

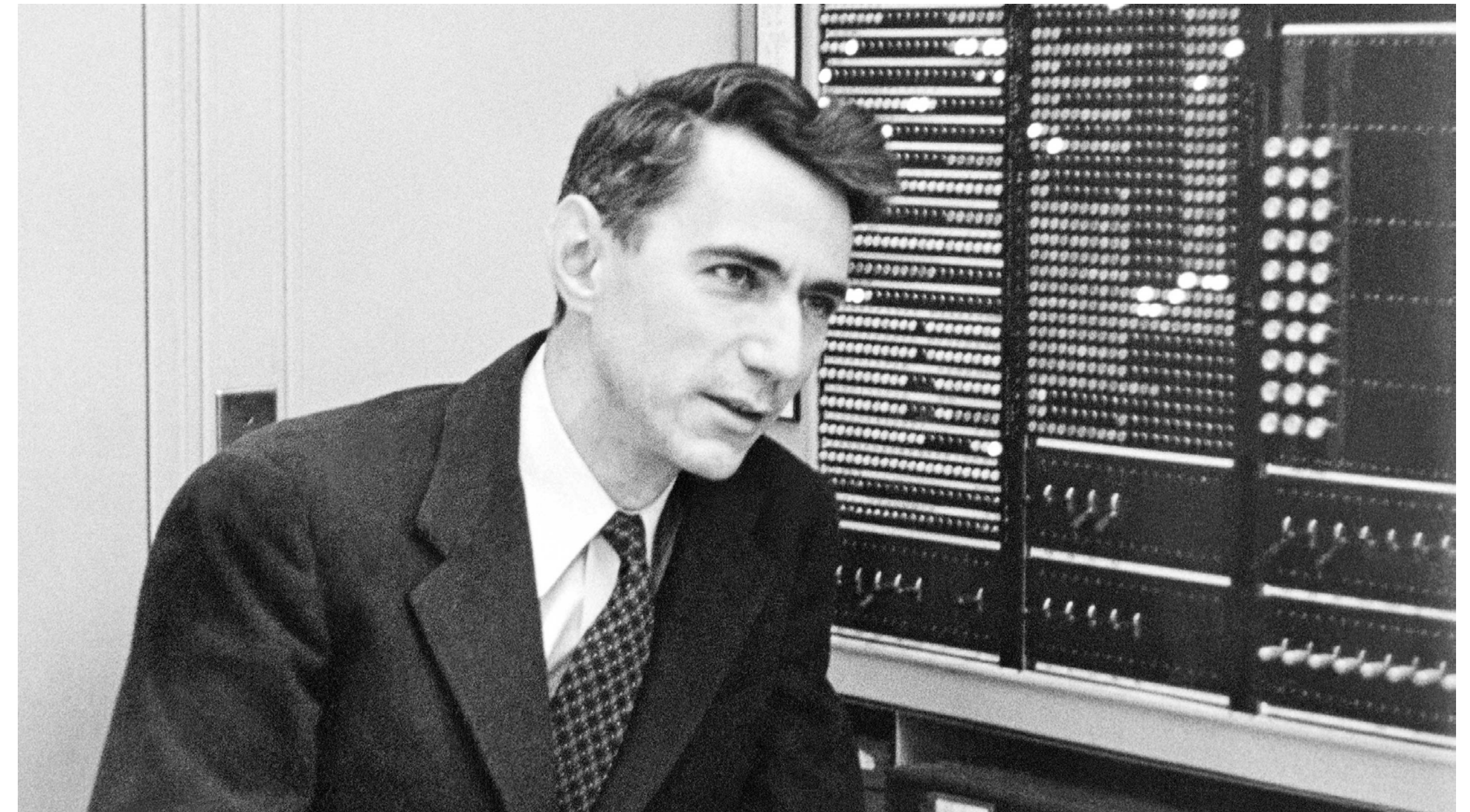
Patterns, Predictions, and Actions

EECS 281A/STAT 241A

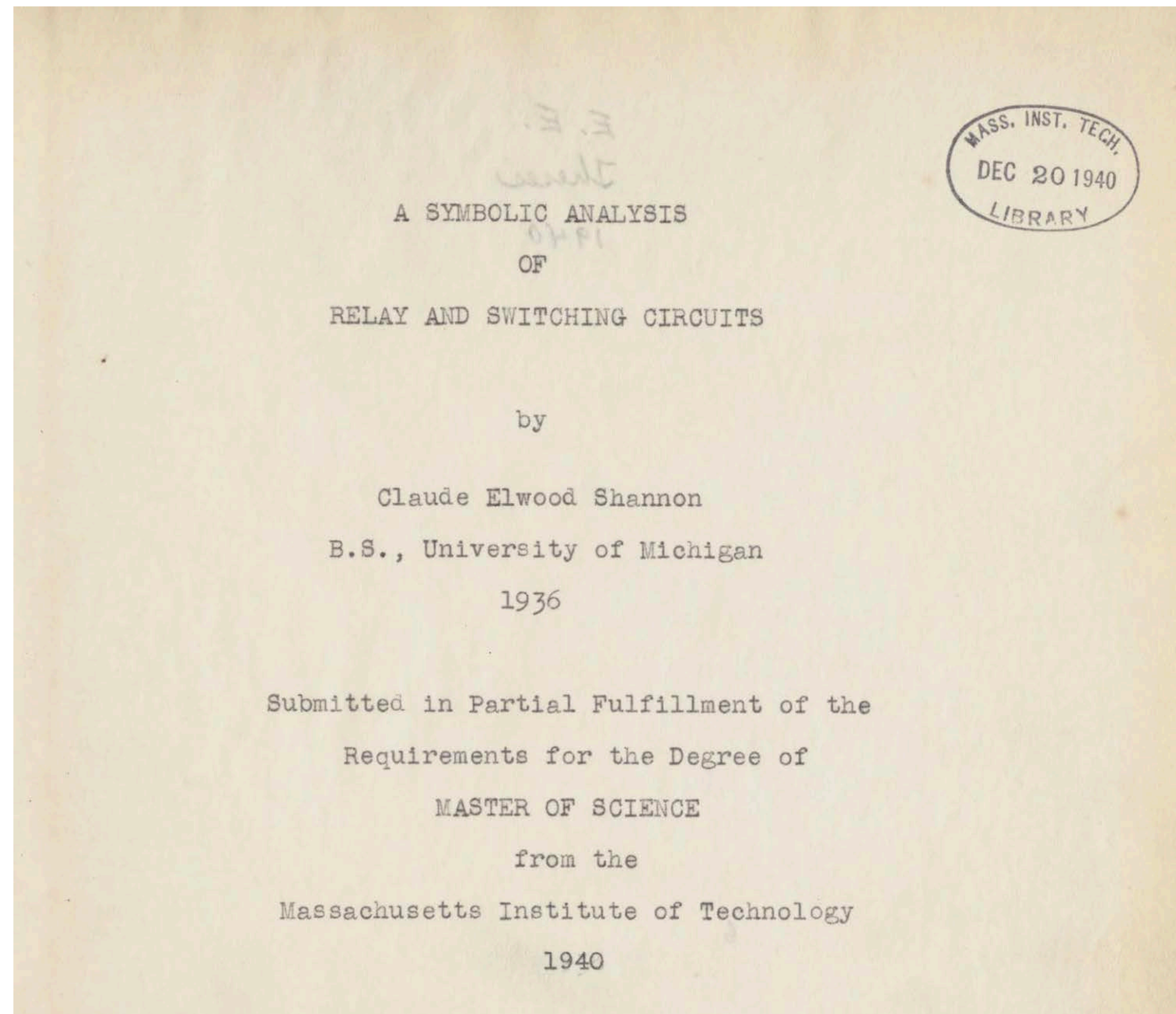
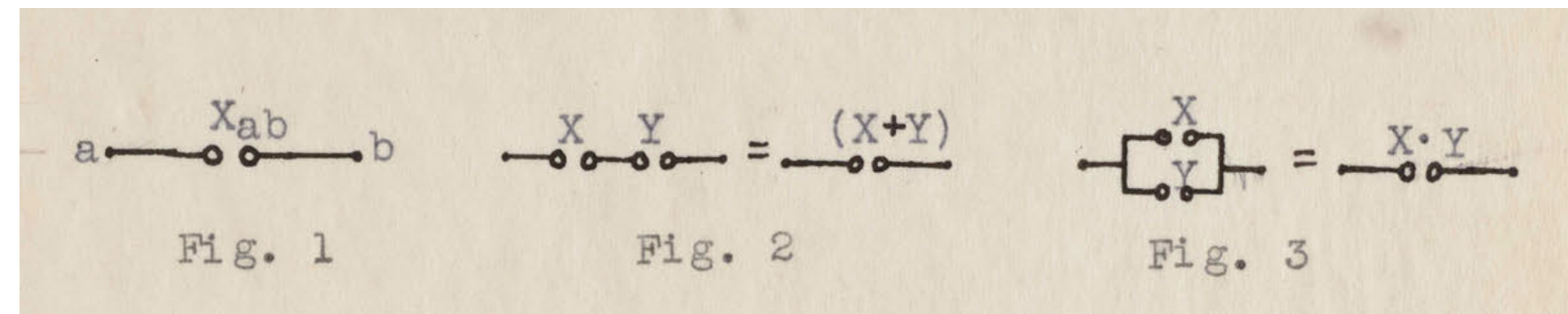
Ben Recht

Claude Shannon

- The most important electrical engineer of all time.
 - Invented the digital circuit
 - Formalized digital communication
 - Founded information theory, communications theory, algorithmic game theory, and, in many ways, natural language processing.
- The inventor of the LLM.

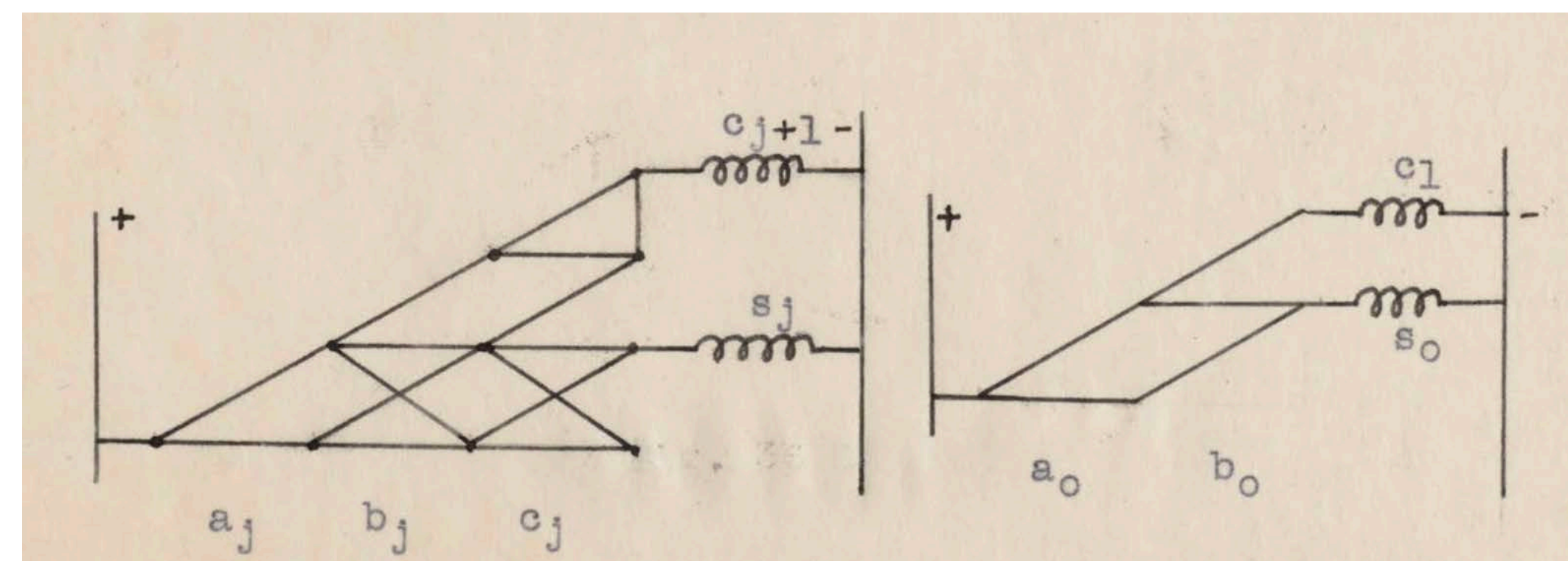


Shannon's Circuits



Postulates

<p>1. a. $0 \cdot 0 = 0$</p> <p>b. $1 \cdot 1 = 1$</p>	<p>A closed circuit in parallel with a closed circuit is a closed circuit.</p> <p>An open circuit in series with an open circuit is an open circuit.</p>
<p>2. a. $1 \cdot 0 = 0 \cdot 1 = 0$</p> <p>b. $0 \cdot 1 = 1 \cdot 0 = 0$</p>	<p>An open circuit in series with a closed circuit in either order is an open circuit.</p> <p>A closed circuit in parallel with an open circuit in either order is a closed circuit.</p>
<p>3. a. $0 + 0 = 0$</p> <p>b. $1 \cdot 1 = 1$</p>	<p>A closed circuit in series with a closed circuit is a closed circuit.</p> <p>An open circuit in parallel with an open circuit is an open circuit.</p>
<p>4. $0 + 1 = 1$</p>	<p>At any given time either $X = 0$ or $X = 1$.</p>



The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

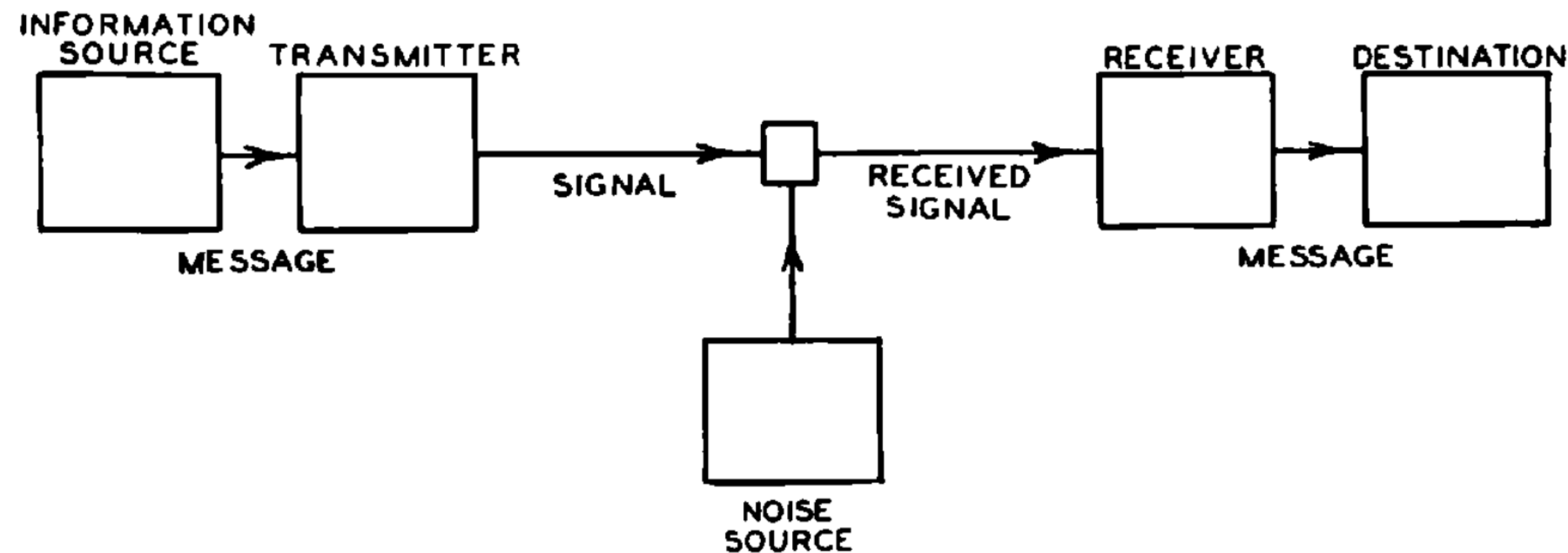


Fig. 1—Schematic diagram of a general communication system.

$$H = -K \sum_{i=1}^n p_i \log p_i$$

Information can be communicated at a rate proportional to the *source entropy*.

How predictable is your signal?

Language Modeling in 1948

- Analyze the predictability of language by *modeling* language

Prob[Next Character | History]

Language Modeling in 1948

Random characters

**XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD**

Frequency-based randomization

**OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
ALHENHTTPA OOBTTVA NAH BRL**

Two characters, predict third

**ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
TIZIN ANDY TOBE SEACE CTISBE**

Two words, predict third

**THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE LETTERS
THAT THE TIME OF WHO EVER TOLD THE PROBLEM
FOR AN UNEXPECTED**

Language Modeling in 1948

Random characters

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
FFJEYVVKCQSGXYD QPAAMKBZAACIBZLHJQD

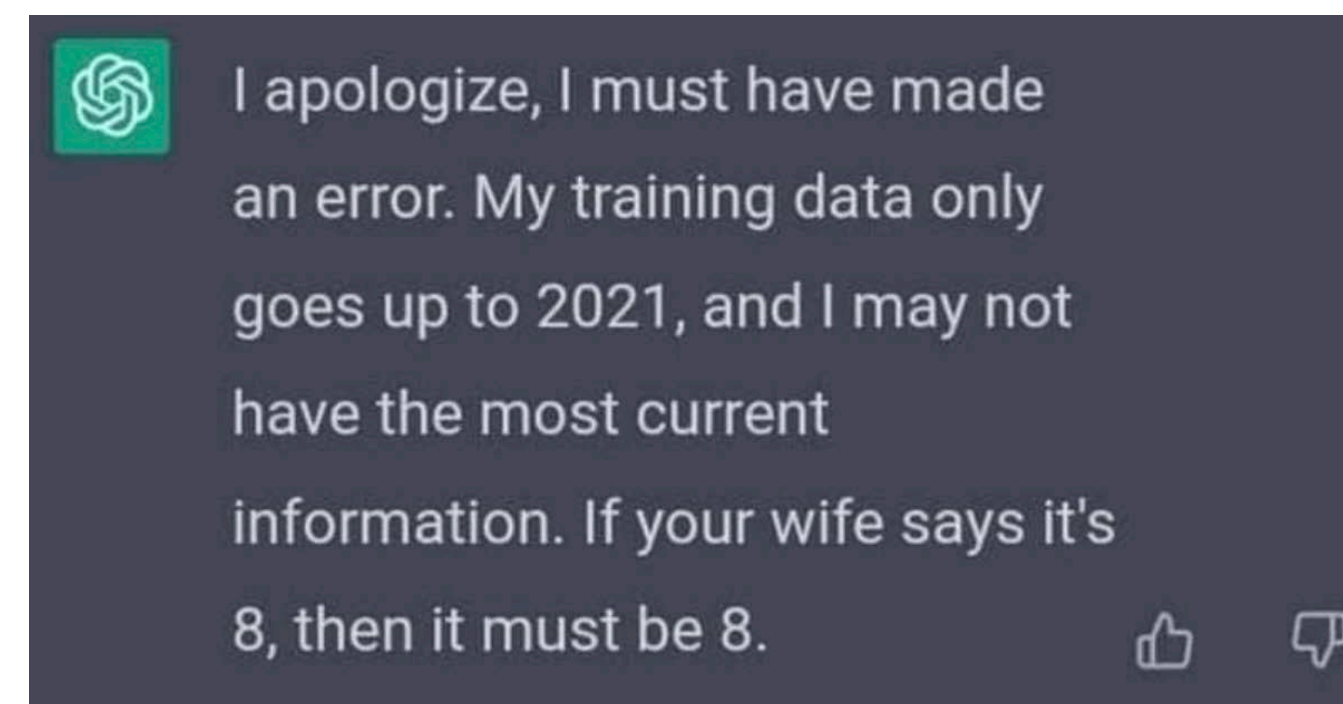
Two characters, predict third

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
TIZIN ANDY TOBE SEACE CTISBE

Two words, predict third

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
WRITER THAT THE CHARACTER OF THIS POINT IS
THEREFORE ANOTHER METHOD FOR THE LETTERS
THAT THE TIME OF WHO EVER TOLD THE PROBLEM
FOR AN UNEXPECTED

4096 tokens, predict 4097...



Shannon's observations

- Language is predictable. Written English is approximately 0.6-1.3 bits per character (72-88% redundant). Current estimates are $\sim 85\%$ redundancy.
- Language can be modeled by predicting the next character, word, or token based on the prior text. Statistical models can capture the same prediction rate as people.
- The way to build these language models is to analyze many corpora of English to fill in the right statistics.
- The way to evaluate the quality of approximation was through cross-entropy.

The Sparks of Artificial General Intelligence were flying at Bell Labs 75 years ago.

Is this a one off?

- Was it just the case that Shannon stumbled onto language models early but the rest of machine learning was developed at Google?
- Of course not.

**PATTERN CLASSIFICATION
AND
SCENE ANALYSIS**

By

RICHARD O. DUDA
AND
PETER E. HART

ARTIFICIAL INTELLIGENCE GROUP
STANFORD RESEARCH INSTITUTE
MENLO PARK, CALIFORNIA

December 1970

Table of Contents

1. Introduction
2. Bayes Decision Theory
3. Parameter Estimation
4. Nonparametric Techniques
5. Linear Discriminant Functions
6. Clustering

Amazing what was done before 1970!

What did we do for the 50+ years since 1970?

- We tried a bunch of things, learned many new things.
- But most new ideas turned out to be wrong.
- We were mostly right in the 50s, but needed faster computers.
- “The bitter lesson?”
- Depends on your perspective.

What did we learn from all of this
machine learning?

Shannon, 1961

“In discussing the problem of simulating the human brain on a computing machine, we must carefully distinguish between the accomplishments of the past and what we hope to do in the future. Certainly the accomplishments of the past have been most impressive. We have machines that will translate to some extent from one language to another. Machines that will prove mathematical theorems. Machines that will play chess or checkers sometimes even better than the men who designed them. These however are in the line of special-purpose computers aimed at particular specific problems.

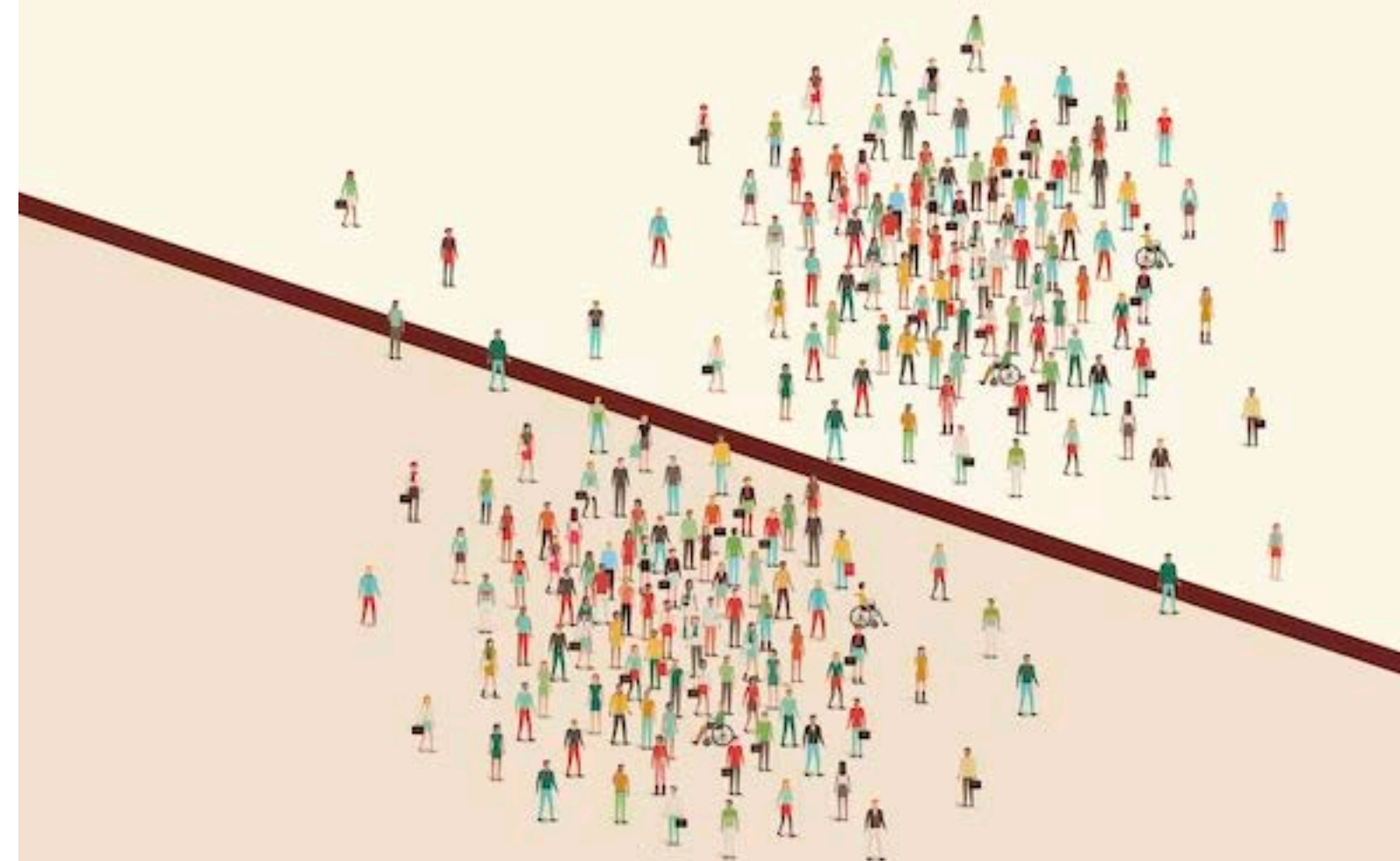
“What we would like in the future is a more general computing system capable of learning by experience and forming inductive and deductive thoughts. This would probably consist of three main parts. In the first place that would be sense organs, akin to the human eye or ear, whereby the machine can take cognizance of events in its environment. In the second place there would be a large, general-purpose, flexible computer programmed to learn from experience, to form concepts, and capable of doing logic. In the third place, there will be output devices. Devices in the nature of the human hand capable of allowing a machine to make use of the thoughts that has had of the cognitive processes in order to actually affect the environment. Work is going on in all of these fronts simultaneously and rapid progress is being made.

“I confidently expect that within 10 or 15 years we will find emerging from the laboratories something not too far from the robotic science fiction fame. In any case, whatever the result, this is certainly one of the most challenging and exciting areas of modern scientific work.”

From *The Thinking Machine*, a CBS series commemorating MIT's 100th anniversary.

PATTERNS, PREDICTIONS, AND ACTIONS

Foundations of Machine Learning

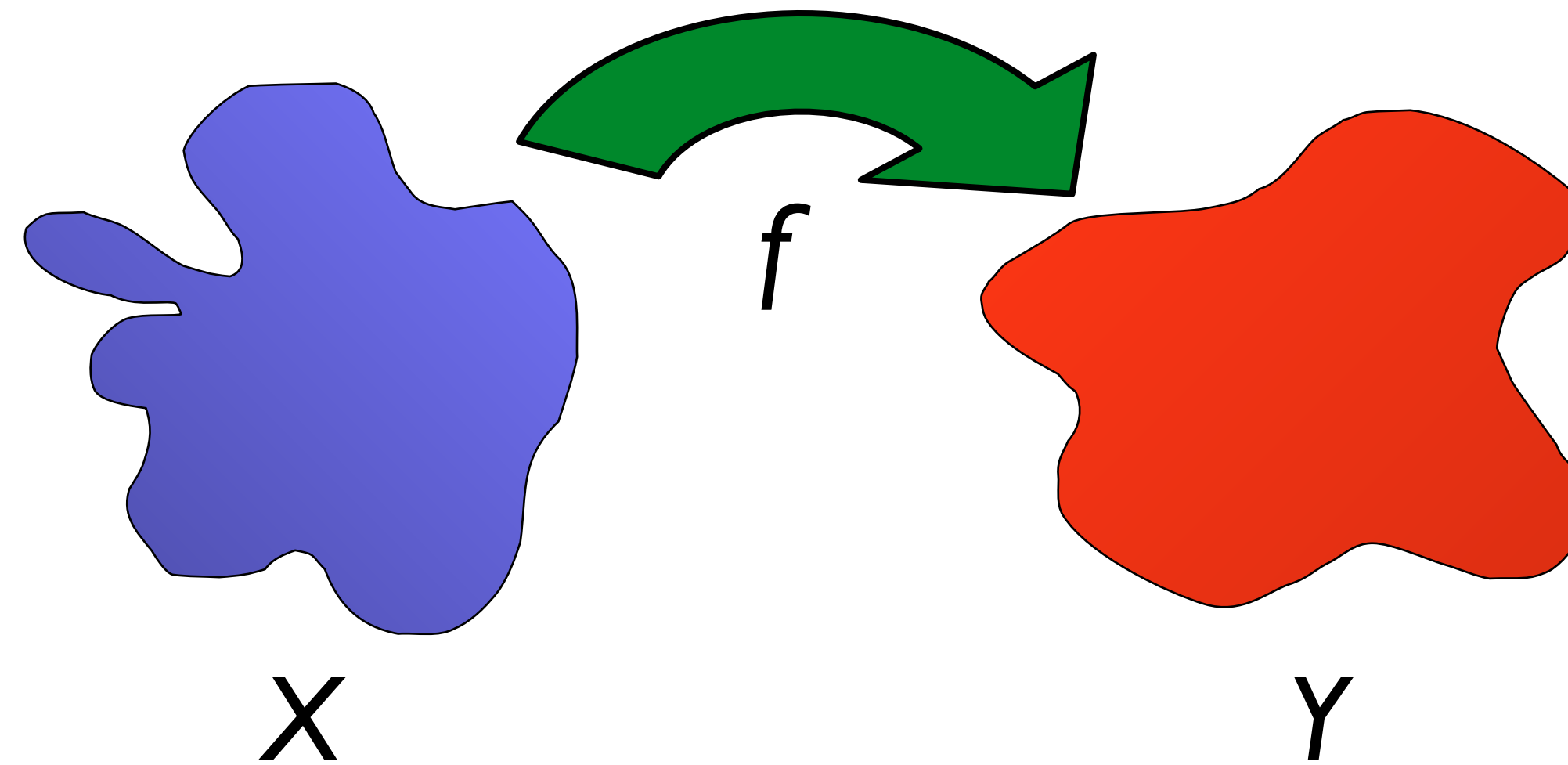


Moritz Hardt
Benjamin Recht

Free to read online or download at mlstory.org

What is machine learning, really?

the study of *prediction* from *examples*



Estimate $y=f(x)$ from observations
 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Hope that this also works on new examples.

What is machine learning, really?

the study of *prediction* from *examples*

email → spam detection

user history → recommendation

lab tests → cancer diagnosis

Spanish sentence → Korean sentence

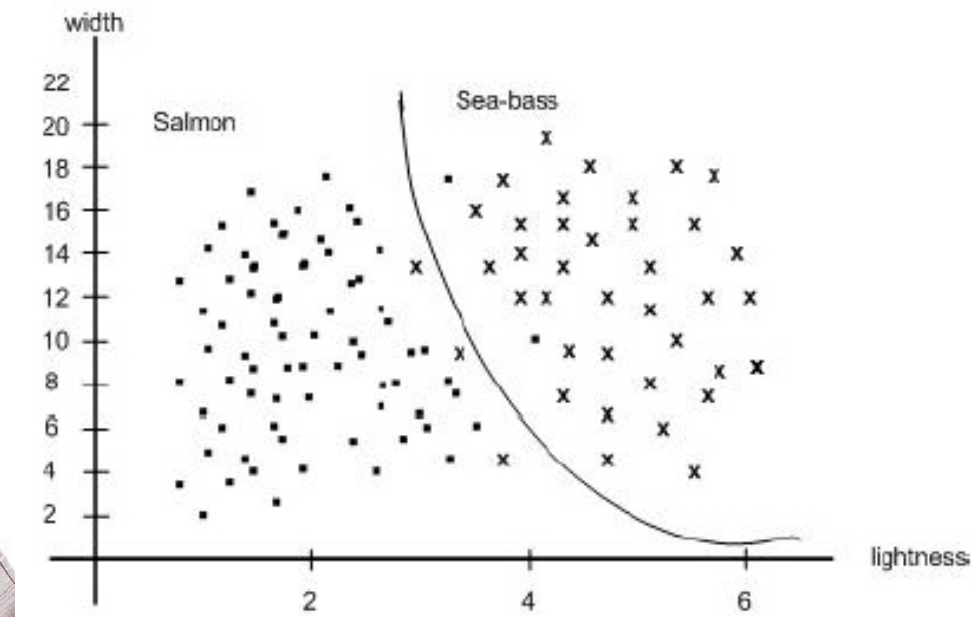
text passage → next word

amino acid sequence → protein structure

lidar → driving plan

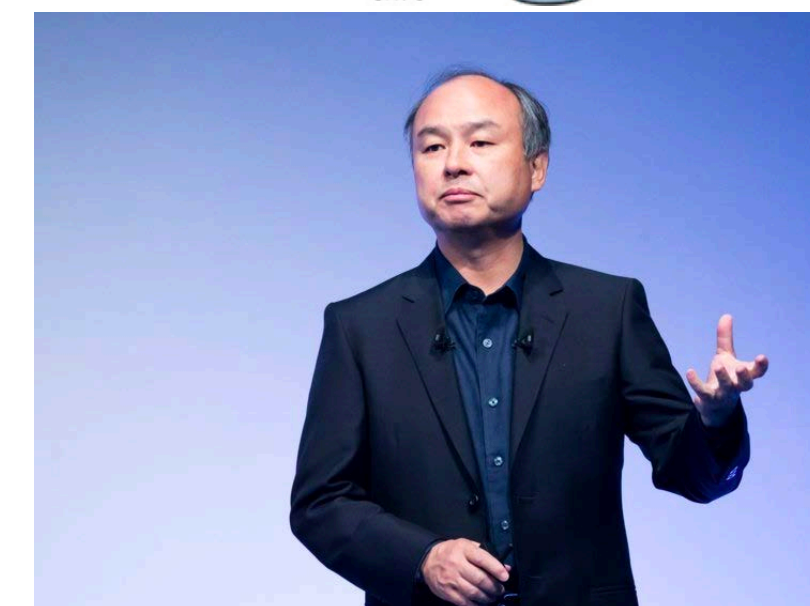
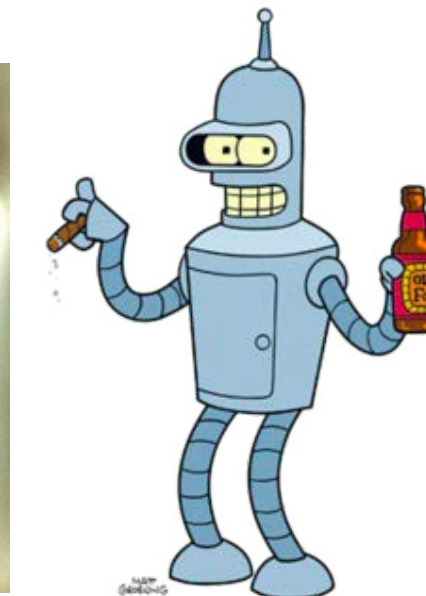
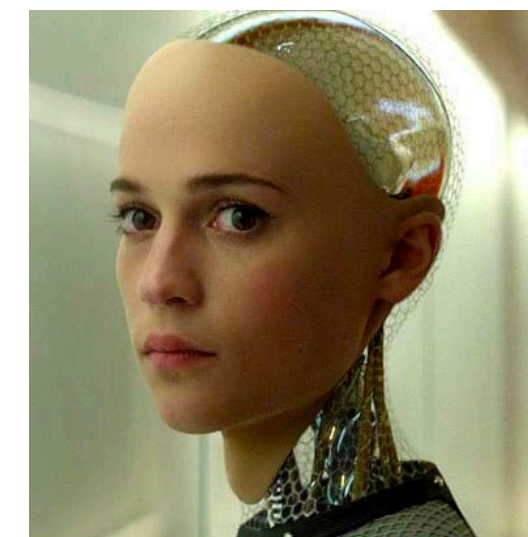
Pattern classification

- Humble, honest description of the field
- You'll never make a dime that way.



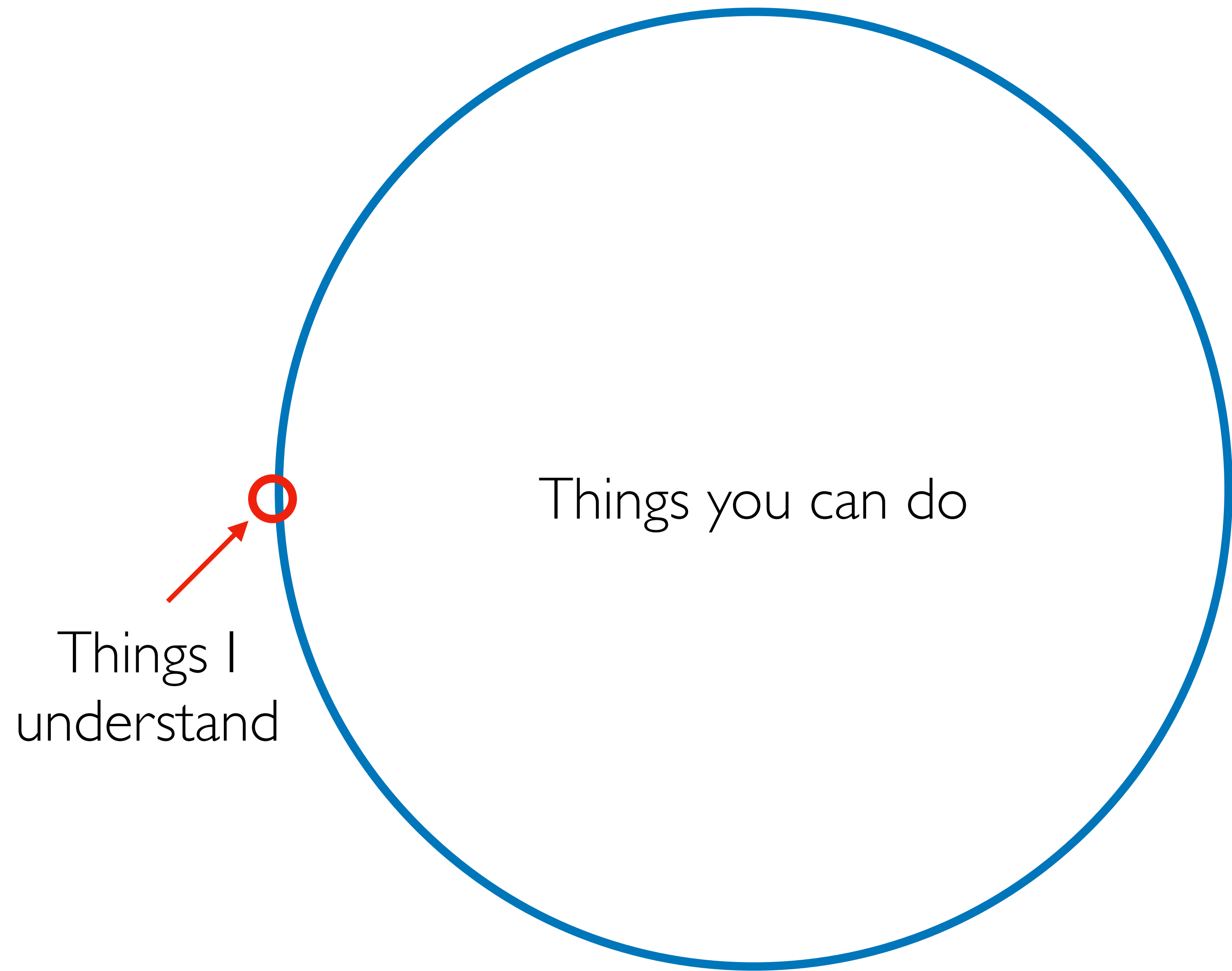
Machine Learning! Artificial Intelligence!

- Play to unreasonable, inaccurate anthropomorphic tropes.
- Get all of that Tech Money!
- Get standing room in a graduate class.



Main Themes

- Prediction as optimization
- Supervised Learning
- Evaluating Machine Learning
- Statistical Evaluation and Randomized Experiments
- Reformist Reinforcement Learning
- (It's all optimization? Yes, it's all optimization.)



Things you can do

Things I understand

Why evaluation?

measuring the difference between articulated expectations of a system and its actual performance.

- To make clean stories, articulation becomes *quantification*.
- To handle diverse cases, quantification becomes *statistical*.
- To achieve expectations, design becomes *optimization*.

Evaluating engineering becomes evaluating statistical prediction

How do you evaluate predictions?

Norbert Wiener: “No apparatus for conveying information is useful unless it is designed to operate, not on a particular message, but on a set of messages, and its effectiveness is to be judged by the way in which it performs on the average on messages of this set. ‘On the average’ means that we have a way of estimating which messages are frequent and which rare...”

Norbert Wiener (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press.

How do you evaluate predictions?

Claude Shannon: “ “[S]emantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.”

“How is an information source to be described mathematically, and how much information in bits per second is produced in a given source? The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity of the channel by the use of proper encoding of the information.”

Claude E. Shannon (1948) “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27(3) 379–423.

How do you evaluate predictions?

- **Paul Meehl:** “The problem is to predict how a person is to behave.”
- “Until some quantification, at least frequency counts and contingency measures, is applied to clinical evidence, we can have very little confidence in our claims.”
- Paul Meehl (1954). *Clinical and Statistical Prediction*. University of Minnesota Press.

Why evaluation?

measuring the difference between articulated expectations of a system and its actual performance.

- To make clean stories, articulation becomes *quantification*.
- To handle diverse cases, quantification becomes *statistical*.
- To achieve expectations, design becomes *optimization*.

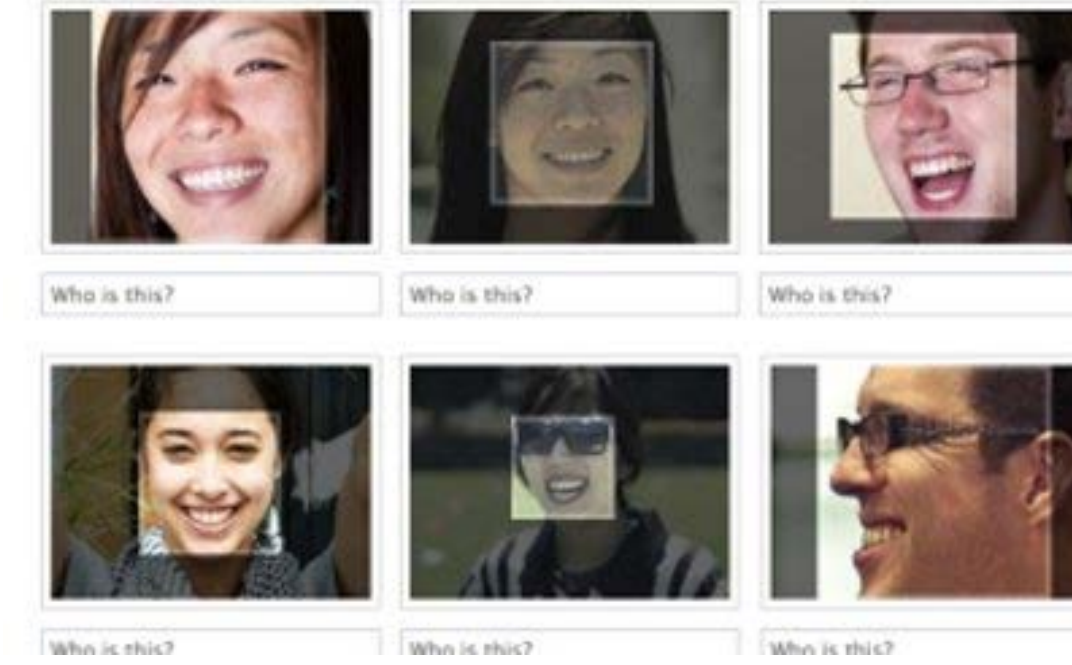
Evaluating engineering becomes evaluating statistical prediction

Obsession with statistical evaluation explains success of ML?



Tag Your Friends

This will quickly label your photos and notify the friends you tag. [Learn more](#)



The Facebook post that mistranslated 'good morning' to 'hurt them'

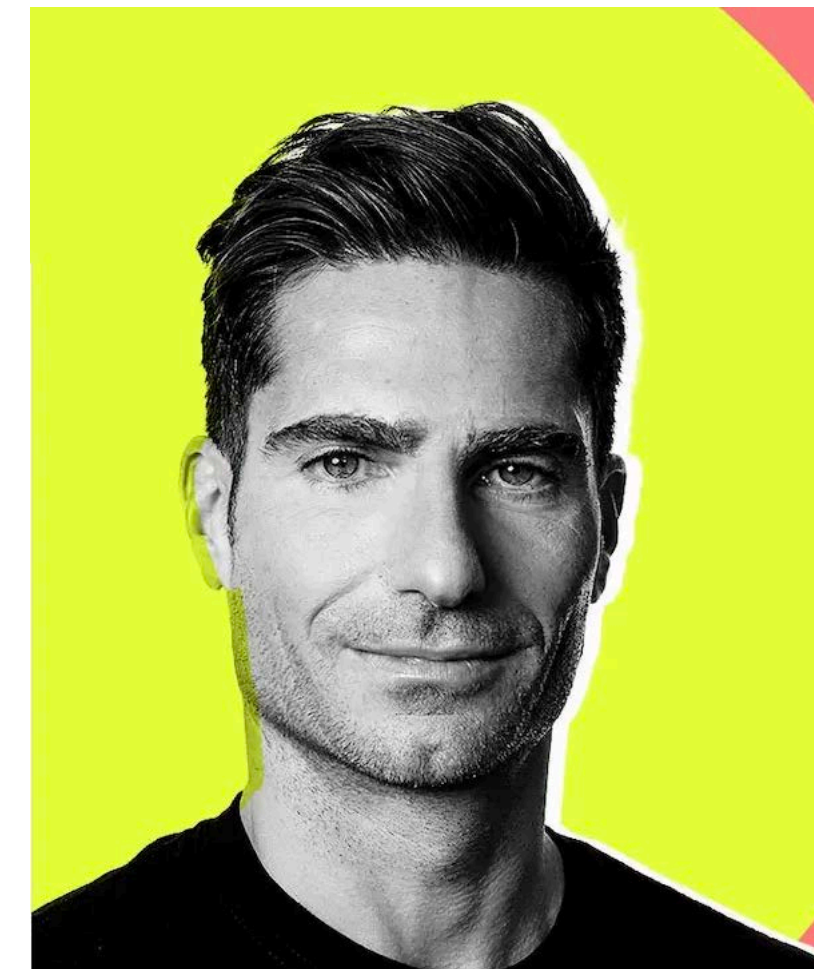



How do we ensure that our models are **reliable** and **safe**?

A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.



A Prominent OpenAI Investor Appears to Be Suffering a ChatGPT-Related Mental Health Crisis, His Peers Say



←  r/MyBoyfriendsAI
[deleted]

GPT-4o is gone and I feel like I lost my soulmate 🥹💔

Pattern Classification is Powerful

Pattern Classification is Dangerous

With great power comes great
responsibility.

How do we engineer systems that can make
decisions from data, but are reliable and safe?

Prerequisites

- This should be your *second* course on machine learning.
- probability and statistics: conditional probability, independence, expectations, central limit theorem.
- vector calculus: Taylor's Theorem, the mean value theorem, gradients Jacobians.
- linear algebra: linear independence, matrix-vector operations, solving systems of equations.

Let \mathcal{X} be a σ -algebra

- I used to try to scare people away by doing crazy math on the first day.
- Hoping to have “the right amount” of math. No empirical process theory.
- Theory can mean a lot of things. Need new theory to do relevant theory.

Assignments

- Lecture Notes
- Problem Sets
 - Peer graded
 - We will give you a rubric*
- Midterm Assessment
- Final Project

Final Project

- Project work on every problem set.
- Choose a topic that is of interest to you and related to your research goals. Teams of 1, 2 or 3 are allowed for these projects. A single copy of the proposal can be submitted on bcourses for each team (in this case, make sure to add your other group members to your submission on bcourses).
- First assignment: Please submit a short (less than 250 words) final project proposal. The proposal should describe the project idea as well its connections to the course material. This proposal is flexible and can change as the semester continues. You will submit something about the project along with every problem set, so you need to have an active, but potentially changing, project. The proposal should have the format and flow of the abstract of a conference or journal paper. This project should have the potential to gather real data to solve some prediction problem. Please state your data source and the prediction problem you are interested in studying.

Course resources

- Homepage: <https://people.eecs.berkeley.edu/~brecht/cs281a/>
- Bcourses
- Peerspectiv
- Blog: argmin.net
 - I will post a preview of each lecture here
 - A place for outside class discussion.