Fairness and Privacy in Machine Learning

Elliot Creager
AIMS
May 17, 2020

Resources

Slides due to Elliot Creager, Rich Zemel, Toni Pitassi, and David Madras

Course materials (AIMS) - https://github.com/ecreager/ammi-fairness-and-privacy

Course materials (UofT) - http://www.cs.toronto.edu/~toni/Courses/Fairness/fair.html

Overview

➔ **Fairness**

Motivation and approaches

Fair classification and representation learning

➔ **Privacy**

Motivation and basic mechanisms

Threat models for non-private ML

Private logistic regression and neural nets

Part 1: Fairness in ML

WHY WAS I NOT SHOWN THIS AD?

FAIRNESS IN AUTOMATED DECISIONS

Algorithmic unfairness: Algorithms are pervasive, high-stakes, high-impact

Need more than just ”accuracy”

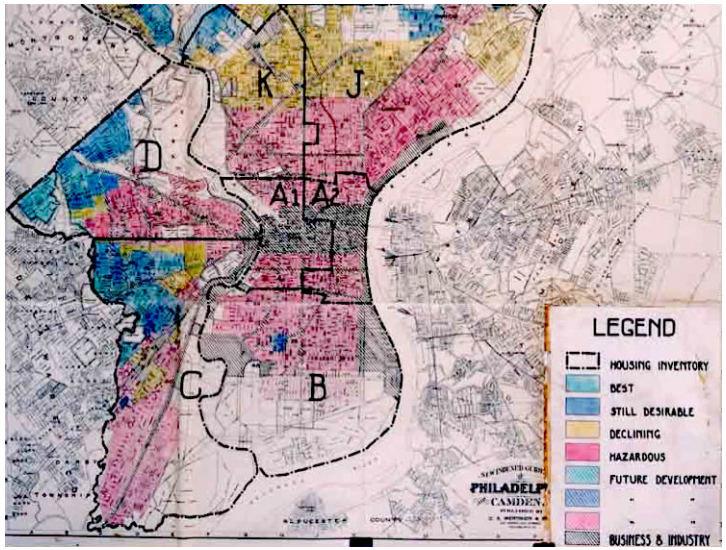
CONCERN: DISCRIMINATION

- Population includes minorities
  - Ethnic, religious, medical, geographic
- Protected by law, policy, ethics
- (If) we cannot completely control our data, can we regulate how it is used, how decisions are made based on it?

Forms of Discrimination

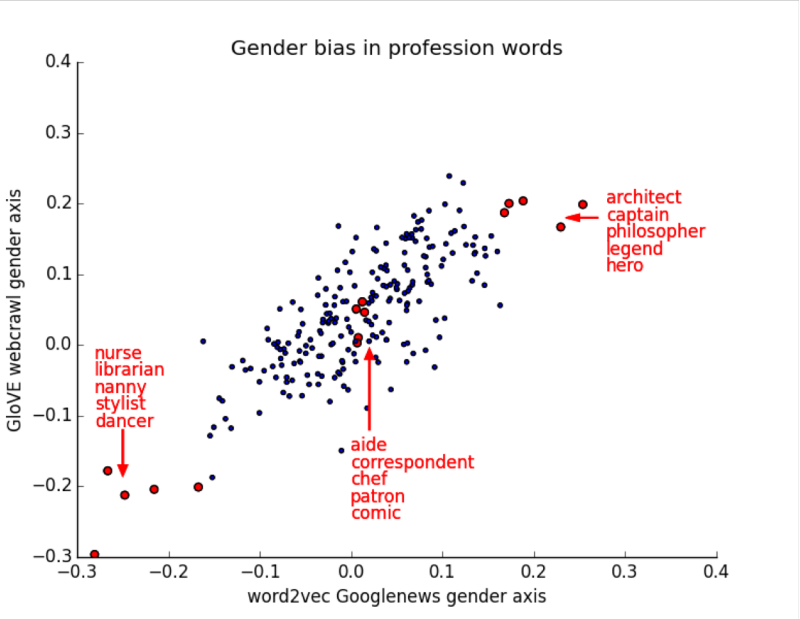
- *Steering* minorities into higher rates (advertising)

- *Redlining:* deny service, change rates based on area

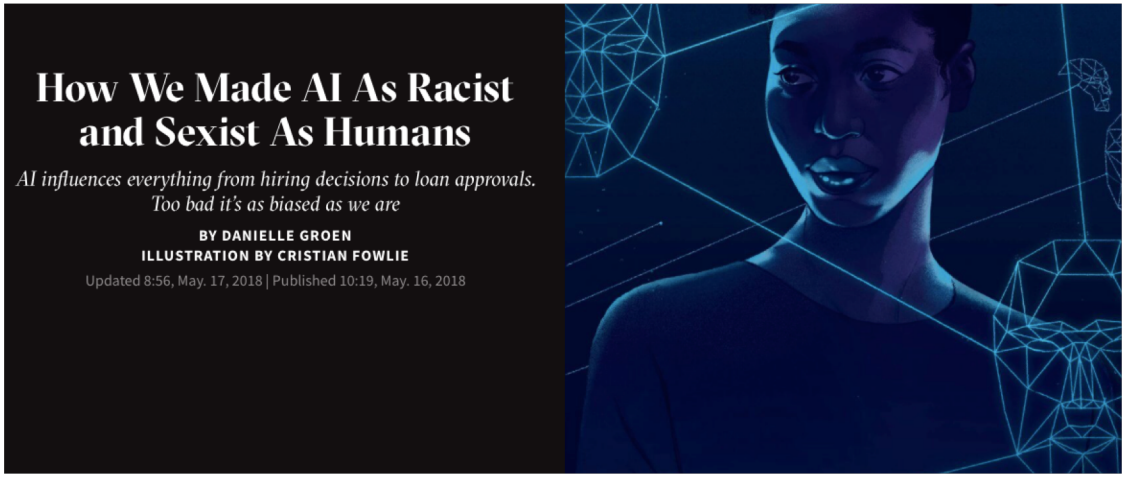


- *Self-fulfilling prophecy:* select less qualified to "justify" future discrimination

Unfairness in Machine Learning?



Joy Buolawmini



The Walrus, 2018

Fairness in ML: Goals

Identify and mitigate bias in ML-based decision-making, in all aspects of data pipeline

Sources of Bias/Discrimination

DATA

- Imbalanced / impoverished data
- Labeled data imbalance (more data on white recidivism outcomes)
- Labeled data incorrect / noisy (historical bias)

MODEL

- ML prediction error imbalanced
- Compound injustices (Hellman)

FAIR CLASSIFICATION

Explosion of fairness research over last five years

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction
- A, a sensitive attribute (race, gender, age, socio-economic status)

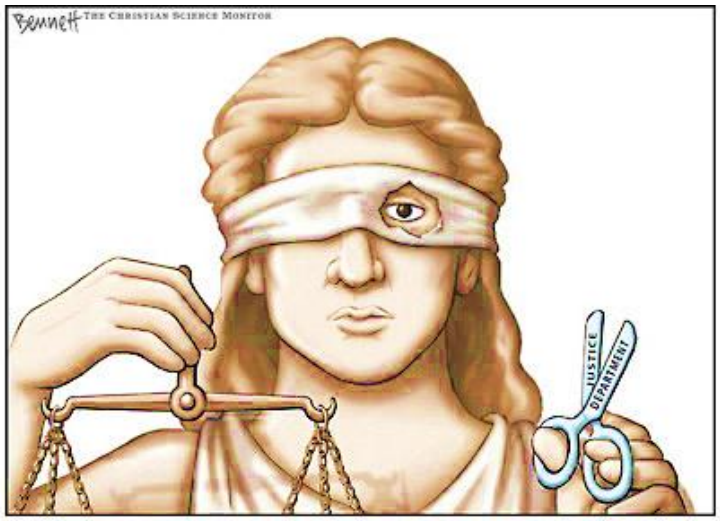
We want to learn a classifier that is:

- accurate
- fair with respect to A

FAIRNESS THROUGH AWARENESS

Dwork, Hardt, Pitassi, Reingold, Zemel, 2012

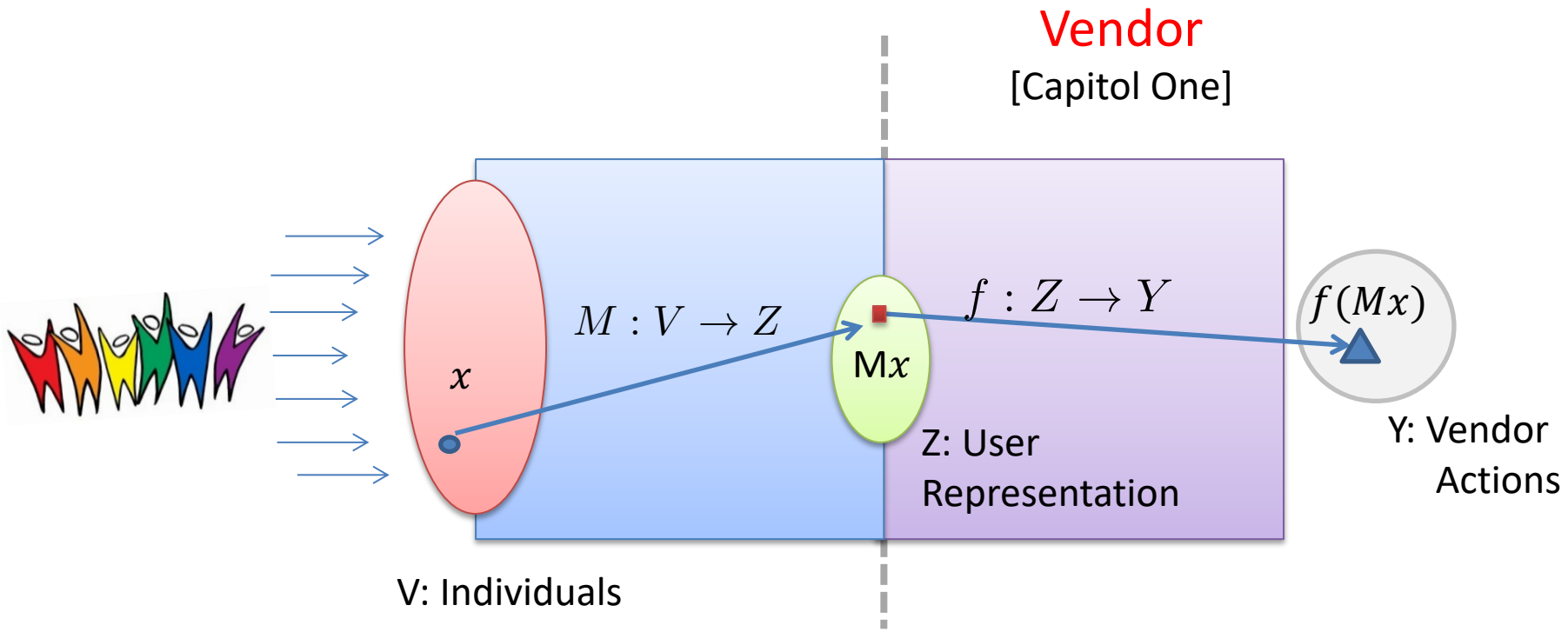
Goal: Assign each individual *a* representation *by being aware of membership in group A*



(1). **Individual Fairness**: Treat similar individuals similarly

(2). **Group Fairness:** equalize two groups (A=1 = minority; A=0 is majority) at the level of outcomes (statistical parity)

General Framework



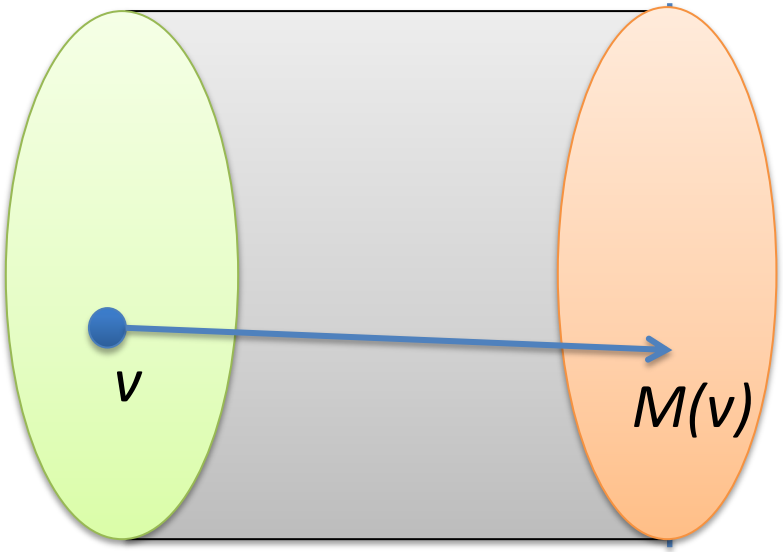
Our goal:

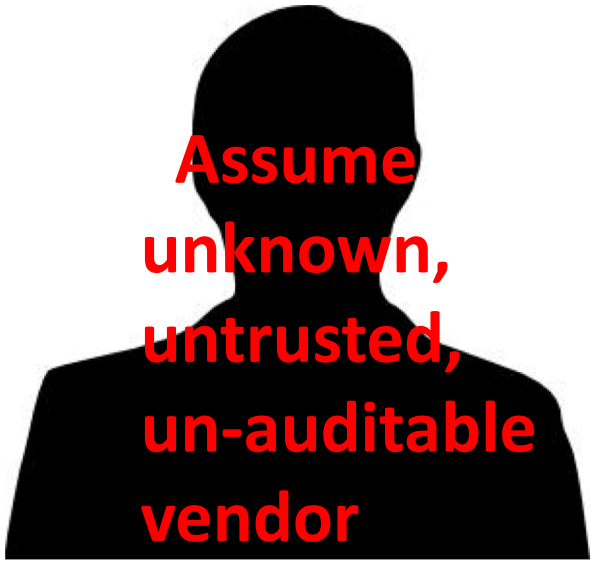
Achieve Fairness in the representation step

Ad Network (with society oversight)

Ad Network

$M : V \rightarrow Z$

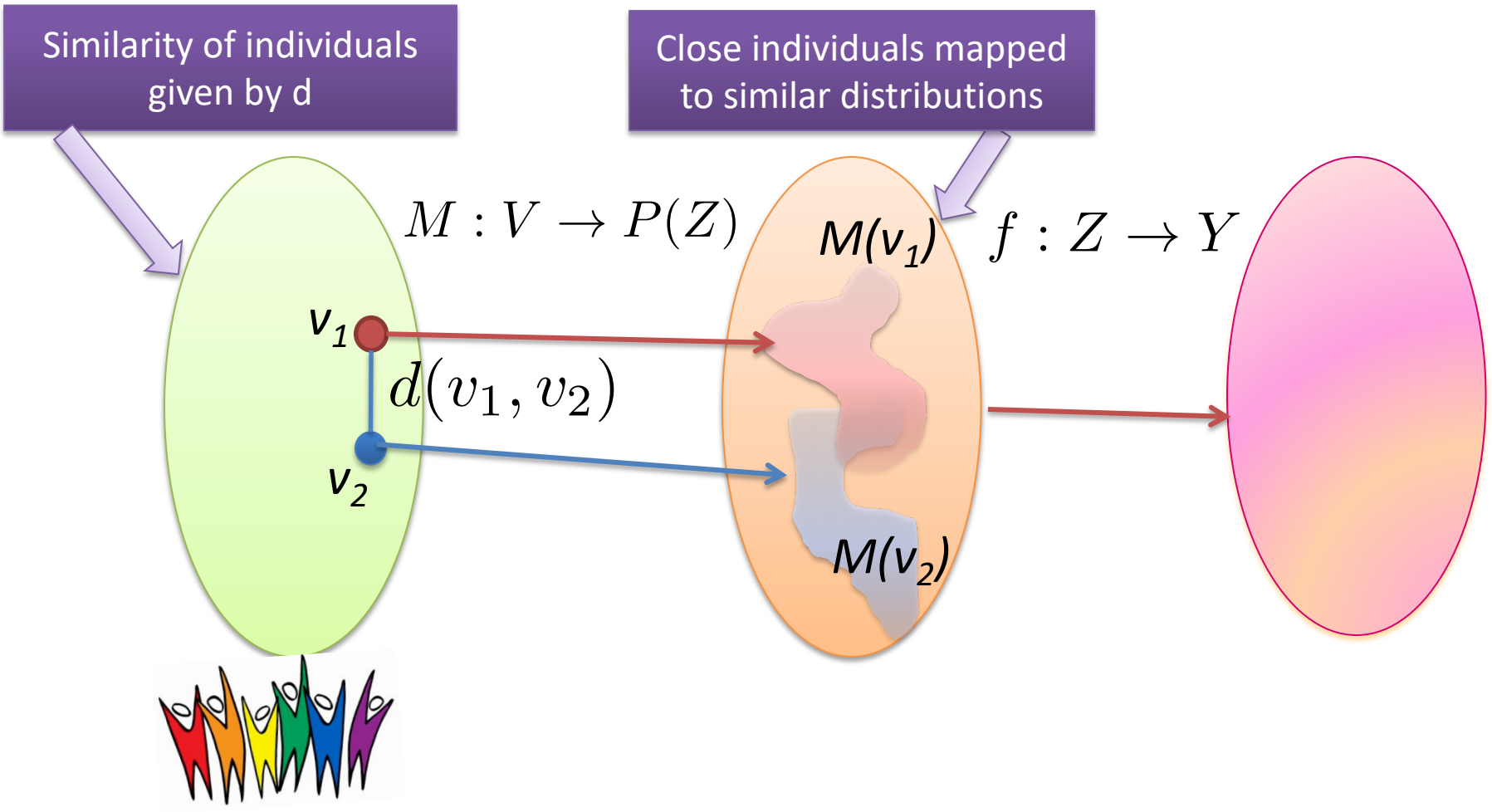




V: Individuals

Z: representation

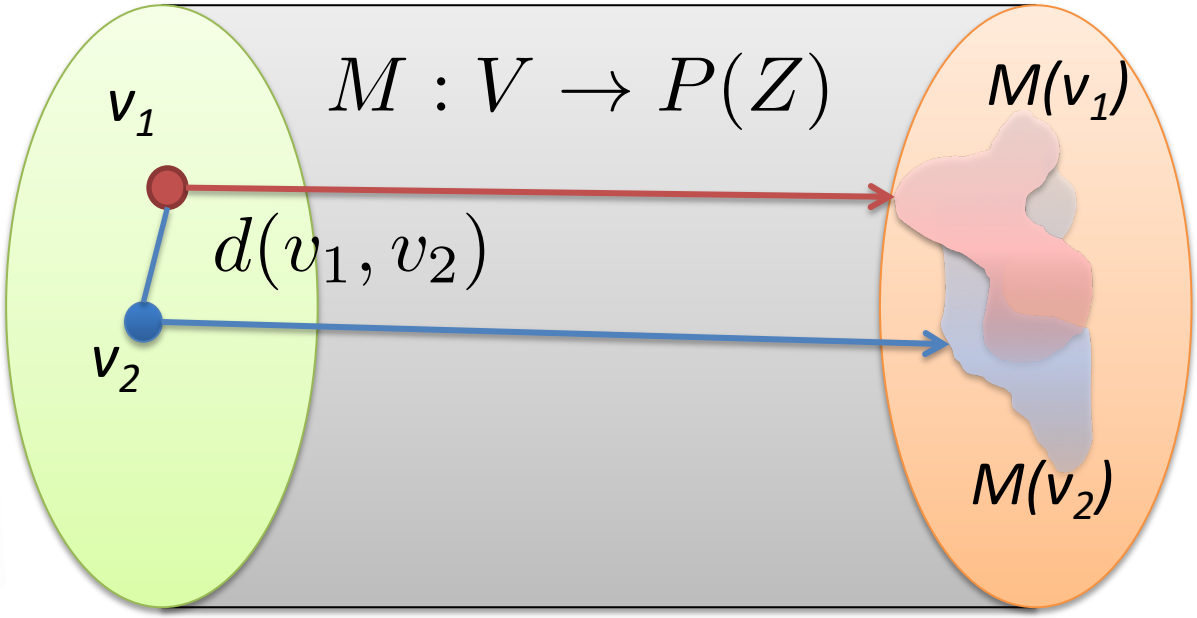
Our Approach: Define a randomized mapping that "blends people with the crowd"



V: Individuals *Z*: representations Y: Actions

Metric $d : V \times V \rightarrow \mathbb{R}$

Lipschitz condition $||M(v_1) - M(v_2)|| \leq d(v_1, v_2)$



V: Individuals

Z: Representations

The Metric

- Assume ***task-specific similarity metric***
  - Extent to which two individuals are similar w.r.t. the classification task at hand
- Ideally captures *ground truth*
  - Or, society's best approximation
- Open to public discussion, refinement

Examples: Financial/insurance risk metrics

  - Already widely used (though secret)
- AALIM health care metric
  - health metric for treating similar patients similarly
- Roemer's relative effort metric
  - Well-known approach in economics/political theory

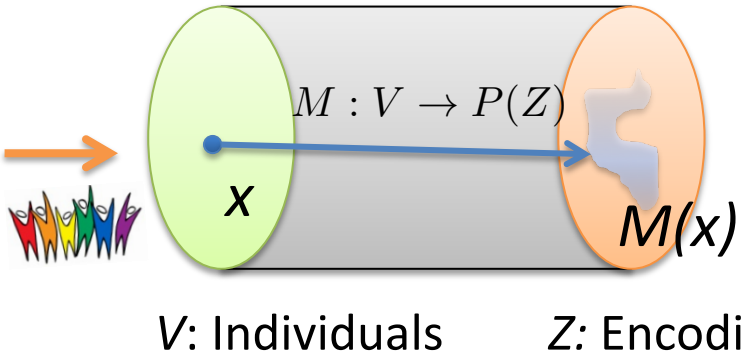
An Algorithm for Fair Classification

utility function
$U: V \times Z \to R$

Metric
$d: V \times V \to R$

Efficient Procedure

d-fair mapping M

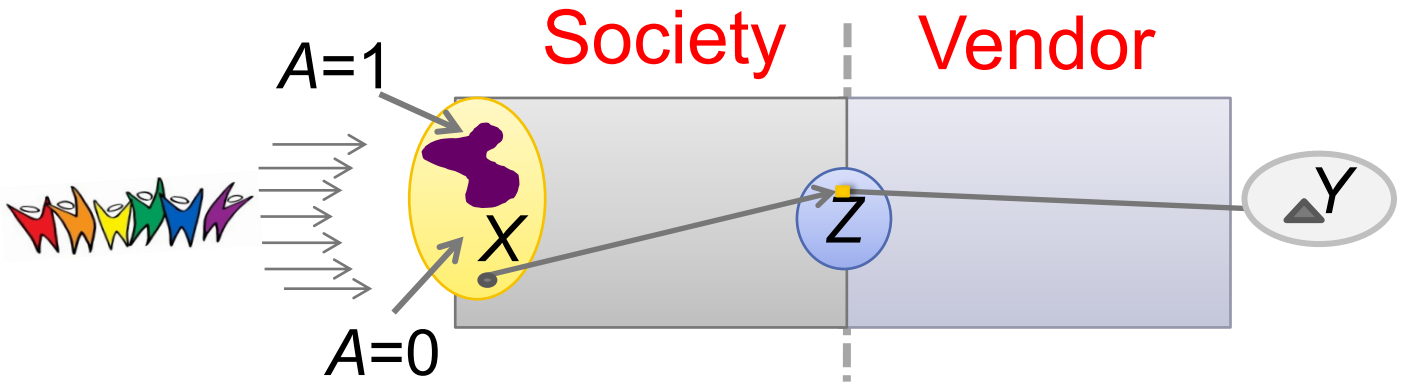


V: Individuals *Z:* Encodings

LP maximizes vendor's expected utility subject to fairness conditions

FAIR REPRESENTATION LEARNING: FRAMEWORK

Zemel. Wu, Swersky, Pitassi, Dwork, 2013



Goal: Learn a mapping from X to distributions over representations *Z that is fair*

Aims for Z:

1. Lose information about A:

$$P[Z=k \mid A=1] = P[Z=k \mid A=0]$$

2. Retain information about X
3. Preserve information for classification so vendor can max utility [decisions Y = g(Z)]

INITIAL FORMULATION

Difficult to jointly optimize:

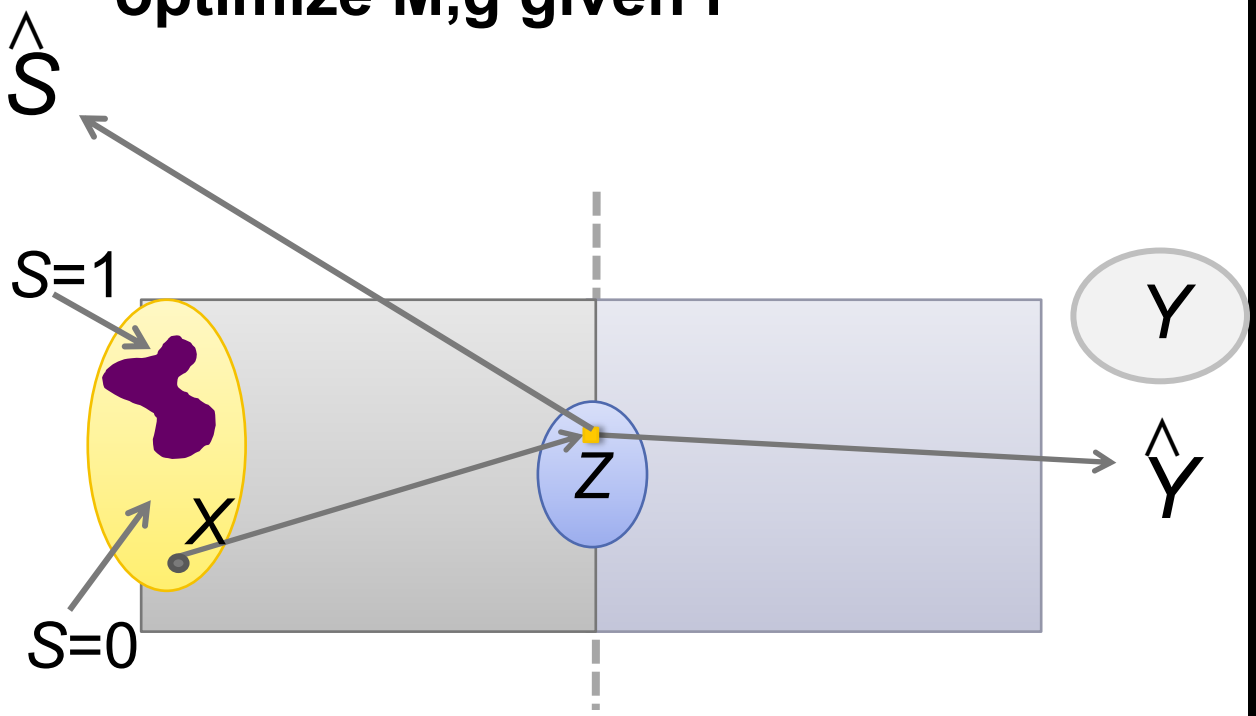
min. |f(Z) – Y|; max. |g(Z) – S| (thwart adversary)

Can alternate:

optimize M,f given g;

optimize M,g given f

But unstable



INSTANTIATING THE MODEL

Key: min. *MI(Z,S)* by forcing *P(Z|S+)* = *P(Z|S-)*

$$P(Z|S) = \int_X P(Z|X,S)P(X|S)dX$$

$$P(Z|S=1) \approx \frac{1}{N^+} \sum_{n=1}^{N^+} P(Z|X,S=1)$$

$$P(Z|S=1) = P(Z|S=0) = P(Z) \Rightarrow$$

Simple tractable formulation:

$$MI(Z,S) = 0$$

***Z* is a discrete latent variable**

FULL OBJECTIVE FUNCTION

Learn mapping M(X) to minimize L

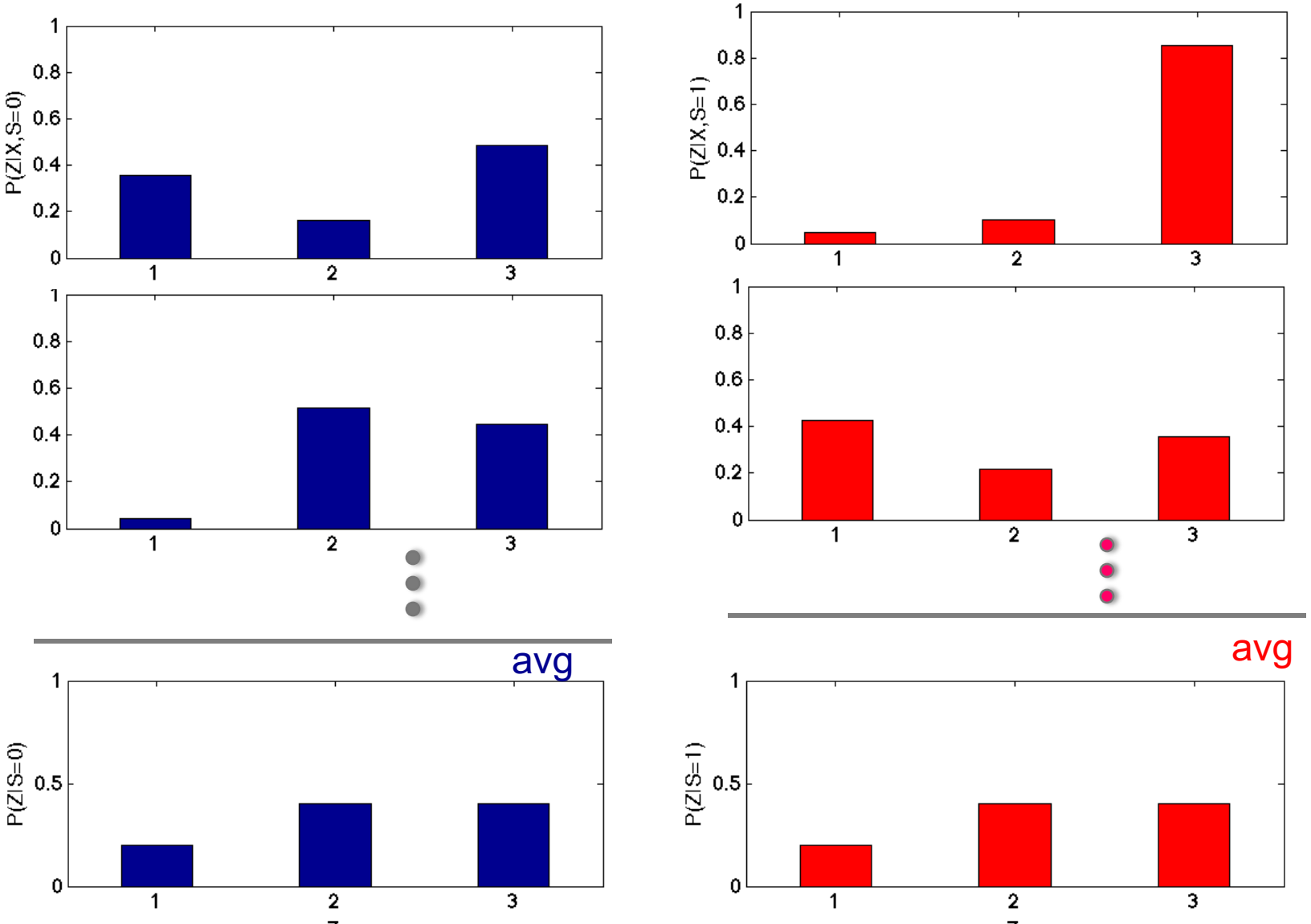
$$P^{+}_{n,k} = P(Z = k|\mathbf{x}, S = 1) = \frac{\exp(\mathbf{x}_n^T \mathbf{w}_k^{+})}{\sum_{k'} \exp(\mathbf{x}_n^T \mathbf{w}_{k'}^{+})}$$

$$L = A_y \cdot L_y + A_z \cdot L_z$$

$$L_z = \sum_k |P_k^{+} - P_k^{-}| \qquad P_k^{+} = P(Z = k|S = 1)$$

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \qquad \hat{y}_n = \sum_k P_{n,k} u_k$$

OBFUSCATING MEMBERSHIP



$$P(Z|S=1) = P(Z|S=0) \Rightarrow MI(Z,S) = 0$$

EXPERIMENTS

1. **German Credit**

   Size: 1000 instances, 20 attributes

   Task: classify as good or bad credit

   Sensitive feature: Age

2. **Adult Income**

   Size: 45,222 instances, 14 attributes

   Task: predict whether or not annual income > 50K

   Sensitive feature: Gender

3. **Heritage Health**

   Size: 147,473 instances, 139 attributes

   Task: predict whether patient spends any nights in hospital

   Sensitive feature: Age

PERFORMANCE METRICS

- **Accuracy**

$$yAcc = 1 - \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$$

- **Discrimination**

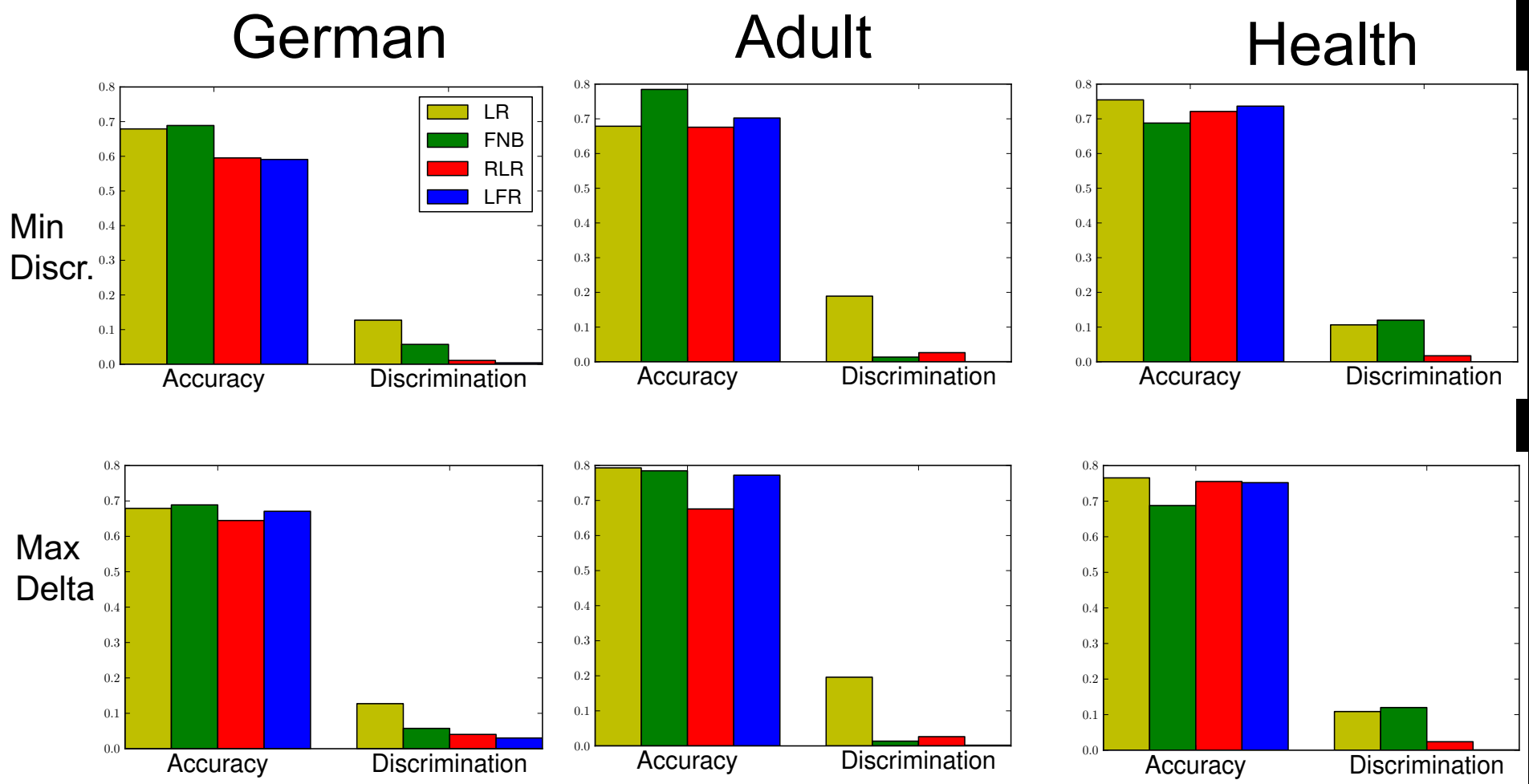
$$yDiscrim = |\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}|$$

ALTERNATIVE APPROACHES

Build fair classifier and force vendor to use it:

- Massage labels to achieve proportional access (FNB) [Kamiran & Calders, 2009]
- Trade off classification error vs. discrimination (RLR) [Kamishima et al, 2011]

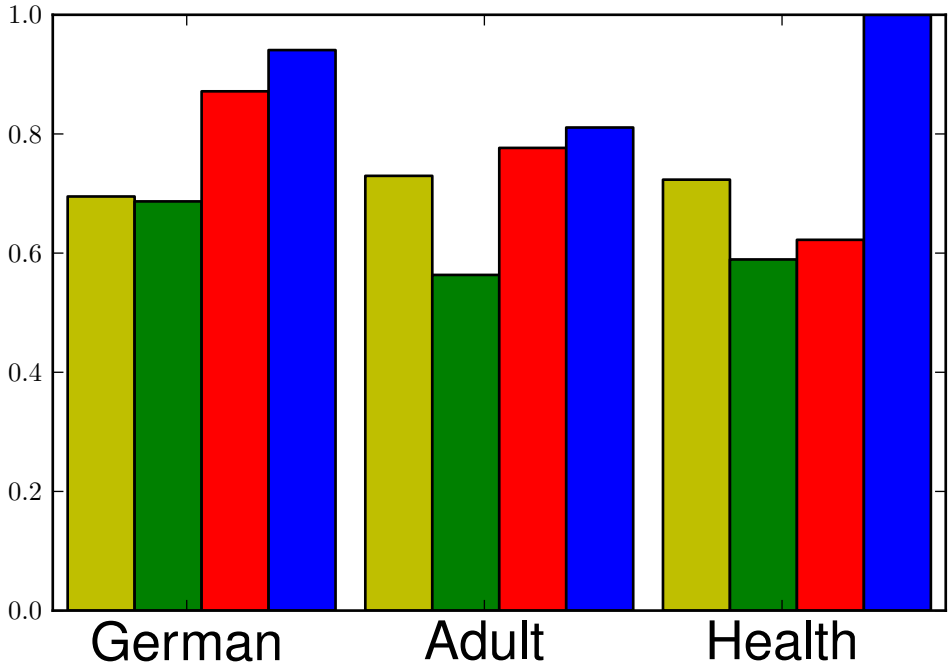
EXPERIMENTAL RESULTS



RESULTS: INDIVIDUAL FAIRNESS

Consistency:

$$yNN = 1 - \frac{1}{N}\sum_{n} |\hat{y}_n - \frac{1}{k}\sum_{j \in kNN(\mathbf{x_n})} \hat{y}_j|$$



EXAMPLE DOMAINS

1. Targeted search/advertising: How do different groups see internet content?
   - Males/females with equal interest, equal p(ad)?
   - (leisure interests; lower paying jobs; credit card rates)

2. Medical testing/diagnosis: decision-making based on tests, that affect p(diagnosis)
   - Applied uniformly to different groups
   - Medical tests for conditions that vary widely between groups

3. Recidivism: risk tools assess p(future-arrest) given history
   - Used in decisions about bail, sentencing, parole
   - Claims of bias based on race against COMPAS risk tool

Common:

1. Algorithm input to decision-maker
2. Attempting to classify individual possesses property: interest; condition; risk
3. Output is a probability

REPRESENTATIONS BEYOND CLUSTERS

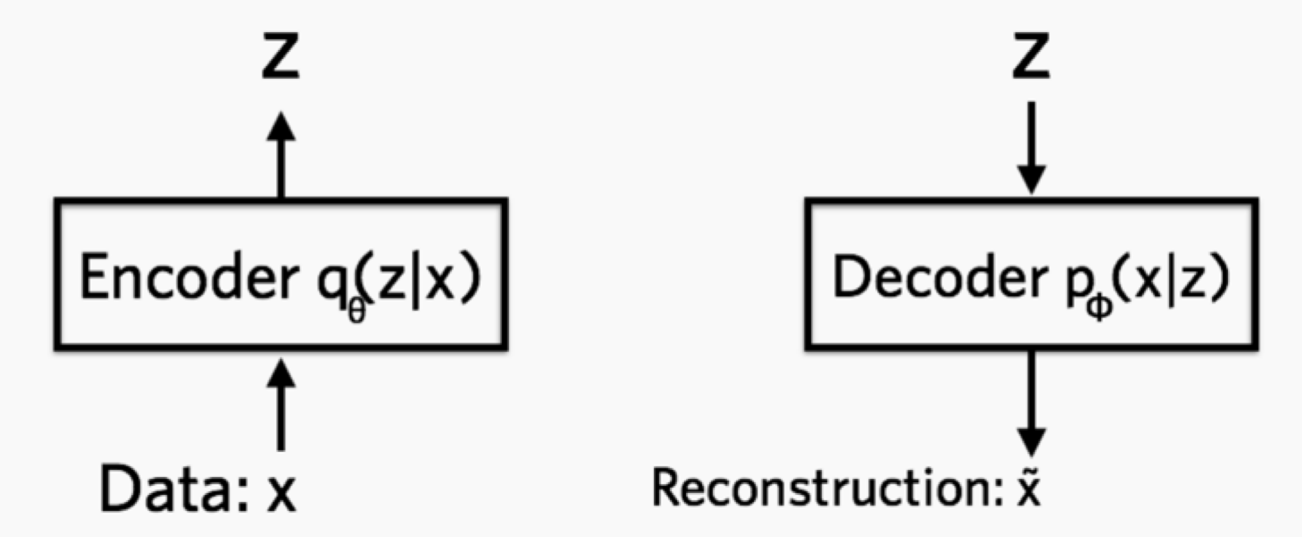
Aim: Replace discrete representation with continuous, multi-dimensional Z

Allow more flexible, nuanced representations

Bring ML arsenal to bear: powerful methods for mapping, embedding in vector spaces: Variational Auto Encoders (VAE)

How to maintain statistical parity in learned representations?

VAE



Re-formulation of autoencoders:

- Each input encoded into a distribution in latent space
- Output prediction obtained by sampling from distribution, mapping through decoder

Allows maximum-likelihood based density modelling:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right)$$

MMD

- Suppose we have access to samples from two probability distributions $X \sim P_A$ and $Y \sim P_B$, how can we tell if $P_A = P_B$?
- Maximum Mean Discrepancy (MMD) is a measure of distance between two distributions given only samples from each. [Gretton 2010]

$$\left\| \frac{1}{N} \sum_{n=1}^{N} \phi(X_n) - \frac{1}{M} \sum_{m=1}^{M} \phi(Y_m) \right\|^2$$

$$= \frac{1}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} \phi(X_n)^\top \phi(X_{n'}) + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \phi(Y_m)^\top \phi(Y_{m'}) - \frac{2}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \phi(X_n)^\top \phi(Y_m)$$

$$= \frac{1}{N^2} \sum_{n=1}^{N} \sum_{n'=1}^{N} k(X_n, X_{n'}) + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} k(Y_m, Y_{m'}) - \frac{2}{MN} \sum_{n=1}^{N} \sum_{m=1}^{N} k(X_n, Y_m)$$

- Our idea: learn to make two distributions indistinguishable

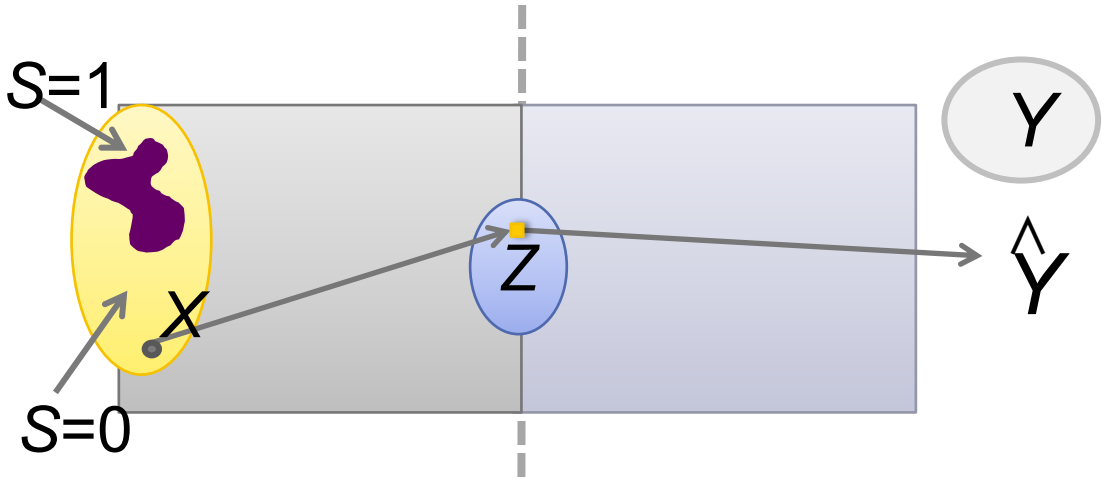
  ➔ small MMD!

VARIATIONAL FAIR AUTOENCODER

VAE with regularizer on latent representations

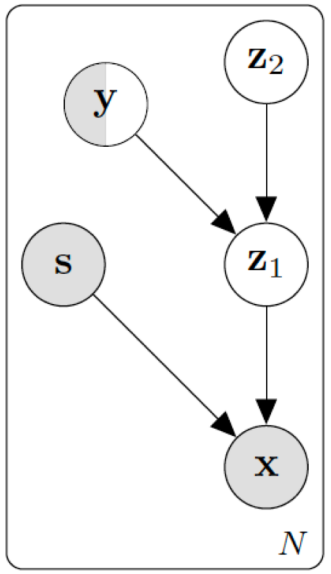
Match higher-order moments, continuous *Z*:

$$\ell_{\text{MMD}}(\mathbf{Z}_{1\mathbf{s}=0}, \mathbf{Z}_{1\mathbf{s}=1}) = \| \mathbb{E}_{\tilde{p}(\mathbf{x}|\mathbf{s}=0)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=0)}[\psi(\mathbf{z}_1)]] - E_{\tilde{p}(\mathbf{x}|\mathbf{s}=1)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=1)}[\psi(\mathbf{z}_1)]] \|^2$$



VARIATIONAL FAIR AUTOENCODER

Extend VAE to include some labels *y* (semi-supervised VAE [Kingma & Welling, 2014]) and "nuisance variable" *s*



Objective -- maximize:

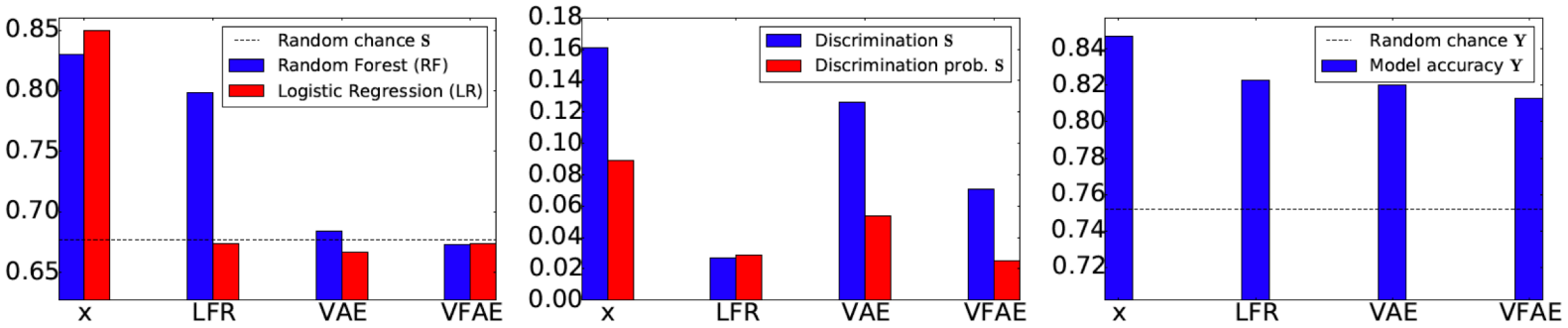
$$\sum_{n=1}^{N_s} \mathbb{E}_{q_\phi(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)}[-KL(q_\phi(\mathbf{z}_{2n}|\mathbf{z}_{1n},\mathbf{y}_n)||p(\mathbf{z}_2)) + \log p_\theta(\mathbf{x}_n|\mathbf{z}_{1n},\mathbf{s}_n)] +$$

$$+ \mathbb{E}_{q_\phi(\mathbf{z}_{2n}|\mathbf{z}_{1n},\mathbf{y}_n)}[-KL(q_\phi(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)||p_\theta(\mathbf{z}_{1n}|\mathbf{z}_{2n},\mathbf{y}_n))]$$

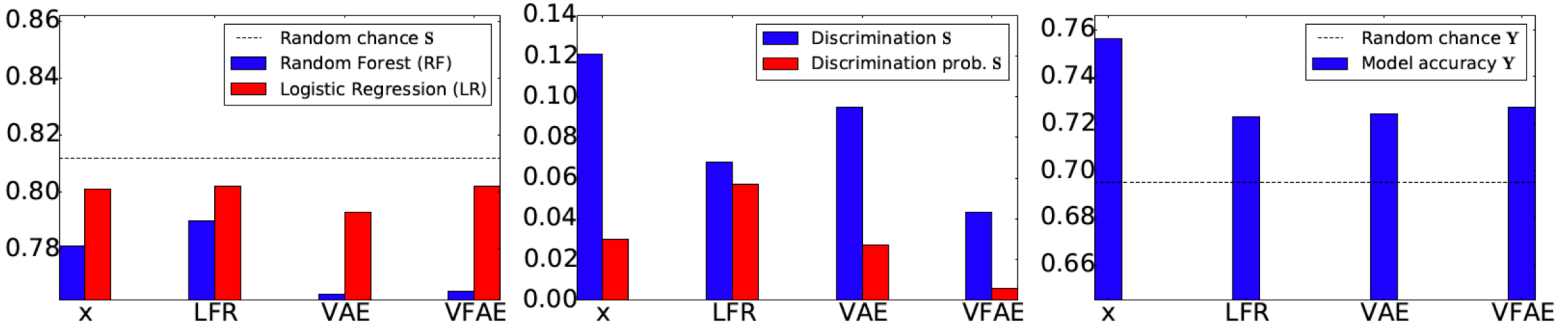
Add for labeled set: $\sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{z}_{1n}|\mathbf{x}_n,\mathbf{s}_n)}[-\log q_\phi(\mathbf{y}_n|\mathbf{z}_{1n})]$

unlabeled set: $\sum_{m=1}^{M} \mathbb{E}_{q_\phi(\mathbf{z}_{1m}|\mathbf{x}_m,\mathbf{s}_m)}[-KL(q(\mathbf{y}_m|\mathbf{z}_{1m})||p(\mathbf{y}_m))$

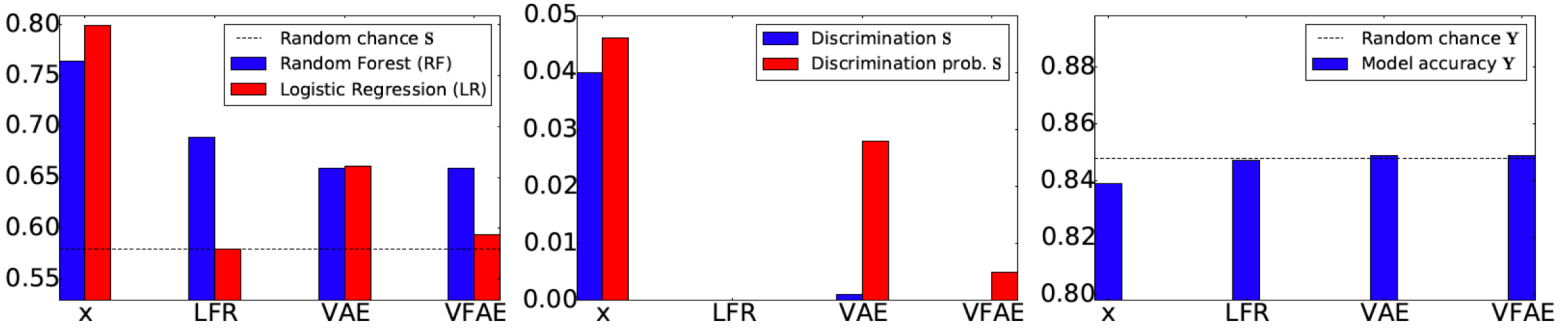
RESULTS



(a) Adult dataset

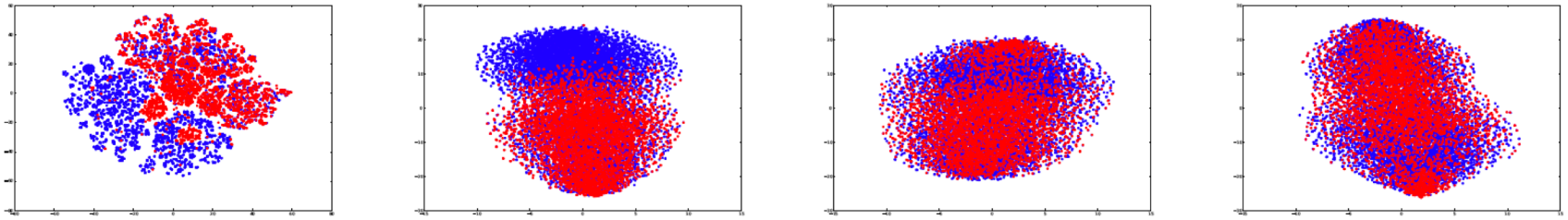


(b) German dataset



(c) Health dataset

RESULTS



(a) (b) (c) (d)

Figure 4: t-SNE (van der Maaten, 2013) visualizations from the Adult dataset on: (a): original $\mathbf{x}$, (b): latent $\mathbf{z}_1$ without $\mathbf{s}$ and MMD, (c): latent $\mathbf{z}_1$ with $\mathbf{s}$ and without MMD, (d): latent $\mathbf{z}_1$ with $\mathbf{s}$ and MMD. Blue colour corresponds to males whereas red colour corresponds to females.

ADAPTING THE FRAMEWORK

The same idea has many other useful applications, e.g.,

- Eliminating demographic discrimination in deciding who should get transplant surgery
- Removing confounds, such as which scanner produced a medical image

Key: Learning to make two (or more) distributions indistinguishable

DOMAIN ADAPTATION

Natural fit: domain adaptation

Make feature representations for source and target domain data indistinguishable

Sentiment classification

- Product reviews (text, tf-idf on words & bigrams)
- Labeled data from source domain, unlabeled data from target domain

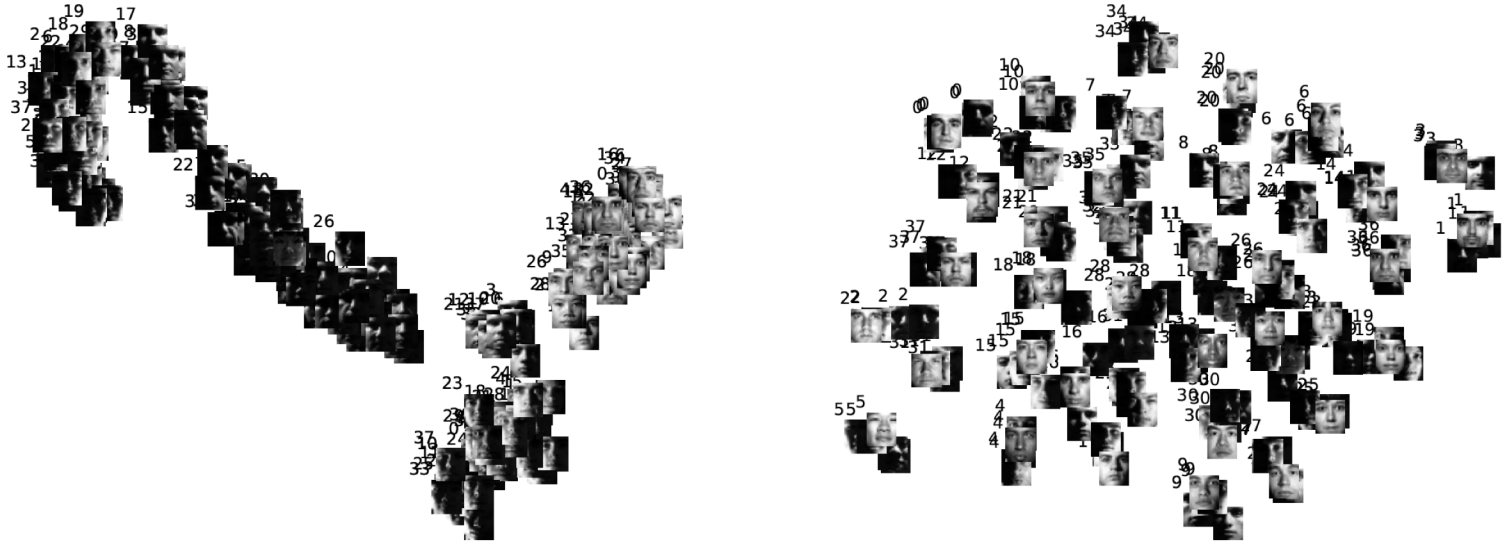
Source - Target	S		Y	
	RF	LR	VFAE	DANN
books - dvd	0.535	0.564	**0.799**	0.784
books - electronics	0.541	0.562	**0.792**	0.733
books - kitchen	0.537	0.583	**0.816**	0.779
dvd - books	0.537	0.563	**0.755**	0.723
dvd - electronics	0.538	0.566	**0.786**	0.754
dvd - kitchen	0.543	0.589	**0.822**	0.783
electronics - books	0.562	0.590	**0.727**	0.713
electronics - dvd	0.556	0.586	**0.765**	0.738
electronics - kitchen	0.536	0.570	0.850	**0.854**
kitchen - books	0.560	0.593	**0.720**	0.709
kitchen - dvd	0.561	0.599	0.733	**0.740**
kitchen - electronics	0.533	0.565	0.838	**0.843**

LEARNING INVARIANT FEATURES

If we have labeled data from all domains, factoring out unwanted domain bias still leads to better generalization.

Make the learned representations invariant to unwanted transformation / variation / bias.

Example: Face identification under different lighting conditions



ADVERSARIAL FAIR LEARNING

Rather than using MMD to ensure learned representation is fair, can use adversarial approach

Adversary takes latent representation (here *R*) as input and attempts to predict *S*, then model minimizes:

$$D_{\theta,\phi}(R,S) = \mathop{\mathbb{E}}_{X,S} S \cdot \log\left(\mathrm{Adv}(R)\right) + (1-S) \cdot \log\left(1-\mathrm{Adv}(R)\right)$$

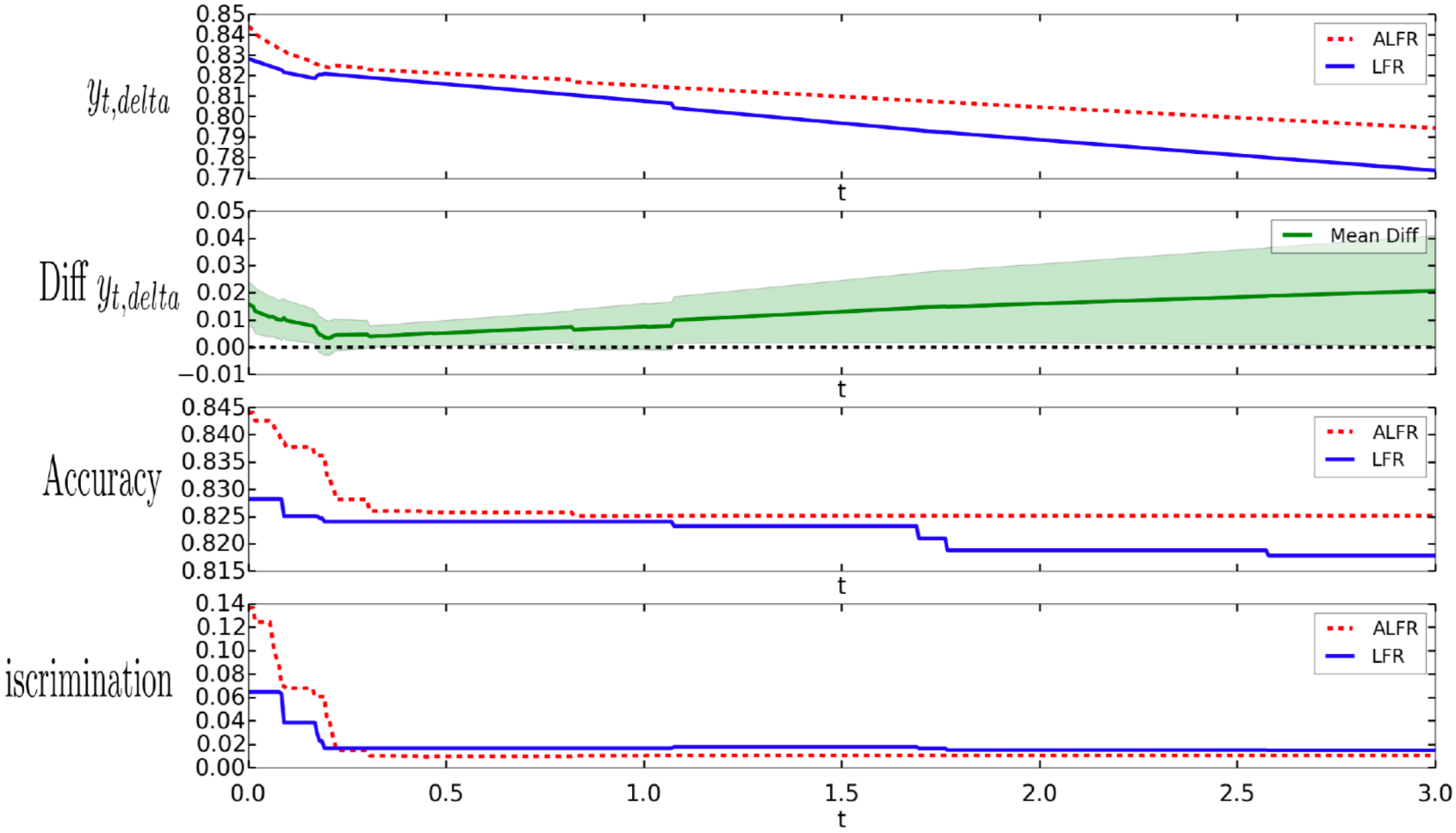
Combine with reconstruction and classification losses to ensure representation retains info about *X, Y*

$$C_\theta(X,R) = \mathop{\mathbb{E}}_{X} \|X - \mathrm{Dec}(R)\|_2^2$$

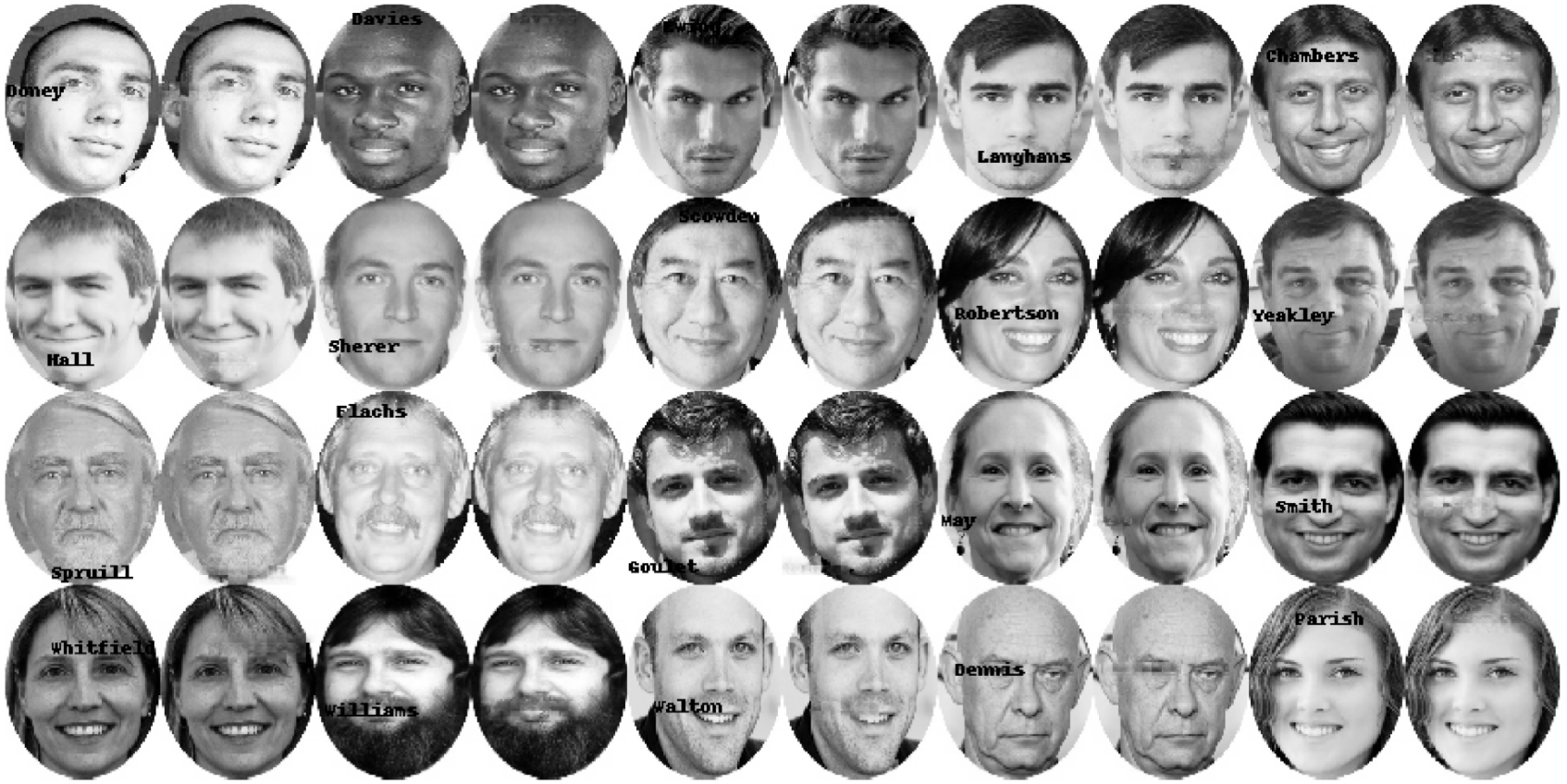
$$E_\theta(R,S|) = -\mathop{\mathbb{E}}_{X,Y} Y \cdot \log\left(\mathrm{Pred}(R)\right) + (1-Y) \cdot \log\left(1-\mathrm{Pred}(R)\right)$$

Censoring Representations with an Adversary: Edwards & Storkey, 2015

RESULTS


$y_{t,delta}$
ALFR
LFR
t
Diff $y_{t,delta}$
Mean Diff
t
Accuracy
ALFR
LFR
t
iscrimination
ALFR
LFR
t

RESULTS


Doney
Davies
Chambers
Langhans
Scowden
Hall
Sherer
Robertson
Yeakley
Flachs
Smith
May
Spruill
Goulet
Parish
Whitfield
Williams
Walton
Dennis

EQUALIZED ODDS / OPPORTUNITY

Both VFAE and AFLR define fairness as statistical parity

Problems with demographic/statistical parity:

- Coarse measure, not about individuals
- May entail large loss in accuracy

Alternative definition: equal opportunity [Hardt, Price, Srebro, 2016]

- Encourage perfect prediction
- But ensure that the prediction errors are balanced between the groups

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in \{0,1\}$$

BACK TO FAIR REPRESENTATIONS

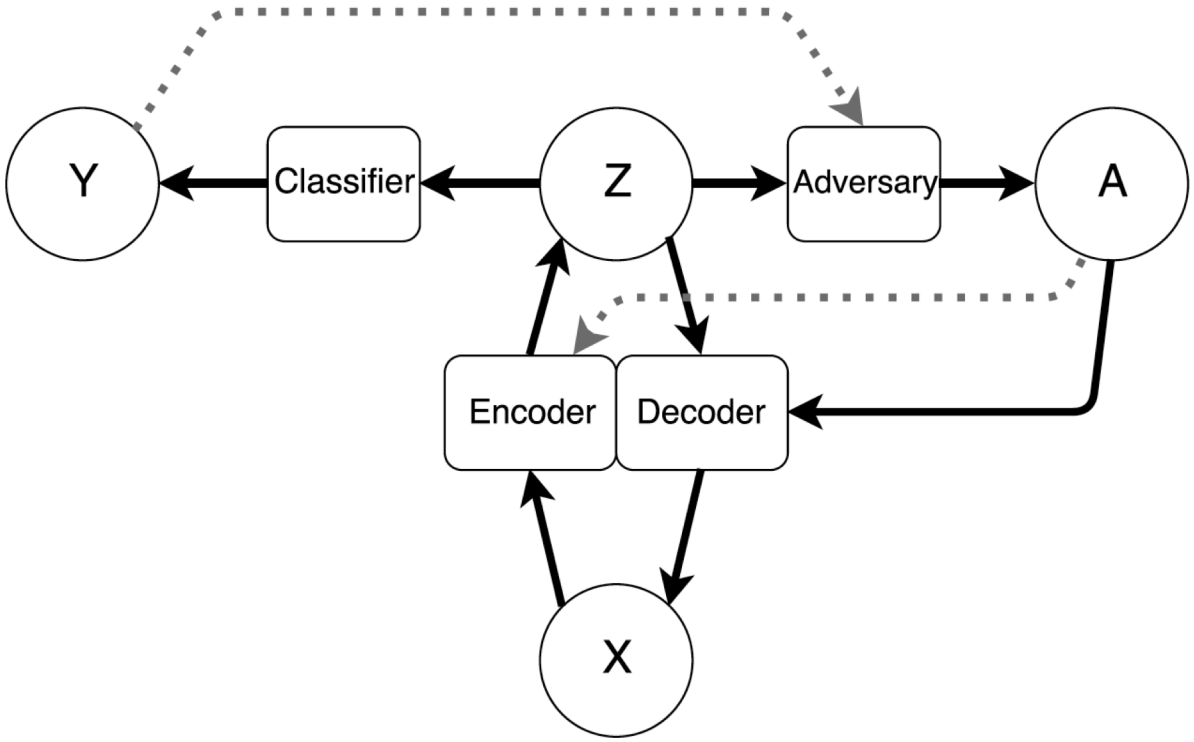
- Minimize unfair targeting of disadvantaged groups by vendors (worse lines of credit, lower paying jobs)
- Aim: form a data representation that ensures fair classifications downstream

- Consider two types of unfair vendors:
  1. The **indifferent** vendor: does not care about fairness, only maximizes utility
  2. The **malicious** vendor: doesn't care about utility, discriminates unfairly

- Good fit to adversarial learning scheme

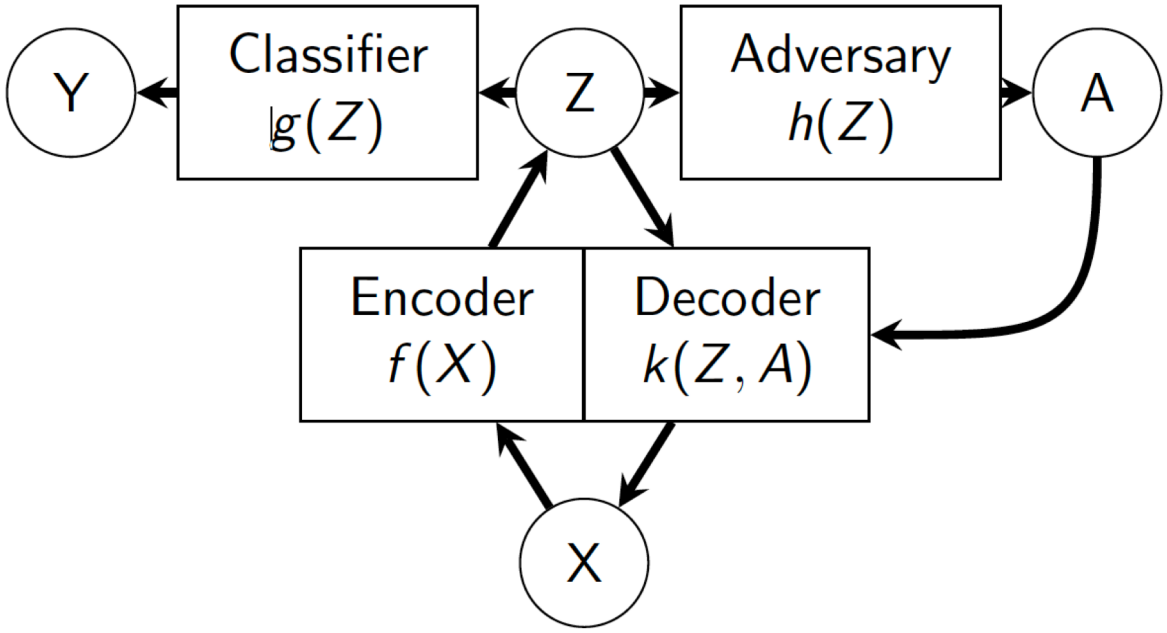
LEARNING ADVERSARIALLY FAIR TRANSFERABLE REPRESENTATIONS

Madras, Creager, Pitassi, Zemel, 2018



- The classifier is indifferent vendor, forcing the encoder to make the representations useful
- The adversary is the malicious vendor, forcing the encoder to hide the sensitive attributes in the representations

ADVERSARIAL LEARNING IN LAFTR

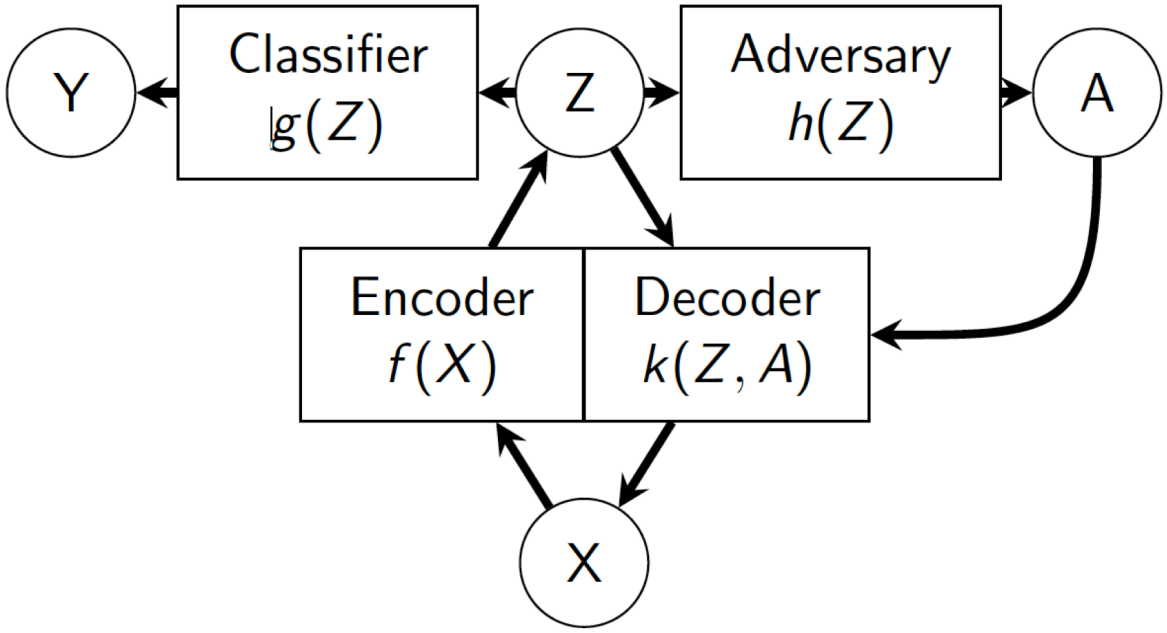


- Our game: encoder-decoder-classifier vs. adversary
- Aim: Learn fair encoder

$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[\mathcal{L}(f,g,h,k)\right]$$

$$\mathcal{L}(f,g,h,k) = \alpha\mathcal{L}_{Class} + \beta\mathcal{L}_{Dec} - \gamma\mathcal{L}_{Adv}$$

ADVERSARIAL OBJECTIVES



Choice of adversarial objective depends on fairness desideratum

- Demographic parity: $\mathcal{L}_{DP}(h) = \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a) \in \mathcal{D}_i} |h(f(x)) - a|$
- Equalized odds: $\mathcal{L}_{EO}(h) = \sum_{i,j \in \{0,1\}^2} \frac{1}{|\mathcal{D}_i^j|} \sum_{(x,a,y) \in \mathcal{D}_i^j} |h(f(x), y) - a|$
- Equal Opportunity: $\mathcal{L}_{EOpp}(h) = \sum_{i \in \{0,1\}} \frac{1}{|\mathcal{D}_i^1|} \sum_{(x,a) \in \mathcal{D}_i^1} |h(f(x)) - a|$

FROM ADVERSARIAL OBJECTIVES TO FAIRNESS DEFINITIONS

In general: pick the right adversarial loss, encourage the right conditional independencies

- Demographic parity encourages $Z \perp A$ to fool adversary
- Equalized odds encourages $Z \perp A \mid Y$ to fool adversary
- Equal opportunity encourages $Z \perp A \mid Y = 1$ to fool adversary

Note that independencies of $Z = f(x)$ also hold for predictions $\hat{Y} = g(Z)$

We show: In the adversarial limit, these objectives guarantee these fairness metrics!

- The key is to connect predictability of A by the adversary $h(Z)$ to unfairness in the classifier $g(Z)$

EXPERIMENTS

Datasets

1. Adult Income

Size: 45,222 instances, 14 attributes

Task: predict whether or not annual income > 50K

Sensitive feature: Gender

2. Heritage Health

Size: 147,473 instances, 139 attributes

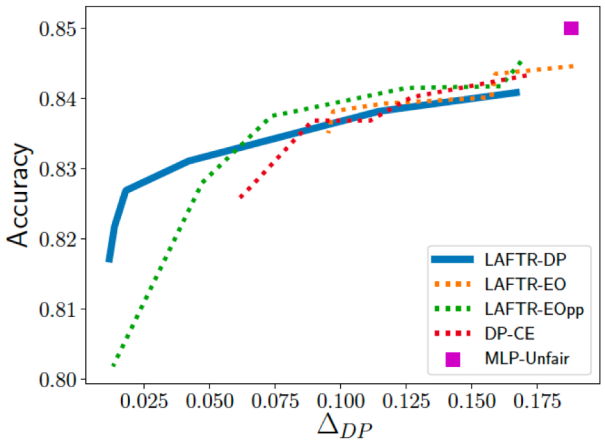
Task: predict patient's Charlson Index (co-morbidity)

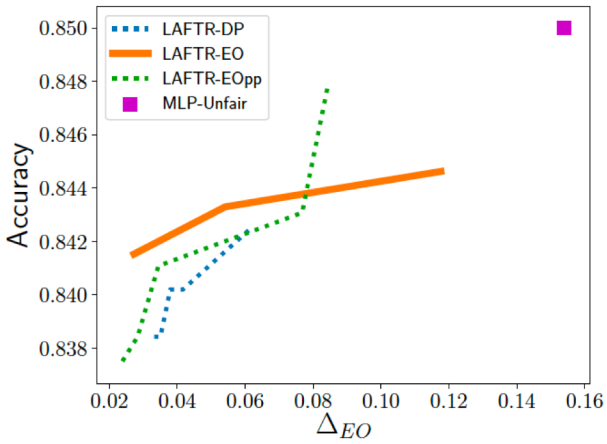
Sensitive feature: Age

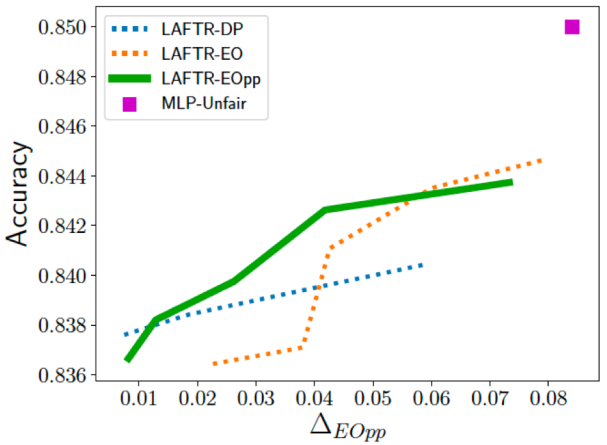
Models

Encoder, classifier, adversary: each single hidden-layer MLP (8; 20 hidden units)

RESULTS: FAIR CLASSIFICATION







- Train with 2-step process to simulate owner → vendor framework
- Tradeoffs between accuracy and fairness metrics produced by different LAFTR loss functions
- Achieves best solutions, wrt fairness-accuracy tradeoff

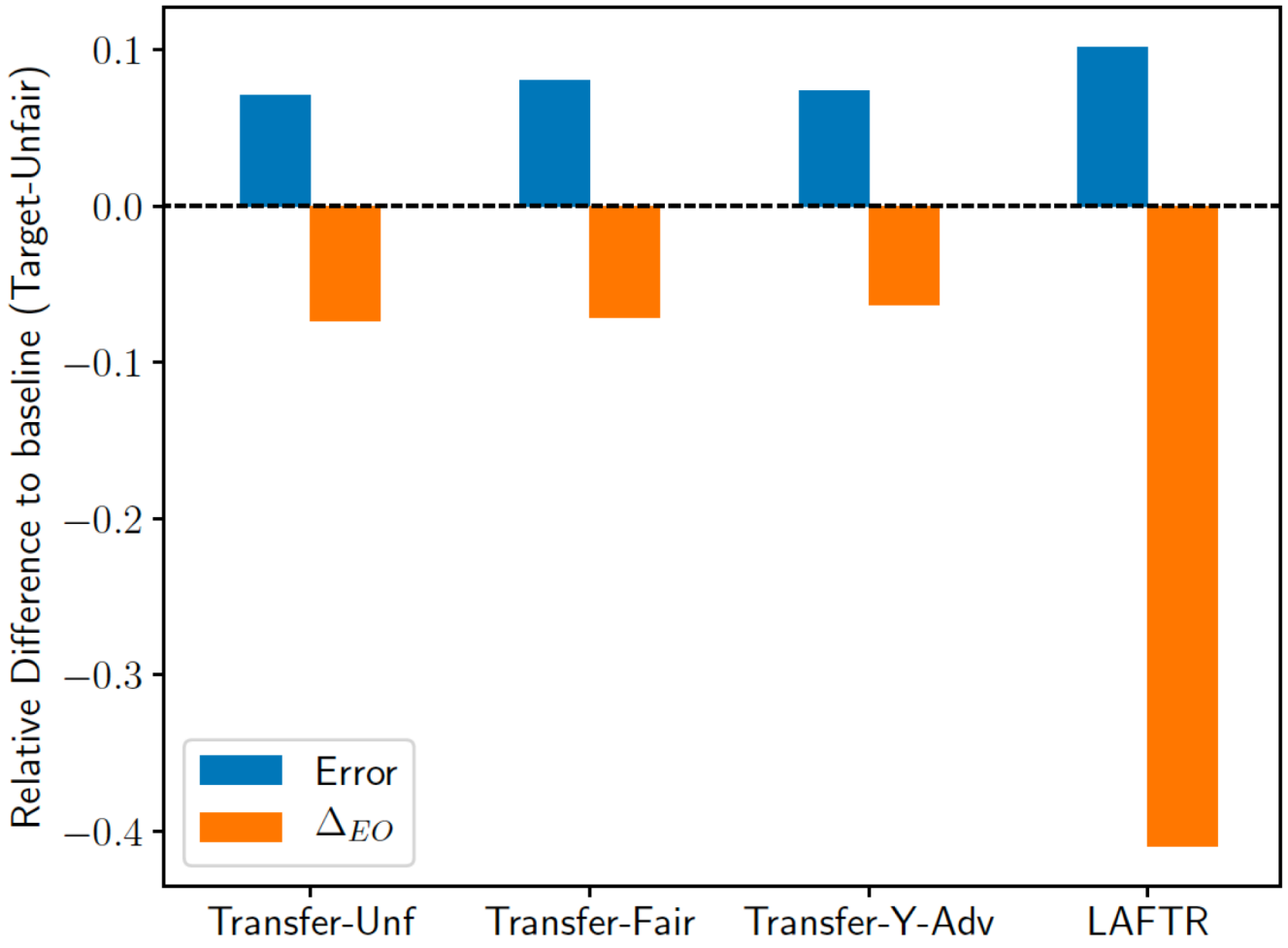
RESULTS: FAIRNESS METRICS

METHOD	Δ_{DP}	Δ_{EO}	Δ_{EOpp}	ACC.
MLP (UNFAIR)	0.381	0.476	0.231	**0.785**
LAFTR-EO	0.152	**0.050**	0.036	0.763
	0.143	**0.052**	0.032	0.752
LAFTR-EOPP	0.087	0.092	**0.010**	0.742
	0.113	0.063	**0.024**	0.735
LAFTR-DP	**0.041**	0.140	0.025	0.731
	0.002	0.196	0.031	0.728

SETUP: FAIR TRANSFER LEARNING

- Downstream vendors will have unknown prediction tasks
- Does fairness transfer?
- We test this as follows:
  1. Train encoder f on data X, with label Y
  2. Freeze encoder f
  3. On new data X', train classifier on top of $f(X')$, with new task label Y'
  4. Observe fairness and accuracy of this new classifier on new task Y'
- Compare LAFTR encoder f to other encoders
- We use Heritage Health dataset
  - Y is Charlson comorbidity index > 0
  - Y' is whether or not a certain type of insurance claim was made
  - Check for fairness w.r.t. age

RESULTS : FAIR TRANSFER LEARNING



Fair transfer learning on Health dataset. Down is better in both metrics.

ALTERNATIVE FORMULATIONS

Rather than an (un)fairness regularizer, can set up as constrained optimization problem

$$\max_{\phi \in \Phi} I_q(\mathbf{x}; \mathbf{z}|\mathbf{u}) \qquad \text{s.t. } I_q(\mathbf{z}; \mathbf{u}) < \epsilon$$

Learning Controllable Fair Representations (2018) by Song et al.

- Hard to compute and optimize these mutual information terms
- Propose tractable approximations, bounds to optimize
- Solve the dual

ALTERNATIVE FORMULATIONS

Another popular approach is to adjust the input data, by removing features or pre-processing

- Data preprocessing techniques for classification without discrimination (2011), Kamiran & Calders
- Certifying and removing disparate impact (2015), Feldman et al.
- Optimized data pre-processing for discrimination prevention, Calmon et al.
- The case for process fairness in learning: Feature selection for fair decision ,aking, Grgić-Hlača et al.

Part 2: Privacy in ML

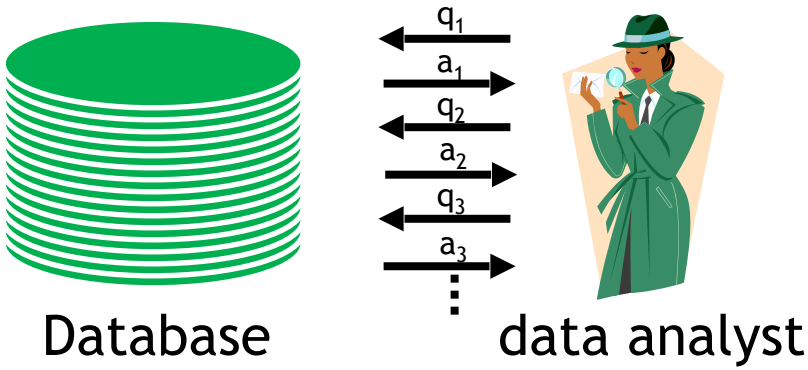
Why Privacy?

Microsoft: tool for diagnosing pancreatic cancer by monitoring Bing queries

Netflix: film recommender algorithm "anonymized"

Model inversion attacks:
train ML model using sensitive information
hackers can invert model to recover very sensitive individual info (credit card number)

Privacy-Preserving Data Analysis

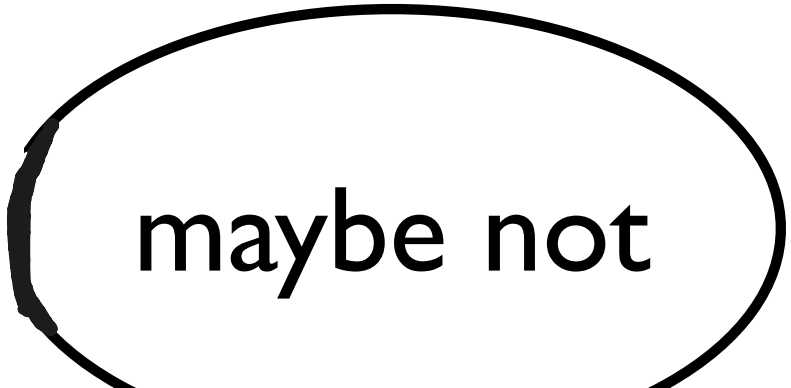


- Census, epidemic detection based on OTC drug purchases; analysis of loan application data for evidence of discrimination,…
- 50+ year old problem

What analyses on a database might violate privacy? What analyses are privacy-preserving?

what to promise?

delete identifying information



Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	...nicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Figure 1: ...e-identifying anonymous data by linking to external data

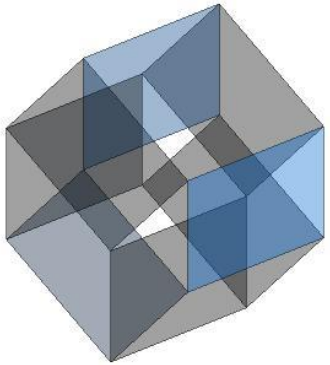
Public voter dataset

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

Curse of Dimensionality

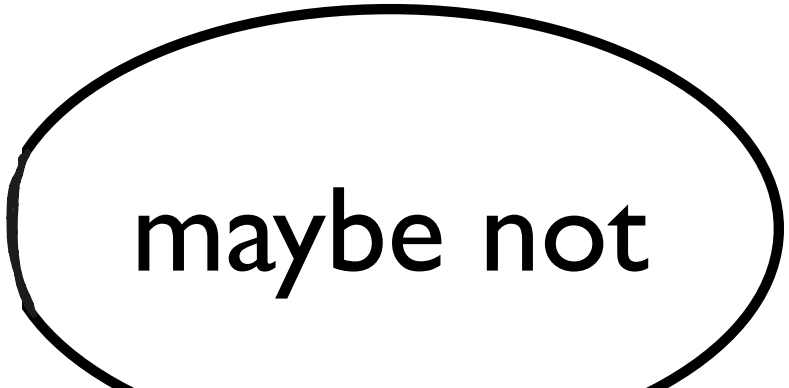
Aggarwal (VLDB 2005)



- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info ⇒ k-anonymized datasets are useless

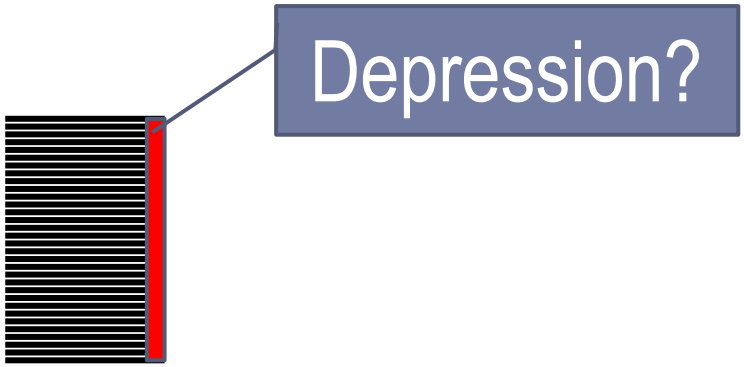
what to promise?

only ask questions that pertain to large populations



The Statistics Masquerade

- Differencing Attack
  - *How many members of House of Representatives have sickle cell trait?*
  - *How many members of House, other than the Speaker, have the trait?*
- Needle in a Haystack
  - Determine presence of an individual's genomic data in GWAS case group
- The Big Bang attack
  - Reconstruct "depression" bit column



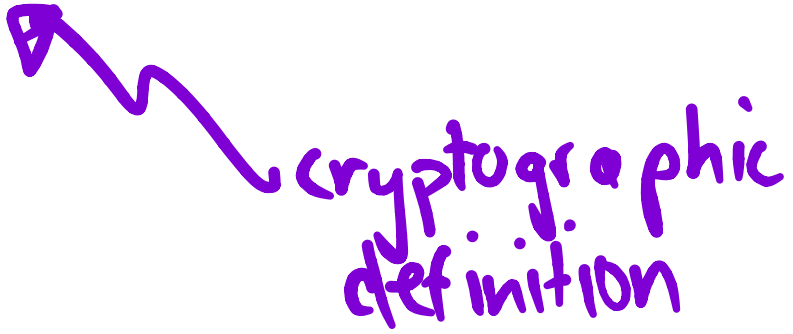
Dinur and Nissim'03 Homer+'08

Fundamental Law of Info Recovery

- "Overly accurate" estimates of "too many" statistics is blatantly non-private.

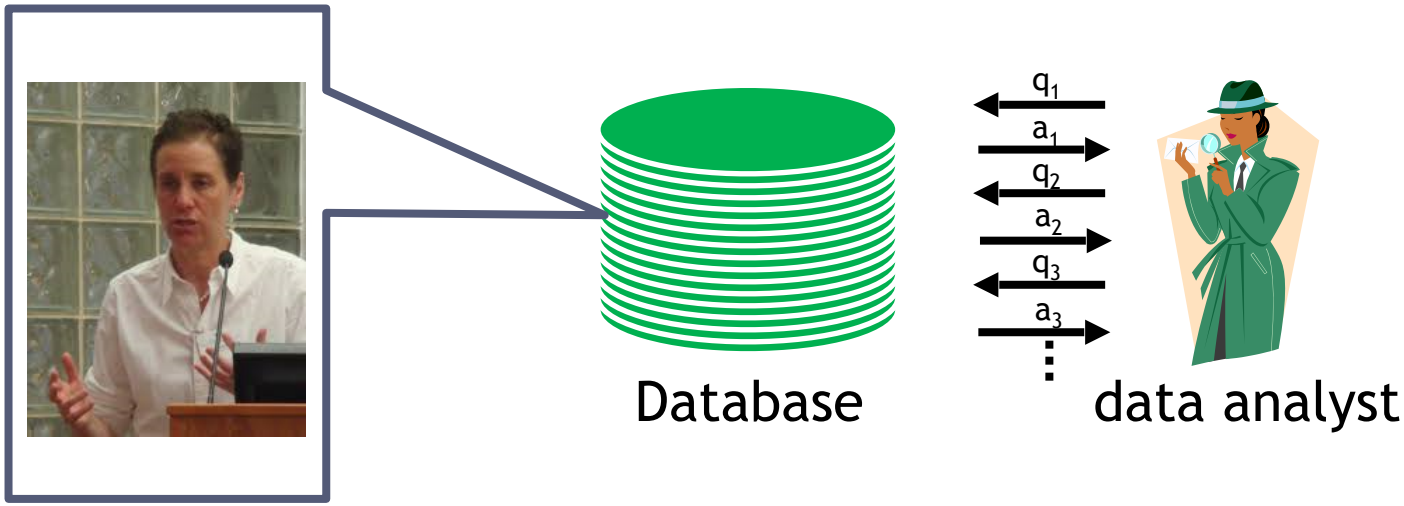
what to promise?

access to the output should not enable one to learn anything about an individual that could not be learned without access



is this desirable?

Privacy-Preserving Data Analysis?

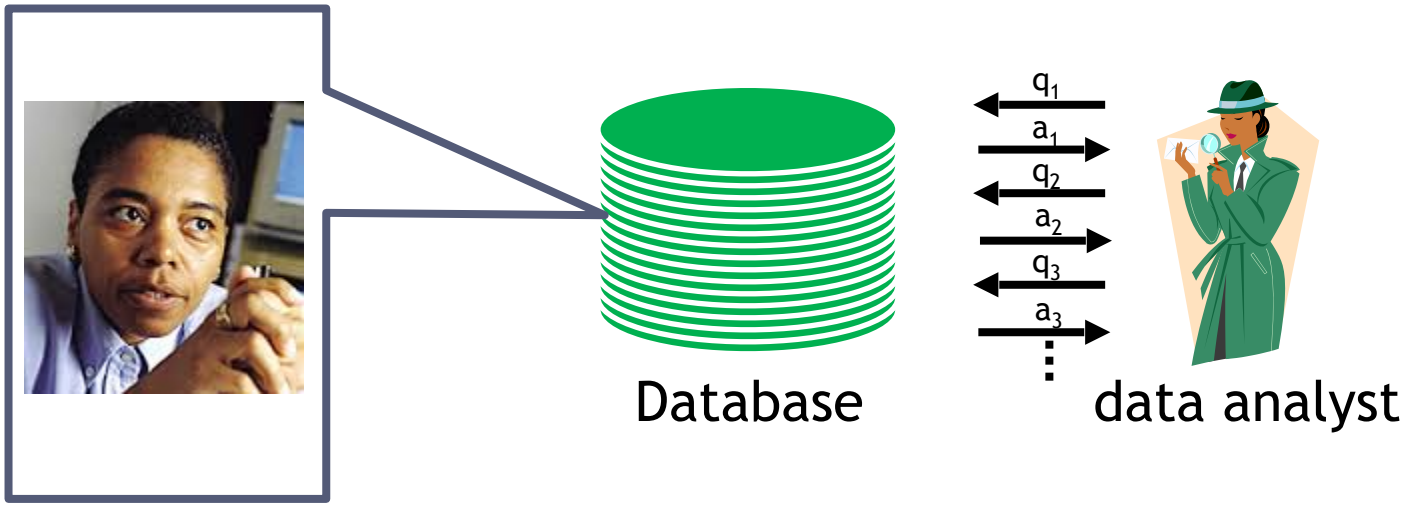


- "Can't learn anything new about Helen"?
- Then what is the point?

what to promise?

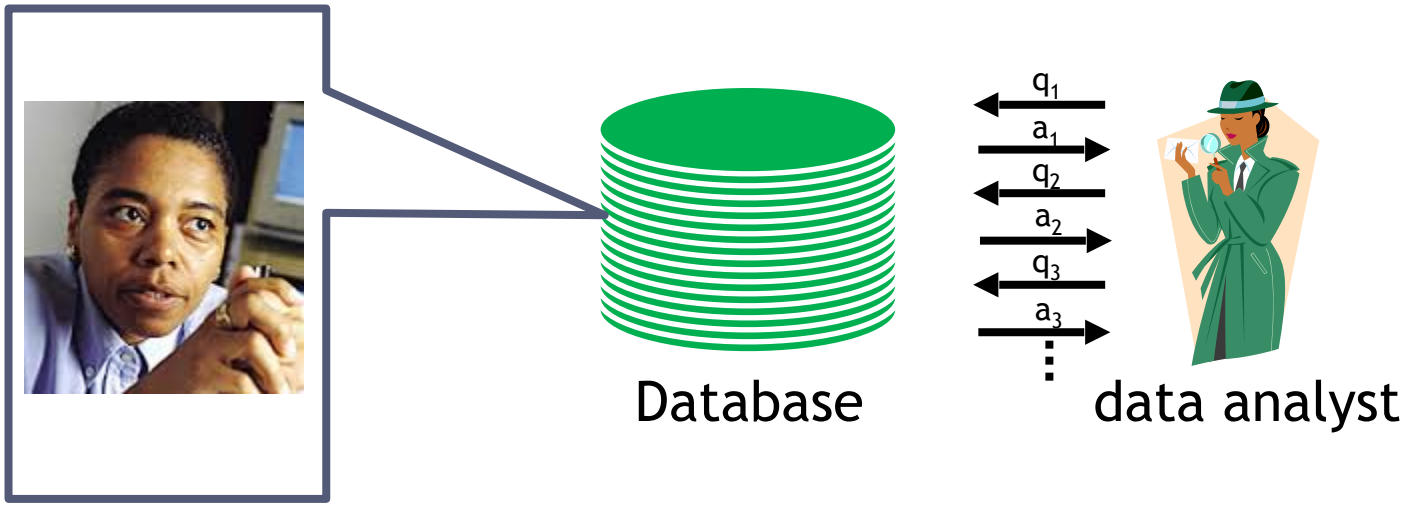
access to the output should not enable one to learn much more about an individual than could be learned via the same analysis omitting that individual from the database

Privacy-Preserving Data Analysis?



- Ideally: learn same things if Helen is replaced by another random member of the population ("stability")

Privacy-Preserving Data Analysis?



- Stability preserves Helen's privacy AND prevents over-fitting
- Privacy and Generalization are aligned!

statistical database model

X set of possible entries/rows

one row per person

database x a set of rows; $x \in \mathbb{N}^{|X|}$

(histogram)

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N

neighboring databases

what's a small change?

require nearly identical behavior on neighboring databases differing by the addition or removal of a single row:

$$||x - y||_1 \leq 1$$

for $x, y \in \mathbb{N}^{|X|}$

differential privacy

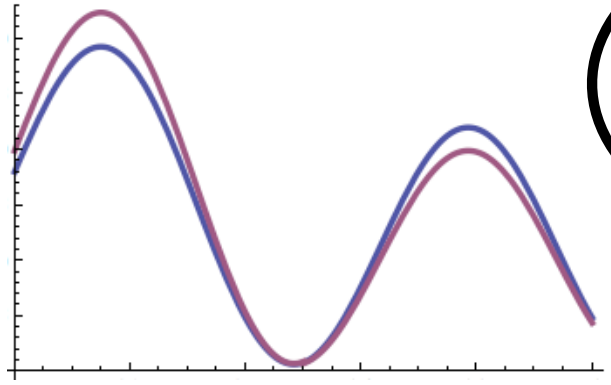
[DinurNissim03, DworkNissimMcSherrySmith06, Dwork06]

ε-Differential Privacy for algorithm M:

for any two neighboring data sets x_1, x_2, differing by the addition or removal of a single row

any $S \subseteq$ range(M),

$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$

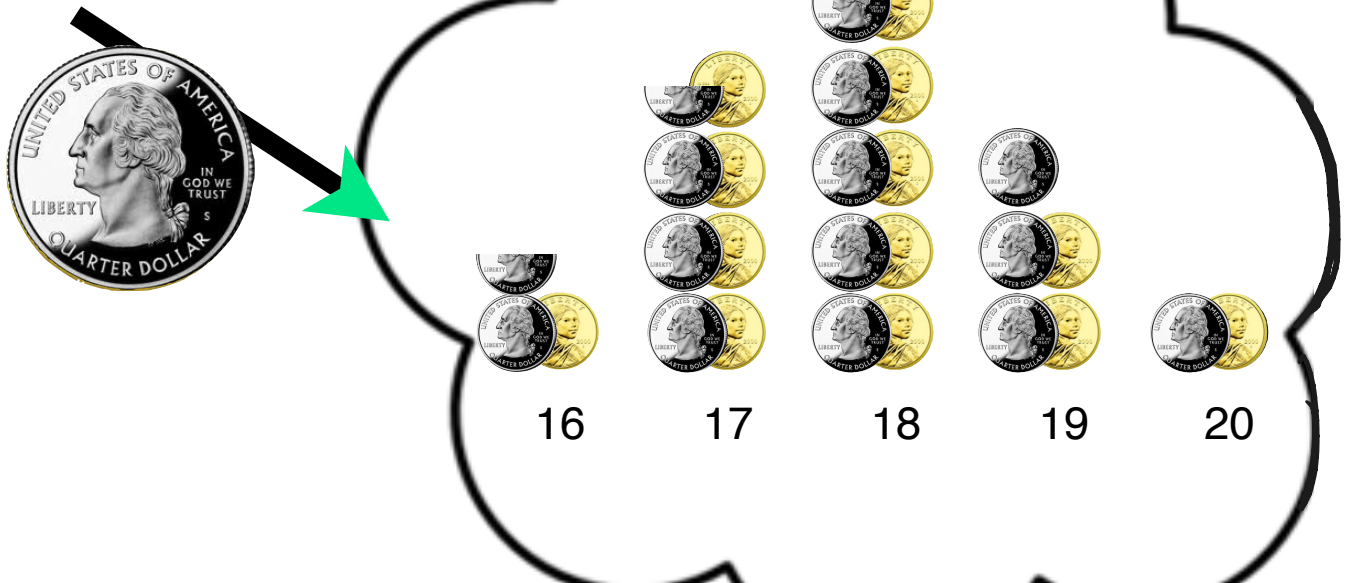


$e^{\varepsilon} \sim (1 + \varepsilon)$

differential privacy

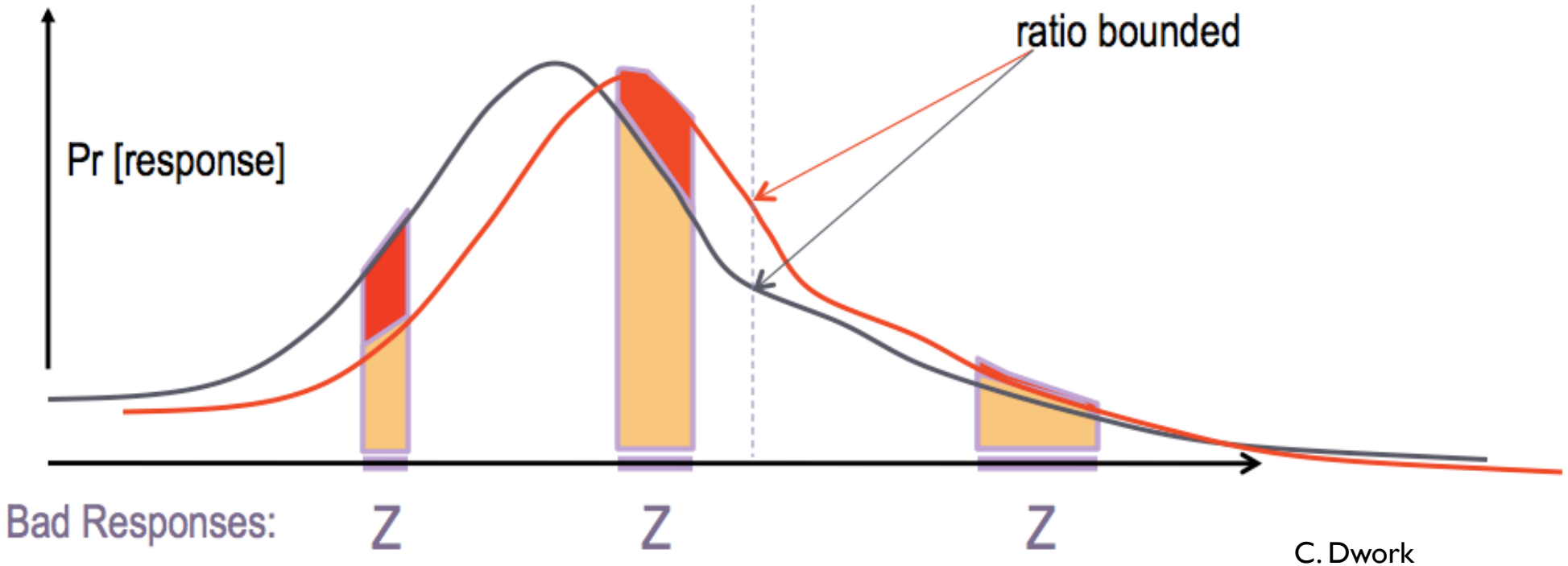
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
~~Ellen Jones~~	~~4/24/59~~	~~F~~	~~160~~	~~Y~~	~~Y~~
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



differential privacy

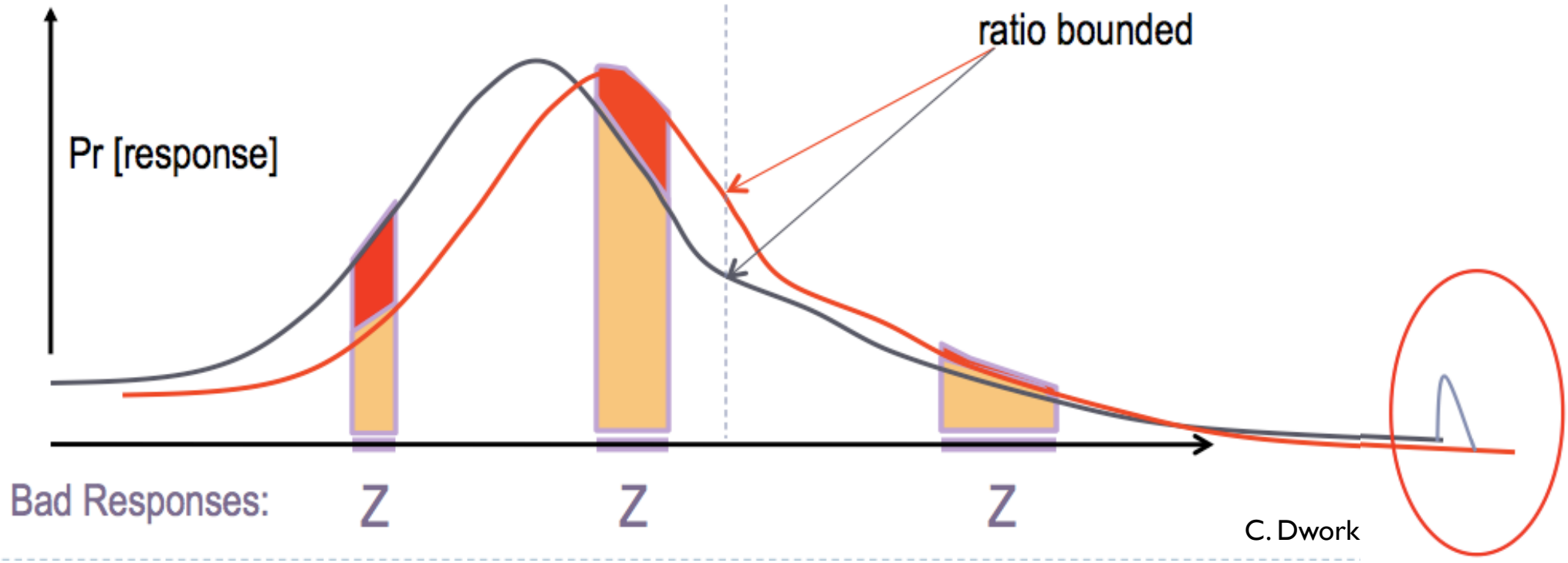
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$



C. Dwork

(ε, δ)-differential privacy

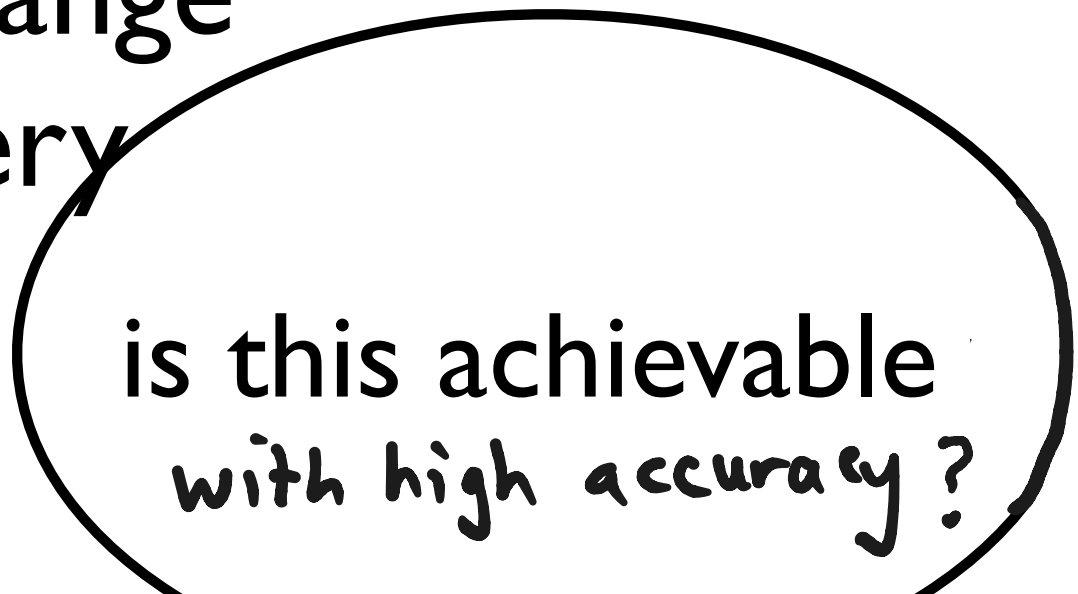
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S] + \delta$$



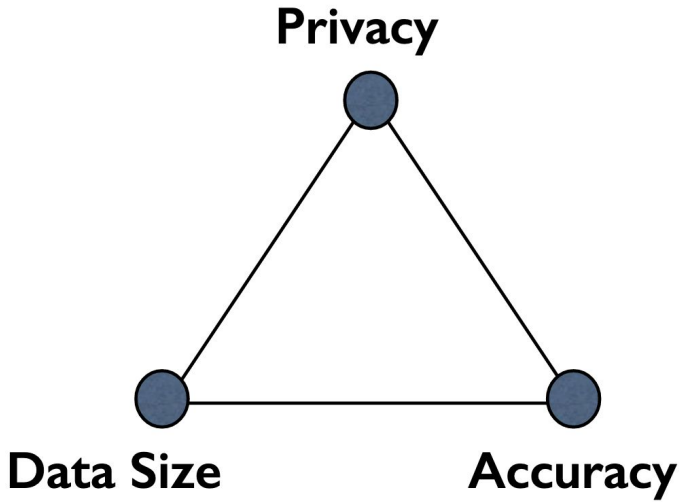
differential privacy

$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

promise: if you leave the database, no outcome will change probability by very much

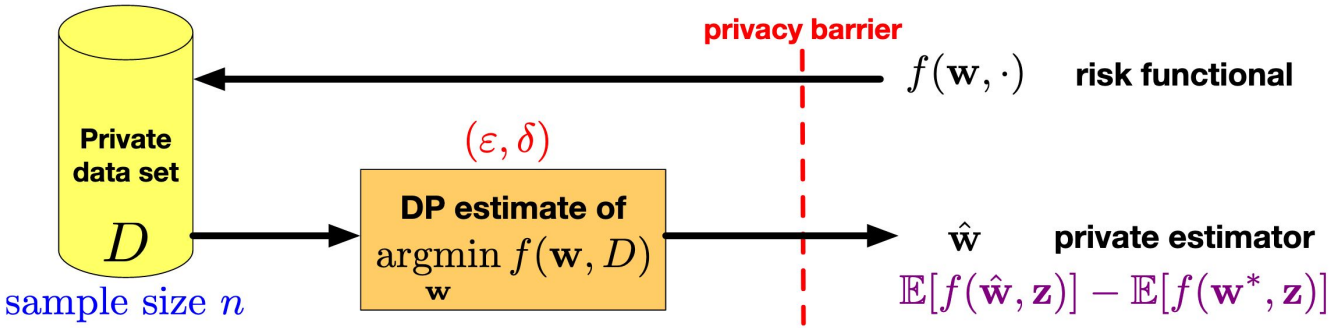


Tradeoffs in DP+ML



[Chaudhuri & Sarwate 2017 NIPS tutorial]

Tradeoffs in DP+ML



Statistical estimation: estimate a parameter or predictor using private data that has good expected performance on future data.

[Chaudhuri & Sarwate 2017 NIPS tutorial]

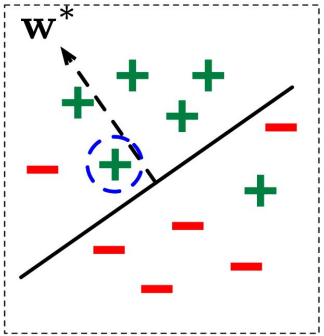
Goal: Good privacy-accuracy-sample size tradeoff

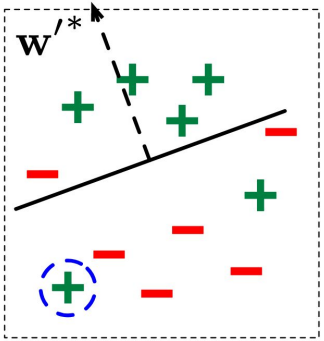
Private Empirical Risk Min.

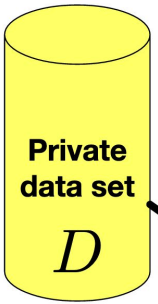
$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i)) + \lambda R(\mathbf{w})$$

- *Empirical Risk Minimization* (ERM) is a common paradigm for prediction problems.
- Produces a predictor **w** for a label/response *y* given a vector of features/covariates **x**.
- Typically use a convex loss function and regularizer to "prevent overfitting."

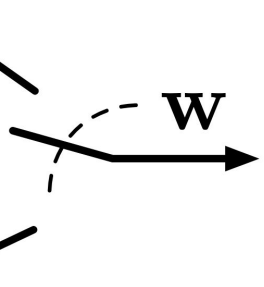
Private ERM







easy for adversary to tell the difference between D and D'



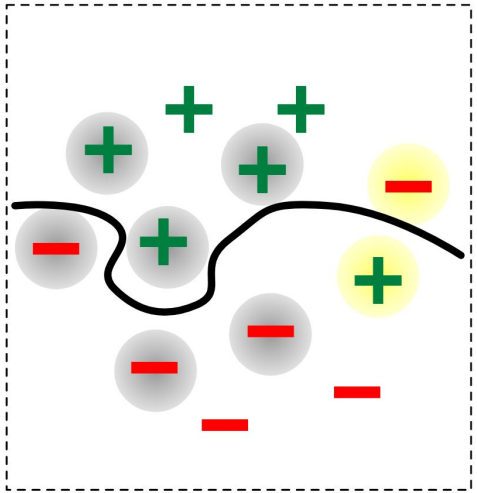
[CMS11, RBHT12]

[Chaudhuri & Sarwate 2017 NIPS tutorial]

Kernel Approaches Even Worse

- Kernel-based methods produce a classifier that is a function of the data points.
- Even adversary with black-box access to w could potentially learn those points.

$$\mathbf{w}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$



[Chaudhuri & Sarwate 2017 NIPS tutorial]

[CMS11]

Privacy & Learning Compatible

- Good learning algorithms *generalize* to the population distribution, not individuals.
- *Stable learning algorithms* generalize [BE02].
- Differential privacy can be interpreted as a form of stability that also implies generalization [DFH+15,BNS+16].
- Two parts of the same story:
  Privacy implies generalization asymptotically.
  Tradeoffs between privacy-accuracy-sample size for finite *n*.

[Chaudhuri & Sarwate 2017 NIPS tutorial]

Revisiting ERM

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i)) + \lambda R(\mathbf{w})$$

- Learning using (convex) optimization uses three steps:

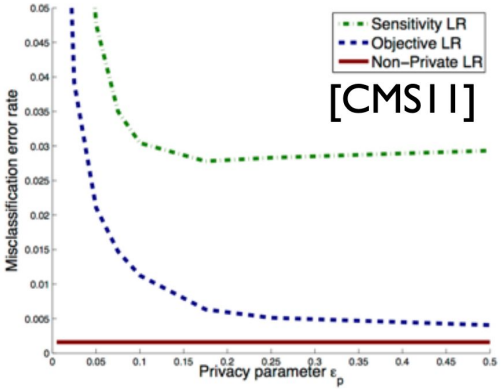
1. read in the data — **input perturbation**
2. form the objective function — **objective perturbation**
3. perform the minimization — **output perturbation**

We will discuss this one -->

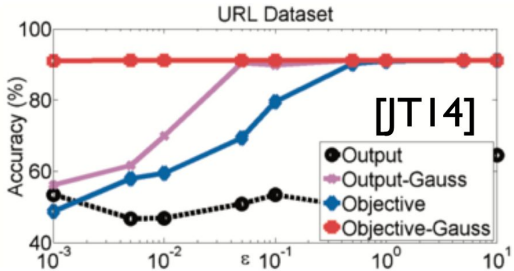
- We can try to introduce privacy in each step!

[Chaudhuri & Sarwate 2017 NIPS tutorial]

Typical Empirical Results



(a) Regularized logistic regression, KDDCup99

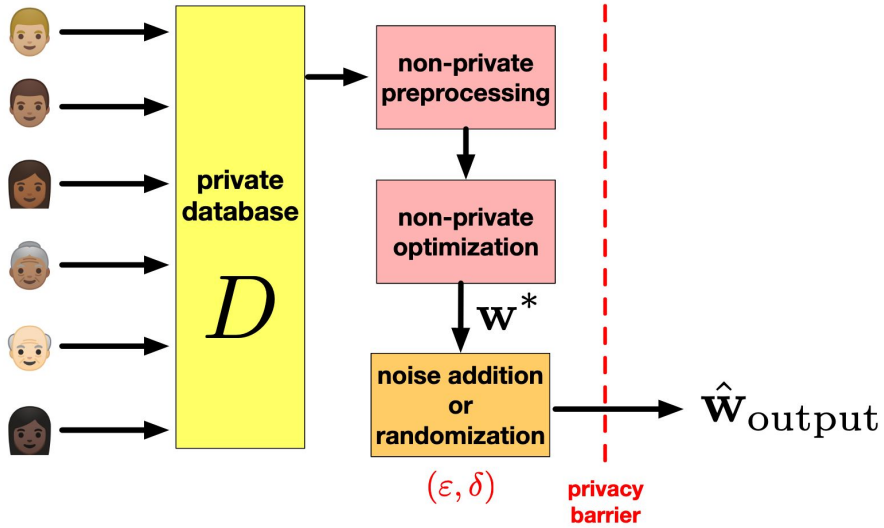


In general:

- **Objective** perturbation empirically outperforms output perturbation.
- **Gaussian** mechanism with (ε, δ) guarantees outperform **Laplace** - like mechanisms with ε-guarantees.
- **Loss vs. non-private methods** is very dataset-dependent.

[Chaudhuri & Sarwate 2017 NIPS tutorial]

Output Perturbation



- Compute the minimizer and add noise.
- Does not require re-engineering baseline algorithms

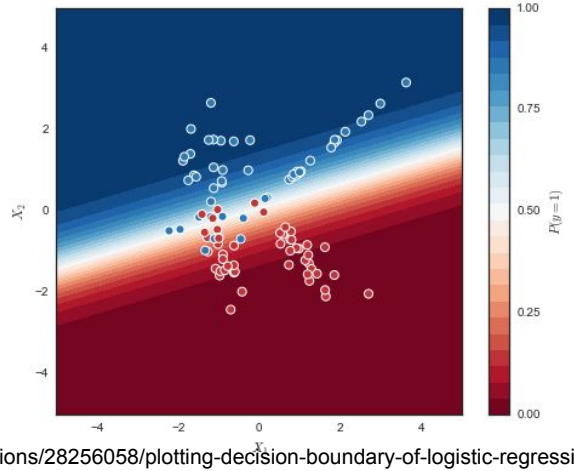
Noise depends on the sensitivity of the argmin.

[CMS11, RBHT12]

Private Logistic Regression

$$L(w) = \sum_{i=1}^{N} \text{CrossEntropy}(\sigma(w^T x_i), y_i) + \frac{1}{2}\lambda w^T w$$

What is $\Delta L(w)$, the sensitivity of the loss?



https://stackoverflow.com/questions/28256058/plotting-decision-boundary-of-logistic-regression

Private Logistic Regression

$$L(w) = \sum_{i=1}^{N} \underbrace{\text{CrossEntropy}(\sigma(w^T x_i), y_i)}_{\text{convex}} + \underbrace{\frac{1}{2}\lambda w^T w}_{\lambda-\text{strongly convex}}$$

$h(w)$ is convex iff $\nabla^2 h(w) \succcurlyeq 0$ (its Hessian is positive semi-definite)

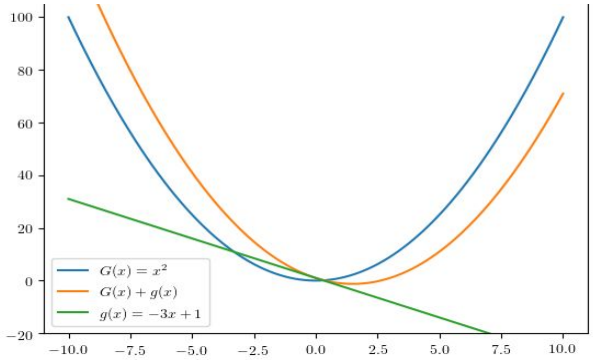
$h(w)$ is $\lambda-$strongly convex iff $(\nabla h(x) - \nabla h(y))^T (x - y) \geq \lambda ||x - y||_2^2$

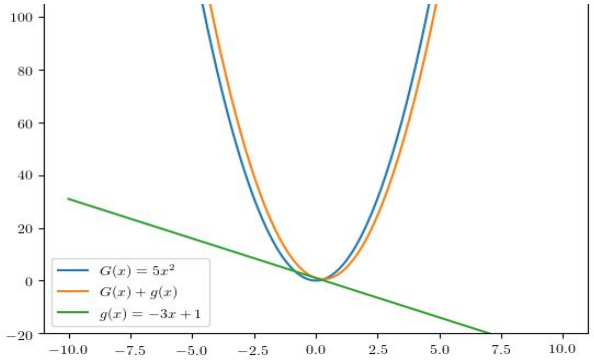
The sum of a convex function and a λ-strongly convex function is λ-strongly convex.

Private Logistic Regression

Lemma 7 *Let $G(\mathbf{f})$ and $g(\mathbf{f})$ be two vector-valued functions, which are continuous, and differentiable at all points. Moreover, let $G(\mathbf{f})$ and $G(\mathbf{f}) + g(\mathbf{f})$ be λ-strongly convex. If $\mathbf{f}_1 = \operatorname{argmin}_{\mathbf{f}} G(\mathbf{f})$ and $\mathbf{f}_2 = \operatorname{argmin}_{\mathbf{f}} G(\mathbf{f}) + g(\mathbf{f})$, then*

$$\|\mathbf{f}_1 - \mathbf{f}_2\| \leq \frac{1}{\lambda} \max_{\mathbf{f}} \|\nabla g(\mathbf{f})\| .$$





Private Logistic Regression

$h(w)$ is $\lambda-$strongly convex iff $(\nabla h(x) - \nabla h(y))^T (x - y) \geq \lambda ||x - y||_2^2$

The sum of a convex function and a λ-strongly convex function is λ-strongly convex.

$$L(w) = \frac{1}{N} \sum_{i \in \{1..N\}} \text{CrossEntropy}(\sigma(w^T x_i), y_i) + \frac{1}{2} \lambda w^T w$$

$$L'(w) = \left(\frac{1}{N} \sum_{i \in \{1..N\} \setminus j} \text{CrossEntropy}(\sigma(w^T x_i), y_i) \right) + \text{CrossEntropy}(\sigma(w^T x_{j'}), y_{j'}) + \frac{1}{2} \lambda w^T w$$

$$L'(w) = L(w) + g(w)$$

$$g(w) = \frac{1}{N} \left(\text{CrossEntropy}(\sigma(w^T x_{j'}), y_{j'}) - \text{CrossEntropy}(\sigma(w^T x_j), y_j) \right)$$

Assuming $||x|| \leq 1$, we have $\nabla g(w) \leq \frac{2}{N}$

$$\rightarrow \Delta L = \frac{2}{N\lambda}$$

Private Logistic Regression

Algorithm:

sensitivity $\Delta L = \frac{2}{N}$

1) solve $w^* = \arg\min_w L(w)$

2) draw η with $p(\eta) \propto \exp(-\frac{\Delta L}{\epsilon}||\eta||)$ (vector analogue of Laplace draw)

3) return $w = w^* + \eta$

Private Non-Convex Learning

For non-convex losses:

- We don't know whether our optimizer will (asymptotically) find the global optimum
- We can not bound the loss function at the optimum or anywhere else

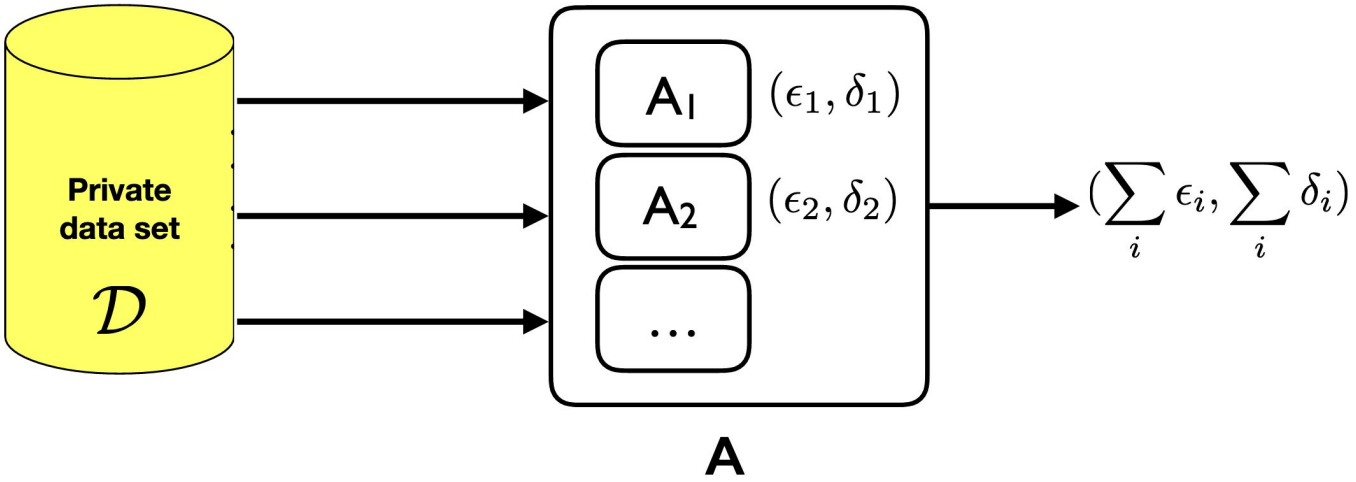
Instead we will make every step of the iterative optimization algorithm private, and somehow account for the overall privacy loss at the end

Private Non-Convex Learning

To discuss DPSGD algorithm, we need to understand

- Basic composition of mechanisms
- “Advanced” composition in an adaptive setup
- Privacy amplification by subsampling
- Per-example gradient computation

Recall Composition Theorem



Total privacy loss is the sum of privacy losses

(Better composition possible — coming up later)

Advanced Composition

What is composition?

- Repeated use of DP algorithms on the same database
- Repeated use of DP algorithms on the different databases that nevertheless may contain shared information about individuals

We can show that the privacy loss over all possible outcomes has a Markov structure, which hints at a better composition

Advanced Composition

Theorem 3.20 (Advanced Composition)**.** For all $\varepsilon, \delta, \delta' \geq 0$, the class of (ε, δ)-differentially private mechanisms satisfies $(\varepsilon', k\delta + \delta')$-differential privacy under k-fold adaptive composition for:

$$\varepsilon' = \sqrt{2k \ln(1/\delta')}\varepsilon + k\varepsilon(e^{\varepsilon} - 1).$$

Privacy by Subsampling

Lemma 3 (Amplification via sampling) If A is 1 –differentially private, then for any $\epsilon \in (0,1)$, $A'(\epsilon, \cdot)$ is 2ϵ –differentially private.

Suppose A is a 1 –differentially private algorithm that expects data sets from a domain D as input. Consider a new algorithm A', which runs A on a random subsample of $\approx \epsilon n$ points from its input:

Proof: Fix an event S in the output space of A', and two data sets x, x' that differ by a single individual, say $x = x' \cup \{i\}$.

Consider a run of A' on input x. If i is not included in the sample T, then the output is distributed the same as a run of A' on $x' = x \setminus \{i\}$, since the inclusion of i in the sample is independent of the inclusion of other elements. On the other hand, if i is included in the sample T, then the behavior of A on T is only a factor of e off from the behavior of A on $T \setminus \{i\}$. Again, because of independence, the distribution of $T \setminus \{i\}$ is the same as the distribution of T conditioned on the omission of i. For a set $T \subseteq D$, let p_T denote the distribution of $A(T)$. In symbols, we have that for any event S:

$$p_x(S \mid i \notin T) = p_{x'}(S) \quad \text{and} \quad p_x(S \mid i \in T) \in e^{\pm 1} p_{x'}(S).$$

We can put the pieces together, using the fact that i is in T with probability only ϵ:

$$\begin{aligned} p_x(S) &= (1-\epsilon) \cdot p_x(S \mid i \notin T) + \epsilon \cdot p_x(S \mid i \in T) \\ &\leq (1-\epsilon) \cdot p_{x'}(S) + \epsilon \cdot e \cdot p_{x'}(S) \\ &= (1 + \epsilon(e-1)) p_{x'}(S) \\ &\leq \exp(2\epsilon) \cdot p_{x'}(S) \end{aligned}$$

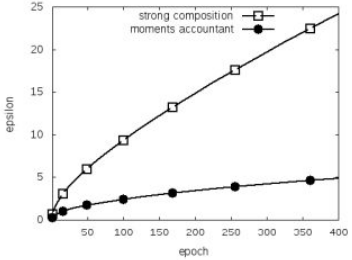
We can get a similar lower bound:

$$\begin{aligned} p_x(S) &= (1-\epsilon) \cdot p_x(S \mid i \notin T) + \epsilon \cdot p_x(S \mid i \in T) \\ &\geq (1-\epsilon) \cdot p_{x'}(S) + \epsilon \cdot \tfrac{1}{e} \cdot p_{x'}(S) \\ &= (1 - \epsilon(1 - e^{-1})) \cdot p_{x'}(S) \\ &\geq \exp(-\epsilon) \cdot p_{x'}(S) \end{aligned}$$

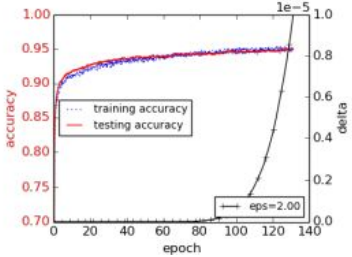
The last inequality uses the fact that $\epsilon \leq 1$. □

https://adamdsmith.wordpress.com/2009/09/02/sample-secrecy/

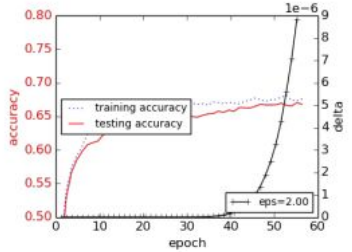
Moments accountant improves bounds



MNIST epoch vs accuracy/privacy



CIFAR-10 epoch vs accuracy/privacy



Differentially Private SGD

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

[Abadi et al. 2016]

Guarantees final parameters don't depend too much on individual training examples

Gaussian noise added to the parameter update at every iteration

Privacy loss accumulates over time

The "moments accountant" provides better empirical bounds on (ε,δ)

Differentially Private SGD

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

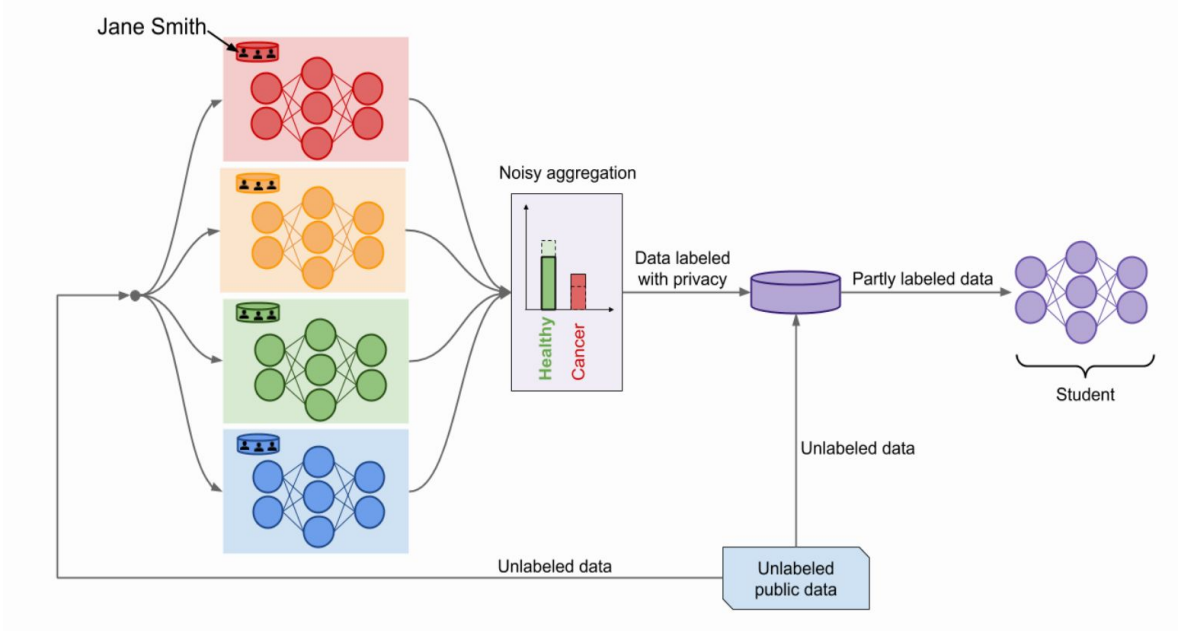
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

← when can we efficiently compute per-example gradients?

[Abadi et al. 2016]

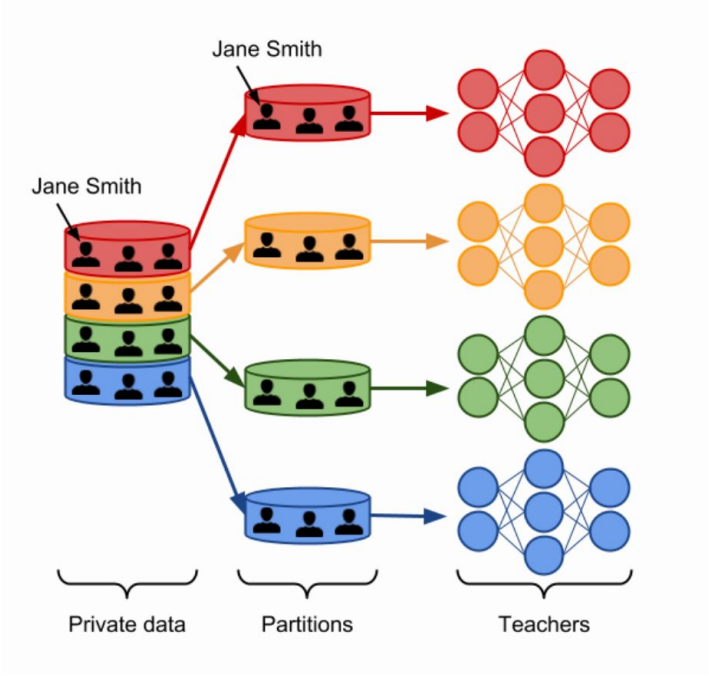
PATE



Private Aggregation of Teacher Ensembles [Papernot et al 2017, Papernot et al 2018]

Key idea: instead of adding noise to gradients, add noise to *labels*

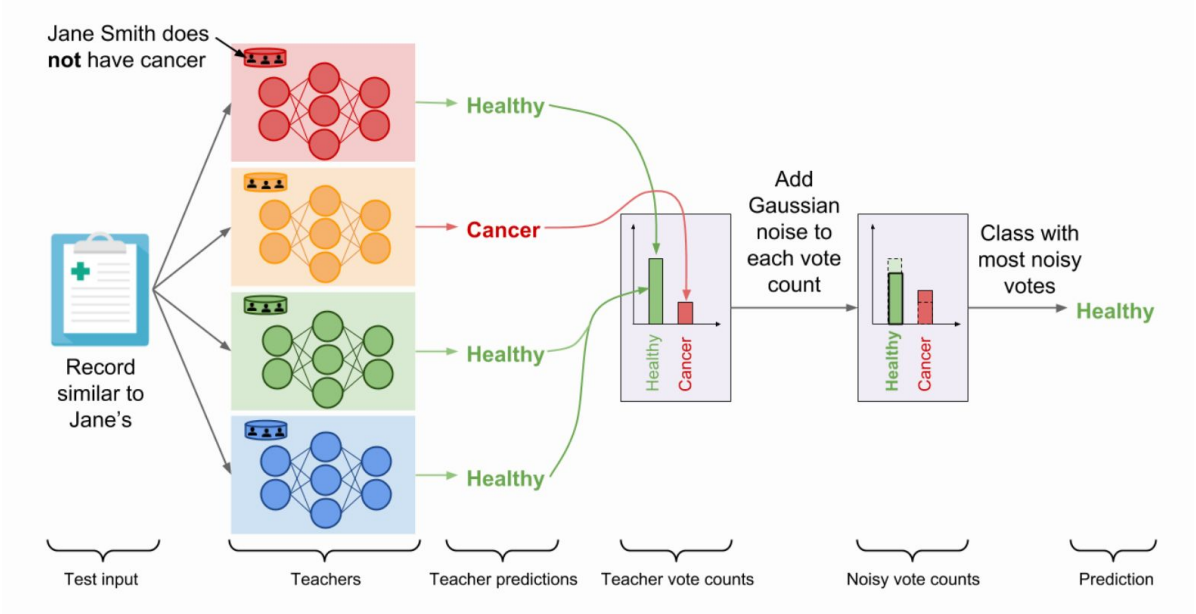
PATE



Start by partitioning private data into disjoint sets

Each teacher trains (non-privately) on its corresponding subset

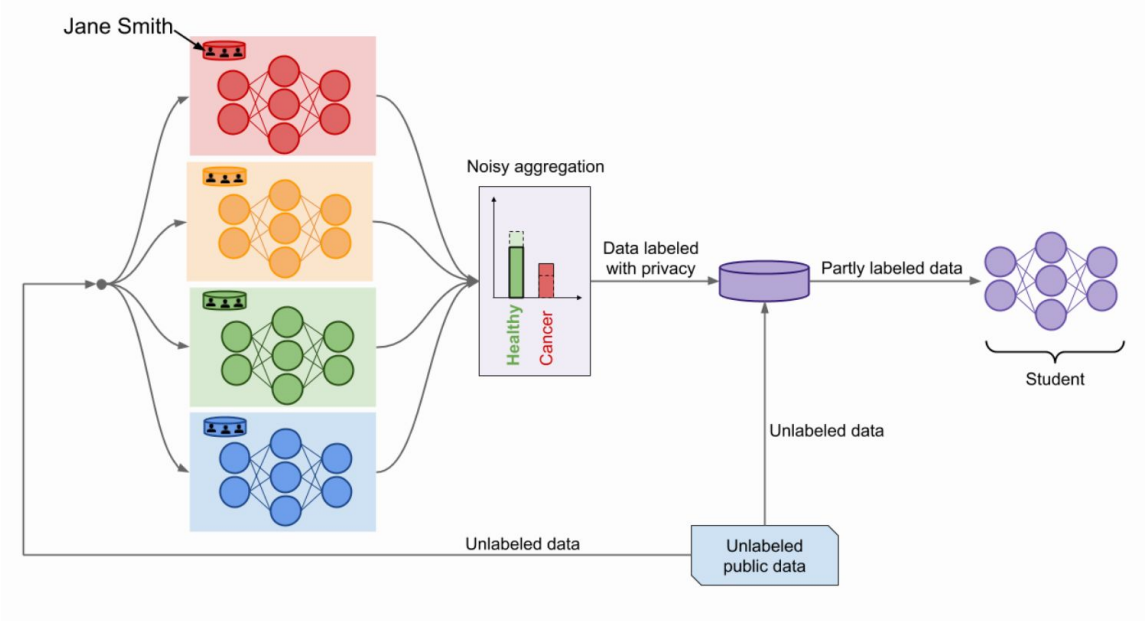
PATE



Private predictions can now be generated via the exponential mechanism, where the "score" is computed with an election amongst teachers - output the noisy winner

We now have private inference, but we lose privacy every time we predict. We would like the privacy loss to be constant at test time.

PATE

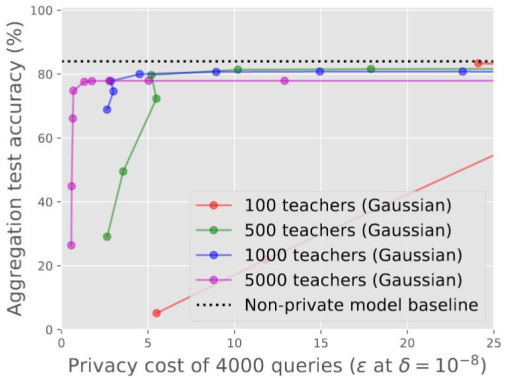


We can instead use the noisy labels provided by the teachers to train a student

We leak privacy during training but at test time we lose no further privacy (due to post-processing thm)

Because the student should use as few labels as possible, unlabeled public data is leveraged in a semi-supervised setup.

PATE



Dataset	Aggregator	Queries answered	Privacy bound ε	Accuracy Student	Accuracy Baseline
MNIST	LNMax (Papernot et al., 2017)	100	2.04	98.0%	99.2%
	LNMax (Papernot et al., 2017)	1,000	8.03	98.1%	
	Confident-GNMax (T=200, σ_1=150, σ_2=40)	286	**1.97**	**98.5%**	
SVHN	LNMax (Papernot et al., 2017)	500	5.04	82.7%	92.8%
	LNMax (Papernot et al., 2017)	1,000	8.19	90.7%	
	Confident-GNMax (T=300, σ_1=200, σ_2=40)	3,098	**4.96**	**91.6%**	
Adult	LNMax (Papernot et al., 2017)	500	2.66	83.0%	85.0%
	Confident-GNMax (T=300, σ_1=200, σ_2=40)	524	**1.90**	**83.7%**	
Glyph	LNMax	4,000	4.3	72.4%	82.2%
	Confident-GNMax (T=1000, σ_1=500, σ_2=100)	10,762	2.03	**75.5%**	
	Interactive-GNMax, two rounds	4,341	**0.837**	73.2%	

https://arxiv.org/pdf/1802.08908.pdf

Thanks!