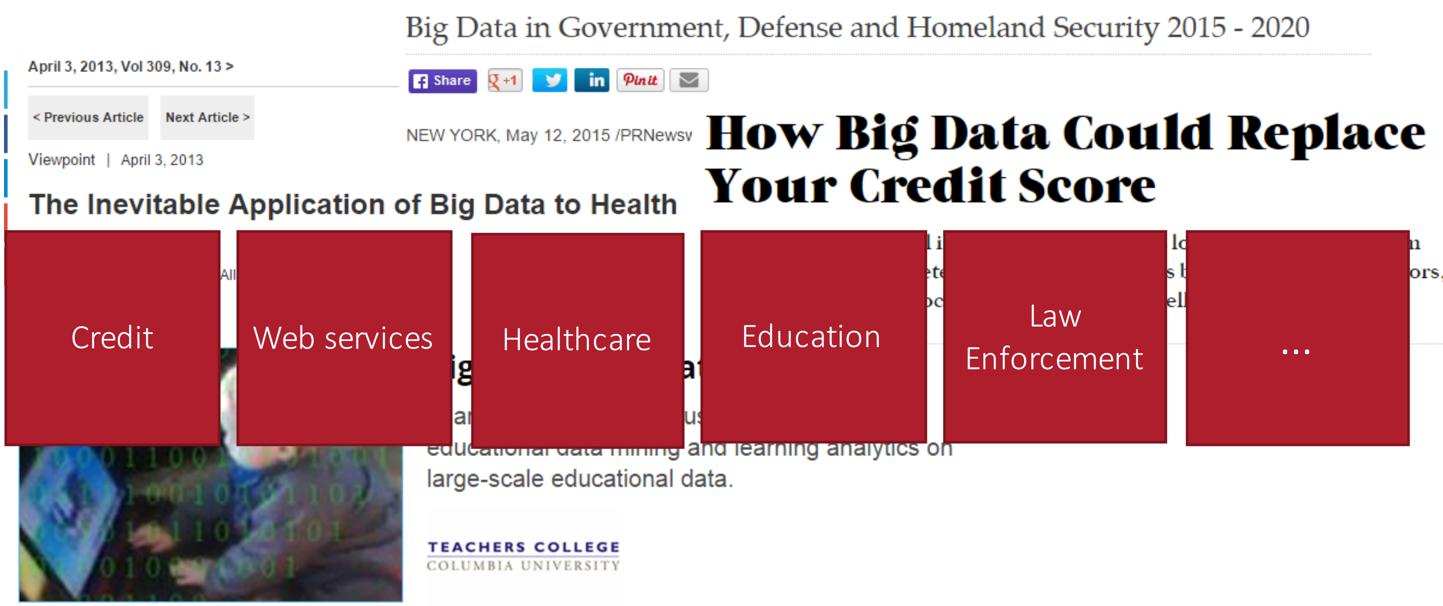
Algorithmic Accountability via Quantitative Input Influence

Anupam Datta

18734: Foundations of Privacy

Fall 2016

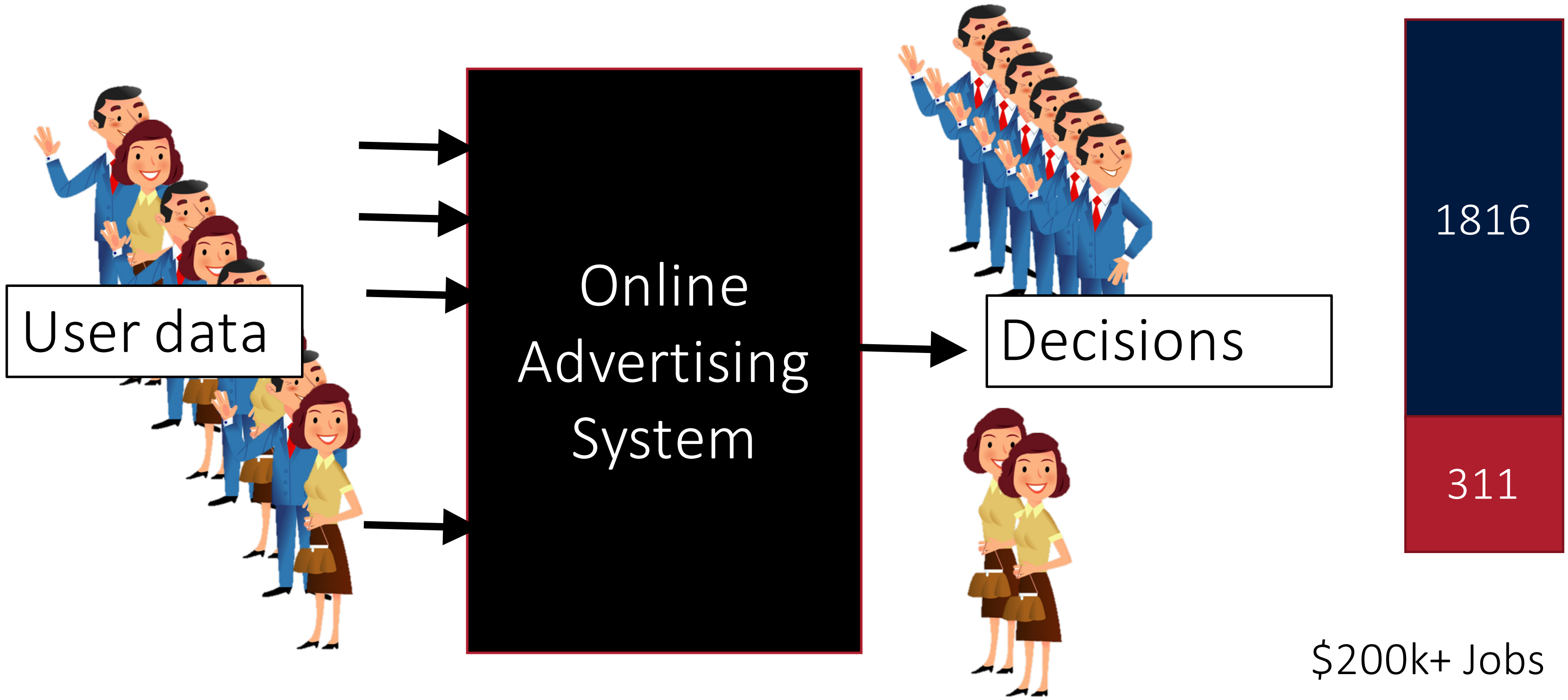
Big Data Systems are Ubiquitous



- Credit
- Web services
- Healthcare
- Education
- Law Enforcement
- …

Big Data Systems Threaten Fairness

Explicit Use [Datta, Tschantz, Datta 2015]



Big Data Systems Threaten Privacy

Proxy Use [Datta, Fredrikson, Ko, Mardziel, Sen 2016]

Using pregnancy status for marketing [Target 2012]

Pregnant?

Accountable Big Data Systems

- Oversight to detect violations and explain behaviors
- Correction to prevent future violations

Use Restrictions in Big Data Systems

Do not use a protected information type (explicit or proxy use) for certain purposes with some exceptions

- Non-discrimination:
  - Do not use race or gender for employment decisions; business necessity exceptions
- Use Privacy:
  - Do not use health information for purposes other than those of healthcare context; exceptions for law enforcement

Formalizing Explicit Use | Decisions with Explanations [Datta, Sen, Zick 2016]

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

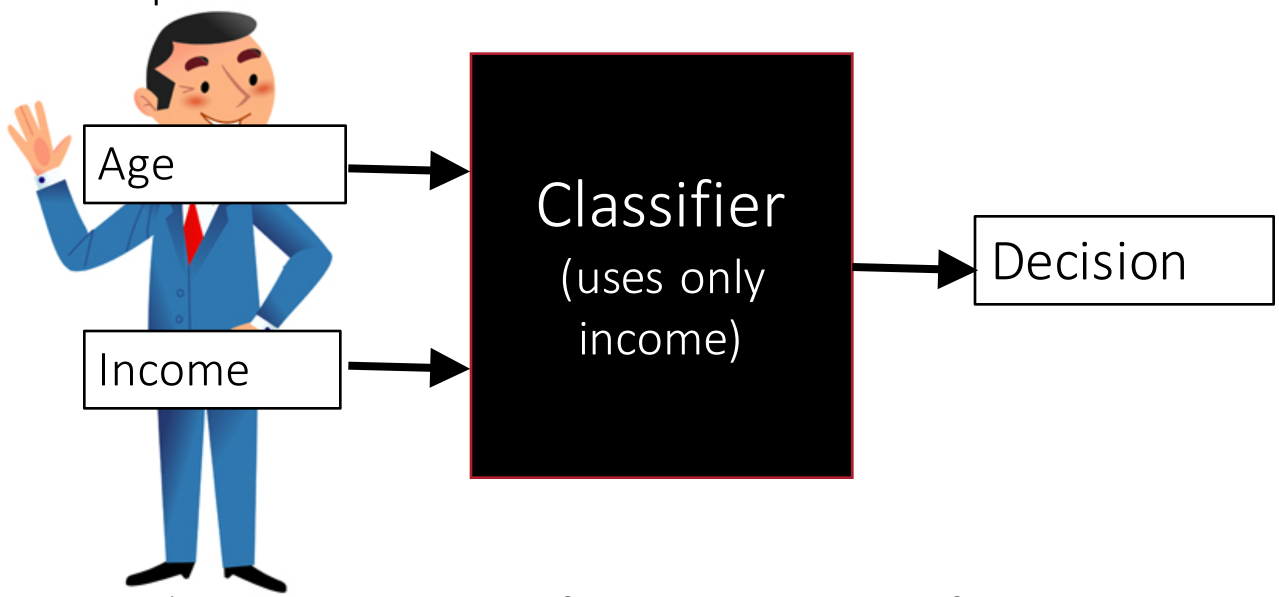
Age	27
Workclass	Private
Education	Preschool
Marital Status	Married
Occupation	Farming-Fishing
Relationship to household income	Other Relative
Race	White
Gender	Male
Capital gain	$41310
…..	

Negative Factors:
Occupation
Education Level

Positive Factors:
Capital Gain

Challenge | Correlated Inputs

Example: Credit decisions



Conclusion: Measures of association not informative

Challenge | General Class of Transparency Queries

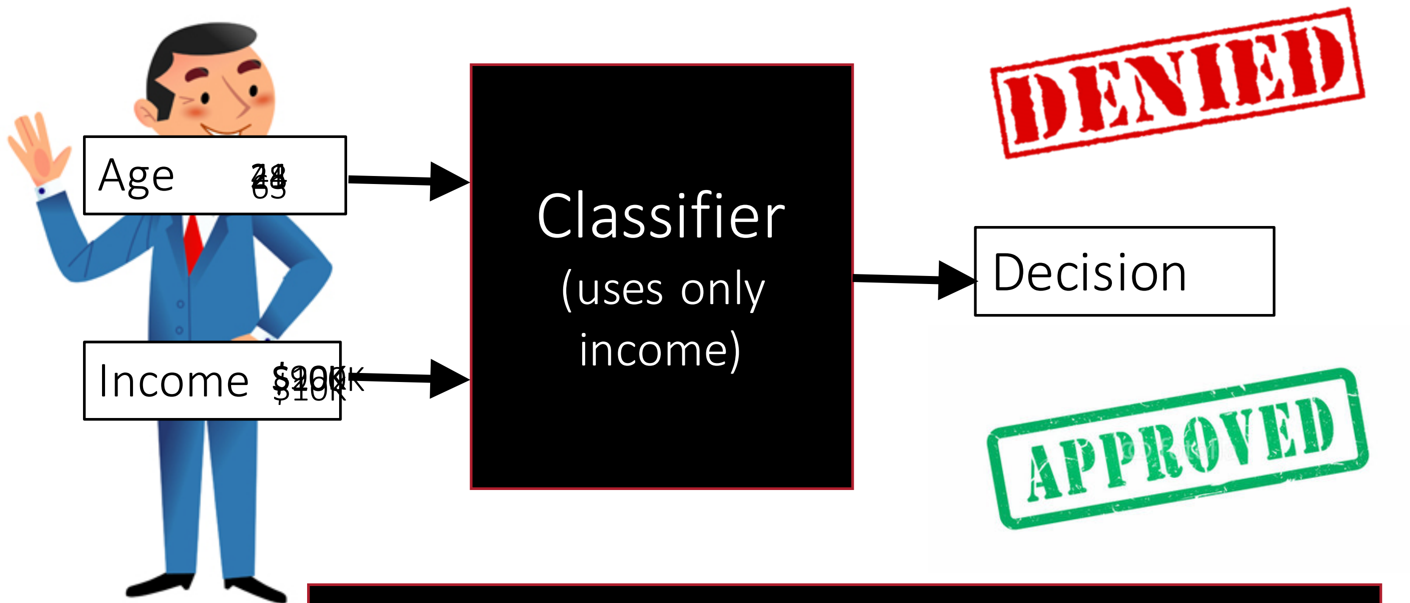
Individual	*Which input had the most influence in my credit denial?*
Group	*What inputs have the most influence on credit decisions of women?*
Disparity	*What inputs influence men getting more positive outcomes than women?*

Result | Quantitative Input Influence (QII)

A technique for measuring the influence of an input of a system on its outputs.

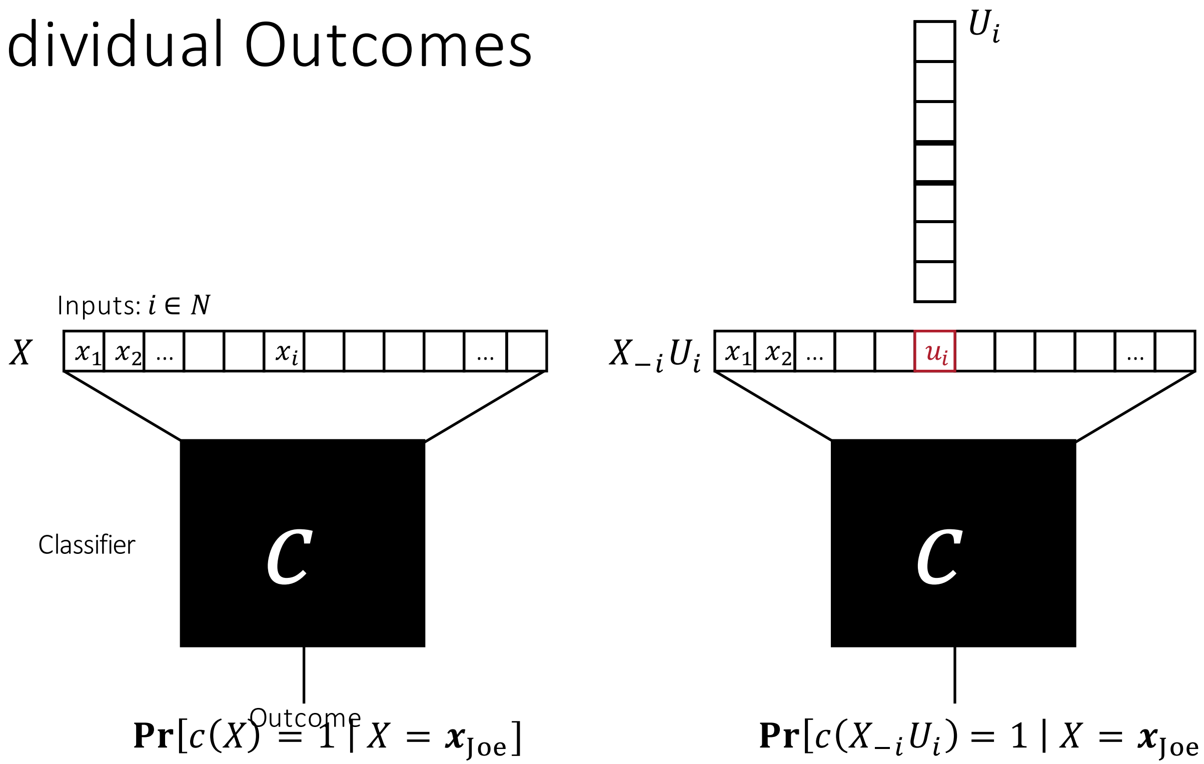
Causal Intervention	Deals with *correlated inputs*
Quantity of Interest	Supports a general class of transparency queries

Key Idea 1| Causal Intervention



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

QII for Individual Outcomes



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

Key Idea 2 | Quantity of Interest

- Various statistics of a system:
  - Classification outcome of an individual

$$\mathbf{Pr}[c(X) = c(\boldsymbol{x}_0) \mid X = \boldsymbol{x}_0] - \mathbf{Pr}[c(X_{-i}U_i) = c(\boldsymbol{x}_0) \mid X = \boldsymbol{x}_0]$$

  - Classification outcomes of a group of individuals

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

  - Disparity between classification outcomes of groups

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

QII | Definition

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

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

Result | Quantitative Input Influence (QII)

A technique for measuring the influence of an input of a system on its outputs.

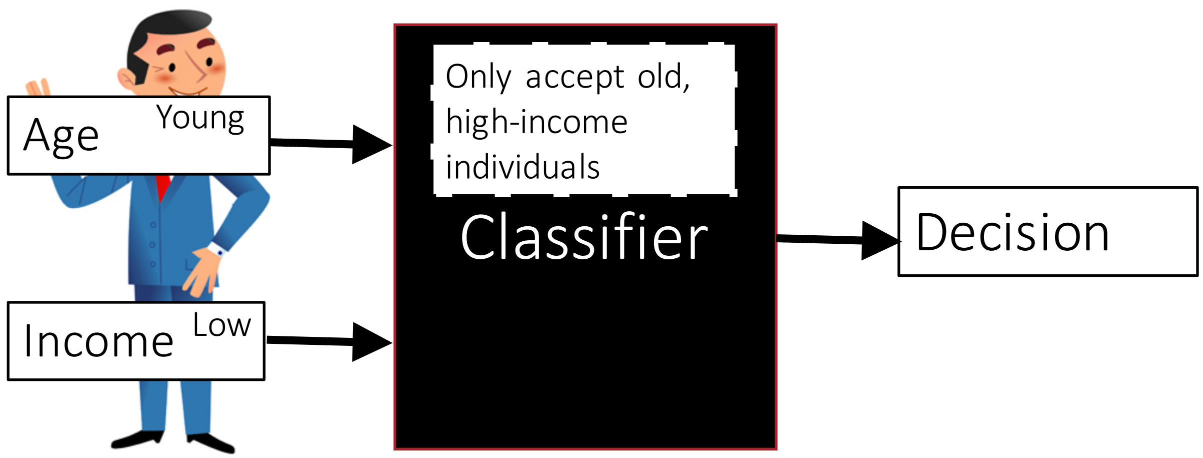
Causal Intervention	Deals with *correlated inputs*
Quantity of Interest	Supports a general class of transparency queries

Result | Quantitative Input Influence (QII)

A technique for measuring the influence of an input of a system on its outputs.

Causal Intervention	Deals with *correlated inputs*
Quantity of Interest	Supports a general class of transparency queries

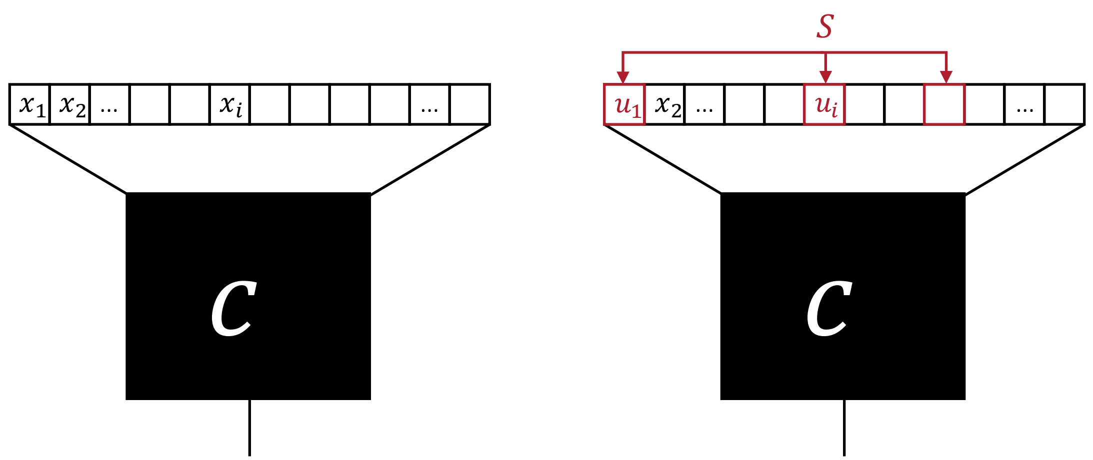
Challenge | Single Inputs have Low Influence



Naïve Approach | Set QII

Replace X_S with a independent random value from the joint distribution of inputs $S \subseteq N$.

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



Marginal QII

- Not all features are equally important within a set.
- *Marginal QII*: Influence of age and income over only income.

$$\iota(\{\text{age}, \text{income}\}) - \iota(\{\text{income}\})$$

- But age is a part of many sets!

$$\iota(\{\text{age}, \text{gender}, \text{job}\}) - \iota(\{\text{gender}, \text{job}\})$$

$$\iota(\{\text{age}\}) - \iota(\{\})$$

$$\iota(\{\text{age}, \text{gender}\}) - \iota(\{\text{gender}\})$$

$$\iota(\{\text{age}, \text{job}\}) - \iota(\{\text{job}\})$$

$$\iota(\{\text{age}, \text{gender}, \text{job}\}) - \iota(\{\text{gender}, \text{job}\})$$

$$\iota(\{\text{age}, \text{gender}, \text{income}\}) - \iota(\{\text{gender}, \text{income}\})$$

$$\iota(\{\text{age}, \text{gender}, \text{income}, \text{job}\}) - \iota(\{\text{gender}, \text{income}, \text{job}\}$$

Key Idea 3| Set QII is a Cooperative Game

- Cooperative game
  - set of agents
  - value of subsets

Voting

Revenue Sharing

Input Influence
agents → features
value → influence

Shapley Value

- [Shapley'53] For cooperative games, the only aggregation measure that satisfies symmetry, dummy, and monotonicity is:

$$\phi_i(N, v) = \sum_{\boxed{S \subseteq N \setminus \{i\}}} \frac{|S|!\,(n - |S| - 1)!}{n!} m_i(S)$$

- Need to compute sum over all subsets of features:
  - Efficient approximation by randomly sampling sets

Details | Set QII is a Cooperative Game

- Cooperative game $< N, v(S) >$
  - N: set of agents
  - $v(S)$: value of set S
- Examples of cooperative games:
  - Voting: $v(S)$ = does motion pass if voters in S vote yes?
  - Revenue sharing: $v(S)$ = revenue earned by agents in S
- Our setting: $v(S)$ is the Set QII $\iota(S)$

- Define a value $\phi_i(v)$, measuring importance of i in game v.
  - By aggregating marginal contributions: $m_i(S) = v(S \cup \{i\}) - v(S)$

Details| Set QII is a Cooperative Game

- Marginal contribution: $m_i(S) = v(S \cup \{i\}) - v(S)$
- Axioms of influence:
- *Symmetry*:
  - For all i, j and $S \subseteq N \setminus \{i, j\}$, $v(S \cup \{i\}) = v(S \cup \{j\})$, implies $\phi_i(v) = \phi_j(v)$.
- *Dummy*
  - For all $i, S \subseteq N$, $v(S \cup \{i\}) = v(S)$, implies $\phi_i(v) = 0$.
- *Monotonicity*
  - For two games v_1, v_2, if for all S, $m_i(S, v_1) \geq m_i(S, v_2)$, then $\phi_i(v_1) \geq \phi_i(v_2)$.

- Only the Shapley value satisfies all three:

$$\phi_i(N, v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} m_i(S)$$

Experiments | Test Applications

arrests

Predictive policing using the National Longitudinal Survey of Youth (NLSY)

- Features: Age, Gender, Race, Location, Smoking History, Drug History
- Classification: History of Arrests
- ~8,000 individuals

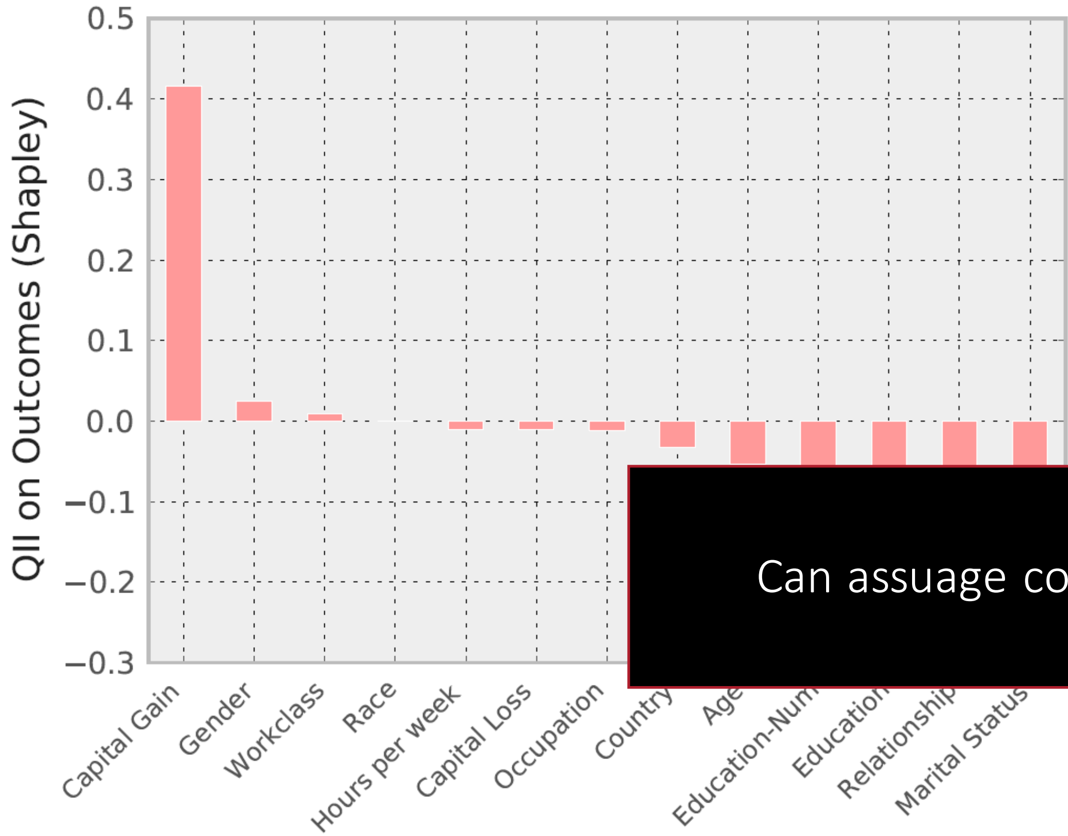
income

Income prediction using a benchmark census dataset

- Features: Age, Gender, Relationship, Education, Capital Gains, Ethnicity
- Classification: Income >= 50K
- ~30,000 individuals

Implemented with Logistic Regression, Kernel SVM, Decision Trees, Decision Forest

Personalized Explanation | Mr X

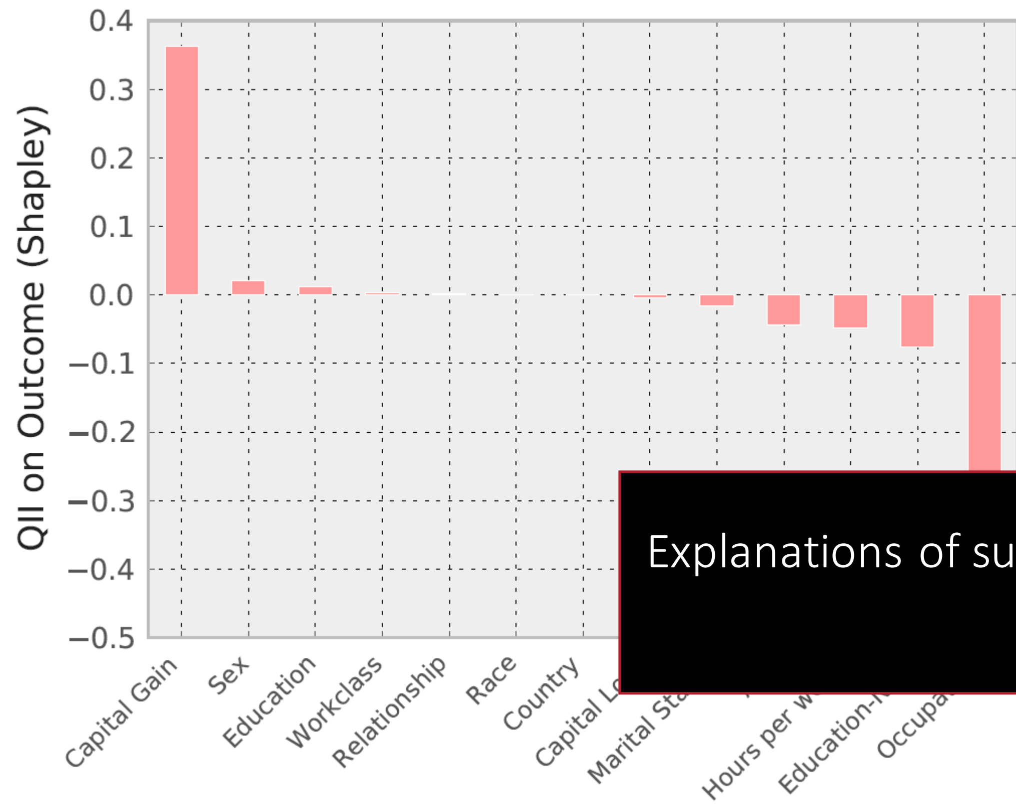


Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344

Can assuage concerns of discrimination.

income

Personalized Explanation | Mr Y



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

Explanations of superficially similar people can be different.

income

Result | Quantitative Input Influence (QII)

A technique for measuring the influence of an input of a system on its outputs.

Causal Intervention	Deals with *correlated inputs*
Quantity of Interest	Supports a general class of transparency queries
Cooperative Game	Computes joint and marginal influence
Performance	QII measures can be approximated efficiently

Related Work: QII

- Randomized Causal Intervention
  - Feature Selection: Permutation Importance [Breiman 2001]
  - Importance of Causal Relations [Janzing et al. 2013]
  - *Do not consider marginal influence or general quantities of interest*

- Associative Measures
  - Quantitative Information Flow: Appropriate for secrecy
  - FairTest [Tramèr et al. 2015]
  - *Correlated inputs hide causality*

- Interpretability-by-design
  - Regularization for simplicity (Lasso)
  - Bayesian Rule Lists [Letham et al. 2015]
  - *Potential loss in accuracy*

Related Work: Accountability for Use Restrictions

- Accounting for proxies and their causal use is missing
  - Usage control in computer security, Sandhu and Park 2002
  - Information accountability, Weitzner et al. 2008
  - Audit algorithms for privacy policies, Garg, Jia, Datta 2011
  - Enforcing purpose restrictions in privacy policies, Tschantz, Datta, Wing 2012
  - Privacy compliance of big data systems, Sen, Guha, Datta, Rajamani, Tsai, Wing 2014

- Fairness in big data systems
  - Group fairness [Feldman+ 2015]: detection and repair of disparate impact; does not account for proxy usage in general
  - Individual fairness [Dwork et al. 2011]: focus on correctness by construction not accountability

Use Restrictions in Big Data Systems

Do not use a protected information type (explicit or proxy use) for certain purposes with some exceptions

Accountable Big Data Systems

- Oversight to detect violations and explain behaviors
- Correction to prevent future violations

- Usage Privacy:
  - Do not use health information for purposes other than those of healthcare context; exceptions for law enforcement

Thanks!