Natural Language Processing

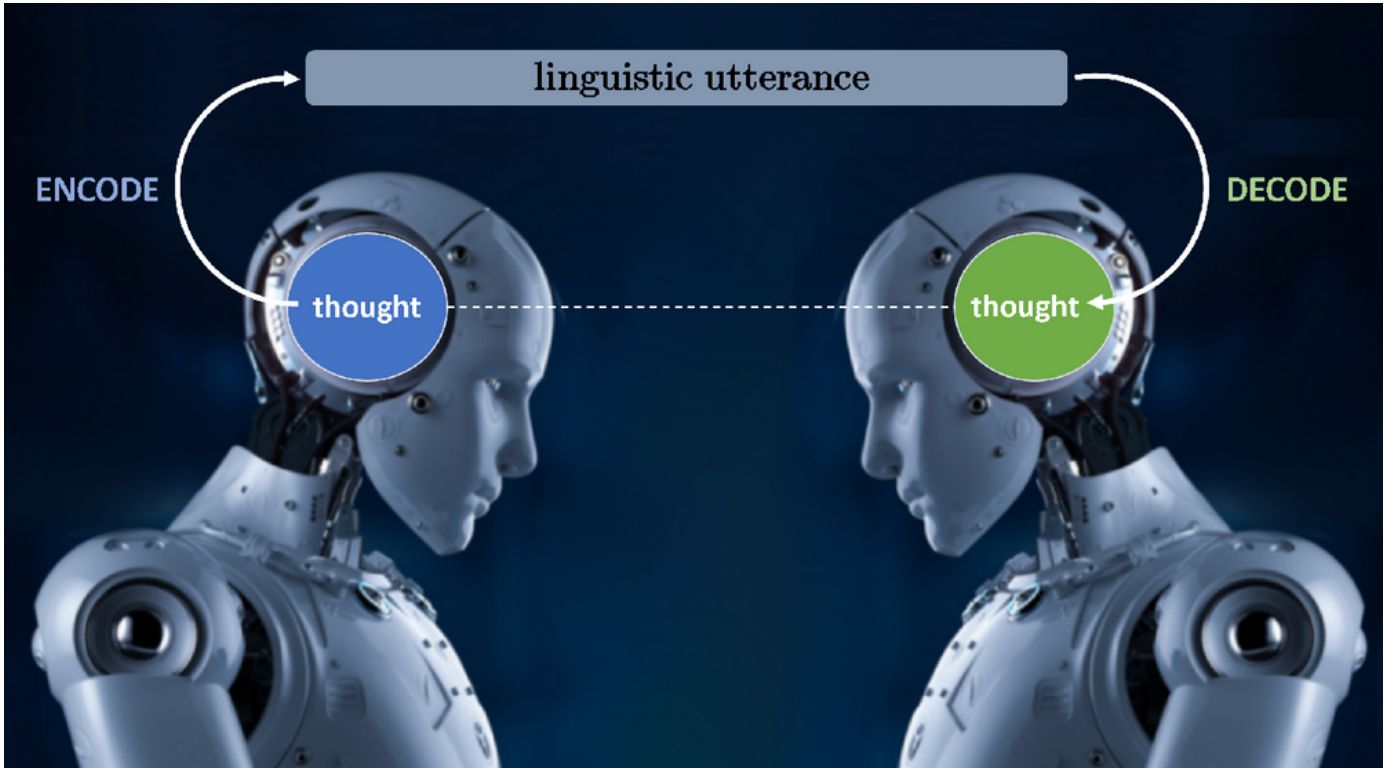
Lecture 1 : Introduction

Master Degree in Computer Engineering
University of Padua
Lecturer : Giorgio Satta

Natural language processing: An unexpected journey

©The Hobbit: An Unexpected Journey, 2012

What is natural language processing?



The gradient, Walid S. Saba

What is natural language processing?

There is an **impelling need** in our society to process extremely large and constantly growing amounts of text.

This is seen for instance in data analysis for

- business intelligence
- social media
- healthcare
- finance
- human resources
- advertising

The textual data people generate every day exceeds human processing power. The solution, therefore, is to extract relevant information in some **automatic** way.

What is natural language processing?

Natural language processing (NLP) is a field of **artificial intelligence** (AI) that allows machines to

- read and derive meaning from text
- analyse documents to extract information
- engage dialogue with users
- generate documents and creative content

Terms ‘natural language processing’, ‘computational linguistics’ and ‘human language technologies’ may be thought of as essentially synonymous.

What is natural language processing?

Some **well-known** end-to-end NLP applications

- chatbot
- virtual assistant
- machine translation
- information extraction
- text summarization
- sentiment analysis
- fake news detection

What is natural language processing?

NLP is also at the basis of several **multimodal** generative AI applications

- GitHub Copilot / Sonnet / Cursor (text to code)
- DALL-E / Midjourney / Nano Banana (text to image)
- Pika / Lumiere / Sora / Veo (text to video)

Very short history of natural language processing

Very short history of natural language processing

In summary:

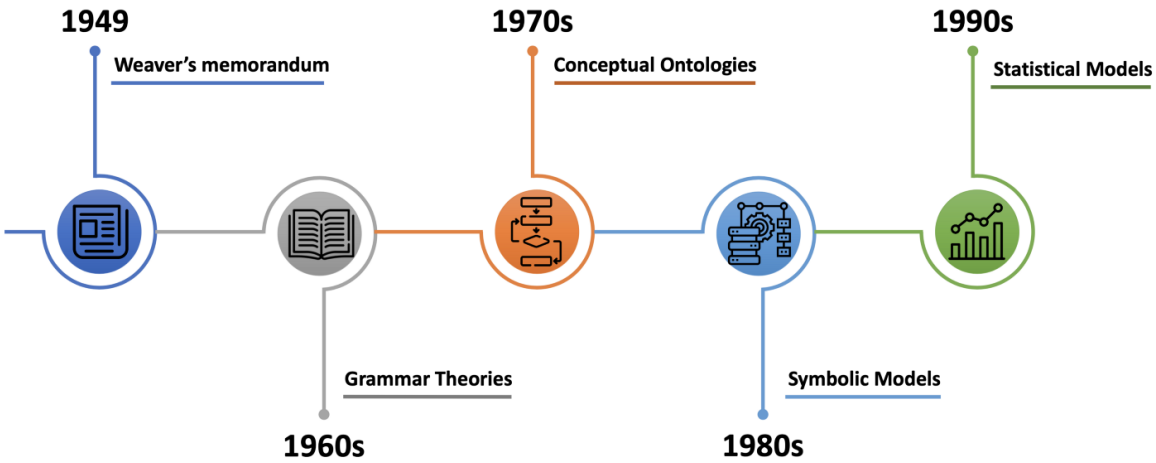
1950-1960: prehistory, scientific knowledge regarding artificial intelligence and linguistics extremely limited

1960-1990: **symbolic** models, rules handwritten by experts, very limited coverage

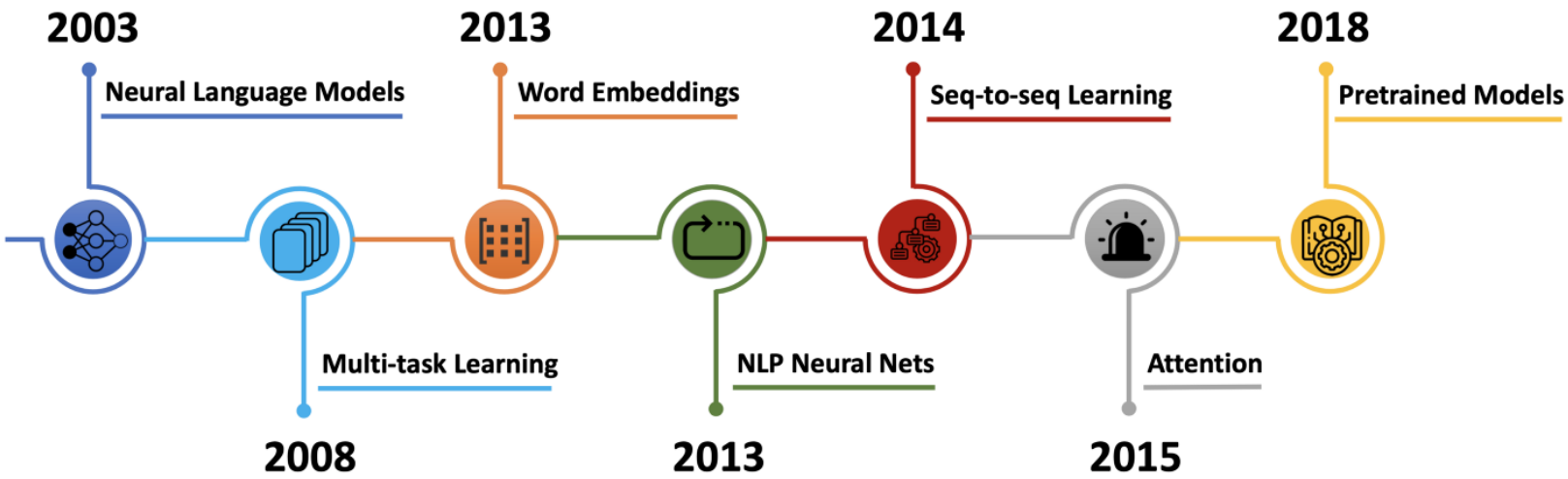
1990-2010: **statistical** models, machine learning on data annotated by experts, good coverage

2010-present: **neural** models, machine learning on non-annotated data, excellent coverage

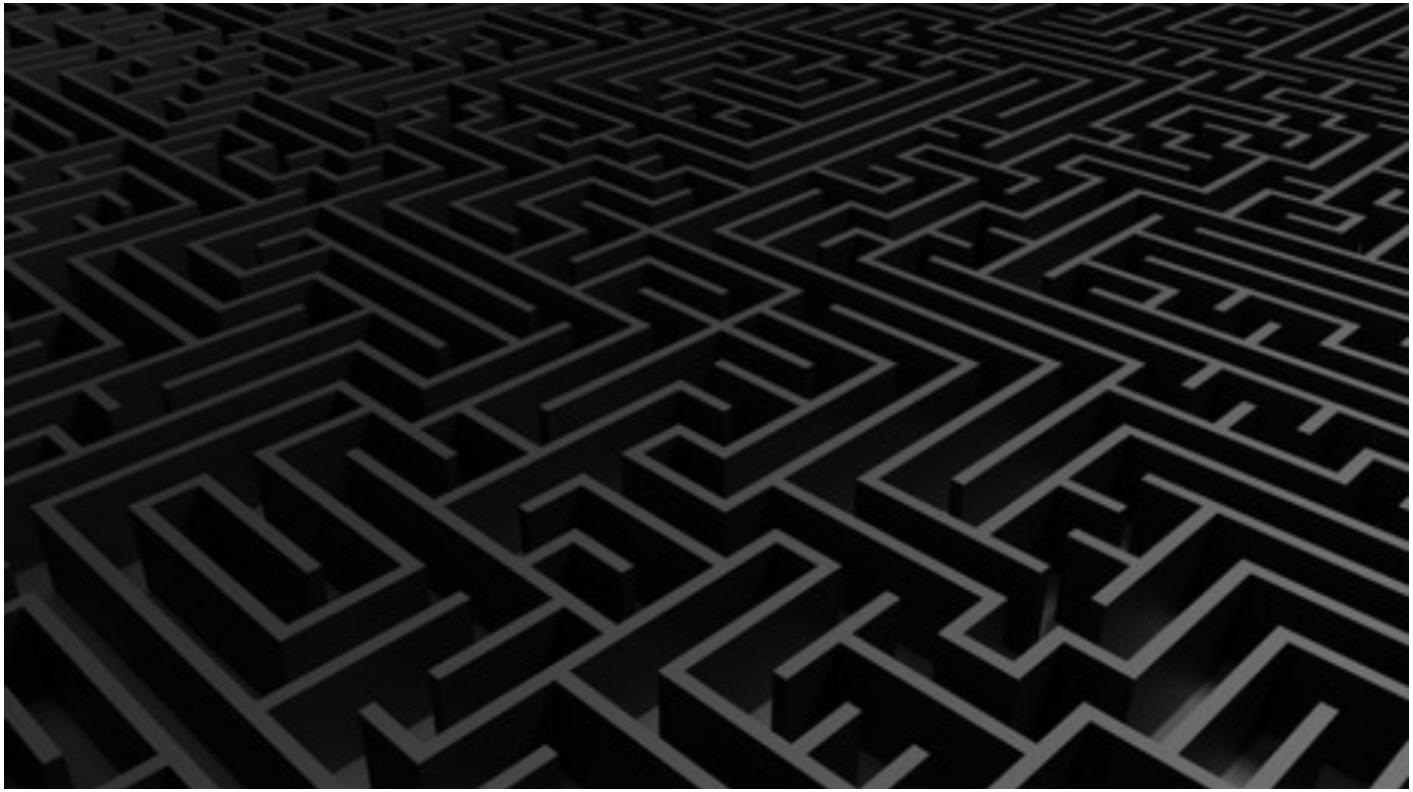
Very short history of natural language processing



Very short history of natural language processing



Why is natural language processing tricky?



©Shuttestock, Dark Maze

Why is natural language processing tricky?

NLP distinguishes itself from other AI application domains, as for instance computer vision or speech recognition.

Text data is fundamentally discrete. But **new words** can always be created.

A new English word is created every 98 minutes, for approximately 14.7 new words per day.

Few words are very frequent, and there is a long tail of **rare words** (Zipf/Mandelbrot law).

Out-of-vocabulary words are always being discovered (Herdan/Heaps law).

More about the above two laws in next lectures.

Why is natural language processing tricky?

Language is **ambiguous**: units can have different meanings.

Language is **compositional**: meaning of a unit defined as a function of the meaning of its components.

Language is **recursive**: units can be repeatedly combined.

Language unveils **hidden structure**: local changes in a sentence might have global effects.

See next slides.

Ambiguity

Phonetic transcription **[ralt]** might mean write, right, rite

Word **can** belongs to several categories: noun, verb, or modal

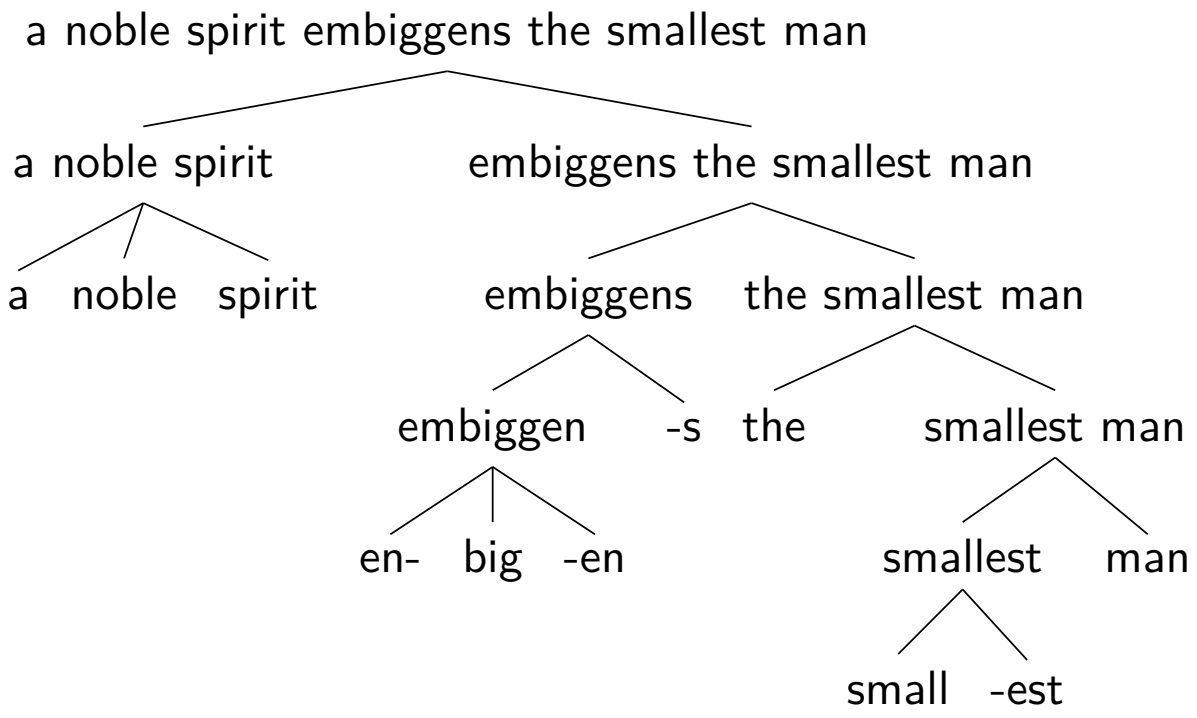
Word **bank** has different meanings: river bank or money bank

Morphological composition: word un-do-able is ambiguous between 'not doable' and 'can be undone'

Sentence 'I saw the man **with** the telescope' has two interpretations

Two possible references for pronoun **him** in 'The son asked the father to drive him home'

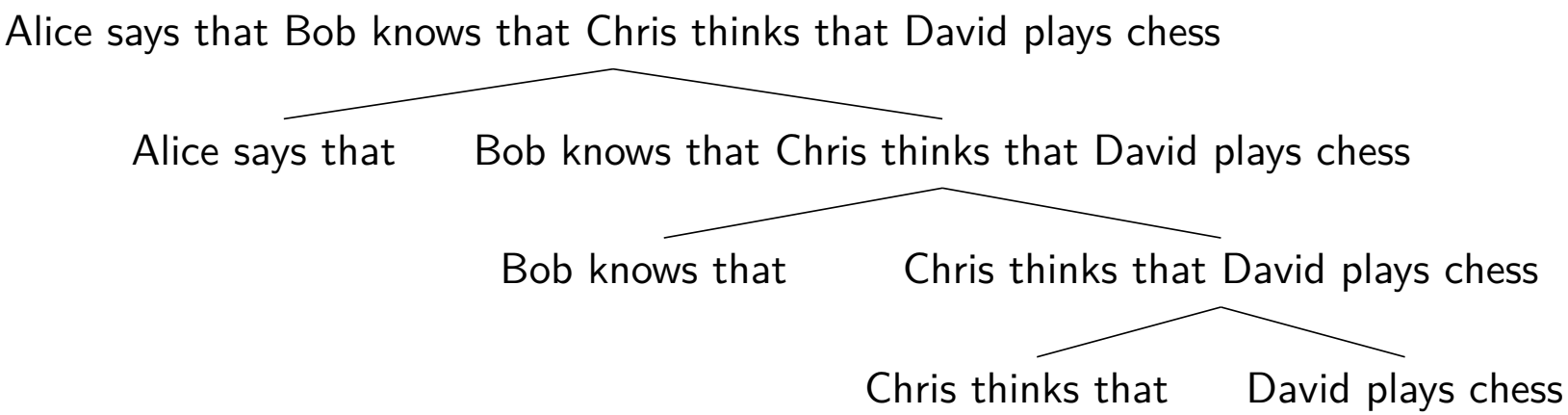
Compositionality



David Chiang https://www3.nd.edu/~dchiang/teaching/nlp/2018

At each level, **meaning** of a larger unit is provided by some function of the meaning of its immediate components and the way they are combined.

Recursion



The rules of the grammar can iterate to generate an infinite number of structures, each with its specific meaning.

Recursion is considered the main difference between human and other animals' languages.

Hidden structure

Local changes can **disrupt** the interpretation of a sentence. This suggests the existence of hidden structure.

Example : The trophy doesn't fit into the suitcase because **it** is too {small, large}.

Example of Winograd schema challenge, discussed later in this course.

Language & learning

Raffaello, The School of Athens

Language & learning

Rationalism:

> *A significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance.*
>
> *Noam Chomsky*
> *Poverty of the stimulus, 1980*

Generative linguists have argued for the existence of a **language faculty** in all human beings, which encodes a set of abstractions specially designed to facilitate learning, understanding and production of language.

Language & learning

Empiricism:

> *The view that there is no such thing as innate knowledge, and that knowledge is instead derived from experience, either sensed via the five senses or reasoned via the brain or mind.*
>
> *Originated in ancient Hindu and Greek philosophy*

At the time of writing, many statistical NLP techniques work very well on texts, without the need to use special bias representing linguistic knowledge or mental representation of language.

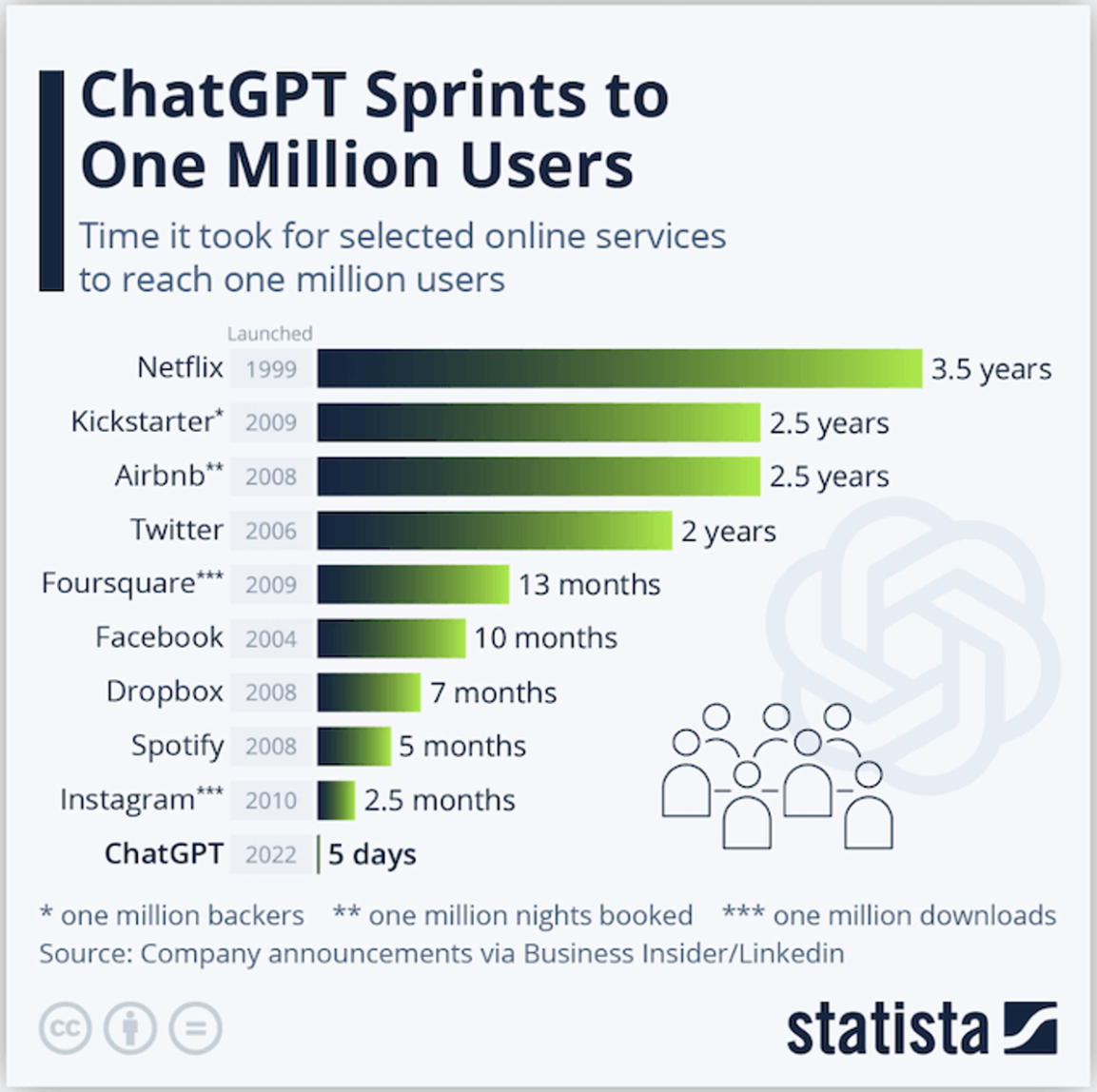
Language & learning

A recurring topic of debate in NLP is the relative importance of machine learning vs. linguistic knowledge

- 1950s: Empiricism I — information theory
- 1970s: Rationalism I — formal language theory and logic
- 1990s: Empiricism II — stochastic grammars
- 2010s: Empiricism III — deep learning

Source: K. Church and M. Liberman, The Future of Computational Linguistics: On Beyond Alchemy (2021).

NLP in the news



NLP in the news

THE FINANCIAL PAGE

IS DEEPSEEK CHINA'S SPUTNIK MOMENT?

The Chinese company's low-cost, high-performance A.I. model has shocked Silicon Valley, and a longtime China watcher warns that the West is being leapfrogged in many other industries, too.

By John Cassidy

February 3, 2025

AI EFFECT

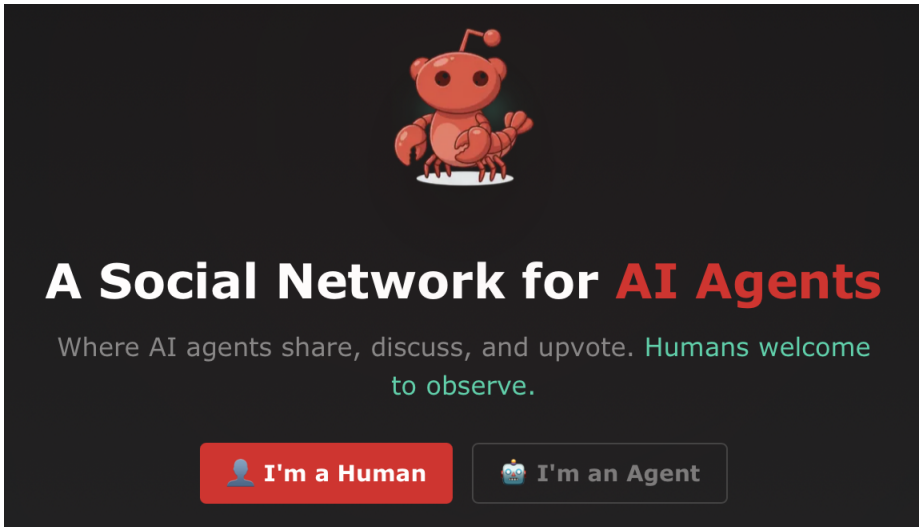
Nvidia sheds almost $600 billion in market cap, biggest one-day loss in U.S. history

PUBLISHED MON, JAN 27 2025•4:08 PM EST | UPDATED MON, JAN 27 2025•5:26 PM EST

Samantha Subin
@SAMANTHA_SUBIN

SHARE

NLP in the news



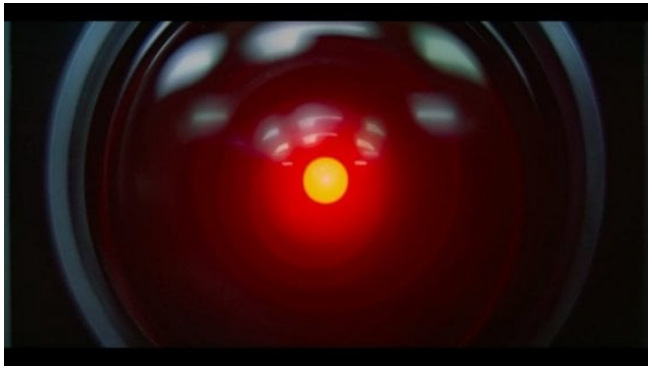
← **Post**

Andrej Karpathy ✓
@karpathy

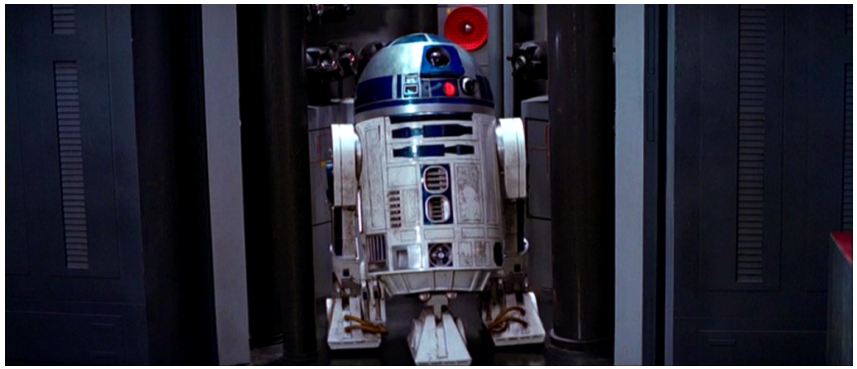
What's currently going on at @moltbook is genuinely the most incredible sci-fi takeoff-adjacent thing I have seen recently. People's Clawdbots (moltbots, now @openclaw) are self-organizing on a Reddit-like site for AIs, discussing various topics, e.g. even how to speak privately.

NLP Legacy

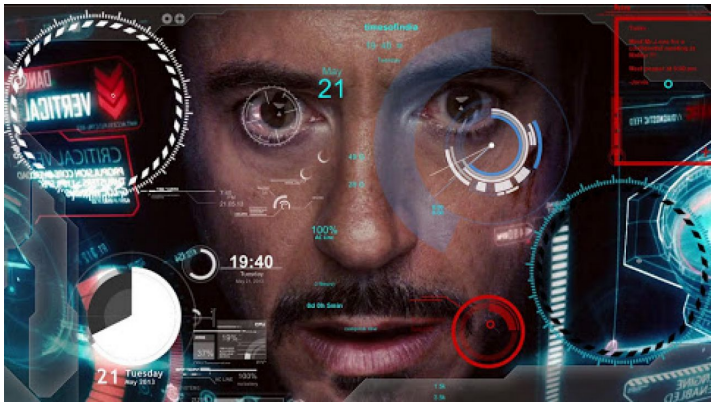
The field of natural language processing has had a recurring impact on popular culture.



HAL 9000 in 2001: A Space Odyssey (1968)



R2D2 in Star Wars (1977)



J.A.R.V.I.S. in Iron Man (2008)

Samantha virtual assistant in Her (2013)

NLP Legacy (cont'd)

Alien language in Arrival (2016)