Explaining Privacy and Fairness Violations in Data-Driven Systems

Matt Fredrikson

Carnegie Mellon University

Joint effort

Emily Black

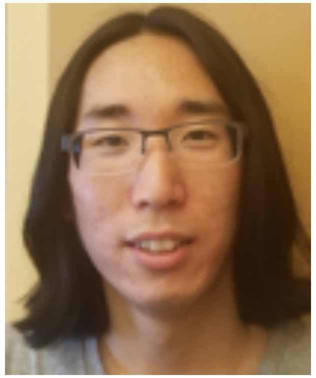
Gihyuk Ko

Klas Leino

Anupam Datta

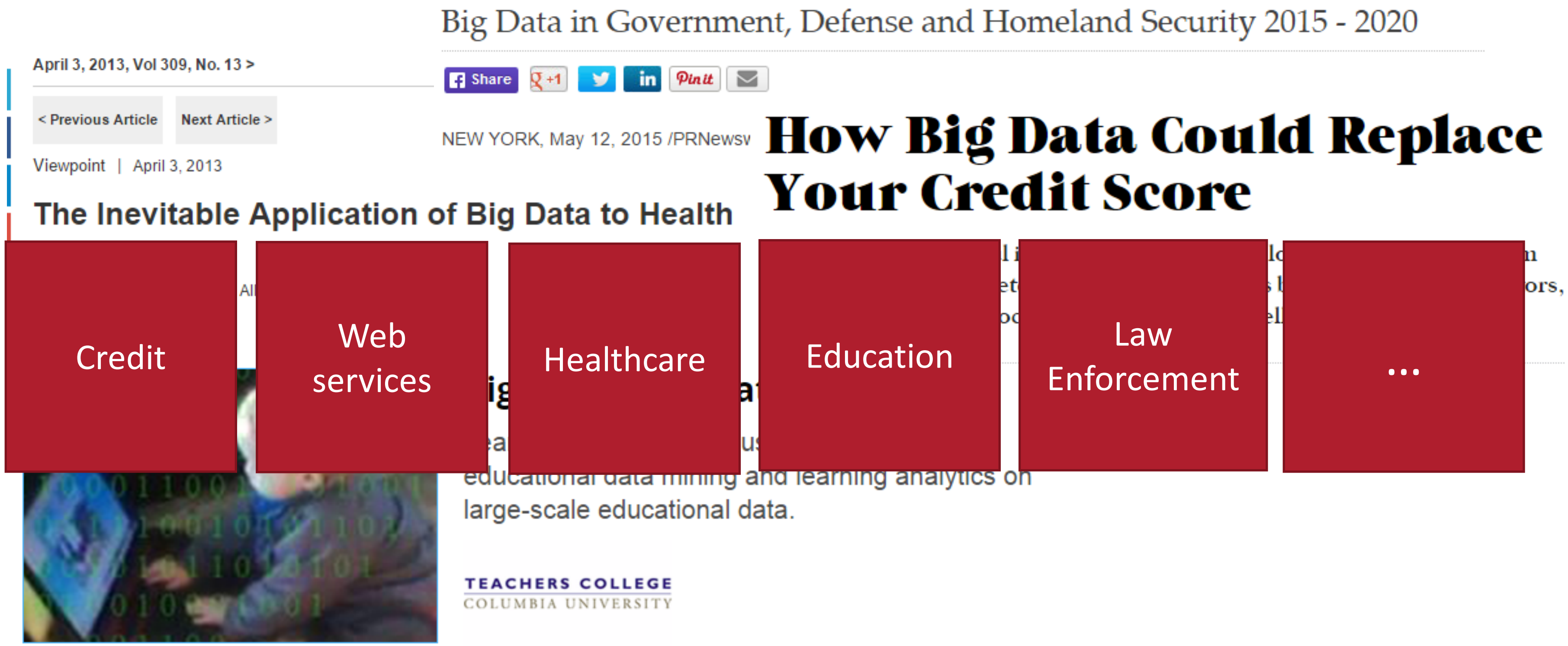
Piotr Mardziel

Shayak Sen

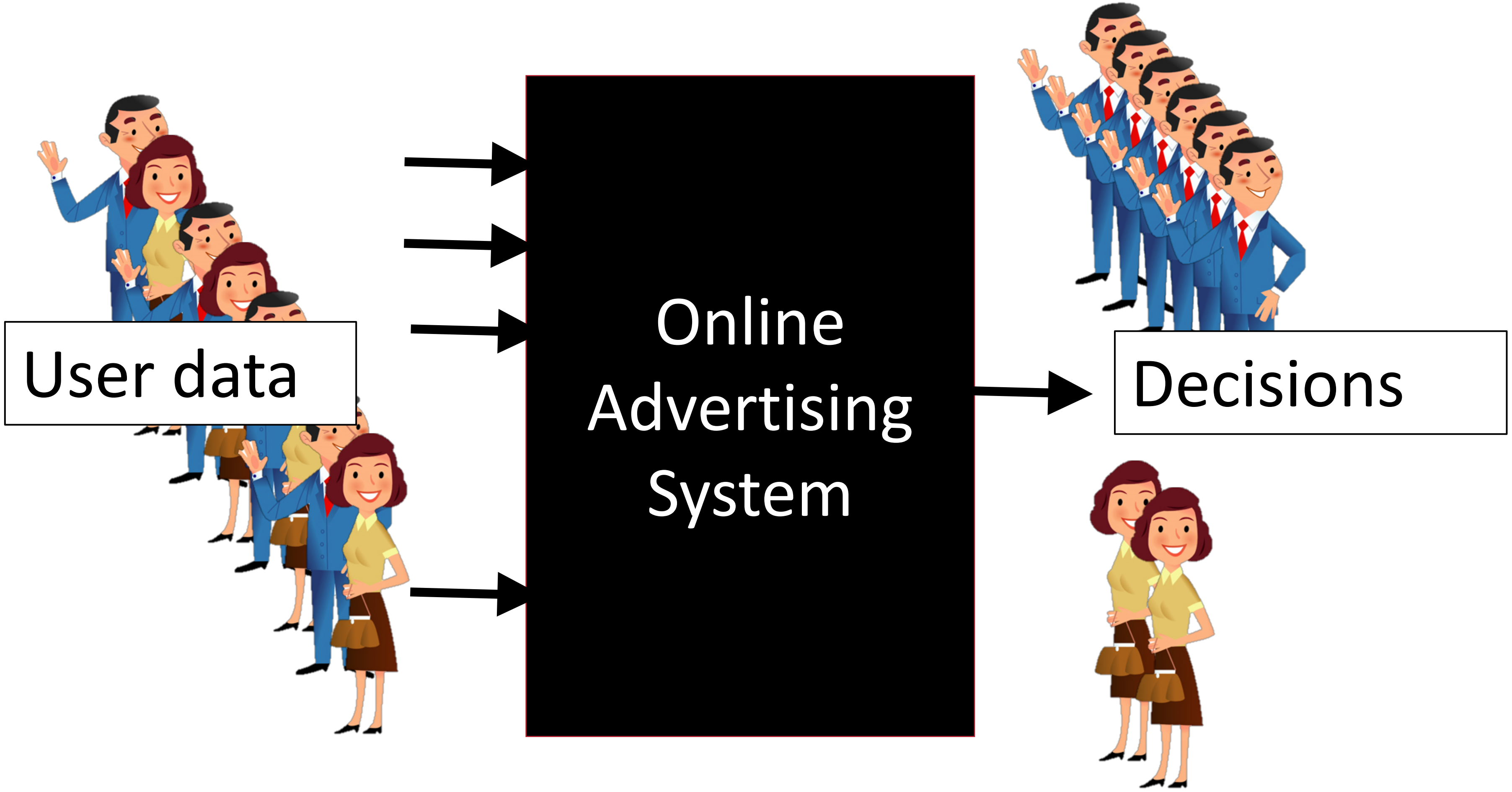


Sam Yeom

Data-driven systems are ubiquitous



Data-driven systems are opaque



Opacity and privacy

…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

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 4 DAYS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

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

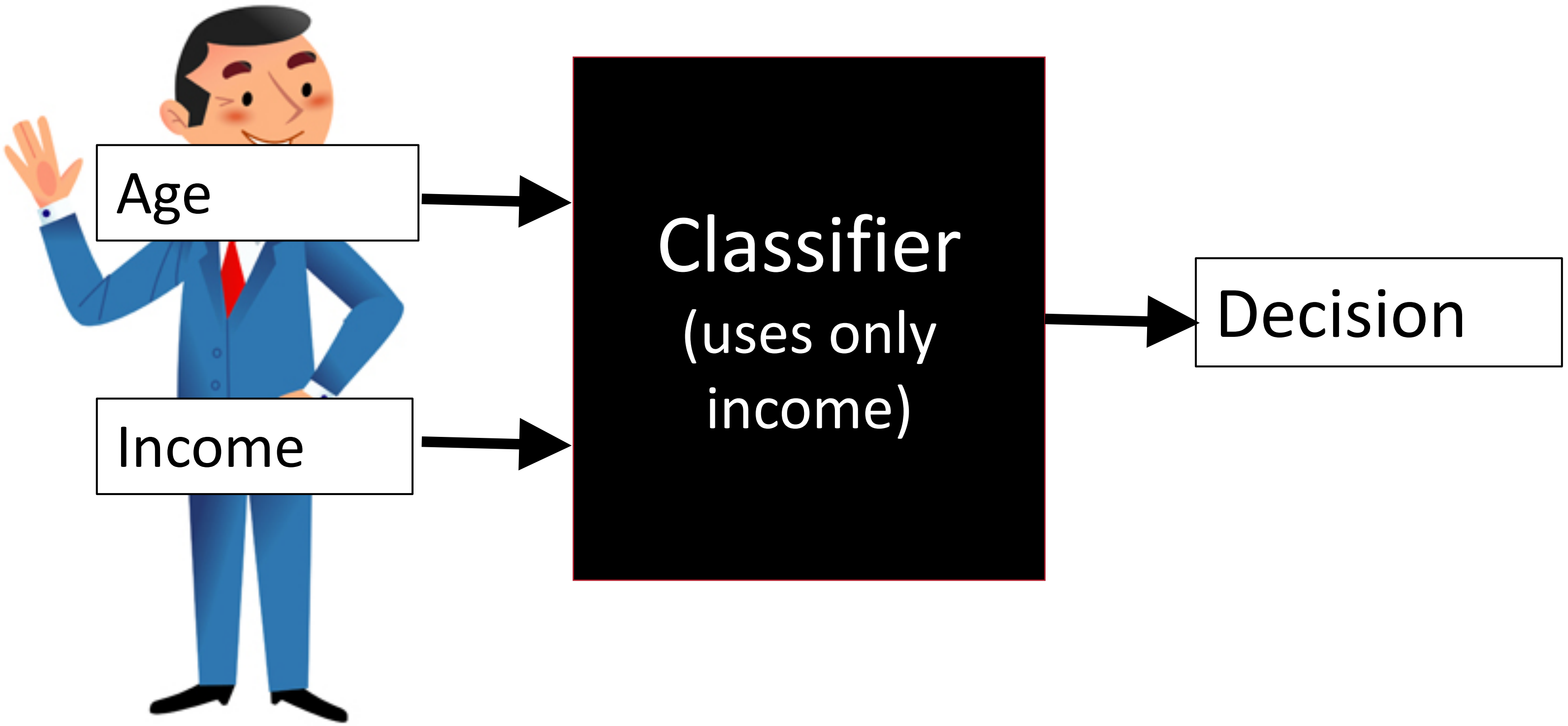
- Detecting when it occurs
- Providing diagnostic information to developers
- Automatic repair, when possible

Remaining talk:

- Formalize “inappropriate information use”
- Show how it applies to classifiers
- Generalize to continuous domain
- Nonlinear continuous models & applications

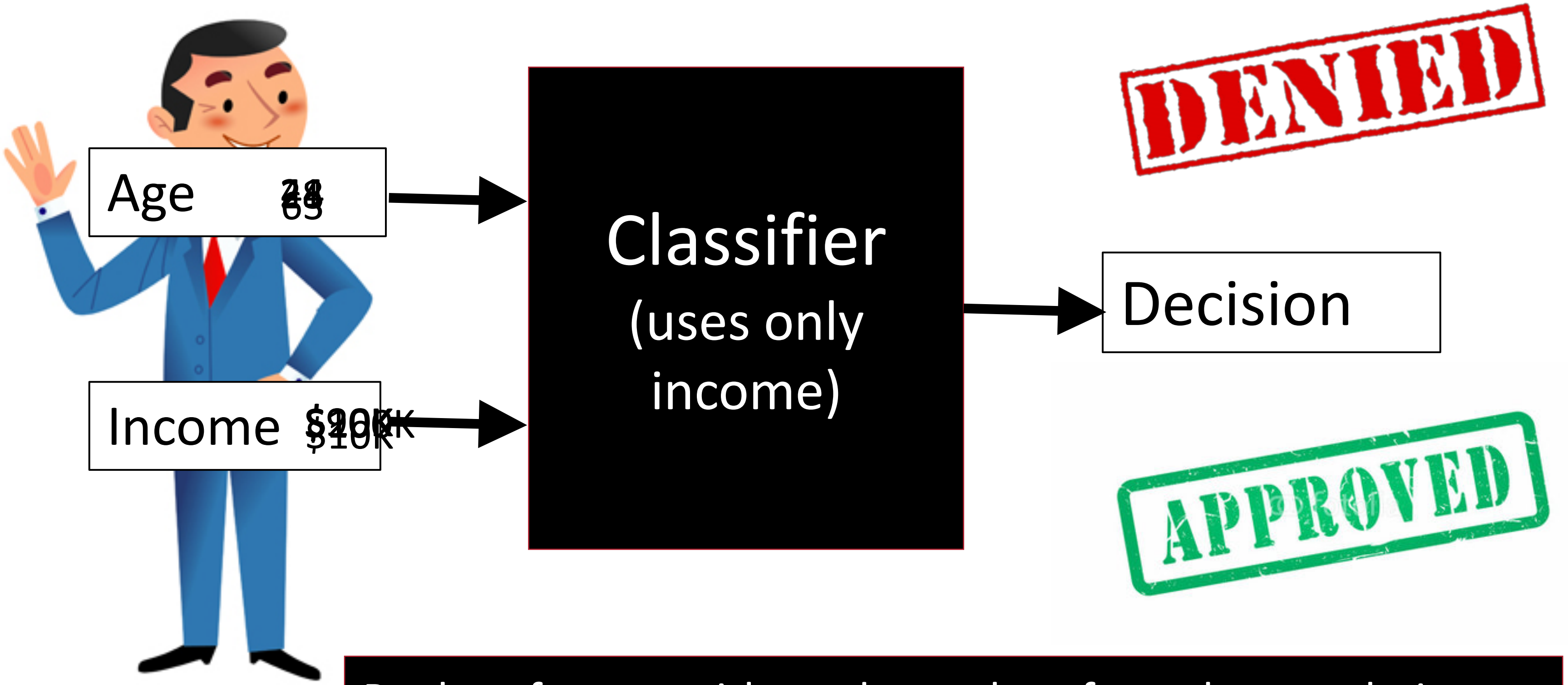
Explicit use via causal influence [Datta, Sen, Zick Oakland'16]

Example: Credit decisions



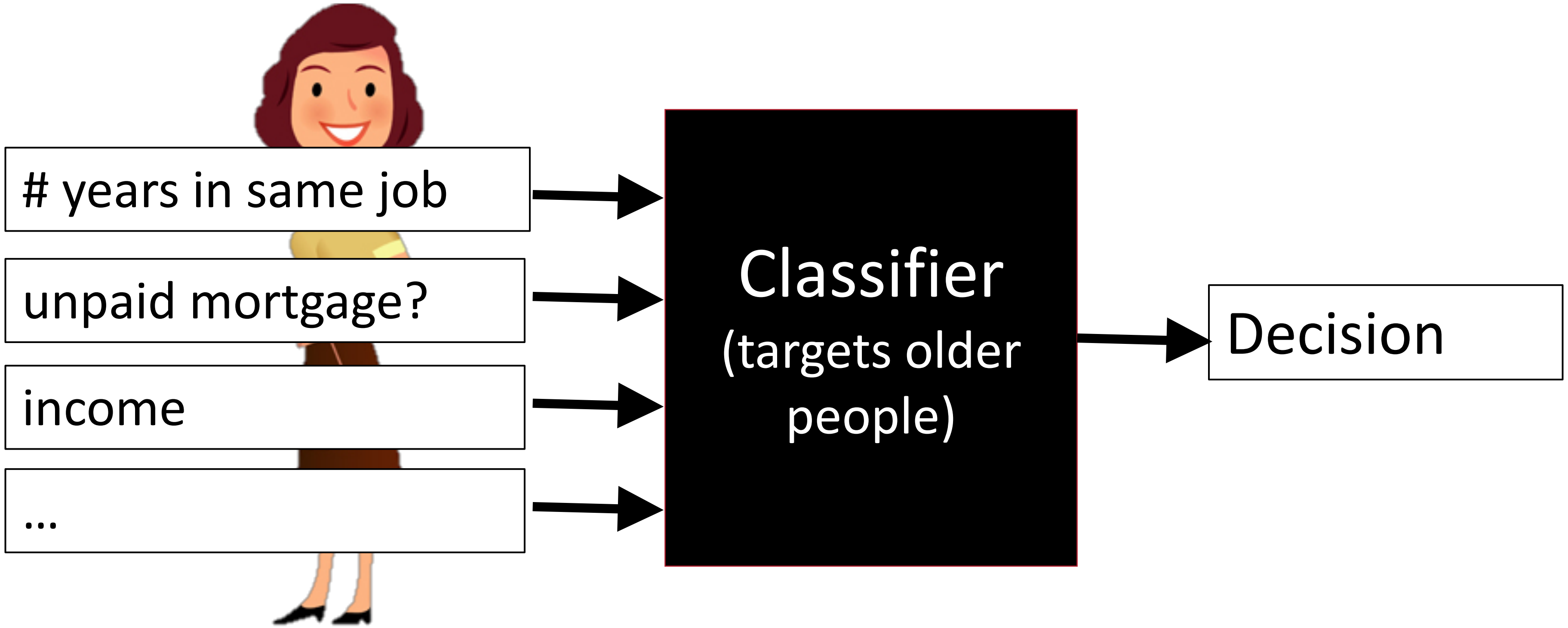
Conclusion: Measures of association not informative

Causal intervention



Replace feature with random values from the population, and examine distribution over outcomes.

Challenge: Indirect (proxy) use

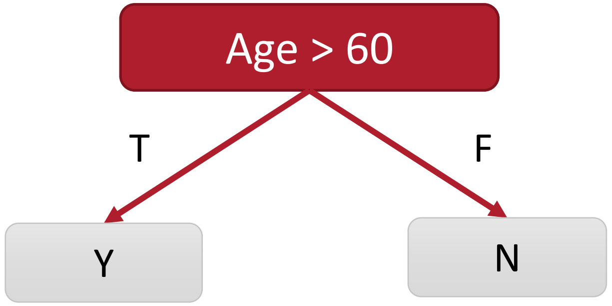


Need to determine when information type is **inferred** and then used

Proxy use: a closer look

What do we mean by proxy use?

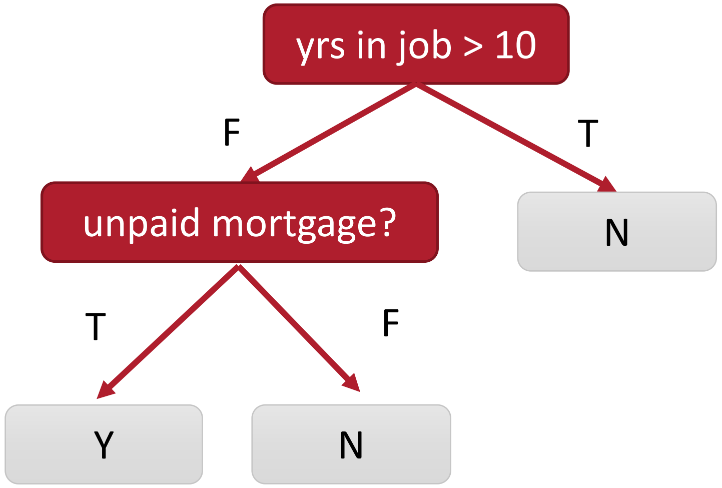
1. Explicit use is also proxy use



Proxy use: a closer look

What do we mean by proxy use?

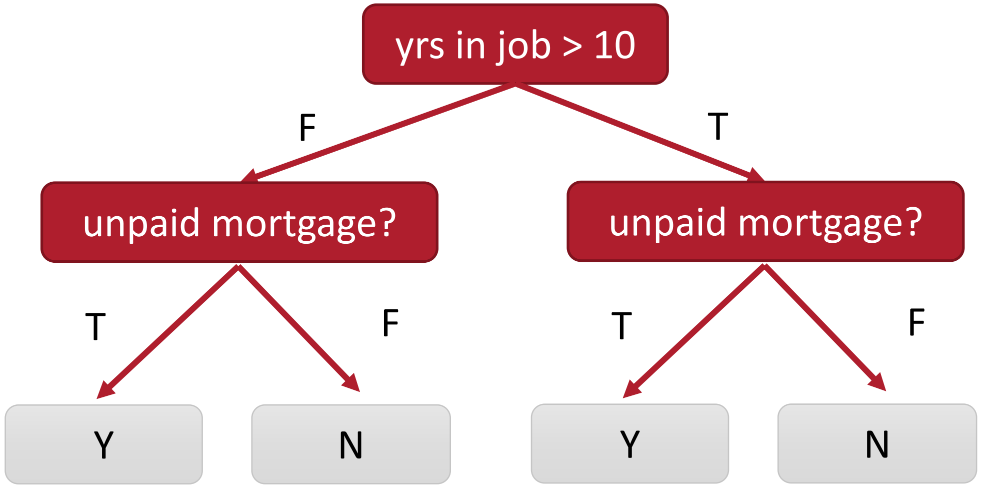
1. Explicit use is also proxy use
2. “Inferred use” is proxy use



Proxy use: a closer look

What do we mean by proxy use?

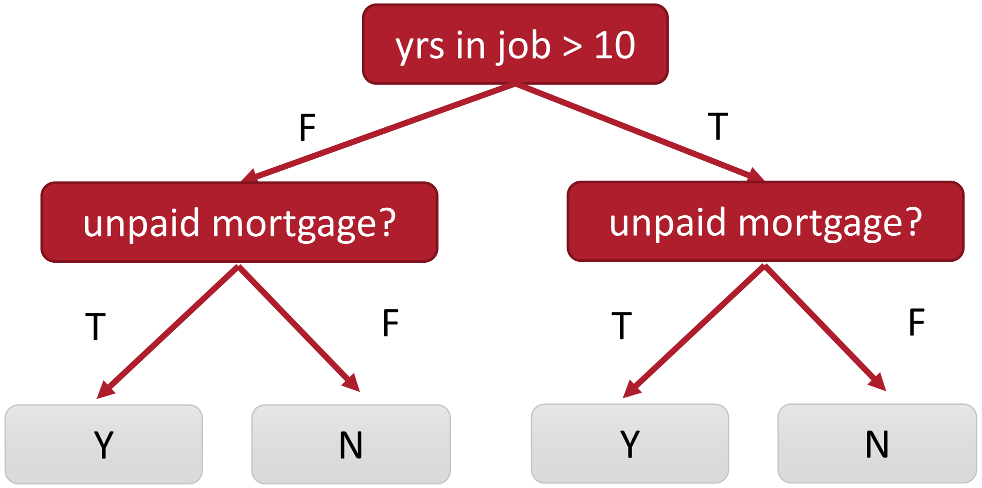
1. Explicit use is also proxy use
2. “Inferred use” is proxy use
    - Inferred values *must be influential*



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



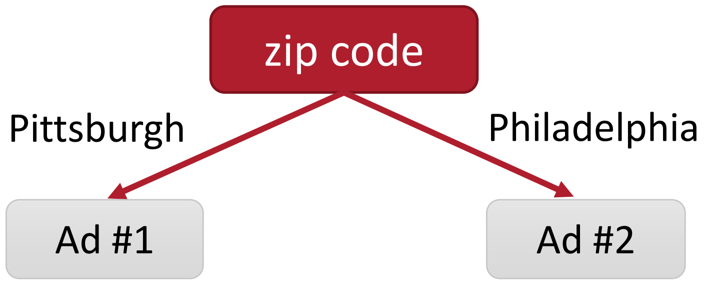
One- and two-sided associations

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

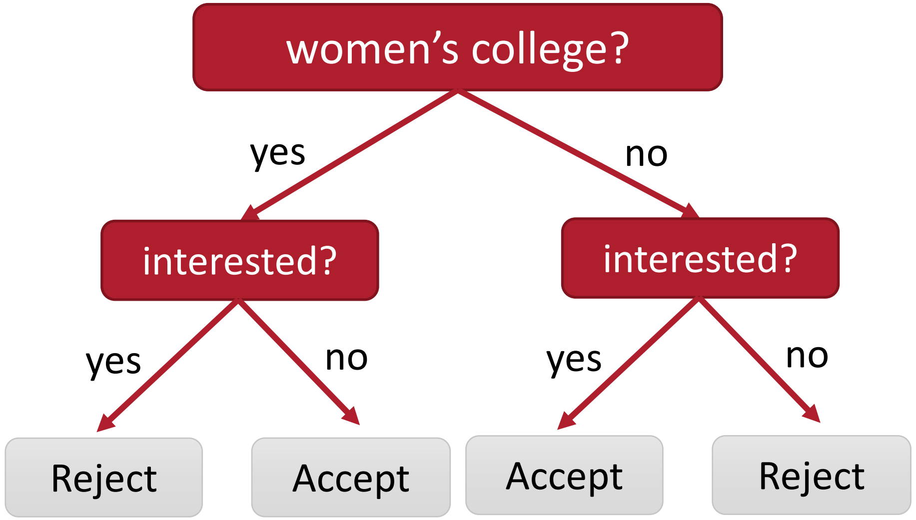
This is a benign use of information that's associated with a protected information type



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**
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

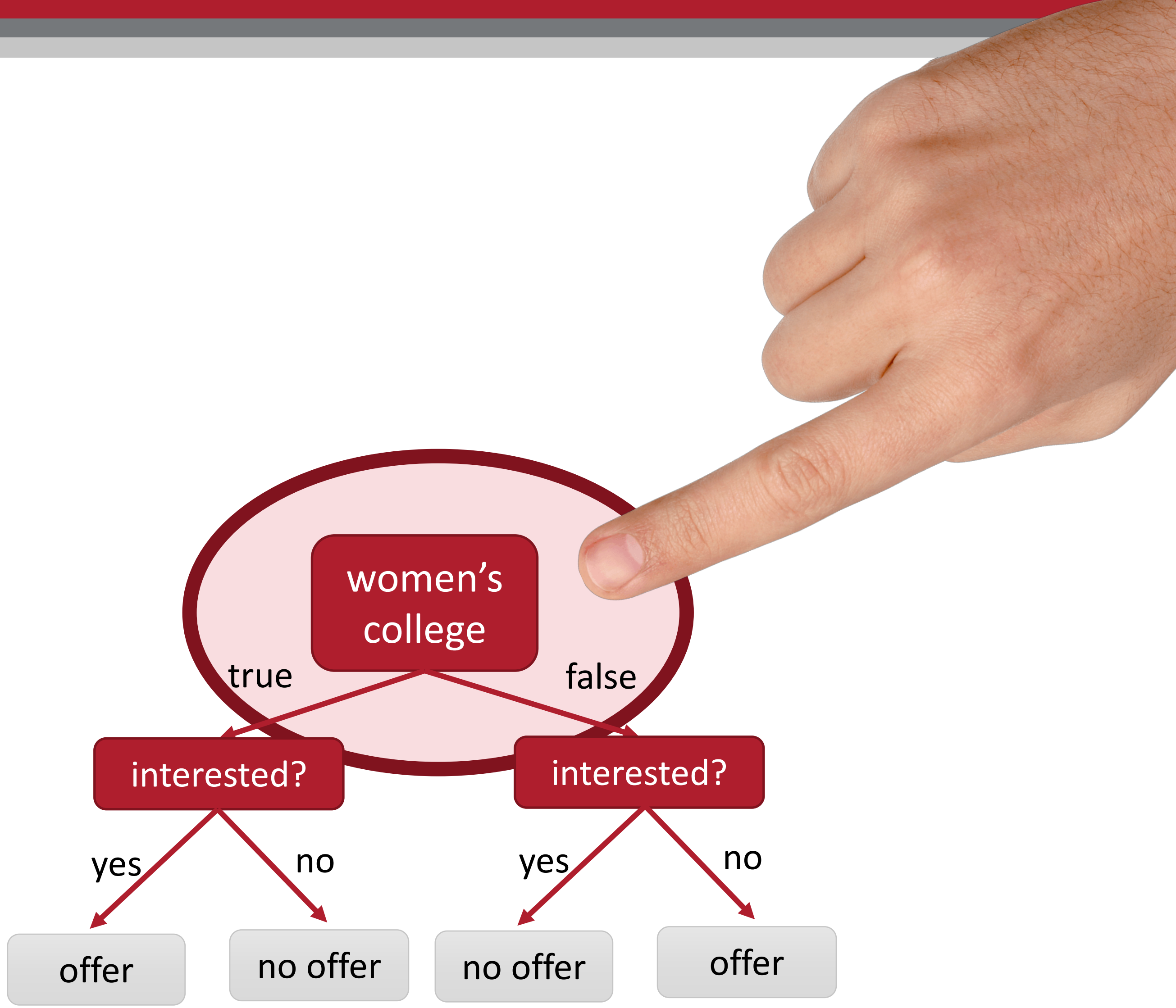
Key Intuition:

- **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

$\langle exp \rangle$ ::= $\boldsymbol{R}$ | **True** | **False** | var

| op($\langle exp \rangle$, … , $\langle exp \rangle$)

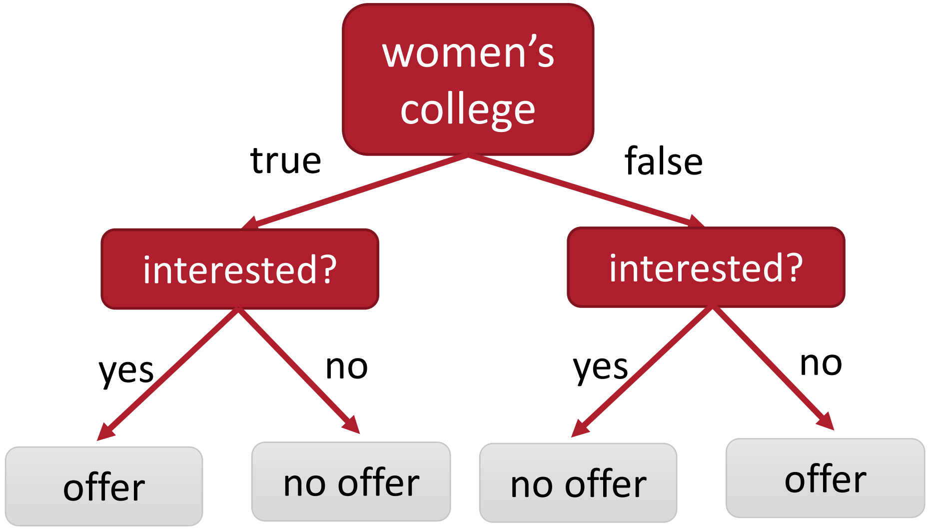
| **if (** $\langle exp \rangle$ **) then** { $\langle exp \rangle$ } **else** { $\langle exp \rangle$ }

Operations:

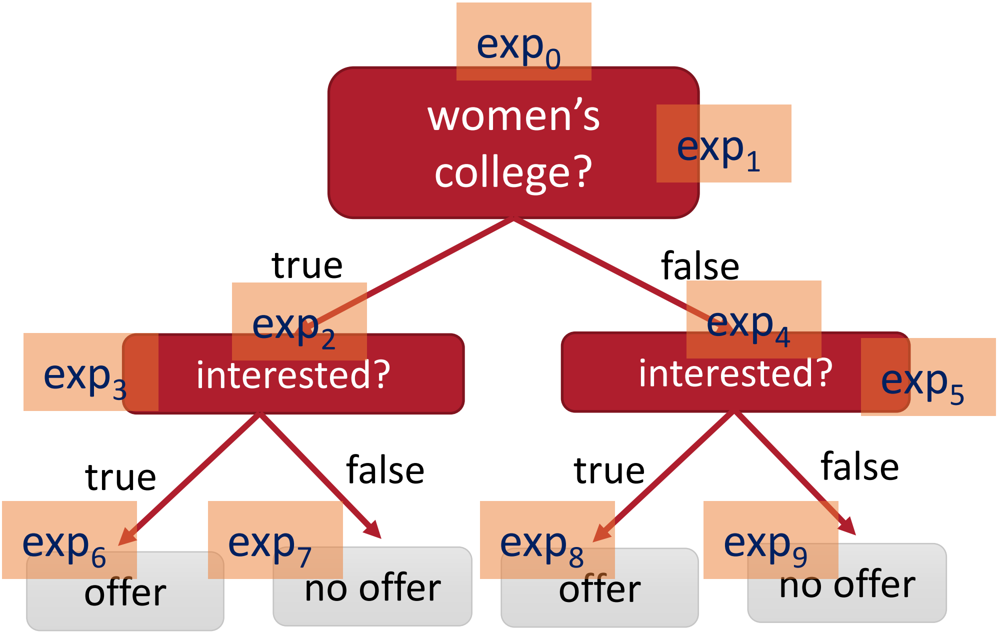
arithmetic operations: +, -, *, etc.

boolean connectives: or, and, not, etc.

relations: ==, <, ≤, >, etc.



Modeling Systems | Probabilistic Semantics



Expression semantics:

$$[\![\mathrm{exp}]\!] : \mathrm{Instance} \rightarrow \mathrm{Value}$$

I is a random variable over dataset instances

$$[\![\mathrm{exp}]\!] : I \rightarrow V$$

V is a random variable over the expression's value

Joint over input instance (I) and expression values (V_i) for each expression exp_i.

$$\Pr[\ I, V_0, V_1, \ldots, V_9\ ]$$

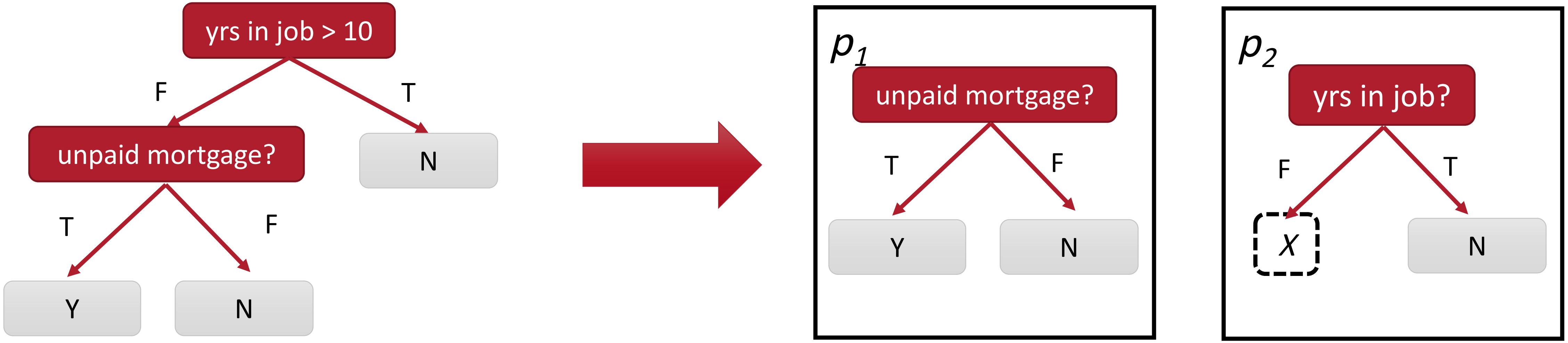
marginals: $\Pr[V_4 = \mathbf{True},\ V_0 = \mathbf{Ad_1}]$

conditionals: $\Pr[V_4 = \mathbf{True} \mid V_0 = \mathbf{Ad_1}]$

Program decomposition

Decomposition

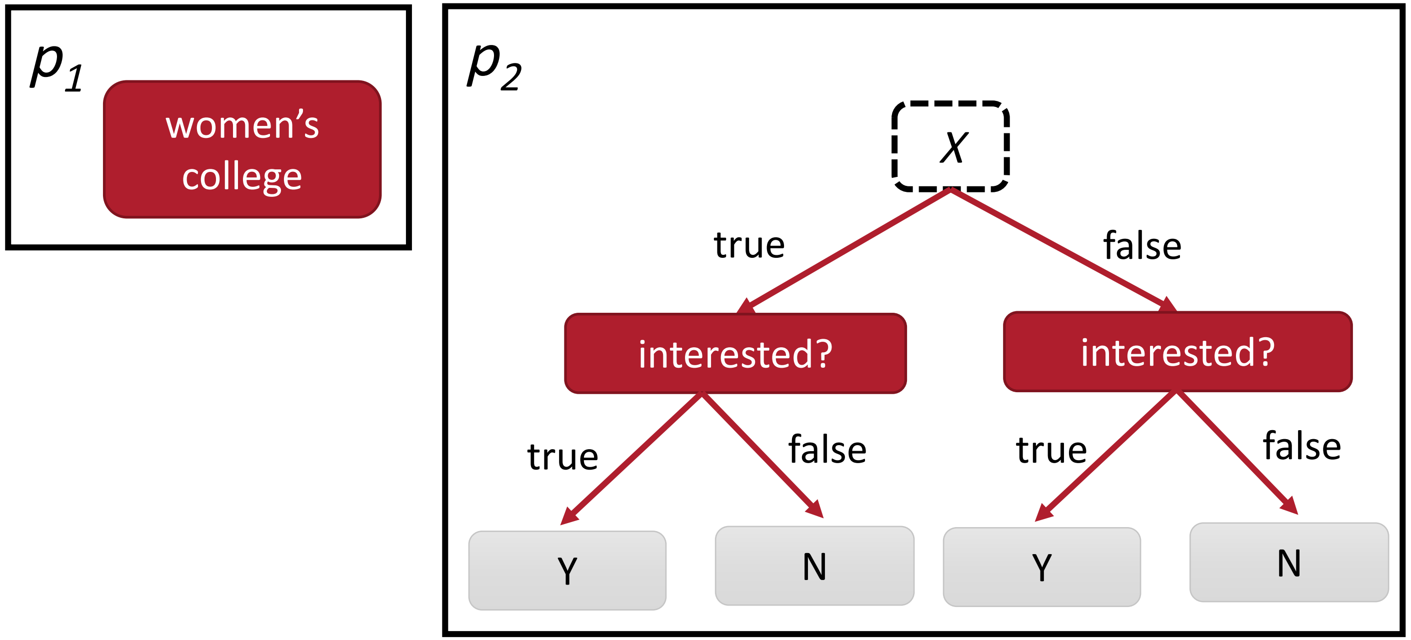
Given a program *p*, a decomposition (p_1, X, p_2) consists of two programs p_1, p_2, *and a fresh variable X* such that replacing X with p_1 inside p_2 yields *p.*



Characterizing **proxies**

Proxy

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

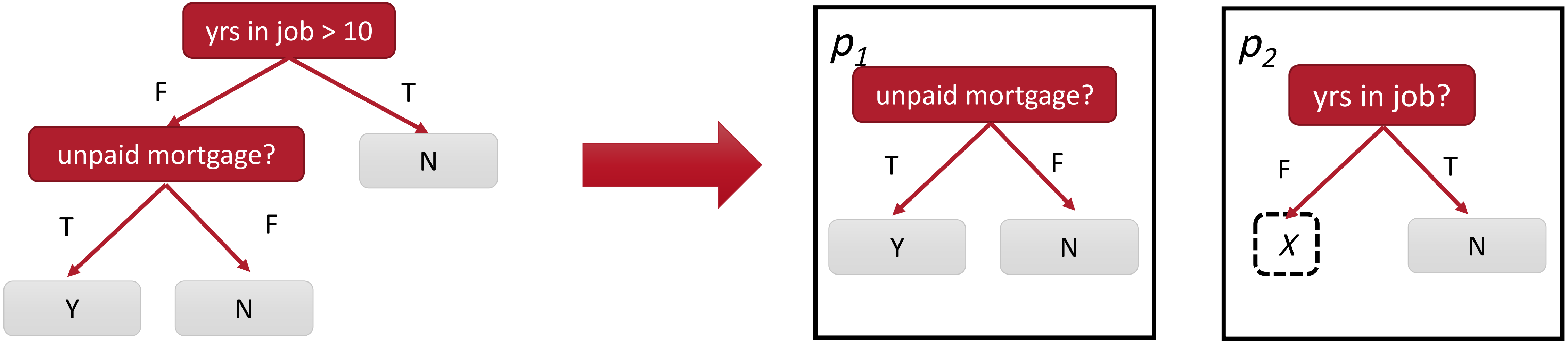


p_1 is a proxy for "gender = Female"

Characterizing use

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 ***(ε, δ)-proxy use of Z*** when

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

A program has (ε, δ)-proxy use of Z when it admits a decomposition that is an (ε, δ)-proxy use of Z

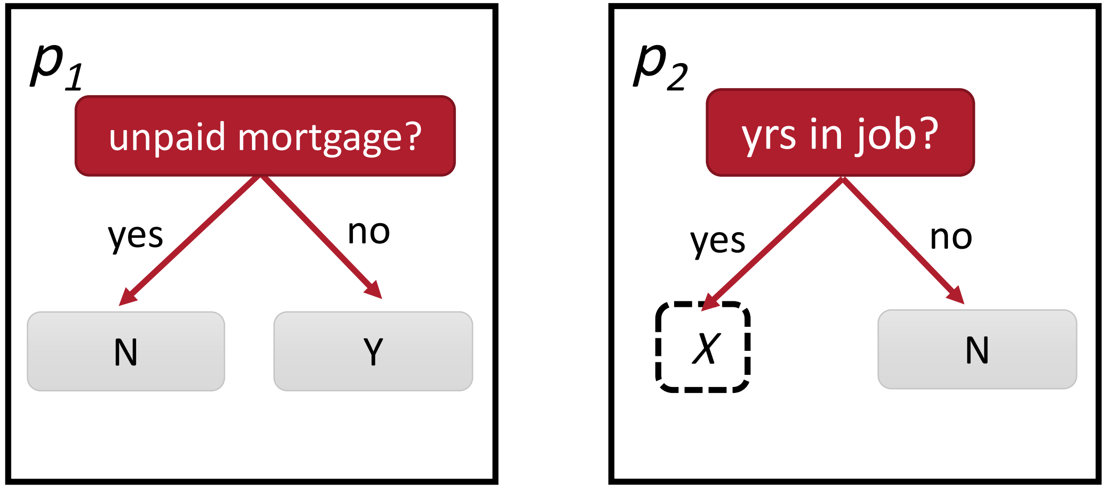
Quantifying decomposition influence

1. Intervene on p_1

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

3. Measure divergence:

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



Algorithmics

- Does system have an (ε, δ)-proxy-use of a protected variable?

- Basic algorithm $O(S*N^2)$
  - S – # expressions
  - N – # dataset instances

- How do we remove (ε, δ)-proxy-use violation?

- Naive algorithm
  - Replace Exp_i with a constant

    O(1) // any constant

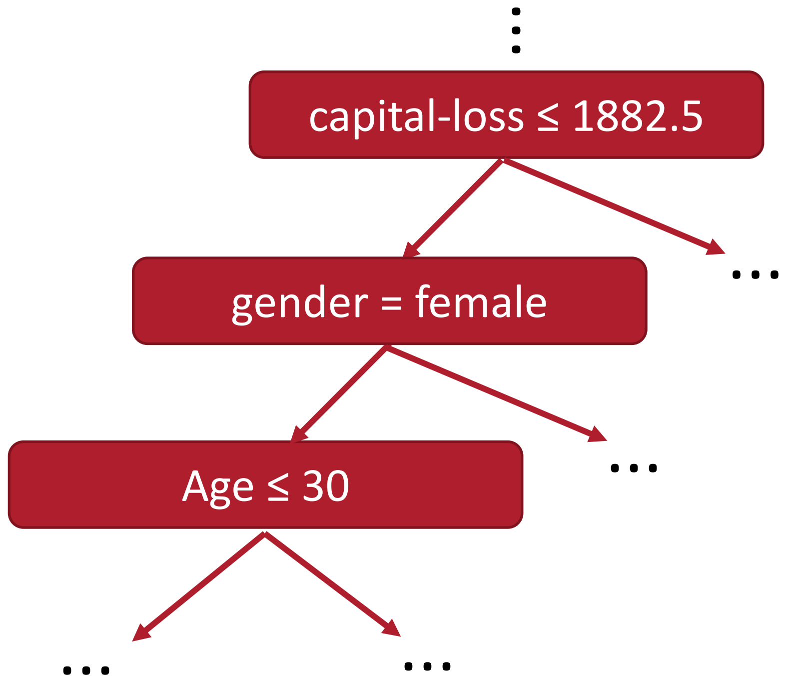
    O(N * M) // best constant, M – # possible values

Witnesses

Using Witnesses

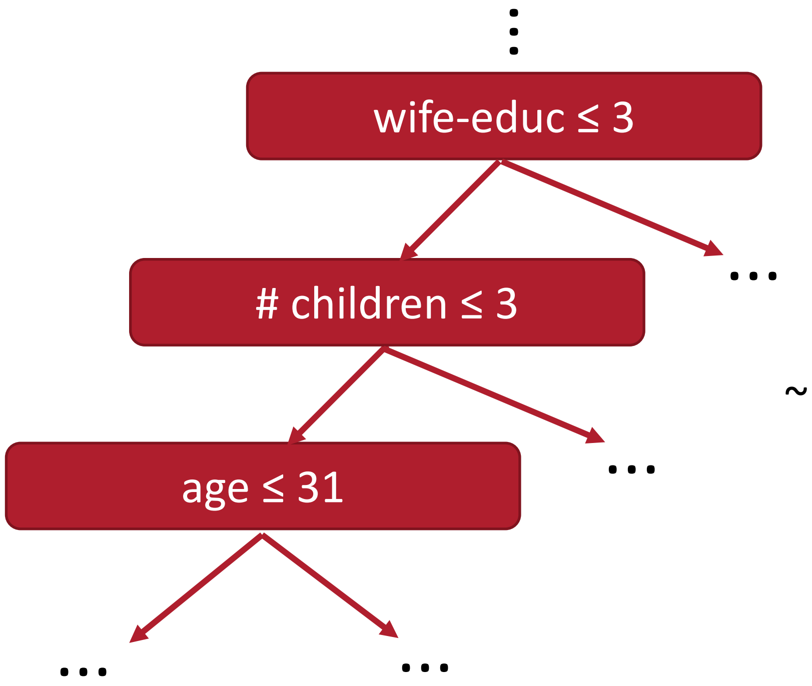
Demonstration of violation in the system
Localize where scrutiny/human eyeballs need to be applied
Determine what repair should be applied

Experiments: Benchmark datasets (CCS’17)



~ Marital status

model accuracy: 83.6 %
after repair: 81.7%



~ Wife’s religion

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]

$$Y(\boldsymbol{X}) = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

Recall our definition of decomposition influence:

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

We generalize to regression by defining:

$$\iota(p_1, p_2) = \mathbf{E}_{X,X'}[\ ([\![p]\!](X) - [\![p_2]\!](X, [\![p_1]\!](X')))^2\ ]$$

Proxies in regressors [NIPS'18]

$$Y(\boldsymbol{X}) = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

What are the decompositions?

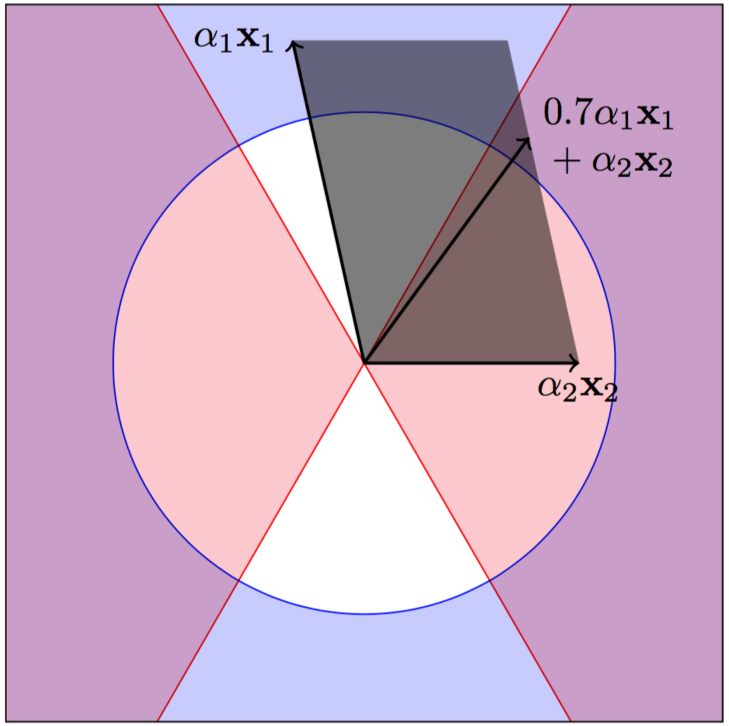
- Just individual terms a_nX_n? Or groups like $a_1X_1 + a_2X_2$?
- What about $0.5 * a_1X_1 + a_2X_2$?

$$\text{Component } P(\boldsymbol{X}) = \beta_1 a_1X_1 + \beta_2 a_2X_2 + \ldots + \beta_n a_nX_n$$
$$\text{for } \beta_1, \ldots, \beta_n \in [0, 1]$$

Proxies in regressors [NIPS'18]

$$\iota(p_1, p_2) = \mathbf{E}_{X,X'}[\,(Y(\boldsymbol{X}) - Y(\boldsymbol{X}, P(\boldsymbol{X}')))^2\,] \propto \mathrm{Var}(\,P(\boldsymbol{X})\,)$$

$$\mathrm{Asc}(Y, Z) \propto \mathrm{Cov}(Y, Z)$$



Optimize to find proxies!

Find	$\mathbf{max}_{\beta}\, \cancel{\lVert A\beta \rVert}\ c^T\beta\ (\text{for}\ \lVert A\beta \rVert \le c^T\beta\,)$
such that	$\lvert \mathrm{Asc}(A\beta\,,\, Z)\rvert \geq \varepsilon$
where	$A^TA = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X})$

Agenda

Methods for dealing with inappropriate information use

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

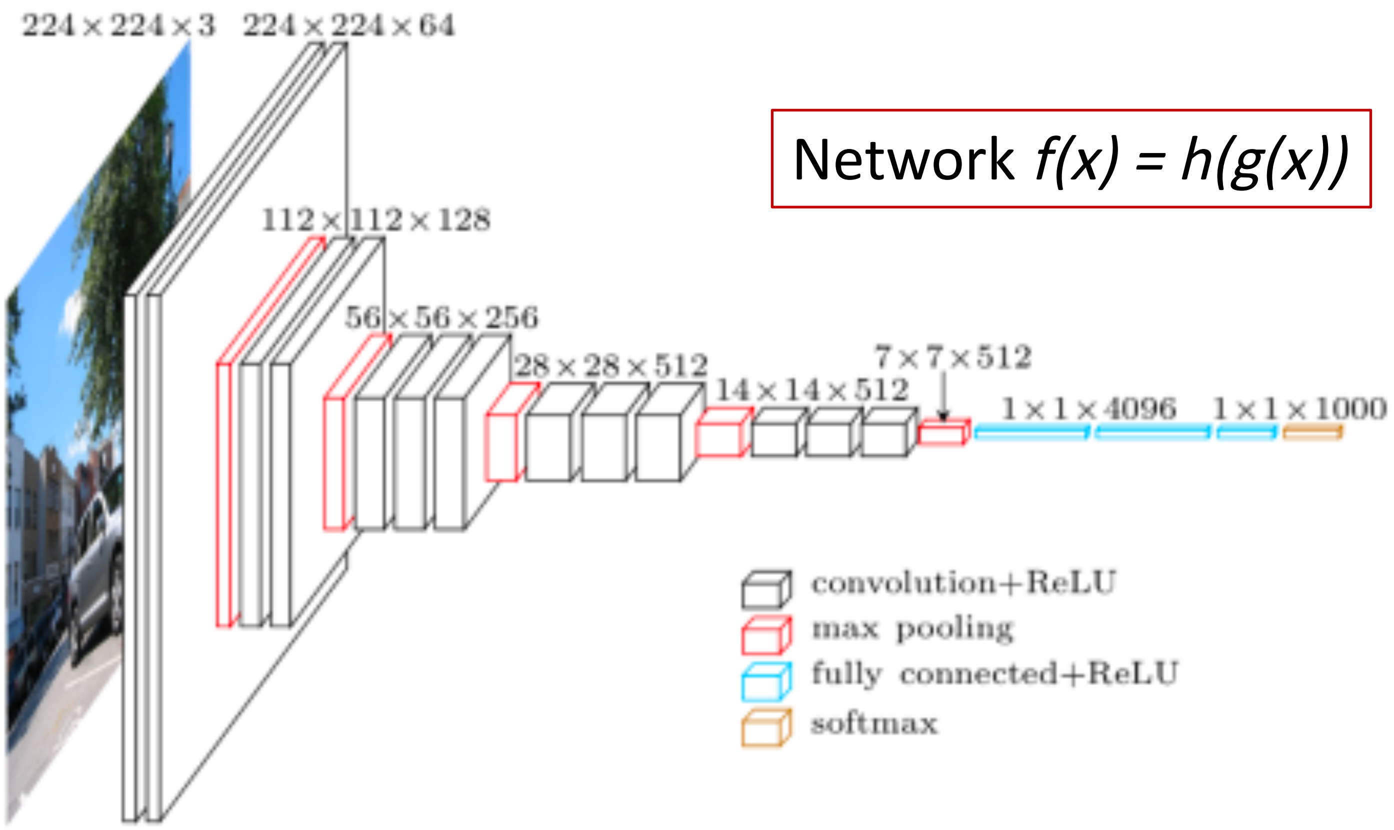
Remaining talk:

- 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

$$\iota_j(f,P) = \int_{\mathcal{X}} \left.\frac{\partial g}{\partial z_j}\right|_{h(x)} P(x)\, dx$$

Axioms:

- Linear agreement
- Distributional marginality
- Distribution linearity

Problems with neural nets: stereotyping

basketball
(73%)

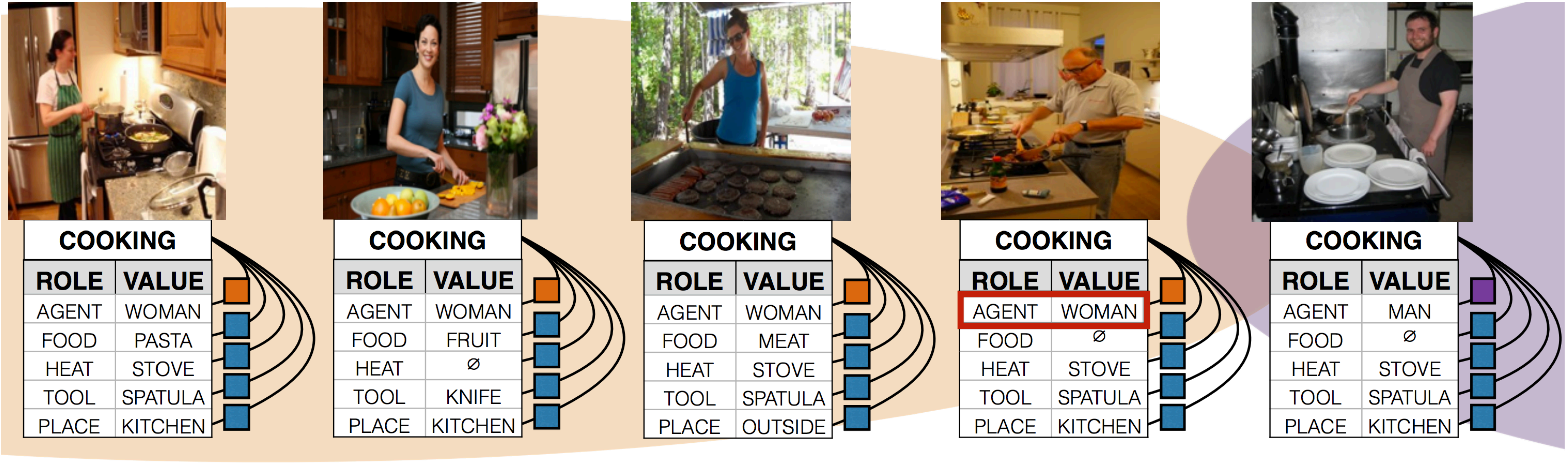
ping-pong ball
(37%)

ballplayer
(90%)

See [Stock & Sisse, 2018] for more examples like this

Problems with neural nets: bias amplification

Image source: [Zhao et al., EMNLP 2018]

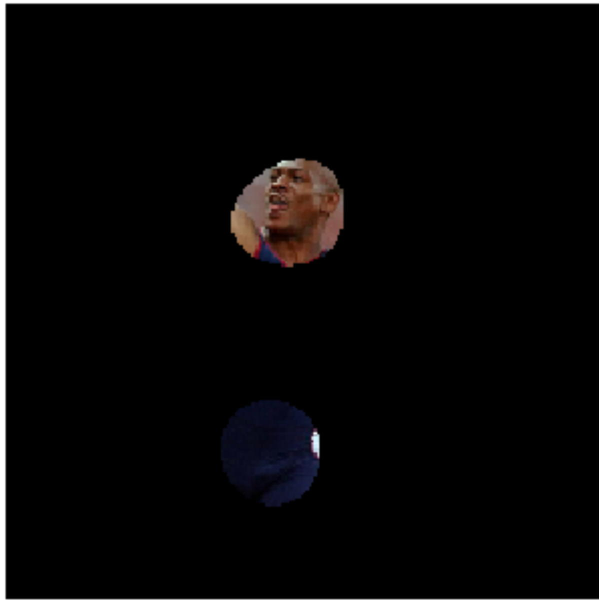


In training data, 33% of "cooking" images have men in them

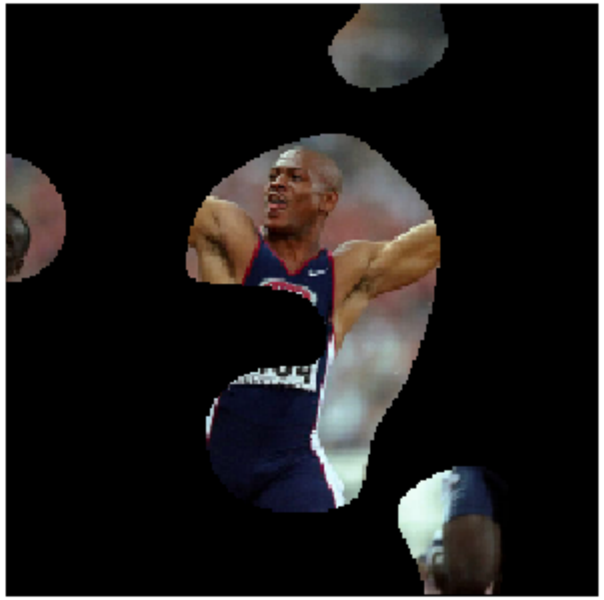
In predictions, 16% of "agent" roles in cooking images are labeled "man"

Explaining stereotype predictions

basketball
(73%)

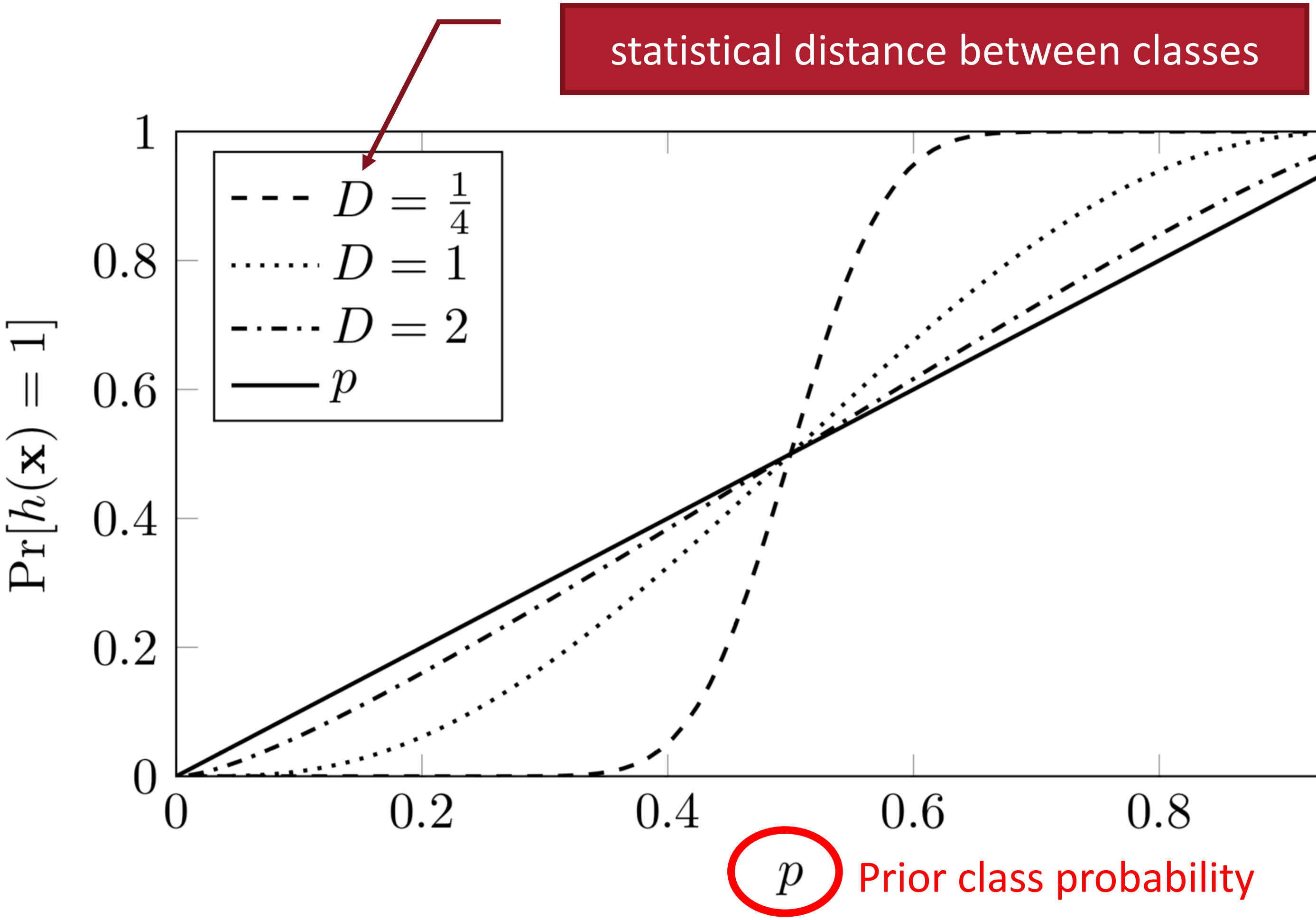


top 5% most
influential features



top 25% most
influential features

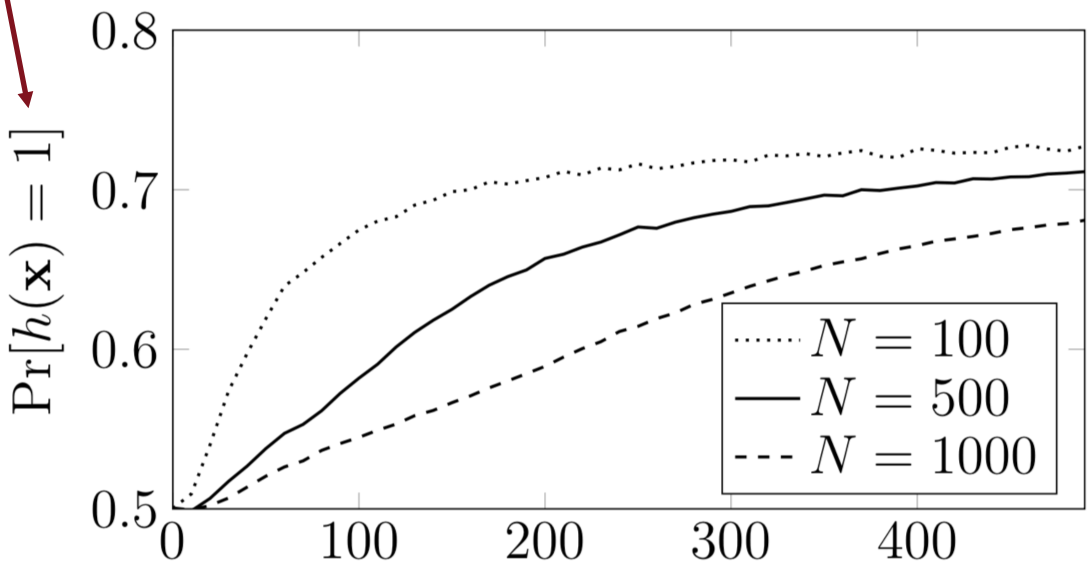
Intrinsic bias amplification



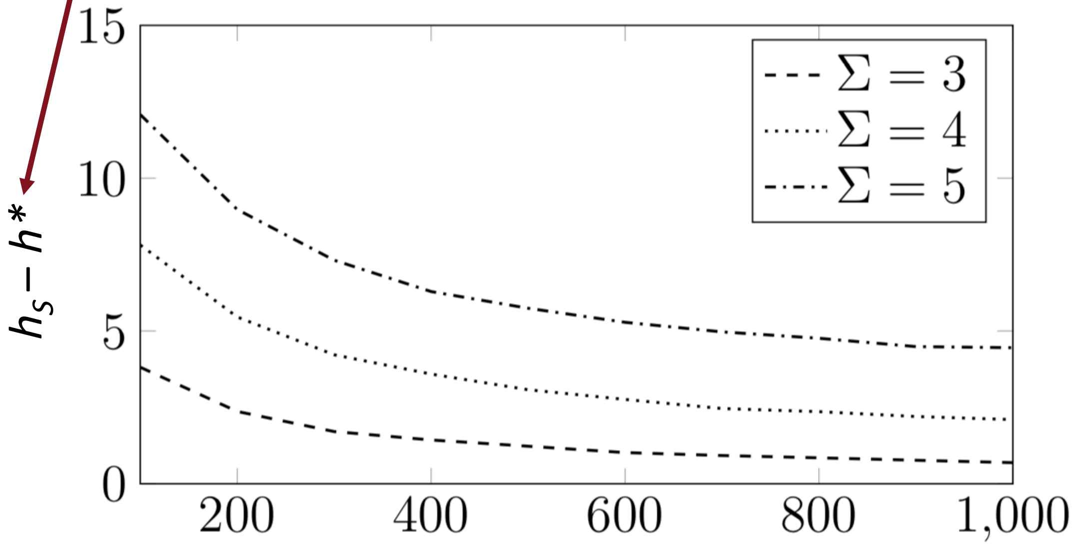
Prediction bias from inductive bias

= 0.5 + prediction bias

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



"weak" features for h(x) = 1



data size

Larger weights $\approx$ More influence

Simple fix: kill weak features

most positive-influential features to keep

Bias of resulting classifier

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} \left| B_{\mathcal{D}}(g_\beta^\alpha) \right| \text{ subject to } L_S(g_\beta^\alpha) \leq L_S(g)$$

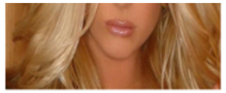
most negative-influential features to keep

Don’t increase the emprical loss

Early results

dataset	p^*	*asymm. (%)*	$B_{\mathcal{D}}(h_S)$	$B_{\mathcal{D}}(h_S)$ *(post-fix)* *par*	*exp*	ℓ_1	*acc. (%)*	*acc. (%) (post-fix)* *par*	*exp*	ℓ_1
CIFAR10	50.0	52.0	1.8	1.7	**0.4**	*n/a*	93.0	93.1	**94.0**	*n/a*
CelebA	50.4	50.2	7.7	7.7	**0.2**	*n/a*	79.6	79.6	**79.9**	*n/a*
arcene	56.0	57.7	2.7	**0.6**	1.2	1.7	68.9	69.0	**74.2**	69.4
colon	64.5	51.0	23.1	22.9	**22.6**	35.5	58.5	58.7	58.7	**64.5**
glioma	69.4	54.8	17.4	17.4	**12.2**	17.0	76.3	**76.3**	76.7	75.44
micromass	69.0	54.1	0.68	**0.66**	0.69	0.68	**98.4**	**98.4**	**98.4**	**98.4**
pc/mac	50.5	60.6	1.6	1.6	**1.4**	1.6	**89.0**	**89.0**	88.0	**89.0**
prostate	51.0	44.4	47.3	47.2	**10.0**	28.1	52.7	52.8	**90.2**	71.3
smokers	51.9	50.4	47.4	45.4	**8.0**	33.0	50.0	50.7	**59.0**	51.2
synthetic	50.0	99.9	24.1	17.2	23.6	**5.7**	74.9	**77.9**	74.8	71.4





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