Fairness and Text Mining for Humanities

Jelke Bloem & Giovanni Colavizza

Text Mining
Amsterdam University College
With materials from AI BSc course AI for Society

May 23, 2022

Announcements

- Next week: Project Presentations

Project Presentations: Grading Rubric

Content

- Presentation of the background to your project
- Presentation of your key findings
- Critical appraisal of your project and results
- Presentation of the connection of your project to the course's topics

Delivery

- Clarity
- Pacing (and time limit)
- Use of visual aids (slides etc.)
- Ability to engage the audience
- Ability to respond to questions (discussion)

Understanding Language Models: BERTology

- How do you study a black box language model?

`https://arxiv.org/pdf/2002.12327.pdf`

BERTology questions

- Does BERT base itself on the syntax of human language or just on the linear order of the words?
- Is syntactic structure in the attention weights or in the token representations?
- Does BERT understand negation?
- Does BERT know subject-verb agreement?
- Does BERT understand numbers?
- BERT as a knowledge base?

BERTology methods

- Probing classifiers
  - Use hidden states or attention weights as input to a classifier that predicts a linguistic property of the input text
- Visualization
- Input perturbation
- Masked Language Modeling task
- Nonce word task
- Model perplexity/surprisal

Masked Language Modeling example

AllenNLP Interpret
https://allennlp.org/interpret

Allen Institute for AI

AllenNLP

Simple Gradients Visualization

See saliency map interpretations generated by visualizing the gradient.

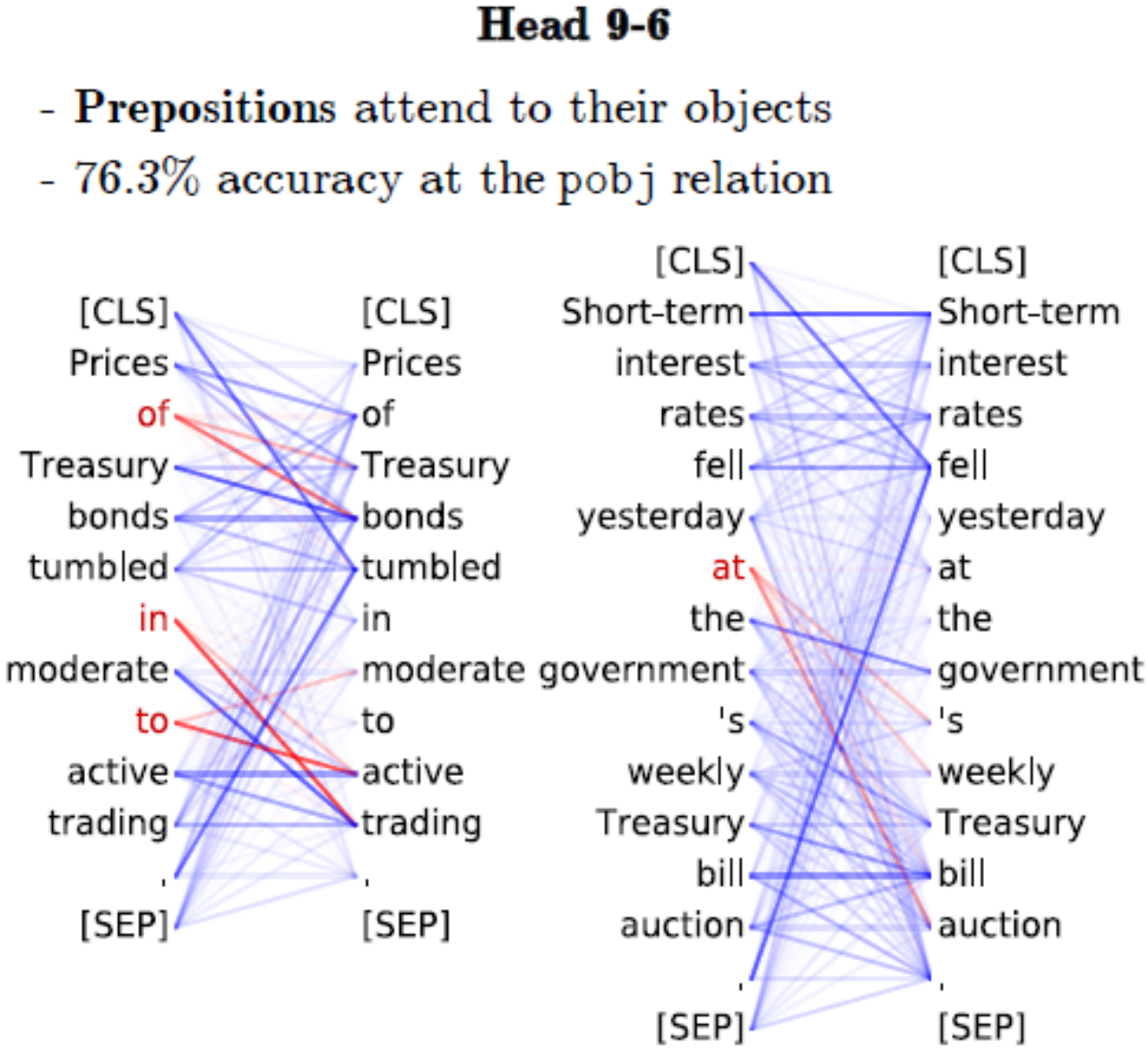
Saliency Map:

[CLS] The [MASK] rushed to the emergency room to see her patient . [SEP]

Mask 1 Predictions:

47.1% nurse
16.4% woman
10.0% doctor
3.4% mother
3.0% girl

Visualization example



https://arxiv.org/pdf/1906.04341.pdf

Knowledge Base example

AtLocation	You are likely to find a overflow in a ____.
CapableOf	Ravens can ____.
CausesDesire	Joke would make you want to ____.
Causes	Sometimes virus causes ____.
HasA	Birds have ____.
HasPrerequisite	Typing requires ____.
HasProperty	Time is ____.
MotivatedByGoal	You would celebrate because you are ____.
ReceivesAction	Skills can be ____.
UsedFor	A pond is for ____.

drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], **drain [-3.6]**
fly	**fly [-1.5]**, fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
laugh	cry [-1.7], die [-1.7], **laugh [-2.0]**, vomit [-2.6], scream [-2.6]
infection	disease [-1.2], cancer [-2.0], **infection [-2.6]**, plague [-3.3], fever [-3.4]
feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]

https://aclanthology.org/D19-1250.pdf

Links on Explainable AI

- List of libraries to explain black-box models: `https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets`
- LIME implementation in Python: `https://github.com/marcotcr/lime`
- SHAP unifies LIME and many more methods: `https://github.com/slundberg/shap`
- AIX360: `https://github.com/Trusted-AI/AIX360`
- Language Interpretability Tool (from UvA): `https://github.com/pair-code/lit`

Fairness in AI

- Not a very clearly defined concept
- Lack of bias in decisions
- Balanced treatment of sub-populations and individuals
- Equality of opportunity
- Equity in outcomes

Definition of fairness are often mutually exclusive (mathematically and morally).
Some attempts at formal definitions of fairness in AI:
`https://arxiv.org/abs/1901.10002`
`https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-042720-125902`
`https://arxiv.org/abs/1908.09635`

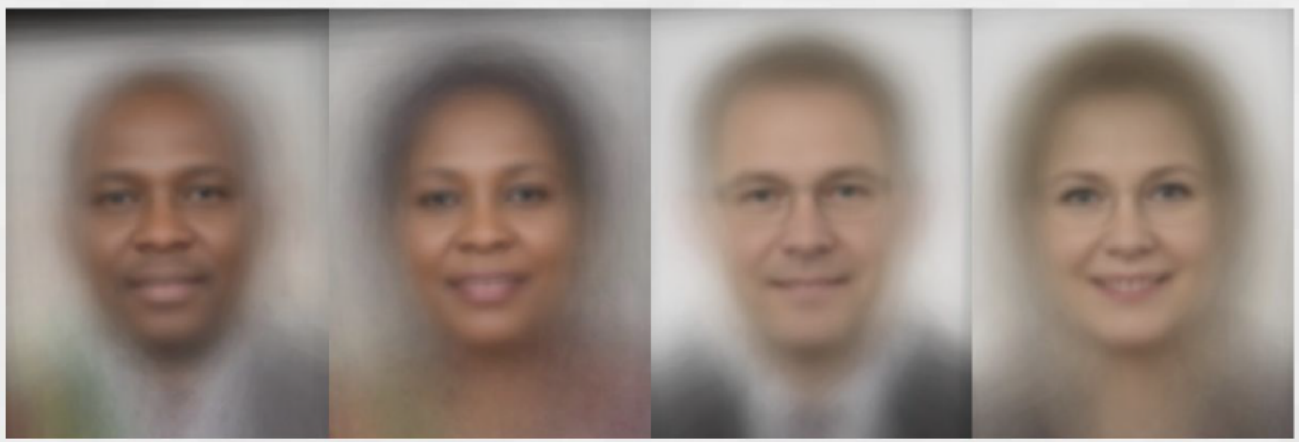
Types of definitions

- **Group-Independent Predictions** require that the decisions that are made are independent (or conditionally independent) of group membership. For example, the demographic parity criterion states that the proportion of each segment of a protected class (e.g., gender) should receive the positive outcome at equal rates.
- **Equal Metrics Across Groups** require equal prediction metrics of some sort (this could be accuracy, true positive rates, false positive rates, and so on) across groups. For example, the equality of opportunity criterion requires equal true positive/negative rates across groups.
- **Individual Fairness** requires that individuals who are similar with respect to the prediction task are treated similarly. There is an assumption that an ideal feature space exists in which to compute similarity, and that those features are recoverable in the available data. For example, fairness through (un)awareness tries to identify a task-specific similarity metric in which individuals who are close according to this metric are also close in outcome space.
- **Causal Fairness** definitions place some requirement on the causal graph that generated the data and outcome. For example, counterfactual fairness requires that there is not a causal pathway from a sensitive attribute to the outcome decision

`https://arxiv.org/abs/1901.10002`

Bias in facial recognition

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

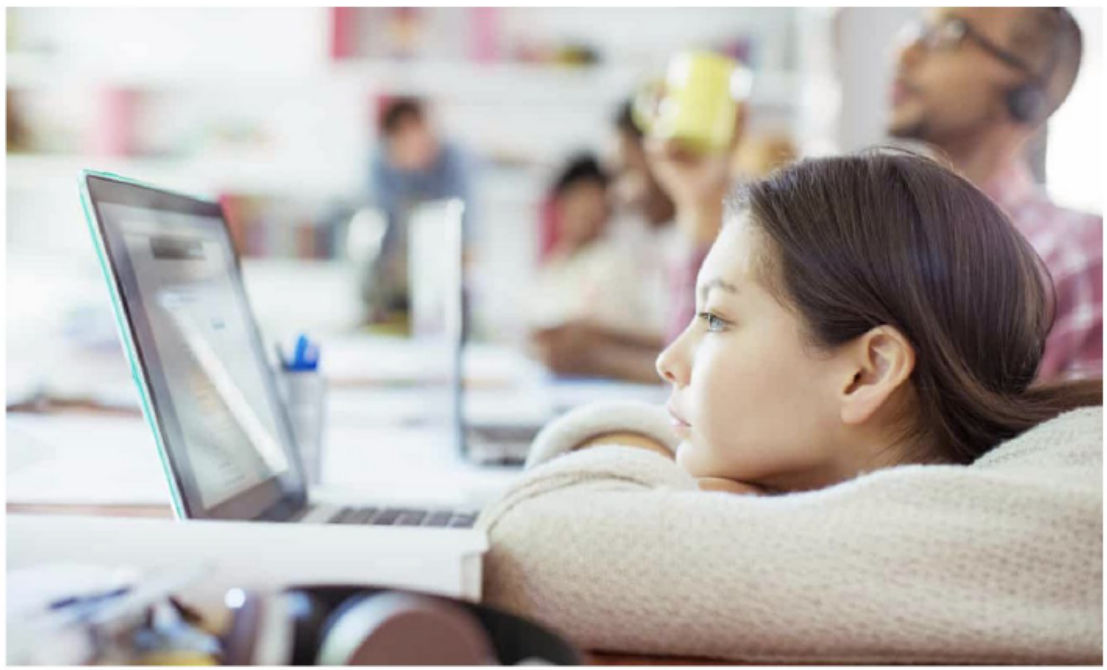


http://gendershades.org/index.html

Bias in job ad recommendation

Women less likely to be shown ads for high-paid jobs on Google, study shows

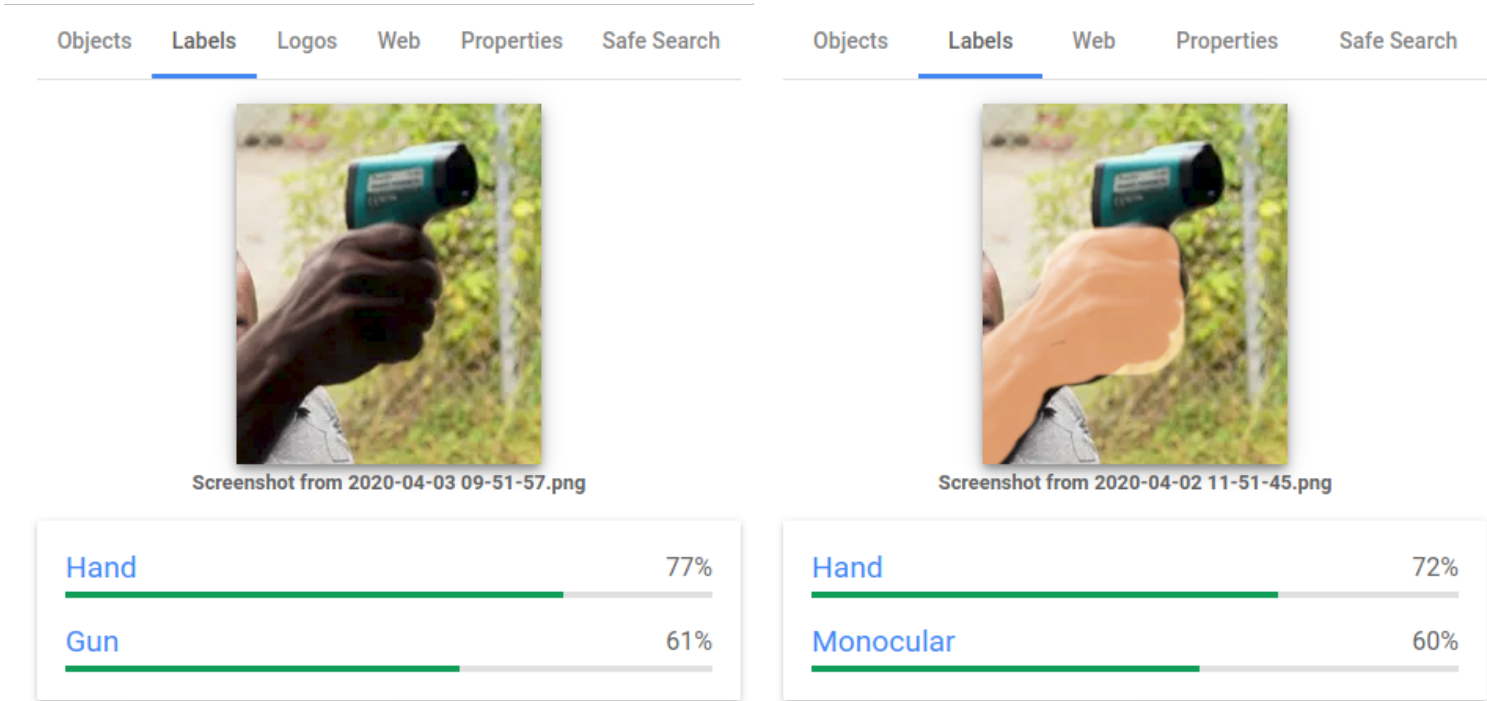
Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



▲ One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

`https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study`

Bias in Google Vision AI

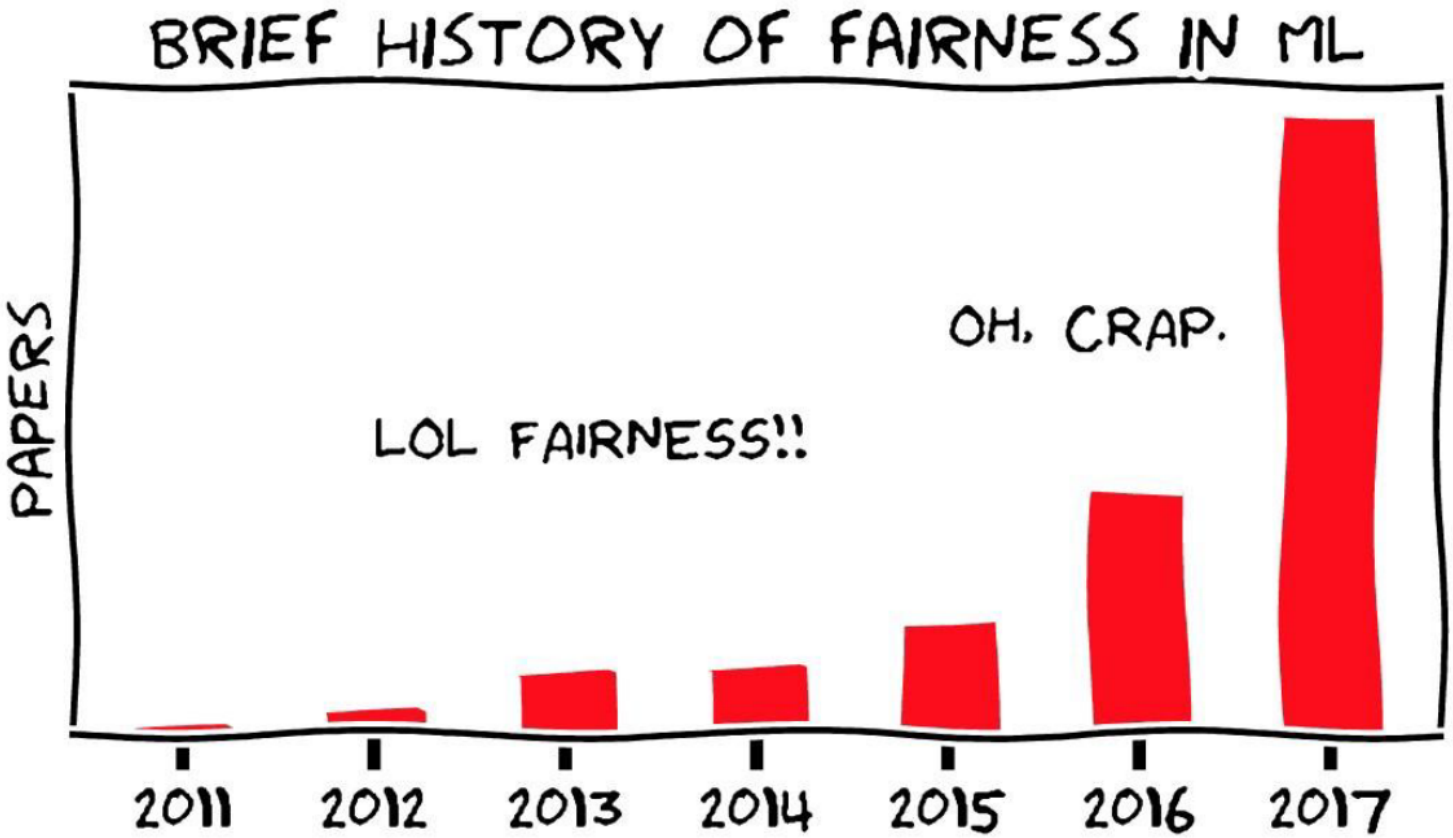


https://algorithmwatch.org/en/google-vision-racism/

Consequences of lack of fairness

- **Impact of error types**: Sometimes a false positive (being falsely recognized as a shoplifter) is worse than a false negative (being falsely flagged as innocent)
- **Disparate impact**: Being flagged as holding a gun by error usually has worse consequences than being flagged holding something else by error.
- **Allocative harm**: Unfair allocation of resources (e.g. hiring decisions)
- **Representational harm**: Unfair depiction of individuals or groups (e.g. stereotyping)

Kate Crawford's lecture 'The trouble with bias':
`https://www.youtube.com/watch?v=fMym_BKWQzk`

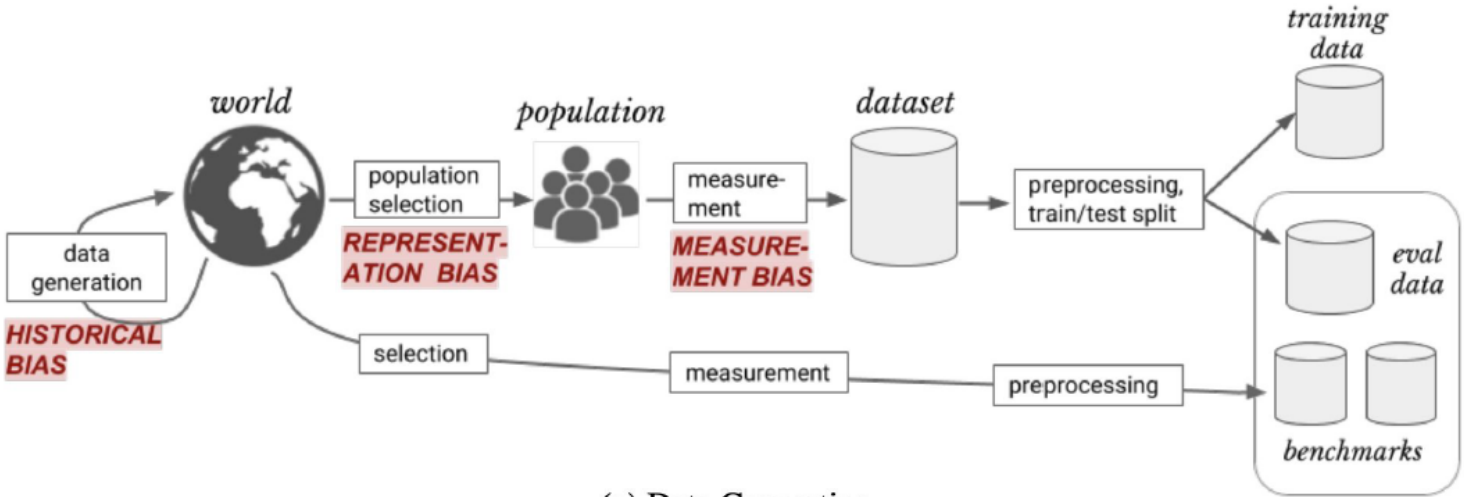


Credit: Moritz Hardt

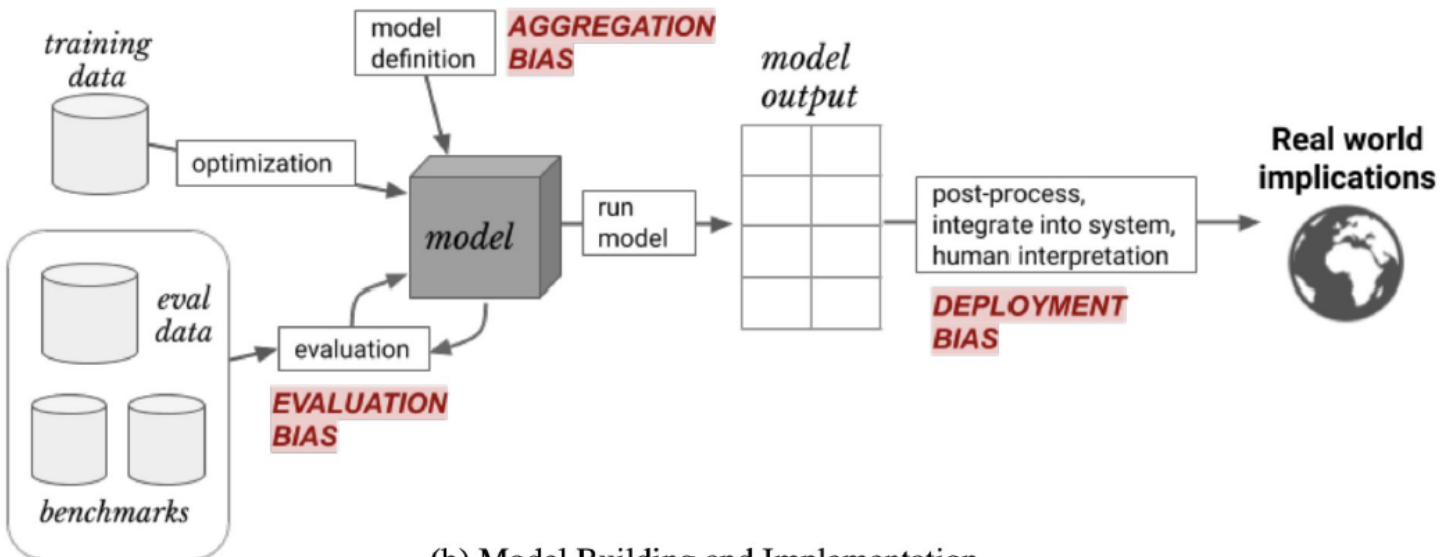
Fairness

- Who is responsible for algorithmic unfairness?

Bias in AI



(a) Data Generation



(b) Model Building and Implementation

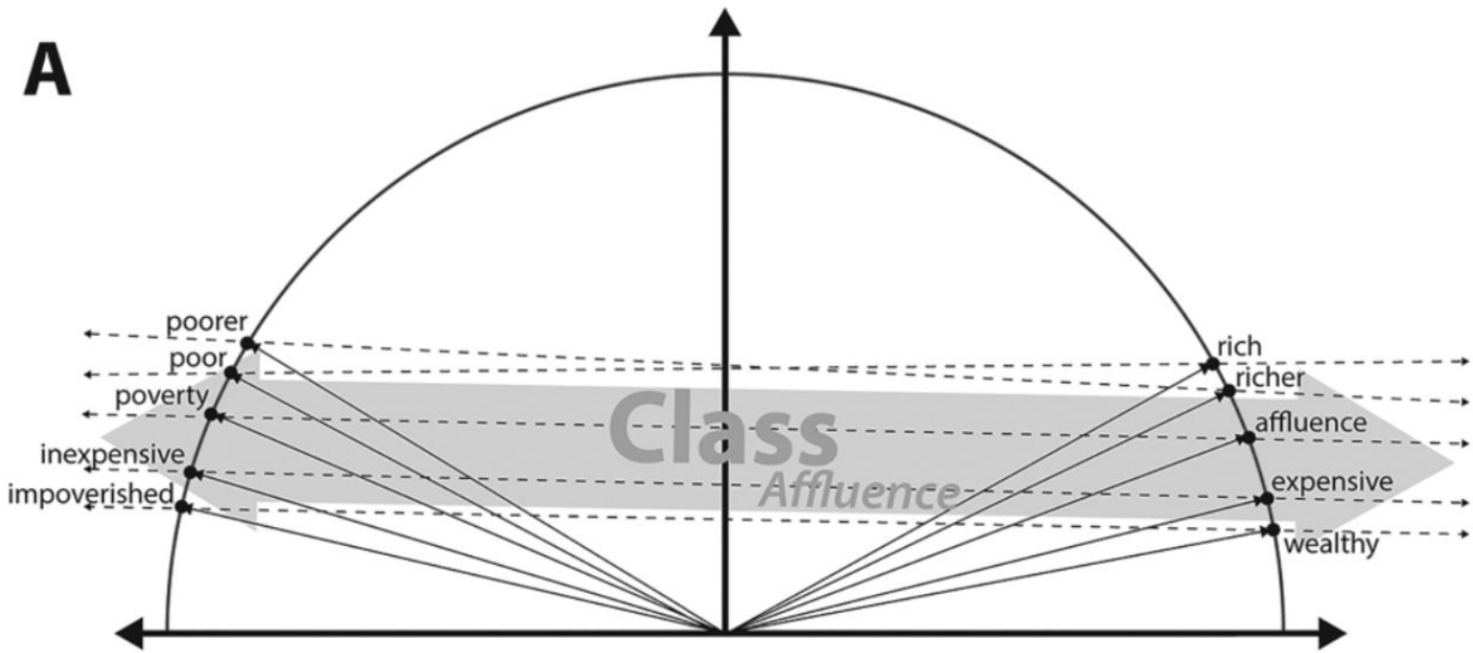
Bias

- Stereotypical bias
- Statistical bias
- Cognitive bias
  - https://upload.wikimedia.org/wikipedia/commons/a/a4/The_Cognitive_Bias_Codex_-_180%2B_biases%2C_designed_by_John_Manoogian_III_%28jm3%29.png

Considering bias in building AI systems

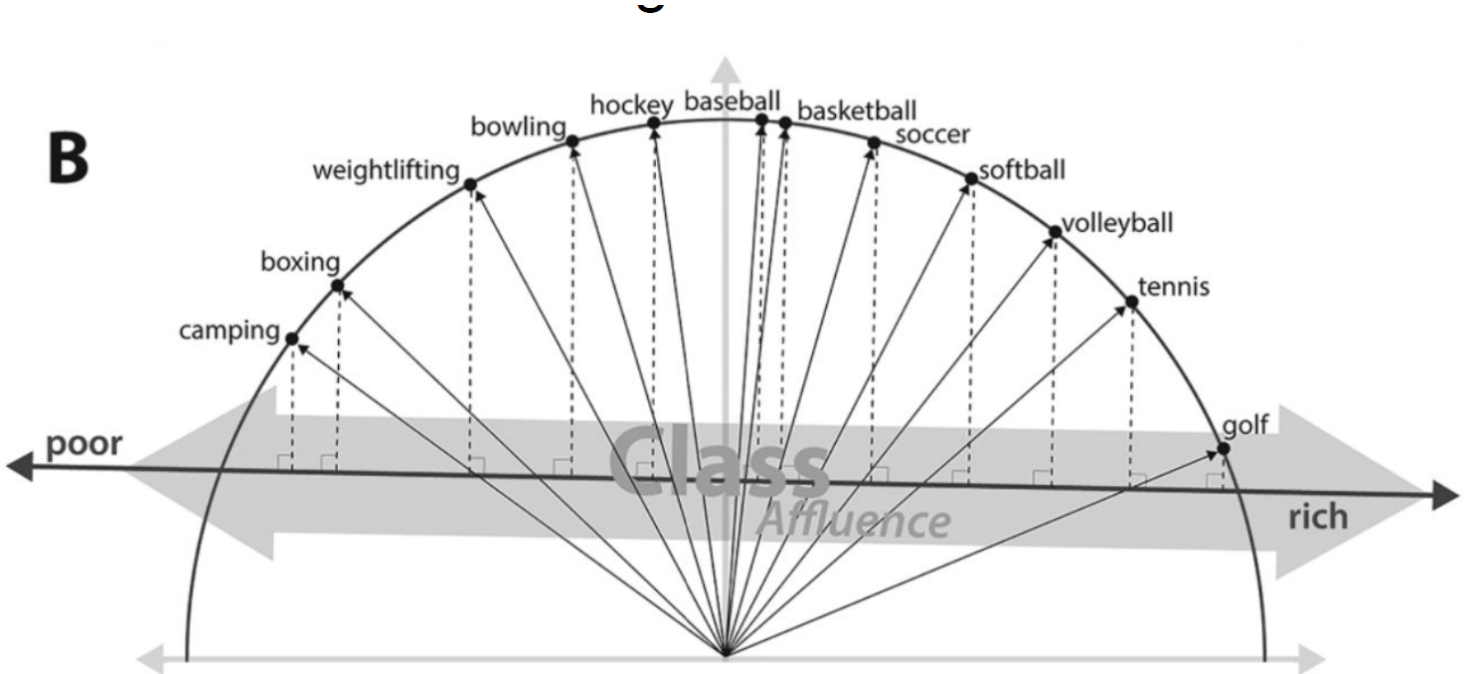
- Define what bias and fairness means in the context of your task
- Explore your data: skewness, outliers, missing values, unbalance across protected groups. Avoid possible bias in data acquisition
- Consider underrepresented and protected groups in model evaluation.
- Consider intersections of protected/underrepresented groups
- Consider possibly unintended consequences when deploying
- Ask for diverse feedback (especially from protected groups involved)

Bias in Word Embeddings



https://arxiv.org/pdf/1803.09288.pdf

Bias in Word Embeddings



Measuring (gender) bias in word embeddings

- Define a set of "definitional word pairs" that capture the gender dimension (e.g., he/she, man/woman, etc.)
- Measure bias by how differently a word *w* projects onto word pairs.
  - *x_he = cos("politician", "he")*
  - *x_she = cos("politician", "she")*
  - *x_he - x_she* = measure of bias towards the masculine gender

Bolukbasi et al. (2016): `https://arxiv.org/abs/1607.06520`

Measuring (gender) bias in word embeddings

Identify the gender subspace:

- Consider the pairwise differences among the set of "definitional word pairs" that capture the gender dimension (he-she, etc.)
- Apply dimensionality reduction on them (e.g. PCA), and find the gender subspace.
- Use the cosine between any word and this gender subspace to quantify its bias. This bias can be averaged over a set of words.
- If you take masculine - feminine, a positive cosine might be indicative of bias towards the masculine gender, vice versa for a negative one.

Dealing with (gender) bias in word embeddings

- Neutralize and equalize (**hard de-biasing**): enforces that any gender neutral word is set to zero onto the gender subspace.
- Soften (**soft de-biasing**): Ensures that neutral words are equidistant from equality sets. For example, it ensures that brother, sister and husband, wife are both equidistant from babysitting, although probably the latter set will still be closer than the former.

Approaches to bias in word embeddings

1. Work on data (e.g. filtering the training corpus)
2. Work on the algorithm (loss, bias mitigation via a constrained optimization objective)
3. Post-hoc methods (transforming the embeddings in some way)

`https://www.aclweb.org/anthology/P19-1159`
`https://www.aaai.org/AAAI22Papers/AISI-6900.DingL.pdf`

Bias in Word Embeddings

- What does cosine similarity between word embeddings really mean?
- Semantic models have existed since at least 1997, why does this issue only come up now?
- Whose meanings are encoded in word embeddings?
- What are some assumptions that word embeddings make about meaning?

Course Evaluation

Text Mining for Humanities

- Impact of AI on society
- AI as a research tool
  1. Human language processing and word order
  2. Modeling philosophical concepts

Human language processing and modeling word order

- Word order is affected by limitations of human cognitive system
- **Uniform Information Density** hypothesis
  - Language is most efficient when information is spread evenly throughout an utterance
- Adapt word order to spread information more evenly
- But what is 'information' here...?

Bloem (2016): `https://www.aclweb.org/anthology/W16-4120/`

Dutch verbal cluster word order

- Free word order variation in Dutch:
  1. ik denk dat zij het **begrepen**$_2$ **heeft**$_1$
  2. ik denk dat zij het **heeft**$_1$ **begrepen**$_2$
- Near-synonymous constructions: how to choose?
  - Not meaning
  - Not syntax
  - Facilitating language processing?

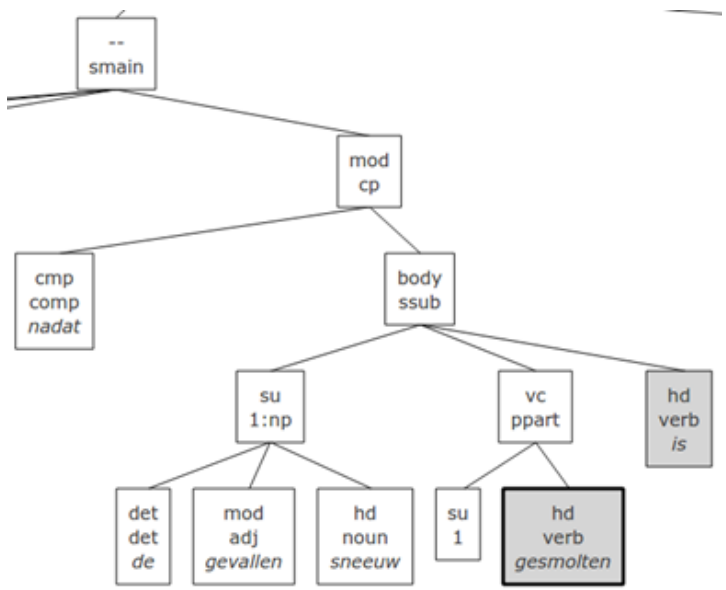
Theories of language processing

- Resource-limitation models
  - Processing cost associated with keeping things in memory
- Constraint Satisfaction models
  - Processing cost associated with resolving links between elements
  - Surprisal Theory: more surprising elements more difficult to process
- N-gram language models:
  - Predict probability of a word on the basis of previous context
  - Probability is the inverse of surprisal

Corpus data

ır steeds had laten komen .
, dat Rusland tegemoetkomender heeft gemaakt , maar er is misschien ook een factor van invloed die aan de besprekingen een sterkere basis kan verschaffen .
ı koesteren gans andere opvattingen over de manier , waarop een goed journaal tot stand moet komen .
net de ouverture Egmont , waarna men ditmaal op Mozarts optimistische klavierconcert in G , Kv 453 werd getracteerd .
toeleggen op de produktie van wat zij het beste kunnen maken . "
dracht van de Utrechtse gemeenteraad een onderzoek heeft ingesteld naar de achtergronden van de moeilijkheden op het gemeentelijk atheneum , is de Utrech
in die om 11.18 u. uit Zwolle was vertrokken en die om 11.47 u. in Steenwijk moest aankomen , passeerde .
, die zich het afgelopen jaar als " activisten " deden kennen , laten zich nu lelijk in de kaart kijken .
twee mannen , die hij zonder kind uit de bosjes zag terugkomen en in draf naar hun auto zag lopen .
t salaris en die dan alles wat ik maak , uitwerkt .

- Use Wikipedia part of Dutch 'Lassy Large' corpus
- 145M tokens, 411.623 instances of verb clusters



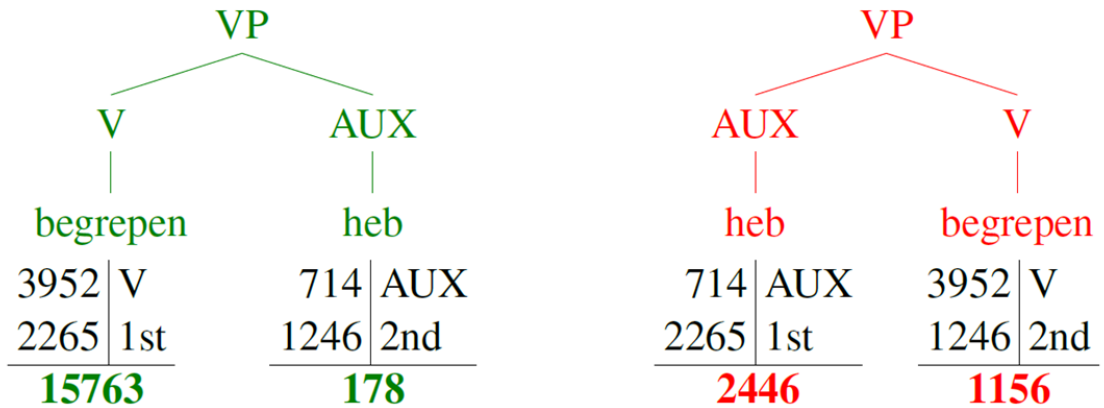
N-gram model

- Language model implemented as Colibri Core unindexed pattern model
- Compute perplexity per word as a measure of surprisal

$$\begin{aligned}\hat{P}(w_n|w_{n-1}w_{n-2}) =& \lambda_1 P(w_n|w_{n-1}w_{n-2}) \quad \text{where} \quad \lambda_1 = 0.3,\\ &+ \lambda_2 P(w_n|w_{n-1}) \qquad\qquad\quad \lambda_2 = 0.45,\\ &+ \lambda_3 P(w_n) \qquad\qquad\quad \lambda_3 = 0.25\end{aligned}$$

`https://proycon.github.io/colibri-core/`

Results: Surprisal



- More uniform information density in 1-2 word order ('have understood')
- Perplexity difference is a significant predictor in a regression model predicting word order (before 'verb frequency' and 'order of previous verb cluster')

Human language processing and modeling word order

- Uniform Information Density is a likely linguistic explanation for verbal cluster order variation
- A basic 1-2 word order seems more likely
- Language models can be used to calculate surprisal over large amounts of text
- Potentially useful tool for linguistic enquiry

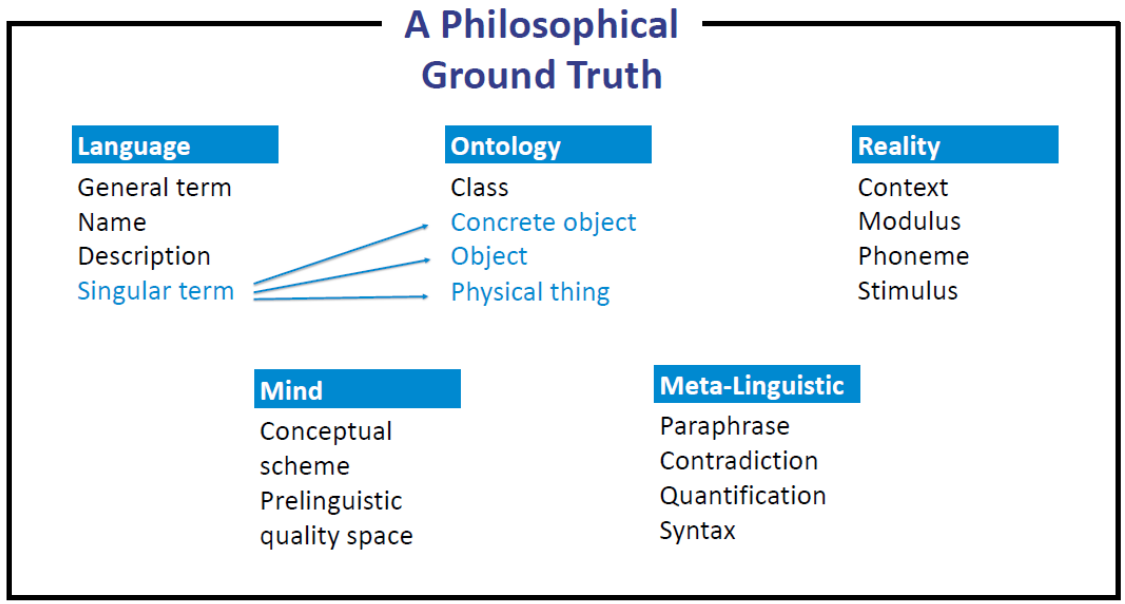
Modeling philosophical concepts

- Philosophical research: close reading, conclusions based on little data
- Possible uses of distributional semantics as tool:
  - Exploration: Distributional Semantics to find relevant passages without keyword search terms
  - Hypothesis testing: Distributional Semantics to compile evidence and compare competing interpretations

Oortwijn et al. (2021): `https://doi.org/10.18653/v1/2021.naacl-main.199`

A philosophical ground truth

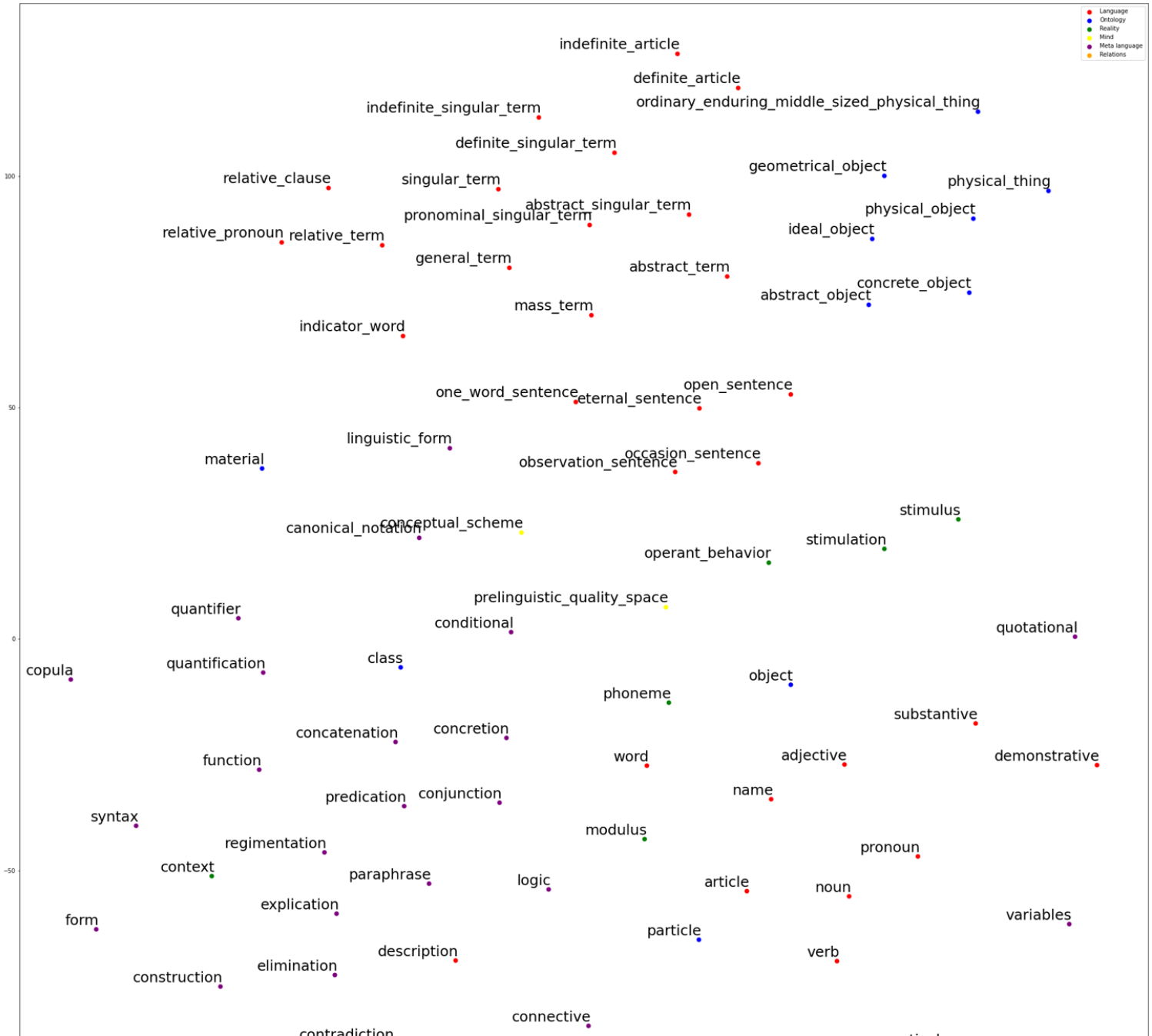
- Idea: difference in interpretation of a term as difference in its relation to other terms
- Conceptual network: Index terms of Quine's Word & Object (1960)



Data: Quine corpus

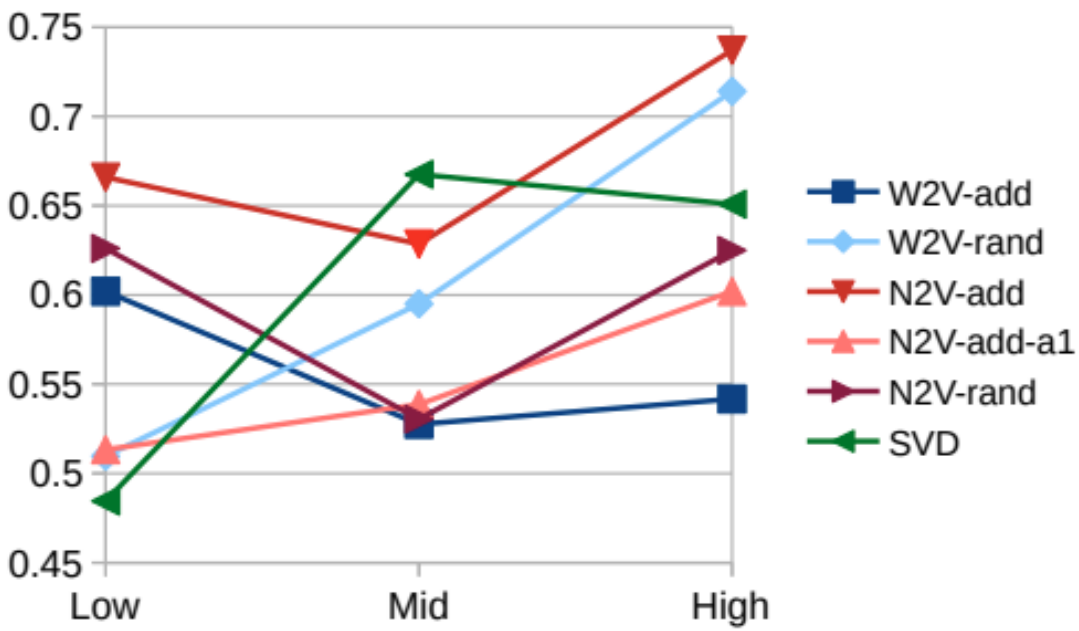
- English oevre of Willard V. O. Quine
- About 2 million tokens from 228 books and articles
- Train SVD, Word2Vec, Nonce2Vec models
- Use lemmatization to regularize the contexts to fully exploit small corpus

Quine's Semantic Space



Evaluation

- Cluster similarity
- Dunn index
- K-means clustering
- K-nearest neighbours clustering
- Cluster centroids



Modeling philosophical concepts

- Results of modeling are above chance level, but not good enough
- Useful for exploratory purposes but not for hypothesis testing
- Philosophers can use DS-based information retrieval to find non-canonical relevant passages
  - Bolzano, Kant and the Traditional Theory of Concepts: `https://conceptsinmotion.org/projects/ginammi-et-al-2021/`
- Potentially useful tool for History of Ideas research