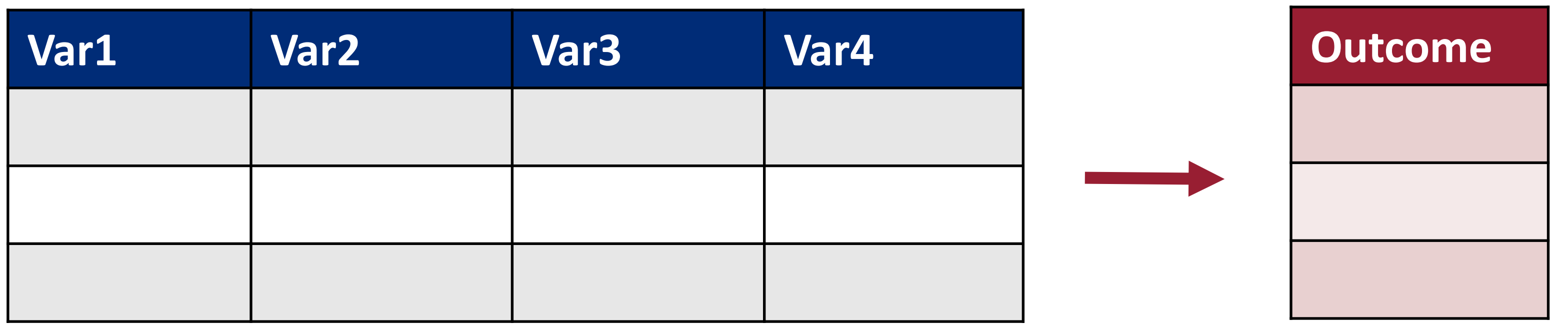
- What is accuracy?
  - Ability to make correct predictions on data you haven't seen

Var1	Var2	Var3	Var4

Outcome

What Influences Accuracy in ML?

- Quantity of data
  - Number of distinct observations



What Influences Accuracy in ML?

- Quantity of data
  - Number of distinct observations

Var1	Var2	Var3	Var4

Outcome

Wharton
University of Pennsylvania

What Influences Accuracy in ML?

- Quantity of data
  - Number of distinct observations
  - Different characteristics about your observations

V1	V2	V3	V4

Outcome

Wharton
University of Pennsylvania

What Influences Accuracy in ML?

- Quantity of data
  - Number of distinct observations
  - Different characteristics about your observations

V1	V2	V3	V4	V5	V6	V7	V8

Outcome

What Influences Accuracy in ML?

- Quantity of data (number of rows)
- Number of features (columns)

What Influences Accuracy in ML?

- Quantity of data
- Number of features
- Lots of other things!
  - Relevance of information to underlying phenomena
    - Umbrellas predict rain better than color of a person's dress
    - Does future data look like past data?
  - Complexity of your model (how we approximate "*f*")
    - Modern developments in "deep learning" allow for very complex models that were historically very hard to train
  - "Feature engineering"
    - Using domain knowledge to create new features from the input data (some of it can be automated but the analyst has a big role to play)

Wharton
UNIVERSITY *of* PENNSYLVANIA

Specific ML Methods: A Deep Dive

- Logistic Regression
- Decision Trees & Random Forests
- Neural Networks
- Model Selection

Note to students: More technical than other parts of this course. But if you have taken a basic stats course on regression modeling, I believe you'll be able to follow.

A Closer Look at ML Methods

- Logistic regression
- Decision trees and random forests
- Neural nets

All different ways of approximating the relationship between X and y

$$f(X) = y$$

Logistic Regression

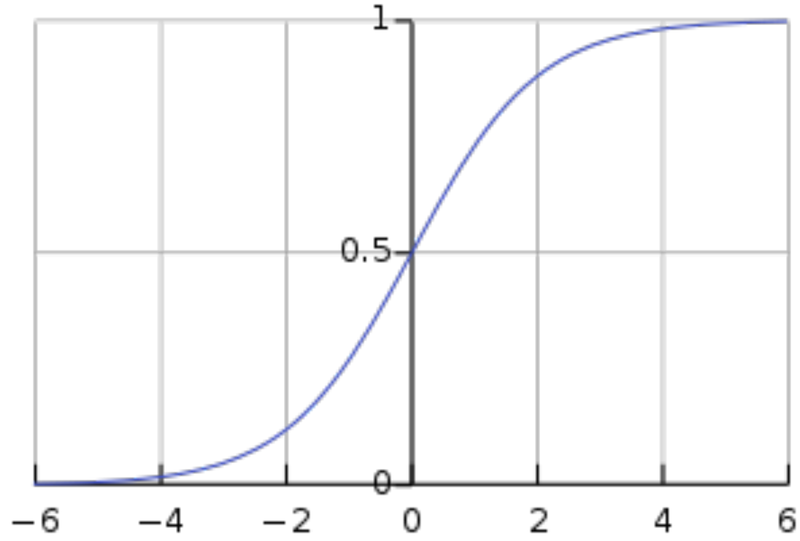
- Used for binary classification, i.e. when outcomes can take only 1 of 2 values (yes/no, click/no click, health/sick)
  - Contrast with continuous variables such as revenue or weight
- Among the most useful and popular methods (along with ordinary least squares) in all of statistics, data science, academia
- Is it really ML?
  - Early development in 19^{th} century

Logistic Regression

- Goal is to estimate the probability of a given outcome, conditional on some variables (x):

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i}$$

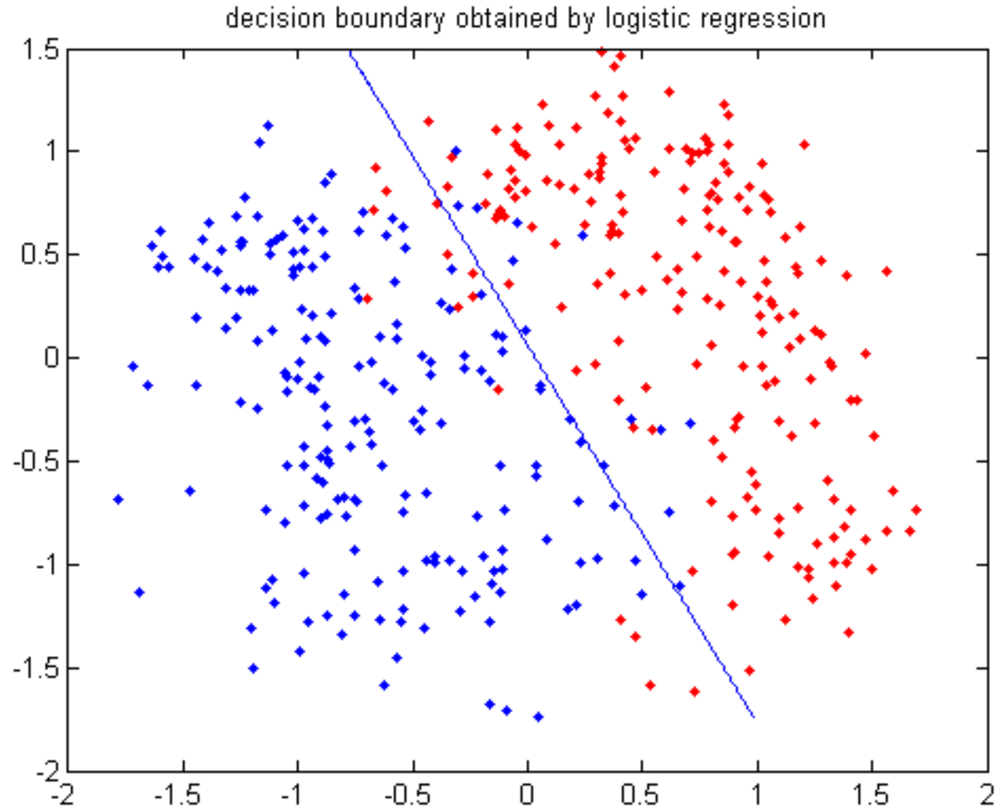
- Logit function constrains probabilities to between 0 and 1



Logistic Regression

- Equivalent to finding the "best-fit" line/plane that separates the data

- Example: how do age and income predict whether a customer will purchase your new product?



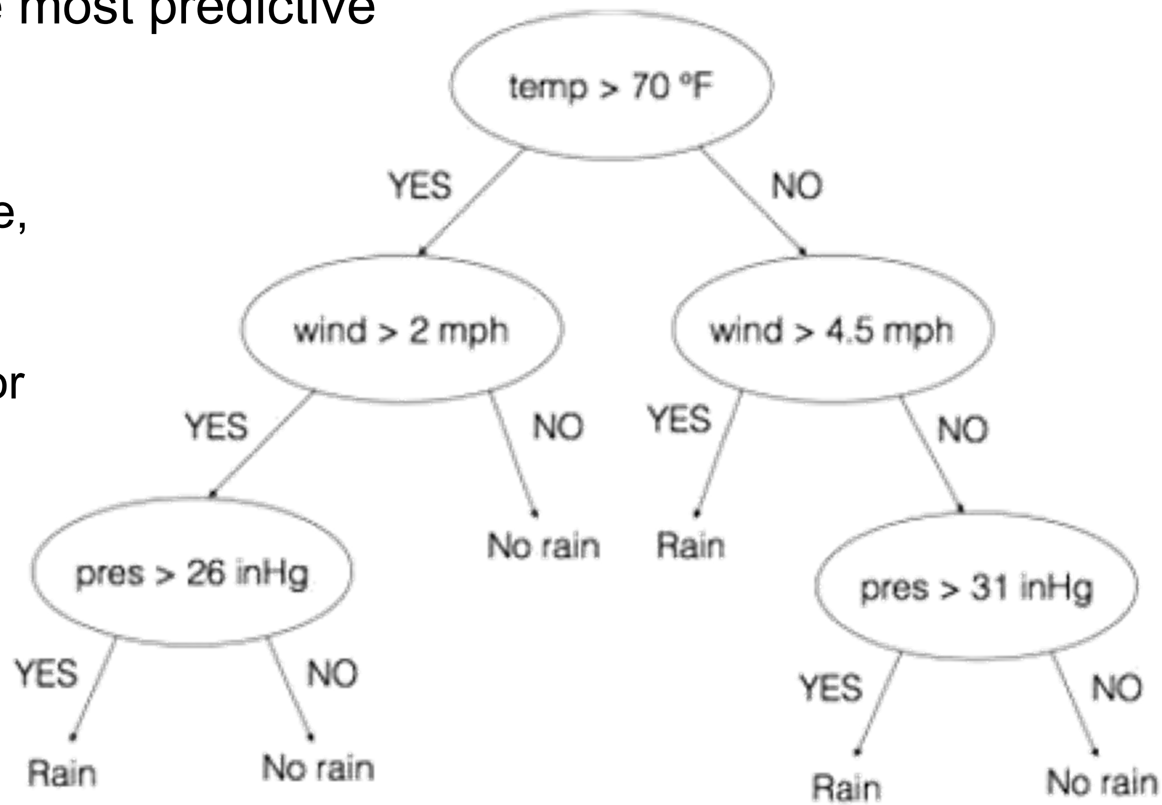
- Red dots: purchase
- Blue dots: no purchase

Decision Trees

- Easy-to-interpret models built by iteratively looking for features in your data that are most predictive
- Example:

  Inputs: Temperature, Wind, Pressure

  Output: Will it rain or not today?



Decision Trees

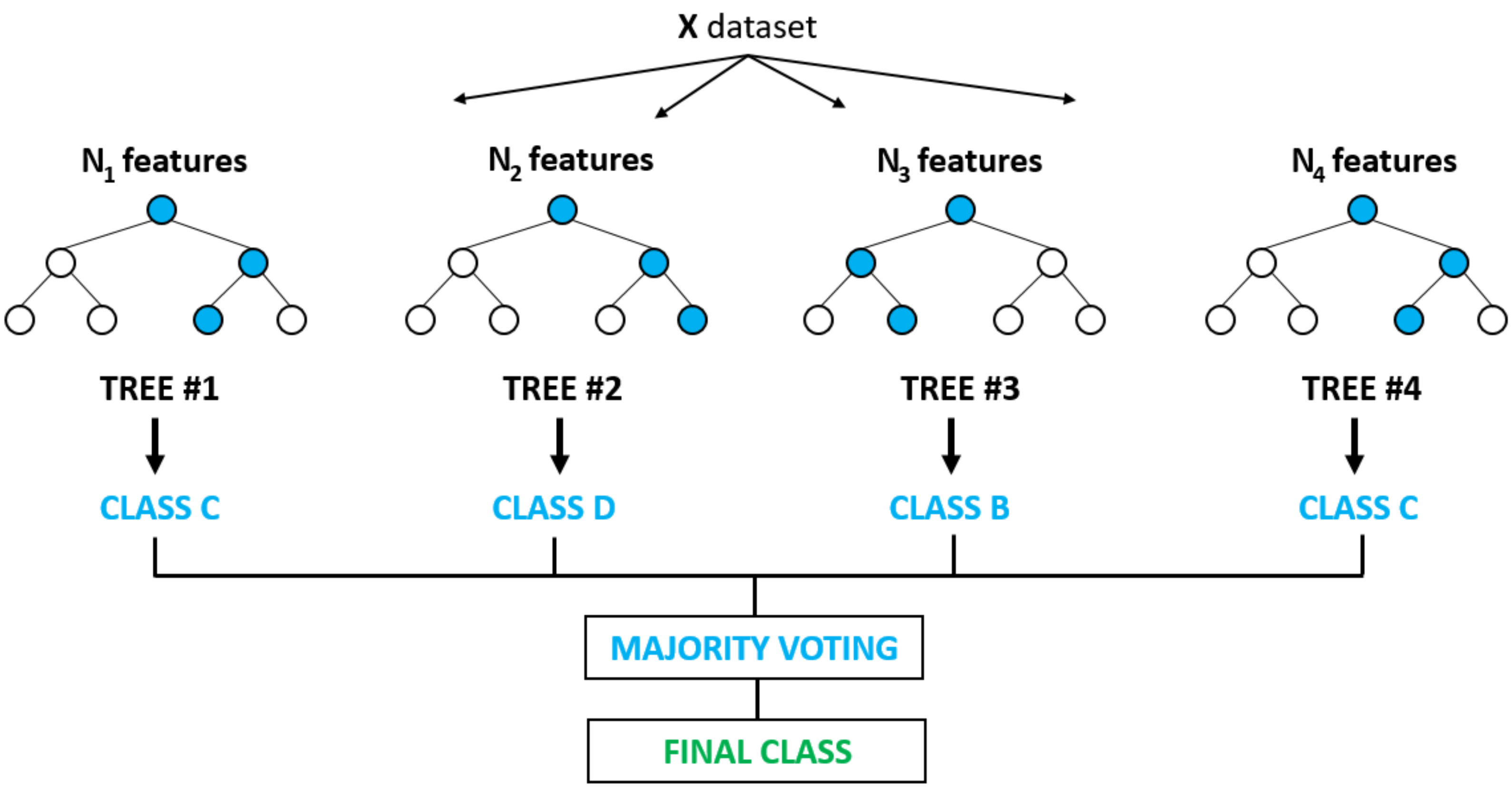
- **Question:** How to decide which variables to split on at each branch? (e.g. consider temperature or pressure first?)
  - Also: how to select what value to split on?
    - e.g. 70 degrees vs 60 degrees

- **Mathematical Answer:** Select variables/splits that minimize "entropy" of the dataset
  - In simple terms, choose a variable/ split that provides the most predictive power at each step
  - A "greedy" algorithm that looks for the best split at each step in the process
  - Continue to split branches on different variables until a desired level of depth or data size limits are reached

Random Forests

- An "ensemble" algorithm that harnesses the power of many decision trees
- Very popular, powerful and relatively simple ML algorithm
- Basic idea:
  - Take many random samples of your dataset
  - For each subset of data, train a decision tree
    - For each node in tree, only use a random subset of features
  - To make final decision, treat each decision tree as a "vote" and choose the prediction with the most votes among all sub-trees

Random Forests

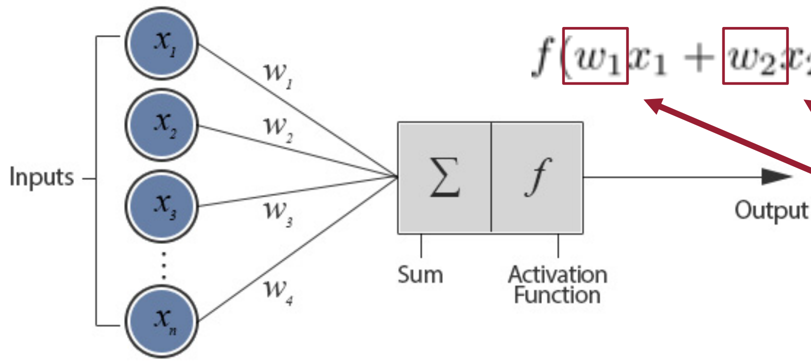


Wharton
University of Pennsylvania

Neural Networks (NNs)

- Loosely inspired by biological neurons
  - Neurons take input from other neurons, apply transformation & pass signal on

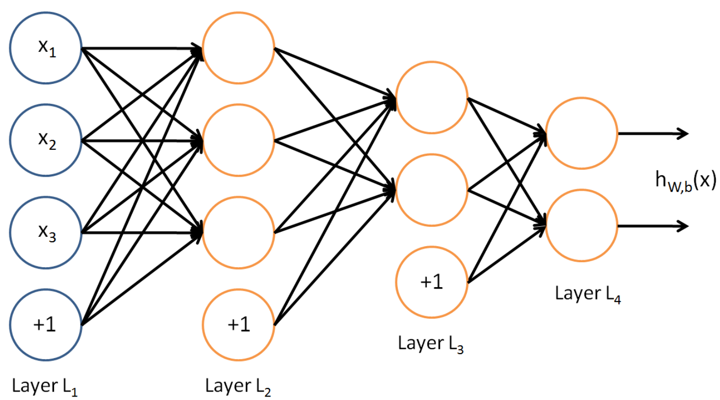
- Single "Neuron"



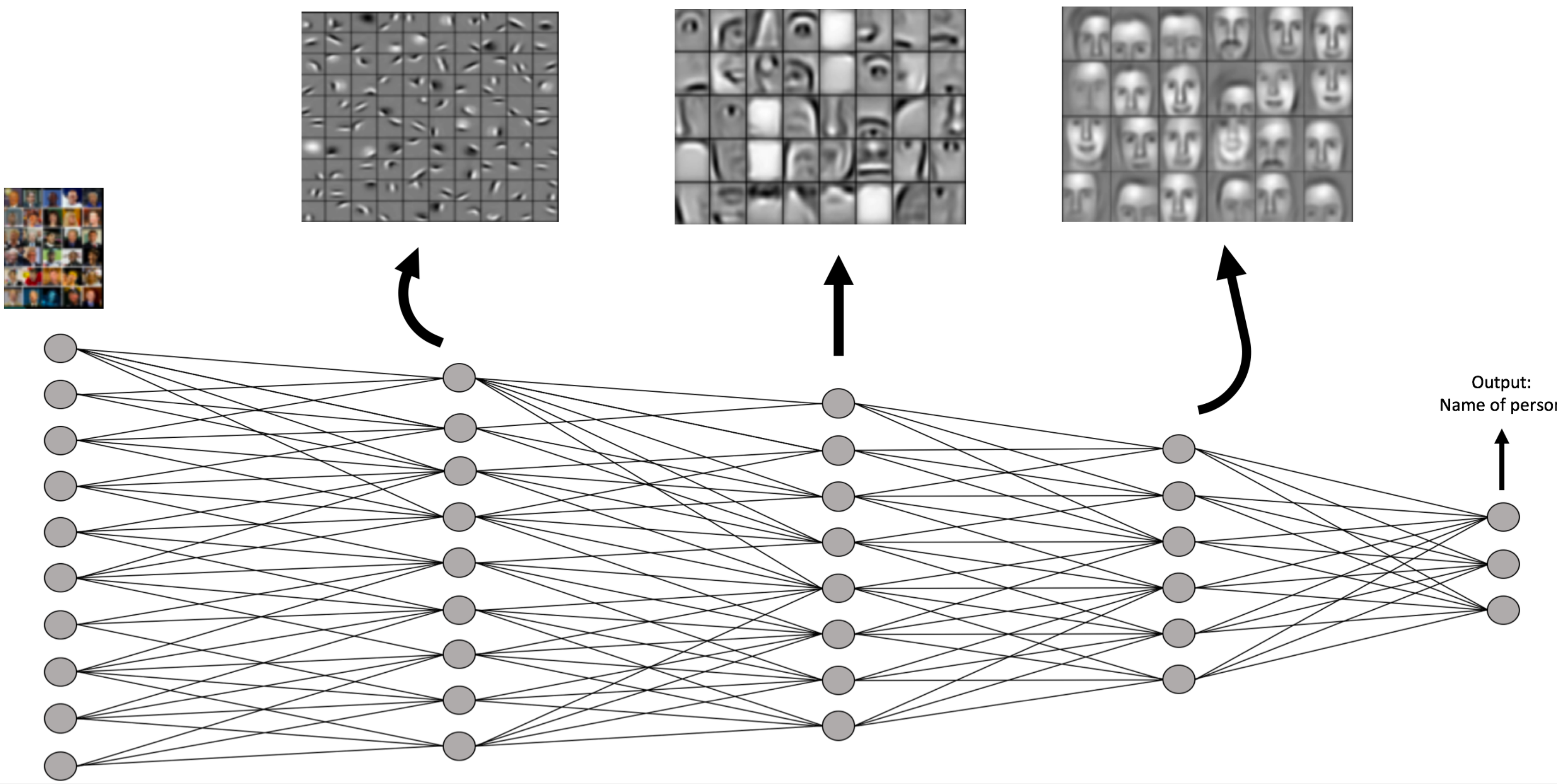
$$f(\boxed{w_1}x_1 + \boxed{w_2}x_2 + \cdots + \boxed{w_n}x_n)$$

"weights"
"parameters"

- Deep Neural Net



Neural Network for Face Recognition



See https://www.youtube.com/watch?v=aircAruvnKk for a video tutorial

Wharton
UNIVERSITY of PENNSYLVANIA

Neural Networks (NNs)

- Layers of a deep neural network
  - Input layer: Neurons correspond to the RGB value of all the pixels in the image
  - Final layer has the outputs we expect (e.g. 3 neurons corresponding to three different people who are going to be identified)
  - Hidden layers in between
  - See https://www.youtube.com/watch?v=aircAruvnKk for a simple video tutorial

Wharton
University of Pennsylvania

Neural Networks (NNs)

- NNs have become very successful in recent years
  - Are often among the best algorithms in top ML competitions (especially with images, audio, video, etc)
- Why?
  - Lots of parameters = ability to build very complex models
  - Recent advances in computation (GPUs) and algorithms (backpropagation) has allowed for more layers (more complex models)
- NNs are hard to understand and interpret
  - We know inputs and outputs, but the middle layers are merely numbers from our point of view
  - Lots of work is being done to “open up” black box models so we can understand what they are doing

Other ML Methods

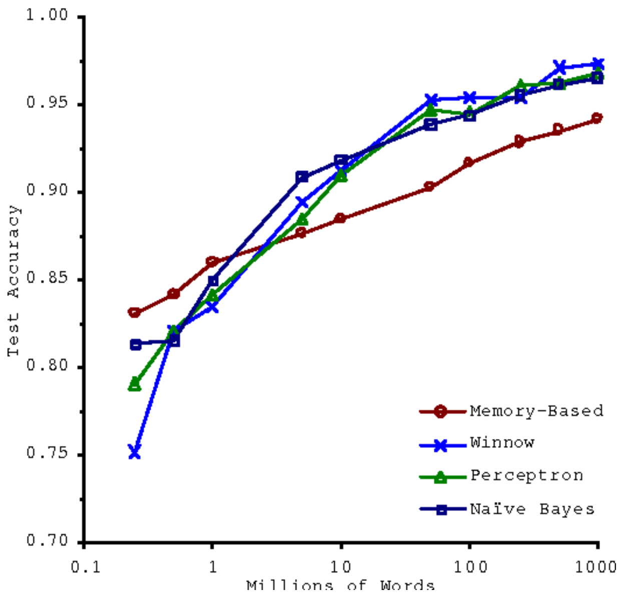
- Boosting
- Support vector machines (SVM)
- Neural Nets: much more complicated & varied than covered here
- Many, many other regression techniques
  - LASSO, Ridge, weighted regression, kernel regression
- Want to learn more?
  - Statistics or Intro CS courses on Machine Learning

Model Selection

- For any prediction problem, there are many algorithms and methods available: Decision Trees, Random Forests, Neural Networks, and more
- Model evaluation and selection is done by evaluating model performance on a validation dataset
  - Holdout validation: Partition available data into a training dataset and a holdout; evaluate model performance on holdout
  - Cross-validation: create a number of partitions (validation datasets) from the training dataset; fit model to the training dataset (sans the validation data); evaluate model against each validation dataset; repeat with each validation set and average results to obtain the cross-validation error.

Data vs Model

- Often Data > Methods
  - Microsoft researchers (Banco and Brill) evaluated performance of multiple models for a language understanding task
  - Varied size of training dataset (up to 1B words)
  - Among modern methods, performance differences between algorithms are relatively small when compared to differences between same algorithms with more/less data

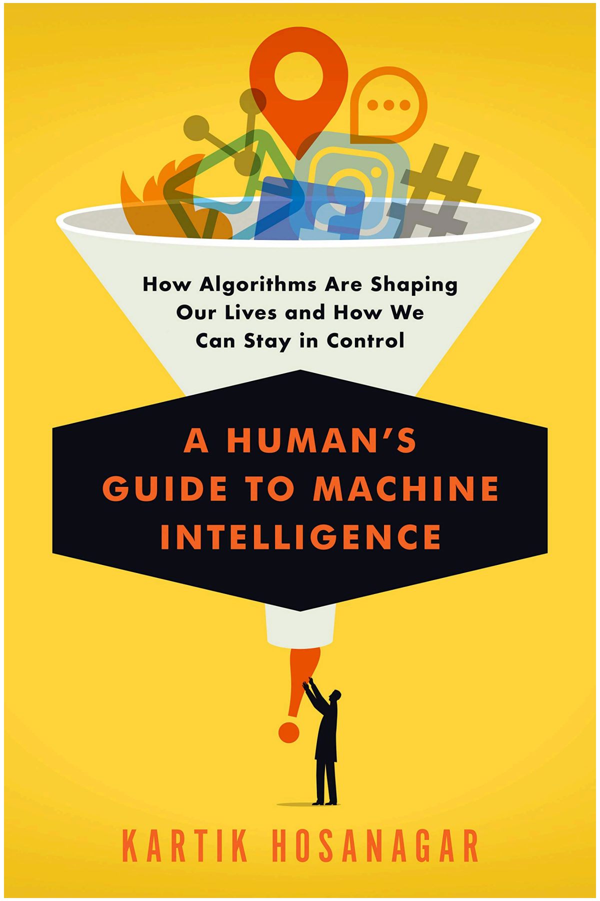


"Unreasonable effectiveness of data" – Peter Norvig (Google)

Sources:

Much of the content for these slides has been taken from the book: “A Human’s Guide to Machine Intelligence” by Kartik Hosanagar

Additional sources are cited at the bottom of individual slides. Image citations can be found in the notes section of individual slides.



Wharton
UNIVERSITY of PENNSYLVANIA