



COS 484: Natural Language Processing

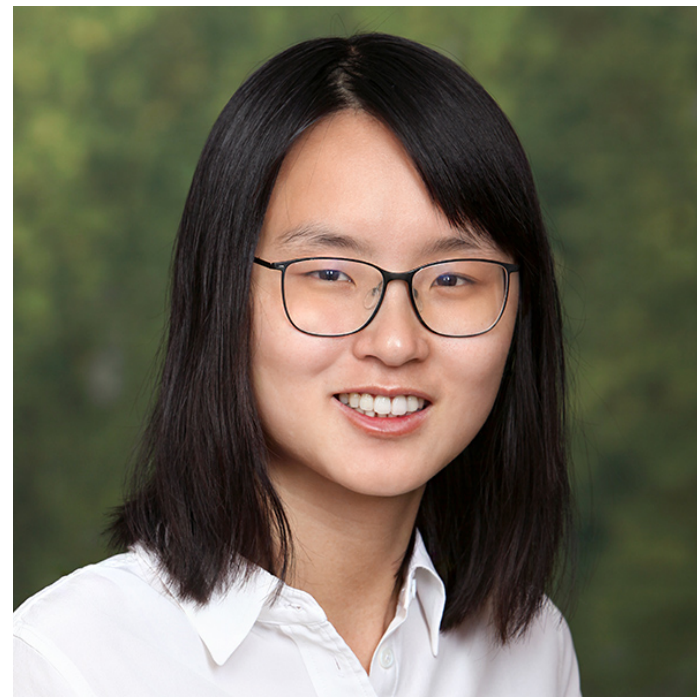
LI: Introduction to NLP

Spring 2023

Logistics and course structure

Course staff

Instructor



Danqi Chen

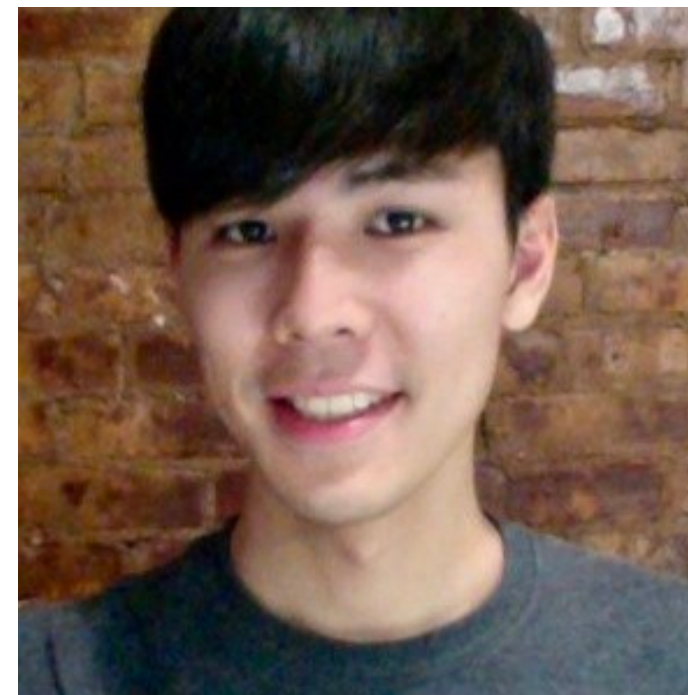
Graduate TAs



Alexander Wettig



Austin Wang



Howard Chen



Samyak Gupta

Logistics

Course webpage: <https://nlp.cs.princeton.edu/cos484/>

- Contains all the detailed information about the course (slides, reading lists, assignments, policy, office hours, etc).
- Canvas will be only used for announcements - make sure you have **notifications** turn on!

Schedule

slides will be available before the class

All readings are optional but can help strengthen your understanding

Lecture schedule is tentative and subject to change. All assignments are due **9:30am EST** before Monday lectures.

Week	Date	Topics	Readings	Assignments
1	Mon (1/30)	Introduction to NLP	1. Advances in natural language processing 2. Human Language Understanding & Reasoning	A0 out
	Wed (2/1)	n-gram language models	J & M 3.1-3.4	
2	Mon (2/7)	Text classification	J & M 4.1-4.8	A0 due, A1 out
	Mon (2/8)	Word embeddings 1	J & M 6.2-6.4, 6.6 Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors	

All assignments are due on Monday at 9:30 before the lecture

Logistics

- **Lectures:** Monday / Wednesday 9:30-10:50 at Bowen Hall 222 (no Zoom support)
- **Precepts:** 1-hour precept every week taught by TAs (optional)
 - Refresher for basic maths, review course materials, Colab/Pytorch tutorial etc
 - Likely Friday - we will send out a poll later today!



Logistics

- Sign up for **Ed** and **Gradescope** today
 - We will use **iClicker** for polls in the class... Will make an announcement before Wednesday!
- **Ed** will be the main forum for all class-related questions and discussion
- We also provide a mailing list cos484-2023@googlegroups.com for emergencies, or personal matters that you don't wish to put in a private Ed post.
 - The mailing list only has instructor + grad TAs. Please don't write to the instructor directly unless you HAVE TO.
- We would like to help in our office hours!
 - All the information will be posted this week on the website
 - We will also have a number of undergraduate TAs' office hours



Course structure

Assignments (40%): 4 + 1 = 5 in total

- **A0 (4%)** is a warm-up assignment and will be released TODAY and due in one week
- **A1, A2, A3, A4:** 9% each - each assignment has 2 or 3 weeks
- Every assignment has a **written** component and a **programming** component based on Colab (A4 requires access to GPUs).
- You will need to know how to program in PyTorch/Numpy.

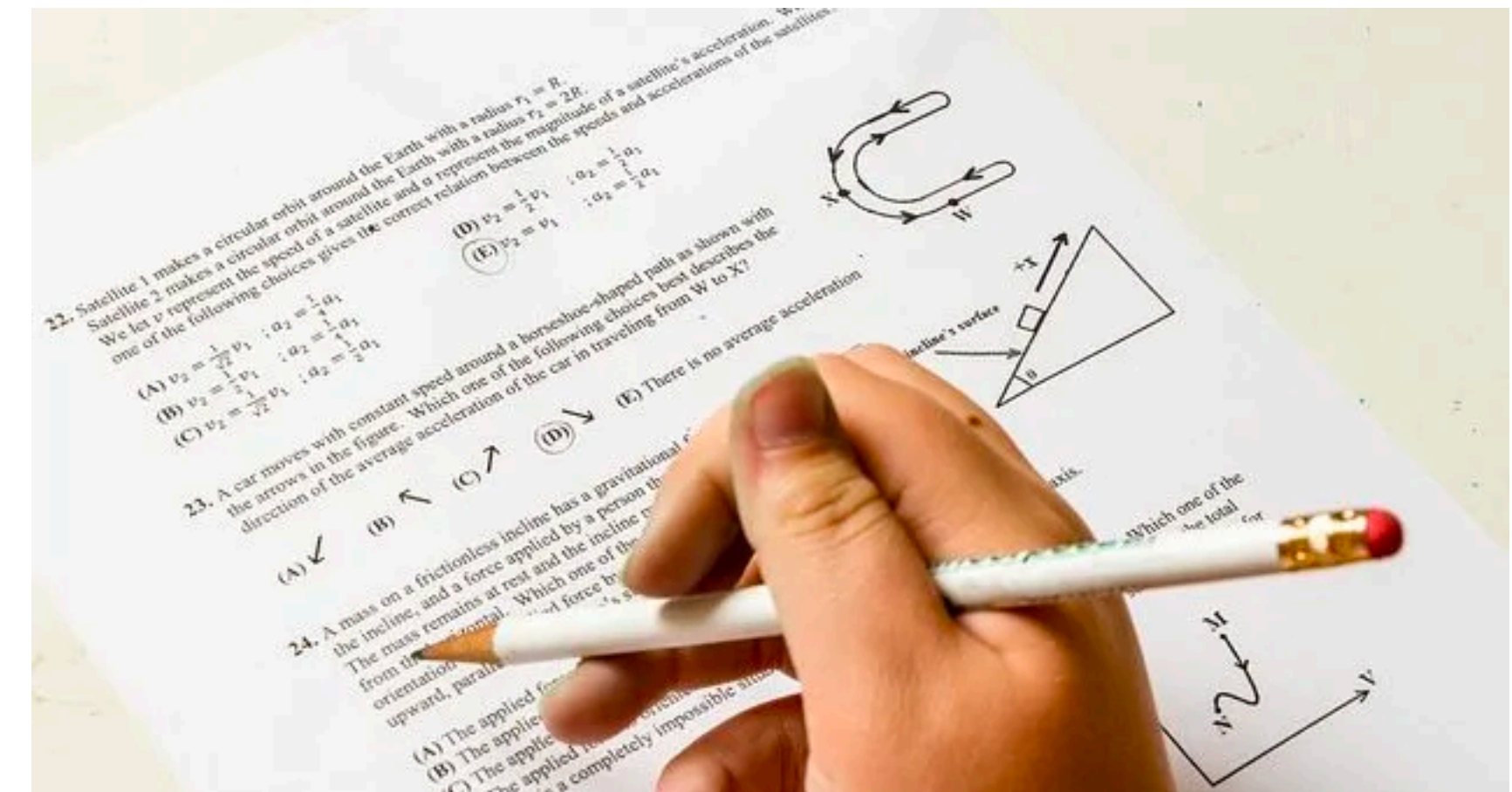
You have **96 free late hours** for all assignments; After that, 10% penalty for each late day (up to a maximum of 3 days beyond which submissions will not be accepted)



Course structure

Midterm (25%)

- A 3-hour timed exam on Gradescope in a 27-hour window (**March 8-9**)
- All the topics up to March 1 will be covered
- No final exam



Course structure

Final project (35%)

- Complete in a team of 3
- Two options: (a) reproducing a state-of-the-art NLP paper (ACL/EMNLP/NAACL 2020-2022) and coming up with your analysis, ablations or innovations (**strongly encouraged**) (b) competing a research project
- Proposal (0%) due on March 20
- Poster presentation tentatively scheduled on May 3 (during reading period)
- Final report due on May 9

Extra bonus (5%) - participation in class and Ed discussions

*No **pre-determined cut-offs for final grades**, will be decided at the end taking into account the performance of the entire class and will be fairly assigned to measure your level of understanding of the subject.*

Textbooks

(NLP is a rapid-moving field...)

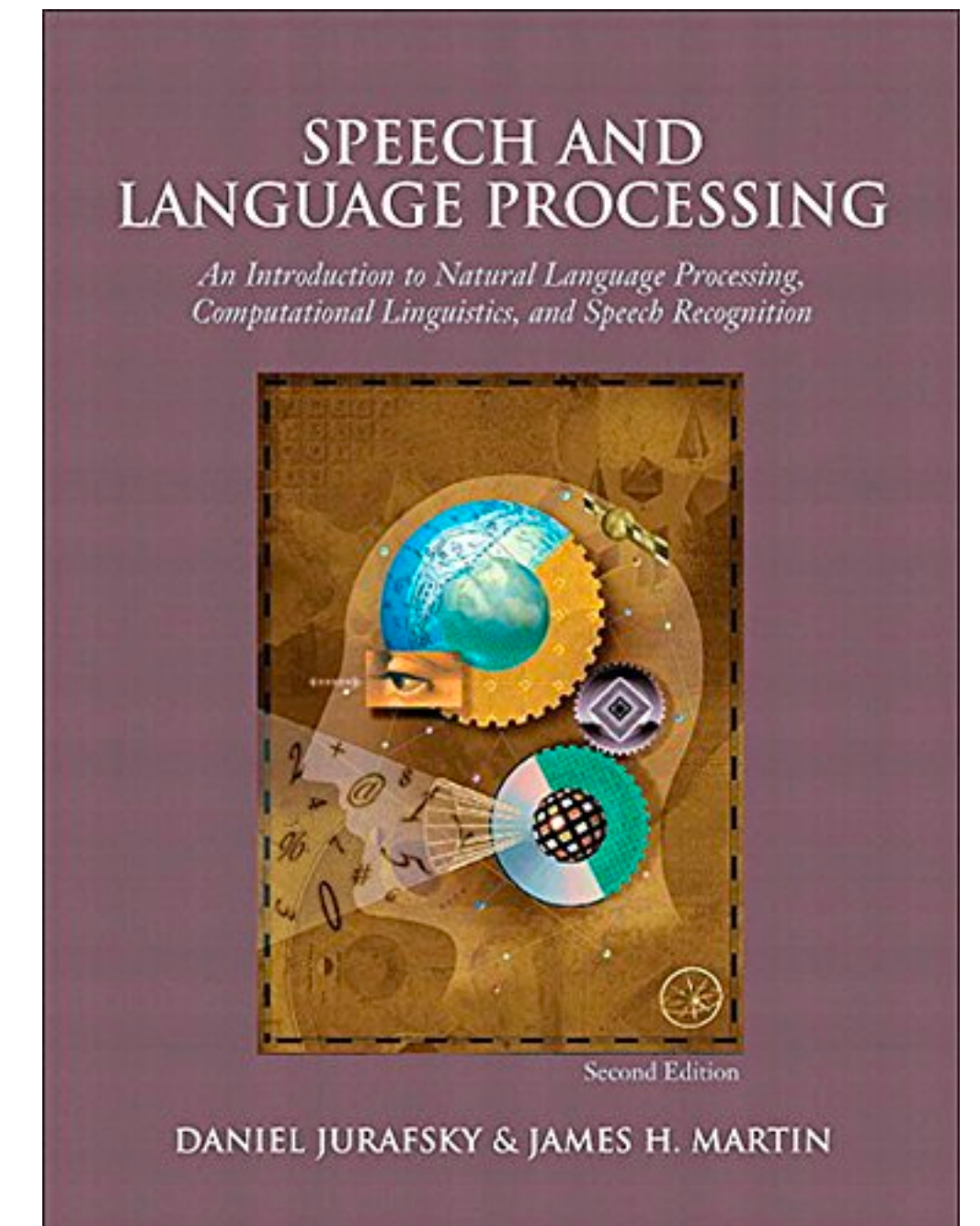
<https://web.stanford.edu/~jurafsky/slp3/>

Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

 Here's our **Jan 7, 2023 draft!** This draft is mostly a bug-fixing and restructuring release, there are no r the applications section earlier, reflecting how we and others tend to teach NLP, and combines the linguisti

A good way to learn about state-of-the-art NLP concepts is through **research papers** and **blog posts**

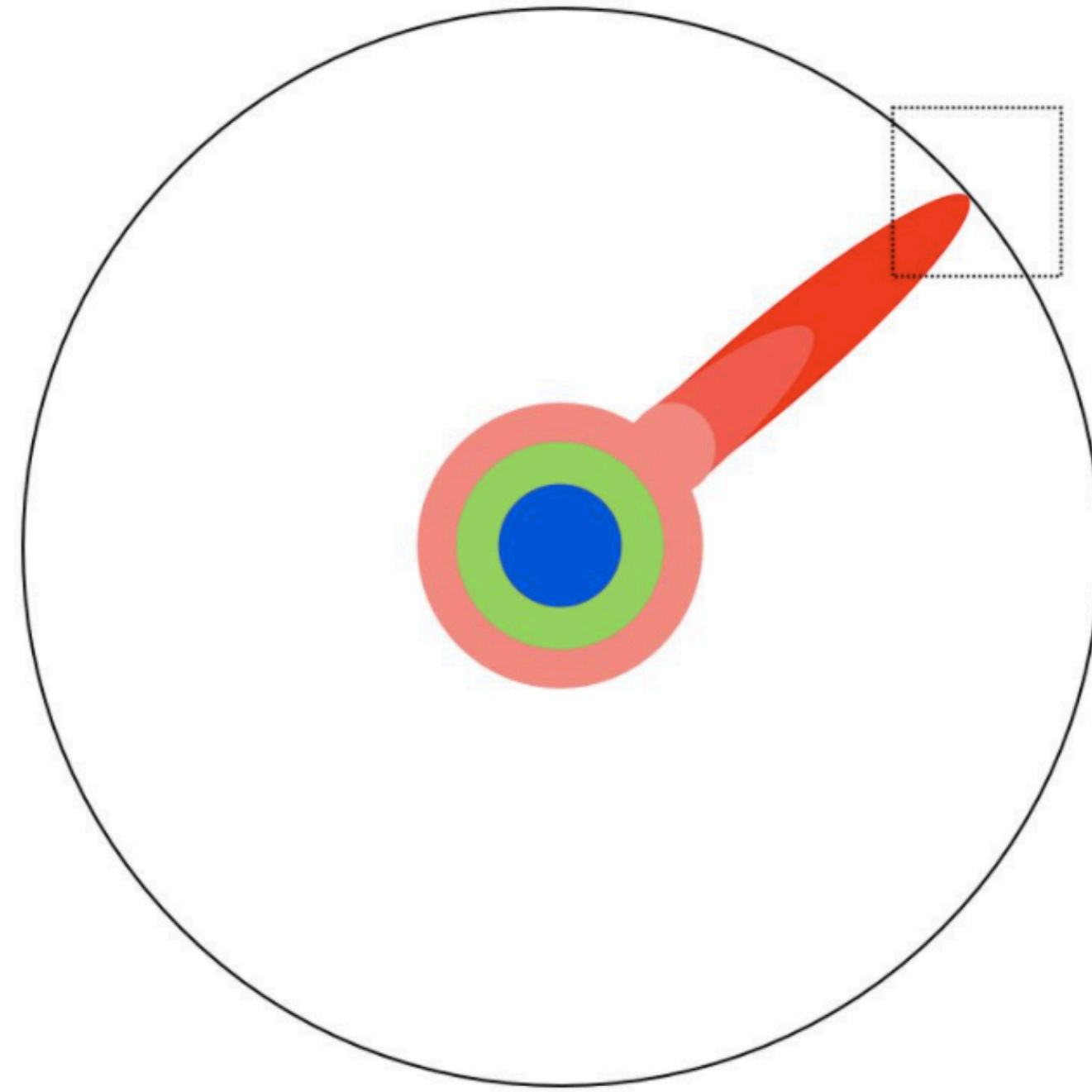



2nd edition ✘

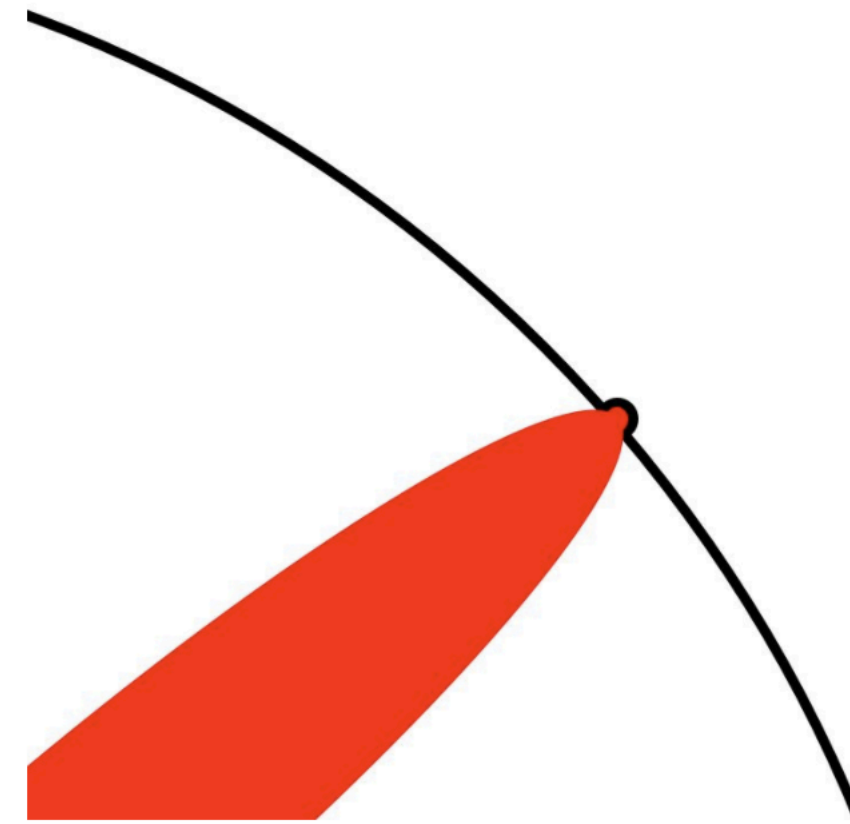
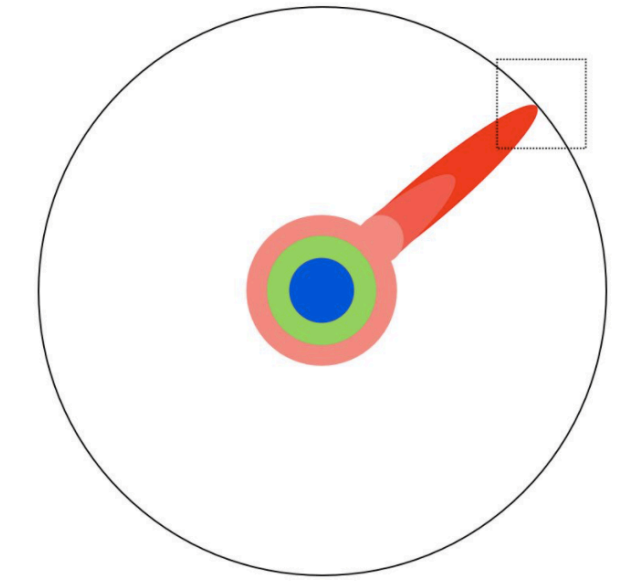
Course goals



- Gain an understanding of the **fundamentals** of different sub-fields within NLP
- Understand **theoretical** concepts and algorithms
- **Hands on experience** building statistical models for language processing
- Carry out an **independent research project** at the end



This is an advanced class



This is a *very* advanced class

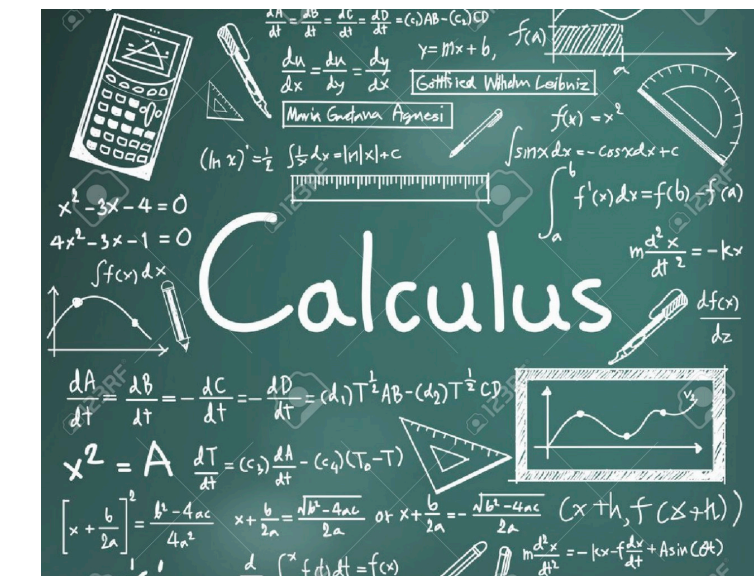
Prerequisites

- **Required: COS324**, knowledge of probability, linear algebra, calculus (A0 will give you a sense)
- Be ready to pick up new ML concepts
- Proficiency in Python: programming assignments and projects will require use of Python, Numpy and PyTorch.

Q. Why is COS324 a prerequisite?

We assume you have learned the following concepts already:

- Language models
- Logistic regression w/ regularization
- Unsupervised vs supervised learning
- Feedforward neural networks, convolutional neural networks
- PyTorch programming
- (A little bit of reinforcement learning)



What is NLP and what is the course about

Natural Language processing

- NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)
- NLP is an interdisciplinary field

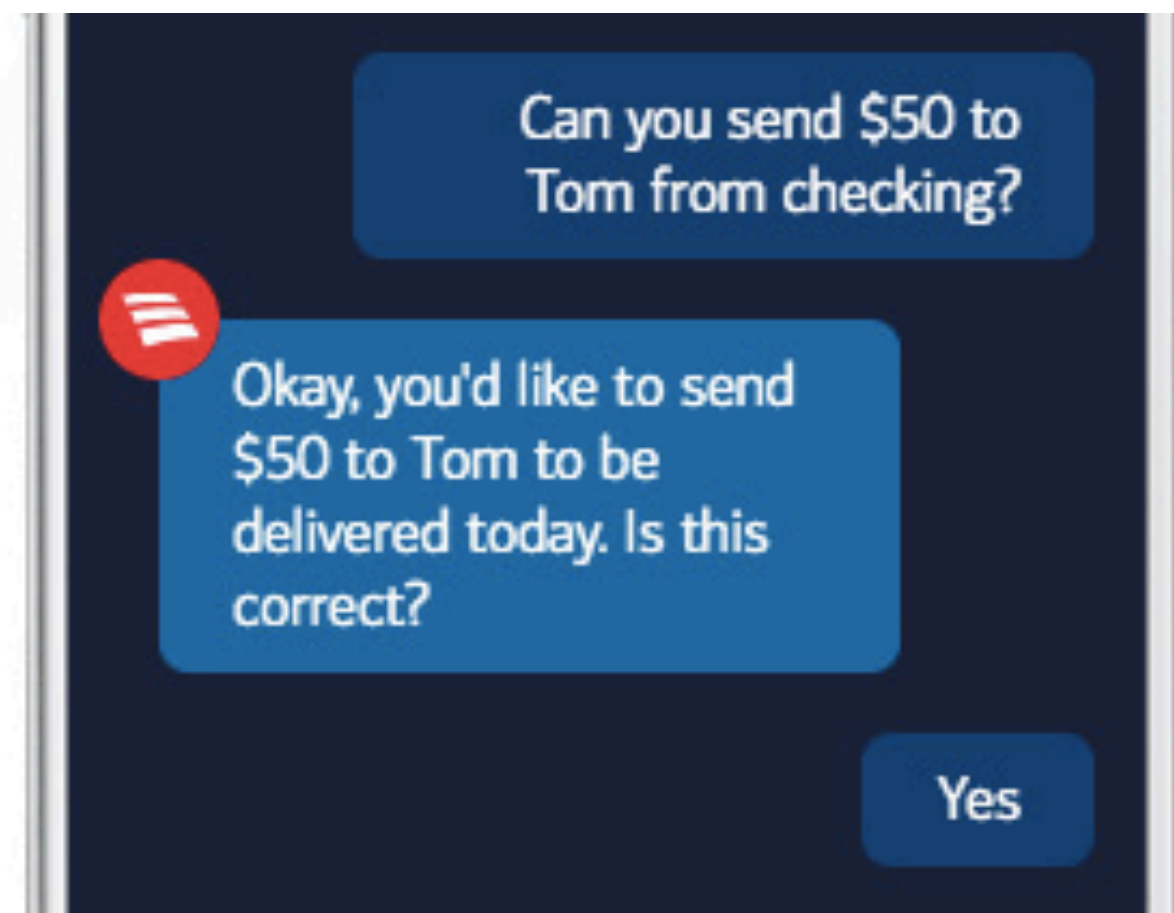


Natural Language processing

- NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)

Communication with humans (ex. personal assistants, customer service)

Access the wealth of information about the world — crucial for AI systems



Banking assistant

ONLINE



OFFLINE

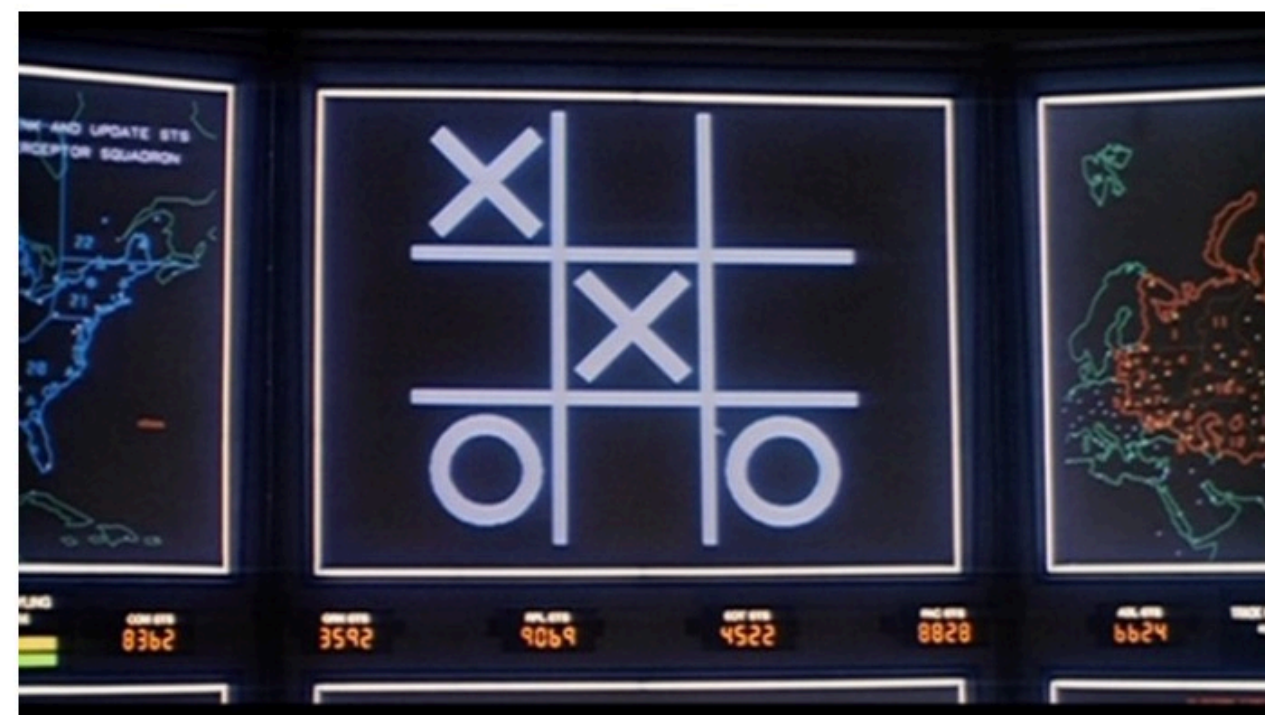
Computer learns to play Civilization by reading the instruction manual

By Matthew Rogers on July 14, 2011 at 5:03 pm | [16 Comments](#)






532 SHARES

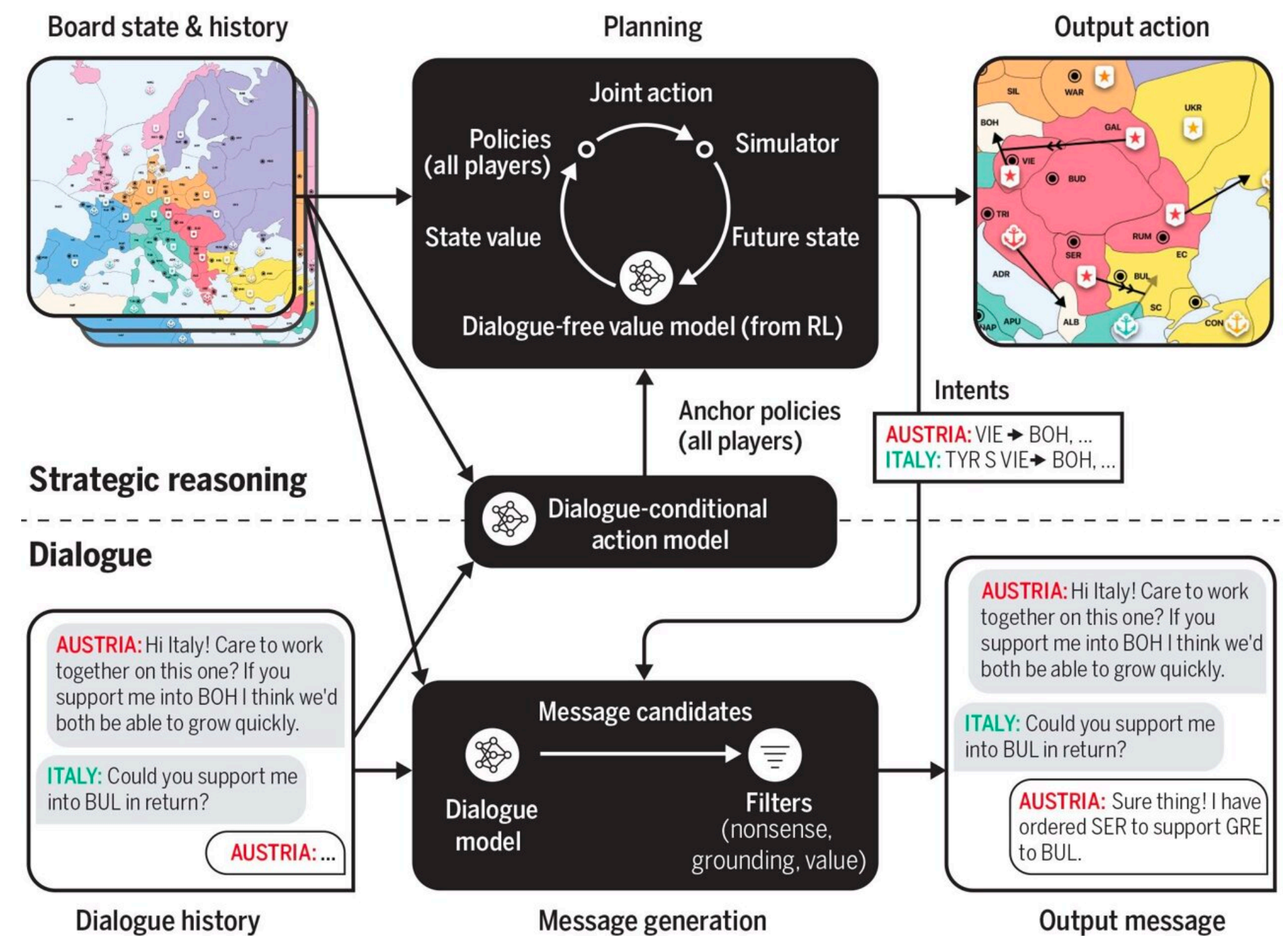


MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

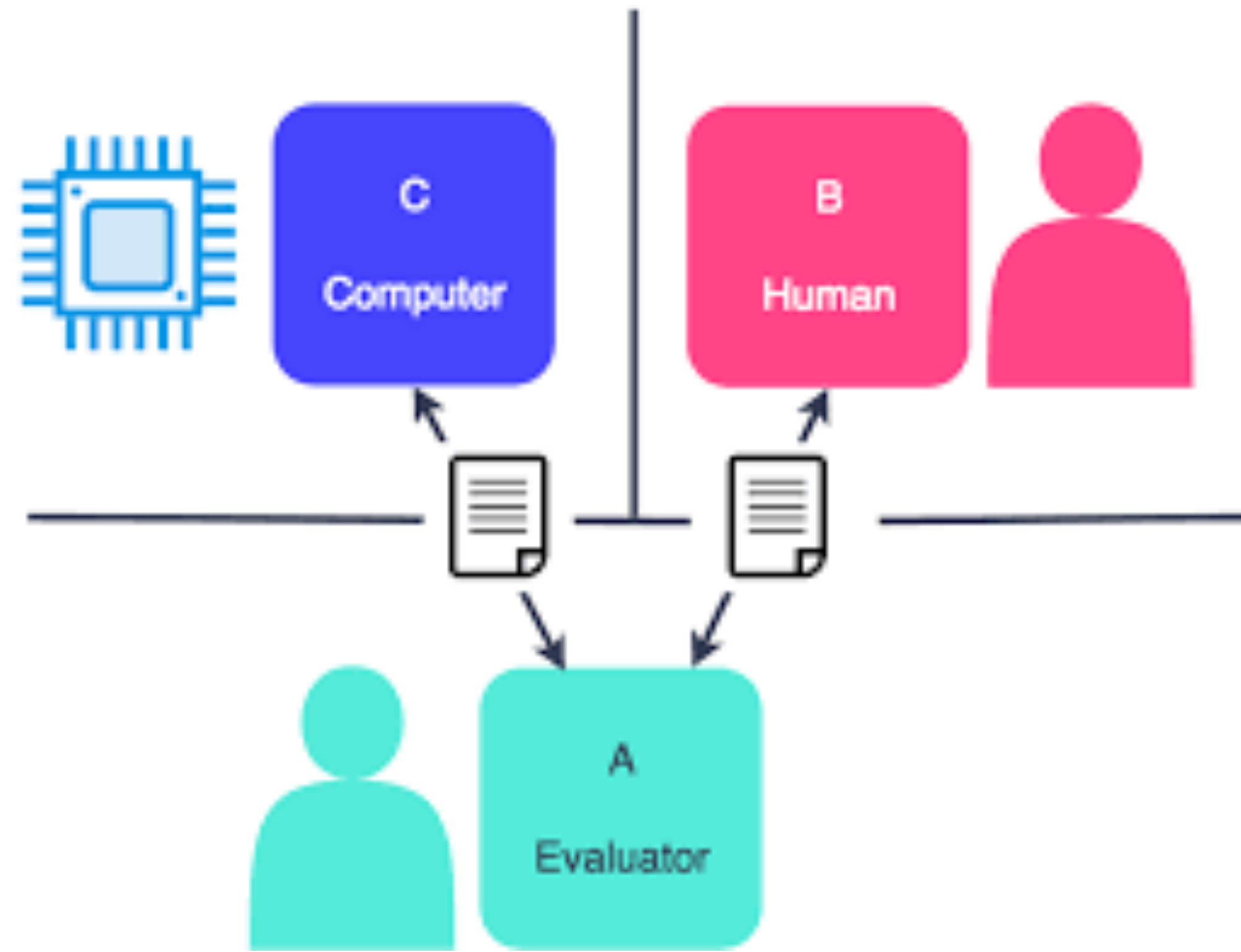
The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization

Meta's New AI Ranked in the Top 10% at the Game 'Diplomacy'—and Human Players Were None the Wiser

By Edd Gent > November 28, 2022



Turing Test

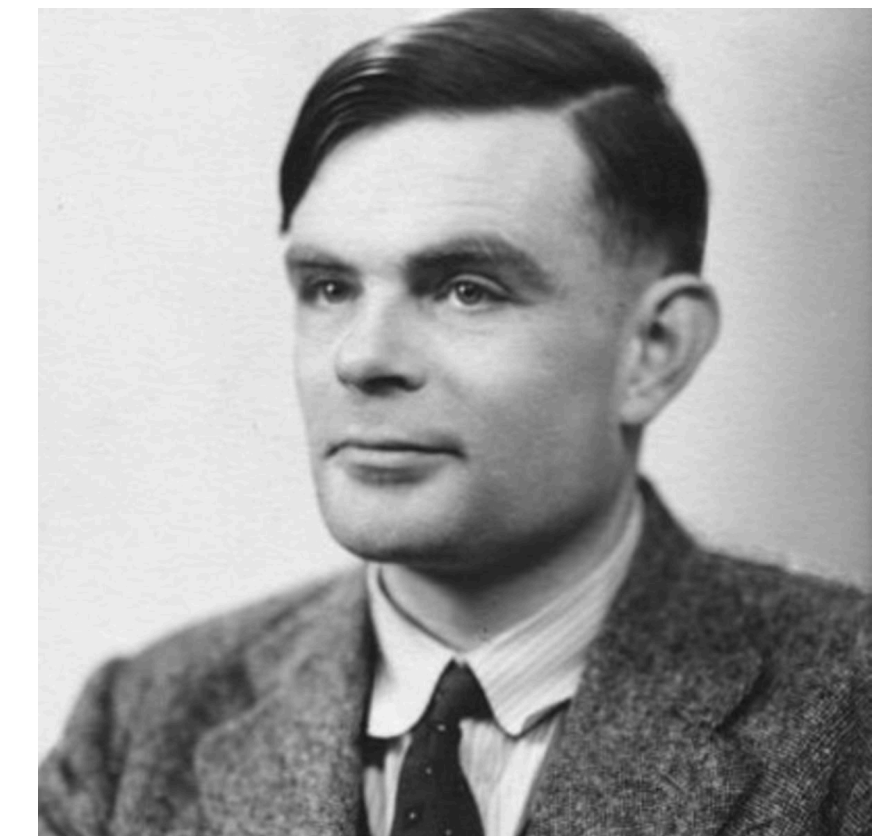


A. M. Turing (1950) *Computing Machinery and Intelligence*. *Mind* 49: 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

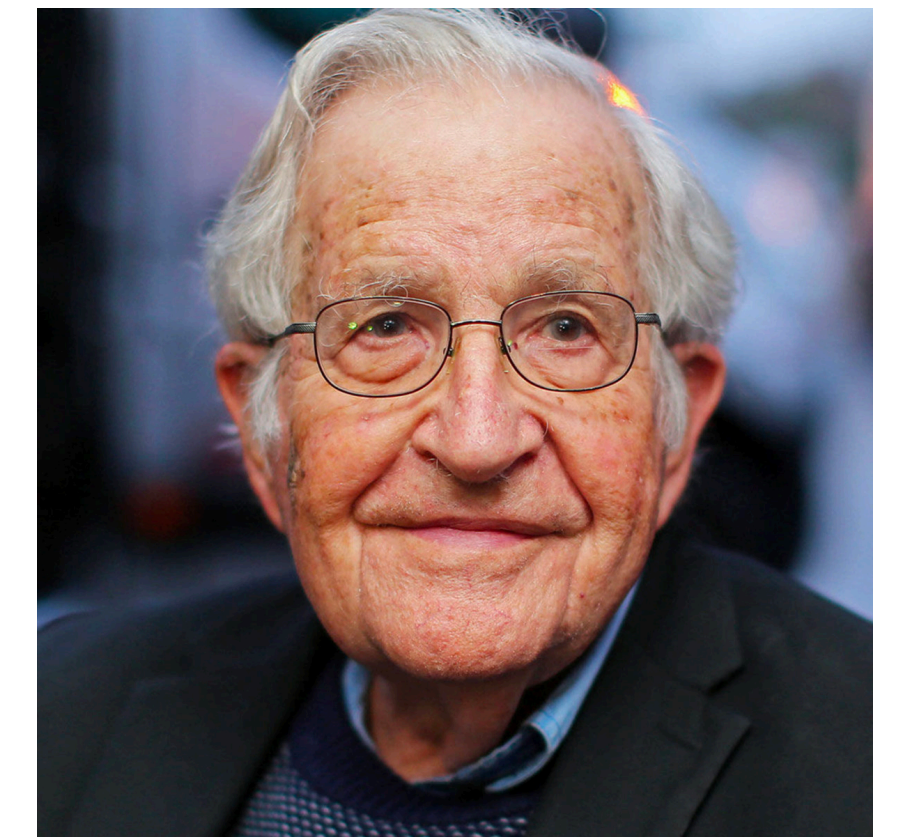
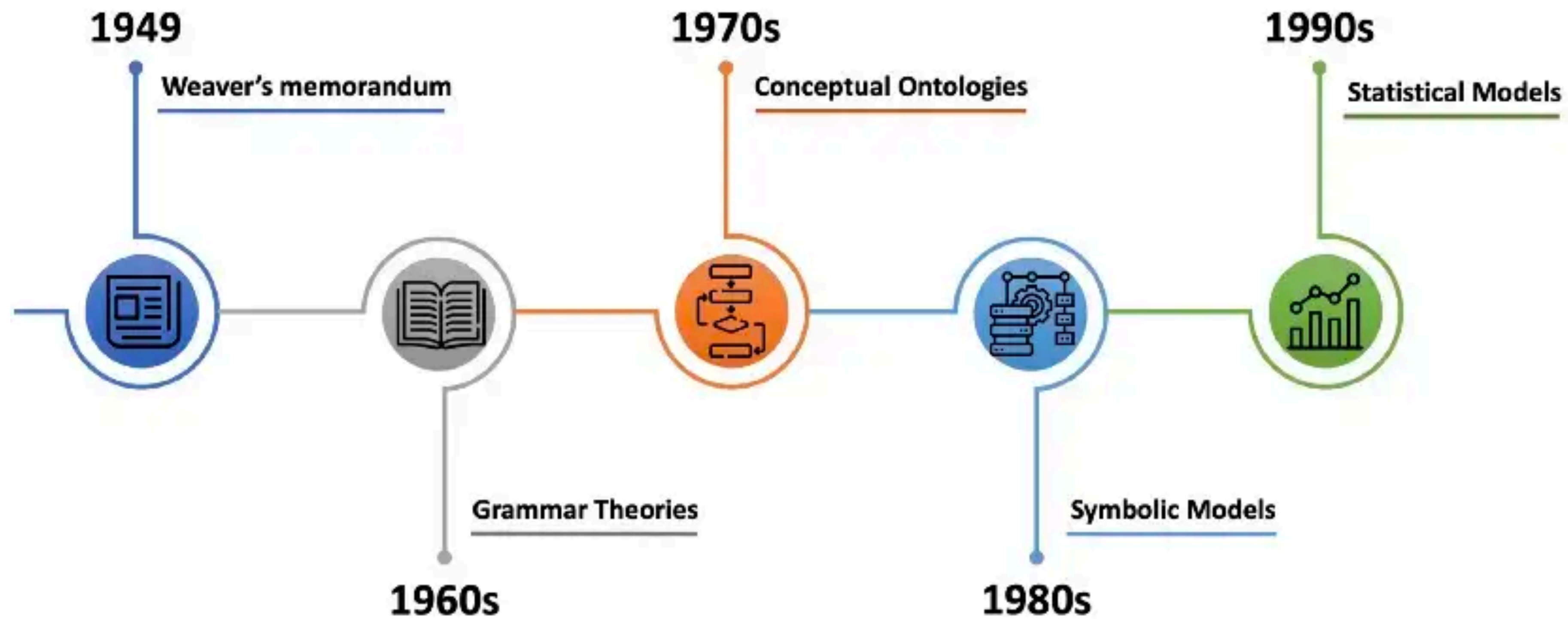
By A. M. Turing

1. The Imitation Game

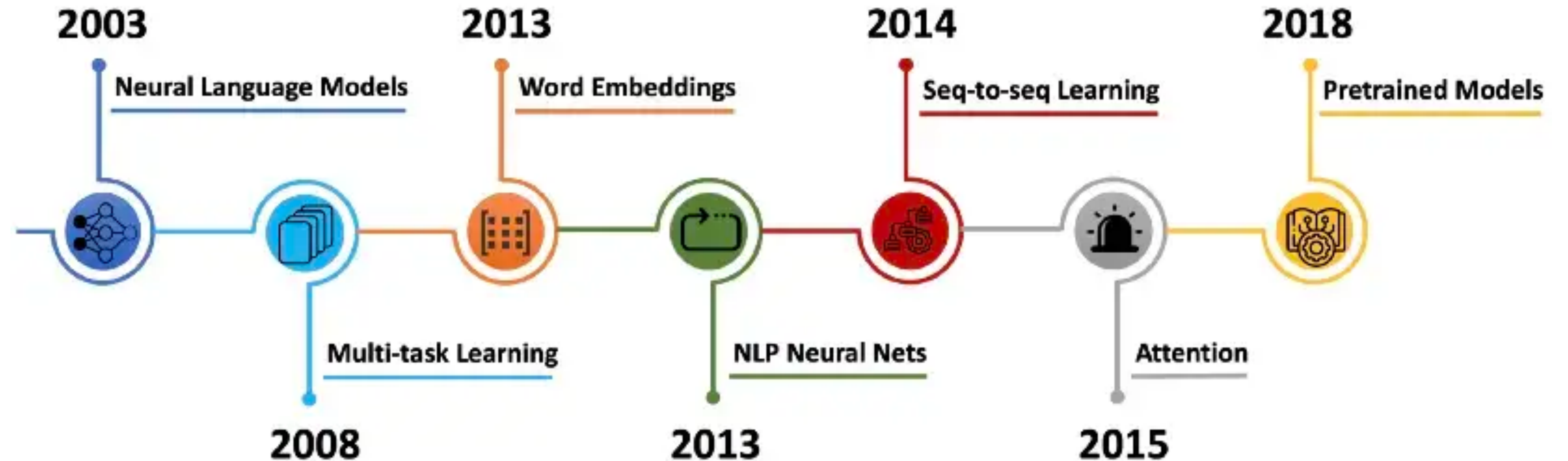


Ability to understand and generate language ~ intelligence

A brief history of NLP



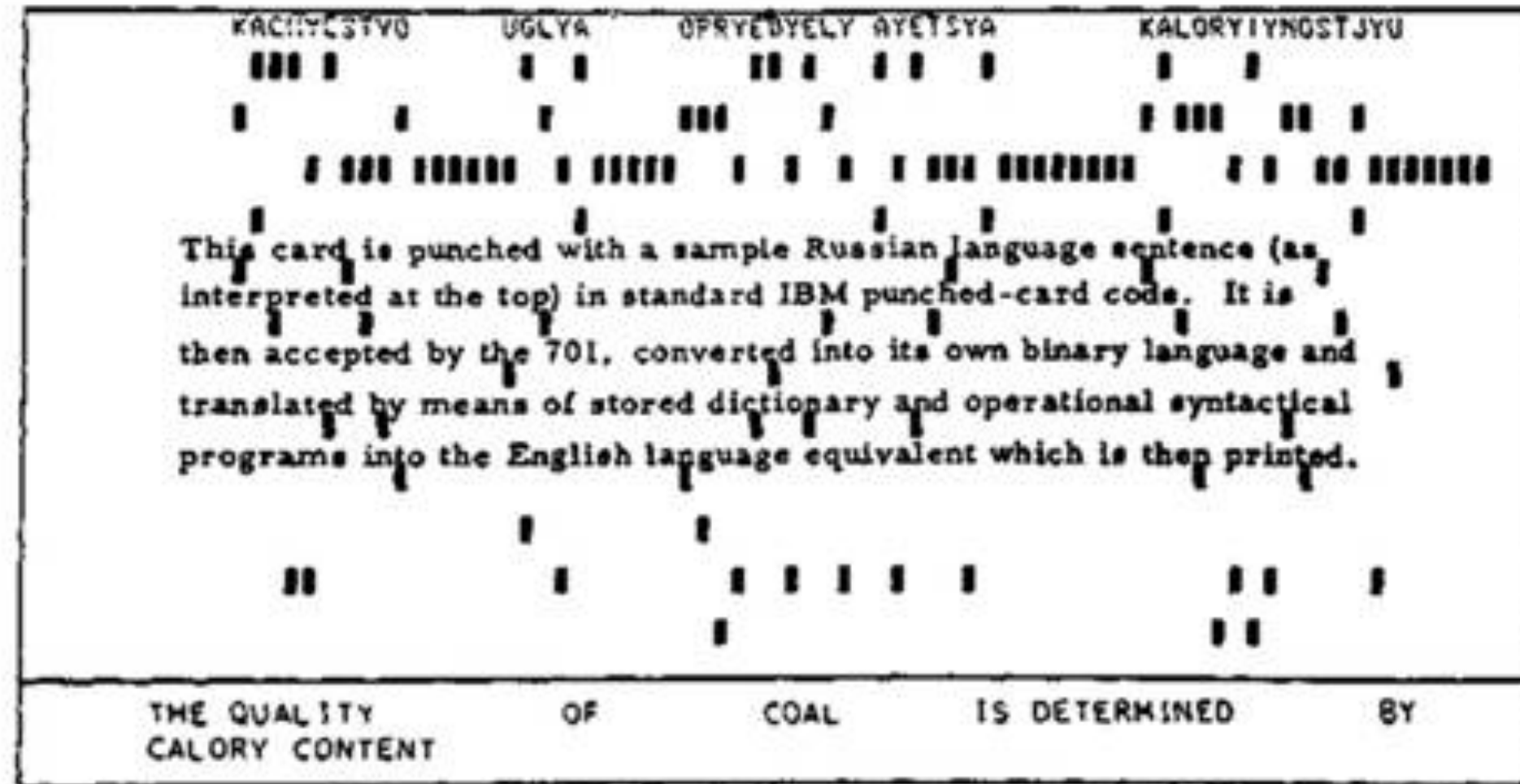
A brief history of NLP



2018: BERT 2019: T5, RoBERTa 2020: GPT-3 2022: ChatGPT

How “neural” is this course?

How it started



Specimen punched card and below a strip with translation, printed within a few seconds

Georgetown experiment 1954

“Within three or five years, machine translation will be a solved problem”

How it's going

DETECT LANGUAGE YORUBA **ENGLISH** MALAY

↔ **HINDI** YORUBA ENGLISH

The man shot the elephant while wearing his pyjamas.

उस आदमी ने हाथी को अपना पजामा पहनाया।

us aadamee ne haathee ko apana pajaama pahanaaya.

DETECT LANGUAGE **HINDI** ENGLISH YORUBA

↔ HINDI YORUBA **ENGLISH**

उस आदमी ने हाथी को अपना पजामा पहनाया।

The man put his pajamas on the elephant.

Text Documents Websites

DETECT LANGUAGE HINDI **ENGLISH** KOREAN

↔ ENGLISH **HINDI** KOREAN

The man shot the elephant while wearing his pajamas

आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी

aadamee ne apana pajaama pahane hue haathee ko golee maar dee

51 / 5,000

51 / 5,000

Text Documents Websites

DETECT LANGUAGE ENGLISH **HINDI** KOREAN

↔ HINDI **ENGLISH** KOREAN

आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी

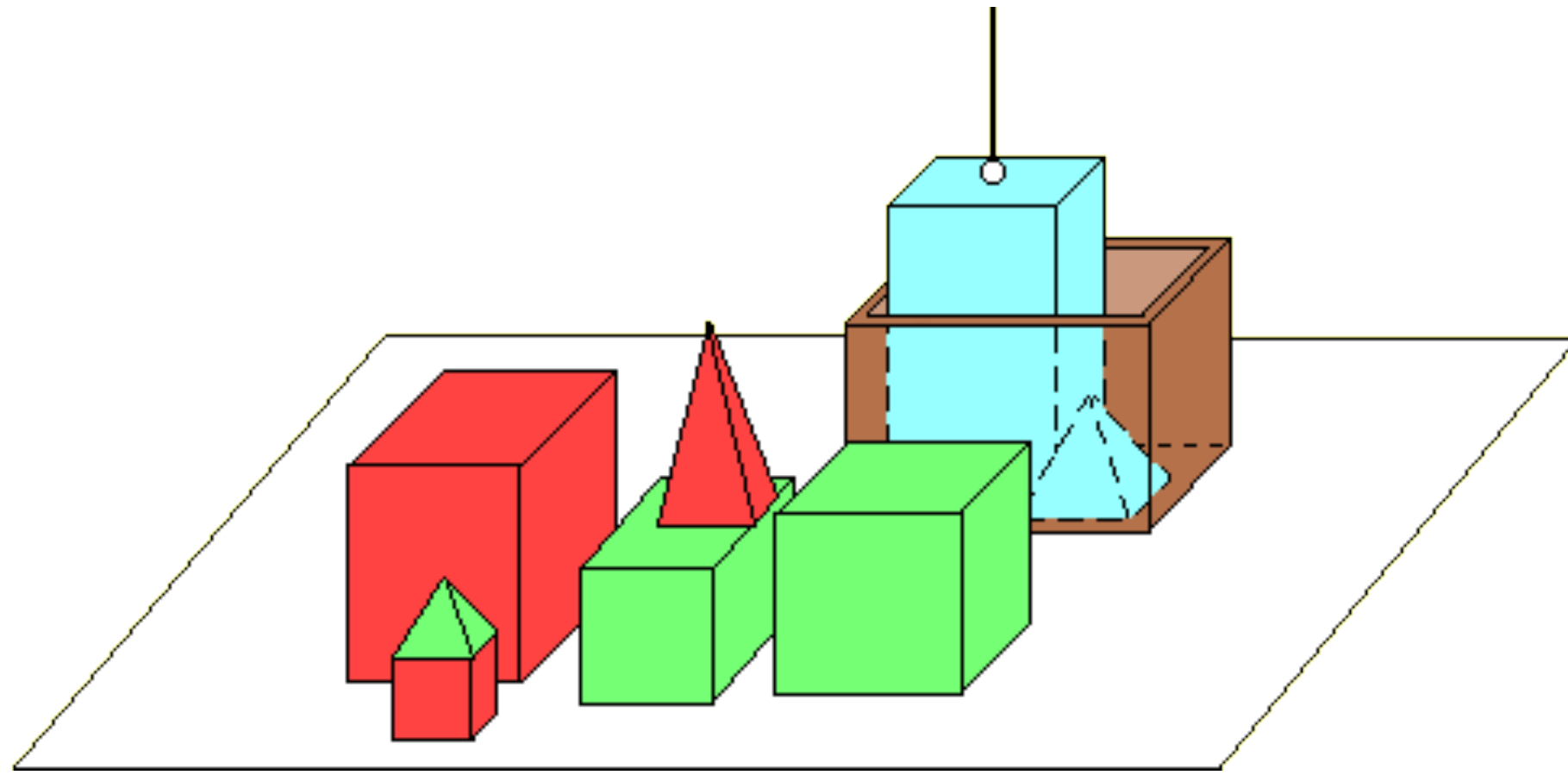
Man shoots elephant wearing his pajamas

aadamee ne apana pajaama pahane hue haathee ko golee maar dee

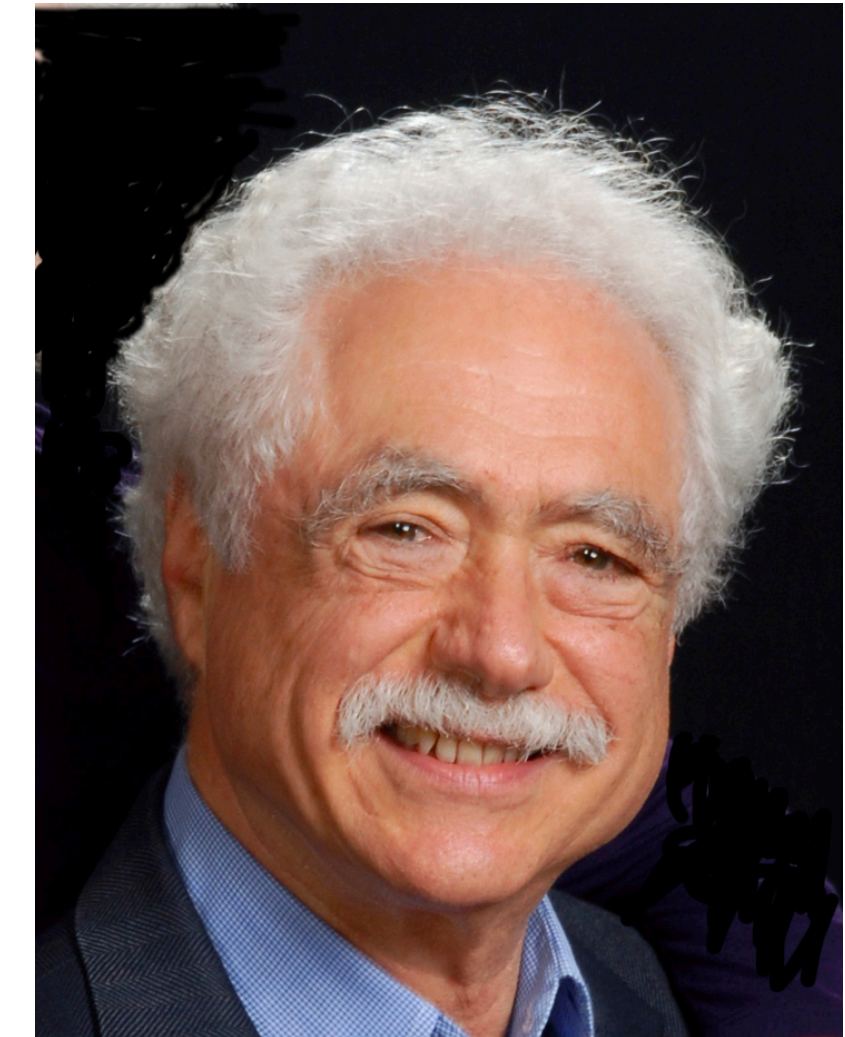
47 / 5,000 अ

47 / 5,000 अ

(Based on Google Translate result in 2023-01)



SHRDLU,
1968

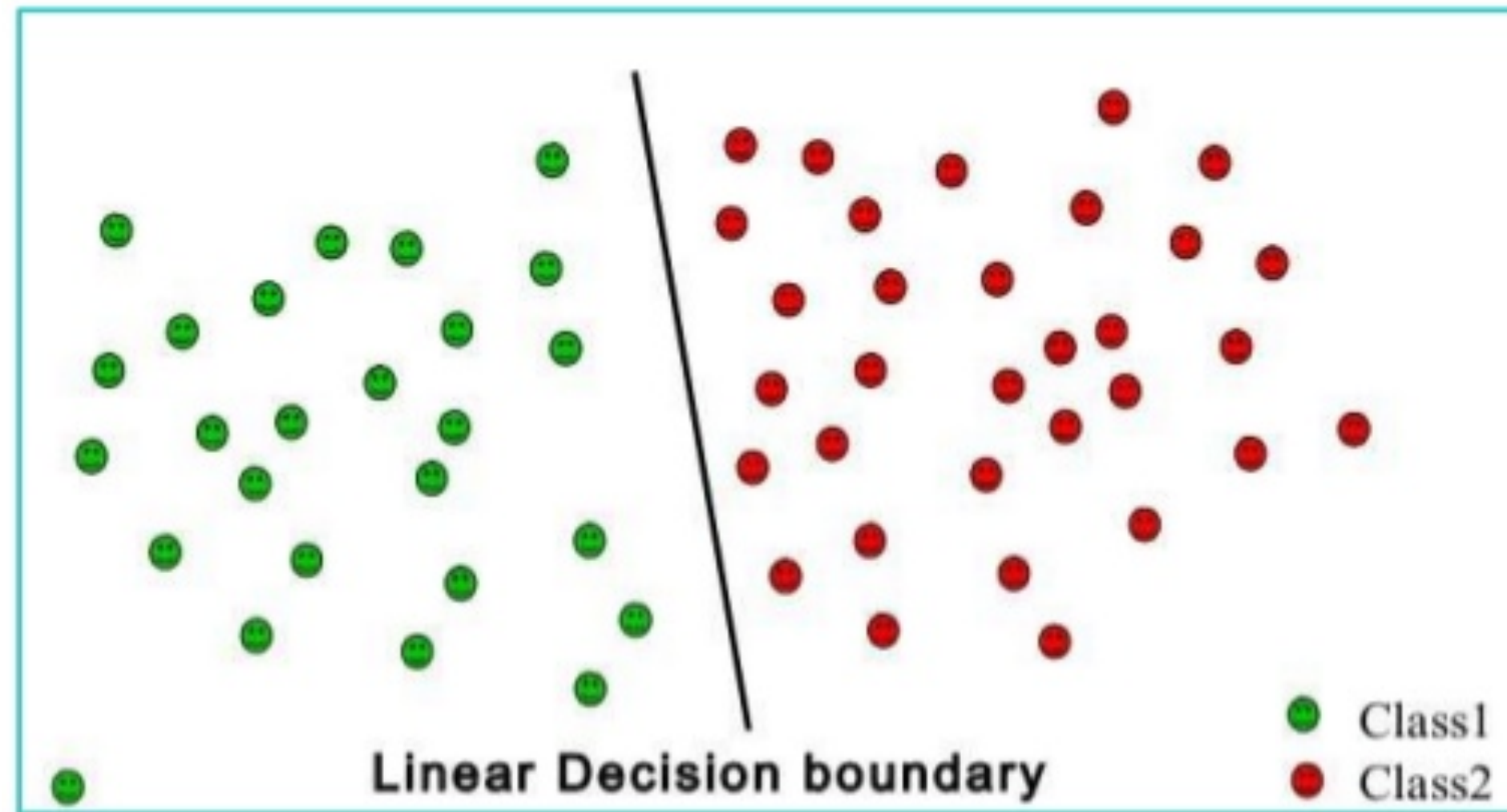


> How many red
blocks are there?
- **THREE OF THEM**

> Pick up the red
block on top of a
green one
OK.

- Rule-based, requiring extensive programming
- Limited domain

Statistical learning

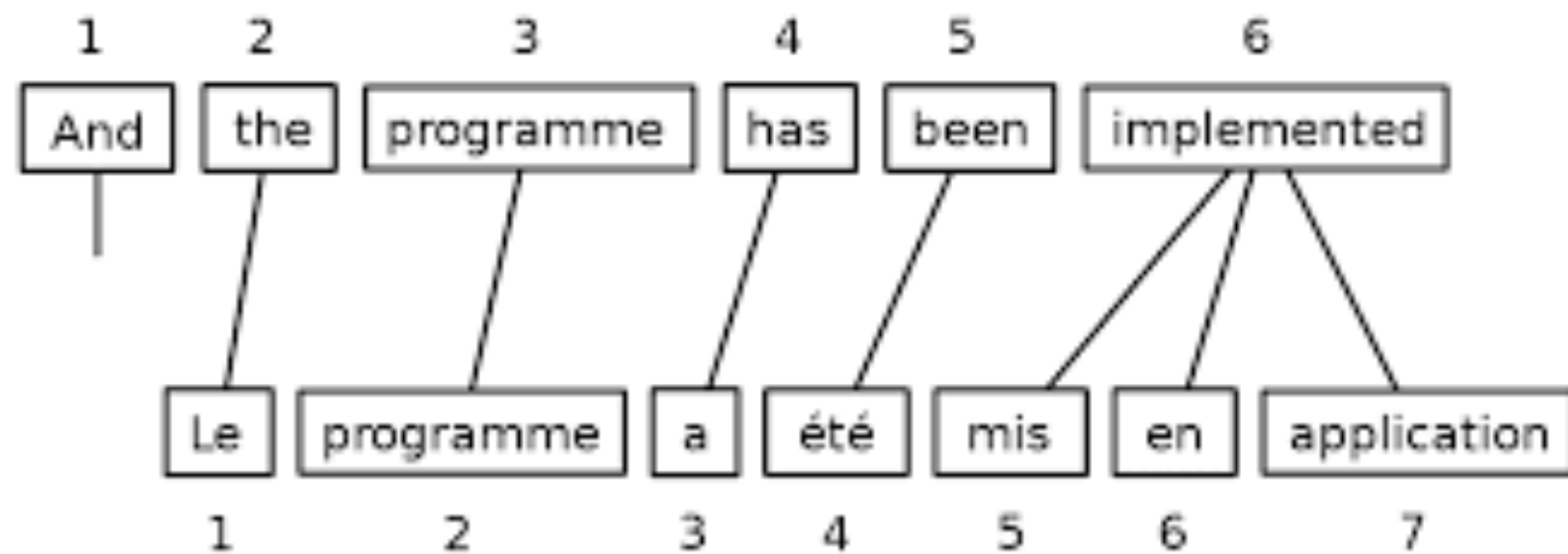


- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

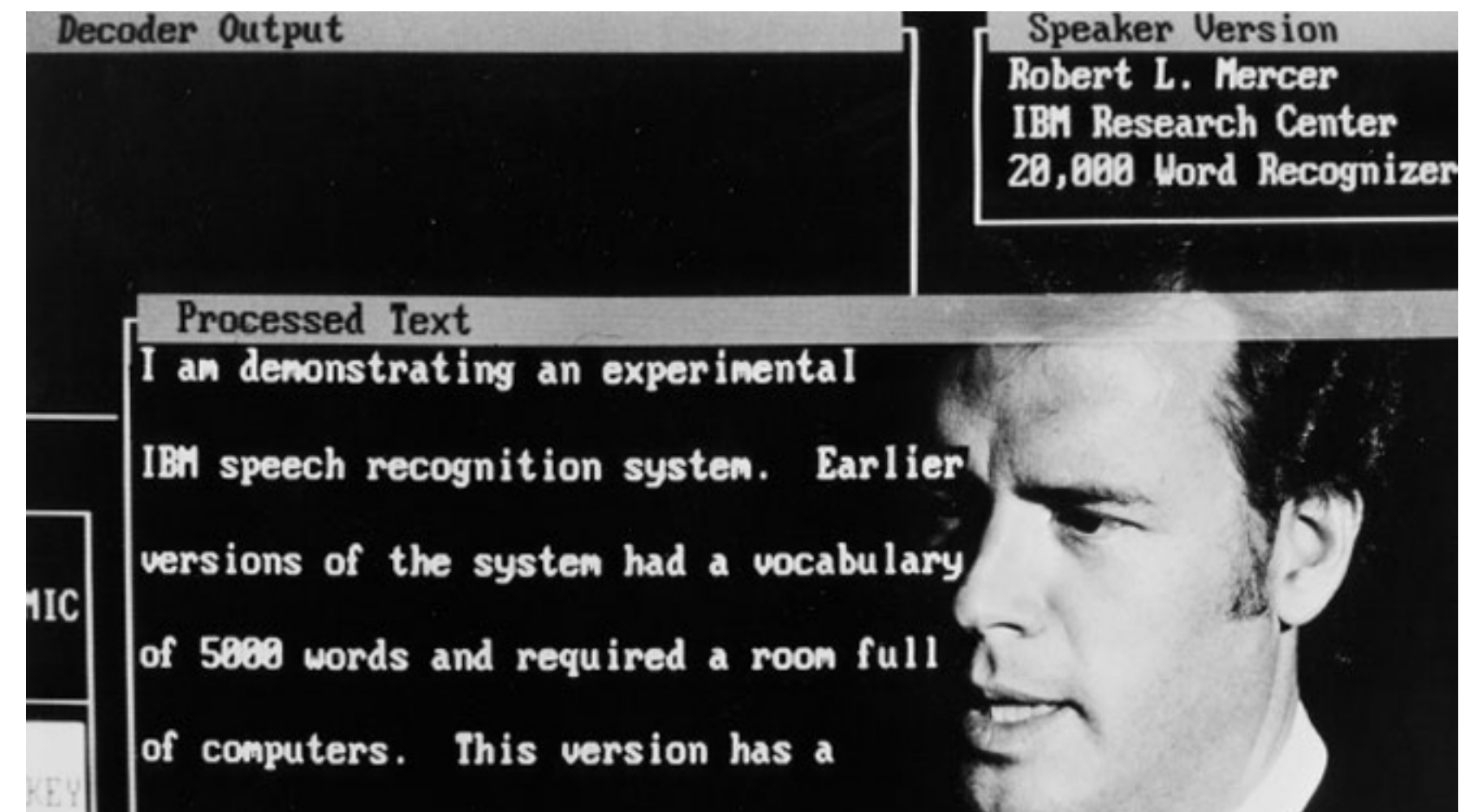
Unsupervised vs. supervised?

Statistical learning

IBM translation models



Speech recognition



Anytime a linguist leaves the group the (speech) recognition rate goes up
- Fred Jelinek 1998

The era of deep learning

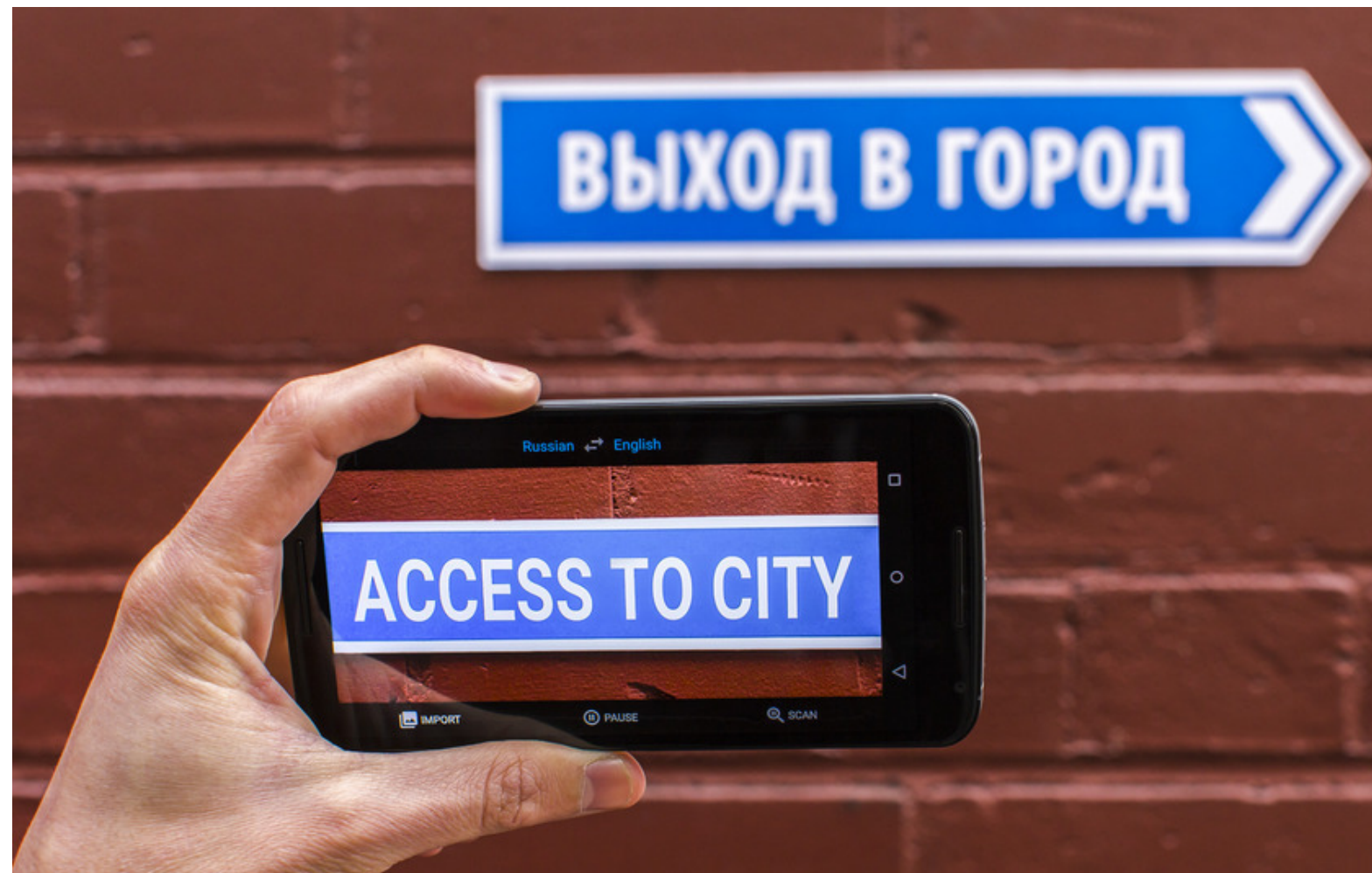
- Significant advances in core NLP technologies
- **Essential ingredient:** large-scale supervision, lots of compute
- Reduced manual effort - less/zero **feature engineering**



GPU



TPU



36M sentence pairs

Russian: Машинный перевод - это круто!

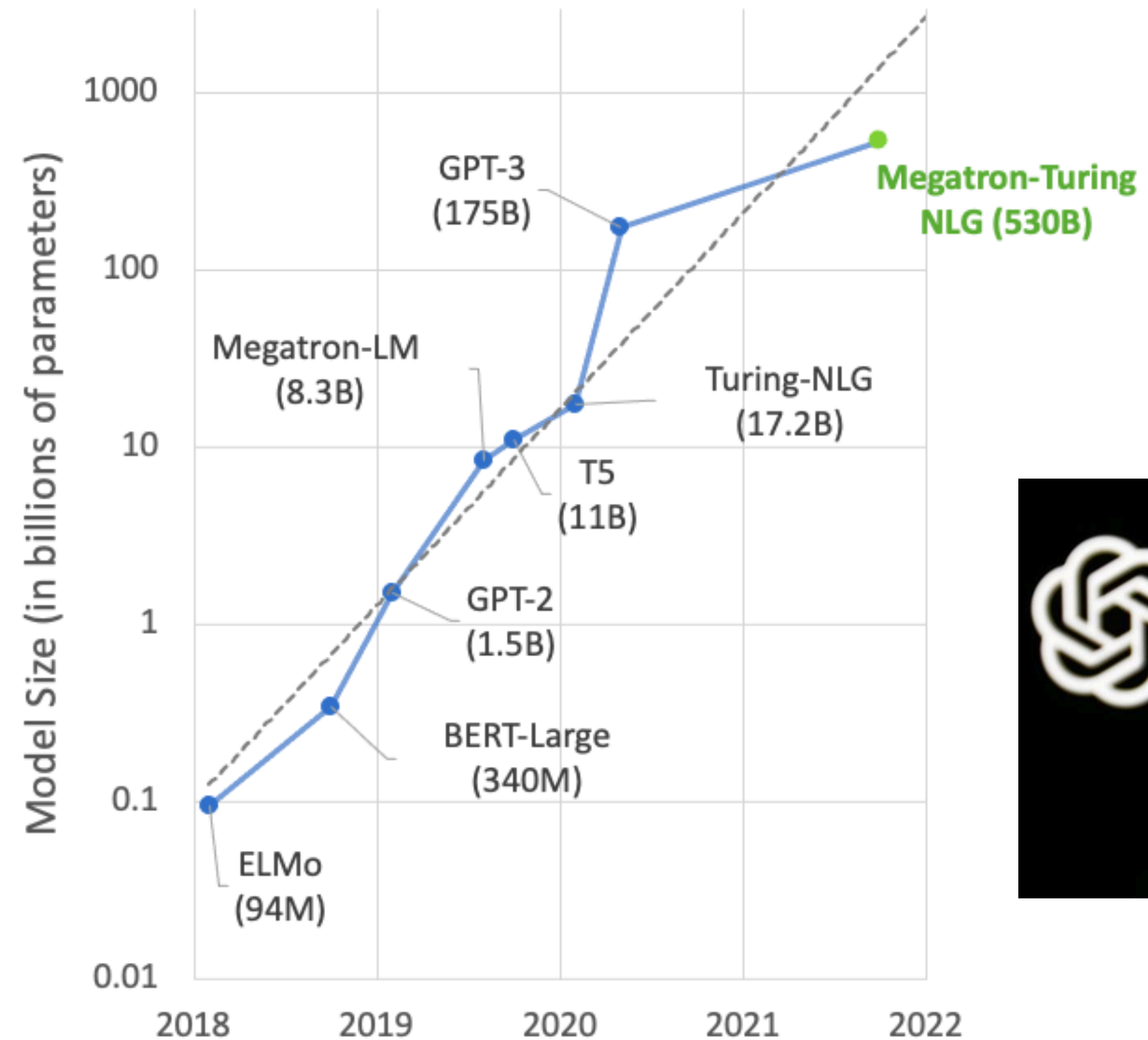


English: Machine translation is cool!

The era of pre-training / LLMs



BERT, ELMo, ERNIE...



- Leverages a lot of unlabeled text
- Model size increased by $10^3 - 10^5$ x in parameters

ChatGPT examples



What NLP topics should I teach in 2023?



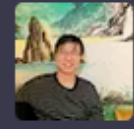
Here are some NLP topics that are popular and relevant in 2023:



1. Transformers (BERT, GPT, etc.)
2. Transfer Learning in NLP
3. Multi-lingual NLP
4. Attention Mechanisms
5. Adversarial Training in NLP
6. NLP for Question Answering
7. Sentiment Analysis
8. Text Generation (summarization, translation, etc.)
9. Named Entity Recognition
10. Dialogue Systems

Note: The field of NLP is constantly evolving, so it's important to stay up-to-date with recent advances and adjust your teaching content accordingly.

ChatGPT examples



what do you know about the Princeton NLP group?



The Princeton NLP group is a research group at Princeton University that focuses on the study of natural language processing (NLP) and its applications. The group is led by Professor Christopher Manning, who is a leading expert in the field of NLP. The group has published numerous research papers and has developed several widely-used NLP tools, including the Stanford Parser and Stanford Named Entity Recognizer. The group also collaborates with other research groups and organizations, such as the Allen Institute for Artificial Intelligence and Google, on NLP research projects.



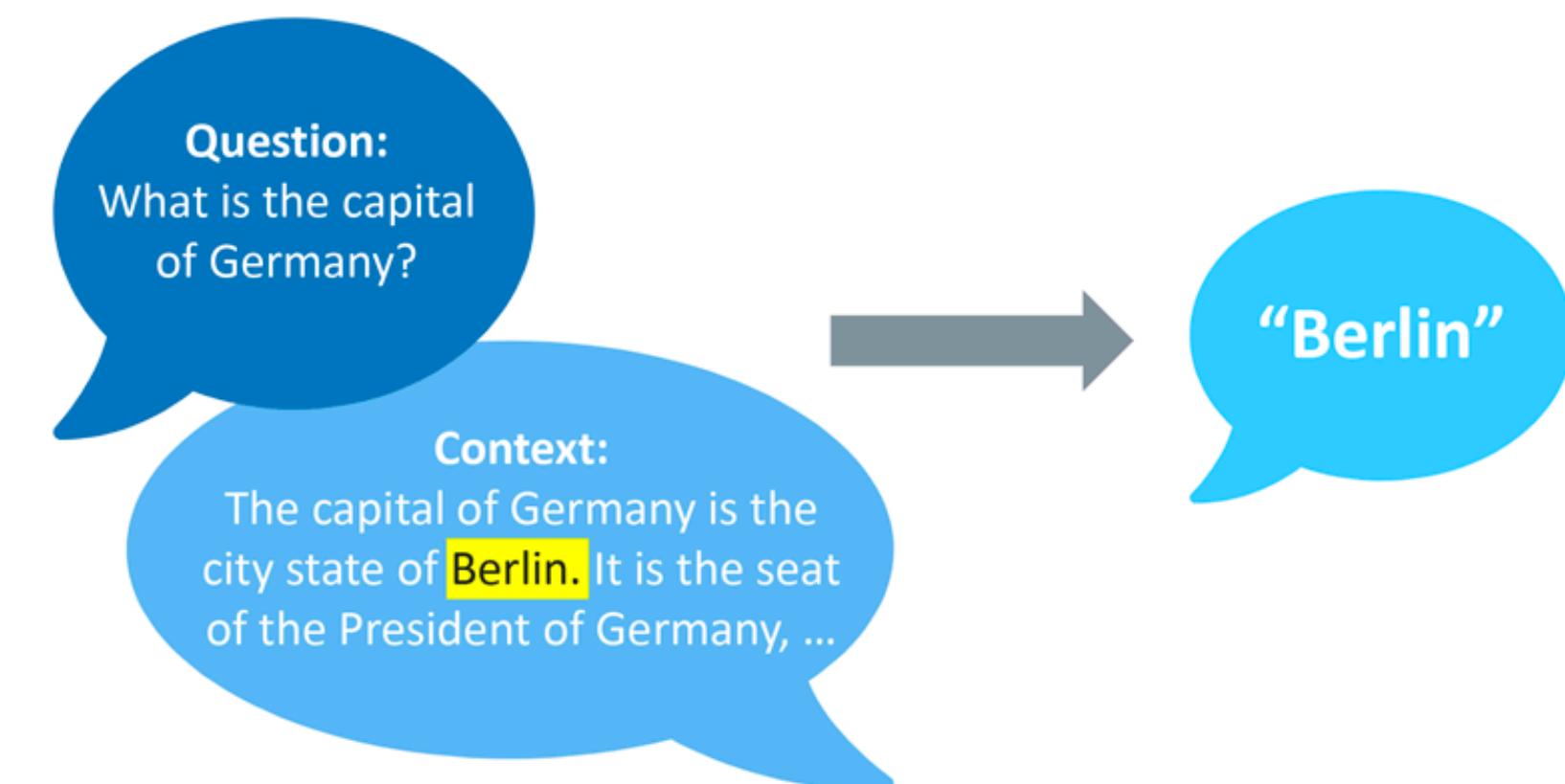
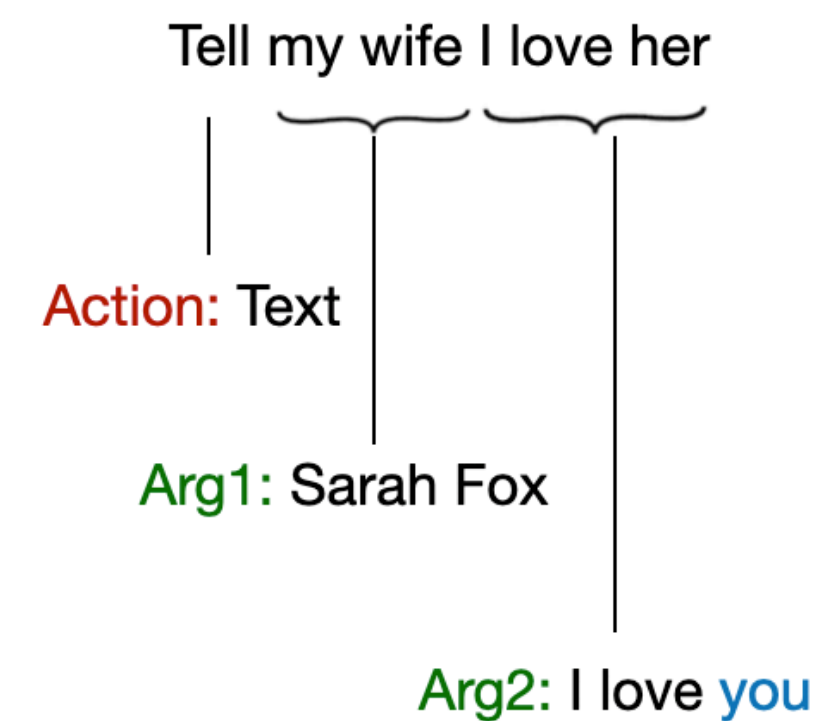
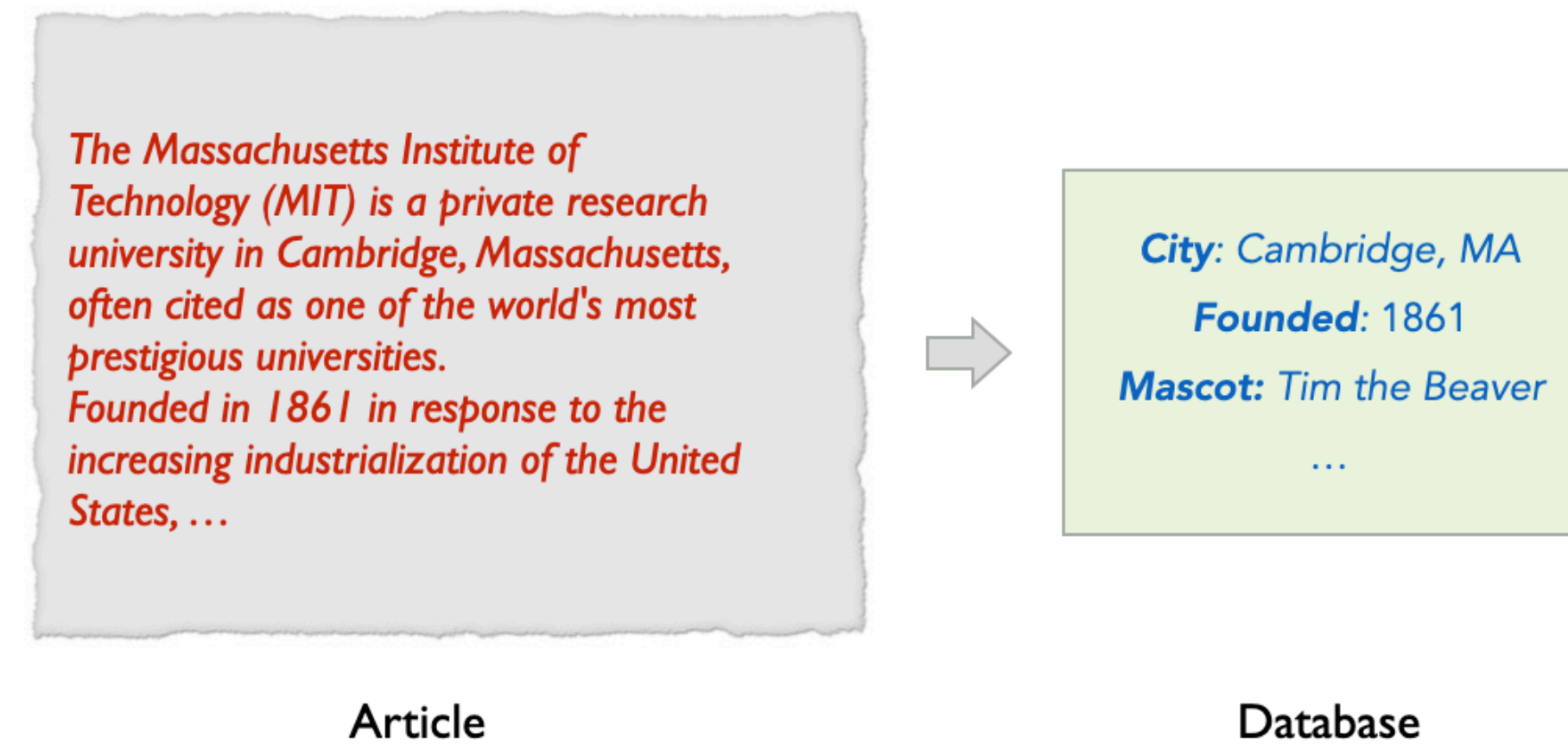
Why do you hallucinate so much?



As an AI language model, I don't have the capacity to hallucinate as I don't possess emotions, sensations, or consciousness. I can only generate responses based on patterns in the text I was trained on.

NLP subfields

- Machine translation
- Question answering
- Information extraction
- Text summarization
- Dialogue
- Semantic parsing
- Syntactic parsing
- Language grounding
-



Topics we will cover in this class

- We will focus on fundamentals, important concepts rather than concrete applications
- The class will be a mix of statistical and neural NLP approaches (>70% neural)
 - N-gram language models
 - Text classification
 - Word embeddings × 2
 - Sequence models × 2
 - Parsing × 2
 - Neural networks for NLP
 - Recurrent neural networks × 2
 - (neural) machine translation × 2
 - **Transformers** × 2
 - BERT, ELMo, ...
 - **Large language models** + more guest lectures!
 - Question answering
 - Language grounding

Before midterm

After midterm

Why is language difficult to understand?

Why is language difficult to understand?

- Ambiguous
- Dialects
- Accents
- listener has to infer - pragmatics
- humor, sarcasm, irony
- context, dependencies

Lexical ambiguity

The fisherman went to the *bank*.

bank¹

/bɑNGk/ 

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.

"willows lined the bank"

synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; [More](#)

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.

"I paid the money straight into my bank"

synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

One word can mean several different things

Lexical ambiguity

The fisherman went to the *bank*. He deposited some money.

bank¹

/bɑŋk/ 

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.

"willows lined the bank"

synonyms: [edge](#), [side](#), [shore](#), [coast](#), [embankment](#), [bankside](#), [levee](#), [border](#), [verge](#), [boundary](#),
[margin](#), [rim](#), [fringe](#); [More](#)

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.

"I paid the money straight into my bank"

synonyms: [financial institution](#), [merchant bank](#), [savings bank](#), [finance company](#), [trust company](#),

Word sense disambiguation

Lexical variations



**ACCORDING TO THE THESAURUS,
"THEY'RE HUMID, PREPOSSESSING
HOMOSAPIENS WITH FULL SIZED AORTIC
PUMPS" MEANS "THEY'RE WARM, NICE
PEOPLE WITH BIG HEARTS."**

Several words can mean the same thing!

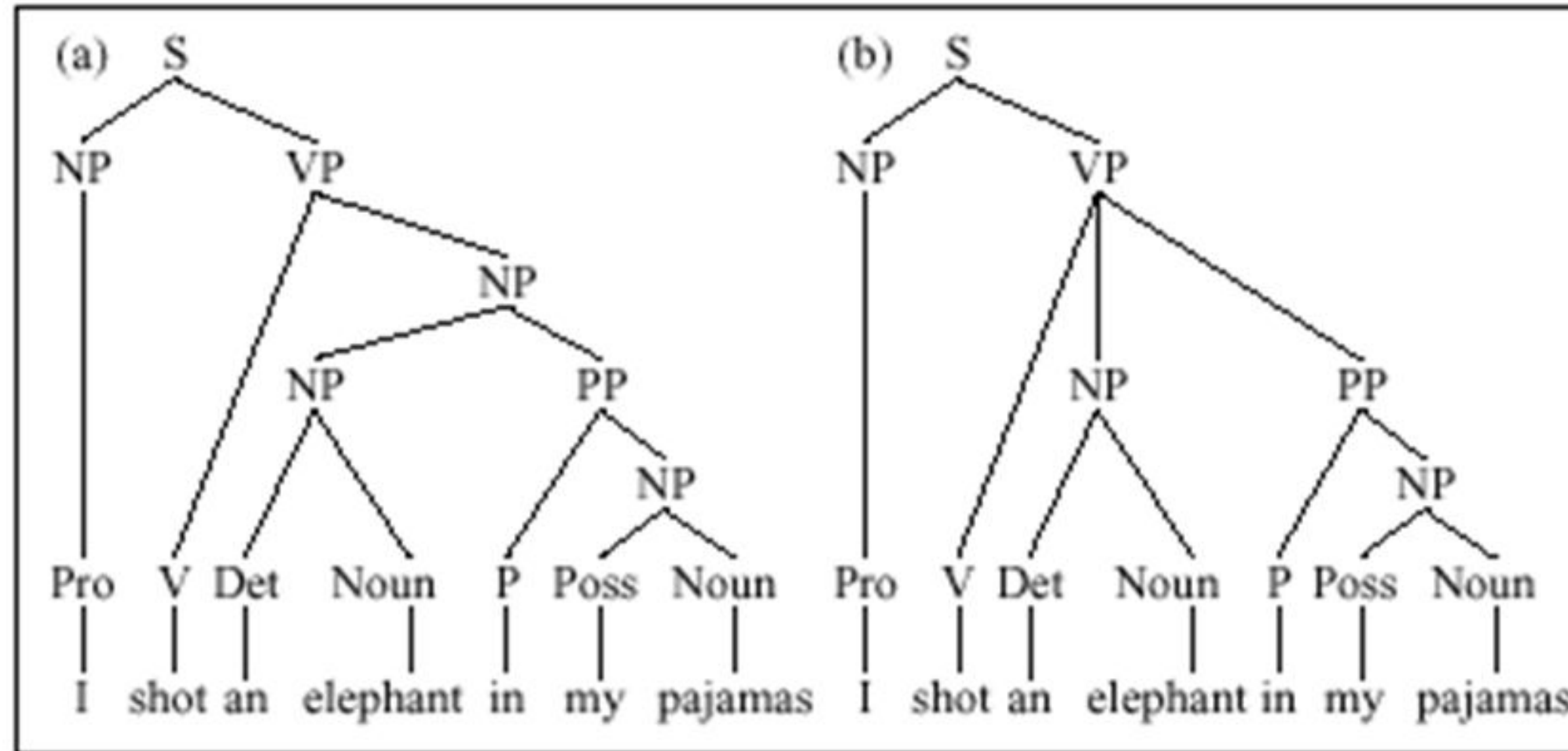
Comprehending word sequences

- My brother went to the park near my sister's house
 - Park my went house near to sister's my brother the
 - "My brother went park near sister's house"?
 - The old man the boat
 - Fat people eat accumulates
- Implicit structure in all languages
 - Coarse-to-fine levels (recursive)
 - What are some good data structures to represent this?

Garden Path sentence

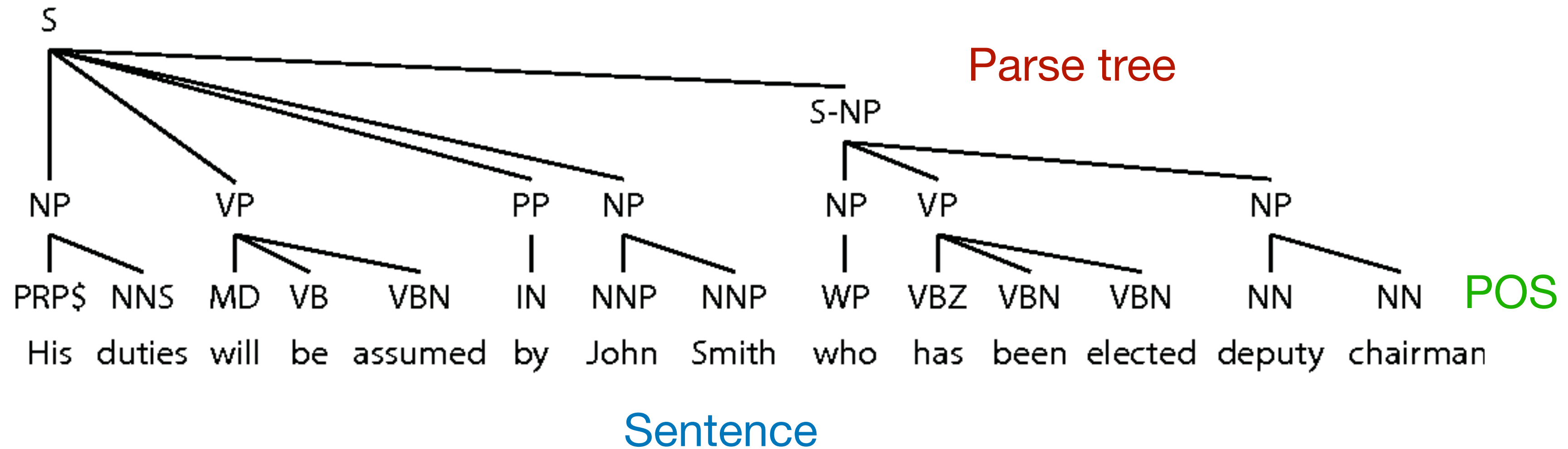
Syntactic ambiguity

I shot an elephant in my pajamas



Human language is full of such examples!

Syntactic parsing



Online tools: <http://nlp.stanford.edu:8080/corenlp/>

Discourse ambiguity

- The man couldn't lift his son because **he** was so **heavy**.
- The man couldn't lift his son because **he** was so **weak**.

What does "he" refer to?

- The city councilmen refused the demonstrators a permit because **they** **feared** violence.
- The city councilmen refused the demonstrators a permit because **they** **advocated** violence.

What does "they" refer to?

Anaphora resolution

Next lecture: n-gram language models