MAX-PLANCK-GESELLSCHAFT

What is Machine Learning?

MLSS 2013

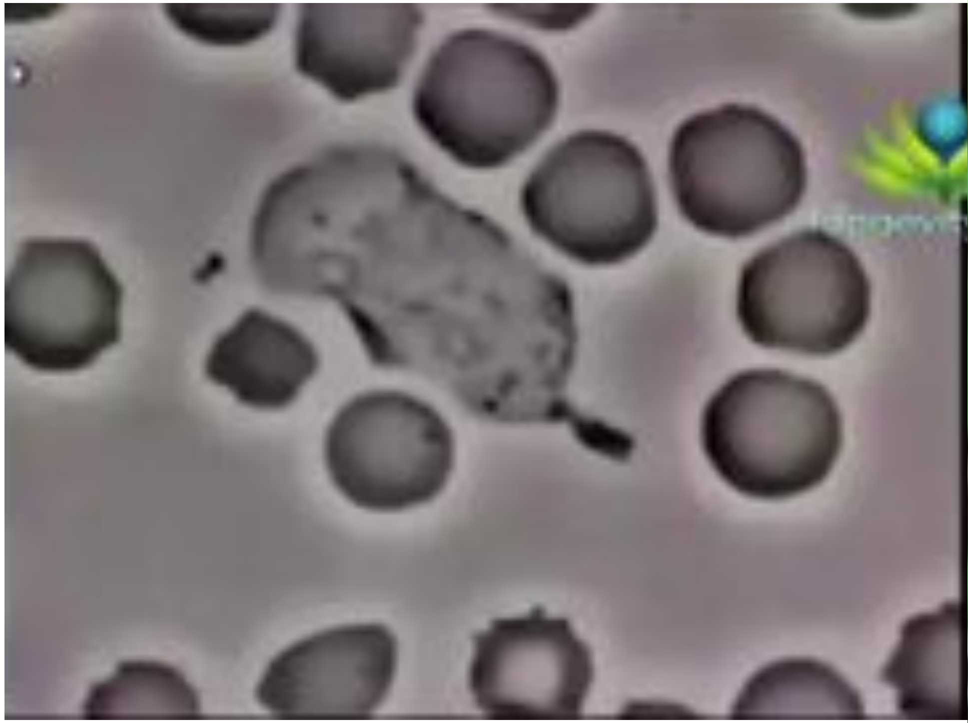
Bernhard Schölkopf
Empirical Inference Department
MPI for Intelligent Systems
Tübingen

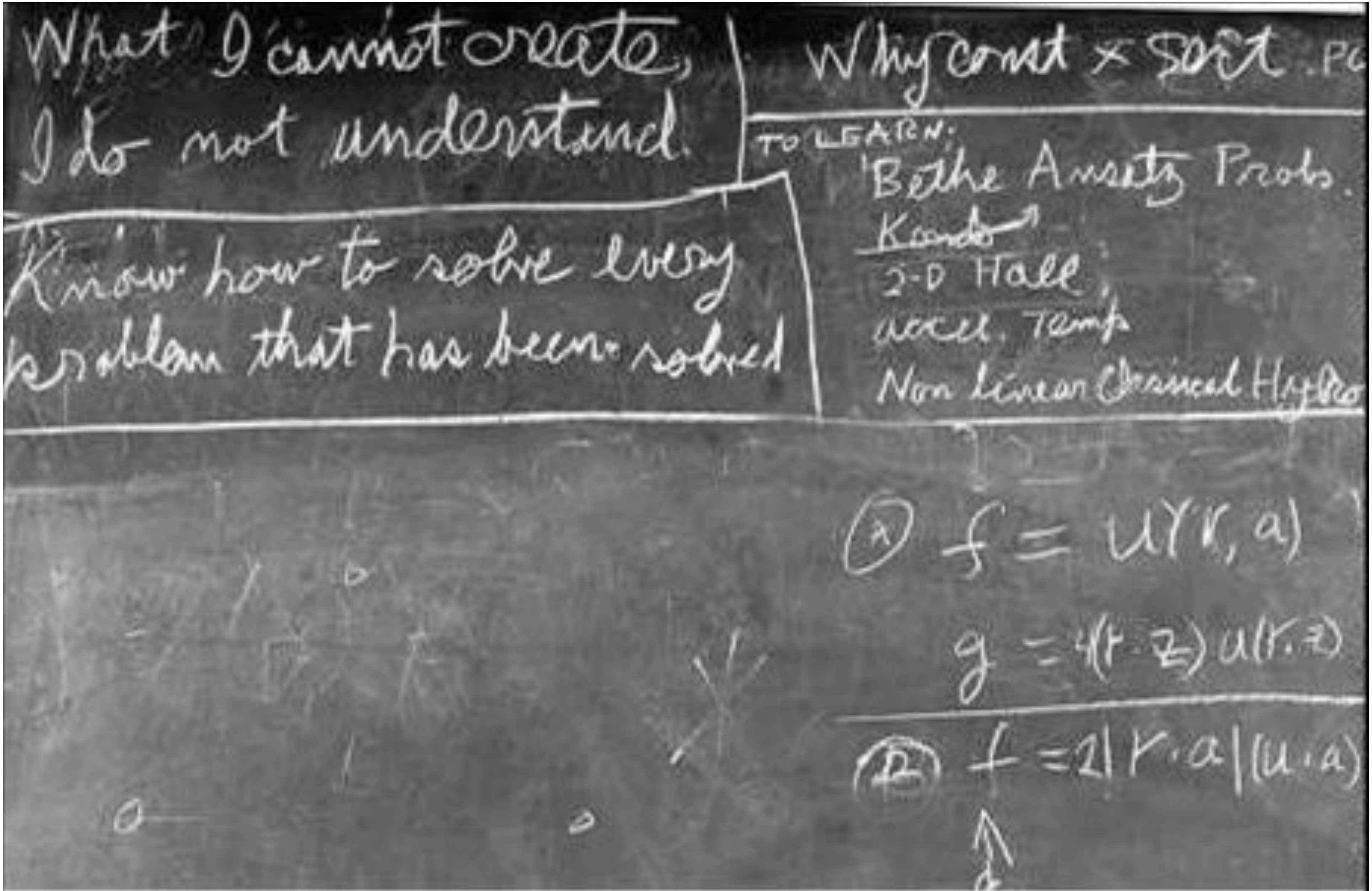
Two parts:

1. some basic ideas of machine learning and inference
2. some interesting aspects of its history (subjective..)

In order to act successfully in a complex environment, biological systems have developed sophisticated **adaptive behaviors** through learning and evolution.

What is adaptive behavior?





What I cannot create,
I do not understand.
Know how to solve every
problem that has been solved
Why const × sort
TO LEARN:
Bethe Ansatz Probs.
Kondo
2-D Hall
accel. Temp
Non Linear Classical Hydro

The pinnacle of adaptive behavior is learning and intelligence.

We try to build intelligent systems in order to understand their organizing principles.

The best way to build them at present seems to be **learning and inference,** and we have made substantial progress understanding learning and inference as organizing principles of intelligent behavior.

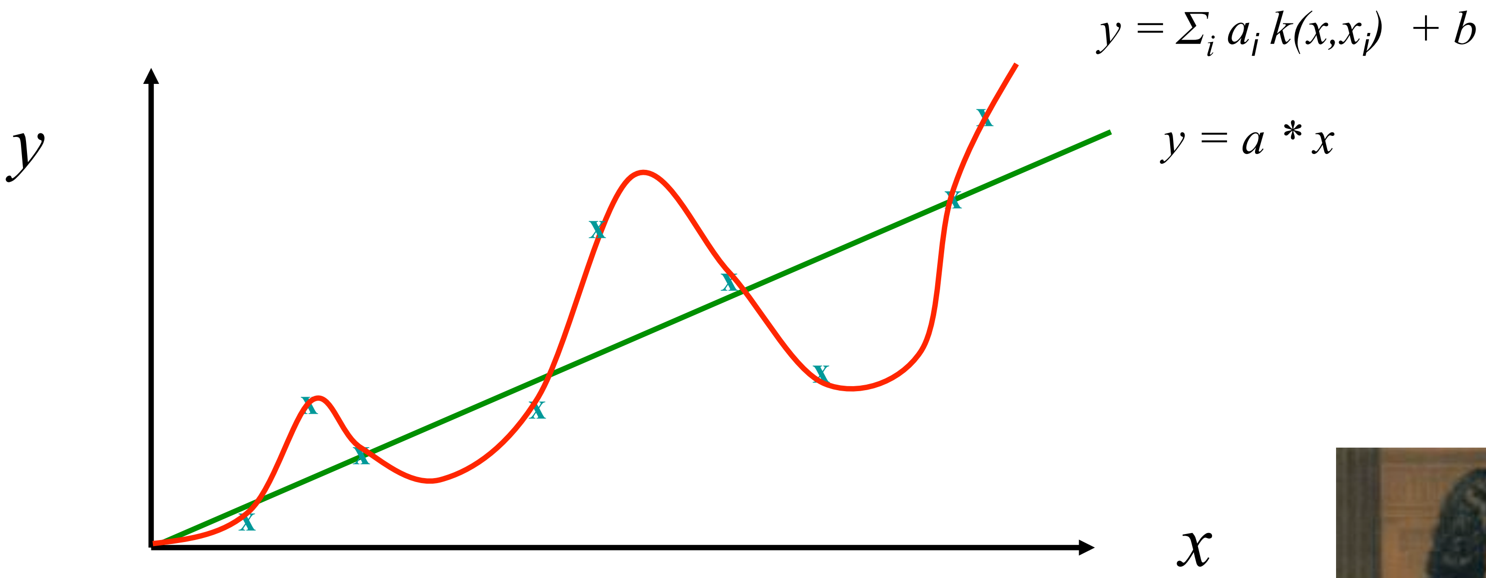
Two definitions of learning

(1) Learning is the acquisition of knowledge about the world. *Kupfermann (1985)*

(2) Learning is an adaptive change in behavior caused by experience. *Shepherd (1988)*

Empirical Inference

- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference



Leibniz, Weyl, Chaitin

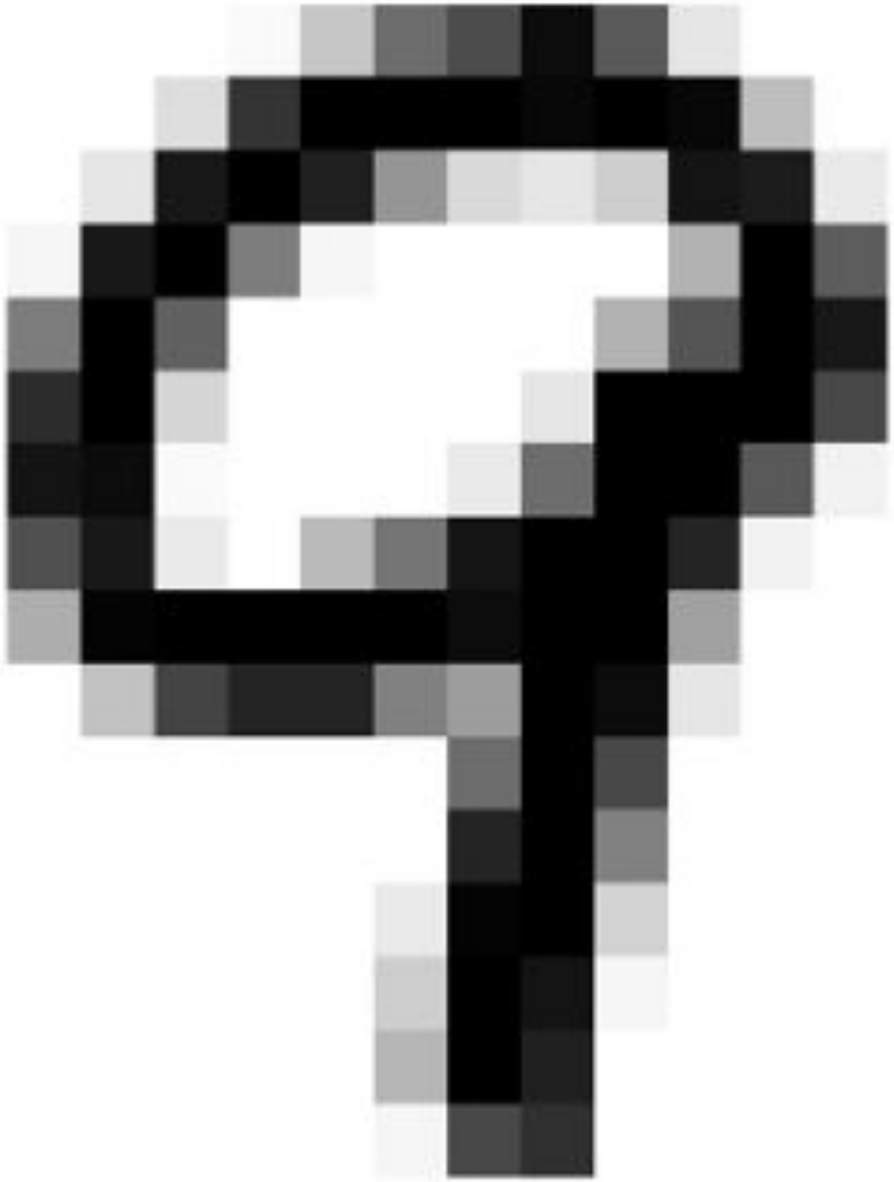
Empirical Inference

- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference

"If your experiment needs statistics [inference]*,*
you ought to have done a better experiment." (Rutherford)

Empirical Inference, II

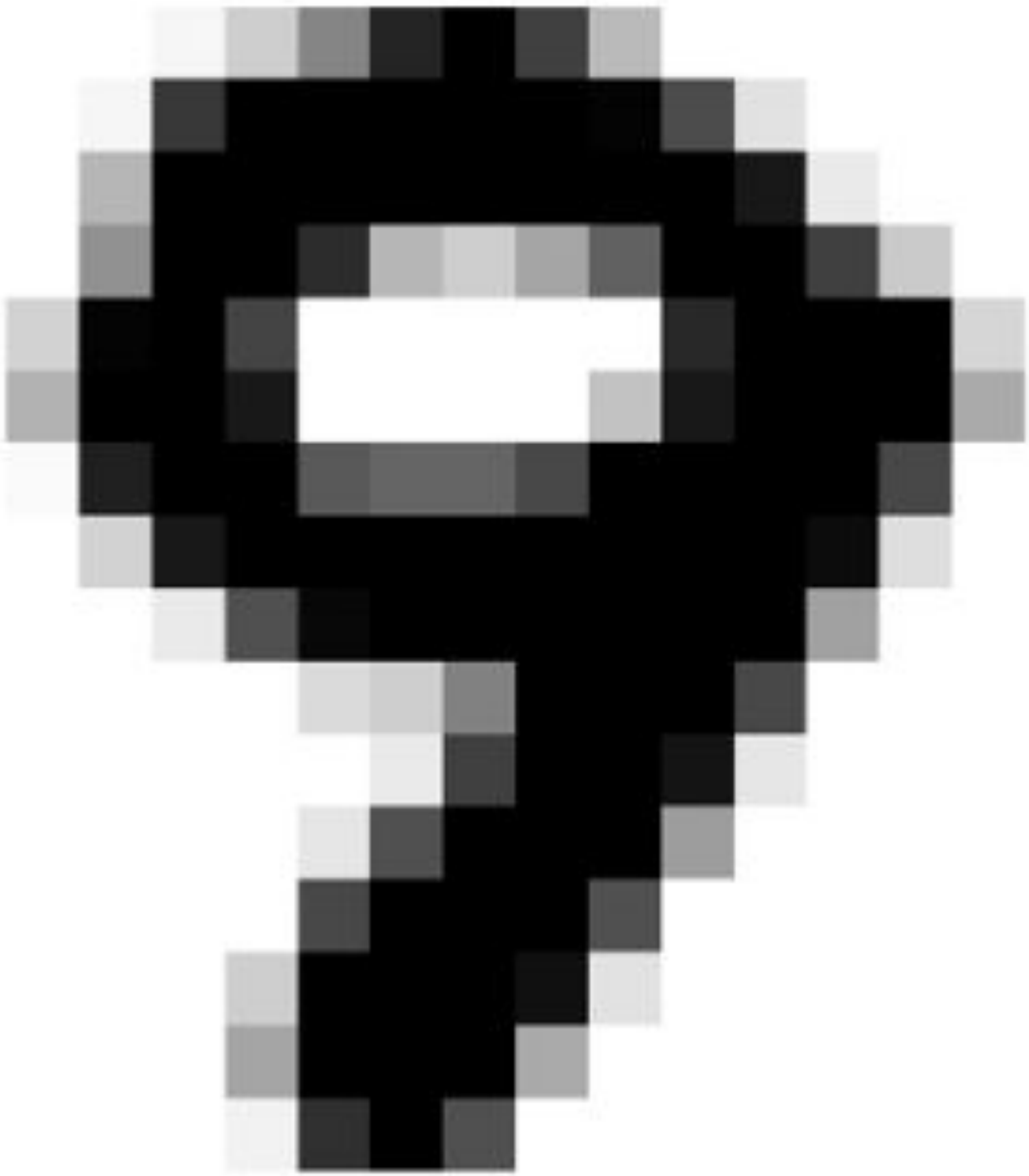
- Example 2: perception



9

9

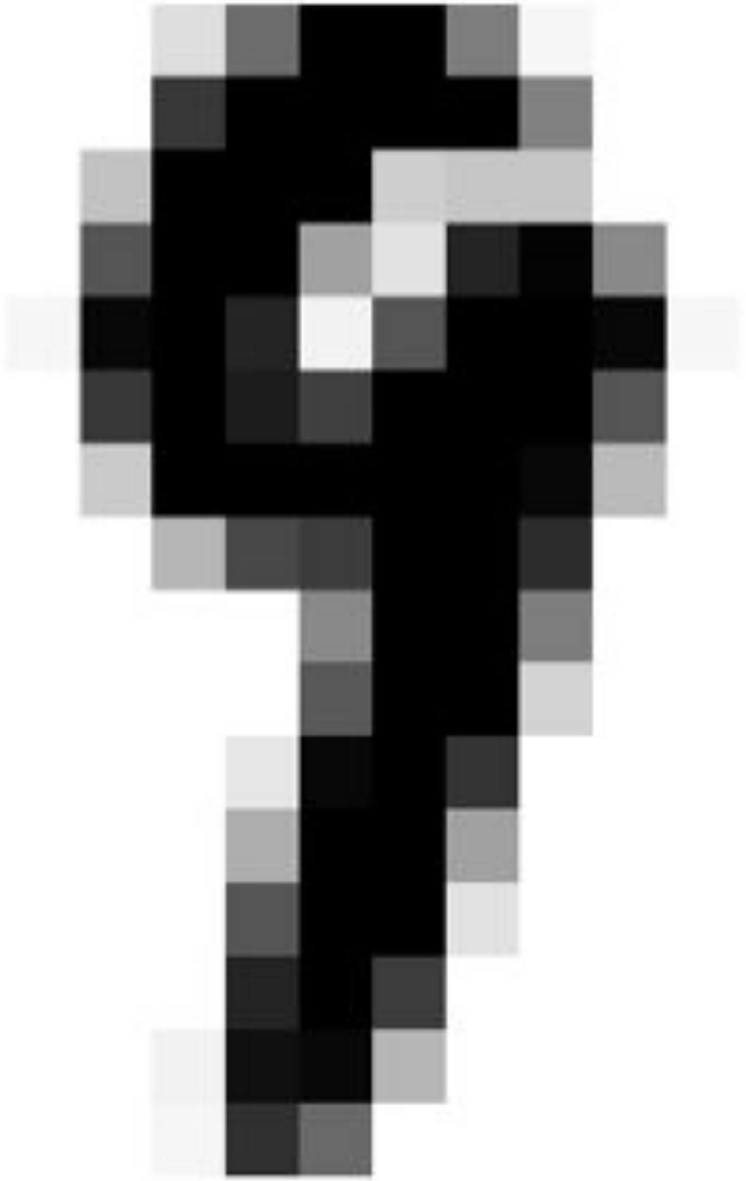
8



9

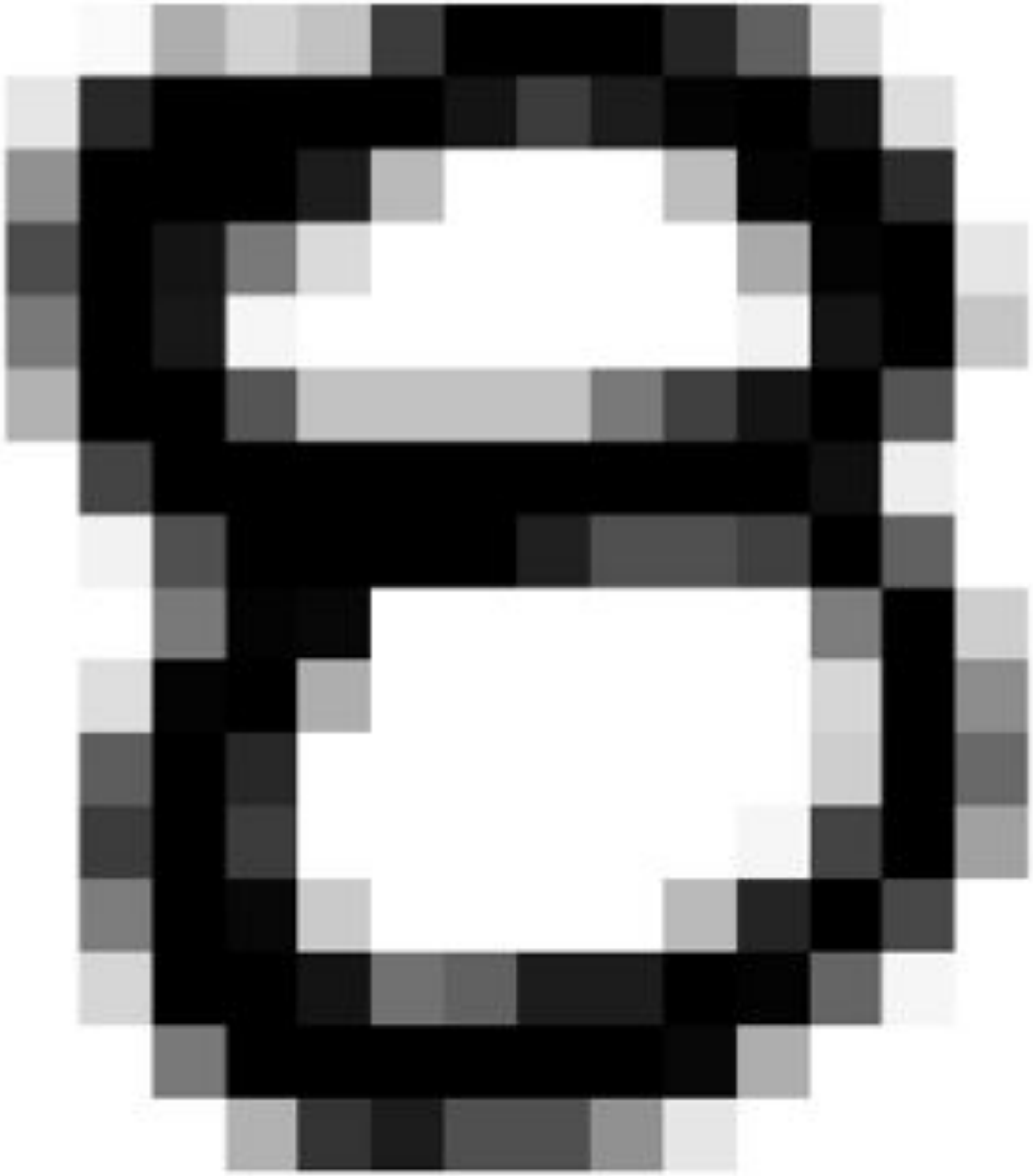
8

8



9

8



8

9

9

9

8

9

8

8

9

8

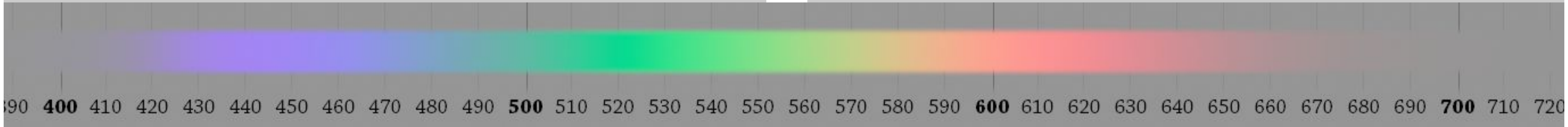
8

9

Empirical Inference, II

- Example 2: perception

"The brain is nothing but a statistical decision organ"
(H. Barlow)



x

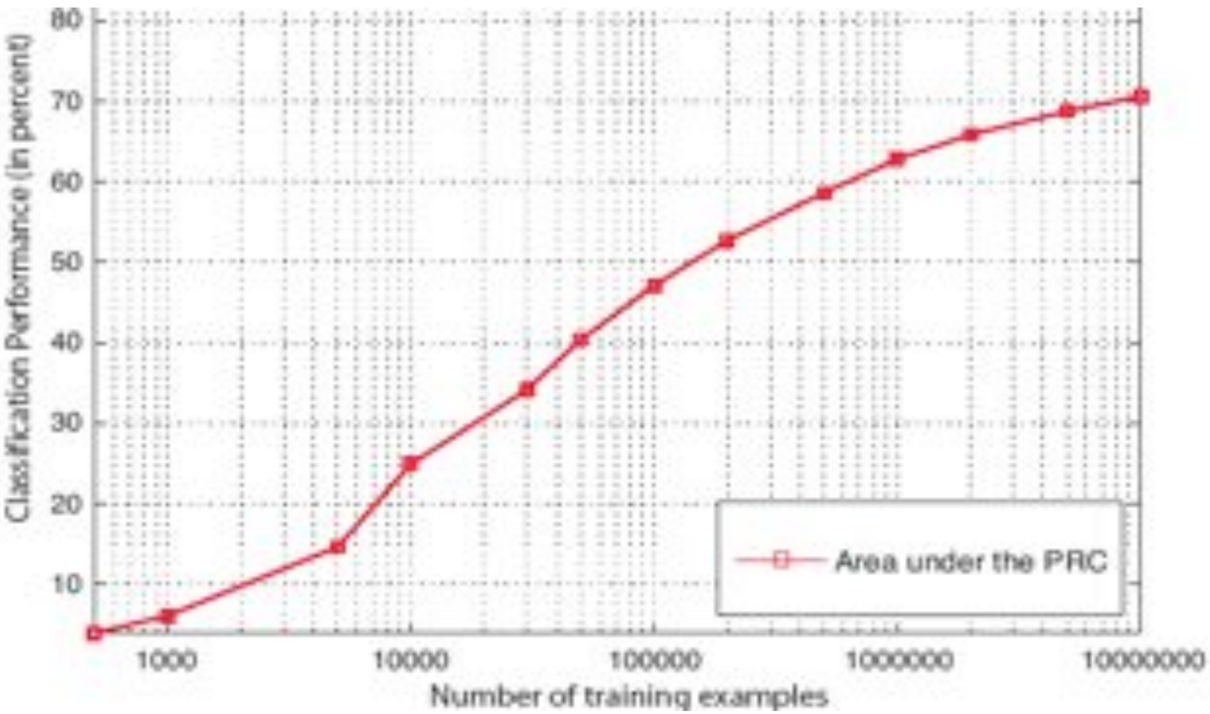
- Vision as unconscious inference *(Helmholtz)*

The elephants seem to be of different size

FIFA World Cup, Germany vs. England, June 27, 2010

The ball appears to be in the goal.

Hard Inference Problems



Sonnenburg, Rätsch, Schäfer, Schölkopf, 2006, Journal of Machine Learning Research

Task: classify human DNA sequence locations into {acceptor splice site, decoy} using 15 Million sequences of length 141, and a Multiple-Kernel Support Vector Machines.

PRC = Precision-Recall-Curve, fraction of correct positive predictions among all positively predicted cases

- High dimensionality – *consider many factors simultaneously to find the regularity*
- Complex regularities – *nonlinear, nonstationary, etc.*
- Little prior knowledge – *e.g., no mechanistic models for the data*
- Need large data sets – *processing requires computers and automatic inference methods*

Generalization

- *observe* 1, 2, 4, 7,..
- What's next?

+1 +2 +3

- 1,2,4,7,11,16,…: $a_{n+1}=a_n+n$ ("lazy caterer's sequence")
- 1,2,4,7,12,20,…: $a_{n+2}=a_{n+1}+a_n+1$
- 1,2,4,7,13,24,…: "Tribonacci"-sequence
- 1,2,4,7,14,28: divisors of 28
- 1,2,4,7,1,1,5,…: decimal expansions of π=3,14159… and e=2,718… interleaved *(thanks to O. Bousquet)*
- **The On-Line Encyclopedia of Integer Sequences**: >600 hits…

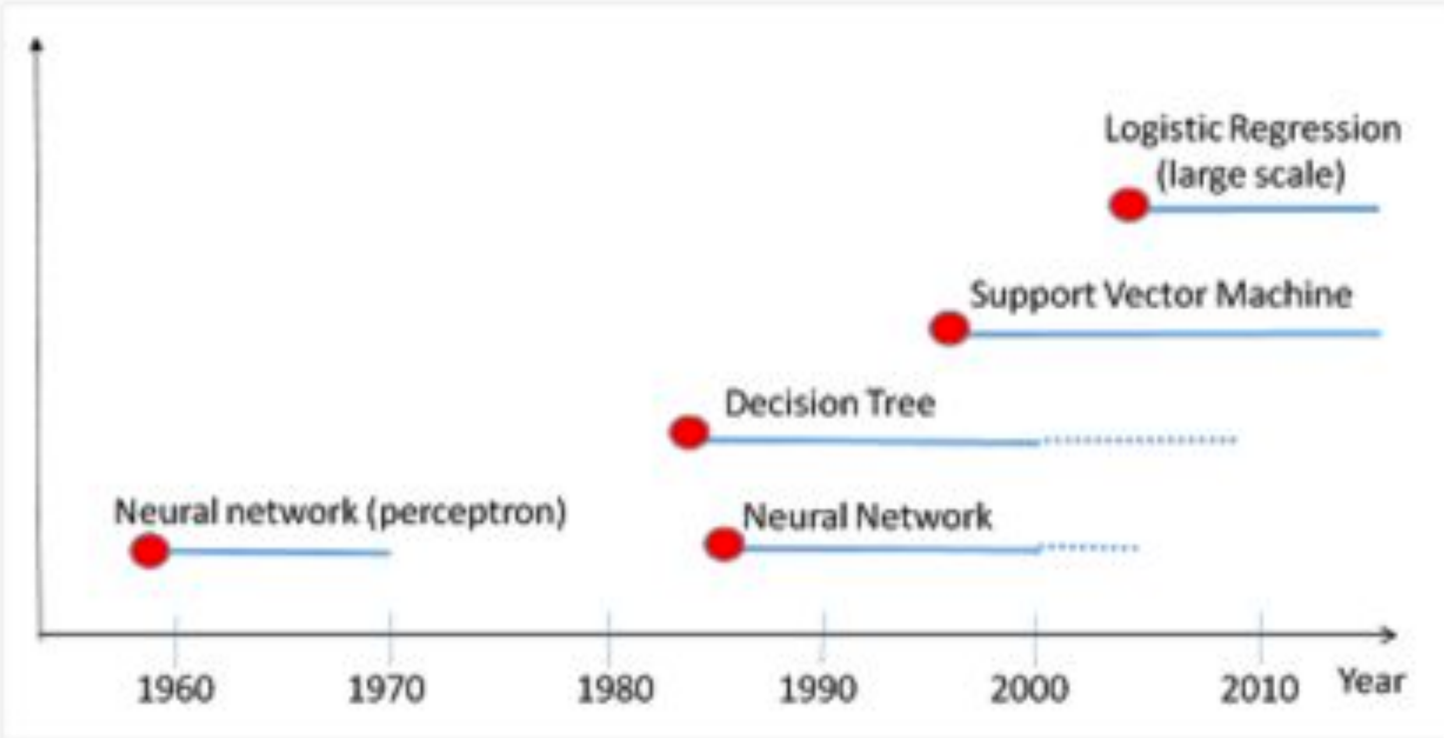
Generalization, II

- Question: which continuation is correct ("generalizes")?
- *Answer:* there's no way to tell *("induction problem")*

- Question of statistical learning theory: how to come up with a law that generalizes *("demarcation problem")*

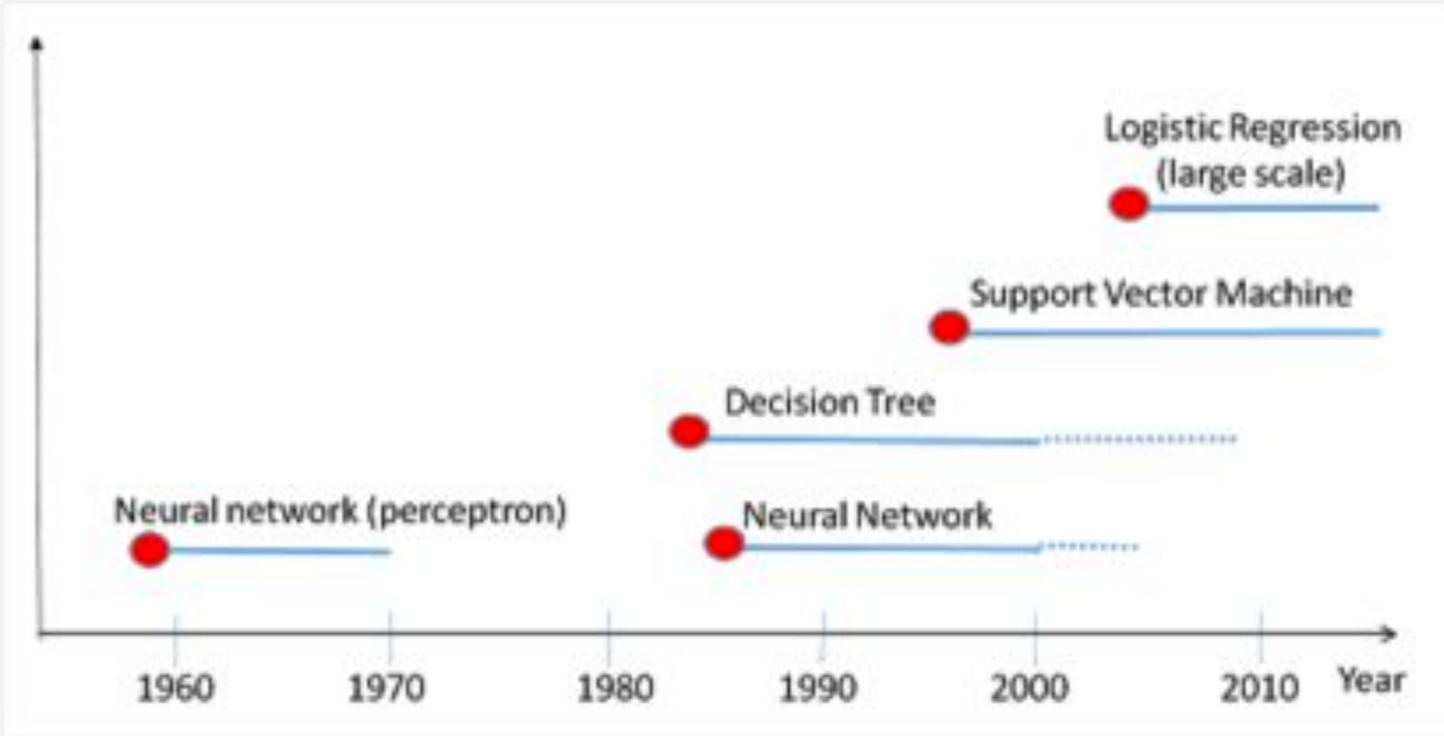
  [i.e.: a law that will probably do almost as well in the future as it has done in the past]

History of Machine Learning



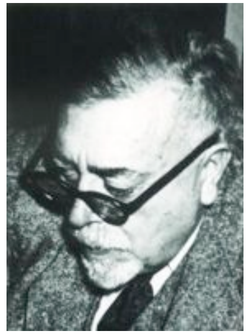
Source: “About data mining: History of Machine Learning”
www.aboutdm.com

History of Machine Learning, II



- History is highly subjective -data mining point of view
- ML / Neural Net / Pattern Recognition point of view
  (sources, images: *Olazaran, 1996, Pias, 2006*)

http://owpdb.mfo.de/detail?photo_id=4520



Cybernetics – 1940s/50s

- Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine* (1948)
- study of control and information processing (rather than energy) in animals and machines
- Macy Conferences 1946-53: “*Circular Causal and Feedback Mechanisms in Biological and Social Systems”*. Birth of cybernetics and cognitive science
- John von Neumann, Alan Turing, Claude Shannon

Tagungsplanung Notizzettel von Warren McCulloch zur Konferenz von 1953

1953, April 22nd, 23rd and 24th / J. Macy. Found's
Plaza 7-7705

+ Bateson — Telegram
+ Bavelas
+ Bigelow → Von Karman – Probably –
+ Brosin
— Lorente not coming
+ L.K. Frank
+ Gerard — Part.
+ Hutchinson
+ Kluver
+ Kubie
+ McCulloch
No Mead — Part. ← + Margins 4 an
+ Northrop
+ Pitts
No Rosenblueth – ? Telegram
+ Savage
+ Fremont-Smith ← Smith
+ Teuber
+ von Bonin
+ von Foerster

+ Bowman
+ Quastler, Dr. Prof. Henry Q. Assoc. Prof. of Physiology Univ. of Illinois 252 Engineering hall Urbana, Illinois
+ Wiesner

+ Bar-Hillel
+ Carnap Instit. for Advanced Study

No Turing Dr. Alan Turing Dept. of Applied Math Univ. of Manchester Manchester 13, England

No Piaget Faculte des Sciences Universite de Geneve Geneve, Switzerland

Alternates
+ Shannon ▶ Bell Tel. Dr. Claude Shannon Bell Tel. Murray Hill, N. J.
Dr. Nathan ~~Rochester~~ IBM Poughkeepsie
Dr. Y. R. Chao Univ. of Cal. Berkeley
+

Murray Hill
Bell Labs –
Summit 66000

Hollymeade
Adlington Rd
Wilmslow
Cheshire

Dear McCulloch,

It was very gratifying and tempting to get your invitation to the Macy meeting. You have certainly got a wonderful collection of people together. If it were in Europe I should certainly try to make it, but I am really rather a stay-at-home type. Unfortunately also it is during term time, and I am doubtful if I could get permission to be away.

Yours sincerely

A. M. Turing


Dr. med. Georg Bayr
Kybernetik
und
homöopathische
Medizin
Haug Verlag


H. R. RAPP
Mensch
Gott
und Zahl
Kybernetik im Horizont
der Theologie
FURCHE

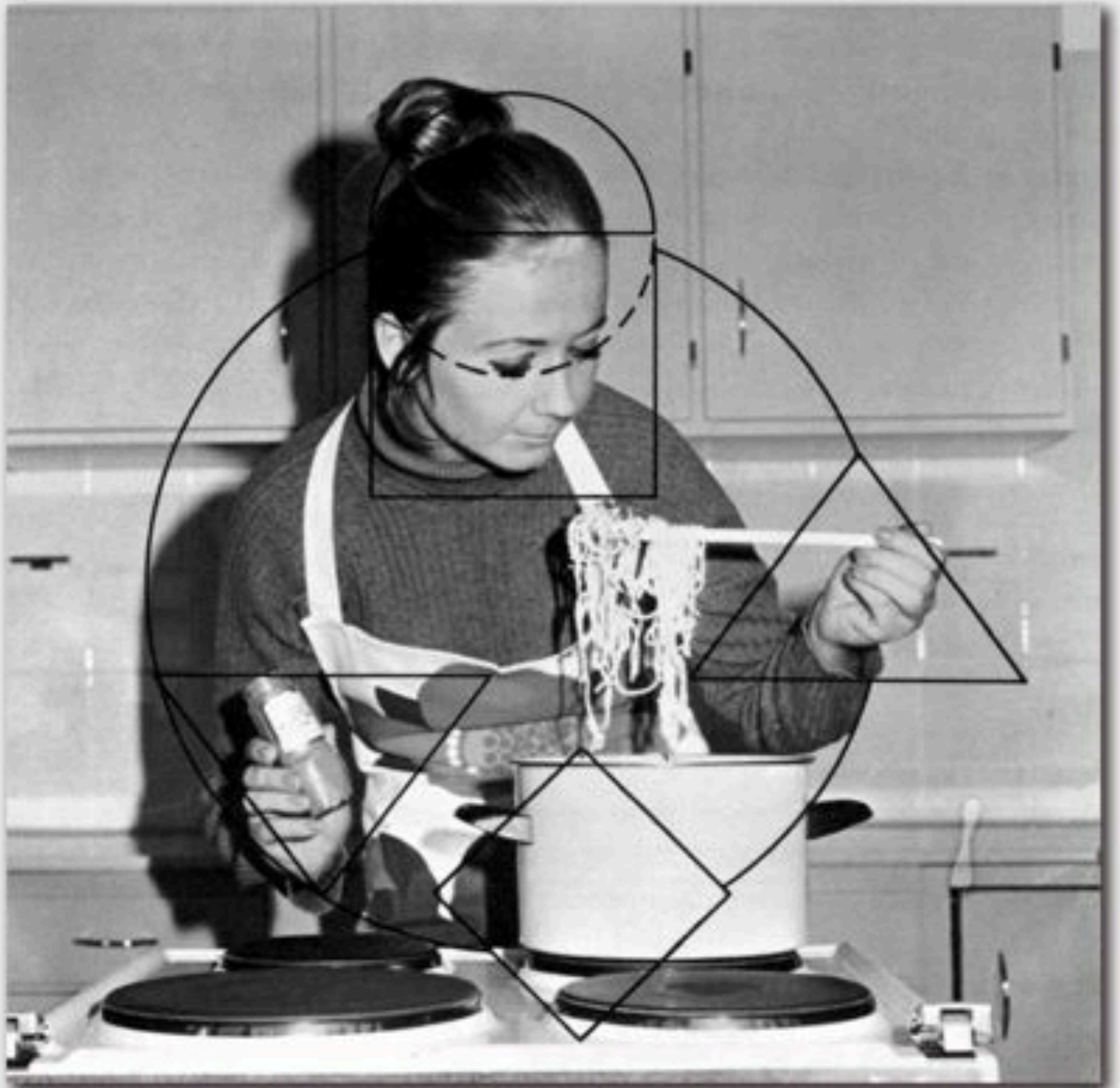
4.

Was die moderne Frau
von der Kybernetik wissen
muß

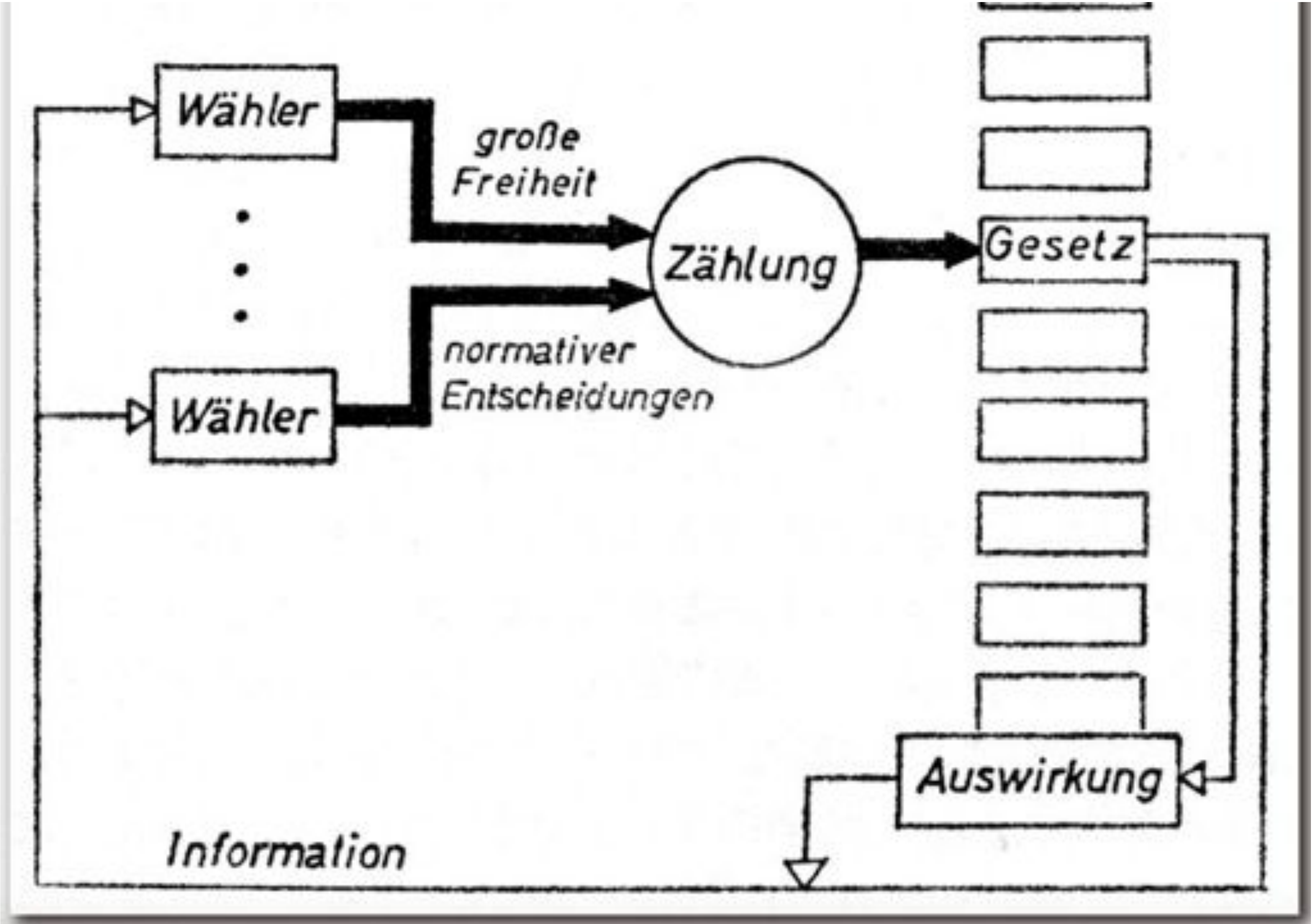
Rudi
Wetzel

Die Geheimnisse des Rechenautomaten

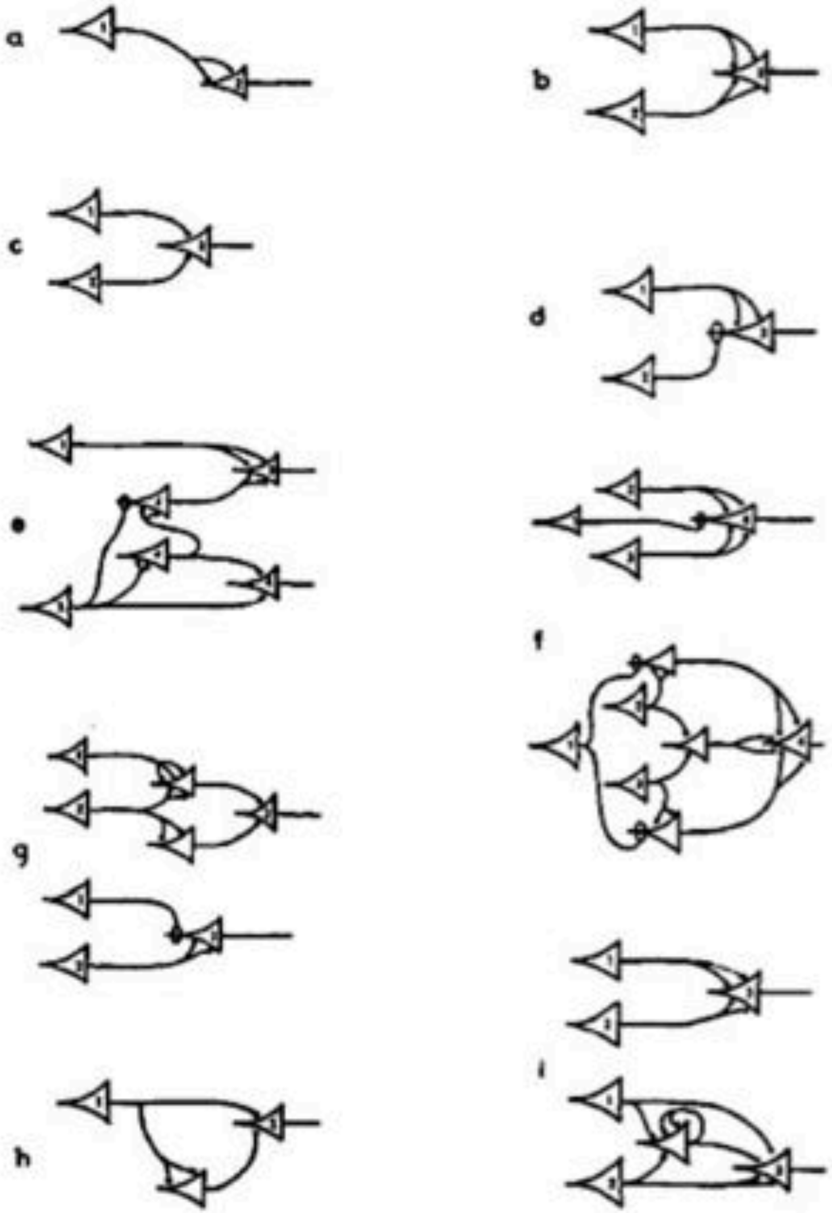
Diagrammatische Evidenzen kybernetischer Regelkreise – Kochen (1970)



Bernhard Schölkopf


Wähler
große
Freiheit
Zählung
Gesetz
normativer
Entscheidungen
Wähler
Auswirkung
Information

- Project Cybersyn at Allende Government (1971-73)
- Stafford Beer



Neurophysiologische Entsprechungen

a. Präszisierung
b. Disjunktion
c. Konjunktion
d. verknüpfte Negation
e. ?
f. relative Inhibition
g. (oben) Löschung
(unten) absolute Inhibition
h. zeitliche Summation
j. Regeneration (Lernen)

- Neural Nets (1950s)
- McCulloch-Pitts "formal Neurons", networks can emulate Universal Turing machine
- Hebb. Connectionism
- Ashby's "Homeostat"

REFERENCES

1. Ashby, W. Ross, Design for a brain Wiley & Sons, New York, 1954.

2. McCulloch, W. S., and Pitts, W., "A logical calculus of the ideas immanent in nervous activity" Bulletin of Math. Biophysics 5 (1943) pp. 115-133.

3. Rosenblatt, Frank, The Perceptron - A theory of statistical separability in cognitive systems Cornell Aeronautical Laboratory, Inc. Report No. VG-1196-G-1, January 1958.

4. Von Neumann, John, The computer and the brain Yale University Press, New Haven, 1958.

MAX-PLANCK-GESELLSCHAFT

Rosenblatt's Perceptron (1957)

(**8**), and Minsky (**13**). A relatively small number of theorists, like Ashby (**1**) and von Neumann (**17, 18**), have been concerned with the problems of how an imperfect neural network, containing many random connections, can be made to perform reliably those functions which might be represented by idealized wiring diagrams. Unfortunately, the language of symbolic logic and Boolean algebra is less well suited for such investigations. The need for a suitable language for the mathematical analysis of events in systems where only the gross organization can be characterized, and the precise structure is unknown, has led the author to formulate the current model in terms of probability theory rather than symbolic logic.

Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN [1]

F. ROSENBLATT
Cornell Aeronautical Laboratory

Rosenblatt's (1958) [perceptron] schemes quickly took root, and soon there were perhaps as many as a hundred groups, large and small, experimenting with the model either as a 'learning machine' or in the guise of 'adaptive' or 'self-organizing' networks or 'automatic control' systems.[13]

(Minsky & Papert, 1969)

Other groups included:

- Bernard Widrow (Stanford)
- Charles Rosen (Stanford Research Institute, SRI)

FIGURE 2
Perceptron

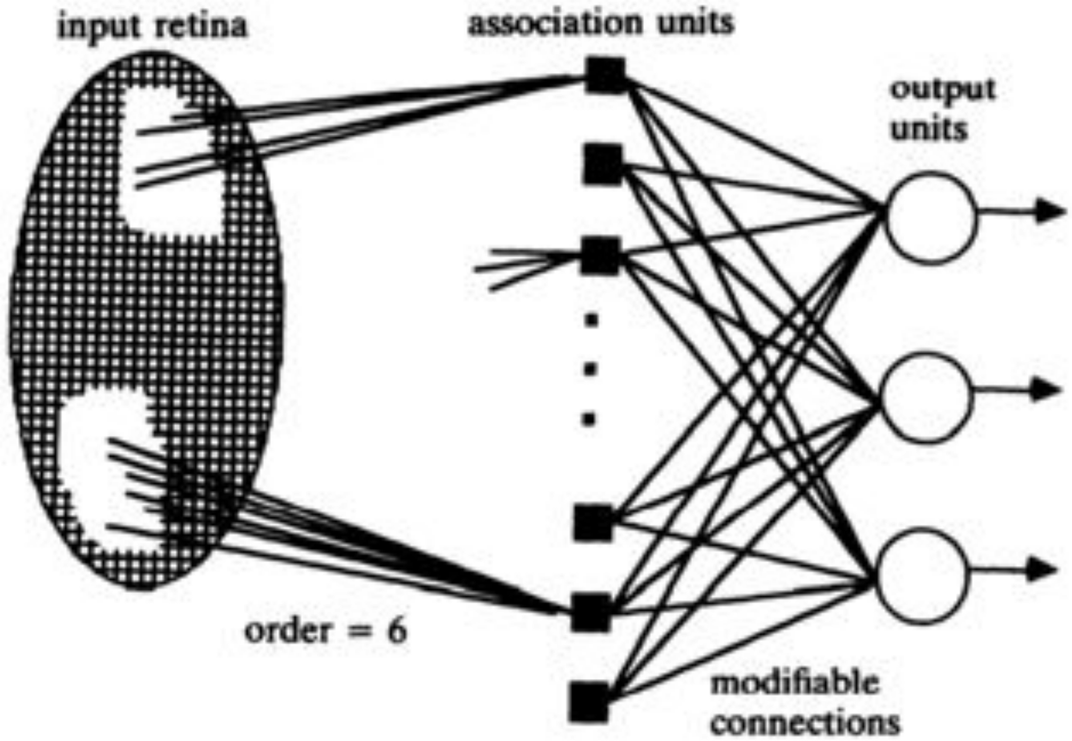
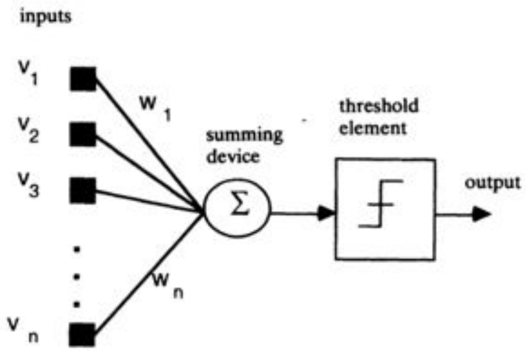


FIGURE 3
Processing Unit



- first layer: fixed weights
- second layer: adjustable weights (Perceptron learning rule)

Percepton Convergence Theorem *(Novikoff, 1962)*

Theorem 11.1: Perceptron Convergence Theorem: *Let* **F** *be a set of unit-length vectors. If there exists a unit vector* $\mathbf{A}^*$ *and a number* $\delta > 0$ *such that* $\mathbf{A}^* \cdot \mathbf{\Phi} > \delta$ *for all* $\mathbf{\Phi}$ *in* **F**, *then the program*

START: Set **A** to an arbitrary $\mathbf{\Phi}$ of **F**.

TEST: Choose an arbitrary $\mathbf{\Phi}$ of **F**, and
if $\mathbf{A} \cdot \mathbf{\Phi} > 0$ go to TEST;
otherwise go to ADD.

ADD: Replace **A** by $\mathbf{A} + \mathbf{\Phi}$.
Go to TEST.

will go to ADD *only a finite number of times.*

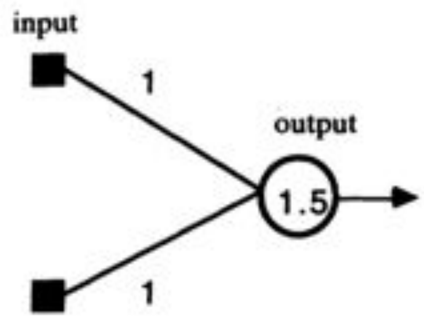
Some readers might be amused to note that the proof of this theorem does not use any assumptions of finiteness of the set **F** or the dimension of the vector space. This will not be true of later sections where the compactness of the unit sphere plays an apparently essential role.

from Minsky & Papert (1969)

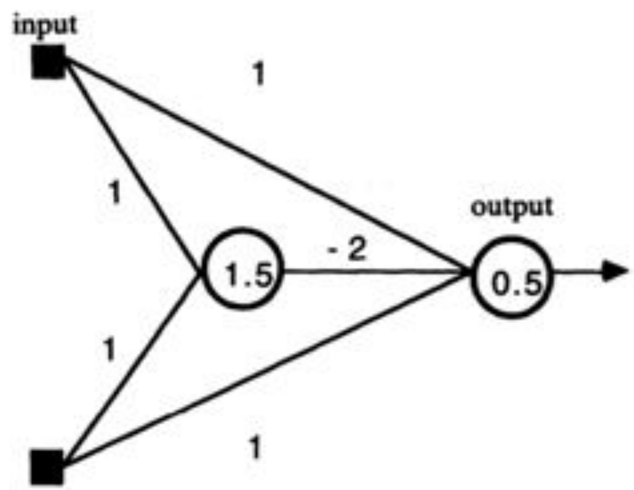
A particular task that could **not** be learnt *(Hawkins, 1961)*

- ‘and’ can be done with one layer, ‘xor’ requires a cascade. However, no training algorithm for this exists.

5.1 ‘And’ Function



5.2 ‘Exclusive Or’ Function



Perceptron limitations recognized by Rosenblatt

- excessive learning time
- figure-ground separation
- recognition of topological relationships and abstract concepts
- training multi-layer systems
- technology/size

The models which conceive of the brain as a strictly digital, Boolean algebra device, always involve either an impossibly large number of discrete elements, or else a precision of the 'wiring diagram' and synchronization of the system which is quite unlike the conditions observed in a biological nervous system.[43]

- NYT on a 1958 Press conference (Rosenblatt & ONR):

> The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence. Later perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech and writing in another language, it was predicted.[17]

- Hecht-Nielssen (1990)

> The campaign was waged by means of personal persuasion by Minsky and Papert and their allies, as well as by limited circulation of an unpublished technical manuscript (which was later de-venomized and, after further refinement and expansion, published in 1969 as the book *Perceptrons*).[20]

The “XOR Affair”

- Minsky & Papert (1969): *Perceptrons*

Perceptrons have been widely publicized as ‘‘pattern recognition’’ or ‘‘learning’’ machines and as such have been discussed in a large number of books, journal articles, and voluminous ‘‘reports.’’ Most of this writing ... is without scientific value. (p. 4)

[We] became involved with a somewhat therapeutic compulsion: to dispel what we feared to be the first shadows of a ‘‘holistic’’ or ‘‘Gestalt’’ misconception that would threaten to haunt the fields of engineering and artificial intelligence... (p. 20)

There is no reason to suppose that any of these virtues carry over to the many layered version. Nevertheless, we consider it to be an important research problem to elucidate (or reject) our intuitive judgement that the extension is sterile. (p. 231)

MAX-PLANCK-GESELLSCHAFT

cited after Pollack’s book review of the new edition of “Perceptrons” (1988)

- Minsky & Papert recall (1988/89):

In the late 1950s and early 1960s, after Rosenblatt's work, there was a great wave of neural network research activity. There were maybe thousands of projects. For example Stanford Research Institute had a good project. But nothing happened. The machines were very limited. So I would say by 1965 people were getting worried. They were trying to get money to build bigger machines, but they didn't seem to be going anywhere. That's when Papert and I

There was *some* hostility in the energy behind the research reported in *Perceptrons*. . . . Part of our drive came, as we quite plainly acknowledged in our book, from the fact that funding and research energy were being dissipated on . . . misleading attempts to use connectionist methods in practical applications.[46]

Technical content of *Perceptrons*

- assume 1 output neuron
- restrict the "order" of the association units
- parity cannot be solved unless the order equals the whole retina
- similar for figure/ground (connectedness)
- both can easily be solved using serial algorithms

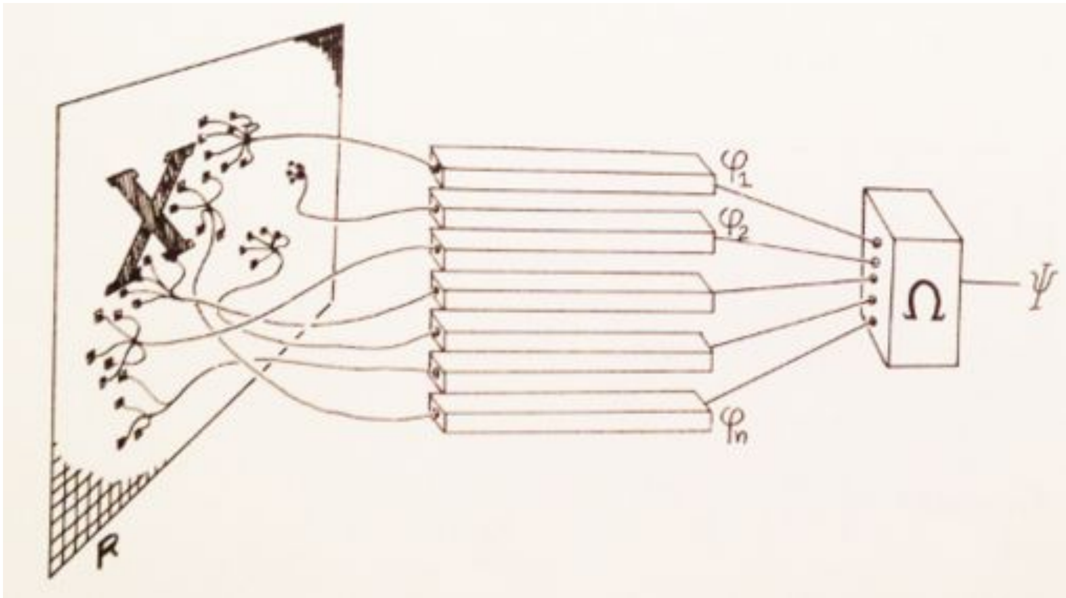
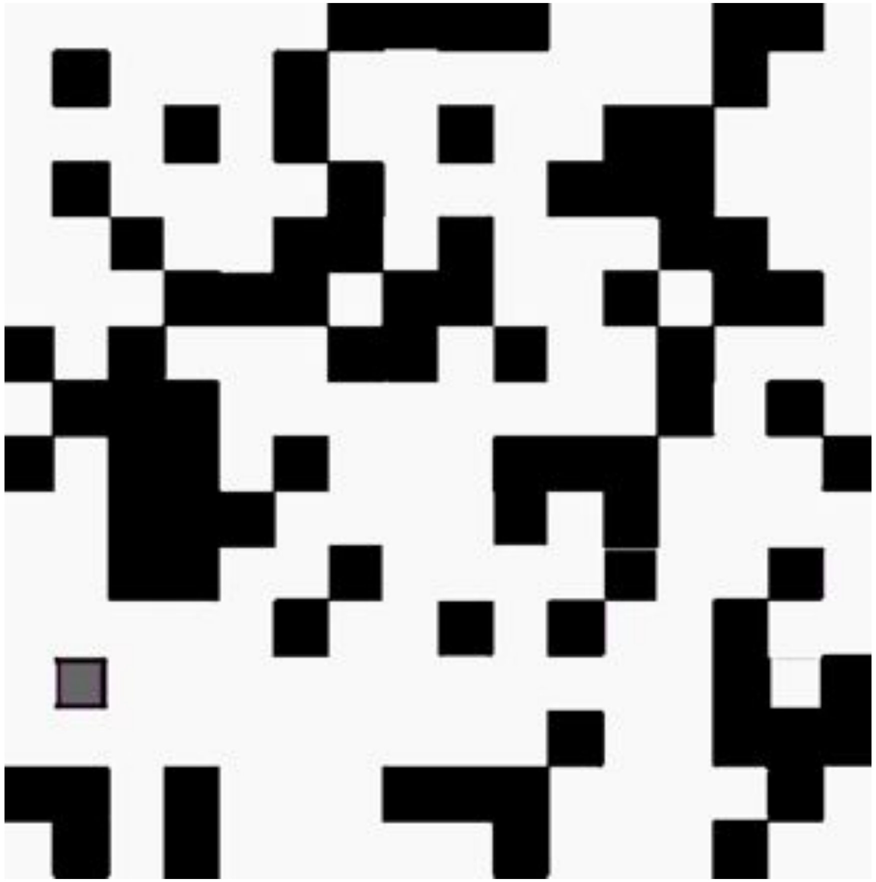


Figure: Minsky & Papert (1969)

Importance of parity and figure-ground



Importance of parity and figure-ground

FIGURE 7

Reception of “Perceptrons”

Another indication of this difference of perspective is Minsky and Papert’s concern with such predicates as *parity* and *connectedness*. Human beings cannot perceive the parity of large sets (is the number of dots in a newspaper photograph *even* or *odd*?), nor connectedness (on the cover of Minsky and Papert’s book are two patterns; one is connected, one is not. It is virtually impossible to determine by visual examination which is which). Rosenblatt

H. D. Block: A Review of “Perceptrons”. 1970

This is a great book. To understand this judgment, and why I am willing to make it at so early a date, is not so simple. For the book is many things,

A. Newell: *A step toward the understanding of information processes*. **Science**, 1969

I should remark, perhaps, that I am not an unbiased witness, although I trust I have kept my wits about me in examining the book. For I share with Minsky and Papert a common view of the appropriate shaping of computer science into a disciplined field of inquiry. And I see no need to give other than my true assessment of the potential role of this book in that shaping.

ALLEN NEWELL

Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213

Did `Perceptrons' kill Perceptrons?

When I first saw the book, years and years ago, I came to the conclusion that they had defined the idea of a perceptron sufficiently narrowly so that they could prove that it couldn't do anything. I thought that the book was relevant, in the sense that it was good mathematics. It was good that somebody did that,

Widrow (1989)

Extensions used by Rosenblatt & others

- two layers of association units
- feedback connections within layers

The end of perceptrons *Olazaran (1996)*

- the importance of the `arithmetic ideal' in science
- the competing paradigm was gaining momentum:
  - digital computers became available (1950s)
  - development of high-level programming languages (some of them developed by AI people, e.g. IPL and LISP)
  - early successes of symbolic AI: General Problem Solver, Logic Theorist, STUDENT (Minsky: *"STUDENT...understands English"*), Chess systems
  - of the major groups, only Rosenblatt continued, but died in 1971
  - ARPA decided to back symbolic AI and cut off neural nets
  - The defeat of neural nets helped legitimitize symbolic AI:

> **The principal body of evidence for the symbolic hypothesis that we have not considered [so far in this paper] is negative evidence: the absence of specific competing hypotheses as to how intelligent activity might be accomplished — whether by man or by machine.[78]**

(Newell & Simon, 1976)

Symbolic AI

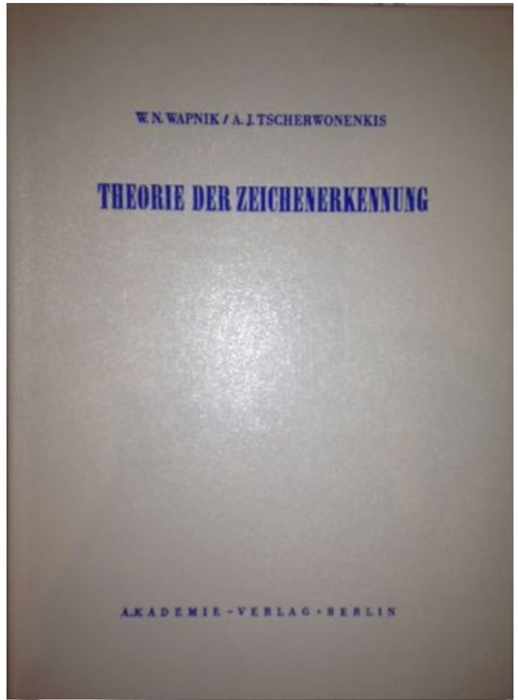
- Symbolic AI (Dartmouth Summer School, 1956): intelligence is a process of manipulating discrete symbols; *John McCarthy, Allen Newell, Herb Simon, Marvin Minsky*
- helped the transformation of "computers" into symbol processing systems

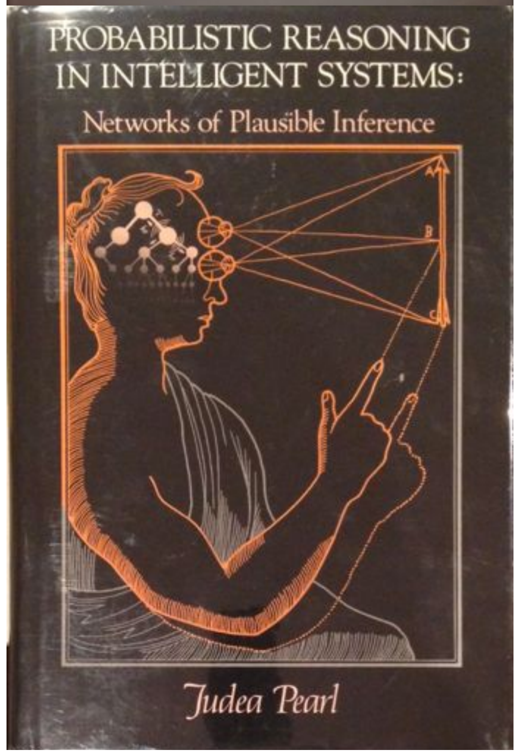
- *Herb Simon*, following the success of his program General Problem Solver (1957), predicted that within 10 years,
  - A computer would be world champion in chess.
  - A computer would discover and prove an important new mathematical theorem.
  - A computer would write music of considerable aesthetic value
  - Most theories in psychology will take the form of computer programs.
- Hubert Dreyfus dismissed this in *Alchemy and AI (1965)*

- Herb Simon won a Turing award (1975, with Newell) and a Nobel Prize (1978).

Machine Learning in Exile

- neural nets were almost extinct, very few continued
  - Kohonen, Hinton, Amari, Grossberg…
- Statistical Learning Theory
  - *Vapnik & Chervonenkis (ca. 1968-1982)*
- Expert Systems / knowledge representation were made probabilistic
  - *Judea Pearl (1988)*
  - *this gave birth to Bayesian nets*





The Return of Neural Nets

- symbolic AI was doing well at chess, but failed miserably at speech and vision
- computing becomes a commodity

- the PDP group was formed by **psychologists** *Rumelhart & McClelland*
- the field attracted other **physicists**, e.g. *John Hopfield (1982)* (Ising model)
- Boltzmann machine *(Hinton, Sejnowski)*

The Return of Neural Nets, II

- Back-propagation, mid-1980s *(Rumelhart, Hinton, Williams, LeCun, Werbos, Amari)*
- *Minsky & Papert (1988, Perceptrons, 2nd ed.):*

We have the impression that many people in the connectionist community do not understand that this [back-propagation] is merely a particular way to compute a gradient and have assumed instead that back-propagation is a new learning scheme that somehow gets around the basic limitations of hill-climbing.

situation. . . . We fear that its [back-propagation's] reputation also stems from unfamiliarity with the manner in which hill-climbing methods deteriorate when confronted with larger-scale problems. In any case, little good can come from statements like 'as a practical matter, GD leads to solutions in virtually every case' or 'GD can, in principle, learn arbitrary functions'. Such pronouncements are not merely technically wrong; more significantly, the pretense that problems do not exist can deflect us from valuable insights that could come from

Probability, Statistics, and Machine Learning

- Laplace. Introduced Bayes' Theorem / inverse probability in the general form and applied it to celestial mechanics.
- Gauss.
- Solomonoff (1950s): probabilistic AI
- MCMC (1980s)
- PAC (1984)
- first UAI (1985)
- first NIPS (1987)
- Probabilistic foundations for ML (1990s) – *MacKay, Neal, Jordan, Hinton, Bishop,…*
- SVMs (1990) – *Vapnik et al.*

Generalized Portrait and Kernel Methods

- Vapnik proposed the ‘generalized portrait algorithm’
- p.d. kernels first used by *Hilbert (1904)*
- *Grace Wahba (since 1970)*
- *Duda & Hart (1973):* “The familiar functions of mathemati
eigenfunctions of symmetric kernels, and their use is often suggested for the construction of potential functions. However, these suggestions are more appealing for their mathematical beauty than their practical usefulness.”
- used to prove convergence of the potential function method *(Aizerman, Braverman, & Rozonoer, 1964)*
- Generalized Portrait method *(Vapnik & Chervonenkis, 1974)*

MAX-PLANCK-GESELLSCHAFT

• **used in Optimal Margin Classifiers** ***(Boser, Guyon & Vapnik)*****, Soft Margin Classifiers / Support Vector Networks** ***(Cortes & Vapnik***

Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

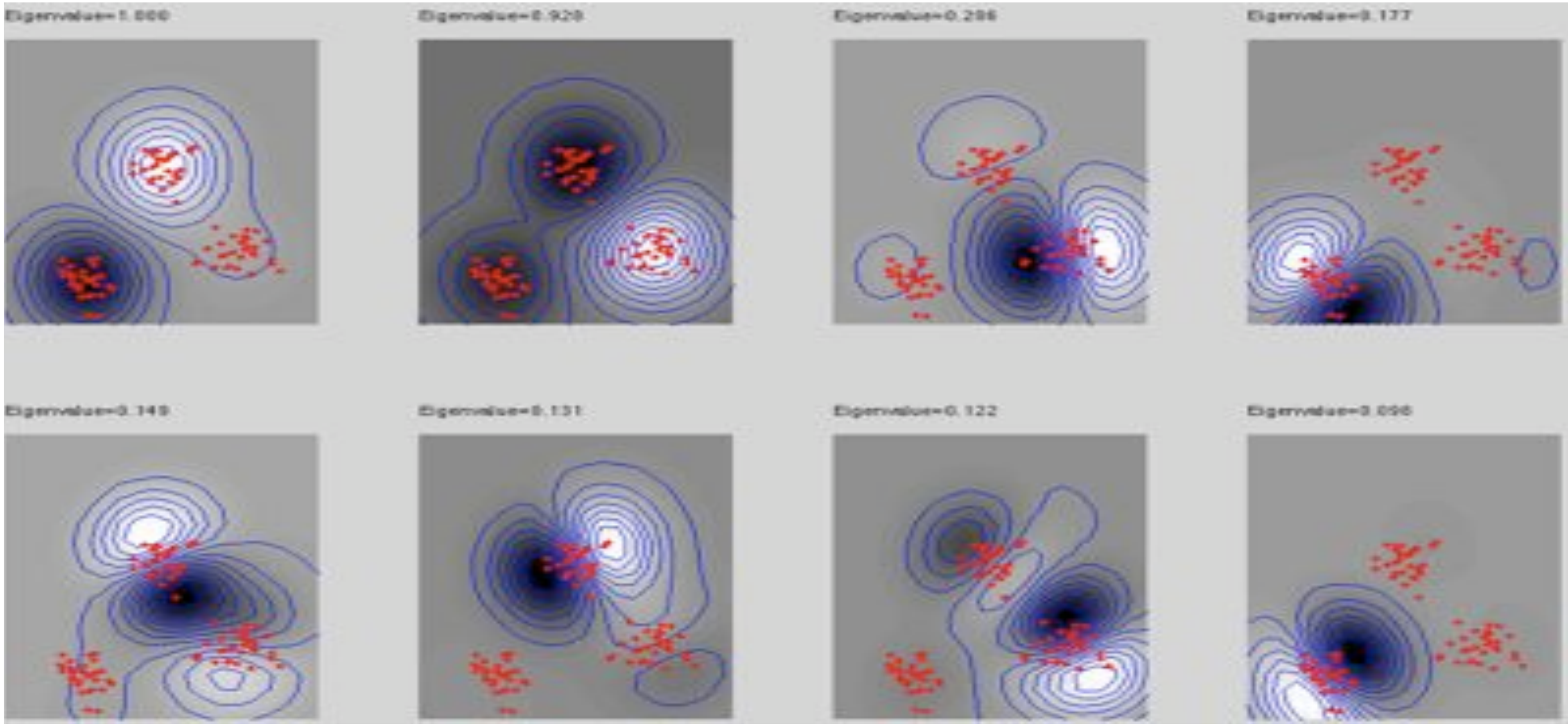
CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

Editor: Lorenza Saitta

Abstract. The *support-vector network* is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data.

High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.

• kernelization works for arbitrary dot product algorithms, e.g. KPCA *(Schölkopf, Smola & Müller, 1997; Burges 1998)* --- ***"kernel trick"***



- kernelization does not require vectorial data *(Schölkopf, 1997)*

Conclusion

- Where is Machine Learning heading today?
  - technical issues: optimization, structured data, efficient learning and inference, sparsity, …
  - integration of/with domain knowledge
  - learning control
  - learning in physical (synthetic or hybrid) systems
  - learning in environments populated by agents
  - learning in nonstationary settings; causal learning
- What is machine learning?
  - “not statistics”
  - a young discipline
  - the only understood organizing principle of intelligent systems