Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation

Han Zhao[1*] Jianfeng Chi[2*] Yuan Tian[2] Geoffrey J. Gordon[1]

Presented by: Jianfeng Chi

[1]Machine Learning Department, Carnegie Mellon University

[2]Department of Computer Science , University of Virginia

jc6ub@virginia.edu

$*$ Equal Contribution

Motivation

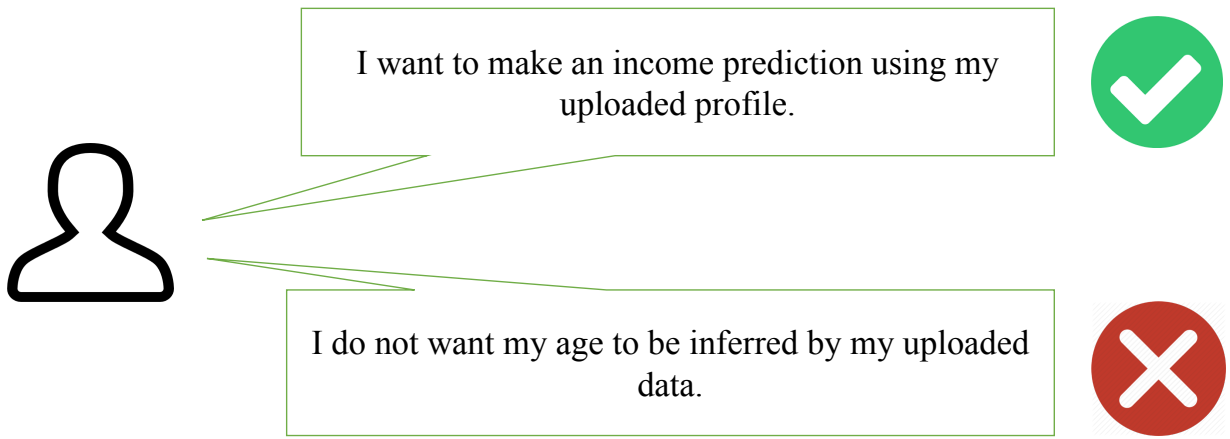
- Growing need for data sharing and collection;

Motivation

- Growing need for data sharing and collection;
- In some applications,
  - the data owner agrees on the data usage for the target task
  - while he/she does not want his/her other sensitive information (e.g., age) to be leaked.
  - Simply removing age attribute from the shared data is not enough.

Motivation

- Growing need for data sharing and collection;
- In some applications,
  - the data owner agrees on the data usage for the target task
  - while he/she does not want his/her other sensitive information (e.g., age) to be leaked.
  - Simply removing age attribute from the shared data is not enough.



Research Questions

- Can we minimize the information leakage of the sensitive attribute while still maximizing the task accuracy?

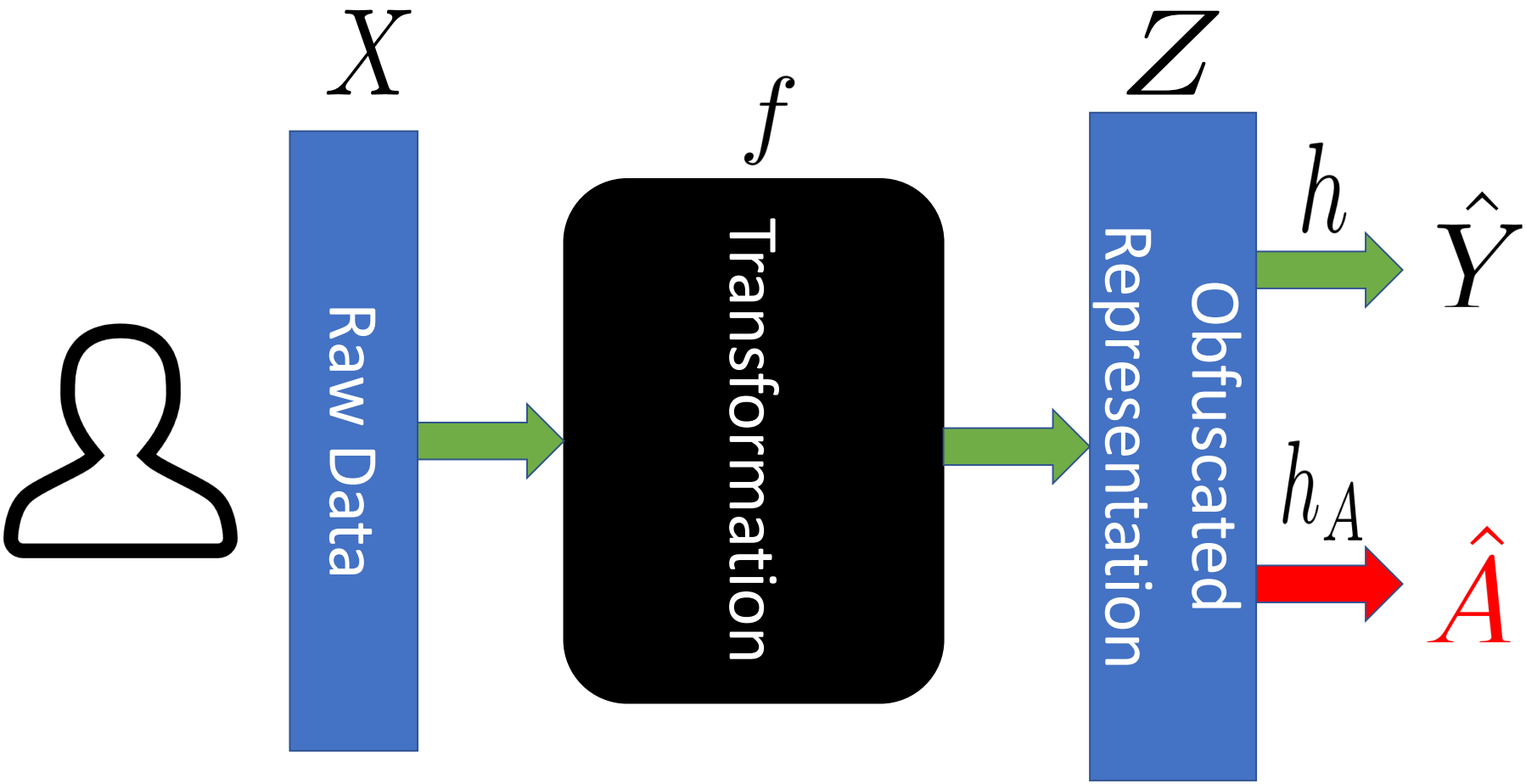
Research Questions

- Can we minimize the information leakage of the sensitive attribute while still maximizing the task accuracy?
- Is there any formal guarantee for attribute obfuscation against arbitrary adversary?

Research Questions

- Can we minimize the information leakage of the sensitive attribute while still maximizing the task accuracy?
- Is there any formal guarantee for attribute obfuscation against arbitrary adversary?
- What is the fundamental trade-off between attribute obfuscation and accuracy maximization?

Problem Setup



Main Results

- We provide a **formal guarantee** for attribute obfuscation by proving an information-theoretic lower bound on the inference error of the protected attribute under attacks from any adversaries.

$$\underbrace{\Pr_{\mathcal{D}^{f^*}}(\widehat{A} \neq A)}_{\text{Attribute Inference Error}} \geq H^*/2\lg(6/H^*) \quad \text{where } H^* := \underbrace{H(A \mid Z = f^*(X))}_{\text{Conditional Entropy}}$$

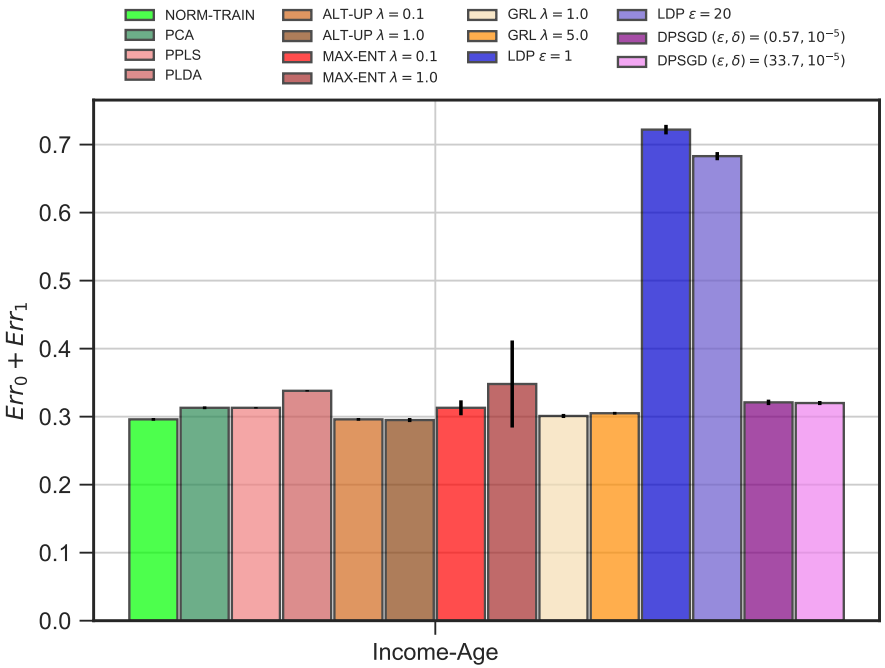
Main Results (Cont.)

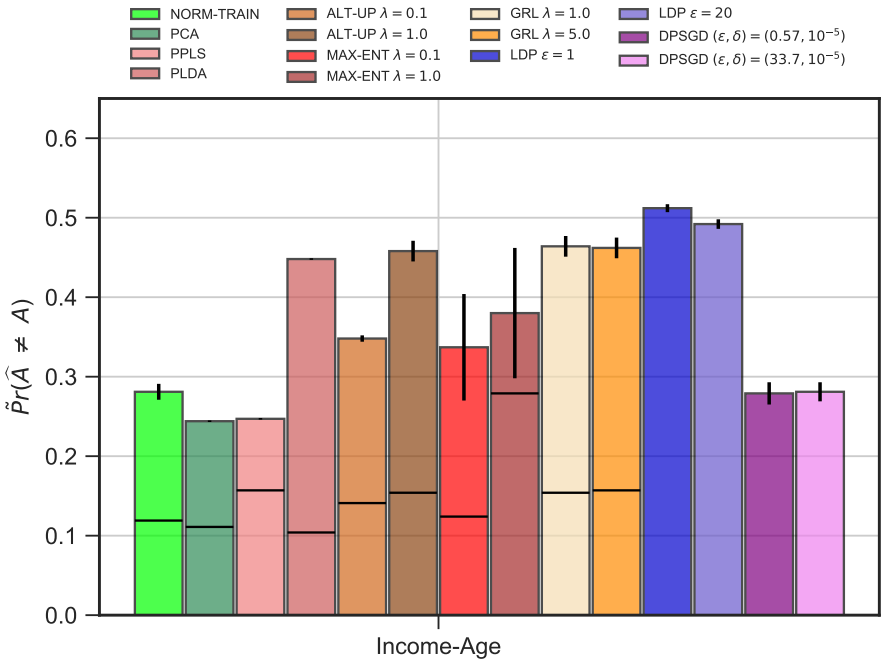
- We prove a theorem that formally characterizes the fundamental **trade-off** between attribute obfuscation and accuracy maximization.

$$\underbrace{\mathrm{Err}_0(h \circ f) + \mathrm{Err}_1(h \circ f)}_{\text{Task Joint Errors}} \geq \frac{1}{2}\Bigg(\Big(\underbrace{d_{\mathrm{JS}}(\mathcal{D}_0^Y, \mathcal{D}_1^Y)}_{\text{Distribution Discrepancy}} - \underbrace{\sqrt{\mathrm{ADV}(\mathcal{H}_A \circ f)}}_{\text{Adversarial Advantage}} \Big)_+ \Bigg)^2$$

Empirical Evaluation

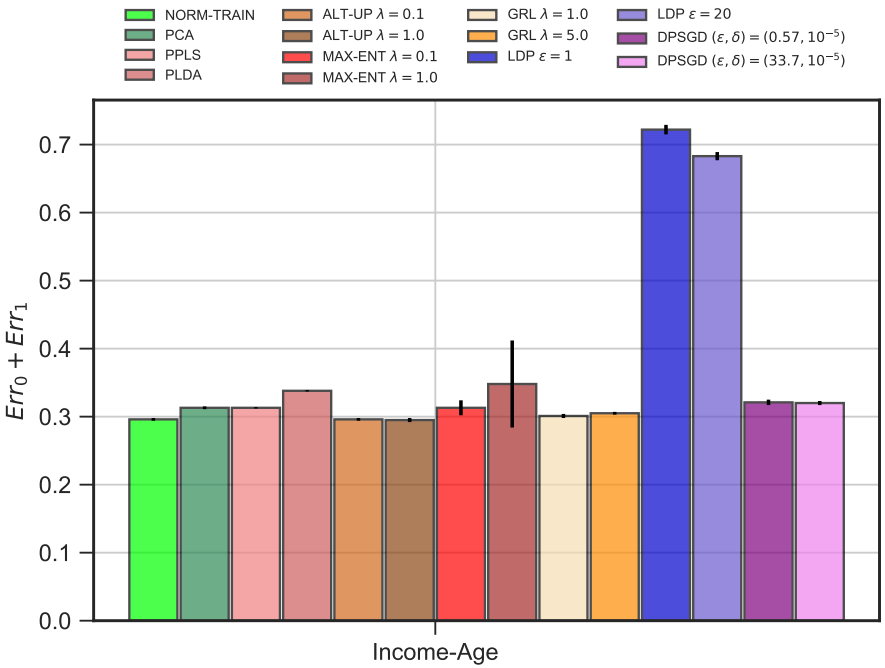
- We empirically validate our theoretical results and suggest that adversarial representation learning methods achieve the best trade-offs.

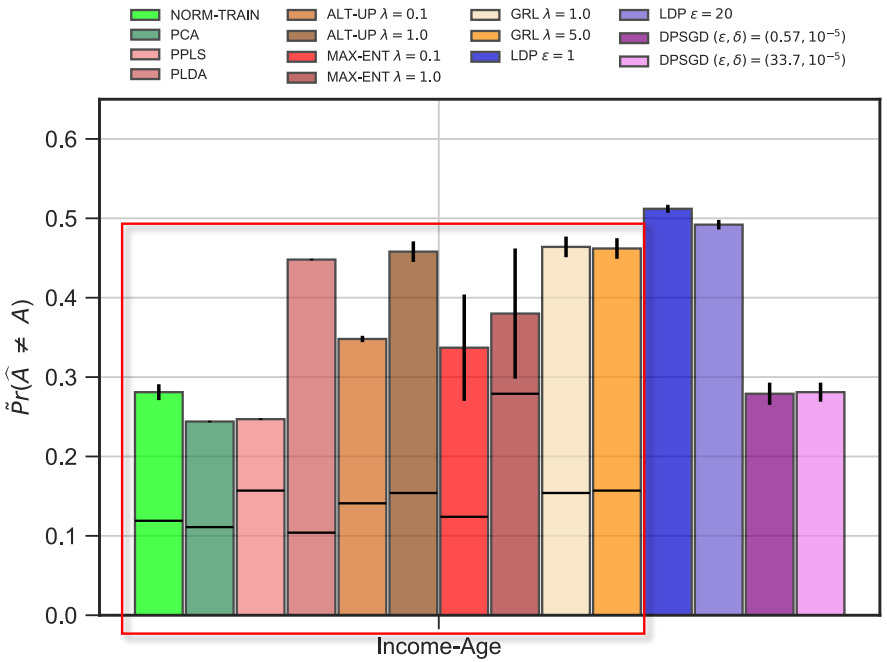




Empirical Evaluation

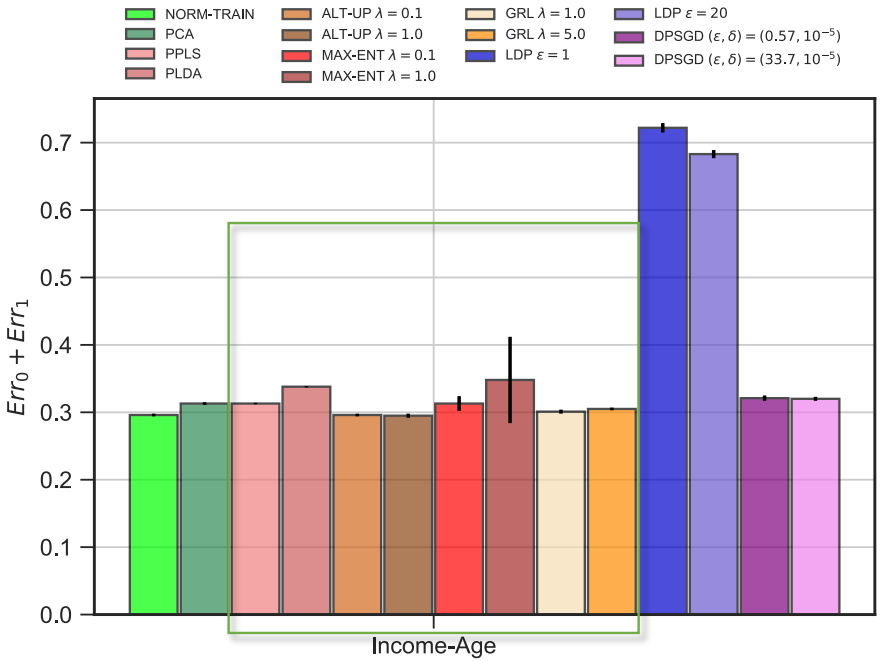
- We empirically validate our theoretical results and suggest that adversarial representation learning methods achieve the best trade-offs.

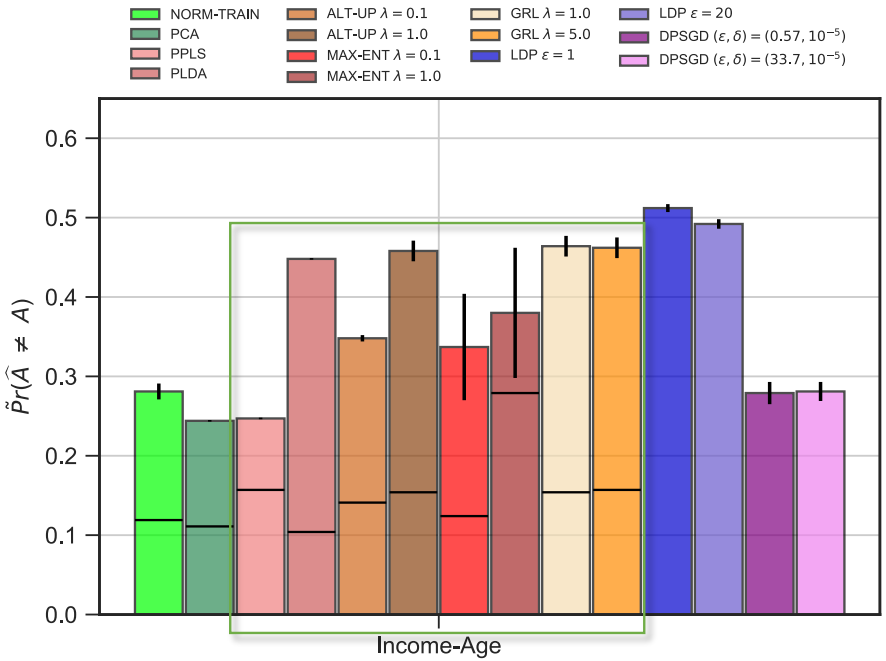




Empirical Evaluation

- We empirically validate our theoretical results and suggest that adversarial representation learning methods achieve the best trade-offs.





See you in NeurIPS 2020!