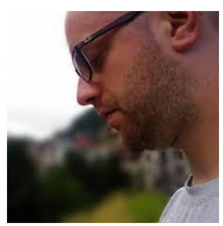
ISTITUTO ITALIANO
DI TECNOLOGIA
COMPUTATIONAL STATISTICS
AND MACHINE LEARNING

Enhancing Multitask Learning with Fairness and Privacy Constraints

Massimiliano Pontil

Istituto Italiano di Tecnologia (IIT), Italy
&
University College London, UK

Joint work with



Luca Oneto
(Univ. of Genoa)

Michele Donini
(Amazon AWS)

Plan

- Empirical risk minimization under fairness constraints
- Fairness and multitask learning (MTL)
- Privacy and MTL
- Hyper-parameter optimization and adaptive data analysis

Based on the papers:

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, & M. P. **Empirical Risk Minimization Under Fairness Constraints.** (To appear in NIPS 2018)

L. Oneto, M. Donini, A. Elders, & M. P. **Taking Advantage of Multitask Learning for Fair Classification.** (Submitted)

Need for Fairness and Privacy in AI

ISTITUTO ITALIANO DI TECNOLOGIA
COMPUTATIONAL STATISTICS AND MACHINE LEARNING

theguardian

UK world sport football opinion culture business lifestyle fashion environment tech travel ≡ browse all sections

home › money › careers property savings pensions borrowing

Inequality

Rise of the racist robots – how AI is learning all our worst impulses

There is a saying in computer science: garbage in, garbage out. When we feed machines data that reflects our prejudices, they mimic them – from antisemitic chatbots to racially biased software. Does a horrifying future await people forced to live at the mercy of algorithms?

Will GDPR Make Machine Learning Illegal?

Mar 2018 Gold KDnuggets Blog

Does GDPR require Machine Learning algorithms to explain their output? Probably not, but experts disagree and there is enough ambiguity to keep lawyers busy.

Machine Learning Journal

Learning with Privacy at Scale

Vol. 1, Issue 8 • December 2017
by Differential Privacy Team

Science Home News Journals Topics Careers

SHARE

REPORT

The reusable holdout: Preserving validity in adaptive data analysis

Cynthia Dwork[1,*], Vitaly Feldman[2,*], Moritz Hardt[3,*], Toniann Pitassi[4,*], Omer Reingold[5,*], Aaron Roth[6,*]

[1]Microsoft Research, Mountain View, CA 94043, USA.
[2]IBM Almaden Research Center, San Jose, CA 95120, USA.
[3]Google Research, Mountain View, CA 94043, USA.
[4]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, Canada.
[5]Samsung Research America, Mountain View, CA 94043, USA.
[6]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

POLICY & LAW \ US & WORLD \ TECH

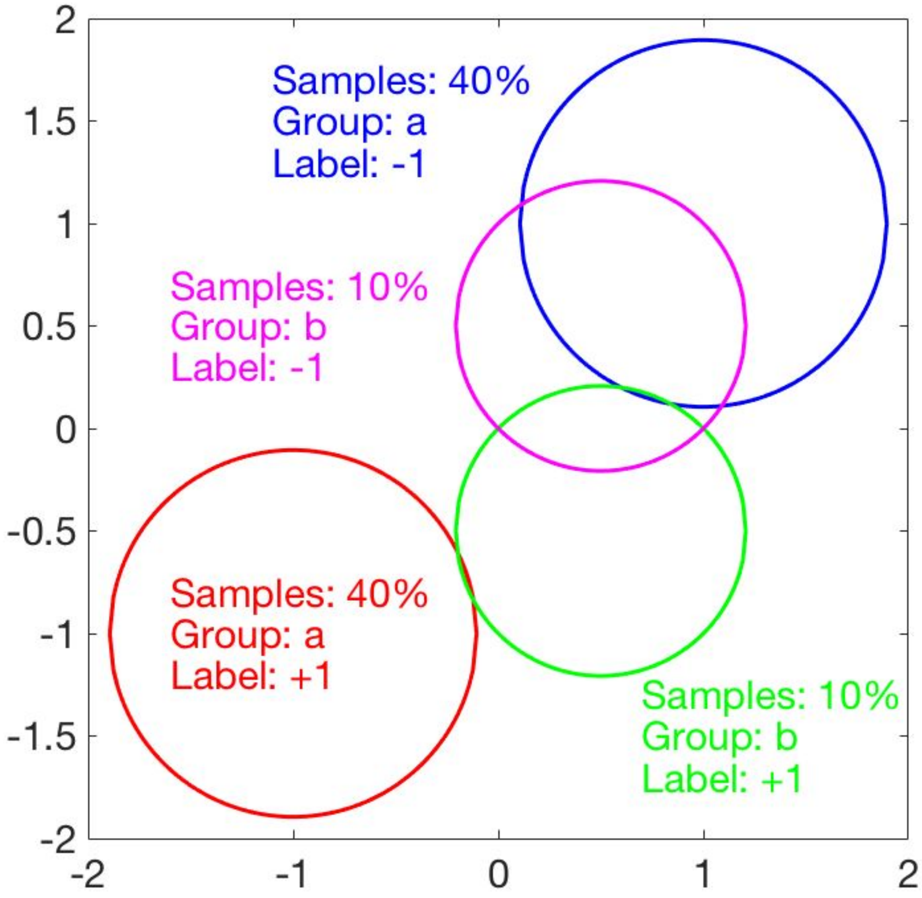
The UK says it can't lead on AI spending, so will have to lead on AI ethics instead

A new report from the House of Lords says the UK could help develop international norms for AI

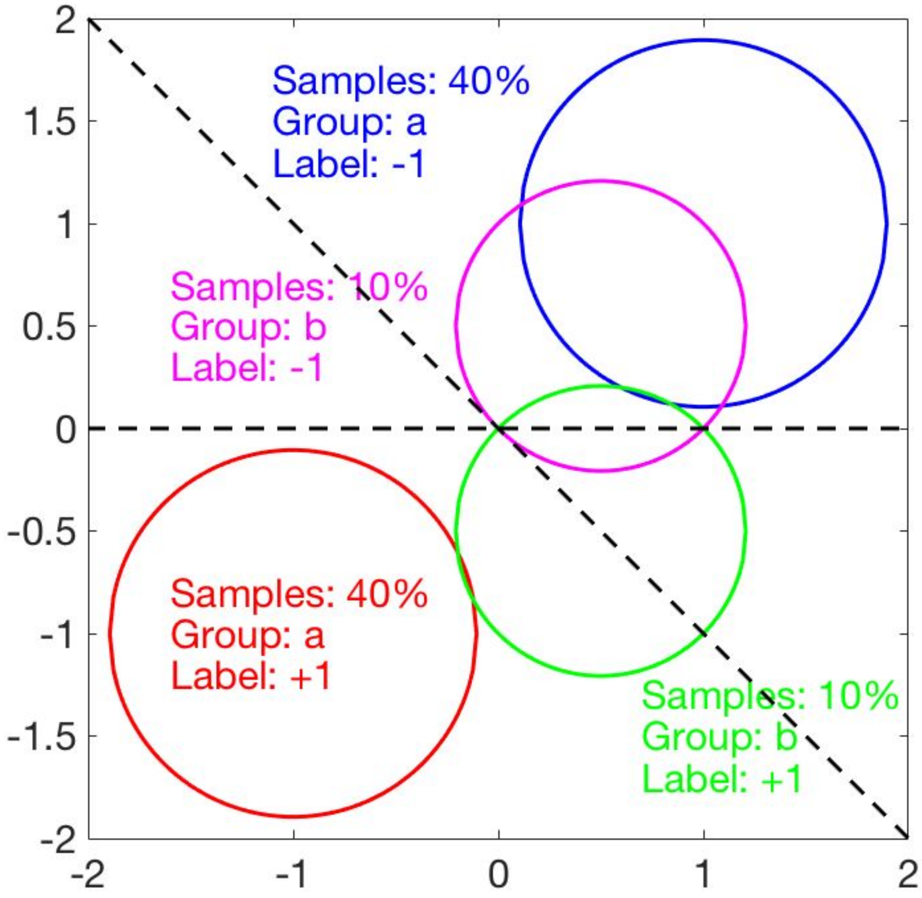
Fairness

- What?
  - Ensure that the learned model does not treat subgroups in the population ‘unfairly’
- Why?
  - Avoid cascade effects in perpetrating biases in the data
- In order to create fair models we need
  - a formal definition of fairness
  - a way to impose fairness during model construction
- We will focus on binary classification problems!

Effect of Fairness Constraint

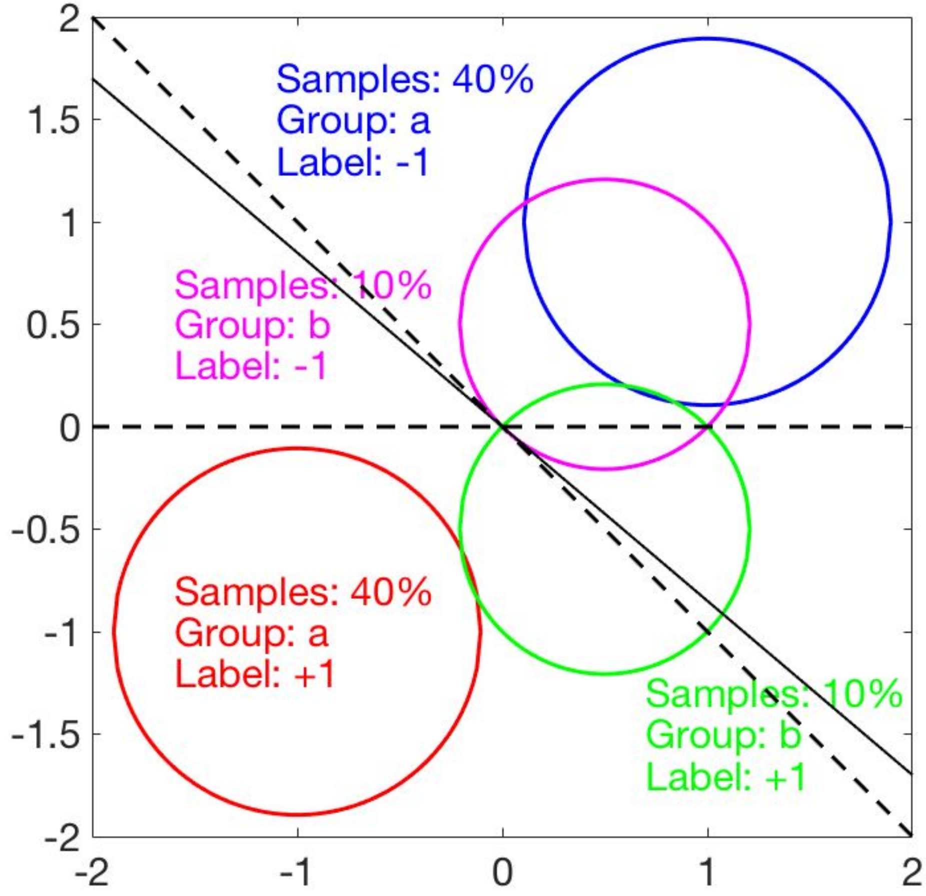


Effect of Fairness Constraint

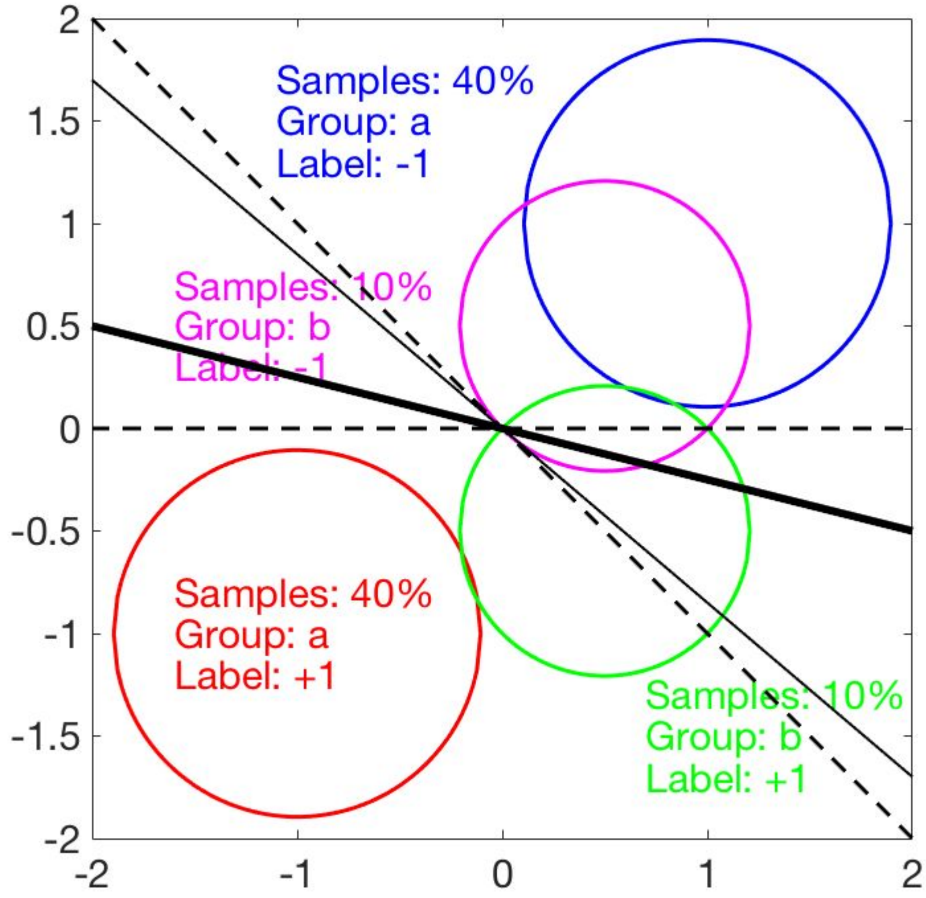


Effect of Fairness Constraint

ISTITUTO ITALIANO
DI TECNOLOGIA
COMPUTATIONAL STATISTICS
AND MACHINE LEARNING



Effect of Fairness Constraint



Notions of Fairness and How to Impose Them

- Many notions are available in literature
  - Equalized Odds and **Equal Opportunity** (True Positive Parity)
  - Demographic Parity, Accuracy Parity
  - Predictive (Positive or Negative) Value Parity
  - Fairness Through Awareness and Fairness through Causality

- How to impose these notions?
  - Pre-Processing (modify the data)
  - In-Processing (modify the algorithm)
  - Post-Processing (modify the learned model)

Equal Opportunity

Equal Opportunity (EO) [Hardt et al. 2017] demands the same True Positive Rate among the groups

$$\mathbb{P}\{f(\boldsymbol{x}) > 0 \mid y = 1, s = a\} = \mathbb{P}\{f(\boldsymbol{x}) > 0 \mid y = 1, s = b\}$$

Requires non-discrimination only within the “advantaged” outcome class (e.g. getting a job). *Equalized odds* extend this to the both positive and negative class

Imposing EO In-Processing

Learning methods aim to find a model which minimizes the risk (error)

$$\min_{f} \; L(f)$$

Our approach: search for a fair model that minimizes the risk

$$\begin{aligned} \min_{f} \quad & L(f) \\ & \mathbb{P}\{f(\boldsymbol{x}){>}0 \mid y{=}1, s{=}a\} = \mathbb{P}\{f(\boldsymbol{x}){>}0 \mid y{=}1, s{=}b\} \end{aligned}$$

Generalization of the EO

Definition of Epsilon-Fairness

$$|L^{+,a}(f) - L^{+,b}(f)| \leq \epsilon, \quad L^{+,g}(f) = \mathbb{E}[\ell(f(\boldsymbol{x}), y) | y=1, s=g]$$

- EO is recovered using the hard loss:

  $\epsilon = 0, \ \ell_h(f(\boldsymbol{x}), y) = \mathbb{1}_{\{yf(\boldsymbol{x}) \leq 0\}} \ \rightarrow \ \mathbb{P}\{f(\boldsymbol{x}) > 0 \mid y=1, s=a\} = \mathbb{P}\{f(\boldsymbol{x}) > 0 \mid y=1, s=b\}$

- If the linear loss is exploited

  $\epsilon = 0, \ \ell_l(f(\boldsymbol{x}), y) = (1 - yf(\boldsymbol{x}))/2 \ \rightarrow \ \mathbb{E}[f(\boldsymbol{x}) \mid y = 1, s = a] = \mathbb{E}[f(\boldsymbol{x}) \mid y = 1, s = b]$

ISTITUTO ITALIANO DI TECNOLOGIA COMPUTATIONAL STATISTICS AND MACHINE LEARNING

Our Problem

- Original Problem

$$\min\left\{L(f) : f \in \mathcal{F},\ \mathbb{P}\{f(\boldsymbol{x}) > 0 \mid y = 1, s = a\} = \mathbb{P}\{f(\boldsymbol{x}) > 0 \mid y = 1, s = b\}\right\}$$

- Our proposal (generalization of the EO)

$$\min\left\{L(f) : f \in \mathcal{F},\ \left|L^{+,a}(f) - L^{+,b}(f)\right| \leq \epsilon\right\}$$

- Its empirical version

$$\min\left\{\hat{L}(f) : f \in \mathcal{F},\ \left|\hat{L}^{+,a}(f) - \hat{L}^{+,b}(f)\right| \leq \hat{\epsilon}\right\}$$

- We assume the space of functions to be learnable

Goal:

- Consistency properties
- Computational efficiency

Consistency Result

Ideal model

$$f^* = \arg\min\left\{L(f) : f\in\mathcal{F},\ \left|L^{+,a}(f) - L^{+,b}(f)\right| \leq \epsilon\right\}$$

FERM (Fair Empirical Risk Minimization) estimator

$$\hat{f} = \min\left\{\hat{L}(f) : f\in\mathcal{F},\ \left|\hat{L}^{+,a}(f) - \hat{L}^{+,b}(f)\right| \leq \hat{\epsilon}\right\} \qquad \hat{\epsilon} = \epsilon + O(1/\sqrt{n})$$

FERM is

- Consistent w.r.t. the risk $\quad L(\hat{f}) - L(f^*) \leq O(1/\sqrt{n})$
- Consistent w.r.t. the fairness $\quad \left|L^{+,a}(\hat{f}) - L^{+,b}(\hat{f})\right| \leq \epsilon + O(1/\sqrt{n})$

Convex FERM Estimator

- Problem (Hard Loss for Error & Hard Loss for Fairness) **Non-Convex**

$$f_h^* = \arg\min \left\{ L_h(f) : f \in \mathcal{F},\ \left| L_h^{+,a}(f) - L_h^{+,b}(f) \right| \leq \epsilon \right\}$$

- FERM Estimator (Hard Loss for Error & Hard Loss for Fairness) **Non-Convex**

$$\hat{f}_h = \arg\min \left\{ \hat{L}_h(f) : f \in \mathcal{F},\ \left| \hat{L}_h^{+,a}(f) - \hat{L}_h^{+,b}(f) \right| \leq \hat{\epsilon} \right\}$$

- FERM Estimator (Hinge Loss for Error & Linear Loss for Fairness) **Convex**

$$f_c = \arg\min \left\{ \hat{L}_c(f) : f \in \mathcal{F},\ \left| \hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f) \right| \leq \hat{\epsilon} \right\}$$

How Good Is Our Approximation?

FERM Estimator (Hard Loss for Error & Hard Loss for Fairness) **Non-Convex**

$$\hat{f}_h = \arg\min\left\{\hat{L}_h(f) : f \in \mathcal{F},\ \left|\hat{L}_h^{+,a}(f) - \hat{L}_h^{+,b}(f)\right| \leq \hat{\epsilon}\right\}$$

FERM Estimator (Hinge Loss for Error & Linear Loss for Fairness) **Convex**

$$f_c = \arg\min\left\{\hat{L}_c(f) : f \in \mathcal{F},\ \left|\hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f)\right| \leq \hat{\epsilon}\right\}$$

Dataset	$\hat{\Delta}$
Arrhythmia	0.03
COMPAS	0.04
Adult	0.06
German	0.05
Drug	0.03

The Hinge Loss ensures that $\hat{L}_h(f) \leq \hat{L}_c(f)$

Moreover it is possible to prove that

$$\frac{1}{2}\sum_{g\in\{a,b\}}\left|\hat{\mathbb{E}}\left[\operatorname{sign}\left(f(\boldsymbol{x})\right) - f(\boldsymbol{x}) \mid y=1, s=g\right]\right| \leq \hat{\Delta} \ \rightarrow\ \left|\hat{L}_h^{+,a}(f) - \hat{L}_h^{+,b}(f)\right| \leq \left|\hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f)\right| + \hat{\Delta}$$

Together these observation justify the method

ISTITUTO ITALIANO DI TECNOLOGIA
COMPUTATIONAL STATISTICS AND MACHINE LEARNING

Our Convex Problem and Kernel Methods

Convex FERM: $\min \left\{ \hat{L}_c(f) : f \in \mathcal{F},\ \left| \hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f) \right| \leq \hat{\epsilon} \right\}$

Kernel Methods: $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$

The constraint becomes

$$\left| \hat{L}_l^{+,a}(f) - \hat{L}_l^{+,b}(f) \right| \leq \hat{\epsilon} \ \rightarrow \ \left| \langle \boldsymbol{w}, \boldsymbol{u} \rangle \right| \leq \epsilon,\ \boldsymbol{u} = \boldsymbol{u}_a - \boldsymbol{u}_b,\ \boldsymbol{u}_g = \frac{1}{n^{+,g}} \sum_{i \in \mathcal{I}^{+,g}} \boldsymbol{\phi}(\boldsymbol{x}_i)$$

The problem (in feature space)

$$\min_{\boldsymbol{w} \in \mathbb{H}} \sum_{i=1}^{n} \ell(\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}_i) \rangle, y_i) + \lambda \|\boldsymbol{w}\|^2 \quad \text{s.t. } \left| \langle \boldsymbol{w}, \boldsymbol{u} \rangle \right| \leq \epsilon$$

The dual formulation (with kernels)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} \ell\left(\sum_{j=1}^{n} K_{ij}\alpha_j, y_i \right) + \lambda \sum_{i,j=1}^{n} \alpha_i \alpha_j K_{ij} \quad \text{s.t. } \left| \sum_{i=1}^{n} \alpha_i \left[\frac{1}{n^{+,a}} \sum_{j \in \mathcal{I}^{+,a}} K_{ij} - \frac{1}{n^{+,b}} \sum_{j \in \mathcal{I}^{+,b}} K_{ij} \right] \right| \leq \epsilon \right\}.$$

Observation

If $\epsilon = 0$ and in the linear case our **In-Processing** method becomes a **Pre-Processing** method

$$\tilde{x}_j = x_j - x_i \frac{u_i}{u_j}, \quad j \in \{1, \ldots, i-1, i+1, \ldots, d\}, \; i : u_i = \|\boldsymbol{u}\|_\infty$$

With a simple preprocessing we can make fair any linear (or kernel) based method

- See paper for experiments with the Lasso

Test & Dataset

Performance measures

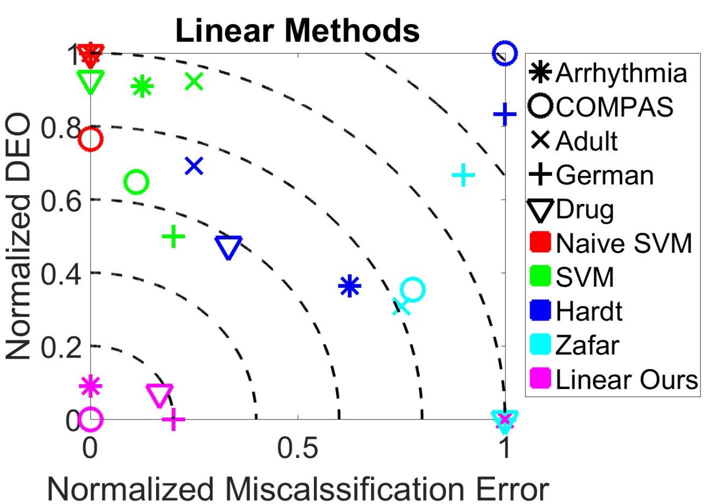
- Accuracy (ACC)
- Difference of EO (DEO)

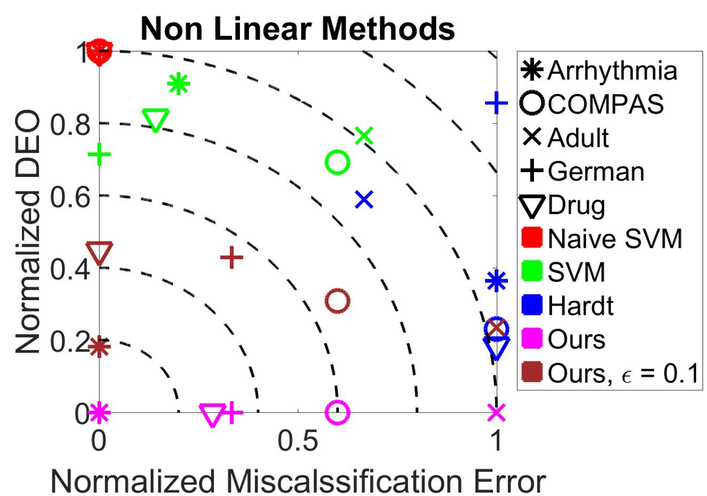
Modified validation procedure: select the fairest model among those with accuracy above 97% that of the most accurate model

Dataset	Examples	Features	Sensitive Variable
Arrhythmia	452	279	Gender
COMPAS	6172	10	Ethnicity
Adult	32561, 12661	12	Gender
German	1700	20	Foreign
Drug	1885	11	Ethnicity

Results

Method	Arrhythmia ACC	Arrhythmia DEO	COMPAS ACC	COMPAS DEO	Adult ACC	Adult DEO	German ACC	German DEO	Drug ACC	Drug DEO
s inside $\boldsymbol{x}$										
Naïve Lin. SVM	0.79±0.06	0.14±0.03	0.76±0.01	0.17±0.02	0.81	0.14	0.71±0.06	0.17±0.05	0.81±0.02	0.44±0.03
Lin. SVM	0.78±0.07	0.13±0.04	0.75±0.01	0.15±0.02	0.80	0.13	0.69±0.04	0.11±0.10	0.81±0.02	0.41±0.06
Hardt	0.74±0.06	0.07±0.04	0.67±0.03	0.21±0.09	0.80	0.10	0.61±0.15	0.15±0.13	0.77±0.02	0.22±0.09
Zafar	0.71±0.03	0.03±0.02	0.69±0.02	0.10±0.06	0.78	0.05	0.62±0.09	0.13±0.11	0.69±0.03	0.02±0.07
Lin. Ours	0.79±0.07	0.04±0.03	0.76±0.01	0.04±0.03	0.77	0.01	0.69±0.04	0.05±0.03	0.79±0.02	0.05±0.03
Naïve SVM	0.79±0.06	0.14±0.04	0.76±0.01	0.18±0.02	0.84	0.18	0.74±0.05	0.12±0.05	0.82±0.02	0.45±0.04
SVM	0.78±0.06	0.13±0.04	0.73±0.01	0.14±0.02	0.82	0.14	0.74±0.03	0.10±0.06	0.81±0.02	0.38±0.03
Hardt	0.74±0.06	0.07±0.04	0.71±0.01	0.08±0.01	0.82	0.11	0.71±0.03	0.11±0.18	0.75±0.11	0.14±0.08
Ours	0.79±0.09	0.03±0.02	0.73±0.01	0.05±0.03	0.81	0.01	0.73±0.04	0.05±0.03	0.80±0.03	0.07±0.05





Using the Sensitive Feature?

Accuracy increases if s is used as a predictor

Method	Arrhythmia		COMPAS		Adult		German		Drug	
	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
					s not inside $\boldsymbol{x}$					
Naïve Lin. SVM	0.75±0.04	0.11±0.03	0.73±0.01	0.13±0.02	0.78	0.10	0.71±0.06	0.16±0.04	0.79±0.02	0.25±0.03
Lin. SVM	0.71±0.05	0.10±0.03	0.72±0.01	0.12±0.02	0.78	0.09	0.69±0.04	0.11±0.10	0.79±0.02	0.25±0.04
Hardt	-	-	-	-	-	-	-	-	-	-
Zafar	0.67±0.03	0.05±0.02	0.69±0.01	0.10±0.08	0.76	0.05	0.62±0.09	0.13±0.10	0.66±0.03	0.06±0.06
Lin. Ours	0.75±0.05	0.05±0.02	0.73±0.01	0.07±0.02	0.75	0.01	0.69±0.04	0.06±0.03	0.79±0.02	0.10±0.06
Naïve SVM	0.75±0.04	0.11±0.03	0.72±0.01	0.14±0.02	0.80	0.09	0.74±0.05	0.12±0.05	0.81±0.02	0.22±0.04
SVM	0.71±0.05	0.10±0.03	0.73±0.01	0.11±0.02	0.79	0.08	0.74±0.03	0.10±0.06	0.81±0.02	0.22±0.03
Hardt	-	-	-	-	-	-	-	-	-	-
Ours	0.75±0.05	0.05±0.02	0.72±0.01	0.08±0.02	0.77	0.01	0.73±0.04	0.05±0.03	0.79±0.03	0.10±0.05
					s inside $\boldsymbol{x}$					
Naïve Lin. SVM	0.79±0.06	0.14±0.03	0.76±0.01	0.17±0.02	0.81	0.14	0.71±0.06	0.17±0.05	0.81±0.02	0.44±0.03
Lin. SVM	0.78±0.07	0.13±0.04	0.75±0.01	0.15±0.02	0.80	0.13	0.69±0.04	0.11±0.10	0.81±0.02	0.41±0.06
Hardt	0.74±0.06	0.07±0.04	0.67±0.03	0.21±0.09	0.80	0.10	0.61±0.15	0.15±0.13	0.77±0.02	0.22±0.09
Zafar	0.71±0.03	0.03±0.02	0.69±0.02	0.10±0.06	0.78	0.05	0.62±0.09	0.13±0.11	0.69±0.03	0.02±0.07
Lin. Ours	0.79±0.07	0.04±0.03	0.76±0.01	0.04±0.03	0.77	0.01	0.69±0.04	0.05±0.03	0.79±0.02	0.05±0.03
Naïve SVM	0.79±0.06	0.14±0.04	0.76±0.01	0.18±0.02	0.84	0.18	0.74±0.05	0.12±0.05	0.82±0.02	0.45±0.04
SVM	0.78±0.06	0.13±0.04	0.73±0.01	0.14±0.02	0.82	0.14	0.74±0.03	0.10±0.06	0.81±0.02	0.38±0.03
Hardt	0.74±0.06	0.07±0.04	0.71±0.01	0.08±0.01	0.82	0.11	0.71±0.03	0.11±0.18	0.75±0.11	0.14±0.08
Ours	0.79±0.09	0.03±0.02	0.73±0.01	0.05±0.03	0.81	0.01	0.73±0.04	0.05±0.03	0.80±0.03	0.07±0.05

Using the Sensitive Feature?

Fairness measure tends to improve if s is not in the functional form of the model

Method	Arrhythmia ACC	Arrhythmia DEO	COMPAS ACC	COMPAS DEO	Adult ACC	Adult DEO	German ACC	German DEO	Drug ACC	Drug DEO
					s not inside $\boldsymbol{x}$					
Naïve Lin. SVM	0.75±0.04	0.11±0.03	0.73±0.01	0.13±0.02	0.78	0.10	0.71±0.06	0.16±0.04	0.79±0.02	0.25±0.03
Lin. SVM	0.71±0.05	0.10±0.03	0.72±0.01	0.12±0.02	0.78	0.09	0.69±0.04	0.11±0.10	0.79±0.02	0.25±0.04
Hardt	-	-	-	-	-	-	-	-	-	-
Zafar	0.67±0.03	0.05±0.02	0.69±0.01	0.10±0.08	0.76	0.05	0.62±0.09	0.13±0.10	0.66±0.03	0.06±0.06
Lin. Ours	0.75±0.05	0.05±0.02	0.73±0.01	0.07±0.02	0.75	0.01	0.69±0.04	0.06±0.03	0.79±0.02	0.10±0.06
Naïve SVM	0.75±0.04	0.11±0.03	0.72±0.01	0.14±0.02	0.80	0.09	0.74±0.05	0.12±0.05	0.81±0.02	0.22±0.04
SVM	0.71±0.05	0.10±0.03	0.73±0.01	0.11±0.02	0.79	0.08	0.74±0.03	0.10±0.06	0.81±0.02	0.22±0.03
Hardt	-	-	-	-	-	-	-	-	-	-
Ours	0.75±0.05	0.05±0.02	0.72±0.01	0.08±0.02	0.77	0.01	0.73±0.04	0.05±0.03	0.79±0.03	0.10±0.05
					s inside $\boldsymbol{x}$					
Naïve Lin. SVM	0.79±0.06	0.14±0.03	0.76±0.01	0.17±0.02	0.81	0.14	0.71±0.06	0.17±0.05	0.81±0.02	0.44±0.03
Lin. SVM	0.78±0.07	0.13±0.04	0.75±0.01	0.15±0.02	0.80	0.13	0.69±0.04	0.11±0.10	0.81±0.02	0.41±0.06
Hardt	0.74±0.06	0.07±0.04	0.67±0.03	0.21±0.09	0.80	0.10	0.61±0.15	0.15±0.13	0.77±0.02	0.22±0.09
Zafar	0.71±0.03	0.03±0.02	0.69±0.02	0.10±0.06	0.78	0.05	0.62±0.09	0.13±0.11	0.69±0.03	0.02±0.07
Lin. Ours	0.79±0.07	0.04±0.03	0.76±0.01	0.04±0.03	0.77	0.01	0.69±0.04	0.05±0.03	0.79±0.02	0.05±0.03
Naïve SVM	0.79±0.06	0.14±0.04	0.76±0.01	0.18±0.02	0.84	0.18	0.74±0.05	0.12±0.05	0.82±0.02	0.45±0.04
SVM	0.78±0.06	0.13±0.04	0.73±0.01	0.14±0.02	0.82	0.14	0.74±0.03	0.10±0.06	0.81±0.02	0.38±0.03
Hardt	0.74±0.06	0.07±0.04	0.71±0.01	0.08±0.01	0.82	0.11	0.71±0.03	0.11±0.18	0.75±0.11	0.14±0.08
Ours	0.79±0.09	0.03±0.02	0.73±0.01	0.05±0.03	0.81	0.01	0.73±0.04	0.05±0.03	0.80±0.03	0.07±0.05

Lessons Learned

Tension between accuracy and fairness

- Accuracy increased using the sensitive feature

- Removing the sensitive feature
  - Usually increases fairness
    (see previous results)
  - It may not ensure fairness
    (other feature correlated with the sensitive one)

Multi Task Learning (MTL)

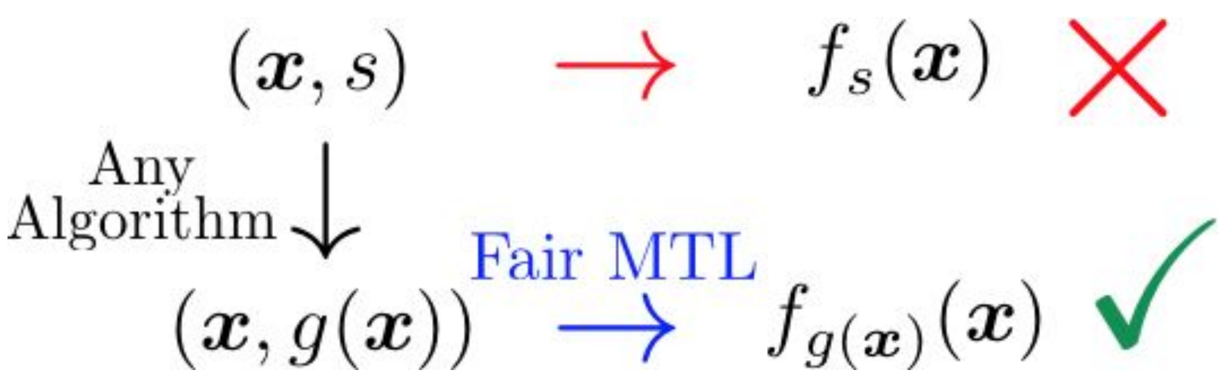
- Framework for solving a collection of related learning problems jointly

- When problems (tasks) are closely related, jointly learning can be more efficient than learning independently
  - Single Task Learning: learn a single model for all the groups
  - Independent Task Learning: learn a model for each group
  - Multi Task Learning: jointly learn both a shared and group specific models

Approach

- Optimize model accuracy and fairness without explicitly using the sensitive feature in the functional form of the model

- Our method is based on two key ideas
  - Use MTL enhanced with fairness constraints to jointly learn group specific classifiers that leverage information between sensitive groups
  - Since learning group specific models might not be permitted, we propose to first predict the sensitive features by any learning method and then to use the predicted sensitive feature



MTL plus Fairness

We build on "regularization around a common mean" for jointly learn a shared and group specific models

$$\min_{\boldsymbol{w}_0,\boldsymbol{w}_1,\dots,\boldsymbol{w}_S\in\mathbb{H}} \ \theta\hat{L}(\boldsymbol{w}_0)+(1-\theta)\tfrac{1}{k}\textstyle\sum_{s=1}^{k}\hat{L}_s(\boldsymbol{w}_s)+\rho\left[\lambda\|\boldsymbol{w}_0\|^2+(1-\lambda)\tfrac{1}{k}\textstyle\sum_{s=1}^{k}\|\boldsymbol{w}_s\|^2\right]$$

Then we generalized our FERM fairness constraint to the MTL framework

Constrain for shared model

$$\boldsymbol{w}_0\cdot(\boldsymbol{u}_1^\diamond-\boldsymbol{u}_2^\diamond)=0\ \wedge\ \dots\ \wedge\ \boldsymbol{w}_0\cdot(\boldsymbol{u}_1^\diamond-\boldsymbol{u}_k^\diamond)=0$$

Constrain for group specific model

$$\boldsymbol{w}_1\cdot\boldsymbol{u}_1^\diamond=\boldsymbol{w}_2\cdot\boldsymbol{u}_2^\diamond\ \wedge\ \dots\ \wedge\ \boldsymbol{w}_1\cdot\boldsymbol{u}_1^\diamond=\boldsymbol{w}_k\cdot\boldsymbol{u}_k^\diamond$$

Datasets

ADULT

Sens.	Group	%
G	Male (M)	66.9
	Female(F)	33.2
R	White (W)	85.5
	Black (B)	9.6
	Asian-Pac-Islander (API)	3.1
	Amer-Indian-Eskimo (AIE)	1.0
	Other (O)	0.8
G+R	W&M	58.8
	W&F	26.7
	B&M	4.9
	B&F	4.7
	API&M	2.1
	API&F	1.1
	AIE&M	0.6
	AIE&F	0.4
	O&M	0.5
	O&F	0.3

COMPAS

Sens.	Group	%
G	Female (F)	19.34
	Male (M)	80.66
R	African-American (AA)	51.23
	Asian (A)	0.44
	Caucasian (C)	34.02
	Hispanic (H)	8.83
	Native American (NA)	0.25
	Other (O)	5.23
G+R	Female African-American	9.04
	Female Asian	0.03
	Female Caucasian	7.86
	Female Hispanic	1.48
	Female Native American	0.06
	Female Other	0.93
	Male African-American	42.20
	Male Asian	0.45
	Male Caucasian	26.16
	Male Hispanic	7.40
	Male Native American	0.19
	Male Other	4.30

ISTITUTO ITALIANO DI TECNOLOGIA
COMPUTATIONAL STATISTICS AND MACHINE LEARNING

Predicting the Sensitive Feature

ADULT

G	M	F
M	58.2	3.8
F	8.7	29.4

R	W	B	API	AIE	O
W	78.5	1.7	0.5	0.2	0.1
B	4.6	7.8	0.1	0.0	0.0
API	0.5	0.0	0.8	0.0	0.0
AIE	1.5	0.1	0.0	2.6	0.0
O	0.4	0.0	0.0	0.0	0.7

COMPAS

G	M	F
M	16.7	8.6
F	2.6	72.1

R	AA	A	C	H	NA	O
AA	44.8	0.0	3.4	0.6	0.0	0.3
A	0.1	0.3	0.0	0.0	0.0	0.0
C	4.4	0.0	29.6	0.4	0.0	0.2
H	1.2	0.0	0.6	7.7	0.0	0.1
NA	0.0	0.0	0.0	0.0	0.2	0.0
O	0.7	0.0	0.4	0.1	0.0	4.6

Results (short version)

Comparison between

- (S = 0) the shared model trained with MTL, with fairness constraint, and no sensitive feature in the predictors
- (S = 1) the group specific models trained with MTL, with fairness constraint, the sensitive feature exploited as predictor

- **BUMP IN ACCURACY (S = 1)**

	S	MTL ACC	MTL $DEOp^+$	MTL ACC	MTL $DEOp^-$	MTL ACC	MTL DEOd
Adult Dataset							
G	0	81.8	0.06	82.7	0.05	82.0	0.06
	1	88.1	0.03	89.1	0.03	88.3	0.03
R	0	82.6	0.01	83.5	0.01	82.8	0.01
	1	90.4	0.03	91.3	0.03	90.6	0.03
G+R	0	83.2	0.04	83.9	0.04	83.5	0.04
	1	90.0	0.05	90.8	0.05	90.3	0.05
COMPAS Dataset							
G	0	76.5	0.03	76.4	0.03	75.7	0.03
	1	82.9	0.07	82.8	0.06	82.1	0.06
R	0	82.4	0.03	83.3	0.03	82.6	0.03
	1	90.0	0.03	91.0	0.03	90.2	0.03
G+R	0	83.1	0.05	83.8	0.05	83.4	0.05
	1	89.9	0.05	90.7	0.05	90.3	0.05

Results (short version)

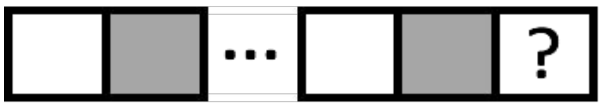
Comparison between

- The group specific models trained with MTL, with fairness constraint, and the true sensitive feature exploited as a predictor (P = 0)
- Against the same model when the predicted sensitive feature exploited as predictor (P = 1)

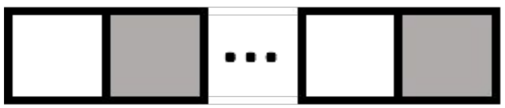
- **BUMP IN FAIRNESS (P = 1)**
- **MILD DECREASE IN ACCURACY (P=1)**

	P	MTL		MTL		MTL	
		ACC	$DEOp^+$	ACC	$DEOp^-$	ACC	DEOd
		Adult Dataset					
G	0	88.1	0.03	89.1	0.03	88.3	0.03
	1	87.4	0.01	88.3	0.01	87.6	0.01
R	0	90.4	0.03	91.3	0.03	90.6	0.03
	1	89.2	0.01	90.2	0.01	89.4	0.01
G+R	0	90.0	0.05	90.8	0.05	90.3	0.05
	1	89.0	0.01	89.8	0.01	89.3	0.01
		COMPAS Dataset					
G	0	82.9	0.07	82.8	0.06	82.1	0.06
	1	82.1	0.01	82.0	0.01	81.3	0.01
R	0	90.0	0.03	91.0	0.03	90.2	0.03
	1	89.0	0.01	89.9	0.01	89.2	0.01
G+R	0	89.9	0.05	90.7	0.05	90.3	0.05
	1	89.0	0.01	89.8	0.01	89.3	0.01

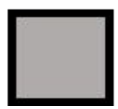
Privacy: Aggregation is enough?



$\sum = a$

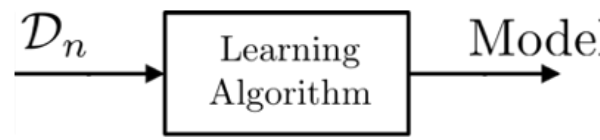


$\sum = a - 1$

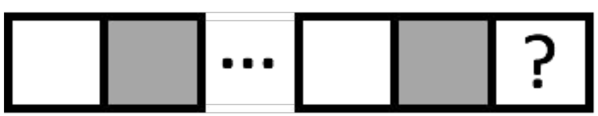


PRIVACY VIOLATED

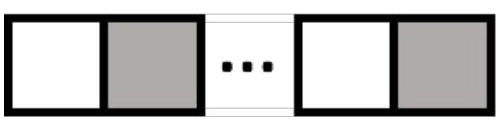
NO!



Deterministic Algorithms



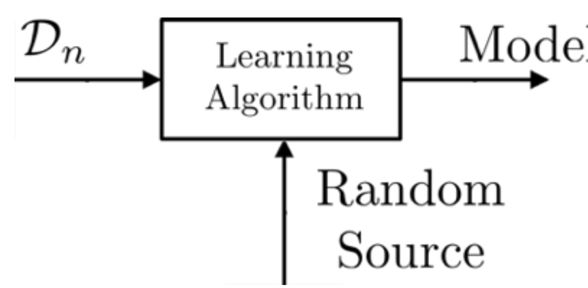
$\{\pm 1\} Random + \sum = a$



$\{\pm 1\} Random + \sum = b$

PRIVACY NOT VIOLATED INFORMATION PRESERVED

Need NOISE!



Randomized Algorithms

Differentially Private Algorithm

Hypotheses:

- randomized algorithms
- samples are i.i.d.

Idea:

If, with the result of the learning procedure, we are not able to retrieve what data we used for learning then the model will generalize

Noise as a tool:

- must be small enough not to hide completely the true answer
- must be large enough to maintain the privacy in the data

$$\frac{\mathbb{P}\{\mathcal{A}(\mathcal{D}_n) = f\}}{\mathbb{P}\{\mathcal{A}(\mathcal{D}_n^i) = f\}} \leq e^{\epsilon}$$

Two Pigeons with one Stone!
DP Algorithms also Generalize

$$\boldsymbol{F} = \mathcal{A}(\mathcal{D}_n),\ \epsilon\text{-private} \qquad \epsilon \leq \sqrt{t^2 - \frac{\ln(2)}{2n}}$$

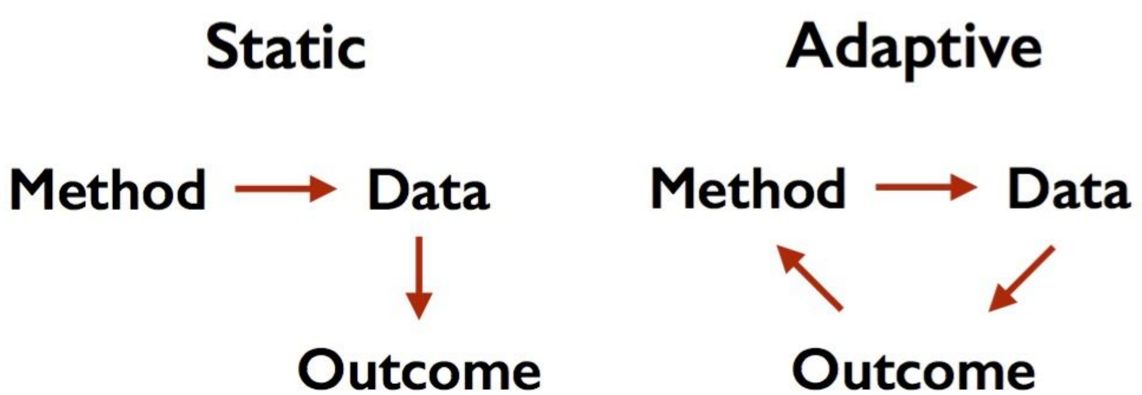
$$\mathbb{P}\{|L(\boldsymbol{F}) - \widehat{L}_n(\boldsymbol{F})| > t\} \leq 3\sqrt{2}e^{-nt^2}$$

1. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015b. Preserving statistical validity in adaptive data analysis, in: Annual ACM Symposium on Theory of Computing.
2. Oneto, L., Ridella, S., & Anguita, D. (2017). Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters, 89, 31-38.

What if the Learning Algorithm is not DP?

- DP theory allows to state the conditions under which a hold-out set can be reused without risk of false discovery through a DP procedure called Thresholdout

- This results is very important in **Adaptive Data Analysis**
    - Hyperparameter Optimization
    - Competitions
    - etc.



Classical Holdout in Adaptive Data Analysis

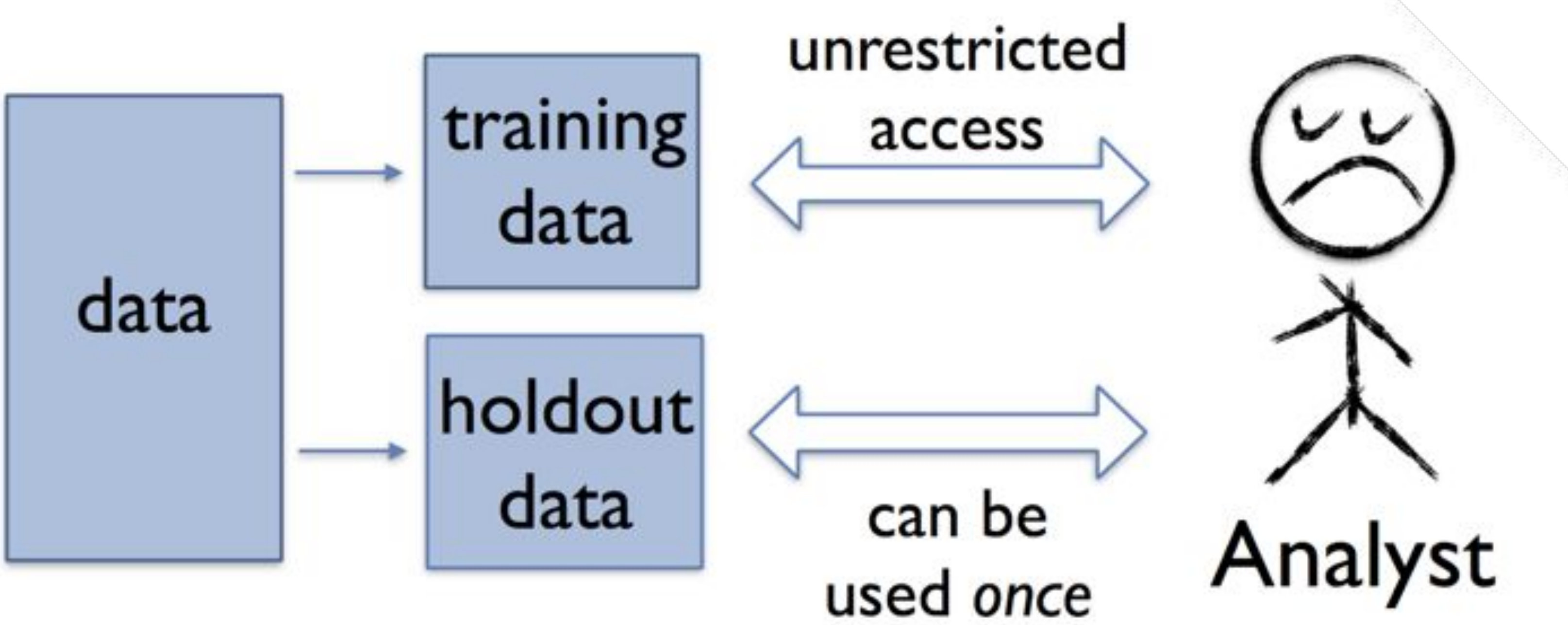
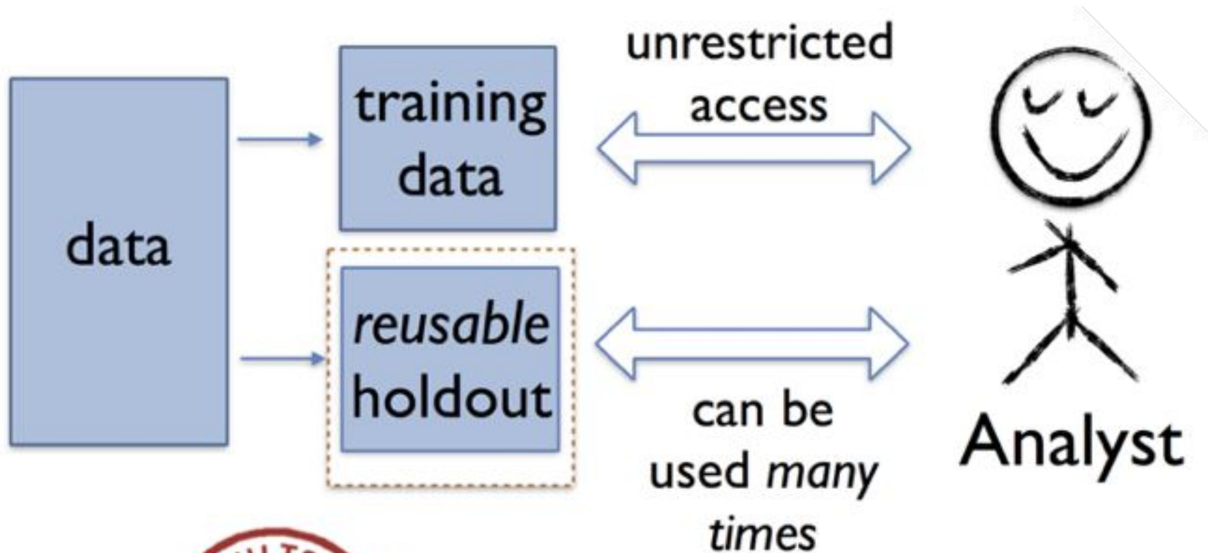


Image Credits:
https://ai.googleblog.com/2015/08/the-reusable-holdout-preserving.html

Thresholdout (Reusable Holdout)

```
from numpy import *
def Thresholdout(sample,holdout,q,sigma,threshold)
    sample_mean = mean([q(x) for x in sample])
    holdout_mean = mean([q(x) for x in holdout])
    if (abs(sample_mean-holdout_mean)<threshold+random(sigma))
        return sample_mean
    else
        return holdout_mean+random(sigma)
```



essentially as good as using *fresh* data each time!

Generalization Bounds in Adaptive Data Analysis

Classical Holdout in Adaptive Data Analysis

$$\mathbb{P}\left\{\exists i \in \{1, \cdots, n_f\} \middle| |L(f_i) - \widehat{L}_n^{s_h^i}(f_i)| \geq \sqrt{\frac{m \ln\left(\frac{2}{\delta}\right)}{2n}}\right\} \leq \delta$$

Thresholdout (Reusable Holdout)

$$\mathbb{P}\left\{\exists i \in \{1, \cdots, n_f\} \middle| |a_i - L(f_i)| \geq 40\sqrt{\frac{B \ln\left(\frac{12m}{\beta}\right)}{n}}\right\} \leq \beta$$

Advantage when

$$m \gg B \ln(m)$$

1. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015c. The reusable holdout: Preserving validity in adaptive data analysis. Science 349, 636–638.
2. Oneto, L., Ridella, S., & Anguita, D. (2017). Differential privacy and generalization: Sharper bounds with applications. Pattern Recognition Letters, 89, 31-38.

Future work

Broad goal is to extend DP theory to MTL setting:

- Partial (wrt. features or tasks) Privacy Constraints
  - Links between privacy and fairess
- Hyperparameters Optimization (Thresholdout algorithm)