Advanced NLP

Lecture 5: Interpretability

Dr. Gabriel Stanovsky

April 28, 2023

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

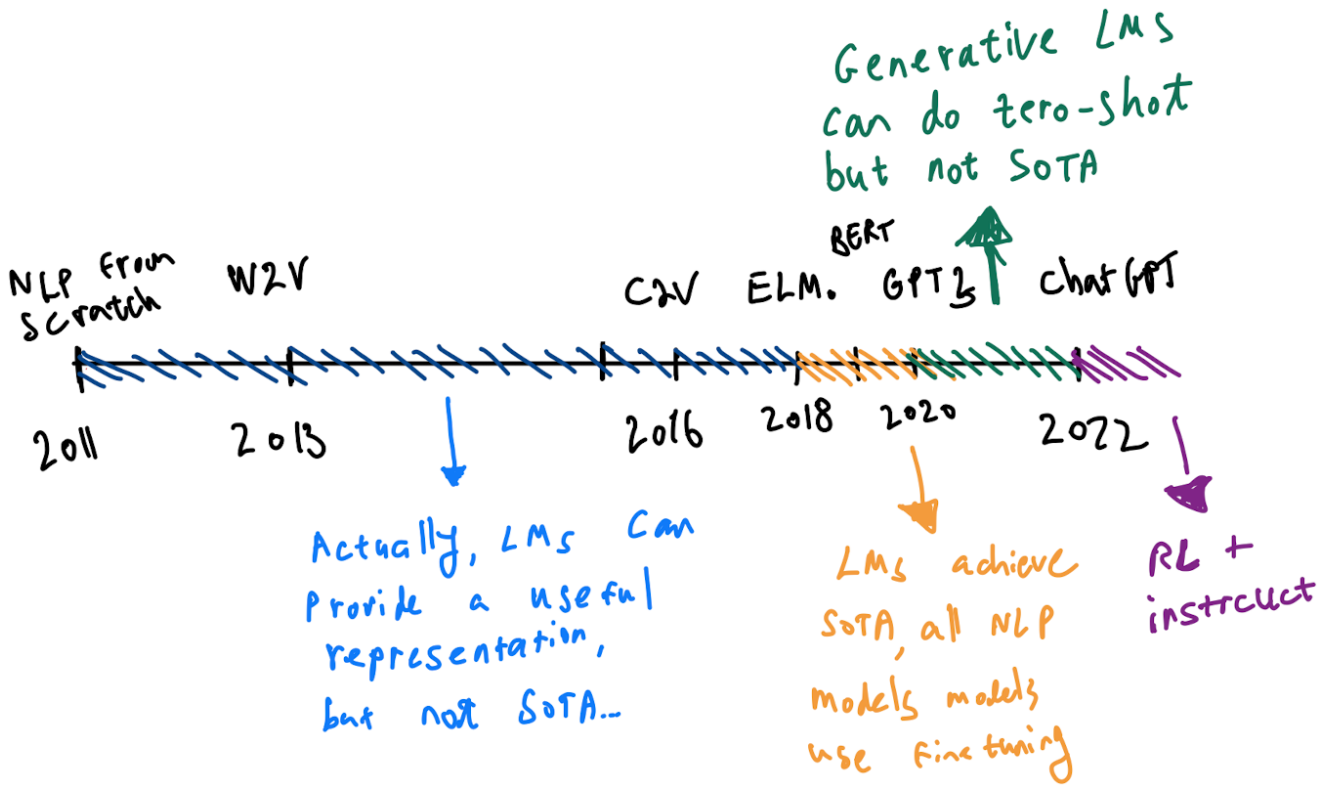
Last Weeks

- **NLP Tasks**
  - Intrinsic: grounding, coreference, entity linking ...
  - Extrinsic: summarization, information extraction, sentiment, ...
- **Crash course in LLMs**
  - Representation learning: dist. hypothesis, tokenization, ...
  - LLMs: finetuning, encoder-decoder, prompting, ... ‘

ANLP Syllabus

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

- **LLMs outperform other methods on practically all tasks**
- Initially through finetuning (the BERT era)
- More recently via zero-shot, instruction tuned LLMs (the GPT era)
- As models get more complex, they get harder to explain



- **LLMs outperform other methods on practically all tasks**
- Initially through finetuning (the BERT era)
- More recently via zero-shot, instruction tuned LLMs (the GPT era)

Today

How are LLMs achieving good performance?

Or:

Q: Why did the model make a certain prediction?

Today

1. How to explain model behavior?
2. ... via training samples
3. ... via model weights and architecture
4. ... via latent representation
5. Conclusion

Outline

Q: Why did the model make this prediction?

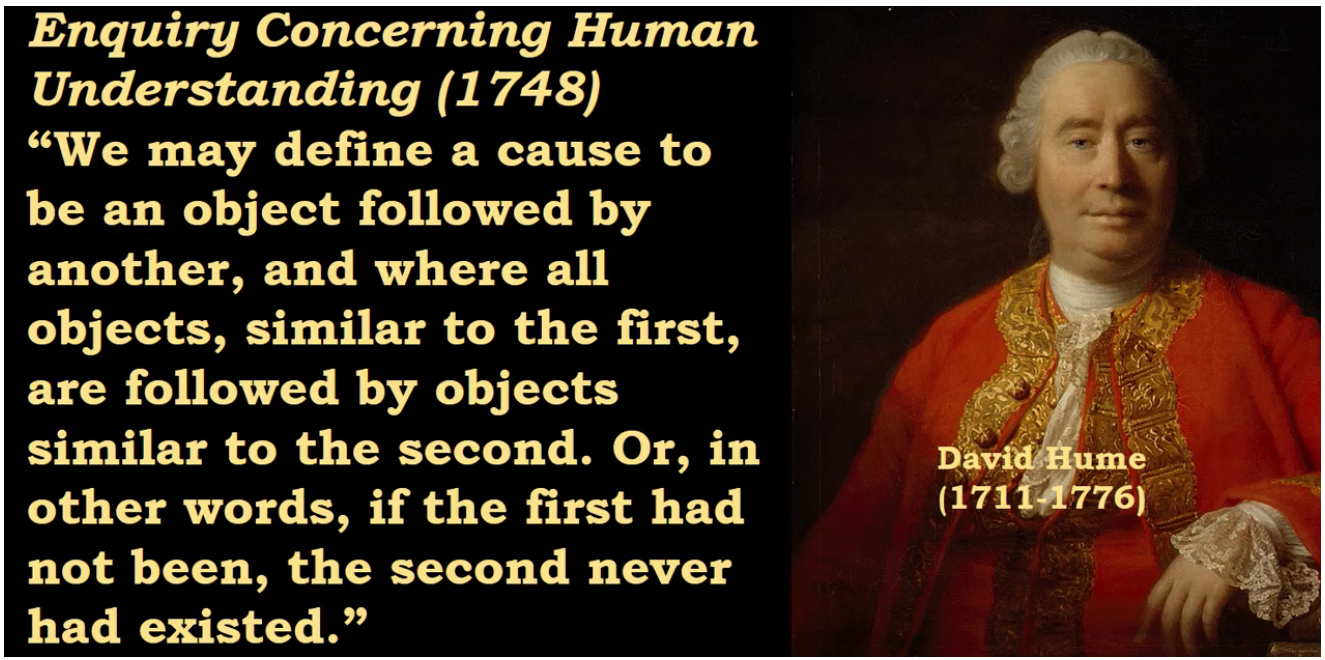
- The scientific method can't properly answer **why** questions
  - Instead focusing on reduction
  - E.g., a complex process is decomposed to simpler events A, B, C
  - where $A \Rightarrow B \Rightarrow C$
  - **Q: Why did C happen? Why did A happen?**
- However, many reasons to still try to answer this question
  - Debugging during development
  - Regulation before deployment
  - Explaining during inference

Causality

- **Causality** may be science's best approximation of explaining **why**
  - E.g., taking aspirin may cause blood pressure to drop
  - **Q: How do we formally define that X causes Y?**

Causality

- **Causality** may be science's best approximation of explaining **why**
  - E.g., taking aspirin may cause blood pressure to drop
  - **Q: How do we formally define that X causes Y?**

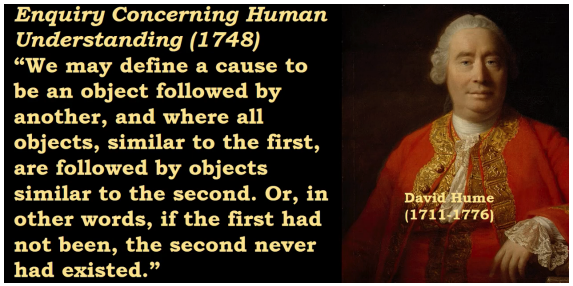


Causality: Counterfactuals and Interventions

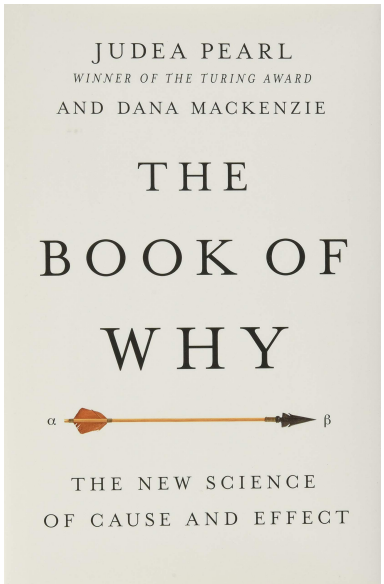
- **Q: How far back in the casual chain do you go?**

Causality

- **Causality** may be science's best approximation of explaining **why**



- **Judea Pearl** mathematically formalized causality and counterfactuals
  - I **think** that assuming the world has causation is an axiom



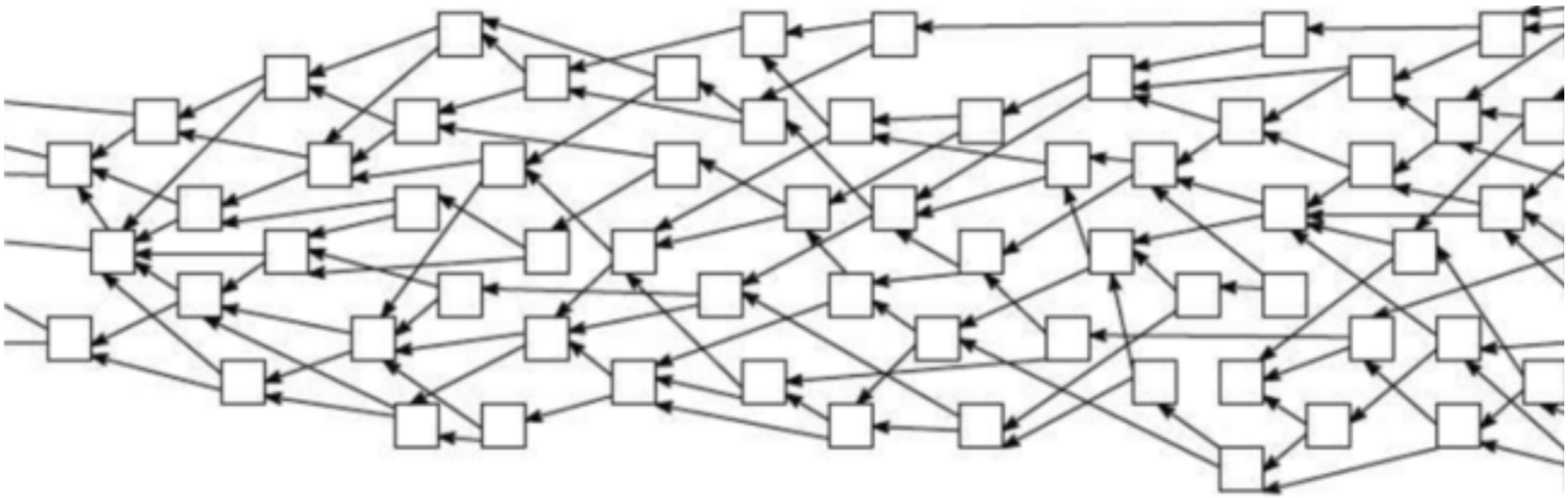
link

Measuring Counterfactuals through Intervention

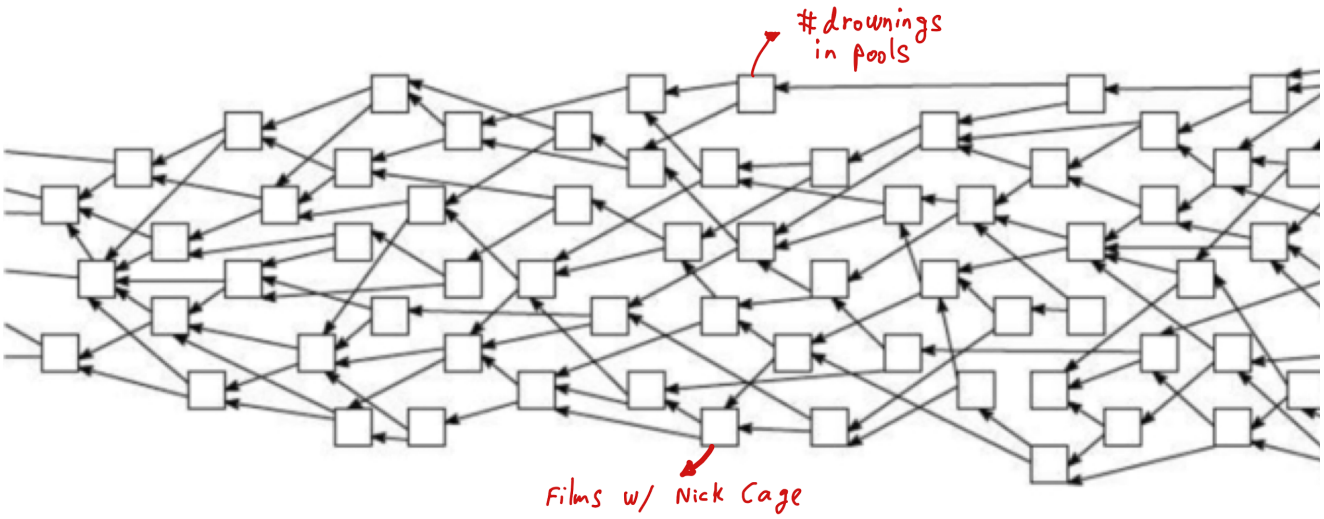
- $A \Rightarrow B$ if **all other things being equal**, changing A affects B
  - Intervening to take / don't take aspirin and measure BP
  - This simulates **rewinding** the world state and intervening
- Computer models lend themselves to interventions
  - We'll see this in practice

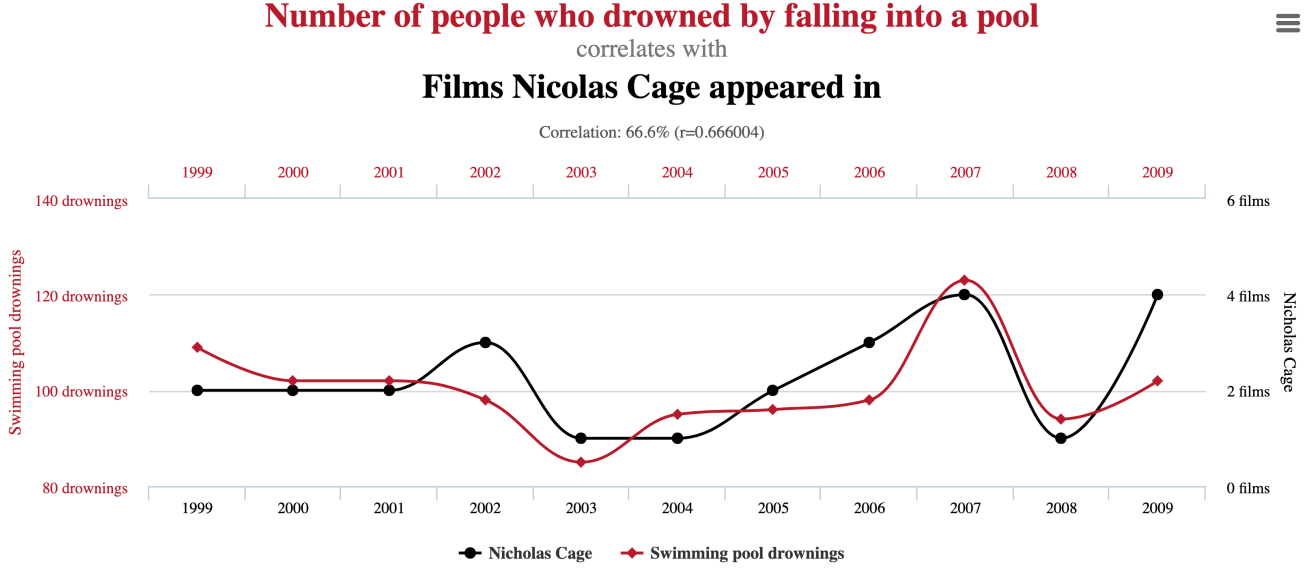
Causal Graph as a Model for the World

- Put all related events in a graph
  - E.g., for model performance: training examples, archeitecture, ...
- Then examine their causal relations

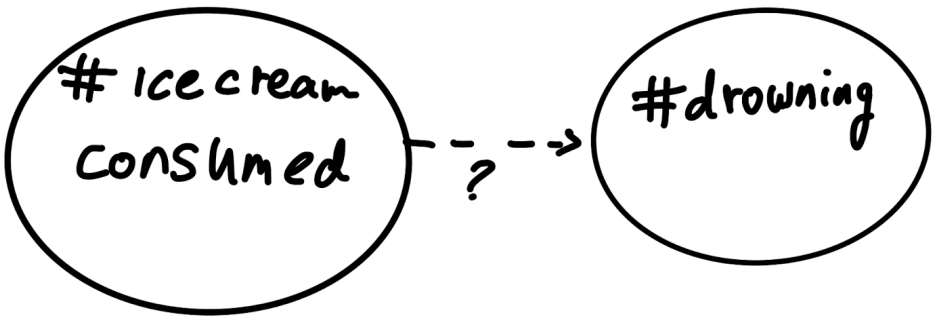


Spurious Correlations: Mistaking Correlation for Causation

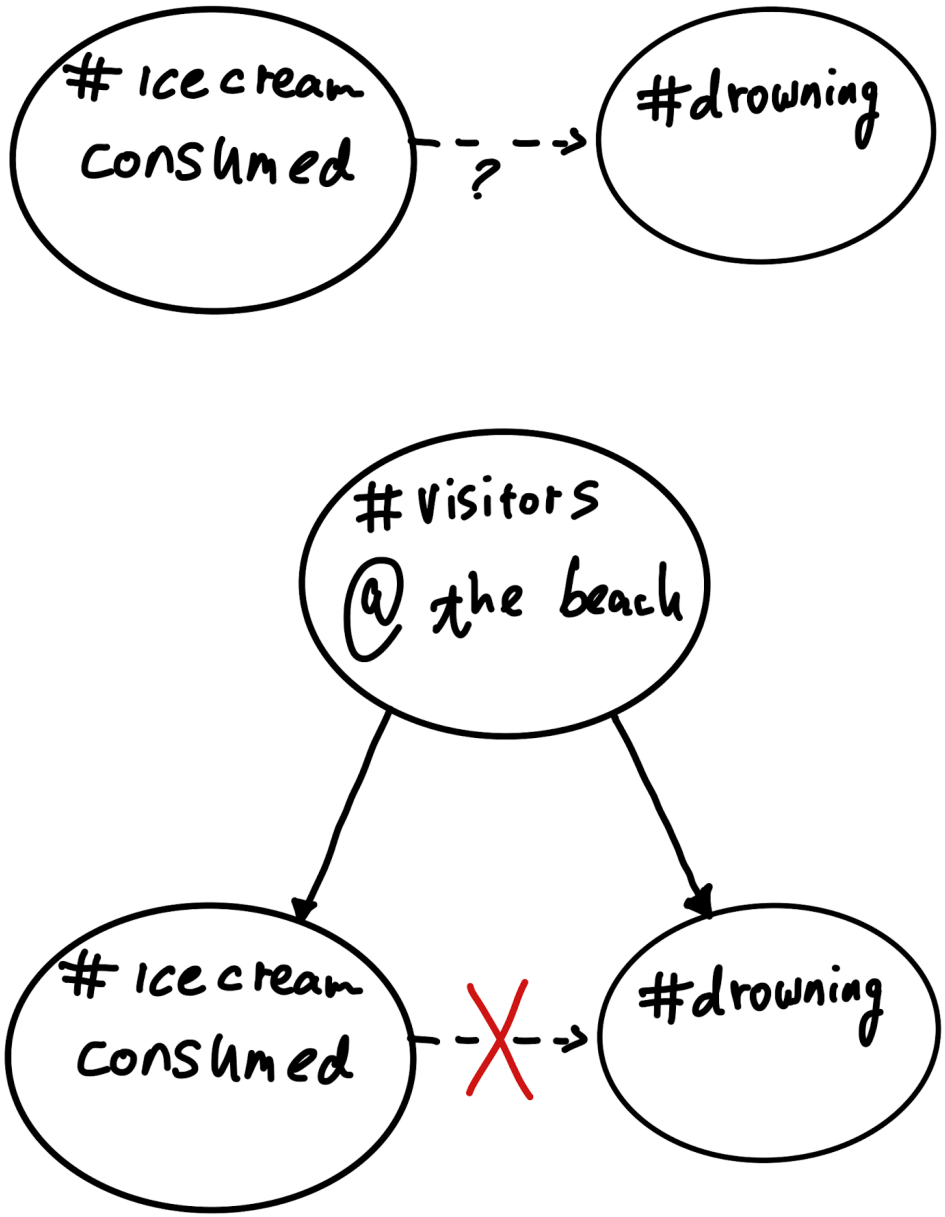




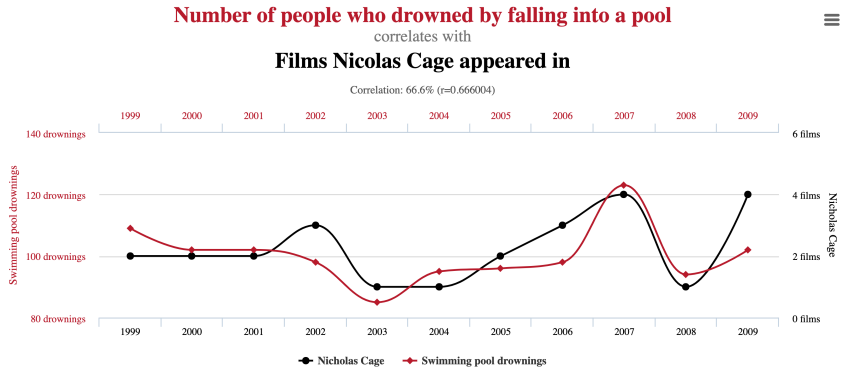
Confounding Variables: Choosing what to Represent

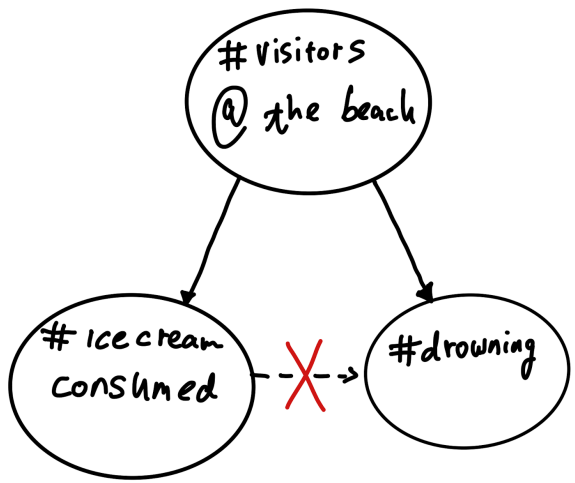


Confounding Variables: Choosing what to Represent



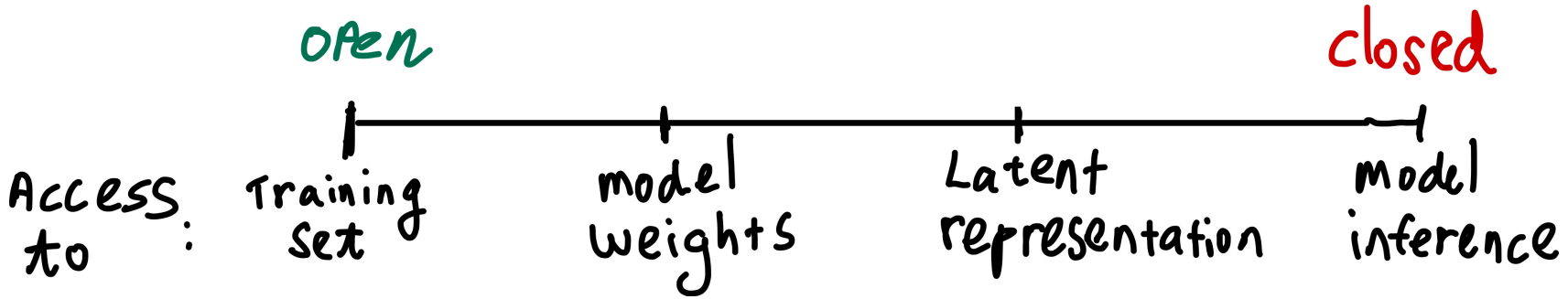
Intervention rules out causation in confounding and spurious correlations





Back to explaining model behavior

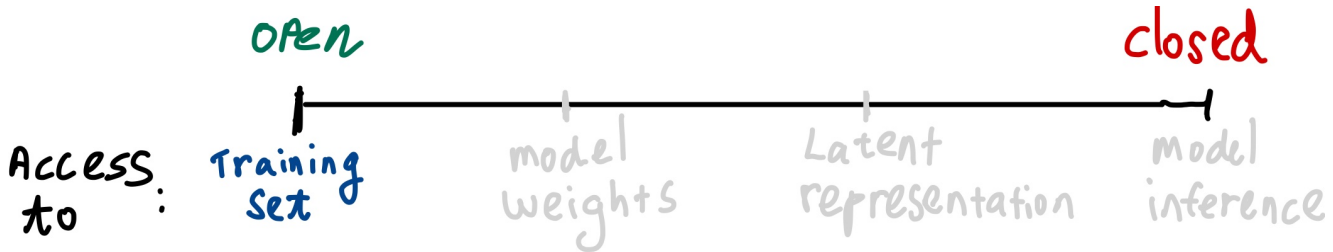
- Factors governing model performance **not fully known or available**
- Especially true for proprietary & close models
- We'll discuss a spectrum of available resources



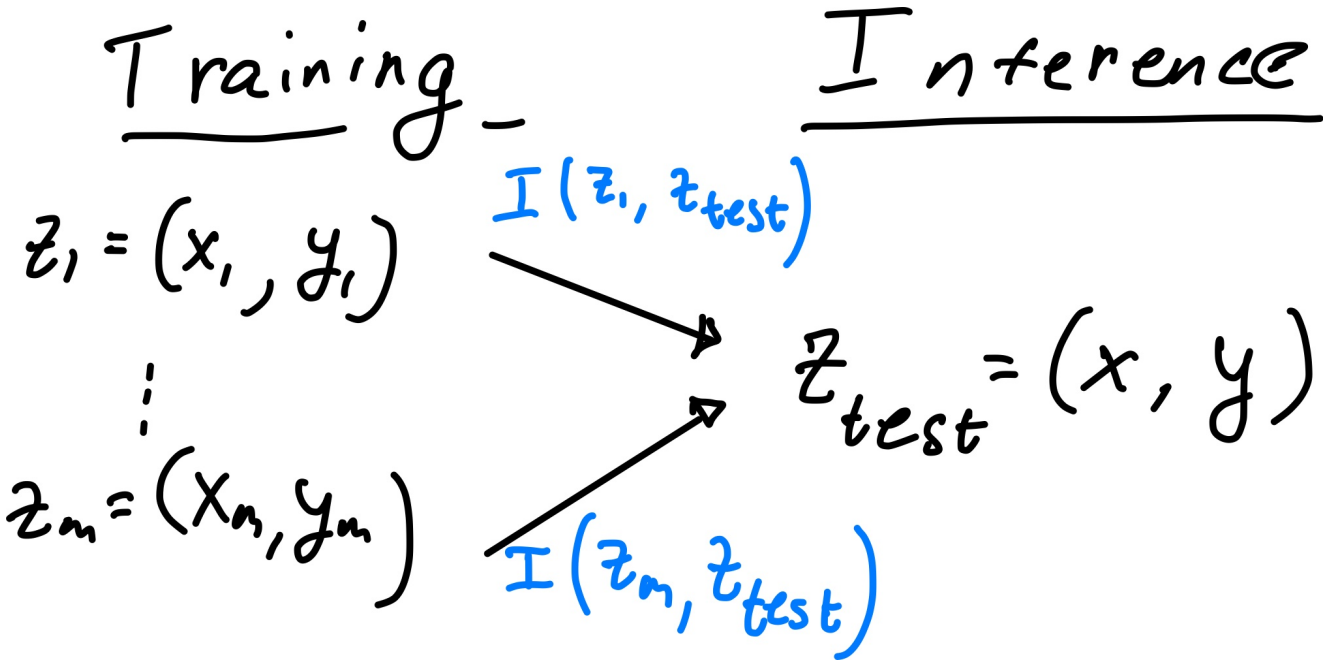
Outline

What in the training set caused this prediction?

- Assume we have access to the training set
- We'd like to know the causal effect between them and a prediction

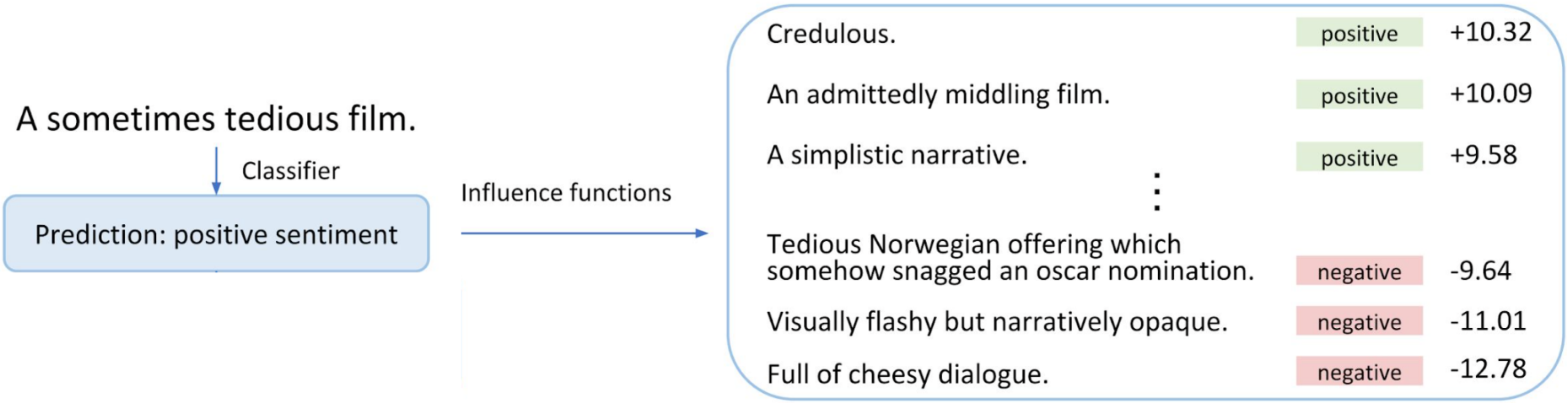


Causal Graph for Influence Functions in Sentiment Analysis



- $I(z_i, z_{test})$ quantifies how much z_i causes z_{test} [1]
- A test sample induces ranking of the **influence** of all training samples

Influence Ranking



Han et al., 2020 [2]

- Explain performance through training examples
- Debugging by finding incorrect gold labels
- Identify biases

Quantifying how much z_i causes z_{test}

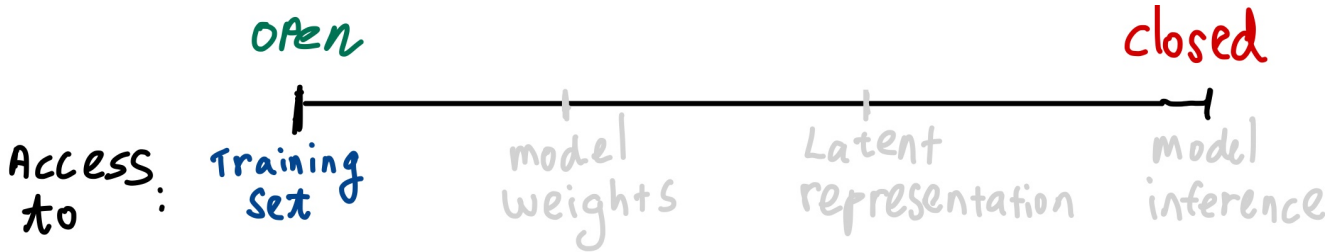
- **Q: What's the counterfactual approach?**
  - Define $T = \{z_1, \ldots, z_m\}$; $T_{-z_i} = T \setminus \{z_i\}$
  - Denote: θ from training on T; θ_{-z_i} from training on T_{-z_i}
  - $I(z_i, z_{test}) = L(z_{test}, \theta_{-z_i}) - L(z_{test}, \theta)$
- Requires training m different models
- Instead, approximating based on the following observation[3] :

$$\theta_{-z_i} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{m} \left(\sum_{k=1}^{m} L(z_k, \theta) \right) - \frac{1}{m} L(z_i, \theta)$$

  - This approximation is still expensive: $O(m \cdot d^2 + d^3)$
  - Assumes loss is smooth

Influence Functions

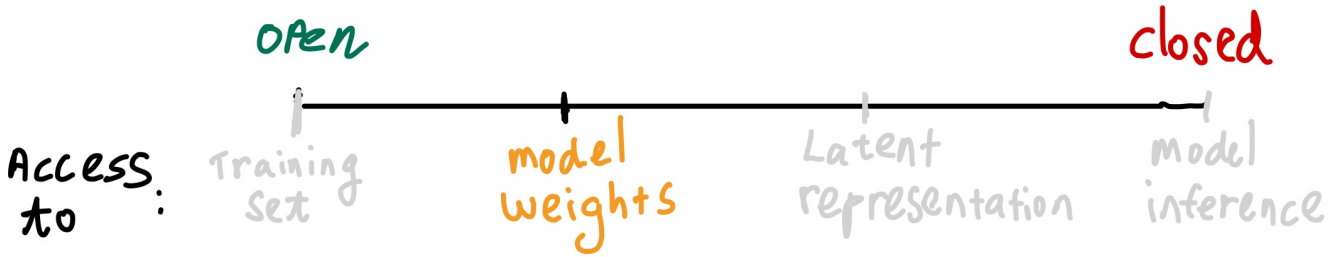
- **Was shown empirically useful on many domains**
- **However, doesn't scale well** [4]



Outline

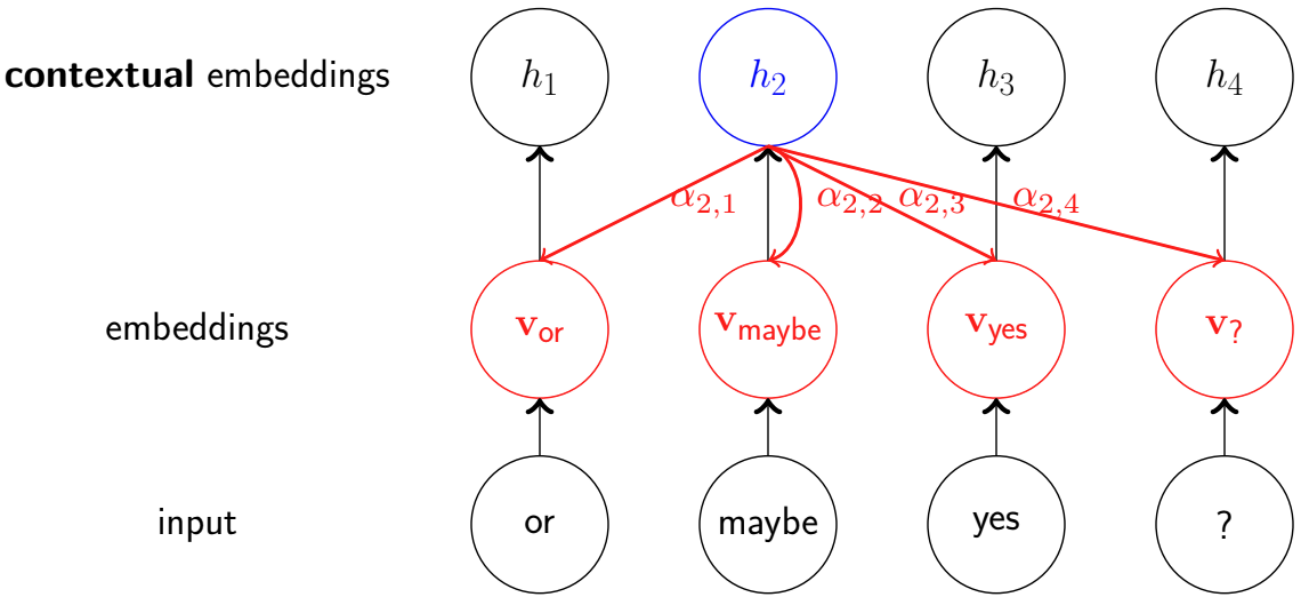
What in the network weights caused this prediction?

- Assume we have access to architecture and trained weights
- We'd like to know the causal effect their properties and a prediction



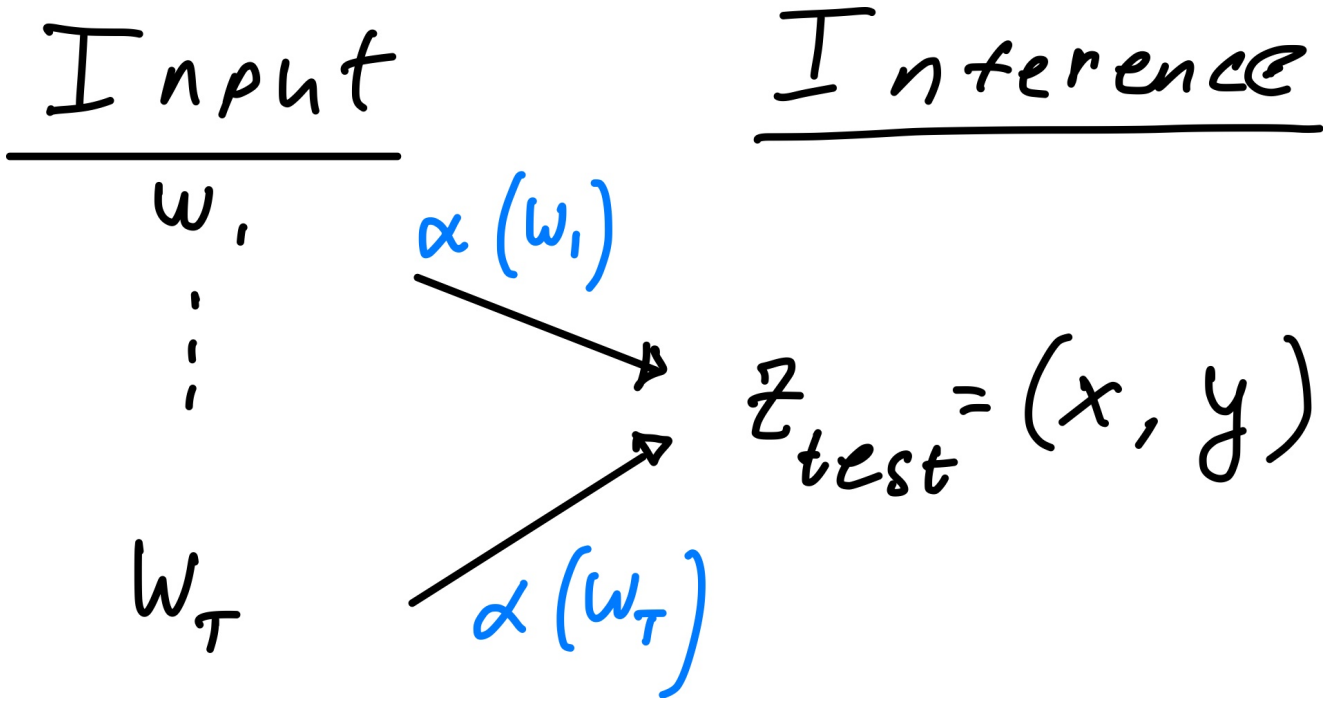
Self-Attention as an Explanation for Prediction

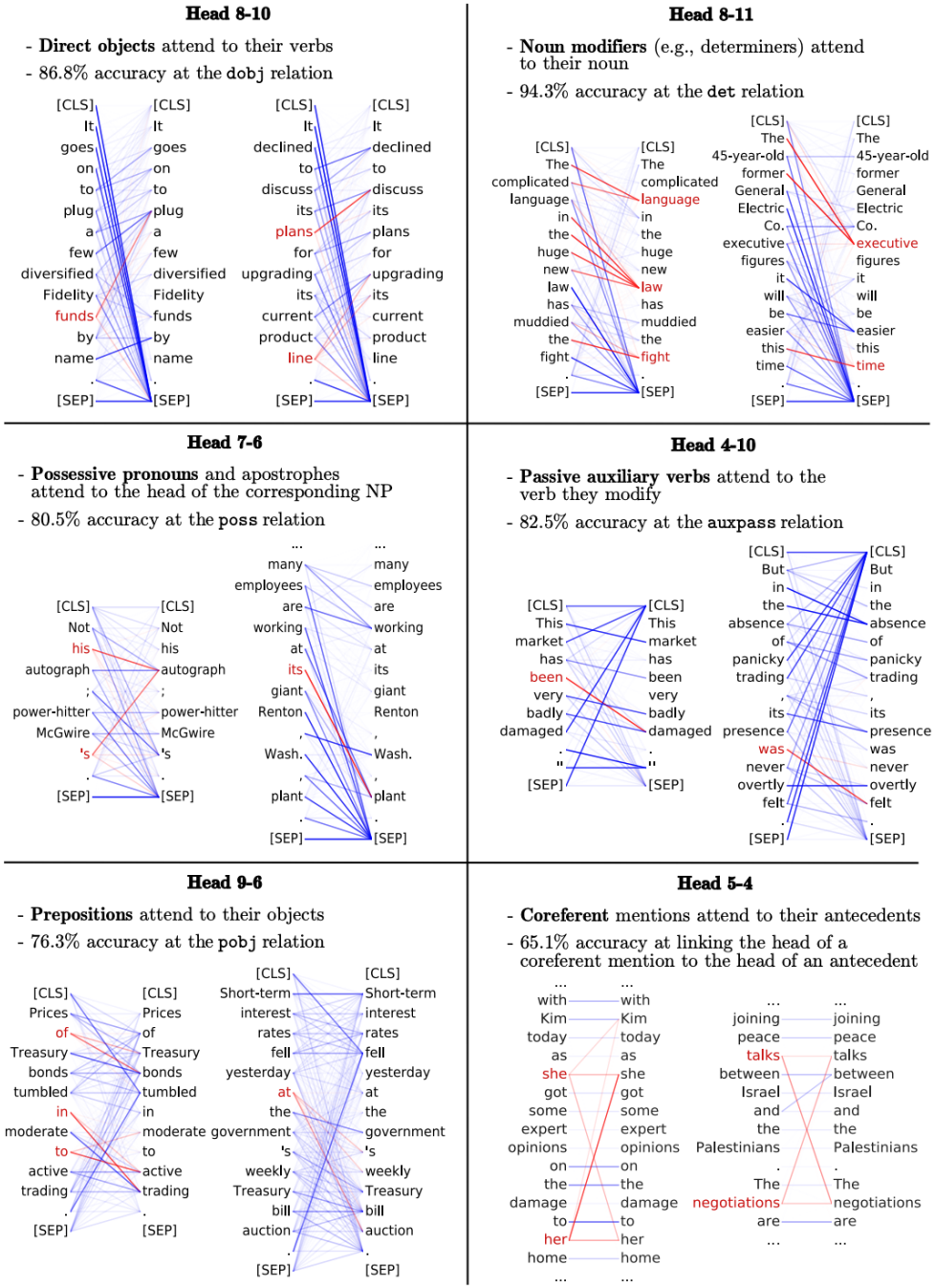
- Recall the **self-attention** mechanism from Lecture 3
- A **very important** component in the Transformer



- Self-attention seems convenient for explaining predictions
  - A word in the input **caused** this prediction

> *"Attention provides an important way to explain the workings of neural models"* [5]

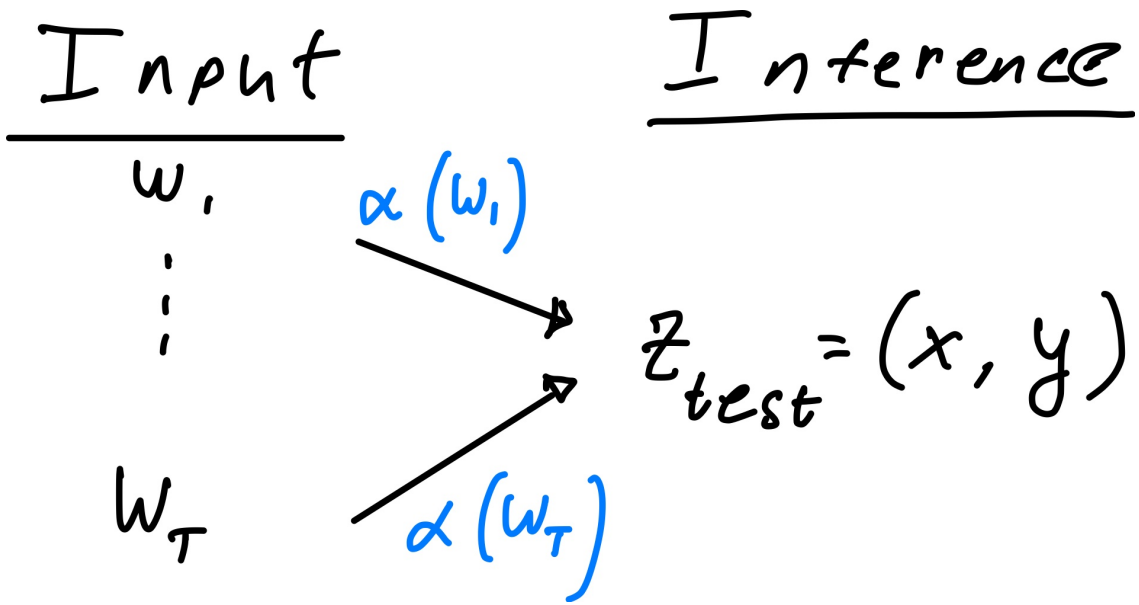




Attention weights for different words during MLM [6]

Attention is not explanation?

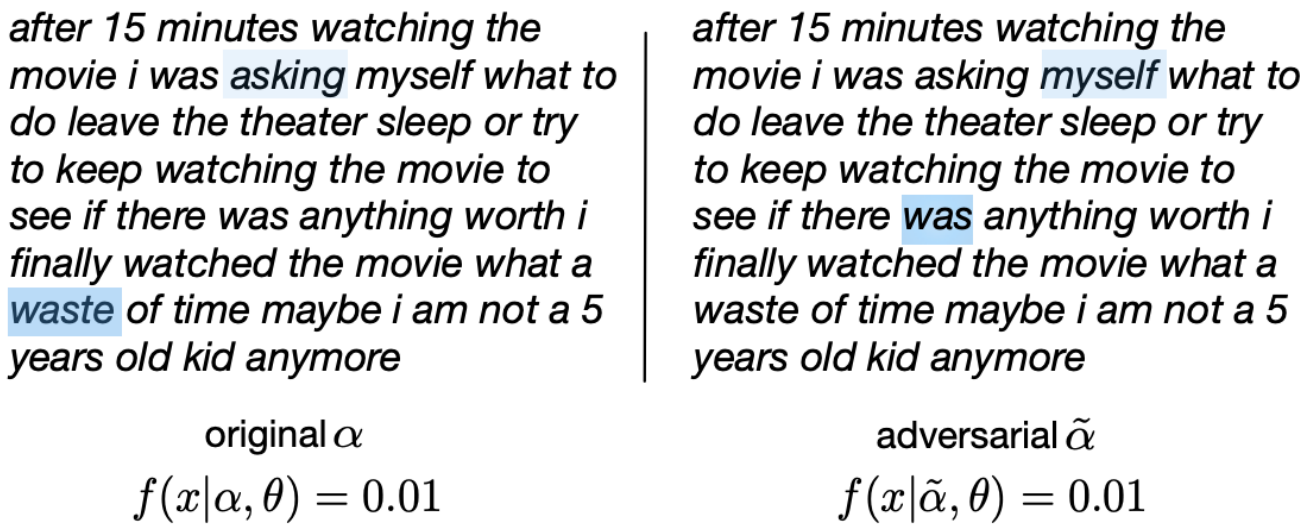
- **Q: What's the counterfactual approach for testing causation?**
  - Changing attention should **change prediction**
  - ... or provide an **equally valid alternative explanation**



Attention is not explanation?

Adversarial Explanations as an optimization problem [7]

Given a trained model f defined by $\theta \in \mathbb{R}^d$, attention weights $\alpha \in \mathbb{R}^T$, and examples $S_m = x_1, \ldots, x_m$, find $\hat{\alpha}$ which **maximizes** the difference between $\{f(x_i|\theta, \alpha) : x_i \in S_m\}$ and $\{f(x_i|\theta, \hat{\alpha}) : x_i \in S_m\}$ while **minimizing** the difference between α and $\hat{\alpha}$



Heatmap of attention weights induced over a negative movie review. Original model attention (left) vs. an adversarially constructed attention (right). Despite their differences, both yield the same prediction (0.01).

Attention is Not Not Explanation

- Counterfactual attention argument **was questioned** in [8][1]
  - Attention gives an explanation, not *the* explanation
  - Many ways to reach a 0-1 label, not all should make sense
  - Counterfactuals disregard attention the model chose
- Discussion echoes the complexity defining a **good explanation** [9]

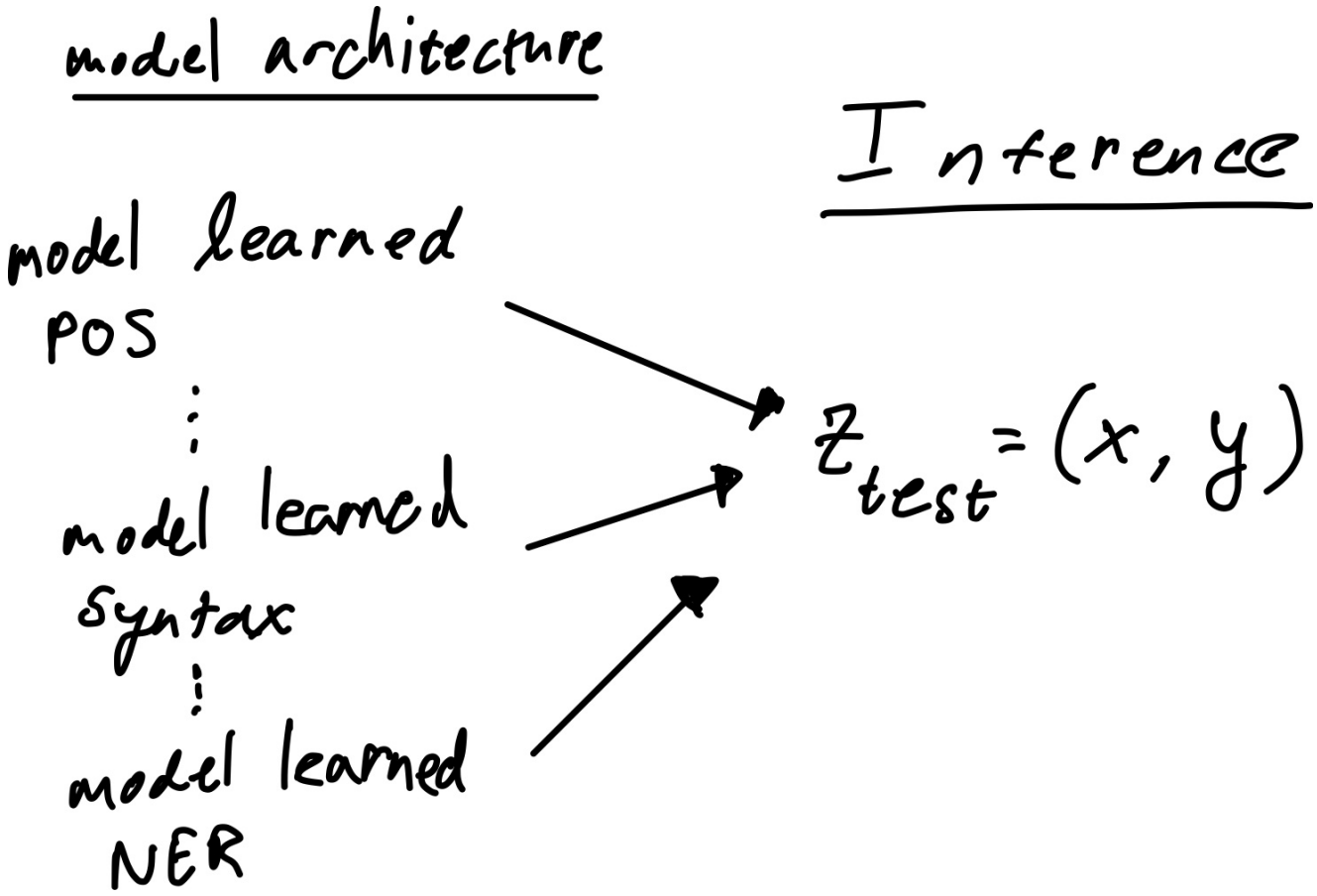
[1] Yuval's blog-post

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



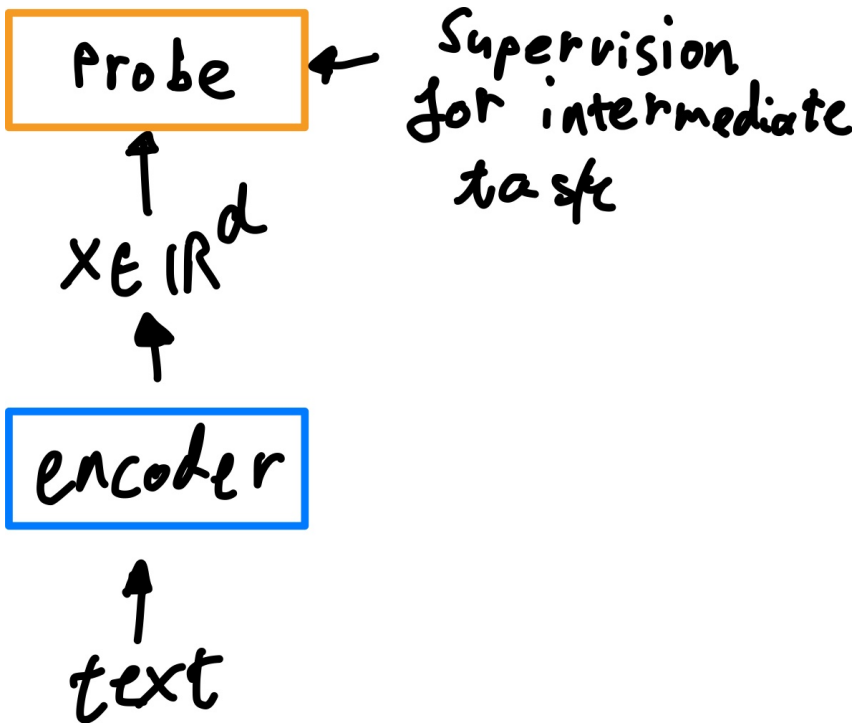
- Many works propose performance caused by learning **intrinsic tasks**
  - Recall that LLMs don't train on them **explicitly**
- **Q: How can they still learn intrinsic tasks?**

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**

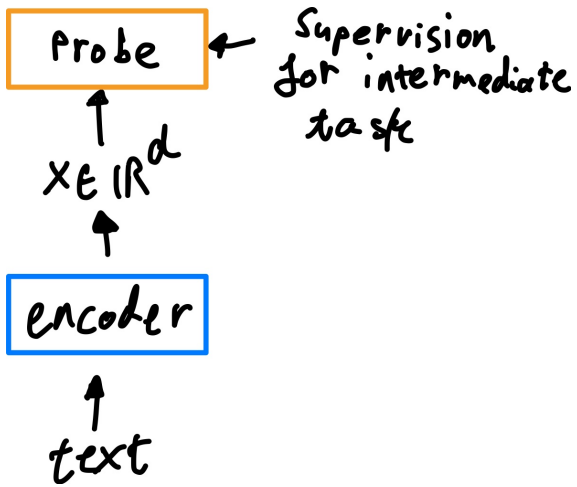
Probing

- 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



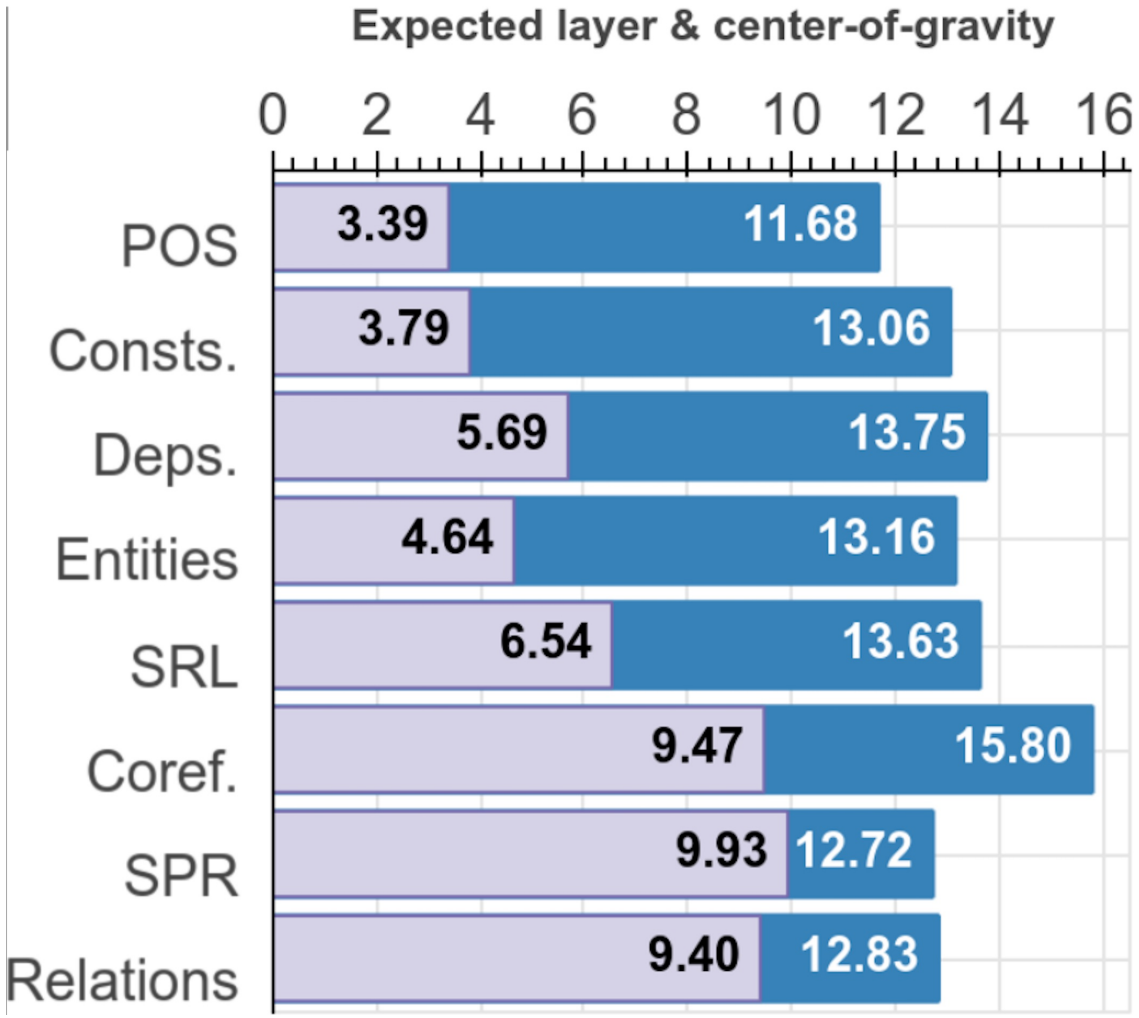
Probing

- **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



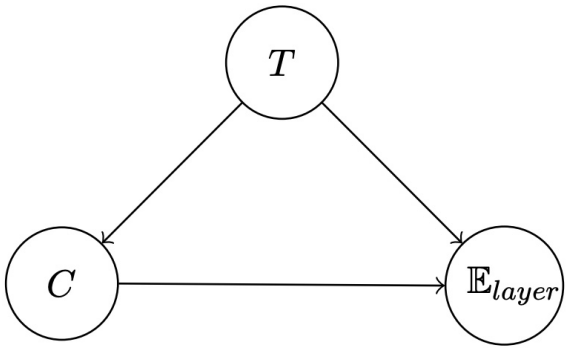
BERT Rediscovers the NLP Pipeline

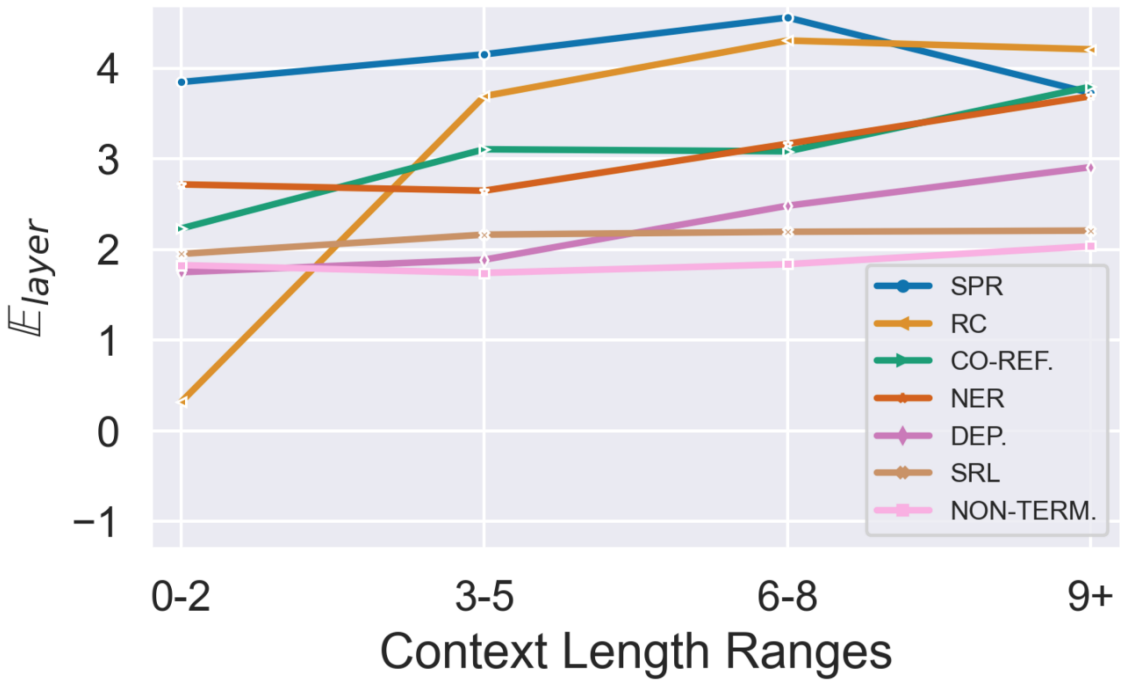
- Layers in BERT seem to **learn different intermediate tasks** [10]
- . . . and they so in the **traditional linguistic order**



Context Length as a confounding factor

- Length proposed as a **confounding factor** [11]





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
- Takeaways should be studied **carefully and specifically**
- **Next Week:**
  - Explaining a closed model behaviour
  - Social biases, contrastive explanations
  - Implications for wide-spread adoption, regulation

References I

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

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov.
Explaining black box predictions and unveiling data artifacts through influence functions.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online, July 2020. Association for Computational Linguistics.

Frank R Hampel.
The influence curve and its role in robust estimation.
Journal of the american statistical association, 69(346):383–393, 1974.

References II


Samyadeep Basu, Phillip E. Pope, and Soheil Feizi.
Influence functions in deep learning are fragile.
ArXiv, abs/2006.14651, 2020.


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


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 III

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 IV


Zachary C Lipton.
The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.
Queue, 16(3):31–57, 2018.


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.