

An intro to language data

NLP Week 1

Plan for today

1. What is NLP and what is this class?

1.1. What we will cover

1.2. Class format

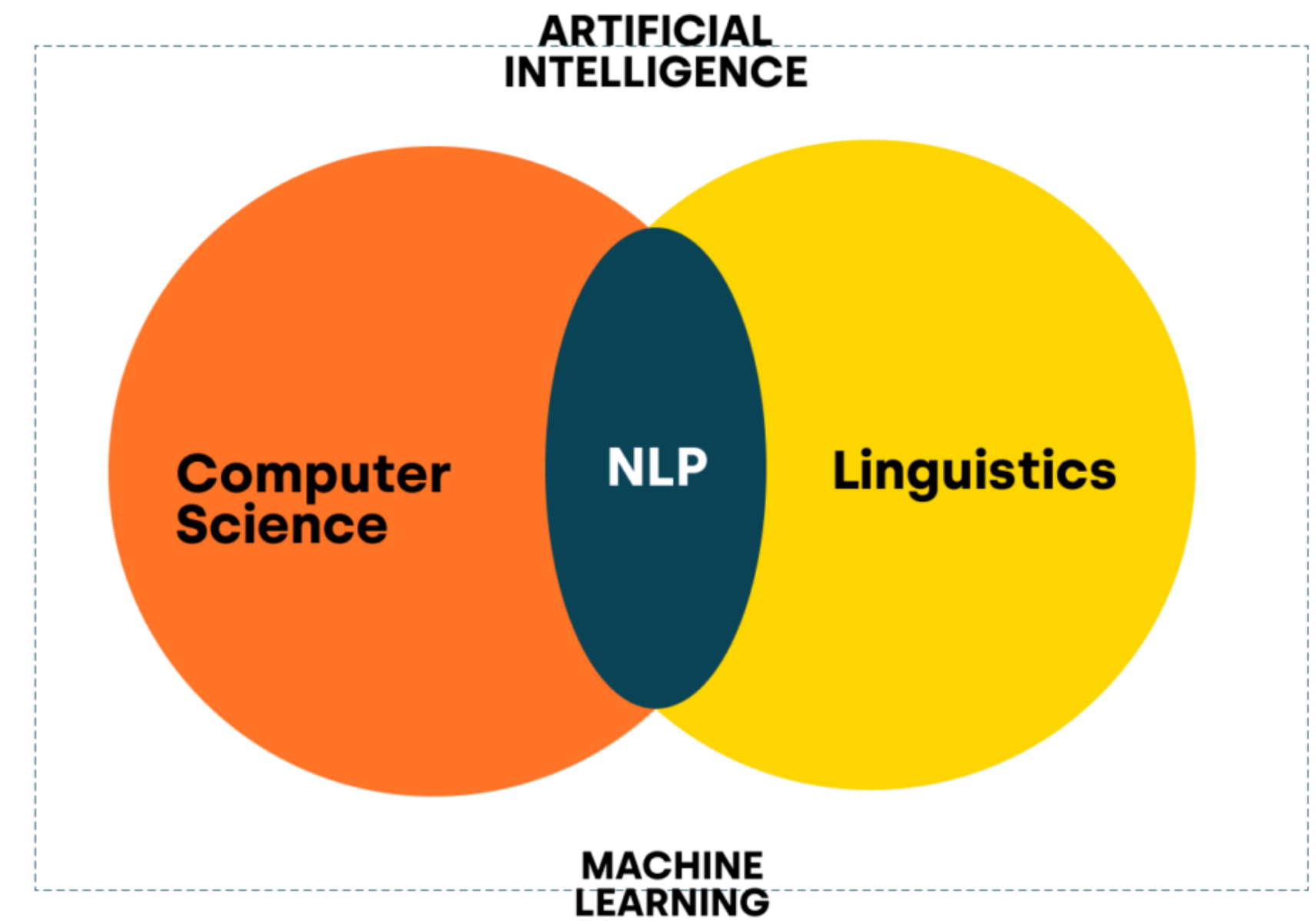
1.3. Evaluations

1.4. Respect and expectations

2. What is linguistic data and why is it so special?

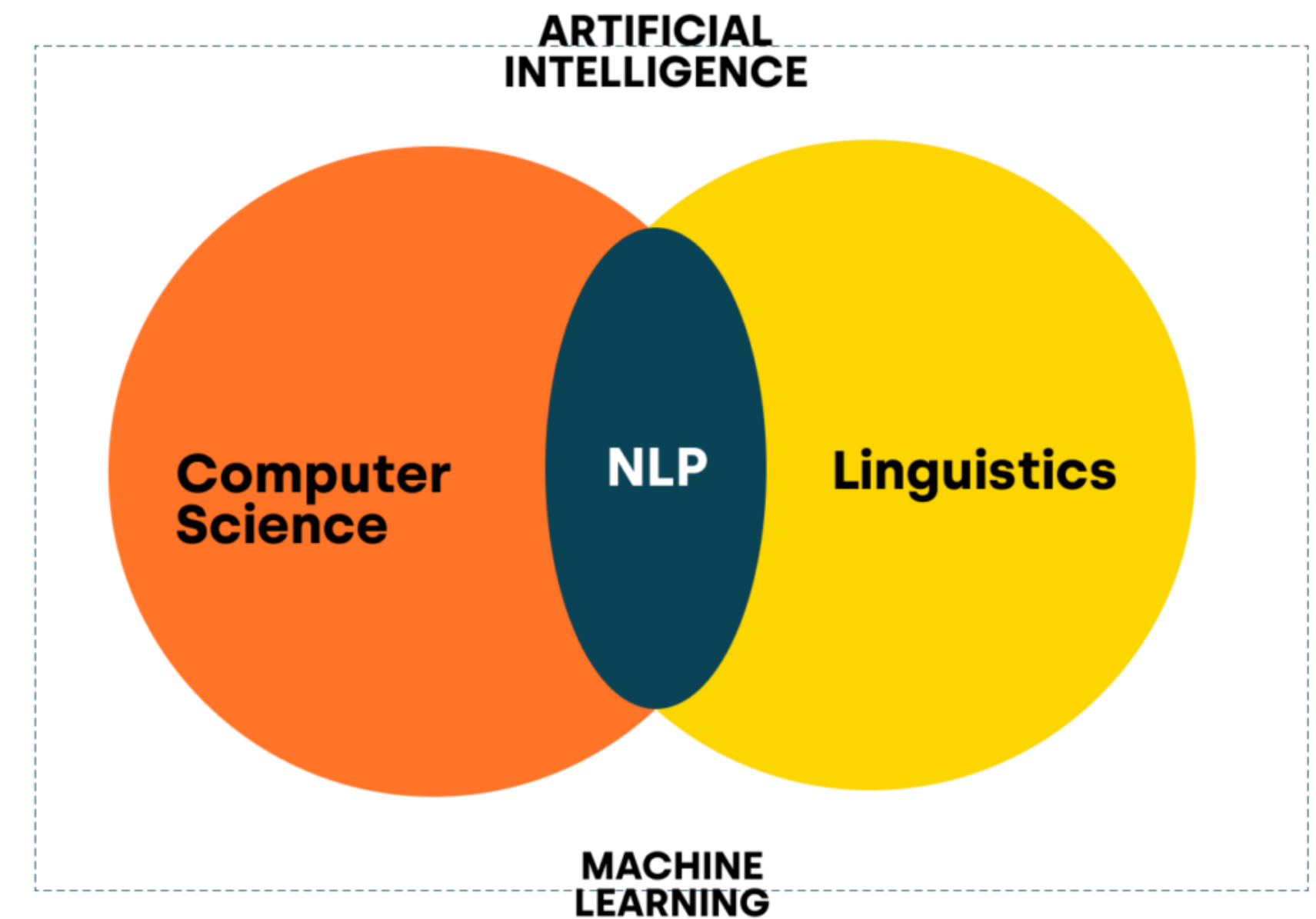
3. *Group exercises:* Setting up your coding environment

What is NLP?



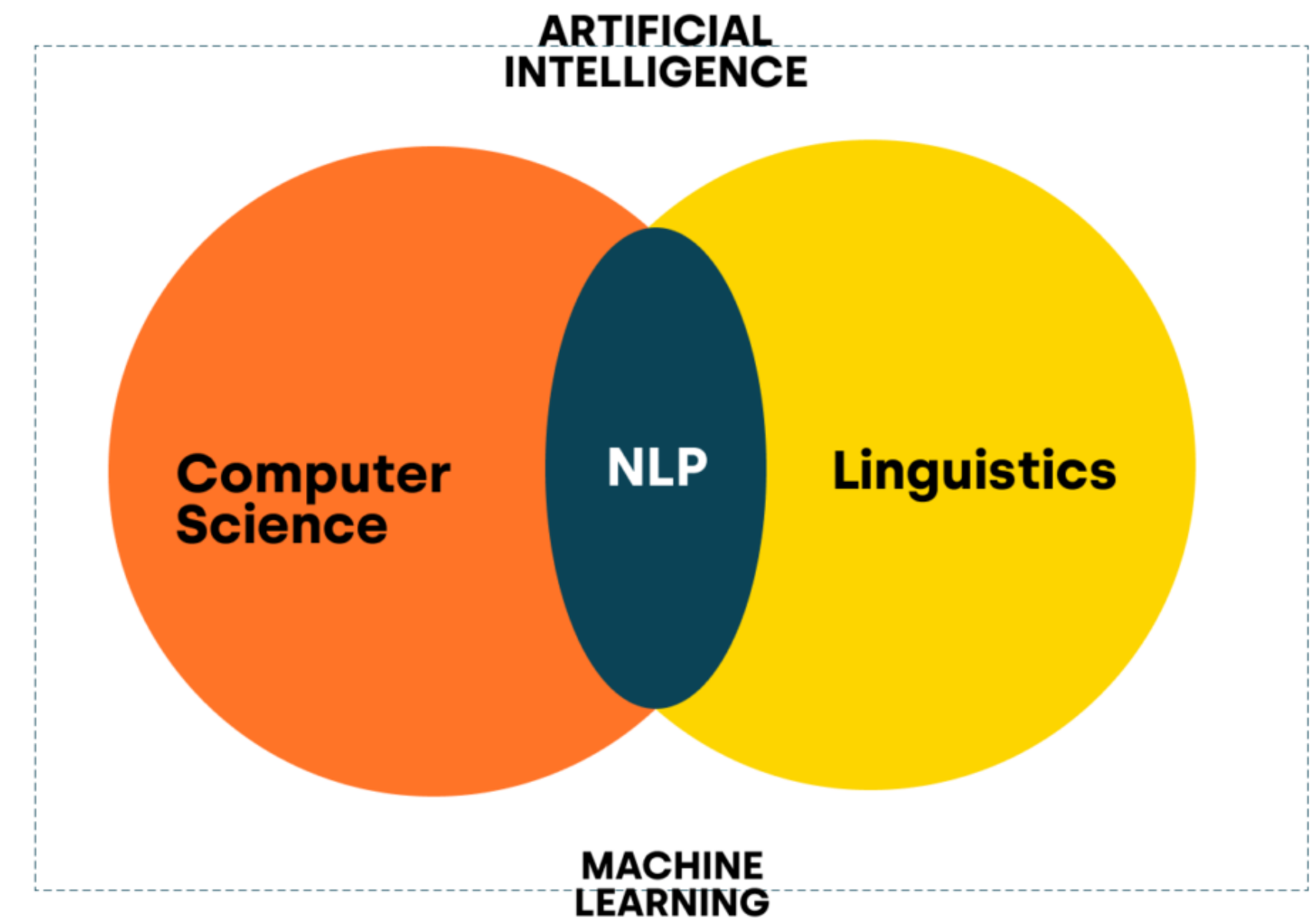
What is NLP?

- NLP = Natural language + processing



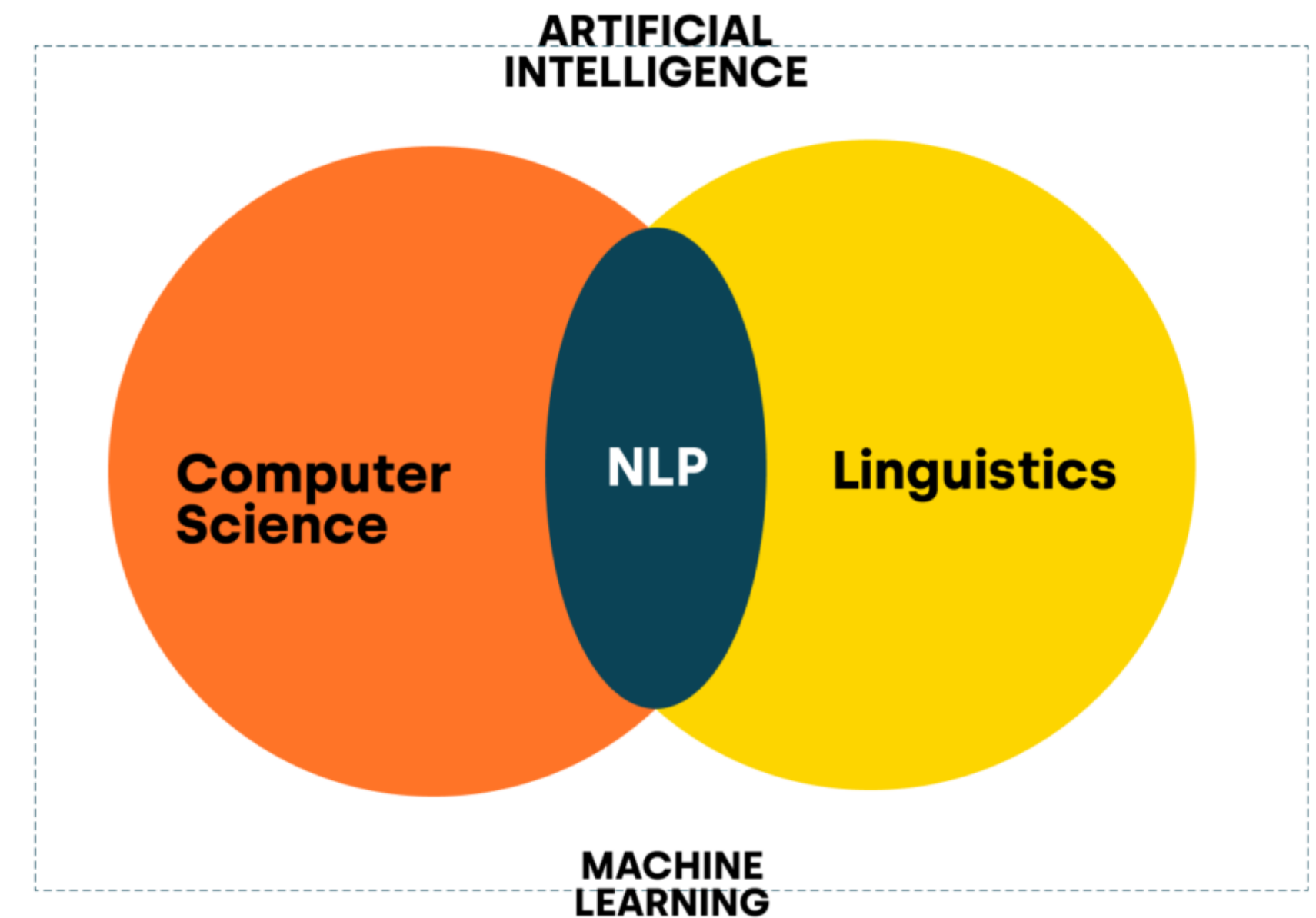
What is NLP?

- NLP = Natural language + processing
- **Definition:** An interdisciplinary field of linguistics and computer science that develops and studies algorithms to ***learn, process*** and ***generate*** natural language data.



What is NLP?

- NLP = Natural language + processing
- **Definition:** An interdisciplinary field of linguistics and computer science that develops and studies algorithms to ***learn, process*** and ***generate*** natural language data.
- Today, most prominent algorithms are machine learning, and in particular neural network based.



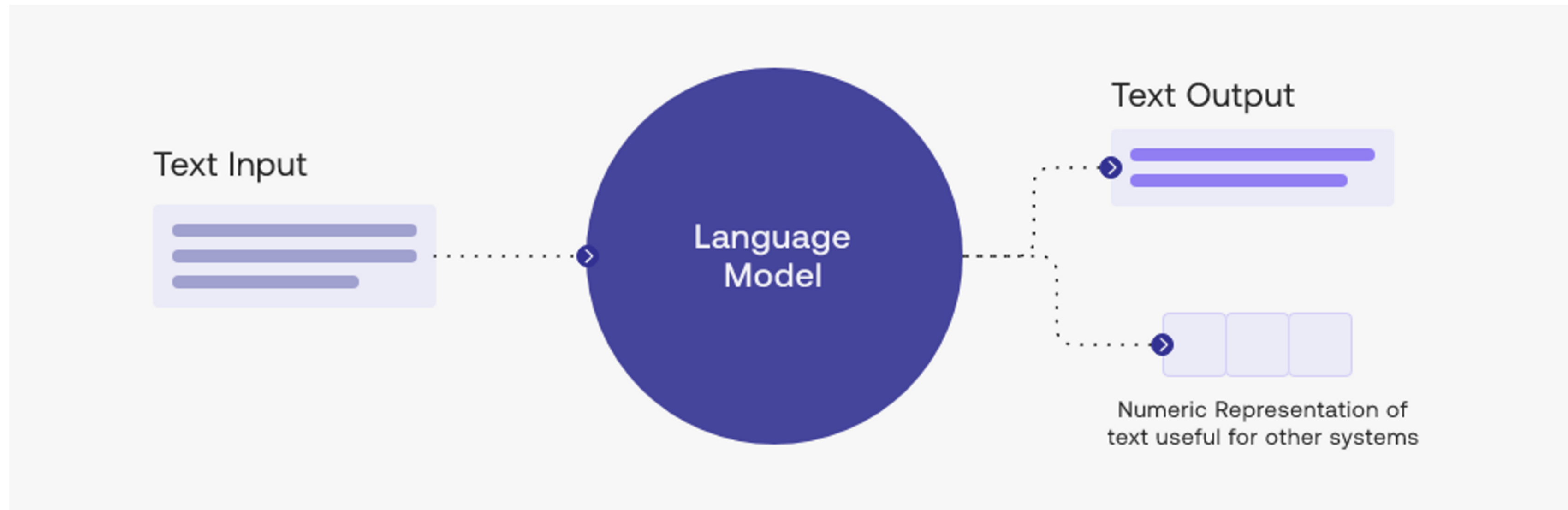
What we will cover:

What we will cover:

LANGUAGE MODELS


What we will cover:

LANGUAGE MODELS



What we will cover:

LANGUAGE MODELS

- 
- **Bag-of-words** — topic models and naive classification
 - **N-gram language models (Markov models)** — Probabilistic models of language in context
 - **Hidden Markov models** — part-of-speech tagging
 - **Distributed representations** — vector semantics
 - **Recurrent neural language models, LSTMs** — language generation
 - **Transformers and masked language modeling** — language learning
 - **Encoder models** — semantic search and RAG (retrieval augmented generation)
 - **Encoder-decoder models** — text translation/text summarisation

Class format

- Every class time:
 - First half — lecture on new material
 - [15 minutes break]
 - Second half — hands on group coding exercises (bring laptops to class!)
- **Elephant in the room** - Me! Maternity leave as of March 10th
- Your in good hands! Subsequent classes will be given by **Dr. Marius Mosbach**, a very cool and knowledgeable NLP Postdoc from Mila - Quebec AI Institute.



Evaluations

- **Two types of evaluations:** Assignments (50%) and Exams (50%)
- **Assignment 1 (25%)** - Due ... All applied coding problems
- **Midterm exam (25%)** - TBD ... written theoretical problems
- **Assignment 2 (25%)** - Due ... All applied coding problems
- **Final exam (25%)** - TBD ... written theoretical problems

Respect and expectations

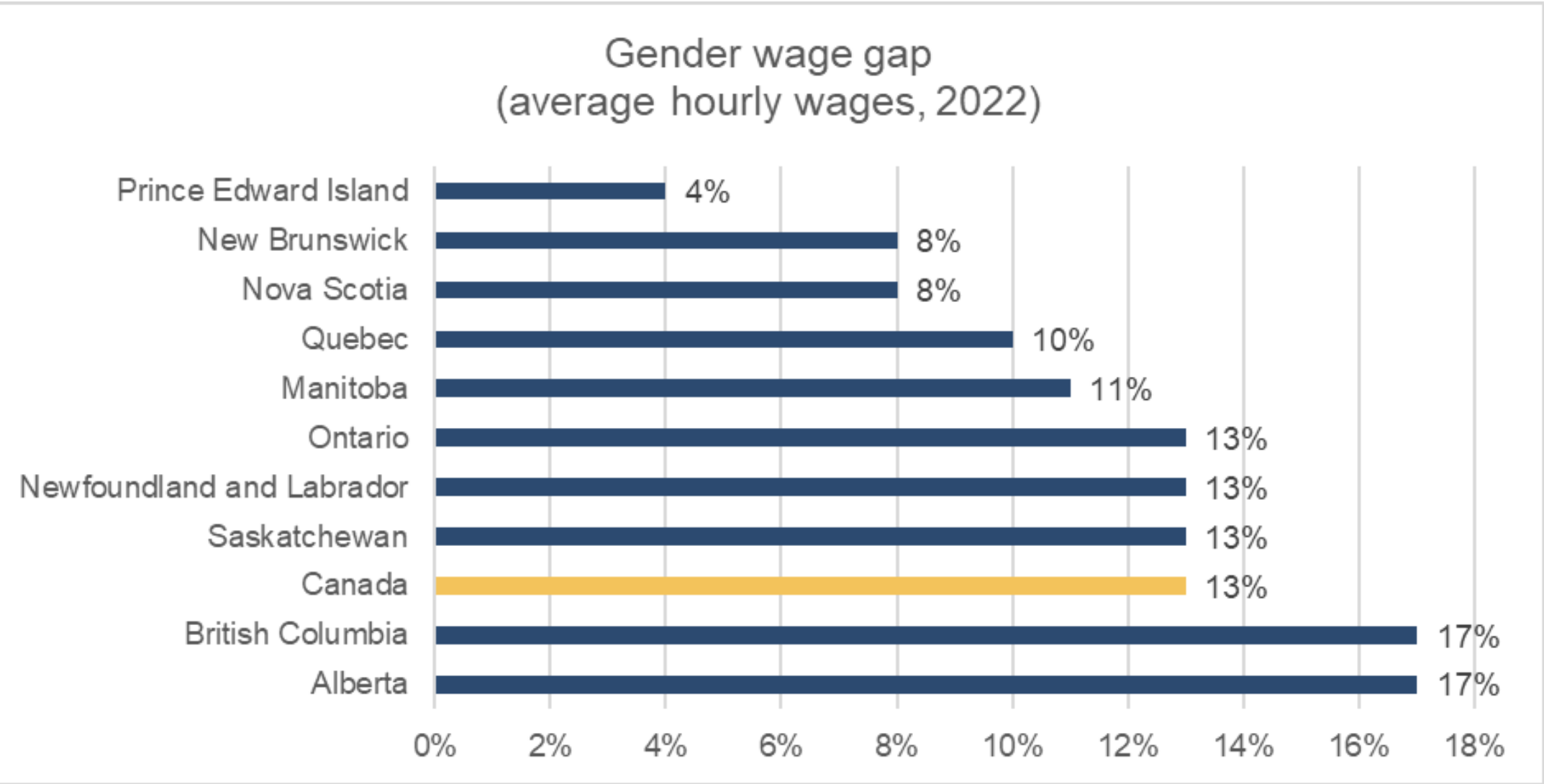
- **What to call me in class:** Professor or Professor Portelance.
- **Email policy:** Please allow up to 2 working days for all email responses. I will not respond on weekends.
- **Group work policy:** You may work on assignments in groups of up to 3 people. If you work with others, you must write their names where indicated. You must still each submit your own assignment and will be individually graded.
- **Late assignment policy:** It's minus 15% of your assignment grade per late day, no exceptions. You will be given assignments 3 weeks in advance before their due date — plan accordingly, there is no reason you should be late.

Now the fun part.

Linguistic data... why so special?

Mortgage Rates Hold Near 7%

The average rate on a 30-year fixed mortgage edged lower to 6.78%. It was one basis point higher last week.

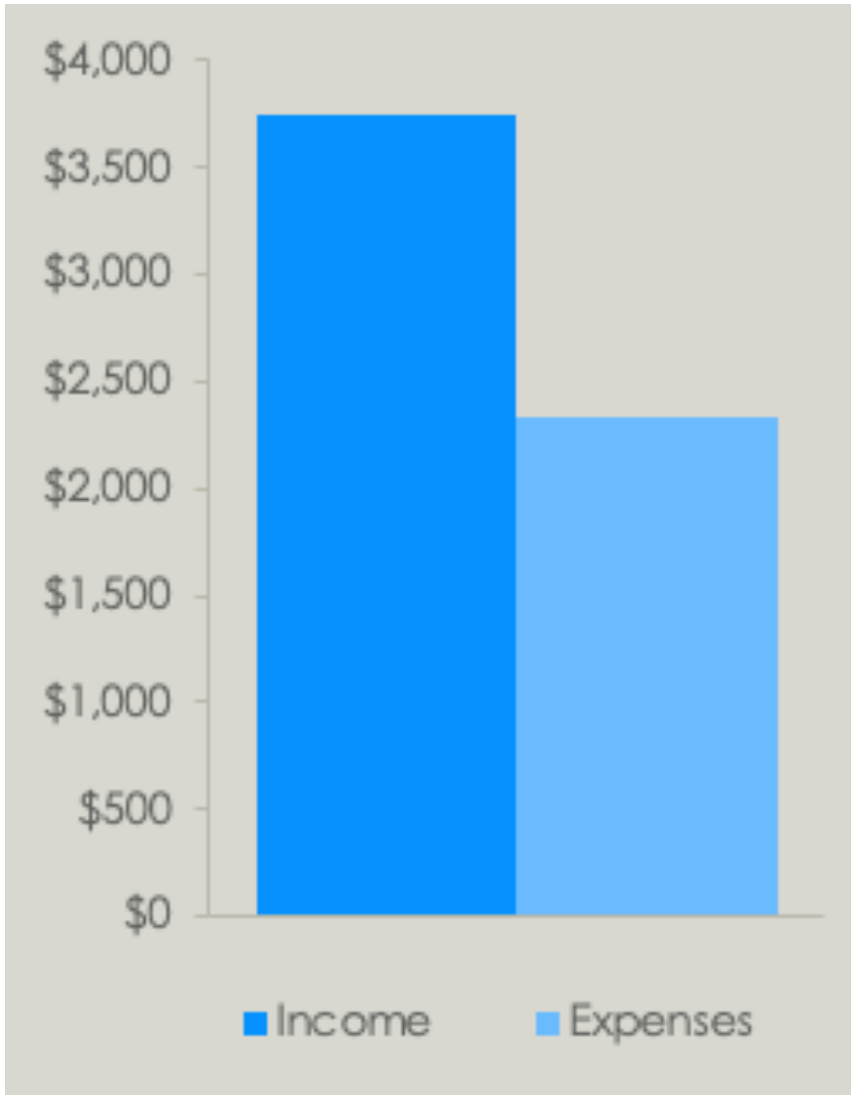


Year	Population
1860	56,802
1870	149,473
1880	233,959
1890	298,997
1900	342,782
1910	416,912
1920	506,676
1930	634,394
1940	634,536
1950	775,357
1960	740,316
1970	715,674
1980	678,974
1990	723,959
2000	776,733

CHAPTER 1

Loomings

Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It



Linguistic data... why so special?

Language is a structured data type!

Linguistic data... why so special?

This is a sentence composed of words.

-> [This, is, a, sentence, composed, of, words]

Words are composed of morphemes.

-> Word-s are compos-ed of morpheme-s

Morpheme: the smallest meaningful constituents within a linguistic expression and particularly within a word. Eg. Morpho-logi-c-al

Linguistic data... why so special?

Document > sections > paragraphs > sentences > morphemes > characters

Linguistic data... why so special?

Language is a hierarchical data type!

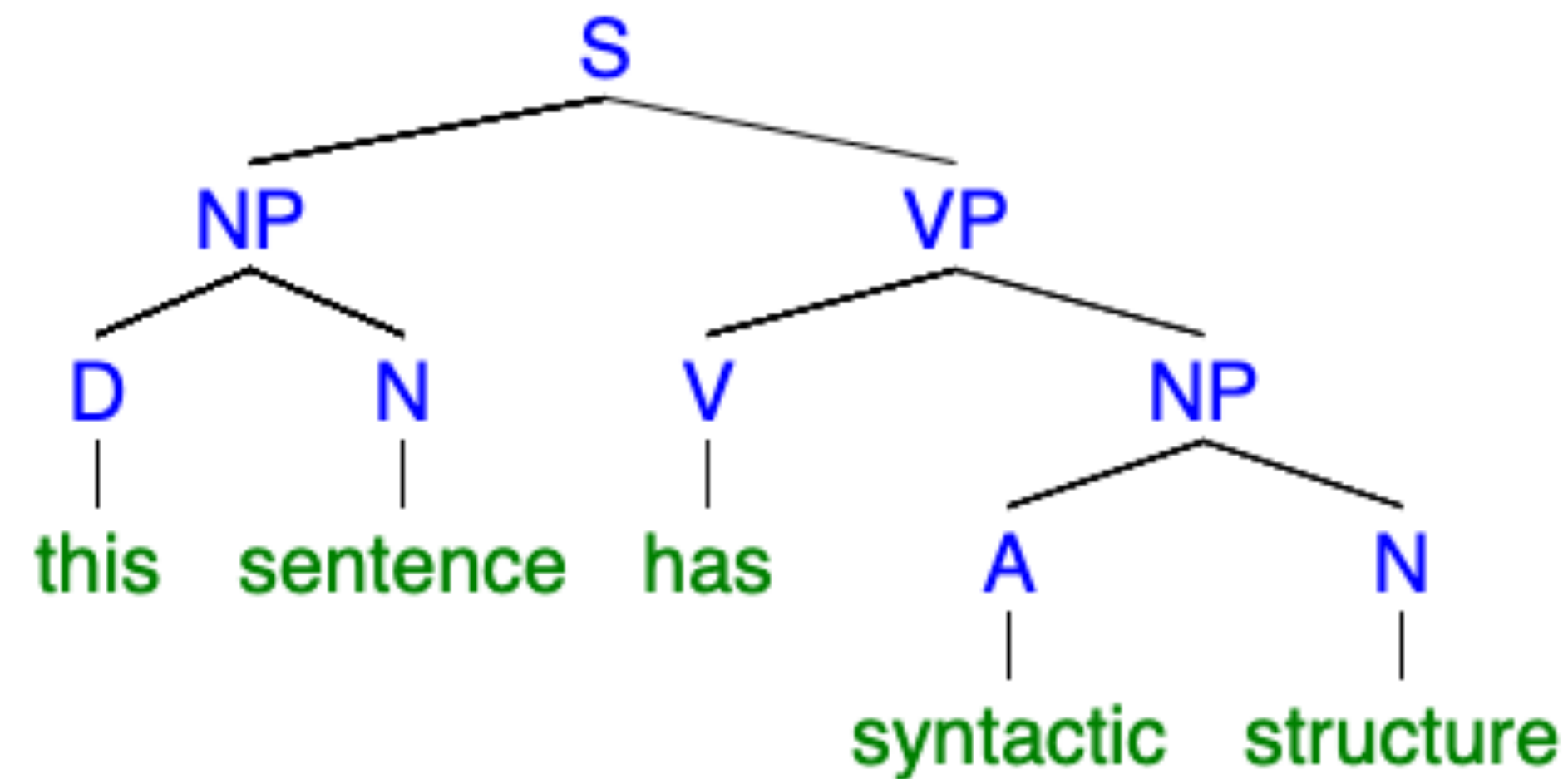
Linguistic data... why so special?

There are word types or categories: Nouns, verbs, adjectives, adverbs, function words

Not all categories are equal.

Linguistic data... why so special?

The structure of a sentence is called **syntax**.

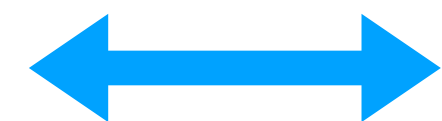


Linguistic data... why so special?

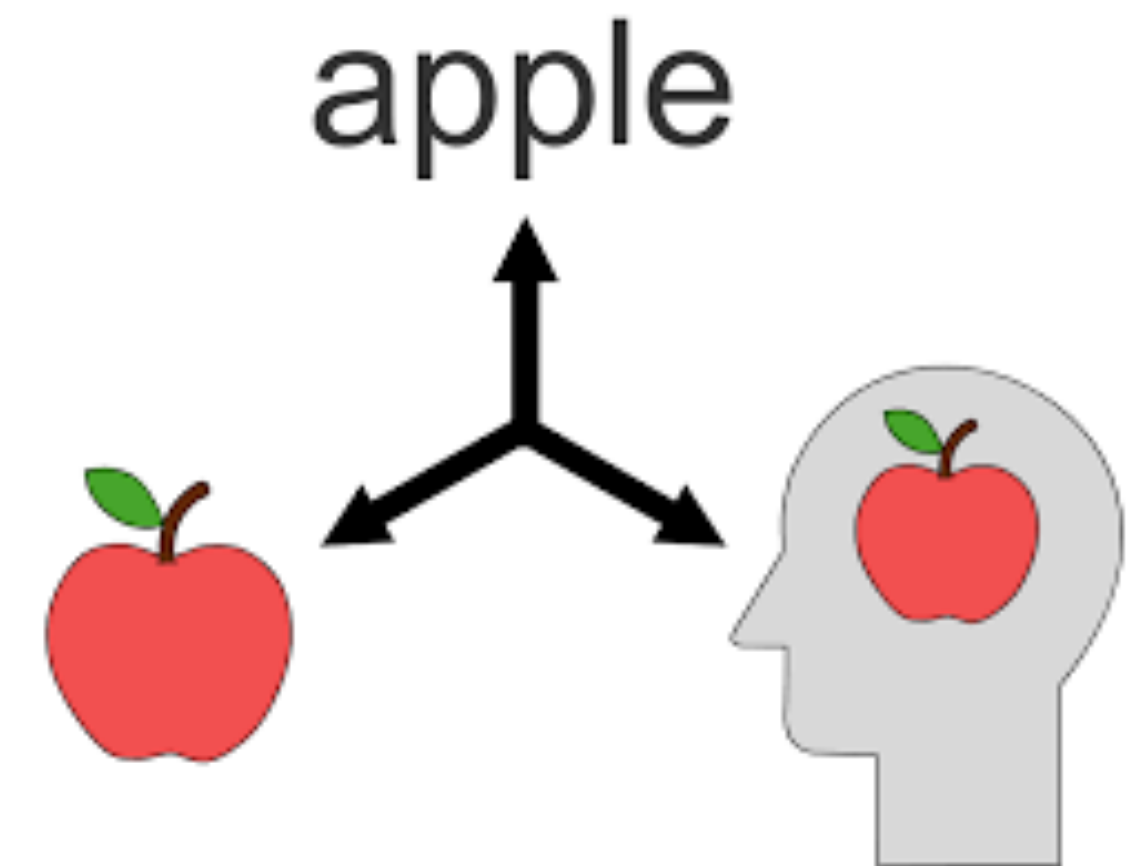
Language is a symbolic data type!

Linguistic data... why so special?

The meaning of a sentence is called **semantics**.

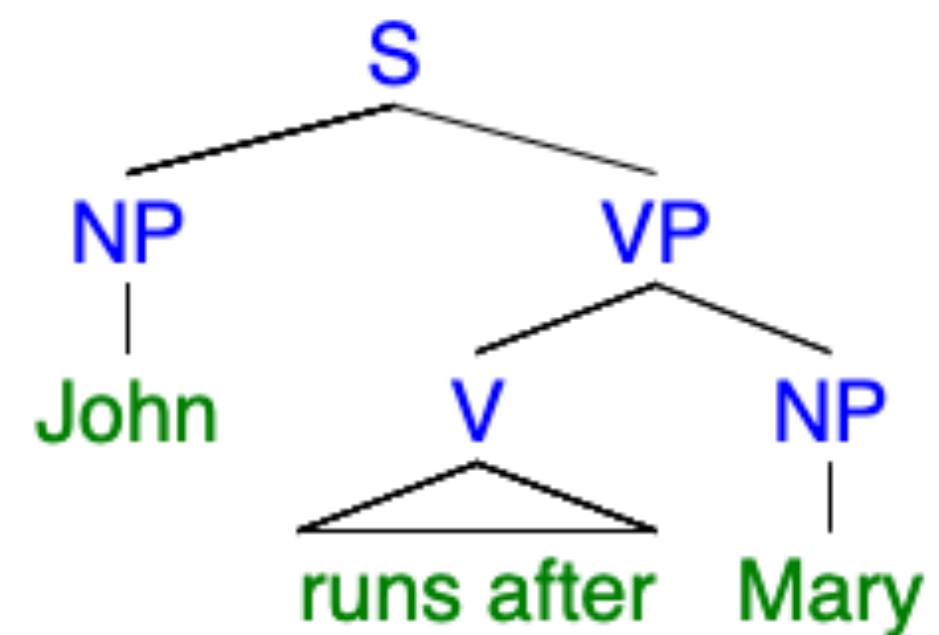


John runs after Mary.

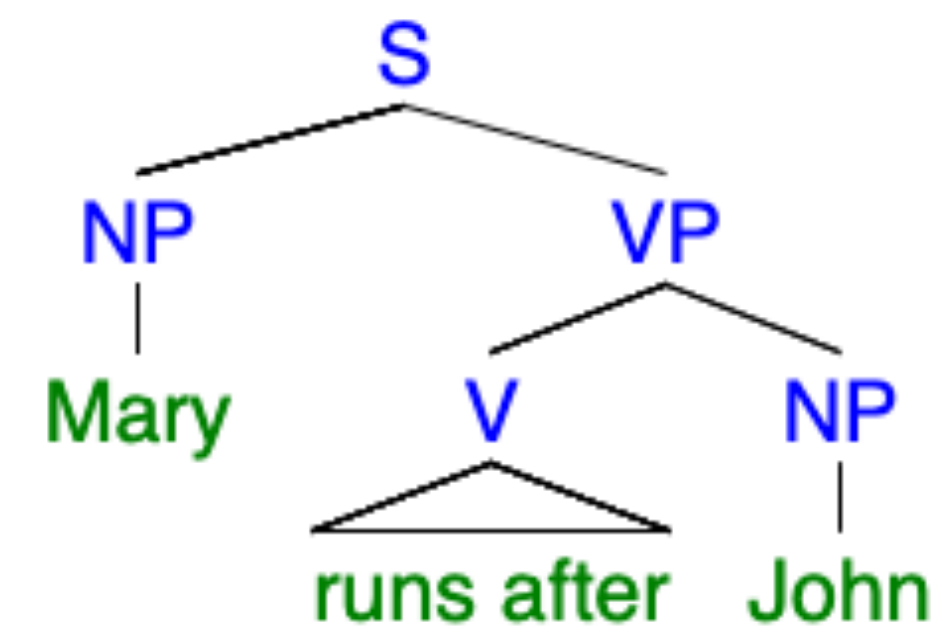


Linguistic data... why so special?

There exists a mapping between structure and meaning.



VS



Linguistic data... why so special?


Language is very different from any other data out there. Science still doesn't have an answer to:

- How do we *acquire/learn* our language?
- How do we *understand* our language?
- How do we *produce* our language?

So how are we supposed to find algorithms that do just that?

What we will cover:

LANGUAGE MODELS

- 
- **Bag-of-words** — topic models and naive classification
 - **N-gram language models (Markov models)** — Probabilistic models of language in context
 - **Hidden Markov models** — part-of-speech tagging
 - **Distributed representations** — vector semantics
 - **Recurrent neural language models, LSTMs** — language generation
 - **Transformers and masked language modeling** — language learning
 - **Encoder models** — semantic search and RAG (retrieval augmented generation)
 - **Encoder-decoder models** — text translation/text summarisation

[15 minute break]

Setting up your environment

1. Go to the course GitHub repo and follow getting started instructions

www.github.com/evaportelance/HEC-NLP

2. Open exercises/week 1 and start running code!