



Applied Natural Language Processing

Info 256

Lecture 7: Testing (Feb 12, 2019)

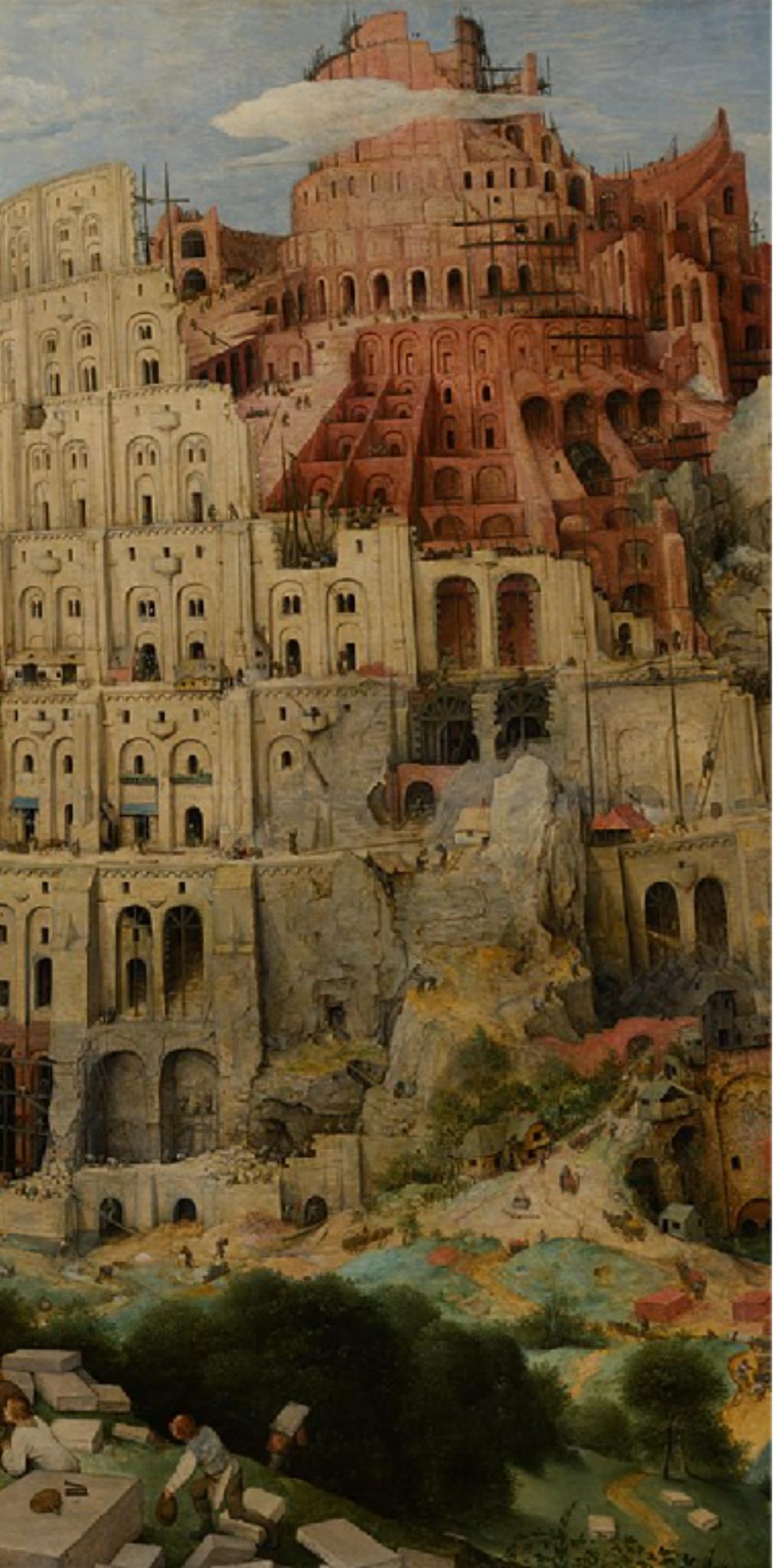
David Bamman, UC Berkeley

Significance in NLP

- You develop a new method for text classification; is it better than what comes before?
- You're developing a new model; should you include feature X? (when there is a cost to including it)
- You're developing a new model; does feature X reliably predict outcome Y?

Evaluation

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



Classification

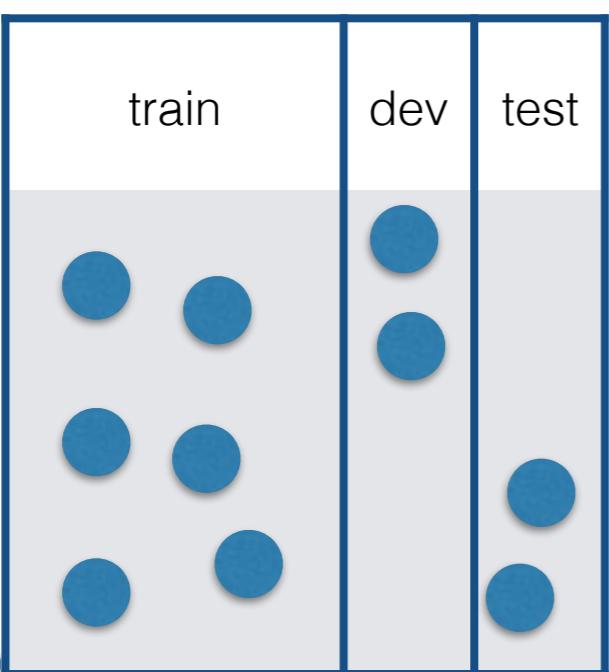
A mapping h from input data $\textcolor{magenta}{x}$ (drawn from instance space \mathcal{X}) to a label (or labels) $\textcolor{magenta}{y}$ from some enumerable output space \mathcal{Y}

$\textcolor{magenta}{\mathcal{X}}$ = set of all documents
 $\textcolor{magenta}{\mathcal{Y}}$ = {english, mandarin, greek, ...}

$\textcolor{magenta}{x}$ = a single document
 $\textcolor{magenta}{y}$ = ancient greek

\mathcal{X}

instance space



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.
 - Text classification: accuracy, precision, recall, F1
 - Phrase-structure parsing: PARSEVAL (bracketing overlap)
 - Dependency parsing: Labeled/unlabeled attachment score
 - Machine translation: BLEU, METEOR
 - Summarization: ROUGE
 - Language model: perplexity

Metrics

- Downstream tasks that use NLP to predict the natural world also have metrics:
 - Predicting presidential approval rates from tweets.
 - Predicting the type of job applicants from a job description.
 - Conversational agent

Binary classification



System B

Truth

puppy

fried chicken

		puppy	fried chicken
puppy	6	3	
	2	5	

Accuracy = 11/16 = 68.75%

Multiclass confusion matrix

		Predicted (\hat{y})		
		Dem	Repub	Indep
True (y)	Dem	100	2	15
	Repub	0	104	30
	Indep	30	40	70

Accuracy

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

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

True (y)

Predicted (\hat{y})

	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

Precision

$\text{Precision}(\text{Dem}) =$

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

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

True (y)

		Predicted (\hat{y})		
		Dem	Repub	Indep
True (y)	Dem	100	2	15
	Repub	0	104	30
Indep	30	40	70	

Recall

Recall(Dem) =

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

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

True (y)

	Dem	Repub	Indep
Dem	100	2	15
Repub	0	104	30
Indep	30	40	70

F score

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Ablation test

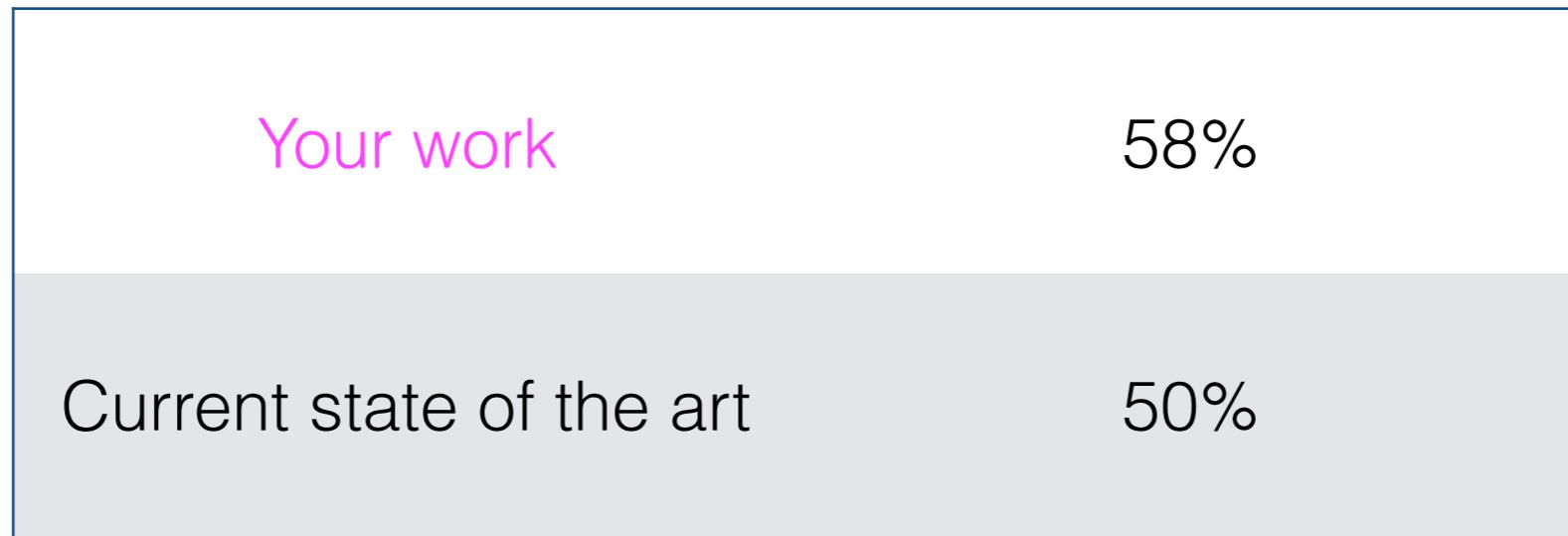
- To test how important individual features are (or components of a model), conduct an ablation test
 - Train the full model with all features included, conduct evaluation.
 - Remove feature, train reduced model, conduct evaluation.

Ablation test

	Dev.	Test	
Our tagger, all features	88.67	89.37	
independent ablations:			
–DISTSIM	87.88	88.31	(–1.06)
–TAGDICT	88.28	88.31	(–1.06)
–TWORTH	87.51	88.37	(–1.00)
–METAPH	88.18	88.95	(–0.42)
–NAMES	88.66	89.39	(+0.02)
Our tagger, base features	82.72	83.38	
Stanford tagger	85.56	85.85	
Annotator agreement	92.2		

Table 2: Tagging accuracies on development and test data, including ablation experiments. Features are ordered by importance: test accuracy decrease due to ablation (final column).

Significance



- 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?

Hypotheses

hypothesis

The average income in two sub-populations is different

Web design A leads to higher CTR than web design B

Self-reported location on Twitter is predictive of political preference

Your system X is better than state-of-the-art system Y

Null hypothesis

- A claim, assumed to be true, that we'd like to test (because we think it's wrong)

hypothesis

H_0

The average income in two sub-populations is different

The incomes are the *same*

Web design A leads to higher CTR than web design B

The CTR are the *same*

Self-reported location on Twitter is predictive of political preference

Location has *no* relationship with political preference

Your system X is better than state-of-the-art system Y

There is *no* difference in the two systems.

Hypothesis testing

- If the null hypothesis were true, how likely is it that you'd see the data you see?

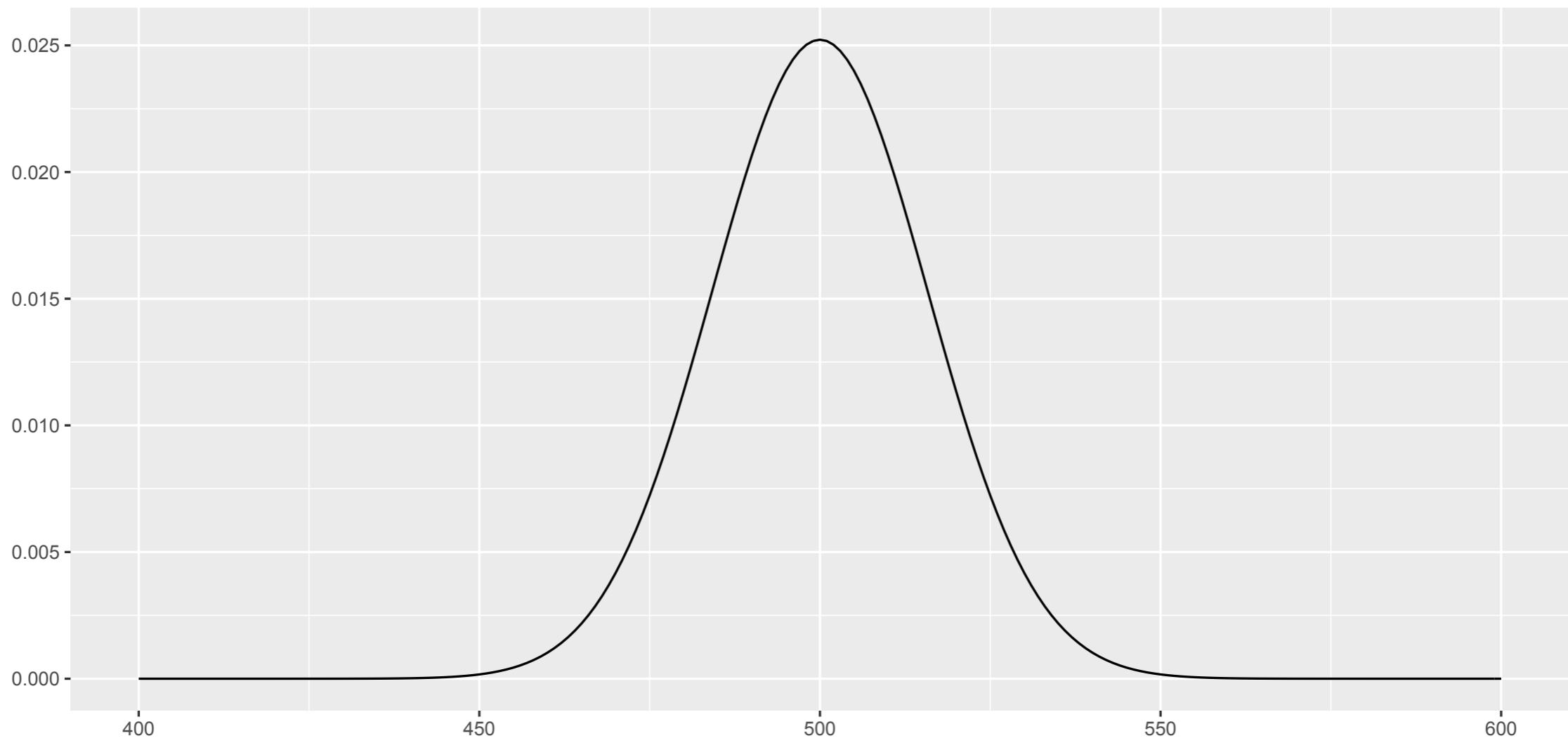
Hypothesis testing

- 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)

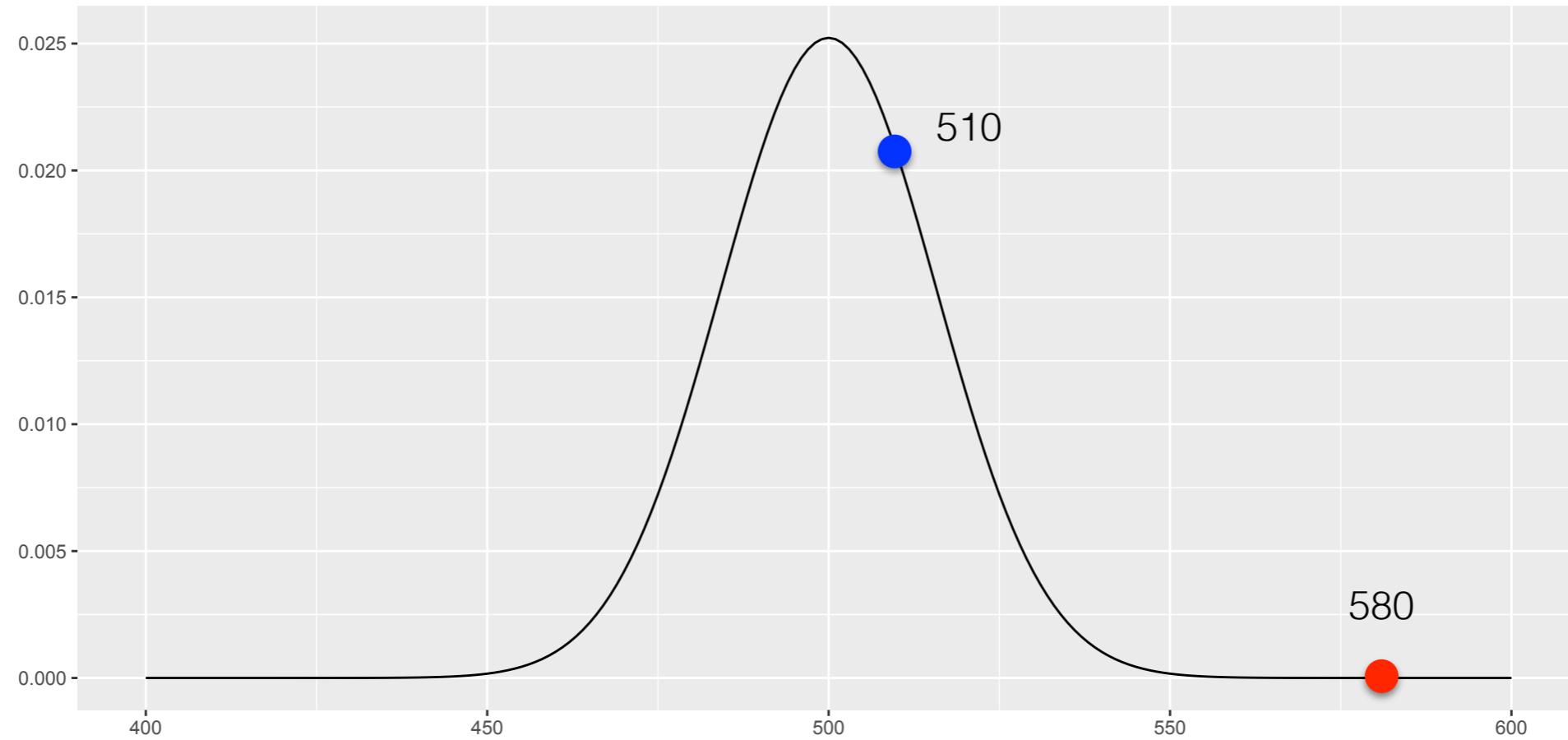
Example



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

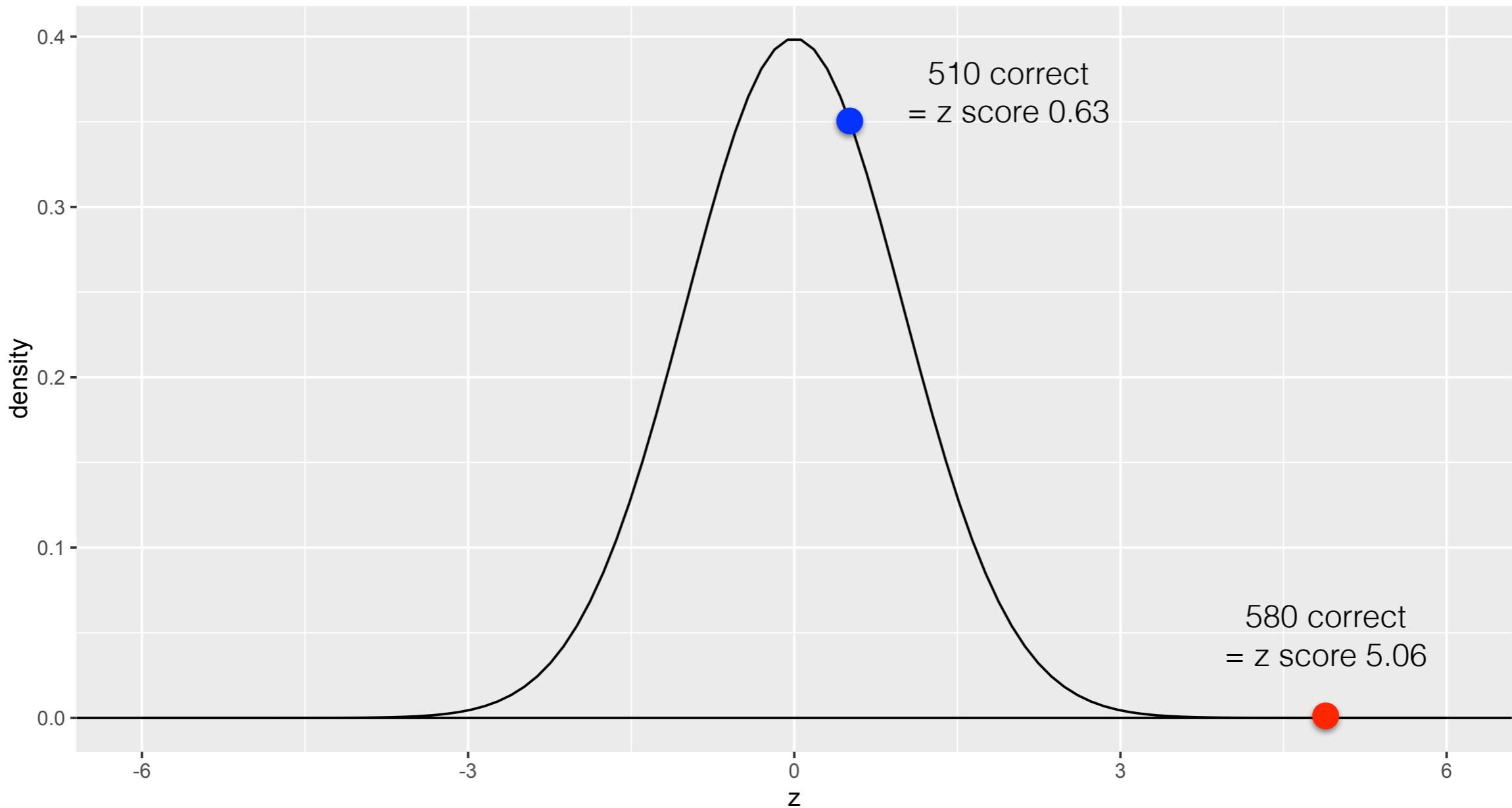
- The form we assume for the null hypothesis lets us quantify that level of surprise.
- We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size n ; for large n , we can often make a normal approximation.

Z score

$$Z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

For Normal distributions, transform into standard normal (mean = 0, standard deviation = 1)

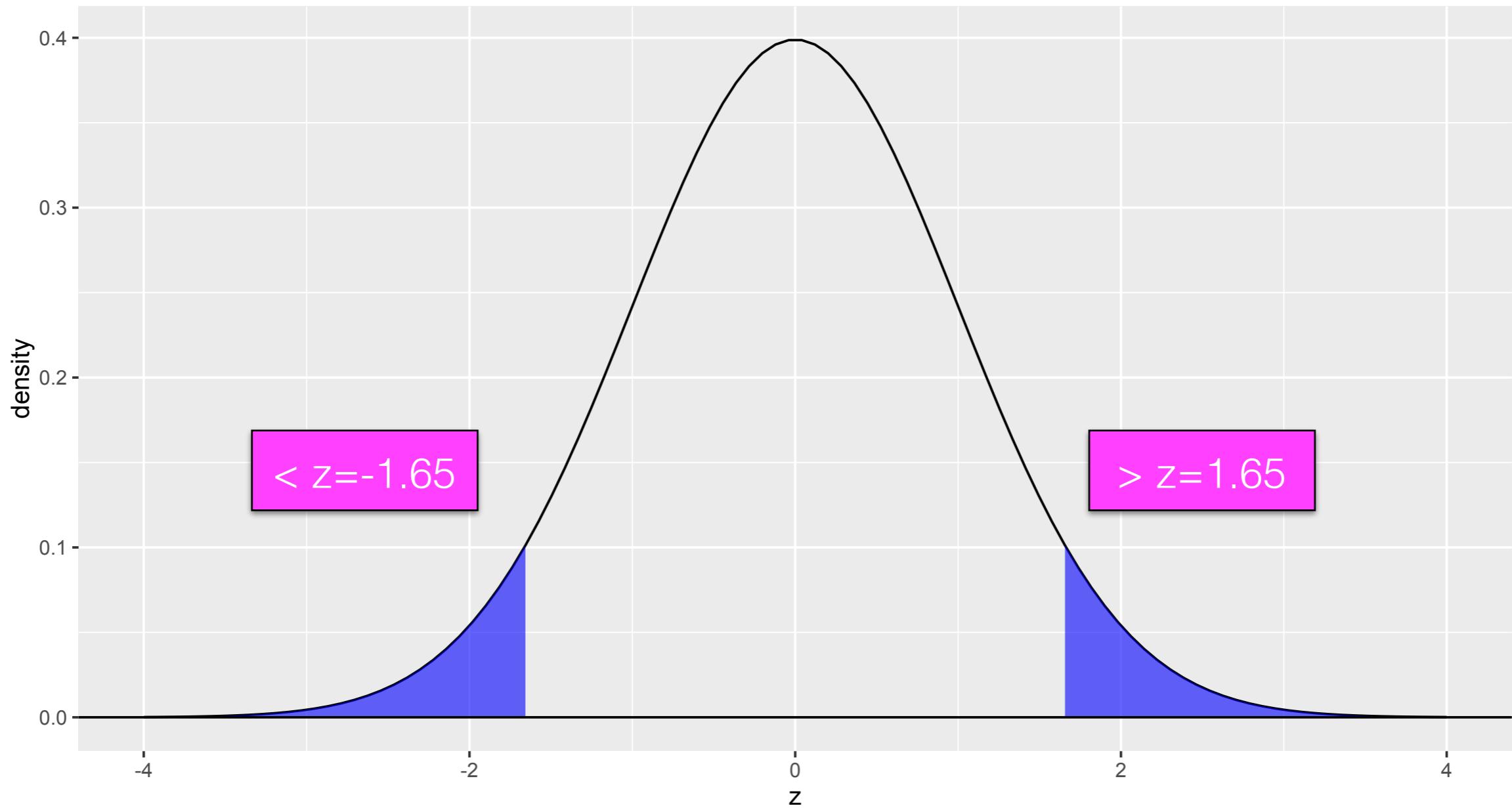
Z score



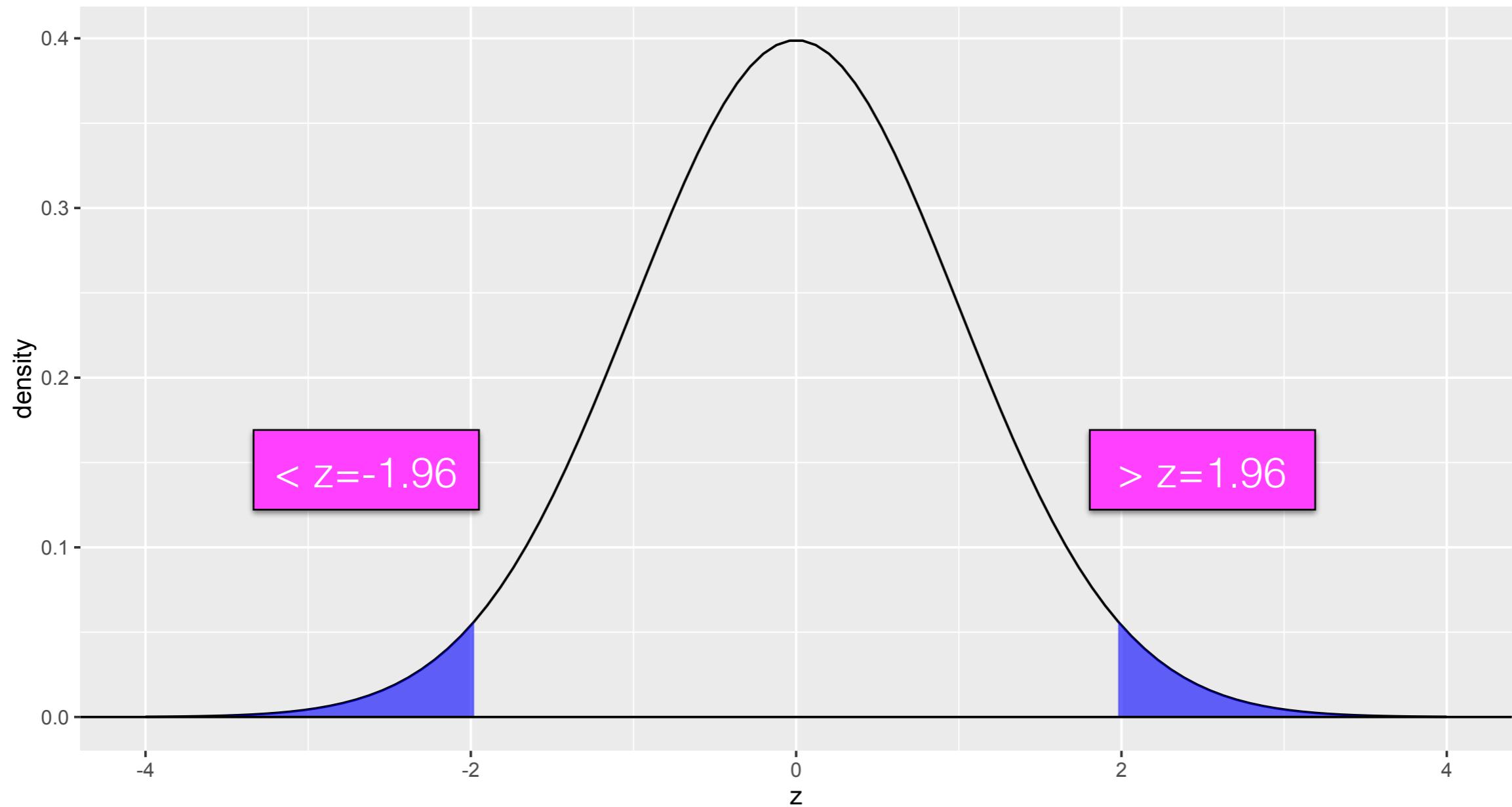
Tests

- We will define “unusual” to equal the most extreme areas in the tails

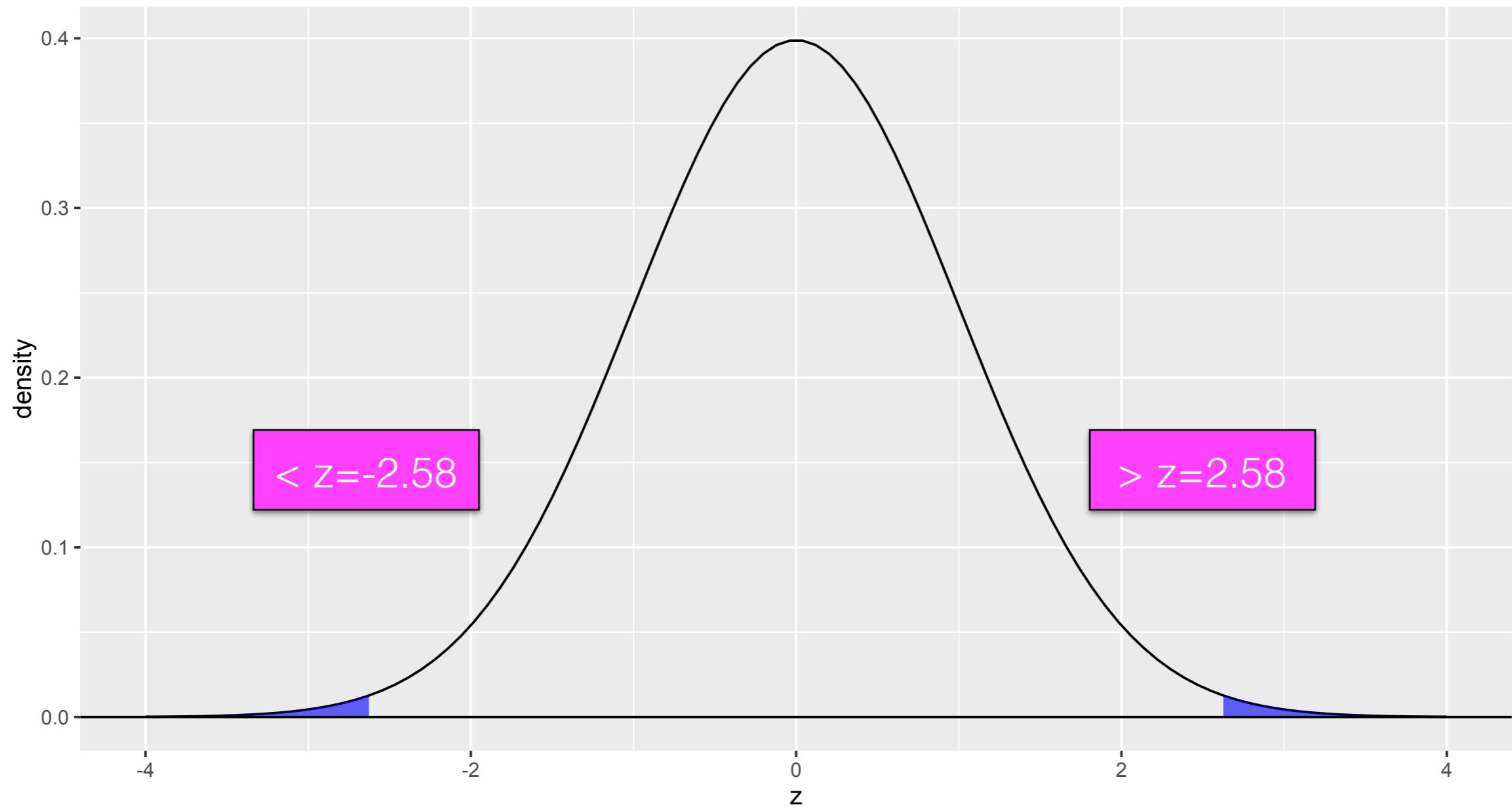
least likely 10%



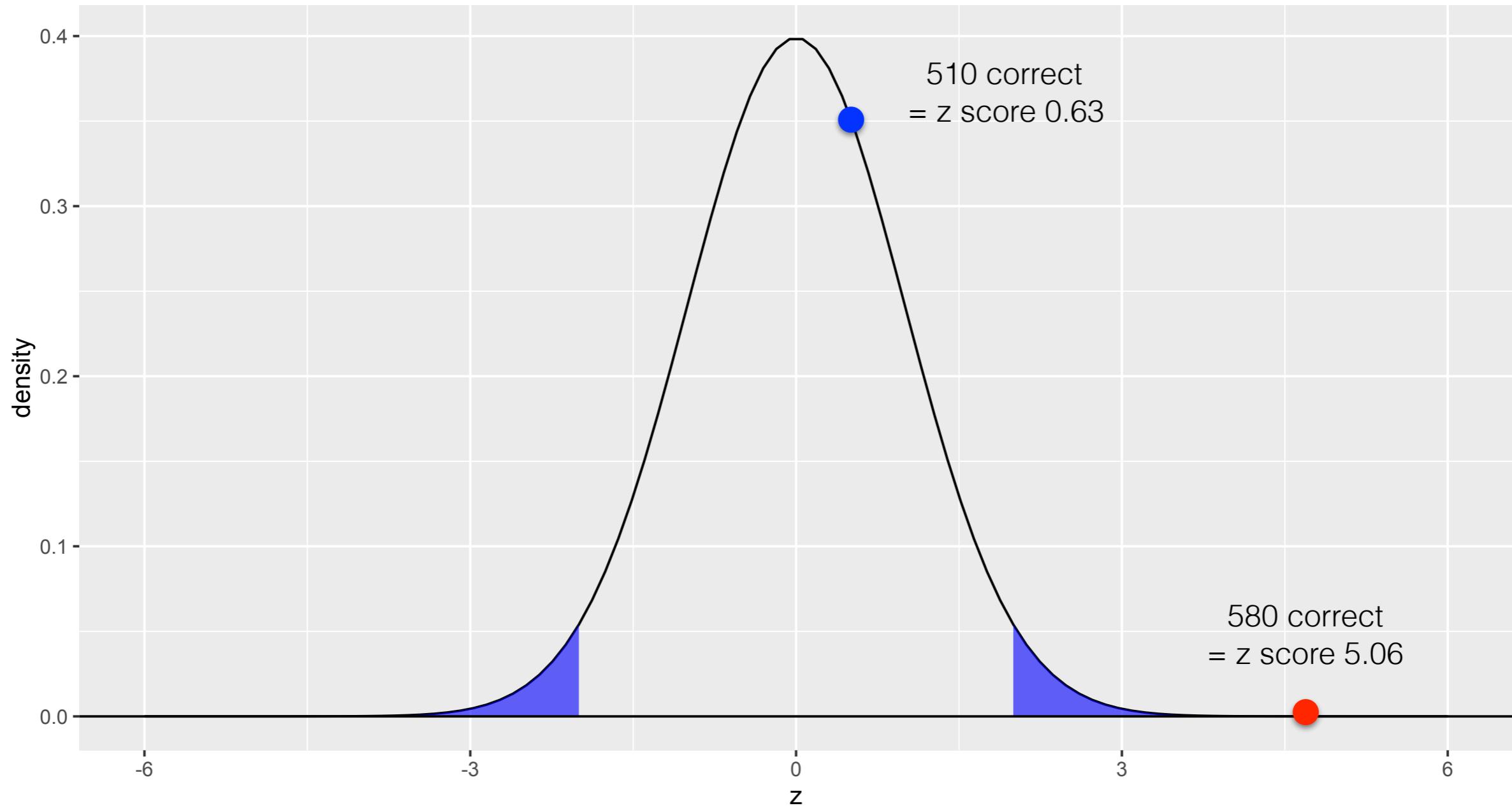
least likely 5%



least likely 1%



Tests

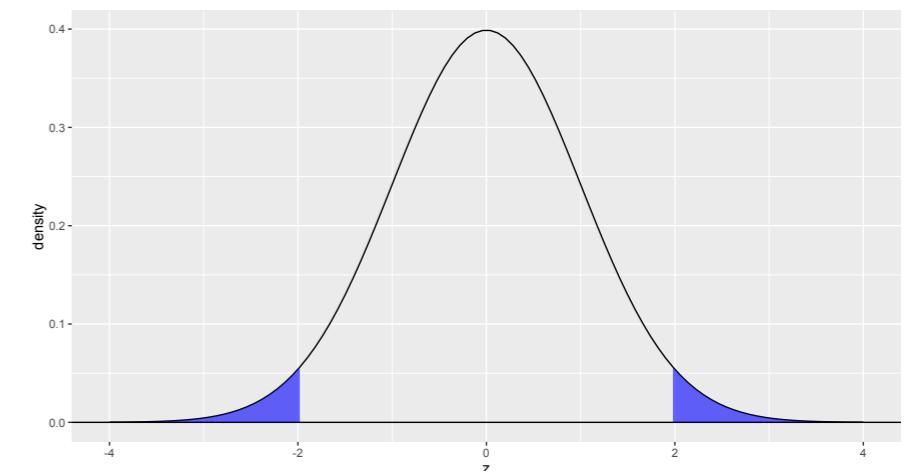


Tests

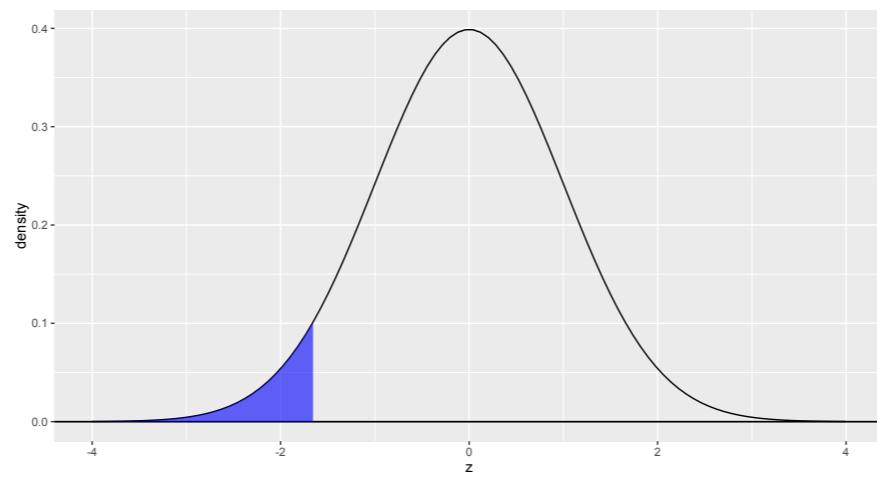
- Decide on the level of significance α . $\{0.05, 0.01\}$
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

Tails

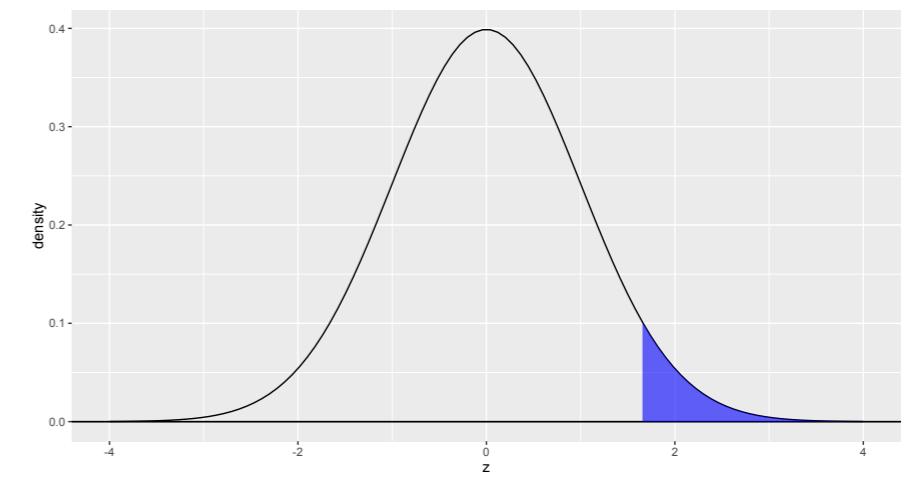
- Two-tailed tests measured whether the observed statistic is **different** (in either direction)
- One-tailed tests measure difference **in a specific direction**
- All differ in where the rejection region is located; $\alpha = 0.05$ for all.



two-tailed test



lower-tailed test



upper-tailed test

p values

A p value is the probability of observing a statistic at least as extreme as the one we did **if the null hypothesis were true.**

- Two-tailed test $p\text{-value}(z) = 2 \times P(Z \leq -|z|)$
- Lower-tailed test $p\text{-value}(z) = P(Z \leq z)$
- Upper-tailed test $p\text{-value}(z) = 1 - P(Z \leq z)$

Errors

		Test results	
		keep null	reject null
Truth	keep null	Type I error α	
	reject null	Type II error β	Power

Errors

- Type I error: we reject the null hypothesis but we shouldn't have.
- Type II error: we don't reject the null, but we should have.

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

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

Confidence intervals

- Even in the absence of specific test, we want to quantify our uncertainty about any metric.
- Confidence intervals specify a range that is likely to contain the (unobserved) population value from a measurement in a sample.

Confidence intervals

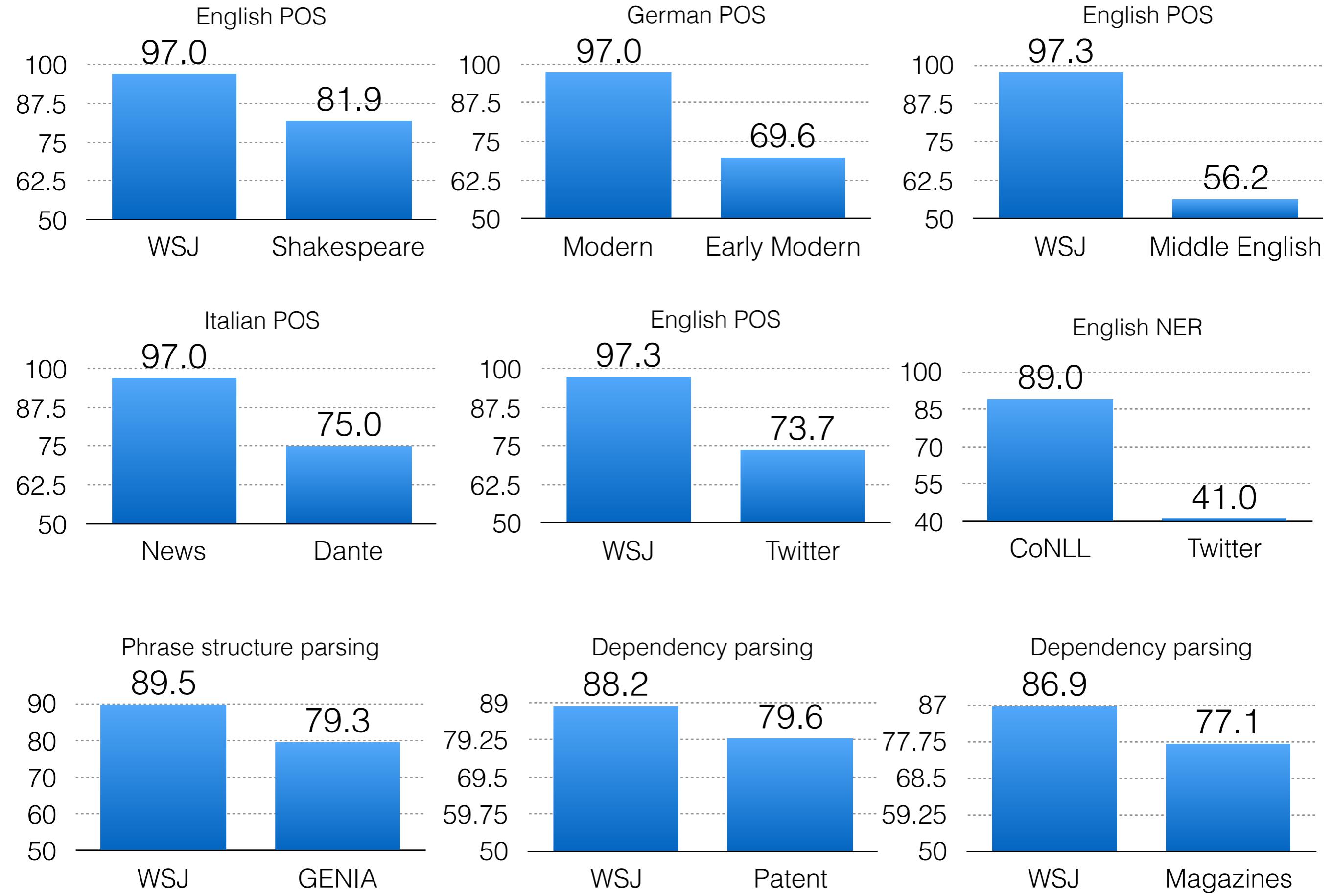
Binomial confidence intervals (again using Normal approximation):

- p = rate of success (e.g., for binary classification, the accuracy).
- n = the sample size (e.g., number of data points in test set).
- z_α = the critical value at significance level α .
 - 95% confidence interval: $\alpha = 0.05$; $z_\alpha = 1.96$
 - 99% confidence interval: $\alpha = 0.01$; $z_\alpha = 2.58$

$$p \pm z_\alpha \sqrt{\frac{p(1 - p)}{n}}$$

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



Takeaways

- At a minimum, always evaluate a method on the domain you're using it on
- When comparing the performance of models, quantify your uncertainty with significant tests/confidence bounds
- Use ablation tests to identify the impact that a feature class has on performance.

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



Google King of United States

Web Maps Images Shopping Videos More Search tools

About 460,000,000 results (0.72 seconds)

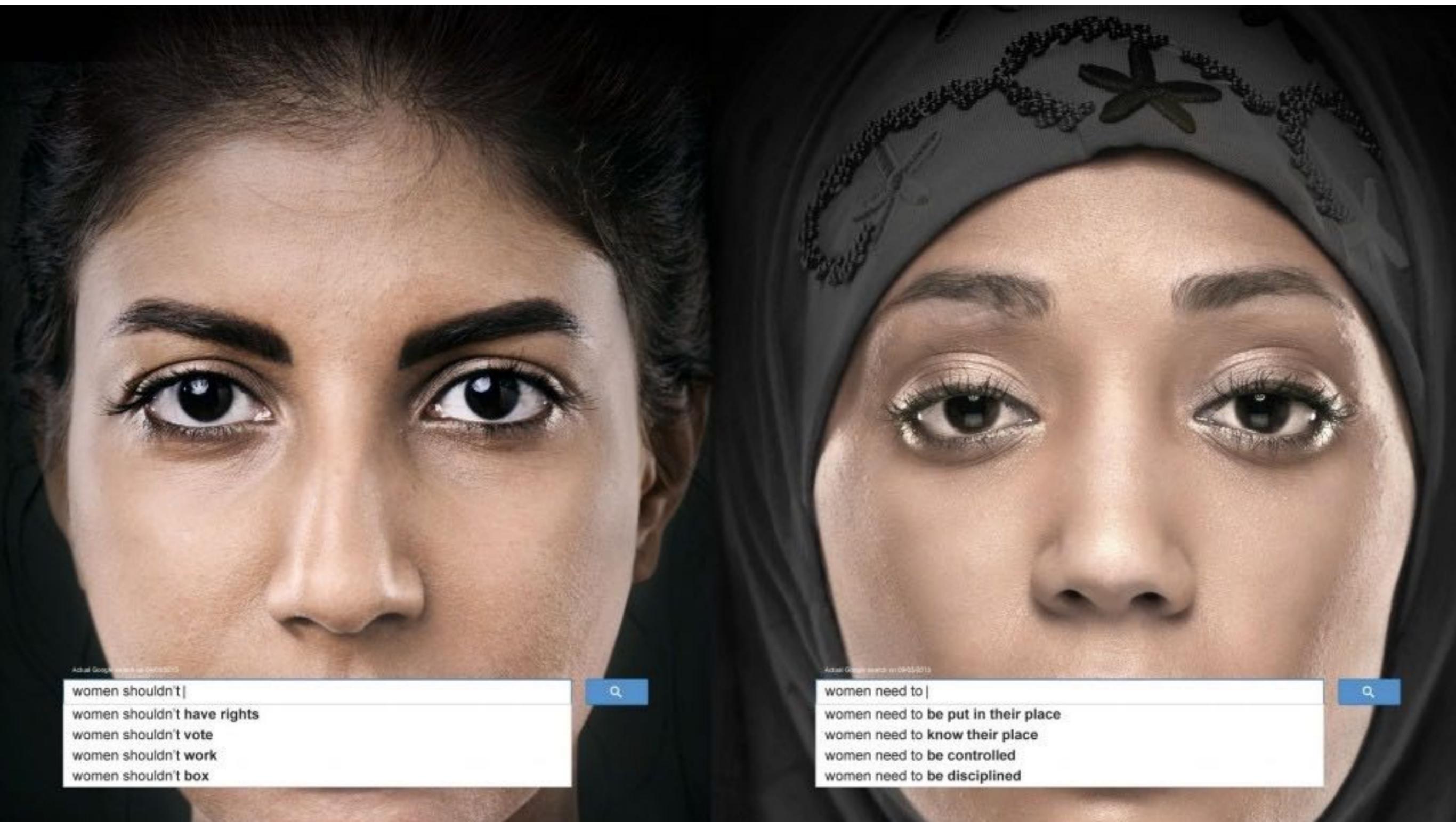
All Hail King **Barack Obama**, Emperor Of The United States Of America!



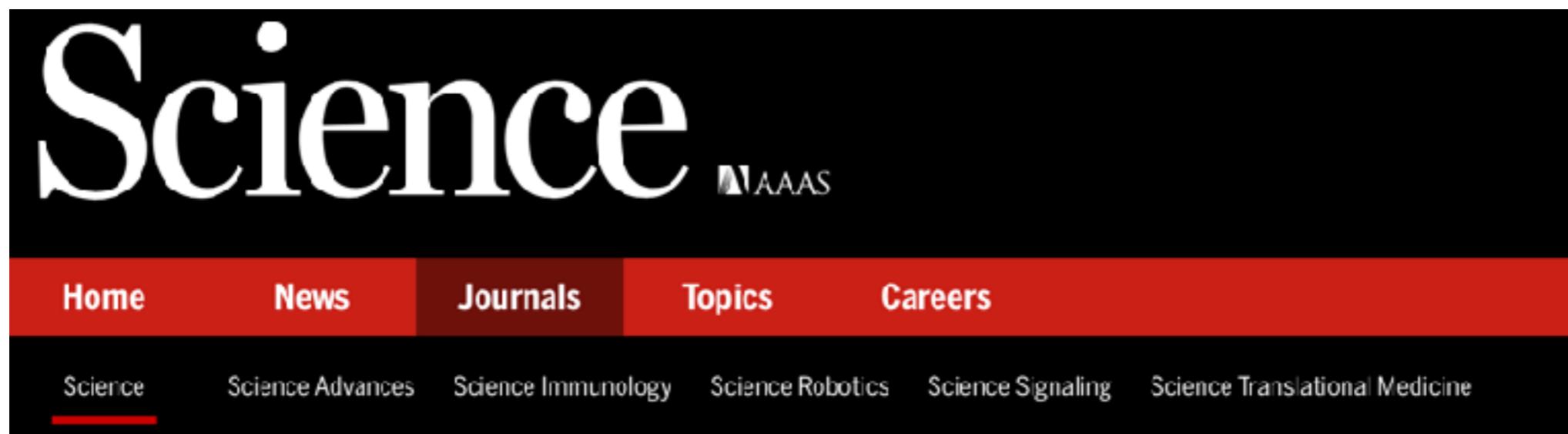
All Hail King Barack Obama, Emperor Of The United States ...
www.breitbart.com/.../All-Hail-King-Barack-Obama-Emperor-Of-... Breitbart

Feedback

Language Modeling



Vector semantics



The image shows the header of the Science journal website. The top half has a black background with the word "Science" in large white serif letters and "AAAS" in smaller white letters to its right. Below this is a red navigation bar with five tabs: "Home", "News", "Journals", "Topics", and "Careers". The "Home" tab is highlighted with a red underline. Underneath the red bar is a black footer bar containing six links: "Science", "Science Advances", "Science Immunology", "Science Robotics", "Science Signaling", and "Science Translational Medicine".

SHARE REPORT



0



Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}



+ See all authors and affiliations

13

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230



Peer Reviewed
← see details

Article

Figures & Data

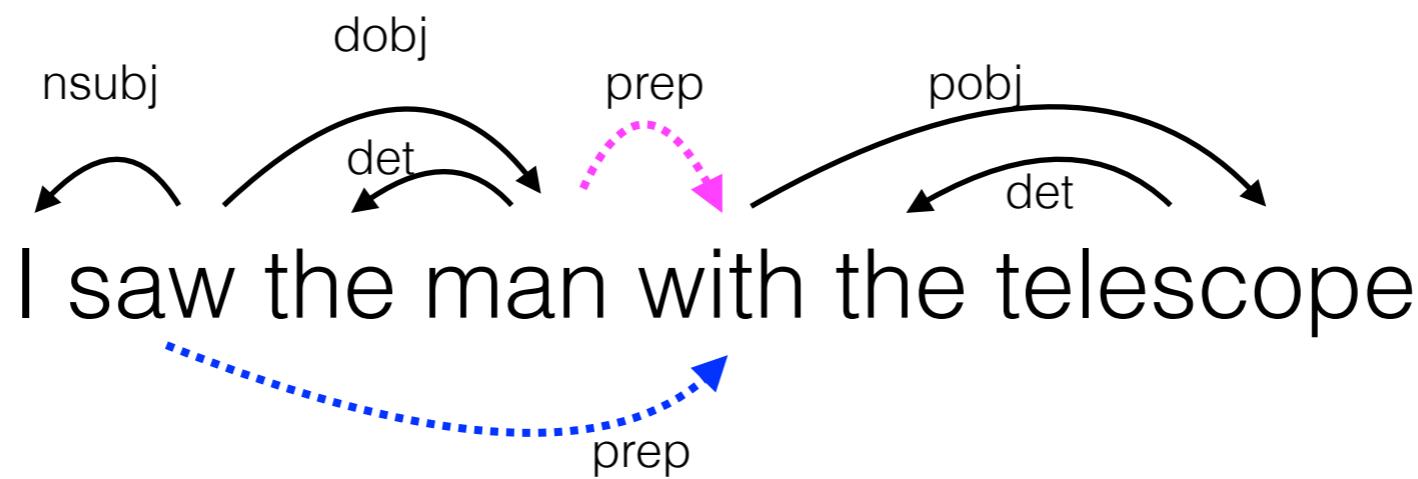
Info & Metrics

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

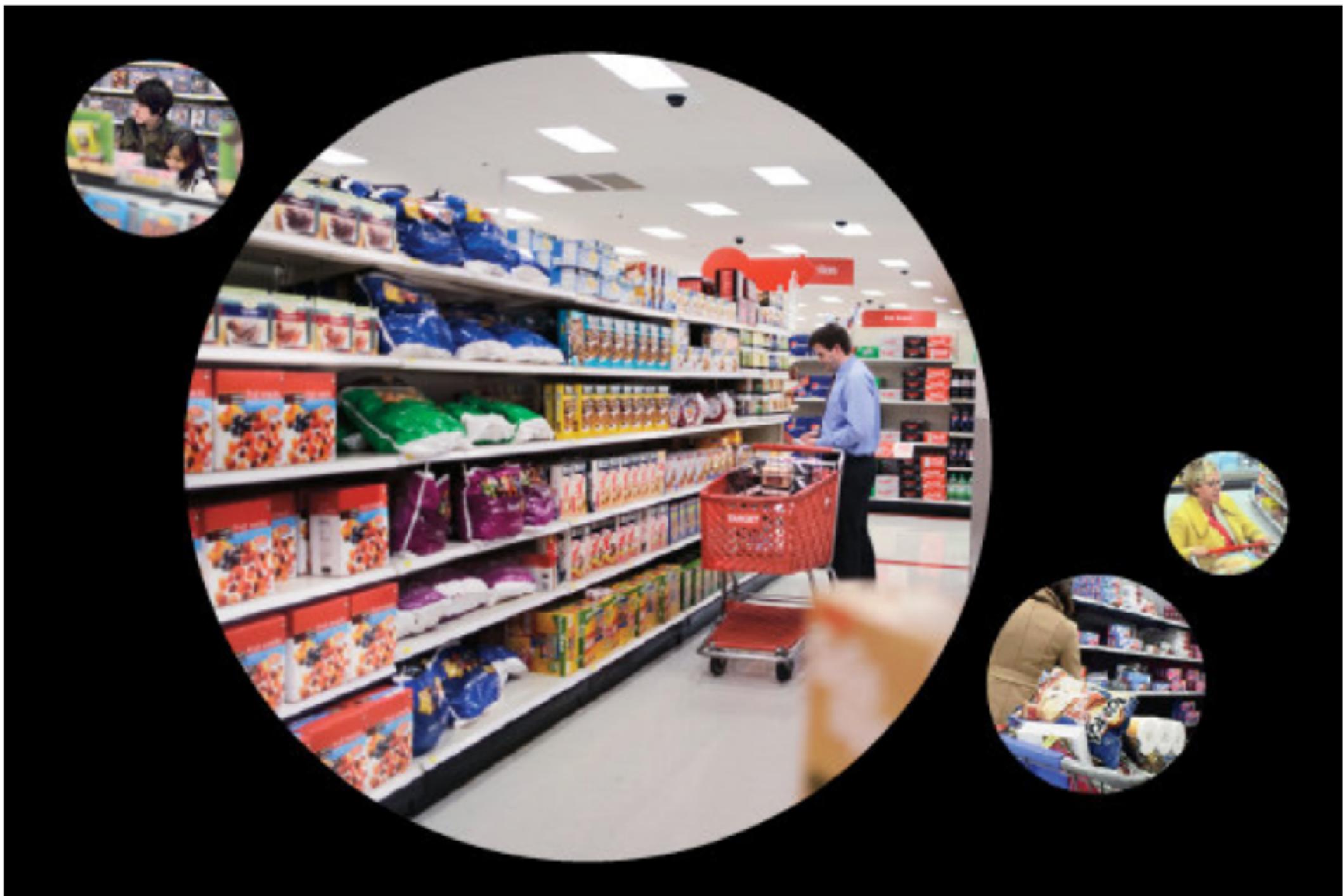


- NLP often operates on text divorced from the context in which it is uttered.
- It's now being used more and more to reason about **human behavior**.

Privacy

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

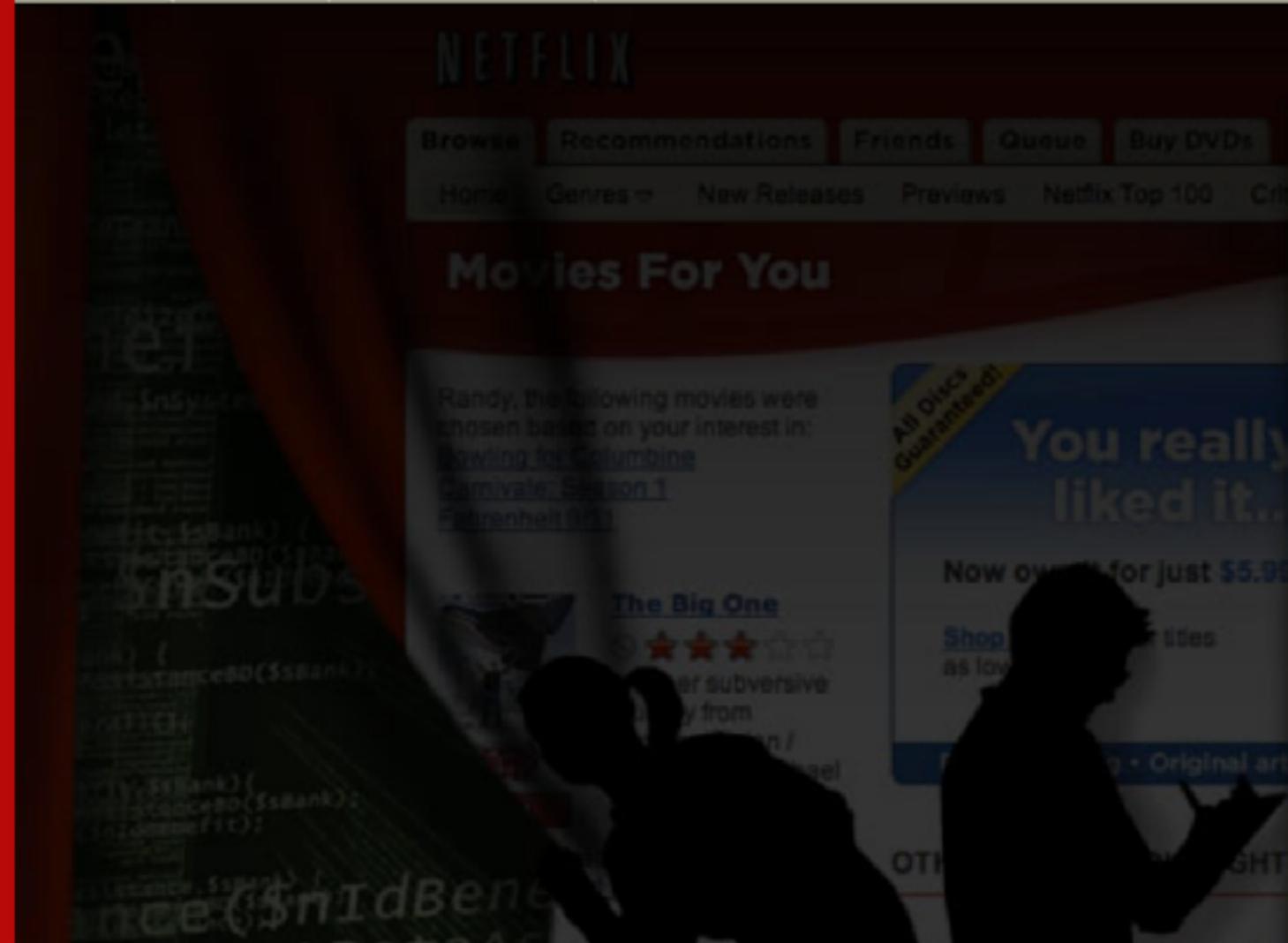




Netflix Prize

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 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