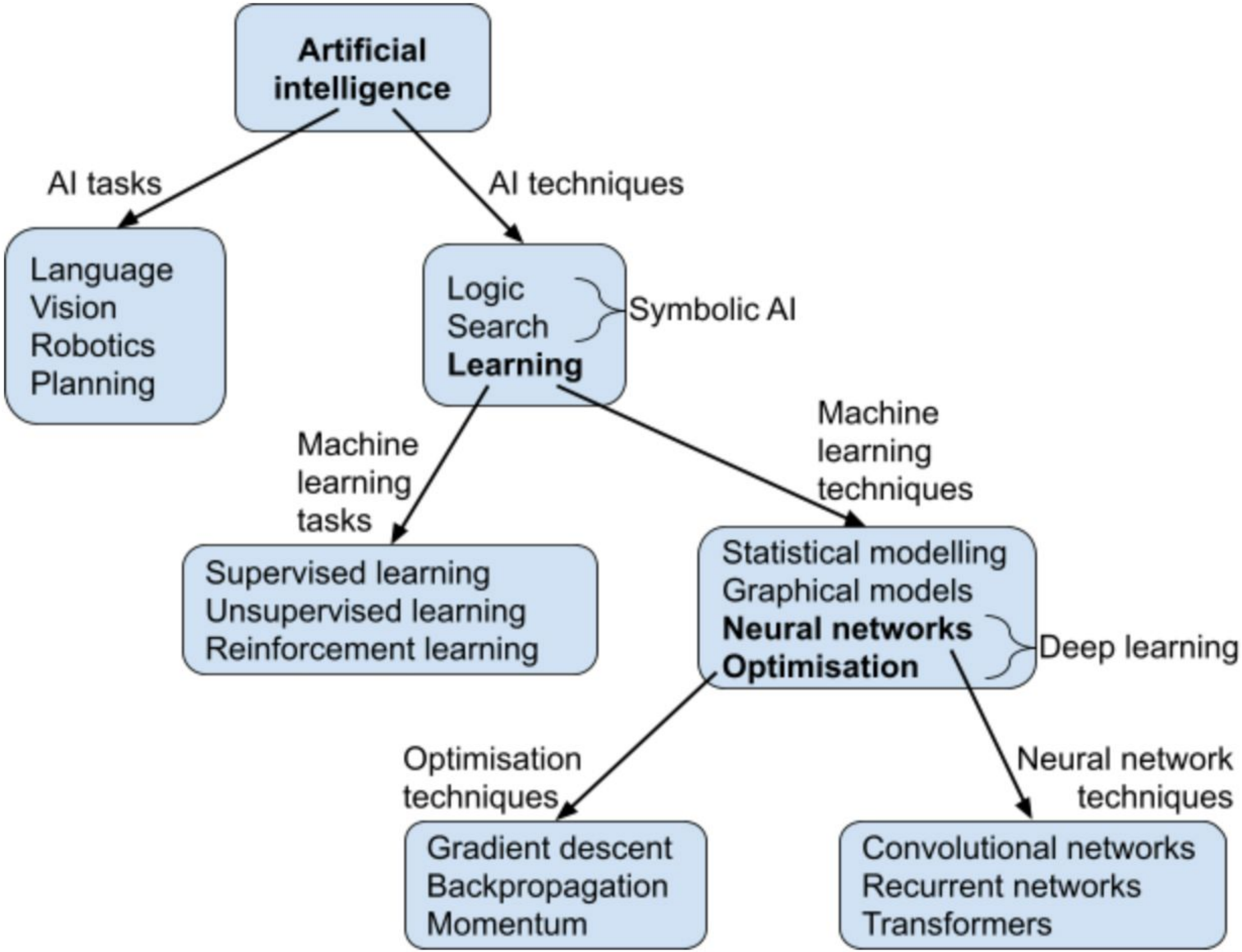
AGI Safety Fundamentals

•••

Session 0: Introduction to machine learning



Artificial intelligence

AI tasks

Language
Vision
Robotics
Planning

AI techniques

Logic
Search
Learning

Symbolic AI

Machine learning tasks

Supervised learning
Unsupervised learning
Reinforcement learning

Machine learning techniques

Statistical modelling
Graphical models
Neural networks
Optimisation

Deep learning

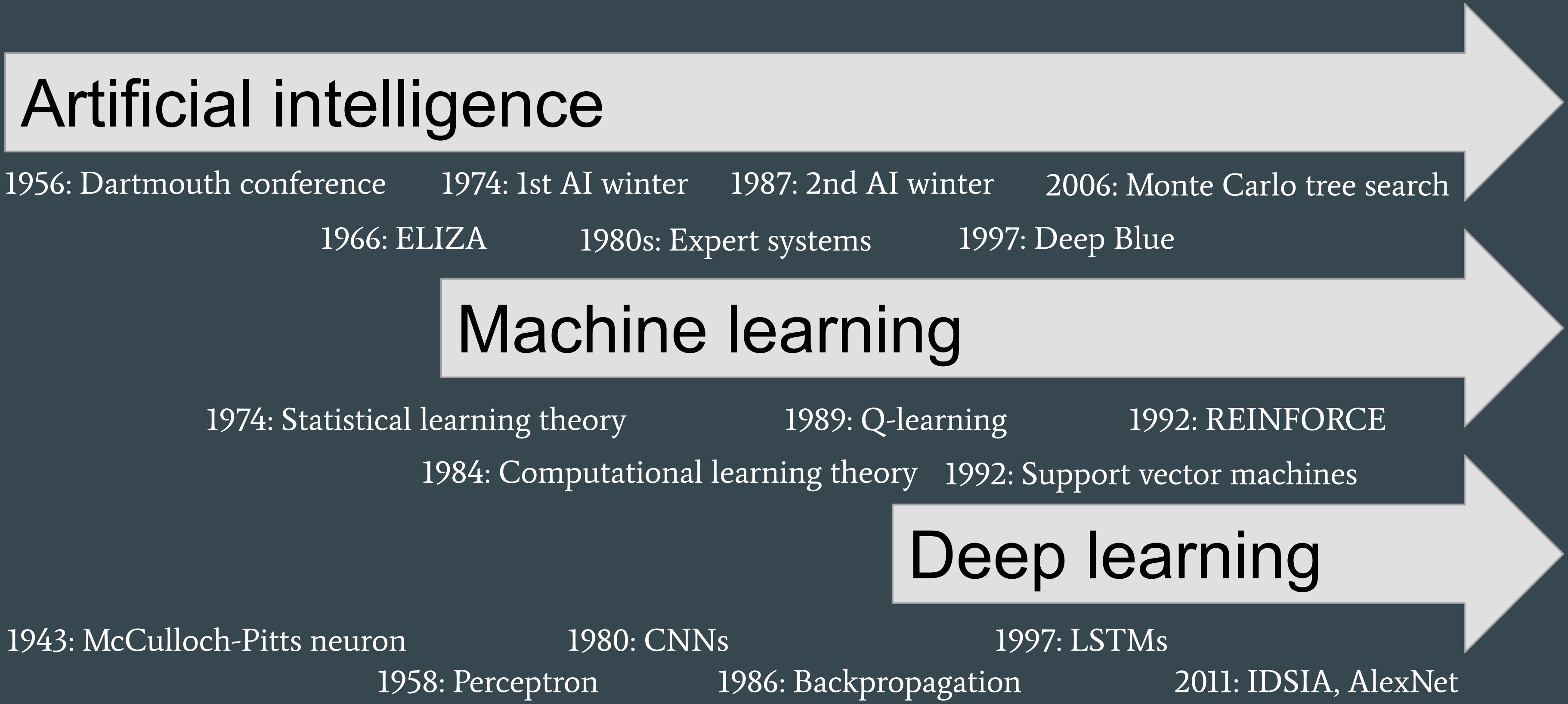
Optimisation techniques

Gradient descent
Backpropagation
Momentum

Neural network techniques

Convolutional networks
Recurrent networks
Transformers

A brief timeline of AI



Artificial intelligence

1956: Dartmouth conference

1966: ELIZA

1974: 1st AI winter

1980s: Expert systems

1987: 2nd AI winter

1997: Deep Blue

2006: Monte Carlo tree search

Machine learning

1974: Statistical learning theory

1984: Computational learning theory

1989: Q-learning

1992: REINFORCE

1992: Support vector machines

Deep learning

1943: McCulloch-Pitts neuron

1958: Perceptron

1980: CNNs

1986: Backpropagation

1997: LSTMs

2011: IDSIA, AlexNet

Supervised learning as function approximation

Start with (x, y) datapoints

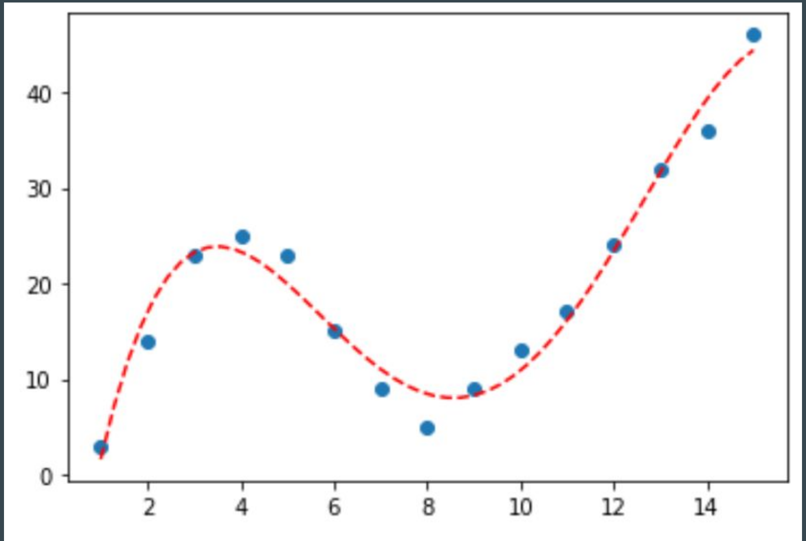
E.g. given age, predict happiness

Define a class of models, and associated parameters

E.g. $y = ax^3 + bx^2 + cx + d$

Then learn the best parameters.

E.g. find a, b, c, d to make curve fit data



Supervised learning as function approximation

Start with (x, y) datapoints

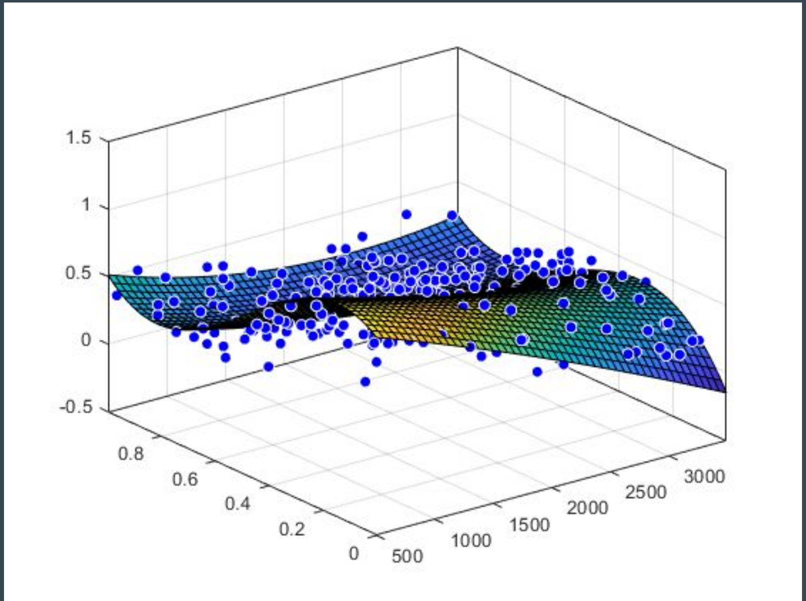
E.g. given age, predict happiness

Define a class of models, and associated parameters

E.g. $y = ax^3 + bx^2 + cx + d$

Then learn the best parameters.

E.g. find a, b, c, d to make curve fit data

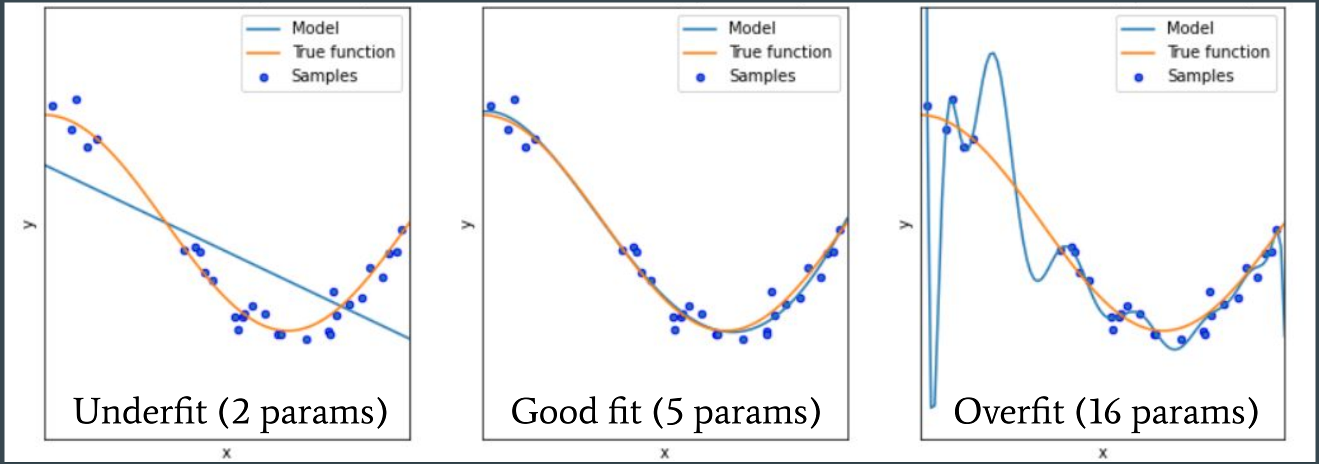


Learning parameters

Repeatedly:

- Score model parameters on some training set data
- Update parameters to perform better on that training set data

Possible results:



Underfit (2 params)

Good fit (5 params)

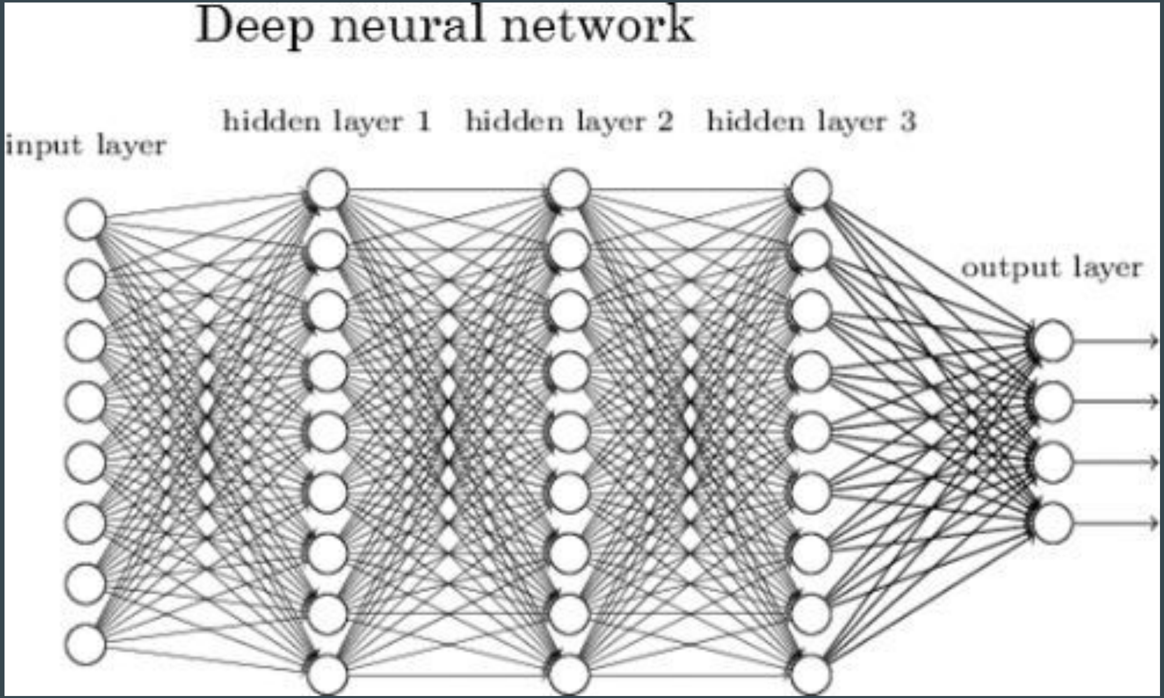
Overfit (16 params)

Deep learning as a new paradigm

Inputs: words, pixels, sensors

Output layers: predictions, generated samples, behaviour

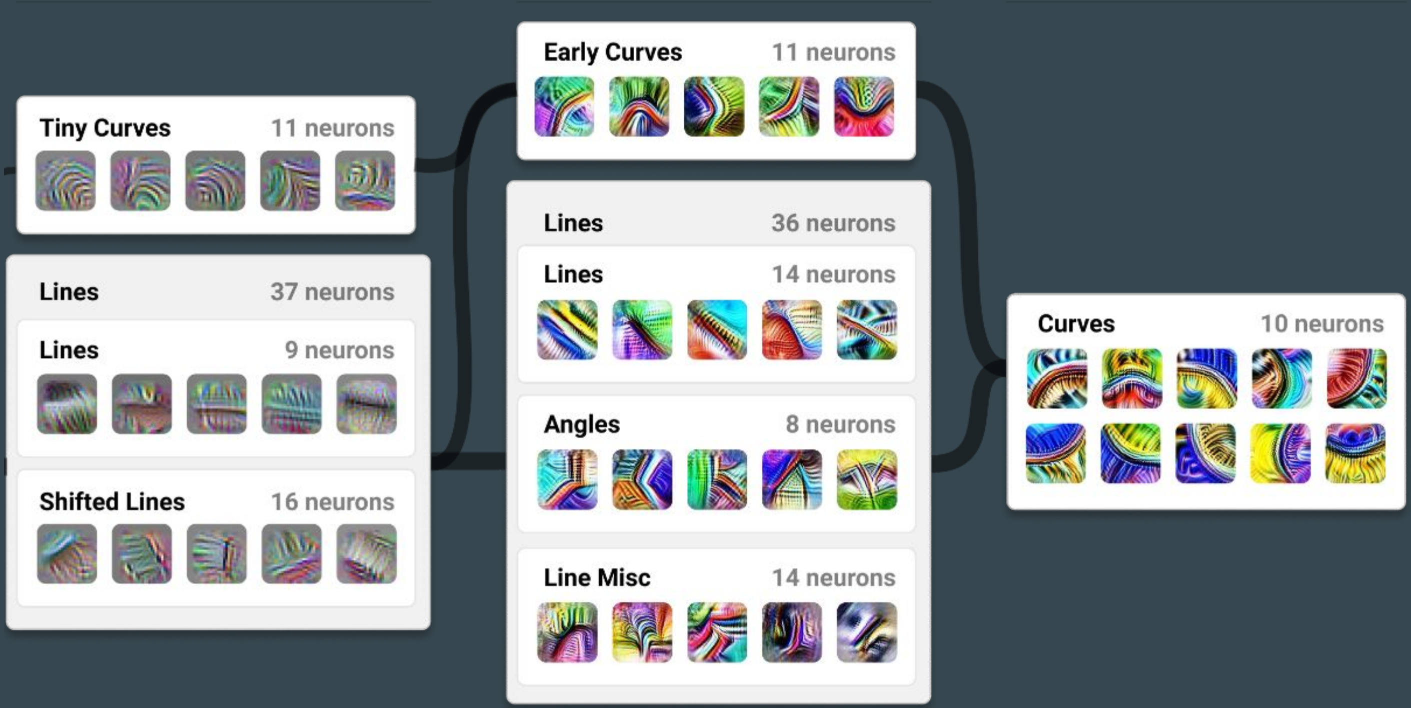
Neural networks have billions of parameters, but don't overfit nearly as much as expected!



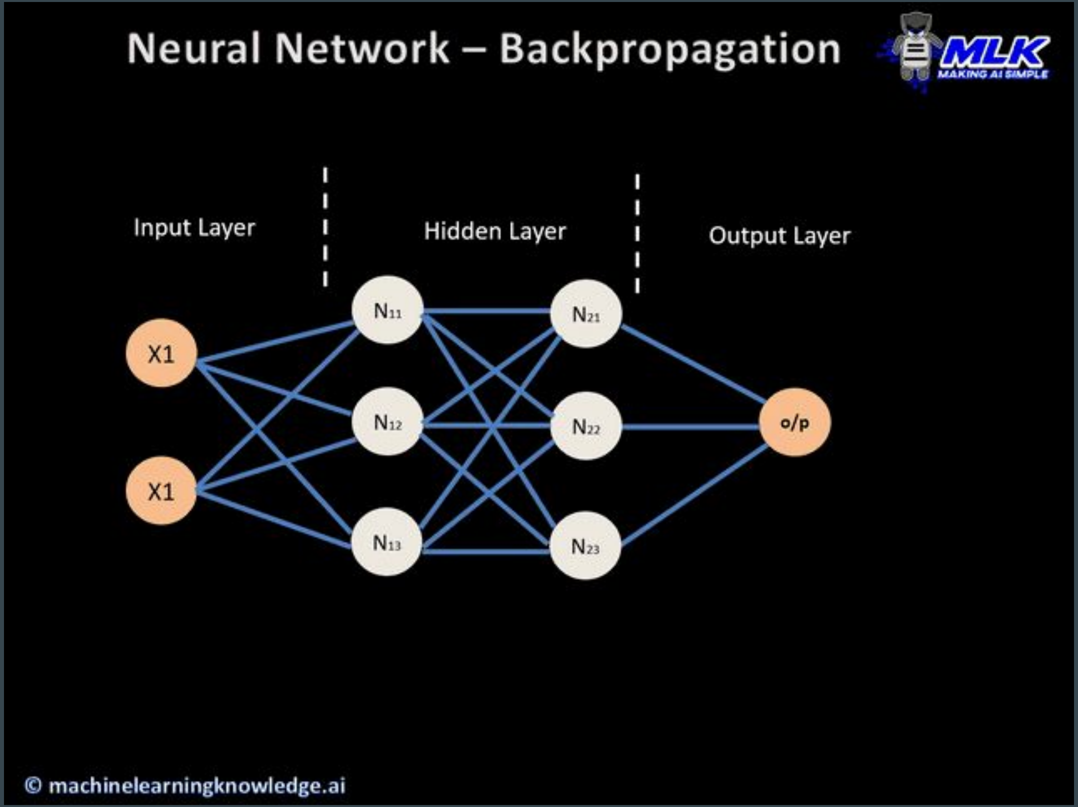
Deep learning as representation learning

Hidden layers: higher-level representations

Brains do the same! (Although with more complex neural architectures.)

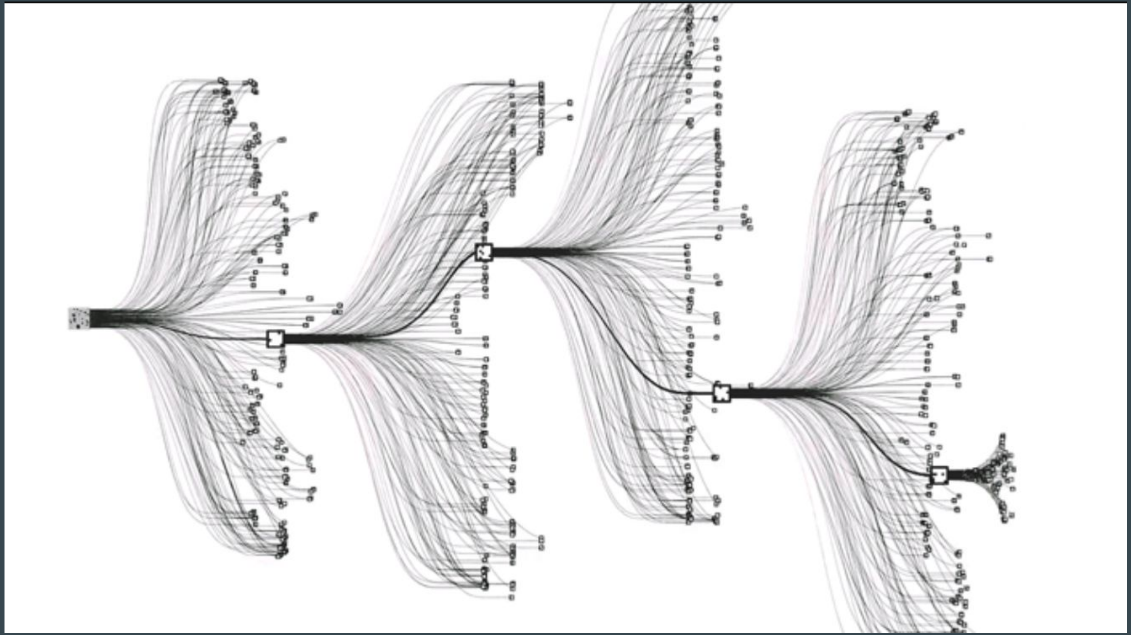


Training neural networks via backpropagation



Reinforcement learning as credit assignment

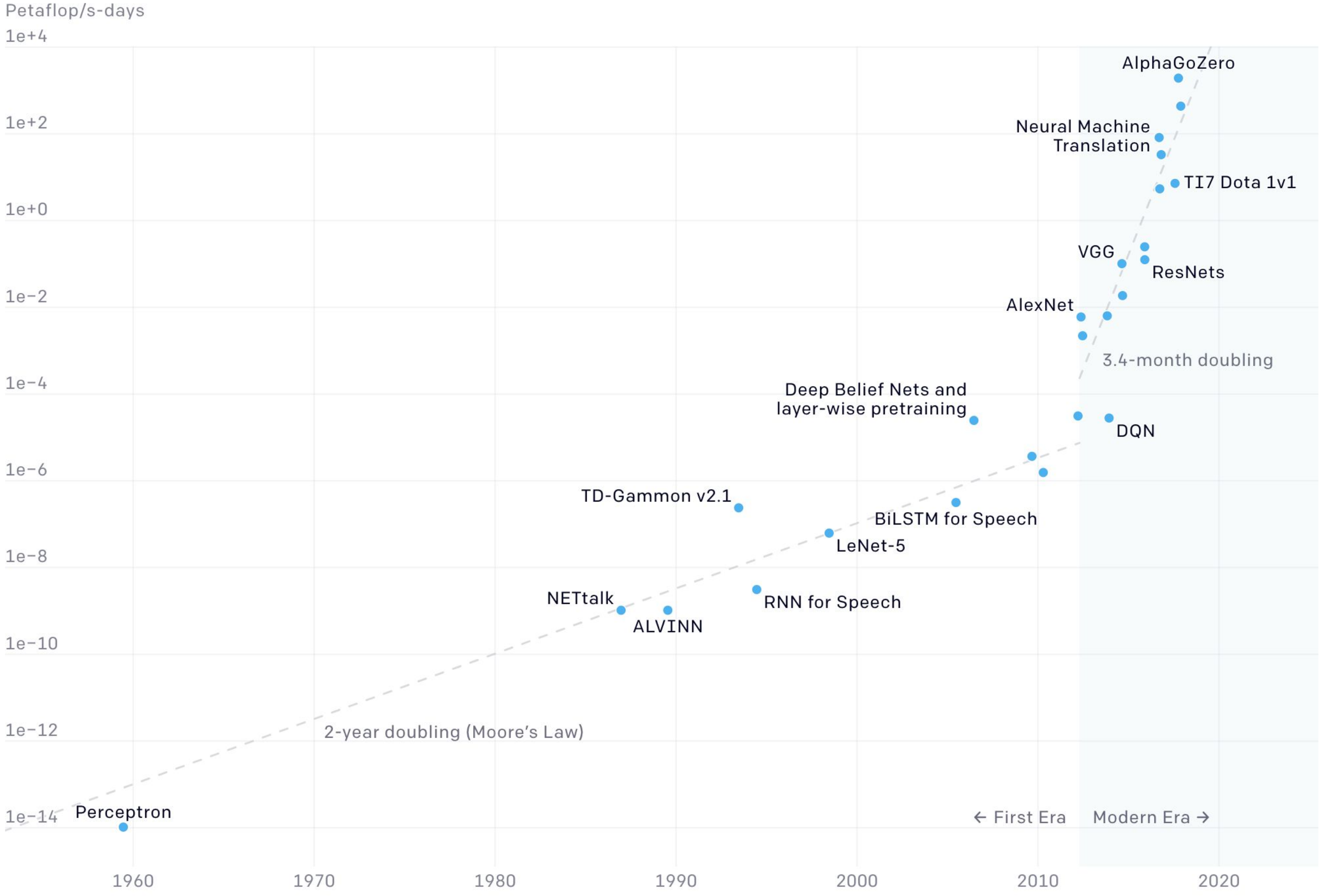
- What happens when feedback is given much after action?
- Need to calculate which actions were responsible for which feedback (credit assignment)
- Algorithm for doing so in the exercises



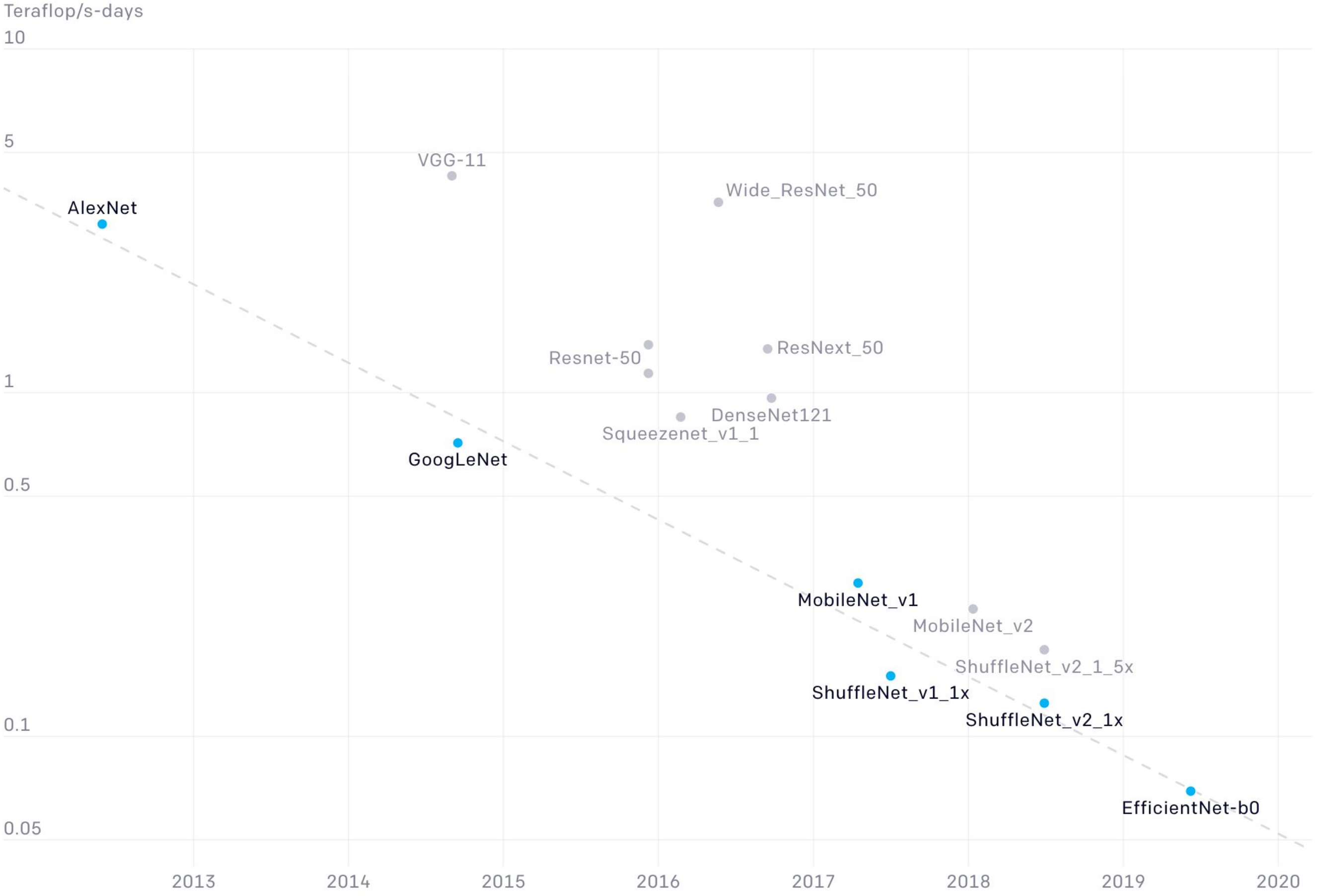
Three ingredients of AI progress

- Compute
- Algorithms
- Data

Two Distinct Eras of Compute Usage in Training AI Systems



44x less compute required to get to AlexNet performance 7 years later (log scale)



Key points in modern deep learning

- 2012: AlexNet wins ImageNet image recognition competition
- 2014: Deep reinforcement learning used to play Atari games
- 2016: AlphaGo beats Lee Sedol at Go
- 2017: Transformers released
- 2019: AlphaStar and OpenAI Five beat professionals at Starcraft and DOTA
- 2020: GPT-3 released
- 2020: AlphaFold solves protein folding
- 2021: OpenAI Codex released