Machine Learning

lanyanyan@ict.ac.cn

Self-Introduction: Yanyan Lan(兰艳艳)

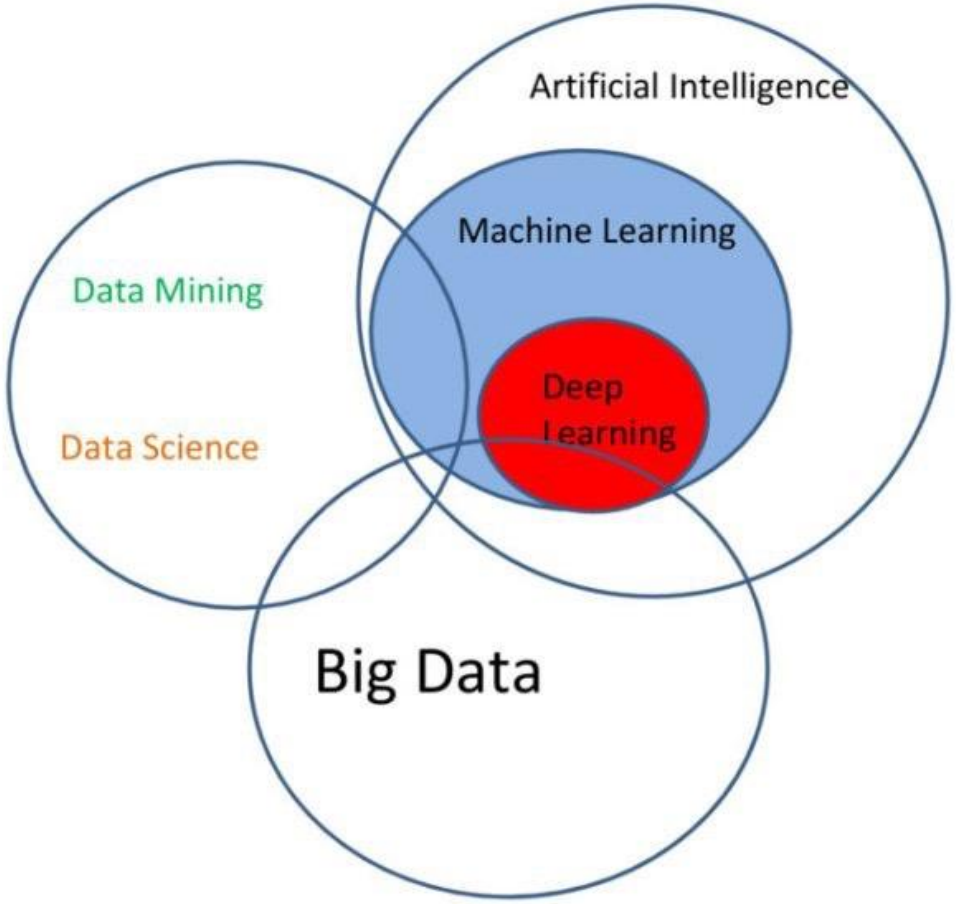
- Associate Professor in Institute of Computing Technology, Chinese Academy of Sciences, a research group leader in (CAS) Key Lab of Network Data Science and Engineering
- Research Interest: Machine Learning, Information Retrieval, Data Mining, and Applications.
  - 2006-2012: Learning to Rank (Search & Recommendation)
  - 2013-2018: Deep Learning for Text Data Analysis (Search, QA and Dialogue)
  - 2018 till now: Learning to Reasoning (QA and Dialogue)
- Homepage: www.bigdatalab.ac.cn/~lanyanyan
- 12 hours' course
  - Statistical Learning Framework, Supervised Learning, SVM

Survey

Some Suggestions

- Teaching
  - Top 30%
  - English Slides
  - Elements
- Requirements
  - Knowledge of basic computer science principles and skills
  - Familiarity with the probability theory
  - Familiarity with linear algebra
- Target
  - Basic: Exam
  - Engineering: Practice
  - Research: Theory + Practice

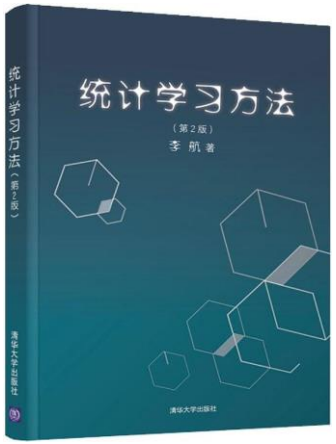
Relations with Other Fields

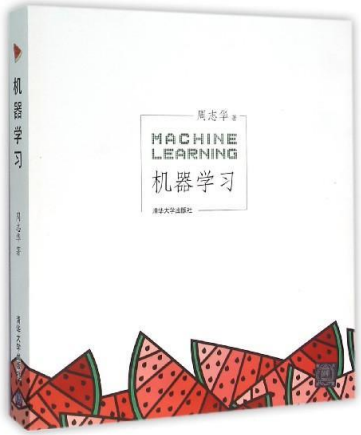


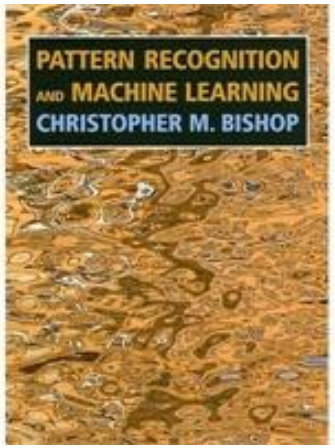
Courses, Books, and Conferences

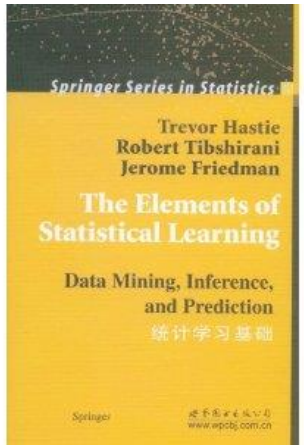
Machine Learning

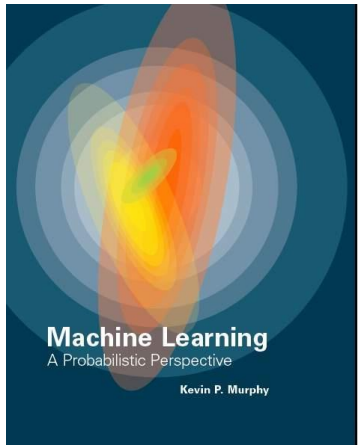
Neural Information Processing Systems Foundation











Courses, Books, and Conferences

- Stanford: CS229

  http://cs229.stanford.edu
- Berkeley: CS 189/289A

  https://people.eecs.berkeley.edu/~jrs/189/
- CMU: 10701

  http://www.cs.cmu.edu/~epxing/Class/10701/
- Cambridge: Information Theory, Pattern Recognition and Neural Networks

  http://www.inference.phy.cam.ac.uk/mackay/itprnn/

What is Machine Learning?

- Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.
- 1959, Arthur Samuel: "Field of study that gives computers the ability to learn without being explicitly programmed".
- Tom M. Mitchell: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". (PAC theory)

Defining the Learning Task

Improve on task, T, with respect to

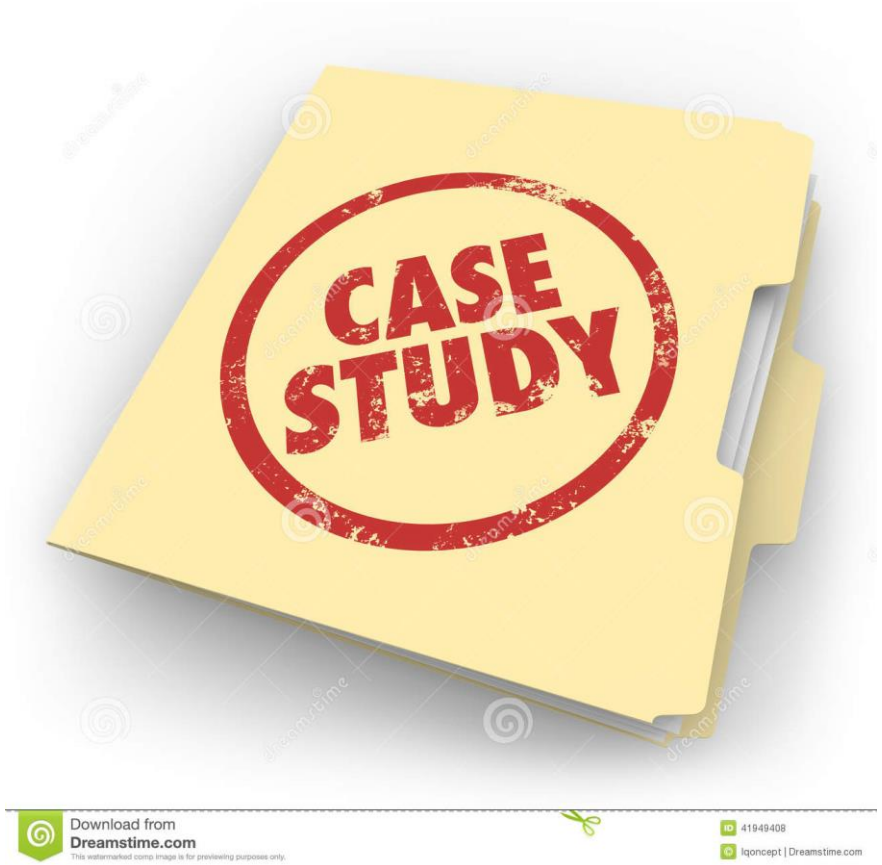
performance metric, P, based on experience, E.

T: Categorize email messages as spam or legitimate.
P: Percentage of email messages correctly classified.
E: Database of emails, some with human-given labels

T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors
P: Average distance traveled before a human-judged error
E: A sequence of images and steering commands recorded while observing a human driver.

T: Playing checkers
P: Percentage of games won against an arbitrary opponent
E: Playing practice games against itself



A Learning Comic

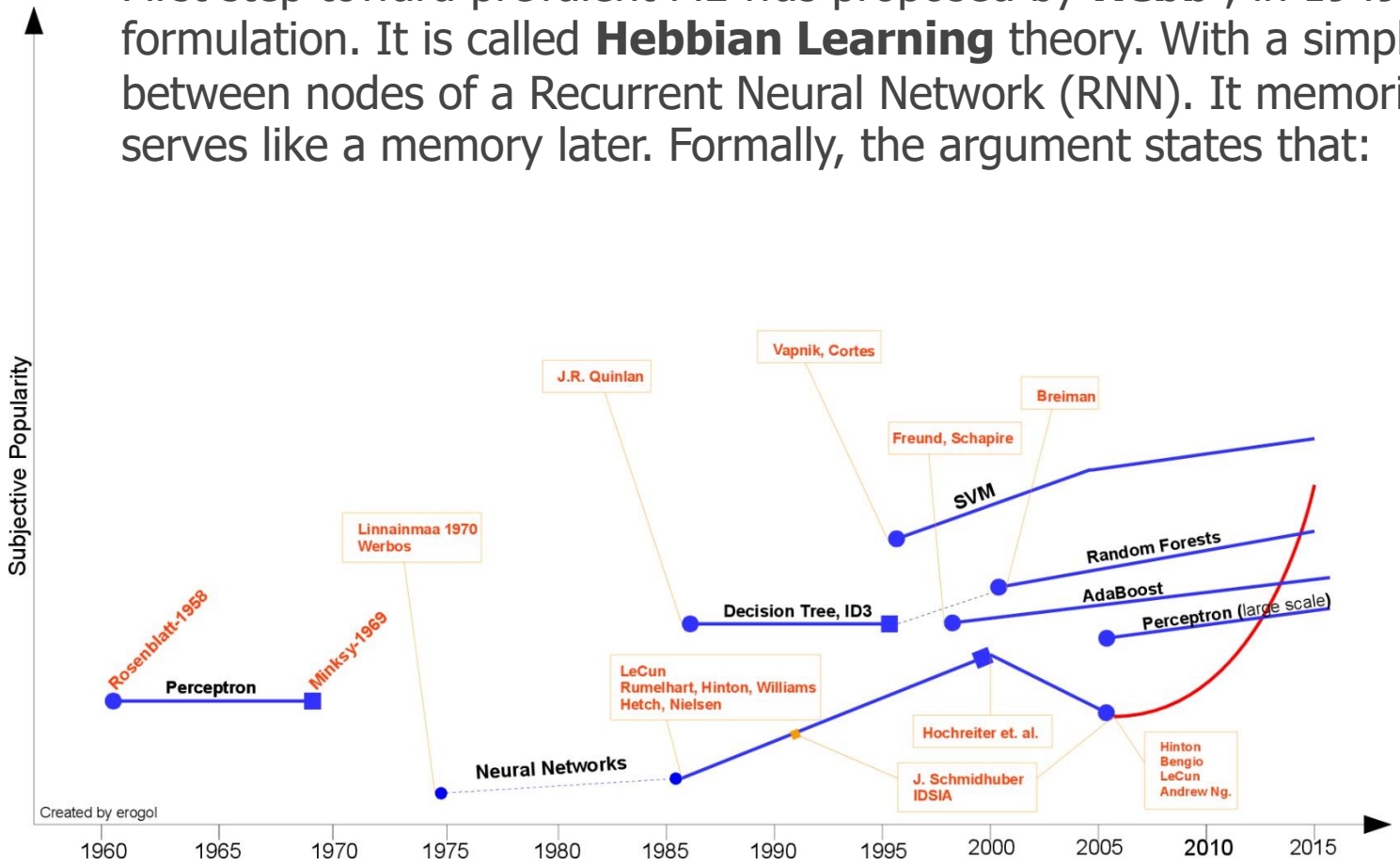
- Experience: Training Data
- What is to be learned: Target Function
- Learning Algorithm: how to infer the target function from the experience
- Evaluation: Test Data

Categories of Machine Learning

- Supervised Learning: direct labeled training data
- Unsupervised Learning: unlabeled training data
- Semi-Supervised Learning: unlabeled + labeled training data
- Reinforcement Learning: indirect labeled training data
- Transfer Learning: training and test data are Non IID
- Multi-Task Learning: multiple task share representation
- Active Learning: actively choose training data

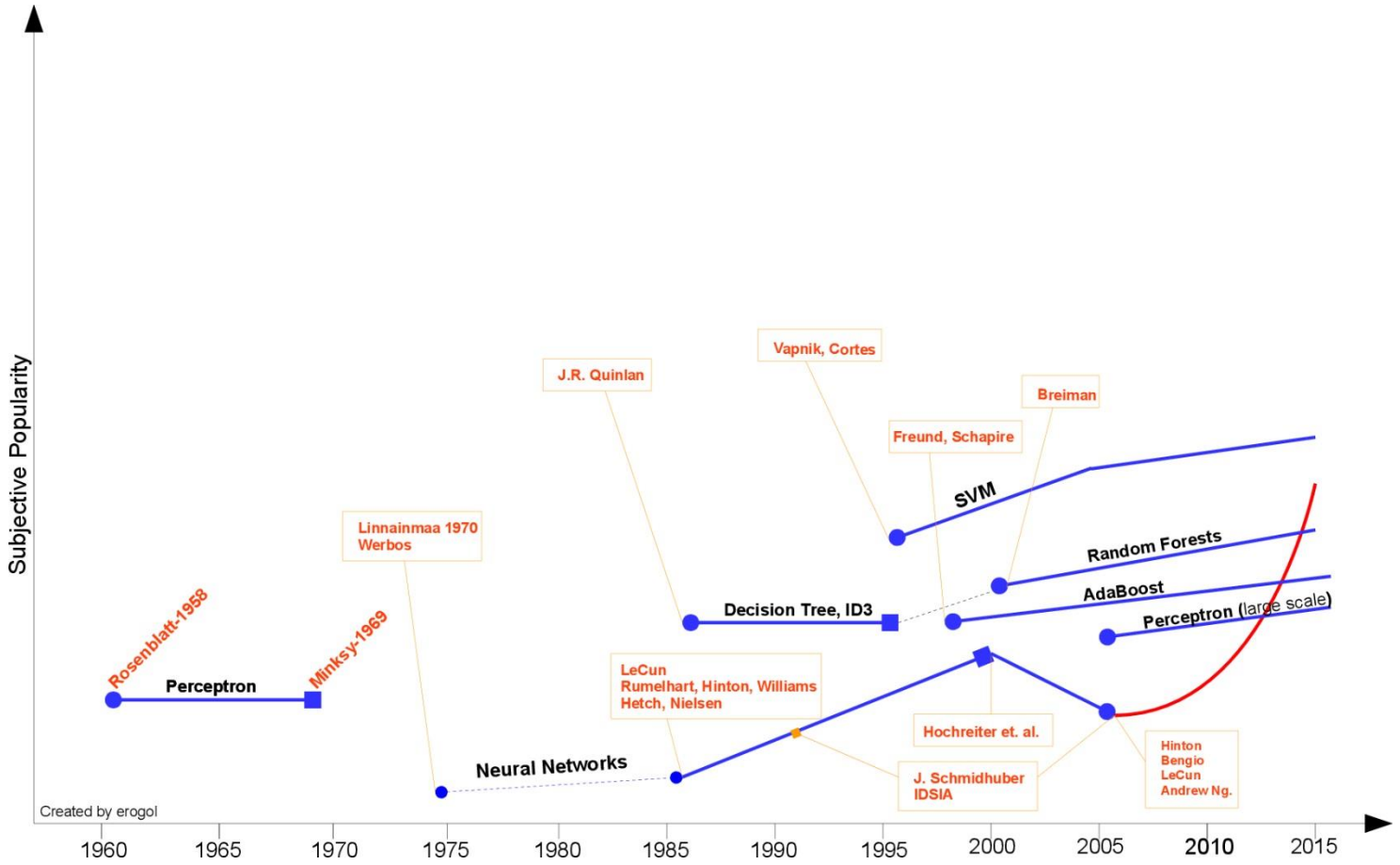
History of Machine Learning (From Eren Golge's Blog)

First step toward prevalent ML was proposed by **Hebb** , in 1949, based on a neuropsychological learning formulation. It is called **Hebbian Learning** theory. With a simple explanation, it pursues correlations between nodes of a Recurrent Neural Network (RNN). It memorizes any commonalities on the network and serves like a memory later. Formally, the argument states that:



Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability.... When an axon of cell *A* is near enough to excite a cell *B* and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that *A*'s efficiency, as one of the cells firing *B*, is increased.

History of Machine Learning (From Eren Golge's Blog)



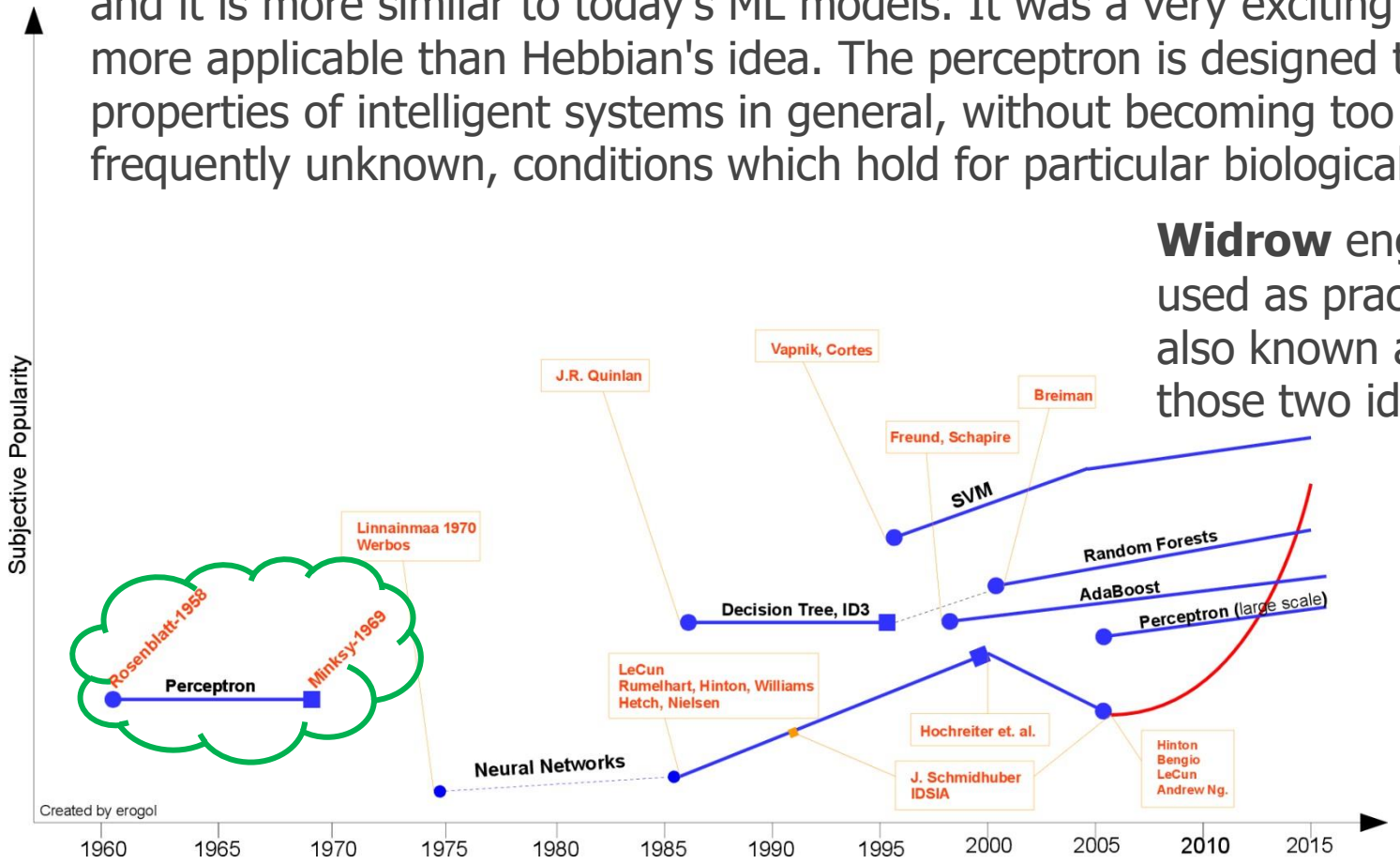
Arthur Samuel

In 1952 , **Arthur Samuel** at IBM, developed a program playing **Checkers** . The program was able to observe positions and learn a implicit model that gives better moves for the latter cases. Samuel played so many games with the program and observed that the program was able to play better in the course of time.

History of Machine Learning (From Eren Golge's Blog)

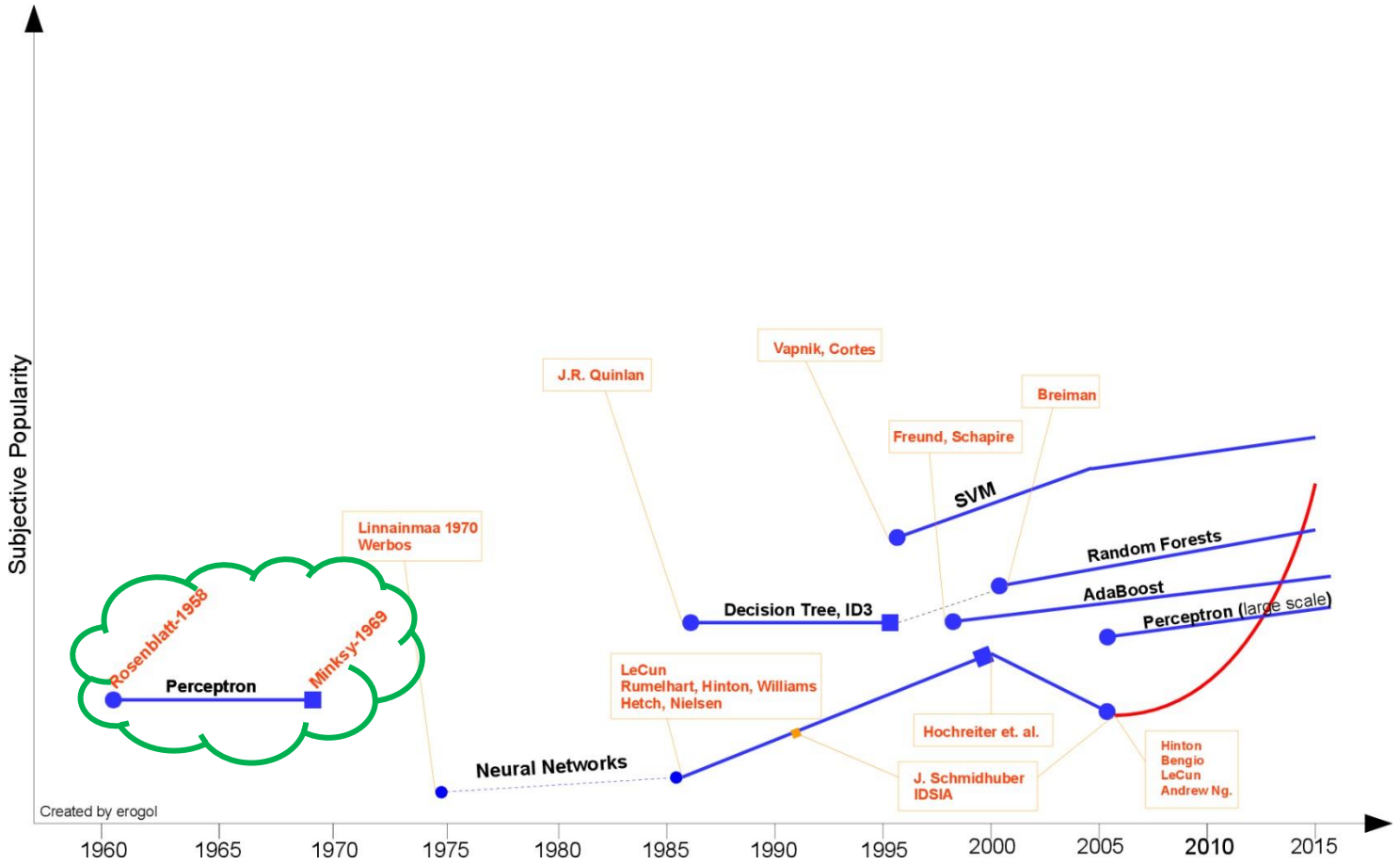
In 1957 , **Rosenblatt's Perceptron** was the second model proposed again with neuroscientific background and it is more similar to today's ML models. It was a very exciting discovery at the time and it was practically more applicable than Hebbian's idea. The perceptron is designed to illustrate some of the fundamental properties of intelligent systems in general, without becoming too deeply enmeshed in the special, and frequently unknown, conditions which hold for particular biological organisms.

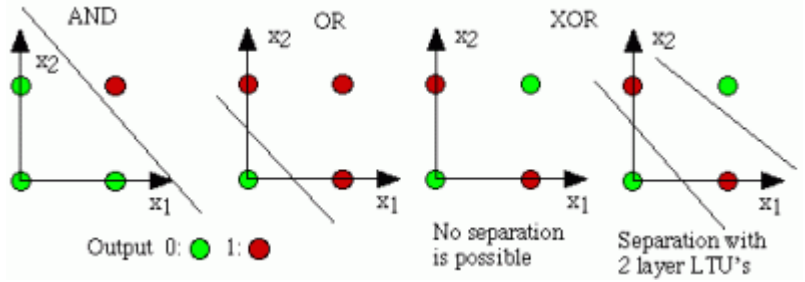
Widrow engraved **Delta Learning rule** that is then used as practical procedure for Perceptron training. It is also known as **Least Square** problem. Combination of those two ideas creates a good linear classifier.



F. Rosenblatt

History of Machine Learning (From Eren Golge's Blog)

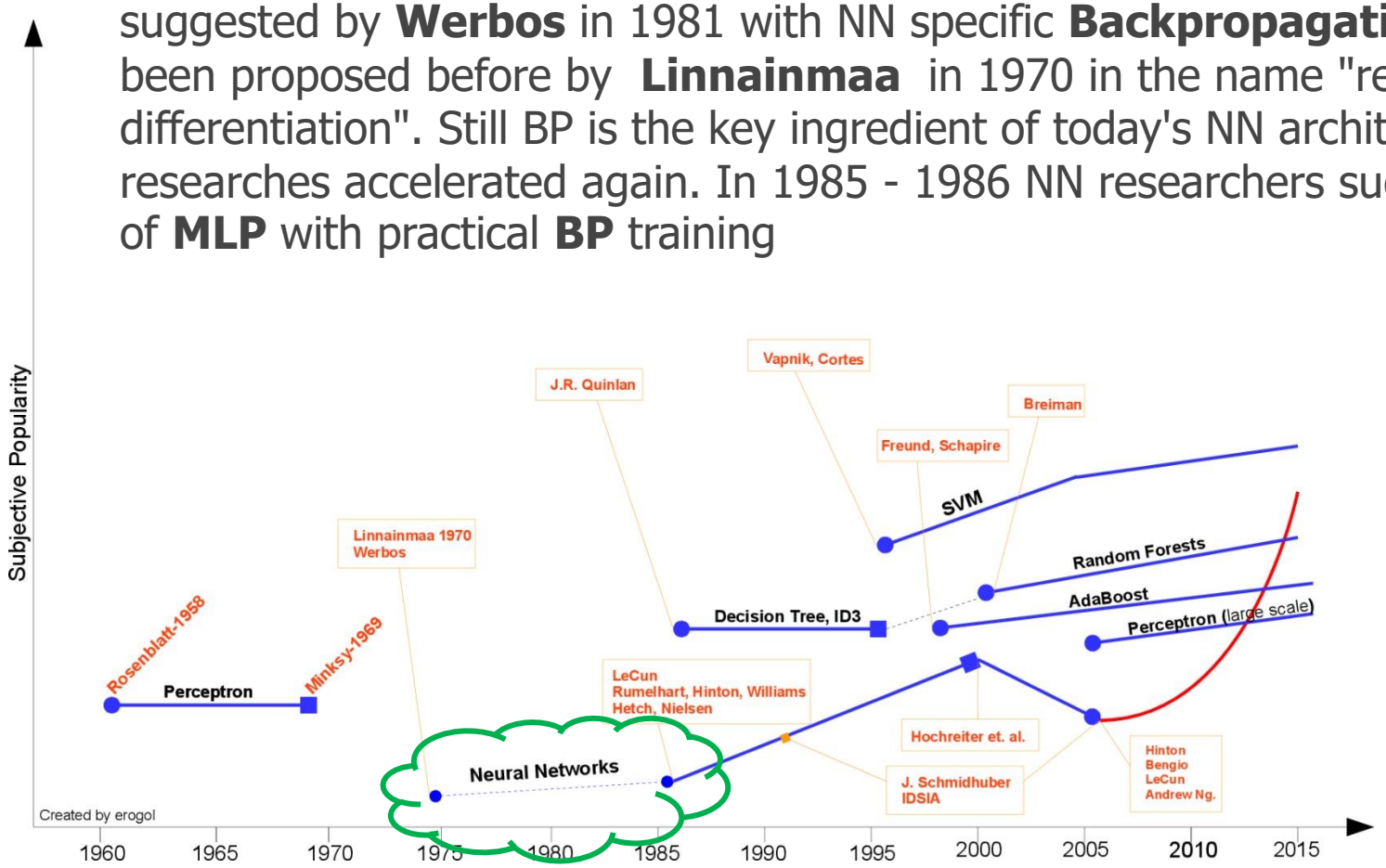


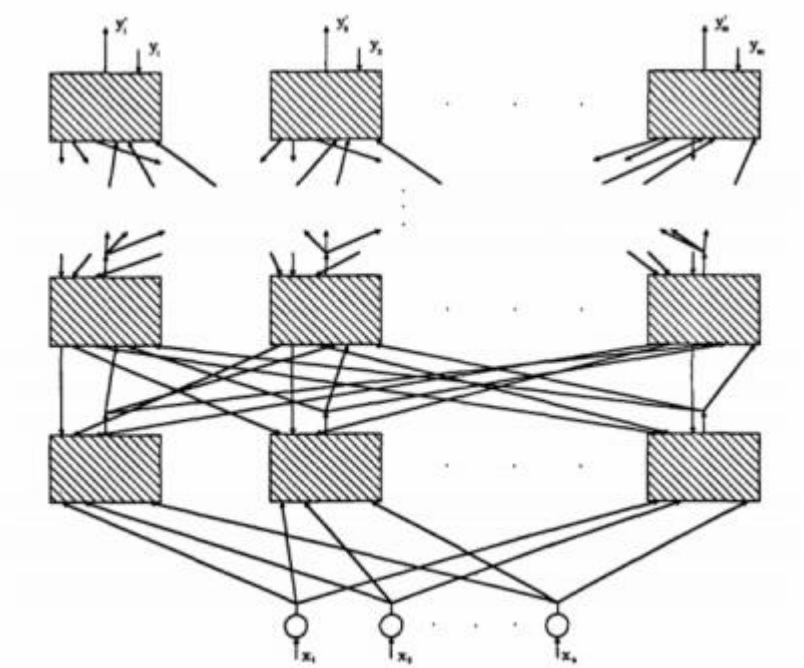


However, Perceptron's excitement was hinged by **Minsky** in 1969 . He proposed the famous **XOR** problem and the inability of Perceptrons in such linearly inseparable data distributions. It was the Minsky's tackle to NN community. Thereafter, NN researches would be dormant up until 1980s

History of Machine Learning (From Eren Golge's Blog)

There had been not to much effort until the intuition of **Multi-Layer Perceptron (MLP)** was suggested by **Werbos** in 1981 with NN specific **Backpropagation(BP)** algorithm, albeit BP idea had been proposed before by **Linnainmaa** in 1970 in the name "reverse mode of automatic differentiation". Still BP is the key ingredient of today's NN architectures. With those new ideas, NN researches accelerated again. In 1985 - 1986 NN researchers successively presented the idea of **MLP** with practical **BP** training

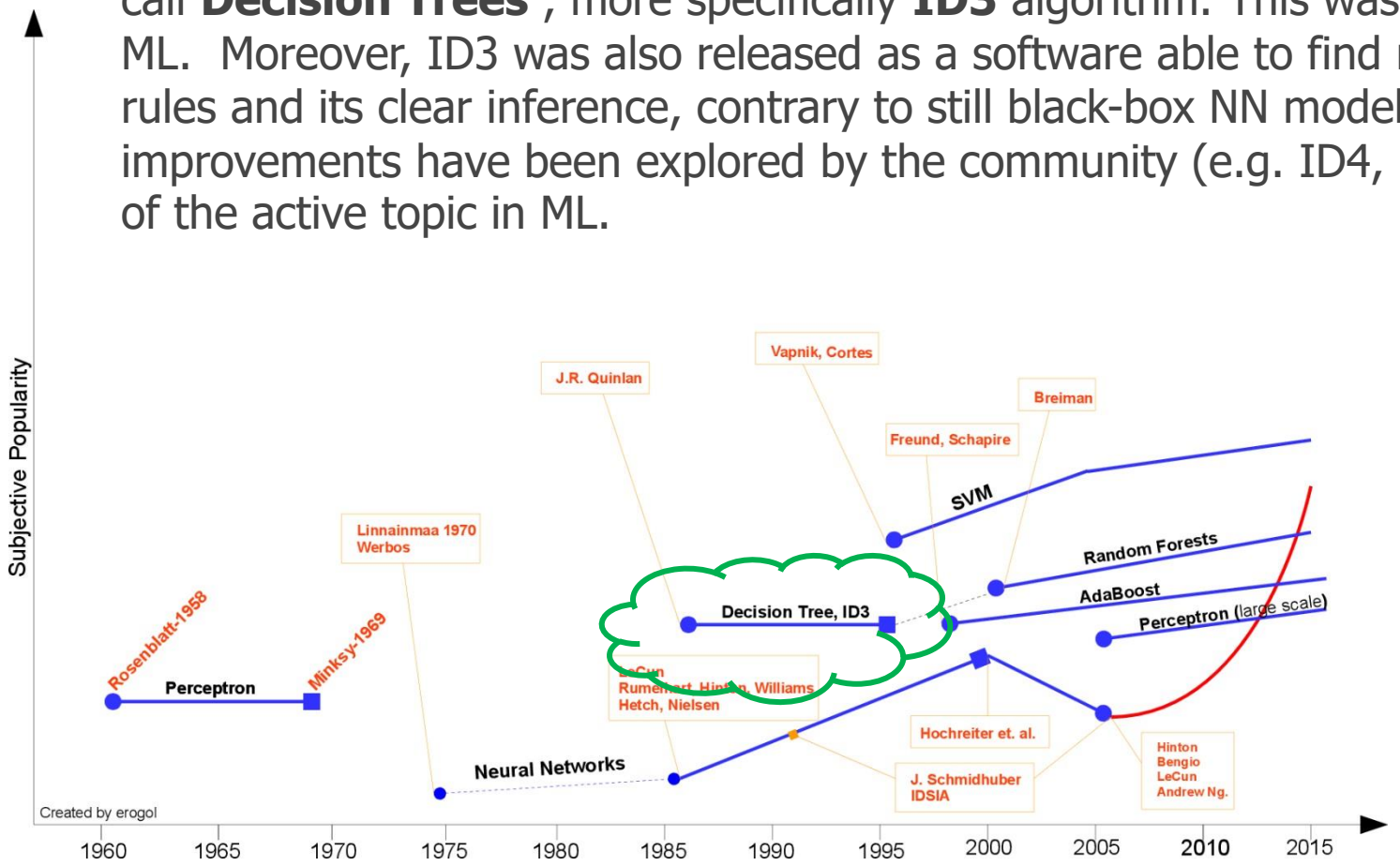




Hecht-Nielsen, Robert. "Theory of the backpropagation neural network." *Neural Networks, 1989. IJCNN., International Joint Conference on* . IEEE, 1989.

History of Machine Learning (From Eren Golge's Blog)

At the another spectrum, a very-well known ML algorithm was proposed by **J. R. Quinlan** in 1986 that we call **Decision Trees** , more specifically **ID3** algorithm. This was the spark point of the another mainstream ML. Moreover, ID3 was also released as a software able to find more real-life use case with its simplistic rules and its clear inference, contrary to still black-box NN models. 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 topic in ML.



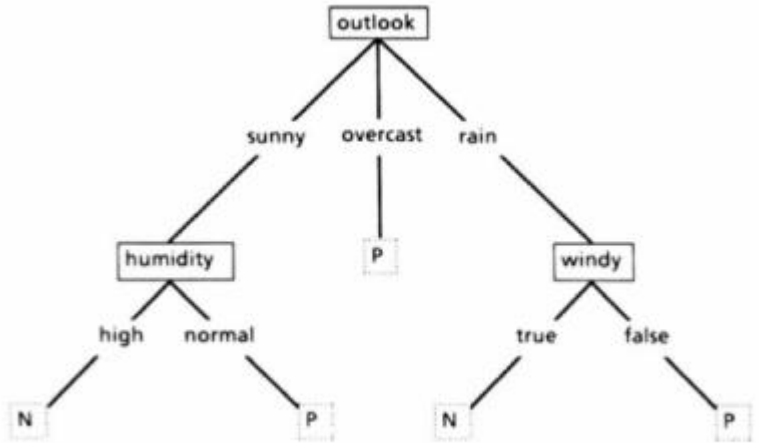
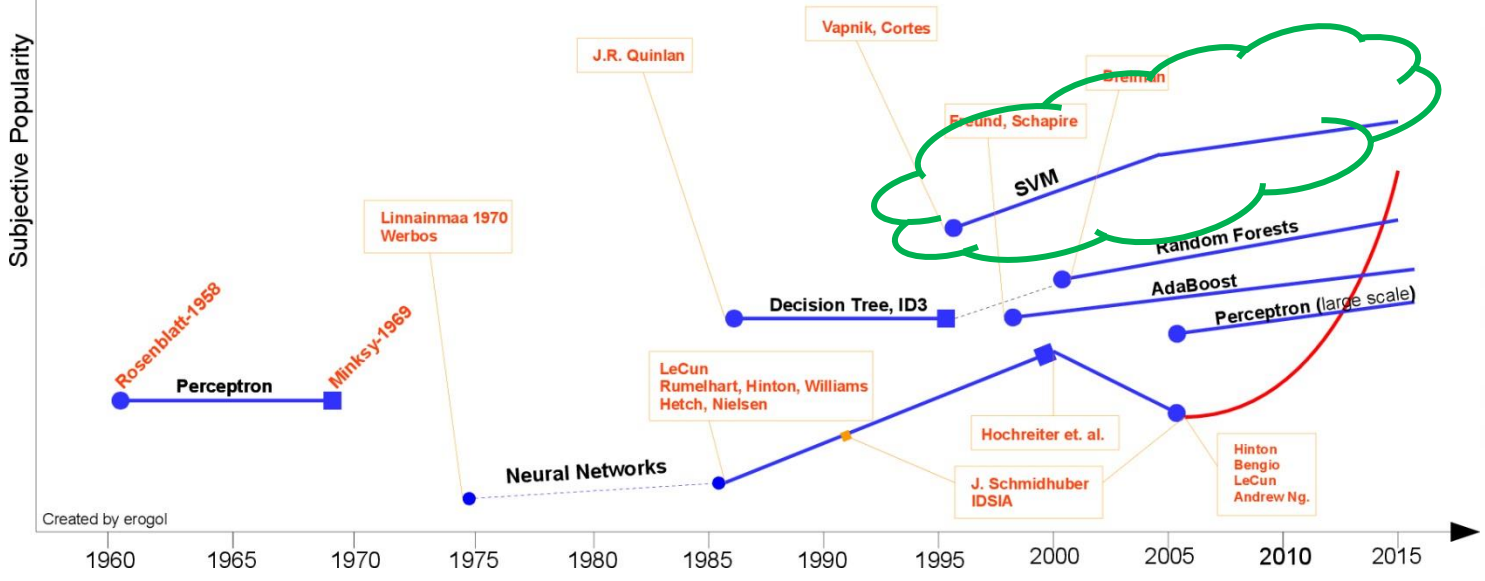


Figure 2. A simple decision tree

Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

History of Machine Learning (From Eren Golge's Blog)

One of the most important ML breakthrough was **Support Vector Machines** (Networks) (SVM), proposed by **Vapnik and Cortes** in **1995** with very strong theoretical standing and empirical results. That was the time separating the ML community into two crowds as NN or SVM advocates. However the competition between two community was not very easy for the NN side after **Kernelized** version of SVM by **near 2000s** .(I was not able to find the first paper about the topic), 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. Therefore, it could find large push from different disciplines causing very rapid theoretical and practical improvements.



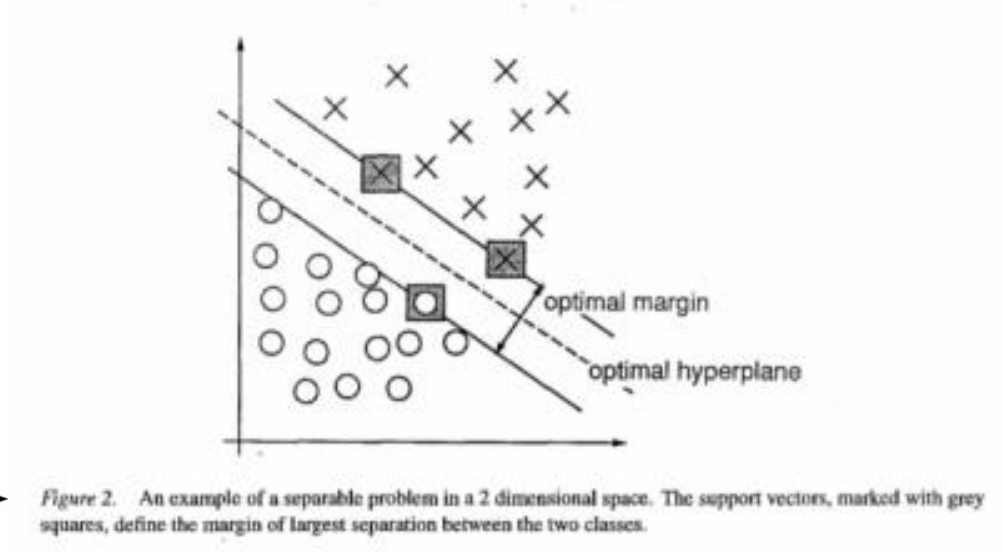
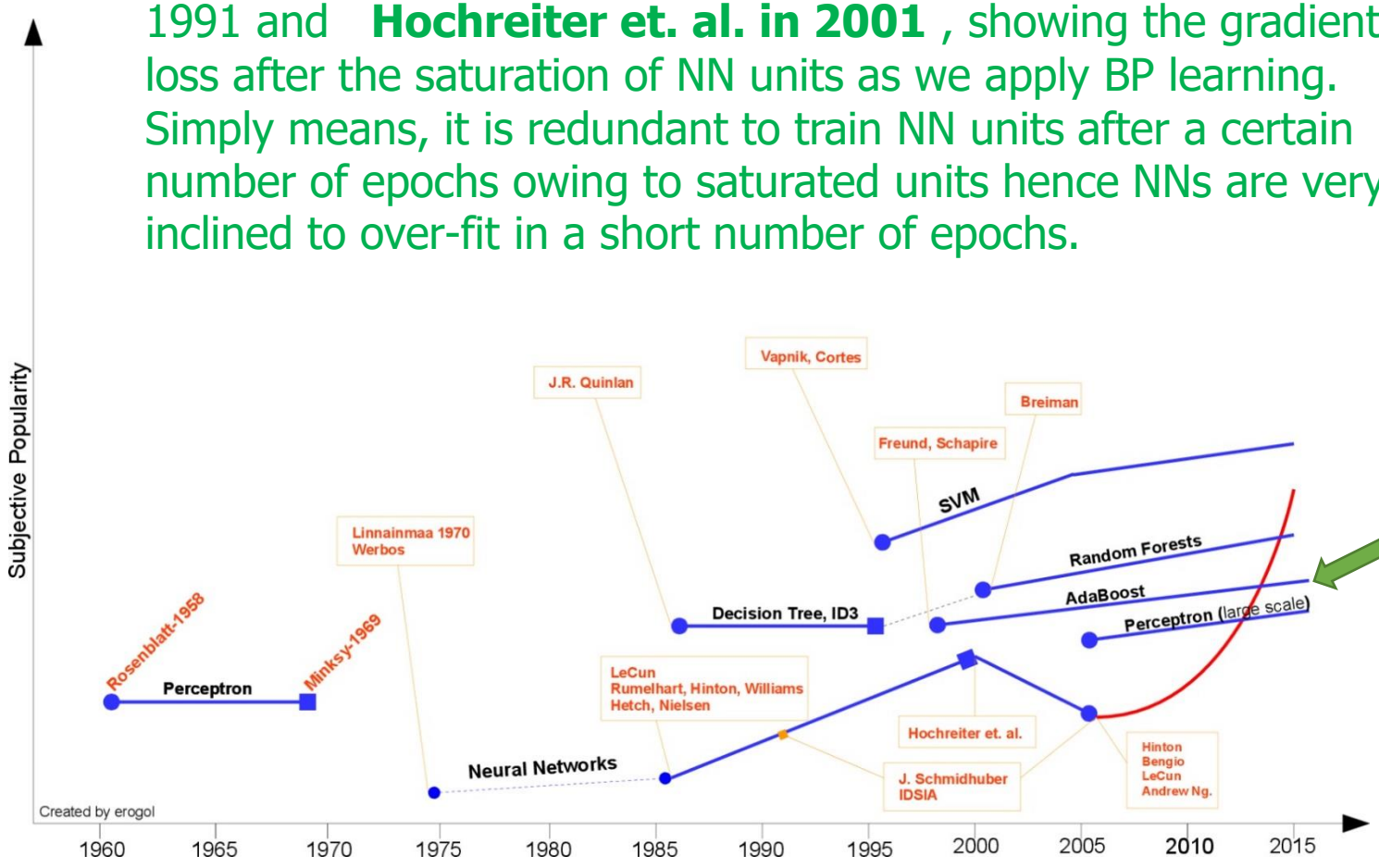


Figure 2. An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes.

Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.

History of Machine Learning (From Eren Golge's Blog)

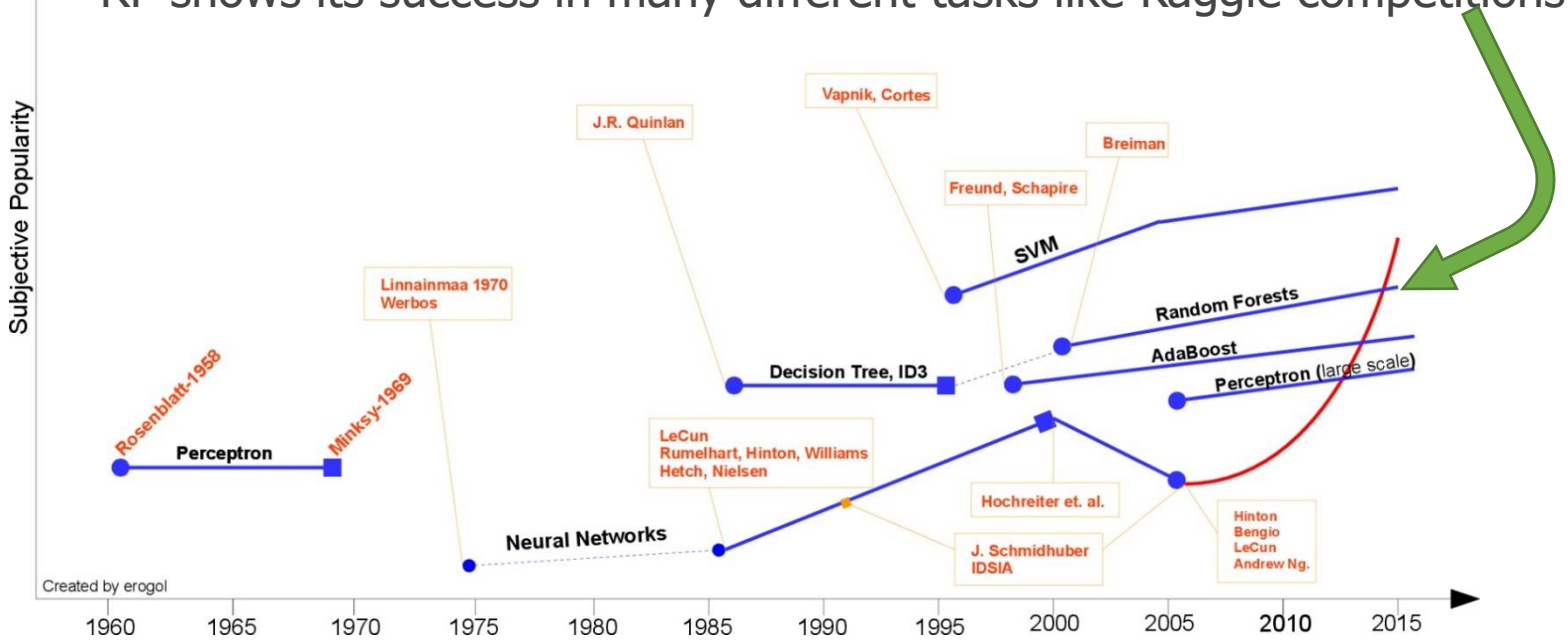
NN took another damage by the work of Hochreiter's thesis in 1991 and **Hochreiter et. al. in 2001** , showing the gradient loss after the saturation of NN units as we apply BP learning. Simply means, it is redundant to train NN units after a certain number of epochs owing to saturated units hence NNs are very inclined to over-fit in a short number of epochs.



Little before, another solid ML model was proposed by **Freund and Schapire** in **1997** prescribed with boosted ensemble of weak classifiers called **Adaboost.** This work also gave the Godel Prize to the authors at the time. Adaboost trains weak set of classifiers that are easy to train, by giving more importance to hard instances. This model still the basis of many different tasks like face recognition and detection.

History of Machine Learning (From Eren Golge's Blog)

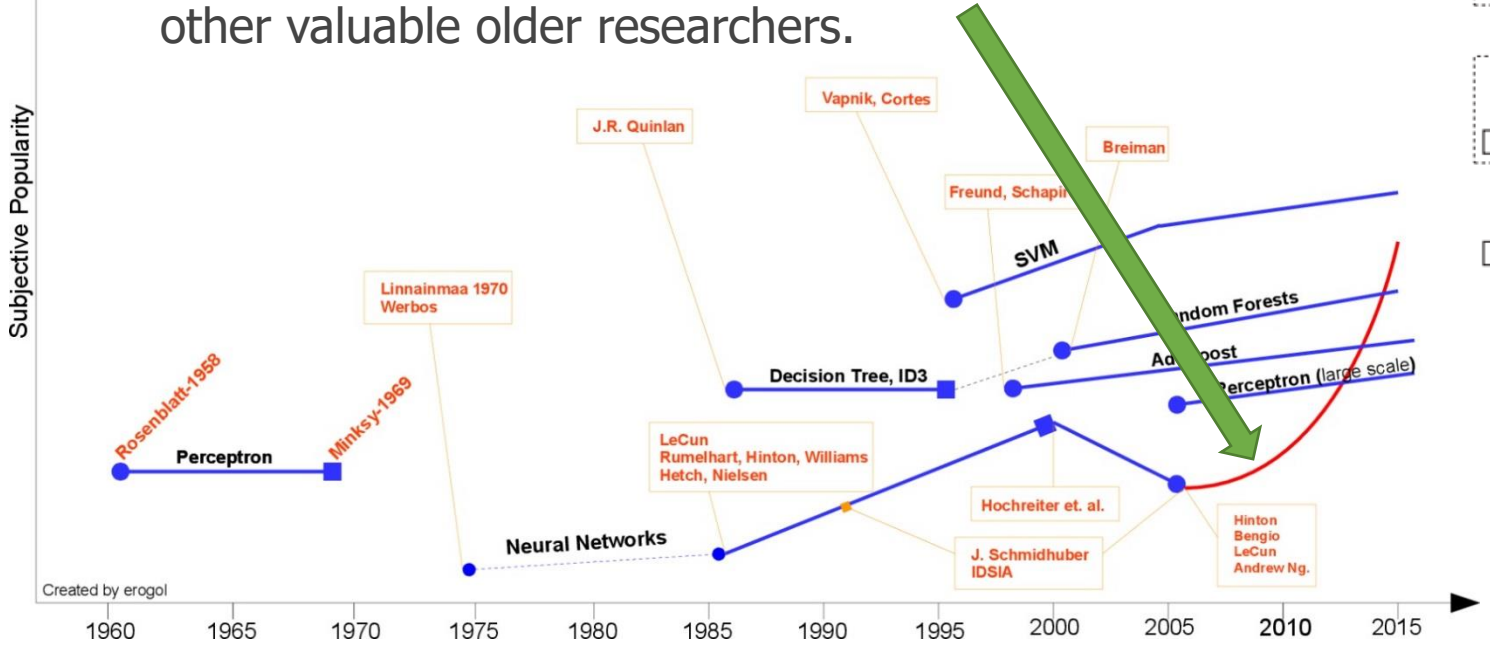
Another ensemble model explored by **Breiman** in **2001** that ensembles multiple decision trees where each of them is curated by a random subset of instances and each node is selected from a random subset of features. Owing to its nature, it is called **Random Forests(RF)** . RF has also theoretical and empirical proofs of endurance against over-fitting. Even AdaBoost shows weakness to over-fitting and outlier instances in the data, RF is more robust model against these caveats.(For more detail about RF, refer tomy old post.). RF shows its success in many different tasks like Kaggle competitions as well.

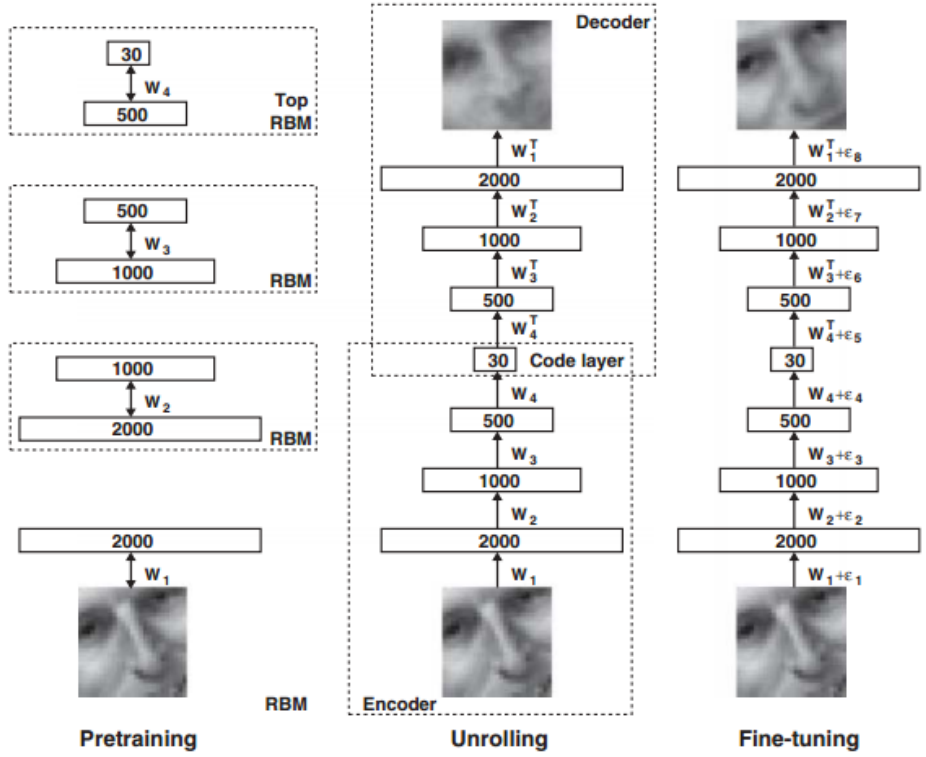


Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large.

History of Machine Learning (From Eren Golge's Blog)

As we come closer today, a new era of NN called **Deep Learning** has been commerced. This phrase simply refers NN models with many wide successive layers. The 3rd rise of NN has begun roughly in **2005** with the conjunction of many different discoveries from past and present by recent mavens Hinton, LeCun, Bengio, Andrew Ng and other valuable older researchers.

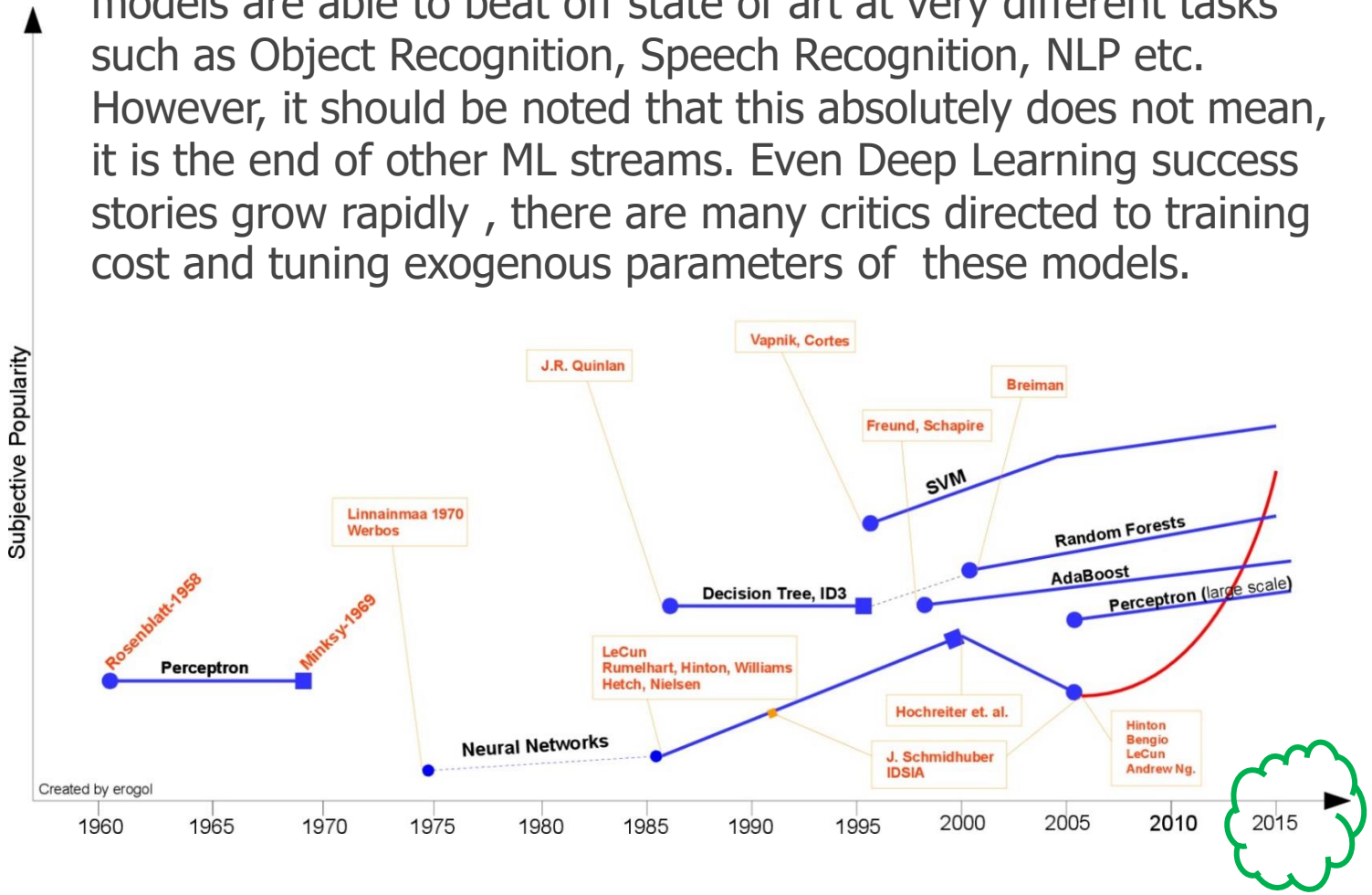




Hinton, G. E. and Salakhutdinov, R. R
Reducing the dimensionality of data with neural networks.
Science, Vol. 313. no. 5786, pp. 504 - 507, 28 July 2006..

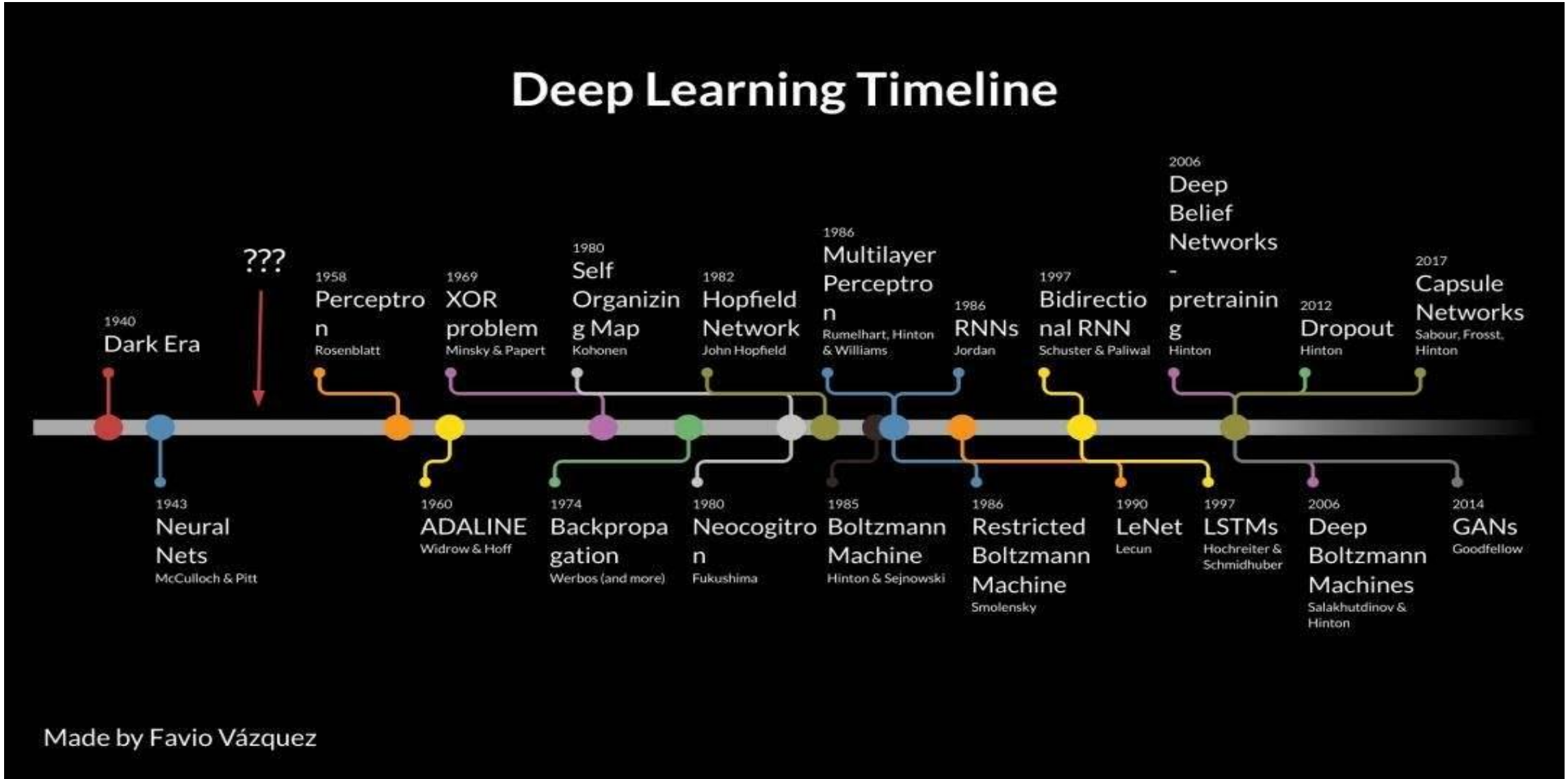
History of Machine Learning

With the combination of all those ideas and non-listed ones, NN models are able to beat off state of art at very different tasks such as Object Recognition, Speech Recognition, NLP etc. However, it should be noted that this absolutely does not mean, it is the end of other ML streams. Even Deep Learning success stories grow rapidly , there are many critics directed to training cost and tuning exogenous parameters of these models.

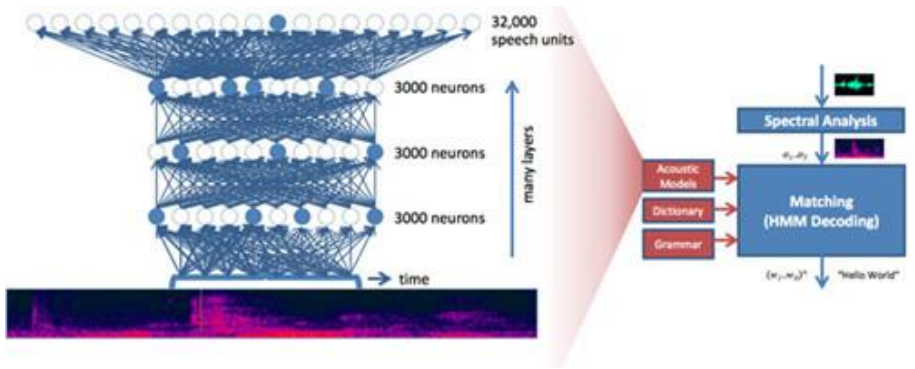


After the growth of WWW and Social Media, a new term, **BigData** emerged and affected ML research wildly. Because of the large problems arising from BigData , many strong ML algorithms are useless for reasonable systems (not for giant Tech Companies of course). Hence, research people come up with a new set of simple models that are dubbed **Bandit Algorithms** (formally predicated with **Online Learning**) that makes learning easier and adaptable for large scale problems.

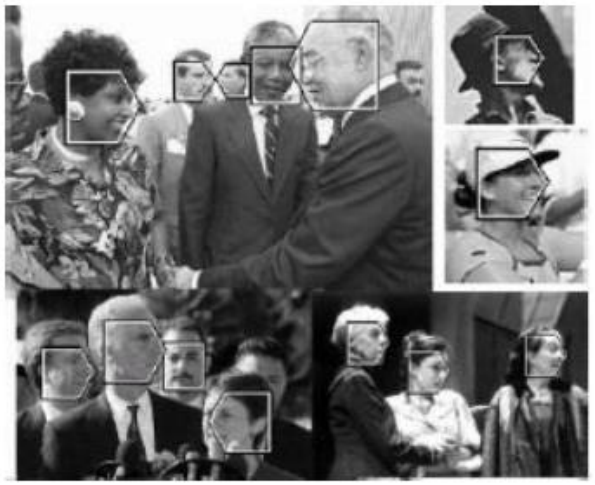
History of Deep Learning

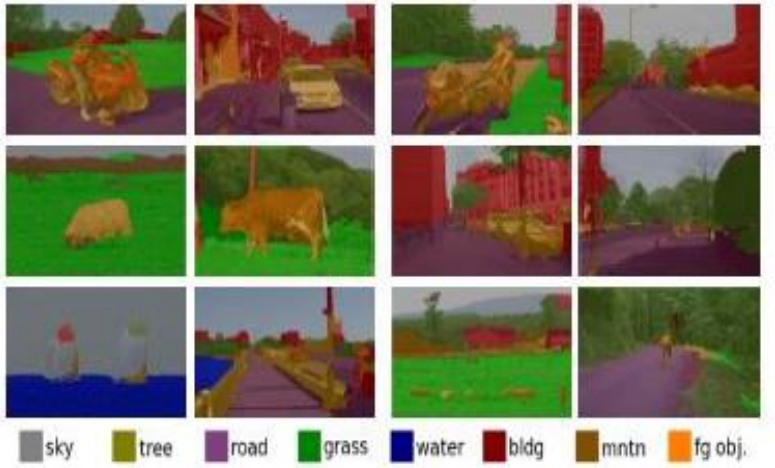


Applications: Speech Recognition

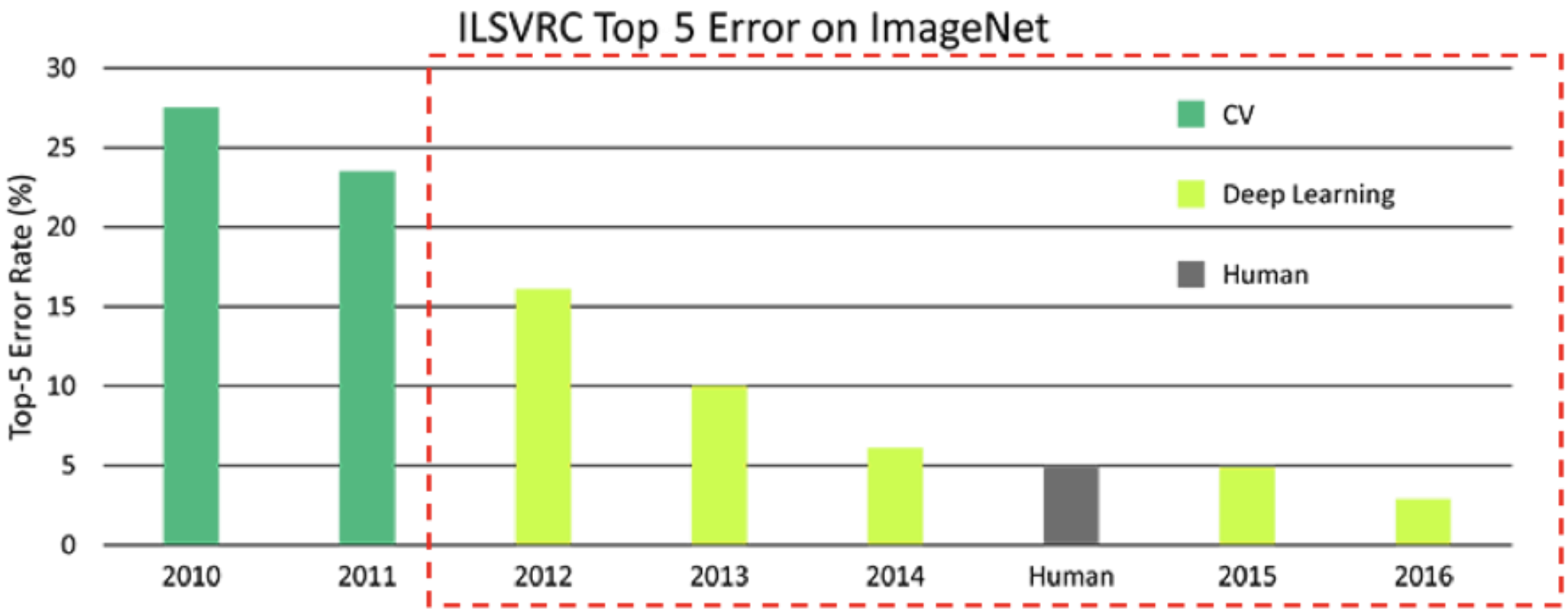


Applications: Computer Vision





Success of Deep Learning: Computer Vision



The introduction of Deep Learning techniques drove performance on image categorization from 30% error rates in 2010, down to <2% in 2017

CV: Deefake

Deepfake is a technique for human image synthesis based on artificial intelligence.

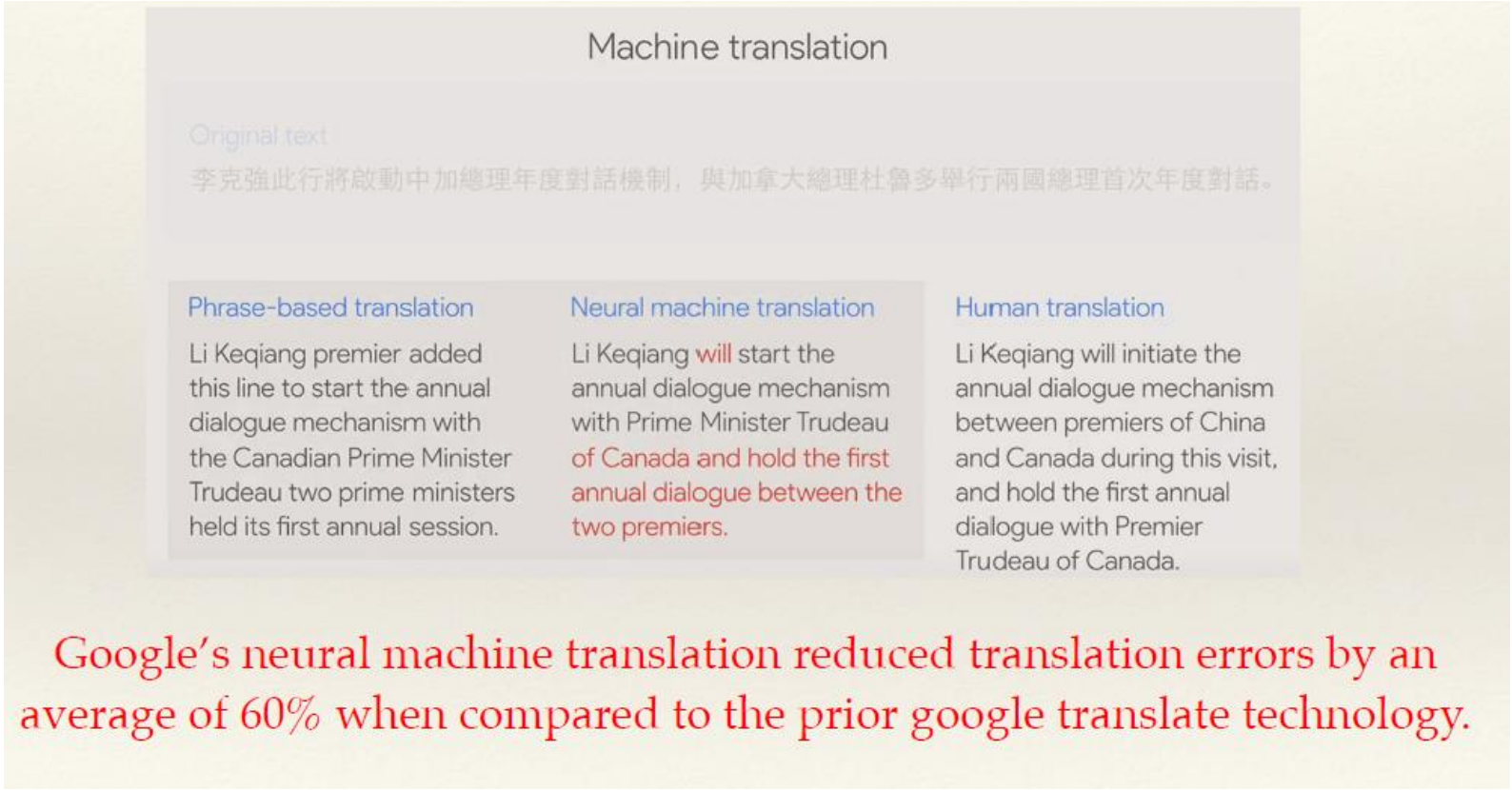
Combine and superimpose existing images and videos onto source images or videos using a machine learning technique known as generative adversarial network.

Applications: Web Search

Applications: Recommender System

Applications: Social Computing

Applications: Natural Language Processing



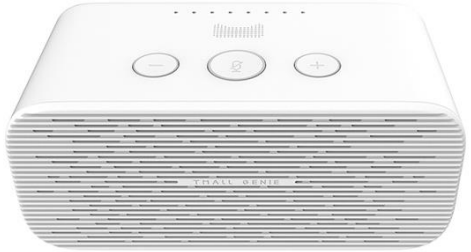
Applications: Natural Language Processing

Siri

Alexa

Cortana

Xiaomi

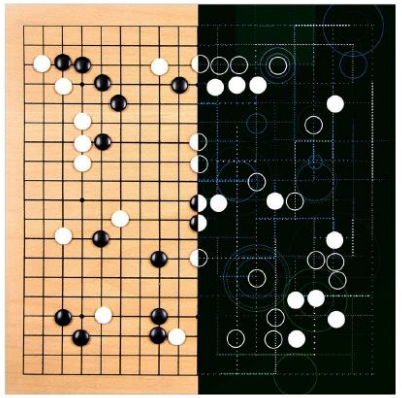


Tmall

Google Assistant

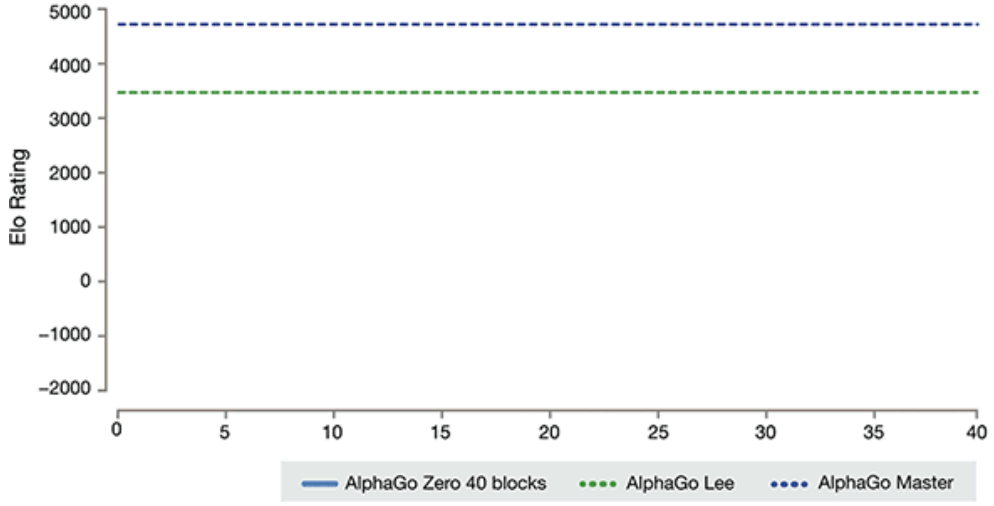
Applications: Game

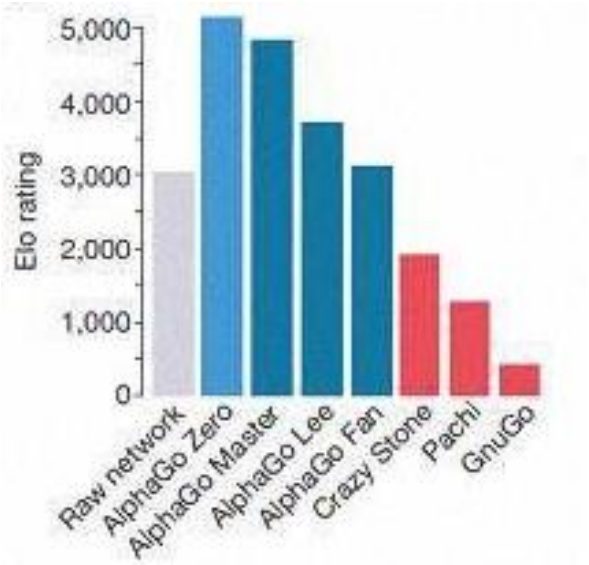
AlphaGo



TimeLine of AlphaGo

- 2016.1.27 AlphaGo Fan vs fan hui : 5-0
- 2016.3.9 – 2016.3.15 AlphaGo Lee vs li shishi : 4-1
- 2017.5.23 – 2017.5.27 AlphaGo Master vs ke jie : 3-0
- Now AlphaGo Zero





Applications: Game

- 2010: the first AI competition.
- 2012: The use of persistent file storage enables the Bot to learn throughout the competition. Bots have emerged that can predict results.
- 2016: studying starcraft AI is no longer about beating human players, but about how it makes decisions in more complex environments.
- 2019: AlphaStar vs TLO, Mana: 10-1

More Applications

Human Society Affairs


5 min Break