

CS4/5780

Introduction to

Machine Learning

Christopher De Sa
Computer Science

Lecture 1 — Spring 2022
January 24



credit: xkcd

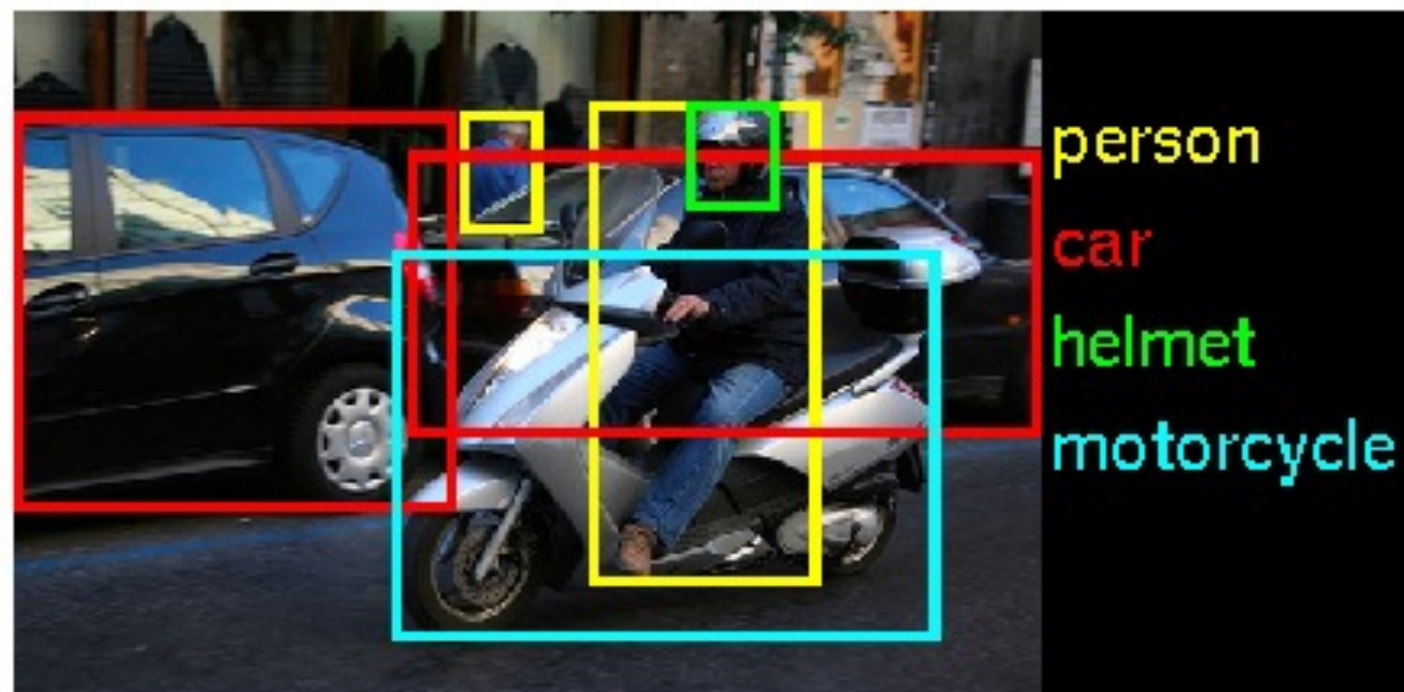
What is machine learning?

- Roughly: programs that **improve** with **experience**.
 - A subfield of **artificial intelligence**, which refers to programs that demonstrate “intelligence” in some sense
- “Learning” by analogy with how **humans also get better with experience**



Why should I care about **CS 4/5780** ?

- Machine learning is doing great things!



Object recognition



**Speech
recognition**



Robotics



Game playing

This class teaches the principles behind all this exciting technology...plus the subject is just fun intrinsically

How will CS4/5780 work?

Now for some slides on

**administrative
course details!**

This baby says:

BORING!

Course Overview — Lectures

- Remote for the first two weeks!
- PowerPoint for now
 - Because a slideshow is better for zoom lectures
- Will transition to a blackboard/document camera format when we go in-person
 - This lets us better take advantage of the space



Course Overview — Websites

- **Homepage:** <https://www.cs.cornell.edu/courses/cs4780/>
 - All the publicly-facing course stuff is here
- **Canvas:** <https://canvas.cornell.edu>
 - For things that need to be limited to students enrolled in the course
- **Vocareum:** <http://vocareum.com>
 - A service we use for our programming assignments
- **Ed:** <https://edstem.org/us/courses/19541/>
 - All discussion happens here



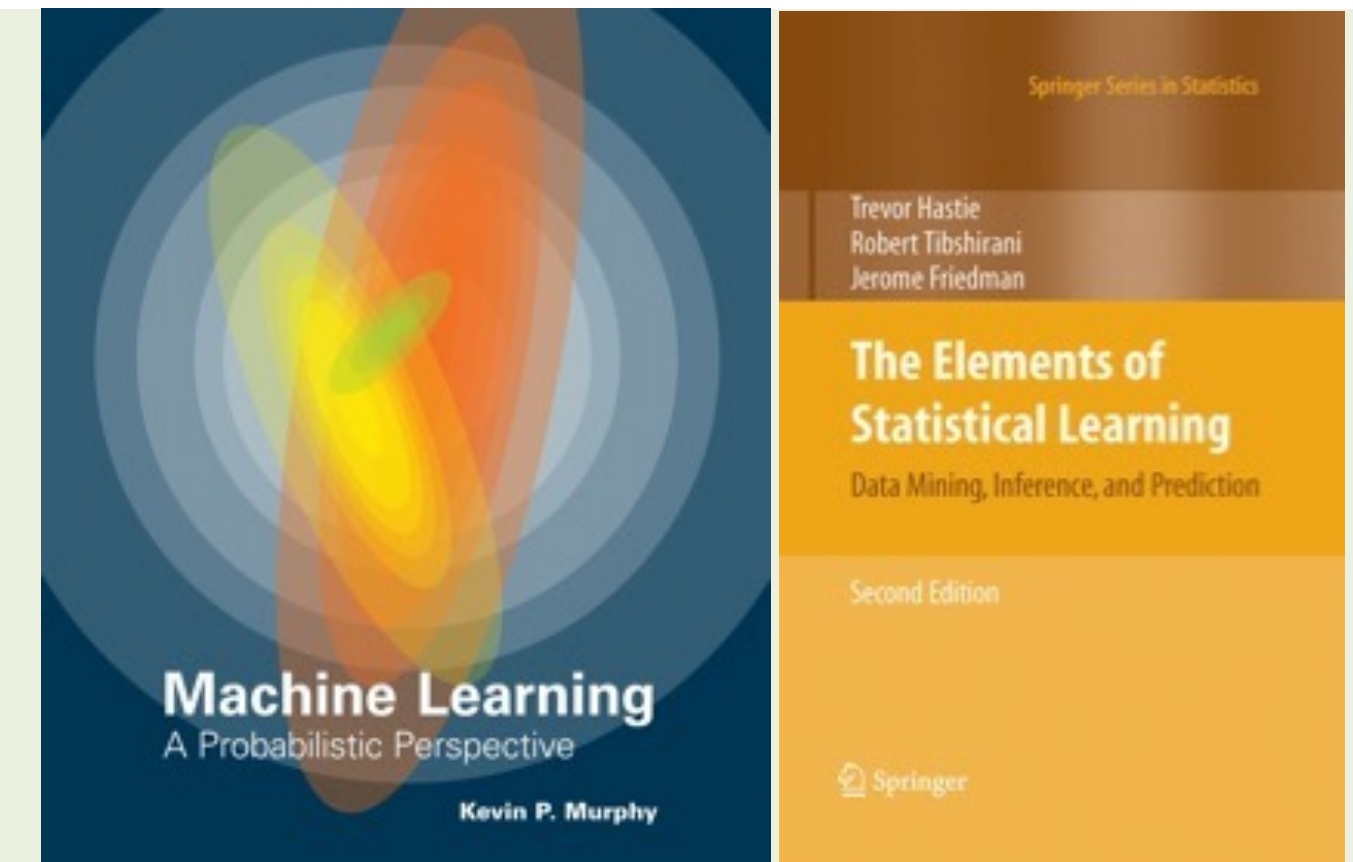
Course Overview — Where can I get help?

- **TA Office Hours:** Every day
 - All zoom for now, hybrid after first in-person lecture (Feb 7).
 - Schedule will be on calendar on course webpage.
 - You can leave feedback via a link on the course webpage.
- **My Office Hours:** 2–3PM on Wednesdays
 - Also zoom for now, in future in Gates 450
- Post **all questions** on ED (you can make them private)
 - Do not email directly (except in an emergency)

Course Texts/Resources — What can I read?

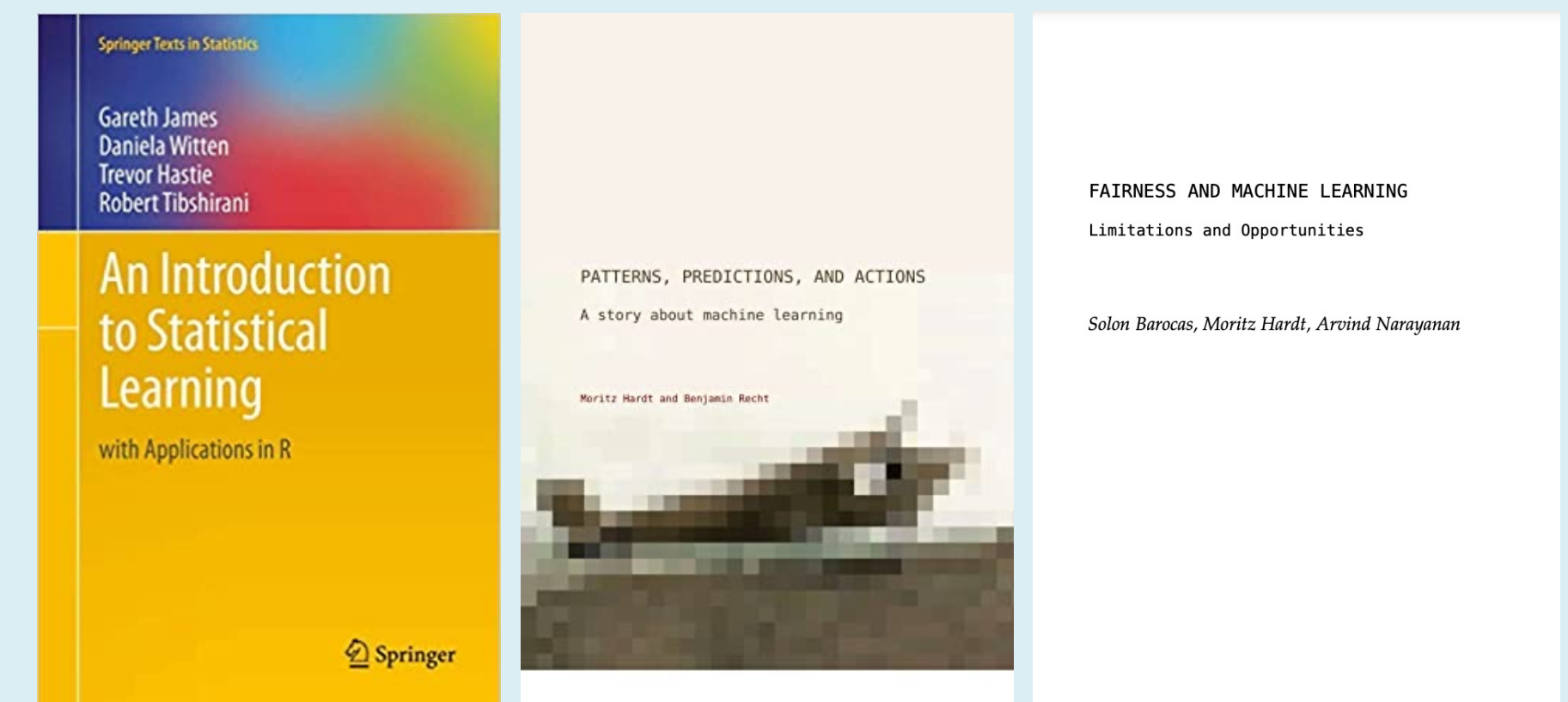
Main Course Texts

- Machine Learning a Probabilistic Perspective
 - K.P. Murphy
- The Elements of Statistical Learning
 - Hastie, Tibshirani, Friedman



Additional Texts

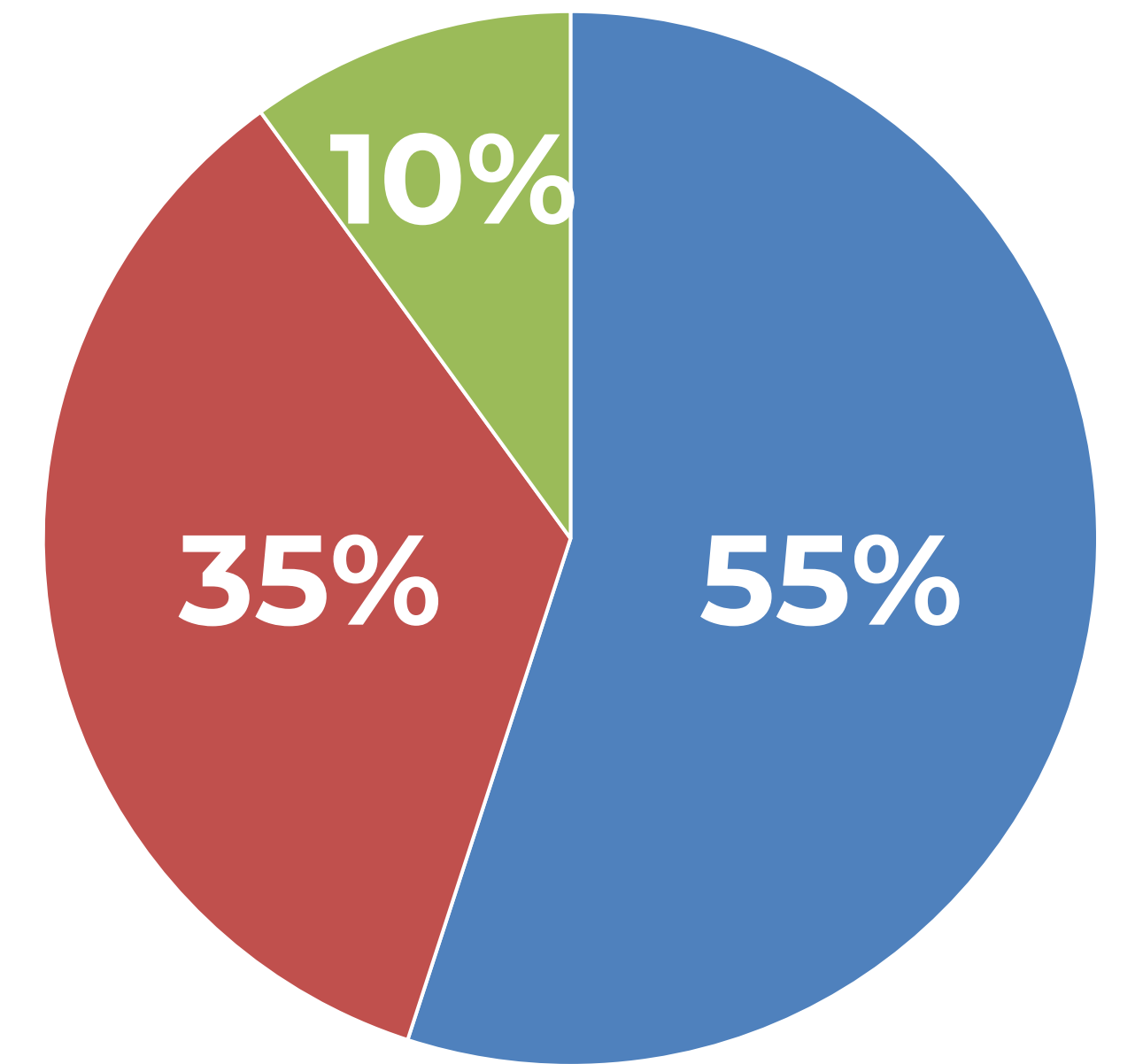
- An Introduction to Statistical Learning
 - James, Witten, Hastie, and Tibshirani
- Patterns, Predictions, and Actions
 - Hardt and Recht
- Fairness and Machine Learning
 - Barocas, Hardt, and Narayanan



Also: Background and programming resources on the website.

Course Breakdown 5780

- (55% or 45%) Theory: Midterm + Final
 - Closed book, No personal notes, no cheat sheets!
- (0% or 10%) Theory: Homeworks
 - Up to 5 members in each team
 - Preparation for exam
 - **Must submit a faithful attempt**
- 35% Programming Assignments
 - Up to 2 members in each team
 - 2 days extension per team per project
 - Autograder (unlimited resubmissions)
 - Extra credit if you win contests
- 10% Paper Comprehension (mandatory)
 - Original Research Papers in ML
 - Canvas Quizzes

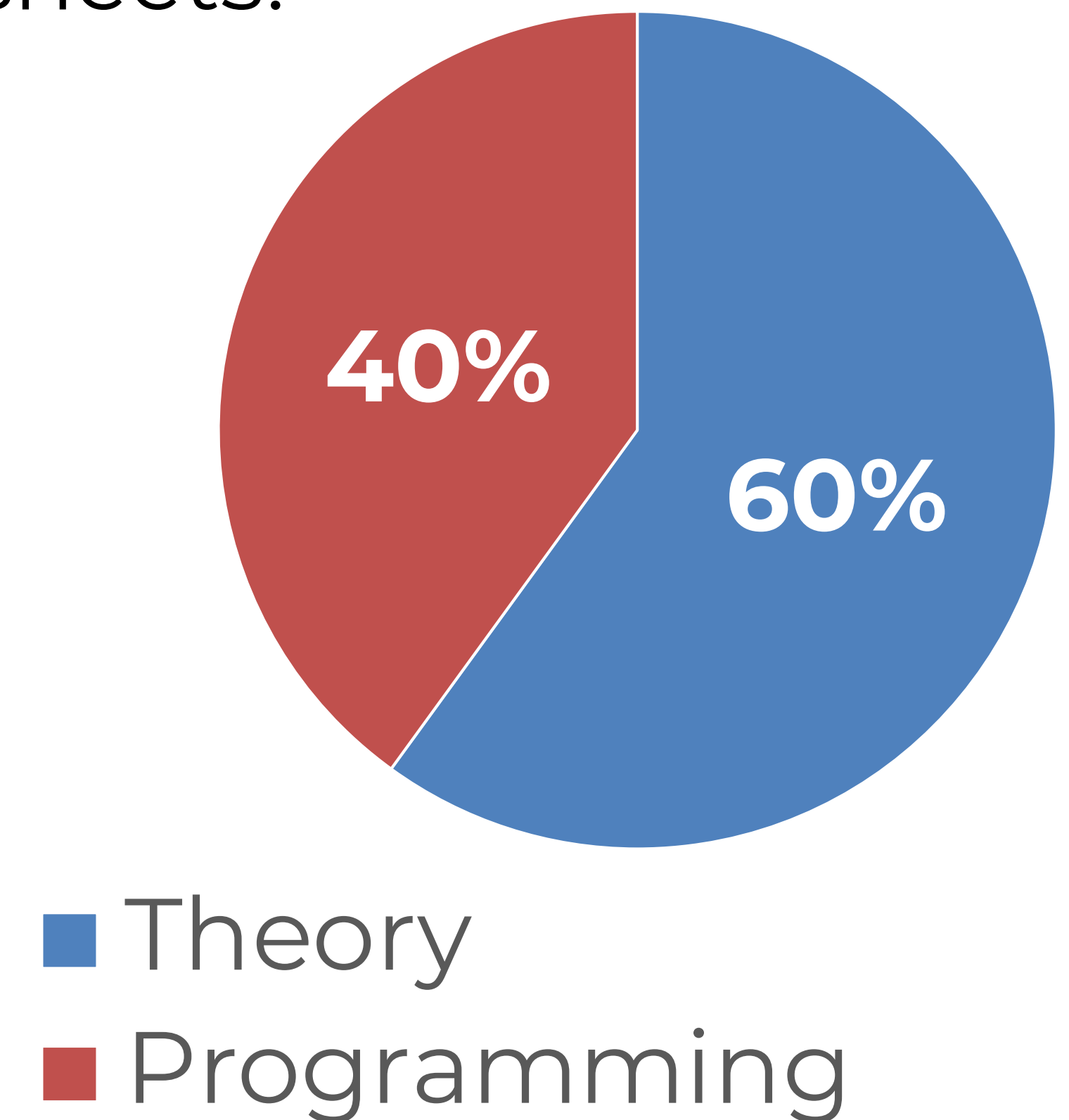


- Theory
- Programming
- Paper Comprehension

Course Breakdown 4780

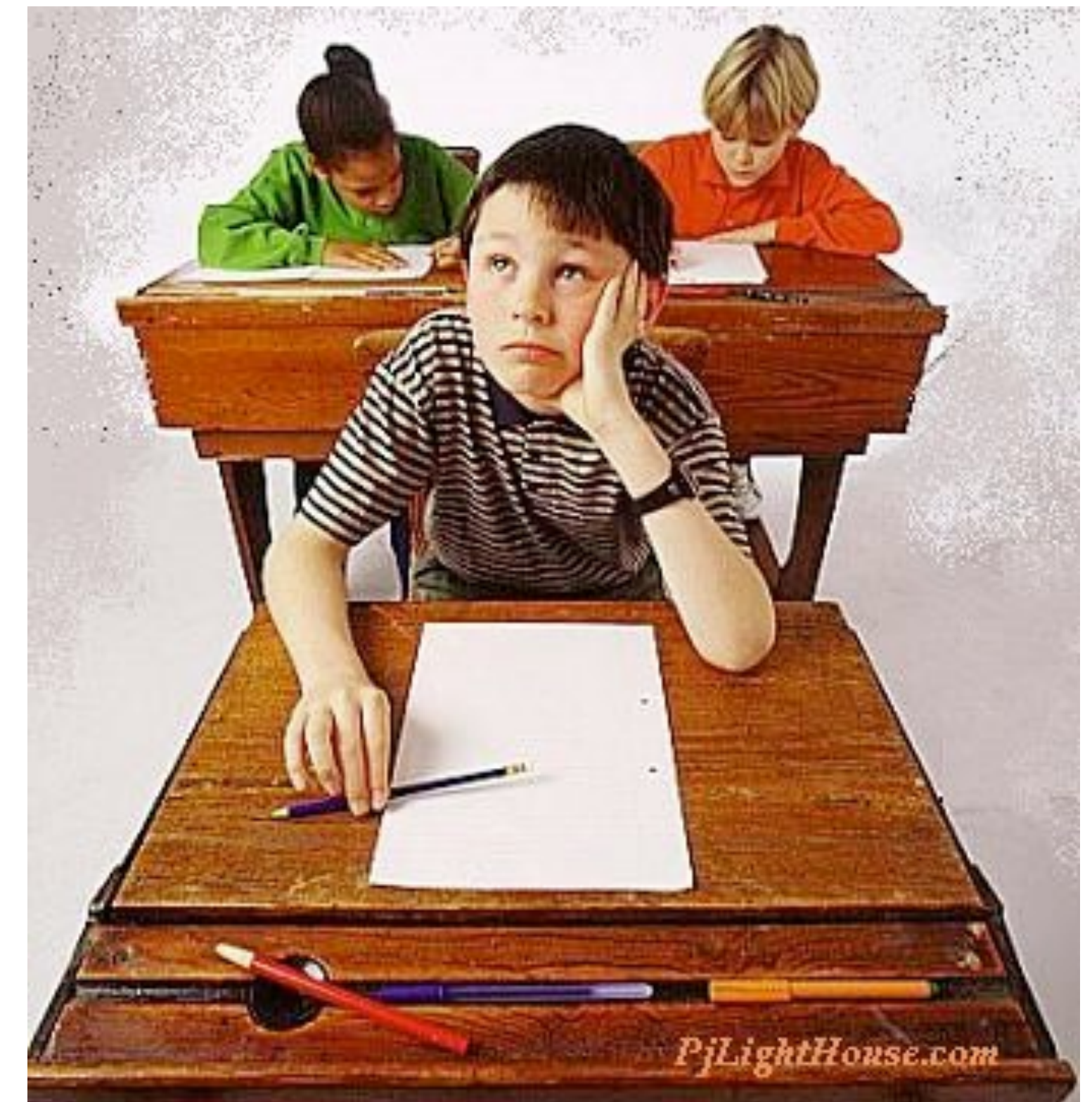
...or you can choose the 5780 grading scheme.

- (60% or 50%) Theory: Midterm + Final
 - Closed book, No personal notes, no cheat sheets!
- (0% or 10%) Theory: Homeworks
 - Up to 5 members in each team
 - Preparation for exam
 - **Must submit a faithful attempt**
- 40% Programming Assignments
 - Up to 2 members in each team
 - 2 days extension per team per project
 - Autograder (unlimited resubmissions)
 - Extra credit if you win contests



Placement Exam

- Due Wednesday, February 2nd.
- Taken on Canvas under “Quizzes.”
- Get started early!
- It is there for your own protection!
- This is how you get a Vocareum invite!
- Take it even if you are on the waitlist!



Study Groups (2–5 people)

- You **must** join a study group by the end of next week.
- Find people on **Ed**
- This course will cover non-trivial material, learning in a group makes it easier and more fun!



<http://vocareum.com>

- Pass placement exam → get account
- There will be 8(+2) projects
 - You have (roughly) 2 weeks for each project
 - Unlimited submits until deadline
- Costs \$30 :-)

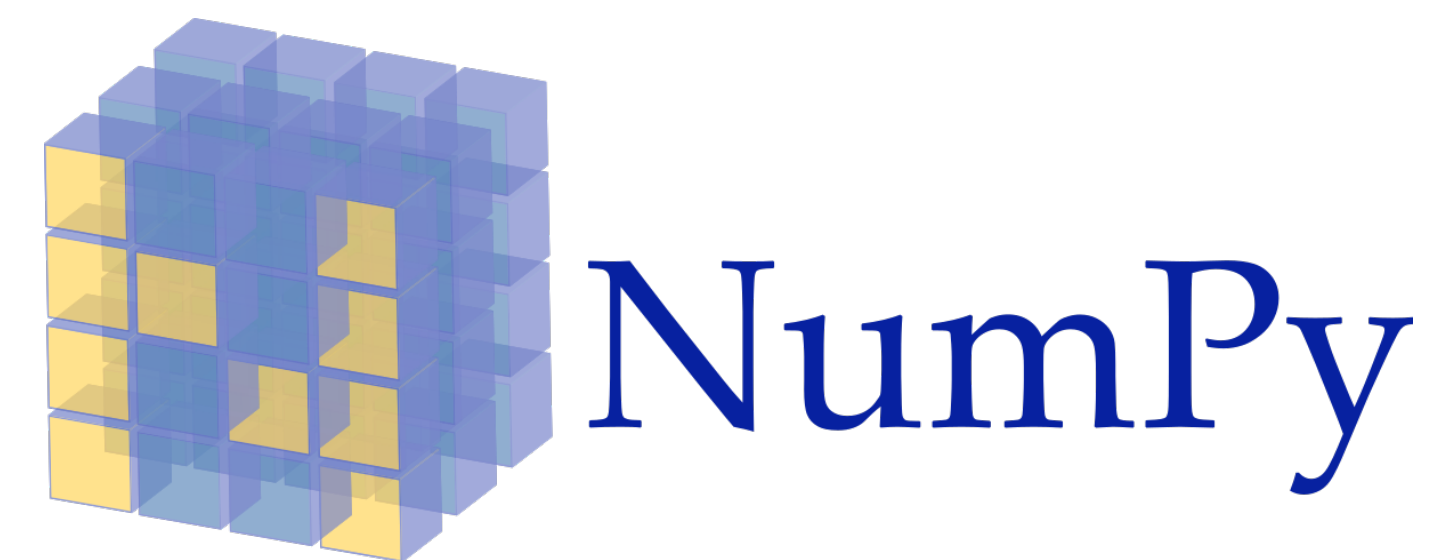
The screenshot shows the Vocareum website interface. At the top, there's a navigation bar with 'Vocareum', 'My Classes', 'Help', and 'Quirin'. Below this, the course 'CS4780 / CS5780 Spring 2017' is selected. A 'MyGrade' button is present. The main content area shows 'Slip days used: 0 of 2'. Under the 'Project 0: Getting Started' section, there's an 'Assignment' tab. The assignment text reads: 'Project 0: Getting Started. This is a first assignment to familiarize yourself with our infrastructure. Further instructions on completing the assignment can be found in the Jupyter notebooks by clicking on "My Work" below. You may complete the assignment in either Julia or Python by filling out that language's respective Jupyter notebook. You are NOT required to complete the assignment in both languages.' Below this is a 'Submission' section stating 'You have submitted the assignment.' and a 'My Work' button. A 'Details' table shows submission information:

Details	
Last submitted:	Jan-18-2017 10:38:24 am EST
Submission count:	4
Due date:	None

A 'Leaderboard' button is located at the bottom right of the submission details.

Placement Exam II (Project -1)

- NUMPY proficiency test
- Will turn into your own cheat-sheet
- Please take it seriously — this is for your own good
- We will go through the solutions in dedicated sections



Autograder

- Pick a (secret) name for your team for the leaderboard
 - (top right corner – click on your login)
- **Only text between #<GRADED> and #</GRADED> will be graded**
- Be sure to **form teams before you get started!**
 - Otherwise it creates headaches for the Vocareum system

Course Topics

We will cover

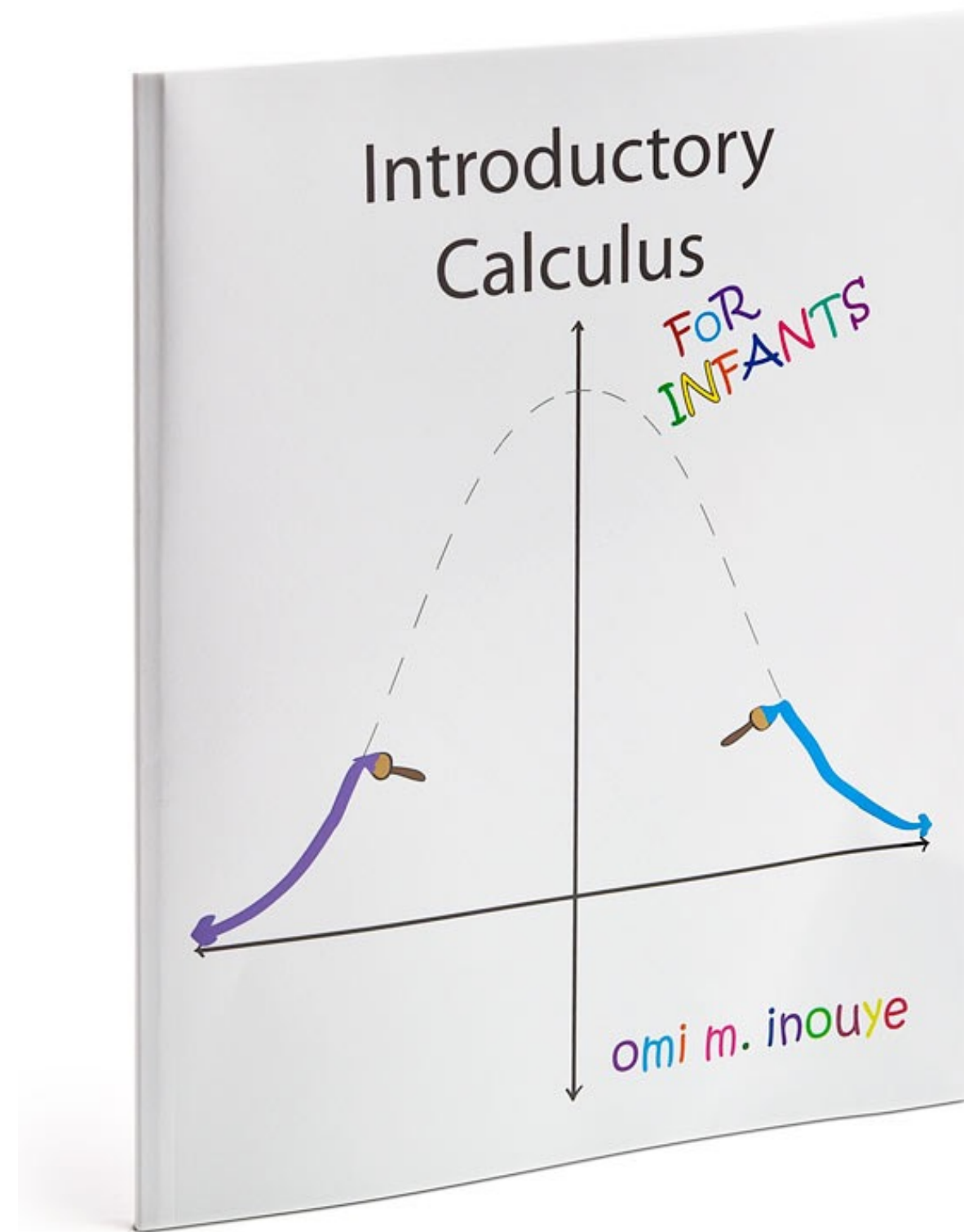
- Parametric learning
- Non-parametric learning
- Empirical risk minimization
- Unsupervised learning
- Bias/Variance Trade-off
- Boosting
- Support Vector Machines
- Deep Learning

We will not cover

- Graphical Models
- Reinforcement Learning
- Genetic Programming

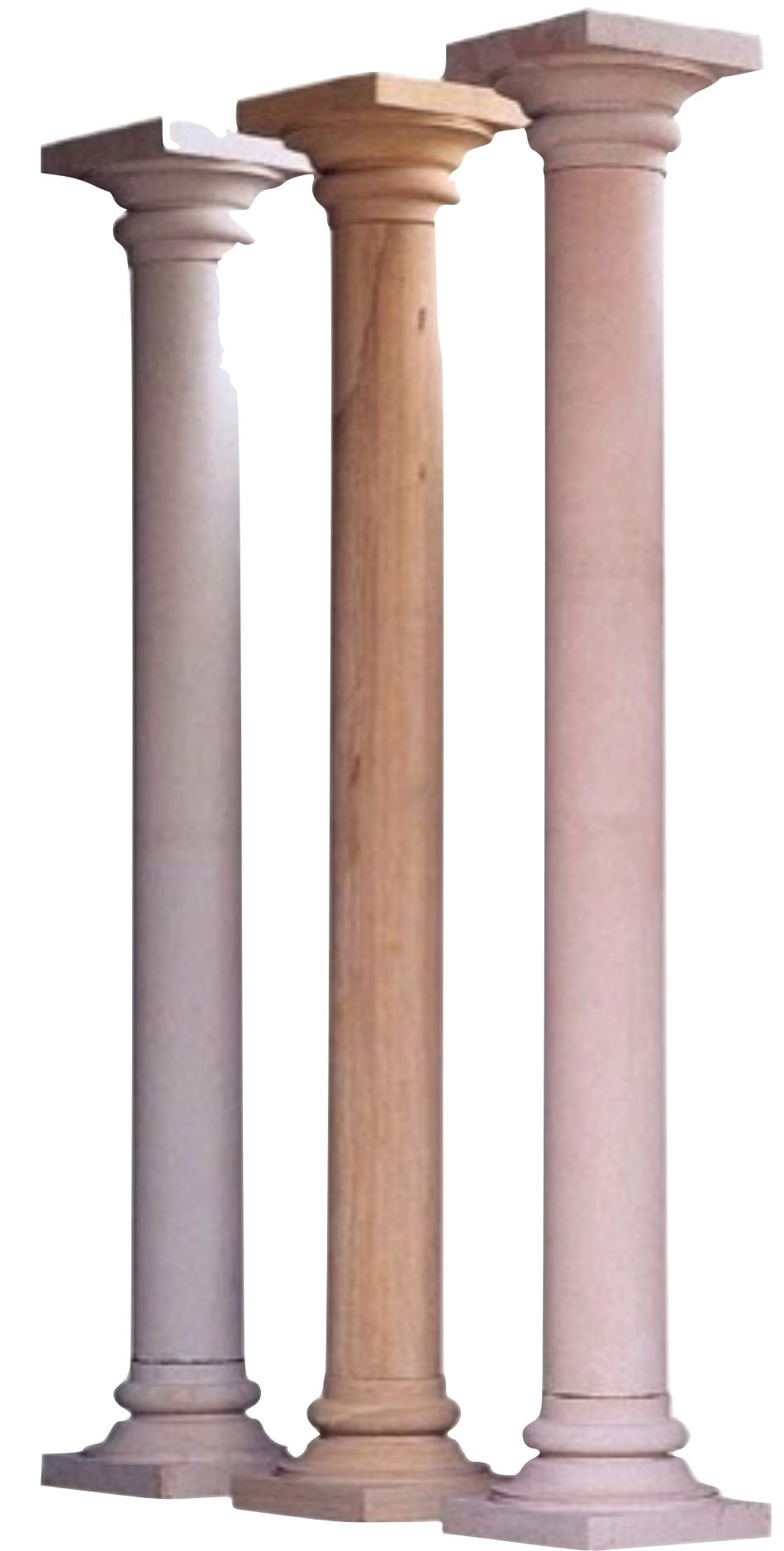


Prerequisites



Prerequisites

- Three pillars of ML:
 - **Statistics / Probability**
 - **Linear Algebra**
 - **Multivariate Calculus**
- Should be familiar with all, confident in at least $1/3$, ideally $2/3$.
- TAs might be able to give recitations on some topics if needed.
 - But don't rely on it.



About this course

- **Take this course if ...**

- you are interested in Machine Learning
- you are comfortable with a decent amount of mathematics
- you are proficient at programming

- **Don't take this course if ...**

- matrices scare you
- you don't remember how to take derivatives
- you want an easy A

- You cannot take this course if you fail the placement exam.
 - In that case, take appropriate prerequisites and come back next year.

A selection of student comments from past years

“It's mostly a math class”

“The topics were pretty complicated and difficult to understand quickly. I would have preferred a slightly slower pace.”

“...Requires a good knowledge in math and derivatives.”

“Huge work load, excessive at times, but that's just the nature of the course”

“great course, but prepare to work your butt off.”

“A TON of work, but mostly worth it for a very valuable skill.”

Academic Integrity Policy

- **Zero tolerance policy**
 - all occurrences will be reported
 - this will result in a **fair academic integrity hearing**
- We **actively look for academic conduct violations**
 - The autograder checks for plagiarism
- *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



Questions?

Let's get back to the course content!

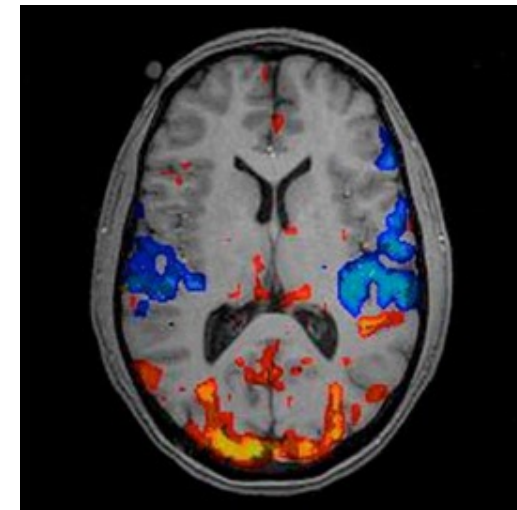
What is machine learning?

- Roughly: programs that **improve** with **experience**.
 - A subfield of **artificial intelligence**, which refers to programs that demonstrate “intelligence” in some sense
- “Learning” by analogy with how **humans also get better with experience**

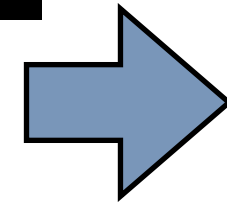


Traditional Computer Science

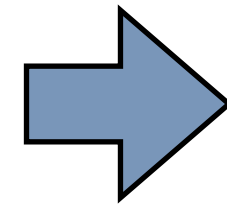
Traditional CS:



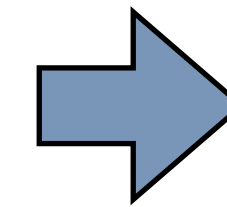
Data



Program



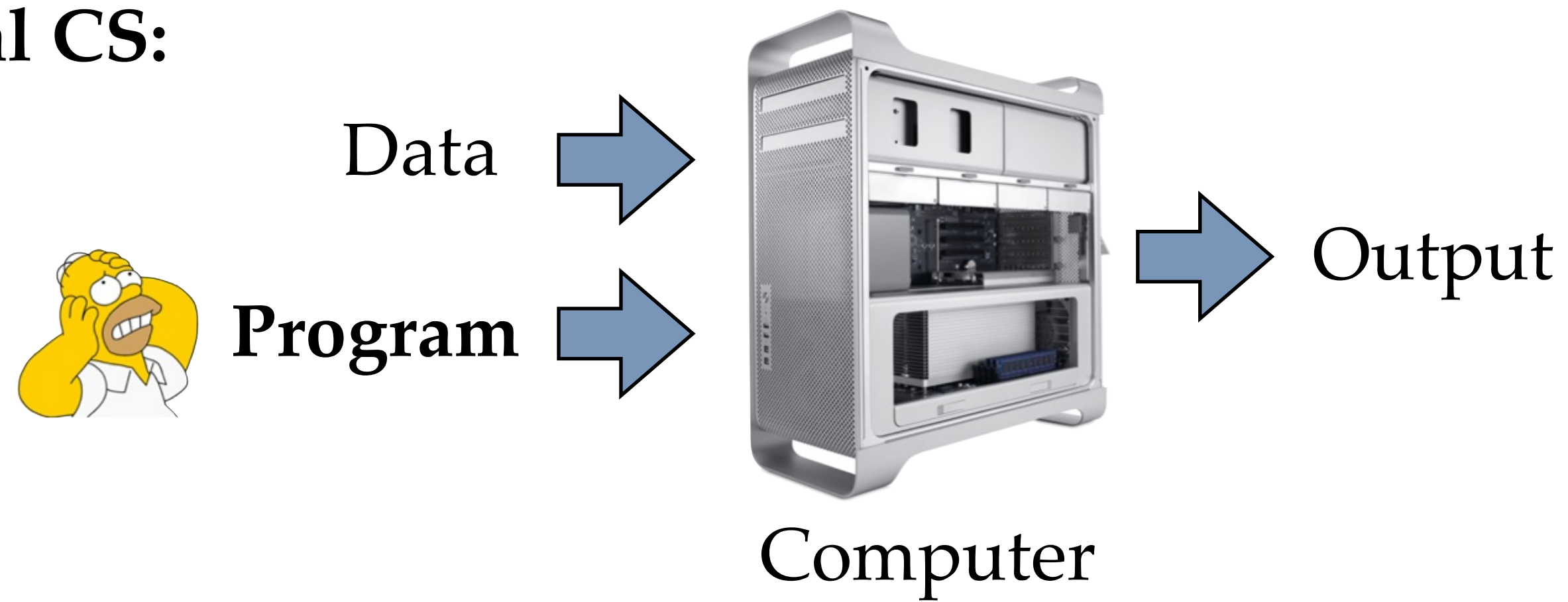
Computer



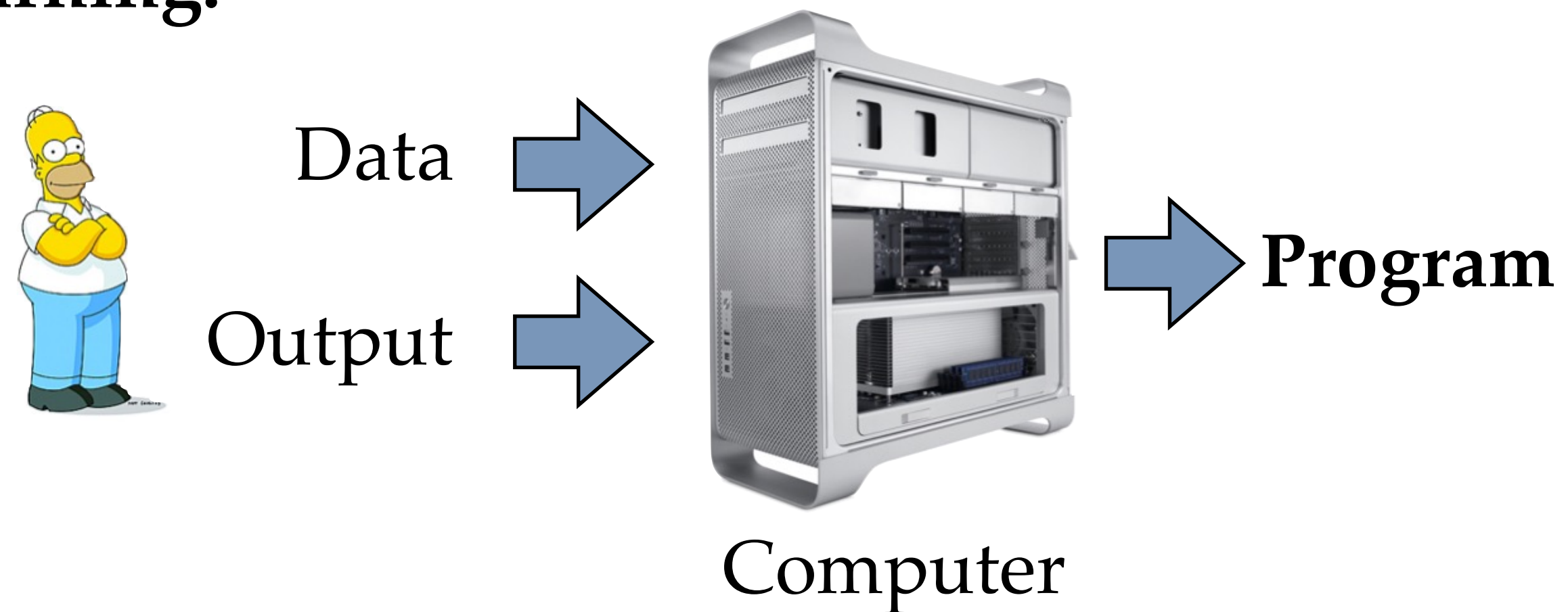
Output

Machine Learning

Traditional CS:

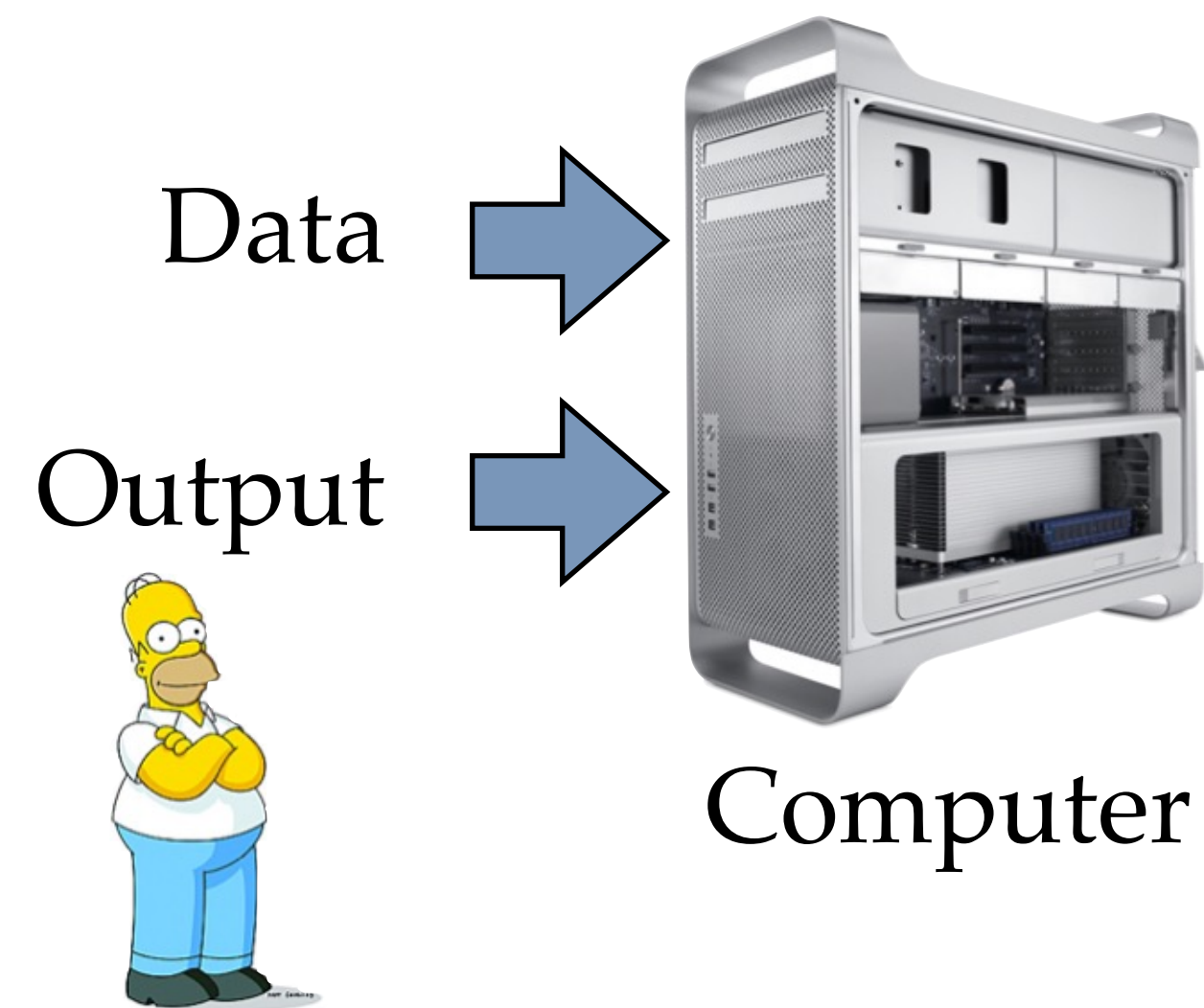


Machine Learning:

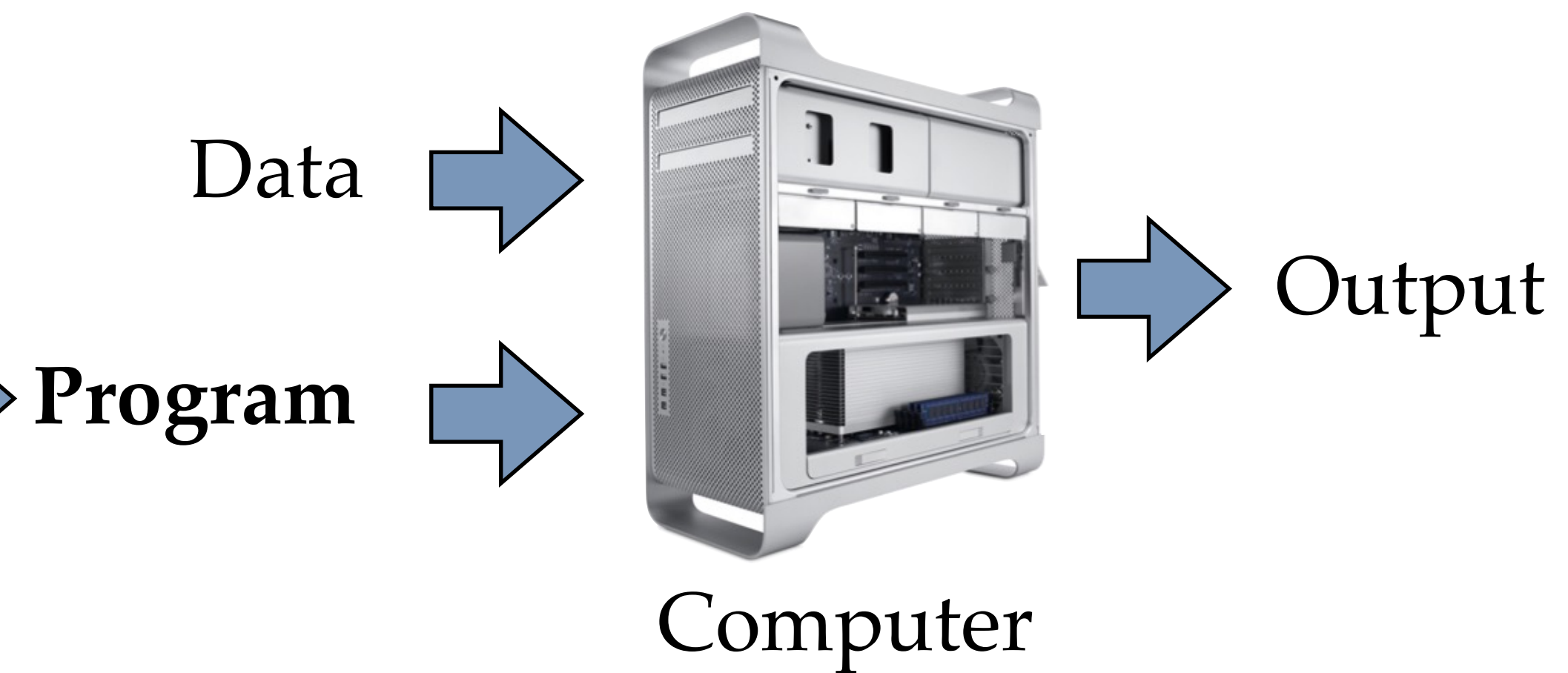


Machine Learning

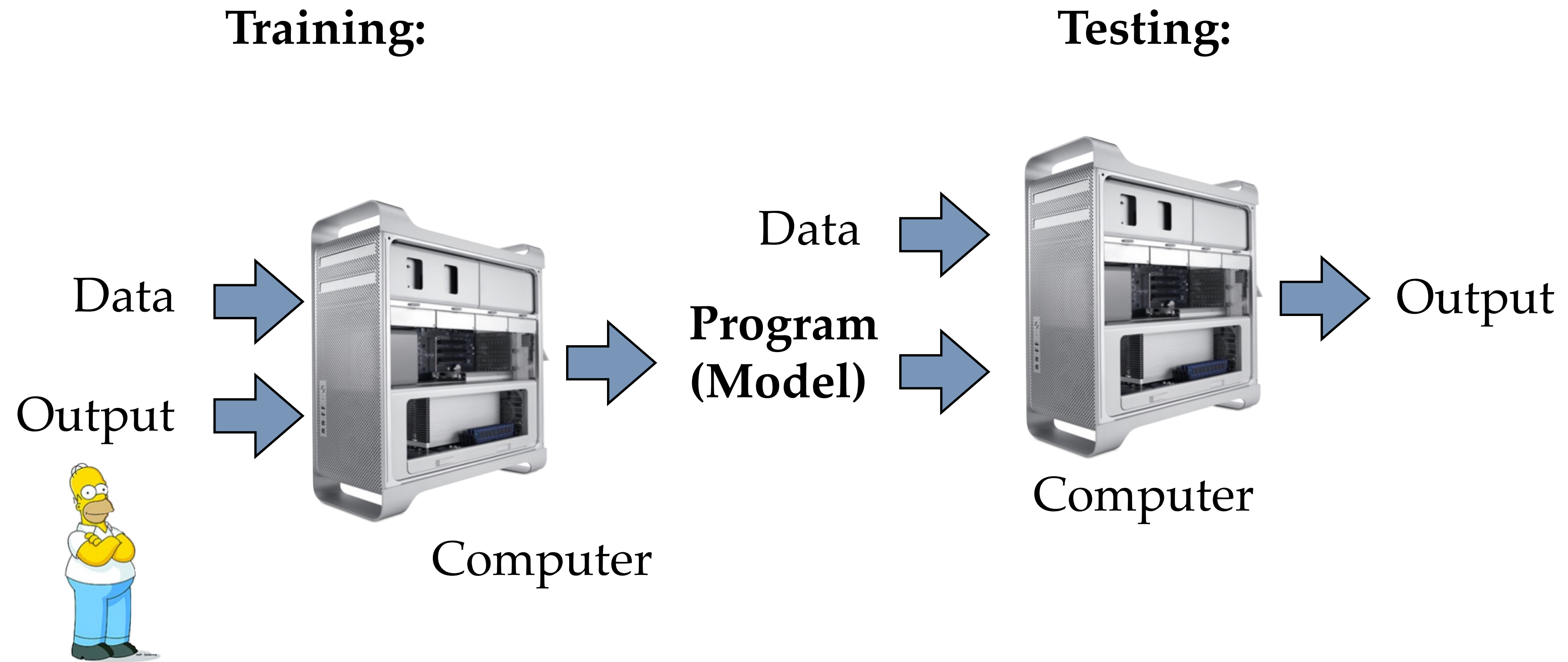
Machine Learning:

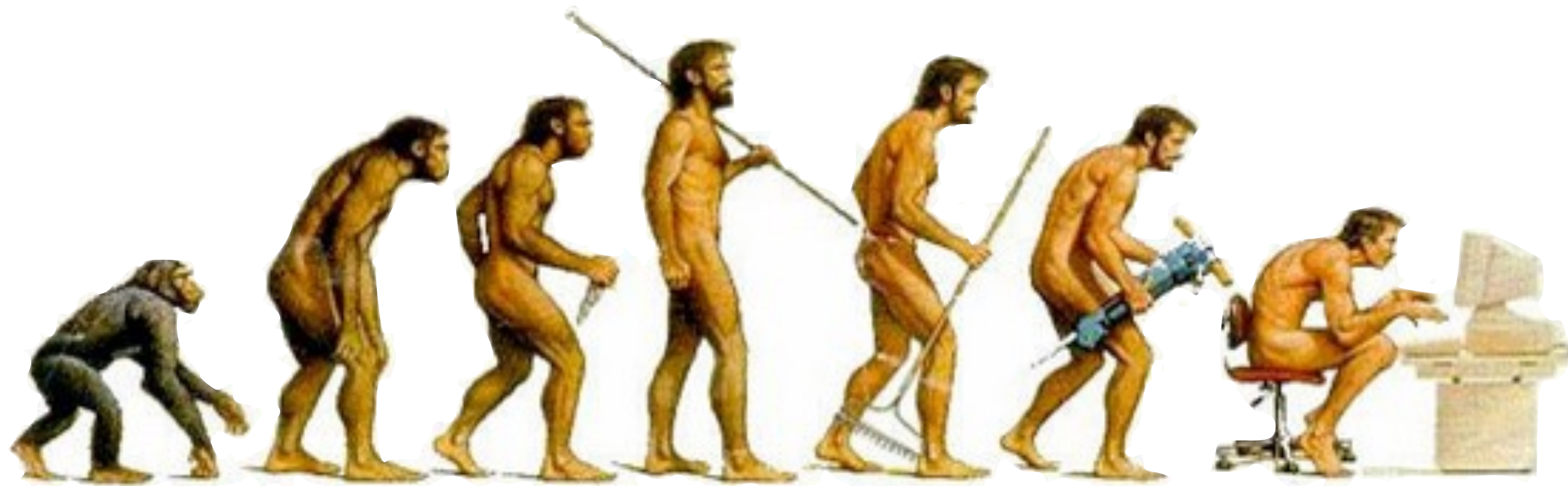


Traditional CS:



Machine Learning





A (very brief) History of ML



Alan Turing

The Turing Test, 1950

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

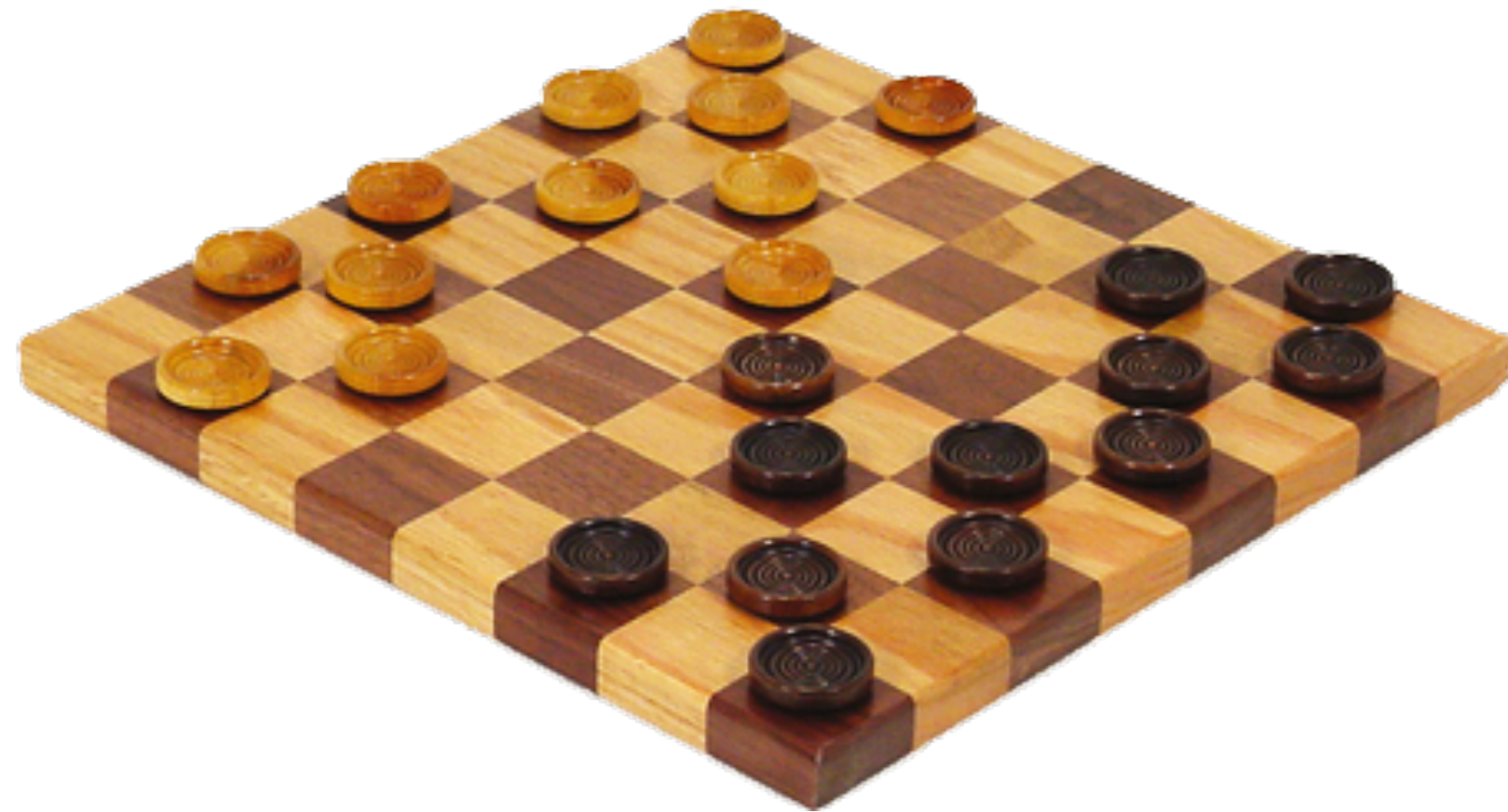




Arthur Samuel

Checkers Program, 1952

Created a Checkers-playing program that got better overtime.



Also introduced the term
“Machine Learning.”

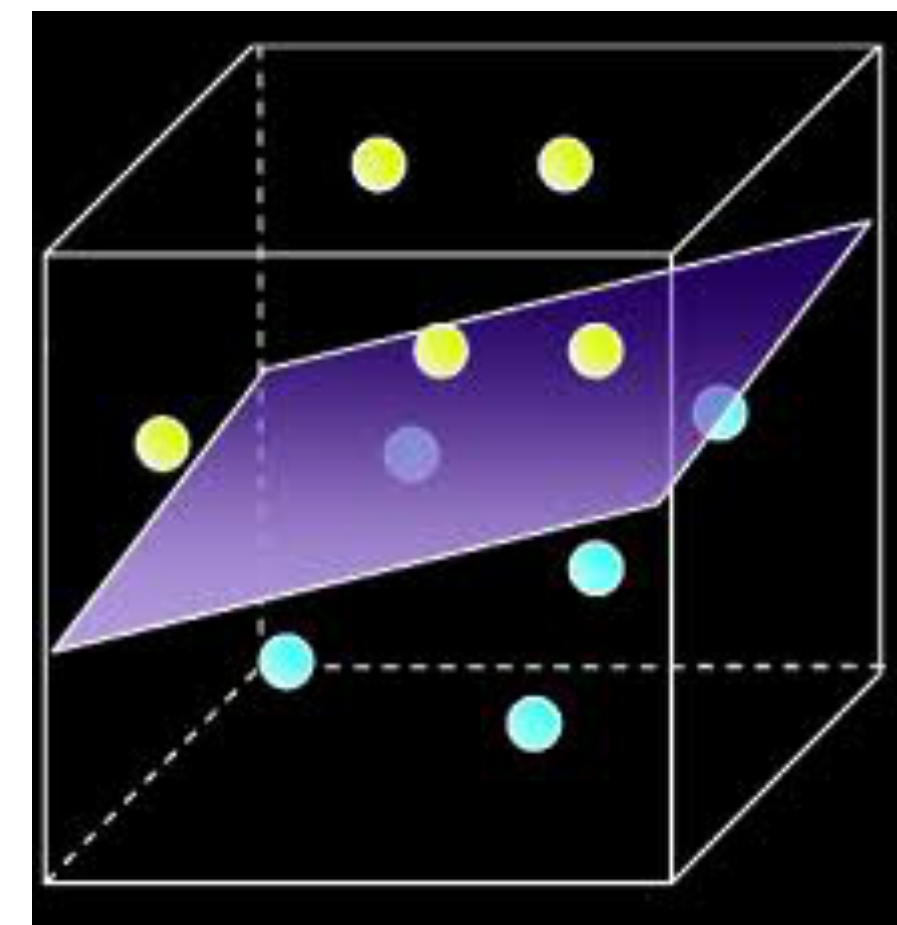
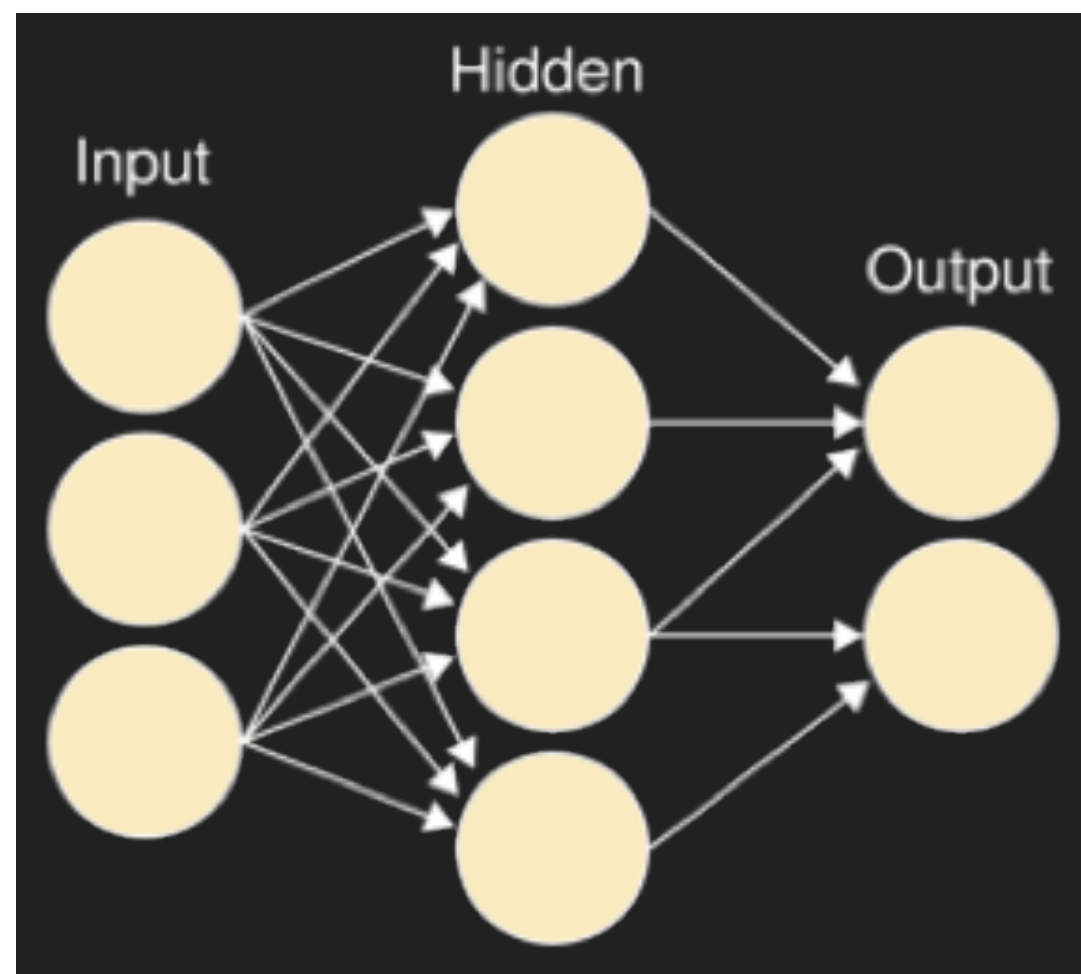


Perceptron, 1957

Predecessor of deep networks.

Frank Rosenblatt
@ Cornell!

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.

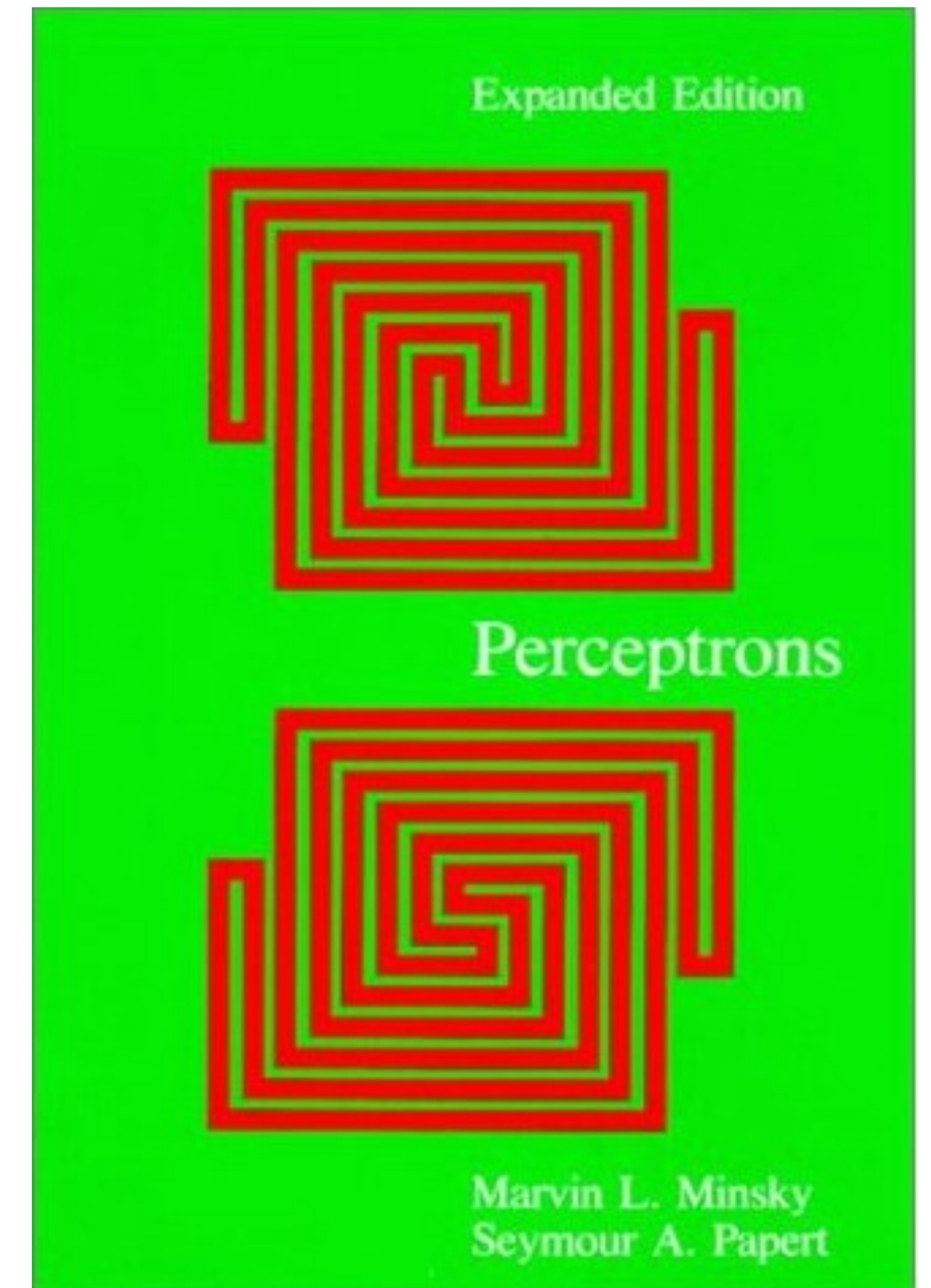
Perceptron, 1957

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.”

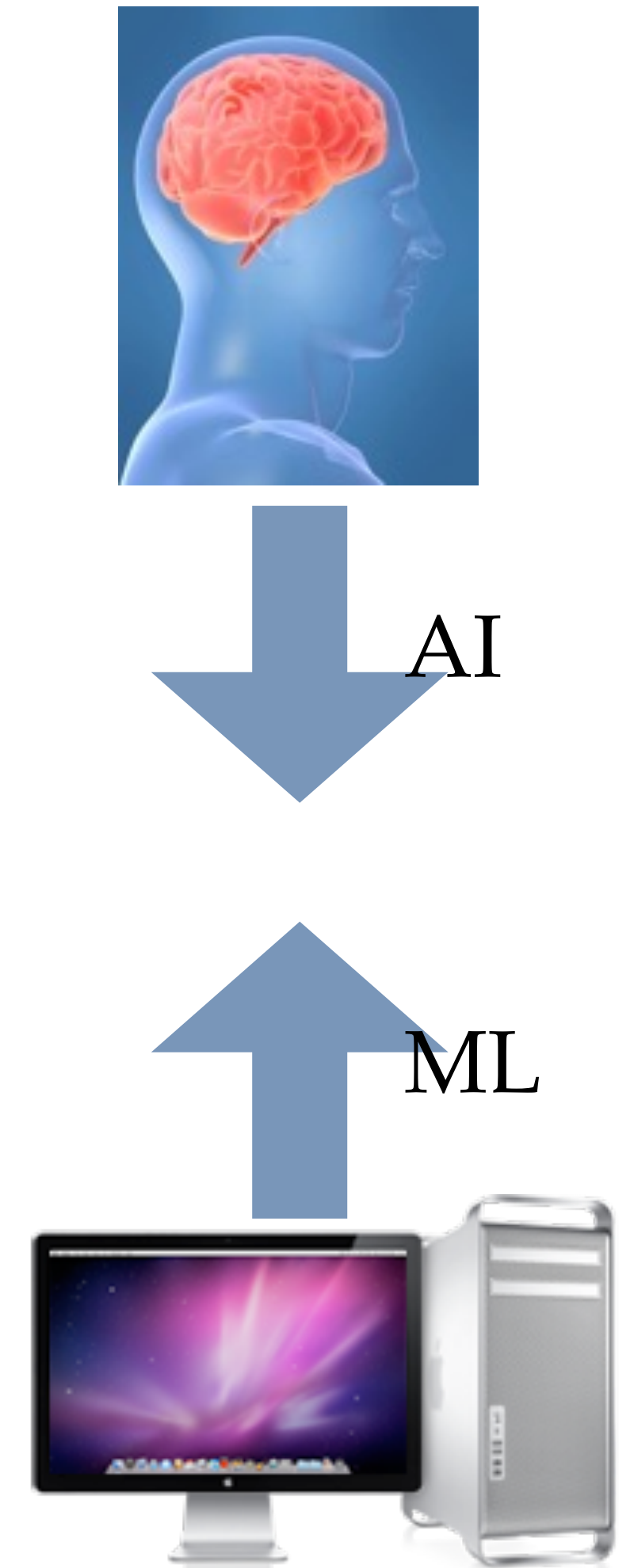
AI Winter (1974-1980)

- (1969) Minsky & Papert “killed” AI
 - simple XOR dataset
- 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 teaches **itself** how to play so well!!!
- Algorithm found new techniques that people had erroneously ruled out.



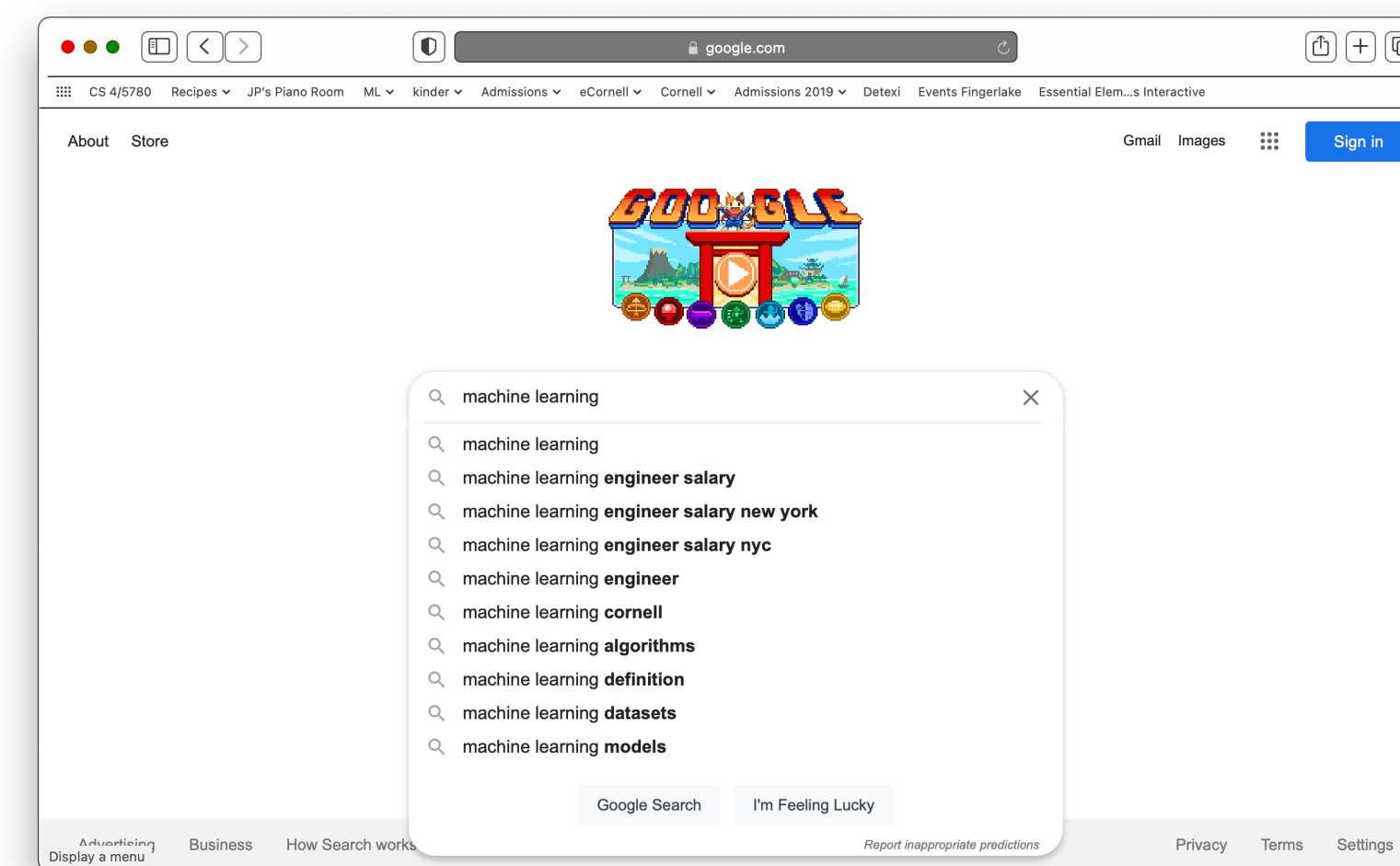
Deep Blue (1997)

- 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)



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.



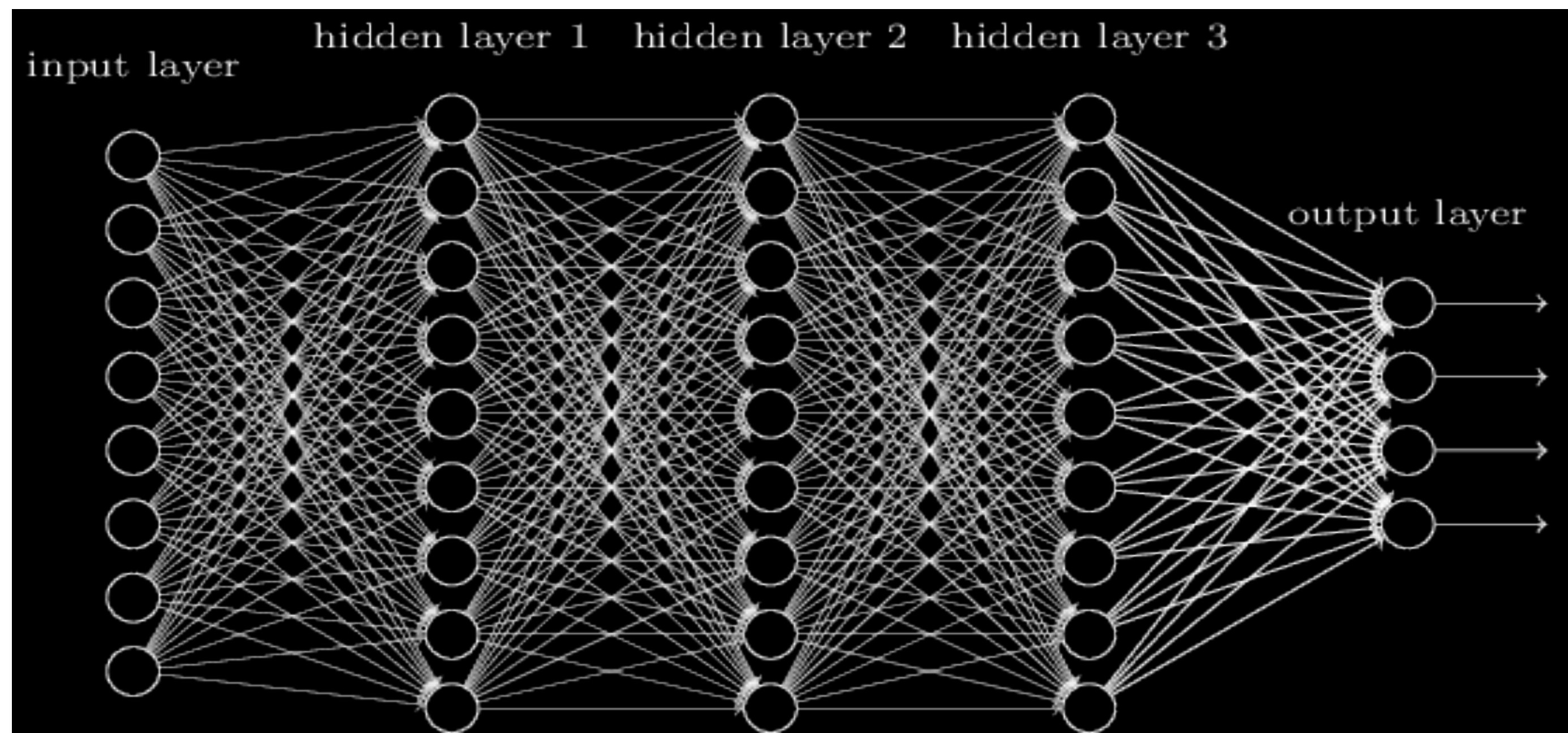
\$\$\$

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
- Big shock
 - China invests heavily in AI research
- Beginning of "AI arms race"

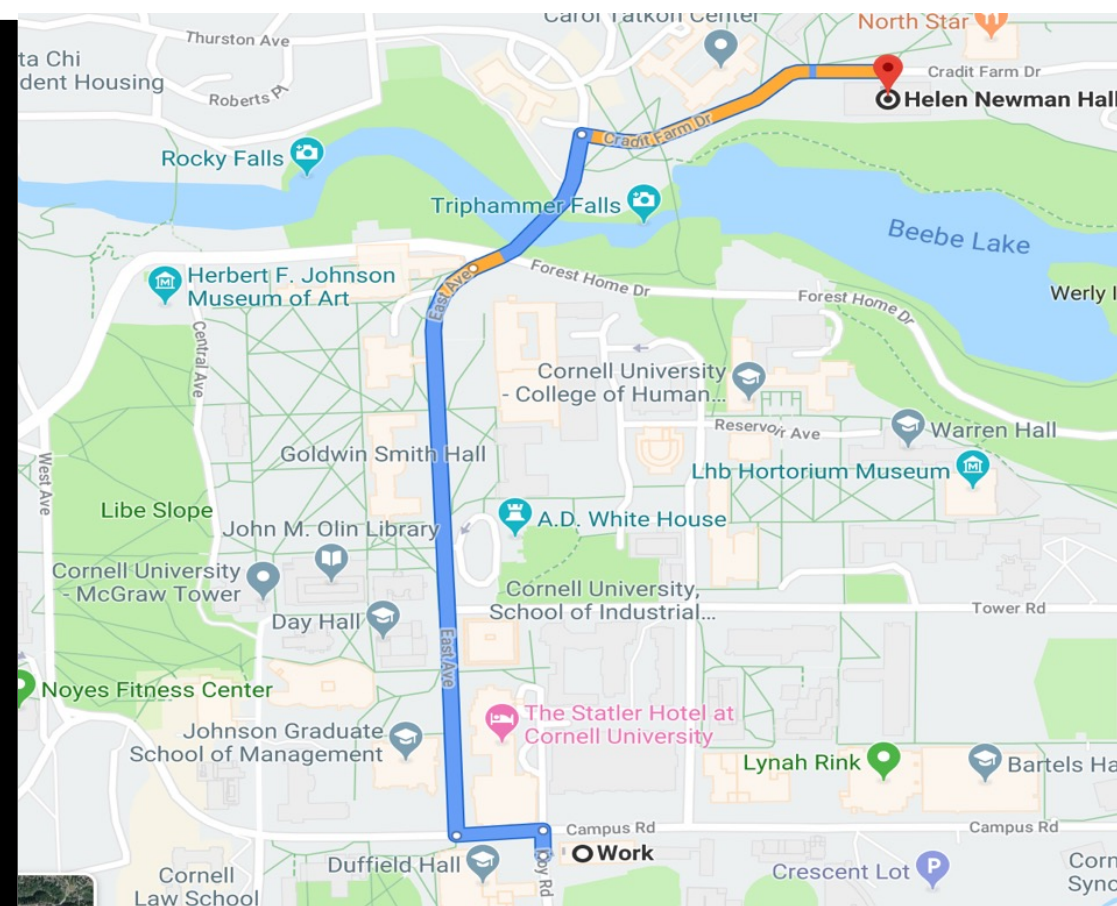


Today We're Surrounded by Machine Learning

Google Translate

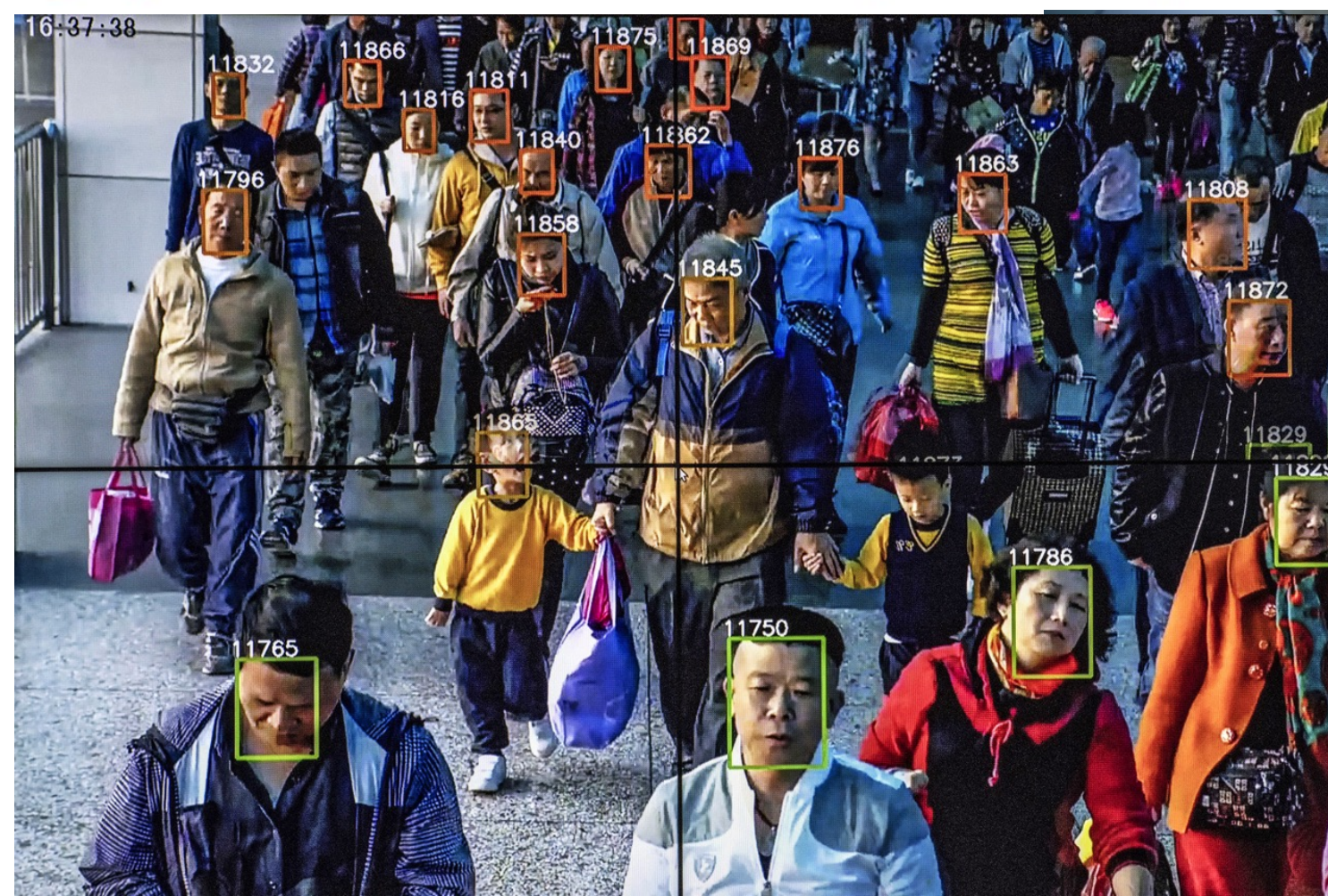
machine learning

فراگیری ماشین



Azure Machine Learning | Create Your Free Account Today

[Ad azure.microsoft.com/Services/MachineLearning](https://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:** 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



Data privacy / misuse

Learning models leak training data
(Fredrickson et al. '15)

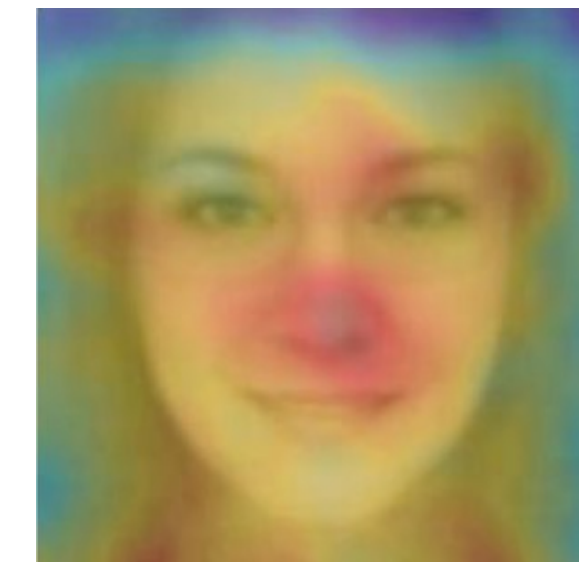
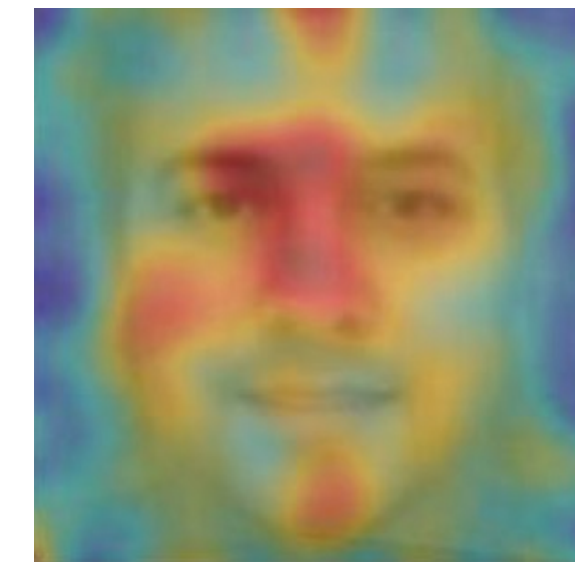


Leaked data



Real image

Learning algorithms detect sexual orientation better than people
(Wang & Kosinski'17)



Formal definitions of data privacy:

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



Latanya
Sweeney



Cynthia Dwork



Frank
McSherry



Kobbi Nissim



Adam Smith

Robust and Secure ML

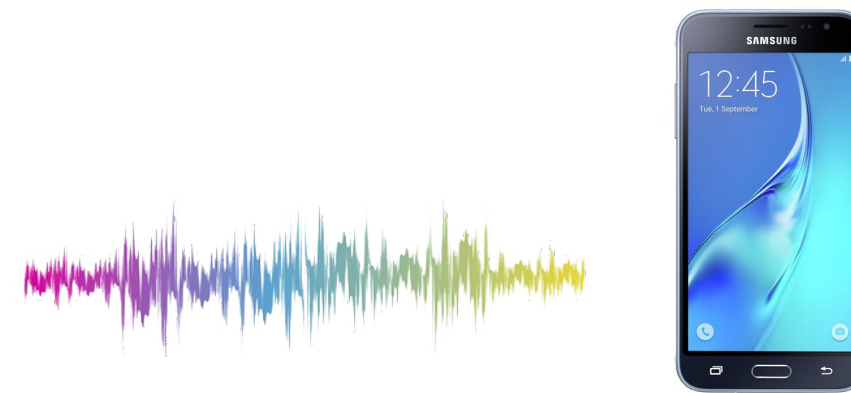
Image Recognition

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?

Learning and the Society

- Bad dynamics perpetuate and worsen stereotypes and biases.
- Who carries the burden of bad prediction?
- How to design good dynamics?

The Best Algorithms Struggle to Recognize Black Faces Equally

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

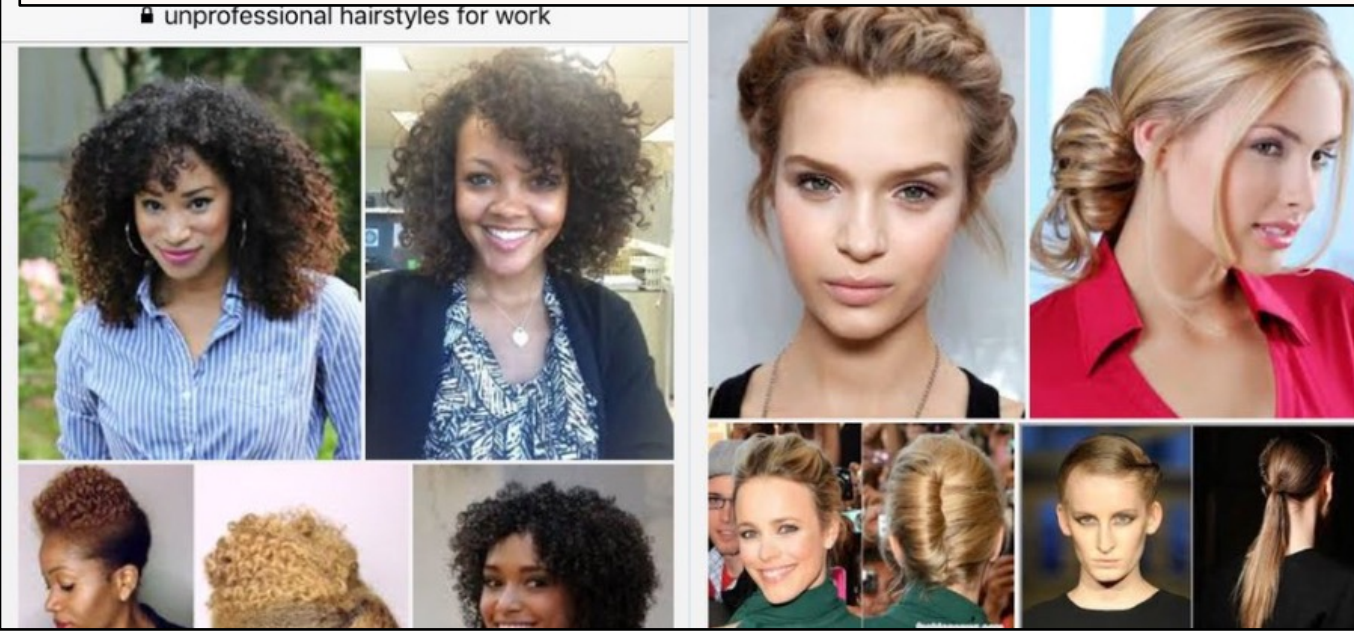
Gender and racial bias found in Amazon's facial recognition technology (again)

How Amazon Accidentally Invented a Sexist Hiring Algorithm

A company experiment to use artificial intelligence in hiring inadvertently favored male candidates.

Do Google's 'unprofessional hair' results show it is racist?

unprofessional hairstyles for work



When an Algorithm Helps Send You to Prison

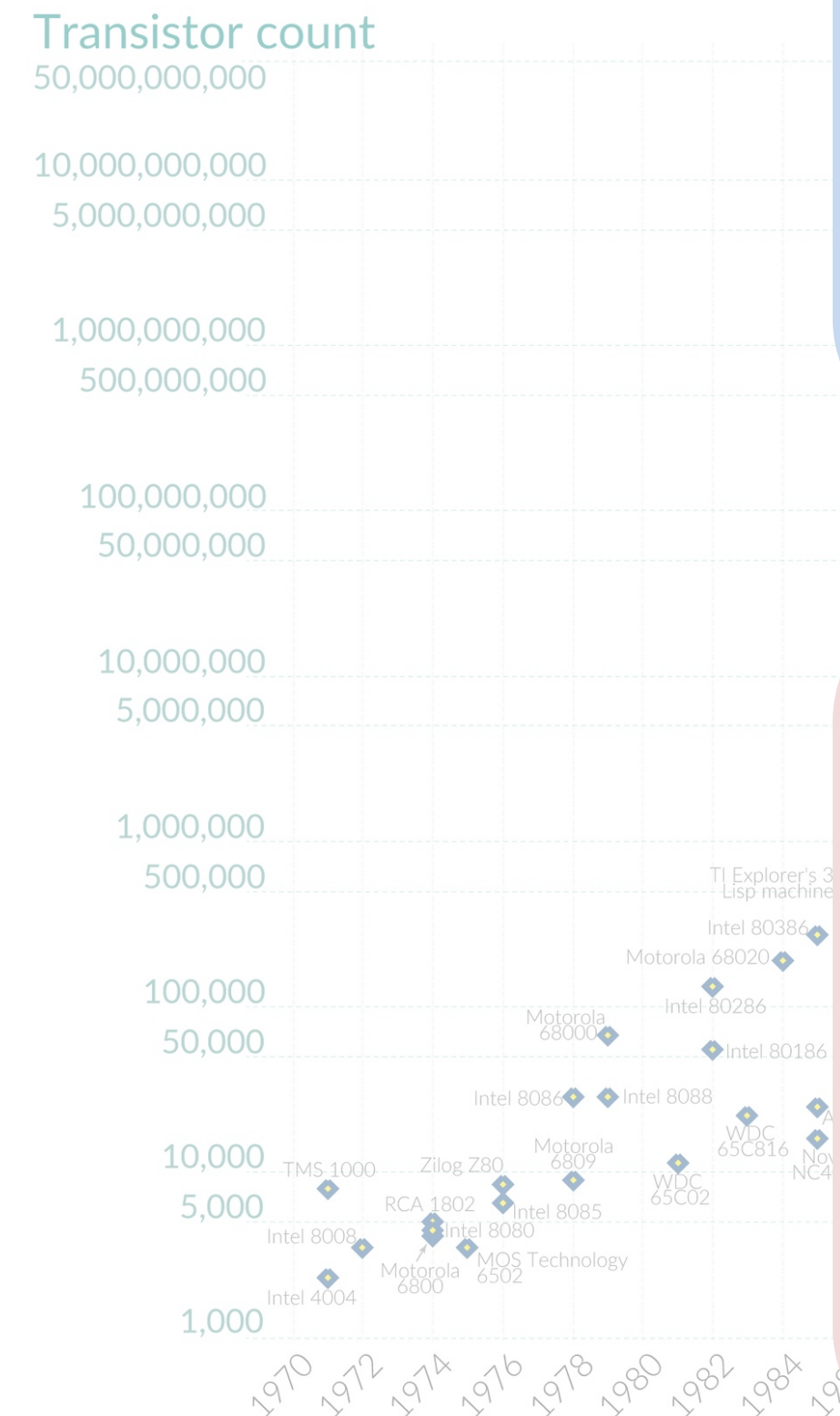
By Ellora Thadaney Israni



Will AI take over the world?

The

Moore's Law: The number of tra
Moore's law describes the empirical regularity that the nu
This advancement is important for other aspects of techn



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org - Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Good news:

AI is nowhere near to general Intelligence (no real progress)

Bad news:

AI doesn't have to be smarter than us to be harmful

ing Papers
Law?



AI/ML is different from humans

- Where and how we 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 use is appropriate

To-do action items

- Finish Placement Exam!
 - Start Project -1
- Find study group
- Autograder Setup:
 - Find Teammate
 - Start Project 0