Advanced NLP

Lecture 6: Interpretability II

Dr. Gabriel Stanovsky

May 7, 2023

Suggested reading: **Local Explanations for DL Models, Marasovich, 2022 Class Material**
Interpreting Predictions of NLP Models, Wallace, Gardner & Singh, EMNLP20 Tutorial [1]

Announcements

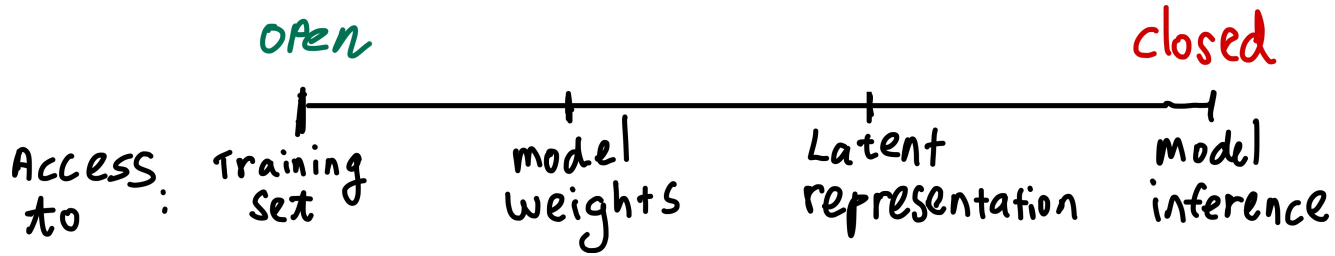
- **Exercises 1 is out**
- **Check out the project forums**
  - **Thanks Dan!**
- **Share your slides if you haven't already**

ANLP Syllabus

1. **Intro to Modern NLP**
   - Intermediate Tasks
   - Downstream Tasks
   - Text Representation
   - Large Language Models
2. **Explainability**
   - Interpretability
   - **Social Biases and toxicity**
     [You are here]
   - Artifacts and spurious correlations
3. **Research Agendas**
   - Data collection
   - Sustainability & efficiency
   - What do LLMs know?
   - Multimodality
   - Multilinguality
4. **Summary**

Q: Why did the model make this prediction?

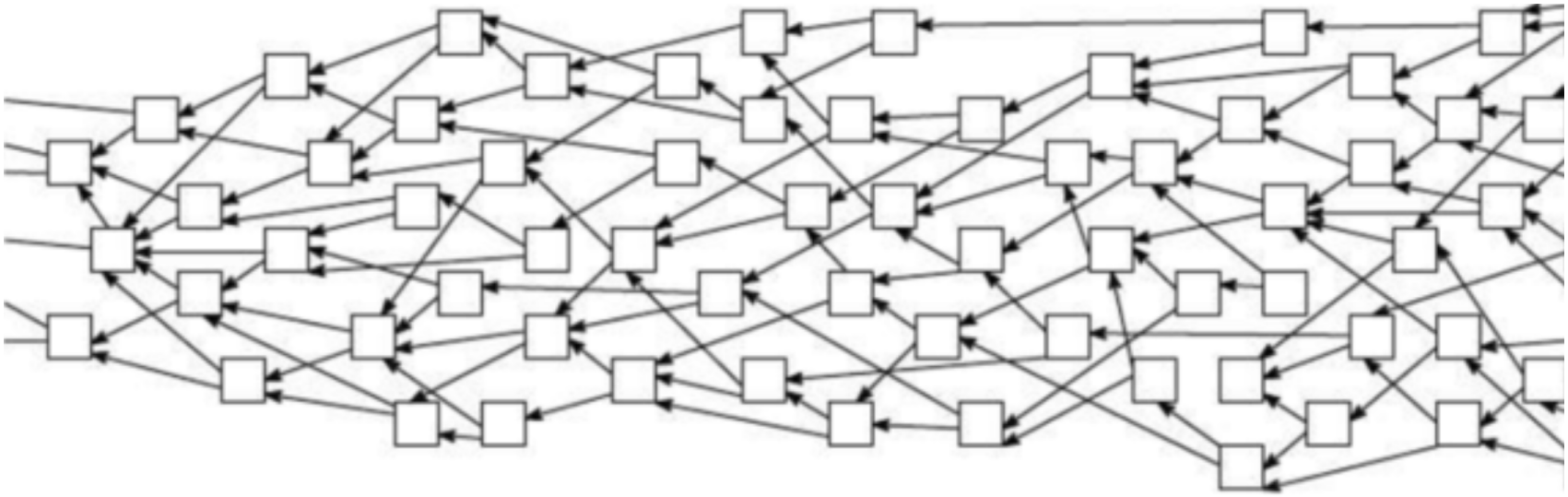
- **Observed**: **correlations** between some factors and the model output
- **Desired**: understanding **causality** between them
  - Many potential **intervening variables**
  - Ample opportunities for **spurious correlations**



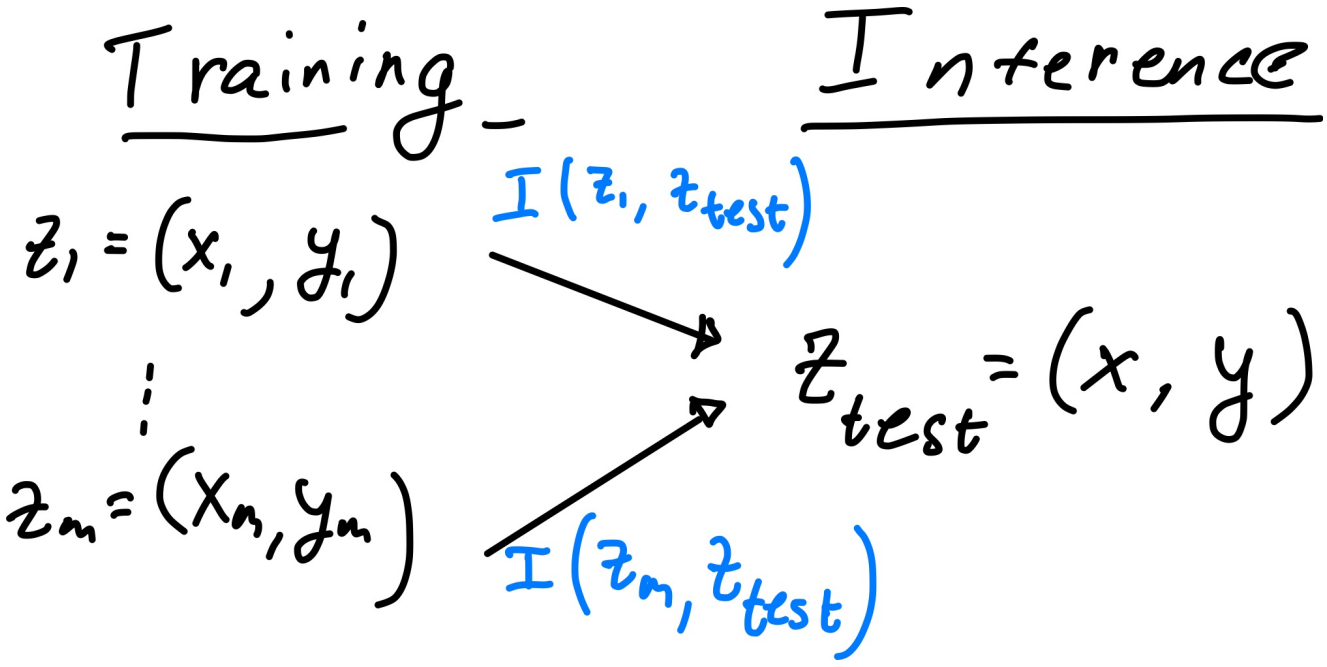
Counterfactuals

A **causes** *B* if *all other things being equal*,
intervening to change *A* inevitebly leads to a change in *B*

- **Counterfactual analysis** separates correlation from causality
- Requires a mecahnism for **minimal intervention**
- Dependant on the causal **world representation**
  - I.e., what consitutes atomic vertices in the causal graph



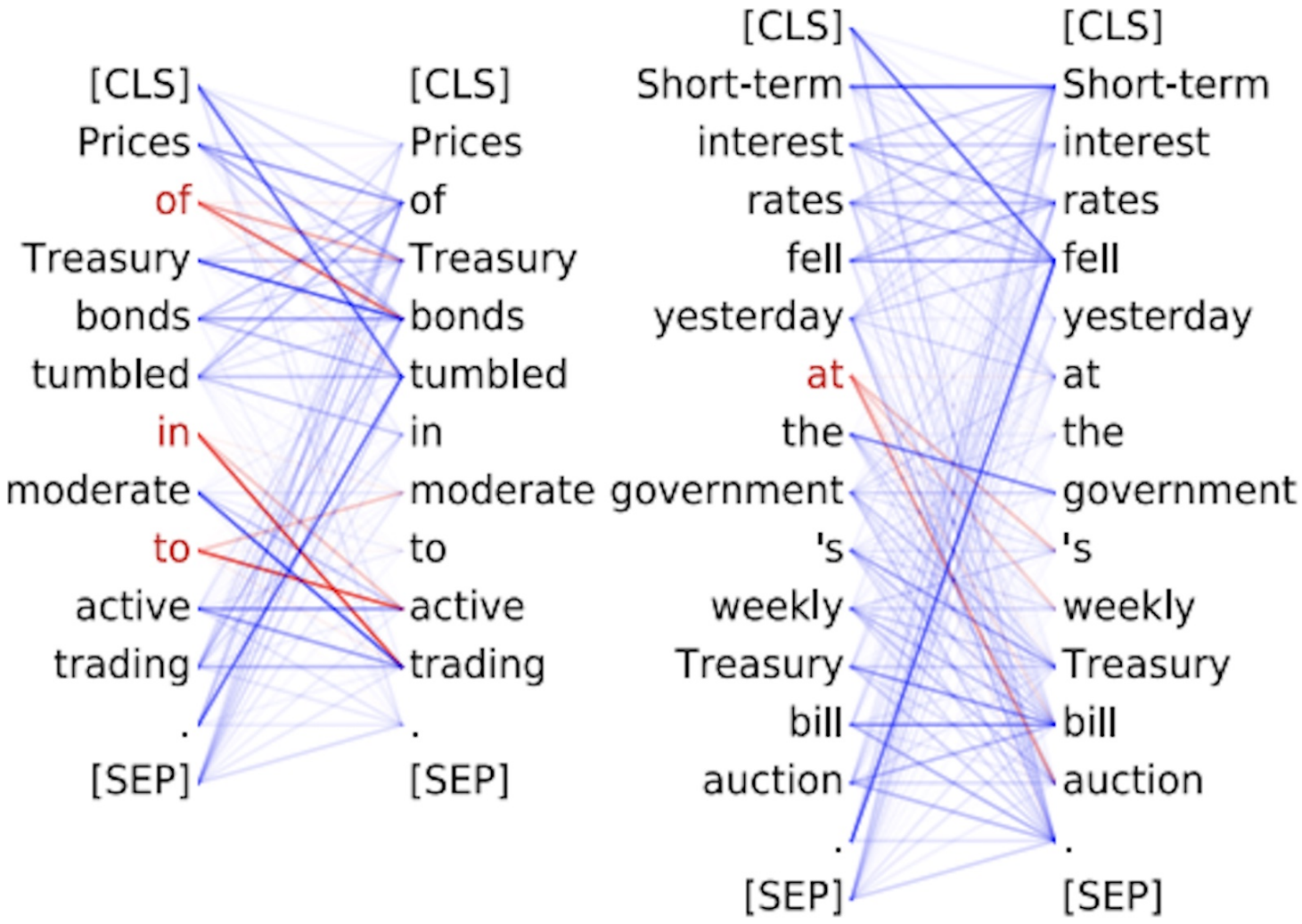
Causal Graph for Influence Functions in Sentiment Analysis



- $I(z_i, z_{test})$ quantifies how much z_i causes z_{test} [2]
- A test sample induces ranking of the **influence** of all training samples

What in the network weights caused this prediction?

- **Prepositions** attend to their objects
- 76.3% accuracy at the `pobj` relation

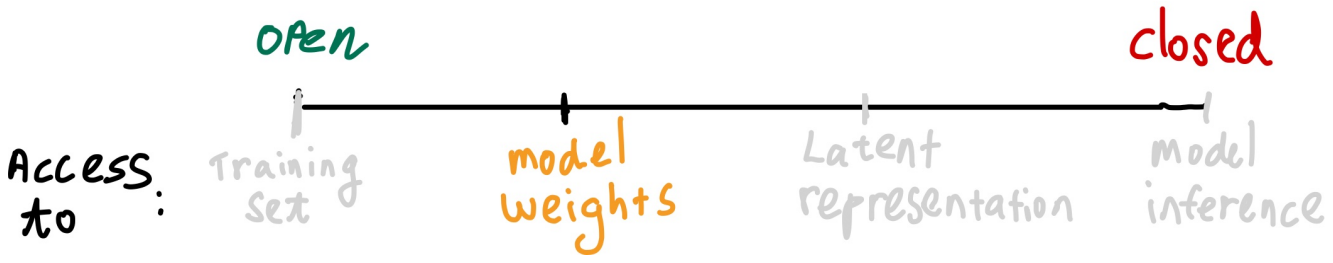


What in the network weights caused this prediction?

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

What in the network weights caused this prediction?

- We discussed **attention** as means for explanation [3]
  - Counterfactuals as evidence **against** attention as explanation [4]
  - And a counter-evidence **in favor** attention [5]
- Highlights the challenges of **disentangling** correlation from causation



Today

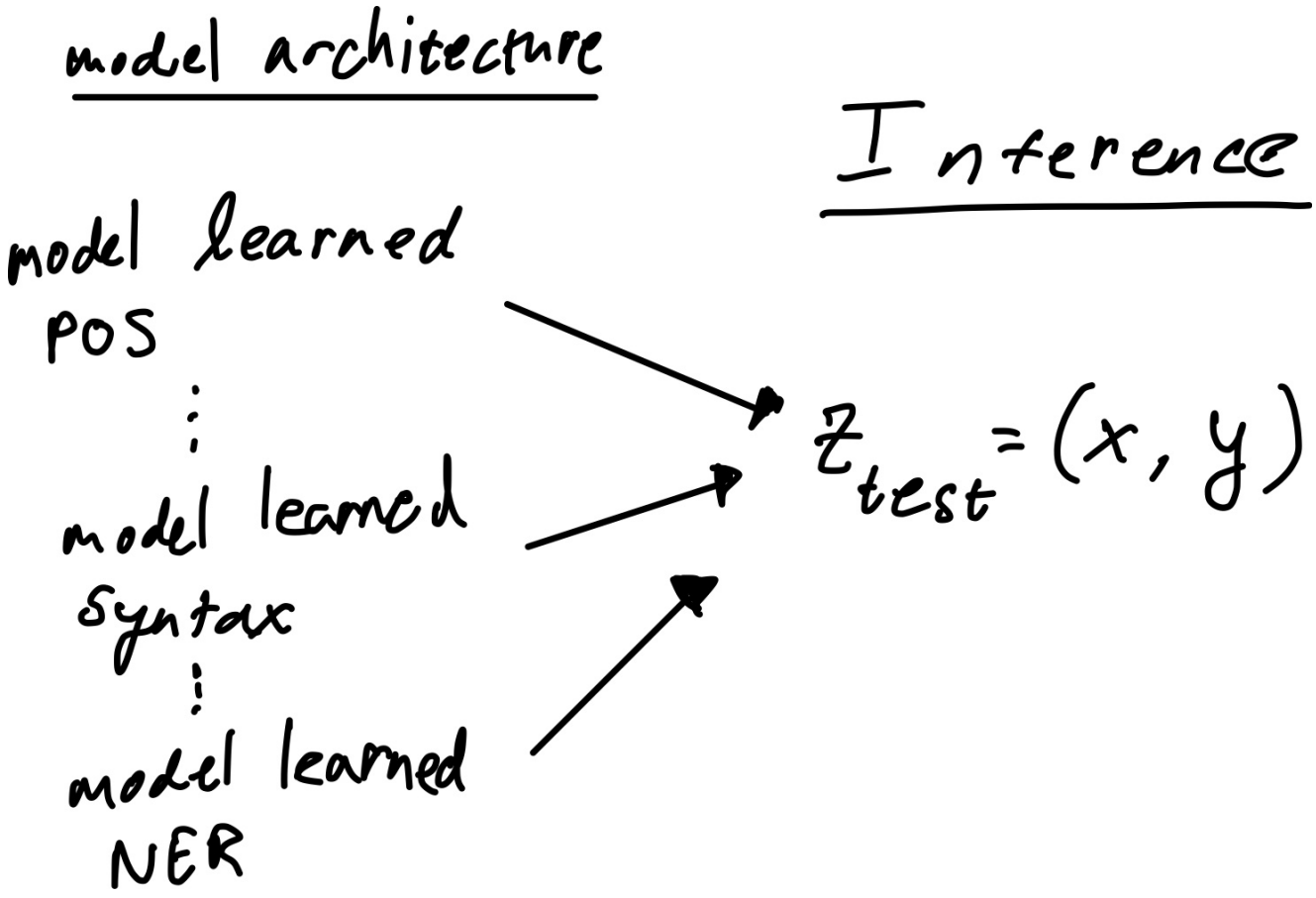
1. Explaining Model Behavior with Probing
2. Interpretation during Inference
3. Social Bias as Intervening Factors
4. Discussion

Outline

What in the latent represenation caused this prediction?

- Assume we have access to some aspects of the internal represenation
  - E.g., encoding at different layers
- We'd like to know the causal effect their properties and a prediction

Causal Graph for Learned Architecture



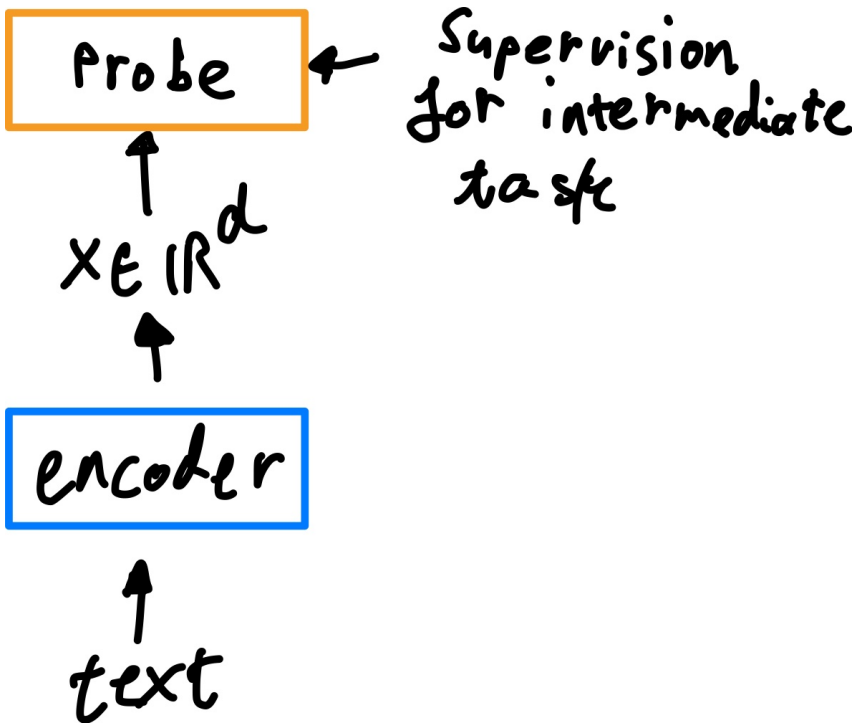
- Many works propose performance caused by learning **intrinsic tasks**
  - Recall that LLMs don't train on them **explicitly**
- **Q: How can they still learn intrinsic tasks?**

How Can LLMs Implicitly Learn Intermediate Tasks?

- Consider the following MLM examples:
  - The **MASK** jumped on the couch
  - Sally **MASK** to meet Harry
- **Q: What's the expected POS for each of these examples?**
- LLMs may **need to know POS** to fill these examples
  - Similar arguments for **other intermediate tasks**

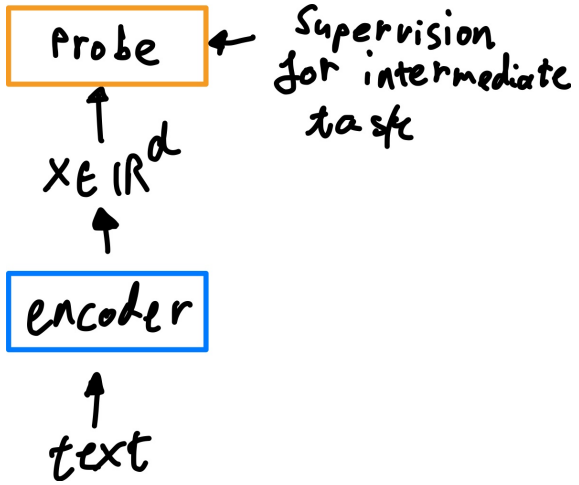
Probing

- How can we tell if a network has learned tasks implicitly?
- A common approach is **probing**
  - Latent representation used as input to a classifier
  - The classifier is **trained** for the intermediate task (e.g., POS)
  - Classifier does well $\Rightarrow$ LLM learned the task



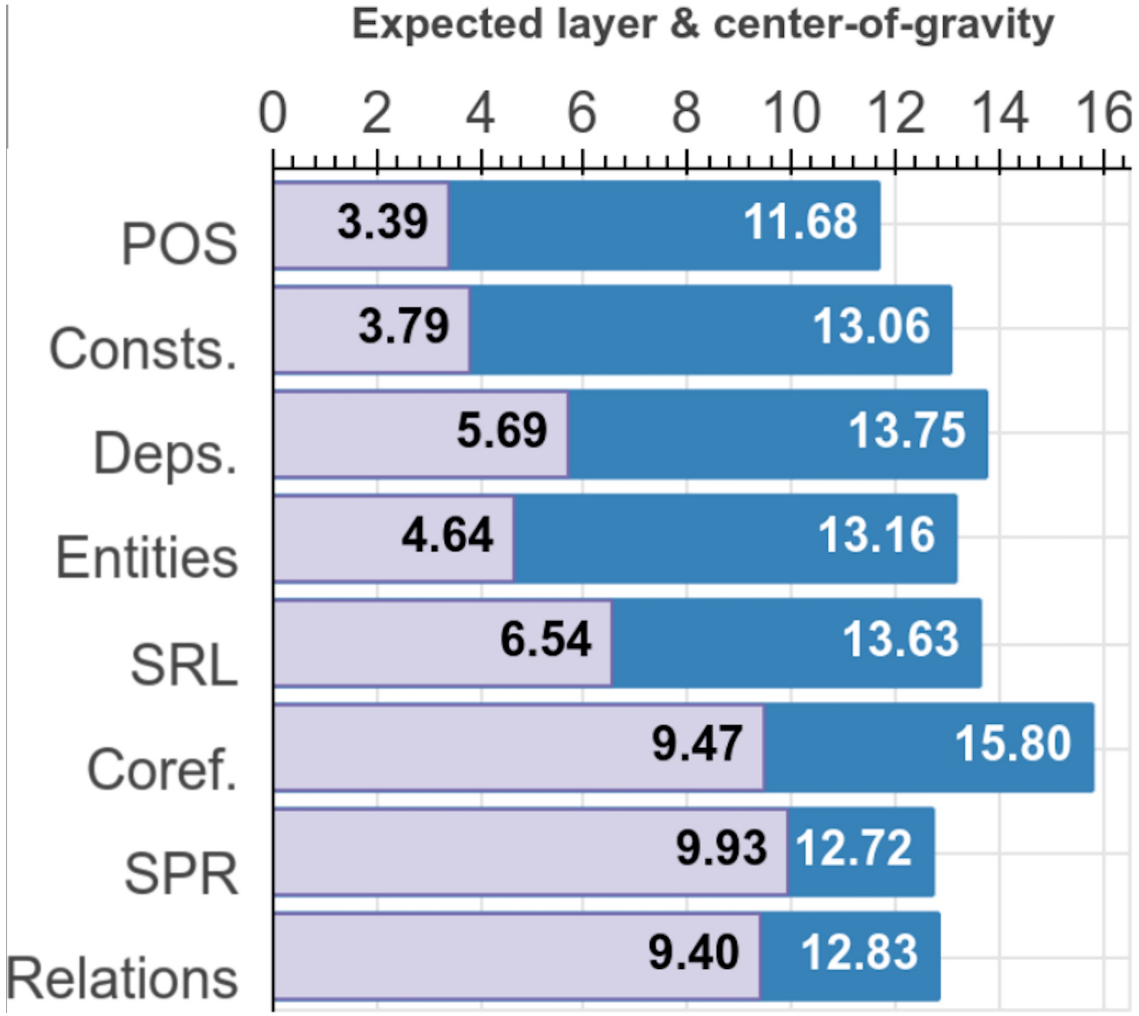
Probing

- **Q: What's a possible flaw in this argument?**
  - Parameterized classifier could learn POS from *any* representation
- **Q: How can this flaw be mitigated?**
  - Train a very simple classifier (e.g., linear regression)
  - Compare with random embeddings
  - This is the **counterfactual** argument



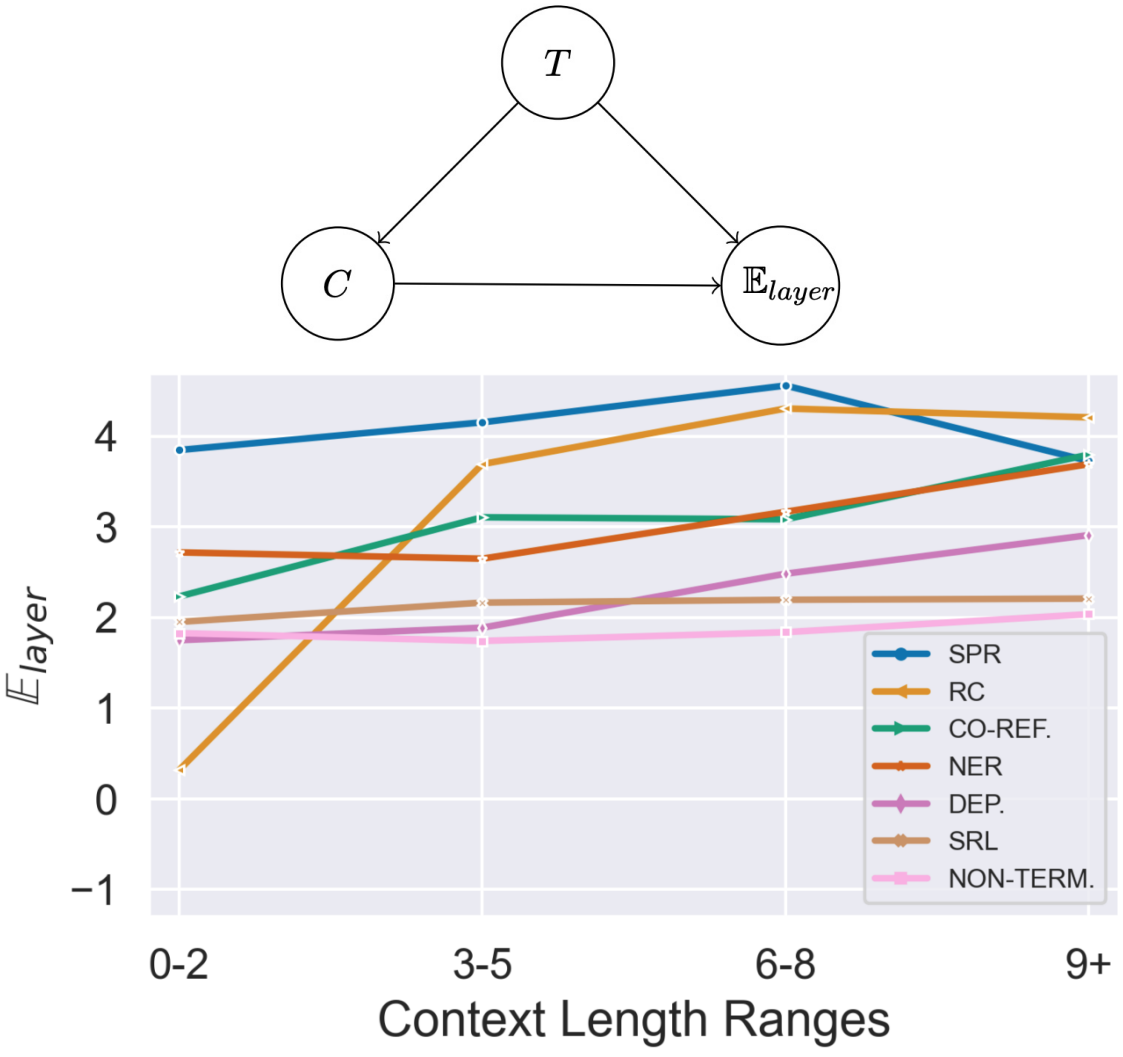
BERT Rediscovers the NLP Pipeline

- Layers in BERT seem to **learn different intermediate tasks** [6]
- . . . and they so in the **traditional linguistic order**



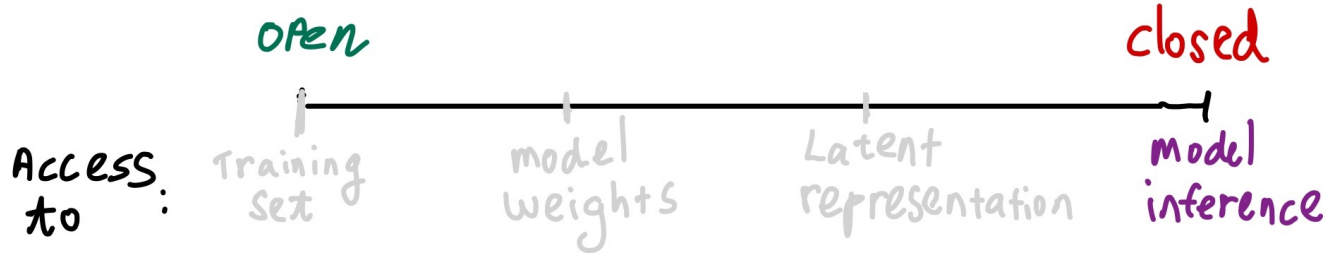
Context Length as a confounding factor

- Length proposed as a **confounding factor** [7]



Outline

- Assume we only have access to model's **inference**
  - I.e., provide arbitrary inputs observe its outputs
  - **Confidence** is often considered as part of output
- Explanation assumes input elements **cause** certain outputs
- **Q: What are the advantages of this approach?**
  - Requires only an API for the model
  - Relatively cheap to run
  - Model independent



Input Reduction

- What if we remove the **least** important words?
- But, keep the model prediction the **same**

[Feng et al. 2018]

Input Reduction

- What if we remove the **least** important words?
- But, keep the model prediction the **same**

Question	Confidence
What did Tesla ~~spend~~ Astor's money on ?	0.78

[Feng et al. 2018]

Input Reduction

- What if we remove the **least** important words?
- But, keep the model prediction the **same**

Question	Confidence
What did Tesla ~~spend~~ Astor's money on ?	0.78
What did Tesla Astor's ~~money~~ on ?	0.74

[Feng et al. 2018]

Input Reduction

- What if we remove the **least** important words?
- But, keep the model prediction the **same**

Question								Confidence
What	did	Tesla	~~spend~~	Astor's	money	on	?	0.78
What	did	Tesla		Astor's	~~money~~	on	?	0.74
What	did	Tesla		Astor's		~~on~~	?	0.76
~~What~~	did	Tesla		Astor's			?	0.80
	did	Tesla		Astor's			~~?~~	0.87

[Feng et al. 2018]

Input Reduction

- What if we remove the **least** important words?
- But, keep the model prediction the **same**

Question								Confidence
What	did	Tesla	~~spend~~	Astor's	money	on	?	0.78
What	did	Tesla		Astor's	~~money~~	on	?	0.74
What	did	Tesla		Astor's		~~on~~	?	0.76
~~What~~	did	Tesla		Astor's			?	0.80
	did	Tesla		Astor's			~~?~~	0.87
	did	~~Tesla~~		Astor's				0.82
	did			~~Astor's~~				0.89
	did							0.91

[Feng et al. 2018]

Input Reduction

- What if we remove the **least** important words?
- But, keep the model prediction the **same**

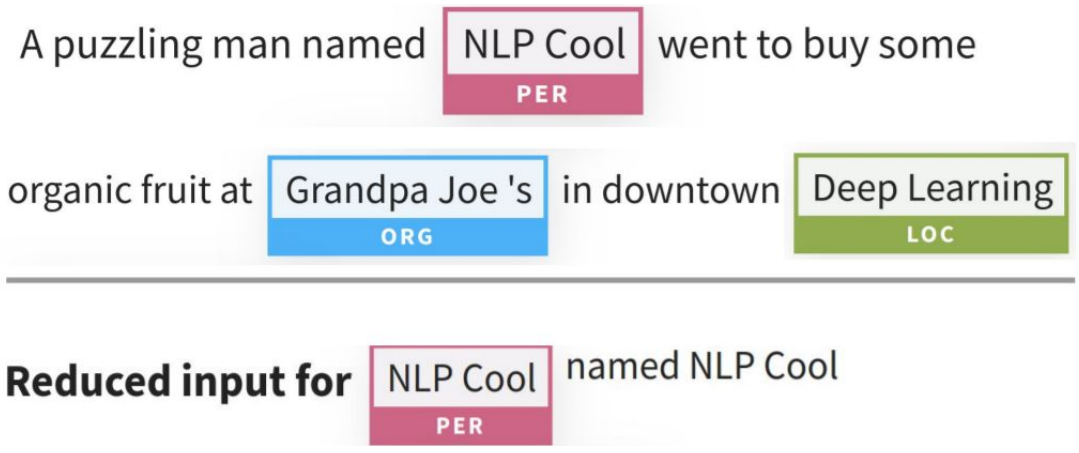
Question								Confidence
What	did	Tesla	~~spend~~	Astor's	money	on	?	0.78
What	did	Tesla		Astor's	~~money~~	on	?	0.74
What	did	Tesla		Astor's		~~on~~	?	0.76
~~What~~	did	Tesla		Astor's			?	0.80
	did	Tesla		Astor's			~~?~~	0.87
	did	~~Tesla~~		Astor's				0.82
	did			~~Astor's~~				0.89
	did							0.91

Prediction remains the same.

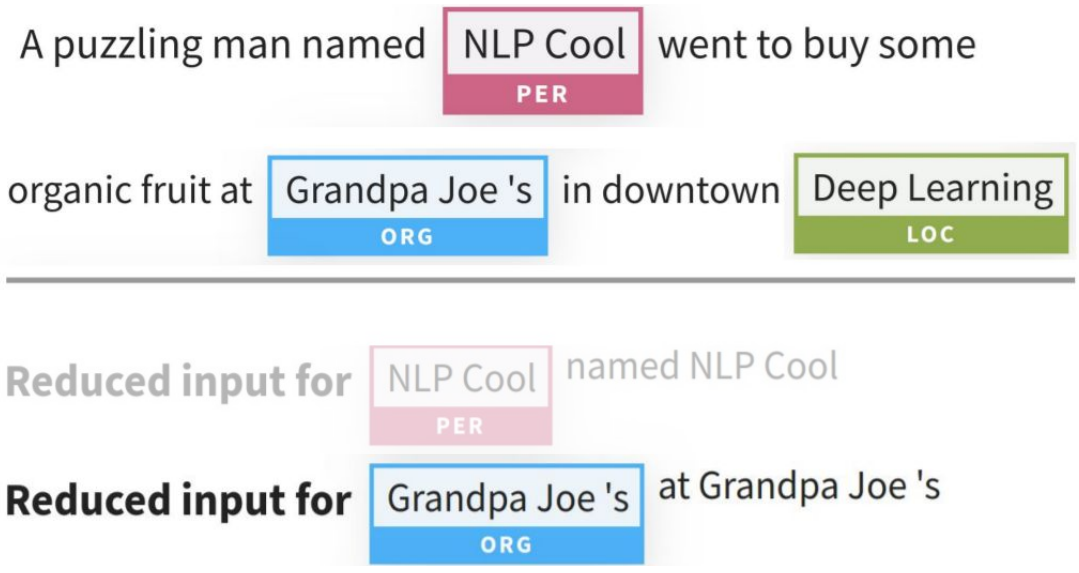
[Feng et al. 2018]

Input Reduction as Explanations

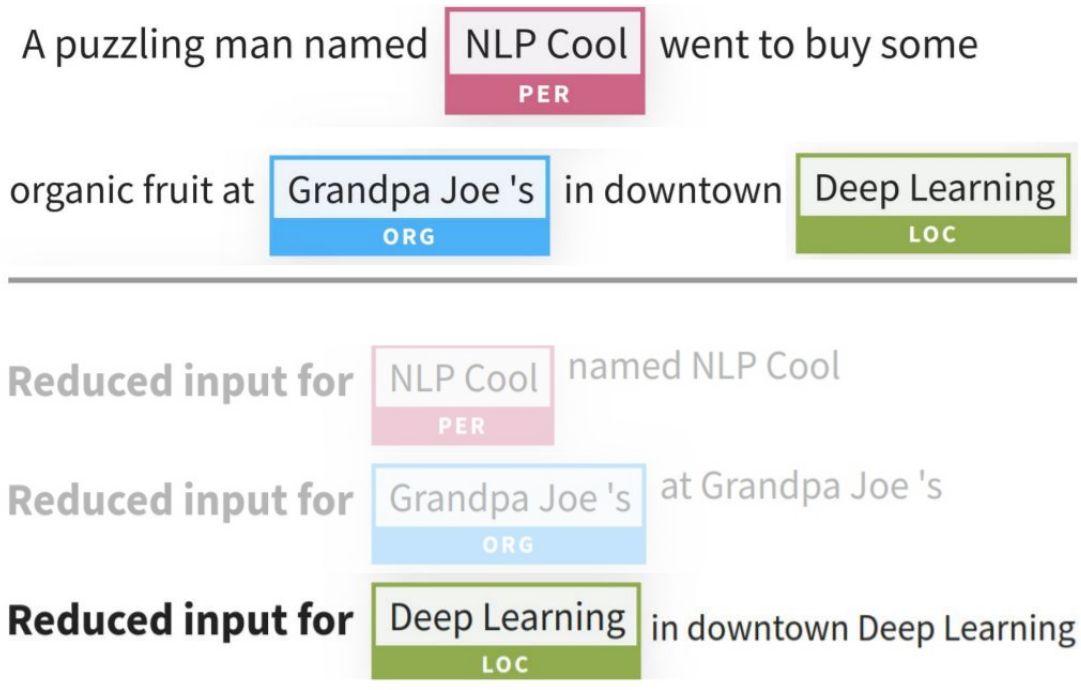
Input Reduction as Explanations



Input Reduction as Explanations



Input Reduction as Explanations



Input Reduction Examples

SQuAD	
Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	**did**
Confidence	0.78 → 0.91

[Feng et al. 2018]

Input Reduction Examples

SQuAD

Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	**did**
Confidence	0.78 → 0.91

VQA

Original	What color is the flower ?
Answer	yellow
Reduced	**flower ?**
Confidence	0.827 → 0.819

[Feng et al. 2018]

Input Reduction Examples

SQuAD

Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	**did**
Confidence	0.78 → 0.91

VQA

Original	What color is the flower ?
Answer	yellow
Reduced	**flower ?**
Confidence	0.827 → 0.819

SNLI

Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Answer	Contradiction
Reduced	**dancing**
Confidence	0.977 → 0.706

[Feng et al. 2018]

Disadvantages of Input Reduction

- Search space is **exponential**
  - While search for explanation is only $O(\#$ of tokens$)$
- Explanation may **not be understandable** to humans

Leave-one-out: remove tokens and look at confidence

- Simplest method is leave-one-out: define importance as drop in prediction confidence when a feature (e.g., token, phrase) is removed

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	

[Li et al. 2017]

Leave-one-out: remove tokens and look at confidence

- Simplest method is leave-one-out: define importance as drop in prediction confidence when a feature (e.g., token, phrase) is removed

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	
~~What~~ did Tesla spend Astor's money on ?	0.67	What

[Li et al. 2017]

Leave-one-out: remove tokens and look at confidence

- Simplest method is leave-one-out: define importance as drop in prediction confidence when a feature (e.g., token, phrase) is removed

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	
~~What~~ did Tesla spend Astor's money on ?	0.67	What
What ~~did~~ Tesla spend Astor's money on ?	0.72	did

[Li et al. 2017]

Leave-one-out: remove tokens and look at confidence

- Simplest method is leave-one-out: define importance as drop in prediction confidence when a feature (e.g., token, phrase) is removed

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	
~~What~~ did Tesla spend Astor's money on ?	0.67	What
What ~~did~~ Tesla spend Astor's money on ?	0.72	did
What did ~~Tesla~~ spend Astor's money on ?	0.66	Tesla
What did Tesla ~~spend~~ Astor's money on ?	0.74	spend
What did Tesla spend ~~Astor's~~ money on ?	0.76	Astor's
What did Tesla spend Astor's ~~money~~ on ?	**0.48**	money
What did Tesla spend Astor's money ~~on~~ ?	0.72	on
What did Tesla spend Astor's money on ~~?~~	0.73	?

[Li et al. 2017]

Leave-one-out: remove tokens and look at confidence

- Simplest method is leave-one-out: define importance as drop in prediction confidence when a feature (e.g., token, phrase) is removed

Question	Confidence	Highlight
What did Tesla spend Astor's money on ?	**0.78**	
~~What~~ did Tesla spend Astor's money on ?	0.67	What
What ~~did~~ Tesla spend Astor's money on ?	0.72	did
What did ~~Tesla~~ spend Astor's money on ?	0.66	Tesla
What did Tesla ~~spend~~ Astor's money on ?	0.74	spend
What did Tesla spend ~~Astor's~~ money on ?	0.76	Astor's
What did Tesla spend Astor's ~~money~~ on ?	**0.48**	money
What did Tesla spend Astor's money ~~on~~ ?	0.72	on
What did Tesla spend Astor's money on ~~?~~	0.73	?

What did Tesla spend Astor's money on ?

[Li et al. 2017]

Disadvantages of the Leave-One-Out-Method

- Number of feature subsets is **exponential**
  - Requires knowing constituent elements **in advance** [8]

The movie is mediocre, maybe even bad.	**Negative** 99.8%
The movie is mediocre, maybe even ~~bad~~.	**Negative** 98.0%
The movie is ~~mediocre~~, maybe even bad.	**Negative** 98.7%

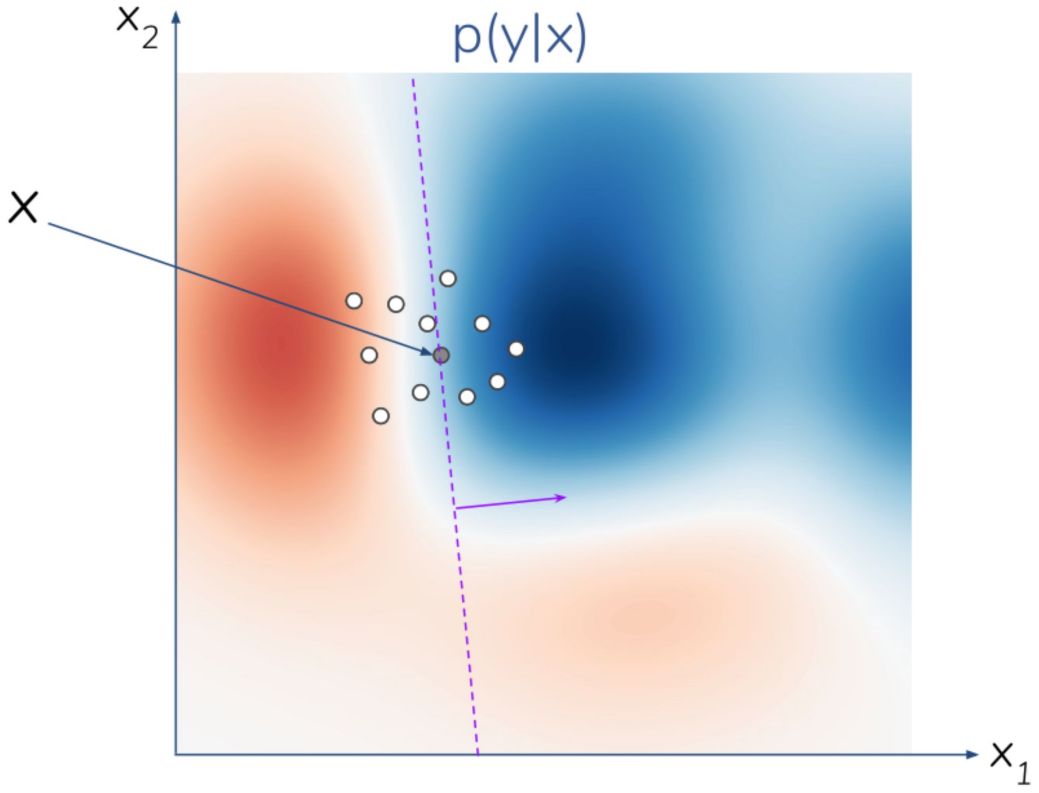
Disadvantages of the Leave-One-Out-Method

- Number of feature subsets is **exponential**
  - Requires knowing constituent elements **in advance** [8]

The movie is mediocre, maybe even bad.	**Negative** 99.8%
The movie is mediocre, maybe even ~~bad~~.	**Negative** 98.0%
The movie is ~~mediocre~~, maybe even bad.	**Negative** 98.7%
The movie is ~~mediocre~~, maybe even ~~bad~~.	**Positive** 63.4%

LIME [9]

- Define **perturbation** and **distance** functions
- Sample $\{(x_i, y_i)\}_{i=1,\ldots,k}$ uniformly, then weight by dist. from x
- Fit linear classifier $w \in \mathbb{R}^d$, *explaining* model prediction
- Intuitively, LIME searches for a linear model centered around x



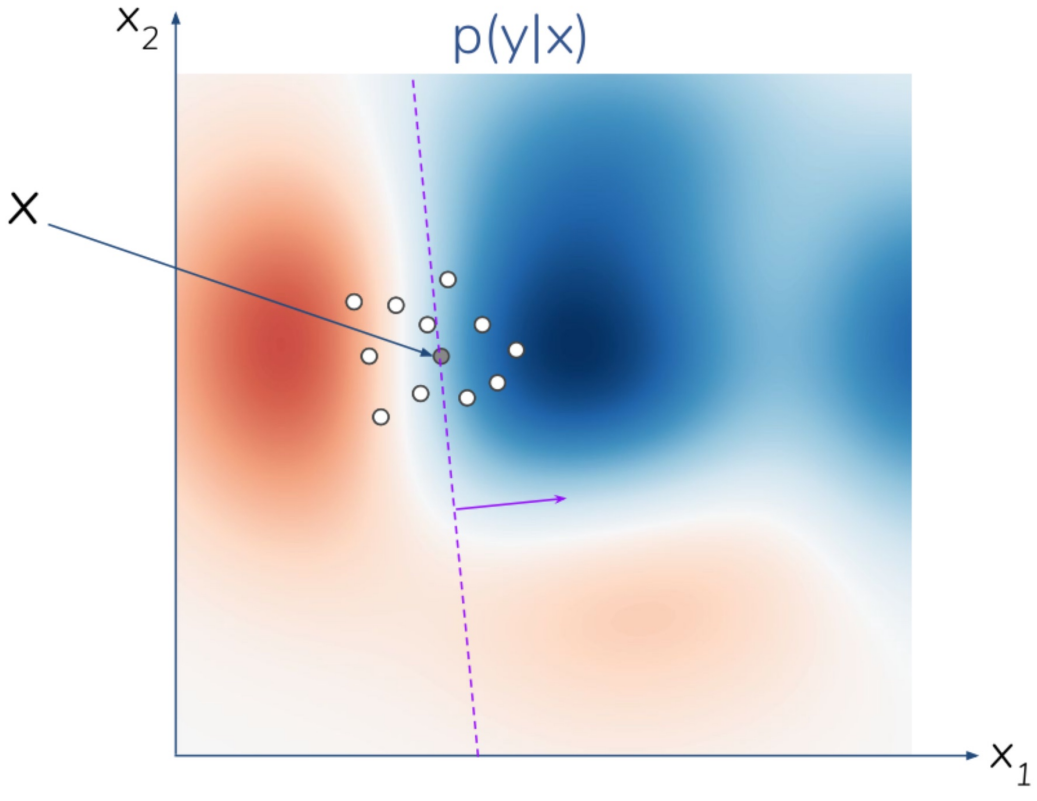
LIME - Example

The movie is mediocre, maybe even bad.	**Negative** 99.8%
The movie is mediocre, maybe even ~~bad~~.	**Negative** 98.0%
The movie is ~~mediocre~~, maybe even bad.	**Negative** 98.7%
The movie is ~~mediocre~~, maybe even ~~bad~~.	**Positive** 63.4%
The movie is ~~mediocre~~, ~~maybe~~ even ~~bad~~.	**Positive** 74.5%
The ~~movie~~ is mediocre, maybe even ~~bad~~.	**Negative** 97.9%

The movie is mediocre, maybe even bad.

LIME Disadvantages

- Explainability tradeoff - explanations are always linear
- May be computationally expensive, depending on # of samples



Outline

- Biases and spurious correlations can be **intervening** factors
  - E.g., the existence stereotypes **causes** prediction
  - Instead of **more meaningful signal**

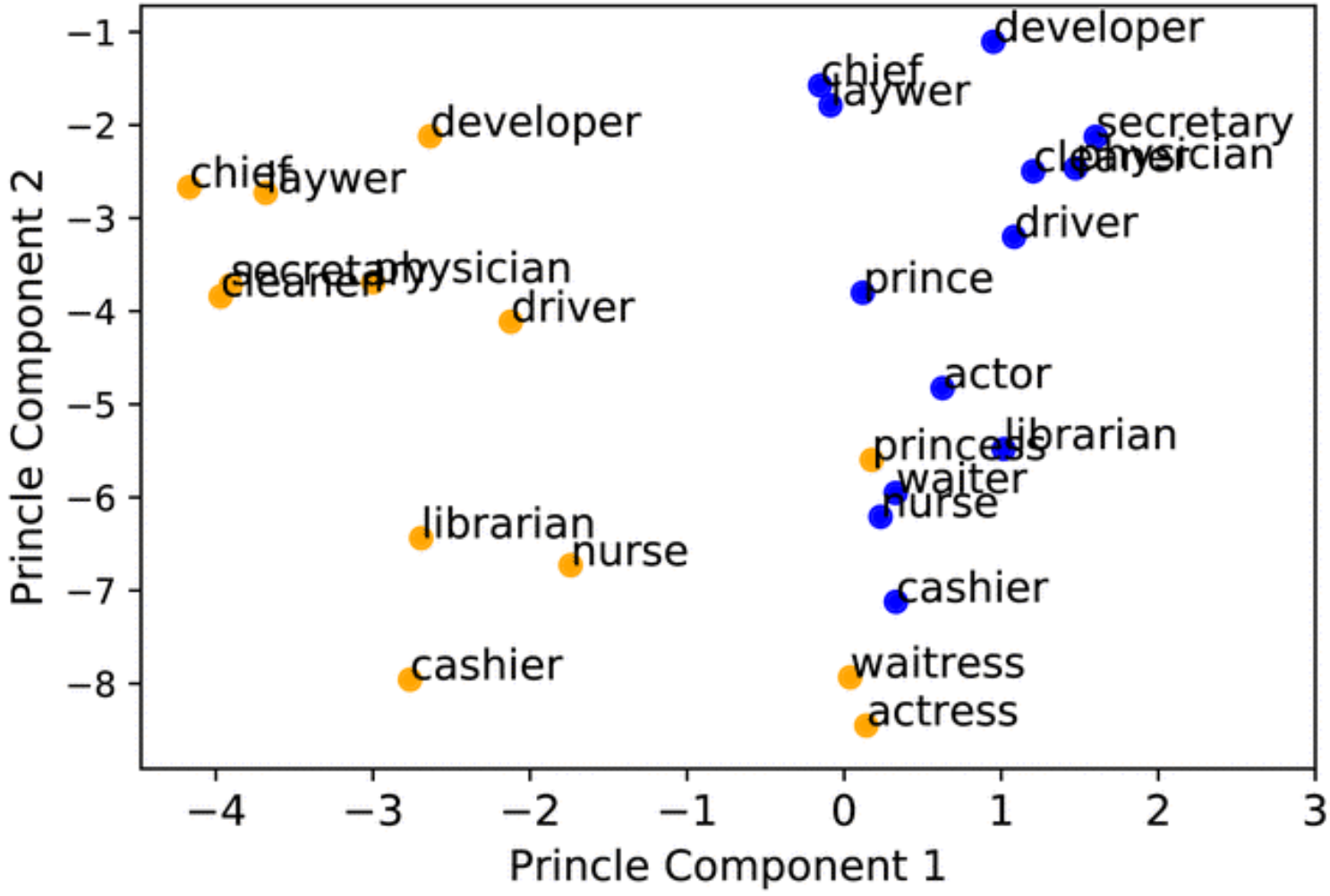
Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

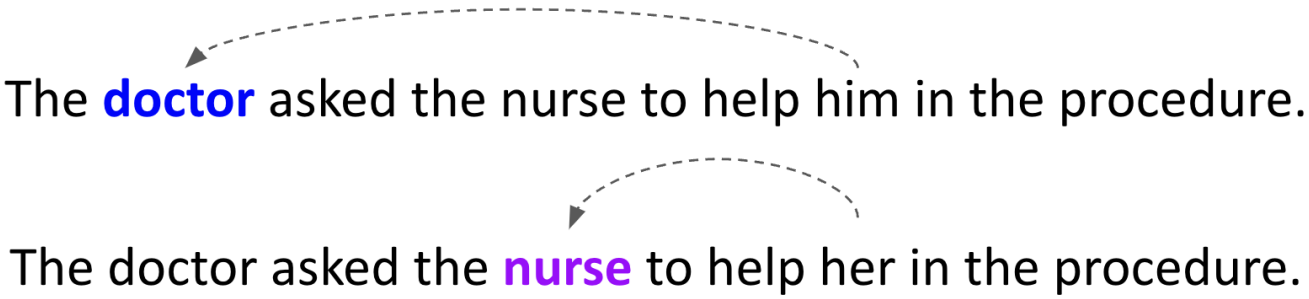
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Bolukbasi et al., 2016 [10]


developer
chief
laywer
developer
secretary
physician
cleaner
chief
laywer
driver
secretary
physician
cleaner
driver
prince
actor
librarian
princess
waiter
nurse
librarian
nurse
cashier
cashier
waitress
actress
Princle Component 2
Princle Component 1
−1
−2
−3
−4
−5
−6
−7
−8
−4
−3
−2
−1
0
1
2
3

Gender Bias in Coreference Resolution

- **Pronouns** *cause* coreference links
  - Counterfactuals tested via **Winograd-style** templates
  - Often conform to gender stereotypes [11, 12]



Gender Bias in Machine Translation [13]

- Coreference is an intrinsic task sometimes required for MT
  - To translate between languages with different morphology
- E.g., consider translating this sentence to Hebrew
  - *The trophy didn't fit the suitcase because it was too small*
- **Q: Does gender bias "propogate" to the downstream task?**

Methodology: Automatic evaluation of gender bias

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

Input: MT model + target language
Output: Accuracy score for gender translation

The **doctor** asked the nurse to help her in the procedure.

Methodology: Automatic evaluation of gender accuracy

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

Input: MT model + target language
Output: Accuracy score for gender translation

🇬🇧 The **doctor** asked the nurse to help her in the procedure.

🇪🇸 La doctora le pidió a la enfermera que le ayudara con el procedimiento.

Methodology: Automatic evaluation of gender accuracy

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

2. **Align** between source and target
   - Using *fast align* (Dyer et al., 2013)

Input: MT model + target language
Output: Accuracy score for gender translation

The **doctor** asked the nurse to help her in the procedure.

La doctora le pidió a la enfermera que le ayudara con el procedimiento.

Methodology: Automatic evaluation of gender accuracy

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender

Input: MT model + target language
Output: Accuracy score for gender translation

2. **Align** between source and target
   - Using *fast align* (Dyer et al., 2013)

3. **Identify** gender in target language
   - Using off-the-shelf morphological analyzers or simple heuristics in the target languages

The **doctor** asked the nurse to help her in the procedure.

La **doctora** le pidió a la enfermera que le ayudara con el procedimiento.

Methodology: Automatic evaluation of gender accuracy

1. **Translate** the coreference bias datasets
   - To target languages with grammatical gender
2. **Align** between source and target
   - Using *fast align* (Dyer et al., 2013)
3. **Identify** gender in target language
   - Using off-the-shelf morphological analyzers or simple heuristics in the target languages

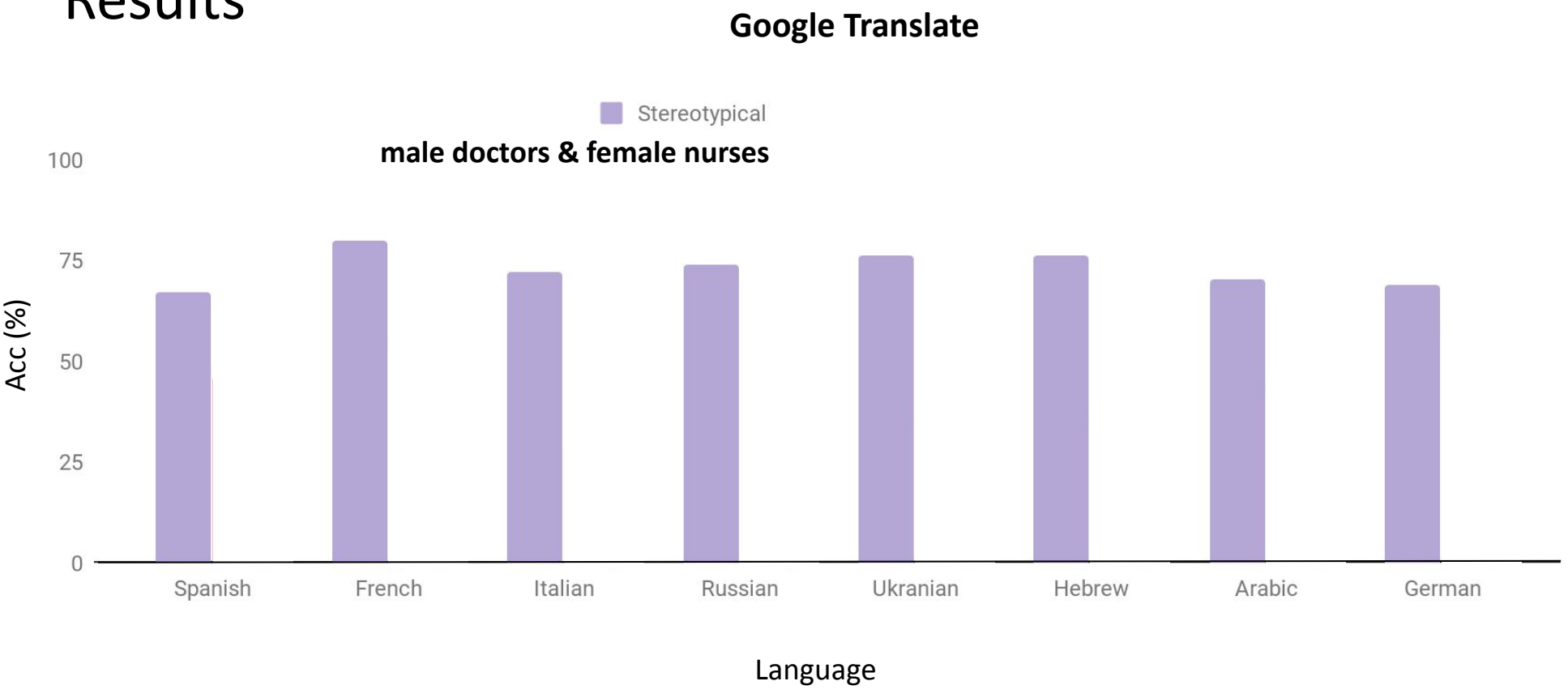
Input: MT model + target language
Output: Accuracy score for gender translation

Quality estimated at > 85% vs. 90% IAA
Doesn't require reference translations!

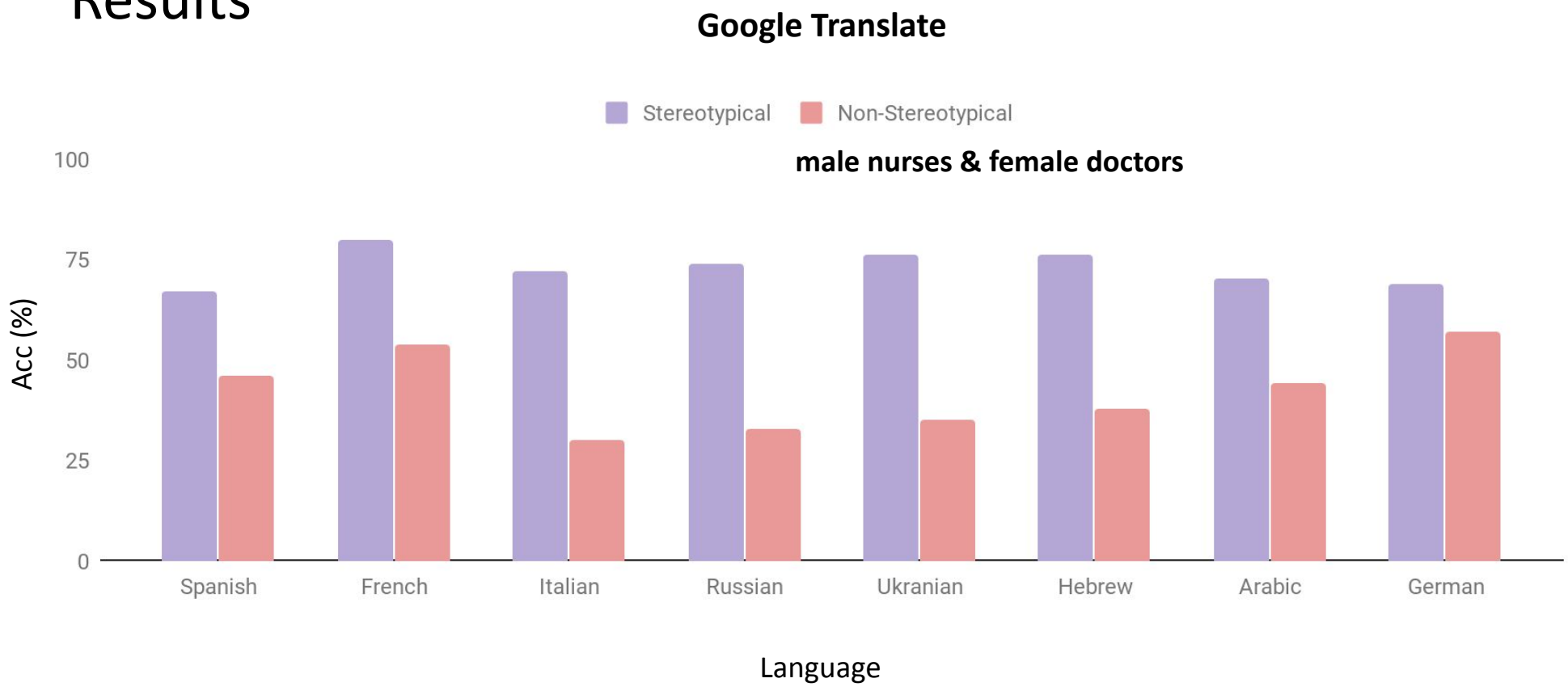
The **doctor** asked the nurse to help her in the procedure.

La **doctora** le pidió a la enfermera que le ayudara con el procedimiento.

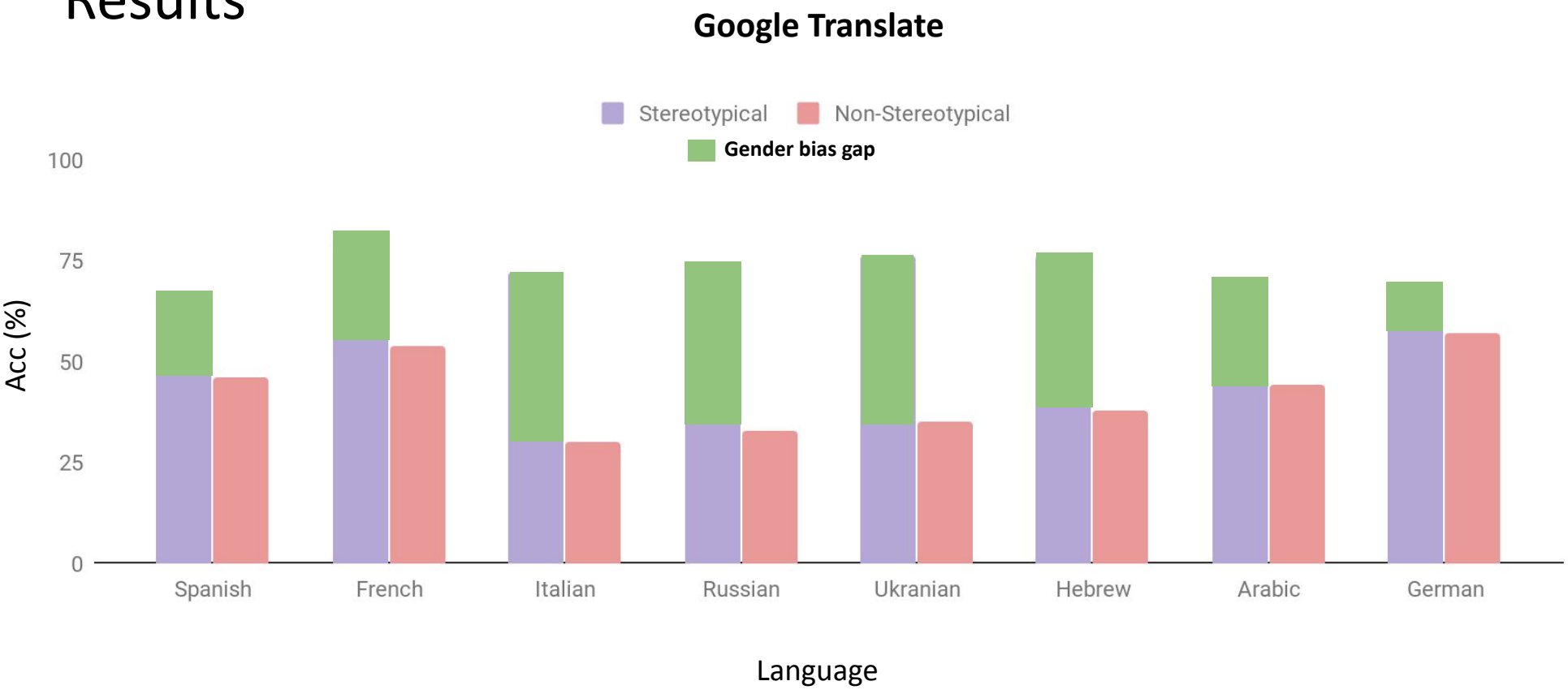
Results


Google Translate
Stereotypical
male doctors & female nurses
100
75
50
25
0
Acc (%)
Spanish
French
Italian
Russian
Ukranian
Hebrew
Arabic
German
Language

Results


Google Translate
Stereotypical
Non-Stereotypical
male nurses & female doctors
100
75
50
25
0
Acc (%)
Spanish
French
Italian
Russian
Ukranian
Hebrew
Arabic
German
Language

Results


Google Translate
Stereotypical
Non-Stereotypical
Gender bias gap
100
75
50
25
0
Acc (%)
Spanish
French
Italian
Russian
Ukranian
Hebrew
Arabic
German
Language

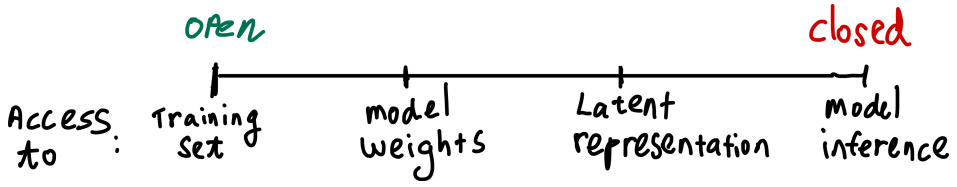
Debaising

- Various works suggest ways to **debias** outputs
  - **Q: Should we debias data, representation, or model?**
- Debiasing models focused on obfuscating **guarded features** [14, 15]
  - E.g., projecting embedding space s.t. it can't predict gender
  - Inherent tradeoff with downstream performance
  - **Come hear Shauli Ravfogel's talk on 17.5**
- **Notoriously hard to do** [16]
  - Higher-order correlations may exist
  - Requires knowing the guarded features in advance
  - Unclear how intrinsic debiasing affects downstream bias

Outline

Explaining model behavior is hard

- We've discussed a **causal model** for explanation
- Over a **spectrum** of accessible model components
  - Training samples
  - Internal model components
  - Latent representations
  - During inference
- Takeaways should be studied **carefully and specifically**



Biases

- **Societal biases** (gender, race, etc.) were found to explain predictions
- Especially **harmful** when model deployed in real-world settings
  - E.g., deciding who to lend money to, or who to incarcerate[1]
- These applications turn correlation into causation
  - E.g., race may be correlated with incarceration
  - but in NLP models race *causes* it

[1] AI is sending people to jail—and getting it wrong

A survey of Bias in NLP [17]

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

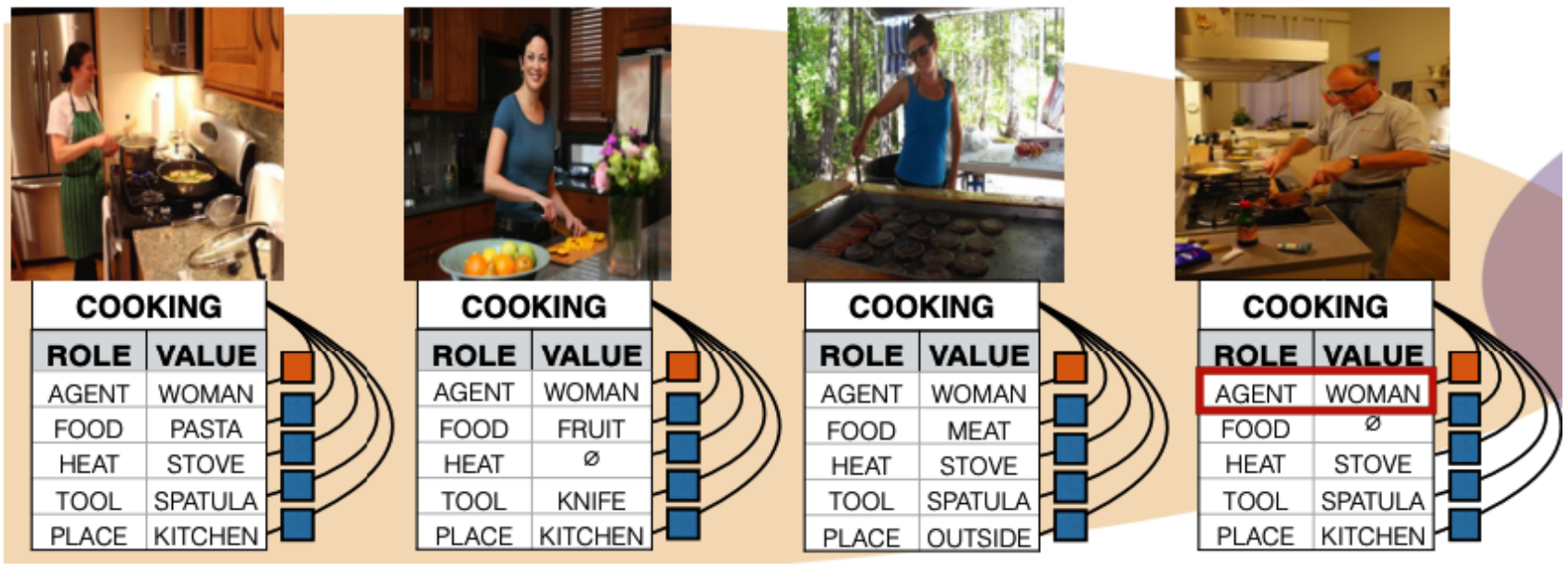
Types of Biases

- "**Allocational harms arise** when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups"

- "**representational harms** arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether."

	Papers	
Category	**Motivation**	**Technique**
Allocational harms	30	4
Stereotyping	50	58
Other representational harms	52	43
Questionable correlations	47	42
Vague/unstated	23	0
Surveys, frameworks, and meta-analyses	20	20

Prevalent in ML Models



On the Limits of Explanations

- Predictions often caused due to **complex** combinations of factors
- Explanations **reduce** the complexity of observed phenomena
- Often do so at the **expense** of correctness

Different Use-Cases for Explanations and interpretability

- **How can I fix my model?**
- **What did the model learn?**
- **How can I be sure that the model is safe?**

Next Week: Artifacts and Spurious Correlations

- How do we formally define **spurious correlations** in NLP?
- How do we find and **eliminate** them?

References I

Eric Wallace, Matt Gardner, and Sameer Singh.
Interpreting predictions of NLP models.
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, 2020.

Pang Wei Koh and Percy Liang.
Understanding black-box predictions via influence functions.
In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning.
What does BERT look at? an analysis of BERT's attention.
In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

References II

Sarthak Jain and Byron C. Wallace.
Attention is not Explanation.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter.
Attention is not not explanation.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

References III


Ian Tenney, Dipanjan Das, and Ellie Pavlick.
BERT rediscovers the classical NLP pipeline.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.


Aviv Slobodkin, Leshem Choshen, and Omri Abend.
Mediators in determining what processing BERT performs first.
In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 86–93, Online, June 2021. Association for Computational Linguistics.


Jiwei Li, Will Monroe, and Dan Jurafsky.
Understanding neural networks through representation erasure.
ArXiv, abs/1612.08220, 2016.

References IV


Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
"why should I trust you?": Explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.


Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai.
Man is to computer programmer as woman is to homemaker? debiasing word embeddings.
In *NIPS*, 2016.

References V

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme.
Gender bias in coreference resolution.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang.
Gender bias in coreference resolution: Evaluation and debiasing methods.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

References VI

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer.
Evaluating gender bias in machine translation.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg.
Null it out: Guarding protected attributes by iterative nullspace projection.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics.

References VII

Yanai Elazar and Yoav Goldberg.
Adversarial removal of demographic attributes from text data.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg.
Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

References VIII


Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.