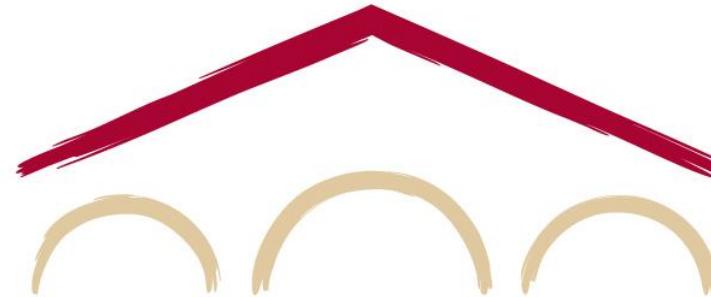


History of Natural Language Processing

CS224N/Ling284



Christopher Manning

Lecture 1

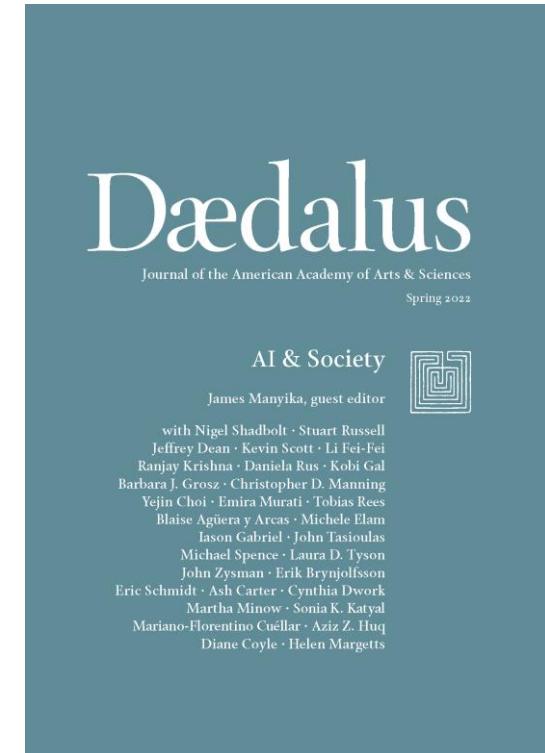
Christopher D. Manning: Human Language Understanding & Reasoning

Four eras of NLP

- **1940–1969:** Early Explorations
- **1970–1992:** Hand-built symbolic NLP systems, of increasing formalization
- **1993–2012:** Statistical or Probabilistic NLP and then more general Supervised ML for NLP
- **2013–now:** Deep Learning or Artificial Neural Networks; Unsupervised or Self-Supervised NLP; Reinforcement Learning

Only somewhat entwined fields

- NLP / Computational Linguistics
- AI / Neural Networks



Dædalus 151(2): 127–138 (Spring 2022)
<https://www.amacad.org/publication/human-language-understanding-reasoning>

1. Early Explorations

1940–1969

Machine Translation: The origin of NLP/Computational Linguistics

“Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’ ”

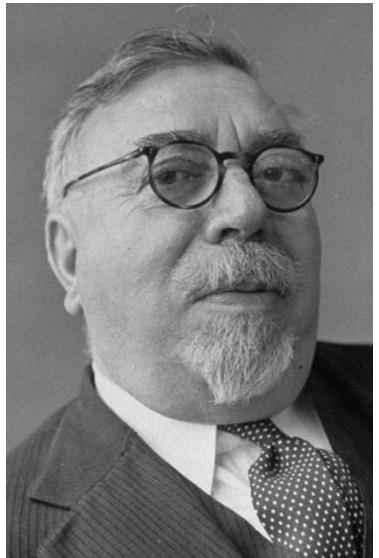
– Warren Weaver (1955:18, quoting a letter he wrote in 1947)





“When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’” – Warren Weaver, March 1947

Weaver was a mathematician & engineer known for his work as a science funder at the Rockefeller Foundation and OSR&D (US Govt WWII science funder) and for coauthoring an approachable Info Theory intro with Shannon



“... as to the problem of mechanical translation, I frankly am afraid that the [semantic] boundaries of words in different languages are too vague ... to make any quasi-mechanical translation scheme very hopeful.”
– Norbert Wiener, April 1947

Wiener: MIT originator of **cybernetics**, which sought to tie together communication, control, and feedback in living things and computers.

"All the News
That's Fit to Print"

The New York Times.

LATE CITY EDITION

Fair and cold today. Increasing cloudiness and cold tomorrow.
Temperature Range Today—Max., 35; Min., 22
Temperatures Yesterday—Max., 36; Min., 26
Full U. S. Weather Bureau Report, Page 41

VOL. CIII., No. 35,048.

Entered as Second-Class Matter,
Post Office, New York, N. Y.

Copyright, 1954, by The New York Times Company.

NEW YORK, FRIDAY, JANUARY 8, 1954.

Times Square, New York 36, N. Y.
Telephone Lackawanna 4-1000

FIVE CENTS

PIER STRIKE CURB SOUGHT IN COURTS AS TIE-UPS SPREAD

Walkout on Another Brooklyn
Dock Stirs Fear of 'Flash'
Stoppage Throughout Port

VOTE DISPUTE CONTINUES

New Questions Delay Sending
Poll Report to Capital—City
Tightens Its Precautions

By A. H. RASKIN

Shipping employers sought court
help yesterday to halt the spread
of Brooklyn pier strikes before
they exploded into a port-wide
shutdown.

The injunction moves came
after a second dock had been
closed by a work stoppage. The
walkout by seventy-five members
of the old International Long-
shoremen's Association stopped
loading of the British freighter
City of Barcelona at Pier 2, Bush
Terminal.

It occurred on the sixth day of
a strike by 150 American Fed-
eration of Labor dock workers,
which has prevented removal of
cargo from two Isbrandtsen Line
freighters at Pier 29, Brooklyn.
Although the walkouts had no
apparent connection with one
another, shipping executives ex-
pressed fear that they might pro-
vide the spark for a general
tie-up of the strife-ridden harbor.

City Widens Precautions

Mayor Wagner took new steps
to perfect the city's preparations
for possible violence in the event
of a waterfront strike on a broad

White Miners Veto Rhodesia Color Bar

By ALBION ROSS
Special to THE NEW YORK TIMES

JOHANNESBURG, South
Africa, Jan. 7—White miners at
the American-controlled Roan
Antelope Mine in Northern
Rhodesia have shaken the min-
ing and industrial world of
southern Africa by voting for
removal of the color bar.

The mine's white workers,
representing roughly one-fourth
of the white persons in the
Northern Rhodesian copper indus-
try, voted to strike out the
first clause of their agreement
with the copper producers. The
clause limits employment in the
higher paid and more
skilled jobs to white men.

In the new goldfields of
Ondendaalrus, Orange Free
State, Gert Lombaard, speak-
ing for the white South African
Gold Miners Union, now under
Nationalist leadership, called
on white miners to resist this

Continued on Page 5, Column 5

MRS. LUCE TO RUSH TO ITALY IN CRISIS

Ambassador Cuts Visit to U.S.
by Week to Report Events
—Vatican Backs Pella

Special to THE NEW YORK TIMES

WASHINGTON, Jan. 7—Mrs.
Clare Boothe Luce, United States
Ambassador to Rome, will return
to her post ahead of schedule be-
cause of the Italian Government
crisis.

Mrs. Luce, who returned to
Washington shortly before Jan. 1,
plans to leave New York for

MAJOR ASKS STUDY OF STATE TAX SHIFT TO AID CITY INCOME

Urge Dewey to Join Move for
'Proper' Sharing of Levies
as Way to End Problems

Text of address by the Mayor
is printed on Page 14.

By PAUL CROWELL
Mayor Wagner offered last
night to join with Governor
Dewey in the appointment of a
commission to consider the ad-
visability of giving the city a
'proper share' of the unlimited
taxing powers now exercised by
the state.

He said the purpose of such a
commission would be to "study
all state aid with a view to abol-
ishing it and returning to us the
taxing powers from which the
money is now obtained."

The Mayor's offer was made
in an address on city finances at
the annual public-service-awards
dinner of the Citizens Budget
Commission at the Plaza Hotel.
He declared that his proposal
looked to a long-range solution
of the city-state fiscal problem
and could not, in all probability,
bear any fruit in time to help the
city prepare its 1954-55 budget.

"I said almost a year ago," the
Mayor declared, "that I was tired
of the city going to Albany for
a handout. I said that I agreed
with the Governor—and this was
in the winter of 1952-53—that
the state should get out of the
collection agency business. I
added that as a substitute for it
the state should return to the lo-
calities their proper share of the
taxing powers themselves; that
they should give us a decent share

of the Smith Act ban on conspiracy to
advocate overthrow of the Gov-
ernment cannot vote, hold public
office, or be elected to Federal or

PRESIDENT WOULD CUT SPENDING, KEEP SECURITY AND PROSPERITY, END CITIZENSHIP OF SUBVERSIVES

'AKIN TO TREASON'

Eisenhower Maps Plan
to Deal With Reds
Guilty of Plots

By W. H. LAWRENCE
Special to THE NEW YORK TIMES
WASHINGTON, Jan. 7—President
Eisenhower proposed today
to strip United States citizenship
from Communists convicted in
the future of conspiring to ad-
vocate the overthrow of Govern-
ment by force and violence.

His advocacy of a modern-day
"men without a country" status
for native-born or naturalized
American conspirators surprised
and delighted Congress.

The Communist conspiracy, the
President said, is "akin to treason"
and should be dealt with as
such. Any person convicted of
conspiring armed overthrow of
Government, he proposed, should
be "treated as having, by such
act, renounced his allegiance to
the United States and forfeited
his United States citizenship."

This goes well beyond the pres-
ent deprivation of civil rights
suffered by those convicted of a
felony. Under Federal law, those
convicted of violating the Smith
Act ban on conspiracy to ad-
vocate overthrow of the Gov-
ernment cannot vote, hold public
office, or be elected to Federal or



Associated Press Wirephoto

THE PRESIDENT REPORTS TO CONGRESS: President Eisenhower as he delivered his annual State of the Union message yesterday at a joint session of Congress. Behind him on the dais are Vice President Richard M. Nixon, left, and Speaker Joseph W. Martin Jr.

MESSAGE ON UNION

Eisenhower Proposes
Amendment to Give
18-Year-Olds Vote

Text of State of Union Message
is printed on Page 10.

By JAMES RESTON
Special to THE NEW YORK TIMES
WASHINGTON, Jan. 7—Presi-
dent Eisenhower, in his second
State of the Union Message, to-
day asked the almost evenly di-
vided Eighty-third Congress to
moderate but sustain the foreign
and welfare policies of the New
Deal era.

He proposed to cut the military
budget, to cut taxes in relation
to Government expenditures, and
to reduce the guarantees to farm-
ers very slowly. The cuts, how-
ever, were designed to reconcile
security and solvency without
jeopardizing the collective secu-
rity or prosperity policies of the
past.

The message contained only
two surprises.

One was an appeal for a law
that would take away the citizen-
ship of any American hereafter
convicted in the courts of "con-
spiring to advocate the over-
throw of this Government by
force or violence."

Russian Is Turned Into English By a Fast Electronic Translator

By ROBERT K. PLUMB

A public demonstration of what is believed to be the first successful use of a machine to translate meaningful texts from one language to another took place here yesterday afternoon.

This may be the culmination of centuries of search by scholars for "a mechanical translator." So far the system has a vocabulary of only 250 words. But there are no foreseeable limits to the number of words that the device can store or the number of languages it can be directed to translate.

Scholars and scientists who worked on it believe that within a few years the system may greatly increase communication, particularly in technical subjects, by making translation quick, accurate and easy.

The demonstration was at the headquarters of the International Business Machines Corporation, 590 Madison Avenue. It is the result of cooperative research by

scientists of the corporation and scholars of the Georgetown University Institute of Languages and Linguistics in Washington.

The "mechanical" part of the translation system, which is mostly electronic, is a standard commercial model of the largest International Business Machines "stock" computer. This device, called the IBM Type 701 Electronic Data Processing Machine, was put on the market last April. Since then twelve of the machines have been sold to commercial, military and university computation laboratories.

The "literary" part of the system is a mechanical model of language devised at Georgetown by Prof. Leon Dostert and Dr. Paul Garvin. The corporation's share in the project was conducted by Dr. Cuthbert C. Hurd, director of its Division of Applied Science.

In the demonstration, a girl

Continued on Page 5, Column 2

Russian Is Turned Into English By a Fast Electronic Translator

By ROBERT K. PLUMB

A public demonstration of what is believed to be the first successful use of a machine to translate meaningful texts from one language to another took place here yesterday afternoon.

This may be the culmination of centuries of search by scholars for "a mechanical translator." So far the system has a vocabulary of only 250 words. But there are no foreseeable limits to the number of words that the device can store or the number of languages it can be directed to translate.

Scholars and scientists who worked on it believe that within a few years the system may greatly increase communication, particularly in technical subjects, by making translation quick, accurate and easy.

The demonstration was at the headquarters of the International Business Machines Corporation, 590 Madison Avenue. It is the result of cooperative research by

scientists of the corporation and scholars of the Georgetown University Institute of Languages and Linguistics in Washington.

The "mechanical" part of the translation system, which is mostly electronic, is a standard commercial model of the largest International Business Machines "stock" computer. This device, called the IBM Type 701 Electronic Data Processing Machine, was put on the market last April. Since then twelve of the machines have been sold to commercial, military and university computation laboratories.

The "literary" part of the system is a mechanical model of language devised at Georgetown by Prof. Leon Dostert and Dr. Paul Garvin. The corporation's share in the project was conducted by Dr. Cuthbert C. Hurd, director of its Division of Applied Science.

In the demonstration, a girl

Continued on Page 5, Column 2

LANGUAGE DEVICE TRANSLATES FAST

Continued From Page 1

operator typed out on a keyboard the following Russian text in English characters:

"Mi pyeryedayem mislyi pos-
ryedstvom ryechy."

The machine printed a translation almost simultaneously:

"We transmit thoughts by
means of speech."

The operator did not know Russian. Again she typed out the meaningless (to her) Russian words:

"Velyichyna uglj opryedyely-
ayetsya otnosheniyem diyini-
dugi k radiyusu."

And the machine translated it to:

"Magnitude of angle is deter-
mined by the relation of length
of arc to radius."

Guided by Language Code

Several short messages, within the 250-word range of the device, were tried. Included were brief statements in Russian about politics, law, mathematics, chemistry, metallurgy, communications and military affairs. The sentences were turned into good English without human intervention.

The heart of the system is the mechanical model of language devised at Georgetown. There the scholars first assembled 250-word vocabulary in Russian covering the seven broad fields. Then

they determined the rules of syntax required for a meaningful statement and reduced them to six instructions for the data-processing calculator.

These instructions are introduced into the calculator's short-term electrostatic "memory" with punch cards. The cards tell the machine how to cope with syntax.

In translating, for instance, a word "A" which precedes a word "B" in Russian, may be reversed in some cases in English. Each of the 250 words is coded for this inversion. Sometimes words must be inserted in the English text, sometimes they must be omitted, following code instructions.

When there are several possible English meanings for a Russian word, the instructions tell the machine to pick out the meaning that best fits the context.

Foreign words are typed on a keyboard that punches out I.B.M. cards. These are fed into the calculator, where they encounter the vocabulary, also punched on cards. On a standard printer meaningful English texts emerge.

According to Dr. Hurd, the calculator is a general-purpose data processing machine not designed specifically for translating. Nev-

ertheless, it has a memory capable of storing roughly 1,000,000 five-letter words. There are 600,000 entries in the latest Webster's unabridged New International dictionary.

Dr. Hurd said that the corporation would now design a machine particularly fit for translating rather than for general computing utility. Such a device should be ready within three to five years, when the Georgetown scholars believe they can complete the "literary" end of the system.

Dr. Dostert and Dr. Garvin said they chose Russian for their first experiments because it was a difficult language and a system that could translate it could handle anything.

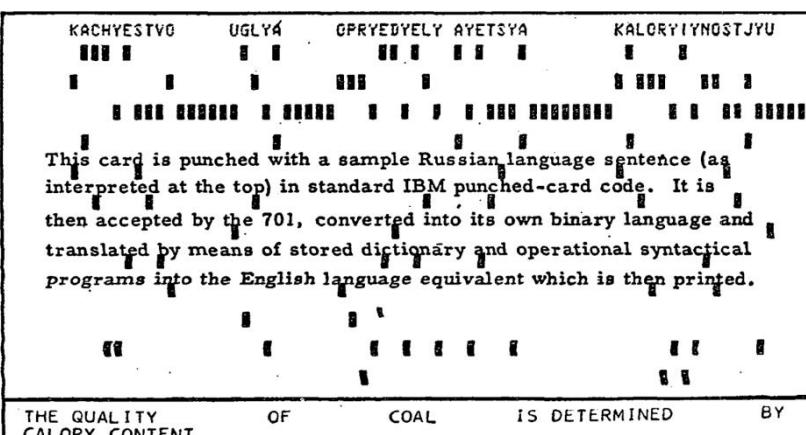
The machine will not accept incoherent statements, Dr. Dostert said. If they are introduced for "translation" the machine balks and rings a bell. And it will ring the bell when it encounters a misprint. It now prints eighty letters in two seconds.

As soon as cards for Russian are completed, sets will be made for German and French. Then other Slavic, Germanic and Romance languages can be set up at will.

Calculator Takes on a New Job: Language Translation



The New York Times
An electronic calculator produced by the International Business Machines Corporation was demonstrated yesterday in a new role: translating Russian phrases into English. The device has a "vocabulary" of 250 Russian words and can be adapted to other languages. Miss Marilyn Polle "types" Russian phrases on I. B. M. punch cards that are fed into the machine.



Above, specimen punch card and below a strip with translation, typed almost simultaneously

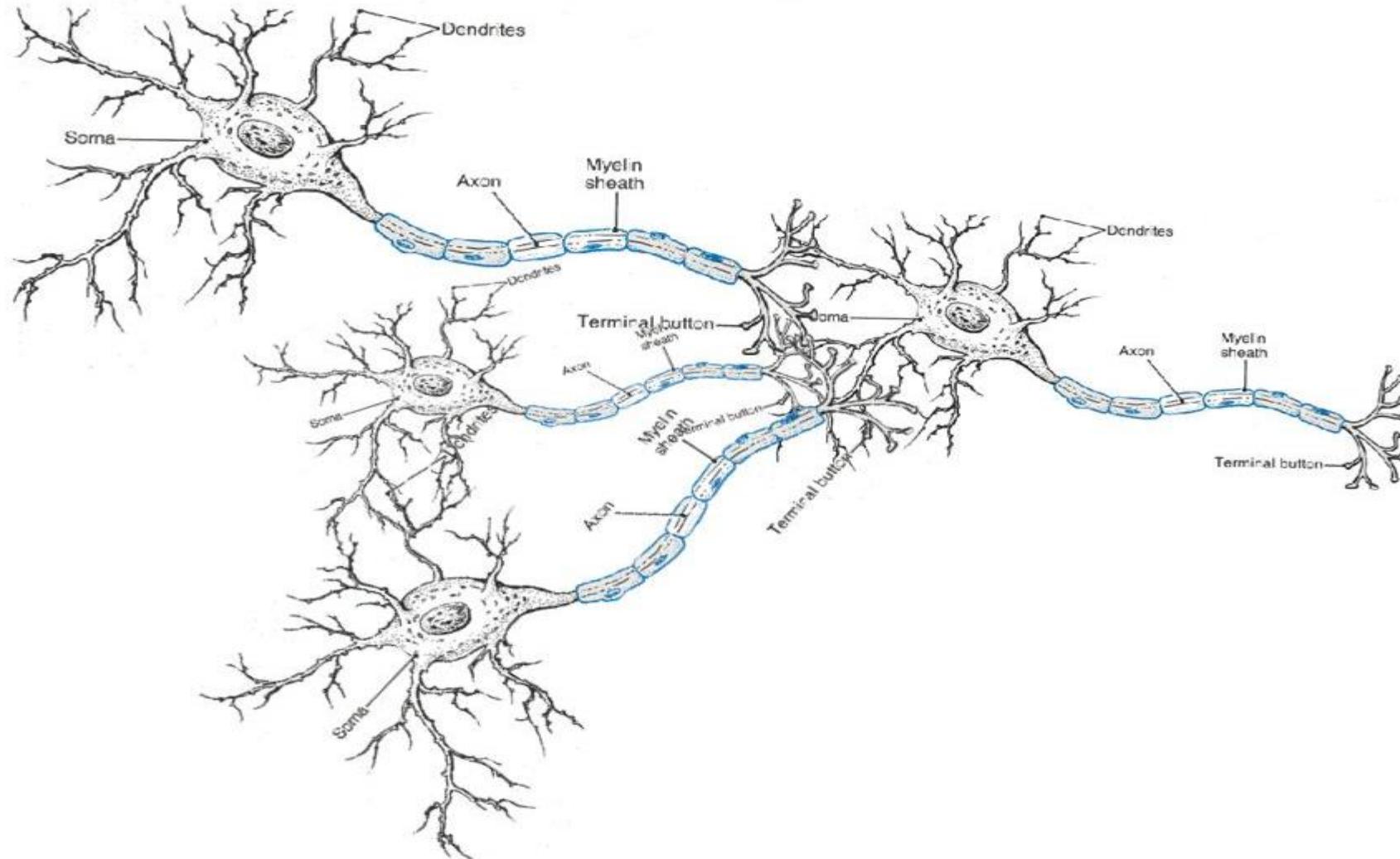
The early history of NLP: MT in the 1950s



AI Begins: The Dartmouth Summer Research Project 1956



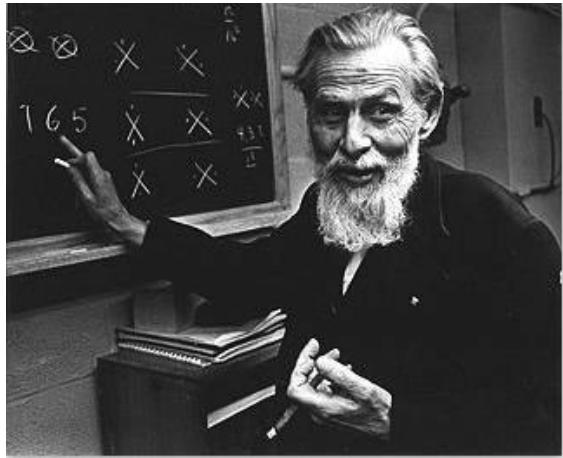
Neural Networks: The inspiration from knowledge of neurons



Warren S. McCulloch

and

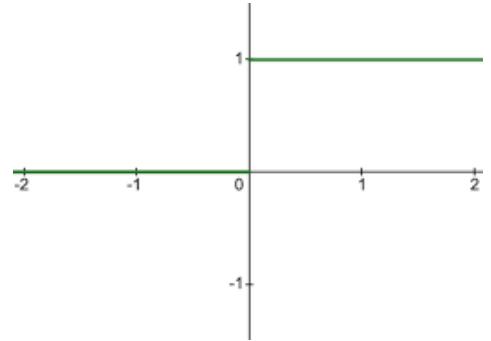
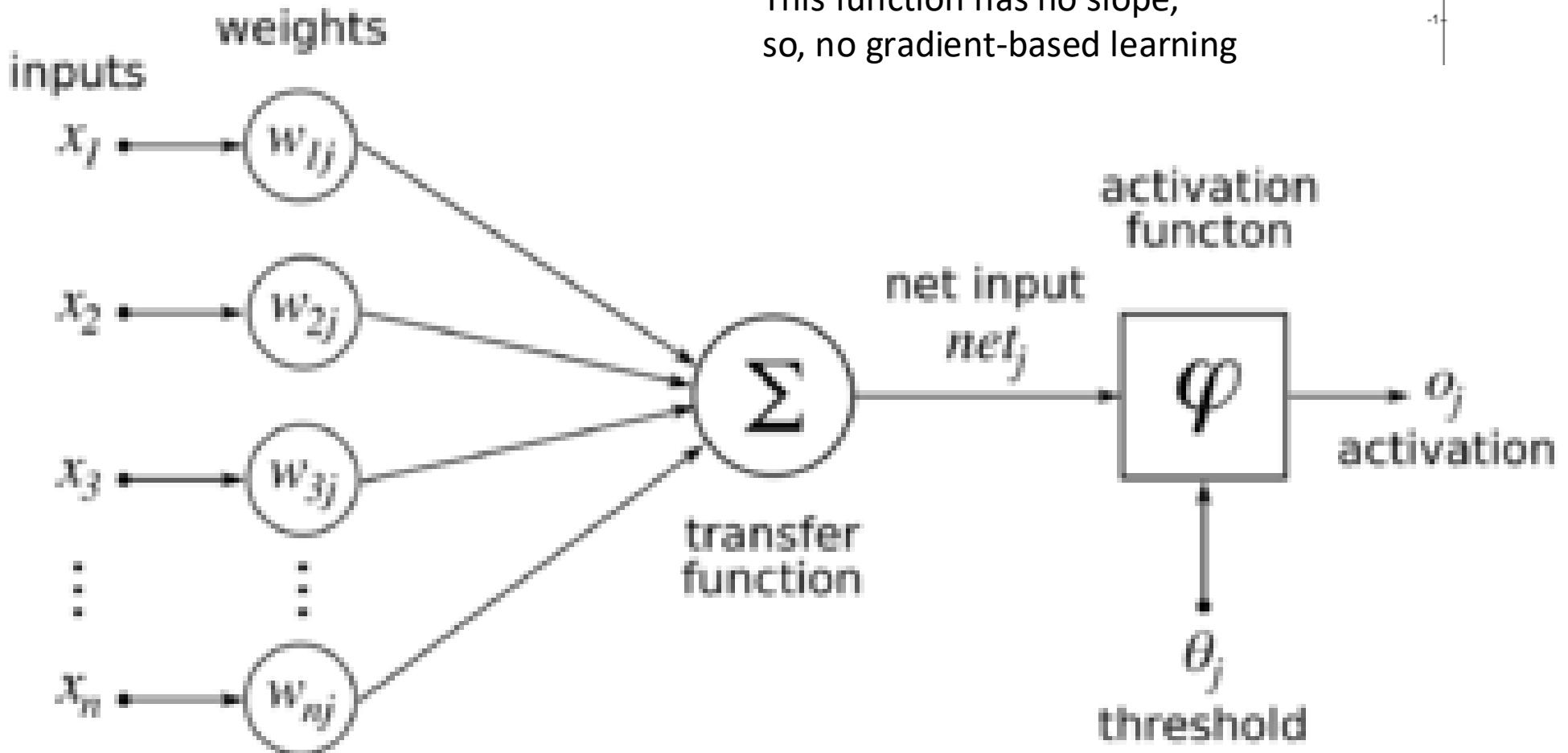
Walter Pitts



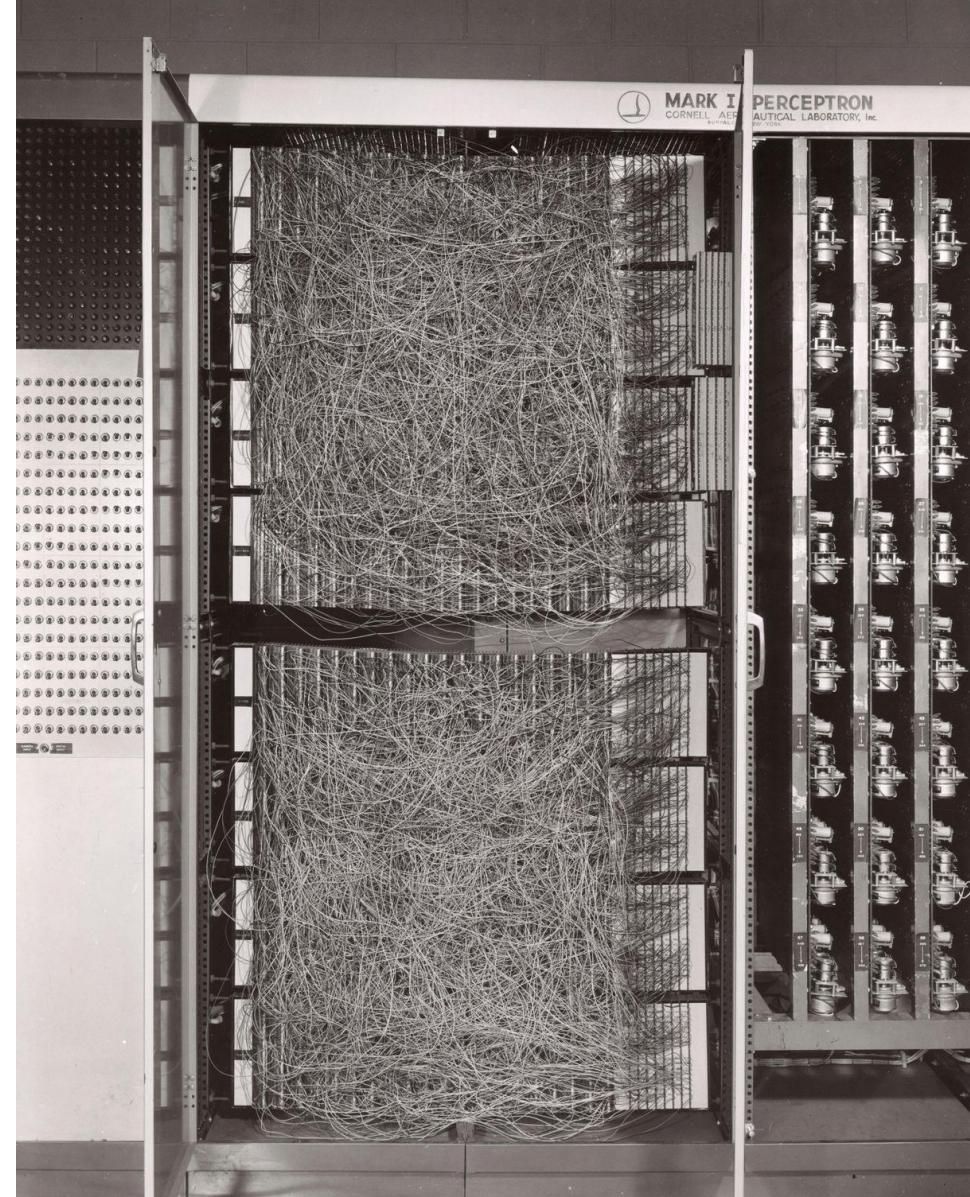
- 1898–1969. Neurophysiologist (MD), psychiatrist, and cyberneticist, moved to MIT in 1951 but Wiener split with him and Pitts in 1952.
- 1923–1969. Self-taught logician* and computational neuroscientist, lived homeless at Chicago and then MIT; died of alcoholism.

In 1943, published “A Logical Calculus of Ideas Immanent in Nervous Activity”, the first mathematical model of a neural network, showing it could implement a Turing Machine

McCulloch-Pitts Neuron



Frank Rosenblatt: The (Mark I) Perceptron



<https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon>

Early AI Hype! New York Times July 8, 1958



NEW NAVY DEVICE LEARNS BY DOING Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

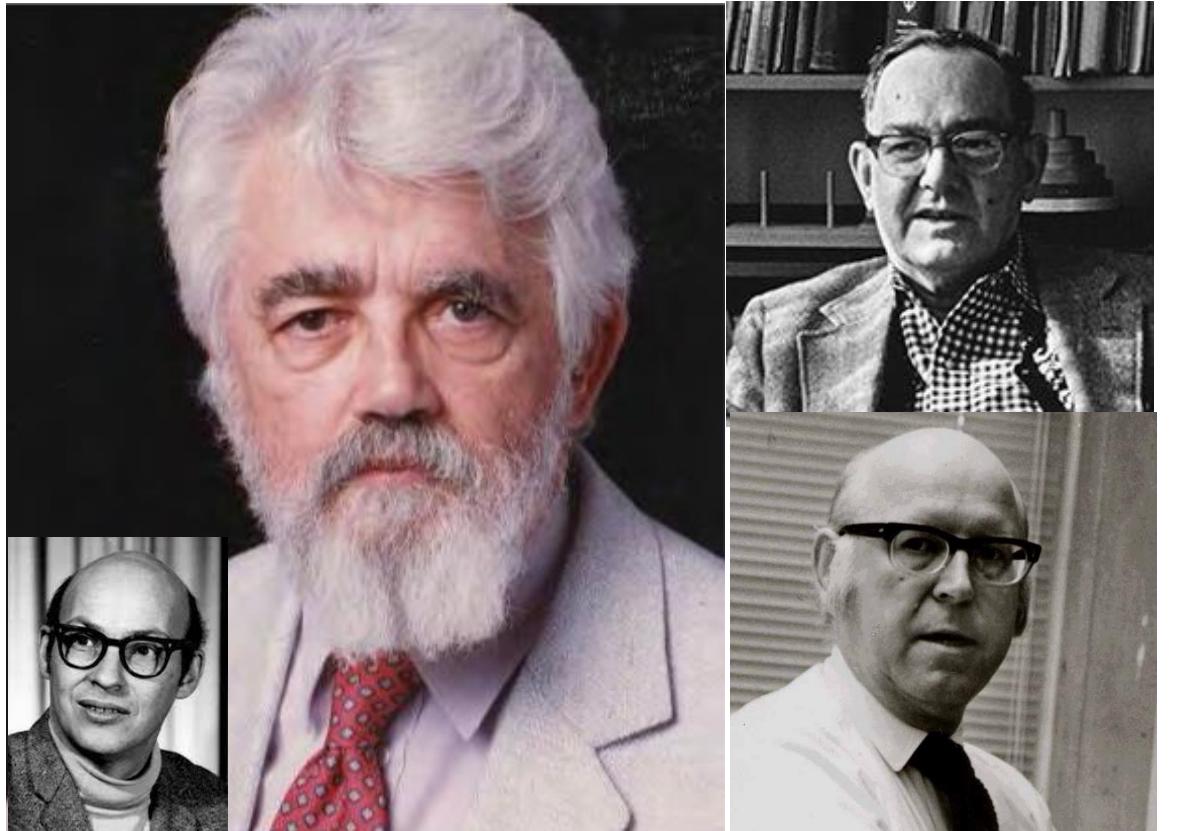
The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

Two visions of artificial intelligence in the 1950s

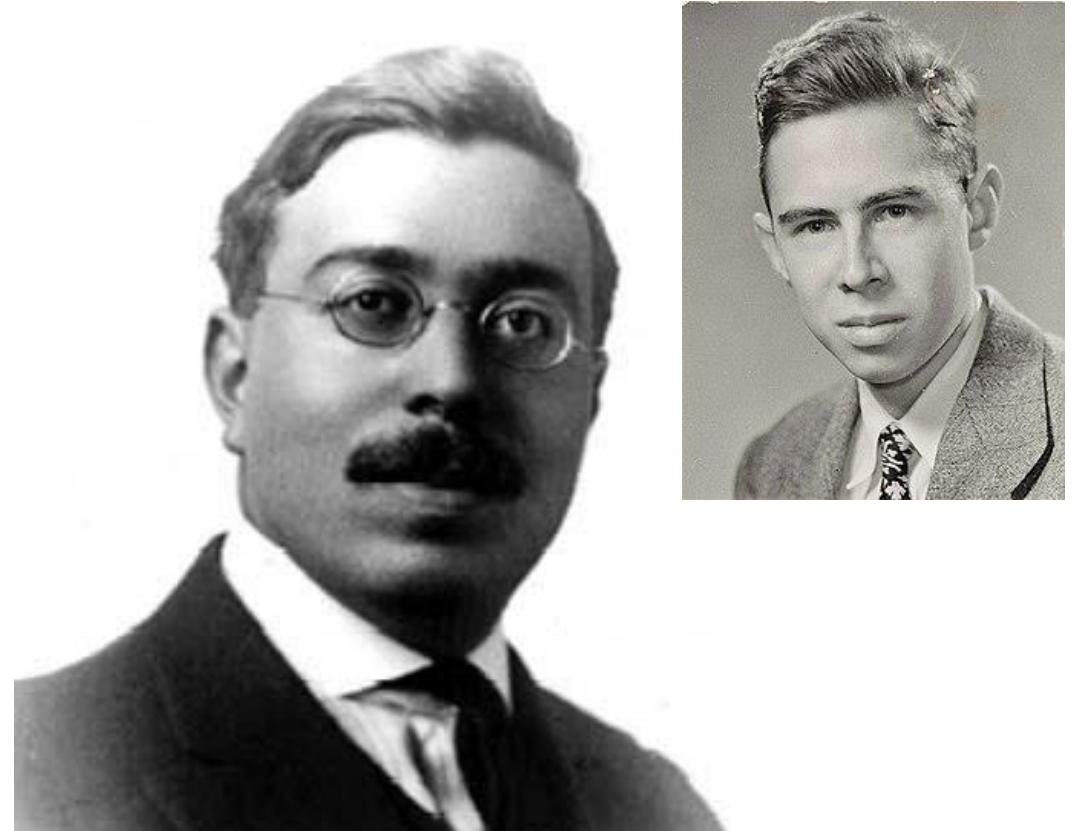
(Symbolic) Artificial Intelligence

Minsky, McCarthy, Simon, Newell



Cybernetics

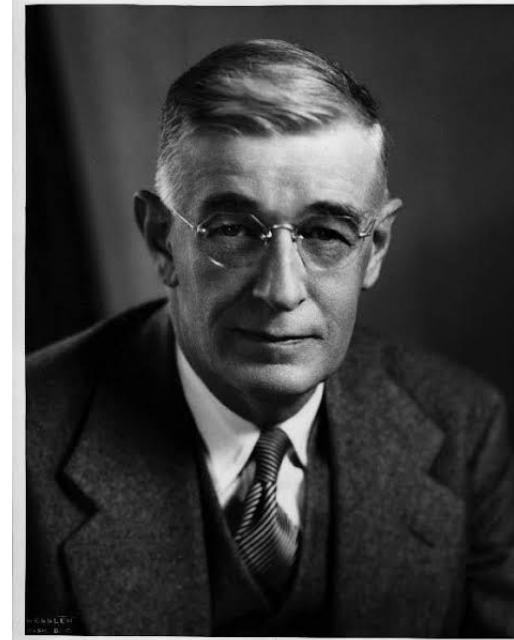
Wiener, Rosenblatt



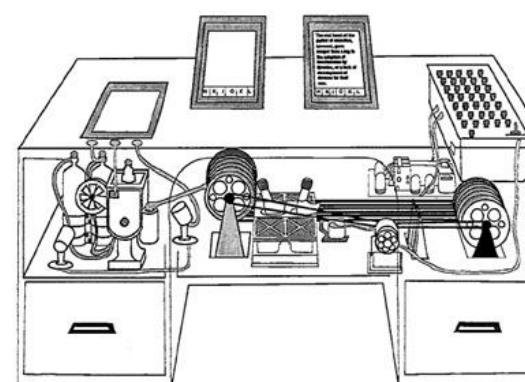
Information Retrieval: Vannevar Bush

Bush (1945): As We May Think

“Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, “memex” will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.



Bush went from being Dean of Engineering at MIT and president of the Carnegie Institution of Washington to chairman of the National Defense Research Committee in WWII (which had huge funding, directing all wartime science, including the Manhattan Project)



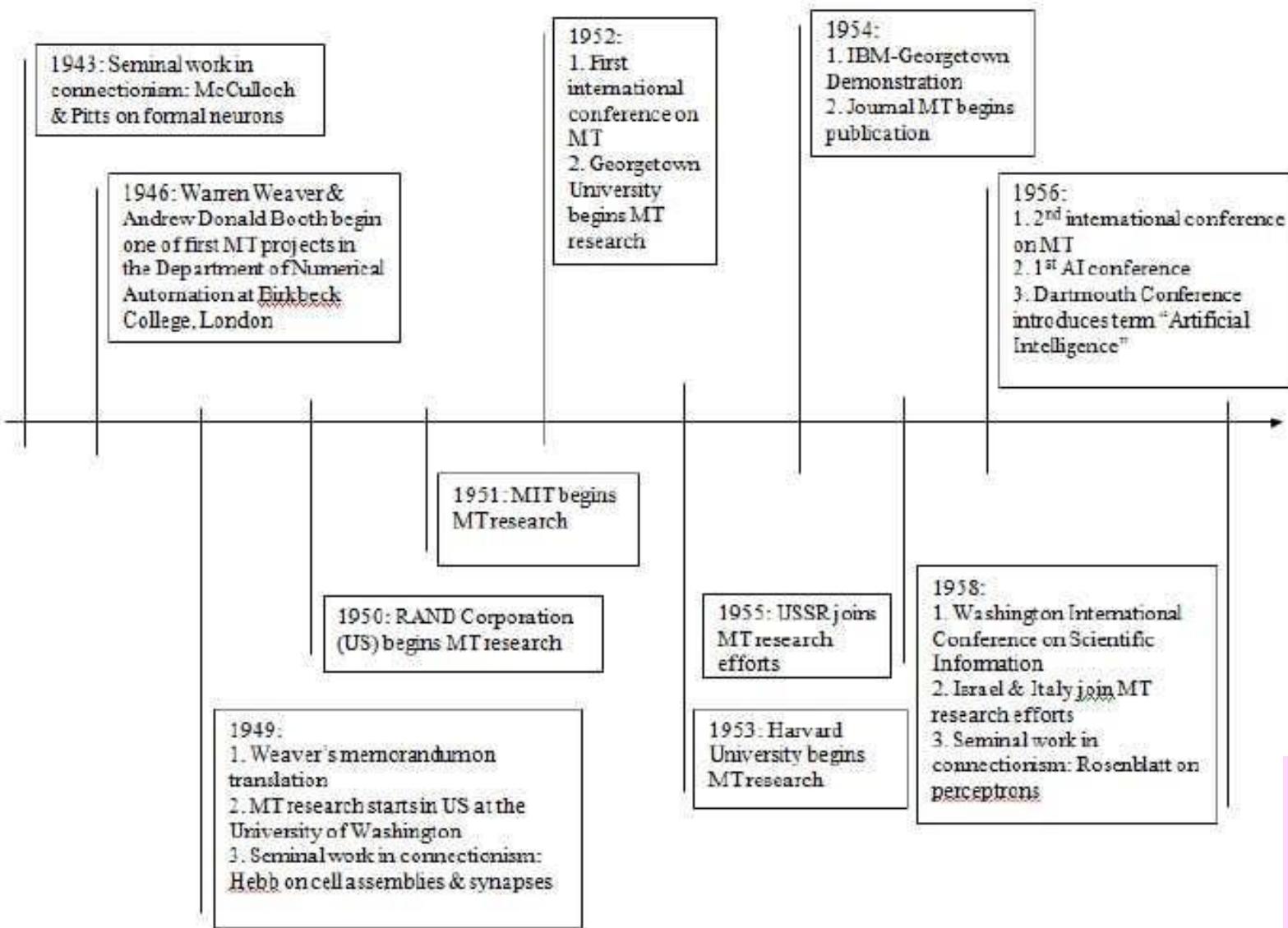
<https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>

Cyril Cleverdon introduces benchmarking

- Cranfield tests (1957–1967)
- Defined the idea of **benchmark tests** for language, with a document collection, queries, and correct answers.
- He had exhaustive answers on document relevance over a small corpus!



The 1950s origins of NLP/Computational Linguistics



- Foundational work on automata, formal languages, probabilistic modeling, and information theory
- First speech systems (Davis et al.)
- MT heavily funded by military — huge overconfidence
- But using machines dumber than a pocket calculator
- Little understanding of syntax, semantics, pragmatics

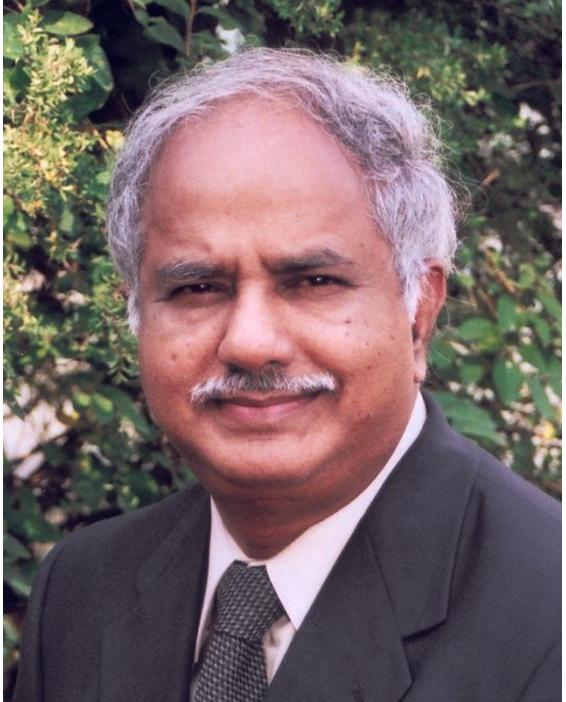
I grabbed these timelines from Ruth Camburn's "A Short History of Computational Linguistics". She was a CSU Fresno Linguistics grad student around 2013.

1960s: David Glenn Hays: ACL founder and dependency parsing



Raj Reddy

(1937–, Stanford PhD student 1963–1966; faculty 1966–1969; CMU; Turing award)



(Selected) Students:

James Gosling, Kai-Fu Lee,
Roni Rosenfeld, Alex Waibel

Joyce Friedman (1928–2018; at Stanford 1965–1968)

cs -109



A COMPUTER SYSTEM FOR WRITING AND TESTING
TRANSFORMATIONAL GRAMMARS¹

FINAL REPORT

JOYCE FRIEDMAN
PRINCIPAL INVESTIGATOR

This research was supported in part by the United
States Air Force Electronic Systems Division, under
Contract F196828-C-0035.



C. Ray
Perrault
↓
James
Allen

David Scott
Warren

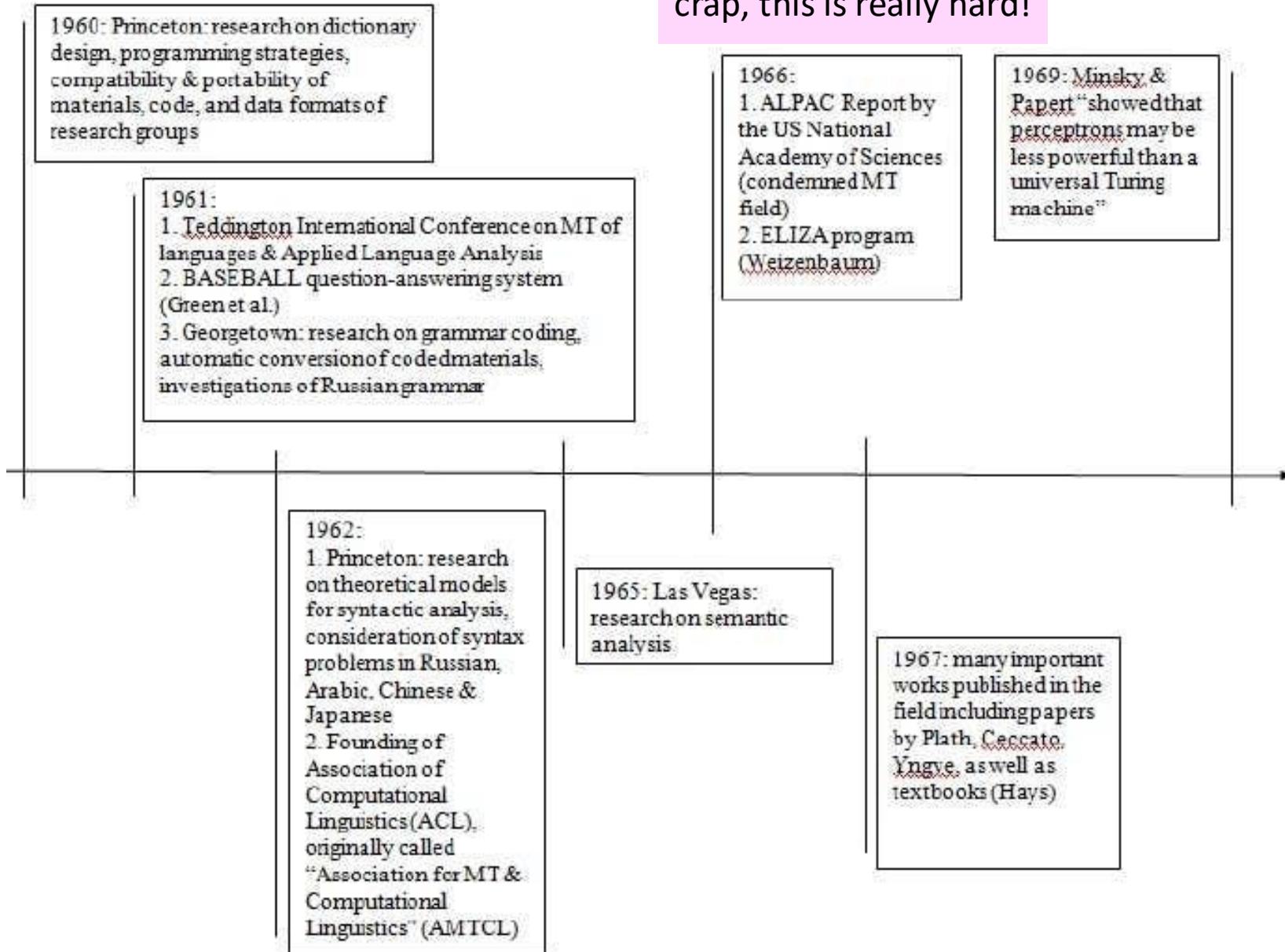


STANFORD UNIVERSITY COMPUTER SCIENCE DEPARTMENT
COMPUTATIONAL LINGUISTICS PROJECT

30 SEPTEMBER 1968



NLP in the 1960s



2. Hand-built demonstration NLP systems, of increasing formalization

1970–1992

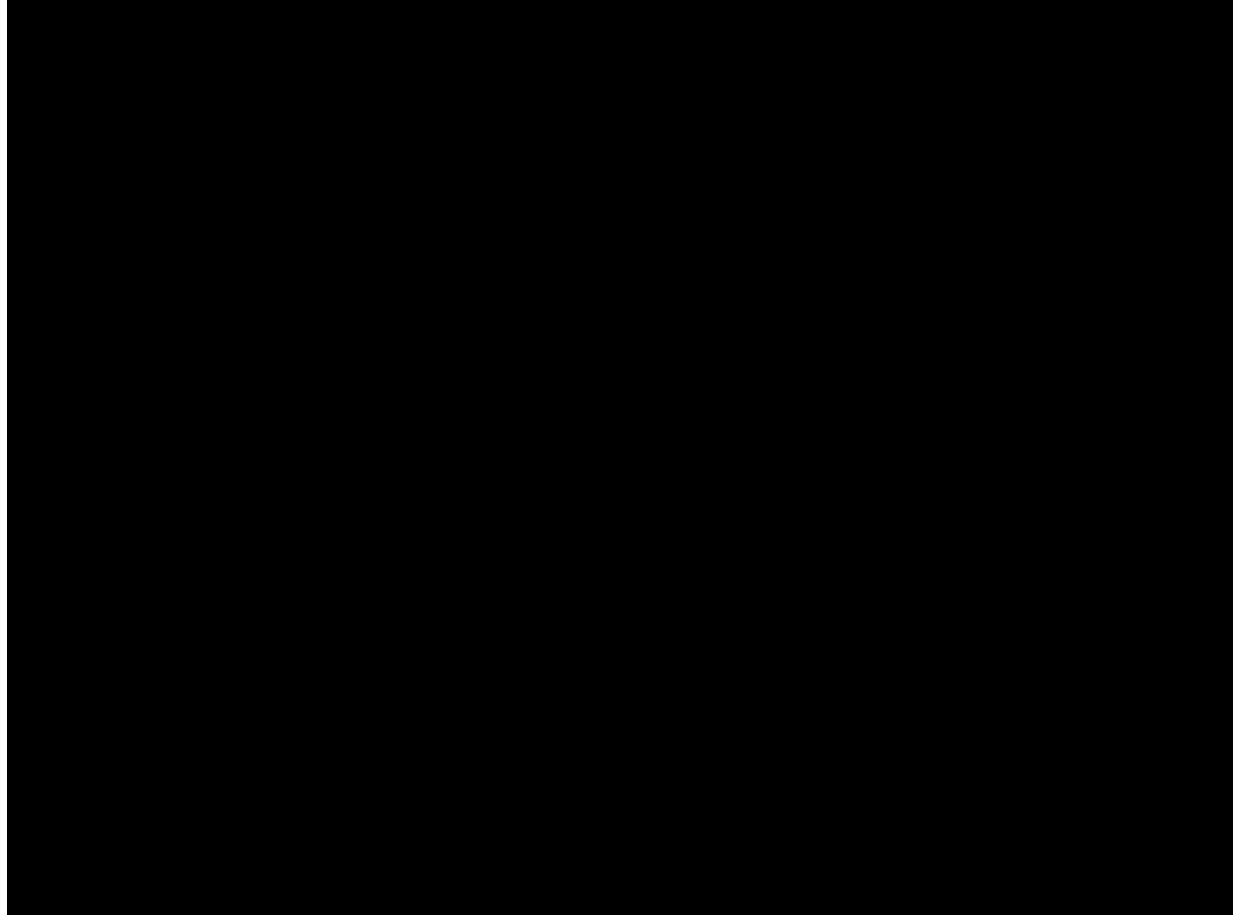
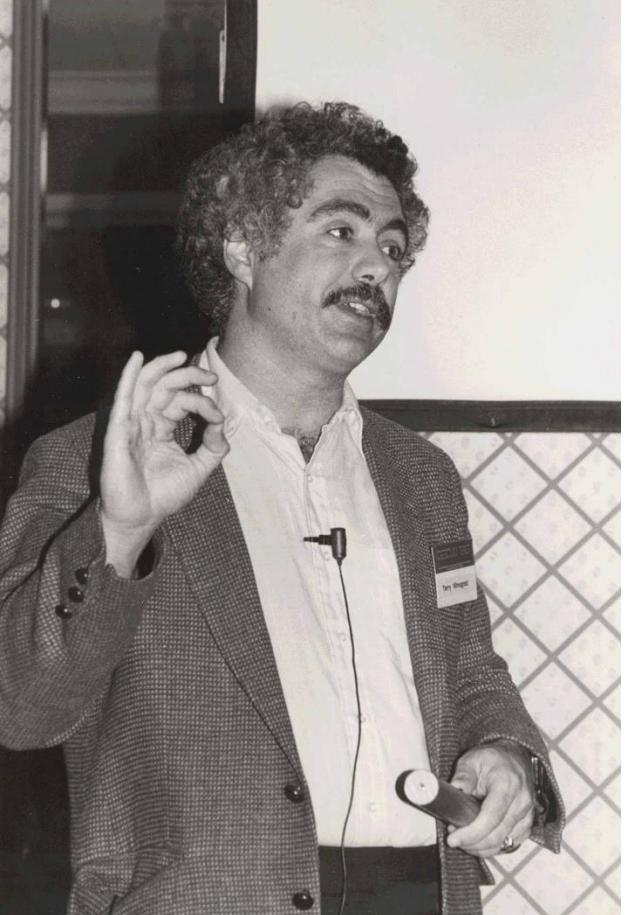
Terry Winograd

(1946–, Stanford faculty 1973–2014; AI/NLP 1973–1984)

Seymour Papert

Terry Winograd

Stuart Shieber



Formal and Unification-based Grammars for NLP: Martin Kay



Fernando Pereira

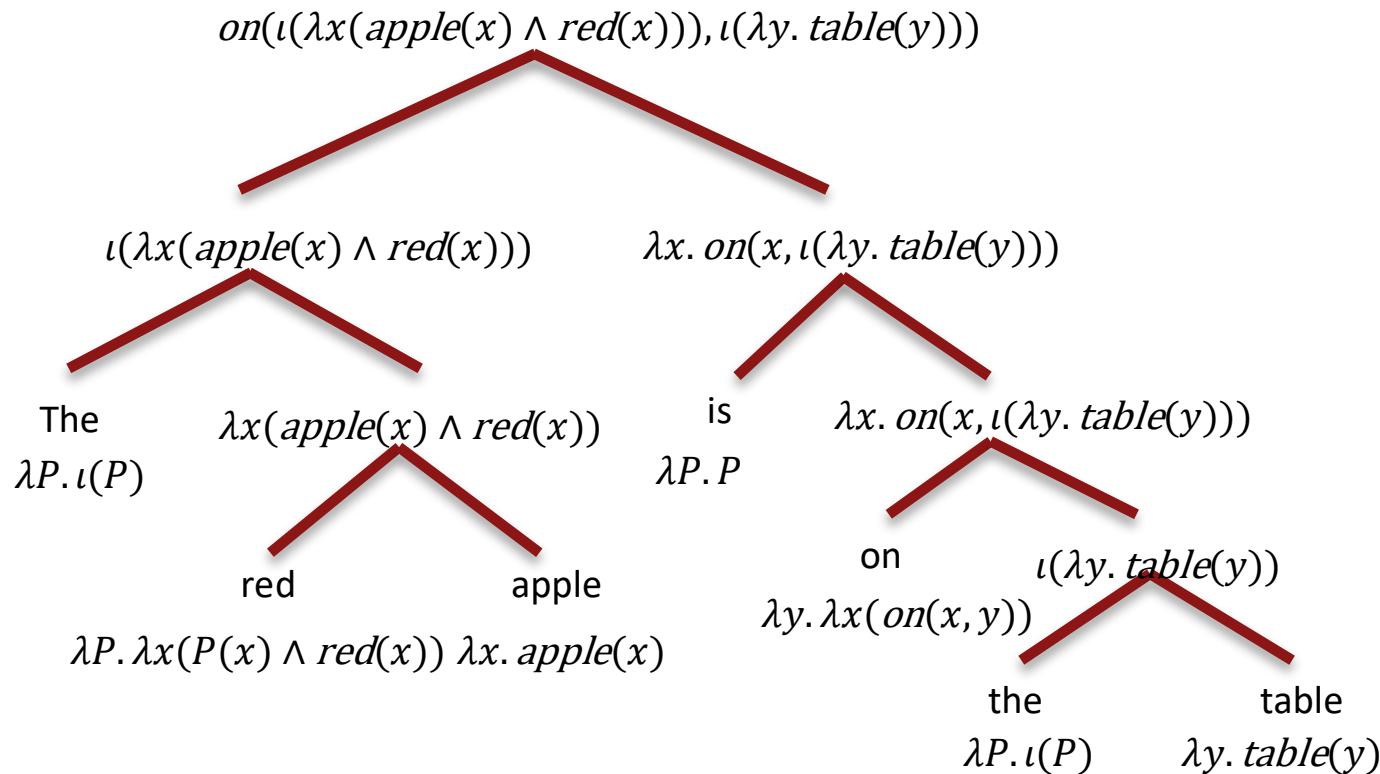


Stuart Shieber



Constructing the meaning of human language: 1967–2017

- Tokenize a sentence into words: **The red apple is on the table**
- Parse it into a tree or graph data structure (using context-free grammars and beyond)
- Construct its meaning by: (i) lexical lookup, then (ii) semantic composition, using a “rule-to-rule” approach working up the tree (e.g., $PP: \alpha(\beta) \rightarrow P: \alpha NP: \beta$)

$$on(\iota(\lambda x. apple(x) \wedge red(x))), \iota(\lambda y. table(y)))$$


NLP and Knowledge Representation: Norvig (1986) Ph.D.

Peter Norvig's thesis: A Unified Theory of Inference for Text Understanding (1986)

A Unified Theory of Inference for Text Understanding

By

Peter Norvig

B.S. (Brown University) 1978

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

in the

GRADUATE DIVISION

OF THE

UNIVERSITY OF CALIFORNIA, BERKELEY

The language analyzed:

In a poor fishing village built on an island not far from the coast of China, a young boy named Chang Lee lived with his widowed mother. Every day, little Chang bravely set off with his net, hoping to catch a few fish from the sea, which they could sell and have a little money to buy bread.

- (a) There is a sea, which surrounds the island, is used by the villagers for fishing, and forms part of the coast of China
- (b) Chang intends to trap fish in his net, which is a fishing net
- (c) The word *which* refers to *the fish*
- (d) The word *they* refers to Chang and his mother

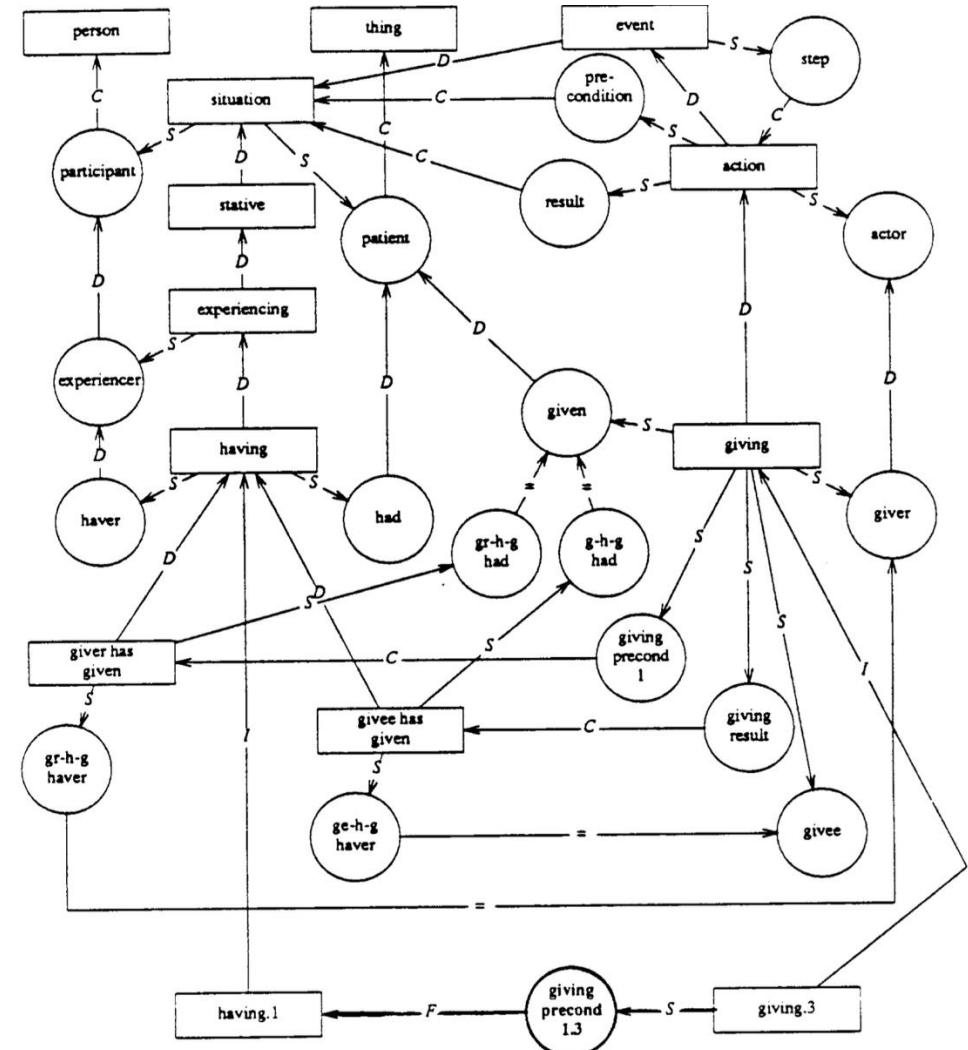


The unified theory of inference

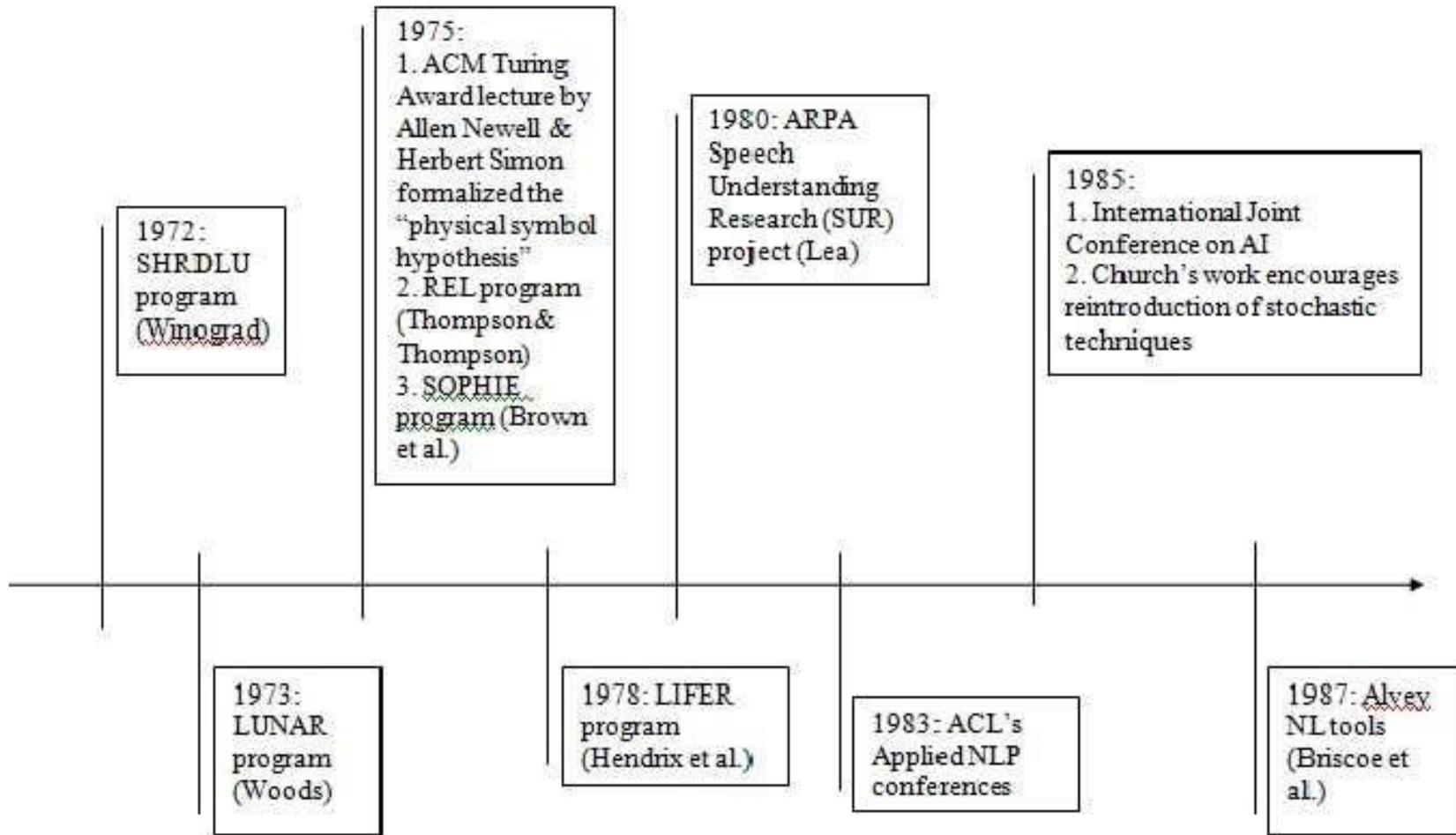
“As we have just seen, a suitable knowledge base is a prerequisite for making proper inferences” (p. 4). It’s built to enable inferences

System had 6 general forms of inference; 2 pairs, so 4 basic types:

1. Elaboration: Filling a slot to connect two entities
 - John got piggybank for REASON have money for REASON buy present
2. Reference Resolution: Hey – it’s coreference!!!
3. View Application: *The Red Sox killed the Yankees*
 - KILLED is not animal; KILLING is viewed as a DEFEAT-CONVINCINGLY
4. Concretization: Infer more specific
 - TRAVELLING in an AUTOMOBILE is an instance of DRIVING



NLP in the 1970s and 1980s

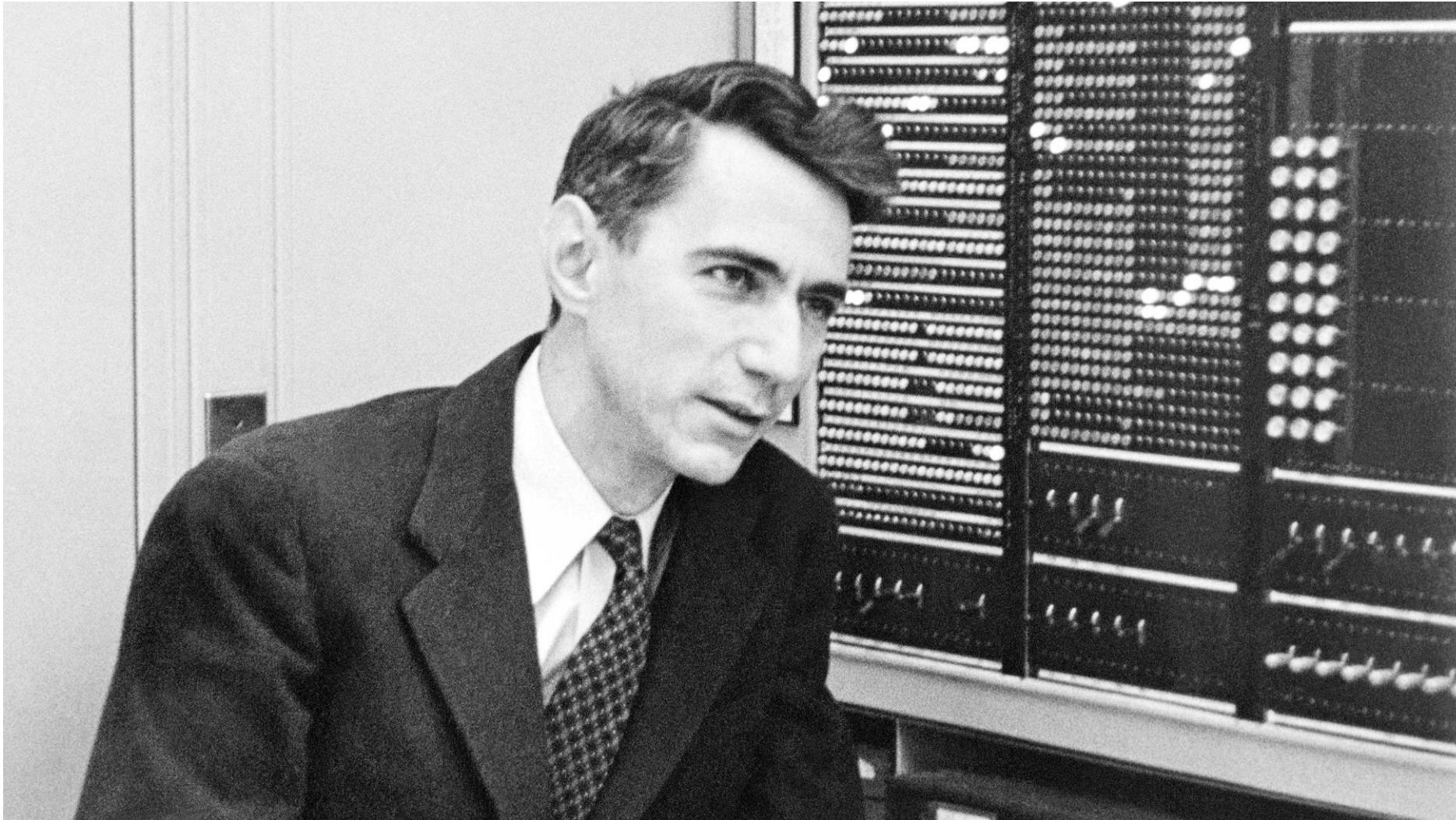


- Foundational work on speech recognition: stochastic modeling, hidden Markov models, the “noisy channel”
- Ideas from this work would later revolutionize NLP!
- Logic programming, rules-driven AI, deterministic algorithms for syntactic parsing (e.g., LFG)
- Increasing interest in natural language understanding: SHRDLU, LUNAR, CHAT-80
- But symbolic AI hit the wall: “AI winter”

3. Statistical or Probabilistic NLP (“StatNLP”) and then more general Supervised ML for NLP

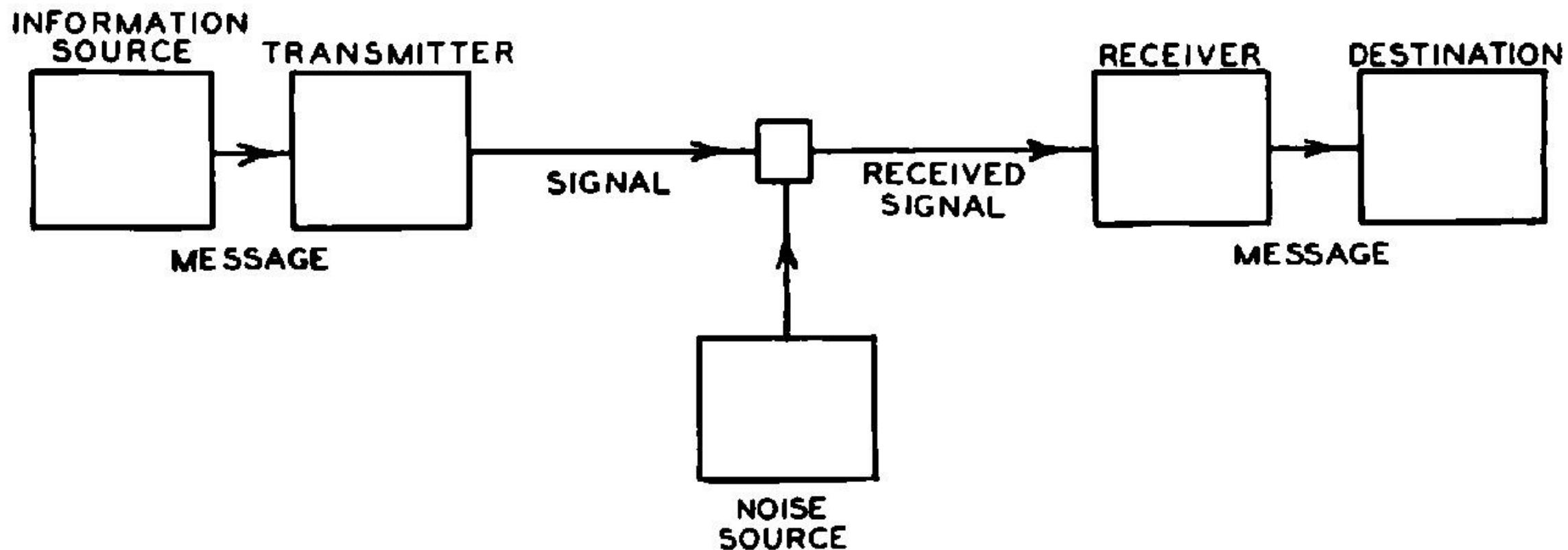
1993–2012

Claude Shannon



A Mathematical Theory of Communication (1948)

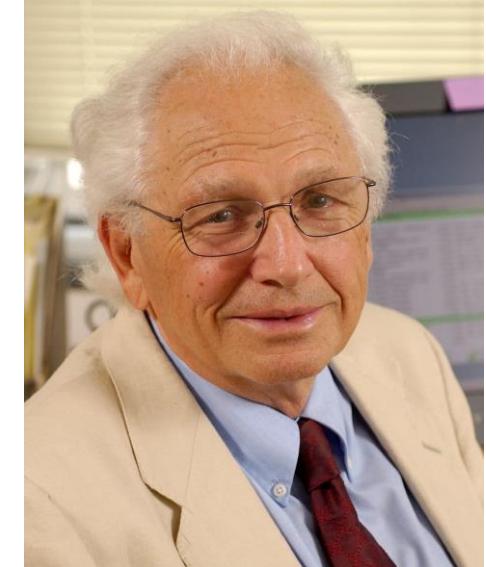
The noisy channel model



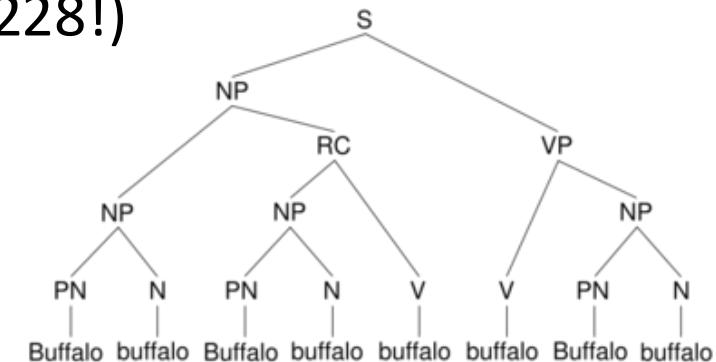
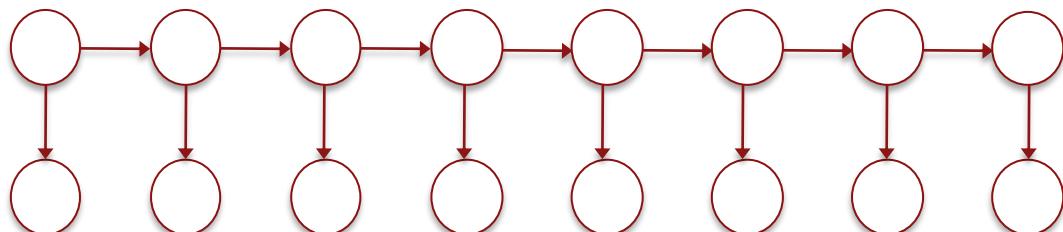
<https://www.quantamagazine.org/how-claude-shannons-information-theory-invented-the-future-20201222/>

The statistical revolution: 1990s

- Influx of new ideas from EE & ASR: probabilistic modeling, corpus statistics, supervised learning, empirical evaluation
- New sources of data: explosion of machine-readable text; human-annotated training data (e.g., the Penn Treebank)
- Annotated data + algorithms + probabilistic prediction
- Lowered expectations: forget full semantic understanding, let's do text cat, part-of-speech tagging, NER, and parsing!
- Tools: Naïve Bayes classifiers, HMMs, PCFGs, CRFs (CS228!)



Fred Jelinek



Probabilistic spelling correction

“She is a stellar and versatile **acress** whose combination of sass and glamour attracts”

Let's estimate probabilities of corrections following one another

$$P(\text{actress} \mid \text{versatile}) = .0019$$

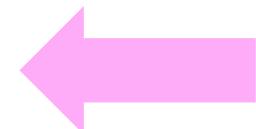
$$P(\text{across} \mid \text{versatile}) = .000092$$

$$P(\text{whose} \mid \text{actress}) = .0043$$

$$P(\text{whose} \mid \text{across}) = .000026$$

$$P(\text{"versatile actress whose"}) = .00019 \times .0043 = 817,000 \times 10^{-10}$$

$$P(\text{"versatile across whose"}) = .000092 \times .000026 = 239 \times 10^{-10}$$



The rise of the machines: 2000s

- Availability of much more powerful machines
- Consolidation of the gains of the statistical revolution
- More sophisticated statistical modeling and machine learning algorithms: MaxEnt, SVMs, Bayes Nets, LDA, etc.
- Big big data: 100x growth of web, massive server farms
- Focus shifting from supervised to *unsupervised* learning
- Revived interest in higher-level semantic applications



4. Deep Learning or Artificial Neural Networks for NLP

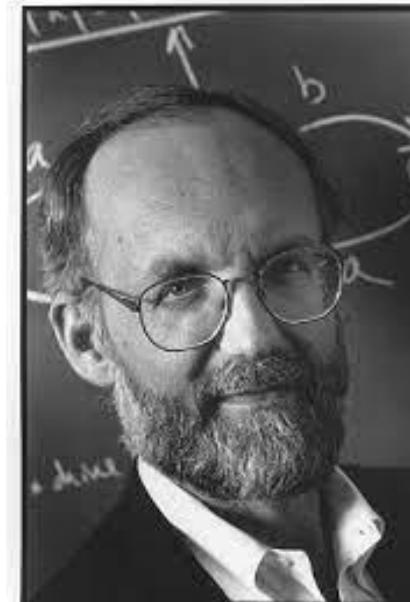
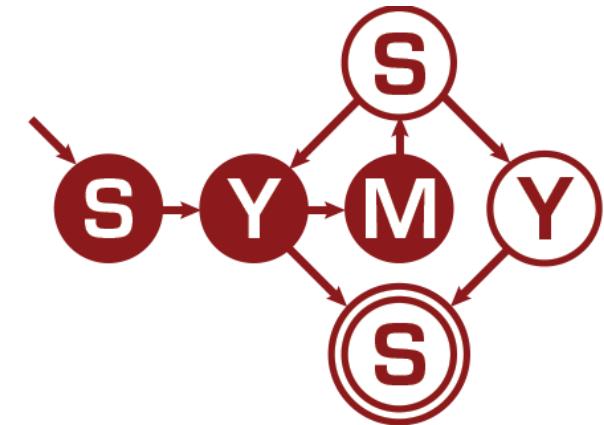
4a. Neural NLP

2013–2021

Symbolic AI vs. “Cybernetics”

Stanford, the home of “Symbolic Systems”

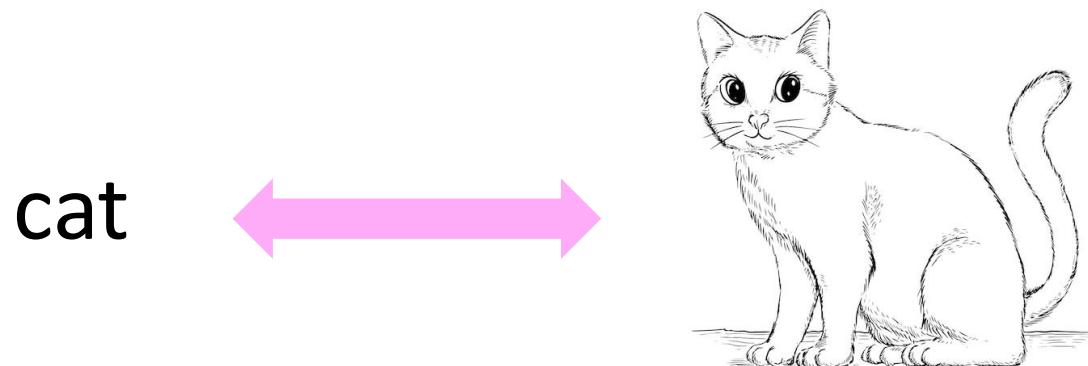
- “Symbolic systems” studies systems of meaningful symbols that represent the world about us—like human languages, logics, and programming languages—and the systems that work with these symbols—like brains, computers, and complex social systems
 - While “cognitive science” focuses on the mind and intelligence as naturally occurring phenomena, Symbolic Systems gives equal focus to human-constructed systems that use symbols to communicate and to represent information



Jon Barwise
(1942–2000)

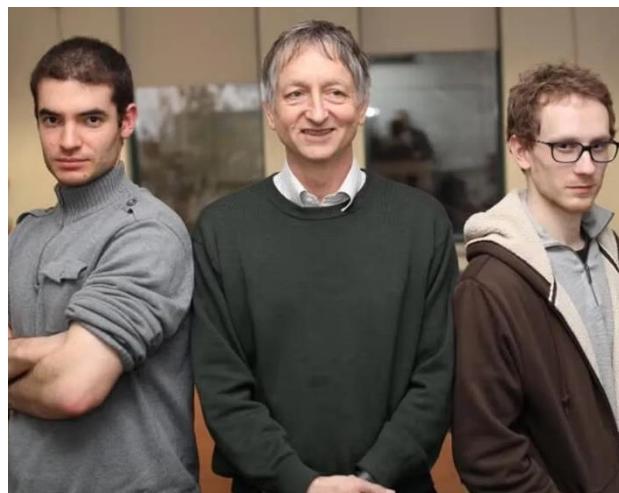
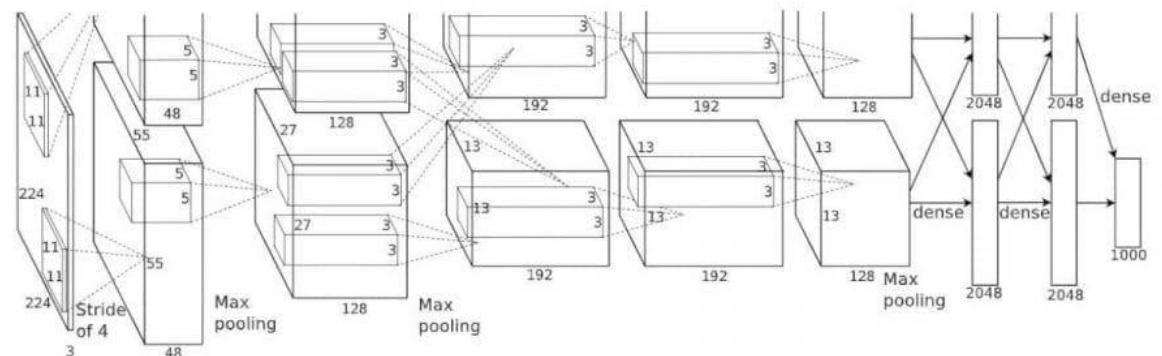
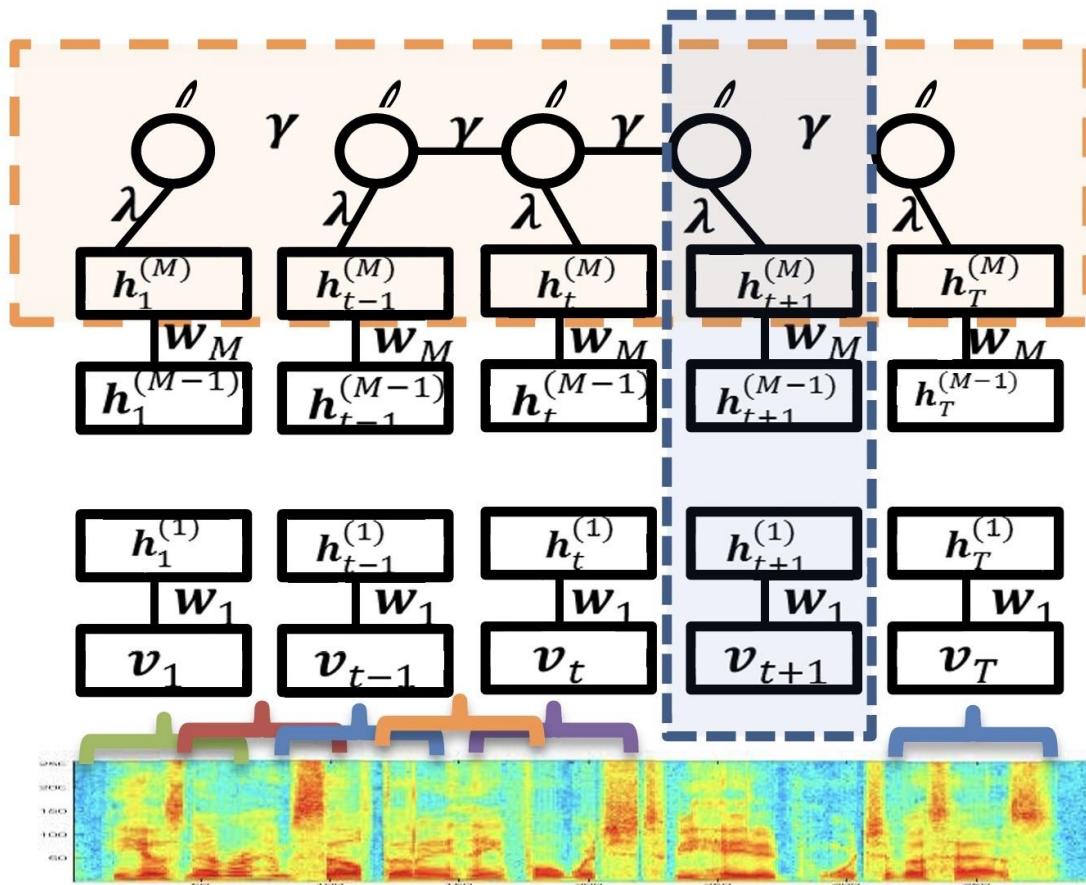
Symbolic systems versus their processors

- Language is the symbolic system par excellence; we should study and make use of its symbolic structure
- This does not show that the main processor of these symbols—the human brain—is implemented as a physical symbol system
- We need not design NLP systems as physical symbol systems
- The brain is more like a neural network model
- Artificial neural network models scale better and can capture the world represented by symbols



Source: <https://www.freepik.com/free-photos-vectors/cat-drawing>

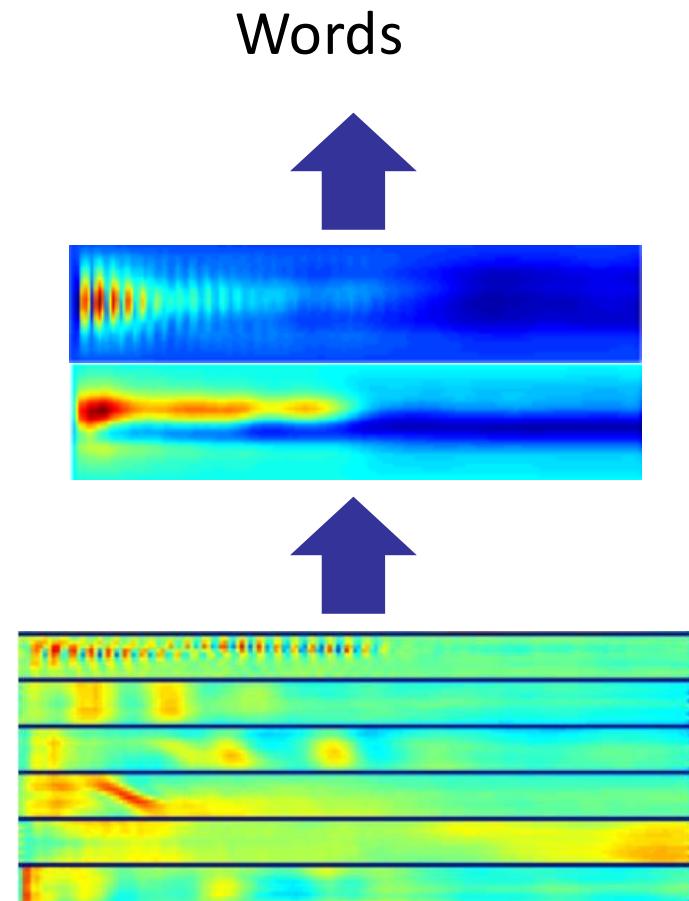
Deep Learning Breakthroughs by Hinton ... at the University of Toronto, 2009–2012



Deep Learning for Speech Recognition

- Speech recognition had first shown the breakthrough success of probabilistic methods: HMMs and GMMs
- The first breakthrough results of “deep learning” on large datasets also happened in **speech recognition**
- George Dahl et al. (2010/2012): Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition

Acoustic model \ WER	RT03S FSH	Hub5 SWB
Traditional GMM (D. et al. 2012)	27.4	23.6
Deep Learning (Dahl et al. 2012)	18.5 (-33%)	16.1 (-32%)
Deep Learning (Saon et al. 2017)	8.0 (-71%)	5.5 (-77%)



“(Artificial) neural (network)” or “deep learning” models for word meaning

We learn to represent a word as a vector of numbers

$$\text{versatile} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Similar vectors = similar meaning

Learn vectors via distributional similarity

How do we learn those word vectors?



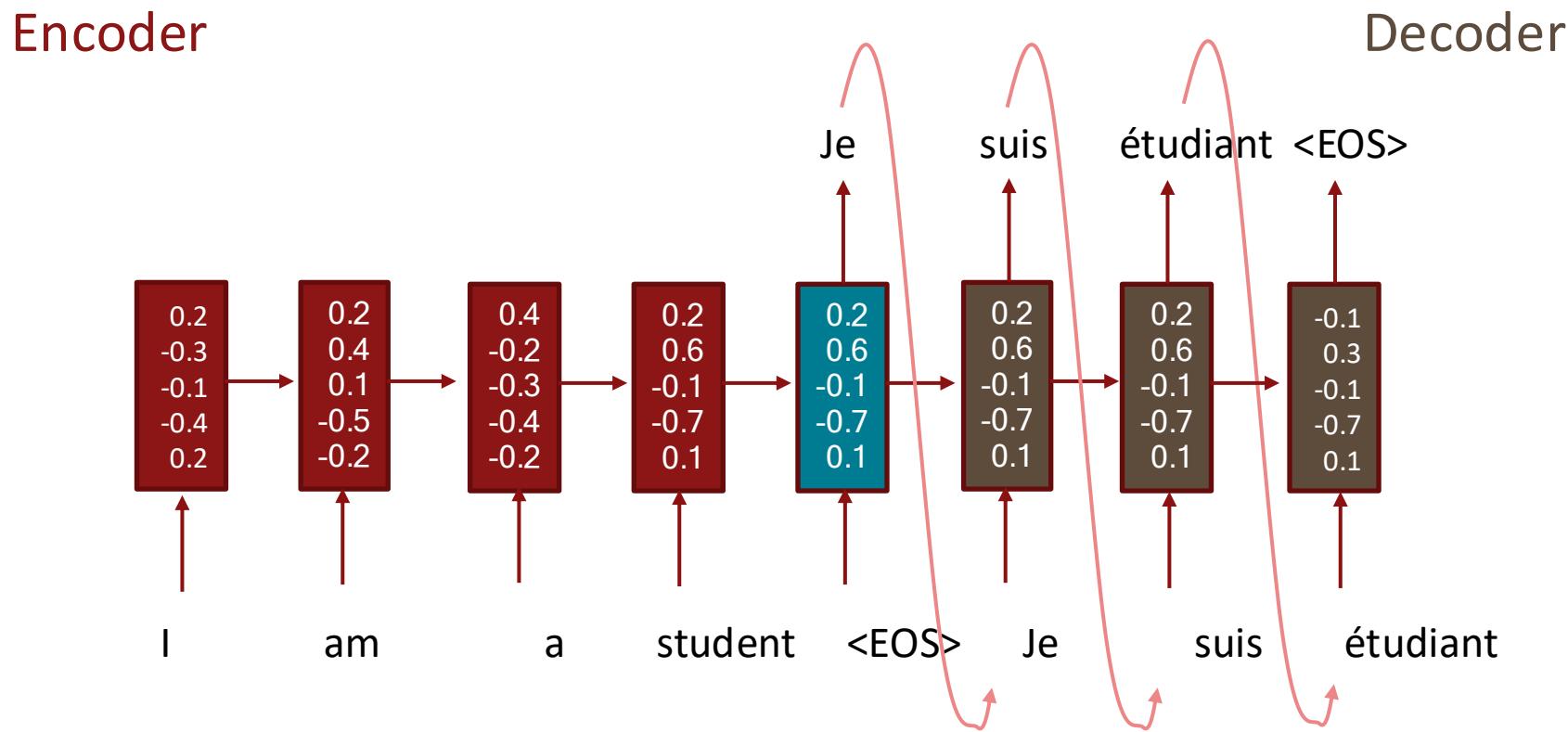
“You shall know a word by the company it keeps” (J. R. Firth 1957: 11)

any devices with a web browser, from laptops and tablets to smart phones
Users can download it for home computers or laptops from Microsoft Update website

↖ These words will represent *laptops* ↗

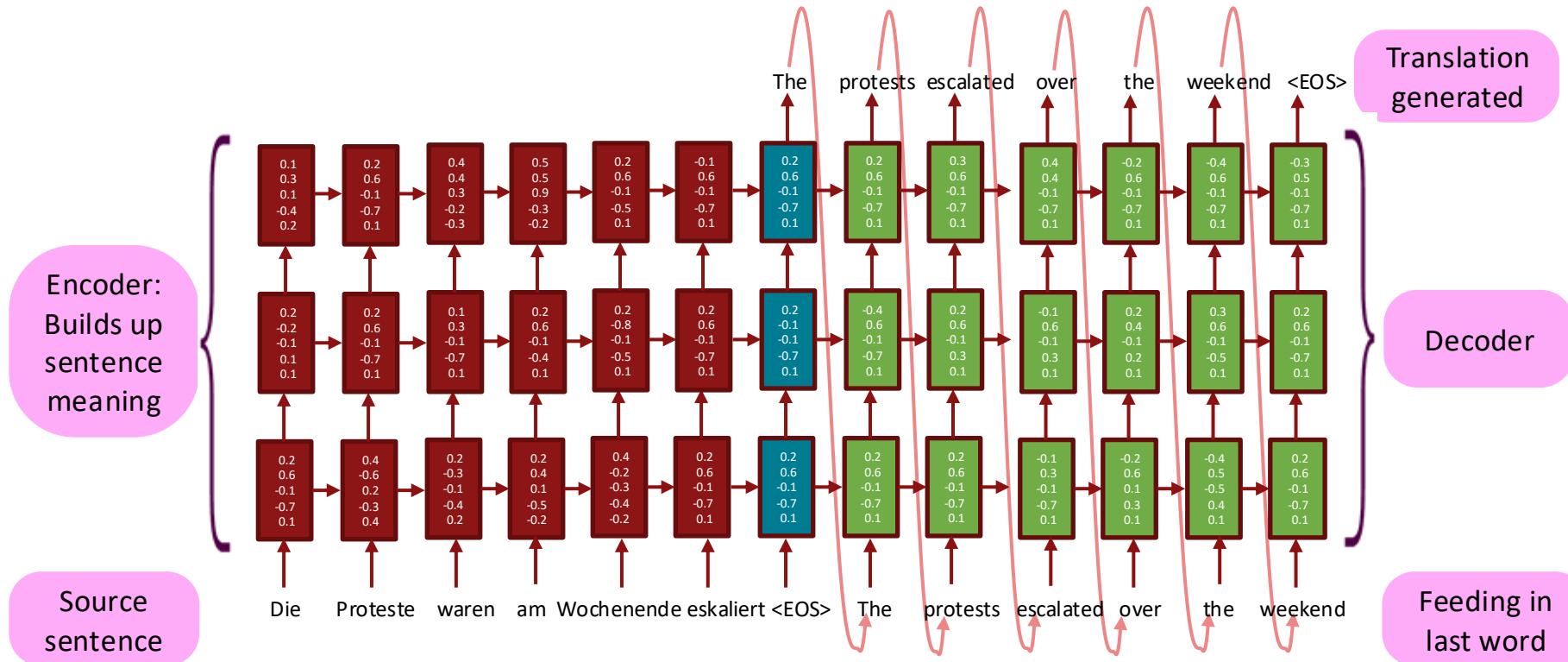
Defining similarity via contextual distributions in texts is one of the most successful ideas of modern computational linguistics

An RNN encoder-decoder network



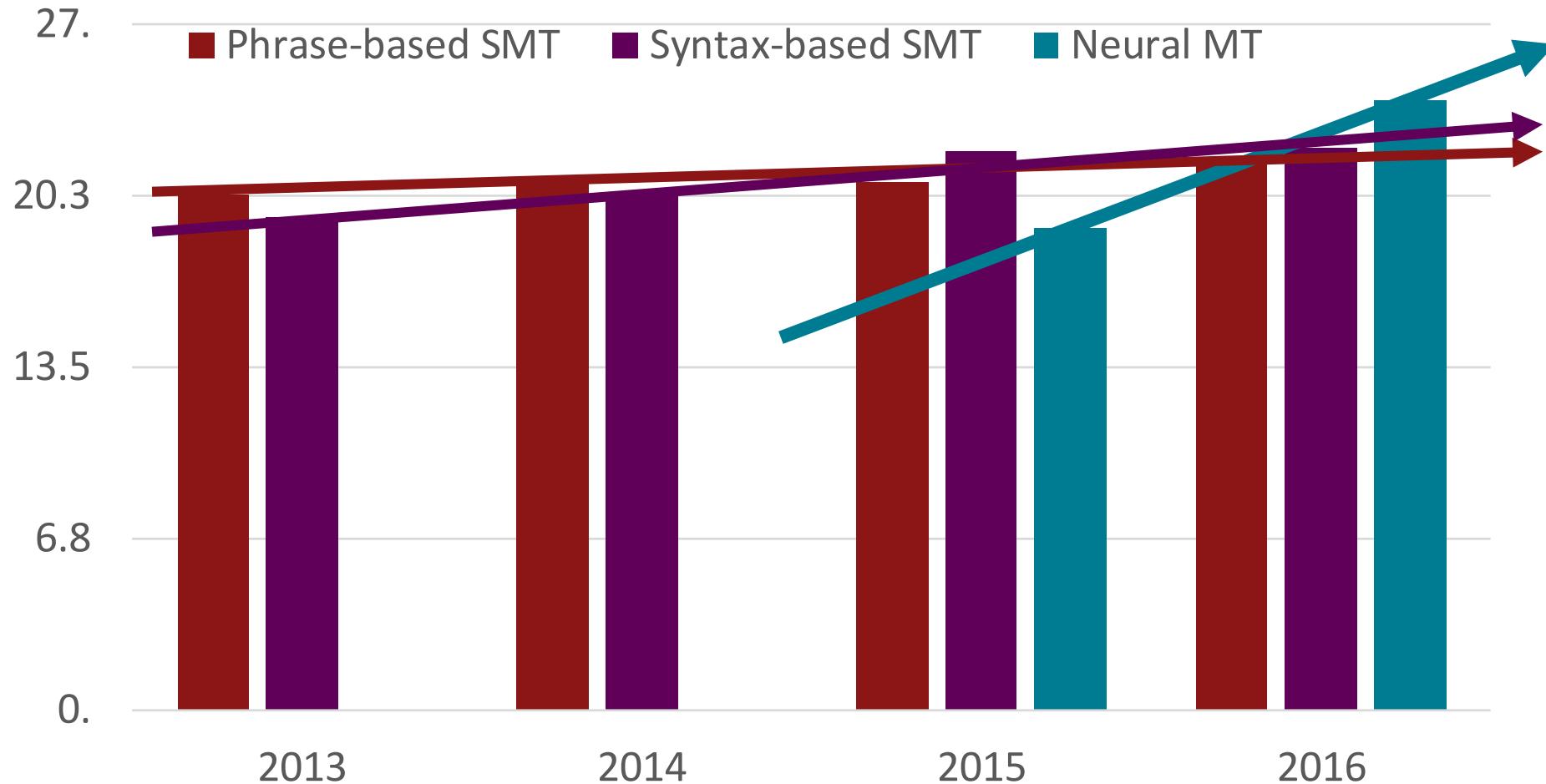
An LSTM encoder-decoder network

[Sutskever et al. 2014]



MT progress over time

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]



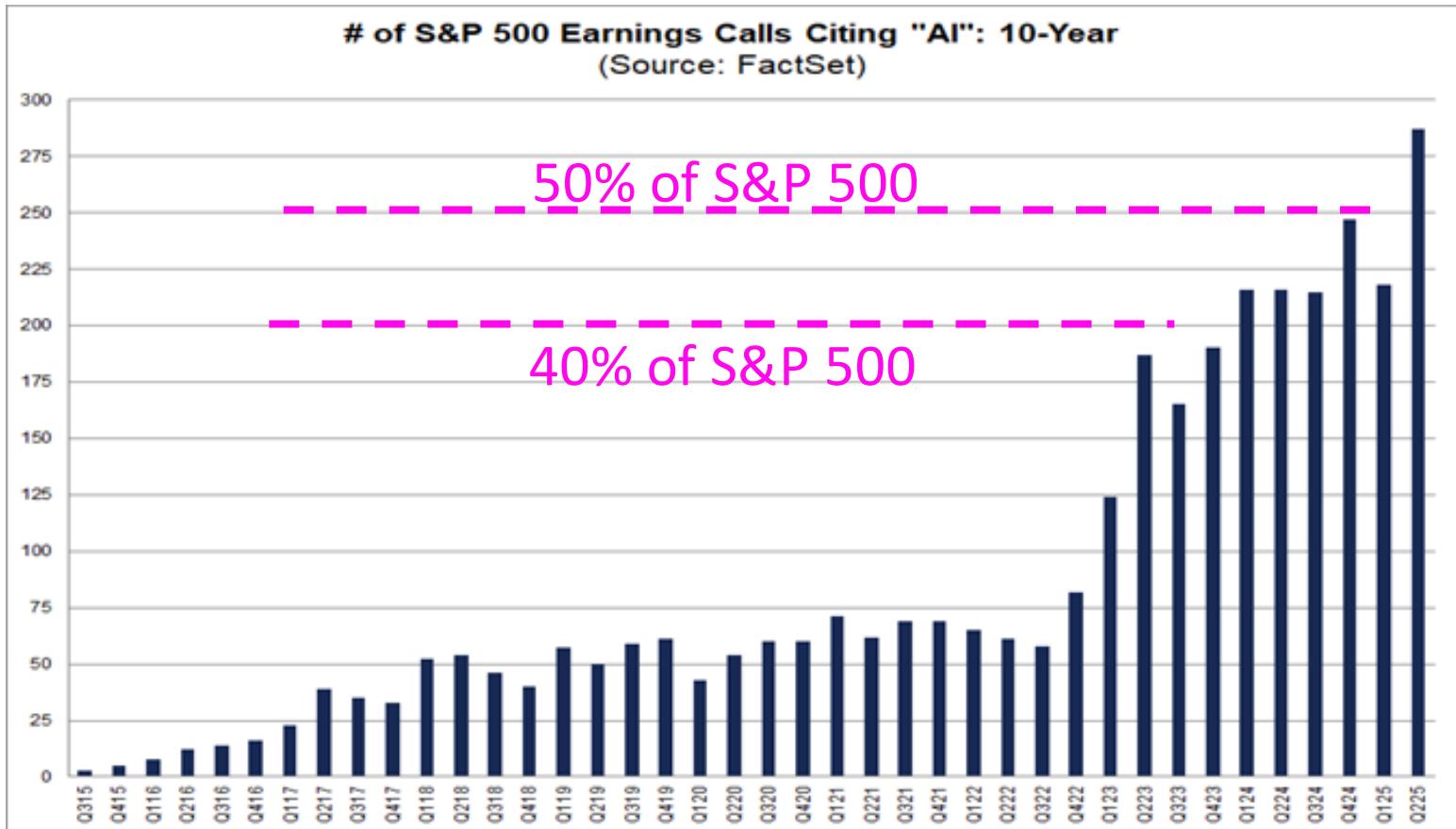
Source: http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

4. Deep Learning or Artificial Neural Networks for NLP

4b. Large Language Models

2022–present

The Rise of Generative AI: S&P 500 Companies Discussing “AI”



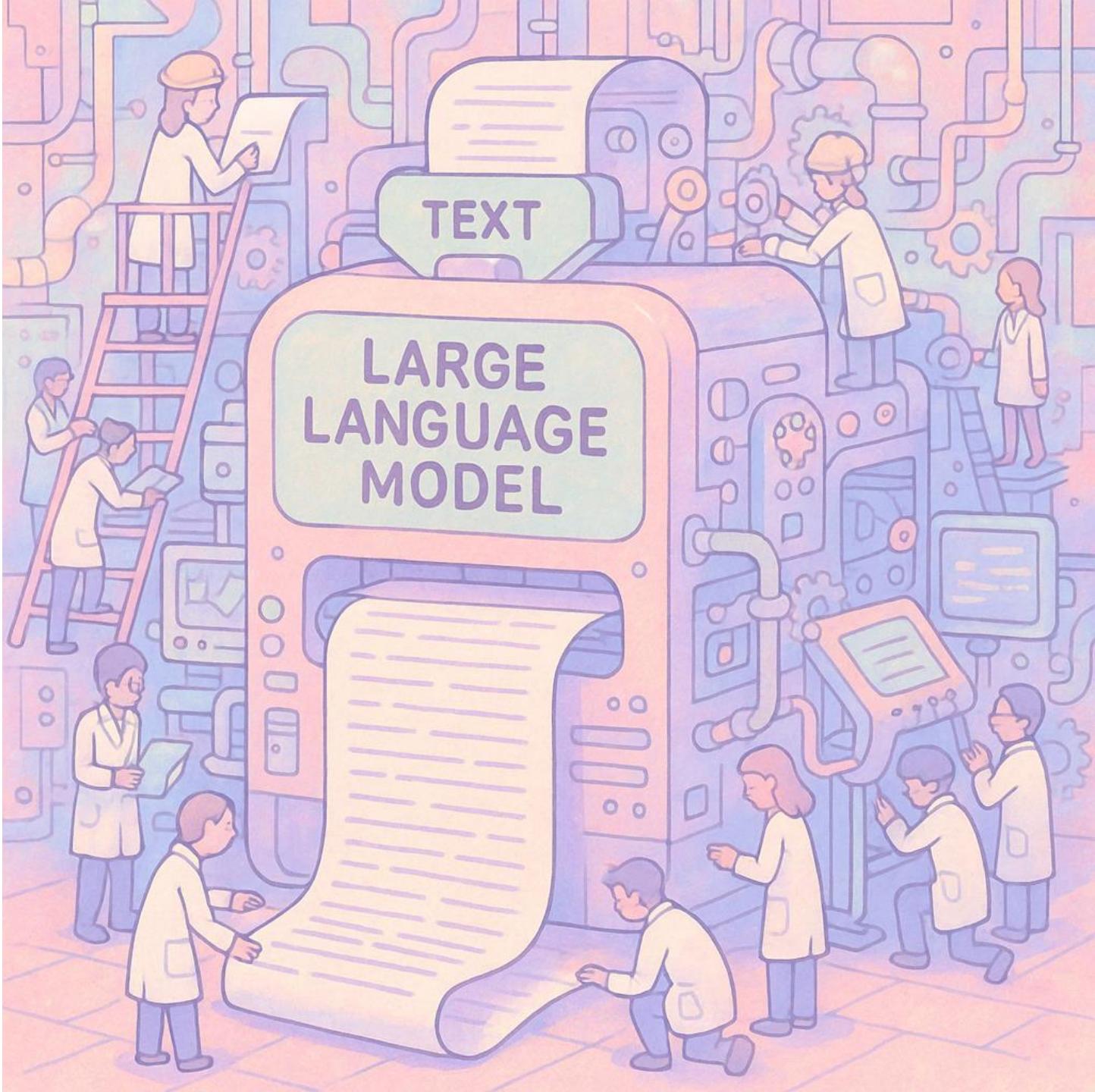
| 2024 | 2025

ChatGPT launch

Large Language Models

Source: GPT 4o

An abstract image in pastel manga style of a large language model as a big machine that takes in text and produces fresh writing on an endless scroll and which is maintained by a large number of scientists and engineers.



The History of Language Models

1913

Andrey A. Markov explores consonant-vowel transition probabilities in Alexander Pushkin novel, *Eugene Onegin*, developing Markov Models



1948

Claude E. Shannon develops *A Mathematical Theory of Information*; explores character/word-level n -gram models, entropy, text generation



1975

Frederick Jelinek's group at IBM defines and names the modern concept of a (probabilistic) “language model” for next token prediction. Used for spelling correction, speech recognition, machine translation....



But this technology wasn't considered a pathway to machines with artificial intelligence: For that we needed models of memory, knowledge representations, planning systems, and reasoning over abstract concepts

The History of Large Language Models

1998	CPAT-Tree-Based Language Models with an Application for Text Verification in Chinese. ROCLing 1998. First use of “LLM” trigram I know of; 200M word corpus
2000	A Neural Probabilistic Language Model. Bengio, Ducharme & Vincent NIPS 2000 First neural language model built on 32 million token corpus, 31K vocab
2007	Large Language Models in Machine Translation. Brants, Popat, Xu, Och and Dean . EMNLP 2007. 2 trillion token corpus n-gram model of up to 5-grams
2018	GPT (Radford, Narasimhan, Salimans & Sutskever) and BERT (Devlin, Chang, Lee & Toutanova). 3.3 billion token corpus
2020-	100+ billion parameter neural language models trained on > 1 trillion tokens: GPT-3, GPT-4, PaLM 2, Llama 3, Nemotron-4,

The History of Large Language Models

1998	CPAT-Tree-Based Language Models with an Application for Text Verification in Chinese. ROCLing 1998. First use of “LLM” trigram I know of; 200M word corpus	
2000	A Neural Probabilistic Language Model. First neural language model. Vincent NIPS 2000, 31K vocab	Not enough compute!
2007	Large Language Model. Och and 5-grams	Not enough model flexibility!
2018	GPT (Radford, Narasimhan, Lee & Toutanova). 3.3 billion parameters. Yang, 2018	Not enough data!
2020-	100+ billion parameter neural language models. GPT-3, GPT-4, PaLM 2, Llama 3,...	LLMs go brrr!

ChatGPT 4o

Write a sonnet praising Wikipedia in which each line begins with the letter W

Wikipedia: A Sonnet of Wonder

With wisdom vast, it spans the world so wide,
Wielding the wealth of knowledge, free to share.
Words weave through time where curious minds confide,
Whispering truths with patience, just and fair.

Wondrous it stands, a beacon shining bright,
Warding off darkness, doubt, and false pretense.
Wrought by the hands of many, day and night,
Witness to learning's boundless recompense.

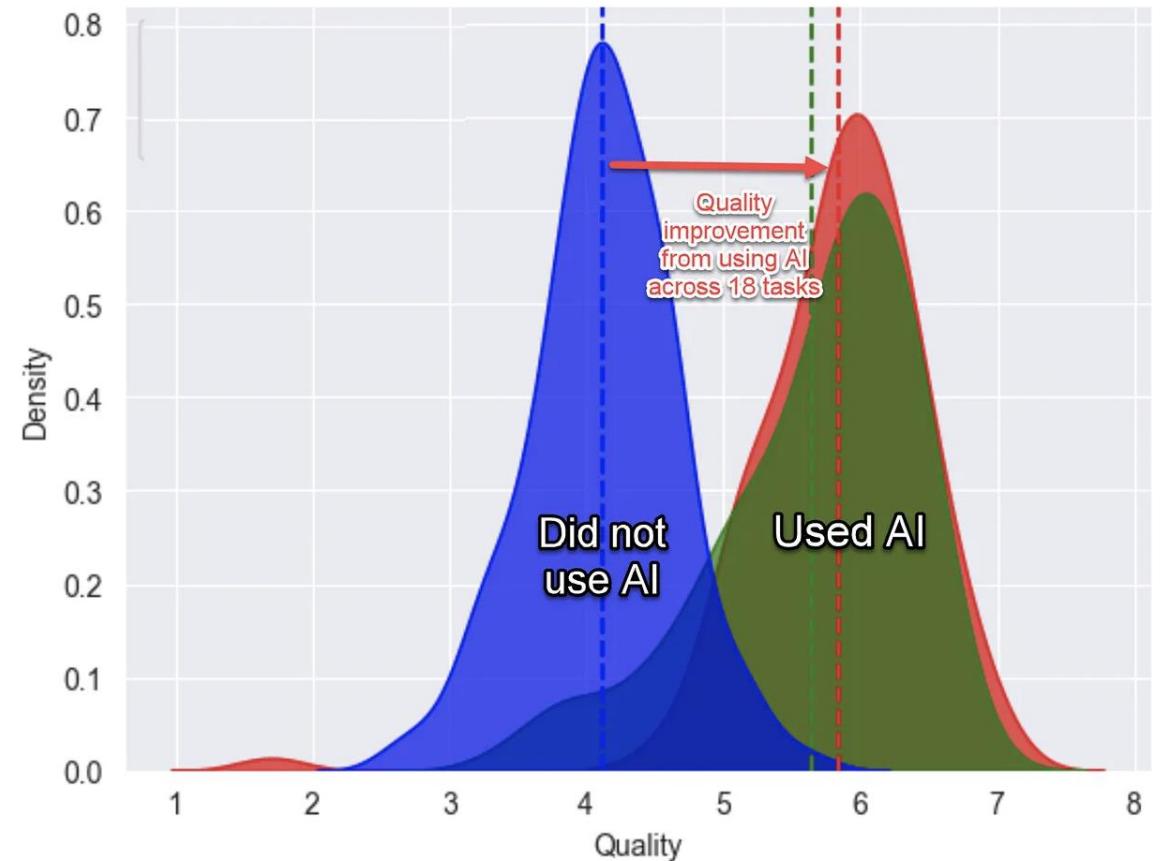
Who dares to claim the heights of thought alone?
Whence comes this trove but hearts in concert joined?
While scholars build on stones that others've known,
Waking the past, their voices intertwine.

Wisdom and will sustain this noble dream,
Wikipedia—our light, our endless stream.

Consultants using GPT-4 outperform non-users

- Consultants (from Boston Consulting Group) using GPT-4 **finished 12.2% more tasks** on average, **completed tasks 25.1% more quickly**, and produced **40% higher quality results** than those not using AI
- Use of an LLM particularly improved the performance of lower-performing humans
- Results vary depending on the task

— Dell'Acqua et al. 2023; Mollick 2023



Distribution of output quality across all tasks. Blue group did not use GPT-4; green and red groups used GPT-4; red group got additional training on how to use AI.

Can GPT-4 write fiction that matches the quality of *New Yorker* fiction?

Good news! No! (At least in 2023....)

GPT-4 is still 3–10x worse at creative writing!

Art or Artifice? Large Language Models and the False Promise of Creativity

TUHIN CHAKRABARTY, Columbia University, USA

PHILIPPE LABAN, Salesforce AI Research, USA

DIVYANSH AGARWAL, Salesforce AI Research, USA

SMARANDA MURESAN, Columbia University, USA

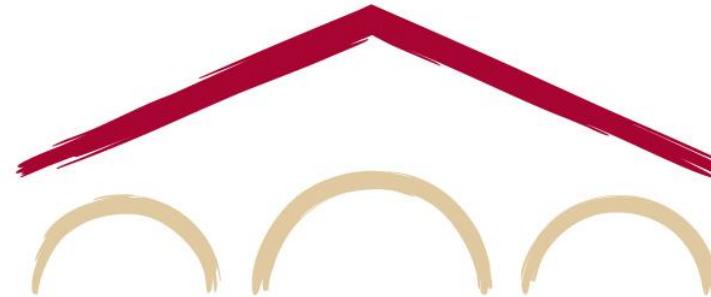
CHIEN-SHENG WU, Salesforce AI Research, USA

Researchers have argued that large language models (LLMs) exhibit high-quality writing capabilities from blogs to stories. However, evaluating objectively the creativity of a piece of writing is challenging. Inspired by the Torrance Test of Creative Thinking (TTCT) [61], which measures *creativity as a process*, we use the Consensual Assessment Technique [3] and propose *Torrance Test of Creative Writing* (TTCW) to evaluate *creativity as product*. TTCW consists of 14 binary tests organized into the original dimensions of Fluency, Flexibility, Originality, and Elaboration. We recruit 10 creative writers and implement a human assessment of 48 stories written either by professional authors or LLMs using TTCW. Our analysis shows that LLM-generated stories pass 3-10X less TTCW tests than stories written by professionals. In addition, we explore the use of LLMs as assessors to automate the TTCW evaluation, revealing that none of the LLMs positively correlate with the expert assessments.

<https://arxiv.org/abs/2309.14556>

History of Natural Language Processing

CS224N/Ling284



Christopher Manning

Lecture 1