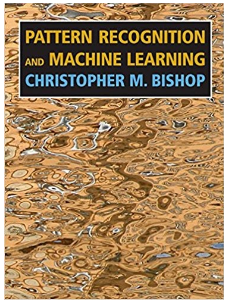
Lecture 1: Machine Learning

Andreas Wichert
Department of Computer Science and Engineering
Técnico Lisboa

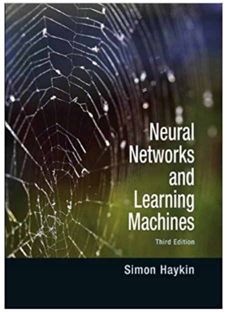
Corpo docente – Alameda/Tagus

- Andreas (Andrzej) Wichert
    - andreas.wichert@tecnico.ulisboa.pt
    - tel: 214233231
    - room: N2 5-7 (Taguspark)
    - http://web.tecnico.ulisboa.pt/andreas.wichert/

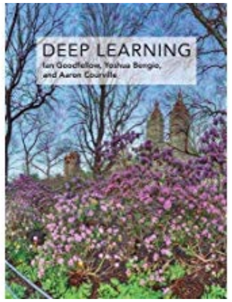
Main Literature



- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
  - https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book



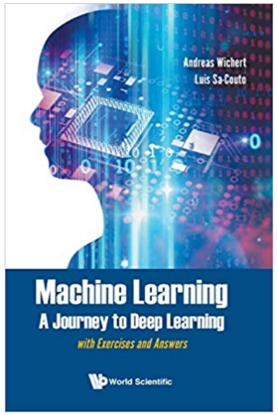
- Simon O. Haykin, Neural Networks and Learning Machine, (3rd Edition), Pearson 2008



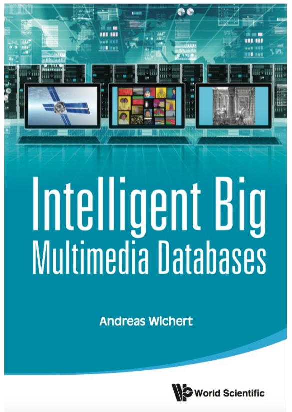
Deep Learning, I. Goodfellow, Y. Bengio, A. Courville
MIT Press 2016

- https://www.deeplearningbook.org

Main Literature

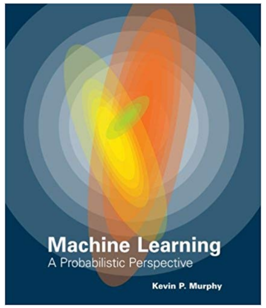


- Machine Learning - A Journey to Deep Learning, A. Wichert, Luis Sa-Couto, World Scientific, 2021

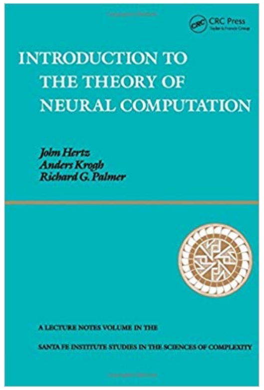


- Intelligent Big Multimedia Databases, A. Wichert, World Scientific, 2015
  - *Preprocessing, Feature Extraction like DFT, Wavelets, will be not covered in the lecture….*

Additional Literature

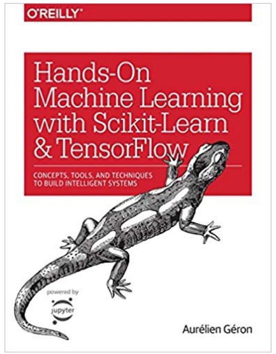


- Machine Learning: A Probabilistic Perspective, K. Murphy, MIT Press 2012



- Introduction To The Theory Of Neural Computation (Santa Fe Institute Series Book 1), John A. Hertz, Anders S. Krogh, Richard G. Palmer, Addison-Wesley Pub. Co, Redwood City, CA; 1 edition (January 1, 1991)
  - *I find this book to be one of the best written mathematical guides for Neural Networks. See Perceptron, Backpropagation…*

Literature Software



- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 1st Edition,   Aurélien Géron , O'Reilly Media; 1 edition (April 9, 2017)
  - https://github.com/amitanalyste/aurelienGeron

- https://scikit-learn.org/stable/index.html

- http://www.numpy.org

I) Outline:

Introduction: What is Machine Learning?

1. Introduction
2. Decision Trees

Mathematical Tools:

3. Probability theory & Information *(Naive Bayes)*
4. Linear Algebra & Optimization *(Simple NN)*

Road to deep learning: Error Minimization (Loss), Regularization, Optimization by Gradient descent

5. Linear Regression & Bayesian Linear Regression
6. Perceptron & Logistic Regression
7. Multilayer Perceptrons

II) Outline

Why do the neural works work :

8. Learning theory, Bias-Variance
9. K-Means, EM-Clustering
10. Kernel Methods & RBF
11. Support Vector Machines

How to use the models:

12. Model Selection

III) Outline

Deep Learning ***solves*** *the problem of* ***high dimensionality*** *which is related to the* ***training database size****!*

13. Deep Learning
14. Convolutional Neural Networks
15. Recurrent Neural Networks

Dimension Reduction:

16. PCA, ICA
17. Autoencoders

IV) Outline

Alternative Road to Machine Learning (Classical Approach):

18. Feature Extraction *(FFT, SFT, Edge Detection)*
19. k Nearest Neighbour & Locally Weighted Regression
20. Ensemble Methods

Probabilistic and Stochastic Approach:

21. Bayesian Networks
22. Stochastic Methods

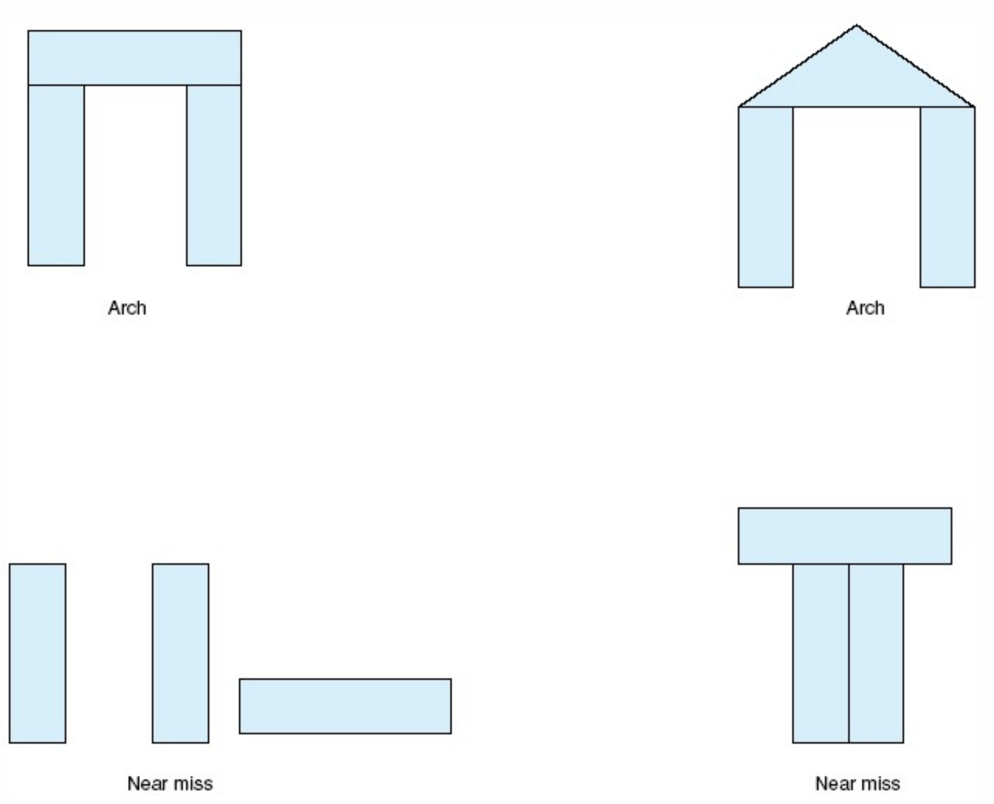
What is machine Learning?

- Parallels between “animals” and machine learning
- Many techniques derived from efforts of psychologist / biologists to make more sense “animal” learning through computational models

Machine Learning

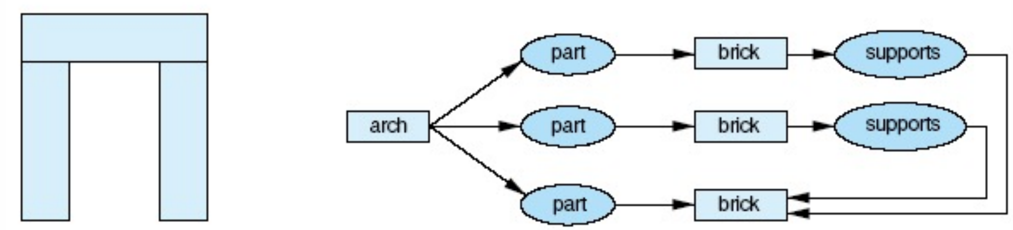
- Statistical Machine Learning
  - Linear Regression
  - Clustering, Self Organizing Maps (SOM)
  - Artificial Neural Networks, Kernel Machines
  - Bayesian Network
- *We will not cover….*
  - *Inductive Learning (ID3)*
  - *Knowledge Learning*
  - *Analogical Learning*
  - *SOAR: Model of Cognition and Learning*

An Example of Symbolical Learning (Patrick Winston-1975)

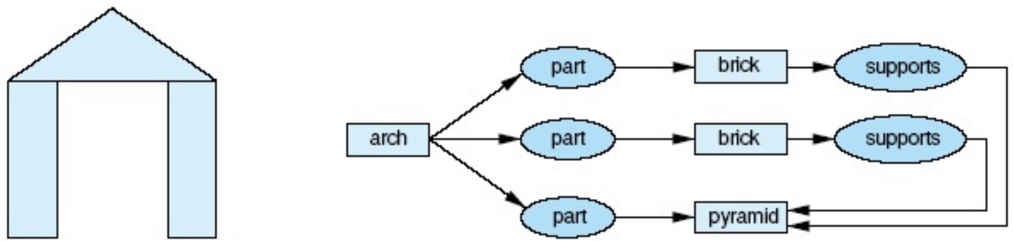


An Example (Patrick Winston-1975)

a. An example of an arch and its network description

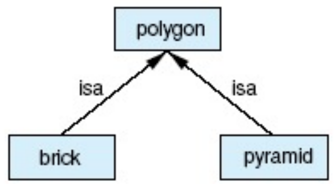


b. An example of another arch and its network description

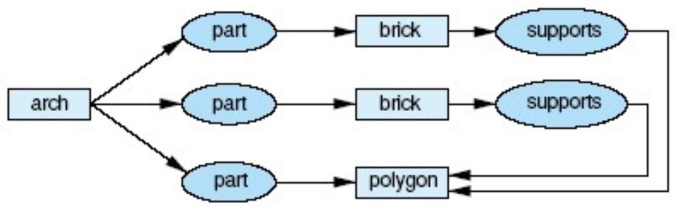


An Example (Patrick Winston-1975)

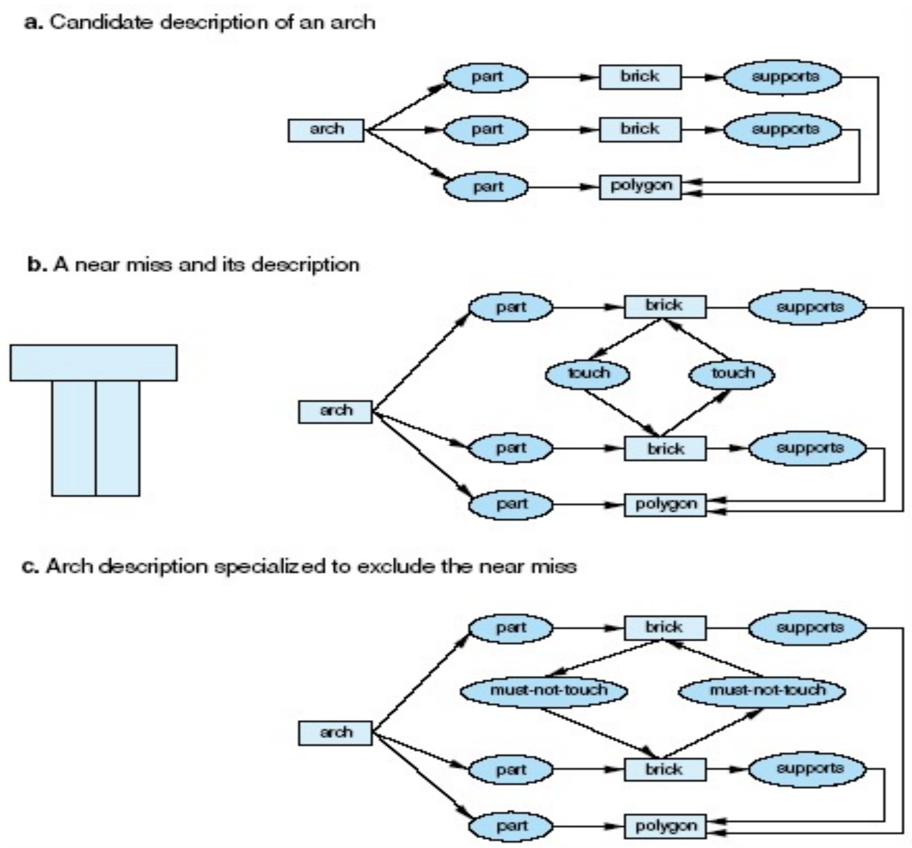
c. Given background knowledge that bricks and pyramids are both types of polygons



d. Generalization that includes both examples



An Example (Patrick Winston-1975)

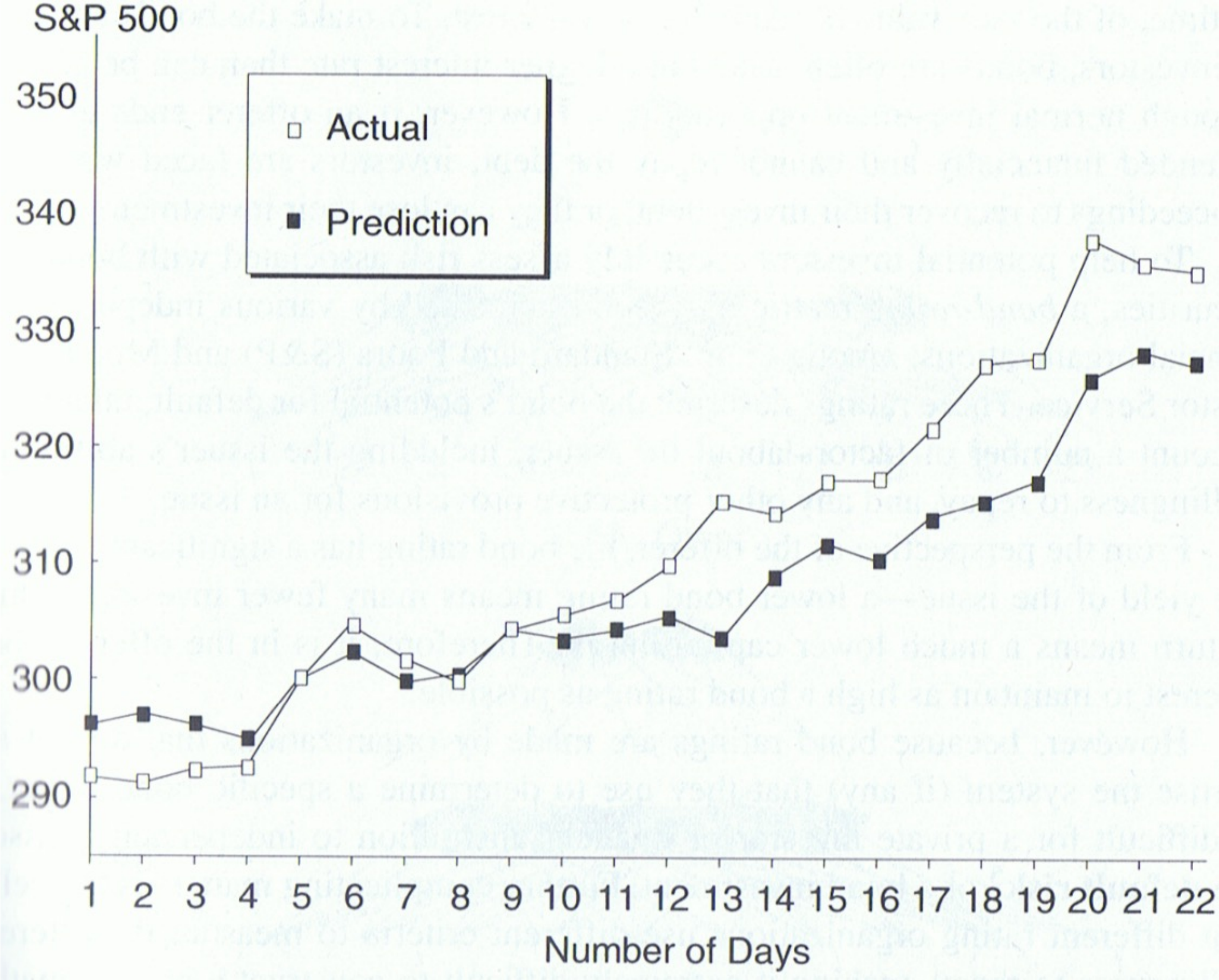


Statistical Machine Learning

- Changes in the system that perform tasks associated with AI
    - Recognition
    - Prediction
    - Planning
    - Diagnosis

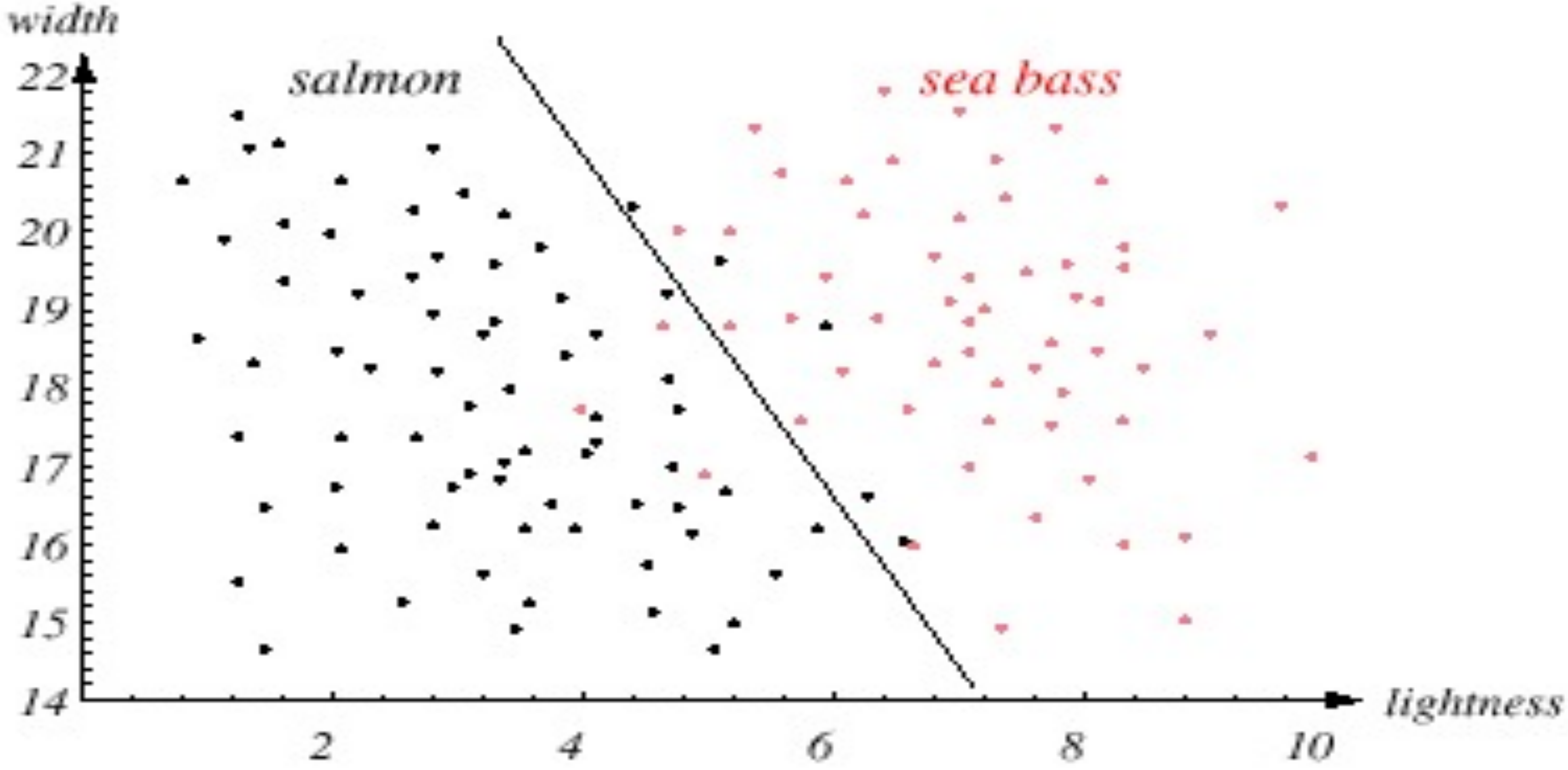
Learning Input output functions

- Supervised
  - With a teacher
- Unsupervised
  - Without a teacher
- Reinforcemet Learning
  - Actions within & responses from the environment
  - Absence of a designated teacher to give positive and negative examples



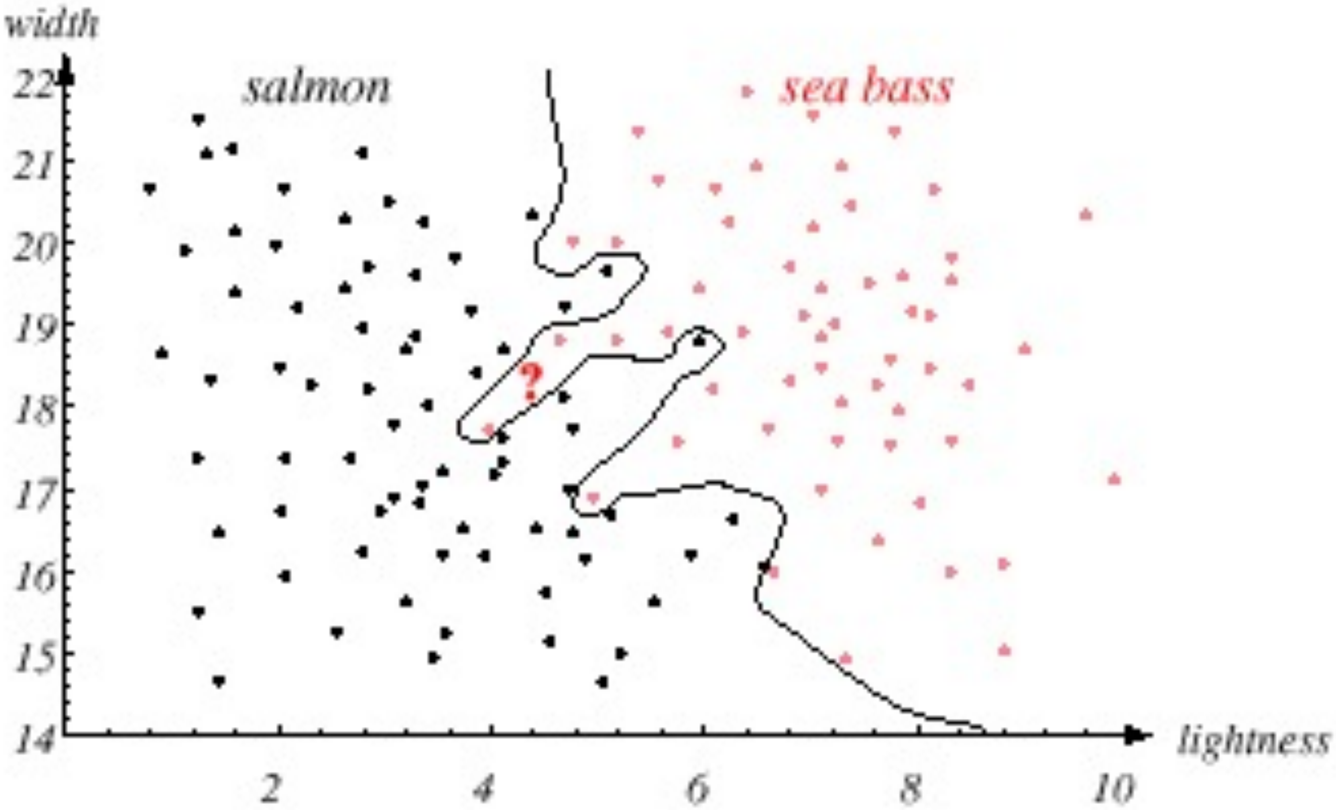
S&P 500
350
340
330
320
310
300
290
Actual
Prediction
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Number of Days


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"


width
salmon
sea bass
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness

- We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such “noisy features”

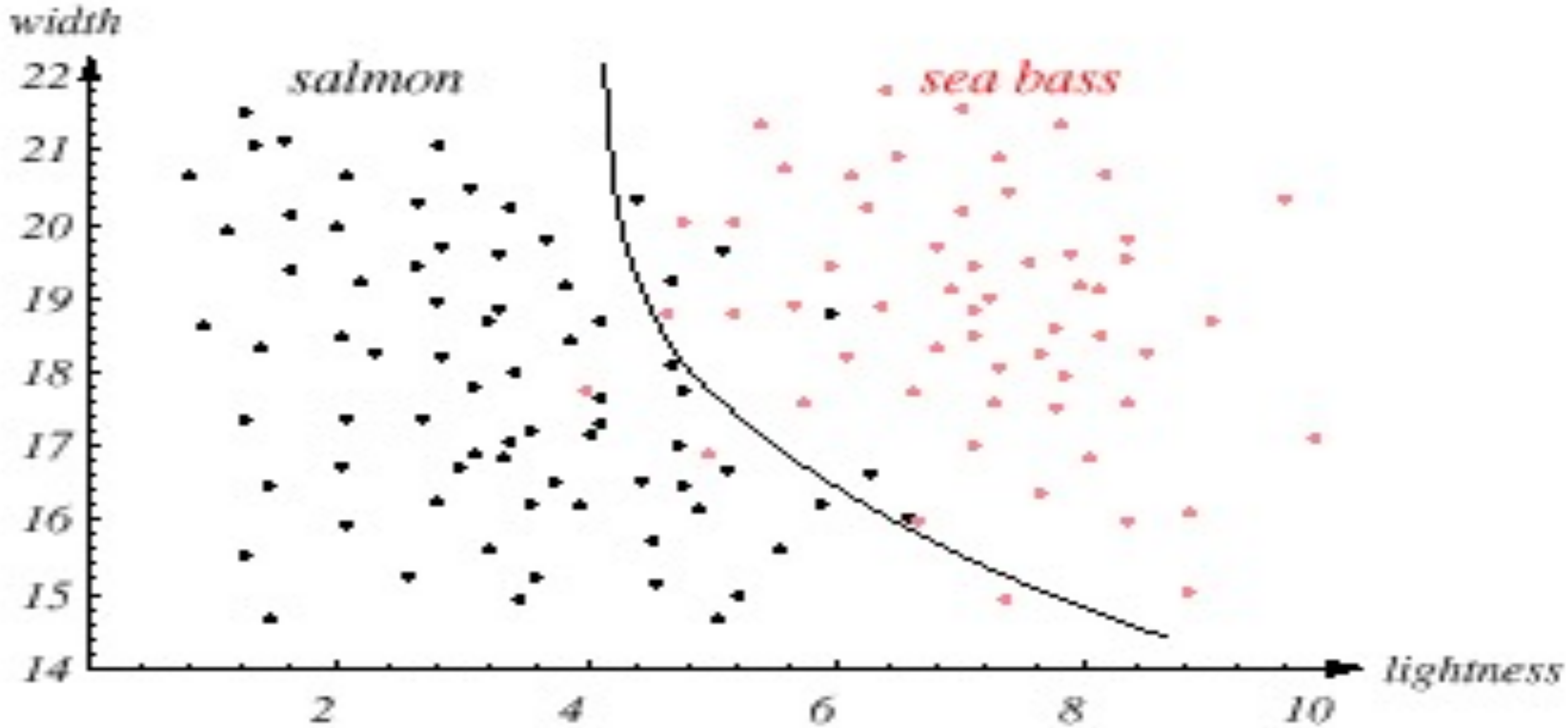
- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:



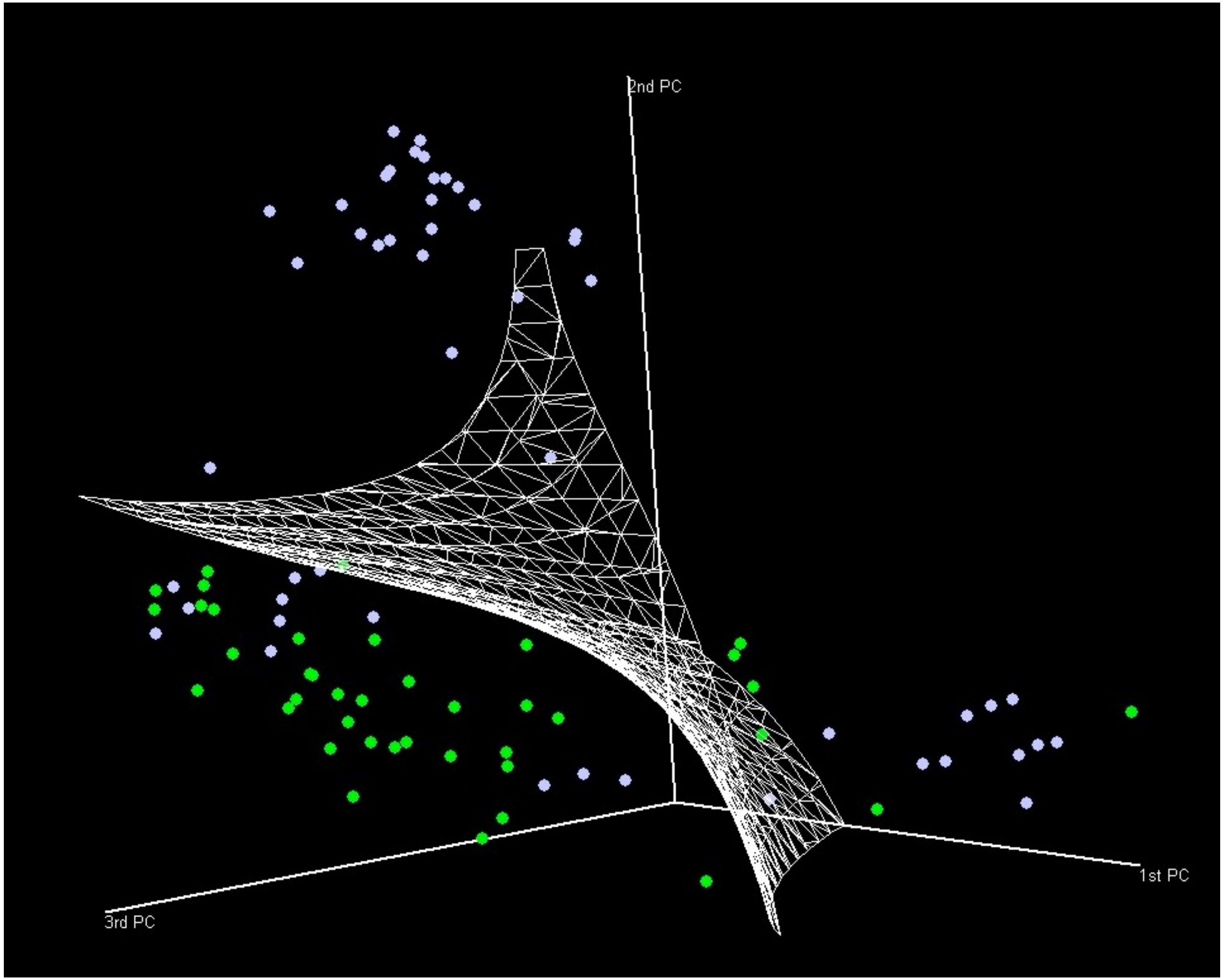
width
salmon
sea bass
?
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness

- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

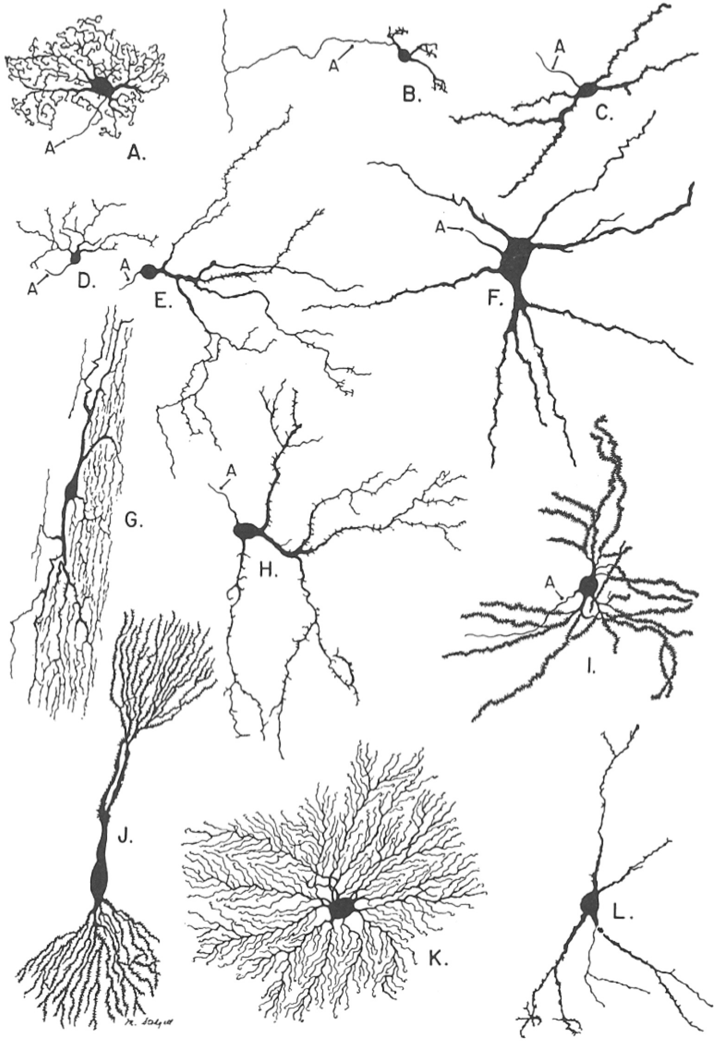
Issue of generalization!

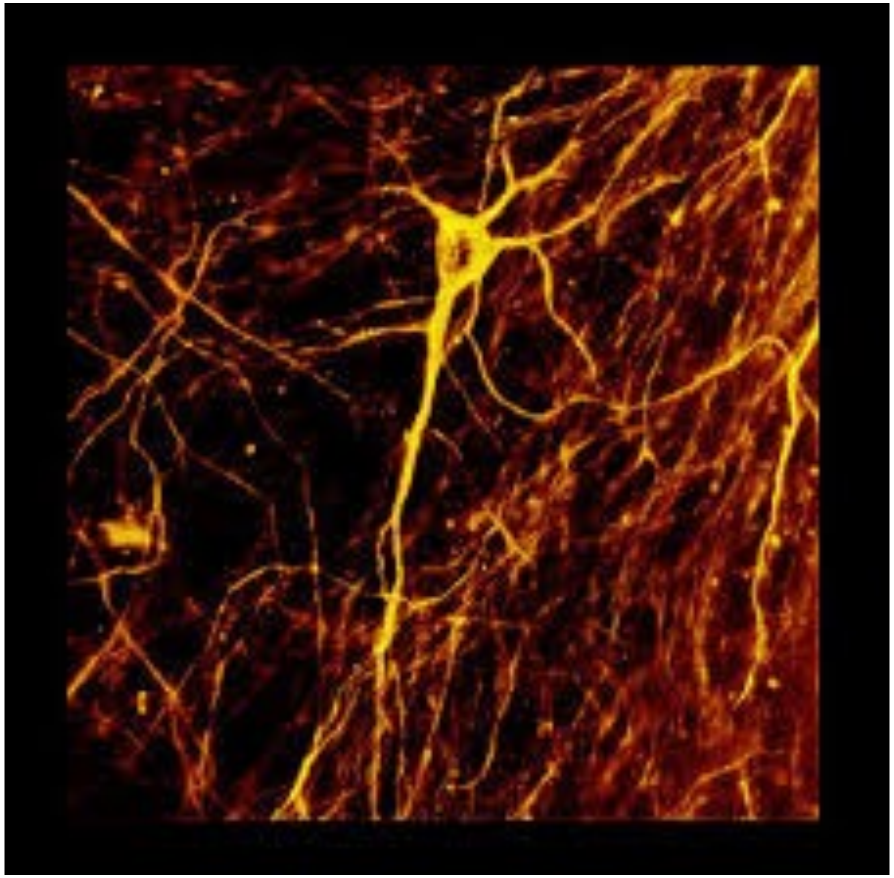


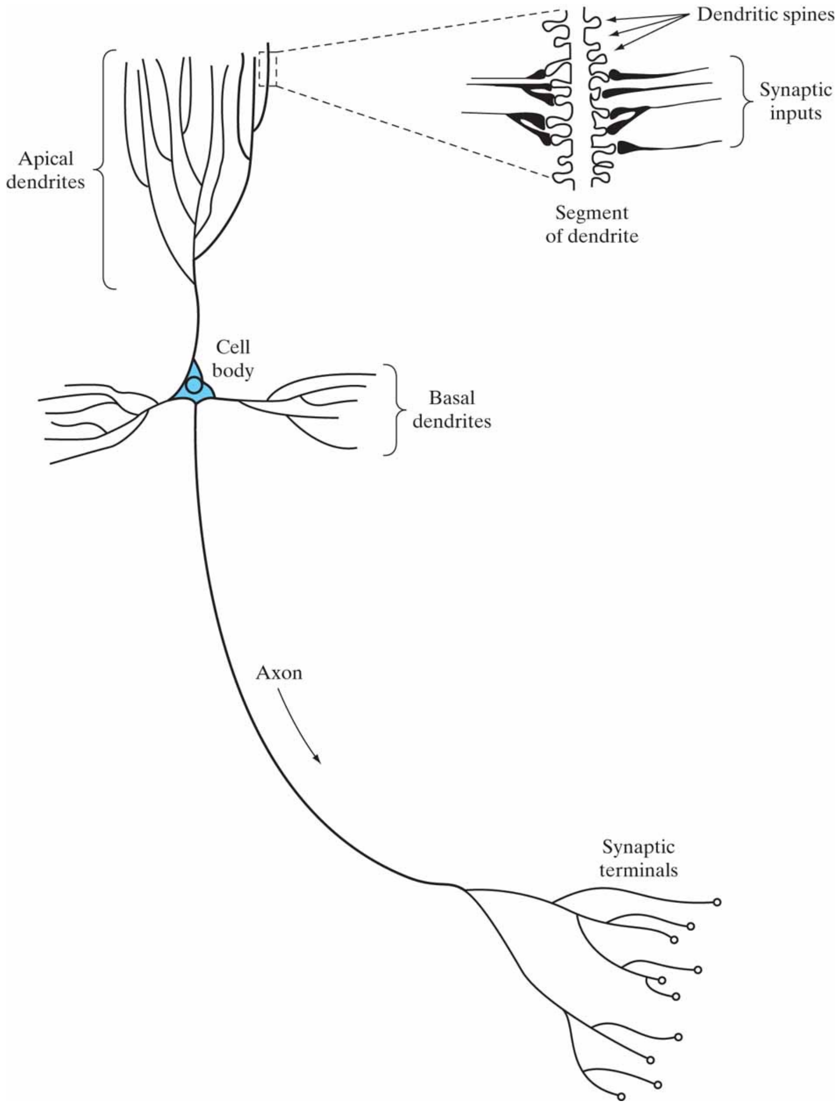
width
salmon
sea bass
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness


2nd PC
1st PC
3rd PC

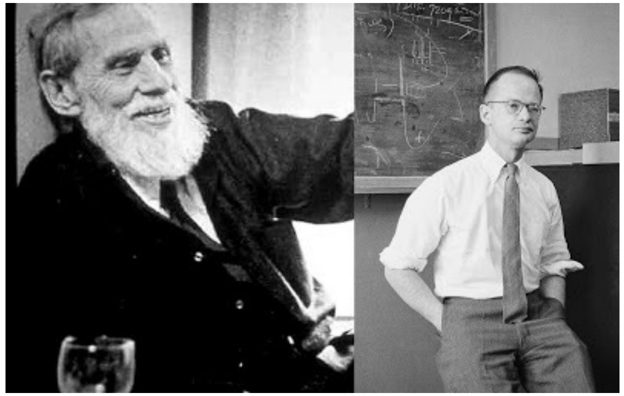
- 10^{40} Neurons
- 10^{4-5} connections per neuron



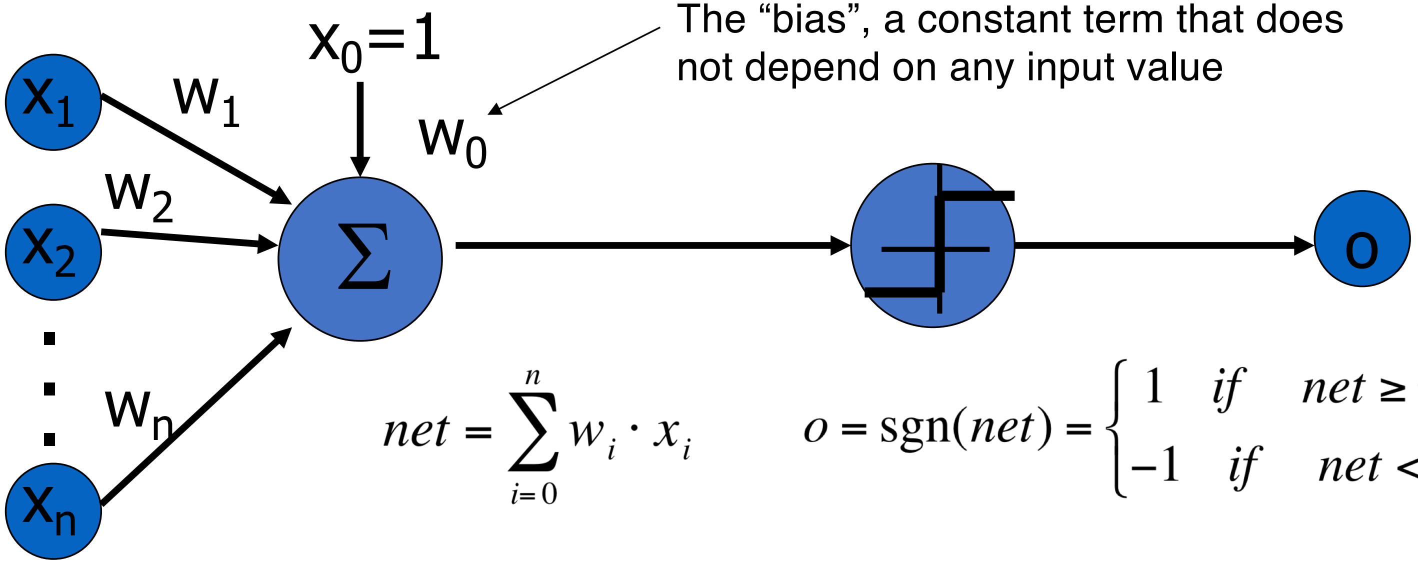



Dendritic spines
Synaptic inputs
Apical dendrites
Segment of dendrite
Cell body
Basal dendrites
Axon
Synaptic terminals

Perceptron (1957)



- Linear threshold unit (LTU)



$$net = \sum_{i=0}^{n} w_i \cdot x_i \qquad o = \mathrm{sgn}(net) = \begin{cases} 1 & \text{if} \quad net \geq 0 \\ -1 & \text{if} \quad net < 0 \end{cases}$$

McCulloch-Pitts model of a neuron (1943)

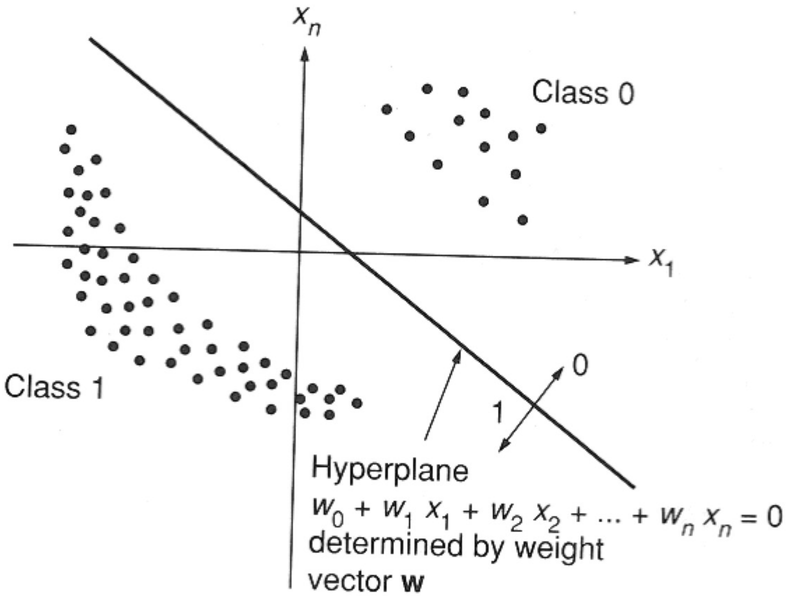
Linearly separable patterns

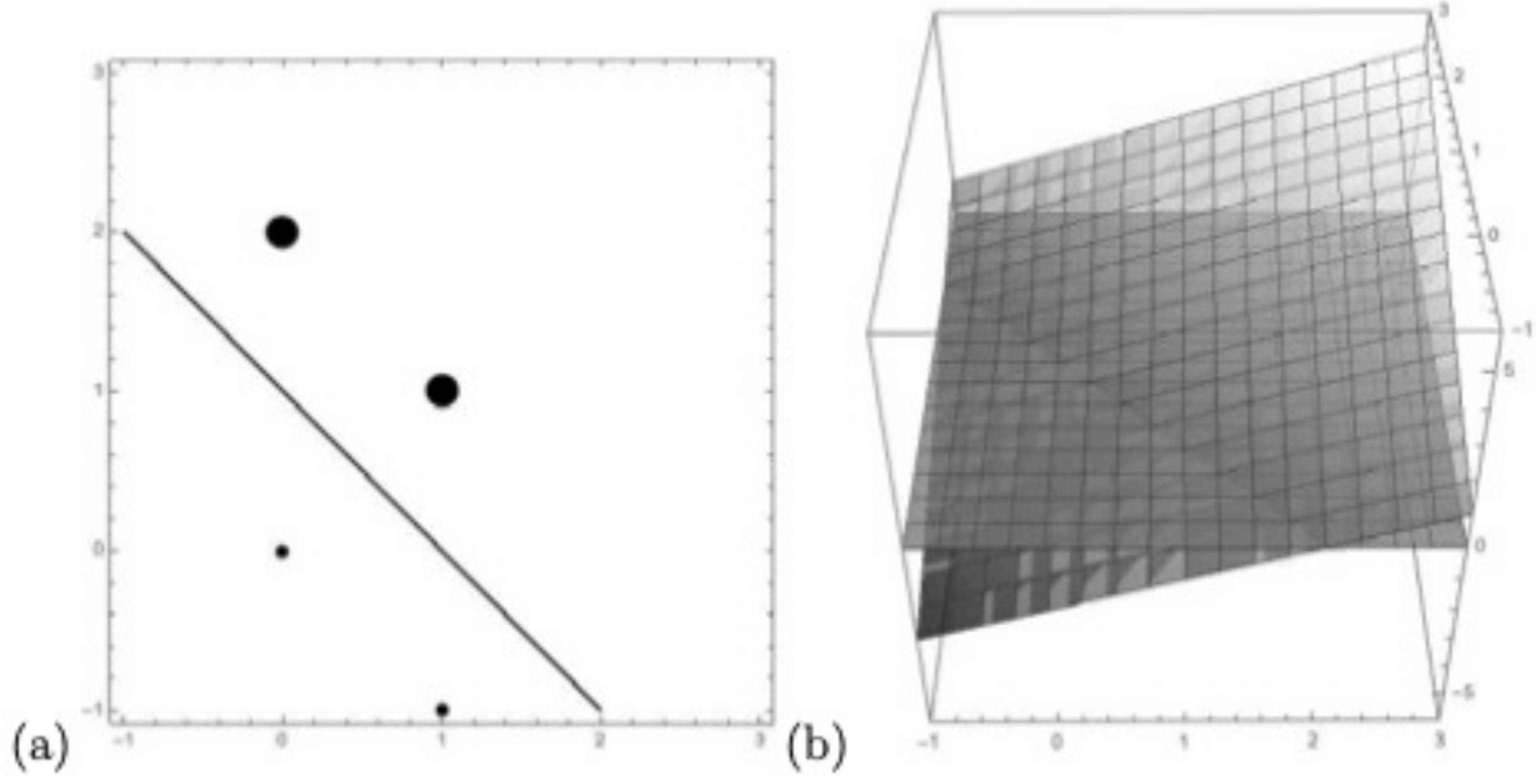
$$o = \mathrm{sgn}(\sum_{i=0}^{n} w_i x_i)$$

$$\sum_{i=0}^{n} w_i x_i > 0 \qquad for \quad C_0$$

$$\sum_{i=0}^{n} w_i x_i \leq 0 \qquad for \quad C_1$$

X_0=*1*, bias...



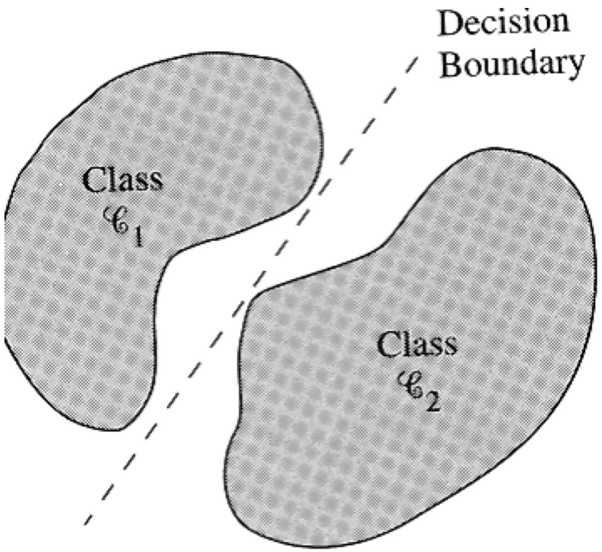


(a) The two classes 1 (indicated by a big point) and −1 (indicated by a small point) are separated by the line $-1 + x1 + x2 = 0.$

(b) The hyperplane $-1+x1+x2 = y$ defines the line for y=0.

- The goal of a perceptron is to correctly classify the set of pattern $\boldsymbol{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,..\boldsymbol{x}_m\}$ into one of the classes C_1 and C_2
- The output for class C_1 is $o=1$ and for C_2 is $o=-1$

  - For $n=2$ ➔



Perceptron learning rule

- Consider linearly separable problems
- How to find appropriate weights
  - Initialize each vector ***w*** to some small *random* values
- Look if the output pattern *o* belongs to the desired class, has the desired value *d*

$$w^{new} = w^{old} + \Delta w \qquad \Delta w = \eta \cdot (d - o) \cdot x$$

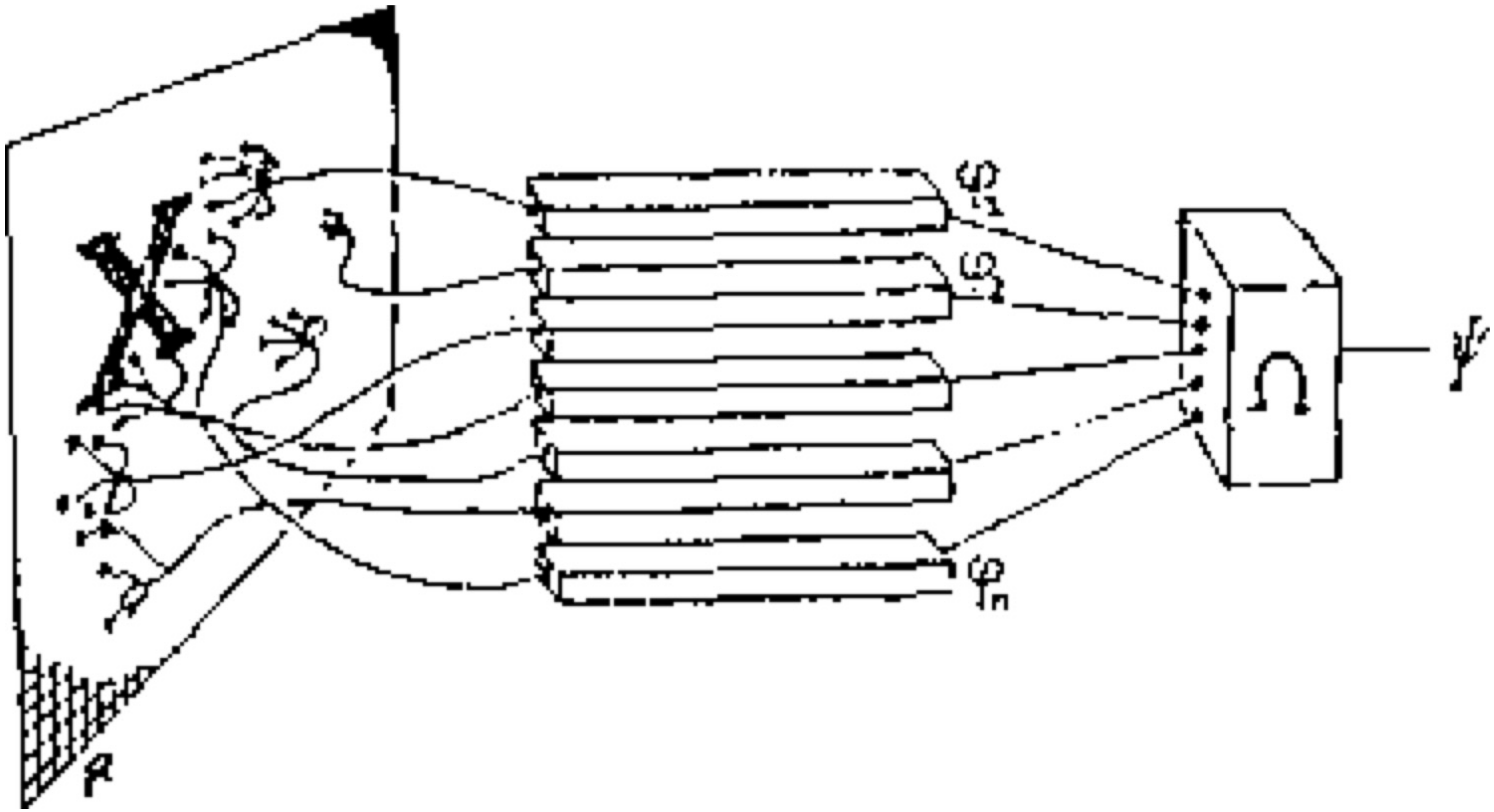
- η is called the **learning rate**
- $0 < \eta \leq 1$

- In supervised learning the network has its output compared with known correct answers
  - Supervised learning
  - Learning with a teacher
- *(d-o)* plays the role of the error signal

Algorithm

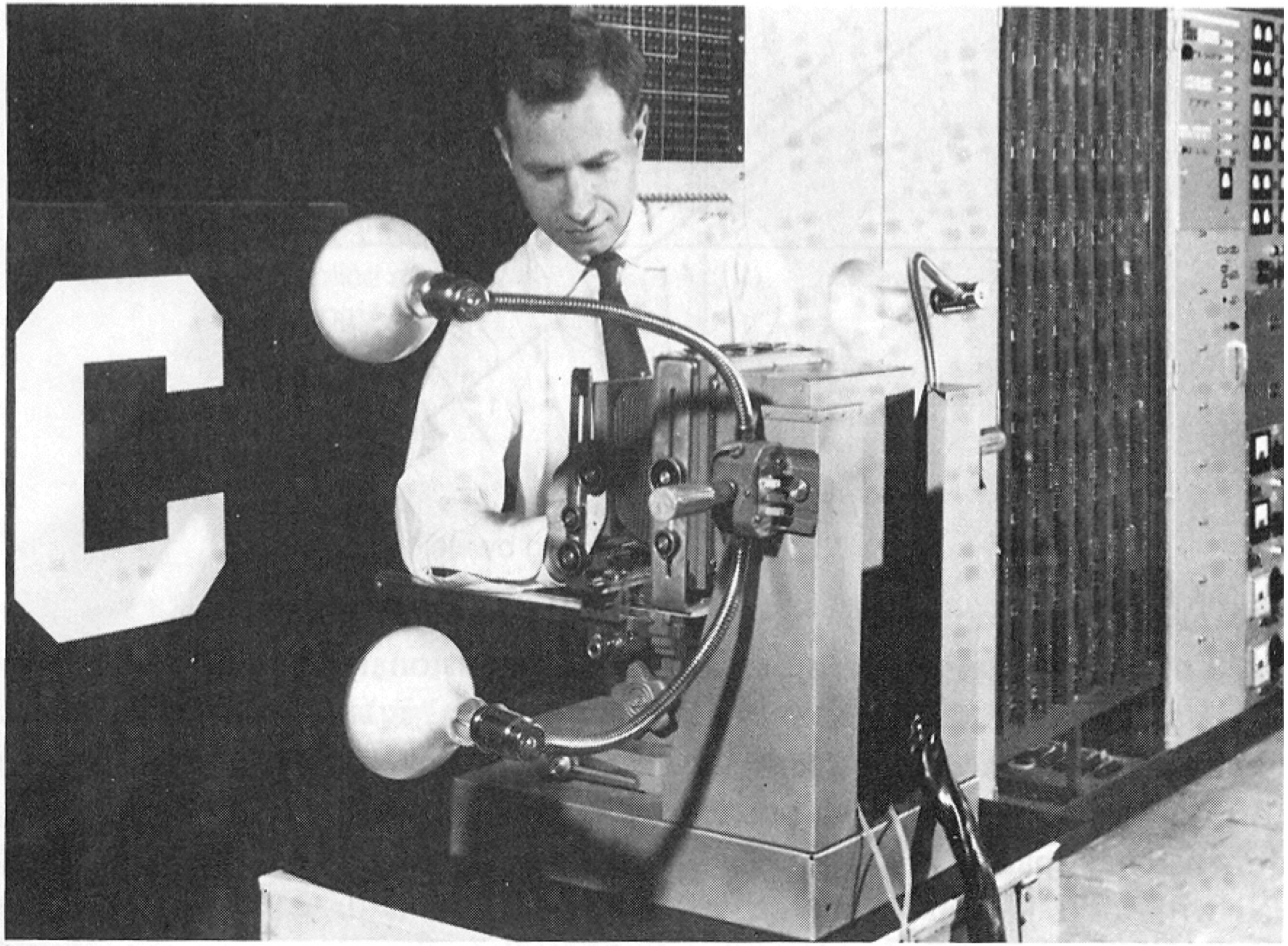
1. iterations=0;
2. $\eta \in (0, 1]$;
3. Initialise all the weights $w_0, w_, \cdots, w_D$ to some random values;
4. Choose a pattern $\mathbf{x}_k$ out of the training set;
5. Compute $net_k = \sum_{i=1}^{D} w_j \cdot x_{k,j} + w_0 = \langle \mathbf{x}_k | \mathbf{w} \rangle + w_0 \cdot x_0$;
6. Compute the output by the activation function $o_k = sgn(net_k)$;
7. Compute $\Delta w_j = \eta \cdot (t_k - o_k) \cdot x_{k,j}$;
8. Update the weights $w_j = w_j + \Delta w_j$;
9. iterations++;
10. If no change in weights for all training set or maximum number of iteration THEN STOP ELSE GOTO 4;

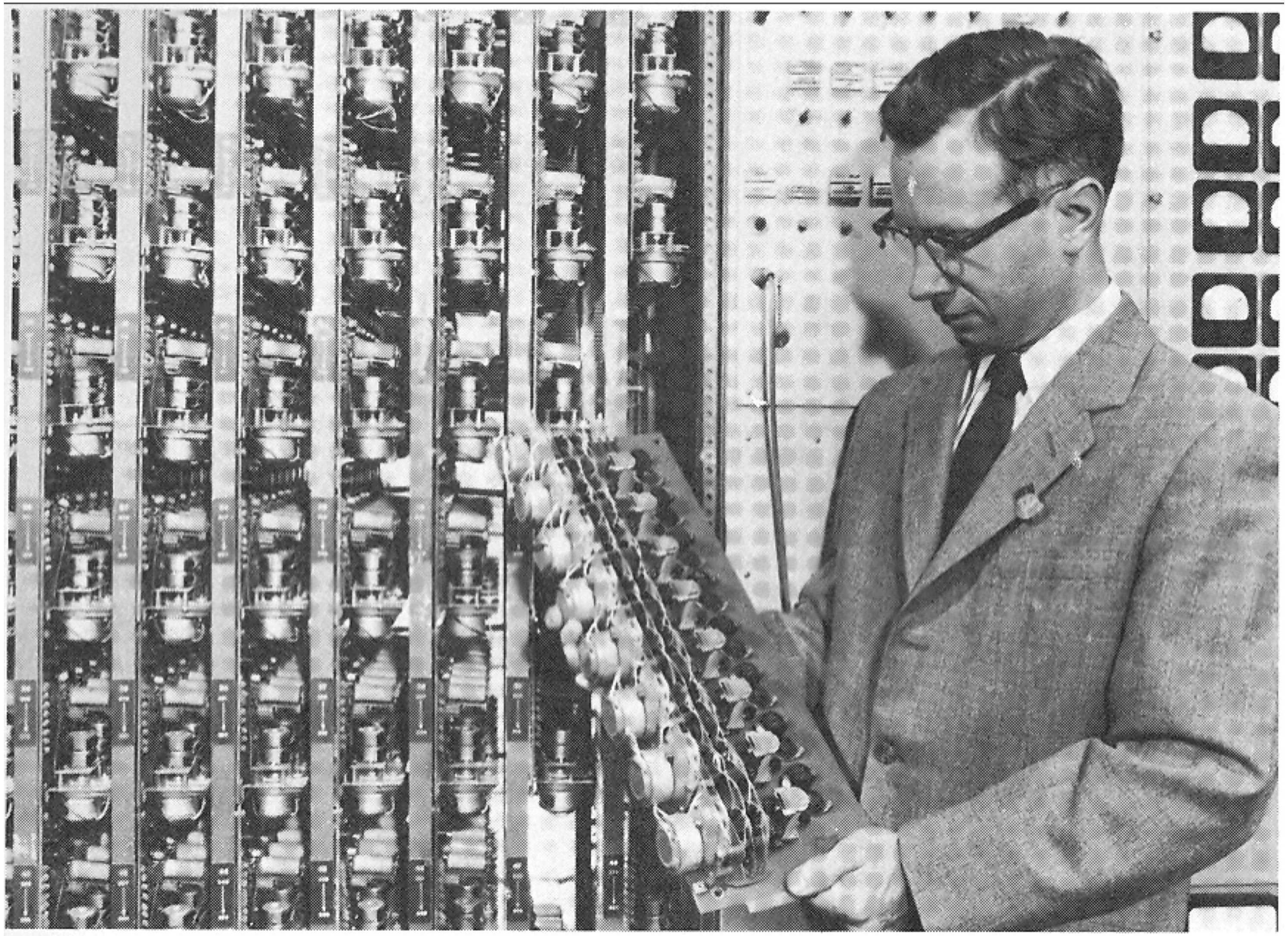
Constructions

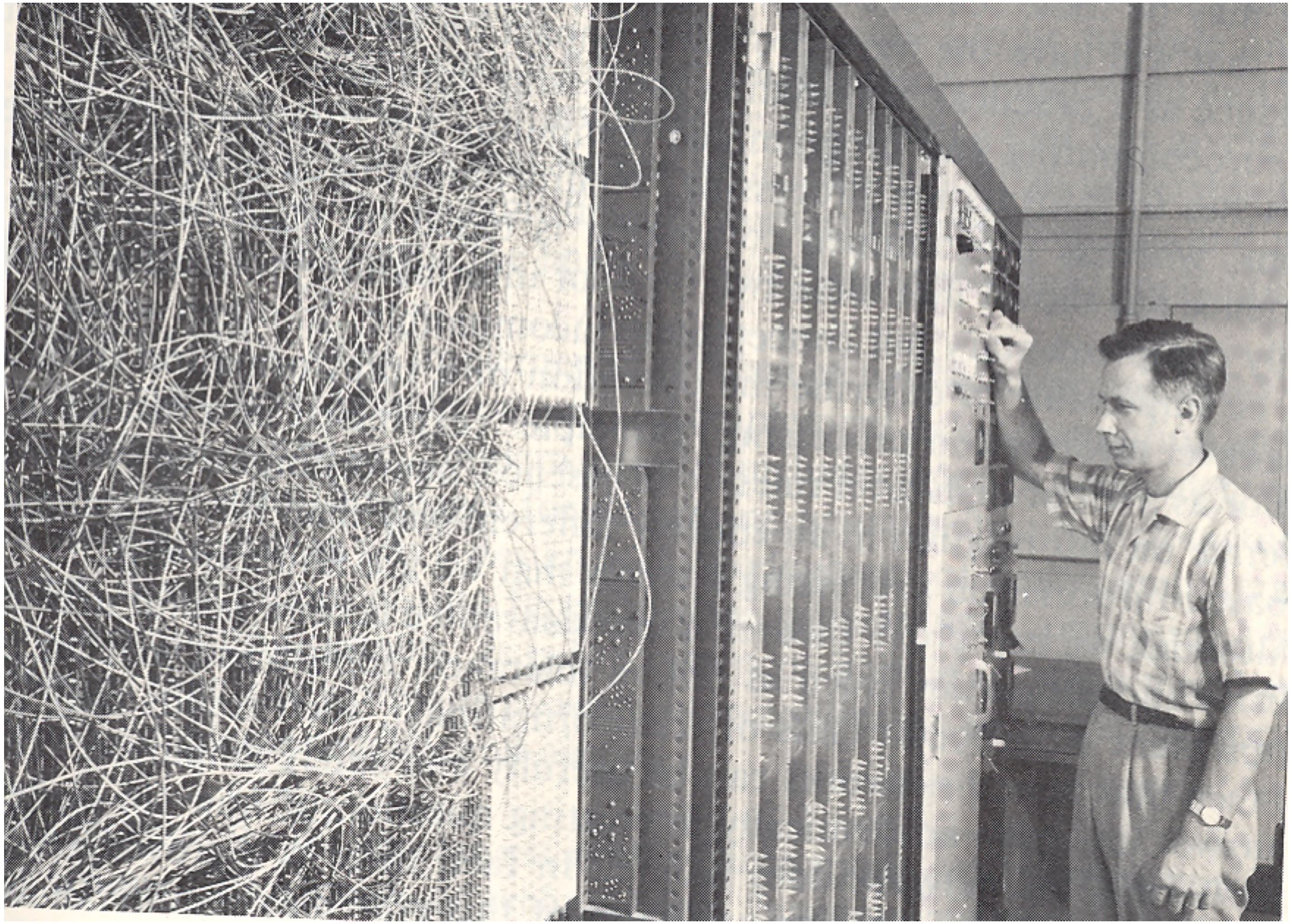


Frank Rosenblatt

- 1928-1971


C

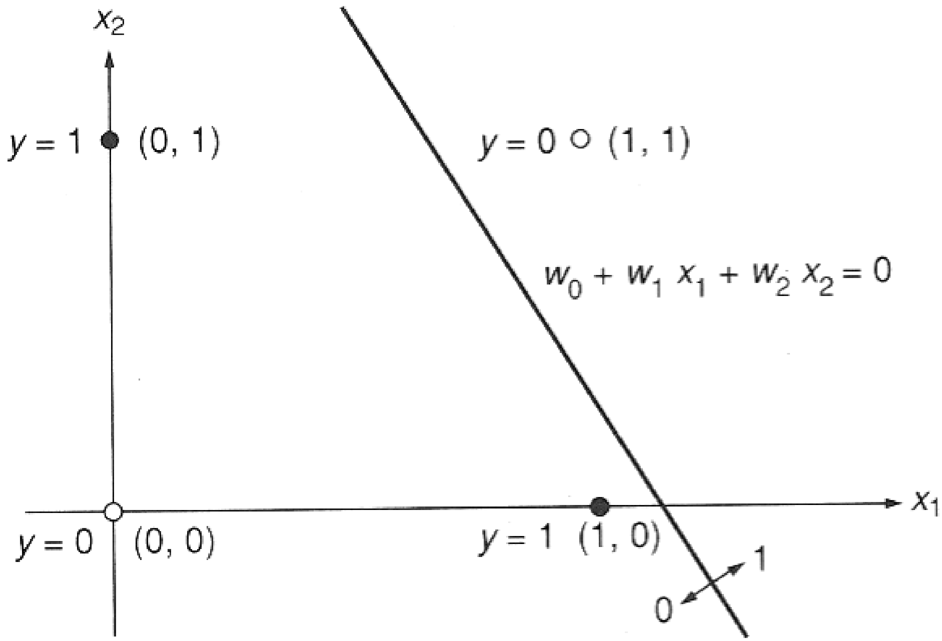




- Rosenblatt's bitter rival and professional nemesis was Marvin Minsky of Carnegie Mellon University
- Minsky despised Rosenblatt, hated the concept of the perceptron, and wrote several polemics against him
- For years Minsky crusaded against Rosenblatt on a very nasty and personal level, including contacting every group who funded Rosenblatt's research to denounce him as a charlatan, hoping to ruin Rosenblatt professionally and to cut off all funding for his research in neural nets

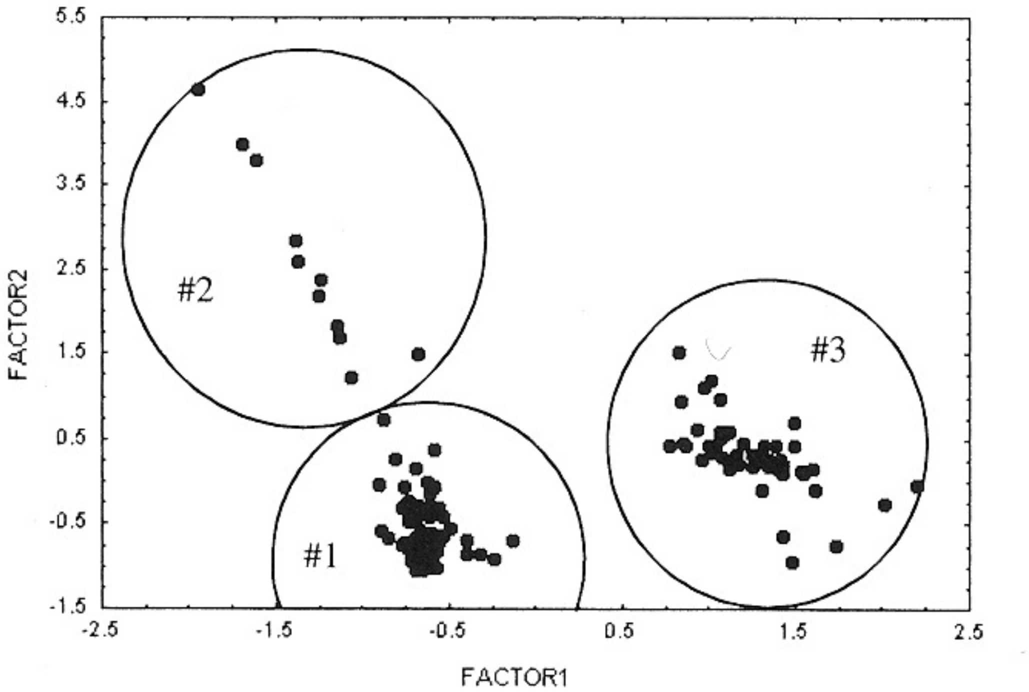
XOR problem and Perceptron

- By Minsky and Papert in mid 1960



k Means Clustering (Unsupervised Learning)

- The standard algorithm was first proposed by Stuart Lloyd in 1957

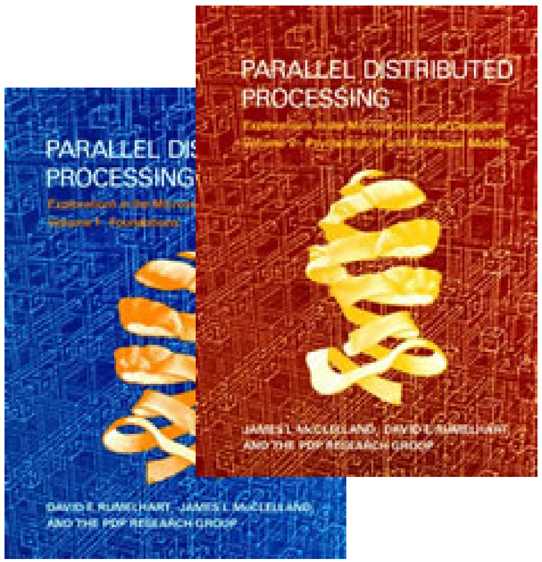


Back-propagation (1980)

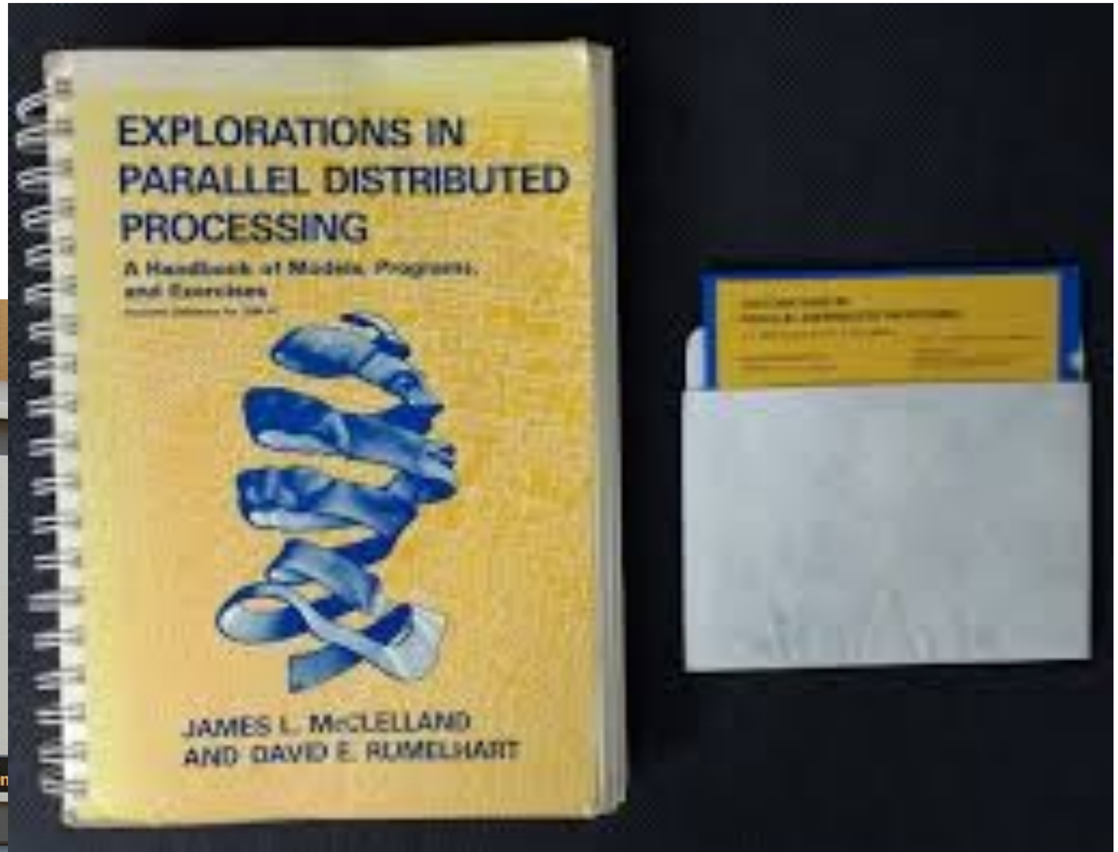
- Back-propagation is a learning algorithm for multi-layer neural networks
- It was invented independently several times
  - Bryson an Ho [1969]
  - Werbos [1974]
  - Parker [1985]
  - **Rumelhart et al. [1986]**

Parallel Distributed Processing - Vol. 1
Foundations
David E. Rumelhart, James L. McClelland and the PDP Research Group

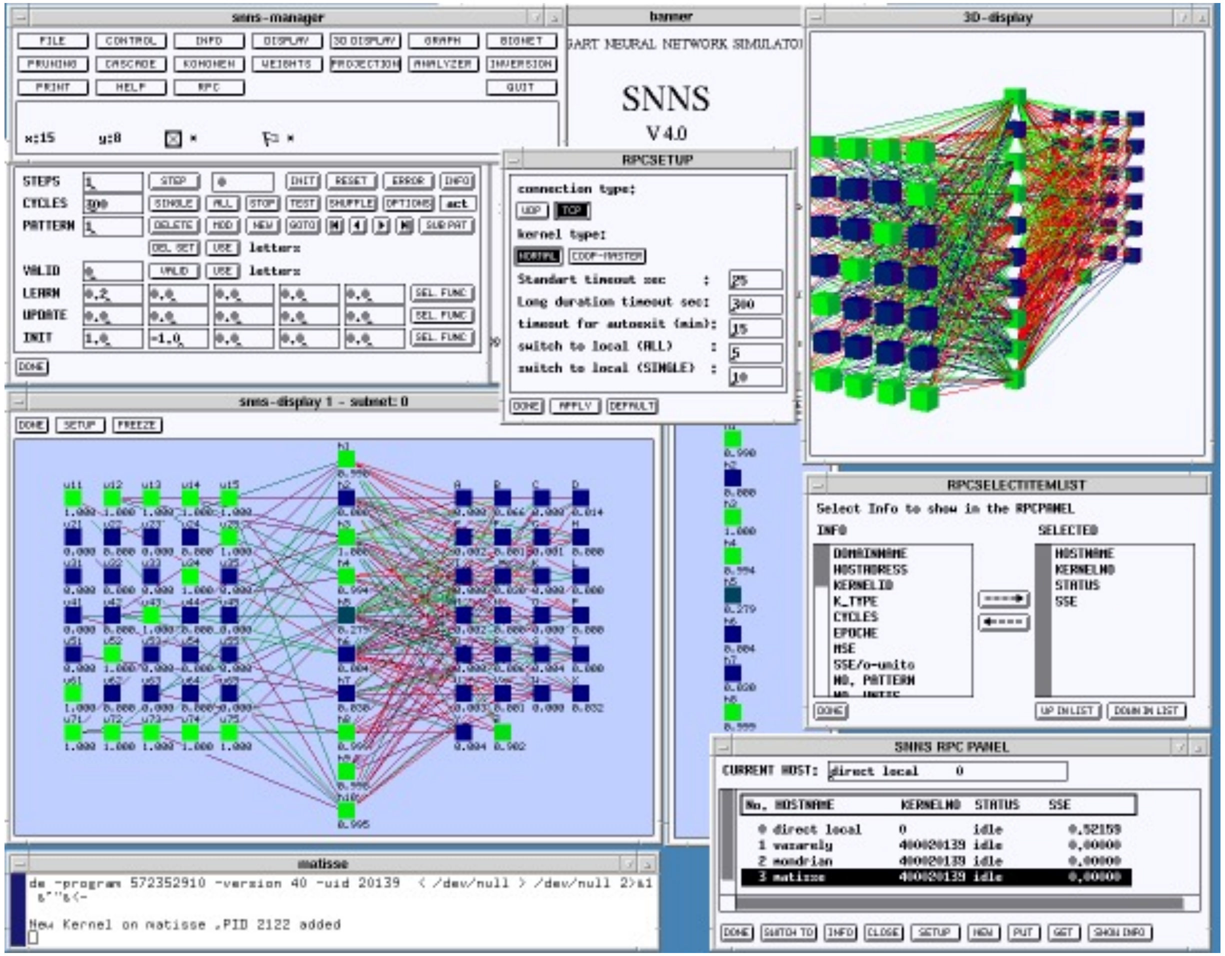
What makes people smarter than computers? These volumes by a pioneering neurocomputing.....

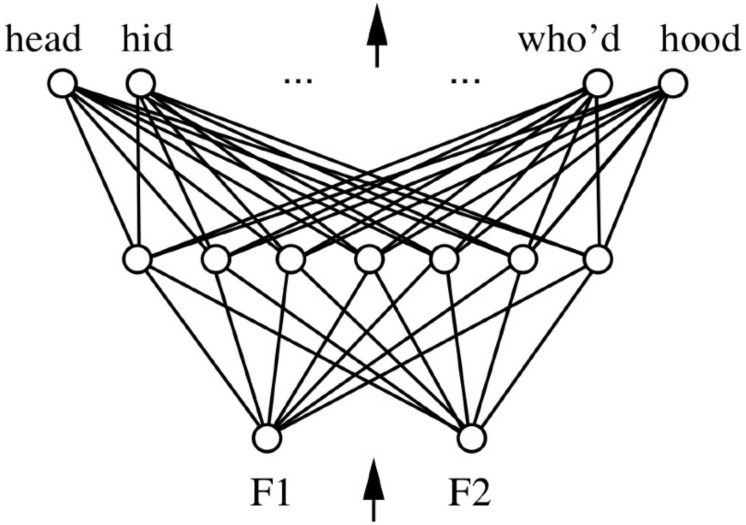


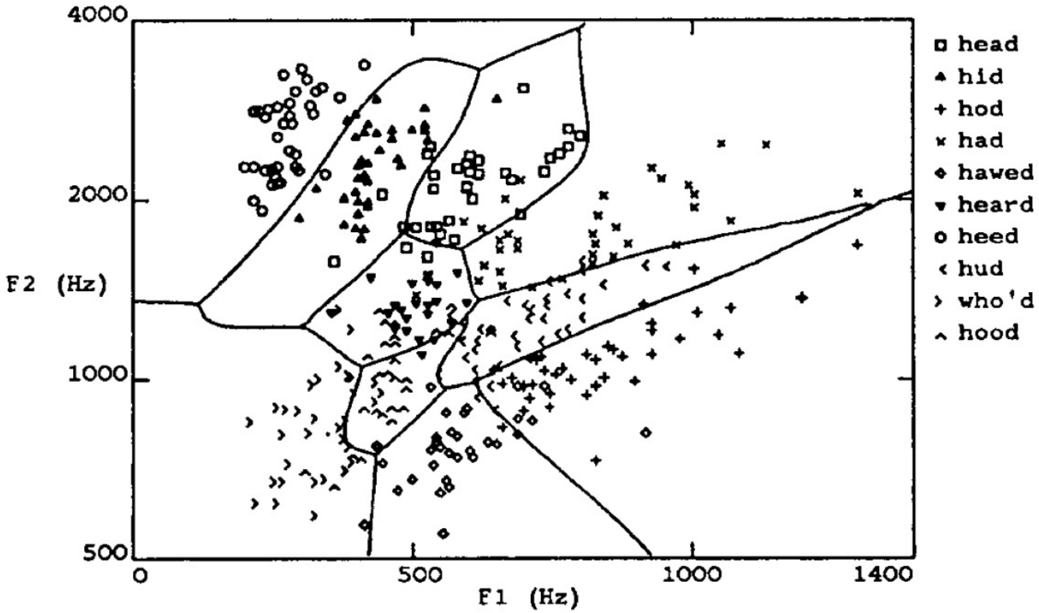
The good old days…



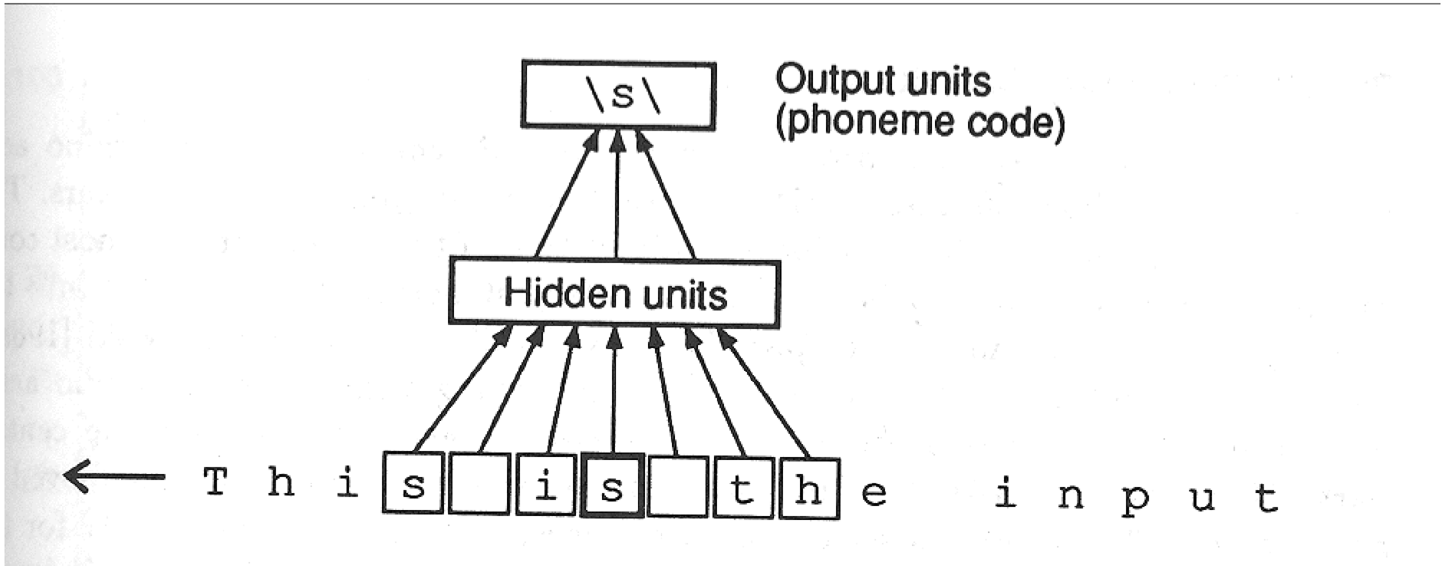
Everyone was doing Back-propagation….

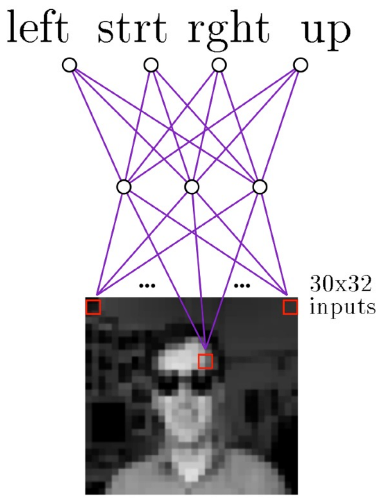



head
hid
...
...
who'd
hood
F1
F2


4000
2000
1000
500
F2 (Hz)
0
500
1000
1400
F1 (Hz)
head
hid
hod
had
hawed
heard
heed
hud
who'd
hood

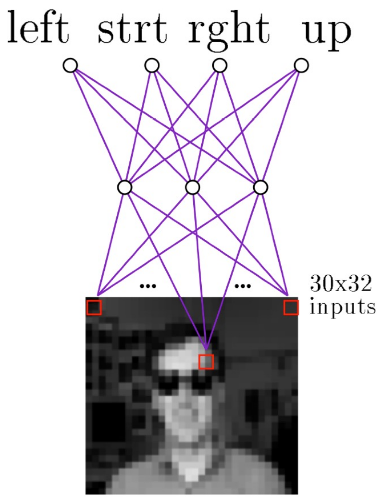
NETtalk Sejnowski et al 1987



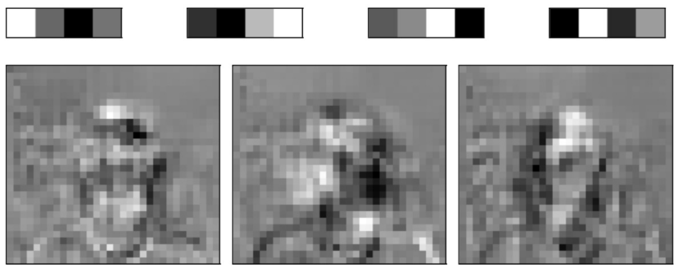


Typical input images

90% accurate learning head pose, and recognizing 1-of-20 faces



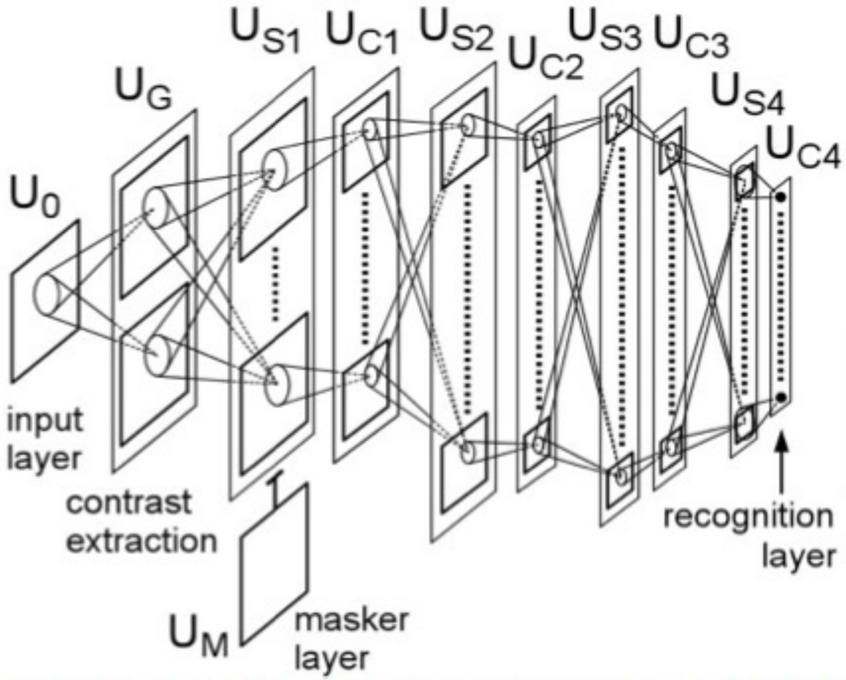
Learned Weights

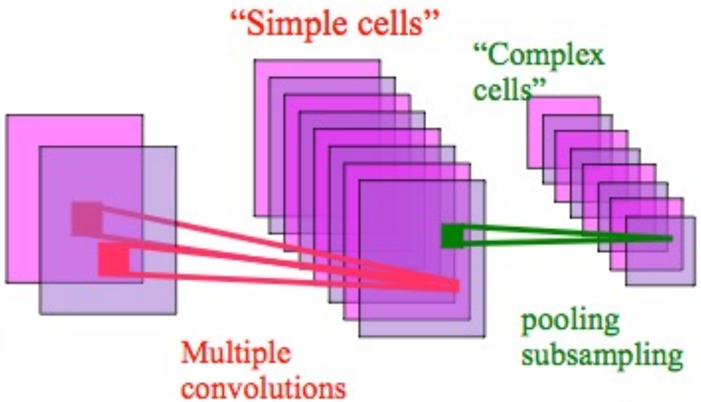


Typical input images

http://www.cs.cmu.edu/~tom/faces.html

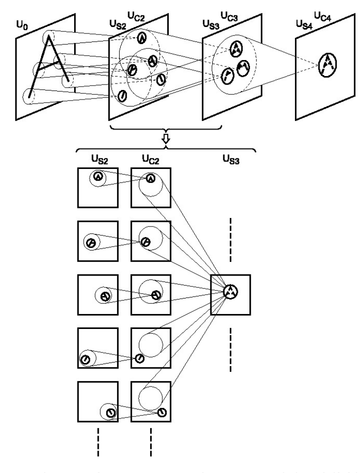
- [Hubel & Wiesel 1962]:
  - simple cells detect local features
  - complex cells "pool" the outputs of simple cells within a retinotopic neighborhood.

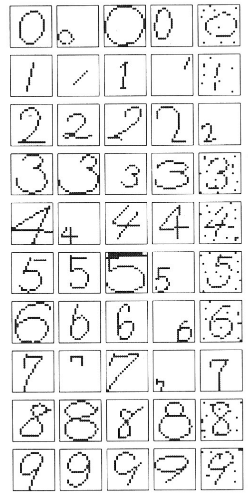




Cognitron & Neocognitron [Fukushima 1974-1982]

Kunihiko Fukushima

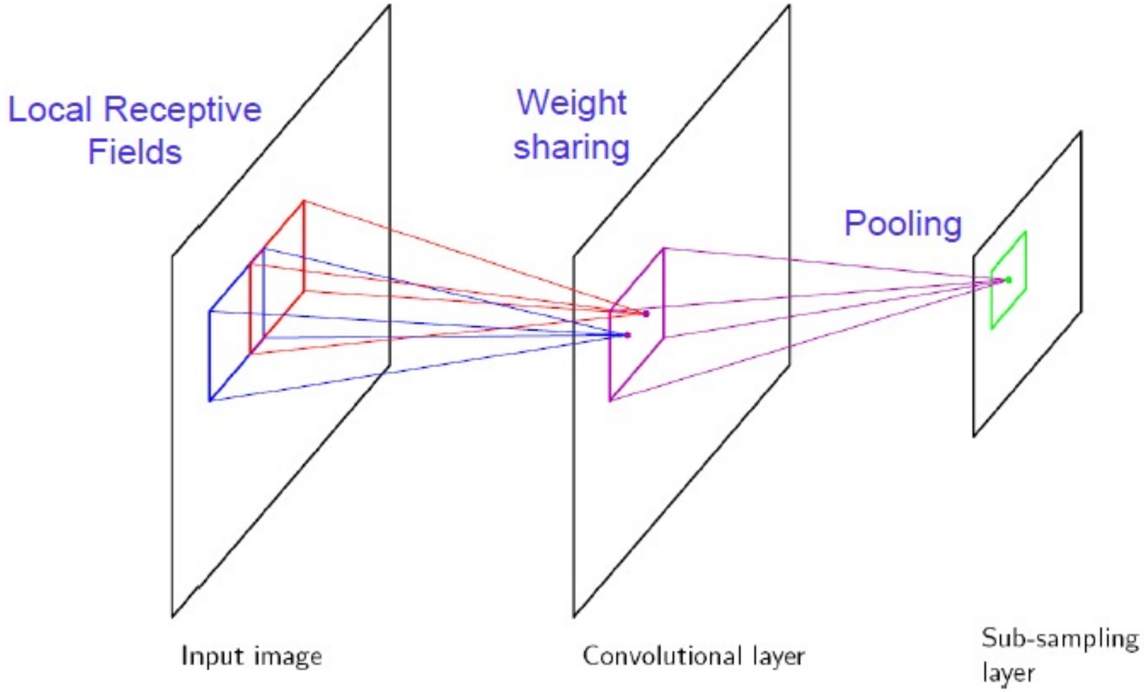


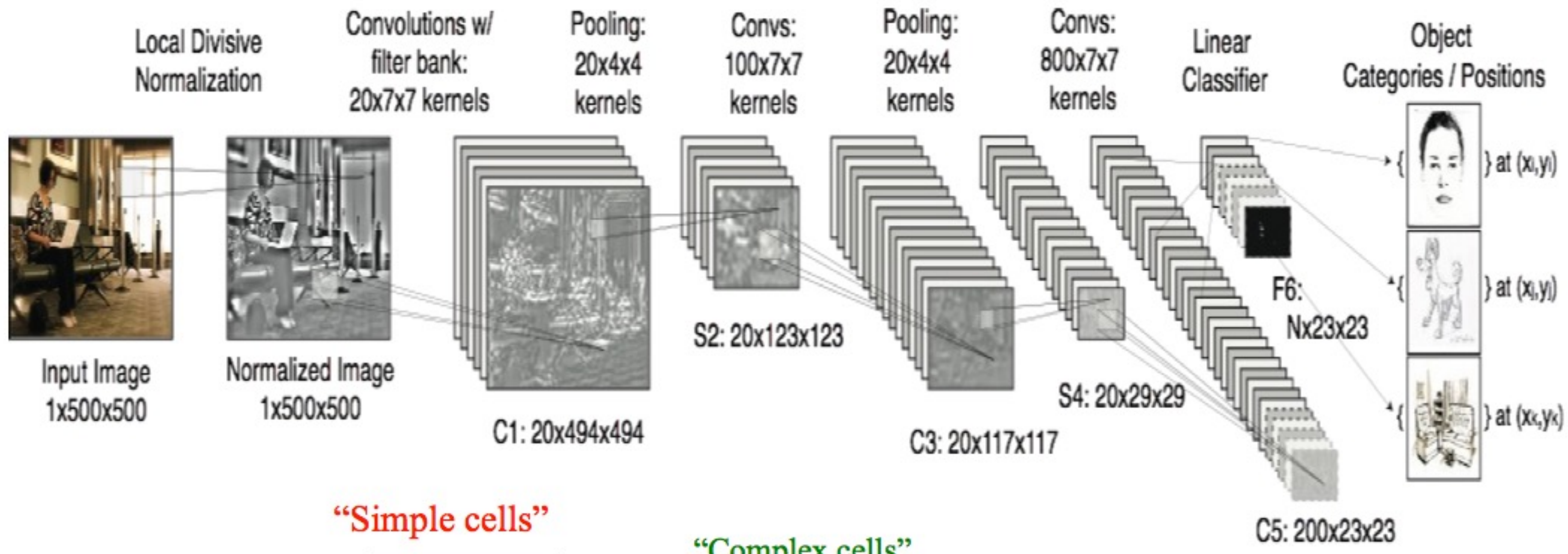


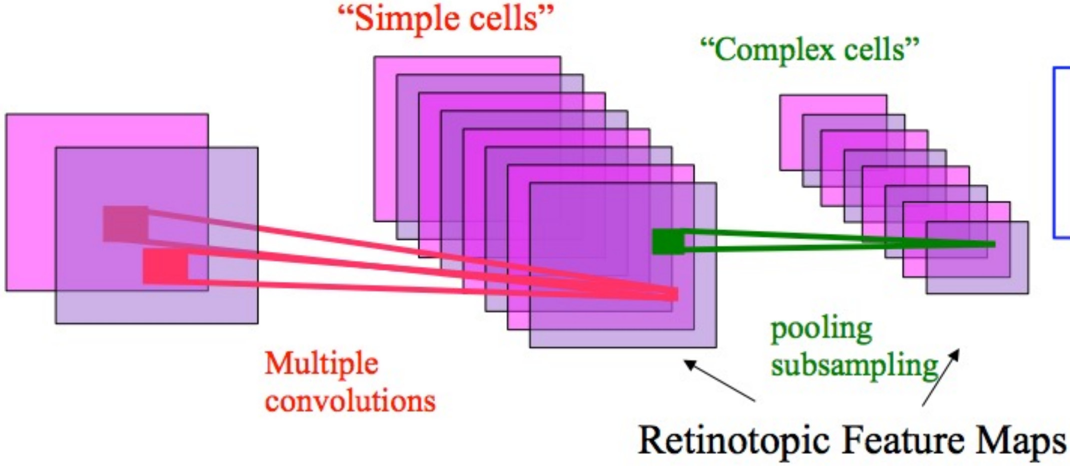
Kunihiko Fukushima received a B.Eng. degree in electronics in 1958 and a PhD degree in electrical engineering in 1966 from Kyoto University, Japan. He was a professor at Osaka University from 1989 to 1999, at the University of Electro-Communications from 1999 to 2001, at Tokyo University of Technology from 2001 to 2006; and a visiting professor at Kansai University from 2006 to 2010. Prior to his Professorship, he was a Senior Research Scientist at the NHK Science and Technology Research Laboratories. He is now a Senior Research Scientist at Fuzzy Logic Systems Institute (part-time position), and usually works at his home in Tokyo.

Convolutional Neural Networks

(LeCun et al., 1989)





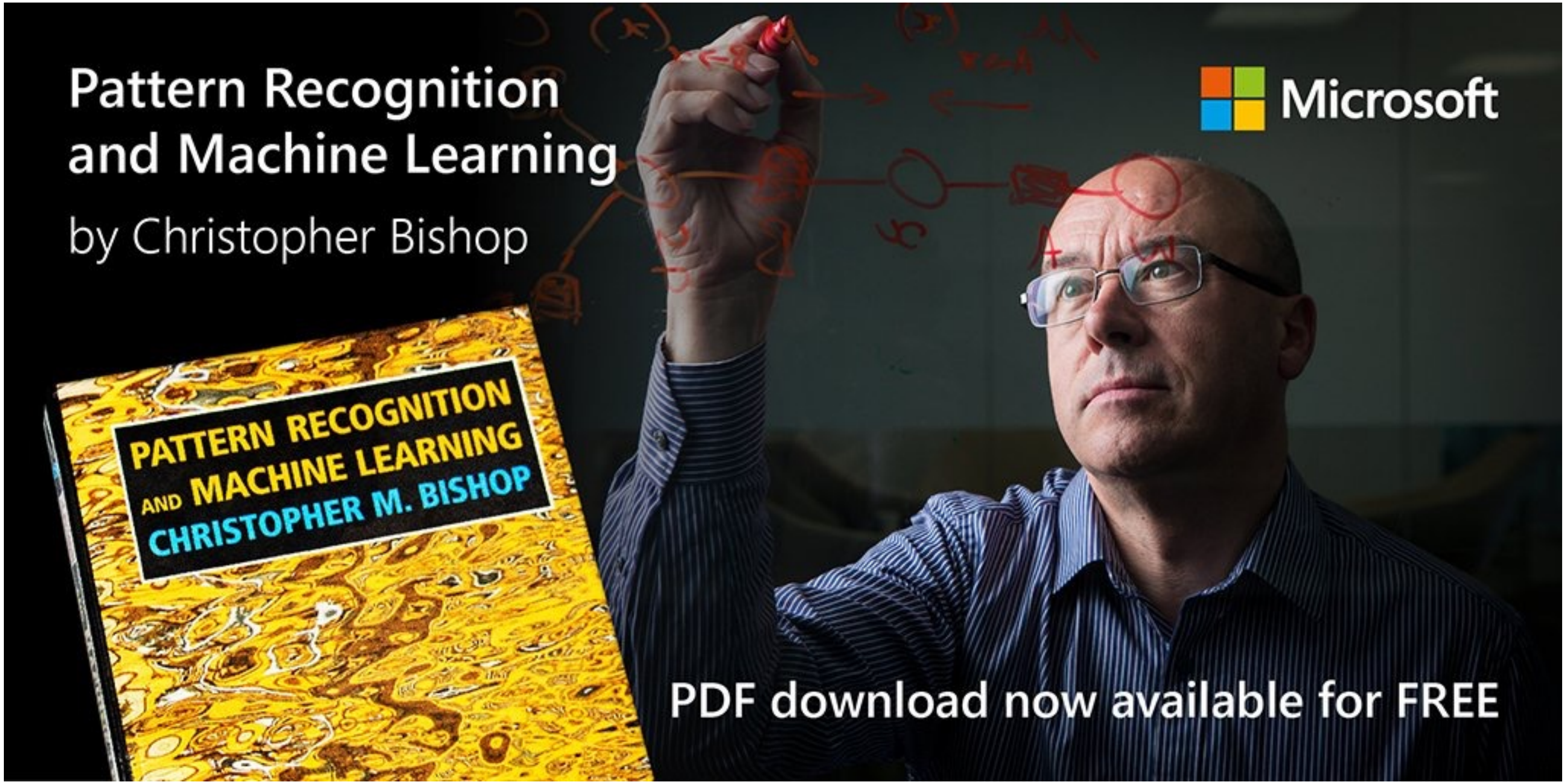


- **Training is supervised**
- **With stochastic gradient descent**

[LeCun et al. 89]

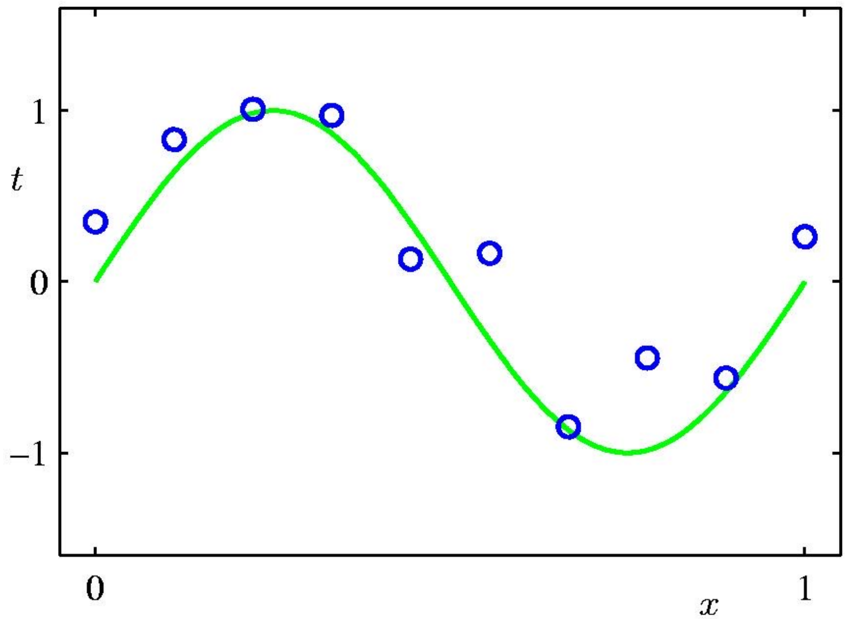
[LeCun et al. 98]

2006



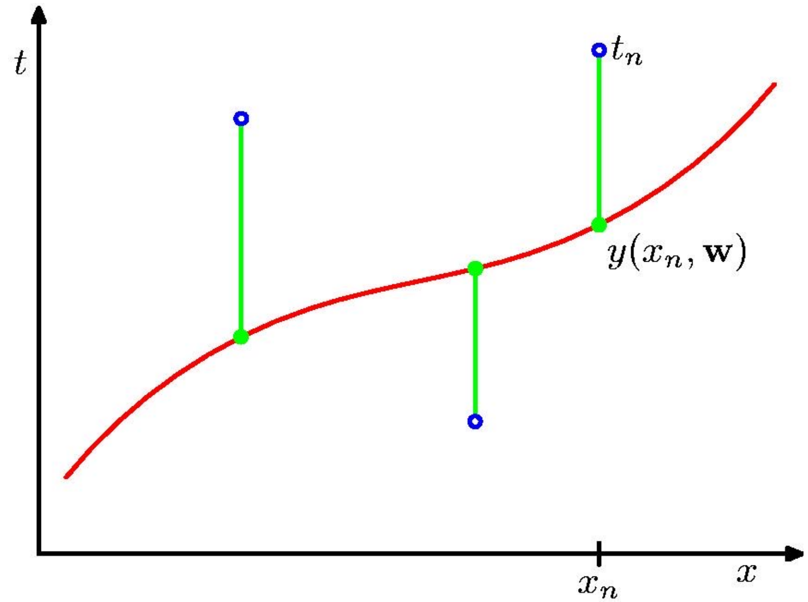
Pattern Recognition
and Machine Learning
by Christopher Bishop
Microsoft
PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
PDF download now available for FREE

Polynomial Curve Fitting



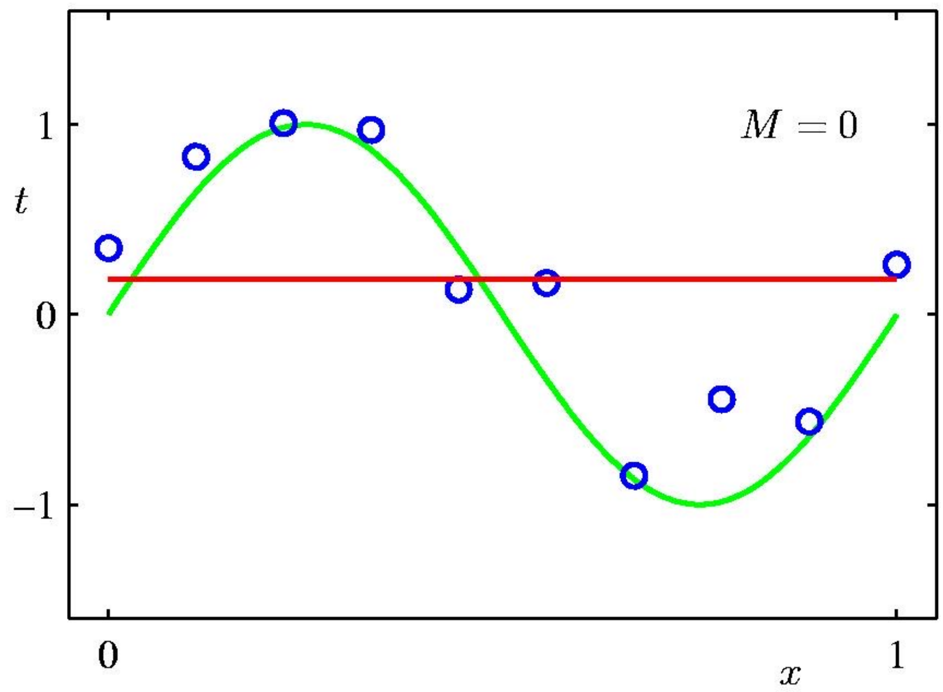
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

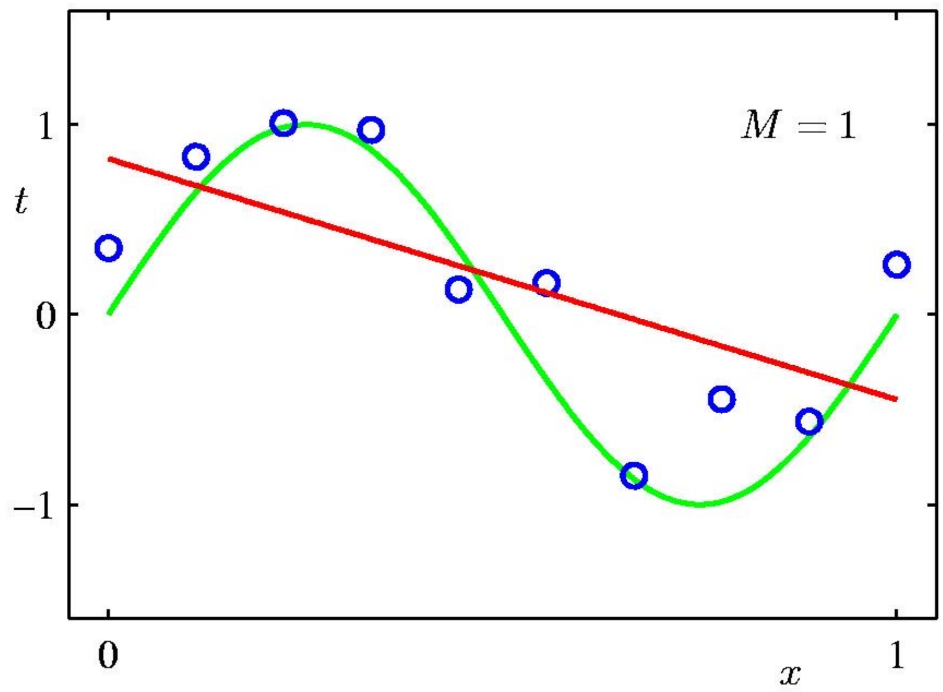


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

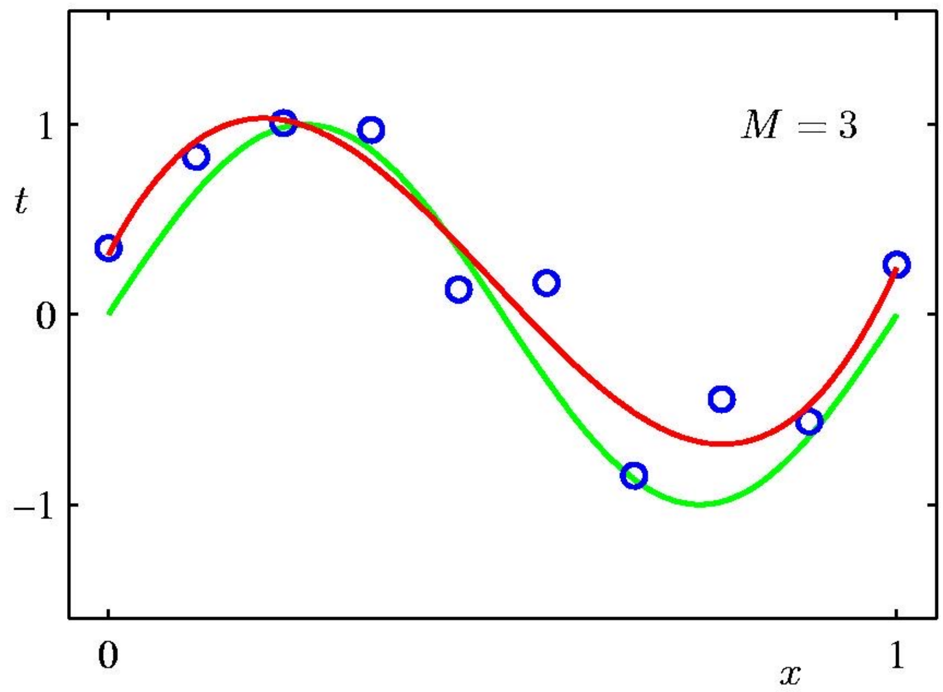
0^{th} Order Polynomial



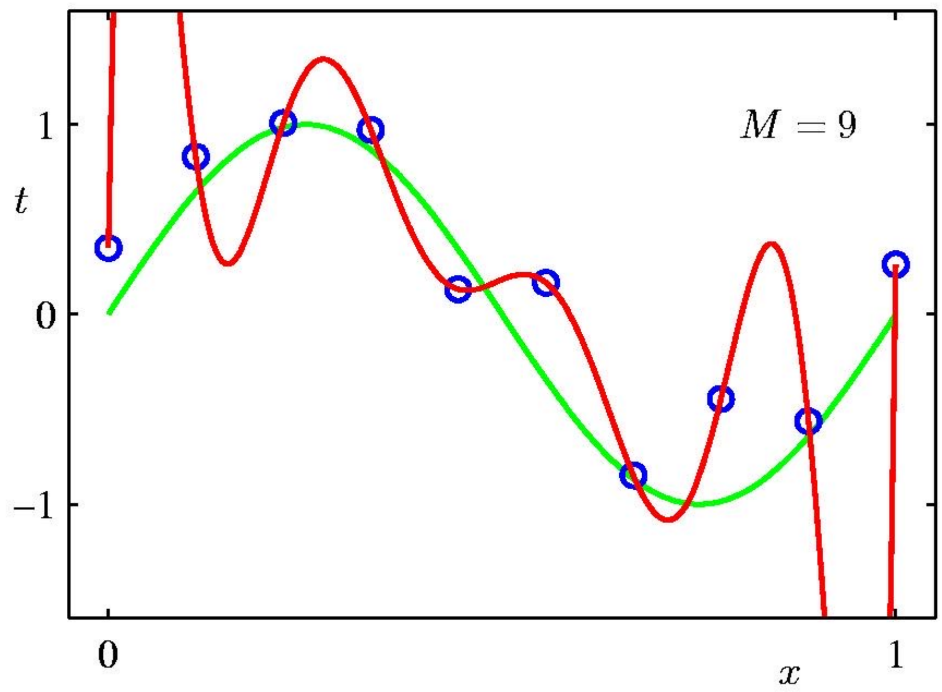
1^{st} Order Polynomial



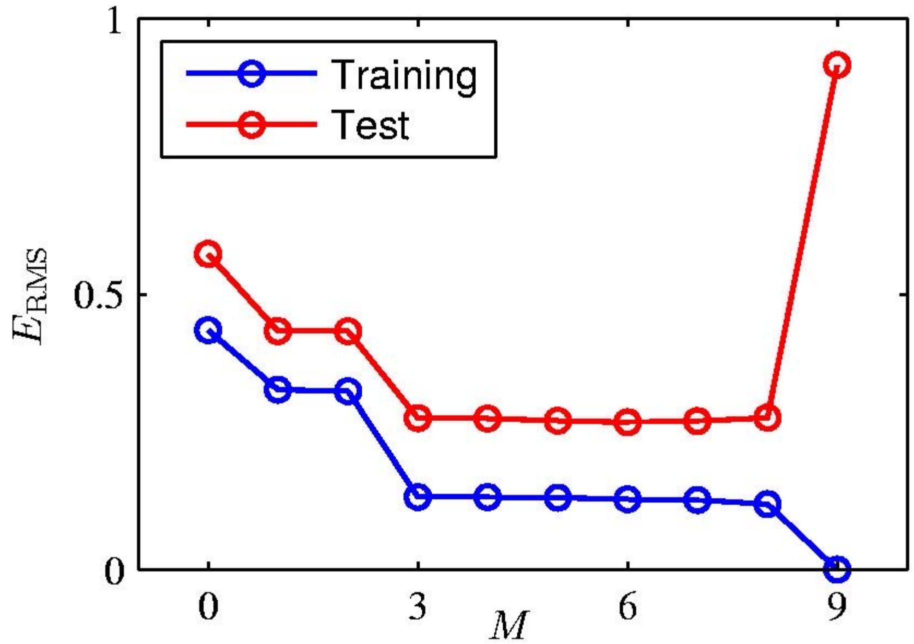
3rd Order Polynomial



9th Order Polynomial



Over-fitting



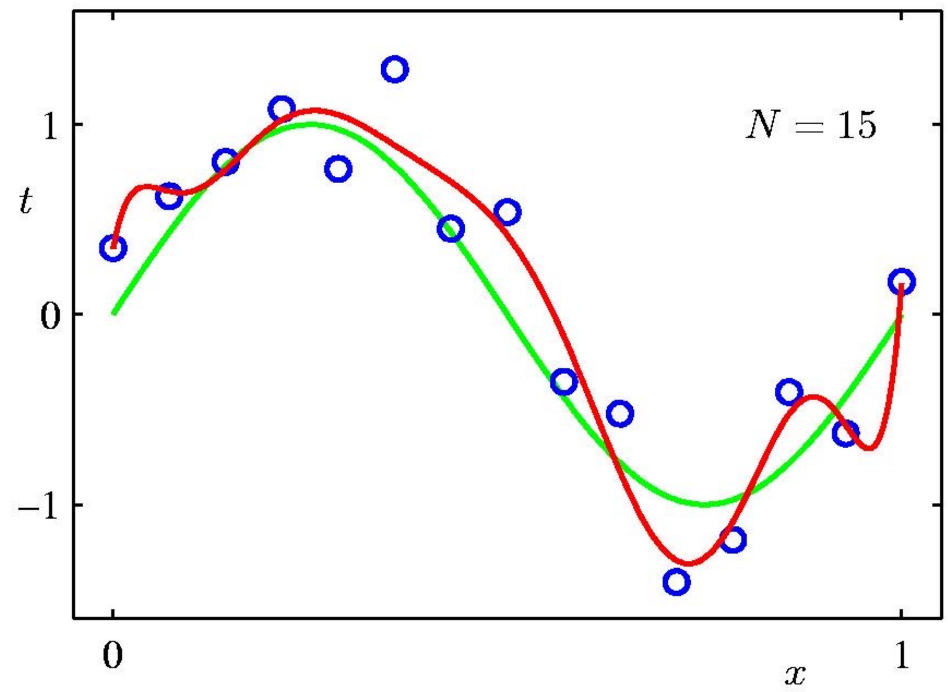
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

Polynomial Coefficients

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

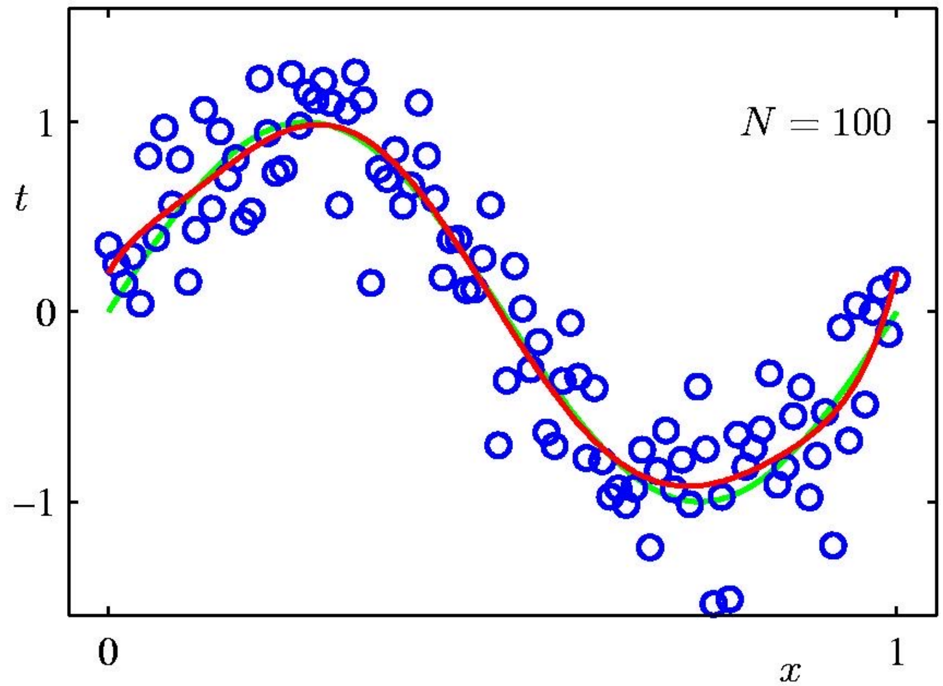
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9^{th} Order Polynomial

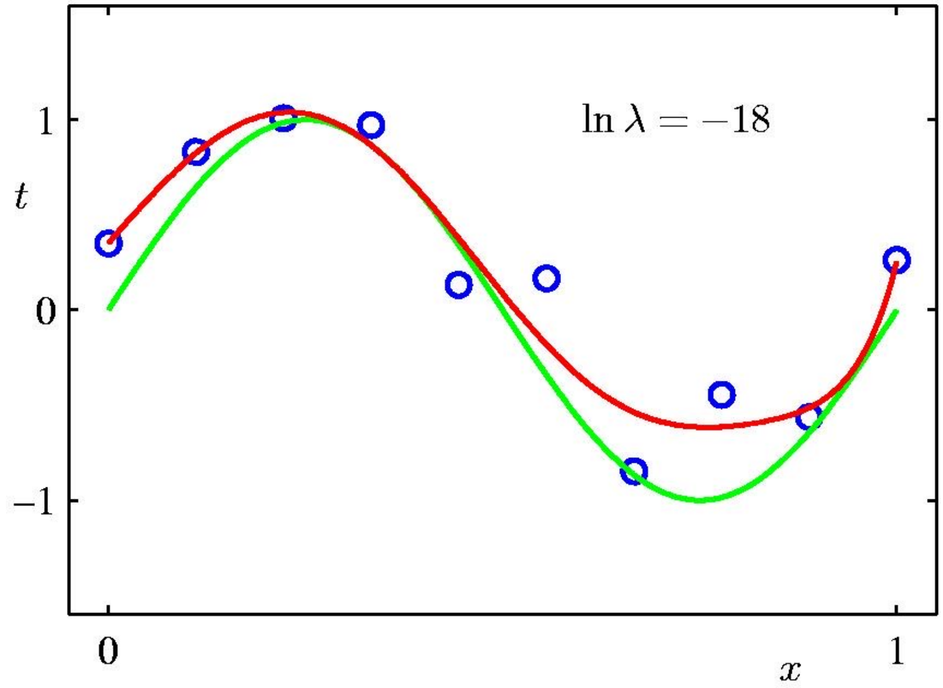


Regularization

Penalize large coefficient values

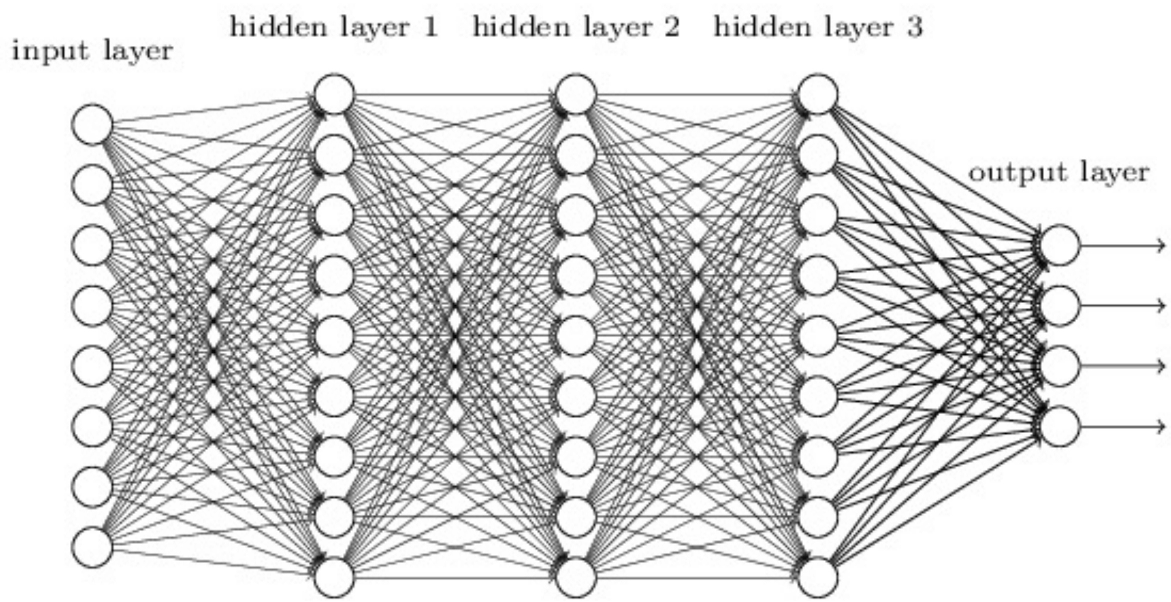
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Regularization: $\ln \lambda = -18$



Problem of Local Minima

- The immediate solution to this is to build networks with more hidden layers with regularization
- “Deep Learning”…
- Déjà vu?



Artificial intelligence pioneer (Geoffrey Hinton) says we need to start over

- Back-propagation still has a core role in AI's future.
- Entirely new methods will probably have to be invented
- "I don't think it's how the brain works," he said. "We clearly don't need all the labeled data.

What is an „A“ ?

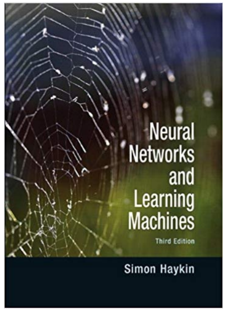
- What makes something similar to something else (specifically what makes, for example, an uppercase letter 'A' recognisable as such)

- Metamagical Themas, Douglas Hoffstader, Basic Books, 1985

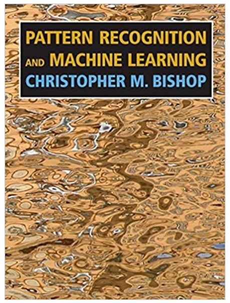
- What is the essence of dogness or house-ness?
- What is the essence of 'A'-ness?
- What is the essence of a given person's face, that it will not be confused with other people's faces?
  - How to convey these things to computers, which seem to be best at dealing with hard-edged categories--categories having crystal-clear, perfectly sharp boundaries?

- What Next?
- Example of what is machine learning: Decision Trees

Literature

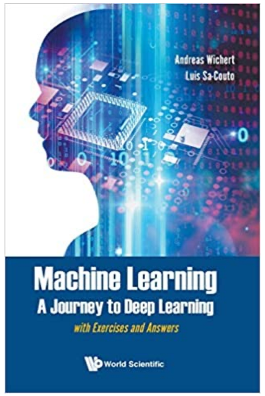


- Simon O. Haykin, Neural Networks and Learning Machine, (3rd Edition), Pearson 2008
  - Chapter 1



- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
  - Section *1.1*

Literature



- Machine Learning - A Journey to Deep Learning, A. Wichert, Luis Sa-Couto, World Scientific, 2021
  - Chapter 1