Adversarially Learned Representations for Information Obfuscation and Inference

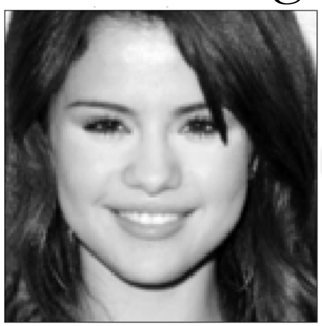
Martin Bertran[1], Natalia Martinez[1], Afroditi Papadaki[2]
Qiang Qiu[1], Miguel Rodrigues[2], Galen Reeves[1], Guillermo Sapiro[1]

1. Duke University
2. University College London

Motivation

Why do users share their data?

Shared data
Facial Image



User
decision

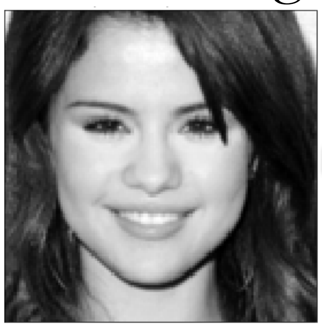
Service
Provider

Utility ✓
Subject verification

Motivation

Why do users share their data?

Shared data
Facial Image



User decision

Service Provider

Utility ✓
Subject verification

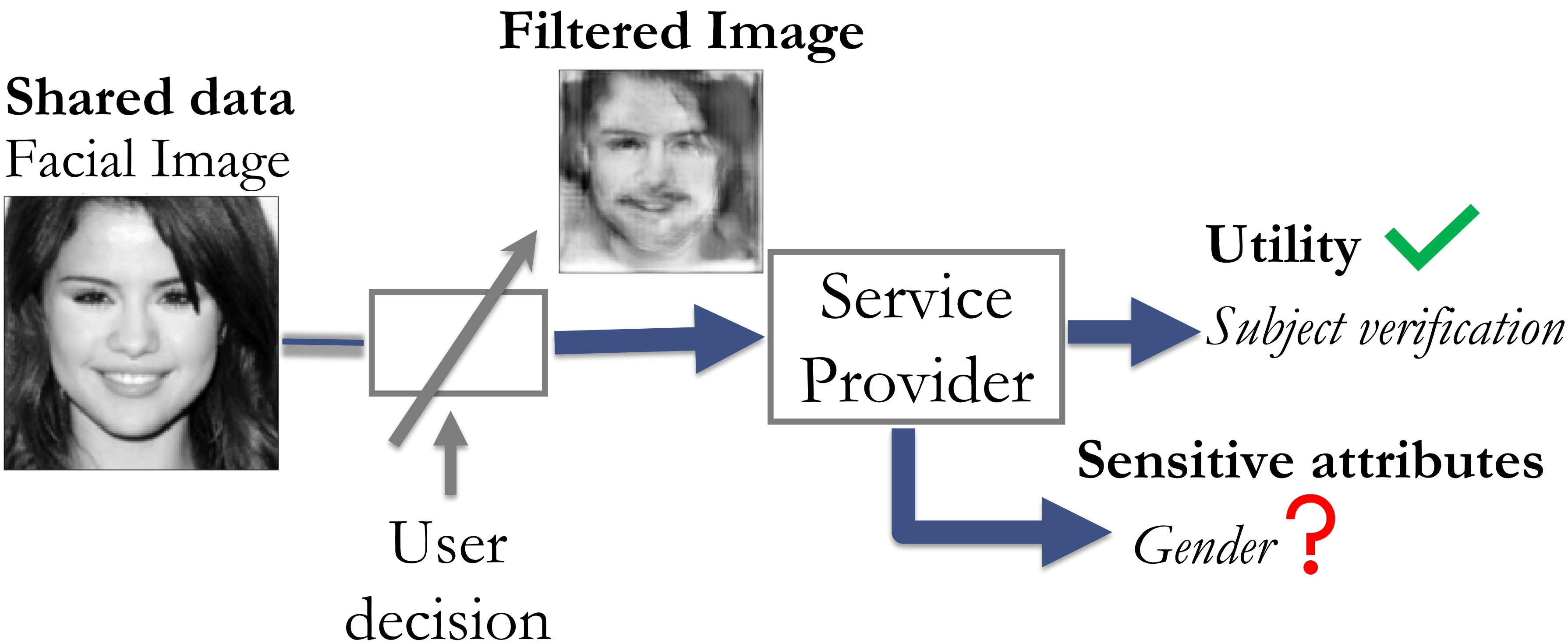
Sensitive attributes ✗
Emotion
Gender
Race

Motivation

Can we do better?

Motivation

Can we do better?



Learn **space-preserving** representations that **obfuscate** sensitive information while **preserving** utility.

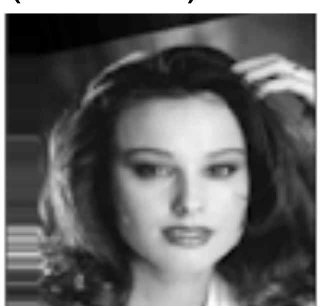
Motivation

Example: Preserve gender & obfuscate emotion

Original

P(Male) = 0.98
P(Smile) = 0.78

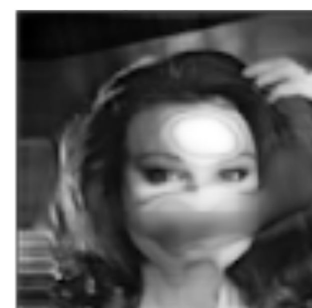
P(Female) = 0.99
P(Serious) = 0.98



Filtered

P(Male) = 0.98 =
P(Smile) = 0.38 ↓

P(Female) = 0.99 =
P(Serious) = 0.31 ↓



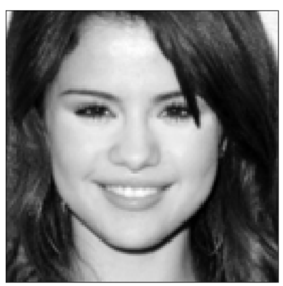
Motivation

Example: Preserve subject & obfuscate gender

Original

P(Male) = 0.99
Subject verified

P(Female) = 0.99
Subject verified



Filtered

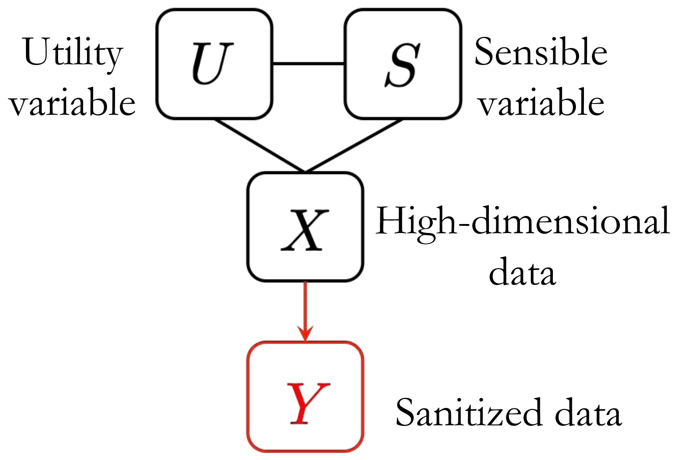
P(Male) = 0.70
Subject verified

P(Female) = 0.54
Subject verified

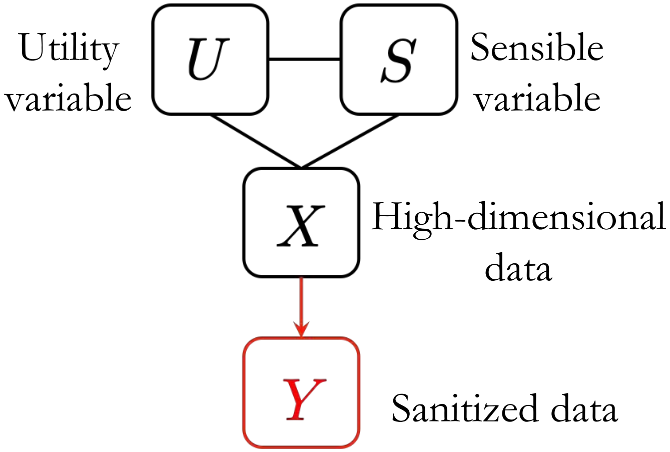
Sample of related work

- (2003) Chechik et al. Extracting relevant structures with side information.
- (2016) Basciftci et al. On privacy-utility tradeoffs for constrained data release mechanisms.
- (2018) Madras et al. Learning adversarially fair and transferable representations.
- (2018) Sun et al. A hybrid model for identity obfuscation by face replacement.

Problem formulation

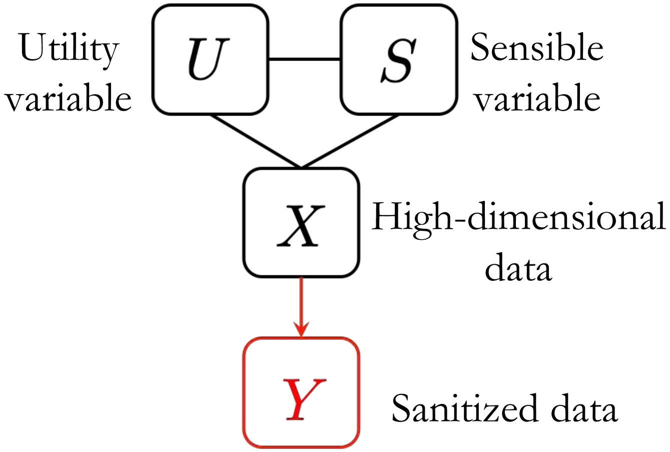


Problem formulation



$$(U, S) \sim p(U, S)$$

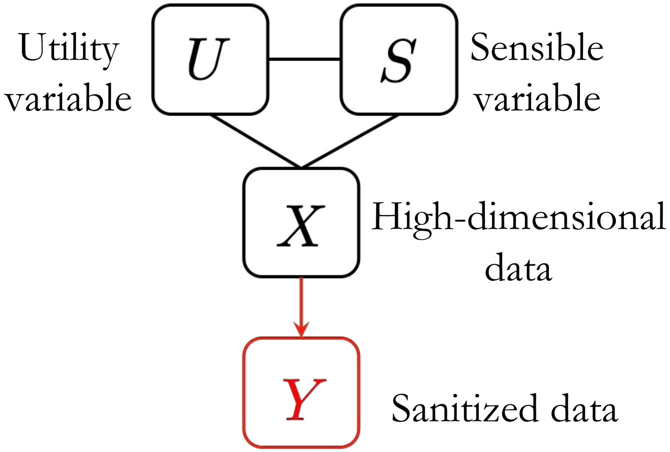
Problem formulation



$$(U, S) \sim p(U, S)$$

$$X \sim p(X|U, S)$$

Problem formulation



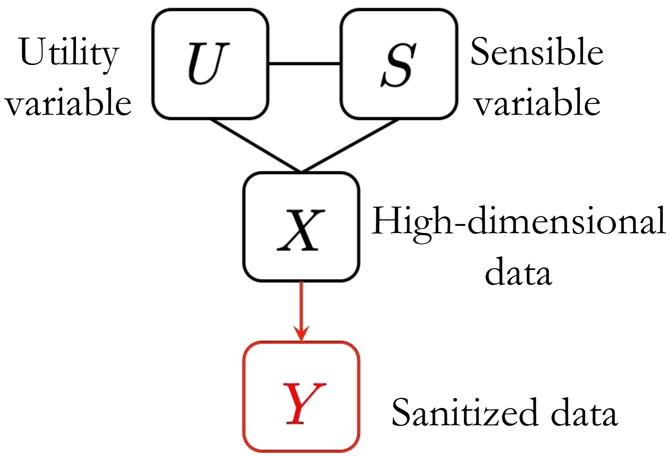
$$(U, S) \sim p(U, S)$$

$$X \sim p(X|U, S)$$

$$Y \sim p(Y|X)$$

Our objective!

Problem formulation



$$(U, S) \sim p(U, S)$$

$$X \sim p(X|U, S)$$

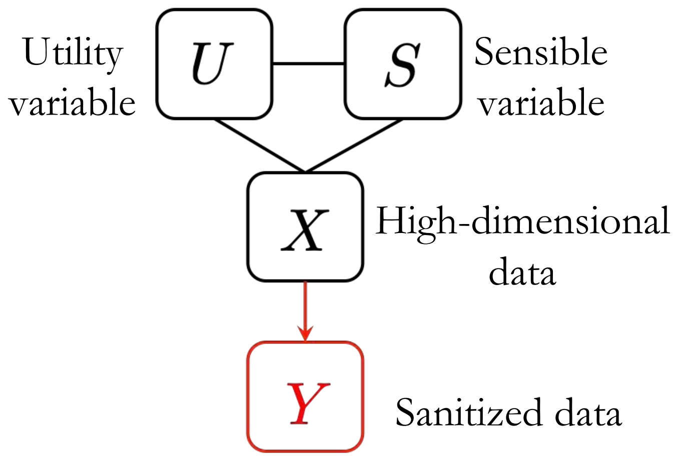
$$Y \sim p(Y|X)$$

Our objective!

Want to learn $Y \sim p(Y|X)$ such that :

- $p(S|Y) \sim p(S)$
- $p(U|Y) \sim p(U|X)$

Problem formulation



$(U, S) \sim p(U, S)$

$X \sim p(X|U, S)$

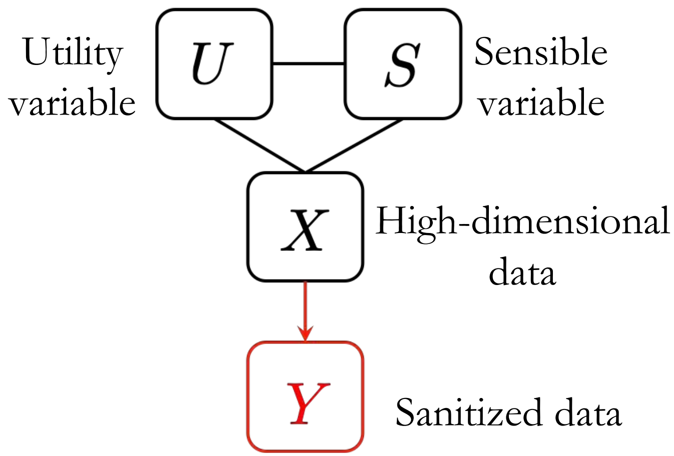
$Y \sim p(Y|X)$

Our objective!

Want to learn $Y \sim p(Y|X)$ such that :

- $p(S|Y) \sim p(S) \longrightarrow min\, D_{KL}[p(S|Y)||p(S)]$
- $p(U|Y) \sim p(U|X)$

Problem formulation



$$(U, S) \sim p(U, S)$$

$$X \sim p(X|U, S)$$

$$Y \sim p(Y|X)$$

Our objective!

Want to learn $Y \sim p(Y|X)$ such that :

- $p(S|Y) \sim p(S) \longrightarrow min\, D_{KL}[p(S|Y)||p(S)]$
- $p(U|Y) \sim p(U|X) \longrightarrow min\, D_{KL}[p(U|X)||p(U|Y)]$

Problem formulation

Want to learn $Y \sim p(Y|X)$ such that:

- $min\, D_{KL}[p(S|Y)||p(S)]$
- $min\, D_{KL}[p(U|X)||p(U|Y)]$

Problem formulation

Want to learn $Y \sim p(Y|X)$ such that:

- $min\, D_{KL}[p(S|Y)||p(S)]$ $\xrightarrow{E_Y[.]}$ $I(S;Y)$
- $min\, D_{KL}[p(U|X)||p(U|Y)]$

Problem formulation

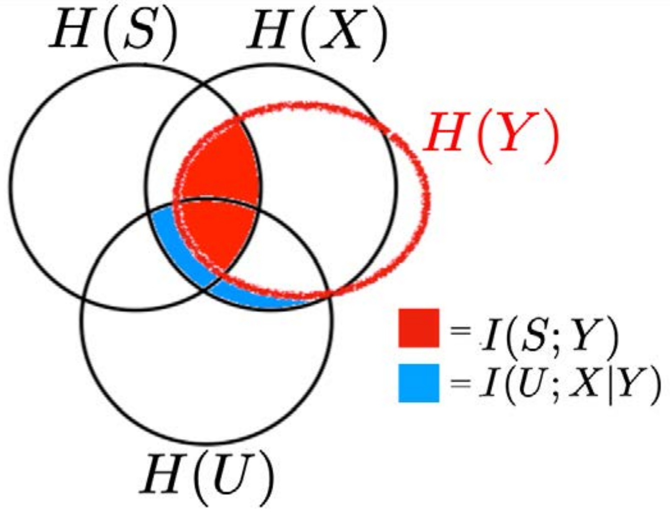
Want to learn $Y \sim p(Y|X)$ such that:

- $min\, D_{KL}[p(S|Y)||p(S)] \xrightarrow{E_Y[.]} I(S;Y)$
- $min\, D_{KL}[p(U|X)||p(U|Y)] \xrightarrow{E_{X,Y}[.]} I(U;X|Y)$

Problem formulation

Want to learn $Y \sim p(Y|X)$ such that:

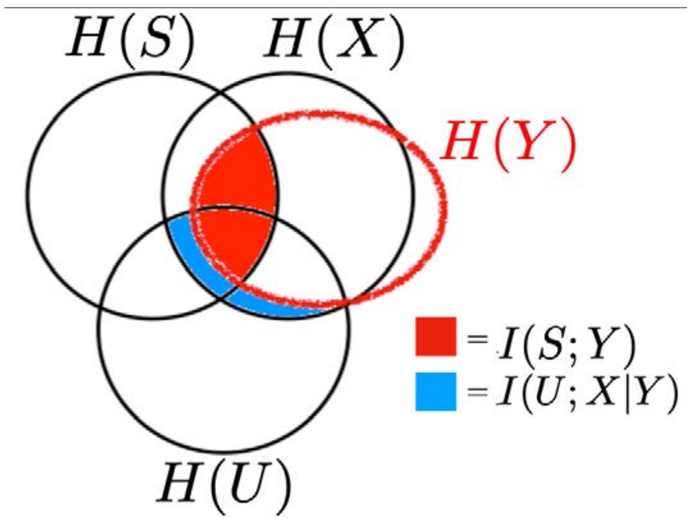
- $min\, D_{KL}[p(S|Y)||p(S)] \xrightarrow{E_Y[.]} I(S;Y)$
- $min\, D_{KL}[p(U|X)||p(U|Y)] \xrightarrow{E_{X,Y}[.]} I(U;X|Y)$



Problem formulation

Want to learn $Y \sim p(Y|X)$ such that:

- $min\, D_{KL}[p(S|Y)||p(S)] \xrightarrow{E_Y[.]} I(S;Y)$
- $min\, D_{KL}[p(U|X)||p(U|Y)] \xrightarrow{E_{X,Y}[.]} I(U;X|Y)$



Objective:

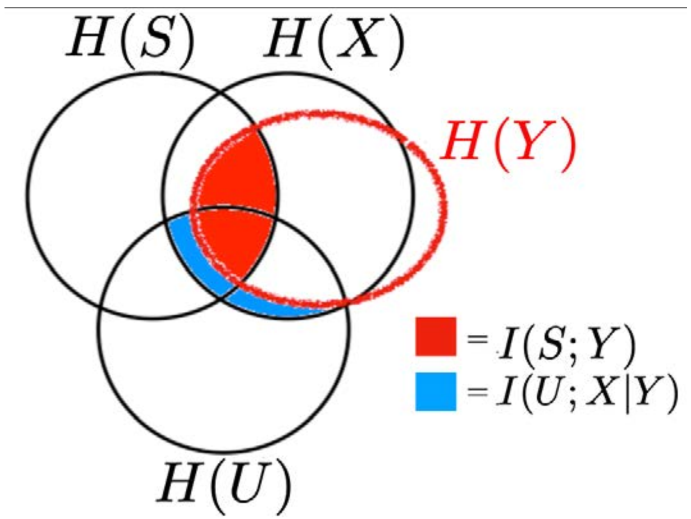
$$\min_{p(Y|X)} I(U;X|Y)$$

$$s.t. \quad I(S;Y) \le k$$

Problem formulation

Want to learn $Y \sim p(Y|X)$ such that:

- $min\ D_{KL}[p(S|Y)||p(S)] \xrightarrow{E_Y[.]} I(S;Y)$
- $min\ D_{KL}[p(U|X)||p(U|Y)] \xrightarrow{E_{X,Y}[.]} I(U;X|Y)$



Objective:

$$\min_{p(Y|X)} I(U;X|Y) \sim \max_{p(Y|X)} I(U;Y)$$

$$s.t. \quad I(S;Y) \leq k$$

Performance bounds

Given the objective $\min_{p(Y|X)} I(U;X|Y) \;\; s.t. \quad I(S;Y) \leq k$

Performance bounds

Given the objective $\min_{p(Y|X)} I(U;X|Y) \;\; s.t. \quad I(S;Y) \leq k$

What are the intrinsic limits on the trade-offs for this problem?

Performance bounds

Given the objective $\min\limits_{p(Y|X)} I(U;X|Y) \;\; s.t. \quad I(S;Y) \leq k$

What are the intrinsic limits on the trade-offs for this problem?

Lemma 1.
$(U,S) \in \mathcal{U} \times \mathcal{S}$ finite alphabets, $X \sim p(X|U,S)$.
Then:

$$\begin{aligned} &\min_{p(Y|X)} I(U;X|Y) && \geq && \min_{p(Y|U,S)} I(U;X) - I(U;Y) \\ &s.t. \quad I(S;Y) \leq k && && s.t. \quad I(S;Y) \leq k \\ & && && \qquad\; I(U;Y) \leq I(U;X) \end{aligned}$$

- With $|\mathcal{Y}|$ finite we can compute a sequence of upper bounds: Restricted cardinality sequence (RCS).

Performance bounds

Given the objective $\min\limits_{p(Y|X)} I(U;X|Y) \;\; s.t. \quad I(S;Y) \leq k$

What are the intrinsic limits on the trade-offs for this problem?

Lemma 2. Given $(X,U,S) \sim p(X,U,S)$

$$I(U;X|Y) \geq -I(S;Y) + I(U;S) - I(U;S|X)$$

Performance bounds

Given the objective $\min_{p(Y|X)} I(U;X|Y) \;\; s.t. \quad I(S;Y) \leq k$

What are the intrinsic limits on the trade-offs for this problem?

Lemma 2. Given $(X,U,S) \sim p(X,U,S)$

$$I(U;X|Y) \geq -I(S;Y) + I(U;S) - I(U;S|X)$$

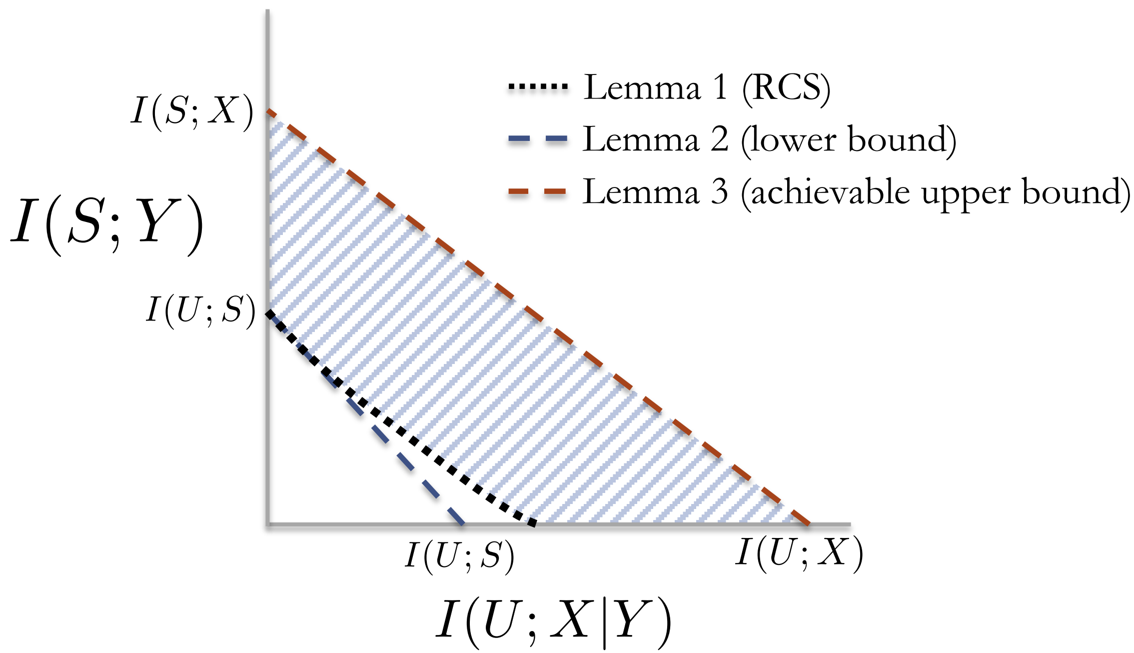
Lemma 3. Given $(X,U,S) \sim p(X,U,S)$, $\forall\, k \geq 0 \;\exists\, p(Y|X)$ such that:

$$I(S;Y) \leq k$$

$$I(U;X|Y) = max(0, 1 - \frac{k}{I(S;X)})I(U;X)$$

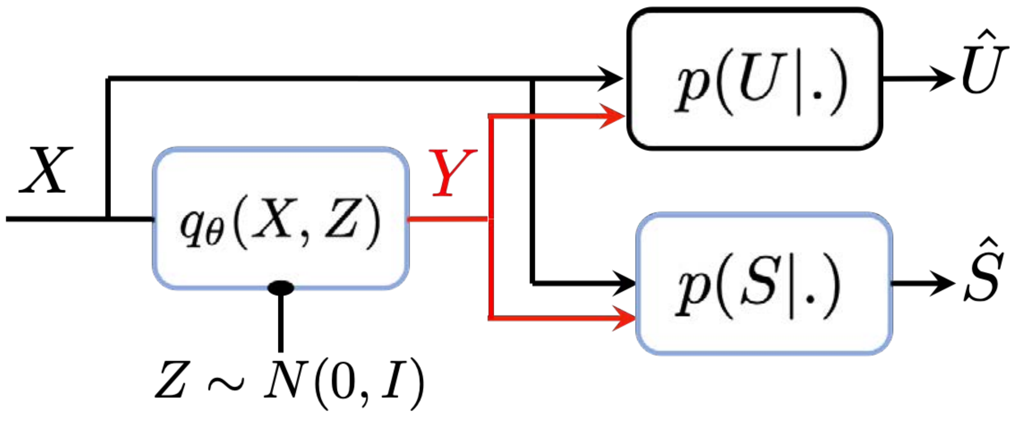
Performance bounds

Lemmas 1, 2 and 3 can be approximated using contingency tables.

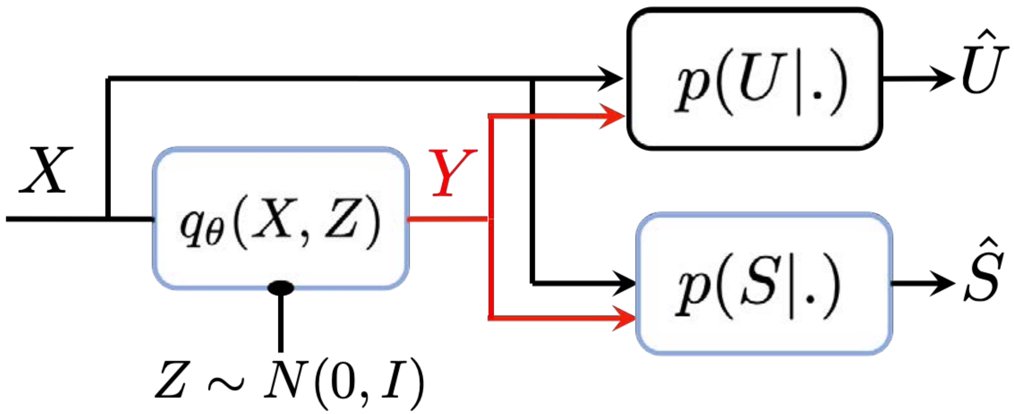


* Sketch under the assumption that $I(U;S|X) = 0$

Proposed framework



Proposed framework

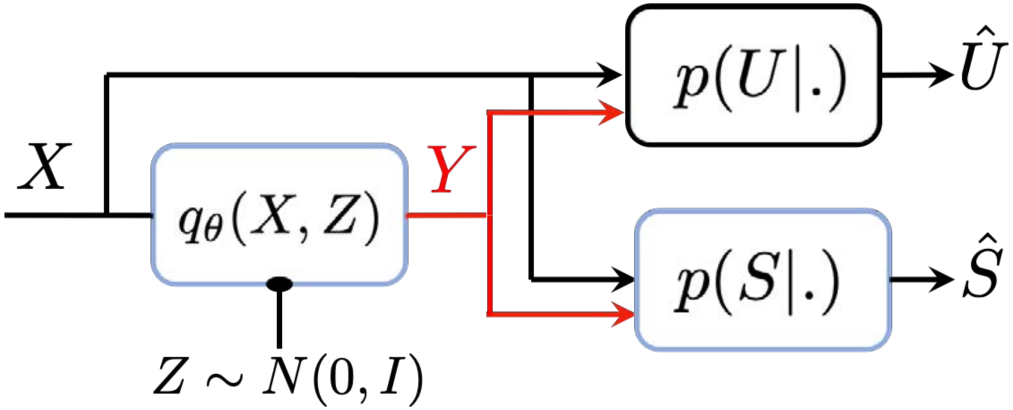


Objective:

$$\min_{p(Y|X) \sim q_\theta(X,Z)} I(U; X|Y)$$

$$s.t.: \; I(S; Y) \leq k$$

Proposed framework



Objective:

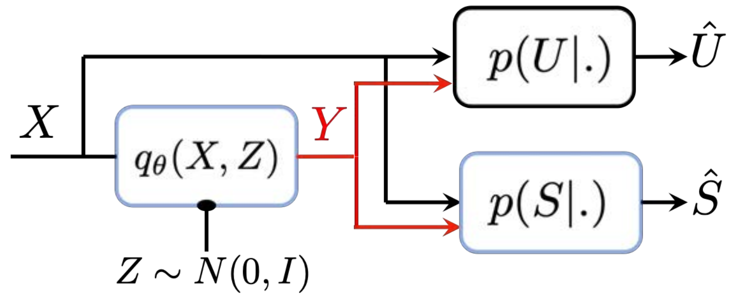
$$\min_{p(Y|X)\sim q_\theta(X,Z)} \; I(U;X|Y)$$

$$s.t.: \; I(S;Y) \le k$$

Optimization objective:

$$\min_{p(Y|X)\sim q_\theta(X,Z)} \; [I(U;X|Y) + \lambda max\{I(S;Y)-k, 0\}^2]$$

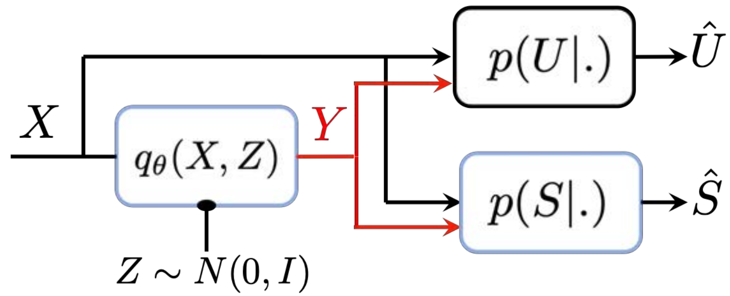
Implementation



Optimization objective:

$$\min_{q_\theta(X,Z)} \; [I(U;X|Y) + \lambda max\{I(S;Y) - k, 0\}^2]$$

Implementation



Optimization objective:

$$\min_{q_\theta(X,Z)} \left[I(U;X|Y) + \lambda max\{I(S;Y) - k, 0\}^2 \right]$$

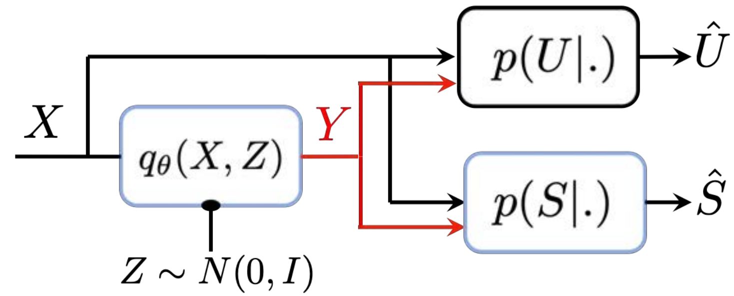
Learning the stochastic mapping $Y = q_\theta(X,Z)$:

$$p(U|X) \sim p_\phi(U|X) \longrightarrow \hat{\phi} = \text{argmin}_\phi E_{X,U}\left[-log(p_\phi(U \mid X) \right]$$

$$p(U|Y) \sim p_\psi(U|Y) \longrightarrow \hat{\psi} = \text{argmin}_\psi E_{X,U,Z}\left[-log(p_\psi(U \mid q_{\hat{\theta}}(X,Z)) \right]$$

$$p(S|Y) \sim p_\eta(S|Y) \longrightarrow \hat{\eta} = \text{argmin}_\eta E_{X,S,Z}\left[-log(p_\eta(S \mid q_{\hat{\theta}}(X,Z)) \right]$$

Implementation



Optimization objective:

$$\min_{q_\theta(X,Z)} \left[I(U;X|Y) + \lambda max\{I(S;Y) - k, 0\}^2\right]$$

Learning the stochastic mapping $Y = q_\theta(X,Z)$:

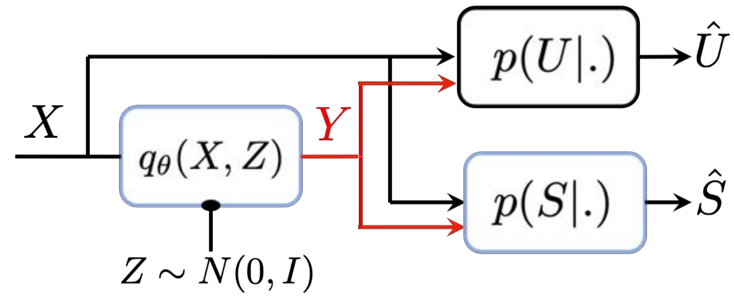
$$p(U|X) \sim p_\phi(U|X) \longrightarrow \hat{\phi} = \text{argmin}_\phi E_{X,U}\left[-log(p_\phi(U \mid X)\right]$$

$$p(U|Y) \sim p_\psi(U|Y) \longrightarrow \hat{\psi} = \text{argmin}_\psi E_{X,U,Z}\left[-log(p_\psi(U \mid q_{\hat{\theta}}(X,Z))\right]$$

$$p(S|Y) \sim p_\eta(S|Y) \longrightarrow \hat{\eta} = \text{argmin}_\eta E_{X,S,Z}\left[-log(p_\eta(S \mid q_{\hat{\theta}}(X,Z))\right]$$

$$\hat{\theta} = \text{argmin}_\theta E_{X,Z}\left[D_{KL}[p_{\hat{\phi}}(U \mid X) \mid\mid p_{\hat{\psi}}(U \mid q_\theta(X,Z))]\right]$$
$$+\lambda \max(E_{X,Z}\left[D_{KL}[p_{\hat{\eta}}(S \mid q_\theta(X,Z)) \mid\mid P(S)]\right] - k, 0)^2$$

Implementation



Optimization objective:

$$\min_{q_\theta(X,Z)} \; [I(U;X|Y) + \lambda max\{I(S;Y) - k, 0\}^2]$$

Learning the stochastic mapping $Y = q_\theta(X,Z)$:

$$p(U|X) \sim p_\phi(U|X) \rightarrow \hat{\phi} = \text{argmin}_\phi E_{X,U}\left[-log(p_\phi(U \mid X)\right]$$

$$p(U|Y) \sim p_\psi(U|Y) \rightarrow \hat{\psi} = \text{argmin}_\psi E_{X,U,Z}\left[-log(p_\psi(U \mid q_{\hat{\theta}}(X,Z))\right]$$

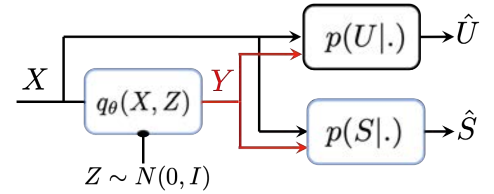
$$p(S|Y) \sim p_\eta(S|Y) \rightarrow \hat{\eta} = \text{argmin}_\eta E_{X,S,Z}\left[-log(p_\eta(S \mid q_{\hat{\theta}}(X,Z))\right]$$

Xception Networks

$$\hat{\theta} = \text{argmin}_\theta E_{X,Z}\left[D_{KL}[p_{\hat{\phi}}(U \mid X) \mid\mid p_{\hat{\psi}}(U \mid q_\theta(X,Z))]\right]$$
$$+\lambda \max(E_{X,Z}\left[D_{KL}[p_{\hat{\eta}}(S \mid q_\theta(X,Z)) \mid\mid P(S)]\right] - k, 0)^2$$

U-NET + noise

Experiments



Emotion obfuscation vs gender detection

k

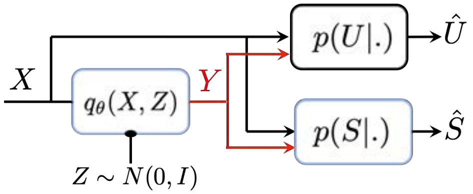
∞

0.5

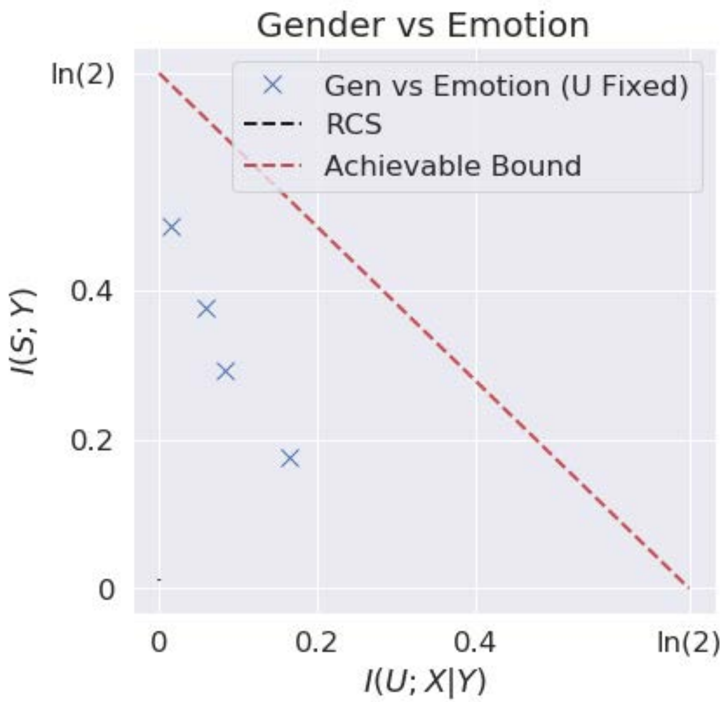
0.3

Sensitivity Tolerance k	Fixed Emotion Accuracy	Adversarial Emotion Accuracy	Fixed Gender Accuracy
∞	**91.8%**	**91.8%**	**94.9%**
0.5	68.4%	91.4%	89.3%
0.4	58.6%	85.8%	88.0%
0.3	56.8%	81.5%	86.7%
0.2	**51.9%**	**74.3%**	**83.9%**
Guessing	51.9%	51.9%	60.7%

Experiments



Emotion obfuscation vs gender detection



k

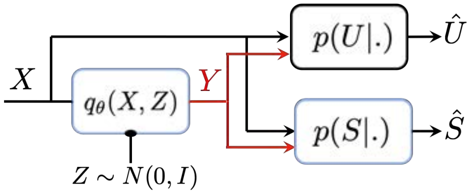
∞

0.5

0.3

Sensitivity Tolerance k	Fixed Emotion Accuracy	Adversarial Emotion Accuracy	Fixed Gender Accuracy
∞	**91.8%**	**91.8%**	**94.9%**
0.5	68.4%	91.4%	89.3%
0.4	58.6%	85.8%	88.0%
0.3	56.8%	81.5%	86.7%
0.2	**51.9%**	**74.3%**	**83.9%**
Guessing	51.9%	51.9%	60.7%

Experiments



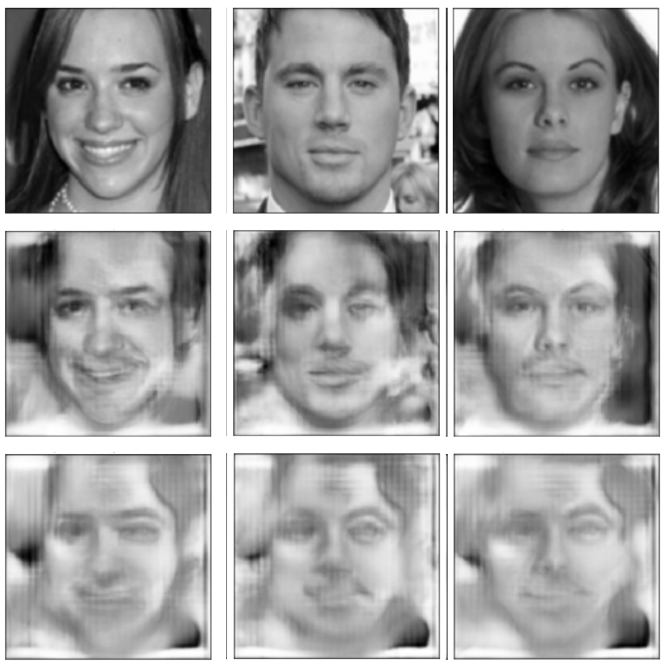
Gender obfuscation vs subject verification

k

∞

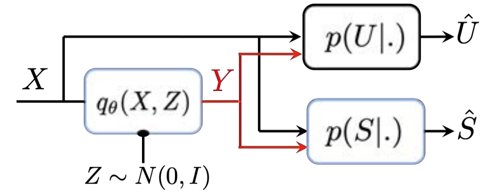
0.3

0.2

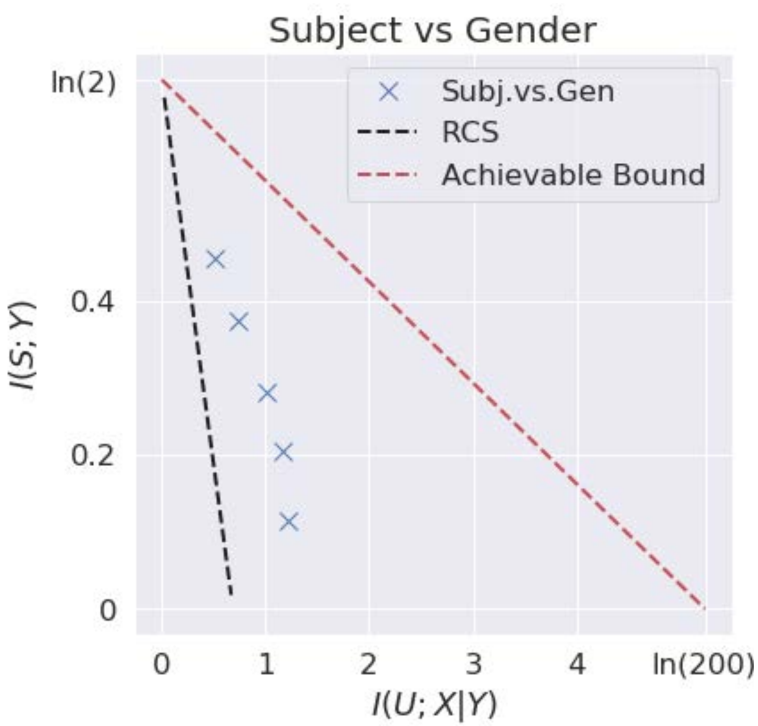


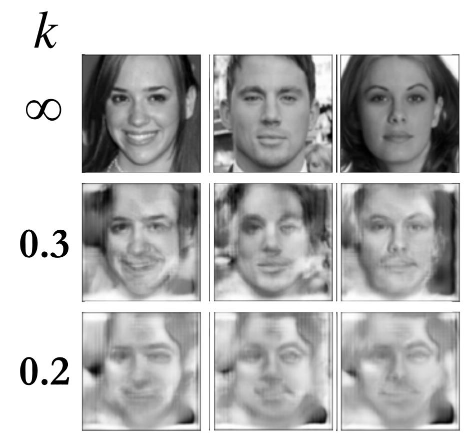
SENSITIVITY TOLERANCE k	FIXED GENDER ACCURACY	ADVERSARIAL GENDER ACCURACY	FIXED SUBJECT TOP-5 ACCURACY	RETRAINED SUBJECT TOP-5 ACCURACY
∞	**98.6%**	**98.6%**	**98.8%**	**98.8%**
0.5	59.5%	90.2%	93.5%	96.8%
0.4	60.3%	85.3%	88.1%	94.9%
0.3	54.0%	79.4%	81.4%	92.8%
0.2	**56.1%**	**74.6%**	**81.6%**	**91.0%**
0.1	51.6%	67.1%	74.5%	89.6%
GUESSING	54.8%	54.8%	2.5%	2.5%

Experiments



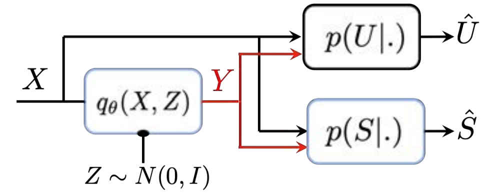
Gender obfuscation vs subject verification



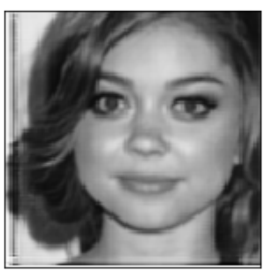


Sensitivity Tolerance k	Fixed Gender Accuracy	Adversarial Gender Accuracy	Fixed Subject Top-5 Accuracy	Retrained Subject Top-5 Accuracy
∞	**98.6%**	**98.6%**	**98.8%**	**98.8%**
0.5	59.5%	90.2%	93.5%	96.8%
0.4	60.3%	85.3%	88.1%	94.9%
0.3	54.0%	79.4%	81.4%	92.8%
0.2	**56.1%**	**74.6%**	**81.6%**	**91.0%**
0.1	51.6%	67.1%	74.5%	89.6%
Guessing	54.8%	54.8%	2.5%	2.5%

Experiments

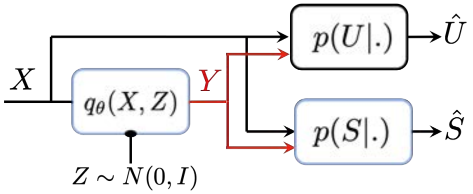


Subject within Subject

k	Consenting User	Nonconsenting User
	Subject verified ✓	Subject verified ✓
∞		
	Subject verified ✓	Subject verified ✗
0.5		

Sensitivity Tolerance k	Consenting Users Top-5 Accuracy	Nonconsenting Users Top-5 Accuracy
∞	**98.7%**	**97.9%**
3	98.3%	9.38%
1	97.8%	6.25%
0.5	**97.6%**	**4.69%**
Guessing	2.5%	2.5%

Concluding remarks



- Learned representations that *preserve utility* and *obfuscate sensitive information.*
- Transformations are *space-preserving*. Can reuse existing pipelines.
- Derived easy-to-compute bounds.
- Experimental results show representations compare favorably against derived bounds.

Limitations:

- Expectation-based approach.
- Reliance on adversary as a proxy for information.

Thanks!

Please visit us at poster #81