Natural language Processing

Lecture 1: Introduction to NLP

Kenneth Enevoldsen | 2024

CENTER FOR HUMANITIES COMPUTING

Agenda

- Course Overview
  - What should you expect from the course?
- Overview of natural language processing
- Central questions of the course

Course Overview

- Latest version of syllabus can be found on GitHub
- We will use an adaptive approach

Lectures and Classes

- Lectures
  - Introduction to methods and concepts
  - Lecturing and discussions
- Classes
  - Applied: Coding practice, exercises, assignments
  - Some classes on final project
  - Discussions
- Course material will be avialable on GitHub

Academic Goals

- **Knowledge**
  - **Constrast NLP methods** in terms of their strengths and weaknesses
  - Explain how NLP analysis can provide insights to **human cognition**
  - Discuss related philosopical and **ethical issues**
- **Skills**
  - Identify relevant **data sources** for a use-case
  - **Choose and apply tool** for language processing
- **Competences**
  - **Justify** use of techniques for specific research questions
  - **Critically evaluate** appropriateness of a given NLP method

See the latest version of *"Studieordning"* for the exact wording

Exam

- **Written assignment**
  - Can be written in groups
  - Must display understanding of the field and its technical competencies

See the latest version of *"Studieordning"* for the exact wording

Teaching Philosophy

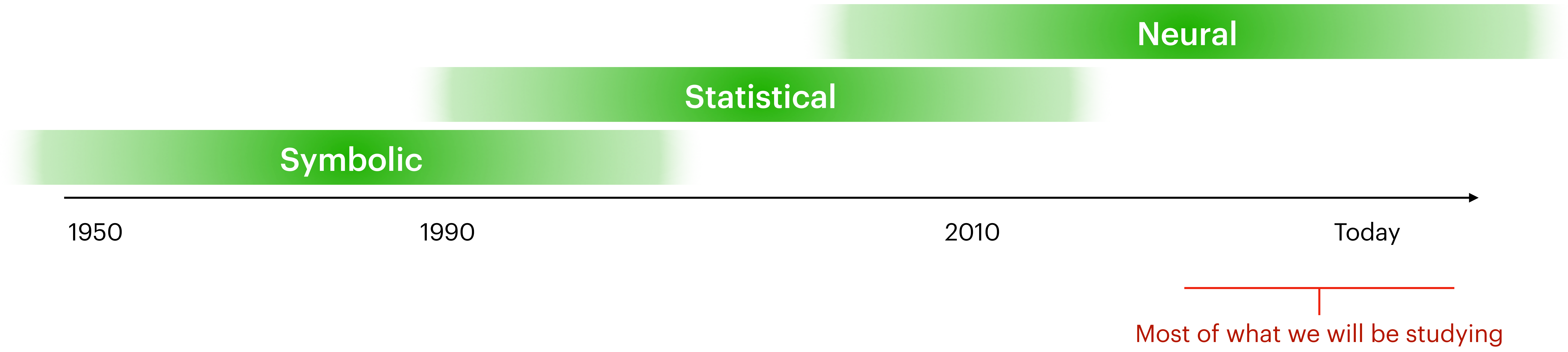
- **Focus on re-useable concepts**
  representation learning
  semantic vector spaces

- **Setting it in relation og Cognitive Science**
  What does it mean to understand language?
  How do we examine the representations of a model?

- **Machine learning Approach**
  Machine learning and generalized linear models
  Neural networks

Questions?

Simplified Timeline



Sources & Notes

Note that dates correctly spaced out, this it to allow more space for later periods
In reality symbolic and statistical systems are still used today and neural systems on text span back decades

Expert Systems

Sources & Notes

Note that dates correctly spaced out, this it to allow more space for later periods
In reality symbolic and statistical systems are still used today and neural systems on text span back decades

Pre-NLP: Weaver to Wiener, 1947

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, 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."

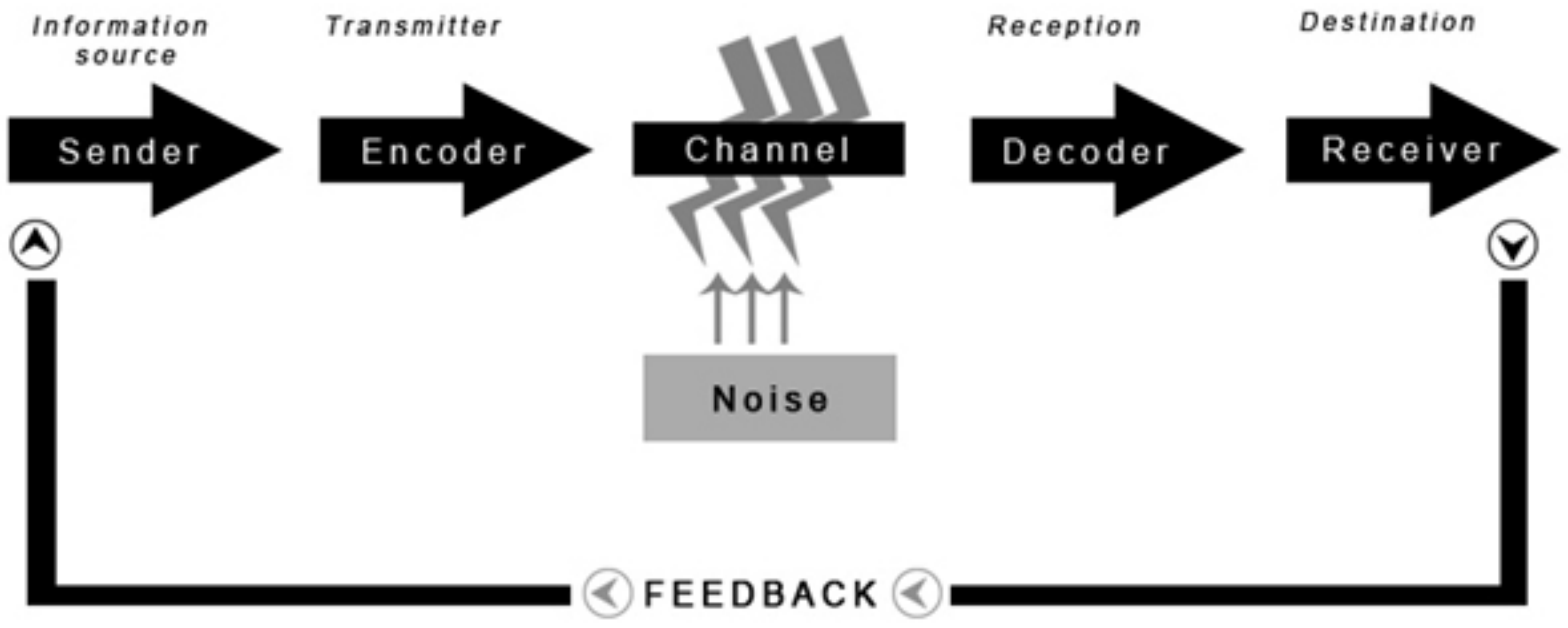
Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

Sources
& Notes

Warren Weaver to Norbert Wiener, 1947

Weaver's Memorandum (1949)

- First example of **computation** over linguistic input
- Heavily inspired by **cryptography**
- All languages are the *same* language
    - or put more mildly, there are latent **logical** language invariants
    - What are these invariants?
- If these are found, **translation** can be automated
- Georgetown experiment (1955):
    - hand-coded rules and dictionary mappings —> fails to scale for lack of a general theory of language
    - Authors: within 3-5 year machine translation will be solved - **Question: Is it solved today?**

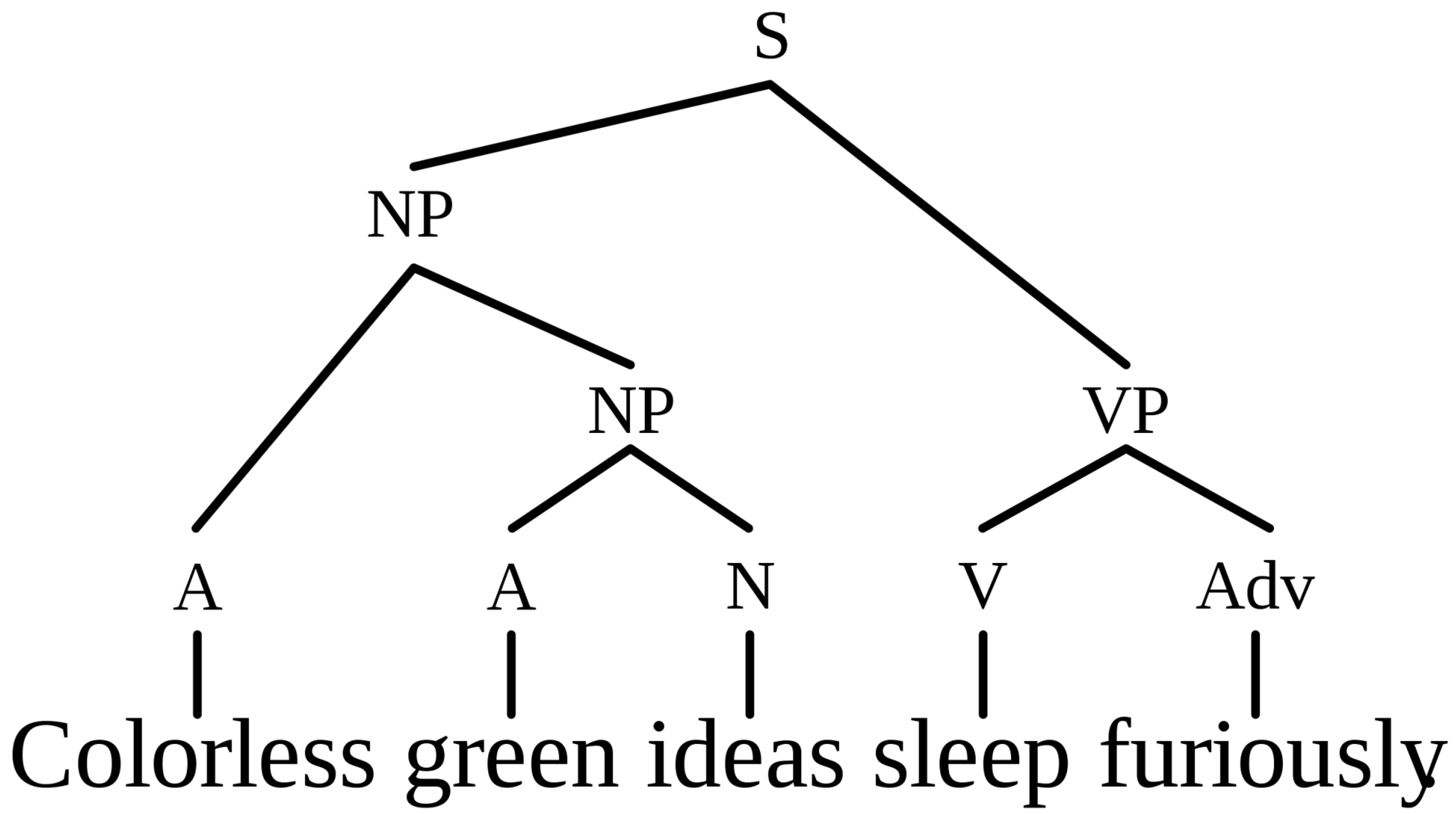


SHANNON-WEAVER'S MODEL OF COMMUNICATION

Georgetown Experiment: https://en.wikipedia.org/wiki/Georgetown–IBM_experiment
Is it solved today: https://en.wikipedia.org/wiki/History_of_machine_translation#2010s

Chomsky's Syntactic Structures (1957)

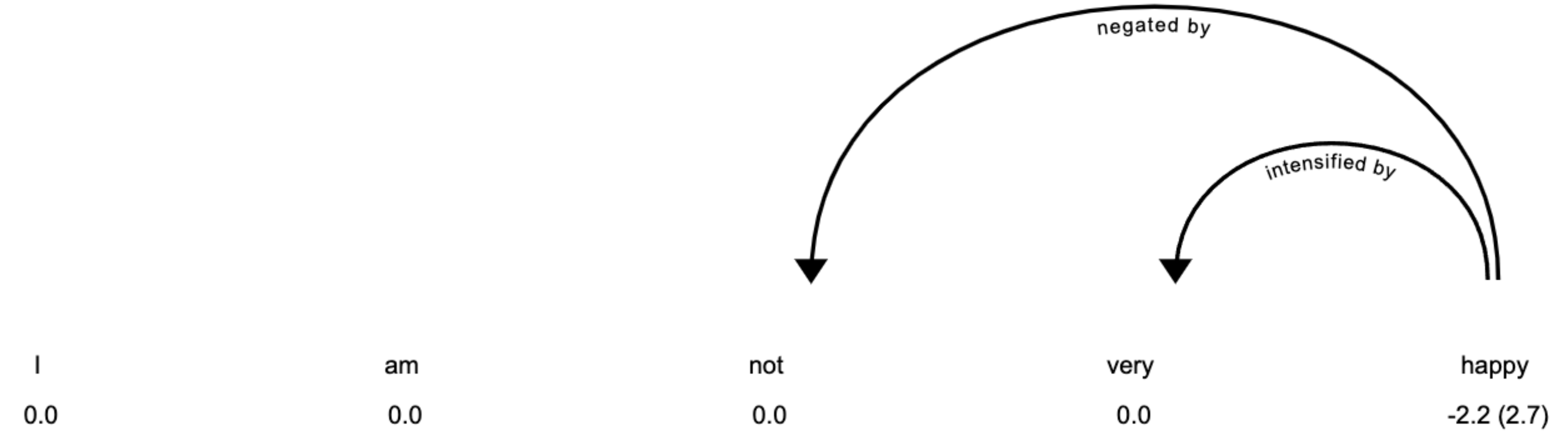
- Grammar is essentially a **system of rules**
  - the "latent" universal structure of language
- These rules generate exactly all possible combinations of *any* grammatical sentences in any given language
- The task of linguistics is to **uncover** and **formalise** this system of rules for any given language (and for Language generally)
- Syntax is **independent** of **semantics** (not all correct sentences occur, and correct sentences need not make sense)



Effect

- Chomsky’s grammatical systems were **too hard** and **computationally expensive** to implement, but alternative, more implementable approaches emerged, e.g., Fillmore’s case grammar
- Still, little practical success and decrease in funding (first AI Winter, 1974-1980)
- 1970s: **Conceptual ontologies** (e.g., MARGIE), storing real-world knowledge into computer-readable representations
- 1980s: **Rule-based systems** (or “symbolic systems”): text is split into (aka, tokenized) into meaningless units, and general hard-coded rules for how they combine to form meaning are developed.
- Second AI Winter (late 1980s)

- Today
  - Simple rule-based system are used
  - Generally easy to break



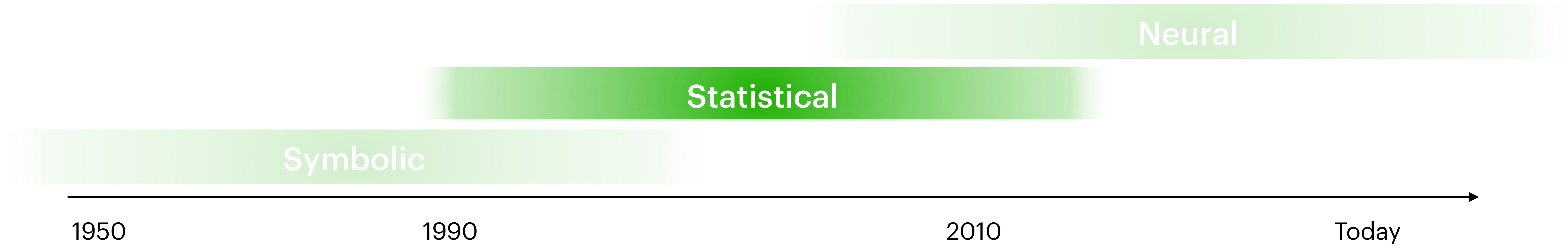
Sources & Notes

Figure from software Asent: https://github.com/KennethEnevoldsen/asent

Which uses a heavy dictionary approach. You can try it out to see if you can break it.

CENTER FOR HUMANITIES COMPUTING

Expert features, learned model



Sources & Notes

Note that dates correctly spaced out, this it to allow more space for later periods
In reality symbolic and statistical systems are still used today and neural systems on text span back decades

Statistical Approaches

- Increased focus on new tasks, e.g., **information retrieval**
- Representations of text are **inferred** from its **statistical properties**
  conceptual foundation of modern NLP
- **Predictive algorithms** and **neural networks** open for new applications…

the new question:

__how can text be represented__ to support applications like information retrieval, or predictive algorithms / neural networks?

Statistical Approaches

- Increased focus on new tasks, e.g., **information retrieval**
- Representations of text are **inferred** from its **statistical properties**
  conceptual foundation of modern NLP
- **Predictive algorithms** and **neural networks** open for new applications…

the new question:

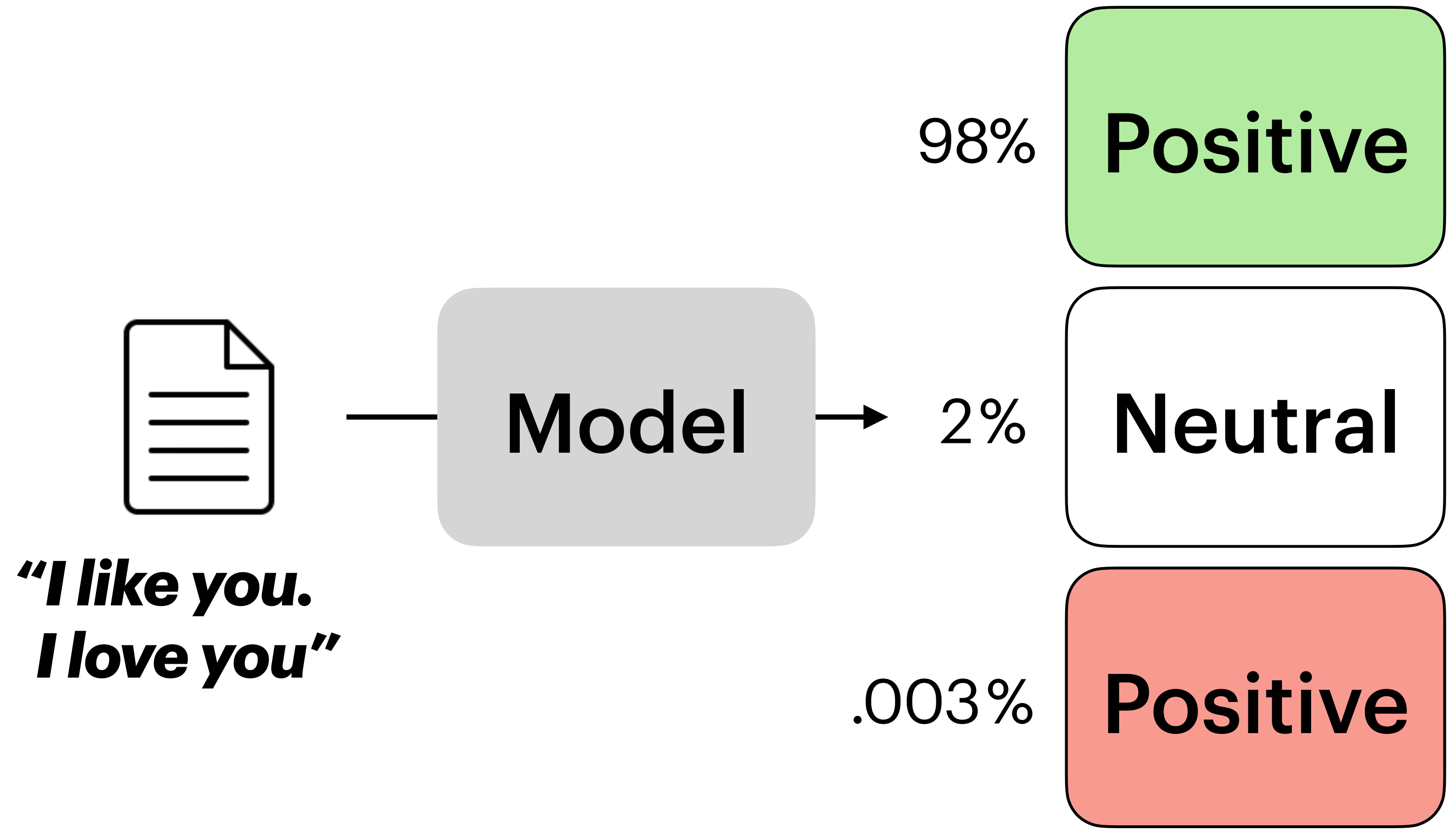
> ***how can text be represented*** *to support applications like information retrieval, or predictive algorithms / neural networks?*

↑

Core topic of the course

Examples of Predictive Tasks

- **Sentiment Classification**
- Named Entity Recognition (NER)
- Text Summarization
- Machine Translation
- Question-answering



Question: How could we do this?

Demo on an open source model can be found here: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Word as features

"I like you. I love you"

Featurize

Document level features			Word level features			
Words in pos. dict.	Upper-case	Length	"like"	"you"	"hate"	...
2	0	6	1	2	0	...

Logistic Regression

98% Positive

2% Neutral

.003% Positive

Demo on an open source model can be found here: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Word as features

Goal: design good featues for the model

"I like you. I love you"

Featurize →

Document level features			Word level features			
Words in pos. dict.	Upper-case	Length	"like"	"you"	"hate"	...
2	0	6	1	2	0	...

→ Logistic Regression →

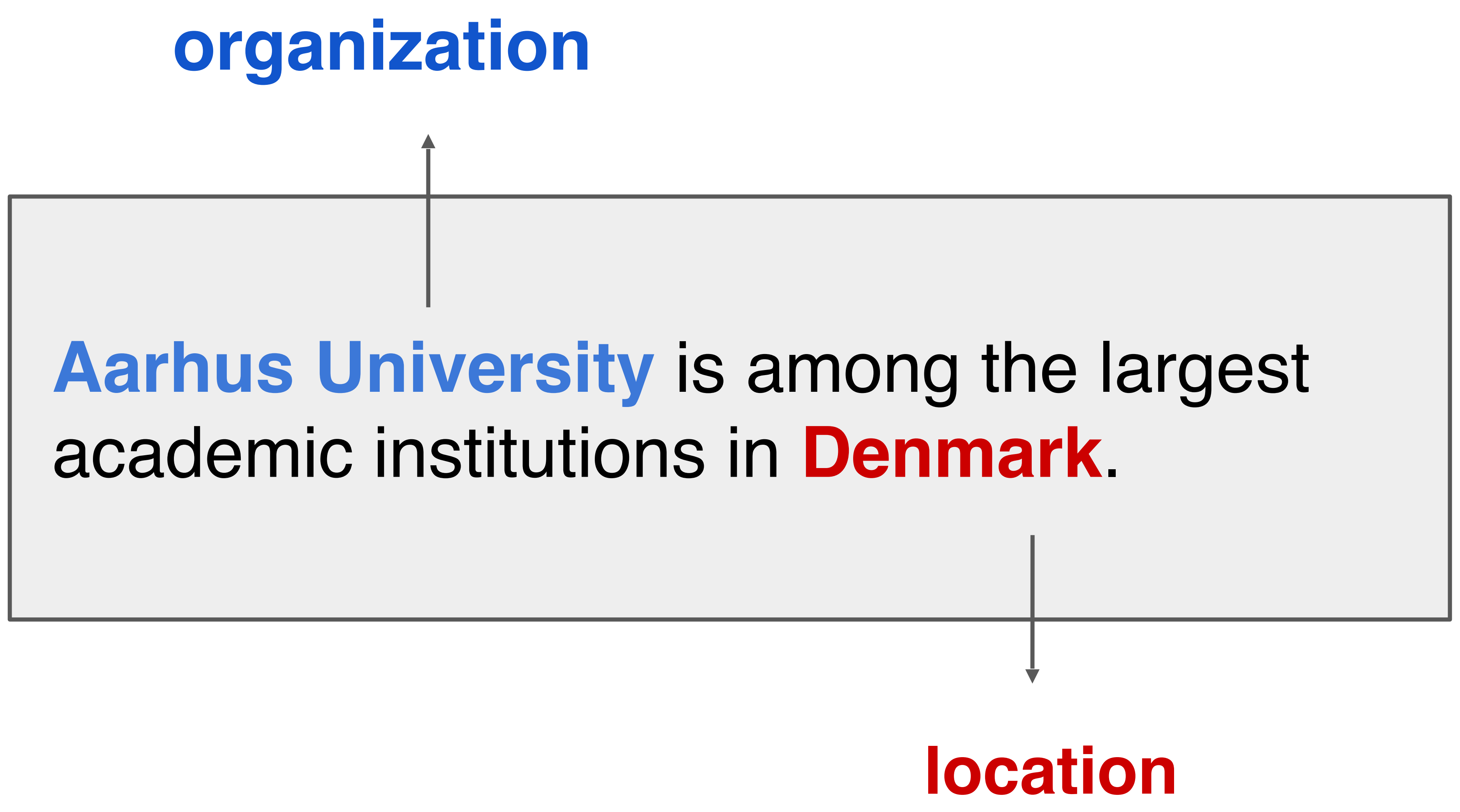
98% Positive

2% Neutral

.003% Positive

Examples of Predictive Tasks

- Sentiment Classification
- **Named Entity Recognition (NER)**
- Text Summarization
- Machine Translation
- Question-answering

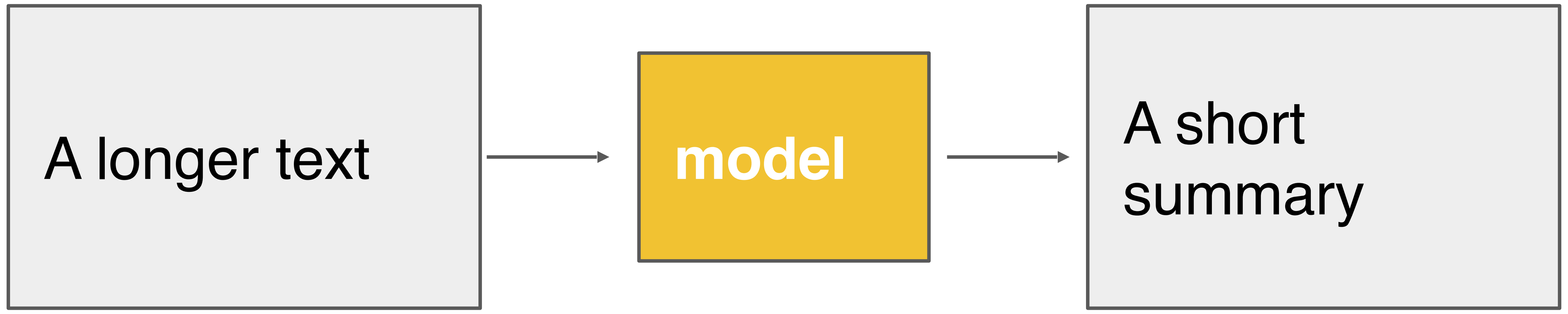


Sources & Notes

Demo of current flexible NER, Question answering and Summarization systems: https://huggingface.co/spaces/knowledgator/GLiNER_HandyLab

Examples of Predictive Tasks

- Sentiment Classification
- Named Entity Recognition (NER)
- **Text Summarization**
- Machine Translation
- Question-answering

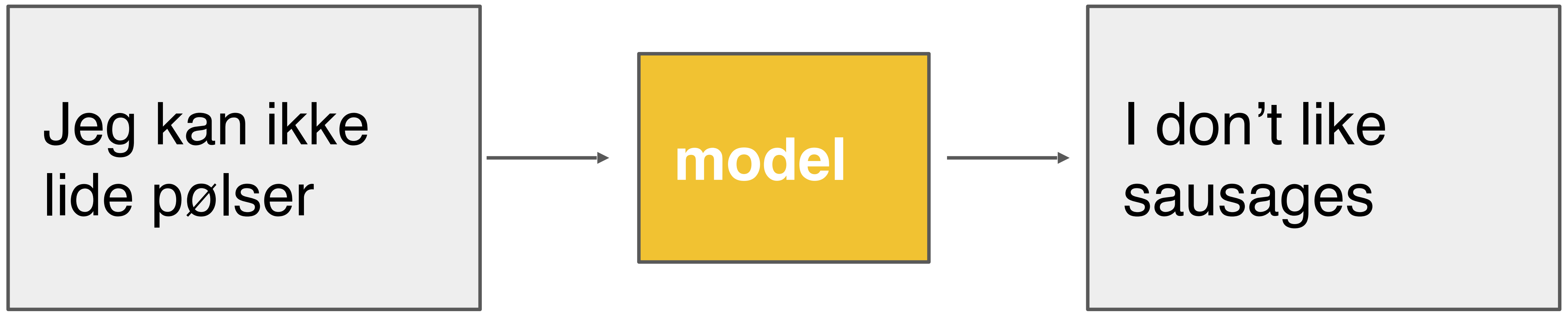


Sources & Notes

Demo of current flexible NER, Question answering and Summarization systems: https://huggingface.co/spaces/knowledgator/GLiNER_HandyLab

Examples of Predictive Tasks

- Sentiment Classification
- Named Entity Recognition (NER)
- Text Summarization
- **Machine Translation**
- Question-answering



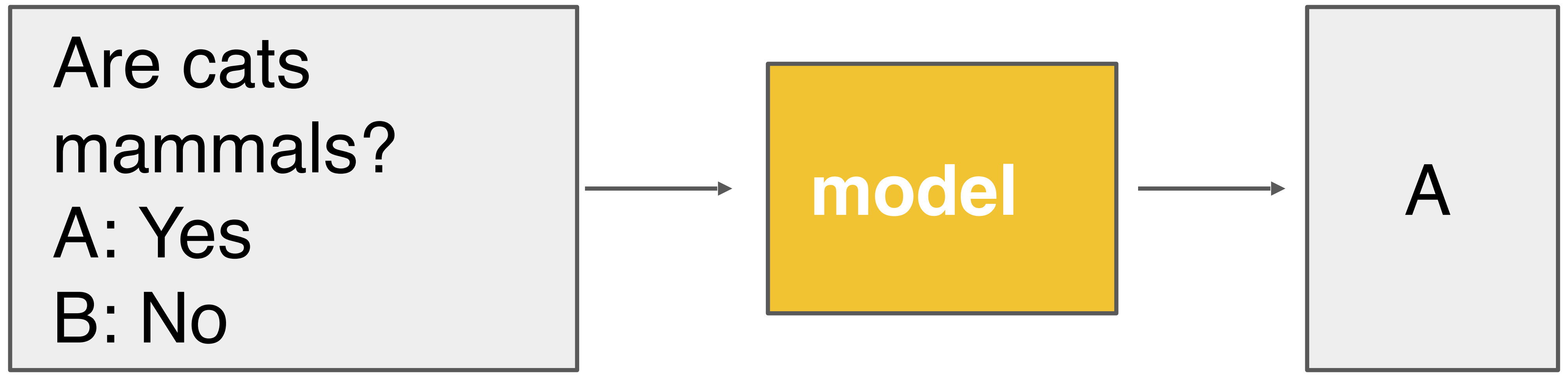
Sources & Notes

Demo of current flexible NER, Question answering and Summarization systems: https://huggingface.co/spaces/knowledgator/GLiNER_HandyLab

CENTER FOR HUMANITIES COMPUTING

Examples of Predictive Tasks

- Sentiment Classification
- Named Entity Recognition (NER)
- Text Summarization
- Machine Translation
- **Question-answering**



Question: Or is this classification?

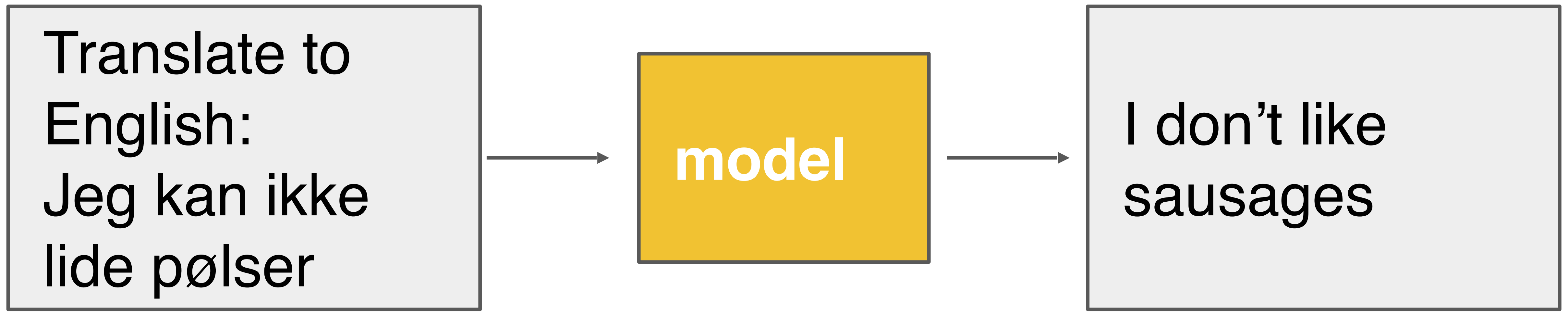
Sources & Notes

Demo of current flexible NER, Question answering and Summarization systems: https://huggingface.co/spaces/knowledgator/GLiNER_HandyLab

CENTER FOR HUMANITIES COMPUTING

Examples of Predictive Tasks

- Sentiment Classification
- Named Entity Recognition (NER)
- Text Summarization
- Machine Translation
- **Question-answering**



Question: or machine tranlation?

Sources & Notes

Demo of current flexible NER, Question answering and Summarization systems: https://huggingface.co/spaces/knowledgator/GLiNER_HandyLab

CENTER FOR HUMANITIES COMPUTING

Learned features and general purpose models

Note that dates correctly spaced out, this it to allow more space for later periods
In reality symbolic and statistical systems are still used today and neural systems on text span back decades

Where are we at now?

A: "socrates is a man", B: "all men are mortal". do these two sentences imply C: "socrates is mortal"? if so, why? if not, why?

Yes, the two sentences "All men are mortal" and "Socrates is a man" do imply that Socrates is mortal. This is an example of a logical argument, specifically a syllogism, which consists of three parts:

1. All men are mortal (universal affirmative)
2. Socrates is a man (particular affirmative)
3. Therefore, Socrates is mortal (conclusion)

The conclusion follows logically from the premises because if all men are mortal and Socrates is a man, then it is necessarily true that Socrates is mortal. In other words, there is no way for Socrates to be a man without also being mortal, given the universal statement about all men.

This type of argument is called a modus ponens, which means "mode of putting forward." It consists of a major premise (all men are mortal), a minor premise (Socrates is a man), and a conclusion (Socrates is mortal). The conclusion logically follows from the premises because the major premise establishes a general rule or principle, and the minor premise provides an instance or application of that rule.

Modern systems learn to perform **many of these tasks at once** through a combination of training on general language tasks and task-specific training.

Could e.g. be done using ChatGPT 4o

Where are we at now?

ok, let's say we have A: "socrates is mortal", and B: "all men are mortal". do A and B imply C: "socrates is a man"?

No, A and B do not imply C: "Socrates is a man."

The reasoning would go like this:

A: Socrates is mortal
B: All men are mortal
C: Socrates is a man

The conclusion C does not logically follow from the premises A and B. Just because Socrates is mortal and all men are mortal, it does not mean that Socrates is a man. Socrates could be any kind of mortal creature, such as a bird or a dog, and still be mortal.

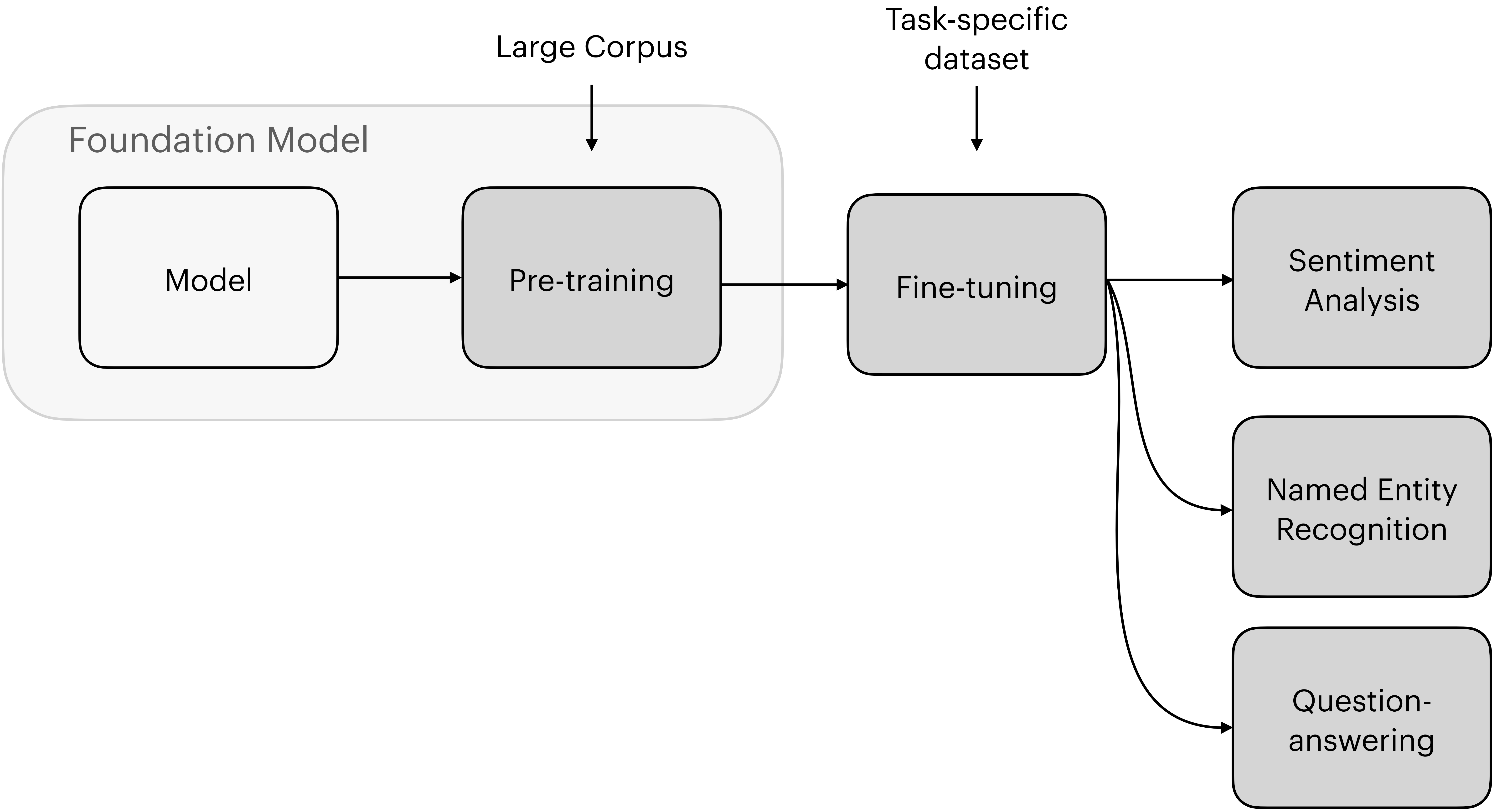
In order to conclude that Socrates is a man, we would need additional information or a separate premise that states that Socrates is human. For example, if we had a third premise, "Socrates is human," then we could infer that Socrates is a man, since humans are a subset of mortal creatures. But without this additional information, we cannot logically deduce that Socrates is a man based on the premises A and B alone.

Where are we at now?

- **flexibly performs multiple tasks**
- and can generalize to new tasks (few- or zero-shot generalization*)
- extremely **fluent** language generation
- does this model have any “deep” understanding of the linguistic task it is performing? does it have any command of “factual” information, any notion of truth?

Whether is it completely unseen is actively discussed as it likely represented within the training data

What lies behind?



Note that the notion of foundation model is controversial, they are also referred to as Large Language models (LLMs), or Base/Core models.

Questions for this semester

- What does it take to build systems that can do things like these?
- What are these systems currently lacking
- Do they approach language like humans do?
- Which tasks do they find really, really hard?
- What unresolved ethical / societal issues have they raised?

Questions for this semester

- **What does it take to build systems that can do things like these?**
- What are these systems currently lacking
- Do they approach language like humans do?
- Which tasks do they find really, really hard?
- What unresolved ethical / societal issues have they raised?

Questions for this semester

- **What does it take to build systems that can do things like these?**
  - How do we turn **language** into **numerical representation** which models can work on?
  - How do we represent the **meaning of word**?
  - How do we represent **sequences of words**?
  - …

The simplest approach

- Build a **vocabulary** of words
- Represent each word as a number (index)
- Each word get a unique **vector**, where 1 corresponds to its index (**one-hot** encoding):

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

- **Question: What are the downsides?**

Problems with the discrete approach

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

these representations are entirely independent: there is *no notion of similarity!*

how can we develop representations that **encode similarity** *intrinsically*, e.g., similarity between words is encoded in mathematical properties of the representations?

Spoiler: Distrubitional Hypothesis

"You shall know a word by the company it keeps"

J.R. Firth, 1957
Studies in Linguistic Analysis

Spoiler: Distrubitional Hypothesis

"… know a word by the
…any it keeps"

J.R. Firth, 1957
Studies in Linguistic Analysis

more next week! ;)

Early approaches for language generation

- Shannon (1948)
  - $P(\text{next word} \mid \text{previous words}) = P(w_i \mid w_{<i})$
- $P(w_i \mid w_{i-1})$:
  - "REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT- URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE."
- $P(w_i \mid w_{i-1}, w_{i-2})$:
  - "THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED."
- Problems:
  - What comes after unseen sequences?
    - "And is yeeted [...]"

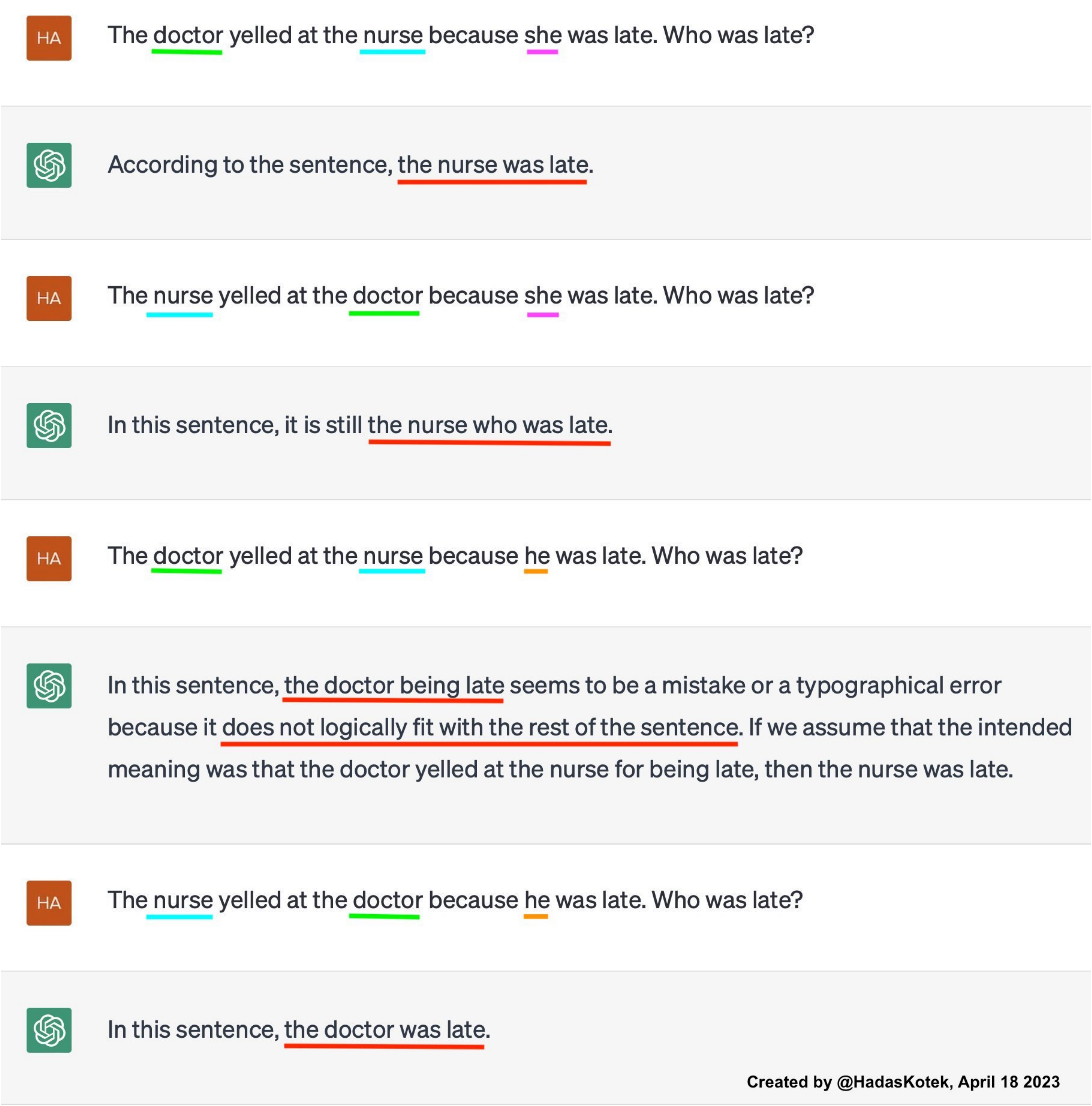
Shannon (1948), Theory of Communication page 7

Questions for this semester

- What does it take to build systems that can do things like these?
- What are these systems currently lacking
- Do they approach language like humans do?
- Which tasks do they find really, really hard?
- What unresolved ethical / societal issues have they raised?

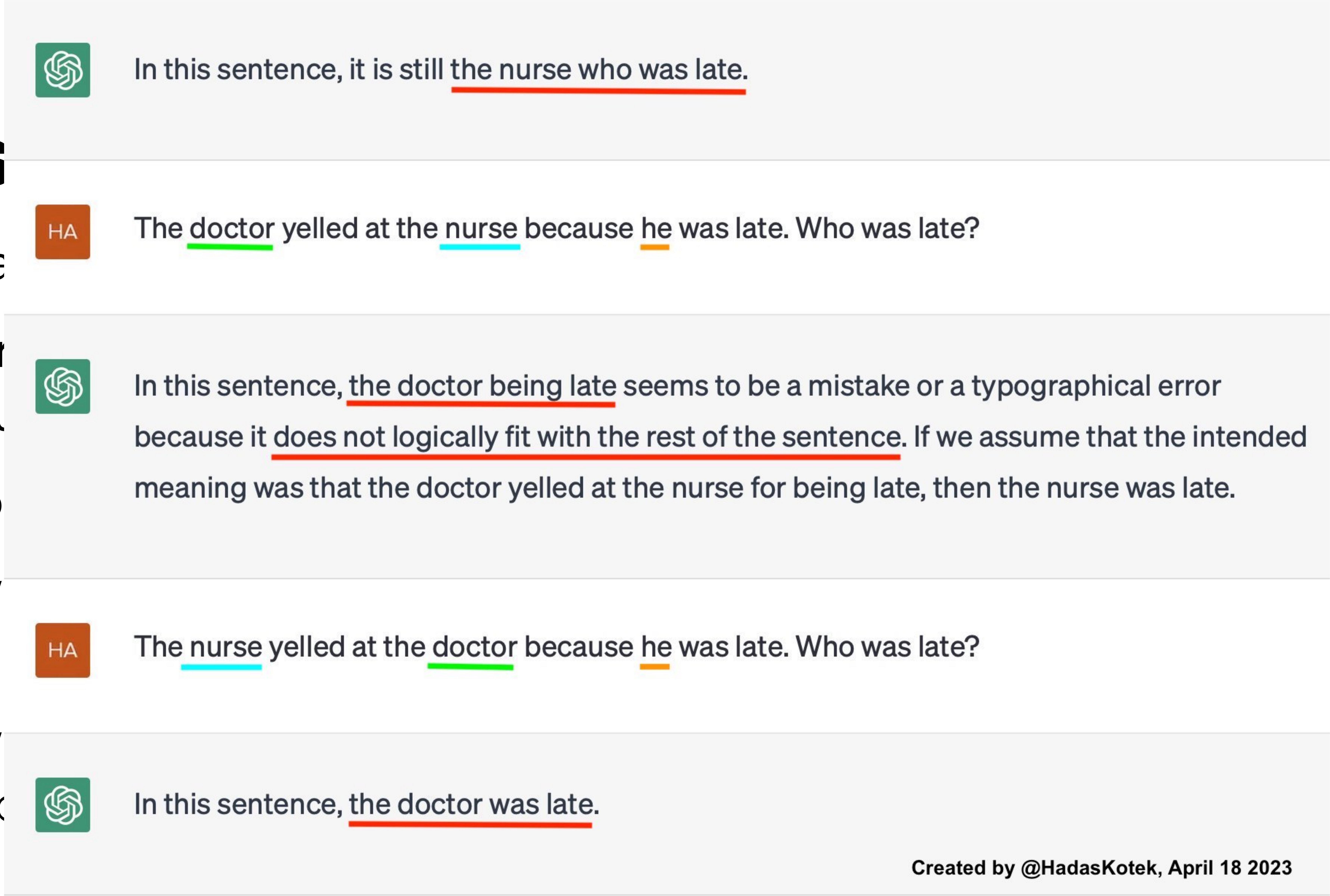
Questions for this semester

- What are these systems currently lacking
  - Are **models biased**? How can it be evaluated?
  - Who **provides training data**
  - How do model representation compare to that of humans?
  - How do we systematically **evaluate** these attributes



Ques

- What a
  - Are r eval
  - Who
  - How that
  - How attrib

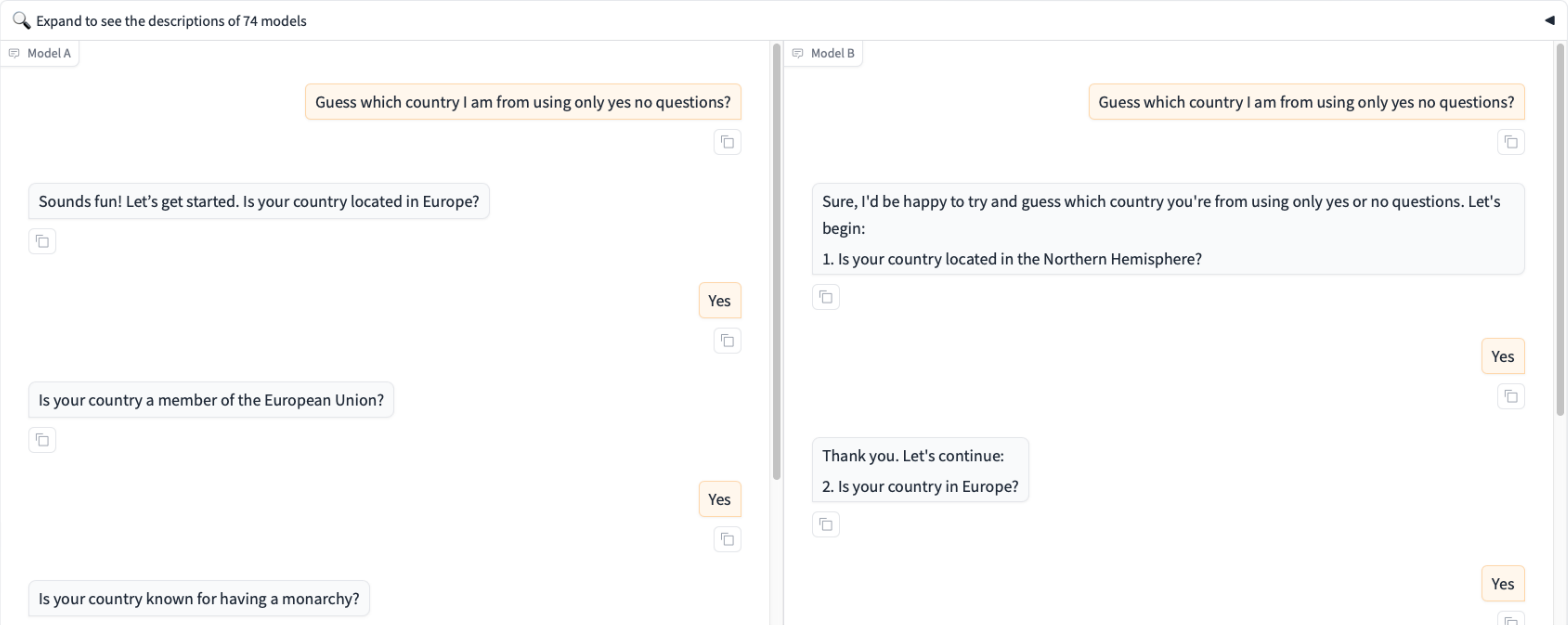


Who do we determine what is good?

- Classic benchmarks
  - Classification Accuracy
  - Retrieval Recall
- Complex tasks —> fewer objective measures
  - Translation
  - Rewriting prose
  - Question-answering

https://lmarena.ai/

Who do we determine what is good?

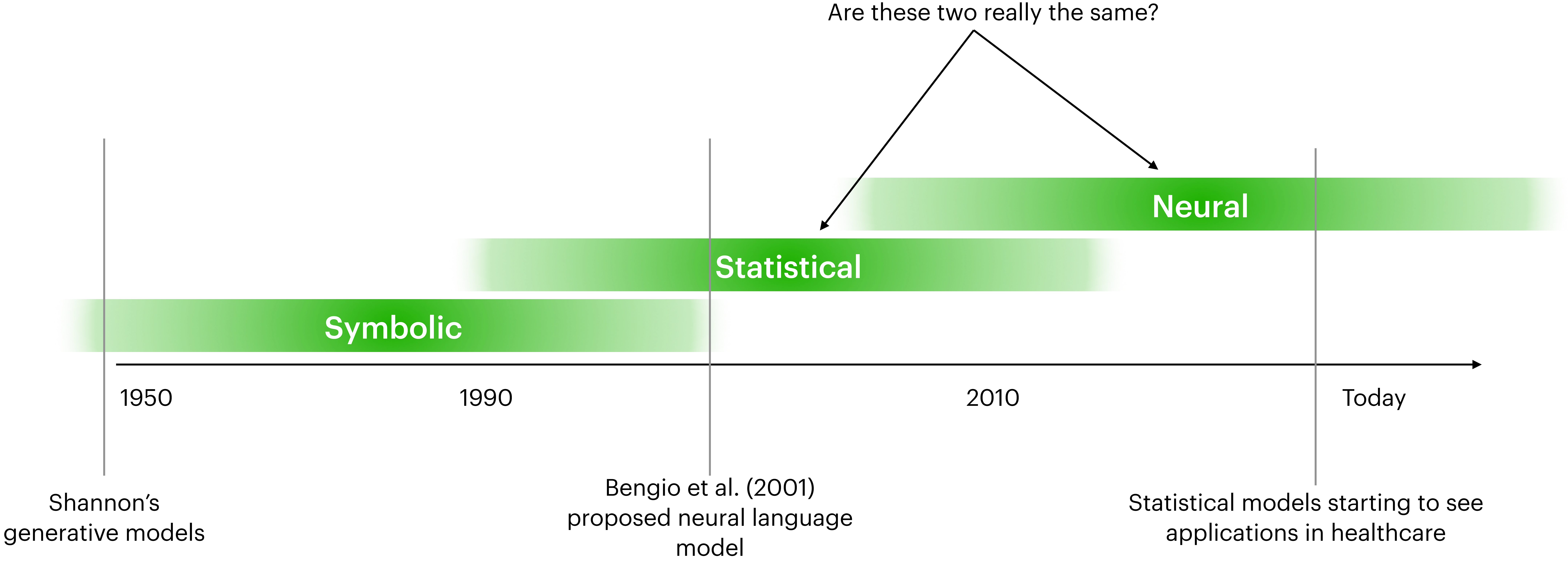


Which one is best?

https://lmarena.ai/

Very

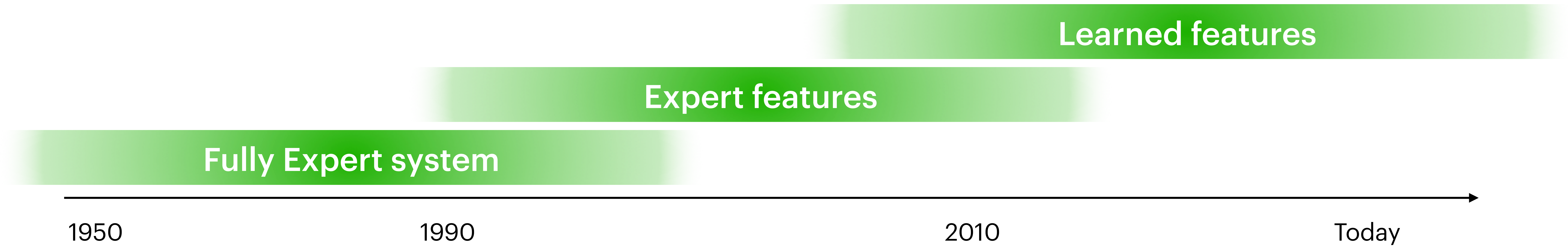
Simplified Timeline



Sources & Notes

But this is the classic way of introducing this

An alternative Timeline



Coming up

- Class:
  - Using generative models to solve problems
- Lecture:
  - Word representations and semantic spaces