# Raw Data

## Recap

**Cornell CS 5740: Natural Language Processing**
**Yoav Artzi, Spring 2023**

# Raw Data

## Lexical Semantics and Word Embeddings

- Surface forms vs. senses
- Discrete vs. vector-based (distributional) representations
- Sparse vs. dense representations
- Inducing word meaning from raw data
  - Self-supervised learning
  - Approximations
  - Surface- vs. syntax-based
- Context-free (word2vec) vs. context-dependent (BERT) embeddings

# Raw Data

## Language Models

- Approximation of context: Markov assumption
- Count-based n-gram models vs. neural estimators
- Smoothing: generalization vs. just following the training data
- Unknown words
- Tokenization and unseen events
- Evaluation and use
- Decoding
- Scaling: how and impacts
- Auto-regressive LMs vs. masked LMs

# Raw Data

## Neural Architectures

- MLP, even for sequence problems
- Transformers
  - Context weighted-sum
  - Relation to bag-of-words
  - Computational costs
- Encoders vs. decoders