

Causal Interpretations of Black Box Models

Dylan Randle, Bill Zhang

Outline

1. Background and Motivation
2. Partial Dependence Plots and Possible Causal Interpretation
3. Review of Graphical Models and Structural Equation Model
4. Back-Door Adjustment
5. Mediation Analysis & Individual Conditional Expectation
6. Example Applications

Mathematical Setup

- Assume some process $Y = f(X, e)$ where e is some random noise
- f is the “function of nature”, and we approximate it with g
- g is often chosen to be a “black box” (for performance) which is hard to interpret
 - “Data modeling culture” assume parametric form, often interpretable
 - “Algorithmic modeling culture” train complex models (e.g. random forest, neural nets) to maximize performance, often black box



Two Different Goals in Analysis

1. **Prediction:** predict Y given X

- a. **Associational:** “How many people take aspirin when they have a headache?”

2. **Science:** extracting information about the law of nature “ f ”

- a. **Counterfactual** (causes of effects): “My headache has gone. Is it because I took aspirin?”
- b. **Interventional** (effects of causes): “I have a headache. Will it help if I take aspirin?”

Notions of feature importance

“What is the importance of component p of X ?”

1. Impact on a model's prediction

- Coefficient of (normalized) variable in linear regression
- Analysis of variance

2. Impact on predictive accuracy

- Random permutation of feature (general)
- Contribution to decreasing impurity (tree-specific)

3. Causality

- If we were to “intervene” on feature X_i , how much would Y change

Here we focus on **causality**, which is concerned with **science** — not just prediction

Contribution

Successful causal interpretation requires:

1. A good predictive model g of the law of nature f
2. Satisfying the back-door condition regarding causal structure
3. Visualizations like PDP and ICE

Partial Dependence Plot (PDP)

- Assume we have a model $Y = g(x)$
- We want to know the dependence of $g(x)$ on a subset x_S of features
- Let x_C denote the complement set of features
- The PDP is the **expectation of g** taken over the marginal of x_C

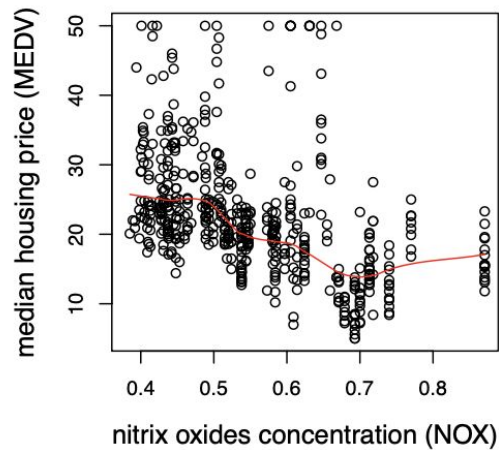
$$g_S(x_S) = E_{X_C}[g(x_S, X_C)] = \int g(x_S, x_C) dP(x_C)$$

We “marginalize over” x_C

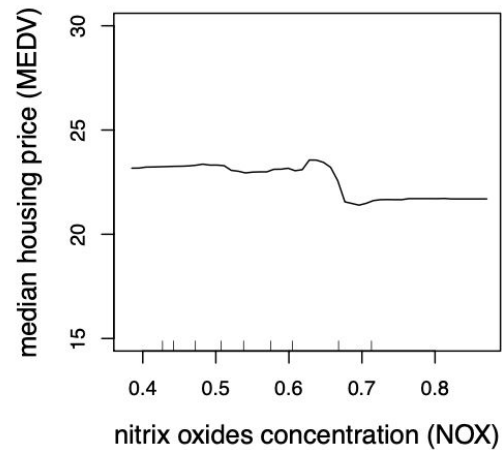
$$\bar{g}_S(x_S) = \frac{1}{n} \sum_{i=1}^n g(x_S, X_{iC})$$

In practice, take the average for a fixed x_S

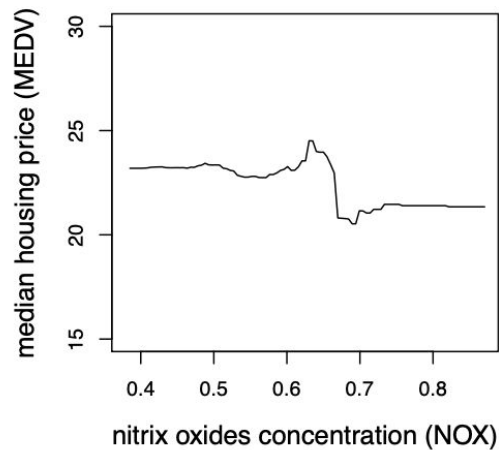
Scatter plot



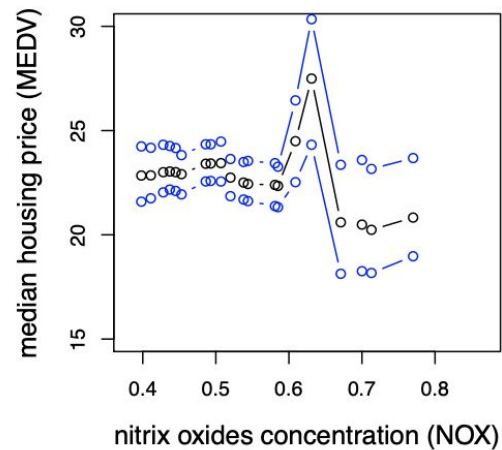
Random Forest



GBM



BART



A Curious Coincidence

- Assume $g(x)$ is the expectation of the response Y
- Assume the conditioning set C is the complement set of S
- Then the formula for PDP is identical to the “back-door” adjustment in causal inference

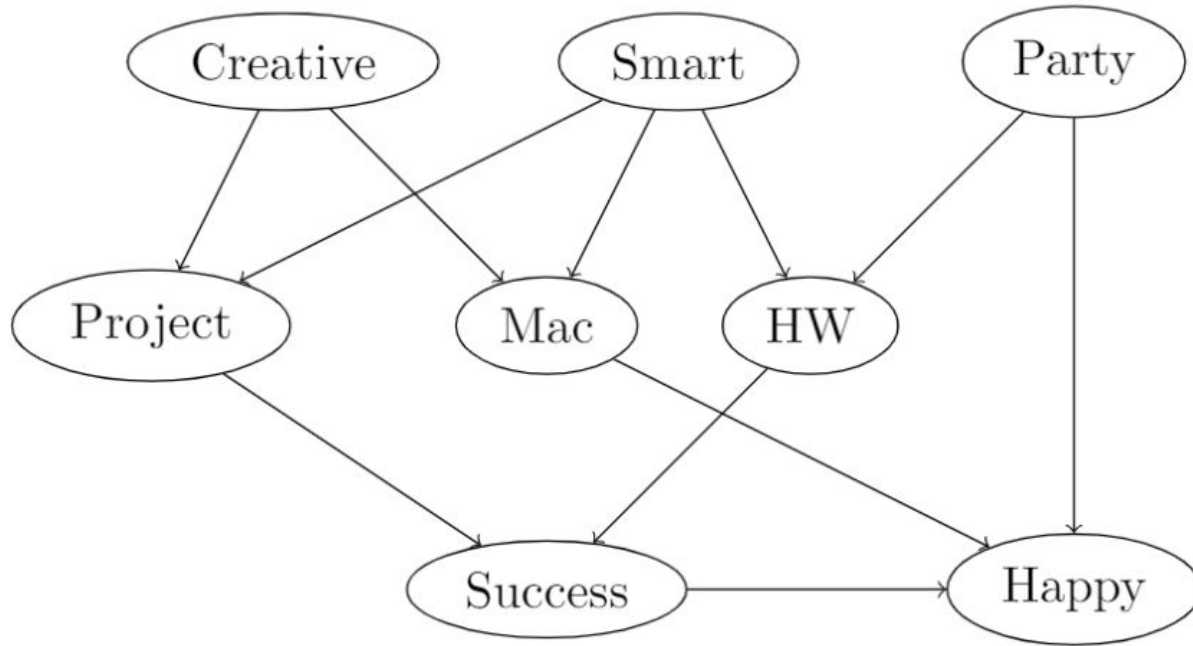
$$E[Y|do(X_S = x_S)] = \int E[Y|X_S = x_S, X_C = x_C] dP(x_C)$$

$$\implies E[Y|do(X_S = x_S)] = g_S(X_S)$$

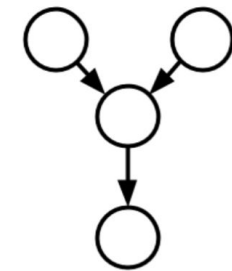
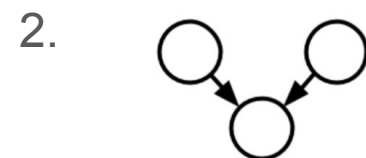
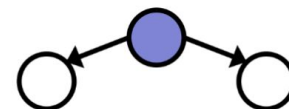
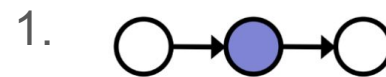
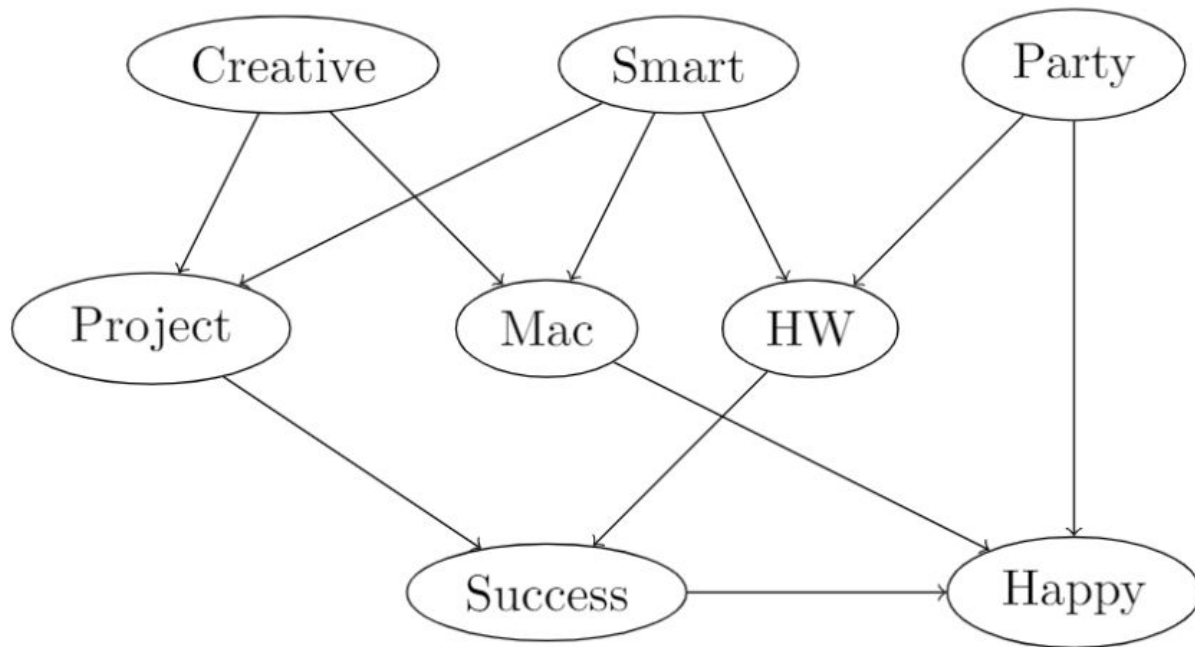
A Causal Interpretation of Black Box Models?

- This coincidence suggests that PDP is perhaps an unintended attempt to causally interpret black-box models
- If true, this would be a big deal as PDP plots are easy to compute and widely implemented (e.g. in scikit-learn)
- Now we will see under what circumstances a causal interpretation can be made with PDP

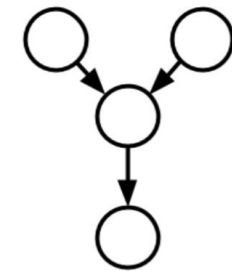
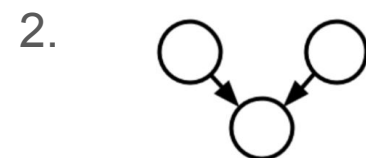
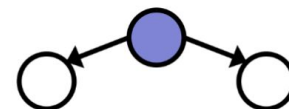
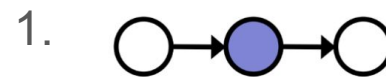
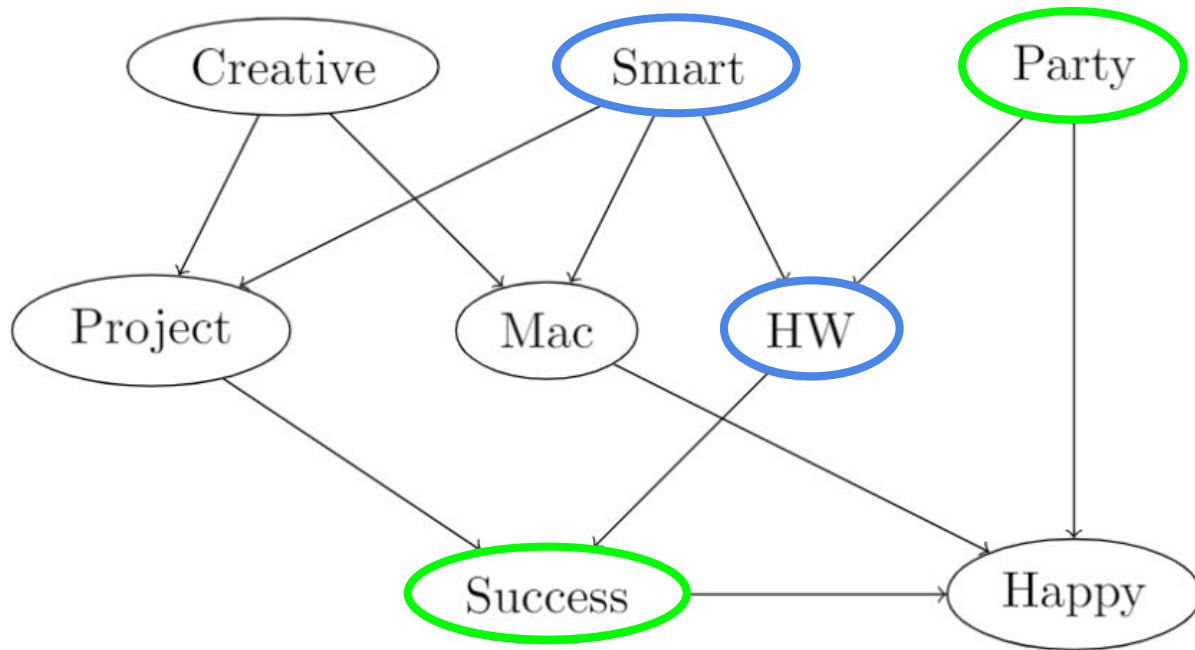
Review: Graphical Models



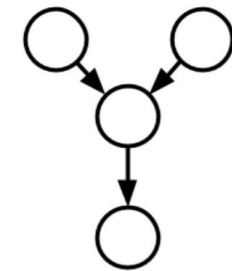
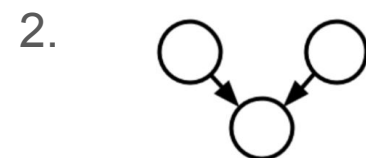
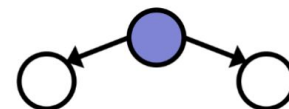
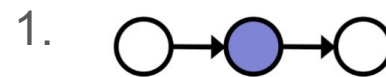
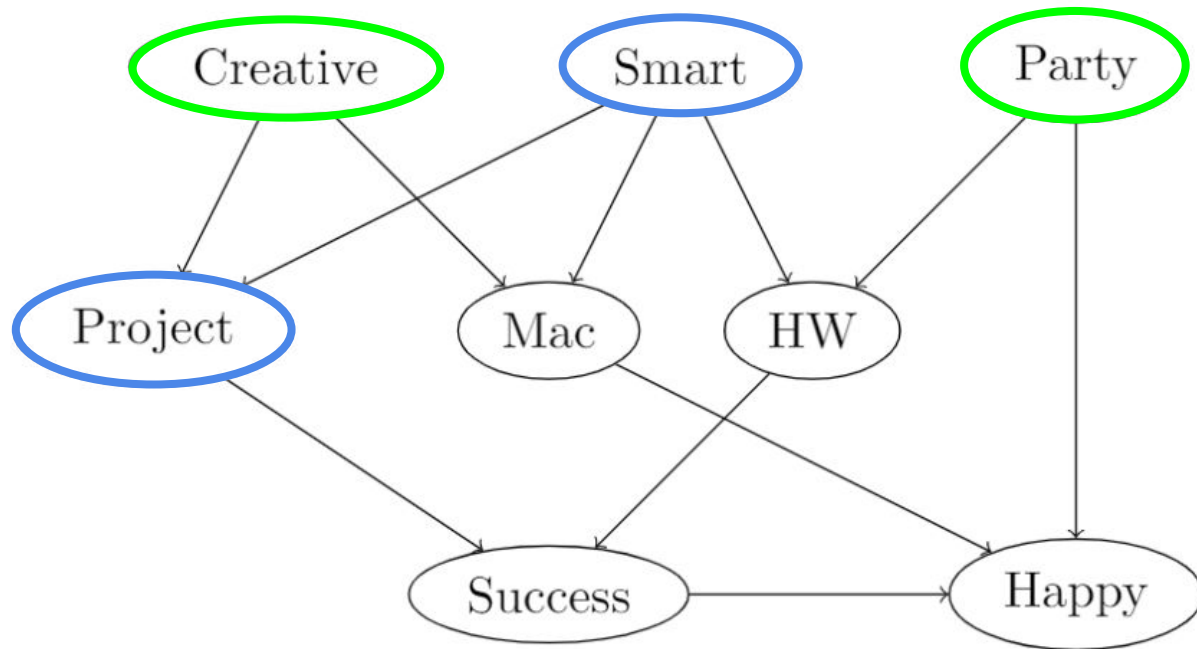
Review: D-Separation



Review: D-Separation



Review: D-Separation



Structural Equation Model (NPSEM)

- Let $G = (V, E)$ be a directed acyclic graph
- Here $V = \{X_1, X_2, X_3, \dots, X_p, Y\}$
- Assume each variable generated by system of nonlinear equations f and random noise ϵ

$$Y = f(\text{pa}(Y), \epsilon_Y),$$
$$X_j = f_j(\text{pa}(X_j), \epsilon_j),$$

- Where $\text{pa}(Y)$ is the parent set of Y
- These are structural models and convey causality

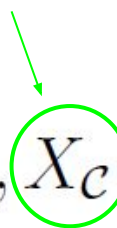
Structural (Law of Nature)



Equation	Regression	Causation
$\text{Grade} = \alpha + \beta (\text{Hours studied}) + \epsilon$	Yes	Yes
$\text{Hours studied} = \alpha' + \beta' (\text{Grade}) + \epsilon'$	Yes	No

The Back-Door Adjustment Formula

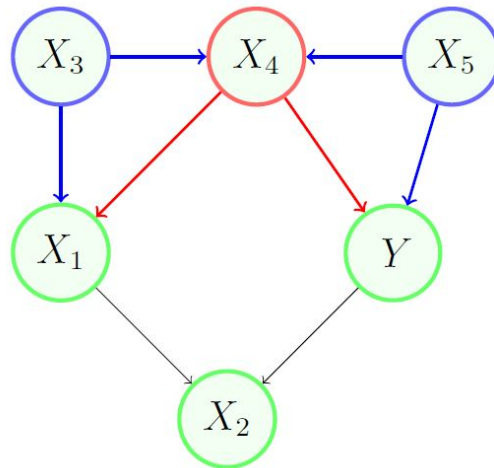
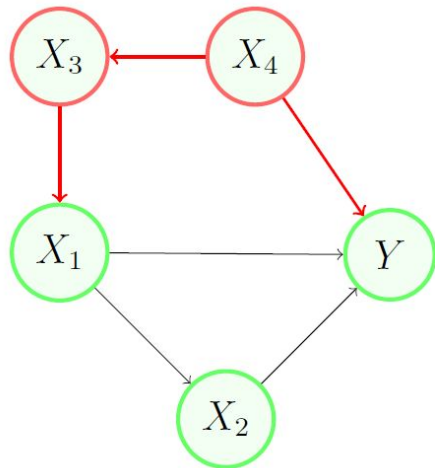
What criteria does X_C need to satisfy?

$$P(Y|do(X_S = x_S)) = \int P(Y|X_S = x_S, X_C = x_C) dP(x_C).$$


The Back-Door Criterion

1. No node in X_C is a descendant of X_S
2. X_C d-separates (blocks) every “back-door” path between X_S and Y

Here, a “back-door” path is any undirected path that has an arrow into X_S



Thus, PDP estimates **causal effects** of X_S
on Y ...

Only if the **complement set C** satisfies the
back-door criterion...

Otherwise, **not** causally interpretable!

Boston Housing Data

- X_S = nitrix oxides concentration (NOX, air pollution level)
- Y = median value of homes
- X_C = crime rate, average number of rooms, etc. (all other predictors)

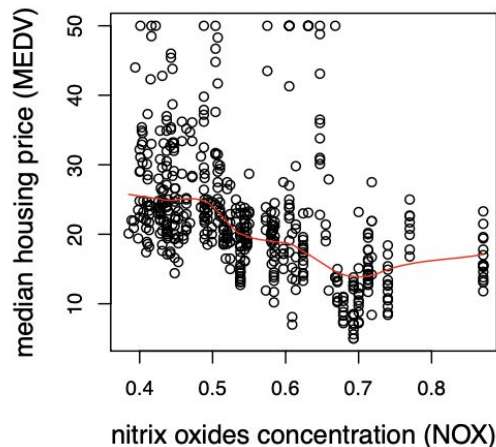
Then:

- **Assume** NOX is not a cause of other predictors X_C
- **Assume** the other predictors X_C block all back-door paths

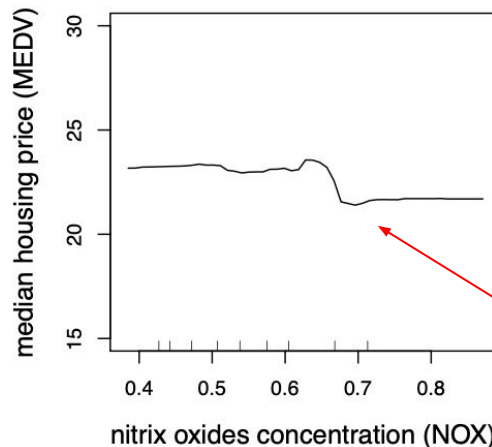
Then:

- PDP estimates causal effect!

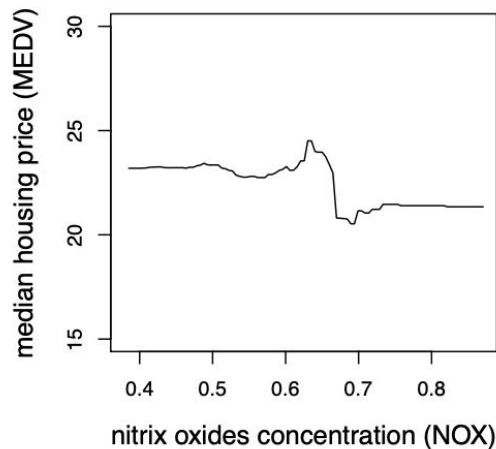
Scatter plot



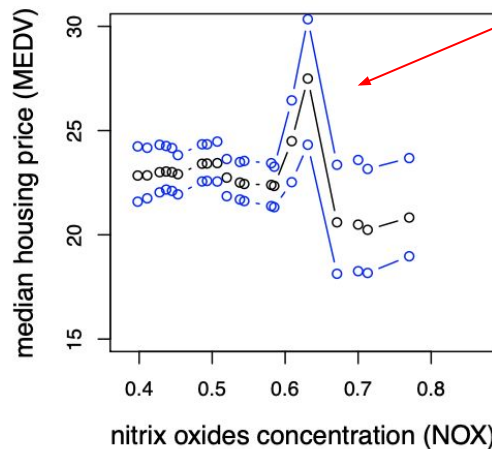
Random Forest



GBM



BART



Actually causal? Further investigation required...

Mediation Analysis

- What if we don't know the causal model, and it is hard to determine the appropriate X_C ?
- Assume we know some variables in C are causal descendants of X_S
- Measure how impact of X_S on Y is **mediated** through some X_M
- NPSEM:

$$X_{\mathcal{M}} = h(X_S, X_C, \epsilon_{\mathcal{M}})$$

$$Y = f(X_S, X_{\mathcal{M}}, X_C, \epsilon)$$

Mediation Analysis

- For some fixed x_S and x'_S
- **Total effect** is the total causal impact of X_S on Y

$$TE = E[f(x_S, h(x_S, X_C, \epsilon_M), X_C, \epsilon)] - E[f(x'_S, h(x'_S, X_C, \epsilon_M), X_C, \epsilon)]$$

- **Controlled direct effect** is the causal impact of X_S on Y given $X_S = x_S$

$$CDE(x_M) = E[f(x_S, x_M, X_C, \epsilon)] - E[f(x'_S, x_M, X_C, \epsilon)]$$

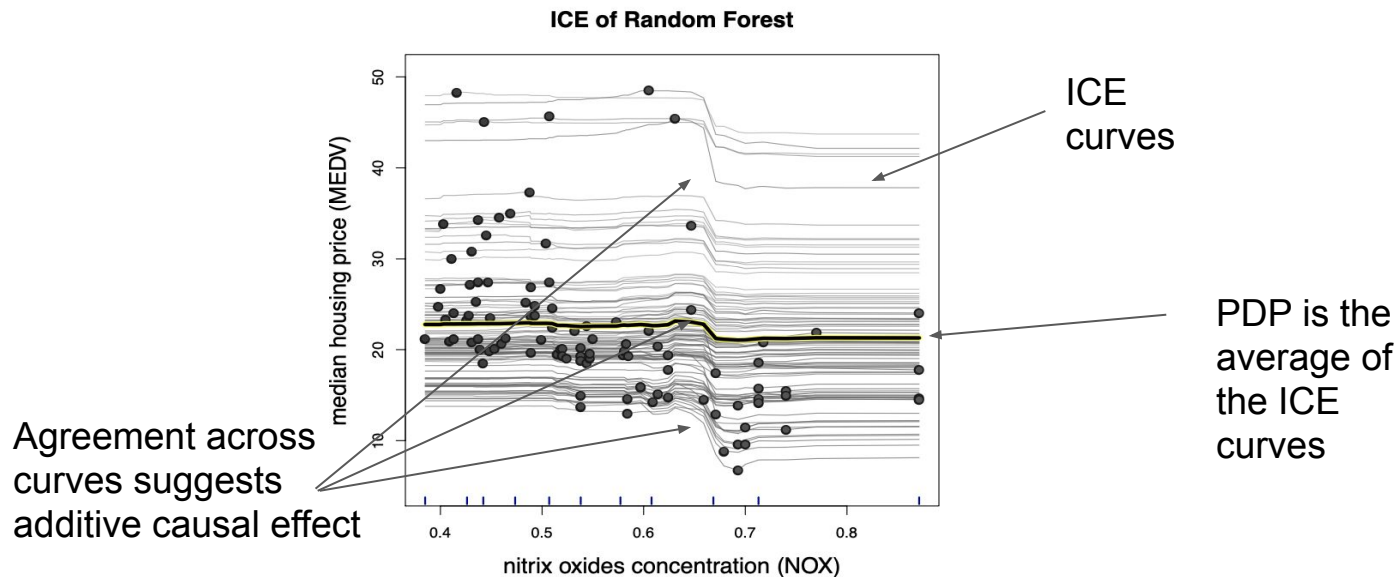
Mediation Analysis

- If C satisfy back-door criterion, PDP visualizes total effect
 - Here C are more generally any set of variables that satisfy back-door criterion, not necessarily the complement of S
- ICE essentially plots $CDE(x_M)$ for many x_M
- If the causal effect is **additive**, then CDE does not depend on mediation variables X_M

$$CDE(x_M) \equiv f_S(x_S) - f_S(x'_S)$$

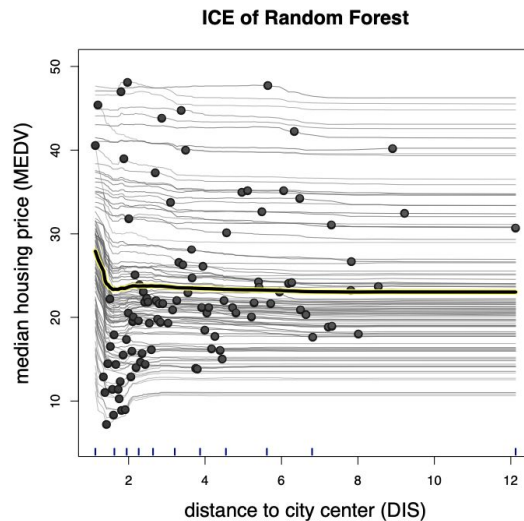
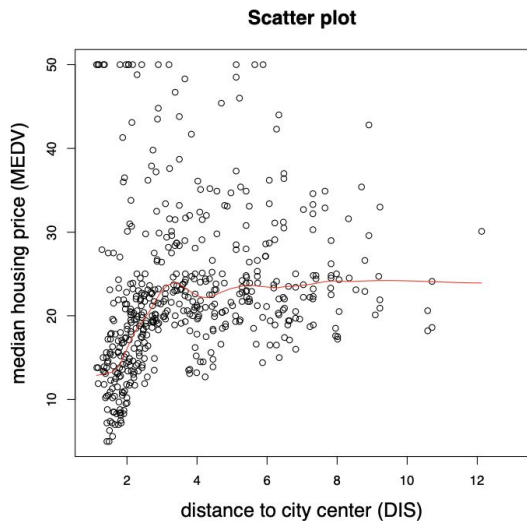
Individual Conditional Expectation (ICE)

- Plots each individual curve (for each X_C) instead of averaging over them
- If curves are in agreement, evidence for additive causal relationship (i.e. evidence CDE does not depend on mediation variables)



Boston housing, continued

- Median price (MEDV) vs distance to city center (DIS)
- Scatter plot shows increasing price further away (likely **indirect**, e.g. higher crime rates closer to city center reduces prices)
- ICE shows that the **direct effect** of DIS has an opposite trend



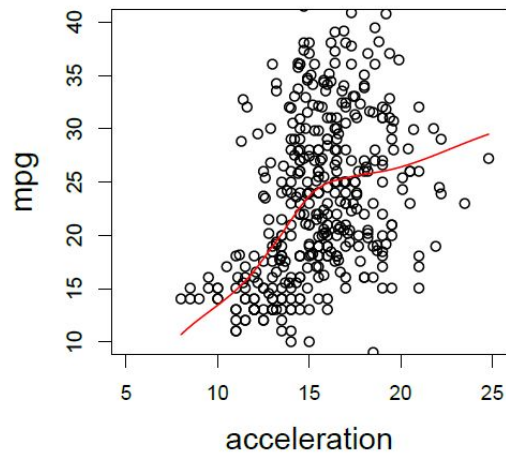
Automobile MPG Data

- X_S = maximum acceleration
- Y = miles per gallon (MPG)
- X_C = number of cylinders, displacement, weight, etc. (all other predictors)

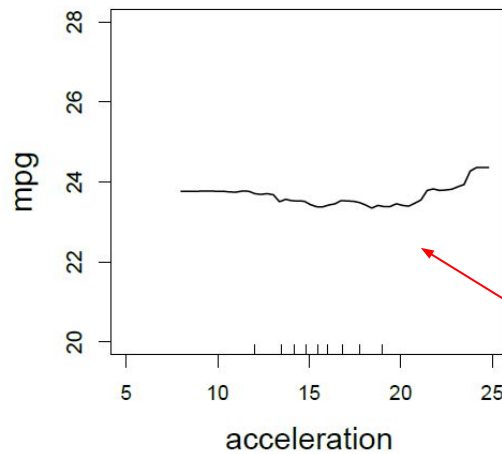
Then:

- **Assume** that acceleration is a causal descendant of all other variables

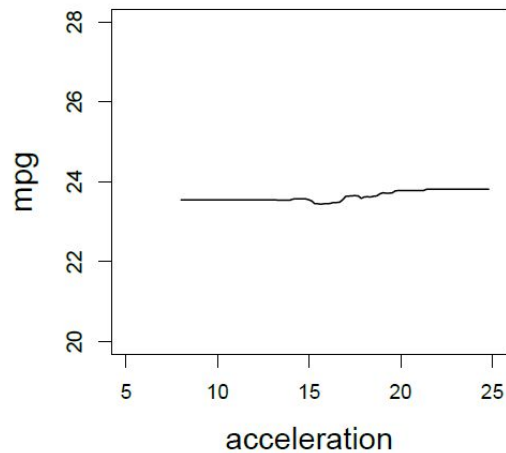
Scatter plot



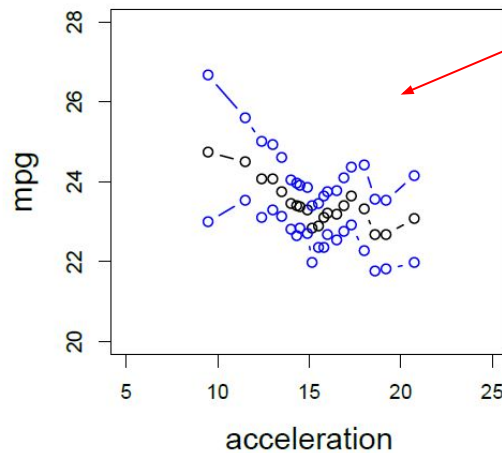
Random Forest



GBM



BART



Conflicting interpretations
of causal relationship

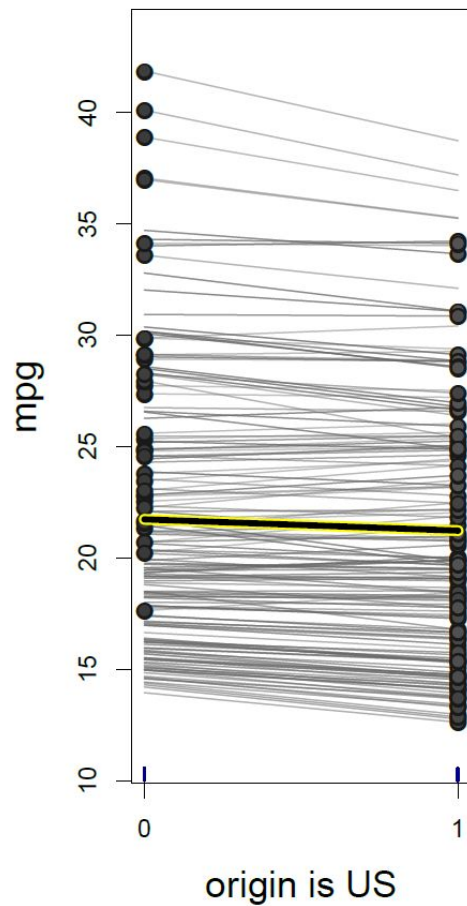
Automobile MPG Data

- X_S = place of origin (US, Europe, or Japan)
- Y = miles per gallon (MPG)
- X_C = number of cylinders, displacement, weight, etc. (all other predictors)

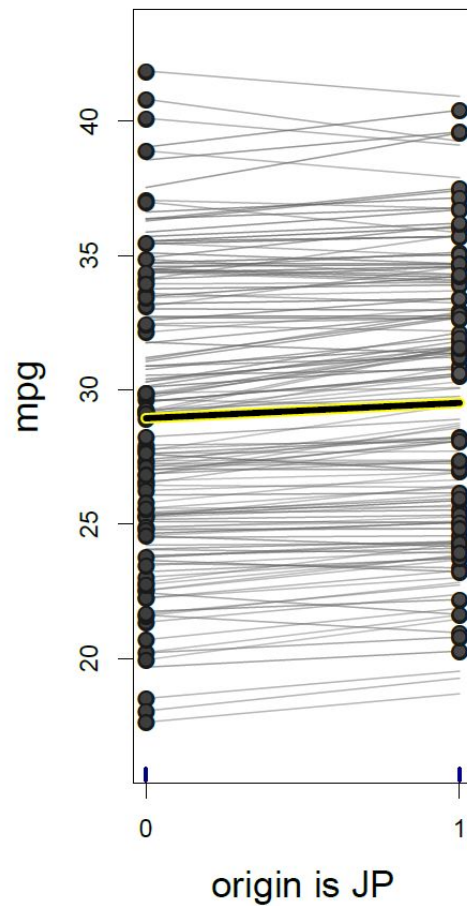
Then:

- **Assume** that origin is a causal ancestor of all other variables

ICE of Random Forest



ICE of Random Forest



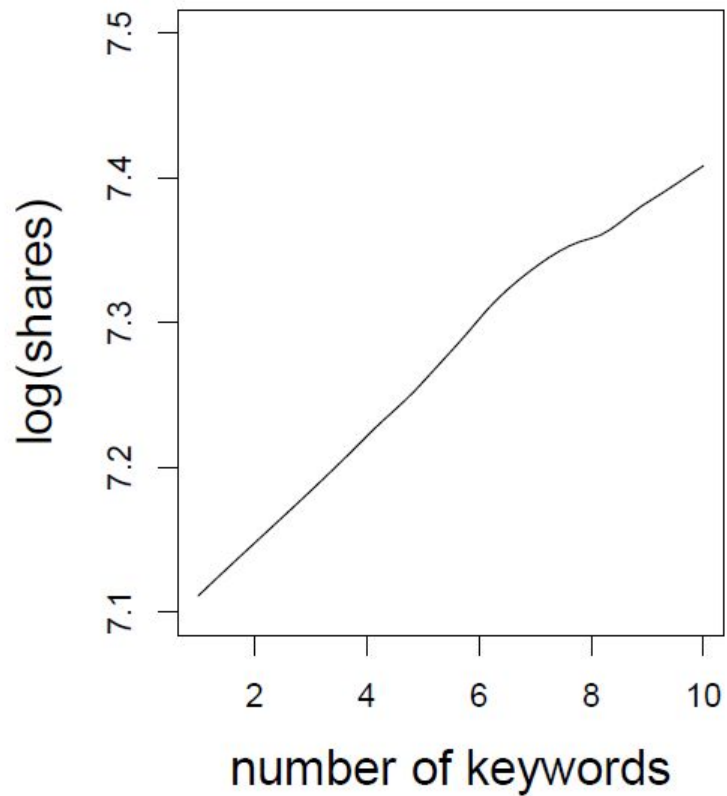
Online News Popularity Data

- X_S = number of keywords, title sentiment polarity
- Y = number of shares on social media
- X_C = all other predictors

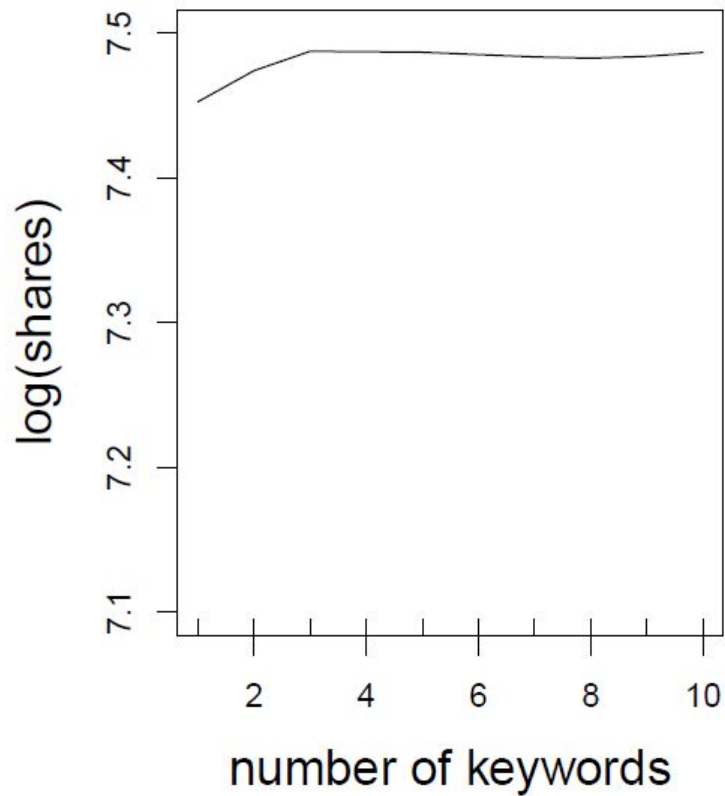
Then:

- **Assume** that they are causal descendants of all other variables because they are determined close to time of publication

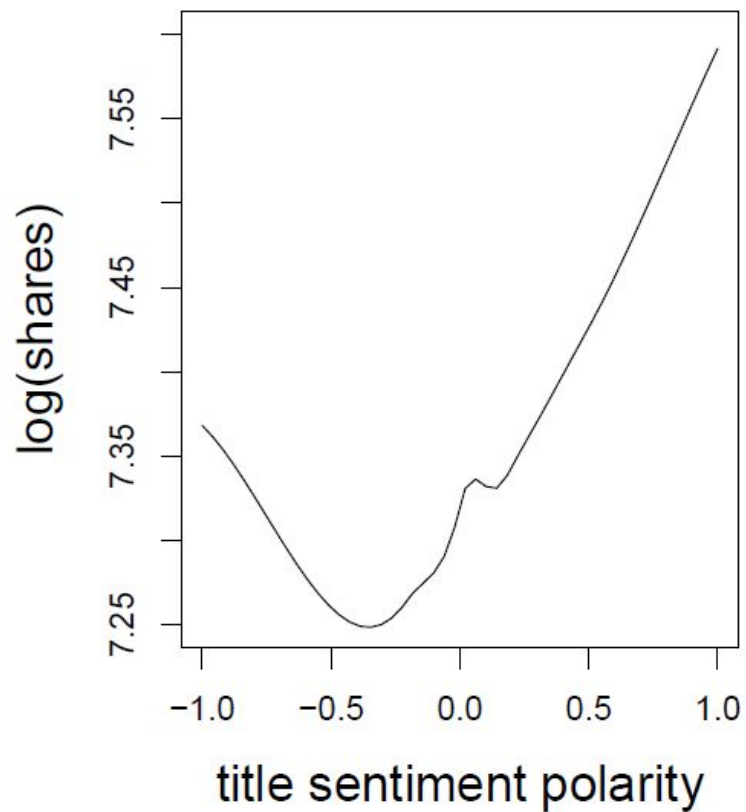
LOESS (conditional expectation)



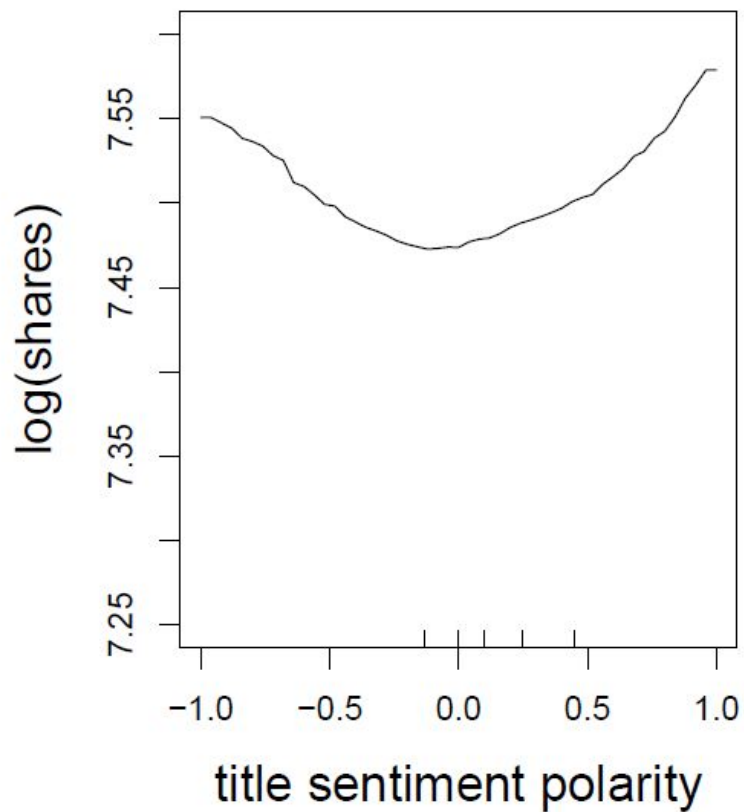
Random Forest (partial dependence)



LOESS (conditional expectation)



Random Forest (partial dependence)



In Conclusion

Successful causal interpretation requires:

1. A good predictive model g of the law of nature f
2. Satisfying the back-door condition regarding causal structure
3. Visualizations like PDP and ICE

Questions?

Learning Cost-Effective and Interpretable Treatment Regimes

Sean McGrath

sean_mcgrath@g.harvard.edu

Parth Mehta

parthmehta@college.harvard.edu

Alexandra (Ola) Zitek

zyteka@mit.edu

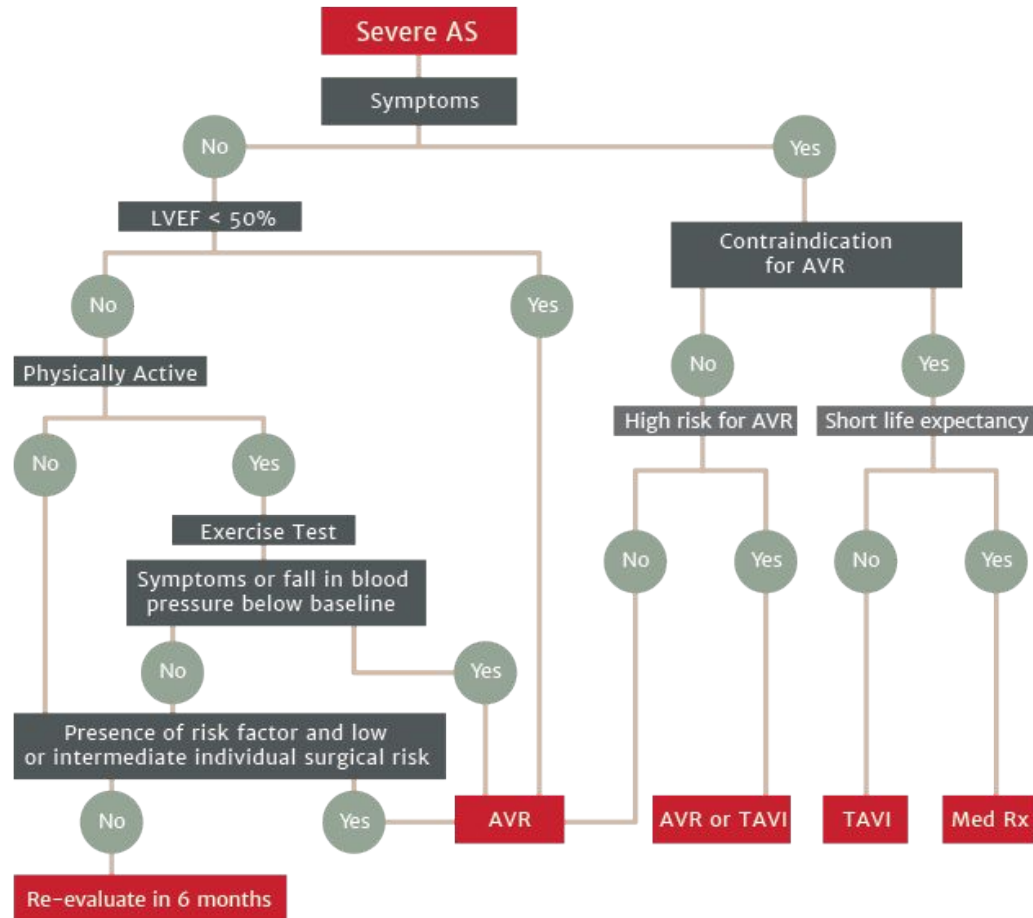
Himabindu Lakkaraju

Cynthia Rudin

Contributions

- Main contribution: Novel method for learning treatment regimes that maximize patient outcomes **and minimize costs** (in gathering information, treatment)
- Other contributions:
 - **Interpretable** treatment regimes
 - First application to adjudication bail decisions
 - Performed better than state-of-the-art baselines

Motivation



Related Work: Treatment Regimes

- *Treatment Regime*: A function that maps patient characteristics to an assigned treatment

If Spiro-Test=Pos and Prev-Asthma=Yes and Cough=High then C

Else if Spiro-Test=Pos and Prev-Asthma =No then Q

Else if Short-Breath =Yes and Gender=F and Age \geq 40 and Prev-Asthma=Yes then C

Else if Peak-Flow=Yes and Prev-RespIssue=No and Wheezing =Yes, then Q

Else if Chest-Pain=Yes and Prev-RespIssue =Yes and Methacholine =Pos then C

Else Q

Related Work: Treatment Regimes

- *Treatment Regime*: A function that maps patient characteristics to an assigned treatment
- *Optimal Treatment Regime*: A treatment regime that maximizes the average patient outcomes had all patients followed that regime.
- Challenges in estimating the effectiveness of treatment regimes:
 - Correlation \neq Causation
 - What assumptions are necessary for the effectiveness of treatment regimes?

Related Work: Treatment Regimes

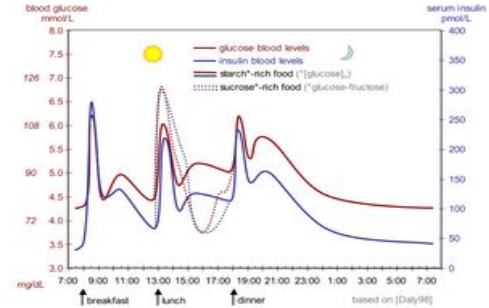
- Regression-based approaches
 - Model condition distribution of outcome given the past treatment and covariates.
 - Select the treatment regime that maximizes the expected outcome
- Policy-search-based methods (classification-based approaches)
 - Estimate marginal mean of outcome for all treatments regimes
 - Select the treatment regime that maximizes the expected value
 - E.g., adjustment by inverse propensity score weighting
- Limitations:
 - These approaches don't incorporate costs of gather subject information and costs of treatment!
 - Very few of these approaches produce intelligible regimes

Other Related Work

- Subgroup analyses
 - If / how treatment effects vary across subgroups of individuals
- Interpretable models
 - Many classes of models proposed for better interpretability (e.g., decision lists, decision sets)
 - Haven't been developed for and applied to model treatment effects

Framework: Input data

1. X_i



, ...)

2. Treatment and outcome of subject i , where $a_i \in \mathcal{A}$ and $y_i \in \mathbb{R}$

3. Assessment cost d and treatment cost d'

Framework: Treatment regime

The rules in π partition the dataset \mathcal{D} into $L + 1$ groups: $\{\mathcal{R}_1, \mathcal{R}_2 \cdots \mathcal{R}_L, \mathcal{R}_{\text{default}}\}$. A group \mathcal{R}_j , where $j \in \{1, 2, \cdots L\}$, is comprised of those subjects that satisfy c_j but do not satisfy any of $c_1, c_2, \cdots c_{j-1}$:

$$\mathcal{R}_j = \left\{ \mathbf{x} \in [\mathcal{V}_1 \cdots \mathcal{V}_p] \mid \text{satisfy}(\mathbf{x}, c_j) \wedge \bigwedge_{t=1}^{j-1} \neg \text{satisfy}(\mathbf{x}, c_t) \right\}. \quad (1)$$

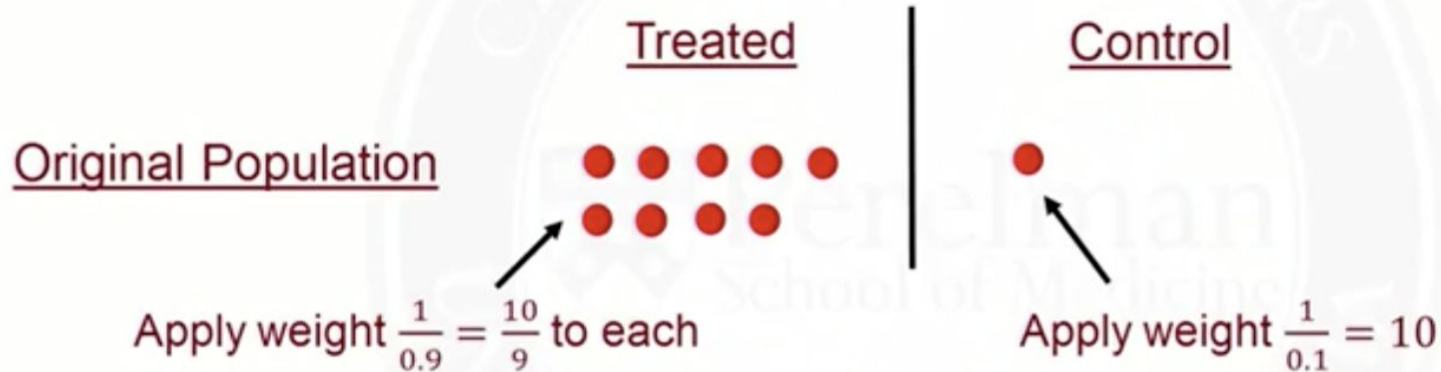
Framework: Expected outcome

$$p(E|C) \neq p(E|do(C))$$

$$P\left(\left[\text{Image of nail polish box}\right] \mid \left[\text{Image of natural nails}\right]\right) \neq P\left(\left[\text{Image of nail polish box}\right] \mid \left[\text{Image of painted nails}\right]\right)$$

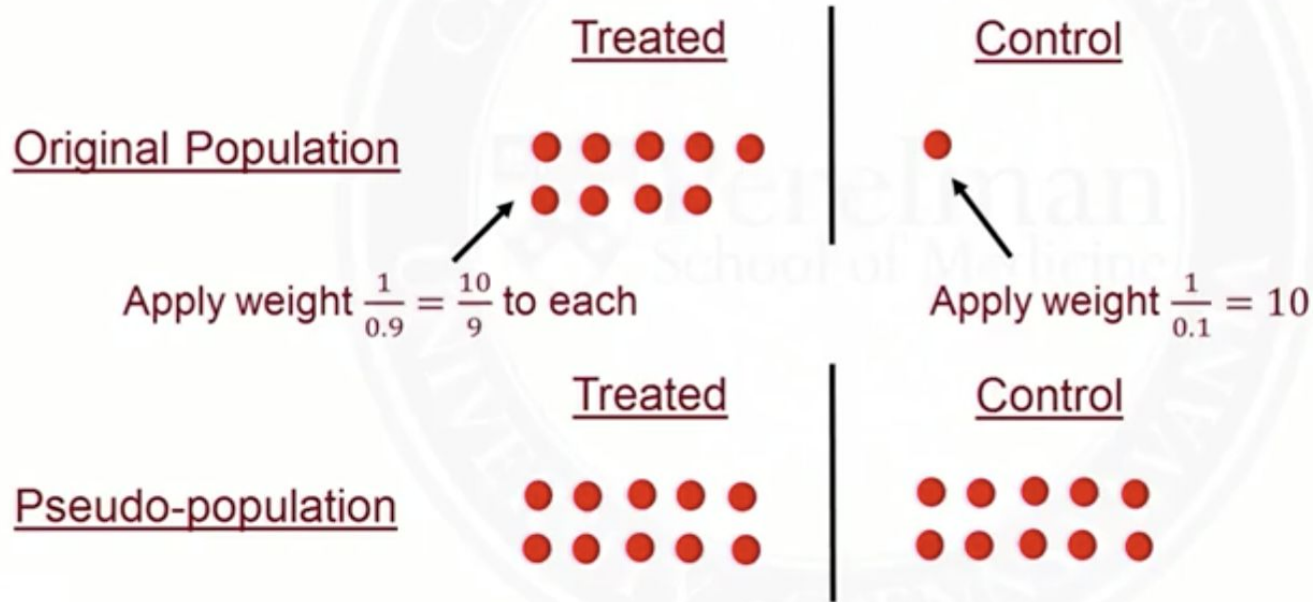
Framework: Expected outcome

- ◆ Suppose $P(A=1|X)=0.9$



Framework: Expected outcome

- ◆ Suppose $P(A=1|X)=0.9$



Framework: Expected outcome

$$g_1(\pi) = \frac{1}{N} \sum_{i=1}^N \sum_{a \in \mathcal{A}} o(i, a), \text{ where} \quad (4)$$

$$o(i, a) = \left[\frac{\mathbb{1}(a_i = a)}{\hat{\omega}(\mathbf{x}_i, a)} (y_i - \hat{y}(\mathbf{x}_i, a)) + \hat{y}(\mathbf{x}_i, a) \right] \mathbb{1}(\pi(\mathbf{x}_i) = a).$$

Framework: Expected outcome

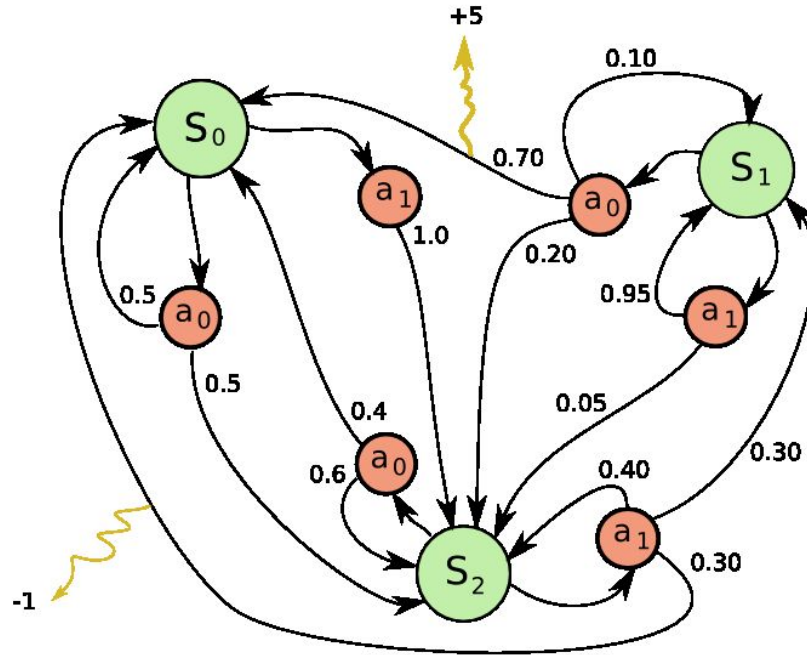
Expected assessment cost and expected treatment cost:

$$g_2(\pi) = \frac{1}{N} \sum_{i=1}^N \psi(\mathbf{x}_i).$$

$$g_3(\pi) = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i).$$

$$\arg \max_{\pi \in C(\mathcal{L}) \times \mathcal{A}} \lambda_1 g_1(\pi) - \lambda_2 g_2(\pi) - \lambda_3 g_3(\pi)$$

Framework: Optimizing the objective



Framework: Critiques

Learning Cost-Effective and Interpretable Treatment Regimes

tions of expected outcome, assessment, and treatment costs of a treatment regime π with respect to the dataset \mathcal{D} .

Expected Outcome Recall that the treatment regime π assigns a subject i with characteristics \mathbf{x}_i to a treatment $\pi(\mathbf{x}_i)$. The quality of regime π is partly determined by the expected outcome when all the subjects in \mathcal{D} are assigned treatments according to regime π . The higher the value of such an expected outcome, the better the quality of π . There is, however, one caveat to computing the value of this expected outcome – we only observe the outcome y_i resulting from assigning \mathbf{x}_i to a_i in the data \mathcal{D} , and not any of the counterfactuals. If the regime π assigns a different treatment $a' \neq a_i$ to \mathbf{x}_i , we cannot readily determine the corresponding outcome from the data.

The solutions proposed to compute expected outcomes in settings such as ours can be categorized as: adjustment by regression modeling, adjustment by inverse propensity score weighting, and doubly robust estimation. A detailed treatment of each of these approaches is presented by Lunceford et al. [22]. The success of regression-based modeling and inverse weighting depends heavily on the postulated regression model and the postulated propensity score model respectively. In either case, if the postulated models are not identical to the true models, we have biased (inconsistent) estimates of the expected outcome. On the other hand, doubly robust estimation combines the above approaches in such a way that the estimated value of the expected outcome is unbiased as long as one of the postulated models is identical to the true model and there are no unmeasured confounders. The doubly robust estimator for the expected outcome of regime π , denoted by $g_1(\pi)$, can be written as:

$$g_1(\pi) = \frac{1}{N} \sum_{i=1}^N o(i, a), \quad \text{where} \quad (4)$$

$$o(i, a) = \begin{bmatrix} \mathbb{1}(a_i = a) \\ \hat{\omega}(\mathbf{x}_i, a) \end{bmatrix} (y_i - \hat{g}(\mathbf{x}_i, a)) + \hat{g}(\mathbf{x}_i, a) \Big] \mathbb{1}(\pi(\mathbf{x}_i) = a).$$

$\hat{\omega}(x_i, a)$ denotes the probability that the subject i with characteristics \mathbf{x}_i is assigned to treatment a in the data \mathcal{D} . $\hat{\omega}$ represents the propensity score model. In practice, we fit a multinomial logistic regression model on \mathcal{D} to learn this function. Similarly, $\hat{g}(\mathbf{x}_i, a)$ denotes the predicted outcome when a subject characterized by \mathbf{x}_i is assigned to a treatment a . \hat{g} is learned in our experiments by fitting a linear regression model on \mathcal{D} prior to optimizing for the treatment regime. Note that our framework does not impose any constraints on the functional forms of \hat{g} and $\hat{\omega}$, i.e., \hat{g} and $\hat{\omega}$ could be modeled using any suitable technique.

Expected Assessment Cost Recall that there are assessment costs associated with each subject. These costs are

governed by the characteristics that will be used in assessing the subject's condition and recommending a treatment. The assessment cost of a subject i treated using the regime π is given in Eqn. 3. The expected assessment cost across the entire population can be computed as:

$$g_2(\pi) = \frac{1}{N} \sum_{i=1}^N v(\mathbf{x}_i). \quad (5)$$

It is important to ensure that our learning process favors regimes with smaller values of expected assessment cost. Keeping this cost low also ensures that the learned decision list is sparse, which assists with interpretability.

Expected Treatment Cost The treatment cost for a subject i who is assigned treatment using a regime π is given in Eqn. 2. The expected treatment cost across the entire population can be computed as:

$$g_3(\pi) = \frac{1}{N} \sum_{i=1}^N c(\mathbf{x}_i). \quad (6)$$

The smaller the expected treatment cost of the regime, the more desirable it is in practice. We present the complete objective function below.

Complete Objective We assume access to the following inputs: 1) the observational data \mathcal{D} ; 2) a set \mathcal{FP} of frequently occurring patterns in \mathcal{D} . Recall that each pattern corresponds to a conjunction of one or more predicates. An example of such a pattern is “Age $\geq 40 \wedge$ Gender=Female”. In practice, such patterns can be obtained by running a frequent pattern mining algorithm such as Apriori [2] on the set \mathcal{D} ; 3) a set of all possible treatments \mathcal{A} .

We define the set of all possible (pattern, treatment) tuples as $\mathcal{L} = \{(c, a) | c \in \mathcal{FP}, a \in \mathcal{A}\}$, and $C(\mathcal{L})$ as the set of the permutations of all possible subsets (excluding the null set) of \mathcal{L} . An element in \mathcal{L} can be thought of as a rule in a decision list and an element in $C(\mathcal{L})$ can be thought of as a list of rules in a decision list (without the default rule). We then search over all elements in the set $C(\mathcal{L}) \times \mathcal{A}$ to find a regime that maximizes the expected outcome (Eqn. 4) while minimizing the expected assessment (Eqn. 5), and treatment costs (Eqn. 6) all of which are computed over \mathcal{D} . Our objective function can be formally written as:

$$\arg \max_{s \in C(\mathcal{L}) \times \mathcal{A}} \lambda_1 g_1(\pi) - \lambda_2 g_2(\pi) - \lambda_3 g_3(\pi) \quad (7)$$

where g_1, g_2, g_3 are defined in Eqns. 4, 5, 6 respectively, and λ_1 and λ_2 are non-negative weights that scale the relative influence of the terms in the objective.

Theorem 1 The objective function in Eqn. 7 is NP-hard. (Please see appendix for details.)

Framework: Critiques

Hima



Cynthia



Framework: Critiques



Framework: Critiques

approaches in such a way that the estimated value of the expected outcome is unbiased as long as one of the postulated models is identical to the true model and there are no unmeasured confounders. The doubly robust estimator for

Framework: Critiques

On Multi-Cause Causal Inference with Unobserved Confounding:
Counterexamples, Impossibility, and Alternatives¹

Alexander D'Amour
Google AI

Framework: Critiques

On Multi-Cause Causal Inference with Unobserved Confounding:
Counterexamples, Impossibility, and Alternatives¹

Alexander D'Amour
Google AI

Algorithmic Decision Making
in the Presence of Unmeasured Confounding

Jongbin Jung
Stanford University

Ravi Shroff
New York University

Framework: Critiques

On Multi-Cause Causal Inference with Unobserved Confounding:
Counterexamples, Impossibility, and Alternatives¹

Alexander D'Amour
Google AI

Algorithmic Decision Making
in the Presence of Unmeasured Confounding

Jongbin Jung
Stanford University

Ravi Shroff
New York University

Multiple Causal Inference with Latent Confounding

Rajesh Ranganath¹ Adler Perotte²

Framework: Critiques

On Multi-Cause Causal Inference with Unobserved Confounding:
Counterexamples, Impossibility, and Alternatives¹

Alexander D'Amour
Google AI

Algorithmic Decision Making
in the Presence of Unmeasured Confounding

Jongbin Jung
Stanford University

Ravi Shroff
New York University

The Blessings of Multiple Causes

Yixin Wang
Department of Statistics
Columbia University
yixin.wang@columbia.edu

David M. Blei
Department of Statistics
Department of Computer Science
Columbia University
david.blei@columbia.edu

Multiple Causal Inference with Latent Confounding

Rajesh Ranganath¹ Adler Perotte²

Framework: Critiques

The Blessings of Multiple Causes

Yixin Wang
Department of Statistics
Columbia University
yixin.wang@columbia.edu

On Multi-Cause Causal Inference
Counterexamples, Impossibility

Alexander I
Google

Algorithmic Inference
in the Presence of Unobserved Causes

Jongbin Jung
Stanford University

The Deconfounded Recommender: A Causal Inference Approach to Recommendation

Yixin Wang
Columbia University

Dawen Liang
Netflix Inc.

Laurent Charlin*
Mila, HEC Montréal

David M. Blei
Columbia University

Science
a.edu

Multiple Causal Inference with Latent Confounding

Ravi Shroff
New York University

Rajesh Ranganath¹ Adler Perotte²

Experimental Evaluation: Datasets

- Bail decisions: 86K Defendants
 - Characteristics
 - Decisions made
 - Outcome
- Asthma: 60K Patients
 - Demographics, symptoms, health history
 - Test results
 - Medications - quick relief or long-term controller drugs
 - Outcome
- Characteristics and treatments had costs

Experimental Evaluation: Baselines

- Outcome Weighted Learning (OWL)
 - Weighted classification problem
 - Uses all characteristics
- Modified Covariate Approach (MCA)
 - Modified covariates to capture interactions between characteristics and treatments
 - Minimizes number of characteristics required
 - Do not explicitly reduce costs
- Interpretable and Parsimonious Treatment Regime Learning (IPTL)
 - Produces interpretable decision lists to maximize outcome
 - Minimizes number of characteristics required
 - Do not explicitly reduce costs

Experimental Evaluation: Setting

- Cross-validation to choose hyperparameters
 - Maximized average outcome and satisfied cost constraints
 - Is this a good way to choose hyperparameters?
- Ran for 50K iterations

Experimental Evaluation: Metrics

- Metrics:
 - **Average outcome**
 - **Average assessment costs**, based on characteristics chosen
 - **Average number of characteristics**
 - **Average treatment costs**, based on treatments chosen
 - **List length** of decision lists
- Higher outcome is better
- Lower is better for the rest

Experimental Evaluation: Results

- **CITR** maximized average outcome

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
IPTL	77.6	14.53	35.23	8.57	9	71.88	18.58	11.83	7.87	8
MCA	73.4	19.03	35.48	12.03	-	70.32	19.53	12.01	10.23	-
OWL (Gaussian)	72.9	28	35.18	13	-	71.02	25	12.38	16	-
OWL (Linear)	71.3	28	34.23	13	-	71.02	25	12.38	16	-
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Results

- **CITR** minimized costs
 - IPTL and MCA do not explicitly reduce costs

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
IPTL	77.6	14.53	35.23	8.57	9	71.88	18.58	11.83	7.87	8
MCA	73.4	19.03	35.48	12.03	-	70.32	19.53	12.01	10.23	-
OWL (Gaussian)	72.9	28	35.18	13	-	71.02	25	12.38	16	-
OWL (Linear)	71.3	28	34.23	13	-	71.02	25	12.38	16	-
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Results

- **CITR** minimized average number of characteristics
 - IPTL and MCA do aim to minimize this
 - OWL always uses all characteristics

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
IPTL	77.6	14.53	35.23	8.57	9	71.88	18.58	11.83	7.87	8
MCA	73.4	19.03	35.48	12.03	-	70.32	19.53	12.01	10.23	-
OWL (Gaussian)	72.9	28	35.18	13	-	71.02	25	12.38	16	-
OWL (Linear)	71.3	28	34.23	13	-	71.02	25	12.38	16	-
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Results

- **CITR** minimized list length
 - Only relevant to IPTL

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
IPTL	77.6	14.53	35.23	8.57	9	71.88	18.58	11.83	7.87	8
MCA	73.4	19.03	35.48	12.03	-	70.32	19.53	12.01	10.23	-
OWL (Gaussian)	72.9	28	35.18	13	-	71.02	25	12.38	16	-
OWL (Linear)	71.3	28	34.23	13	-	71.02	25	12.38	16	-
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Ablation Study

- Incrementally remove terms from objective function
- **CITR - No Treat**
 - No treatment cost
- **CITR - No Assess**
 - No assessment cost
- **CITR - Outcome**
 - Excluding both costs

Experimental Evaluation: Ablation Study

- Outcome improves when ignoring costs

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
CITR - No Treat	80.5	8.93	34.48	7.57	7	77.39	14.02	12.87	7.38	7
CITR - No Assess	81.3	13.83	32.02	9.86	10	78.32	18.28	12.02	8.97	9
CITR - Outcome	81.7	13.98	34.49	10.38	10	79.37	18.28	12.88	9.21	9
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Ablation Study

- Costs increase when ignoring costs

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
CITR - No Treat	80.5	8.93	34.48	7.57	7	77.39	14.02	12.87	7.38	7
CITR - No Assess	81.3	13.83	32.02	9.86	10	78.32	18.28	12.02	8.97	9
CITR - Outcome	81.7	13.98	34.49	10.38	10	79.37	18.28	12.88	9.21	9
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Ablation Study

- Number of characteristics and list length increases

Bail Dataset						Asthma Dataset				
	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len	Avg. Outcome	Avg. Assess Cost	Avg. Treat Cost	Avg. # of Characs.	List Len
CITR	79.2	8.88	31.09	6.38	7	74.38	13.87	11.81	7.23	6
CITR - No Treat	80.5	8.93	34.48	7.57	7	77.39	14.02	12.87	7.38	7
CITR - No Assess	81.3	13.83	32.02	9.86	10	78.32	18.28	12.02	8.97	9
CITR - Outcome	81.7	13.98	34.49	10.38	10	79.37	18.28	12.88	9.21	9
Human	69.37	-	33.39	-	-	68.32	-	12.28	-	-

Experimental Evaluation: Ablation Study

- Should we include ablation studies even in simple cases like these?

Experimental Evaluation: Qualitative Analysis

- Expensive methacholine rarely used
- Spirometry test is used - worth the costs
- Factors in asthma history

If Spiro-Test=Pos and Prev-Asthma=Yes and Cough=High then C

Else if Spiro-Test=Pos and Prev-Asthma =No then Q

Else if Short-Breath =Yes and Gender=F and Age \geq 40 and Prev-Asthma=Yes then C

Else if Peak-Flow=Yes and Prev-RespIssue=No and Wheezing =Yes, then Q

Else if Chest-Pain=Yes and Prev-RespIssue =Yes and Methacholine =Pos then C

Else Q

Experimental Evaluation: Qualitative Analysis

- Completely avoids expensive mental illness and drug tests
- Uses demographics before personal information

If Gender=F and Current-Charge =Minor and Prev-Offense=None then RP

Else if Prev-Offense=Yes and Prior-Arrest =Yes then RC

Else if Current-Charge =Misdemeanor and Age \leq 30 then RC

Else if Age \geq 50 and Prior-Arrest=No, then RP

Else if Marital-Status=Single and Pays-Rent =No and Current-Charge =Misd. then RC

Else if Addresses-Past-Yr \geq 5 then RC

Else RP

Do the experiments back up the claims?

- Pros:
 - Demonstrates that outcome and costs are optimized for two different datasets
 - Compares to several baselines
 - Includes ablation study
- Cons:
 - Not necessarily best for real use, domain expert input would be good

Thank you

