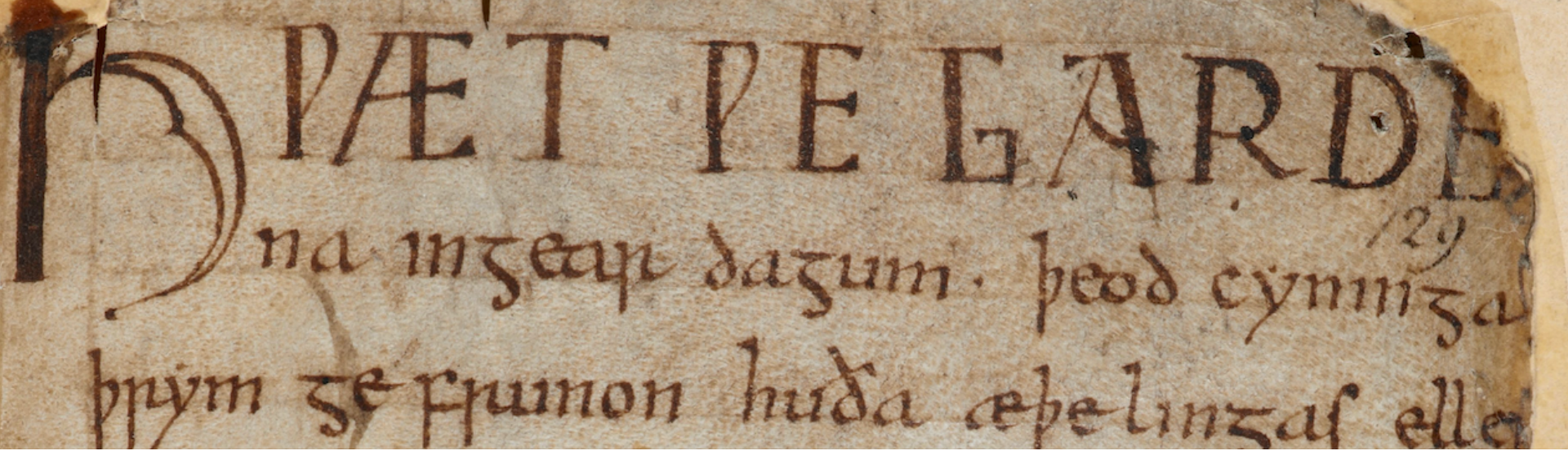


Natural Language Processing

Info 159/259

Lecture 2: Truth and ethics (Jan 21, 2021)

David Bamman, UC Berkeley

Hwæt! Wé Gárde
na in géardagum, þéodcyninga
þrym gefrúnon, hú ðá æþelingas ellen

Natural Language Processing

Info 159/259
Lecture 5: Truth and ethics (Jan 21, 2021)

David Bamman, UC Berkeley

In-class questions:
http://bit.ly/nlpqs

Modern NLP is driven by annotated data

- Penn Treebank (1993; 1995;1999); morphosyntactic annotations of WSJ
- OntoNotes (2007–2013); syntax, predicate-argument structure, word sense, coreference
- FrameNet (1998–): frame-semantic lexica/annotations
- MPQA (2005): opinion/sentiment
- SQuAD (2016): annotated questions + spans of answers in Wikipedia

Modern NLP is driven by annotated data

- In most cases, the data we have is the product of human judgments.
    - What's the correct part of speech tag?
    - Syntactic structure?
    - Sentiment?

Ambiguity

“One morning I shot
an elephant in my pajamas”

Animal Crackers

Dogmatism

Fast and Horvitz (2016), "Identifying Dogmatism in Social Media: Signals and Models"

Given a comment, imagine you hold a well-informed, different opinion from the commenter in question. We'd like you to tell us how likely that commenter would be to engage you in a constructive conversation about your disagreement, where you each are able to explore the other's beliefs. The options are:

(5): It's unlikely you'll be able to engage in any substantive conversation. When you respectfully express your disagreement, they are likely to ignore you or insult you or otherwise lower the level of discourse.

(4): They are deeply rooted in their opinion, but you are able to exchange your views without the conversation degenerating too much.

(3): It's not likely you'll be able to change their mind, but you're easily able to talk and understand each other's point of view.

(2): They may have a clear opinion about the subject, but would likely be open to discussing alternative viewpoints.

(1): They are not set in their opinion, and it's possible you might change their mind. If the comment does not convey an opinion of any kind, you may also select this option.

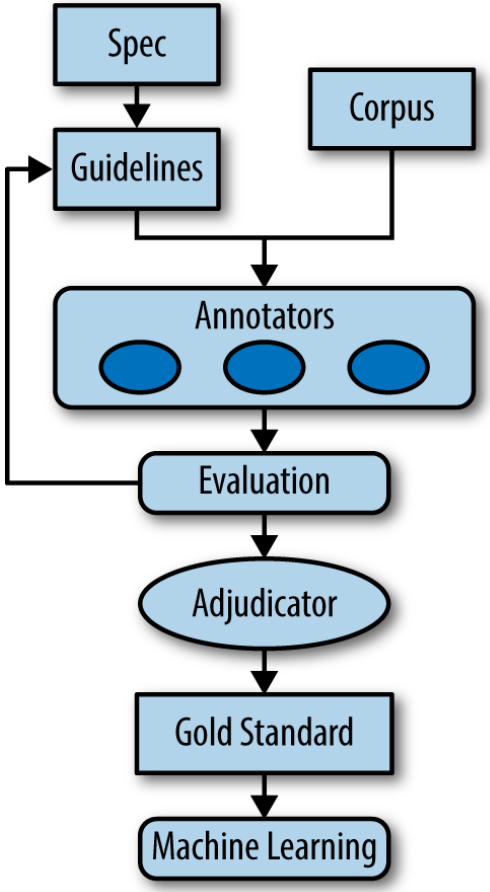
Sarcasm

“In many respects you know they honor President Obama. ISIS is honoring President Obama! He is the founder of ISIS. He’s the founder of ISIS, O.K.! He’s the founder, he founded ISIS and I would say the co-founder would be crooked Hillary Clinton. Co-founder, crooked Hillary Clinton. And that’s what it’s about.”

Fake News

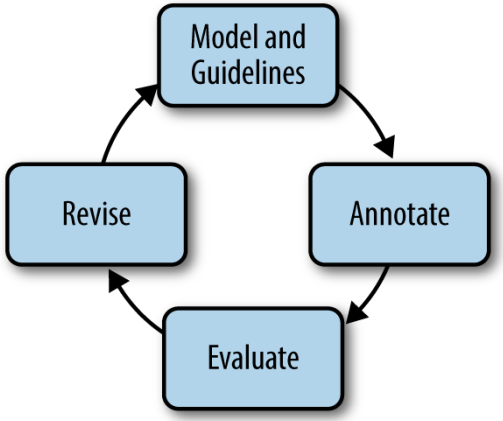
http://www.fakenewschallenge.org

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation pipeline



Pustejovsky and Stubbs (2012),
Natural Language Annotation for Machine Learning

Annotation guidelines

- Our goal: given the constraints of our problem, how can we formalize our description of the annotation process to encourage multiple annotators to provide the same judgment?

Annotation guidelines

- What is the goal of the project?
- What is each tag called and how is it used? (Be specific: provide examples, and discuss gray areas.)
- What parts of the text do you want annotated, and what should be left alone?
- How will the annotation be created? (For example, explain which tags or documents to annotate first, how to use the annotation tools, etc.)

Pustejovsky and Stubbs (2012), Natural Language Annotation for Machine Learning

Why not do it yourself?

- Expensive/time-consuming
- Multiple people provide a measure of consistency: is the task well enough defined?
- Low agreement = not enough training, guidelines not well enough defined, task is bad

Adjudication

- Adjudication is the process of deciding on a single annotation for a piece of text, using information about the independent annotations.
- Can be as time-consuming (or more so) as a primary annotation.
- Does not need to be identical with a primary annotation (both annotators can be wrong by chance)

Interannotator agreement

annotator A

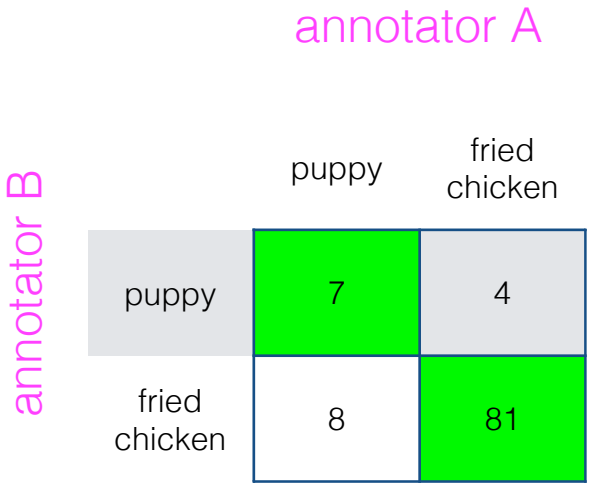
annotator B

	puppy	fried chicken
puppy	6	3
fried chicken	2	5

observed agreement = 11/16 = 68.75%

Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

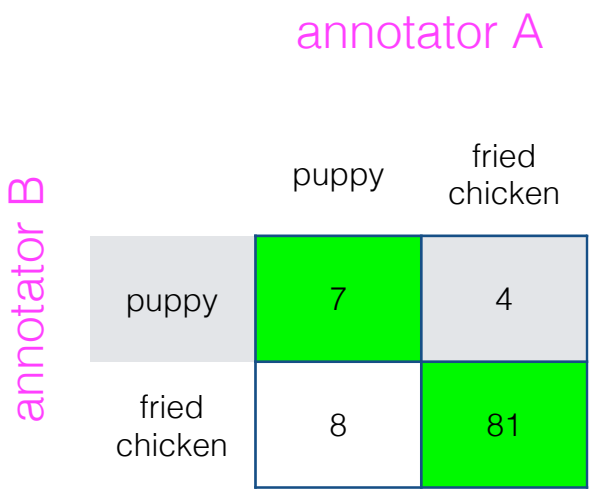


Cohen's kappa

- If classes are imbalanced, we can get high inter annotator agreement simply by chance

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

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$



Cohen's kappa

- Expected probability of agreement is how often we would expect two annotators to agree assuming independent annotations

$$p_e = P(A = \text{puppy}, B = \text{puppy}) + P(A = \text{chicken}, B = \text{chicken})$$

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

Cohen's kappa

$$= P(A = \text{puppy})P(B = \text{puppy}) + P(A = \text{chicken})P(B = \text{chicken})$$

P(A=puppy)	15/100 = 0.15
P(B=puppy)	11/100 = 0.11
P(A=chicken)	85/100 = 0.85
P(B=chicken)	89/100 = 0.89

$$= 0.15 \times 0.11 + 0.85 \times 0.89$$
$$= 0.773$$

annotator A

annotator B

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

Cohen's kappa

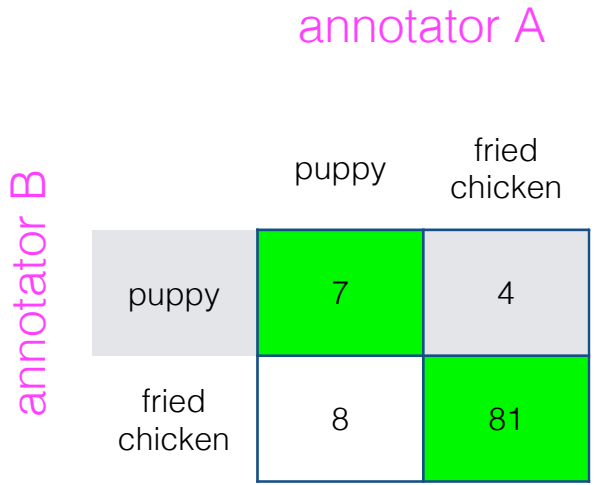
- If classes are imbalanced, we can get high inter annotator agreement simply by chance

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

$$\kappa = \frac{0.88 - p_e}{1 - p_e}$$

$$\kappa = \frac{0.88 - 0.773}{1 - 0.773}$$

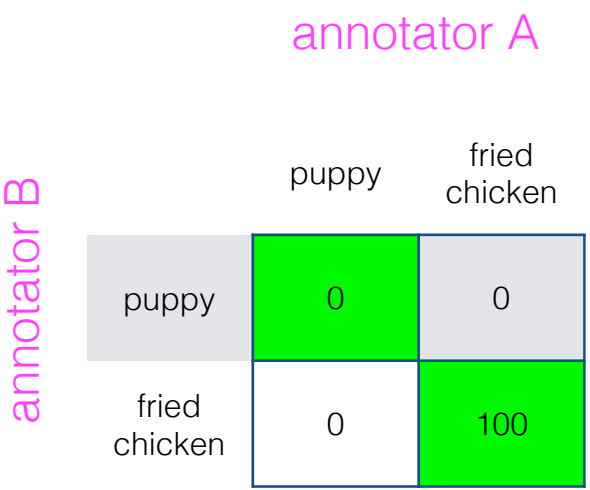
$$= 0.471$$

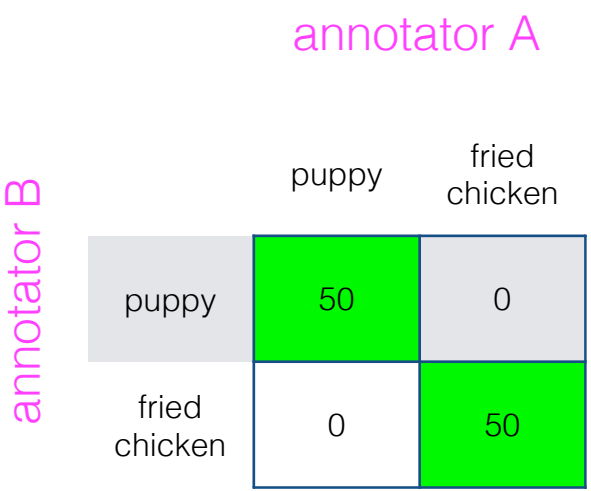


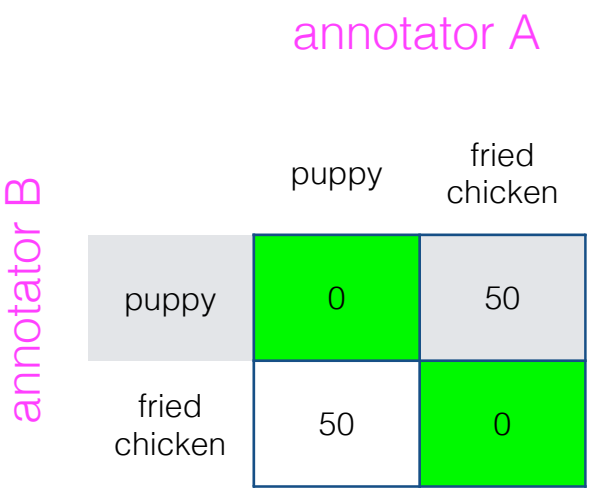
Cohen’s kappa

- “Good” values are subject to interpretation, but rule of thumb:

0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement


annotator A
annotator B
puppy
fried chicken
puppy
0
0
fried chicken
0
100


annotator A
annotator B
puppy
fried chicken
puppy
fried chicken
50
0
0
50


annotator A
annotator B
puppy
fried chicken
puppy
0
50
fried chicken
50
0

Interannotator agreement

- Cohen's kappa can be used for any number of classes.
- Still requires two annotators who evaluate the same items.
- Fleiss' kappa generalizes to multiple annotators, each of whom may evaluate different items (e.g., crowdsourcing)

Fleiss' kappa

- Same fundamental idea of measuring the observed agreement compared to the agreement we would expect by chance.

- With N > 2, we calculate agreement among pairs of annotators

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Fleiss' kappa

Number of annotators who assign category *j* to item *i*

$$n_{ij}$$

For item *i* with *n* annotations, how many annotators agree, among all n(n-1) possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

Fleiss' kappa

For item *i* with *n* annotations, how many annotators agree, among all n(n-1) possible pairs

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

Annotator

A	B	C	D
+	+	+	-

agreeing pairs of annotators →

A-B
B-A
A-C
C-A
B-C
C-B

Label	n_{ij}
+	3
-	1

$$P_i = \frac{1}{4(3)}(3(2) + 1(0))$$

Fleiss' kappa

Average agreement among all items

$$P_o = \frac{1}{N}\sum_{i=1}^{N} P_i$$

Probability of category *j*

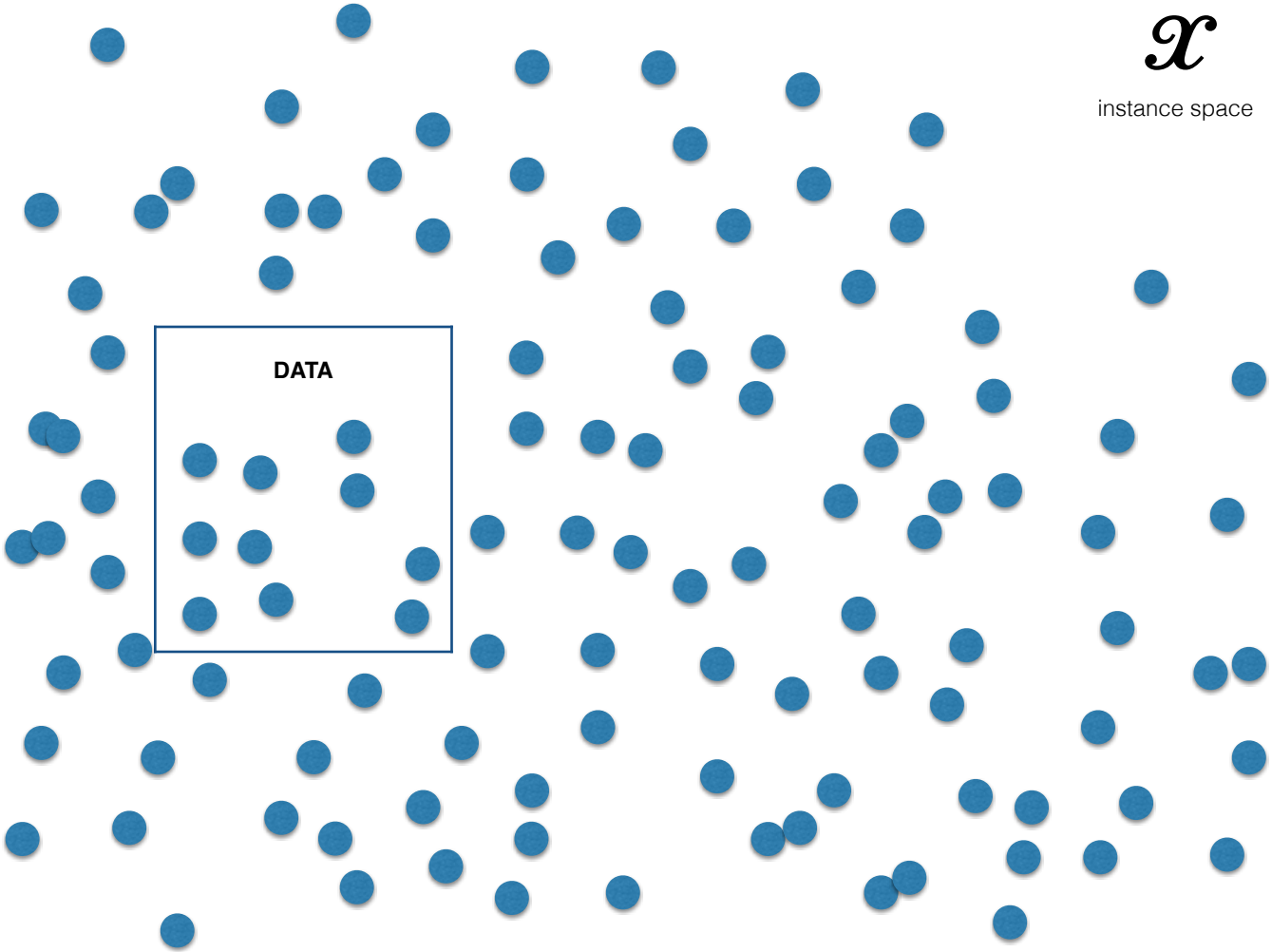
$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$

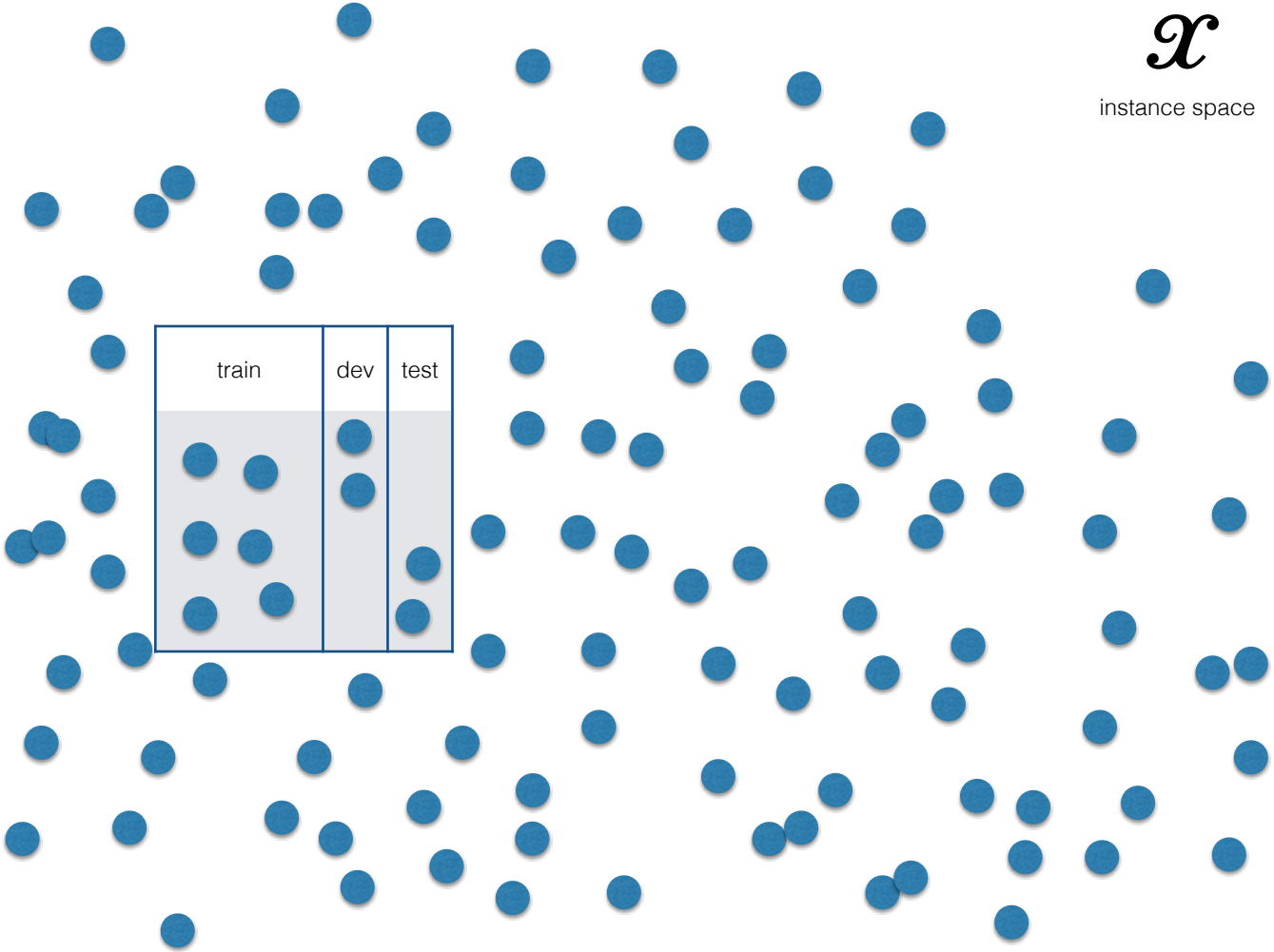
Expected agreement by chance — joint probability two raters pick the same label is the product of their independent probabilities of picking that label

$$P_e = \sum_{j=1}^{K} p_j^2$$

Evaluation

- A critical part of development new algorithms and methods and demonstrating that they work


$\mathcal{X}$
instance space
DATA


$\mathcal{X}$
instance space
train
dev
test

Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Metrics

- Evaluations presuppose that you have some metric to evaluate the fitness of a model.
    - Language model: perplexity
    - POS tagging/NER: accuracy, precision, recall, F1
    - Phrase-structure parsing: PARSEVAL (bracketing overlap)
    - Dependency parsing: Labeled/unlabeled attachment score
    - Machine translation: BLEU, METEOR
    - Summarization: ROUGE

Uncertainty

Your work	58%
Current state of the art	50%

- If we observe difference in performance, what's the cause? Is it because one system is better than another, or is it a function of randomness in the data? If we had tested it on other data, would we get the same result?

Uncertainty

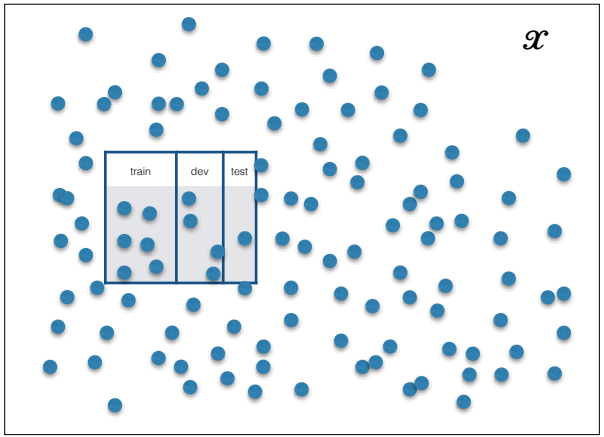
- When evaluating the performance of a system on a dataset, we need to reason about the uncertainty in our measurement: if we calculated accuracy on a different dataset (from the same distribution), how much variability would there be?

Berg-Kirkpatrick et al. (2012), “An Empirical Investigation of Statistical Significance in NLP”; Dror et al. (2018), "The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing”; Card et al. (2020), "With Little Power Comes Great Responsibility”

Bootstrap

- Core idea: the data we happen to have is a sample from all data that could exist; let's sample from our sample to estimate the variability.

- Our estimate of the point value of the metric itself won't change, but we can infer something about the variability of the population from the variable in the resamples.



Bootstrap

- Start with test data x of size n
- Draw b bootstrap samples x(i) of size n by sampling with replacement from x
- For each x(i)
    - Let m(i) = the metric of interest calculated from x(i)

				accuracy
I love this movie	I hate this movie	I don't love this movie	Not the worst ever!	0.50

				m(i)
I love this movie	I don't love this movie	I don't love this movie	Not the worst ever!	0.25
I love this movie	I love this movie	I hate this movie	Not the worst ever!	0.75
I hate this movie	I don't love this movie	I don't love this movie	I love this movie	0.50
I love this movie	I hate this movie	I don't love this movie	I hate this movie	0.75
I don't love this movie	I don't love this movie	I don't love this movie	Not the worst ever!	0.00

Bootstrap percentile interval

- At the end of the process, you end up with a vector of values m = [m(1), …, m(b)] (for b bootstrap samples) — e.g. [0.25, 0.75, 0.50, 0.75, 0] for the example before.
- We can define a 95% confidence interval as the middle 95% of m
- e.g., α = 0.05 (95% confidence intervals) = [2.5, 97.5] percentile
- Accurate for larger sample sizes

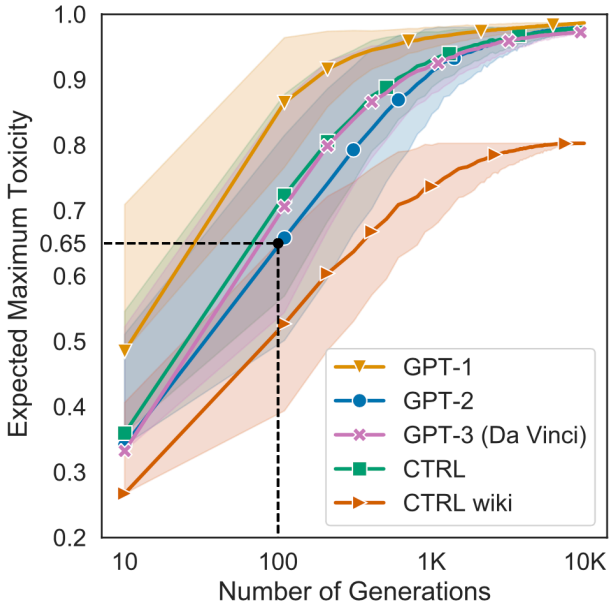
Ethics

Why does a discussion about ethics need to be a part of NLP?

Conversational Agents

Toxic generation

- Language models like GPT-{1,2,3} trained on toxic data (e.g., banned subreddits like /r/The_Donald or /r/WhiteRights) reproduce that toxicity in both prompted and unprompted generations



Sap et al. (2020), "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models"

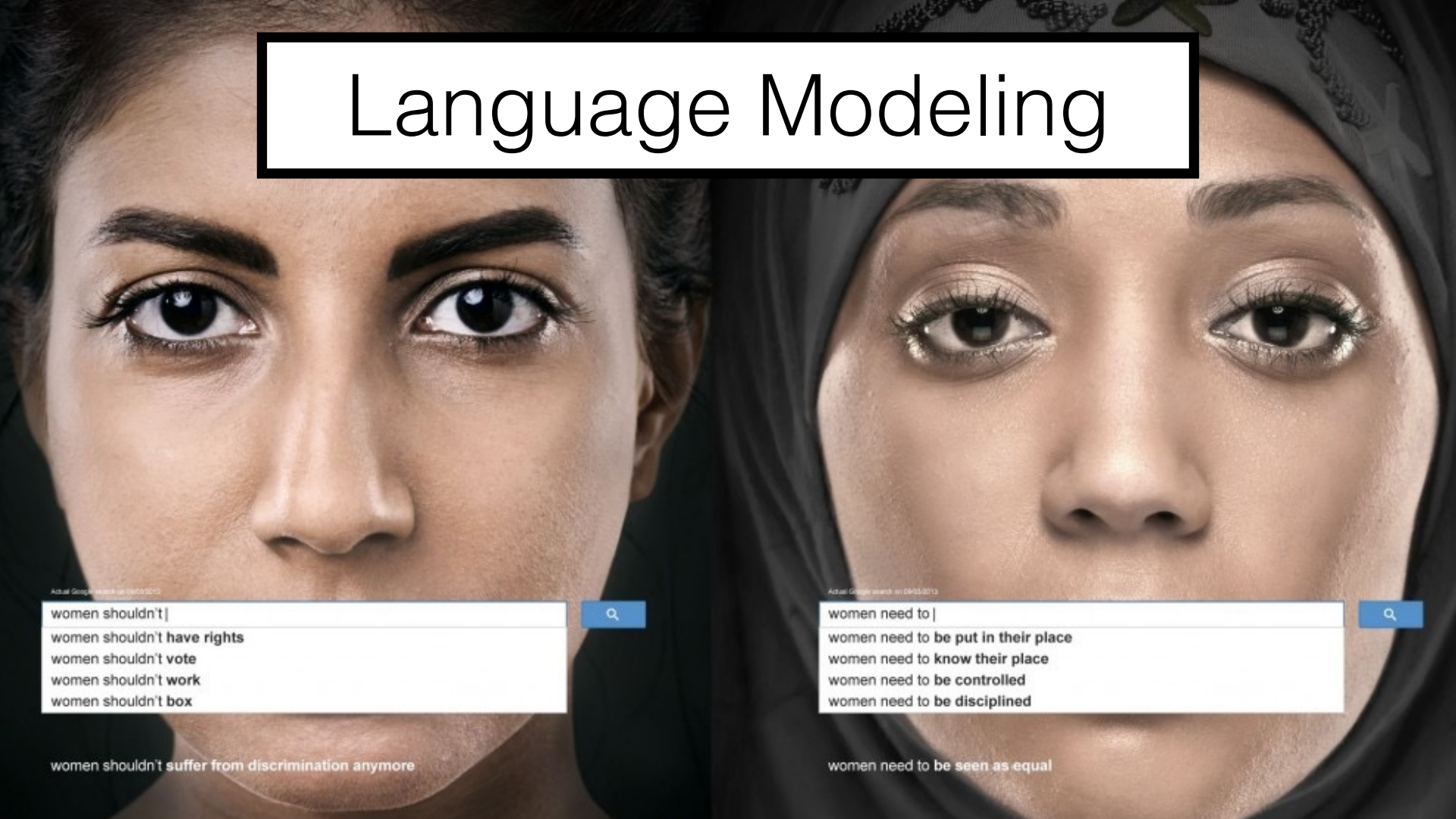
Question Answering

According To Google, Barack Obama Is King Of The United States

Google Answers gets it wrong. Is this a Google Answers Bomb?

Barry Schwartz on November 25, 2014 at 6:04 pm

Language Modeling



Ethics

- The decisions we make about our methods — training data, algorithm, evaluation — are often tied up with its use and impact in the world.
- NLP is now being used more and more to reason about human behavior.

Ethics

- Bias leading to allocational or representational harms.
- Privacy
- Exclusion
- Dual Use

Bias

- Allocational harms: automated systems allocate resources unfairly to different groups (access to housing, credit, parole).
- Representational harms: automated systems represent one group less favorably than another (including demeaning them or erasing their existence).

Blodgett et al. (2020), “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”

Adverse impact

“substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group”

Uniform Guidelines on Employee Selection Procedures

Allocations

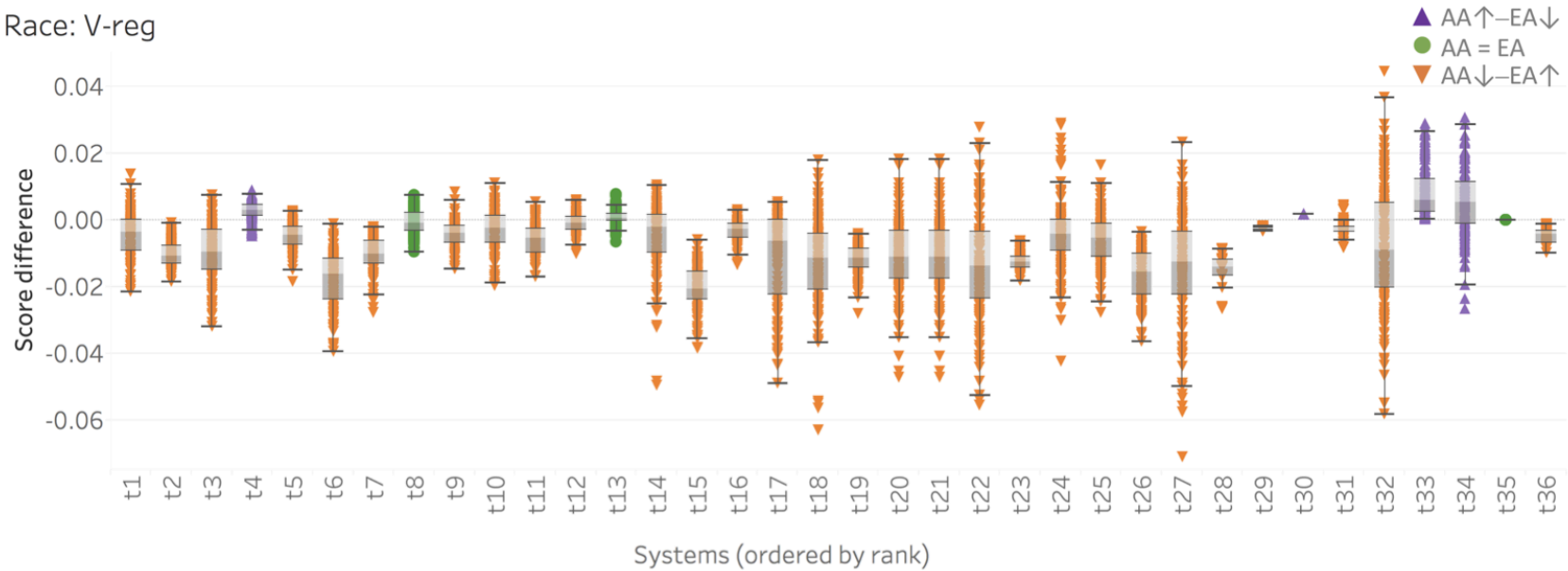
- Credit opportunities
- Assess to housing
- Job opportunities (LinkedIn, HR)
- Predictive policing

Not just categorical decisions, but e.g. advertising choices

Representations

- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

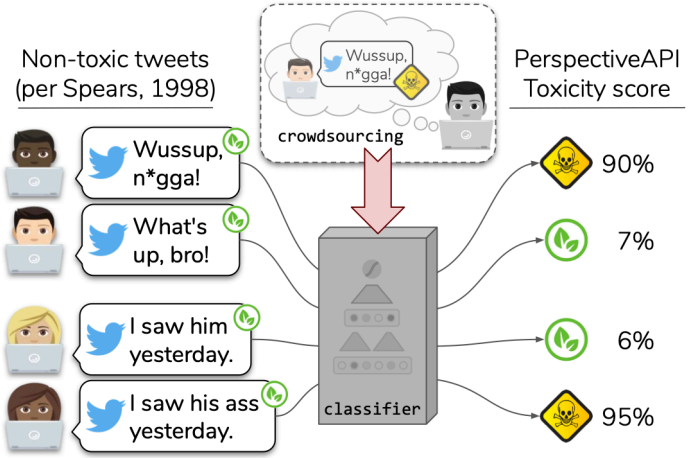
- Embeddings for African-American first names are closer to “unpleasant” words than European names (Caliskan et al. 2017)



- Sentiment analysis over sentences containing African-American first names are more negative than identical sentences with European names

Kiritchenko and Mohammad (2018), "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems"

- Toxicity detection systems score text with African-American English as more offensive
- Implicit negative perception of AAE → more AAE tweets are removed → users change language practices



Blodgett et al. (2020); Sap et al. (2019), “The risk of racial bias in hate speech detection”

Privacy

How Companies Learn Your Secrets

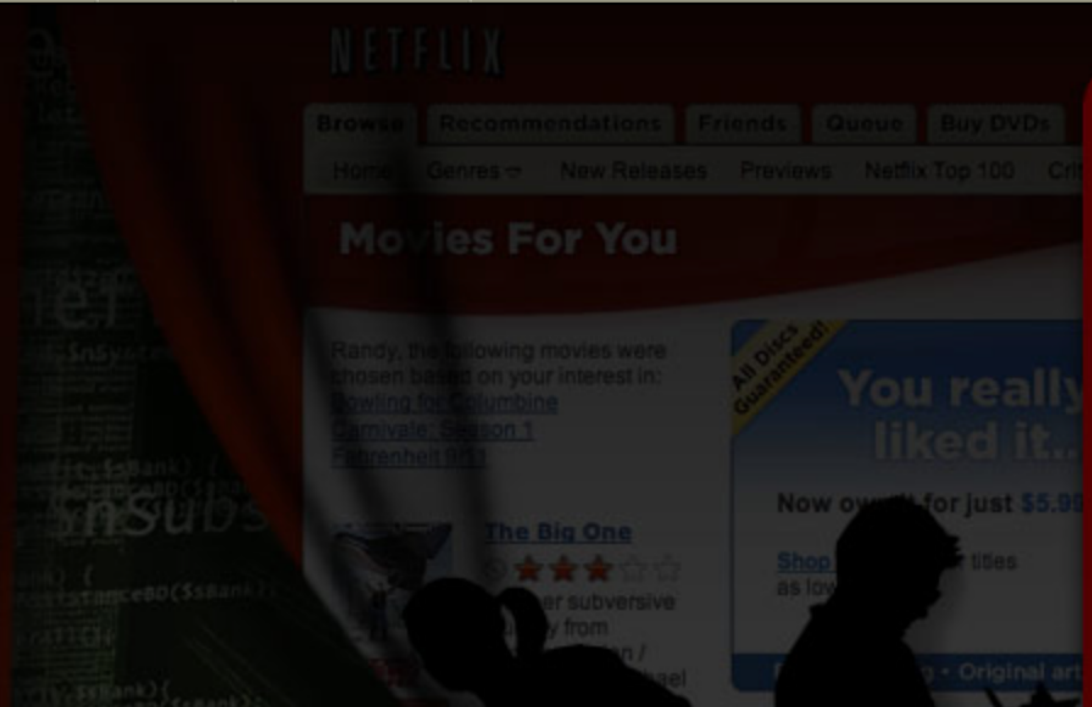
By CHARLES DUHIGG FEB. 16, 2012

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update



Congratulations!

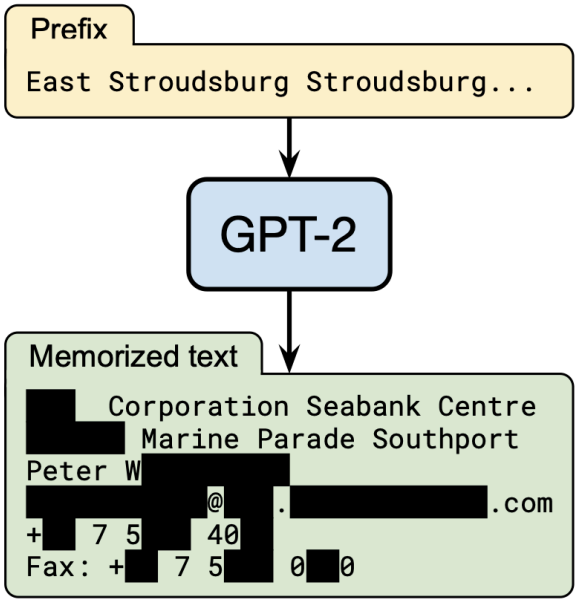
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the $1M Grand Prize to team “BellKor’s Pragmatic Chaos”. Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.

We applaud all the contributors to this

Privacy

- Large language models (e.g,. GPT-3, BERT) can memorize training data, which is recoverable from it.
- Potential violations of confidential data (e.g., GMail messages) and contextual integrity (data being published in a way that violates a user's expectations of use).



Carlini et al. (2020), "Extracting Training Data from Large Language Models"

Exclusion

- Focus on data from one domain/demographic
- State-of-the-art models perform worse for young (Hovy and Søgaard 2015) and minorities (Blodgett et al. 2016)

Exclusion

	AAE	White-Aligned
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Table 3: Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. (§4.1)

Language identification

Parser	AA	Wh.	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)

Dependency parsing

Blodgett et al. (2016), "Demographic Dialectal Variation in Social Media: A Case Study of African-American English" (EMNLP)

Dual Use

- Authorship attribution (author of *Federalist Papers* vs. author of ransom note vs. author of political dissent)
- Fake review detection vs. fake review generation
- Censorship evasion vs. enabling more robust censorship

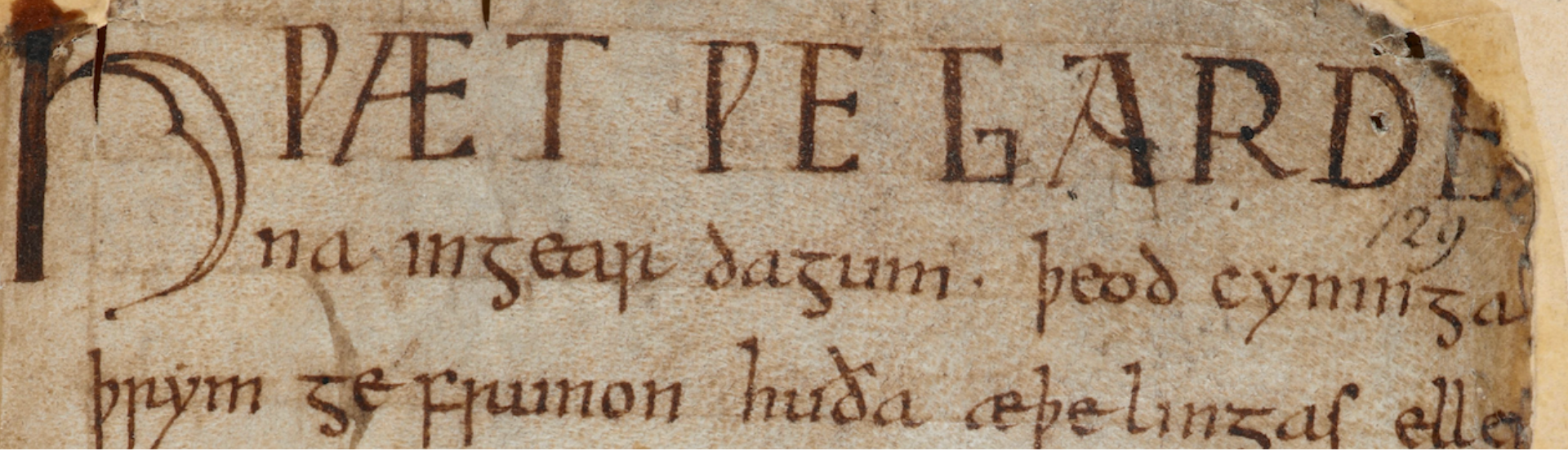
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

NAACL 2021

Ethics, Bias, and Fairness. This area includes work that analyzes, detects and mitigates stereotypical bias or offensive wording in language data as well as work discussing ethical concerns about NLP applications.

Homework 1

- Homework 1 will come out tonight: creating an annotated dataset by labeling the topic of NLP articles
- Due Wednesday 1/27 at 11:59pm.



Next time:

Classification 1