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 $x_S$
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

1. 

2. 
Review: D-Separation
Structural Equation Model (NPSEM)

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

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

● Where $\text{pa}(Y)$ is the parent set of $Y$
● These are structural models and convey causality
**Structural (Law of Nature)**

\[
\text{Grade} = \alpha + \beta (\text{Hours studied}) + \epsilon
\]

Yes  Yes

\[
\text{Hours studied} = \alpha' + \beta' (\text{Grade}) + \epsilon'
\]

Yes  No
The Back-Door Adjustment Formula

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

What criteria does \( X_C \) need to satisfy?
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

- \( X_S \) = nitrioxides 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!
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:

$$X_M = h(X_S, X_C, \epsilon_M)$$

$$Y = f(X_S, X_M, X_C, \epsilon)$$
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 $X_M$

$$\text{CDE}(x_M) \equiv f_S(x_S) - f_S(x'_S)$$
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).

Agreement across curves suggests additive causal effect.

ICE curves

PDP is the average of the ICE curves
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
Conflicting interpretations of causal relationship
Automobile MPG Data

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

Then:

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

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 judication 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>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 ≠ 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. $X_i = (\ldots, a_i, a_{i+1}, \ldots)$

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

3. Assessment cost $d$ and treatment cost $d'$
The rules in $\pi$ 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 $c_1, c_2, \cdots c_{j-1}$:

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

(1)
Framework: Expected outcome

\[ p(E|C) \neq p(E|do(C)) \]
Suppose $P(A=1|X)=0.9$

Original Population

Apply weight $\frac{1}{0.9} = \frac{10}{9}$ to each

Treated

Apply weight $\frac{1}{0.1} = 10$

Control
Framework: Expected outcome

- Suppose $P(A=1|X)=0.9$

Original Population

- Apply weight $\frac{1}{0.9} = \frac{10}{9}$ to each

Pseudo-population

Treated

Control

Treated

Control

Apply weight $\frac{1}{0.1} = 10$
Framework: Expected outcome

\[
g_1(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in A} o(i, a), \text{ where}
\]

\[
o(i, a) = \left[ \frac{1(\hat{\omega}(\hat{\pi} = a)) (y_i - \hat{y}(x_i, a)) + \hat{y}(x_i, a)}{\hat{\omega}(x_i, a)} \right] 1(\pi(x_i) = a).
\]
Framework: Expected outcome

Expected assessment cost and expected treatment cost:

\[ g_2(\pi) = \frac{1}{N} \sum_{i=1}^{N} \psi(x_i). \]
\[ g_3(\pi) = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i). \]

\[ \arg\max_{\pi \in C(\mathcal{L}) \times A} \lambda_1 g_1(\pi) - \lambda_2 g_2(\pi) - \lambda_3 g_3(\pi) \]
Framework: Optimizing the objective
Learning Cost-Effective and Interpretable Treatment Regimes

Expected Outcome: Recall that the treatment regime $\nu$ assigns a subject $s$ with characteristics $x$, to a treatment $\tau$ if $\nu(s, x) = \tau$. The expected outcome $\nu(s, x)$ is defined as the expected outcome when all the subjects in $\mathcal{D}$ are assigned the $\nu(s, x)$ treatment according to $\nu$. This is the value of the expected outcome, the better the quality of $\nu$. There is, however, an aspect to comparing the value of this expected outcome — we only observe the outcome $y_{s\nu}$ resulting from assigning $s$ to $\nu(s, x)$ in the data $\mathcal{D}$, and not $\nu$. If the regime $\nu$ assigns a different treatment $\nu'(s, x)$ to $s$, we cannot readily determine the corresponding outcome from the data.

The solutions proposed to compare 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 such approaches is provided by Lunceford et al. (22). The success of regression-based modeling and inverse weighting depends heavily on the specified regression model and the specified propensity score model respectively. In either case, if the propensity scores are not identical in the two studies, we have found that the results are sensitive to the model assumptions. 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 proposed models is correct. Thus, under suitable circumstances, the doubly robust estimator can be very robust to the model assumptions.

Doubly Robust Estimator: For the expected outcomes of regime $\nu$, denoted by $g_\nu(x)$, we have

$$g_\nu(x) = \sum_{s \in \mathcal{D}} \mathbb{I}(s, x) y_{s\nu}, \quad \text{where} \quad \mathbb{I}(s, x) = \mathbb{I}(\nu(s, x) = \nu(s, x)) $$

where $\mathbb{I}(s, x) = 1$ if the probability that the subject $s$ with characteristics $x$ is assigned to treatment $\nu$ in the data $\mathcal{D}$, and $\mathbb{I}(s, x) = 0$ if it is assigned to treatment $\nu'$. A doubly robust estimator is defined as the expected outcome when a subject characterized by $x$, is assigned to a treatment $\nu$. It is defined as an estimator by fitting a linear regression model on $\mathcal{D}$ prior to optimizing the treatment regime.

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 $s$ treated using the regime $\nu$ is given in Eq. 3. The expected assessment cost across the entire population can be computed as:

$$g_{a}(x) = \sum_{s \in \mathcal{D}} a(s, x) $$

It is important to note that our learning process favors regimes with smaller expected assessment cost. Keeping this cost low also decreases the shared decision list size, which assists with interpretability.

Expected Treatment Cost: The treatment cost for a subject $s$ is assigned using a regime $\nu$, it is given in Eqs. 2. The expected treatment cost across the entire population can be computed as:

$$g_{c}(x) = \sum_{s \in \mathcal{D}} c(s, x) $$

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

Objective Function: We assume access to the following inputs: $\mathcal{D}$, the observational data $\mathcal{D}$; $\mathcal{D}_f$, the set of FD of frequently occurring outcomes in $\mathcal{D}$; recall that each patient corresponds to a combination of one or more individuals. As examples of such a patient is “Age $\geq 40$ and Gender:Female”. In practice, such patients can be obtained by running a frequent pattern mining algorithms such as Apriori (21) or the set $\mathcal{D}_f$, is a set of all possible treatment regimes $\mathcal{A}$.

We define the set of all possible outcomes, treatment regimes $\mathcal{A} = \{f(\mathcal{A}) = \emptyset, \emptyset \in \mathcal{A}\}$, and $\mathcal{A}_f$ as the set of combinations of all possible outcomes (including the null set $\emptyset$). An element $\emptyset$ in $\mathcal{A}_f$ can be thought of as a rule in a decision list and an element in $\mathcal{A}_f$ can be thought of as a list of rules in a decision list (without the default rule). We then search over all the elements in the set $\mathcal{A}_f$ and find a regime that maximizes the expected outcome $\mu_{\nu}(x)$ while minimizing the expected assessment $\mu_{a}(x)$ and treatment costs $\mu_{c}(x)$ of which are computed over $\mathcal{D}$.

Our objective function can be formally written as:

$$\arg\max_{\nu \in \mathcal{A}} \mu_{\nu}(x) - \lambda_1 \mu_{a}(x) - \lambda_2 \mu_{c}(x)$$

where $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ are defined in Eq. 4, 5, 6, respectively, and $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ are non-negative weights that scale the relative influence of the terms in the objective.

Theorem 1 The objective function in Eq. 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 Alternatives

Alexander D'Amour
Google AI
Framework: Critiques

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

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 Alternatives

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 Ranganath
Adler Perotte
Framework: Critiques

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

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 Ranganath
Adler Perotte

Rajesh Ranganath
Adler Perotte
Framework: Critiques

The Blessings of Multiple Causes
Yixin Wang
Department of Statistics
Columbia University
yixin.wang@columbia.edu

The Deconfounded Recommender:
A Causal Inference Approach to Recommendation

Algorithmic I
in the Presence of Un

Yixin Wang
Columbia University
Dawan Liang
Netflix Inc.
Laurent Charlin*
Mila, HEC Montréal
David M. Blei
Columbia University

Multiple Causal Inference with Latent Confounding

Jongbin Jung
Stanford University
Ravi Shroff
New York University

Rajesh Ranganath ¹
Adler Perotte ²
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
Experimental Evaluation: Results

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
<td>31.09</td>
<td>6.38</td>
<td>7</td>
<td>74.38</td>
<td>13.87</td>
<td>11.81</td>
<td>7.23</td>
<td>6</td>
</tr>
<tr>
<td>IPTL</td>
<td>77.6</td>
<td>14.53</td>
<td>35.23</td>
<td>8.57</td>
<td>9</td>
<td>71.88</td>
<td>18.58</td>
<td>11.83</td>
<td>7.87</td>
<td>8</td>
</tr>
<tr>
<td>MCA</td>
<td>73.4</td>
<td>19.03</td>
<td>35.48</td>
<td>12.03</td>
<td>-</td>
<td>70.32</td>
<td>19.53</td>
<td>12.01</td>
<td>10.23</td>
<td>-</td>
</tr>
<tr>
<td>OWL (Gaussian)</td>
<td>72.9</td>
<td>28</td>
<td>35.18</td>
<td>13</td>
<td>-</td>
<td>71.02</td>
<td>25</td>
<td>12.38</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>OWL (Linear)</td>
<td>71.3</td>
<td>28</td>
<td>34.23</td>
<td>13</td>
<td>-</td>
<td>71.02</td>
<td>25</td>
<td>12.38</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
<td>33.39</td>
<td>-</td>
<td>-</td>
<td>68.32</td>
<td>-</td>
<td>12.28</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Experimental Evaluation: Results

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

<table>
<thead>
<tr>
<th></th>
<th>Bail Dataset</th>
<th>Asthma Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
</tr>
<tr>
<td>IPTL</td>
<td>77.6</td>
<td>14.53</td>
</tr>
<tr>
<td>MCA</td>
<td>73.4</td>
<td>19.03</td>
</tr>
<tr>
<td>OWL (Gaussian)</td>
<td>72.9</td>
<td>28</td>
</tr>
<tr>
<td>OWL (Linear)</td>
<td>71.3</td>
<td>28</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
</tr>
</tbody>
</table>
Experimental Evaluation: Results

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
<td>31.09</td>
<td>6.38</td>
<td>7</td>
<td>74.38</td>
<td>13.87</td>
<td>11.81</td>
<td>7.23</td>
<td>6</td>
</tr>
<tr>
<td>IPTL</td>
<td>77.6</td>
<td>14.53</td>
<td>35.23</td>
<td>8.57</td>
<td>9</td>
<td>71.88</td>
<td>18.58</td>
<td>11.83</td>
<td>7.87</td>
<td>8</td>
</tr>
<tr>
<td>MCA</td>
<td>73.4</td>
<td>19.03</td>
<td>35.48</td>
<td>12.03</td>
<td>-</td>
<td>70.32</td>
<td>19.53</td>
<td>12.01</td>
<td>10.23</td>
<td>-</td>
</tr>
<tr>
<td>OWL (Gaussian)</td>
<td>72.9</td>
<td>28</td>
<td>35.18</td>
<td>13</td>
<td>-</td>
<td>71.02</td>
<td>25</td>
<td>12.38</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>OWL (Linear)</td>
<td>71.3</td>
<td>28</td>
<td>34.23</td>
<td>13</td>
<td>-</td>
<td>71.02</td>
<td>25</td>
<td>12.38</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
<td>33.39</td>
<td>-</td>
<td>-</td>
<td>68.32</td>
<td>-</td>
<td>12.28</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
<td>31.09</td>
<td>6.38</td>
<td>11.81</td>
</tr>
<tr>
<td>IPTL</td>
<td>77.6</td>
<td>14.53</td>
<td>35.23</td>
<td>8.57</td>
<td>11.83</td>
</tr>
<tr>
<td>MCA</td>
<td>73.4</td>
<td>19.03</td>
<td>35.48</td>
<td>12.03</td>
<td>10.23</td>
</tr>
<tr>
<td>OWL (Gaussian)</td>
<td>72.9</td>
<td>28</td>
<td>35.18</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>OWL (Linear)</td>
<td>71.3</td>
<td>28</td>
<td>34.23</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
<td>33.39</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Bail Dataset</th>
<th></th>
<th></th>
<th></th>
<th>Asthma Dataset</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>List Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
<td>31.09</td>
<td>6.38</td>
<td>7</td>
<td>74.38</td>
<td>13.87</td>
<td>11.81</td>
<td>7.23</td>
<td>6</td>
</tr>
<tr>
<td>CITR - No Treat</td>
<td>80.5</td>
<td>8.93</td>
<td>34.48</td>
<td>7.57</td>
<td>7</td>
<td>77.39</td>
<td>14.02</td>
<td>12.87</td>
<td>7.38</td>
<td>7</td>
</tr>
<tr>
<td>CITR - No Assess</td>
<td>81.3</td>
<td>13.83</td>
<td>32.02</td>
<td>9.86</td>
<td>10</td>
<td>78.32</td>
<td>18.28</td>
<td>12.02</td>
<td>8.97</td>
<td>9</td>
</tr>
<tr>
<td>CITR - Outcome</td>
<td>81.7</td>
<td>13.98</td>
<td>34.49</td>
<td>10.38</td>
<td>10</td>
<td>79.37</td>
<td>18.28</td>
<td>12.88</td>
<td>9.21</td>
<td>9</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
<td>33.39</td>
<td>-</td>
<td>-</td>
<td>68.32</td>
<td>-</td>
<td>12.28</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Experimental Evaluation: Ablation Study

- Costs increase when ignoring costs

<table>
<thead>
<tr>
<th></th>
<th>Bail Dataset</th>
<th>Asthma Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
</tr>
<tr>
<td>CITR - No Treat</td>
<td>80.5</td>
<td>8.93</td>
</tr>
<tr>
<td>CITR - No Assess</td>
<td>81.3</td>
<td>13.83</td>
</tr>
<tr>
<td>CITR - Outcome</td>
<td>81.7</td>
<td>13.98</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
</tr>
</tbody>
</table>
Experimental Evaluation: Ablation Study

- Number of characteristics and list length increases

<table>
<thead>
<tr>
<th></th>
<th>Bail Dataset</th>
<th>Asthma Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>CITR</td>
<td>79.2</td>
<td>8.88</td>
</tr>
<tr>
<td>CITR - No Treat</td>
<td>80.5</td>
<td>8.93</td>
</tr>
<tr>
<td>CITR - No Assess</td>
<td>81.3</td>
<td>13.83</td>
</tr>
<tr>
<td>CITR - Outcome</td>
<td>81.7</td>
<td>13.98</td>
</tr>
<tr>
<td>Human</td>
<td>69.37</td>
<td>-</td>
</tr>
</tbody>
</table>
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 and 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