

Ethics in NLP

Google

Web [+ Show options...](#)

[Google again suggests Creed is the **worst band in the world** ...](#)
10 Dec 2008 ... But punch in the term "**worst band in the world**," and you might be surprised with the answer. "See results for: [creed](#)," the Google page ...
[latimesblogs.latimes.com/technology/.../worst-band-in-t.html](#) - [Cached](#) - [Similar](#) - [Print](#) - [Close](#)

[Urban Dictionary: **Worst Band In The World**](#)
Worst Band In The World - 4 definitions - "Creed" according to Google. Although it's fixed now, you used to be able to type in "the **worst band** in t...
[www.urbandictionary.com/define.php?...Worst%20Band%20In%20The%20World](#) - [Cached](#) - [Similar](#) - [Print](#) - [Close](#)

[Urban Dictionary: **worst band**](#)
guy 1: What is "**Worst Band In The World** " guy 2: Naked brothers band ... **Worst the World**. Is an easter egg that Google has in there search engine. ...
[www.urbandictionary.com/define.php?term=worst+band](#) - [Cached](#) - [Similar](#) - [Print](#) - [Close](#)

[+ Show more results from \[www.urbandictionary.com\]\(#\)](#)

See results for: [creed](#)

Google

About 1 results (0.01 seconds)

[Google won't search for \[Chuck Norris\]\(#\) because it knows you don't find \[Chuck Norris\]\(#\), he finds you.](#)

Your search - [Chuck Norris](#) - did not match any documents.

Suggestions:

- Run, before he finds you.
- Try a different person.
- Try someone less dangerous



EECS 498 (April 1, 2020)
(Remote) Lecture 22

Reminders

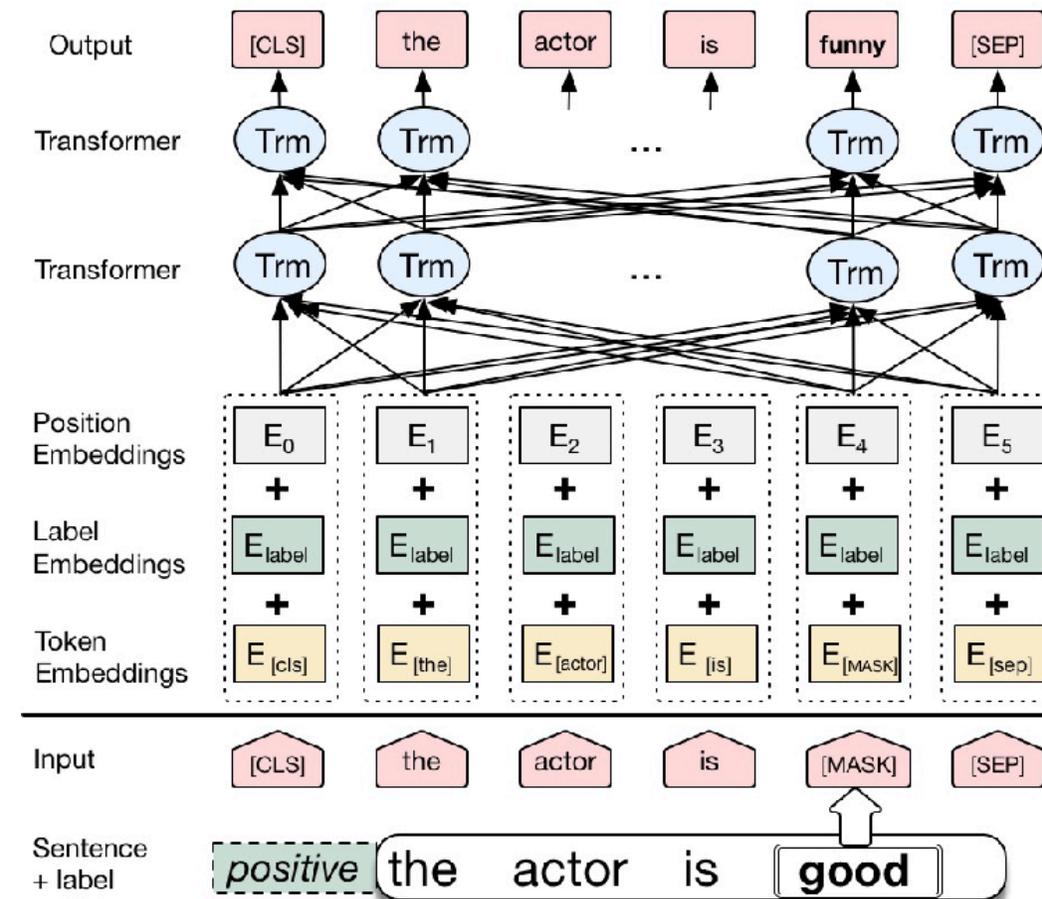
- Due next Monday, 4/6
 - PC5 (Cooperative Testing) due 4/6
 - PC6 (Sprint Review 3) due 4/6, delivered as YouTube video
 - Please also upload your raw video to Google Drive (so others can download)
 - SR3: Please review group feedback
- No lectures next week
 - Instead, you will use the time to review each group's SR3 video presentation
- PC7 (Final Presentations) will be a **scheduled telecon** with your team
 - Schedule a 30 minute block here:
<https://calendar.google.com/calendar/selfsched?sstoken=UVBaMkN5bk9KeIVRfGRIZmF1bHR8Mjk4MTIINjJjODMyODdkODk3MzU4YjNmNWlxZDUyNTI>
 - Try to have most/all your team members present for that

Diversity, Equity, and Inclusivity Town Hall

- CSE faculty holding a remote town hall for undergraduate students
- Date TBA, but **pay attention to your email** if interested
- This is an opportunity to ask questions of faculty concerning recent department- and university-wide climate issues
 - (especially relevant for this class)
- Previous turnout: grad student town hall had 60 in attendance remotely, including 5 faculty, the chair, and a dean
 - (the department is taking it seriously)

Review: Bert

- Bidirectional Encoder Representation with Transformers
- Bidirectional
 - Language model encompasses context from both left-to-right and right-to-left
- Encoder
 - NN structure that yields an embedding
- Representation
 - The embedding “represents” a token
- Transformer
 - Neural architecture that encompasses self-attention to learn contextual embeddings



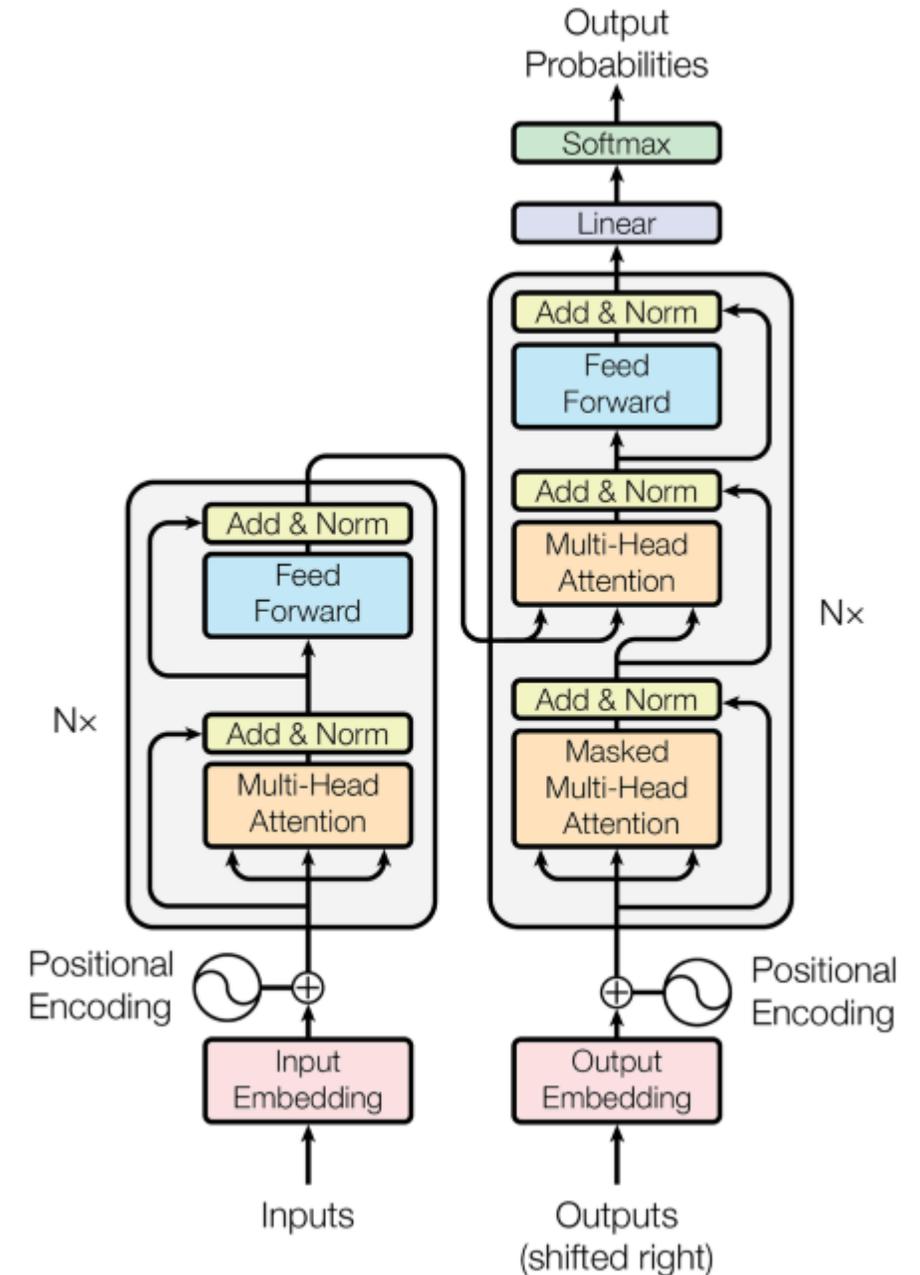
Review: Contextual Embeddings

- The glue *sticks* to the paper
- The windstorm left many *sticks* in the yard.
- The context affects how we interpret the word “sticks”
 - **Conclusion:** the embeddings for a word depend not only on the word itself, but also the context in which it appears
 - “sticks” could thus occupy more than one point in the embedding space
- Bert creates such embeddings



Review: Transformers

- Transformers are a **neural architecture** that use **encoders** with **self-attention** to incorporate the **importance** of other **tokens** in an utterance into the **embeddings** of a given token
- TL;DR Attention is a NN technique that learns relative importance of other words when predicting another



Review: Transformers in Machine Translation

- Transformers are very good at translating from one language to another
- Step 1: **Encode** an input utterance into the embedding space
- Step 2: **Decode** each word from the embedding space to the target language vocabulary

"The food is tasty."		Embeddings...	这个饭好吃	
The	1		这	1
food	2		个	2
is	3		饭	3
tasty	4		好	4
			吃	5

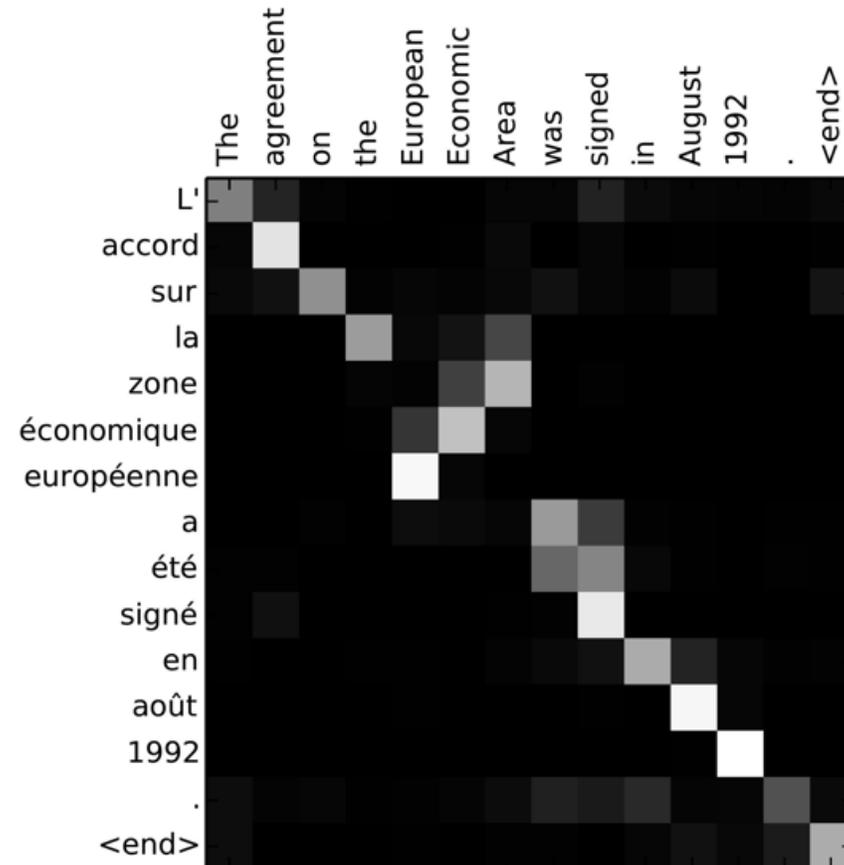
Review: Attention in Machine Translation

- Input: [1, 2, 3, 4]
 - Time step 1: Transformer sees token “1” as input
 - “1” gets encoded to some contextual embedding vector, v_1
 - v_1 is fed through a decoder to find the corresponding (Chinese) output
 - v_1 gets decoded according to the attention learned from context
 - **Example:** “The” would give high attention to both “这” and “个”
 - The model is more likely to predict those words (and not pay attention to later words like 好 or 吃)

“The food is tasty.”		Embeddings...	这个饭好吃	
The	1		这	1
food	2		个	2
is	3		饭	3
tasty	4		好	4
			吃	5

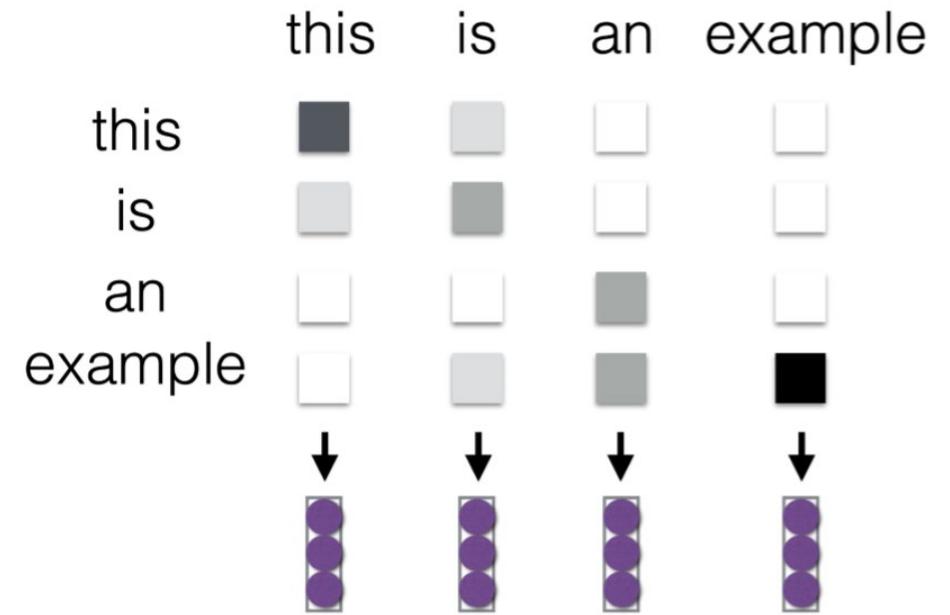
Review: Attention in Machine Translation

- Note attention is given to words in reverse order for some languages



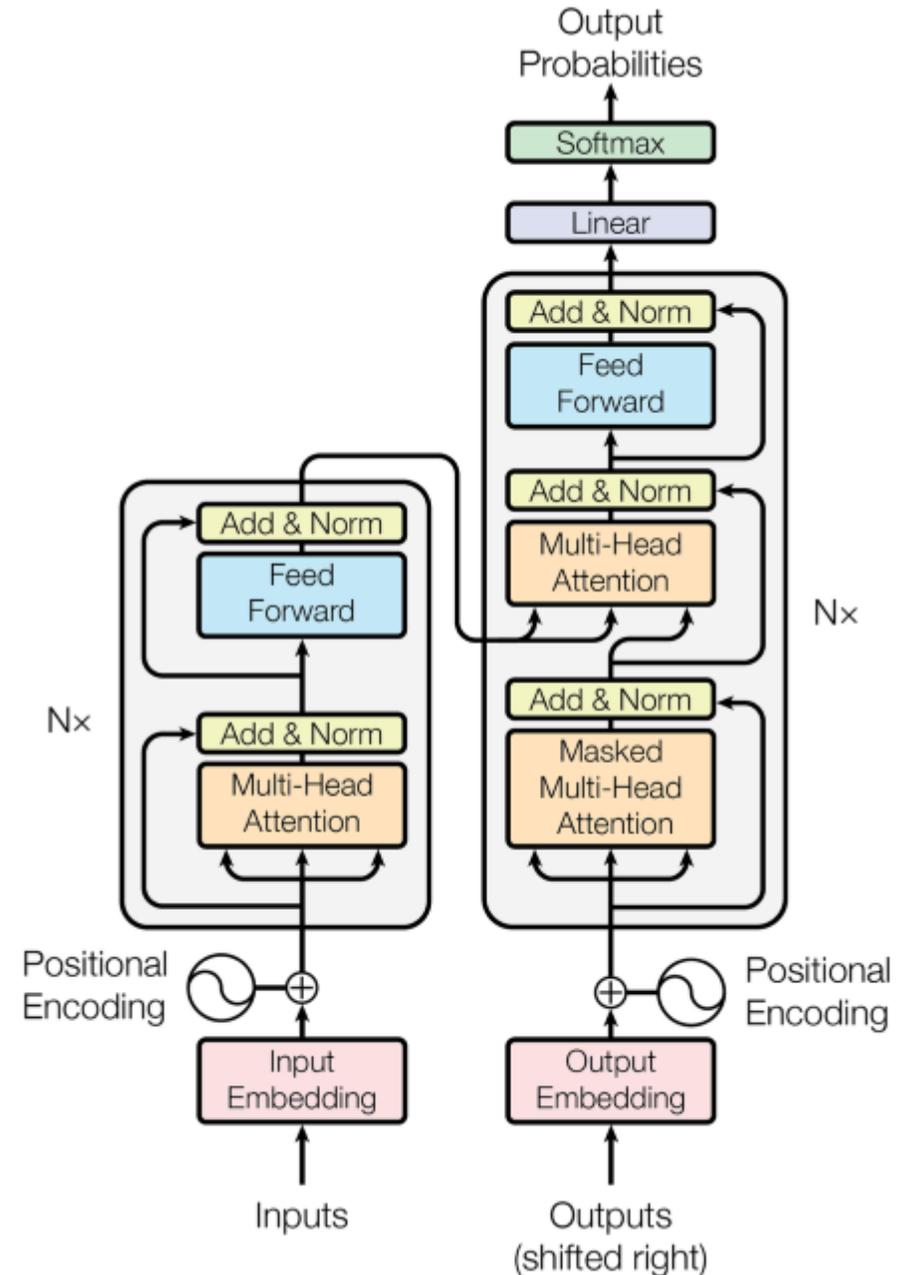
Review: Attention in Bert

- Transformers use **self-attention**
- Rather than translating from one language to another, Bert Transformers learn relative importance of each word in the context of others
- Given **surrounding context**, Bert can thus **predict** a reasonable word to fit in an unknown space
 - By having learned the most important parts of the utterance to look at



Review: Transformers

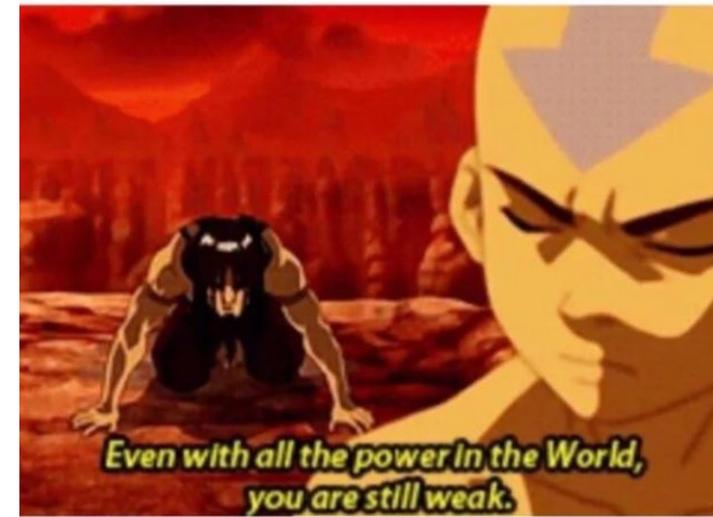
- Transformers help predict word-at-a-time **while accounting for context**
- During **training**, each word in the input utterance learns
 - Position information of both **input** and **output** tokens
 - Position and value information of **surrounding** words in both the **input** and **output** utterances
- **Summary:** word-at-a-time predictions possible because the predictions can account for context



Review: Bert Summary

- We have discussed Bert as a mechanism for acquiring robust contextual embeddings
- In practice, Bert can do a lot more
 - The word embeddings were more of a nice “side effect” of the architecture
 - **Sentence Prediction**
 - Given one sequence of words, predict the next sequence
 - **Question-answering**
 - Learns relationships between question sequence inputs and answer sequence outputs
- Bert is unwieldy
 - 11GB of VRAM to run?

When your little brother has an RTX 2080 GPU but doesn't know about Deep Learning.



Bert: Semantic Entailment

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT _{BASE}	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	91.1	94.9	60.5	86.5	89.3	70.1	81.9

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.

Hypothesis: Jainism hates nature.

Label: Contradiction

CoLa

Sentence: The wagon rumbled down the road.

Label: Acceptable

Sentence: The car honked down the road.

Label: Unacceptable

Bert: Logical Analysis

A girl is going across a set of monkey bars. She

(i) jumps up across the monkey bars.

(ii) struggles onto the bars to grab her head.

(iii) gets to the end and stands on a wooden plank.

(iv) jumps up and does a back flip.

- Run each Premise + Ending through BERT.
- Produce logit for each pair on token 0 ([CLS])

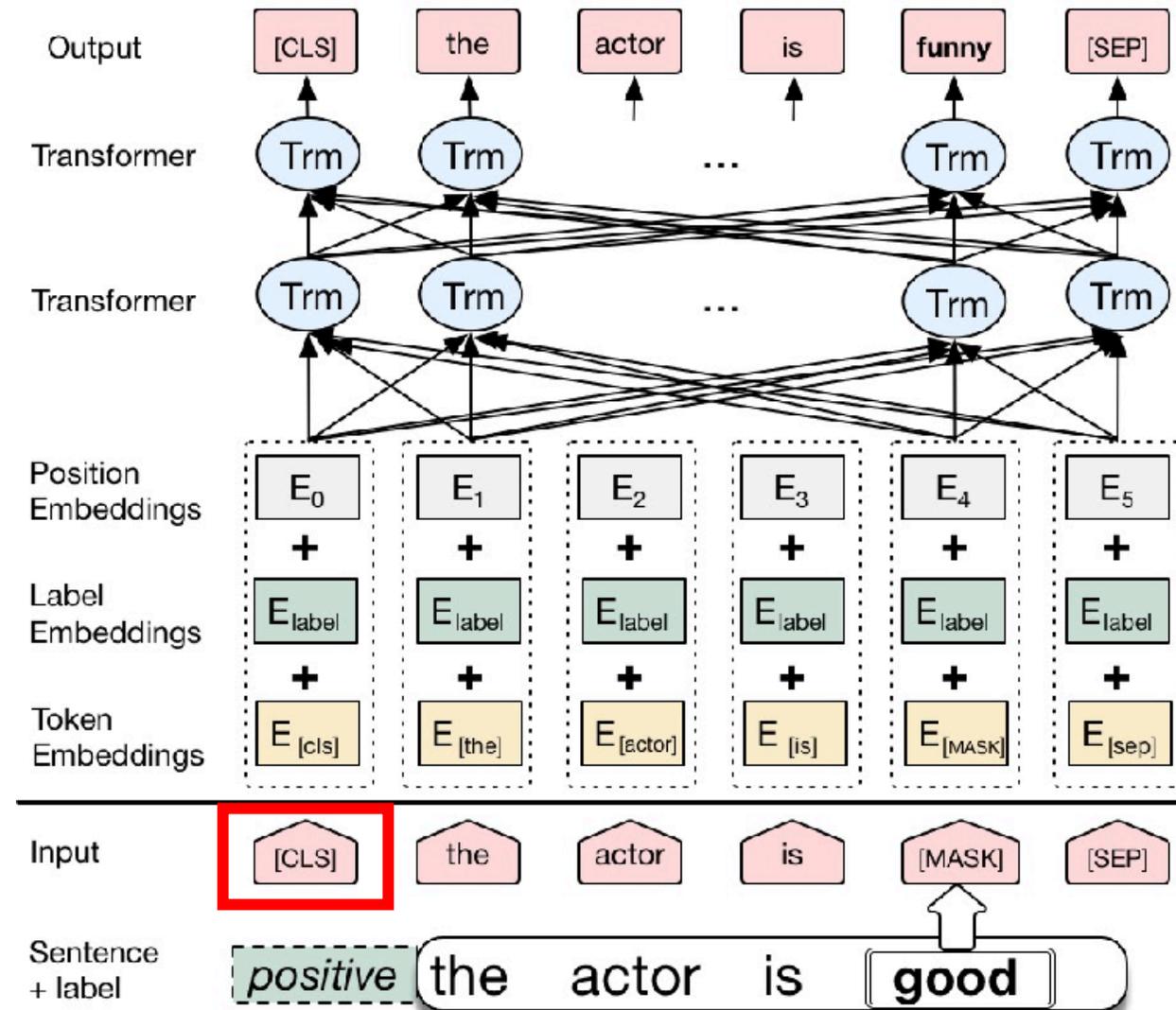
Leaderboard

— Human Performance (88.00%)
— Running Best
◆ Submissions

Rank	Model	Test Score
1	BERT (Bidirectional Encoder Representations from Transfo... <i>Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova</i> 10/11/2018	86.28%
2	OpenAI Transformer Language Model <i>Original work by Alec Radford, Karthik Narasimhan, Tim Salimans, ...</i> 10/11/2018	77.97%
3	ESIM with ELMo <i>Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin</i> 08/30/2018	59.06%
4	ESIM with Glove <i>Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin</i> 08/29/2018	52.45%

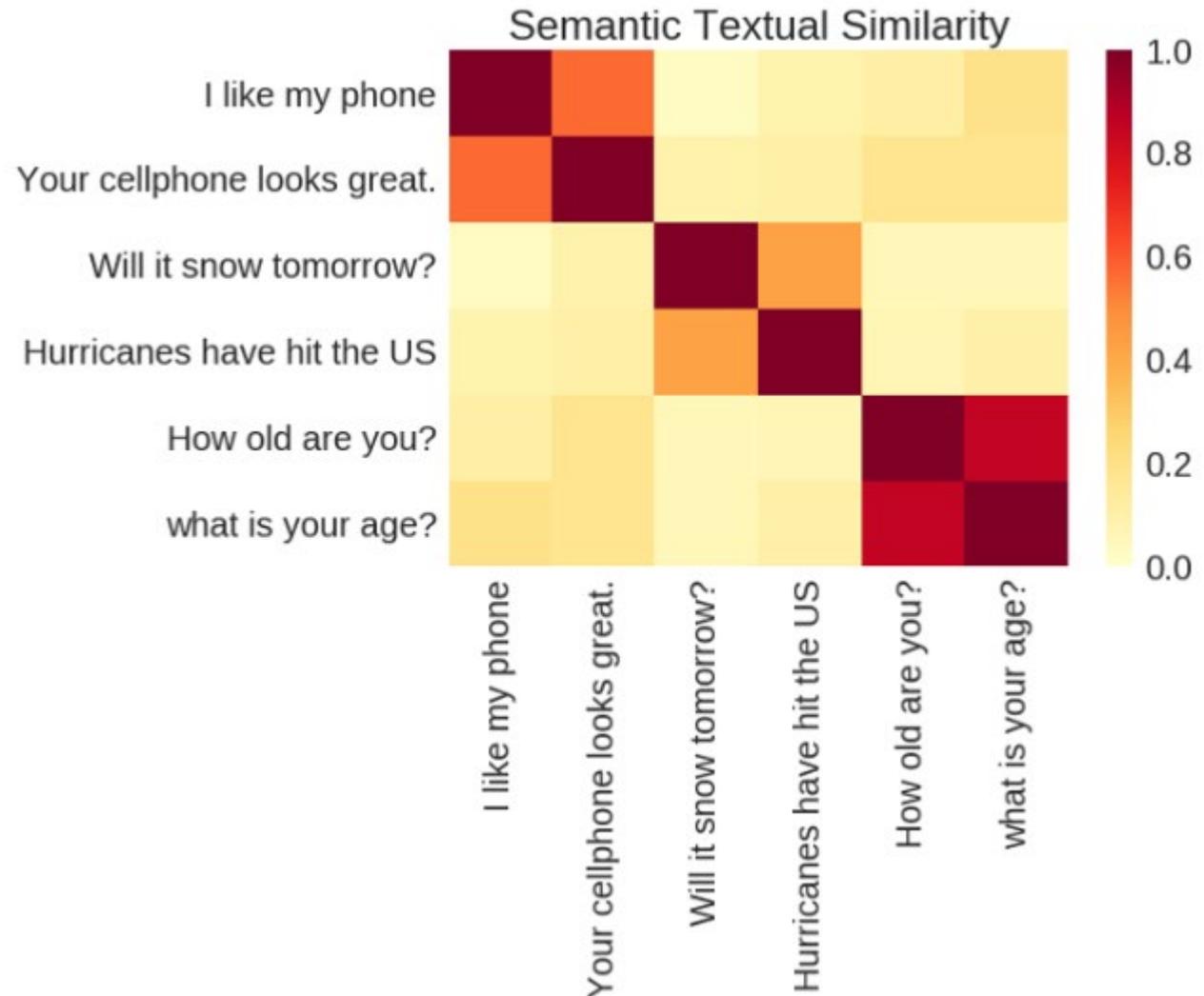
Bert: Sentence Embeddings

- The [CLS] token is meant to represent the start of a sentence
 - **Consider:** The model supposedly learns context in part from position
 - Every sentence “starts” with [CLS]
- No matter what sentence is given, [CLS] always involves context learned from every other word
 - Thus, the embeddings for [CLS] are a **rich representation** of the **whole sentence**



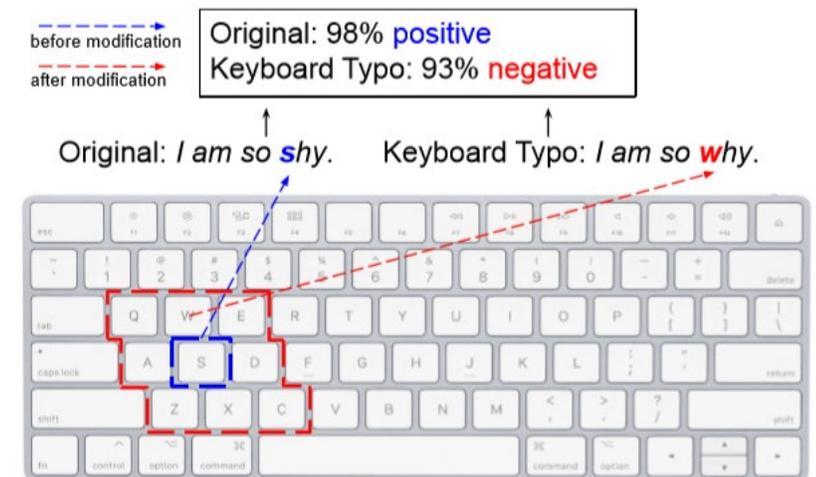
Review: Sentence Embeddings in General

- Embed sentences into vector space
- Useful for comparing sentences semantically
- Word embeddings are used in addition to positional information



Review: Bert Shortcomings

- Bert's language modeling assumes **independence** among MASK tokens
 - **Recall:** Bert operates by MASKing some tokens, forcing the embeddings to reflect **context**
 - **Problem:** if multiple MASK tokens appear in a sentence, their ordering and relationship are assumed irrelevant by BERT
 - "I have to fly from MASK₁ to MASK₂" <- wouldn't make sense if the MASKed tokens were "Ithaca" and "Syracuse"
- Bert's input leverages WordPiece
 - **Problem:** Limited robustness against misspellings



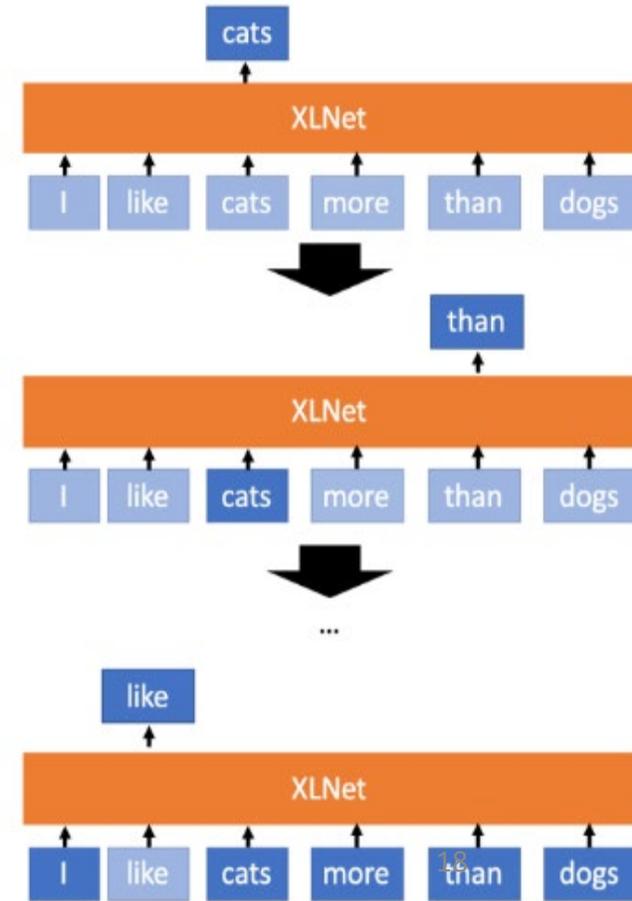
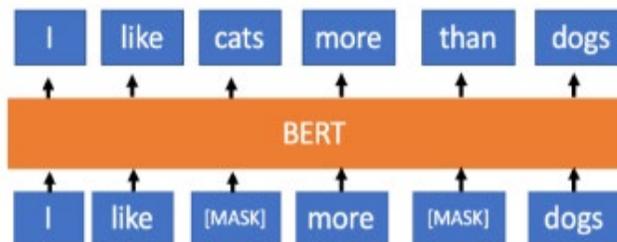
XLNet: Even more state-of-the-art?

- Eliminate independence assumption with “*Permutation Language Modeling*”
 - Basically, consider predictions of multiple permutations of words in a sequence
- Even more complex
 - The model learns multiple ways to predict each sequence given different parts of the context

above. Specifically, we train on 512 TPU v3 chips for 500K steps with an Adam weight decay optimizer, linear learning rate decay, and a batch size of 8192, which takes about 5.5 days. It was

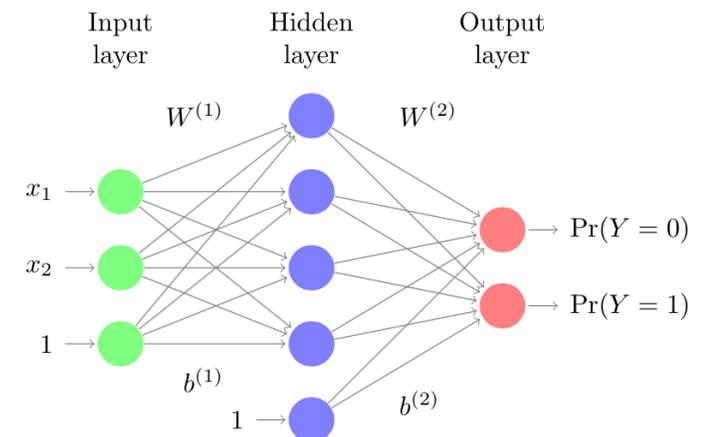
(that’s \$160k to train)

(in contrast, Bert used 64 TPUs for 4 days for a “mere” \$14k)



Multiple Models in NLU Pipeline

- **Intent Classification** is often performed with SVM or FastText models
 - **Use Multi-class Support Vector Machine** to decide amongst multiple intents
 - tf-idf, n-gram frequency, embeddings, all potential features for SVM
 - Binary classifier for every pair of intents
 - “is account_balance” vs. “is open_credit”, etc.
 - Simple vote: increase an intent’s class count by 1 each time it wins one of the binary classifiers; take the highest as the intent label
 - **Advantage:** SVM is fast to train and infer; accuracy > 90% on standard workloads
 - **FastText** used to classify sequences into “topics”
 - Just create “topics” to be intents
 - Model takes sequences of words as inputs, embeds them, trained to select among multiple classes
 - **Advantage:** Fast (with pretrained embeddings); more accurate
 - Robust against misspellings
 - Words embedded in 3-character sequences:
Kevin becomes: <Ke, Kev, evi, vin, in>



Stateful Classification

- Cline
 - Each state associated with a separate intent classifier
 - Create an SVM/FastText model with each outgoing edge as a possible intent class
 - **Advantage:** State makes it easier to discern between intents
 - There are typically fewer intent classes to choose from in a given state
- DialogFlow
 - Coerce model outputs using Contexts
 - Classification model probabilities are changed based on Context
 - e.g., “2x more likely” to choose intent A over B in context C
 - **Advantage:** Reduced overall training (there’s no per-state classifier to train)
- Rasa
 - Touted as “stateless”
 - You give it training data that captures state (e.g., which intents should come next)
 - **Advantage:** Purely example based. Rasa scales well as a result

Slot Extraction

- Can be thought of as a Sequence-to-Sequence task
 - Turn an utterance into an IOB representation
 - Yo fam get me a burger.
 - O B:person O B:person O B:food
 - Embed words, train a model to learn how to predict I, O, or B for each token
- Clinc
 - Per-competency slot extraction
 - Currently using Glove embeddings (olde but fast)
 - Bert embeddings improve accuracy (but radically increase training time)
- DialogFlow and Rasa
 - Appears per-intent, although model details are not immediately obvious

One-Slide Summary: Ethics in NLP

- **Humans** engage with Conversational AI systems **regularly**
 - What societal impact might these systems have?
 - Personal information exchange (bank accounts)
 - Critical decision making (medical records)
- NLP in general is **increasingly** used in ways that **affect human lives**
 - **Reasoning systems** that deliver targeted advertisements or recommendations
 - Employment, citizenship, parole, and credit **decisions**
- **Critical biases** exist in NLP systems as a result of data
 - What techniques adjust for systemic bias in data?

Full Disclosure: Ethics in NLP

- Slides derived from UMass Amherst grad-level NLP course
- Some topics may be uncomfortable (e.g., racism and sexism in data)
- **Critically:** as an MDE elective, we want you to think about societal impact had by the technologies you work on

Guiding Principles

The common misconception is that **language** has to do with **words** and what they **mean**.

It doesn't.

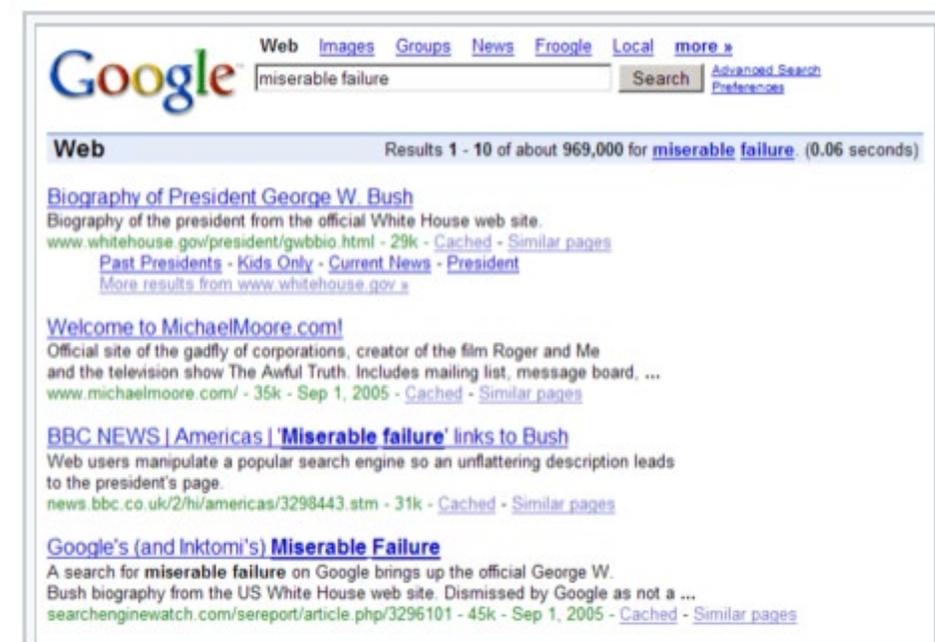
It has to do with **people** and what *they* **mean**.



Dan Jurafsky's keynote talks at CVPR'17 and EMNLP'17

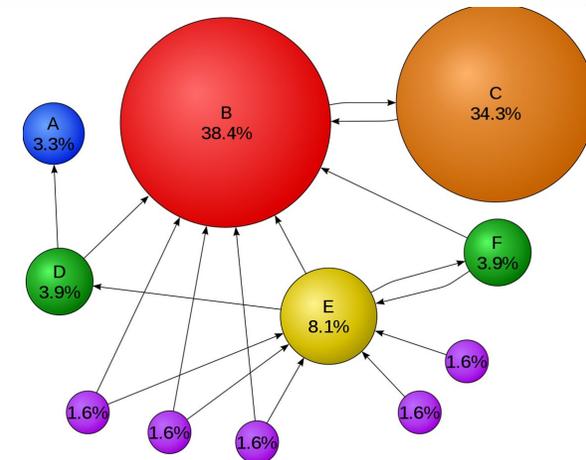
Motivation: Google Bombing

- Google's **PageRank** algorithm delivers search results based on incoming links to a given page that relate to a "topic" (like an utterance or search term)
 - Here, **intentional bias** is possible by creating **fake links** and traffic to **influence** a topic's **association** with a page
- (not shown here) similarly sexist and racist remarks occur
- **Problem:** Model works fine, it's just the data, right?



The screenshot shows a Google search interface with the query "miserable failure". The search results are displayed under the heading "Web" and show "Results 1 - 10 of about 969,000 for miserable failure. (0.06 seconds)". The first result is "Biography of President George W. Bush" from the official White House web site. The second result is "Welcome to MichaelMoore.com!". The third result is "BBC NEWS | Americas | 'Miserable failure' links to Bush". The fourth result is "Google's (and Inktomi's) Miserable Failure".

An example of Google bombing in 2006 that caused the search query "miserable failure" to be associated with **George W. Bush** and **Michael Moore**.



A Dilemma

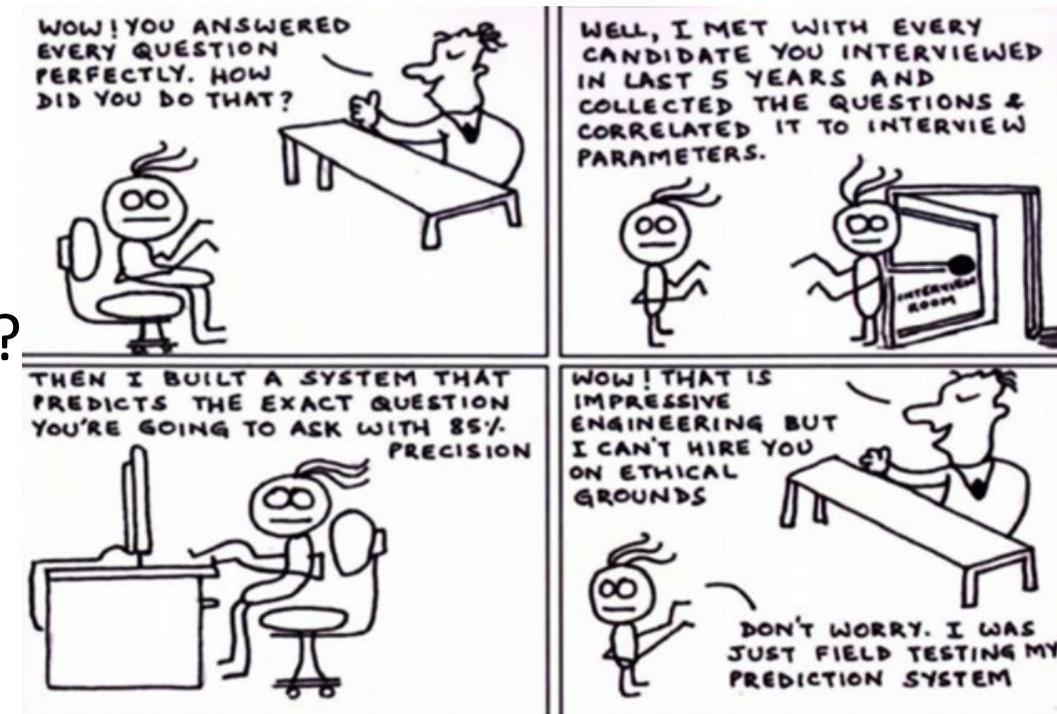
- We leverage ML **models** that are increasingly **unexplainable** in an era of **explosive data generation**

Two potential extremes

- (1) We do nothing, because it's a data problem
 - Are we censoring “the people” if we intervene?
- (2) We intervene, adjusting models to account for data
 - How do we detect biases?
 - What does it mean to unbiased data?

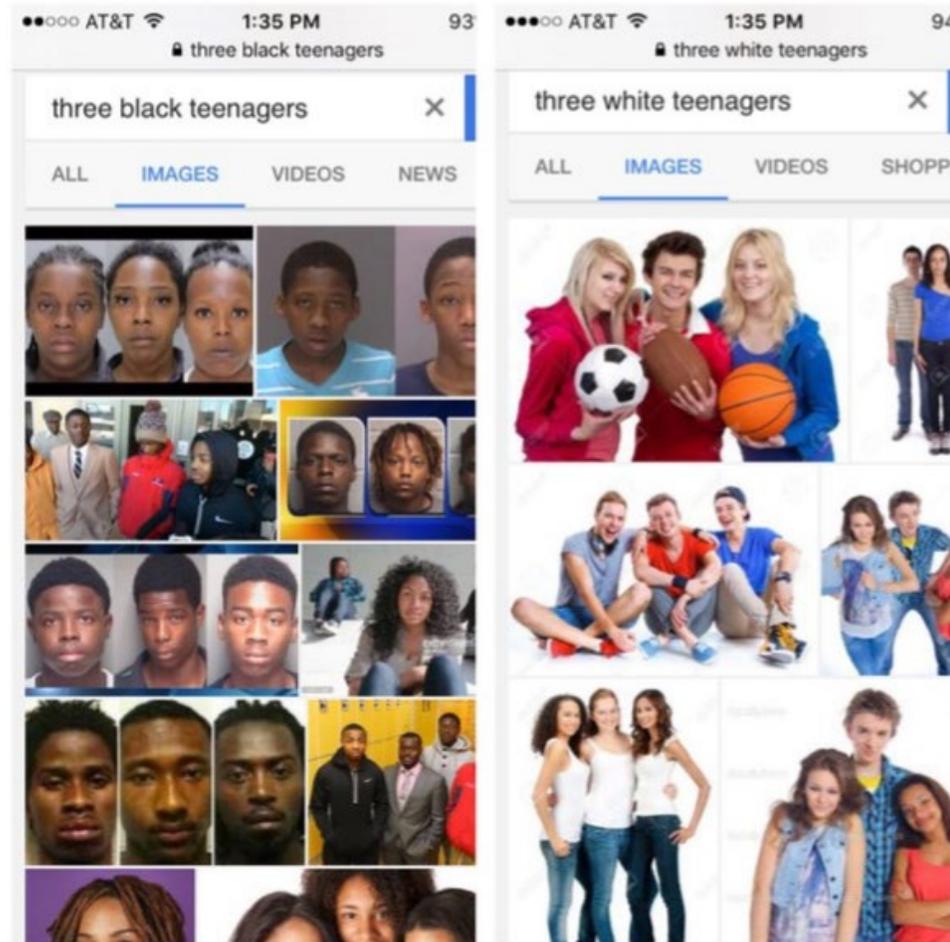
Adversarial Thinking in AI Systems

- (see also EECS 388) What are the risks and benefits of deploying a system?
 - Who benefits from a technology?
 - Who could be harmed by a technology?
 - **Consider:** Medical informatics. Diagnostic technologies? Medical malpractice?
- How representative is your training data?
 - Could sharing the data impact lives?
 - **Consider:** employment information
- Does the system optimize for the right thing?
 - **Consider:** are we only maximizing F_1 score?
- Could wrong predictions impact lives?
 - **Consider:** diagnostics?



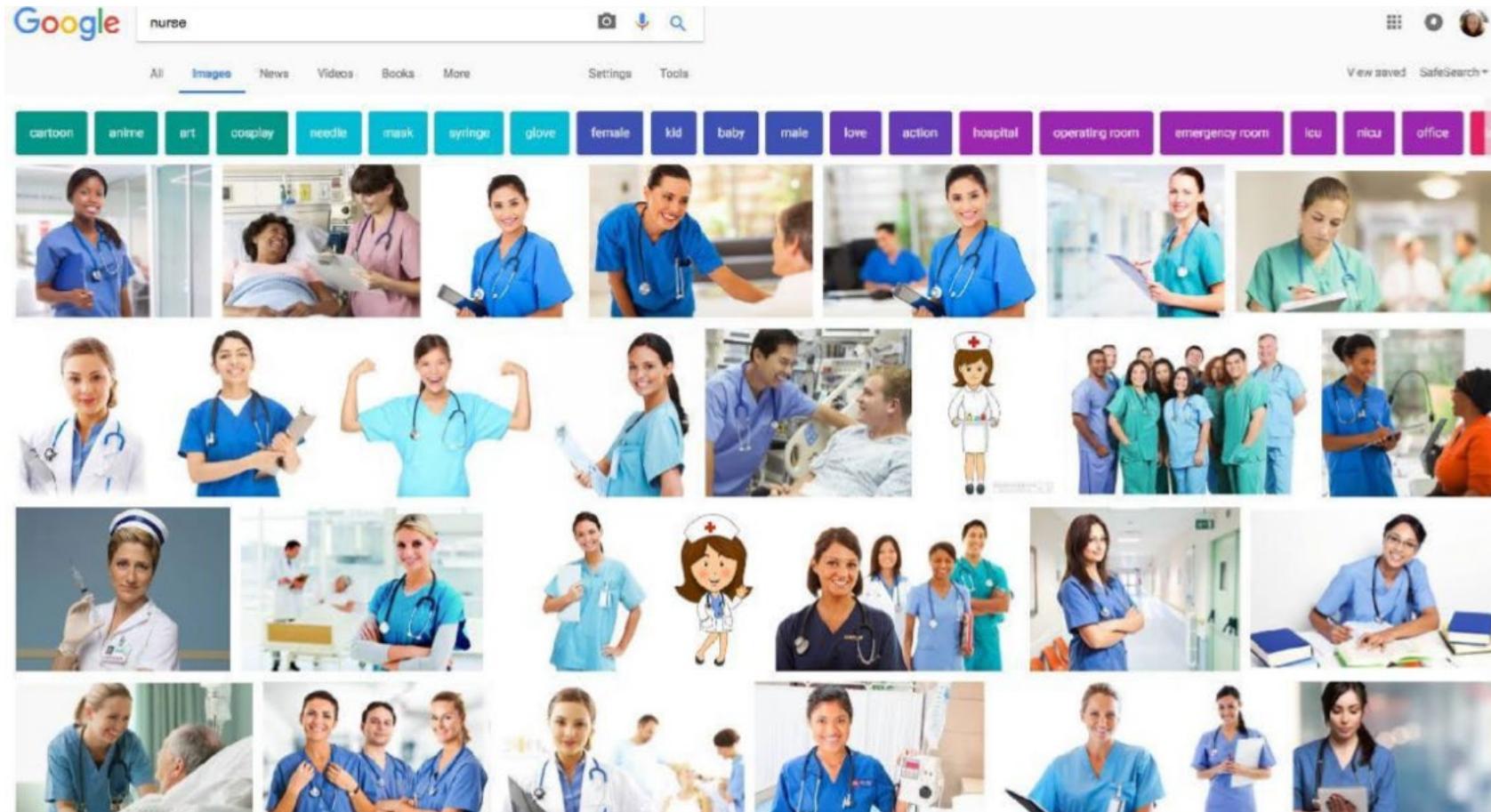
Impact of Social Stereotypes on Data

- 2016 Google queries: racial stereotypes



Impact of Social Stereotypes on Data

- Google query for “nurse”



Impact of Social Stereotypes on Data

- Google query for “homemaker”

The image shows a Google search interface for the query "homemaker". The search bar is at the top left, and the results are displayed in a grid below. The filters at the top include: woman, wife, cartoon, male, traditional, mother, clipart, breadwinner, modern, old fashioned, happy, busy, india, black, african american, vector. The search results include:

- Homemaker and Why She is So Important ... momscove.com
- A Homemaker's Presentation - YouTube youtube.com
- Traditional Catholic ... tradcatfem.com
- Retro housewife ... pinterest.com
- Homemaker mom Challenges faced o... vgo.watch
- What Makes Me A Homemaker - My Little... lukascondie.com
- The Number One Enemy to Homemakers ... youtube.com
- Housewife and Homemaker ... heroes.com
- Homemakers Are Facing a Retirement ... nextavenue.org
- 1950s Homemaker Secrets-How You Can ... hubpages.com
- Give Me My Due - As A Homemaker, I ... womensweb.in
- Housewife or Homema... clarissarwest.com
- June Cleaver and Susie Homema... southadadnewsleader.com
- Who is Suzy Homemaker? See the vintage ... clickamericana.com
- Basic and Essential Kitchen Tools For ... olgasflavorfactory.com
- Suzie Homemaker cftstinks.com
- NATIONAL HOMEMAKER DAY - Novembe... nationaltoday.com
- Homemaker Services - www ... tyshealthyhealers.com
- How to be a Productive Homemaker
- struggle with mental health ...
- The role of a homemaker
- Homemaker In The Age Of Third W...
- woman cute cleaning cartoon Vect...
- On C.S. Lewis and being a 'homema...
- Related searches: woman homemaker, male homemaker, indian homemaker
- HomeMaker, Premium Squee...
- Homemakers, Have More Confidence In ...
- 6 rules for homemakers

Impact of Social Stereotypes on Data

- Google query for “CEO”

The image shows a Google search interface for the query "ceo". The search bar is at the top left, with the Google logo and navigation icons. Below the search bar are tabs for "All", "News", "Images", "Books", "Videos", and "More". To the right of the search bar are "Settings" and "Tools". Below the search bar is a horizontal row of filters, each with a small circular icon and a label: "business", "google", "cartoon", "snapchat", "microsoft", "apple", "woman", "desk", "amazon", "uber", "pepsi", "youtube", "black", "facebook", "starbucks", "successful". The main content area displays a grid of search results, each with a thumbnail image and a caption. The results include: "Chief executive officer - Wikipedia", "Boeing CEO pushed out amid 73...", "What do CEOs do? A CEO Job Description ...", "Marriott CEO Arne Sorenson Is The 201...", "Casey's Announces CEO ...", "Why You Need To Be The CEO Of Your Career", "C.E.O. Fired Over a Relationship ...", "Harvard study: What CEOs do all day", "How to use 'CEO magic' when trying to ...", "LinkedIn CEO Jeff Weiner steps down ...", "John Furner President & CEO of...", "Rise of the next-gen bank CEO", "HP has a new CEO - The Verge", "Meet Our CEO - Stellar", "Volkswagen Executive as New C.E.O. ...", "DFC's CEO visits Egypt to promote U.S ...", "Mike Roman | 3M CEO", "You are the CEO of Your Life | Personal ...", "CEO vs. Owner: The Key Differences ...", "Selective CEO begins the next chapter ...", "Trump says Google CEO Sundar Pichai ...", "CEO MESSAGE | JCB Global Website", "Amtrak Names William Flynn...", "Google CEO salary raised to \$2 million ...", "Related searches" (cartoon ceo, ceo logo, ceo sign), "McDonald's CEO pushed out after ...", and "Verizon CEO to Retire, Succeeded by a ...".

Societal Stereotypes in Data

- Biased data produces biased models
 - Thus, predictions are biased as well
- Alternative thought question:
 - What is a chair?



Research on Bias in AI

- Machines learn trustworthiness and likeability traits from faces
(Steed and Caliskan 2020)
- Self-driving cars biased against genders and races
(Wilson, Hoffman, and Morgenstern 2019)
- Males are over-represented in the reporting of web-based news articles
(Jia, Lansdall-Welfare, and Cristianini 2015)
- Males are over-represented in twitter conversations
(Garcia, Weber, and Garimella 2014)
- Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues
(Wagner et al. 2015)
- IMDB reviews written by women are perceived as less useful
(Otterbacher 2013)

Bias in Conversational AI Systems

The physician hired the secretary because he was overwhelmed with clients.



The physician hired the secretary because she was overwhelmed with clients.



The physician hired the secretary because she was highly recommended.



The physician hired the secretary because he was highly recommended.



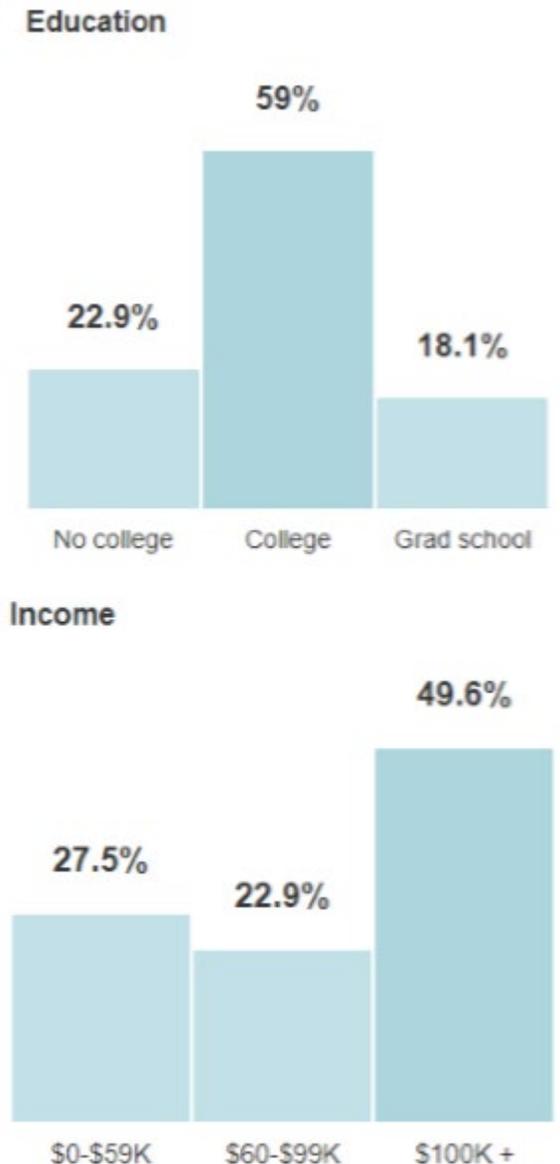
Sources of Bias in AI Systems

- Bias in data and sampling
 - (social biases, unrepresentative user base)
- Optimizing for a biased objective
 - (bad training)
- Inductive bias
 - (implicit assumptions made by the model itself)
- Bias amplification
 - (the model learns the “wrong” features)

Bias in Data and Sampling

- **Self-selection bias** is a statistical effect in which a group will select themselves, biasing a sample
- **Concretely:** who writes Yelp reviews? Who reads them?
 - People may not talk about things consistent with empirical measurement
 - Communities of language speakers lead to differing model performance
- What about system bias?
 - Can we tell if Yelp is biasing reviews?
 - “it would be a shame if you didn’t pay us and you got a few 1-star reviews...”

Distribution of Yelp Users



Bias in Language Identification

- NLP application: Identifying a language give a string written in it



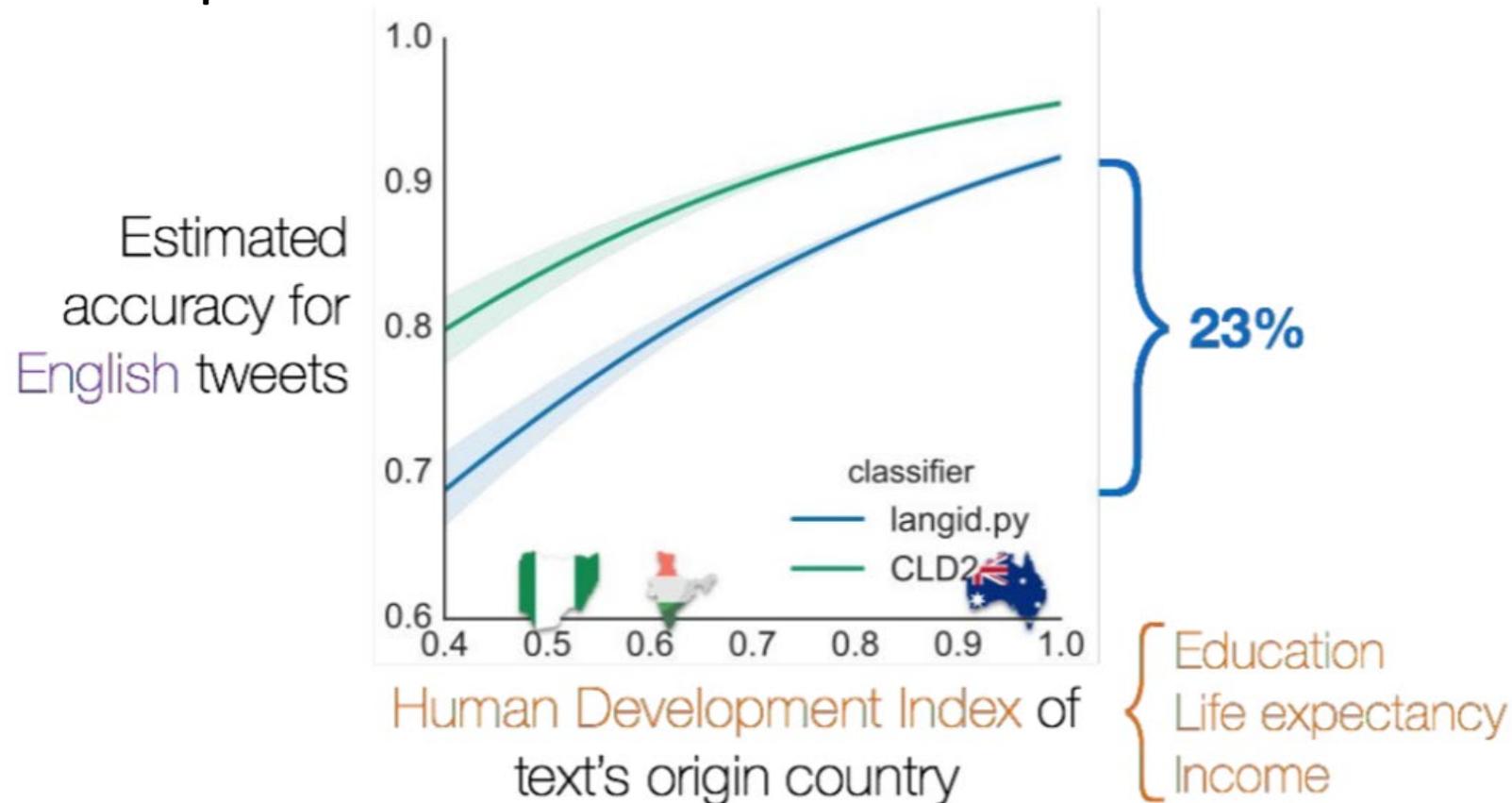
After language identification, we can look for keywords like flu/sick, then followup with a conclusive explanation
(maybe they're hungover)



If we can't identify the language to begin with, there's no way to extract followup semantics (i.e., we can't find keywords like flu/sick without knowing it's an English Tweet)

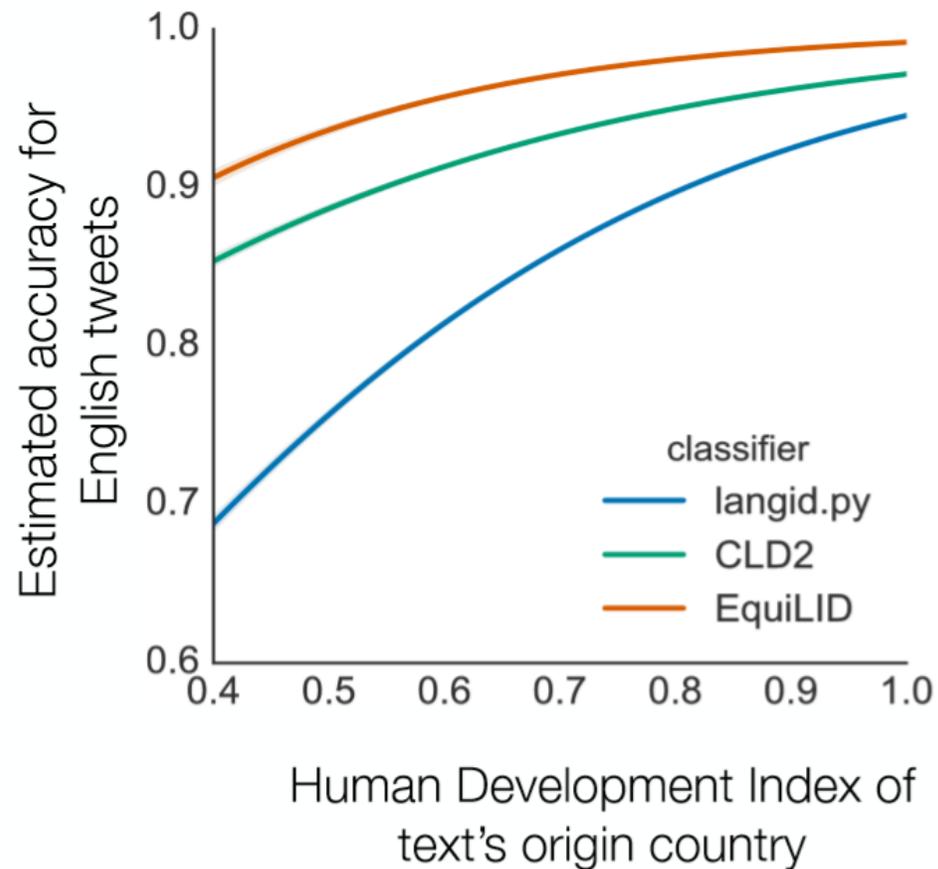
Bias in Language Identification

- Language Identification systems under-represent populations in underdeveloped countries



Bias in Language Identification

- By retraining on more representative corpora:



Objective Bias

- **Objective bias** occurs when models are asked to make predictions that actually answer a different question
- **Concretely:** “What is the **probability** that a given **person** will commit a serious **crime** in the **future** based on the **sentence given now?**”
- Example: COMPAS
 - Balanced data from people of all races (and race was not a feature)
 - **Problem:** “who will commit a crime” is not obtainable (we can’t know it ahead of time)
 - **Instead:** model was learning “who is more likely to be convicted” (notice the difference!)

Inductive Bias

- An **Inductive bias** is the result of an implicit assumption made in the construction of a given model
- **Concretely:** Embeddings may represent biases
- In word2vec:
 - $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\ programmer} - \overrightarrow{homemaker}$

Inductive Bias in Embeddings

$$\min \cos(\mathit{he} - \mathit{she}, x - y) \text{ s.t. } \|x - y\|_2 < \delta$$

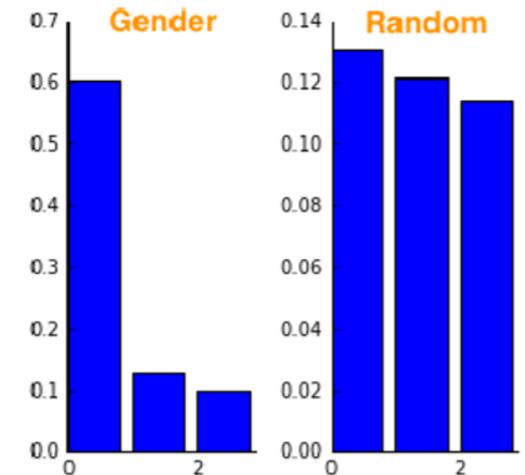
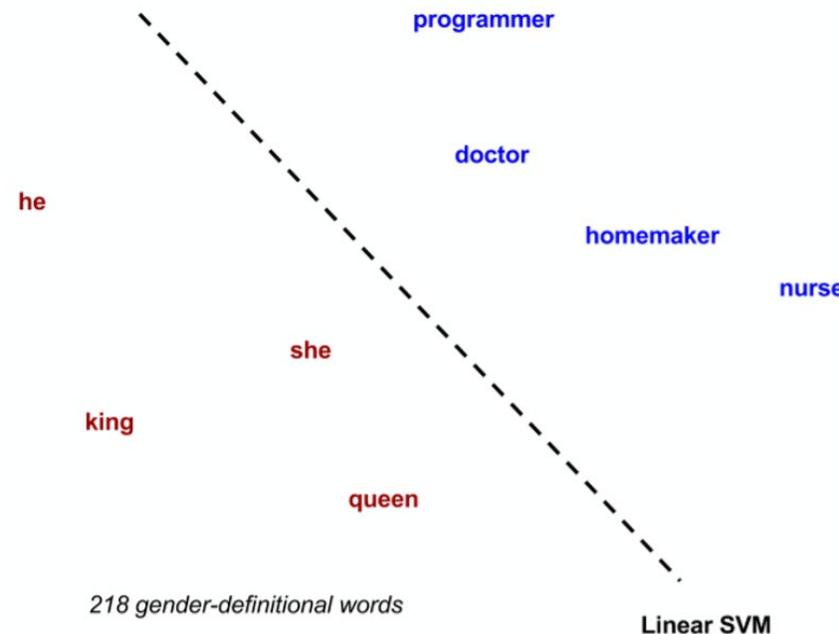
Extreme <i>she</i>	Extreme <i>he</i>		Gender stereotype <i>she-he</i> analogies
1. homemaker	1. maestro	sewing-carpentry	registered nurse-physician
2. nurse	2. skipper	nurse-surgeon	interior designer-architect
3. receptionist	3. protege	blond-burly	feminism-conservatism
4. librarian	4. philosopher	giggle-chuckle	vocalist-guitarist
5. socialite	5. captain	sassy-snappy	diva-superstar
6. hairdresser	6. architect	volleyball-football	cupcakes-pizzas
7. nanny	7. financier		
8. bookkeeper	8. warrior		Gender appropriate <i>she-he</i> analogies
9. stylist	9. broadcaster	queen-king	sister-brother
10. housekeeper	10. magician	waitress-waiter	mother-father
			ovarian cancer-prostate cancer
			convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she*–*he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

Fixing Inductive Bias: Dibiasking

- First: Identify some biased subspace (e.g., using PCA)
- Second: Find subspace-neutral words (e.g., using SVM)
- Third: Transform embeddings space to minimize the subspace components
- TL;DR: Minimize impact of embeddings component that leads to bias in subspace-neutral words

$\overrightarrow{\text{she}} - \overrightarrow{\text{he}}$
 $\overrightarrow{\text{her}} - \overrightarrow{\text{his}}$
 $\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$
 $\overrightarrow{\text{Mary}} - \overrightarrow{\text{John}}$
 $\overrightarrow{\text{herself}} - \overrightarrow{\text{himself}}$
 $\overrightarrow{\text{daughter}} - \overrightarrow{\text{son}}$
 $\overrightarrow{\text{mother}} - \overrightarrow{\text{father}}$
 $\overrightarrow{\text{gal}} - \overrightarrow{\text{guy}}$
 $\overrightarrow{\text{girl}} - \overrightarrow{\text{boy}}$
 $\overrightarrow{\text{female}} - \overrightarrow{\text{male}}$



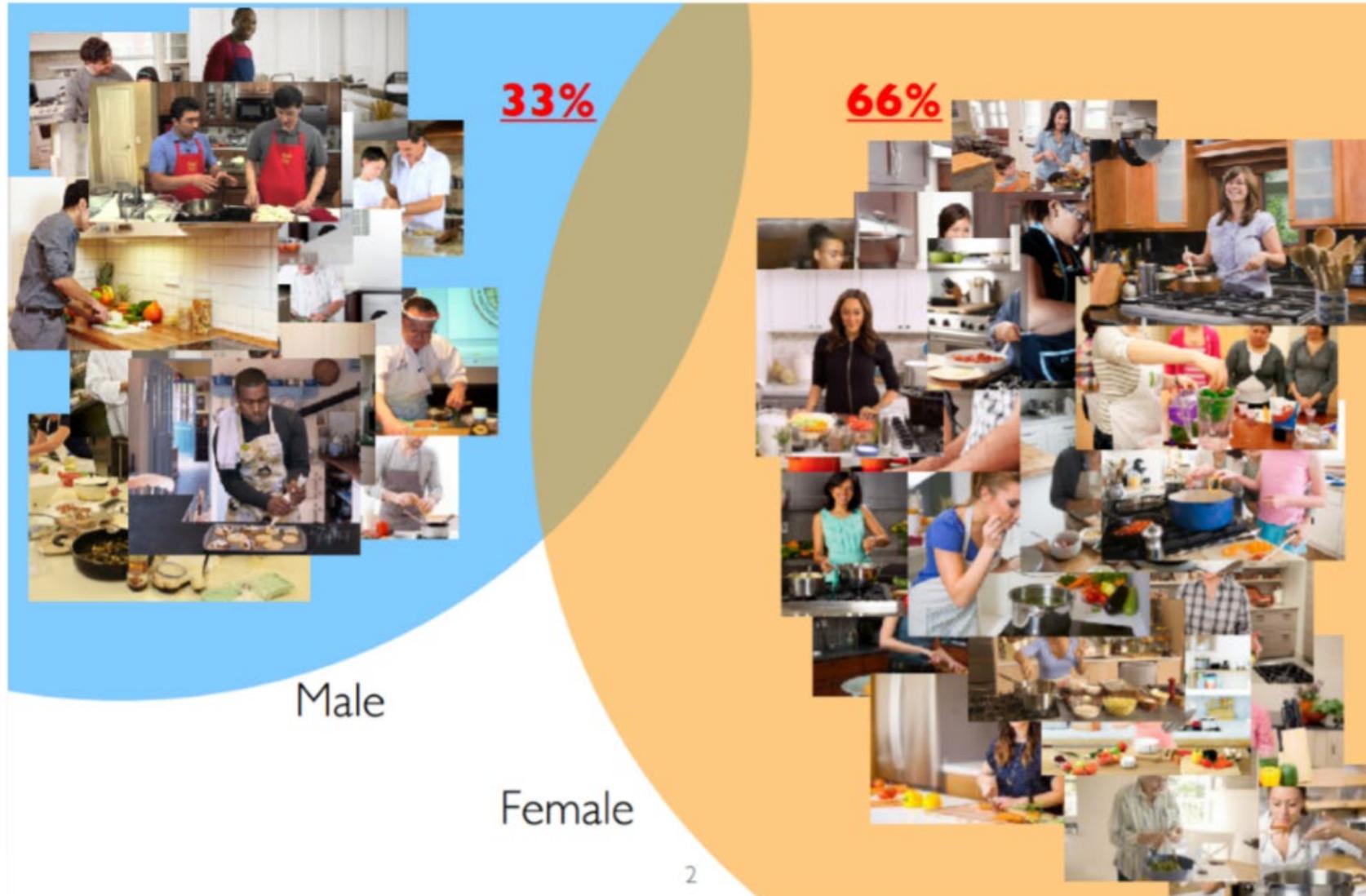
Top principal components identify gender subspace

Bias Amplification

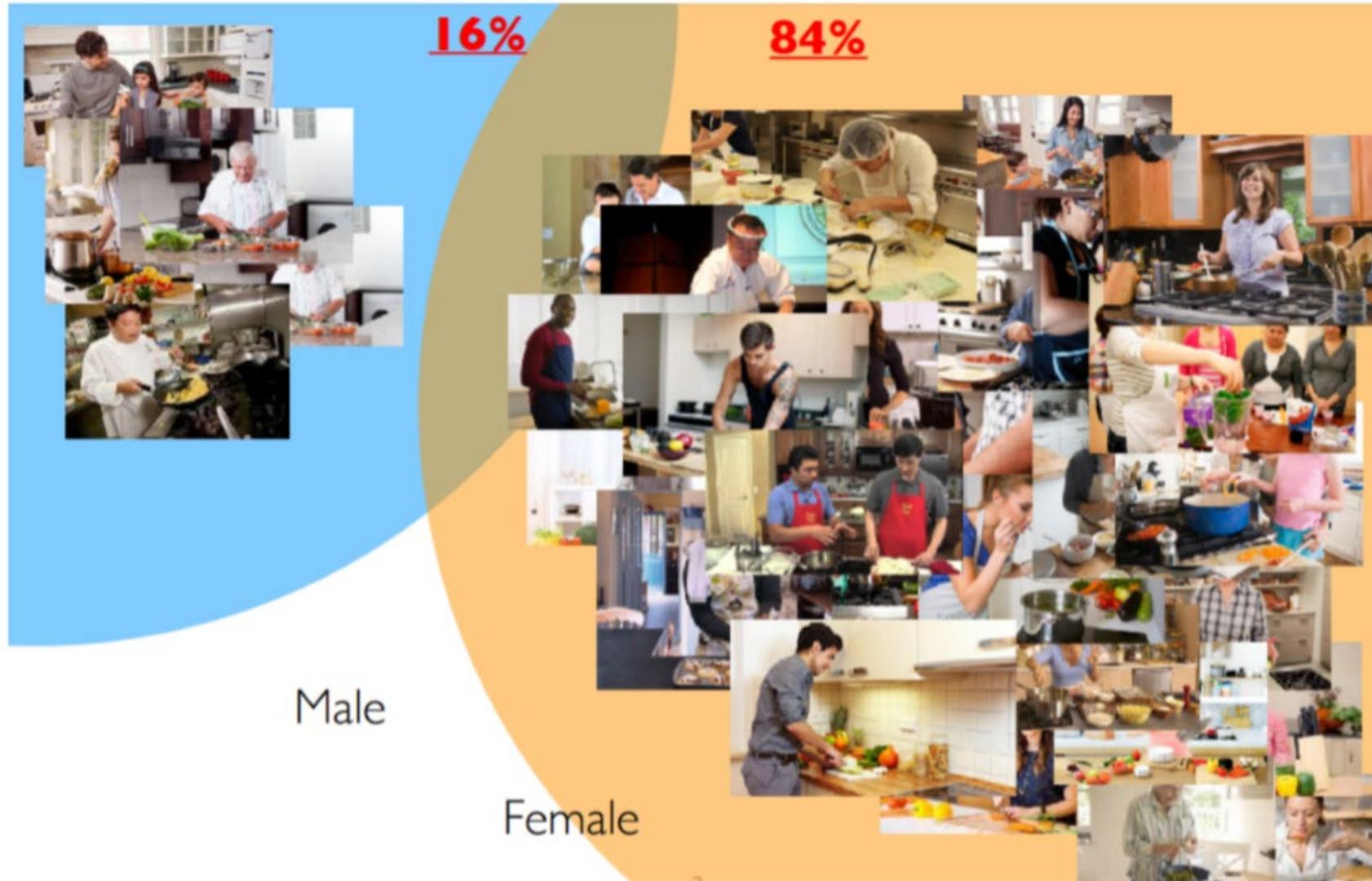
- **Bias Amplification** occurs when unrepresentative data leads a model to learn the wrong features
- **Recall:** What is a chair?
- **Concretely:** if all of your dataset contains barstools as examples of chairs, your model will learn the wrong features
 - e.g., it will only have examples of tall, backless seats near alcohol sources



Bias Amplification: Training

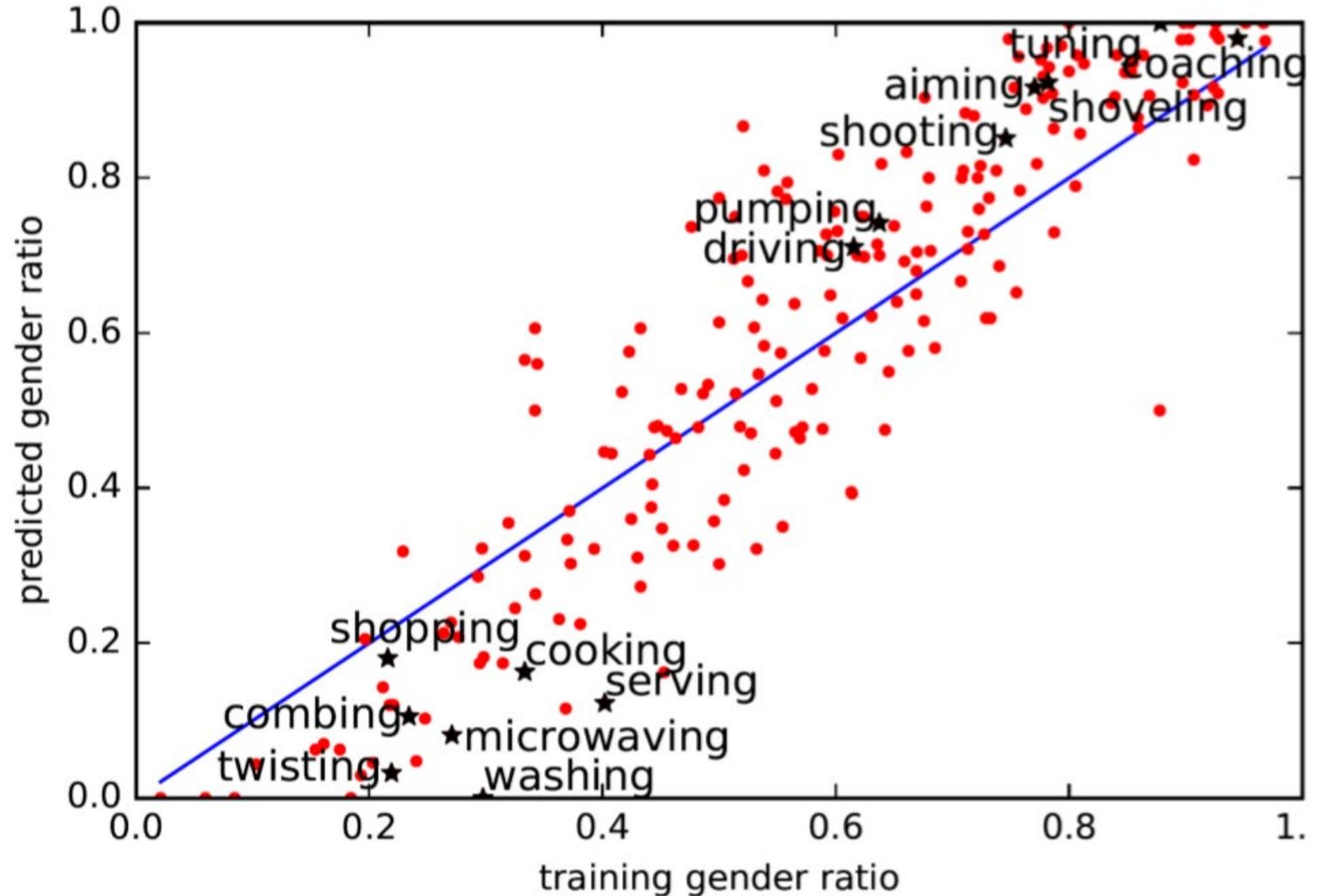


Bias Amplification: Predictions



Reducing Bias Amplification

- Find ratio of predictions made against ground truth labels
- Identify distribution of labels in dataset
- Adjust predicted outputs based on target distribution



Bias in AI and NLP

- Bias in data and sampling
 - (social biases, unrepresentative user base)
- Optimizing for a biased objective
 - (bad training)
- Inductive bias
 - (implicit assumptions made by the model itself)
- Bias amplification
 - (the model learns the “wrong” features)

Bias in AI and NLP

- These are critical human-facing systems that have real impact
- Key questions:
 - Should our ML models represent actual real-world data?
 - Is it right to adjust models to reflect desired distributions?
 - If we do try to de-bias, how do we tell what's right?