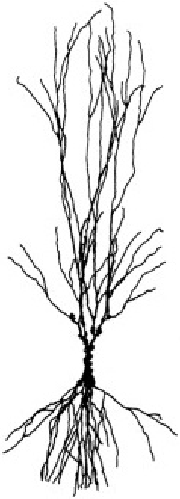
Neural networks, 1900 – 1990

Yuxi Liu

UC Berkeley

2024-11-21

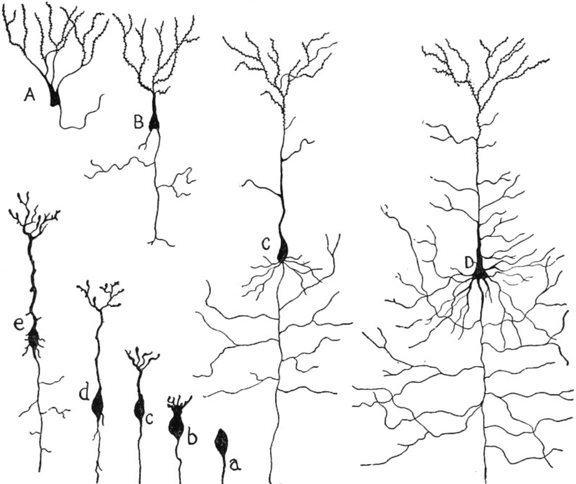
The thunderbolt steers all things.
— Heraclitus



Cajal (1900s)

Santiago Ramón y Cajal (1852–1934)

- Spanish neuroanatomist.
- Discovered neurons: the neural network is not just a network.



Cajal (1900s)

All neural networks are feedforward...

Cajal (1900s)

All neural networks are feedforward... except

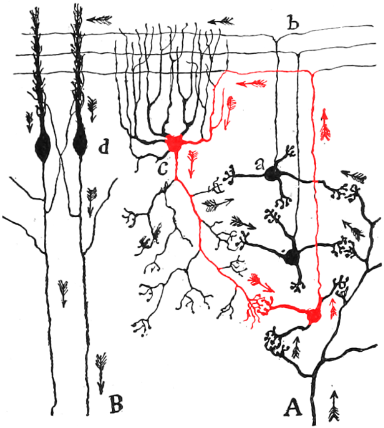
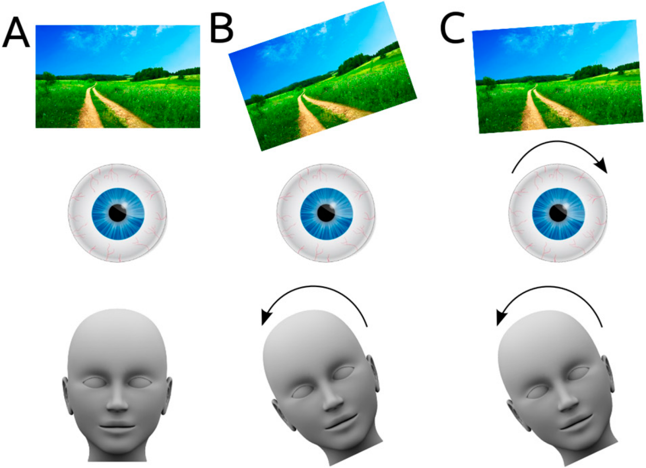


Figure: First observation of recurrent network (in the cerebellum). [Ramón y Cajal, 1909, figure 103, page 149]

Vestibulo-Ocular Reflex (VOR)



de No (1930s)

Rafael Lorente de Nó (1902–1990)

- Student of Cajal
- Discovered RNN and ResNet in the VOR.
- Cajal: For the sake of your career, don't publish about RNN, because nobody would believe you!...

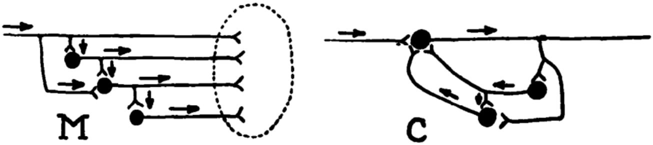


Figure: RNN and ResNet in the VOR. [de Nó, 1938]

de No (1930s)

Rafael Lorente de Nó (1902–1990)

- Student of Cajal
- Discovered RNN and ResNet in the VOR.
- Cajal: For the sake of your career, don't publish about RNN, because nobody would believe you!... so he published it after Cajal died (1934).

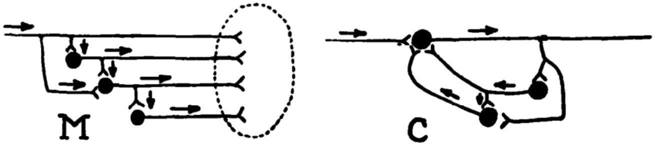


Figure: RNN and ResNet in the VOR. [de Nó, 1938]

Hebb (1949)

- Donald O. Hebb (1904–1985)
- Hebbian learning: "fire together, wire together"
- "Reverberation" in sub-networks

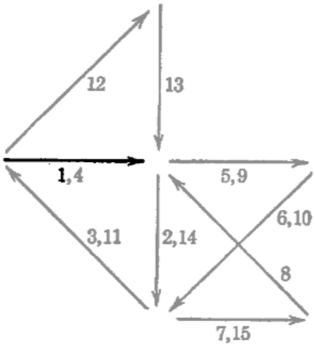
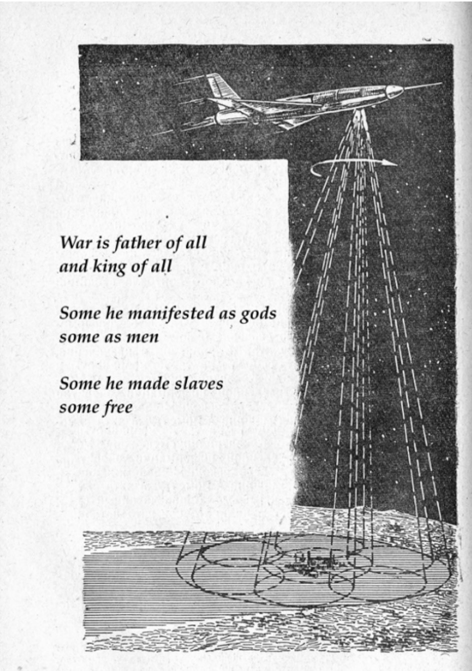


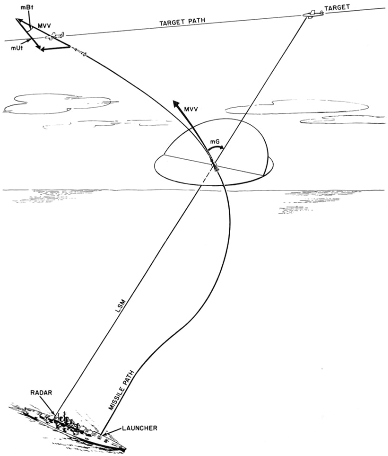
Figure: [Hebb, 2002]


War is father of all
and king of all

Some he manifested as gods
some as men

Some he made slaves
some free

Fire control



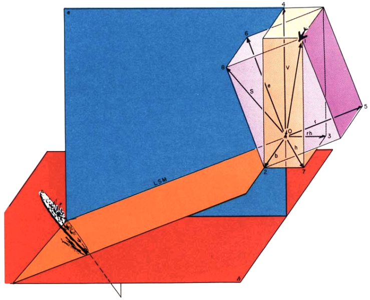


Figure 10—Relative Missile Displacement About Line of Sight to Missile in Stable Coordinates.

Table 21

	Relative Guidance	Missile	Target	To Aiming Position	Capture
Displacement at capture or thereafter	0–1 **(mM)o**	**(mMm)o**	**(mMt)o**	**mM4**	**mM6**
Displacement perpendicular to vertical plane through LSM	0–2 **mMb**	**mMbm**	**mMbt**	**mMb4**	**mMb6**
Displacement in horizontal in vertical plane through LSM	0–3 **(mMrh)o**	**(mMrhm)o**	**(mMrht)o**	**mMrh4**	**mMrh6**
Displacement in vertical in vertical plane through LSM	0–4 **(mMv)o**	**(mMvm)o**	**(mMvt)o**	**mMv4**	**mMv6**
Displacement along LSM	0–5 **(mMr)**	**(mMrm)**	**(mMrt)**	**mMr4**	**mMr6**

Figure: Fire control as linear regression

Fire control

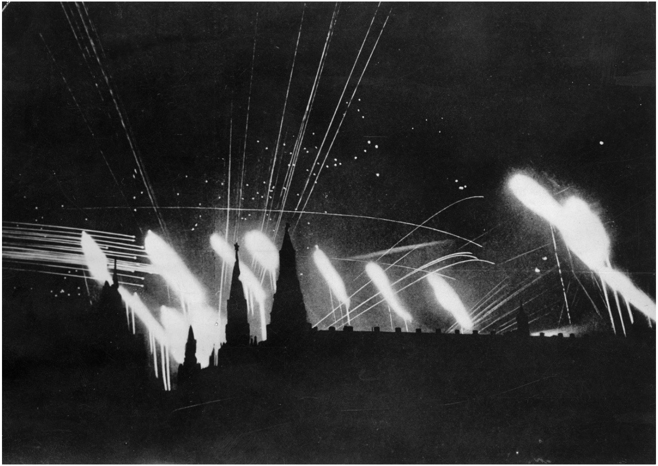


Figure: Fire control as linear regression

Fire control

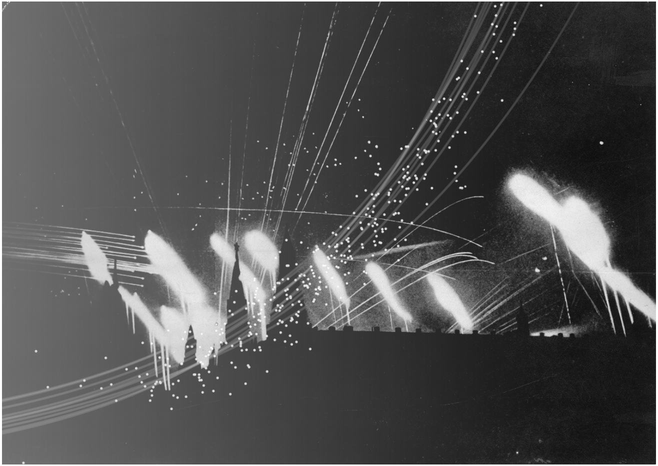


Figure: Fire control as kernelized linear regression

Norbert Wiener (1941)

- Antiaircraft fire as negative feedback control
- $$\text{Loss} = \int \|\vec{x}_{\text{aircraft}}(t) - \vec{x}_{\text{bullet}}(t)\|^2 dt$$
- Minimize loss by kernelized least squares.
- Kernel estimated statistically from pilot data and gunner data.

  It is necessary to assimilate the different parts of the system to a single basis, either human or mechanical. Since our understanding of the mechanical elements of gun pointing appeared to us to be far ahead of our psychological understanding, we chose to try to find a mechanical analogue of the gun pointer and the airplane pilot.
  — [Wiener, 2017, page 407]

Norbert Wiener (1941)

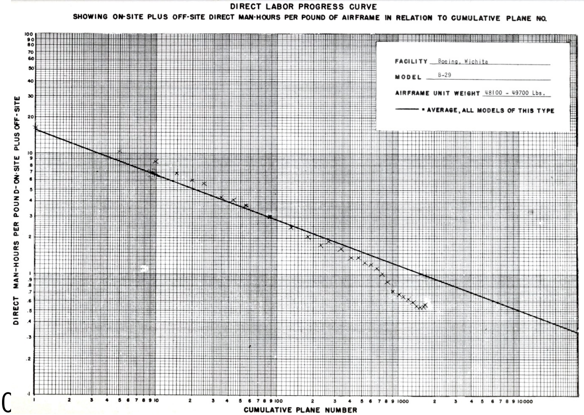
- Aircraft prediction as noisy communication
- Past trajectory of the aircraft is a noisy channel for the future trajectory.
- Minimize mean-squared error by Wiener filtering
- Communication = Control

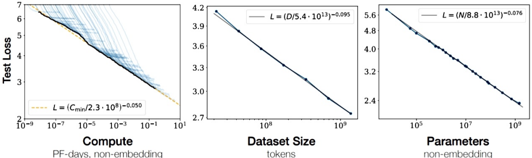
  This book represents an attempt to unite the theory and practice of two fields of work which are of vital importance in the present emergency... of time series in statistics and of communication engineering.

  — [Norbert Wiener, 1966, page 1]

Bonus: scaling law

- Scaling curve of B-29 at the Boeing Wichita.
- Scaling curves for LLM. [Kaplan et al., 2020]





Bonus: piecewise linear regression

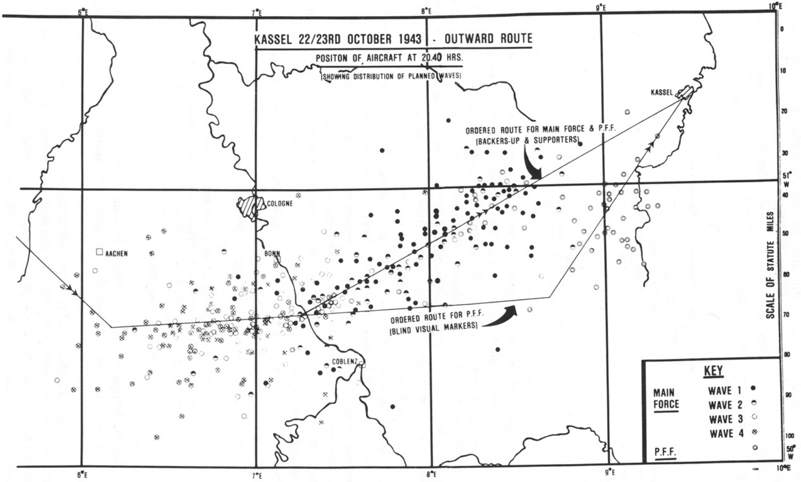


Figure: Kassel bombing raid.

Bonus: Fire control VOR

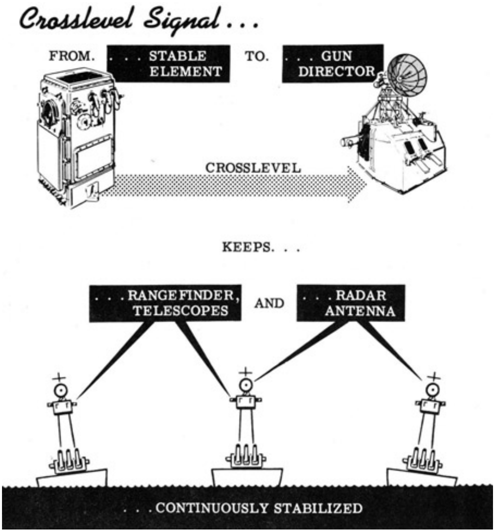


Figure: Fire control with vestibulo-ocular reflex

McCulloch & Pitts (1943)

- Warren McCulloch (1898–1969): neuro-philosopher.
- Walter Pitts (1923–1969): tragic logician.
- Met in 1941. Started working on a mathematical theory of thinking.
- Inspired by cybernetics, causal loops, memory, learning, neural logic, hallucination, epilepsy...
- Went to the Macy conferences. (de No was there too)
- [McCulloch and Pitts, 1943]: Proposed artificial neural network.

  What is a number, that a man may know it, and a man, that he may know a number?
  — *Warren McCulloch*

McCulloch & Pitts (1943)

- Universality: Any boolean function is computed by some feedforward network.
- Feedforward network with Hebbian learning is equivalent to a recurrent network without Hebbian learning.

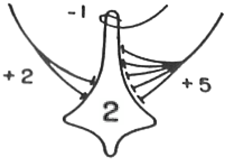


Figure: A single MP neuron.

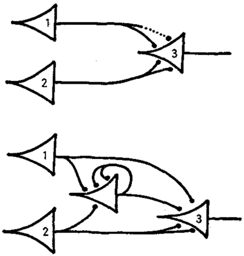


Figure: Hebbian learning is equivalent to recurrence.

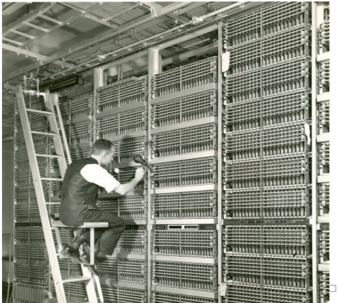
Why was causal loop unacceptable?

I have no firm answers, but I have some ideas...

- Reflexes are simpler to study for neuroscientists.
- Feedback is much harder to solve mathematically.
- Loops threaten linear causality: cause $\rightarrow$ effect.
- Dominance of personalities: Cajal, Pavlov, Skinner...

Comment: Progress is possible

- Early 20th century, the brain was often analogized to a telephone switchboard
- Now, analogized to a neural network
- Some commentators sneer that our analogy is nothing but techno-fashion
- But telephone switchboards are feedforward, while neural network have feedback.
- Lesson: There is progress!



1950s neural networks

Multiple neural networks in the early 1950s. Most notable is Minsky's SNARC (1952): maze-running rat with 40 neurons, trained by reinforcement Hebbian learning. Every time the operator presses a reward button, the synaptic strengths change by Hebbian learning.

Figure: Shannon and his robot mouse.

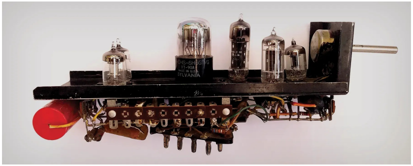
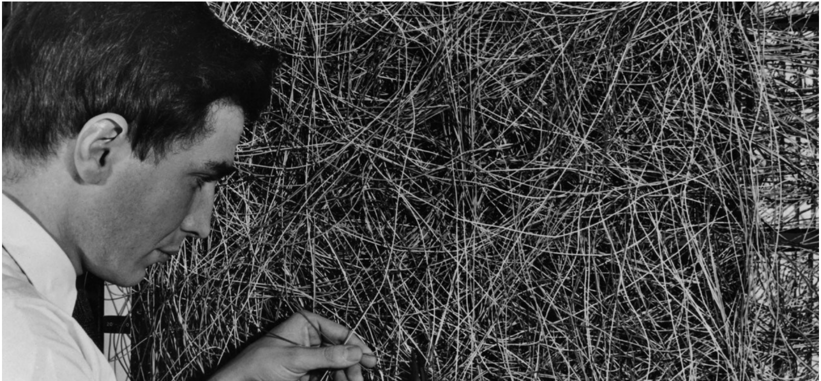


Figure: A SNARC neuron.

Perceptron

- Frank Rosenblatt (1928–1971): Did most of his work during 1957–1964.
- [Rosenblatt, 1958] and [Rosenblatt, 1962]: multilayered perceptrons.
- A perceptron is a binary neuron: $\theta(\sum_i w_i x_i + b)$, where θ is the 0-1 activation function.



Example

- First layer fixed.
- Second layer has both residual connection and learned connection. The learned weights increase by Hebbian learning, and decay exponentially.
- Third layer by the perceptron learning rule.

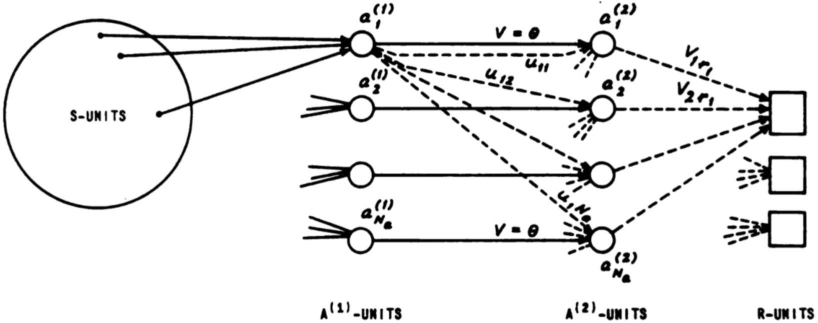


Figure: A multilayered perceptron. [Rosenblatt, 1962, page 347]

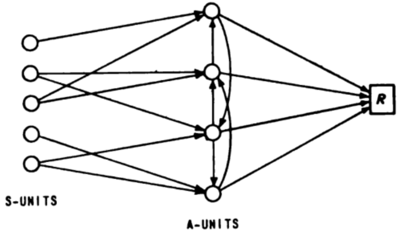


Figure: Hopfield network? (Figure 47)

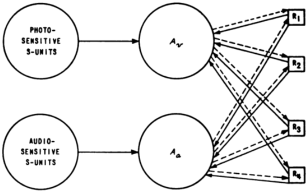


Figure 58 ORGANIZATION OF A DUAL MODALITY PERCEPTRON, WITH 4 R-UNITS (BROKEN LINES INDICATE VARIABLE-VALUED CONNECTIONS)

Figure: Multimodality. (Figure 58)

Tobermory

- Speech recognition. Named after a talking cat.
- 2 hidden layers, 12,000 weights. Takes up an entire room.
- Built during 1961–1967. When completed, already slower than simulation on IBM machines.

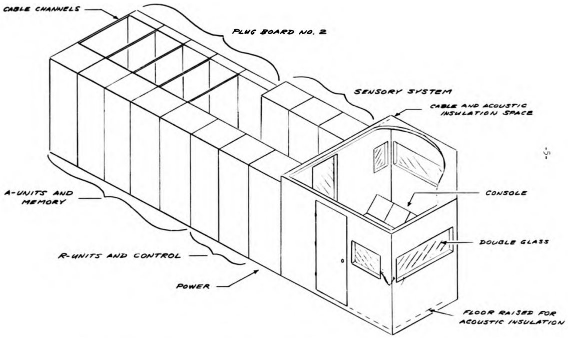


Figure: Schematic for Tobermory.

MINOS

- MINOS project (1958–1967), at Stanford Research Institute (SRI).
- Character recognition for military applications.
- MINOS II (1963): analog image $\to$ 100 binary hidden neurons by 100 optical masks, then a linear layer to 63 binary outputs by perceptron learning rule.

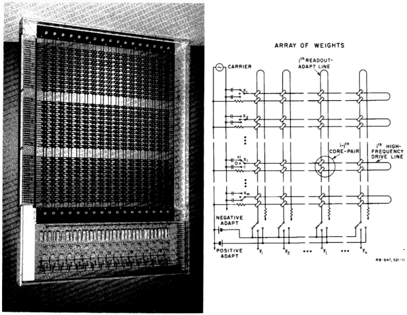


Figure: Weight matrix for MINOS II (1962). World's first TPU?

MINOS

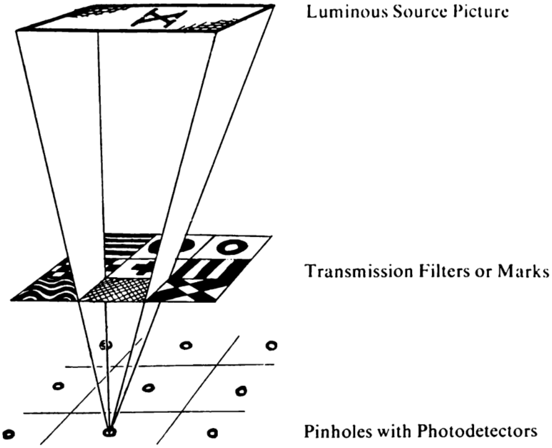


Figure: Using a list of hand-designed photomasks to featurize an image into a list of binary bits.

ADALINE

ADALINE (adaptive linear), single binary perceptron, trained by gradient descent on squared error $(\sum_i w_i x_i + b - y)^2$. Developed by Bernard Widrow's group.

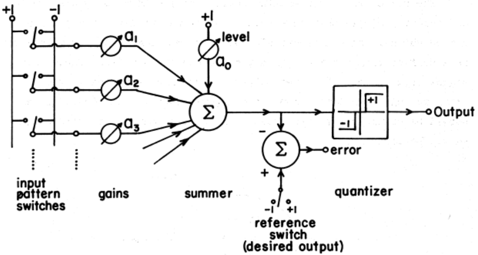


Figure: Circuit diagram of ADALINE.

ADALINE

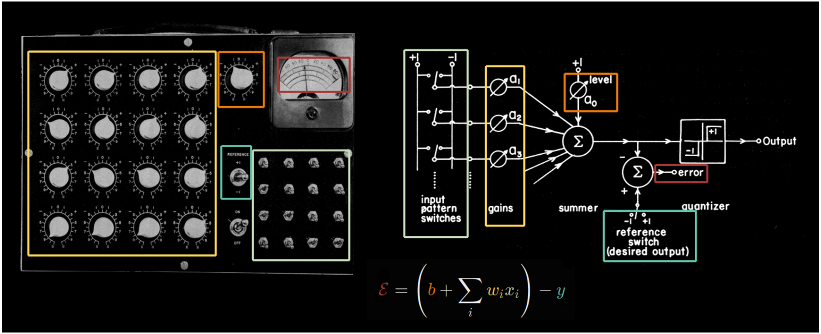


Figure: 16 input switches and 1 output switch. Manually operated.

ADALINE

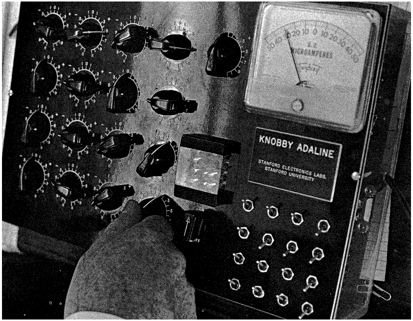


Figure: Widrow performing brain surgery on his first-born.

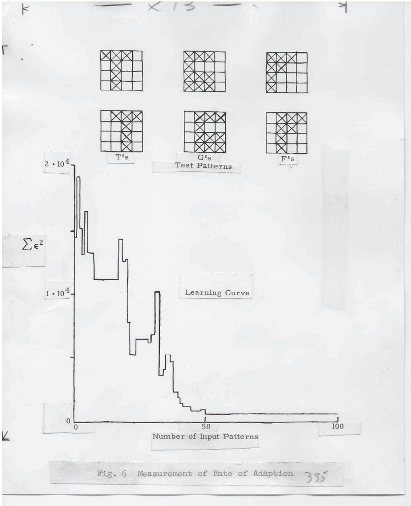


Figure: Input and learning curve.

MADALINE

They also tried MADALINE (many ADALINE), with many hacky rules, but never gradient descent (because of the binary activation).

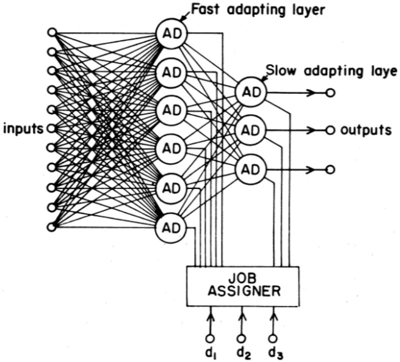
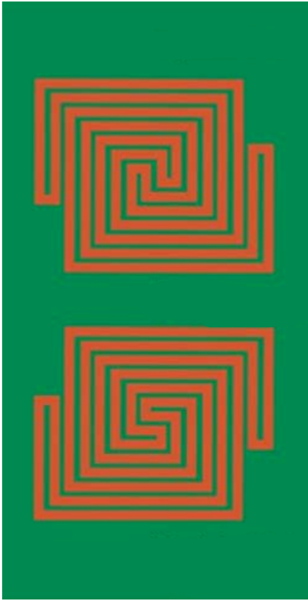


Figure 11. A Two-Layer Network of Adaptive ADALINES.

Figure: MADALINE from 1962. The "JOB ASSIGNER" learning rule is too hacky to explain. [Widrow, 1962]

It would seem that "Perceptrons" has much the same role as The Necronomicon — that is, often cited but never read.

— Marvin Minsky



XOR myth

The myth of XOR: Neural networks were abandoned after *Perceptrons* (Minsky and Papert, 1969), which showed that a single-layered perceptron cannot learn the XOR function. It was revived in the 1980s after the development of multilayered perceptrons.

XOR myth

The myth of XOR: Neural networks were abandoned after *Perceptrons* (Minsky and Papert, 1969), which showed that a single-layered perceptron cannot learn the XOR function. It was revived in the 1980s after the development of multilayered perceptrons.

But...

- *Obviously* XOR is impossible for single-layered perceptrons.
- McCulloch and Pitts (1943) already showed MLP as Turing-complete.
- Electric engineers were hand-designing MLP ("linear threshold logic") as Boolean logic modules.
- Rosenblatt, Widrow, SRI were all training MLP in the 1960s.

What is the real story?

MLP fail

Everyone failed to train MLP.
Rosenblatt had "back-propagating errors", continuous activation, and multiple trainable layers, but never gradient descent.
The other two groups did not fair any better.

> *We never succeeded in developing an algorithm for adapting both layers... Backprop to me is almost miraculous.*
> *— Bernard Widrow, 1994 interview. [Rosenfeld and Anderson, 2000, page 60–61]*

> *I recall Charles Rosen, the leader of our group, sitting in his office with yellow quadrille tablets hand-simulating his ever-inventive schemes for making weight changes; none seemed to work out.*
> *— Nils Nilsson (of SRI) [Nilsson, 2009, section 29.4]*

Minsky & Papert (1960s)

- Met in 1963. Started collaborating to dunk on large neural networks.
- Circulated their working notes at conferences.
- Also, on-stage debates. "Many remember as great spectator sport the quarrels Minsky and Rosenblatt had".
- Published their (in)famous book *Perceptrons* in 1969.

Why Minsky?

Minsky: General algorithms never scale, only a Society of Mind Works

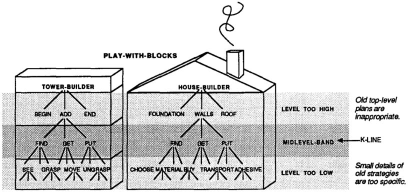


Figure: Society of Mind [Minsky, 1988, section 8.6]

I had the naïve idea that if one could build a big enough network, with enough memory loops, it might get lucky ... either this was a bad idea or it would take thousands or millions of neurons to make it work, and I couldn't afford to try to build a machine like that.

— *[Bernstein, 1981]*

Why Papert?

A new computer culture ... would require a new social construction of the computer, with a new set of intellectual and emotional values more like those applied to harpsichords than hammers. ... Epistemological pluralism is a necessary condition for a more inclusive computer culture.

— *[Turkle and Papert, 1990]*

Papert: General algorithms are unjust and exclusionary to other ways of knowing. **Con**struction, not **In**struction

So what really caused the winter?

- Minsky and Papert's book: "our book put a stop to those" [Bernstein, 1981].
- Funding turned away from neural nets to the more symbolic methods.
- Lack of people
  - McCulloch (1969), Pitts (1969), Rosenblatt (1971) died
  - Widrow gave up MADALINE; focussed on ADALINE applications
  - Ted Hoff went to Intel to invent microprocessor
  - SRI turned to logical AI
- No backprop.

Why no backprop?

- Bad luck.
- Fear of local optima.
- Everyone "knew" that real neurons are binary, so artificial neurons must also be binary.
- They thought of machine learning as boolean function synthesis.
- ... but I don't know for sure. See my post for details.

`https://yuxi-liu-wired.github.io/essays/posts/backstory-of-backpropagation/`

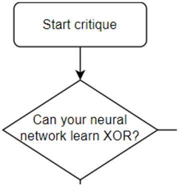
Why the XOR myth?

Why do people keep teaching the XOR myth? My guess...

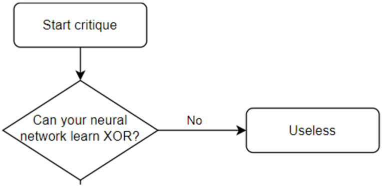
- Professors need to explain what happened during the 1970s in an introductory AI course.
- The actual XOR problem requires one to already understand perceptrons.
- The XOR problem is a convenient excuse.
- Thus – the XOR myth!

What is the actual XOR problem?

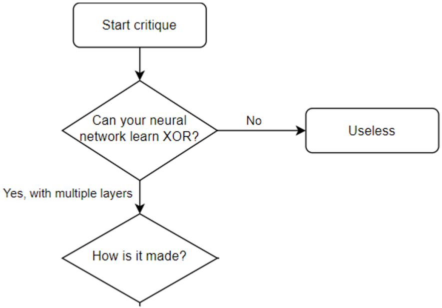
The XOR problem is real, despite the myth. See this flowchart...



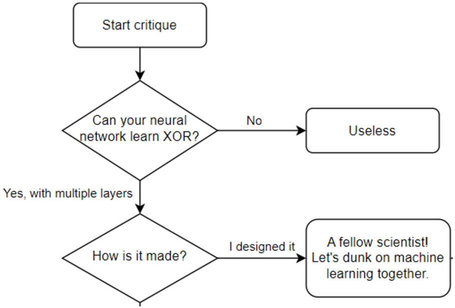
What is the actual XOR problem?



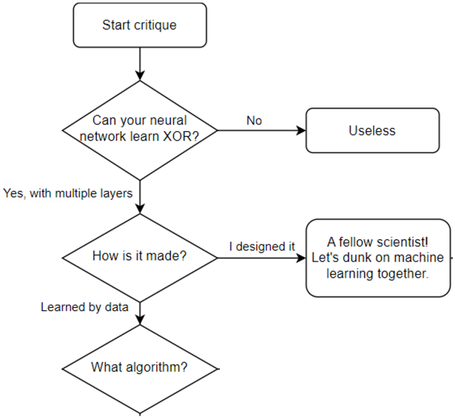
What is the actual XOR problem?



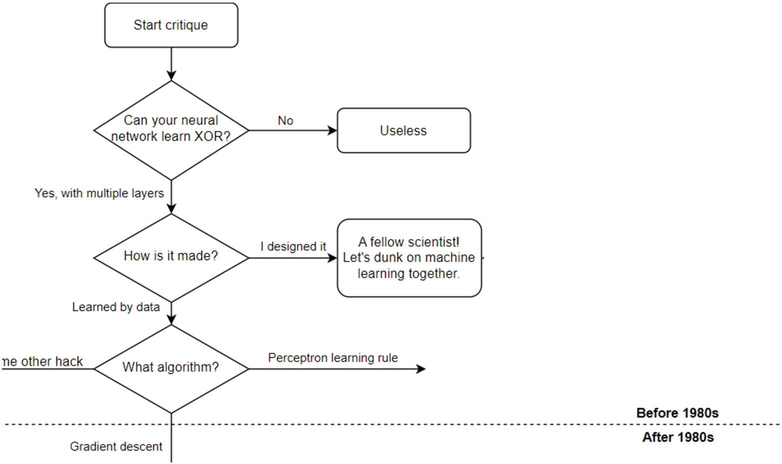
What is the actual XOR problem?



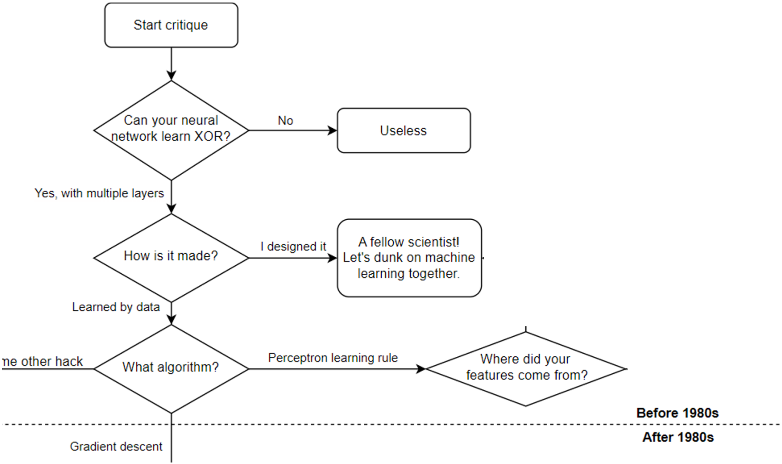
What is the actual XOR problem?



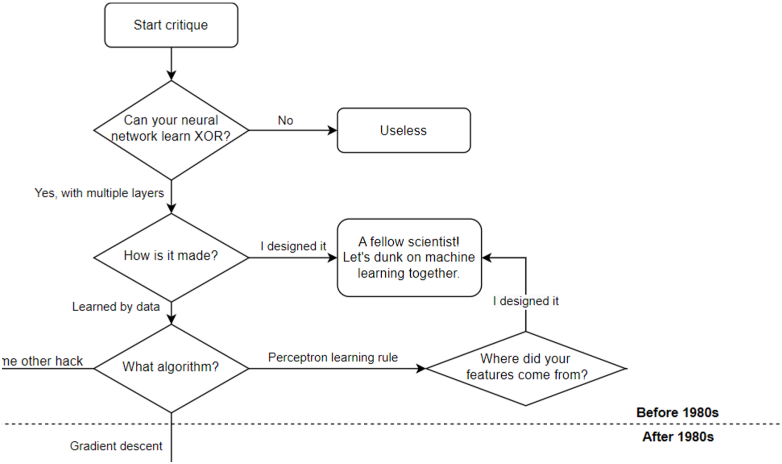
What is the actual XOR problem?



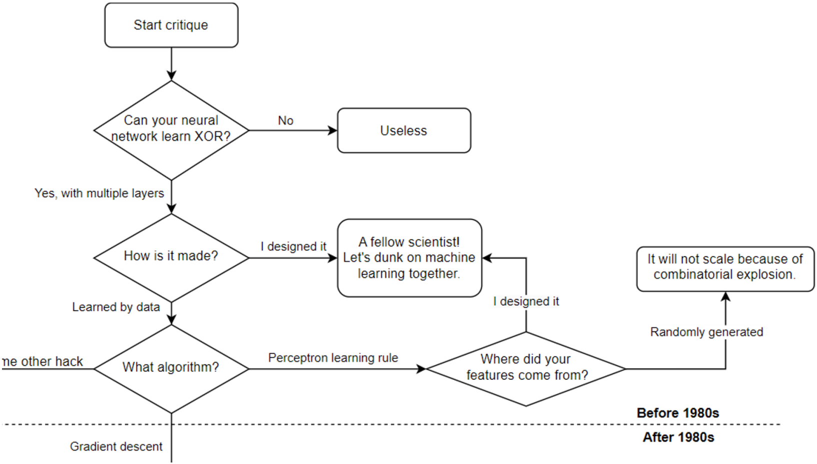
What is the actual XOR problem?



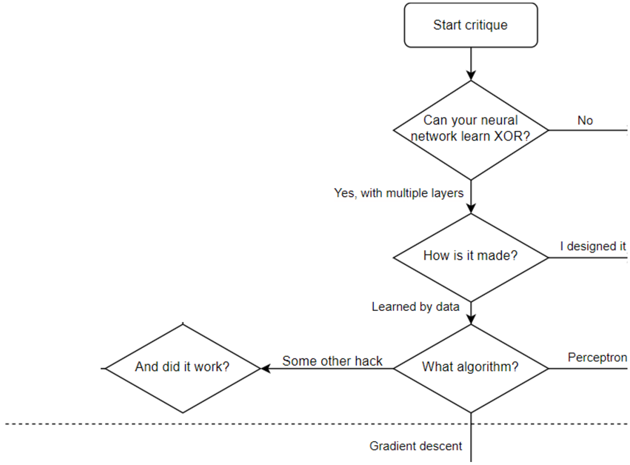
What is the actual XOR problem?



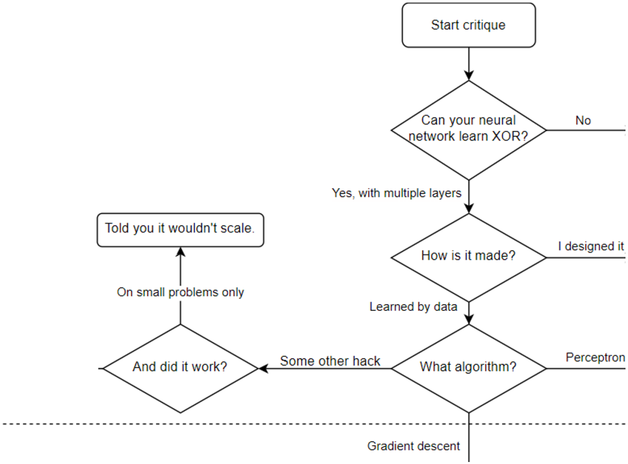
What is the actual XOR problem?



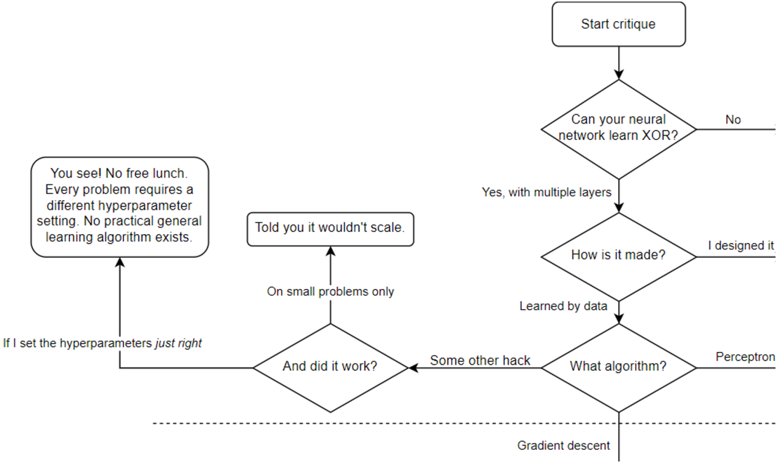
What is the actual XOR problem?



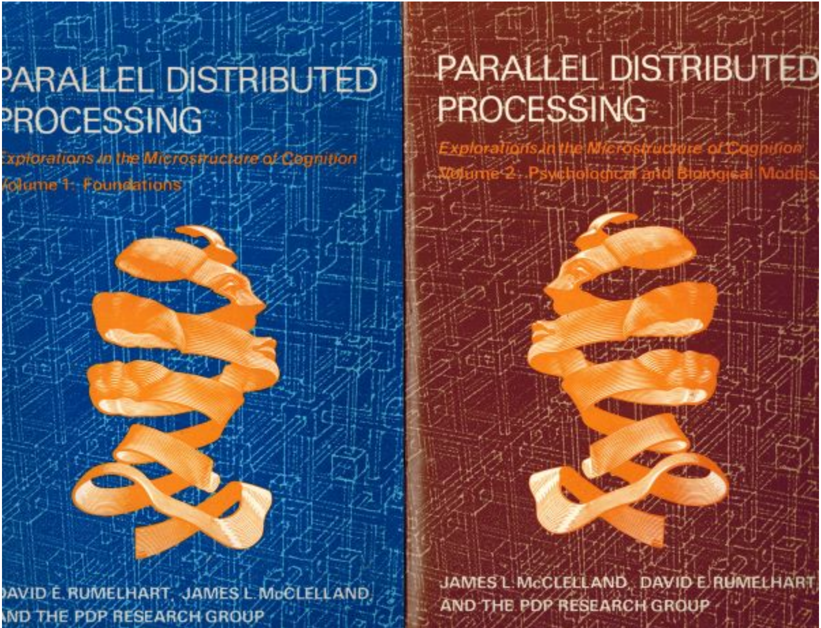
What is the actual XOR problem?



What is the actual XOR problem?



Neural Network Spring (1980s)



Spring has sprung (1980s)

Why did neural networks emerge again in the 1980s?

- Backprop developed multiple times, 1970s.
- But really hard to be known (just ask Paul Werbos).
- Also, widespread cheap compute (like PDP-11).

Figure: PDP-11 16-bit minicomputer (1970–1997)

Paul Werbos

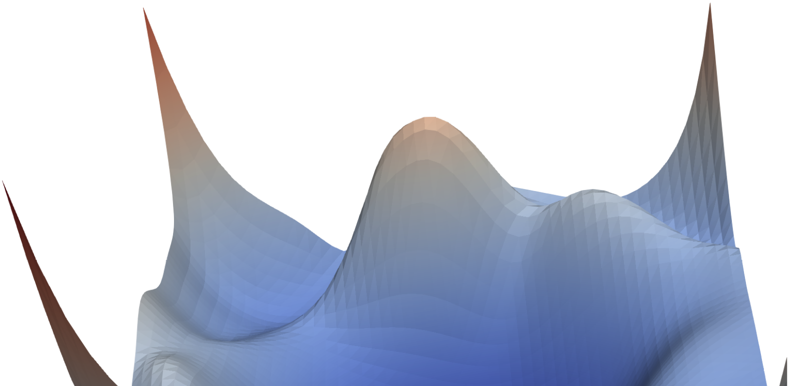
- Discovered backprop in 1972 during PhD, by mathematifying Freud's "psychic energy".
- Committee unsure of the math (or the Freud), "You've got to prove some theorems first."
- Had to find a supporting advisor.
- Stephen Grossberg: "It's already been done before. Or else it's wrong. I'm not sure which of the two, but I know it's one of the two."
- Jerome Lettvin: "You're saying that there's motive and purpose in the human brain [like a Freudian, but] you cannot take an anthropomorphic view of the human brain."
- Marvin Minsky: "Everybody knows a neuron is a 1-0 spike generator... [Your model] is totally crazy."

Paul Werbos

- So he simplified it down to piecewise linear activation.
- PhD committee: "too trivial to be worthy of a Harvard PhD".
- Lost funding; lived in the slums. "I remember getting the shakes from inadequate nutrition."
- Found a supporting advisor (Karl Deutsch, who wrote *The Nerves of Government*).
- Finished PhD in 1974. Thesis contained backprop.
- A decade of misadventures working in the federal bureaucracy.
- Finally published backprop in 1982.

Boltzmann machines

- Consider n neurons connected by wires. Overall state of system is $X = (x_1, \ldots, x_n)$.
- Energy of the system is $E(X)$.
- Probability of the system is $Pr(X) = \frac{e^{-\beta E(X)}}{\sum_{X'} e^{-\beta E(X')}}$.
- Many beautiful theorems. Physicists love this.

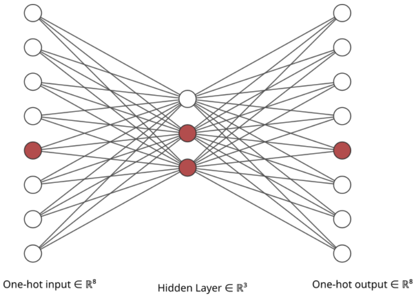


Rumelhart and Hinton

- In 1982, David Rumelhart re-discovered backprop and told Geoffrey Hinton about it.
- Hinton resisted backprop for **3 whole years** before relenting.
- Hinton: "It couldn't break symmetry." – Rumelhart: "Use noise."
- Hinton: "It gets stuck in local minima, unlike Boltzmann machine."
- Turns out Boltzmann machines also get stuck in local minima, oops.
- Boltzmann machines are beautiful, but they just don't work!
- Hinton's last resort: backpropagation (1985).

Rumelhart and Hinton

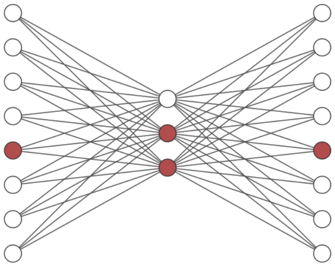
- Asked his students: "Who wants to implement this?"
- Students "thoroughly indoctrinated into Boltzmann machines" Refused to work on it, so Hinton programmed it himself.
- Tried it on an 8-3-8 autoencoder task. The weights looked weird.



- "Turns out backprop's not that good after all."...

Rumelhart and Hinton

- Asked his students: "Who wants to implement this?"
- Students "thoroughly indoctrinated into Boltzmann machines" Refused to work on it, so Hinton programmed it himself.
- Tried it on an 8-3-8 autoencoder task. The weights looked weird.



- "Turns out backprop's not that good after all."...wait, the loss is zero?!
- Published backprop (1985).

Further controversies

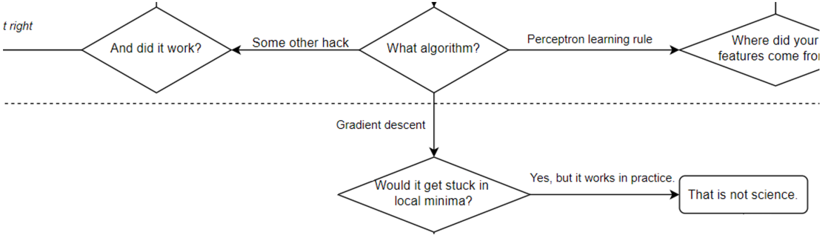
- The *Past-Tense Debate*, about whether statistical language models can truly learn linguistic rules.
- Started with a chapter in *Parallel Distributed Processing* (1986); lasted well into 2000s. Died of exhaustion. [Seidenberg and Plaut, 2014]
- Not very important now, but angered a lot of Chomskyan linguists.
- May be analogous to the modern debate about LLM?

Further controversies

- Minsky and Papert re-published *Perceptrons* (1988). Added over 40 pages of commentary re-rejecting large neural networks.
- Seems like people ignored Minsky and Papert this time.

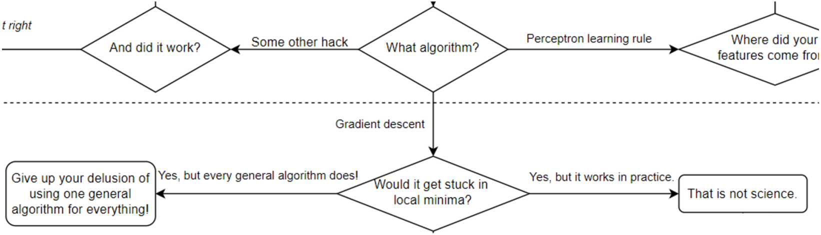
Further controversies

- Minsky and Papert re-published *Perceptrons* (1988). Added over 40 pages of commentary re-rejecting large neural networks.
- Seems like people ignored Minsky and Papert this time.



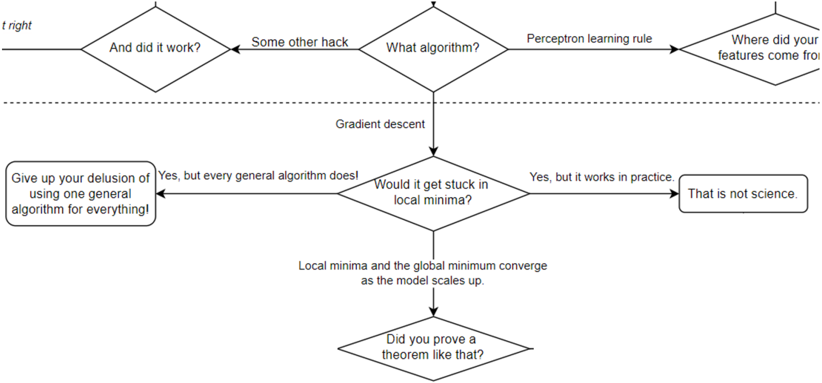
Further controversies

- Minsky and Papert re-published *Perceptrons* (1988). Added over 40 pages of commentary re-rejecting large neural networks.
- Seems like people ignored Minsky and Papert this time.



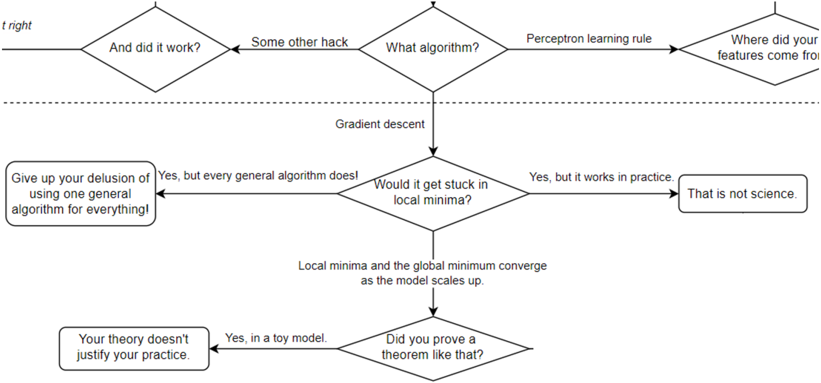
Further controversies

- Minsky and Papert re-published *Perceptrons* (1988). Added over 40 pages of commentary re-rejecting large neural networks.
- Seems like people ignored Minsky and Papert this time.



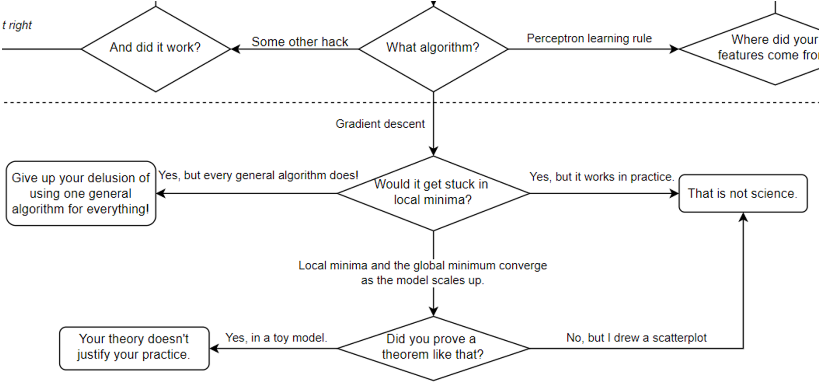
Further controversies

- Minsky and Papert re-published *Perceptrons* (1988). Added over 40 pages of commentary re-rejecting large neural networks.
- Seems like people ignored Minsky and Papert this time.



Further controversies

- Minsky and Papert re-published *Perceptrons* (1988). Added over 40 pages of commentary re-rejecting large neural networks.
- Seems like people ignored Minsky and Papert this time.



Some (bitter) lessons

History is useless

- You can draw lessons from history, but everyone else also can draw the opposite ones.
- Use history to help you make history. Accuracy is overrated.
- History for history's sake is for old professors (that's Schmidhuber) and procrastinating PhD students (that's me).

  We need it for life and action, not as a convenient way to avoid life and action... to value its study beyond a certain point mutilates and degrades life.
  — Nietzsche, On the Use and Abuse of History for Life

Ideas can't be lost

- Ideas can't be lost.
- People will keep repeating the same ideas until it starts working, then it will not be lost again.
- Not the first inventor who gets the credit, but the last re-inventor.
- History is a cenotaph, useless for practitioners.

  Those who cannot remember the past are condemned to repeat it.
  — *George Santayana*

Ideas can't be lost

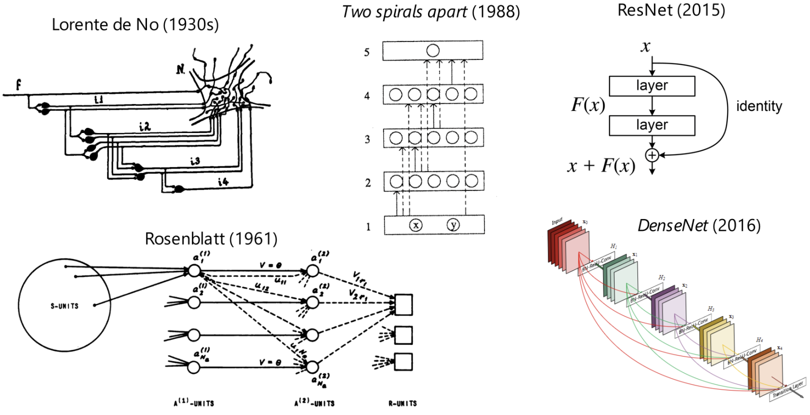
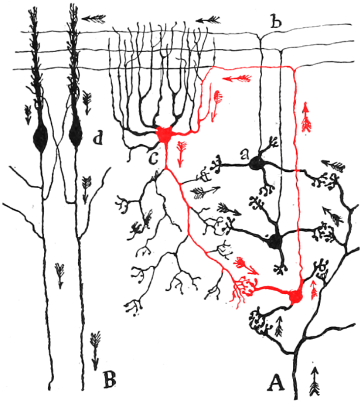


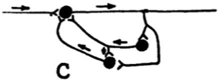
Figure: Reinventions of the residual network.

Ideas can't be lost

Cajal (1900s)



Lorente de No (1930s)



McCulloch and Pitts (1943)

Figure: Reinventions of the recurrent network.

Ideas can't be lost

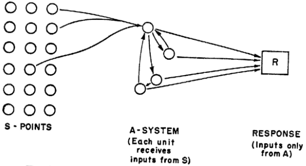


Fig. 4. Typical connections for a cross-coupled perceptron.

Rosenblatt (1960)

FEATURES ASSIGNED TO NODES

	1	2	3	4	5	6
	NAME	HEIGHT	AGE	WEIGHT	HAIR	EYES
−1	SMITH	TALL	OLD	THIN	BROWN	BLUE
+1	JONES	SHORT	YOUNG	FAT	BLOND	BROWN

NODES

MEMORIES	1	2	3	4	5	6
A	+1	+1	+1	−1	−1	−1
B	+1	−1	+1	+1	−1	+1
C	+1	+1	−1	+1	−1	−1

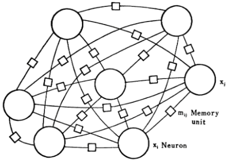
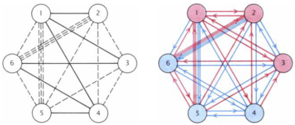


Figure 1. Associative Memory as a Neural Network.

Nakano (1971)



Hopfield (1982)

Figure: Reinventions of the Hopfield network.

Compute matters more

- For each elegant idea, there is an equal and opposite one.
- Benchmarks speak louder than ideas.
- Good ideas are patient seeds, waiting for the soil of compute.
- cheap compute → cheap experiments → better ideas

  Which one helps, which doesn't help. Let's rip it out. Let's replace it with this... All of these components of the Transformer were the output of this extremely high-paced, iterative trial and error. [Levy, 2024]

  For months, they toyed with various ways to add more layers and still get accurate results. After a lot of trial and error, the researchers hit on a system they dubbed "deep residual networks". [Linn, 2015]

It just keeps bittering

- Even with backprop, not much was possible in the 1970s. Ted Hoff was right to leave neural networks for Intel. But how many are like him?
- Minsky never accepted large neural networks.
  - "We have not found (by thinking or by studying the literature) any other really interesting class of multilayered machine." (1969)
  - "You're not working on the problem of general intelligence. You're just working on applications." (2006)
- Noam Chomsky never accepted statistical language modeling.
  - "the notion of 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (1969)
  - "[Large language models] have achieved zero... GPT-4 will be exactly the same. It'll use even more energy and achieve exactly nothing, for the same reasons. So there's nothing to discuss." (2022)

It just keeps bittering

- Even Geoffrey Hinton remained unhappy about backprop.
- "I was always a bit disappointed. I mean, intellectually, backpropagation wasn't nearly as satisfying as Boltzmann machines. It's not just because I didn't think of it. I think it's because it didn't have the nice probabilistic interpretation." (1995)
- Also he was doing Deep Belief Networks (multilayered Boltzmann machines) up until 2009.
- And so, the bitter lesson continues...

Teaser

Not the End of History! I'm working on the history of the *second* neural network winter – should appear this December.

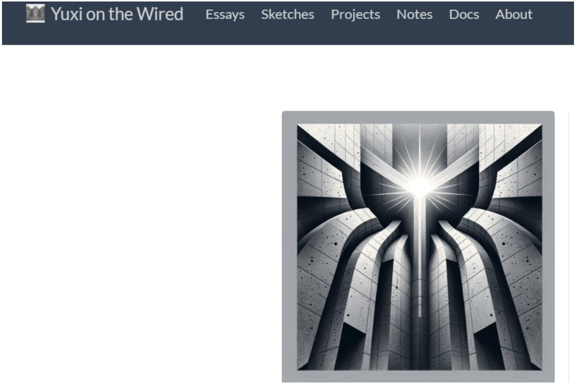


Figure: Watch this space!

Further reading

- https://yuxi-liu-wired.github.io/essays/posts/backstory-of-backpropagation/
- https://yuxi-liu-wired.github.io/essays/posts/reading-perceptron-book/
- https://yuxi-liu-wired.github.io/essays/posts/perceptron-controversy/
- https://yuxi-liu-wired.github.io/sketches/posts/history-neural-networks-talk-notes/

Bernstein, J. (1981).
Marvin Minsky's Vision of the Future.
The New Yorker.

de Nó, R. L. (1938).
Analysis of the activity of the chains of internuncial neurons.
Journal of Neurophysiology, 1(3):207–244.

Hebb, D. O. (2002).
The Organization of Behavior : A Neuropsychological Theory.
Psychology Press.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020).
Scaling Laws for Neural Language Models.

Levy, S. (2024).
8 Google Employees Invented Modern AI. Here's the Inside Story.

Wired.

Linn, A. (2015).
Microsoft researchers win ImageNet computer vision challenge.

McCulloch, W. S. and Pitts, W. (1943).
A logical calculus of the ideas immanent in nervous activity.
The bulletin of mathematical biophysics, 5(4):115–133.

Minsky, M. (1988).
The Society of Mind.
A Touchstone Book. Simon and Schuster, New York, 6. pb-pr edition.

Nilsson, N. J. (2009).
The Quest for Artificial Intelligence.
Cambridge University Press, Cambridge ; New York, 1st edition edition.

Norbert Wiener (1966).
Extrapolation, Interpolation, and Smoothing of Stationary Time Series.

Ramón y Cajal, S. (1909).
Histologie du système nerveux de l'homme et des vertébrés, volume II.
Paris : A. Maloine.

Rosenblatt, F. (1958).
The perceptron: A probabilistic model for information storage and organization in the brain.
Psychological review, 65(6):386.

Rosenblatt, F. (1962).
Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, volume 55.
Spartan books Washington, DC.

Rosenfeld, E. and Anderson, J. A., editors (2000).
Talking Nets: An Oral History of Neural Networks.
The MIT Press, reprint edition edition.

Seidenberg, M. S. and Plaut, D. C. (2014).

Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate.
Cognitive Science, 38(6):1190–1228.

Turkle, S. and Papert, S. (1990).
Epistemological Pluralism: Styles and Voices within the Computer Culture.
Signs: Journal of Women in Culture and Society, 16(1):128–157.

Widrow, B. (1962).
Generalization and information storage in networks of adaline neurons.
Self-organizing systems, pages 435–461.

Wiener, N. (2017).
Norbert Wiener–a Life in Cybernetics.
The MIT Press, Cambridge, Massachusetts.