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

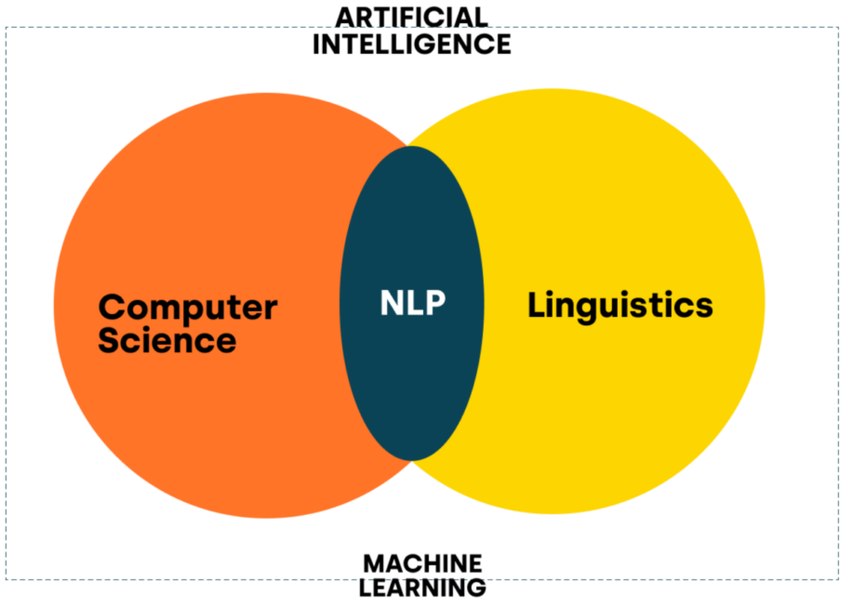
    1.4. Respect and expectations

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

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

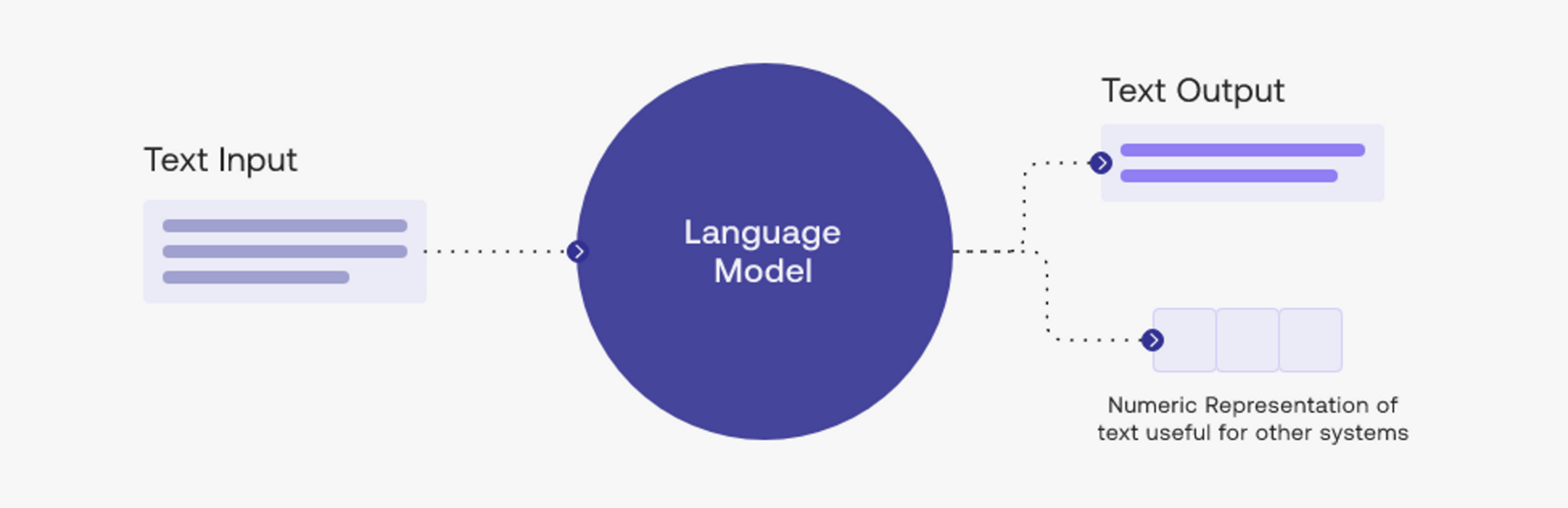
What is NLP?

- **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:

LANGUAGE MODELS



What we will cover:

LANGUAGE MODELS

- **N-gram language models** — Probabilistic models of language in context
- **Embeddings** — vector semantics
- **Recurrent neural language models, LSTMs** — neural networks for language
- **Transformers** — large language model (LLM) architectures
- **Encoder models** — scoring tasks
- **Encoder-decoder models** — generative tasks
- **Finetuning and in-context learning** — current foundation model paradigm

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!)

Evaluations

- **Two types of evaluations:** Assignments (50%) and Exams (50%)

- **Assignment 1 (25%)** - Due Feb 23 ..all applied coding problems
- **Midterm exam (25%)** - March XX ..written theoretical questions
- **Assignment 2 (25%)** - Due April 13 ..all applied coding problems
- **Final exam (25%)** - April XX ..written case study 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.

Generative AI use

This is a class about how generative AI works.

If you wish to, you may use existing generative AI for code generation on exercises and assignments. But do not change any existing code!

Note : for exams, you will only be allowed one page of notes, no laptops. There may be questions requiring you to write pseudo-code, so make sure you actually understand the code you submit using generative AI before using it (And obviously make sure it runs!!!!)

Now the fun part.

Linguistic data… why so special?

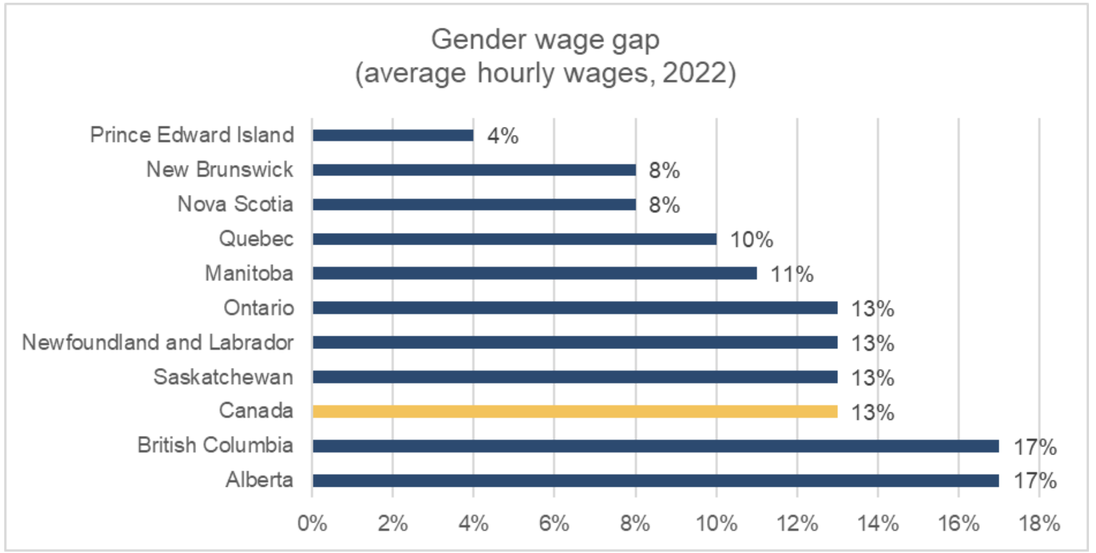
Mortgage Rates Hold Near 7%

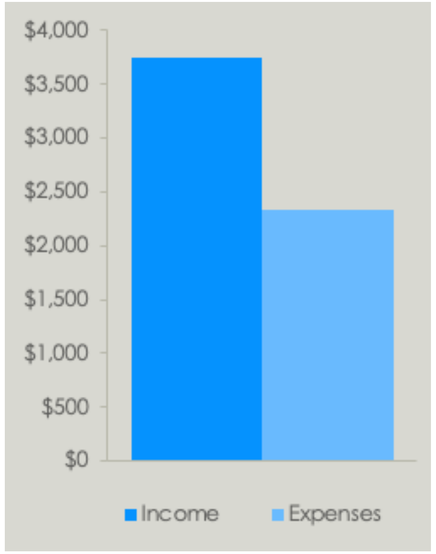
The average rate on a 30-year fixed mortgage edged lower to 6.78%. It was one basis point higher last week.

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





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

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!

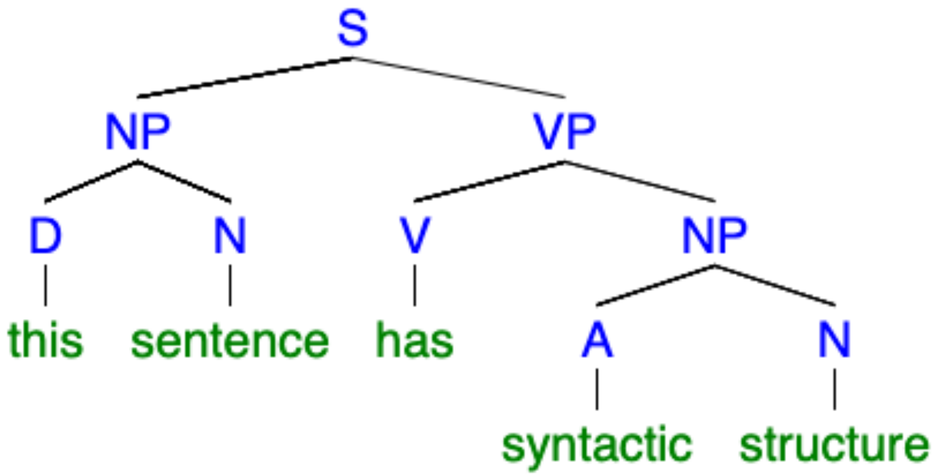
Linguistic data… why so special?

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.**



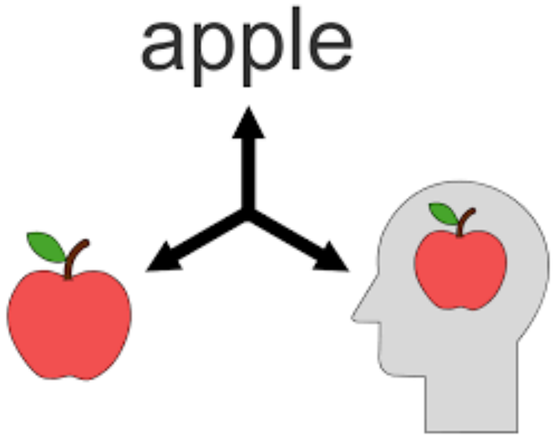
Linguistic data… why so special?

Language is a symbolic data type!

Linguistic data… why so special?

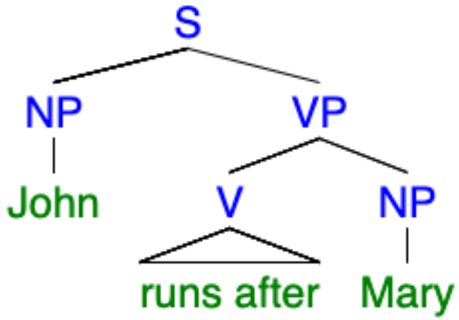
The meaning of a sentence is called **semantics.**

John runs after Mary.

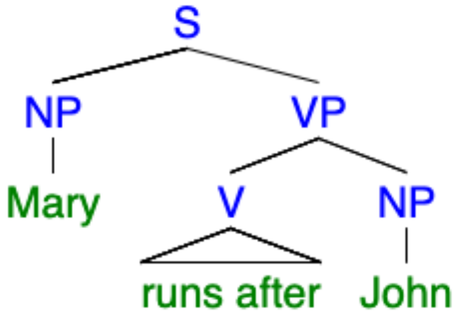


Linguistic data… why so special?

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:

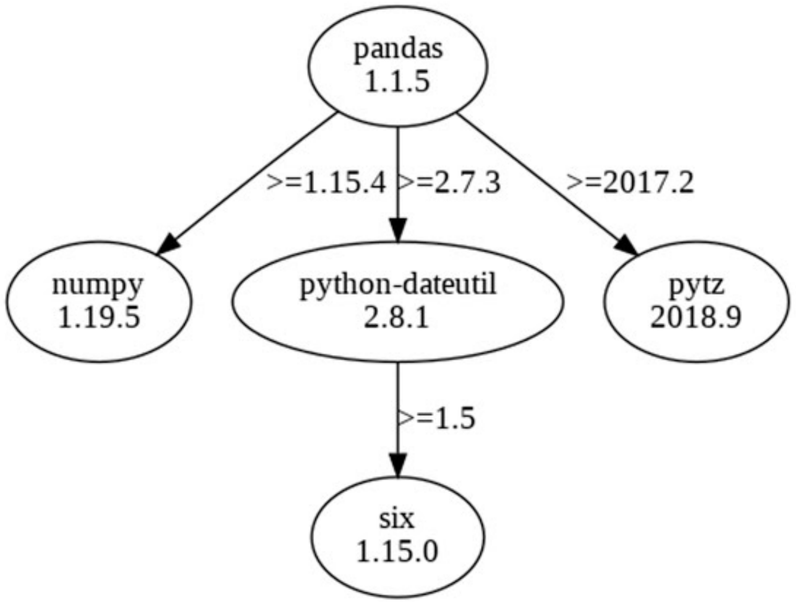
LANGUAGE MODELS

- **N-gram language models** — Probabilistic models of language in context
- **Embeddings** — vector semantics
- **Recurrent neural language models, LSTMs** — neural networks for language
- **Transformers** — large language model (LLM) architectures
- **Encoder models** — scoring tasks
- **Encoder-decoder models** — generative tasks
- **Finetuning and in-context learning** — current foundation model paradigm

[15 minute break]

What is a package manager?

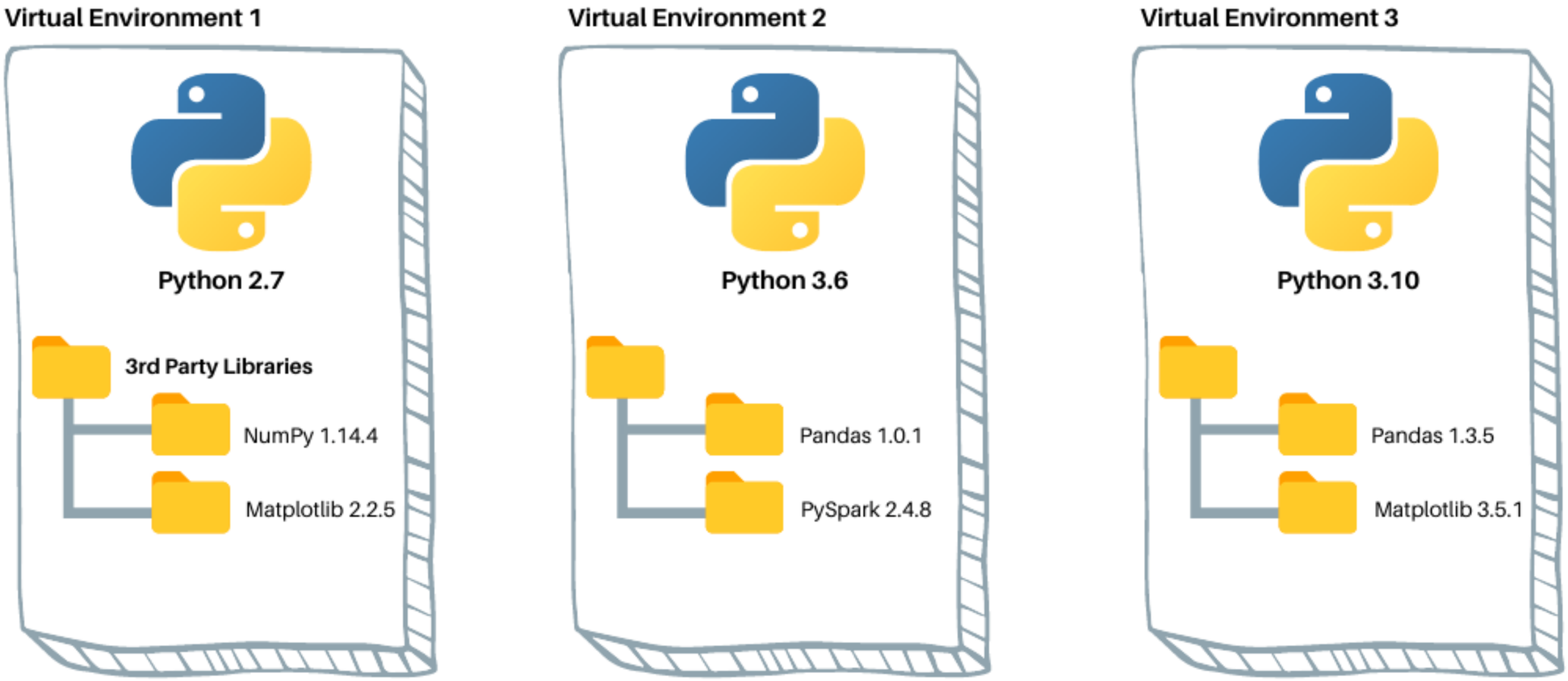
- A tool that automates the process of installing, upgrading, configuring, and removing Python software packages and their dependencies.
- Packages are external libraries, modules, or frameworks that extend Python's core functionality, so you can include pre-written code into projects without writing everything from scratch.
- Package managers are essential for dealing with complex package dependencies and conflicts.
- Eg. `pip` or `pip3`



What is a virtual environment?

- A Python virtual environment is a self-contained directory that contains project-specific dependencies (packages and their versions)
- It is isolated from other project environments or the global Python installation on your system, i.e. a package installed in one environment will not affect another.
- It has its own dedicated python interpreter, i.e. you can have a different version of Python or `pip` than the default one on your system.
- They help avoid dependency conflicts and encourage reproducibility.
- E.g. `venv` or `virtualenv`

What is a virtual environment?



What is an environment manager?

- An environment/project manager combines package management with virtual environments, allowing you to easily create different environments and package dependencies for different projects.
- As a tool, it keeps track of all your environment and helps with keeping projects separate and clean.
- Egs: Anaconda, Poetry, uv.

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!