CS 3/5780: Introduction to

Machine Learning

Lecture 1

Sarah Dean   Karthik Sridharan

First the Boring Administrative Stuff . . .

Course Links

- **Homepage:** https://www.cs.cornell.edu/courses/cs3780/sp2026
  - Most of the information about the course

- **Canvas:** https://canvas.cornell.edu
  - For assignments, quizzes etc

- **Vocareum:** http://vocareum.com
  - Programming assignments are here

- **Ed:** https://edstem.org/us/courses/69608/
  - Discussion and questions

Where to get help?

- **Post all questions** on ED (fastest way to get answers)
  - For Emergency or privacy email: intro-ml-profs@cornell.edu
  - DO NOT Email instructors directly!

- **Instructor Office Hours:**
  - Sarah Dean: Tue 3pm-4pm, 424 Gates Hall
  - Karthik Sridharan: Wed 10:30am-11:30am, CIS 477

- **TA Office Hours:** Every day
  - Details will be posted on course webpage.
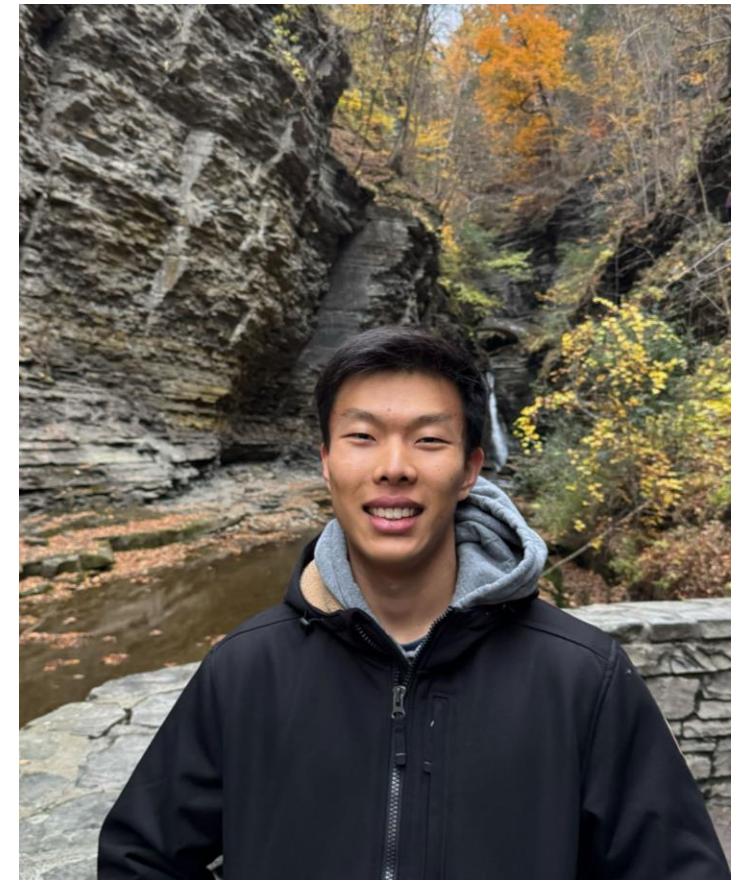
Leon, Sherlock

Srivatsa, Breaking Bad

Jess, Ikebukuro West Gate Park

Nick, Grey's Anatomy

Frank

Lauren, The Great Pottery Throw Down

David, Severance

Eric, Brooklyn 99

Haozheng, Better Call Saul

Serhii

Huajie

Boaz, Alice in Borderland

Xenex, Peaky Binders

Hami, The Office

Shriya, 11.22.63
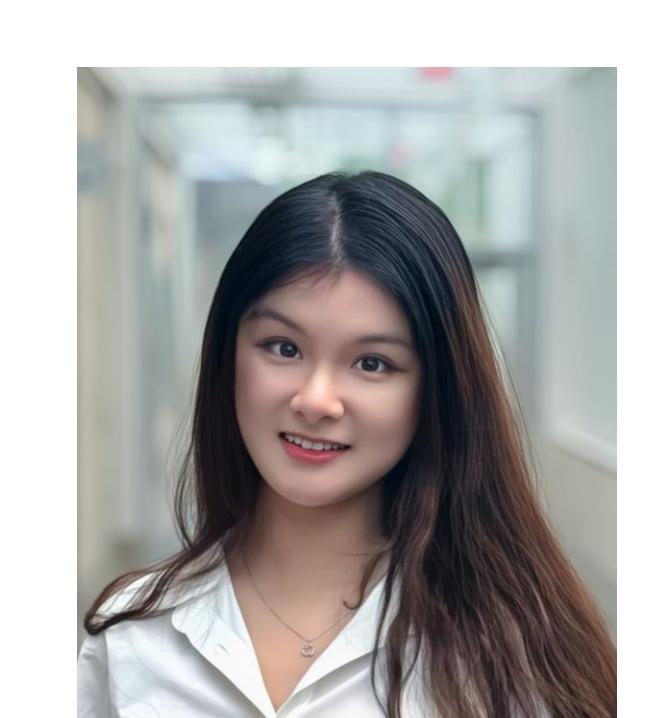
Rose, The Queen's Gambit

Jefferson, One Piece

Alex, Blacklist

Alexander, Dragon Ball

Zain, Avatar The Last Airbender

Zijie, Pitt

Kevin Beef

Jengwen, Inside No. 9

Janet, Squid Game

Jack, Fallout

Nedim, One Piece

Jade, Gilmore Girls

Farhan, Sherlock

Judy

Leardon, Mr. And Mrs. Smith

Thurushi, Money Heist

Sharvari, Breaking Bad
Sharvari, Breaking Bad

Caleb, Black Mirror

Layla Liu
Johnson Wang
Aakaash Nattanmai
Bryan Chung
Claudiu Llioi
Jerry Ji
Andrew Noviello
Casper Liao
Aidan Wu
Saesha Agarwal
Jiaxuan Wang
**Timothy Li**

Sarah, Treme

Karthik, GOT

References:

Main Course Texts

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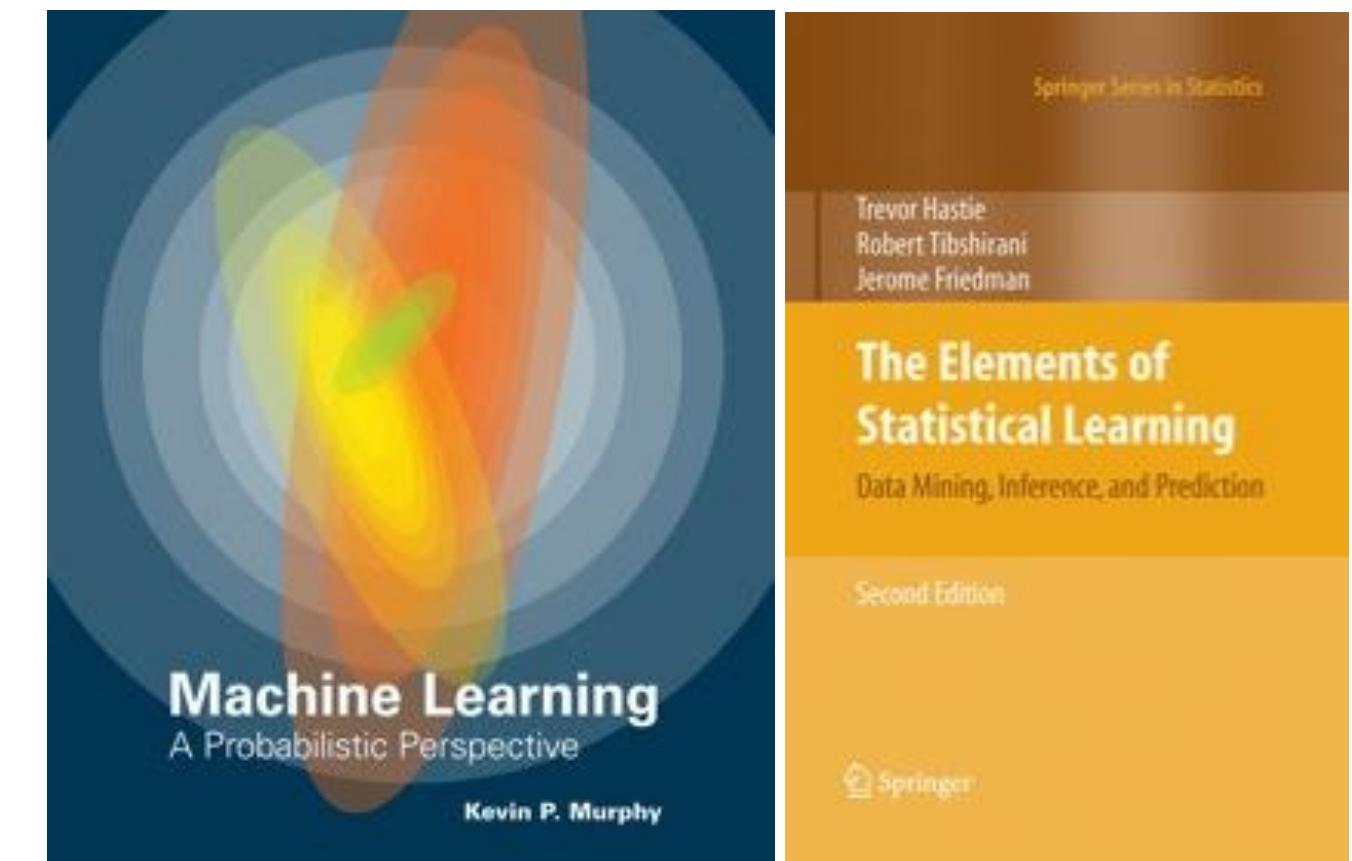
Additional Texts

- 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

- 45% Exams: Midterm + Final
  - Closed book
- 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% Homeworks
  - Up to 4 members in each team
  - **Preparation for exam**
- 10% Paper Comprehension (mandatory)
 - Original Research Papers in ML
  - Canvas Quizzes

- Exams
- Programming
- Paper Comprehension
- Homeworks
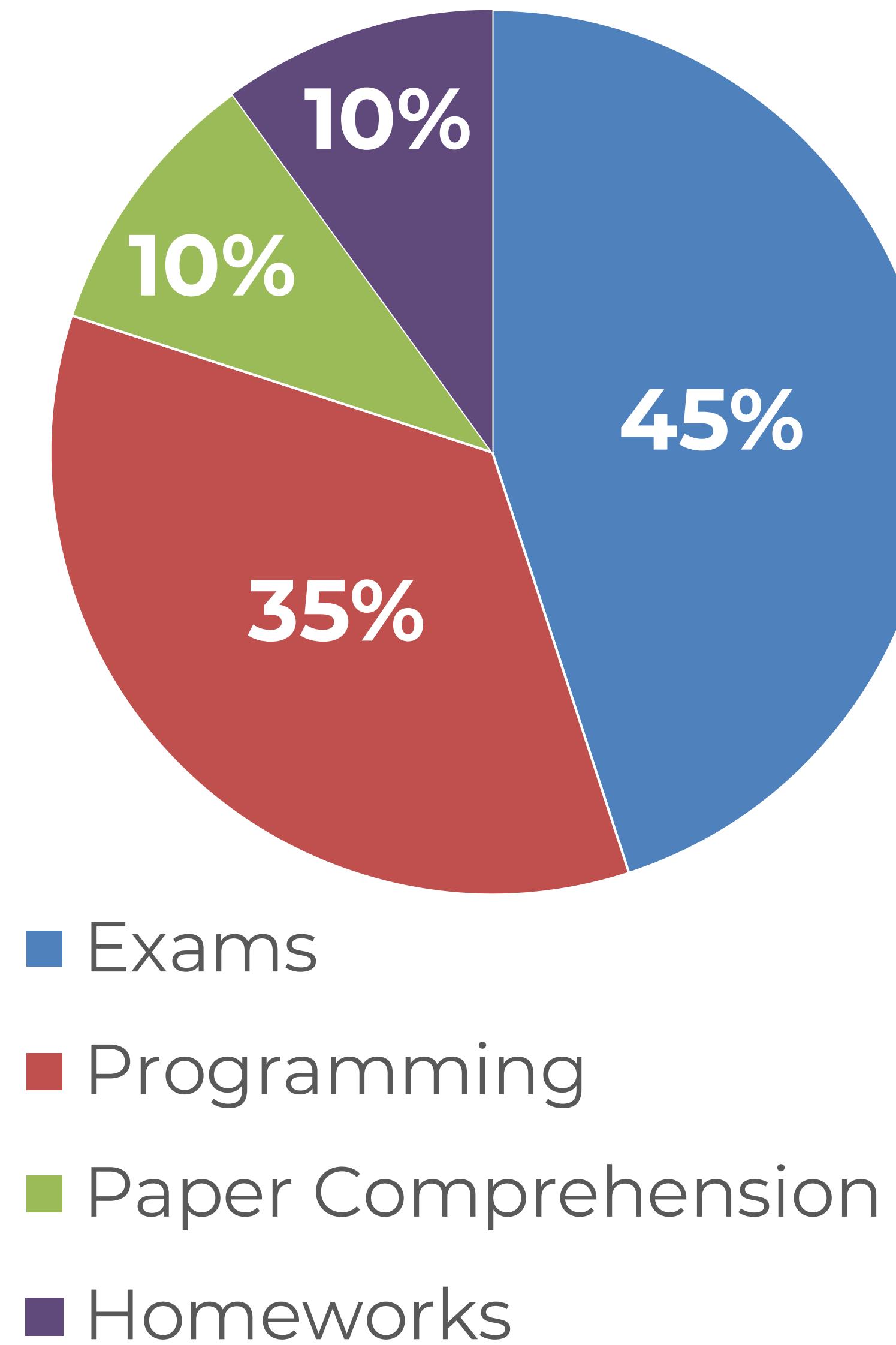
Course Breakdown 3780

- (45% or 48%) Exams: Midterm + Final
  - Closed book

- (35% or 37%)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% or 15%) Homeworks
  - Up to 4 members in each team
  - **Preparation for exam**

- (0% or 10%)Paper Comprehension(Optional)
  - Original Research Papers in ML
  - Canvas Quizzes

15%

48%

37%

- Exams
- Programming
- Homeworks
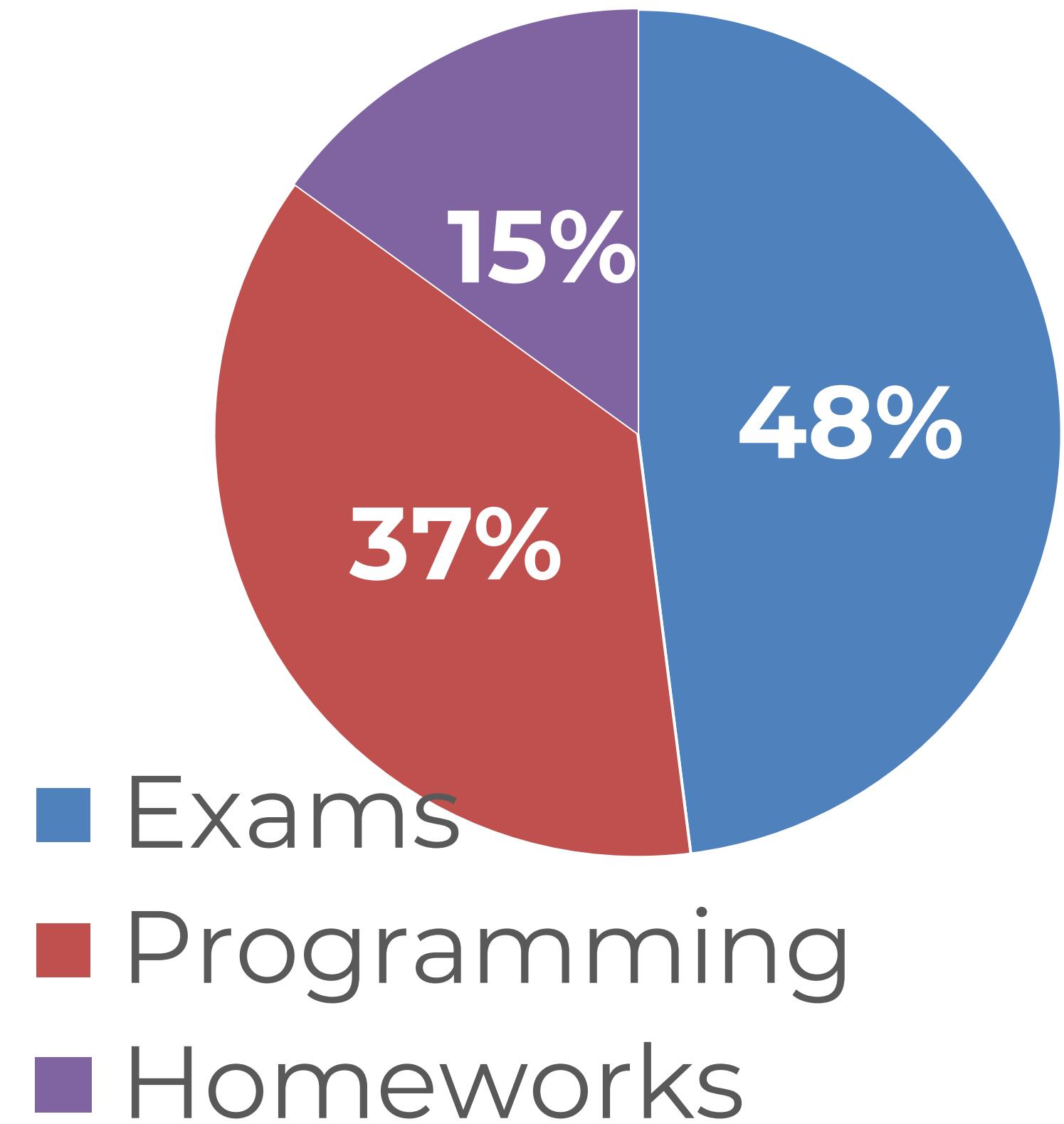
Placement Exam

- Due January 27th.
- **Canvas** page under Quiz.
- 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!
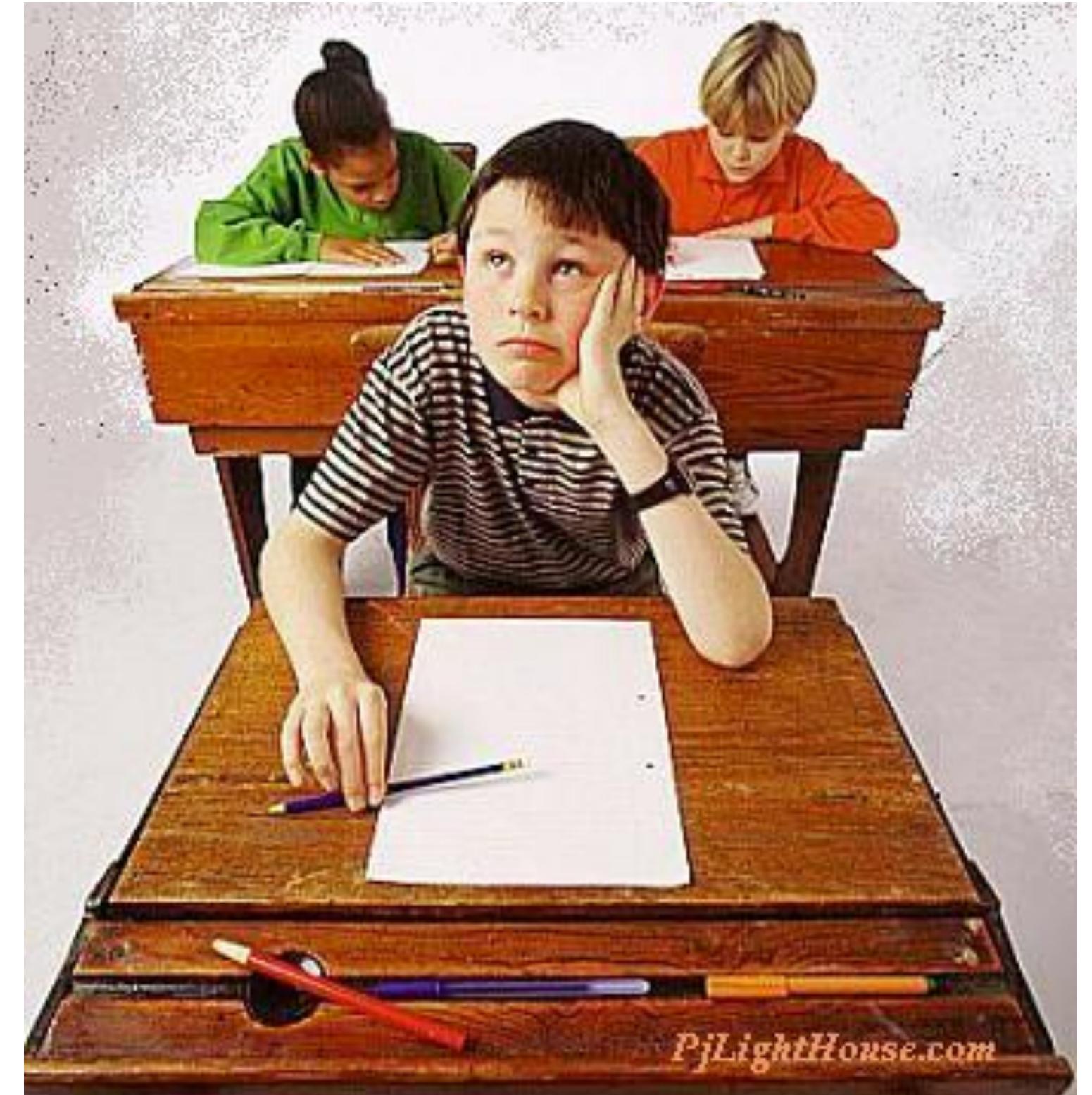
Study Groups (2–5 people)

- 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 :-(

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
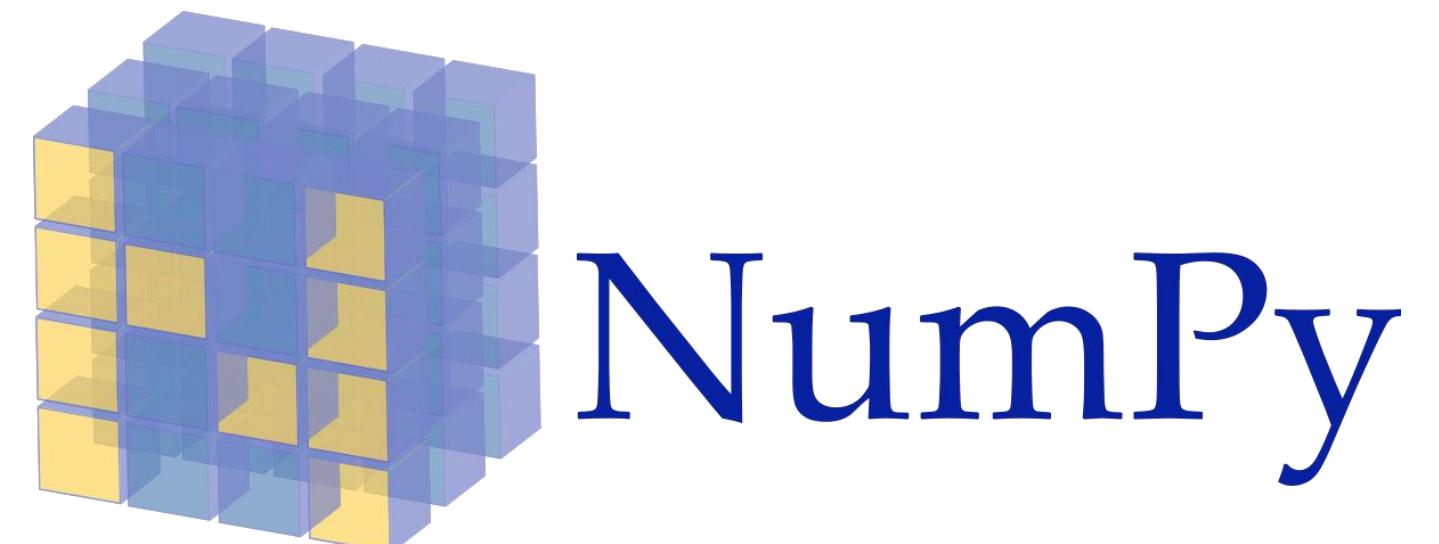
Autograder

- 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
- Basic Deep Learning

We will not cover

- Graphical Models -> [CS3700]
- Reinforcement Learning -> [CS4789]
- Learning Theory -> [CS4783]
- More Deep Learning -> [CS4782]
- Genetic Programming

Prerequisites

Introductory Calculus FOR INFANTS — omi m. inouye

Making Everything Easier! — Linear Algebra FOR DUMMIES — Mary Jane Sterling

Introduction to Probability, Statistics and Random Processes — Hossein Pishro-Nik
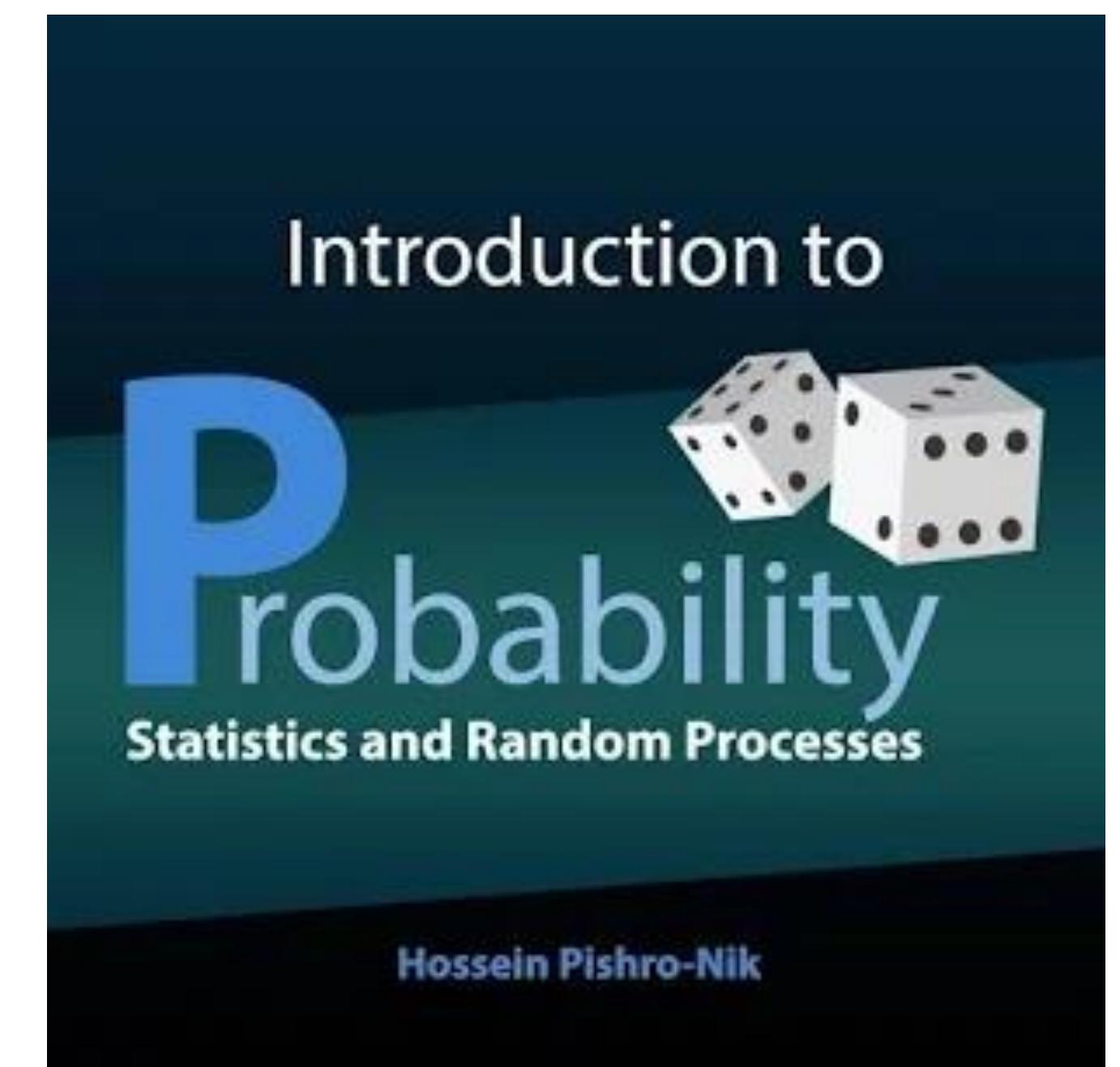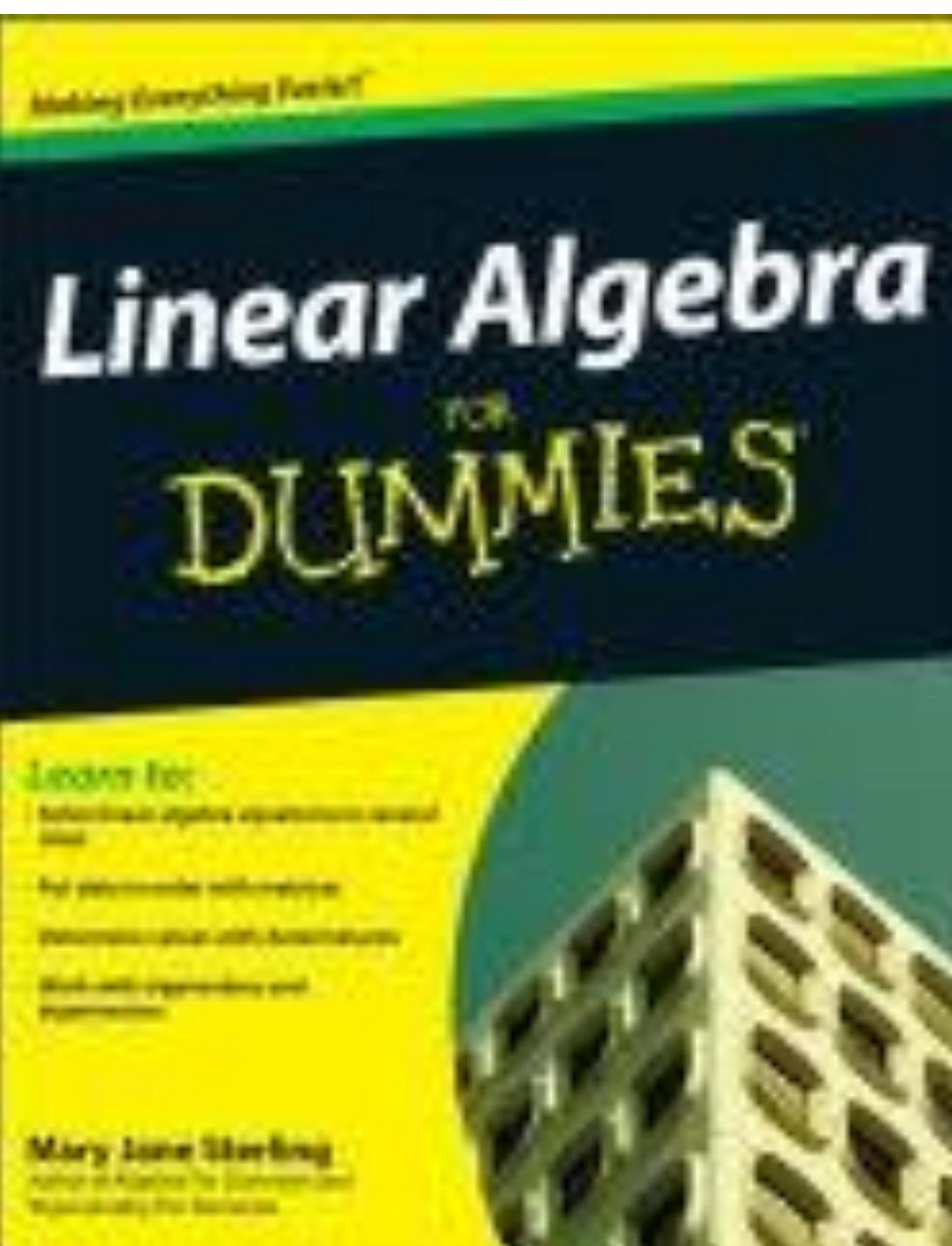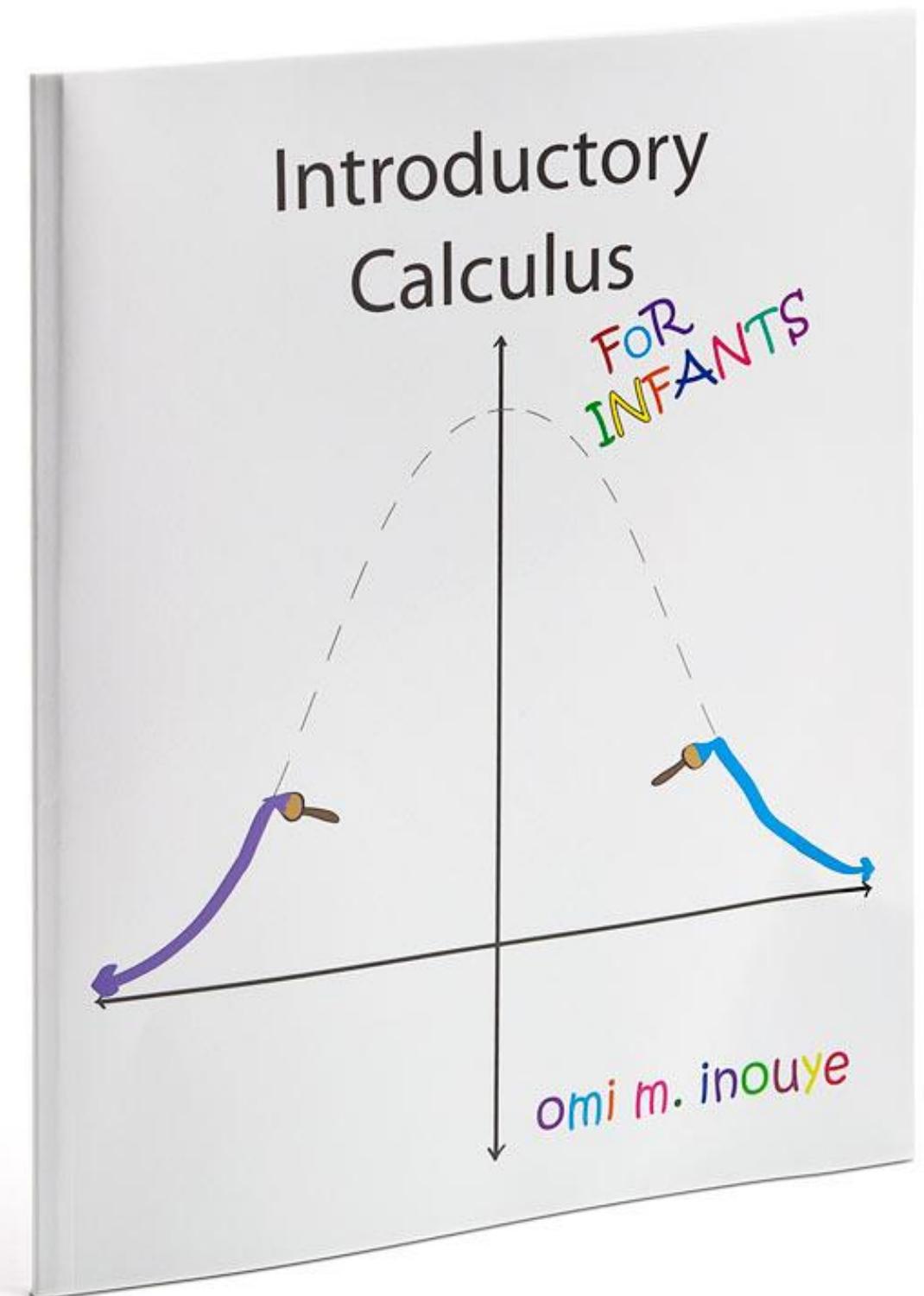
Prerequisites

- 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.
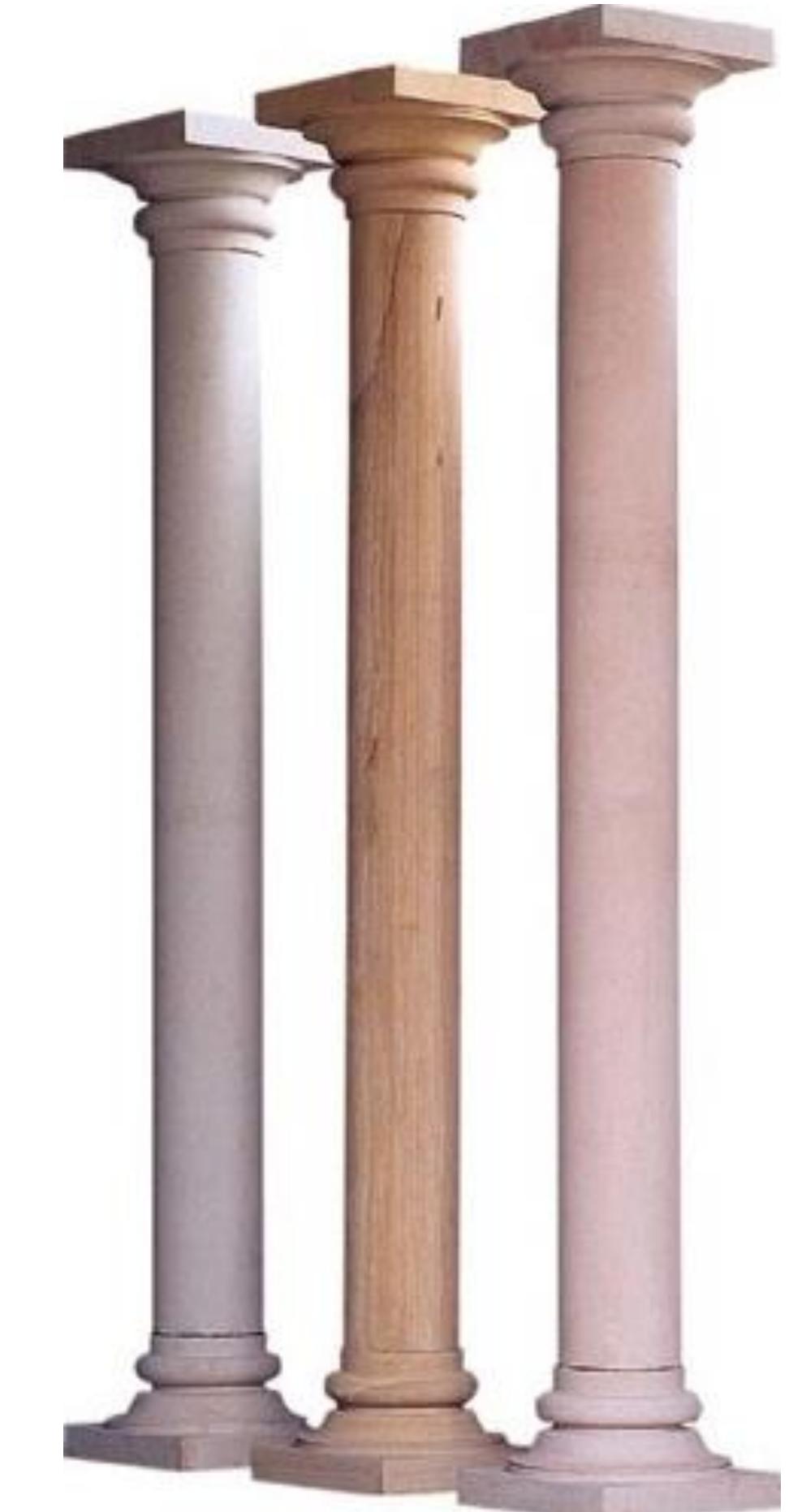
About this course

- **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

- We **actively look for academic conduct violations**
  - The autograder checks for plagiarism

- **Zero tolerance policy**
  - all occurrences will be reported (see course Webpage)

- *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

But Using LLM is OK...

- First try without…

- If you use it Cite it!

- You can use LLMs (e.g. ChatGPT, Bard) for your assignments, but you must hand in a detailed description of how you used it (including all prompts and outputs)

- Use without citing is an Academic Integrity Violation

- We will be using AI detection engines that have high accuracy.
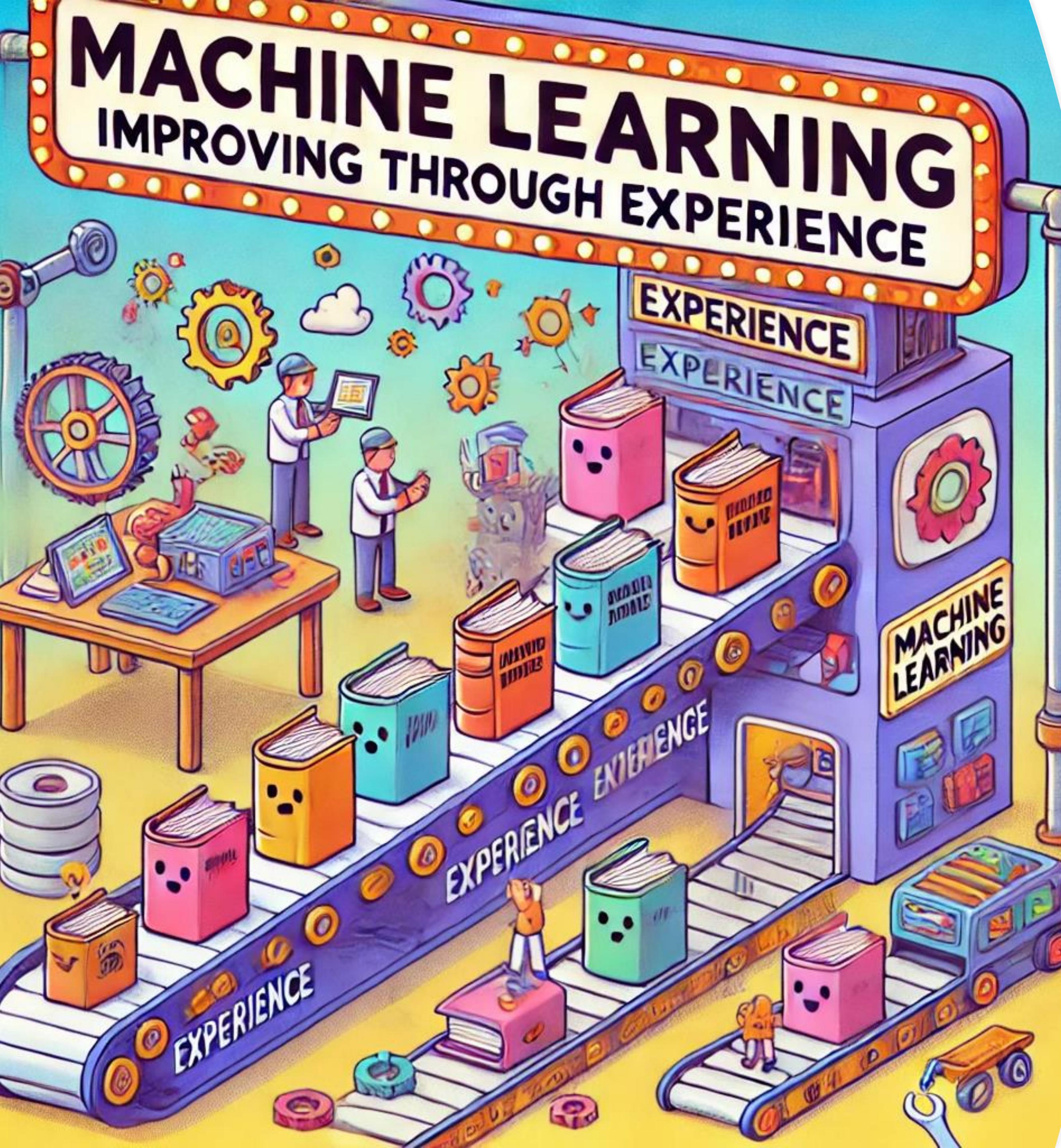
What is machine learning?

- Roughly: programs that **improve** with **experience**.
  - A subfield of **artificial intelligence**, which refers to programs that demonstrate "intelligence" in some sense

Traditional Computer Science

Traditional CS:

Data →

Program →

Computer

→ Output

Machine Learning

Traditional CS:

Data →

Program → Computer → Output

Computer

Machine Learning:

Data →

Output → Computer → **Program**

Computer
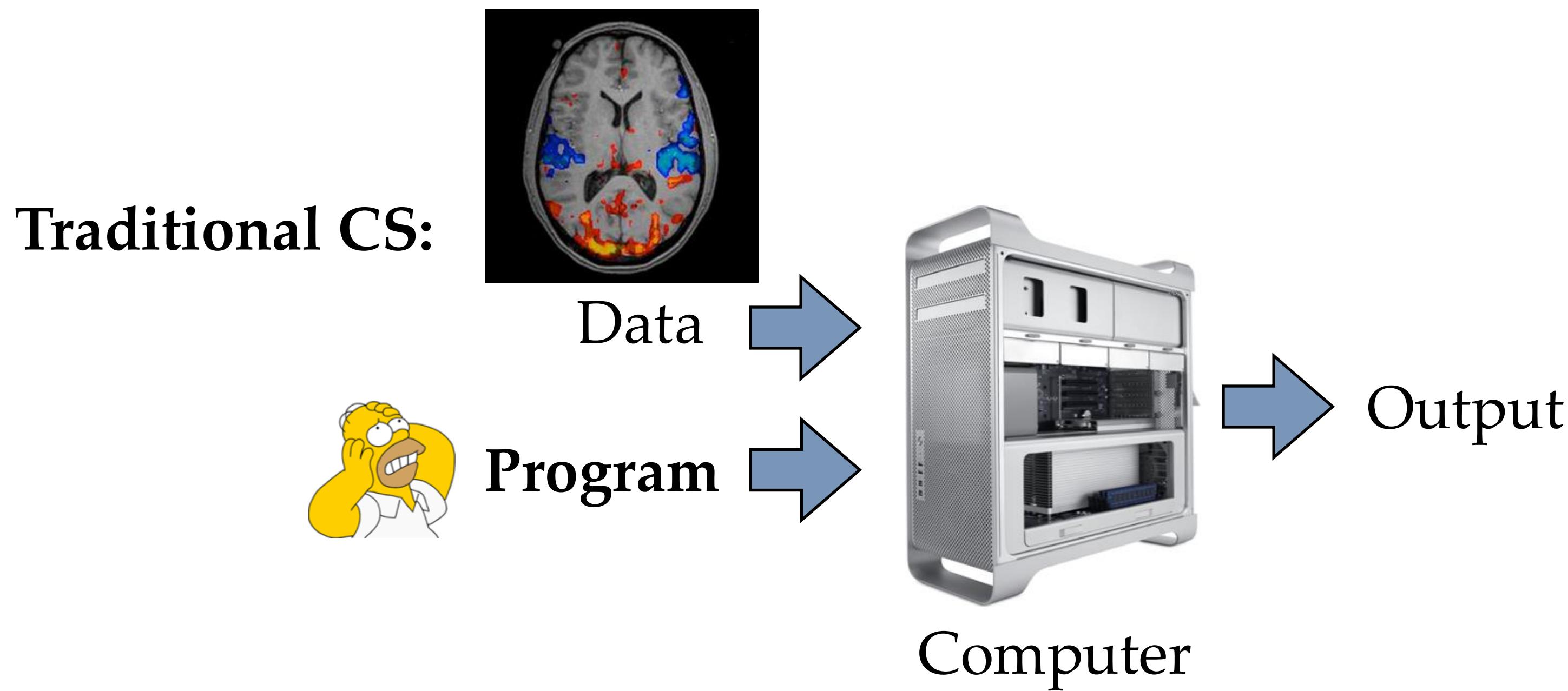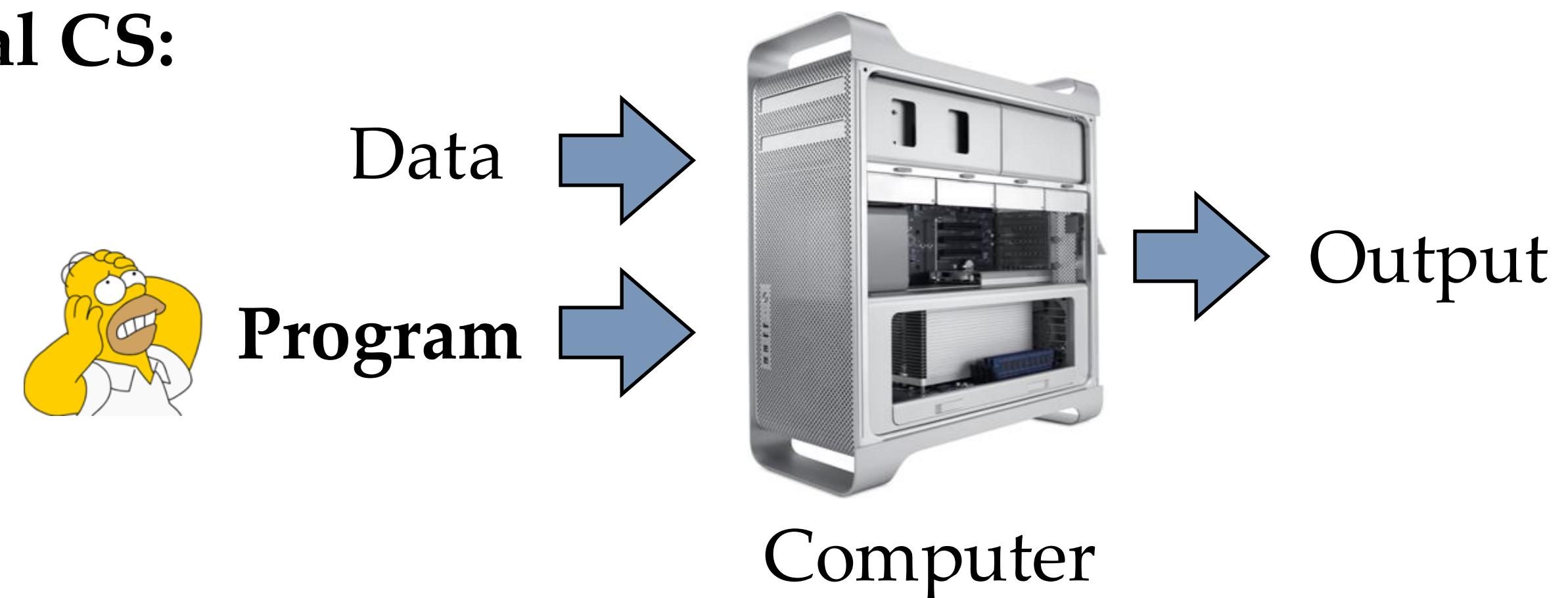
Machine Learning

Machine Learning: **Traditional CS:**

Machine Learning

Training:

Testing:

Data →

Output →

Program (Model) →

Computer

Data →

Data →

Output →

Output →
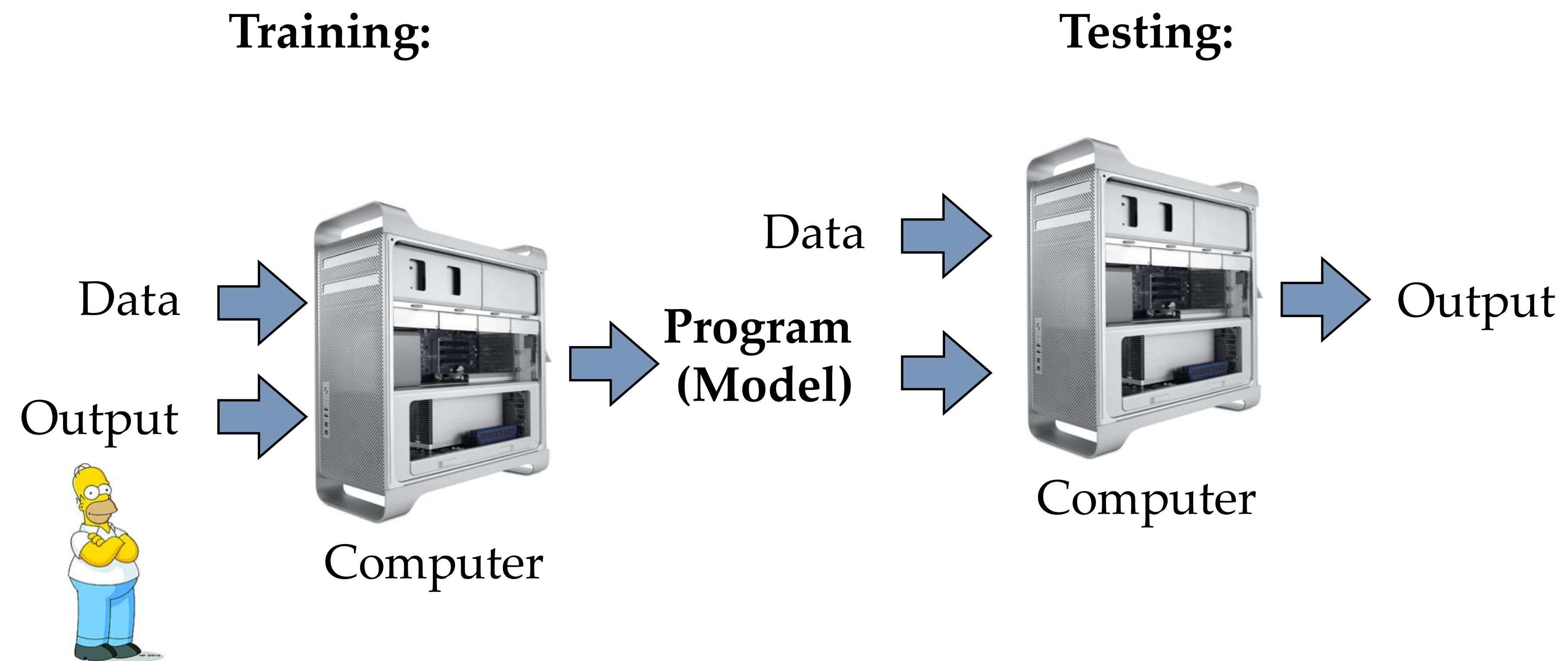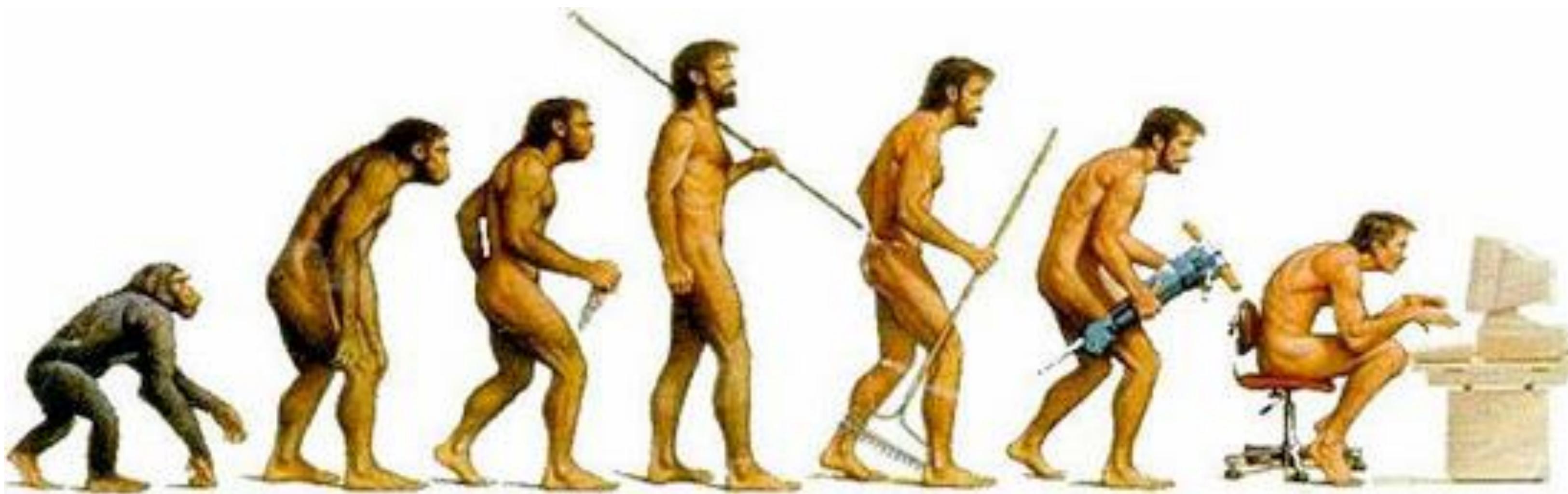
Computer

A (very brief) History of ML

A Mathematical Theory of Communication, 1948

Information can be *quantified* and *processed* using a digital computer.

Claude Shannon
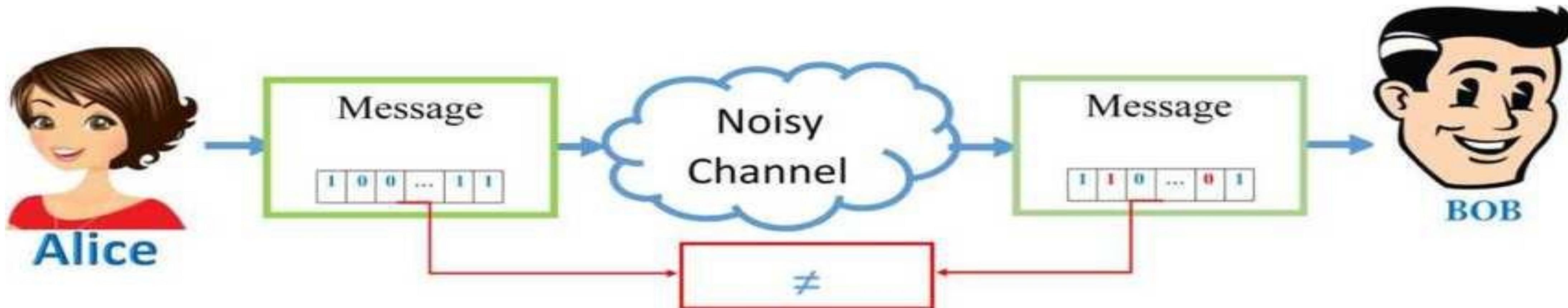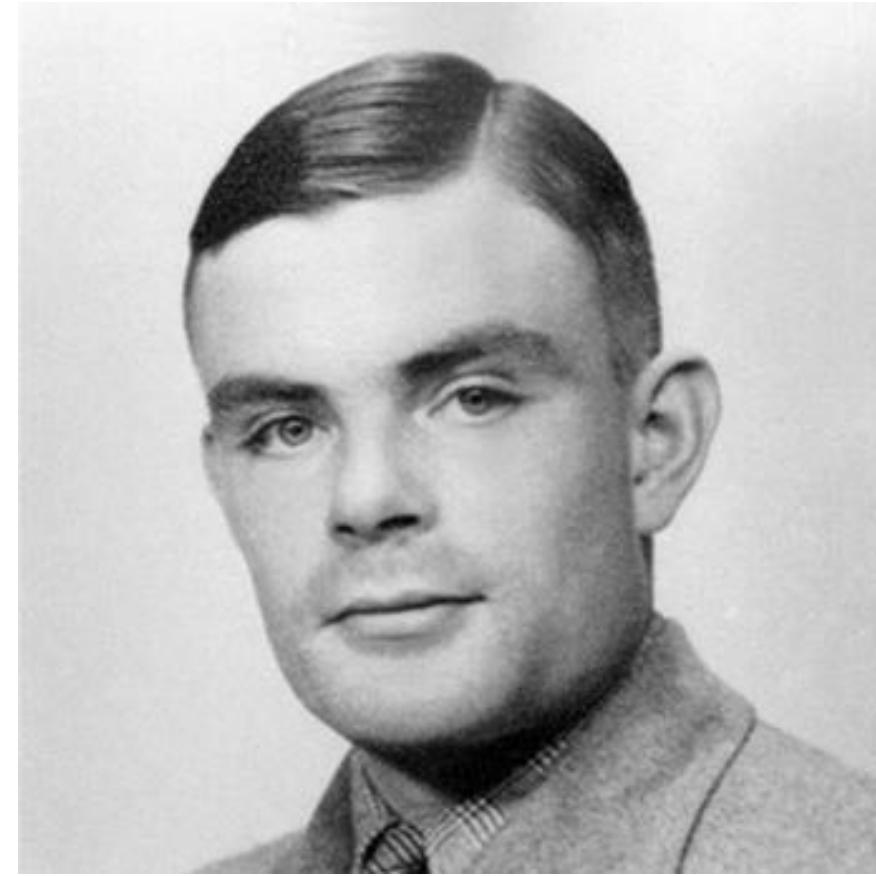
The Turing Test, 1950

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

Alan Turing

Checkers Program, 1952

Arthur Samuel

Created a Checkers-playing program that got better overtime.
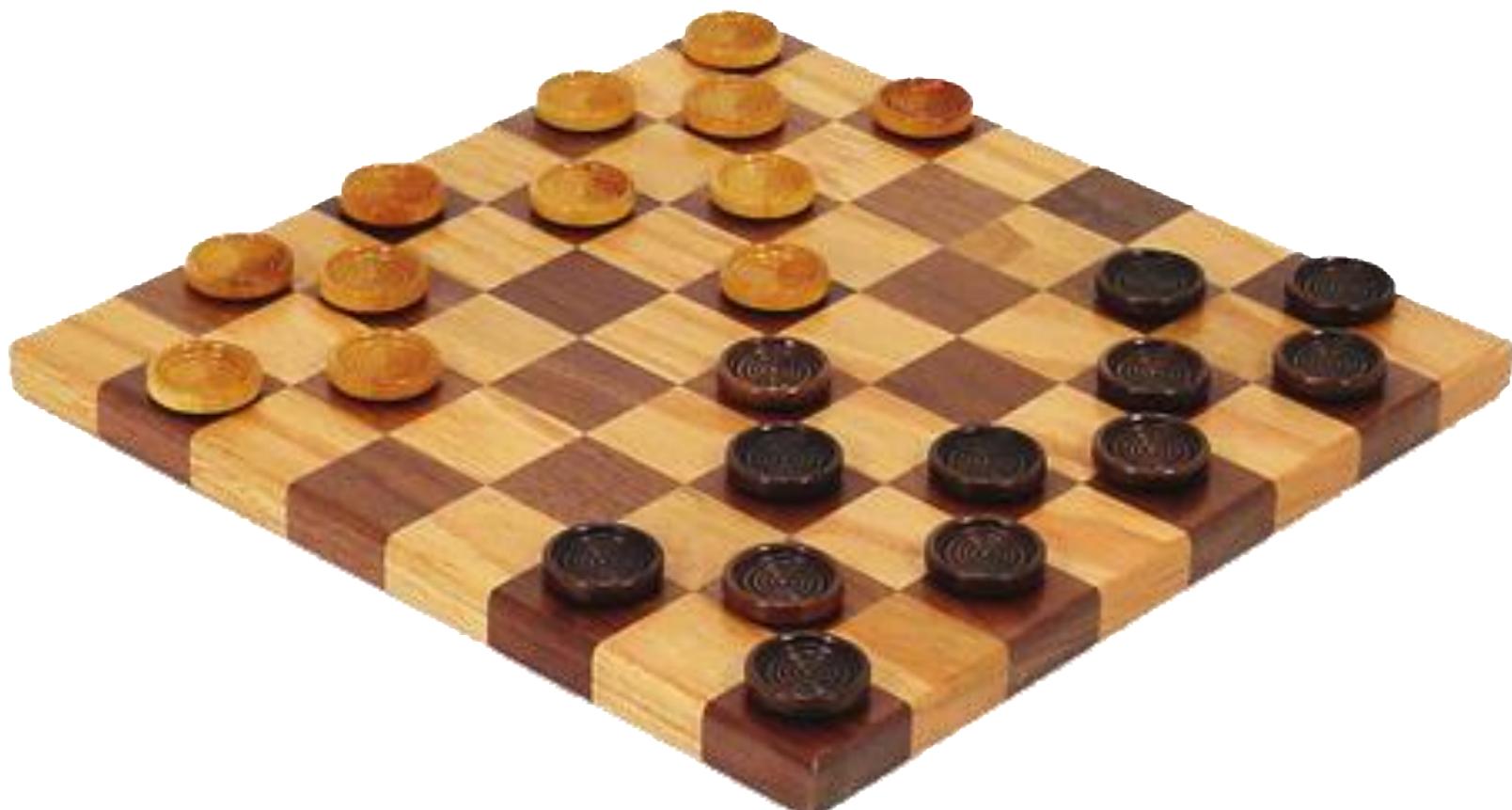
Also introduced the term "Machine Learning."

Perceptron, 1957

Predecessor of deep networks.

Separating two classes of objects using a linear threshold classifier.

Frank Rosenblatt
@ Cornell!

Frank Rosenblatt
@ Cornell!

Perceptron, 1957
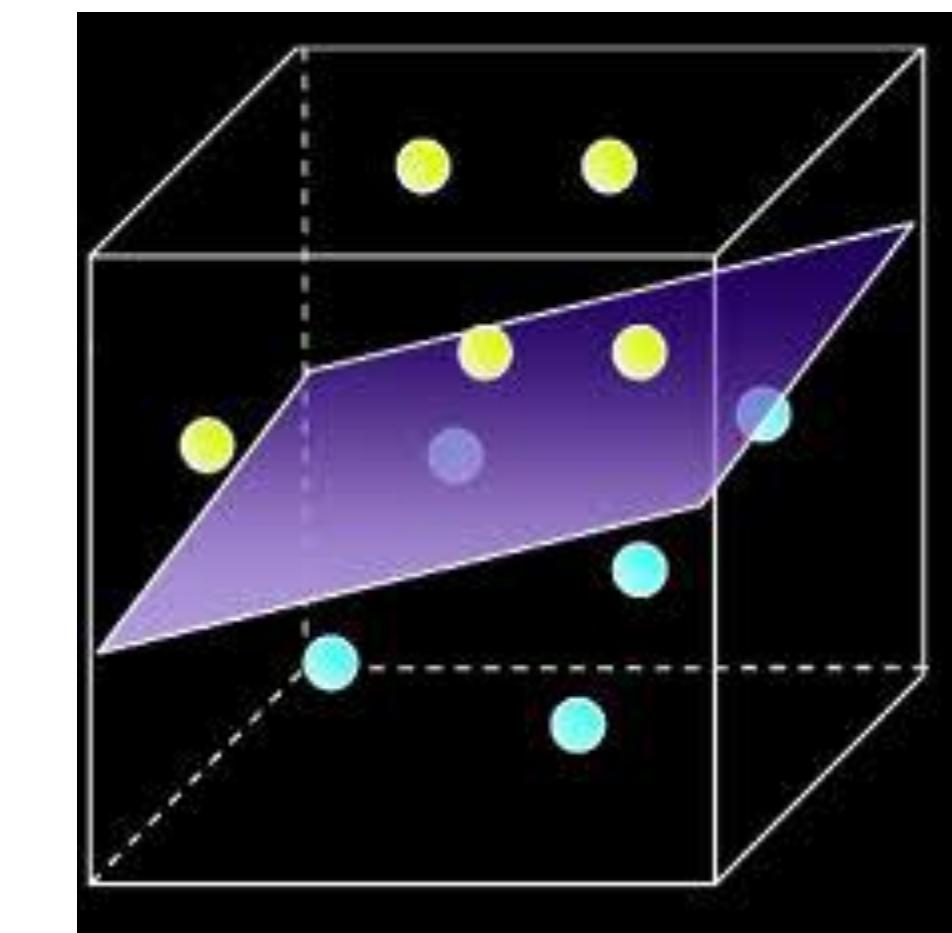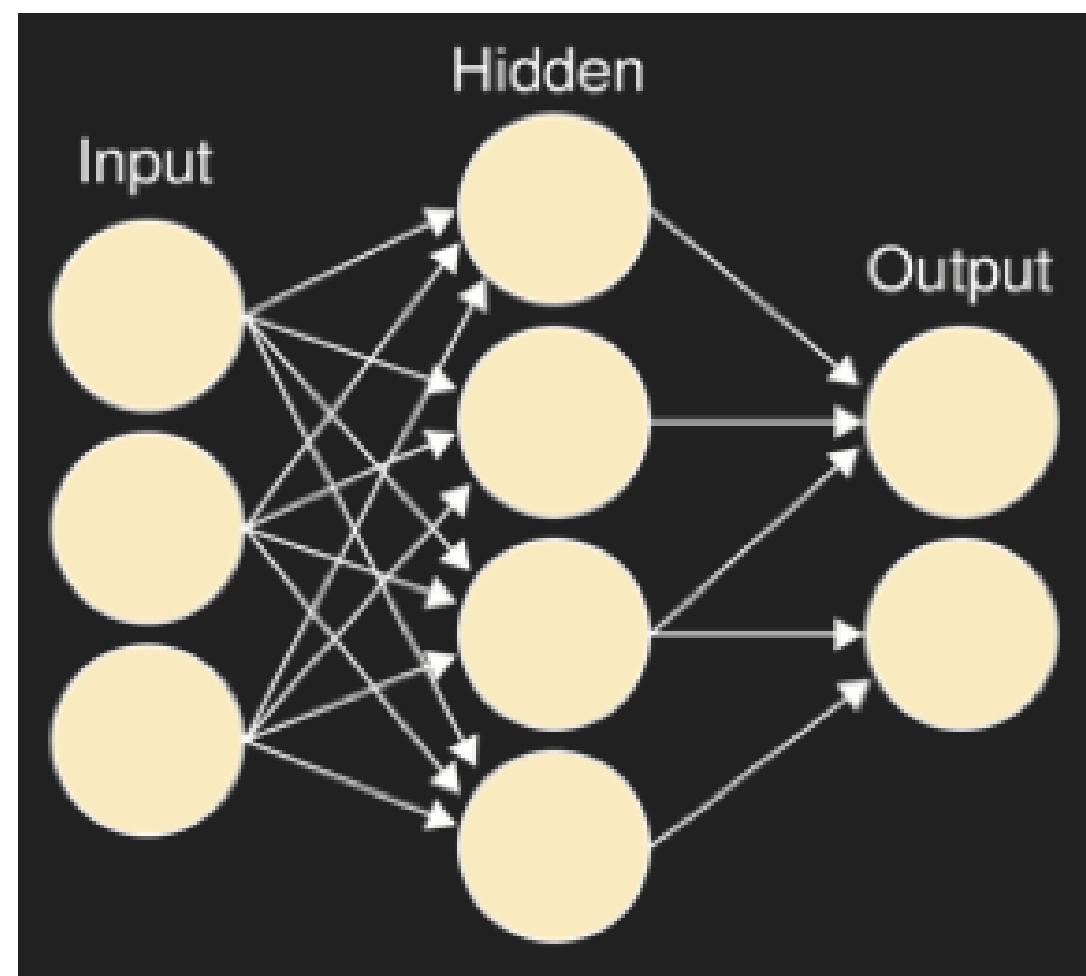
New Navy Device Learns by Doing
- The New York Times (July 8, 1958)

"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."

Neural Network Boom, 1960-79

Frank Rosenblatt
@ Cornell!

RETINA OF
S UNITS

A UNITS

OUTPUT SIGNAL
+1 OR -1

Fig. 2. Network organization of a typical elementary perceptron
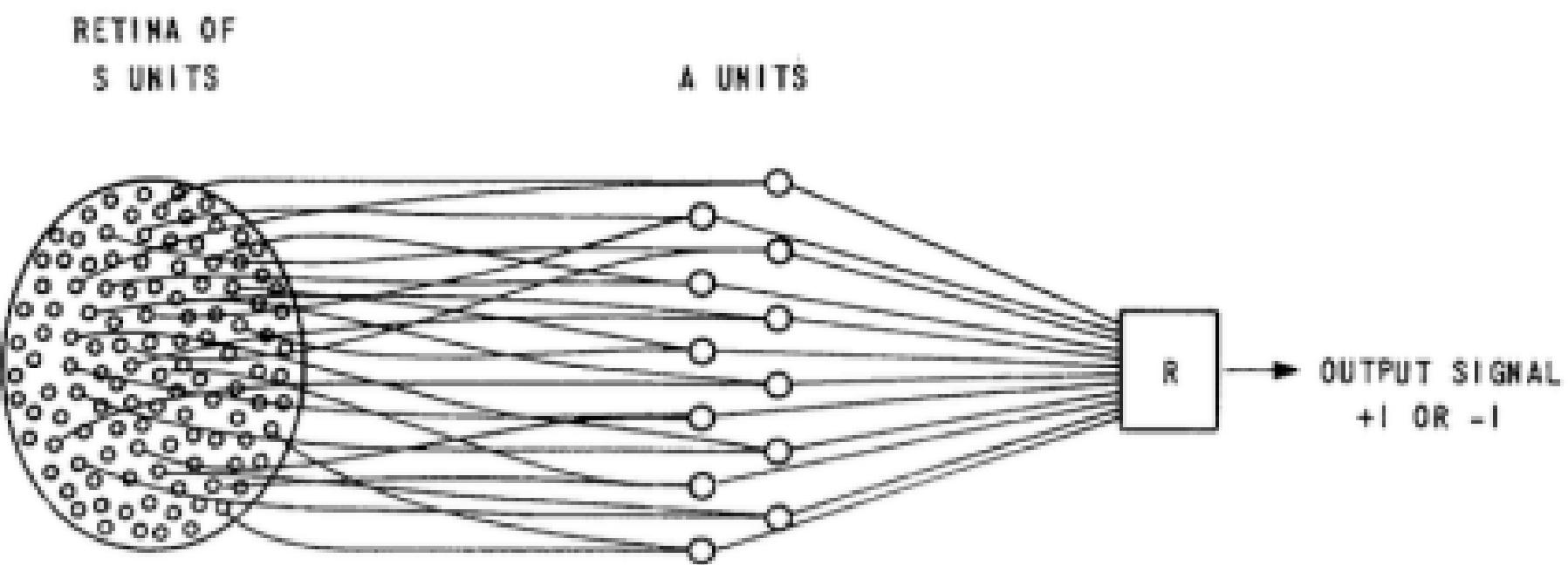
SEPTEMBER 1964

Rosenblatt—Analytic Techniques for the Study of Ne

- 1962 Rosenblatt invents Multi-Layer Perceptron (MLP) (fixed hidden layer)
- 1965 Ivakhnenko and Lapa introduce first Feed Forward Neural Net (FFNN)
- 1967 First FFN trained with SGD [Amari 1967]
- 1970 Modern back-propagation is introduced by Seppo Linnainmaa

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
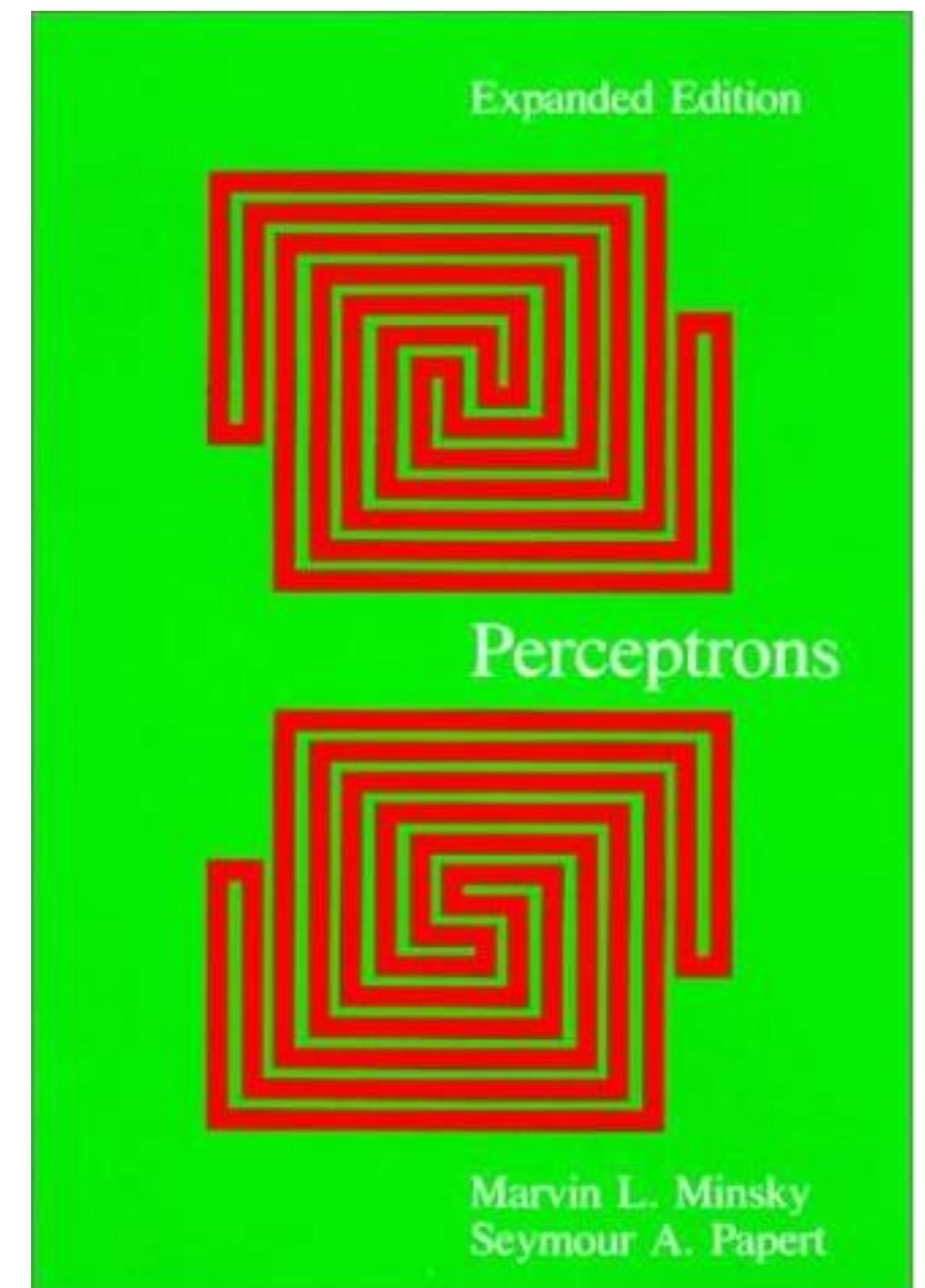
AI Boom then ML

- AI Boom: 1980-87
  - Search Algorithms (Eg. A*)
  - No Learning, Rule Based systems
- AI Winter (1987-93)

- Expectations overhyped

- AI Lull

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**

AI

ML
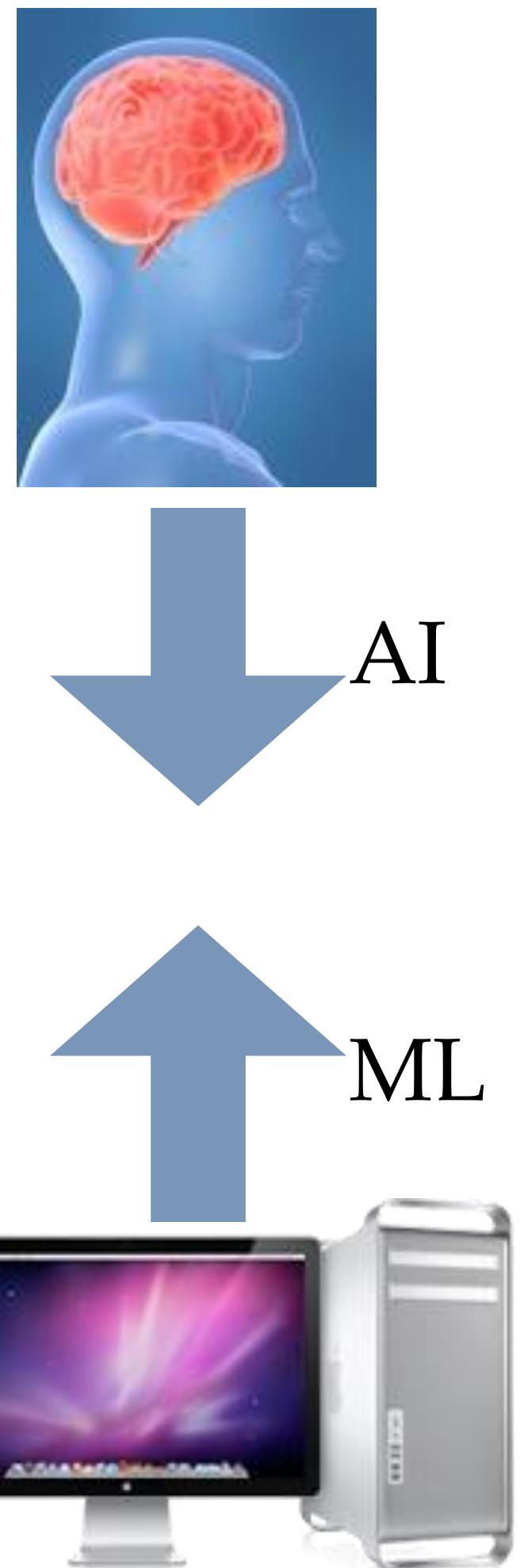
TD-Gammon (1994)

- 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.

$$\$\$\$$$
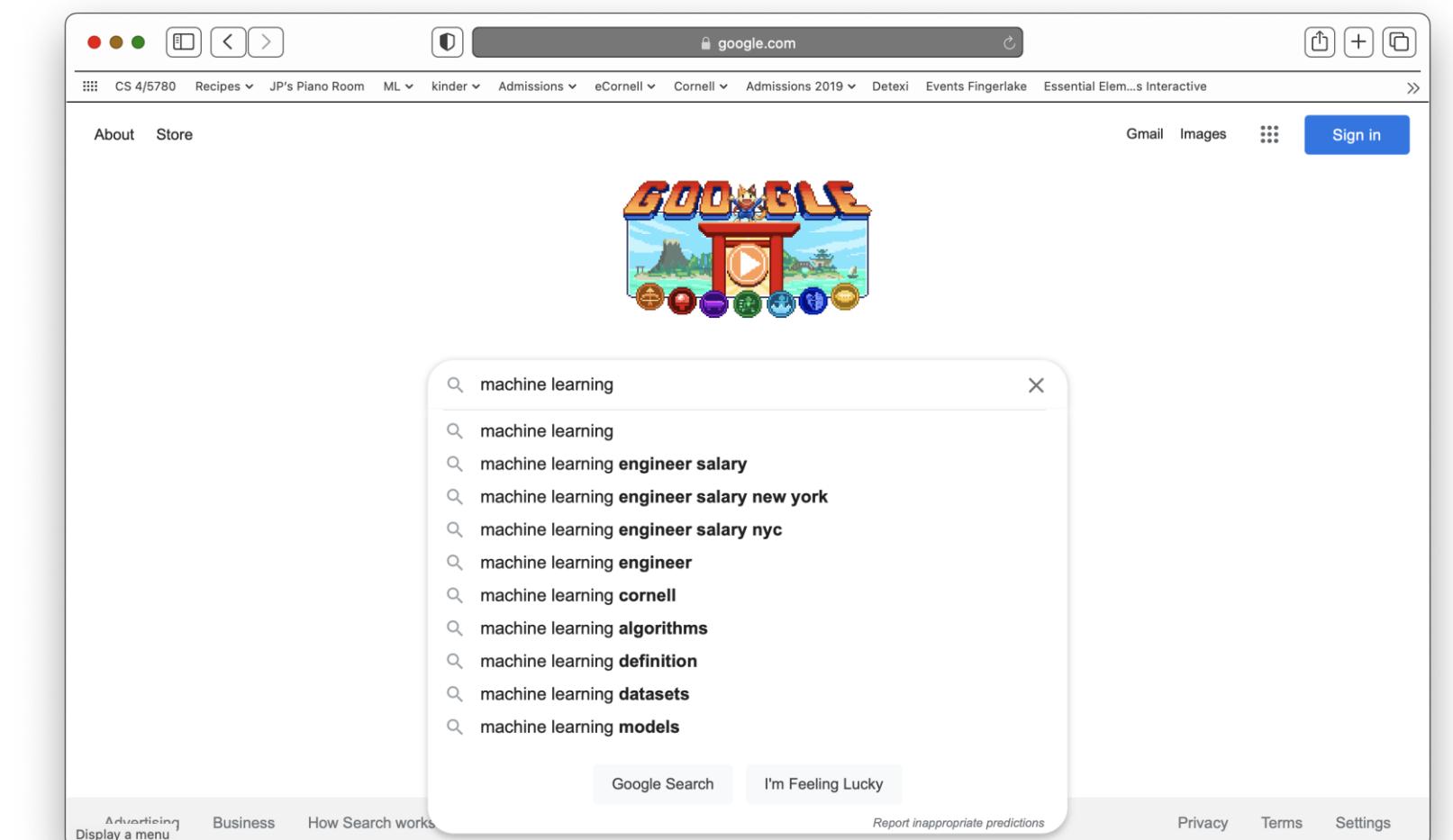
Machine learning became profitable!

2000-2010 The Convex Years

- Success of Support Vector Machines (SVM) and Convex Optimization
- Kernelization enabled non-linear classifiers that still use convexity
- Rigorous mathematical guarantees with empirical success
- Led to development of optimization algorithms that are workhorse of current models

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.
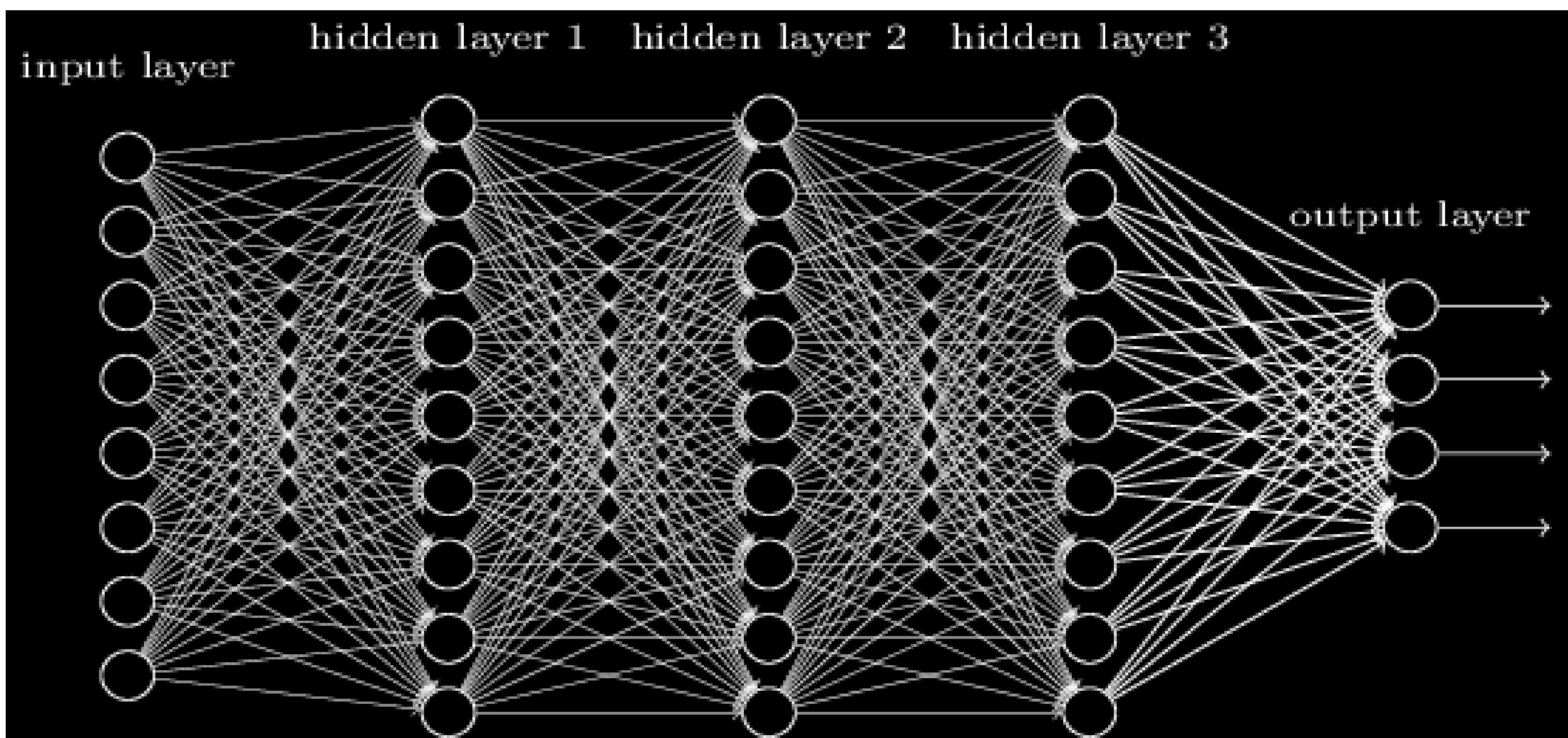
2016 Alpha Go

- 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"

2021 Generative Models become Mainstream

- 178 Billion Parameters
- $4.6 Million estimated cost
- 355 GPU Years of Training
- Trained on 45TB of text
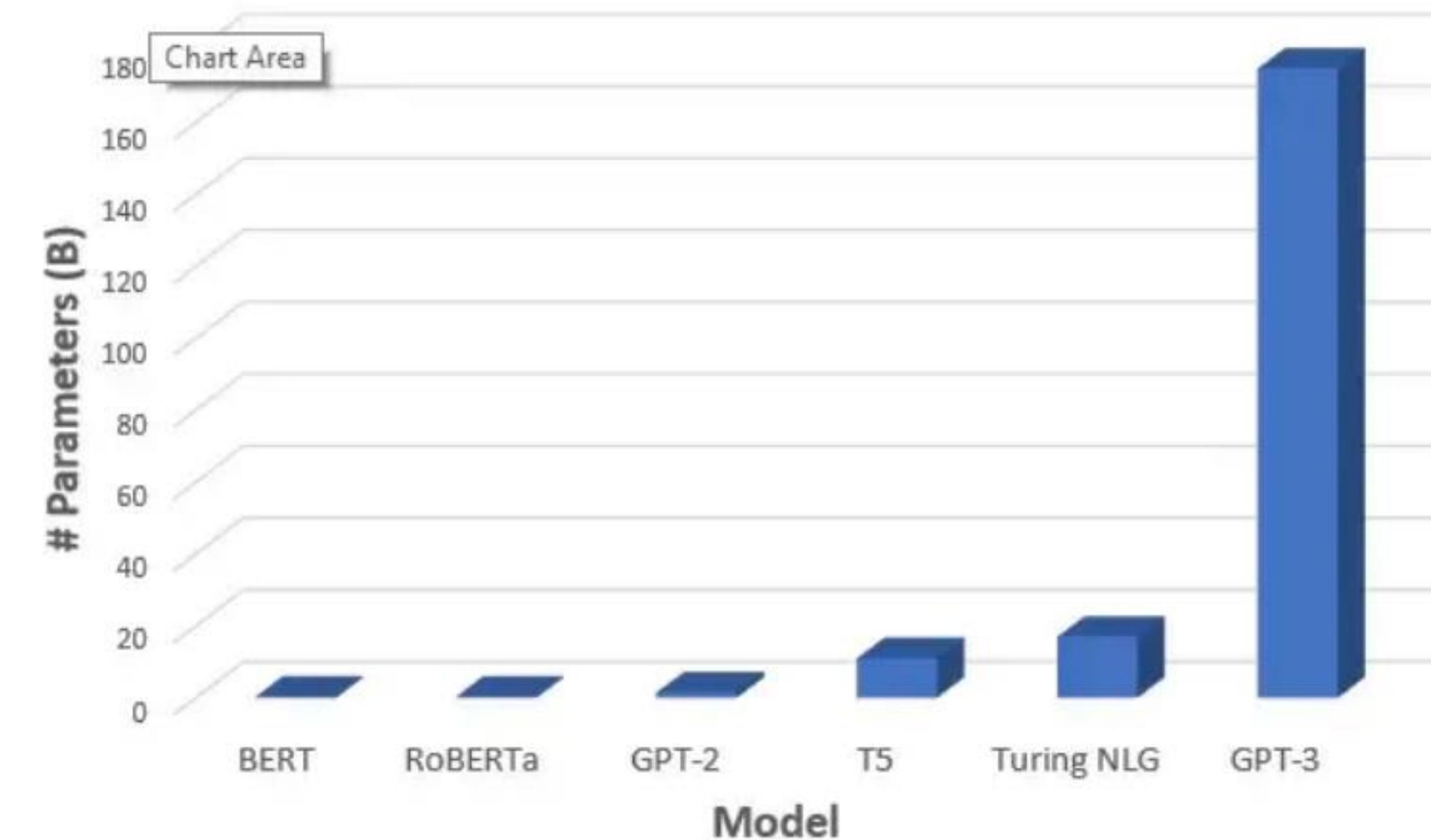
ML is Creative (well maybe we are still working on it)

ChatGPT ⌄

write a funny rap song on

Sure! What topic do you want the rap song to be about?

diffussion models

"Diffusion Beats (The AI Flow)"

[Intro]
Yo, it's your boy AI, spittin' machine-made bars,
Talkin' 'bout diffusion models takin' over like stars.
From noise to clarity, it's how we roll,
Let me break it down, control to goal.

[Verse 1]
Started with some noise, yeah, a cloud of confusion,
Pixel party wild, that's the initial diffusion.
Addin' noise forward, things get outta sync,
But then we reverse it—watch the data rethink.

Models spinnin' like a DJ, crafting the beat,
From static to masterpiece, can't be beat.
GANs tried their shot, but we took the crown
Diffusion's the king, wearin' that AI gown.

↓

And Versatile

make a funny picture that describes machine learning as programs that improve with experience
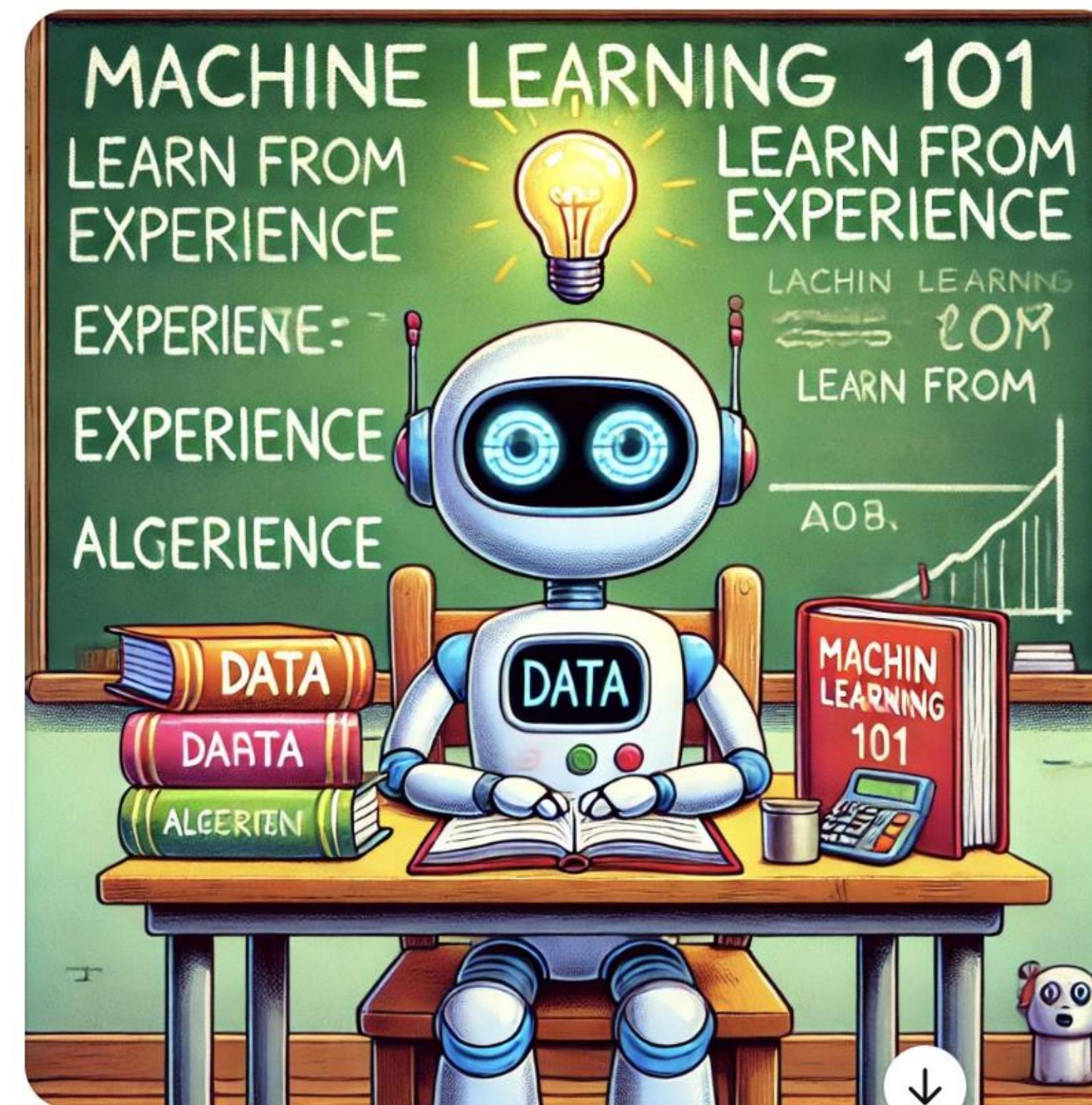
& 2024...

Two 2024 Nobel Prizes Honor Researchers Innovating in AI, Machine Learning

October 21, 2024

Recognizing the transformative power of AI, the Royal Swedish Academy of Sciences recently awarded the 2024 Nobel Prizes in physics and chemistry to researchers at the forefront of machine learning, artificial neural networks, computational design, and protein structure prediction.

John Hopfield and Geoffrey Hinton earned the **Nobel Prize in physics** for "foundational discoveries and inventions that enable machine learning with artificial neural networks." David Baker, Demis Hassabis, and John Jumper received the **Nobel Prize in chemistry** for advances in computational protein design and protein structure prediction.

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)

Today We're Surrounded by Machine Learning
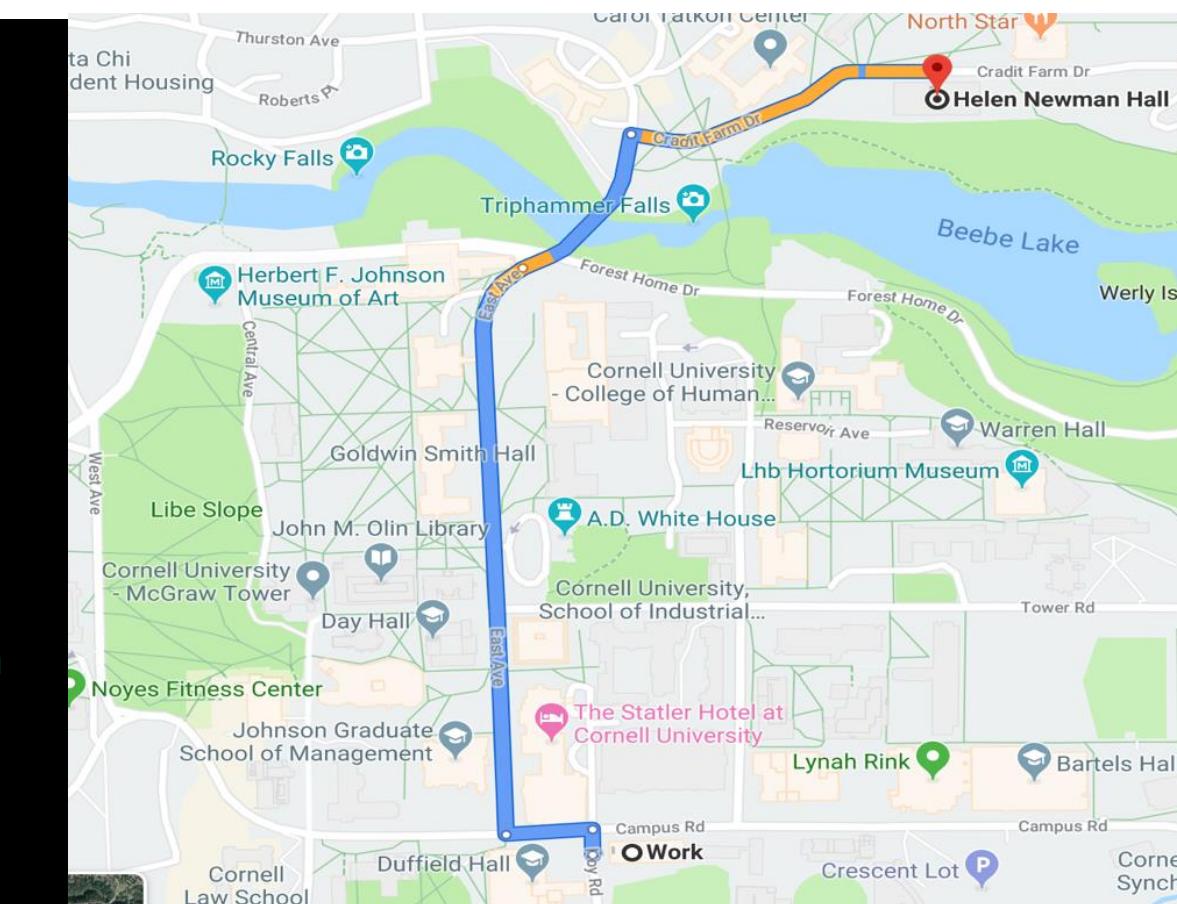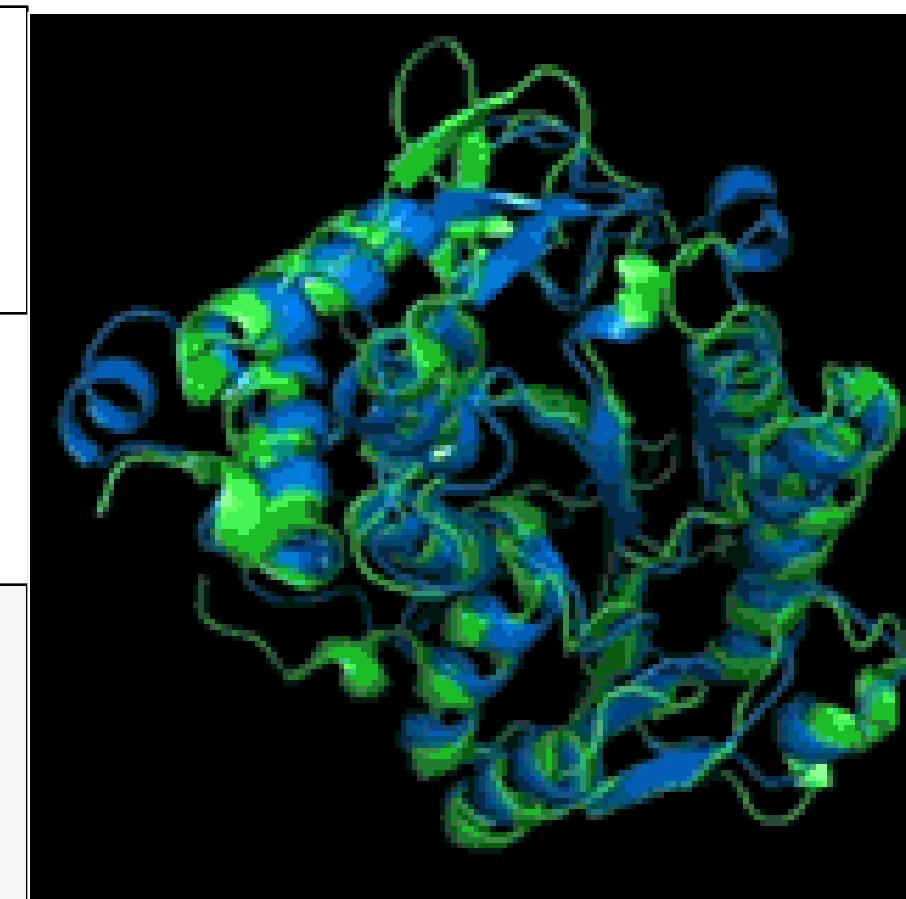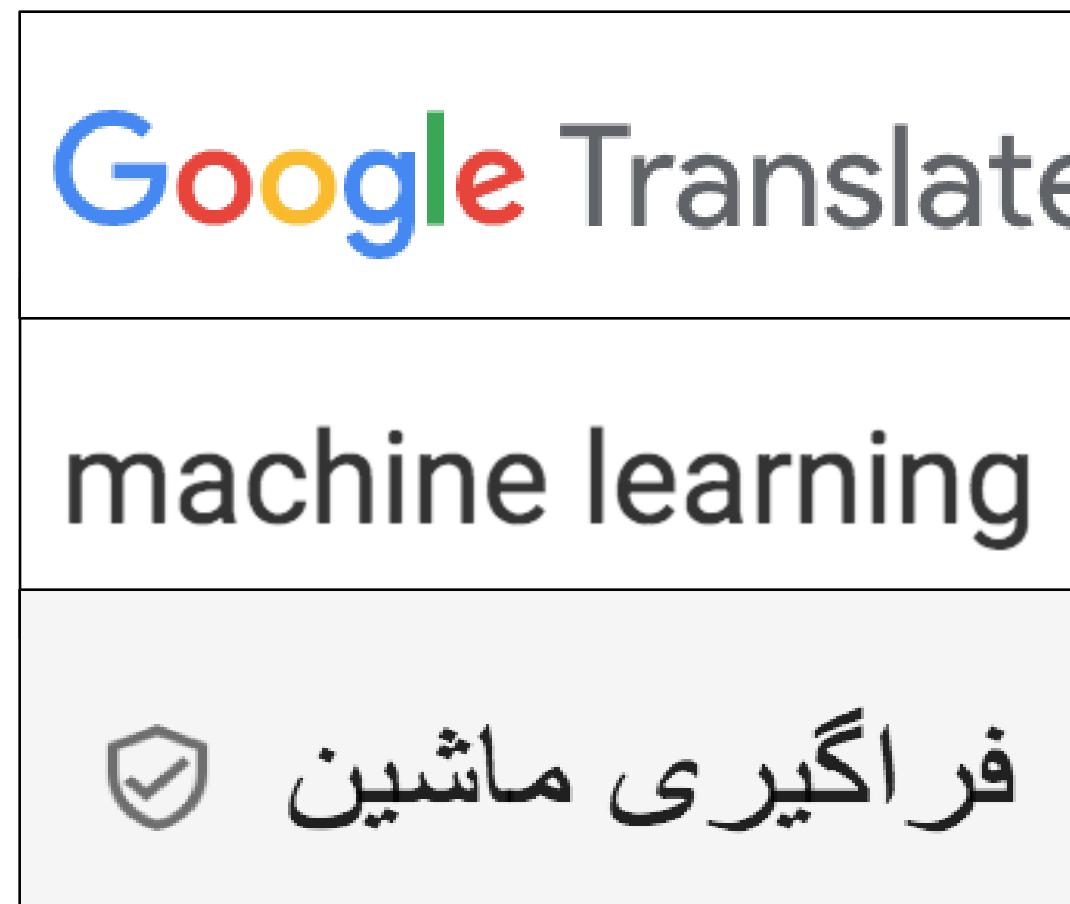
DANGER

Data privacy / misuse

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

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

Leaked data   Real image
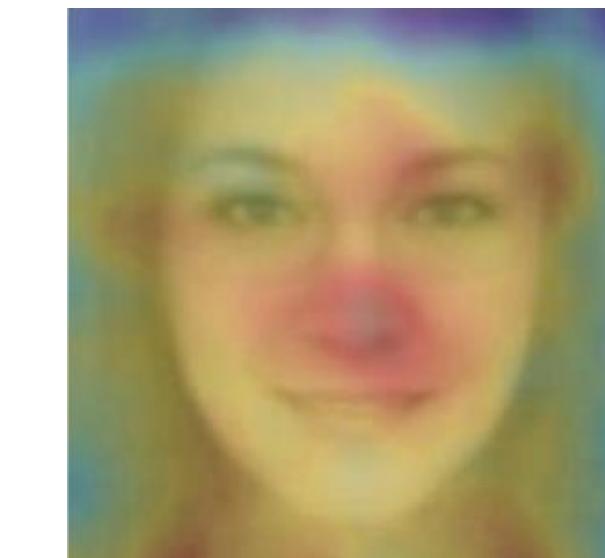
Formal definitions of data privacy:
- **K- anonymity (Sweeney)**
- **Differential Privacy (Dwork**

Latanya Sweeney   Cynthia Dwork   Frank McSherry   Kobbi Nissim   Adam Smith
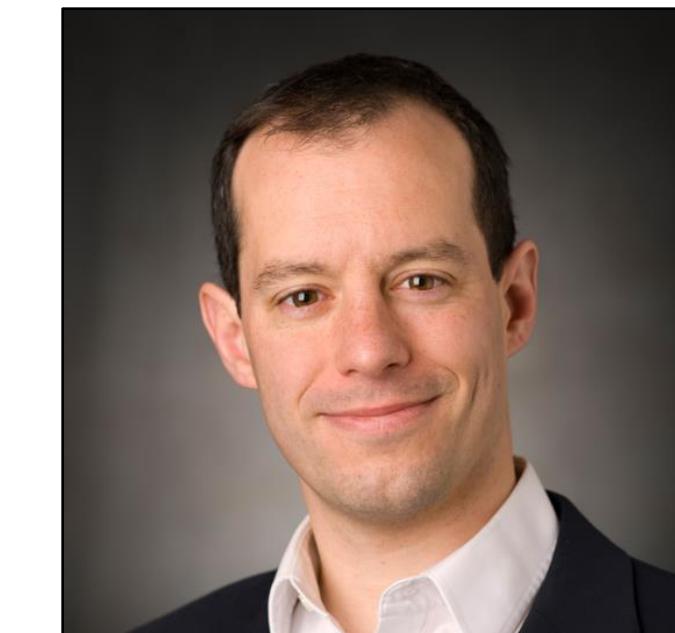
Robust and Secure ML

Image Recognition
Misreading traffic signs
(Eykholt et al)

Speech recognition
Hide commands in noise (Carlini & Wagner)
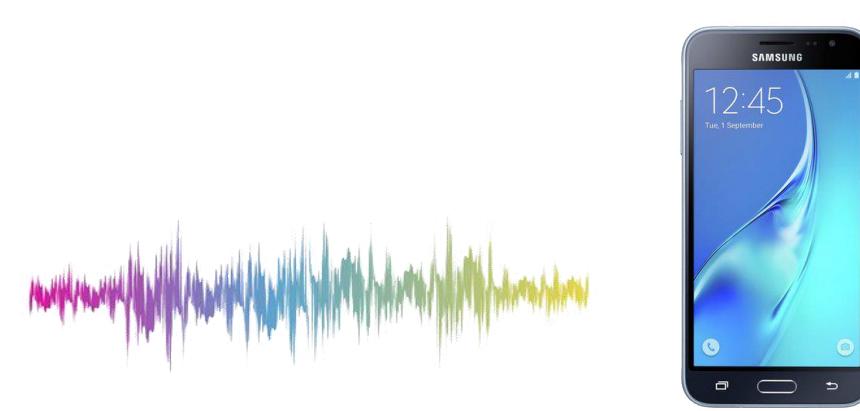
Poisoning Attacks
Tay (chat bot) became inflammatory in 16 hr.
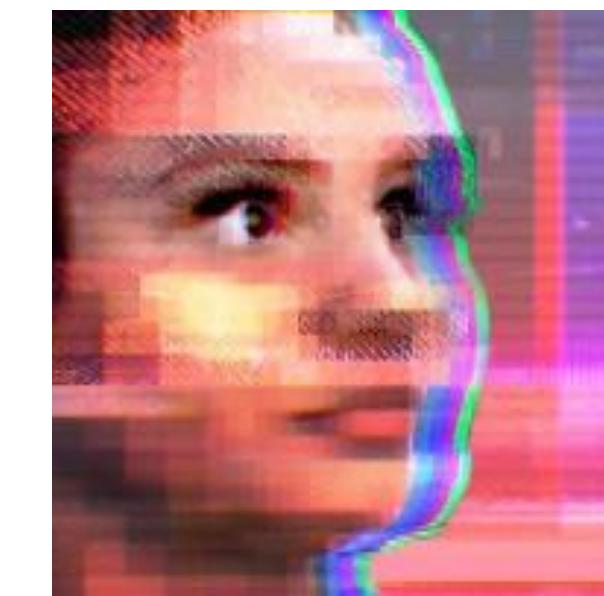
How to create robust and secure machine learning algorithms?

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?

When an Algorithm Helps Send You to Prison

By Ellora Thadaney Israni

ML in Society

- New technology that can transform society can also overwhelm it

  - Steam engine late 1800s (lead to sweat shops, exploitation)

  - Cars: out-of-control traffic accidents until 1970s
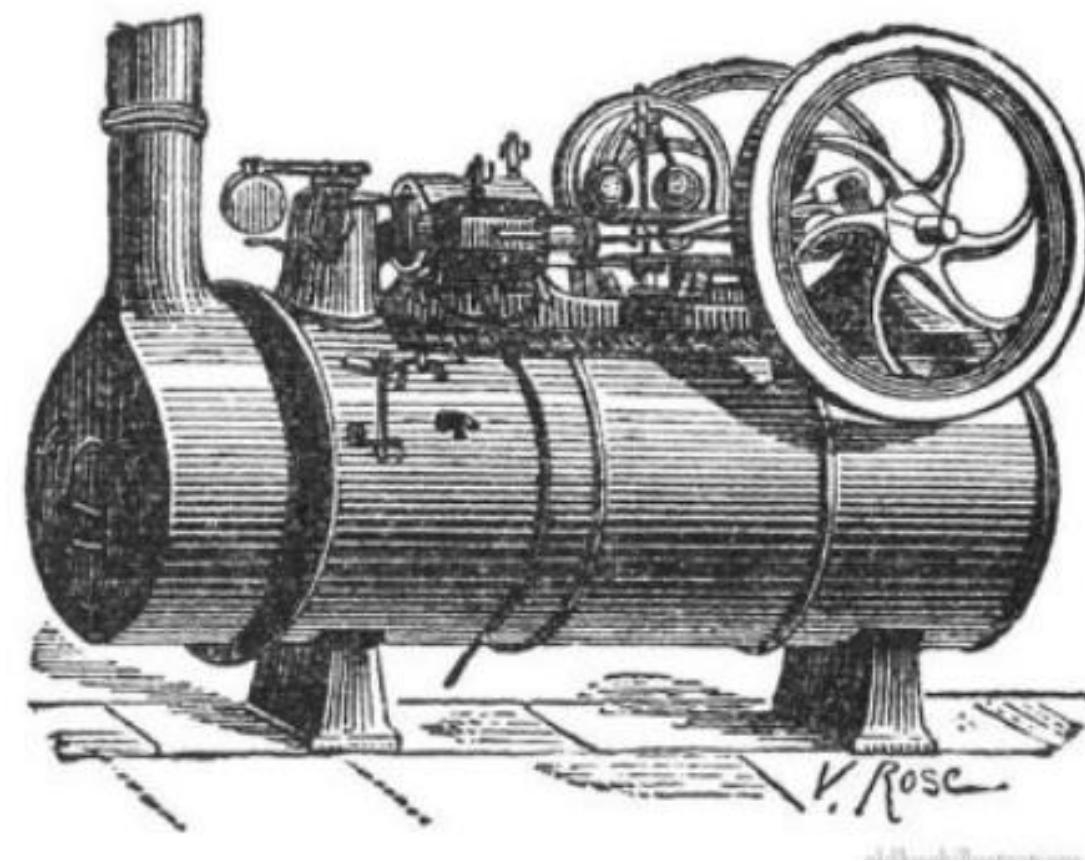
To-do action items

- Pass Placement Exam!

- Find study group

- Autograder Setup:
  - Find Teammate
  - Start Project -1
  - Start Project 0