Explaining Privacy and Fairness Violations in Data-Driven Systems

Matt Fredrikson
Carnegie Mellon University
Joint effort

Emily Black  Gihyuk Ko  Klas Leino

Piotr Mardziel  Shayak Sen  Sam Yeom

Anupam Datta
Data-driven systems are ubiquitous
Data-driven systems are opaque
...able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score.

Take a fictional Target shopper who ... bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug.

There’s, say, an 87 percent chance that she’s pregnant
Opacity and fairness

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin
Inappropriate information use

Both problems can be seen as *inappropriate use of protected information*

• Fairness/discrimination  
  • Use of *race* or *gender* for *employment decisions*  
  • Business necessity exceptions

• Privacy  
  • Use of *health* or *political background* for *marketing*  
  • Exceptions derive from contextual information norms

This is a type of bug!
Agenda

Methods for dealing with inappropriate information use
- Detecting when it occurs
- Providing diagnostic information to developers
- Automatic repair, when possible

Remaining talk:
- Formalize “inappropriate information use”
- Show how it applies to classifiers
- Generalize to continuous domain
- Nonlinear continuous models & applications
Explicit use via causal influence [Datta, Sen, Zick Oakland’16]

Example: Credit decisions

Conclusion: Measures of association not informative
Causal intervention

Replace feature with random values from the population, and examine distribution over outcomes.
Challenge: Indirect (proxy) use

- # years in same job
- Unpaid mortgage?
- Income
- ...

Classifier (targets older people)

Decision

Need to determine when information type is inferred and then used
Proxy use: a closer look

What do we mean by proxy use?

1. Explicit use is also proxy use
Proxy use: a closer look

What do we mean by proxy use?

1. Explicit use is also proxy use
2. “Inferred use” is proxy use
Proxy use: a closer look

What do we mean by proxy use?

1. Explicit use is also proxy use
2. “Inferred use” is proxy use
   • Inferred values must be influential
Proxy use: a closer look

What do we mean by proxy use?

1. Explicit use is also proxy use
2. “Inferred use” is proxy use
   • Inferred values *must be influential*
   • Associations must be two-sided

F Y N yrs in job > 10

F Y N unpaid mortgage?

F Y N unpaid mortgage?
One- and two-sided associations

What happens if we allow one-sided association?

Consider this model:
• Uses postal code to determine state
• Zip code can predict race
• ...but not the other way around

This is a benign use of information that’s associated with a protected information type
What do we mean by proxy use?

1. Explicit use is also proxy use
2. “Inferred use” is proxy use
   • Inferred values must be **influential**
   • Associations must be **two-sided**
3. Output association is unnecessary for proxy use
Towards a formal definition: axiomatic basis

• (Axiom 1: Explicit use) If random variable $Z$ is an influential input of the model $A$, then $A$ makes proxy use of $Z$.

• (Axiom 2: Preprocessing) If a model $A$ makes proxy use of $Z$, and $A'(x) = A(x, f(x))$, then $A'$ also makes proxy use of $Z$.
  • Example: $A'$ infers a protected piece of info given directly to $A$

• (Axiom 3: Dummy) If $A'(x, x') = A(x)$ for all $x$ and $x'$, then $A'$ has proxy use of $Z$ exactly when $A$ does.
  • Example: feature never touched by the model.

• (Axiom 4: Independence) If $Z$ is independent of the inputs of $A$, then $A$ does not have proxy use of $Z$.
  • Example: model obtains no information about protected type
Extensional proxy use axioms are inconsistent

Key Intuition:

• **Preprocessing** forces us to preserve proxy use under function composition
  • But the rest of the model can **cancel out** a composed proxy

• Let $X,Y,Z$ be pairwise independent random variables, and $Y = X \oplus Z$

• Then $A(Y,Z) = Y \oplus Z$ makes proxy use of $Z$ (explicit use axiom)

• So does $A'(Y,Z,X) = Y \oplus Z$ (dummy axiom)

• And so does $A''(Z,X) = A'(X \oplus Z, Z, X)$ (preprocessing axiom)

• But $A''(Z,X) = X \oplus Z \oplus Z = X$, and $X, Z$ are independent...
Syntactic relaxation

• We address this with a **syntactic definition**

• Composition is tied to how the function is represented as a **program**

• **Checking** for proxy use requires access to program internals
Models as Programs

- Expressions that produce a value
- No loops or other complexities
- But often very large

\[ \langle exp \rangle ::= R | \text{True} | \text{False} | \text{var} \\
| \text{op}(\langle exp \rangle, \ldots, \langle exp \rangle) \\
| \text{if} \{ \langle exp \rangle \} \text{then} \{ \langle exp \rangle \} \text{else} \{ \langle exp \rangle \} \]

Operations:
- Arithmetic operations: +, -, *, etc.
- Boolean connectives: or, and, not, etc.
- Relations: ==, <, ≤, >, etc.
Expression semantics:

\[ \llbracket \text{exp} \rrbracket : \text{Instance} \rightarrow \text{Value} \]

\( I \) is a random variable over dataset instances

\[ \llbracket \text{exp} \rrbracket : I \rightarrow V \]

\( V \) is a random variable over the expression’s value

Joint over input instance (\( I \)) and expression values (\( V_i \)) for each expression \( \text{exp}_i \).

- Marginals: \( \Pr[I, V_0, V_1, \ldots, V_9] \)
- Conditionals: \( \Pr[V_4 = \text{True} \mid V_0 = \text{Ad}_1] \)
Decomposition
Given a program $p$, a decomposition $(p_1, X, p_2)$ consists of two programs $p_1, p_2$, and a fresh variable $X$ such that replacing $X$ with $p_1$ inside $p_2$ yields $p$. 
Characterizing proxies

Proxy

Given a decomposition \((p_1, X, p_2)\) and a random variable \(Z\), \(p_1\) is a proxy for \(Z\) if \(\llbracket[p_1]\rrbracket(I)\) is associated with \(Z\).

\(p_1\) is a proxy for “gender = Female”
Influential Decomposition

A decomposition \((p_1, X, p_2)\) is influential if \(X\) can change the outcome of \(p_2\).
Putting it all together

**Proxy Use**

A program $p$ has **proxy use** of random variable $Z$ if there exists an influential decomposition $(p_1, X, p_2)$ of $p$ that is a proxy for $Z$.

This is close to our intuition from earlier

Formally, it satisfies similar axioms:
- Dummy and independence axioms remain largely unchanged
- Explicit use, preprocessing rely on program decomposition instead of function composition
Quantitative proxy use

A decomposition \((p_1, X, p_2)\) is an \((\varepsilon, \delta)\)-proxy use of \(Z\) when

- The association between \(p_1\) and \(Z\) is \(\geq \varepsilon\), and
- \(p_1\)'s influence in \(p_2\), \(\iota(p_1, p_2)\), is \(\geq \delta\)

A program has \((\varepsilon, \delta)\)-proxy use of \(Z\) when it admits a decomposition that is an \((\varepsilon, \delta)\)-proxy use of \(Z\)
Quantifying decomposition influence

1. Intervene on $p_1$

2. Compare the behavior:
   • With intervention
   • As the system runs normally

3. Measure divergence:

   $$\iota(p_1, p_2) = E_{X,X'}[ p(X) \neq p_2(X, p_1(X')) ]$$
Algorithmics

• Does system have an \((\varepsilon, \delta)\)-proxy-use of a protected variable?

• Basic algorithm \(O(S*N^2)\)
  • \(S\) – # expressions
  • \(N\) – # dataset instances

• How do we remove \((\varepsilon, \delta)\)-proxy-use violation?

• Naive algorithm
  • Replace \(\text{Exp}_i\) with a constant
    \(O(1)\) // any constant
    \(O(N * M)\) // best constant, \(M\) – # possible values
Using Witnesses

Demonstration of violation in the system
Localize where scrutiny/human eyeballs need to be applied
Determine what repair should be applied
Experiments: Benchmark datasets (CCS’17)

~ Marital status

- capital-loss ≤ 1882.5
  - gender = female
    - Age ≤ 30
      - ...
      - ...
      - ...

~ Wife’s religion

- wife-educ ≤ 3
  - # children ≤ 3
    - age ≤ 31
      - ...
      - ...
      - ...

model accuracy: 83.6 %
after repair: 81.7%

model accuracy: 61.2 %
after repair: 52.1 %
Agenda

Methods for dealing with inappropriate information use
• Detecting when it occurs
• Providing diagnostic information to developers
• Automatic repair, when possible

Remaining talk:
• Formalize “inappropriate information use”
• Show how it applies to classifiers
• Generalize to continuous domain
• Nonlinear continuous models & applications
Proxies in linear regressors [NIPS’18]

\[ Y(X) = a_1X_1 + a_2X_2 + ... + a_nX_n \]

Recall our definition of decomposition influence:

\[ \iota(p_1, p_2) = \mathbb{E}_{X,X'}[ p(X) \neq p(X', p_1(X')) ] \]

We generalize to regression by defining:

\[ \iota(p_1, p_2) = \mathbb{E}_{X,X'}[ (p(X) - p(X', p_1(X')))^2 ] \]
Proxies in regressors [NIPS’18]

\[ Y(\mathbf{X}) = a_1X_1 + a_2X_2 + \ldots + a_nX_n \]

What are the decompositions?
- Just individual terms \( a_nX_n \)? Or groups like \( a_1X_1 + a_2X_2 \)?
- What about \( 0.5*a_1X_1 + a_2X_2 \)?

Component \( P(\mathbf{X}) = \beta_1a_1X_1 + \beta_2a_2X_2 + \ldots + \beta_na_nX_n \)

for \( \beta_1, \ldots, \beta_n \in [0, 1] \)
Proxies in regressors [NIPS’18]

\[ l(p_1, p_2) = E_{X,X'}[(Y(X) - Y(X, P(X')))^2] \propto \text{Var}(P(X)) \]

\[ \text{Asc}(Y, Z) \propto \text{Cov}(Y, Z) \]

Optimize to find proxies!

Find \( \max_{\beta} \| A\beta \| \leq c^T \beta \) (for \( A\beta \| \leq c^T \beta \))

such that \( |\text{Asc}(A\beta, Z)| \geq \varepsilon \)

where \( A^TA = \text{Cov}(X, X) \)
Agenda

Methods for dealing with inappropriate information use
  • Detecting when it occurs
  • Providing diagnostic information to developers
  • Automatic repair, when possible

Remaining talk:
  • Formalize “inappropriate information use”
  • Show how it applies to classifiers
  • Generalize to continuous domain
  • Nonlinear continuous models & applications
Distributional Influence: proxies in neural nets

Feature extractor $z = h(x)$  
Classifier $g(z)$

Network $f(x) = h(g(x))$

Distribution over inputs

$$i_j(f, P) = \int x \left[ \frac{\partial g}{\partial z_j} \right]_{h(x)} P(x) \, dx$$

Axioms:
- Linear agreement
- Distributional marginality
- Distribution linearity
Problems with neural nets: stereotyping

See [Stock & Sisse, 2018] for more examples like this
Problems with neural nets: bias amplification

In training data, 33% of “cooking” images have men in them.
In predictions, 16% of “agent” roles in cooking images are labeled “man”
Explaining stereotype predictions

basketball (73%)

top 5% most influential features

top 25% most influential features
Intrinsic bias amplification

statistical distance between classes

Prior class probability
Prediction bias from inductive bias

\[ = 0.5 + \text{prediction bias} \]

Difference between learned (\(h_s\)) and optimal (\(h^*\)) weight (averaged)

Larger weights \(\approx\) More influence
Simple fix: kill weak features

\[ \alpha^*, \beta^* = \arg \min_{\alpha, \beta} \left| B_D(g_\beta^\alpha) \right| \text{ subject to } L_S(g_\beta^\alpha) \leq L_S(g) \]

- # most positive-influential features to keep
- Bias of resulting classifier
- # most negative-influential features to keep
- Don’t increase the empirical loss
Early results

<table>
<thead>
<tr>
<th>dataset</th>
<th>$p^*$</th>
<th>asymm. (%)</th>
<th>$B_D(h_S)$</th>
<th>$B_D(h_S)$ (post-fix)</th>
<th>acc. (%)</th>
<th>acc. (%) (post-fix)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$B_D(h_S)$</td>
<td>par exp $\ell_1$</td>
<td></td>
<td>par exp $\ell_1$</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>50.0</td>
<td>52.0</td>
<td>1.8</td>
<td>1.7 0.4 n/a</td>
<td>93.0</td>
<td>93.1 94.0 n/a</td>
</tr>
<tr>
<td>CelebA</td>
<td>50.4</td>
<td>50.2</td>
<td>7.7</td>
<td>7.7 0.2 n/a</td>
<td>79.6</td>
<td>79.6 79.9 n/a</td>
</tr>
<tr>
<td>arcene</td>
<td>56.0</td>
<td>57.7</td>
<td>2.7</td>
<td>0.6 1.2 1.7</td>
<td>68.9</td>
<td>69.0 74.2 69.4</td>
</tr>
<tr>
<td>colon</td>
<td>64.5</td>
<td>51.0</td>
<td>23.1</td>
<td>22.9 22.6 35.5</td>
<td>58.5</td>
<td>58.7 58.7 64.5</td>
</tr>
<tr>
<td>glioma</td>
<td>69.4</td>
<td>54.8</td>
<td>17.4</td>
<td>17.4 12.2 17.0</td>
<td>76.3</td>
<td>76.3 76.7 75.44</td>
</tr>
<tr>
<td>micromass</td>
<td>69.0</td>
<td>54.1</td>
<td>0.68</td>
<td>0.66 0.69 0.68</td>
<td>98.4</td>
<td>98.4 98.4 98.4</td>
</tr>
<tr>
<td>pc/mac</td>
<td>50.5</td>
<td>60.6</td>
<td>1.6</td>
<td>1.6 1.4 1.6</td>
<td>89.0</td>
<td>89.0 88.0 89.0</td>
</tr>
<tr>
<td>prostate</td>
<td>51.0</td>
<td>44.4</td>
<td>47.3</td>
<td>47.2 10.0 28.1</td>
<td>52.7</td>
<td>52.8 90.2 71.3</td>
</tr>
<tr>
<td>smokers</td>
<td>51.9</td>
<td>50.4</td>
<td>47.4</td>
<td>45.4 8.0 33.0</td>
<td>50.0</td>
<td>50.7 59.0 51.2</td>
</tr>
<tr>
<td>synthetic</td>
<td>50.0</td>
<td>99.9</td>
<td>24.1</td>
<td>17.2 23.6 5.7</td>
<td>74.9</td>
<td>77.9 74.8 71.4</td>
</tr>
</tbody>
</table>
Summary

Methods for dealing with inappropriate information use

• Detecting when it occurs
• Providing diagnostic information to developers
• Automatic repair, when possible

Progress:

• Formalize “inappropriate information use” as proxy use
• Generalized to continuous domain and neural networks
• Algorithms for detection and diagnosis
• Explanation-based repair methods