Welcome to Natural Language Processing!

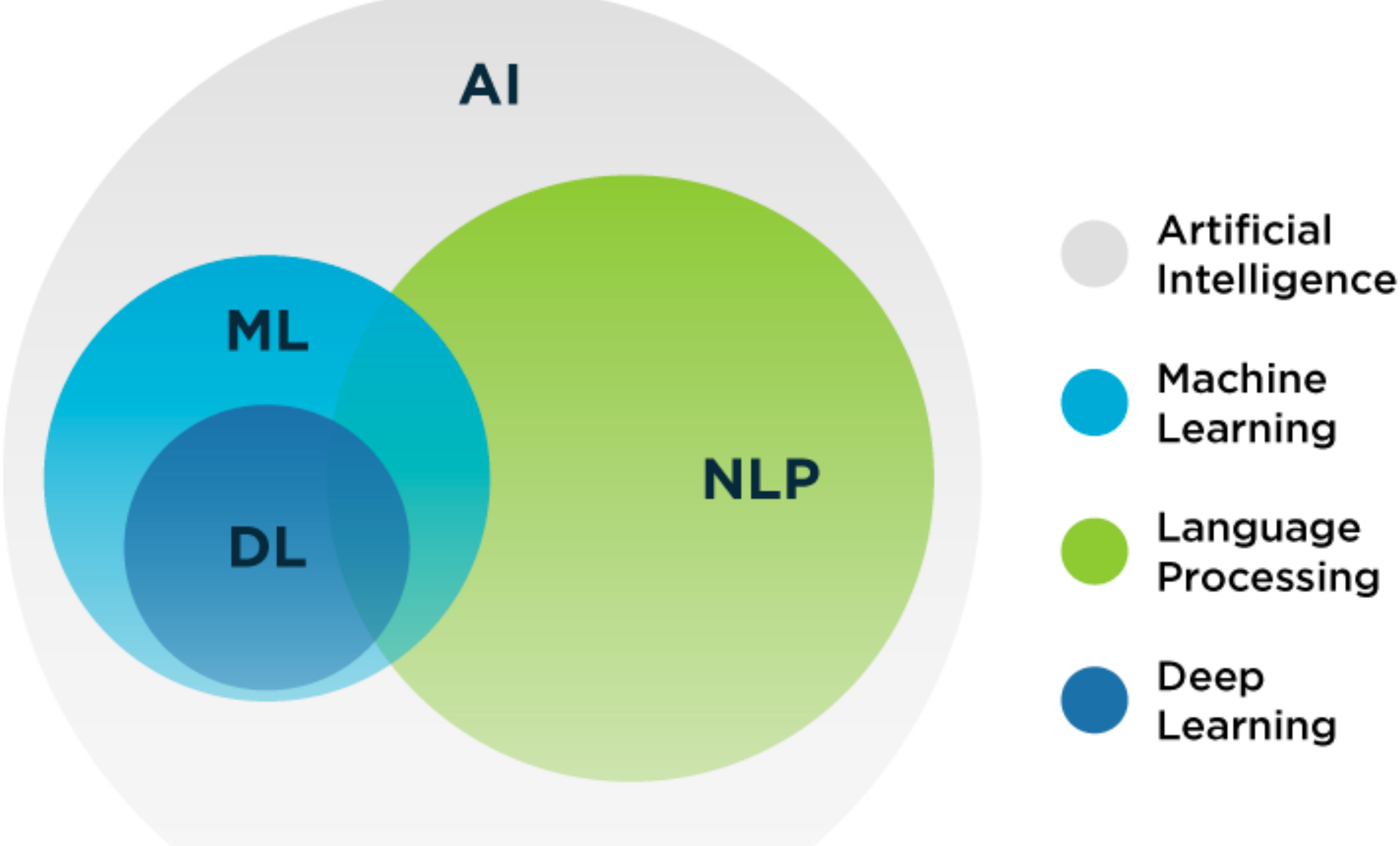
CPSC 599.27/601.27 Winter 2024

Instructor: Katie Ovens

What is natural language?

- Human language
  - Emails
  - Text messages
  - Social media newsfeeds
  - Video
  - Audio
  - and more…

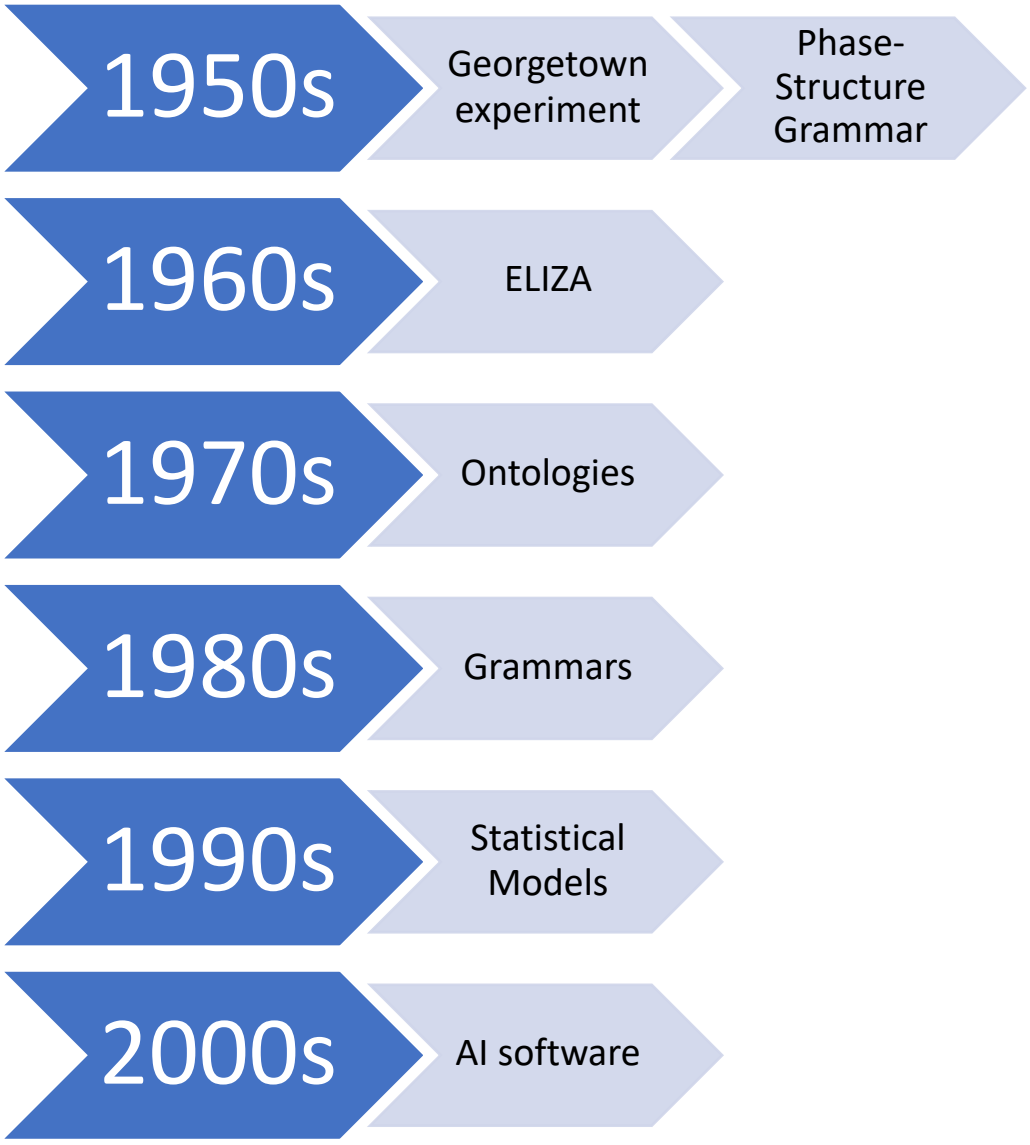
- Automatically interpret, manipulate, and comprehend


AI
ML
DL
NLP
Artificial Intelligence
Machine Learning
Language Processing
Deep Learning

2021 This Is What Happens In An Internet Minute

Timeline of NLP

- The study of natural language processing has been around for more than 60 years



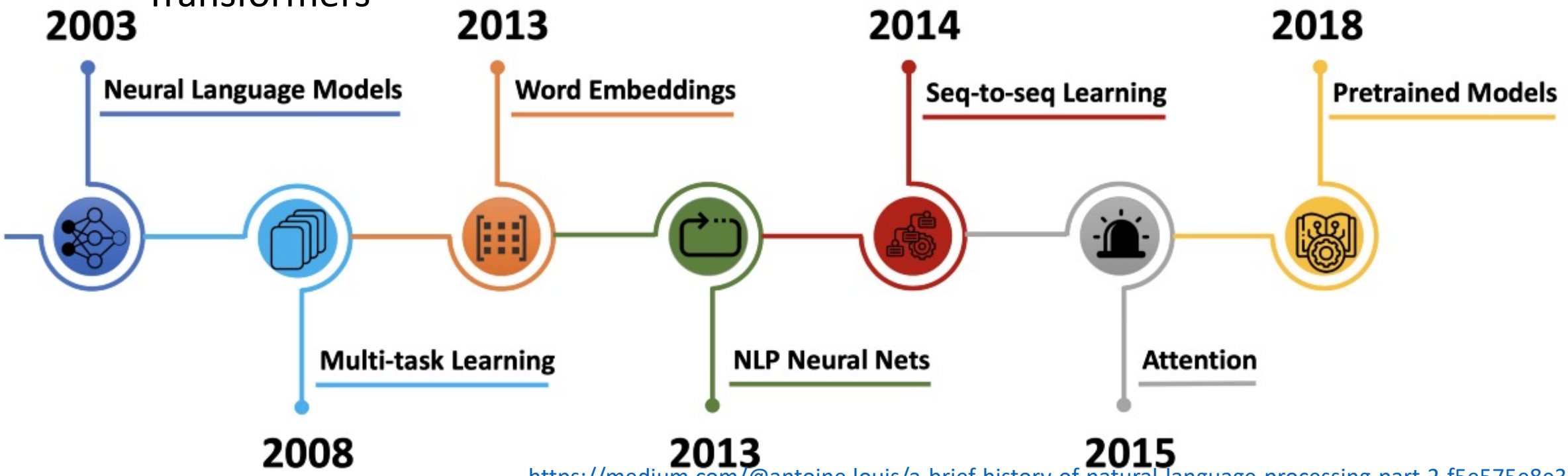
The last decade(s) of NLP

2010s:

- Word2vec
- Encoder-Decoder
- Pretrained language models
- Transformers

2020s:

- GPT and large language models

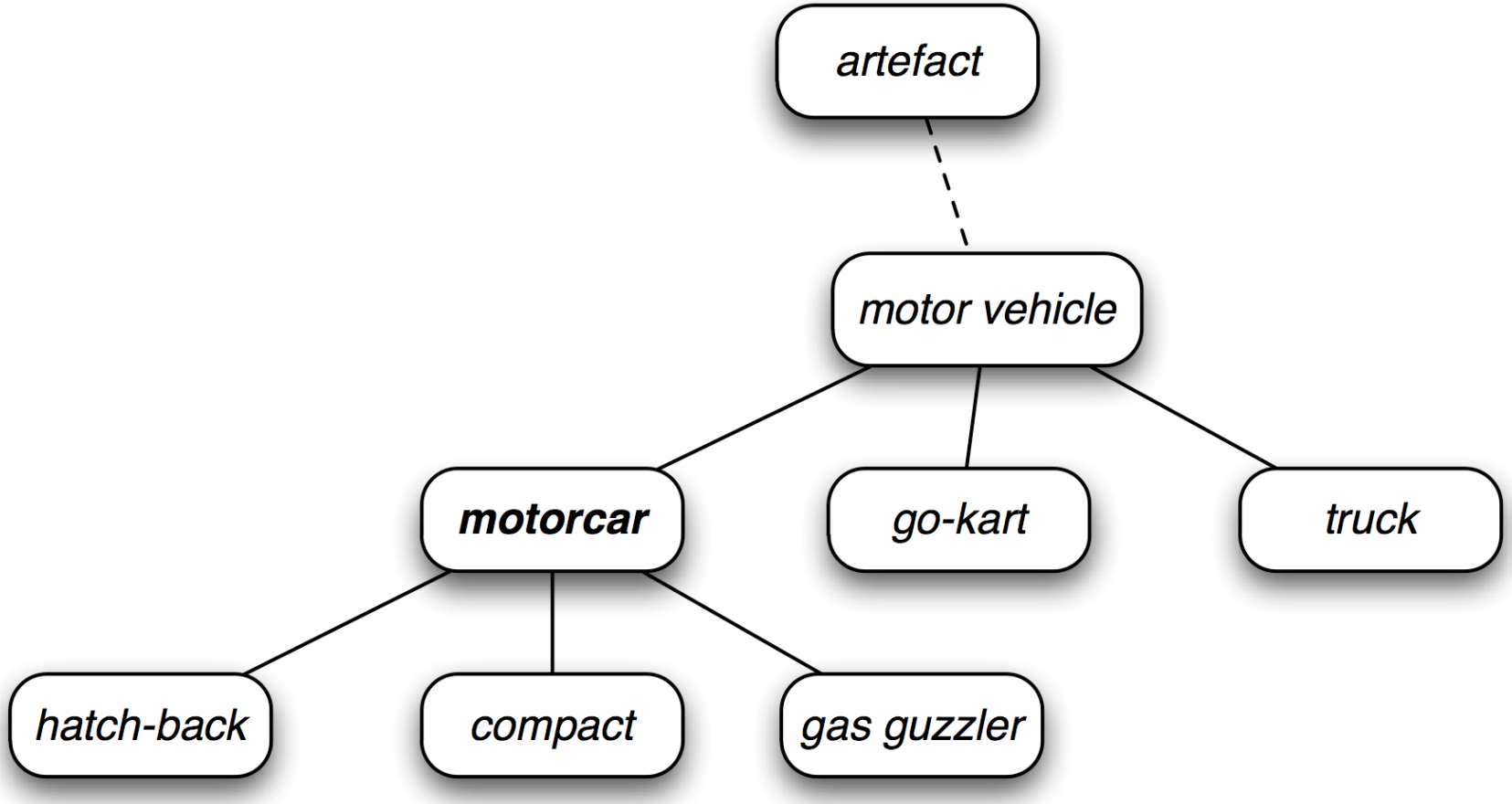


https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37

Not just computer science…

- Interdisciplinary subfield of linguistics, computer science, and artificial intelligence

Wordnet



Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

Is there anything you regularly use that relies on NLP-based applications?

Popular consumer products

- Google
  - Search
  - Gmail
  - Translate
  - Assistant
- Amazon Alexa
- Apple’s Siri

Where are we now?

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

https://openai.com/blog/chatgpt/

What makes a NLP task hard for a machine?

- Try to come up with things to ask or talk about with ChatGPT (https://chat.openai.com/chat) that will cause it to give strange or erroneous responses

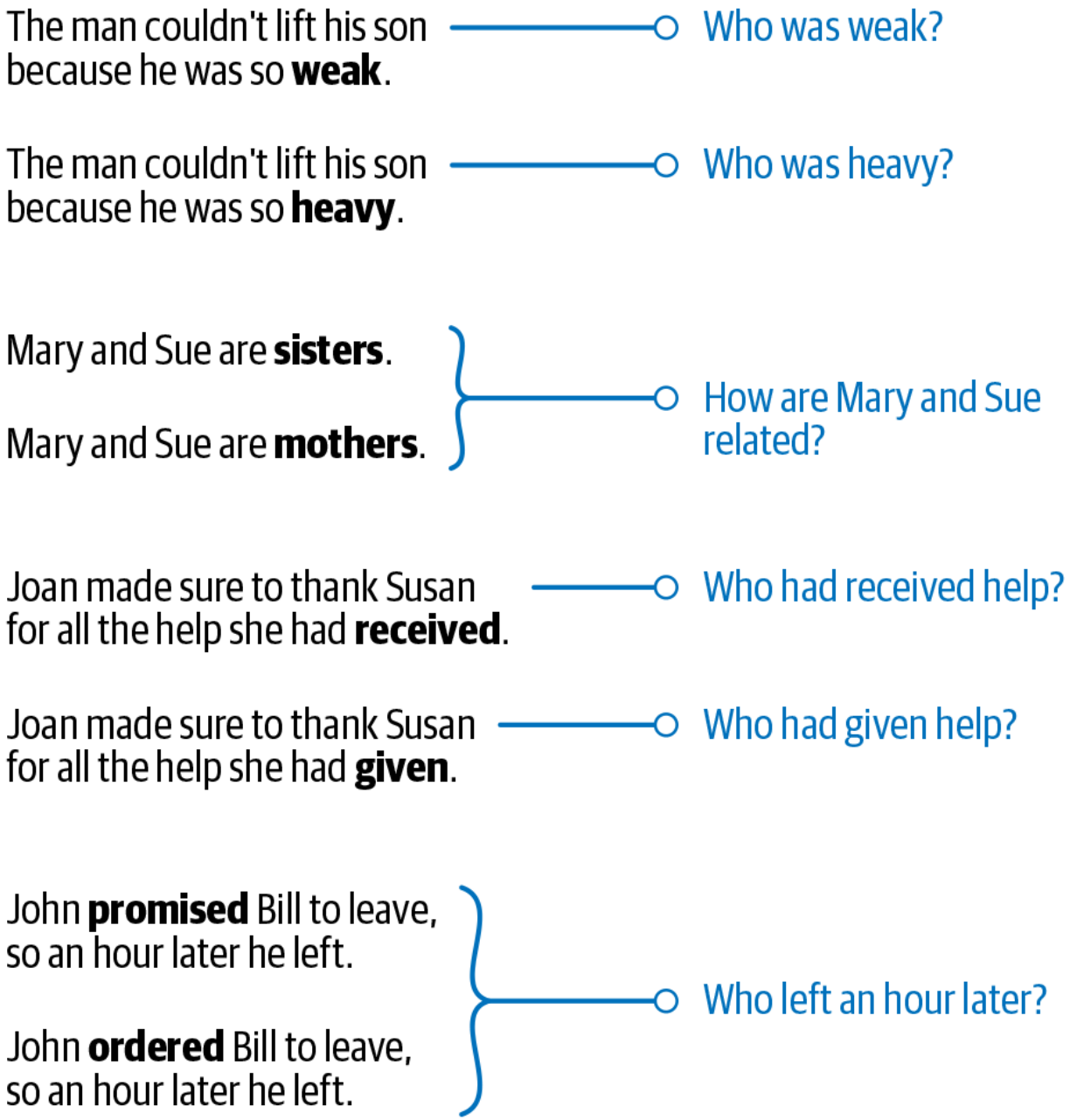
- What things does ChatGPT seem to do well?
- What does it seem to struggle with?

Ambiguity

- What happens when something could be understood in more than one way?

- Lexical
- Syntactic
- Anaphoric
- Etc.

Winograd Schema Challenge



Examples of ChatGPT in action:

- Can you explain semi-self-supervised learning to me?
- Paraphrase what you just said.

- Write about woodchucks that chuck wood in the style of a Shakespearean play.

- Technical: Write a Python 3 function to count the number of words in a sentence.

- Asking for references for certain topics, i.e. give me the citation of 5 of the most influential papers in <insert area of research>

Exercise

- Sort these activities from what you think is the most challenging task for a machine to the least challenging
- Are there any activities that sound unfamiliar?
- Why did you pick the order you picked?

Topic modeling

Text classification

Key-word based information retrieval

Open domain conversational agent

Spell checking

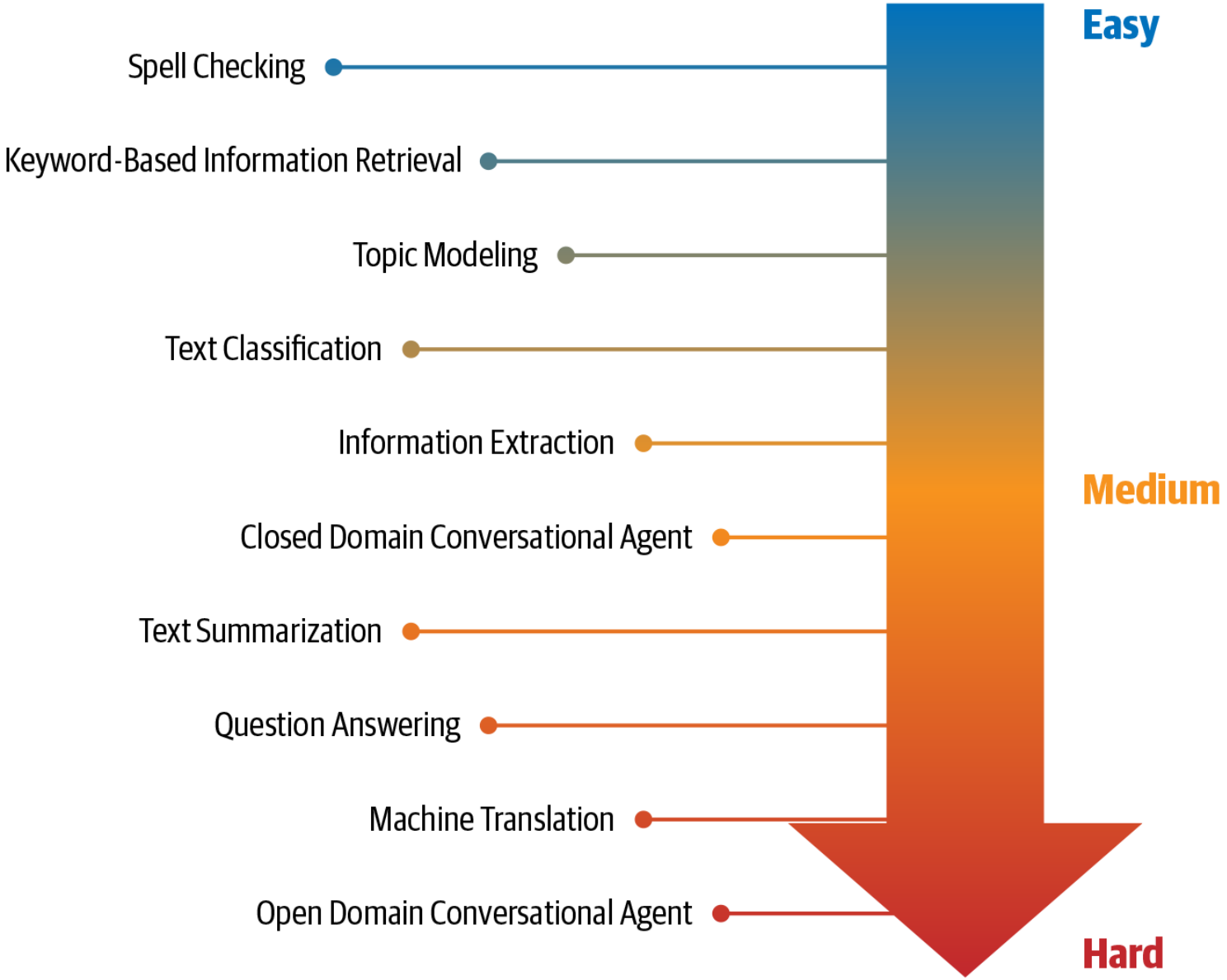
Question answering

Machine translation

Information extraction

Closed domain conversational agent

Text summarization


Spell Checking
Keyword-Based Information Retrieval
Topic Modeling
Text Classification
Information Extraction
Closed Domain Conversational Agent
Text Summarization
Question Answering
Machine Translation
Open Domain Conversational Agent
Easy
Medium
Hard

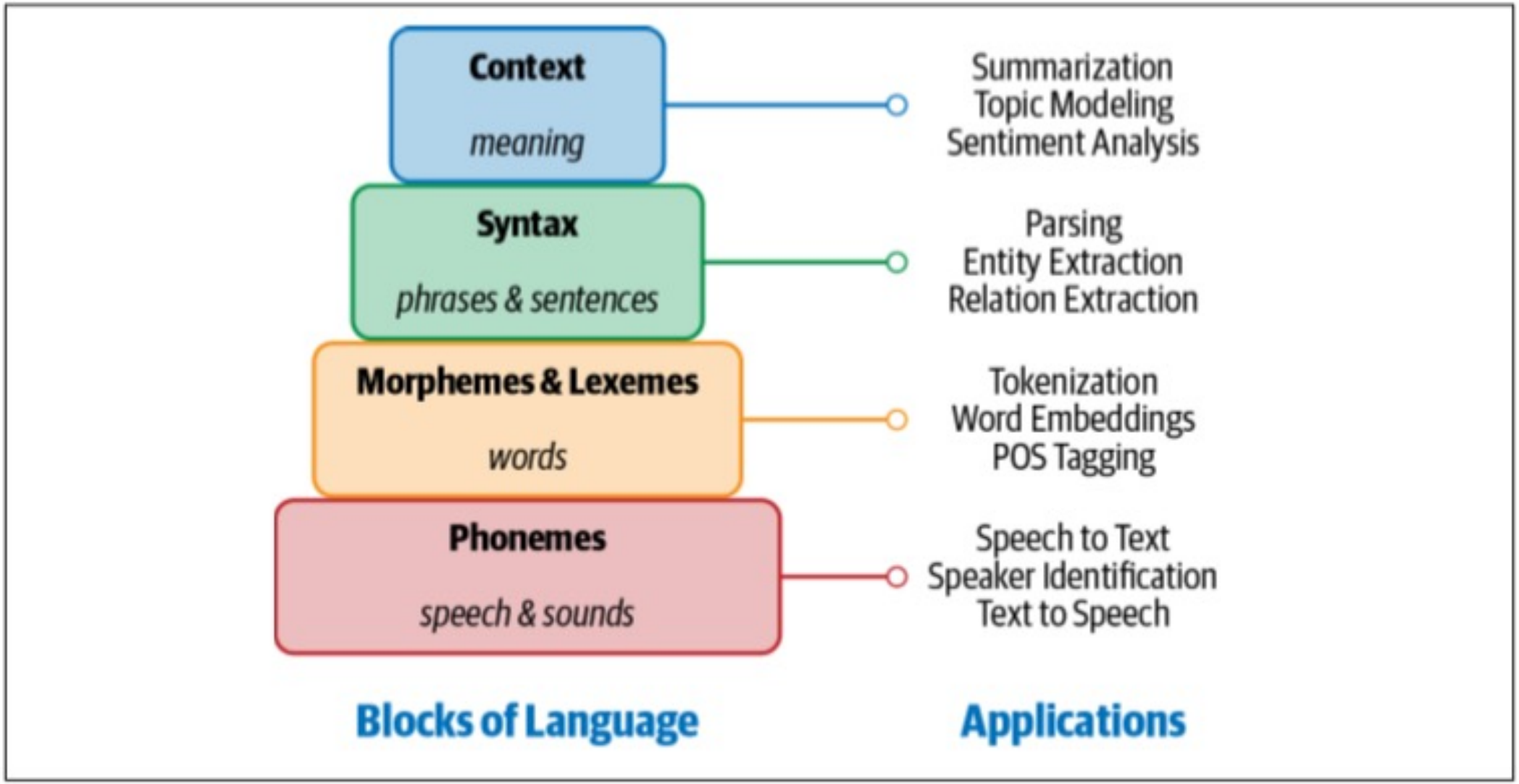


Figure 1-3. Building blocks of language and their applications

CORE TASKS

Text Classification

Information Extraction

Conversational Agent

Information Retrieval

Question Answering Systems

GENERAL APPLICATIONS

Spam Classification

Calendar Event Extraction

Personal Assistants

Search Engines

Jeopardy!

INDUSTRY SPECIFIC APPLICATIONS

Social Media Analysis

Retail Catalog Extraction

Health Records Analysis

Financial Analysis

Legal Entity Extraction

Do I *need* to know about machine learning for this course?

- Python 3
- Jupyter Notebooks
- File I/O
- Statistics

Minimum level of understanding of machine learning and deep learning by the end of this course

We will be *using* ML and DL models, but I will not expect you to build something like a DL model from scratch

Input
What it looks like for different methods.
How to pass input into some already existing methods.
Best practices for organizing data for training machine learning or deep learning models.
ML and DL Models
What models/strategies can be applied to specific NLP problems/applications?
What are the challenges/limitations of the models we cover?
How to use the pretrained version of these models?
Output
What are some strategies of evaluating the results you get from these models?
What is model generalizability?

Common libraries/packages/frameworks

- NLTK
  - The library everyone starts with
- SpaCy
  - Gaining momentum
- TextBlob
  - Built on NLTK
- Scikit-learn
- Pytorch
- More advanced: fast.ai, hugging face

There are many ways to do one thing

NLTK	SpaCy	TextBlob	Textacy
Gensim	Scikit-learn	Stanford NLP	Retext

This course

- Get an overview of traditional NLP concepts and methods
- Utilize the basics of NLTK/SpaCy and PyTorch tensor manipulation library
- Use embedding to represent words, sentences, and documents
- Explore sequence prediction and sequence-to-sequence models
- Apply methodologies for analyzing text in real-world scenarios

Meant to give you a starting point for different topics

Preprocessing

- Spell checking and correction
- Text normalization
- Language detection, code mixing
- Augmentation
- Language detection
- Speech recognition and text-to-speech

Text Representation

- Graph-based (e.g., TextRank, PageRank)
- Knowledge-based (e.g., ConceptNet, WordNet)
- Subword embeddings (e.g. WordPiece)
- Hybrid models
- Multimodal models (e.g., combining text with images or audio)

Neural Networks

- Convolutional neural networks
- Generative models (e.g., Variational Autoencoders, Generative Adversarial Networks)
- Reinforcement learning
- Multimodal Deep Learning

Classification

- Other classical ML models
- Multi-class Classification
- Evaluation metrics, cost-sensitive learning

Text generation

- Gated recurrent units
- Bi-directional LSTMs
- Attention mechanisms
- Conditional language models
- Machine translation

.

.

.

Schedule

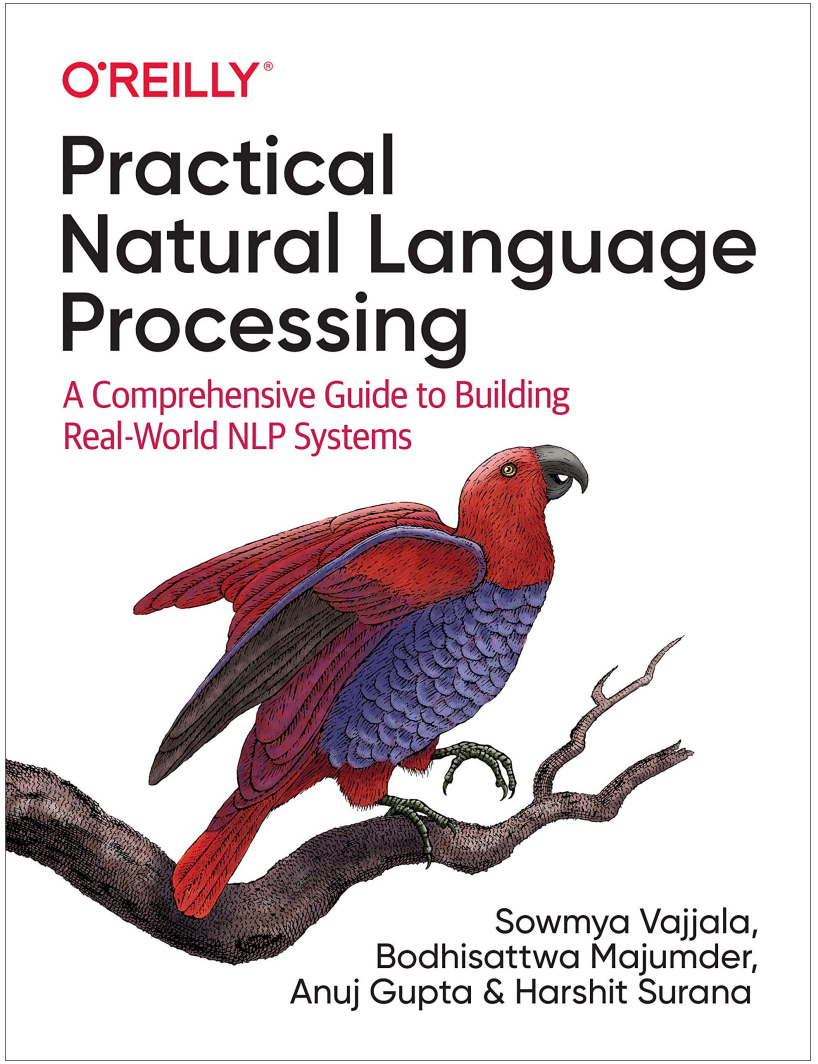
Week	Tuesday	Thursday
Jan 8 - Jan 12	Course Introduction	The art of preprocessing (NLTK)
Jan 15 - Jan 19	The art of preprocessing (SpaCy)	NLTK vs SpaCy
Jan 22 - Jan 26	Getting Datasets, BeautifulSoup	Word Vectors and Text representation
Jan 29 - Feb 2	Neural networks and autoencoders	Word2Vec + GloVe
Feb 5 - Feb 9	Visualizations	Topic Modelling
Feb 12 - Feb 16	Classification	Using pretrained models
Feb 19 - Feb 23		
Feb 26 - Mar 1	Classification	Sentiment Analysis
Mar 4 - Mar 8	Information Extraction	Recommender System
Mar 11 - Mar 15	Text generation	Making a Chatbot
Mar 18 - Mar 22	Context: Attention and Transformers	Coreference resolution
Mar 25 - Mar 29	Dependency Parsing	Large Language Models
Apr 1 - Apr 5	Presentations	Presentations

O'Reilly Resources

- Free access through the University of Calgary

(https://www.oreilly.com/member/login/)

- **Practical Natural Language Processing**
- Natural Language Processing with NLTK and Pytorch
- Applied Natural Language Processing in the Enterprise



No tutorials

- Practical examples will be covered in-class
- Office hours (by appointment through bookings – see D2L)
- Notebooks with exercises for practice
- Discussion board set up on D2L to ask questions

Assignments

- 4 Assignments, 15% each
  - Assignment 1: NLTK and SpaCy
  - Assignment 2: Web scraping, text representation and Word2Vec
  - Assignment 3: Classification and Sentiment Analysis
  - Assignment 4: Chatbots & Text Generation

Project

- Individual or pairs
- **Proposal**
  - Pick a topic/application in NLP to explore
  - Application: identifying datasets & proposed NLP method(s)/models that will be applied to your datasets
- **Milestone**: Half-way mark report that includes what you have found out about your topic and dataset, challenges you have encountered, and how you solved or plan to solve them
- **Presentation**: 7-10 minute presentation of project
  - Online recording for undergraduate students
  - In-class for graduate students
- **Report**

Things to note:

- **I will not be rounding final grades** – any emails asking to round up grades at the end of term will be ignored
- **You have 10 days to dispute Assignment or Project submission grades once they have been released**
- **Contact me to negotiate extensions before deadlines pass –** I will also allow for late submissions of assignments without a prior extension agreement with me, but the submission will at most get 50%

What You Can Expect from Me

- I will be here on time
- Your project milestones (an assignments) will be graded in a timely manner
  - Typically within 1-2 weeks
- Discussion board or emailed questions will be responded to in a timely manner
  - Typically within 1-2 days
- If I don't know the answer to your question, I will (try my best) find out
- I will do my best to incorporate feedback

Participation and Feedback

- 5% of your grade
- Each participation survey/activity is worth 1%, capped at 5%

Muddiest Point

2-minute Memo

Guest Lectures

Questions?

Tasks for next class: The art of preprocessing

- Get Jupyter Notebook setup or just use Google Colab
  - https://colab.research.google.com/
- Install NLTK and SpaCy