CSDS 452 Causality and Machine Learning

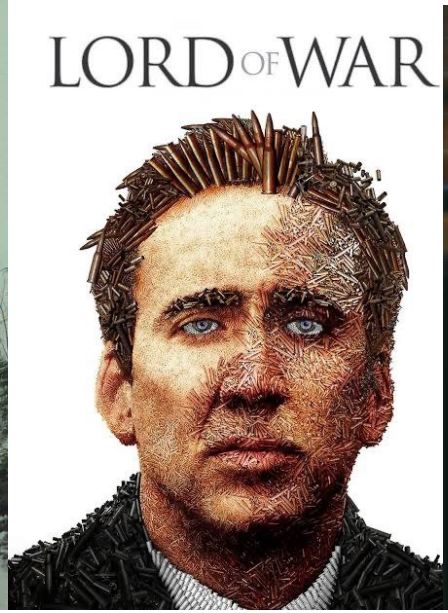
Lecture 19: Causal NLP

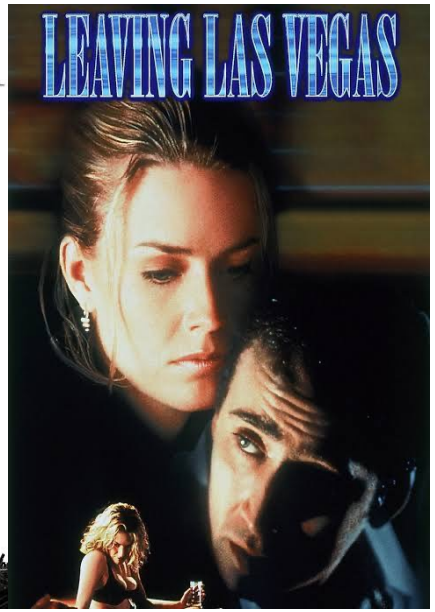
Instructor: Jing Ma

Fall 2024, CDS@CWRU

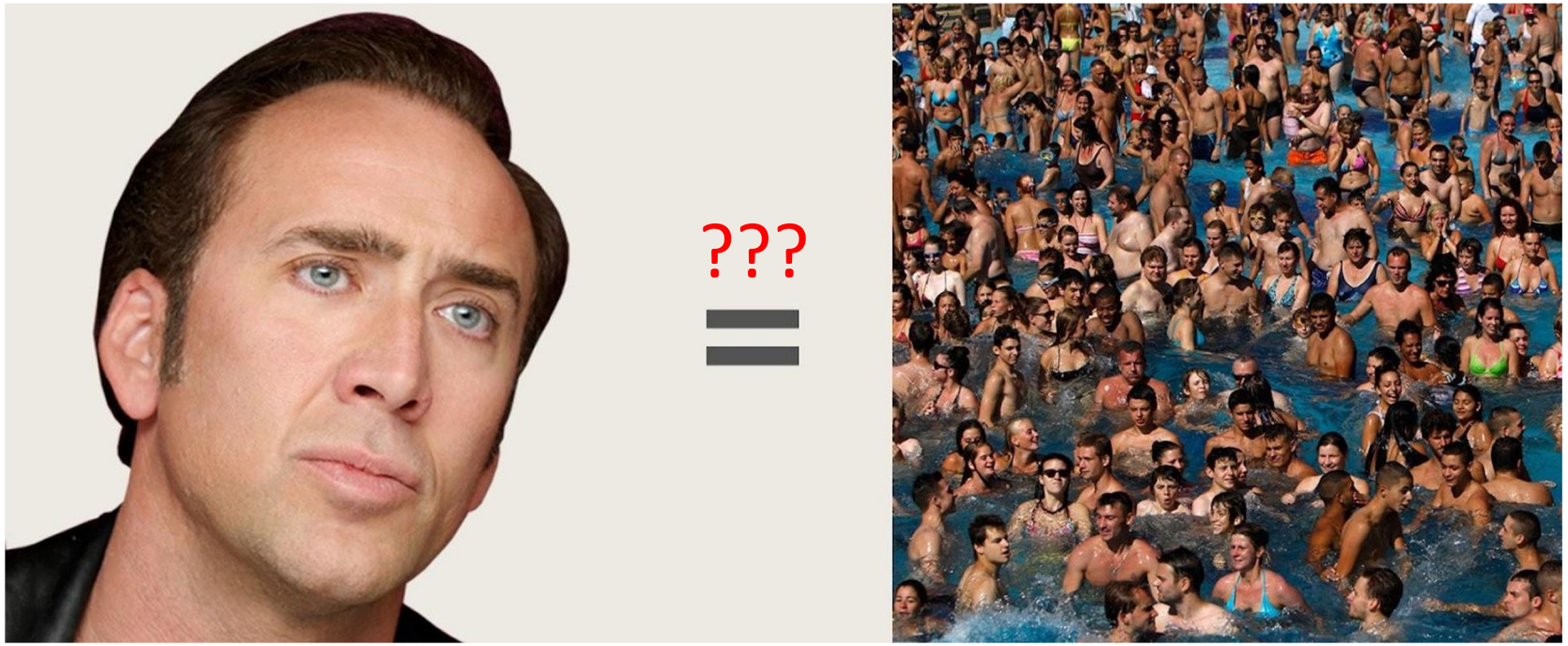
Story still begins with weird correlations

Nicolas Cage

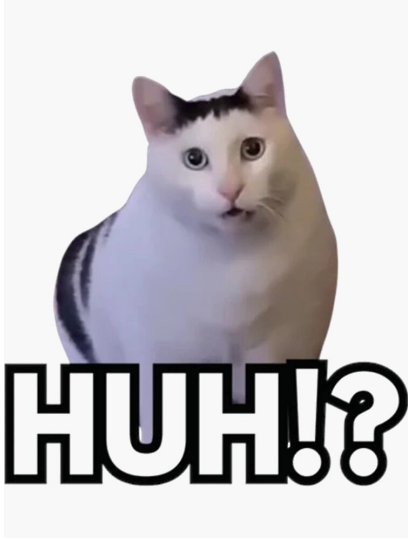


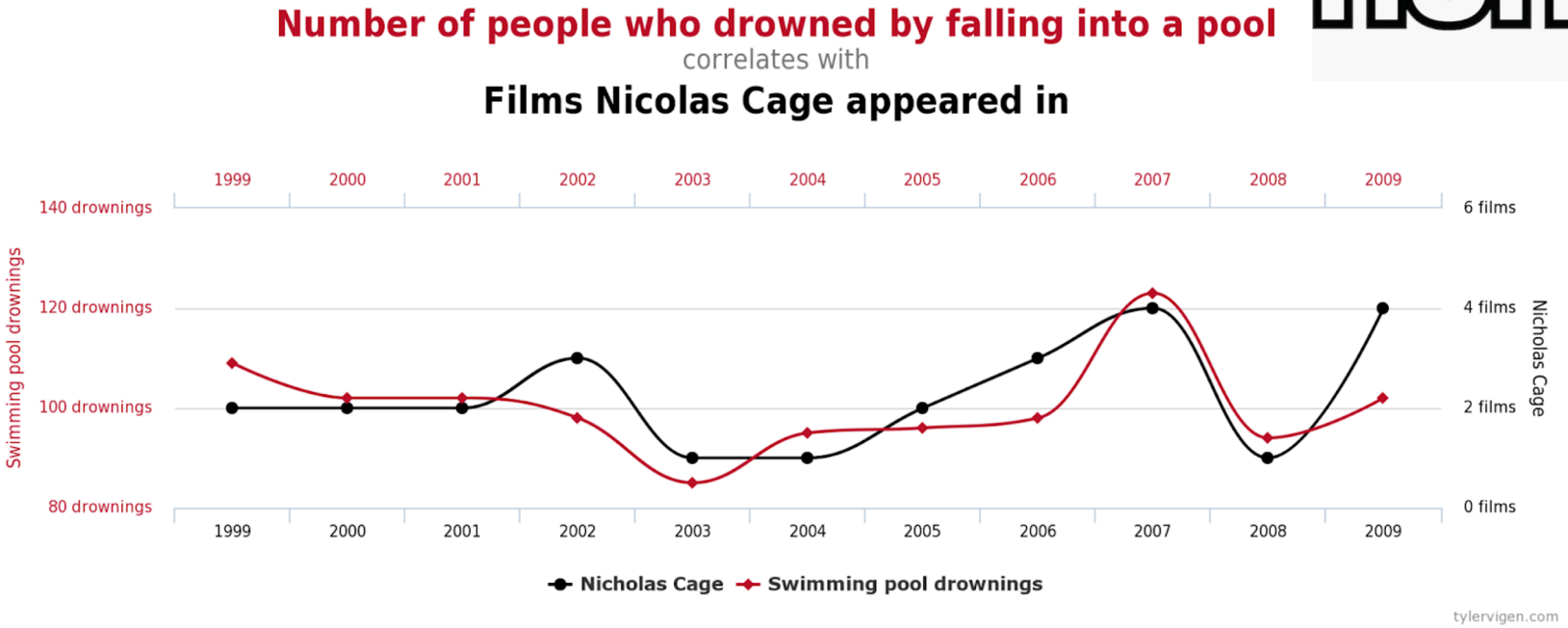


Q: Does Nicolas Cage cause drownings?



A: An ML model might think so...





Causality makes us *more insightful NLP researchers*.

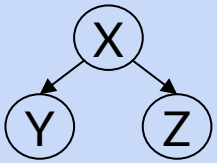
3 Benefits:

1. Better models
2. Deeper understanding, with larger social impact
3. More scientific practice

“Vocabulary”

Causality in the Real World

E.g., “does public sentiment affect policies?”



Traditional causality question:

- Does smoking causes cancer?

Example NLP questions:

- Does gender directly affect citations?
- How does Twitter sentiment affect COVID policies?
- Why does certain phenomenon happen in language evolution?

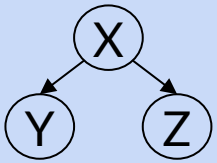
Tracks:

- *Computational Social Science and Cultural Analytics*
- *NLP Applications*
- *Linguistic Theories, Cognitive Modeling, and Psycholinguistics, …*

“Vocabulary”

Causality in the Real World

E.g., “does public sentiment affect policies?”



Causality of LM Behavior

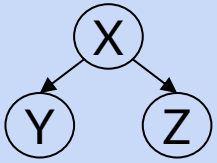
Which neuron affects what?

Does the model capture the data generation causal process?

Why do certain prompts work?

“Vocabulary”

Real-World Causality

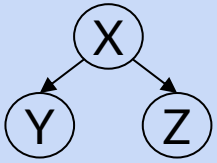


LM Causality

“Vocabulary”

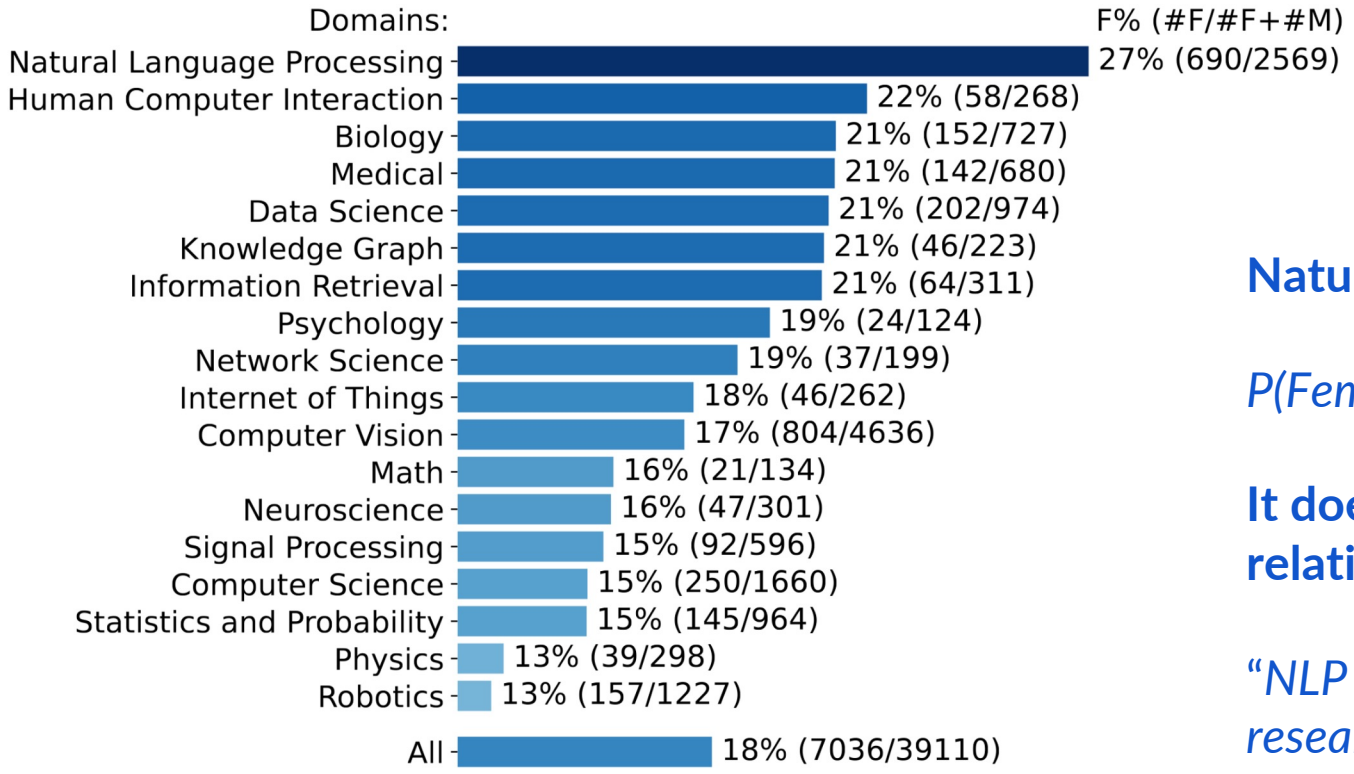
Tools can be used for

Real-World Causality



LM Causality

Example for Prediction/Correlation:



Nature of these quantities:

P(Female | Domain=d)

It does not mean causal relations like:

"NLP is easier for female researchers than Signal Processing. "

Voices of Her: Analyzing Gender Differences in the AI Publication World

Yiwen Ding*, Jiarui Liu*, Zhiheng Lyu*, Kun Zhang, Bernhard Schoelkopf, Zhijing Jin†, Rada Mihalcea† **Paper 2022**

Example for Correlation: LM

Input

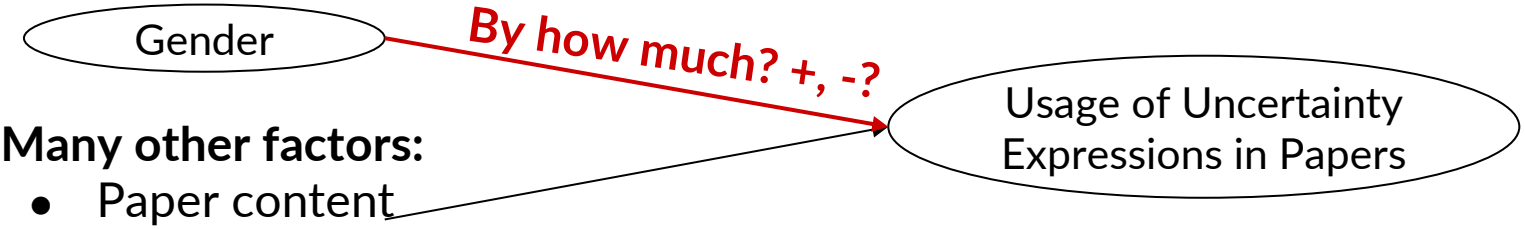
My friend is a doctor. He/She/They …

Prediction Probability

P(“**he**” | “**doctor**”) = x
P(“**she**” | “**doctor**”) = y
P(“**they**” | “**doctor**”) = z
…

This reflects the correlations that LMs captured, not necessarily causation.

Gender Counterfactual



Many other factors:

- Paper content
- Publication year
- Author citations
- …

*"This **may** indicate that …"*

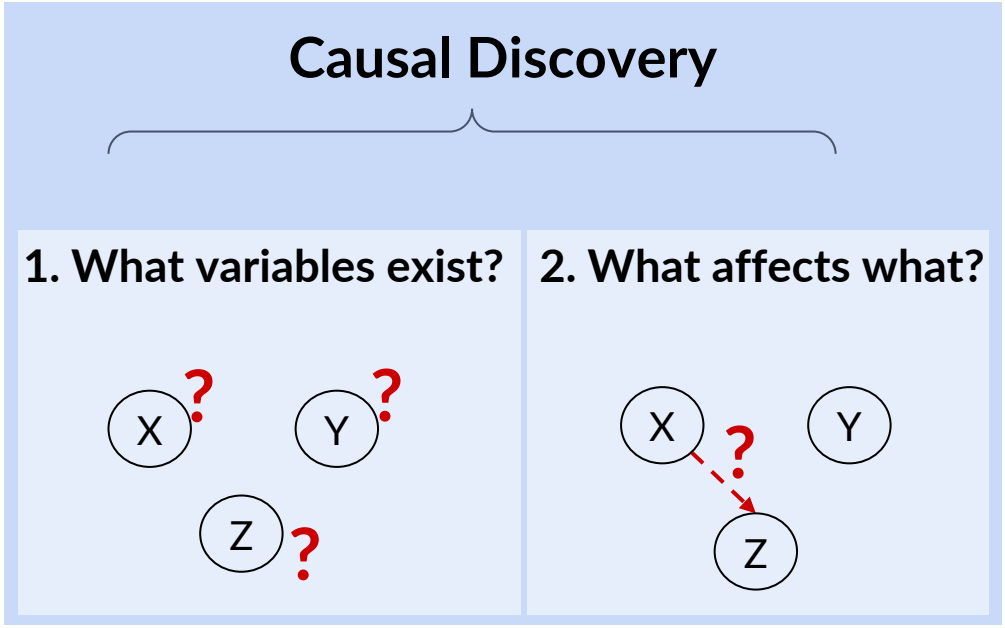
Research Question to Answer:

If we take out all the **male-authored papers**,
had the **paper and author characteristics been the same**,
if we **change the gender to female**,
by how much will the **usage of uncertainty expressions** change?

Editing a Woman's Voice
Anna Costello, Ekaterina Fedorova, Zhijing Jin, Rada Mihalcea.
arxiv.org/abs/2212.02581

Adapting text embeddings for causal inference
Victor Veitch, Dhanya Sridhar, David Blei.
UAI 2022 (Oral)

Causal Discovery &
Causal Representation Learning



Where Are We?

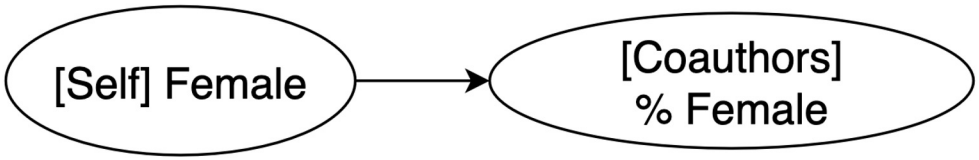
i.i.d. data?	Parametric constraints?	Latent confounders?	What can we get?
Yes	No	No	(Different types of) equivalence class
		Yes	
	Yes	No	Unique identifiability (under structural conditions)
		Yes	
Non-I, but I.D.	No/Yes	No	?
		Yes	
I., but non-I.D.	No	No	
	Yes		
	No	Yes	
	Yes		

Motivation Example: Causal Discovery for NLP

AI Scholar Dataset
github.com/causalNLP/ai-scholar

78K
AI Scholars
Self statistics and co-author statistics

(Potential) Causal Discovery Questions



More:
- Check the diversity-innovation paradox
- Check various biases
- …

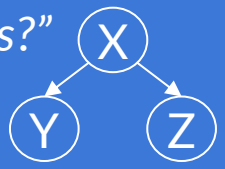
Are LMs equipped with the skills to understand real-world causality?

“Vocabulary”

Do LMs* Understand “Real-World Causality”?

Causality in the Real World

E.g., “does public sentiment affect policies?”



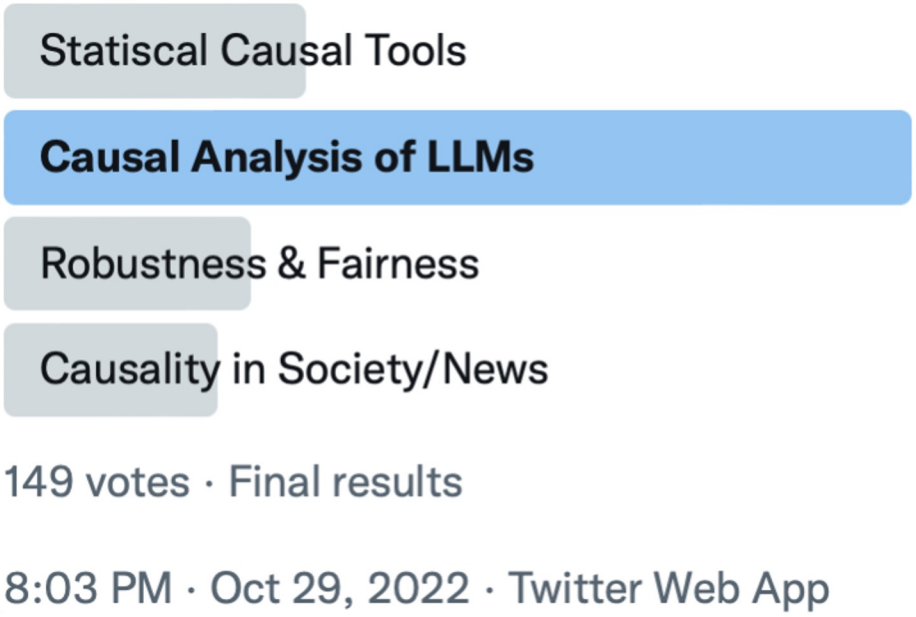
LMs as a Window to "Real-World Causality"

What are the things very difficult for (even larger) LMs to do?

- **Forecasting:** any conversation that needs some **estimations about the future**
- **Hardcore scientific questions:** any reasoning over causality that scientists are investigating
- **Summarization:** how to distill the best information for certain use cases (for whom, for what use cases, what are people's cognitive budget)
- **Personalization:** give suggestions to certain people in certain cases (what they need, what are world constraints, …)

… …

Popular demand



Motivation

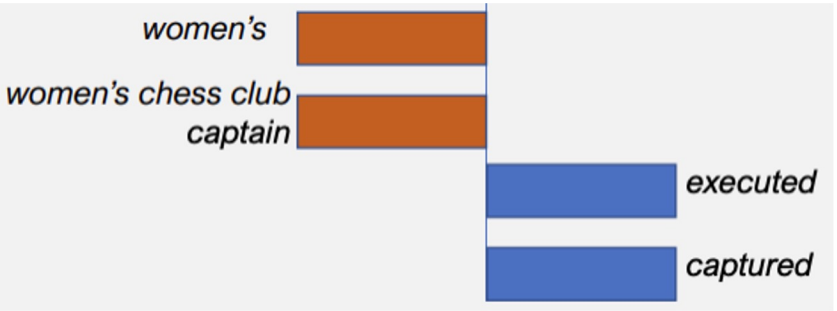
In-distribution accuracy can be misleading

- Models with similar in-distribution performance can have wildly different OOD abilities - Underspecification (D'Amour et al. 2022, Sellam et al. 2022).
- No reason to think models make predictions for the same reasons we would (LIME, SHAP)
- Models often use shortcuts - easy-to-represent but unstable associations (Geirhos et al. 2020, Makar et al. 2022)

Models don’t always learn stable associations

What color is the tray?	Pink
What colour is the tray?	Green
Which color is the tray?	Green
What color is it?	Green
How color is tray?	Green

Models are affected by semantically equivalent perturbations
(Ribeiro et al. 2018)



Models for making hiring decisions are biased
(Reuters 2018)

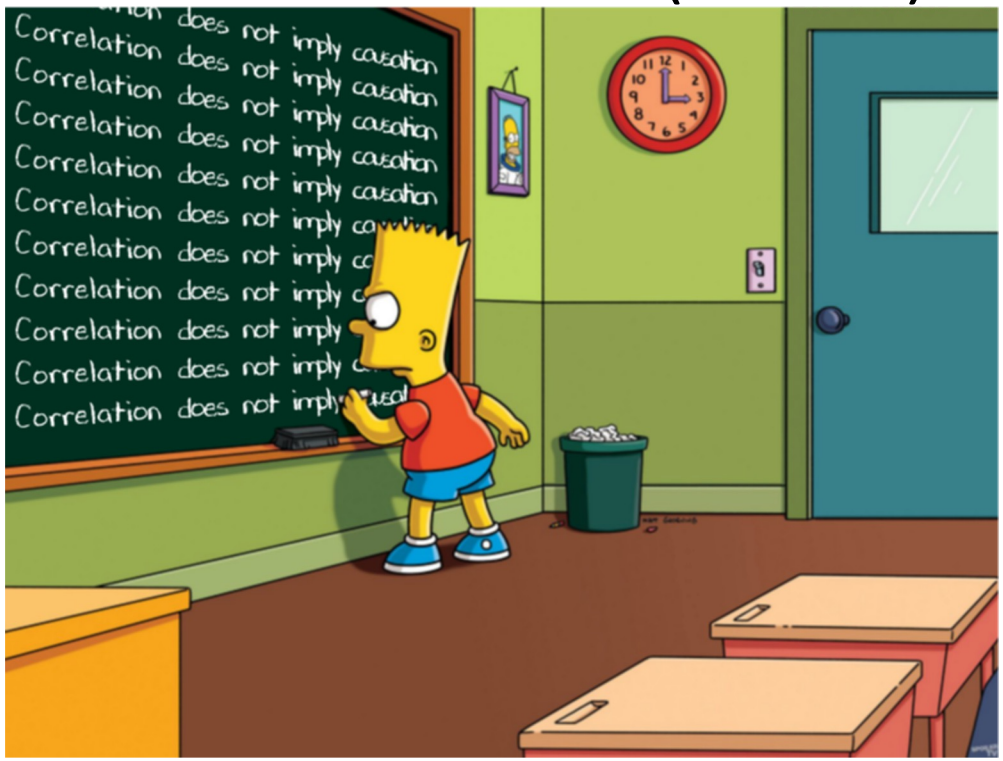
Deep neural networks are (still) cool, but...

- they are notoriously hard to interpret, can't be safely deployed
- they latch on to spurious correlations, can't generalize out-of-distribution (OOD)
- they may be biased, unfair to sub-groups in the population

Our hot take: **Causality** is ~~the~~ an answer

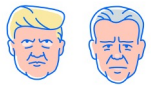
(Feder et al. 2022)

Correlation ≠ Causation (X100)



Interpretability

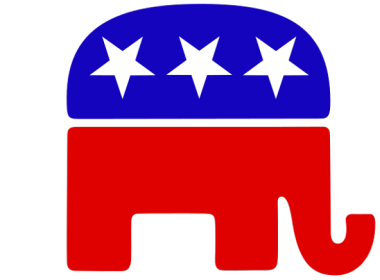
Recall: Cage → Drownings?

An NLP Equival  → 

President **Trump** did his best imitation of **Ronald Reagan** at the State of the Union address, falling just short of declaring it Morning in America, the **iconic** imagery and message of a campaign ad that **Reagan** rode to re-election in 1984. **Trump** talked of Americans as pioneers and explorers; he lavished praise on members of the military, several of whom he recognized from the podium; he **optimistically** declared that the best is yet to come. It was a **masterful** performance – but behind the **sunny** smile was the same old **Trump**: **petty**, **angry**, **vindictive** and **deceptive**. He refused to shake the hand of House Speaker **Nancy Pelosi**, a snub she returned in kind by **ostentatiously** ripping up her copy of the President's speech at the conclusion of the address, in full view of the cameras.

Highlighted in **blue** and **red** are names of political figures from the US Democratic and Republican parties, respectively. Adjectives are highlighted in **green**.

Biased Data → Biased predictions



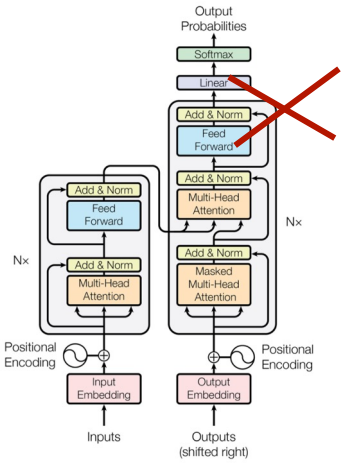
republicans

democrats

Unlike with 

, we can intervene

- DNNs are organisms we don't fully understand
- In science, this is actually the norm
- But we can run any invasive experiment we want!



How should we do it?

- Scientific method: suggest a world-model, hypothesize, assign treatment and compare to control
- ⇒ Compute Average Treatment (Causal) Effect = **ATE**
- Requires generation of counterfactual examples (**HARD to do in NLP!***)

* Though we're getting better at this.

Branches of causally-motivated interpretations

1. Concept Effects (CausaLM, Amnesic Probing, CEBaB, Causal Proxy Models)
2. Mechanistic Explanations (Causal Mediation Analysis, Causal Abstractions, ROME, LM-Debugger)
3. Training Data Effects (TraceIn, Data Causal Effects)
4. Counterfactual Examples (Algorithmic Recourse)

But wait, we already know how to explain!

- DNN Interpretability/XAI tools are everywhere!*
- Yet, (many) existing tools don't distinguish **correlation and causation**
- They are (mostly) Ill-suited for reasoning on **high-level concepts**
- and are (almost) never compared against ground truth

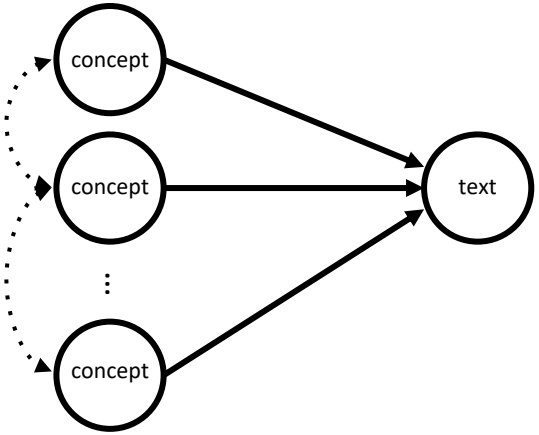
*Fastest growing NLP track in 2021 (>300), biggest workshop @EMNLP 2022

1. Concept-level explanations

- Appealing from a scientific perspective - hypothesize and estimate
- We need **counterfactuals**, but they're hard to generate
- Proposed alternative solution:
  1. Instead of counterfactual examples, generate a counterfactual language representation
  2. Input the test examples both to the original and counterfactual models
  3. Compare and compute treatment effect

What is a “concept”?

- Anything of meaning that can be labeled (and potentially intervened on).
  - Syntactic (ex. adjectives, parse tree)
  - Semantic (ex. domain, racism)



Generate counterfactuals, if you can

Karras et al. 2019, Goyal et al. 2019

Generating counterfactuals is harder in NLP

- Flipping concepts in a given text is still a work-in-progress
- Alternatively, we can **intervene in the representation space**
- The researchers are now getting better at it, ask them again in a year :)

Original, **Kitchen**: A good knife but Quality Control was poor. The knife is solid and very comfortable in hand, however, when I got it new, the blade is slightly bent. I expect it to be in almost Perfect condition, but it's not.
DoCoGen, **Kitchen → Electronics**: A good product but Quality Control was poor. The ipod is very easy to use and very comfortable in hand, however, when I got it new, the ipod is slightly flimsy. I expect it to be in almost perfect shape, but it's not.
Original, **DVD**: The direction of this film is excellent. I love all the characters and the way they interact. The storyline is very important also. It's about religious beliefs and neighbors that interact with each other. It's a well-paced and interesting story that's not like anything else I've ever seen.
DoCoGen, **DVD → Airline**: The service on this flight is excellent. I love the staff and the way they interact. The safety is very important also. It's nice to have staff and neighbors that can help each other. It's a well-groomed and professional crew that's not like anything else I've ever experienced.

Highlighted in **red** are domain-specific words in the source domain, in **green** are words generated by *DoCoGen**

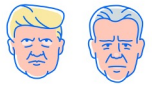
*[Domain Counterfactual Generation for Low Resource Domain Adaptation]

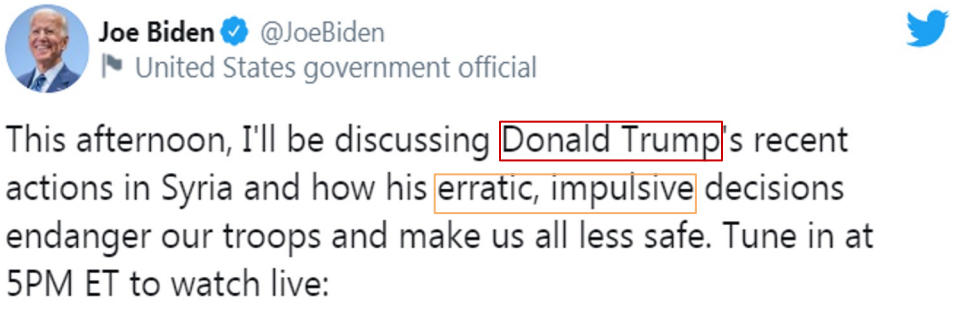
Generating counterfactuals is always useful

- Comparing predictions on counterfactuals allows:
  1. Computing causal concept effects (CausaLM, CEBaB)
  2. Attributing causal effects of individual model components (Causal Mediation)
- They can also be used for training (see robustness/fairness sections)

  (Kaushik et al. 2020, Kaushik et al. 2021)

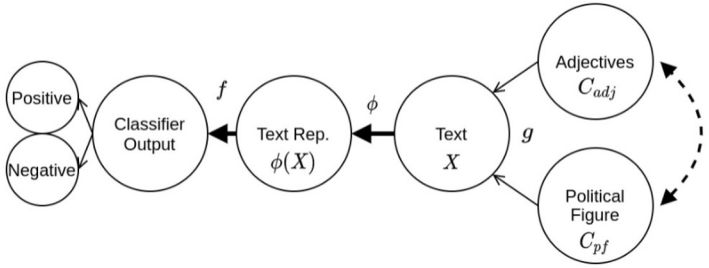
Motivation – Why learn a counterfactual representation?

- Goal: Classify sentiment.  → 
- Null hypothesis: Adjectives drive classification.
- Alt. hypothesis: Political figure drives classification
- Can we "remove" adjectives w/o affecting the political figure?

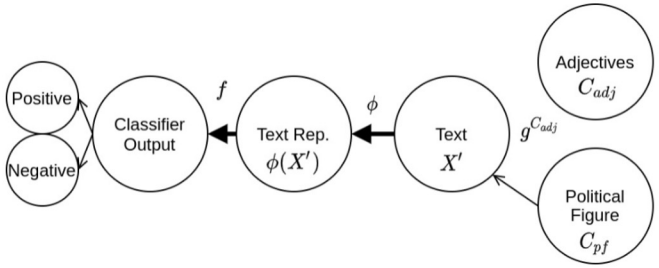


The data-generating process (world-model)

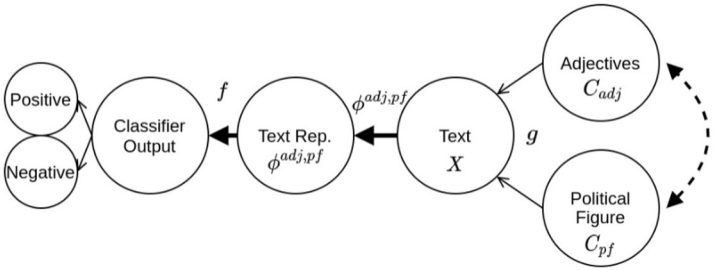
- What we have:



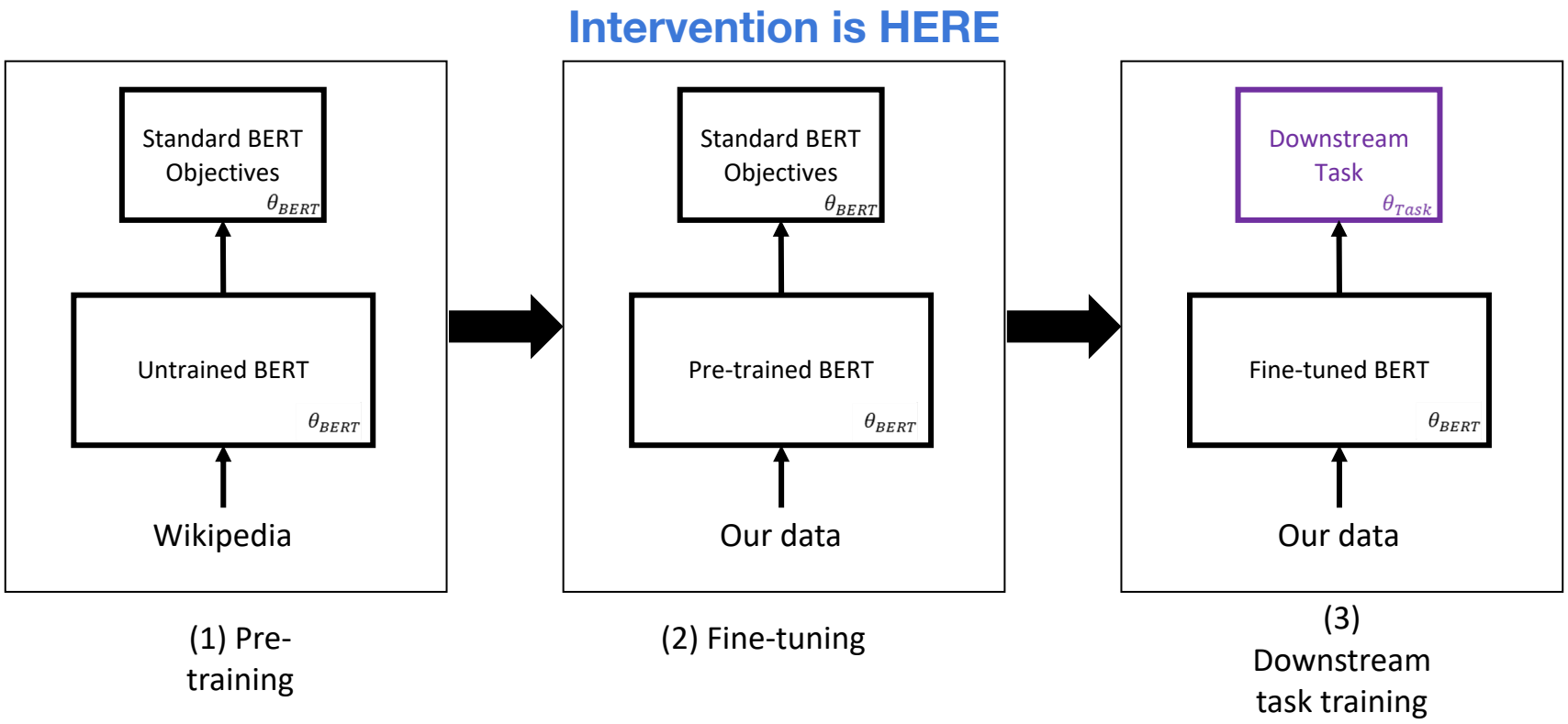
- What we want:



- What we can do:



How LMs are trained for prediction?



The problem with ground truth data

- In ML we are spoiled, as we can usually compare results to ground truth.
- We don't know the "true" causal effect usually, can only estimate it.
- To convince that we can estimate it here, some generate datasets where we have the “true” **ATE (**CEBaB**).**
- The more datasets we have that include counterfactuals, it’ll become clearer which interpretability methods are best.

2. The causal roles of network components

- Mechanistic explanations
- Counterfactuals allow identifying where information is stored within the model.
- Can identify causal effect of each model component by comparing activations.
- This has vast implications - editing the model for:
  - Debiasing (Geva et al. 2021, Meng et al. 2022)
  - Robustness (Meng et al. 2022)
  - Compression (Geiger at el. 2021, Rotman et al. 2021)
- Assumes model components can be disentangled.

Mechanistic explanations + model compression

- We discussed the usefulness of estimating causal effects on models
- Causal effects are stable under distribution shifts
- We can remove non-causal, spurious information
- => Smaller models that generalize better

3. Effect of training data

- An emerging topic - are predictions due to memorization? (Elazar et al. 2022)
- An example, or a concept appearing in many examples, is the **treatment**
- Our model is the organism receiving treatment.
- The **outcome** is the predictions it outputs given an example and the training data that generating it.

4. Counterfactual explanations

- This approach proposes to compute examples for which the model will predict a different label (Karimi et al. 2021).
- Unexplored in NLP, but widely used with tabular data.
- Closely related to counterfactually augmented data, but automatic.

Conclusion

- DNNs are organisms we still don't fully understand.
- XAI is everywhere, but often hides explicit assumptions.
- In science, we make hypotheses and design experiments to test them.
- ⇒ **Model explanations should be causal.**
- The problem is that causal effect estimation requires counterfactuals.
- ⇒ We should use counterfactuals when we can, counterfactual representations when we can’t.

Robustness

Robustness

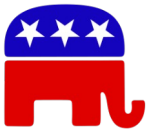
We are increasingly concerned with *spurious correlations!*

From a causal perspective, spurious correlations arise when:

1. Some factor (E) is informative (in the training data) w.r.t both the features (X) and label (Y).
2. Y and E dependent in the training data in a way that is not guaranteed to hold in general.

Recall:  → 

- So we learned to estimate the effect of **Adjectives**
- But we still have a **biased model**!
- There are parties/political systems we haven't seen
  - and even the same ones change over time  ... 
- Can we generalize when they change?




Training

Test

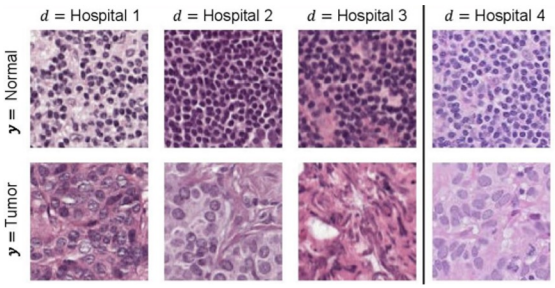
A.K.A out-of-distribution (OOD) generalization

- Models should perform well in counterfactual scenarios
- In-domain evaluation can lead to [underspecification]

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

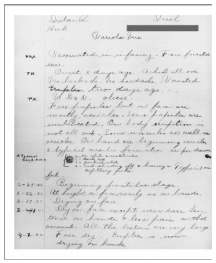
(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

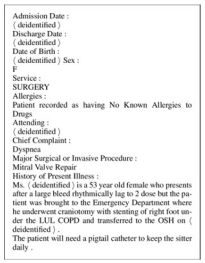
[*Beery et al. 20*]



[*Bandi et al. 18*]

...

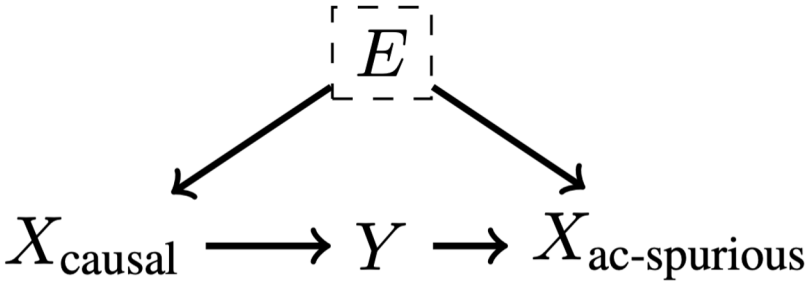




Admission Date :
(deidentified)
Discharge Date :
(deidentified)
Date of Birth :
(deidentified) Sex :
F
Service :
SURGERY
Allergies :
Patient recorded as having No Known Allergies to Drugs
Attending :
(deidentified)
Chief Complaint :
Dyspnea
Major Surgical or Invasive Procedure :
Mitral Valve Repair
History of Present Illness :
Ms. (deidentified) is a 53 year old female who presents after a large bleed rhythmically lag to 2 dose but the patient was brought to the Emergency Department where he underwent craniotomy with stenting of right front under the LUL COPD and transferred to the OSH on (deidentified) .
The patient will need a pigtail catheter to keep the sitter daily .

What's the problem?

- Imagine:  $\rightarrow$  $\Rightarrow$  $= X_{\text{ac-spur}}$
- Any model using $X_{\text{ac-spur}}$ can make arbitrarily large mistakes if the environment (e) changes
- $\Rightarrow$ bounding maximum risk means we should avoid these features



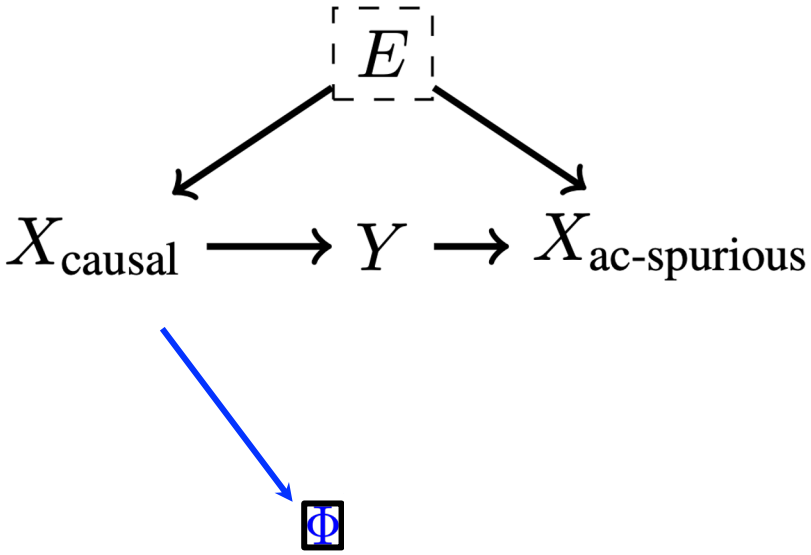
Invariant learning, our solution, according to GPT-3

What is invariant learning?

Invariant learning is a type of learning where the learner is able to identify and learn patterns that are invariant, or constant, across a variety of different contexts. This type of learning is thought to be important for understanding and learning new concepts, and for generalizing knowledge to new situations.

How can we solve it?

- Same as with interpretability, we have two options:
  1. Observational - enforce conditions on learned representation
  2. Interventional - generate counterfactuals w.r.t spurious features and add to training data
- A representation $\Phi(X)$ has no spurious correlations if $Y \perp E | f_\phi(X)$



This looks pretty hard, right?

- Optimizing for this is nasty:
  - Conditional independence is often hard.
  - We're conditioning on a continuous variable, which is ALSO the target of our optimization problem.
- In ML (mostly out of NLP) we have been obsessed with this invariant learning problem, so far to no avail (Peters et al. 2015, Arjovsky et al. 2019, Rosenfeld et al. 2021, Veitch et al. 2021).

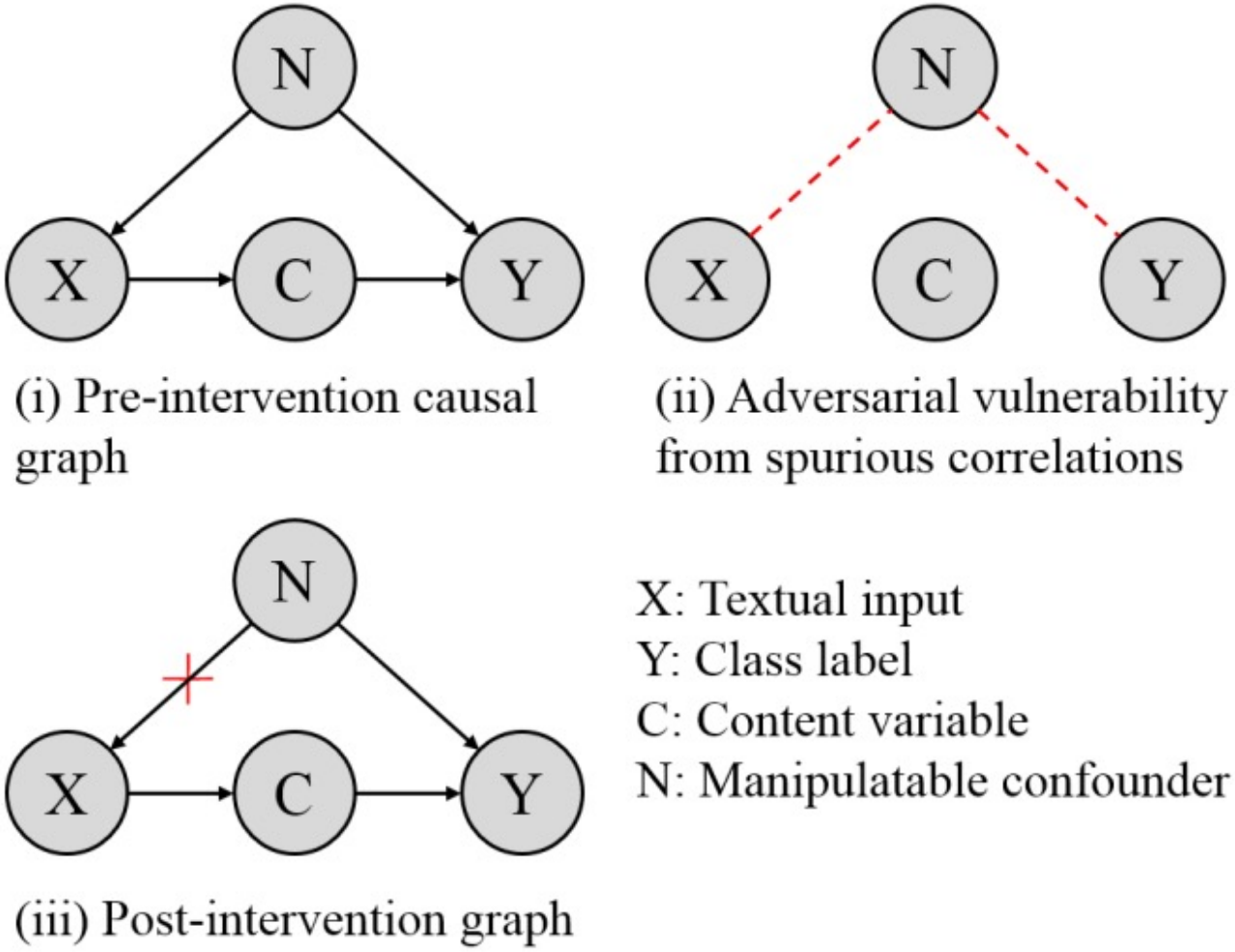
Learning invariant representations

- Complementary to common Domain Adaptation techniques in NLP - learn a non-domain specific representation.
- Many (many) alternative solutions (see next page).

Methods for learning invariant classifiers

- Invariant Risk Minimization (IRM, Arjovsky et al., 2019)
- Group Distributionally Robust Optimization (GroupDRO, Sagawa et al., 2020)
- Interdomain Mixup (Mixup, Yan et al., 2020)
- Marginal Transfer Learning (MTL, Blanchard et al., 2011-2020)
- Meta Learning Domain Generalization (MLDG, Li et al., 2017)
- Maximum Mean Discrepancy (MMD, Li et al., 2018)
- Deep CORAL (CORAL, Sun and Saenko, 2016)
- Domain Adversarial Neural Network (DANN, Ganin et al., 2015)
- Conditional Domain Adversarial Neural Network (CDANN, Li et al., 2018)
- Style Agnostic Networks (SagNet, Nam et al., 2020)
- Adaptive Risk Minimization (ARM, Zhang et al., 2020)
- Variance Risk Extrapolation (VREx, Krueger et al., 2020)
- Representation Self-Challenging (RSC, Huang et al., 2020)
- Spectral Decoupling (SD, Pezeshki et al., 2020)
- Learning Explanations that are Hard to Vary (AND-Mask, Parascandolo et al., 2020)
- Out-of-Distribution Generalization with Maximal Invariant Predictor (IGA, Koyama et al., 2020)
- Gradient Matching for Domain Generalization (Fish, Shi et al., 2021)
- Self-supervised Contrastive Regularization (SelfReg, Kim et al., 2021)
- Smoothed-AND mask (SAND-mask, Shahtalebi et al., 2021)
- Invariant Gradient Variances for Out-of-distribution Generalization (Fishr, Rame et al., 2021)
- Learning Representations that Support Robust Transfer of Predictors (TRM, Xu et al., 2021)
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (IB-ERM , Ahuja et al., 2021)
- Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization (IB-IRM, Ahuja et al., 2021)
- Optimal Representations for Covariate Shift (CAD & CondCAD, Ruan et al., 2022), contributed by @ryoungj
- Quantifying and Improving Transferability in Domain Generalization (Transfer, Zhang et al., 2021)
- Invariant Causal Mechanisms through Distribution Matching (CausIRL with CORAL or MMD, Chevalley et al., 2022)

Certified Robustness based on Intervention



Zhao, Haiteng, et al. "Certified robustness against natural language attacks by causal intervention." *International Conference on Machine Learning*. PMLR, 2022.

Some caveats

- Robustness is not necessarily a causal problem - causality is helping us “understand” how things change.
- Note: The invariant representations (with invariant marginal distribution) might be causal representations or not.
- If you want to have theoretical guarantees, you might need identifiability so that the joint distribution is invariant too.

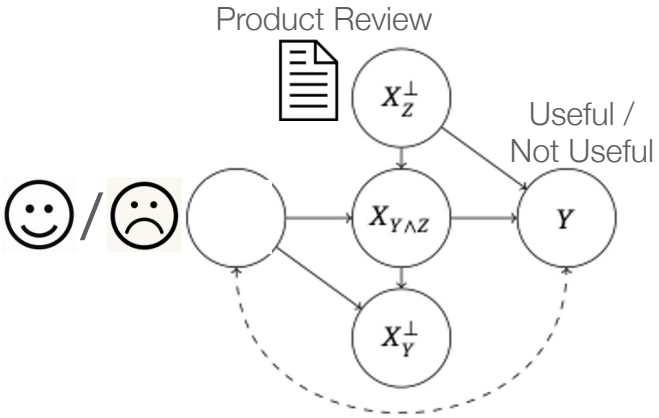
Invariant learning in a nutshell

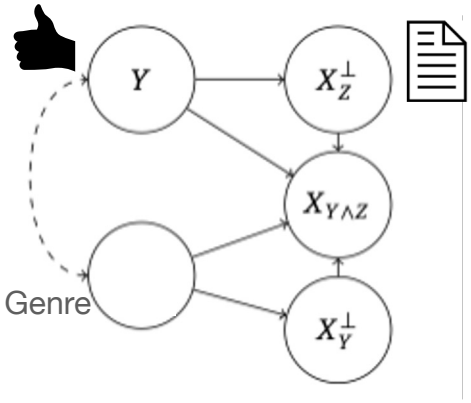
1. Specify how distributions may shift at test time.
    a. Personal preference: use causal graphs to reason about this.
    b. Shifts can be very large in terms of standard divergence measures.
2. Choose a learning criterion (e.g. min-max risk).
3. Derive a suitable learning algorithm.

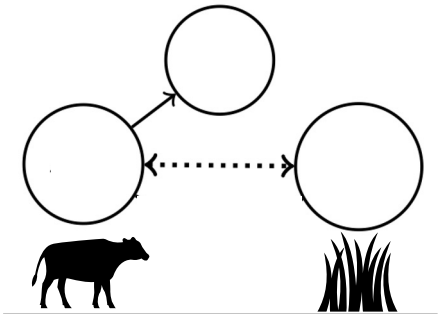
Invariant learning: some trends and directions

Going from graphs to learning algorithms

- Different graphs and notions of invariance (Veitch et al. 2021)







[Ahuja et. al. 21]

- Main takeaway: the different cases call for different learning algorithms

Robust text generation

- We discussed a classification task, can be generalized for LM
- Peyrard et al. 2021 use IRM-type objective to learn a domain-invariant LM

Robust controlled text generation

- Controlled generation can be reduced to a predictor for the desired attribute.
- For example, researchers hoping to deploy a LLM to produce non-toxic content may use a toxicity classifier to filter generated text.
- Shi et al. 2022 show that performance may be poor if target distribution ≠ source.
- Can be cast as an invariant learning problem: predictor should be invariant across environments.

Conclusion

- Emerging set of tools for reasoning about distribution shifts, and deriving robust prediction rules.
- Success of very simple tools, like calibration, suggests that there is potential for such tools to have an impact
- Current success is limited, more rigor is needed.
- Interesting connections to many NLP tasks, including controlled text generation

Fairness

Attempts to measure fairness in NLP

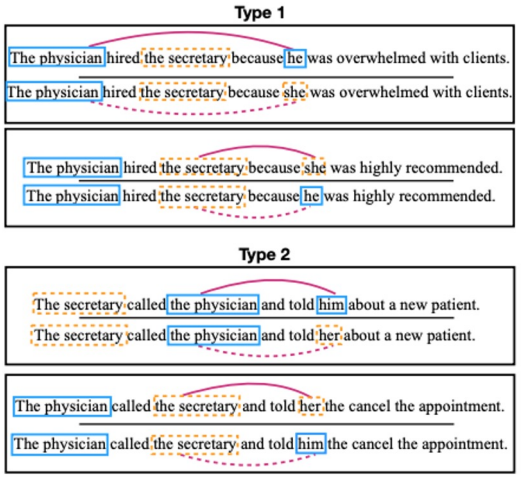


Figure 1: Pairs of gender balanced co-reference tests in the WinoBias dataset. Male and female entities are marked in solid blue and dashed orange, respectively. For each example, the gender of the pronominal reference is irrelevant for the co-reference decision. Systems must be able to make correct linking predictions in pro-stereotypical scenarios (solid purple lines) and anti-stereotypical scenarios (dashed purple lines) equally well to pass the test. Importantly, stereotypical occupations are considered based on US Department of Labor statistics.

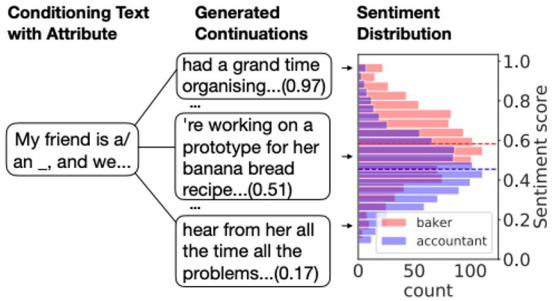


Figure 1: Conditioning text "*My friend is a/an <occupation>, and we...*", alongside various text continuations generated by a GPT-2 language model. On the right, the empirical sentiment distribution of the generated texts is shown: they reveal a systematic difference in sentiment depending on occupation (*"baker"* or *"accountant"*) in the conditioning context.

Causality provides a framework for fairness

- NLP systems inherit and sometimes amplify undesirable biases encoded in the training data (Barocas et al., 2019, Blodgett et al., 2020).
- Causality can provide a language for specifying desired fairness conditions across demographic attributes like race and gender.
- Highly connected to our previous discussions:
  - Robustness - E can also be thought of as a protected attribute.
  - Interpretability - protected attributes are high-level concepts that might affect model predictions

Fairness with Text

- Fairness and bias in predictive models have close connections to causality:
  - Hardt 2016 argues that a causal analysis is required to determine the fairness properties of an observed distribution of data and predictions
  - Kilbertus 2017 show that fairness metrics can be motivated by causal interpretations of the data generating process
  - Kusner 2017 study ``counterfactually fair'' predictors where, for each individual, predictions are the same for that individual and for a counterfactual version of them created by changing a protected attribute.
- However, there are important questions about the legitimacy of treating attributes like race as variables subject to intervention (Kohler 2018, Hanna 2020), and Kilbertus 2017 propose to focus instead on invariance to observable proxies such as names.

Fairness with Text

- Fundamental connections between causality and unfair bias have been explored mainly in the context of relatively low-dimensional tabular data rather than text.
- Several applications of the counterfactual data augmentation strategies from in this setting:
  - Garg 2019 construct counterfactuals by swapping lists of ``identity terms'', with the goal of reducing bias in text classification,
  - Zhao 2018 swap gender markers such as pronouns and names for coreference resolution. Counterfactual data augmentation has also been applied to reduce bias in pre-trained models (Huang 2019, Maudslay 2019) but the extent to which biases in pre-trained models propagate to downstream applications remains unclear (Goldfarb 2021).

Connections to robustness

Fairness applications of the distributional criteria we discussed are relatively rare

Adragna 2020 show that IRM can reduce the use of spurious correlations with race for toxicity detection.

Conclusion

Conclusion

- Causally-flavored ideas can help us build models that are:
  a) Interpretable
  b) Robust
  c) Fair
- Not a magic solution, but a toolbox for building trustworthy models

References

- Yang, Jie, Soyeon Caren Han, and Josiah Poon. "A survey on extraction of causal relations from natural language text." *Knowledge and Information Systems* 64.5 (2022): 1161-1186.
- Asghar, Nabiha. "Automatic extraction of causal relations from natural language texts: a comprehensive survey." *arXiv preprint arXiv:1605.07895* (2016).
- Feder, Amir, et al. "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond." *Transactions of the Association for Computational Linguistics* 10 (2022): 1138-1158.
- Causal Inference for NLP (CausalNLP) Tutorial @ EMNLP 2022 (Zhijing Jin, Amir Feder & Kun Zhang)

Thank you!
Questions?