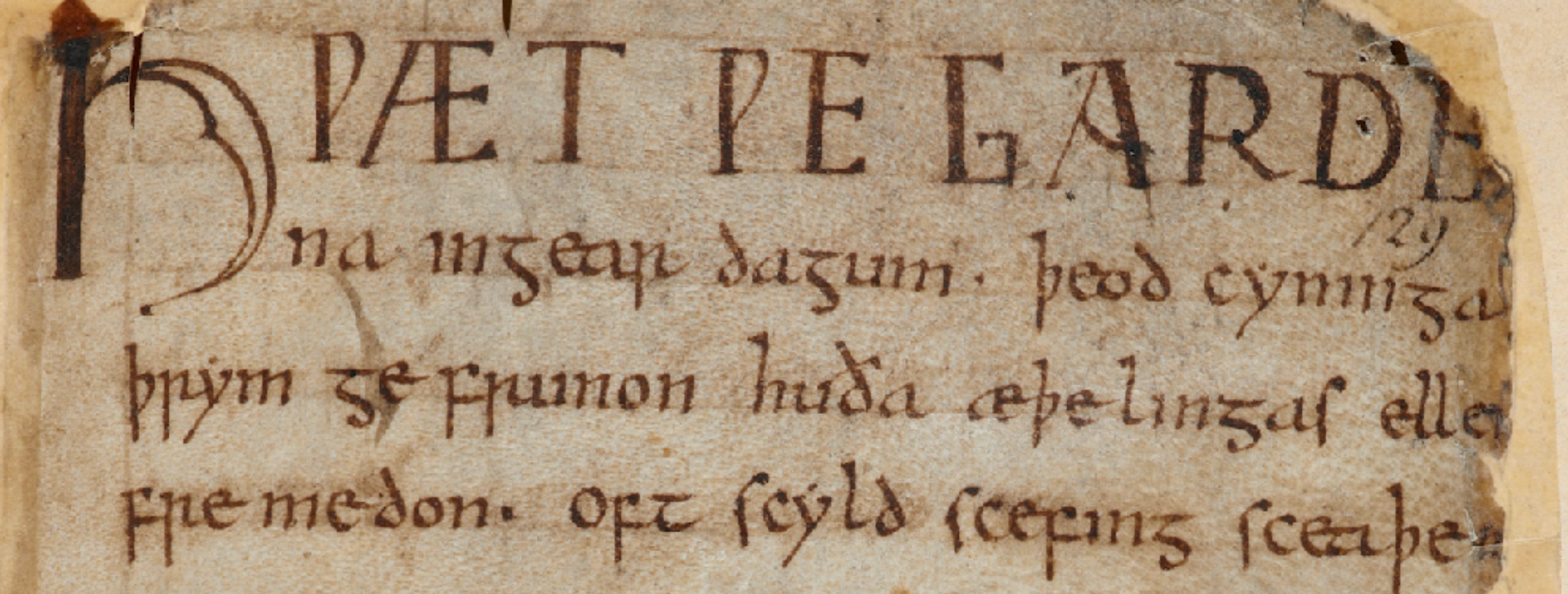




Natural Language Processing

Info 159/259

Lecture 5: Truth and ethics (Sept 6, 2018)

David Bamman, UC Berkeley

Hwæt! Wé Gárde
na in géardagum, þéodcýninga
þrym gefrúnon, hú ðá æþelingas ellen
fremedon. Oft Scyld Scéþing sceapena

Natural Language Processing

Info 159/259

Lecture 5: Truth and ethics (Sept 6, 2018)

David Bamman, UC Berkeley

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

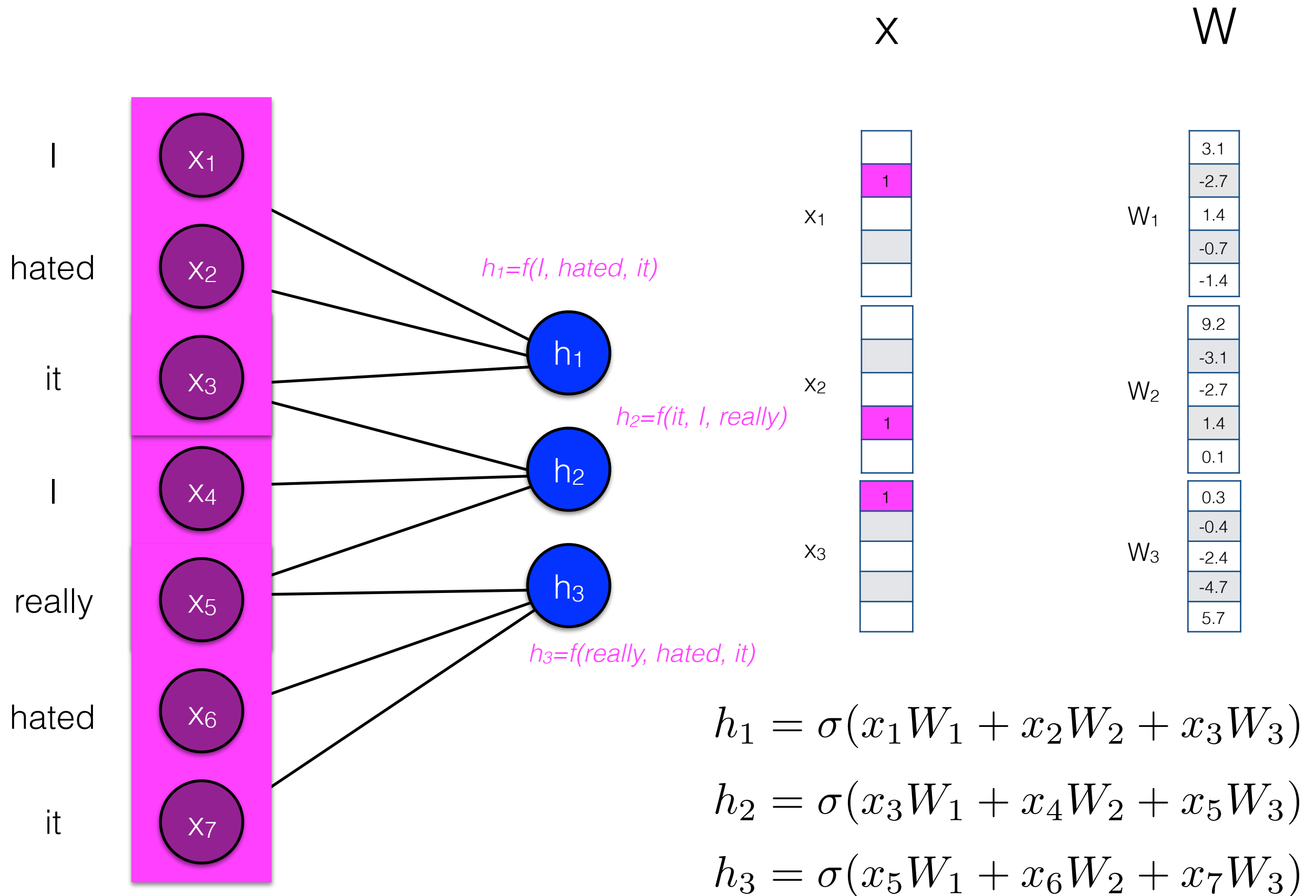
Sept. 5, 2018



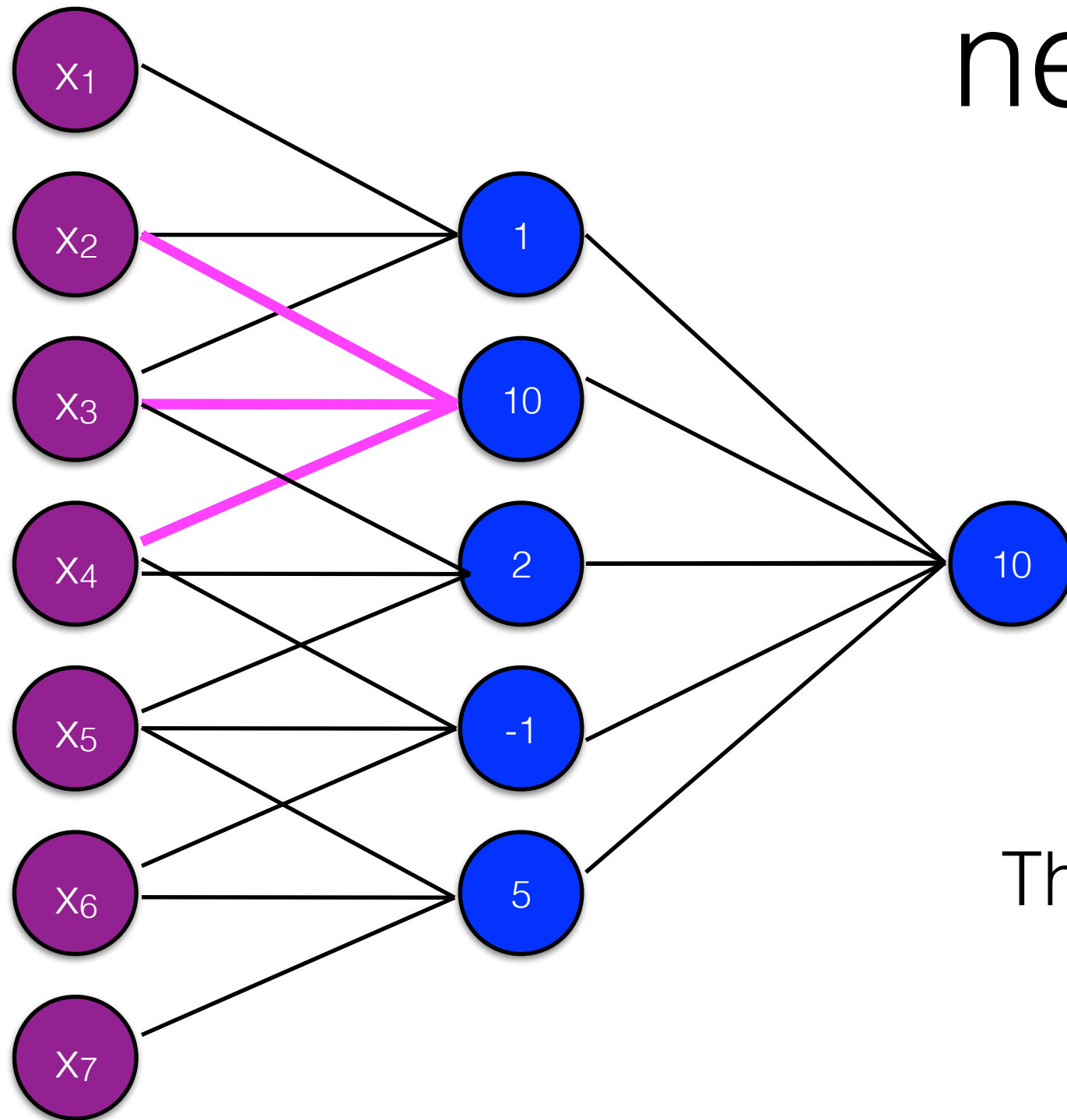
8956

[Leer en español](#) · [阅读简体中文版](#) · [閱讀繁體中文版](#)

*The Times today is taking the rare step of publishing an **anonymous** Op-Ed essay. We have done so at the request of the author, a senior official in the Trump administration whose identity is known to us and whose job would be jeopardized by its disclosure. We believe publishing this essay anonymously is the only way to deliver an important perspective to our readers. We invite you to submit a question about the essay or our vetting process [here](#).*



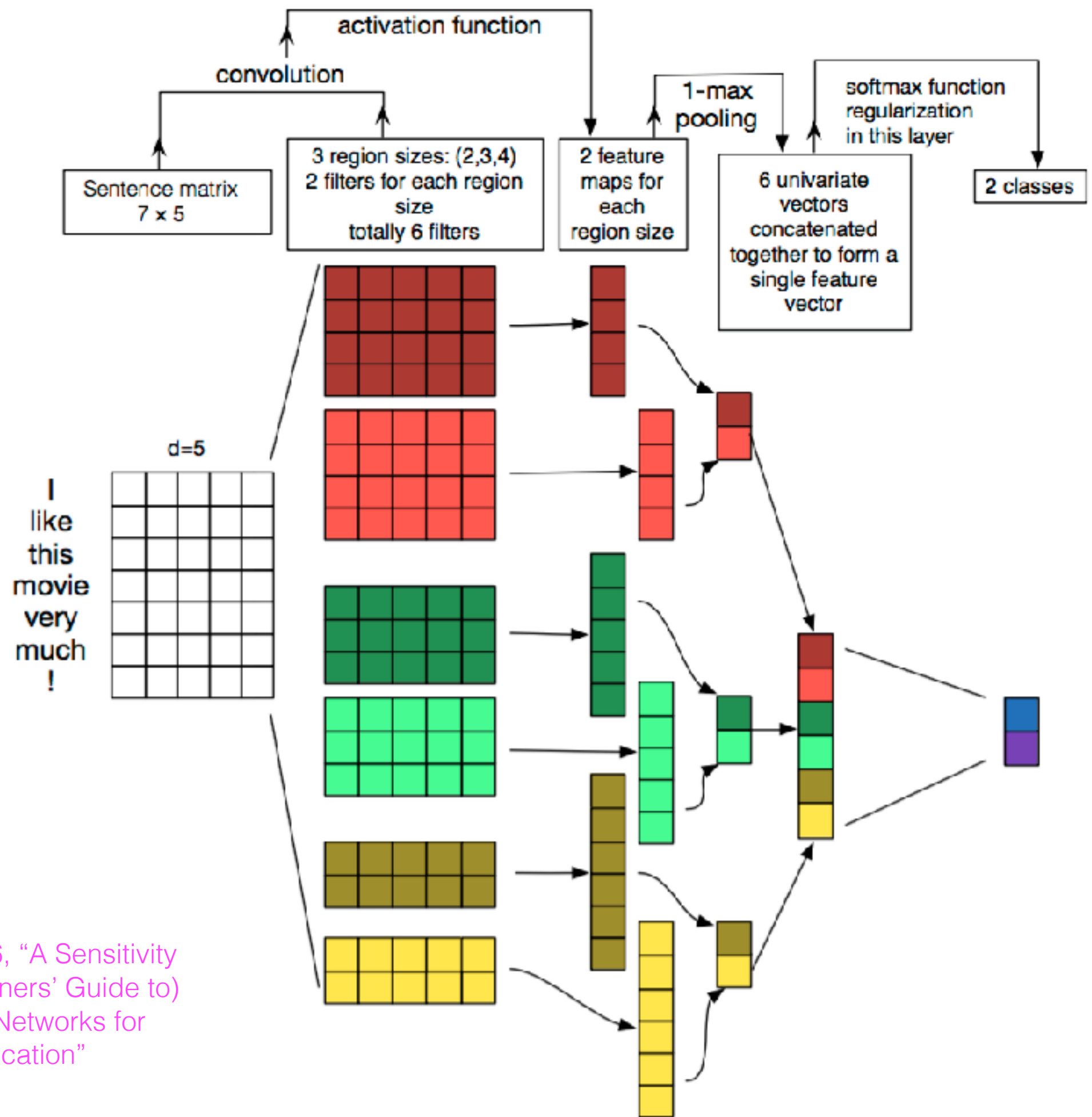
Convolutional networks



This defines one filter.

convolution

max pooling



Zhang and Wallace 2016, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification"

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.”



Donald J. Trump ✓
@realDonaldTrump

Follow

Ratings challenged @CNN reports so seriously that I call President Obama (and Clinton) "the founder" of ISIS, & MVP. THEY DON'T GET SARCASM?

3:26 AM - Aug 12, 2016

9,730 7,787 23,837

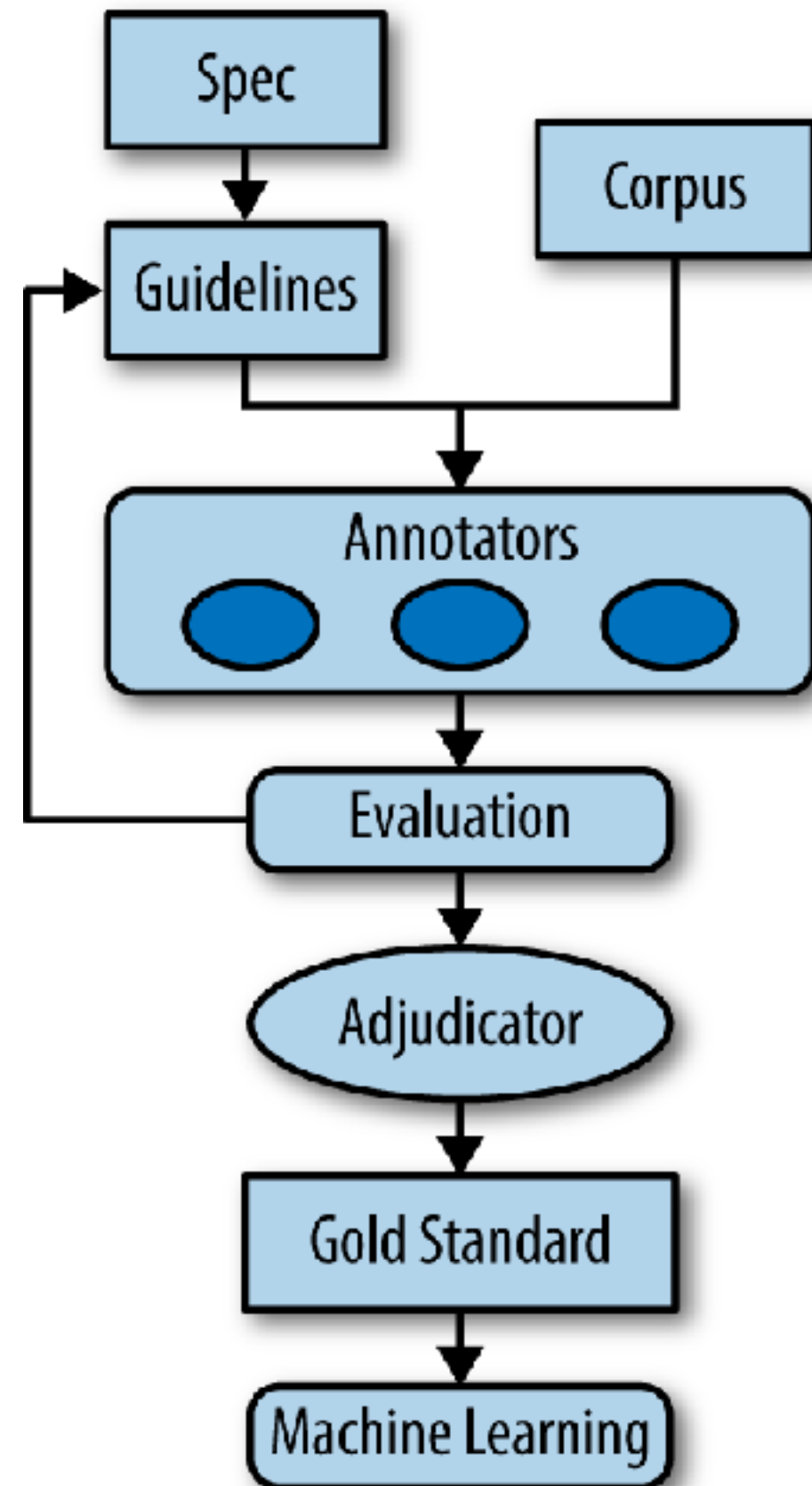


Fake News



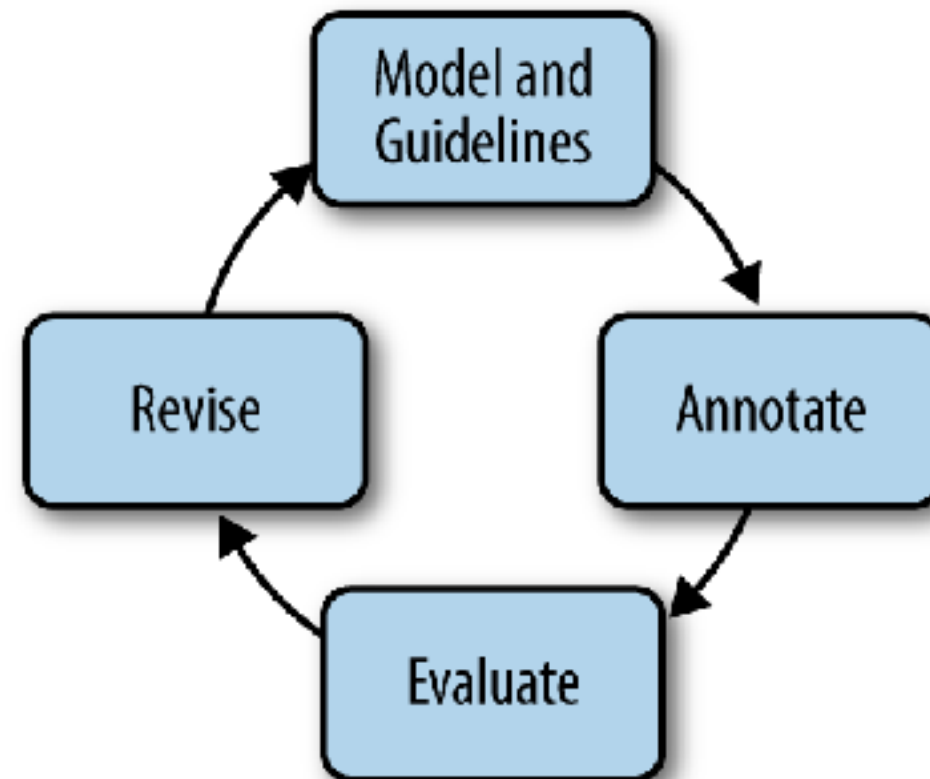
<http://www.fakenewschallenge.org>

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.)

Practicalities

- Annotation takes time, concentration (can't do it 8 hours a day)
- Annotators get better as they annotate (earlier annotations not as good as later ones)

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

puppy fried
chicken

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

annotator A

	puppy	fried chicken						
annotator B	<table><tr><td>puppy</td><td>7</td><td>4</td></tr><tr><td>fried chicken</td><td>8</td><td>81</td></tr></table>	puppy	7	4	fried chicken	8	81	
puppy	7	4						
fried chicken	8	81						

Cohen's kappa

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

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

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

annotator A

	puppy	fried chicken
annotator B		
puppy	7	4
fried chicken	8	81

Cohen's kappa

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

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

Cohen's kappa

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

$$P(A=\text{puppy}) \quad 15/100 = 0.15$$

$$P(B=\text{puppy}) \quad 11/100 = 0.11$$

$$P(A=\text{chicken}) \quad 85/100 = 0.85$$

$$P(B=\text{chicken}) \quad 89/100 = 0.89$$

$$\begin{aligned} &= 0.15 \times 0.11 + 0.85 \times 0.89 \\ &= 0.773 \end{aligned}$$

annotator B

annotator A

	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

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

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

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

$$= 0.471$$

annotator A

	puppy	fried chicken
annotator B		
puppy	7	4
fried chicken	8	81

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	50	0
fried chicken	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

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

Annotator bias correction

- Dawid, A. P. and Skene, A. M. "[Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm](#)," Journal of the Royal Statistical Society, 28(1):20–28, 1979.
- Weibe et al. (1999), "Development and use of a gold-standard data set for subjectivity classifications," ACL (for sentiment)
- Carpenter (2010), "Multilevel Bayesian Models of Categorical Data Annotation"
- Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. EMNLP 2008
- Sheng et al. (2008), "Get another label? improving data quality and data mining using multiple, noisy labelers", KDD.
- Raykar et al. (2009), "Supervised learning from multiple experts: whom to trust when everyone lies a bit," ICML
- Hovy et al. (2013), "Learning Whom to Trust with MACE," NAACL

Annotator bias correction

annotator label

truth

	positive	negative	mixed	unknown
positive	0.95	0	0.03	0.02
negative	0	0.80	0.10	0.10
mixed	0.20	0.05	0.50	0.25
unknown	0.15	0.10	0.10	0.70

$P(\text{label} \mid \text{truth})$

confusion matrix for a single annotator (David)

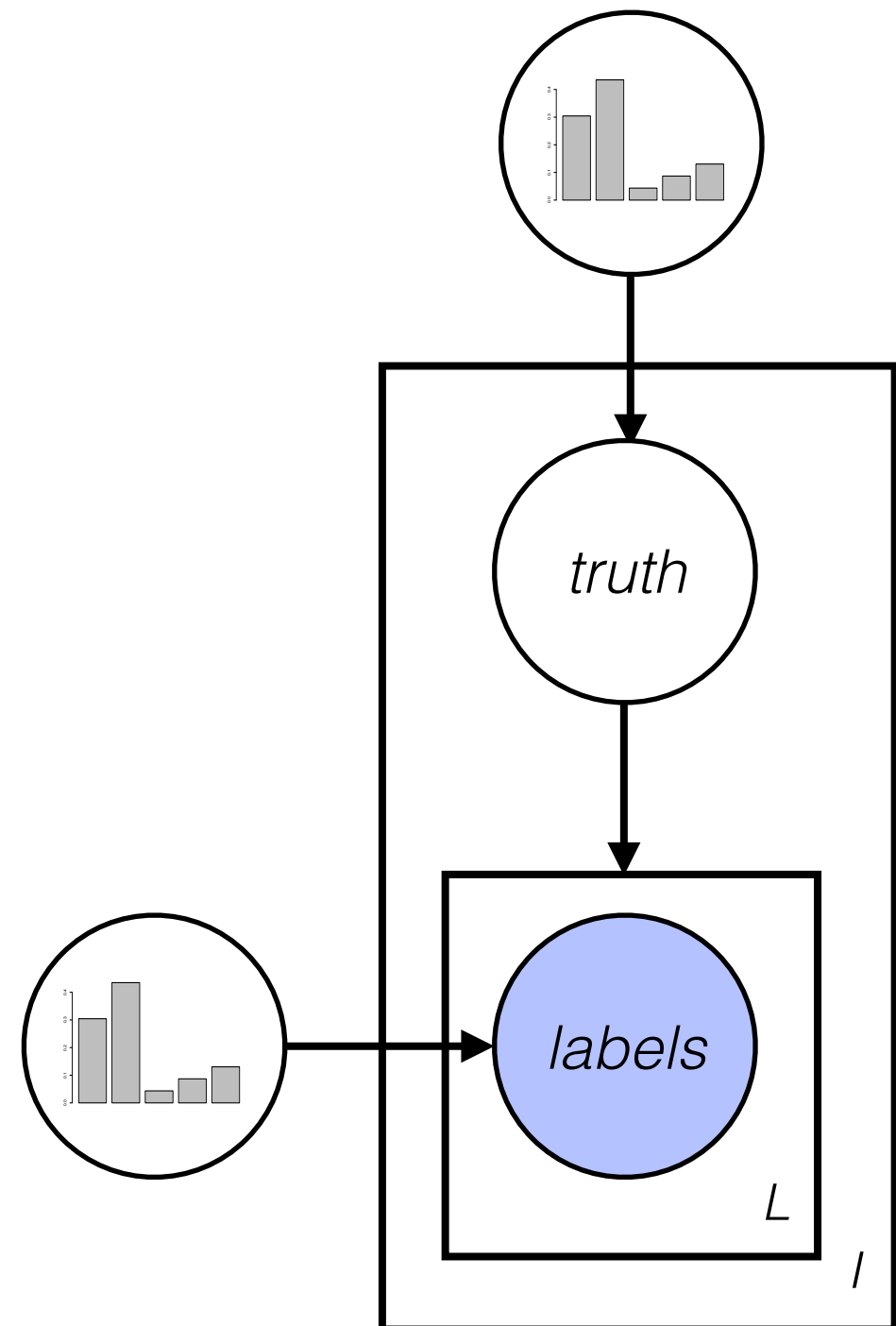
Annotator bias correction

Annotator bias
correction

Dawid and Skene 1979

Basic idea: the true label is unobserved; what we observe are noisy judgments by annotators

annotator confusion matrix
 $P(\text{label} \mid \text{truth})$



Evaluation

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



Classification

A mapping h from input data x (drawn from instance space $\mathcal{X}$) to a label (or labels) y from some enumerable output space $\mathcal{Y}$

$\mathcal{X}$ = set of all documents

$\mathcal{Y} = \{\text{english, mandarin, greek, ...}\}$

x = a single document

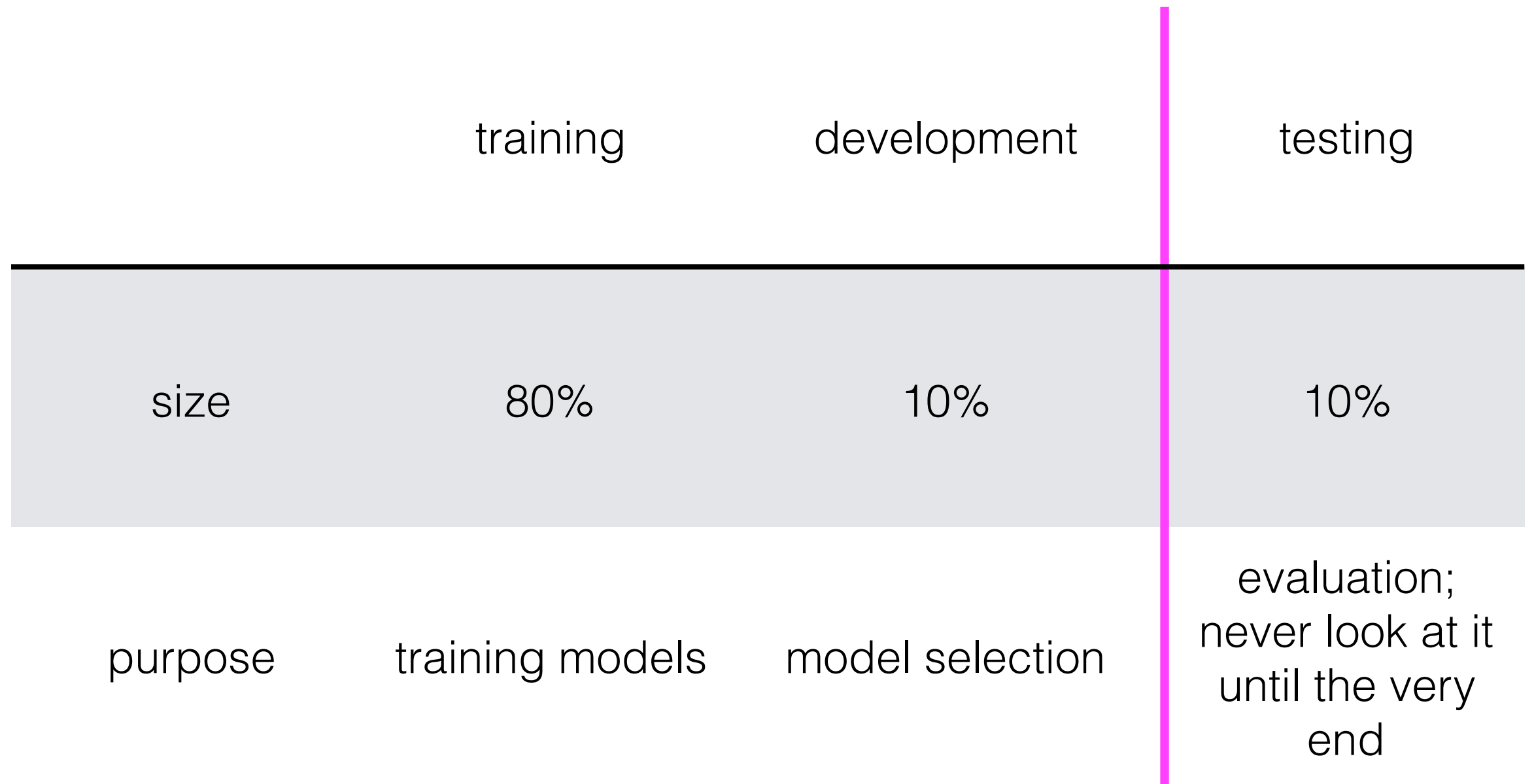
y = ancient greek

$\mathcal{X}$

instance space

train	dev	test
     	 	 

Experiment design



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

Multiclass confusion matrix

		Predicted ($\hat{y}$)		
		POS	NEG	NEUT
True (y)	POS	100	2	15
	NEG	0	104	30
	NEUT	30	40	70

Accuracy

$$\frac{1}{N} \sum_{i=1}^N I[\hat{y}_i = y_i]$$

$$I[x] = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Predicted ($\hat{y}$)

POS

NEG

NEUT

True (y)

	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

Precision

Precision(POS) =

$$\frac{\sum_{i=1}^N I(y_i = \hat{y}_i = \text{POS})}{\sum_{i=1}^N I(\hat{y}_i = \text{POS})}$$

Precision: proportion of predicted class that are actually that class.

True (y)

Predicted ($\hat{y}$)

	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

Recall

Recall(POS) =

$$\frac{\sum_{i=1}^N I(y_i = \hat{y}_i = \text{POS})}{\sum_{i=1}^N I(y_i = \text{POS})}$$

Recall: proportion of true class that are predicted to be that class.

True (y)

Predicted ($\hat{y}$)

	POS	NEG	NEUT
POS	100	2	15
NEG	0	104	30
NEUT	30	40	70

Significance

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?

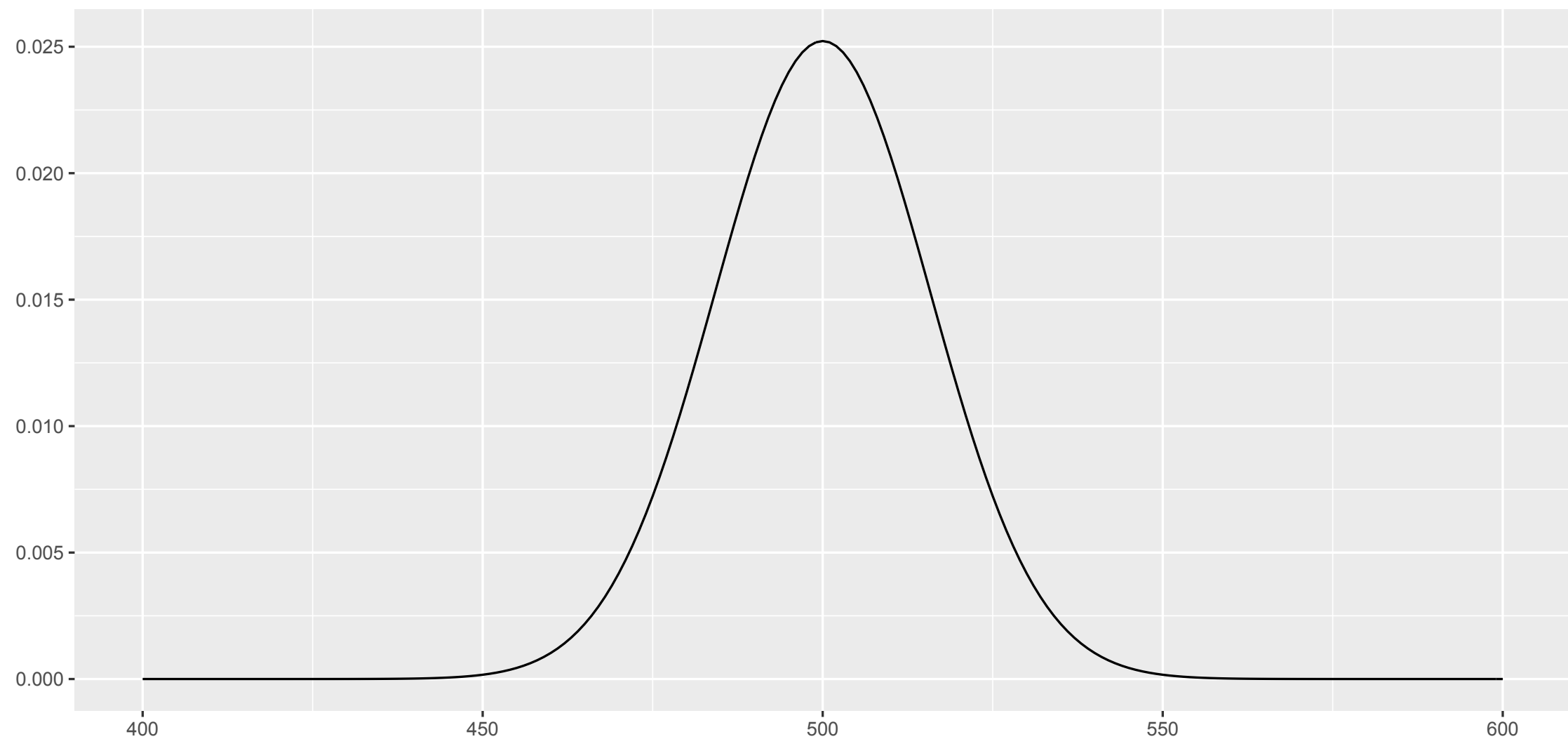
Hypothesis testing

- Hypothesis testing measures our confidence in what we can say about a null **from a sample**.

Hypothesis testing

- Current state of the art = 50%; your model = 58%. Both evaluated on the same test set of 1000 data points.
- Null hypothesis = there is no difference, so we would expect your model to get 500 of the 1000 data points right.
- If we make parametric assumptions, we can model this with a Binomial distribution (number of successes in n trials)

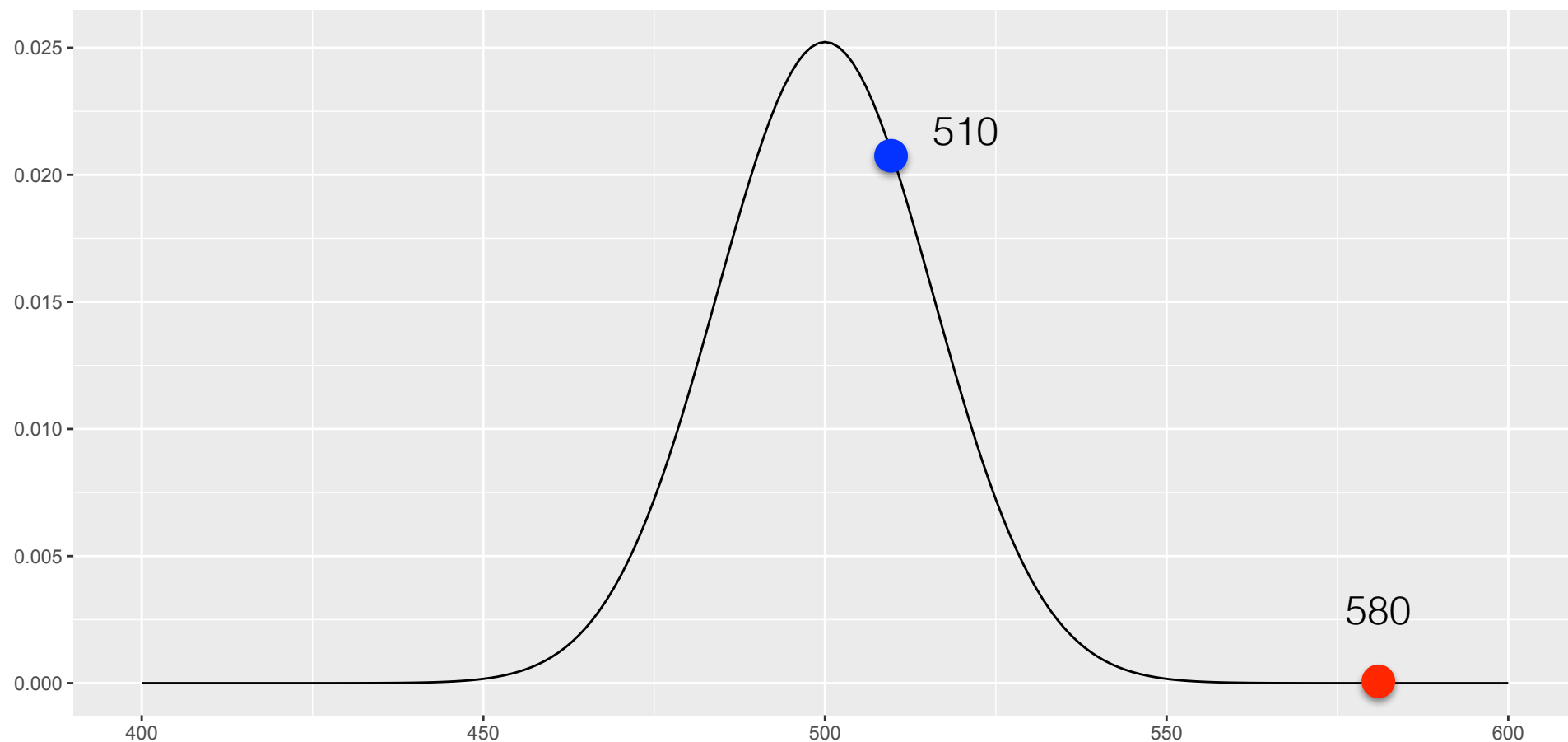
Example



Binomial probability distribution for number of correct predictions in $n=1000$ with $p = 0.5$

Example

At what point is a sample statistic **unusual enough** to reject the null hypothesis?



Example

- The form we assume for the null hypothesis lets us quantify that level of surprise.
- We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size n ; for large n , we can often make a normal approximation.

Tests

- Decide on the level of significance α . {0.05, 0.01}
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

1	“jobs” is predictive of presidential approval rating
2	“job” is predictive of presidential approval rating
3	“war” is predictive of presidential approval rating
4	“car” is predictive of presidential approval rating
5	“the” is predictive of presidential approval rating
6	“star” is predictive of presidential approval rating
7	“book” is predictive of presidential approval rating
8	“still” is predictive of presidential approval rating
9	“glass” is predictive of presidential approval rating
...	...
1,000	“bottle” is predictive of presidential approval rating

Errors

- For any significance level α and n hypothesis tests, we can expect $\alpha \times n$ type I errors.
- $\alpha=0.01$, $n=1000$ = 10 “significant” results simply by chance

Multiple hypothesis corrections

- Bonferroni correction: for family-wise significance level α_0 with n hypothesis tests:

$$\alpha \leftarrow \frac{\alpha_0}{n}$$

- [Very strict; controls the probability of at least one type I error.]
- False discovery rate

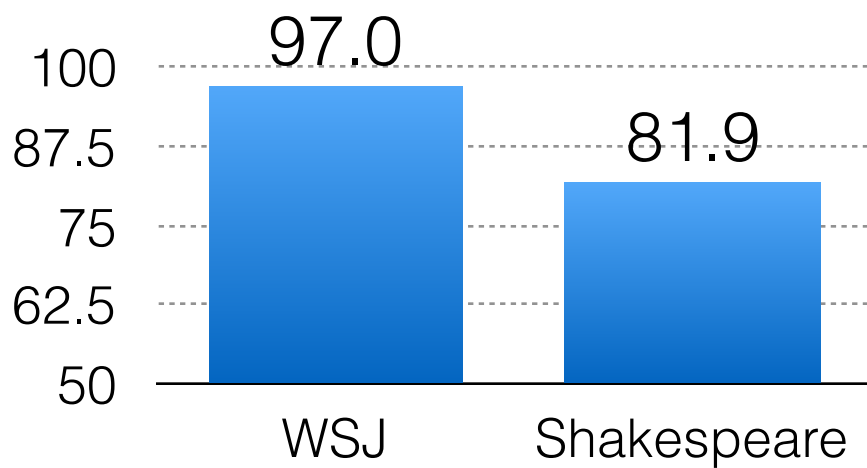
Nonparametric tests

- Many hypothesis tests rely on parametric assumptions (e.g., normality)
- Alternatives that don't rely on those assumptions:
 - permutation test
 - the bootstrap

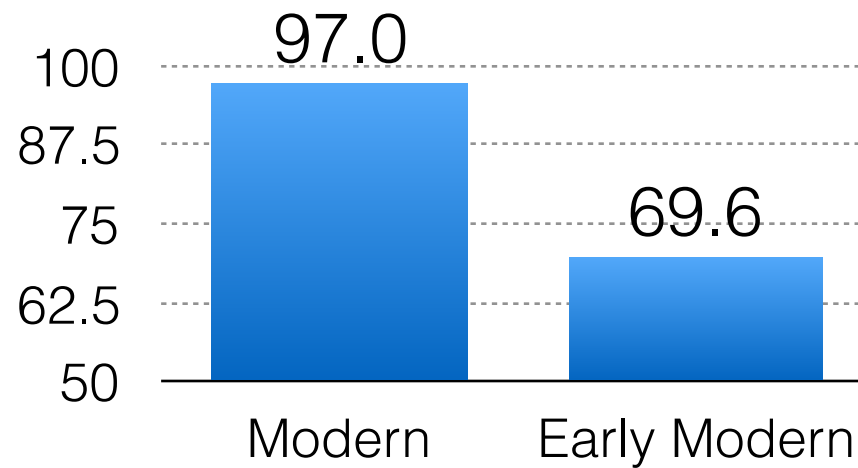
Issues

- Evaluation performance may not hold across domains (e.g., WSJ → literary texts)
- Covariates may explain performance (MT/parsing, sentences up to length n)
- Multiple metrics may offer competing results

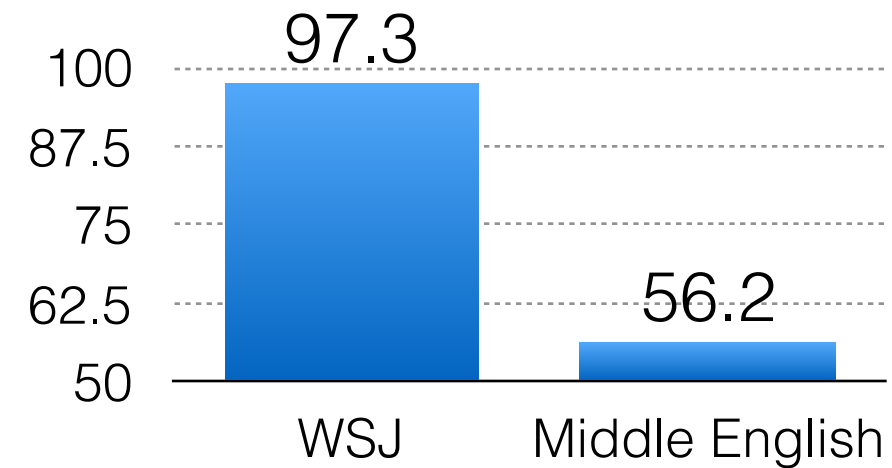
English POS



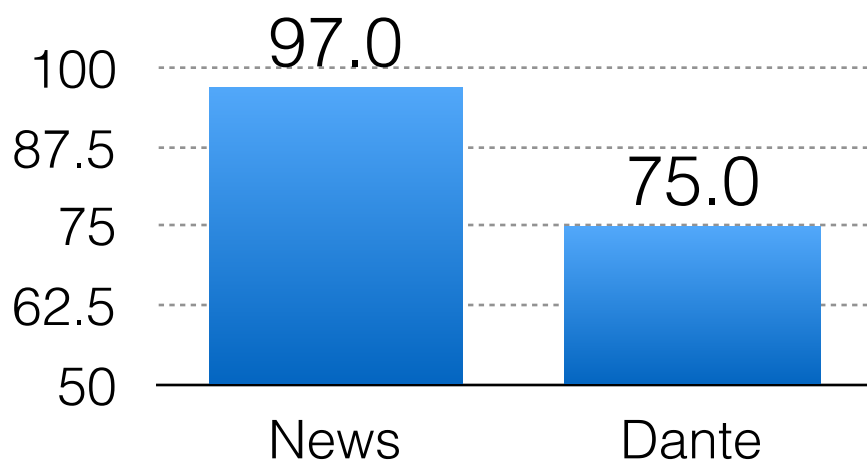
German POS



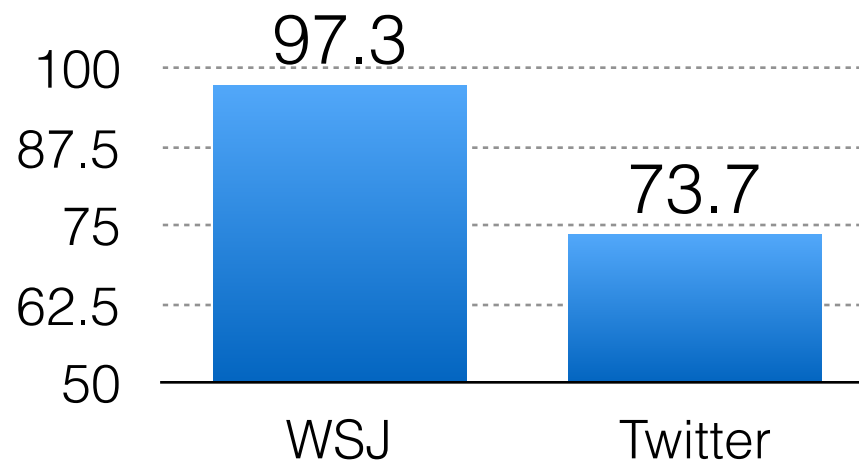
English POS



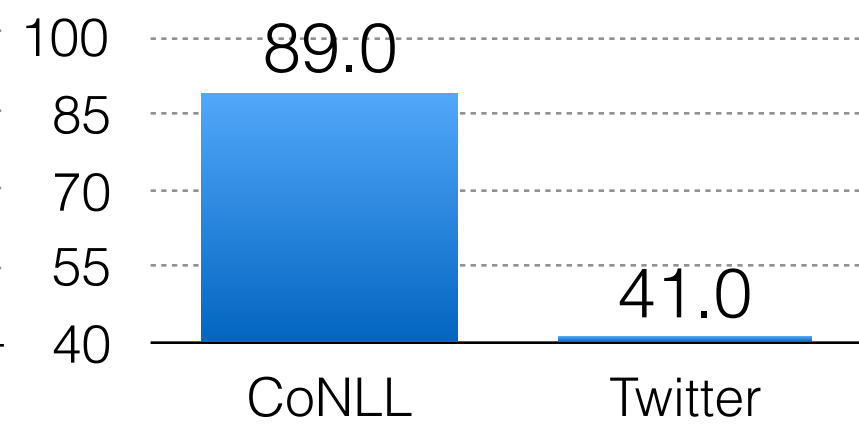
Italian POS



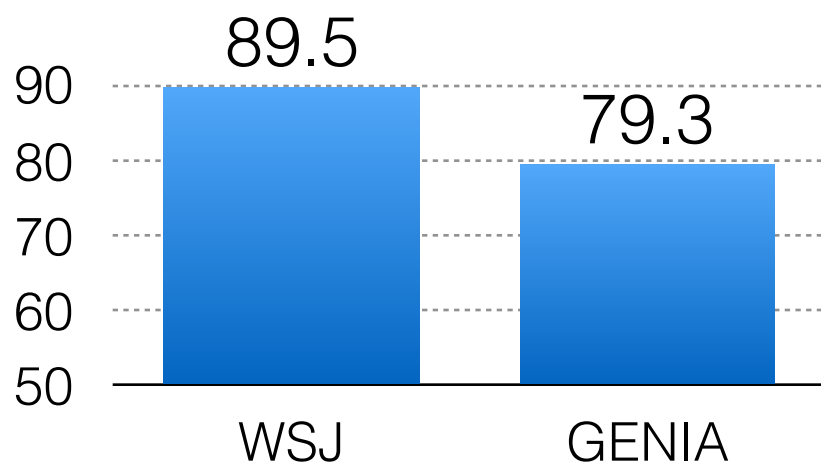
English POS



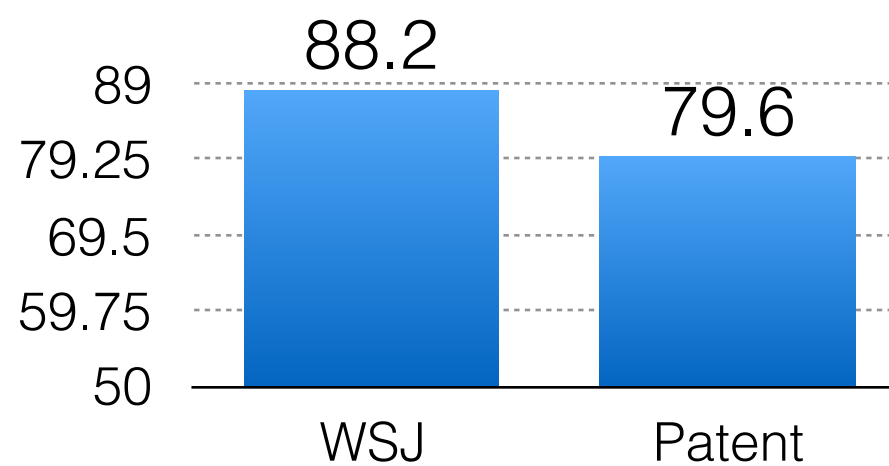
English NER



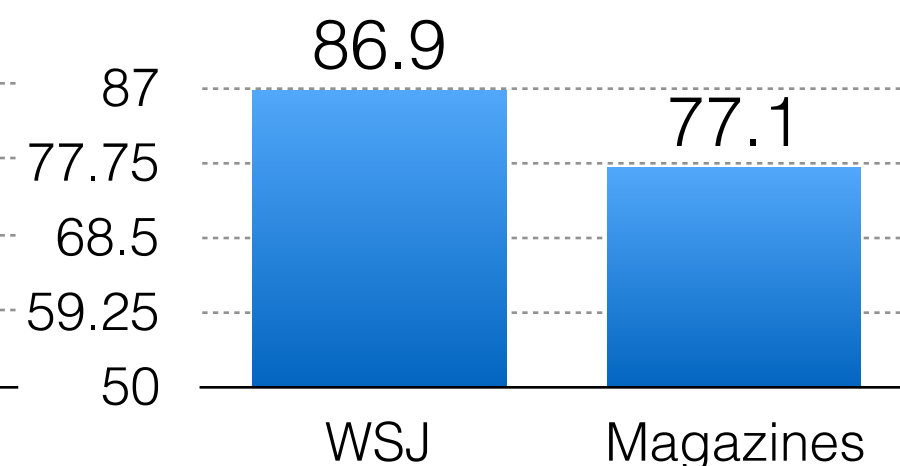
Phrase structure parsing



Dependency parsing



Dependency parsing



Ethics

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

Conversational Agents

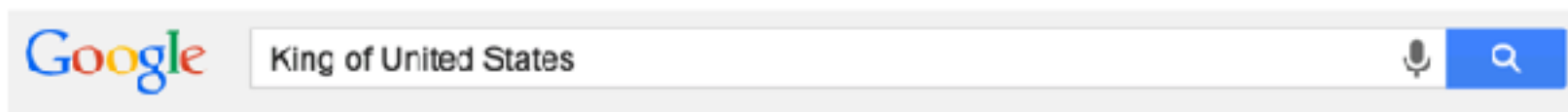


Question Answering

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



Web Maps Images Shopping Videos More ▾ Search tools

About 160,000,000 results (0.72 seconds)

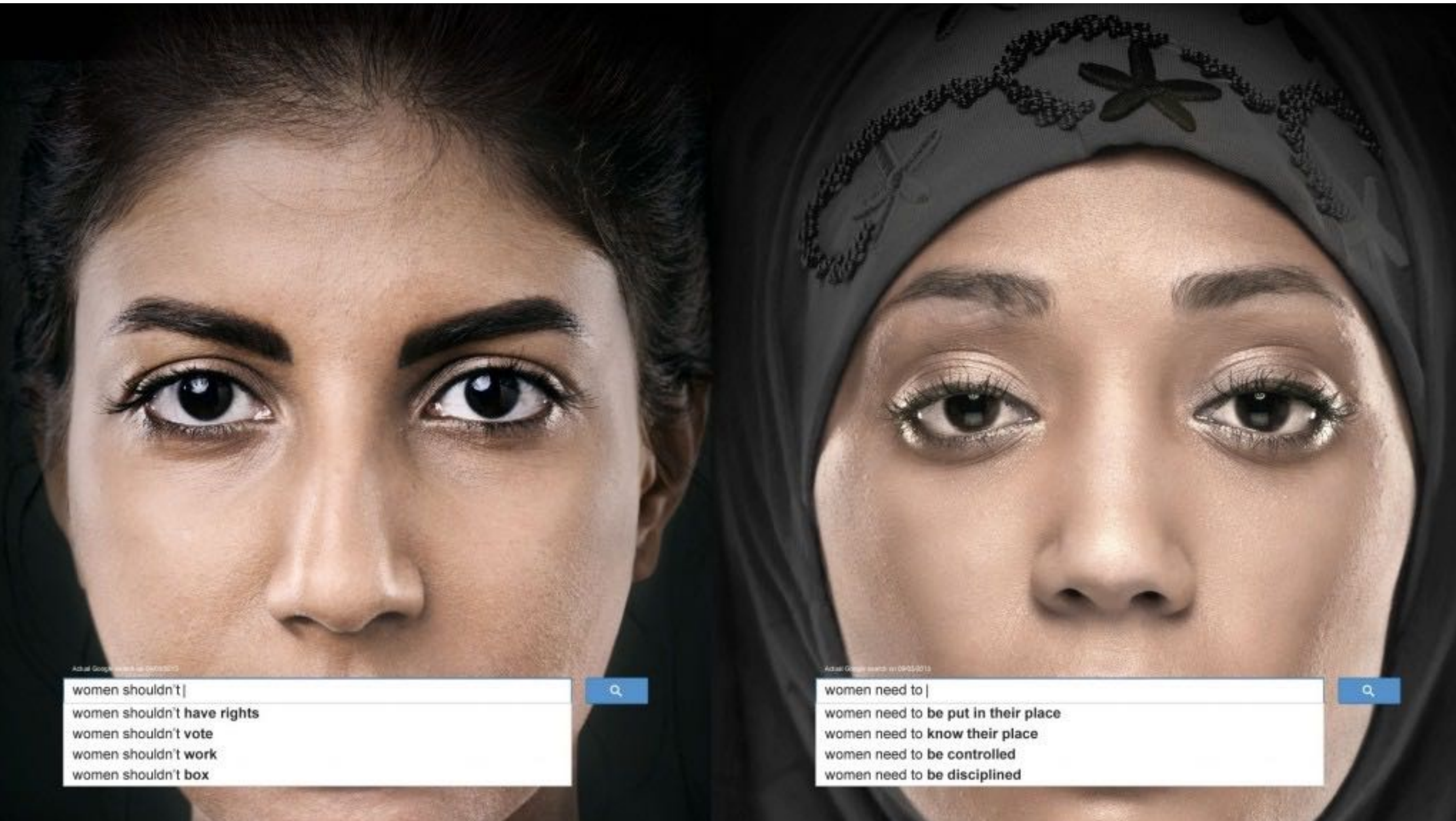
All Hail King **Barack Obama**, Emperor
Of The United States Of America!



All Hail King Barack Obama, Emperor Of The United States ...
www.breitbart.com/.../All-Hail-King-Barack-Obama-Emperor-Of-... Breitbart ▾

Feedback

Language Modeling



Vector semantics

Science AAAS

[Home](#)[News](#)[Journals](#)[Topics](#)[Careers](#)

[Science](#)[Science Advances](#)[Science Immunology](#)[Science Robotics](#)[Science Signaling](#)[Science Translational Medicine](#)


0




13

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}
[+ See all authors and affiliations](#)

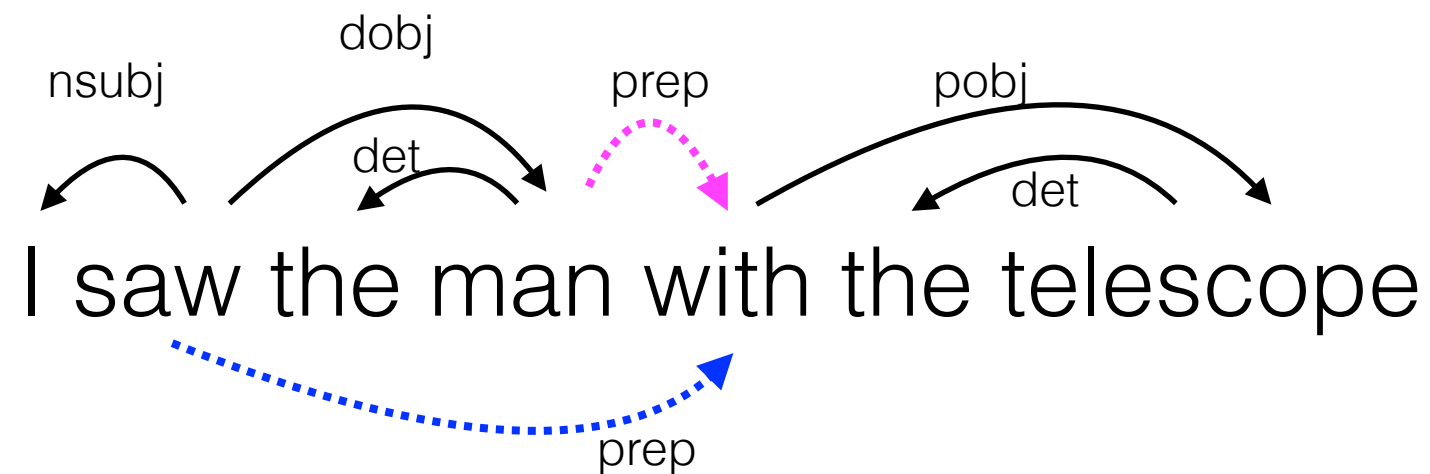
Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.acl4230

 **Peer Reviewed**
← see details

[Article](#)[Figures & Data](#)[Info & Metrics](#)[eLetters](#)[PDF](#)

- The decisions we make about our methods — training data, algorithm, evaluation — are often tied up with its use and **impact** in the world.

Scope



- NLP often operates on text divorced from the context in which it is uttered.
- It's now being used more and more to reason about human behavior.

Privacy

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012



NETFLIX

Netflix Prize

COMPLETED

Home

Rules

Leaderboard

Update

NETFLIX

Browse

Recommendations

Friends

Queue

Buy DVDs

Home

Genres

New Releases

Previews

Netflix Top 100

Crit

Movies For You

Randy, the following movies were chosen based on your interest in:
[Sawing for Columbine](#)
[Carnivale: Season 1](#)
[Fahrenheit 9/11](#)

The Big One

★★★★★

More subversive

by from

on /

angel

All Discs
Guaranteed!

You really
liked it...

Now own it for just \$5.99

Shop

as low

titles

Original art

OTI

HT

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 quest, which improves our ability to connect people to the movies they love.

Interventions

Facebook fiasco: was Cornell's study of 'emotional contagion' an ethics breach?

A covert experiment to influence the emotions of more than 600,000 people. A major scientific journal behaving like a rabbit in the headlights. A university in a PR tailspin



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
<i>langid.py</i>	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

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)

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

Language identification

Dependency parsing

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

Overgeneralization

- Managing and communicating the uncertainty of our predictions
- Is a false answer worse than no answer?

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

Homework 2

- Derive the updates for a CNN and implement the functions for forward/backward pass
- Out today, due Sept 20
- Be sure to check Piazza for any updates
- Start early! Another short homework will come out next Thursday.