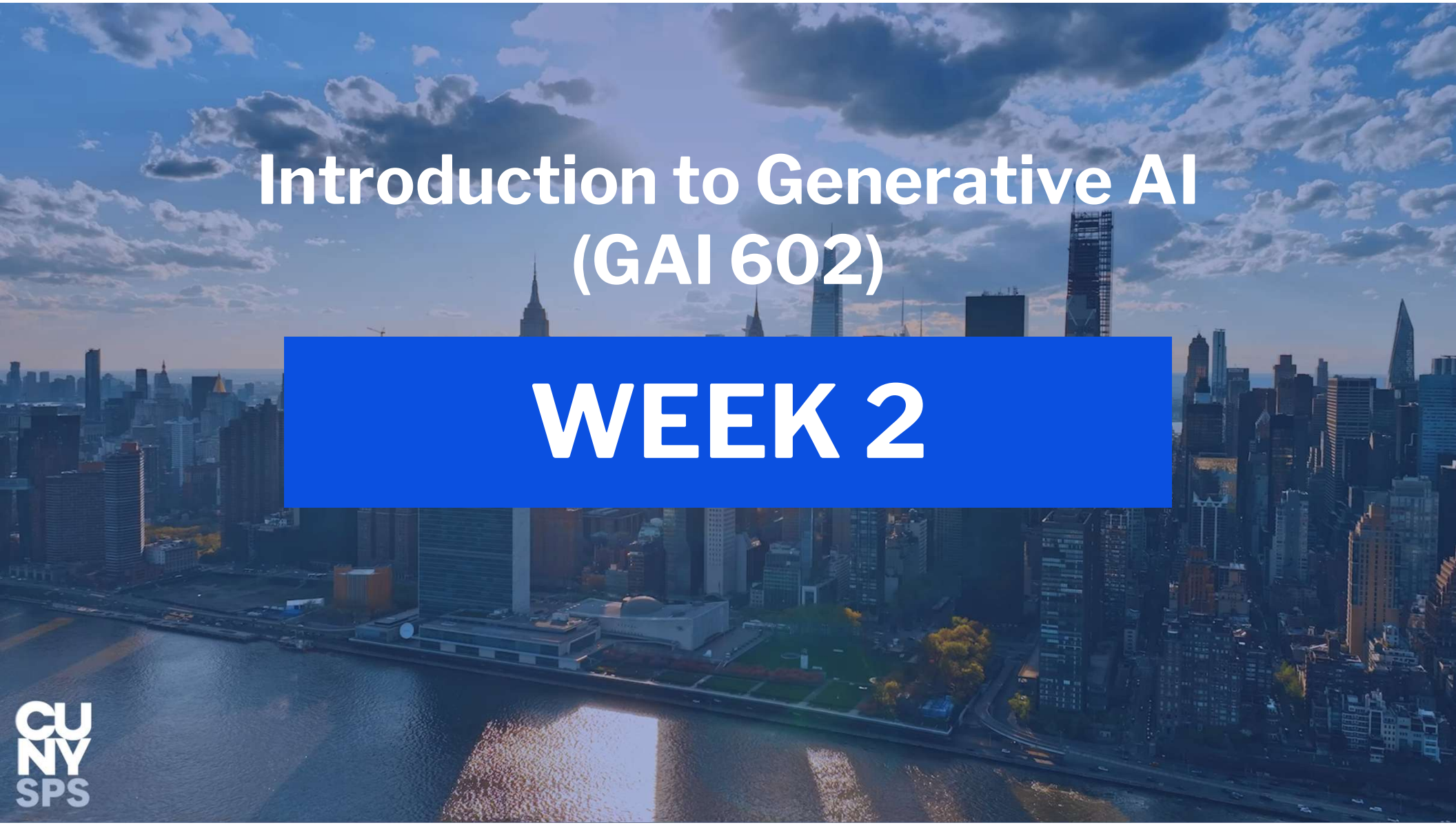
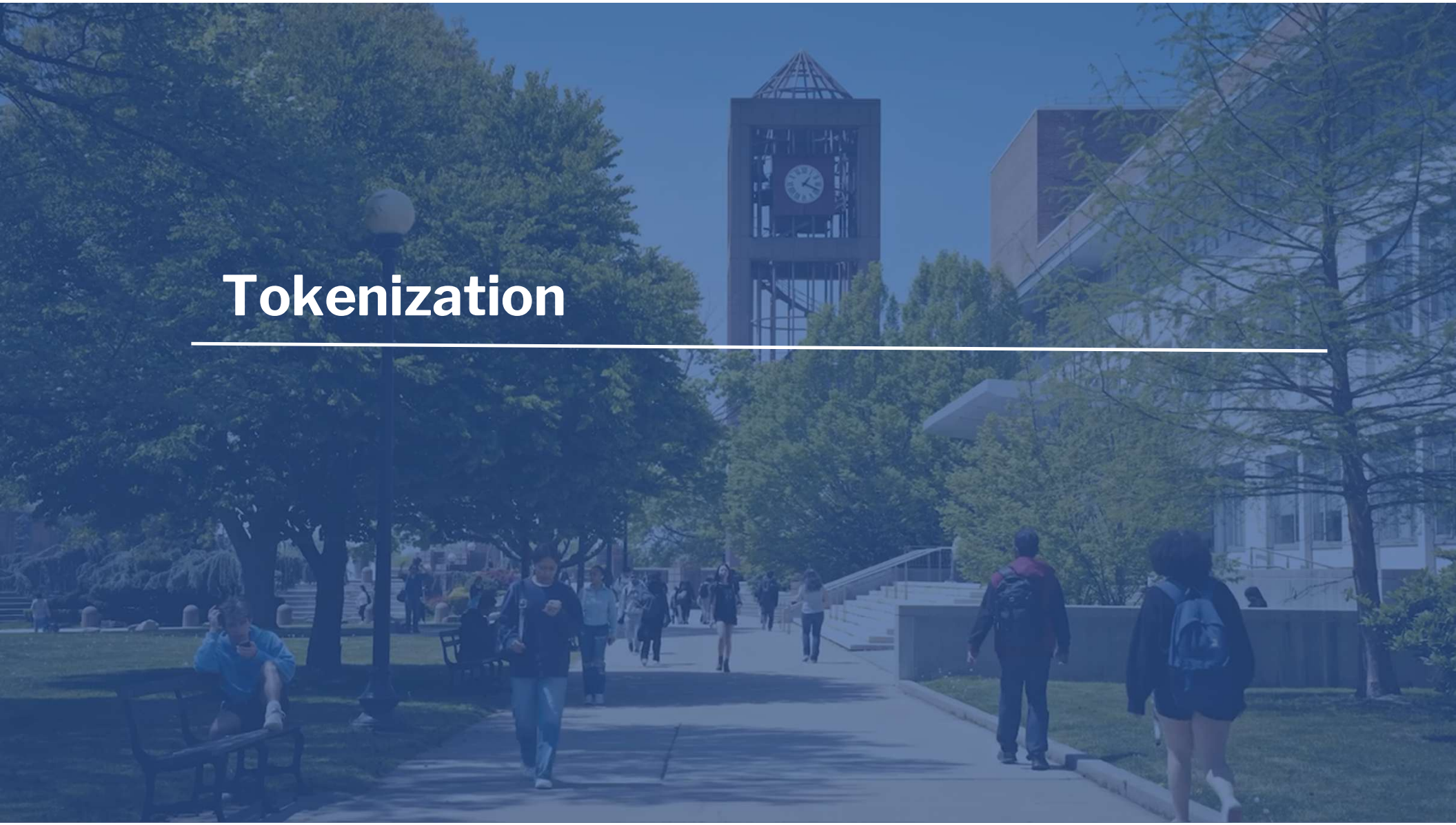
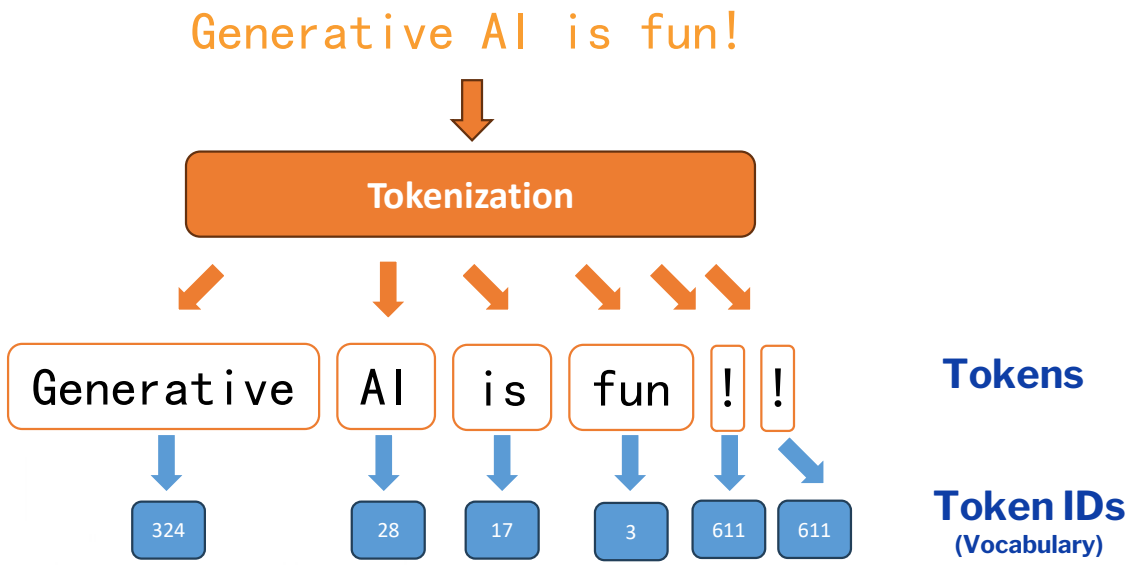

Introduction to Generative AI
(GAI 602)
WEEK 2
CU
NY
SPS


Tokenization

Tokenization

Tokenization is required for:

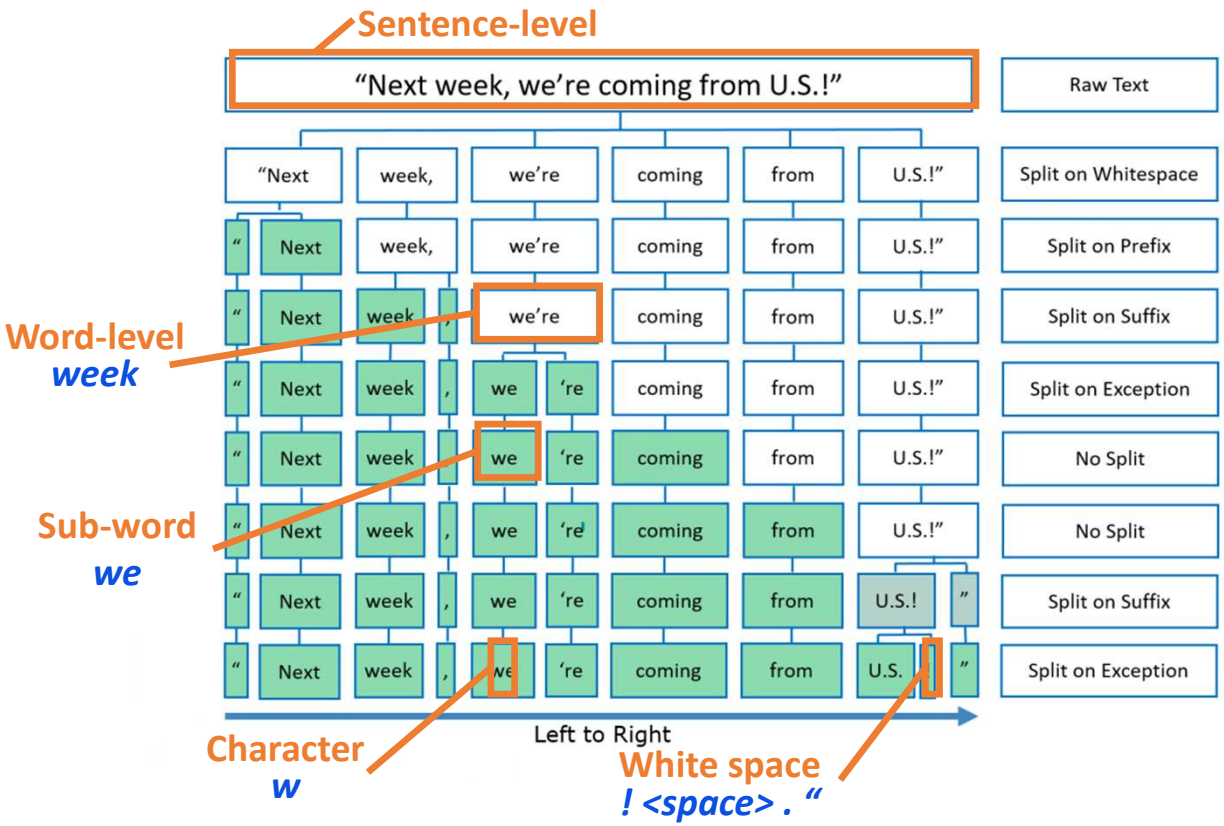
- Numerical representation (computers need numbers)
- Structure unstructured data (structure text)
- Surface meaning (text is messy)
- Efficien



Tokenization Options

1. **Sentence-level**
2. **Whitespace**
3. **Word**
4. **Sub-word (WordPiece)**
5. **Character-level**
6. **Rule-based**

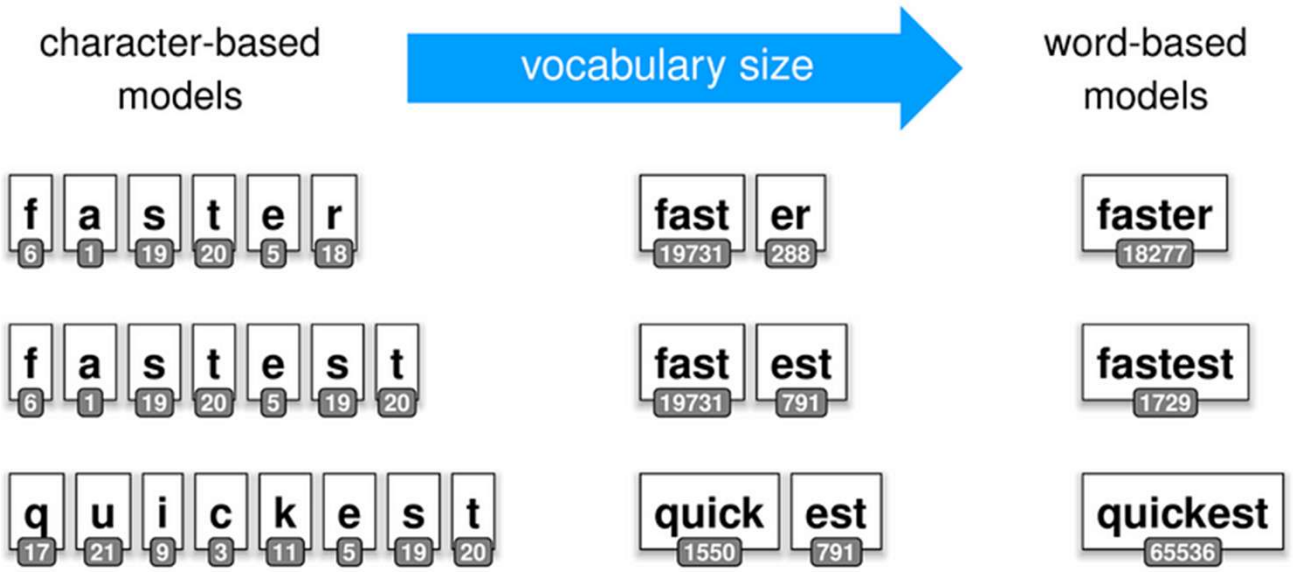
... and many more



Vocabulary Size

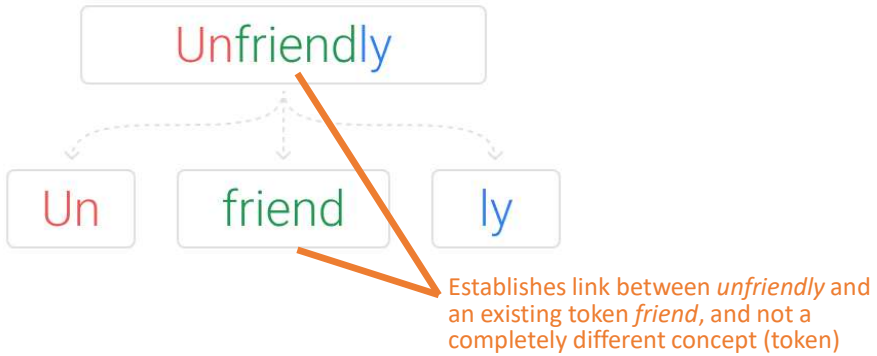
A vocabulary is the set of all tokens (characters, words, sentences, etc.) that are used to represent and analyze text.

There is a trade-off between token granularity (retention of meaning) and vocabulary size.

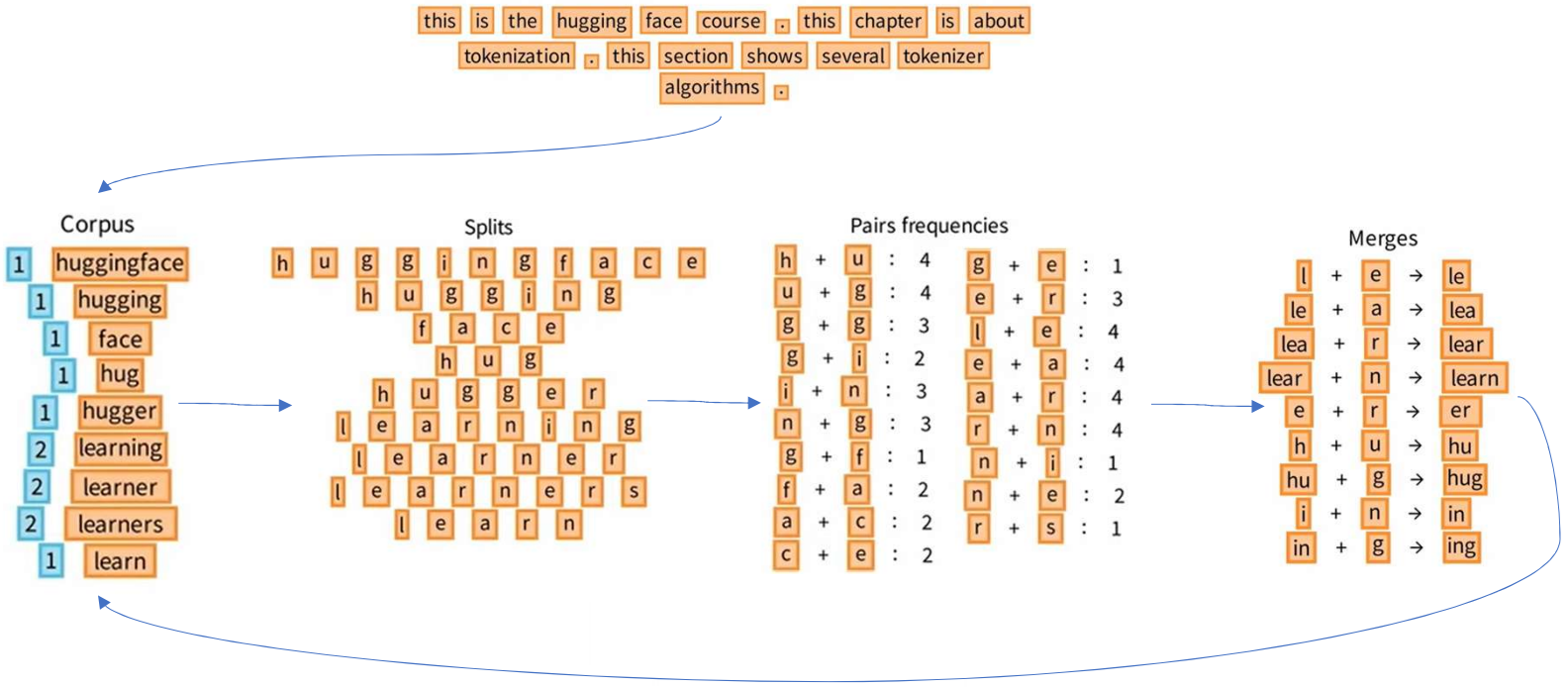


Sub-word

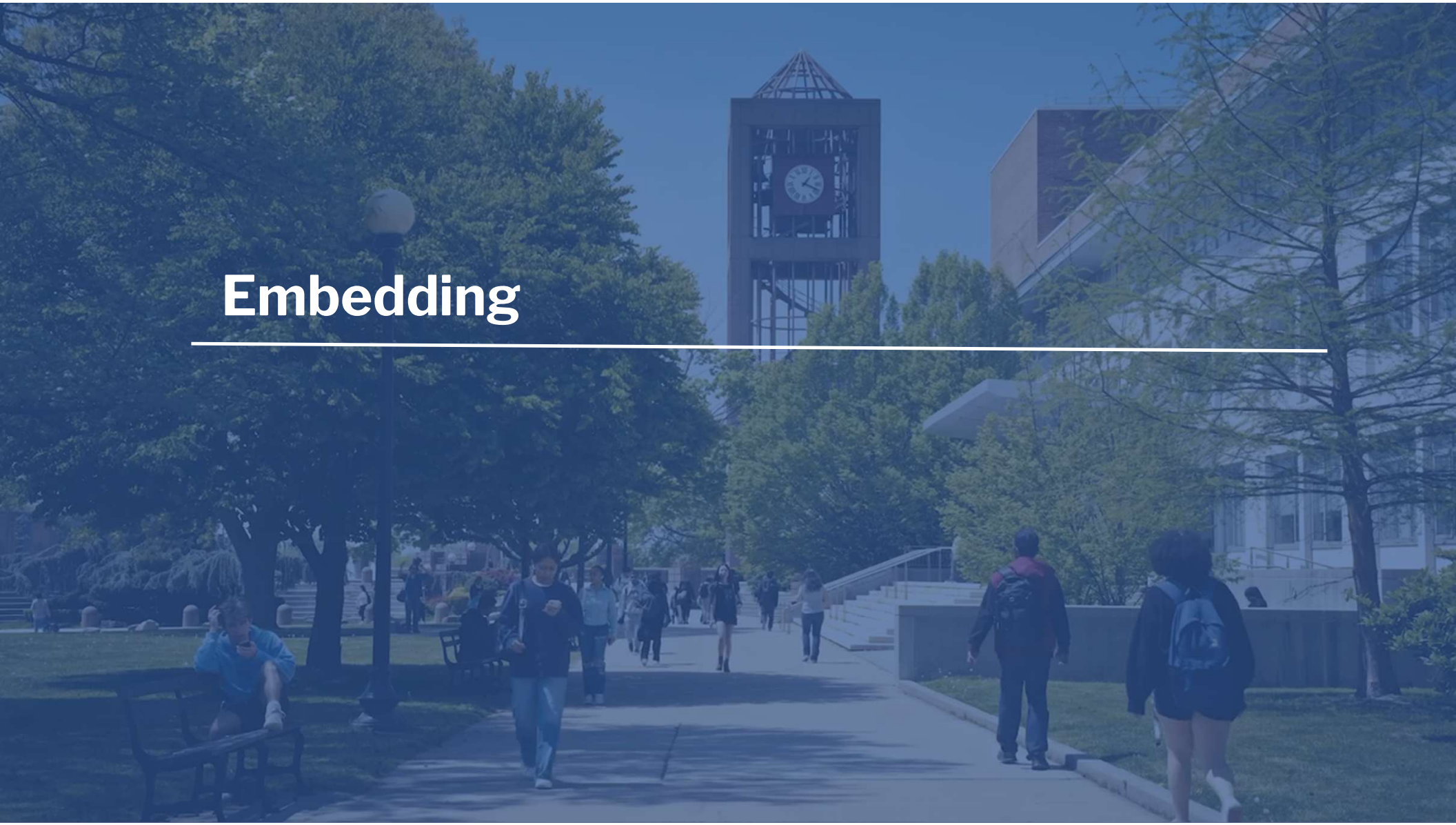
Sub-word is most popular, best balance of vocabulary size and retention of meaning



Byte-pair Encoding (BPE)



Source: Huggingface Byte Pair Encoding Tokenization - https://www.youtube.com/watch?v=HEikzVL-lZU


Embedding

Word Meaning

A word only has meaning in the context of other words

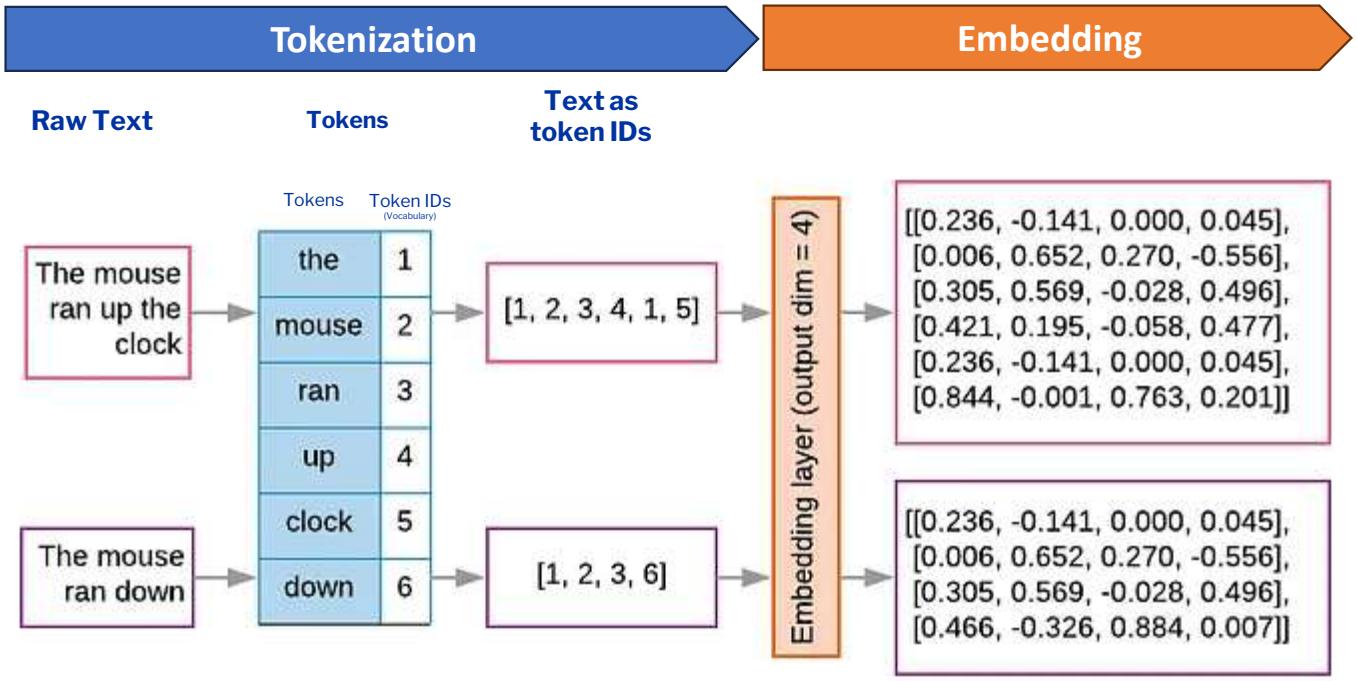
Words may have many meanings (polysemy).
The meaning of a word depends on its context.

An Example: ***bow***

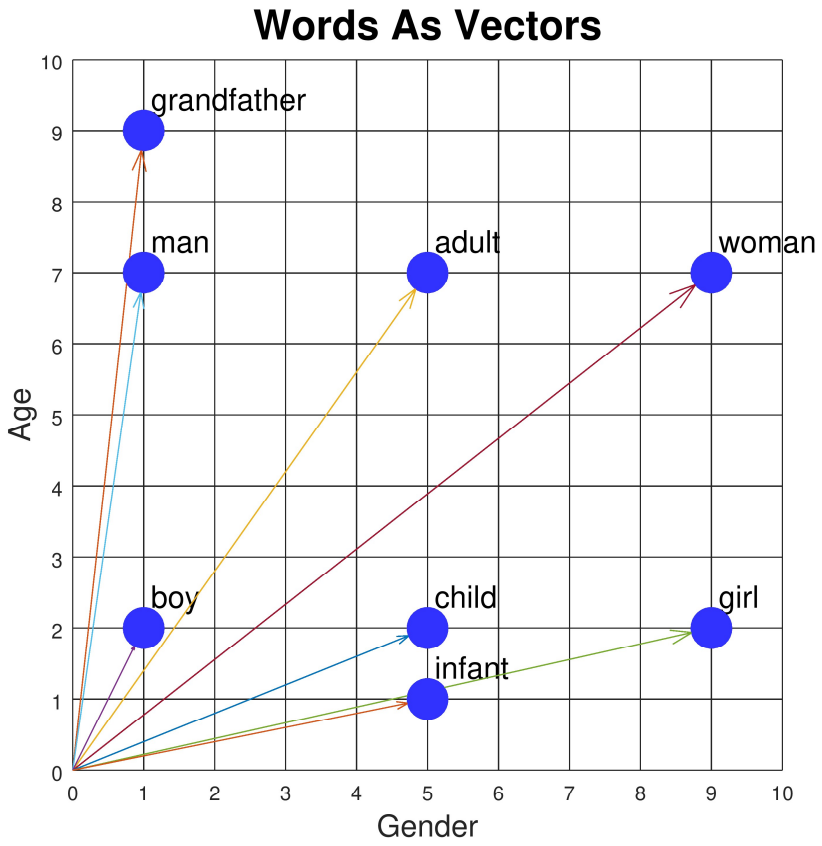
- the front of a ship
- to bend forward in respect
- a weapon that shoots arrows
- a decorative knot
- used to play a violin

All of these have the ***same token*** (and token ID)
but they have ***different meanings***

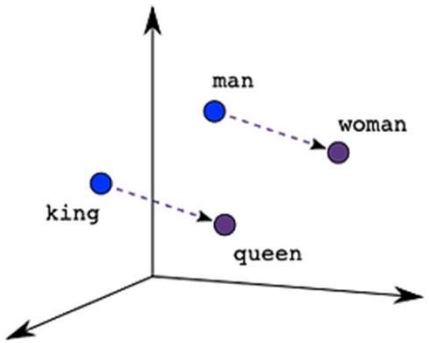
Embedding


Tokenization
Embedding
Raw Text
Tokens
Text as token IDs
Tokens
Token IDs (Vocabulary)
The mouse ran up the clock
the 1
mouse 2
ran 3
up 4
clock 5
down 6
[1, 2, 3, 4, 1, 5]
The mouse ran down
[1, 2, 3, 6]
Embedding layer (output dim = 4)
[[0.236, -0.141, 0.000, 0.045],
[0.006, 0.652, 0.270, -0.556],
[0.305, 0.569, -0.028, 0.496],
[0.421, 0.195, -0.058, 0.477],
[0.236, -0.141, 0.000, 0.045],
[0.844, -0.001, 0.763, 0.201]]
[[0.236, -0.141, 0.000, 0.045],
[0.006, 0.652, 0.270, -0.556],
[0.305, 0.569, -0.028, 0.496],
[0.466, -0.326, 0.884, 0.007]]

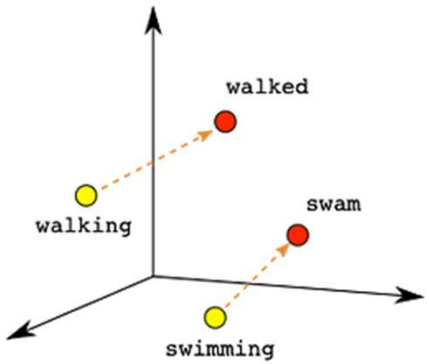
Word2Vec


Words As Vectors
grandfather
man
adult
woman
boy
child
girl
infant
Age
Gender
0
1
2
3
4
5
6
7
8
9
10

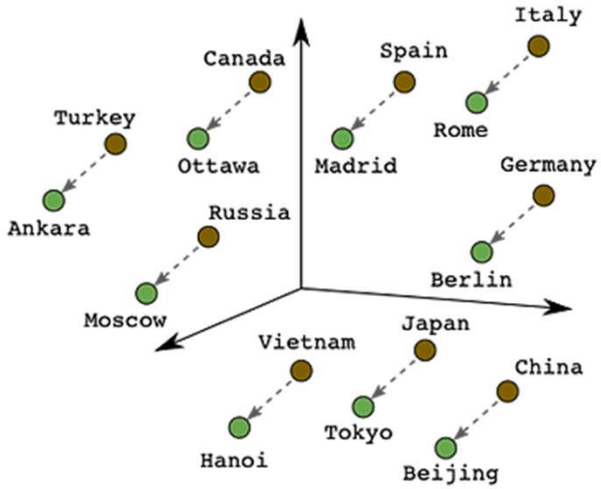
Embedding



Male-Female

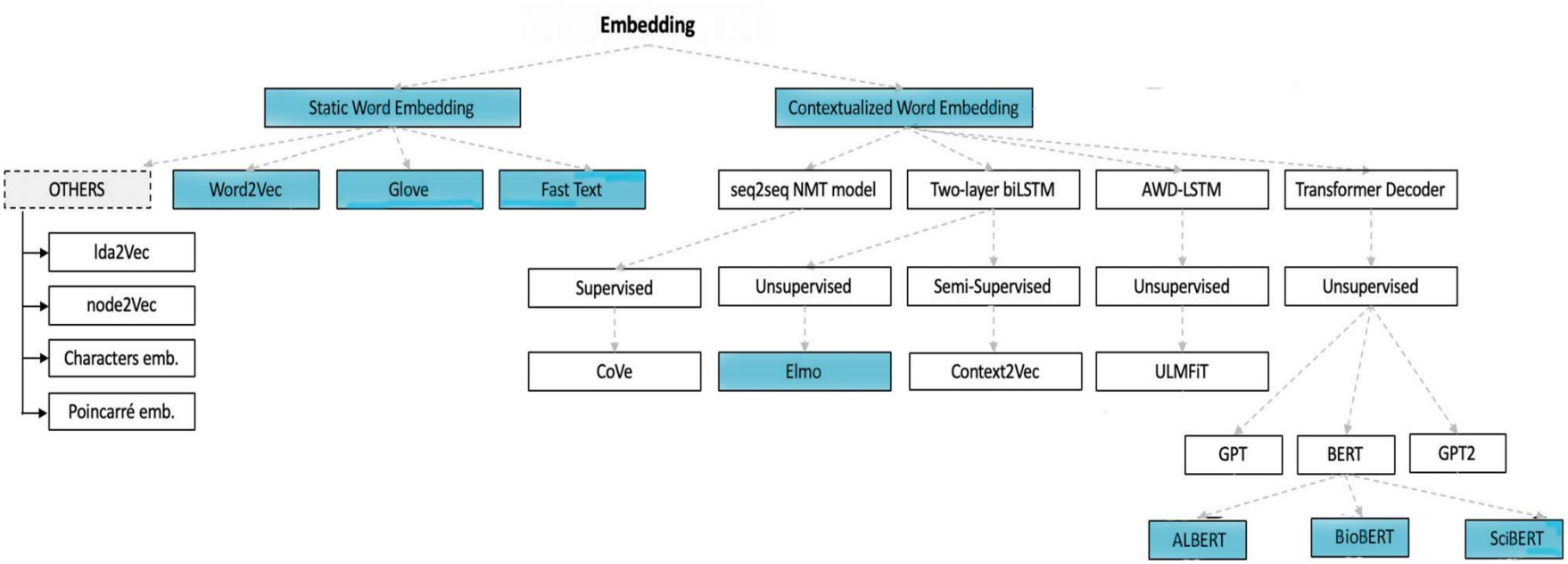


Verb Tense



Country-Capital

Embedding Techniques



Source: Mohiuddin Md Abdul Qudar and Vijay Mago. 2020. A Survey on Language Models.


Natural Language Processing (NLP)

Natural Language Processing Tasks

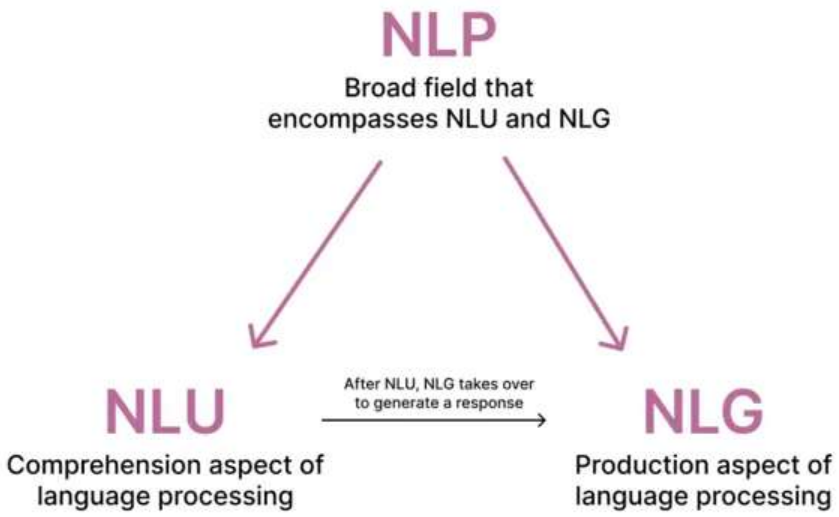
NLP is the process through which **AI is taught to understand the rules and syntax of language**, programmed to develop complex algorithms to represent those rules, and then made to **use those algorithms to carry out specific tasks** like these.

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation	Clause boundary detection	Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

Source: Mobidev

Natural Language Understanding & Generation

The difference between **Natural Language Processing (NLP), Natural Language Understanding (NLU), and Natural Language Generation (NLG)**

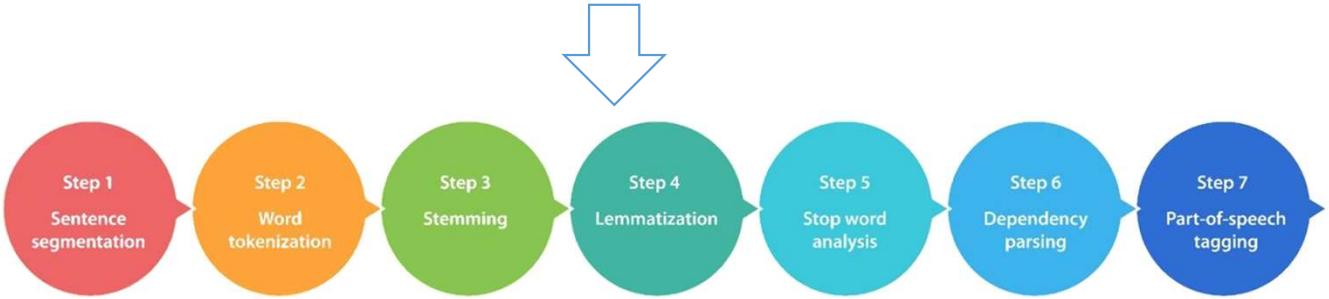


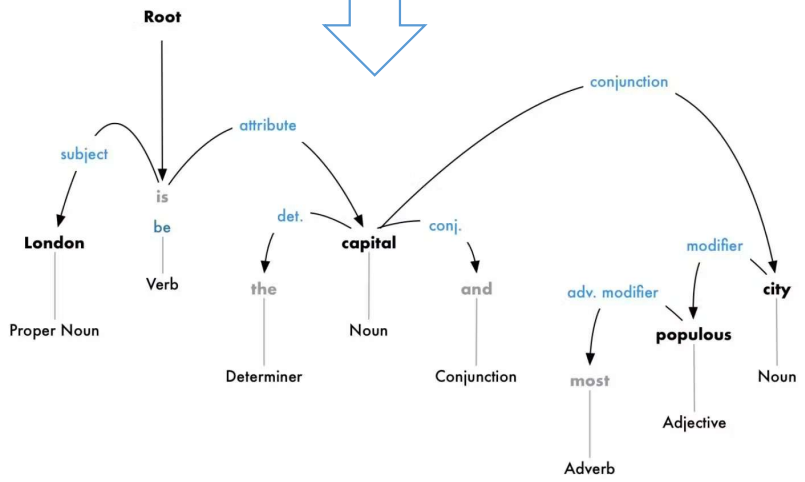
NLU	NLP	NLG
It is a narrow concept.	It is a broader concept.	It is a limited concept.
If we only talk about an understanding text, then it is enough.	But if we want more than understanding, such as decision-making, then it comes into play.	It generates a human-like manner text based on the structured data.
It is a subset of NLP.	It is a combination of it and NLG for conversational Artificial Intelligence problems.	It is a subset of NLP.
It is not necessarily that what is written or said is meant to be the same. There can be flaws and mistakes. It ensures that it will infer correct intent and meaning even if data is spoken and written with some errors. It is the ability to understand the text.	But if we talk about NLP, it is about how the machine processes the given data, such as making decisions, taking action, and responding to the system. It contains the whole End-to-end process. It doesn't need to have it every time.	It generates structured data, but the generated text is not necessarily easy for humans to understand. Thus, NLG ensures that it will be human-understandable.
It reads data and converts it to structured data.	It converts unstructured data to structured data.	NLG writes structured data.

Source: Botpress & Xenonstack

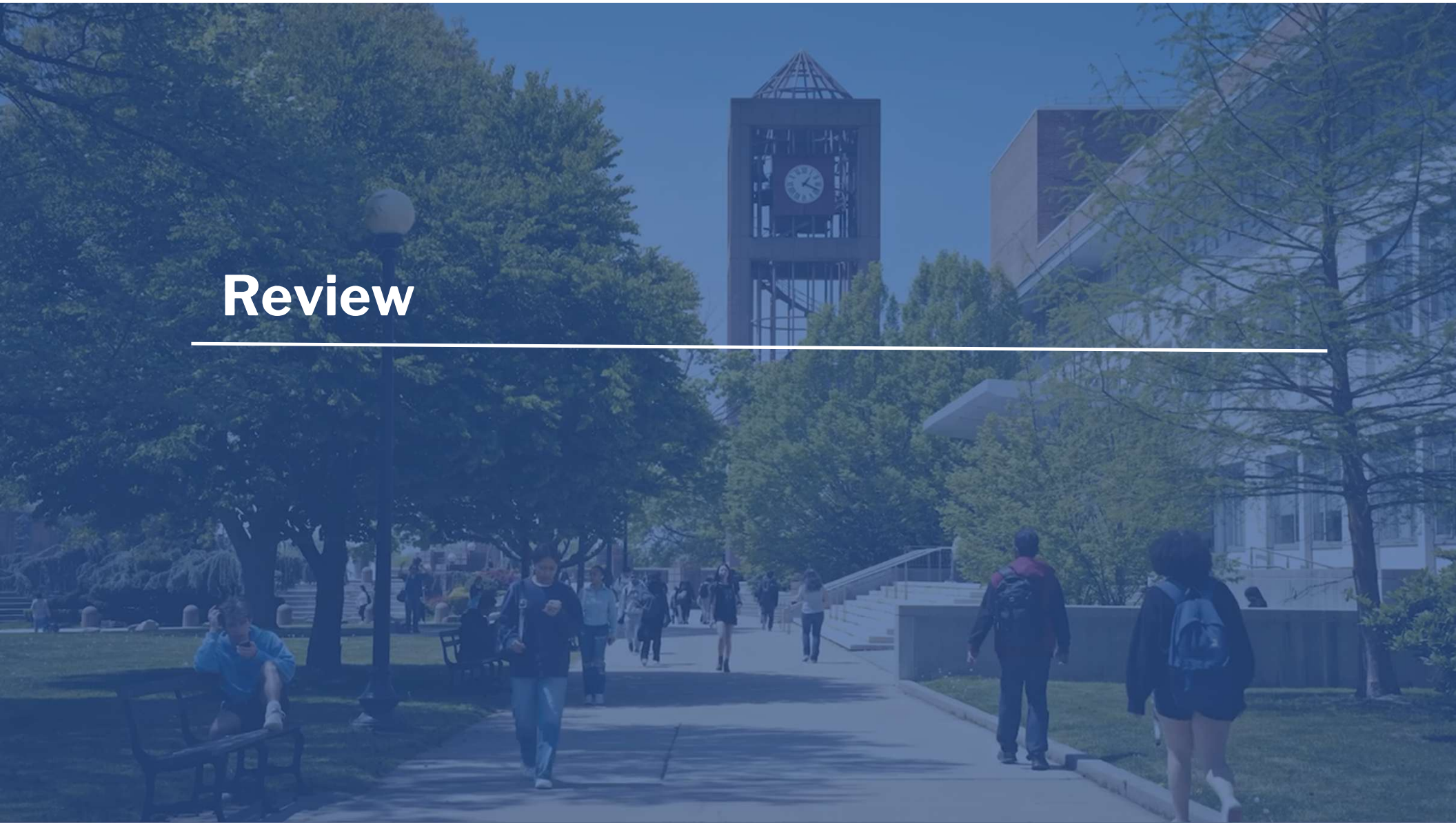
NLP Pipeline

"London is the capital and most populous city of England and the United Kingdom."





Source: Turing



Review

This week we covered

Course Objectives/Topics (Objectives

1. What is tokenization and why it matters
2. Bag-of-words vs. modern embedding techniques
3. Sub-word tokenization methods (e.g., byte pair encoding, Word Piece)
4. Language representation and vocabulary construction

Concepts

Tokenization

Bag-of-words


CUNY
School of
Professional Studies