CMPT 413/713: Natural Language Processing

Conclusion

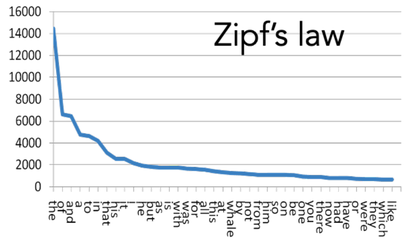
Spring 2024

2024-04-10

What is natural language?

- Way for humans to communicate with each other
  - Use it to express feelings, transmit information, and store knowledge
- Discrete, combinations of symbols
- Learnable by all humans

Key challenges of understanding language



- Ambiguity (at multiple levels)

I saw her duck



- Scale
- Sparsity (Long tail)
- Variation
- Expressivity
- Unmodeled Variables (Implicit knowledge / Context)
- Unknown representation (representation)

What has this course covered?

Representations

- Embeddings
- Compositional
- Structured

Methods

- Language Models
- Statistical Methods
- Neural Models
- Dynamic programming

LLMs

- Emergent capabilities
- Prompting
- Few-shot / In-context learning
- Efficient fine-tuning

Applications and tasks

- Machine translation, text generation, question answering, grounding

What has this course covered?

Representations

- Embeddings
- Compositional
- Structured

Methods

- Language Models
- Statistical Methods
- Neural Models
- Dynamic programming

LLMs

- Emergent capabilities
- Prompting
- Few-shot / In-context learning
- Efficient fine-tuning

Applications and tasks

- Machine translation, text generation, question answering, grounding

Key Questions

- How to **represent** language?
- How to **compose** meaning?
- How do we **parse** the “surface form” of language to a representation that a computer can process?
- How do we take an encoded representation and **generate** text from it?
- What are the key methods / algorithms for **learning** these representations?

Methods we covered

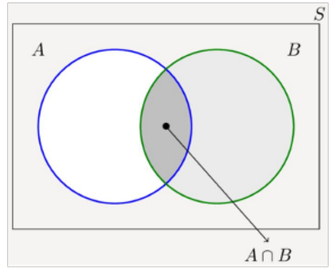
Rule-based methods

```
(0 YOU 0 ME)                                [pattern]
→
(WHAT MAKES YOU THINK I 3 YOU) [transform]
```

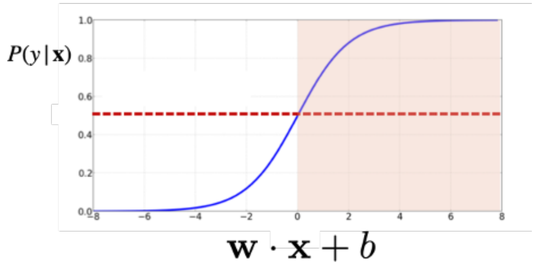
```
You hate me
WHAT MAKES YOU THINK I HATE YOU
```

Use when you have no data
Good way to have something to bootstrap from

Statistical methods



Naive Bayes

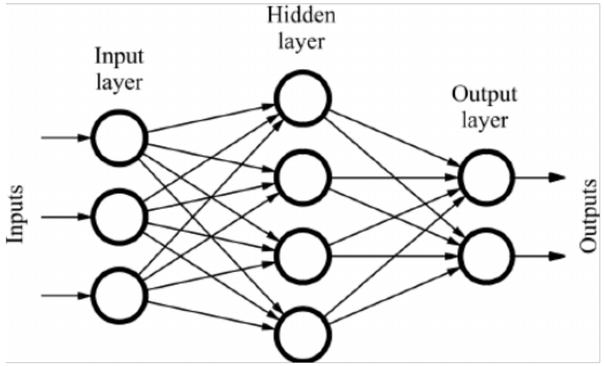


Logistic regression

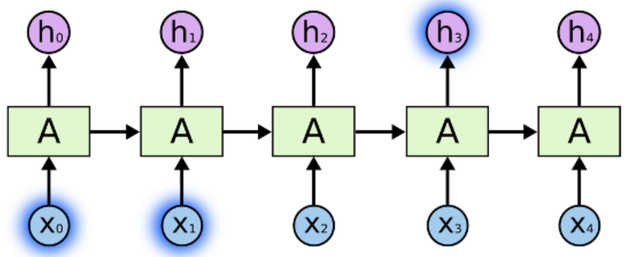
Simple baselines for text classification
Solid understanding helps understand NNs

Neural networks for NLP

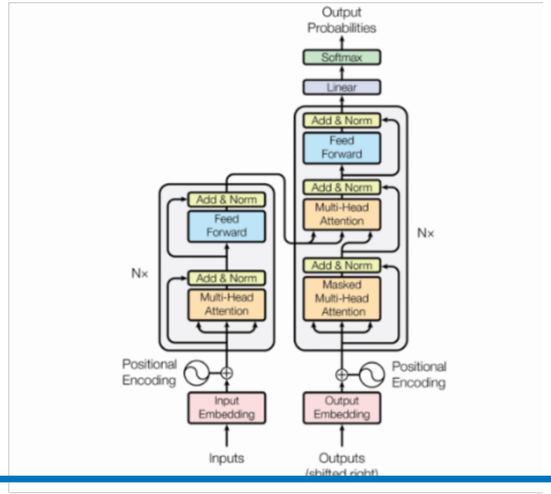
Feed-forward NNs



Recurrent networks (RNNs)

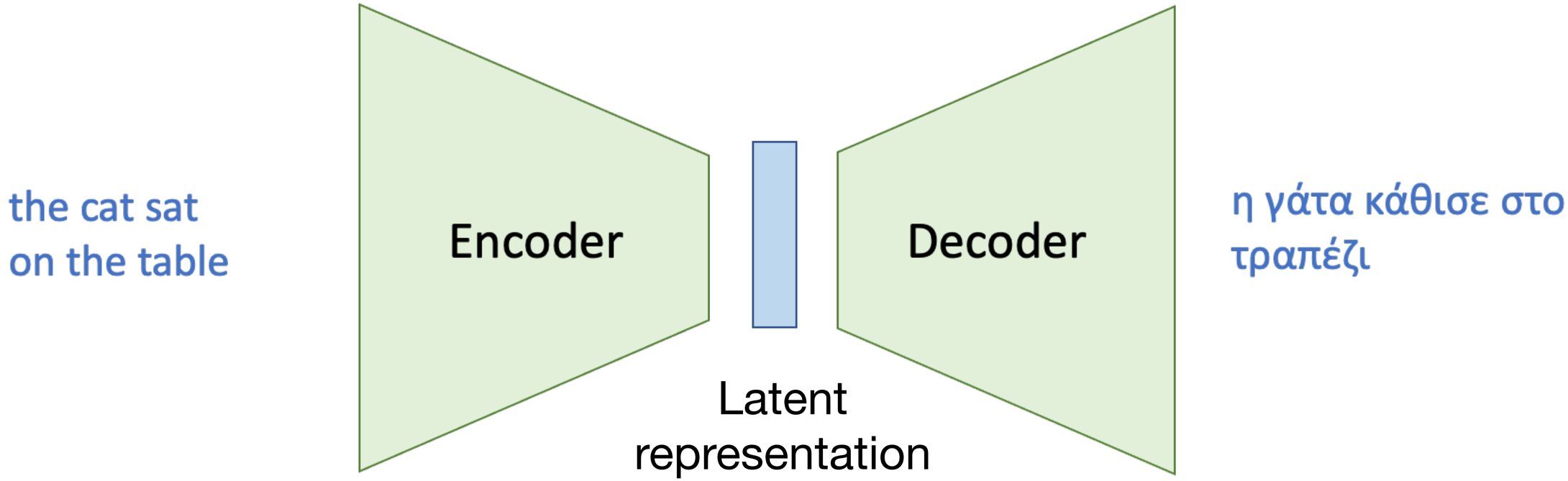


Transformers



Data hungry, compute intensive →Use pretrained models and then fine-tune on limited data

Encoder-Decoder Model



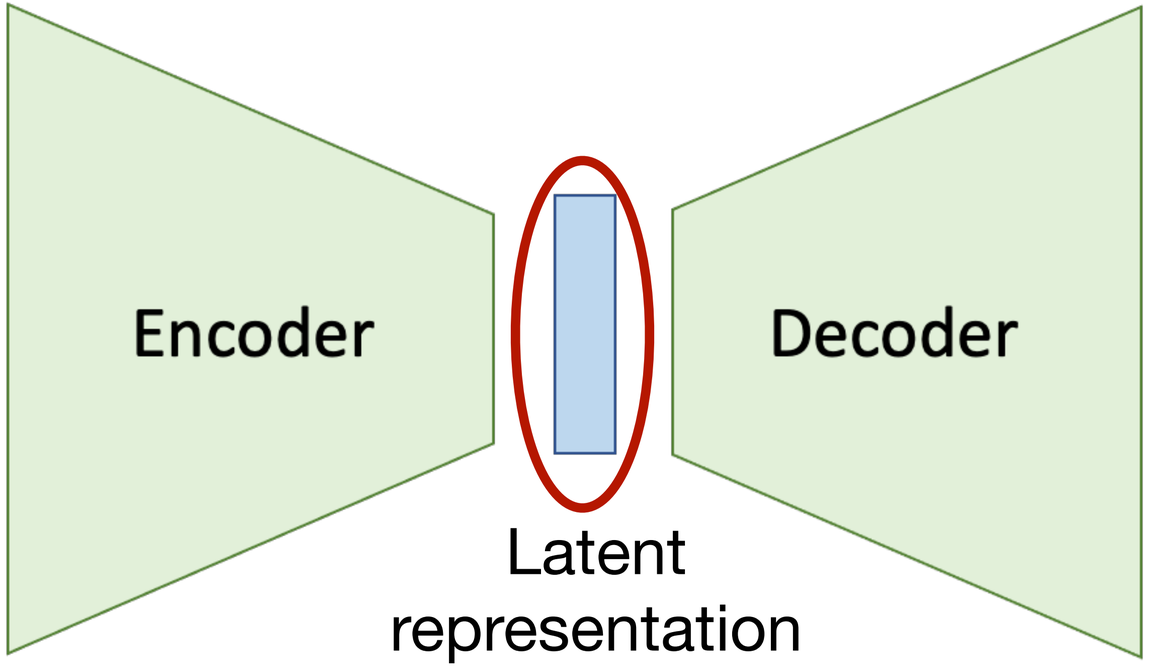
Understanding what is said
(encoding, parsing, feature extraction)

Deciding what to say
(decoding, generating)

How to represent language?

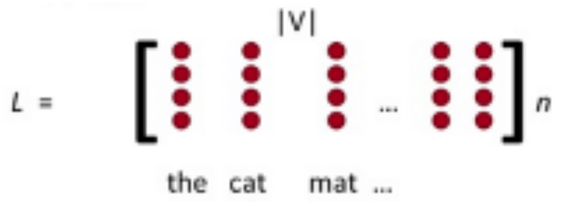
the cat sat
on the table

Lambda expression:
((λ x . (λ y . x)) (λ z . z)) (λ a . a)

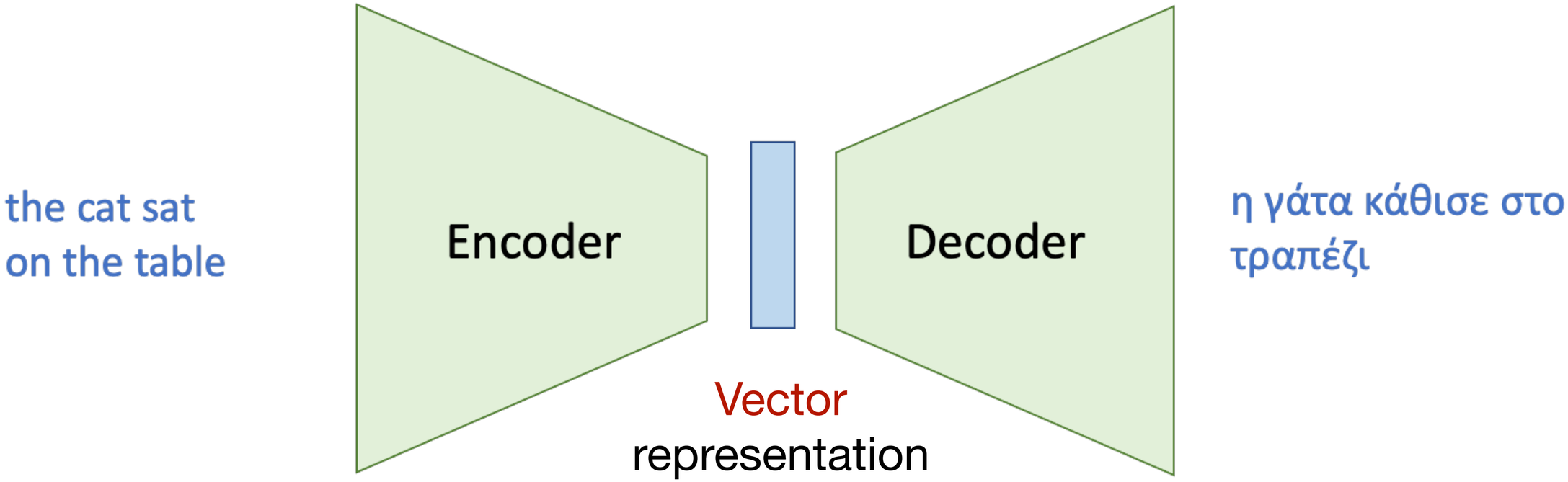


η γάτα κάθισε στο
τραπέζι

What kind of representation should go here?
Structured? Unstructured? Grounded?



Vector based representation

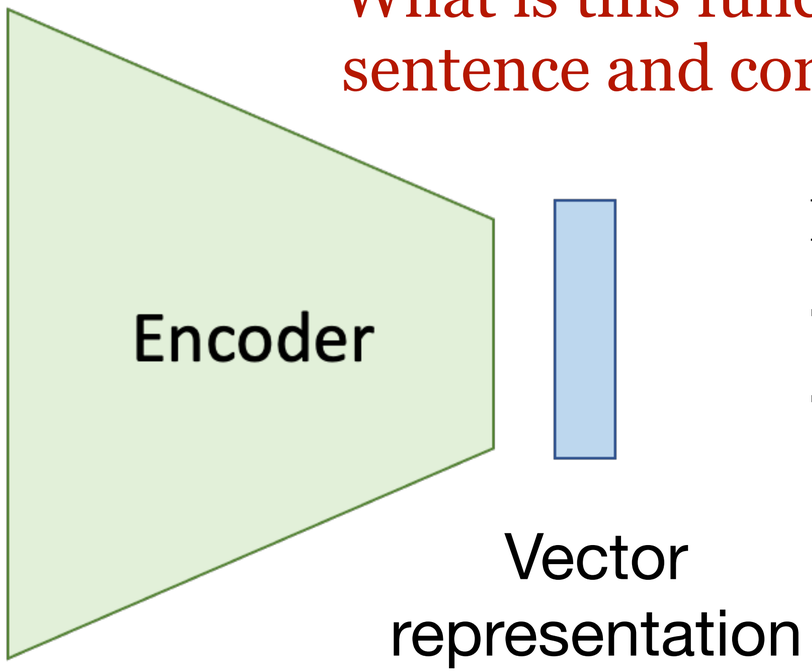


Common choice: flat unstructured vector of real numbers

- easy to work with
- well defined mathematical operations
- can be used to measure if two things are similar or not

How to encode language?

What is this function that takes the sentence and converts it into a vector?



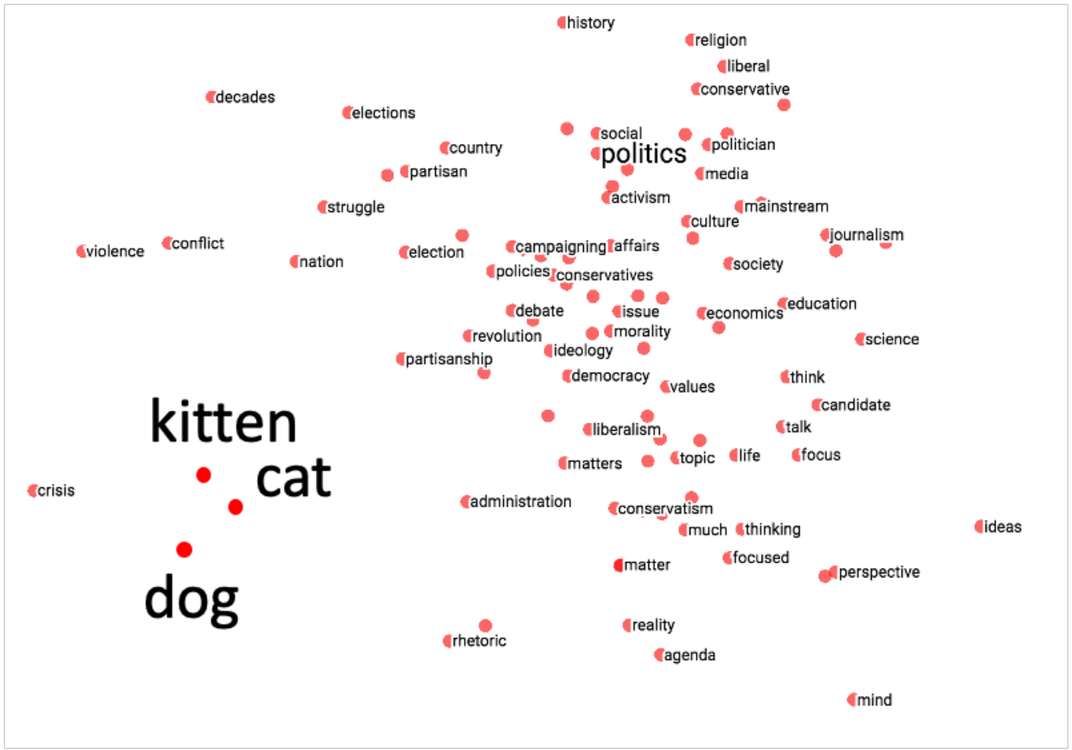
Properties we want

- Be able to represent a single word
- Be able to represent the meaning of a sentence as a composition of the meaning of each word

Understanding what is said
(encoding, parsing, feature extraction)

How to learn representations for words?

"cat" "dog"



Use context to represent words

…government debt problems turning into	*banking*	*crises as happened in 2009…*
…saying that Europe needs unified	*banking*	*regulation to replace the hodgepodge…*
…India has just given its	*banking*	*system a shot in the arm…*

These context words will represent ***banking***

how to learn a dense vector representation that can be used to predict words from their context

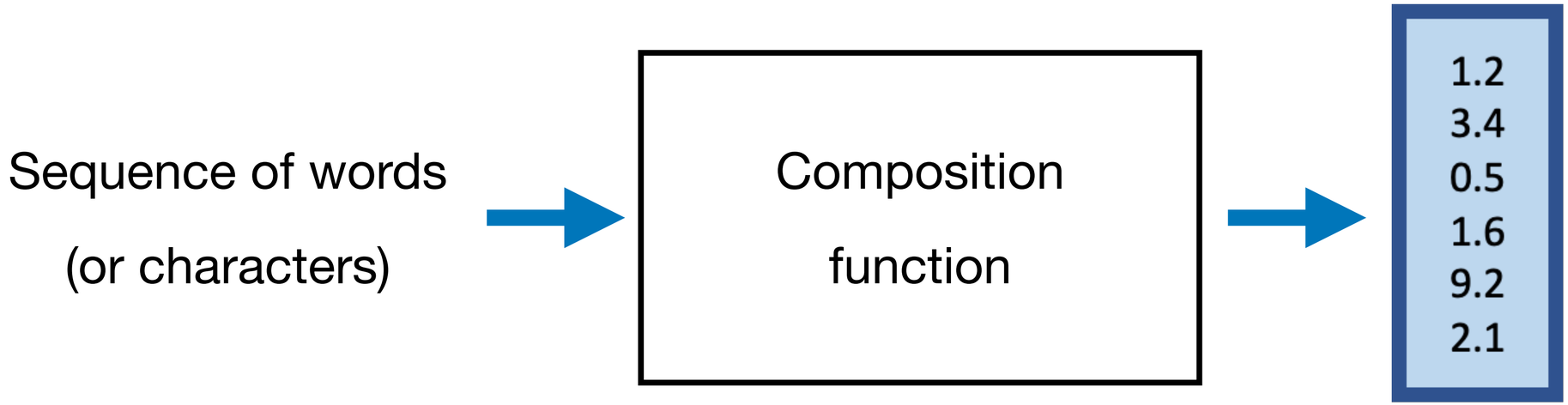
- Predict what goes next

  The cat sat on the _____ LM

- Fill in the blank

  A bottle of _____ is on the table. MLM

How to compose language?

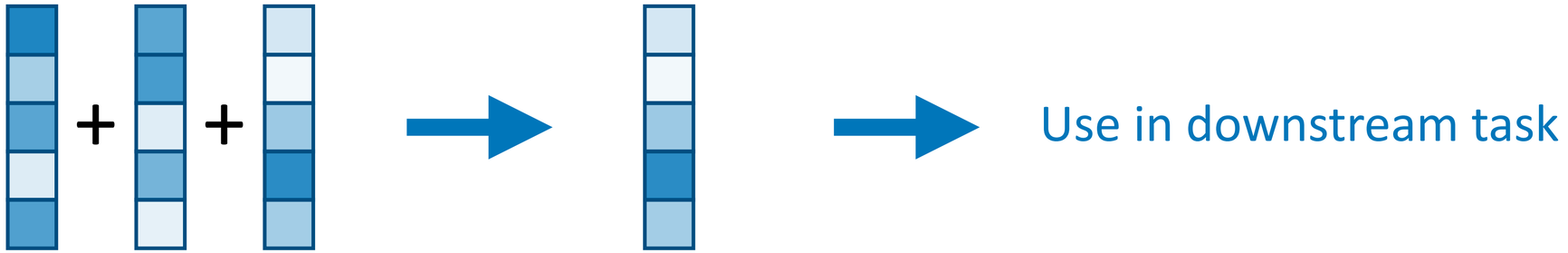


Simplest: take weighted average

Example:

Movie was terrible

Sentence representation



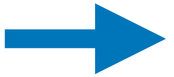
Use in downstream task

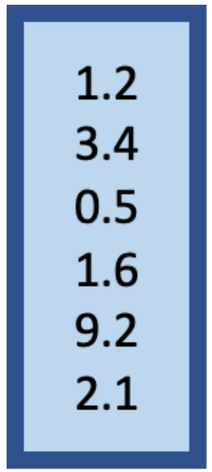
How to compose language?

Sequence of words

(or characters)

Composition function





Other ways to obtain

Sentence representation

Bag of Words (BOW)

Window BoW (bag of ngrams)

RNN

Transformers

CNN

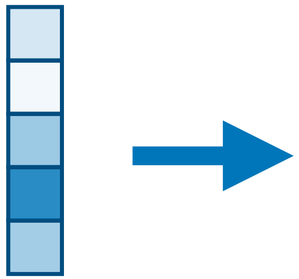
GraphNN

Throughout the course

- started with simple models and then covered more complex models

How to use the learned representation?

What do we do with this vector representation after we get it?



Use in downstream task

Vector representation

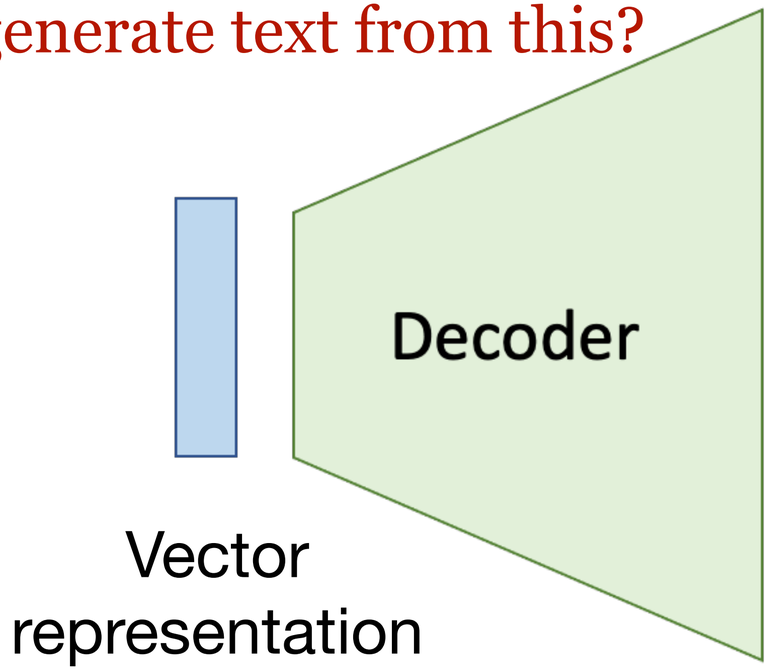
Do prediction: classify as spam/not spam

Use to generate text (also known as **decoding**)

Can also be used to generate other stuff (like images!)

How to decode language?

How do we generate text from this?



Auto-regressive text-generation

- Predict one word at time
- Big classification problem
- Can predict word from vocabulary or copy from somewhat else

Different training strategies

- Maximum likelihood estimation (cross-entropy loss for classification)

Different decoding strategies

- Greedy, Beam search, Sampling

Deciding what to say
(decoding, generating)

What has this course covered?

Representations

- Embeddings
- Compositional
- Structured

Methods

- Language Models
- Statistical Methods
- Neural Models
- Dynamic programming

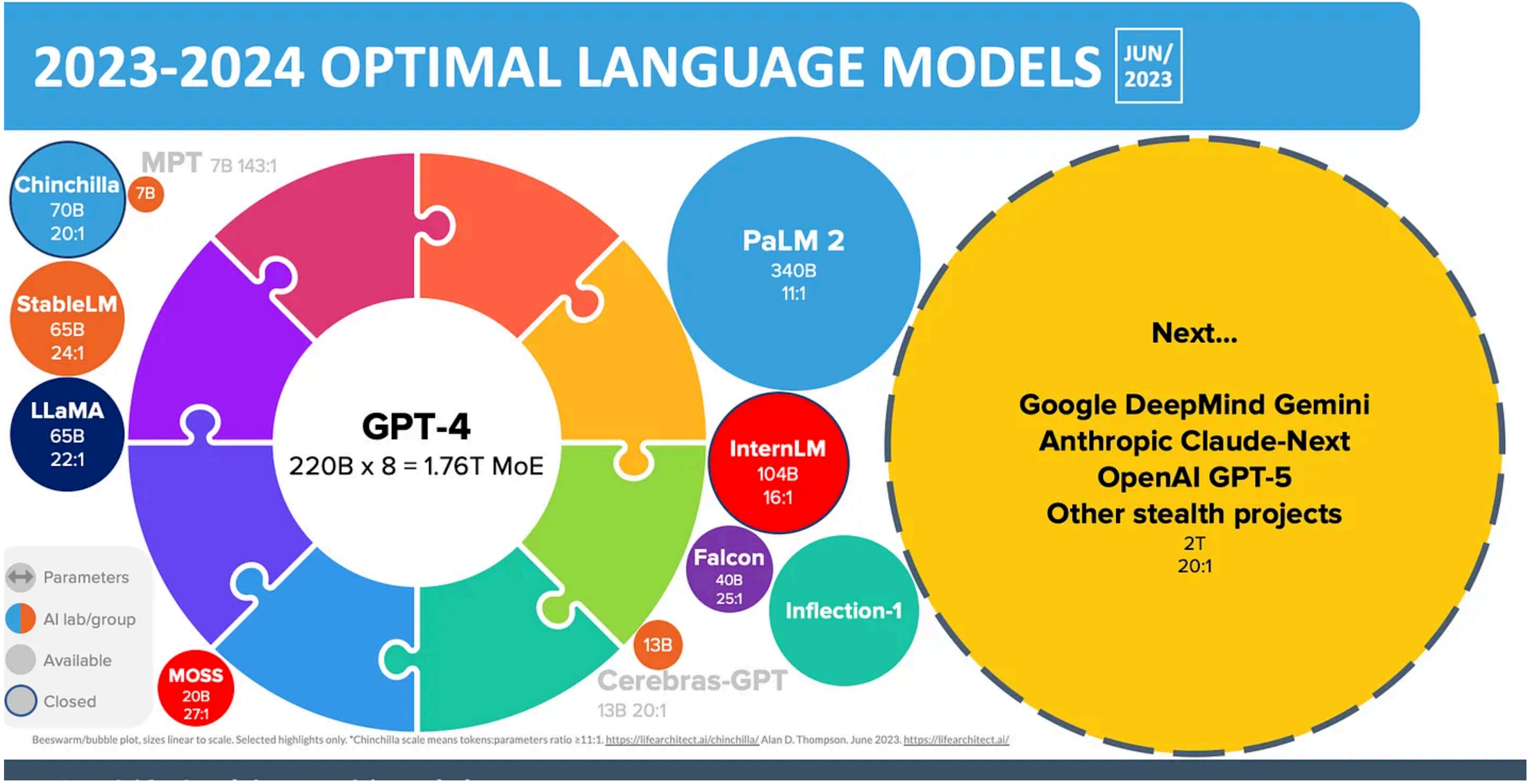
LLMs

- Emergent capabilities
- Prompting
- Few-shot / In-context learning
- Efficient fine-tuning

Applications and tasks

- Machine translation, text generation, question answering, grounding

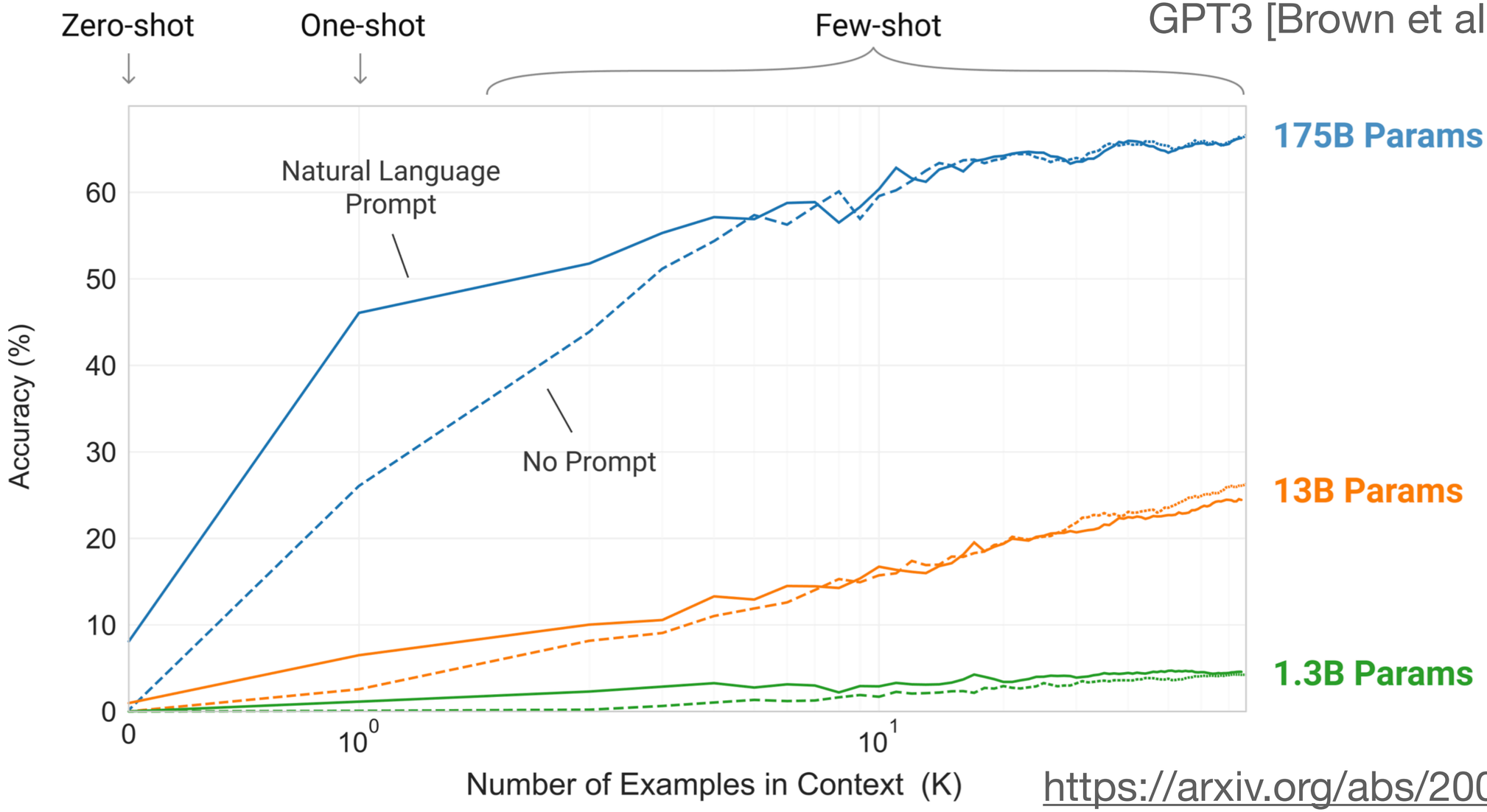
Rise of large pretrained models



https://lifearchitect.ai/models

Emergence of capabilities in LLMs

GPT3 [Brown et al. 2020]



https://arxiv.org/abs/2005.14165

SoTA methods: can use zero-shot/few-shot prompting, or fine-tune

What has this course covered?

Representations

- Embeddings
- Compositional
- Structured

Methods

- Language Models
- Statistical Methods
- Neural Models
- Dynamic programming

LLMs

- Emergent capabilities
- Prompting
- Few-shot / In-context learning
- Efficient fine-tuning

Applications and tasks

Limited coverage of applications

- Machine translation, text generation, question answering, grounding

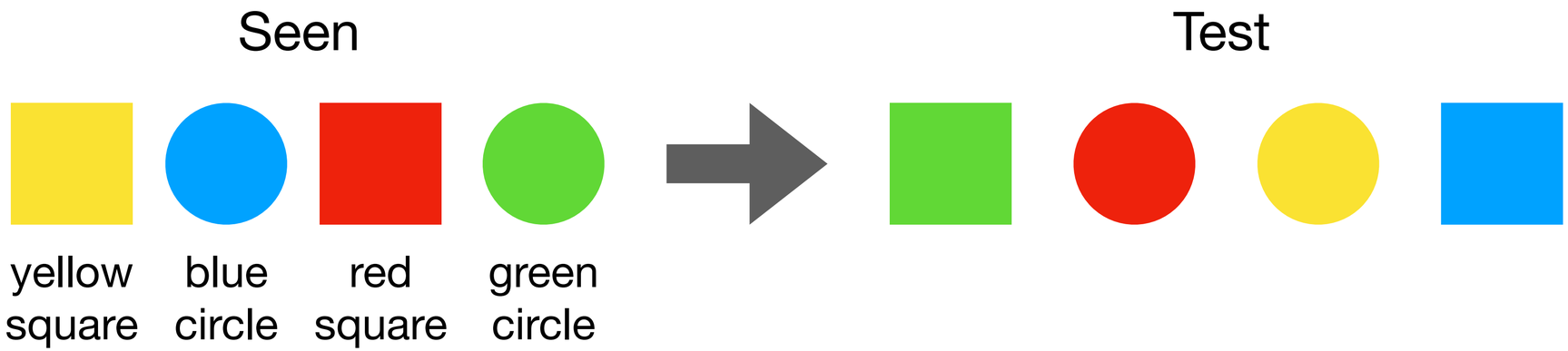
Topics in NLP research

- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Generation and Summarization
- Information Extraction
- Question Answering
- Information Retrieval
- Language Resources and Evaluation
- Language and Vision
- Linguistic and Psycholinguistic Aspects of CL
- Machine Learning for NLP
- Machine Translation
- NLP for Web, Social Media and Social Sciences
- NLP-enabled Technology
- Phonology, Morphology and Word Segmentation
- Semantics
- Sentiment Analysis and Opinion Mining
- Spoken Language Processing
- Tagging, Chunking, Syntax and Parsing
- Text Categorization and Topic Models

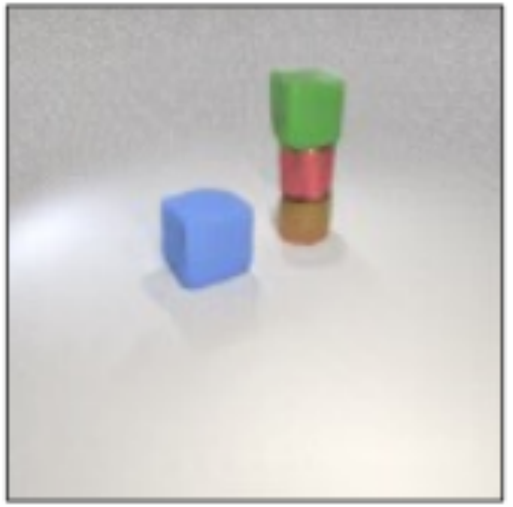
Remaining challenges and topics in NLP

- Understanding and interpreting these large data-driven models
- Compositional generalization
- Grounded natural language understanding
- Unsupervised and few-shot learning
- Language acquisition and emergence of language
- Building practical systems
- Improving evaluation
- Bias, trust, and ethics
- Efficient models
- Integrating domain-specific knowledge into NLP models

Compositional generalization

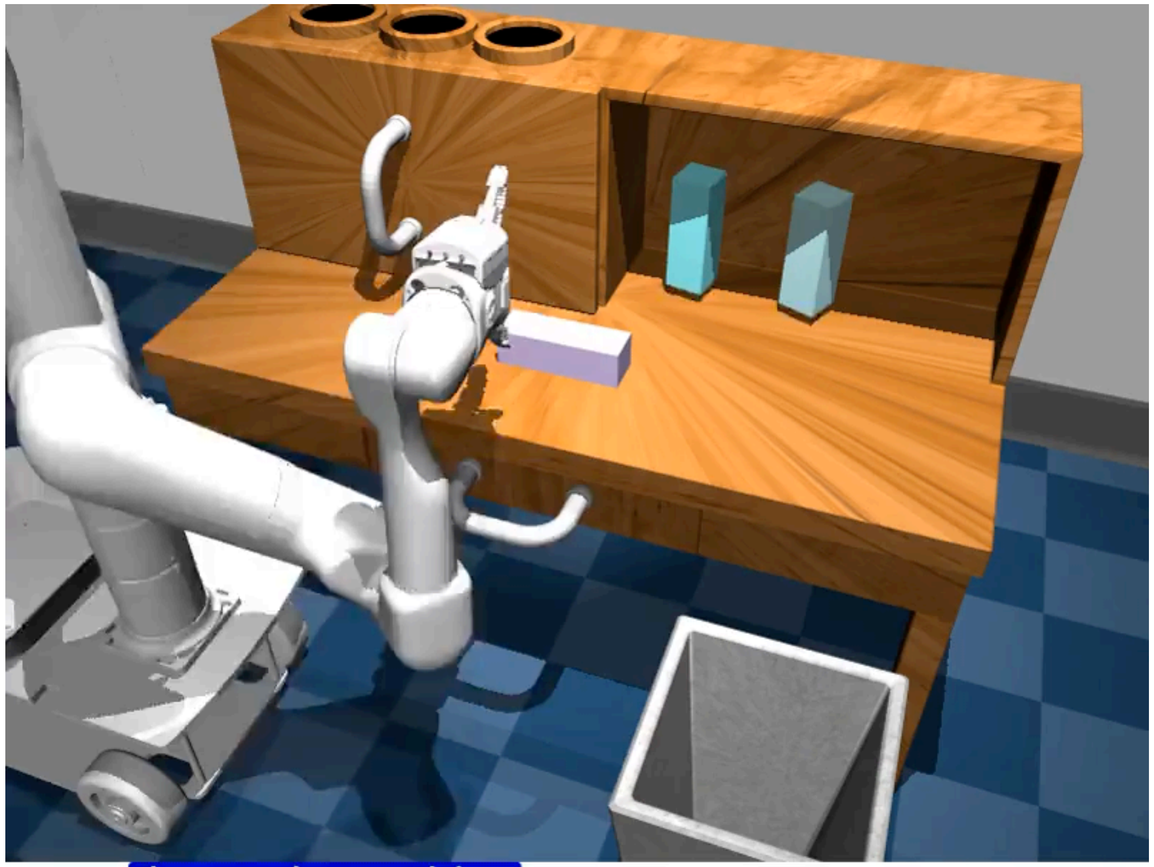


A small red metal cylinder ***below***
a small green rubber cube
A small red metal cylinder ***to the right of***
a large blue rubber cube
A small red metal cylinder ***above***
a small brown metal cylinder



Learning to Compose Visual Relations, Liu et al, NeurIPS 2021

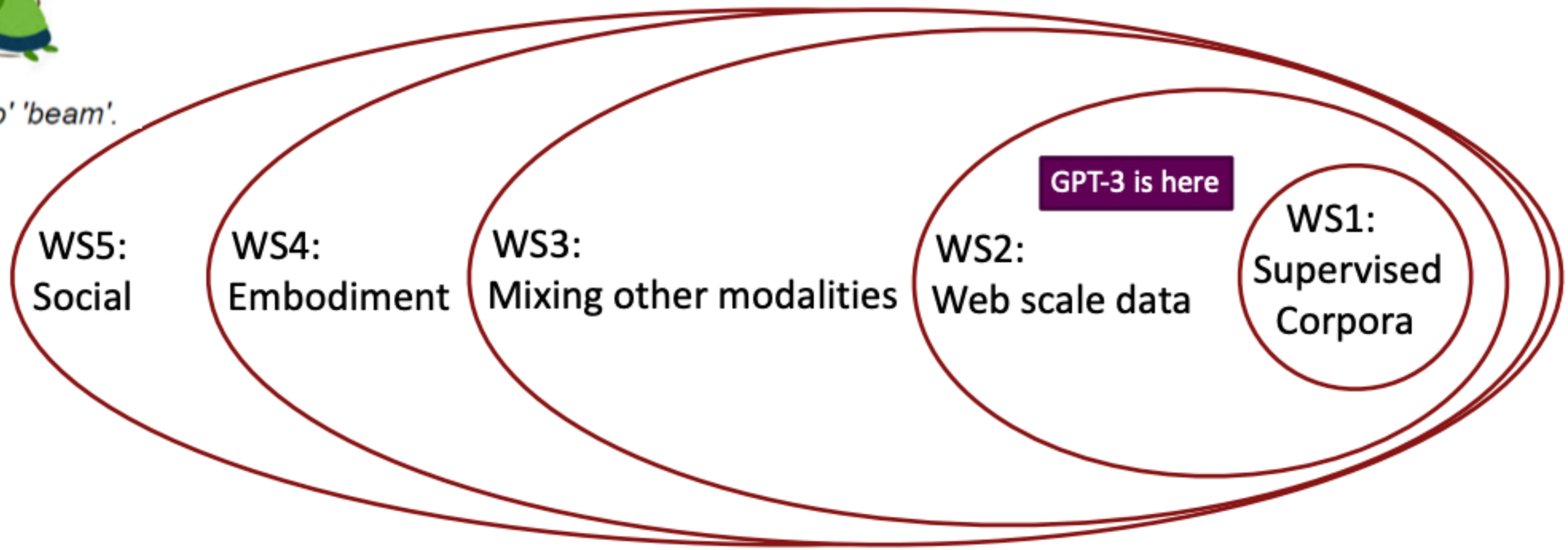
Grounded NLP / RoboNLP



Language Conditioned Imitation Learning over Unstructured Data (https://language-play.github.io/)
Lynch and Sermanet, RSS 2021

Going beyond passive, supervised NLP

'block' 'pillar' 'slab' 'beam'.



Experience Grounds Language (https://www.youtube.com/watch?v=cQasYLUC_00)
Bisk et al, EMNLP 2020