高等机器学习

课程导论

刘铁岩
微软亚洲研究院

Microsoft

清華大學
Tsinghua University

Who Am I?

1994 | 1998 | 2003 | 2008 | 2012 | 2017 | 2019

考入清华
电子工程系

本科毕业

博士毕业 加入MSRA

在信息检索和机器学习领域发表200
余篇论文，他引数万次，H-index=49

IEEE Fellow

Elsevier 最高引中国学者

师从张旭东教授
从事信号处理领域的研究 → 从事信息检索领域的研究 → 从事机器学习领域的研究

ACM杰出科学家

中国AI英雄风云榜
技术创新人物

JVCIR
最高引论文

SIGIR
最佳论文

出版第一
本英文书
Learning to
Rank

Springer
计算机类十大
华人作者

ACML
最佳论文

AMiner全球
最具影响力学者
（信息检索）

Prediction Tasks

Taking photos: predict types of images and adjust focus, brightness, and contrast

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:

IF

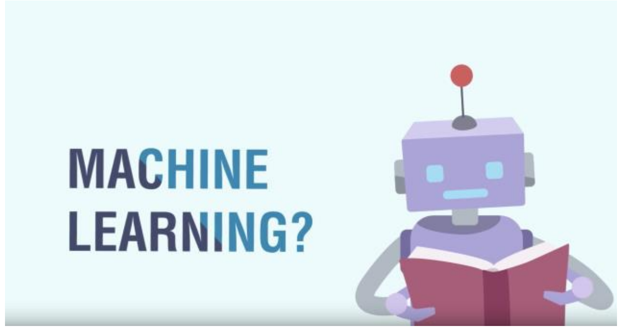
(conditioned on a pattern)

THEN

(take an action)

IF	THEN
You read sports news yesterday	Recommend sports news to you today
The picture contains people	Put the focus on their faces
There are many packages from Beijing to Shanghai last week	Pre-allocate more tracks from Beijing to Shanghai this week

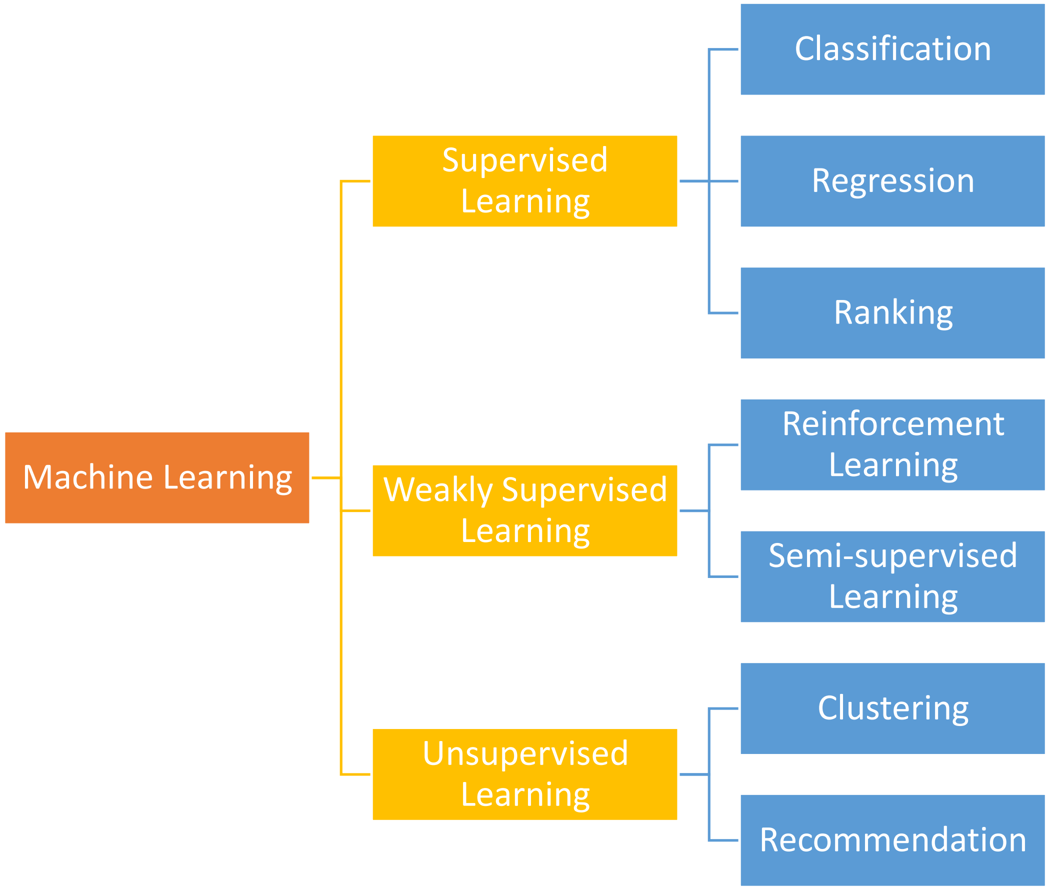
Limitation of Rule-based Solution



- 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

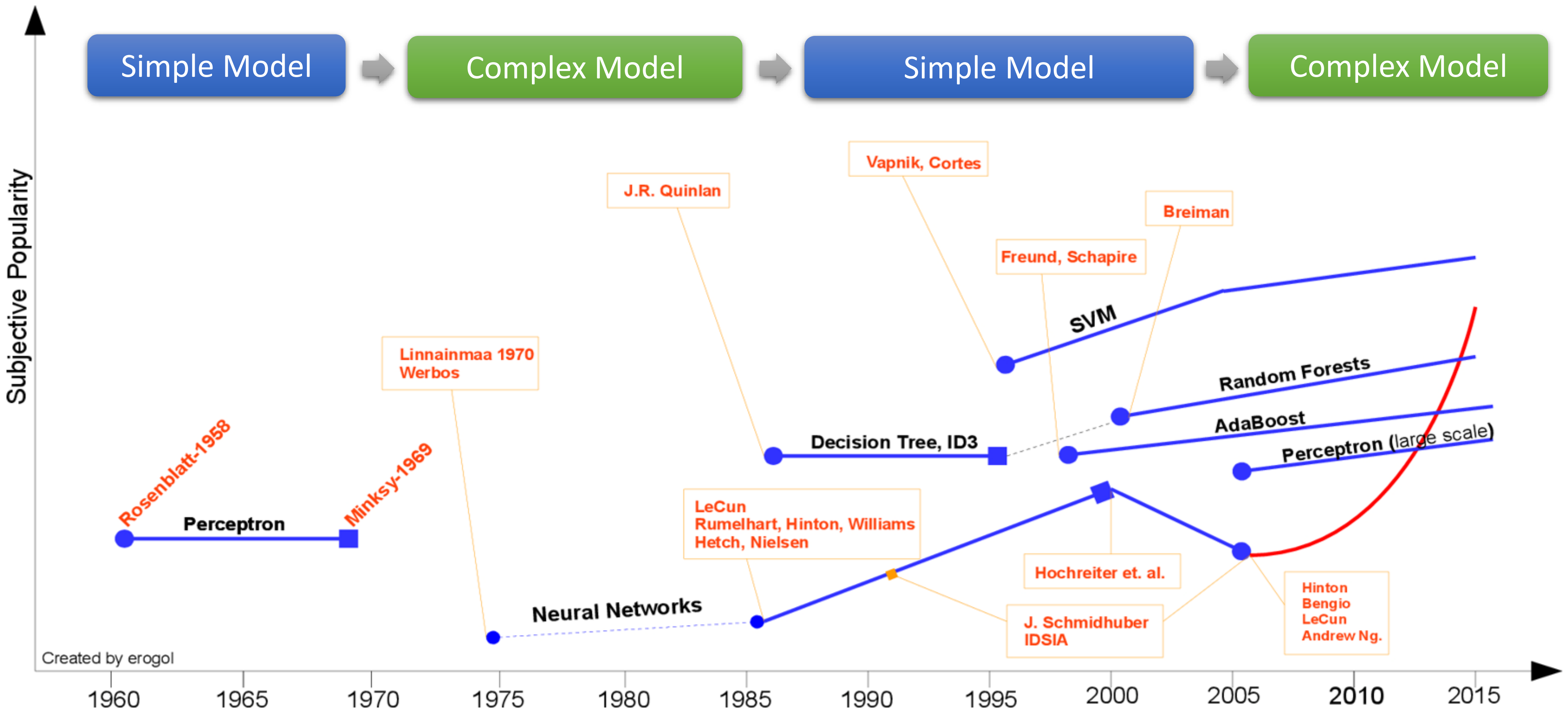
Optimization　Generalization　Objective function

$$\omega^* = \underset{\omega\in\boxed{\boldsymbol{\Omega}}}{\boxed{\arg\min}} \sum_{\substack{x_i\in\boxed{X},y_i\in\boxed{Y} \\ (x_i,y_i)\sim P}}^{i=1,\ldots,\boxed{N\to\infty}} \boxed{\boldsymbol{L}}(f_\omega(x_i), y_i)$$

Constrained model space　Data: Input space　Data: Output space

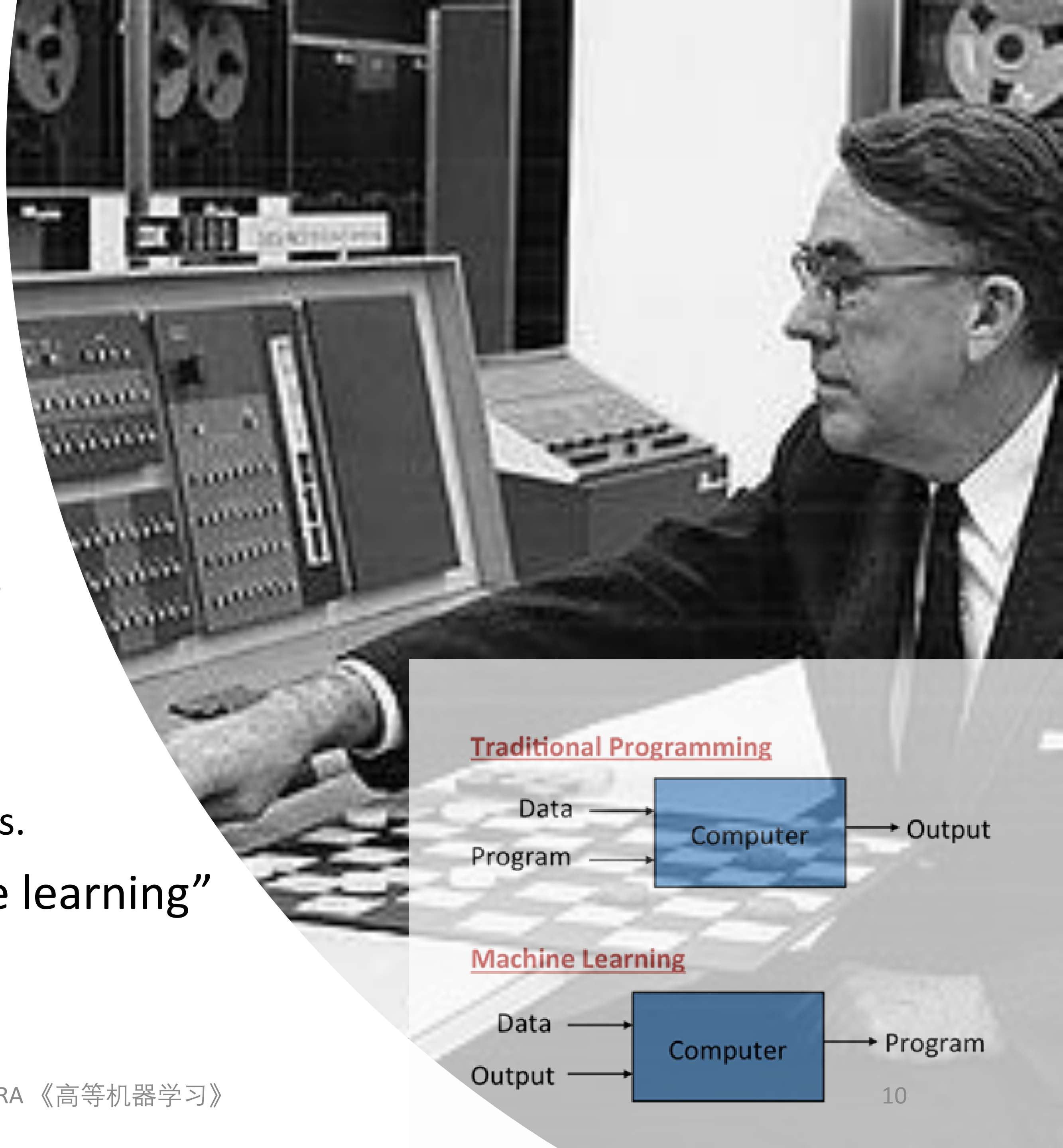
A Brief History of Machine Learning

A Brief History of Machine Learning



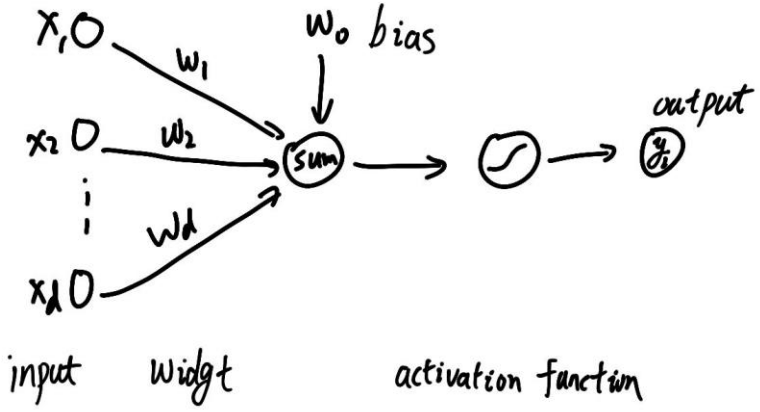
Arthur Samuel

- 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.

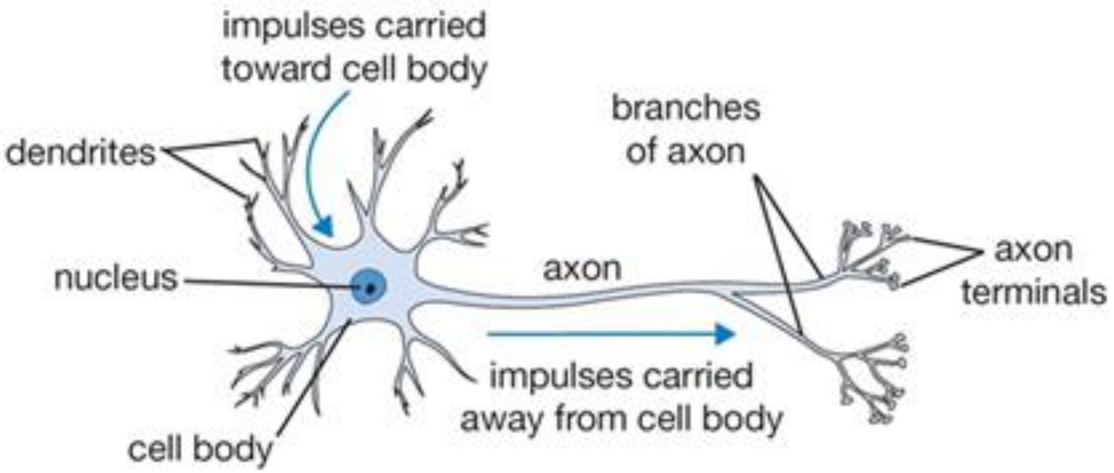


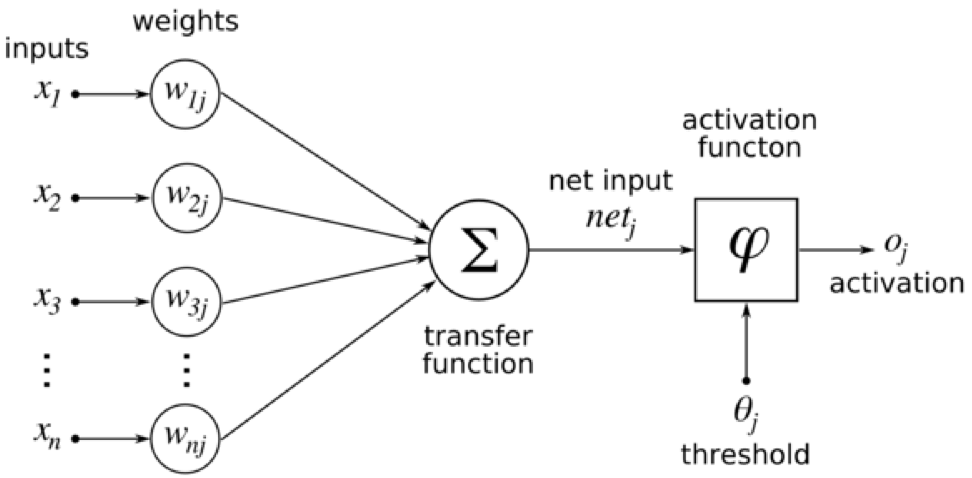
Frank Rosenblatt

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



Perceptron


impulses carried
toward cell body
dendrites
branches
of axon
nucleus
axon
axon
terminals
impulses carried
away from cell body
cell body


inputs
weights
x_1
x_2
x_3
x_n
w_{1j}
w_{2j}
w_{3j}
w_{nj}
Σ
transfer
function
net input
net_j
activation
functon
φ
o_j
activation
θ_j
threshold

Marvin Minsky

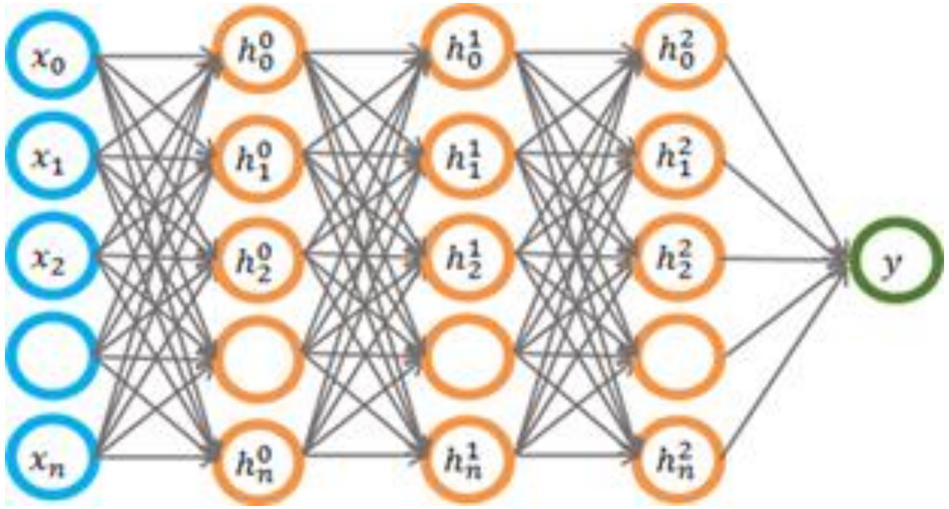
- In 1969, Minsky proposed the famous **XOR** problem and the inability of *Perceptron* in such linearly inseparable data distributions.
- It was the Minsky's tackle to the NN community. Thereafter, NN researches would be dormant up until 1980s.

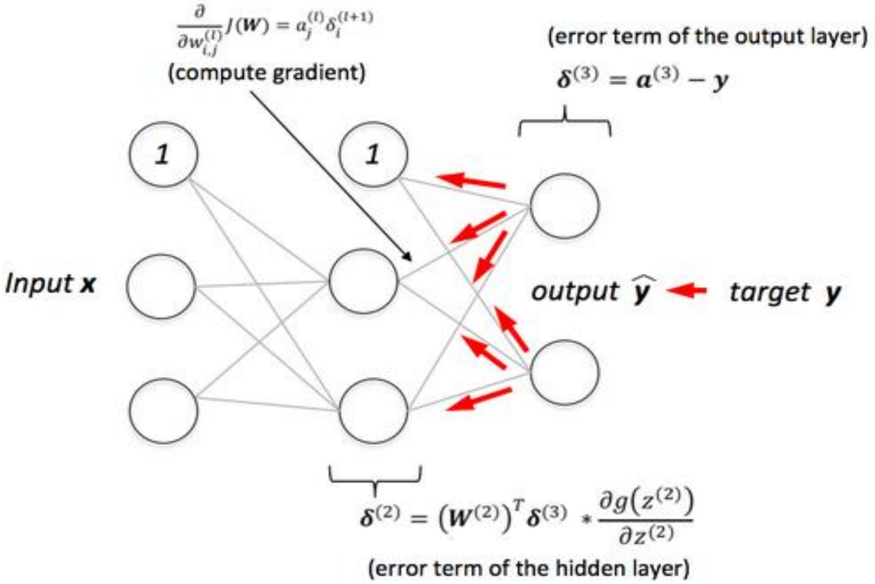
Perceptron is too simple, more complicated models are needed to handle complex problems…

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.

Multi-layer Perceptron / Deep Neural Networks





Universal Approximation Theorem

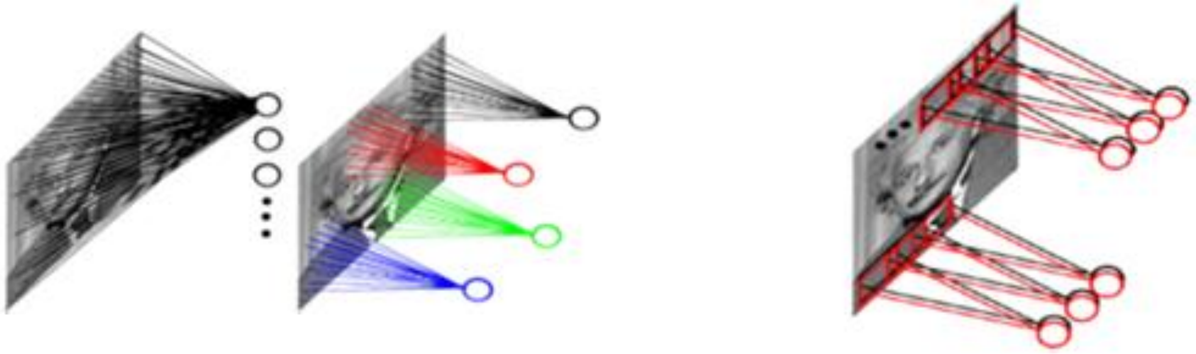
- A feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets of R^n, under mild assumptions on the activation function.

Geoffrey Hinton, Yan LeCun, Jurgen Schmidhuber

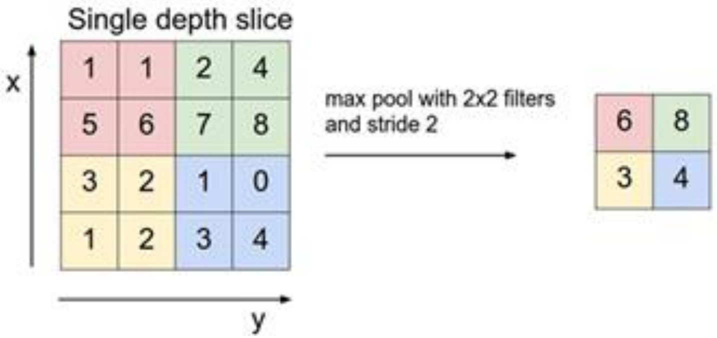
- Geoffrey Hinton contributed a lot to the practical backpropagation algorithms (1986) and Boltzmann Machines (1983).
- Yan LeCun was the first to train a convolutional neural network on images of handwritten digits (1986).
- Jurgen Schmidhuber invented a new type of recurrent neural network called Long 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.

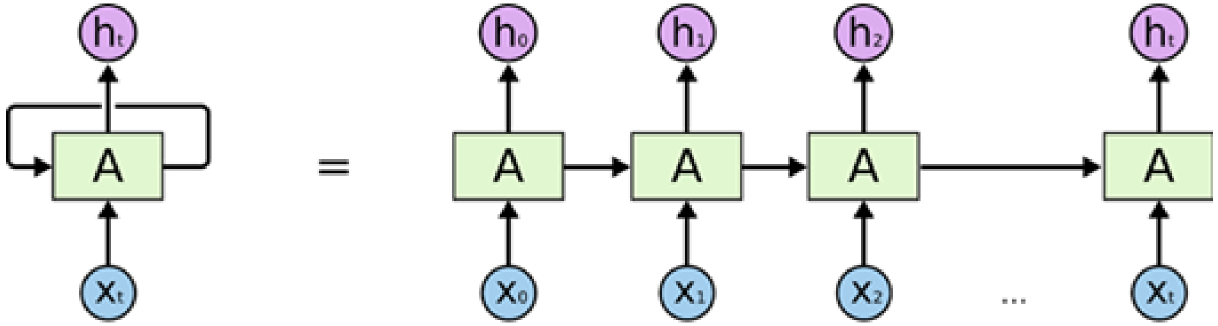


- Convolution:
  – Local connection, pattern recognition
- Weight sharing and pooling
  – Invariance
  – Parameter efficiency



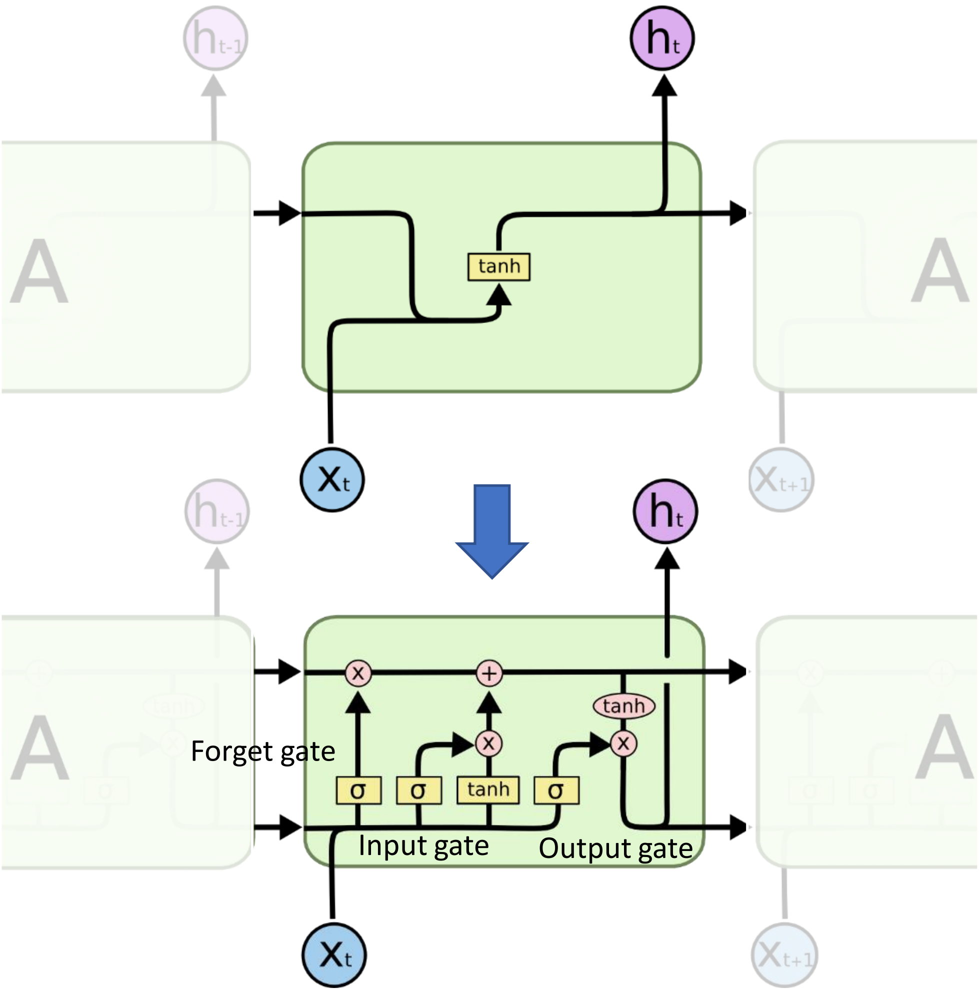
Recurrent Neural Networks (RNN)

- 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 $(x_t, \ldots, x_1)$ 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



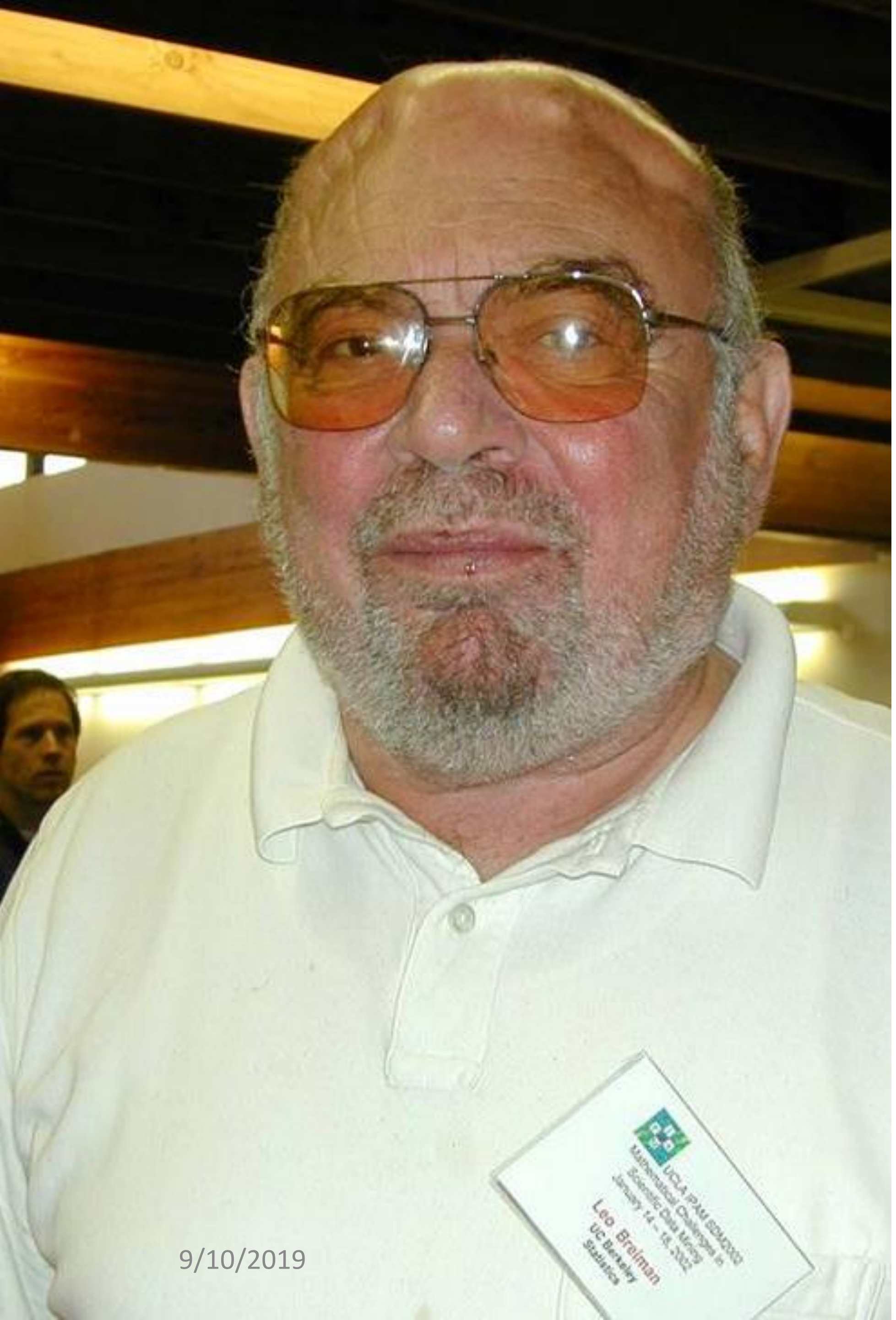
Neural networks are black boxes, and therefore difficult to interpret…

Ross Quinlan

- Decision trees were proposed by Ross Quinlan in 1986, 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

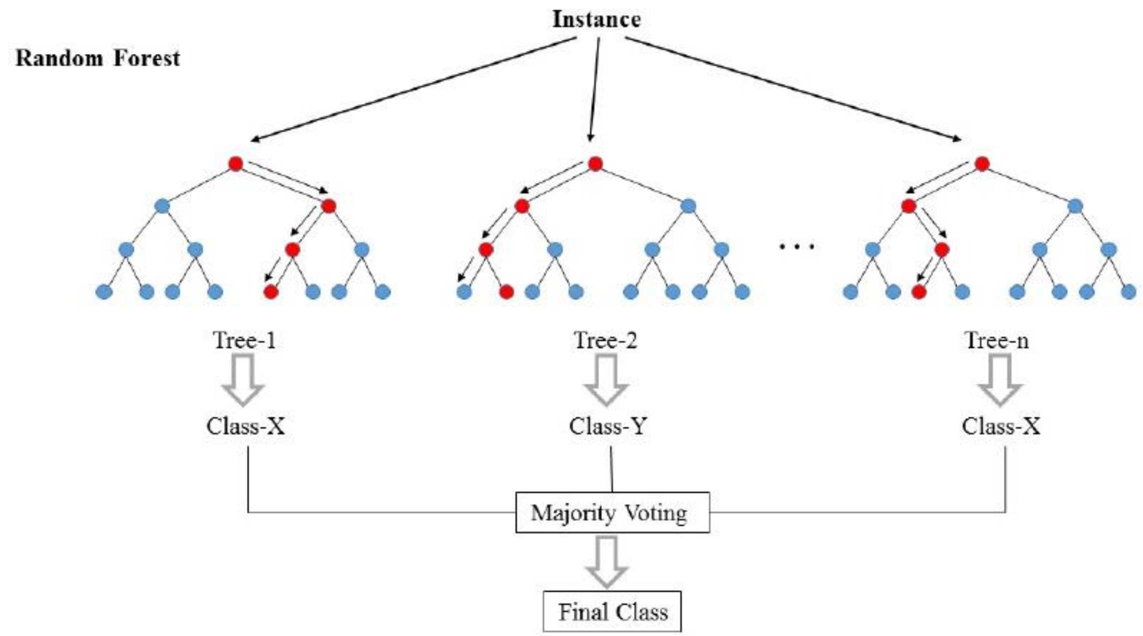


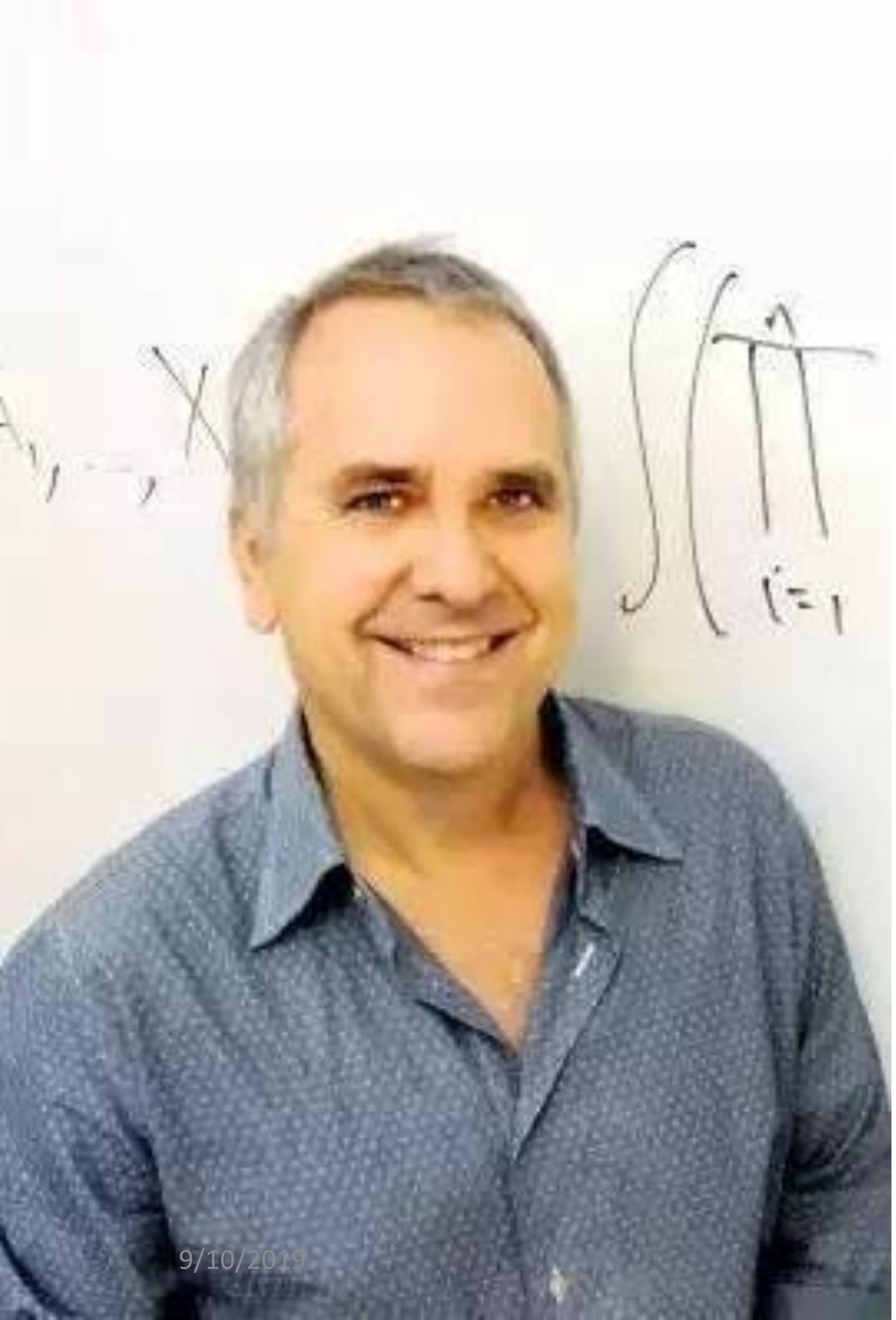
Leo Breiman

- 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



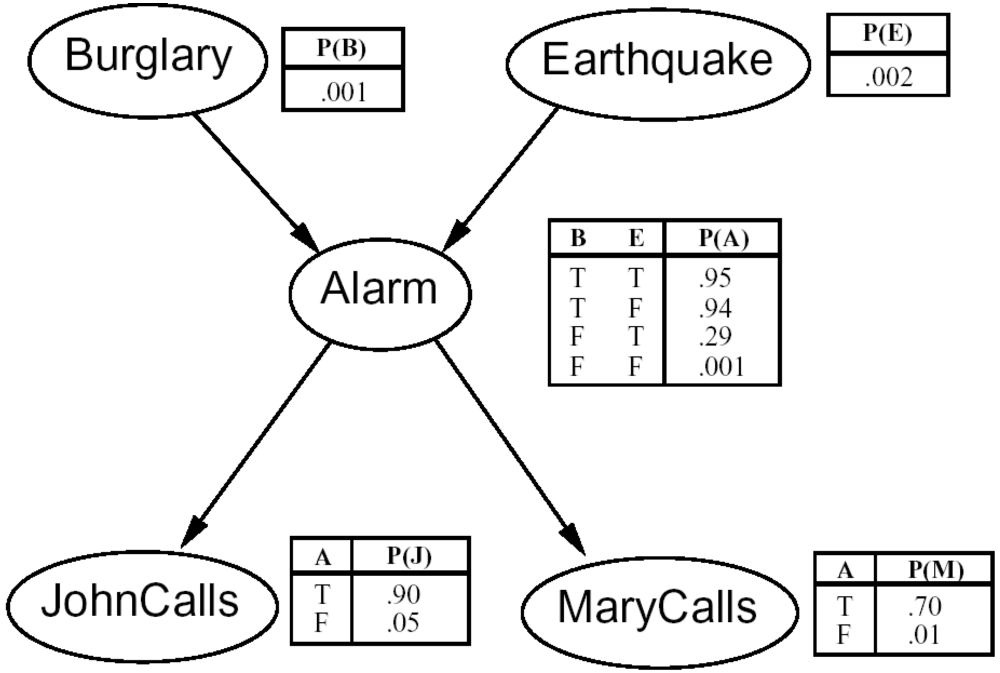


Michael Jordan

- 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).

Bayesian Networks

- 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



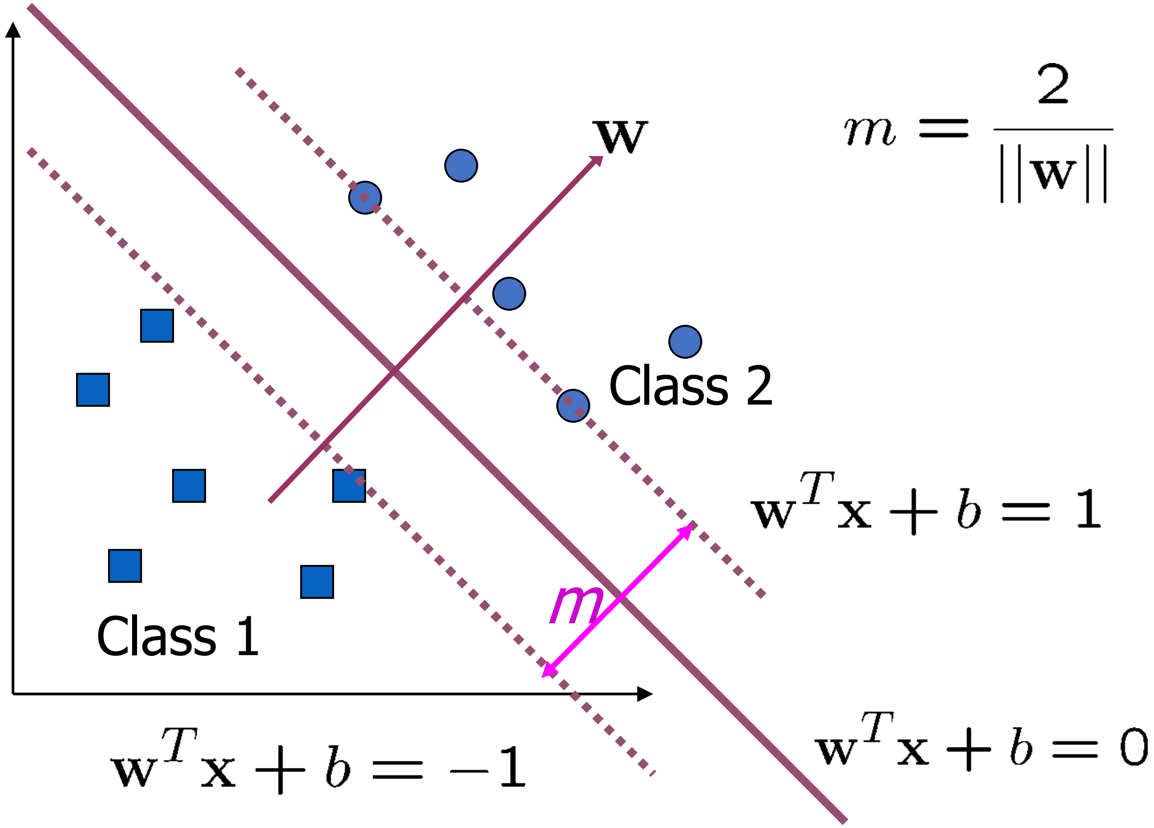
Neural networks are data-hungry. When there are only small number of training data, they will overfit …

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*

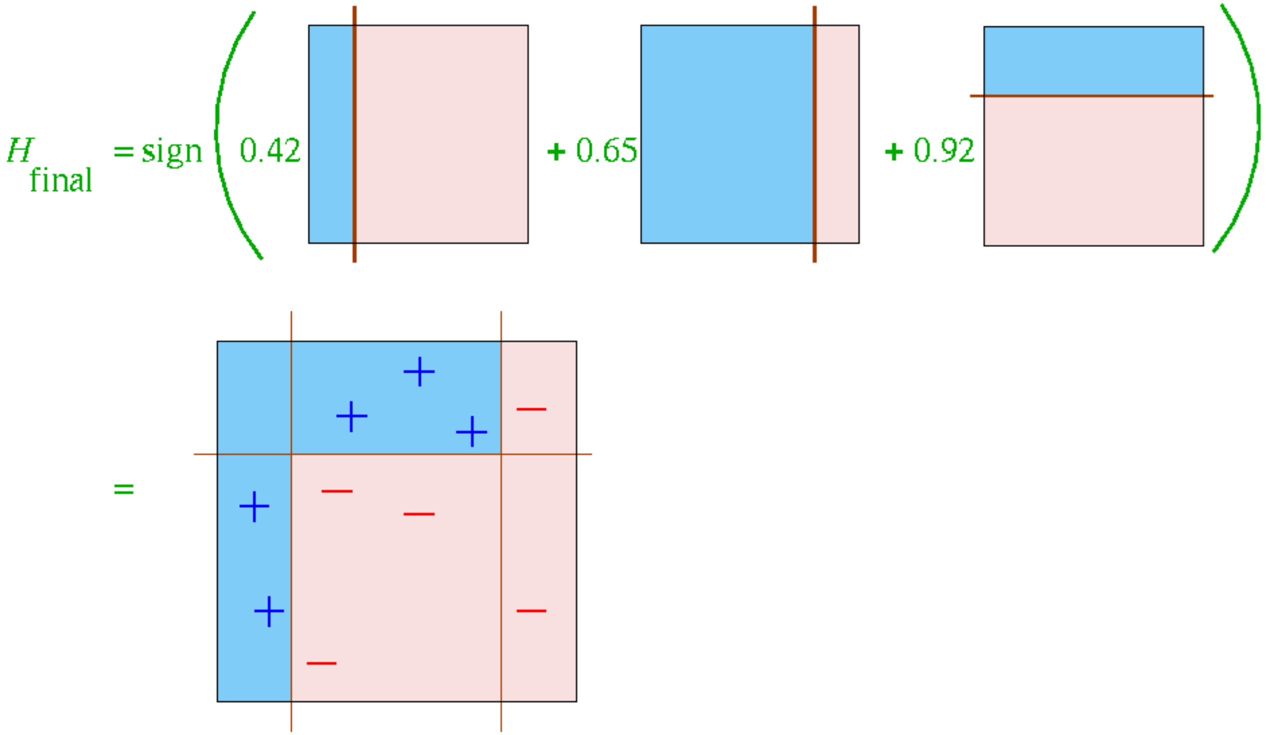


Yoav Freund & Robert Schapire

- 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.

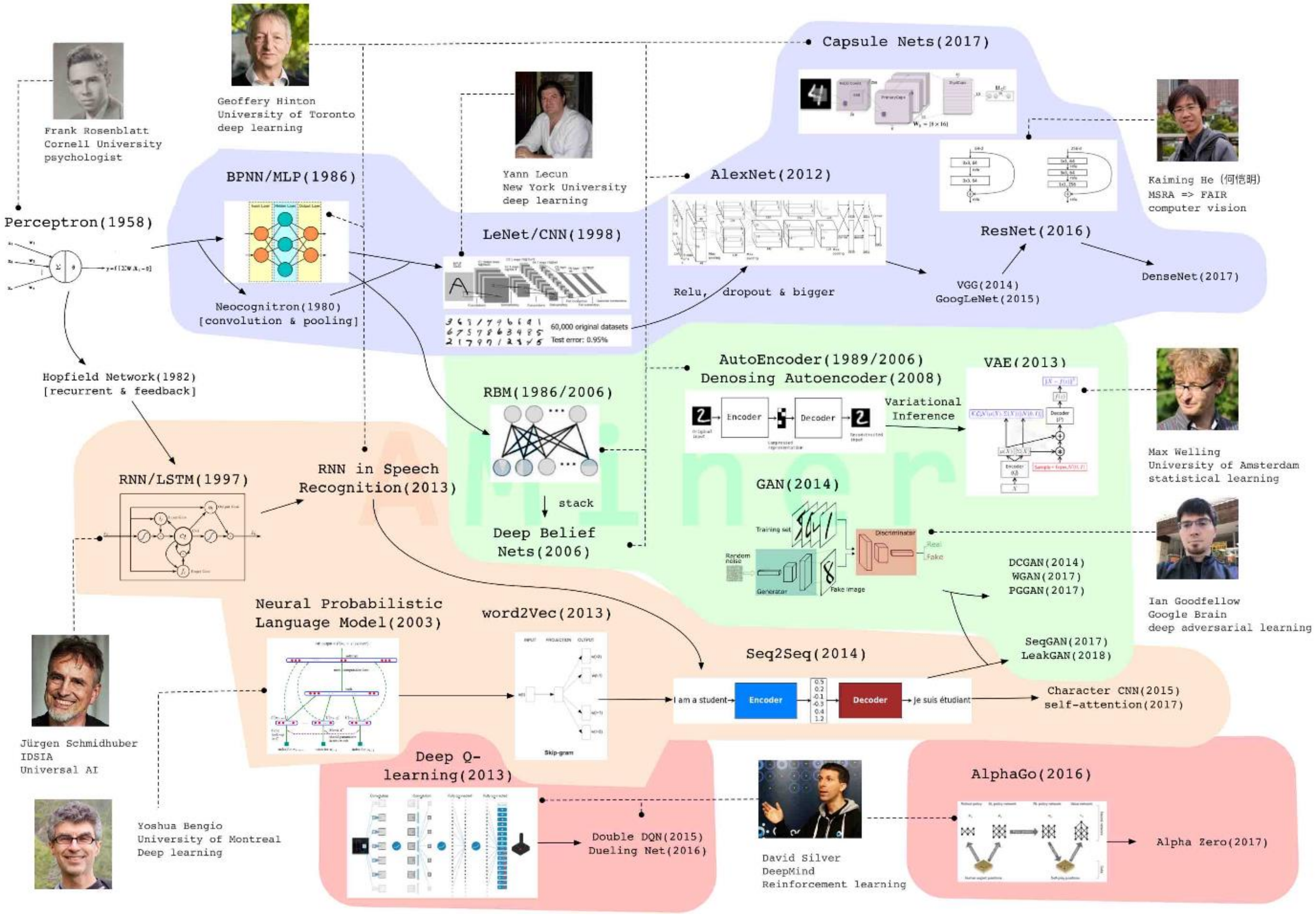
Boosting

- 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



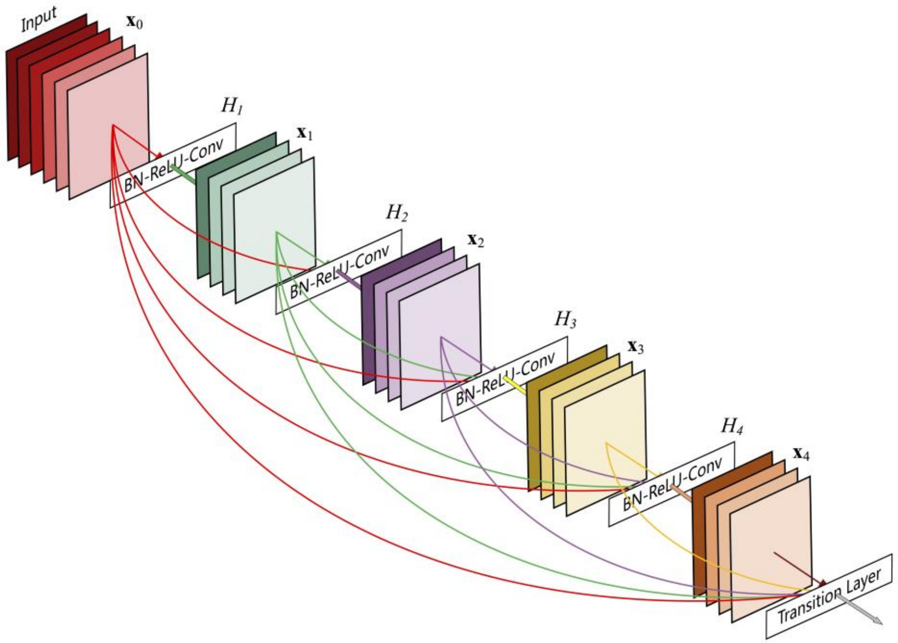
In today’s big-data era, sufficient training data make the outstanding expressiveness of neural networks a huge advantage ...

Revival of Neural Networks (Deep Learning)

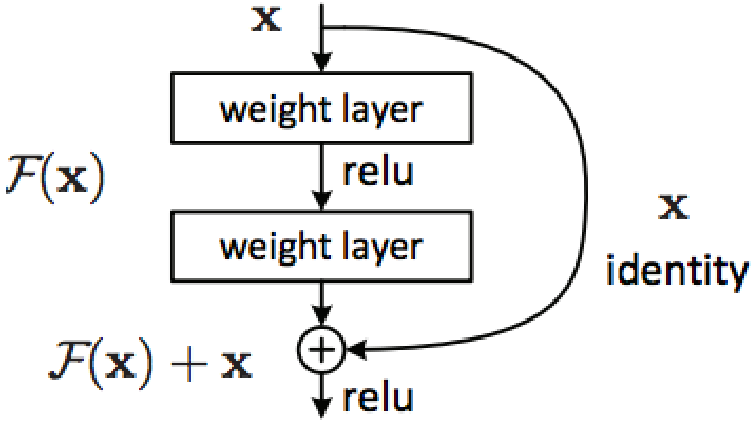


Very Deep Neural Networks

- DenseNet



ResNet



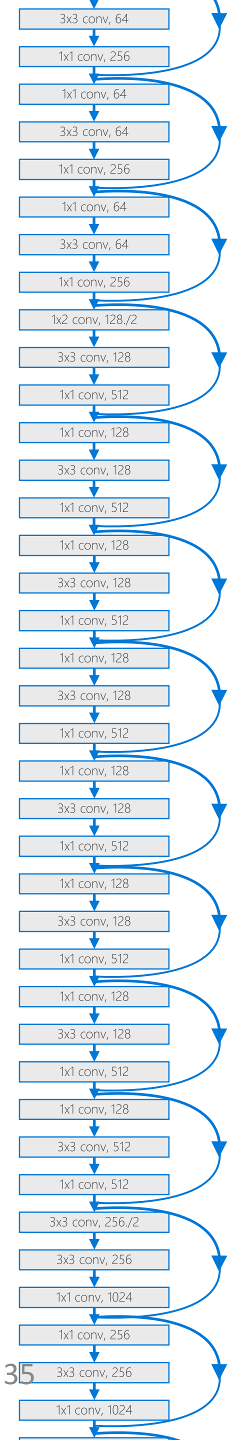
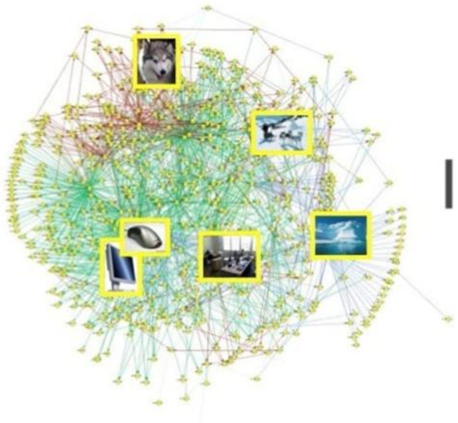
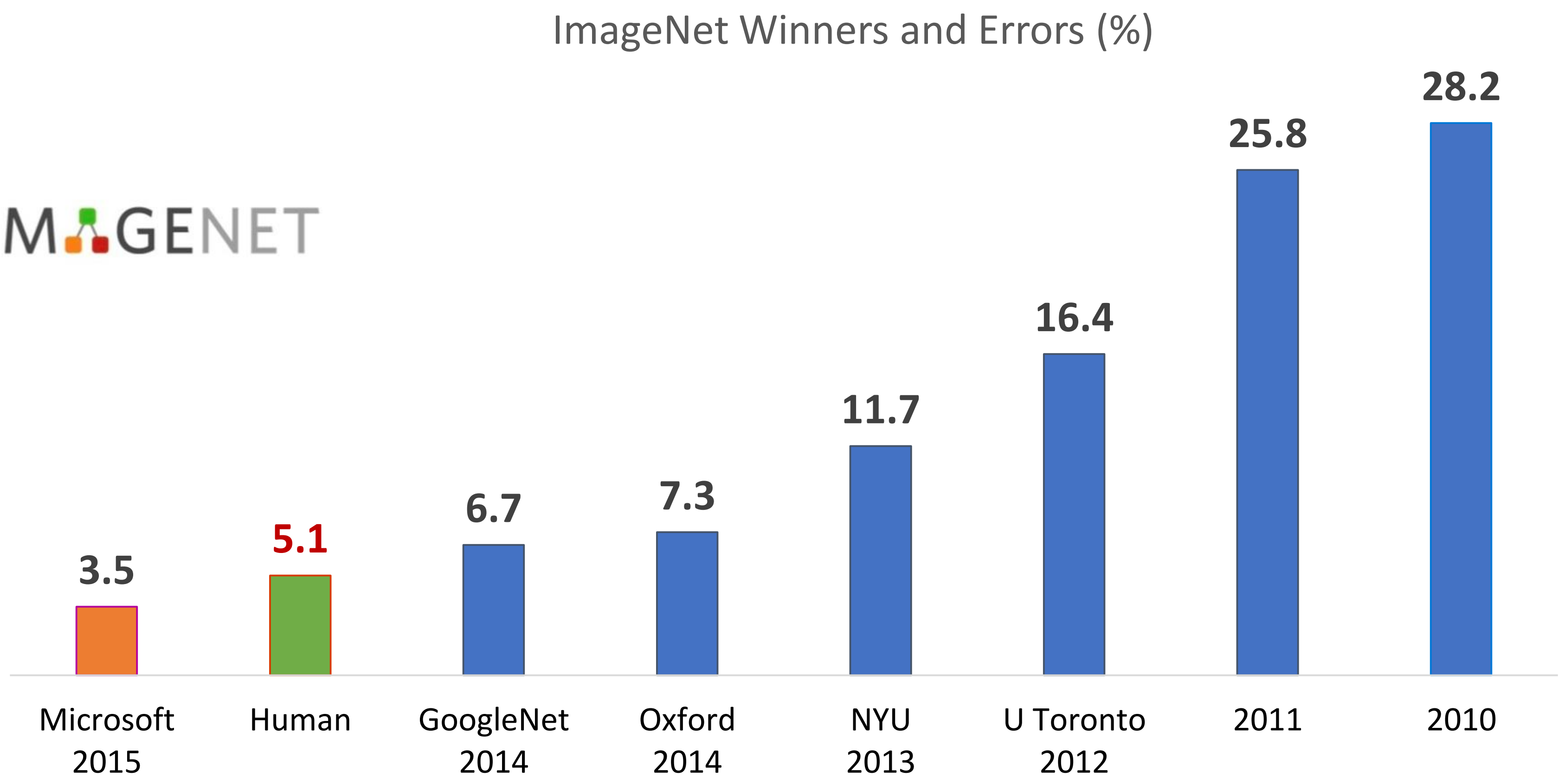


Image Recognition

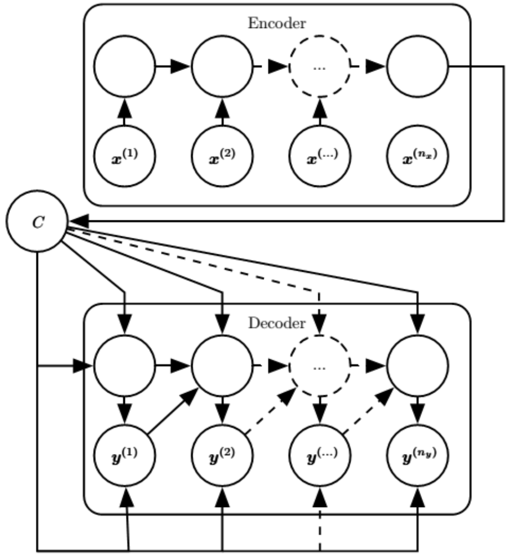


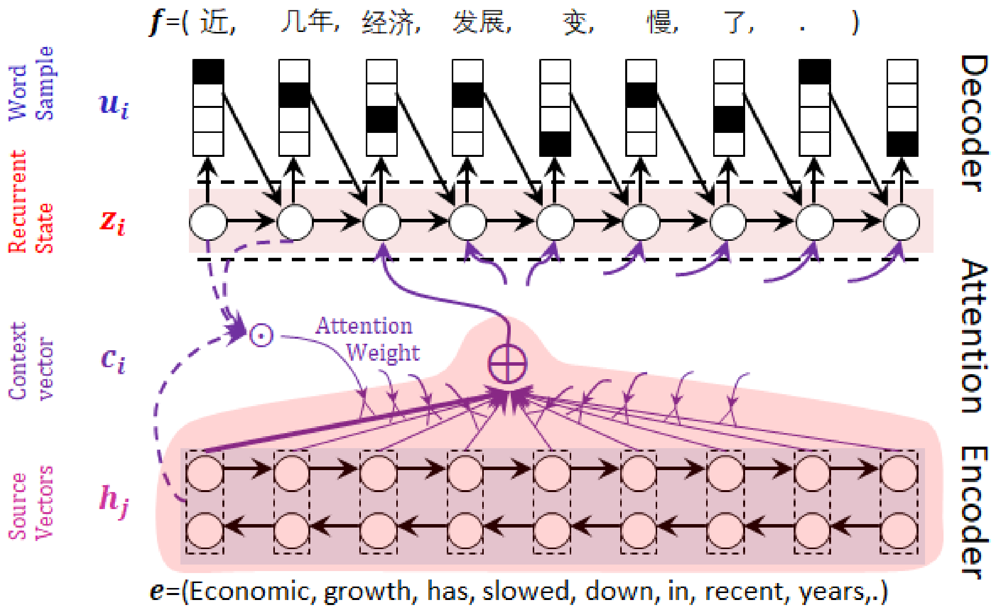
IMAGENET



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

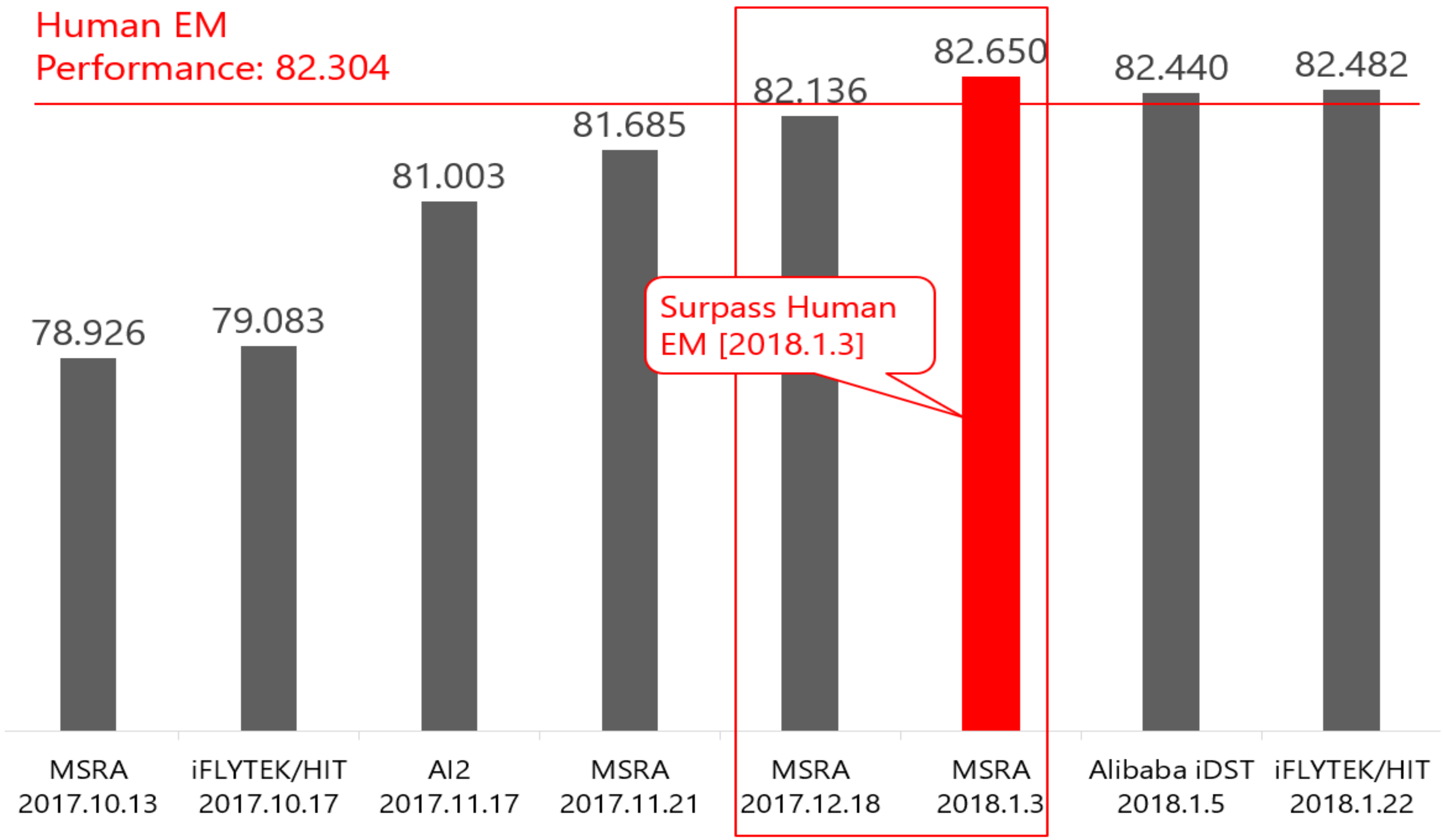
March 14, 2018 | Allison Linn

Subjective score: 69.5

Human: 69.0

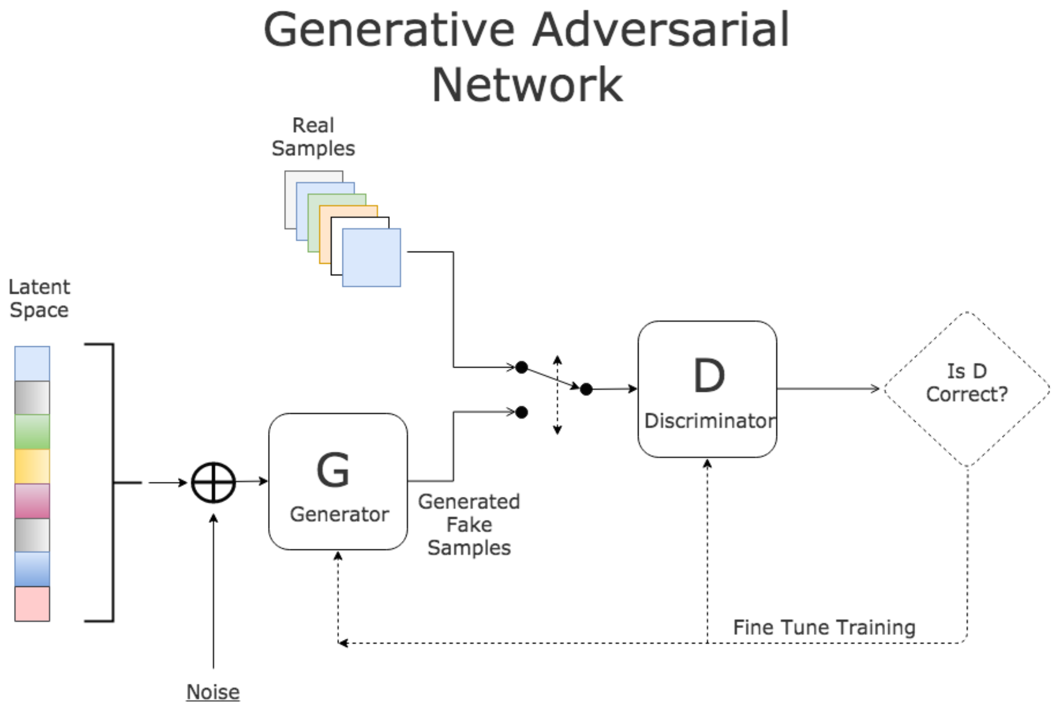
Reading Comprehension

- SQuAd



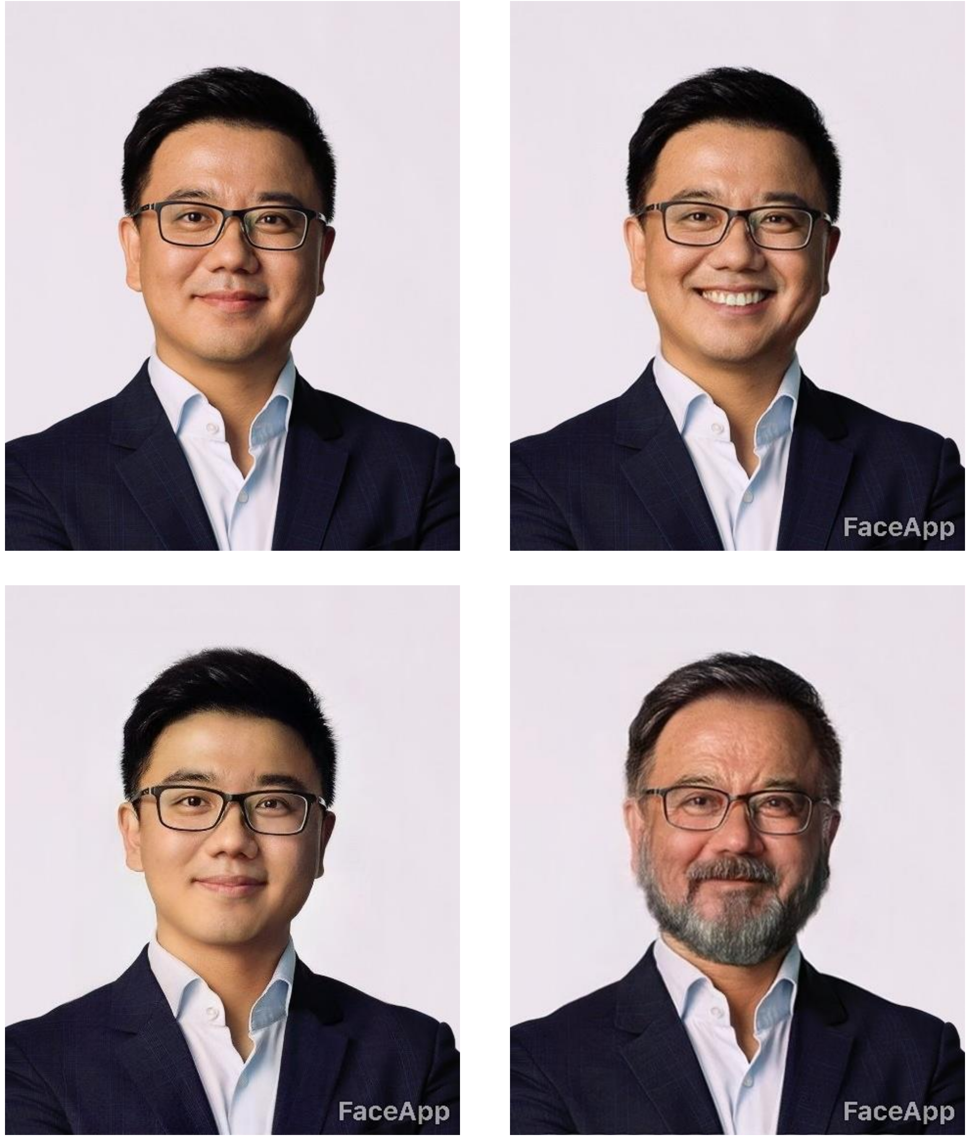
Generative Adversarial Networks

Ian Goodfellow (together with Bengio) proposed Generative Adversarial Networks (GAN) in 2010. Now GAN has been applied to computer vision, speech, and languages, and is the state-of-the-art of generative models.



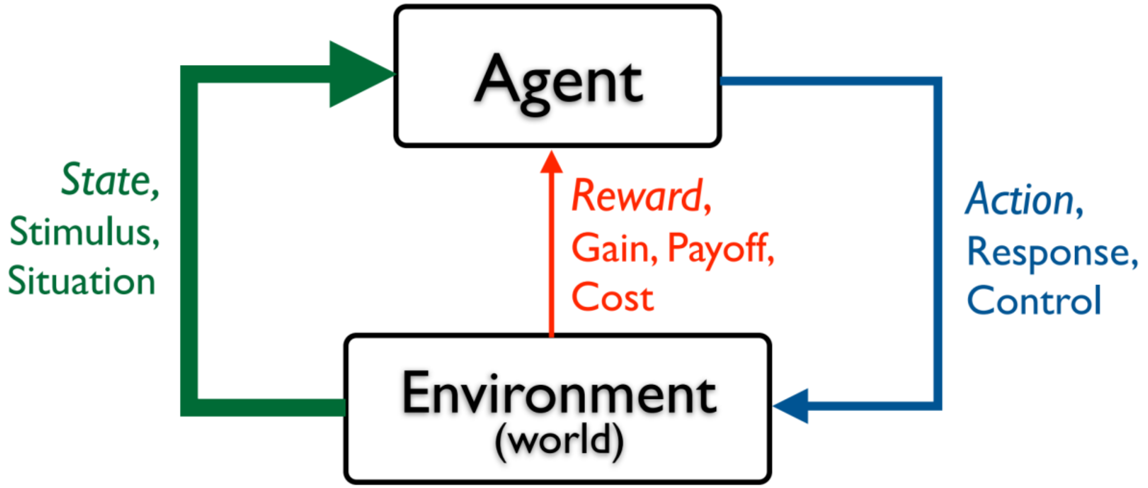
- Generator captures the data distribution
- Discriminator estimate the probability that a sample came from the training data rather than the generator

Deep Fake

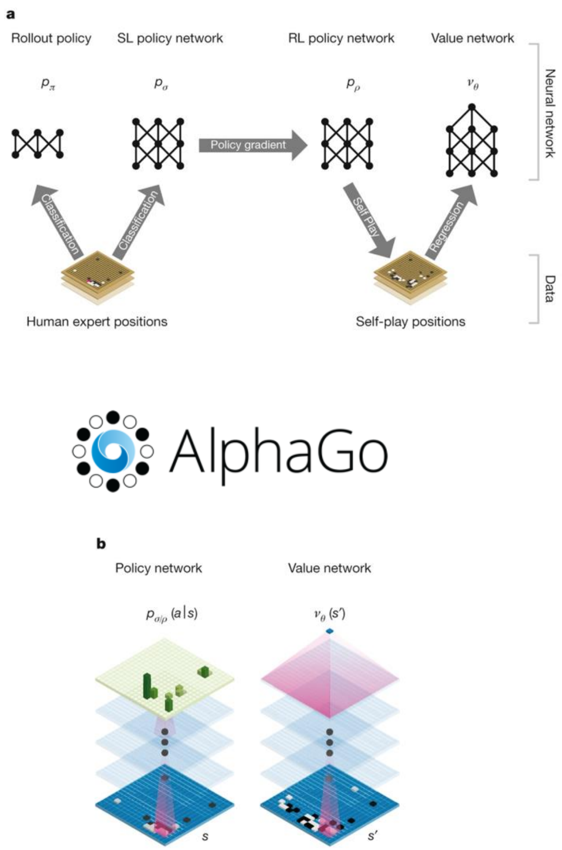


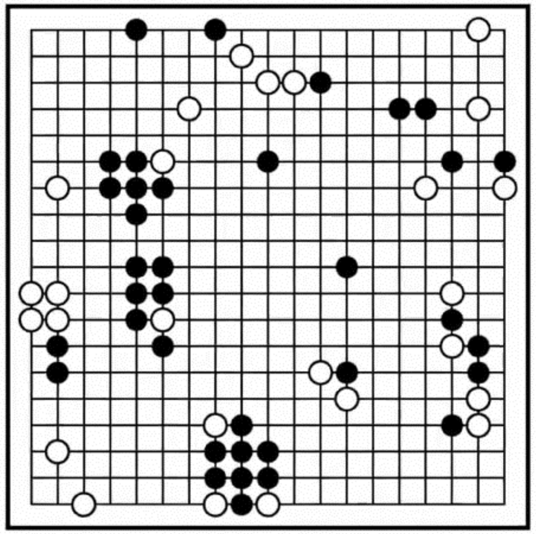
Deep Reinforcement Learning

- 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



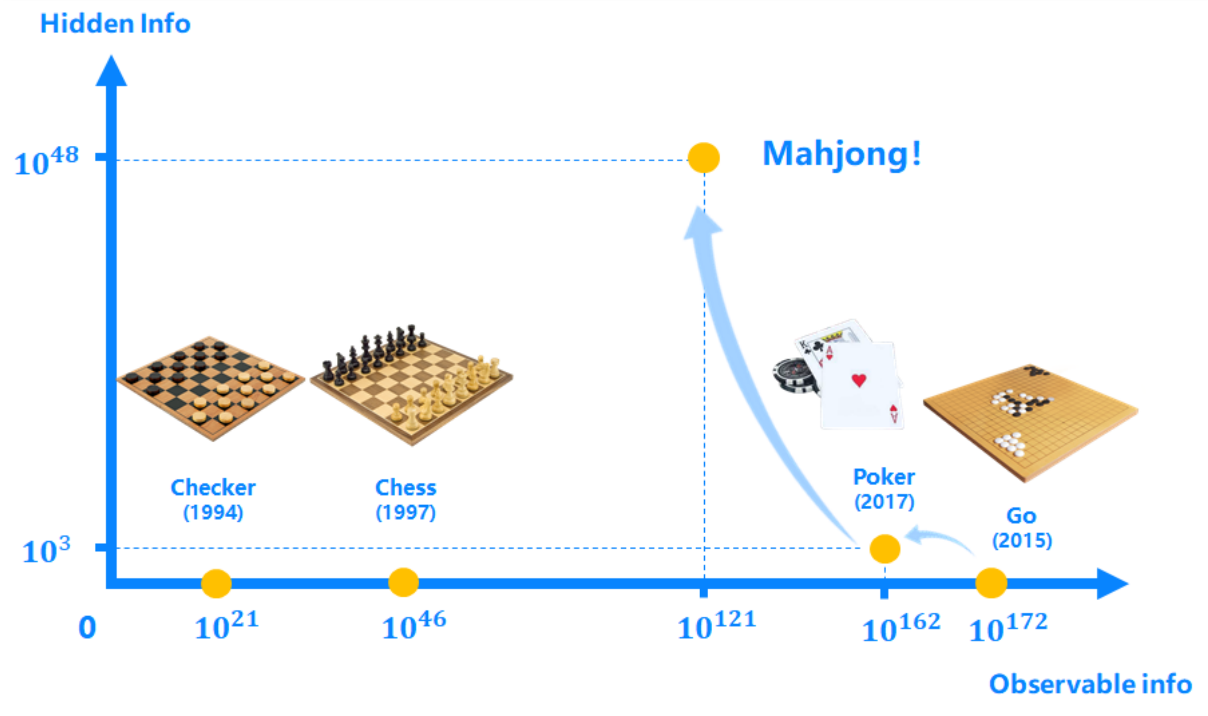


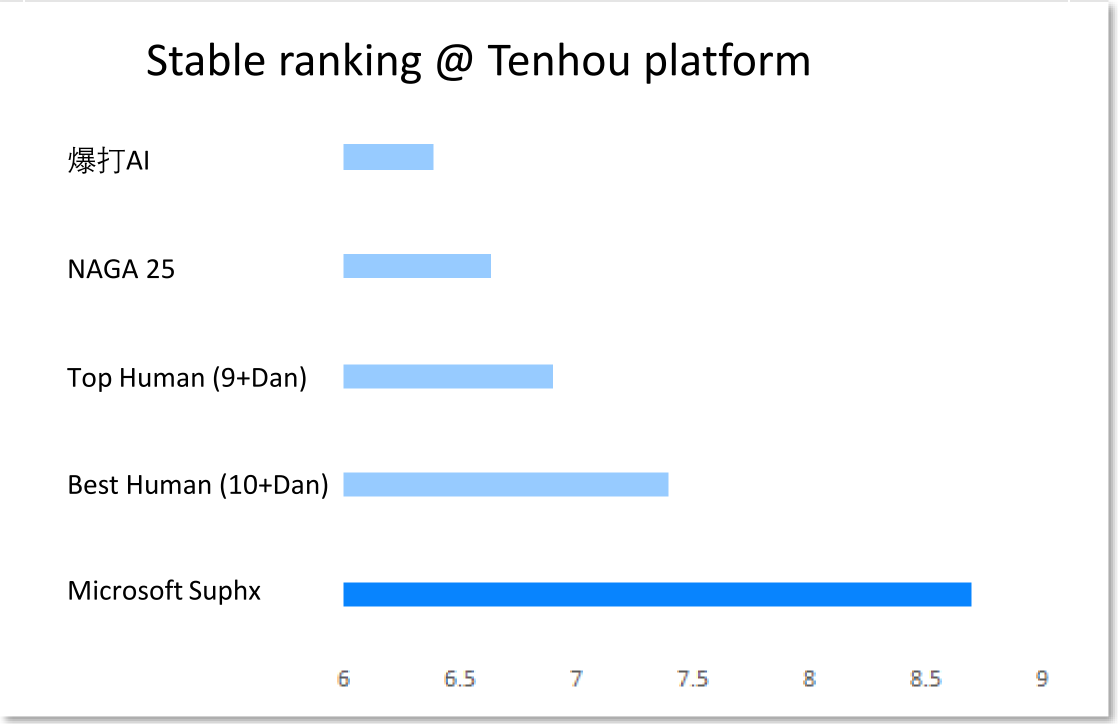
4:1 against Sedol Lee

3:0 against Jie Ke

Mahjong Playing: Suphx

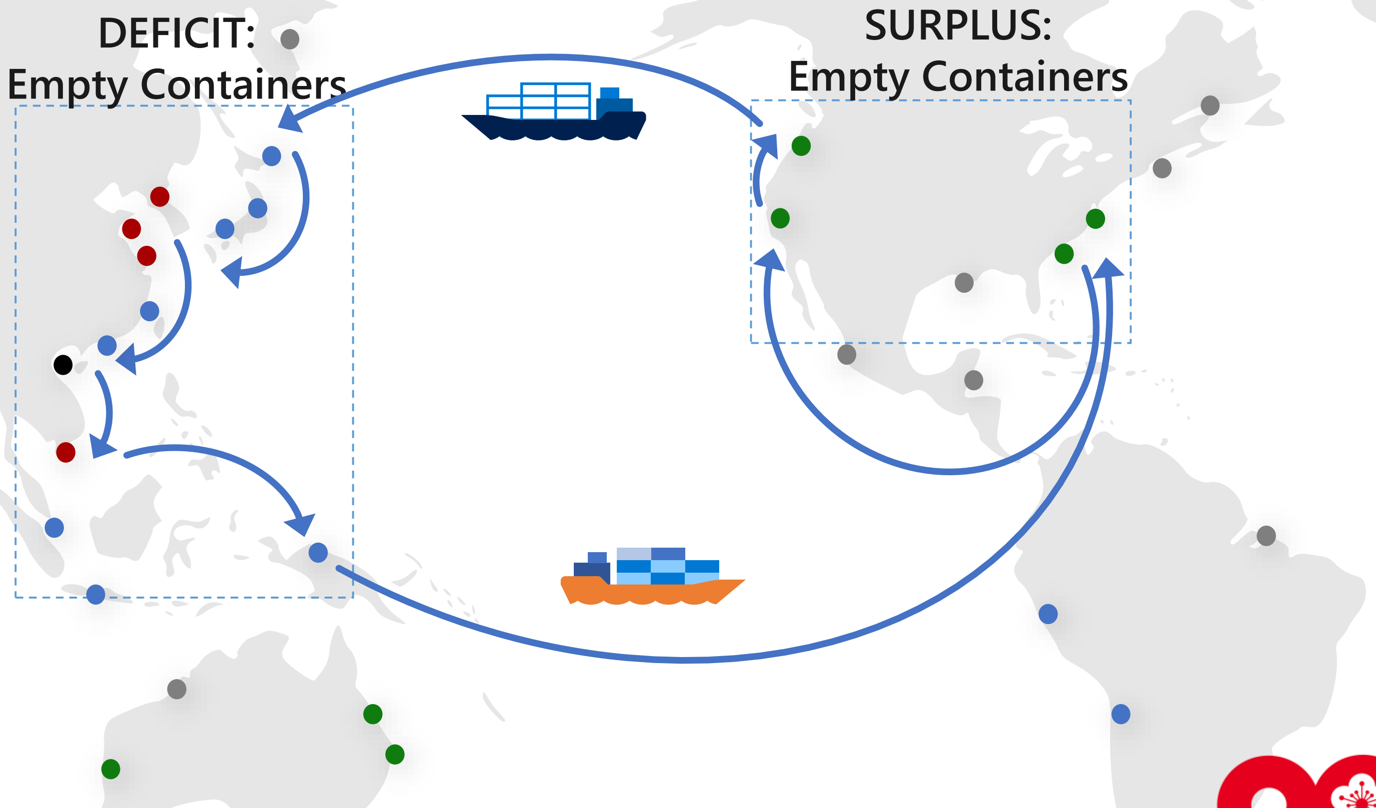
- Deep RL + Oracle critic + Policy adaptation





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

THE CHALLENGE

Challenges of Machine Learning

- Relying on big training data
- Relying on big computational power
- Lack of interpretability (blackbox learning)
- Omitting many key factors of human intelligence

Relying on Big Training Data

- Cannot live without huge amount of human-labeled training data

Tasks	Typical training data
Image classification	Millions of labeled images
Speech recognition	Thousands of hours of annotated voice data
Machine translation	Tens of millions of bilingual sentence pairs
Go playing	Tens of millions of expert moves

Human labeling is very costly; not to mention that for many applications, it is simply impossible to obtain large-scale labeled data (e.g. rare diseases, minority languages)

Relying on Big Computation

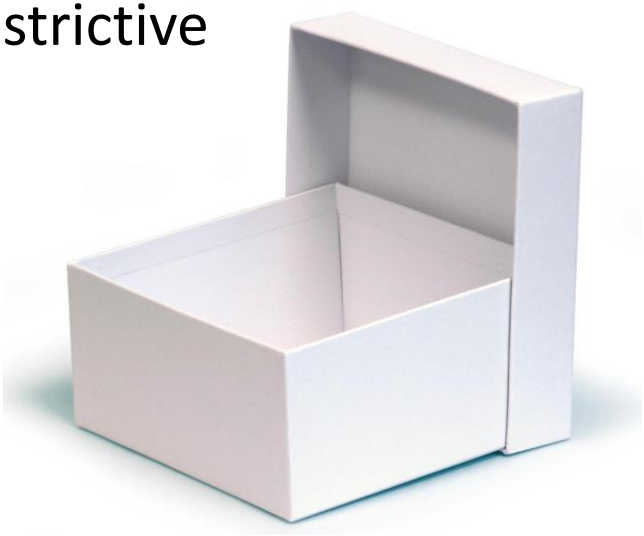
- Big data + big model + heavy learning algorithms → big computational cost for both training and inference

Tasks	Time
Image classification (ResNet)	8 K80, 3 weeks
Machine translation (Google)	96 K80, 6 days
AlphaGo inference: distributed version	1,202 CPUs and 176 GPUs
BERT for pre-training	64 TPUs, 4 days
BigGAN for image synthesis	256 TPUs, 2 days
XLNet for pre-training	512 TPU, 2.5 days

- Monopolization and Matthew effect

Lack of Interpretability

- Deep models are like black boxes
  - Predictions and decisions are not explainable for most deep models
  - Once a DNN model with billions of parameters makes a mistake, it is difficult to diagnose
  - Applications to some domains are therefore restrictive



Far from Human Intelligence

- Current deep learning does not reveal why human can be much more intelligent than animals
  - Bigger brain?
  - The role of language?
  - Knowledge accumulation and transfer?
  - Social collaboration?
  - Teaching system?
  - Consciousness?

Still a lot of research to do…

What will you learn from this course?

Course Outline

- **第一部分：机器学习基础 (6学时)**
  - 第一讲：课程导论（1学时） - （刘铁岩，张旭东）9.12
  - 第二讲：机器学习框架（4学时） - （刘铁岩，张辉帅，李亚利）9.12/19
  - 第三讲：机器学习流程（1学时） - （柯国霖，陈健生）9.19
- **第二部分：高级机器学习模型 (9学时)**
  - 第四讲：深度学习（6学时） - （刘铁岩，张辉帅，李亚利）9.26/ 10.10
  - 第五讲：梯度提升树（1.5学时） - （柯国霖，陈健生）10.17
  - 第六讲：生成模型（1.5学时） - （刘畅，陈健生）10.17
- **第三部分：高级机器学习应用 (9学时)**
  - 第七讲：计算机视觉（3学时） - （秦涛，陈健生）10.24
  - 第八讲：自然语言处理（3学时） - （秦涛，张卫强）10.31
  - 第九讲：金融科技（1.5学时） - （边江，张卫强）11.7
  - 第十讲：生物信息学（1.5学时） - （邵斌，张卫强）11.7

Course Outline

- **第四部分：机器学习前沿 (15学时)**
  - 第十一讲：强化学习/机器博弈（4学时） - （秦涛，陈健生）11.14/21
  - 第十二讲：元学习/教学相长（2学时） - （夏应策，李勇）11.21
  - 第十三讲：对抗学习（3学时） - （贺笛，夏应策，李勇）11.28
  - 第十四讲：对偶学习（1学时） - （夏应策，李勇）12.5
  - 第十五讲：迁移学习（2学时） - （王晋东，李勇）12.5
  - 第十六讲：模型压缩/边缘计算（1学时） - （郑书新，李勇）12.12
  - 第十七讲：分布式机器学习（2学时） - （陈薇，李勇）12.12
- **第五部分：课程总结（6学时）**
  - 第十八讲：课程大作业汇报会（5学时） - （刘铁岩，秦涛，张旭东）12.19/26
  - 第十九讲：机器学习的技术发展趋势（1学时） - （刘铁岩主持圆桌会议）12.26

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

Evaluations

- Class attendance (20%)
  - Check-in
  - Classroom test
- 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
  - Design new machine learning solutions and conduct experiments
  - Write project reports and make presentations

Course Projects: Algorithm Design

	Focus	Task
A1	Design a better NN structure: outperforms SoTA models	Machine translation
A2		Text summarization
A3		Mahjong tile prediction
A4		Protein Contact Prediction
A5	Design a light/compact model: reduce at least 90% parameters while keeping accuracy	Machine translation
A6		Text pre-training
A7		Image super resolution
A8		Image classification
A9	Design an incremental learning algorithm for GBDT	Display advertising
A10	Design a better distributional reinforcement learning algorithm	Atari games
A11	Semantic variational auto-encoder	Image generation

Course Projects: Public Challenges

	Challenges	Description
C1	SQuAD2.0	The Stanford Question Answering Dataset
C2	CoQA	A Conversational Question Answering Challenge
C3	ARC	AI2 Reasoning Challenge
C4	GNQ	Google Natural Questions
C5	RACE	Reading Comprehension Dataset

Course Projects: Theoretical Analysis

	Focus	Description
T1	General deep learning	Reduce the over-parameterization requirement for training deep neural networks
T2		Build PAC-Bayesian generalization bound with normalized flat minima
T3	Transfer learning	Investigate the task/dataset similarity in transfer learning
T4	Dual learning	Derive a generalization bound for dual semi-supervised learning
T5		Derive a tighter generalization bound for dual supervised learning

References

1) Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press
2) Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning, Springer.
3) Christopher M Bishop, Pattern Recognition and Machine Learning, Springer
4) Andrew Ng, Machine Learning Yearning.
5) 周志华，机器学习，清华出版社
6) … …

Thanks