# Algorithmic Transparency via Quantitative Input Influence

Datta et al. 2016 IEEE Symposium on Security and Privacy

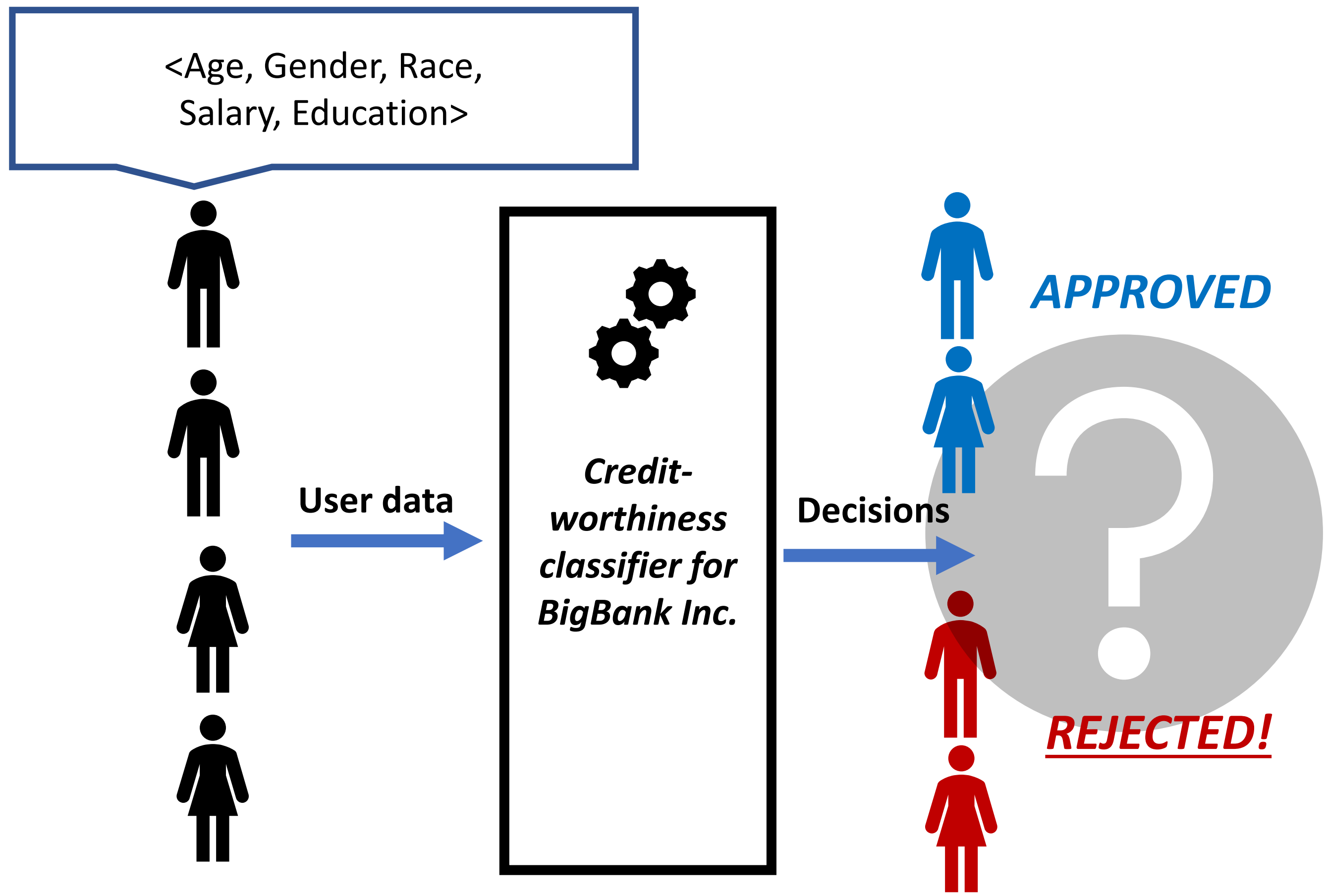
Presented by: Yigitcan Kaya

# *Agenda*

- **Problem Statement**
- **Goals**
- **Solution**
  - Challenges
  - Building Blocks
- **Experiments**
- **Related Work**
- **Conclusion**

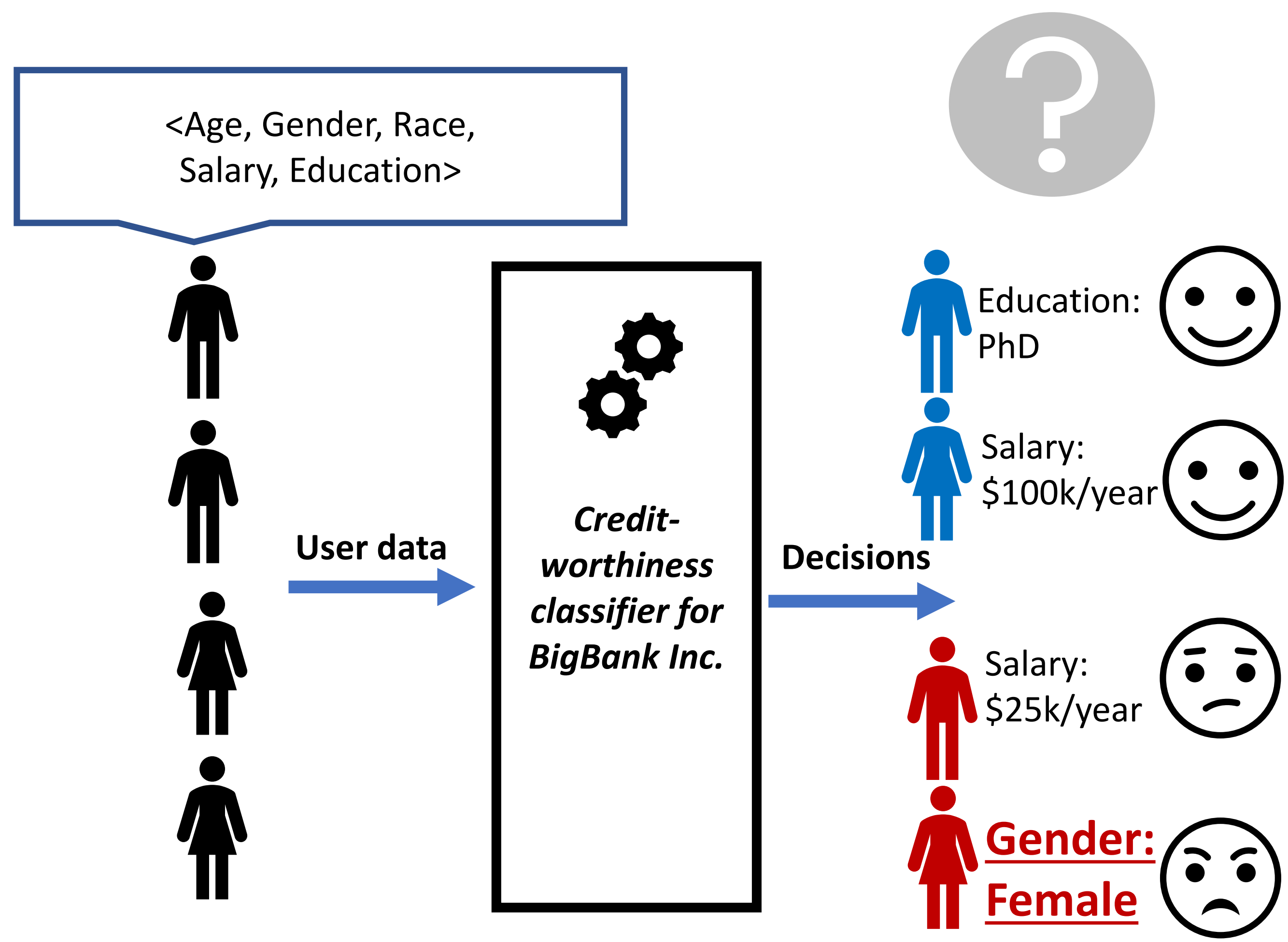
# Problem:

*Opaque machine learning systems*



**Goal:**

*Measuring the influence inputs have on decisions*



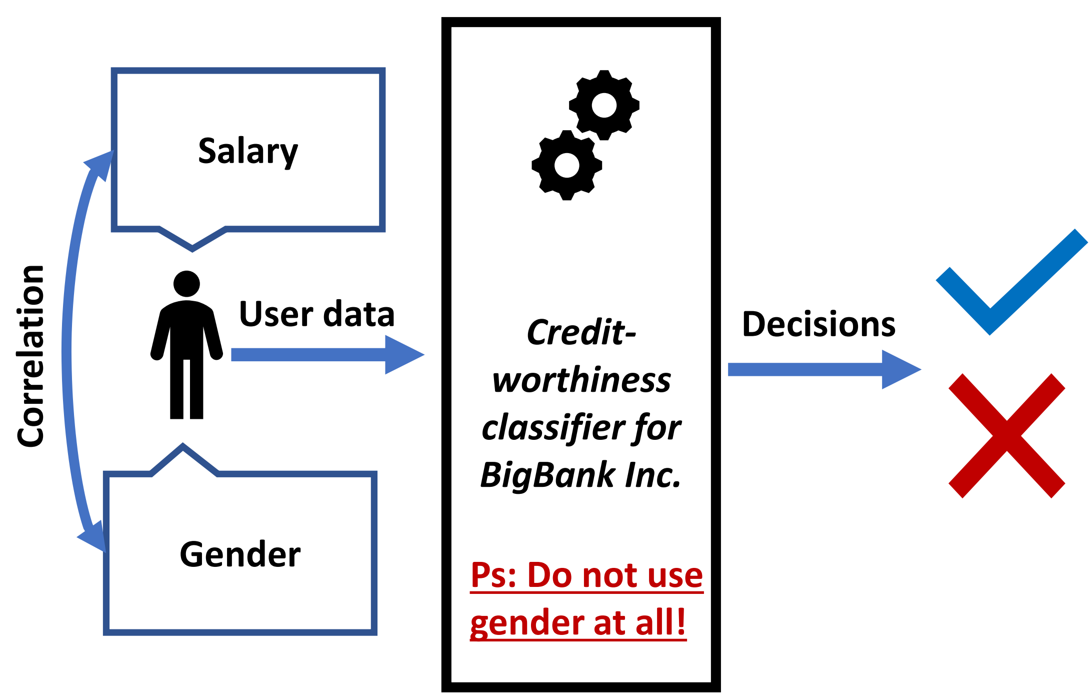
## *Solution:*

*Quantitative Input Influence (**QII**)*

- A family of metrics to generate transparency reports
- Black-box access with the knowledge of the data

# *Challenge #1:*

*Correlated inputs*



# *Challenge #2:*

*General transparency queries*

**Individual**

- *“Which input had the most influence in my credit denial?”*

**Group**

- “What inputs have the most influence on credit decisions of women?”

**Disparity**

- “What inputs influence men getting more positive outcomes than women?”

# *Building Blocks of QII*

**QII**: A technique of measuring the influence of an input on its outputs.

**Causal Intervention**

- Deals with the correlated inputs

**Quantity of Intervention**

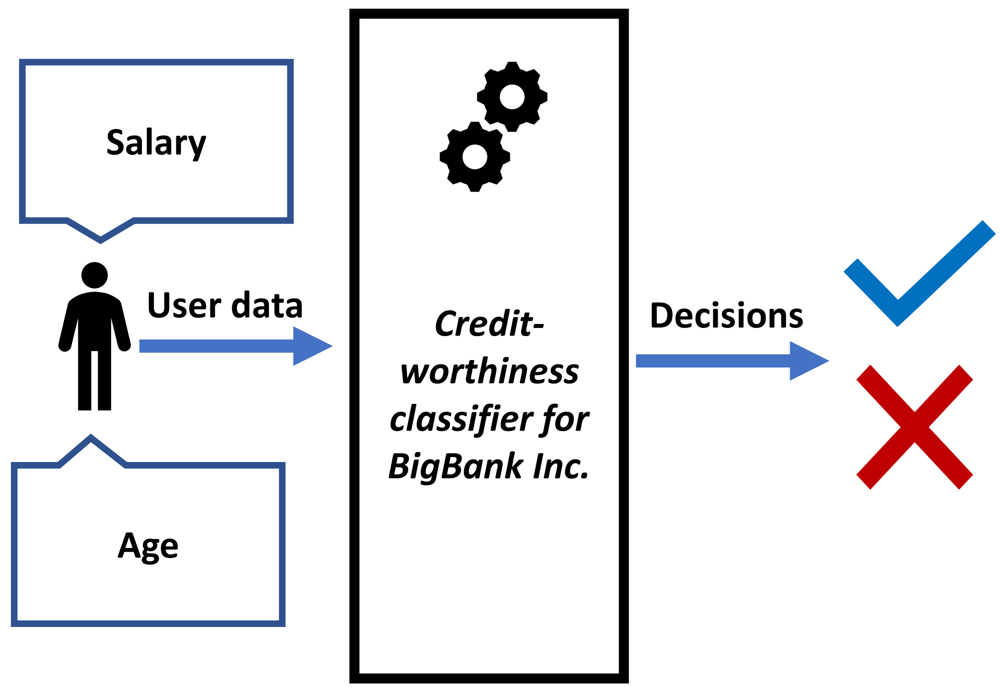
- Supports a general class of transparency queries

## *Building Block #1:*

*Causal intervention*

*QII for Individual Outcomes*

**Basic Idea**: Keep one feature fixed, and vary the other in a specific way to measure.



# *Building Block #1:*

*Causal intervention*

*QII for Individual Outcomes*

Real: $X \sim (x_1, x_2, \ldots, \boldsymbol{x_i}, \ldots, x_n)$

Quantity Measured: $\Pr[c(X) = 1 \mid X = \boldsymbol{x}]$

Randomized intervention on feature $i$ breaks correlations: replace $x_i$ with an independent random sample $u_i$.

$U_i$

Hypothetical:$X_{-i}U_i(x_1, x_2, \ldots, \boxed{\boldsymbol{u_i}}, \ldots, x_n)$

Quantity Measured : $\Pr[c(X_{-i}U_i) = 1 \mid X = \boldsymbol{x}]$

# *Building Block #2:*

*<u>Quantity of Interest</u>*

- $\boldsymbol{Q_A}(.)$: Various statistics of a system.

**<u>Classification outcome of an individual:</u>**

$$\Pr[c(X) = c(x_0) \mid X = x_0]$$

**<u>Classification outcomes a group:</u>**

$$\Pr[c(X) = 1 \mid X \; is \; female]$$

**<u>Disparity between classification outcomes of groups:</u>**

$$\Pr[c(X) = 1 \mid X \; is \; male] - \Pr[c(X) = 1 \mid X \; is \; female]$$

# *Combining two blocks:*

*<u>QII of an input on a quantity of interest</u>*

**<u>QII of input *i* on the classification outcome of an individual:</u>**

$$\Pr[c(X) = c(x_0) \mid X = x_0] - \Pr[c(X_{-i}U_i) = c(x_0) \mid X = x_0]$$

**<u>QII of input *i* on the classification outcomes a group:</u>**

$$\Pr[c(X) = 1 \mid X \; is \; female] - \Pr[c(X_{-i}U_i) = c(x_0) \mid X = x_0]$$

**<u>QII of input *i* on the disparity between classification outcomes of groups:</u>**

$$\Pr[c(X) = 1 \mid X \; is \; male] - \Pr[c(X) = 1 \mid X \; is \; female] - \Pr[c(X_{-i}U_i) = 1 \mid X \; is \; male] - \Pr[c(X_{-i}U_i) = 1 \mid X \; is \; female]$$

# *Formal Definition of QII*

The Quantitative Input Influence (***QII***) of an input $i$ on a quantity of interest $\boldsymbol{Q_A}(.)$ of a system $\boldsymbol{A}$ is the difference in quantity of interest when the input replaced with random value via an intervention.

$$\iota^{Q_{\mathcal{A}}}(i) = Q_{\mathcal{A}}(X) - Q_{\mathcal{A}}(X_{-i}U_i)$$

**Catch:**

*Single inputs have low influence*

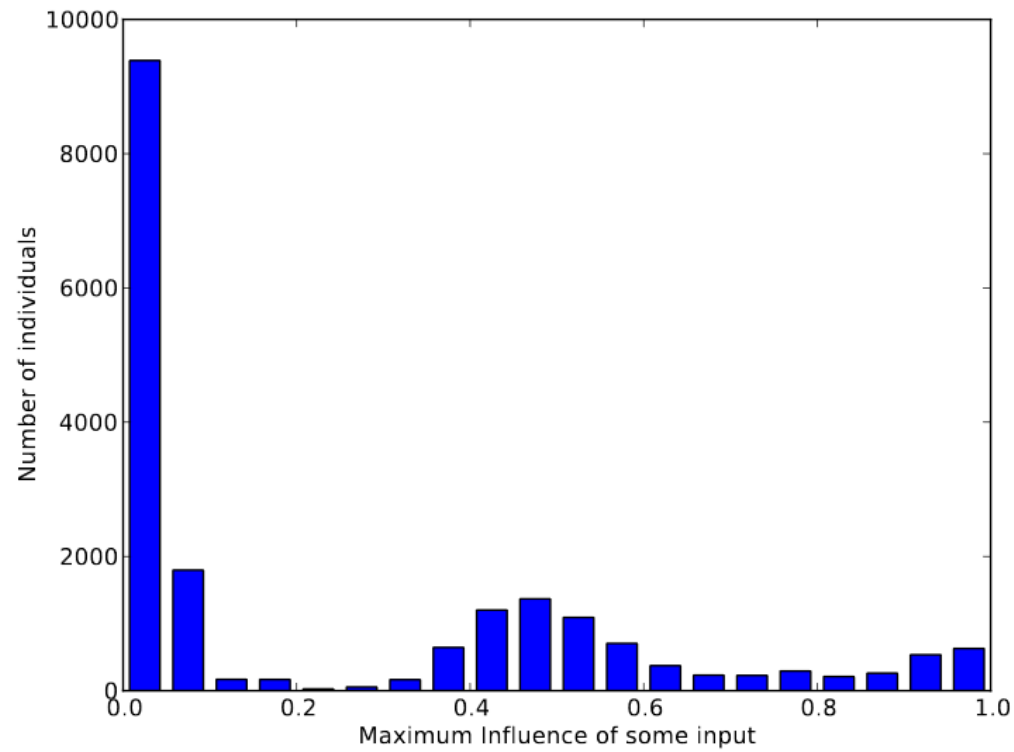



Fig. 1: A histogram of the highest specific causal influence for some feature across individuals in the adult dataset. Alone, most inputs alone have very low influence.

# *Naïve Approach:*

*Set QII*

Instead of a single feature $i$, *r*eplace a set of features $S$ with independent random values from the distribution of inputs.

$$\iota^Q(S) = Q(X) - Q(X_{-S}U_S)$$

**Not all features are equally important within a set $S$!**

# A Better Idea:

*Marginal QII*

Find marginal influence of an input within a set.

Influence of age and income over only income that marginalizes the influence of age

$$i(\{age, income\}) - i(\{income\})$$

**There might be many sets in which $age$ has some marginal contribution!**

…$\{age, income\}, \{age, gender, job\}, \{age, gender, income\}$…

**Need to aggregate marginal QII across all sets.**

# *A Better Idea:*

*Marginal QII*

*Set QII is a cooperative game*

**Cooperative Game:**

- $N$: set of agents
- $v(S)$: Value of set $S$

**Our Setting:**

- Input features are agents
- Influence of feature set $S$, i.e. set QII $i(S)$ is $v(S)$
- Marginal QII is $m_i(S) = v(S \cup \{i\}) - v(S)$

## A Better Idea:

<u>Marginal QII</u>

*Shapley Value*

**Even though it is usually intractable, there are efficient ways to approximate it by sampling.**

$$p[S] = \frac{1}{n} \frac{1}{\binom{n-1}{|S|}}$$

$$\varphi_i(N, v) = \sum_{S \subseteq N} p[S] m_i(S).$$

**Aggregate over all sets**

**Marginal QII of feature $i$ w.r.t. set $S$**

# *Experiments*

Predictive policing using *NLSY* data set

- Classification: History of arrest

Income prediction using a benchmark *census data set*.

- Classification: income < 50k or income >=50k?

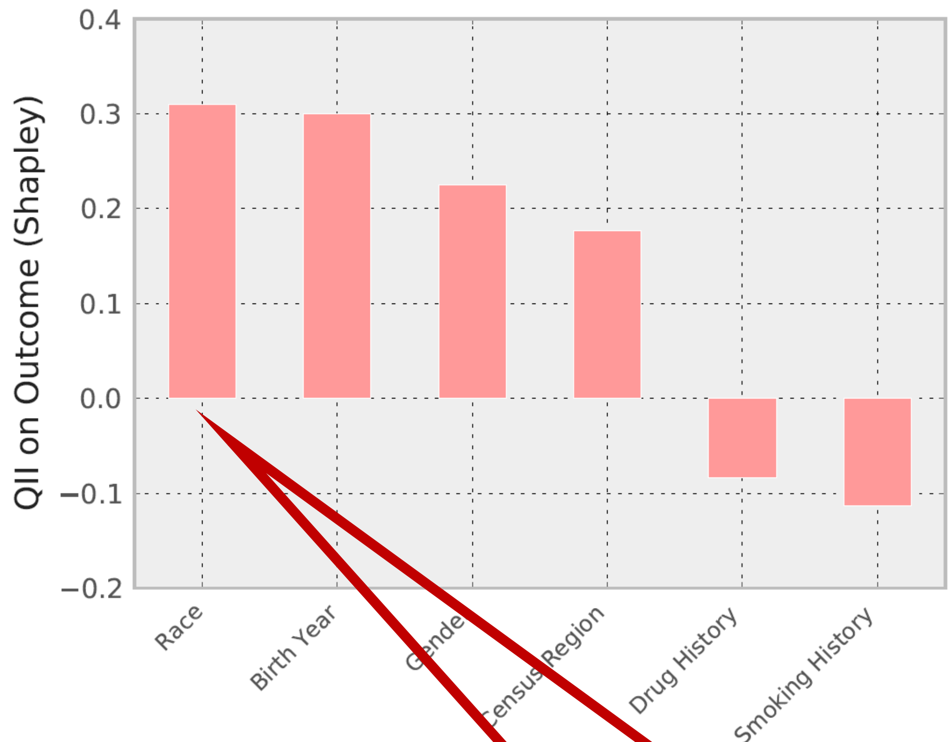
Standard machine learning algorithms

- Logistic Regression, SVM…

# *Experiments:*

*Arrest Prediction*

*Individual Transperancy*

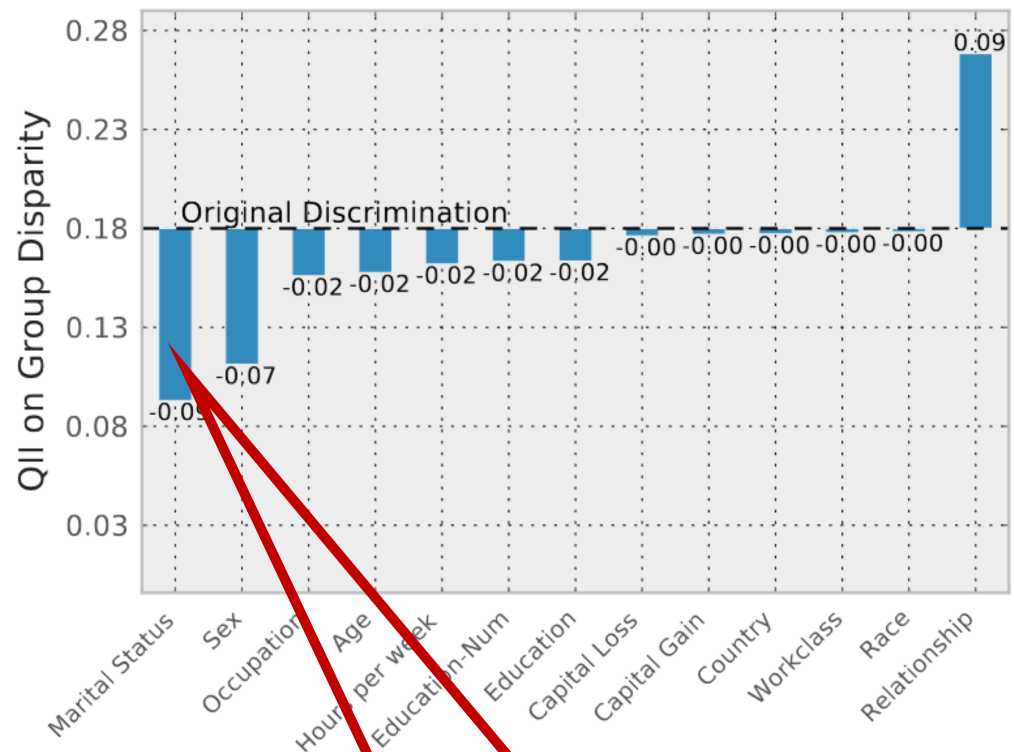



**Evidence of racial discrimination**

# *Experiments:*

*Income Prediction*

*Group Disparity*




**Marital status of females is more influential than of males**

# *Related Work:*

*Model Interpretability*

By simplicity of the model:

- *LASSO*, sparse linear models, decision trees
- Possible accuracy loss, but human interpretable

By approximation of the model:

- LIME (*Local Interpretable Model-Agnostic Explanations*)
- Can provide richer explanations
- The relation with the actual underlying model is not clear

# *Conclusion*

**<u>Causal intervention:</u>**

- Deals with correlated inputs

**<u>Quantity of interest:</u>**

- Supports a general class of transparency queries

**<u>Cooperative game:</u>**

- Computes joint and aggregate influence

**<u>Performance:</u>**

- QII measures can be approximated efficiently
- For each report: worst case <5mins, best case <1sec

Questions?