Computational Methods for Linguists

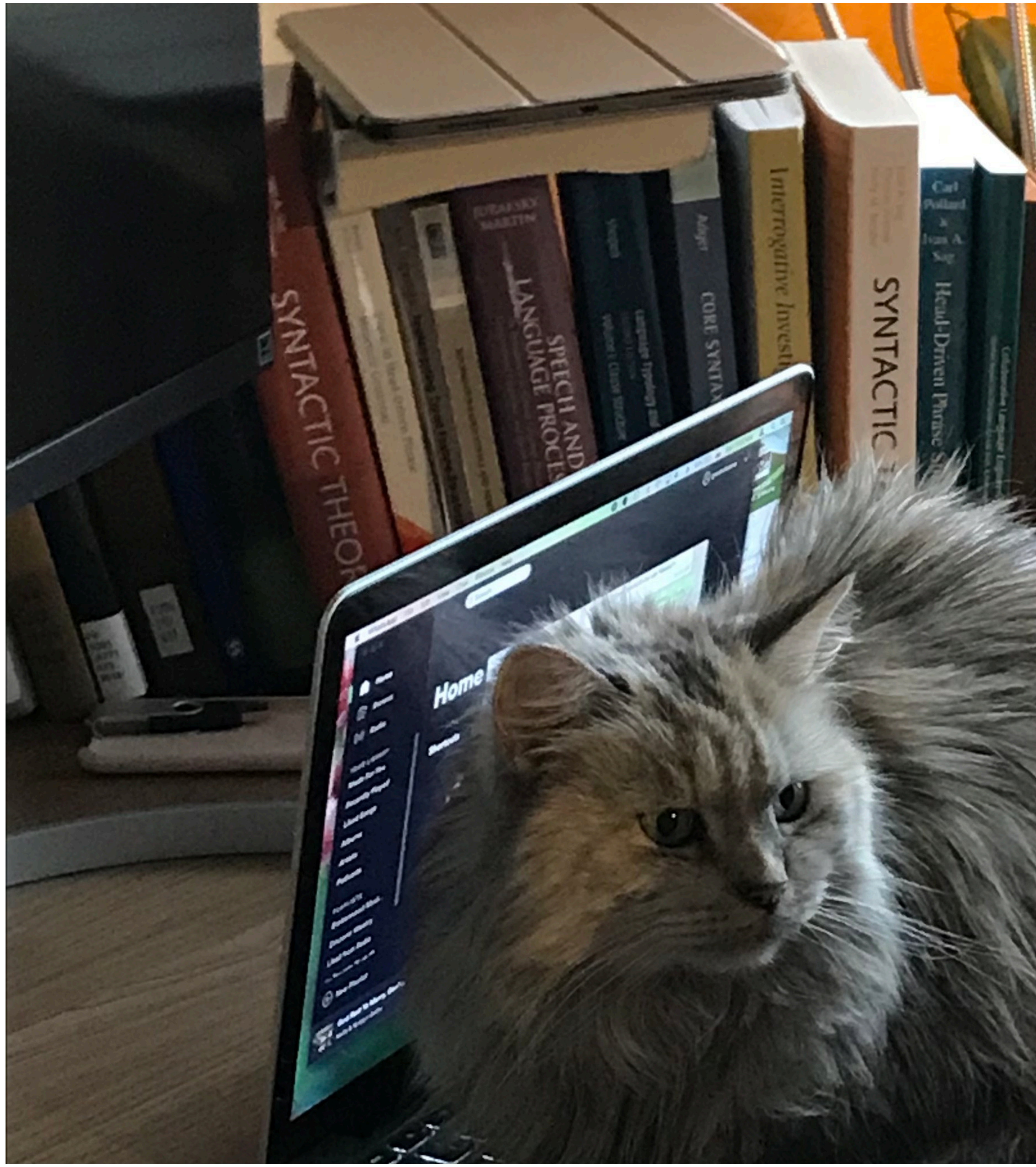
Ling 471

Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)

05/25/21

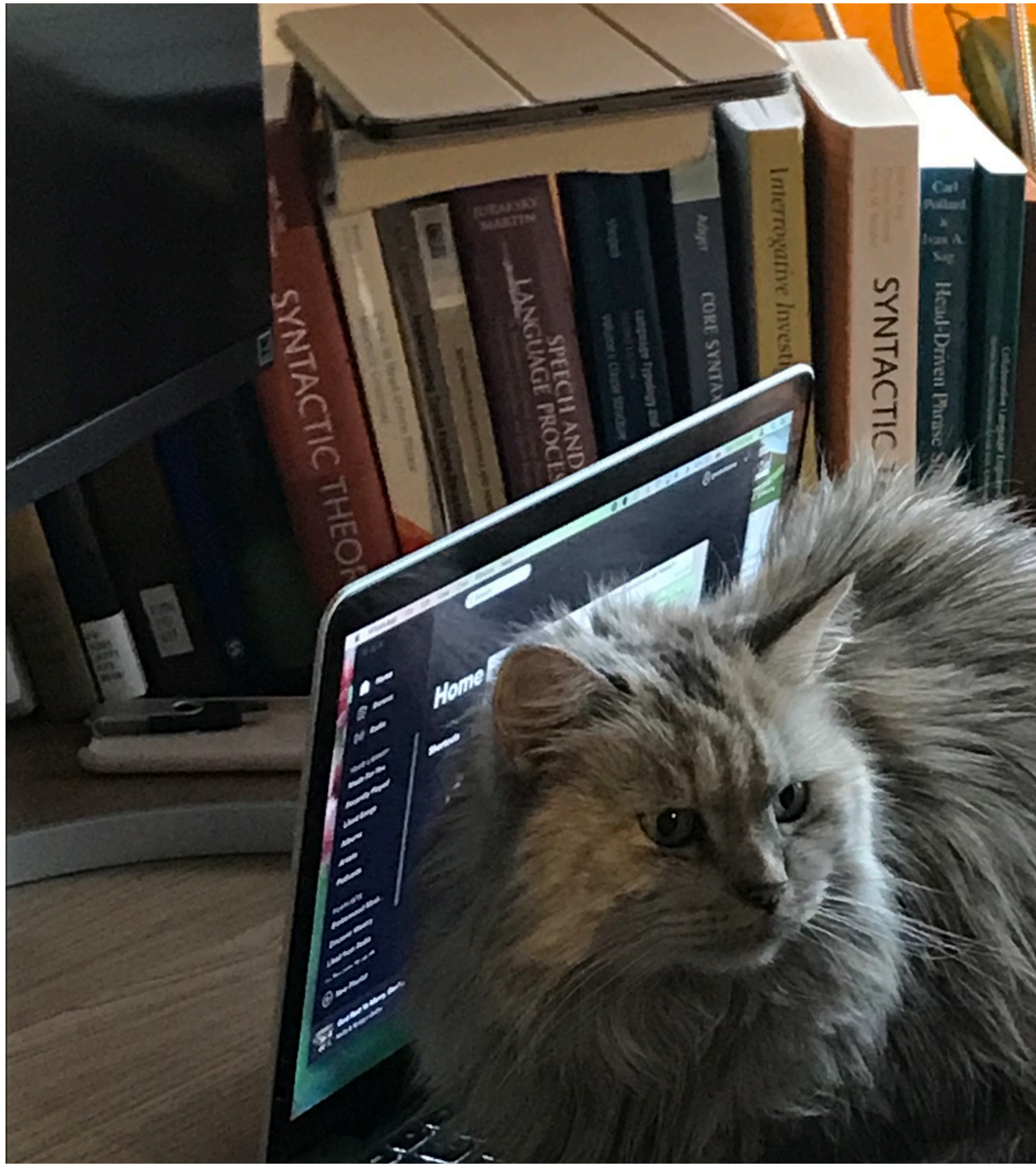
Reminders
and announcements

- HW4 due today
  - Will publish HW5 later today
- Presentation topic **suggestions** due today
  - No late submissions for any of the presentation portions
    - (because otherwise the presentations can't happen on time)



Reminders
and announcements

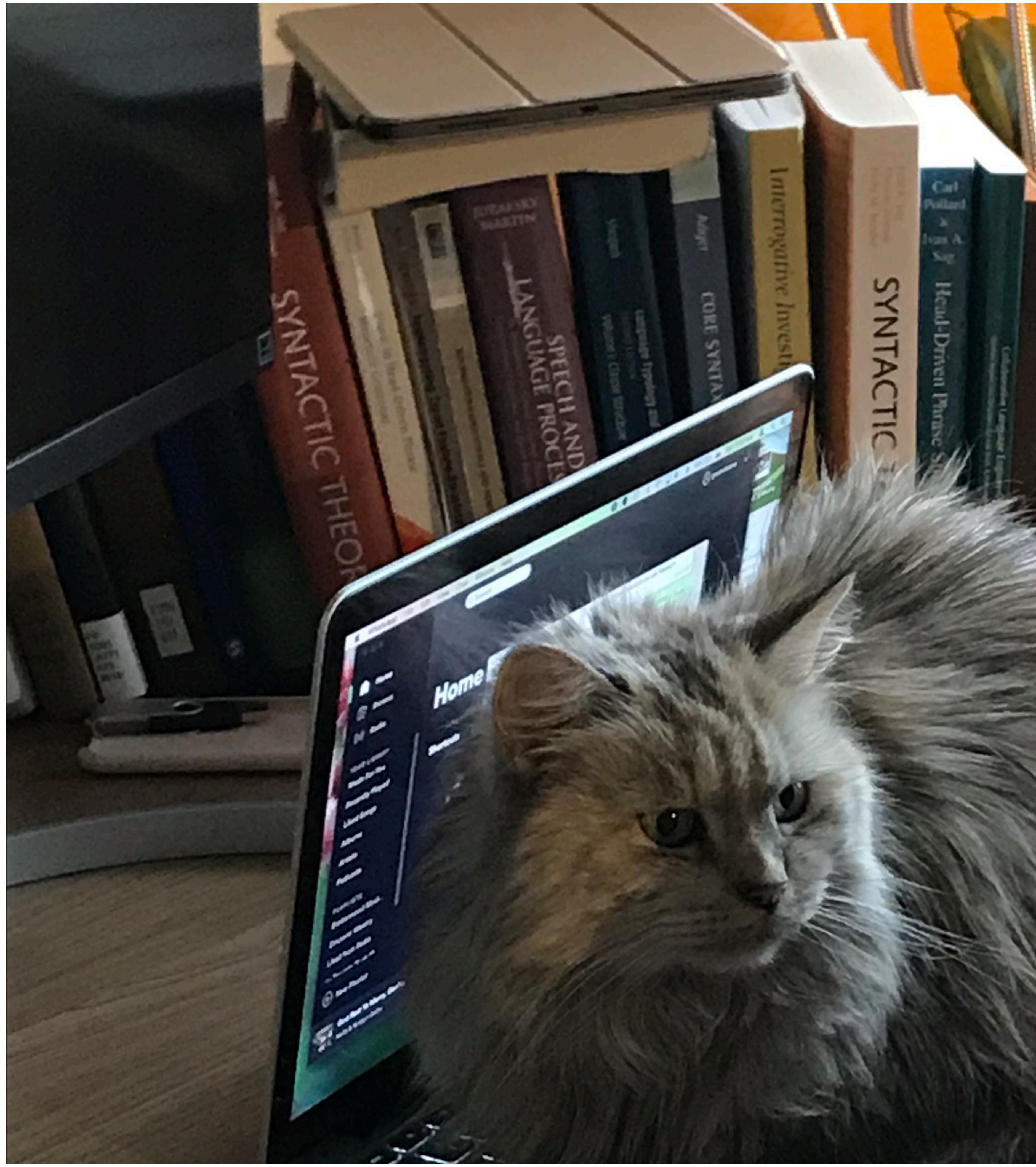
- HW5:
  - I will ask you to do something I didn't teach you
  - We'll help!
  - But it's important to learn to figure out programming stuff that noone showed to you :)
  - other than that, should be just rearranging your old code + plots!



Presentations

Presentations
on Canvas

- **Two** discussion boards
  - Topics
  - Feedback
- **Two** submission areas
  - Prerecorded videos
    - only for those who **chose** prerecorded
  - Final **slides** submissions
    - This one is **officially graded**
    - All others are not graded **but if you miss one of them, you may lose the entire grade, too!**
    - **Strict** deadlines, **no late submissions**
    - Giving feedback to others is kind of optional but very highly encouraged
    - Remember you have participation adjustment…



No late submissions for any of the presentation-related stuff!

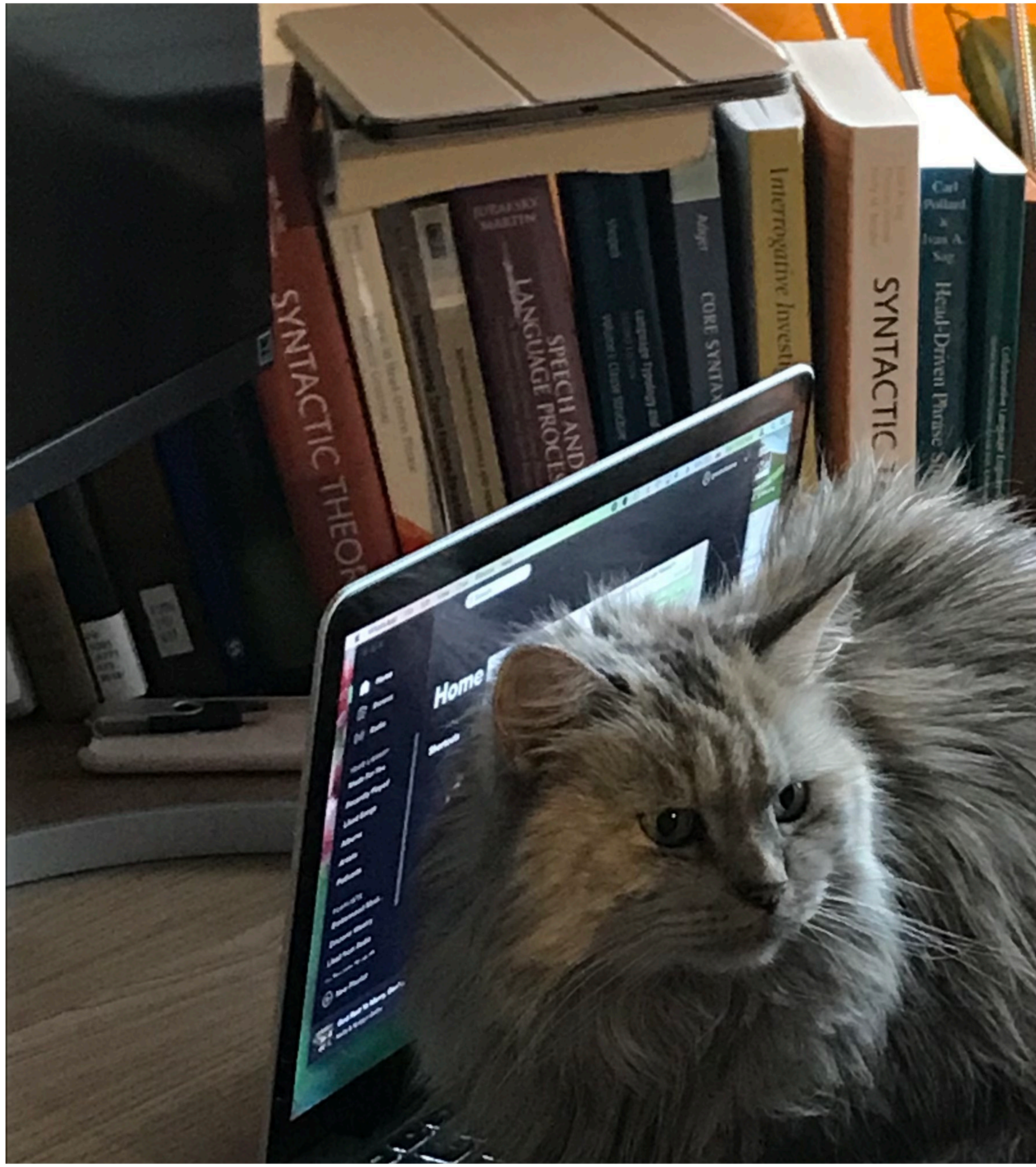
If you miss any presentation-related submission, you may lose 15% of the grade!

Only one presentation-related submission is officially graded but it can’t happen if any of the previous ones is missed!

Presentations

spec

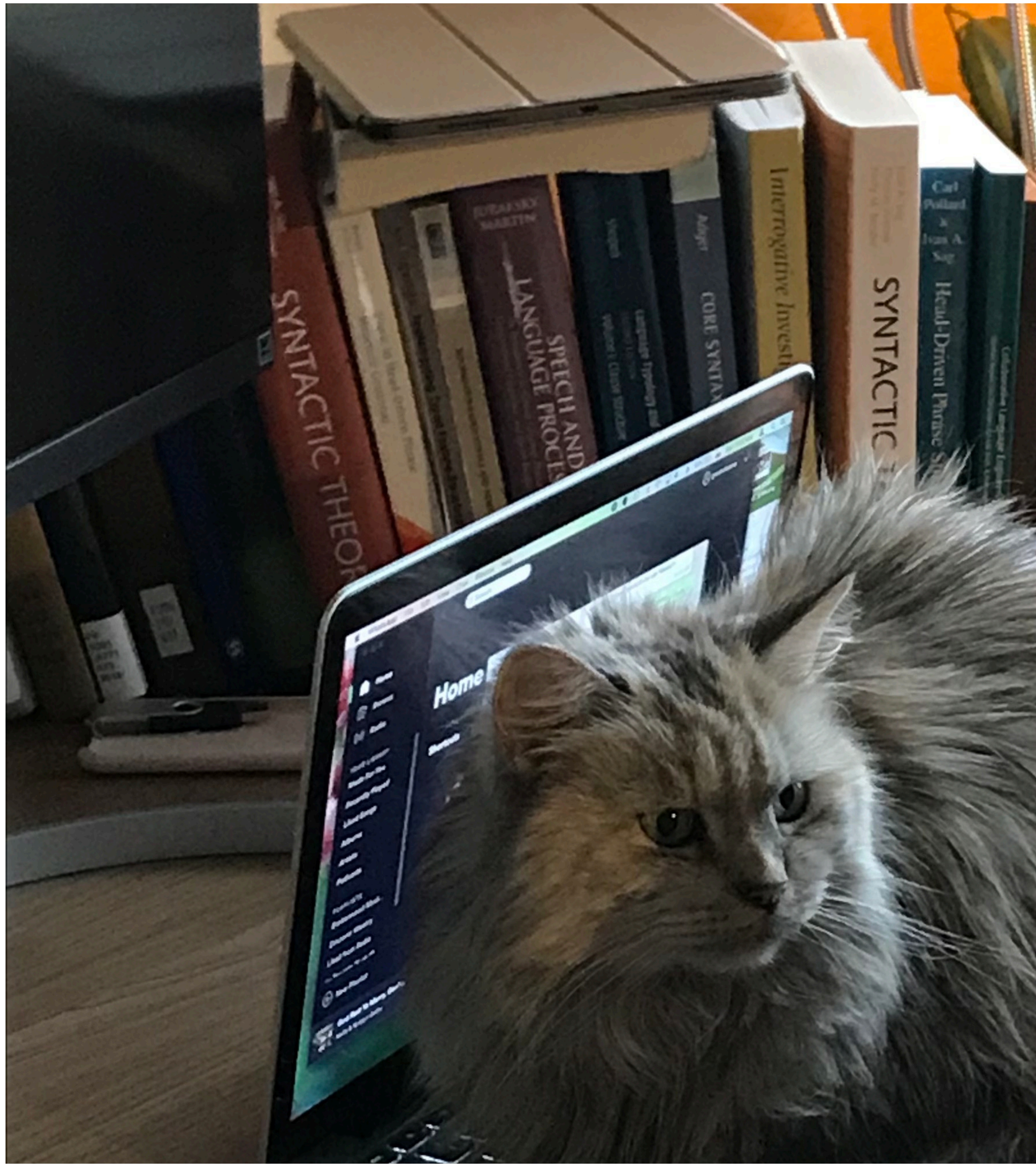
- **5 minutes :**
  - Paper/project **title**, **authors**, **year**, **publication**
  - **what** was done and **why**
  - why this is **interesting**
  - **social impact** (or lack thereof)
  - anything else you want to say
- Recommended: **no more than 3 dense** content slides or 4-5 “**sparse**” ;) slides
  - Avoid dense slides unless presenting bullet by bullet (and even then)
  - But if you have “sparse” slides, make sure to **rehearse** because you don’t want to run out of time.



Presentations

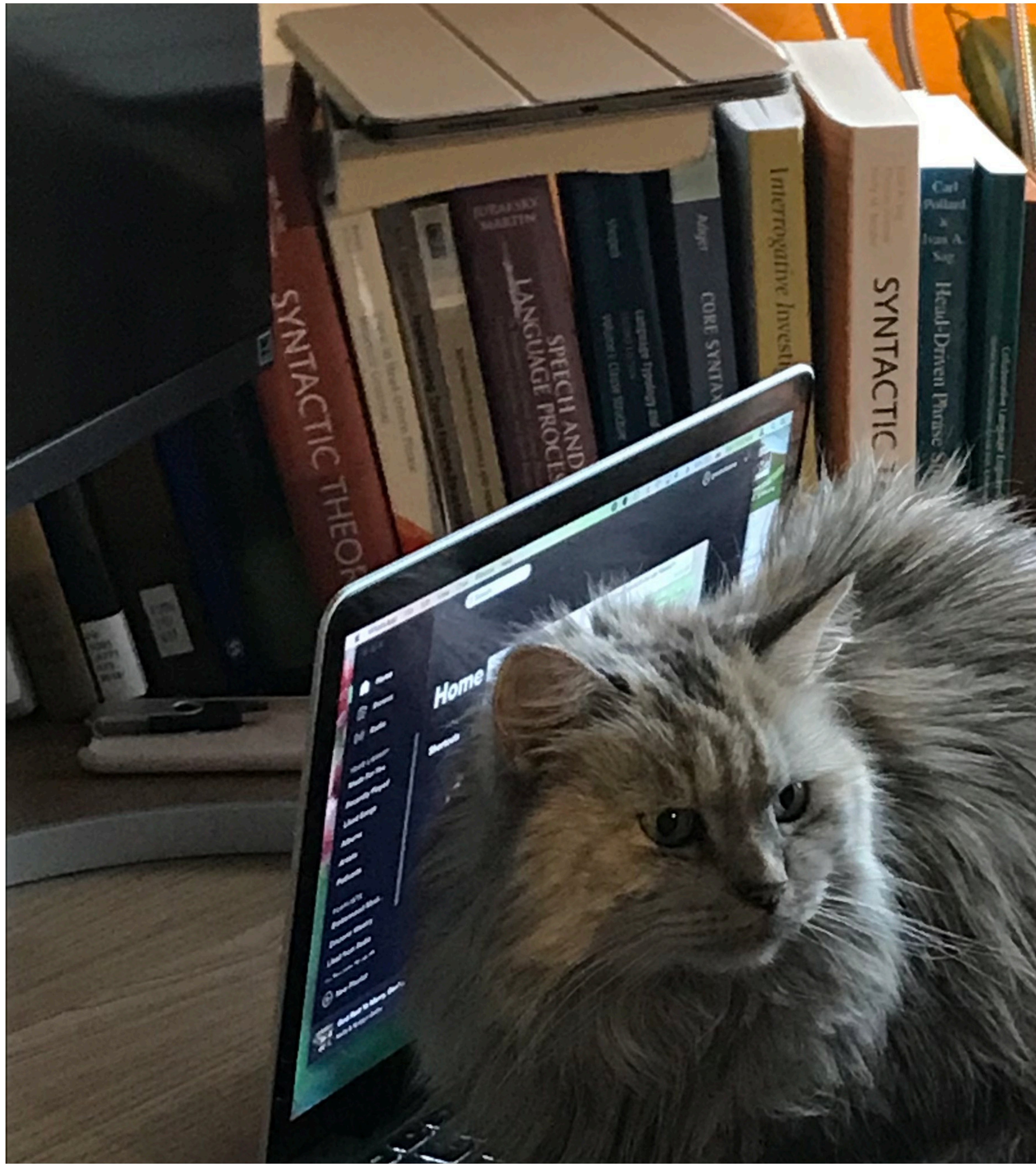
live vs. prerecorded

- Prerecorded: due earlier (May 30)
- Can be played during class with you attending, or not
- If most people want prerecorded and not played during class:
  - Can move the remaining session (one or both) entirely to Canvas discussion board
  - ...that depens on how soon I get the information from everyone :)
- Please fill out the quiz: https://canvas.uw.edu/courses/1465777/quizzes/1452037



Presentations
live

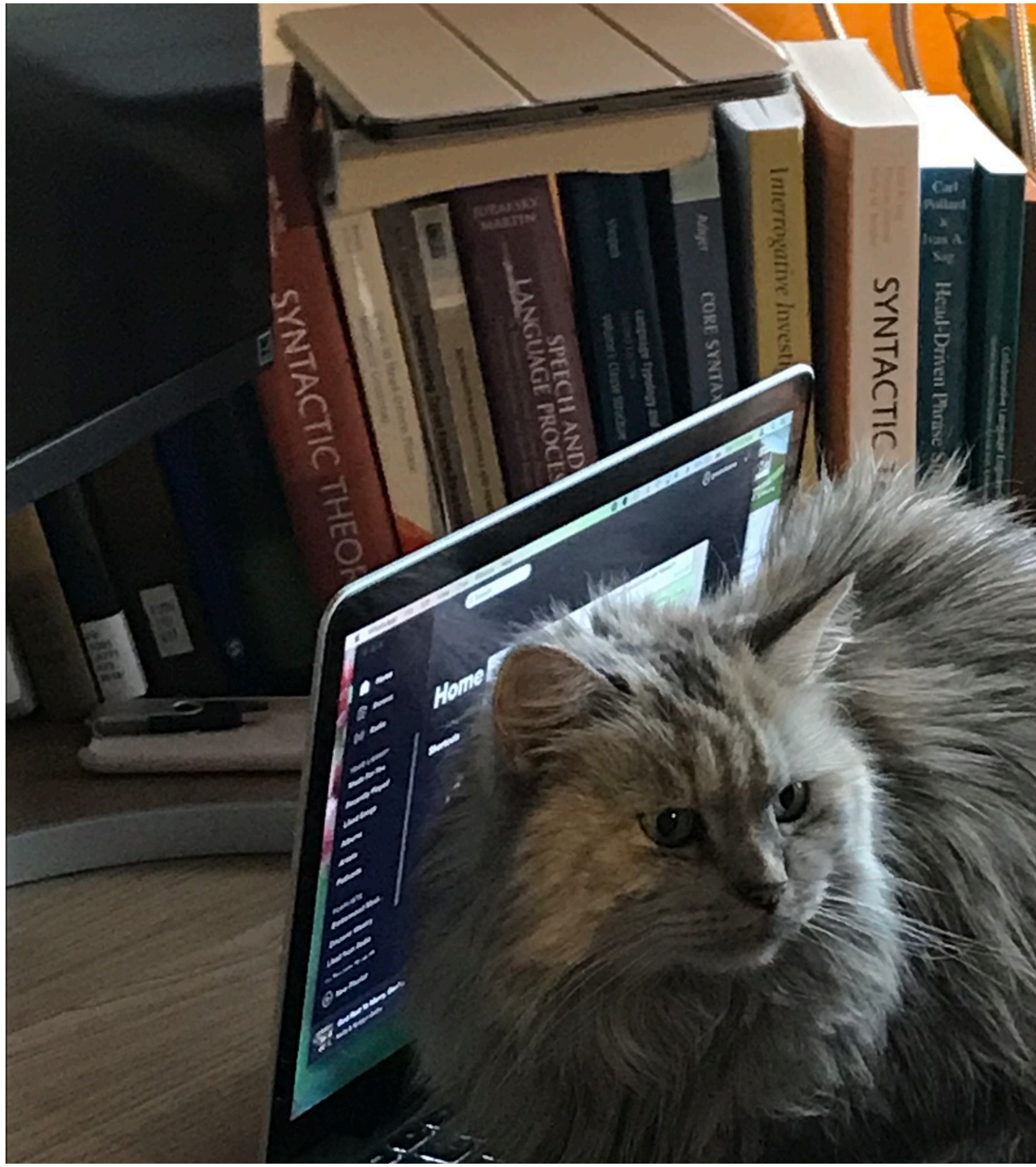
- Canvas quiz: indicate “live”
  - Please do this ASAP but def. **by Friday May 28**
- I will email you your time slot for June 1/3
- Present for 5 minutes + 2-4 minutes questions/feedback
- Take note of feedback during class (or rely on the recording later)
- Monitor your entry on the Feedback discussion board



Presentations

prerecorded

- **Canvas quiz:** indicate “prerecorded”
  - Please do this ASAP but def. **by Friday May 28**
- Record the presentation **by May 30** (mp4) and submit to Canvas area
  - The Canvas area says “assignment does not count towards final grade”
  - That’s because it is addressing the feedback that counts
  - **If we don’t watch** your presentation, you **won’t be able to get** this portion of your grade
  - If you choose prerecorded and don’t submit by May 30, you will **lose 15% of your grade!**
- **Canvas quiz:** indicate whether attending live or not and want the recording played during class:
  - If yes: I will email you your slot. We will all watch live and give feedback
  - If you cannot attend live on June 1 and 3 **or** don’t want the presentation played, I will **post** your recording on **Canvas** for people to watch and we will all be leaving some feedback there


SYNTACTIC THEORY
SPEECH AND LANGUAGE PROCESSING
CORE SYNTAX
Interrogative Investigations
SYNTACTIC
Head-Driven Phrase Structure
Home

Plan for today

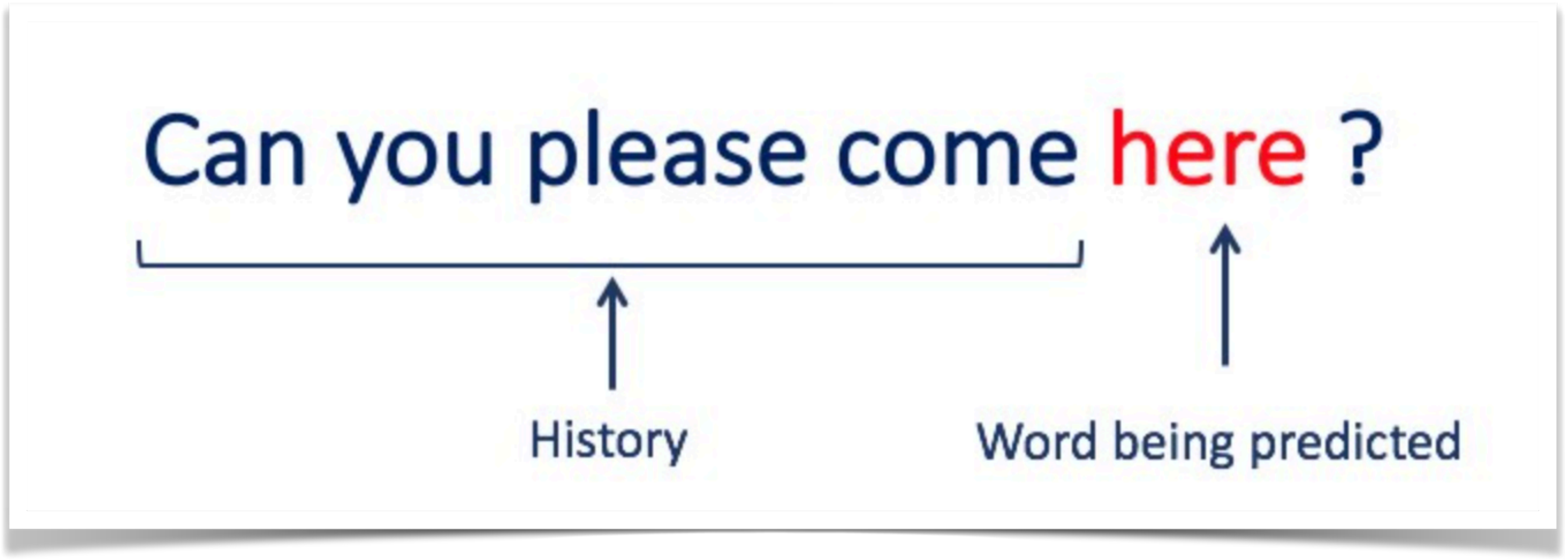
- How are neural models used?
- Class content recap
- Data science and linguistic corpora
- Linguistics and data science
- Presentation procedure

Language models:
How are they used?

Language Models

How are they used?

- Neural and N-gram LMs:
  - output the probability of a **word** given **context**
  - Why is this interesting?
    - “Probability of text”
      - P(“Can you please come here”) is **high**
    - You can do language generation!
    - Neural: similar words have similar vectors
      - **Generalization** over context!
      - Translation, summarization, sentiment…

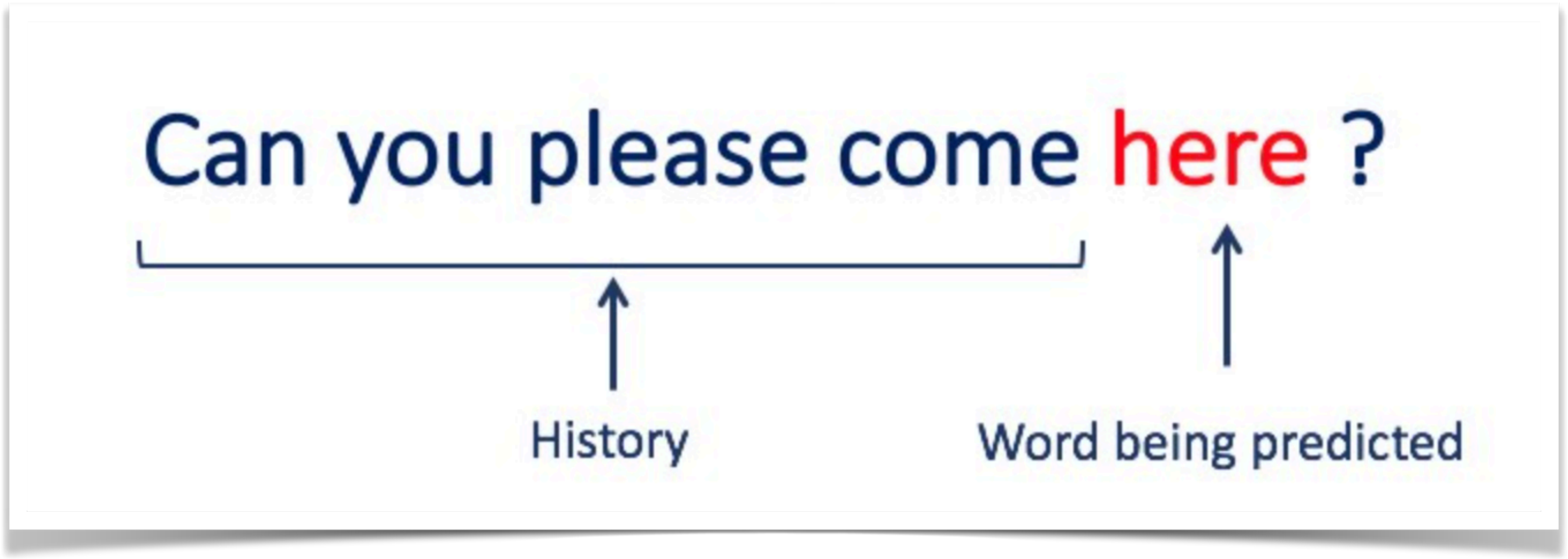


https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/

Language Models

How are they used?

- How to apply to e.g. our IMDB **classification** task?
  - Can use probabilities to get a probability of the entire text
    - ...given a label: P(review|POS)
      - find P(POS|review)
      - that's e.g. **Naive Bayes** which relies on N-grams
  - Neural: Can **adapt** the net to **serve** as a classifier
    - by modifying the **output layer**
    - point is, you still compute the **P(review/text)**
- **Bottom line:** The net trains probabilities of words given contexts
  - on top of this, you can stick any number of classifiers/ technics which rely on probabilities

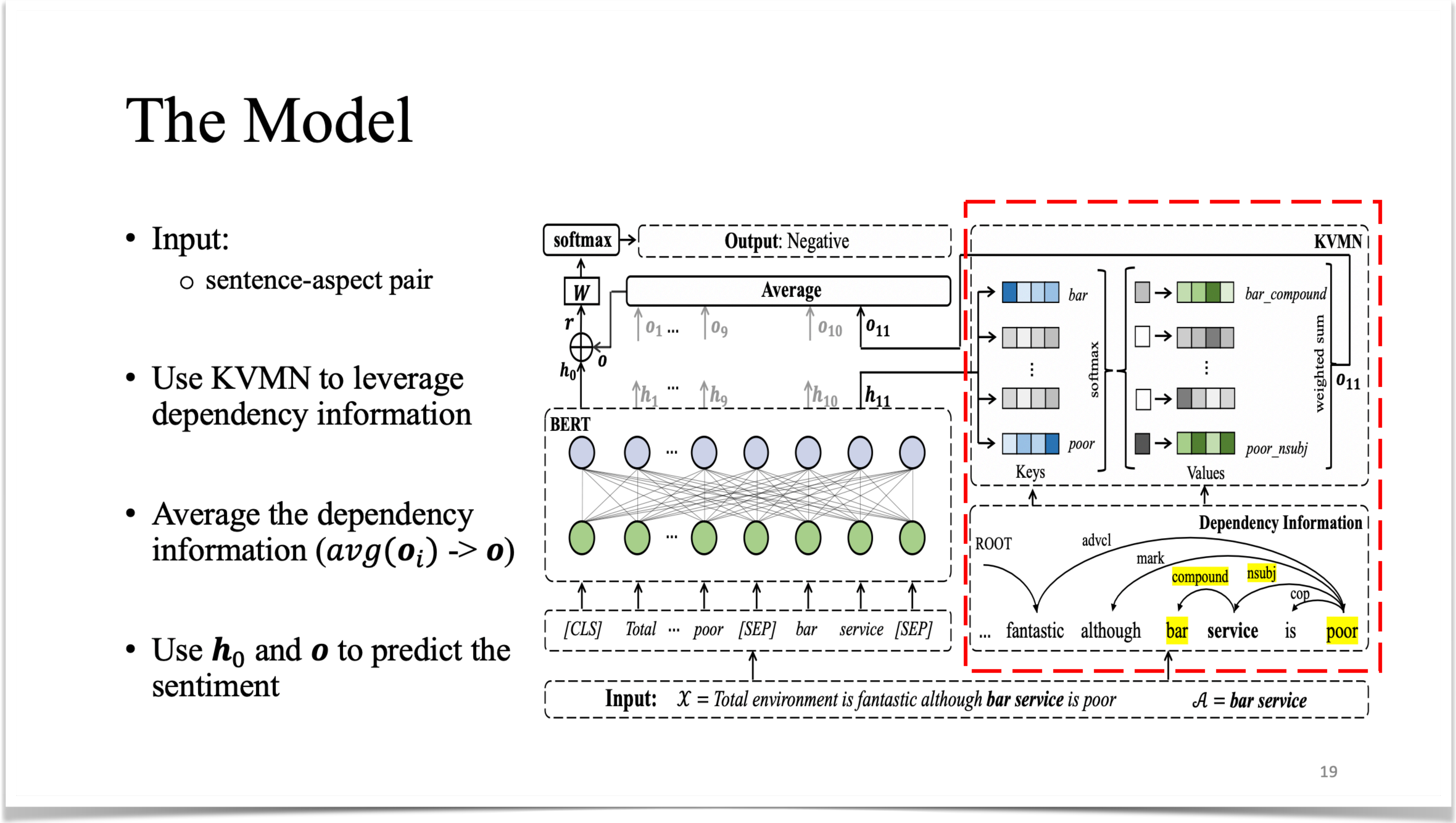


https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/

Language models

How are they used?

- Our own **Yuanhe**:
  - Encode **syntactic** info
  - **Integrate** it into a neural net
  - Become **more accurate** in **nuanced** sentiment analysis
- Yuanhe's talk is available in recording
  - it is quite **accessible**
  - Access info posted on Canvas
    - https://canvas.uw.edu/courses/1465777/discussion_topics/6208885



The Model

- Input:
  - sentence-aspect pair
- Use KVMN to leverage dependency information
- Average the dependency information ($avg(\boldsymbol{o}_i)$ -> $\boldsymbol{o}$)
- Use $\boldsymbol{h}_0$ and $\boldsymbol{o}$ to predict the sentiment

softmax
Output: Negative
KVMN
W
Average
r
h_0
o
o_1 ... o_9 o_{10} o_{11}
h_1 ... h_9 h_{10} h_{11}
BERT
[CLS] Total ... poor [SEP] bar service [SEP]
bar
poor
Keys
softmax
bar_compound
poor_nsubj
Values
weighted sum
o_{11}
Dependency Information
ROOT
advcl
mark
compound
nsubj
cop
... fantastic although bar service is poor
Input: $\mathcal{X}$ = Total environment is fantastic although bar service is poor
$\mathcal{A}$ = bar service

19

Yuanhe Tian's talk in UW Linguistics Treehouse Lab on May 21 2021
Tian, Y., Chen, G., & Song, Y. (2021)

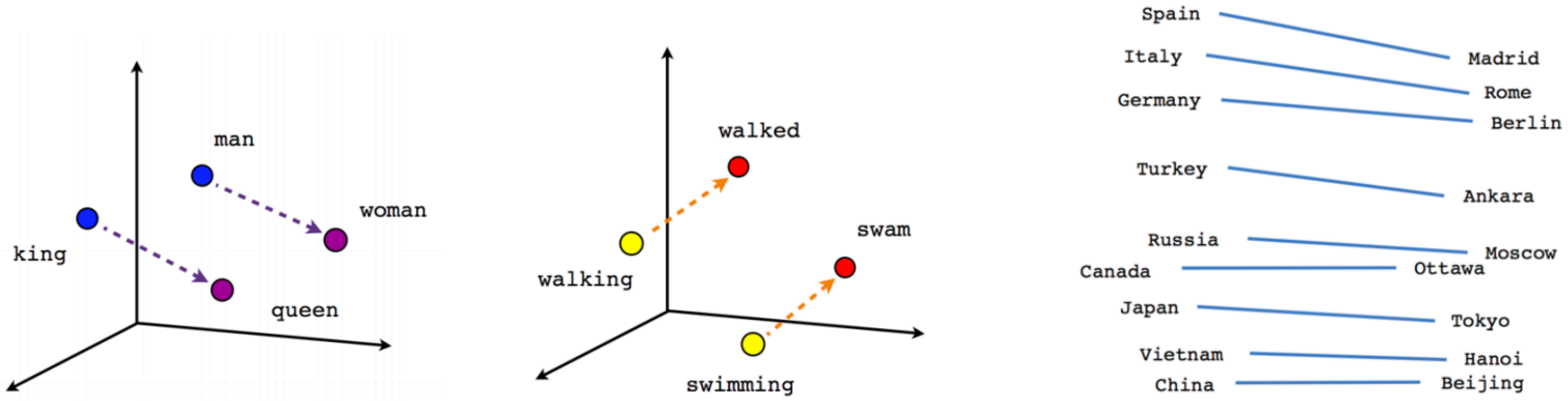
Many things (including people) can be represented as a vector

...and then similarity between them will be measurable :)

Neural models

how are they used?

- Everything can be a **vector**:
  - Word vectors
  - Text vectors
  - Content vectors
  - People vectors
  - Behavior vectors
- What to do with them:
  - Obtain (**train**)
  - Measure **similarity**
  - **Predict** most probable outcomes
    - even if you haven't seen exactly the same features in training!



Male-Female

Verb tense

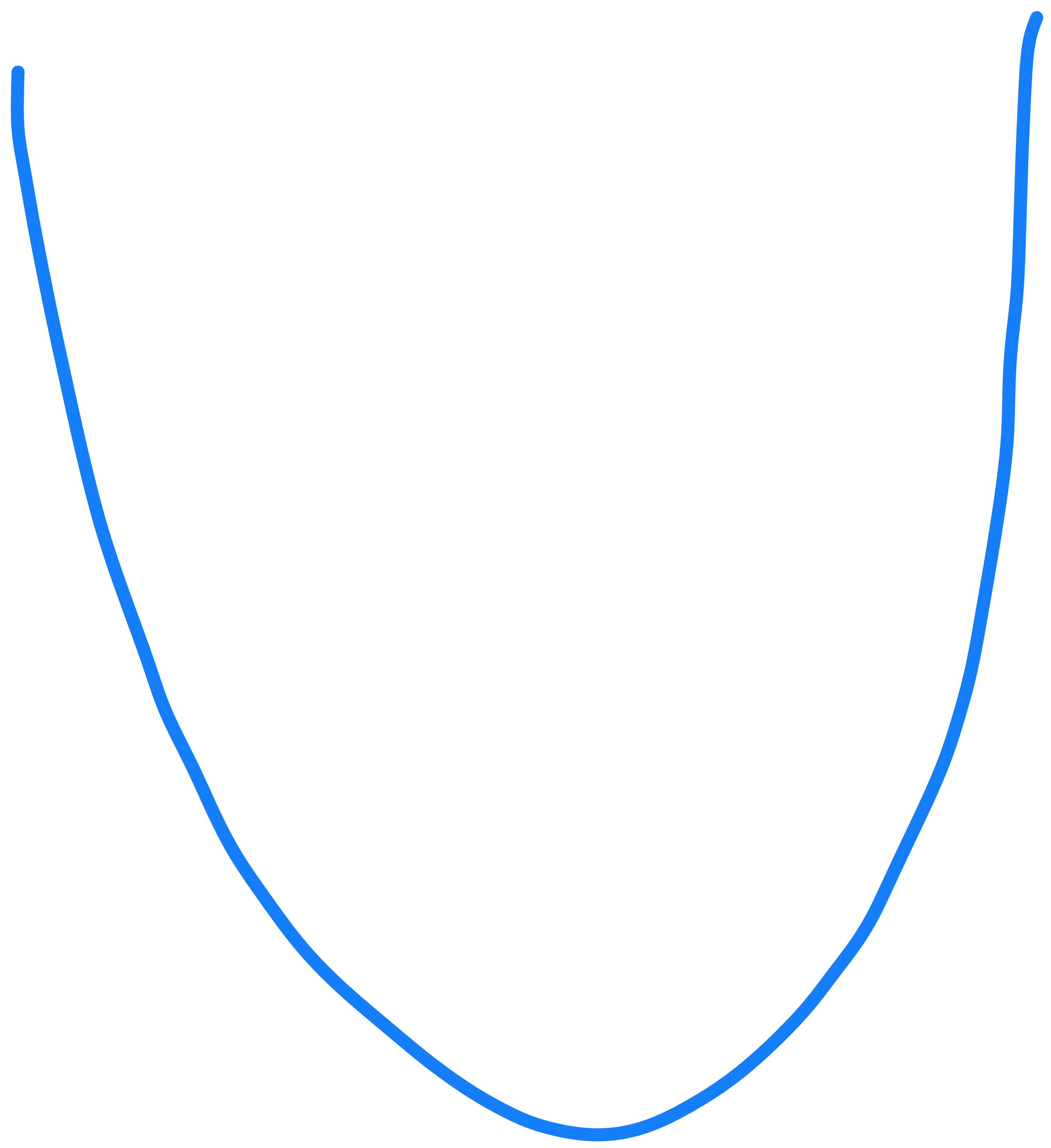
Country-Capital

https://towardsdatascience.com/creating-word-embeddings-coding-the-word2vec-algorithm-in-python-using-deep-learning-b337d0ba17a8

Class recap

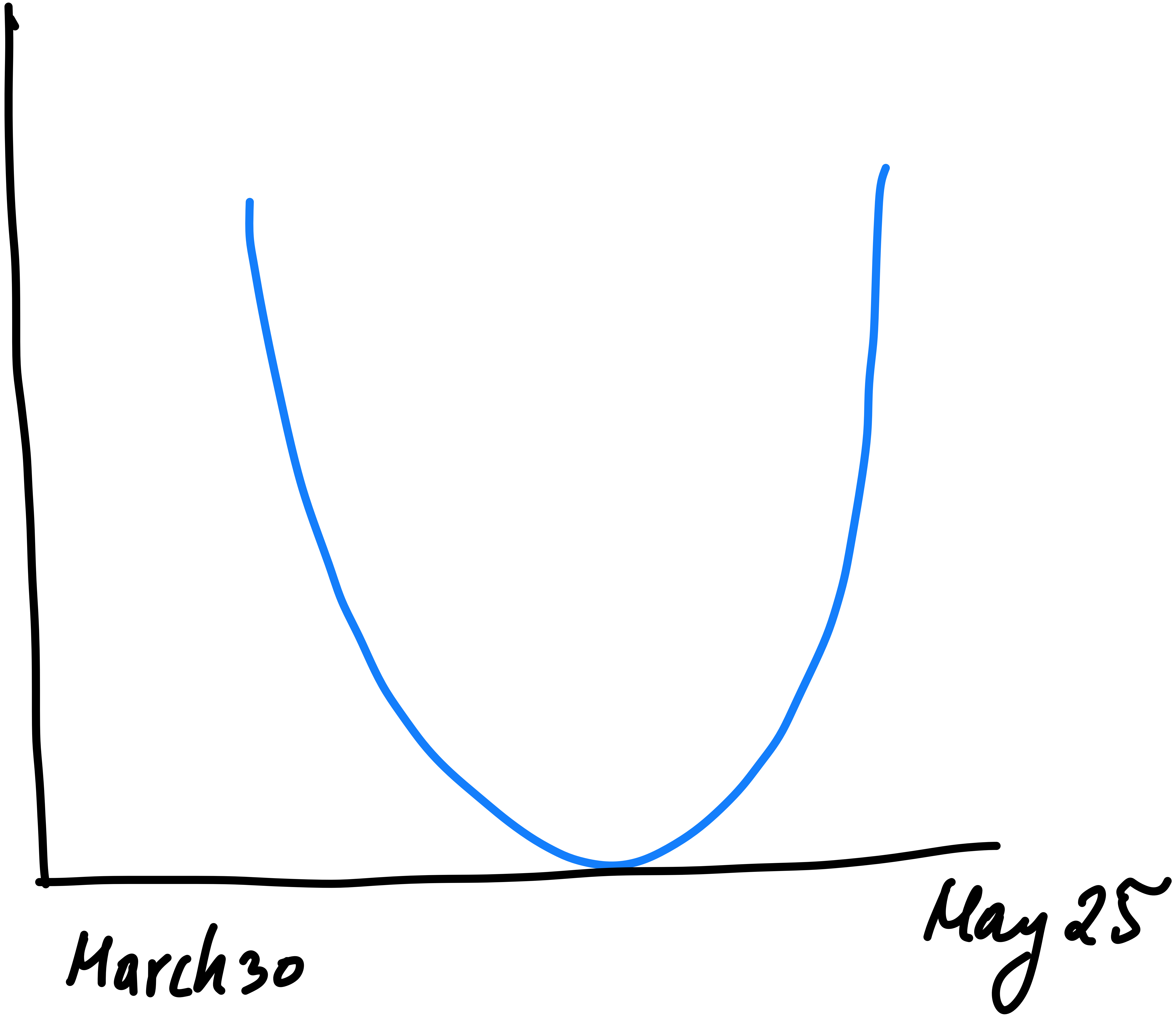
We have come a long way!

- This is a function y = x^2
- It is also the shape of our class :)



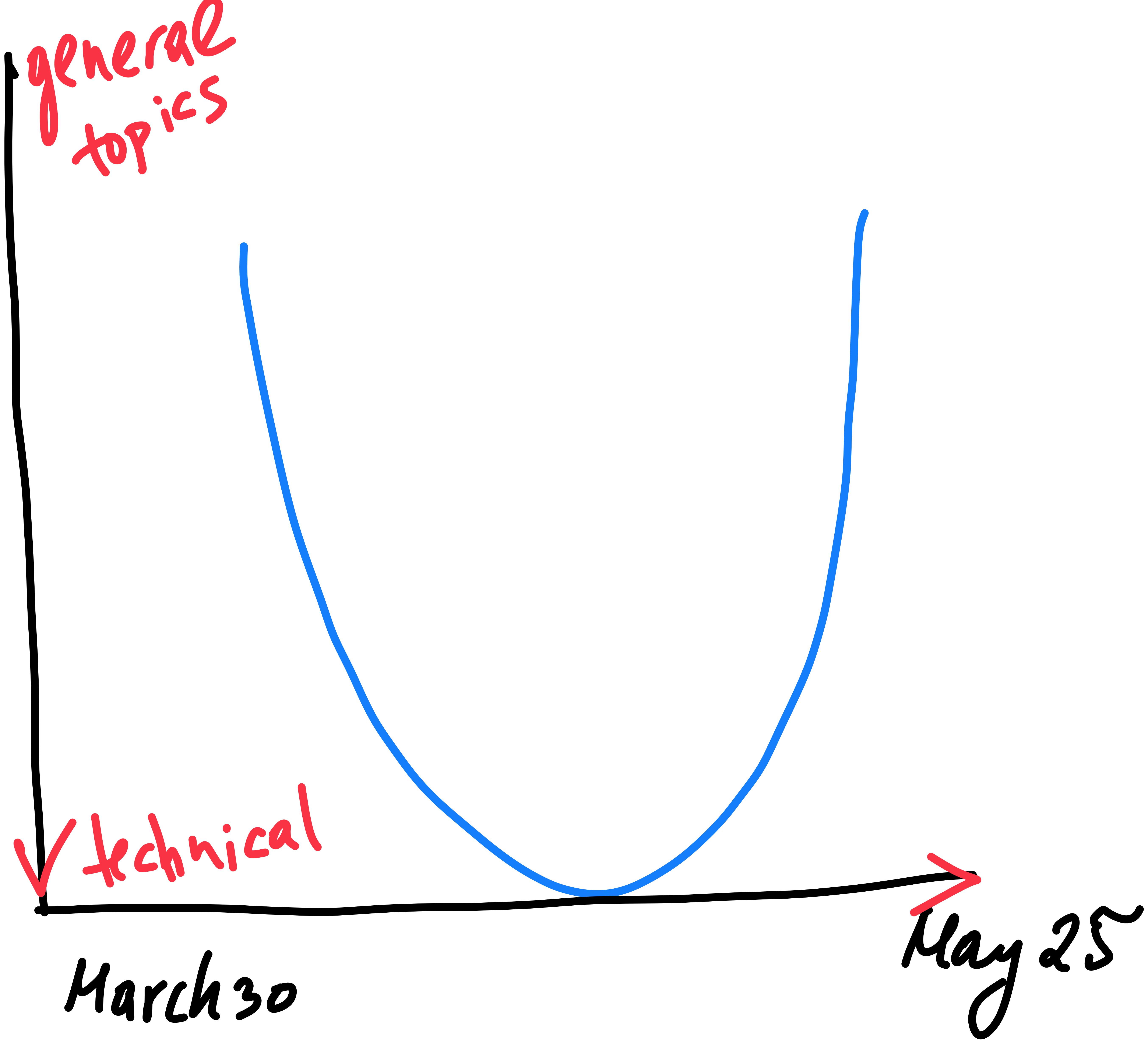
We have come a long way!

- This class content was U-shaped



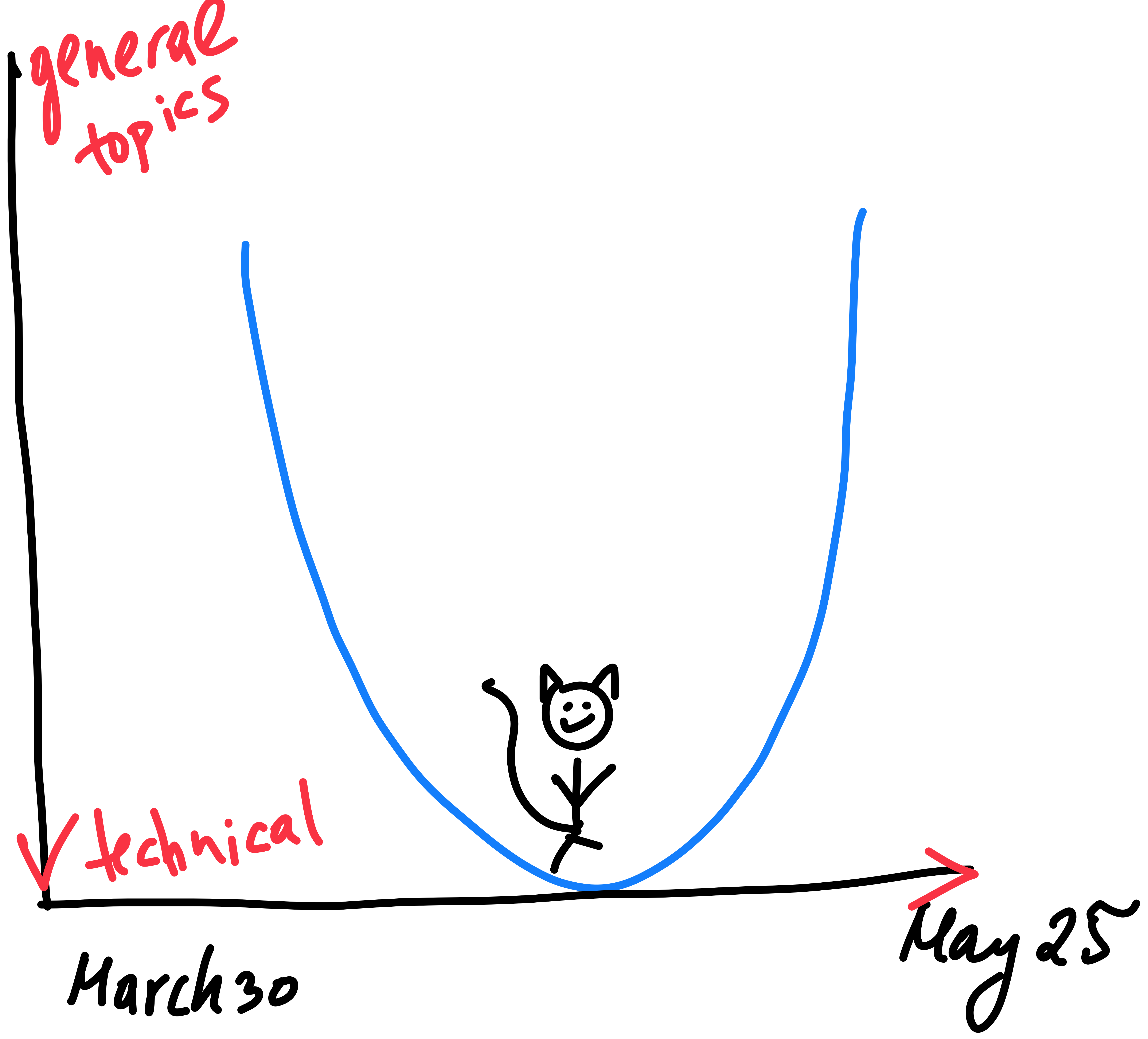
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world



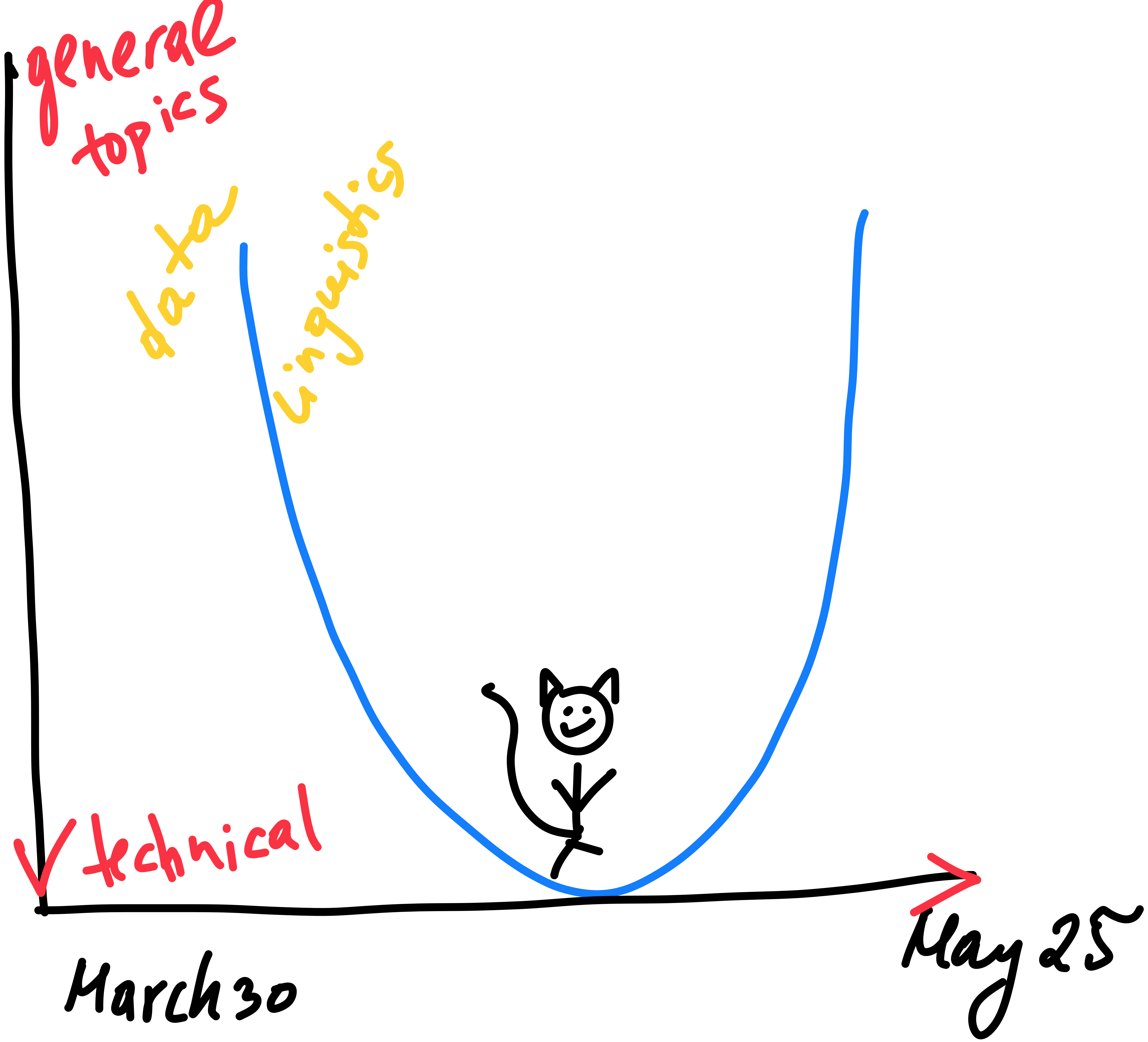
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we’ve been mastering a variety of **technical** concepts and skills



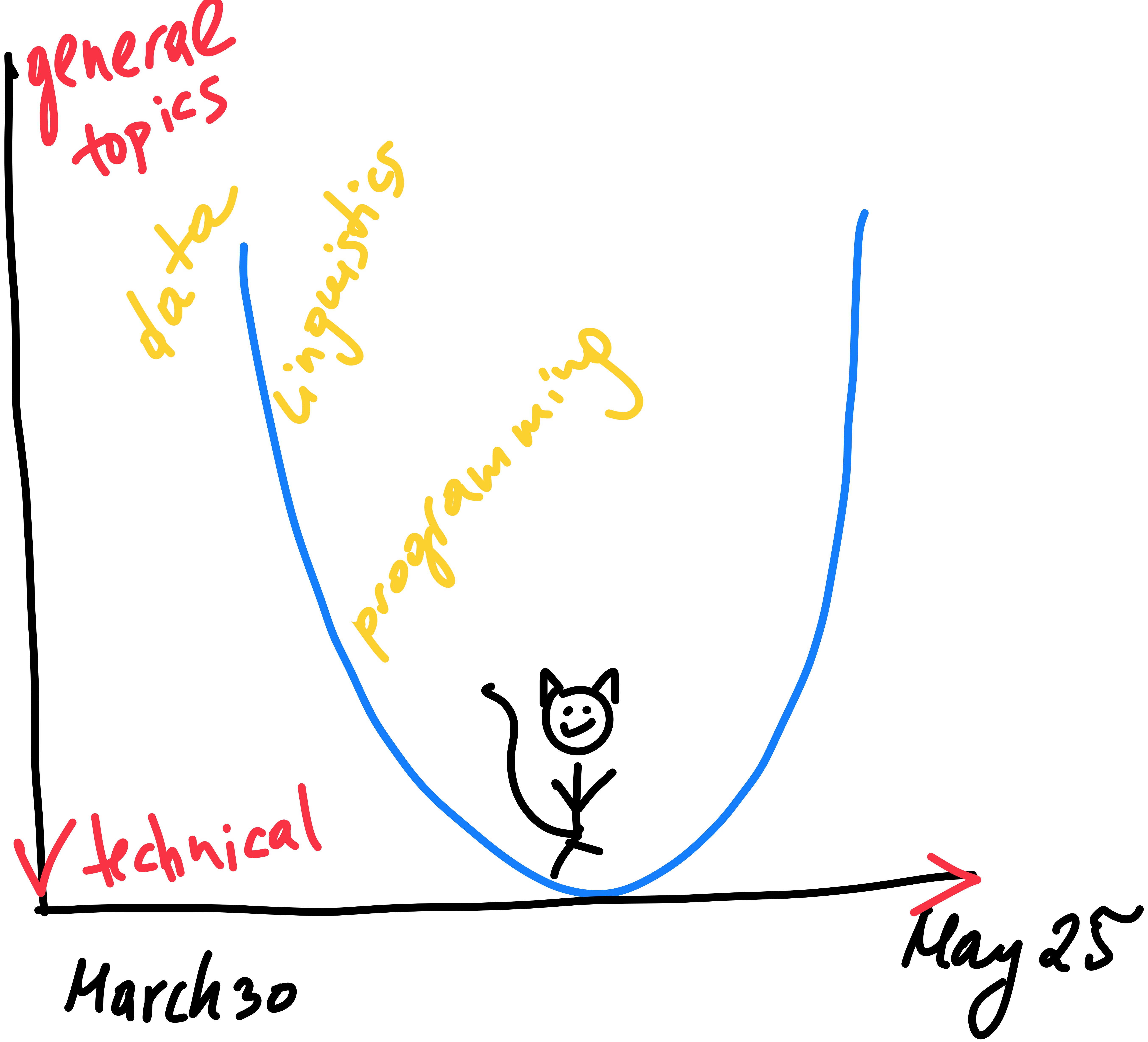
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we've been mastering a variety of **technical** concepts and skills



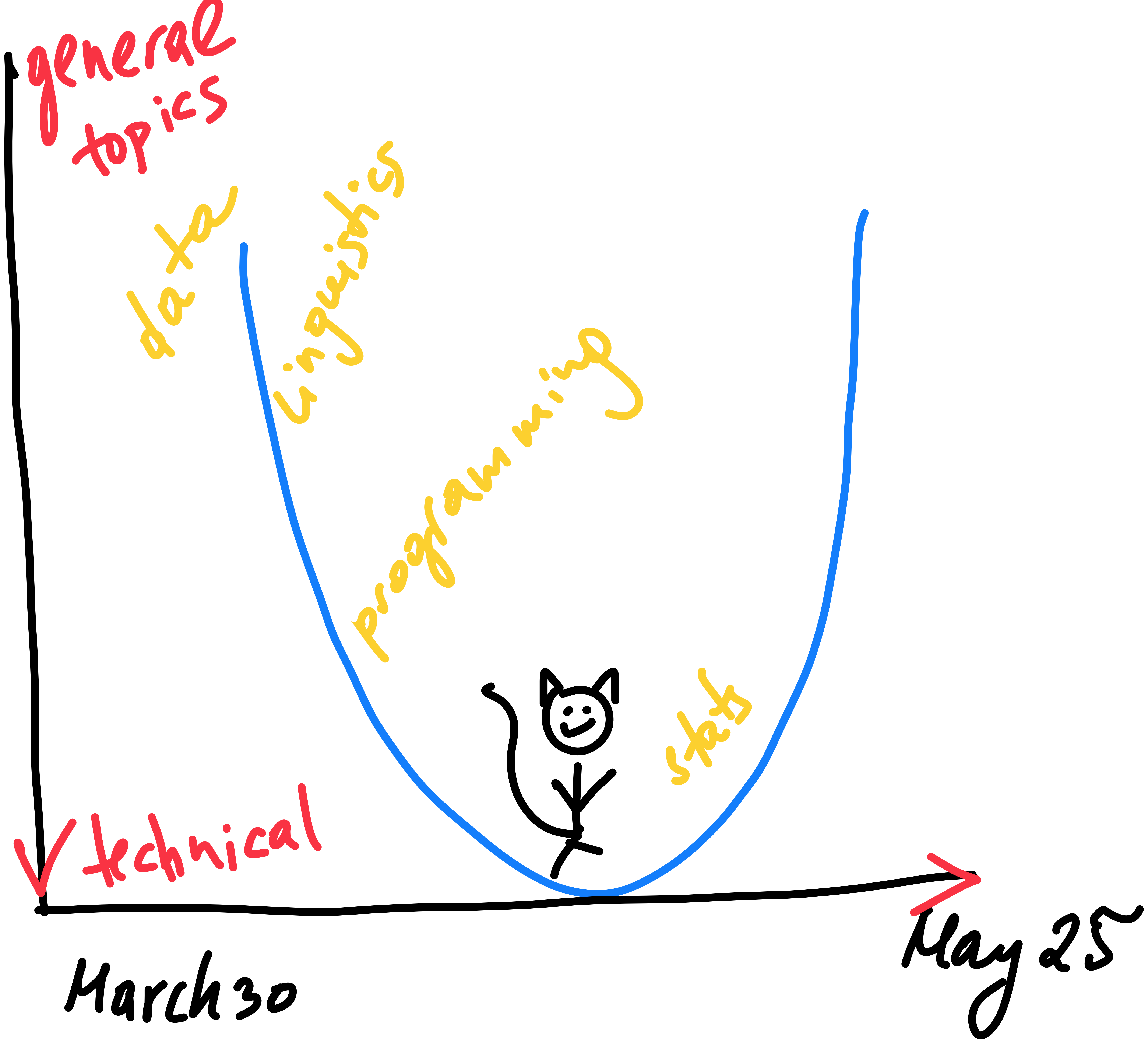
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we've been mastering a variety of **technical** concepts and skills



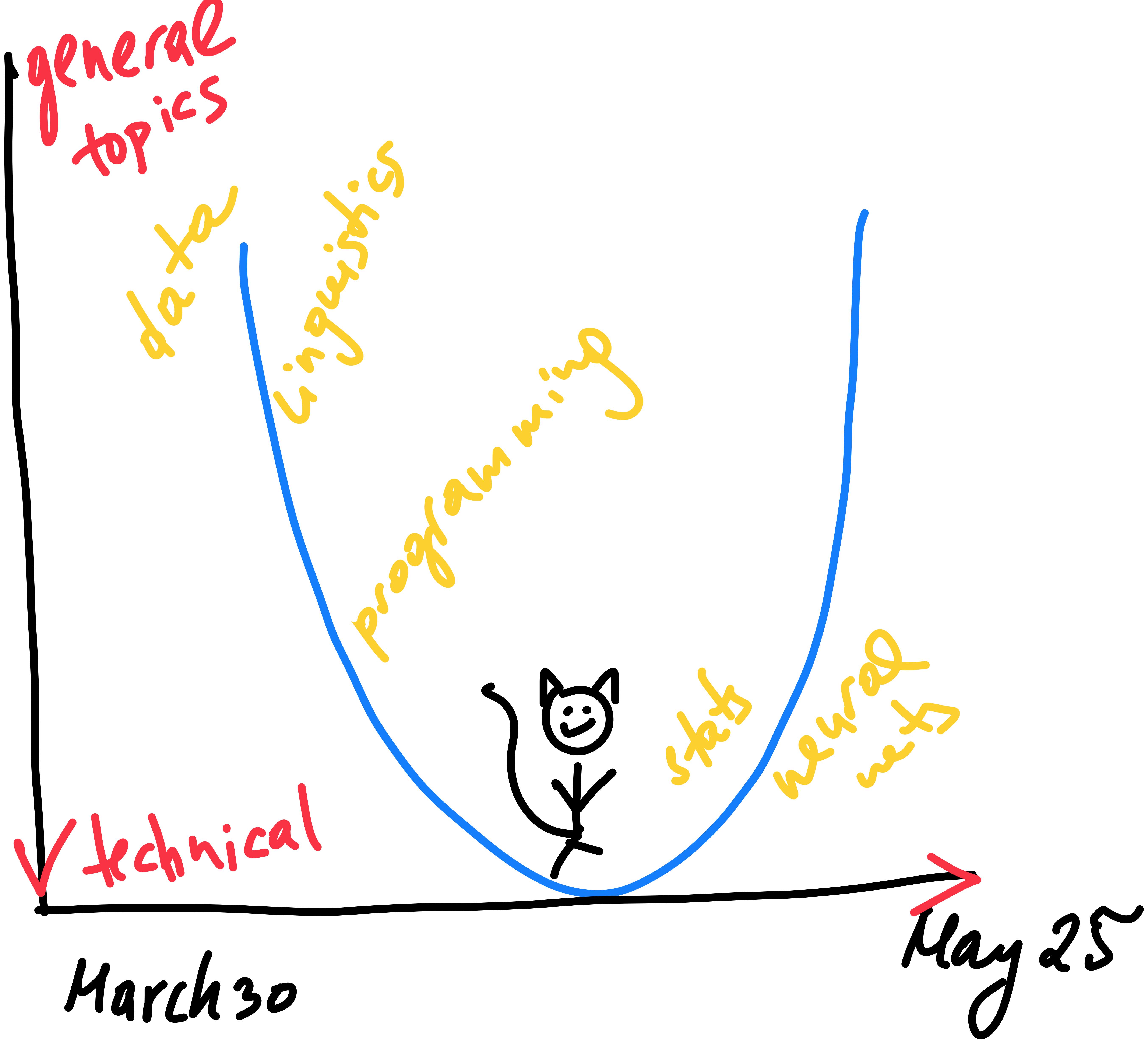
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we've been mastering a variety of **technical** concepts and skills



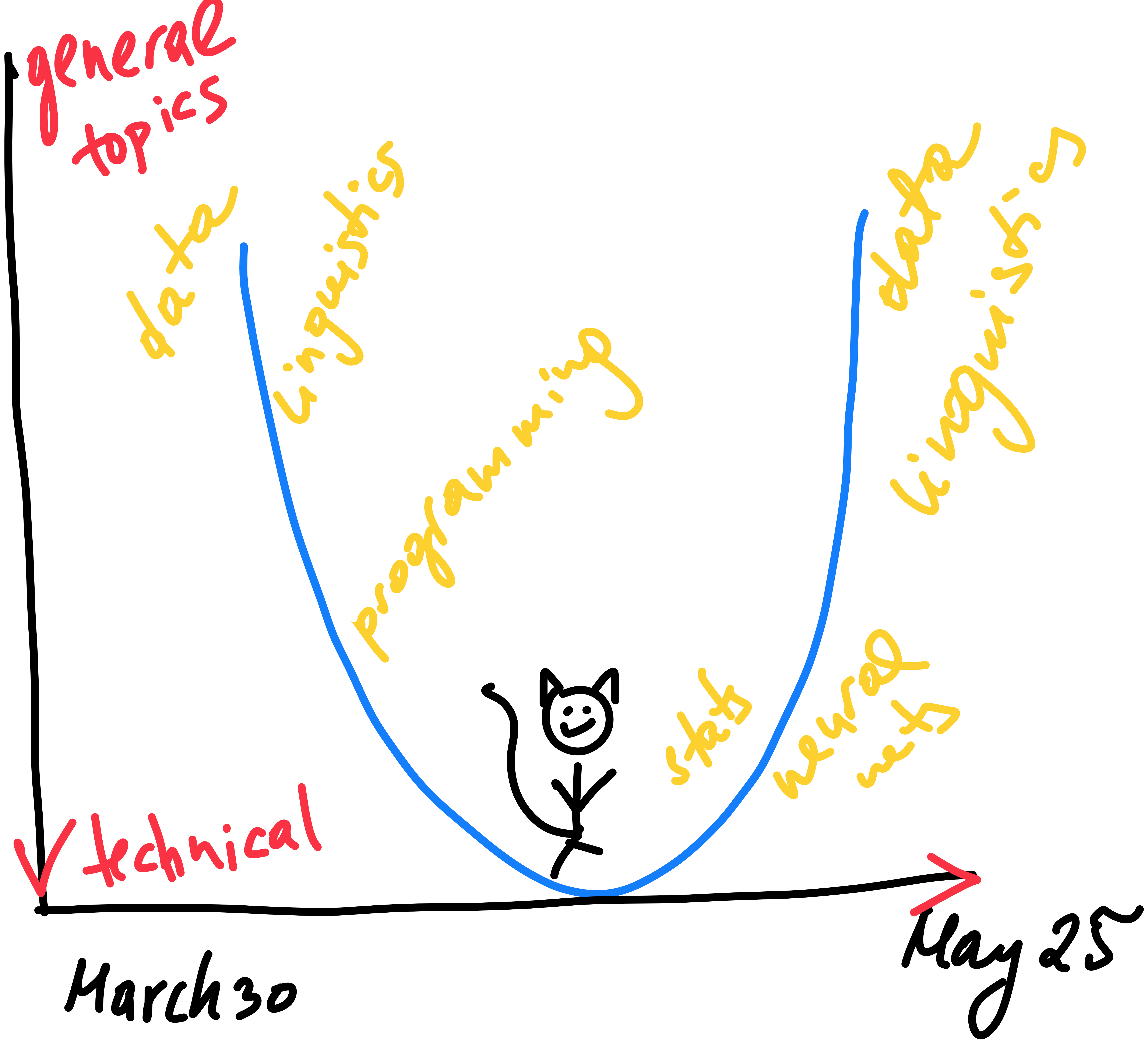
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we've been mastering a variety of **technical** concepts and skills



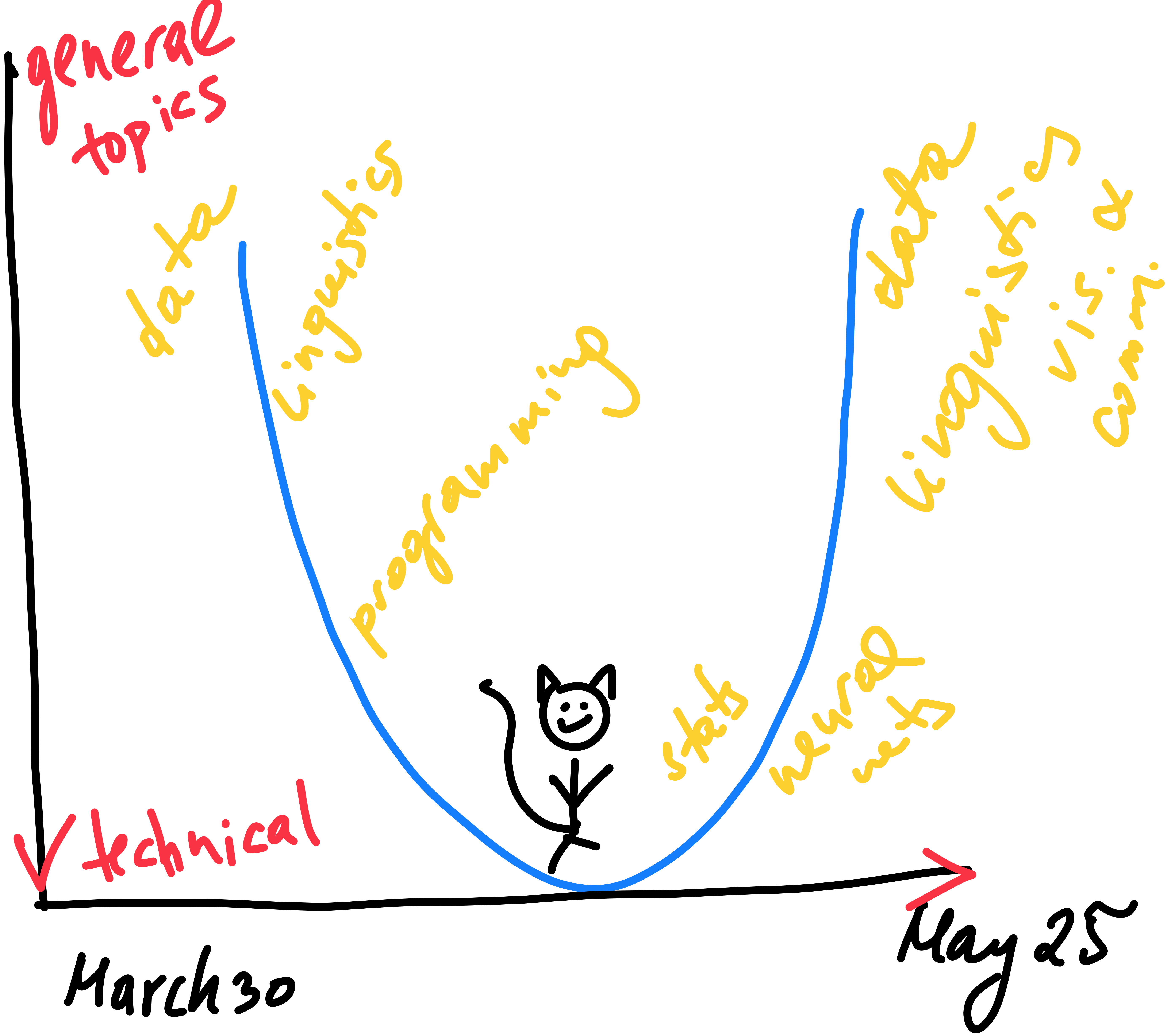
We have come a long way!

- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we've been mastering a variety of **technical** concepts and skills



We have come a long way!

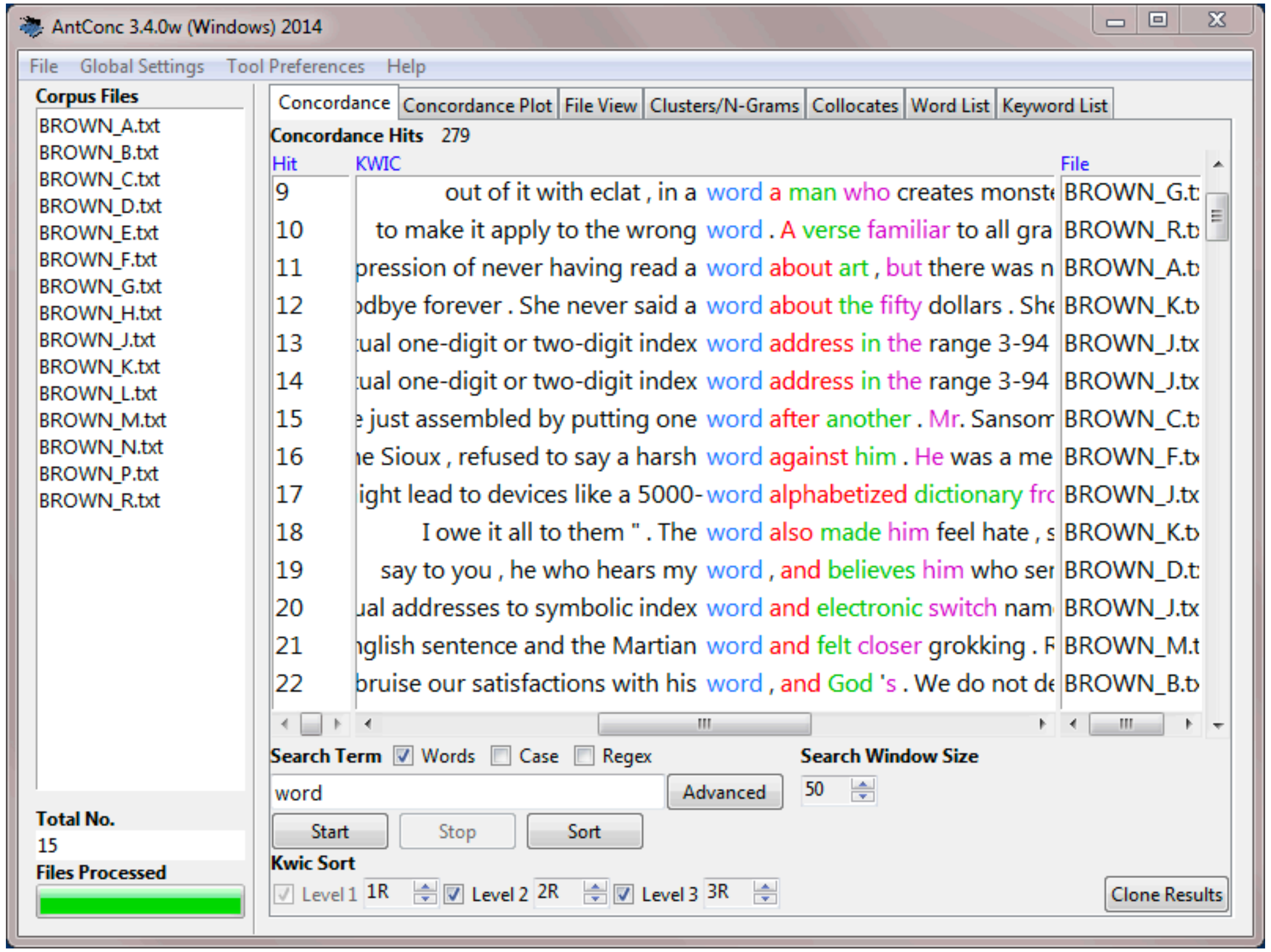
- We started and we finish with topics requiring **synthesizing** knowledge about the world
- …while for the most part, we've been mastering a variety of **technical** concepts and skills
- This was **not** easy!
- **Congratulations** on making it all the way back up!
- Synthesizing knowledge effectively is also **very** hard (much **harder**, in fact :))
    - …So try to do good **presentations** :)



Data Science and Linguistic Corpora

Linguistic Corpora

- Corpus:
  - A (large) collection of texts
  - ...annotated or unannotated
- Corpus search:
  - typically, using RegEx!



Corpus Files

BROWN_A.txt
BROWN_B.txt
BROWN_C.txt
BROWN_D.txt
BROWN_E.txt
BROWN_F.txt
BROWN_G.txt
BROWN_H.txt
BROWN_J.txt
BROWN_K.txt
BROWN_L.txt
BROWN_M.txt
BROWN_N.txt
BROWN_P.txt
BROWN_R.txt

Total No.
15

Files Processed

Concordance Hits 279

Hit	KWIC	File
9	out of it with eclat , in a word a man who creates monste	BROWN_G.t:
10	to make it apply to the wrong word . A verse familiar to all gra	BROWN_R.t›
11	pression of never having read a word about art , but there was n	BROWN_A.t›
12	›dbye forever . She never said a word about the fifty dollars . She	BROWN_K.t›
13	:ual one-digit or two-digit index word address in the range 3-94	BROWN_J.tx
14	:ual one-digit or two-digit index word address in the range 3-94	BROWN_J.tx
15	› just assembled by putting one word after another . Mr. Sansom	BROWN_C.t›
16	ıe Sioux , refused to say a harsh word against him . He was a me	BROWN_F.t›
17	ight lead to devices like a 5000-word alphabetized dictionary frc	BROWN_J.tx
18	I owe it all to them " . The word also made him feel hate , s	BROWN_K.t›
19	say to you , he who hears my word , and believes him who ser	BROWN_D.t:
20	ıal addresses to symbolic index word and electronic switch nam	BROWN_J.tx
21	ıglish sentence and the Martian word and felt closer grokking . F	BROWN_M.t
22	bruise our satisfactions with his word , and God 's . We do not de	BROWN_B.t›

Search Term ☑ Words □ Case □ Regex

word

Search Window Size 50

Kwic Sort

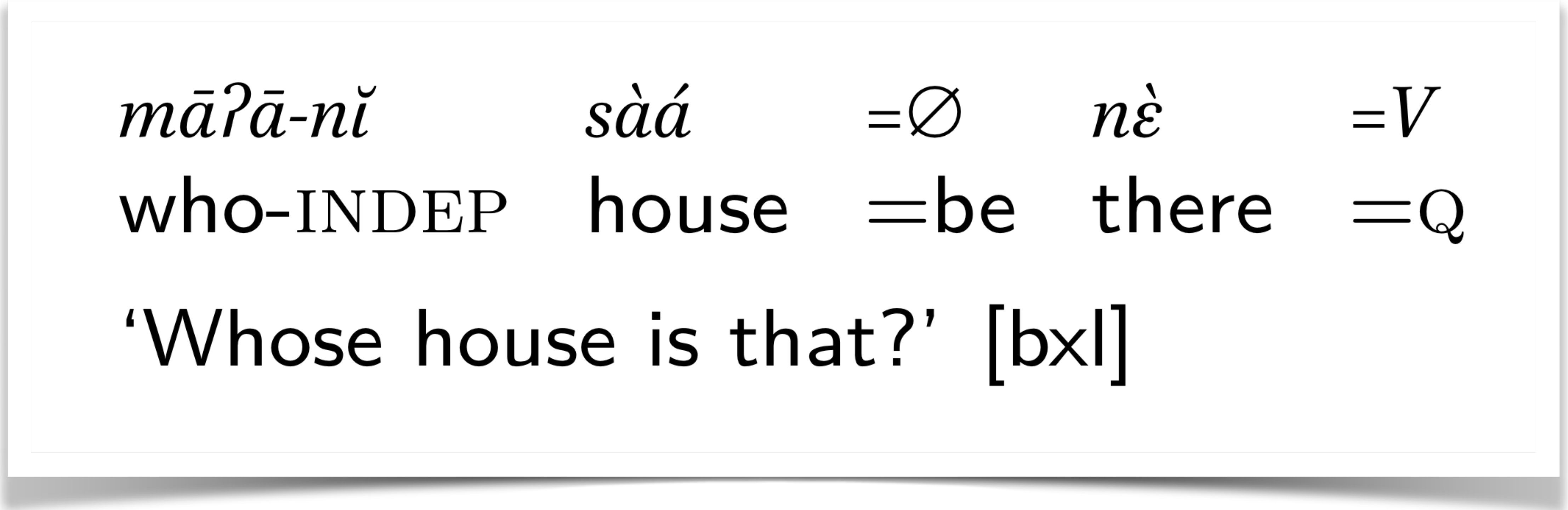
☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

https://allaboutcorpora.com/corpus-software-2

Annotated data

In linguistics

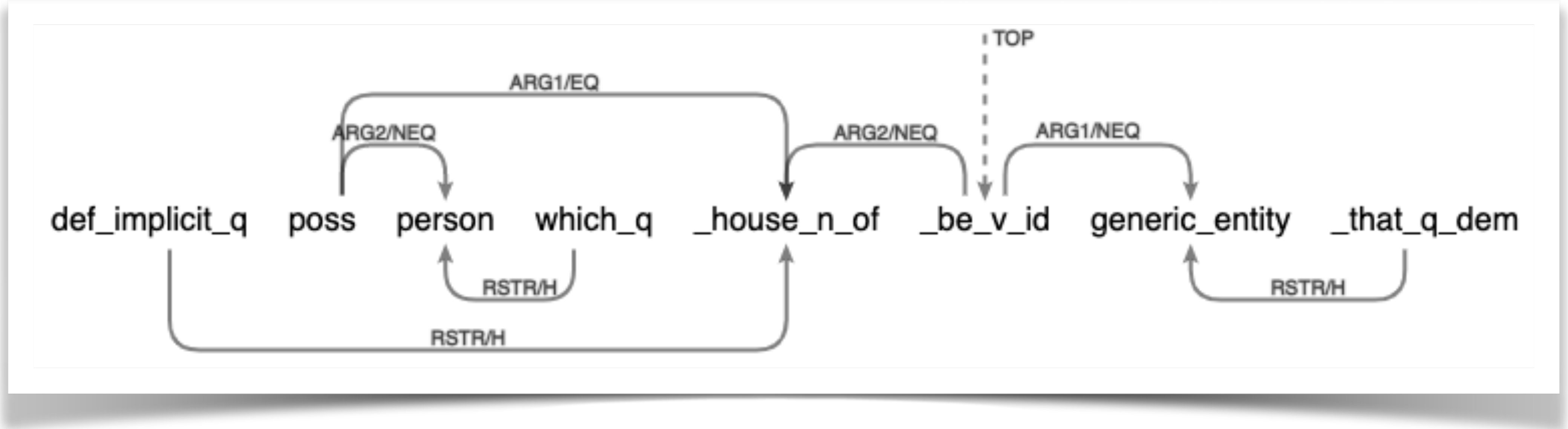
- Recorded speech and text associated with sociolinguistic variables:
    - Gender, age, geographic region…
- Interlinearized Glossed Text
    - Linguistic **analysis** and **annotation**
- **Structural** annotations
- What about syntax trees?
- In **computational** linguistics?
    - In **NLP**, emphasis on **raw** data
        - Why?
            - NLP is a computer science discipline
            - Deep learning



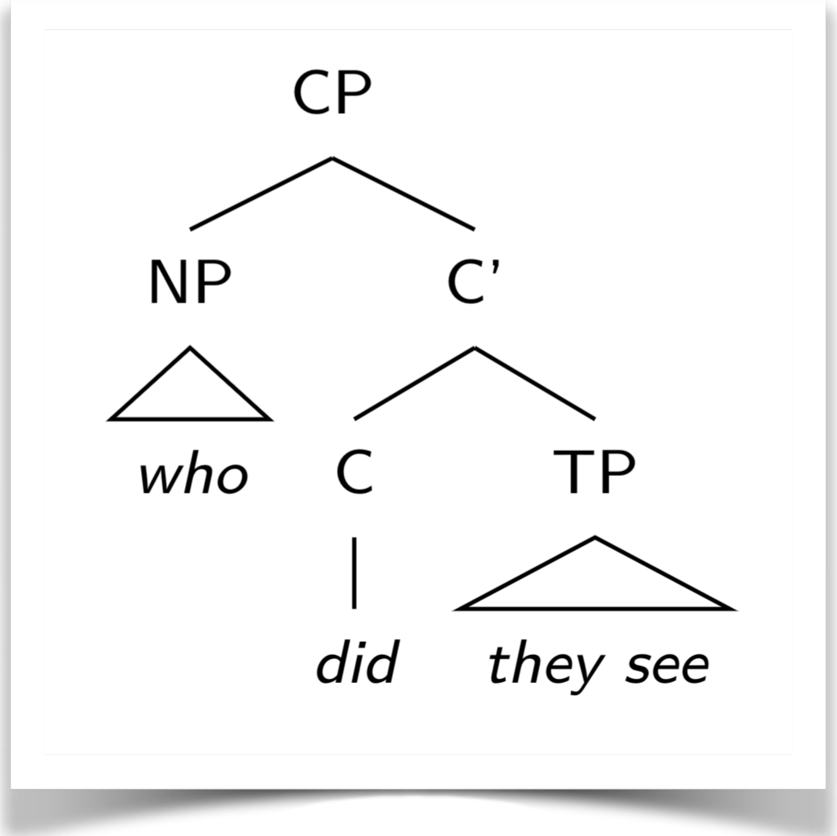
māʔā-nǐ	*sàá*	=∅	*nɛ̀*	=*V*
who-INDEP	house	=be	there	=Q

'Whose house is that?' [bxl]

Heath 2017. *A grammar of Jalkunan (Mande)*

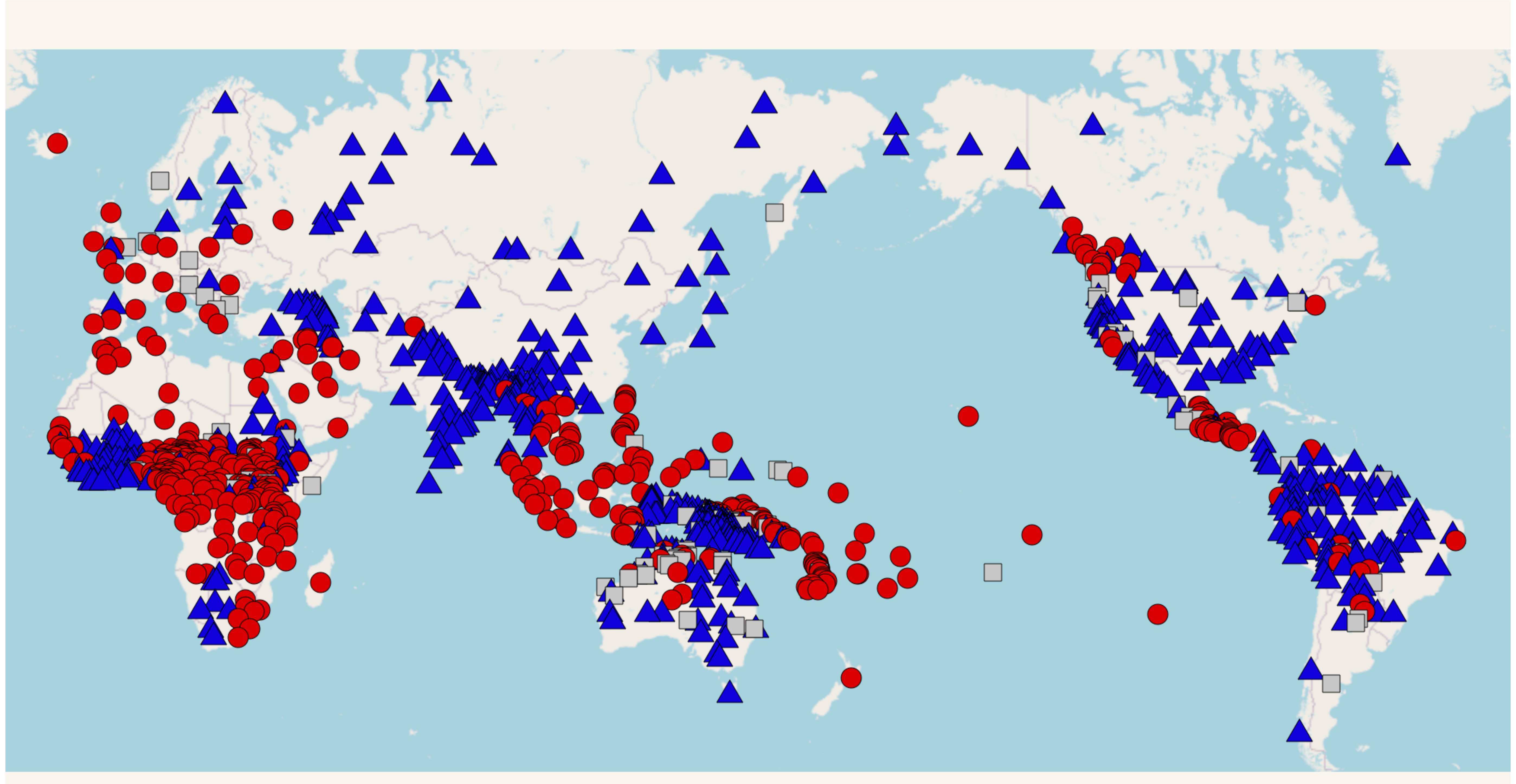


A *dependency graph* for the above sentence
One of the **most** useful data formats in NLP!



Linguistics
and data science

- "Corpus linguistics"
  - Various subfields; statistical analysis over large texts
- Sociolinguistics
  - Statistically significant correlations between sociolinguistic variables
- Historical ("dyachronic") linguistics
- Linguistic typology
- What else?
  - Almost **everything, potentially**
    - So long as the data can be **managed**



https://wals.info/feature/86A?v1=t00d&v3=sccc#2/21.0/152.9 (Dryer, 2005. WALS. Order of Genitive and Noun)

Data in linguistics

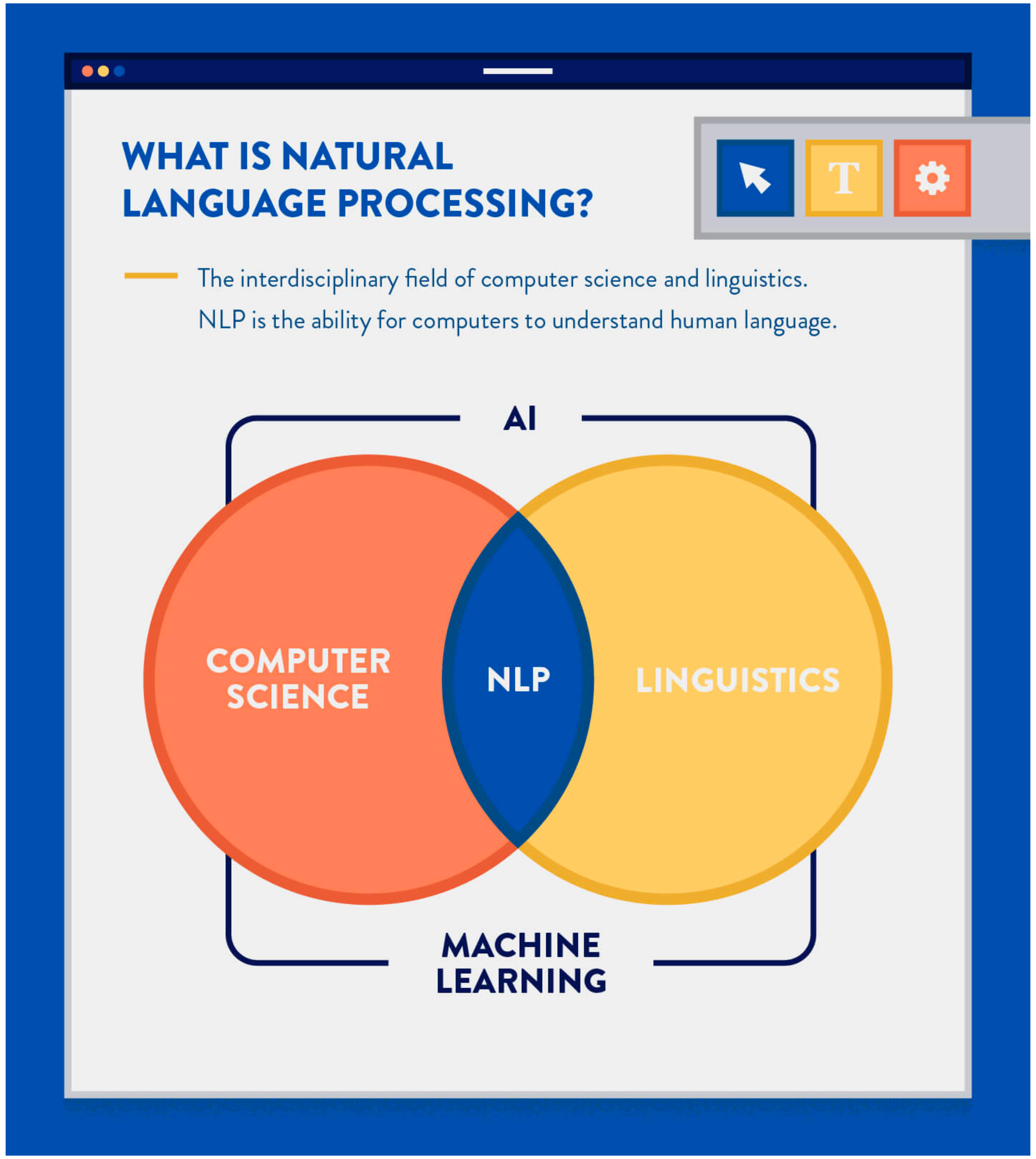
and data science

- Data science usually means LOTS of data
    - Why?
- Which areas of linguistics have LOTS of data?
    - **Most of them**
        - ...potentially

https://depts.washington.edu/ldplab/

Corpora
in NLP and Data Science

- **All** NLP technology is **trained** on corpora
- Much of NLP tech is **tested** on corpora, too…
  - …which need to be cleaned, stored, maintained, preprocessed…
  - …sometimes annotated
  - Our IMDB dataset is a corpus
    - Is it annotated?



WHAT IS NATURAL LANGUAGE PROCESSING?

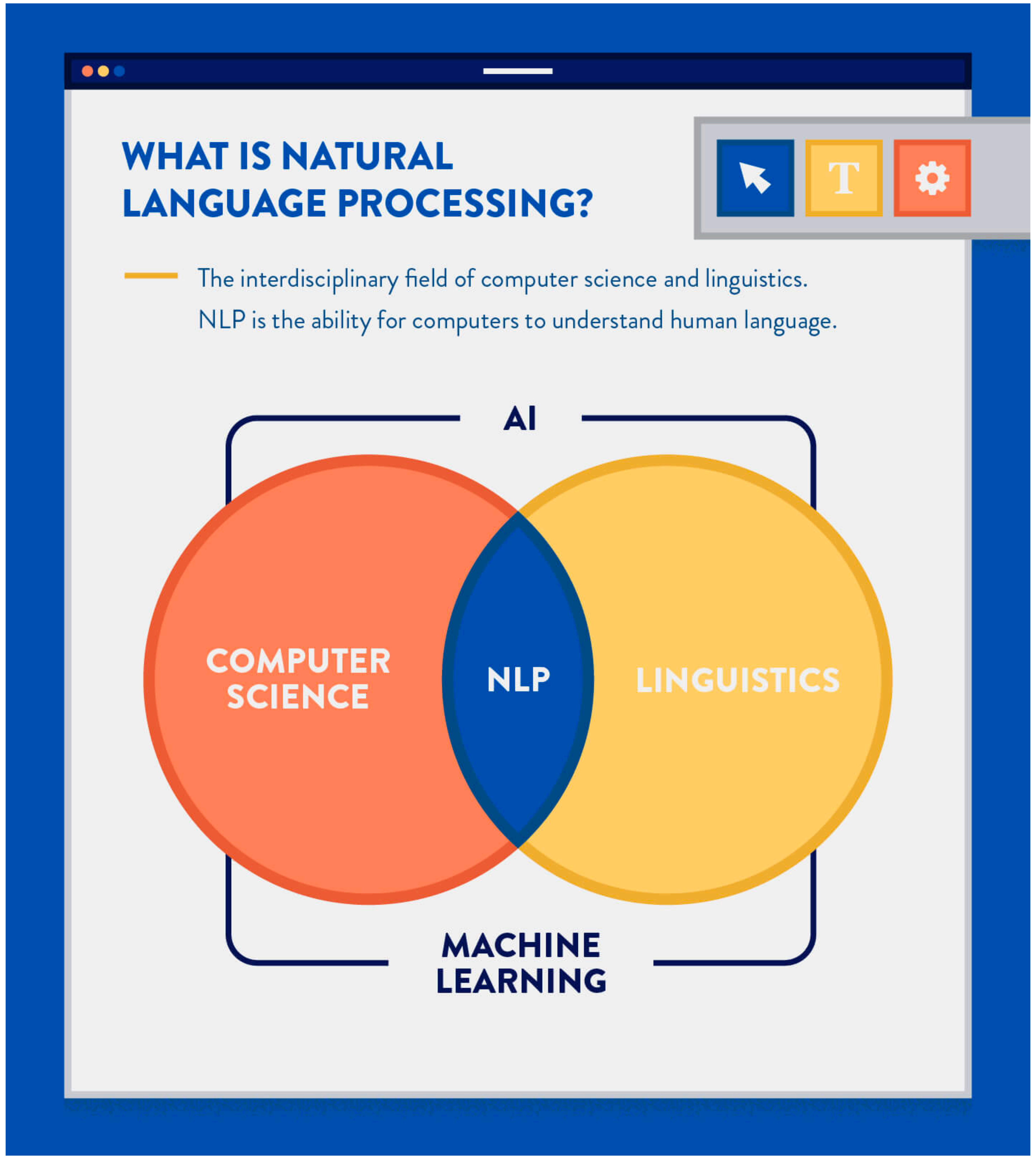
The interdisciplinary field of computer science and linguistics.
NLP is the ability for computers to understand human language.

https://clevertap.com/blog/natural-language-processing/
NB: Olga **doesn't** think computers can "understand" human language!
(But that's a philosophical debate.)

Linguistics

in NLP and Data Science

- *“The dog bit the man”*
- *“The man bit the dog”*
    - How to create different vectors for these?
    - People do add syntactic information to embeddings
    - Out of scope for us but one of the most important current developments



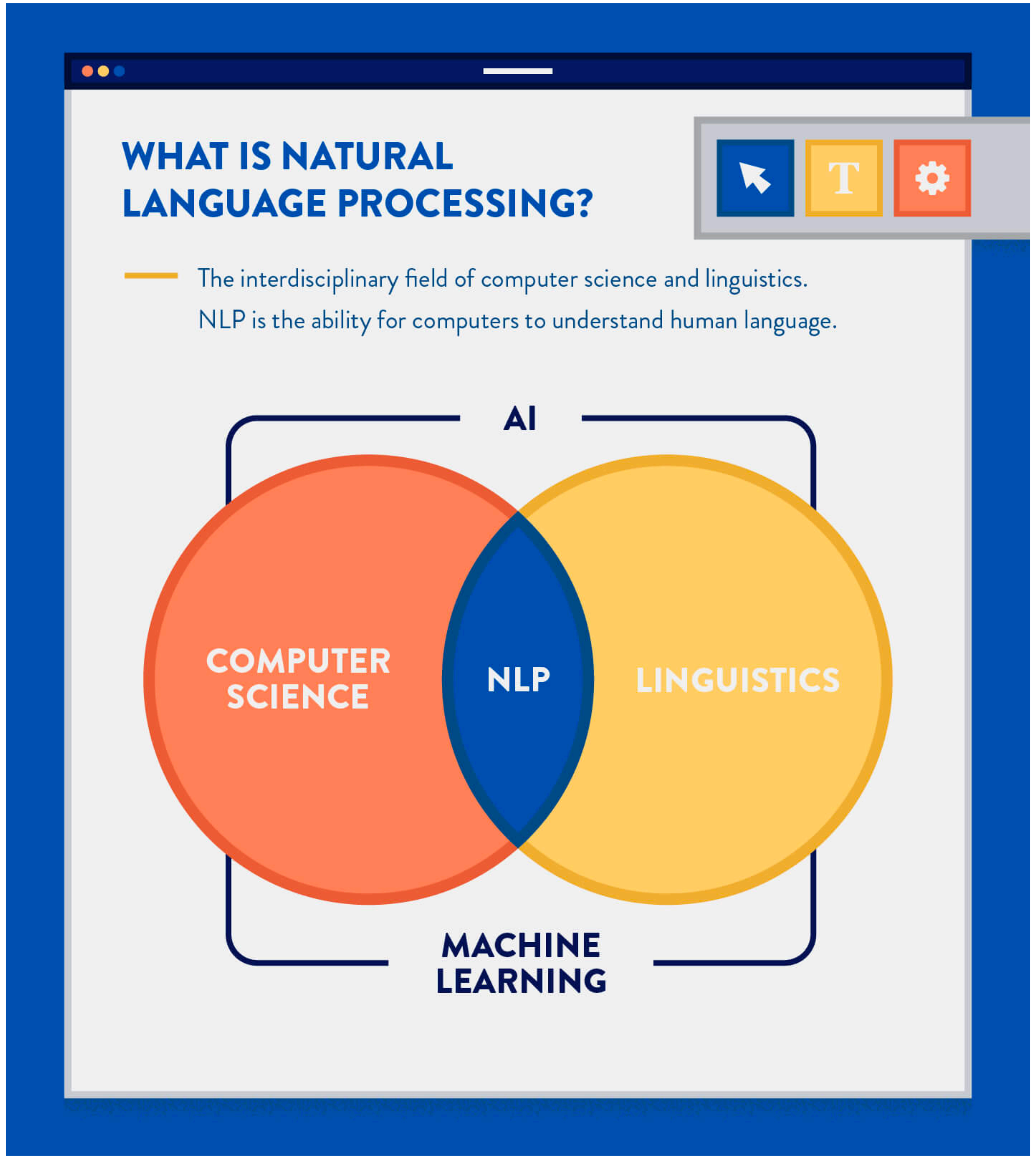
WHAT IS NATURAL LANGUAGE PROCESSING?

The interdisciplinary field of computer science and linguistics. NLP is the ability for computers to understand human language.

https://clevertap.com/blog/natural-language-processing/
NB: Olga **doesn’t** think computers can “understand” human language!
(But that’s a philosophical debate.)

Corpora and Data Science
in/for linguistics

- NLP:
  - Find **patterns** in language data
  - Perform language **tasks**
  - Performing a task well may (or may not) lead to **insights** about **faculties** needed to perform it
  - **Must** use corpora
- Linguistics:
  - Learn something **systematic** about human language
  - May or may not use corpora



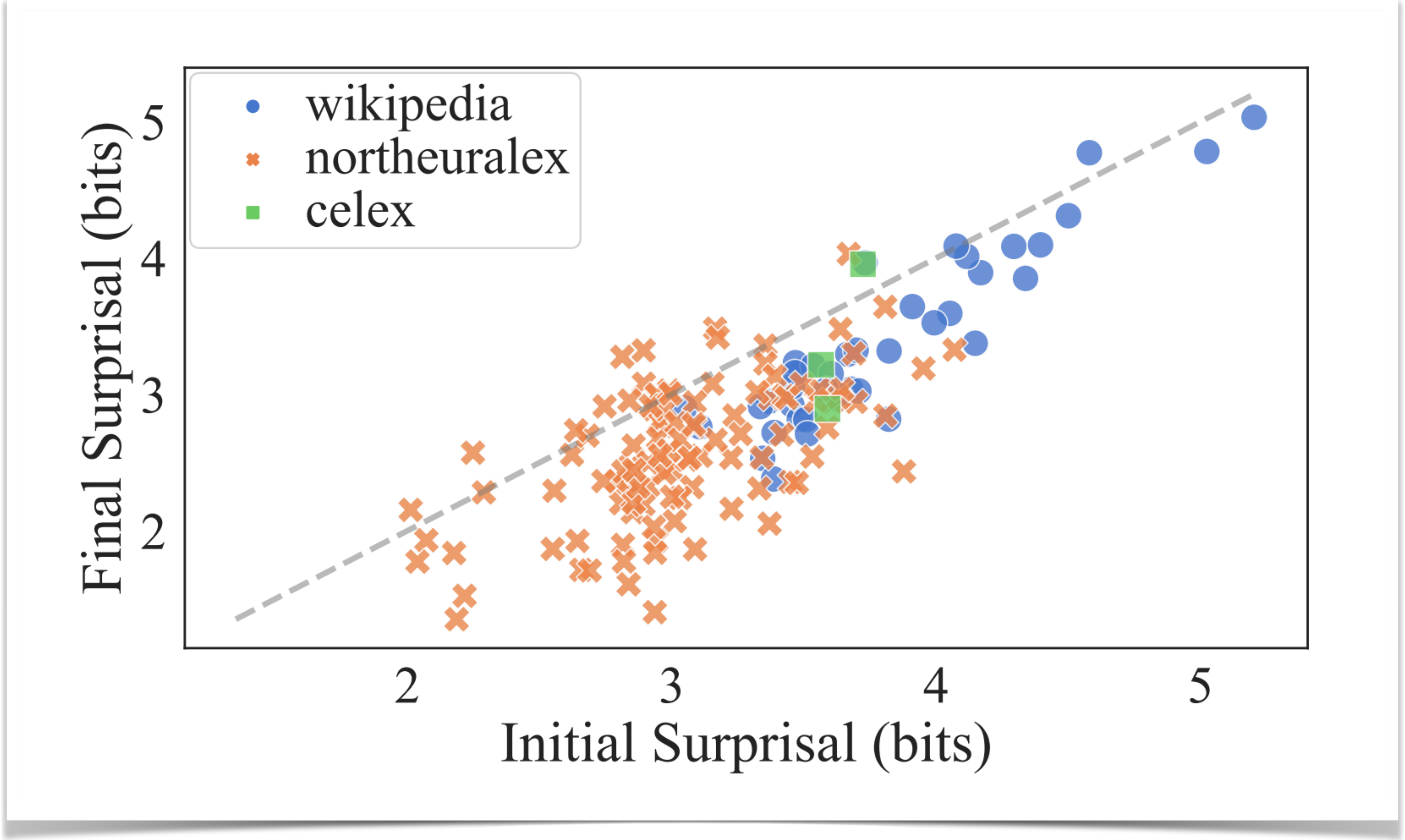
WHAT IS NATURAL LANGUAGE PROCESSING?

The interdisciplinary field of computer science and linguistics. NLP is the ability for computers to understand human language.

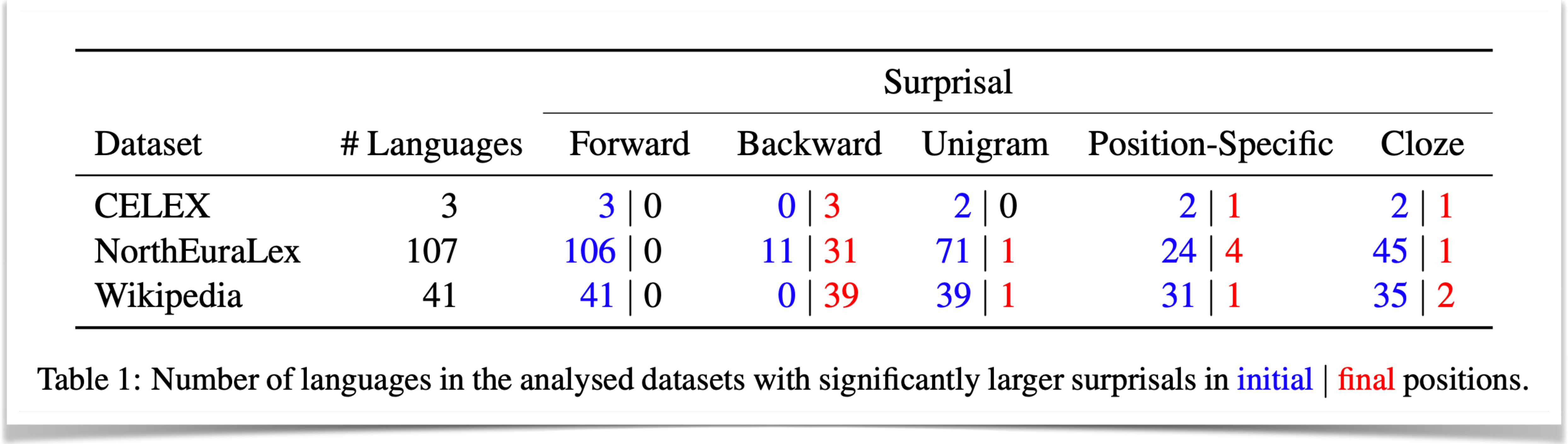
https://clevertap.com/blog/natural-language-processing/
NB: Olga **doesn't** think computers can "understand" human language!
(But that's a philosophical debate.)

Case study: Disambiguatory Signals
Pimentel et al. 2021

- Conjecture:
  - **More** information at the **beginning** of words than at the end
- **Theoretical** evidence:
  - **Information theory**
    - Info "gain"
    - Based on a **mathematical** notion of "**surprise**" (how easily **predictable**?)
    - Related concept: "**entropy**"
  - **This** paper: Probability distributions over phonological possibilities; DL networks



https://www.aclweb.org/anthology/2021.eacl-main.3.pdf

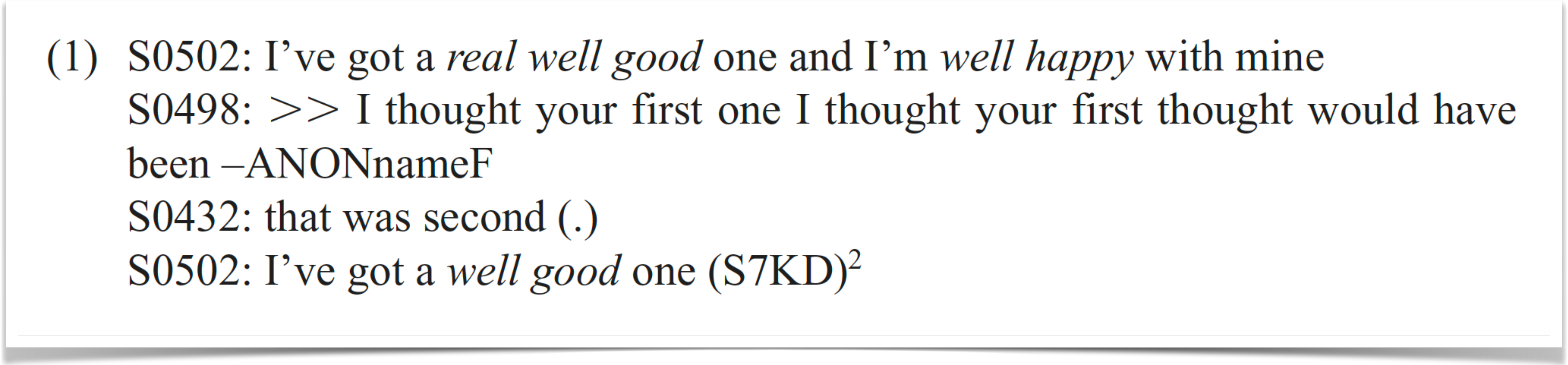


<table>
<tr><th></th><th></th><th colspan="5">Surprisal</th></tr>
<tr><th>Dataset</th><th># Languages</th><th>Forward</th><th>Backward</th><th>Unigram</th><th>Position-Specific</th><th>Cloze</th></tr>
<tr><td>CELEX</td><td>3</td><td>3 | 0</td><td>0 | 3</td><td>2 | 0</td><td>2 | 1</td><td>2 | 1</td></tr>
<tr><td>NorthEuraLex</td><td>107</td><td>106 | 0</td><td>11 | 31</td><td>71 | 1</td><td>24 | 4</td><td>45 | 1</td></tr>
<tr><td>Wikipedia</td><td>41</td><td>41 | 0</td><td>0 | 39</td><td>39 | 1</td><td>31 | 1</td><td>35 | 2</td></tr>
</table>

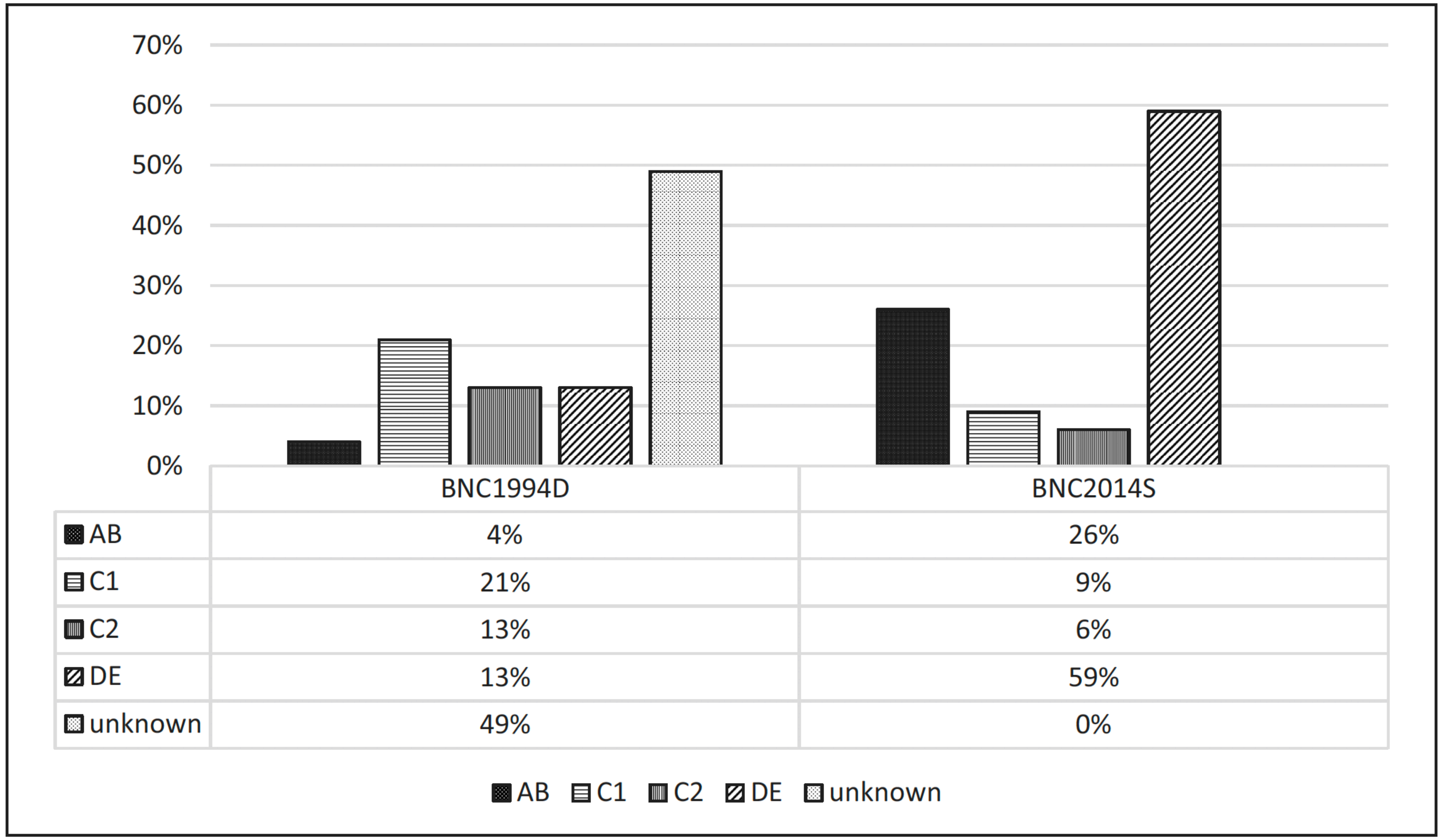
Table 1: Number of languages in the analysed datasets with significantly larger surprisals in initial | final positions.

Case study: language change “That’s well good” Aijmer, 2021

- Example of “small-scale” data science
  - ...but not in terms of what corpus had to be searched!



(1) S0502: I’ve got a *real well good* one and I’m *well happy* with mine
S0498: >> I thought your first one I thought your first thought would have been –ANONnameF
S0432: that was second (.)
S0502: I’ve got a *well good* one (S7KD)[2]

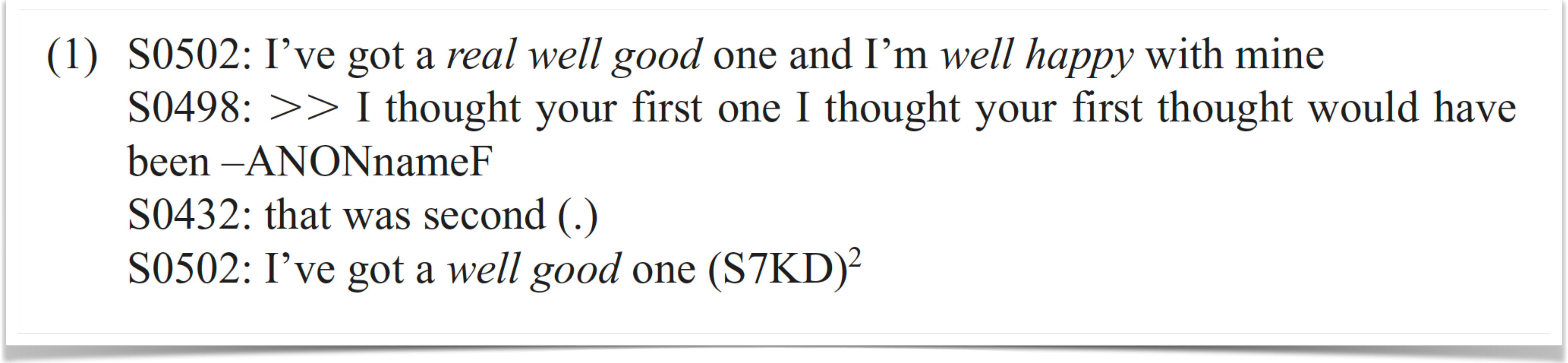


	BNC1994D	BNC2014S
▩ AB	4%	26%
▤ C1	21%	9%
▥ C2	13%	6%
▨ DE	13%	59%
▦ unknown	49%	0%

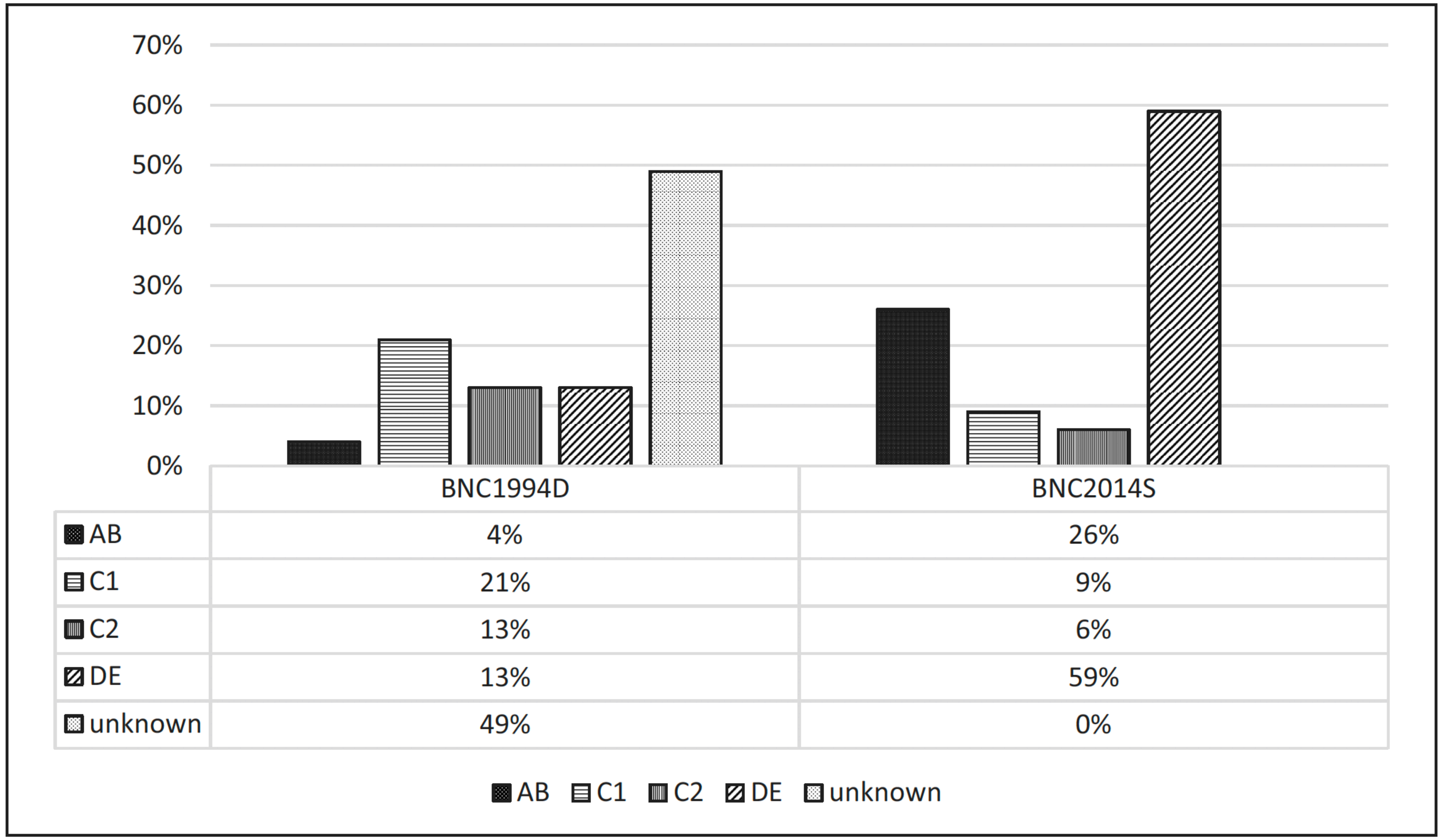
Figure 2. Relative Frequencies of *Well* across Speaker Groups Classified with Regard to Social Grade in BNC1994D and BNC2014S (Percentages)

Case study: language change “That’s well good” Aijmer, 2021

- Study of frequency od “well” as an intensifier
  - standard: “very”, “really”
  - Etymological aside:
    - “well” derives from “will”
    - something one was willing was supposed to be “true” :)



(1) S0502: I’ve got a *real well good* one and I’m *well happy* with mine
S0498: >> I thought your first one I thought your first thought would have been –ANONnameF
S0432: that was second (.)
S0502: I’ve got a *well good* one (S7KD)[2]

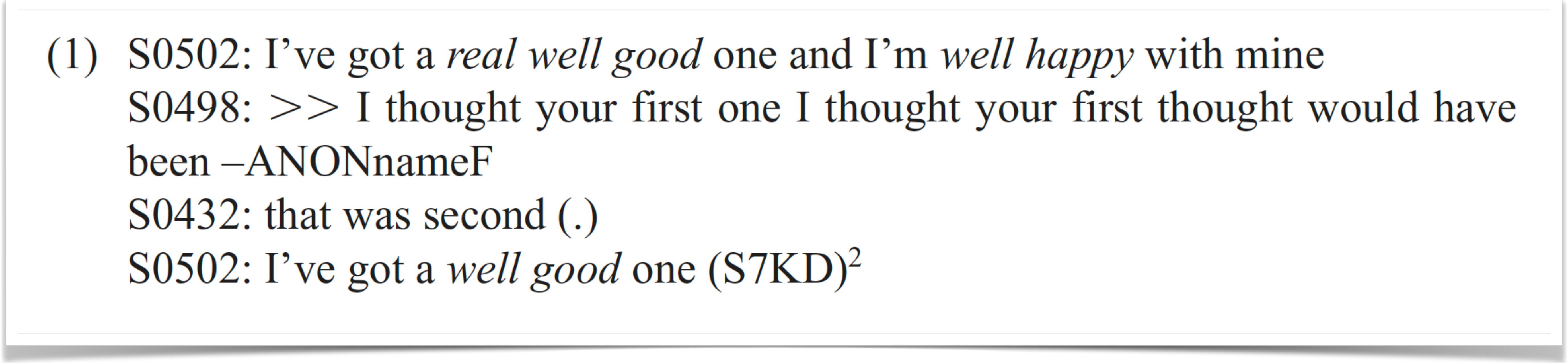


	BNC1994D	BNC2014S
■AB	4%	26%
▤C1	21%	9%
▥C2	13%	6%
▨DE	13%	59%
▩unknown	49%	0%

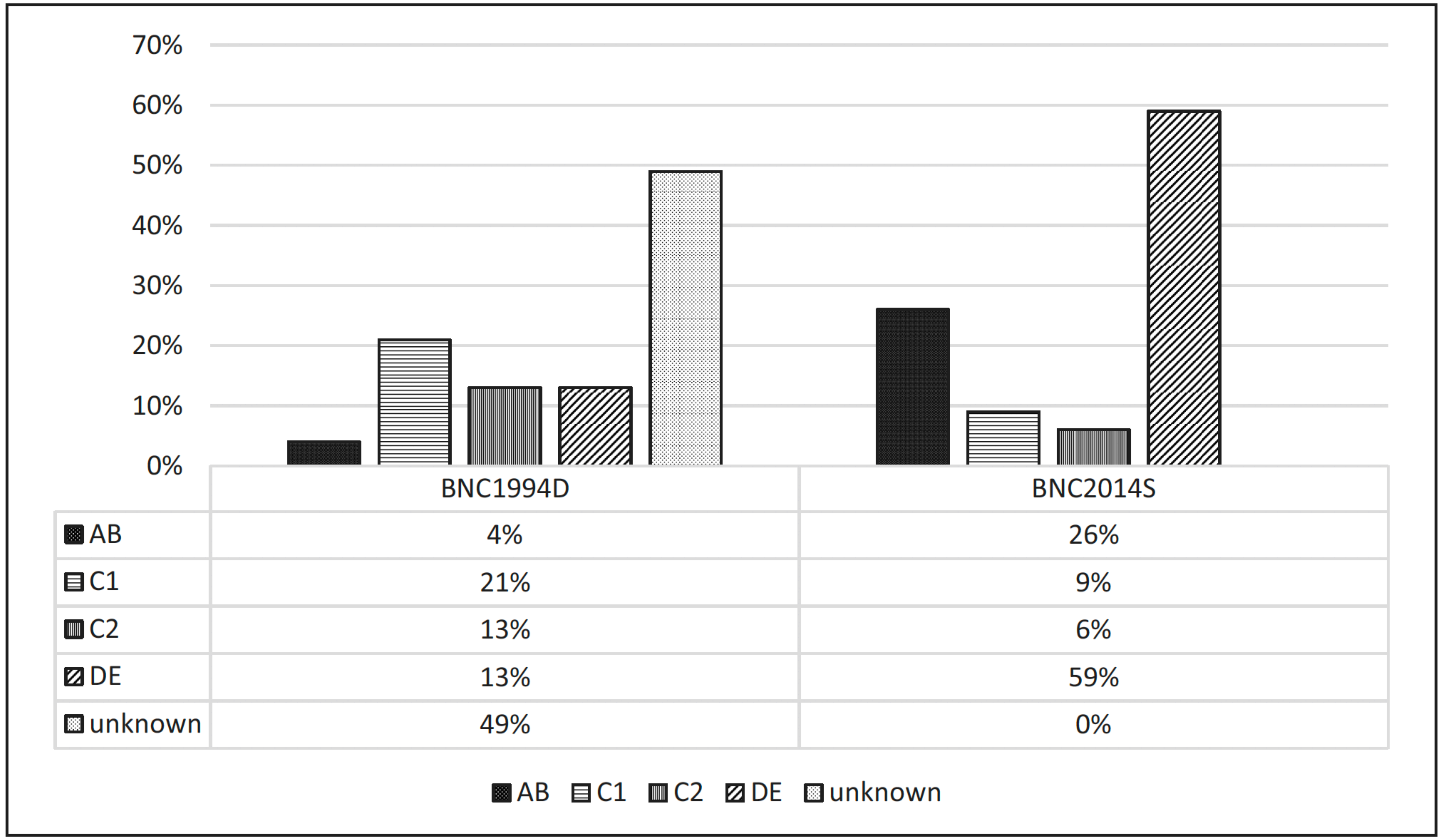
Figure 2. Relative Frequencies of *Well* across Speaker Groups Classified with Regard to Social Grade in BNC1994D and BNC2014S (Percentages)

Case study: language change “That’s well good” Aijmer, 2021

- “Well” was common as an intensifier before
- And may be becoming more common now
- Several questions:
  - Do women use it more than men
  - Do young people use it more than older people
  - Regional variation
  - Socio-economic class



(1) S0502: I’ve got a *real well good* one and I’m *well happy* with mine
S0498: >> I thought your first one I thought your first thought would have been –ANONnameF
S0432: that was second (.)
S0502: I’ve got a *well good* one (S7KD)[2]

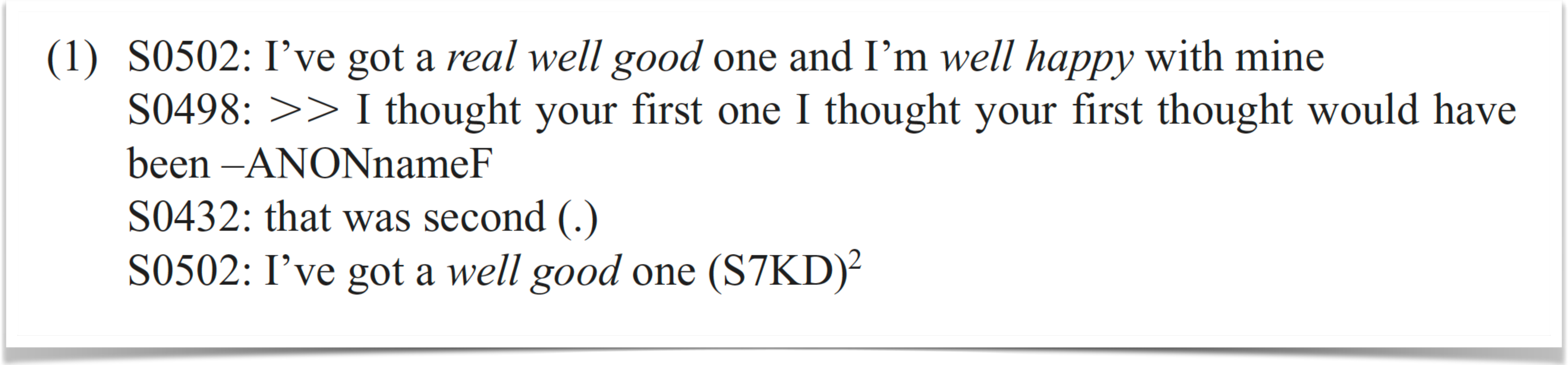


	BNC1994D	BNC2014S
■ AB	4%	26%
▤ C1	21%	9%
▥ C2	13%	6%
▨ DE	13%	59%
▩ unknown	49%	0%

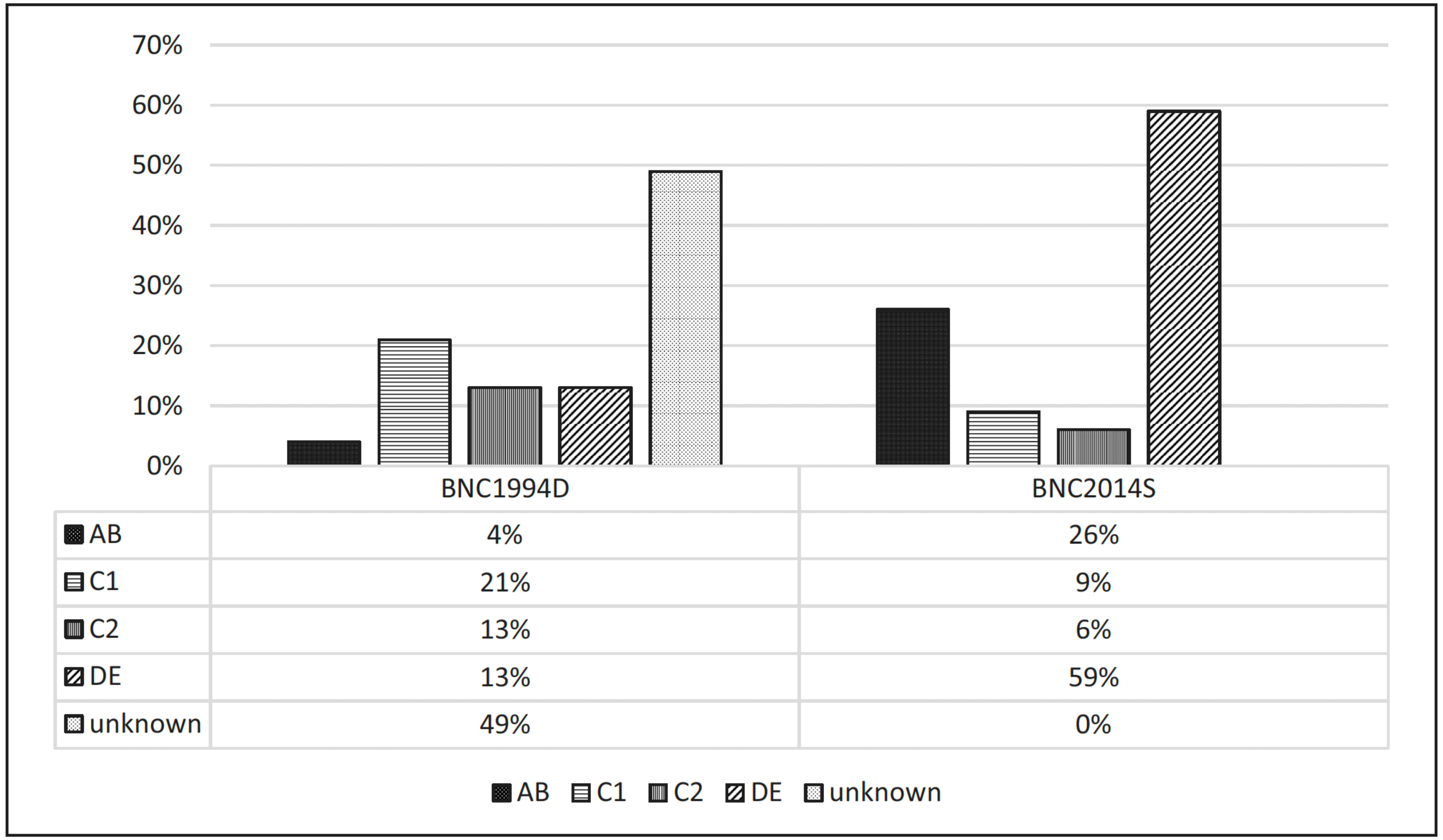
Figure 2. Relative Frequencies of *Well* across Speaker Groups Classified with Regard to Social Grade in BNC1994D and BNC2014S (Percentages)

Case study: language change "That's well good" Aijmer, 2021

- **Not** an example of NLP-style data science
- But **is** data science in the sence that a **huge** corpus had to be processed
- NB: NLP techniques can be used for **better** search in corpora



(1) S0502: I've got a *real well good* one and I'm *well happy* with mine
S0498: >> I thought your first one I thought your first thought would have been –ANONnameF
S0432: that was second (.)
S0502: I've got a *well good* one (S7KD)[2]

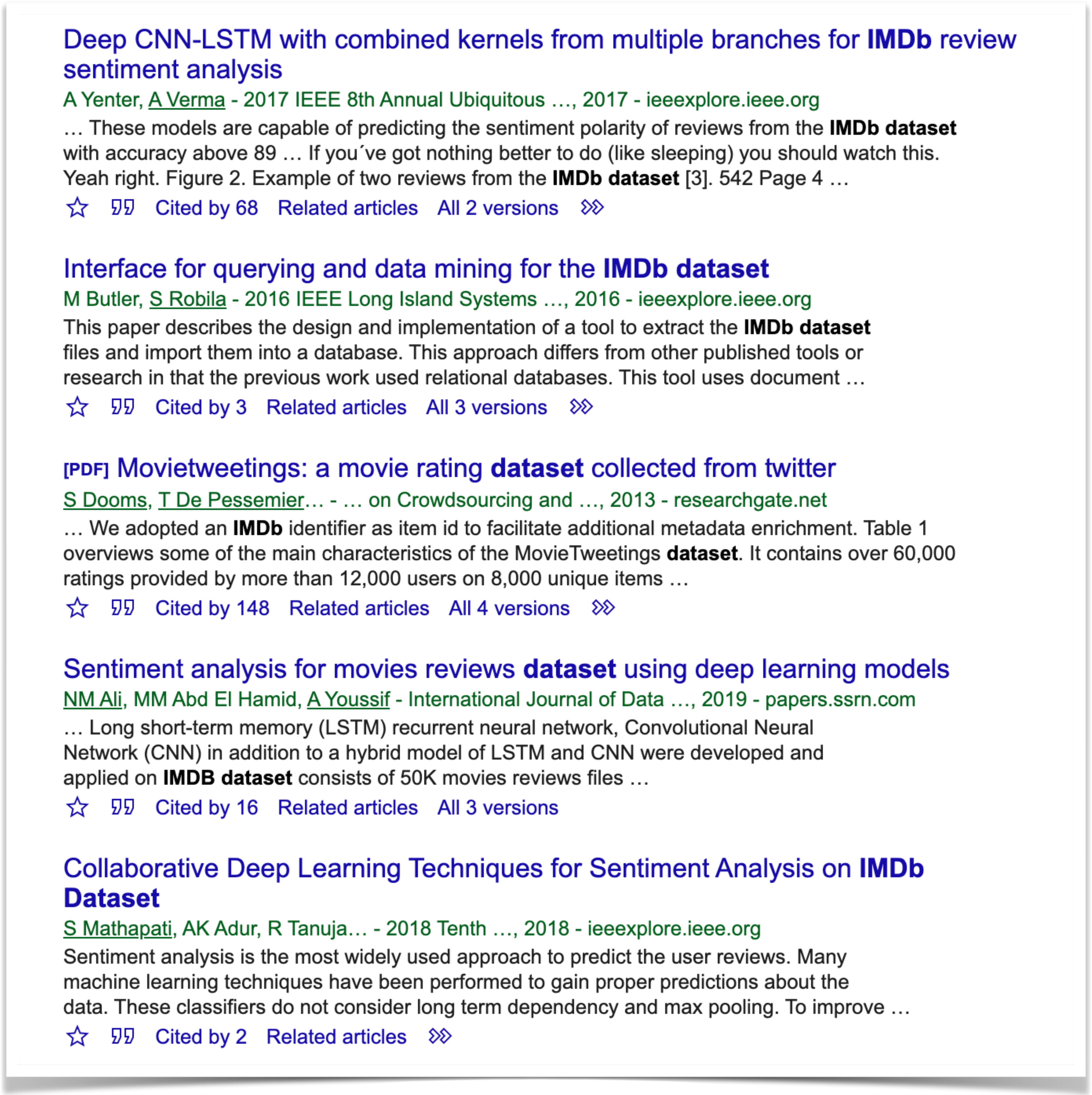


	BNC1994D	BNC2014S
▩ AB	4%	26%
▤ C1	21%	9%
▥ C2	13%	6%
▨ DE	13%	59%
▧ unknown	49%	0%

Figure 2. Relative Frequencies of *Well* across Speaker Groups Classified with Regard to Social Grade in BNC1994D and BNC2014S (Percentages)

IMDB data science

- Task: Sentiment analysis
  - positive or negative?
- Procedure: Train and Test on IMDB
- Goal: Generalize from there
  - still not achieved :)



Deep CNN-LSTM with combined kernels from multiple branches for **IMDb** review sentiment analysis
A Yenter, A Verma - 2017 IEEE 8th Annual Ubiquitous …, 2017 - ieeexplore.ieee.org
… These models are capable of predicting the sentiment polarity of reviews from the **IMDb dataset** with accuracy above 89 … If you´ve got nothing better to do (like sleeping) you should watch this. Yeah right. Figure 2. Example of two reviews from the **IMDb dataset** [3]. 542 Page 4 …
☆ ❞ Cited by 68 Related articles All 2 versions »

Interface for querying and data mining for the **IMDb dataset**
M Butler, S Robila - 2016 IEEE Long Island Systems …, 2016 - ieeexplore.ieee.org
This paper describes the design and implementation of a tool to extract the **IMDb dataset** files and import them into a database. This approach differs from other published tools or research in that the previous work used relational databases. This tool uses document …
☆ ❞ Cited by 3 Related articles All 3 versions »

[PDF] Movietweetings: a movie rating **dataset** collected from twitter
S Dooms, T De Pessemier… - … on Crowdsourcing and …, 2013 - researchgate.net
… We adopted an **IMDb** identifier as item id to facilitate additional metadata enrichment. Table 1 overviews some of the main characteristics of the MovieTweetings **dataset**. It contains over 60,000 ratings provided by more than 12,000 users on 8,000 unique items …
☆ ❞ Cited by 148 Related articles All 4 versions »

Sentiment analysis for movies reviews **dataset** using deep learning models
NM Ali, MM Abd El Hamid, A Youssif - International Journal of Data …, 2019 - papers.ssrn.com
… Long short-term memory (LSTM) recurrent neural network, Convolutional Neural Network (CNN) in addition to a hybrid model of LSTM and CNN were developed and applied on **IMDB dataset** consists of 50K movies reviews files …
☆ ❞ Cited by 16 Related articles All 3 versions

Collaborative Deep Learning Techniques for Sentiment Analysis on **IMDb Dataset**
S Mathapati, AK Adur, R Tanuja… - 2018 Tenth …, 2018 - ieeexplore.ieee.org
Sentiment analysis is the most widely used approach to predict the user reviews. Many machine learning techniques have been performed to gain proper predictions about the data. These classifiers do not consider long term dependency and max pooling. To improve …
☆ ❞ Cited by 2 Related articles »

Google Scholar search results for "IMDB dataset"

IMDB data science

types of data science

- Throw a new architecture at it
  - use as **benchmark**
- Create a new interface for querying
- Create new (similar) datasets
- The above, **especially** the **first** one, are the typical types of NLP research with language data
- Data Science:
  - Similar, **but**
  - Trying to make sense of the numbers more

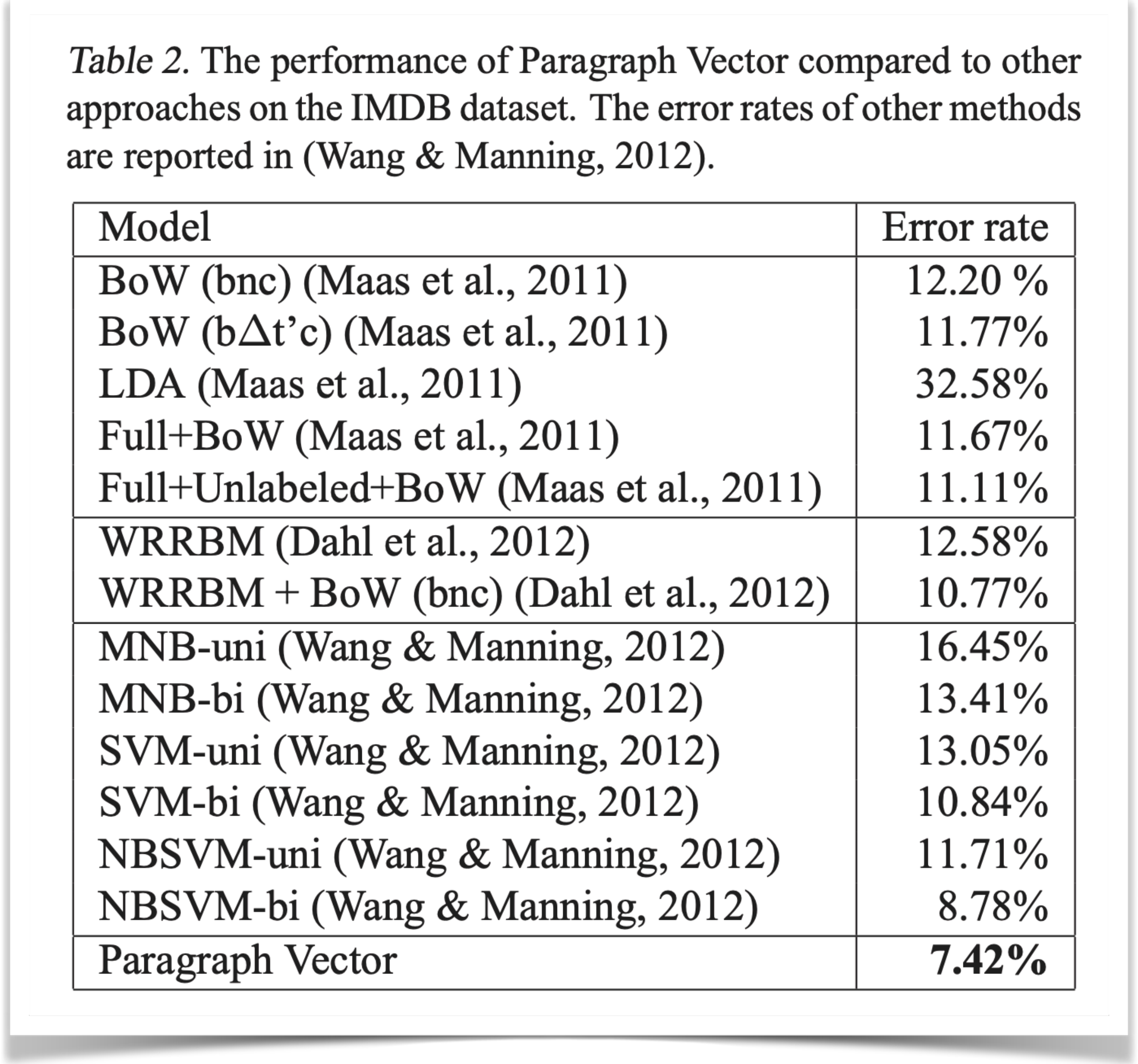
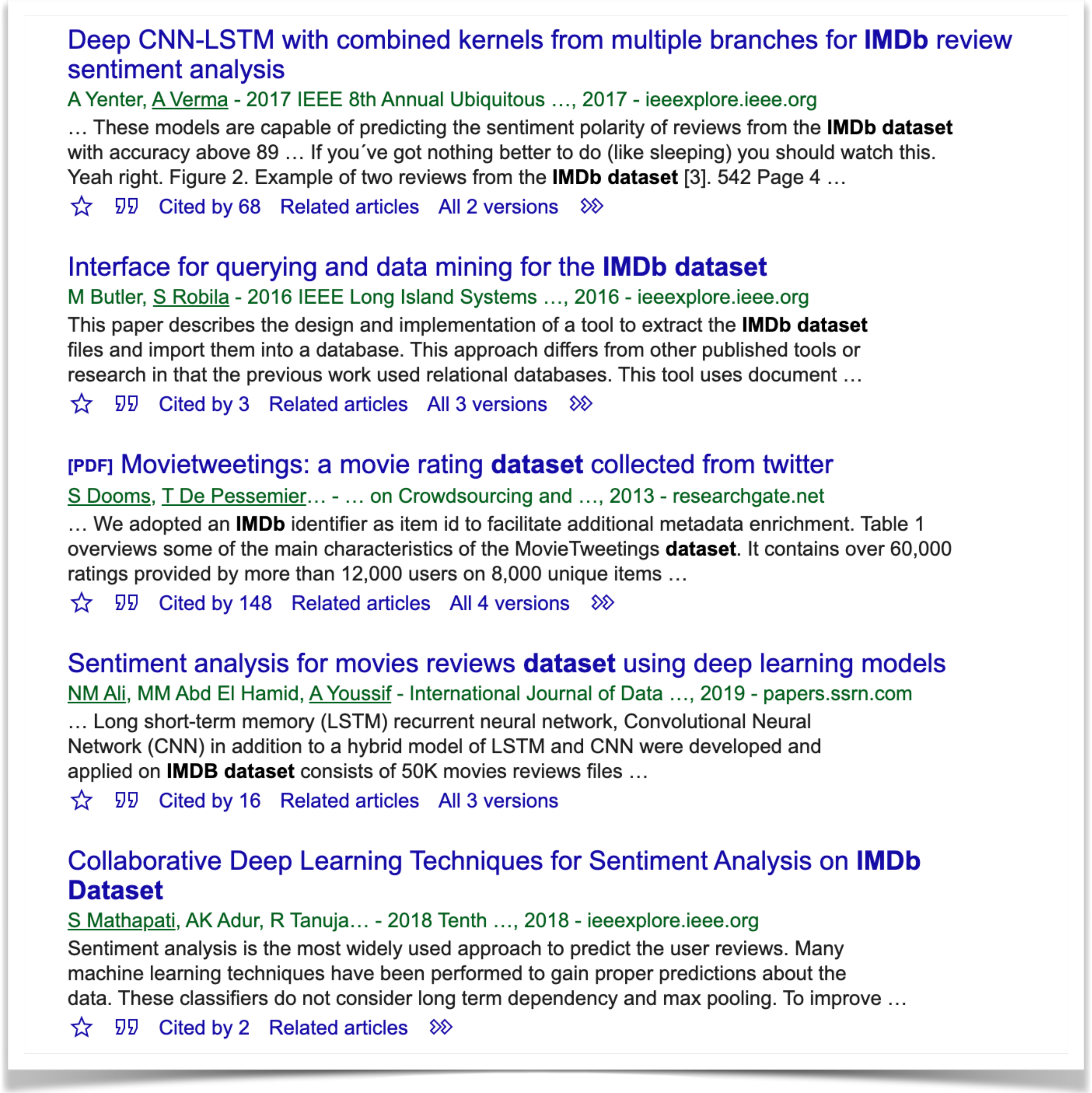


Table 2. The performance of Paragraph Vector compared to other approaches on the IMDB dataset. The error rates of other methods are reported in (Wang & Manning, 2012).

Model	Error rate
BoW (bnc) (Maas et al., 2011)	12.20 %
BoW (bΔt'c) (Maas et al., 2011)	11.77%
LDA (Maas et al., 2011)	32.58%
Full+BoW (Maas et al., 2011)	11.67%
Full+Unlabeled+BoW (Maas et al., 2011)	11.11%
WRRBM (Dahl et al., 2012)	12.58%
WRRBM + BoW (bnc) (Dahl et al., 2012)	10.77%
MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector	**7.42%**

Le and Mikolov 2014



Deep CNN-LSTM with combined kernels from multiple branches for **IMDb** review sentiment analysis
A Yenter, A Verma - 2017 IEEE 8th Annual Ubiquitous …, 2017 - ieeexplore.ieee.org
… These models are capable of predicting the sentiment polarity of reviews from the **IMDb dataset** with accuracy above 89 … If you´ve got nothing better to do (like sleeping) you should watch this. Yeah right. Figure 2. Example of two reviews from the **IMDb dataset** [3]. 542 Page 4 …
☆ ❞ Cited by 68 Related articles All 2 versions »

Interface for querying and data mining for the **IMDb dataset**
M Butler, S Robila - 2016 IEEE Long Island Systems …, 2016 - ieeexplore.ieee.org
This paper describes the design and implementation of a tool to extract the **IMDb dataset** files and import them into a database. This approach differs from other published tools or research in that the previous work used relational databases. This tool uses document …
☆ ❞ Cited by 3 Related articles All 3 versions »

[PDF] Movietweetings: a movie rating **dataset** collected from twitter
S Dooms, T De Pessemier… - … on Crowdsourcing and …, 2013 - researchgate.net
… We adopted an **IMDb** identifier as item id to facilitate additional metadata enrichment. Table 1 overviews some of the main characteristics of the MovieTweetings **dataset**. It contains over 60,000 ratings provided by more than 12,000 users on 8,000 unique items …
☆ ❞ Cited by 148 Related articles All 4 versions »

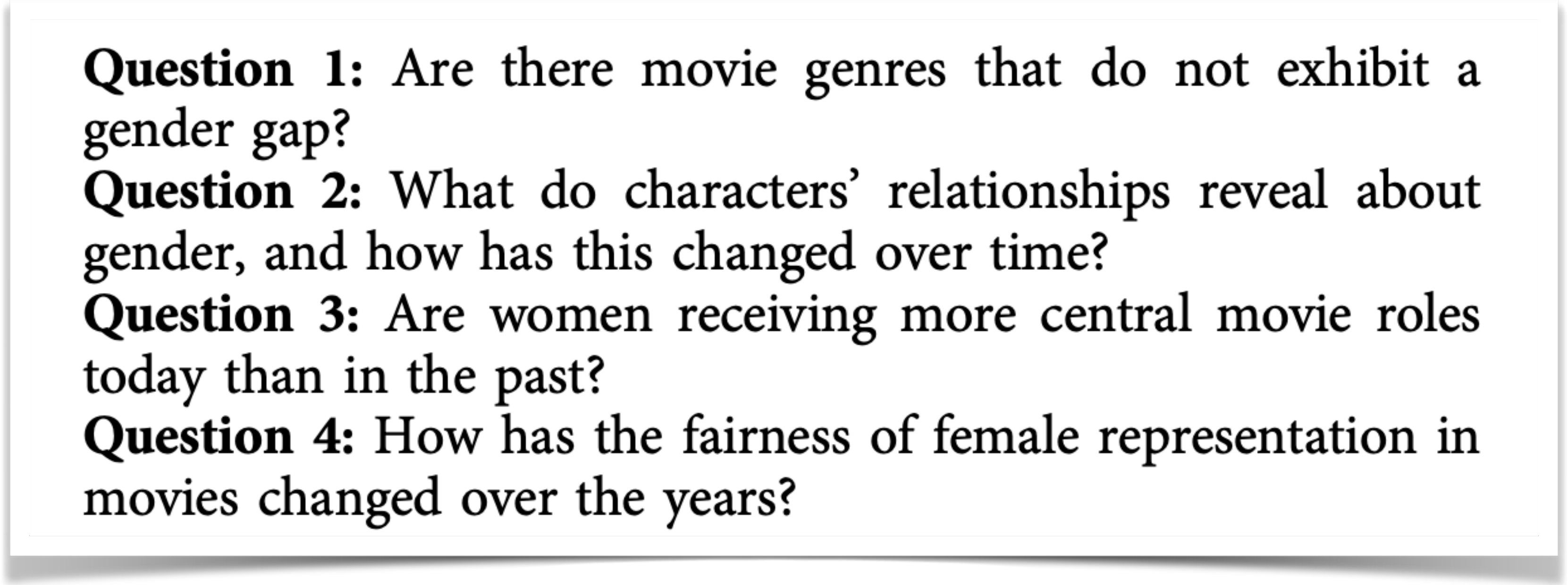
Sentiment analysis for movies reviews **dataset** using deep learning models
NM Ali, MM Abd El Hamid, A Youssif - International Journal of Data …, 2019 - papers.ssrn.com
… Long short-term memory (LSTM) recurrent neural network, Convolutional Neural Network (CNN) in addition to a hybrid model of LSTM and CNN were developed and applied on **IMDB dataset** consists of 50K movies reviews files …
☆ ❞ Cited by 16 Related articles All 3 versions

Collaborative Deep Learning Techniques for Sentiment Analysis on **IMDb Dataset**
S Mathapati, AK Adur, R Tanuja… - 2018 Tenth …, 2018 - ieeexplore.ieee.org
Sentiment analysis is the most widely used approach to predict the user reviews. Many machine learning techniques have been performed to gain proper predictions about the data. These classifiers do not consider long term dependency and max pooling. To improve …
☆ ❞ Cited by 2 Related articles »

Google Scholar search results for "IMDB dataset"

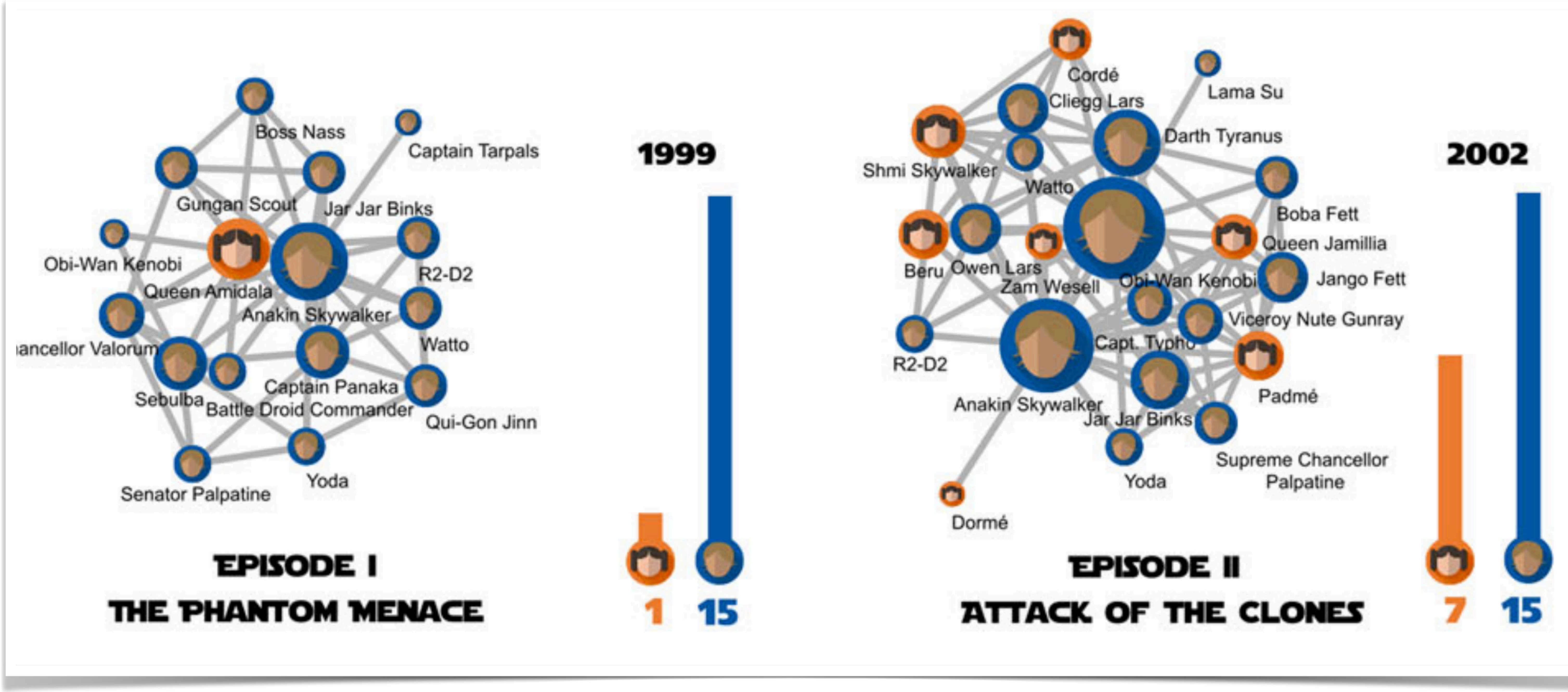
Using data science to understand the film industry's gender gap

Kagan et al. 2020



Question 1: Are there movie genres that do not exhibit a gender gap?
Question 2: What do characters' relationships reveal about gender, and how has this changed over time?
Question 3: Are women receiving more central movie roles today than in the past?
Question 4: How has the fairness of female representation in movies changed over the years?

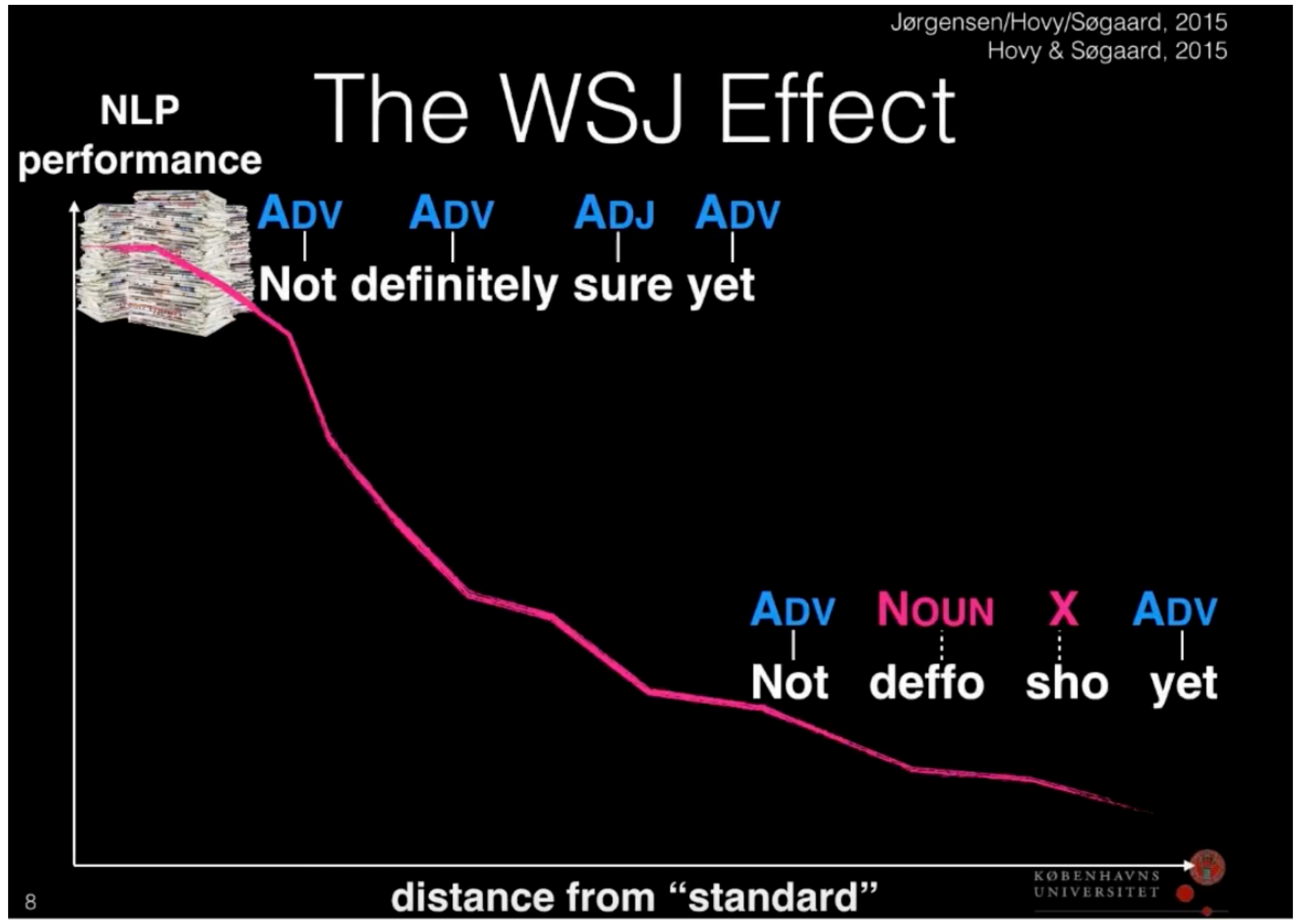
- Exploration of gender-related questions in film
- IMDB used to collect various info e.g. character lists
- **Inquiry into social questions using NLP techniques**
- Lots of cool visualizations :)



https://www.nature.com/articles/s41599-020-0436-1.pdf?origin=ppub

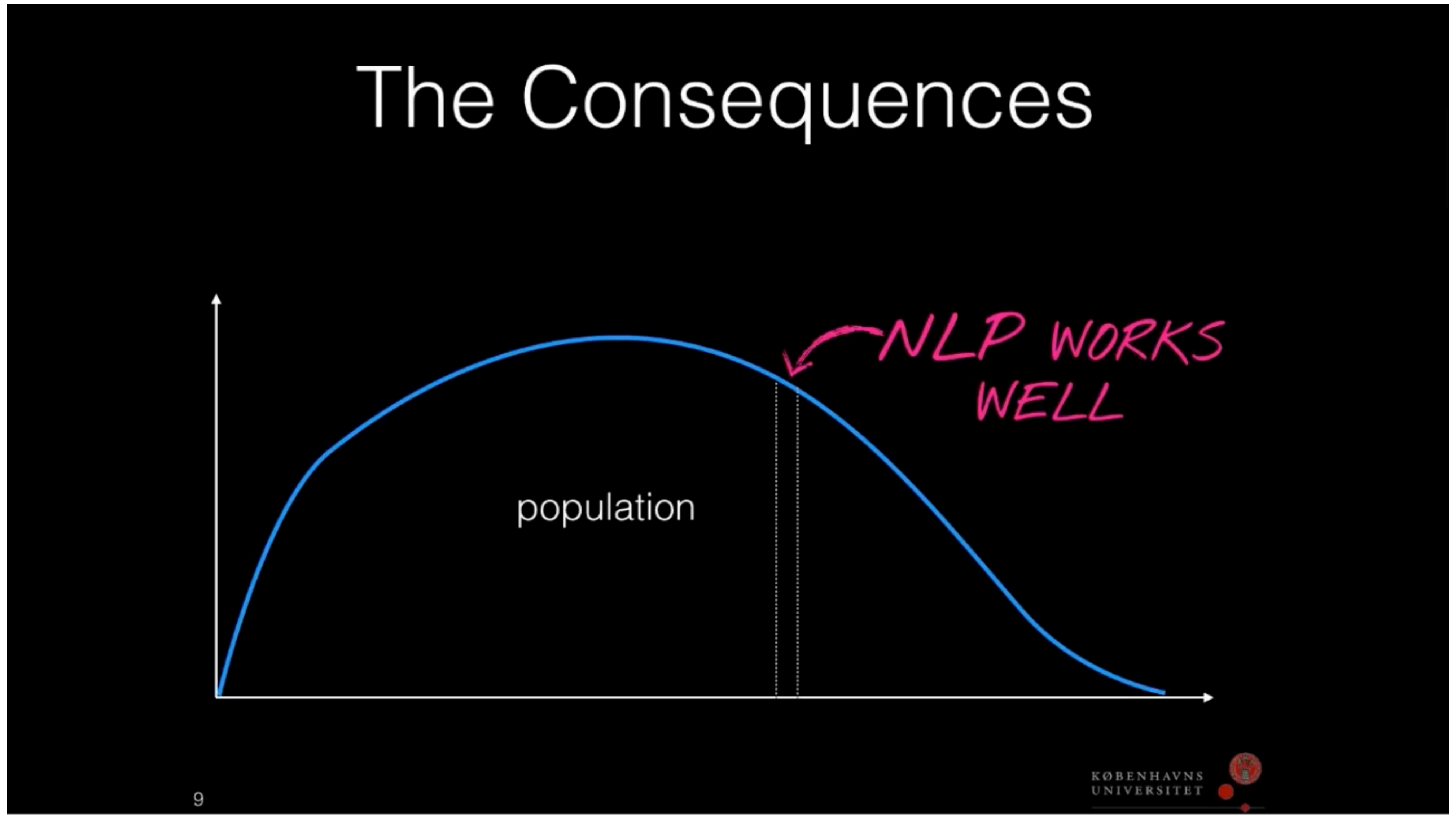
Social impact
of NLP

- is a highly politicized thing :)
- But, stuff does have social impact and it’s good to try and think about it every now and then
- “Bias”: systems trained on data which does not represent all people equally well may discriminate against those who are represented less
- ...and, them being systems, they won’t be held accountable
- Recommended: *Weapons of math destruction* by O’Neil (2016)
- But: No guarantee that any specific measure will make matters better and not worse



Jørgensen/Hovy/Søgaard, 2015
Hovy & Søgaard, 2015

The WSJ Effect



The Consequences

https://hch19.cl.uni-heidelberg.de/program/slides/l/HCH19_lecture_Dirk_Hovy.pdf

Lecture survey in the chat!