U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

LIACC

M.EIC

Natural Language Processing

Henrique Lopes Cardoso

FEUP / LIACC

hlc@fe.up.pt

Teaching Staff, Webpage and Classes

Henrique Lopes Cardoso

(hlc@fe.up.pt)

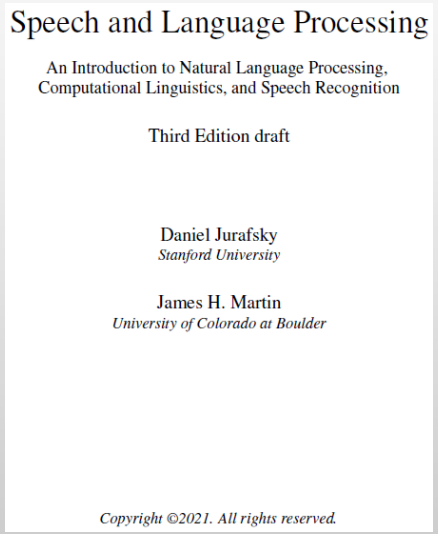
Room i121 (AI lab)

- 3h classes, with
  - Theoretical topics
  - Hands-on exercises
  - Project development
- Course webpage in Moodle
  - Materials (slides, resources)
  - Assignments
  - Forum

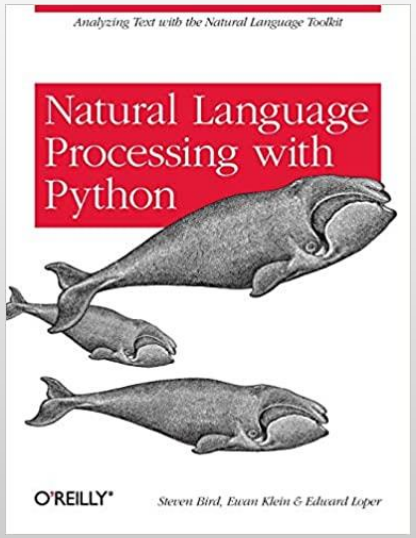
Syllabus

- Introduction to NLP: definitions, tasks, and applications.
- Basic text processing: regular expressions, tokenization, normalization, lemmatization, stemming, segmentation.
- Language models: n-grams.
- Text classification: bag-of-words, Naive Bayes, feature engineering; generative and discriminative classifiers.
- Sequence models: hidden Markov models, conditional random fields; POS-tagging and named entity recognition.
- Vectorized representations of words: lexical semantics, word embeddings.
- Neural networks in NLP: neural language models, RNNs, encoder-decoder networks, attention, Transformers.
- Contemporary research in NLP.

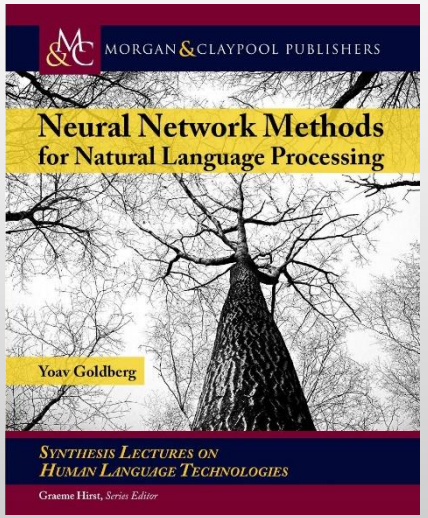
Bibliography and Tools

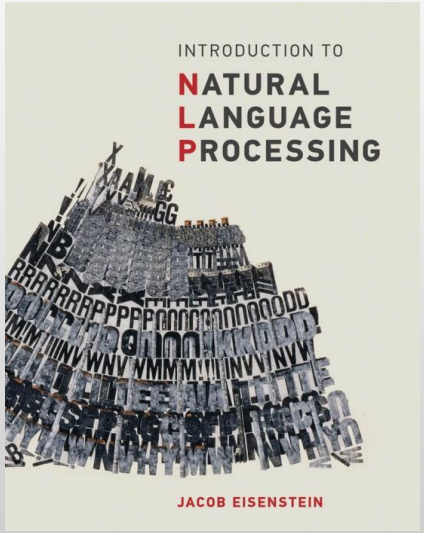


https://web.stanford.edu/~jurafsky/slp3/



https://www.nltk.org/book/





NLTK

https://www.nltk.org/

Stanza

https://stanfordnlp.github.io/stanza/

spaCy

https://spacy.io/

Hugging Face

https://huggingface.co/

Evaluation

- 2 practical assignments (2x6/20)
  - Text classification
  - More advanced task or techniques
- 1 oral presentation on a recent research trend (2/20)
  - Topics and literature to be provided
- Final exam (6/20)
  - Moodle, mostly multiple-choice

Class Plan

date	topic	evaluation
2023-02-10	Intro. Basic text processing	
2023-02-17	Language models	
2023-02-24	Text classification	
2023-03-03	Logistic Regression	
2023-03-10	Sequence labeling	
2023-03-17	Word embeddings	
2023-03-24	Invited talk. Work monitoring	
2023-03-31	**1st assignment presentations**	x
2023-04-07	Easter	
2023-04-14	Neural Networks in NLP	
2023-04-21	Transformers: pre-train and fine-tune	
2023-04-28	**Research direction presentations**	x
2023-05-05	**Research direction presentations**	x
2023-05-12	Academic week	
2023-05-19	Invited talk. Work monitoring	
2023-05-26	**2nd assignment presentations**	x

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

LIACC

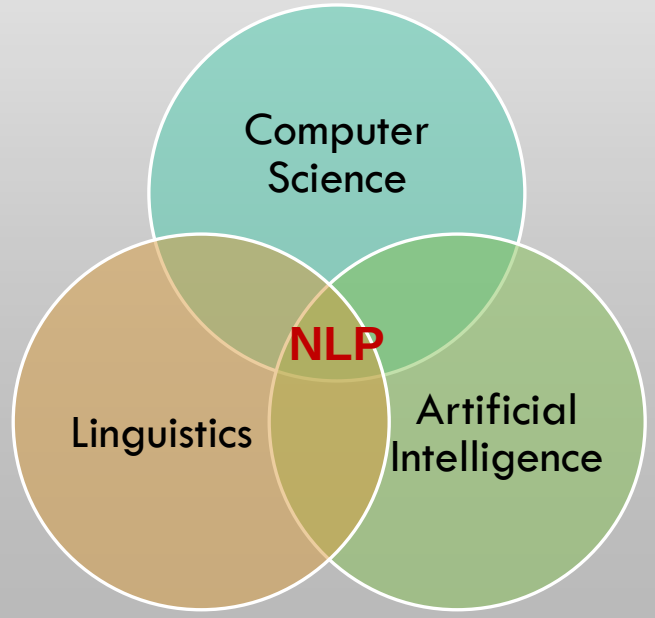
Natural Language Processing (NLP)

definitions, applications, tasks, resources, machine learning approaches, NLU, NLG

Natural Language Processing

Natural language processing (**NLP**) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora.

[Wikipedia]



Some NLP Applications

- Machine Translation
  - Based on multilingual textual corpora
  - Text translation and multilingual real-time conversations

DeepL

- Speech-to-Text/Text-to-Speech
  - Convert spoken language to written text and vice versa
  - Voice control, domotics, readers, …

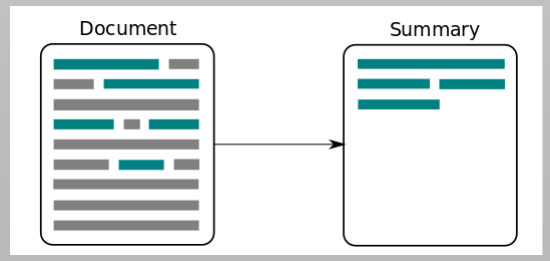
amazon alexa

Some NLP Applications

- Sentiment Analysis and Opinion Mining
  - Determine polarity about specific topics
  - Identify trends of public opinion in social media
  - Analyze product reviews

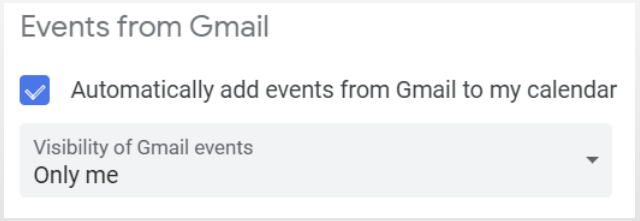
Attributes:
- zoom
- affordability
- size and weight
- flash
- ease of use

- Text Summarization
  - Build a summary out of a long text



Some NLP Applications

- Information Extraction
  - Extract relevant entities from text
  - Event identification, “add to calendar” features



Subject: extra NLP class
Date: September 25, 2020
To: Henrique Lopes Cardoso

Dear Henrique, we’re having an extra NLP class tomorrow, from 10:00-11:30, via Zoom.
-HLC

Event: extra NLP class
Date: Sept-26-2020
Start: 10:00am
End: 11:30am
Where: Zoom

- Question Answering
  - Automatically answer questions posed in natural language
  - IBM Watson won *Jeopardy!* on 2011

Some NLP Applications

- Fact Checking and Fake News Detection
  - Given a claim, collect evidence to check if it is true
  - Given a news article, check whether it is accurate

FEVER

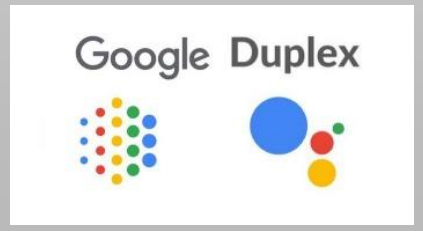
- Argument Mining and Debate Portals
  - Extract arguments that expose a certain position
  - Aggregate pros and cons for a debatable topic
  - Debate on a given topic

PRO DEBATEPEDIA CON

BRITANNICA PROCON.ORG | RELIABLE. NONPARTISAN. EMPOWERING.

Some NLP Applications

- Natural Language Generation
  - Generate a narrative from data
  - Generate source code from natural language descriptions
  - Language models
- Conversational AI (Chatbots)
  - Conversational interfaces
  - Human-like voice assistants



NLP Tasks

- Most NLP tasks aim at making it easier for machines to understand natural language

- Tokenization
  - Split a sentence into tokens (words)

```
That U.S.A. poster-print costs $12.40...
['That', 'U.S.A.', 'poster-print', 'costs', '$12.40', '...']
```

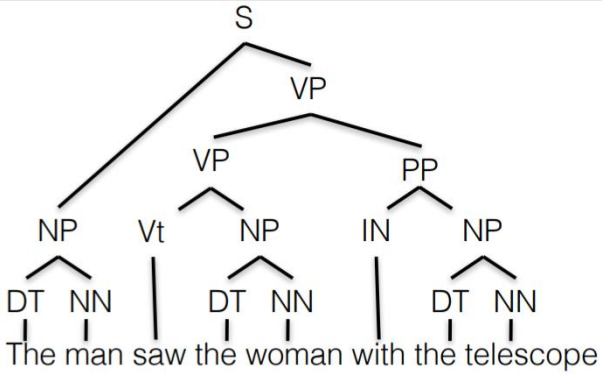
- Sentence breaking
  - Split a text into sentences

```
Hello. Are you Mr. Smith? I've finished my M.Sc. on Informatics!
['Hello.',
 'Are you Mr. Smith?',
 'I've finished my M.Sc. on Informatics!']
```

- Part-of-Speech (POS) tagging
  - Determine the role category for each word in a sentence

NLP Tasks

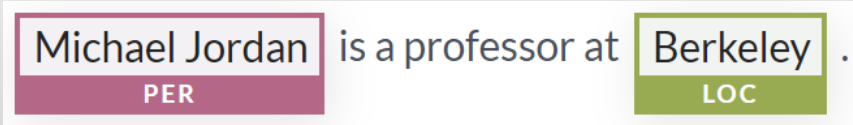
- Syntax parsing
  - Determine the parse tree (grammatical analysis) of a sentence



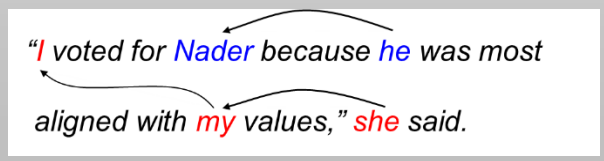
- Word sense disambiguation
  - Select the meaning of words in a context

A mouse is a mammal.
My mouse is broken.

- Named-entity recognition (NER)
  - Determine which items in a text map to entities (people, institutions, places, dates, …)



- Co-reference resolution
  - Determine which words (“mentions”) refer to the same objects (“entities”)



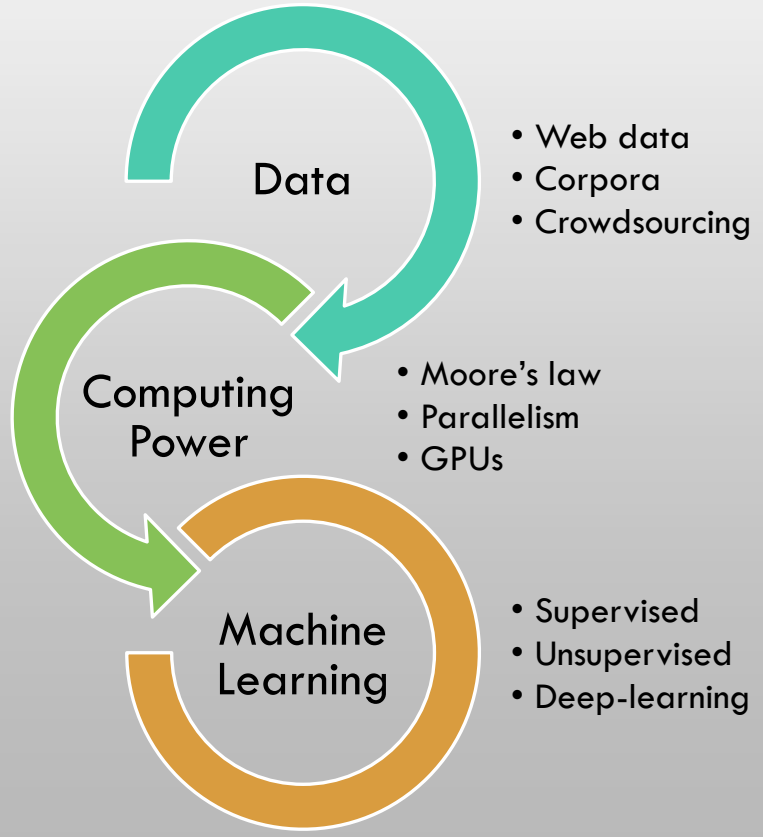
- …

Language Resources

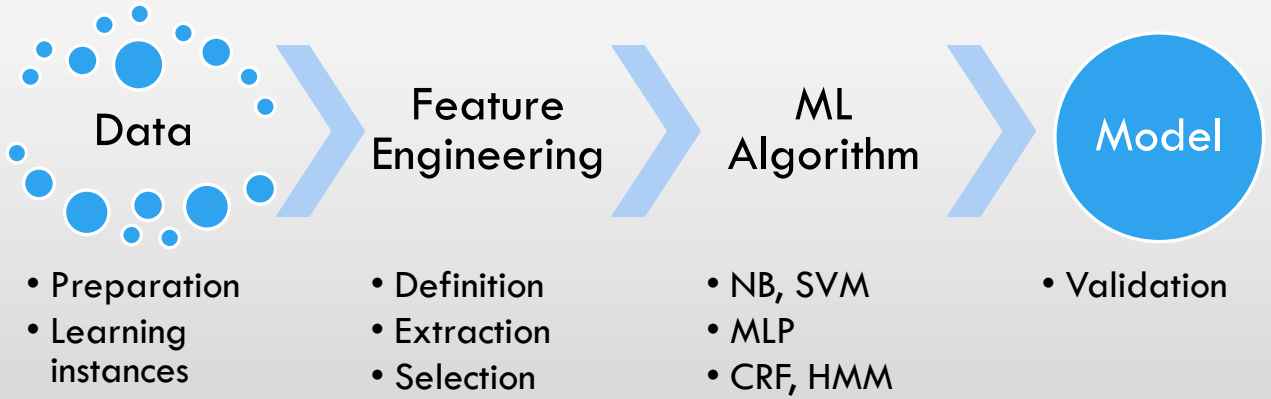
- Lexical databases: WordNet, CONTO.PT, WordNet.pt, …
  - Synsets, word-sense pairs
  - Semantic relations: hypernym/hyponym, meronym/holonym, troponym, entailment, …
- TreeBanks: PDTB, CSTNews, …
  - Text corpora annotated with discourse or semantic sentence structures
- Knowledge graphs: Google, DBpedia, …
  - Entity-predicate relations
- Lexicons: SentiWordNet, SocialSent, SentiLex, VADER, …
  - Words connoted with specific classes (+/-, objectivity, …)
- Word embeddings: word2vec, GloVe, fastText, …
  - Distributed representations of words
- Language Models: ELMo, BERT, …
- Annotated datasets for several NLP tasks
  - Often released under “shared-tasks”, such as those at SemEval or CLEF
- …

Statistical NLP

- Data-driven statistical techniques overtook knowledge (grammar) based methods

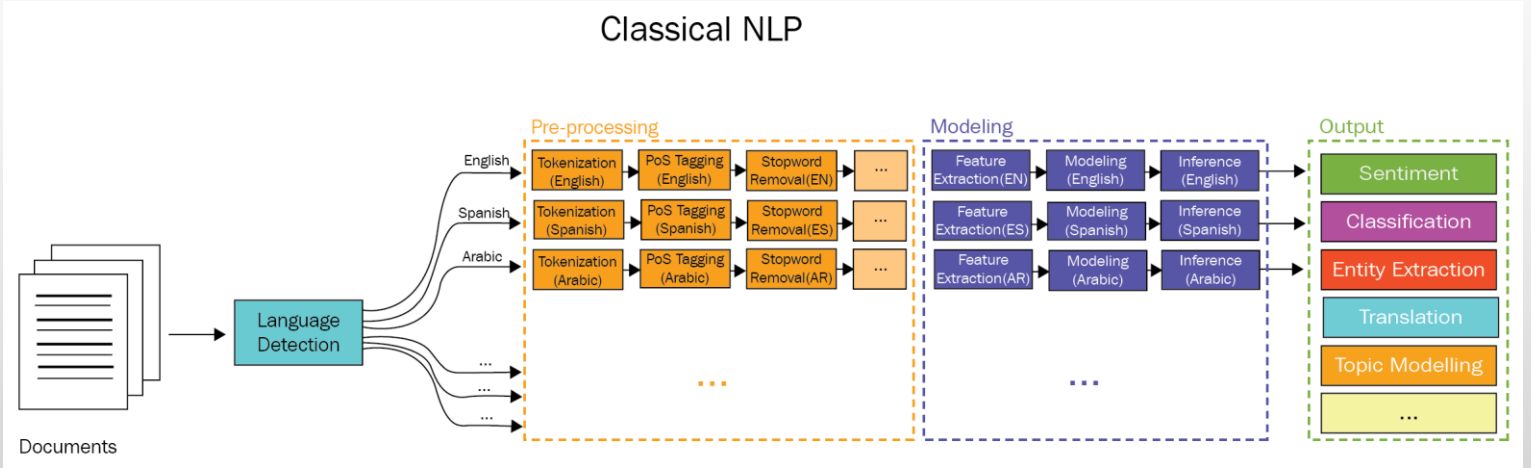


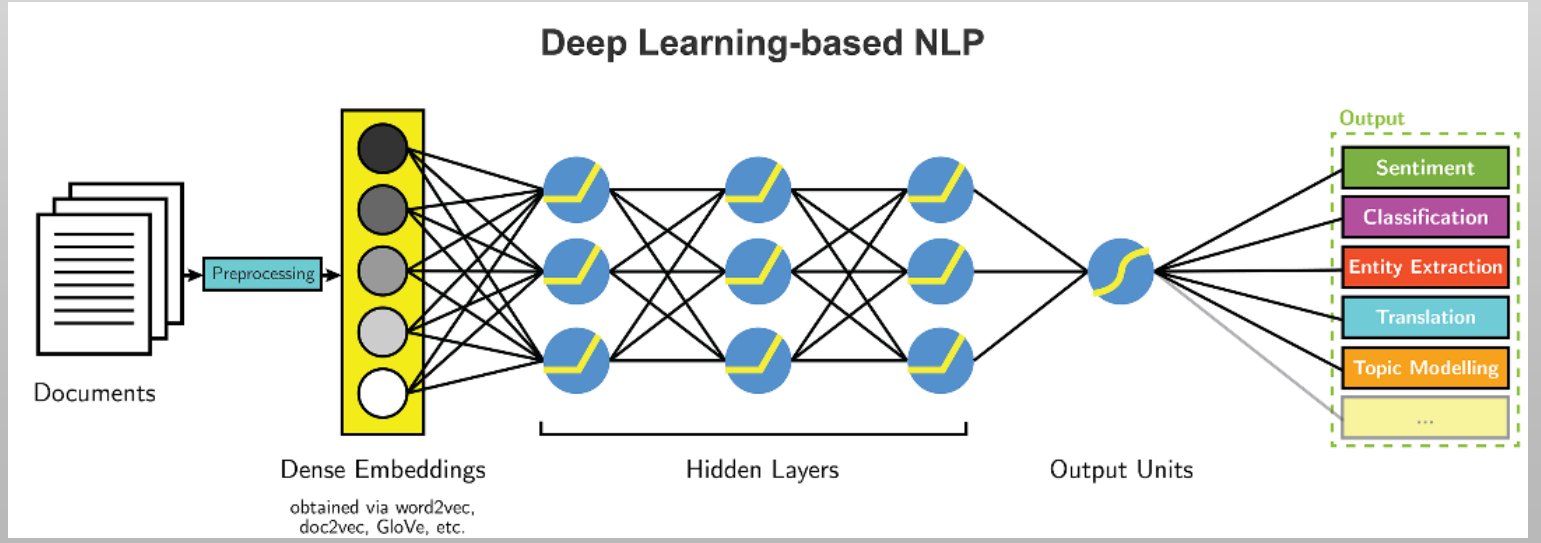
Machine Learning in NLP



- Common linguistic features used in NLP
  - *Lexical*: BoW, TF-IDF, n-grams, word stems, …
  - *Syntactic*: part-of-speech (POS) tagging, parsing, …
  - *Grammatical*: verb tenses, number, gender, …
  - *Semantic*: word similarities, relations, embeddings, …
  - *Structural*: paragraphs, sentence length, document sections, distance metrics, …

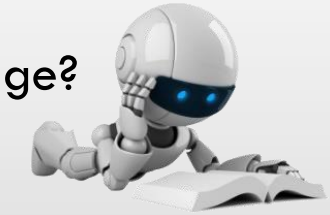
Classical vs Deep Learning NLP





Natural Language Understanding

- Can computers *understand* natural language?



- 2011: IBM Watson, a question answering computer system, won *Jeopardy!*

  - Q&A technology takes a question expressed in natural language and returns a precise answer
  - Does Watson *understand* the questions?

Natural Language Understanding

- Why is NLU difficult?
  - Ambiguity ("Red tape holds up new bridges", "Hospitals sued by 7 Foot Doctors")
  - Non-standard language use (e.g. in Twitter)
  - Segmentation issues ("the New York-New Haven Railroad")
  - Idiomatic expressions ("throw in the towel", "dark-horse candidate", …)
  - Neologisms ("unfriend", "retweet", …)
  - World knowledge ("Mary and Sue are sisters" / "Mary and Sue are mothers")
  - Tricky entity names ("Where is *A Bug's Life* playing?", "*Let it Be* was recorded …")
  - …
- Need knowledge about language and knowledge about the world!

Polícia cerca prédio com índios no Rio

Liminar que garantia posse do imóvel de 1862 aos indígenas foi cassada; governo quer demoli-lo para fazer estacionamento e instalações da Copa

ESPECIAL DE 10 PÁGINAS: TUDO SOBRE A PANDEMIA P. 4 A 13

ESCOLAS ÀS CEGAS COM TESTES RÁPIDOS

GOVERNO ENCOMENDOU 400 MIL UNIDADES, MAS OS DIRETORES NÃO SABEM COMO FAZER O RASTREIO. PAIS APLAUDEM A MEDIDA

Carta de condução vai poder ser usada apenas no telemóvel

Documentação do automóvel, incluindo o seguro e a inspeção, também passa a formato digital

Se as polícias não tiverem meios eletrónicos de leitura é preciso ir mostrar papéis à esquadra Pág

Há 87 concelhos do país que estão a escapar ao vírus

Densidade populacional e distância face a aeroportos beneficiam interior

Reportagem em Mesão Frio e Mondim de Basto, onde ainda não há infetados P.

San Jose cops kill man with knife

Ex-college football player, 23, shot 9 times allegedly charged police at fiancee's home

shortly after she called a suicide intervention hotline in hopes of getting Watkins medical

ed help from police."

She said Watkins was on the sidewalk in front of the home when two

ing for their safety and defense of their life, fired at the suspect."

On the police radio,

Students Cook & Serve Grandparents

On Thursday, September 9, Gorman School hosted the first annual Grandparent's Day.

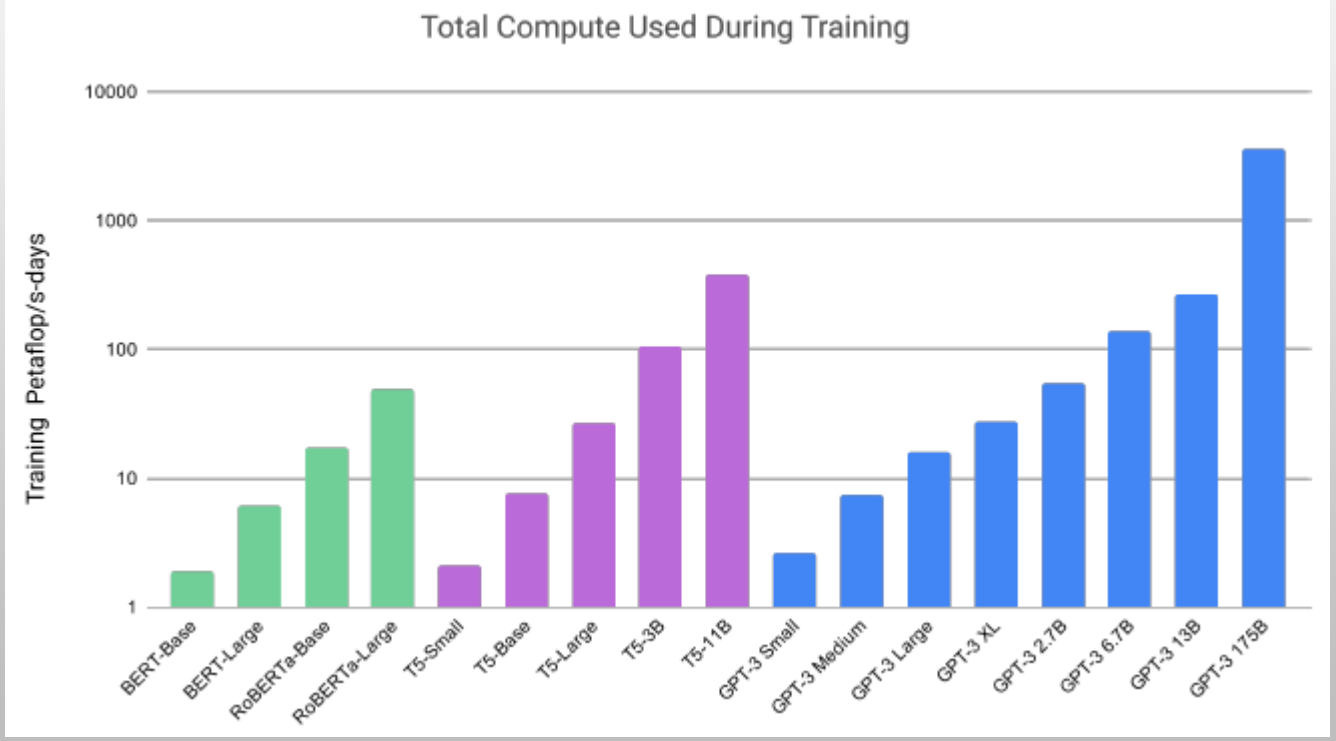
Natural Language Generation

- The process of transforming structured data into natural language
  - *Data-to-text*: generate textual summaries of databases and data sets (weather, finance, business, …)
  - Integrated into business intelligence and analytics platforms
- Other application areas: automated journalism, chatbots, question-generation, …
  - … and fake news?

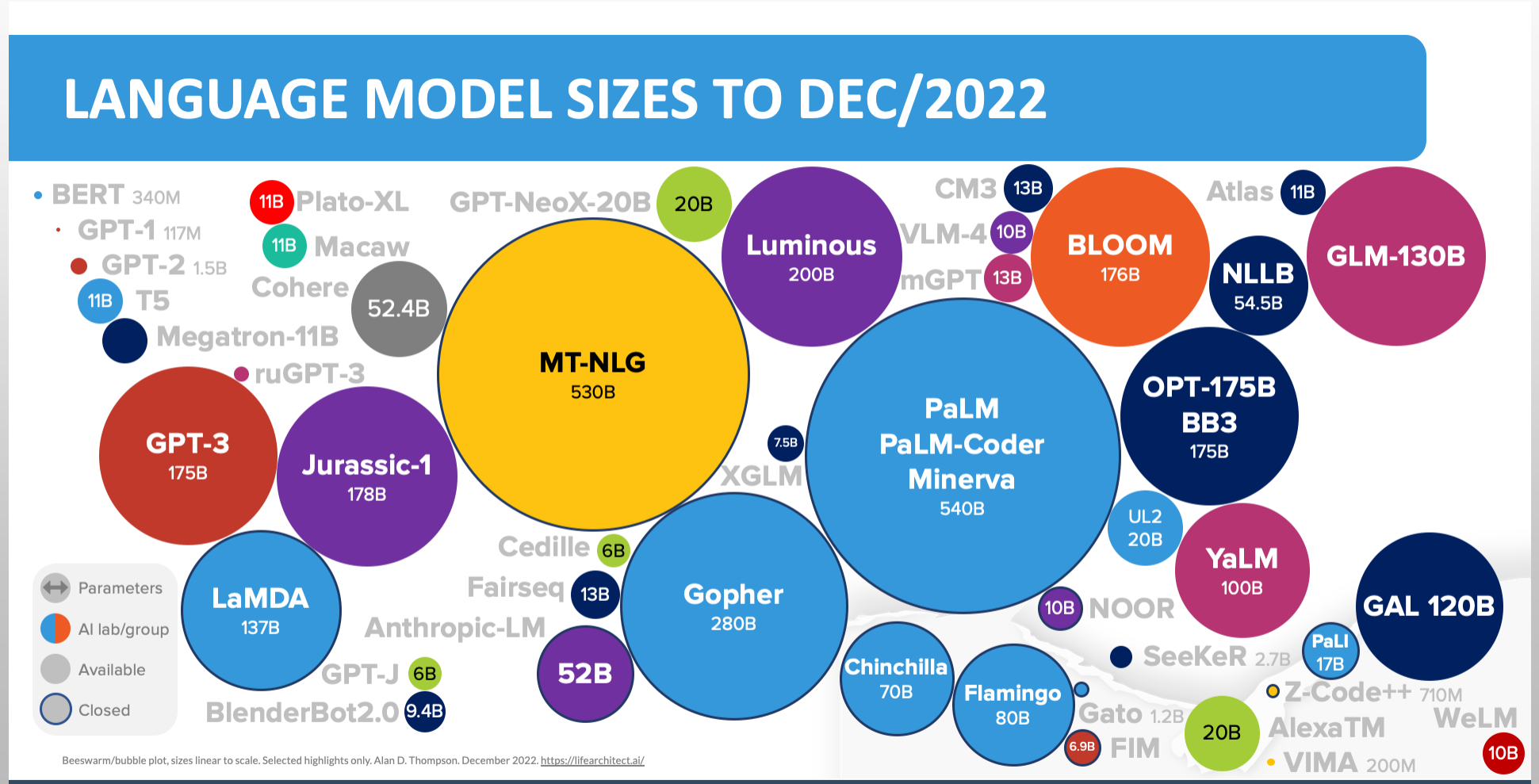
- 2019: OpenAI announces GPT-2
  - A large language model with 1.5 billion parameters, trained on 8 million web pages
  - Generates “convincing” news articles and product reviews (but it cannot write “true” articles)
  - Doesn’t understand what it generates

Language Models

- 2020: OpenAI announces GPT-3
  - A larger language model with 175 billion parameters
  - Evaluated on over two dozen NLP datasets, as well as several novel tasks
  - Scaling laws for neural language models: bigger models are more sample efficient



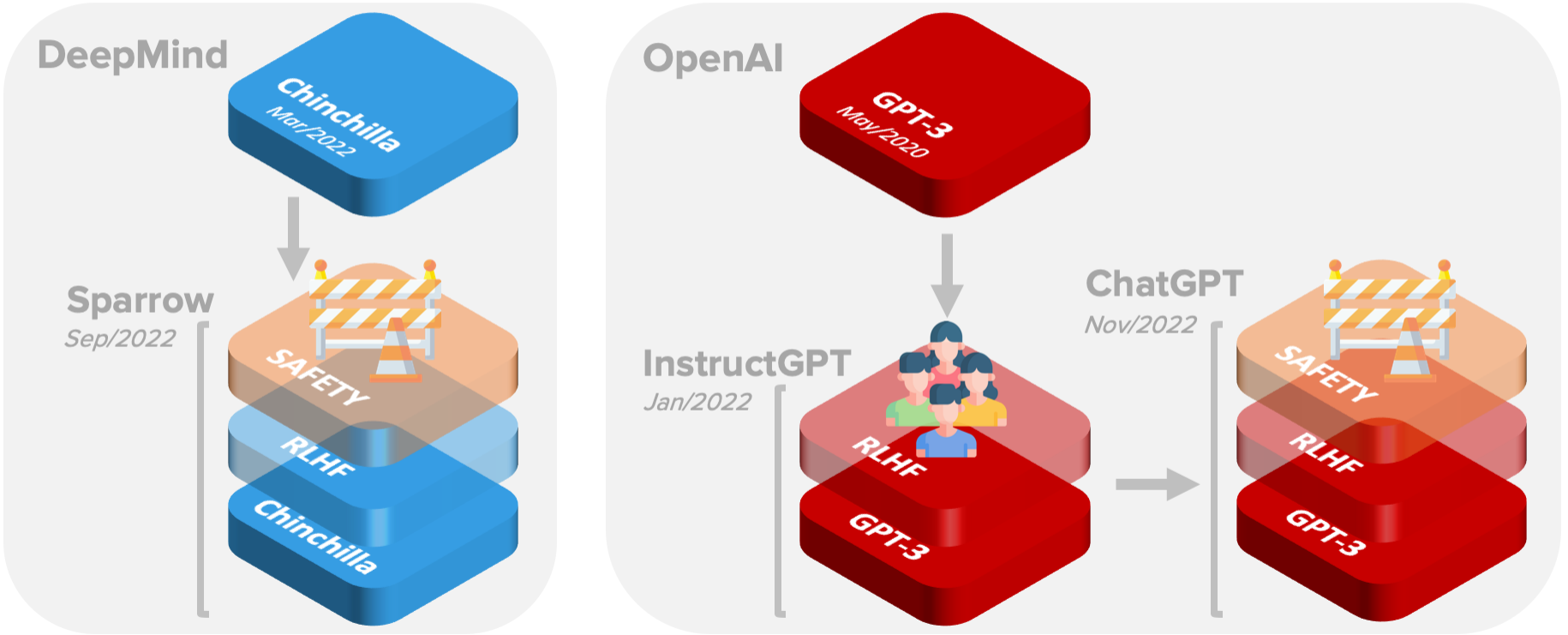
Language Models



LifeArchitect.ai/models

Language Models

CHATGPT VS SPARROW: DIALOGUE MODELS



Not to scale. Alan D. Thompson. December 2022. https://lifearchitect.ai/

LifeArchitect.ai/models