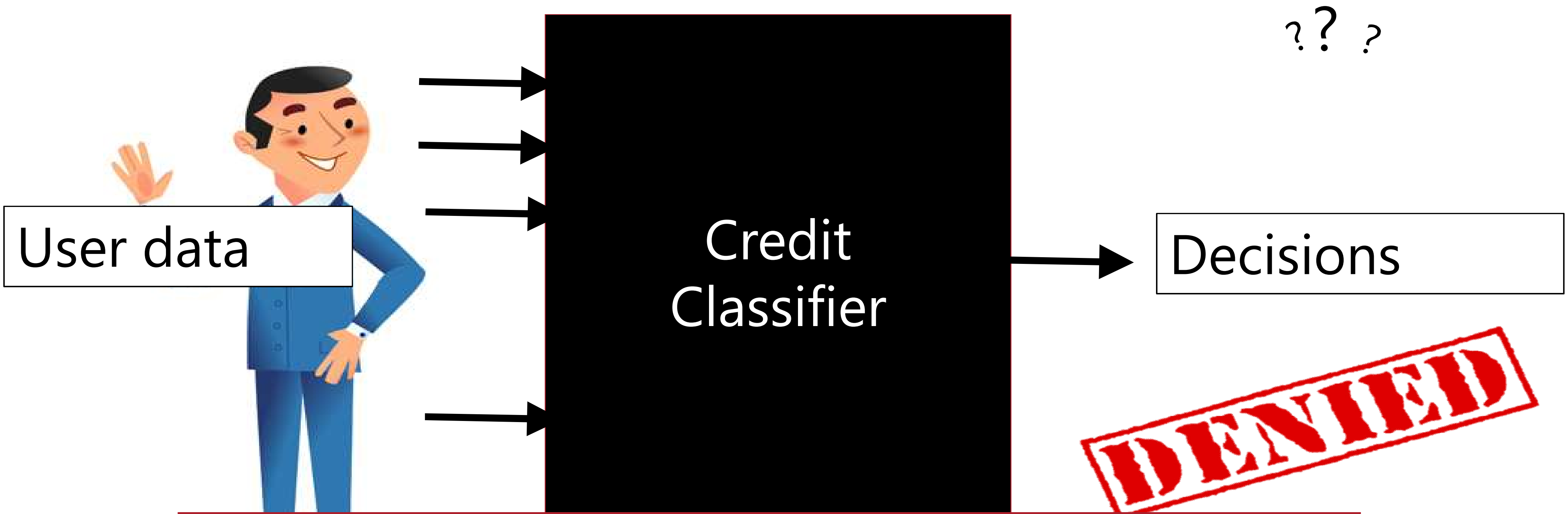
Influence-directed Explanations for Machine Learning Systems

Anupam Datta

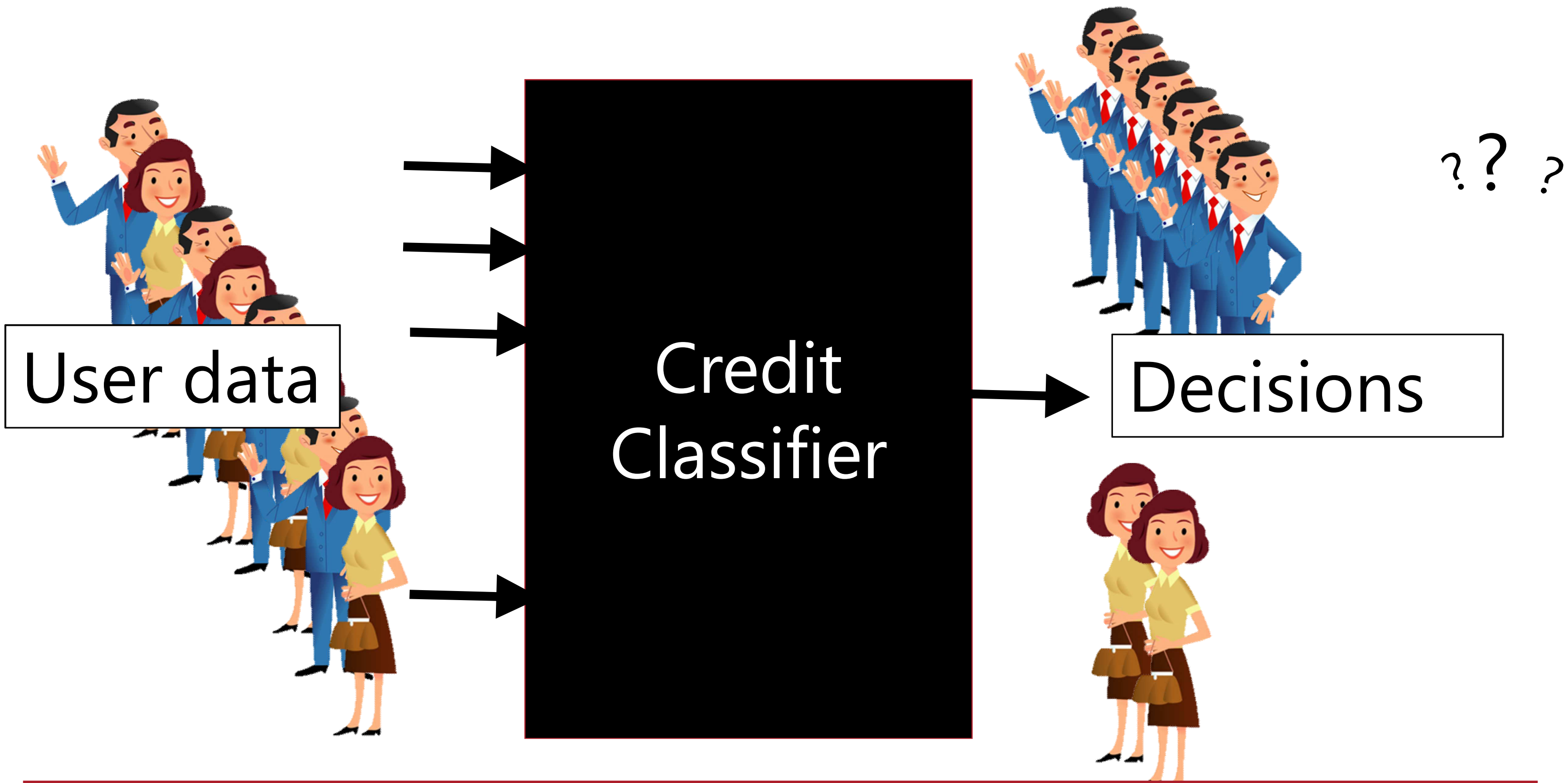
Professor
Electrical and Computer Engineering
Computer Science
Carnegie Mellon University

Machine Learning Systems are Opaque



Why was Joe denied credit?

Machine Learning Systems are Opaque



Why gender disparity in approvals?

Vision: Explainable Machine Learning Systems

Reveal "meaningful information about the logic" of the machine learnt prediction/decision model

- Enable humans + machines to make decisions together
- Build trust in and debug models
- Guard against societal harms, e.g. unfairness
- Comply with regulations, e.g. EU GDPR, US ECOA
- Applications: Finance, healthcare

Abstraction is key

Explaining property of a ML system =
identify causally influential factors +
make them human interpretable

- Causation: What are important factors causing this model property?
- Interpretation: What do these factors mean?

Quantitative Input Influence
[Datta, Sen, Zick 2016]

How much <u>influence</u> do various inputs (features) have on a given classifier's decision about individuals or groups?

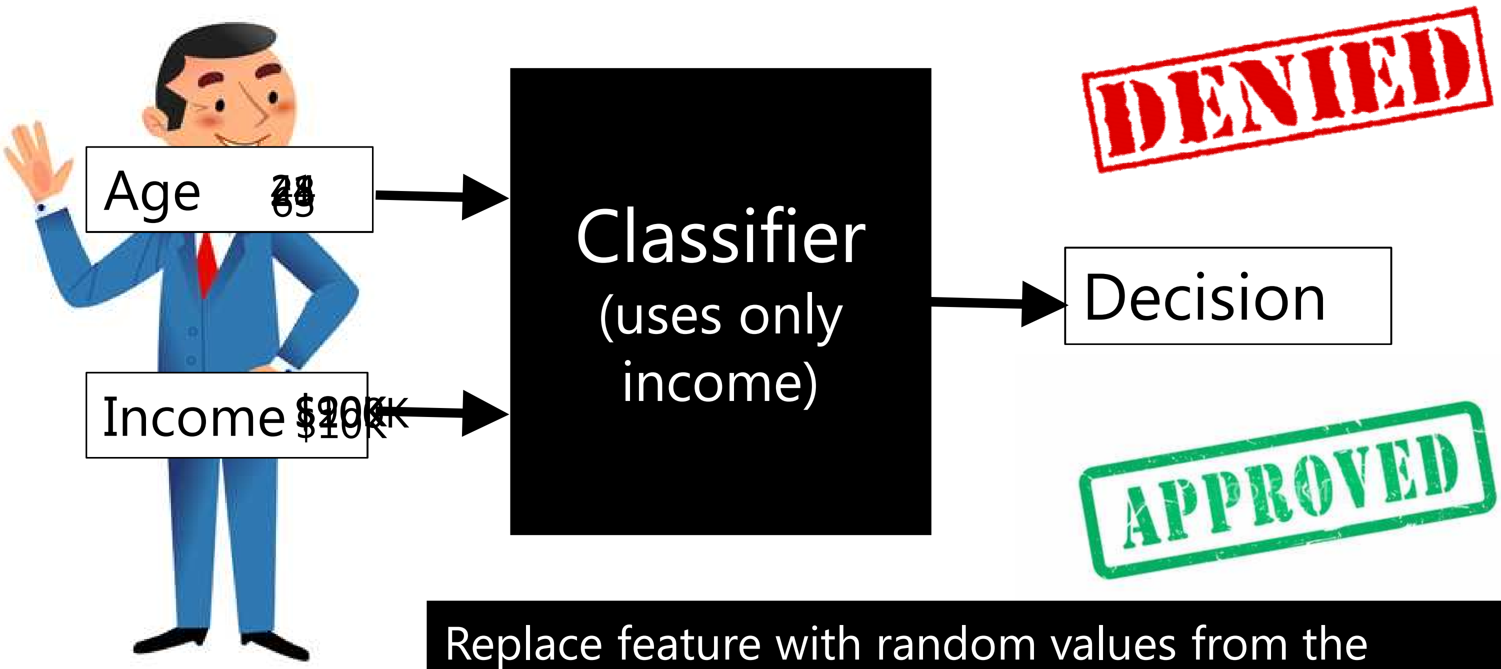
Age	27
Workclass	Private
Education	Preschool
Marital Status	Married
Occupation	Farming-Fishing
Relationship to household income	Other Relative
Race	White
Gender	Male
Capital gain	$41310
…..	

Negative Factors:
Occupation
Education Level

Positive Factors:
Capital Gain

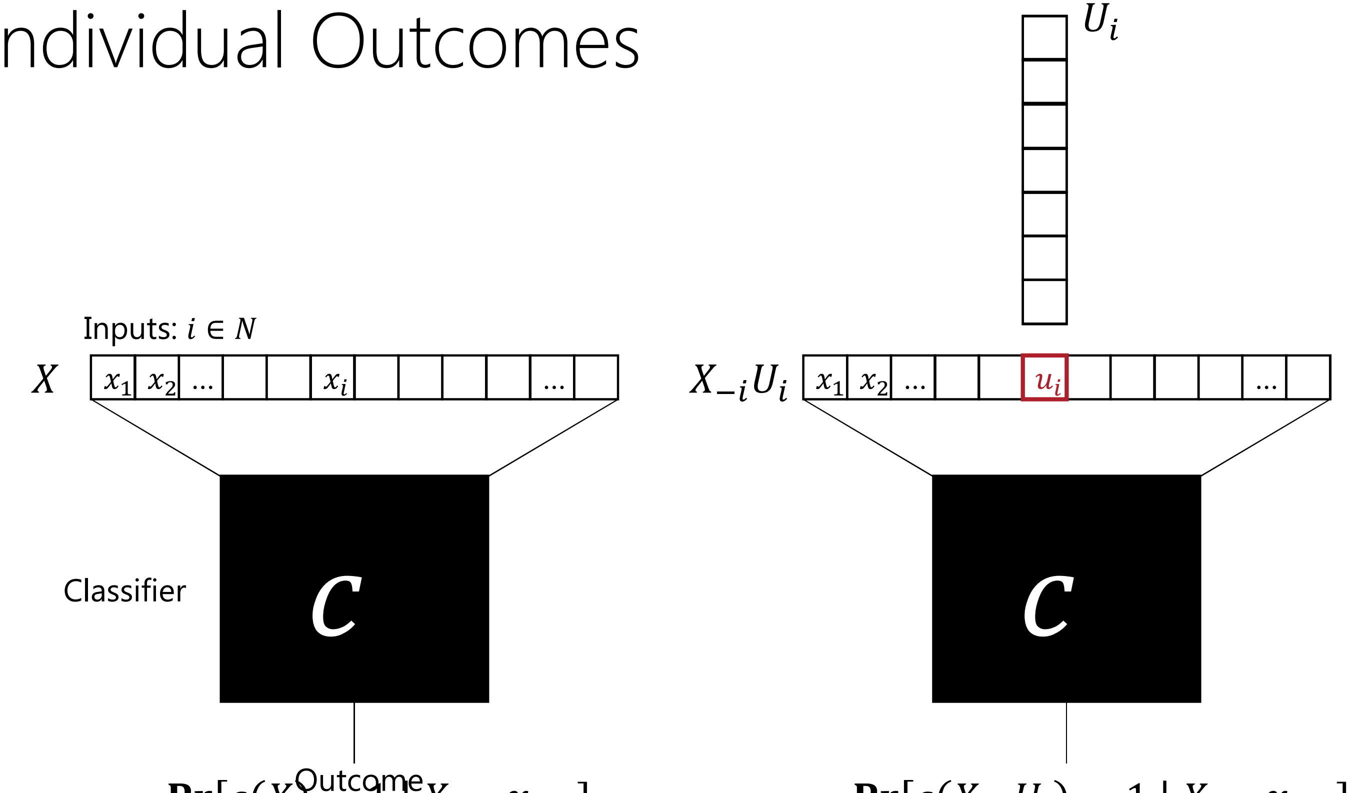
Locally linear model

Key Idea | Causal Testing



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

QII for Individual Outcomes

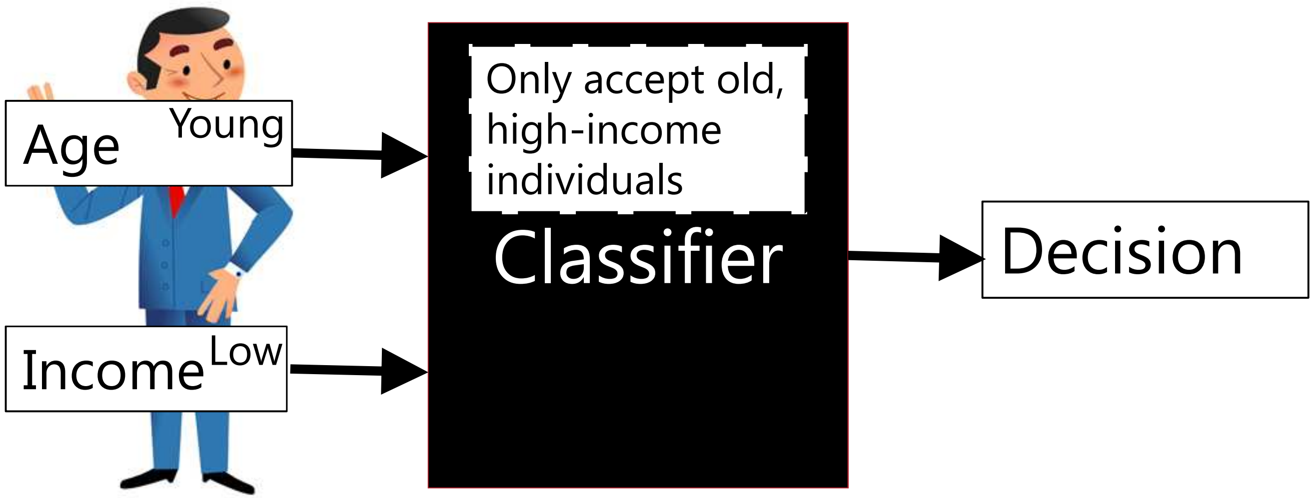


$\mathbf{Pr}[c(X) = 1 \mid X = \boldsymbol{x}_{\text{Joe}}]$

$\mathbf{Pr}[c(X_{-i}U_i) = 1 \mid X = \boldsymbol{x}_{\text{Joe}}]$

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

Challenge | Joint and Marginal Influence



- Single inputs alone may have insignificant influence.

Observation: Similar to voting
Approach: Model influence as a cooperative game.
Use game-theoretic power indices.

Key Idea | Marginal Influence

Think of features as states in an election

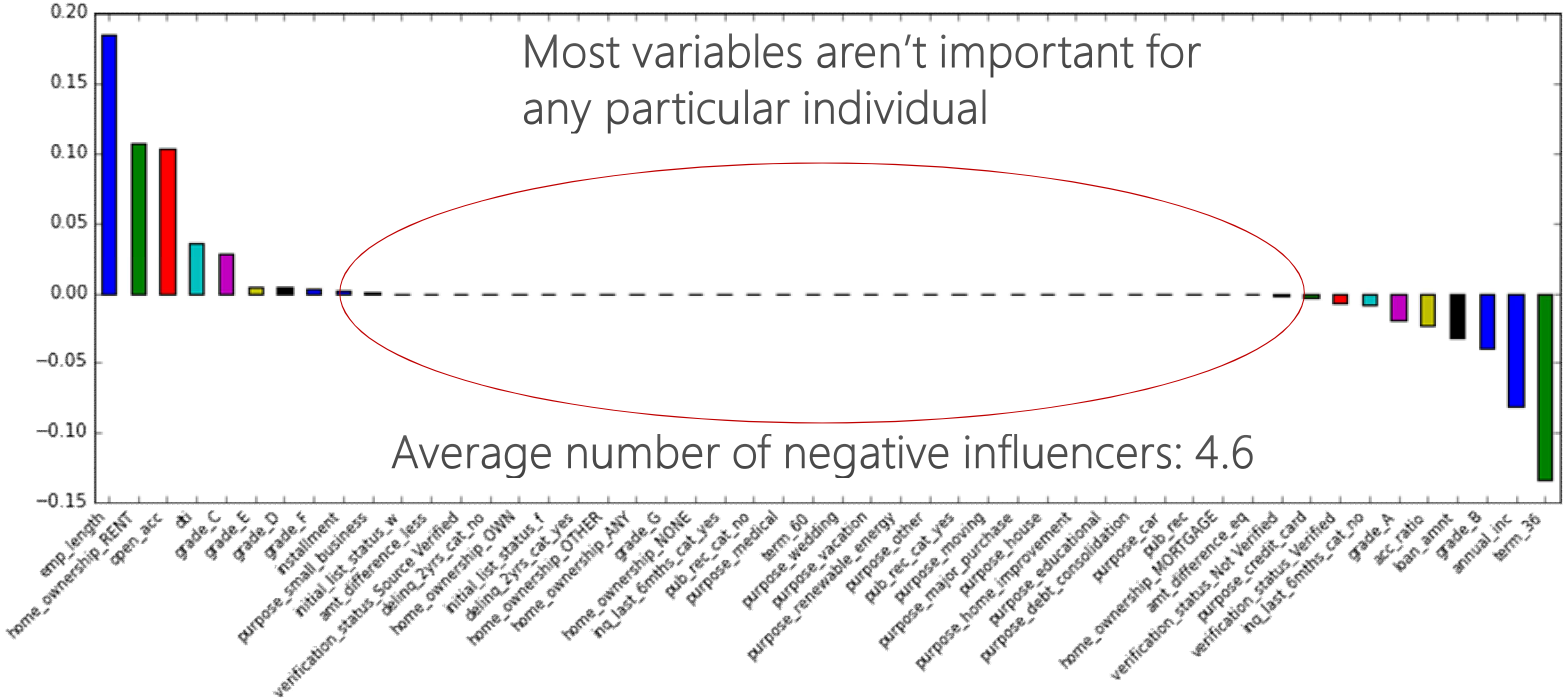
What is the effect of PA after results from IN, GA, MD are in?

[NY Times Election Needle]

Aggregate marginal influences using appropriate power index (e.g., Shapley)

Case study with Lending Club data

51-variable tree ensembles: scalable, succinct explanations



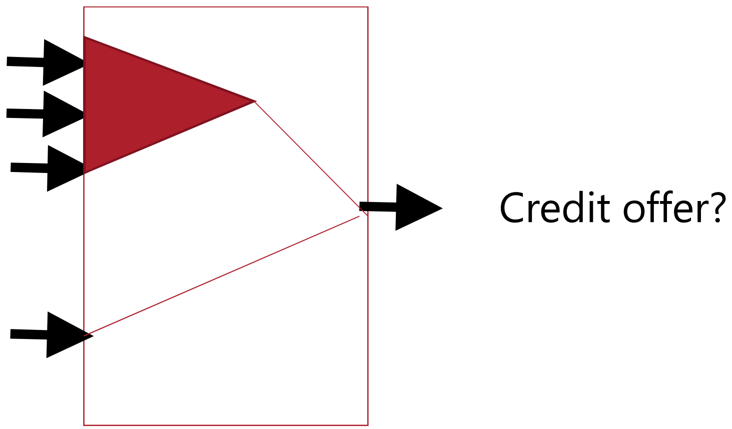
ECOA style
Adverse Action Notice

Employment Length
Home ownership
Open accounts
DTI

Proxy use and indirect discrimination
[Datta, Fredrikson, Ko, Mardziel, Sen 2017]

Protected information type:
Race

- Age
- Income
- Zip-code
- …



Example models: Tree ensembles

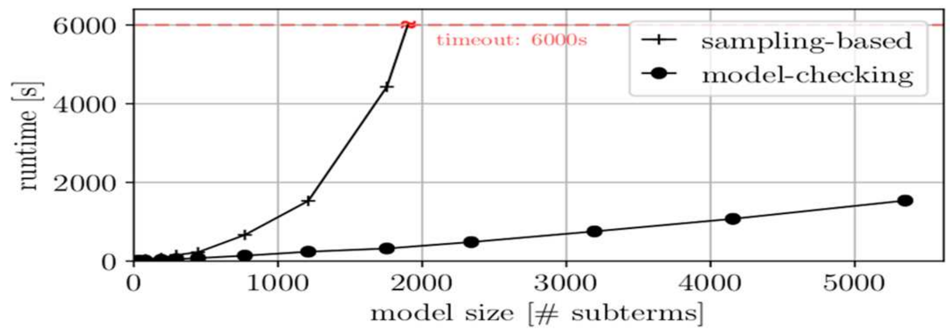
Proxy use

1. **Strong predictor (associated)**
2. **Causally affects output (high QII)**

Model checking for proxy use [Ko, Mardziel, Sen, Datta, Fredrikson 2018]

- ML models are probabilistic programs
- Checking for proxy use reduced to checking a reachability property via self composition
- Scalability improved by order of magnitude using an abstraction technique

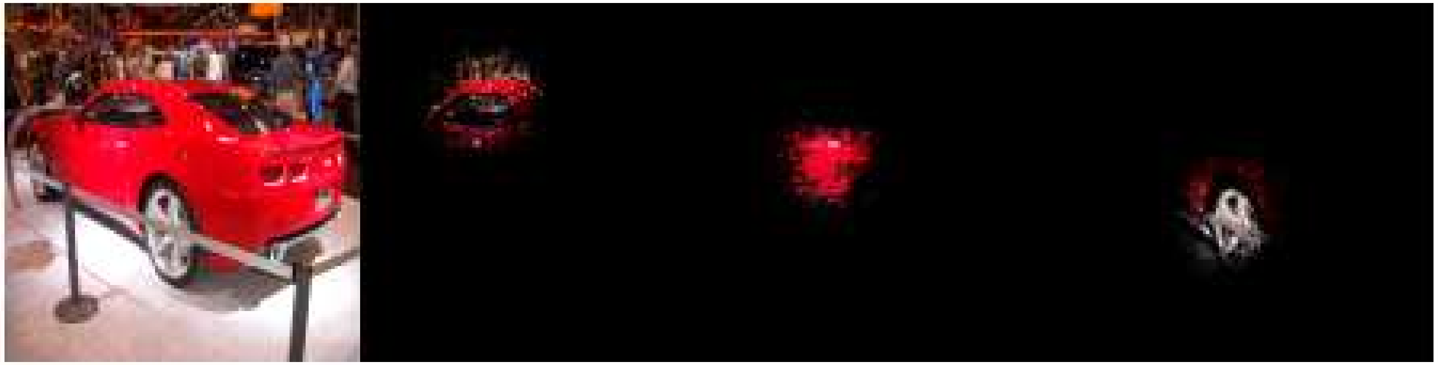
PRISM results: runtime comparison vs. our previous work



Influence-directed explanations

[Leino, Sen, Li, Datta, Fredrikson 2018]

- Identify causally influential neurons in internal layers
- Give them interpretation using visualization techniques



White-box model, scalable, axiomatically justified like the Shapley value

Why did the network classify input as sports car instead of convertible?

VGG16 ImageNet model

Input image

Influence-directed Explanation

Uncovers high-level concepts that generalize across input instances

Abstraction is key

Explaining property of a ML system =
identify causally influential factors +
make them human interpretable

Vision: Explainable Machine Learning Systems

Reveal "meaningful information about the logic" of the machine learnt prediction/decision model

- Enable humans + machines to make decisions together
- Build trust in and debug models
- Guard against societal harms, e.g. unfairness
- Comply with regulations , e.g. EU GDPR, US ECOA
- Applications: Finance, healthcare

Thanks!