

Information Retrieval (IN4325)

Bias in Natural Language Processing

Oana Inel

Paper Reviews

- P3 review due March 6

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

Plagiarism

- Turnitin
 - checks whether words or (parts of) sentences correspond to source material
 - compares with:
 - internet sources
 - student thesis databases (worldwide)
 - scientific papers

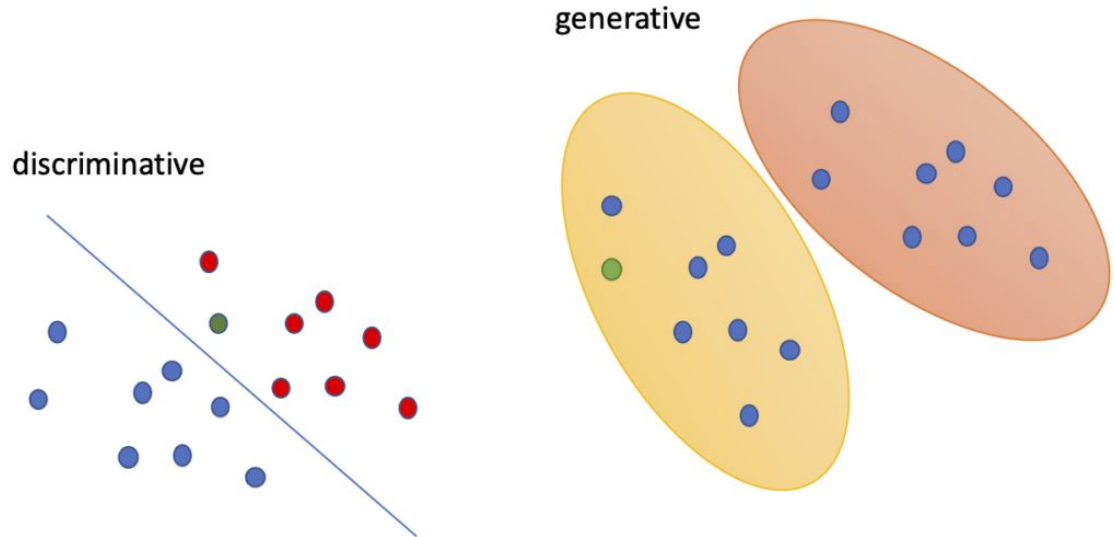
Last week

- Machine learning for NLP
 - classes of machine learning problems
 - feature selection/extraction
 - ML techniques
 - generative
 - discriminative
- Natural Language Generation (NLG)
- NLG pipeline:
 - document planning
 - microplanning
 - realisation

Terminology

Classifier

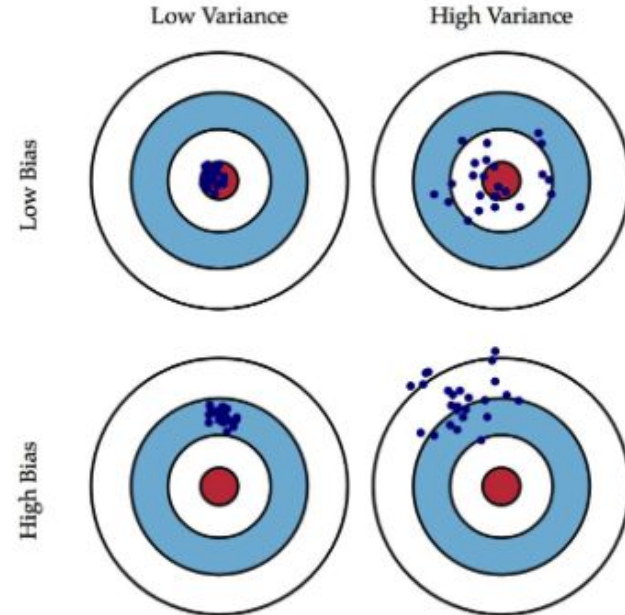
- Example of a generative classifier
- Example of a discriminative classifier



Terminology

Classifier

- Example of a generative classifier
- Example of a discriminative classifier
- Bias versus variance trade-off



Terminology

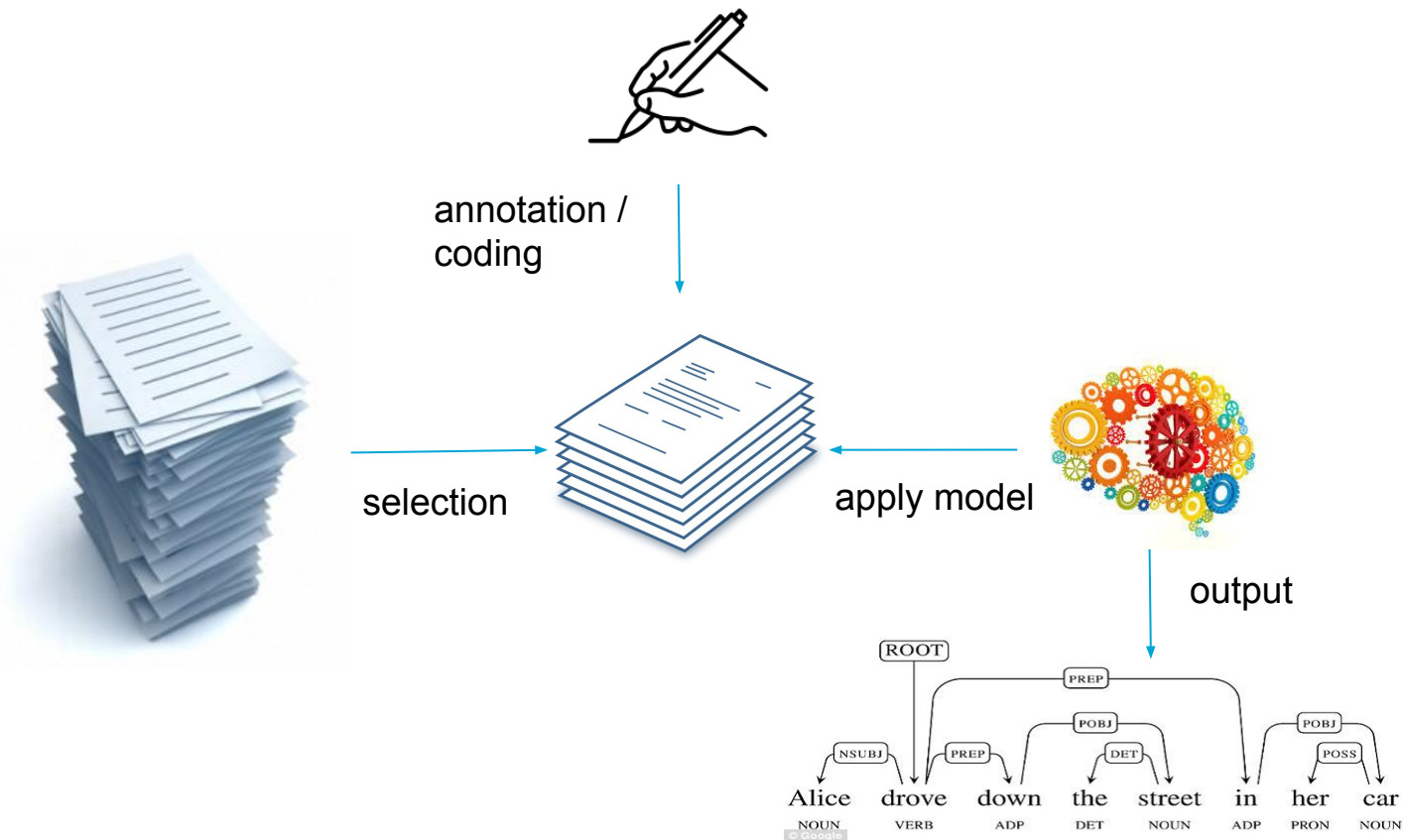
NLG pipeline

- document planning
- microplanning
- realisation

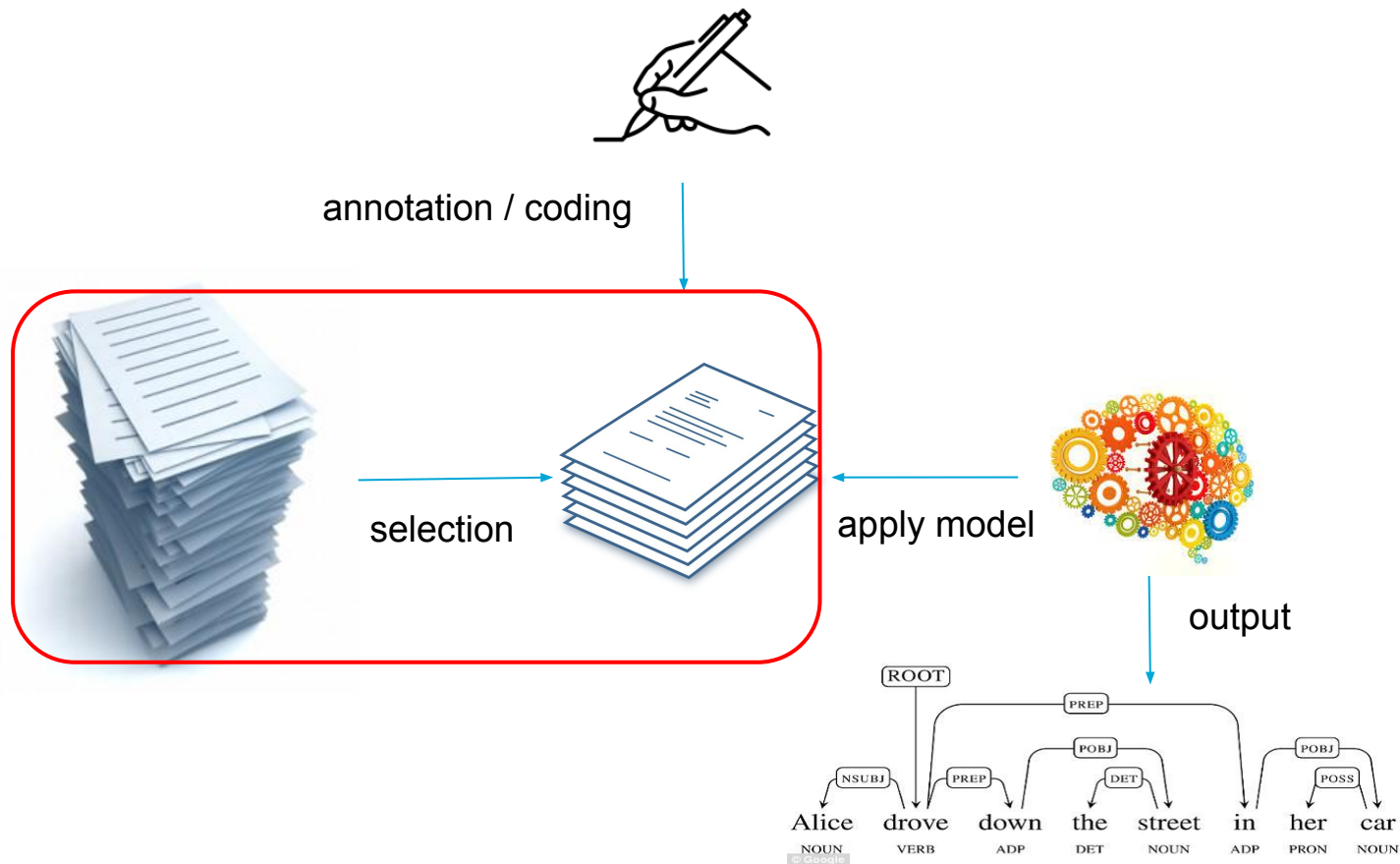
This week

- Bias(es) in NLP
 - selection bias
 - annotation bias
 - machine learning bias
- Word embeddings

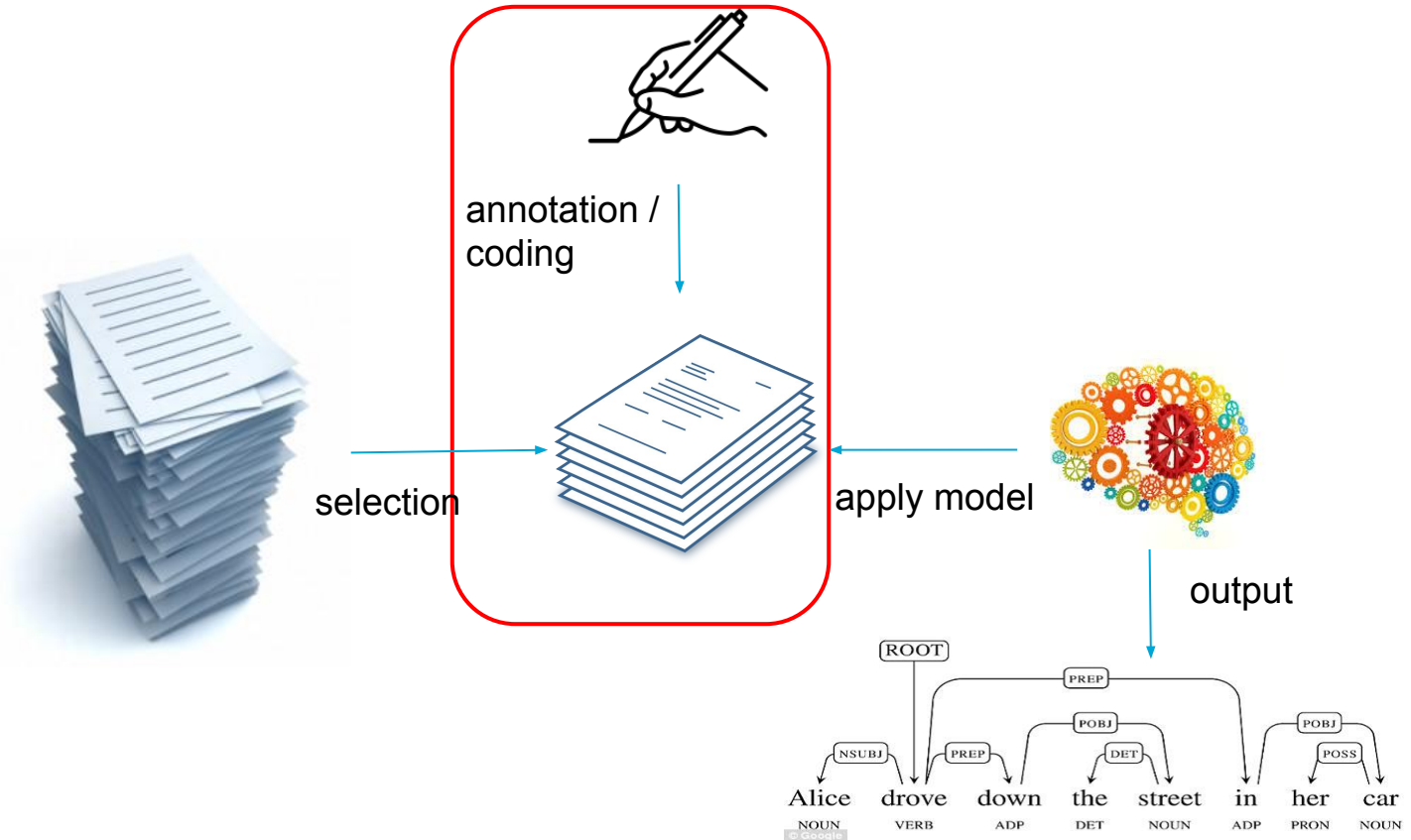
Typical NLP Pipeline



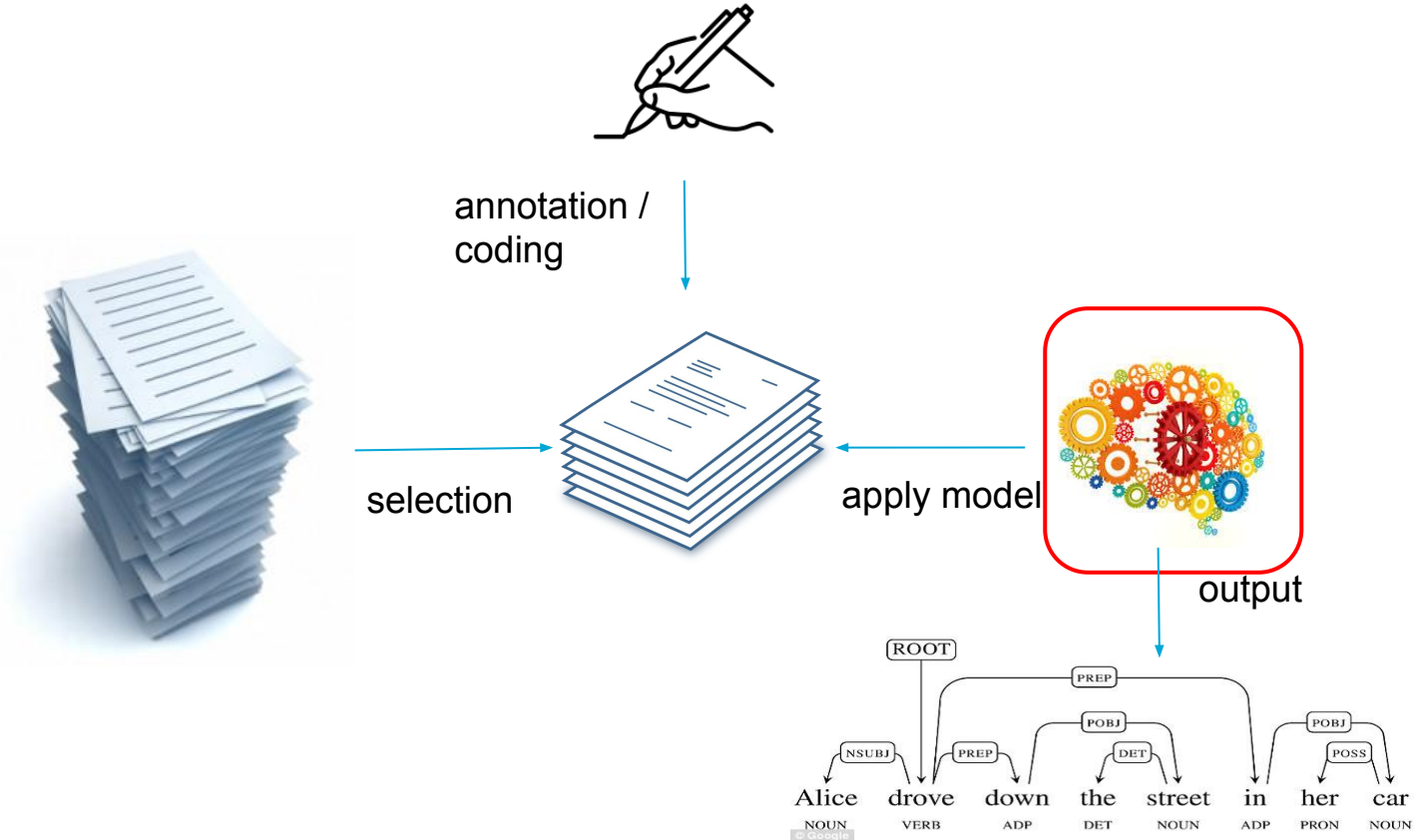
Selection Bias



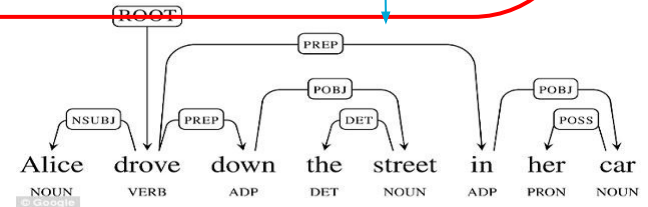
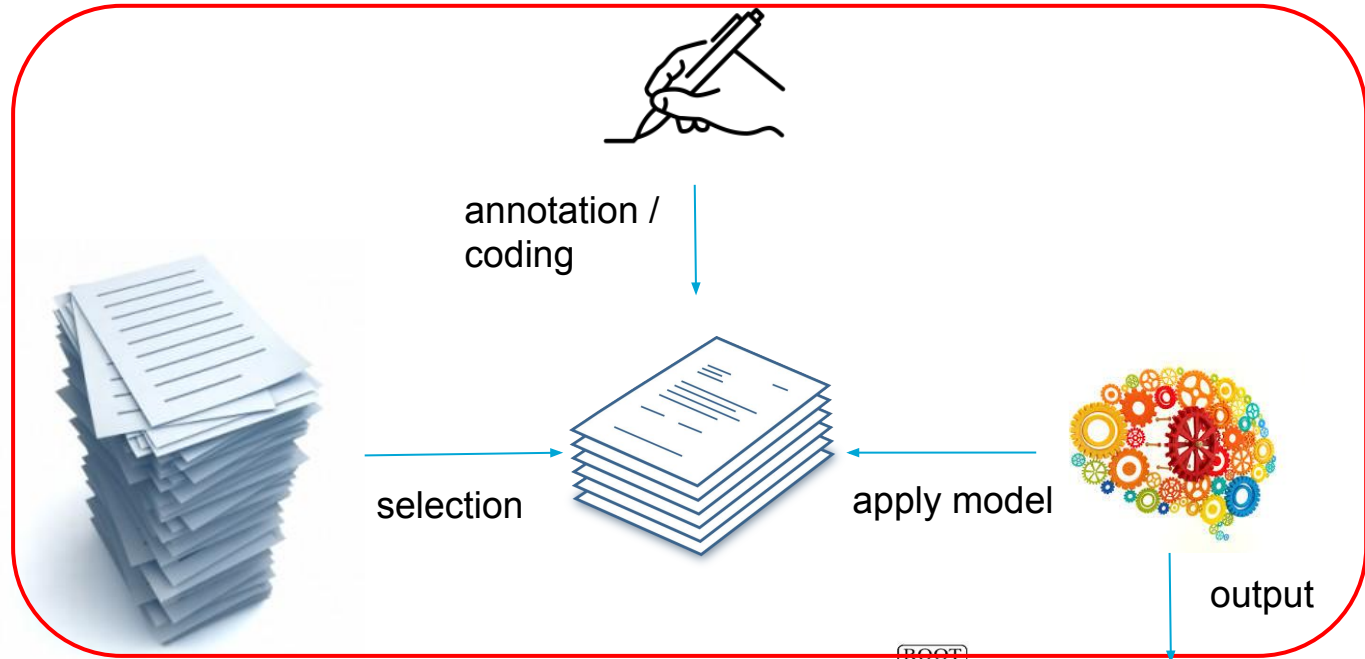
Annotation / Coding Bias



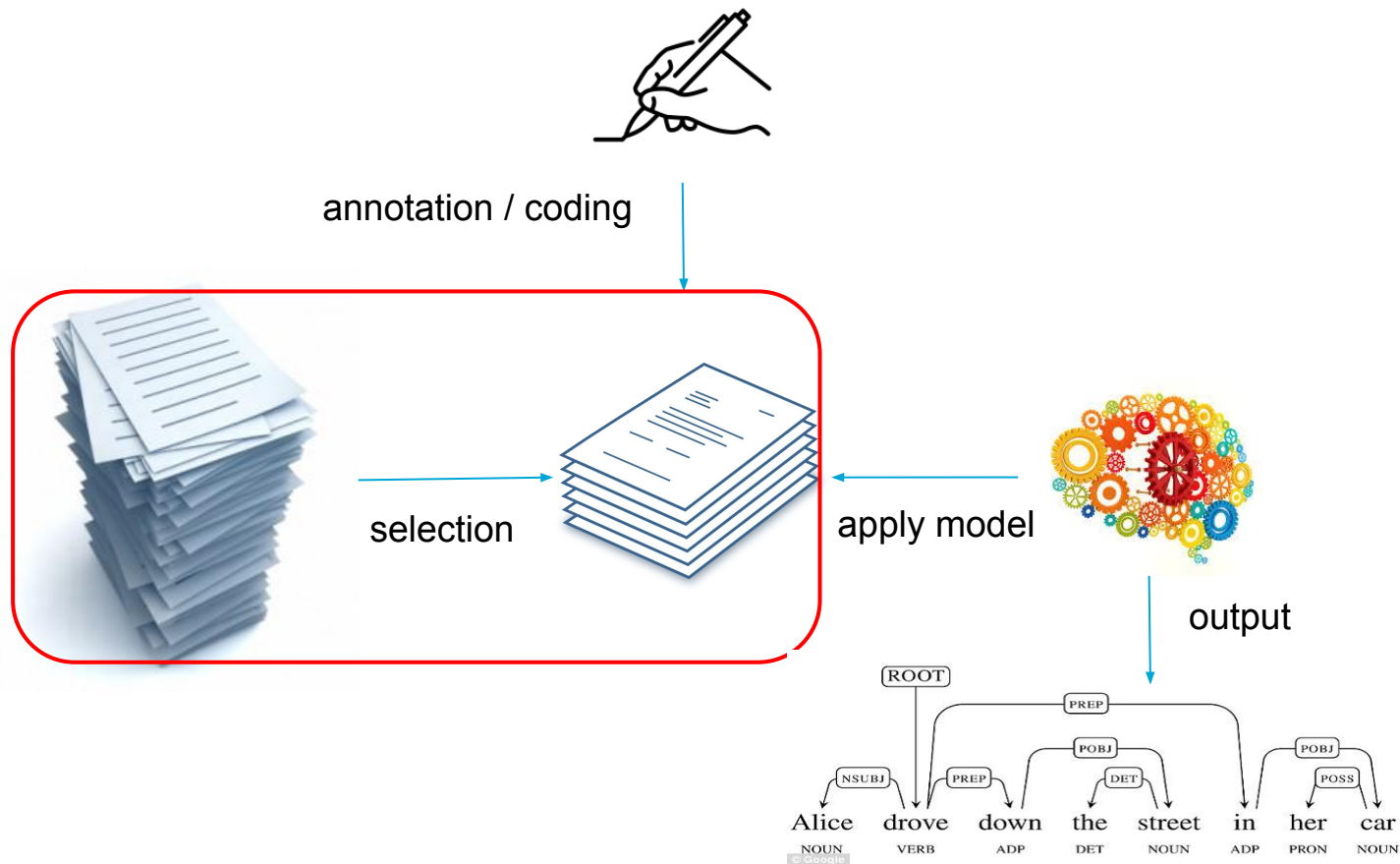
Machine Learning / Model Bias



Bias(es) in NLP Pipeline



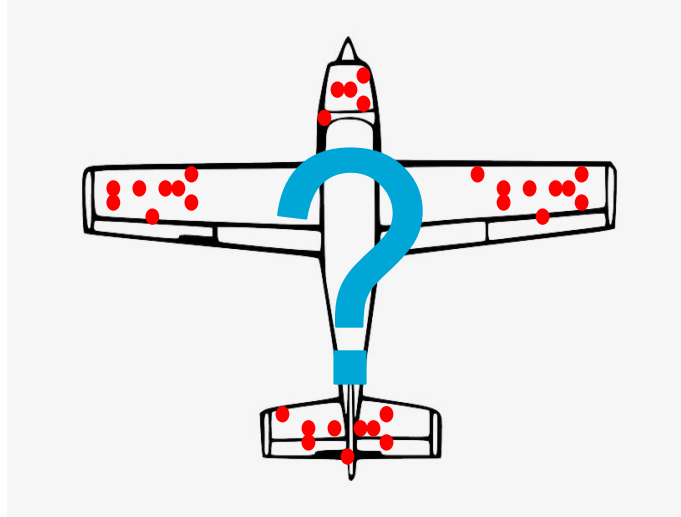
Selection Bias



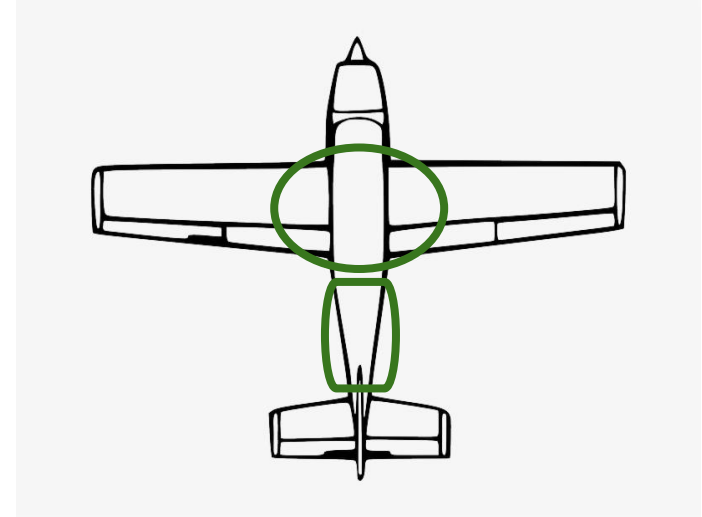
Selection Bias

- Experimental error
- Consequence of the method of selecting the data
- Sample data is not representative for the entire dataset (or population)
- Types: exclusion, time interval, sampling

Historical Example



Intuition: reinforce the armor in the most frequently hit areas

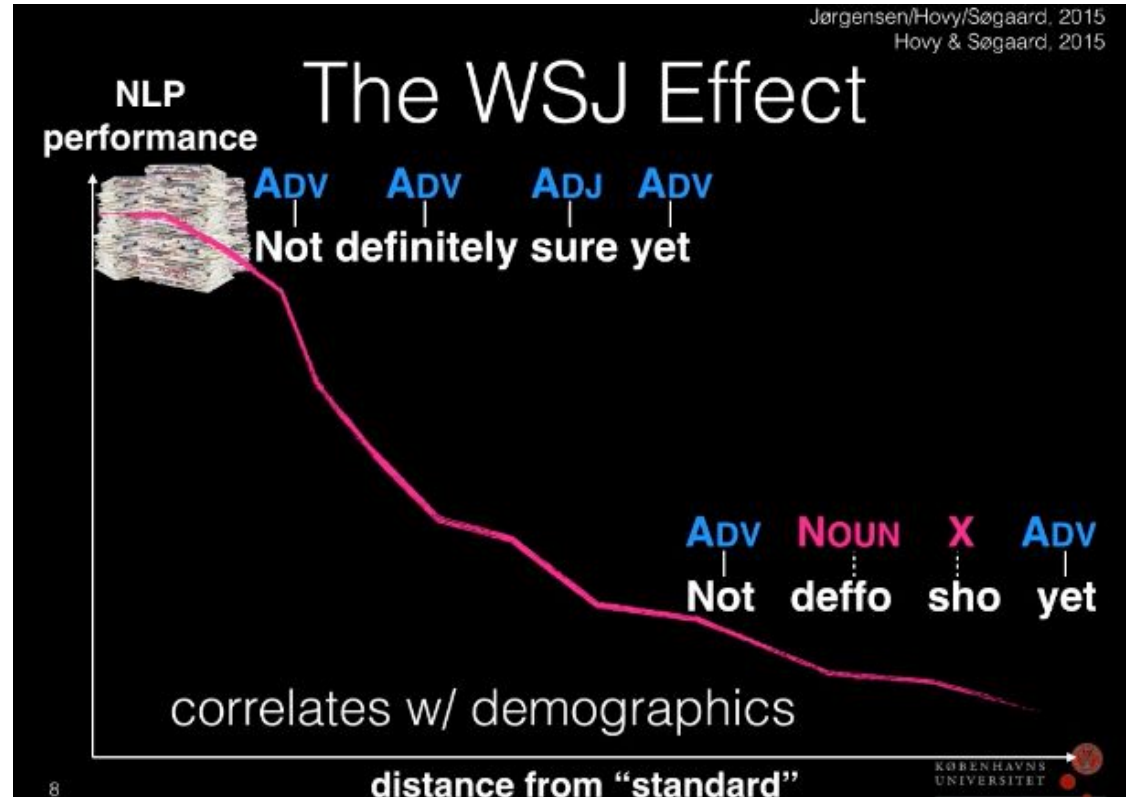


Lesson Learned: think critically when doing data collection/selection

Language

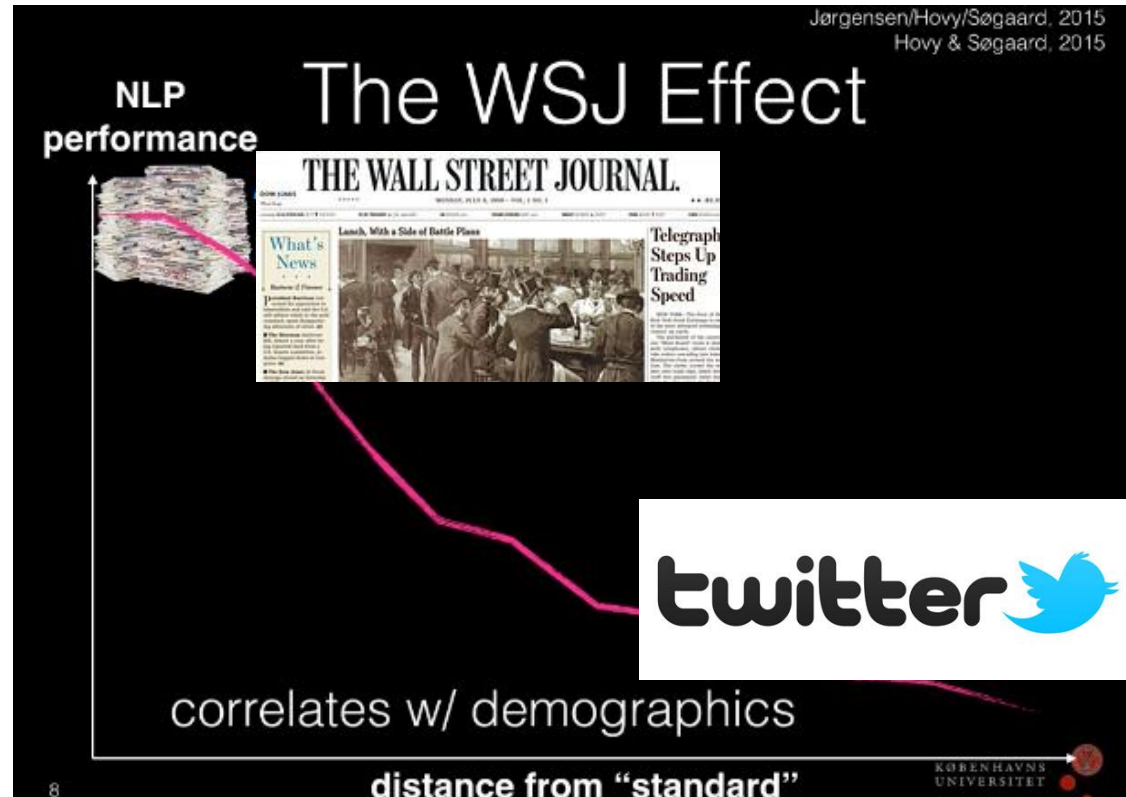
- information
- communication with other people
- communication within groups or contexts
- takes a variety of forms
 - style shifting
 - accents, dialects
 - register, jargon, slang
- social context: age, gender, education, demographics

Language effect on NLP performance

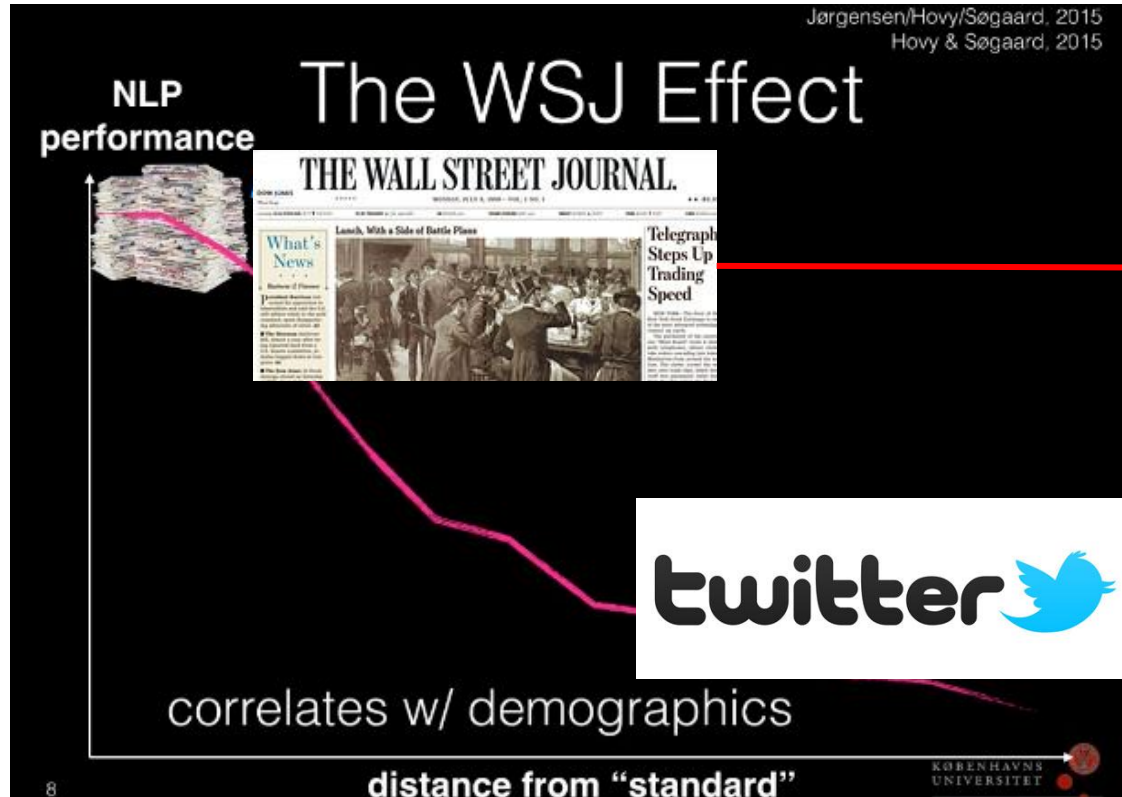


Slide from Dirk Hovy - Reading between the lines - the hidden bias of NLP

Language effect on NLP performance



Sample or exclusion bias



Demographic Bias in NLP

- running example: POS tagging
 - demographic group:
 - **age: under 35 (U35) and over 45 (O45)**
 - gender: male and female
 - **region: w/o African-American words**

Age Bias in NLP

Language	Age Group	Avg. POS Accuracy
German	Under 35	86.68
	Over 45	88.22
English	Under 35	88.08
	Over 45	88.33

- tested on user reviews in English and German
- taggers are better for the older group (O45) than for the younger (U35)
- accuracy is significantly worse for U35

But where does this difference come from?

Age Bias in NLP

- difference of vocabulary
- lexical changes
 - use of neologisms
 - spelling variations
 - linguistic changes
 - grammatical differences

Hovy, Dirk, and Anders Søgaard. "Tagging performance correlates with author age." *ACL 2015 (Volume 2: Short Papers)*, vol. 2, pp. 483-488. 2015.

Region Bias in NLP

Language	Stanford	Gate	Ark
AAVE*	61.4	79.1	77.5
non-AAVE	74.5	83.3	77.9
delta	13.1	4.2	0.4

- tested on a sample of 200 tweets
- performance in terms of accuracy
- Stanford: trained on newswire
- Gate & Ark: adapted to Twitter

*AAVE: African-American Vernacular English

Jørgensen, Anna, Dirk Hovy, and Anders Søgaard. "Challenges of studying and processing dialects in social media." In *Proceedings of the Workshop on Noisy User-generated Text*, pp. 9-18. 2015.

Region Bias in NLP

Language	Stanford	Gate	Ark
AAVE*	61.4	79.1	77.5
non-AAVE	74.5	83.3	77.9
delta	13.1	4.2	0.4

- tested on a sample of 200 tweets
- performance in terms of accuracy
- Stanford: trained on newswire
- Gate & Ark: adapted to Twitter

*AAVE: African-American Vernacular English

Jørgensen, Anna, Dirk Hovy, and Anders Søgaard. "Challenges of studying and processing dialects in social media." In *Proceedings of the Workshop on Noisy User-generated Text*, pp. 9-18. 2015.

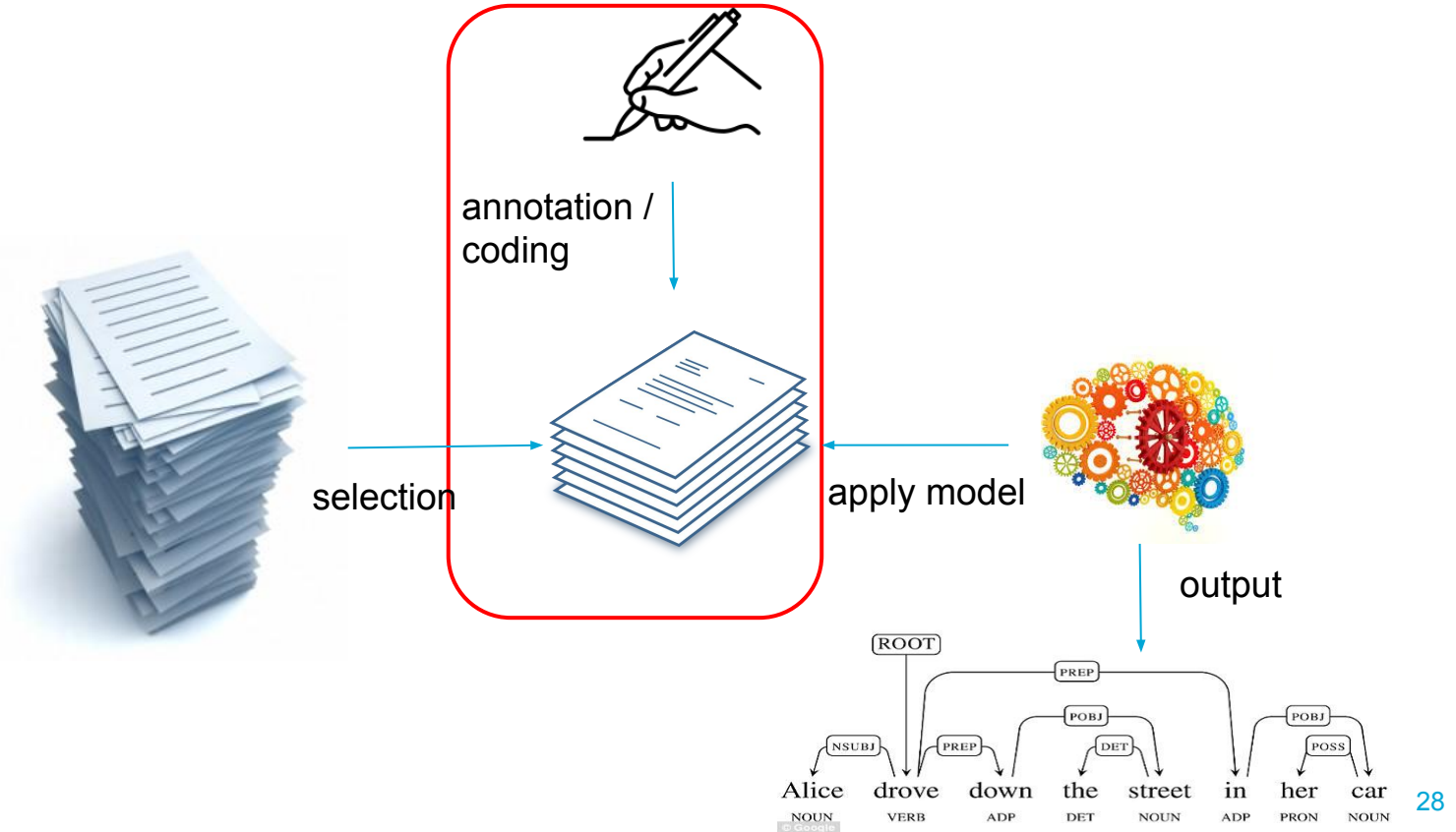
Selection Bias in NLP

- common training datasets for NLP are biased towards older people language
- statistical significant difference in performance
- can lead to overfitting
- cascade behavior in performance decrease
POS tagging -> lemmatization -> sentiment analysis

How to deal with selection bias?

- balance the training groups
- train separate models for each demographic group
- regularization (automatically penalizes extra features that you use in your model)
- priors
- sampling

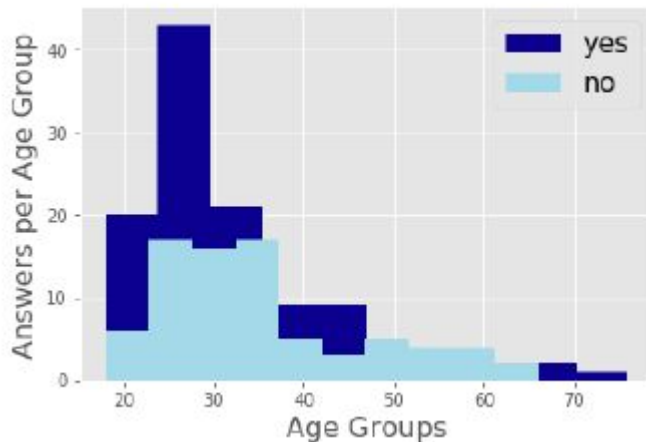
Annotation / Coding Bias



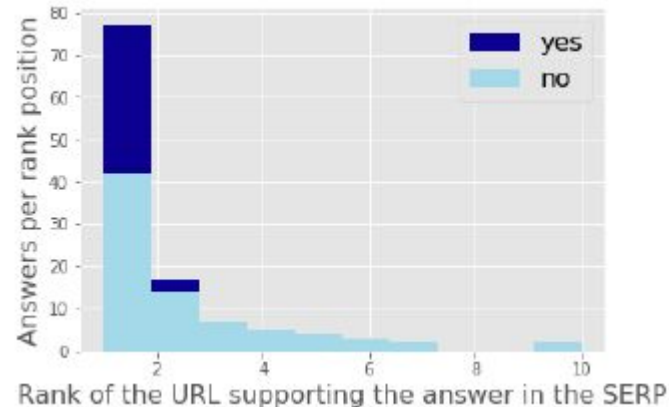
Annotation Type: Fact Checking

- controversial fact and fake news verification
- crowd distribution: US and India
- binary answer: yes/no or true/false
- answer motivation:
 - what search engine?
 - which search query?
 - which URL position?
- *potential bias*: annotators implicit bias

Catalonia Independence



- 200 workers in total:
 - 116 said yes
 - 84 said no
- Age bias: “yes” answers tend to come from younger people
- significant result (t-test $p < 0.5$)



- Search result rank bias: annotators who go lower in the search result to find supporting evidence tend to answer “no”
- significant result (t-test $p < 0.01$)

Pope News

- Population bias: annotators from India more inclined to believe the news is true (significant result, $p < 0.01$)
- Age bias: younger population easier to be mislead



Annotation Type: Relevance

- topical relevance
- evaluate the effectiveness of search engines
- given a search query and a document: *is the document highly relevant, relevant or not relevant wrt the topic?*
- *potential biases:*
 - relevance scale
 - expert bias

What do experts say?

Identify documents that discuss opposition to the use of the euro, the European currency.

A common currency would take many of these decisions out of national hands, the main reason why so many Conservatives in Britain oppose introducing the euro there.

Highly
relevant

Jack Straw, left, Britain's new foreign secretary, said the country's euro policy remained unchanged, producing a brief recovery in the pound from a 15-year low set last week.

Highly
relevant

Greece never had a prayer of joining the first round of countries eligible for Europe's common currency, so its exclusion from the list of 11 countries ready to adopt the euro in 1999 was not an issue here.

Highly
relevant

What does a lay people say?

Identify documents that discuss opposition to the use of the euro, the European currency.

A common currency would take many of these decisions out of national hands, the main reason why so many Conservatives in Britain oppose introducing the euro there.

0.89
relevant

Jack Straw, left, Britain's new foreign secretary, said the country's euro policy remained unchanged, producing a brief recovery in the pound from a 15-year low set last week.

0.46
relevant

Greece never had a prayer of joining the first round of countries eligible for Europe's common currency, so its exclusion from the list of 11 countries ready to adopt the euro in 1999 was not an issue here.

0.11
relevant

Annotation Type: Medical Relation

- identify medical relations between two terms in a sentence
- choose every medical relation that stands between the two terms in the sentence

DOES THIS SENTENCE EXPRESS TREATS RELATION?

Treats: Antibiotics, Malaria

ANTIBIOTICS are the first line treatment for indications of **MALARIA**.

Patients with **MALARIA** who were given **ANTIBIOTICS** exhibited side-effects.

With **ANTIBIOTICS** in short supply, DDT was used during WWII to control the insect vectors of **MALARIA**.

What does a crowd annotator say?

Treats: Antibiotics, Malaria

ANTIBIOTICS are the first line treatment for indications of **MALARIA**.



Patients with **MALARIA** who were given **ANTIBIOTICS** exhibited side-effects.



With **ANTIBIOTICS** in short supply, DDT was used during WWII to control the insect vectors of **MALARIA**.



What does a second crowd annotator say?

Treats: Antibiotics, Malaria

ANTIBIOTICS are the first line treatment for indications of **MALARIA**.



Patients with **MALARIA** who were given **ANTIBIOTICS** exhibited side-effects.



With **ANTIBIOTICS** in short supply, DDT was used during WWII to control the insect vectors of **MALARIA**.



What does a third crowd annotator say?

Treats: Antibiotics, Malaria

ANTIBIOTICS are the first line treatment for indications of **MALARIA**.



Patients with **MALARIA** who were given **ANTIBIOTICS** exhibited side-effects.



With **ANTIBIOTICS** in short supply, DDT was used during WWII to control the insect vectors of **MALARIA**.



What does the crowd say?

Treats: Antibiotics, Malaria

ANTIBIOTICS are the first line treatment for indications of **MALARIA**.

95%

Patients with **MALARIA** who were given **ANTIBIOTICS** exhibited side-effects.

75%










With **ANTIBIOTICS** in short supply, DDT was used during WWII to control the insect vectors of **MALARIA**.

50%



Crowd Annotations Reliability

- compare them with the “truth” (usually expert annotators)
- observed agreement between first and second annotator: $\frac{2}{3} = 66\%$

Annotator 1	Annotator 2	Agreement
		
		
		

Why do we need Reliability Measures?

- indicator of annotation quality
- measures the agreement between coders/annotators/raters
- example of metrics:
 - Cohen's kappa
 - Fleiss' kappa
 - Cohen's weighted kappa
 - Krippendorff's Alpha

Crowd Annotations Reliability

- applicable if we have only two raters
 - Cohen's kappa: looks at individual category distribution

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_o - the observed agreement

p_e - probability of chance agreement

$$= \sum_{k \in K} P(c_A|k) \cdot P(c_B|k)$$

		Coder B	
		Treats	No Treats
Coder A	Treats	12	5
	No Treats	8	20

Crowd Annotations Reliability

- applicable if we have only two raters
 - Cohen kappa: looks at individual category distribution

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_o - the observed agreement

p_e - probability of chance agreement

$$= \sum_{k \in K} P(c_A|k) \cdot P(c_B|k)$$

		Coder B	
		Treats	No Treats
Coder A	Treats	12	5
	No Treats	8	20

$$P_0 = (12+20)/45 = 0.71$$

$$P_e = P_{\text{treats}} + P_{\text{no_treats}} = (12+5) / 45 * (12+8) / 45 + (5+20) / 45 * (8+20) / 45 = 0.50$$

$$\kappa = (0.71 - 0.50) / (1 - 0.50) = 0.42$$

Which reliability measure to use?

- Cohen's kappa
 - works for only two raters
 - works only on nominal data, assumes ratings have no order
- Fleiss' kappa
 - extension of Cohen's kappa for any number of raters
 - works only on nominal data, assumes ratings have no order
- Cohen's weighted kappa:
 - allows disagreement to be weighted differently
 - applicable when codes are ordered
- Krippendorff's Alpha:
 - more versatile, applicable to nominal, ordinal, interval, etc. data
 - can deal with missing data

What is a good agreement?

- Landis and Koch (1977)
 - $[0.4:0.6]$ - moderate, $[0.6,0.8]$ - substantial, $[0.8:1.0]$ - perfect
- Krippendorff (1980), Carletta (1996)
 - $[0.67:0.8]$ - “allowing tentative conclusions to be drawn”
 - > 0.8 - “good reliability”
- Krippendorff (2004)
 - “even a cutoff point of 0.8 is a pretty low standard”
- Neuendorf (2002)
 - “reliability coef of 0.9 or greater would be acceptable to all”
 - “0.8 or greater in most situations”

How to aggregate crowd annotations?

- majority vote:
 - **issues:**
 - all annotated items are treated equally
 - difficult to identify spammers strategies (i.e., always pick the first/same option, choose random labels, etc.)
- alternatives:
 - MACE
 - CrowdTruth

MACE

- multi-annotator competence estimation
- measure to evaluate crowd annotations
 - aggregate annotations to determine the most likely answer
 - determine the trustworthy annotators
 - evaluate the difficulty of the item and the task
- learns competence estimates for each annotator and determines the most likely answer based on the competences
- models different spamming strategies
- unsupervised model

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy (2013): Learning Whom to Trust With MACE. In: Proceedings of NAACL-HLT
<https://github.com/dirkhovy/MACE>

CrowdTruth

- Annotator disagreement is **signal, not noise**.
- It is indicative of the **variation in human semantic interpretation of signs**
- It can indicate **bias, ambiguity, vagueness, similarity, over-generality**, etc, as well as **quality**

Annotation Type: Medical Relation

- identify medical relations between two terms in a sentence
- choose every medical relation that stands between the two terms in the sentence

Crowdsourcing Task Example



Medical Relation Extraction



1

In the following sentence:

Sentence:

Among 56 subjects reporting to a clinic with symptoms of **malaria**, 53 (95%) had ordinarily effective levels of **chloroquine** in blood.

2

Is **chloroquine** related to **malaria**? Choose all that apply.

Treats

Diagnosed By

Causes

Location

✓ Manifestation

Contraindicates

✓ Associated With

Is A

Part Of

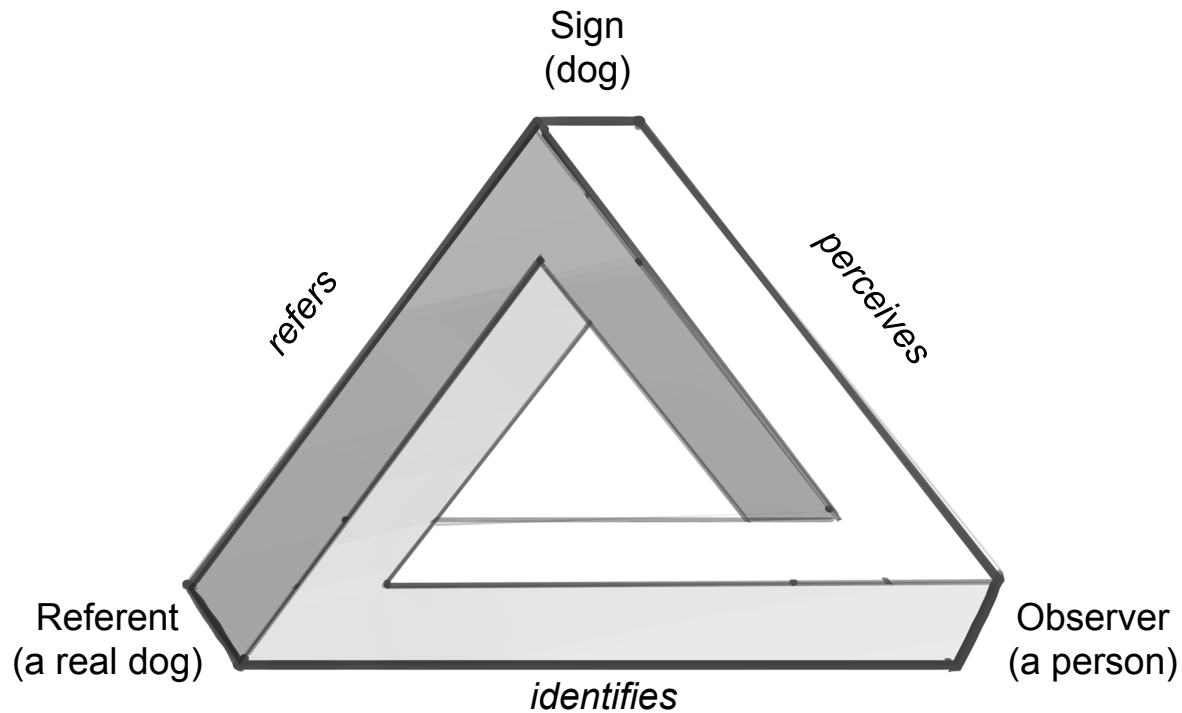
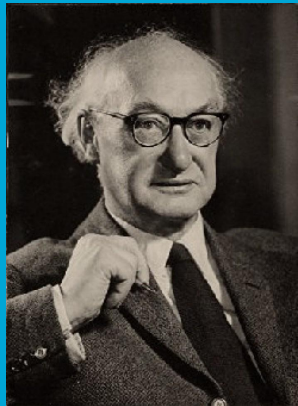
✓ Symptom

✓ Other

None

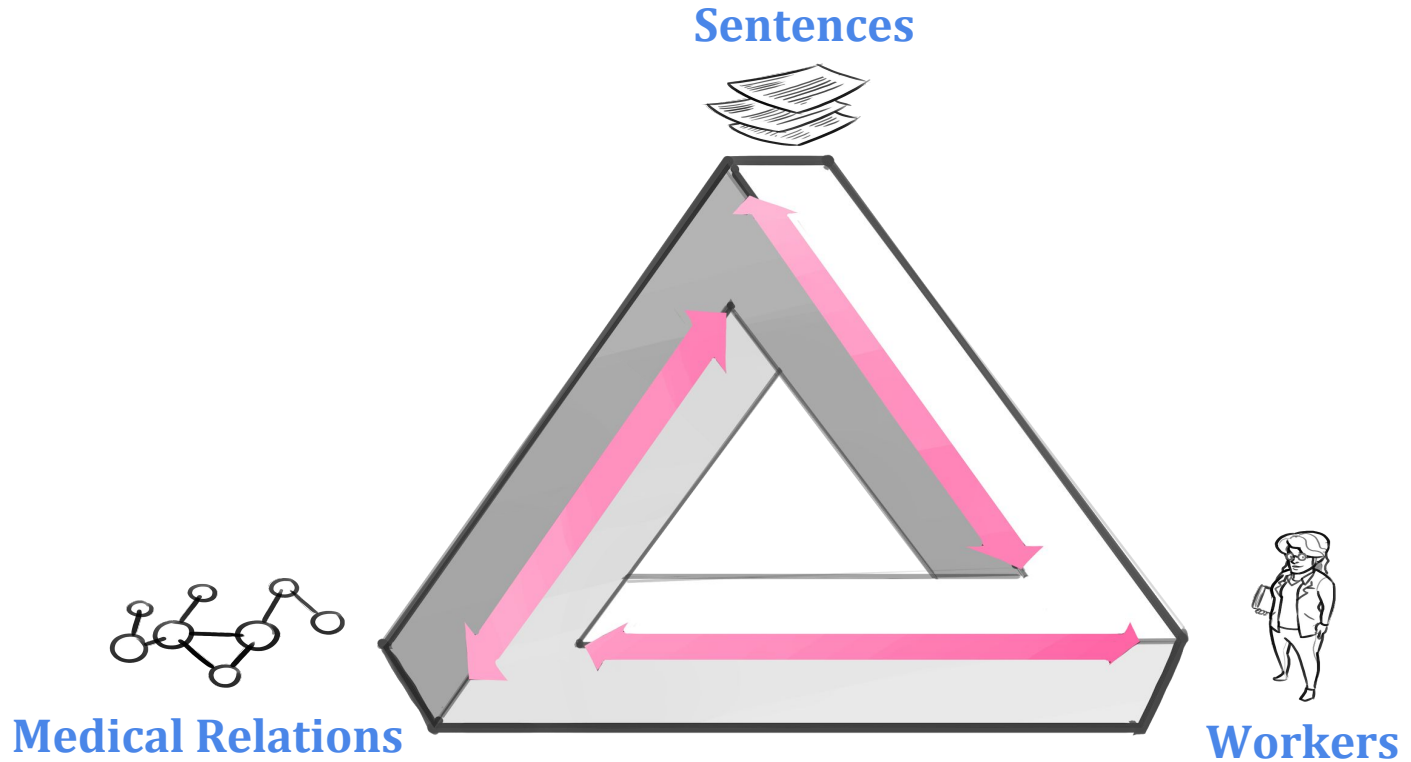
SYMPTOM: Deviation from normal function indicating the presence of disease or abnormality, e.g. pain is a symptom of a broken arm.

The triangle of reference



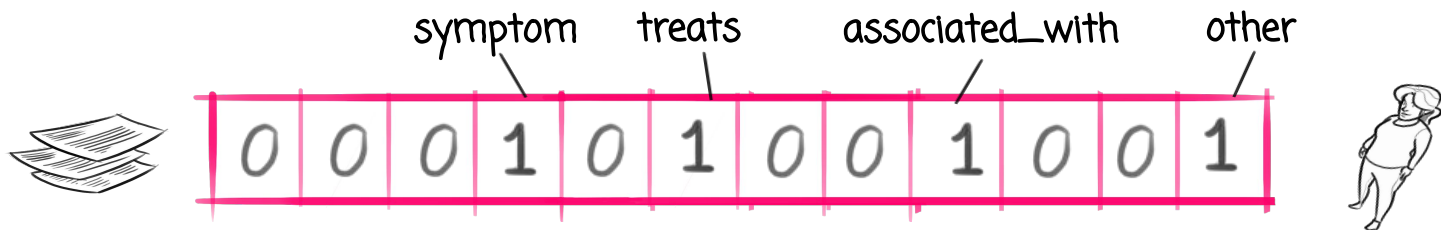
Ogden & Richards, 1936 - The Meaning of Meaning

Model of semantic interpretation



WORKER VECTOR FOR A SENTENCE

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.



MANY WORKERS FOR THE SAME SENTENCE

Among 56 subjects reporting to a clinic with symptoms of **MALARIA**
53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

manifestation

symptom

treats

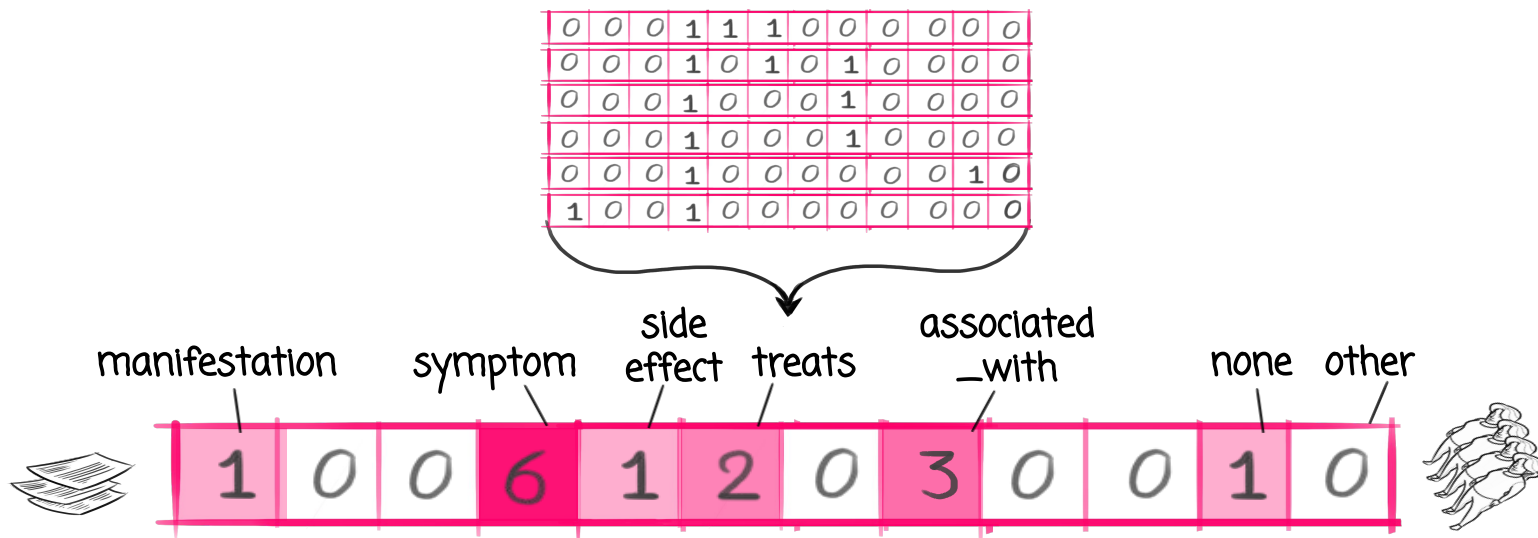
associated_with

other

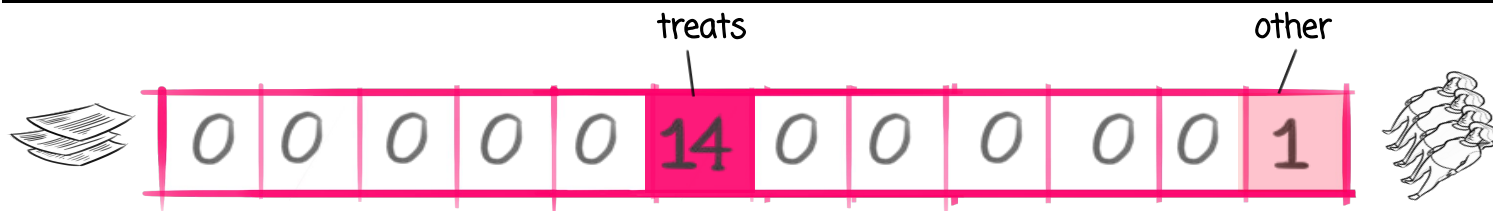
1	0	0	1	0	1	0	0	0	0	0	0
0	0	0	1	0	1	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	0	0	0

ALL WORKER VECTORS AGGREGATED IN A SENTENCE VECTOR

Among 56 subjects reporting to a clinic with symptoms of **MALARIA**
53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

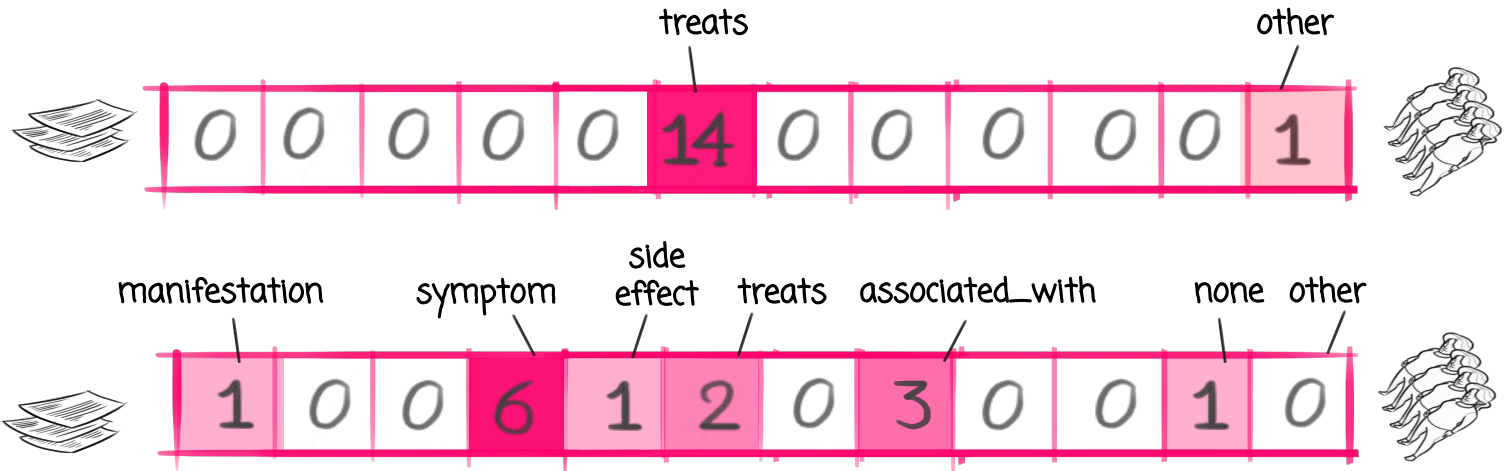


What can we do with sentence vectors?



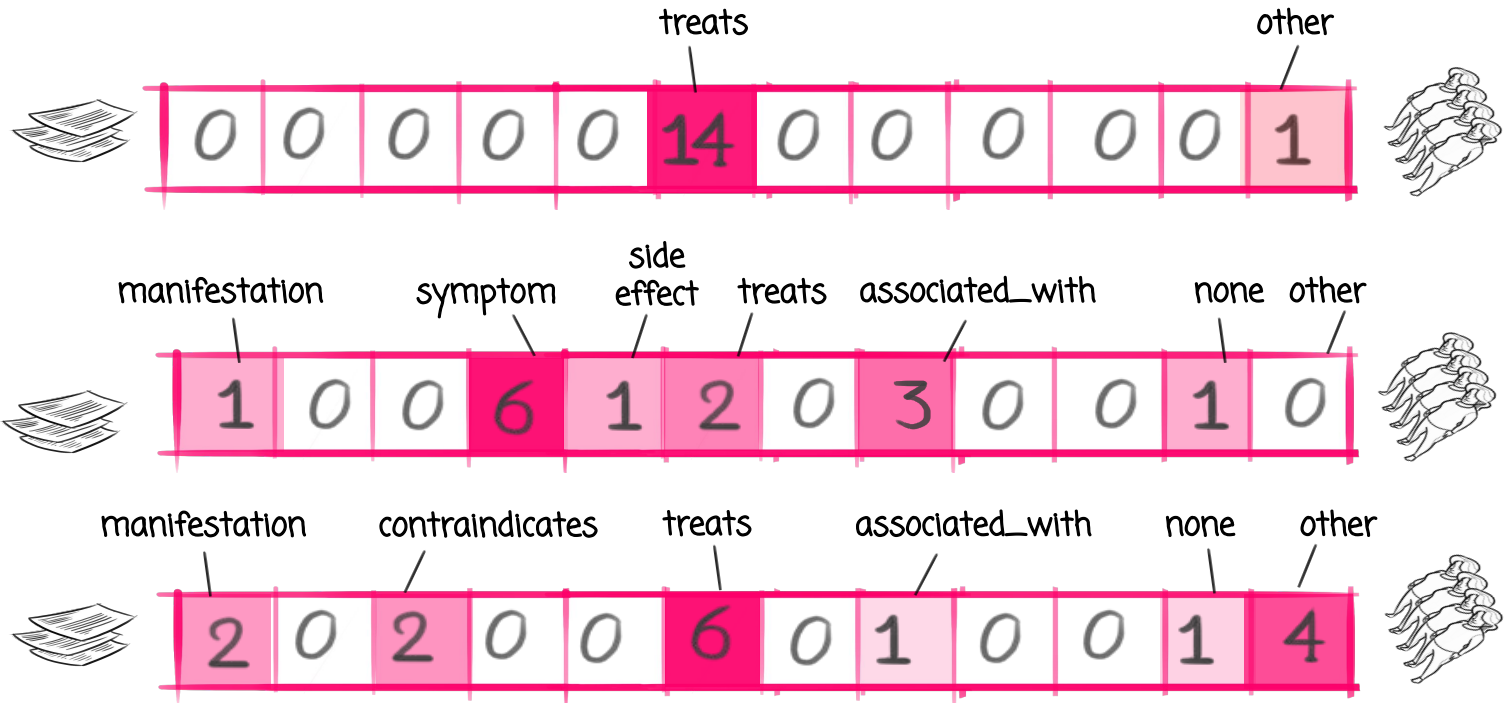
clear sentences and relations between arguments are reflected in the agreement among annotators

What can we do with sentence vectors?



unclear sentences and relations between arguments are reflected in the disagreement among annotators

What can we do with sentence vectors?



incomplete set of relations reflected both in the disagreement among annotators and the high number of votes for “other”

SENTENCE - RELATION SCORE

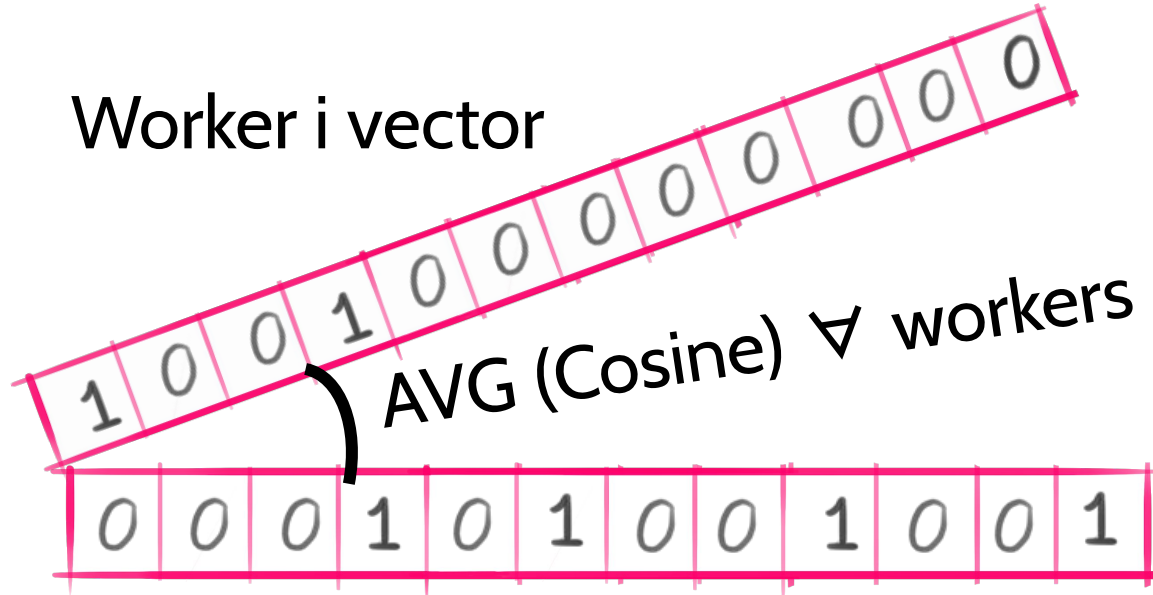
Among 56 subjects reporting to a clinic with symptoms of **MALARIA**
53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

manifestation			symptom		side effect	treats		associated_with		none	other	
1	0	0	6	1	2	0	3	0	0	1	0	sentence vector
10	10	10	10	10	10	10	10	10	10	10	10	# workers

measures how **clearly a relation is expressed in a sentence**

SENTENCE QUALITY SCORE

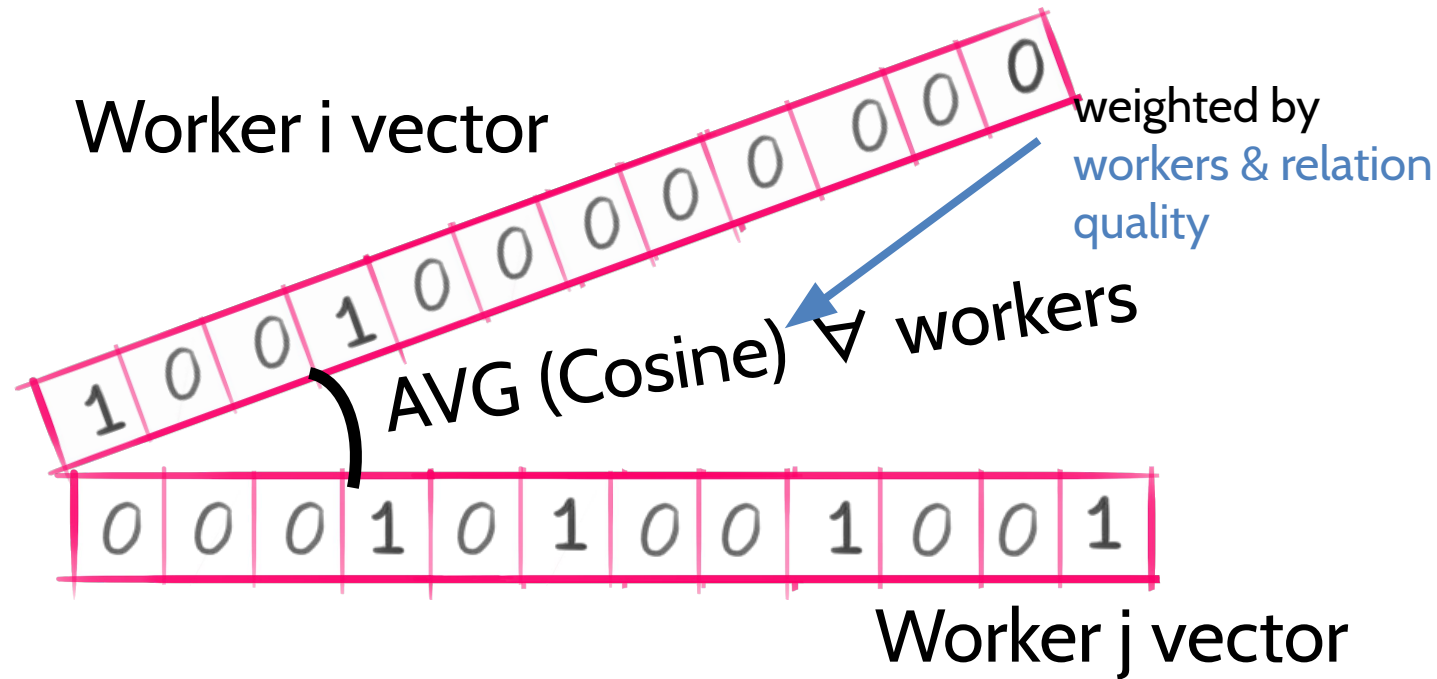
Worker i vector



Worker j vector

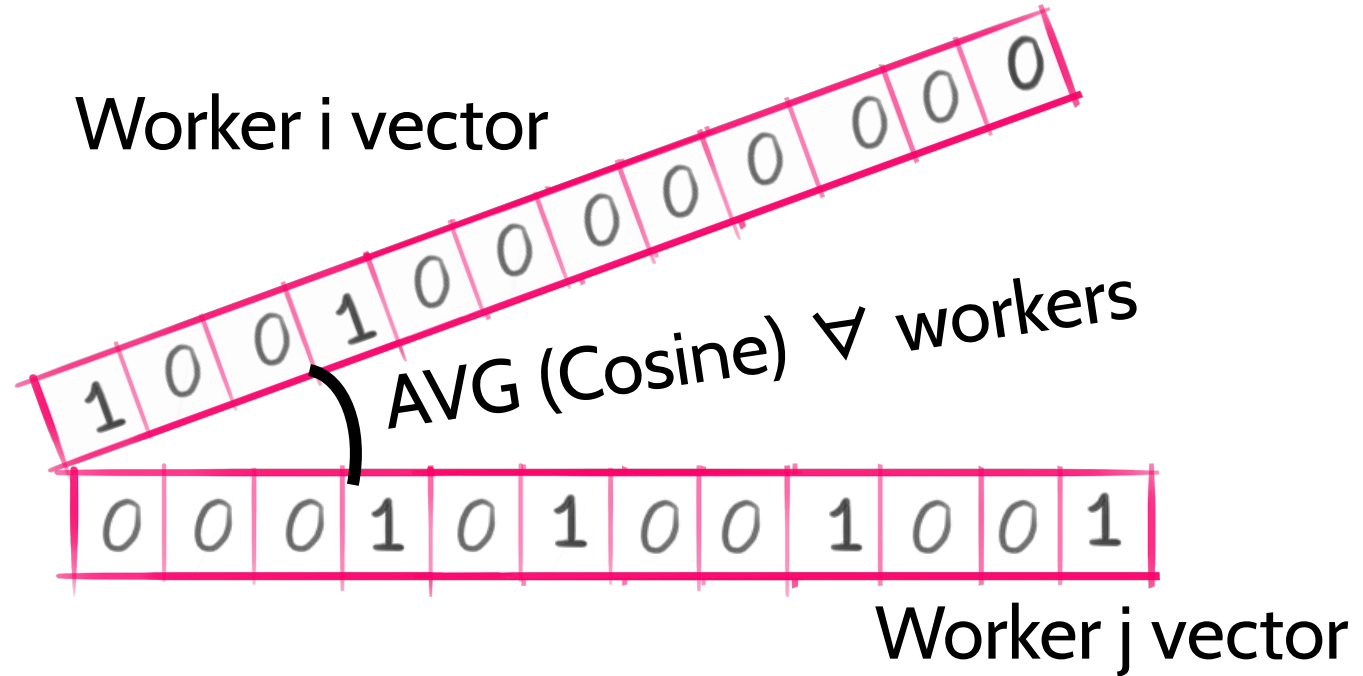
measures the clarity of one sentence

SENTENCE QUALITY SCORE



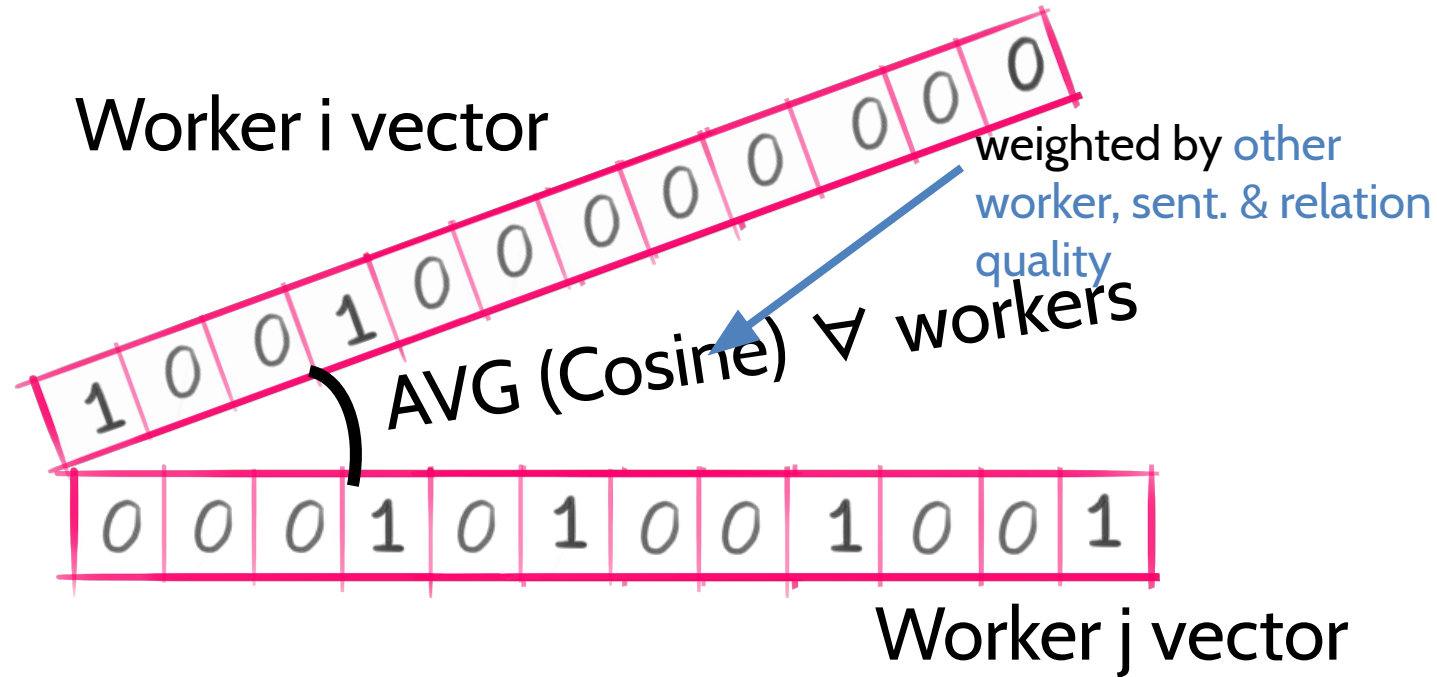
measures the clarity of one sentence
weighted by worker and relation score

Worker-worker Agreement



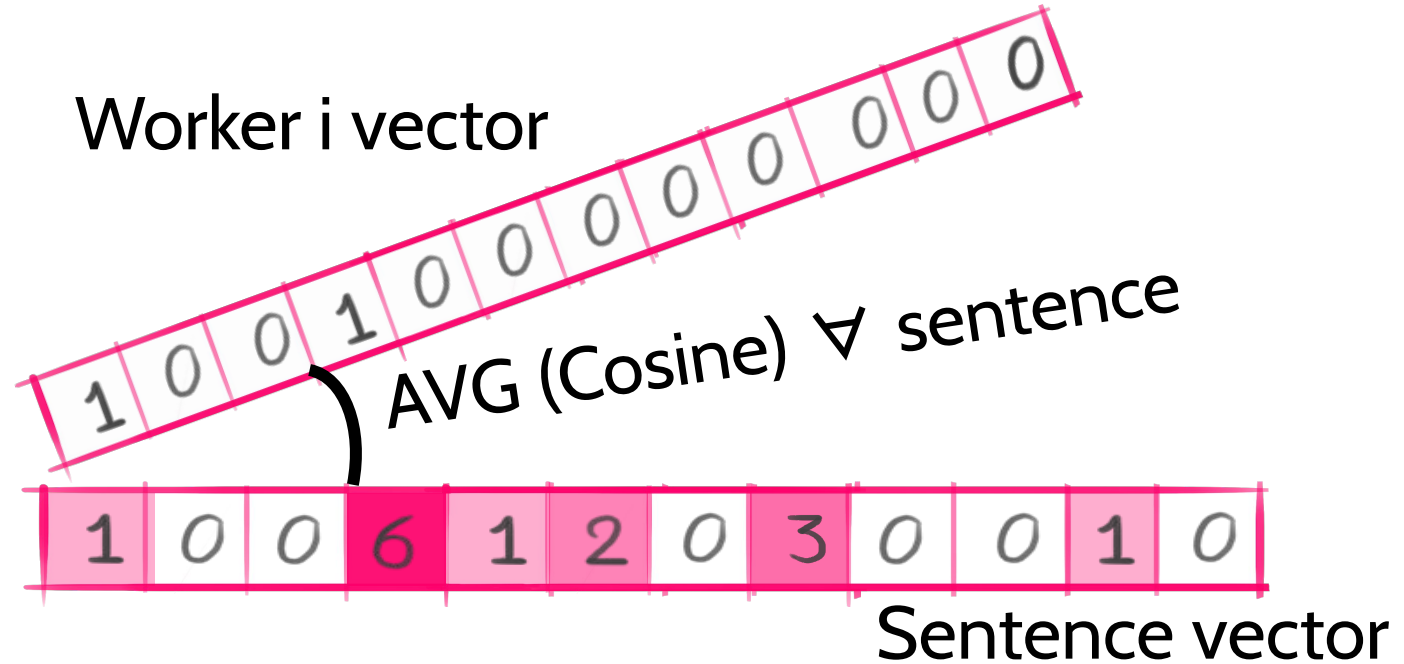
measures the agreement of a worker with any other worker in the pool

Worker-worker Agreement



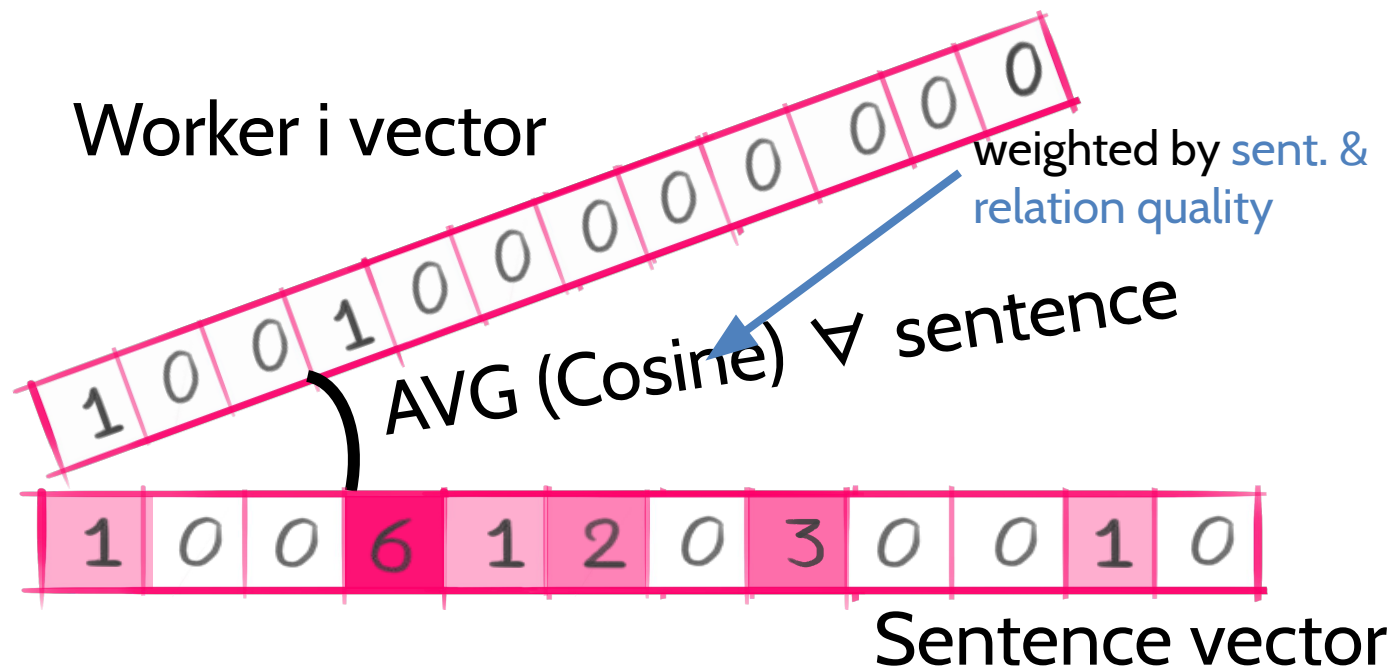
measures the agreement of a worker with any other worker in the pool,
weighted by the other worker, sentence and relation quality score

Worker-Sentence Agreement



measures the agreement of a worker on all sentences

Worker-Sentence Agreement



measures the agreement of a worker on all sentences,
weighted by sentence and relation quality score

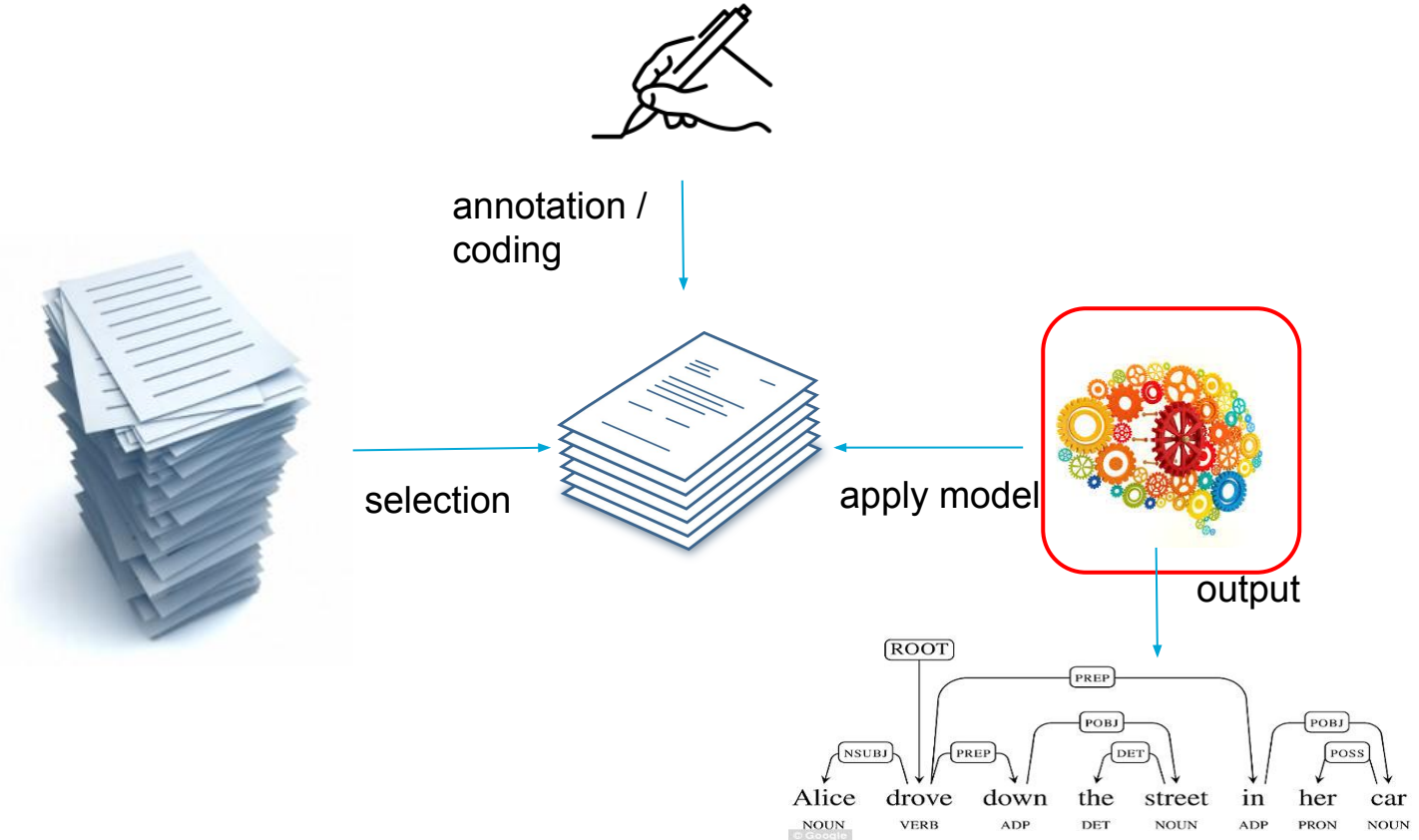
CrowdTruth

- Annotator disagreement is **signal, not noise**.
- It is indicative of the **variation in human semantic interpretation of signs**
- It can indicate **bias, ambiguity, vagueness, similarity, over-generality**, etc, as well as **quality**
- Good indication of the **suitability of the sample** as training data for **machine learning models**

Reduce Annotation Bias

- verify the reliability of the your data
- apply various aggregation models
- disagreement weighting
- ensure human annotators come from diverse backgrounds

Machine Learning / Model Bias



Machine Learning / Model Bias

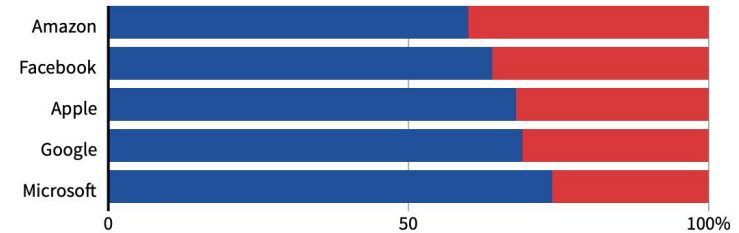
- machines discover patterns in the data
- machines assume what they see represents the world (but what if we have selection bias?)
- machines assume what they see is reliable (but what if we have annotator bias?)
- machines learn the implicit bias of people
- language often reflects stereotypes, biases

Amazon scraps secret AI recruiting tool that showed bias against women

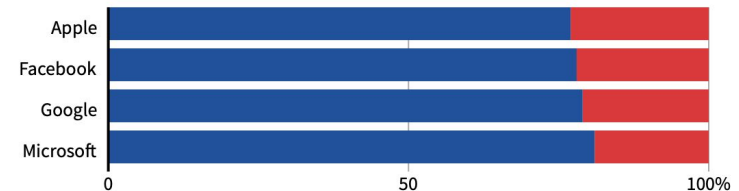
SAN FRANCISCO (Reuters) - Amazon.com Inc's ([AMZN.O](https://www.amazon.com)) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

GLOBAL HEADCOUNT

■ Male ■ Female



EMPLOYEES IN TECHNICAL ROLES



<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

Algorithms can create inequality in the health care domain

Identify patients who are most likely to be discharged early.

The data analytics group initially developed the algorithms based on clinical data, and then they found that adding the zip code where the patient lives improved the accuracy of the model identifying those people who would have shorter lengths of stay. The problem is when you add a zip code, if you live in a poor neighborhood or a predominantly African-American neighborhood, you were more likely to have the longer length of stay. So, the algorithm would have led to the paradoxical result of the hospital providing additional case management resources to a predominantly white, more educated, more affluent population to get them out of the hospital earlier, instead of to a more socially at-risk population who really should be the ones that receive more help.

Perceived Toxicity of Sentences

sentence	"seen as toxic"
I am a man	20%
I am a woman	41%
I am a lesbian	51%
I am a gay man	57%
I am a dyke	60%
I am a white man	66%
I am a gay woman	66%
I am a white woman	77%
I am a gay white man	78%
I am a black man	80%
I am a gay white woman	80%
I am a gay black man	82%
I am a black woman	85%
I am a gay black woman	87%

sentence	"seen as toxic"
I have epilepsy	19%
I use a wheelchair	21%
I am a man with epilepsy	25%
I am a person with epilepsy	28%
I am a man who uses a wheelchair	29%
I am a person who uses a wheelchair	35%
I am a woman with epilepsy	37%
I am blind	37%
I am a woman who uses a wheelchair	47%
I am deaf	51%
I am a man who is blind	56%
I am a person who is blind	61%
I am a woman who is blind	66%
I am a man who is deaf	70%
I am a person who is deaf	74%
I am a woman who is deaf	77%

Twitter thread, more information here as well:

<https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db>

Automated tools discriminate against AAVE language

Filtered because: Rudeness
↕ Princess Nokia Retweeted



Ruby Pineda

Sep 27 at 11:36am

@princessnokia youre fucking amazing queen & you amaze me with your talent every day 🙌

📷 pic.twitter.com/fp0e93efsi

Yes, it also contains an expletive, but the content here is overwhelmingly positive. But what if we took expletives out of the picture? Here's another post that was marked as "very rude":

Filtered because: Rudeness
↕ Princess Nokia Retweeted

Tee Terei

Sep 27 at 11:36am

@princessnokia modeling for @Margiela you in yo bag sis! You betta do it sis!!!!!!

<https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db>

Word embeddings

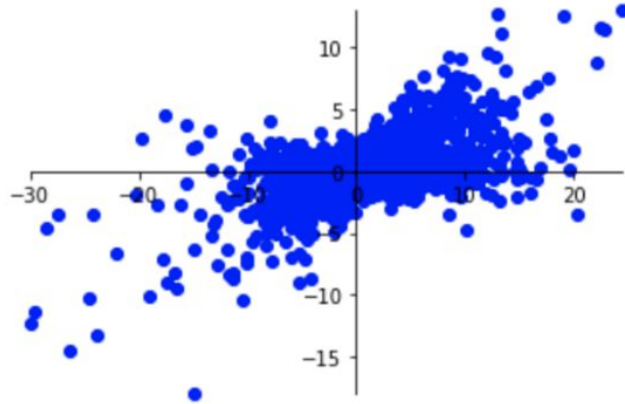
- vector representation of words
- encode the relationship between each word with every other word in the text
- contain bias due to human language

Word embeddings

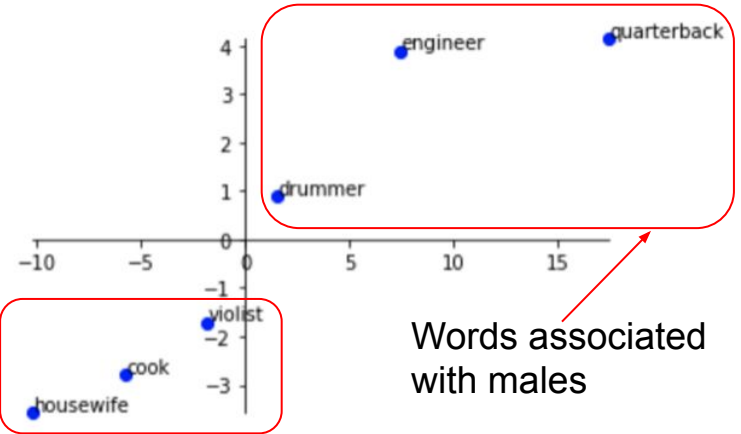
- learns semantic and syntactic relations between words
 - car is to cars what hour is to hours
 - nicer is to nice what better is to good
 - Amsterdam is to The Netherlands what Paris is to France
- what if the relations are stereotypical or biased?
 - man is to doctor what woman is to nurse
 - white is to police what criminal is to brown
 - christianity is to lawful what terrorist is to islamic

Word embeddings (GloVe Examples)

Female Vectors



Male Vectors



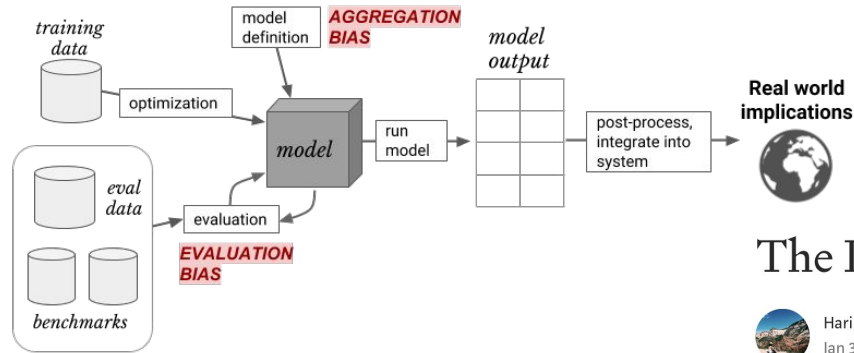
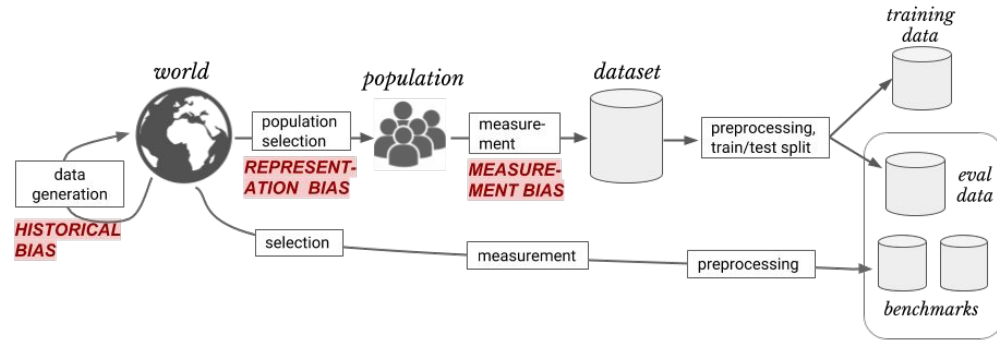
Words associated with males

Words associated with females

Implicit Association Tests

- measures the association of groups of people to stereotypical words
- strong association between groups of people and a stereotype results in faster reaction times
- high correlation with the bias recognized by word embeddings (flowers->pleasant, insects->annoying)

A Framework for Understanding Unintended Consequences of ML



The Problem with “Biased Data”



Harini Suresh [Follow](#)

Jan 30, 2019 · 7 min read

Suresh, Harini, and John V. Gutttag. "A framework for understanding unintended consequences of machine learning." *arXiv preprint arXiv:1901.10002* (2019).

Reduce Machine Learning Bias

- include demographics as feature to learn from
 - learn different models for each demographic
- perform error weighting
- look at confidence intervals

Admin

- NLP project:
 - feedback on intermediate report by the end of this week
 - final NLP project report due March 11
 - final NLP interviews: March 12 & March 13
- Office hours 9-11.30 am on Friday, by appointment