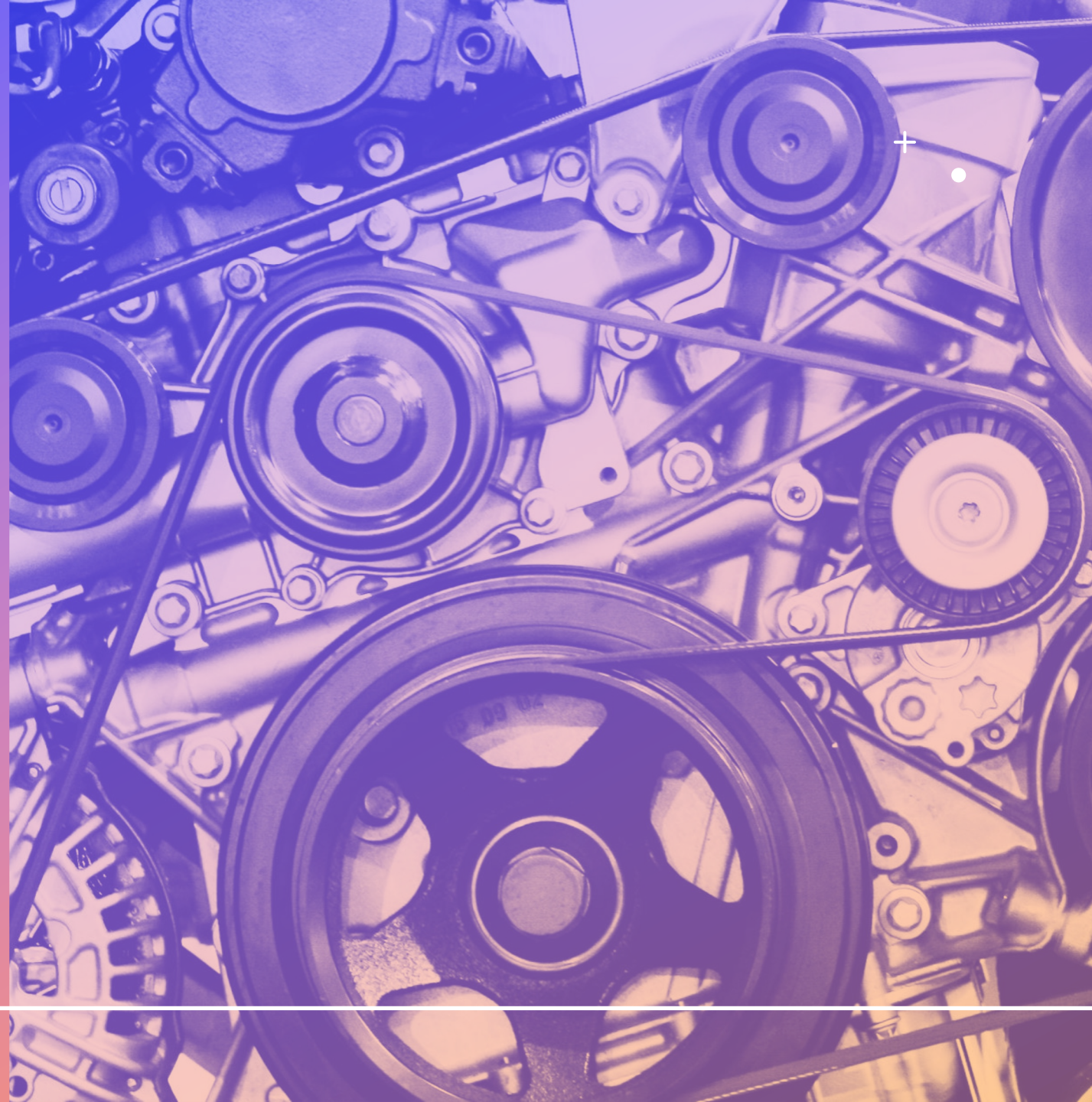


CAUSES WITH MANY MOVING PARTS

Ricardo Silva, UCL

Workshop on AI, Causality and Personalized Medicine

Leibniz AI Lab, September, 2022



First, the Many Moving Parts of this Work



Limor Gultchin



Jean Kaddour



Matt Kusner



Qi Liu



David Watson



Caroline Zhu



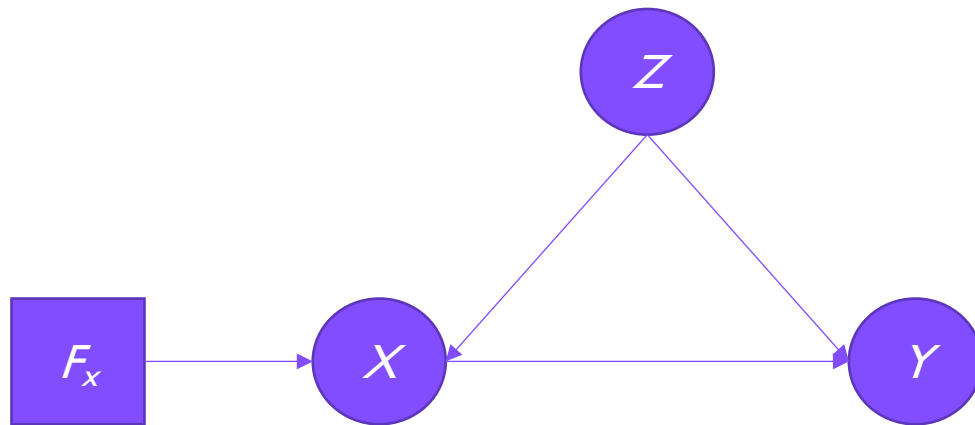
SCOPE

Setup

- We will start from the most vanilla case, with
 - a treatment X
 - an outcome Y
 - pre-treatment covariates Z
 - don't get too emotionally attached to this notation, I *will* be inconsistent!
- **But: treatment X won't be a scalar. It can be a fixed-size vector or a more structured object.**
- **It is not necessarily the case we can perfectly control X .**

Structural assumptions and graphical notation

- Squares are intervention variables
- Circles are random variables



Conditional ignorability + (sometimes) exogeneity

$$Y \perp\!\!\!\perp F_x \mid \{X, Z\}$$

$$Z \perp\!\!\!\perp F_x$$

“**How** X is chosen is **irrelevant** once we know **which** value X took (along a relevant set of background variables).”

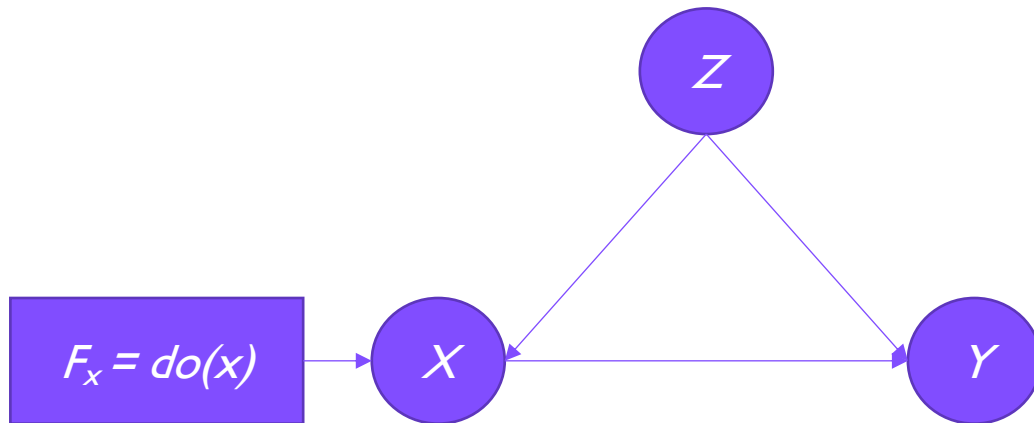
“ F_x is **external** to the system.”

What does F_x mean?

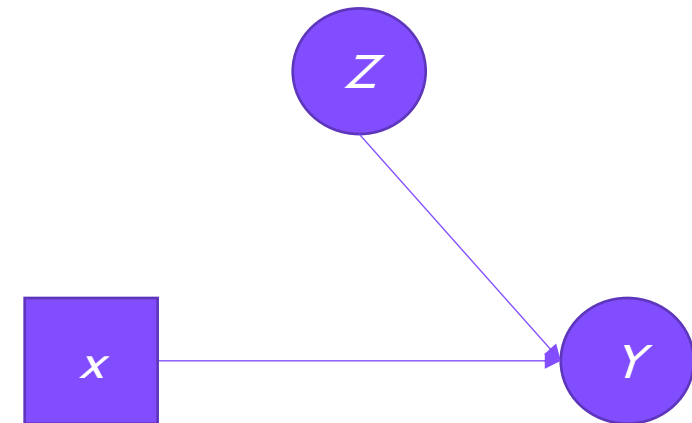
- Mathematically, just an index denoting a type of *external* intervention. Not a random variable. Independence statements still well-defined.
 - Spirtes et al., Pearl, AP Dawid
- Graphically, it has no ancestors.
- Operationally, it can mean any regime-switch indicator, in-sample or out-of-sample.

What does F_x mean?

- Pearl's *do* operator/Spirtes et al.' *set* operator etc. can be understood as a special values in the domain of F_x
 - That is, perfect control of treatment variable.



Explicitly value shown



Alternative graphical representation

Tasks

- Estimating conditional average treatment effects (CATEs).

$$\mathbb{E}[Y \mid do(X = x), Z = z] - \mathbb{E}[Y \mid do(X = x'), Z = z]$$

$$\mathbb{E}_{F_x = f_x}[Y \mid Z = z] - \mathbb{E}_{F_x = f'_x}[Y \mid Z = z]$$

Tasks

- Predicting future outcomes yet to happen under F_x .

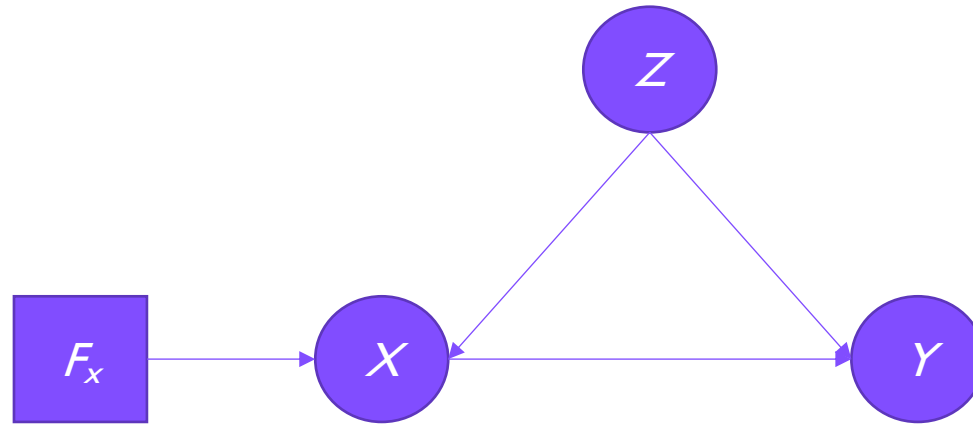
$$\mathbb{E}[Y \mid do(X = x), Z = z]$$

$$\mathbb{E}_{F_x = f_x}[Y \mid Z = z]$$

From assumptions and data to estimates

- $Y \perp\!\!\!\perp F_x \mid \{X, Z\}$ implies

$$\mathbb{E}_{F_x=do(x)}[Y \mid Z = z] = \mathbb{E}_{F_x=idle}[Y \mid X = x, Z = z] = \mathbb{E}[Y \mid X = x, Z = z]$$



- But we know that plain regression is kind of rubbish for CATE.

Example: linear case

- Outcome model:

$$Y = \alpha x + \beta^T z + \epsilon$$

- Parameter of interest is α :

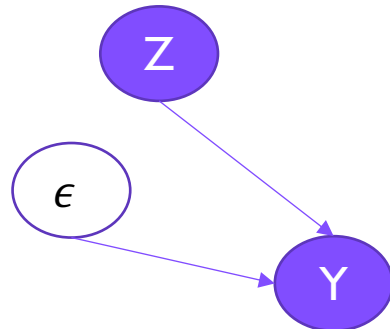
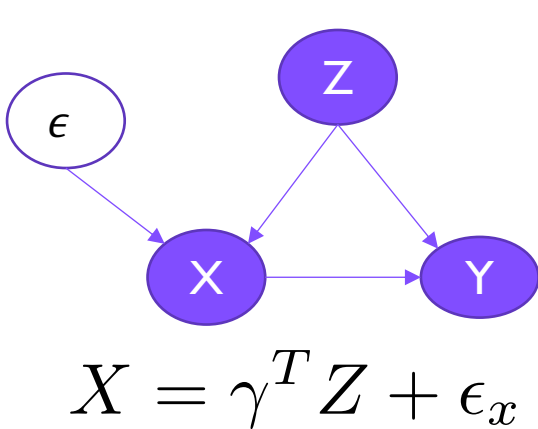
$$\mathbb{E}[Y \mid do(X = x), z] - \mathbb{E}[Y \mid do(X = x - \Delta), z] = \alpha \Delta$$

- That's a drop in the ocean of parameters. If we regularize the regression, that coefficient gets “very” biased.

Hahn et al. (2018). “Regularization and confounding in linear regression for treatment effect estimation”. *Bayesian Analysis*.

A solution

- When X and Z are independent, we can mathematically show the bias disappears, regression is fine.
 - But observational studies are the perfect storm.
- Solution? Write the problem by making-up a **representation** that is independent of Z but is still informative of treatment.



$$\begin{aligned} Y &= \alpha(X - \gamma^T Z) + (\beta + \alpha\gamma)^T Z + \epsilon_y \\ &= \alpha R_x + \beta_\alpha Z + \epsilon_y \end{aligned}$$

What now?

- I will describe problems **where X is structured**, and how this type of residual creation can be fruitfully exploited when we want to learn a **representation** of X .
- Then I'll move on to problems where, on top of structure, **$do(x)$ is itself undefined**. What would be the point of saying X is a cause of anything?

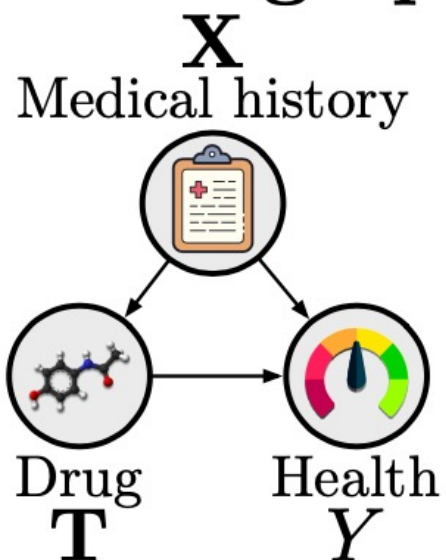


STRUCTURED INTERVENTION NETWORKS

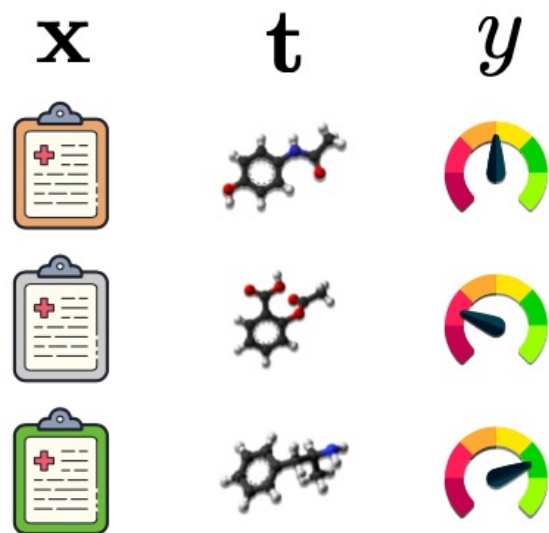
From Kaddour et al. (2021), “Causal effect inference for structured treatments”.
NeurIPS.

Motivation

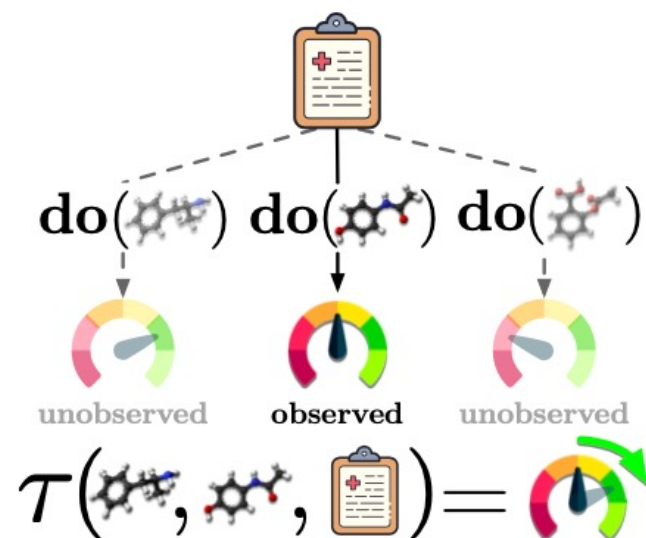
Causal graph



Observed data



CATE



Challenges and perspectives

- Are you joking? Dimensionality is high. What about the lack of overlap problem?
 - Fingers crossed that we are in a domain where it's possible to learn a representation of the treatment that is tractable.
- Aren't many of these problems effectively of categorical treatment? Isn't there a pre-defined set of molecules anyway?
 - Indeed, but it doesn't mean we cannot learn a representation of the treatment that is more tractable than one-hot encoding.

Challenges and perspectives

- We can tap into existing machinery for dealing with structured data, e.g. graph neural networks.
- However, we are still interested in contrasts only (CATEs), so need to be predictive of a treatment by itself.
- We could define a baseline treatment (e.g., no drug at all), but in what follows we will not.

Example

- Before we go into the methodological details, it is good to illustrate what the end product is.
- We will describe a simulation study based on real-data that provides covariates and treatments. Propensities and outcomes are simulated.
- Comparisons will include structured regression methods, like graph neural networks.

Data

- The Cancer Genomic Atlas (TCGA) simulation
 - 4,000 gene expression measurements of cancer patients as covariates
 - 10,000 sampled molecules from the QM9 dataset as treatments
- Simulated propensities: ignores structure completely. Treatments are selected by a logistic regression model.

$$p(\mathbf{T} | \mathbf{x}) = \text{softmax}(\kappa \mathbf{W}^\top \mathbf{X}), \text{ where } \mathbf{W} \in \mathbb{R}^{|\mathcal{T}| \times d}, \forall i, j : W_{ij} \sim \mathcal{U}[0, 1]$$

- Problem gets “easier” (not necessarily) as $\kappa \rightarrow 0$.

Data

- What goes in those molecules?
 - A **relational graph** indicating which atom is linked to which atom
 - **Edge attribute**: one of four classes (*single, double, triple* and *aromatic*)
 - **Vertex attributes**: 78 atom features

Data

- Simulated outcome:
 - for each covariate instance \mathbf{x} , use its projection into the first 8 principal components
 - for each molecule, use 8 structural properties \mathbf{z} defined by QM9
 - A baseline is also added, which is a synthetic linear function

$$Y = 10\mu_0(\mathbf{x}) + 0.01\mathbf{z}^\top \mathbf{x}^{(\text{PCA})} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

- The methods are not given any information about either the structure of the propensity, or the outcome models

Evaluation metrics

- Given two treatment levels t and t' , and covariates \mathbf{x} , define:

$$\tau(t', t, \mathbf{x}) = \mu_{t'}(\mathbf{x}) - \mu_t(\mathbf{x}) = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}, \mathbf{T} = t'] - \mathbb{E}[Y | \mathbf{X} = \mathbf{x}, \mathbf{T} = t].$$

- For a fixed pair (t, t') , define the Unweighted and **Weighted** Precision in Estimation of Heterogeneous Effects as

$$\epsilon_{\text{UPEHE}(\text{WPEHE})} \triangleq \int_{\mathcal{X}} \left(\hat{\tau}(t', t, \mathbf{x}) - \tau(t', t, \mathbf{x}) \right)^2 p(t | \mathbf{x}) p(t' | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

Evaluation metrics

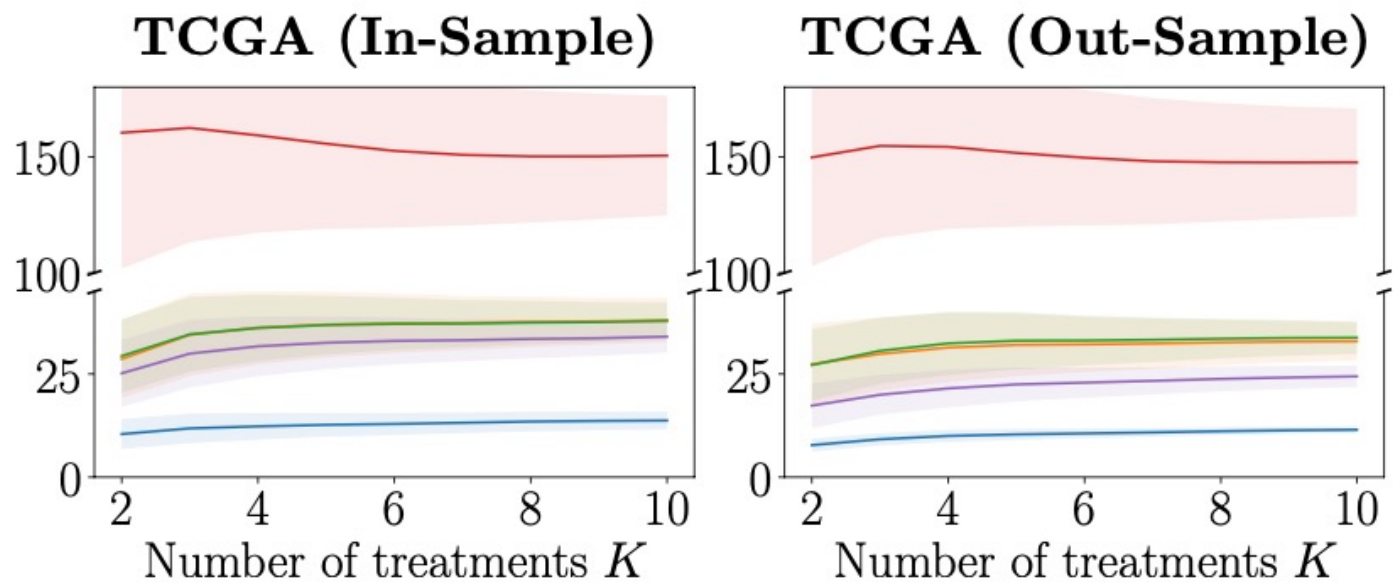
- U/WPEHE @ K :
 - We take the top K treatments (by decreasing order of propensity score) and compute the U/WPEHE for all $K(K - 1)/2$ pairs.
- In-sample vs out-of-sample:
 - In the in-sample case, pairs (x, t) are taken from the data in which we fit the model, t' is chosen from the top K , outcome in the data is taken as the expected response $\mu(x, t)$.

Baselines

- **Zero**: just report zero effect!
- **CAT**: regression with categorical encoding of treatments
- **GNN**: regression with a graph neural network (GNN)
- **GraphITE**: a GNN-based CATE estimator that does regression with a penalization of dependence between X and T representations
- **SIN**: our method, a GNN-based variation of the R-Learner (to be described)

Results

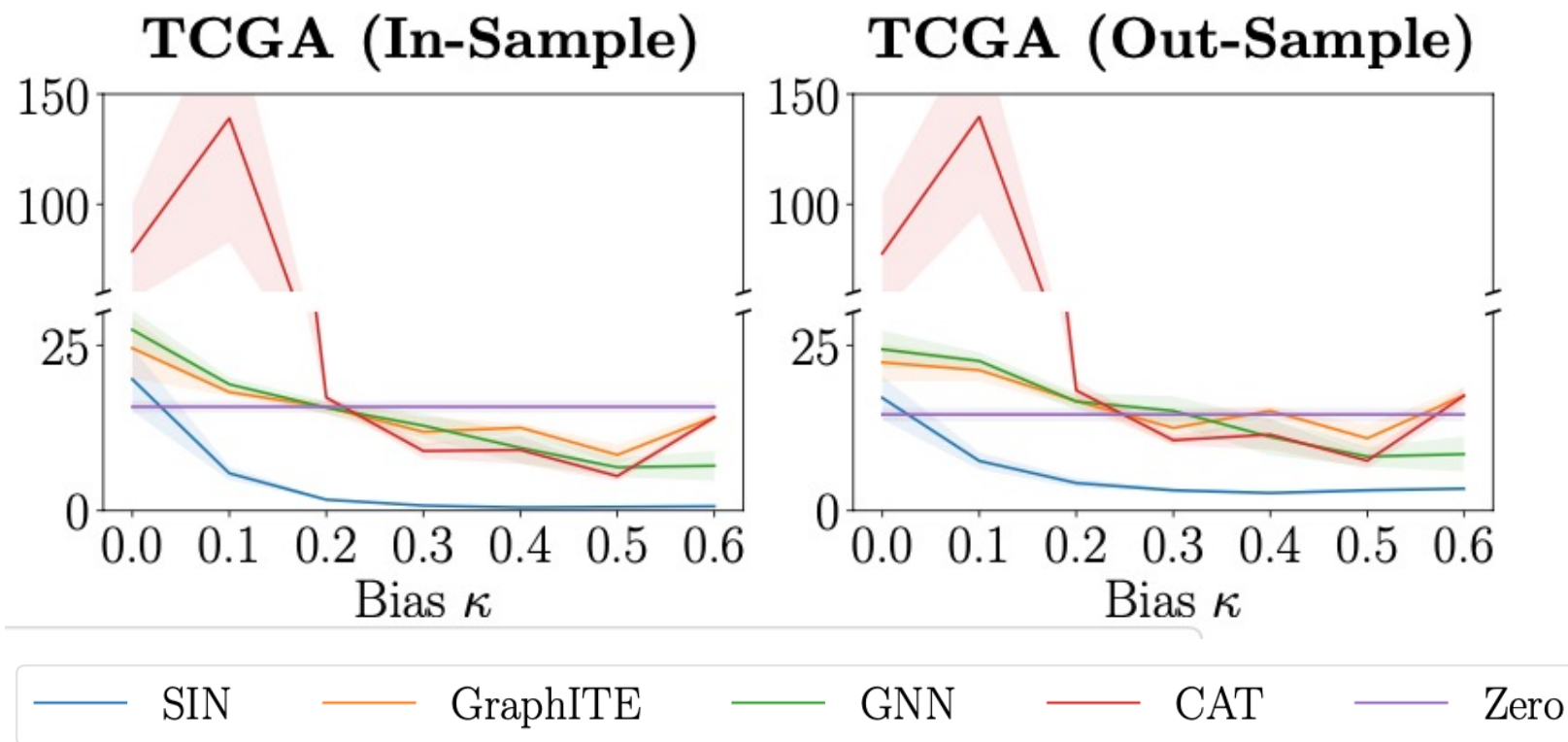
WPEHE@K, 10 trials



— SIN — GraphITE — GNN — CAT — Zero

Results

WPEHE@6 with increasing bias



So, what's the idea?

- SInNs, *Structured Intervention Networks*, have a built-in way of decoupling representations of the treatment from covariates (unlike GraphITE, which is penalization-based).
- It explores the structure of the treatment, instead of dealing with (vanishingly small) partitions of the data by category of treatment.
- It taps in whatever structured data regression method we pick (e.g., GNNs for graph data).

Ingredient 1: Robinson's decomposition

- The main trick we will rely on
 - Exploited in many CATE estimators, including the R-Learner (Nie and Wager, Biometrika, 2020), which is closely related
- We will describe first the well-studied binary treatment case, T in $\{0, 1\}$.

$$Y = f(\mathbf{X}, T) + \varepsilon \equiv \mu_0(\mathbf{X}) + T \times \tau_b(\mathbf{X}) + \varepsilon;$$

Robinson's decomposition

Propensity score: $e(\mathbf{x}) \triangleq p(T = 1 | \mathbf{x})$

Conditional mean outcome: $m(\mathbf{x}) \triangleq \mathbb{E}[Y | \mathbf{x}] = \mu_0(\mathbf{x}) + e(\mathbf{x})\tau_b(\mathbf{x})$

CATE @ \mathbf{x} : $\tau_b(\mathbf{x}) \triangleq \tau(1, 0, \mathbf{x})$

From the above:

$$Y - m(\mathbf{X}) = (T - e(\mathbf{X}))\tau_b(\mathbf{X}) + \varepsilon,$$

Estimator:

$$\hat{\tau}_b(\cdot) = \arg \min_{\tau_b} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\tilde{y}_i - \tilde{t}_i \times \tau_b(\mathbf{x}_i) \right)^2 + \Lambda(\tau_b(\cdot)) \right\}$$

Pseudo-data:

$$\tilde{y}_i \triangleq y_i - \hat{m}(\mathbf{x}_i) \text{ and } \tilde{t}_i \triangleq t_i - \hat{e}(\mathbf{x}_i)$$

R-Learner

- Learns pseudo-data in a separate dataset.
- Fits CATE function with independent data points.
 - It is possible to do **cross-fitting** i.e. use the entire data as long as any data used to form the pseudo-points is independent of the data used in the respective final regression
- Any machine learning method can be used in any of the functions

Beyond binary treatments

- Adopt the following (product) form:

$$Y = g(\mathbf{X})^\top h(\mathbf{T}) + \varepsilon$$

This is general enough
under reasonable function
spaces

- CATE function now given by:

$$\tau(\mathbf{t}', \mathbf{t}, \mathbf{x}) = g(\mathbf{x})^\top \left(h(\mathbf{t}') - h(\mathbf{t}) \right)$$

- **Propensity features** at play: $e^h(\mathbf{x}) \triangleq \mathbb{E}[h(\mathbf{T}) \mid \mathbf{x}]$

Learning problem

- SIN regression:

$$Y - m(\mathbf{X}) = g(\mathbf{X})^\top \left(h(\mathbf{T}) - e^h(\mathbf{X}) \right) + \varepsilon.$$

- If we knew the propensity features:

$$\hat{g}(\cdot), \hat{h}(\cdot) \triangleq \arg \min_{g, h} \left\{ \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{m}(\mathbf{X}_i) - g(\mathbf{X}_i)^\top \left(h(\mathbf{T}_i) - \hat{e}^h(\mathbf{X}_i) \right) \right)^2 + \Lambda(g(\cdot)) \right\}$$

- But we don't know them! If we were to fix $h(\cdot)$, though, we could get some *quasi-oracle property* by fitting propensity features by regression and then plugging them in.

Ingredient 2: Saddle-point algorithm

- $\hat{m}(x)$ is OK, this is just the regression of outcome on the covariates.
- After that:
 1. Fix propensity features estimate $\hat{e}^h(x)$
 2. Learn treatment representation $\hat{h}(t)$ and covariate representation $\hat{g}(x)$
 3. Given $\hat{h}(t)$ learn $\hat{e}^h(x)$ as $\mathbb{E}[\hat{h}(T) \mid X]$
 4. Iterate


$$\hat{g}(\cdot), \hat{h}(\cdot) \triangleq \arg \min_{g, h} \left\{ \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{m}(\mathbf{X}_i) - g(\mathbf{X}_i)^\top \left(h(\mathbf{T}_i) - \hat{e}^h(\mathbf{X}_i) \right) \right)^2 + \Lambda(g(\cdot)) \right\}$$

SIN Summary

- There is no theory that the SIN algorithm will converge, but in practice we haven't had problems.
- At least compared to the off-the-shelf idea of GraphITE, of just trying to learn a decoupling between T and X , Robinson's decomposition seems to give a sizeable pay off.
- Python + PyTorch code fully available at <https://github.com/JeanKaddour/SIN>

Coming up next

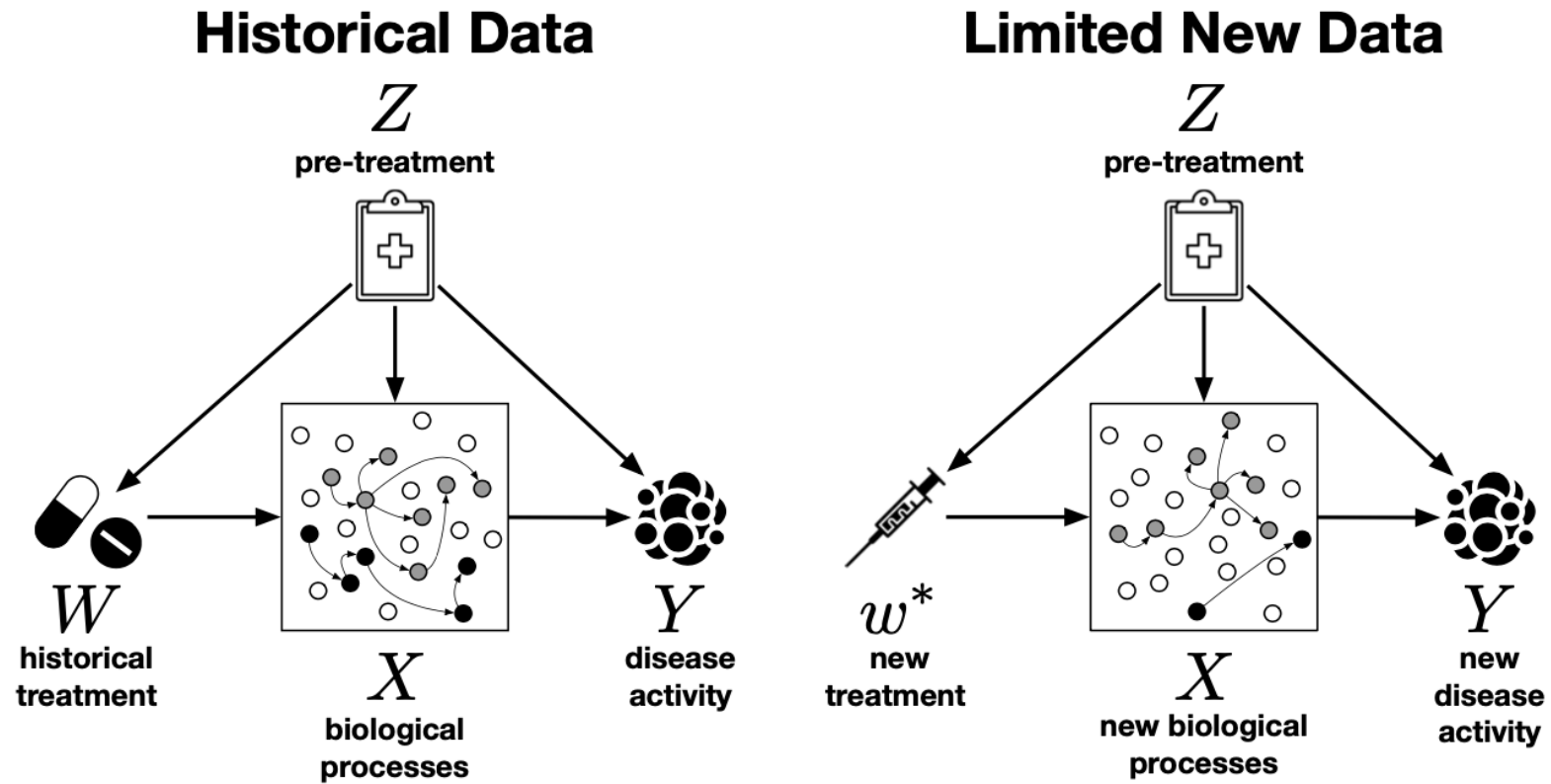
What if the interventions are not well-defined?



CAUSES WITHOUT CONTROL

From Gultchin et al. (2021), “Operationalizing complex causes: a pragmatic view of mediation”. ICML.

Motivation



Causal and constitutive relationships

- As we have seen with SIN, we can think of elements of a cause as having **constitutive** relationships wrt a “global” treatment variable.
- But it doesn't mean we can fully control all of it at once.
- Moreover, in the same way the notion of intervention requires knowing what's “inside” and “outside” a causal system, the notion of treatment requires postulating what's constitutive and what internally causal.
 - A wheel doesn't “cause a car”, but a failure of the O-rings of space shuttle Challenger did make it explode.

The fog of causation...

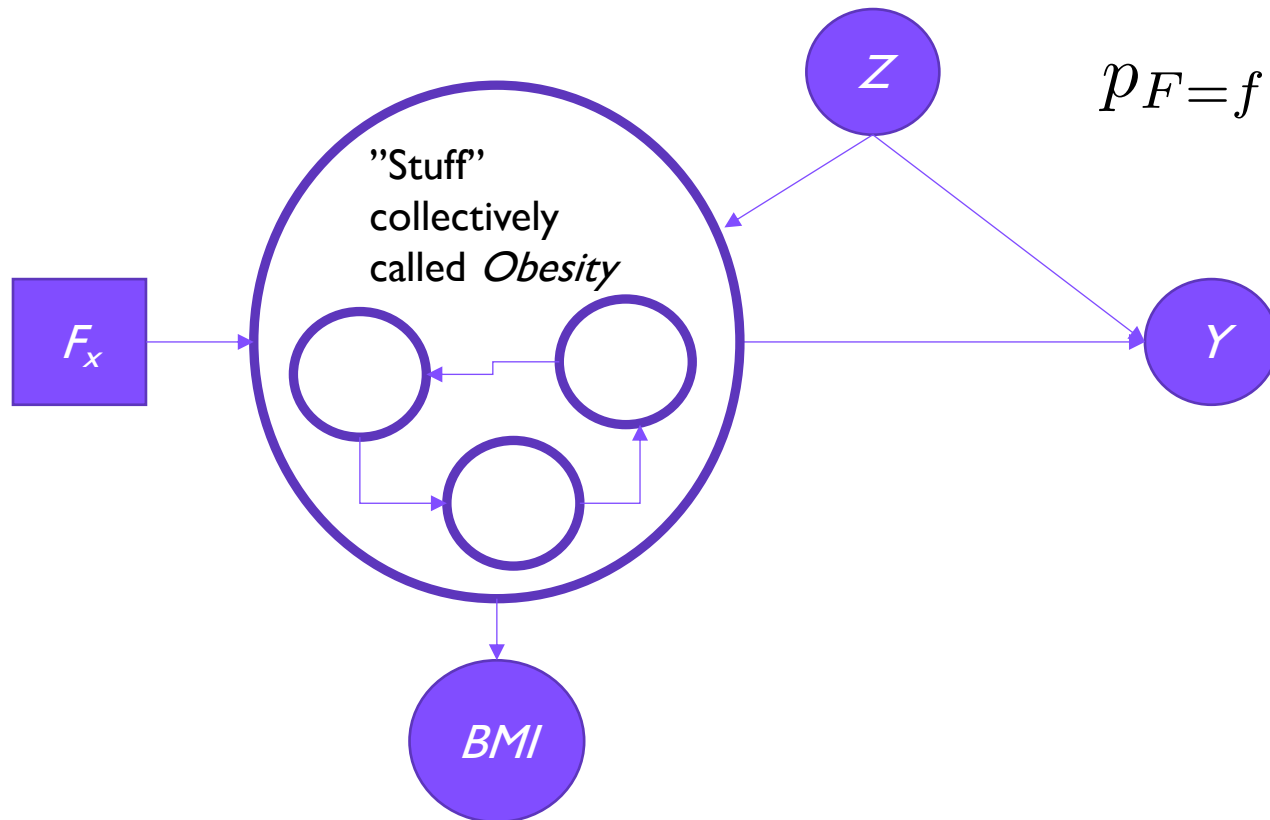
- Different people apply for jobs with different CVs, some of them get job offers.
- How to design a CV to maximize job prospects?
- Do "word-level" interventions make sense? Layout level?
- What about "Theseus' ship" problems?

...and beyond

- What does it mean to say “temperature increases will cause the melting of polar caps”. What is temperature?
- To what extent does it mean to say “obesity causes heart disease”?
- What about constructs on social science in general? Inflation, democratization levels etc.

A first sketch

- Obesity: a useful concept to the extent it captures a range of interventions?



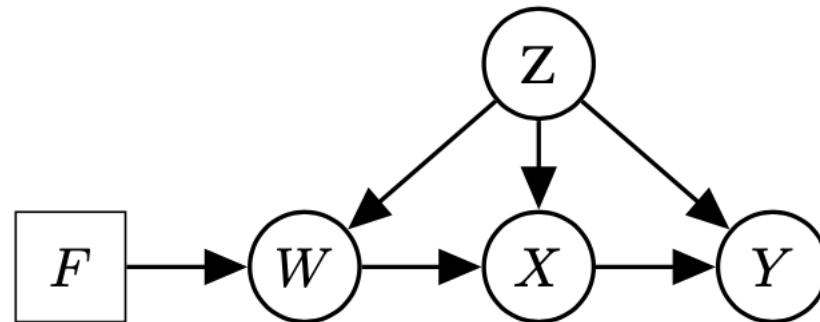
$$p_{F=f}(y \mid bmi, z) \approx p_{F=f'}(y \mid bmi, z) ?$$

What if f is “exercise > 3h a week” and f' is “cut this rebel scum’s hand with a light saber?”



Useful causation from invariance

- Q. To which extent is useful to have a non-trivial variable X postulated as a *cause* of Y when $do(x)$ is not defined?
- A. To the extent its relation with outcome Y is invariant to changes in *actionable* variables W , a variable for which $do(w)$ is defined.



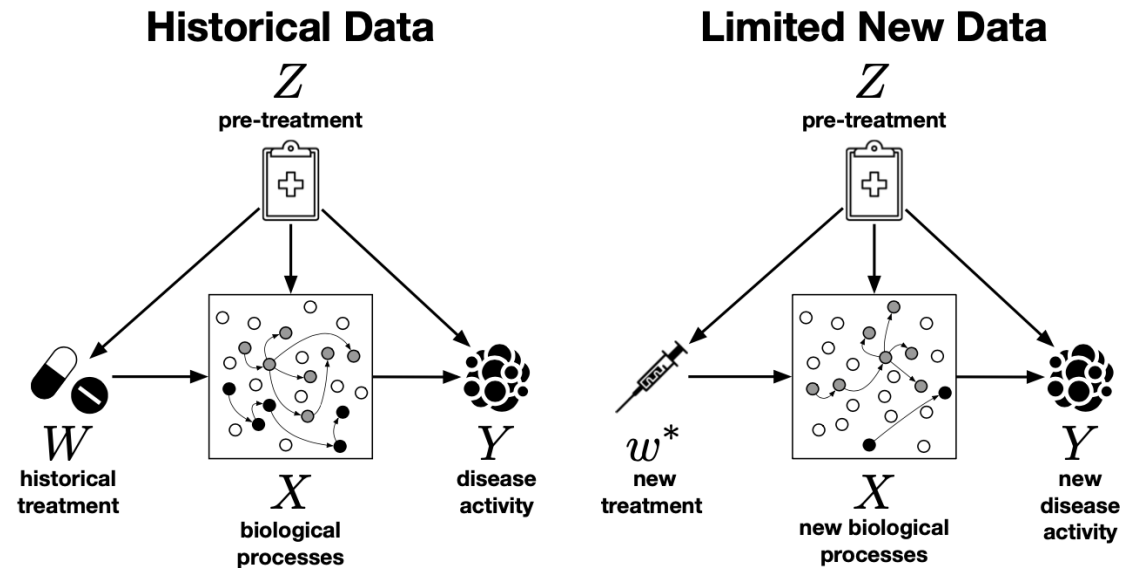
What goes in F ?

$F \in \mathcal{F} \equiv \{idle, do(W = w_1), do(W = w_2), \dots, do(W = f_1(X)), do(W = f_2(X)), \dots\}$

- The conditional independencies postulated are predicated on the support/sample space of W .

Implications

If we postulate/test that X shields W from Y (given Z) then we can predict that will happen under values w^* of W for which we may have knowledge of $p(x | w, z)$ even if no data on (w^*, y) has been jointly collected before.



Goals

- Predict Y from unseen $W = w^*$ and Z . In this case, we marginalize X given W and Z .
- Provide insights on “what in X ” “causes Y ”, postulated as transformations of X which “mediate” W .

Setup

- Model structure

$$Y = \theta_0 + \sum_{i=1}^d \theta_i \phi_i(X, Z) + \epsilon,$$

- Prediction

$$\mathbb{E}[Y \mid do(w), z] = \theta_0 + \sum_{i=1}^d \theta_i \mathbb{E}[\phi_i(X, Z) \mid w, z].$$

Learn this by supervised learning

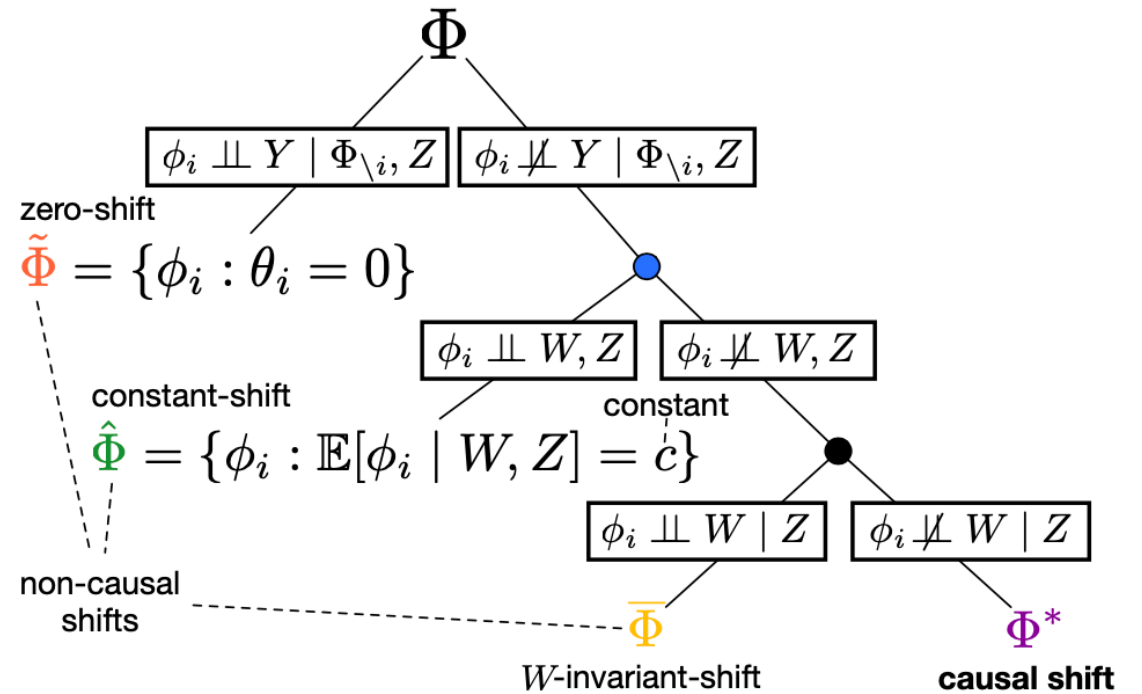


Pragmatic mediation

- Discover features ϕ such that

(i) $\phi_i(X, Z) \not\perp W \mid Z,$

(ii) $\phi_i(X, Z) \not\perp Y \mid \{\Phi \setminus i, Z\}$



Experiments

Table 1. General description of experimental setups.

| | ImagePert | Humicroedit | DREAM5 |
|----------|---|---|-----------------------------------|
| Z | pre-perturbation image | original news headline (GloVe avg. vector) | baseline gene expression |
| W | location of normal distribution for perturbation | new entity edit (GloVe vector) | transcription factor out-degree |
| X | post-perturbation image | edited headline (GloVe avg. vector) | post-intervention gene expression |
| Φ | convolution windows over X | funniness hypotheses (Hossain et al., 2019) | change in kernel eigengene |
| Y | intensity of pixels, linear combination of Φ | funniness score, via linear combination of Φ | linear combination of Φ |

Image perturbation example

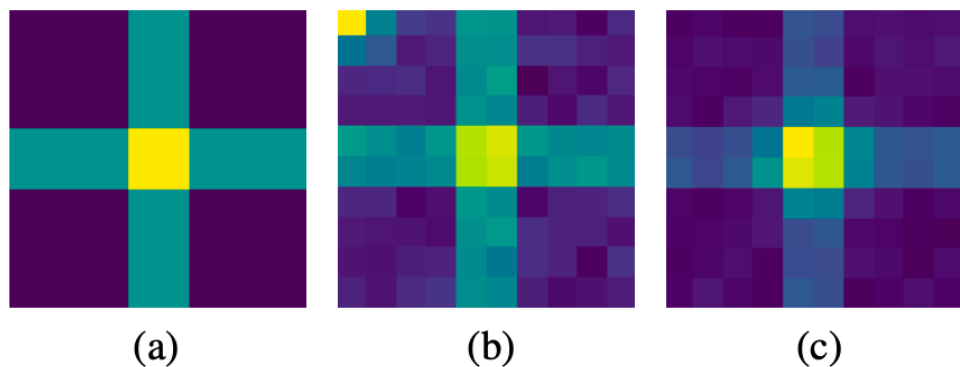
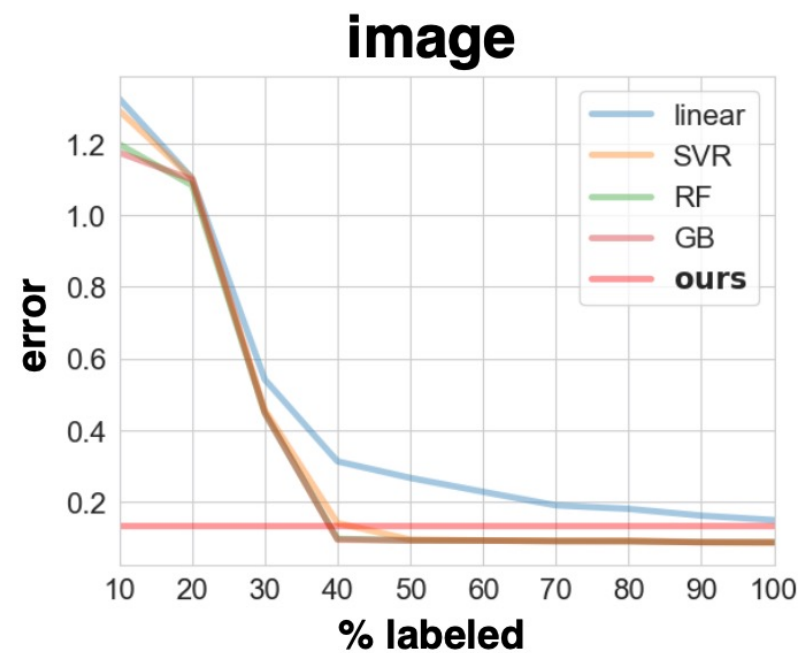
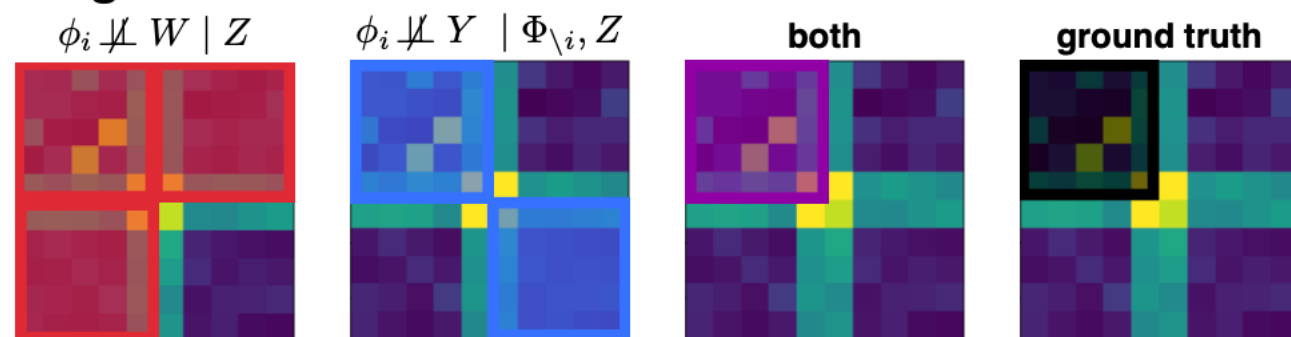


Figure 4. Visual example for image perturbation dataset. (a) Example image (z). (b) Same image, post-perturbation (x , in response to w). (c) Same image under a new perturbation regime, which we leave for the test set (x' , in response to w').



image



Humorous edits example

- To which extent changing a single word in a news headline make it humorous?
- Z : original headline
- W : word to change, then aggregate to “topic”
- X : resulting headline, encoded as linguistic features (sentiment, cosine distance of vector representation of original and replacement word etc.)
- Y : in the original, judgements of humorousness. Here, synthetic

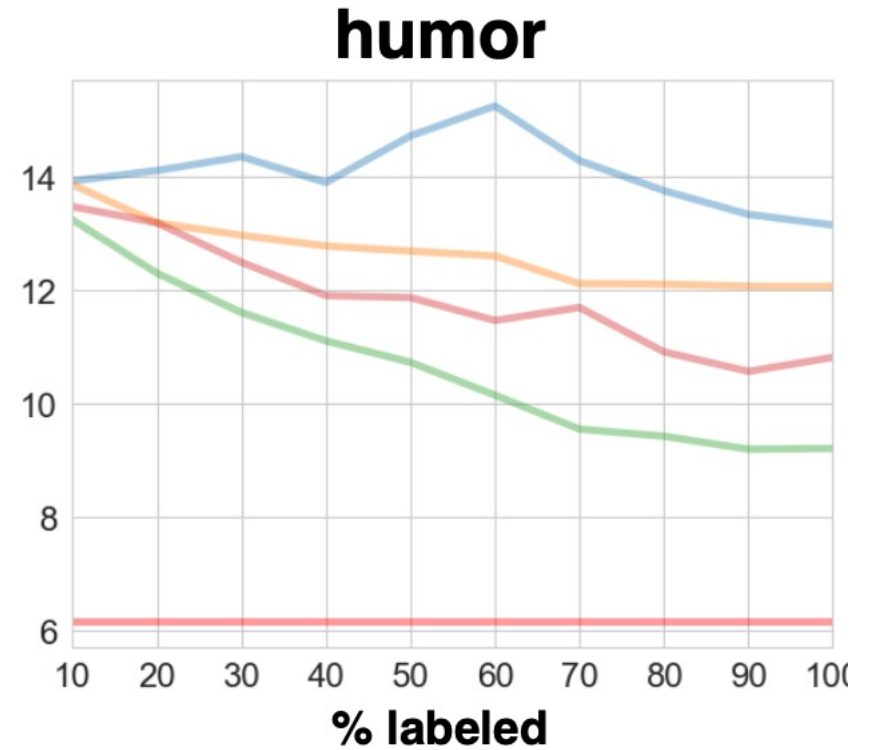
Humorous edits example

ORIGINAL: [Eric Trump](#): Those Who **Oppose** My [Dad](#) Are ' Not Even [People](#) '
EDITED: Eric Trump: Those Who **support** My Dad Are ' Not Even People '
 Substitute:

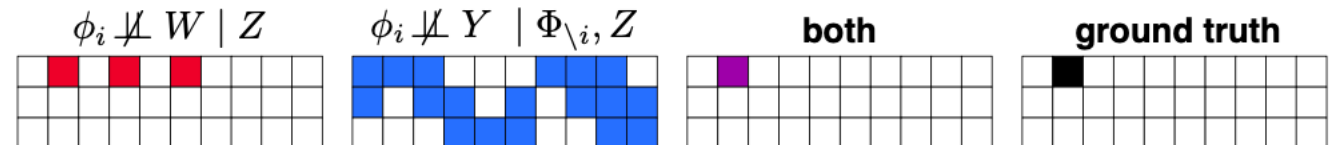
(a) The Headline Editing Task.

Orig: EU says **summit** with Turkey provides no answers to concerns
Edit: EU says **gravy** with Turkey provides no answers to concerns

0 (Not Funny)
 1 (Slightly Funny)
 2 (Moderately Funny)
 3 (Funny)



humor



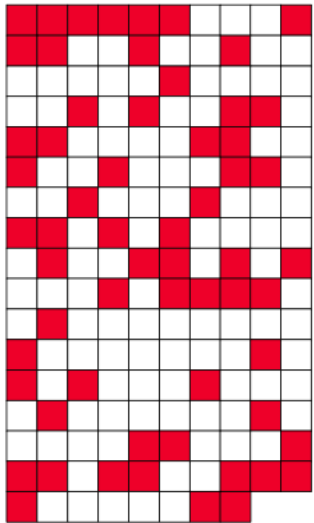
Gene expression example

- Semi-simulated from the DREAM5 challenge.
- Z : baseline expression data
- W : gene knock output action
- X : post-intervention expression, with features being differences in top eigenvectors of Z and X
- Y again is a synthetic response

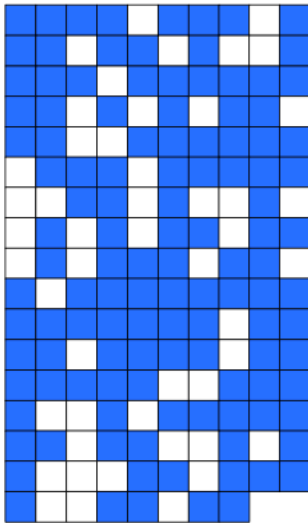
Gene expression example

genomics

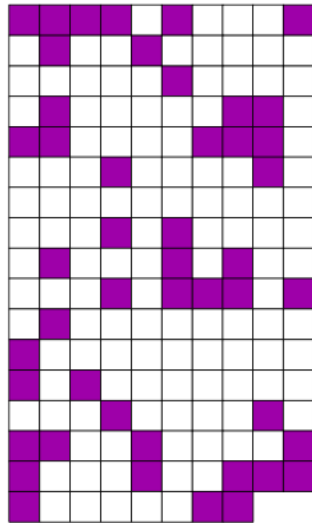
$\phi_i \not\perp W | Z$



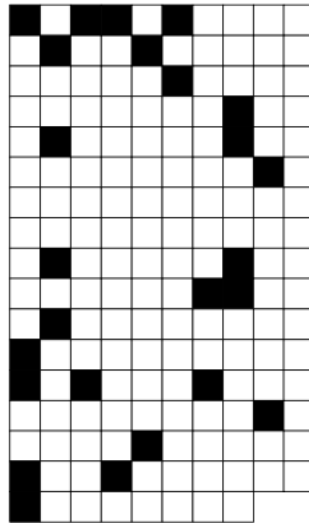
$\phi_i \not\perp Y | \Phi_{\setminus i}, Z$



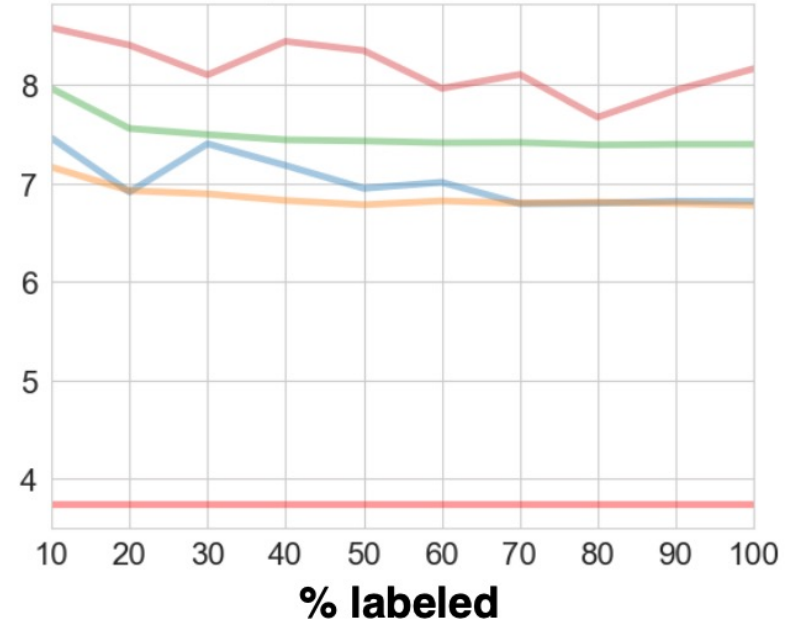
both



ground truth



genomics



Pragmatic mediation: summary

- Invariance under intervention continues to be the guiding principle.
- The lack of a perfect intervention on some X does not mean we cannot profit from the concept of “ X as a cause”: it just means we can’t rely on a default value (“do(x)”) for our intervention space F_x .
- More links to domain adaptation should be explored. On the other hand, it is less clear though how to streamline what we learn about the selection of pragmatic mediators.
- Code available at <https://github.com/limorigu/ComplexCauses>

Conclusion

- Causal inference from observational data has practical limits we won't see in e.g. supervised learning.
- We can still take a leaf from progress in that area to bring new life into classical problems of causal inference.
- More in the spectrum of constitutive vs causal should be explored. Much in medicine, for instance, is construct-based (“syndrome” etc.) but operationalization of constructs as causal factors is not just bean-counting. It has practical implications.



THANK YOU