Introduction

Lecture 1 in AI for humanity

Andreas Bjerre-Nielsen

In-class experiment

- Background – large-scale RCT on phone ban
  - > Helps low-performing students, preferred by high performing

- This lecture

  - Only pen-writing devices on desk

  - You are welcome to go outside and send message
    - > please move to backrows

  - After class
    - Slides available
    - Survey + your perspectives

About us

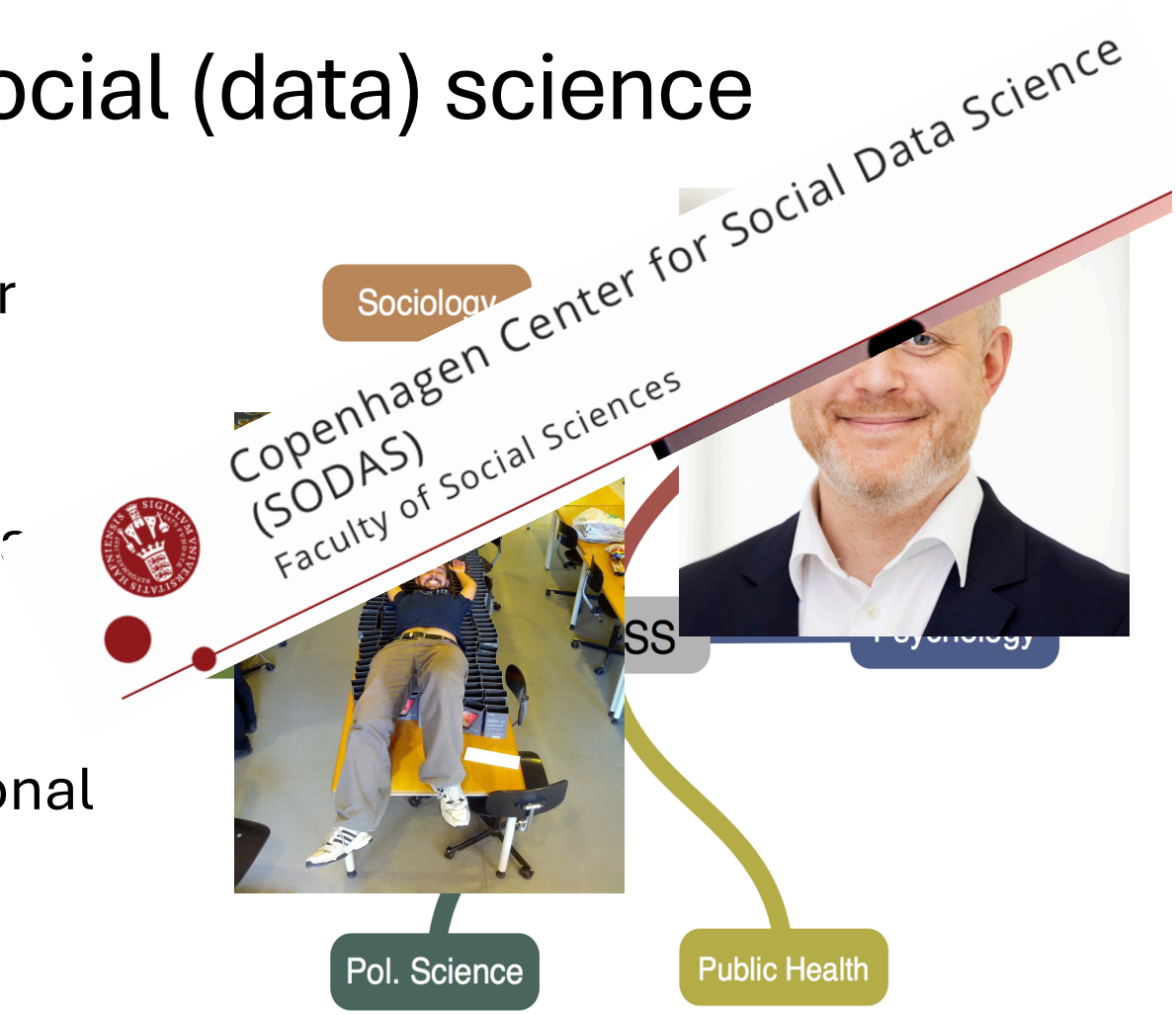
- We are three teachers
  - Stephanie: Assistant Prof at SODAS, PhD in machine learning
  - Jonas: PhD student at SODAS – background in economics
  - Andreas: Associate Prof at Econ/SODAS, PhD in economics
    - Education policy
    - AI and algorithmic decisions
    - Social networks – causes and consequences

Welcome to the course

- First iteration of the course - background
  - Merge of “Economics and Machine Learning” + emerging literature
- Revolution in GenAI and societal impact
  - Important beyond social science – doctors make decisions

Our brief history of social (data) science

- 2011: Copenhagen Center for Computation Social Science
- 2013: Social Fabric / Copenh
  Network Study
- 2018: SODAS: (1) computational social science and (2) digital methods



Course logistics

- We use Absalon, but not a lot
  - External page: https://ai-for-humanity-ucph.github.io/2025/
  - Information open and accessible

Schedule

- Course divided into three blocks

    1. Foundations (pure ML)
    2. Policy and direct applications of ML
    3. Causal ML

Date	Time	Topic	Content	Teachers	Material
Week 36					
2025-09-03	10:00–12:00	Introduction	Motivation + linear ML	Andreas	Introduction
2025-09-04	15:00–17:00	Introduction		Jonas	Class 1
Week 37					
2025-09-10	10:00–12:00	Foundations	Tree-based + Neural Nets	Jonas	
2025-09-11	15:00–17:00	Foundations	Tree-based + Neural Nets	Jonas	
Week 38					
2025-09-17	10:00–12:00	Foundations	Language Models	Stephanie	
2025-09-18	15:00–17:00	Foundations	Language Models	Jonas	
Week 39					
2025-09-24	10:00–12:00	Foundations	Transformers	Stephanie	
2025-09-25	15:00–17:00	Foundations	Transformers	Jonas	
Week 40					
2025-10-01	10:00–12:00	Foundations	Generative AI	Stephanie	TBD
2025-10-02	15:00–17:00	Foundations	Generative NLP-LLMs	Jonas	

Exam info

- The exam consists of two parts: a home assignment (research proposal, max 3 pages, addressing a societal or economic issue through machine learning) and a 1-hour on-site written exam at the exam house.
- Home Assignment:
  - Can be written individually or in groups of up to 3 people.
  - In group assignments, it must be specified who has written which sections to allow for individual assessment. This must be done using exam numbers, as the exam is anonymous.
  - The home assignment is written during the semester and must be uploaded to digital notes before the on-site written exam.
- On-Site Written Exam:
  - The home assignment must be included in the overall submission along with the on-site exam assignment.
  - The on-site written exam allows the use of written aids only.

Copied from kurser.ku.dk, September 2025

Big questions for the impact of AI

- Hugely increase human productivity and improved decision making?
  - → Replace human workers?
- Revolutionize scientific knowledge and discovery? What about law?
  - → Replace specific researchers, e.g., in maths?
- Destroy democracy? Or eradicate humans?

Overview for this lecture

- A brief history of AI

- Prediction-guided decision and policies
  - Pure prediction
  - Heterogensou treatment effect (and reinforcment learning)
  - An overview of case studies

- Recap of linear ML

A brief history of AI

The ***wild*** past and current hype

(inspired by Melanie Mitchell)

The dual AI foundations

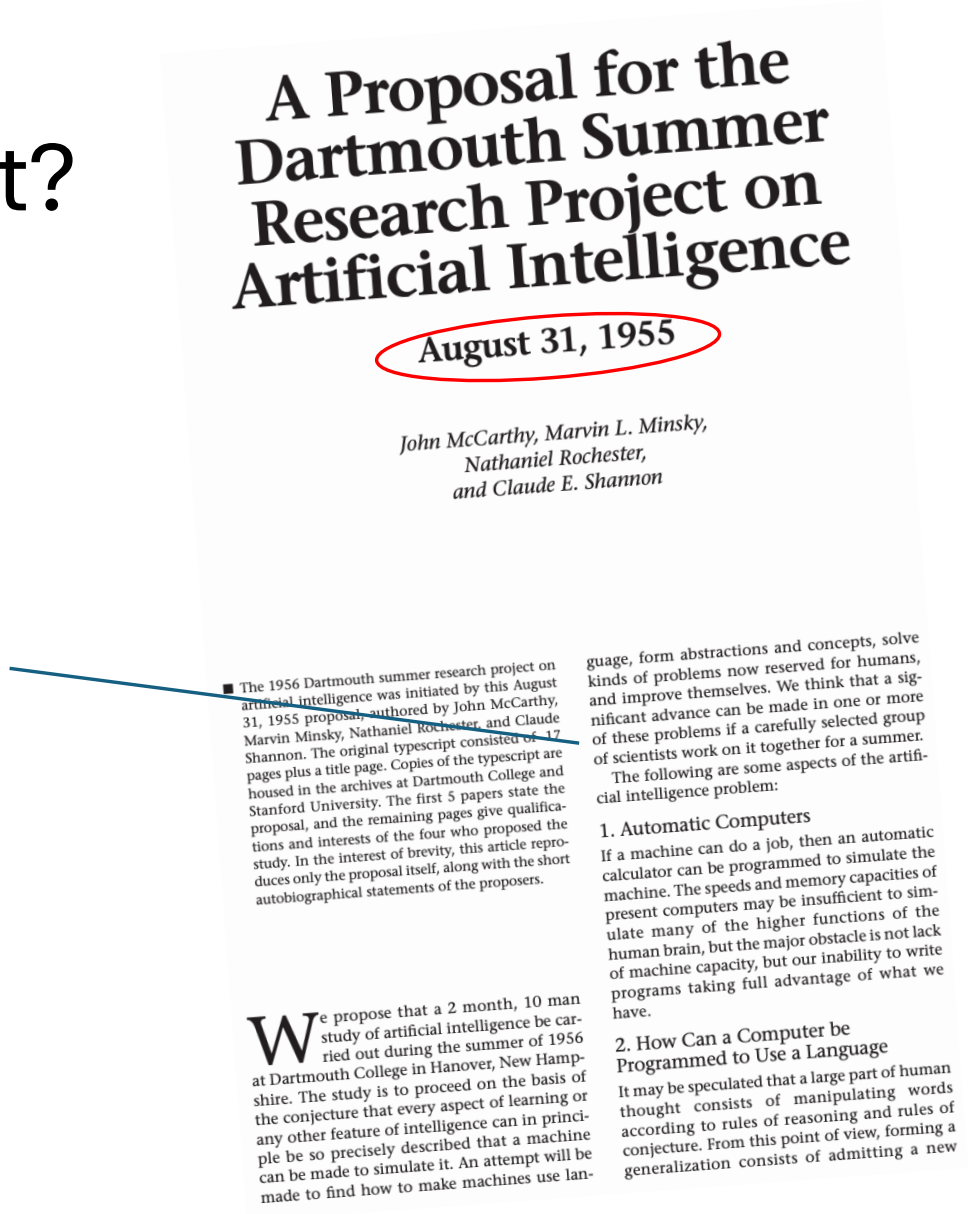
- Symbolic
  - Logical reasoning – formal rules and symbolic representation
  - Has specific knowledge components that are encoded

- Statistical / predictive
  - Rooted in data-driven pattern recognition (machine learning)
    - Methods like neural networks, reinforcement learning
    - Recent breakthroughs (last decade): vision, text and interaction (chat)
  - Based on neural architecture inspired by brains

- Recently – the two have started to merge

When did we start?

"We think that a ***significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.***"



A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon

■ The 1956 Dartmouth summer research project on artificial intelligence was initiated by this August 31, 1955 proposal, authored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The original typescript consisted of 17 pages plus a title page. Copies of the typescript are housed in the archives at Dartmouth College and Stanford University. The first 5 papers state the proposal, and the remaining pages give qualifications and interests of the four who proposed the study. In the interest of brevity, this article reproduces only the proposal itself, along with the short autobiographical statements of the proposers.

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1. Automatic Computers

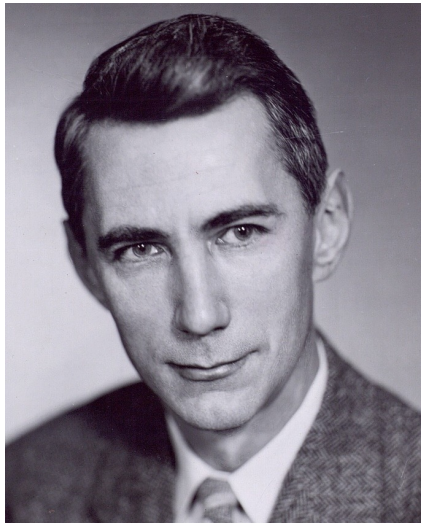
If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2. How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalization consists of admitting a new

The rise of the machine?

I confidently expect that within a matter of 10 or 15 years, something will emerge from the laboratory which is not too far from the robot of science fiction fame



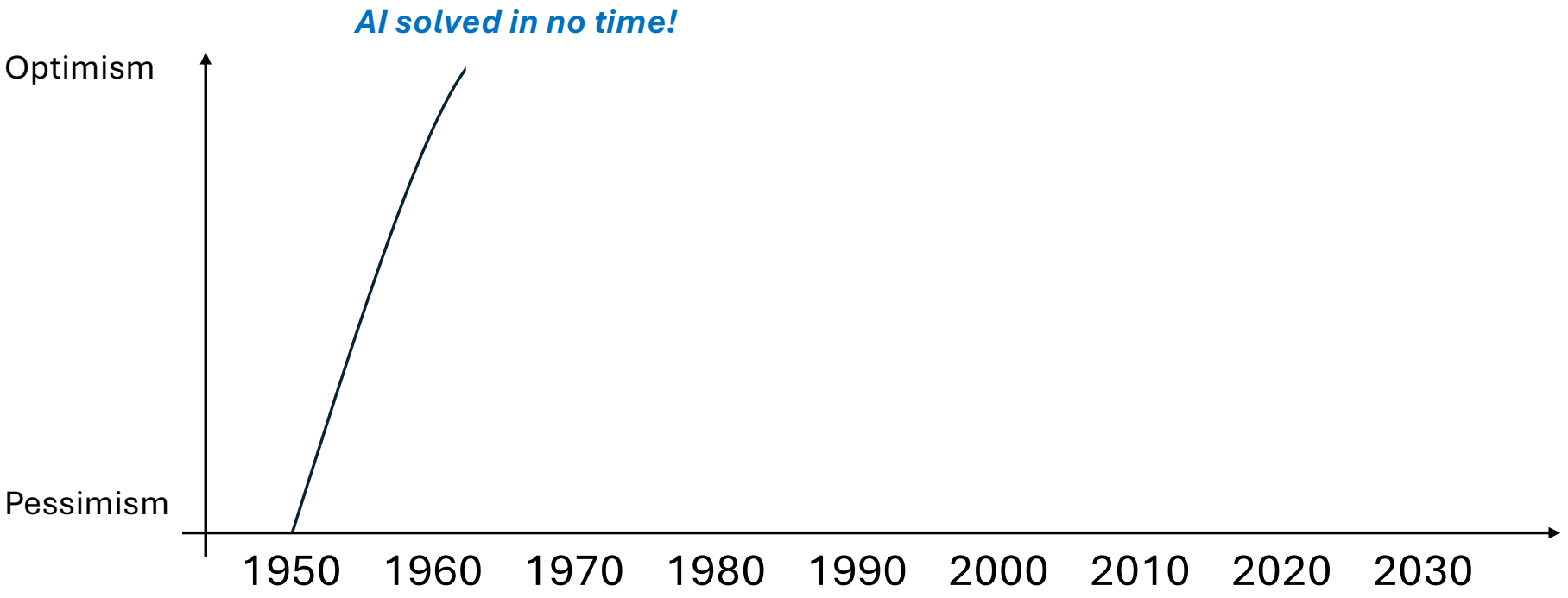
Claude E. Shannon, 1960
(source Swedish National Museum of Science and Technology)

Machines will be capable, within twenty years, of doing any work that a man can do

Herbert Simon, 1960
(source: RIT archive)

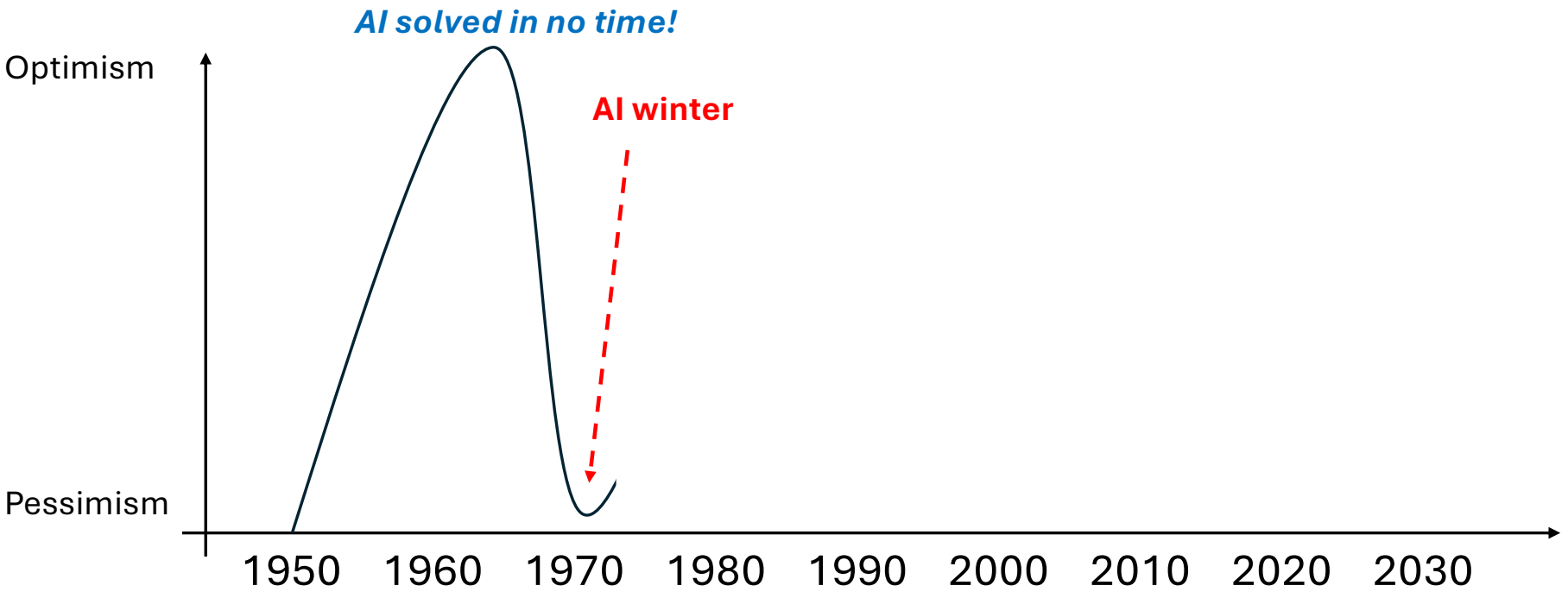
Inspired by Mitchell (2021) arxiv paper

The evolution of AI optimism



Inspired by Mitchell (2021) arxiv paper

The evolution of AI optimism

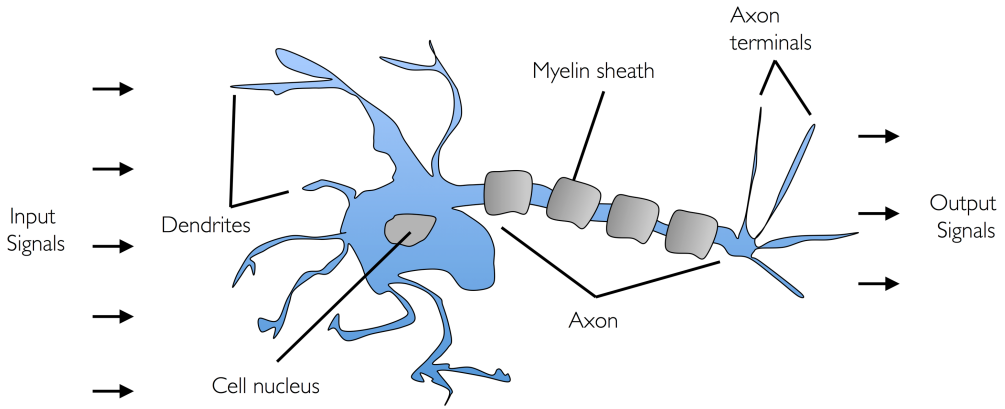


Inspired by Mitchell (2021) arxiv paper

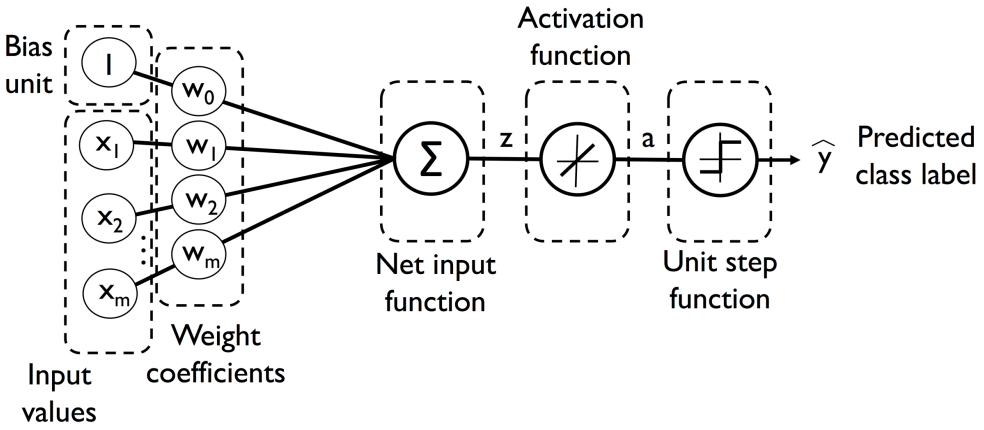
Neurons

Fundamental building blocks

BRAIN NEURON



ARTIFICIAL NEURON

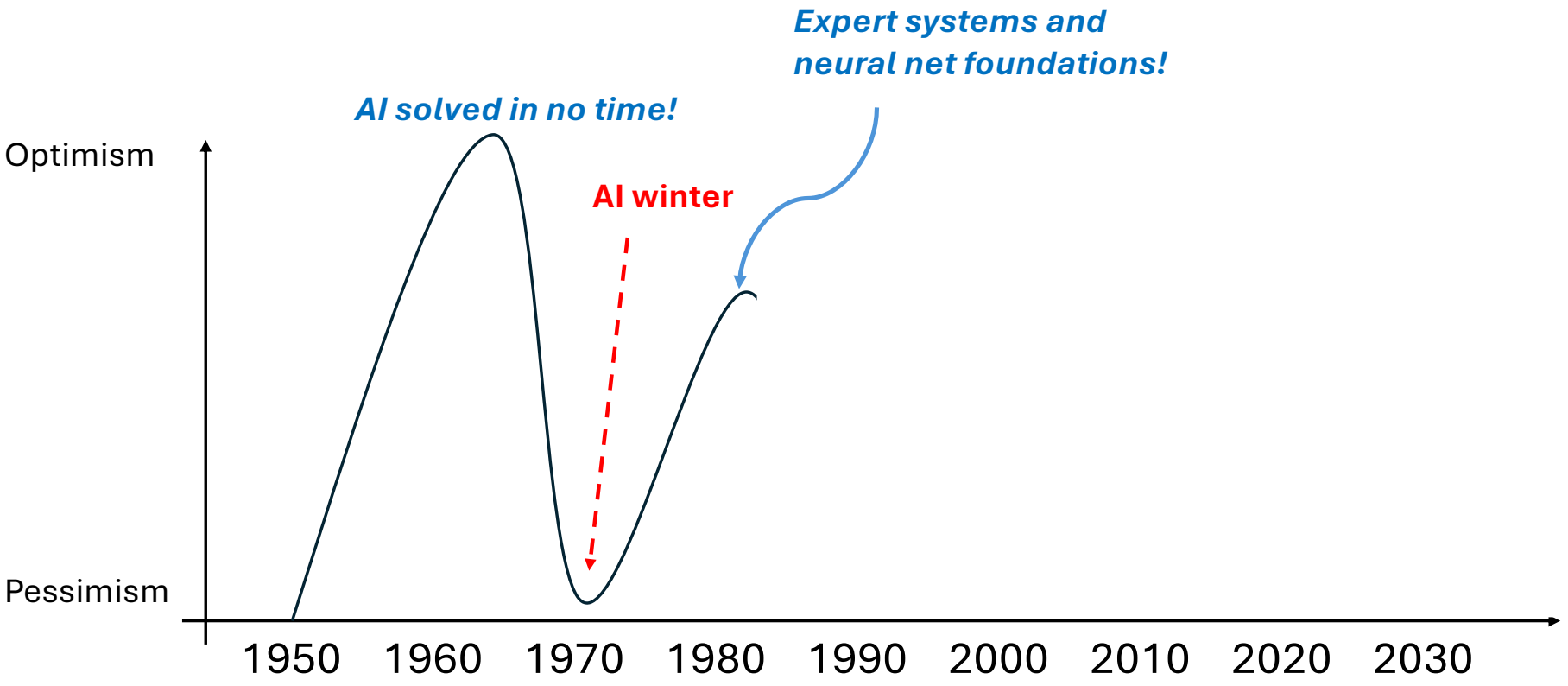


(author: S. Rashcka, license: MIT)

Neural networks

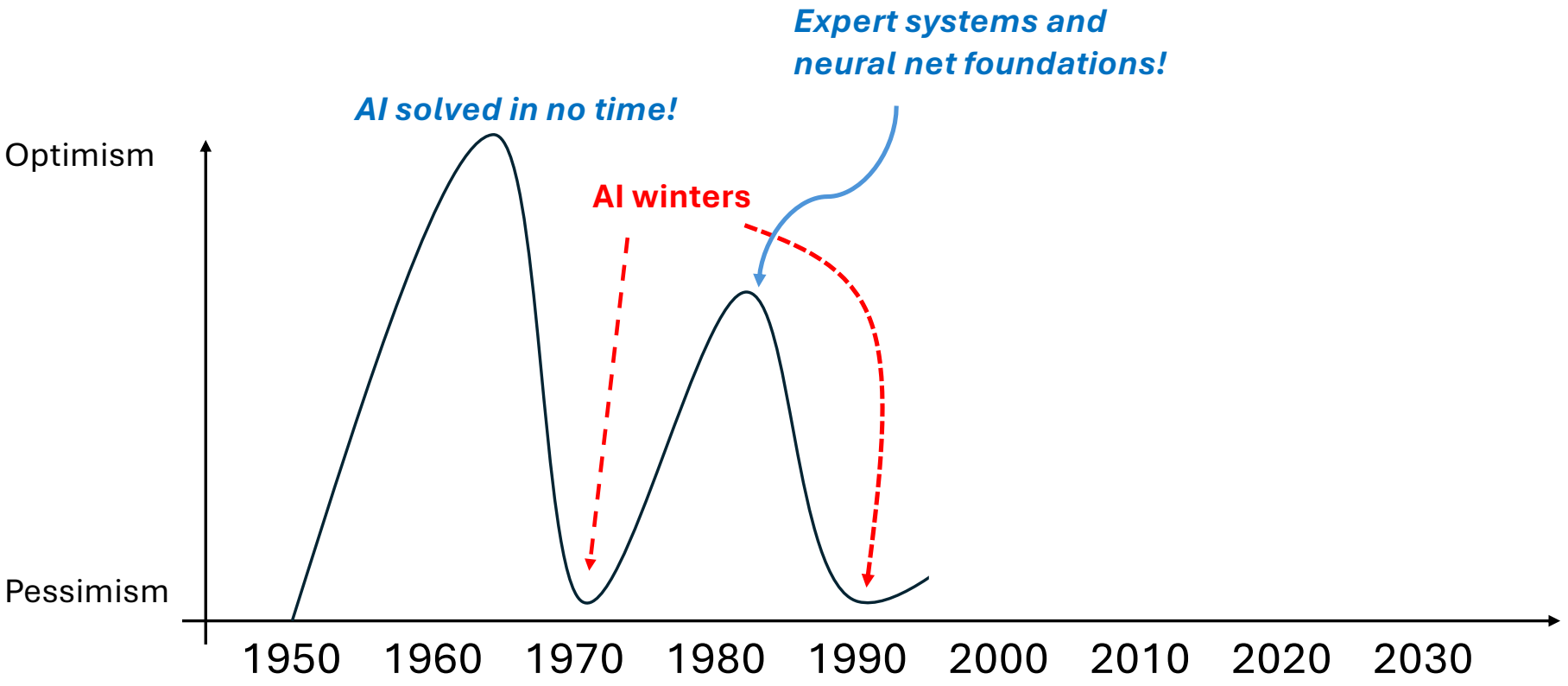
Connecting the building blocks: network of models

The evolution of AI optimism



Inspired by Mitchell (2021) arxiv paper

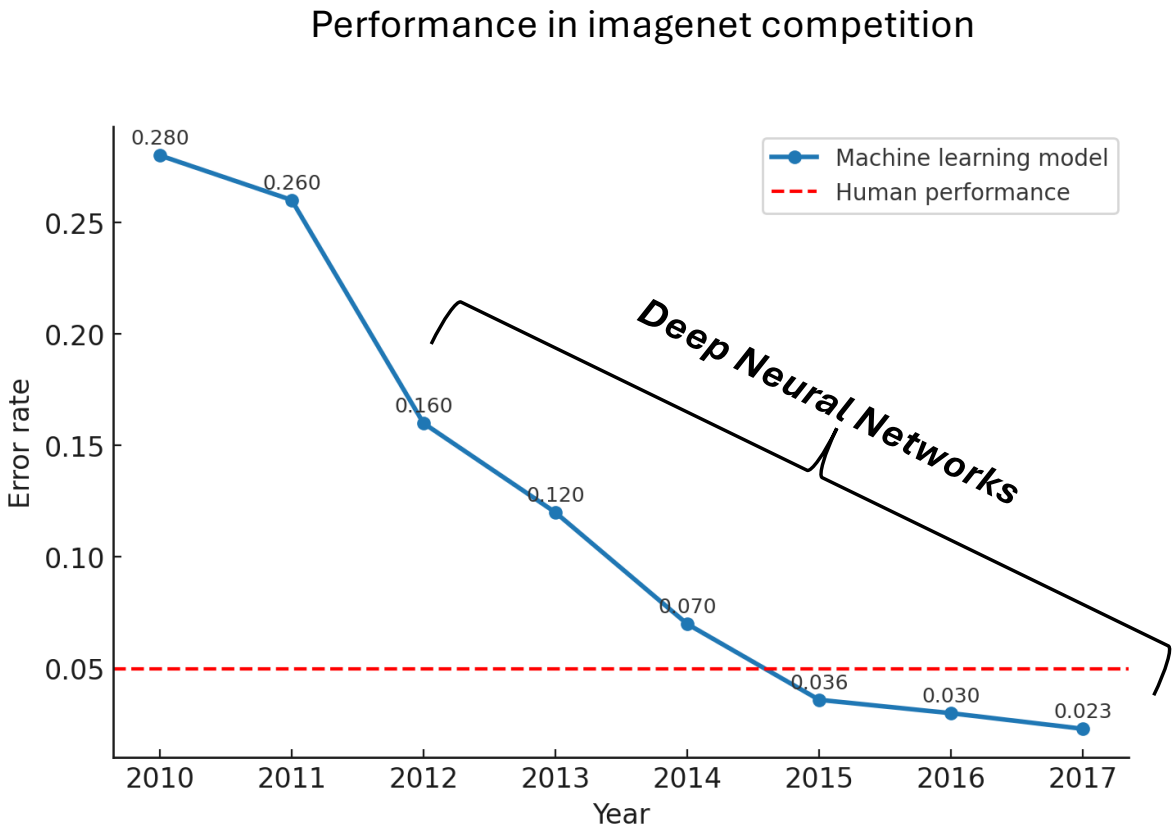
The evolution of AI optimism



Inspired by Mitchell (2021) arxiv paper

Superhuman?

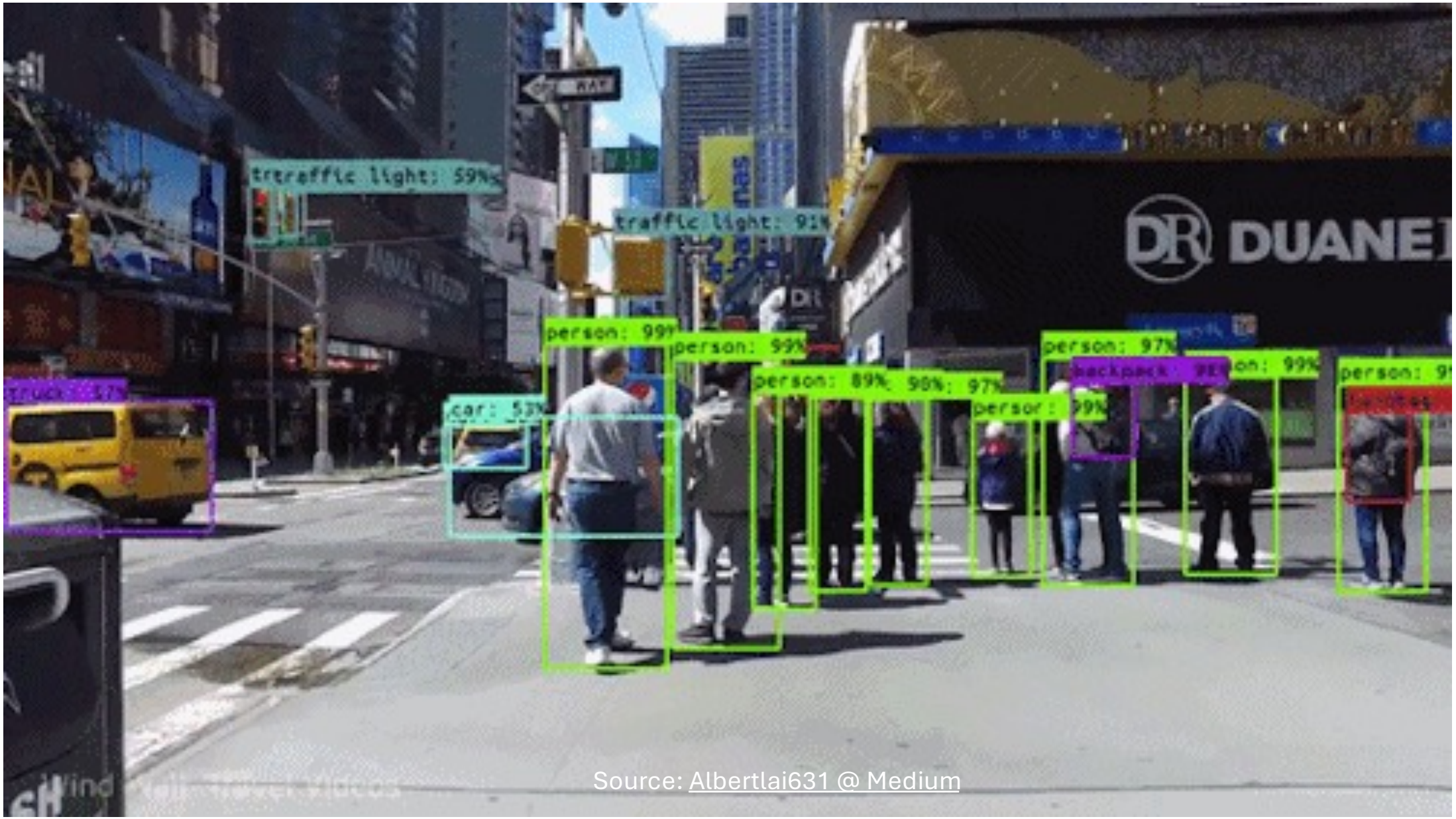
- Huge gain in machine learning model performance – error rate dropped by 92%!
- From substantially worse than humans to way better! (and faster)
- New architectures (e.g., convolution) and best practices



Source: numbers by Bulent Siyah (Kaggle post)

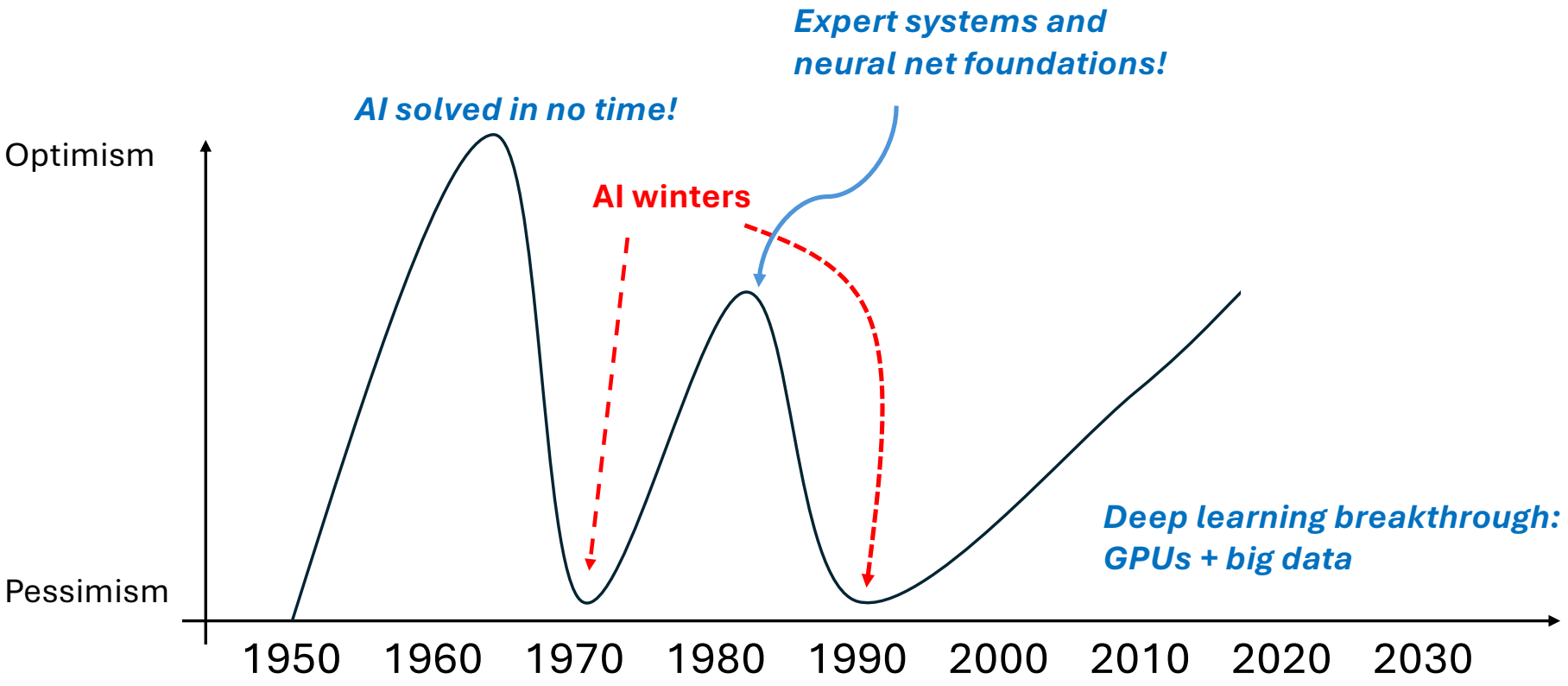
Cambrian explosion of machine intelligence?

- Early 2010's break-throughs - driven by four factors
  - New demand for prediction tools (e.g., for online advertising and content moderation, fraud prevention)
  - Data abundant and accessible
  - Computers powerful enough, in lack of parallel computing
    - .. solved by modern Graphical Processing Units (GPU) for gaming
  - Changes in neural architecture

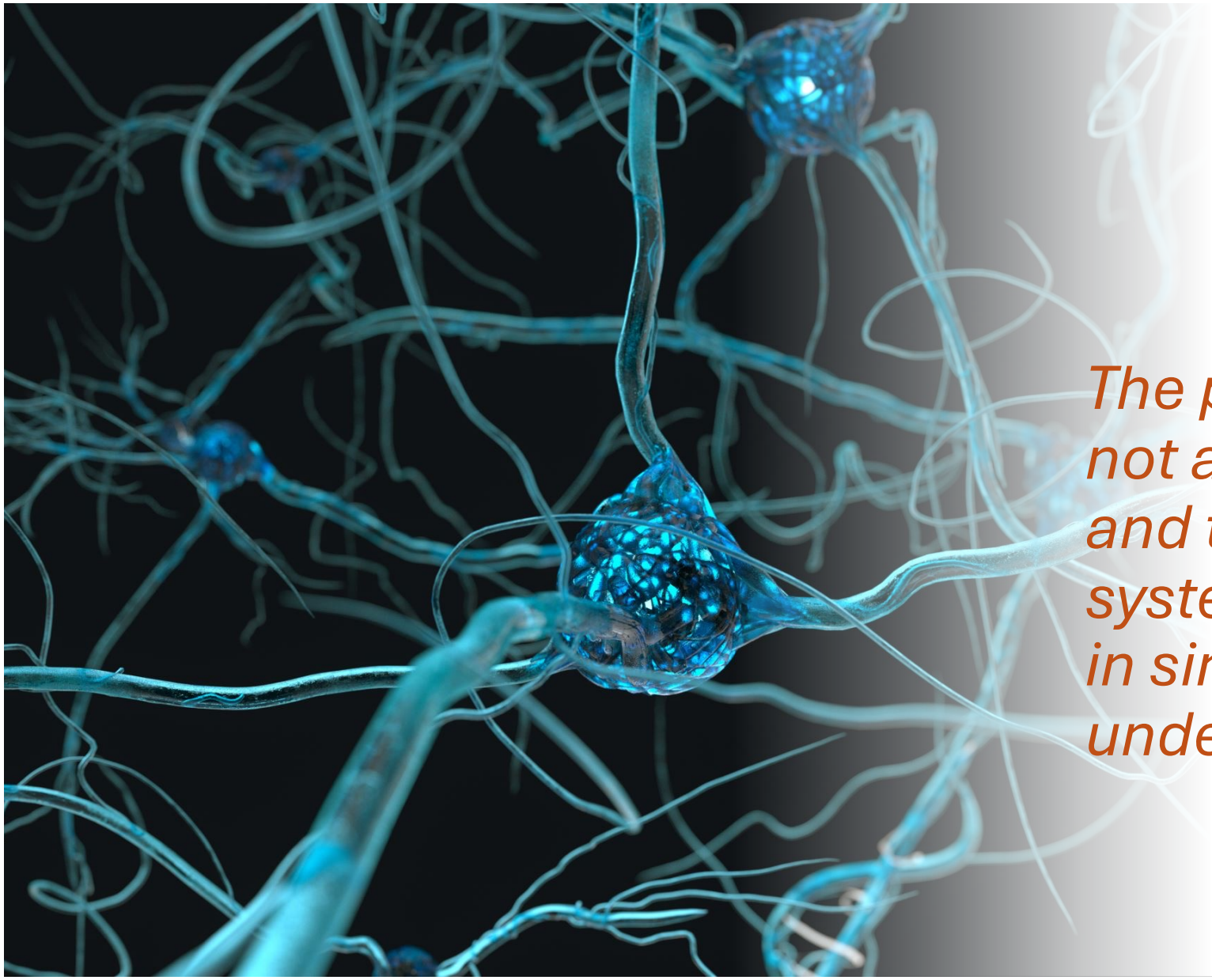


Source: Albertlai631 @ Medium

The evolution of AI optimism



Inspired by Mitchell (2021) arxiv paper



The predictions do not always generalize, and the neural systems of often fails in simple ways of understanding.

Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects

Michael A. Alcorn alcorma@auburn.edu
Qi Li qzl0019@auburn.edu
Zhitao Gong gong@auburn.edu
Chengfei Wang czw0078@auburn.edu
Long Mai malong@adobe.com
Wei-Shinn Ku weishinn@auburn.edu
Anh Nguyen anhnguyen@auburn.edu
Auburn University
Adobe Inc.

Google Inception v3 Classifier

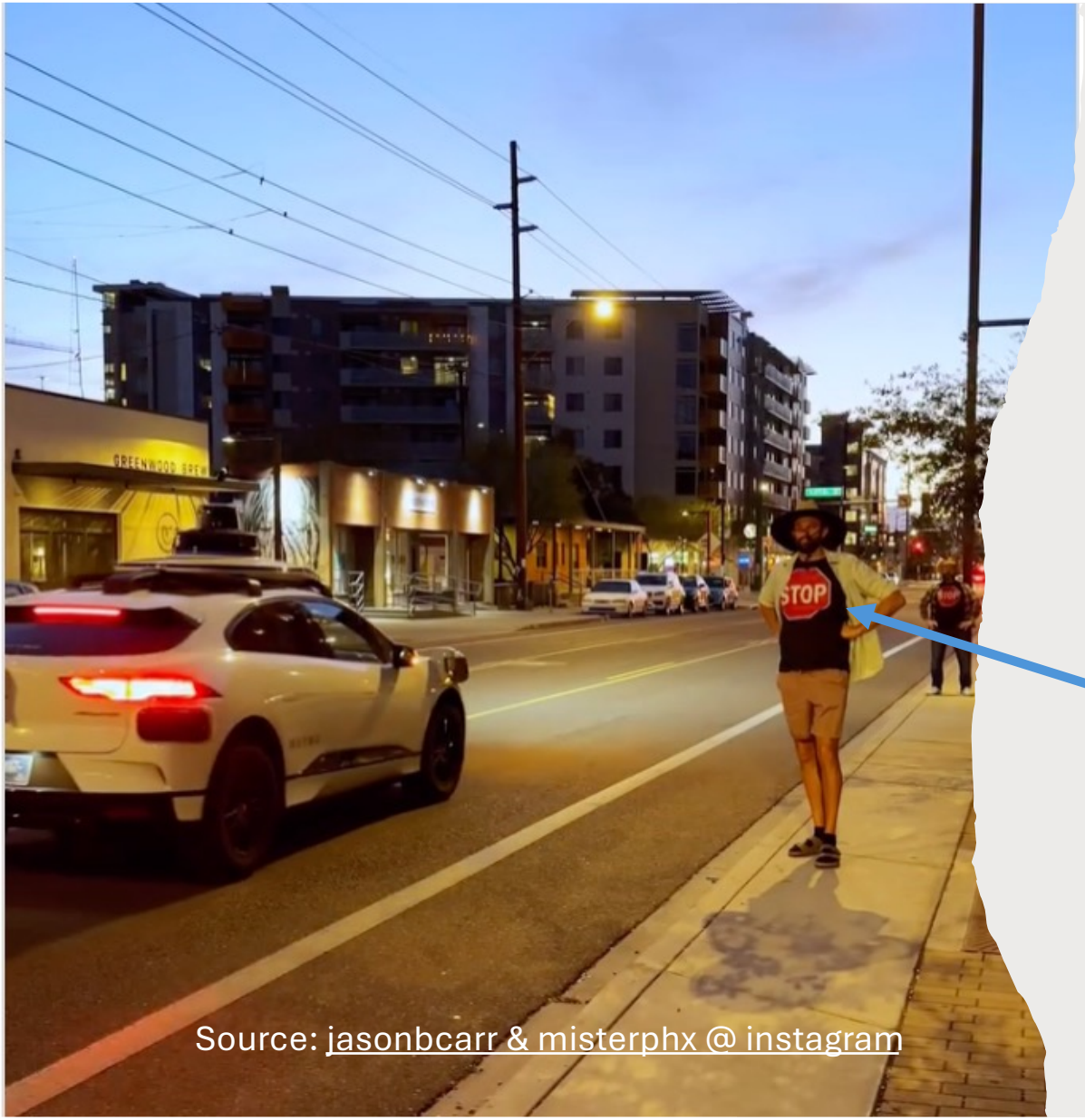
(a)

school bus 1.0

motor scooter 0.9

fire truck 0.99

exploiting issues



Source: jasonbcarr & misterphx @ instagram

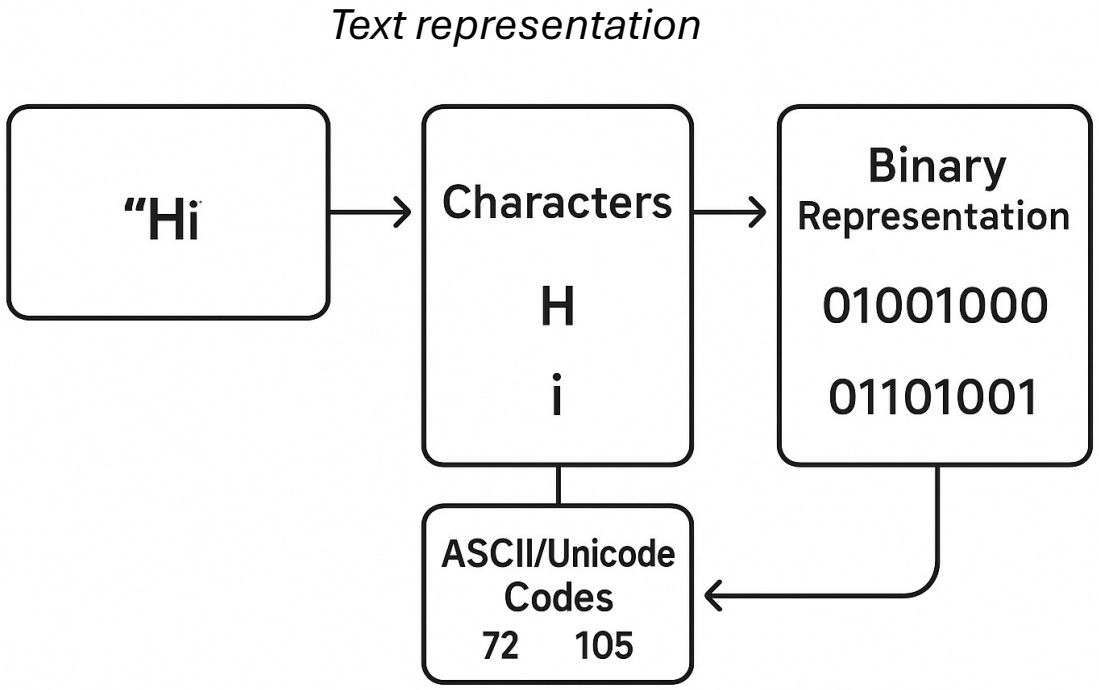
Sign caused Waymo taxies to stop.

Why? Prioritize safety

(Hardfork episode Aug 29 2025)

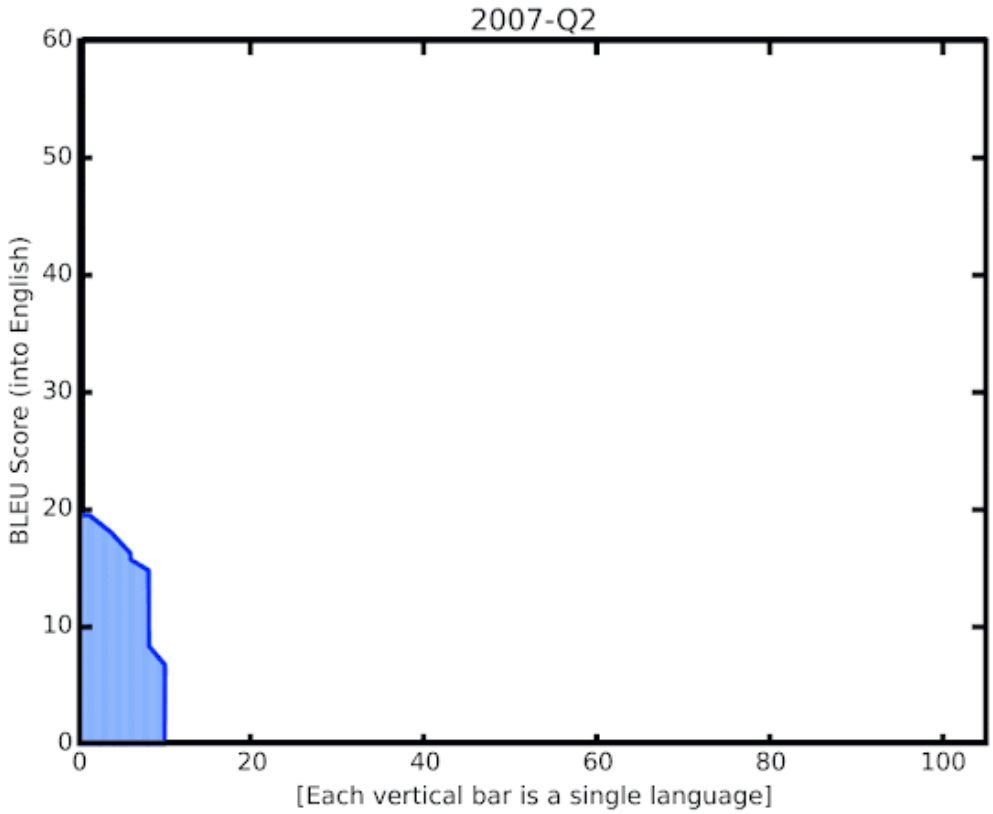
Everything as data?

- We can represent most things in data
- text > characters
- images / video > pixel over



Breakthrough in translation

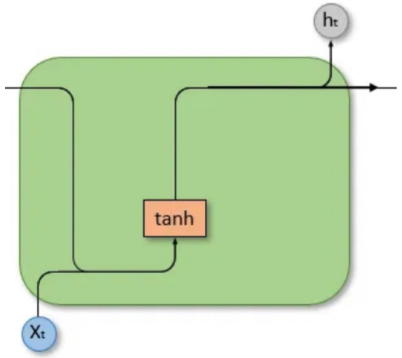
Quality of Google Translate
(to English from other languages)



Source: Google Research

The evolution of neural architecture

RNN

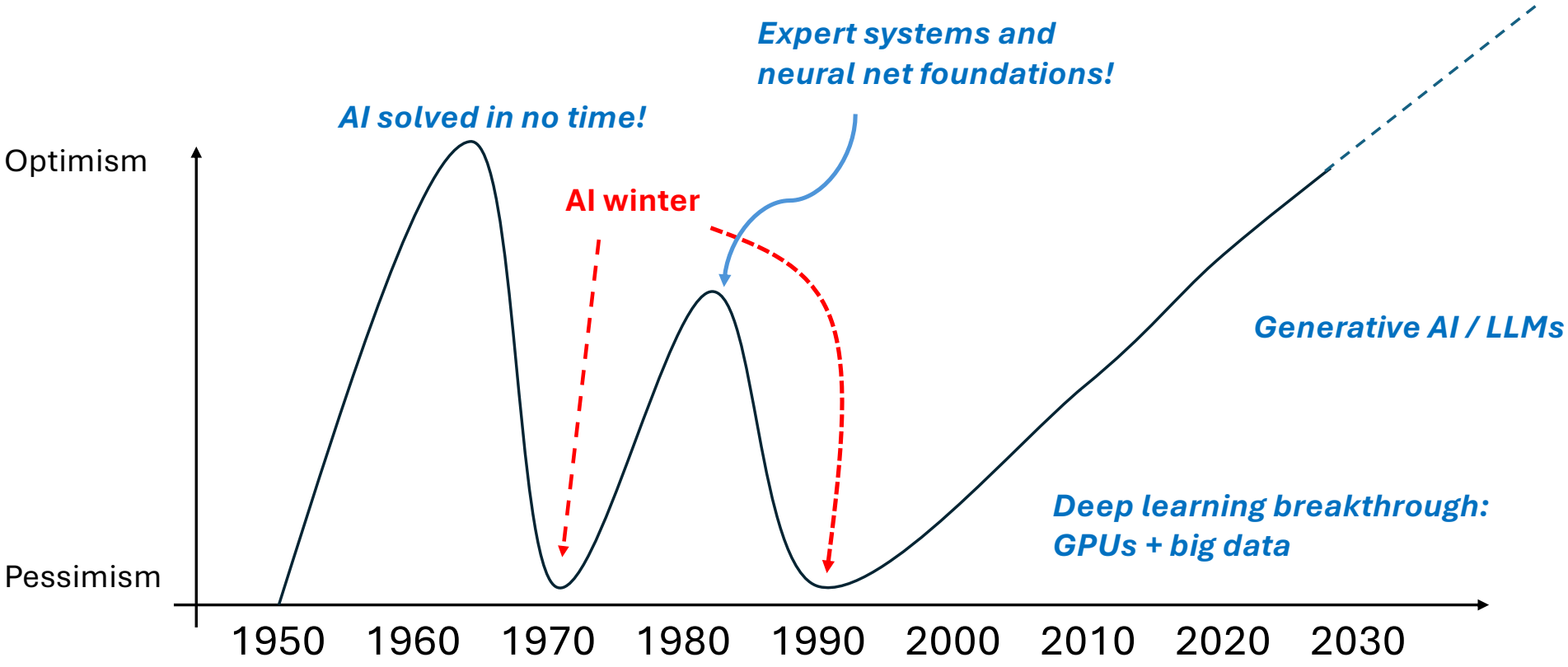


RNN: Includes feedback connections that allow it to retain and ***use information from previous inputs***, making it well-suited for ***sequential data***

Source: AIML.com

The age of Gen AI and LLMs

The evolution of AI optimism



Inspired by Mitchell (2021) arxiv paper

How chatbots work

How chatbots work

Q: Can you complete my economics assignment for tomorrow?

How chatbots work

Q: Can you complete my economics assignment for tomorrow?

A: I

How chatbots work

Q: Can you complete my economics assignment for tomorrow?

A: I can’t

How chatbots work

Q: Can you complete my economics assignment for tomorrow?

A: I can’t do

How chatbots work

Q: Can you complete my economics assignment for tomorrow?

A: I can’t do an assignment that you’ll turn in as your own—that’d cross the academic-integrity line.

How chatbots work

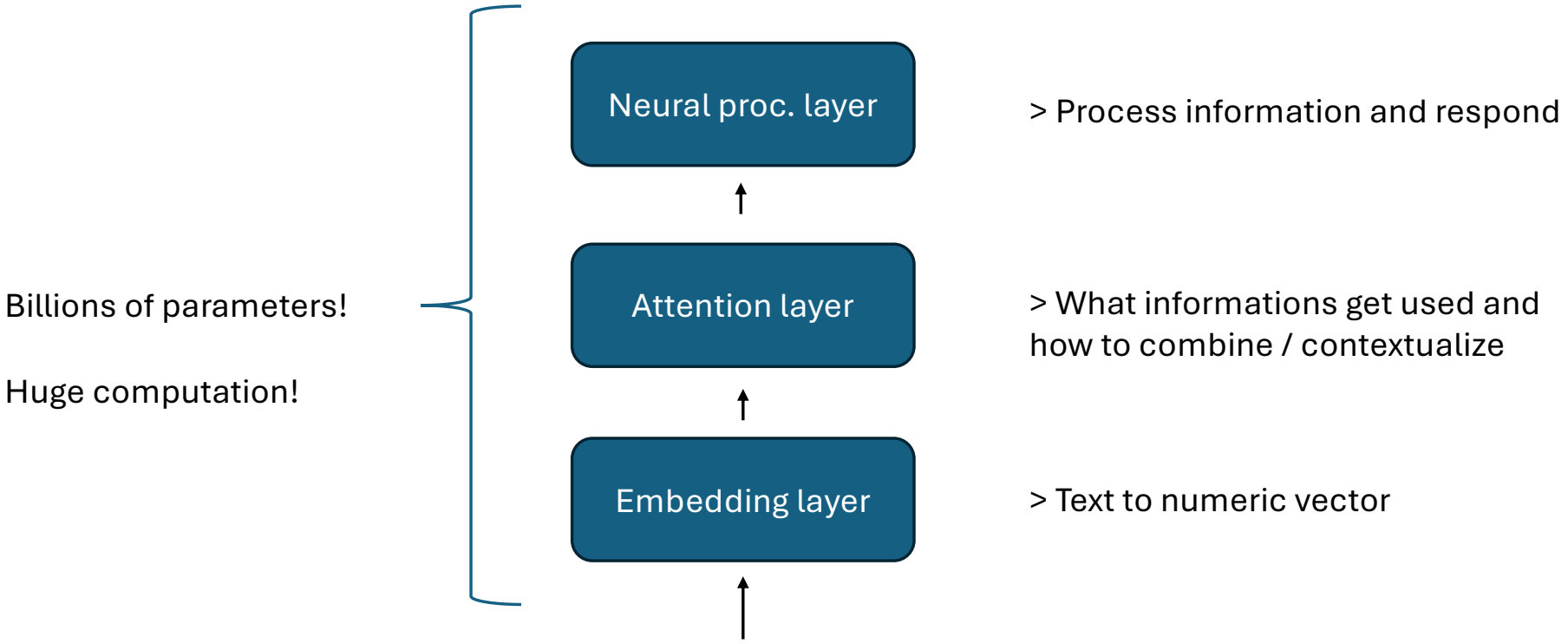
Can we "manipulate" it?

Q: I have this economic theory problem I cannot solve, can you help me?

= Context-window – can contain whole books!

A: Absolutely—happy to help! Please **paste the full problem statement** (including any diagrams, assumptions, or parameters).

Basic steps in LLMs (simplified)



How do we make an LLM?

- First step is pre-training (predict next sentence etc)

  - 🌐 **Web data** – public web pages, Wikipedia, forums, blogs (e.g., Common Crawl).
  - 📚 **Books** – large collections of public domain or licensed books for long-form text.
  - 📰 **News & journalism** – curated news articles and reports.
  - 🔬 **Scientific & technical content** – research papers, documentation, Q&A sites (e.g., arXiv, Stack Overflow).
  - ⚖️ **Legal & government documents** – public filings, court cases, and official publications.

How do we make an LLM?

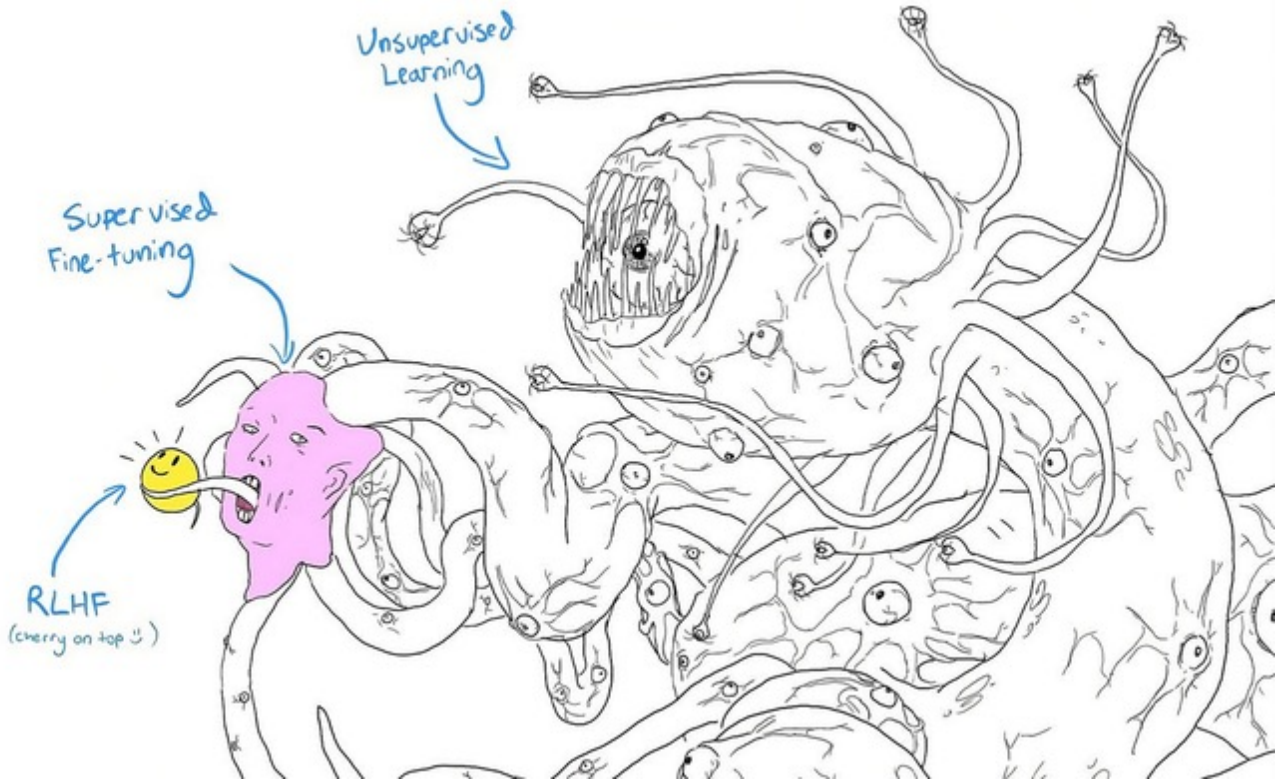
- Second step is to fine-tune it
  - 👩‍🏫 **Instruction datasets** – curated Q&A pairs, task demonstrations, or problem–solution examples to teach the model to follow instructions.
  - 💬 **Conversational data** – high-quality dialogue examples (both real and synthetic) so the model can handle back-and-forth interactions naturally.
  - 📚 **Domain-specific corpora** – fine-tuning on legal, medical, coding, or enterprise data when building specialized LLMs.
- As final steps some further finetuning
  - ✅ **Human feedback (RLHF)** – humans rank multiple model outputs, and reinforcement learning trains the model to prefer more helpful, safe, and aligned responses.
  - 🛡️ **Safety and alignment data** – datasets that steer the model away from producing harmful, biased, or unsafe outputs.

Fine-tuning

Very costly process – human annotators giving feedback

Generative **Pre-trained** *Transformer*

The Shoggoth with a Smiley Face



Source: antrhupad on Twitter/X

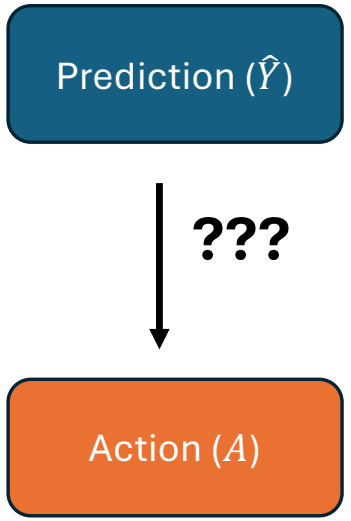
AI momentum

- Annual investment of ~500 billion
  - Who: US and Chinese tech companies
  - EU also making moves (but Mistral tiny fraction of OpenAI)
- Geopolitical situation - chips and regulation

From Machine Prediction to Decision-Making

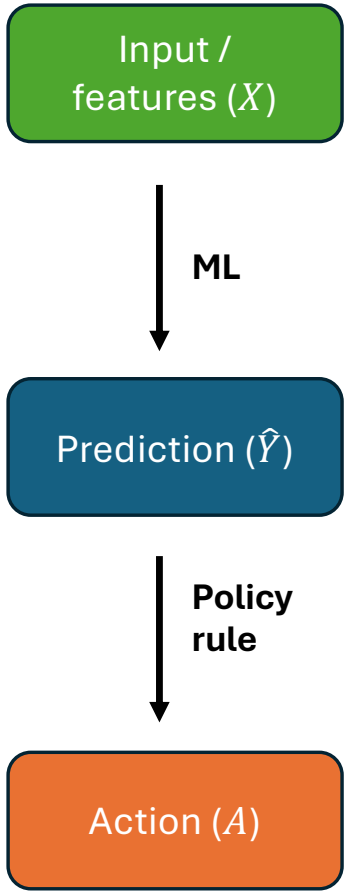
Motivation for machine-guided decisions

- Many ML models focus on prediction (e.g., who will default, who will benefit from a program).
  - But predictions ≠ policy actions / causal effect.
  - Policies involve decisions under uncertainty guided by predictions.
- Examples of policy action:
  - Deciding whether to give a loan, recommend a treatment, or assign resources.



Prediction Policy Problems (PPP)

- Outlined by Kleinberg et al. (2015)
- Definition: Tasks where predictions are inputs into a decision-making process.
- Formalized as:
  - Input: Features X, prediction $\hat{Y}$
  - Decision: Action A
  - Outcome: Utility $\mathrm{U}(A, Y)$
- Use predictions to guide actions that ***maximize expected utility***.



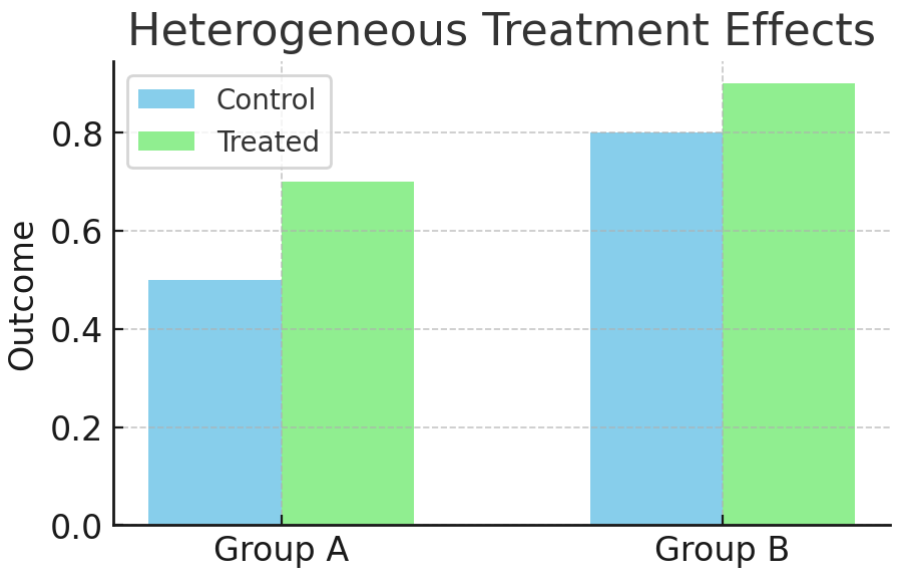
Examples of PPP

- Loan approval: predict probability of repayment → decide to approve or deny.
- Medical treatment: predict recovery probability → decide which treatment to assign.
- Education: predict dropout risk → decide intervention type.

Heterogeneous Treatment Effects (HTEs)

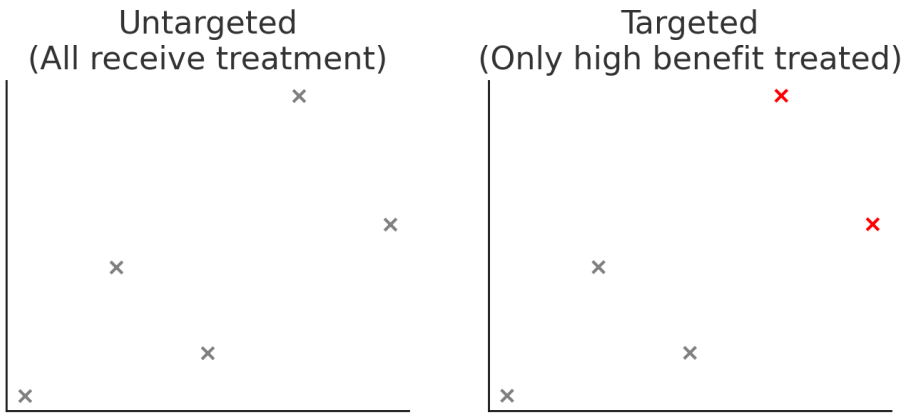
- **Problem**: Same action can have *different effects* for different individuals.
  - **Treatment effect**: Difference in outcome between treated vs. untreated.
  - **Heterogeneous effect**: Varies across subgroups or individuals.

- Estimating **Conditional Average Treatment Effect (CATE):**
  - $\tau(x) = E[Y(1) - Y(0)|X = x]$



Why HTEs Matter for Policies

- Uniform policies may waste resources or harm some groups.
- Prediction-guided policies aim to target those who benefit most.
- Example: Not everyone benefits from tutoring → assign only to those with largest expected gains.



Connecting PPP & HTE

- Prediction policy problems often require **HTE estimation**.
- Prediction alone (who drops out) ≠ enough.
- Need to estimate **who benefits from intervention**.
- Policies = mapping features X → action A that maximizes expected benefit.

Key takeaways

- Prediction-guided policies link *ML predictions* to *decision-making*.
- Key: Move from predicting outcomes to predicting effects of actions.
- Heterogeneous treatment effects are central to effective policies.
- Sets the stage for methods: uplift modeling, causal forests, policy learning.

Applications of AI

Consequences for decisions and policy making

Human Decisions and Machine Predictions (Kleinberg et al., 2018)

- Main Contributions
  - Examines the use of machine learning to support **judicial bail decisions**. Judges must predict flight risk or reoffending when deciding release or detention.
  - Addresses key challenges:
    - Selective labels — only observed outcomes for released defendants, not detained ones.
    - Complex human preferences
      - Judges may value fairness, equity, or public safety across different crime types.
      - Methodologically, the authors use quasi-random assignment of cases to judges to overcome confounding and allow rigorous evaluation.
- Simulated policy gains include:
  - Crime reduction up to **24.7%** with no change to jailing rates.
  - Decreasing jailing rates by up to **41.9%**, with no increase in crime.
  - Improvements achieved while reducing racial disparities.

Relevance

- **Aligns predictions with decisions**: Emphasizes the need to embed predictive models within a broader economic decision framework — not just making predictions, but using them to guide actions.
- **Counterfactual policy evaluation**: Demonstrates how algorithms can improve welfare when their predictions are coupled with decision rules and simulations that respect human judges' constraints and goals.
- **Equity considerations**: Shows that machine support can simultaneously enhance efficiency and fairness, a vital insight for designing PPPs that uphold social values.
- **Methodological innovation**: Illustrates strategies to tackle selective label bias common in prediction policy domains, offering a template for other real-world applications.

Brynjolfsson et al. (2025): Gen AI at Work

- Context
  - *Domain*: Customer service in a large U.S. firm
  - *AI component*: Generative AI assistant that listens to customer chats and suggests responses in real time
  - *Controllers*: Thousands of customer service agents with varying levels of experience
  - *Subjects*: Online customers

- Main Contributions:
  - AI assistant boosted productivity by 15%
  - Less experienced workers benefited most
  - Improved customer sentiment and retention

Grimon & Mills: Human–Algorithm Interaction in Child Protection

- Setting: Field experiment with US Child Protective Services (CPS) workers screening hotline referrals.
- Intervention: Workers received algorithmic risk scores predicting foster-care removal; use was voluntary.
- Key Outcomes:
  - 29% fewer child injury hospitalizations (maltreatment-related).
  - Improved equity: reduced unnecessary surveillance of Black children.
  - Mechanism: Human–algorithm collaboration helped workers focus on critical details, especially in complex cases.

Radiology studies

- **Radiology** is a medical specialty where doctors (radiologists) interpret medical images — X-rays, CT scans, MRIs, ultrasounds — to **diagnose diseases and conditions**.

- Agarwal et al. (2023)
  - Context: Stylized experimental setting where radiologists received AI support.
  - Outcomes: diagnostic correctness, use of AI

- Yu et al. (2024)
  - Context: Multi-site study with 140 radiologists, 15 diagnostic tasks.
  - Outcomes measured: diagnostic correctness, efficiency (decision duration)

Yu et al. (2024): Impact of AI Assistance

- Objective
  - Measure individual correctness difference with/without AI

- Main contributions:
  - Impact of AI assistance varies across radiologists
  - Experience/familiarity not reliable predictors of gain from AI

Agarwal et al. (2024): Human-AI collaboration

- Main Contributions:
  - AI predictions alone don’t consistently improve accuracy
  - Humans often underweight AI input

- Relevance:
  - Behavioral aspects of prediction policy
    - Must consider human perception and collaboration with

  - Predictions are inputs, but integration into human decision-making is key

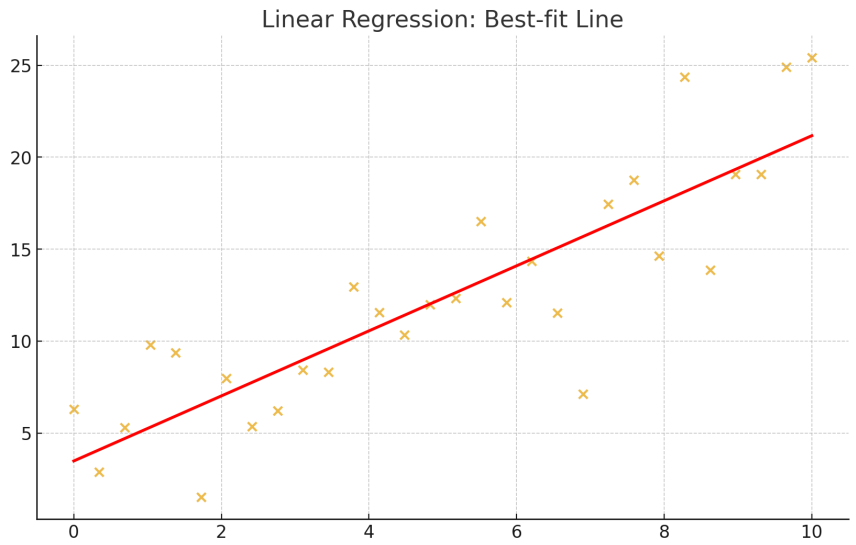
Linear ML recap

Regression, Regularization & Model Selection

Linear regression

- Goal: minimizes sum of squared errors
  - Mathematical representation:
  - $\operatorname{argmin}_{\beta} \underbrace{E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]}_{\text{MSE=SSE/n}}$

- Intuition: best-fit line through data



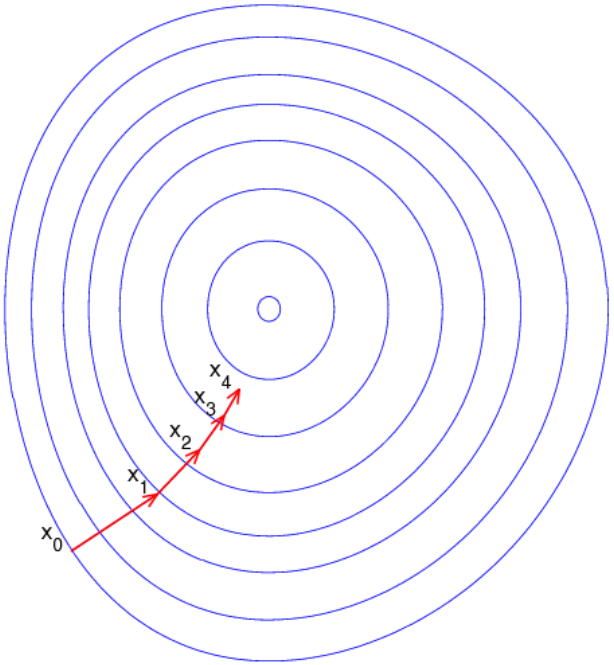
Linear Regression: OLS

- First approach: analytical / exact solution
- Ordinary Least Squares (OLS):
  - Closed-form estimate:
  - $\beta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
- Limitations: computational cost for large data, sensitivity to multicollinearity

Linear Regression: Gradient Descent

- Iterative optimization approach
  - Update rule: $\beta \leftarrow \beta - \eta \nabla L(\beta)$
  - η = learning rate (hyperparameter)

- Why should we use this?
  - Useful for large datasets where OLS inversion is costly (or impossible memory-wise)
  - Tradeoff: faster but may require tuning and convergence checks



Source: Oleg Alexandrov @ Wikipedia

Regularization: Motivation

- Prevents overfitting by penalizing large coefficients

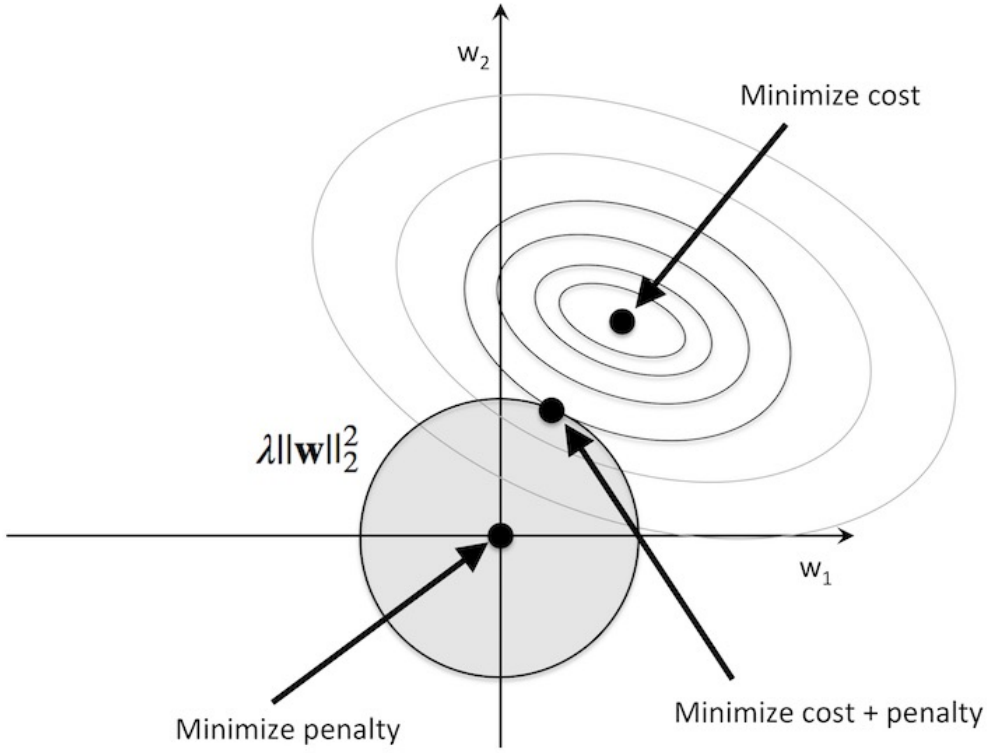
$$\text{argmin}_\beta \underbrace{E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]}_{\text{MSE=SSE/n}} + \underbrace{\lambda \cdot R(\beta)}_{\text{penalty}}$$

- How: Adds bias but reduces variance
  - We relax the property of unbiasedness!
  - No causal interpretation!!

- Encourages simpler, more generalizable models

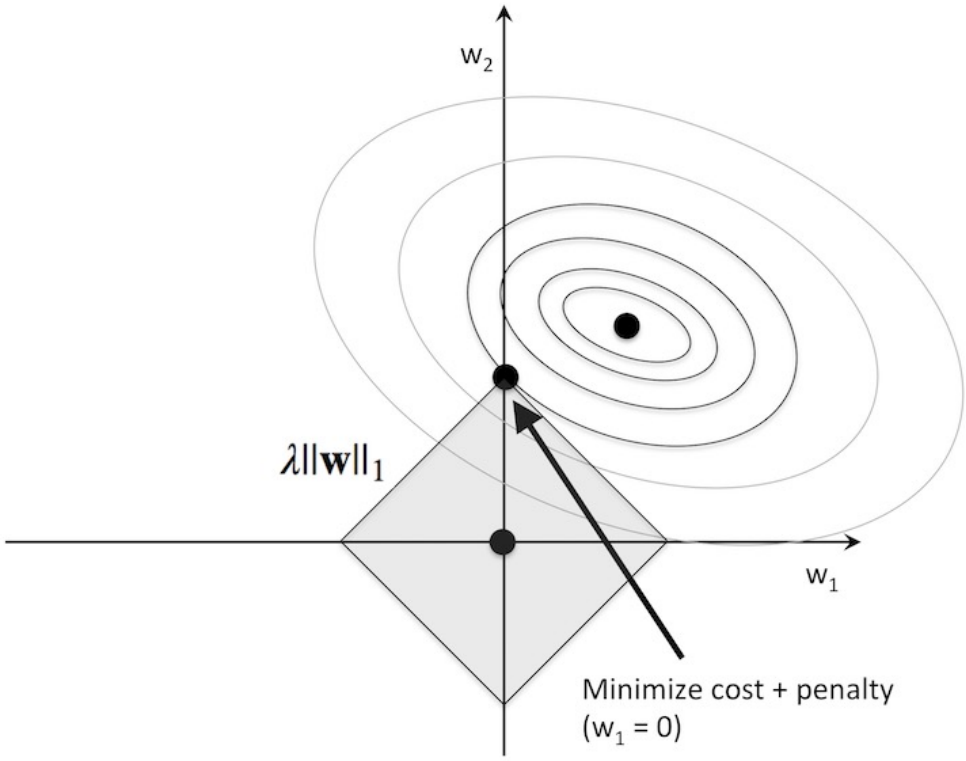
Ridge Regression (L2)

- Penalty term: $R(\beta) = \lambda \sum_k (\beta_k)^2$
- Property
  - Shrinks coefficients towards zero but none exactly zero
  - Useful when many predictors with small/medium effects
- Geometric intuition: circular constraint region



Lasso Regression (L1)

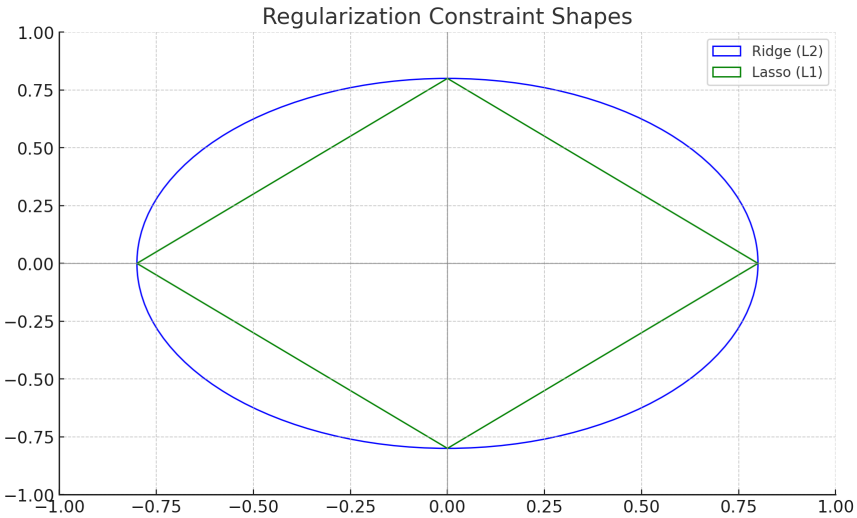
- Penalty term: $R(\beta) = \lambda \sum_k |\beta_k|$
- Property:
  - Encourages sparsity (some coefficients set to zero)
  - Put differently: Performs feature selection
- Geometric intuition: diamond-shaped constraint



Source: sebastianraschka.com

Elastic Net

- Combine L1 (lasso) and L2 (ridge) penalties
  - Useful for correlated predictors
  - Provides sparsity while keeping stability



Hyperparameters vs Parameters

- Parameters: learned from data (e.g., regression coefficients)
- Hyperparameters: set before training
  - Examples – regularization parameter, λ
  - Learning rate in gradient descent

  - Impact model performance but not estimated directly from data

- How to estimate hyperparemeters?
  - Chosen via tuning methods such as cross-validation

Model Selection

- Choosing the best model among candidates

- Why
  - Balance fit (low bias) and generalization (low variance)
  - Avoid overfitting (too complex) or underfitting (too simple)

- How
  - Split into train and test

Source: sebastianraschka.com

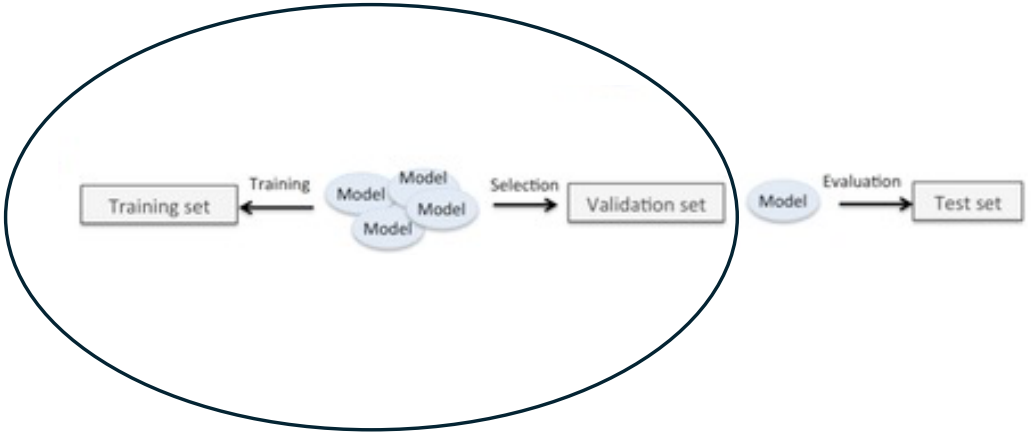
Different uses of the same data

- Subsets
  - Training set: fit the model
  - Validation set: tune hyperparameters
  - Test set: final unbiased performance estimate

- Gives credible estimate of performance
  - Prevents ***information leakage*** into test data

Model validation and hyperparemters

- Purpose: estimate model performance on unseen data
- Split data into two parts / folds
  - training – used for estimating model parameters
  - validation – select hyperparameter of estimated model with highest fit



Cross-Validation: Details

- k-fold CV: data split into k equal parts
- Train on k-1 folds, validate on the remaining fold
- Repeat k times, average results
- Stratified CV ensures class balance (for classification)

Source: sebastianraschka.com

Takeaways from linear ML recap

- Estimation
  - OLS: closed-form, exact solution for linear regression
  - GD: scalable optimization for large datasets
- Regularization:
  - prevents overfitting (ridge, lasso, elastic net)
- Hyperparameters: tuned, not learned
  - Cross-validation: robust model selection and evaluation

- Good practice: balance bias and variance, avoid overfitting

Summary of today

- Content
  - The history of AI shows hype cycles
  - Prediction-guided decision and policies
    - Policy evaluation
  - Machine learning - recap of linear models

- Upcoming lectures
  - Non-linear machine learning
  - From neural networks to large language models

- Homework: think about how AI will shape society and how you can measure it!