高等机器学习

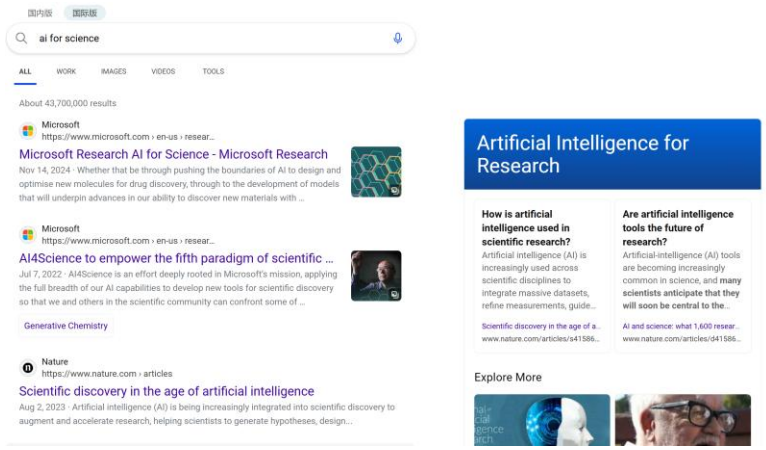
Microsoft

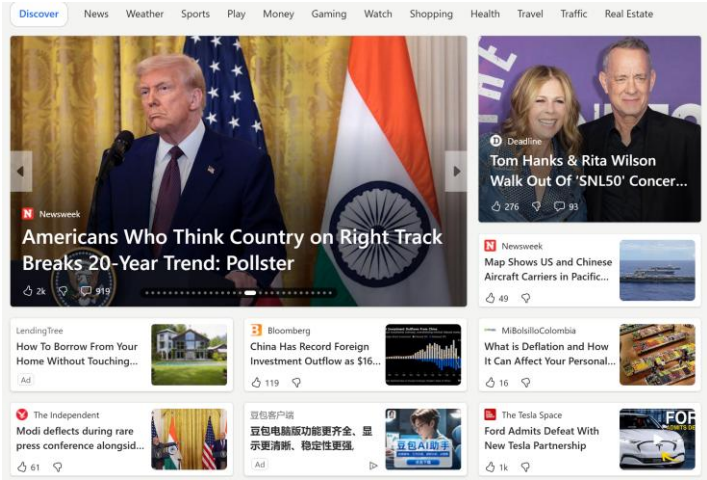
课程导论

秦涛
微软研究院

Why Machine Learning?

Prediction Tasks





Search: Understand user intent, find relevant info, generate and rank results

Express delivery: predict demand and pre-allocate vehicles for package transportation

Reading news: predict interests of users and recommend related news

How to Predict?

- Using hand-crafted rules:

(conditioned on a pattern)

(take an action)

You read sports news yesterday

DeepSeek is very hot recently

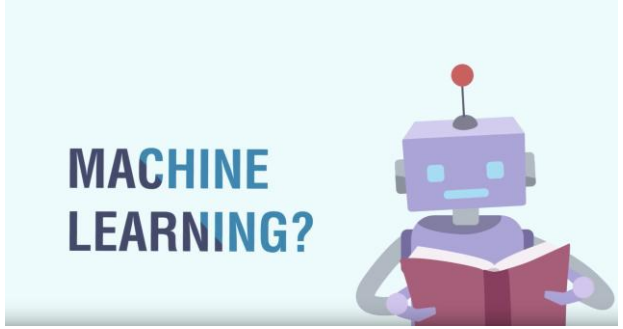
There are many packages from Beijing to Shanghai last week

Recommend sports news to you today

Rank DeepSeek in top positions

Pre-allocate more tracks from Beijing to Shanghai this week

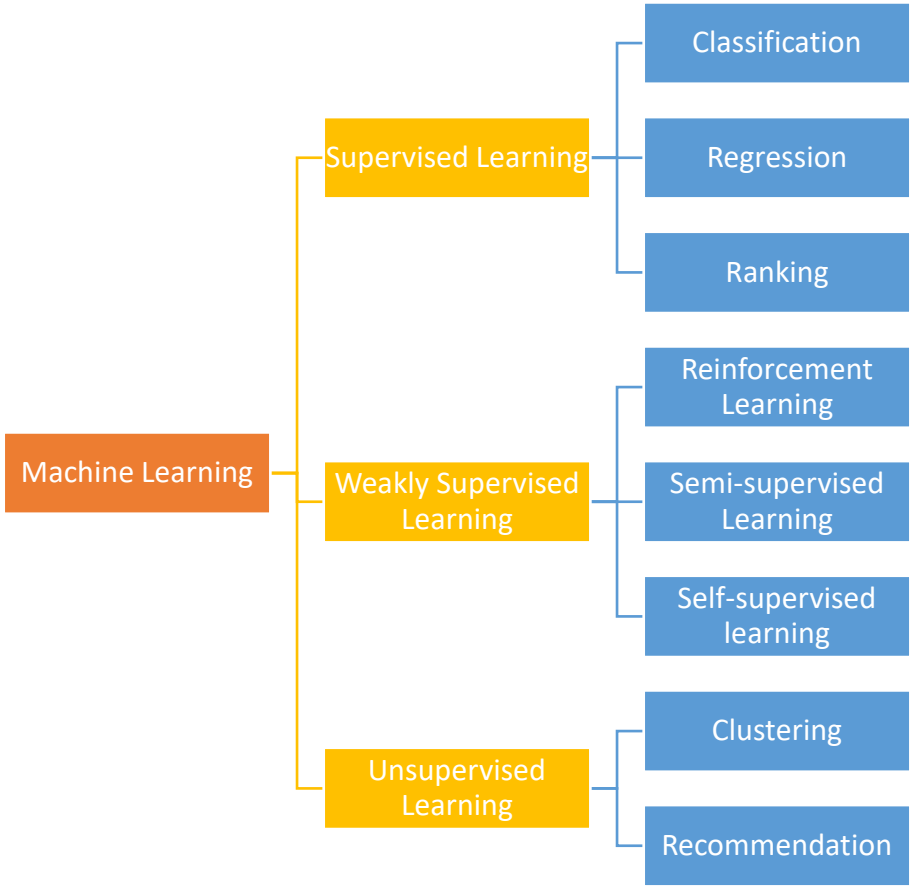
Limitation of Rule-based Solution



- Inaccurate:
  - 80% (regular) vs 20% (exceptional)
- Non-scalable:
  - Human efforts required to deal with new tasks or changes of old tasks
- How to do better?
  - Automatic learning prediction models (patterns → actions) from data

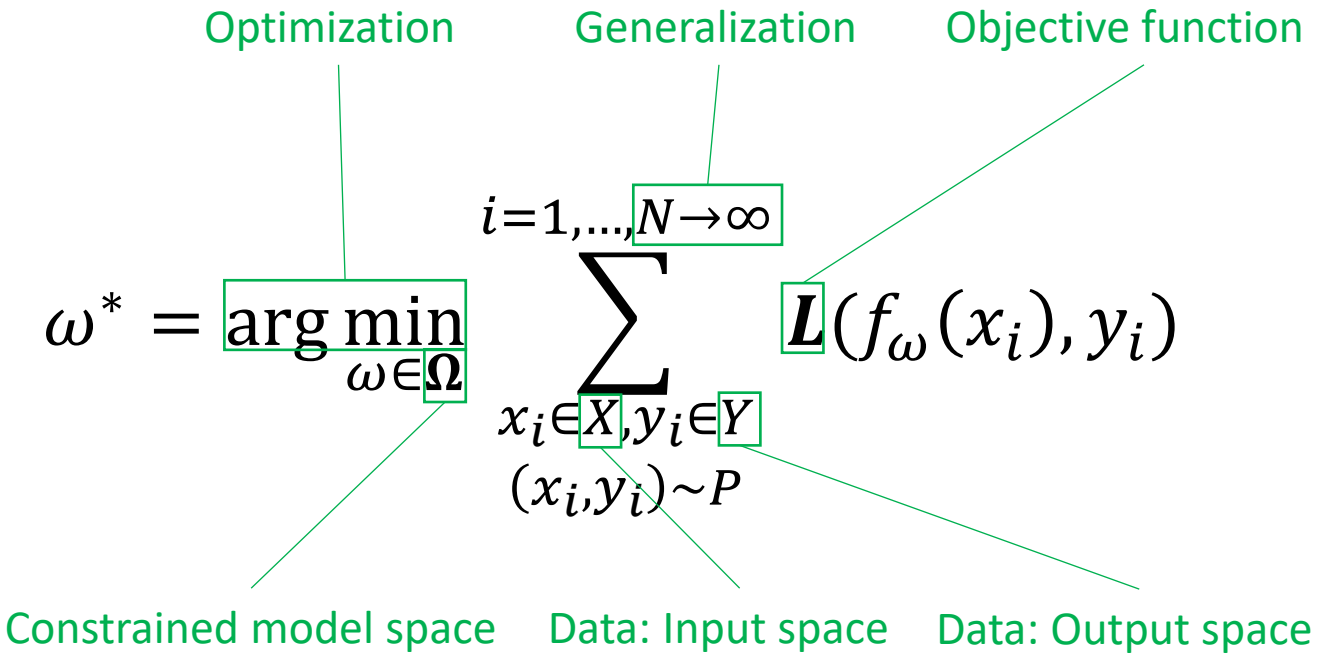
Machine Learning

- **[Narrow]**
  - Machine learning learns a prediction model (pattern → action) from given examples, according to certain objective function, which can be used to deal with future unknown problems of the same kind.
- **[Broad, or AI in general]**
  - Machine learning is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.



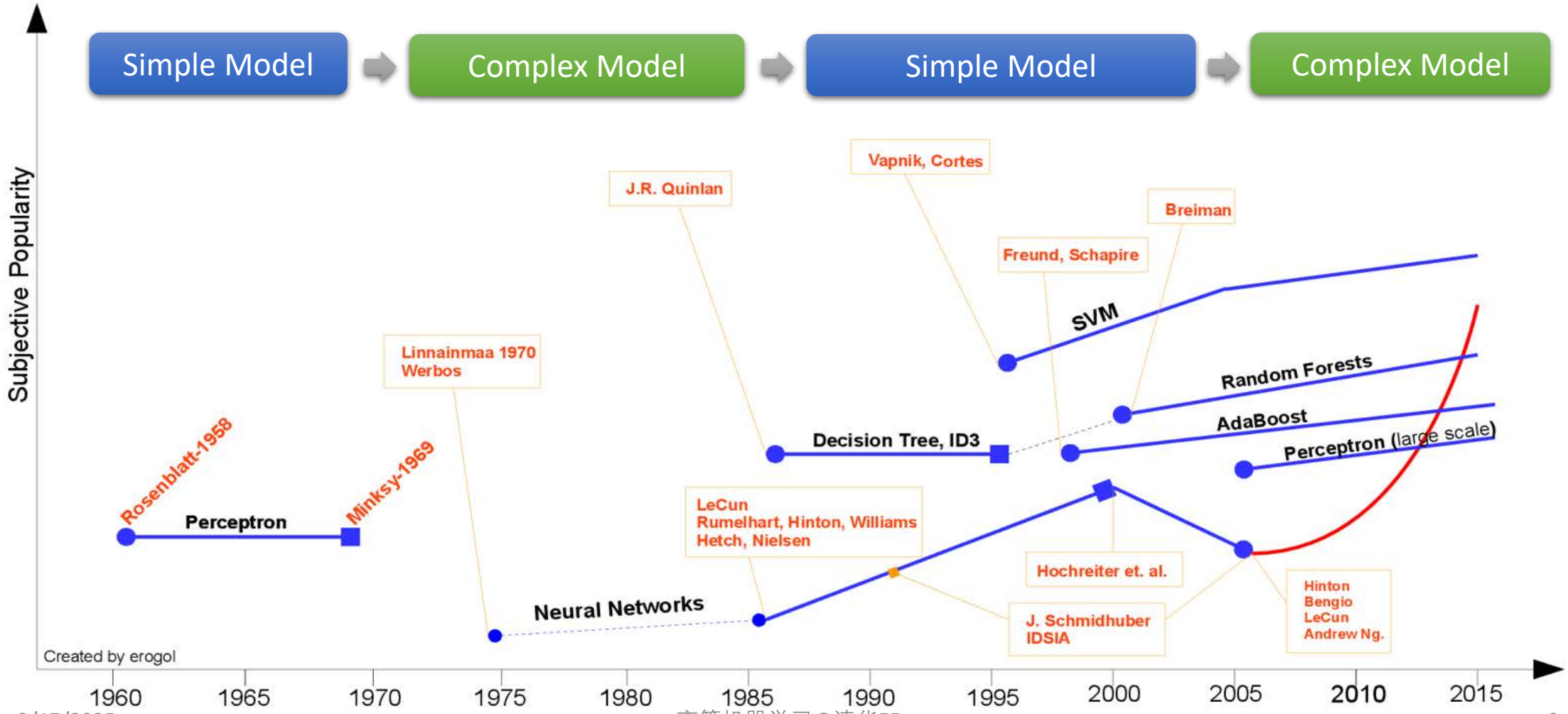
One Formula for (Supervised) Machine Learning

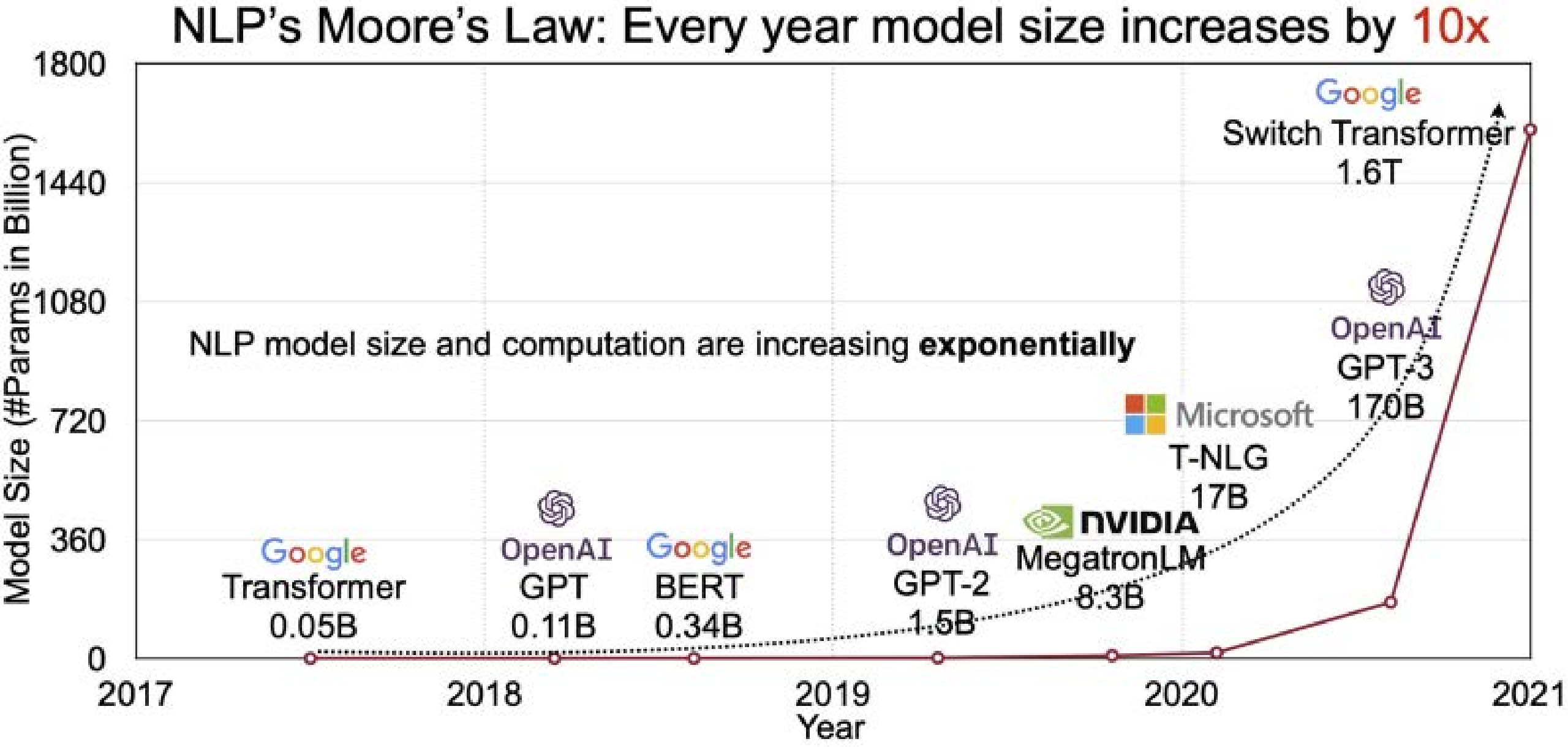
- Learning model from example
- Generalizing to unknown case
- Objective-oriented
- Constrained optimization



A Brief History of Machine Learning

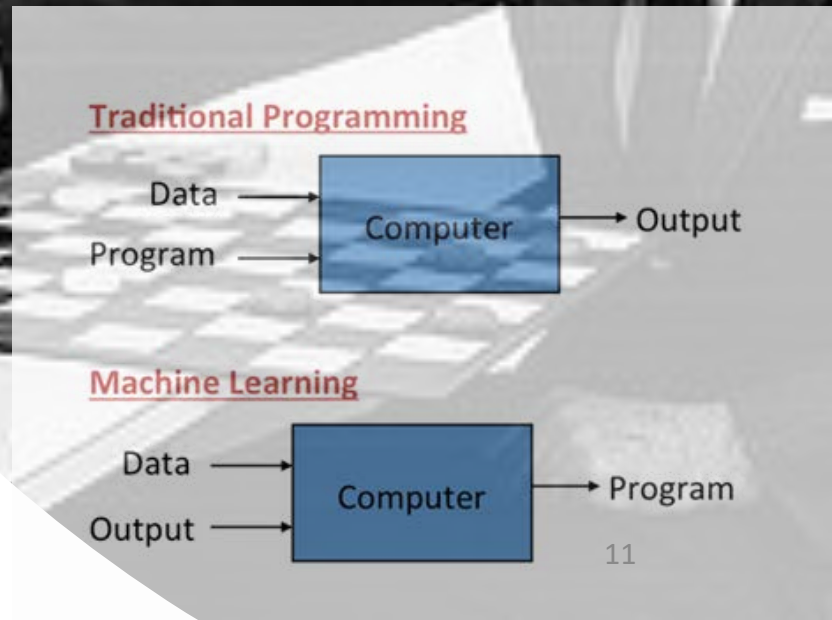
A Brief History of Machine Learning




NLP's Moore's Law: Every year model size increases by 10x
Model Size (#Params in Billion)
1800
1440
1080
720
360
0
2017
2018
2019
2020
2021
Year
NLP model size and computation are increasing exponentially
Google
Transformer
0.05B
OpenAI
GPT
0.11B
Google
BERT
0.34B
OpenAI
GPT-2
1.5B
NVIDIA
MegatronLM
8.3B
Microsoft
T-NLG
17B
OpenAI
GPT-3
170B
Google
Switch Transformer
1.6T

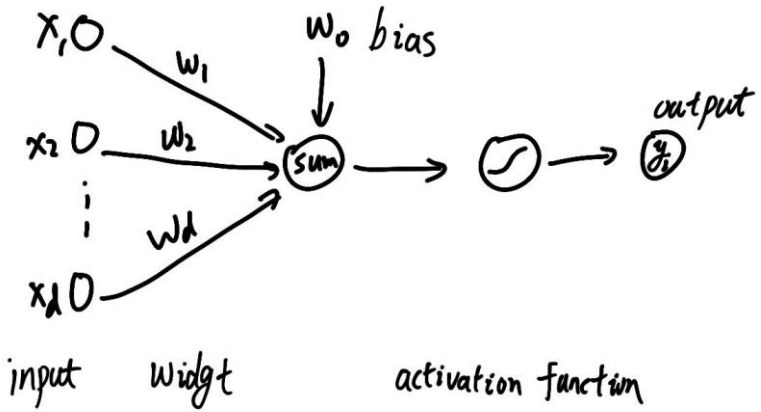
Arthur Samuel

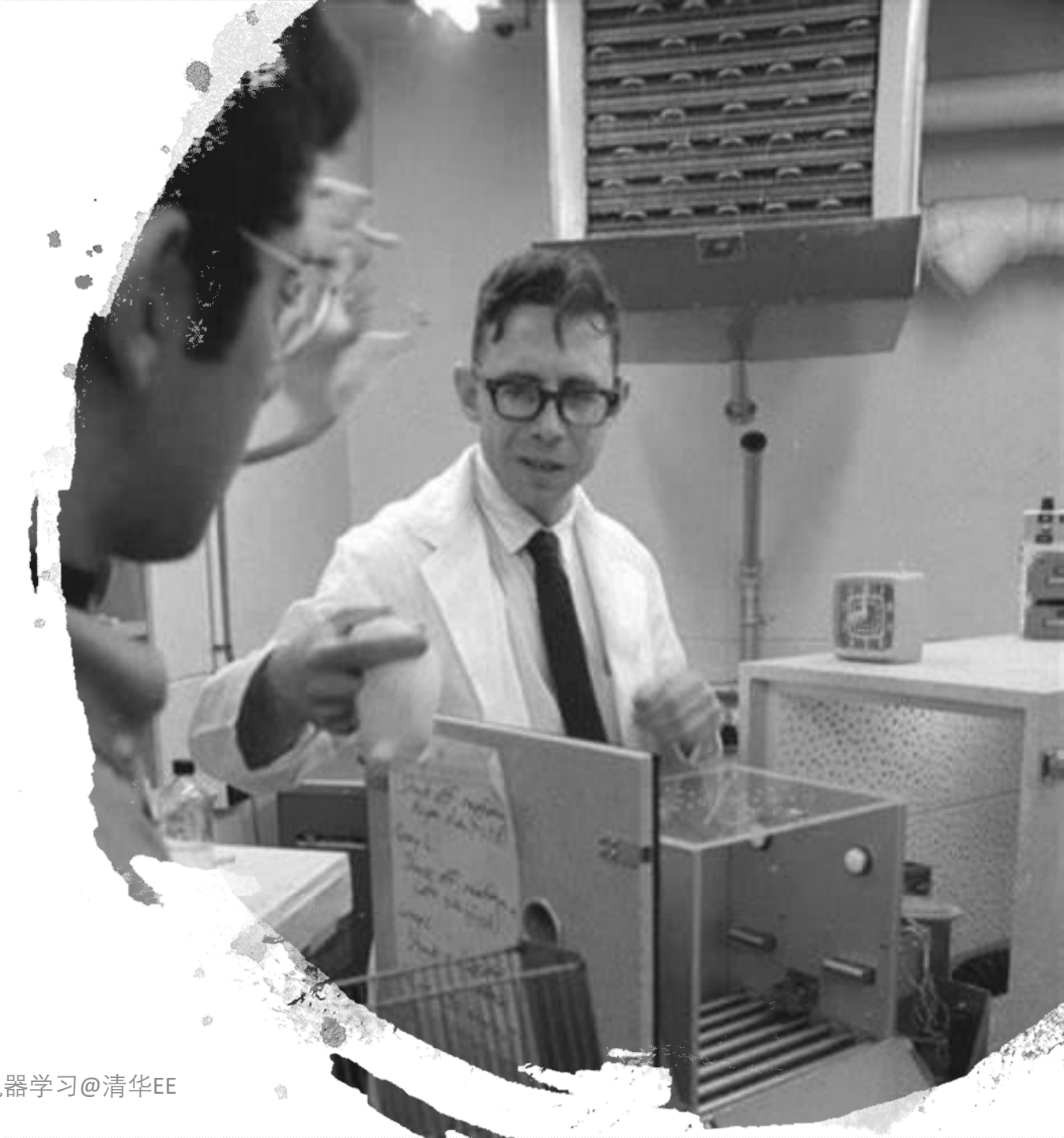
- In 1952, Arthur Samuel, developed a program playing Checkers.
  - The program was able to observe positions and learn an implicit model that gives better moves for the latter cases.
  - With that program, Samuel clamed that machines can go beyond the written codes and learn patterns like human-beings.
- Samuel coined the concept of “machine learning” in 1959.

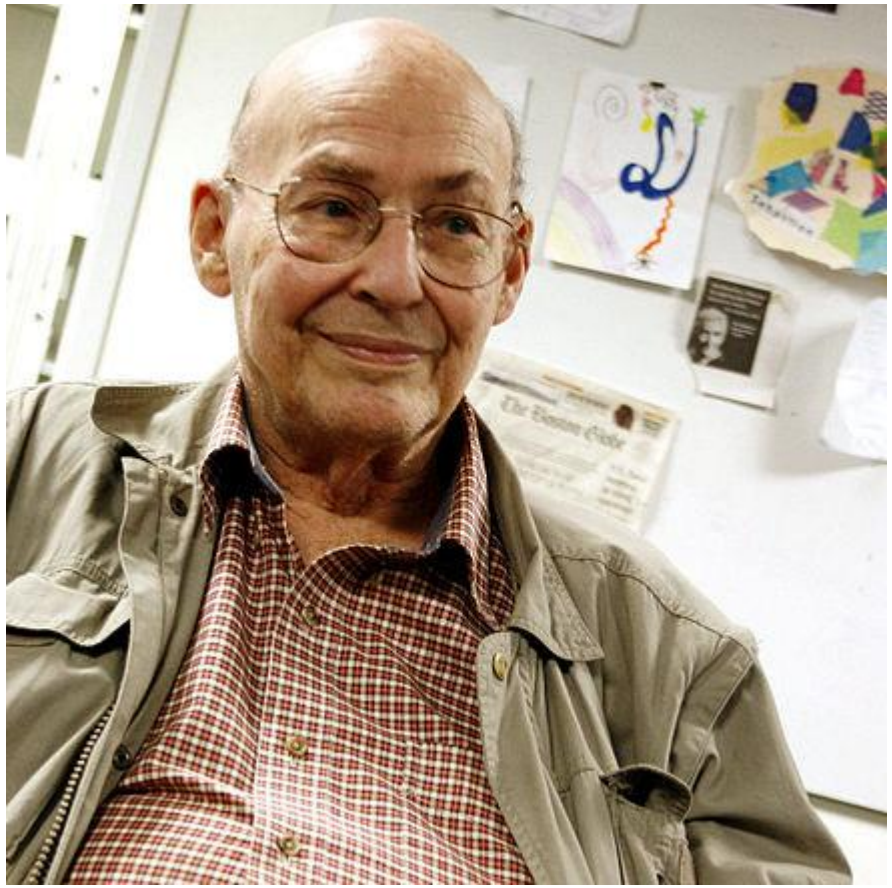


Frank Rosenblatt

- In 1957, Frank Rosenblatt designed the first neural network for computers (the perceptron), which simulates the thought processes of the human brain.

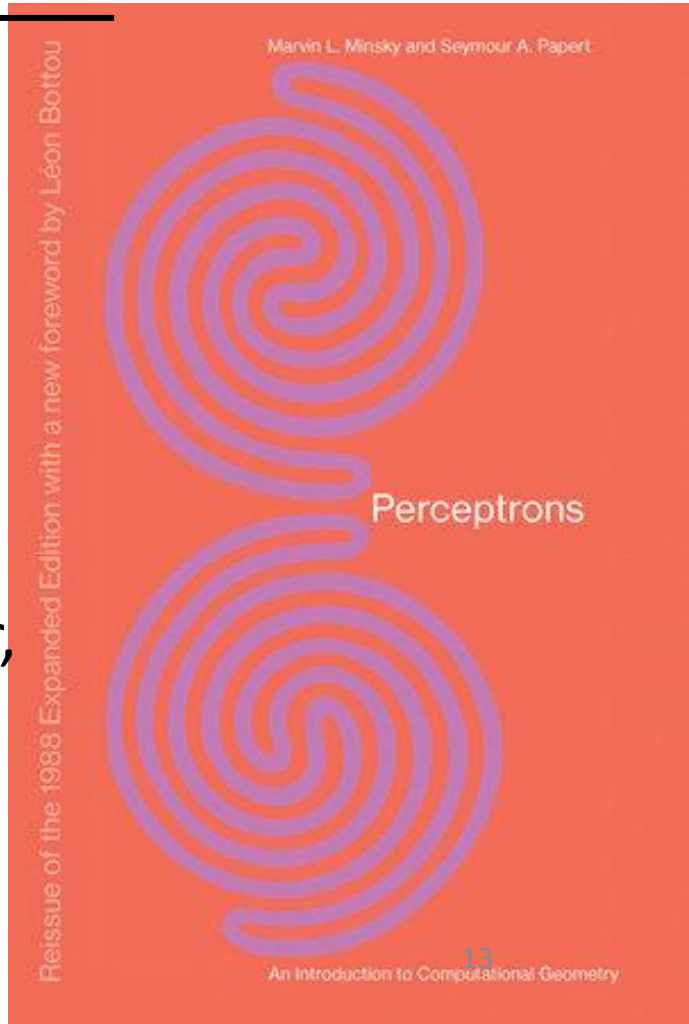






Marvin Minsky

- In 1969, Minsky proposed the famous **XOR** problem and the inability of *Perceptron* in such linearly inseparable data distributions.
- It was the Minsky's tackle to the NN community. Thereafter, NN researches would be dormant up until 1980s.

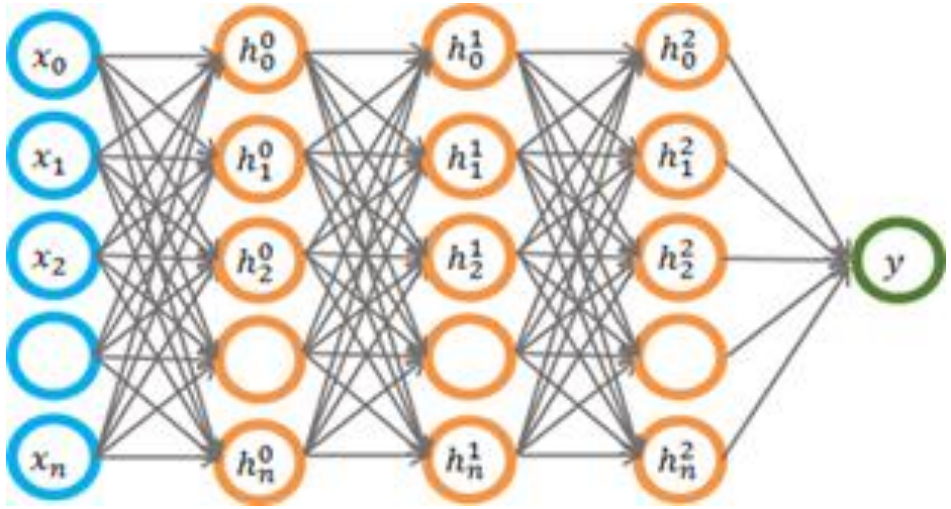


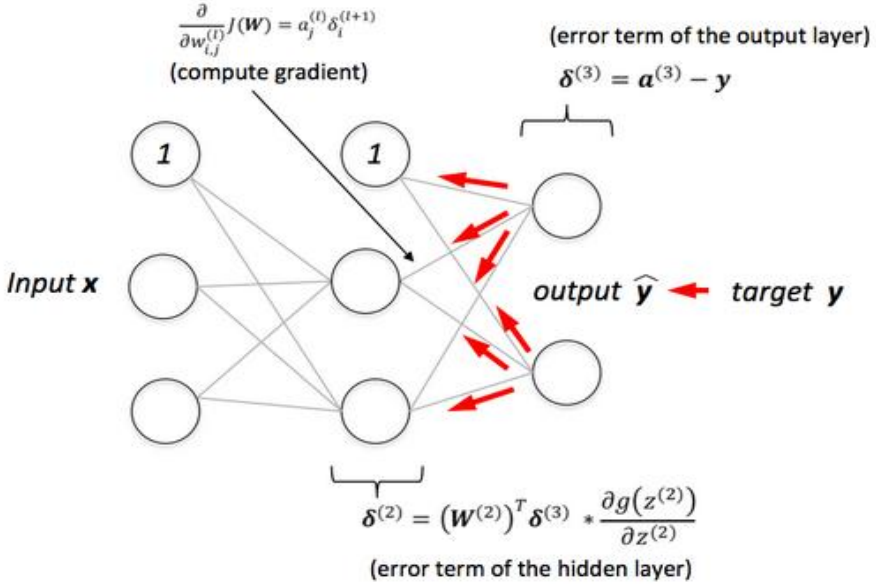
Perceptron is too simple, more complicated models are needed to handle complex problems…

Paul Werbos

- Paul Werbos suggested using Multi-Layer Perceptron (MLP) in 1981, and proposed the Backpropagation (BP) algorithm for training neural networks. This new architecture solved the XOR challenge.
- Following Werbos' new ideas, neural network researchers successively presented different architectures of MLP and a number of BP variants for effective training.
- Was a pioneer of recurrent neural networks

Multi-layer Perceptron / Deep Neural Networks





Universal Approximation Theorem

- A feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets of R^n, under mild assumptions on the activation function.

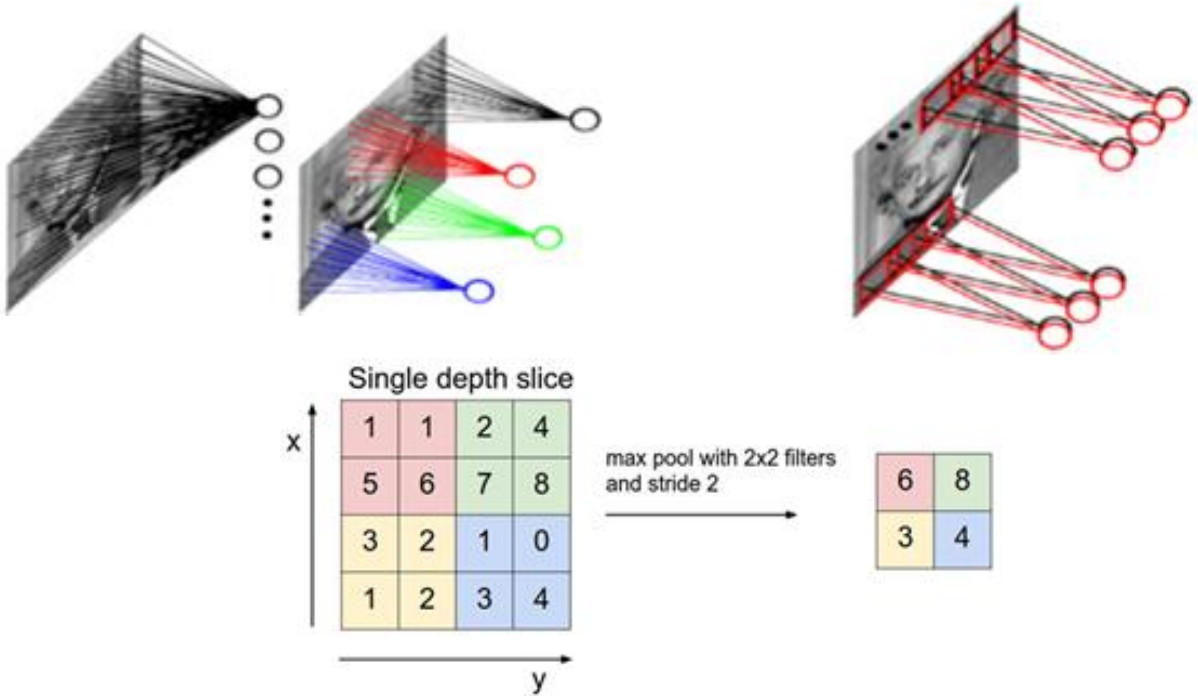
Geoffrey Hinton, Yan LeCun, Jurgen Schmidhuber

- Geoffrey Hinton contributed a lot to the practical backpropagation algorithms (1986) and Boltzmann Machines (1983).
- Yan LeCun was the first to train a convolutional neural network on images of handwritten digits (1986).
- Jurgen Schmidhuber invented a new type of recurrent neural network called Long short-term memory or LSTM (1997), which has its profound impact on speech recognition and natural language processing.

Convolutional Neural Networks

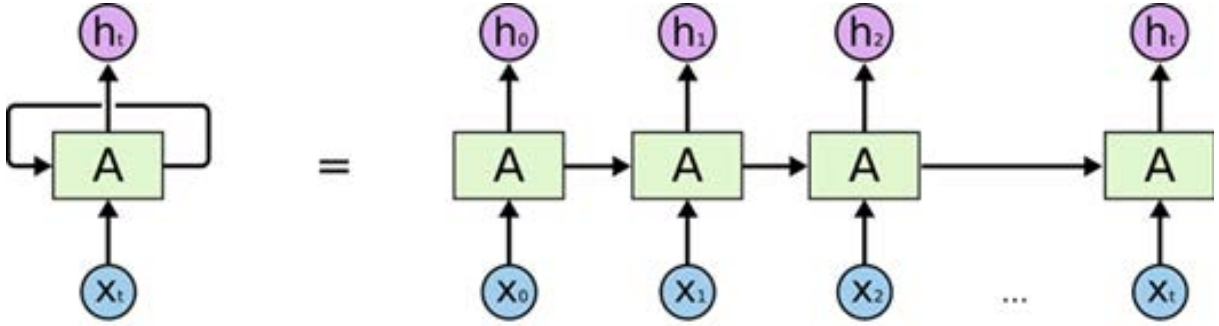
- Inspired by biology:
  - The visual cortex contains cells that are sensitive to small sub-regions, tiled to cover the entire visual field. These cells act as local filters over the input space and are well-suited to exploit the strong spatially local correlation present in natural images.

- Convolution:
  - Local connection, pattern recognition
- Weight sharing and pooling
  - Invariance
  - Parameter efficiency



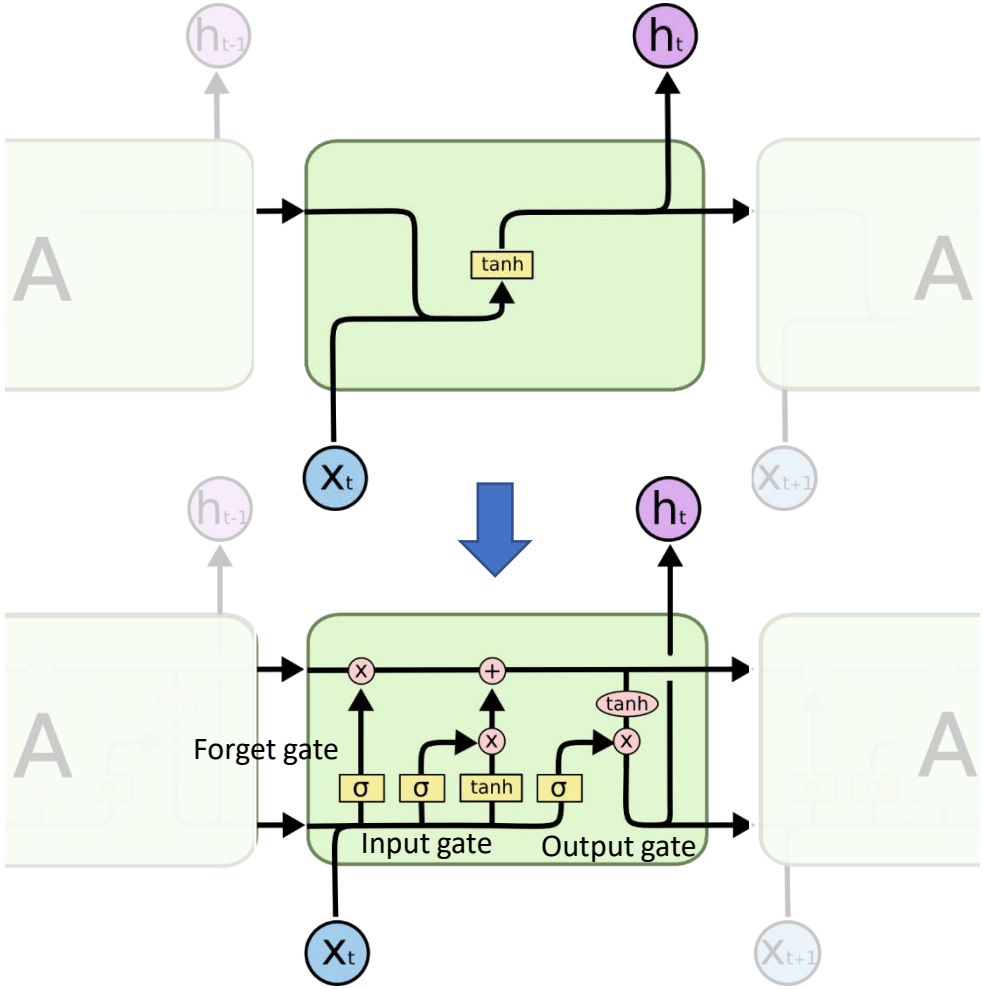
Recurrent Neural Networks (RNN)

- Motivations
  - We don’t throw everything away and start thinking from scratch every time. Our thoughts have persistence. However, standard DNN and CNN do not a mechanism to remember things.
  - RNN contains feedback connection, so the activations can flow round in a loop and enable the networks to do temporal processing and learn sequences. 。
- Model a dynamic system driven by an external signal x
  - $A_t = f(Ux_t + WA_{t-1})$
  - Hidden node A_{t-1} contains information about the whole past sequence
  - function $f(\cdot)$ maps the whole past sequence $(x_t, \dots, x_1)$ to current state A_t



Long Short Term Memory (LSTM)

- Control information flow with gate functions, in order to avoid gradient vanishing or exploding along the long path of RNN
- Three parameterized gates:
  - Forget gate: govern the direct flow across layers
  - Input gate
  - Output gate



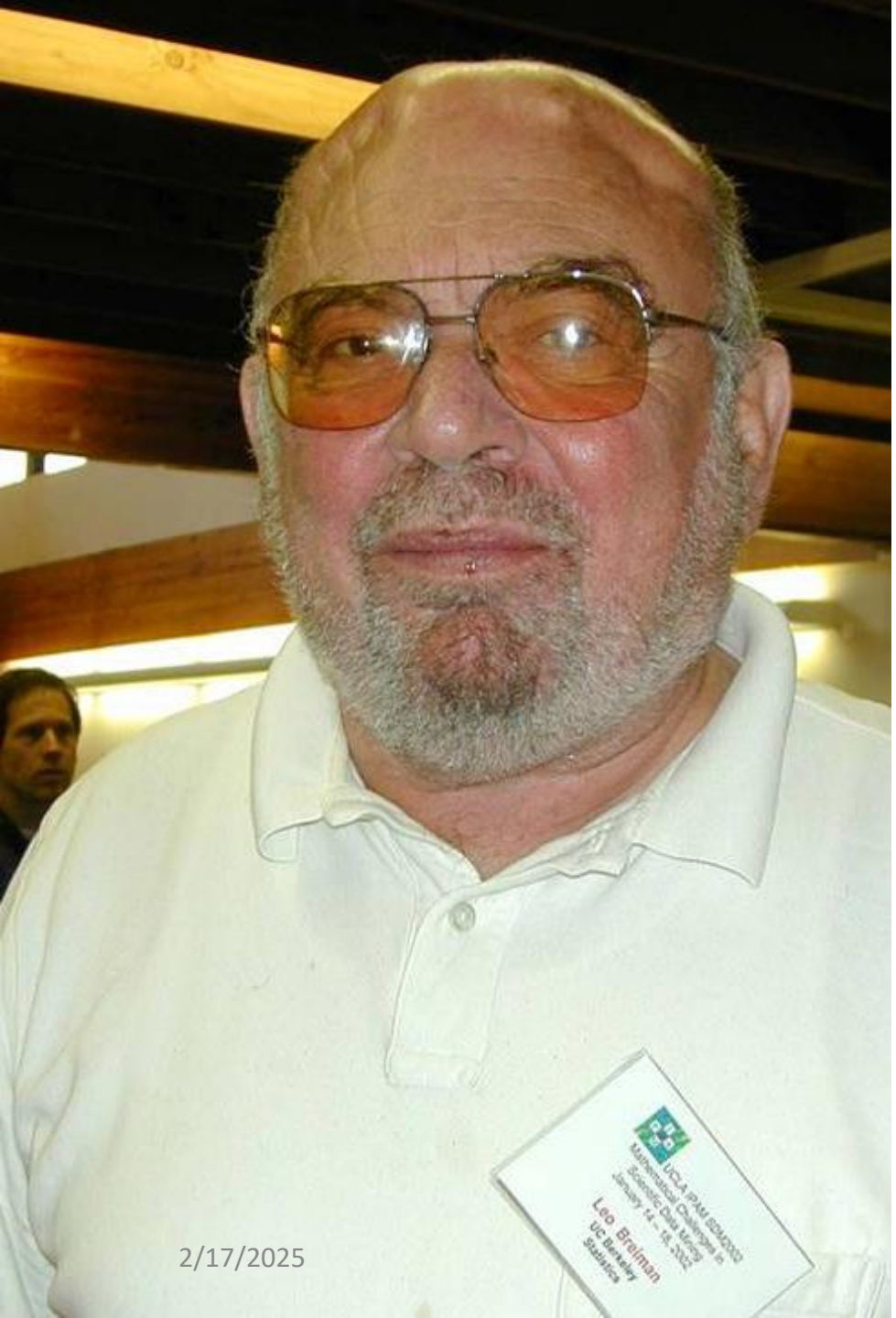
Neural networks are black boxes, and therefore difficult to interpret…

Ross Quinlan

- Decision trees were proposed by Ross Quinlan, more specifically the ID3 algorithm.
- ID3 is able to find more real-life use case with its simplistic rules and its clear inference.
- After ID3, many different alternatives or improvements have been explored by the community (e.g. ID4, Regression Trees, CART ...) and still it is one of the active topics in ML.

Decision Trees

- ID3 Algorithm
  - Take all unused attributes and count their entropy concerning test samples
  - Choose attribute for which entropy is minimum (or, equivalently, information gain is maximum)
  - Make node containing that attribute

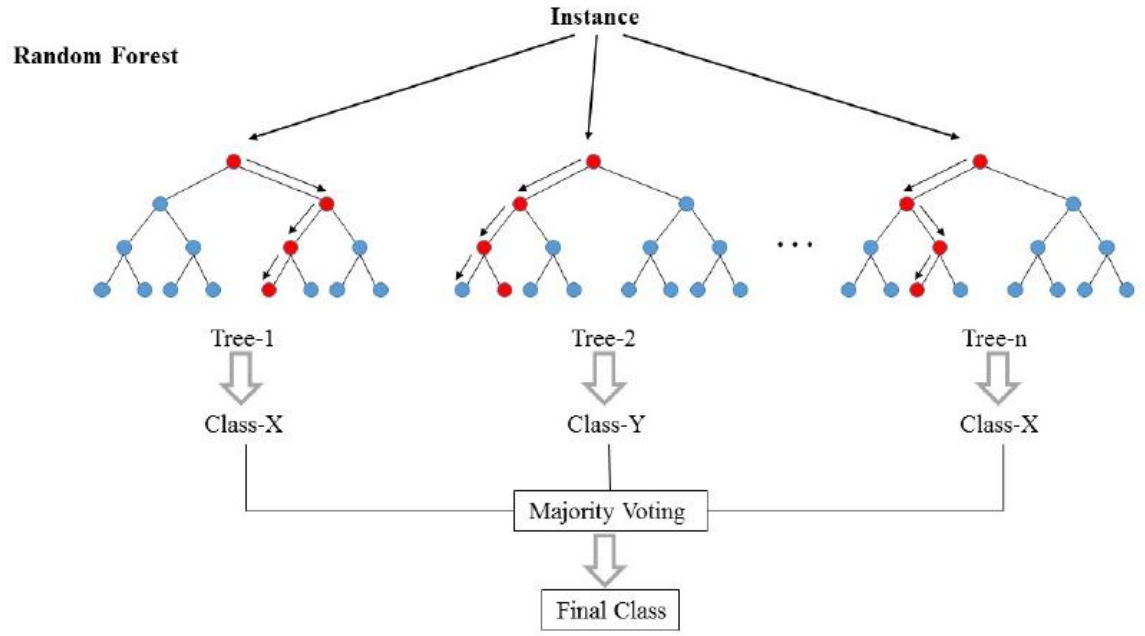


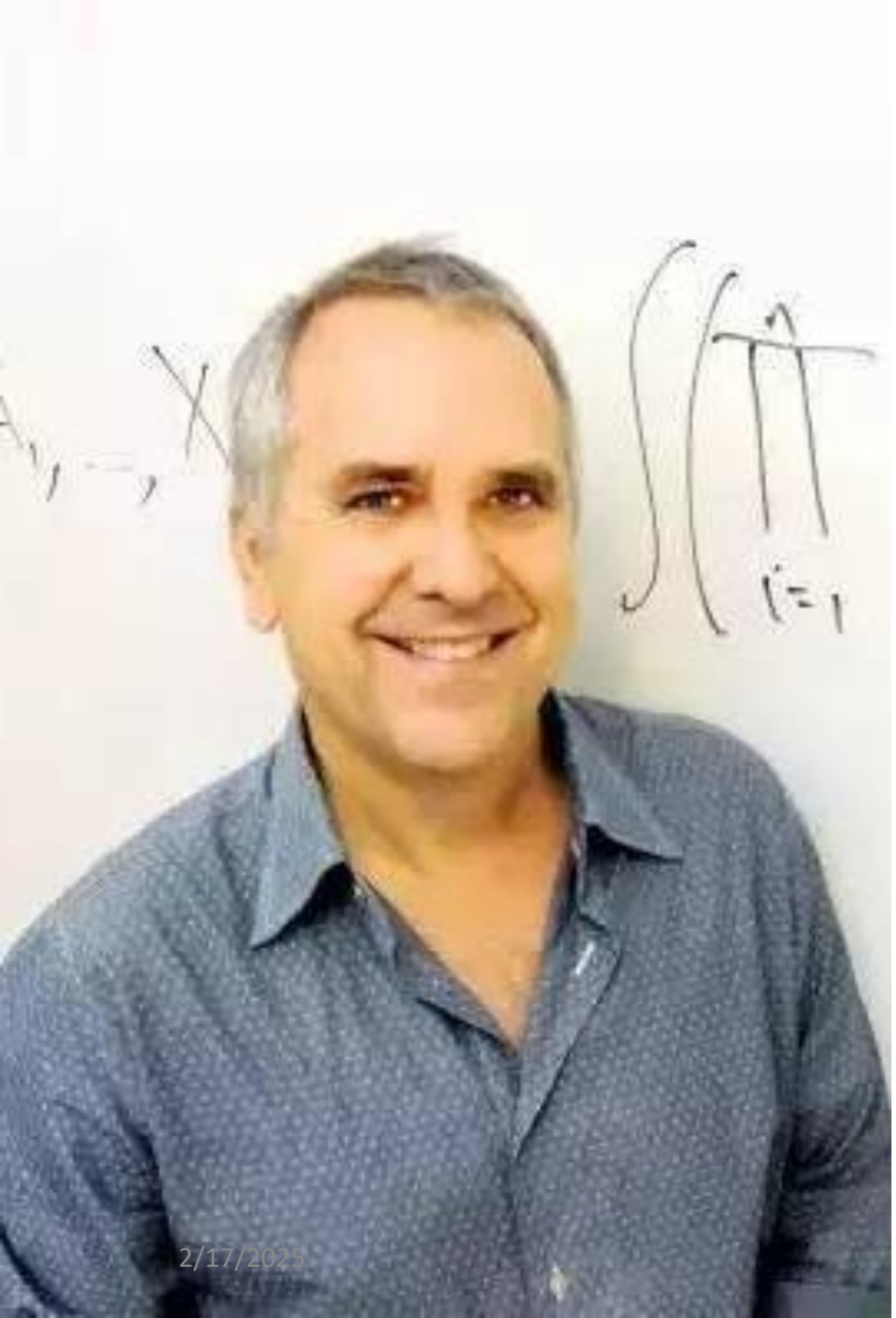
Leo Breiman

- Leo Breiman proposed the Random Forests algorithm in 2001 that ensembles multiple decision trees where each of them is curated by a random subset of instances and each node is selected from a random subset of features.
- RF has theoretical and empirical proofs of endurance against over-fitting
- RF shows its success in many different tasks like Kaggle competitions.

Random Forest

- **Random forest** is an ensemble classifier that consists of many decision trees and outputs the class that is the mode (majority voting) of the class's output by individual trees.
- Principle:
  - Encourage diversity among trees
- Solution:
  - Bagging: Bootstrap aggregation
  - Random decision trees



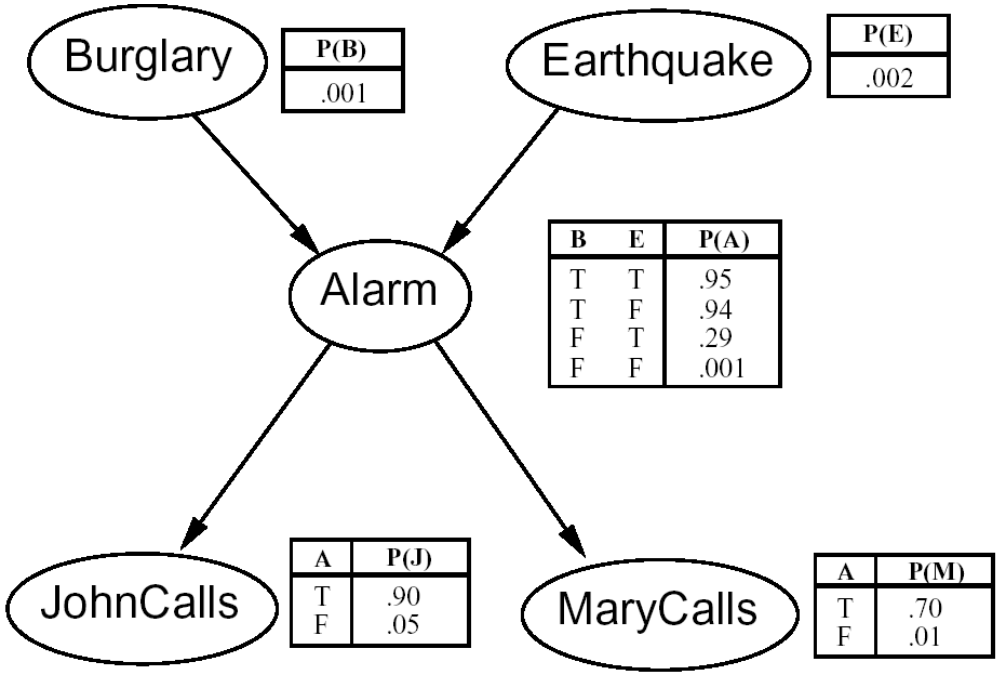


Michael Jordan

- Michael Jordan has wide-spectrum contributions to modern machine learning, especially on Bayesian nonparametric analysis and probabilistic graphical models.
- Many of his students are famous, including Andrew Ng, David Blei, Zoubin Ghahramani, Eric Xing, Percy Liang, and also Yoshua Bengio (postdoc).

Graphical Models

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distribution.
  - Causal Structure
  - Interconnected Nodes
  - Directed Acyclic Links
  - Joint distribution formed from conditional distributions at each node
  - Diagnostic or causal inference



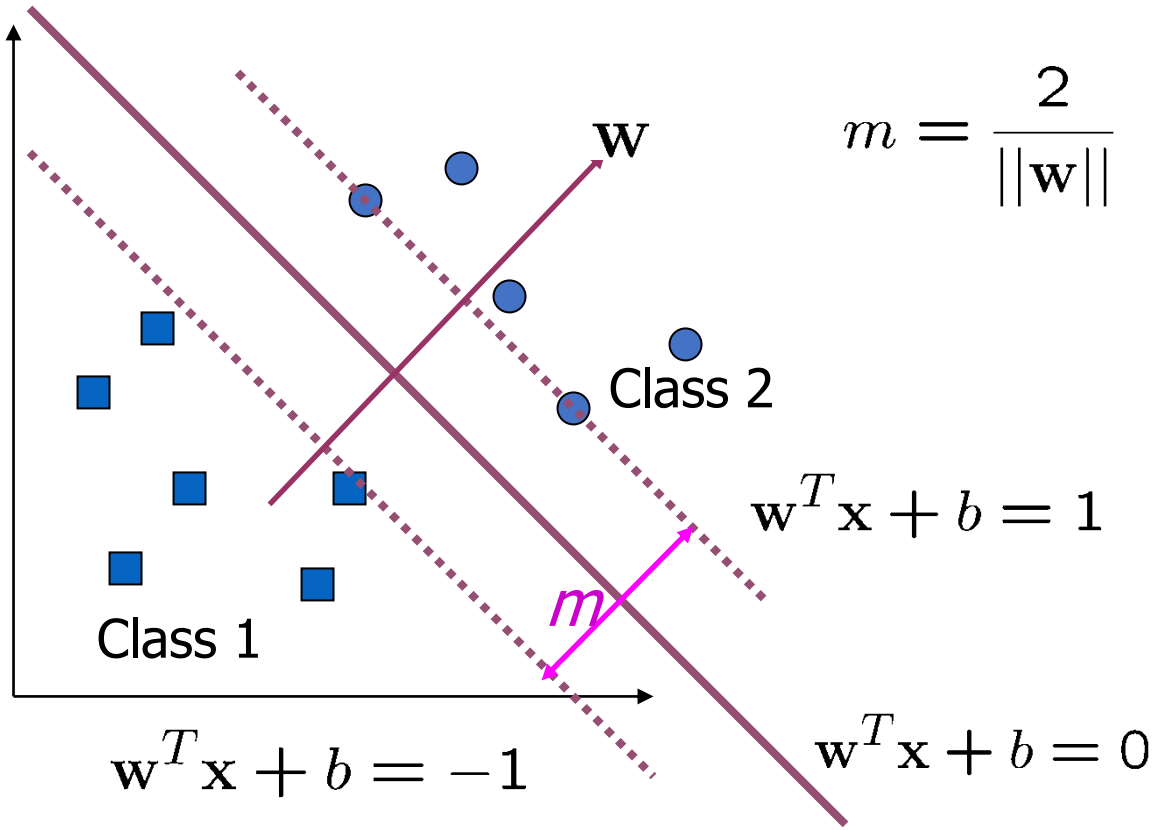
Neural networks are data-hungry. When there are only small number of training data, they will overfit …

Vladimir Vapnik

- Support Vector Machines (SVM) was proposed by Vapnik and Cortes in 1995 with very strong theoretical standing and empirical results.
- SVM got the best of many tasks that were occupied by NN models before. In addition, SVM was able to exploit all the profound knowledge of convex optimization, generalization margin theory and kernels against NN models.
- ML community was separated into two crowds as NN or SVM advocates.

Support Vector Machines

- Basic idea
  - The decision boundary should be as far away from the data of both classes as possible
  - We should maximize the margin *m*
- *SVM could be efficiently solved in its dual form, whose solutions only rely on the so-called support vectors.*
- *SVM could be kernelized to handle non-separable cases*

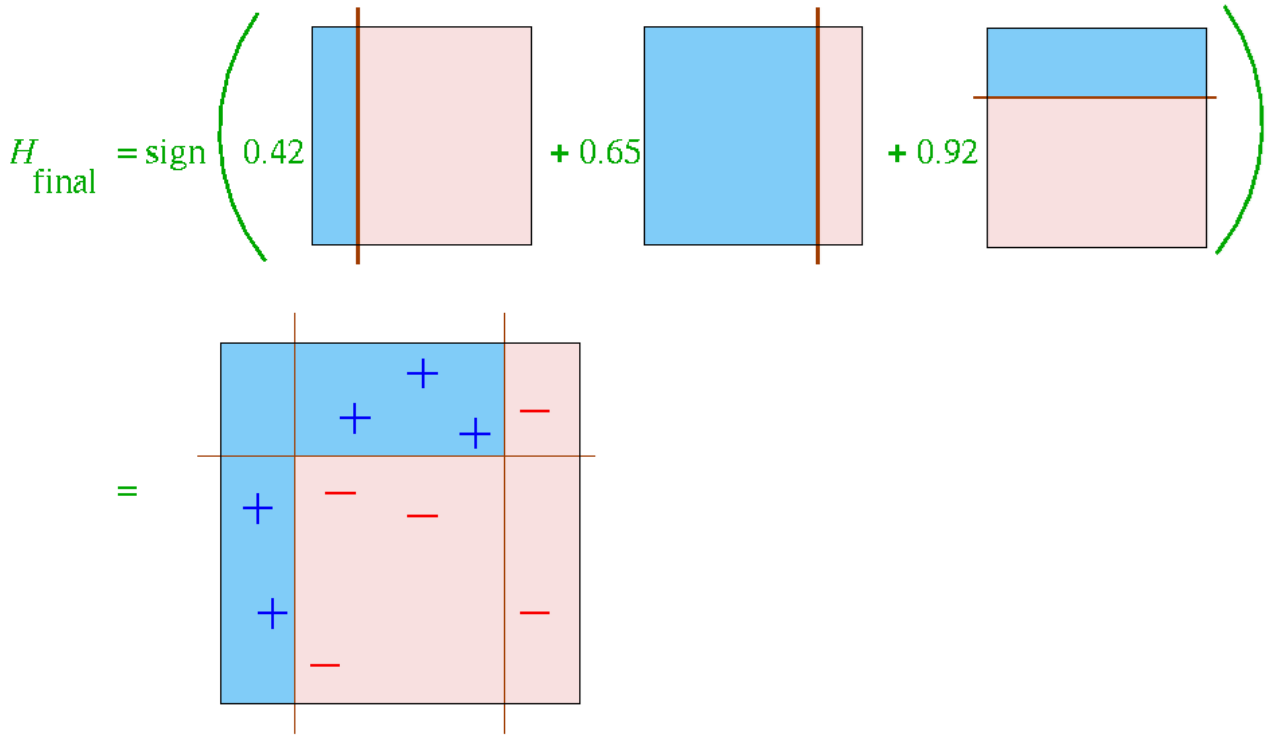


Yoav Freund & Robert Schapire

- Another solid ML model was proposed by Freund and Schapire in 1997 prescribed with boosted ensemble of weak classifiers called Adaboost.
- Adaboost trains weak set of classifiers that are easy to train, by giving more importance to hard instances.
- This model is still the basis of many advanced ML tools like GBDT, and is being actively used in the ML community and related industries.

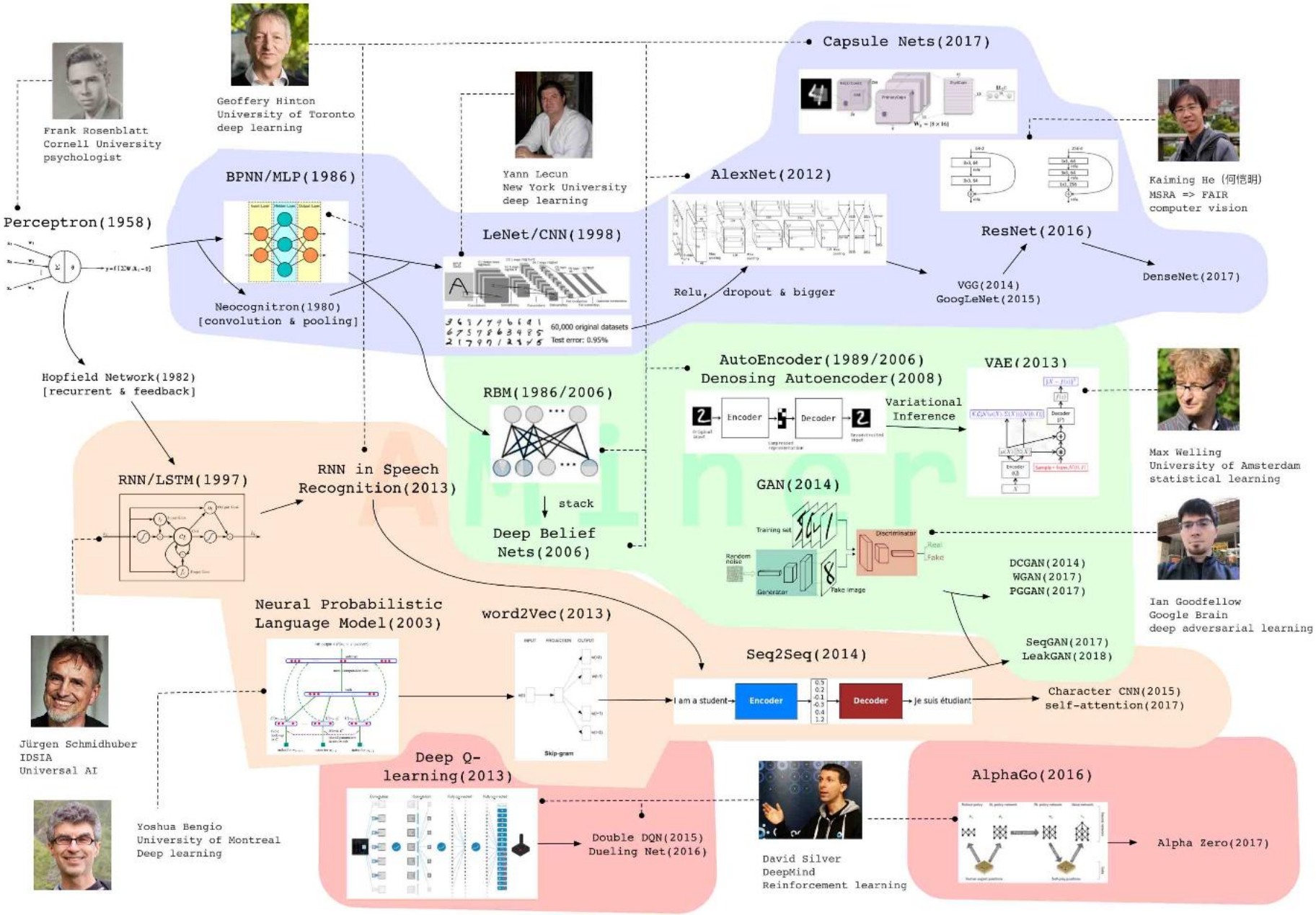
Boosting

- Basic idea:
  - Ask expert (could be “weak” learning algorithm) for rule-of-thumb
  - Assemble set of cases where rule-to-thumb fails (hard cases)
  - Ask expert again for selected set of hard cases (repeat)
  - Combine all rules-of-thumb



In today’s big-data era, sufficient training data make the outstanding expressiveness of neural networks a huge advantage …

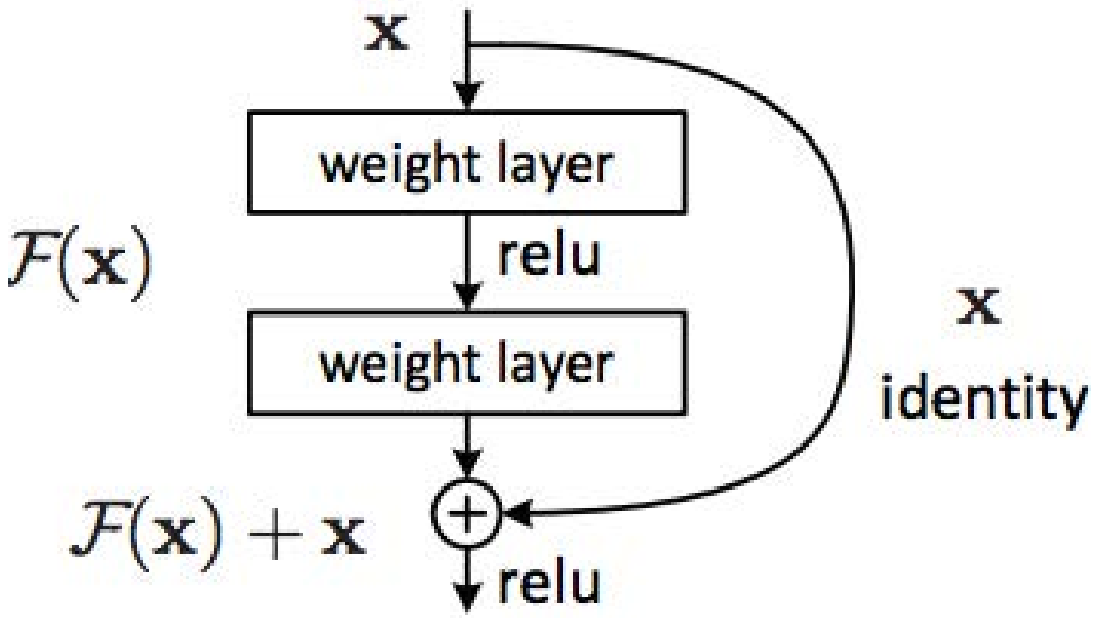
Re ng)


Frank Rosenblatt
Cornell University
psychologist
Geoffery Hinton
University of Toronto
deep learning
Yann Lecun
New York University
deep learning
Capsule Nets(2017)
Kaiming He (何恺明)
MSRA => FAIR
computer vision
BPNN/MLP(1986)
Perceptron(1958)
Neocognitron(1980)
[convolution & pooling]
LeNet/CNN(1998)
60,000 original datasets
Test error: 0.95%
AlexNet(2012)
Relu, dropout & bigger
VGG(2014)
GoogLeNet(2015)
ResNet(2016)
DenseNet(2017)
Hopfield Network(1982)
[recurrent & feedback]
RBM(1986/2006)
stack
Deep Belief
Nets(2006)
AutoEncoder(1989/2006)
Denosing Autoencoder(2008)
Encoder
Decoder
Variational
Inference
VAE(2013)
Max Welling
University of Amsterdam
statistical learning
RNN/LSTM(1997)
RNN in Speech
Recognition(2013)
GAN(2014)
Training set
Random
noise
Generator
Discriminator
Real
Fake
Fake image
DCGAN(2014)
WGAN(2017)
PGGAN(2017)
Ian Goodfellow
Google Brain
deep adversarial learning
Neural Probabilistic
Language Model(2003)
word2Vec(2013)
Skip-gram
Seq2Seq(2014)
I am a student
Encoder
Decoder
Je suis étudiant
SeqGAN(2017)
LeakGAN(2018)
Character CNN(2015)
self-attention(2017)
Jürgen Schmidhuber
IDSIA
Universal AI
Yoshua Bengio
University of Montreal
Deep learning
Deep Q-
learning(2013)
Double DQN(2015)
Dueling Net(2016)
David Silver
DeepMind
Reinforcement learning
AlphaGo(2016)
Alpha Zero(2017)

2018 ACM A.M. Turing Award Laureates

- Geoffrey Hinton
  - Backpropagation
  - Boltzmann machines
  - Improvements to CNNs
- Yoshua Bengio
  - Probabilistic models of sequences
  - High-dimensional word embeddings and attention
  - Generative adversarial networks
- Yann LeCun
  - CNNs
  - Improving backpropagation algorithms
  - Broadening the vision of neural networks

Very Deep Neural Networks



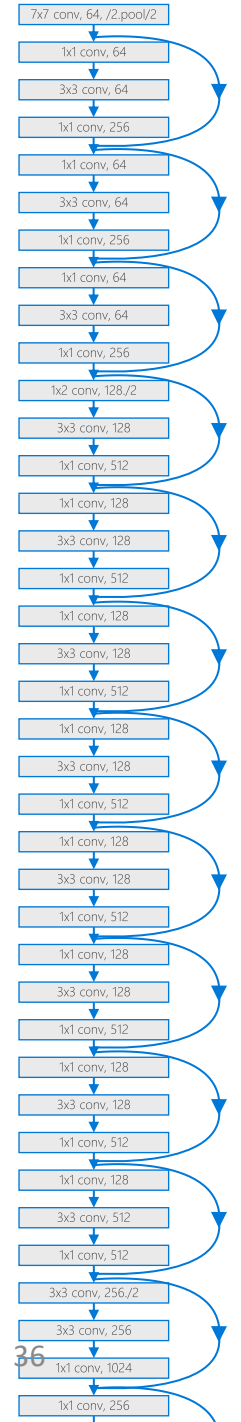
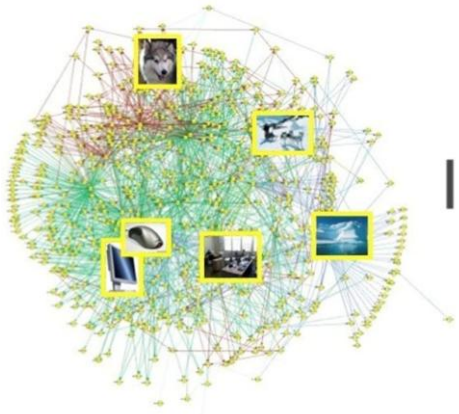
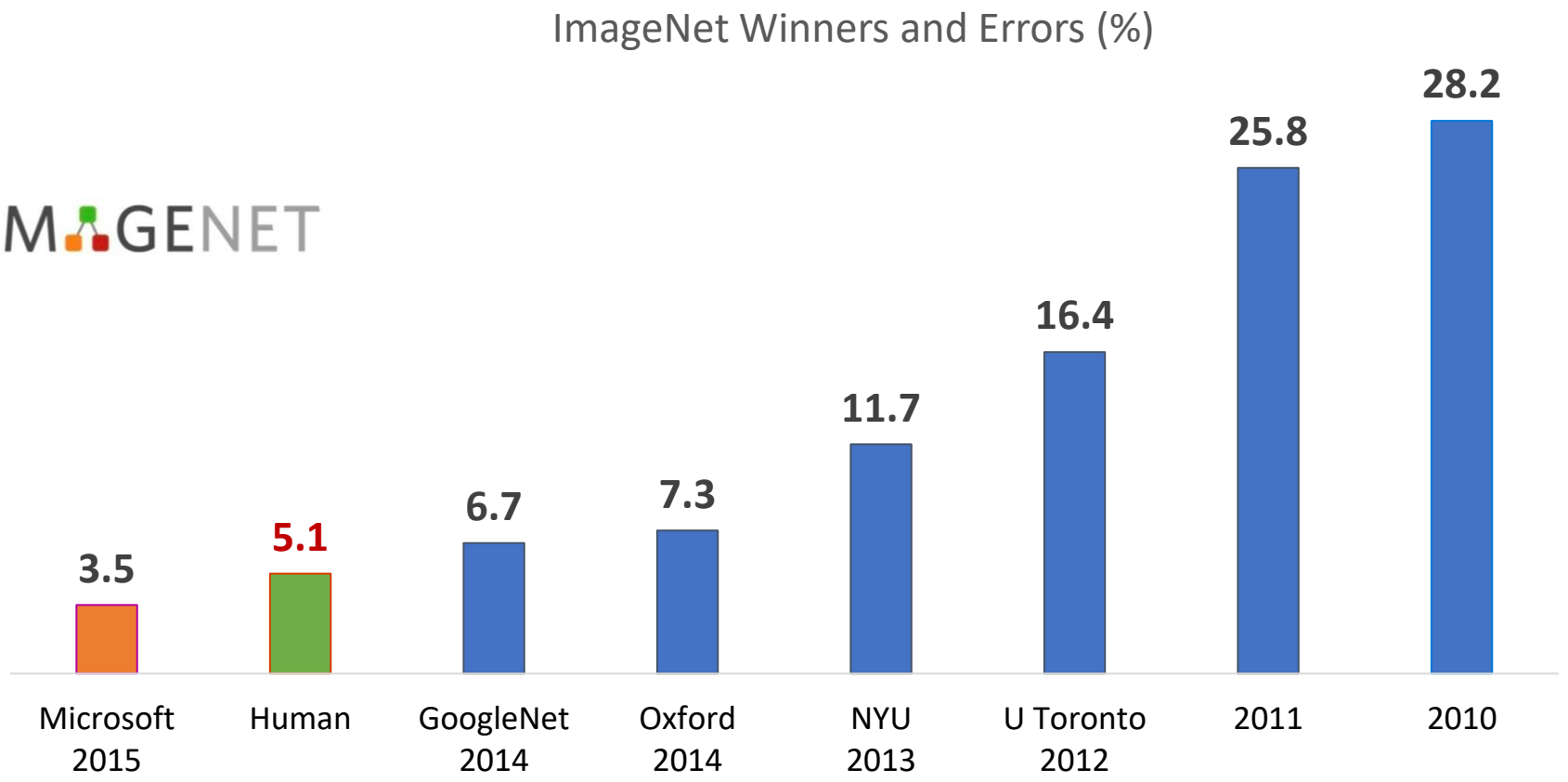


Image Recognition

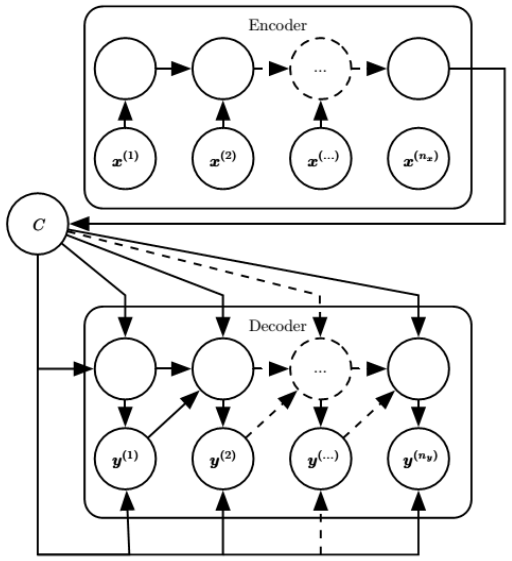


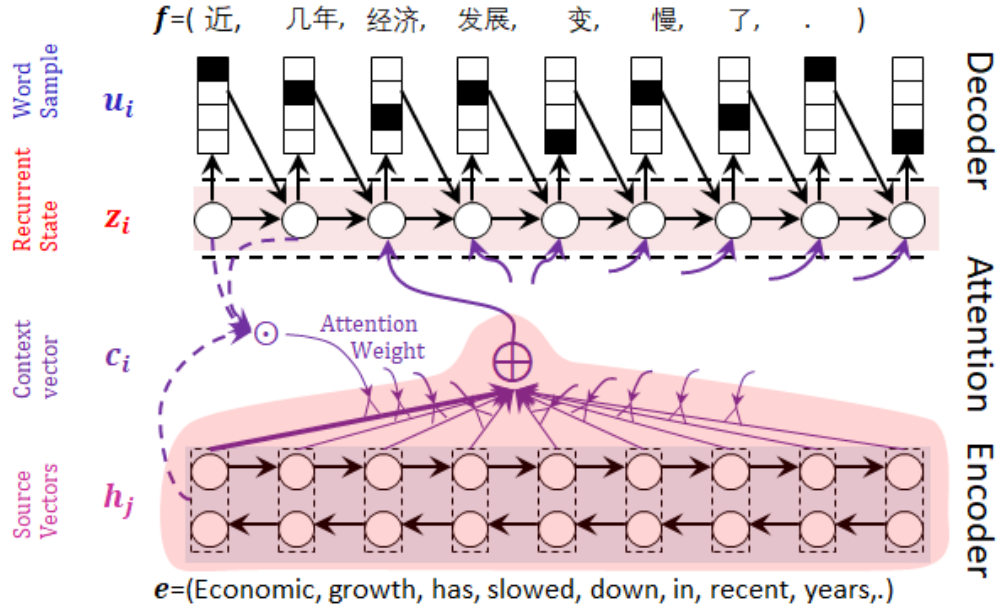
IMAGENET



Encoder-Decoder with Attention Mechanism

Yoshua Bengio made remarkable contributions to neural language model, high-dimensional word embeddings, attention mechanism, and encoder-decoder framework. These works are foundations of deep learning for NLP.





Machine Translation

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

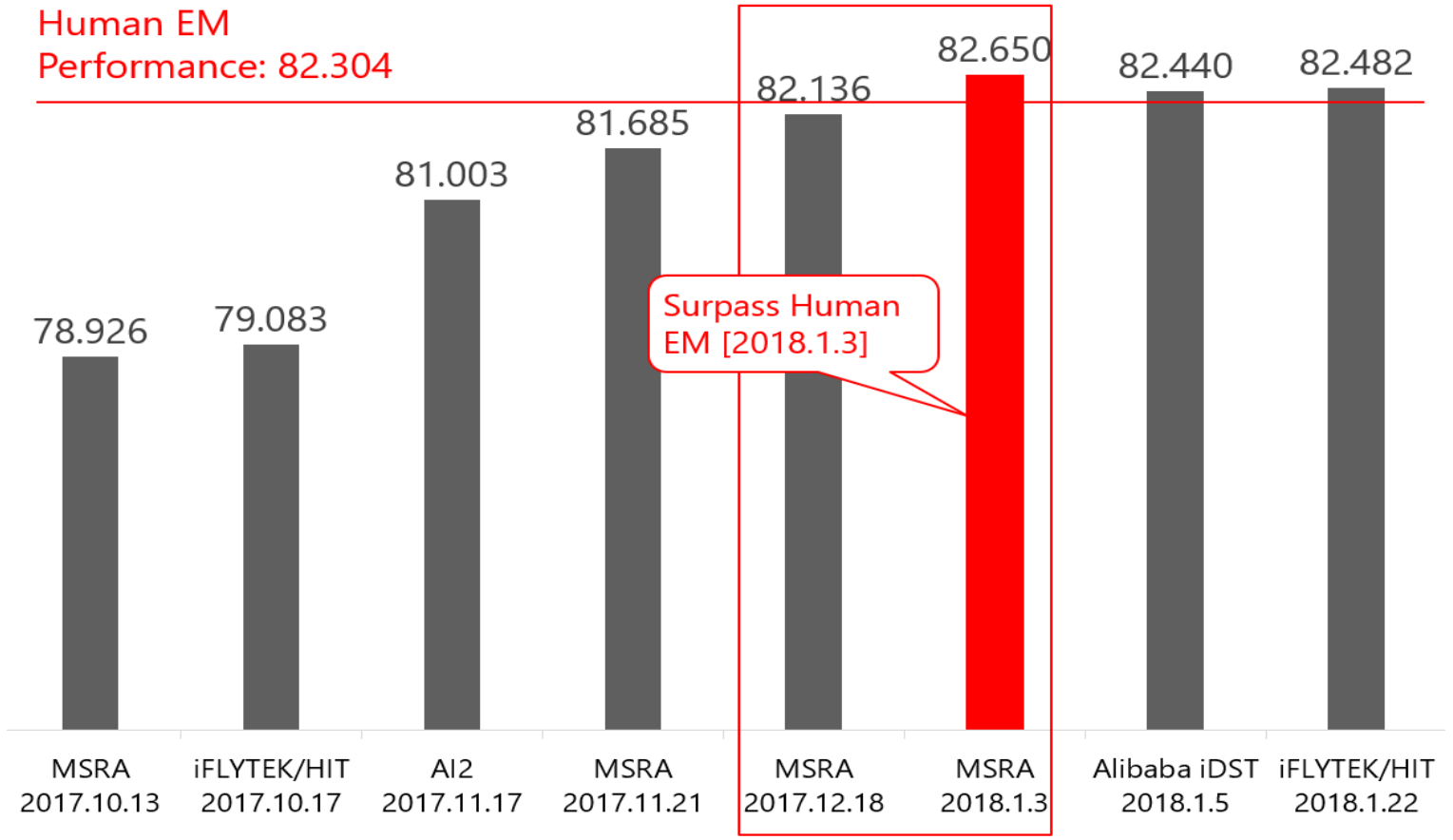
March 14, 2018 | Allison Linn

Subjective score: 69.5

Human: 69.0

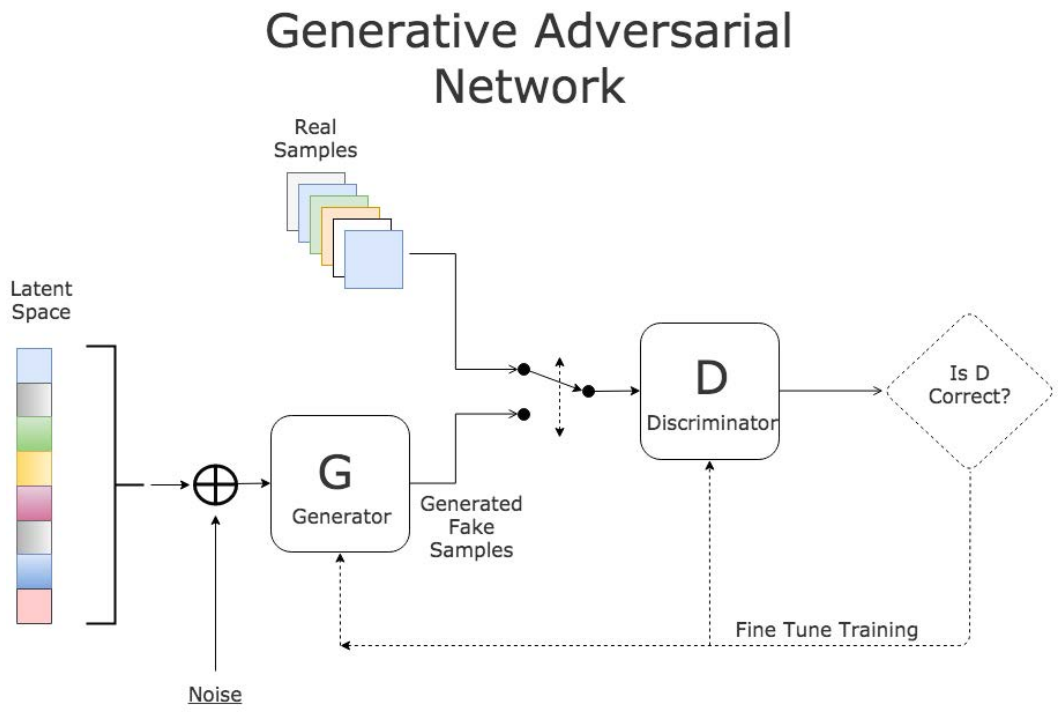
Reading Comprehension

- SQuAd



Generative Adversarial Networks

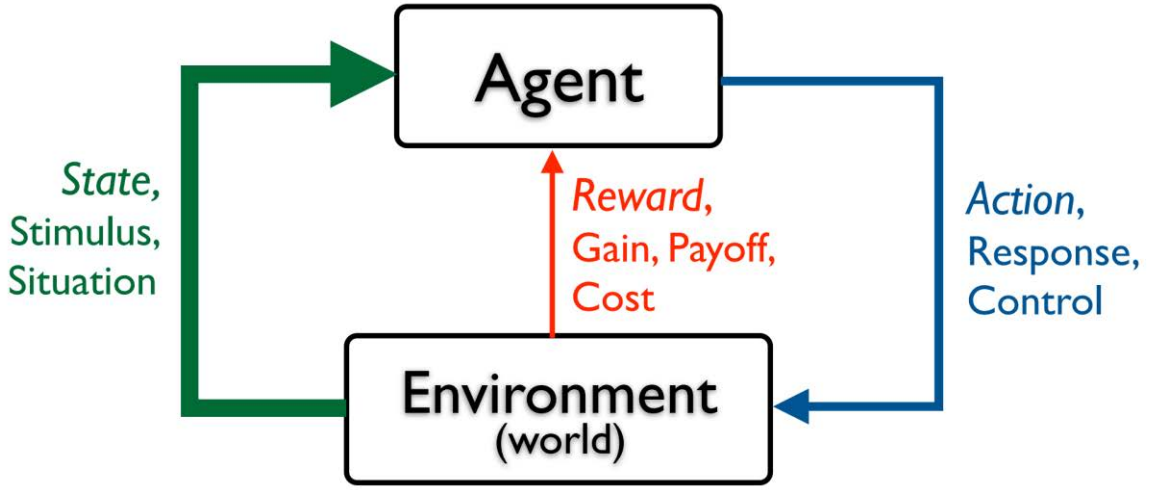
Ian Goodfellow (together with Bengio) proposed Generative Adversarial Networks (GAN) in 2014. GAN has been applied to computer vision, speech, and languages, etc.



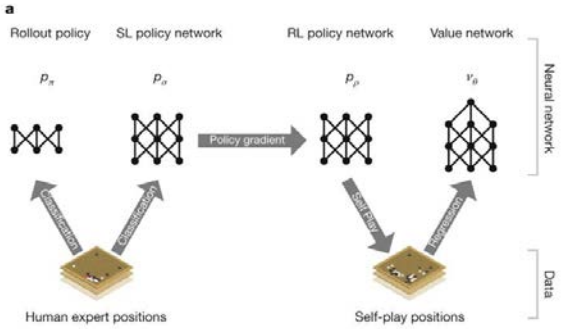
- Generator captures the data distribution
- Discriminator estimate the probability that a sample came from the training data rather than the generator

Deep Reinforcement Learning

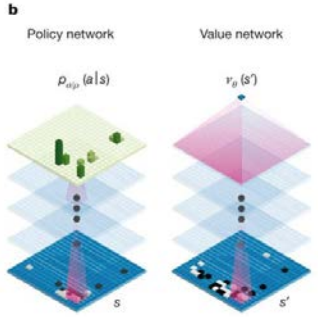
- RL: agent-oriented learning by interacting with an environment to achieve a goal
  - Learning by trial and error, with only delayed evaluative feedback(reward)
  - Agent learns a policy mapping states to actions, in order to maximize its cumulative reward in the long run
- Deep RL:
  - RL defines the objective
  - DL gives the mechanism

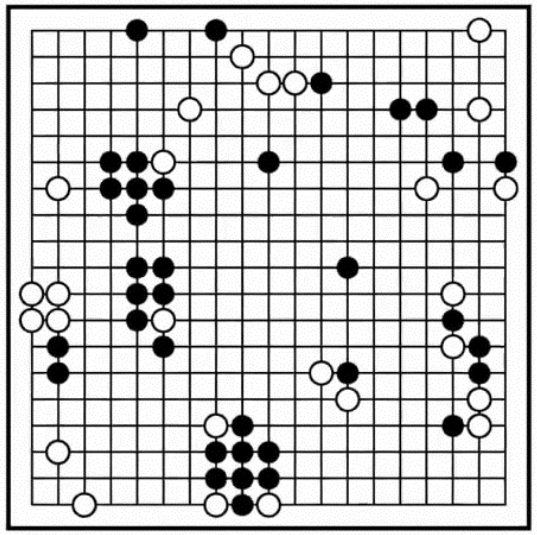


Go Playing - AlphaGo



AlphaGo



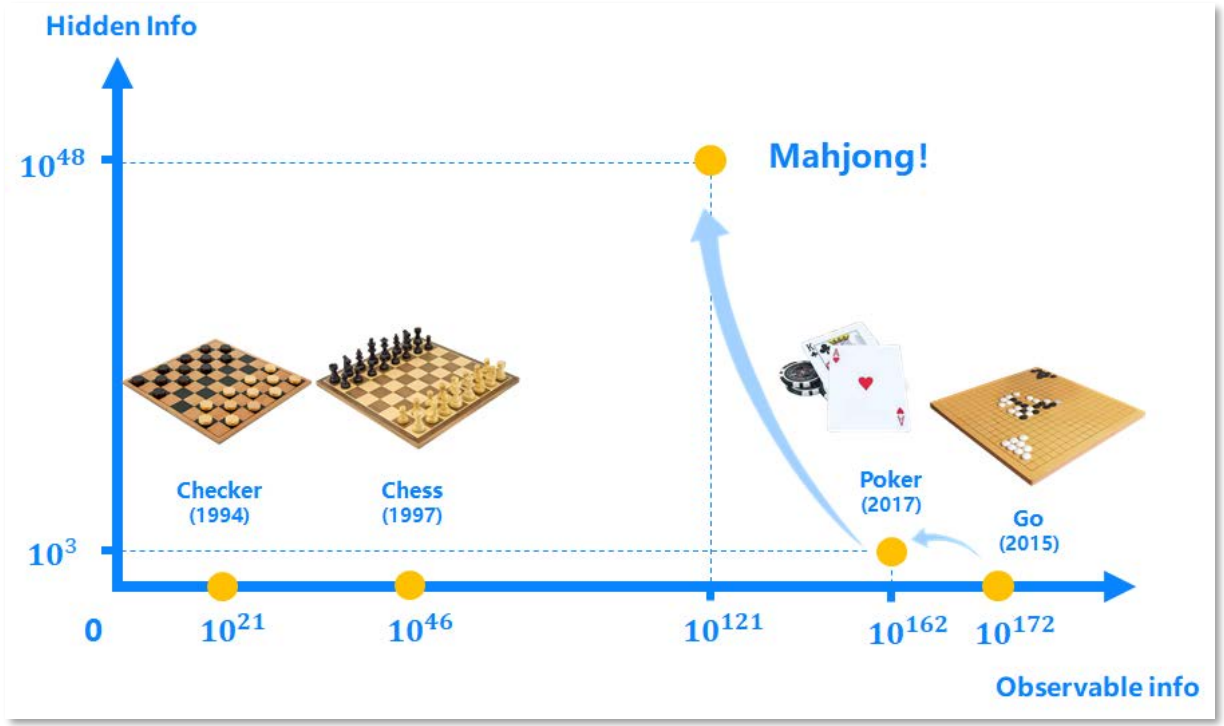


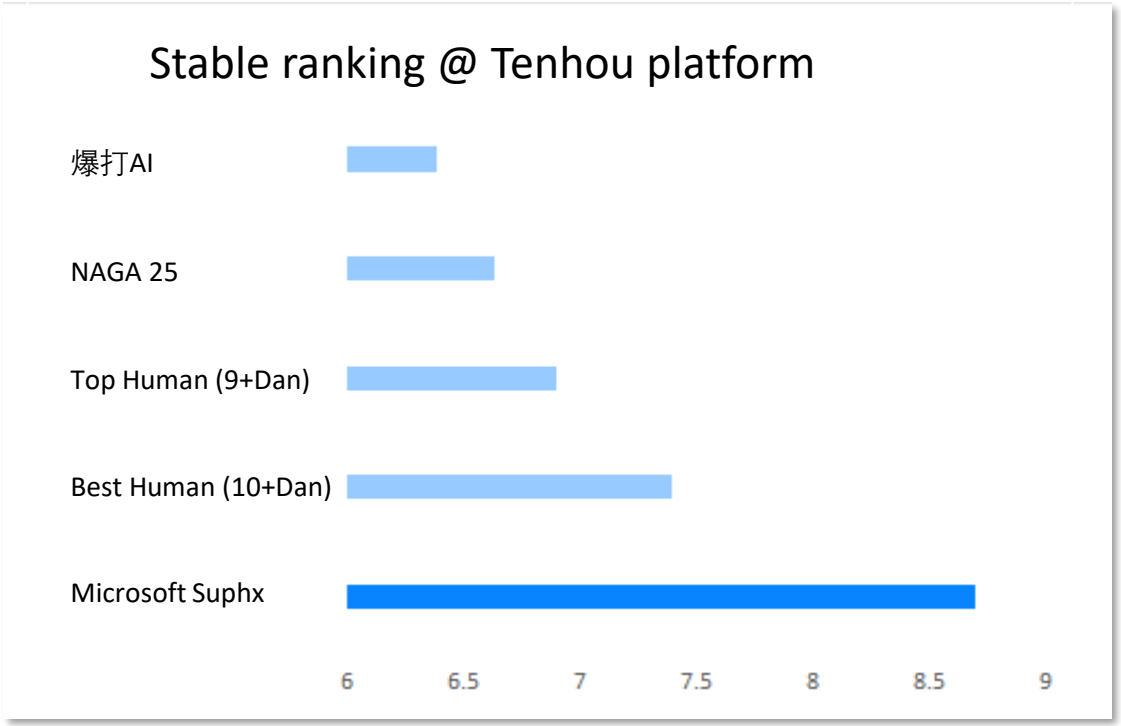
4:1 against Sedol Lee

3:0 against Jie Ke

Mahjong Playing: Suphx

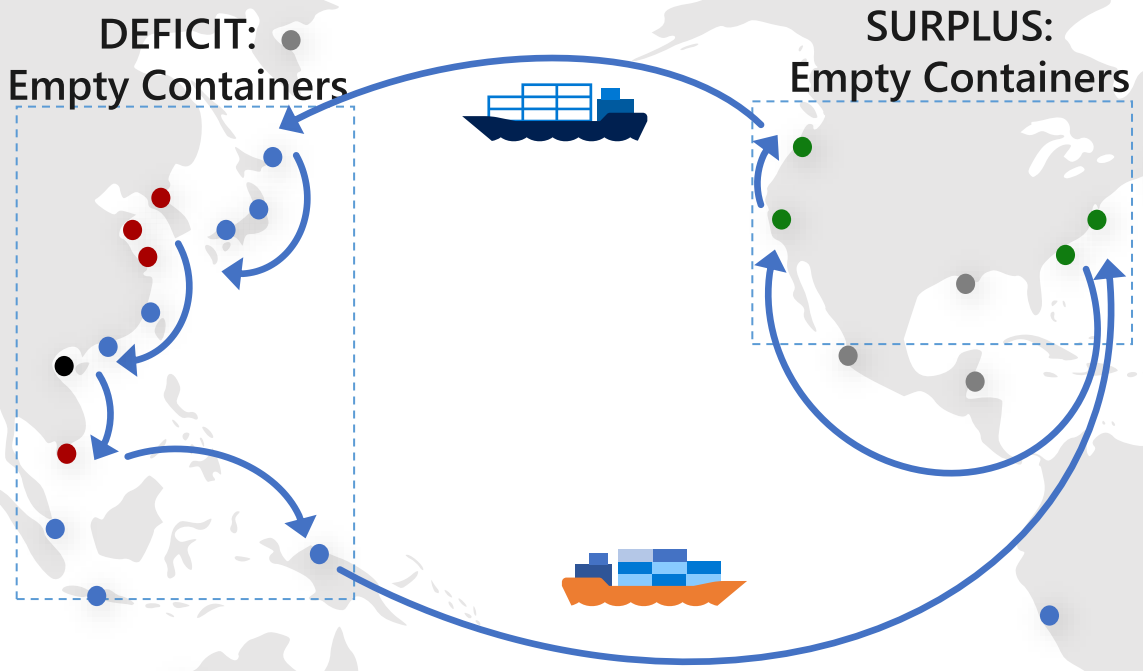
- Deep RL + Oracle critic + Policy adaptation





Container Repositioning

- Use **Coopetitive Learning** to optimize the container repositioning plan (ports and vessels as local agents).
- Outperforms traditional OR-based approaches, in terms of robustness, efficiency, and even fulfillment ratio and operational cost (saving of over 10M USD).



Container Repositioning in Ocean Transportation

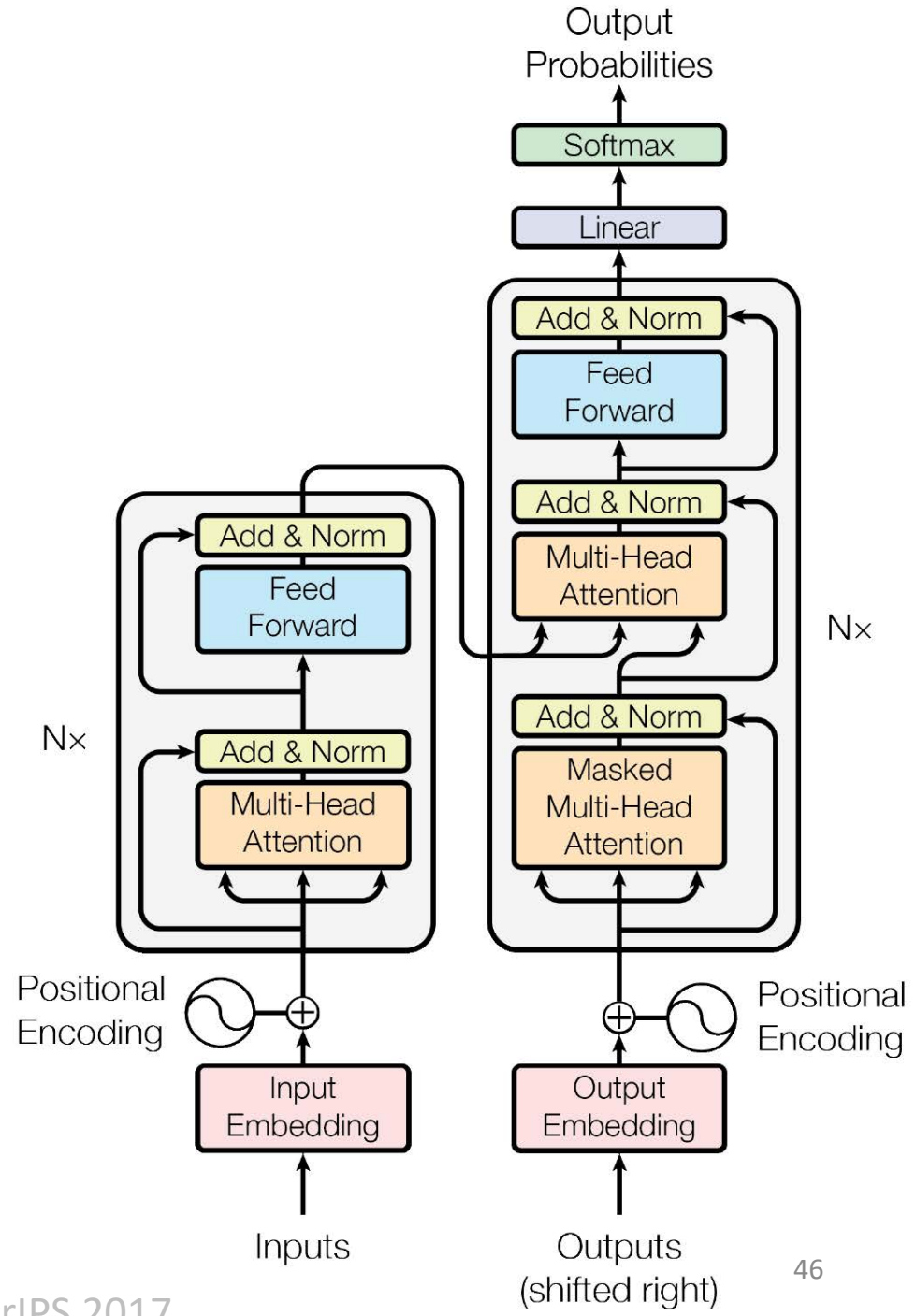
Transformer Networks

- Attentions in the model
  - Encoder-decoder attention layers
  - Self-attention layers in the encoder
  - Masked self-attention layers in the decoder

[CITATION] **Attention is all you need**

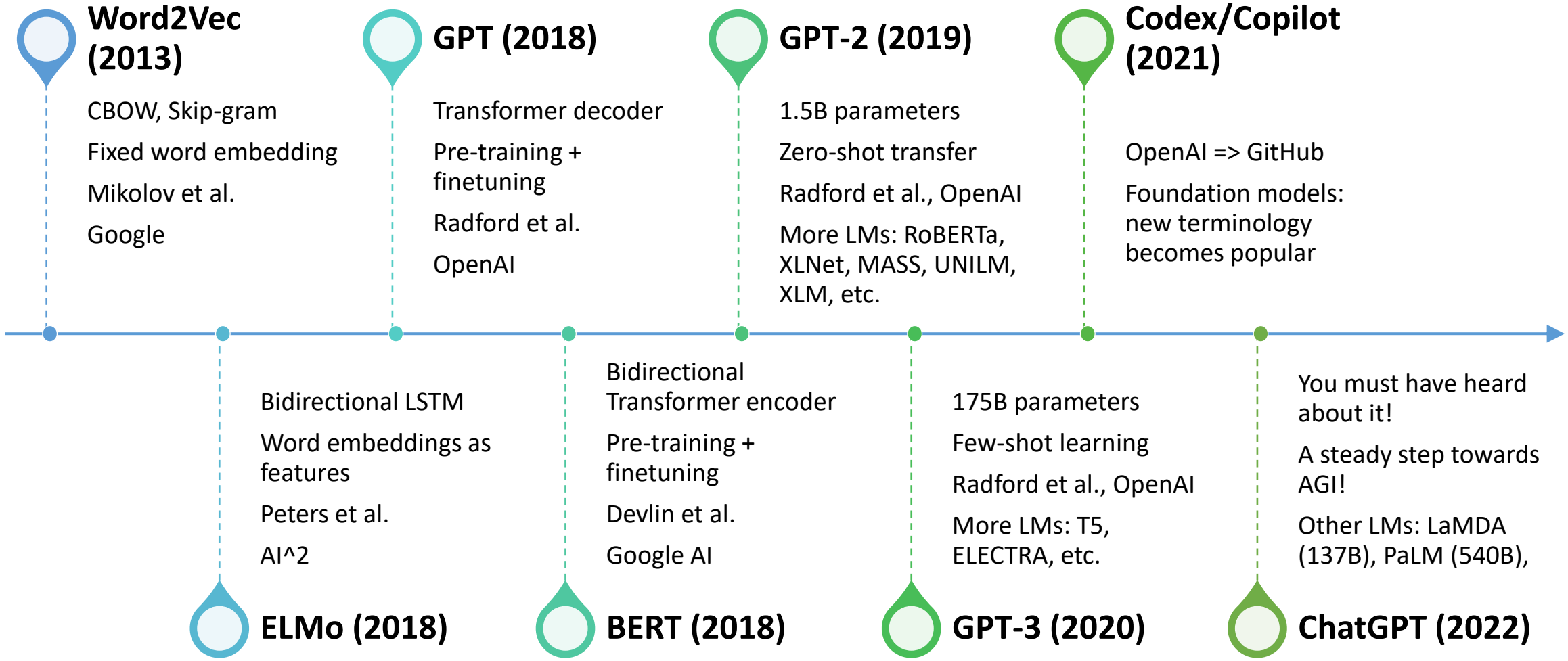
A Vaswani - Advances in Neural Information Processing Systems, 2017

☆ Save Cite Cited by 152962 Related articles



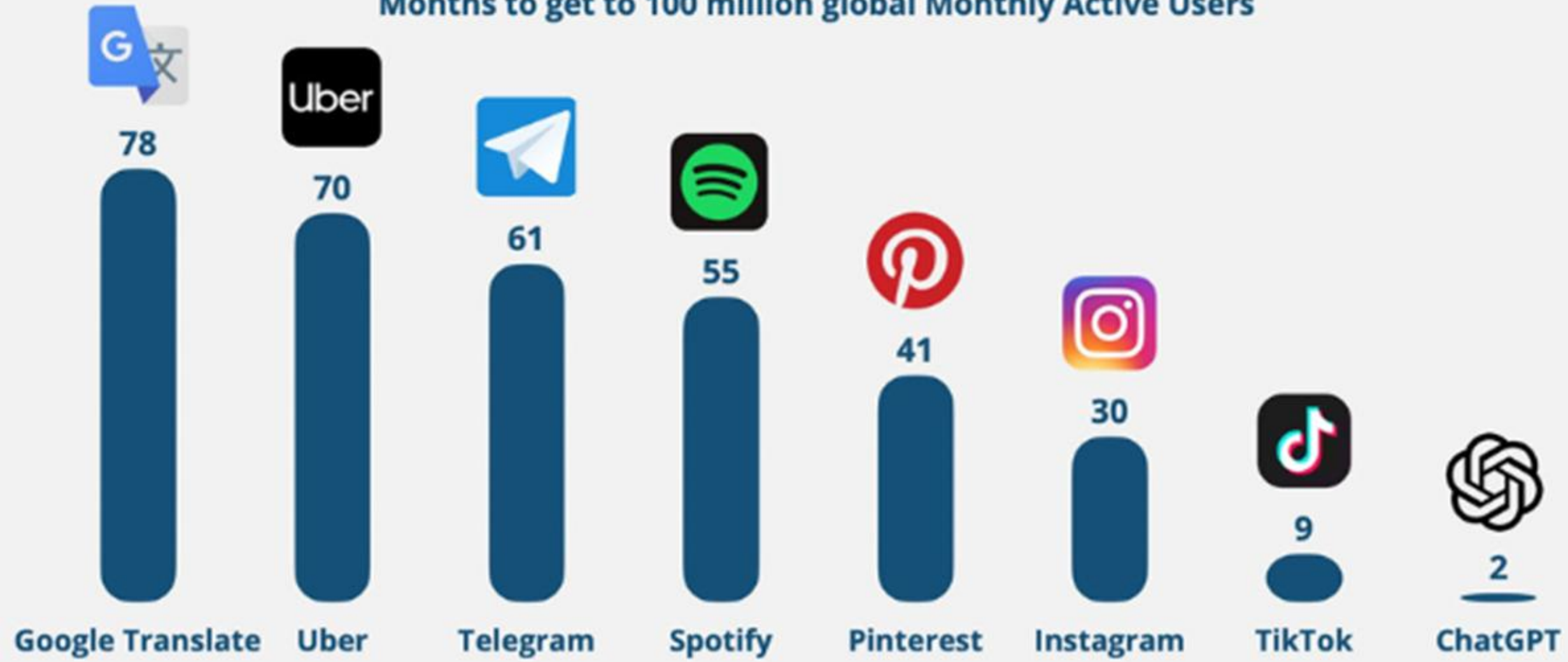
Attention is all you need, NeurIPS 2017

Language models in past several years



Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users



Source: UBS / Yahoo Finance

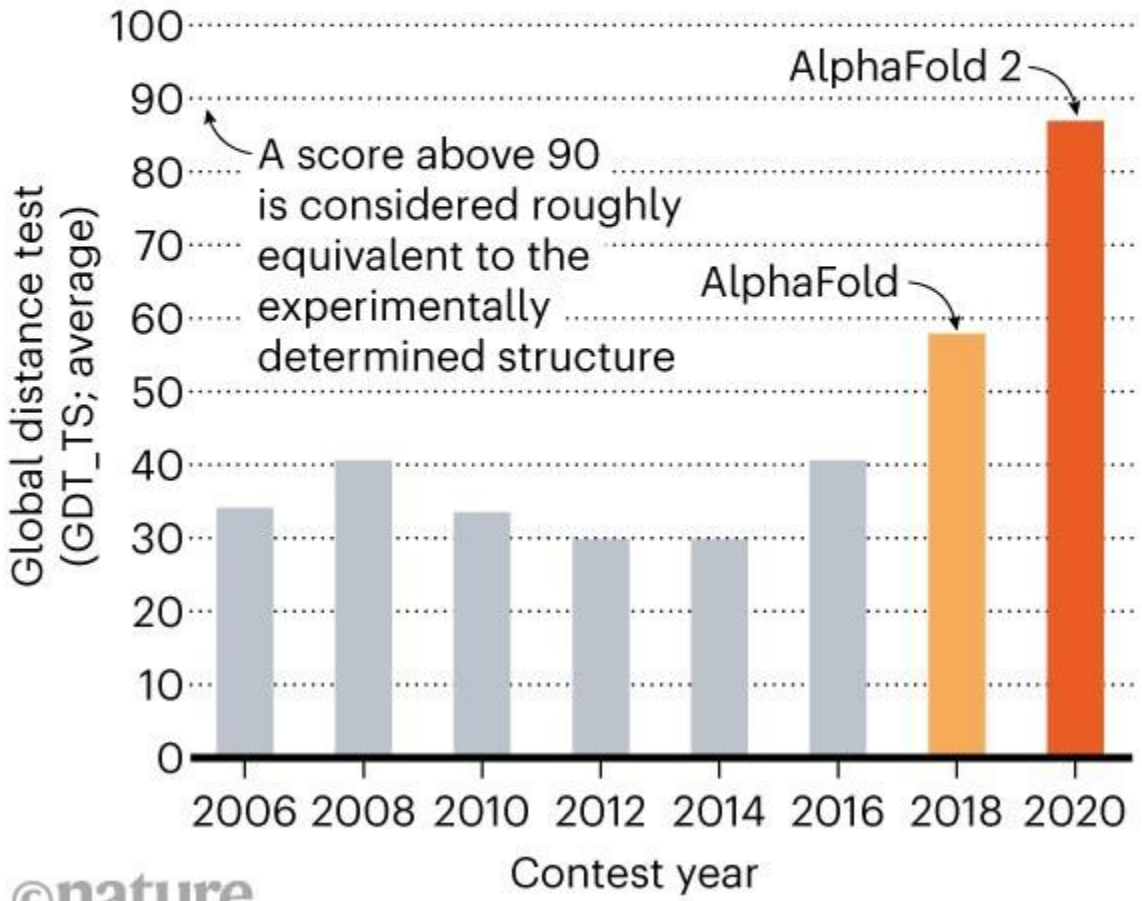
APP ECONOMY INSIGHTS

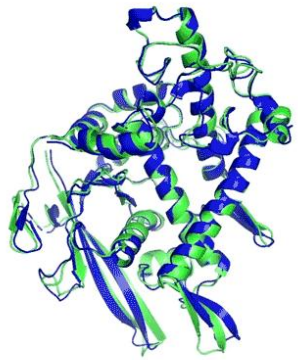
Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

Credit: OpenAI

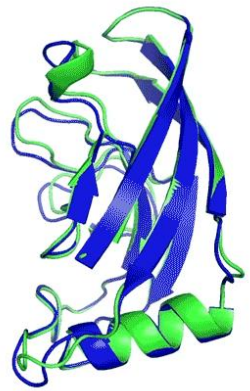
STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



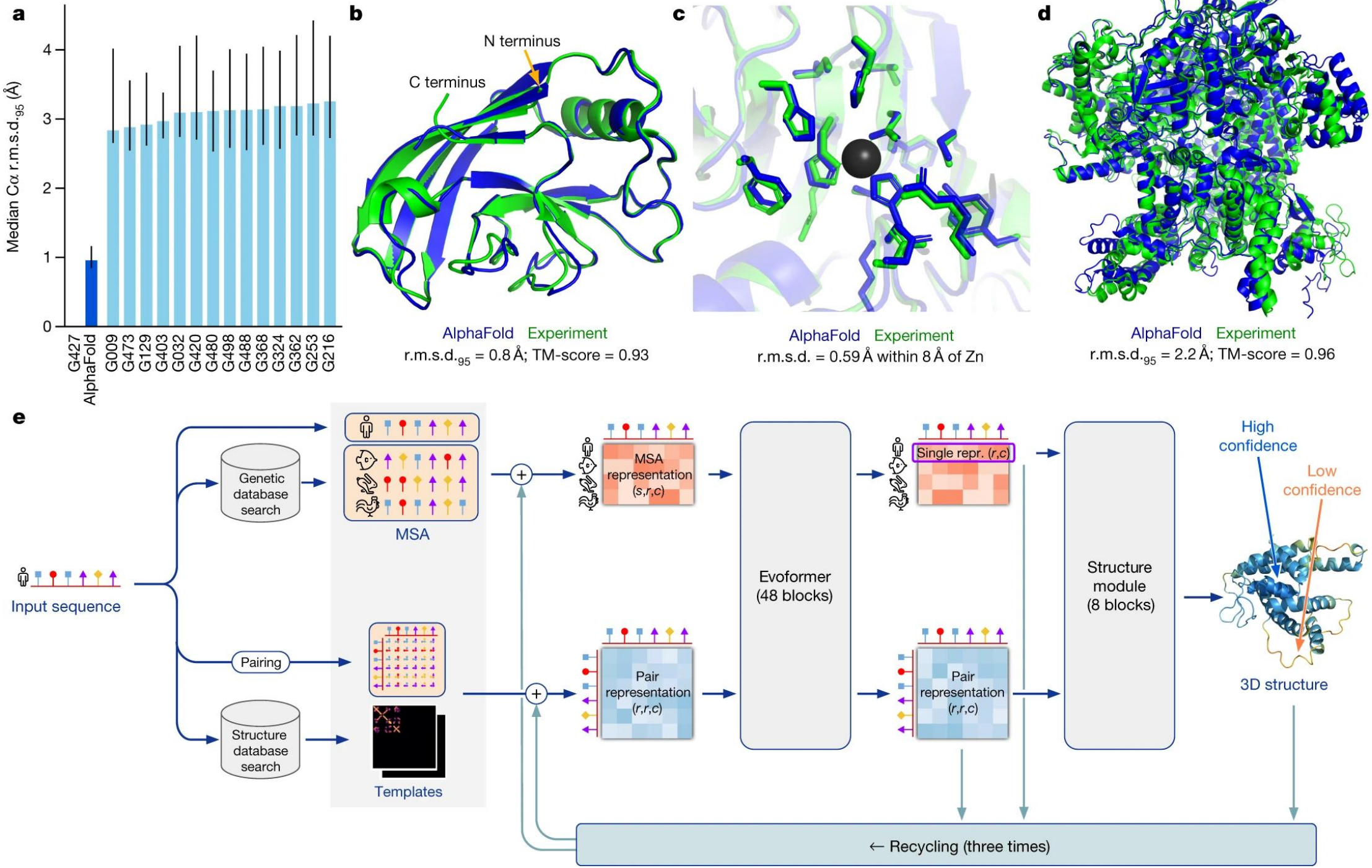


T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

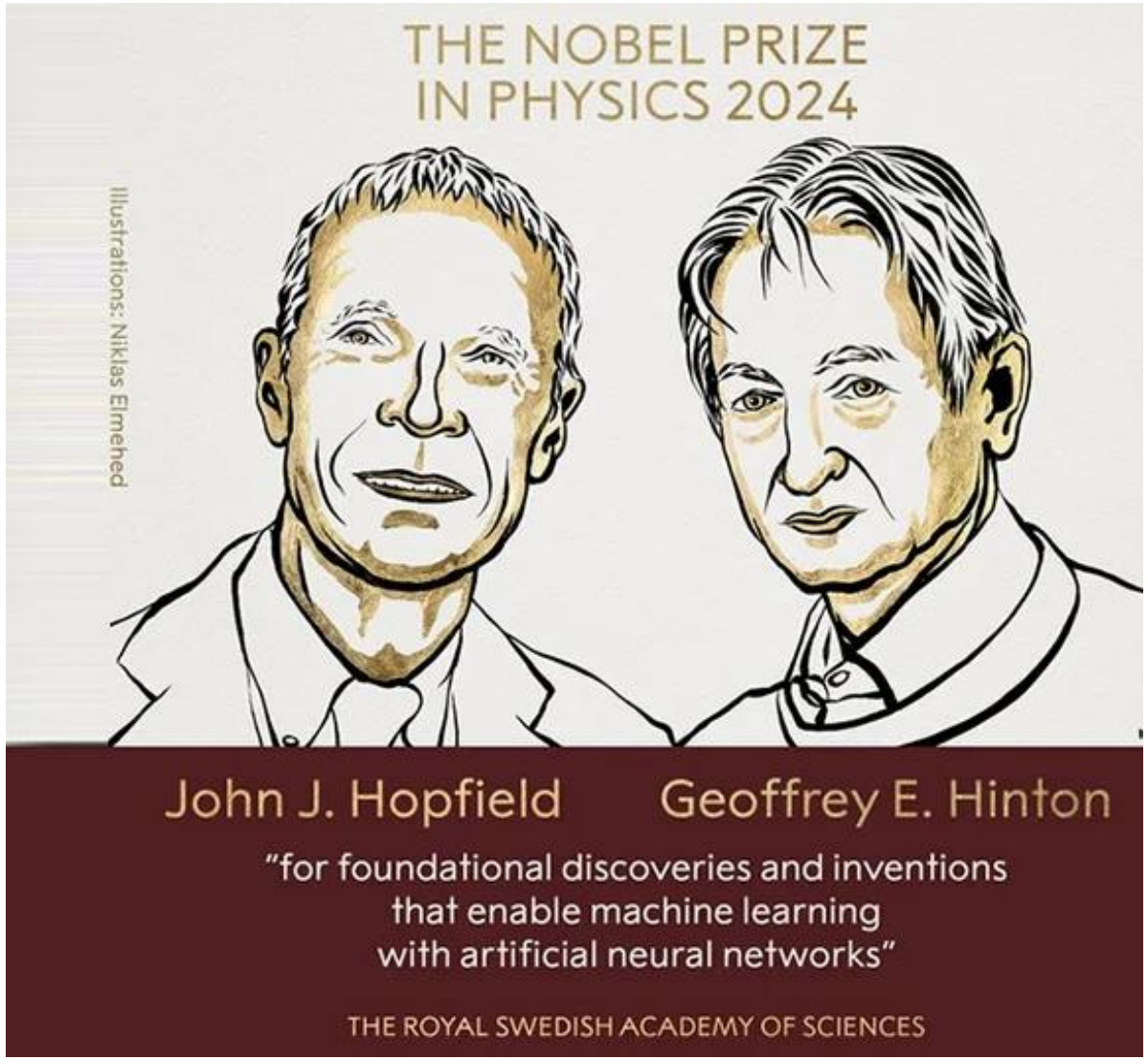


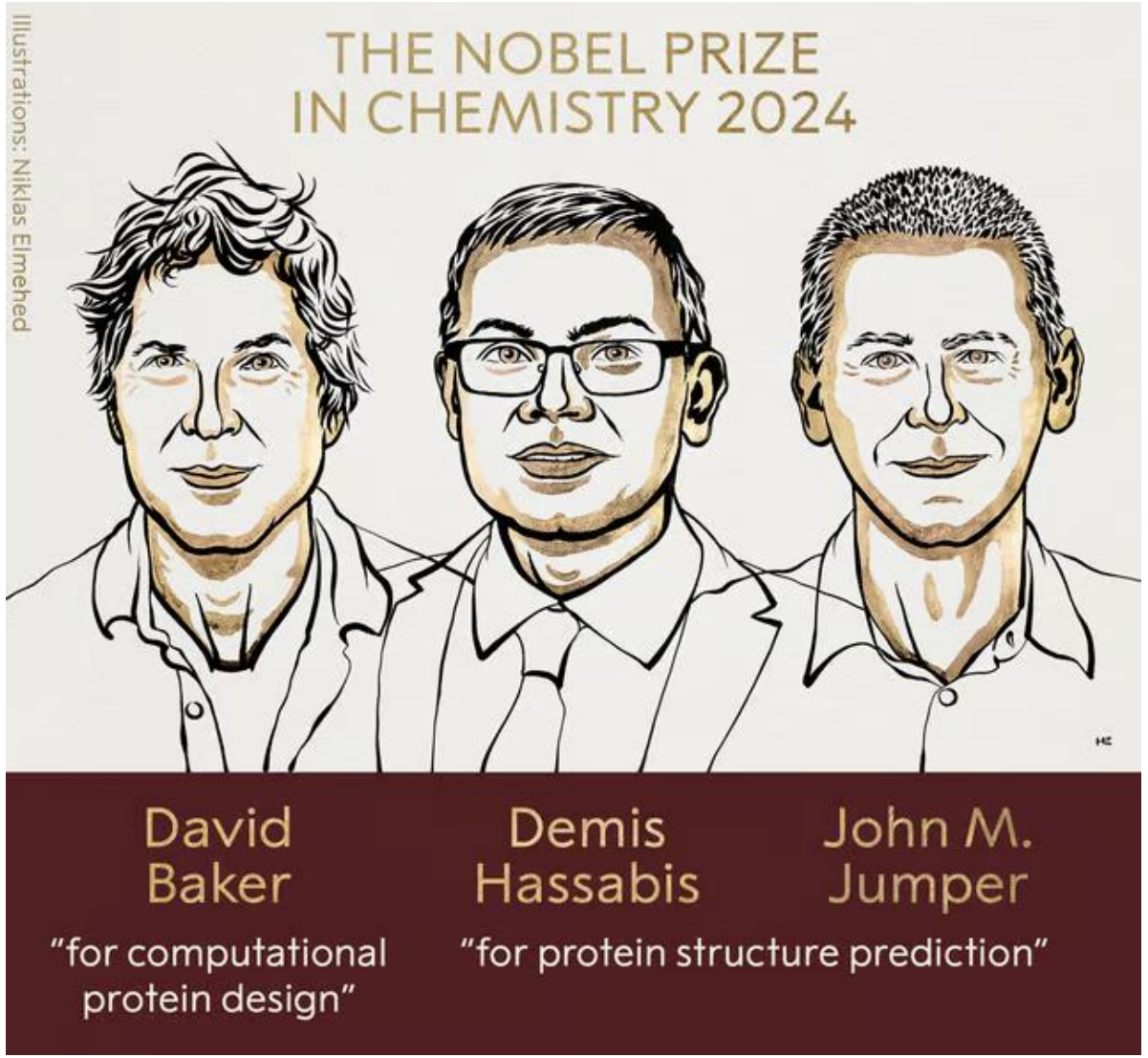
T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction


a
Median Cα r.m.s.d.95 (Å)
0
1
2
3
4
G427 AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.95 = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.95 = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Nobel Prize 2024





超级产品
增长1亿用户所用的时间

注：DeepSeek 包含网站Web/应用App累加不去重，Tiktok 不包含国内版抖音

数据来源：AI产品榜 aicpb.com 感谢邓瑞恒提供作图思路

Pre-Knowledge

CALCULUS

LINEAR ALGEBRA

PROBABILITY THEORY AND STATISTICS

OPTIMIZATION

PROGRAMMING LANGUAGES

Course Requirements

- Be present and on time
- Pay attention, put efforts, study hard!
- Regular paper reading as a habit
- Always hands on – this is not a pure math course
- Collaborate with others on the course projects

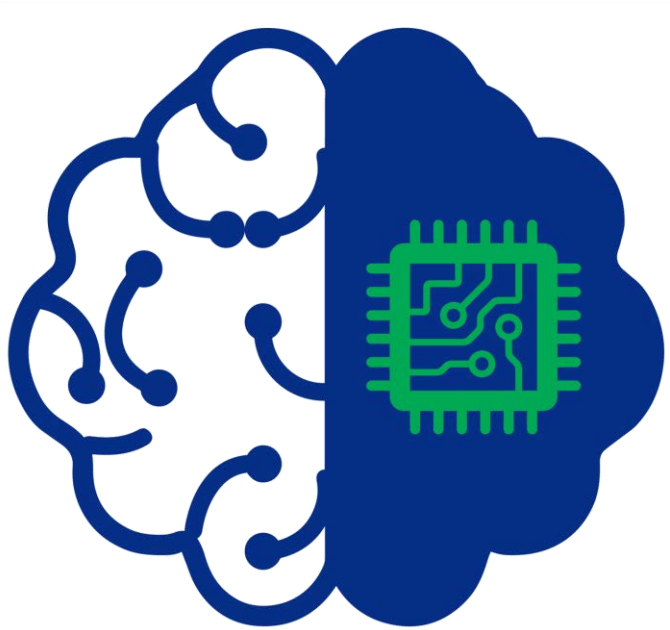
- Ask if you have questions/confusion

Advice for Studying Machine Learning

- Build a strong foundation in mathematics and statistics
- Learn programming
- Start with the basics
- Implement algorithms from scratch
- Work on projects
- Study from reputable resources
- Stay updated with the latest research
- Participate in competitions and challenges
- Collaborate and network
- Embrace continuous learning

Evaluations

- Class attendance (20%)
  - Check-in questions
  - Classroom questions
- Paper reading report (30%)
  - Identify one topic in machine learning
  - Read all related papers in a top conference this year, and write a survey
- Course project (50%)
  - Form a team of 3~5 students
  - Select one project from the list (available next month) or propose by yourself
  - Design new machine learning solutions and conduct experiments
  - Write project reports and make presentations



“Machine Learning” Thinking

Leverage Machine Learning Approach

- Machine Learning ≈ Data-driven optimization

- When and how to leverage the approach?
  - Human expertise does not exist (navigating on Mars)
  - Humans are unable to explain their expertise (speech recognition)
  - Solution needs to be adapted to particular cases (personalization)
  - Knowledge involved is expensive and scares while data or simulations are cheap and adequate (go playing)
  - Theory cannot explain observations
  - … …

Inspirations

- Being more objective and rational
  - Data-driven optimization
  - Objective function, solution space, constrained optimization, avoid overfitting
- Lifelong proactive learning
  - Reinforcement learning: learning from interactions (failure and success) to maximize long-term return
  - Exploration and exploitation trade-off
- Becoming stronger by working as a team
  - Boosting: combination of weak decisions could be very strong

How to do Better with "Machine Learning" Thinking?

- What kinds of professional or personal decision making can be improved?

- Study
- Career development?
- Financial investment?
- Dating?
- … …

References

1) Christopher M Bishop, Hugh Bishop, Deep Learning: Foundations and Concepts, Springer, 2024
2) Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press
3) Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning, Springer
4) Christopher M Bishop, Pattern Recognition and Machine Learning, Springer
5) 张旭东，机器学习，清华出版社，2024
6) 李航，机器学习方法，清华出版社， 2022
7) 周志华，机器学习，清华出版社， 2016

Thanks!

taoqin@microsoft.com

http://research.microsoft.com/users/taoqin/