Machine Learning (CS 4/5780)

Anil Damle and Wen Sun

Administrative stuff…

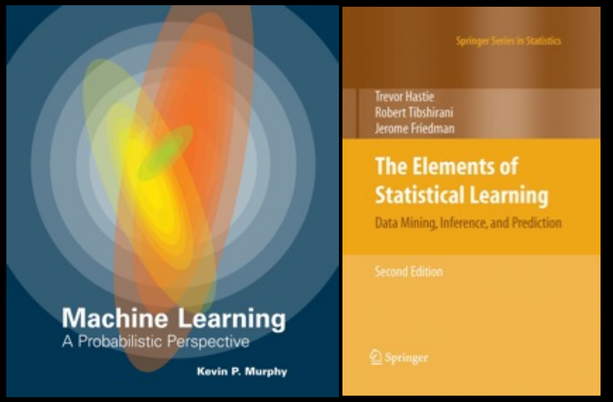
(To get it out of the way)

Course Overview

- Instructors:
  - Anil Damle and Wen Sun
- Homepage:
  - https://www.cs.cornell.edu/courses/cs4780/ (links from here to what you need)
- TAs:
  - Many (more and better than you think)
- Office Hours / Recitations:
  - TA Office Hours: **Every day** (Details will be posted on course webpage.)
  - Leave Feedback: contact Anil and Wen
  - Prof. Office Hours: **TBA**
- Questions:
  - Post all questions on ED (you can make them private)
  - **Do not email directly (except in an emergency or need for privacy)**

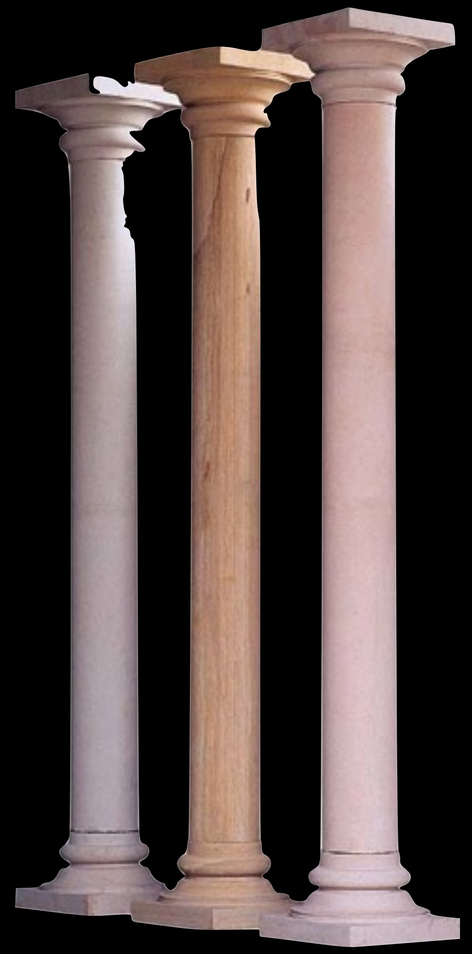
Course resources

- Primary texts
  - *Probabilistic Machine Learning: An Introduction* by K.P. Murphy
  - *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman
- Additional texts
  - *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani
  - *Patterns, Predictions, and Actions* by Hardt and Recht
  - *Fairness and Machine Learning* by Barocas, Hardt, and Narayanan
- Background and programming resources on the website



Prerequisites

- Three pillars of ML:
  - **Statistics / Probability**
  - **Linear Algebra**
  - **Multivariate Calculus**
- Should be confident in at least 1, ideally 2 or 3
- TAs might be able to give recitations on some topics if needed (but don’t rely on it)



Warm up quiz

- Due August 30!!!
- Available later today on Canvas
- **Not** a placement exam; it's designed to help you assess your comfort with prerequisite material
- If you find aspects of the quiz challenging/unfamiliar please use the background resources on the website (we may also run a recitation or two)
- Not part of your final grade, but you **must** complete it

Homework

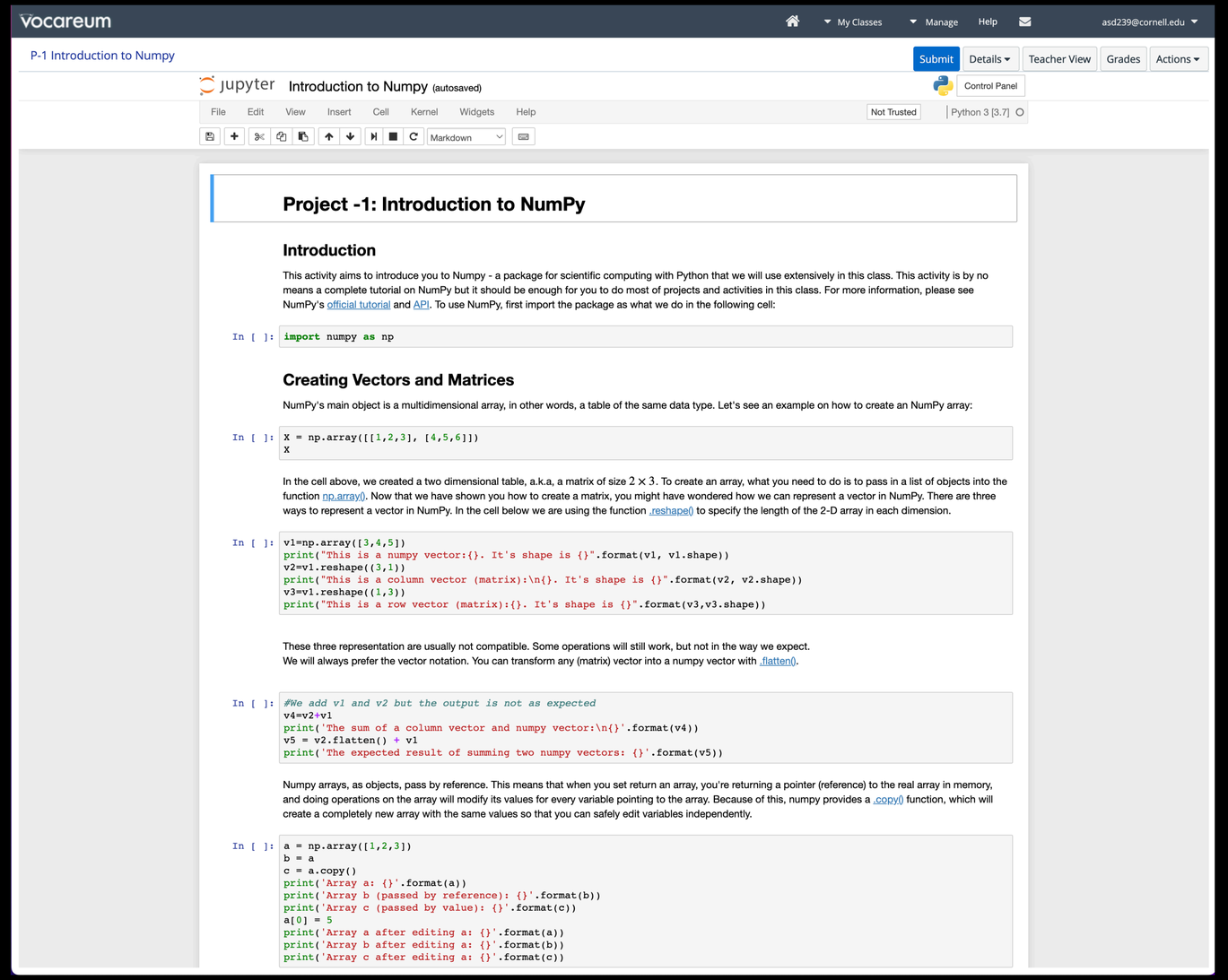
- Roughly 8 assignments (subject to change)
- Due every 1-2 weeks
- Can work in groups of up to **3**
- Submitted via Gradescope
- Primarily theoretical, they reinforce concepts from class and provide practice for the exams

Exams

- Midterm: October 18 at 7:30 pm (room details on website)
- Final: TBA (December 9-17)
- Closed book, no notes, no cheat sheet

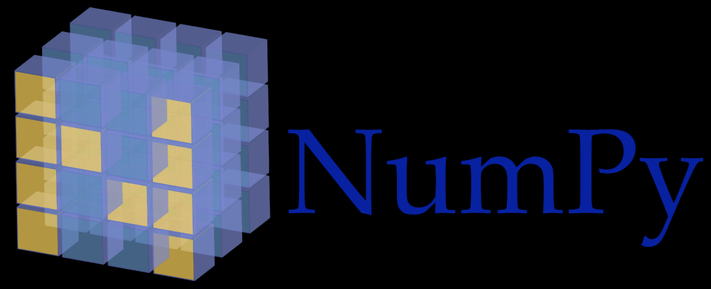
Projects

- vocareum.com
- Enrollment details coming soon
- There will be 8 (+2) projects
- Roughly **2 weeks** per project
- Unlimited submits until deadline
- Costs $30 :-(



Warm up project (Project -1)

- NUMPY proficiency test
- Will turn into your own cheat-sheet
- Please take it seriously, this is for your own good



Vocareum and its autograder

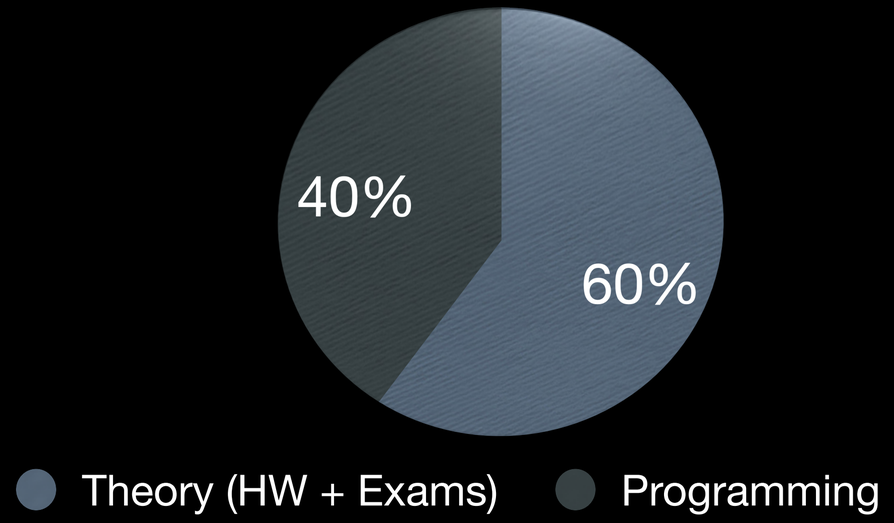
- Important notes:
  - Only text with **#<GRADED>** and **#</GRADED>** will be graded
  - **!!!!You MUST form teams before you get started!!!!!!**

For those in 5780

- Intermittent paper comprehension quizzes
- Read and answer questions on relevant ML papers
- Helps build “research comprehension” in the field
- Quizzes completed on Canvas
- Required for everyone in 5780, if you are in 4780 you can complete them if you like

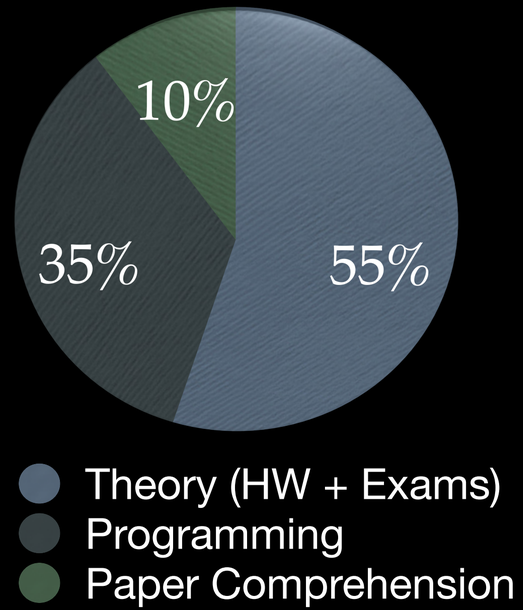
Course Grade Breakdown 4780

- 50% **Theory: Midterm + Final**
  - Closed book
  - No cheat sheets!
  - No personal notes
- 40% **Programming Assignments**
  - Up to **2** members in each team
  - **2** days extension **per team per project**
  - Autograder (unlimited resubmissions)
  - *Extra credit available at times*
- 10% **Homeworks**
  - Up to **3** members in each team
  - Preparation for exam



Course Grade Breakdown 5780

- 45% **Theory: Midterm + Final**
  - Closed book
  - No cheat sheets!
  - No personal notes
- 35% **Programming Assignments**
  - Up to **2** members in each team
  - **2** days extension **per team per project**
  - Autograder (unlimited resubmissions)
  - *Extra credit available at times*
- 10% **Paper Comprehension (mandatory)**
  - Original Research Papers in ML
  - Canvas Quizzes
- 10% **Homeworks**
  - Up to **3** members in each team
  - Preparation for exam



About this course

- Take this course if …
  - you are interested in Machine Learning
  - you are comfortable with a decent amount of mathematics
  - you are not scared of programming
- Don’t take this course if …
  - matrices scare you
  - you don’t remember how to take derivatives
- We discourage taking the course if you find the warm up quiz very unfamiliar and challenging
  - In that case, take appropriate prerequisites and we would love to see you in a future offering

Student comments (truth in advertising)

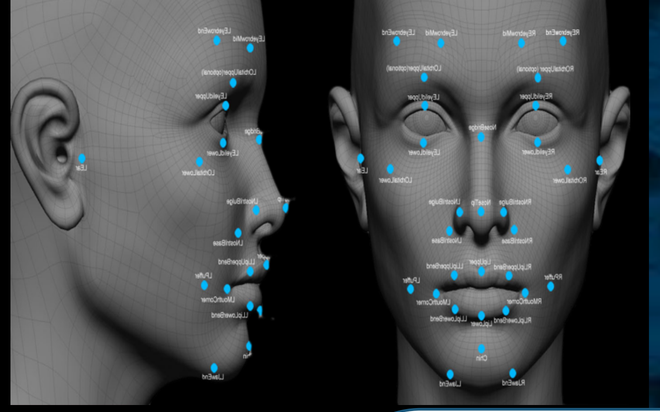
- “[…] Requires a good knowledge in math and derivatives.”
- “A TON of work, but mostly worth it for a very valuable skill.”
- “great course, but prepare to work your butt off.”
- “The topics were pretty complicated and difficult to understand quickly. I would have preferred a slightly slower pace.”
- “It's mostly a math class”

Academic Integrity

- Zero tolerance policy: all occurrences will be reported
- We **actively** look for academic conduct violations
- The autograder checks for plagiarism

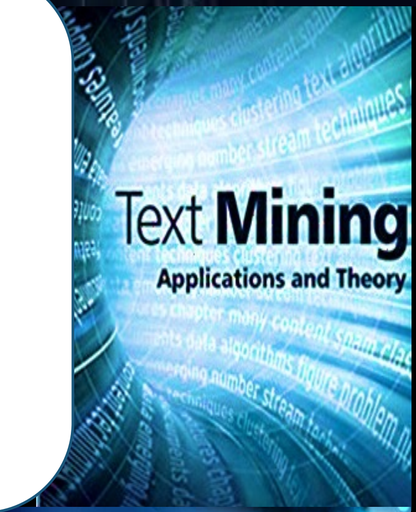
Academic Integrity

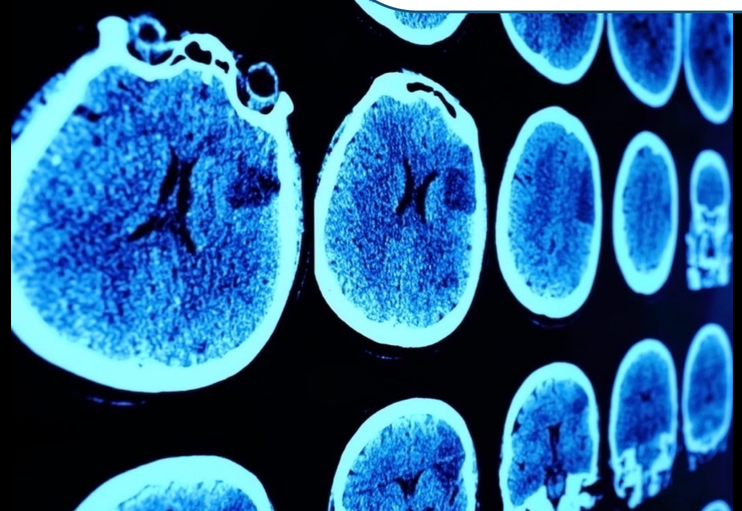
- *Examples:*
  - Most common: Students steal from same source
  - Students post to RentACoder.com or other page
  - Students post solutions on the web
  - Students use solutions from last year's course

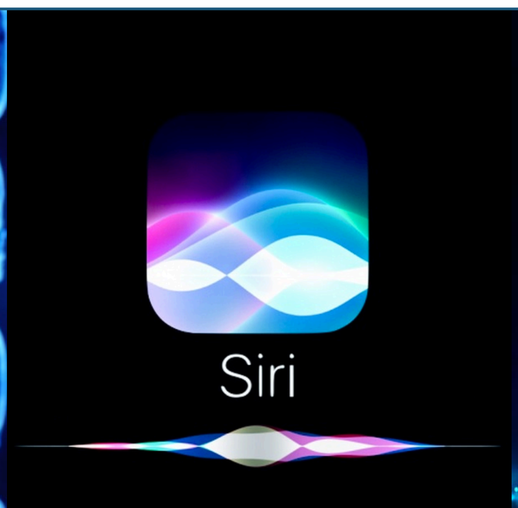


Machine Learning (ML)

Programs/models that improve with experience.

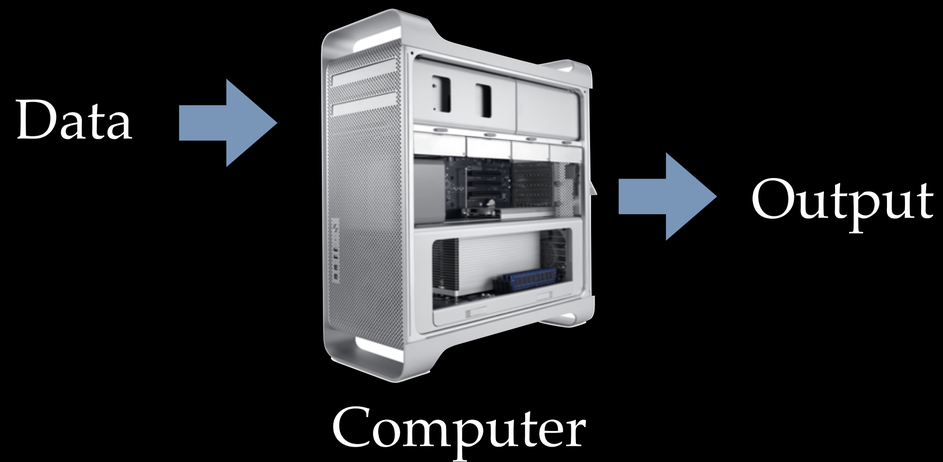






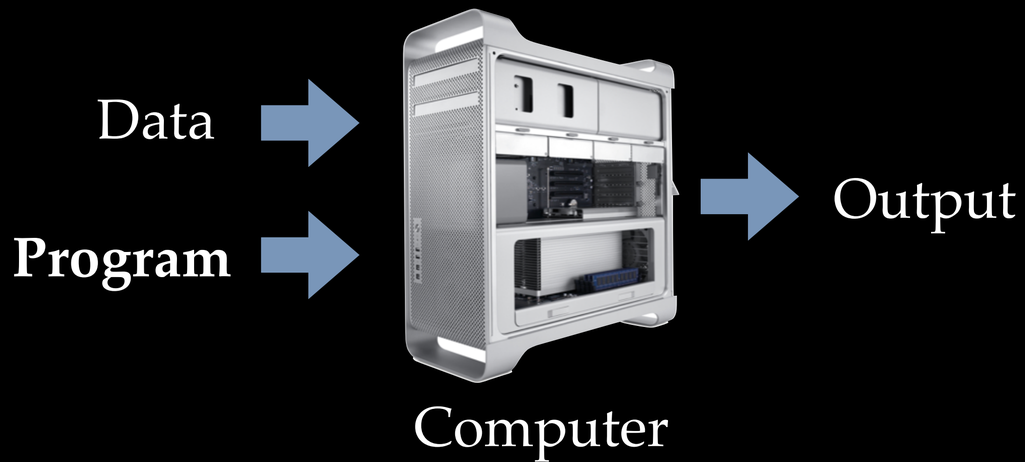
Traditional Computer Science

Traditional CS:



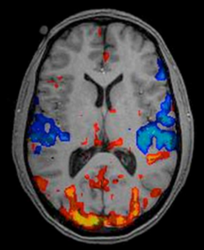
Traditional Computer Science

Traditional CS:



Traditional Computer Science

Traditional CS:



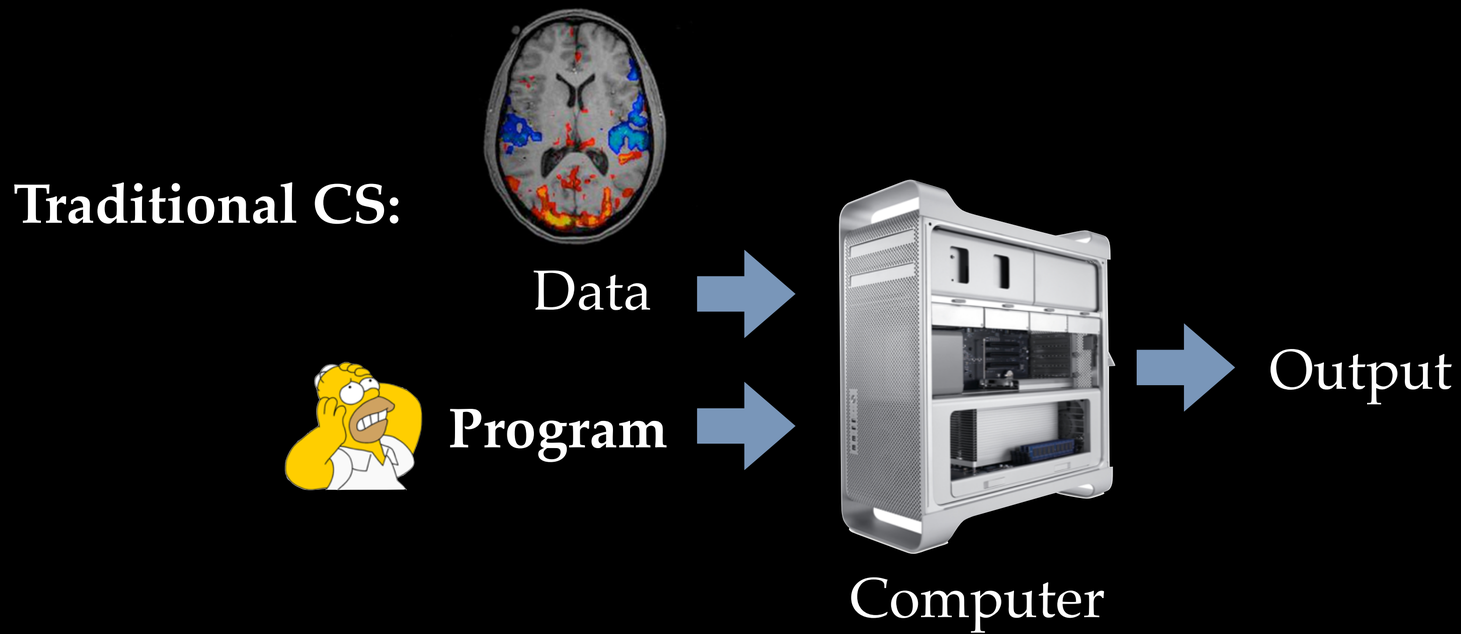
Data

Program

Computer

Output

Traditional Computer Science



Machine Learning

Traditional CS:

Data

Program

Output

Computer

Machine Learning:

Data

Output

Program (Model)

Computer

Machine Learning

Machine Learning:

Data

Output

Computer

Traditional CS:

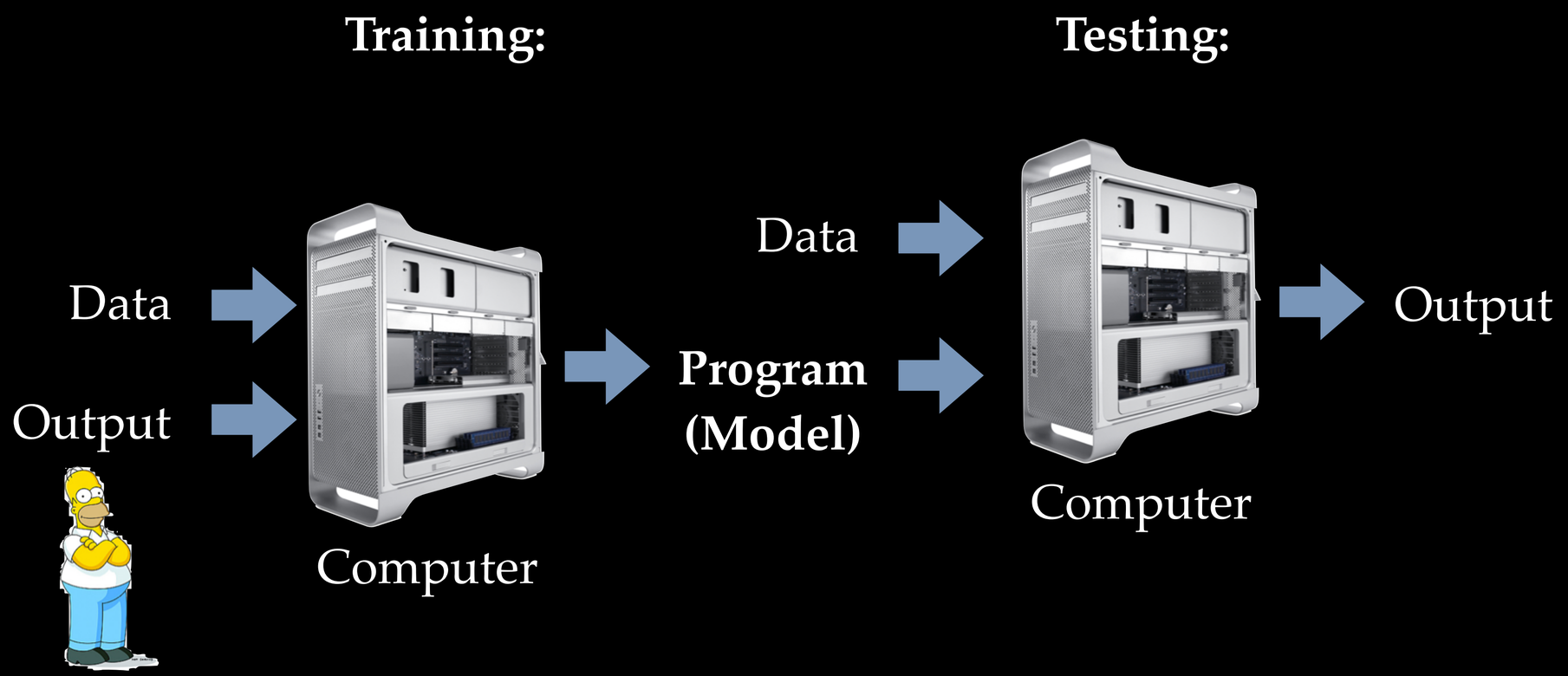
Data

Program (Model)

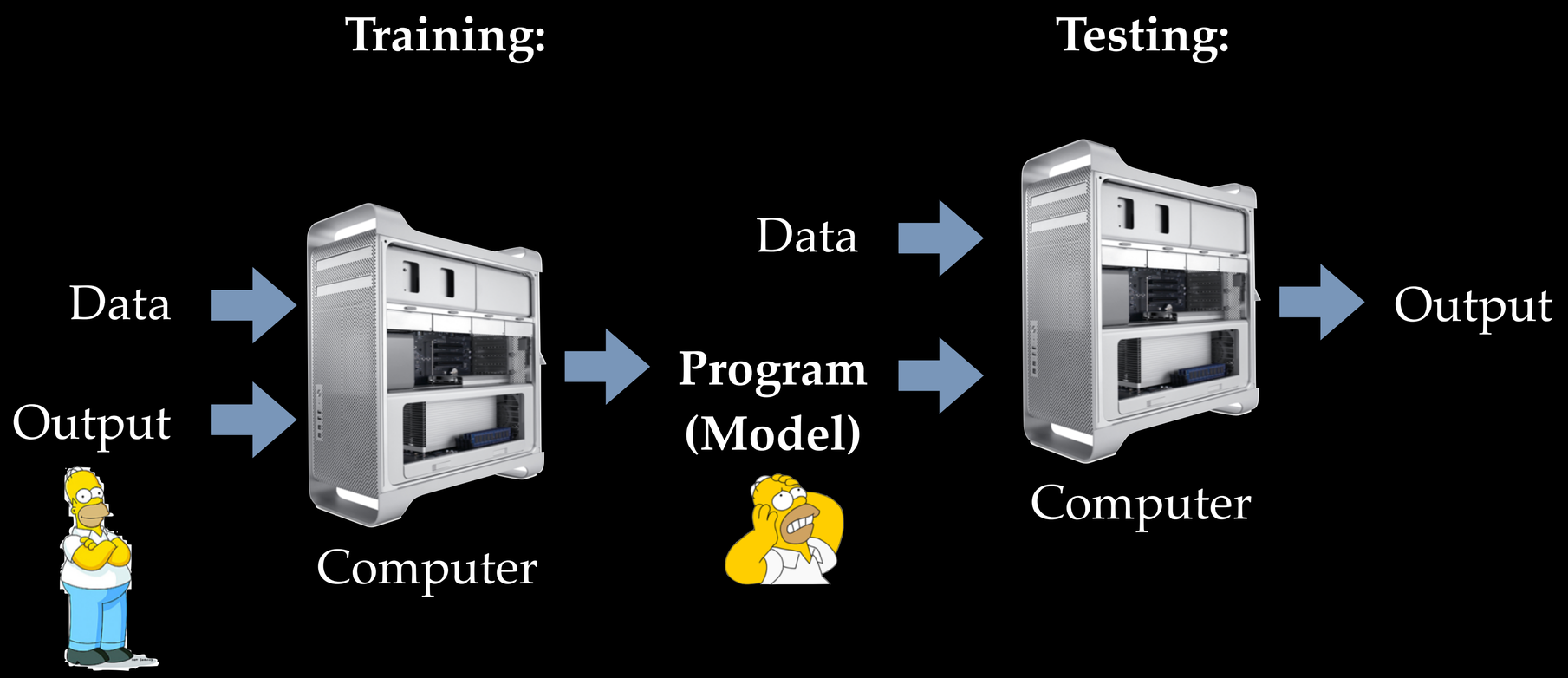
Computer

Output

Machine Learning



Machine Learning

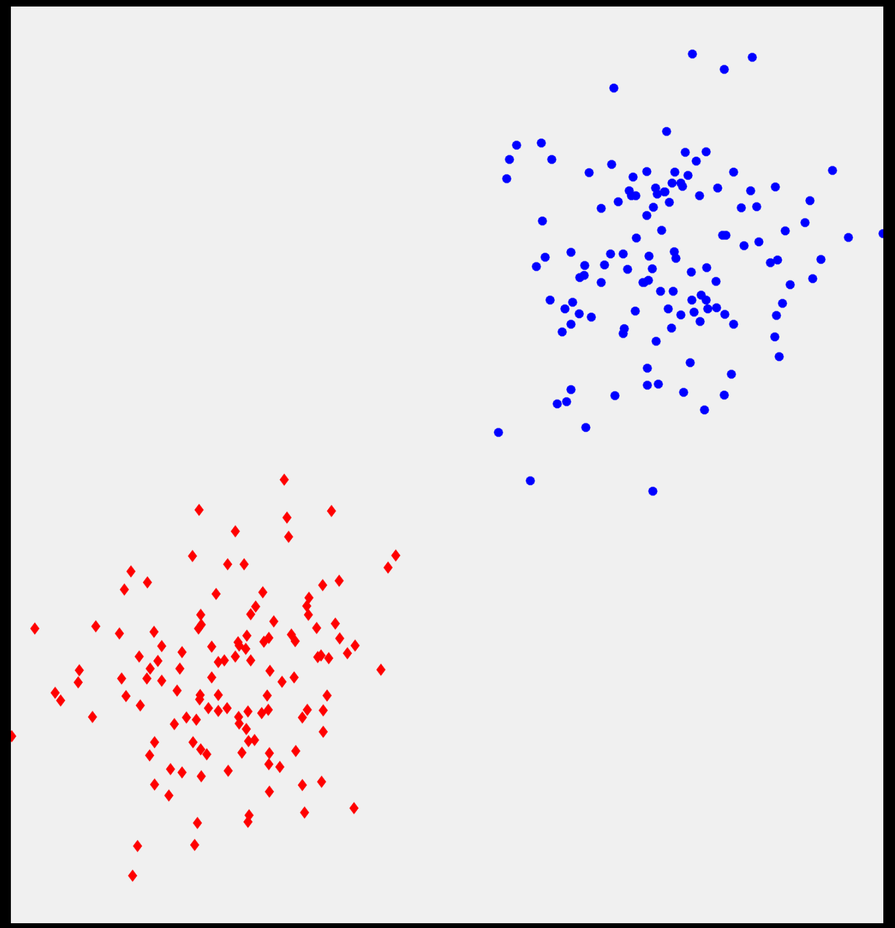


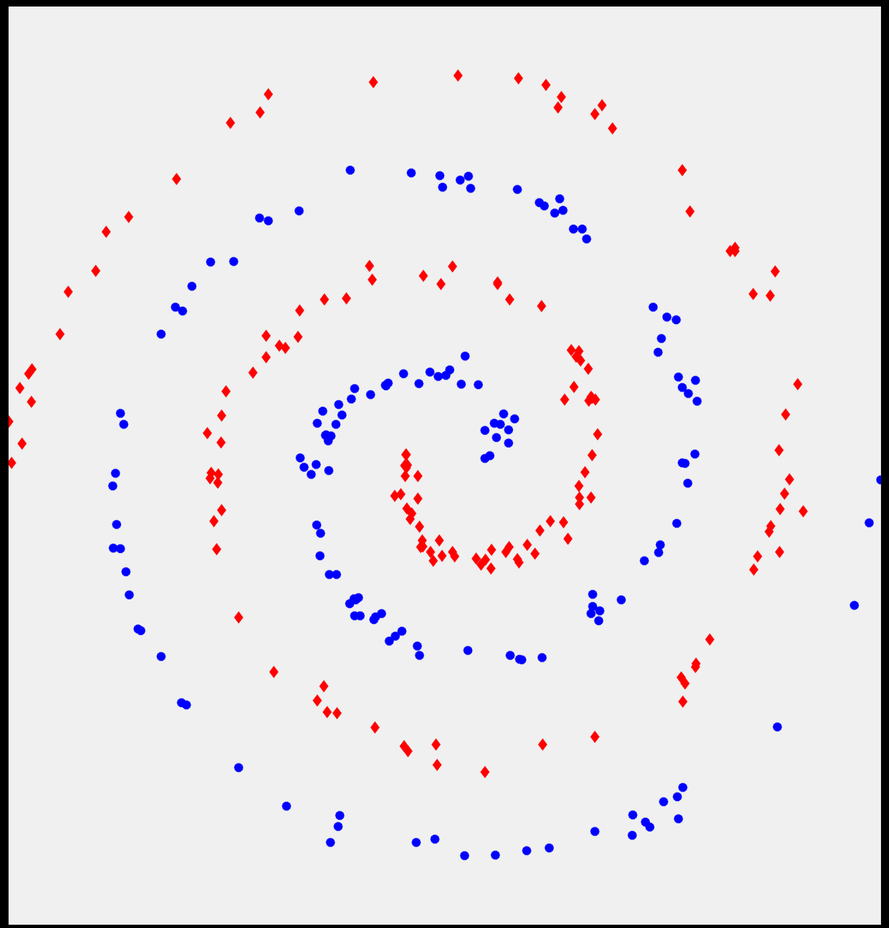
Course topics and an example

Course Topics

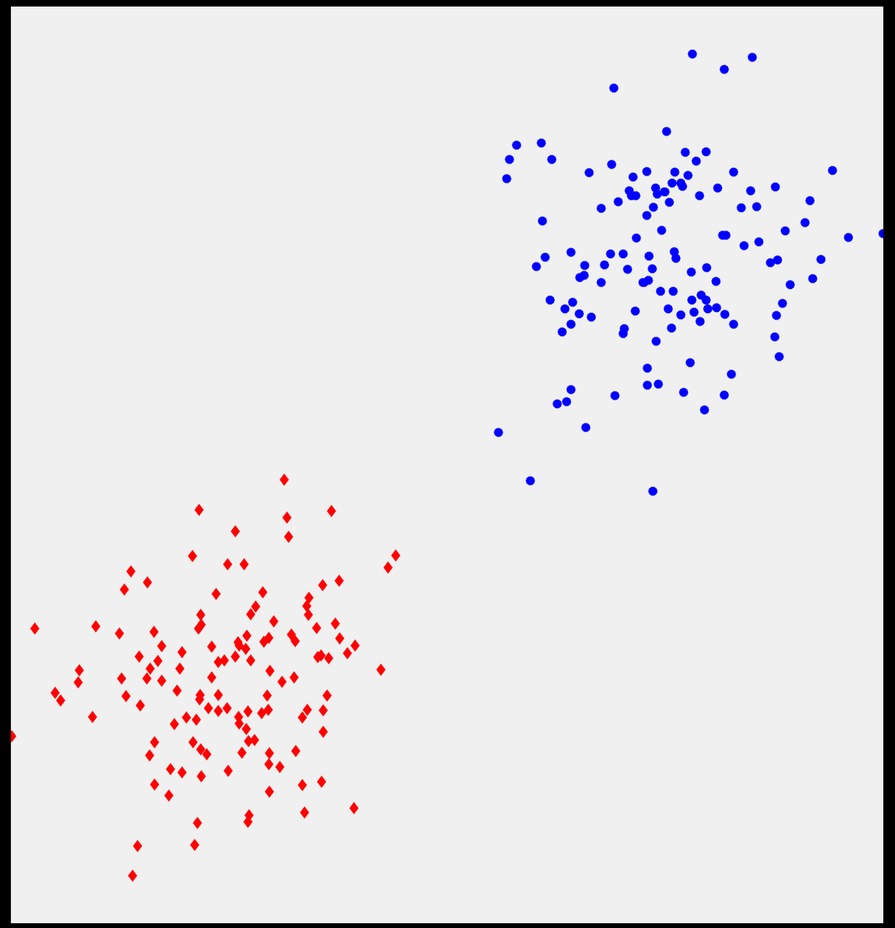
- We will cover:
  - Parametric / Non-parametric learning
  - Empirical Risk Minimization
  - Unsupervised Learning
  - Bias/Variance Trade-off
  - Boosting
  - Support Vector Machines
  - Deep Learning
  - Kernel methods
  - Classification and regression trees
- We will **not** cover:
  - Graphical Models
  - Reinforcement Learning
  - Genetic Programming
  - Gaussian processes

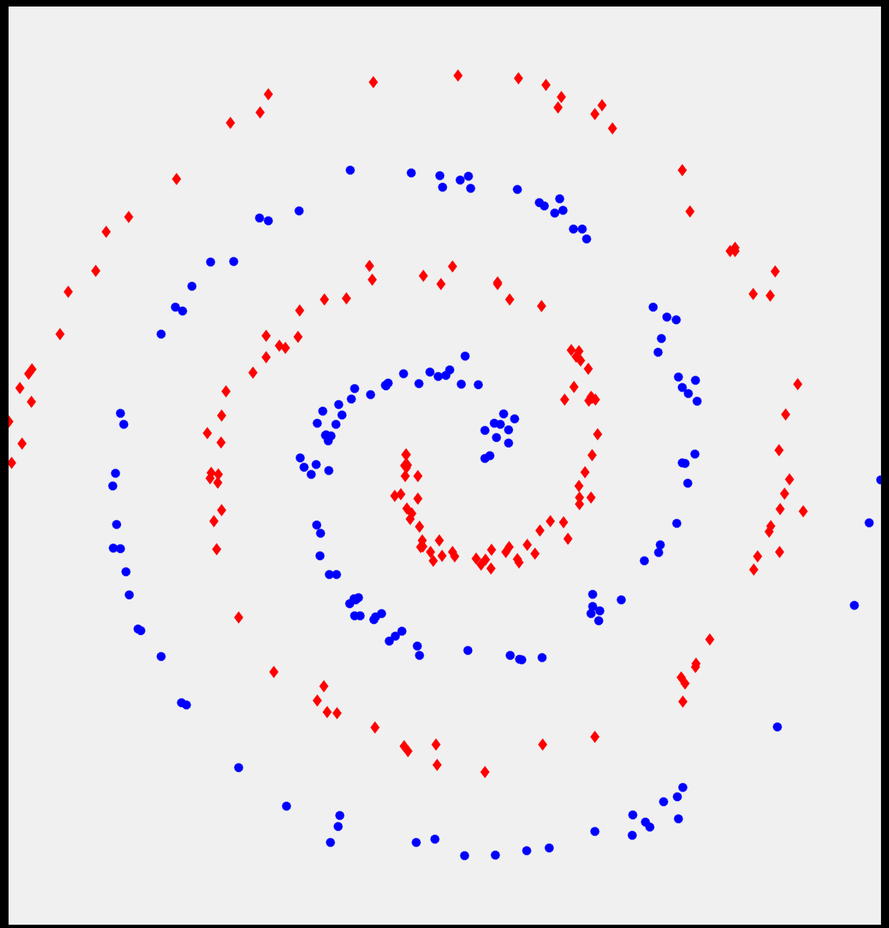
Building a complex classifier



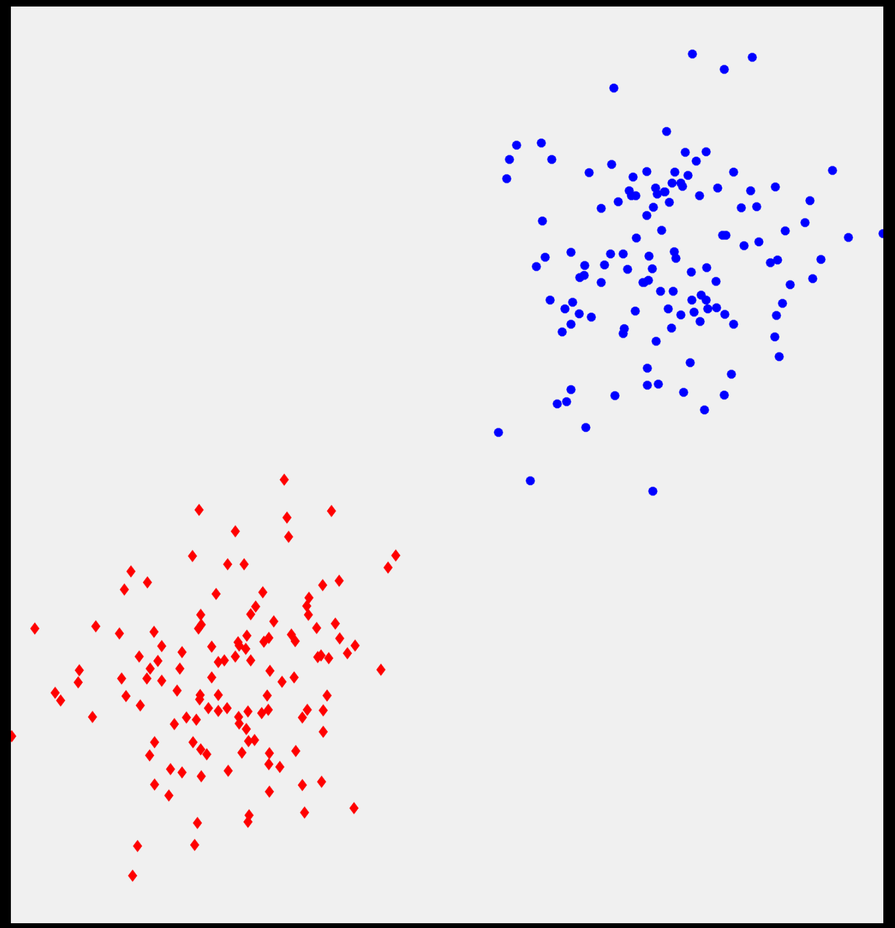


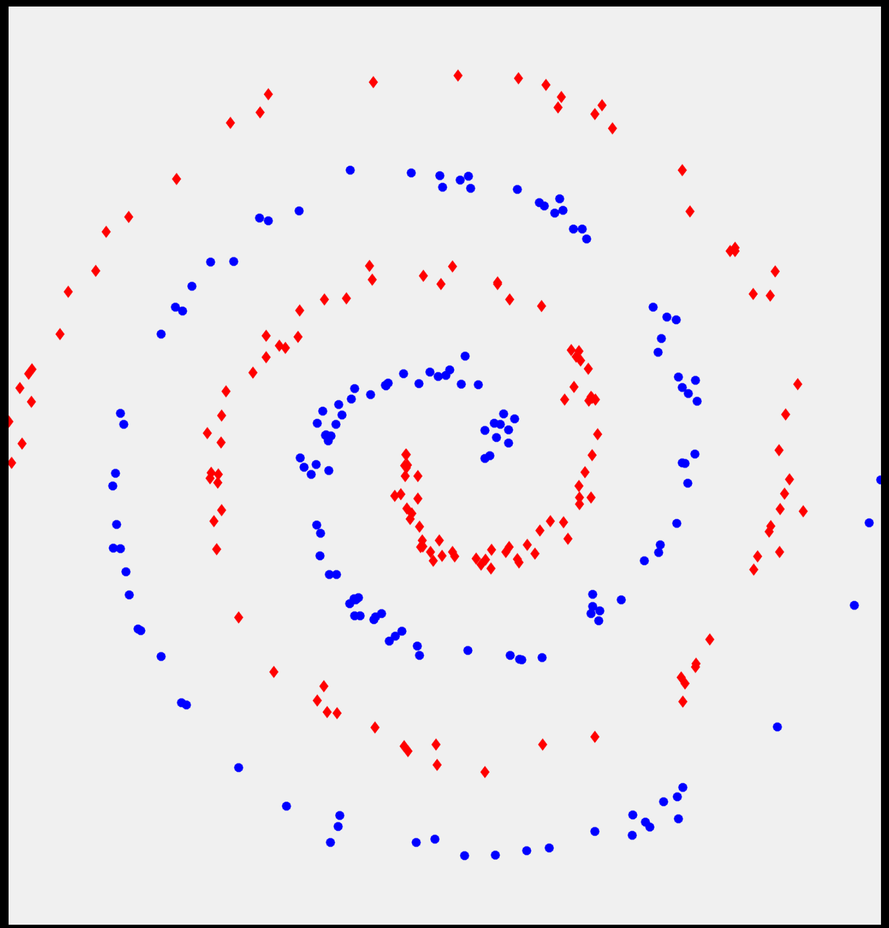
Building a complex classifier





Building a complex classifier





A (very brief) History of ML

Image source: https://xkcd.com/1838/

Alan Turing

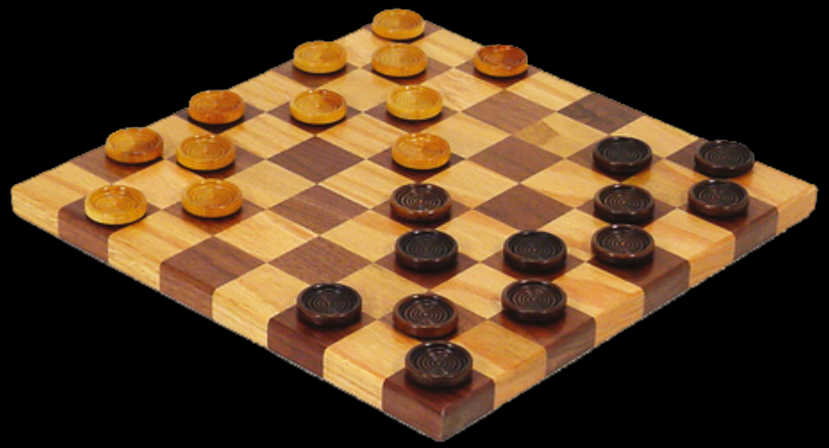
The Turing Test, 1950

A machine is intelligent if its answers are indistinguishable from a human’s

Arthur Samuel

Checkers Program, 1952

Created a Checkers-playing program that got better overtime.



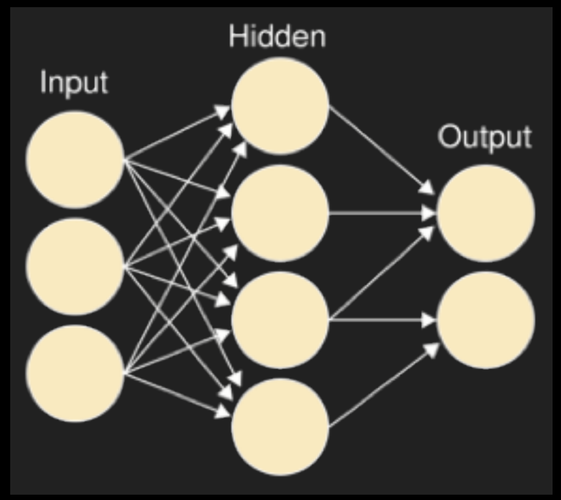
Also introduced the term “Machine Learning.”

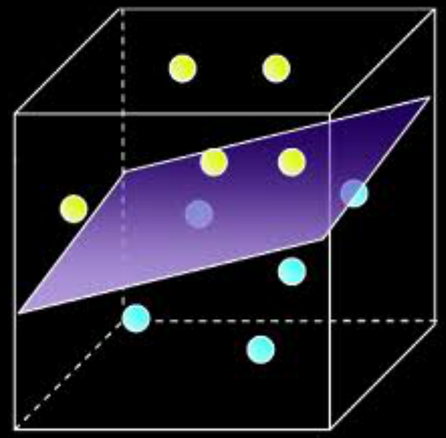
Frank Rosenblatt
@ Cornell!

Perceptron, 1957

Predecessor of deep networks.

Separating two classes of objects using a linear threshold classifier.





Frank Rosenblatt
@ Cornell!

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

The New York Times

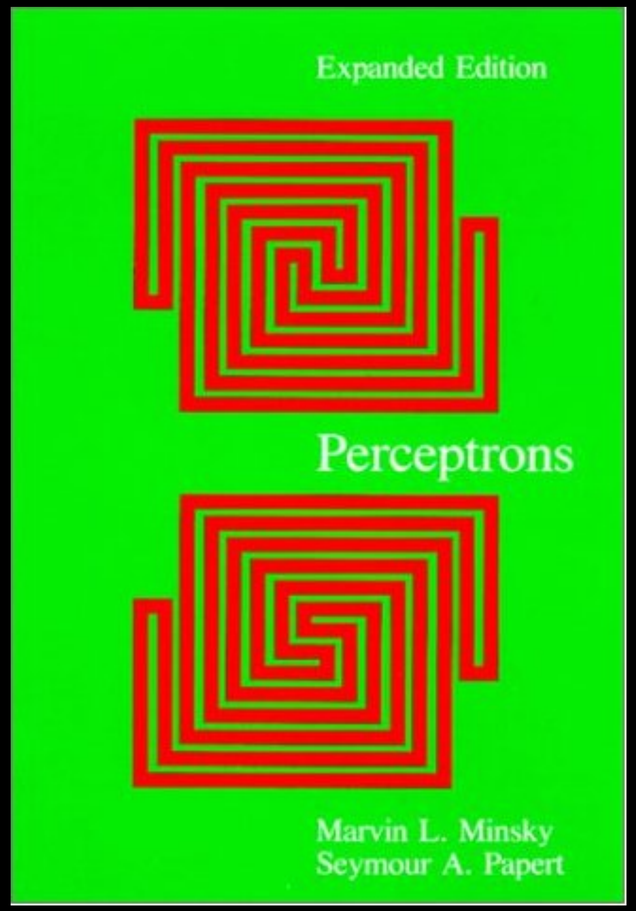
Perceptron, 1957

New Navy Device Learns by Doing
- The New York Times (July 8, 1958)

"Later perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted."

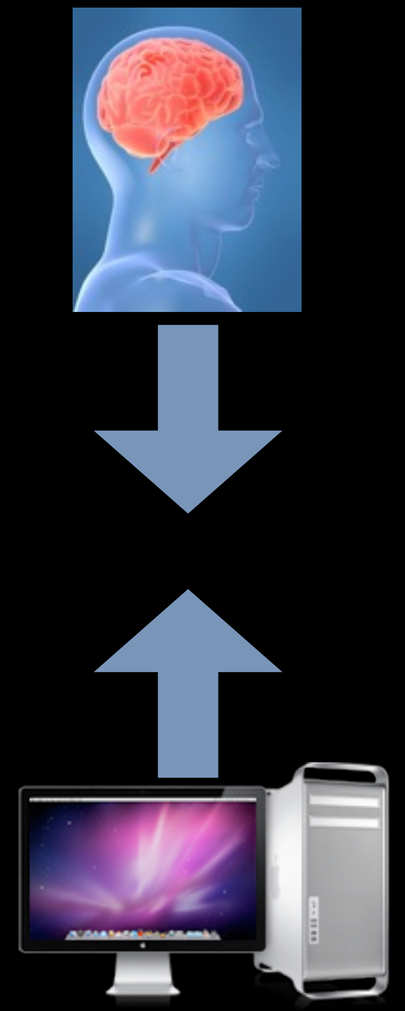
AI Winter (1974-1980)

- (1969) Minsky & Papert “killed” AI
- Burst huge expectation bubble
- Speech understanding / translation fails
- UK and US stop funding AI research



Rebirth as Machine Learning

- Machine Learning:
  - Originally: Mostly a name game to get funding.
- Profound difference:
  - ML: Bottom up, AI: Top down
  - ML: More practical smaller goals
  - Based on **Statistics and Optimization, not Logic**



TD-Gammon (1994)

- Gerry Tesauro (IBM) teaches a neural network to play Backgammon. The net plays 100K+ games **against itself** and beats world champion [Neurocomputation 1994]
- Algorithm found new techniques that people had erroneously ruled out.

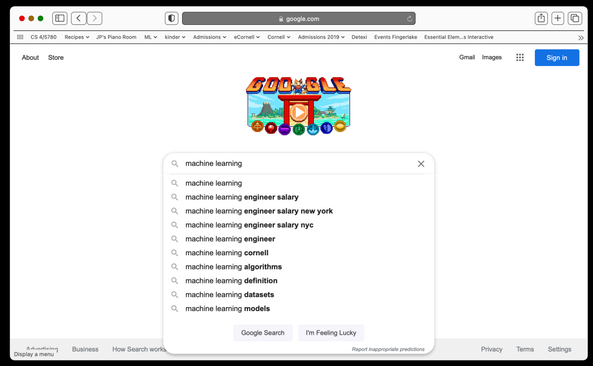
Deep Blue (1997)

- IBM’s Deep Blue wins against Kasparov in chess. Crucial winning move is made due to Machine Learning (G. Tesauro).
  - (Mostly a more classical AI system with a bit of ML)

Expanding the reach, 2000s

Learning to rank

- Powering search engines: Google, Bing, …



Topic Modeling:

- Detecting and organizing documents by subject matter
- Making sense of the unstructured data on the web

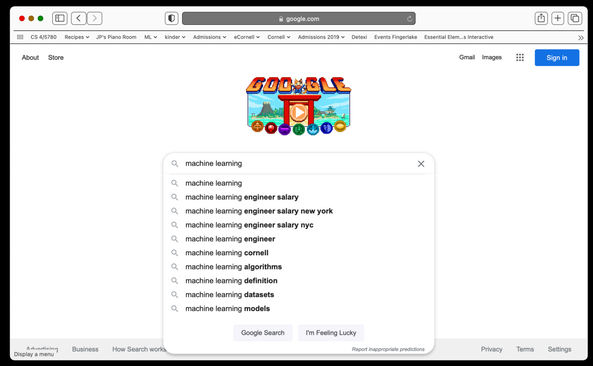
Online economy:

- Ad placement and pricing
- Product recommendation

Expanding the reach, 2000s

Learning to rank

- Powering search engines: Google, Bing, …



Topic Modeling:

- Detecting and organizing documents by subject matter
- Making sense of the unstructured data on the web

Online economy:

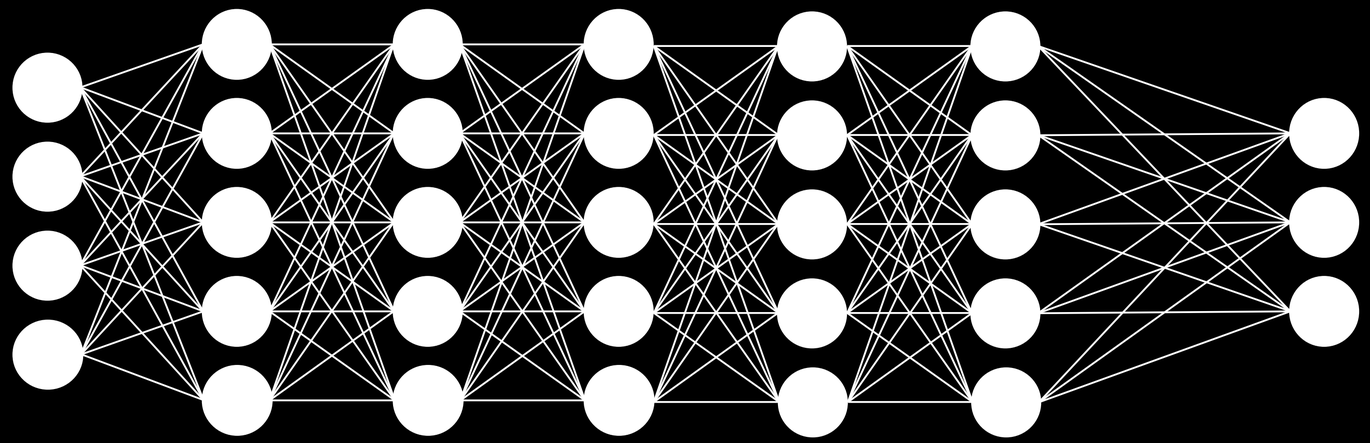
- Ad placement and pricing
- Product recommendation

Machine learning became profitable!

Return of Neural Networks, 2010s

Neural networks return and excel at image recognition, speech recognition, …

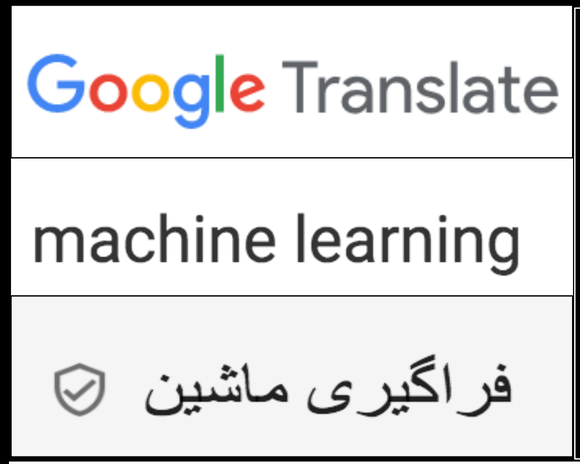
The 2018 Turing award was given to Yoshua Bengio, Geoff Hinton, and Yann LeCun.

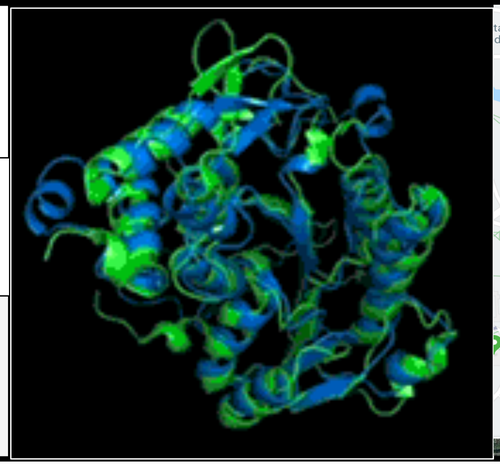


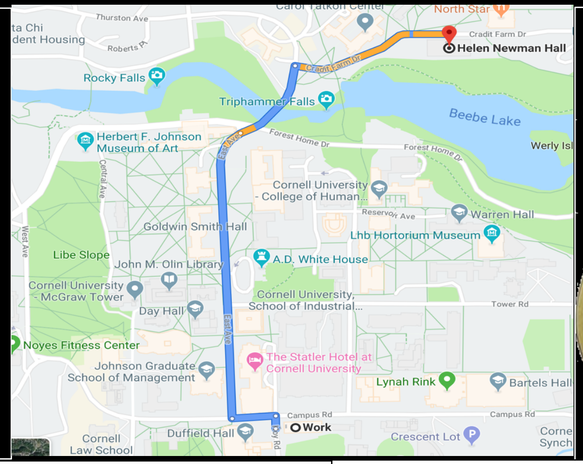
2016 Alpha Go

- 1920 CPUs and 280 GPUs
- Deep Mind’s Alpha Go wins against Lee Sedol 5:1
- Beginning of “AI arms race”

Surrounded by Machine Learning

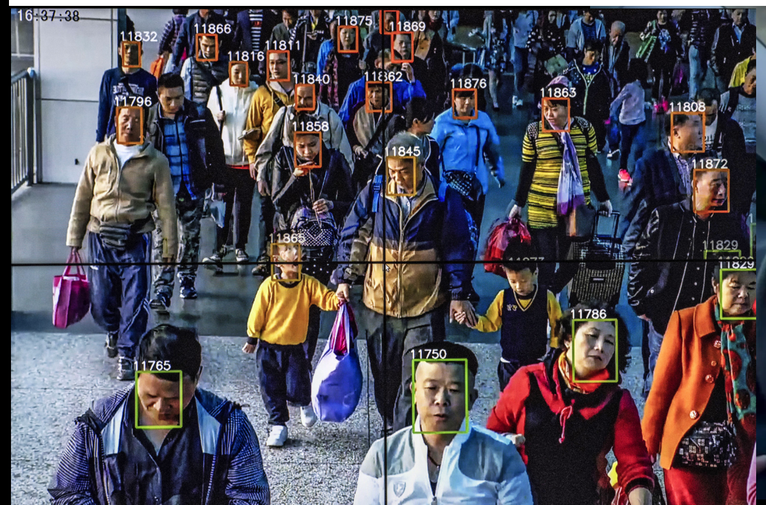






Azure Machine Learning | Create Your Free Account Today

Ad azure.microsoft.com/Services/MachineLearning



When will it stop?

- **Humans learn really well!**
  - So, we know that we can still do a **lot better!**
- However, it is hard. Very few people can design new ML algorithms.
- But many people can use them!

What types of ML are there?

As far as this course is concerned:

- **supervised learning:** Given labeled examples, find the right prediction of an unlabeled example. (e.g. *Given annotated images learn to detect faces.*)
- **unsupervised learning:** Given data try to discover similar patterns, structure, sub-spaces (e.g. *automatically cluster news articles by topic*)
- **reinforcement learning:** Try to learn from delayed feedback (e.g. *robot learns to walk, fly, play chess)*

Outlook

“A breakthrough in machine learning would be worth ten Microsofts.” (Bill Gates, Microsoft)

“It will be the basis and fundamentals of every successful huge IPO win in 5 years.” (Eric Schmidt, Google / Alphabet)

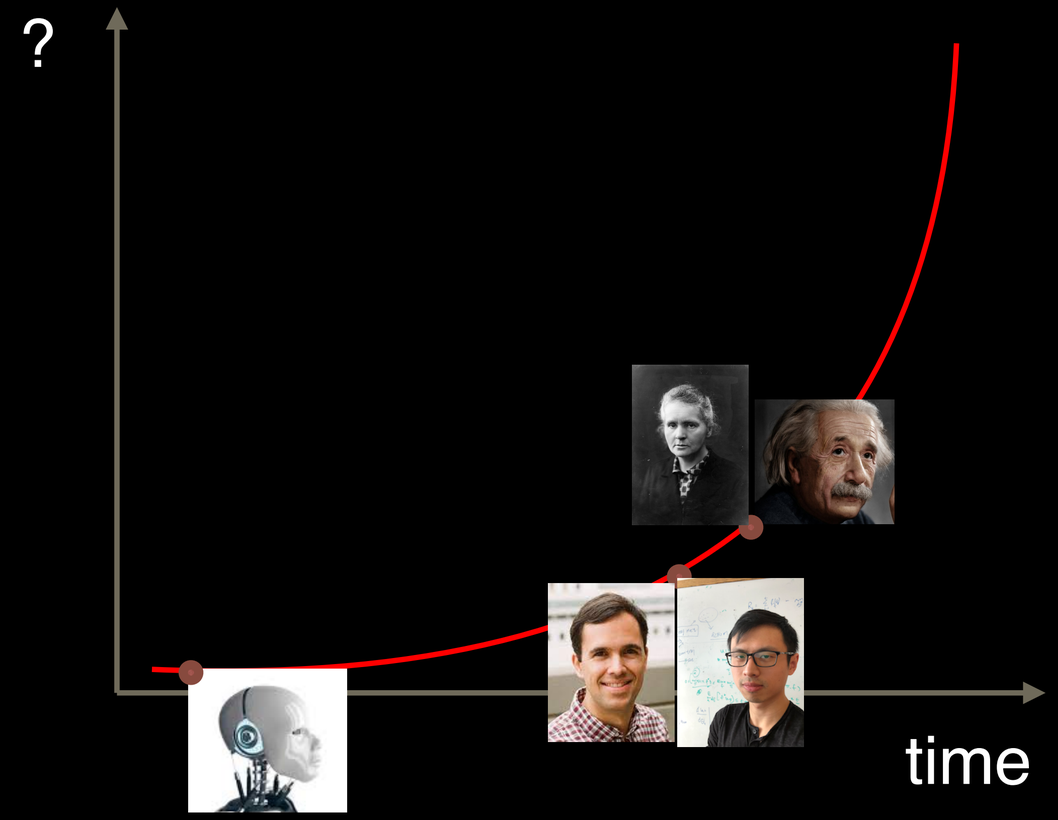
“AI and machine learning are going to change the world and we really have not begun to scratch the surface.” (Jennifer Chayes, UC Berkeley)

“ML is transforming sector after sector of the economy, and the rate of progress only seems to be accelerating.” (Daphne Koller, Stanford / Coursera/ Insitro)

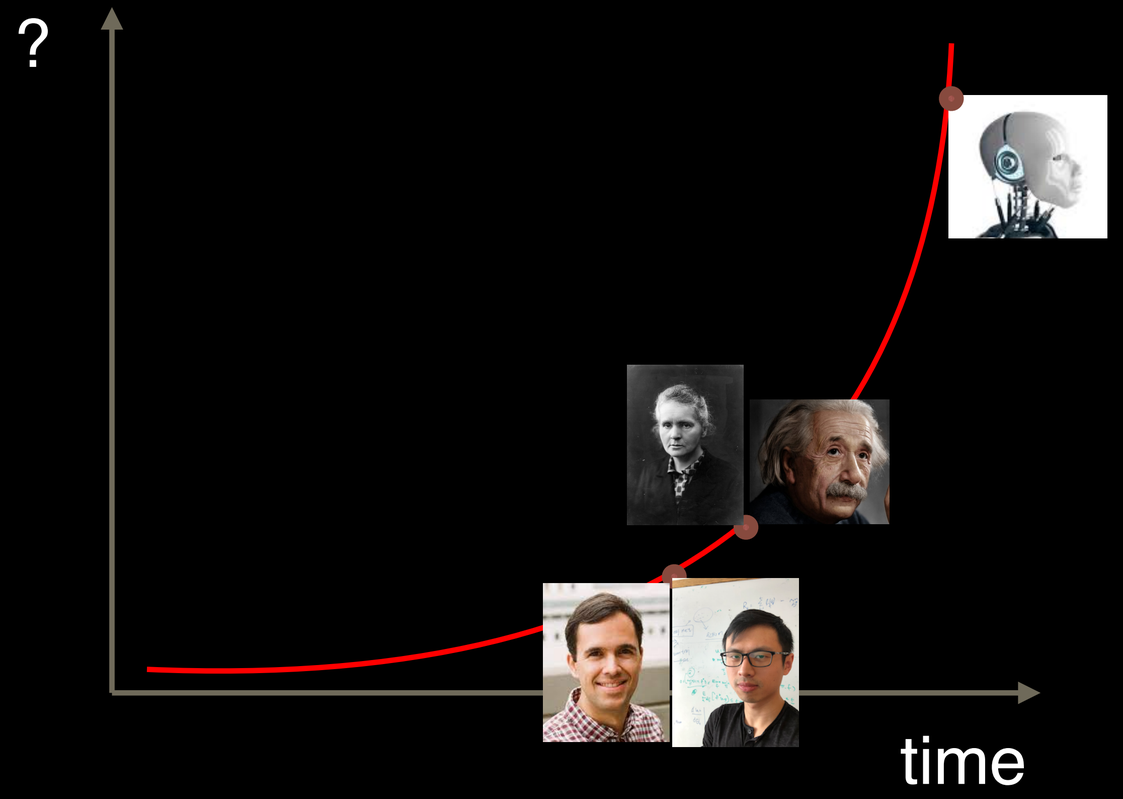
“Machine learning is the next Internet” (Tony Tether, DARPA)


DANGER

Will AI/ML take over the world?



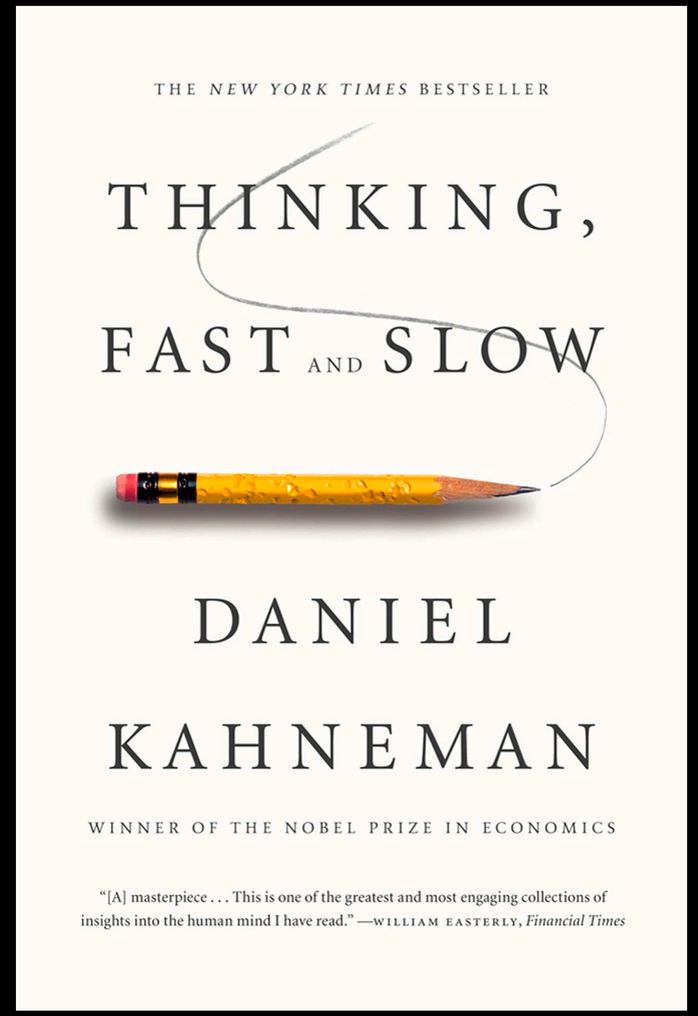
Will AI/ML take over the world?



Will AI take over the world?

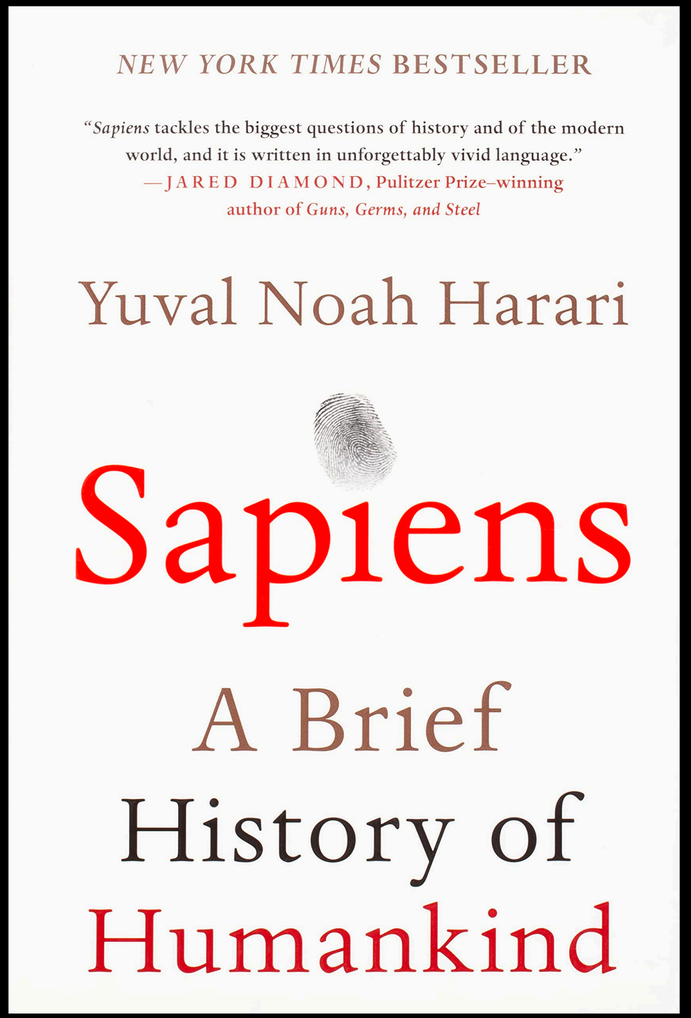
One brain, two systems

- System 1:
  - Subconscious
  - Fast
  - Requires low energy
  - Involuntary
  - Triggers surprise
- System 2:
  - Conscious
  - Slow
  - Expensive
  - Voluntary
  - Requires concentration



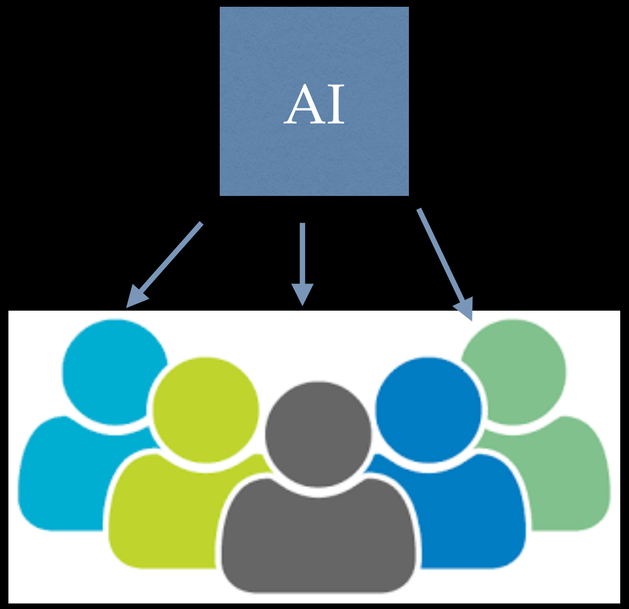
Our brain: strength and weakness

- Weakness:
  - “Unarmed human” was not on the top of the food chain
  - Our system 1 reacts instantly to rudimentary fears /dangers / drives (flight, fight, feeding, etc.)
  - Homo sapiens had no time to adapt to its new position as APEX predator



AI and Online advertising

- Advertisers pay AI companies to induce change in people's behaviors
  - AI "learns" to interact with System 1
    - (e.g. Displays sensational / alarming headlines)
  - Leads to fast clicks/ prolonged engagement (more advertising time)
  - Causes (social) anxiety, fear, undesired behavior, elevates misinformation



Robust and Secure ML

Image Recognition
Misreading traffic signs
(Eykholt et al)

Speech recognition
Hide commands in noise (Carlini & Wagner)

Poisoning Attacks
Tay (chat bot) became inflammatory in 16 hr.

How to create robust and secure machine learning algorithms?

ML and Society

- Bad dynamics, perpetuating and worsening stereotypes and biases
- Who carries the burden of bad prediction?
- How to design good dynamics?

Data privacy

Learning models leak training data

(Fredrickson et al. ‘15)

Leaked data

Real image

Data privacy

Learning models leak training data

(Fredrickson et al. ‘15)

Leaked data

Real image

Formal definitions of data privacy:

- K-anonymity (Sweeney)
- Differential Privacy (Dwork, McSherry, Nissim, Smith).

Latanya Sweeney

Cynthia Dwork

Frank McSherry

Kobbi Nissim

Adam Smith

AI/ML is different from humans

- Where and how we (choose to) deploy AI/ML systems matters
- Good at some things humans are not
- Can have extensive “unforeseen” consequences (both in development and deployment)
- Not a drop in for humans; inappropriate to use in certain settings
- Important to understand what is going on “under the hood” to understand what problems are AI/ML problems and where its use is appropriate