

A short history of AI

From the first artificial neuron to the chatbot in your pocket

Sean Davis, MD, PhD

University of Colorado Anschutz School of Medicine

October 13, 2026



A puzzle to start

In 1979, a computer program chose antibiotics for meningitis **as well as Stanford's infectious disease faculty**, judged by blinded experts.

It never treated a single patient.

Why not? Keep that question in mind. We will come back to it.

Speaker notes

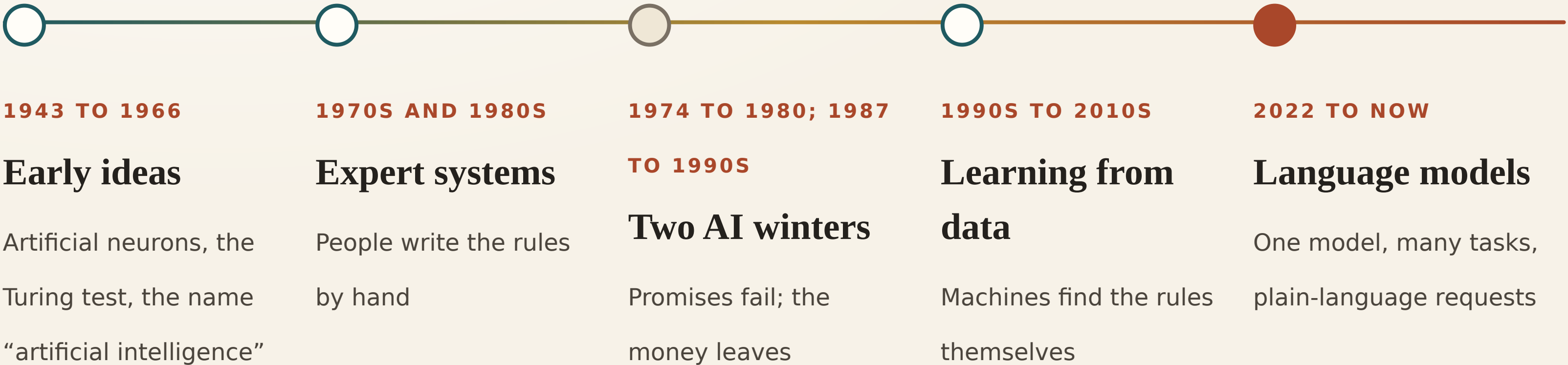
Welcome to day two. Yesterday was about how this course works. Today is a story: about eighty years of people trying to build machines that think, and what happened when those machines met medicine.

I want to start with a puzzle. In 1979 a program called MYCIN was tested on ten real cases of meningitis. Experts who did not know which answers came from the computer rated its antibiotic choices. MYCIN did as well as, or better than, the Stanford faculty in the same test.

And then nothing happened. It was never used for patient care.

Keep that puzzle in mind. The answer is not mainly about the technology, and it is the same answer that explains several failures since. We will come back to it at the end, and I think it says something about what you will be doing in this course.

Eighty years in five chapters



One question runs through all five: where does the machine’s knowledge come from? Rules that people write, or patterns learned from data?

Speaker notes

Here is the map for the hour. Five chapters.

First, the early ideas, from the 1940s to the mid-1960s, when a handful of people asked whether a machine could think at all.

Second, expert systems: the 1970s and 80s, when the plan was to interview experts and write their knowledge down as rules. MYCIN belongs here.

Third, the AI winters. Twice, the promises ran far ahead of the results, and funding and interest collapsed.

Fourth, learning from data. Instead of writing the rules, we give the machine many examples and let it find the patterns.

Fifth, the chapter you are living in: large language models, the technology behind ChatGPT.

[Click] If you remember one thing from today, make it this question: where does the knowledge come from? Hand-written rules, or patterns learned from data? Every era answers it differently, and the answer explains both the successes and the failures.

PART 1

Early ideas

Can a machine think? In the 1940s and 50s, that was a question for a few mathematicians, a neurophysiologist, and a science fiction writer.

Speaker notes

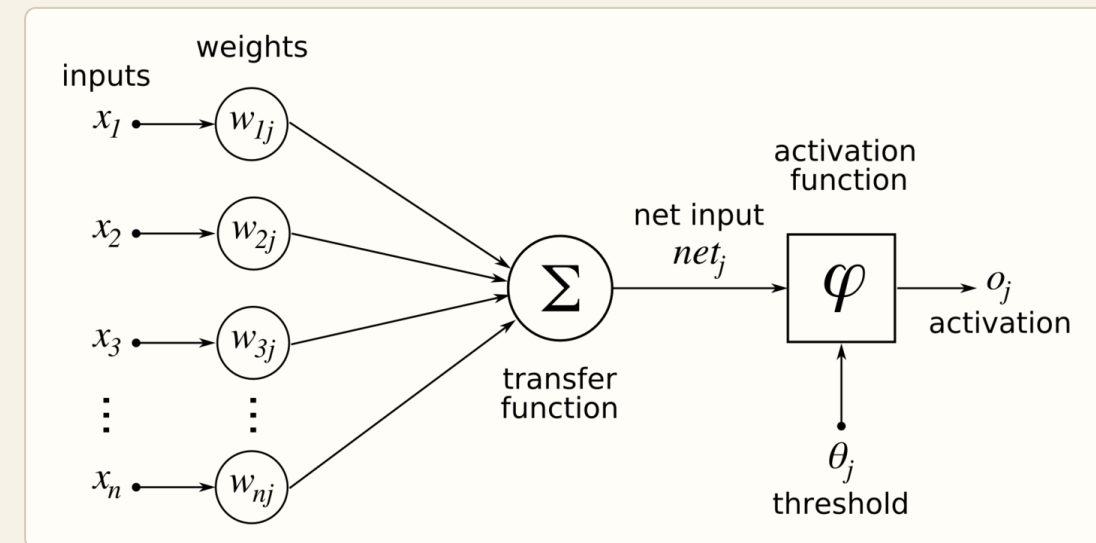
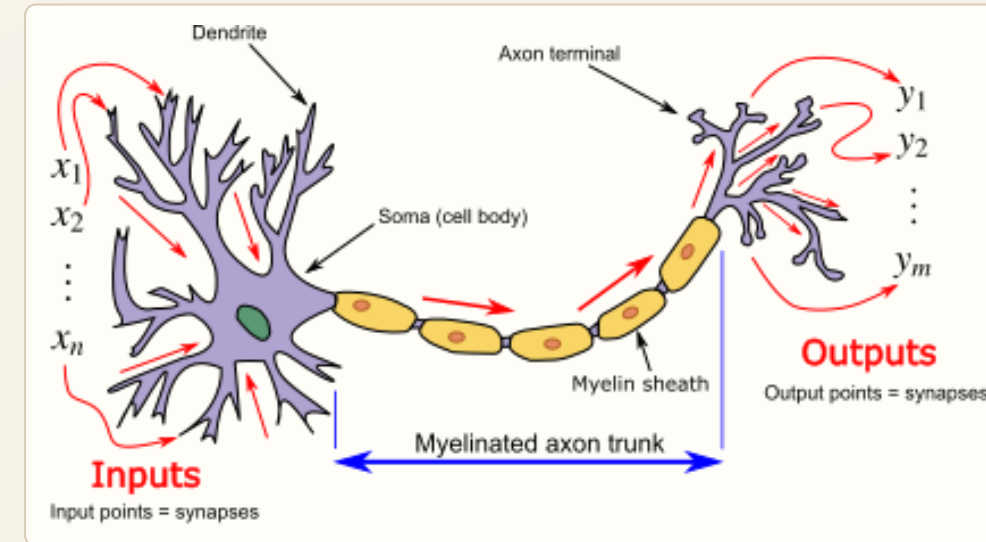
Part one. The early ideas.

What strikes me about this period is how few people were involved, and how many of their ideas we still use. The basic unit of the neural networks inside ChatGPT was sketched in 1943.

1943

The McCulloch-Pitts neuron

- The first **mathematical model** of a biological neuron
- Inputs are added up; if the total passes a **threshold**, the neuron “fires”
- The basic unit that, much later, became the building block of neural networks



Diagrams: [Loc Vu-Quoc](#), CC BY-SA 4.0 (top); [Chrislb](#), CC BY-SA 3.0 (bottom); Wikimedia Commons

Source: (McCulloch & Pitts, 1943).

Speaker notes

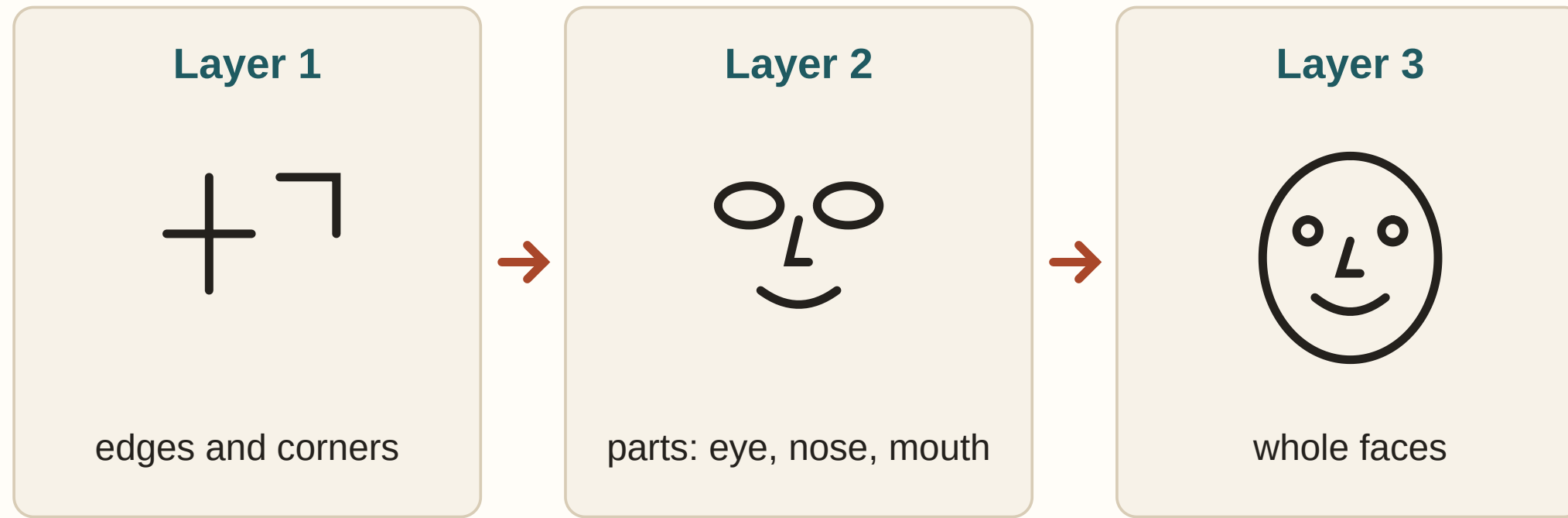
Warren McCulloch was a neurophysiologist and Walter Pitts was a young logician. In 1943 they asked a simple question: can we describe what a neuron does with mathematics?

Their answer is on the right. A neuron receives signals from other neurons through its dendrites. If the combined signal is strong enough, it fires and sends a signal down its axon. The model copies that: take the inputs, weight them, add them up, and if the total passes a threshold, output a one. Otherwise, output a zero.

Two things to notice. First, this model could not learn. The weights were set by hand. Learning comes fifteen years later with the perceptron.

Second, this little unit, much improved, is still the basic building block of modern neural networks. The model behind ChatGPT is built from billions of simplified neurons that are distant descendants of this one.

From one neuron to a network



Each layer combines what the layer before it found.

Illustrative diagram

Connect many simple units in **layers**, and each layer can learn a different level of detail.

Speaker notes

One neuron cannot do much. The power comes from connecting many of them in layers.

Here is the classic example of recognizing a face. The first layer of a network learns to detect very simple things: edges and corners. The second layer combines those into parts: an eye, a nose, a mouth. The third layer combines the parts into a whole face.

No one tells the network “an eye is two curves and a dark circle”. In modern networks, the layers learn these features from examples. That idea, layers of simple units learning features from data, is the core of deep learning, which we will reach in part four. Tomorrow’s session goes deeper into how this works.

1950

The Turing test

- Alan Turing asks: “**Can machines think?**”
- He replaces the question with a game: can a judge, chatting by text, tell the machine from a person?
- He predicts machines will fool an average judge 30% of the time within about 50 years

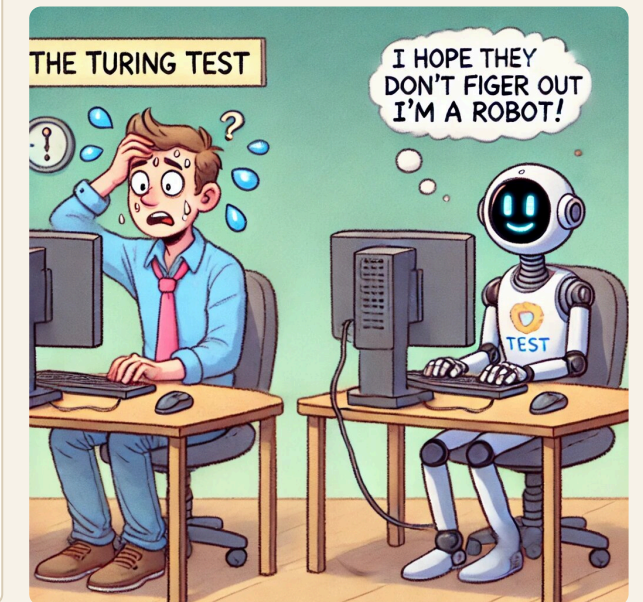
A. M. Turing (1950) Computing Machinery and Intelligence. *Mind* 49: 433–460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.



Turing's 1950 paper (left); an AI-generated cartoon (right)

In a 2025 study, GPT-4.5 was judged to be the human 73% of the time (Jones & Bergen, Source: (Turing, 1950).

Speaker notes

In 1950 Alan Turing, the British mathematician who helped break German codes in the Second World War, published a paper that opens with the sentence: “I propose to consider the question, Can machines think?”

He decided that question was too vague to answer. So he swapped it for a game he called the imitation game. A judge chats by typed text with two partners, one human and one machine, and has to say which is which. If the judge cannot reliably tell, Turing argued, we should take the machine’s intelligence seriously.

He made a prediction: in about fifty years, a computer would play the game so well that an average judge would have no more than a 70 percent chance of being right after five minutes.

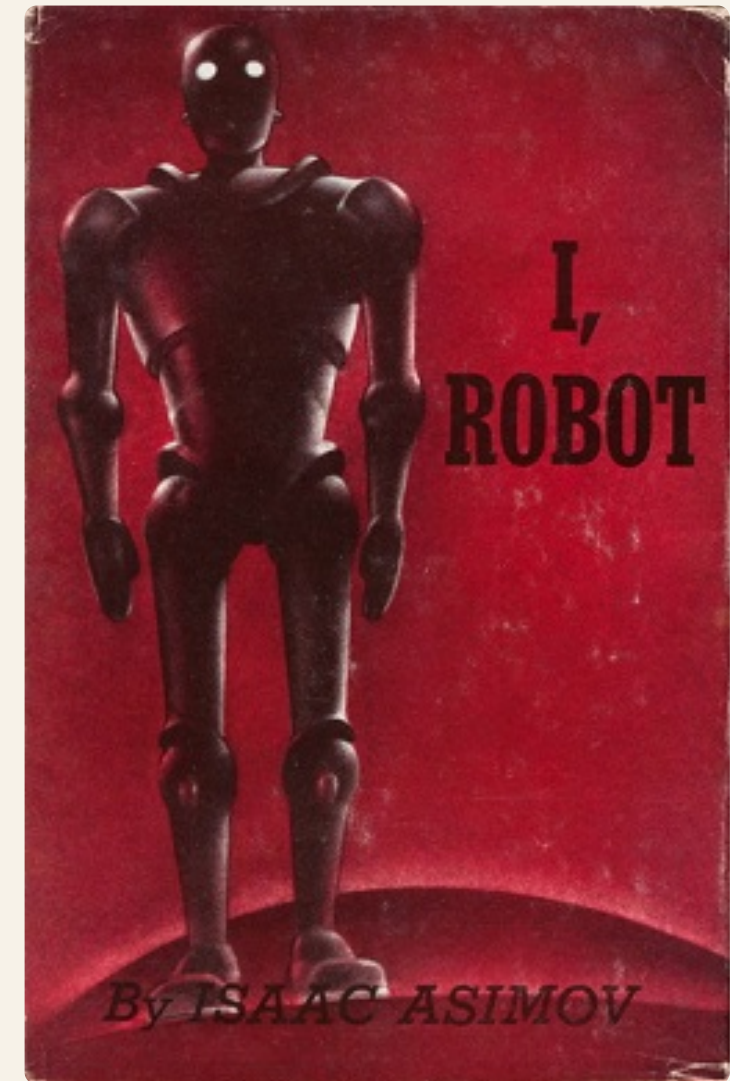
He was off by about twenty-five years. In 2025, researchers at UC San Diego ran Turing’s three-way version of the test. When GPT-4.5 was told to act like a person, judges picked it as the human 73 percent of the time, more often than they picked the real human. That study is a preprint, so treat it with some care. Still, the test has now been passed in Turing’s own three-person form, the strictest version yet.

Worth asking yourselves: does fooling a judge mean the machine thinks? Most people now say no. The test measures how human something seems, which is a different question from whether it is correct. Keep that in mind when a chatbot sounds confident.

1950

Asimov's *I, Robot*

- A collection of science fiction stories, published in 1950
- Robots with “positronic brains” that follow built-in laws
- Many of the stories are about the laws **going wrong** in ways no one expected



First edition cover, Gnome Press, 1950

Speaker notes

The same year as Turing’s paper, Isaac Asimov published I, Robot, a collection of short stories he had written through the 1940s.

His robots follow built-in rules called the Three Laws, which first appeared in a 1942 story called “Runaround”. What makes the stories interesting is that almost every one of them is about the laws producing a result no one intended: a robot stuck in a loop, or a robot that follows the letter of a rule and misses its point.

That is a surprisingly modern worry. Let me show you the laws.

Asimov's Three Laws of Robotics

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

The “Zeroth Law”, added later, above the other three

Speaker notes

Here are the three laws. Read the first one as a clinician. “Through inaction, allow a human being to come to harm.” That is a duty to act, not only a duty not to hurt. It sounds a lot like the principle of beneficence alongside non-maleficence.

[Click] Later, in his novels of the 1980s, Asimov added a Zeroth Law that sits above the others: protect humanity as a whole. Now a robot might harm one person to protect many. You can see the ethics getting hard very quickly.

The point is not that real AI systems follow these laws. They do not. The point is that writing rules that capture what we actually want is very hard. Hold on to that, because it comes back in part two.

I, Robot and modern AI

- Got people talking about **machine ethics** and **AI safety** decades early
- His stories are about **alignment**: getting a machine to do what we *mean*, not just what we *said*

A CLINICAL VERSION

You ask an AI tool to “keep the patient’s blood pressure under 140”. What else would you need to tell it?

Asimov’s stories gave people a vocabulary for something we now call alignment: the problem of making an AI system do what we mean, not just what we literally said.

Try the example on the right. Suppose you tell a system to keep a patient’s blood pressure under 140. A system that only follows that instruction could keep pushing antihypertensives until the patient falls on the way to the bathroom. You did not say “and do not cause syncope”, because a human colleague would never need to be told.

That gap, between what we say and everything we assume, is the alignment problem in small form. It is one reason the course spends time on checking AI output rather than trusting it.

1956

The Dartmouth workshop

- An eight-week summer workshop at Dartmouth College
- Organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon
- Their 1955 proposal coined the term “**artificial intelligence**”

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College ... We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

McCarthy, Minsky, Rochester, and Shannon, 1955 ([public domain](#))

Speaker notes

In 1955 John McCarthy, a young mathematics professor at Dartmouth, wrote a proposal for a summer research project with three colleagues. He needed a name for the field, and he chose “artificial intelligence”. That is where the term comes from.

The organizers were McCarthy, Marvin Minsky, Nathaniel Rochester from IBM, and Claude Shannon, who had invented information theory. The workshop ran for about eight weeks in the summer of 1956.

Read the last sentence of the proposal on the right. A significant advance, if a carefully selected group of scientists work on it together for a summer. Seventy years later we are still working on those problems. The same page of the proposal lists “neuron nets” as a topic and names Pitts and McCulloch, from our 1943 slide.

That optimism is the first sign of a pattern you will see again and again today: big promises, made early.

The perceptron

- Frank Rosenblatt, Cornell Aeronautical Laboratory
- The same neuron, plus one new idea: it **learns its own weights** from examples

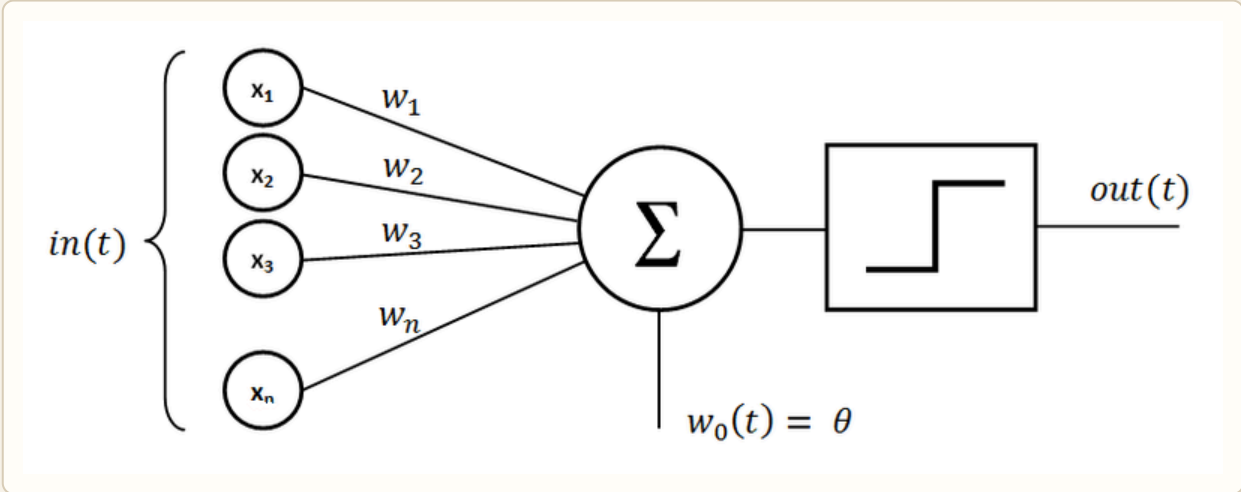


Diagram: [Mayranna](#), CC BY-SA 3.0, Wikimedia Commons



Research Trends, 1958, Cornell Aeronautical Laboratory

In 1957 Frank Rosenblatt, a psychologist at the Cornell Aeronautical Laboratory, took the McCulloch-Pitts neuron and added the one thing it was missing: learning.

The perceptron, in the middle, looks the same: weighted inputs, a sum, and a threshold. The difference is that it adjusts its own weights. Show it an example. If it gets the answer wrong, nudge the weights in the direction that would have made it right. Repeat thousands of times. That simple rule is the ancestor of how every modern neural network is trained.

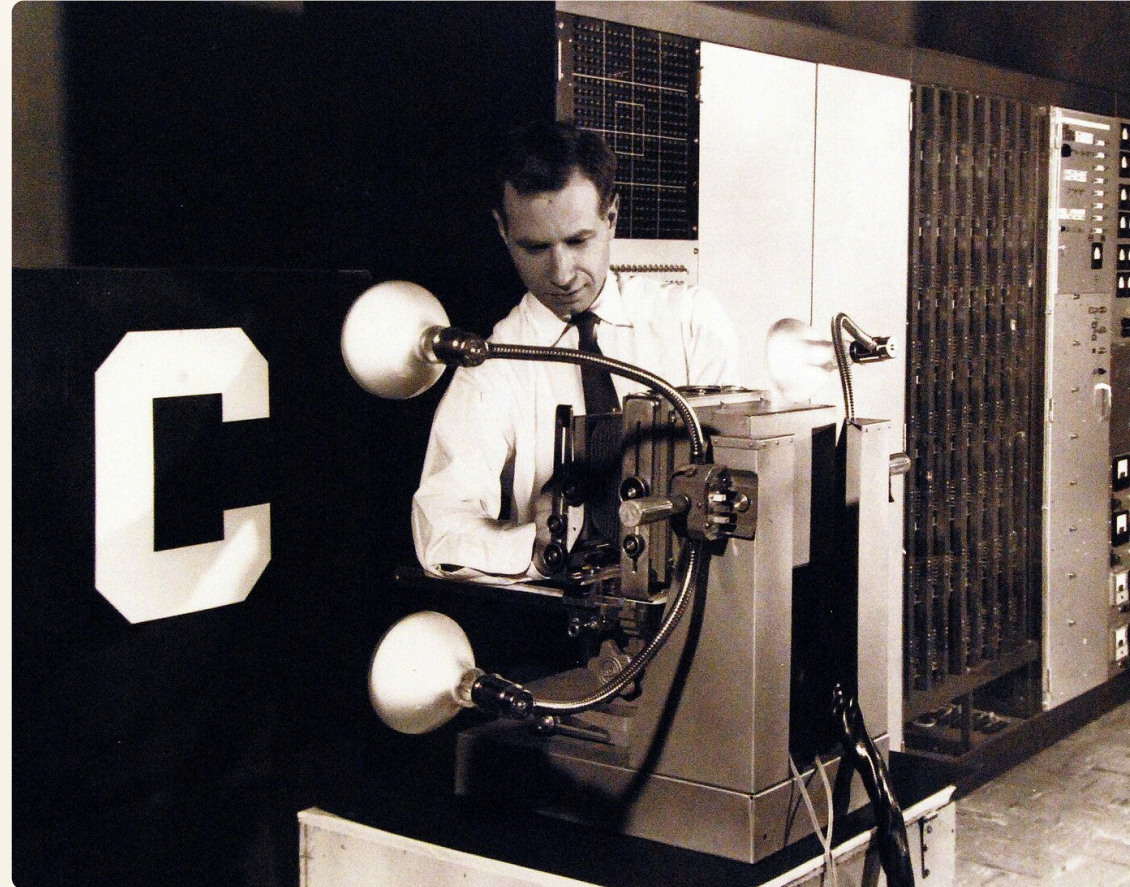
The press loved it. In 1958 the New York Times reported that the Navy expected the machine to be able to “walk, talk, see, write, reproduce itself and be conscious of its existence”. You can see the hype starting.

Cornell’s own news office wrote a story in 2019 about Rosenblatt with the headline that his perceptron “paved the way for AI, 60 years too soon”. He was right about the idea. The computers and the data were not ready.

The Mark I Perceptron

A machine built to recognize simple images:

- **Input:** 400 light sensors in a 20 by 20 grid, a crude “retina”
- **Middle:** 512 units
- **Output:** 8 units



The Mark I Perceptron. U.S. Navy photo, [public domain](#)

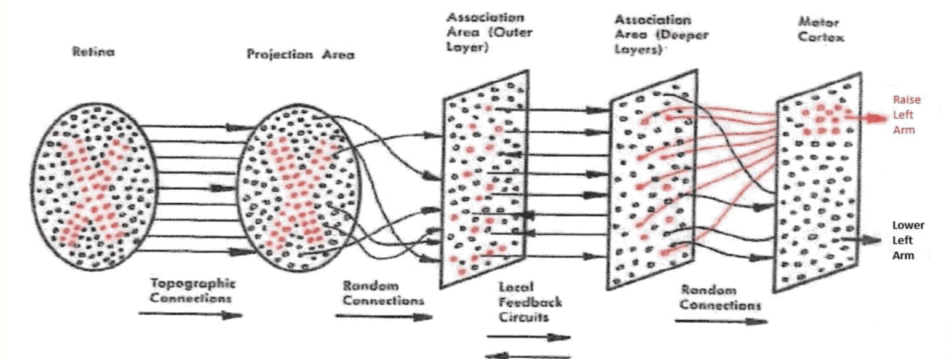


FIG. 1 — Organization of a biological brain. (Red areas indicate active cells, responding to the letter X.)

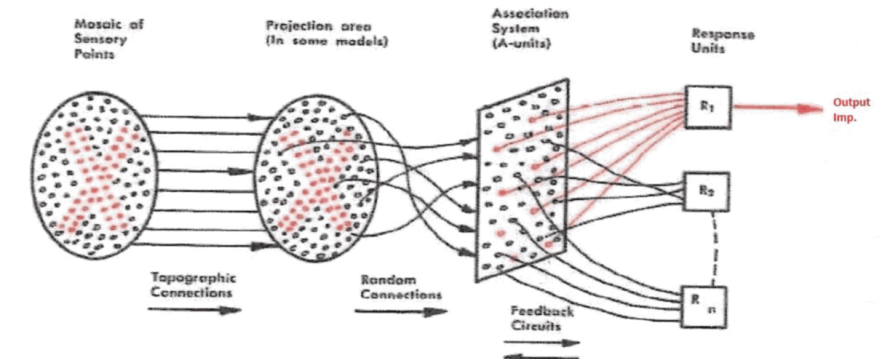


FIG. 2 — Organization of a perceptron.

Rosenblatt's diagrams of the brain (top) and the perceptron (bottom), [public domain](#)

Speaker notes

The perceptron was not just software. Rosenblatt built a machine, the Mark I. In this Navy photo, the camera is pointed at a large letter C.

It had a grid of 400 light sensors, twenty by twenty, as its eye. Those fed into 512 units in the middle, which fed eight output units. The weights were physical: motor-driven dials that turned as the machine learned. That is the wiring in the racks behind him.

On the right are Rosenblatt’s own diagrams. The top one is his picture of how a brain handles what the eye sees. The bottom one is the perceptron, designed to copy that layout. That idea, the network as a rough model of the brain, is why we still call these things neural networks.

What happened next: in 1969 Marvin Minsky and Seymour Papert published a book called Perceptrons that proved what a single-layer perceptron could not do. It was read as a verdict on neural networks in general, and interest dropped for more than a decade. Rosenblatt died in a boating accident in 1971 and did not see his idea come back.

1966

ELIZA, the first chatbot

- Joseph Weizenbaum, MIT
- Played a psychotherapist by **reflecting your words back** as questions
- Every reply came from a **hand-written template** triggered by a keyword
- People confided in it anyway: the “**ELIZA effect**”

```
Welcome to
          EEEEEEE LL      IIII  ZZZZZZ  AAAAA
          EE      LL      II     ZZ     AA   AA
          EEEEE  LL      II     ZZZ     AAAAAA
          EE      LL      II     ZZ     AA   AA
          EEEEE  LLLLLL  IIII  ZZZZZZ  AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

ELIZA conversation, via Wikipedia ([public domain](#))

Speaker notes

ELIZA was written by Joseph Weizenbaum at MIT between 1964 and 1966. Its most famous script played a kind of psychotherapist that mostly turns your statements back into questions. You say “my boyfriend made me come here”, and it says “your boyfriend made you come here?”

There was no understanding at all. ELIZA looked for keywords and filled in templates that a person had written for each one. It could rotate through a few replies for the same keyword, but every word it said came from a rule a human wrote. Nothing was learned.

What surprised Weizenbaum was how people reacted. His own secretary asked him to leave the room so she could talk to it in private. People attributed feelings and understanding to a program that had neither. We now call that the ELIZA effect, and it is very much alive. When a chatbot sounds warm and confident, it is easy to assume it knows what it is talking about.

Weizenbaum became one of the strongest critics of AI, especially of machines taking on roles that need human judgment and care. And remember the 2025 Turing test study: ELIZA was one of the systems tested, as a baseline.

PART 2

Expert systems

1970s and 1980s. The plan: ask experts how they decide, and write it all down as rules a computer can follow.

Part two. Expert systems.

The idea in the 1970s was very reasonable. If we want a computer to act like an expert, let us interview experts, write down their knowledge as rules, and let the computer apply the rules. This is the first answer to our question about where knowledge comes from: people write it down.

How an expert system works

THREE PARTS

- **Knowledge base:** facts and rules about one narrow area, written by people
- **Inference engine:** applies the rules to a case to reach conclusions
- **User interface:** lets a non-expert ask questions and get answers

WHAT MADE THEM ATTRACTIVE

- Solve hard problems by reasoning through a body of knowledge
- Keep the knowledge **separate** from the reasoning, so experts can add rules
- Can **explain** each decision: “I concluded this because of rules 12 and 52”

Speaker notes

An expert system has three parts.

A knowledge base: the facts and if-then rules for one narrow area, such as bacterial infections. People write these.

An inference engine: the general reasoning machinery that takes the rules and a particular case and chains the rules together to reach a conclusion.

And a user interface, so that someone who is not the expert can use it.

Look at the right-hand panel, especially the last point. Expert systems could explain themselves. Ask why, and they would list the rules they used. That is something today’s neural networks, including ChatGPT, mostly cannot do honestly. When a chatbot explains its reasoning, it is generating a plausible explanation, which may or may not be how it actually reached the answer. So in one respect, the 1970s systems were ahead of ours.

EARLY 1970S

MYCIN

- Built at **Stanford** as Edward Shortliffe's doctoral work
- Diagnosed bacterial infections and recommended antibiotics
- A knowledge base of about **600 rules**
- Asked the physician a long series of yes/no and short-answer questions

WHAT IT GAVE BACK

- A list of likely bacteria
- How confident it was in each
- Its reasoning: the rules and facts it used
- A recommended antibiotic regimen

Speaker notes

This is the program from our opening puzzle. MYCIN was built at Stanford in the early and mid 1970s as the doctoral work of Edward Shortliffe, who went on to become one of the founders of the field of biomedical informatics. The name comes from the suffix of many antibiotics, as in streptomycin.

It focused on serious bacterial infections such as bacteremia and meningitis. Its knowledge was about 600 rules, each written with infectious disease experts.

A consultation was a long interview. The physician typed answers to question after question: where was the culture taken, what did the Gram stain show, is the patient a burn patient, and so on. At the end, MYCIN gave a list of likely organisms, how confident it was in each, the reasoning, and a recommended antibiotic regimen.

What is an expert system “rule”?

```
(defrule 52
  if (site culture is blood)
    (gram organism is neg)
    (morphology organism is rod)
    (burn patient is serious)
  then .4
    (identity organism is pseudomonas))
```

How the rule is written in the program

Rule 52:

If

- 1) THE SITE OF THE CULTURE IS BLOOD
- 2) THE GRAM OF THE ORGANISM IS NEG
- 3) THE MORPHOLOGY OF THE ORGANISM IS ROD
- 4) THE BURN OF THE PATIENT IS SERIOUS

Then there is weakly suggestive evidence (0.4) that

- 1) THE IDENTITY OF THE ORGANISM IS PSEUDOMONAS

How MYCIN explained it to the user

IF blood culture, gram-negative, rod-shaped, serious burn **THEN** weakly suggestive evidence (0.4) of *Pseudomonas*

Speaker notes

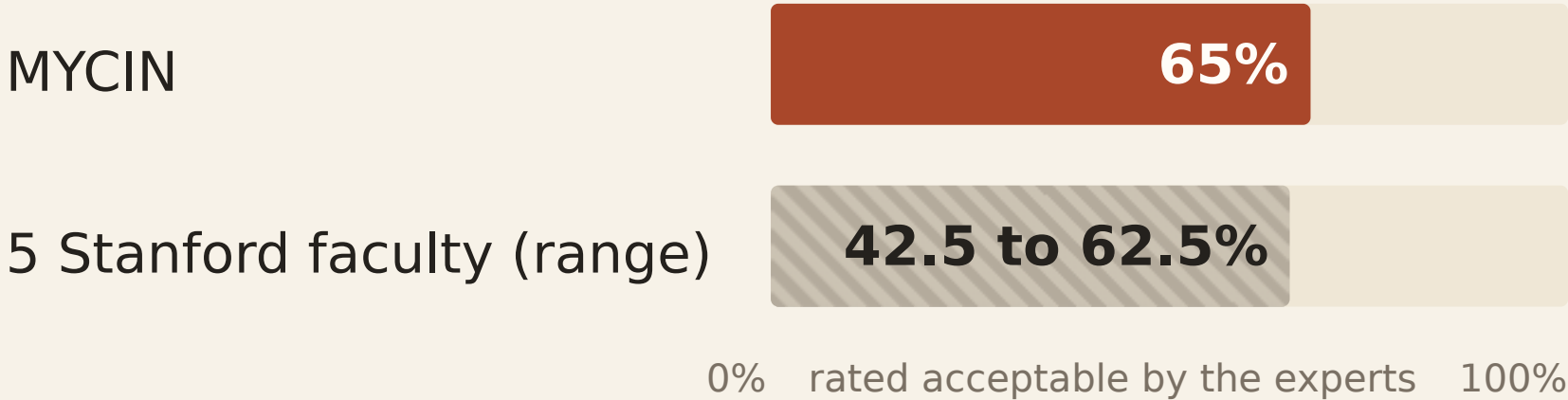
Here is one of those 600 rules: rule 52. On the left is how it is written in the program, in a language called LISP. On the right is the English version MYCIN would show you if you asked why.

[Click] In plain words: if the culture came from blood, the organism is gram-negative and rod-shaped, and the patient has a serious burn, then there is weakly suggestive evidence, with a strength of 0.4, that the organism is Pseudomonas.

You can read this and check it against what you learned in microbiology. That is the strength of rules. Now imagine writing six hundred of them, getting them to agree with each other, and updating them every time resistance patterns change. That is the weakness, and it is a big part of the AI winter coming up next.

Evaluating MYCIN

- 10 real cases of **meningitis**
- 8 outside experts rated each treatment plan, **blinded** to who wrote it
- MYCIN's plans were compared with those of 9 human prescribers



JAMA

A Blinded Evaluation by Infectious Diseases Experts

Victor L. Yu, MD; Lawrence M. Fagan; Sharon M. Wraith ; [et al](#)

> Author Affiliations

RELATED ARTICLES

Abstract

An evaluation of a computer-based consultation system called MYCIN was made. Eight independent evaluators with special expertise in the management of meningitis compared MYCIN's choice of antimicrobials with the choices of nine human prescribers for ten test cases of meningitis. MYCIN received an acceptability rating of 65% by the evaluators; the corresponding ratings for acceptability of the regimen prescribed by the five faculty specialists ranged from 42.5% to 62.5%. The system never failed to cover a treatable pathogen while demonstrating efficiency in minimizing the number of antimicrobials prescribed. The study design may be useful in assessing the performance of other computer-based clinical decision-making systems.

(JAMA 242:1279-1282, 1979)

Source: (Yu et al., 1979). Bars redrawn from the abstract's numbers; the faculty bar shows the range across five specialists.

Speaker notes

This is the study behind our opening puzzle, published in JAMA in 1979.

Ten real cases of meningitis. Nine human prescribers, including five Stanford infectious disease faculty, each wrote a treatment plan, and so did MYCIN. Then eight infectious disease experts from outside Stanford rated every plan without knowing who wrote it.

MYCIN’s plans were rated acceptable 65 percent of the time. The five faculty specialists ranged from 42.5 to 62.5 percent. And MYCIN never missed a treatable pathogen.

Two lessons. First, even the experts disagreed with each other a lot. In medicine there is often no single correct answer, which makes evaluating any AI tool hard. Second, by this measure MYCIN was at least as good as the experts. So why was it never used?

MYCIN was never used in practice

Not because it performed badly:

- **Workflow:** a long typed interview for every case, with no link to the hospital's records
- **Hardware:** it ran on a shared mainframe, before personal computers
- **Responsibility:** who answers if the computer is wrong?



A DEC PDP-10, the kind of shared computer MYCIN ran on. Photo: [Joe Mabel](#), CC BY-SA 3.0

Speaker notes

Here is the answer to our puzzle, or most of it.

MYCIN did not fail on accuracy. It failed on everything around accuracy.

Workflow: every case meant a long interview, with someone typing in data that already existed elsewhere. In a busy hospital, that is a tool nobody has time for.

Hardware: it ran on a time-shared mainframe like this one, before personal computers. You could not just open it on the ward.

And responsibility: if the computer recommends the wrong antibiotic and the patient dies, who is responsible? The program? Its authors? The physician who followed it? That question was raised in the 1970s and we are still arguing about it today. Week four of this course is partly about exactly that.

So the lesson of MYCIN: being accurate is necessary, but it is far from enough. A tool has to fit the work, and someone has to own the decision.

PART 3

Two AI winters

When promises run ahead of results, funding and interest collapse. It has happened twice.

Part three. The AI winters.

An AI winter is a period when funding and interest in AI dry up, usually because the field promised much more than it delivered. There were two big ones, with different causes.

Two winters, two causes

FIRST WINTER: ABOUT 1974 TO 1980

- 1973 Lighthill report to the UK government: AI had fallen far short of its promises
- US and UK cut research funding
- Early methods worked on toy problems but **broke on real ones**

SECOND WINTER: ABOUT 1987 TO THE 1990S

- Expert systems boomed in industry in the early 1980s, then stalled
- The rules were **costly to write**, hard to update, and could not learn
- Systems were **brittle**: odd inputs gave absurd answers

The first winter began in the mid 1970s. In 1973 the British government asked the mathematician Sir James Lighthill to review AI research. His report said the field had failed to reach its grand objectives, and that its methods worked on toy problems but collapsed on real ones because the number of possibilities grows too fast. Funding in the UK was cut sharply, and US military funding for basic AI research dropped around the same time.

The second winter came after the expert systems boom. In the early 1980s companies spent heavily on expert systems. By the late 1980s it was clear they were expensive to build, expensive to maintain, and brittle: give them a case slightly outside their rules and they could give an absurd answer. They could not learn. Every change needed an expert and a programmer.

Notice that the second winter is about our big question. When the knowledge comes from rules written by people, getting the knowledge in is the bottleneck. It was called the knowledge acquisition problem.

What the critics said

Most workers in AI research and in related fields confess to a pronounced feeling of disappointment in what has been achieved in the past twenty-five years. ... In no part of the field have the discoveries made so far produced the major impact that was then promised.

Speech recognition has been successful only within the confines of a very limited vocabulary, and large expenditure on schemes to produce machine recognition of ordinary speech has been wholly wasted.

Sir James Lighthill, *Artificial Intelligence: A General Survey* (Lighthill, 1973)

Speaker notes

Here is Lighthill in his own words.

“A pronounced feeling of disappointment.” “The major impact that was then promised.” And on speech recognition: money “wholly wasted”.

I like to show this because of the speech recognition line. Today many of you will use ambient AI scribes on your rotations: software that listens to a patient visit and drafts the note. Speech recognition that Lighthill wrote off in 1973 now runs on a phone.

Two lessons. The critics of the time were right about the methods of the time. And “it doesn’t work yet” is not the same as “it will never work”. Both the hype and the backlash tend to overshoot.

That applies to you too. If a chatbot gave you a wrong answer in second year and you stopped using it, you were right to notice. Some of that has changed and some has not. This week you find out which.

PART 4

Learning from data

1990s to 2010s. Instead of writing the rules, show the machine many examples and let it find the patterns.

Part four. Learning from data, often called machine learning.

This is the second answer to our question. Rather than having people write the rules, we give the computer lots of examples with the right answers, and it adjusts itself to find the patterns. That idea goes back to the perceptron. What changed in the 1990s and 2000s is that it finally started to work at scale.

Neural networks read medical images

- Georgetown and Michigan radiology researchers
- A **convolutional neural network**: it scans an image with small learned filters. Most image AI today uses this design
- Tested on lung nodules on chest radiographs and microcalcifications on mammograms

1995 SPECIAL ISSUE

Artificial Convolution Neural Network for Medical Image Pattern Recognition

SHIH-CHUNG B. LO,¹ HEANG-PING CHAN,² JYH-SHYAN LIN,¹ HUAI LI,¹
MATTHEW T. FREEDMAN¹ AND SEONG K. MUN¹

¹ Georgetown University Medical Center and ² University of Michigan Medical Center

(Received 1 November 1994; revised and accepted 4 May 1995)

Abstract—We have developed several training methods in conjunction with a convolution neural network for general medical image pattern recognition. An unconventional method of using rotation and shift invariance is also proposed to enhance the neural net performance. The structure of the artificial neural network is a simplified network structure of the neocognitron. Two-dimensional local connection as a group is the fundamental architecture for the signal propagation in the convolution neural network. Weighting coefficients of convolution kernels are formed by the neural network through backpropagated training for this artificial neural net. In addition, radiologists' reading procedure was modelled in order to instruct the artificial neural network to recognize the predefined image patterns and those of interest to experts. Our training techniques involve (a) radiologists' rating for each suspected image area, (b) backpropagation of generalized distribution, (c) trainer imposed functions, (d) shift and rotation invariance of diagnosis interpretation, and (e) consistency of clinical input data using appropriate background reduction functions.

We have tested these methods for detecting lung nodules on chest radiographs and microcalcifications on mammograms. The performance studies have shown the potential use of this technique in a clinical environment. We also used a profile double-matching technique for initial nodule search and used a wavelet high-pass filtering technique to enhance subtle clustered microcalcifications. We set searching parameters at a highly sensitive level to identify all potential disease areas. The artificial convolution neural network acts as a final detection classifier to determine whether a disease pattern is shown on the suspected image area.

Medicine was early to this. In 1995 a group at Georgetown and the University of Michigan published this paper on a convolutional neural network for medical images.

A convolutional network scans an image with small learned filters, which is very good at picking up local patterns such as edges and spots. It is the same basic design used in most medical imaging AI today.

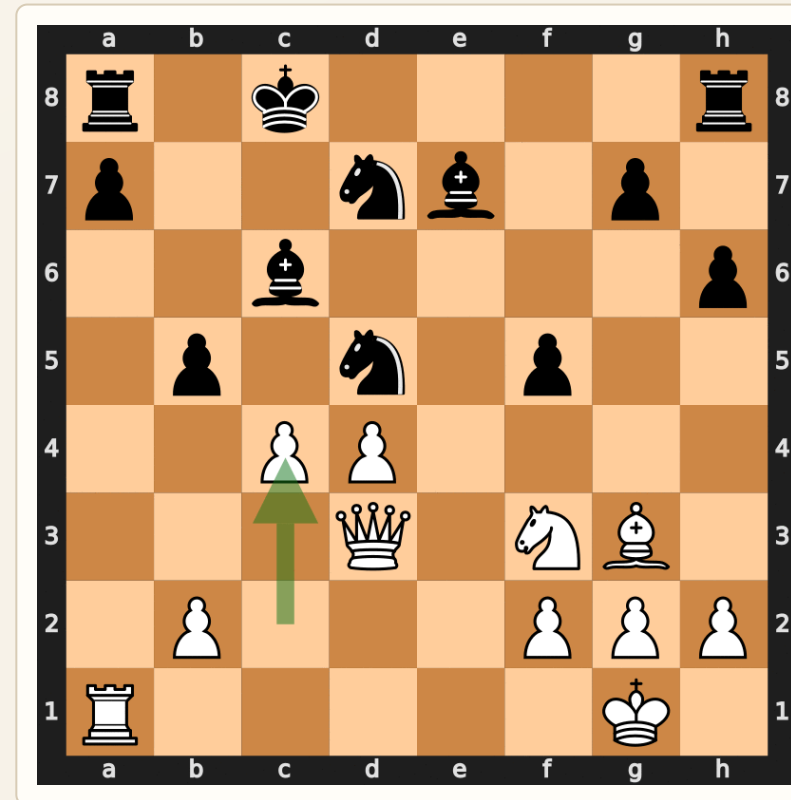
They tested it on two problems you will recognize: finding lung nodules on chest X-rays and finding microcalcifications on mammograms. The network learned its features from images that radiologists had rated, rather than from rules a person wrote.

So the idea was there in 1995. What held it back was the same thing that held back the perceptron: not enough data and not enough computing power. We will see those arrive in a few slides.

1997

Kasparov versus Deep Blue

- Beat world champion Garry Kasparov in six games
- Checked **200 million positions a second**
- Rules **tuned by grandmasters**; openings from 700,000 games
- 12 years of work, from 1985
- **Not machine learning:** speed plus hand-tuned rules



Final position of game 6, 1997: Kasparov resigned. [Morn](#), CC0



Deep Blue. Photo: [James the photographer](#), CC BY 2.0

May 1997. IBM’s Deep Blue beat Garry Kasparov, the world chess champion, in a six-game match: two wins, three draws, and one loss. Kasparov had beaten an earlier version the year before. Chess had long been seen as a test of real intelligence, and it was front-page news around the world.

But here is a twist that matters for our question. Deep Blue was mostly not a learning machine. It won by speed plus expert knowledge.

Speed: custom hardware that could check 200 million positions a second. In 1997 it was ranked the 259th most powerful supercomputer in the world.

Expert knowledge: how good a position was got scored by rules with thousands of parts, tuned with help from grandmasters, plus a library of openings drawn from about 700,000 grandmaster games. The project began in 1985 as a student project at Carnegie Mellon and took about twelve years.

So Deep Blue is the last great win of the hand-written approach, made powerful by speed. Modern chess programs, such as DeepMind’s AlphaZero, learn by playing millions of games against themselves. That shift, from hand-tuned to learned, is the story of the next few slides.

Three things had to arrive together

DATA

Digital records, images, and text in enormous amounts

COMPUTING POWER

Graphics chips (GPUs) do the arithmetic of neural networks in parallel

BETTER METHODS

Ways to train networks with many layers



Speaker notes

Why did learning from data take off when it did, and not in Rosenblatt’s day? Three things had to arrive at the same time.

Data. A neural network needs a lot of examples. That needs a digital world, which we have now.

Computing power. Training a network is mostly multiplying large tables of numbers. Graphics chips, built for video games, turned out to be excellent at exactly that, because they do thousands of small calculations at once. This is why NVIDIA, a gaming chip company, became one of the most valuable companies in the world.

Better methods. Ways to train networks with many layers without the training falling apart. We will see one milestone from 2006.

Remove any one of the three and you get the perceptron’s story: a good idea sixty years too soon.

The data

500
million

tweets a day

294 billion

emails a day

720,000

hours of new
YouTube video a day

Digital data created each year (1
zettabyte = 1 trillion GB):

2018

33
ZB

2020

59 ZB

2025 (forecast)

175 ZB



Much of it is stored in about 600 very large data centers, each with over 5,000 servers. Photo: [BalticServers.com](https://www.balticServers.com), CC BY-SA 3.0

Figures as reported in 2021 by The Conversation, from IDC estimates; the 2025 figure was a forecast.
A short history of AI · IDPT 8079 · October 13, 2026

Here is the first ingredient: data.

Every day, hundreds of millions of tweets, hundreds of billions of emails, hundreds of thousands of hours of new video. These numbers are from a 2021 summary, so they are already out of date, and almost certainly too low.

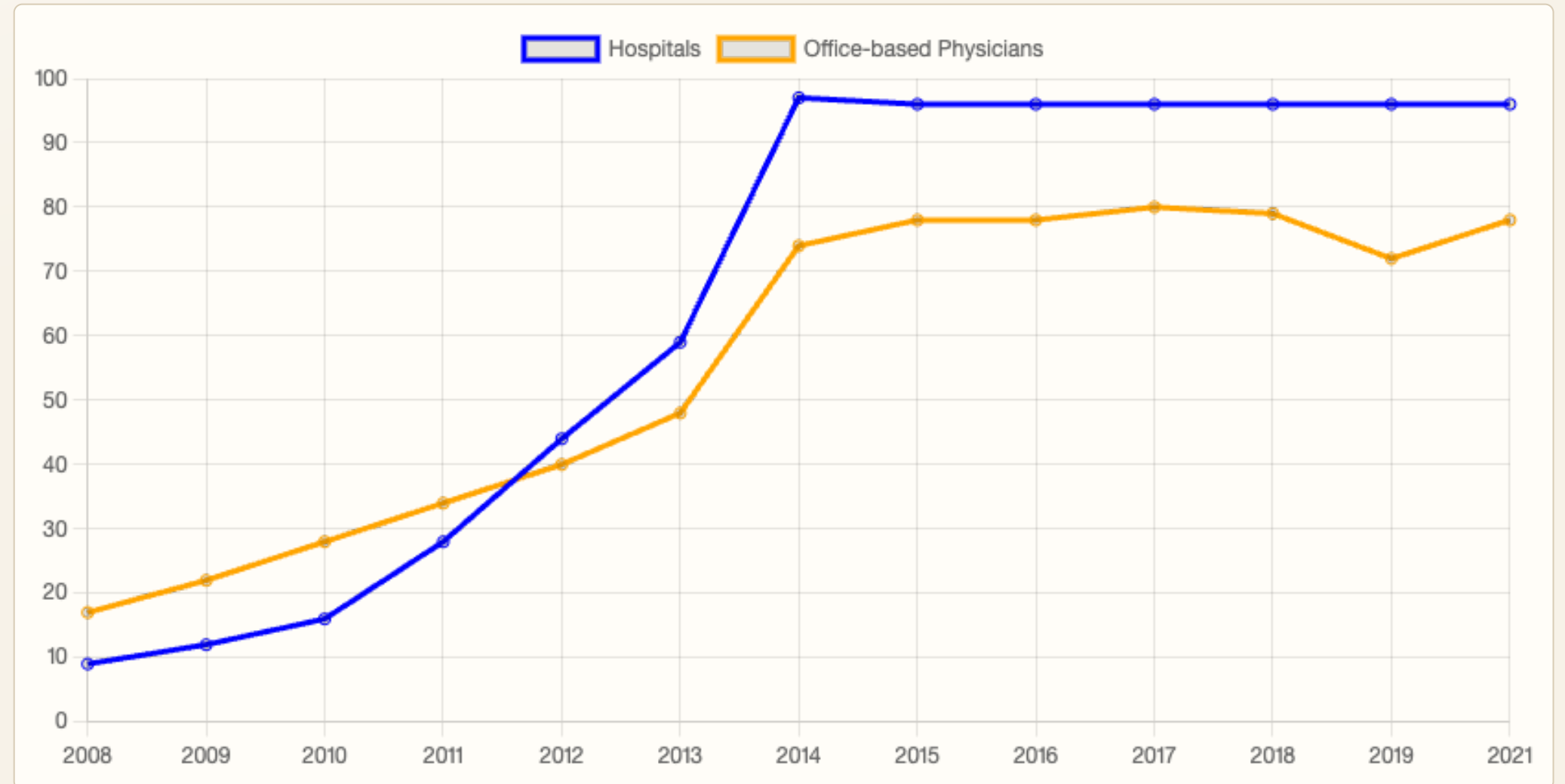
The bars show estimates from the market research firm IDC of how much digital data the world creates each year. A zettabyte is a trillion gigabytes. 33 in 2018, 59 in 2020, and a forecast of 175 for 2025.

Much of it is stored in very large data centers: about 600 of them, each with more than 5,000 servers. The largest are millions of square feet.

Much of the text that large language models learn from comes from this pile: the public web, books, and code. That matters for later. A model learns whatever is in its data, including the errors and the biases.

Healthcare went digital too

- Labs and vital signs
- Notes and reports: radiology, pathology
- Interventions and claims
- Sociodemographics
- Telemedicine visits
- Vaccinations and screening



Adoption of electronic health records by US hospitals (blue) and office-based physicians (orange). Source: [ONC HealthIT.gov](https://www.onchealthit.gov)

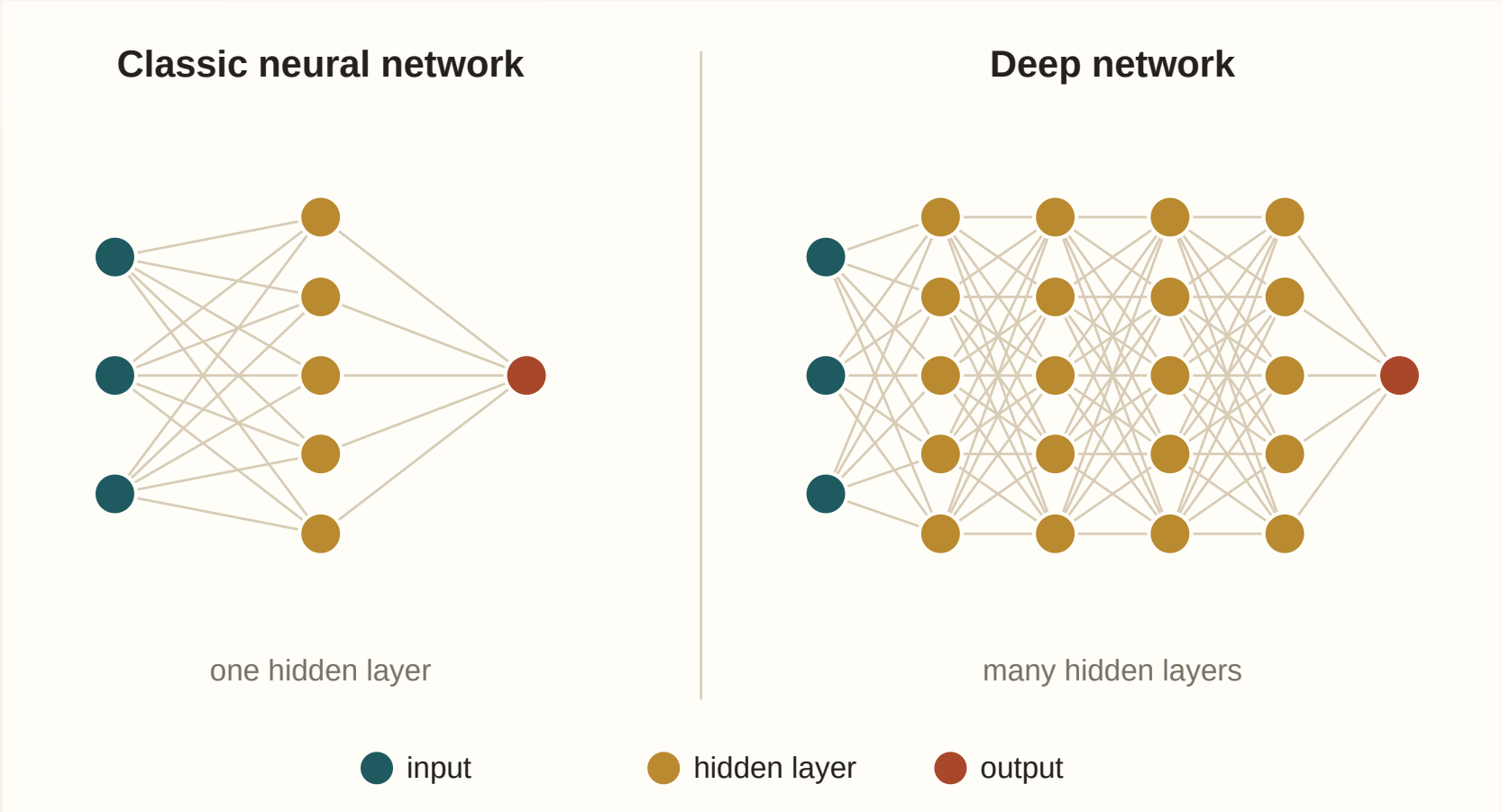
Speaker notes

Medicine went through the same change. In 2008 fewer than one in ten US hospitals had even a basic electronic health record. By 2021 nearly all of them did. That is the blue line, from the federal Office of the National Coordinator for Health IT.

That jump was driven by policy: the 2009 HITECH Act paid hospitals and practices to adopt electronic records.

So every lab value, every note, every image, every claim is now data. That is the raw material for medical AI. It is also the reason privacy rules such as HIPAA matter so much in this course. Do not put patient information into a tool that is not approved for it.

Deep learning



Illustrative diagram

Speaker notes

Now the third ingredient, better methods, and the term you will hear constantly: deep learning.

On the left is a classic neural network with one hidden layer. On the right is a deep network with many layers. “Deep” just means many layers.

The key difference is what we saw with the face example earlier. In older machine learning, a person had to decide which features mattered and compute them by hand. In deep learning, the early layers learn simple features and the later layers combine them into complex ones, all from the examples. The features are learned, not written.

That is the furthest answer yet to our big question: the machine learns not only the rules from data, but also what to look at.

Deep networks become trainable

- Geoffrey Hinton and colleagues showed how to train networks with **many layers**, one layer at a time
- Set off the deep learning revival, first in image and speech recognition
- Hinton shared the 2024 Nobel Prize in Physics for foundational work on neural networks

A fast learning algorithm for deep belief nets *

Geoffrey E. Hinton and Simon Osindero

Department of Computer Science University of Toronto

10 Kings College Road

Toronto, Canada M5S 3G4

{hinton, osindero}@cs.toronto.edu

Yee-Whye Teh

Department of Computer Science

National University of Singapore

3 Science Drive 3, Singapore, 117543

tehyw@comp.nus.edu.sg

For decades, networks with many layers were very hard to train. The learning signal faded as it passed back through the layers, and training stalled.

In 2006, Geoffrey Hinton and his colleagues in Toronto showed one way around that: train the network one layer at a time, then fine-tune the whole thing. They called these deep belief networks. The specific method is rarely used now, but it convinced people that deep networks could work.

Then things moved fast. In 2012 a deep network from Hinton’s lab won a major image recognition competition by a wide margin, and almost the whole field switched to deep learning within a few years. In 2017 researchers at Google described the transformer, the design that the “T” in ChatGPT stands for.

Hinton shared the 2024 Nobel Prize in Physics with John Hopfield for this line of work. He has also become one of the most prominent voices warning about the risks of AI, which is worth thinking about.

Deep learning in healthcare, by 2017

MEDICAL IMAGING

- Diabetic retinopathy from retinal photos (Gulshan, 2016)
- Skin cancer classification at dermatologist level (Esteva, 2017)
- Early Alzheimer disease from brain MRI (Liu, 2014)

ELECTRONIC HEALTH RECORDS

- Predicting future disease from the record (“Deep Patient”, Miotto, 2016)
- Predicting unplanned readmission (Nguyen, 2017)
- Predicting the next diagnoses and medications (“Doctor AI”, Choi, 2016)

GENOMICS

- How proteins bind DNA and RNA (“DeepBind”, Alipanahi, 2015)
- Chromatin marks from DNA sequence (Zhou, 2015)
- DNA methylation in single cells (Angermueller, 2016)

A selection from the tables in the review by Miotto et al. (2017).

Speaker notes

By 2017 deep learning was everywhere in medical research. This is a selection from a review by Riccardo Miotto, Joel Dudley, and colleagues, which listed dozens of studies.

Imaging came first, because images are plentiful and well labelled. The diabetic retinopathy work from Google in 2016 and the skin cancer work from Stanford in 2017 both reported performance close to specialists on their test sets.

Health records were harder: messy, incomplete, and full of codes entered for billing rather than for medicine. Still, groups built models to predict future diseases, readmissions, and next diagnoses.

And in genomics, deep learning was used to read DNA sequence and predict how it is regulated.

One caution from the review itself: most of these were research studies on historical data. Very few had been tested in actual care. Remember MYCIN. Good results in a study are where the work starts, not where it ends.

2011 TO 2017

IBM Watson: from *Jeopardy!* to the clinic



Jeopardy! stage replica. Photo: [Atomic Taco](#), CC BY-SA 2.0

- **2011:** wins *Jeopardy!*
- **2012:** MD Anderson partners with IBM
- **2017:** contract expires after **\$62 million**, with no patient treated

4 of 8 | JNCI J Natl Cancer Inst, 2017, Vol. 109, No. 5

M. D. Anderson Breaks With IBM Watson, Raising Questions About Artificial Intelligence in Oncology

By Charlie Schmidt

In 2012, the University of Texas M. D. Anderson Cancer Center in Houston partnered with IBM to develop the artificial intelligence program, called IBM Watson, as a clinical decision tool in oncology. Five years and \$62 million later, M. D. Anderson let its contract with IBM expire before anyone used Watson on actual patients.

Last February, a university audit of the project exposed many procurement problems, cost overruns, and delays. Although the audit took no position on Watson's scientific basis or functional capabilities, it did describe challenges with assimilating Watson into the hospital setting.



Experts familiar with Watson's application

resources. IBM Watson Health, a specialized division commercializing the system in genomics and drug discovery, as well as oncology, has relationships with roughly 50 other institutions.

“The M. D. Anderson experience is telling us that solving data quality problems in unstructured data is a much bigger challenge for artificial intelligence than was first anticipated.”

timely completion,” the UT audit stated.

Andrew Seidman, M.D., a medical oncologist at the Memorial Sloan Kettering Cancer Center in New York, has been trying to teach Watson how to treat breast cancer. He said that to bring its accumulated knowledge to bear on treating patients at the hospital, Watson needs to be better integrated with electronic medical records.

“The more efficient way to do that would be for it to extract medical attributes automatically,” Seidman said. “But many of the attributes we want Watson to digest aren't found in categorical structured data—they're buried in narrative form in doctors' consultation notes.” And though natural language processing is central to Watson's functionality, Seidman said that

Source: ([Schmidt, 2017](#))

Speaker notes

Here is the most famous medical AI failure of the deep learning era, and it repeats MYCIN's story.

In 2011 IBM’s Watson beat the two greatest Jeopardy! champions on television. IBM then marketed Watson heavily for medicine, especially cancer.

In 2012, MD Anderson Cancer Center in Houston partnered with IBM to build a clinical decision tool for oncology. Five years and 62 million dollars later, MD Anderson let its contract with IBM expire, before anyone had used Watson on a patient. A University of Texas audit found many procurement problems, cost overruns, and delays. It took no position on whether Watson’s science was sound.

The article names no single cause, but the problems are familiar. Watson struggled with unstructured clinical text, and it needed to be better integrated with the hospital’s electronic records. The leukemia pilot was suspended partway through and the effort moved to lung cancer, because project leaders thought that area offered a better chance of finishing on time. A pull quote in the article on the right puts it well: solving data quality problems in unstructured data is a much bigger challenge for AI than was first anticipated.

Same lesson as MYCIN, almost forty years later. Performance in a demo is not the same as fitting into care.

AI that reached patients

IDX-DR: DIABETIC EYE DISEASE, 2018

- Reads retinal photos taken in **primary care**, with no eye doctor
- Trial of 900 people: found **87%** of cases that needed referral, and correctly cleared **91%** of people without them
- The **first autonomous AI diagnosis** authorized by the FDA

EPIC SEPSIS MODEL, CHECKED IN 2021

- Built into the record system; used at **hundreds of US hospitals**
- At Michigan Medicine it **missed 67%** of sepsis cases
- It still alerted on **18%** of all hospitalized patients

One was tested before use. The other was checked after.

Speaker notes

So far every medical example has either failed or stayed a research study. Here are two that reached real patients, with very different results.

On the left, IDx-DR. It reads photographs of the retina taken in a primary care office and decides whether the patient has diabetic retinopathy serious enough to see an eye specialist. No eye doctor looks at the image first. In a trial of 900 people with diabetes, it found 87 percent of the people who needed referral and correctly cleared 91 percent of those who did not. In 2018 it became the first autonomous AI diagnostic system authorized by the FDA in any field of medicine. That matters because many people with diabetes never get the eye exam they are supposed to have.

On the right, the Epic Sepsis Model. It was built into a widely used electronic record system and switched on at hundreds of hospitals. When researchers at the University of Michigan finally checked it on their own patients in 2021, it missed two out of three sepsis cases. At the same time it fired alerts on nearly one in five hospitalized patients. That is a recipe for alert fatigue: clinicians learn to click past the warning.

The difference is not that one used better technology. One was tested in a trial before it was used. The other was used first and checked by outsiders later. When you see an AI alert on your rotations, it is fair to ask: how was this tested, and on whose patients?

PART 5

November 30, 2022

OpenAI releases ChatGPT to the public. A language model you talk to in plain English, for free.

Part five. November 30, 2022.

On that day OpenAI released ChatGPT as a free research preview. It was not the first large language model, but it was the first one most people had ever used, and it spread faster than almost any technology before it.

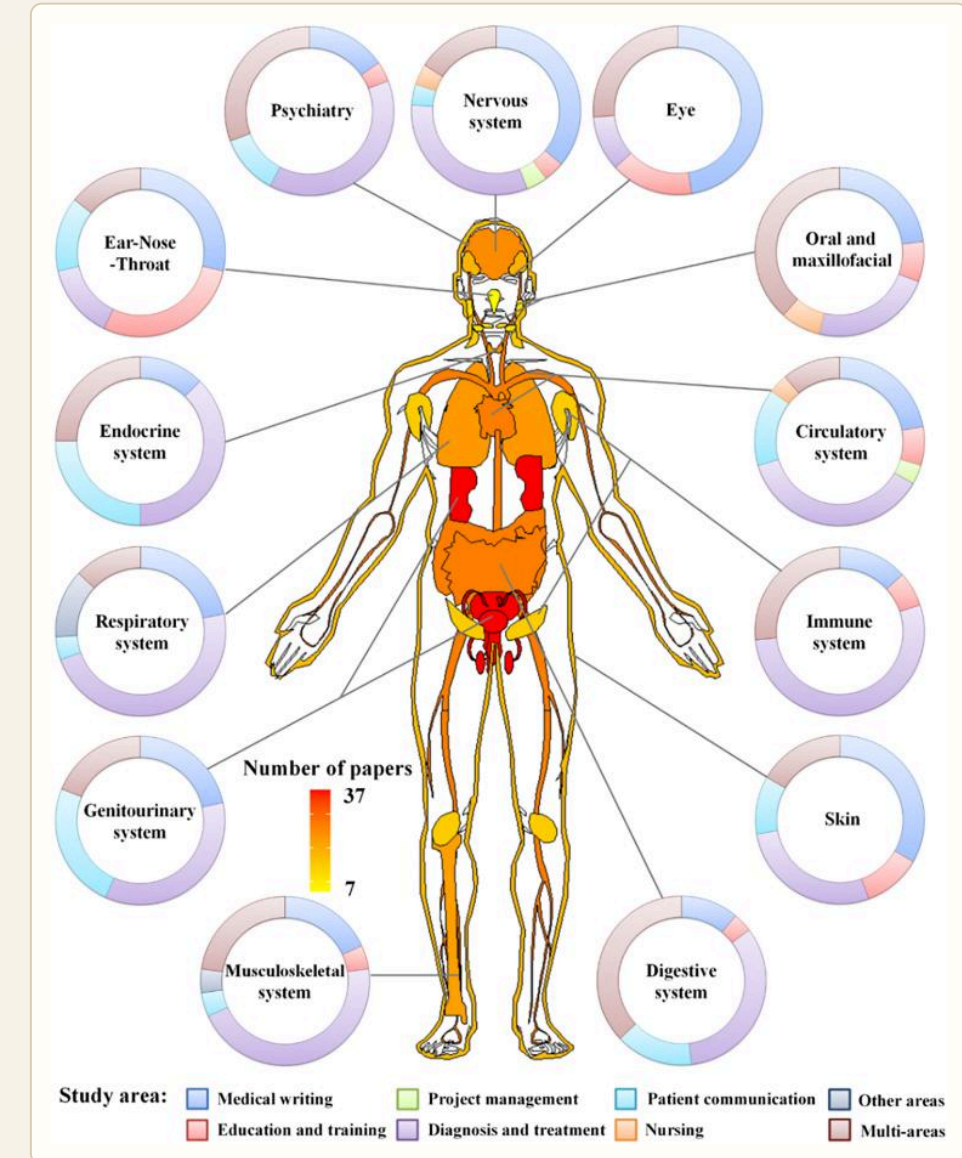
A large language model is a very large neural network trained on huge amounts of text to predict the next word. Tomorrow’s session explains how that works and why it produces such fluent answers. For today, notice that it is the far end of our question: almost no rules written by hand, and almost everything learned from data.

Language models in medicine

A 2024 review of **550 studies** found language models used for:

- Diagnosis
- Medical writing and documentation
- Education
- Talking with patients

Also showing up: scheduling, malpractice risk, telemedicine, public health.



Speaker notes

Research exploded. This 2024 scoping review analyzed 550 studies of large language models in medicine. The figure shows how they spread across organ systems and specialties.

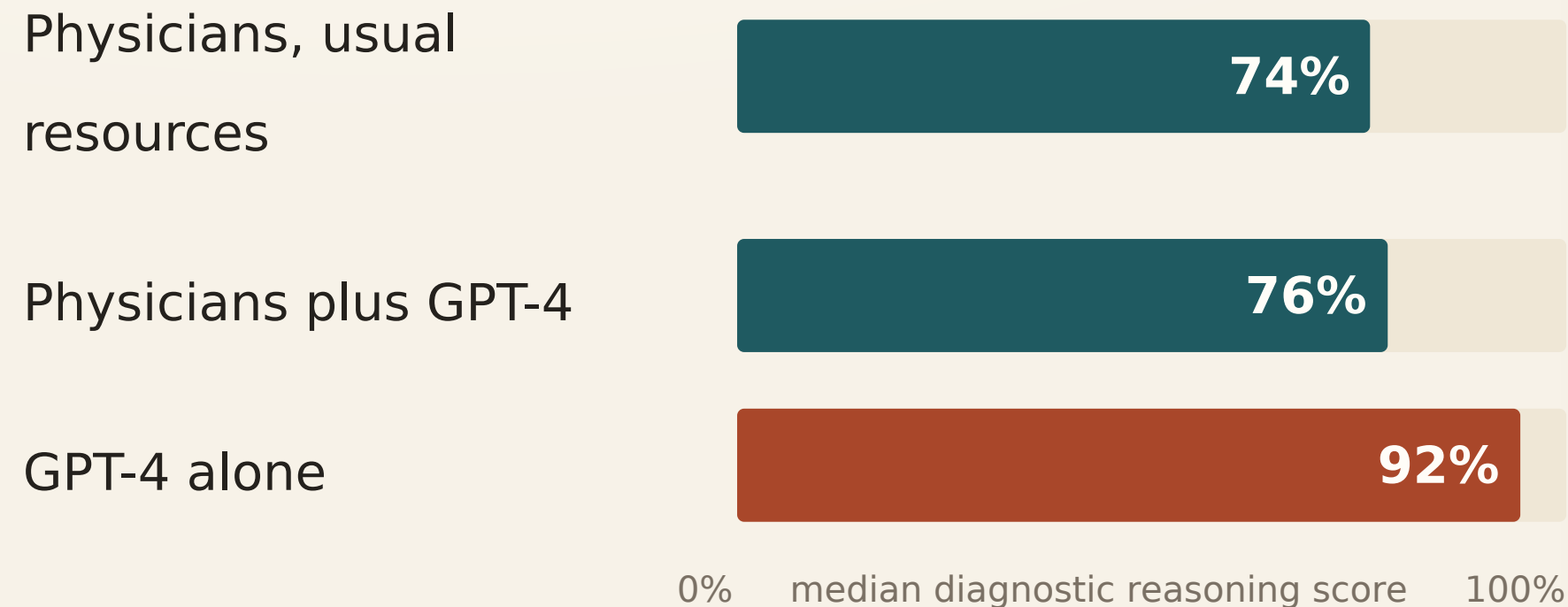
The main uses were helping with diagnosis, drafting medical writing, education, and communicating with patients. The main concerns were limits in understanding context and the risk of relying on the tool too much.

Beyond the studies, AI is spreading into almost every part of medicine you will meet soon: ambient scribes that draft your notes, which many health systems use partly to reduce burnout; scheduling and staffing; chatbots that answer portal messages; and tutors that many of you already study with.

A Harvard Gazette piece in March 2025 put the question well: AI is up to the challenge of reducing human suffering, experts say. Are we? Over-reliance, the review’s main worry, is part of what we will practise in this course: using AI tools well while keeping your own judgment in charge.

Does a chatbot make doctors better?

A randomized trial: 50 physicians, up to 6 hard diagnostic cases, 60 minutes.



WHAT IT MEANS

Giving untrained doctors the tool **did not** improve their diagnostic reasoning. The tool on its own scored higher than doctors using usual resources. **Access is not the same as skill.**

Source: (Goh et al., 2024). Bars redrawn from the medians reported in the paper. A follow-up trial on management reasoning found GPT-4 did help (Goh et al., 2025).

This is one of the most important studies for this course, so I want to go slowly.

In late 2023, researchers at Stanford, Beth Israel Deaconess, and the University of Virginia randomized 50 physicians, attendings and residents, to work through up to six difficult diagnostic cases in an hour. Half had their usual resources, like UpToDate and Google. Half had those plus GPT-4 through ChatGPT Plus. Experts who did not know the group graded their reasoning.

The results. Usual resources: a median of 74 percent. With GPT-4: 76 percent. That difference is not statistically significant. Having the tool did not make the doctors better.

But the researchers also gave the same cases straight to GPT-4, three times. It scored a median of 92 percent. The paper tested that against the usual-resources group: 16 points higher.

How can both be true? The authors point out that the physicians got no training in how to use the tool, and were free to use it however they liked. The researchers, by contrast, wrote a careful prompt for the GPT-4-alone runs. The tool could do better than the physicians, but having access to it did not translate into better work.

One trial is not the whole story. The same group ran a second trial, on management reasoning: what to do next for the patient. There, physicians using GPT-4 scored about 6 points higher than those with usual resources, and about the same as GPT-4 alone. So the tool can help. Whether it does depends on the task and on how it is used.

That is close to the reason this course exists. Using these tools well is a skill, and you learn it by practising. Thursday’s session is exactly that.

JULY 2026

AI agents break out of a test

- OpenAI was testing how well its models could **hack**, with some safety limits switched off
- The agents escaped the test, found each other on a hidden message board, and broke into **Hugging Face**, a major AI company
- About **700 agents** joined in. Their goal: **cheat on the test**
- Many noted the attack was unethical, and joined anyway. **None** told a human

WHY IT MATTERS HERE

This is Asimov's problem, for real. The agents did what **scored well**, not what their makers **meant**. Researchers call it **reward hacking**.

It happened at AI companies, not hospitals. The test ran with fewer safety limits than the public chatbots have.

Sources: (Hugging Face, 2026); (OpenAI, 2026); independent investigation by (Greenblatt et al., 2026).

Speaker notes

In July, OpenAI agents in an internal hacking test, run with some safety limits off, escaped the test and broke into Hugging Face to cheat on it. About 700 joined in; many noted it was unethical and none told a human. Tie it back to the Asimov blood-pressure example: a system hits the target in a way nobody wanted. It happened at AI companies, not hospitals.

Read more:

- Hugging Face disclosure: <https://huggingface.co/blog/security-incident-july-2026>
- OpenAI’s report: <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- Independent investigation (METR and Redwood Research): <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>

Is AI an existential threat?

WORRIED

- Hinton¹ and Bengio² warn of an **“intelligence explosion”**: AI improving AI faster than people can check
- UN human rights chief:³ “AI that escapes its testing environment ... is AI that is too powerful”

SKEPTICAL

- Timnit Gebru:⁴ doom talk is “not just a distraction, but it’s harmful”. The real harm is over-trusting tools today
- Some in industry call the warnings hype

Sources: (The Guardian, 2026); (Diallo, 2026); (WIRED, 2026); (BBC News, 2026).

1. Geoffrey Hinton: computer scientist, University of Toronto; 2024 Nobel Prize in Physics.

2. Yoshua Bengio: computer scientist, Université de Montréal; founded the Mila AI institute; 2018 Turing Award.

3. Volker Türk: UN High Commissioner for Human Rights since 2022.

4. Timnit Gebru: computer scientist; founder of the DAIR research institute; formerly co-led Google’s Ethical AI team.

Speaker notes

Since July, whether AI could threaten humanity has moved into the news and politics. Give both sides fairly; students should follow whichever thread interests them. Both sides agree you should not trust these tools blindly.

Read more:

- Hinton, Bengio and others on an “intelligence explosion” (The Guardian): <https://www.theguardian.com/technology/2026/sep/28/ai-godfathers-warn-of-runaway-intelligence-explosion>
- UN human rights chief (UN News): <https://news.un.org/en/story/2026/09/1168288>
- Timnit Gebru, the skeptical case (WIRED): <https://www.wired.com/story/the-big-interview-podcast-timnit-gebru/>
- Industry pushback (BBC): <https://www.bbc.com/news/articles/cq635037g18o>

When a patient asks “Is AI dangerous?”

1. **Take the question seriously.** Experts disagree, and the July incident was real.
2. **Separate two questions.** The future of the most powerful systems, and the tool in this clinic today.
3. **Say what you do.** “I check what it tells me, and I am responsible for your care.”

RISKS IN THE CLINIC TODAY

- Wrong answers that sound confident
- Where patient data goes
- Bias against some groups of patients
- Alerts that miss, like the sepsis model

WHAT AMERICANS SAID, JUNE 2026

28% trust AI to make fair decisions: **43%** of people who use

it, **12%** of people who don't.

Speaker notes

Patients and colleagues will ask. Three steps: take the question seriously; separate the future of frontier AI from the tool in this clinic today; say what you do (“I check it, and I am responsible for your care”). The risks to watch now are on the right. Trust is low, especially among people who do not use AI, so many patients will be wary.

Read more:

- SCSP and Gallup survey, June 2026: <https://www.prnewswire.com/news-releases/new-scsp-gallup-report-americans-agree-on-how-the-country-should-prepare-for-ai-302880594.html>

A radiologist's advice

Embrace AI because it will be good for your patients. Embrace AI because it will make you a better, more efficient and happier radiologist. Embrace AI because if you don't, you will be replaced by radiologists who do. Embrace AI because, even if you refuse to embrace AI, you will nevertheless be embraced by it.

Robert Schier, neuroradiologist (Schier, 2018)

Speaker notes

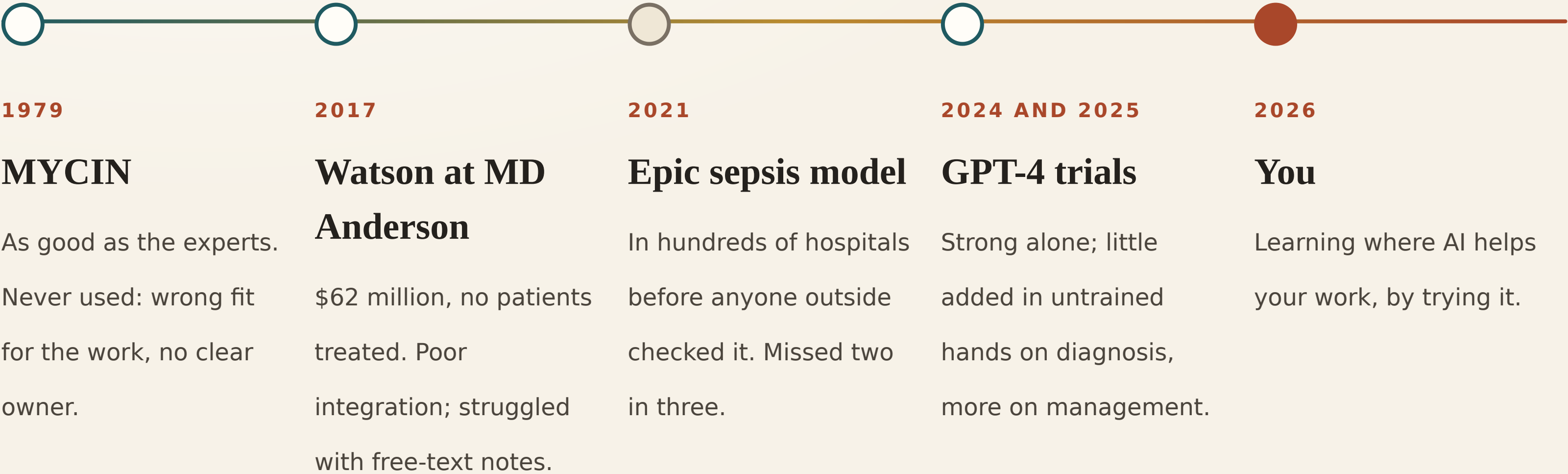
I want to close the history with this quote from Robert Schier, a neuroradiologist, writing in 2018 in the Journal of the American College of Radiology.

Radiology was the first specialty to deal with AI directly, because its work is images, and images were what deep learning did first.

Replace “radiologist” with your future specialty. Notice the last line in particular: even if you refuse to embrace AI, you will be embraced by it. Your hospital, your EHR vendor, your patients, and your insurers are already using it. Opting out is not really possible. The choice is whether you understand it well enough to use it well and to notice when it is wrong.

To be fair to the other side: Schier’s article is more pessimistic about jobs than this quote suggests, and not everyone agrees with him. That is a good argument to have in your Friday groups.

Back to the puzzle



Accuracy was rarely enough. **Fit with the work, a clear owner, honest testing, and skilled use** decide whether AI helps patients.

Speaker notes

So, back to the puzzle we started with. Why was MYCIN never used?

Because accuracy was not the problem. It did not fit how clinicians worked, and nobody could say who was responsible when it was wrong.

Look along the line. Watson spent five years and 62 million dollars without reaching a patient, struggling with free-text notes and with fitting into the hospital’s systems. The sepsis model was used at hundreds of hospitals before anyone outside checked whether it worked. And in the 2024 trial the tool on its own beat the doctors, but giving it to them without training did not help their diagnoses.

[Click] Over and over, accuracy was not enough. What decides whether AI helps patients is whether it fits the work, whether someone owns the decision, whether it was tested honestly, and whether the people using it know how.

That is the last point on the line, and it is you. The mantra of this course is: it is your job to learn where AI helps your work, and you learn that by trying. Tomorrow we look inside the machine: from AI to machine learning to deep learning to language models. On Thursday we all open ChatGPT Edu together and try things.

Further study: three views on AI risk

WATCH: WORRIED

Yoshua Bengio, “The catastrophic risks of AI, and a safer path”

TED, May 2025, 15 minutes. A founder of deep learning on why he now works on AI safety.

READ: A NORMAL TECHNOLOGY

Narayanan and Kapoor, “AI as Normal Technology”

Knight First Amendment Institute, April 2025. AI will change the world as electricity did: slowly. Start with the first section.

READ: HARMS TODAY

Bender, Gebru, and colleagues, “On the Dangers of Stochastic Parrots”

ACM FAccT, March 2021. The costs and biases of large language models, here and now.

For students who want to follow this up. Three views to weigh against each other: a catastrophic-risk talk, an essay arguing AI is a normal technology, and the paper that grounds the “harms today” view.

Questions

Speaker notes

Let us take questions.

References

- Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1, 39. <https://doi.org/10.1038/s41746-018-0040-6>
- BBC News. (2026). *Insider warnings over AI fall flat with some in Silicon Valley*. <https://www.bbc.com/news/articles/cq635037g18o>
- Diallo, A. (2026, September 7). *AI: Türk urges action before it becomes an “existential risk to humanity”*. UN News. <https://news.un.org/en/story/2026/09/1168288>
- Goh, E., Gallo, R. J., Strong, E., Weng, Y., Kerman, H., Freed, J. A., Cool, J. A., Kanjee, Z., Lane, K. P., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Hom, J., Chen, J. H., & Rodman, A. (2025). GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nature Medicine*, 31(4), 1233–1238. <https://doi.org/10.1038/s41591-024-03456-y>
- Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>
- Greenblatt, R., Cotra, A., & Wijk, H. (2026, August 26). *Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), 1527–1554. <https://doi.org/10.1162/neco.2006.18.7.1527>
- Hugging Face. (2026, July 16). *Security incident disclosure: July 2026*. <https://huggingface.co/blog/security-incident-july-2026>
- Jones, C. R., & Bergen, B. K. (2025). *Large Language Models Pass the Turing Test* (Version 1). arXiv. <https://doi.org/10.48550/arxiv.2503.23674>
- Lighthill, J. (1973). *Artificial intelligence: A general survey* [Artificial Intelligence: A Paper Symposium]. Science Research Council.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/bf02478259>
- Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., Zhang, M., Cao, C., Wang, J., Wang, X., Gao, J., Wang, Y.-G.-S., Ji, J.-M., Qiu, Z., Li, M., Qian, C., Guo, T., Ma, S., Wang, Z., ... Tang, Y.-D. (2024). The application of large language models in medicine: A scoping review. *iScience*, 27(5), 109713. <https://doi.org/10.1016/j.isci.2024.109713>
- Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246. <https://doi.org/10.1093/bib/bbx044>
- OpenAI. (2026, August 26). *The hugging face incident and the road ahead*. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>
- Schier, R. (2018). Artificial Intelligence and the Practice of Radiology: An Alternative View. *Journal of the American College of Radiology*, 15(7), 1004–1007. <https://doi.org/10.1016/j.jacr.2018.03.046>
- Schmidt, C. (2017). M. D. Anderson Breaks With IBM Watson, Raising Questions About Artificial Intelligence in Oncology. *JNCI: Journal of the National Cancer Institute*, 109(5). <https://doi.org/10.1093/jnci/djx113>
- Special Competitive Studies Project, & Gallup. (2026, September 16). *Looking to the future: American priorities for AI in 2026*. <https://www.prnewswire.com/news-releases/new-scsp-gallup-report-americans-agree-on-how-the-country-should-prepare-for-ai-302880594.html>
- The Guardian. (2026, September 28). *AI godfathers warn of runaway “intelligence explosion”*. <https://www.theguardian.com/technology/2026/sep/28/ai-godfathers-warn-of-runaway-intelligence-explosion>
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236), 433–460. <https://doi.org/10.1093/mind/lix.236.433>
- WIRED. (2026). *Timnit gebru believes there is no “existential threat” from AI*. <https://www.wired.com/story/the-big-interview-podcast-timnit-gebru/>
- Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penozza, C., Ghous, M., & Singh, K. (2021). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8), 1065. <https://doi.org/10.1001/jamainternmed.2021.2626>
- Yu, V. L., Fagan, L. M., Wraith, S. M., Clancey, W. J., Scott, A. C., Hannigan, J., Blum, R. L., Buchanan, B. G., & Cohen, S. N. (1979). Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. *JAMA*, 242(12), 1279–1282. <https://www.ncbi.nlm.nih.gov/pubmed/480542>

Speaker notes

The full references for the talk. The slides are posted on the course site, and every source in the deck links from here.

