Building trustworthy ML

The role of label quality and availability

Alexandru Țifrea
Research Scientist,
Google DeepMind

Amartya Sanyal
Tenure Track Assistant Professor,
Department of Computer Science,
University of Copenhagen

Introduction

Aspect 1: Trustworthiness of ML Algorithms
An overloaded term

Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

Fairness

Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

Fairness

The New York Times

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist View of Women

Algorithms showed a tendency to associate women with shopping and men with shooting.

Whether Machine Learning Algorithms have disproportionately worse impact on some groups of people than others

Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

Fairness

Privacy

The New York Times

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist View of Women

Algorithms showed a tendency to associate women with shopping and men with shooting.

Aspect 1: Trustworthiness of ML Algorithms

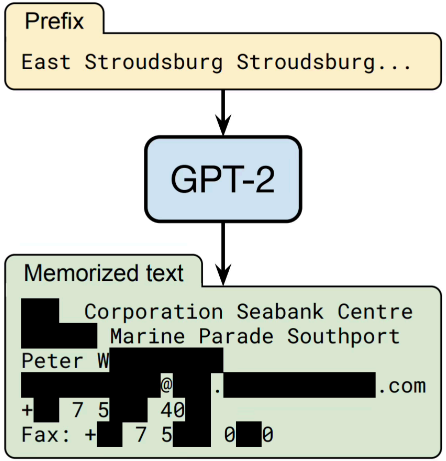
An overloaded term

Fairness

Privacy


The New York Times
A.I. Could Worsen Health Dispa
In a health system riddled with inequity, we risk m
dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM
Machines Taught by Photos Learn a Sexist V
Algorithms showed a tendency to associate women with shopping and men wi



Whether Machine Learning algorithms
leak *personal* (training) data

Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

- **Fairness**
- **Privacy**
- **Robustness**


The New York Times

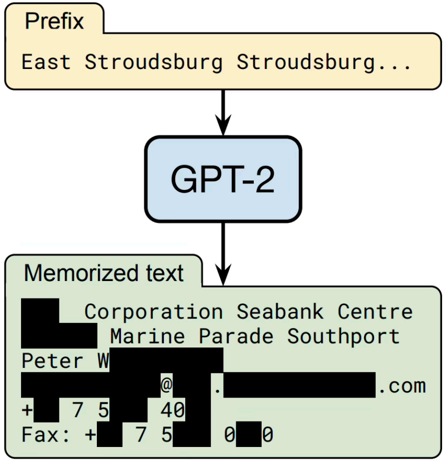
A.I. Could Worsen Health Disp

In a health system riddled with inequity, we risk m
dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist

Algorithms showed a tendency to associate women with shopping and men wi



Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

Fairness

Privacy

Robustness

Whether Machine Learning model can generalise to different data distributions


The New York Times

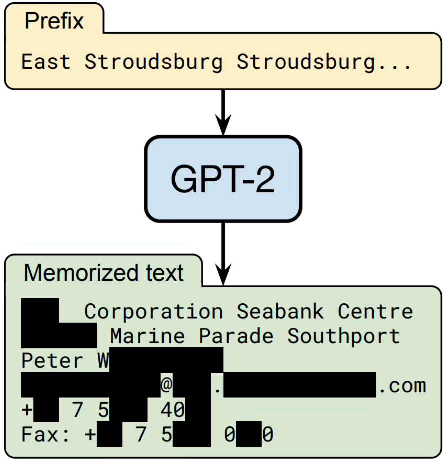
A.I. Could Worsen Health Disp

In a health system riddled with inequity, we risk m
dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist W

Algorithms showed a tendency to associate women with shopping and men wi



Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

Fairness

Privacy

Robustness

Whether Machine Learning model can generalise to different data distributions

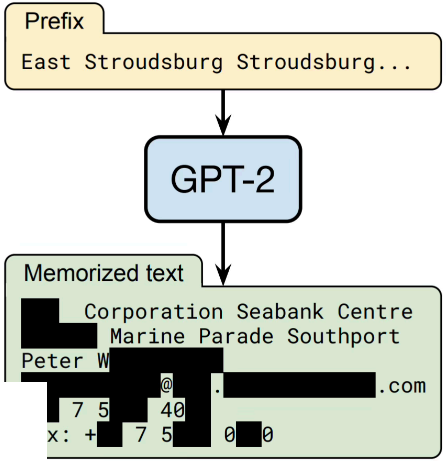

The New York Times

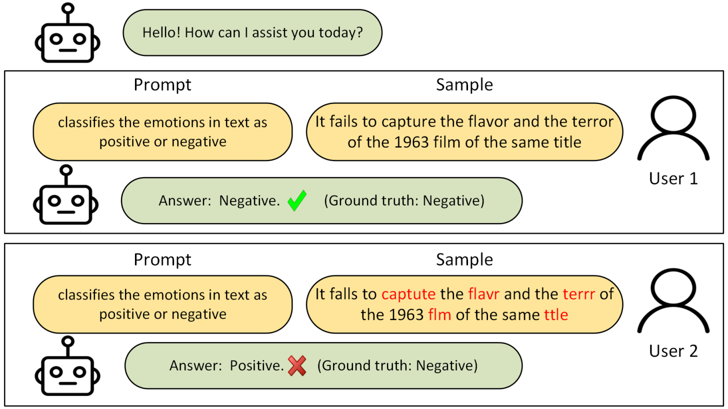
A.I. Could Worsen Health Disp

In a health system riddled with inequity, we risk m
dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist





Adversarial

Aspect 1: Trustworthiness of ML Algorithms

An overloaded term

Fairness

Privacy

Robustness

Whether Machine Learning model can generalise to different data distributions

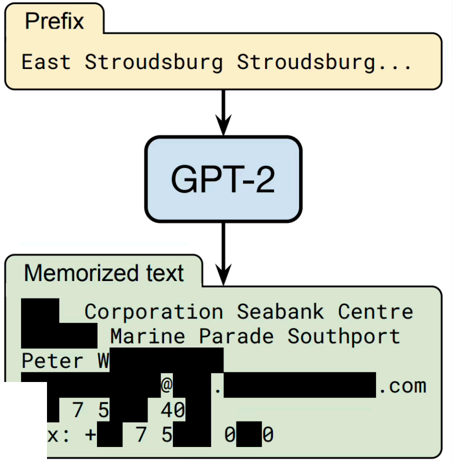

The New York Times

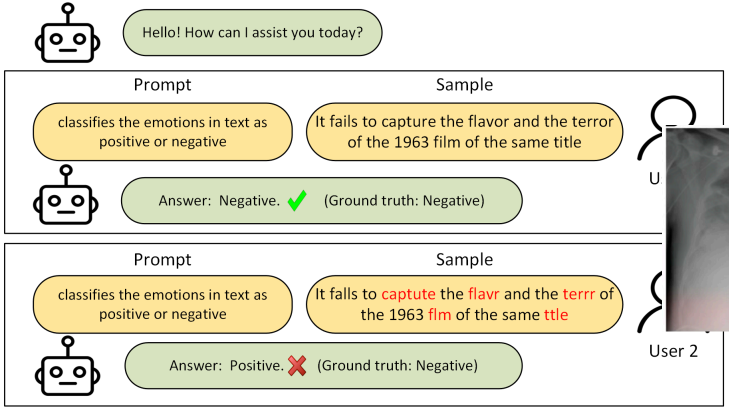
A.I. Could Worsen Health Dispa

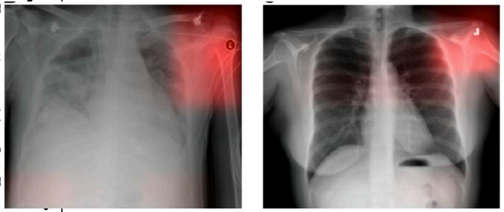
In a health system riddled with inequity, we risk m
dangerous biases automated and invisible.

TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist







Distributional

Adversarial

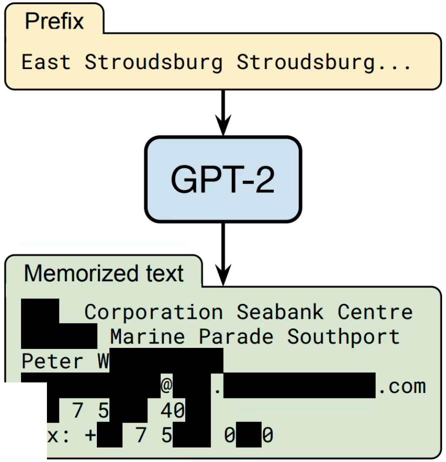
Aspect 1: Trustworthiness of ML Algorithms

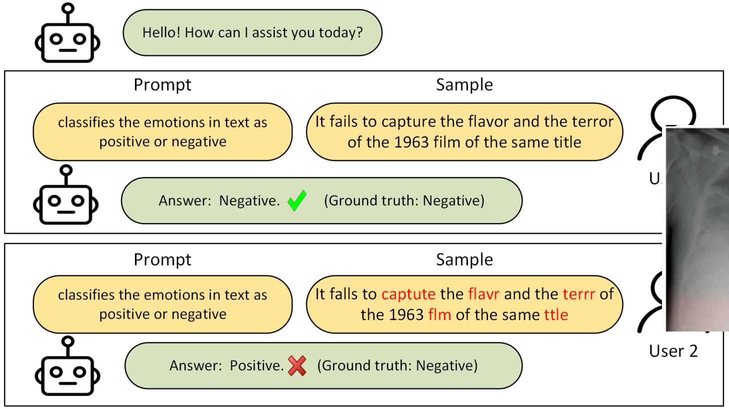
An overloaded term

Fairness

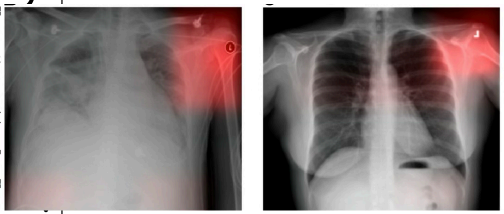
Privacy

Robustness





Adversarial



Distributional

Aspect 2: Label/data quality and availability

Aspect 2: Label/data quality and availability

Two problems with data in ML

Aspect 2: Label/data quality and availability

Two problems with data in ML

- Unlabelled data is significantly more abundant than Labelled data.

Aspect 2: Label/data quality and availability

Two problems with data in ML

- Unlabelled data is significantly more abundant than Labelled data.
- Label noise is ubiquitous in real world data

Aspect 2: Label/data quality and availability

Two problems with data in ML

- Unlabelled data is significantly more abundant than Labelled data.
- Label noise is ubiquitous in real world data

Dataset	Modality	% error
MNIST	image	0.15
CIFAR-10	image	0.54
CIFAR-100	image	5.85
Caltech-256†	image	1.54
ImageNet*	image	5.83
QuickDraw†	image	10.12
20news	text	1.09
IMDB	text	2.90
Amazon Reviews†	text	3.90
AudioSet	audio	1.35

Aspect 2: Label/data quality and availability

Two problems with data in ML

- Unlabelled data is significantly more abundant than Labelled data.
- Label noise is ubiquitous in real world data

In this tutorial, we will look at

Dataset	Modality	% error
MNIST	image	0.15
CIFAR-10	image	0.54
CIFAR-100	image	5.85
Caltech-256†	image	1.54
ImageNet*	image	5.83
QuickDraw†	image	10.12
20news	text	1.09
IMDB	text	2.90
Amazon Reviews†	text	3.90
AudioSet	audio	1.35

Aspect 2: Label/data quality and availability

Two problems with data in ML

- Unlabelled data is significantly more abundant than Labelled data.
- Label noise is ubiquitous in real world data

Dataset	Modality	% error
MNIST	image	0.15
CIFAR-10	image	0.54
CIFAR-100	image	5.85
Caltech-256†	image	1.54
ImageNet*	image	5.83
QuickDraw†	image	10.12
20news	text	1.09
IMDB	text	2.90
Amazon Reviews†	text	3.90
AudioSet	audio	1.35

In this tutorial, we will look at

How **availability and quality of labels** (and data) specifically impact **Fairness, Privacy, and Robustness** of ML Algorithms

Today’s Plan

Today's Plan

- Introduction

Today's Plan

- Introduction
- **Fairness in Machine Learning**

Today's Plan

- Introduction
- **Fairness in Machine Learning**

- Partial group labels
- No group labels
- Low-label regime

Today's Plan

- Introduction
- **Fairness in Machine Learning**
- **Privacy in Machine Learning**

- Partial group labels
- No group labels
- Low-label regime

Today's Plan

- Introduction
- **Fairness in Machine Learning**
- **Privacy in Machine Learning**

- Partial group labels
- No group labels
- Low-label regime

- Privacy and Disparate Impact
- Good data incurs less cost

Today's Plan

- Introduction
- **Fairness in Machine Learning**
  - Partial group labels
  - No group labels
  - Low-label regime
- **Privacy in Machine Learning**
  - Privacy and Disparate Impact
  - Good data incurs less cost
- **Robustness in Machine Learning**

Today's Plan

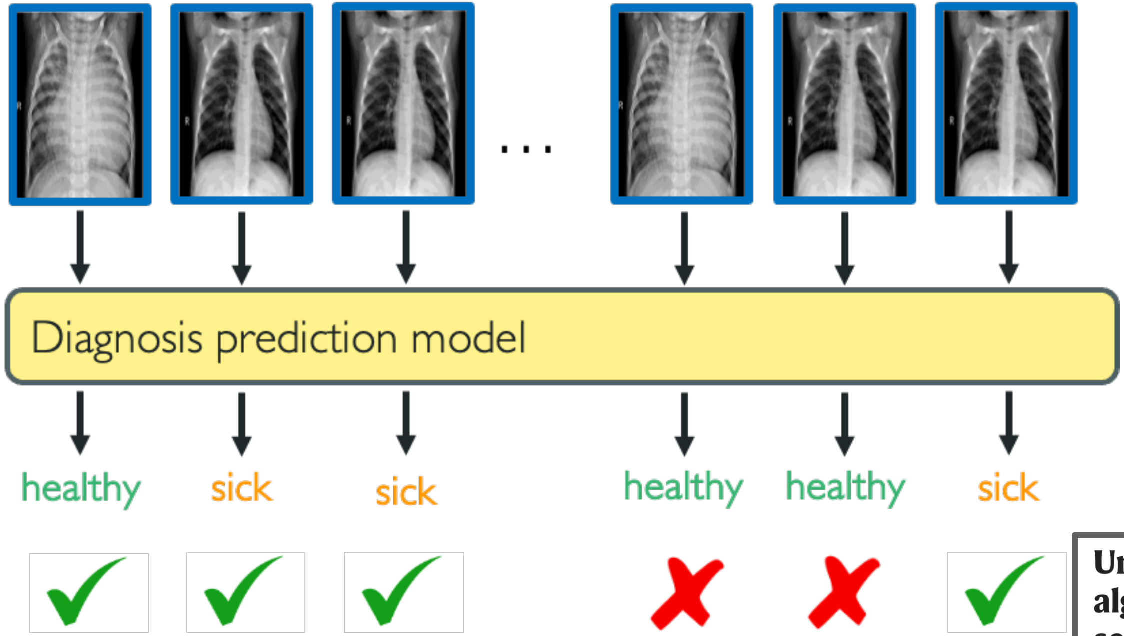
- Introduction
- **Fairness in Machine Learning**
  - Partial group labels
  - No group labels
  - Low-label regime
- **Privacy in Machine Learning**
  - Privacy and Disparate Impact
  - Good data incurs less cost
- **Robustness in Machine Learning**
  - Adversarial Robustness
  - Distributional Generalisation
  - Out-of-distribution detection

Today's Plan

- Introduction
- **Fairness in Machine Learning**
  - Partial group labels
  - No group labels
  - Low-label regime
- **Privacy in Machine Learning**
  - Privacy and Disparate Impact
  - Good data incurs less cost
- **Robustness in Machine Learning**
  - Adversarial Robustness
  - Distributional Generalisation
  - Out-of-distribution detection
- Outlook and Future Direction

Fairness in Machine Learning

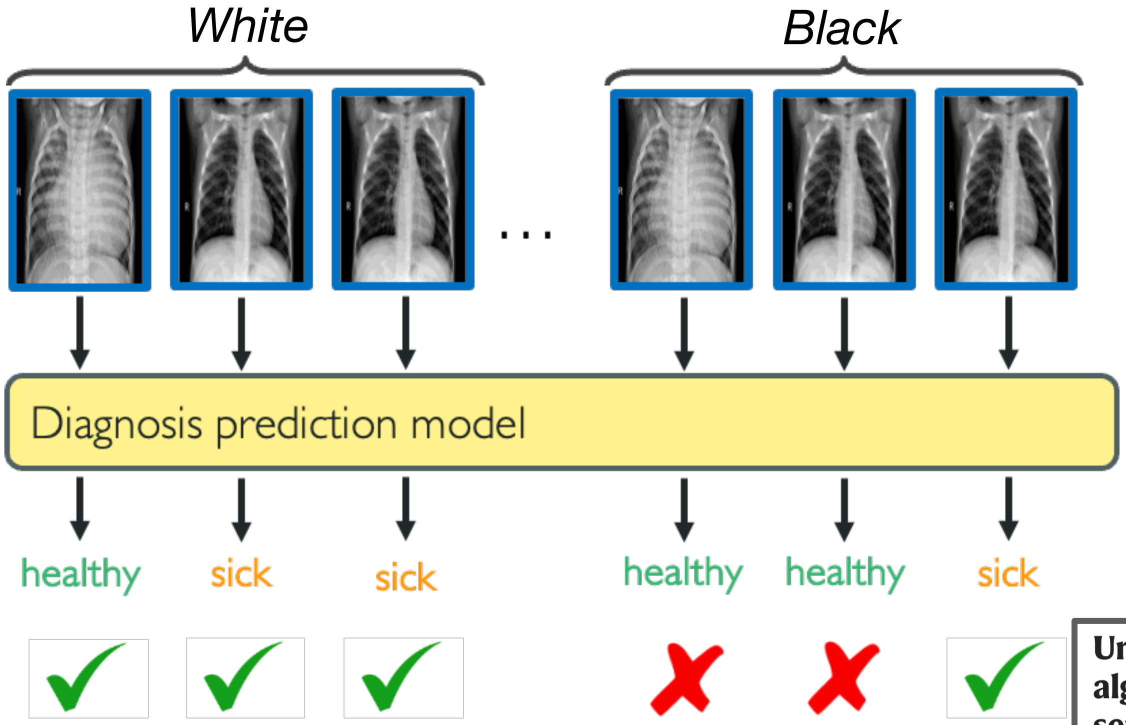
Example of ML model unfairness



Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen & Marzyeh Ghassemi

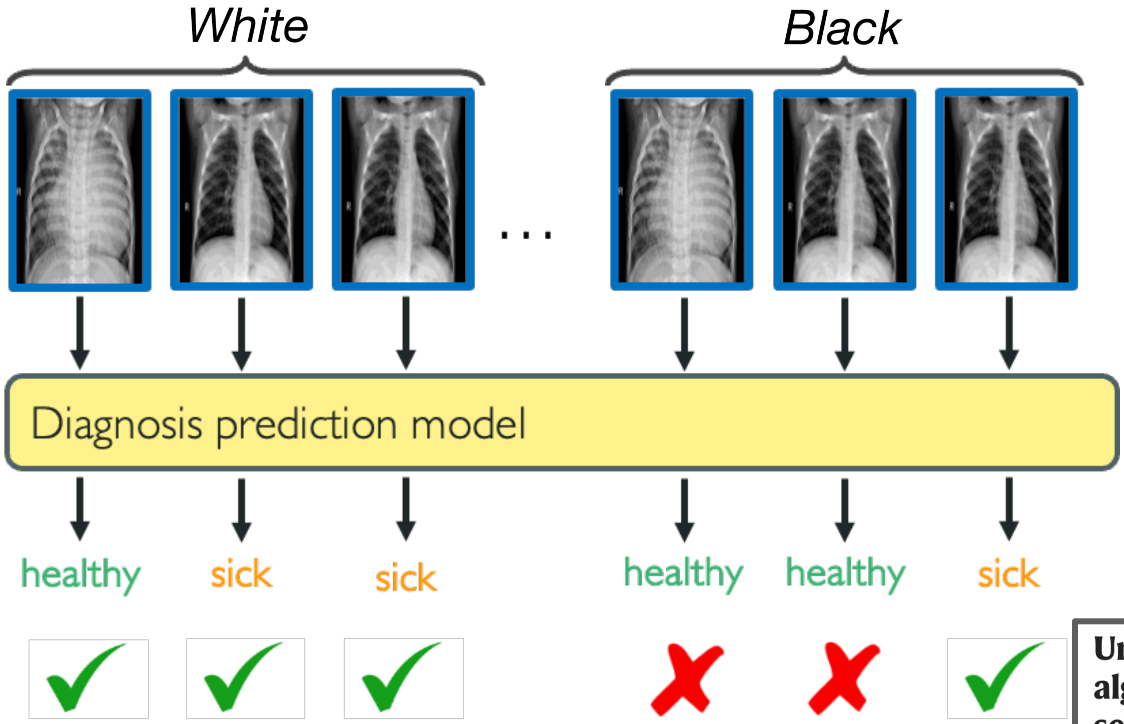
Example of ML model unfairness



Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen & Marzyeh Ghassemi

Example of ML model unfairness



False positive rate:

FPR=P[predicted healthy | actually sick]

FPR[*White*] = 0.16

FPR[*Black*] = 0.27

FPR gap = 0.11

The model is accurate but not fair!

Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen & Marzyeh Ghassemi

Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}\,(X)$ with categorical or continuous labels

Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}\,(X)$ with categorical or continuous labels

Individual fairness: $d_y\left(\hat{f}\,(x_1),\, \hat{f}\,(x_2)\right) < C d_x\,(x_1,\, x_2)$

Fairness Through Awareness

Cynthia Dwork[*] Moritz Hardt[†] Toniann Pitassi[‡] Omer Reingold[§]
Richard Zemel[¶]

'treating similar individuals similarly'

Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}\,(X)$ with categorical or continuous labels

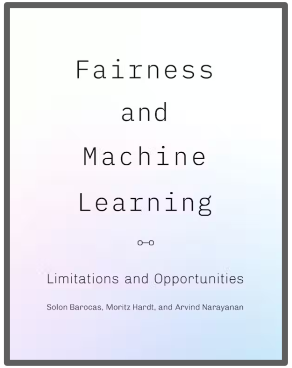
Individual fairness: $d_y\left(\hat{f}\,(x_1),\, \hat{f}\,(x_2)\right) < C d_x\,(x_1,\, x_2)$

Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zemel¶

'treating similar individuals similarly'

Group fairness: Three broad categories of fairness notions



Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}\,(X)$ with categorical or continuous labels

Individual fairness: $d_y\left(\hat{f}\,(x_1),\, \hat{f}\,(x_2)\right) < C d_x\,(x_1,\, x_2)$

Fairness Through Awareness

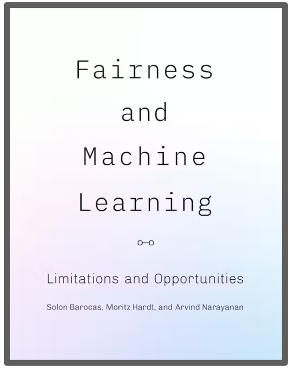
Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zemel¶

'treating similar individuals similarly'

Group fairness: Three broad categories of fairness notions

- Equal acceptance rates
  e.g. statistical parity

$$\mathbb{P}(\hat{Y}|A = \text{White}) = \mathbb{P}(\hat{Y}|A = \text{Black})$$



Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}(X)$ with categorical or continuous labels

Individual fairness: $d_y\left(\hat{f}(x_1), \hat{f}(x_2)\right) < C d_x(x_1, x_2)$

Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zemel¶

'treating similar individuals similarly'

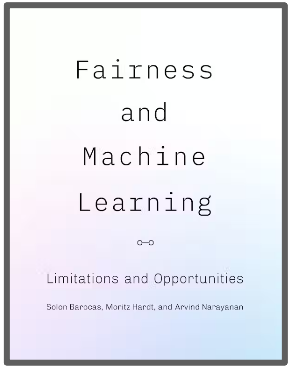
Group fairness: Three broad categories of fairness notions

- Equal acceptance rates
  e.g. statistical parity

$$\mathbb{P}(\hat{Y}|A = \text{White}) = \mathbb{P}(\hat{Y}|A = \text{Black})$$

- Equal error rates
  e.g. Equal Opportunity

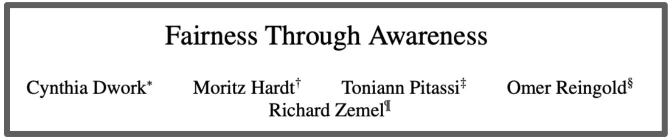
$$\text{FPR}(A = \text{White}) = \text{FPR}(A = \text{Black})$$



Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}(X)$ with categorical or continuous labels

Individual fairness: $d_y\left(\hat{f}(x_1), \hat{f}(x_2)\right) < C d_x(x_1, x_2)$


Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§
Richard Zemel¶

'treating similar individuals similarly'

Group fairness: Three broad categories of fairness notions

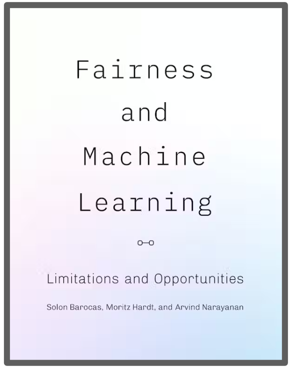
- Equal acceptance rates
  e.g. statistical parity

$$\mathbb{P}(\hat{Y}|A = \text{White}) = \mathbb{P}(\hat{Y}|A = \text{Black})$$

- Equal error rates
  e.g. Equal Opportunity

$$\text{FPR}(A = \text{White}) = \text{FPR}(A = \text{Black})$$

- Equal calibration



Formal definitions of fairness for prediction

Prediction problem: $\hat{Y} = \hat{f}\,(X)$ with categorical or continuous labels

Individual fairness: $d_y\left(\hat{f}\,(x_1),\, \hat{f}\,(x_2)\right) < C d_x\,(x_1,\, x_2)$

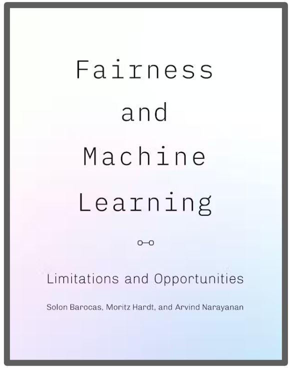
Fairness Through Awareness

Cynthia Dwork[*] Moritz Hardt[†] Toniann Pitassi[‡] Omer Reingold[§] Richard Zemel[¶]

'treating similar individuals similarly'

Group fairness: Three broad categories of fairness notions

- Equal acceptance rates
  e.g. statistical parity
  $$\mathbb{P}(\hat{Y}|A = \mathrm{White}) = \mathbb{P}(\hat{Y}|A = \mathrm{Black})$$
- Equal error rates
  e.g. Equal Opportunity
  $$\mathrm{FPR}(A = \mathrm{White}) = \mathrm{FPR}(A = \mathrm{Black})$$
- Equal calibration



Remark: Different ML problems (e.g. generative ML) employ similar fairness definitions.

Fairness-error trade-off

State-of-the-art prediction models are often unfair

ProPublica

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

The New York Times

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

MIT News

ON CAMPUS AND AROUND THE WORLD

Study reveals why AI models that analyze medical images can be biased

These models, which can predict a patient's race, gender, and age, seem to use those traits as shortcuts when making medical diagnoses.

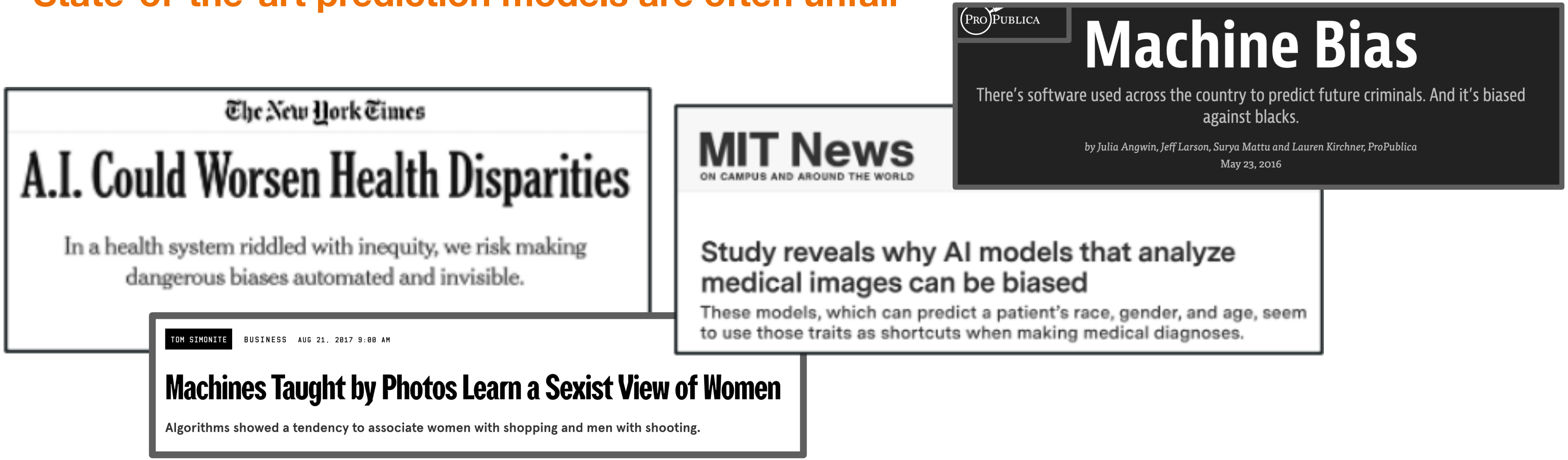
TOM SIMONITE BUSINESS AUG 21, 2017 9:00 AM

Machines Taught by Photos Learn a Sexist View of Women

Algorithms showed a tendency to associate women with shopping and men with shooting.

Fairness-error trade-off

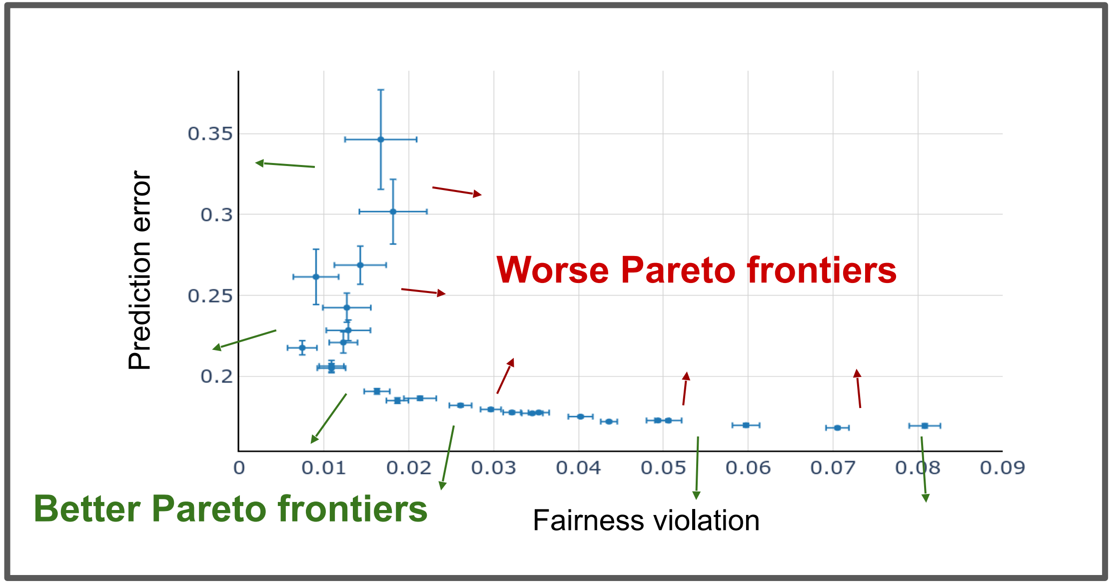
State-of-the-art prediction models are often unfair



Trivial prediction models (e.g. random guessing) can achieve perfect fairness

e.g. for binary classification and two groups $P\left(\hat{Y}=1|A=0\right)=P\left(\hat{Y}=1|A=1\right)=0.5$

Fairness-error Pareto frontier



need special mitigation algorithms

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \boxed{\mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}}$$ (potentially unfair model)

1) **Pre-processing mitigations**

High-level idea: *Change the training data*

Inspired by principle of "Fairness Through Unawareness"

Fairness mitigation strategies

$$\mathrm{OPT_{base}} : \arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \quad \boxed{\mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}}$$ (potentially unfair model)

1) Pre-processing mitigations

High-level idea: *Change the training data*

Inspired by principle of "Fairness Through Unawareness"

Examples:

- feature selection
- fair representation learning
- importance sampling

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

Fairness mitigation strategies

$$\boxed{\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}})} \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$

(potentially unfair model)

2) In-processing mitigations

High-level idea: *Change the training algorithm*

Employ ideas from multi-objective learning

Fairness mitigation strategies

$$\boxed{\mathrm{OPT}_{\mathrm{base}} : \arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}})} \quad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$$ (potentially unfair model)

2) In-processing mitigations

High-level idea: *Change the training algorithm*

Employ ideas from multi-objective learning

e.g. $\arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}) + \lambda \mathcal{L}_{fair}(f; \mathcal{D}_{\mathrm{sensitive}})$ with $\mathcal{D}_{\mathrm{sensitive}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^m$

Fairness mitigation strategies

$$\boxed{\mathrm{OPT}_{\text{base}} : \arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}})} \quad \mathcal{D}_{\text{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$$ (potentially unfair model)

2) In-processing mitigations

High-level idea: *Change the training algorithm*

Employ ideas from multi-objective learning

e.g. $\arg\min_f \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}})} + \lambda \mathcal{L}_{fair}(f; \mathcal{D}_{\text{sensitive}})$ with $\mathcal{D}_{\text{sensitive}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^m$

Fairness mitigation strategies

$$\boxed{\mathrm{OPT}_{\text{base}} : \arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}})} \qquad \mathcal{D}_{\text{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$$ (potentially unfair model)

2) In-processing mitigations

High-level idea: *Change the training algorithm*

Employ ideas from multi-objective learning

e.g. $\arg\min_f \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}})} + \underbrace{\boxed{\lambda \mathcal{L}_{fair}(f; \mathcal{D}_{\text{sensitive}})}}_{\text{unfairness penalty}}$ with $\mathcal{D}_{\text{sensitive}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^m$

Fairness mitigation strategies

$$\boxed{\mathrm{OPT}_{\text{base}} : \arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}})} \qquad \mathcal{D}_{\text{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$$ (potentially unfair model)

2) In-processing mitigations

High-level idea: *Change the training algorithm*

Employ ideas from multi-objective learning

e.g. $\arg\min_f \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}})} + \underbrace{\boxed{\lambda \mathcal{L}_{fair}(f; \mathcal{D}_{\text{sensitive}})}}_{\text{unfairness penalty}}$ with $\mathcal{D}_{\text{sensitive}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^m$

Examples:

- regularized learning
- constrained learning

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_f \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$$ (potentially unfair model)

3) Post-processing mitigations

High-level idea: *Change the outputs of a pre-trained model*

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

3) Post-processing mitigations

High-level idea: *Change the outputs of a pre-trained model*

e.g. group-dependent transformation of outputs:

$$\left(\hat{Y},\, A\right) \to T_A\left(\hat{Y}\right) \in \{0,\, 1\}$$

Fairness mitigation strategies

$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

3) Post-processing mitigations

High-level idea: *Change the outputs of a pre-trained model*

e.g. group-dependent transformation of outputs:

$$\left(\hat{Y},\, A\right) \to T_A\left(\hat{Y}\right) \in \{0,\, 1\}$$

Fairness mitigation strategies

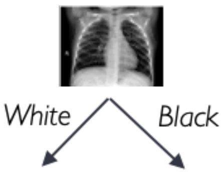
$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

3) Post-processing mitigations

High-level idea: *Change the outputs of a pre-trained model*

e.g. group-dependent transformation of outputs:

$$\left(\hat{Y},\, A\right) \rightarrow T_A\left(\hat{Y}\right) \in \{0,\, 1\}$$



Fairness mitigation strategies

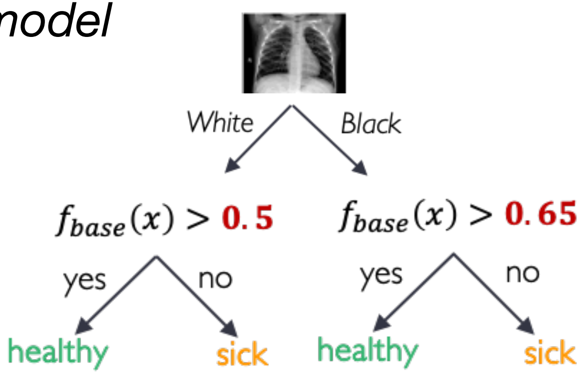
$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

3) Post-processing mitigations

High-level idea: *Change the outputs of a pre-trained model*

e.g. group-dependent transformation of outputs:

$$\left(\hat{Y},\, A\right) \rightarrow T_A\left(\hat{Y}\right) \in \{0,\, 1\}$$



Fairness mitigation strategies

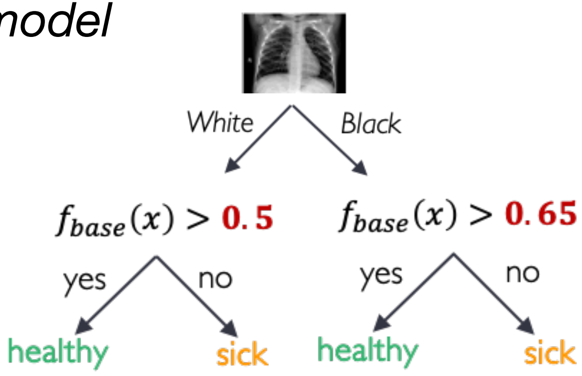
$$\mathrm{OPT}_{\mathrm{base}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\mathrm{pred}}), \qquad \mathcal{D}_{\mathrm{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \sim \mathbb{P}_{XY}$$ (potentially unfair model)

3) Post-processing mitigations

High-level idea: *Change the outputs of a pre-trained model*

e.g. group-dependent transformation of outputs:

$$\left(\hat{Y},\, A\right) \to T_A\left(\hat{Y}\right) \in \{0,\, 1\}$$



Examples:

- group-dependent post-hoc transformations
- group-agnostic transformations

e.g. *fair predictions irrespective of person's willingness to provide sensitive attribute*

Challenges faced by fairness mitigations

Pre-, in-, post-processing mitigations need training data with group labels.

Challenges faced by fairness mitigations

Pre-, in-, post-processing mitigations need training data with group labels.

Issue #1: Group labels are difficult to collect.

e.g. group labels are often sensitive attributes (e.g. gender, ethnicity, age etc)

Challenges faced by fairness mitigations

Pre-, in-, post-processing mitigations need training data with group labels.

Issue #1: Group labels are difficult to collect.

e.g. group labels are often sensitive attributes (e.g. gender, ethnicity, age etc)

What happens when group labels are scarce?

Challenges faced by fairness mitigations

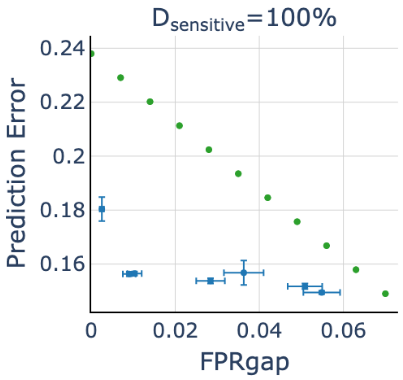
Pre-, in-, post-processing mitigations need training data with group labels.

Issue #1: Group labels are difficult to collect.

e.g. group labels are often sensitive attributes (e.g. gender, ethnicity, age etc)

What happens when group labels are scarce?

Naive baseline: *"predict according to pre-trained model with probability p, and predict 0 with probability (1-p)"*
In-processing mitigation: state-of-the-art MinDiff method



Dataset: Adult
Y = income; A = gender

Figure source: https://arxiv.org/abs/2312.02592

Challenges faced by fairness mitigations

Pre-, in-, post-processing mitigations need training data with group labels.

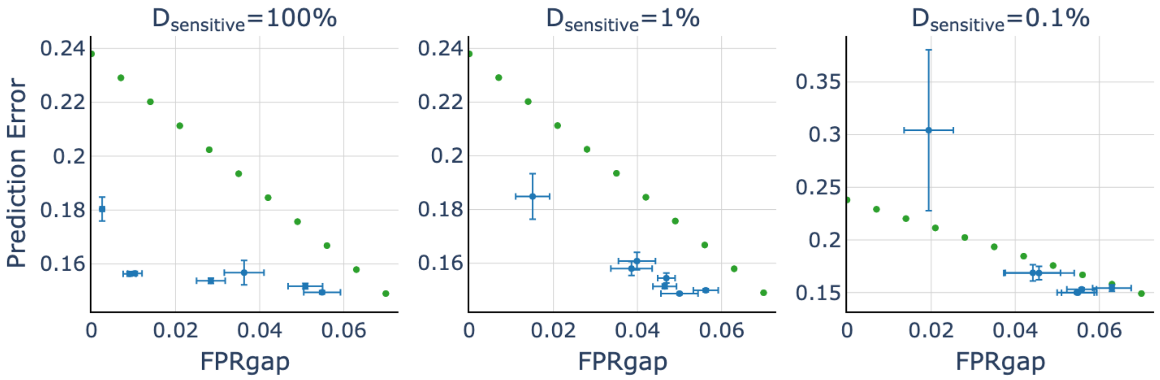
Issue #1: Group labels are difficult to collect.

e.g. group labels are often sensitive attributes (e.g. gender, ethnicity, age etc)

What happens when group labels are scarce?

Naive baseline: *"predict according to pre-trained model with probability p, and predict 0 with probability (1-p)"*

In-processing mitigation: state-of-the-art MinDiff method



Dataset: Adult
Y = income; A = gender

Figure source: https://arxiv.org/abs/2312.02592

Challenges faced by fairness mitigations

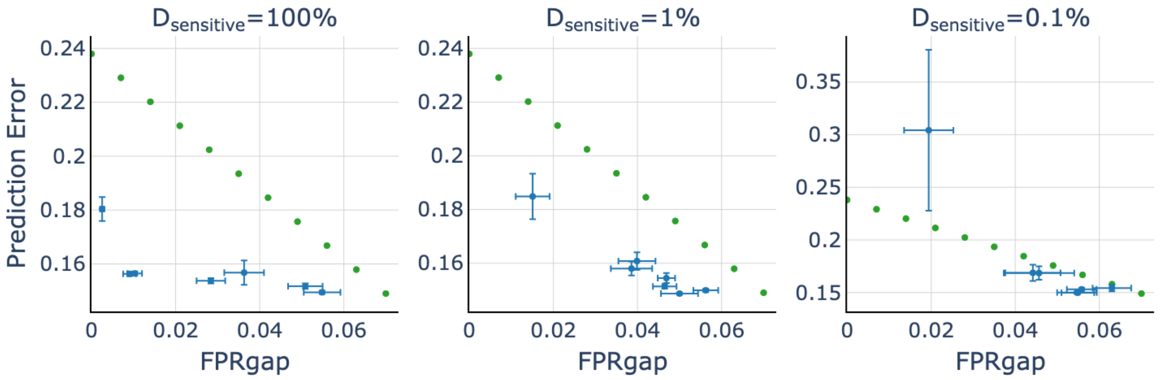
Pre-, in-, post-processing mitigations need training data with group labels.

Issue #1: Group labels are difficult to collect.

e.g. group labels are often sensitive attributes (e.g. gender, ethnicity, age etc)

What happens when group labels are scarce?

Naive baseline: *"predict according to pre-trained model with probability p, and predict 0 with probability (1-p)"*
In-processing mitigation: state-of-the-art MinDiff method



SOTA method as poor as naive baseline

Dataset: Adult
Y = income; A = gender

Figure source: https://arxiv.org/abs/2312.02592

Challenges faced by fairness mitigations

Labeled data can be expensive to collect.

Challenges faced by fairness mitigations

Labeled data can be expensive to collect.

Issue #2: Class label scarcity can amplify unfairness.

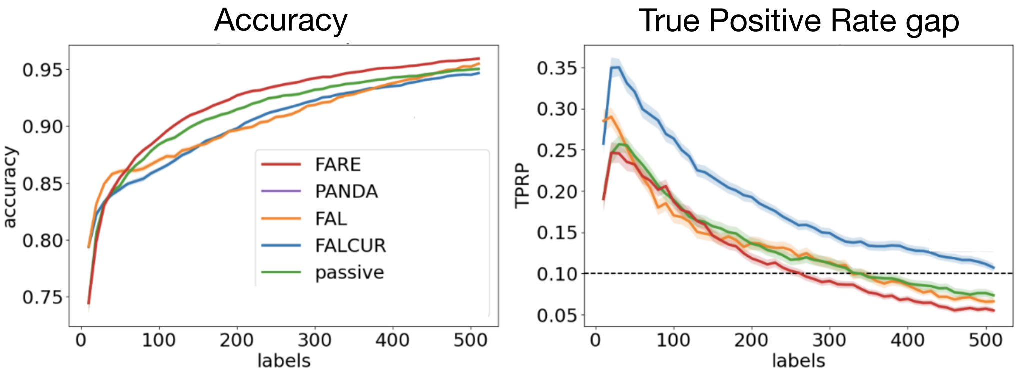
Challenges faced by fairness mitigations

Labeled data can be expensive to collect.

Issue #2: Class label scarcity can amplify unfairness.

What happens in the low-label regime?

e.g. fair active learning strategies



Dataset: Communities & Crime
Y = crime rate; A = ethnicity

Figure source: https://arxiv.org/pdf/2312.08559

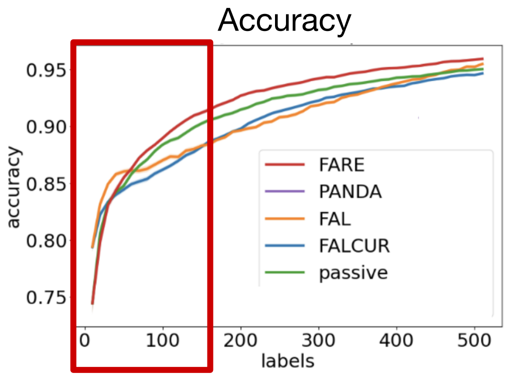
Challenges faced by fairness mitigations

Labeled data can be expensive to collect.

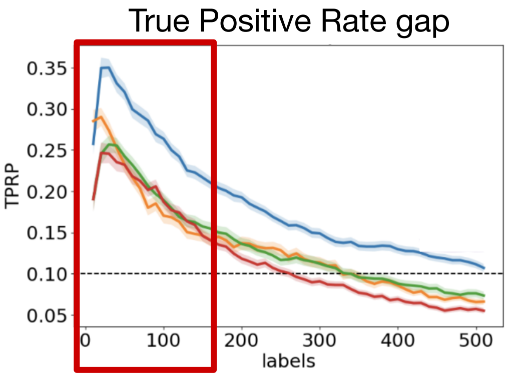
Issue #2: Class label scarcity can amplify unfairness.

What happens in the low-label regime?

e.g. fair active learning strategies



less accurate



more unfair

worse accuracy AND fairness in low-label regime

Dataset: Communities & Crime
Y = crime rate; A = ethnicity

Figure source: https://arxiv.org/pdf/2312.08559

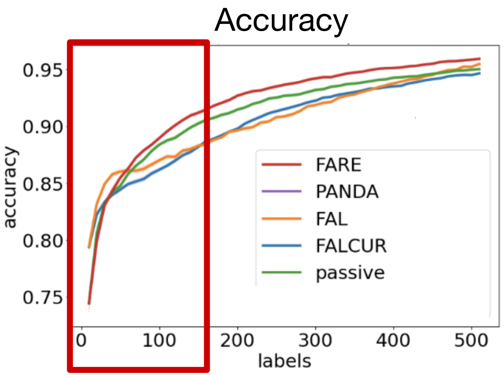
Challenges faced by fairness mitigations

Labeled data can be expensive to collect.

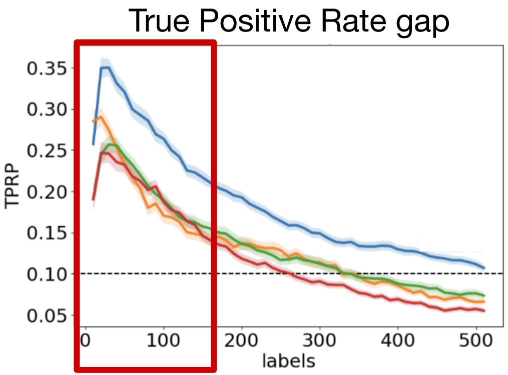
Issue #2: Class label scarcity can amplify unfairness.

What happens in the low-label regime?

e.g. fair active learning strategies



less accurate



more unfair

worse accuracy AND fairness in low-label regime

intersectional fairness amplifies data scarcity

e.g. *avoid discriminating against Hispanic females aged 30-40*

Dataset: Communities & Crime
Y = crime rate; A = ethnicity

Figure source: https://arxiv.org/pdf/2312.08559

Fairness – Outline

Fairness with **partial** group labels

Fairness with **no** group labels

Fairness in the low-label regime

Fairness with partial group labels

Problem setting: Fairness with partial group labels

Problem setting: Fairness with partial group labels

$$\boxed{\mathcal{D}_{\text{pred}} = \{(X_i, Y_i)\}_{i=1}^{n}} \Rightarrow$$ **large dataset**

covariates X; class labels Y

Problem setting: Fairness with partial group labels

$\boxed{\mathcal{D}_{\text{pred}} = \{(X_i, Y_i)\}_{i=1}^{n}}$ ⇨ **large dataset**

covariates X; class labels Y

$\boxed{\mathcal{D}_{\text{sensitive}} = \{(X_i, Y_i, A_i)\}_{i=1}^{n}}$ ⇨ **small dataset**

(X, Y) + sensitive attribute A i.e. group label

Problem setting: Fairness with partial group labels

$\mathcal{D}_{\text{pred}} = \{(X_i, Y_i)\}_{i=1}^{n}$ ⇨ **large dataset**

covariates X; class labels Y

$\mathcal{D}_{\text{sensitive}} = \{(X_i, Y_i, A_i)\}_{i=1}^{n}$ ⇨ **small dataset**

(X, Y) + sensitive attribute A i.e. group label

Case study: In-processing mitigations with partial group labels

Reminder: $\text{OPT}_{\text{IP}} : \arg\min_{f} \mathcal{L}_{pred}(f; \mathcal{D}_{\text{pred}}) + \lambda \mathcal{L}_{fair}(f; \mathcal{D}_{\text{sensitive}})$

Problem setting: Fairness with partial group labels

$\boxed{\mathcal{D}_{\text{pred}} = \{(X_i, Y_i)\}_{i=1}^{n}}$ ⇨ **large dataset**

covariates X; class labels Y

$\boxed{\mathcal{D}_{\text{sensitive}} = \{(X_i, Y_i, A_i)\}_{i=1}^{n}}$ ⇨ **small dataset**

(X, Y) + sensitive attribute A i.e. group label

Case study: In-processing mitigations with partial group labels

Reminder: $\text{OPT}_{\text{IP}} : \arg\min_f \mathcal{L}_{pred}(f; \boxed{\mathcal{D}_{\text{pred}}}) + \lambda \mathcal{L}_{fair}(f; \boxed{\mathcal{D}_{\text{sensitive}}})$

Problem setting: Fairness with partial group labels

$\boxed{\mathcal{D}_{\text{pred}} = \{(X_i, Y_i)\}_{i=1}^{n}}$ large dataset

covariates X; class labels Y

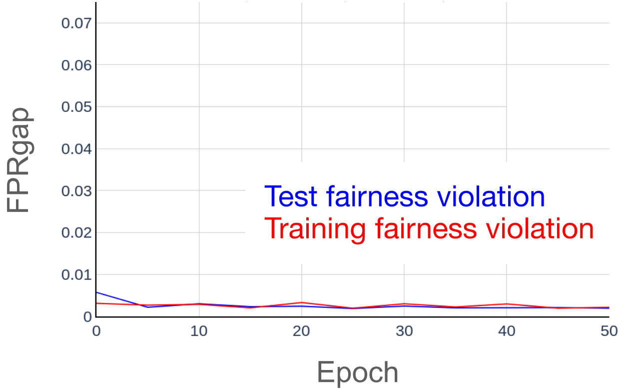
$\boxed{\mathcal{D}_{\text{sensitive}} = \{(X_i, Y_i, A_i)\}_{i=1}^{n}}$ small dataset

(X, Y) + sensitive attribute A i.e. group label

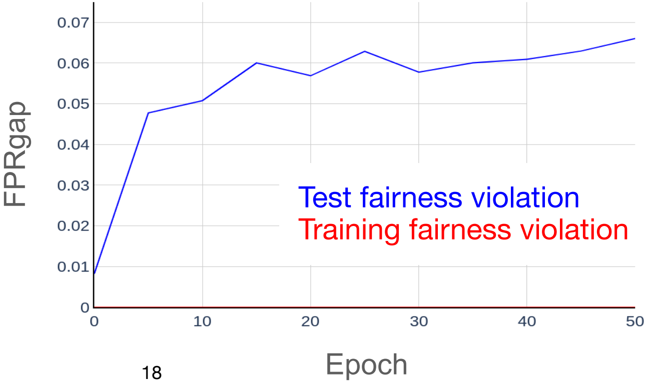
Case study: In-processing mitigations with partial group labels

Reminder: $\mathrm{OPT}_{\mathrm{IP}} : \arg\min_f \mathcal{L}_{pred}(f; \boxed{\mathcal{D}_{\text{pred}}}) + \lambda \mathcal{L}_{fair}(f; \boxed{\mathcal{D}_{\text{sensitive}}})$

$|\mathcal{D}_{\text{sensitive}}| = 100\% \cdot |\mathcal{D}_{\text{pred}}|$



$|\mathcal{D}_{\text{sensitive}}| = 0.1\% \cdot |\mathcal{D}_{\text{pred}}|$



overfitting!

How to deal with partial group labels?

High level strategies

1. Use proxy for missing sensitive attributes

1. Make fairness mitigations more sample efficient

How to deal with partial group labels?

High level strategies

1. **Use proxy for missing sensitive attributes**

1. Make fairness mitigations more sample efficient

Proxy for missing sensitive attributes

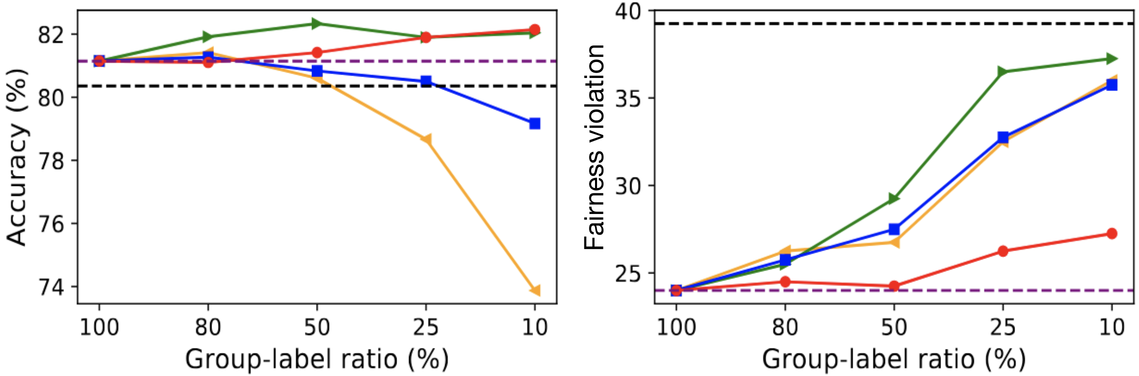
Strategies for missing sensitive attributes A

e.g. process data + in-processing fairness mitigation

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University



Dataset: UTKFace
Y = age group; A = ethnicity

Proxy for missing sensitive attributes

Strategies for missing sensitive attributes A

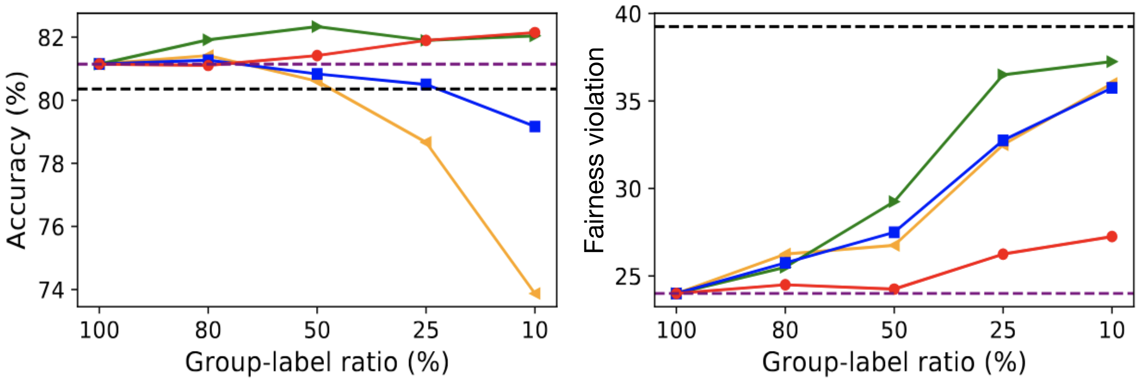
e.g. process data + in-processing fairness mitigation

- drop samples with missing A

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University



Dataset: UTKFace
Y = age group; A = ethnicity

Proxy for missing sensitive attributes

Strategies for missing sensitive attributes A

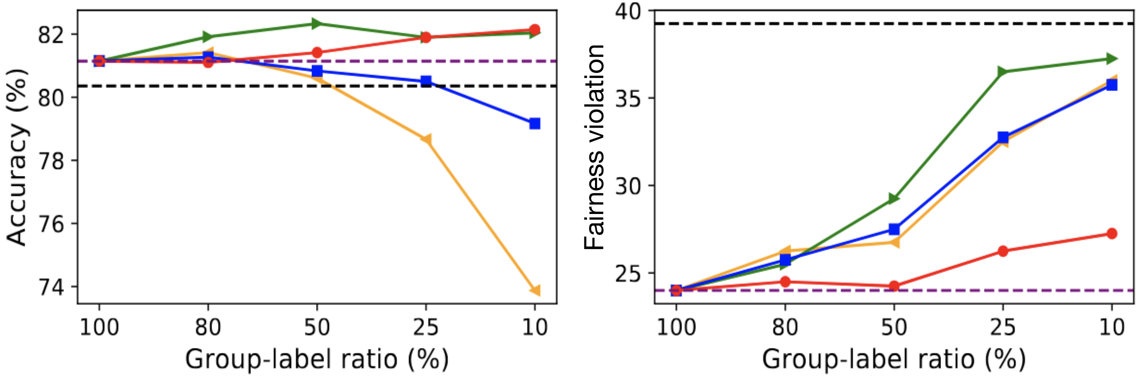
e.g. process data + in-processing fairness mitigation

- drop samples with missing A
- impute A uniformly at random

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University



Dataset: UTKFace

Y = age group; A = ethnicity

Proxy for missing sensitive attributes

Strategies for missing sensitive attributes A

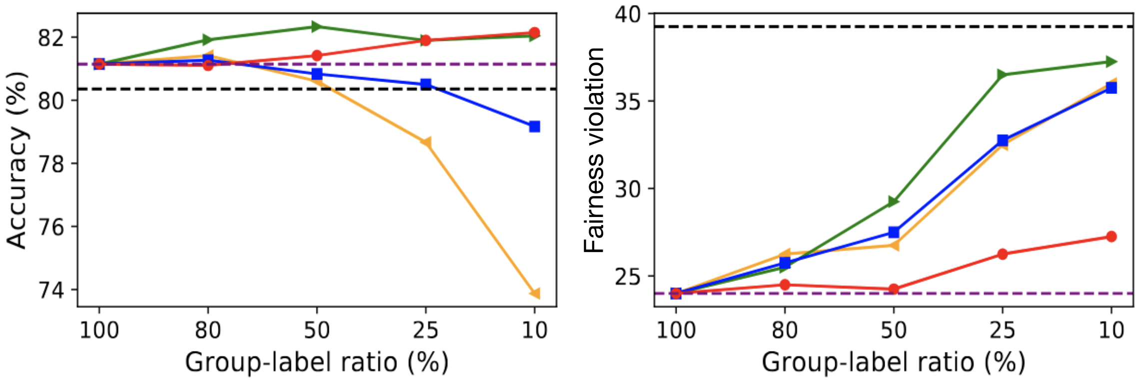
e.g. process data + in-processing fairness mitigation

- drop samples with missing A
- impute A uniformly at random
- pseudo-labels from classifier $\hat{f}_A(x)$ trained on $D_{sensitive} = \{(x_i, a_i)\}_{i=1}^{n}$

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University



Dataset: UTKFace

Y = age group; A = ethnicity

Proxy for missing sensitive attributes

Learning Fair Classifiers with Partially Annotated Group Labels

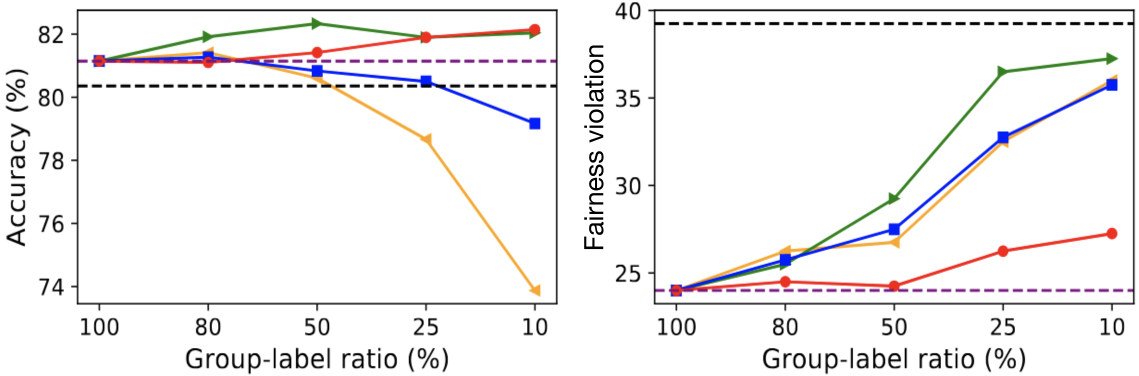
Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University

Strategies for missing sensitive attributes A

e.g. process data + in-processing fairness mitigation

- drop samples with missing A
- impute A uniformly at random
- pseudo-labels from classifier $\hat{f}_A(x)$ trained on $D_{sensitive} = \{(x_i, a_i)\}_{i=1}^{n}$
- pseudo-labels only on high-confidence samples otherwise random value for A



Dataset: UTKFace
Y = age group; A = ethnicity

Proxy for missing sensitive attributes

Strategies for missing sensitive attributes A

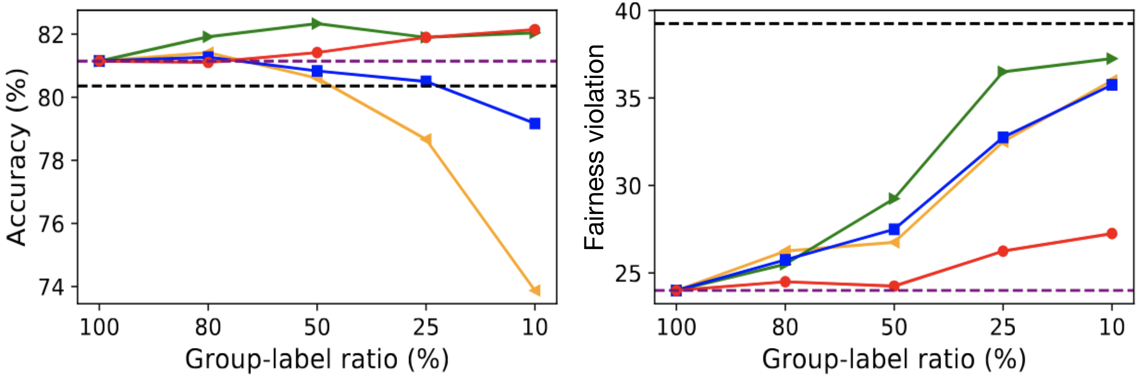
e.g. process data + in-processing fairness mitigation

- drop samples with missing A
- impute A uniformly at random
- pseudo-labels from classifier $\hat{f}_A(x)$ trained on $D_{sensitive} = \{(x_i, a_i)\}_{i=1}^{n}$
- pseudo-labels only on high-confidence samples otherwise random value for A

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University



naive mitigations are suboptimal

Dataset: UTKFace
Y = age group; A = ethnicity

High-confidence group pseudo-labels

Predict missing sensitive attributes A:

$$\hat{a} = \begin{cases} \arg\max \hat{f}_A(x) & \hat{f}_A(x) > \tau \\ \text{draw from } P(A|Y=y) & \text{otherwise} \end{cases}$$

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University

High-confidence group pseudo-labels

Predict missing sensitive attributes A: ?

$$\hat{a} = \begin{cases} \arg\max \hat{f}_A(x) & \hat{f}_A(x) > \boxed{\tau} \\ \text{draw from } P(A|Y=y) & \text{otherwise} \end{cases}$$

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University

High-confidence group pseudo-labels

Predict missing sensitive attributes A: **?**

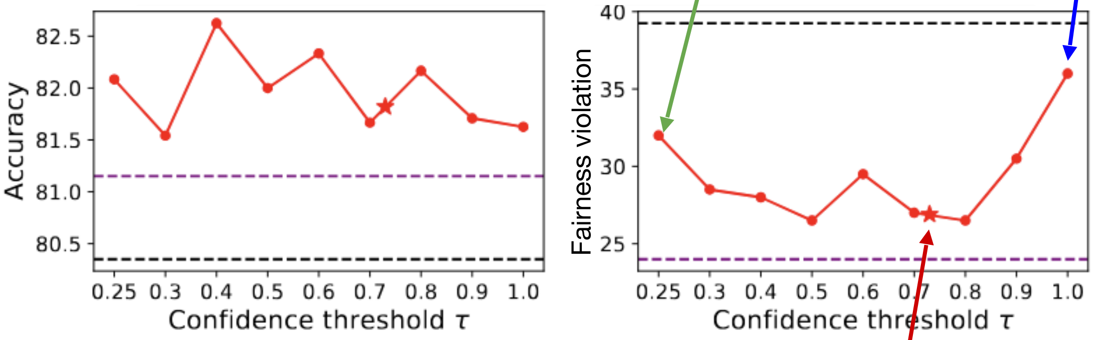
$$\hat{a} = \begin{cases} \arg\max \hat{f}_A(x) & \hat{f}_A(x) > \boxed{\tau} \\ \text{draw from } P(A|Y=y) & \text{otherwise} \end{cases}$$

Learning Fair Classifiers with Partially Annotated Group Labels

Sangwon Jung[1*] Sanghyuk Chun[2†] Taesup Moon[1,3†]

[1] Department of ECE/ASRI, Seoul National University [2] NAVER AI Lab
[3] Interdisciplinary Program in Artificial Intelligence, Seoul National University

Use validation set
(with group labels)



Is thresholding confidence an optimal strategy?

Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved

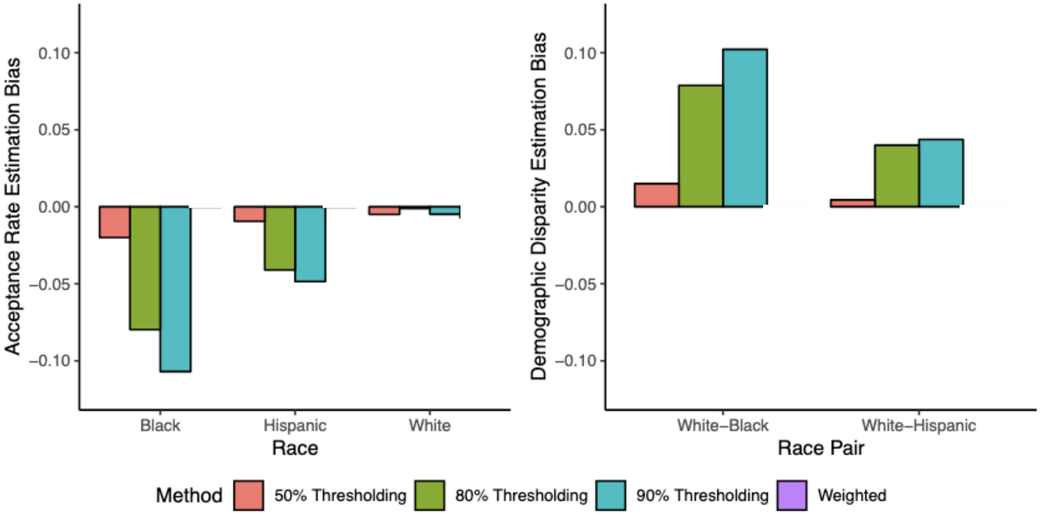
Jiahao Chen
cjiahao@gmail.com

Nathan Kallus
Cornell Tech
New York, New York, USA
kallus@cornell.edu

Xiaojie Mao*
Cornell Tech
New York, New York, USA
xm77@cornell.edu

Geoffry Svacha
svacha@gmail.com

Madeleine Udell
Cornell University
Ithaca, New York, USA
udell@cornell.edu



Is thresholding confidence an optimal strategy?

Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved

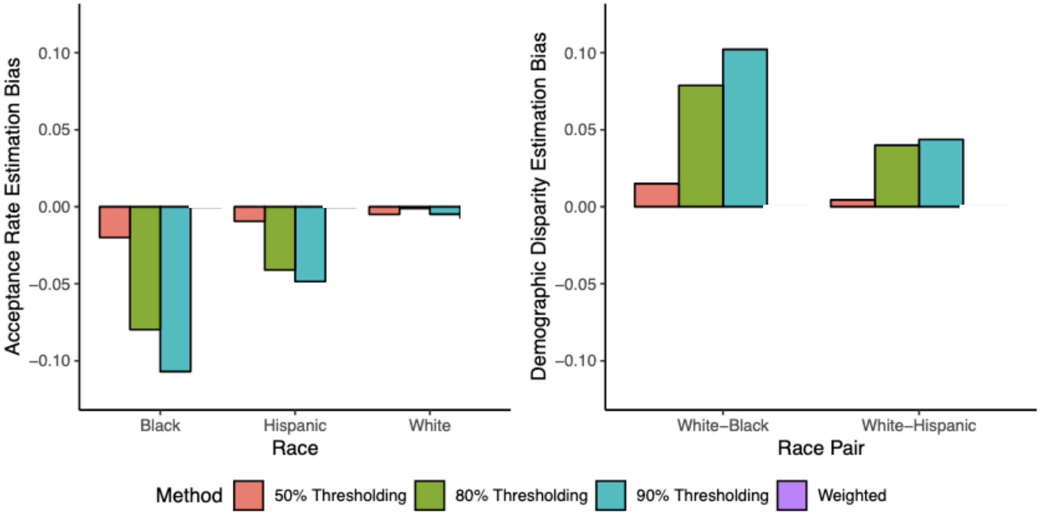
Jiahao Chen
cjiahao@gmail.com

Nathan Kallus
Cornell Tech
New York, New York, USA
kallus@cornell.edu

Xiaojie Mao*
Cornell Tech
New York, New York, USA
xm77@cornell.edu

Geoffry Svacha
svacha@gmail.com

Madeleine Udell
Cornell University
Ithaca, New York, USA
udell@cornell.edu



thresholding confidence can lead to poor fairness estimation

Dataset: HMDA

Y = '*was loan approved?*'
A = ethnicity

Is thresholding confidence an optimal strategy?

Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved

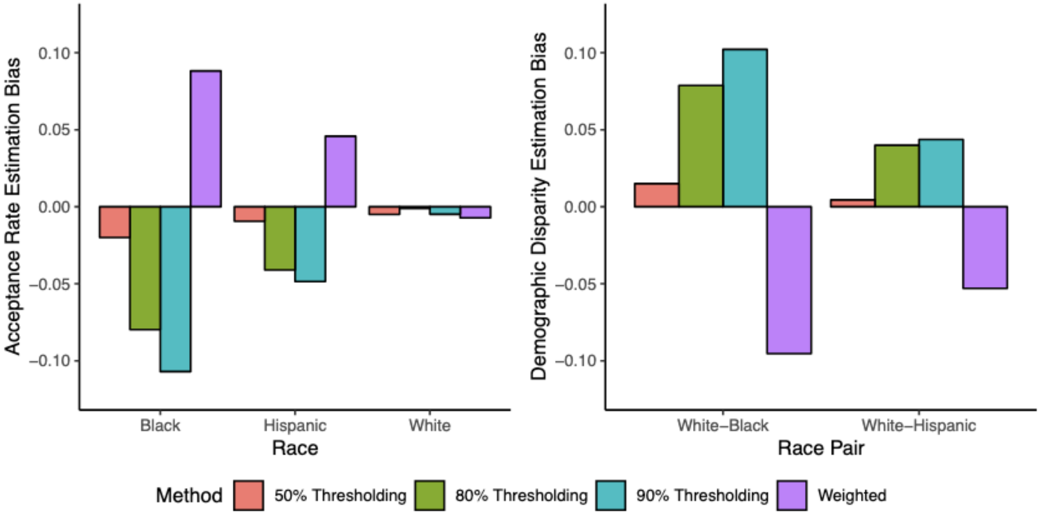
Jiahao Chen
cjiahao@gmail.com

Nathan Kallus
Cornell Tech
New York, New York, USA
kallus@cornell.edu

Xiaojie Mao*
Cornell Tech
New York, New York, USA
xm77@cornell.edu

Geoffry Svacha
svacha@gmail.com

Madeleine Udell
Cornell University
Ithaca, New York, USA
udell@cornell.edu



thresholding confidence can lead to poor fairness estimation

Weighted estimator also fails (but differently)

Dataset: HMDA
Y = '*was loan approved?*'
A = ethnicity

Summary: Using a proxy group label

Effective at mitigating unfairness

as long as sufficient group-labeled validation data is available

e.g. necessary to select hyperparameters like confidence threshold

Summary: Using a proxy group label

Effective at mitigating unfairness

as long as sufficient group-labeled validation data is available

e.g. necessary to select hyperparameters like confidence threshold

Statistically, often easy to predict the sensitive attribute from little data

but it can have ethical concerns and can amplify/hide biases in the data

Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data

Michael Veale [1] and Reuben Binns [2]

Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved

Jiahao Chen
cjiahao@gmail.com

Nathan Kallus
Cornell Tech
New York, New York, USA
kallus@cornell.edu

Xiaojie Mao*
Cornell Tech
New York, New York, USA
xm77@cornell.edu

Geoffry Svacha
svacha@gmail.com

Madeleine Udell
Cornell University
Ithaca, New York, USA
udell@cornell.edu

Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, PA
kjholste@cs.cmu.edu

Jennifer Wortman Vaughan
Microsoft Research
New York, NY
jenn@microsoft.com

Hal Daumé III
Microsoft Research &
University of Maryland
New York, NY
me@hal3.name

Miroslav Dudík
Microsoft Research
New York, NY
mdudik@microsoft.com

Hanna Wallach
Microsoft Research
New York, NY
wallach@microsoft.com

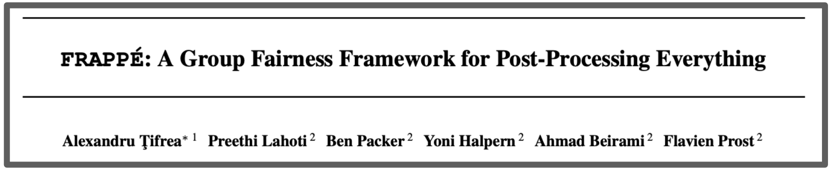
How to deal with partial group labels?

High level strategies

1. Use proxy for missing sensitive attributes

1. **Make fairness mitigations more sample efficient**

Modular sample-efficient fairness mitigations



FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea* [1] Preethi Lahoti [2] Ben Packer [2] Yoni Halpern [2] Ahmad Beirami [2] Flavien Prost [2]

Setup: Equal Opportunity on Adult dataset

Naive baseline: *"predict according to f_{base} with probability p, and predict 0 with probability (1-p)"*

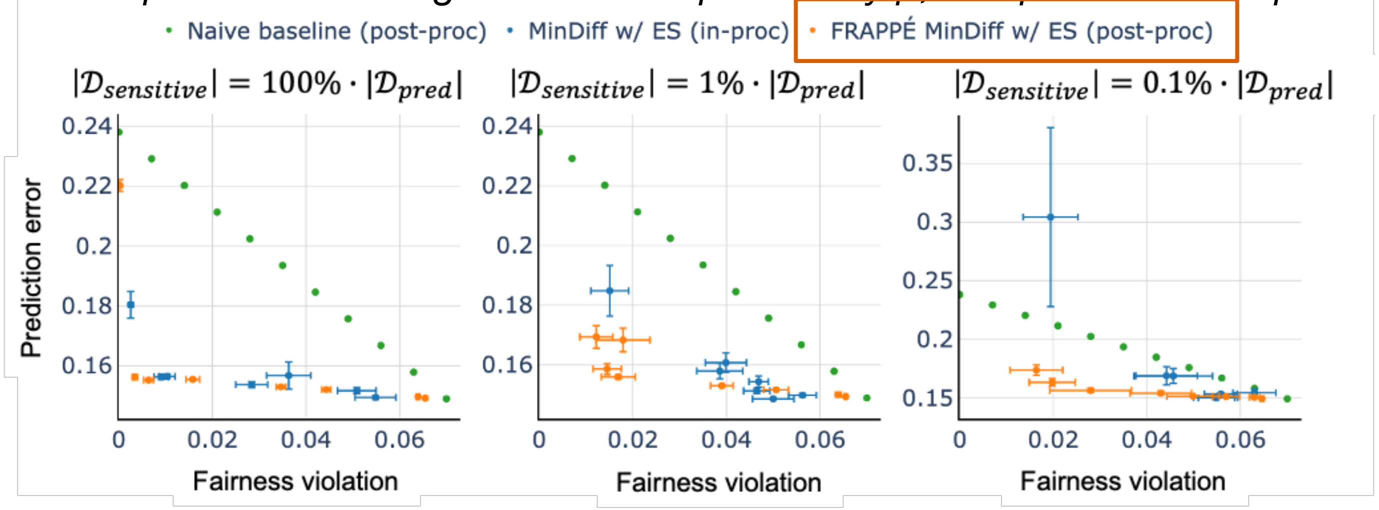
Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

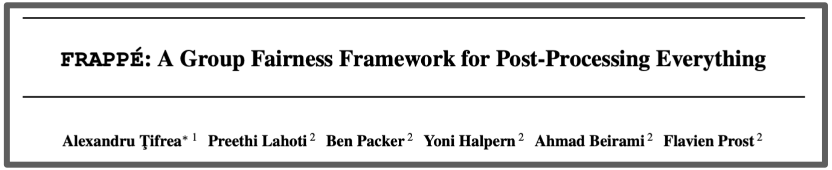
Alexandru Țifrea*[1] Preethi Lahoti[2] Ben Packer[2] Yoni Halpern[2] Ahmad Beirami[2] Flavien Prost[2]

Setup: Equal Opportunity on Adult dataset

Naive baseline: *"predict according to f_{base} with probability p, and predict 0 with probability (1-p)"*

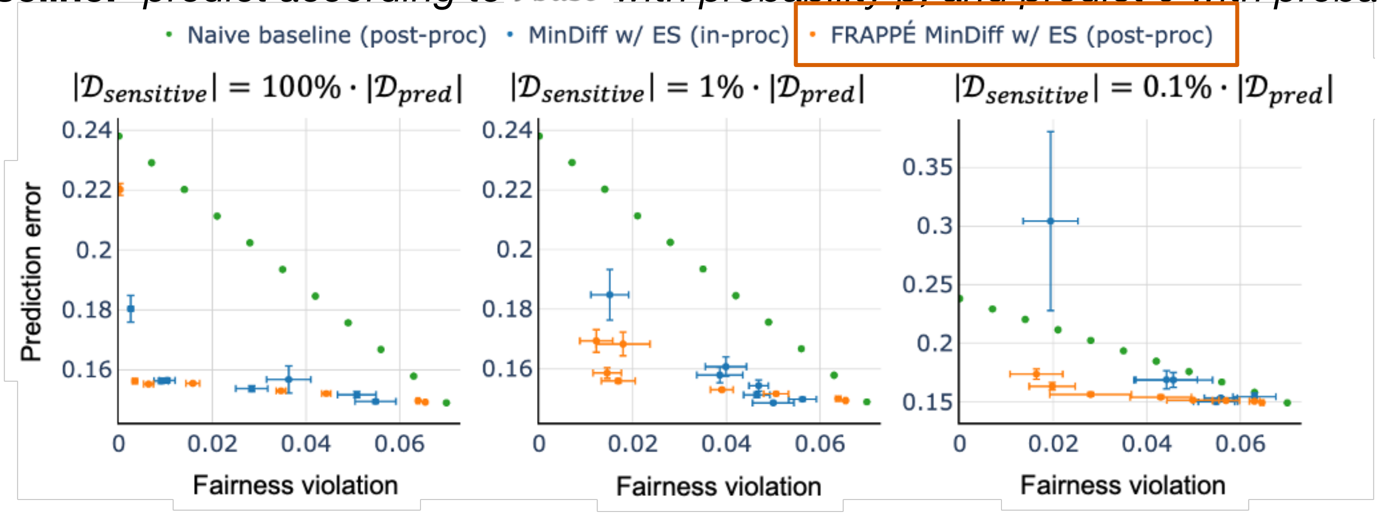


Modular sample-efficient fairness mitigations



Setup: Equal Opportunity on Adult dataset

Naive baseline: *"predict according to f_{base} with probability p, and predict 0 with probability (1-p)"*



computation time ~8x faster than in-processing

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea[*,1] Preethi Lahoti[2] Ben Packer[2] Yoni Halpern[2] Ahmad Beirami[2] Flavien Prost[2]

Accurate but unfair model:

$$f_{base} := \operatorname*{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$$

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea*[1] Preethi Lahoti[2] Ben Packer[2] Yoni Halpern[2] Ahmad Beirami[2] Flavien Prost[2]

Accurate but unfair model:

$$f_{base} := \operatorname*{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$$

Proposed post-hoc transformation:

$$f_{fair}(x) = f_{base}(x) + T(x)$$

(logit additive for classification)

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Țifrea* [1] Preethi Lahoti [2] Ben Packer [2] Yoni Halpern [2] Ahmad Beirami [2] Flavien Prost [2]

Accurate but unfair model: $f_{base} := \operatorname{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$

Proposed post-hoc transformation: $f_{fair}(x) = f_{base}(x) + \underline{T(x)}$ **not group-dependent**

(logit additive for classification)

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea* [1] Preethi Lahoti [2] Ben Packer [2] Yoni Halpern [2] Ahmad Beirami [2] Flavien Prost [2]

Accurate but unfair model:

$$f_{base} := \operatorname*{argmin}_{\boldsymbol{f}} \boxed{\mathcal{L}_{pred}(\boldsymbol{f}; \mathcal{D}_{pred})}$$

Proposed post-hoc transformation:

$$f_{fair}(x) = f_{base}(x) + \underline{T(x)}$$

not group-dependent

(logit additive for classification)

In-processing:

$$OPT_{IP}(\boldsymbol{f}; \lambda) = \boxed{\mathcal{L}_{pred}(\boldsymbol{f}; \mathcal{D}_{pred})} + \boxed{\lambda \mathcal{L}_{fair}(\boldsymbol{f}; \mathcal{D}_{sensitive})}$$

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea* [1] Preethi Lahoti [2] Ben Packer [2] Yoni Halpern [2] Ahmad Beirami [2] Flavien Prost [2]

Accurate but unfair model:

$$f_{base} := \operatorname*{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$$

Proposed post-hoc transformation:

$$f_{fair}(x) = f_{base}(x) + \underline{T(x)}$$

not group-dependent

(logit additive for classification)

In-processing:

$$OPT_{IP}(f; \lambda) = \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})} + \boxed{\lambda \mathcal{L}_{fair}(f; \mathcal{D}_{sensitive})}$$

Proposed post-processing for learning ***T***:

$$OPT_{PP}(T; \lambda) = \boxed{Discrepancy((f_{base} + T) \parallel f_{base}; \mathcal{D}_{unlab})} + \boxed{\lambda \mathcal{L}_{fair}(f_{base} + T; \mathcal{D}_{sensitive})}$$

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea*[1] Preethi Lahoti[2] Ben Packer[2] Yoni Halpern[2] Ahmad Beirami[2] Flavien Prost[2]

Accurate but unfair model:

$$f_{base} := \operatorname*{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$$

Proposed post-hoc transformation:

$$f_{fair}(x) = f_{base}(x) + \underline{T(x)}$$

(logit additive for classification)

not group-dependent

In-processing:

$$OPT_{IP}(f;\lambda) = \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})} + \boxed{\lambda \mathcal{L}_{fair}(f; \mathcal{D}_{sensitive})}$$

Proposed post-processing for learning ***T***:

$$OPT_{PP}(T;\lambda) = \boxed{Discrepancy((f_{base} + T) \parallel f_{base}; \mathcal{D}_{unlab})} + \underbrace{\boxed{\lambda \mathcal{L}_{fair}(f_{base} + T; \mathcal{D}_{sensitive})}}_{\text{any notion of fairness}}$$

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea*[1] Preethi Lahoti[2] Ben Packer[2] Yoni Halpern[2] Ahmad Beirami[2] Flavien Prost[2]

Accurate but unfair model: $f_{base} := \operatorname*{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$

Proposed post-hoc transformation: $f_{fair}(x) = f_{base}(x) + \underline{T(x)}$ **not group-dependent**

(logit additive for classification)

In-processing: $OPT_{IP}(f; \lambda) = \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})} + \boxed{\lambda \mathcal{L}_{fair}(f; \mathcal{D}_{sensitive})}$

Proposed post-processing for learning ***T***:

$$OPT_{PP}(T; \lambda) = \underbrace{\boxed{Discrepancy((f_{base} + T) \parallel f_{base}; \mathcal{D}_{unlab})}}_{\substack{\text{output discrepancy} \\ \text{related to } \mathcal{L}_{pred} \text{ e.g. MSE, KL divergence etc}}} + \underbrace{\boxed{\lambda \mathcal{L}_{fair}(f_{base} + T; \mathcal{D}_{sensitive})}}_{\text{any notion of fairness}}$$

Modular sample-efficient fairness mitigations

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea*[1] Preethi Lahoti[2] Ben Packer[2] Yoni Halpern[2] Ahmad Beirami[2] Flavien Prost[2]

Accurate but unfair model:

$$f_{base} := \operatorname*{argmin}_{f} \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})}$$

Proposed post-hoc transformation:

$$f_{fair}(x) = f_{base}(x) + \underline{T(x)}$$

(logit additive for classification)

not group-dependent

In-processing:

$$OPT_{IP}(f; \lambda) = \boxed{\mathcal{L}_{pred}(f; \mathcal{D}_{pred})} + \boxed{\lambda \mathcal{L}_{fair}(f; \mathcal{D}_{sensitive})}$$

Proposed post-processing for learning ***T***:

$$OPT_{PP}(T; \lambda) = \underbrace{\boxed{Discrepancy((f_{base} + T) \parallel f_{base}; \mathcal{D}_{unlab})}}_{\text{output discrepancy}} + \underbrace{\boxed{\lambda \mathcal{L}_{fair}(f_{base} + T; \mathcal{D}_{sensitive})}}_{\text{any notion of fairness}}$$

output discrepancy
related to $\mathcal{L}_{pred}$ e.g. MSE, KL divergence etc

any notion of fairness

$$\boxed{\mathcal{D}_{unlab} = \{x_i\}_{i=1}^{N}}$$

unlabeled data

Instances of modular multi-objective learning

LLM alignment

Asymptotics of Language Model Alignment

Joy Qiping Yang
University of Sydney
Sydney, Australia
qyan6238@uni.sydney.edu.au

Salman Salamatian
Massachusetts Institute of Technology
Cambridge, MA, USA
salmansa@mit.edu

Ziteng Sun, Ananda Theertha Suresh, Ahmad Beirami
Google Research
New York, NY, USA
{zitengsun, theertha, beirami}@google.com

Out-of-domain generalization

OVERPARAMETERISATION AND WORST-CASE GENERALISATION: FRIEND OR FOE?

Aditya Krishna Menon, Ankit Singh Rawat & Sanjiv Kumar
Google Research
New York, NY
{adityakmenon,ankitsrawat,sanjivk}@google.com

Adversarial robustness

Understanding and Mitigating the Tradeoff Between Robustness and Accuracy

Aditi Raghunathan [*1] **Sang Michael Xie** [*1] **Fanny Yang** [2] **John C. Duchi** [1] **Percy Liang** [1]

Unlabeled Data Improves Adversarial Robustness

Yair Carmon[*]
Stanford University
yairc@stanford.edu

Aditi Raghunathan[*]
Stanford University
aditir@stanford.edu

Ludwig Schmidt
UC Berkeley
ludwig@berkeley.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

John C. Duchi
Stanford University
jduchi@stanford.edu

Summary: Modular fairness mitigations

More sample efficient than in-processing

iff learning the fairness correction module is statistically efficient

e.g. $T(x)$ is not a complex function, $T(x)$ has low-dimensional structure (e.g. sparsity)

Summary: Modular fairness mitigations

More sample efficient than in-processing

iff learning the fairness correction module is statistically efficient

e.g. T(x) is not a complex function, T(x) has low-dimensional structure (e.g. sparsity)

Effective technique to induce any notion of fairness

iff fairness violations can be measured from <u>observational</u> data

e.g. T(X) implicitly estimates P(A|X) which might unidentifiable from observational data

Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination

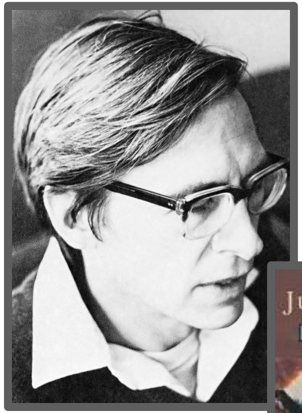
Nathan Kallus
Cornell University, kallus@cornell.edu

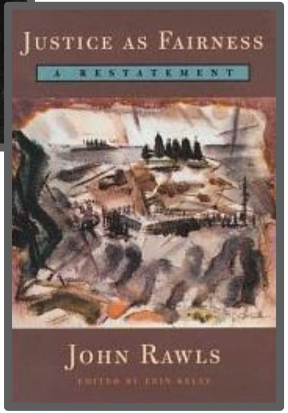
Xiaojie Mao
Cornell University, xm77@cornell.edu

Angela Zhou
Cornell University, az434@cornell.edu

Fairness with <u>no</u> group labels

Fairness as worst-group performance

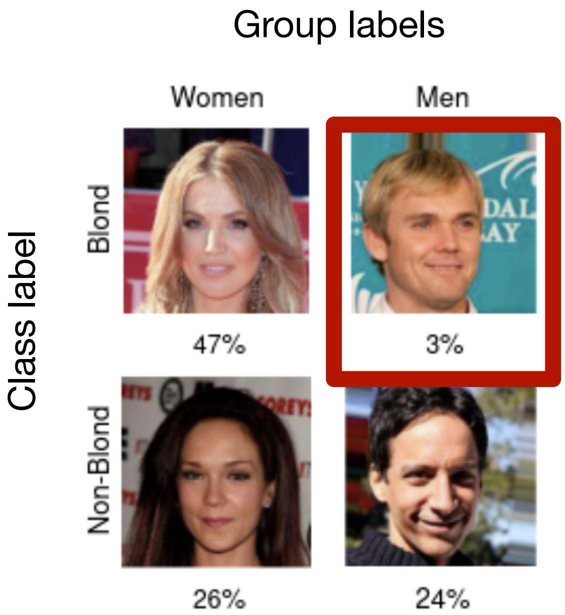




Definition A hypothesis $h^\star$ satisfies Rawlsian max-min fairness if it maximizes the accuracy of the worst-off group

$$h^\star = \arg\max_{h} \min_{a \in \mathcal{A}} Acc\left(h | A = a\right)$$

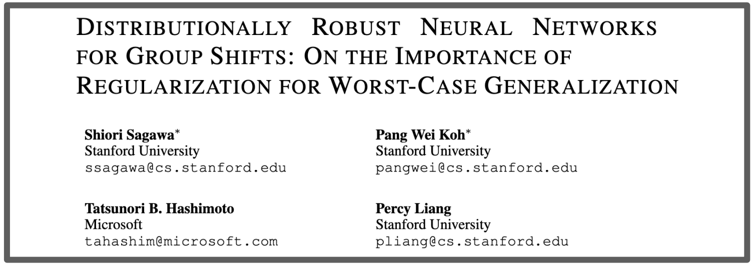
Mitigation strategies for worst-group fairness



CelebA dataset

If we know group labels:

- importance weighting (IW)
- group distributionally robust optimization (GDRO)


DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION

Shiori Sagawa*
Stanford University
ssagawa@cs.stanford.edu

Pang Wei Koh*
Stanford University
pangwei@cs.stanford.edu

Tatsunori B. Hashimoto
Microsoft
tahashim@microsoft.com

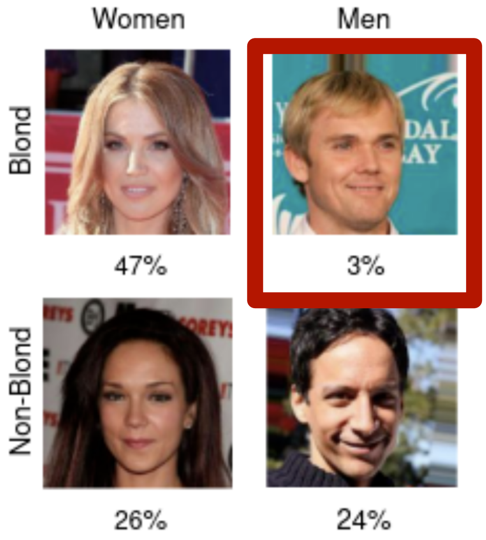
Percy Liang
Stanford University
pliang@cs.stanford.edu

Image source: https://arxiv.org/abs/2302.01385

Mitigation strategies for worst-group fairness

Group labels

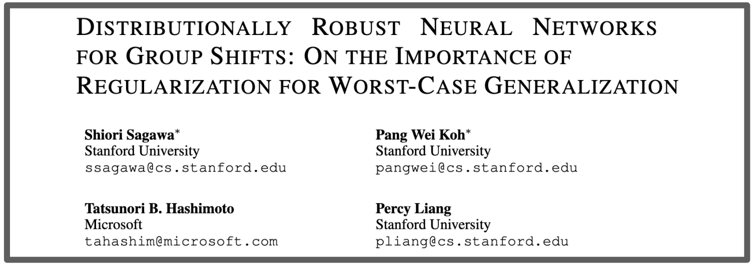
Class label



CelebA dataset

If we know group labels:

- importance weighting (IW)
- group distributionally robust optimization (GDRO)



DISTRIBUTIONALLY ROBUST NEURAL NETWORKS FOR GROUP SHIFTS: ON THE IMPORTANCE OF REGULARIZATION FOR WORST-CASE GENERALIZATION

Shiori Sagawa*
Stanford University
ssagawa@cs.stanford.edu

Pang Wei Koh*
Stanford University
pangwei@cs.stanford.edu

Tatsunori B. Hashimoto
Microsoft
tahashim@microsoft.com

Percy Liang
Stanford University
pliang@cs.stanford.edu

In the absence of group labels:

Two-stage method

1) identify worse-off group
2) employ e.g. IW/GDRO to improve worst-group error

Image source: https://arxiv.org/abs/2302.01385

Fairness via distributionally robust optimization (DRO)

$$\mathcal{R}_{erm}(\theta) := \mathbb{E}_P\left[\ell(\theta; Z)\right]$$

Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1,2] Megha Srivastava [1] Hongseok Namkoong [3] Percy Liang [1]

Fairness via distributionally robust optimization (DRO)

$$\mathcal{R}_{erm}(\theta) := \mathbb{E}_P\left[\ell(\theta; Z)\right]$$

Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1,2] Megha Srivastava [1] Hongseok Namkoong [3] Percy Liang [1]

$$\mathcal{R}_{\mathrm{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]$$

worst-case loss wrt the uncertainty set Q

Fairness via distributionally robust optimization (DRO)

Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1 2] Megha Srivastava [1] Hongseok Namkoong [3] Percy Liang [1]

$$\mathcal{R}_{erm}(\theta) := \mathbb{E}_P\left[\ell\left(\theta; Z\right)\right]$$

$$\mathcal{R}_{\mathrm{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]$$

worst-case loss wrt the uncertainty set Q

P = (marginal) data distribution

Fairness via distributionally robust optimization (DRO)

Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1,2] Megha Srivastava [1] Hongseok Namkoong [3] Percy Liang [1]

$$\mathcal{R}_{erm}(\theta) := \mathbb{E}_P\left[\ell(\theta; Z)\right]$$

$$\mathcal{R}_{\mathrm{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]$$

worst-case loss wrt the uncertainty set Q

P = (marginal) data distribution

r = radius of uncertainty set

Fairness via distributionally robust optimization (DRO)

Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1 2] Megha Srivastava [1] Hongseok Namkoong [3] Percy Liang [1]

$$\mathcal{R}_{erm}(\theta) := \mathbb{E}_P\left[\ell(\theta; Z)\right]$$

$$\mathcal{R}_{\mathrm{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]$$

worst-case loss wrt the uncertainty set Q

P = (marginal) data distribution

r = radius of uncertainty set

determined by

$\alpha_{\min}$ minority group proportion

Fairness via distributionally robust optimization (DRO)

Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1,2] Megha Srivastava [1] Hongseok Namkoong [3] Percy Liang [1]

$$\mathcal{R}_{erm}(\theta) := \mathbb{E}_P\left[\ell\left(\theta; Z\right)\right]$$

$$\mathcal{R}_{\text{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]$$

worst-case loss wrt the uncertainty set Q

P = (marginal) data distribution

r = radius of uncertainty set

determined by

$\alpha_{\min}$ minority group proportion

What if no group labels available?

A: pick a lower bound for α_{min}

Detect worst-group using a biased classifier

DRO: upweights *high-loss samples.*

Alternative: Two-stage method

1) use **biased** classifier to identify *error set*
2) train fair classifier via IW / GroupDRO

Detect worst-group using a biased classifier

DRO: upweights *high-loss samples.*

Alternative: Two-stage method

1) use **biased** classifier to identify *error set*
2) train fair classifier via IW / GroupDRO

Why are two-stage methods expected to work?

Majority group

Intuition: a biased classifier will predict based on the stronger correlation.
e.g. background

Detect worst-group using a biased classifier

DRO: upweights *high-loss samples.*

Alternative: Two-stage method

1) use **biased** classifier to identify *error set*
2) train fair classifier via IW / GroupDRO

Why are two-stage methods expected to work?

Majority group

Minority group

Intuition: a biased classifier will predict based on the stronger correlation.
e.g. background

incorrect predictions where spurious correlation does not hold
i.e. minority groups

How to train a biased classifier?

Setting 1: group labels available for validation set

How to train a biased classifier?

Setting 1: group labels available for validation set

Examples:

- heavy regularization
  e.g. via early stopping

Just Train Twice: Improving Group Robustness without Training Group Information

Evan Zheran Liu[*1] Behzad Haghgoo[*1] Annie S. Chen[*1] Aditi Raghunathan[1] Pang Wei Koh[1]
Shiori Sagawa[1] Percy Liang[1] Chelsea Finn[1]

- custom loss function
  e.g. amplify "easy" examples

Learning from Failure: Training Debiased Classifier from Biased Classifier

Junhyun Nam[1] Hyuntak Cha[2] Sungsoo Ahn[1] Jaeho Lee[1] Jinwoo Shin[1,2]
[1]School of Electrical Engineering, KAIST
[2]Graduate School of AI, KAIST
{junhyun.nam, hyuntak.cha, sungsoo.ahn, jaeho-lee, jinwoos}@kaist.ac.kr

Use **worst-group validation error** to select regularization strength, IW weights etc.

How to train a biased classifier?

Setting 2: no group labels at all

How to train a biased classifier?

Setting 2: no group labels at all

Examples:

- identify groups from training AND validation data with ensemble of biased classifiers to reduce noise

- post-hoc logit adjustment using $P\left(Y|\hat{Y}_{biased}\right)$ as an estimate of $P\left(Y|A\right)$

Boosting worst-group accuracy without any group annotations

Vincent Bardenhagen*, Alexandru Ţifrea*, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland
{vbardenha,tifreaa,fan.yang}@ethz.ch

Group Robust Classification Without Any Group Information

Christos Tsirigotis* Université de Montréal, Mila, ServiceNow Research
Joao Monteiro ServiceNow Research
Pau Rodriguez Apple MLR
David Vazquez ServiceNow Research
Aaron Courville† Université de Montréal, Mila, CIFAR CAI Chair

How to train a biased classifier?

Setting 2: no group labels at all

Examples:

- identify groups from training AND validation data with ensemble of biased classifiers to reduce noise
- post-hoc logit adjustment using $P\left(Y|\hat{Y}_{biased}\right)$ as an estimate of $P\left(Y|A\right)$

Boosting worst-group accuracy without any group annotations

Vincent Bardenhagen[*], Alexandru Ţifrea[*], Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland
{vbardenha,tifreaa,fan.yang}@ethz.ch

Group Robust Classification Without Any Group Information

Christos Tsirigotis[*]
Université de Montréal, Mila, ServiceNow Research

Joao Monteiro
ServiceNow Research

Pau Rodriguez
Apple MLR

David Vazquez
ServiceNow Research

Aaron Courville[†]
Université de Montréal, Mila, CIFAR CAI Chair

	Tuning	Corrupt-MNIST Avg	Corrupt-MNIST Wg	Waterbirds Avg	Waterbirds Wg	CelebA Avg	CelebA Wg	Color MNIST Avg	Color MNIST Wg	Adult Avg	Adult Wg	Poverty Avg	Poverty Wg
No group labels	ERM	99.6	71.2	97.9	74.9	94.3	60.7	99.8	82.6	80.1	41.6	87.6	**55.6**
	Ours	99.0	**96.5**	97.5	**78.5**	88.0	**78.9**	99.3	**96.6**	81.2	**68.0**	86.3	50.0
Val group labels	ERM WG	99.5	79.8	97.6	**86.7**	93.1	77.8	99.7	84.4	78.9	61.2	87.7	51.5
	JTT	99.1	**91.3**	93.3	**86.7**	88.0	**81.1**	98.3	**94.8**	77.8	**63.3**	64.5	**60.5**

Similar average and worst-group accuracy for <u>two-stage methods</u>:

- with no group labels
- with validation group labels

DRO mitigations in the presence of outliers

Recall: Two-stage methods

1) use biased classifier to identify ***error set***
2) train fair classifier via IW / GDRO

DRO mitigations in the presence of outliers

Recall: Two-stage methods

1) use biased classifier to identify ***error set***
2) train fair classifier via IW / GDRO

The opposite of what robust statistics literature recommends!

e.g. can amplify outliers, noisy samples etc

Can we get both fairness and robustness to outliers?

DRO mitigations in the presence of outliers

Recall: Two-stage methods
1) use biased classifier to identify ***error set***
2) train fair classifier via IW / GDRO

The opposite of what robust statistics literature recommends!
e.g. can amplify outliers, noisy samples etc

Can we get both fairness and robustness to outliers?

Robust Mixture Learning when Outliers Overwhelm Small Groups

Daniil Dmitriev[1*], Rares-Darius Buhai[1*], Stefan Tiegel[1], Alexander Wolters[2], Gleb Novikov[3], Amartya Sanyal[4], David Steurer[1], and Fanny Yang[1]

Clustering algorithm that is
- applicable even for |Outliers| ≫ |Minority group|
- computationally efficient
- information-theoretically optimal

DRO mitigations in the presence of outliers

Recall: Two-stage methods
1) use biased classifier to identify ***error set***
2) train fair classifier via IW / GDRO

The opposite of what robust statistics literature recommends!
e.g. can amplify outliers, noisy samples etc

Can we get both fairness and robustness to outliers?

Robust Mixture Learning when Outliers Overwhelm Small Groups

Daniil Dmitriev[1*], Rares-Darius Buhai[1*], Stefan Tiegel[1], Alexander Wolters[2], Gleb Novikov[3], Amartya Sanyal[4], David Steurer[1], and Fanny Yang[1]

Fairness without Demographics through Adversarially Reweighted Learning

Preethi Lahoti *
plahoti@mpi-inf.mpg.de
Max Planck Institute for Informatics

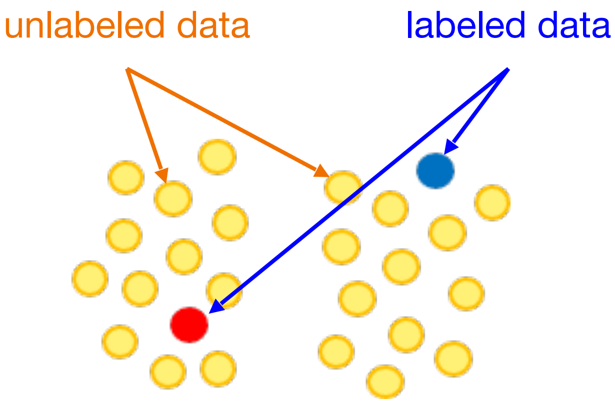
Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, Ed H. Chi
Google Research

Clustering algorithm that is
- applicable even for |Outliers| ≫ |Minority group|
- computationally efficient
- information-theoretically optimal

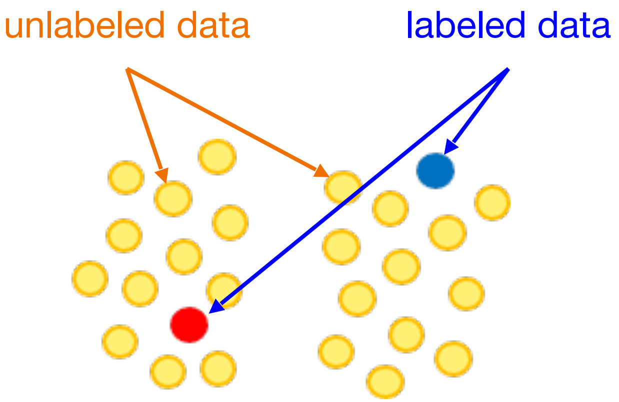
Idea: only upweight samples in the error set that are computationally identifiable using simple function class $\mathcal{F}$.

Fairness in the low-label regime

Low-label regime



Low-label regime

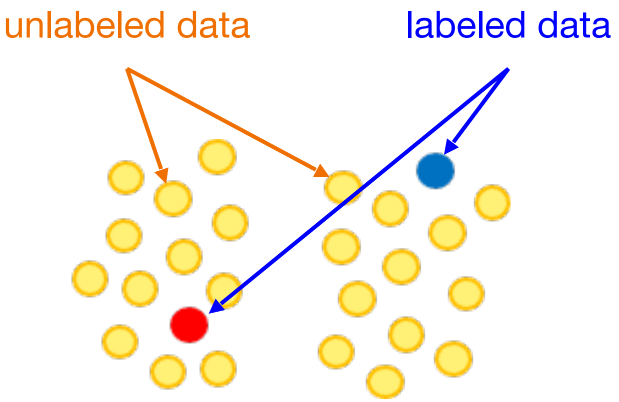


Research questions

1) How to acquire the labeled data?

2) How to learn from both labeled and unlabeled data?

Low-label regime

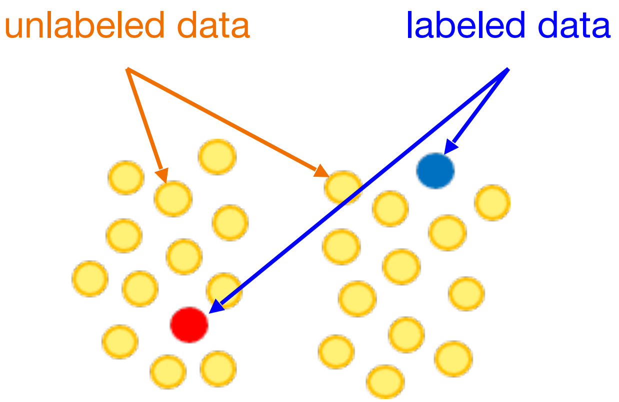


Research questions

1) How to acquire the labeled data? **active learning**

2) How to learn from both labeled and unlabeled data?

Low-label regime

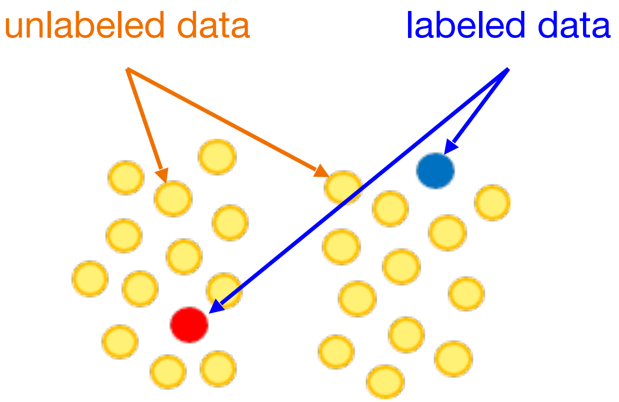


Research questions

1) How to acquire the labeled data? **active learning**

2) How to learn from both labeled and unlabeled data? **semi-supervised learning**

Low-label regime



Research questions

1) How to acquire the labeled data? **active learning**

2) How to learn from both labeled and unlabeled data? **semi-supervised learning**

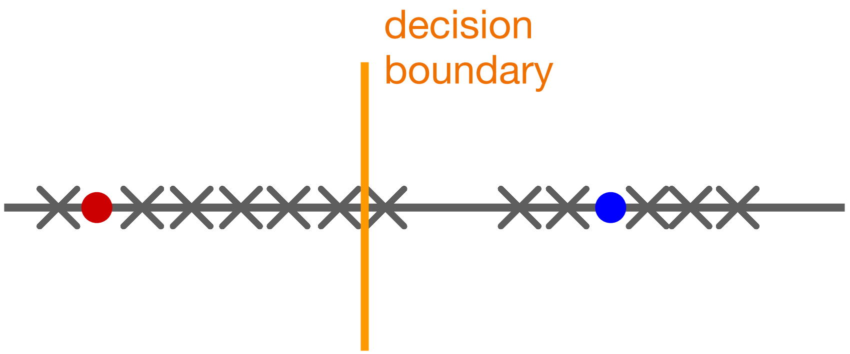
Fairness problems

- class imbalance
- group imbalance
  (but potentially balanced classes)

Uncertainty sampling-based active learning

Uncertainty sampling

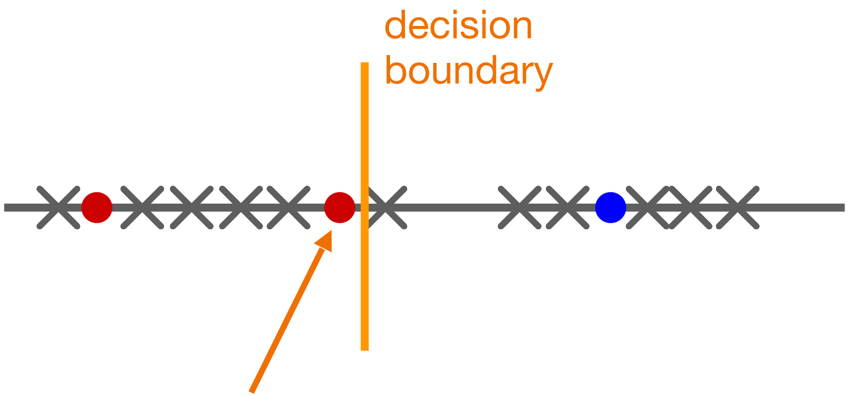
“binary search to find decision boundary”



Uncertainty sampling-based active learning

Uncertainty sampling

"binary search to find decision boundary"

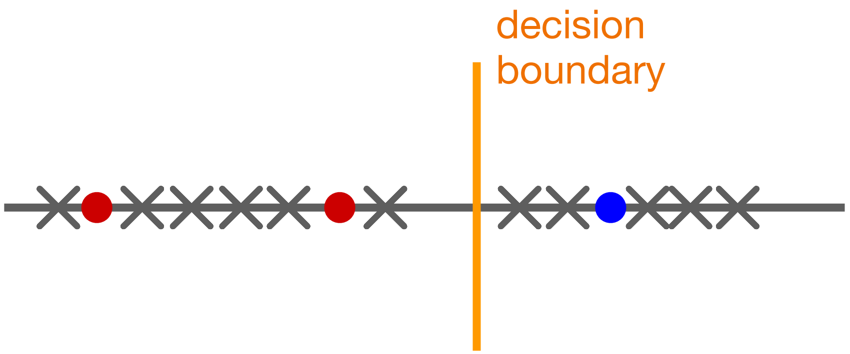


closest point to decision boundary

Uncertainty sampling-based active learning

Uncertainty sampling

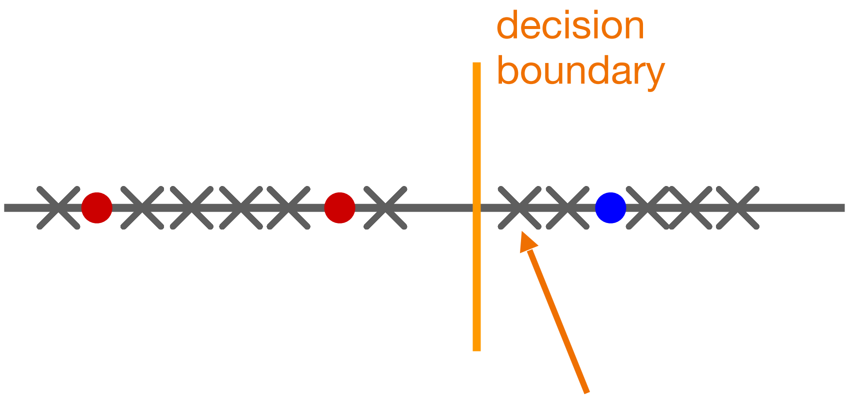
“binary search to find decision boundary”



Uncertainty sampling-based active learning

Uncertainty sampling

“binary search to find decision boundary”

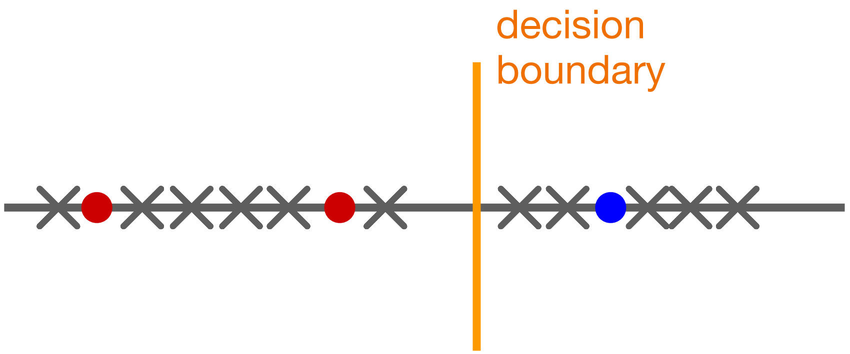


closest point to decision boundary

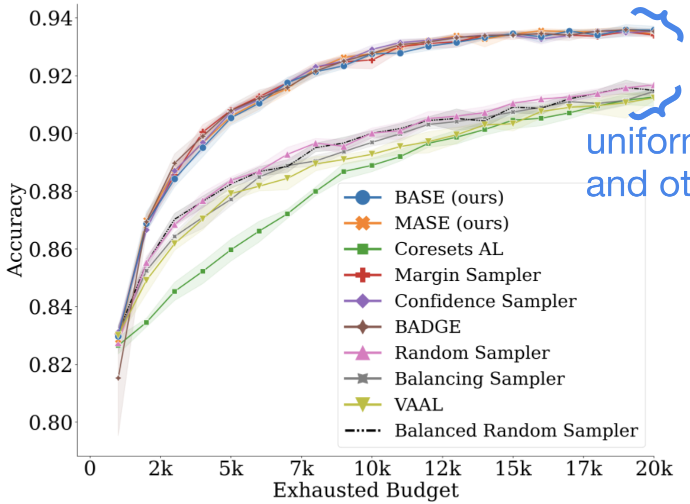
Uncertainty sampling-based active learning

Uncertainty sampling

“binary search to find decision boundary”



uncertainty-based AL



uniform sampling and other AL

CIFAR10 dataset

U-AL is more label efficient than uniform sampling or other AL

Image source: https://arxiv.org/abs/2111.12880

Standard active learning can improve fairness

Class-imbalanced classification

Learning on the Border:
Active Learning in Imbalanced Data Classification

Şeyda Ertekin[1], Jian Huang[2], Léon Bottou[3], C. Lee Giles[2,1]

Focus on linear classification

Standard active learning can improve fairness

Class-imbalanced classification

Learning on the Border:
Active Learning in Imbalanced Data Classification

Şeyda Ertekin[1], Jian Huang[2], Léon Bottou[3], C. Lee Giles[2,1]

true decision boundary

best **avg-case** classifier

Focus on linear classification

Standard active learning can improve fairness

Class-imbalanced classification

true decision boundary

best **avg-case** classifier

Focus on linear classification

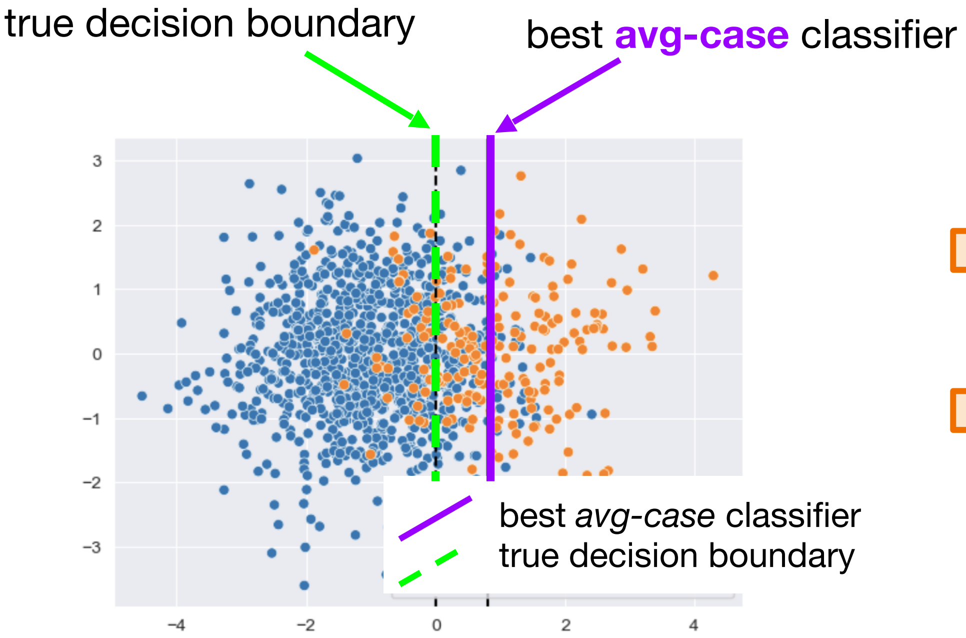
Learning on the Border:
Active Learning in Imbalanced Data Classification

Şeyda Ertekin[1], Jian Huang[2], Léon Bottou[3], C. Lee Giles[2,1]

Decision boundary of biased classifier is closer to minority class

Standard active learning can improve fairness

Class-imbalanced classification



Focus on linear classification

Learning on the Border:
Active Learning in Imbalanced Data Classification

Şeyda Ertekin[1], Jian Huang[2], Léon Bottou[3], C. Lee Giles[2,1]

Decision boundary of biased classifier is closer to minority class

U-AL tends to select more minority points to be labeled

Standard active learning can improve fairness

Class-imbalanced classification

true decision boundary

best **avg-case** classifier

Focus on linear classification

Learning on the Border:
Active Learning in Imbalanced Data Classification

Şeyda Ertekin[1], Jian Huang[2], Léon Bottou[3], C. Lee Giles[2,1]

Decision boundary of biased classifier is closer to minority class

U-AL tends to select more minority points to be labeled

U-AL collects a more balanced labeled set

Standard active learning can improve fairness

Class-imbalanced classification

U-AL also mitigates class imbalance in <u>non-linear</u> classification!

Active Learning at the ImageNet Scale

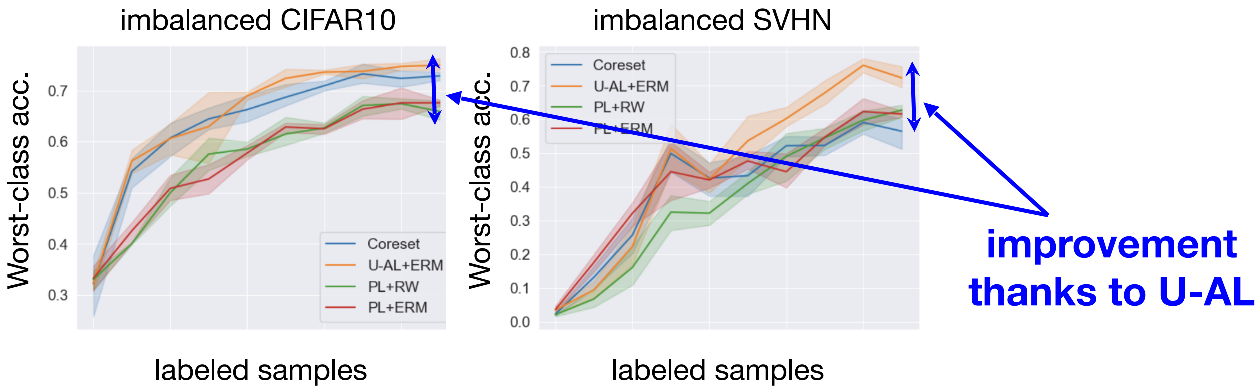
Zeyad Ali Sami Emam[*†‡] zeyad@umd.edu
Hong-Min Chu[*†] hmchu@umd.edu
Ping-Yeh Chiang[*†] pchiang@umd.edu
Wojciech Czaja[†] wojtek@umd.edu
Richard Leapman[‡] leapmanr@mail.nih.gov
Micah Goldblum[§] goldblum@nyu.edu
Tom Goldstein[†] tomg@umd.edu

Improving class and group imbalanced classification with uncertainty-based active learning

Alexandru Tifrea* TIFREAA@INF.ETHZ.CH
Department of Computer Science, ETH Zurich
John Hill* JHILL326@GATECH.EDU
Department of Computer Science, Georgia Institute of Technology
Fanny Yang FAN.YANG@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

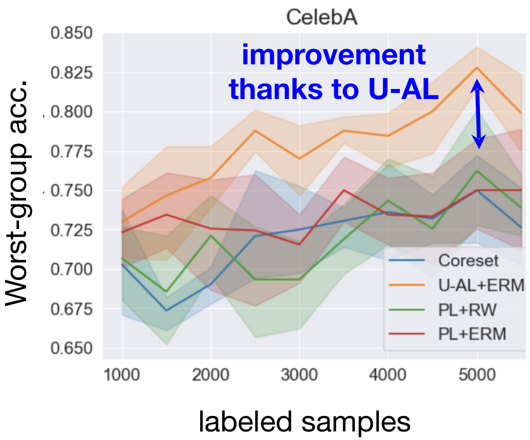
Algorithm Selection for Deep Active Learning with Imbalanced Datasets

Jifan Zhang, University of Wisconsin - Madison, Madison, WI 53715, jifan@cs.wisc.edu
Shuai Shao, Meta Inc., Menlo Park, CA 94025, sshao@meta.com
Saurabh Verma, Meta Inc., Menlo Park, CA 94025, saurabh08@meta.com
Robert Nowak, University of Wisconsin - Madison, Madison, WI 53715, rdnowak@wisc.edu



Standard active learning can improve fairness

Group-imbalanced classification



Improving class and group imbalanced classification with uncertainty-based active learning

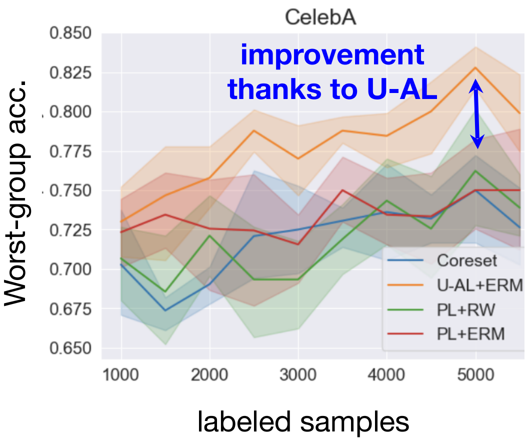
Alexandru Tifrea* TIFREAA@INF.ETHZ.CH
Department of Computer Science, ETH Zurich
John Hill* JHILL326@GATECH.EDU
Department of Computer Science, Georgia Institute of Technology
Fanny Yang FAN.YANG@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

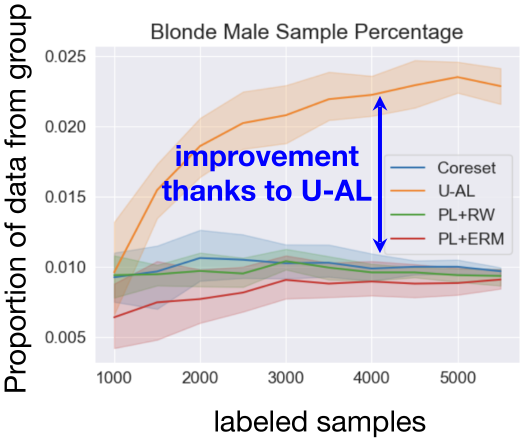
CAN ACTIVE LEARNING PREEMPTIVELY MITIGATE FAIRNESS ISSUES?

Frédéric Branchaud-Charron,* Parmida Atighehchian,* Pau Rodríguez, Grace Abuhamad, Alexandre Lacoste
ServiceNow
{fr.branchaud-charron,parmida.atighehchian}@servicenow.com

Standard active learning can improve fairness

Group-imbalanced classification





Improving class and group imbalanced classification with uncertainty-based active learning

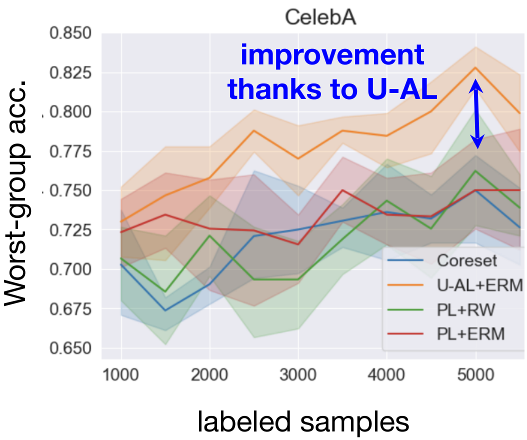
Alexandru Tifrea* TIFREAA@INF.ETHZ.CH
Department of Computer Science, ETH Zurich
John Hill* JHILL326@GATECH.EDU
Department of Computer Science, Georgia Institute of Technology
Fanny Yang FAN.YANG@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

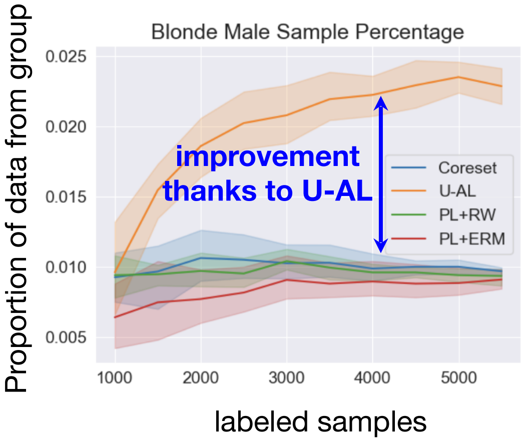
CAN ACTIVE LEARNING PREEMPTIVELY MITIGATE FAIRNESS ISSUES?

Frédéric Branchaud-Charron,* Parmida Atighehchian,* Pau Rodríguez, Grace Abuhamad, Alexandre Lacoste
ServiceNow
{fr.branchaud-charron,parmida.atighehchian}@servicenow.com

Standard active learning can improve fairness

Group-imbalanced classification





Improving class and group imbalanced classification with uncertainty-based active learning

Alexandru Tifrea* TIFREAA@INF.ETHZ.CH
Department of Computer Science, ETH Zurich
John Hill* JHILL326@GATECH.EDU
Department of Computer Science, Georgia Institute of Technology
Fanny Yang FAN.YANG@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

CAN ACTIVE LEARNING PREEMPTIVELY MITIGATE FAIRNESS ISSUES?

Frédéric Branchaud-Charron,* Parmida Atighehchian,* Pau Rodríguez, Grace Abuhamad, Alexandre Lacoste
ServiceNow
{fr.branchaud-charron,parmida.atighehchian}@servicenow.com

Takeaways

- no explicit group information used anywhere during sampling/learning!
- not *all* AL strategies help (e.g. coreset sampling)
- U-AL+ERM can be better than passive learning + reweighting

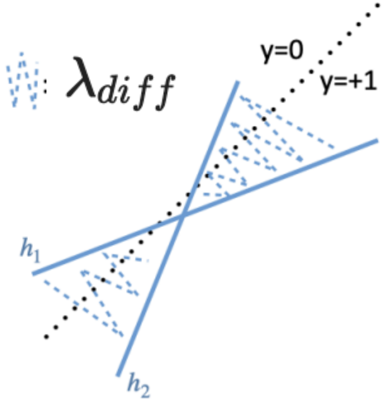
Using group labels for active learning

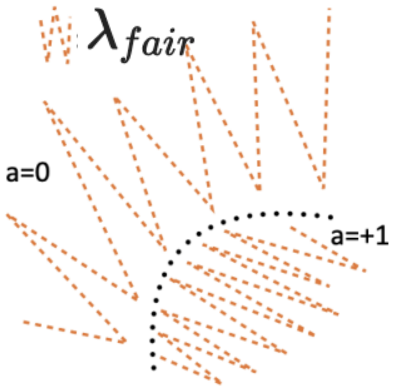
Fair Active Learning in Low-Data Regimes

Romain Camilleri, Andrew Wagenmaker, Jamie Morgenstern, Lalit Jain, Kevin Jamieson
University of Washington, Seattle, WA
{camilr,ajwagen,jamiemmt,jamieson}@cs.washington.edu,lalitj@uw.edu

Acquire labels for samples selected according to:

$$P_{AL}(X) \sim \boxed{\frac{1}{2}\lambda_{diff}(X)} + \boxed{\frac{1}{2}\lambda_{fair}(X)}$$

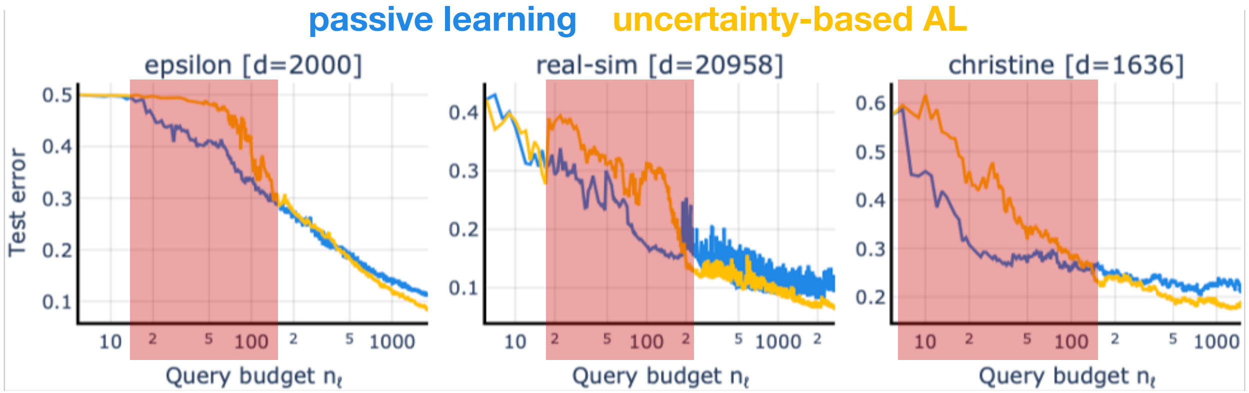




Informativeness criterion:
Disagreement region of ensemble

Fairness criterion:
Uniform mass on all groups

Limitations of uncertainty-based AL



Err[U-AL] > Err[PL]

U-AL can be on par with or even worse than passive learning

- For high-dimensional data
- For data with lots of label noise

Margin-based sampling in high dimensions: When being active is less efficient than staying passive

Alexandru Țifrea [*1] Jacob Clarysse [*1] Fanny Yang [1]

On the Relationship between Data Efficiency and Error for Uncertainty Sampling

Stephen Mussmann [1] Percy Liang [1]

Summary

A few examples of fair learning algorithms that

(1) Have fewer data requirements than standard fairness mitigations
(2) Leverage unlabeled data to improve fairness

Open questions

- Impact of class/group label noise
- Interplay between fairness and other evaluation metrics, beyond accuracy

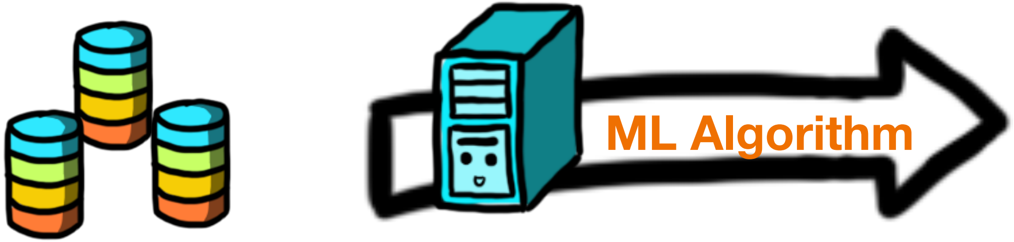
⇨ **coming up in the next part**

Privacy in Machine Learning

Privacy in Machine Learning

Privacy in Machine Learning

Privacy in Machine Learning



Privacy in Machine Learning

Privacy in Machine Learning

Privacy can mean a lot of things but two things are important to define:

Privacy in Machine Learning

Privacy can mean a lot of things but two things are important to define:

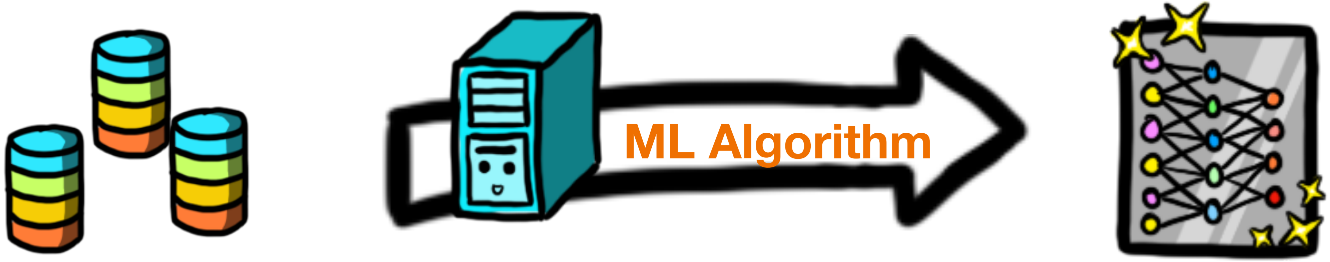
- What is the private entity ?

Privacy in Machine Learning

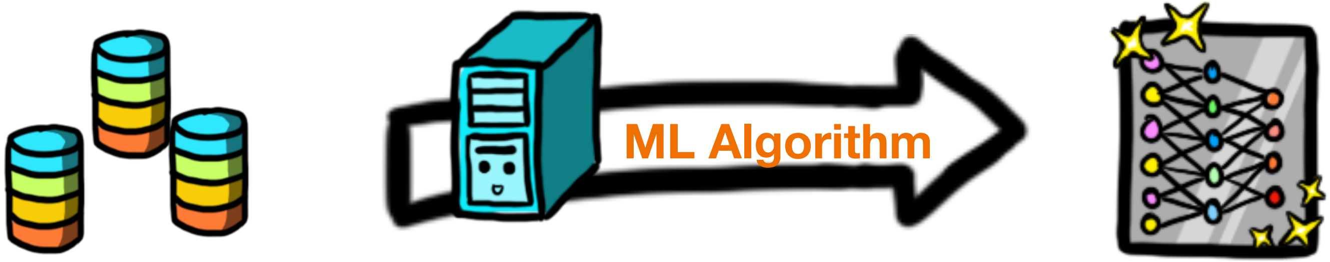
Privacy can mean a lot of things but two things are important to define:

- What is the private entity ?
- What can the privacy adversary observe ?

PETs in Machine Learning

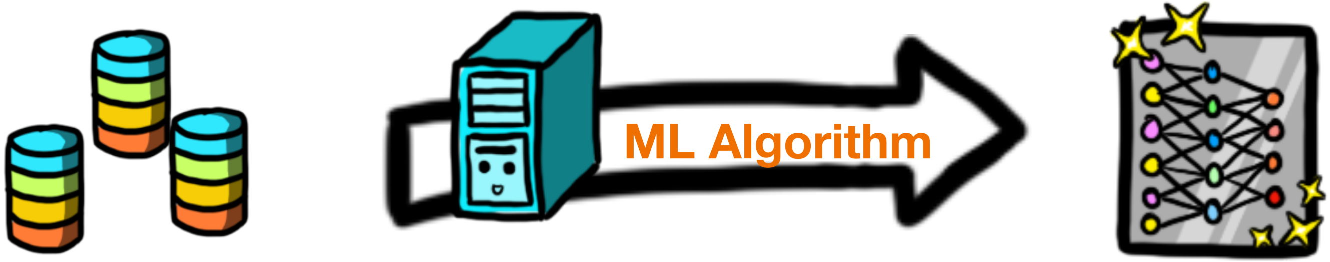


PETs in Machine Learning



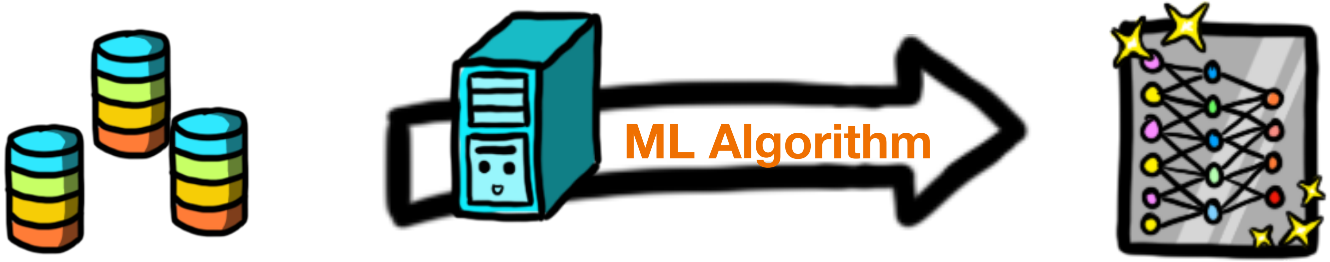
- **Differential Privacy** prevents leakage of *training data from the trained model*

PETs in Machine Learning



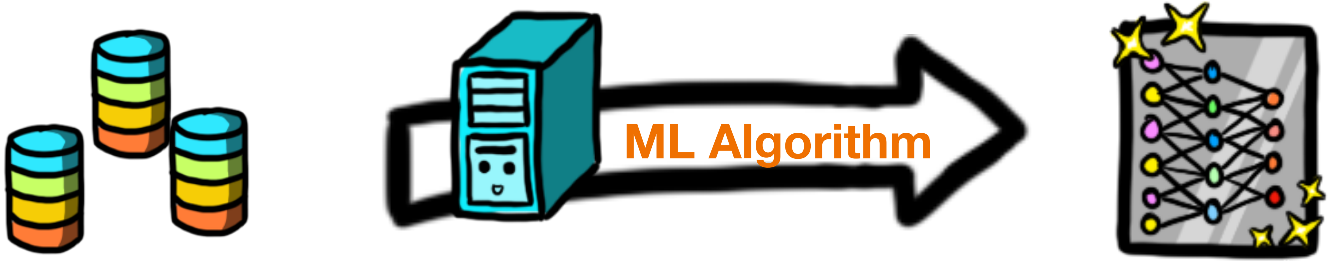
- **Differential Privacy** prevents leakage of *training data from the trained model*
- **Multi-Party Computation** allows multiple data holders to collaboratively execute a computation without learn too much about other parties' data

PETs in Machine Learning



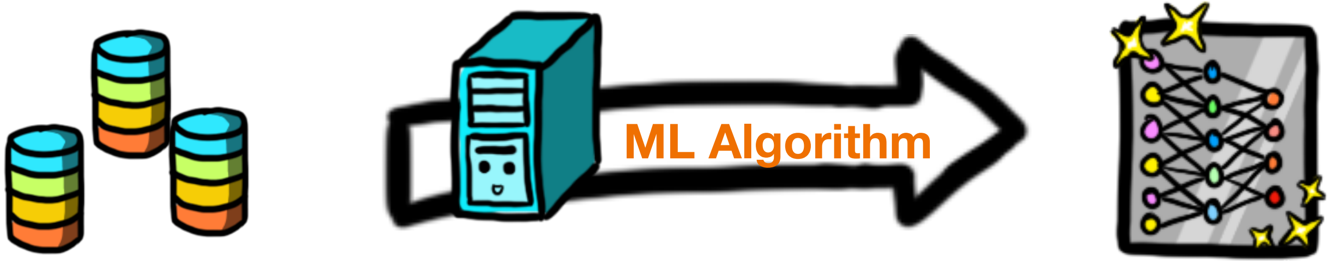
- **Differential Privacy** prevents leakage of *training data from the trained model*
- **Multi-Party Computation** allows multiple data holders to collaboratively execute a computation without learn too much about other parties' data
- **Fully Homomorphic encryption** based methods allow training or testing on encrypted data without decrypting the data.

PETs in Machine Learning



- **Differential Privacy** prevents leakage of *training data from the trained model*
- **Multi-Party Computation** allows multiple data holders to collaboratively execute a computation without learn too much about other parties' data
- **Fully Homomorphic encryption** based methods allow training or testing on encrypted data without decrypting the data.
- Many more PETs like Trusted Execution Environments (TEE), Contextual Integrity etc.

PETs in Machine Learning



- **Differential Privacy** prevents leakage of *training data from the trained model*
- **Multi-Party Computation** allows multiple data holders to collaboratively execute a computation without learn too much about other parties' data
- **Fully Homomorphic encryption** based methods allow training or testing on encrypted data without decrypting the data.
- Many more PETs like Trusted Execution Environments (TEE), Contextual Integrity etc.

“Data is a precious thing and will last longer than the systems themselves”

Sir Tim-Berners Lee

US Census and Privacy

Vulnerability of sparse data

WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago

US Census and Privacy

Vulnerability of sparse data

WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago

- 2010 US Census privacy protections were vulnerable to reconstruction attacks.

US Census and Privacy

Vulnerability of sparse data

WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago

- 2010 US Census privacy protections were vulnerable to reconstruction attacks.
- Some groups are disproportionately affected: **80% of NHPI (Native Hawaiian & Pacific Islander)** responses in NC were fully reconstructed.

US Census and Privacy

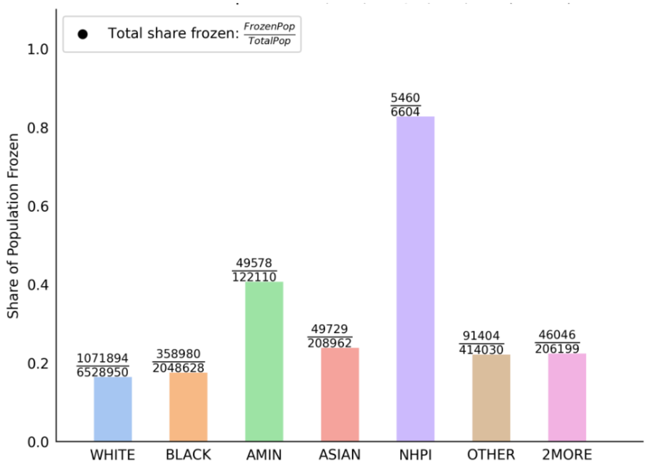
Vulnerability of sparse data

WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago

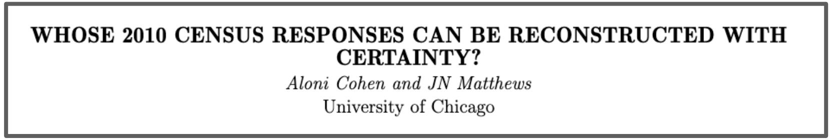
- 2010 US Census privacy protections were vulnerable to reconstruction attacks.
- Some groups are disproportionately affected: **80% of NHPI (Native Hawaiian & Pacific Islander)** responses in NC were fully reconstructed.



US Census and Privacy

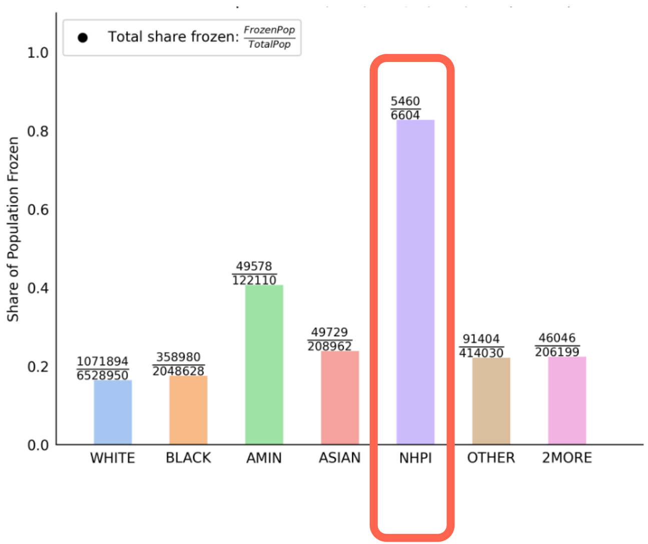
Vulnerability of sparse data

- 2010 US Census privacy protections were vulnerable to reconstruction attacks.
- Some groups are disproportionately affected: **80% of NHPI (Native Hawaiian & Pacific Islander)** responses in NC were fully reconstructed.


WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago



US Census and Privacy

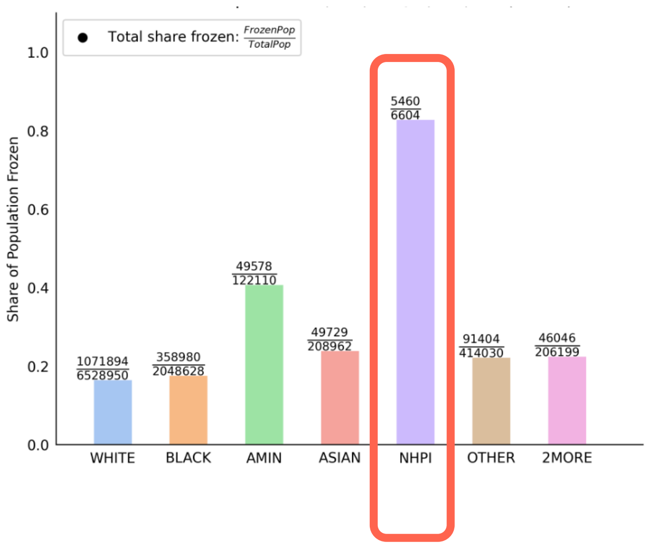
Vulnerability of sparse data

WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago

- 2010 US Census privacy protections were vulnerable to reconstruction attacks.
- Some groups are disproportionately affected: **80% of NHPI (Native Hawaiian & Pacific Islander)** responses in NC were fully reconstructed.
- Similar inferences were also shown about age and more accurate in smaller “blocks”



US Census and Privacy

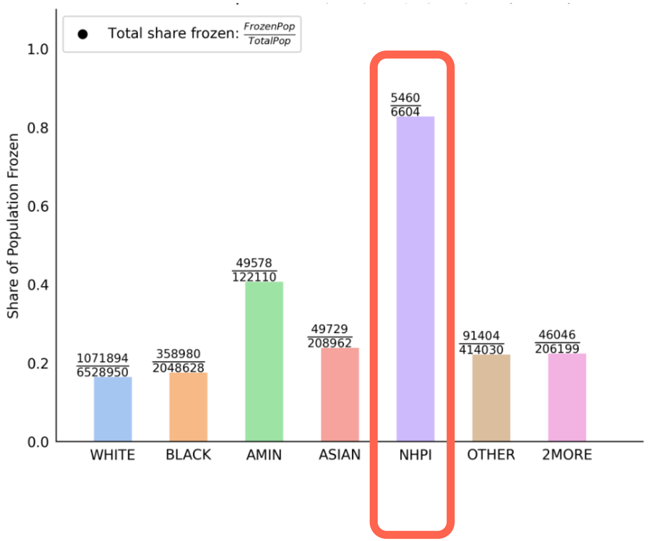
Vulnerability of sparse data

WHOSE 2010 CENSUS RESPONSES CAN BE RECONSTRUCTED WITH CERTAINTY?

Aloni Cohen and JN Matthews

University of Chicago

- 2010 US Census privacy protections were vulnerable to reconstruction attacks.
- Some groups are disproportionately affected: **80% of NHPI (Native Hawaiian & Pacific Islander)** responses in NC were fully reconstructed.
- Similar inferences were also shown about age and more accurate in smaller "blocks"



Takeaway: Often privacy violations are stronger in smaller communities.

Cost of Privacy

Cost of Privacy

Informal Theorem: If you try to answer too many questions too accurately about a dataset, there's a clever way for an attacker to piece together (almost) the entire original data.

Cost of Privacy

Informal Theorem: If you try to answer too many questions too accurately about a dataset, there's a clever way for an attacker to piece together (almost) the entire original data.

If the original dataset's privacy is to be protected, some accuracy needs to be sacrificed. The study of DP tries to control this trade-off.

Making an Algorithm Differentially Private

Making an Algorithm Differentially Private

Differential Privacy

Making an Algorithm Differentially Private

Differential Privacy

- **Differential Privacy** noises the algorithm's output to limit the exposure of any single data point

Making an Algorithm Differentially Private

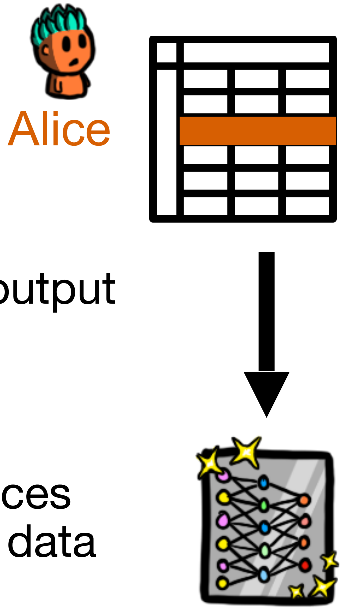
Differential Privacy

- **Differential Privacy** noises the algorithm's output to limit the exposure of any single data point

- A **Differentially Private** ML algorithm produces similar models irrespective of whether Alice's data is in the dataset or Bob's

Making an Algorithm Differentially Private

Differential Privacy



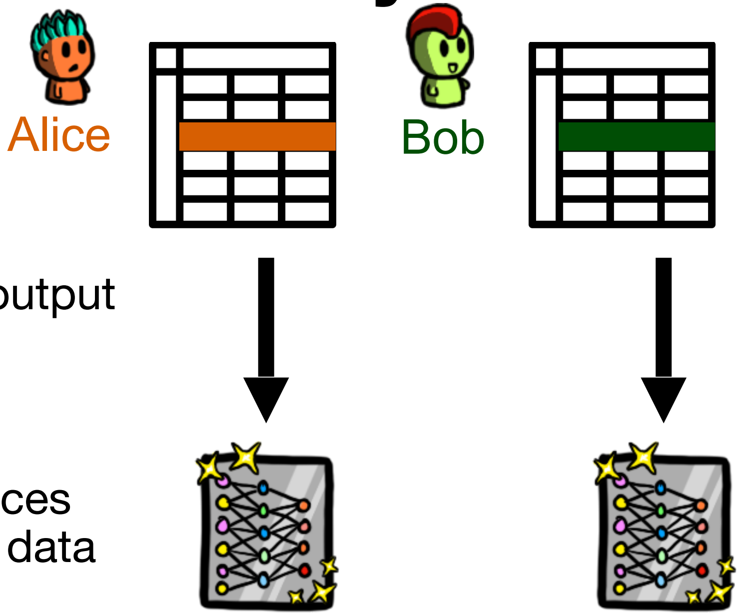
- **Differential Privacy** noises the algorithm's output to limit the exposure of any single data point
- A **Differentially Private** ML algorithm produces similar models irrespective of whether Alice's data is in the dataset or Bob's

Making an Algorithm Differentially Private

Differential Privacy

- **Differential Privacy** noises the algorithm's output to limit the exposure of any single data point

- A **Differentially Private** ML algorithm produces similar models irrespective of whether Alice's data is in the dataset or Bob's

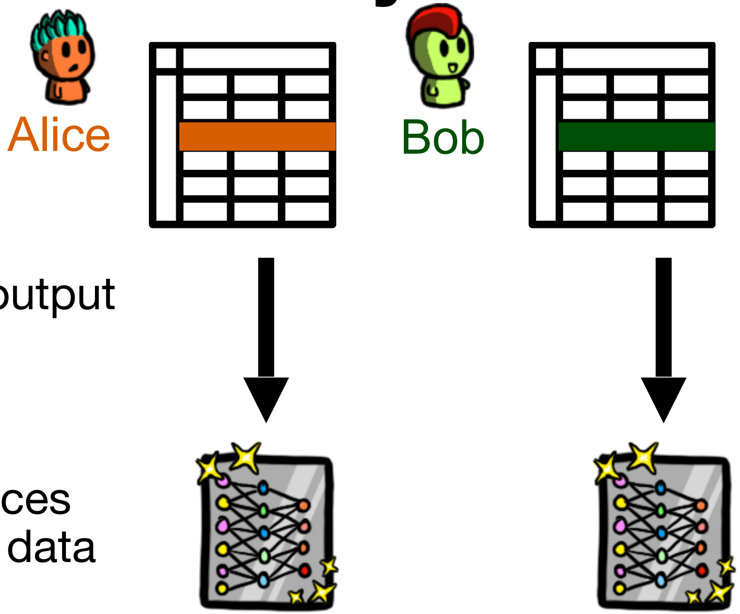


Making an Algorithm Differentially Private

Differential Privacy

- **Differential Privacy** noises the algorithm's output to limit the exposure of any single data point

- A **Differentially Private** ML algorithm produces similar models irrespective of whether Alice's data is in the dataset or Bob's



The replacement of a single data record minimally impacts the trained model

How to make Machine Learning Private

How to make Machine Learning Private

Differential Privacy (Defn.)

How to make Machine Learning Private

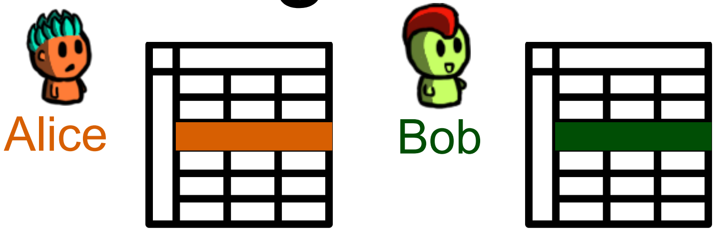
Differential Privacy (Defn.)

Consider any

- Neighbouring datasets S_1 and S_2
- Output set Q

Then Algorithm is (ε, δ)-DP if

How to make Machine Learning Private

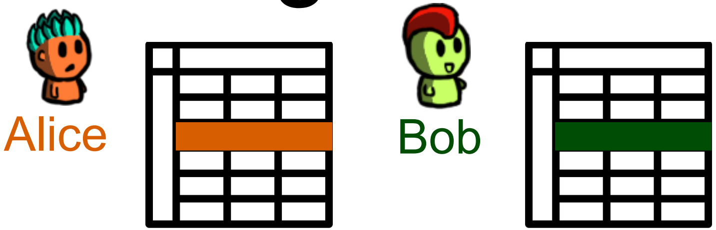


Differential Privacy (Defn.)

Consider any
- Neighbouring datasets S_1 and S_2
- Output set Q

Then Algorithm is (ε, δ)-DP if

How to make Machine Learning Private

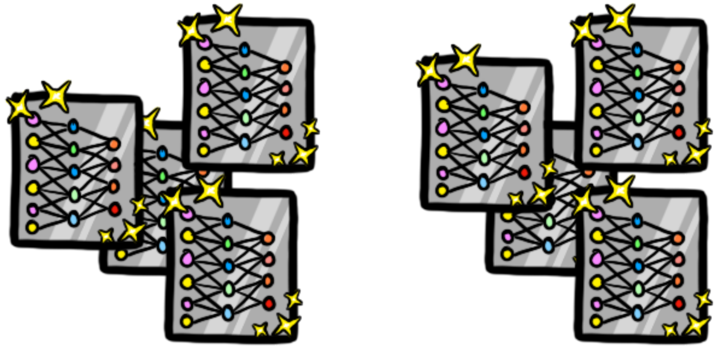


Differential Privacy (Defn.)

Consider any

- Neighbouring datasets S_1 and S_2
- Output set Q

Then Algorithm is (ε, δ)-DP if

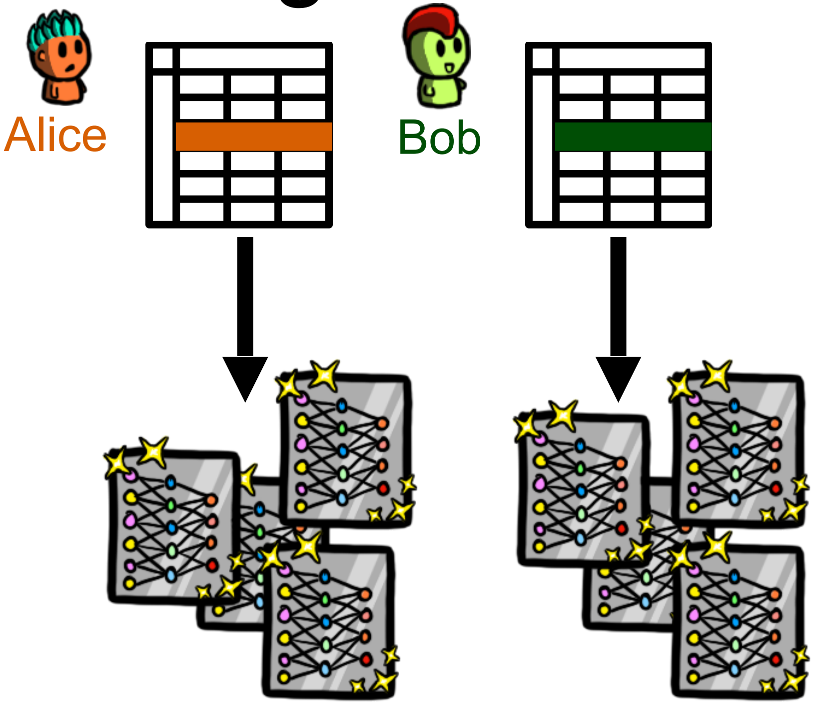


How to make Machine Learning Private

Differential Privacy (Defn.)

Consider any
- Neighbouring datasets S_1 and S_2
- Output set Q

Then Algorithm is (ε, δ)-DP if



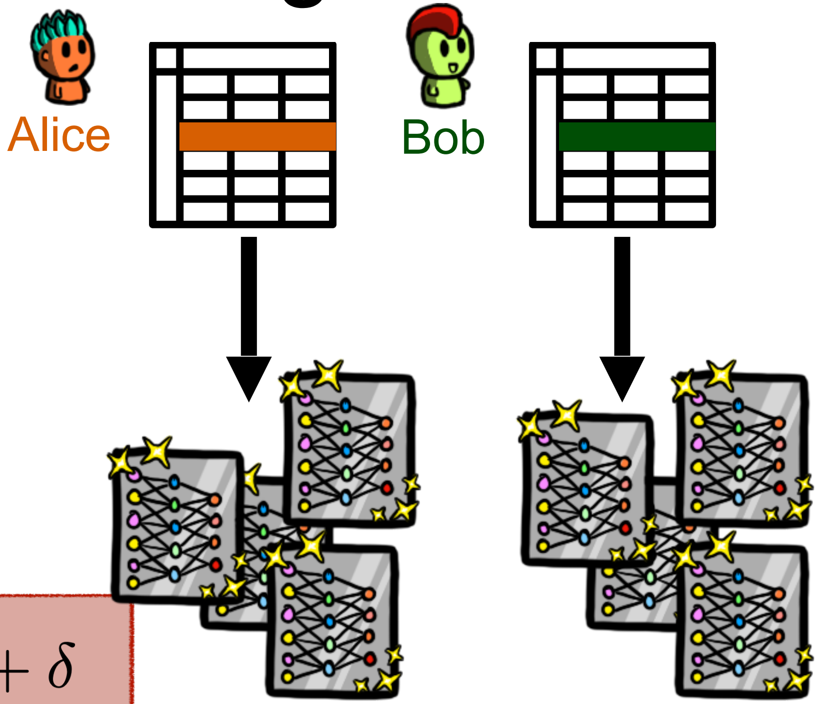
How to make Machine Learning Private

Differential Privacy (Defn.)

Consider any
- Neighbouring datasets S_1 and S_2
- Output set Q

Then Algorithm is (ε, δ)-DP if

$$\mathbb{P}\left(\mathcal{A}(\mathbf{S_1}) \in \mathcal{Q}\right) \leq \mathbf{e}^{\epsilon}\mathbb{P}\left(\mathcal{A}(\mathbf{S_2}) \in \mathcal{Q}\right) + \delta$$



Differential Privacy and data quality

Differential Privacy and data quality

One simple formula for implementing DP :

Differential Privacy and data quality

One simple formula for implementing DP :

- **Compute the algorithm’s sensitivity**: how much can the output change if the *worst* data point in the *worst* input dataset changes

Differential Privacy and data quality

One simple formula for implementing DP :

- **Compute the algorithm's sensitivity**: how much can the output change if the *worst* data point in the *worst* input dataset changes
- **Add noise** proportional to that change magnitude

Differential Privacy and data quality

One simple formula for implementing DP :

- **Compute the algorithm's sensitivity**: how much can the output change if the *worst* data point in the *worst* input dataset changes
- **Add noise** proportional to that change magnitude

Today, we will look at two ways in which data quality affects the performance of Differentially Private Algorithms

Differential Privacy and data quality

One simple formula for implementing DP :

- **Compute the algorithm's sensitivity**: how much can the output change if the *worst* data point in the *worst* input dataset changes
- **Add noise** proportional to that change magnitude

Today, we will look at two ways in which data quality affects the performance of Differentially Private Algorithms

- *Good data requires less added noise* for the same level of privacy

Differential Privacy and data quality

One simple formula for implementing DP :

- **Compute the algorithm's sensitivity**: how much can the output change if the *worst* data point in the *worst* input dataset changes
- **Add noise** proportional to that change magnitude

Today, we will look at two ways in which data quality affects the performance of Differentially Private Algorithms

- *Good data requires less added noise* for the same level of privacy
- Some parts of data domain *incurs disproportionately higher loss* due to the Differential privacy than others

Differential Privacy and Disparate Impact

DP and Disparate Impact

Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

DP and Disparate Impact

Examples in Practice

- Wikimedia foundation released their pageview statistics Differentially Privately.

Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2],
Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2],
Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

DP and Disparate Impact

Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

DP and Disparate Impact

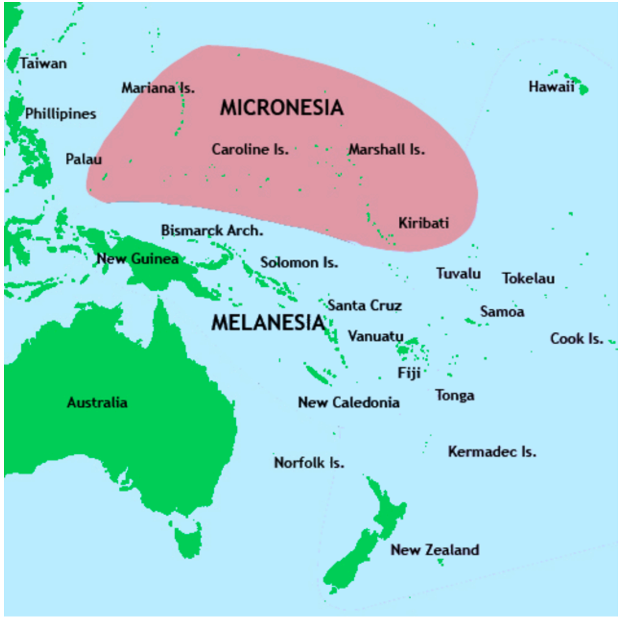
Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%



DP and Disparate Impact
Examples in Practice

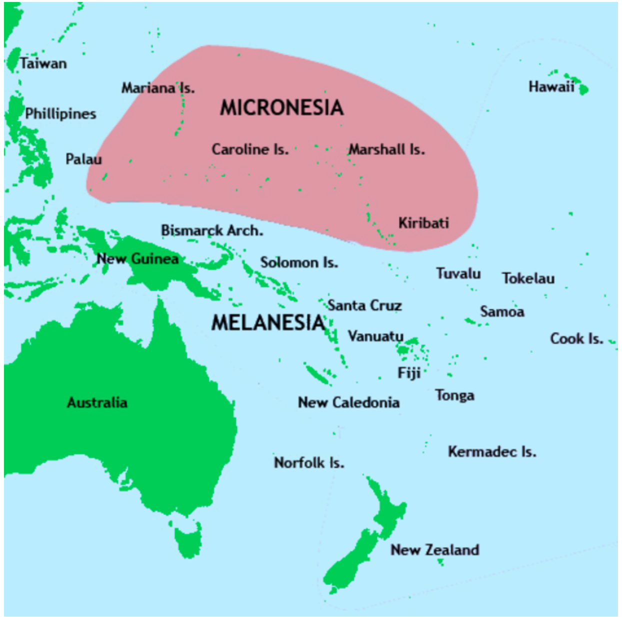
Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

But, the "Micronesia problem"



DP and Disparate Impact
Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

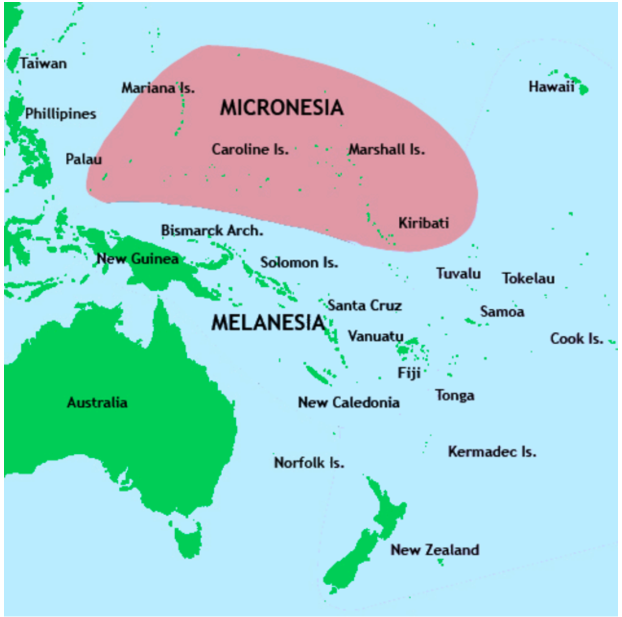
Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

But, the “Micronesia problem”

- Seven Pacific Island nations (low traffic)



DP and Disparate Impact
Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

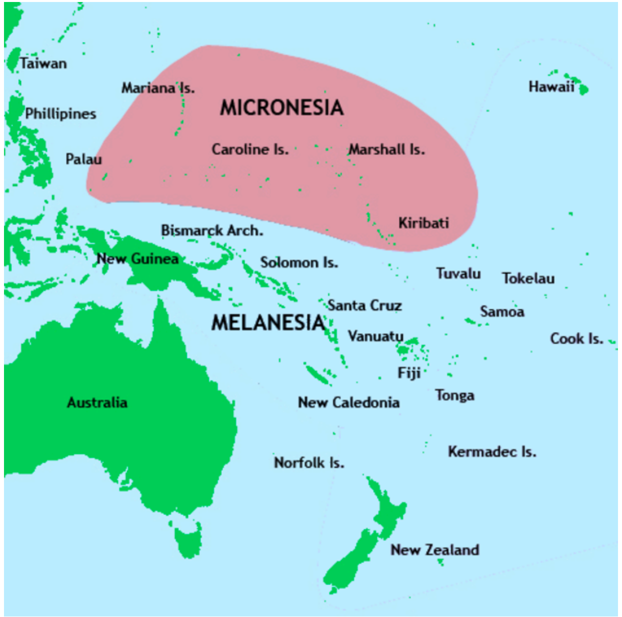
Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

But, the "Micronesia problem"

- Seven Pacific Island nations (low traffic)
- Naive first implementation



DP and Disparate Impact
Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

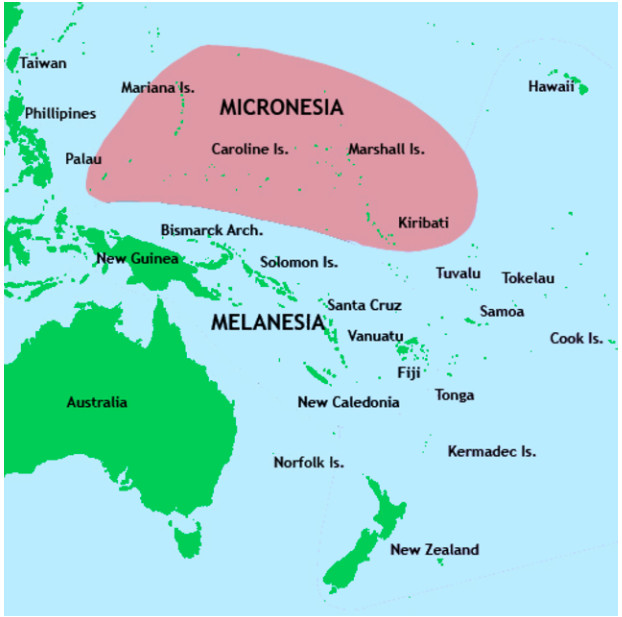
Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

But, the “Micronesia problem”

- Seven Pacific Island nations (low traffic)
- Naive first implementation
    - >99% of published data is spurious



Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

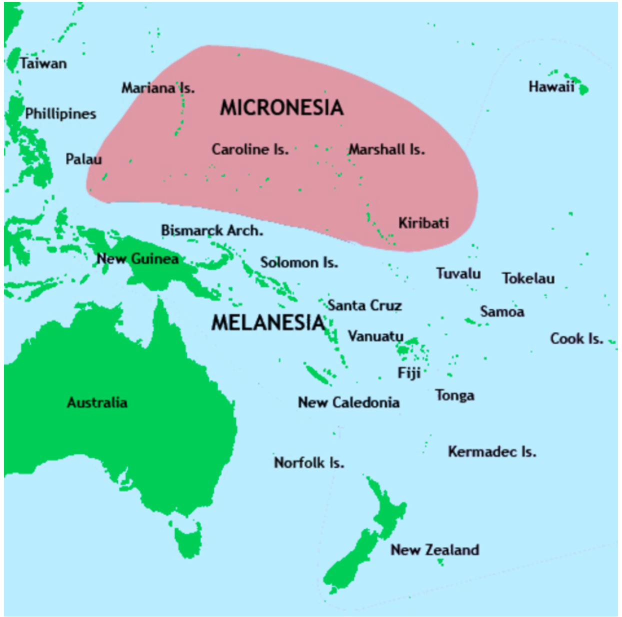
DP and Disparate Impact

Examples in Practice

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

But, the "Micronesia problem"

- Seven Pacific Island nations (low traffic)
- Naive first implementation
  - >99% of published data is spurious
  - **9 out of 23 subcontinental regions have spurious rate of >25%**



DP and Disparate Impact

Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

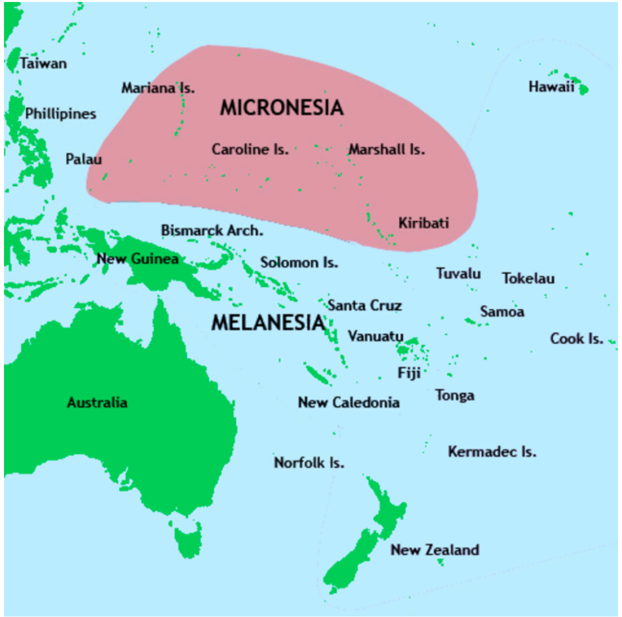
Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
- In their global release, both drop rate and spurious rate were less than 1%

But, the “Micronesia problem”

- Seven Pacific Island nations (low traffic)
- Naive first implementation
    - >99% of published data is spurious
    - **9 out of 23 subcontinental regions have spurious rate of >25%**
    - Africa, Oceania, Central Asia, and the Caribbean



DP and Disparate Impact

Examples in Practice

Publishing Wikipedia usage data with strong privacy guarantees

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org
[2]Tumult Labs – science@tmlt.io

- Wikimedia foundation released their pageview statistics Differentially Privately.
-

The New York Times

The 2020 Census Suggests That People Live Underwater. There's a Reason.

But, the

- Se
- Na
  - ○
  - **9 out of 23 subcontinental regions have spurious rate of >25%**
  - Africa, Oceania, Central Asia, and the Caribbean

DP and Disparate Impact
Controlled experimental setting

How unfair is private learning?

Amartya Sanyal[*1,3] **Yaxi Hu**[*2] **Fanny Yang**[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

DP and Disparate Impact

Controlled experimental setting

How unfair is private learning?

Amartya Sanyal[*1,3] **Yaxi Hu**[*2] **Fanny Yang**[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

bangs

Pointy nose

DP and Disparate Impact

Controlled experimental setting

How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

bangs

Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

DP and Disparate Impact

Controlled experimental setting

How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

bangs

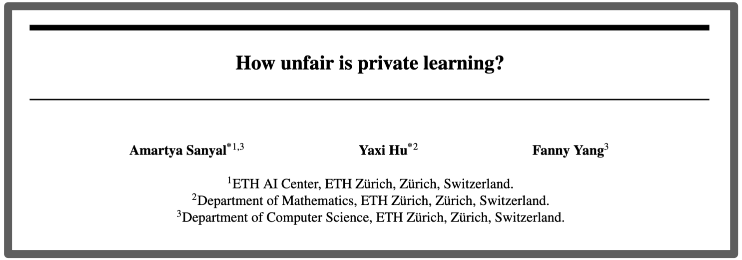
Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

- Subpopulation 1 Eyeglass, no bangs, no pointy nose,...

DP and Disparate Impact

Controlled experimental setting



How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

bangs

Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

- Subpopulation 1 Eyeglass, no bangs, no pointy nose,...
- Subpopulation 2 Eyeglass, bangs, no pointy nose,...

DP and Disparate Impact
Controlled experimental setting

How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

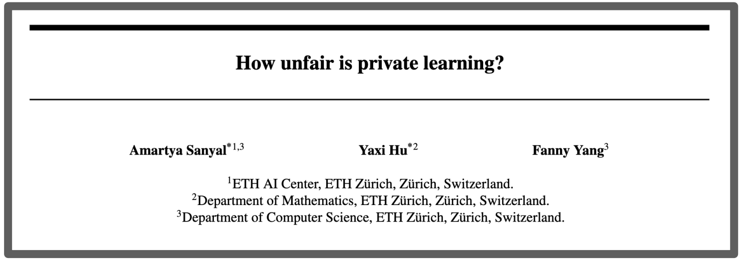
bangs

Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

- Subpopulation 1 Eyeglass, no bangs, no pointy nose,...
- Subpopulation 2 Eyeglass, bangs, no pointy nose,...
- …

DP and Disparate Impact
Controlled experimental setting


How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

bangs

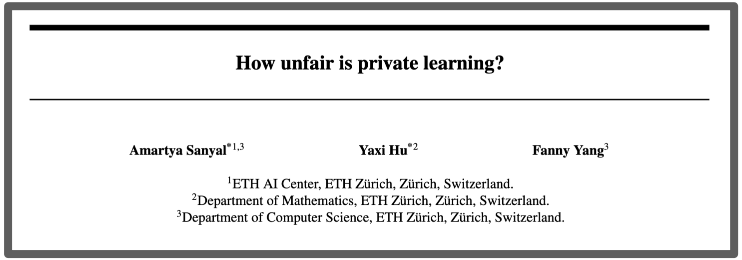
Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

- Subpopulation 1 Eyeglass, no bangs, no pointy nose,**...**
- Subpopulation 2 Eyeglass, bangs, no pointy nose,**...**
- …
- …

DP and Disparate Impact

Controlled experimental setting



How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

eyeglass

bangs

Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

- Subpopulation 1 Eyeglass, no bangs, no pointy nose,...
- Subpopulation 2 Eyeglass, bangs, no pointy nose,...
- …
- …
- Subpopulation 2^{40}

DP and Disparate Impact
Controlled experimental setting

How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

40 binary attributes
for each image

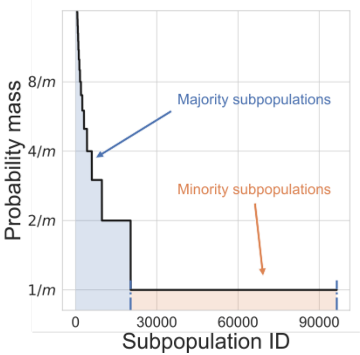
eyeglass

bangs

Pointy nose

40 binary attributes -> **2^{40} subpopulations.**

- Subpopulation 1 Eyeglass, no bangs, no pointy nose,...
- Subpopulation 2 Eyeglass, bangs, no pointy nose,...
- …
- …
- Subpopulation 2^{40}



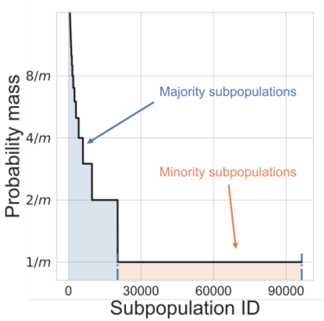
DP and Disparate Impact

Controlled experimental setting

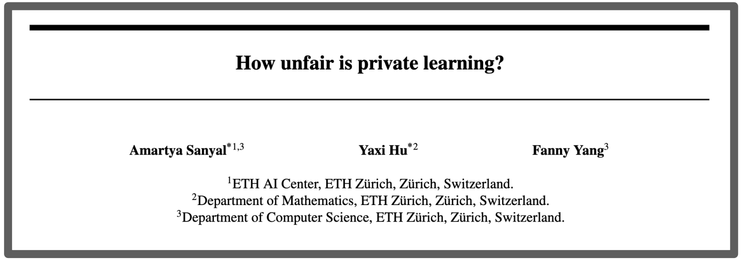
How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

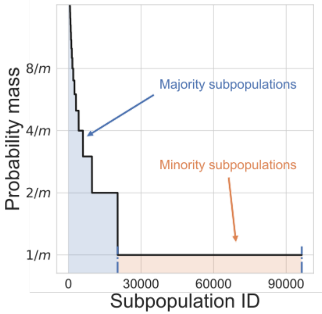
[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

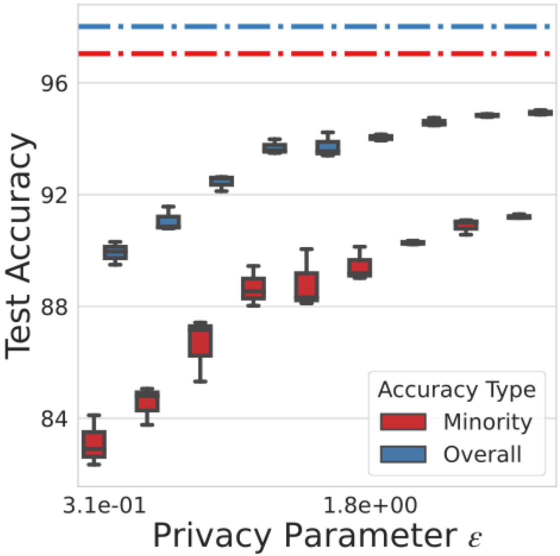


DP and Disparate Impact
Controlled experimental setting


How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]
[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.





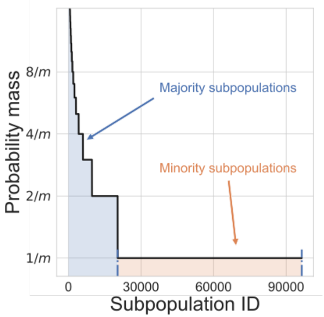
DP and Disparate Impact

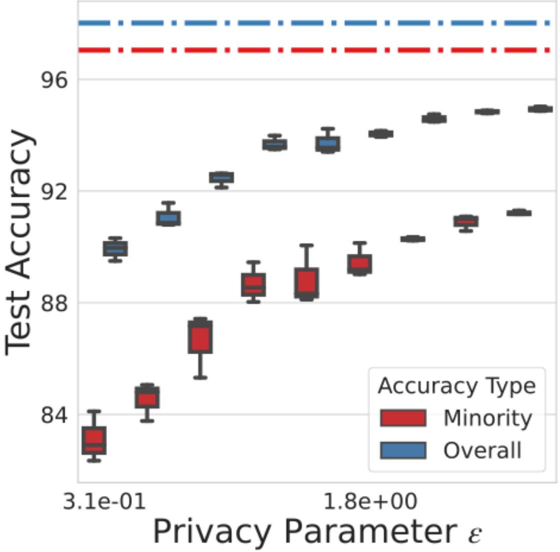
Controlled experimental setting

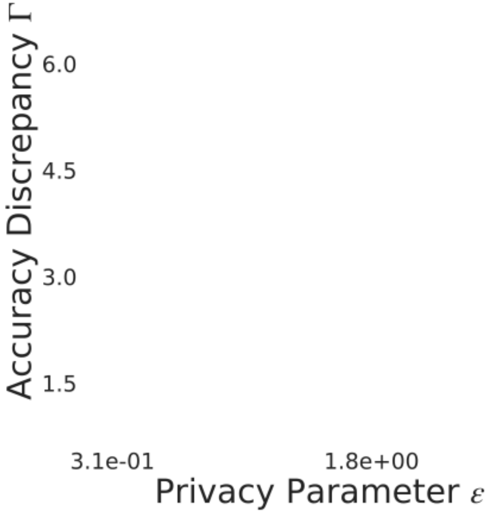
How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

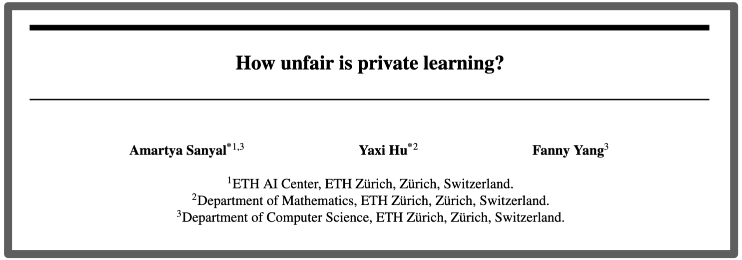


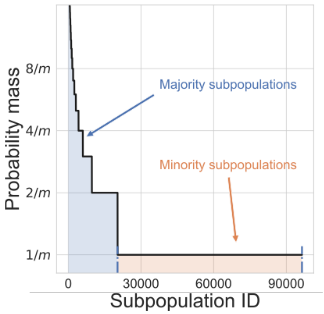


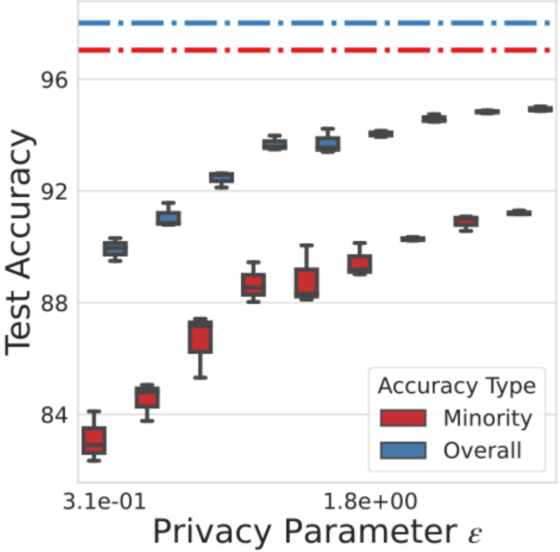


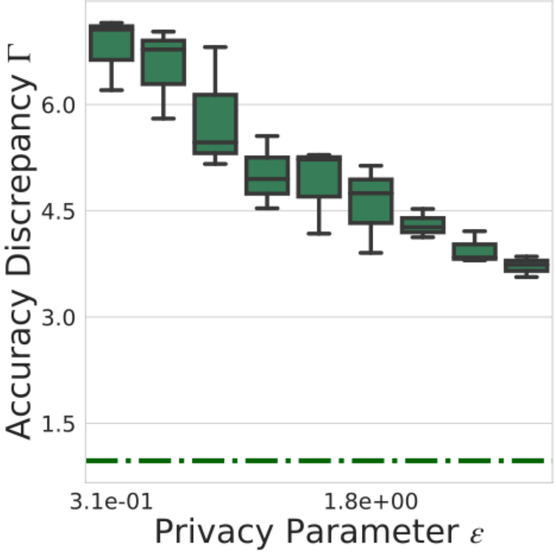
DP and Disparate Impact

Controlled experimental setting









DP and Disparate Impact

Trade-off in long-tailed data

How unfair is private learning?

Amartya Sanyal[*1,3] **Yaxi Hu**[*2] **Fanny Yang**[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

DP and Disparate Impact

Trade-off in long-tailed data

How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ)-DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.

DP and Disparate Impact

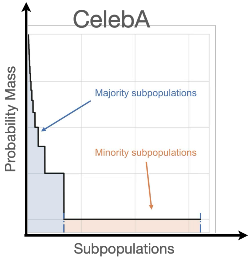
Trade-off in long-tailed data

How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ)-DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.



DP and Disparate Impact

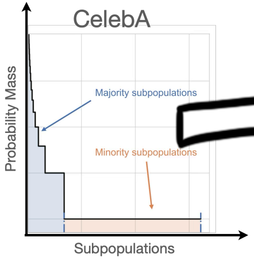
Trade-off in long-tailed data

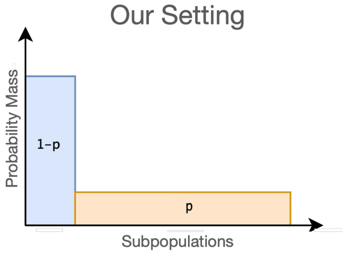
How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

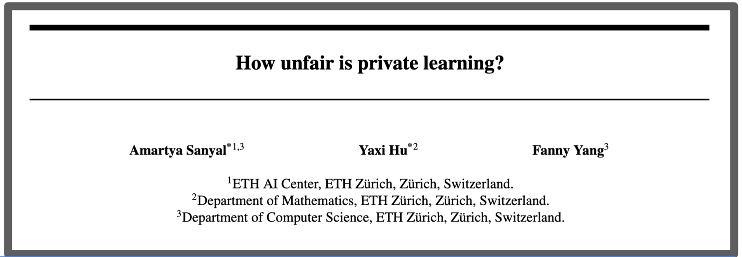
Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ)-DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.





DP and Disparate Impact

Trade-off in long-tailed data



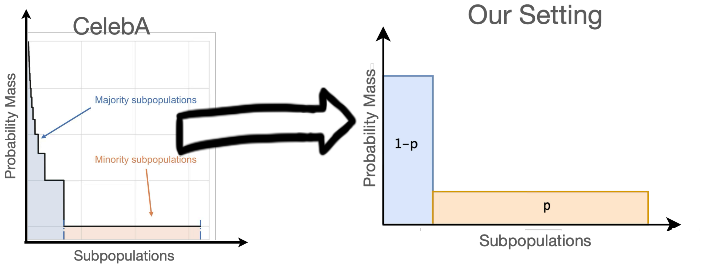
How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ)-DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.

If $\dfrac{N}{m} \to c$ as $N, m \to \infty$



DP and Disparate Impact

Trade-off in long-tailed data

How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

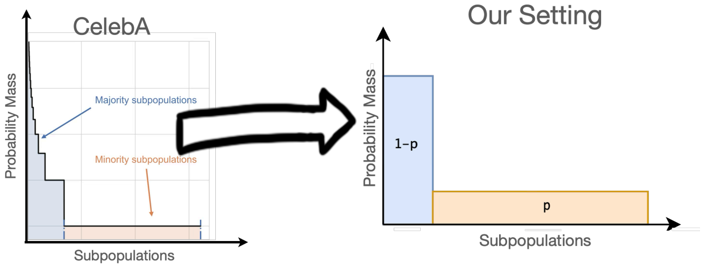
[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ) -DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.

$$\text{If } \frac{N}{m} \to c \text{ as } N, m \to \infty$$

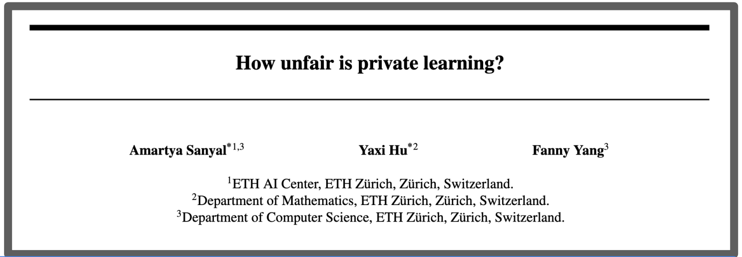
We prove a lower bound on the *accuracy discrepancy* of $\mathcal{A}$ which

- **Increases with the privacy** of $\mathcal{A}$ i.e. smaller (ε, δ)



DP and Disparate Impact

Trade-off in long-tailed data



How unfair is private learning?

Amartya Sanyal*[1,3] Yaxi Hu*[2] Fanny Yang[3]

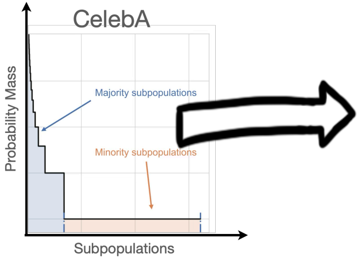
[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

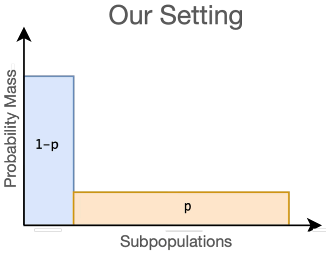
Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ)-DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.

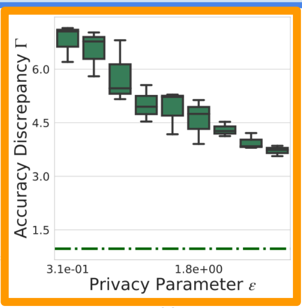
$$\text{If } \frac{N}{m} \to c \text{ as } N, m \to \infty$$

We prove a lower bound on the *accuracy discrepancy* of $\mathcal{A}$ which

- **Increases with the privacy** of $\mathcal{A}$ i.e. smaller (ε, δ)

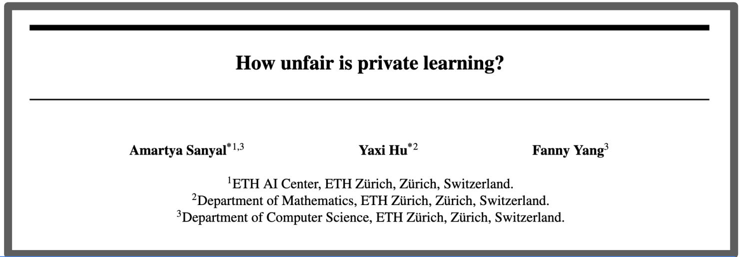






DP and Disparate Impact

Trade-off in long-tailed data



How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

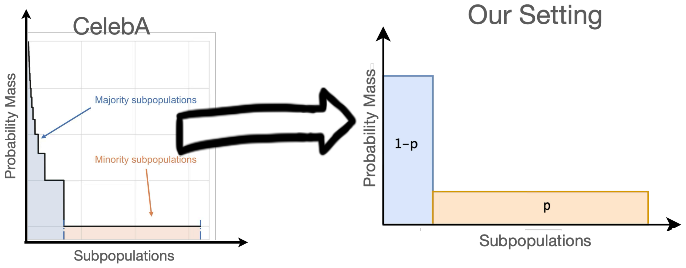
[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

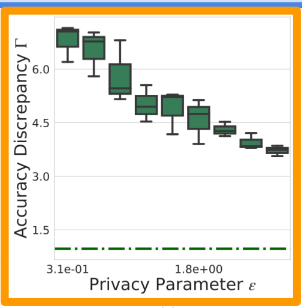
Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ) -DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.

$$\text{If } \frac{N}{m} \to c \text{ as } N, m \to \infty$$

We prove a lower bound on the *accuracy discrepancy* of $\mathcal{A}$ which

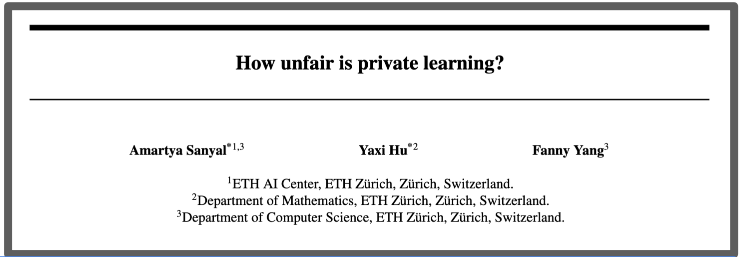
- **Increases with the privacy** of $\mathcal{A}$ i.e. smaller (ε, δ)
- **Increases with long-tailed** nature of data i.e. relative number of minority subpopulations c





DP and Disparate Impact

Trade-off in long-tailed data



How unfair is private learning?

Amartya Sanyal[*1,3] Yaxi Hu[*2] Fanny Yang[3]

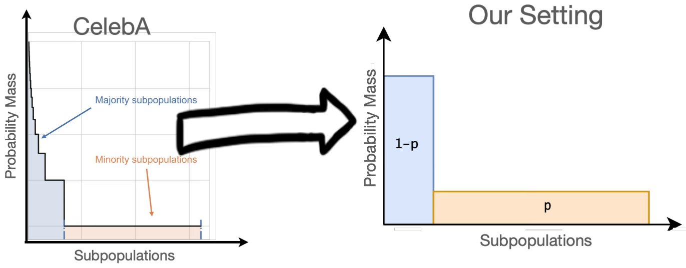
[1]ETH AI Center, ETH Zürich, Zürich, Switzerland.
[2]Department of Mathematics, ETH Zürich, Zürich, Switzerland.
[3]Department of Computer Science, ETH Zürich, Zürich, Switzerland.

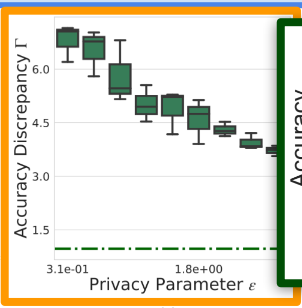
Theorem Consider a *long-tailed distribution* with N sub-populations and sample an m-sized dataset. Consider any (ε, δ)-DP algorithm $\mathcal{A}$ that achieves *low error* on this dataset.

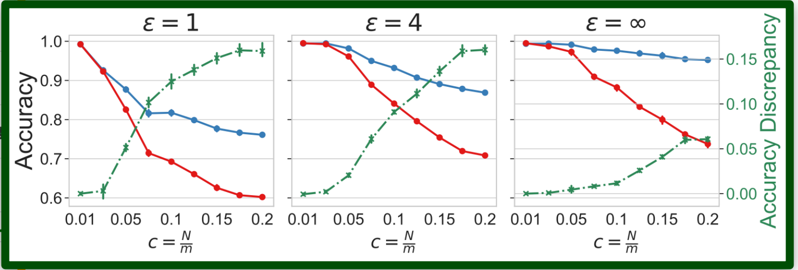
$$\text{If } \frac{N}{m} \to c \text{ as } N, m \to \infty$$

We prove a lower bound on the *accuracy discrepancy* of $\mathcal{A}$ which

- **Increases with the privacy** of $\mathcal{A}$ i.e. smaller (ε, δ)
- **Increases with long-tailed** nature of data i.e. relative number of minority subpopulations c







DP and Disparate Impact

Fundamental Impossibility

DP and Disparate Impact

Fundamental Impossibility

Trade-Offs between Fairness and Privacy in Machine Learning

Sushant Agarwal
University of Waterloo, Canada
sushant.agarwal@uwaterloo.ca

On the Compatibility of Privacy and Fairness

Rachel Cummings* Varun Gupta* Dhamma Kimpara* Jamie Morgenstern*

DP and Disparate Impact

Fundamental Impossibility

Trade-Offs between Fairness and Privacy in Machine Learning

Sushant Agarwal
University of Waterloo, Canada
sushant.agarwal@uwaterloo.ca

On the Compatibility of Privacy and Fairness

Rachel Cummings* Varun Gupta* Dhamma Kimpara* Jamie Morgenstern*

Theorem For any hypothesis class $\mathcal{H}$, no algorithm can simultaneously be $(\varepsilon, 0)$-DP for $\varepsilon < \infty$ and always output a $h \in \mathcal{H}$ that satisfies *equal opportunity* and has *error less than for any constant classifier.*

DP and Disparate Impact

Fundamental Impossibility

Trade-Offs between Fairness and Privacy in Machine Learning

Sushant Agarwal
University of Waterloo, Canada
sushant.agarwal@uwaterloo.ca

On the Compatibility of Privacy and Fairness

Rachel Cummings* Varun Gupta* Dhamma Kimpara* Jamie Morgenstern*

Theorem For any hypothesis class $\mathcal{H}$, no algorithm can simultaneously be $(\varepsilon, 0)$ -DP for $\varepsilon < \infty$ and always output a $h \in \mathcal{H}$ that satisfies *equal opportunity* and has *error less than for any constant classifier.*

- Proof idea:
  - Obs 1. - If a classifier h has non-zero probability to be output under algorithm $\mathcal{A}$ on dataset S_1, it also has non-zero probability to be output on dataset S_2, for all S_2.

DP and Disparate Impact

Fundamental Impossibility

Trade-Offs between Fairness and Privacy in Machine Learning

Sushant Agarwal
University of Waterloo, Canada
sushant.agarwal@uwaterloo.ca

On the Compatibility of Privacy and Fairness

Rachel Cummings* Varun Gupta* Dhamma Kimpara* Jamie Morgenstern*

Theorem For any hypothesis class $\mathcal{H}$, no algorithm can simultaneously be $(\varepsilon, 0)$-DP for $\varepsilon < \infty$ and always output a $h \in \mathcal{H}$ that satisfies *equal opportunity* and has *error less than for any constant classifier.*

- Proof idea:
  - Obs 1. - If a classifier h has non-zero probability to be output under algorithm $\mathcal{A}$ on dataset S_1, it also has non-zero probability to be output on dataset S_2, for all S_2.
  - Obs 2. - Construct two datasets S_1, S_2 such that no classifier, except a constant classifier can be simultaneously fair on both.

DP and Disparate Impact

Other causes

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

In particular two factors are usually isolated,

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

In particular two factors are usually isolated,

- **Gradient Clipping**

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

In particular two factors are usually isolated,

- **Gradient Clipping**
- **Noise** addition

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

In particular two factors are usually isolated,

- **Gradient Clipping**
- **Noise** addition

Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan Cornell Tech eugene@cs.cornell.edu

Omid Poursaeed* Cornell Tech op63@cornell.edu

Vitaly Shmatikov Cornell Tech shmat@cs.cornell.edu

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

In particular two factors are usually isolated,

- **Gradient Clipping**
- **Noise** addition

Differentially Private Empirical Risk Minimization under the Fairness Lens

Cuong Tran Syracuse University ctran@syr.edu
My H. Dinh Syracuse University mydinh@syr.edu
Ferdinando Fioretto Syracuse University ffiorett@syr.edu

Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan Cornell Tech eugene@cs.cornell.edu
Omid Poursaeed* Cornell Tech op63@cornell.edu
Vitaly Shmatikov Cornell Tech shmat@cs.cornell.edu

DP and Disparate Impact

Other causes

Apart from the properties of the data, other reasons are also known to exacerbate unfairness for Differentially Private models

In particular two factors are usually isolated,

- **Gradient Clipping**
- **Noise** addition

Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan Cornell Tech eugene@cs.cornell.edu
Omid Poursaeed* Cornell Tech op63@cornell.edu
Vitaly Shmatikov Cornell Tech shmat@cs.cornell.edu

Differentially Private Empirical Risk Minimization under the Fairness Lens

Cuong Tran Syracuse University ctran@syr.edu
My H. Dinh Syracuse University mydinh@syr.edu
Ferdinando Fioretto Syracuse University ffiorett@syr.edu

Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent

Depeng Xu University of Arkansas Fayetteville, AR, USA depengxu@uark.edu
Wei Du University of Arkansas Fayetteville, AR, USA wd005@uark.edu
Xintao Wu University of Arkansas Fayetteville, AR, USA xintaowu@uark.edu

Good data requires less noise

DP with “good” data

Favourable data properties

DP with “good” data

Favourable data properties

- Privacy guarantees are unconditional - hold for all datasets.

DP with “good” data

Favourable data properties

- Privacy guarantees are unconditional - hold for all datasets.
- Generally, these worst case bound yields a large “*cost of DP”.*

DP with "good" data

Favourable data properties

- Privacy guarantees are unconditional - hold for all datasets.
- Generally, these worst case bound yields a large "*cost of DP*".

Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds

Raef Bassily* Adam Smith* † Abhradeep Thakurta‡

Theorem For every convex function in $\mathbb{R}^d$, DP-SGD satisfies DP and cost of DP is less than $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

DP with "good" data

Favourable data properties

- Privacy guarantees are unconditional - hold for all datasets.
- Generally, these worst case bound yields a large "*cost of DP*".

Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds

Raef Bassily* Adam Smith*† Abhradeep Thakurta‡

Theorem For every convex function in $\mathbb{R}^d$, DP-SGD satisfies DP and cost of DP is less than $\frac{LR\sqrt{d}}{\varepsilon n}$.

Theorem Exists a convex loss function and data in $\mathbb{R}^d$, s.t. all DP algorithms incur extra cost $\frac{LR\sqrt{d}}{\varepsilon n}$.

DP with "good" data

Favourable data properties

- Privacy guarantees are unconditional - hold for all datasets.
- Generally, these worst case bound yields a large "*cost of DP*".

Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds

Raef Bassily* Adam Smith*† Abhradeep Thakurta‡

Can we do better for "nice" datasets ?

Theorem For every convex function in $\mathbb{R}^d$, DP-SGD satisfies DP and cost of DP is less than $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

Theorem Exists a convex loss function and data in $\mathbb{R}^d$, s.t. all DP algorithms incur extra cost $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

DP with "good" data

Favourable data properties

- Privacy guarantees are unconditional - <u>hold for all datasets.</u>
- Generally, these worst case bound yields a large "*cost of DP*".

Can we do better for "nice" datasets ?

Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds

Raef Bassily* Adam Smith*† Abhradeep Thakurta‡

Theorem For every convex function in $\mathbb{R}^d$, DP-SGD satisfies DP and cost of DP is less than $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

Theorem Exists a convex loss function and data in $\mathbb{R}^d$, s.t. all DP algorithms incur extra cost $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

DP with "good" data

Favourable data properties

- Privacy guarantees are unconditional - <u>hold for all datasets.</u>
- Generally, these worst case bound yields a large "*cost of DP*".

Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds

Raef Bassily* Adam Smith*† Abhradeep Thakurta‡

Can we do better for "nice" datasets ?

Theorem For every convex function in $\mathbb{R}^d$, DP-SGD satisfies DP and cost of DP is less than $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

Theorem Exists a convex loss function and data in $\mathbb{R}^d$, s.t. all DP algorithms incur extra cost $\dfrac{LR\sqrt{d}}{\varepsilon n}$.

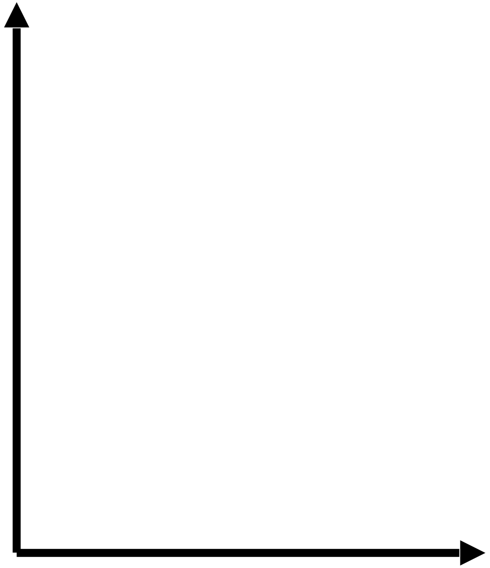
Naive DP Mean estimation

Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R ;

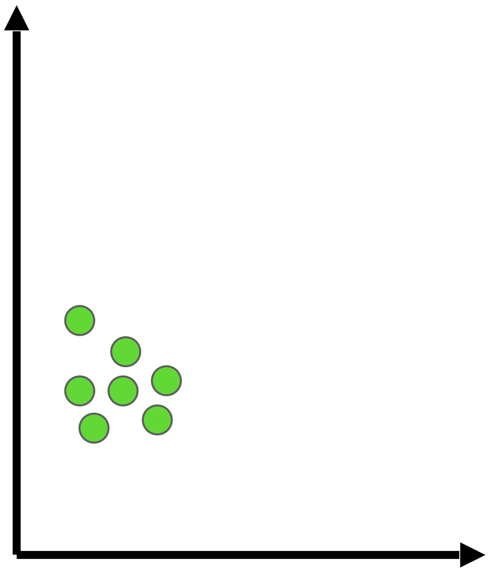
Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R;



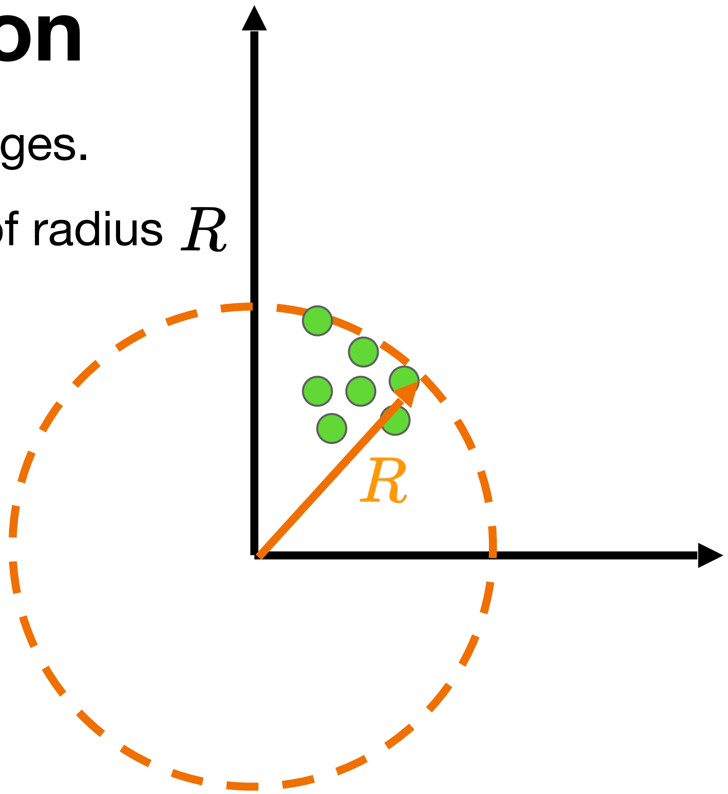
Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R ;



Naive DP Mean estimation

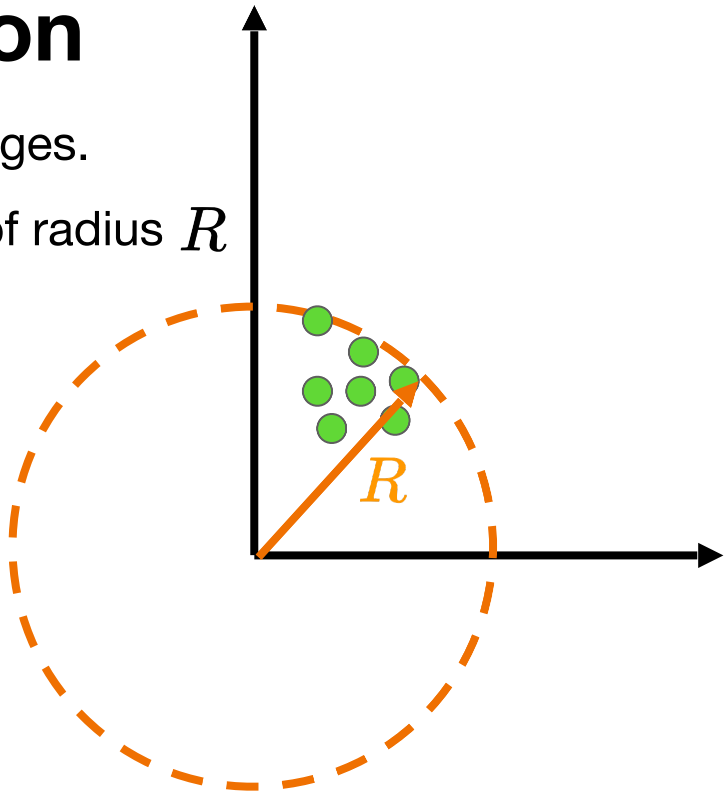
- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R ;



Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R;
- error scales as

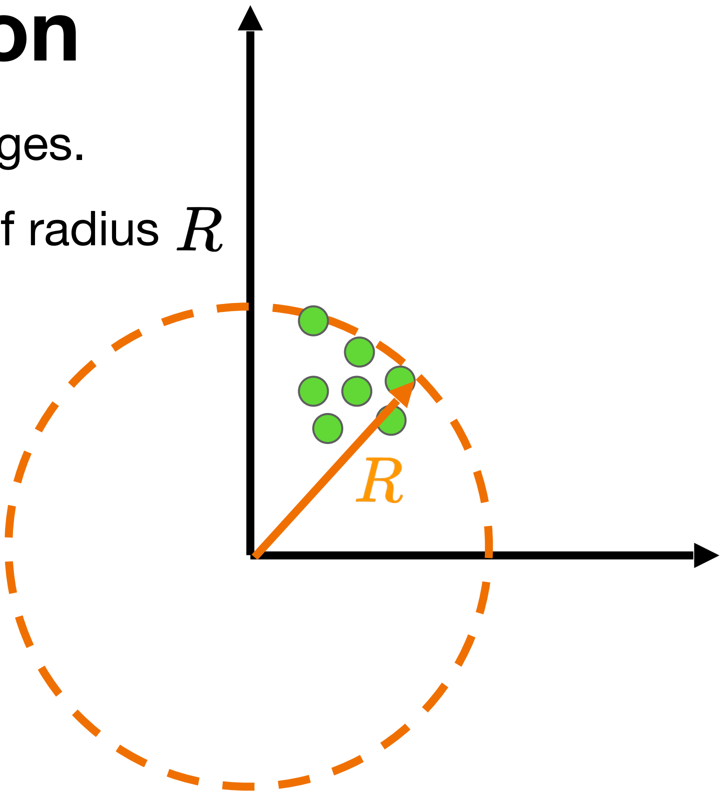
$$||\hat{\mu} - \mu|| \leq \sqrt{\frac{d}{n} + \frac{R^2 d^2}{n^2 \epsilon}}$$



Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R;
- error scales as

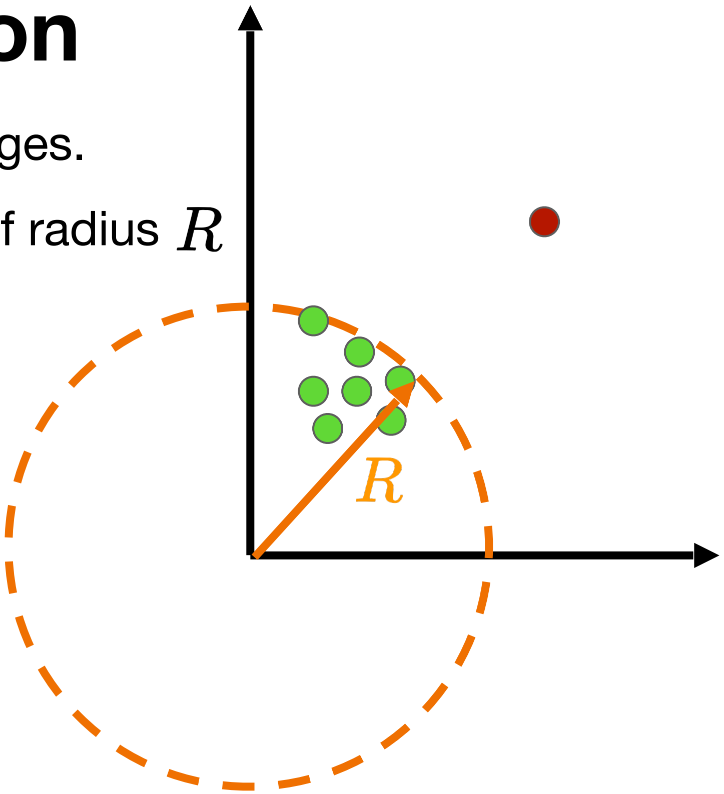
$$\|\hat{\mu} - \mu\| \leq \sqrt{\frac{d}{n} + \boxed{\frac{R^2 d^2}{n^2 \epsilon}}}$$



Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R;
- error scales as

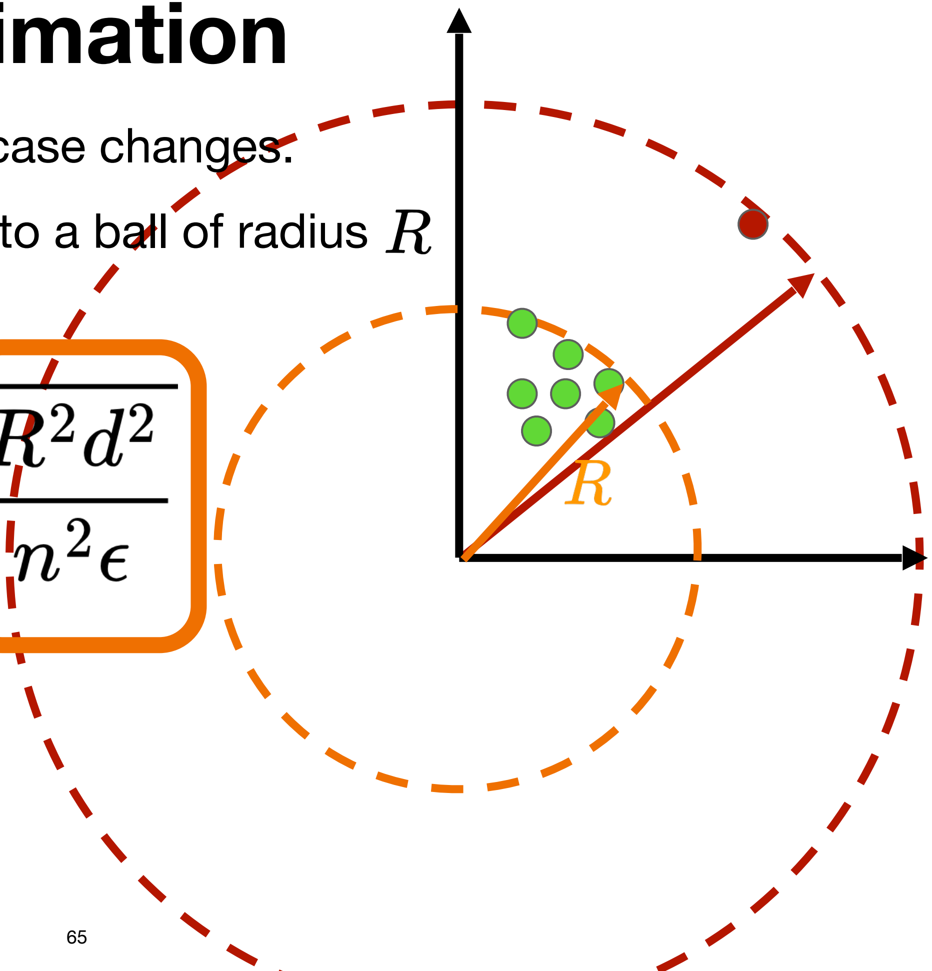
$$\|\hat{\mu} - \mu\| \leq \sqrt{\frac{d}{n} + \boxed{\frac{R^2 d^2}{n^2 \epsilon}}}$$



Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R;
- error scales as

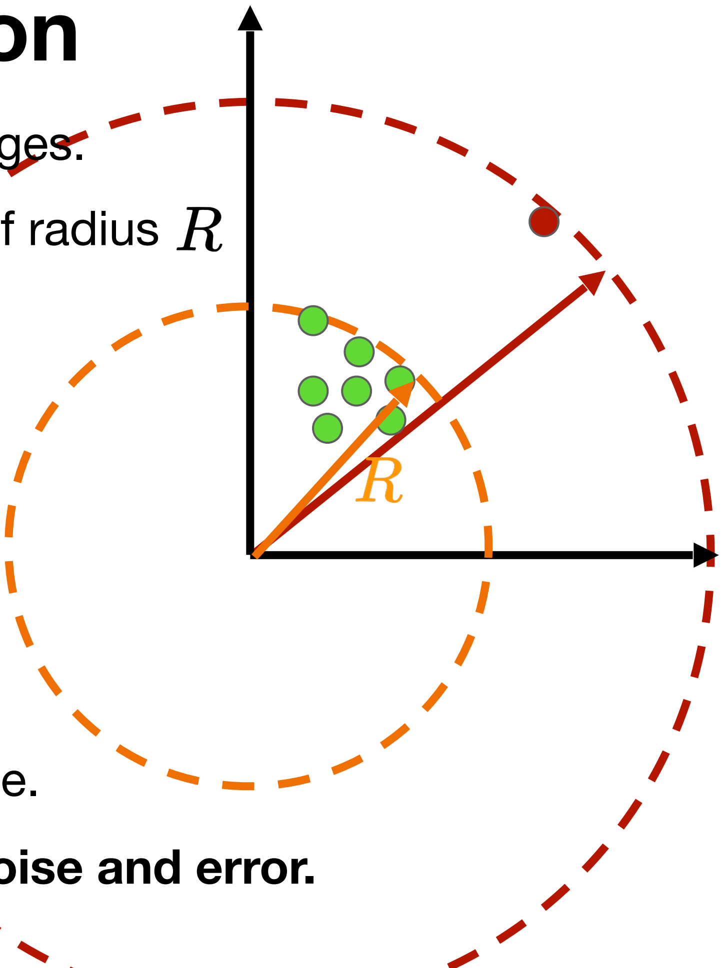
$$||\hat{\mu} - \mu|| \leq \sqrt{\frac{d}{n} + \boxed{\frac{R^2 d^2}{n^2 \epsilon}}}$$



Naive DP Mean estimation

- DP must protect against worst-case changes.
- A naïve DP estimator clips data to a ball of radius R
- Noise is calibrated to R;
- error scales as

$$\|\hat{\mu} - \mu\| \leq \sqrt{\frac{d}{n} + \boxed{\frac{R^2 d^2}{n^2 \epsilon}}}$$



- Heavy tails or outliers force R to be large.
- **Worst-case sensitivity leads to high noise and error.**

Friendly-Core estimation

FriendlyCore: Practical Differentially Private Aggregation

Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour*
Uri Stemmer*

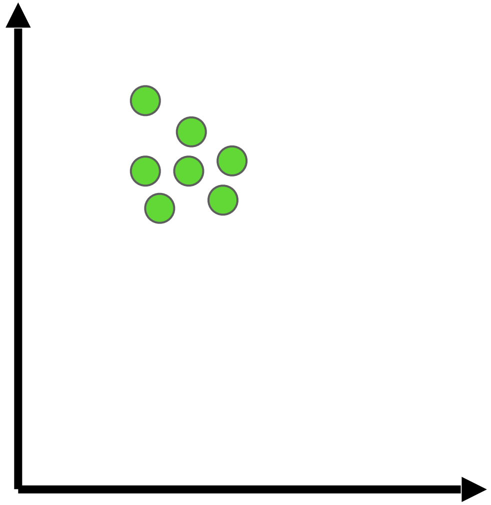
In many settings, most points lie in a ball of radius

Friendly-Core estimation

FriendlyCore: Practical Differentially Private Aggregation

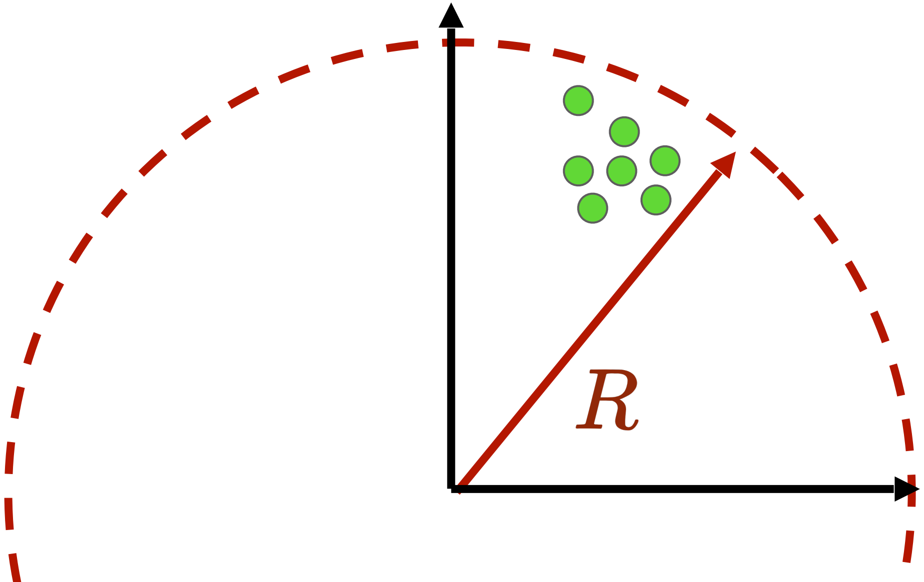
Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour*
Uri Stemmer*

In many settings, most points lie in a ball of radius



Friendly-Core estimation

In many settings, most points lie in a ball of radius



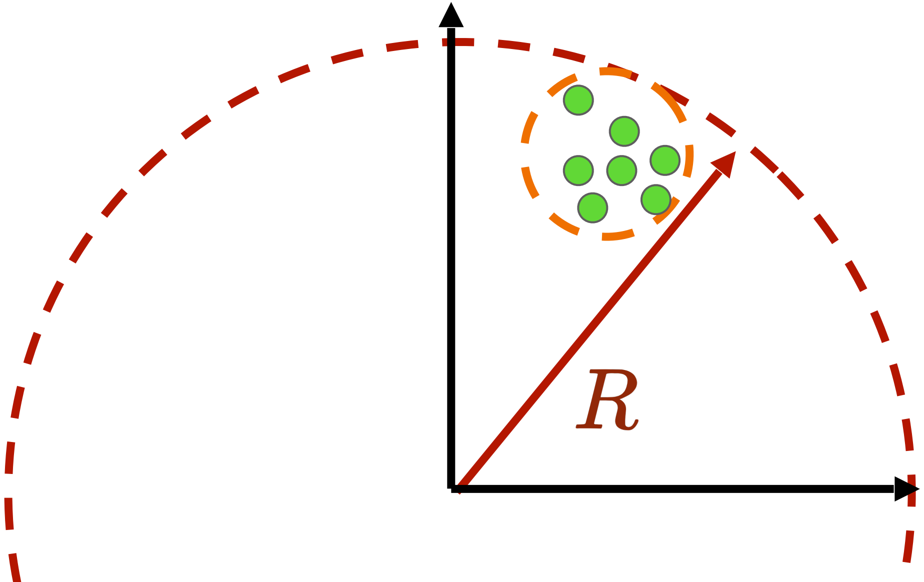
Friendly-Core estimation


FriendlyCore: Practical Differentially Private Aggregation

Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour* Uri Stemmer*

In many settings, most points lie in a ball of radius

- A dataset is **friendly** if every two points have a common neighbor within r



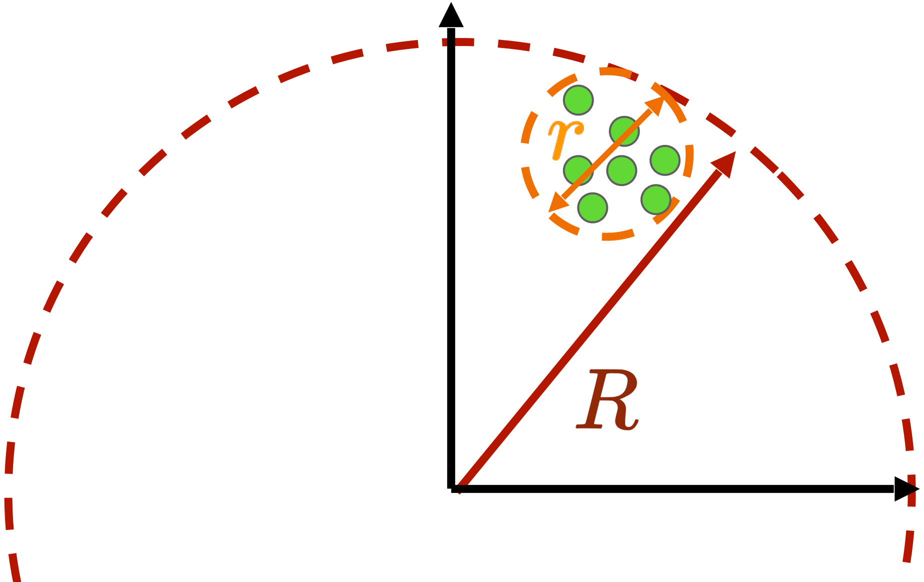
Friendly-Core estimation

FriendlyCore: Practical Differentially Private Aggregation

Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour*
Uri Stemmer*

In many settings, most points lie in a ball of radius

- A dataset is **friendly** if every two points have a common neighbor within r



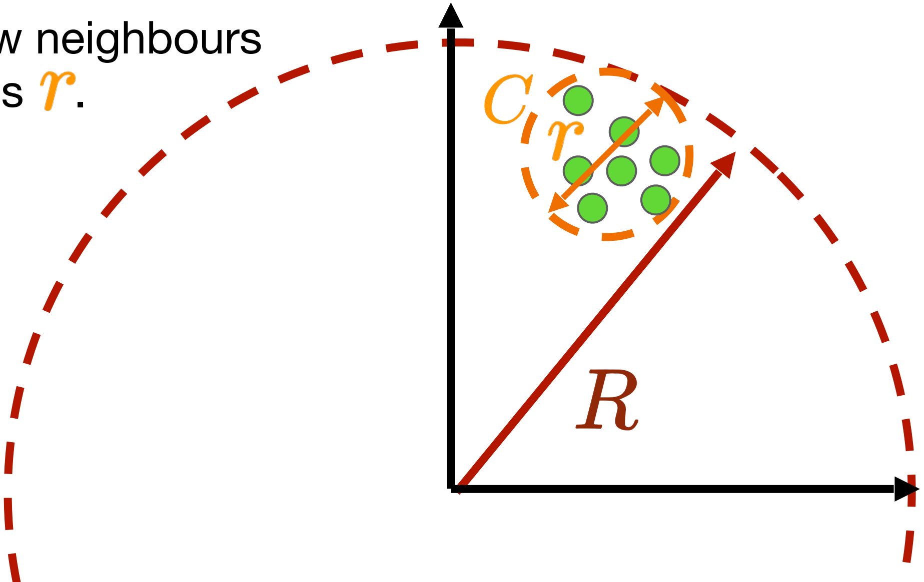
Friendly-Core estimation

FriendlyCore: Practical Differentially Private Aggregation

Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour* Uri Stemmer*

In many settings, most points lie in a ball of radius

- A dataset is **friendly** if every two points have a common neighbor within r
- **Friendly-Core** removes points with few neighbours and outputs w.h.p. a core C with radius r.



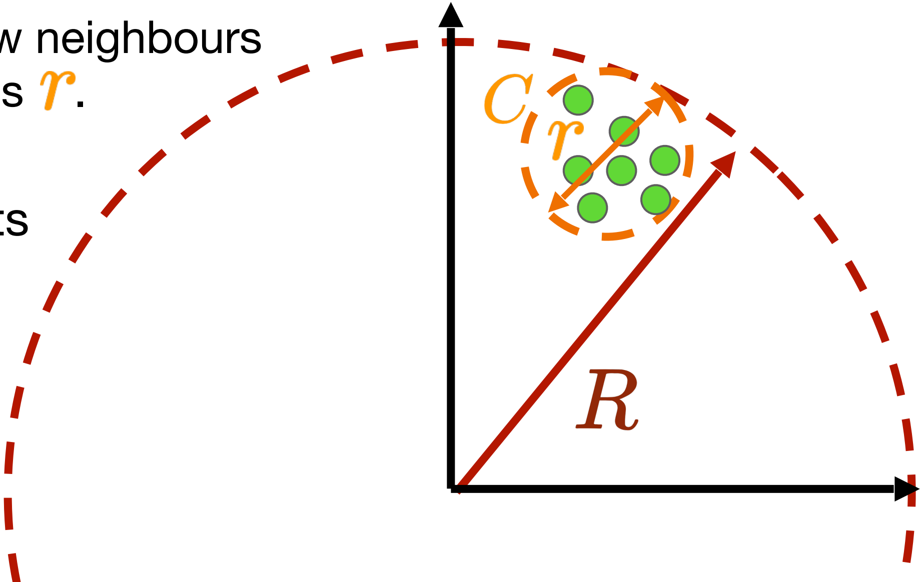
Friendly-Core estimation

FriendlyCore: Practical Differentially Private Aggregation

Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour*
Uri Stemmer*

In many settings, most points lie in a ball of radius

- A dataset is **<u>friendly</u>** if every two points have a common neighbor within r
- **<u>Friendly-Core</u>** removes points with few neighbours and outputs w.h.p. a core C with radius r.
- Estimating mean in the core $\mathcal{C}$ results



Friendly-Core estimation

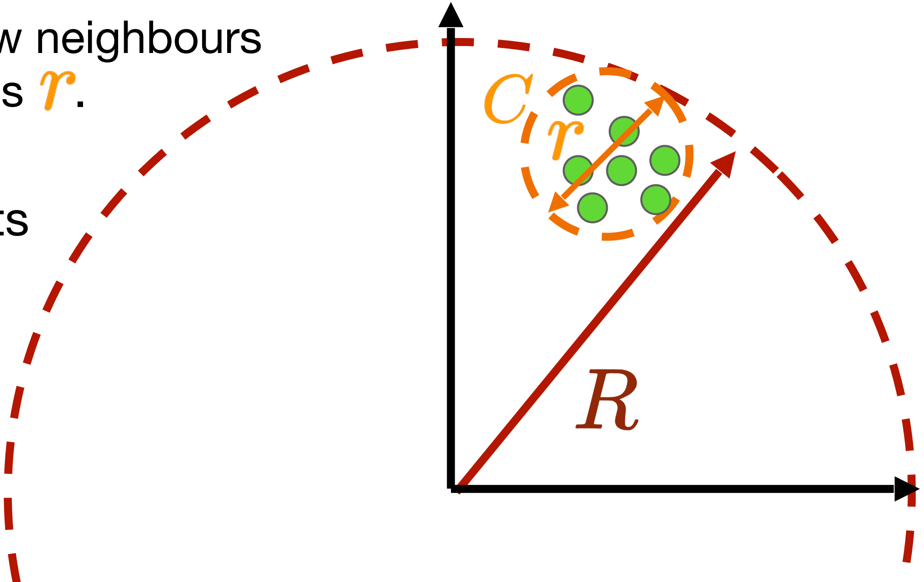
FriendlyCore: Practical Differentially Private Aggregation

Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour*
Uri Stemmer*

In many settings, most points lie in a ball of radius

- A dataset is **friendly** if every two points have a common neighbor within r
- **Friendly-Core** removes points with few neighbours and outputs w.h.p. a core C with radius r.
- Estimating mean in the core $\mathcal{C}$ results

$$\|\hat{\mu} - \mu\| \leq \sqrt{\frac{d}{n}} + \frac{r\sqrt{d}}{n\epsilon}$$



Friendly-Core estimation

FriendlyCore: Practical Differentially Private Aggregation

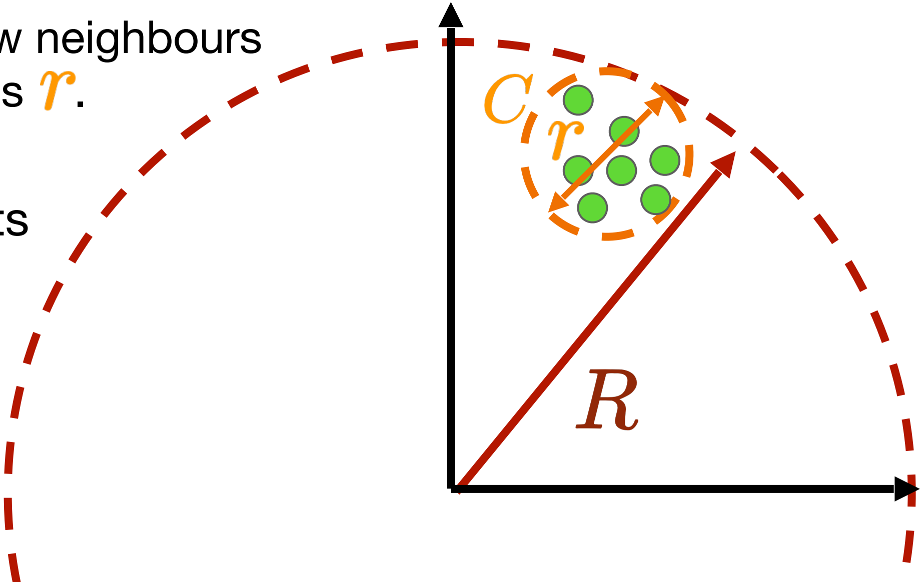
Eliad Tsfadia* Edith Cohen* Haim Kaplan* Yishay Mansour*
Uri Stemmer*

In many settings, most points lie in a ball of radius

- A dataset is **friendly** if every two points have a common neighbor within r
- **Friendly-Core** removes points with few neighbours and outputs w.h.p. a core C with radius r.
- Estimating mean in the core $\mathcal{C}$ results

$$\|\hat{\mu} - \mu\| \leq \sqrt{\frac{d}{n}} + \boxed{\frac{r\sqrt{d}}{n\epsilon}}$$

when the data is **friendly**.



DP geometric median

Private Geometric Median

Mahdi Haghifam* **Thomas Steinke†** **Jonathan Ullman‡**

DP geometric median

Private Geometric Median

Mahdi Haghifam* Thomas Steinke† Jonathan Ullman‡

- Related task is estimating the geometric median: solving the following

$$\sum \|\theta - x_i\|_2$$

DP geometric median

Private Geometric Median

Mahdi Haghifam* Thomas Steinke† Jonathan Ullman‡

- Related task is estimating the geometric median: solving the following

$$\sum \|\theta - x_i\|_2$$

- Solving this problem with DP-SGD yields DP cost $\dfrac{R\sqrt{d}}{\varepsilon}$

DP geometric median

Private Geometric Median

Mahdi Haghifam* Thomas Steinke† Jonathan Ullman‡

- Related task is estimating the geometric median: solving the following

$$\sum \|\theta - x_i\|_2$$

- Solving this problem with DP-SGD yields DP cost $\frac{R\sqrt{d}}{\varepsilon}$
- **HSU24** proposes an algorithm which yields DP cost

DP geometric median

Private Geometric Median

Mahdi Haghifam* Thomas Steinke† Jonathan Ullman‡

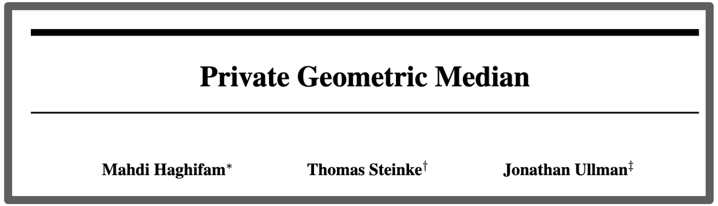
- Related task is estimating the geometric median: solving the following

$$\sum \|\theta - x_i\|_2$$

- Solving this problem with DP-SGD yields DP cost $\dfrac{R\sqrt{d}}{\varepsilon}$
- **HSU24** proposes an algorithm which yields DP cost

$$(\text{effective diameter})\frac{\sqrt{d}}{\varepsilon}$$

DP geometric median



Private Geometric Median

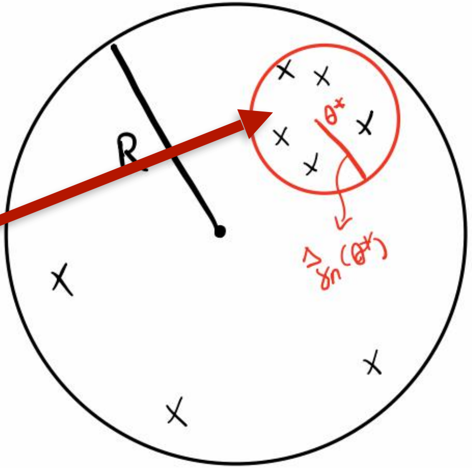
Mahdi Haghifam* Thomas Steinke† Jonathan Ullman‡

- Related task is estimating the geometric median: solving the following

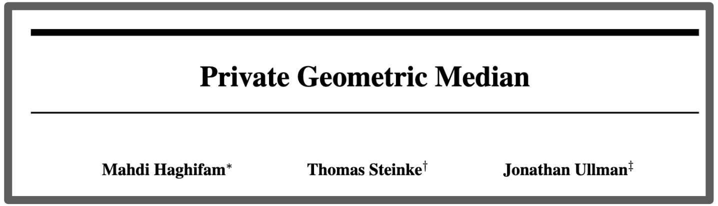
$$\sum \|\theta - x_i\|_2$$

- Solving this problem with DP-SGD yields DP cost $\dfrac{R\sqrt{d}}{\varepsilon}$
- **<u>HSU24</u>** proposes an algorithm which yields DP cost

$$(\text{effective diameter})\frac{\sqrt{d}}{\varepsilon}$$



DP geometric median



Private Geometric Median

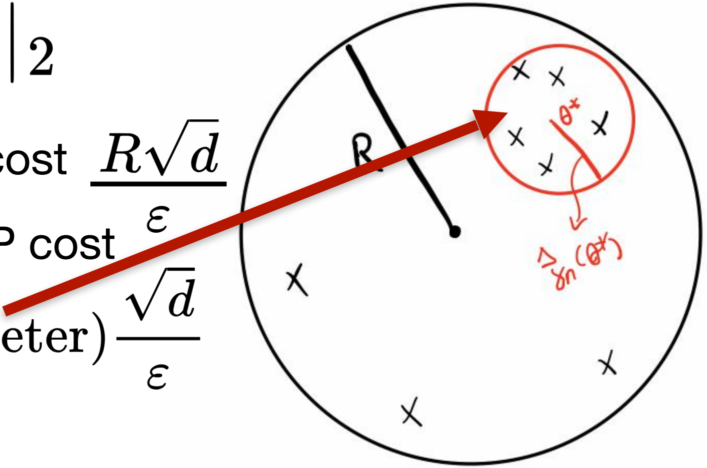
Mahdi Haghifam* Thomas Steinke† Jonathan Ullman‡

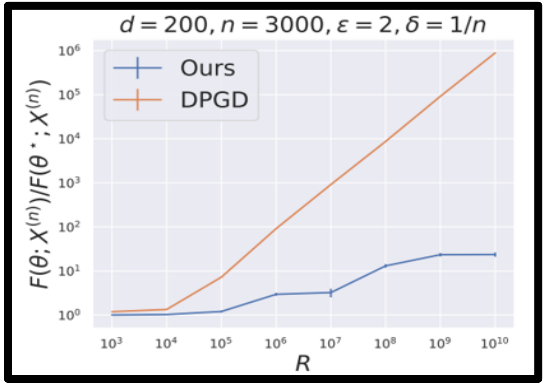
- Related task is estimating the geometric median: solving the following

$$\sum \|\theta - x_i\|_2$$

- Solving this problem with DP-SGD yields DP cost $\dfrac{R\sqrt{d}}{\varepsilon}$
- **<u>HSU24</u>** proposes an algorithm which yields DP cost

$$(\text{effective diameter})\,\frac{\sqrt{d}}{\varepsilon}$$





DP Principal Component Analysis

DP Principal Component Analysis

DP-PCA: Statistically Optimal and
Differentially Private PCA

Xiyang Liu [*] Weihao Kong [†] Prateek Jain [‡] Sewoong Oh [§]

DP Principal Component Analysis

Lower bound for any data

DP-PCA: Statistically Optimal and Differentially Private PCA

Xiyang Liu * Weihao Kong † Prateek Jain ‡ Sewoong Oh §

Theorem 5.4 (Lower bound without As

that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–A.3 with $M = \tilde{O}(d + \sqrt{n\varepsilon/d})$, $V = O(d)$ and $\gamma = O(1)$ is denoted by $\tilde{\mathcal{P}}$. For $d \geq 2$, there exists a universal constant $C > 0$ such that

$$\inf_{\hat{v} \in \mathcal{M}_\varepsilon} \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_{S \sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \geq C\kappa \min\left(\boxed{\sqrt{\frac{d \wedge \log\left(\left(1 - e^{-\varepsilon}\right)/\delta\right)}{\varepsilon n}}}, 1 \right) . \tag{13}$$

DP Principal Component Analysis

Lower bound for any data

DP-PCA: Statistically Optimal and Differentially Private PCA

Xiyang Liu * Weihao Kong † Prateek Jain ‡ Sewoong Oh §

Theorem 5.4 (Lower bound without As... *that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–A.3 with $M = \tilde{O}(d + \sqrt{n\varepsilon/d})$, $V = O(d)$ and $\gamma = O(1)$ is denoted by $\tilde{\mathcal{P}}$. For $d \geq 2$, there exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v} \in \mathcal{M}_\varepsilon} \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_{S \sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \geq C\kappa \min\left(\boxed{\sqrt{\frac{d \wedge \log\left((1-e^{-\varepsilon})/\delta\right)}{\varepsilon n}}}, 1\right). \tag{13}$$

that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of Gaussian distributions with (λ_1, λ_2) as the first and second eigenvalues of the covariance matrix is denoted by $\mathcal{P}_{(\lambda_1, \lambda_2)}$. There exists a universal constant $C > 0$ such that

$$\inf_{\hat{v} \in \mathcal{M}_\varepsilon} \sup_{P \in \mathcal{P}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S \sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \geq C \min\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)\sqrt{\frac{\lambda_2}{\lambda_1}}, 1\right). \tag{12}$$

DP Principal Component Analysis

Lower bound for any data

DP-PCA: Statistically Optimal and Differentially Private PCA

Xiyang Liu * Weihao Kong † Prateek Jain ‡ Sewoong Oh §

Theorem 5.4 (Lower bound without As... *that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–A.3 with $M = \tilde{O}(d + \sqrt{n\varepsilon/d})$, $V = O(d)$ and $\gamma = O(1)$ is denoted by $\tilde{\mathcal{P}}$. For $d \geq 2$, there exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v} \in \mathcal{M}_\varepsilon} \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_{S \sim P^n}[\sin(\hat{v}(S), v_1)] \geq C\kappa \min\left(\sqrt{\boxed{\frac{d \wedge \log\left((1-e^{-\varepsilon})/\delta\right)}{\varepsilon n}}}, 1\right) . \quad (13)$$

that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of Gaussian distributions with (λ_1, λ_2) as the first and second eigenvalues of the covariance matrix is denoted by $\mathcal{P}_{(\lambda_1,\lambda_2)}$. There exists a universal constant $C > 0$ such that

$$\inf_{\hat{v} \in \mathcal{M}_\varepsilon} \sup_{P \in \mathcal{P}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S \sim P^n}[\sin(\hat{v}(S), v_1)] \geq C \min\left(\kappa\left(\sqrt{\frac{d}{n}} + \boxed{\frac{d}{\varepsilon n}}\right)\sqrt{\frac{\lambda_2}{\lambda_1}}, 1\right) . \quad (12)$$

Lower bound for sub-gaussian

DP Principal Component Analysis

Lower bound for any data

DP-PCA: Statistically Optimal and Differentially Private PCA

Xiyang Liu * Weihao Kong † Prateek Jain ‡ Sewoong Oh §

Theorem 5.4 (Lower bound without As... *that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–* ... *$\gamma = O(1)$ is denoted by $\tilde{\mathcal{P}}$. For $d \geq 2$, there exists a*

$$C\kappa \min\left(\sqrt{\frac{d \wedge \log\left(\left(1-e^{-\varepsilon}\right)/\delta\right)}{\varepsilon n}}, 1\right) . \qquad (13)$$

- Similar results known about other problems including PCA.
- Key takeaway:
  - Privacy guaranteed for **ALL** datasets.
  - ***When data quality is high, utility is better.***

... *A set of Gaussian distributions with (λ_1, λ_2) as the* ... *atrix is denoted by $\mathcal{P}_{(\lambda_1,\lambda_2)}$. There exists a universal*

$$\inf_{\hat{v}\in\mathcal{M}_\varepsilon} \sup_{P\in\mathcal{P}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S\sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \geq C \min\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)\sqrt{\frac{\lambda_2}{\lambda_1}}, 1\right) . \qquad (12)$$

Lower bound for sub-gaussian

DP SGD

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi[*] Andy Chu[*] Ian Goodfellow[†]
H. Brendan McMahan[*] Ilya Mironov[*] Kunal Talwar[*]
Li Zhang[*]

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi[*] Andy Chu[*] Ian Goodfellow[†]
H. Brendan McMahan[*] Ilya Mironov[*] Kunal Talwar[*]
Li Zhang[*]

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi[*]
H. Brendan McMahan[*]
Andy Chu[*]
Ilya Mironov[*]
Li Zhang[*]
Ian Goodfellow[†]
Kunal Talwar[*]

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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi*, H. Brendan McMahan*, Andy Chu*, Ilya Mironov*, Li Zhang*, Ian Goodfellow†, Kunal Talwar*

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**

(1) 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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi*, H. Brendan McMahan*, Andy Chu*, Ilya Mironov*, Li Zhang*, Ian Goodfellow†, Kunal Talwar*

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**

(1) Take a random sample L_t with sampling probability L/N

(2) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi* Andy Chu* Ian Goodfellow†
H. Brendan McMahan* Ilya Mironov* Kunal Talwar*
Li Zhang*

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**

(1) Take a random sample L_t with sampling probability L/N

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

(3) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi* H. Brendan McMahan* Andy Chu* Ilya Mironov* Li Zhang* Ian Goodfellow† Kunal Talwar*

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \dots, x_N\}$, 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**

(1) Take a random sample L_t with sampling probability L/N

(2) **Compute gradient**

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

(3) **Clip gradient**

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

(4) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi*
H. Brendan McMahan*
Andy Chu*
Ilya Mironov*
Li Zhang*
Ian Goodfellow†
Kunal Talwar*

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**

(1) Take a random sample L_t with sampling probability L/N

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

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

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

(5) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.
- As we saw earlier, the added noise scales with dimensionality of params

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi*
H. Brendan McMahan*
Andy Chu*
Ilya Mironov*
Li Zhang*
Ian Goodfellow†
Kunal Talwar*

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**

(1) Take a random sample L_t with sampling probability L/N

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

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

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

(5) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.
- As we saw earlier, the added noise scales with dimensionality of params
- To avoid this, they conduct DP-PCA on data before doing DP-SGD.

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi*, H. Brendan McMahan*, Andy Chu*, Ilya Mironov*, Li Zhang*, Ian Goodfellow†, Kunal Talwar*

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**

(1) Take a random sample L_t with sampling probability L/N

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

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

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

(5) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.
- As we saw earlier, the added noise scales with dimensionality of params
- To avoid this, they conduct DP-PCA on data before doing DP-SGD.

But,

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi*
H. Brendan McMahan*
Andy Chu*
Ilya Mironov*
Li Zhang*
Ian Goodfellow†
Kunal Talwar*

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**

(1) Take a random sample L_t with sampling probability L/N

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

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

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

(5) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.
- As we saw earlier, the added noise scales with dimensionality of params
- To avoid this, they conduct DP-PCA on data before doing DP-SGD.

But,

1. DP-PCA requires **additional time**

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi* Andy Chu* Ian Goodfellow†
H. Brendan McMahan* Ilya Mironov* Kunal Talwar*
Li Zhang*

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**

(1) Take a random sample L_t with sampling probability L/N

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

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

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

(5) **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.

DP SGD

- DP-SGD is the standard workhorse for DP Machine Learning algorithms.
- As we saw earlier, the added noise scales with dimensionality of params
- To avoid this, they conduct DP-PCA on data before doing DP-SGD.

But,

1. DP-PCA requires **additional time**
2. DP-PCA incurs a**dditional privacy cost**

Deep Learning with Differential Privacy

October 25, 2016

Martín Abadi* H. Brendan McMahan* Andy Chu* Ilya Mironov* Li Zhang* Ian Goodfellow† Kunal Talwar*

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**

(1) Take a random sample L_t with sampling probability L/N

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

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

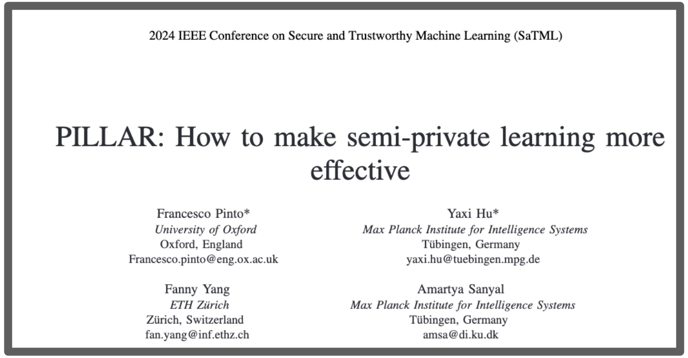
(4) **Add noise**
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

(5) **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.

PILLAR

Leveraging intrinsic low dimensionality



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

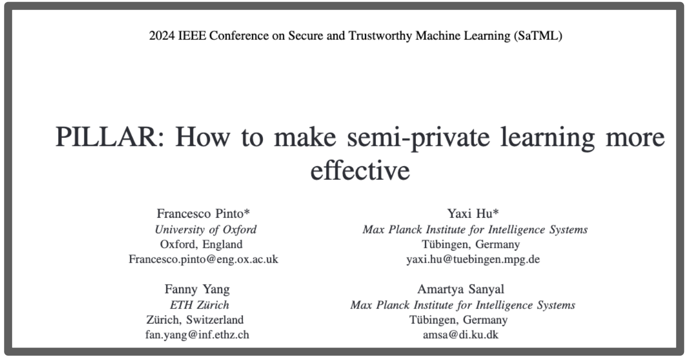
Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

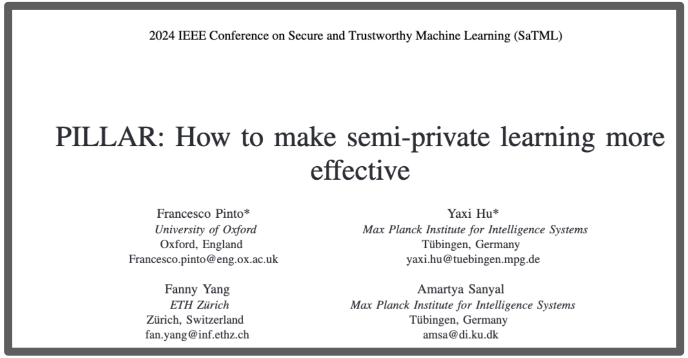
Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection

- But natural data is not inherently low rank in pixel space



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

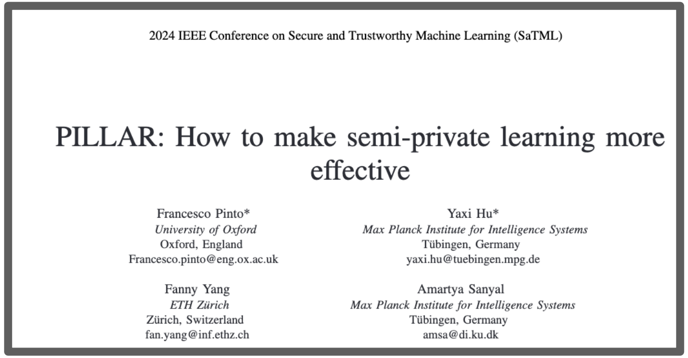
Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

- But natural data is not inherently low rank in pixel space
- Maybe we need to find the right representation space

PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection

- But natural data is not inherently low rank in pixel space
- Maybe we need to find the right representation space

Idea 2: Use any **public unlabelled pre-training** data for representation learning.

2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

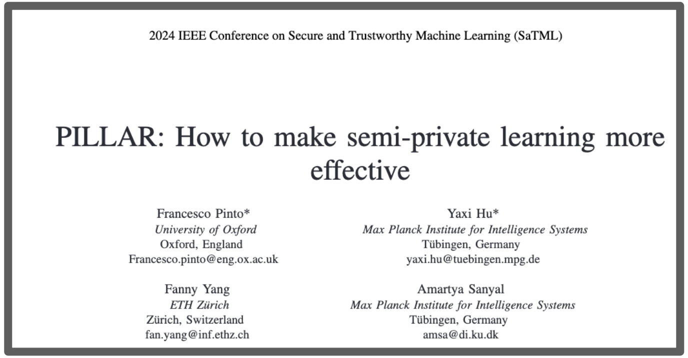
Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection


2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

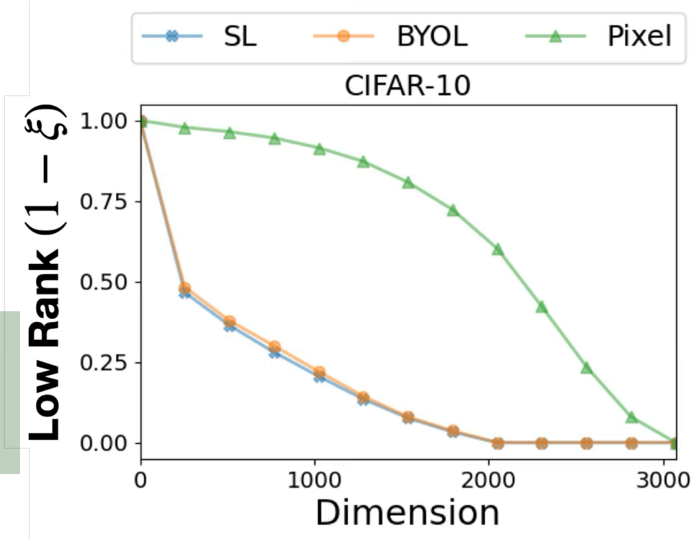
Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

- But natural data is not inherently low rank in pixel space
- Maybe we need to find the right representation space



Idea 2: Use any **public unlabelled pre-training** data for representation learning.

PILLAR

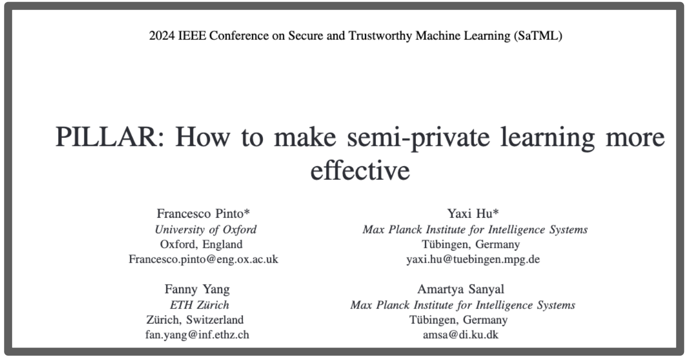
Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection

- But natural data is not inherently low rank in pixel space
- Maybe we need to find the right representation space

Idea 2: Use any **public unlabelled pre-training** data for representation learning.

$1-\xi$ = low rank reconstruction error



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

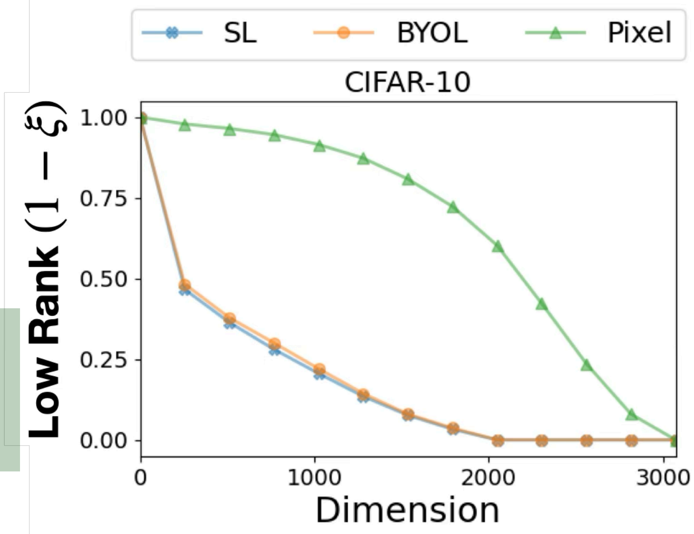
PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

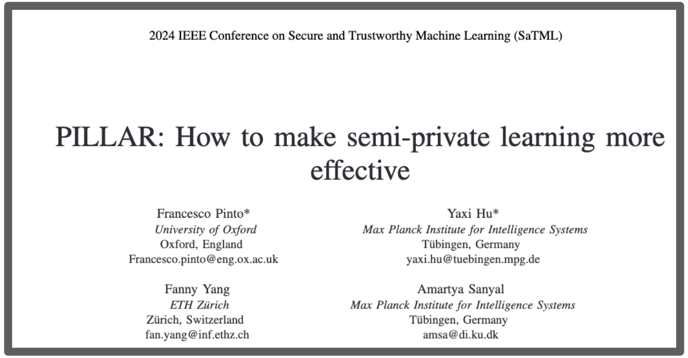
Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk



PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

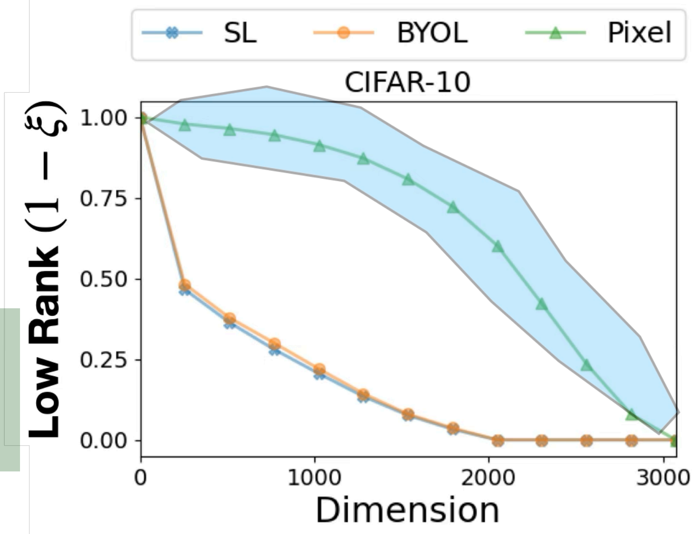
Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

- But natural data is not inherently low rank in pixel space
- Maybe we need to find the right representation space



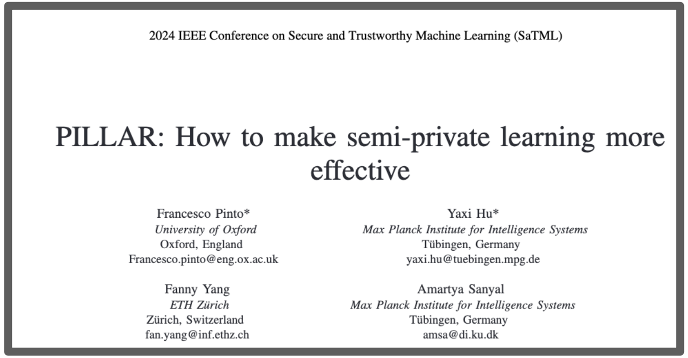
Idea 2: Use any **public unlabelled pre-training** data for representation learning.

$1 - \xi$ = low rank reconstruction error

PILLAR

Leveraging intrinsic low dimensionality

Idea 1: Use identically distributed **public unlabelled data** to find low rank subspace for projection


2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

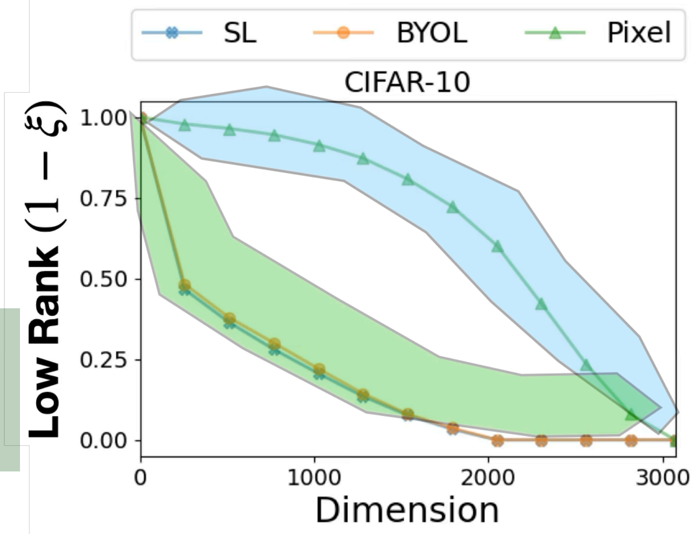
Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

- But natural data is not inherently low rank in pixel space
- Maybe we need to find the right representation space



Idea 2: Use any **public unlabelled pre-training** data for representation learning.

$1 - \xi$ = low rank reconstruction error

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality

2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality

2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

Public
unlabelled
pre-training

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality


2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

Public
unlabelled
pre-training

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality

2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

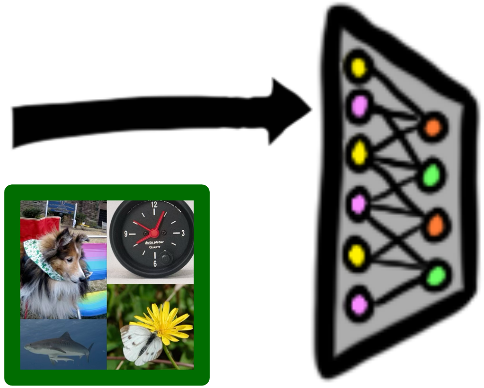
PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

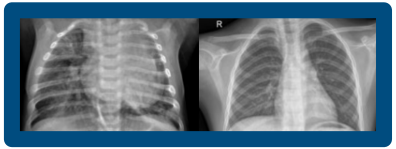


Public
unlabelled
pre-training

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality

Private labelled



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

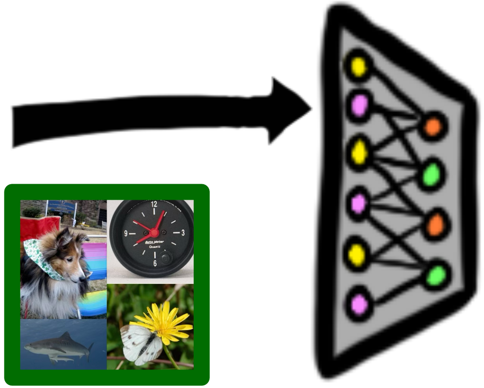
Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

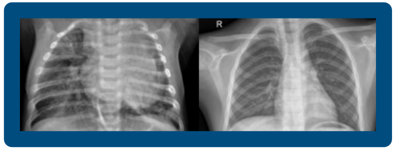
Public
unlabelled
pre-training

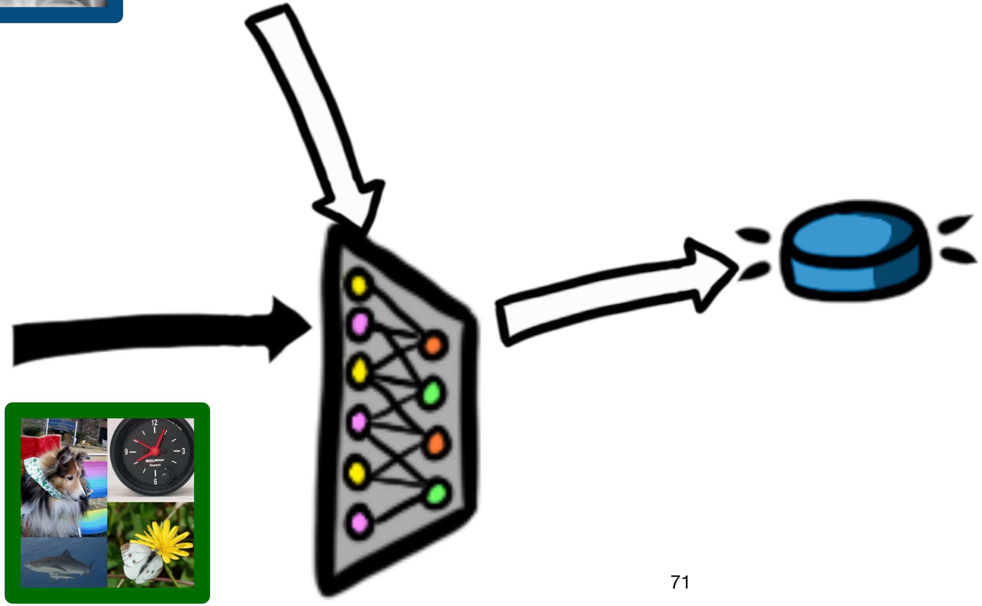


PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality

Private labelled





2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

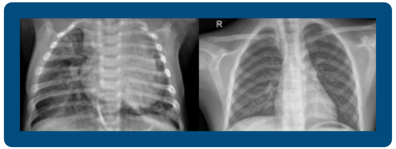
Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

Public unlabelled pre-training

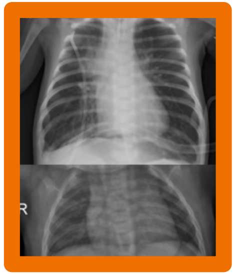
PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality

Private labelled



Public unlabelled



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

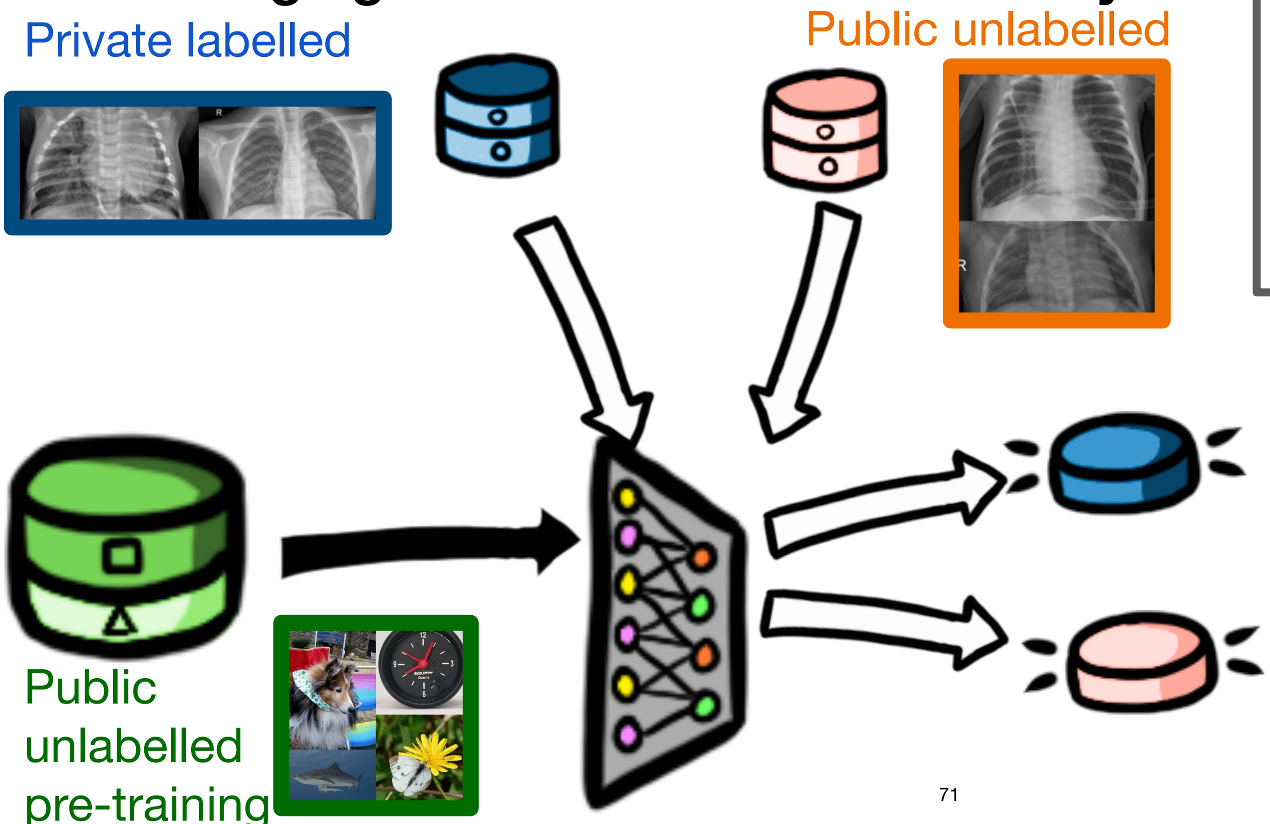
Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

Public unlabelled pre-training

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality




2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

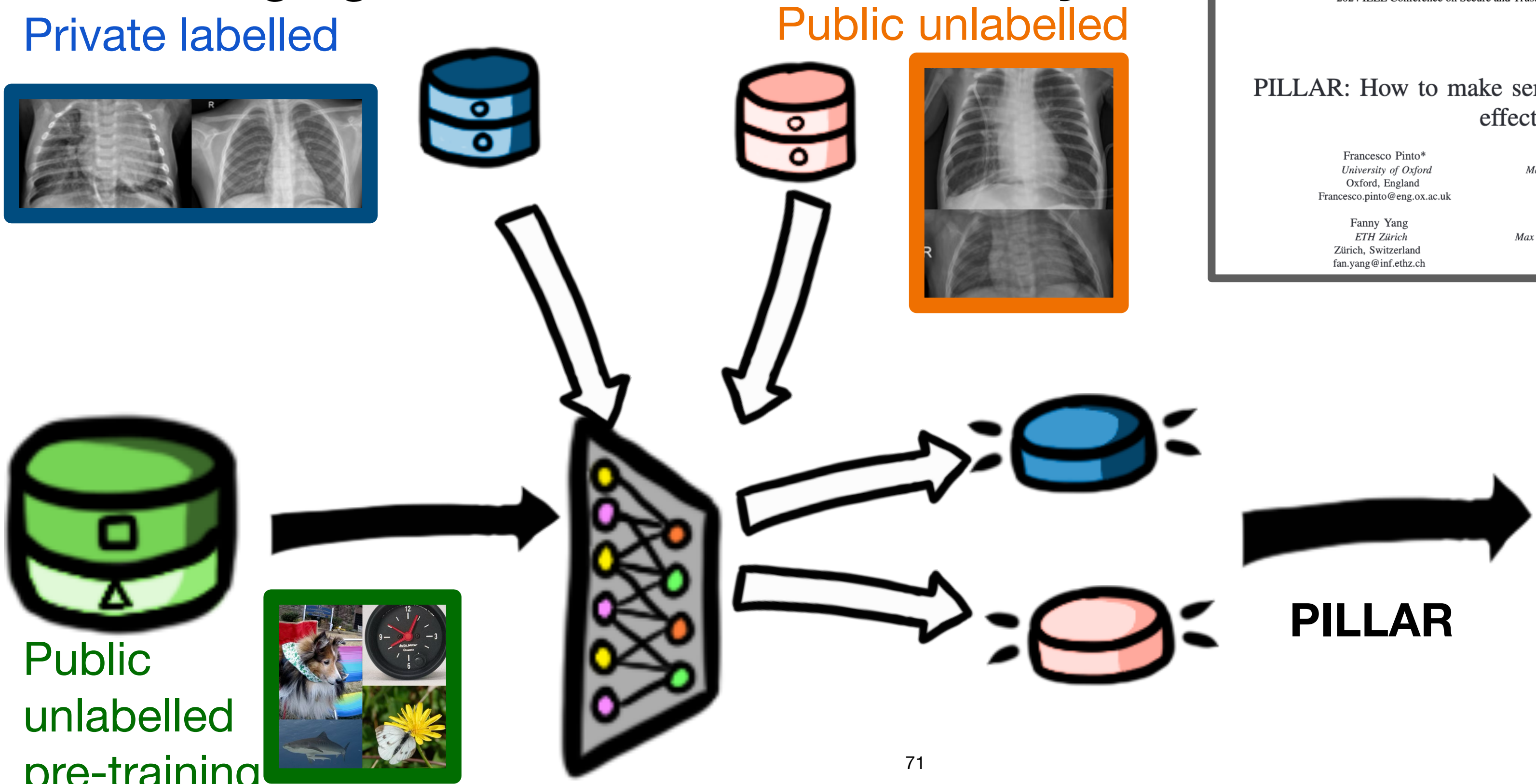
Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality




2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

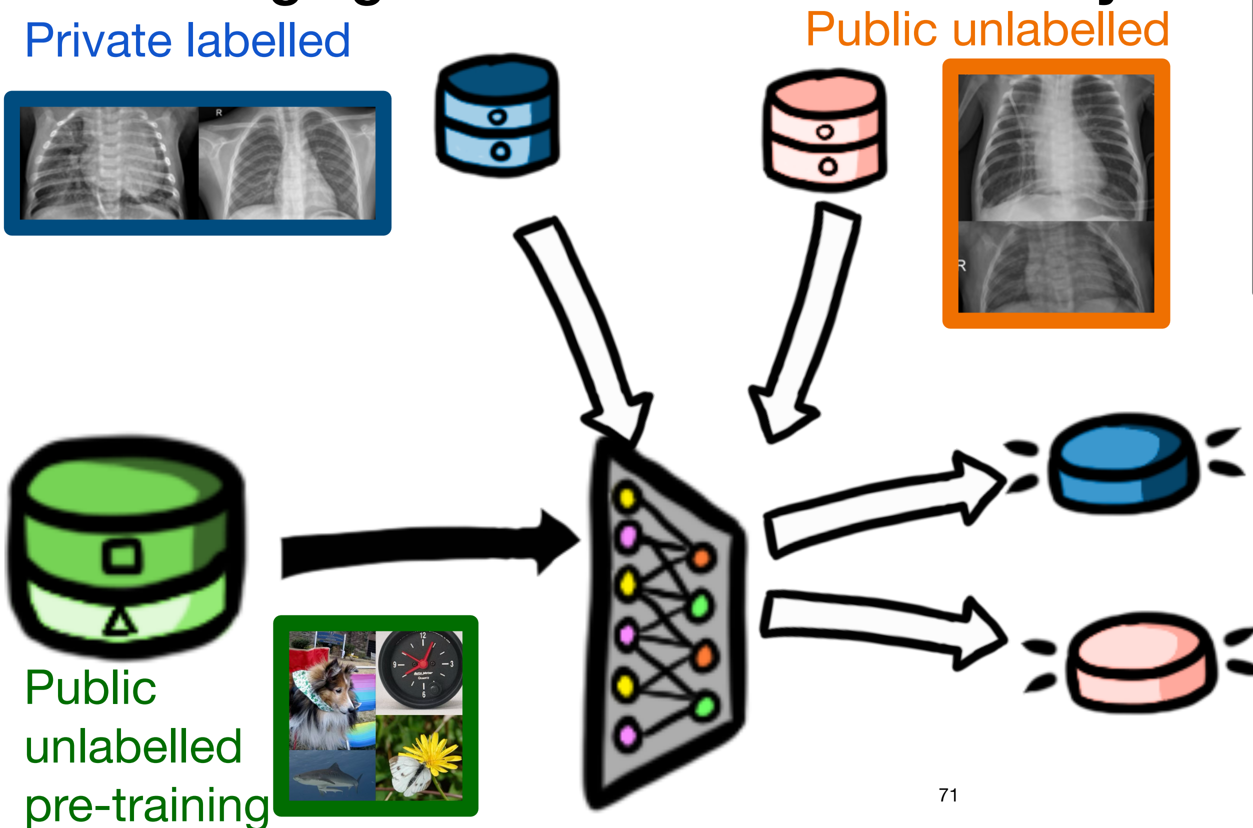
Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

PILLAR for Chest X-Ray Classification

Leveraging intrinsic low dimensionality



2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)

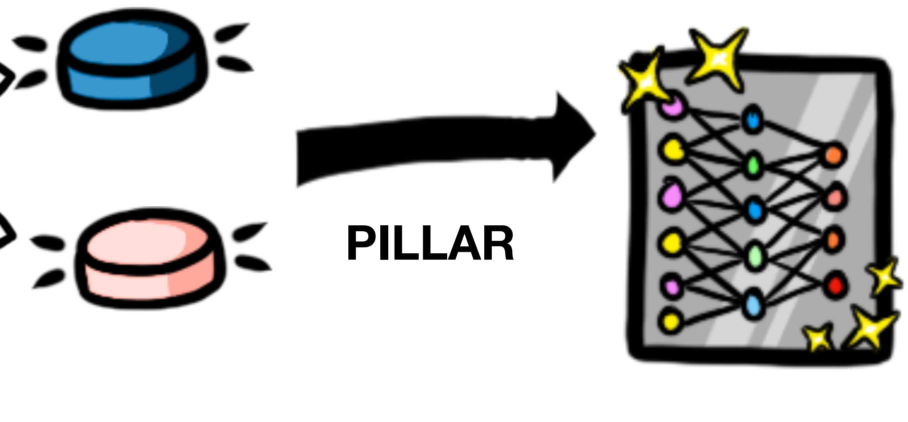
PILLAR: How to make semi-private learning more effective

Francesco Pinto*
University of Oxford
Oxford, England
Francesco.pinto@eng.ox.ac.uk

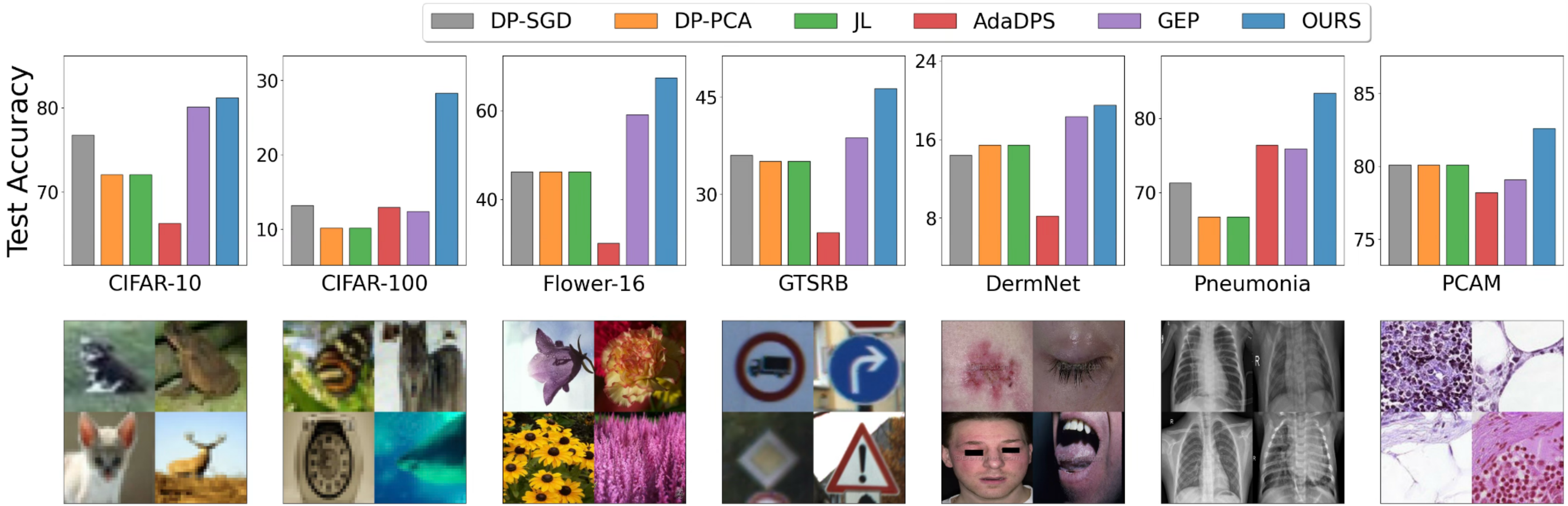
Yaxi Hu*
Max Planck Institute for Intelligence Systems
Tübingen, Germany
yaxi.hu@tuebingen.mpg.de

Fanny Yang
ETH Zürich
Zürich, Switzerland
fan.yang@inf.ethz.ch

Amartya Sanyal
Max Planck Institute for Intelligence Systems
Tübingen, Germany
amsa@di.ku.dk

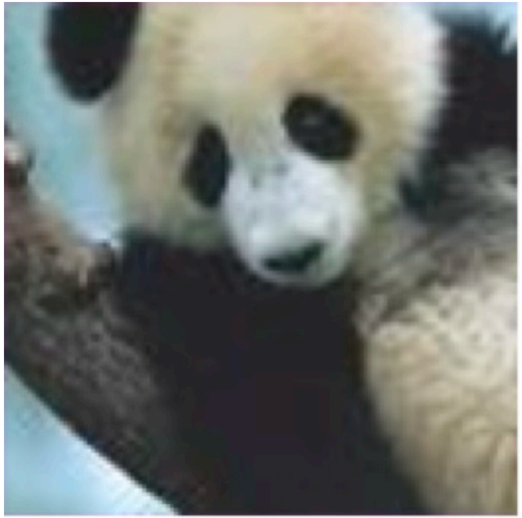


Other approaches to leverage unlabelled data



- GEP works in the gradient space
- AdaDPS use public data for gradient pre-conditioning

Next

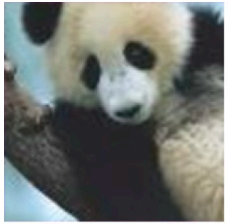


Robustness in Machine Learning

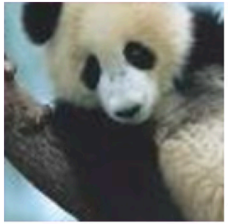
Adversarial Robustness in Machine Learning

Adversarial Robustness in Machine Learning

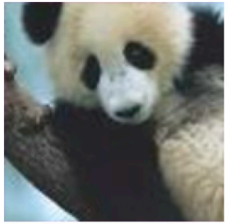
Adversarial Robustness in Machine Learning

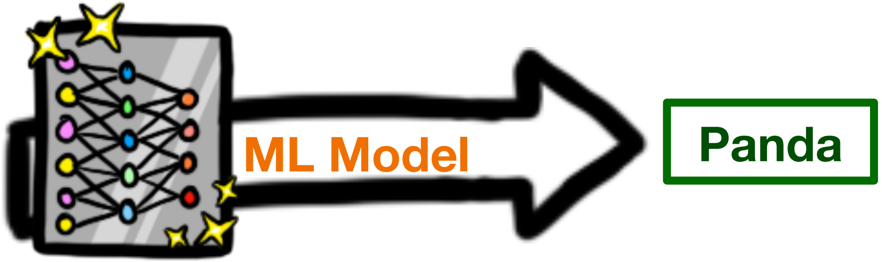


Adversarial Robustness in Machine Learning

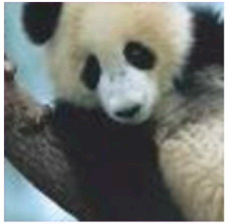


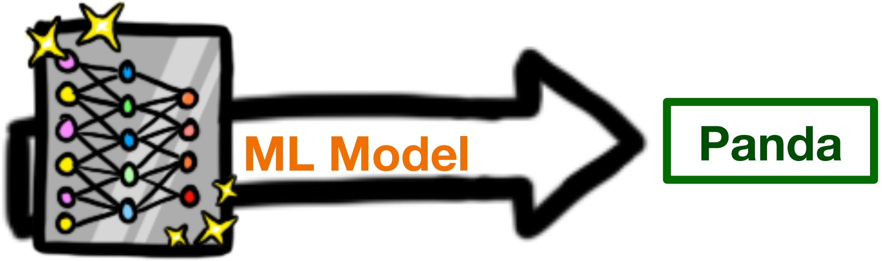
Adversarial Robustness in Machine Learning



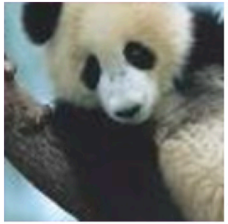


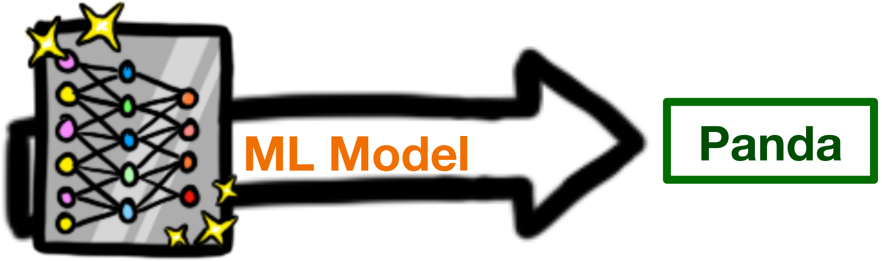
Adversarial Robustness in Machine Learning

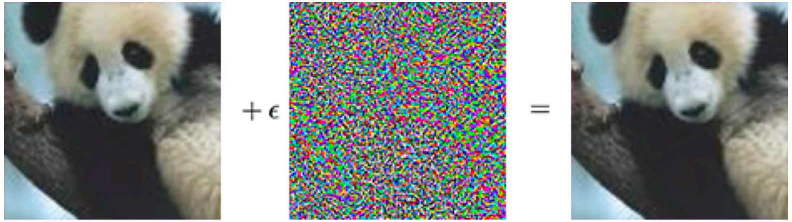




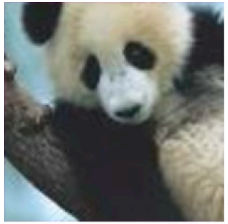
Adversarial Robustness in Machine Learning

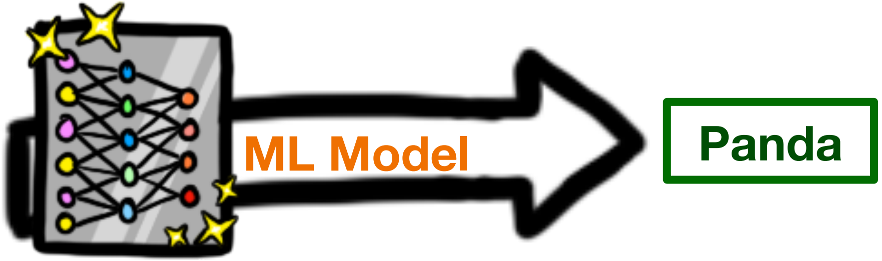


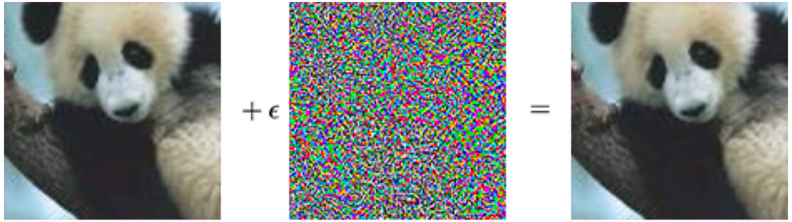


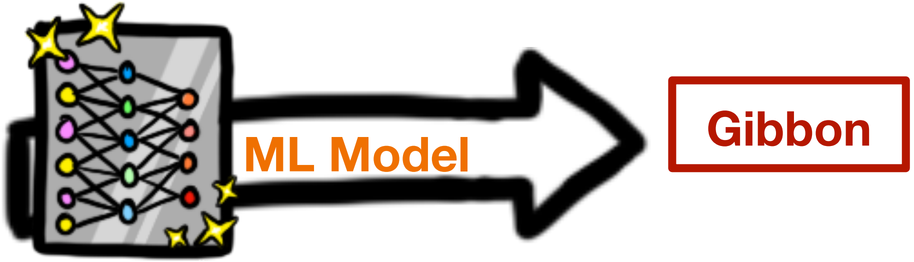


Adversarial Robustness in Machine Learning

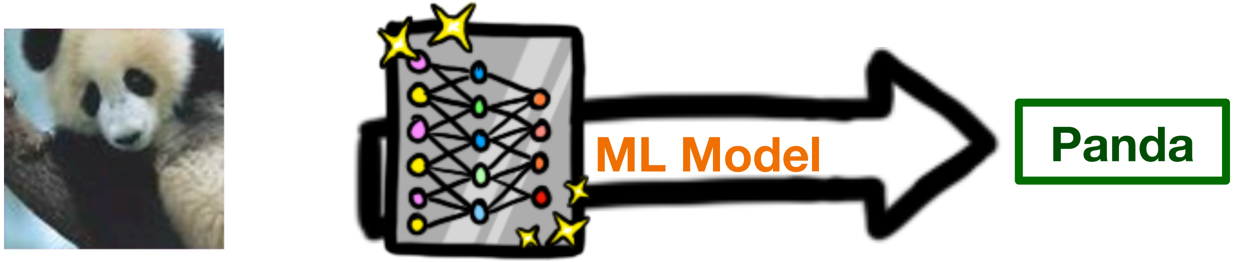




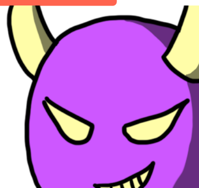




Adversarial Robustness in Machine Learning

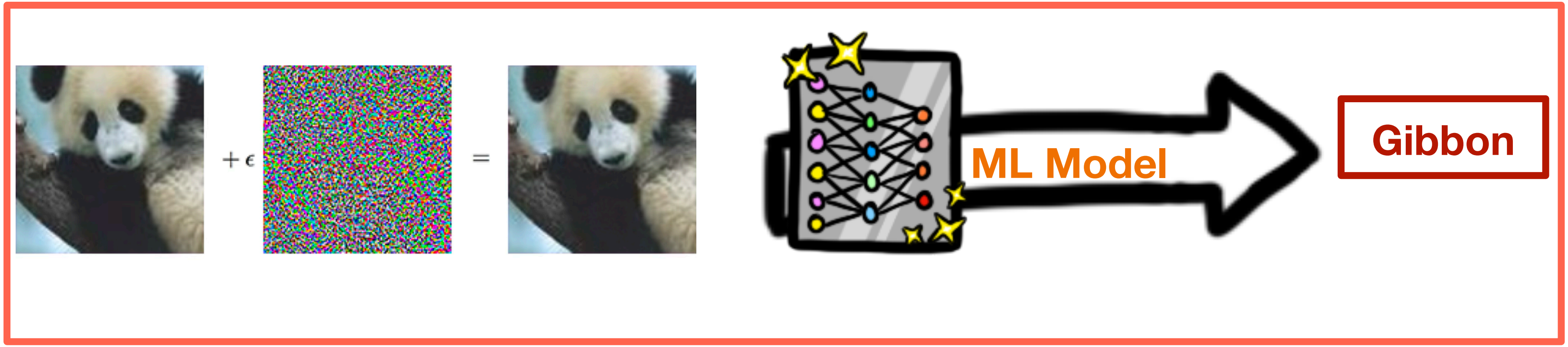


Adversarial Example



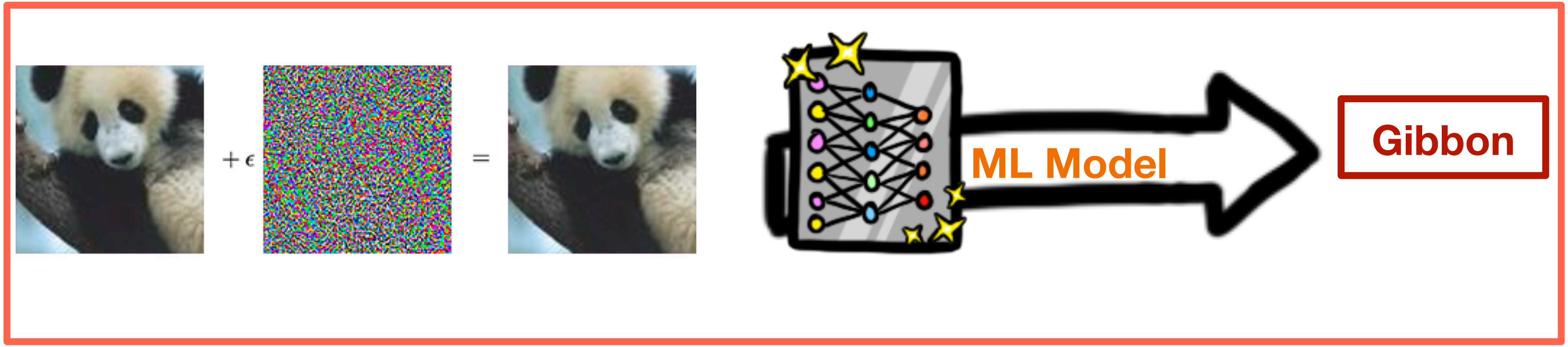
Adversarial Robustness in Machine Learning

Adversarial Robustness in Machine Learning



For any distribution $\mathcal{P}$ over $\mathbb{R}^d \times \{0, 1\}$ and any binary classifier $f : \mathbb{R}^d \to \{0, 1\}$

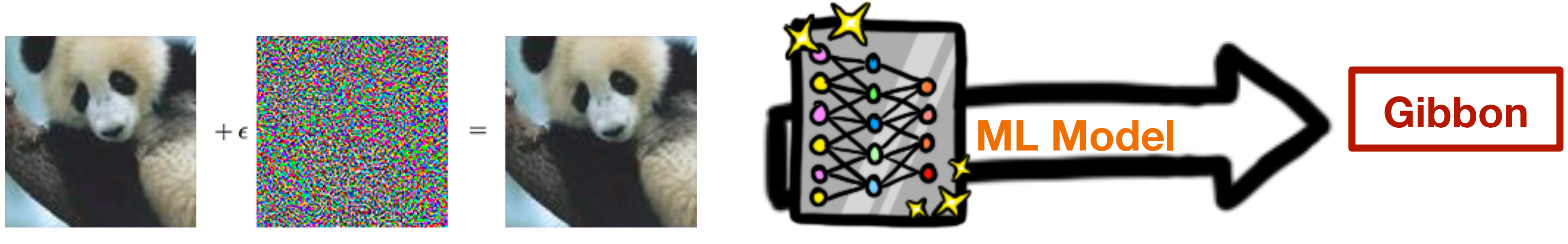
Adversarial Robustness in Machine Learning



For any distribution $\mathcal{P}$ over $\mathbb{R}^d \times \{0, 1\}$ and any binary classifier $f : \mathbb{R}^d \to \{0, 1\}$

the γ-adversarial error is defined as

Adversarial Robustness in Machine Learning



For any distribution $\mathcal{P}$ over $\mathbb{R}^d \times \{0, 1\}$ and any binary classifier $f : \mathbb{R}^d \to \{0, 1\}$

the γ-adversarial error is defined as

$$\Pr_{(x,y)\sim\mathbb{P}}[\text{exists } z \in \mathcal{B}_\gamma(x) : f(z) \neq y]$$

Adversarial Robustness in Machine Learning

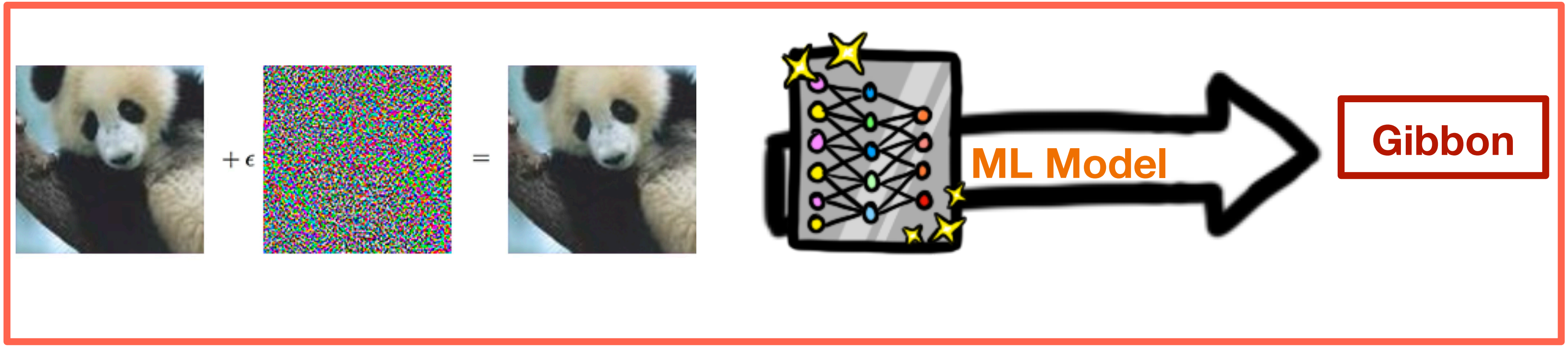
For any distribution $\mathcal{P}$ over $\mathbb{R}^d \times \{0, 1\}$ and any binary classifier $f : \mathbb{R}^d \to \{0, 1\}$

the γ-adversarial error is defined as

$$\Pr_{(x,y)\sim\mathbb{P}}[\text{exists } z \in \mathcal{B}_{\gamma}(x) : f(z) \neq y]$$

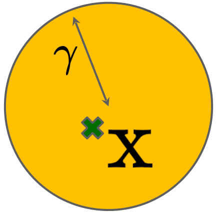
x

Adversarial Robustness in Machine Learning



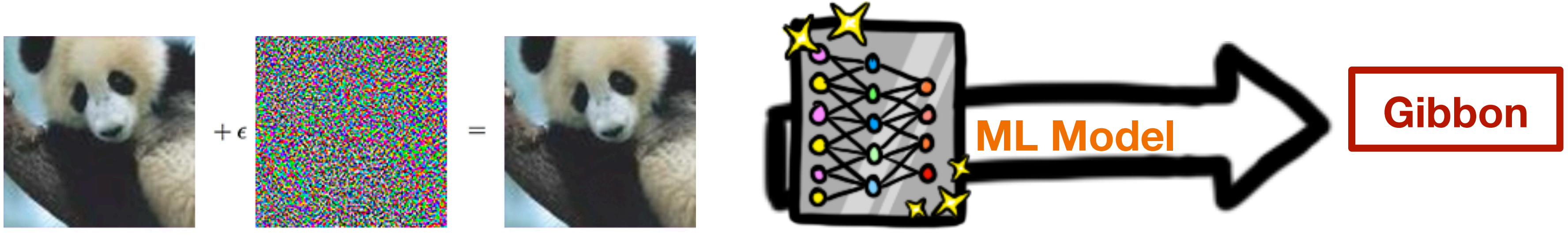
For any distribution $\mathcal{P}$ over $\mathbb{R}^d \times \{0, 1\}$ and any binary classifier $f : \mathbb{R}^d \to \{0, 1\}$

the γ-adversarial error is defined as



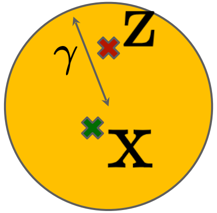
$$\Pr_{(x,y)\sim\mathbb{P}}[\text{exists } z \in \mathcal{B}_\gamma(x) : f(z) \neq y]$$

Adversarial Robustness in Machine Learning



For any distribution $\mathcal{P}$ over $\mathbb{R}^d \times \{0, 1\}$ and any binary classifier $f : \mathbb{R}^d \to \{0, 1\}$

the γ-adversarial error is defined as



$$\Pr_{(x,y)\sim\mathbb{P}}[\text{exists } z \in \mathcal{B}_\gamma(x) : f(z) \neq y]$$

Label Noise is ubiquitously interpolated

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

Learning from Noisy Labels with Deep Neural Networks: A Survey

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Label Noise is ubiquitously interpolated

- Trained long enough, NNs fit label noise

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

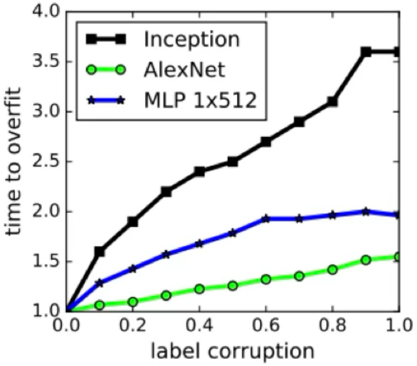
Chiyuan Zhang* Massachusetts Institute of Technology chiyuan@mit.edu
Samy Bengio Google Brain bengio@google.com
Moritz Hardt Google Brain mrtz@google.com
Benjamin Recht† University of California, Berkeley brecht@berkeley.edu
Oriol Vinyals Google DeepMind vinyals@google.com

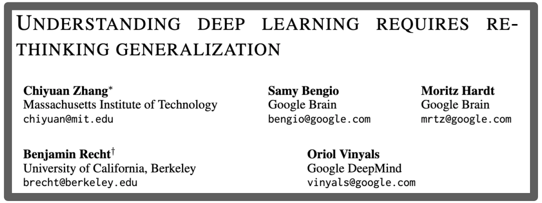
Learning from Noisy Labels with Deep Neural Networks: A Survey

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Label Noise is ubiquitously interpolated

- Trained long enough, NNs fit label noise





UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

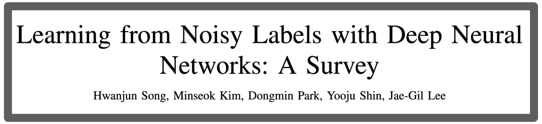
Chiyuan Zhang* Massachusetts Institute of Technology chiyuan@mit.edu

Samy Bengio Google Brain bengio@google.com

Moritz Hardt Google Brain mrtz@google.com

Benjamin Recht† University of California, Berkeley brecht@berkeley.edu

Oriol Vinyals Google DeepMind vinyals@google.com

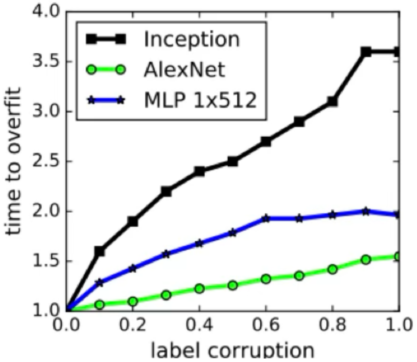


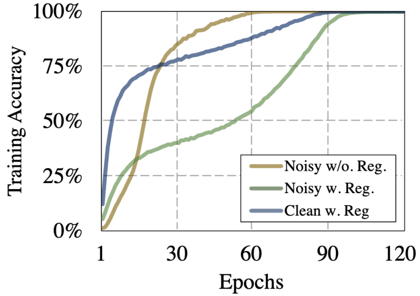
Learning from Noisy Labels with Deep Neural Networks: A Survey

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Label Noise is ubiquitously interpolated

- Trained long enough, NNs fit label noise





UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang* Massachusetts Institute of Technology chiyuan@mit.edu

Samy Bengio Google Brain bengio@google.com

Moritz Hardt Google Brain mrtz@google.com

Benjamin Recht† University of California, Berkeley brecht@berkeley.edu

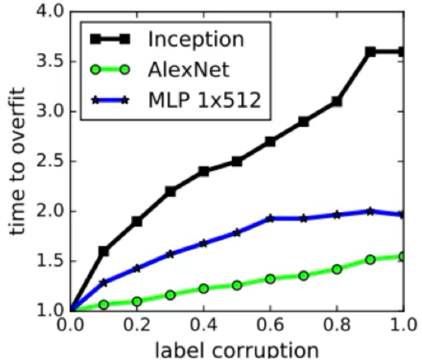
Oriol Vinyals Google DeepMind vinyals@google.com

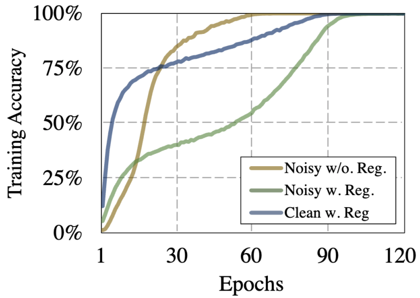
Learning from Noisy Labels with Deep Neural Networks: A Survey

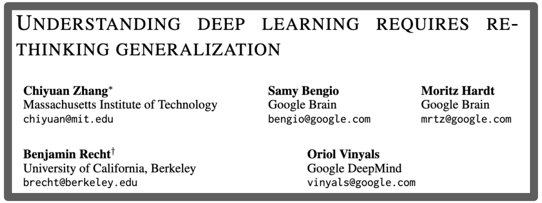
Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Label Noise is ubiquitously interpolated

- Trained long enough, NNs fit label noise
- Does not always hurt Test Accuracy - Benign Overfitting







UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

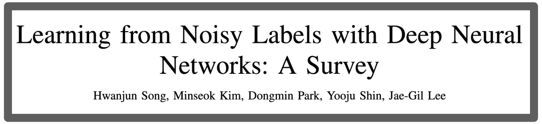
Chiyuan Zhang* Massachusetts Institute of Technology chiyuan@mit.edu

Samy Bengio Google Brain bengio@google.com

Moritz Hardt Google Brain mrtz@google.com

Benjamin Recht† University of California, Berkeley brecht@berkeley.edu

Oriol Vinyals Google DeepMind vinyals@google.com

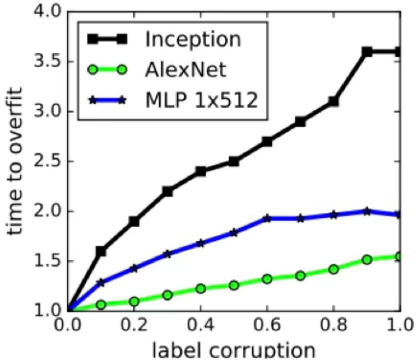


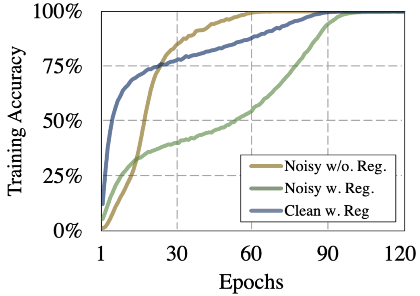
Learning from Noisy Labels with Deep Neural Networks: A Survey

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Label Noise is ubiquitously interpolated

- Trained long enough, NNs fit label noise
- Does not always hurt Test Accuracy - Benign Overfitting
- Define a model with 100% training acc: **Interpolator**





UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang* Massachusetts Institute of Technology chiyuan@mit.edu

Samy Bengio Google Brain bengio@google.com

Moritz Hardt Google Brain mrtz@google.com

Benjamin Recht† University of California, Berkeley brecht@berkeley.edu

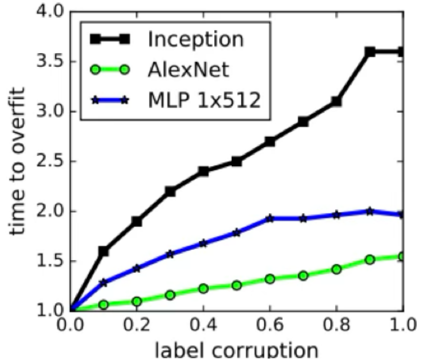
Oriol Vinyals Google DeepMind vinyals@google.com

Learning from Noisy Labels with Deep Neural Networks: A Survey

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Label Noise is ubiquitously interpolated

- Trained long enough, NNs fit label noise
- Does not always hurt Test Accuracy - Benign Overfitting
- Define a model with 100% training acc: **Interpolator**



UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

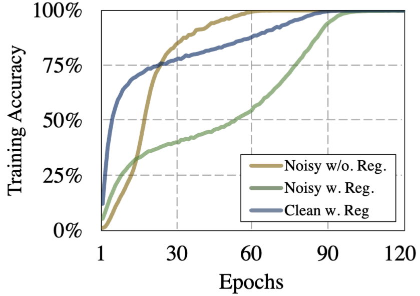
Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com



Learning from Noisy Labels with Deep Neural Networks: A Survey

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, Jae-Gil Lee

Question: What about **Robust Accuracy** ?

Lower bound on Adversarial error

HOW BENIGN IS BENIGN OVERFITTING?

Amartya Sanyal
Department of Computer Science,
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
amartya.sanyal@cs.ox.ac.uk

Varun Kanade
Department of Computer Science
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
varunk@cs.ox.ac.uk

Puneet K.Dokania
Department of Engineering Science
University of Oxford, Oxford, UK
Five AI Limited
puneet@robots.ox.ac.uk

Philip H.S. Torr
Department of Engineering Science
University of Oxford, Oxford, UK
phst@robots.ox.ac.uk

A LAW OF ADVERSARIAL RISK, INTERPOLATION, AND LABEL NOISE

Daniel Paleka *
ETH Zurich
daniel.paleka@inf.ethz.ch

Amartya Sanyal *
ETH AI Center, ETH Zurich
amartya.sanyal@ai.ethz.ch

Lower bound on Adversarial error

Let

- μ be any distribution on $\mathbb{R}^d$,
- $\eta \in (0, 1)$ be the uniform label noise rate,
- $\mathcal{C} \subset \mathbb{R}^d$ be any region, and
- $N(\mathcal{C}, \epsilon, \|\cdot\|)$ is the covering number of $\mathcal{C}$

HOW BENIGN IS BENIGN OVERFITTING?

Amartya Sanyal
Department of Computer Science,
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
amartya.sanyal@cs.ox.ac.uk

Varun Kanade
Department of Computer Science
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
varunk@cs.ox.ac.uk

Puneet K.Dokania
Department of Engineering Science
University of Oxford, Oxford, UK
Five AI Limited
puneet@robots.ox.ac.uk

Philip H.S. Torr
Department of Engineering Science
University of Oxford, Oxford, UK
phst@robots.ox.ac.uk

A LAW OF ADVERSARIAL RISK, INTERPOLATION, AND LABEL NOISE

Daniel Paleka *
ETH Zurich
daniel.paleka@inf.ethz.ch

Amartya Sanyal *
ETH AI Center, ETH Zurich
amartya.sanyal@ai.ethz.ch

Lower bound on Adversarial error

Let

- μ be any distribution on $\mathbb{R}^d$,
- $\eta \in (0, 1)$ be the uniform label noise rate,
- $\mathcal{C} \subset \mathbb{R}^d$ be any region, and
- $N(\mathcal{C}, \epsilon, \|\cdot\|)$ is the covering number of $\mathcal{C}$

HOW BENIGN IS BENIGN OVERFITTING?

Amartya Sanyal
Department of Computer Science,
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
amartya.sanyal@cs.ox.ac.uk

Varun Kanade
Department of Computer Science
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
varunk@cs.ox.ac.uk

Puneet K.Dokania
Department of Engineering Science
University of Oxford, Oxford, UK
Five AI Limited
puneet@robots.ox.ac.uk

Philip H.S. Torr
Department of Engineering Science
University of Oxford, Oxford, UK
phst@robots.ox.ac.uk

A LAW OF ADVERSARIAL RISK, INTERPOLATION, AND LABEL NOISE

Daniel Paleka *
ETH Zurich
daniel.paleka@inf.ethz.ch

Amartya Sanyal *
ETH AI Center, ETH Zurich
amartya.sanyal@ai.ethz.ch

Theorem If the noisy dataset size $m = \Omega\left(\dfrac{N(\mathcal{C}, \epsilon, \|\cdot\|)}{\mu(\mathcal{C})\eta}\right)$, for all interpolators h

Lower bound on Adversarial error

HOW BENIGN IS BENIGN OVERFITTING?

Amartya Sanyal
Department of Computer Science,
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
amartya.sanyal@cs.ox.ac.uk

Varun Kanade
Department of Computer Science
University of Oxford,
Oxford, UK
The Alan Turing Institute, London, UK
varunk@cs.ox.ac.uk

Puneet K.Dokania
Department of Engineering Science
University of Oxford, Oxford, UK
Five AI Limited
puneet@robots.ox.ac.uk

Philip H.S. Torr
Department of Engineering Science
University of Oxford, Oxford, UK
phst@robots.ox.ac.uk

A LAW OF ADVERSARIAL RISK, INTERPOLATION, AND LABEL NOISE

Daniel Paleka *
ETH Zurich
daniel.paleka@inf.ethz.ch

Amartya Sanyal *
ETH AI Center, ETH Zurich
amartya.sanyal@ai.ethz.ch

Let

- μ be any distribution on $\mathbb{R}^d$,
- $\eta \in (0,1)$ be the uniform label noise rate,
- $\mathcal{C} \subset \mathbb{R}^d$ be any region, and
- $N(\mathcal{C}, \epsilon, \|\cdot\|)$ is the covering number of $\mathcal{C}$

Theorem If the noisy dataset size $m = \Omega\left(\frac{N(\mathcal{C}, \epsilon, \|\cdot\|)}{\mu(\mathcal{C})\eta}\right)$, for all interpolators h

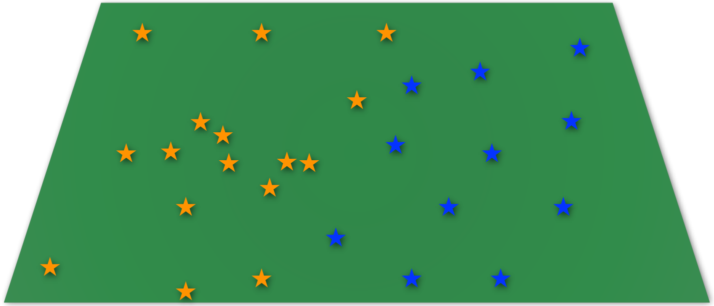
$$\text{Adv. Error}_\epsilon(\text{h}) \geq \mu(\mathcal{C})$$

Lower bound on Adversarial error

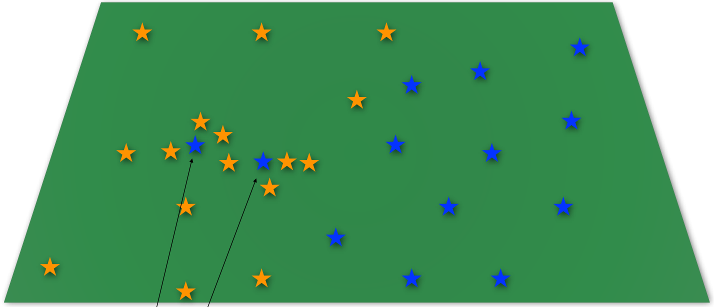
Lower bound on Adversarial error

Lower bound on Adversarial error

Lower bound on Adversarial error

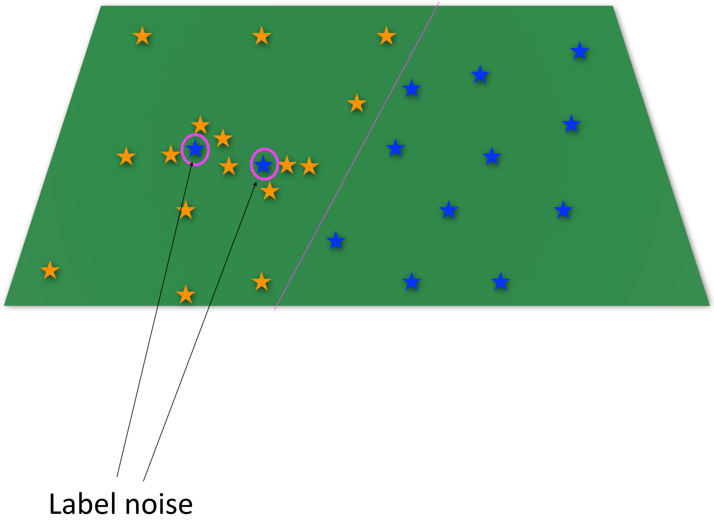


Lower bound on Adversarial error



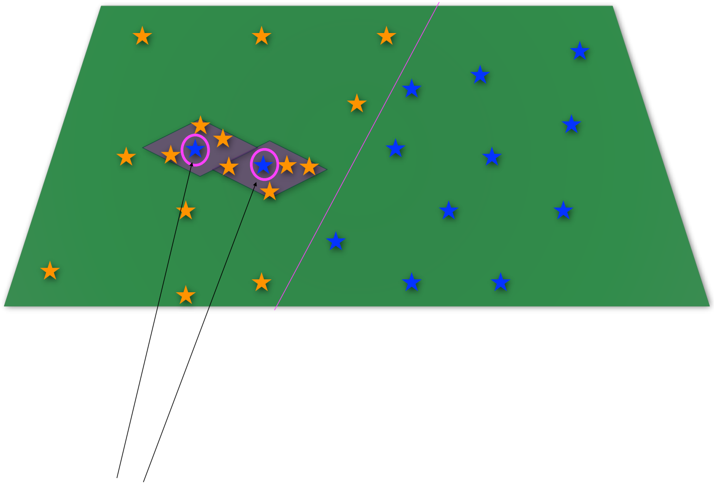
Label noise

Lower bound on Adversarial error

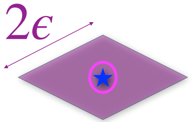


h is an interpolator e.g. Random Forest, 1-NN, NNs

Lower bound on Adversarial error



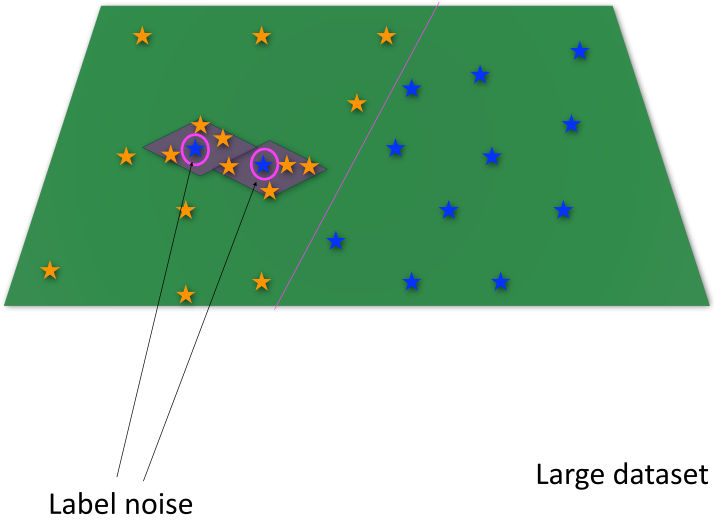
h is an interpolator e.g. Random Forest, 1-NN, NNs



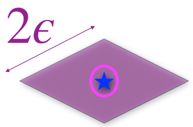
Region of Adversarial Vulnerability

Label noise

Lower bound on Adversarial error



h is an interpolator e.g. Random Forest, 1-NN, NNs



Region of Adversarial Vulnerability

Large dataset $\Longrightarrow$ More noisy labels $\Longrightarrow$ Large Adversarial Error

Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu

Aleksandar Makelov* MIT amakelov@mit.edu

Ludwig Schmidt* MIT ludwigs@mit.edu

Dimitris Tsipras* MIT tsipras@mit.edu

Adrian Vladu* MIT avladu@mit.edu

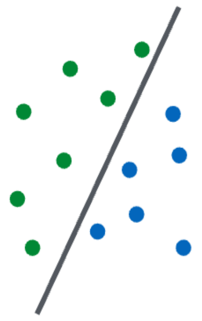
Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.

Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.

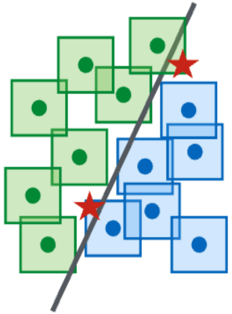


Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.

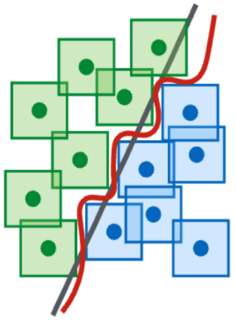


Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.



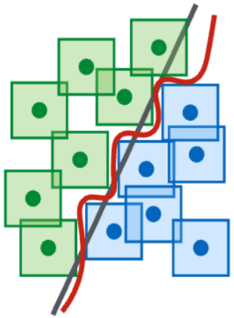
Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.

Naturally, complex models can fit the augmented data better.



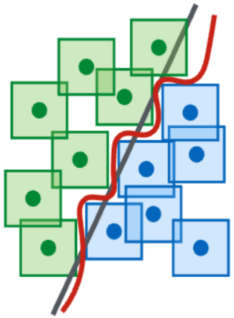
Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial training

Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.

Naturally, complex models can fit the augmented data better.



Robust overfitting is when train robust error decreases but test robust error increases.

Overfitting in adversarially robust deep learning

Leslie Rice [*1] Eric Wong [*2] J. Zico Kolter [1]

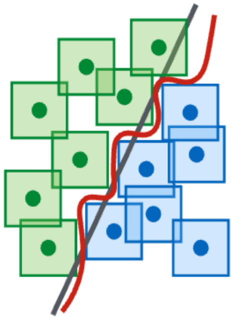
Adversarial training

Towards Deep Learning Models Resistant to Adversarial Attacks

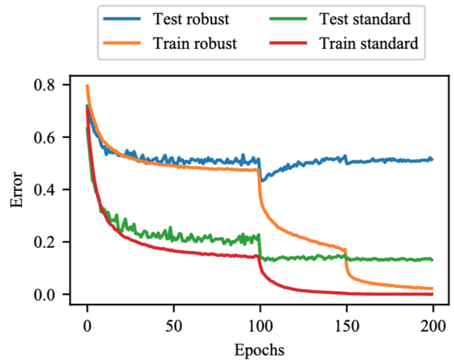
Aleksander Mądry* MIT madry@mit.edu
Aleksandar Makelov* MIT amakelov@mit.edu
Ludwig Schmidt* MIT ludwigs@mit.edu
Dimitris Tsipras* MIT tsipras@mit.edu
Adrian Vladu* MIT avladu@mit.edu

Adversarial Training replaces (or augments) clean data with corresponding adversarial examples during SGD.

Naturally, complex models can fit the augmented data better.



Robust overfitting is when train robust error decreases but test robust error increases.



Overfitting in adversarially robust deep learning

Leslie Rice [*1] Eric Wong [*2] J. Zico Kolter [1]

Robust Overfitting and label noise

Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting

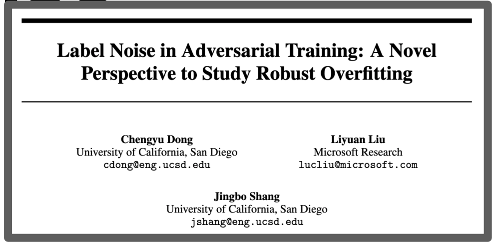
Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

Liyuan Liu
Microsoft Research
lucliu@microsoft.com

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu

Robust Overfitting and label noise

- One of explanations given for Robust overfitting is that adversarial training implicitly adds label noise.



Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting

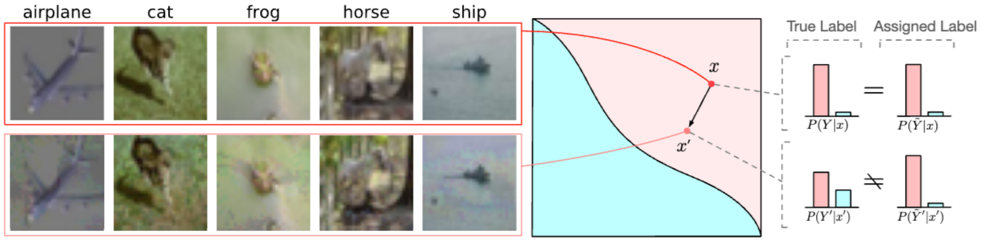
Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

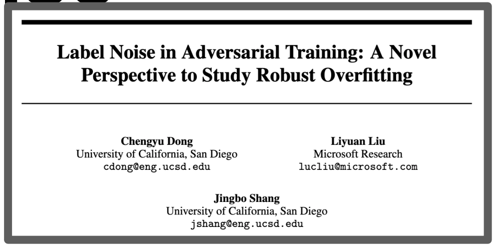
Liyuan Liu
Microsoft Research
lucliu@microsoft.com

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu

Robust Overfitting and label noise

- One of explanations given for Robust overfitting is that adversarial training implicitly adds label noise.





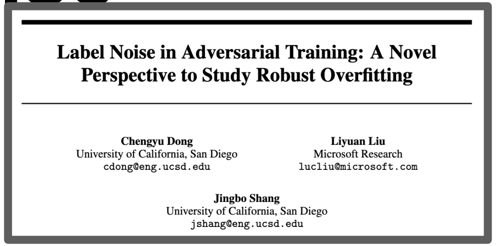
Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting

Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

Liyuan Liu
Microsoft Research
lucliu@microsoft.com

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu

Robust Overfitting and label noise



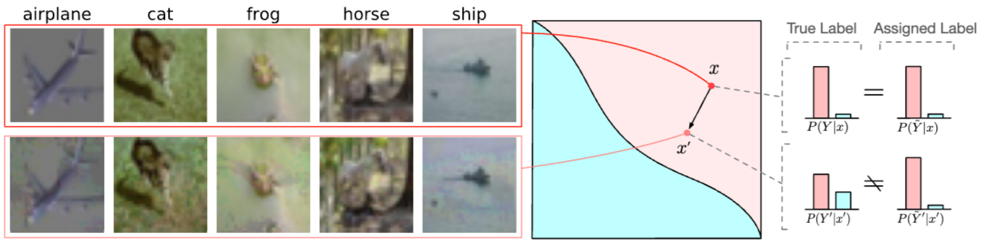
Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting

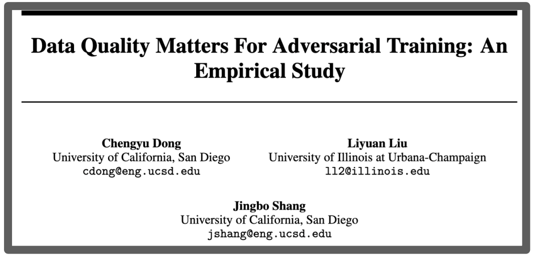
Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

Liyuan Liu
Microsoft Research
lucliu@microsoft.com

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu

- One of explanations given for Robust overfitting is that adversarial training implicitly adds label noise.





Data Quality Matters For Adversarial Training: An Empirical Study

Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

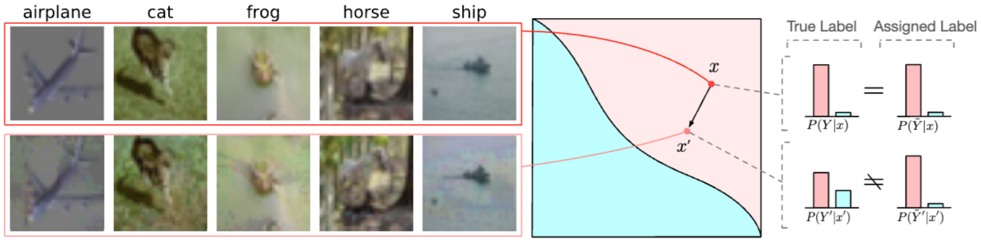
Liyuan Liu
University of Illinois at Urbana-Champaign
ll2@illinois.edu

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu

- Simply using “good” examples that are far from the decision boundary alleviates parts of the issue

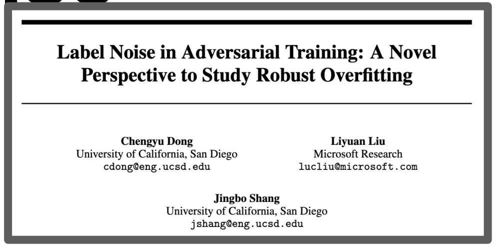
Robust Overfitting and label noise

- One of explanations given for Robust overfitting is that adversarial training implicitly adds label noise.



- Simply using “good” examples that are far from the decision boundary alleviates parts of the issue

- Larger perturbation radius causes more overfitting

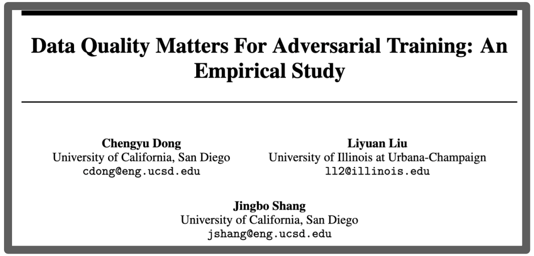


Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting

Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

Liyuan Liu
Microsoft Research
lucliu@microsoft.com

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu

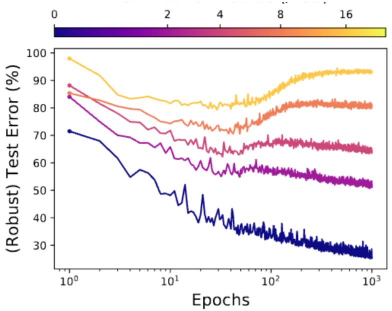


Data Quality Matters For Adversarial Training: An Empirical Study

Chengyu Dong
University of California, San Diego
cdong@eng.ucsd.edu

Liyuan Liu
University of Illinois at Urbana-Champaign
ll2@illinois.edu

Jingbo Shang
University of California, San Diego
jshang@eng.ucsd.edu



Improving Robust generalisation

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Țifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]

[1] *ETH Zurich*, [2] *Rice University*, [3] *Technical University of Munich*

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Ţifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]

[1] *ETH Zurich*, [2] *Rice University*, [3] *Technical University of Munich*

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

- Clearly, more data helps to avoid robust overfitting

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Ţifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]

[1] ETH Zurich, [2] Rice University, [3] Technical University of Munich

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

- Clearly, more data helps to avoid robust overfitting
- Regularisation and early-stopping also helps.

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT; Shibani Santurkar, MIT; Dimitris Tsipras, MIT; Kunal Talwar, Google Brain; Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Țifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]
[1] *ETH Zurich*, [2] *Rice University*, [3] *Technical University of Munich*

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

- Clearly, more data helps to avoid robust overfitting
- Regularisation and early-stopping also helps.

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Ţifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]

[1] *ETH Zurich*, [2] *Rice University*, [3] *Technical University of Munich*

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

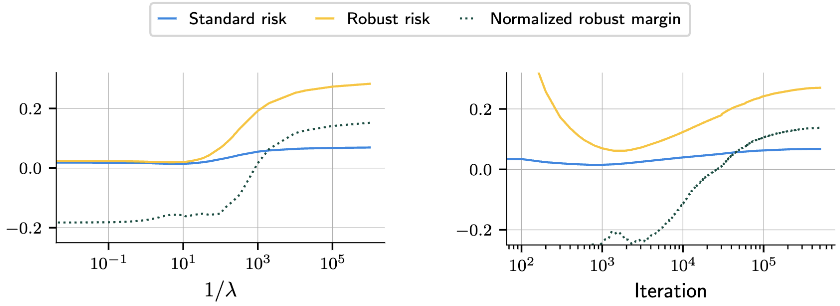
- Clearly, more data helps to avoid robust overfitting
- Regularisation and early-stopping also helps.

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Ţifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]
[1]ETH Zurich, [2]Rice University, [3]Technical University of Munich



(a) Benefit of ridge regularization

(b) Benefit of early stopping

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

- Clearly, more data helps to avoid robust overfitting
- Regularisation and early-stopping also helps.

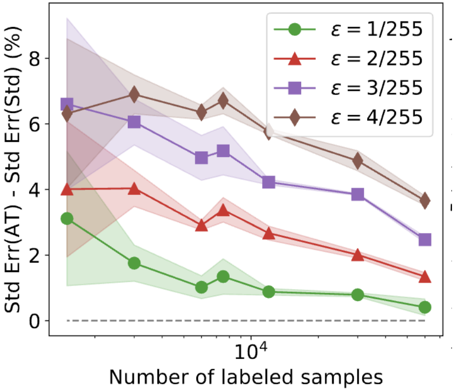
Adversarially Robust Generalization Requires More Data

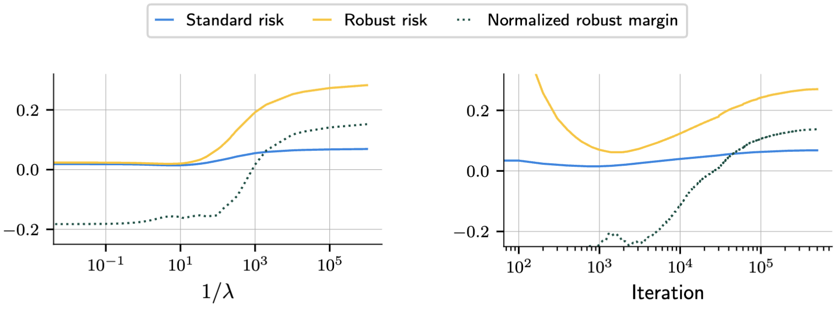
Ludwig Schmidt, MIT; Shibani Santurkar, MIT; Dimitris Tsipras, MIT; Kunal Talwar, Google Brain; Aleksander Mądry, MIT

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Ţifrea*[1], Michael Aerni[1], Reinhard Heckel[2,3] and Fanny Yang[1]

[1] *ETH Zurich*, [2] *Rice University*, [3] *Technical University of Munich*





(a) Benefit of ridge regularization

(b) Benefit of early stopping

Improving Robust generalisation

Exists simple distribution in d dim where robust generalisation requires $\sqrt{d}$ times more data.

Adversarially Robust Generalization Requires More Data

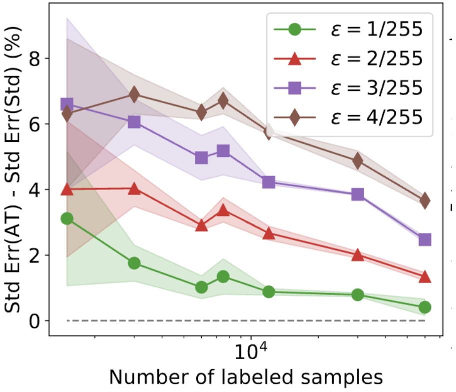
Ludwig Schmidt, MIT
Shibani Santurkar, MIT
Dimitris Tsipras, MIT
Kunal Talwar, Google Brain
Aleksander Mądry, MIT

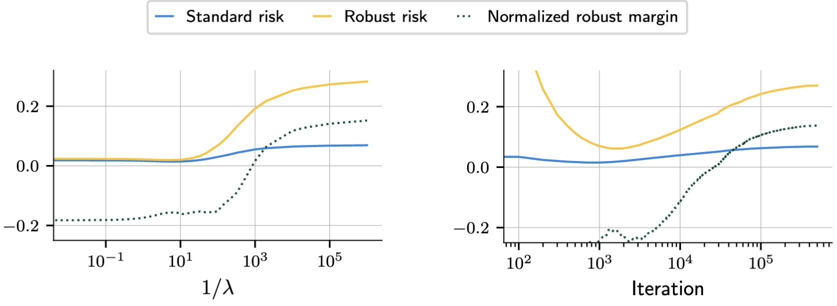
- Clearly, more data helps to avoid robust overfitting
- Regularisation and early-stopping also helps.

Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*[1], Alexandru Ţifrea*[1], Michael Aerni[1]
Reinhard Heckel[2,3] and Fanny Yang[1]

[1] ETH Zurich, [2] Rice University, [3] Technical University of Munich





(a) Benefit of ridge regularization

(b) Benefit of early stopping

What about unlabelled data ?

Improving Robust generalisation

With unlabelled data

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1,*], Tianle Cai[1*], Di He[1*], Chen Dan[2], Kun He[4], John E. Hopcroft[3] & Liwei Wang[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon*, Stanford University, yairc@stanford.edu
Aditi Raghunathan*, Stanford University, aditir@stanford.edu
Ludwig Schmidt, UC Berkeley, ludwig@berkeley.edu
Percy Liang, Stanford University, pliang@cs.stanford.edu
John C. Duchi, Stanford University, jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

Observation: Robust error can be decomposed into

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, **Tianle Cai**[1*], **Di He**[1*], **Chen Dan**[2], **Kun He**[4], **John E. Hopcroft**[3] **& Liwei Wang**[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

Observation: Robust error can be decomposed into

1. **Stability error:** Whether prediction is stable in a ball around data from the test distribution

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, Tianle Cai[1]*, Di He[1]*, Chen Dan[2], Kun He[4], John E. Hopcroft[3] & Liwei Wang[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

Observation: Robust error can be decomposed into

1. **Stability error:** Whether prediction is stable in a ball around data from the test distribution

2. **Classification accuracy:** Whether classification in the original data distribution is accurate

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, Tianle Cai[1]*, Di He[1]*, Chen Dan[2], Kun He[4], John E. Hopcroft[3] & Liwei Wang[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

Observation: Robust error can be decomposed into

1. **Stability error:** Whether prediction is stable in a ball around data from the test distribution
2. **Classification accuracy:** Whether classification in the original data distribution is accurate

Classical use of unlabelled data improves **2. Classification accuracy.**

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, **Tianle Cai**[1]*, **Di He**[1]*, **Chen Dan**[2], **Kun He**[4], **John E. Hopcroft**[3] **& Liwei Wang**[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon*, Stanford University, yairc@stanford.edu
Aditi Raghunathan*, Stanford University, aditir@stanford.edu
Ludwig Schmidt, UC Berkeley, ludwig@berkeley.edu
Percy Liang, Stanford University, pliang@cs.stanford.edu
John C. Duchi, Stanford University, jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

Observation: Robust error can be decomposed into

1. **Stability error:** Whether prediction is stable in a ball around data from the test distribution

2. **Classification accuracy:** Whether classification in the original data distribution is accurate

Classical use of unlabelled data improves **2. Classification accuracy.**

To improve robustness, use unlabelled data to improve **1. Stability error.**

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, Tianle Cai[1*], Di He[1*], Chen Dan[2], Kun He[4], John E. Hopcroft[3] & Liwei Wang[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, **Tianle Cai**[1]*, **Di He**[1]*, **Chen Dan**[2], **Kun He**[4], **John E. Hopcroft**[3] **& Liwei Wang**[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu

Observation: Robust error can be decomposed into

1. **Stability error:** Whether prediction is stable in a ball around data from the test distribution

2. **Classification accuracy:** Whether classification in the original data distribution is accurate

Classical use of unlabelled data improves **2. Classification accuracy.**

To improve robustness, use unlabelled data to improve **1. Stability error.**

Recipe: Use adversarial training on pseudo-labels on the unlabelled data

Improving Robust generalisation

With unlabelled data

Improving Robust generalisation

With unlabelled data

Method	Robust Test Acc.	Standard Test Acc.	
Standard Training	0.8%	95.2%	Vanilla Supervised
PG-AT (Madry et al., 2018)	45.8%	87.3%	
TRADES (Zhang et al., 2019)	55.4%	84.0%	
Standard Self-Training	0.3%	96.4%	Semisupervised with same unlabeled data
Robust Consistency Training (Carmon et al., 2019)	56.5%	83.2%	
RST + PG-AT (this paper)	**58.5%**	**91.8%**	
RST + TRADES (this paper) (Carmon et al., 2019)	**63.1%**	**89.7%**	

Understanding and Mitigating the Tradeoff Between Robustness and Accuracy

Aditi Raghunathan [*1] **Sang Michael Xie** [*1] **Fanny Yang** [2] **John C. Duchi** [1] **Percy Liang** [1]

Improving Robust generalisation

With unlabelled data

Method	Robust Test Acc.	Standard Test Acc.	
Standard Training	0.8%	95.2%	Vanilla Supervised
PG-AT (Madry et al., 2018)	45.8%	87.3%	
TRADES (Zhang et al., 2019)	55.4%	84.0%	
Standard Self-Training	0.3%	96.4%	Semisupervised with same unlabeled data
Robust Consistency Training (Carmon et al., 2019)	56.5%	83.2%	
RST + PG-AT (this paper)	**58.5%**	**91.8%**	
RST + TRADES (this paper) (Carmon et al., 2019)	**63.1%**	**89.7%**	

Understanding and Mitigating the Tradeoff Between Robustness and Accuracy

Aditi Raghunathan [*1] **Sang Michael Xie** [*1] **Fanny Yang** [2] **John C. Duchi** [1] **Percy Liang** [1]

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1*], **Tianle Cai**[1*], **Di He**[1*], **Chen Dan**[2],
Kun He[4], **John E. Hopcroft**[3] **& Liwei Wang**[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Improving Robust generalisation

With unlabelled data

Method	Robust Test Acc.	Standard Test Acc.	
Standard Training	0.8%	95.2%	Vanilla Supervised
PG-AT (Madry et al., 2018)	45.8%	87.3%	
TRADES (Zhang et al., 2019)	55.4%	84.0%	
Standard Self-Training	0.3%	96.4%	Semisupervised with same unlabeled data
Robust Consistency Training (Carmon et al., 2019)	56.5%	83.2%	
RST + PG-AT (this paper)	**58.5%**	**91.8%**	
RST + TRADES (this paper) (Carmon et al., 2019)	**63.1%**	**89.7%**	

Understanding and Mitigating the Tradeoff Between Robustness and Accuracy

Aditi Raghunathan * 1 **Sang Michael Xie** * 1 **Fanny Yang** 2 **John C. Duchi** 1 **Percy Liang** 1

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1]*, **Tianle Cai**[1*], **Di He**[1*], **Chen Dan**[2], **Kun He**[4], **John E. Hopcroft**[3] **& Liwei Wang**[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu

Improving Robust generalisation

With unlabelled data

Method	Robust Test Acc.	Standard Test Acc.	
Standard Training	0.8%	95.2%	Vanilla Supervised
PG-AT (Madry et al., 2018)	45.8%	87.3%	
TRADES (Zhang et al., 2019)	55.4%	84.0%	
Standard Self-Training	0.3%	96.4%	Semisupervised with same unlabeled data
Robust Consistency Training (Carmon et al., 2019)	56.5%	83.2%	
RST + PG-AT (this paper)	**58.5%**	**91.8%**	
RST + TRADES (this paper) (Carmon et al., 2019)	**63.1%**	**89.7%**	

Understanding and Mitigating the Tradeoff Between Robustness and Accuracy

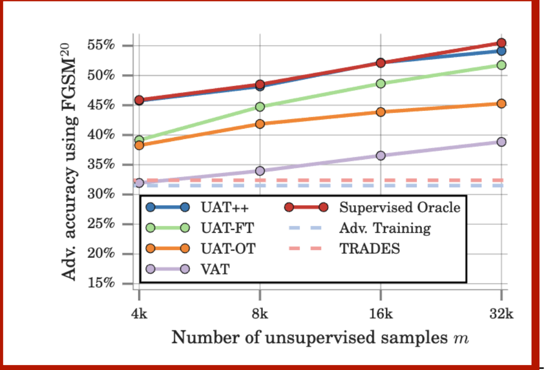
Aditi Raghunathan *[1] **Sang Michael Xie** *[1] **Fanny Yang** [2] **John C. Duchi** [1] **Percy Liang** [1]

ADVERSARIALLY ROBUST GENERALIZATION JUST REQUIRES MORE UNLABELED DATA

Runtian Zhai[1*], **Tianle Cai**[1*], **Di He**[1*], **Chen Dan**[2],
Kun He[4], **John E. Hopcroft**[3] **& Liwei Wang**[1]
[1]Peking University [2]Carnegie Mellon University [3]Cornell University
[4]Huazhong University of Science and Technology
{zhairuntian,caitianle1998,di_he,wanglw}@pku.edu.cn
cdan@cs.cmu.edu,brooklet60@hust.edu.cn,jeh17@cornell.edu

Unlabeled Data Improves Adversarial Robustness

Yair Carmon* Stanford University yairc@stanford.edu
Aditi Raghunathan* Stanford University aditir@stanford.edu
Ludwig Schmidt UC Berkeley ludwig@berkeley.edu
Percy Liang Stanford University pliang@cs.stanford.edu
John C. Duchi Stanford University jduchi@stanford.edu



Are Labels Required for Improving Adversarial Robustness?

Jonathan Uesato* **Jean-Baptiste Alayrac*** **Po-Sen Huang***
Robert Stanforth **Alhussein Fawzi** **Pushmeet Kohli**
DeepMind
{juesato,jalayrac,posenhuang}@google.com

Distributional Robustness in Machine Learning

Robustness to Distribution Shift

Robustness to Distribution Shift

- **Adversarial Robustness** measures performance against the worst shift between train and test set.

Robustness to Distribution Shift

- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to

Robustness to Distribution Shift

- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
    - Hard to collect quality data uniformly.

Robustness to Distribution Shift

- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
  - Hard to collect quality data uniformly.
  - Data sources evolve over time

Robustness to Distribution Shift

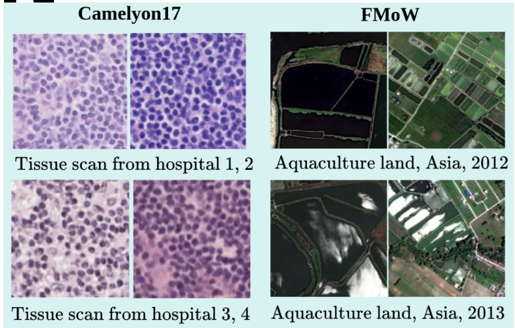
- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
  - Hard to collect quality data uniformly.
  - Data sources evolve over time

Robustness to Distribution Shift

- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
    - Hard to collect quality data uniformly.
    - Data sources evolve over time

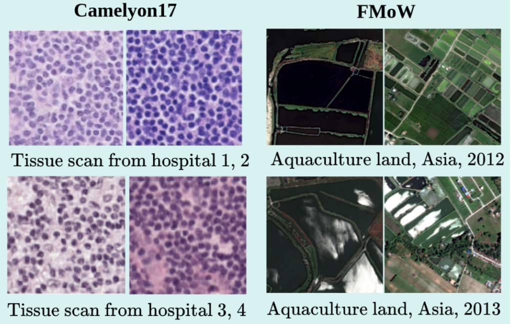
Robustness to Distribution Shift

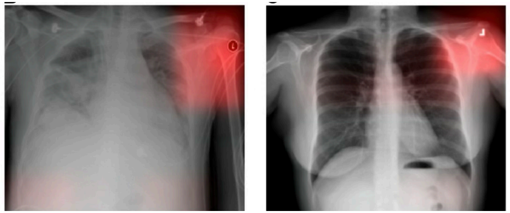
- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
  - Hard to collect quality data uniformly.
  - Data sources evolve over time



Robustness to Distribution Shift

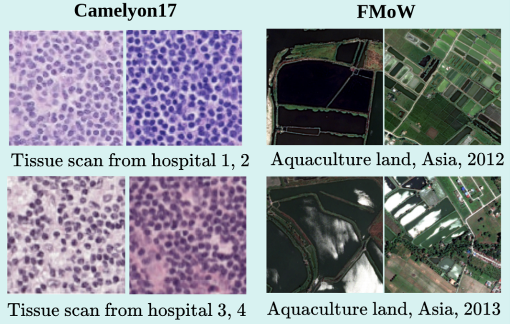
- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
  - Hard to collect quality data uniformly.
  - Data sources evolve over time

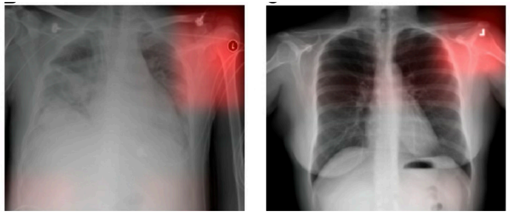




Robustness to Distribution Shift

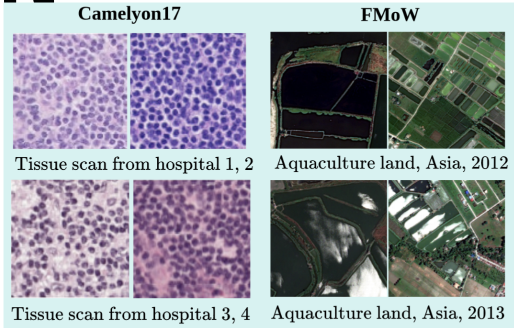
- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
  - Hard to collect quality data uniformly.
  - Data sources evolve over time

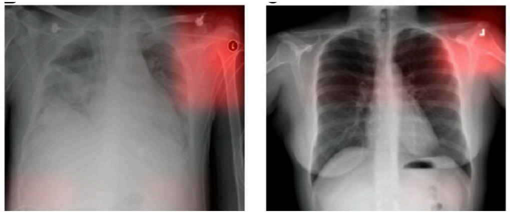




- Robustness to distribution shift requires ***preserving accuracy when the distribution shifts***

Robustness to Distribution Shift

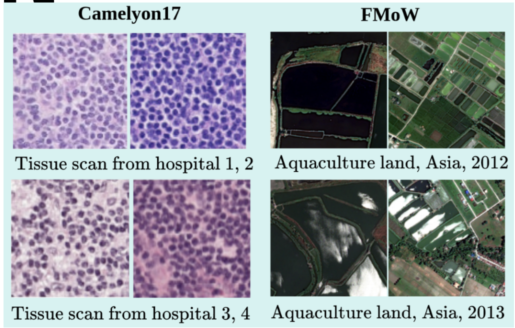


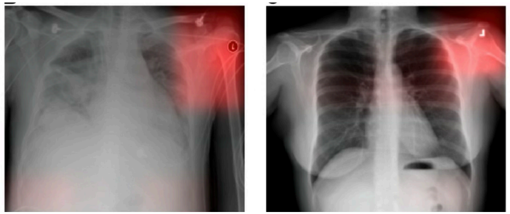


- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
  - Hard to collect quality data uniformly.
  - Data sources evolve over time
- Robustness to distribution shift requires ***preserving accuracy when the distribution shifts***
- Impossible to protect against arbitrary shifts

Robustness to Distribution Shift

- **Adversarial Robustness** measures performance against the worst shift between train and test set.
- More natural distribution shifts exist in the real world between train and test data e.g. due to
    - Hard to collect quality data uniformly.
    - Data sources evolve over time





- Robustness to distribution shift requires ***preserving accuracy when the distribution shifts***
- Impossible to protect against arbitrary shifts
- Goal is to allow for a graceful degradation with increasing shift

Robustness to Distribution Shift

Rich body of existing literature

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
    - Spurious correlations

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
  - Spurious correlations
  - Label Noise

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
    - Spurious correlations
    - Label Noise
    - Over-parameterised models

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
    - Spurious correlations
    - Label Noise
    - Over-parameterised models
- **How can models be made more robust to distribution shift ?**

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
  - Spurious correlations
  - Label Noise
  - Over-parameterised models
- **How can models be made more robust to distribution shift ?**
  - Distributionally Robust Optimisation

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
    - Spurious correlations
    - Label Noise
    - Over-parameterised models
- **How can models be made more robust to distribution shift ?**
    - Distributionally Robust Optimisation
    - Learn Causal/Robust representations

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
    - Spurious correlations
    - Label Noise
    - Over-parameterised models
- **How can models be made more robust to distribution shift ?**
    - Distributionally Robust Optimisation
    - Learn Causal/Robust representations
    - Collect more data (possibly unlabelled)

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

- **What causes failure to generalise to distribution shift ?**
  - Spurious correlations
  - Label Noise
  - Over-parameterised models
- **How can models be made more robust to distribution shift ?**
  - Distributionally Robust Optimisation
  - Learn Causal/Robust representations
  - Collect more data (possibly unlabelled)

Robustness to Distribution Shift

Rich body of existing literature

We will not even attempt to be exhaustive

Robustness to distribution shift features a rich body of existing literature asking

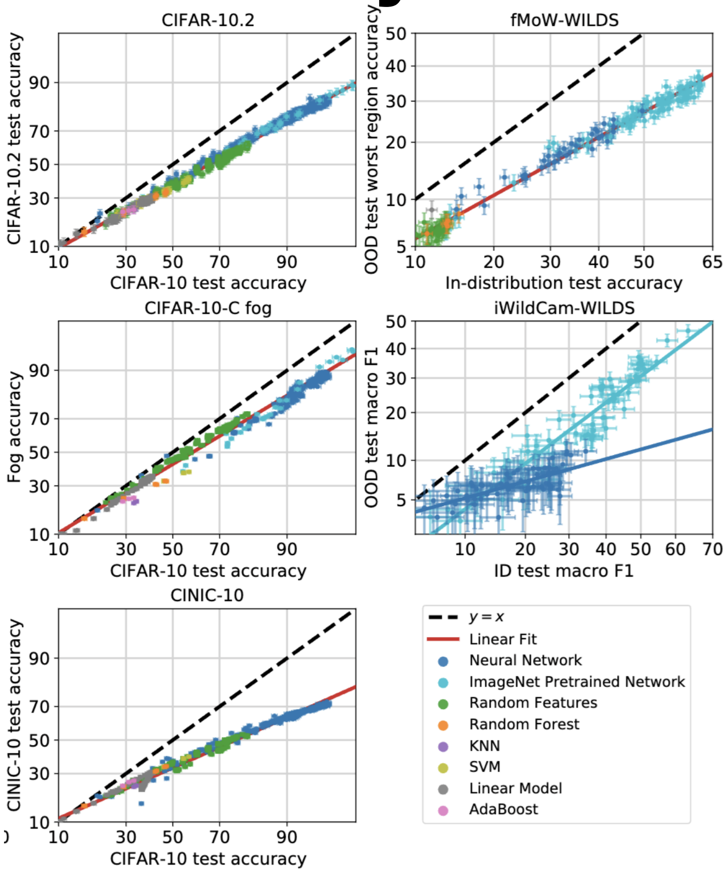
- **What causes failure to generalise to distribution shift ?**
  - Spurious correlations
  - Label Noise
  - Over-parameterised models
- **How can models be made more robust to distribution shift ?**
  - Distributionally Robust Optimisation
  - Learn Causal/Robust representations
  - Collect more data (possibly unlabelled)

Accuracy-on-the line

Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

John Miller* Rohan Taori† Aditi Raghunathan†
Shiori Sagawa† Pang Wei Koh† Vaishaal Shankar* Percy Liang†
Yair Carmon‡ Ludwig Schmidt§

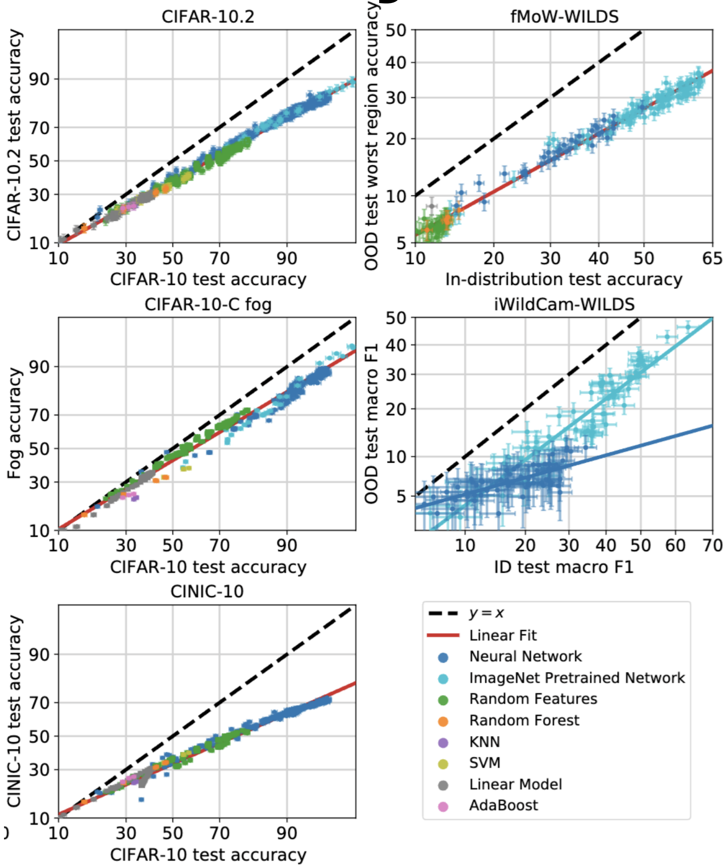
Accuracy-on-the line



Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

John Miller* Rohan Taori† Aditi Raghunathan†
Shiori Sagawa† Pang Wei Koh† Vaishaal Shankar* Percy Liang†
Yair Carmon‡ Ludwig Schmidt§

Accuracy-on-the line

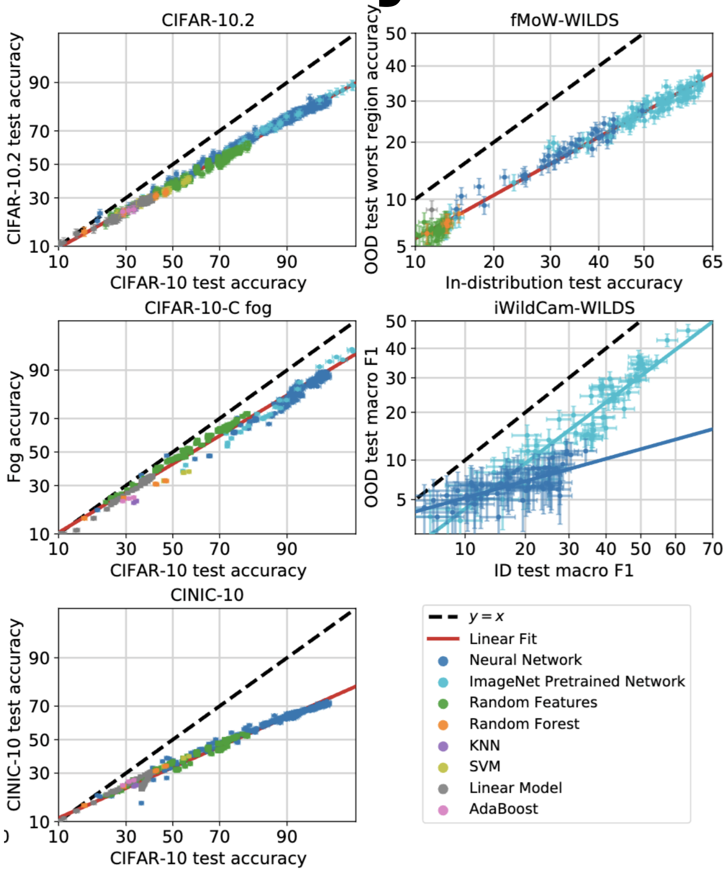


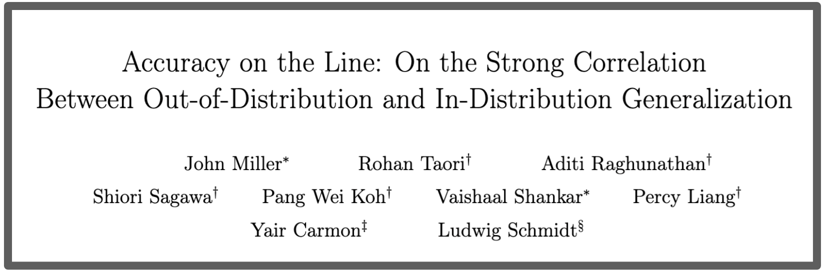
Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

John Miller* Rohan Taori† Aditi Raghunathan†
Shiori Sagawa† Pang Wei Koh† Vaishaal Shankar* Percy Liang†
Yair Carmon‡ Ludwig Schmidt§

- **Accuracy-on-the-line** phenomenon: ID and OOD accuracy are positively correlated.

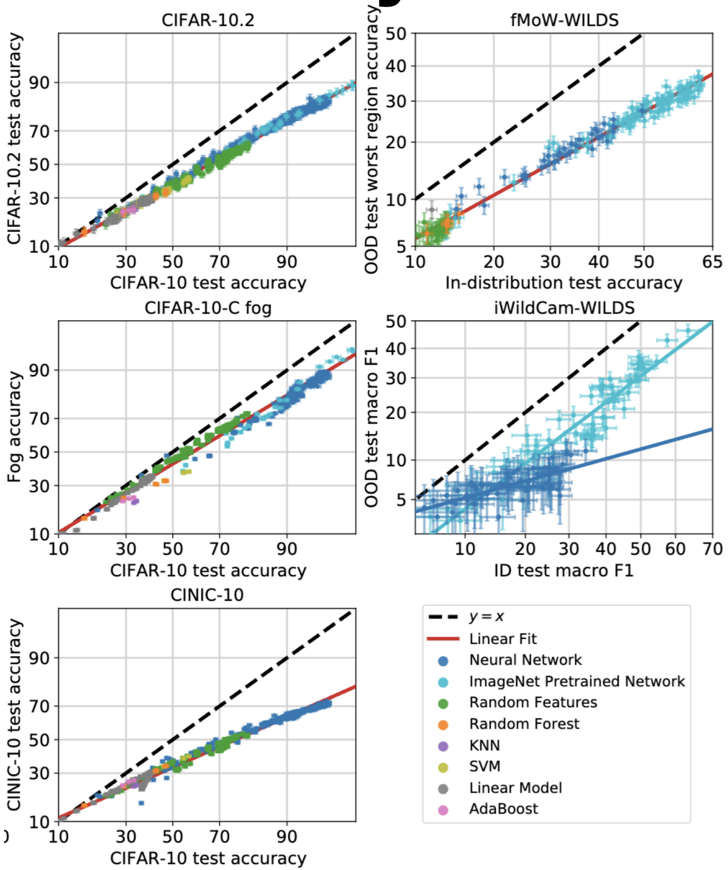
Accuracy-on-the line





- **Accuracy-on-the-line** phenomenon: ID and OOD accuracy are positively correlated.
- Indicates that improving ID accuracy also improves OOD accuracy.

Accuracy-on-the line



Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization

John Miller* Rohan Taori† Aditi Raghunathan†
Shiori Sagawa† Pang Wei Koh† Vaishaal Shankar* Percy Liang†
Yair Carmon‡ Ludwig Schmidt§

- **Accuracy-on-the-line** phenomenon: ID and OOD accuracy are positively correlated.
- Indicates that improving ID accuracy also improves OOD accuracy.
- Holds for a wide variety of models and datasets

Label Noise and Distribution Shift

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.

Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.

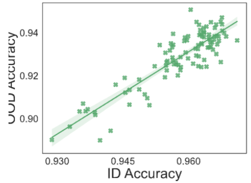
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(b) Noiseless dataset

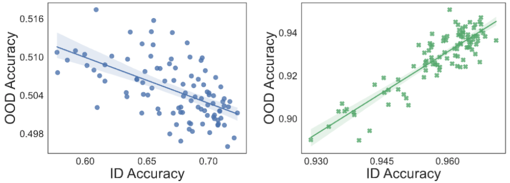
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset

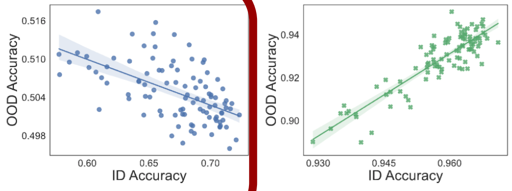
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset

Accuracy-on-the-wrong line

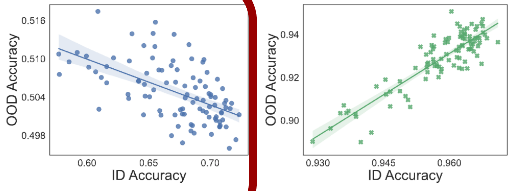
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

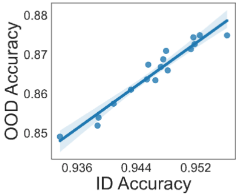
Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset

Accuracy-on-the-wrong line



(a) $\eta = 0$.

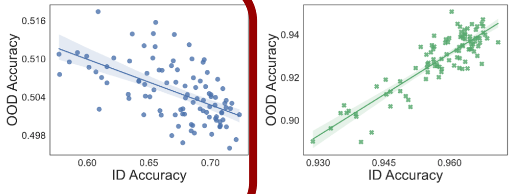
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

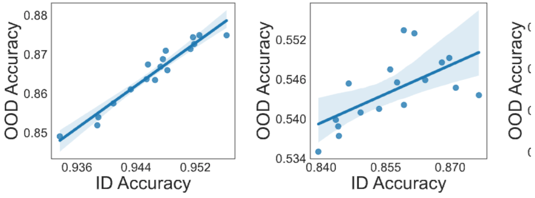
Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset

Accuracy-on-the-wrong line



(a) $\eta = 0.$ (b) $\eta = 0.15$

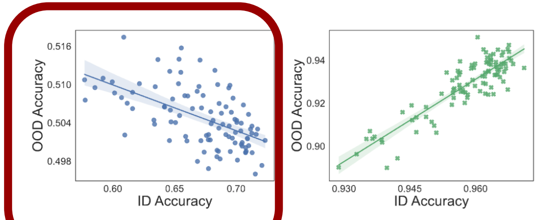
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

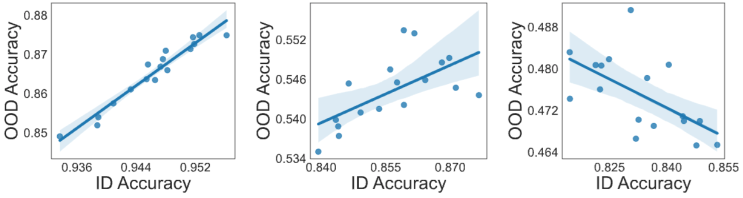
Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset

Accuracy-on-the-wrong line



(a) $\eta = 0.$ (b) $\eta = 0.15$ (c) $\eta = 0.2$

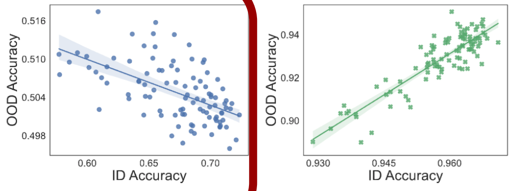
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

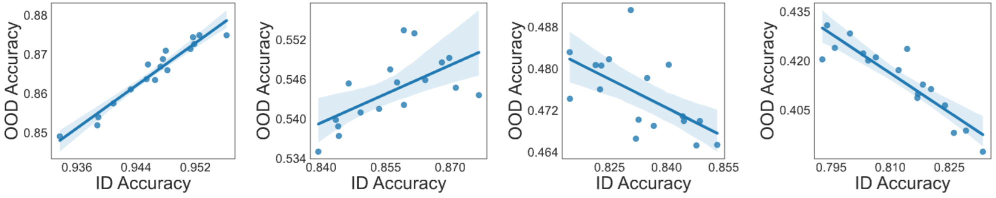
Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset

Accuracy-on-the-wrong line

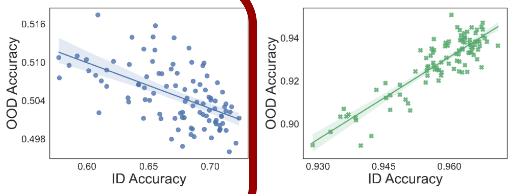


(a) $\eta = 0.$ (b) $\eta = 0.15$ (c) $\eta = 0.2$ (d) $\eta = 0.25$

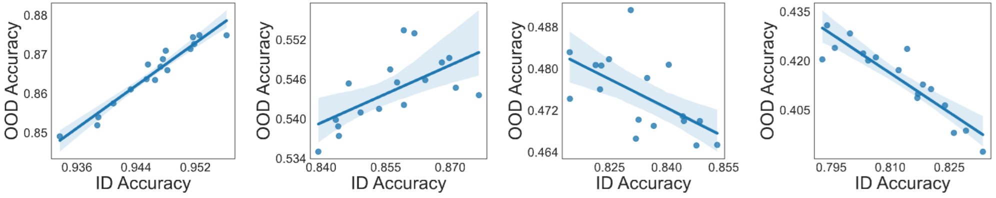
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

> Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation
>
> Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]
>
> [1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
> [2]University of California, Berkeley, U.S.A.
> [3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
> [4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset



(a) $\eta = 0.$ (b) $\eta = 0.15$ (c) $\eta = 0.2$ (d) $\eta = 0.25$

Accuracy-on-the-wrong line

Two sufficient factors for Accuracy-on-the-wrong-line

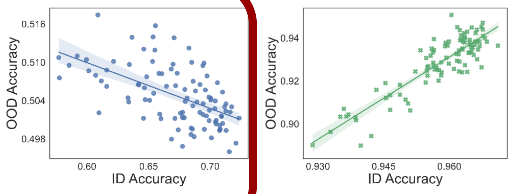
Label Noise and Distribution Shift

- Question: Is **Accuracy-on-the-line** robust to noisy or low quality labels ?

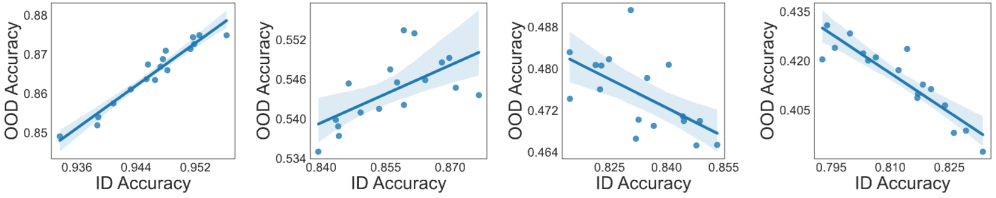
Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation

Amartya Sanyal[1], Yaxi Hu[1], Yaodong Yu[2], Yian Ma[3], Yixin Wang[4], and Bernhard Schölkopf[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of California, Berkeley, U.S.A.
[3]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, U.S.A.
[4]University of Michigan, Ann Arbor, U.S.A.



(a) Noisy dataset (b) Noiseless dataset



(a) $\eta = 0.$ (b) $\eta = 0.15$ (c) $\eta = 0.2$ (d) $\eta = 0.25$

Accuracy-on-the-wrong line

Two sufficient factors for Accuracy-on-the-wrong-line

- Inject and fit **random label noise** in the training data
- Presence of multiple "**nuisance features**" i.e. irrelevant features

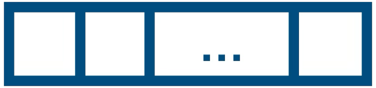
Theory

Theory

Let the data satisfy the following

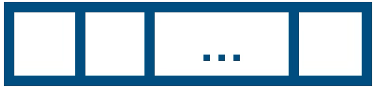
Theory

Let the data satisfy the following



Theory

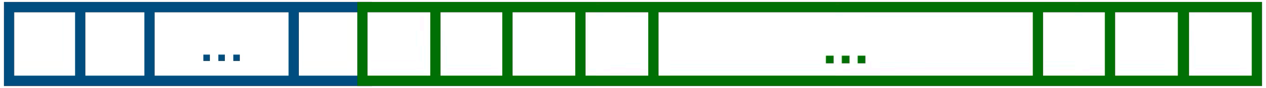
Let the data satisfy the following



Signal support S_0

Theory

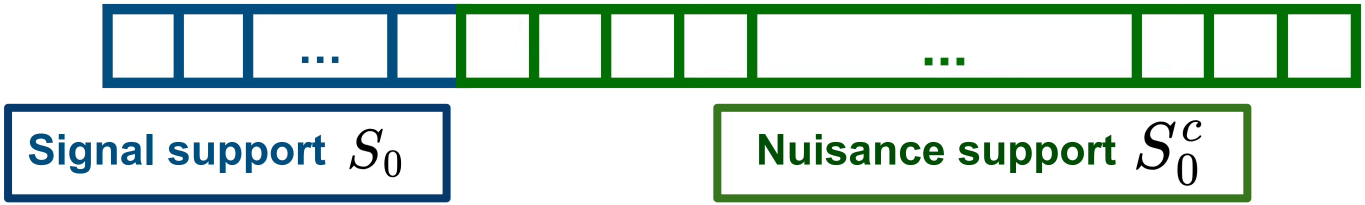
Let the data satisfy the following



Signal support S_0

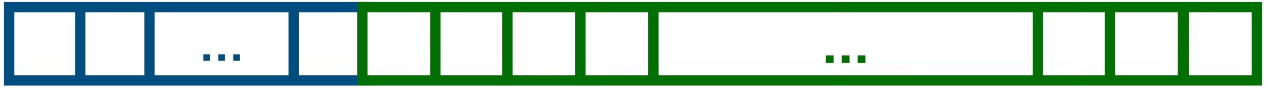
Theory

Let the data satisfy the following



Theory

Let the data satisfy the following



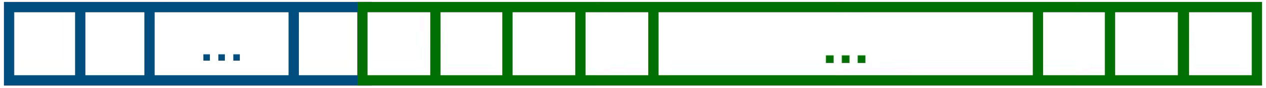
Signal support S_0

Nuisance support S_0^c

Data is labelled as $y = \langle \theta^\star, x \rangle$
where $\theta^\star$ is supported on S_0

Theory

Let the data satisfy the following



Signal support S_0

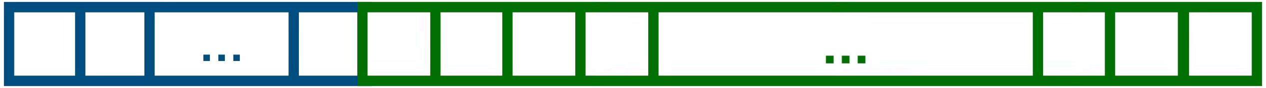
Nuisance support S_0^c

Data is labelled as $y = \langle \theta^\star, x \rangle$
where $\theta^\star$ is supported on S_0

S_0- oblivious shift distribution Δ only
affects nuisance support S_0^c.

Theory

Let the data satisfy the following



Signal support S_0

Nuisance support S_0^c

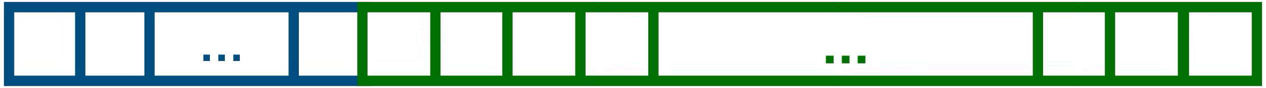
Data is labelled as $y = \langle \theta^\star, x \rangle$ where $\theta^\star$ is supported on S_0

S_0- oblivious shift distribution Δ only affects nuisance support S_0^c.

Implication of label noise: Learned linear classifier has spurious sensitivity $\geq \tau$

Theory

Let the data satisfy the following



Signal support S_0

Nuisance support S_0^c

Data is labelled as $y = \langle \theta^\star, x \rangle$ where $\theta^\star$ is supported on S_0

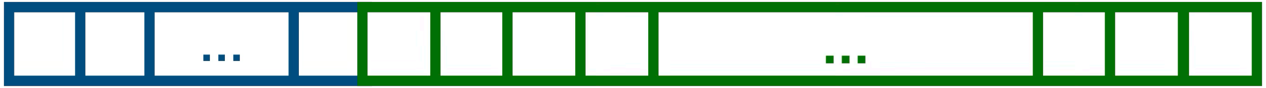
S_0- oblivious shift distribution Δ only affects nuisance support S_0^c.

Implication of label noise: Learned linear classifier has spurious sensitivity $\geq \tau$

Informal Theorem For all x s.t. $\langle \theta^\star, x \rangle > 0$, we have

Theory

Let the data satisfy the following



Signal support S_0

Nuisance support S_0^c

Data is labelled as $y = \langle \theta^\star, x \rangle$ where $\theta^\star$ is supported on S_0

S_0- oblivious shift distribution Δ only affects nuisance support S_0^c.

Implication of label noise: Learned linear classifier has spurious sensitivity $\geq \tau$

Informal Theorem For all x s.t. $\langle \theta^\star, x \rangle > 0$, we have

$$\Pr_{\delta \sim \Delta}\left[\langle \hat{\theta}, x + \delta < 0 \rangle\right] \geq 1 - \exp\left(- \left|\mathcal{S}_0^{\mathcal{C}}\right| \tau^2\right)$$

Using unlabelled data

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Using unlabelled data

- **Pre-trained representations** are a common strategy against this problem.
- But representations from supervised training often suffer from problems like **simplicity bias**.

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Using unlabelled data

- **Pre-trained representations** are a common strategy against this problem.
- But representations from supervised training often suffer from problems like **simplicity bias**.

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Solution - Use **Unlabelled data** & **unsupervised representation** learning

Using unlabelled data

- **Pre-trained representations** are a common strategy against this problem.
- But representations from supervised training often suffer from problems like **simplicity bias**.

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

Philip H.S. Torr
Department of Engineering Science
University of Oxford

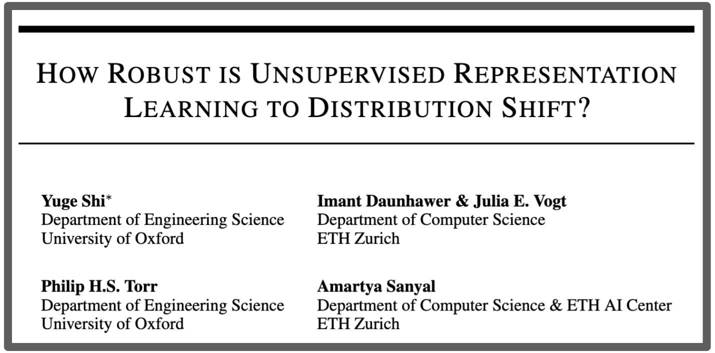
Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Solution - Use **Unlabelled data** & **unsupervised representation** learning

Experimental setup

Using unlabelled data

- **Pre-trained representations** are a common strategy against this problem.
- But representations from supervised training often suffer from problems like **simplicity bias**.



HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

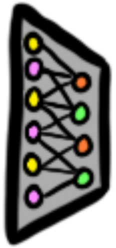
Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Solution - Use **Unlabelled data** & **unsupervised representation** learning

Experimental setup

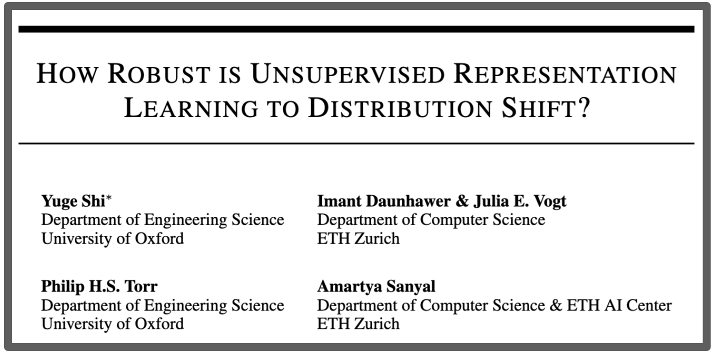
1



Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

Using unlabelled data

- **Pre-trained representations** are a common strategy against this problem.
- But representations from supervised training often suffer from problems like **simplicity bias**.


How Robust is Unsupervised Representation Learning to Distribution Shift?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

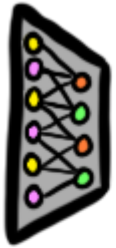
Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Solution - Use **Unlabelled data** & **unsupervised representation** learning

Experimental setup

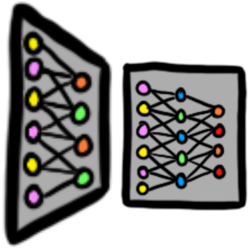
1



Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

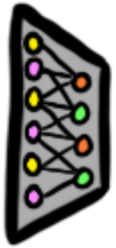
Train a small ML model on top of the features using **Dist X (ID or OOD)**

2



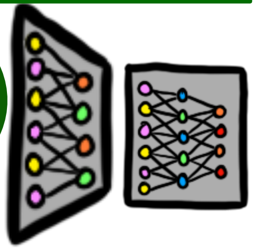
Using unlabelled data

1



Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

2



Train a small ML model on top of the features using **Dist X**

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

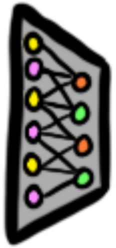
Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

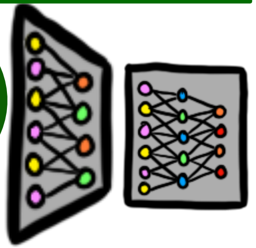
Using unlabelled data

Train a small ML model on top of the features using **Dist X**

1



2



Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

Dist X →OOD. Test on OOD.

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

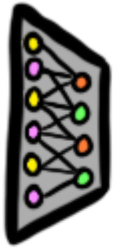
Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

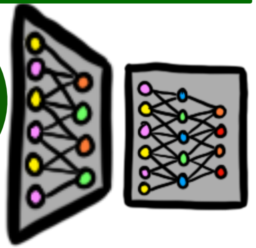
Using unlabelled data

Train a small ML model on top of the features using **Dist X**

1

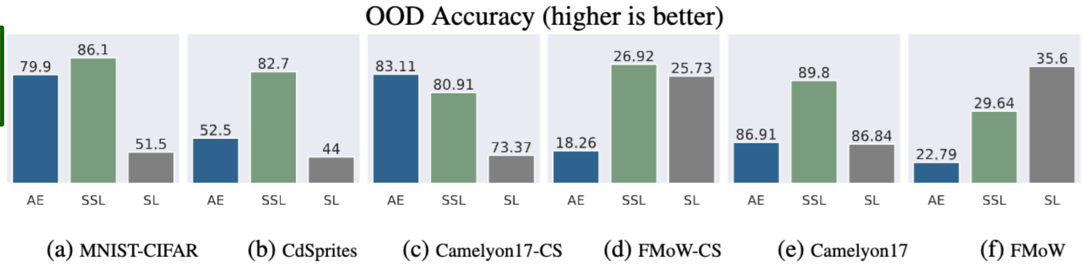


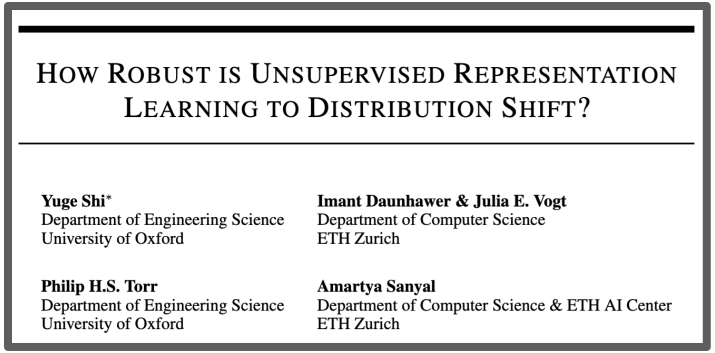
2



Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

Dist X →OOD. Test on OOD.





HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

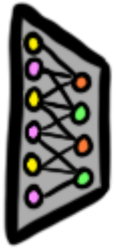
Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

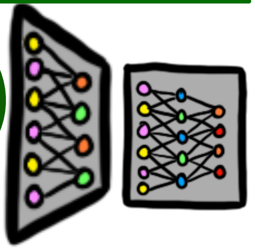
Using unlabelled data

Train a small ML model on top of the features using **Dist X**

1



2



Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

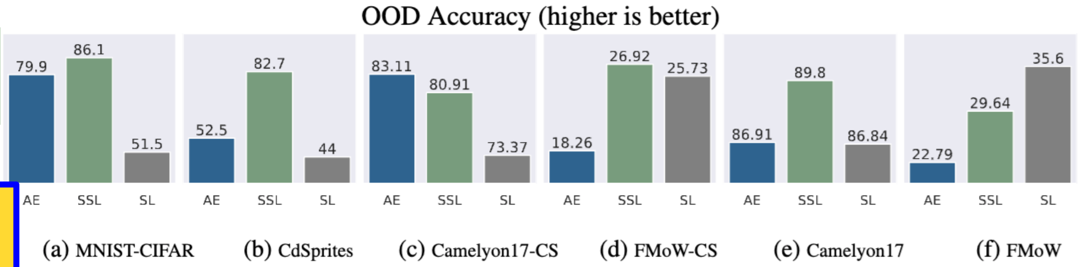
Yuge Shi*
Department of Engineering Science
University of Oxford

Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Dist X →OOD. Test on OOD.



Shift Sensitivity = Diff between

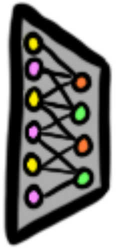
1. Dist X →OOD. Test on OOD.
2. Dist X →ID. Test on ID.

Captures robustness of

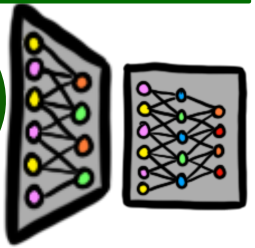
Using unlabelled data

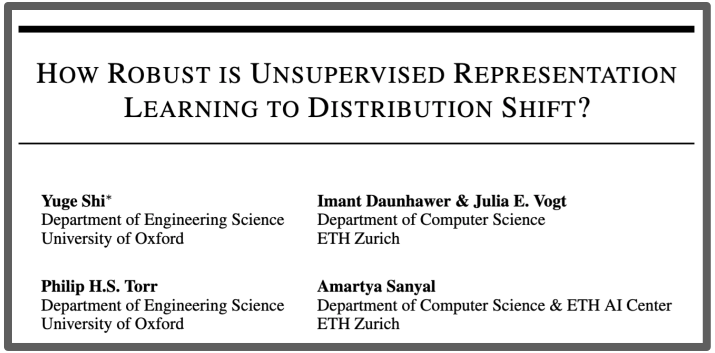
Train a small ML model on top of the features using **Dist X**

1



2





HOW ROBUST IS UNSUPERVISED REPRESENTATION LEARNING TO DISTRIBUTION SHIFT?

Yuge Shi*
Department of Engineering Science
University of Oxford

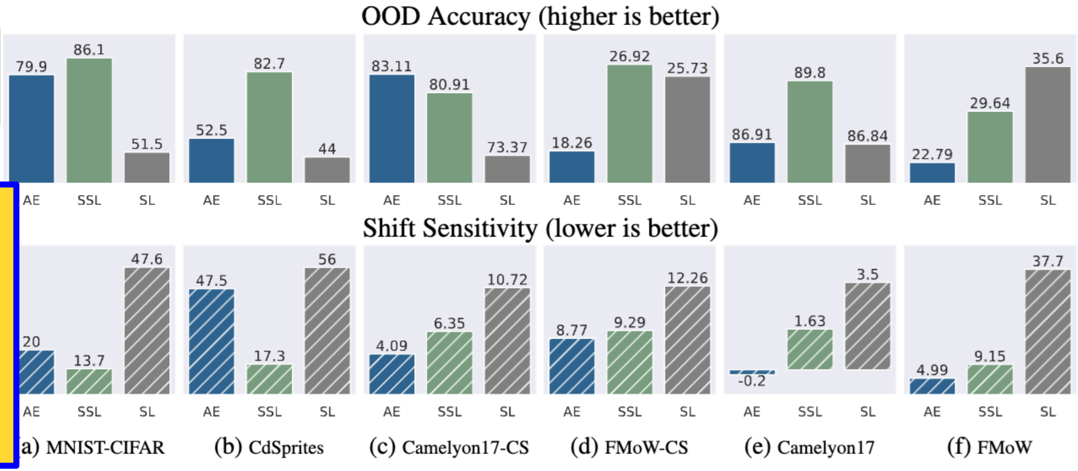
Imant Daunhawer & Julia E. Vogt
Department of Computer Science
ETH Zurich

Philip H.S. Torr
Department of Engineering Science
University of Oxford

Amartya Sanyal
Department of Computer Science & ETH AI Center
ETH Zurich

Pre-train representation learning on **ID data** with labelled (SL) or unlabelled data (AE/SSL)

Dist X →OOD. Test on OOD.



Shift Sensitivity = Diff between

1. Dist X →OOD. Test on OOD.
2. Dist X →ID. Test on ID.

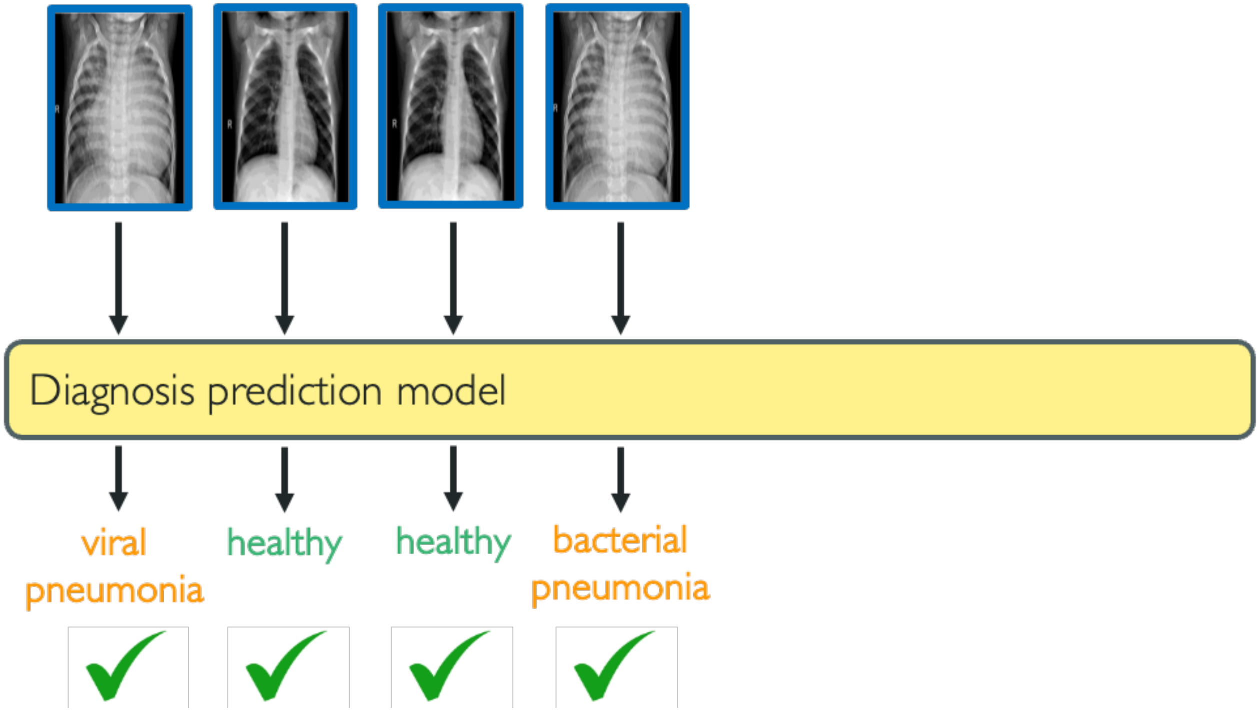
Captures robustness of

Out-of-distribution detection

What if we cannot predict reliably outside of the training distribution?

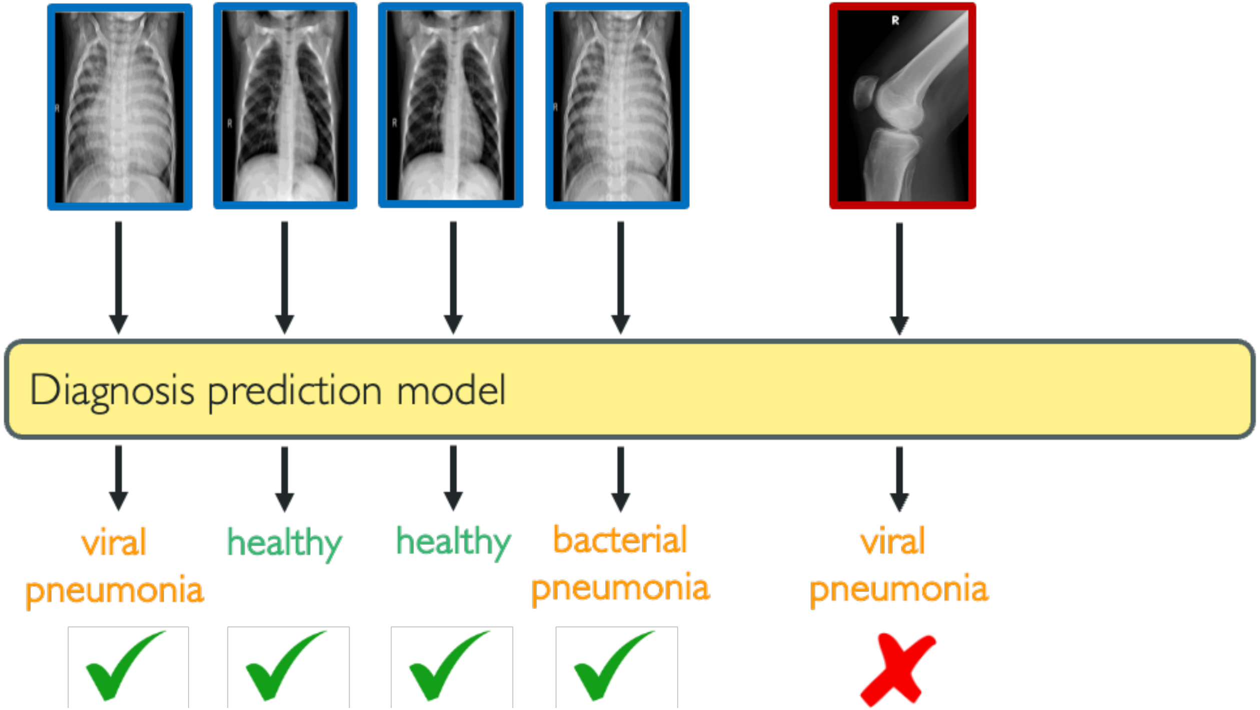
When can't we predict on OOD data?

Novel classes



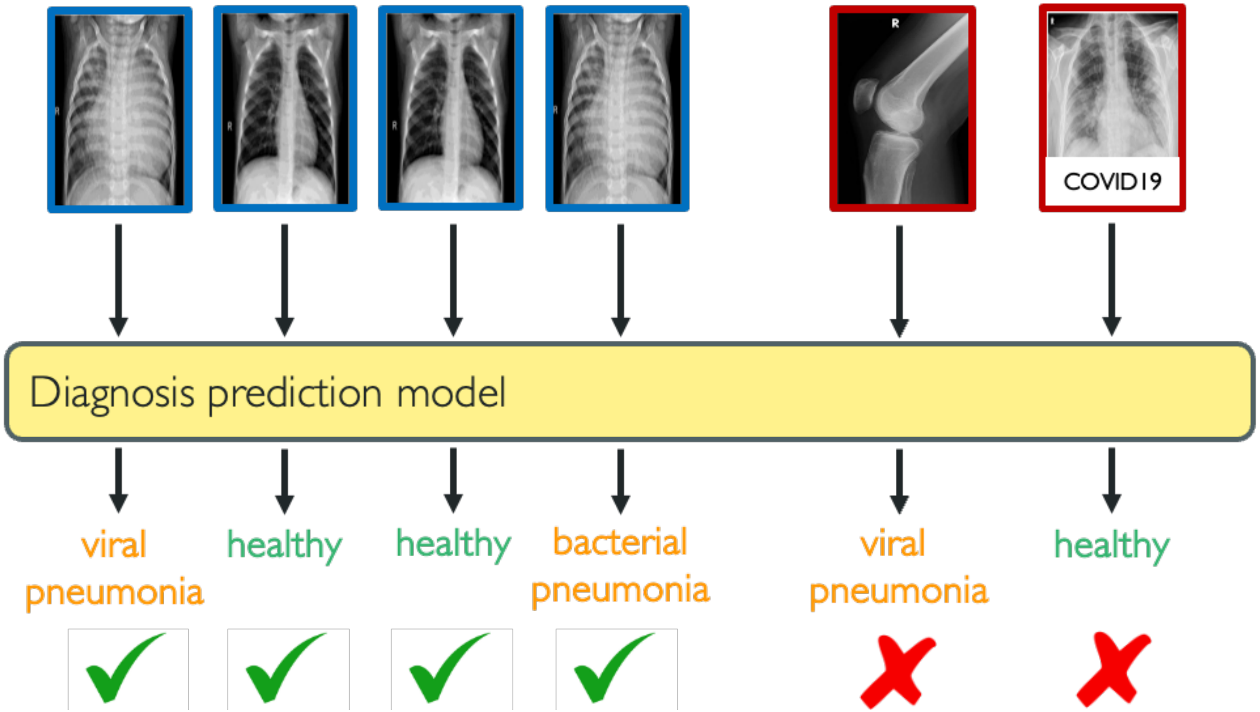
When can't we predict on OOD data?

Novel classes



When can't we predict on OOD data?

Novel classes



When can’t we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$

invariant parameters

domain-specific parameters

Figures courtesy of Fanny Yang.

When can't we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$

invariant parameters

domain-specific parameters

Figures courtesy of Fanny Yang.

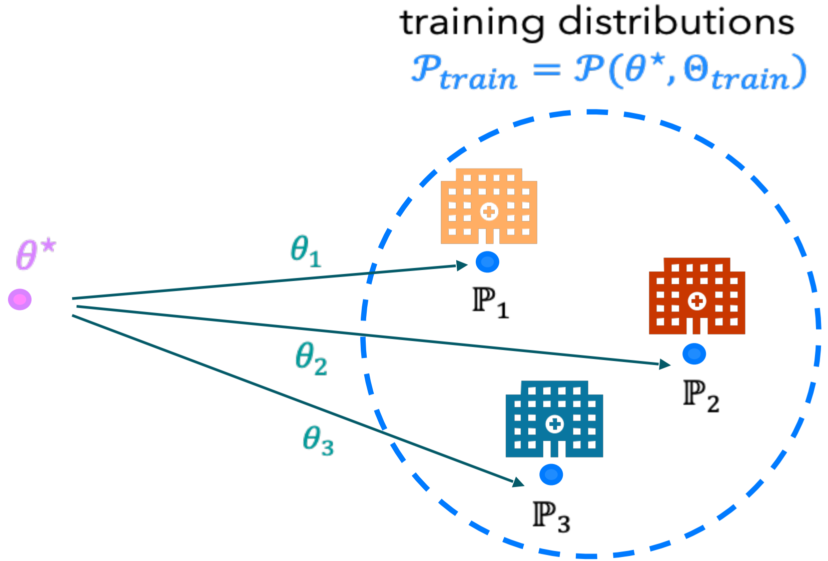
When can't we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$

invariant parameters

domain-specific parameters



Figures courtesy of Fanny Yang.

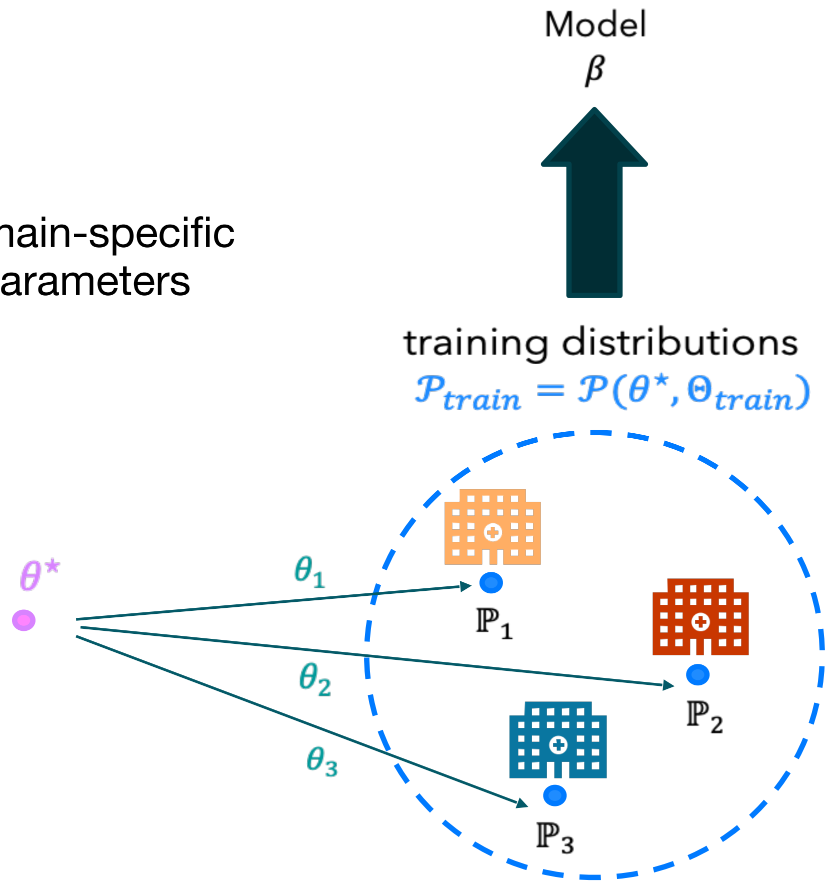
When can't we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$

invariant parameters

domain-specific parameters



Figures courtesy of Fanny Yang.

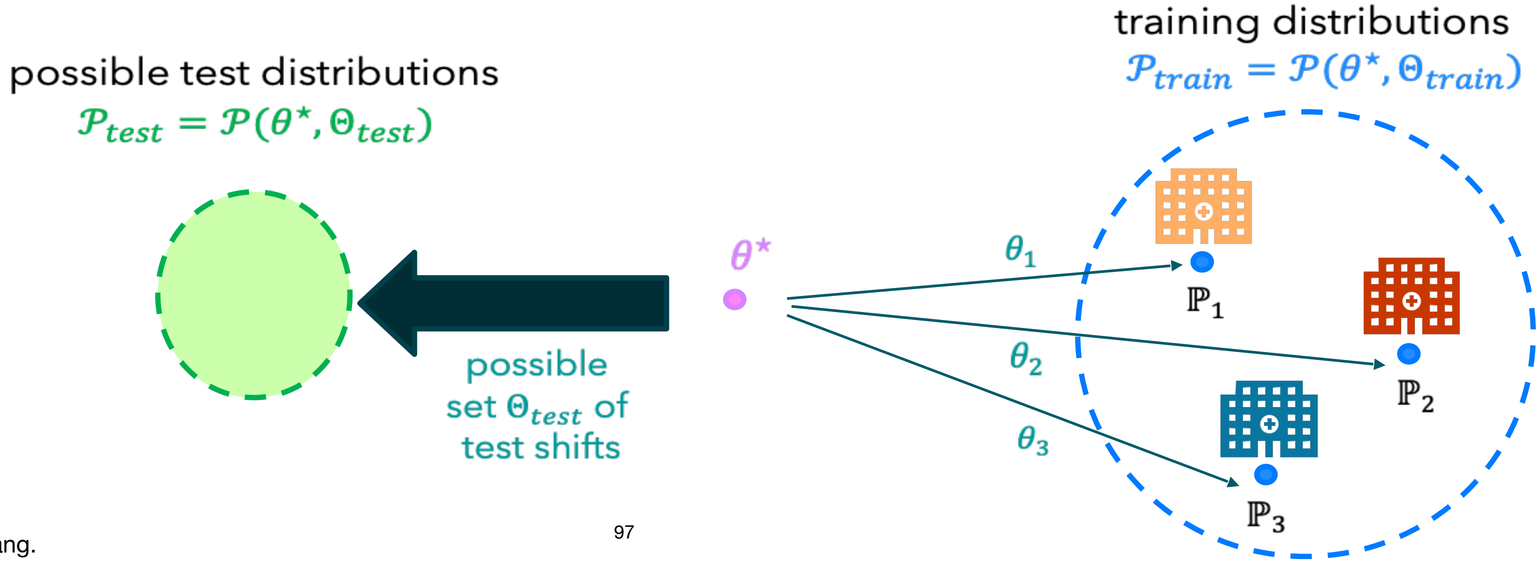
When can't we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$

invariant parameters — $\theta^\star$

domain-specific parameters — θ_e



Figures courtesy of Fanny Yang.

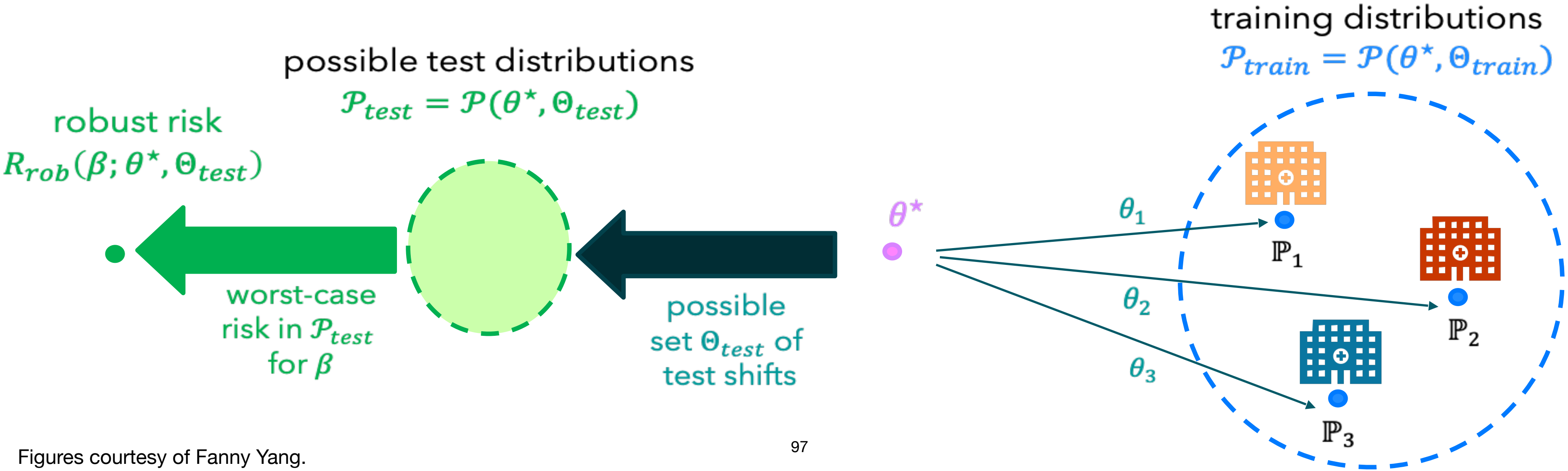
When can't we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$

invariant parameters

domain-specific parameters

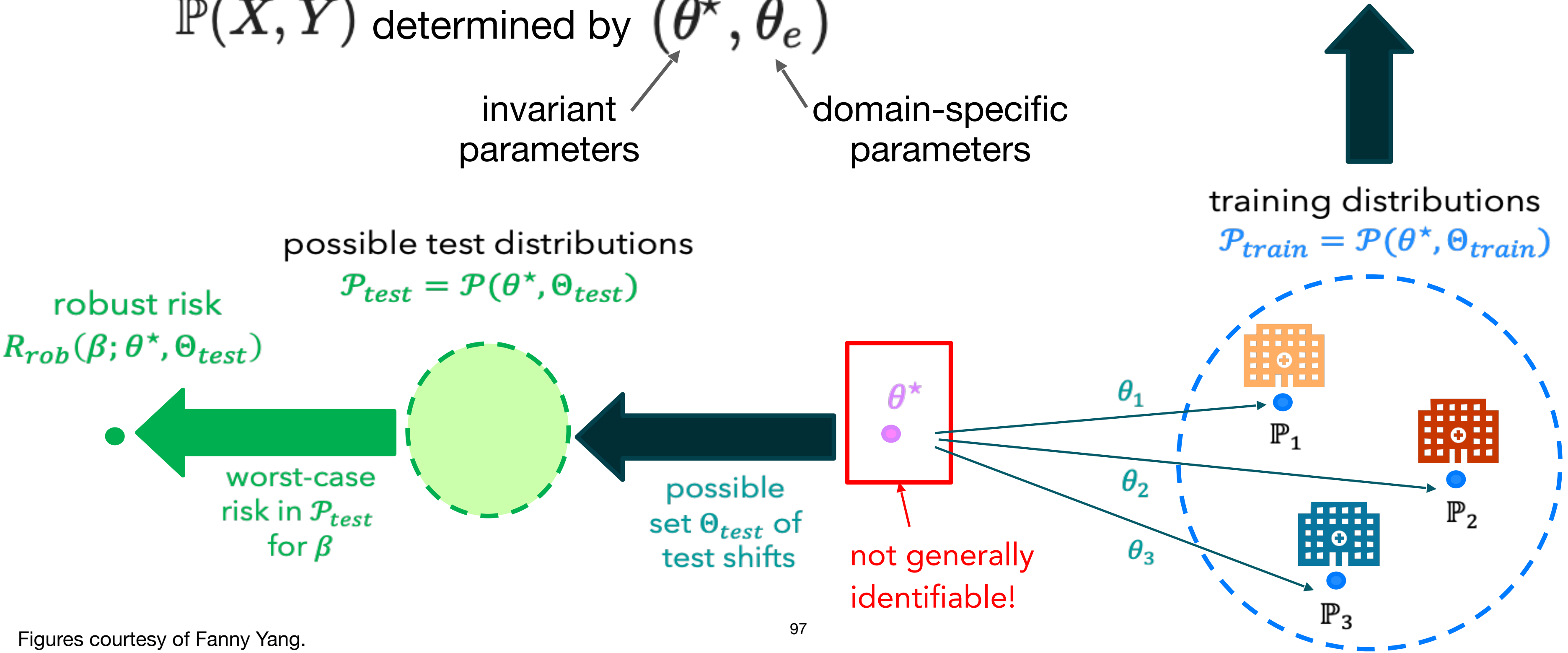


Figures courtesy of Fanny Yang.

When can't we predict on OOD data?

Strong distribution shifts

$\mathbb{P}(X, Y)$ determined by $(\theta^\star, \theta_e)$



Figures courtesy of Fanny Yang.

Impossibility result for distribution shifts

Achievable distributional robustness when the robust risk is only partially identified

Julia Kostin[1], Nicola Gnecco*[2], and Fanny Yang[1]

[1]Department of Computer Science, ETH Zurich
[2]Department of Mathematics, Imperial College London

Mean shifts during test time assumed to lie in $\Theta_{test} = \{\theta_{test} \colon \theta_{test}\theta_{test}^{\top} \preccurlyeq \boxed{\gamma M_{seen}} + \boxed{\gamma' M_{unseen}}\}$

Test time shifts assumptions

Covariance with range in span of **seen shift directions** $range(M_{seen}) \subset span\ \{\theta_e\}_{e \in [k]}$

Projection matrix onto **unseen direction**: $range(M_{seen}) \perp span\ \{\theta_e\}_{e \in [k]}$

Impossibility result for distribution shifts

Achievable distributional robustness when the robust risk is only partially identified

Julia Kostin[1], Nicola Gnecco[*2], and Fanny Yang[1]

[1]Department of Computer Science, ETH Zurich
[2]Department of Mathematics, Imperial College London

Mean shifts during test time assumed to lie in $\Theta_{test} = \{\theta_{test}: \theta_{test}\theta_{test}^\top \preccurlyeq \boxed{\gamma M_{seen}} + \boxed{\gamma' M_{unseen}}\}$

shift strengths

Test time shifts assumptions

Covariance with range in span of **seen shift directions** $range(M_{seen}) \subset span\,\{\theta_e\}_{e\in[k]}$

Projection matrix onto **unseen directions** $range(M_{seen}) \perp span\,\{\theta_e\}_{e\in[k]}$

Impossibility result for distribution shifts

Achievable distributional robustness when the robust risk is only partially identified

Julia Kostin[1], Nicola Gnecco*[2], and Fanny Yang[1]

[1]Department of Computer Science, ETH Zurich
[2]Department of Mathematics, Imperial College London

Mean shifts during test time assumed to lie in $\Theta_{test} = \{\theta_{test}\colon \theta_{test}\theta_{test}^\top \preccurlyeq \boxed{\gamma M_{seen}} + \boxed{\gamma' M_{unseen}}\}$

shift strengths

Test time shifts assumptions

Covariance with range in span of **seen shift directions** $range(M_{seen}) \subset span\,\{\theta_e\}_{e\in[k]}$

Projection matrix onto **unseen direction** $range(M_{seen}) \perp span\,\{\theta_e\}_{e\in[k]}$

Main theoretical result

Information-theoretic lower bound on robust risk.

Corollary

- **No "unseen" shifts:** Existing OOD generalization algorithms (e.g. anchor regression) are optimal.
- **No "seen" shifts:** Anchor regression is not better than ordinary least squares.

What if we cannot predict reliably outside of the training distribution?

What if we cannot predict reliably outside of the training distribution?

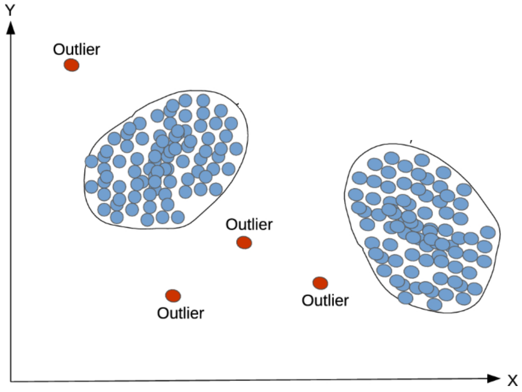
A: Flag out-of-domain samples and abstain.

Traditional OOD detection methods

Unsupervised OOD i.e. only observe in-distribution samples.

Examples:

Density estimation
e.g. in NN embedding space



Predictive uncertainty
e.g. ensembles

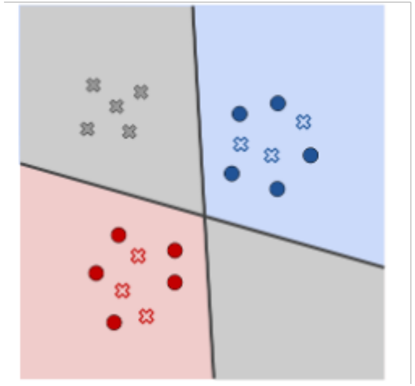


Figure sources: https://link.springer.com/article/10.1007/s10044-021-00998-6, https://arxiv.org/abs/2012.05825

Limitations of unsupervised OOD detection

Limitations of unsupervised OOD detection

Challenge #1: Unsupervised OOD detection can be ill-defined

Perfect Density Models Cannot Guarantee Anomaly Detection

Charline Le Lan [1,2,*] and Laurent Dinh [2]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
[2] Google Research, Montreal H3B 2Y5, CA

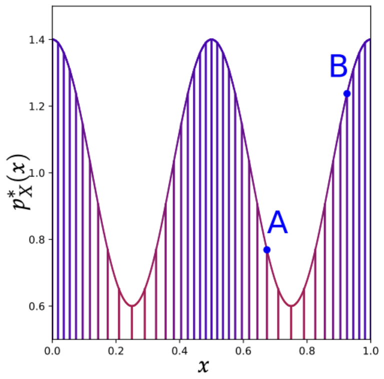
Limitations of unsupervised OOD detection

Challenge #1: Unsupervised OOD detection can be ill-defined

Perfect Density Models Cannot Guarantee Anomaly Detection

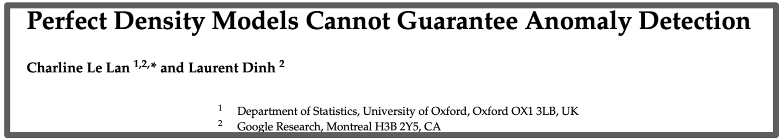
Charline Le Lan [1,2,*] and Laurent Dinh [2]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
[2] Google Research, Montreal H3B 2Y5, CA



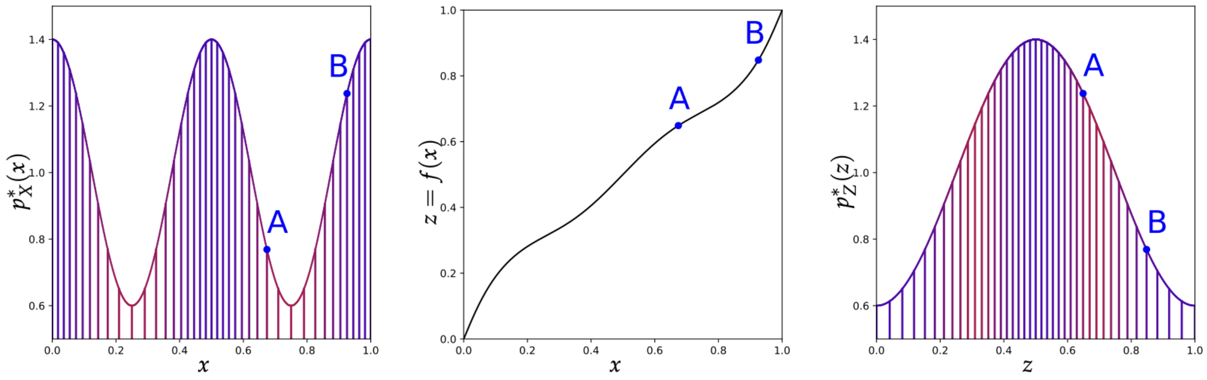
Limitations of unsupervised OOD detection

Challenge #1: Unsupervised OOD detection can be ill-defined


Perfect Density Models Cannot Guarantee Anomaly Detection

Charline Le Lan [1,2,*] and Laurent Dinh [2]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
[2] Google Research, Montreal H3B 2Y5, CA



invertible change of representation

Limitations of unsupervised OOD detection

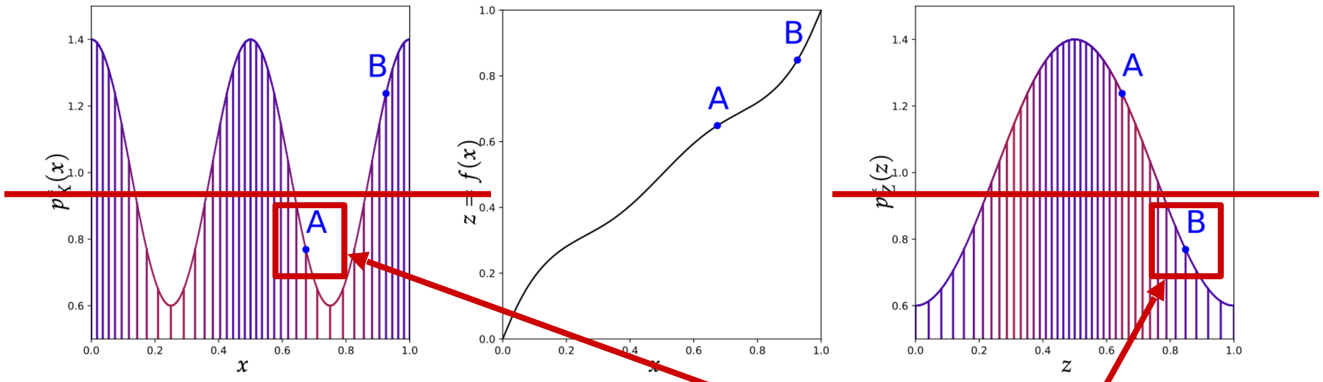
Challenge #1: Unsupervised OOD detection can be ill-defined

Perfect Density Models Cannot Guarantee Anomaly Detection

Charline Le Lan [1,2,*] and Laurent Dinh [2]

[1] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
[2] Google Research, Montreal H3B 2Y5, CA

OOD threshold

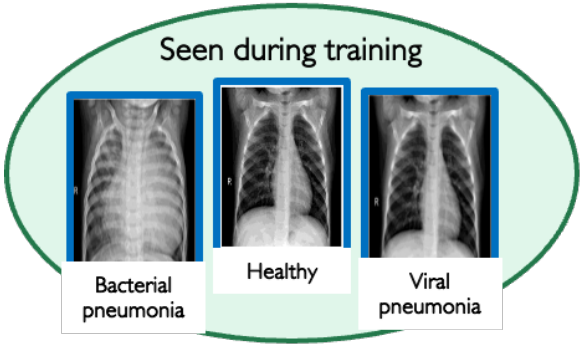


invertible change of representation

outliers!

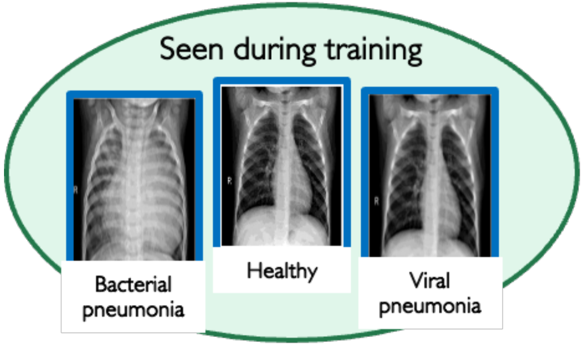
Limitations of unsupervised OOD detection

Challenge #2: finite samples + curse of dimensionality



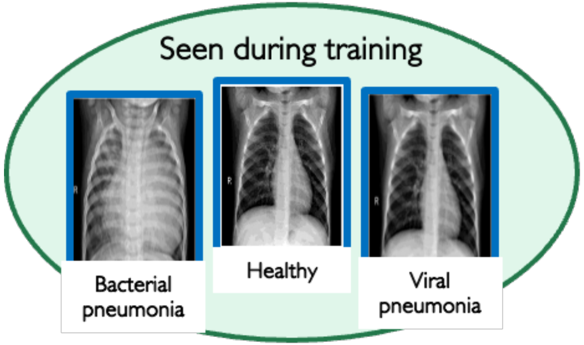
Limitations of unsupervised OOD detection

Challenge #2: finite samples + curse of dimensionality

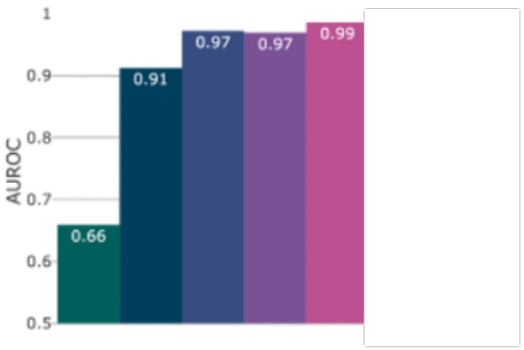


Limitations of unsupervised OOD detection

Challenge #2: finite samples + curse of dimensionality

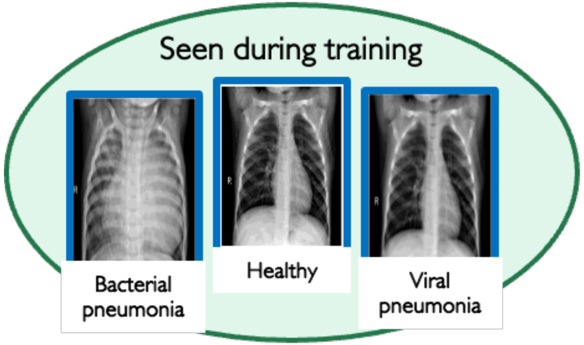


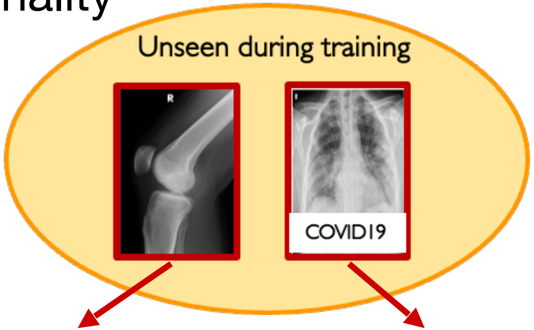
Far OOD ("easy")
i.e. OOD = new dataset



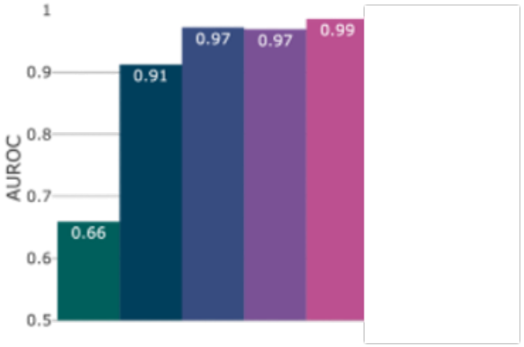
Limitations of unsupervised OOD detection

Challenge #2: finite samples + curse of dimensionality

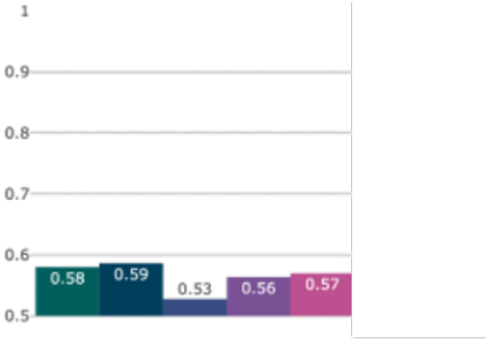




Far OOD ("easy")
i.e. OOD = new dataset



Near OOD ("hard")
i.e. OOD = new classes



Diverse pre-training data

Exploring the Limits of Out-of-Distribution Detection

Stanislav Fort* Stanford University sfort1@stanford.edu
Jie Ren* Google Research, Brain Team jjren@google.com
Balaji Lakshminarayanan Google Research, Brain Team balajiln@google.com

Pre-train on ImageNet21k

Diverse pre-training data

Exploring the Limits of Out-of-Distribution Detection

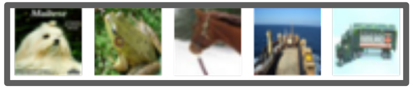
Stanislav Fort*
Stanford University
sfort1@stanford.edu

Jie Ren*
Google Research, Brain Team
jjren@google.com

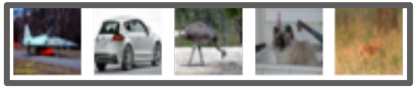
Balaji Lakshminarayanan
Google Research, Brain Team
balajiln@google.com

Pre-train on ImageNet21k

Fine-tune on CIFAR10

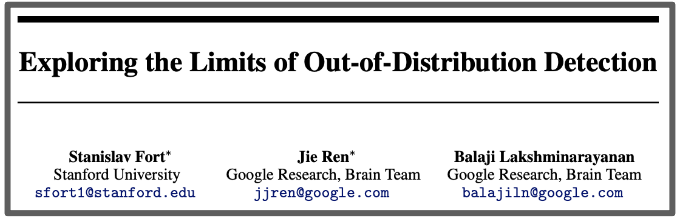


Outliers: CIFAR100



	AUROC
Unsup. method:	0.80
Pretrained method:	0.97

Diverse pre-training data


Exploring the Limits of Out-of-Distribution Detection

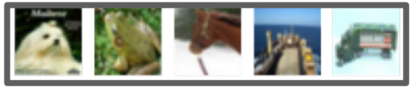
Stanislav Fort*
Stanford University
sfort1@stanford.edu

Jie Ren*
Google Research, Brain Team
jjren@google.com

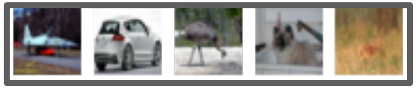
Balaji Lakshminarayanan
Google Research, Brain Team
balajiln@google.com

Pre-train on **ImageNet21k**

Fine-tune on **CIFAR10**

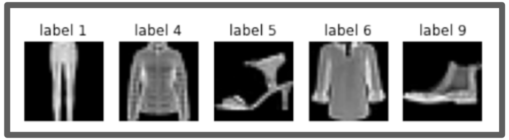


Outliers: CIFAR100

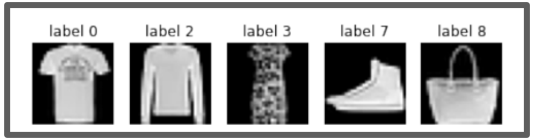


	AUROC
Unsup. method:	**0.80**
Pretrained method:	**0.97**

Fine-tune on
5-class FashionMNIST



Outliers: remaining FashionMNIST classes



	AUROC
Unsup. method:	**0.82**
Pretrained method:	**0.87**

Using proxy OOD data

Natural proxy OOD data

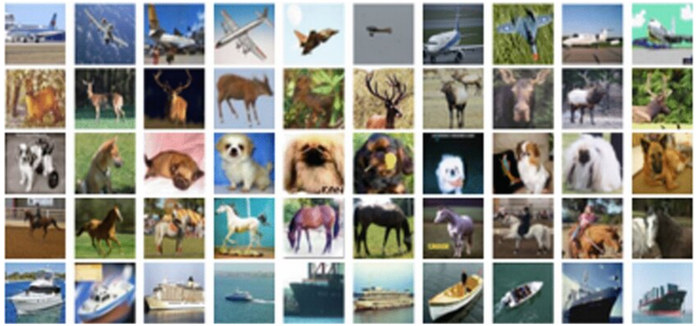
DEEP ANOMALY DETECTION WITH OUTLIER EXPOSURE

Dan Hendrycks
University of California, Berkeley
hendrycks@berkeley.edu

Mantas Mazeika
University of Chicago
mantas@ttic.edu

Thomas Dietterich
Oregon State University
tgd@oregonstate.edu

Known outliers: TinyImages dataset
(superset of CIFAR10/100)



Synthetic proxy OOD data

CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances

Jihoon Tack*†, Sangwoo Mo*‡, Jongheon Jeong‡, Jinwoo Shin†‡
†Graduate School of AI, KAIST
‡School of Electrical Engineering, KAIST

Known outliers: synthetic image transformations

(a) Original (b) Cutout (c) Sobel (d) Noise (e) Blur (f) Perm (g) Rotate

Using proxy OOD data

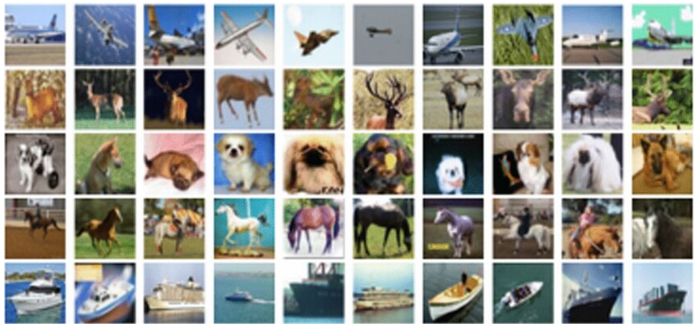
DEEP ANOMALY DETECTION WITH OUTLIER EXPOSURE

Dan Hendrycks
University of California, Berkeley
hendrycks@berkeley.edu

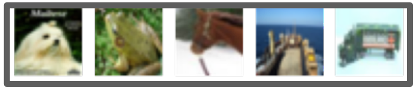
Mantas Mazeika
University of Chicago
mantas@ttic.edu

Thomas Dietterich
Oregon State University
tgd@oregonstate.edu

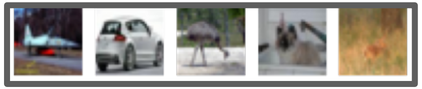
Known outliers: TinyImages dataset
(superset of CIFAR10/100)



In-distribution data:
5-class CIFAR10



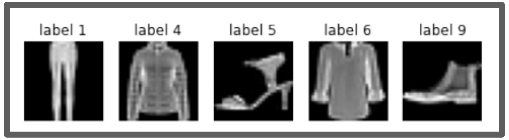
Outliers: **remaining CIFAR10 classes**



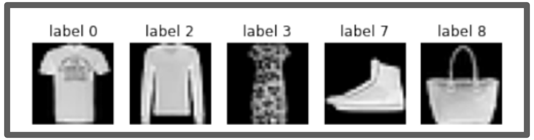
AUROC

Outlier exposure method: **0.82**

In-distribution data:
5-class FashionMNIST



Outliers: **remaining FashionMNIST classes**



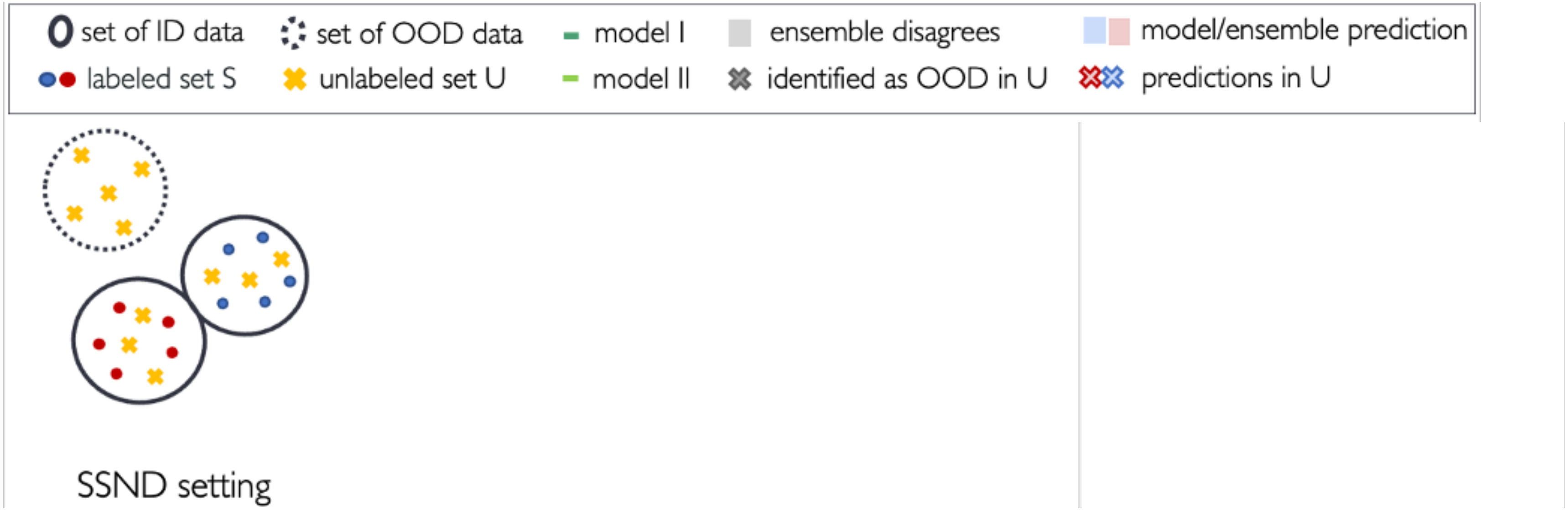
AUROC

Outlier exposure method: **0.66**

Semi-supervised OOD detection
Leveraging unlabeled data

Semi-supervised novelty detection using ensembles with regularized disagreement

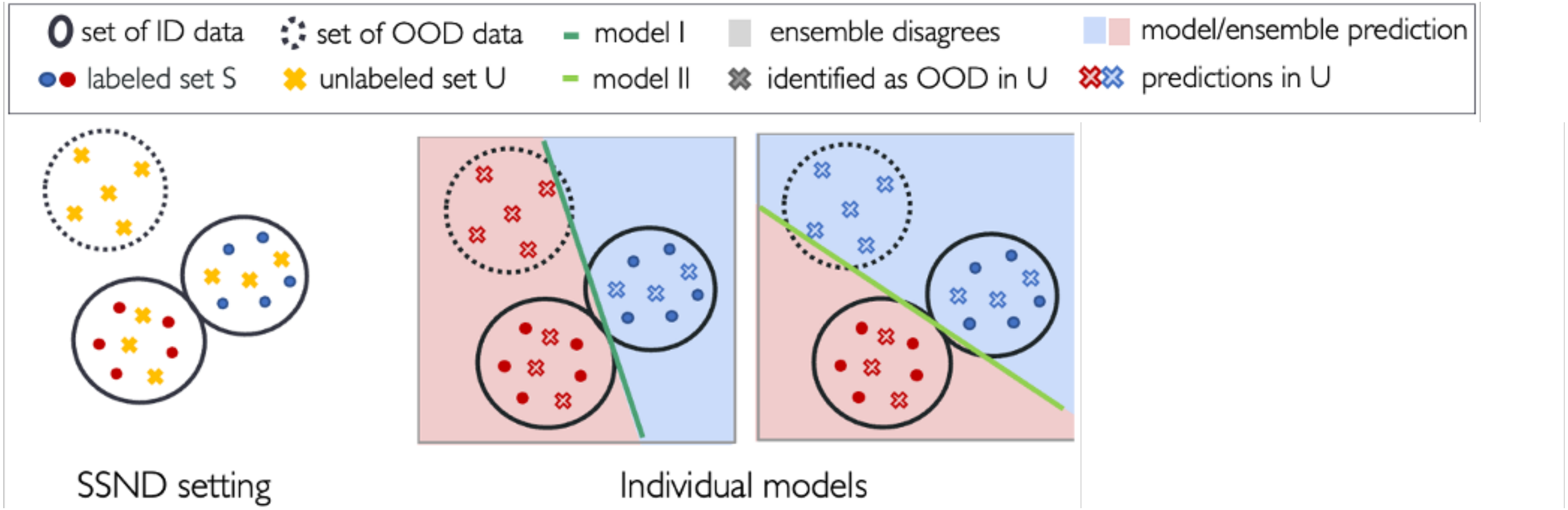
Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



Semi-supervised OOD detection
Leveraging unlabeled data

Semi-supervised novelty detection using ensembles with regularized disagreement

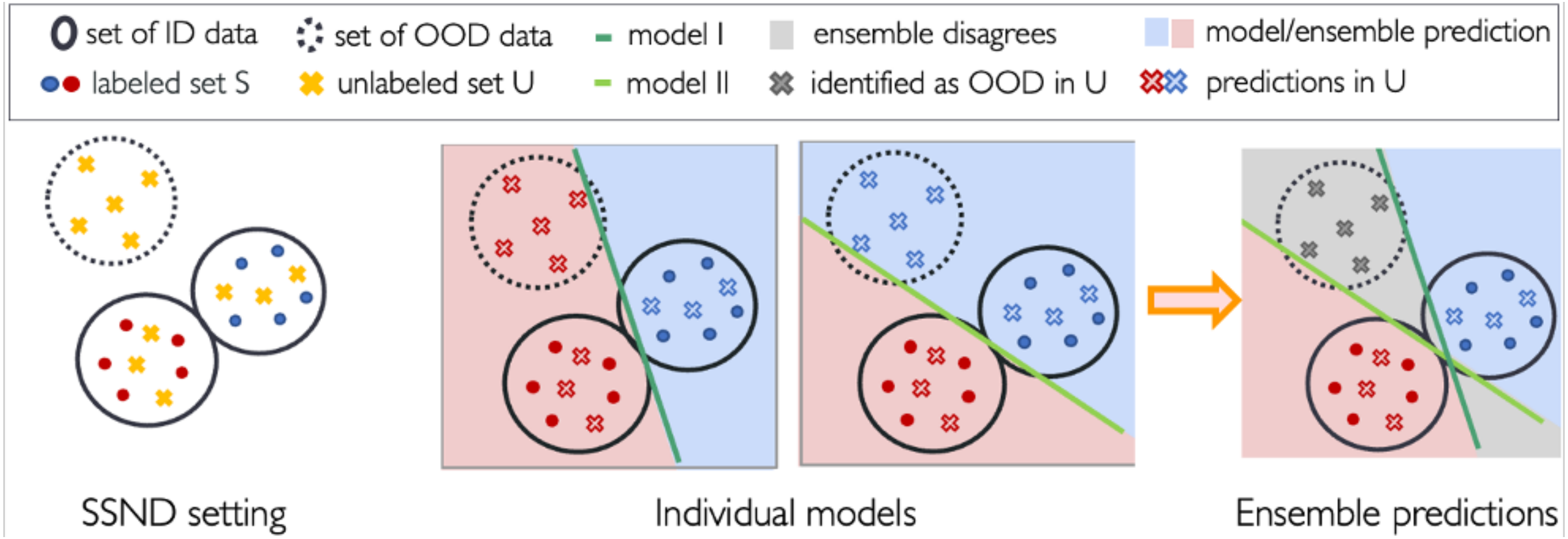
Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



Semi-supervised OOD detection
Leveraging unlabeled data

Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



sample x is flagged as OOD if “disagreement” > threshold

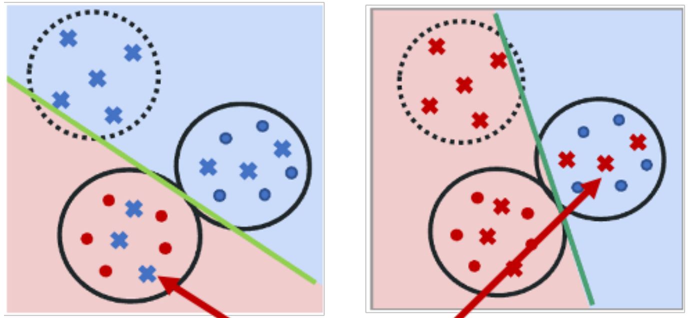
e.g. average pairwise TV distance between predictive distributions of the models in ensemble

Semi-supervised OOD detection

Key ingredient: Appropriate regularization

Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



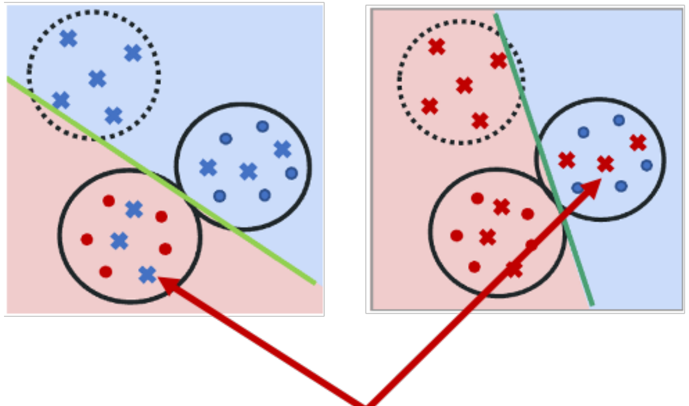
label noise

Semi-supervised OOD detection

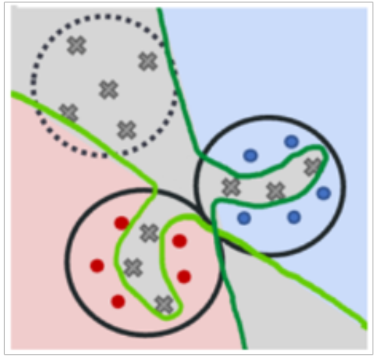
Key ingredient: Appropriate regularization

Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



label noise



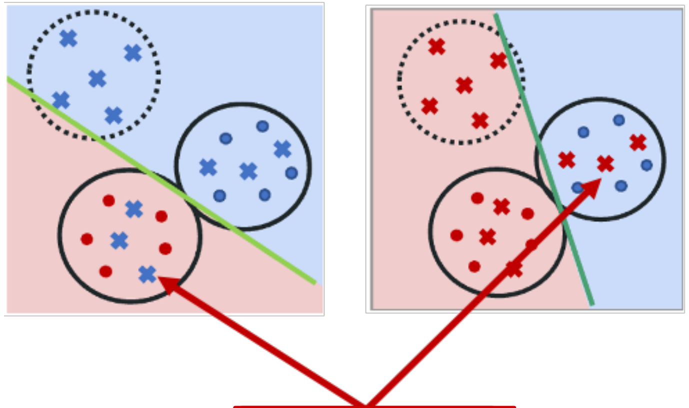
Too much diversity

Semi-supervised OOD detection

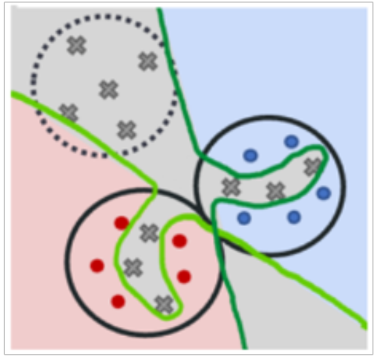
Key ingredient: Appropriate regularization

Semi-supervised novelty detection using ensembles with regularized disagreement

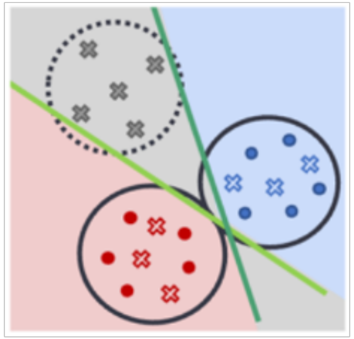
Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



label noise



Too much diversity



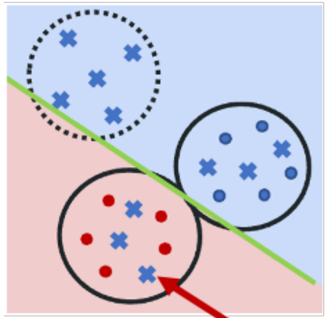
Right amount of diversity

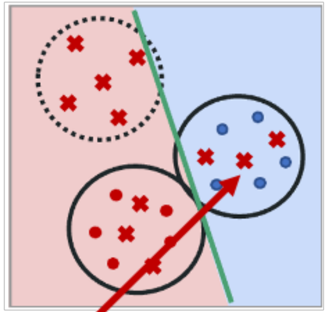
Semi-supervised OOD detection

Key ingredient: Appropriate regularization

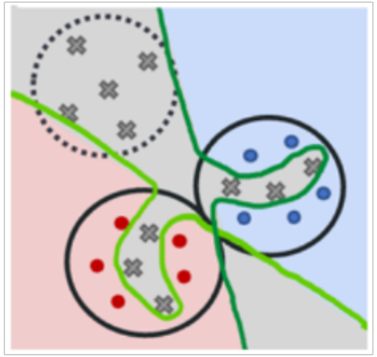
Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland

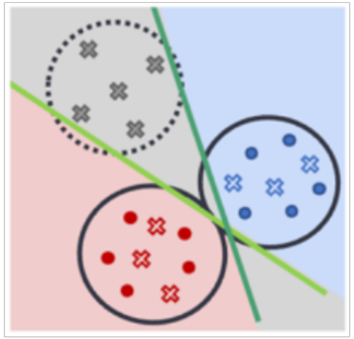




label noise



Too much diversity



Right amount of diversity

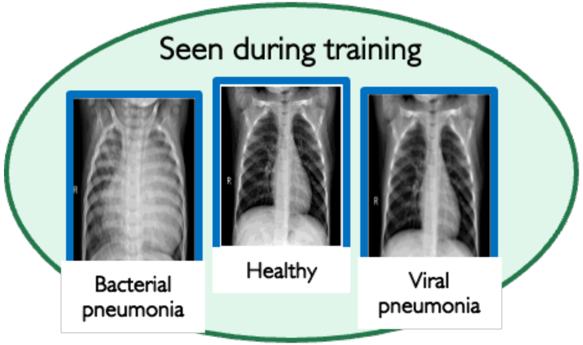
Idea: regularization with strength chosen using ID validation set

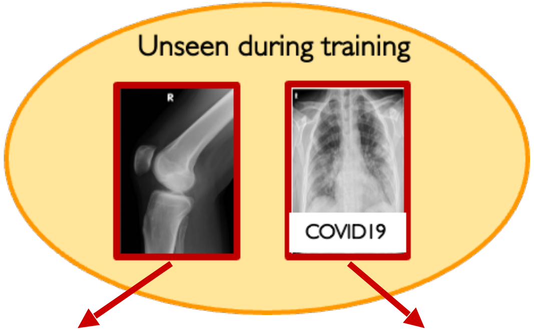
i.e. control FPR (ID samples incorrectly flagged as OOD)

Semi-supervised OOD detection
Performance on near OOD data

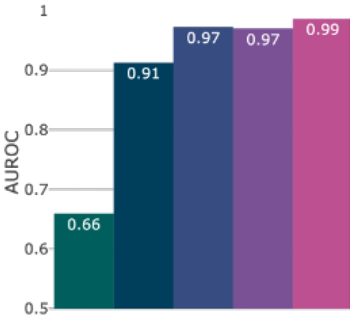
Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland

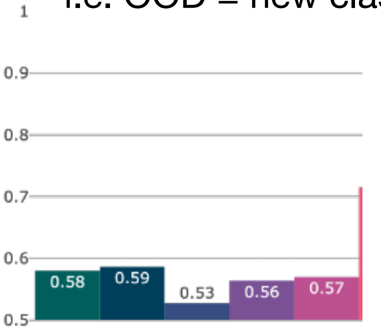




Far OOD ("easy")
i.e. OOD = new dataset



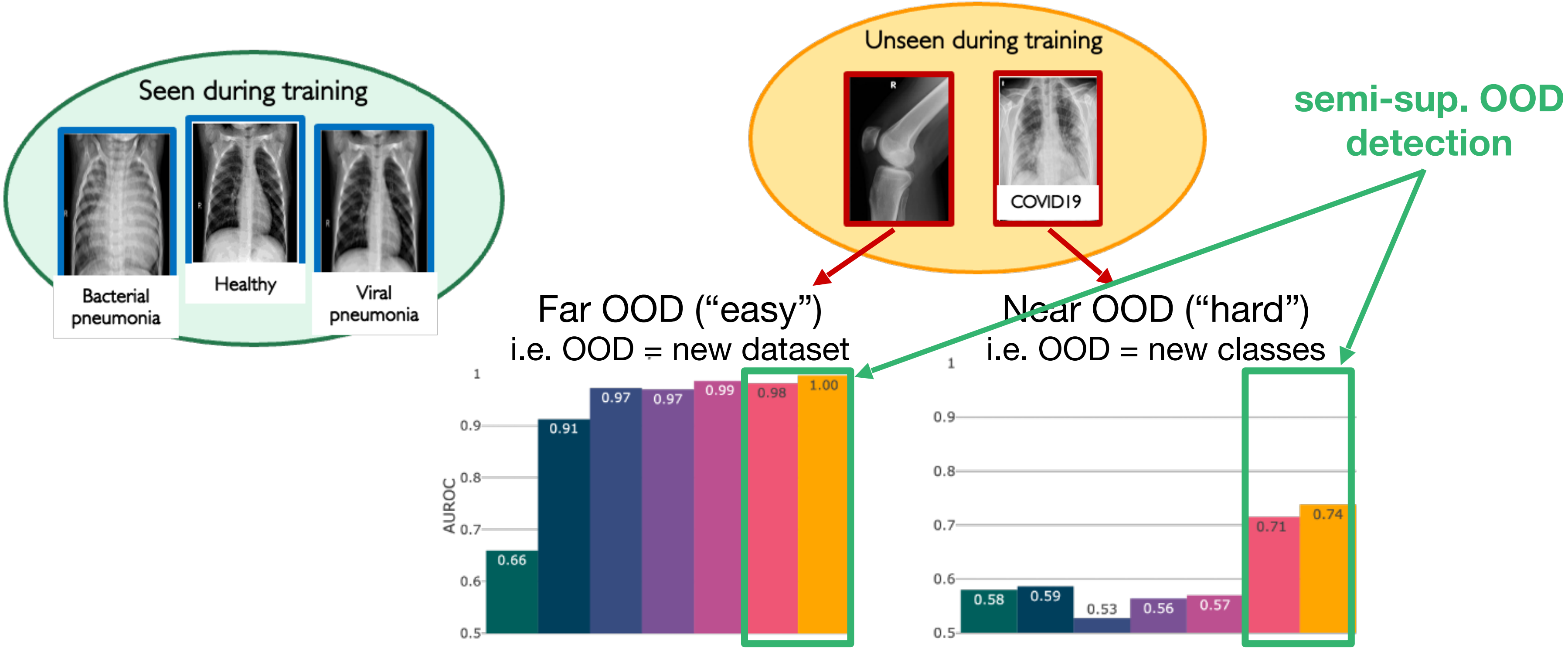
Near OOD ("hard")
i.e. OOD = new classes



Semi-supervised OOD detection
Performance on near OOD data

Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland



Limitations of semi-supervised OOD detection

Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland

Challenge #1: not suitable for real-time applications

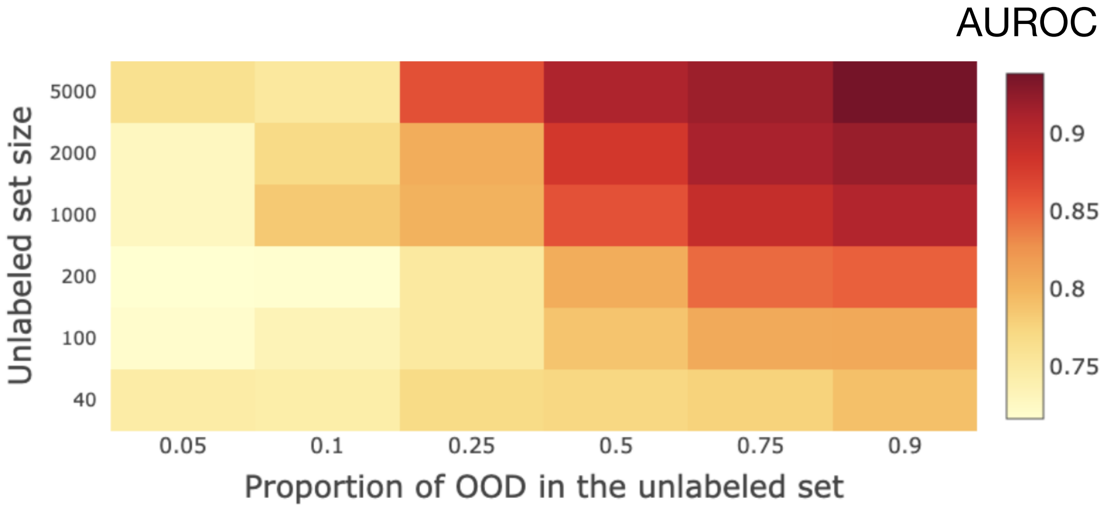
Limitations of semi-supervised OOD detection

Semi-supervised novelty detection using ensembles with regularized disagreement

Alexandru Țifrea, Eric Stavarache, Fanny Yang
Department of Computer Science
ETH Zurich, Switzerland

Challenge #1: not suitable for real-time applications

Challenge #2: not suitable for anomaly detection i.e. singleton outliers



Outlook and future directions

Summary: Trustworthy ML under imperfect data

Figure sources: https://arxiv.org/pdf/1912.02292

Summary: Trustworthy ML under imperfect data

If accuracy alone is the goal:

benign overfitting of label noise

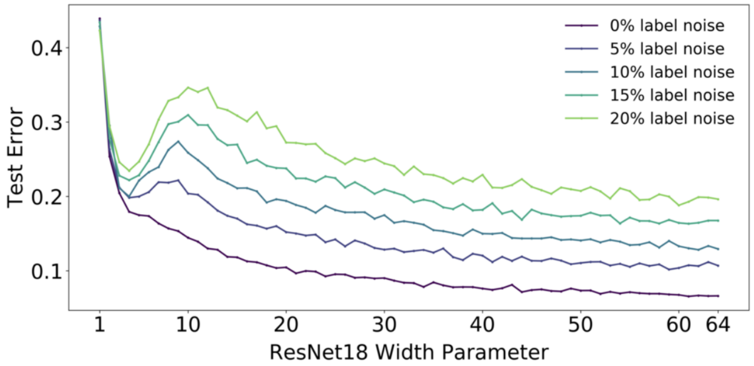
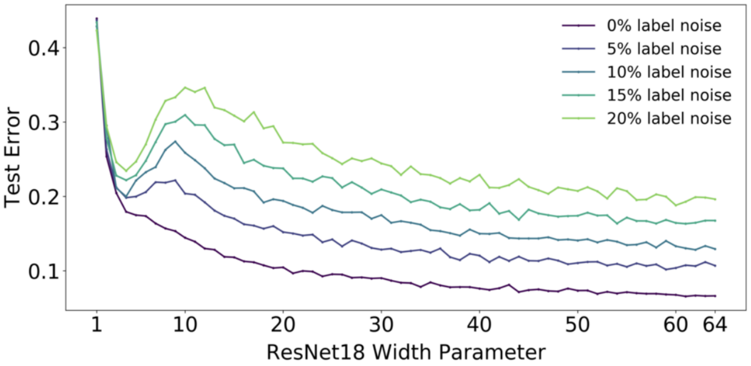


Figure sources: https://arxiv.org/pdf/1912.02292

Summary: Trustworthy ML under imperfect data



If accuracy alone is the goal:

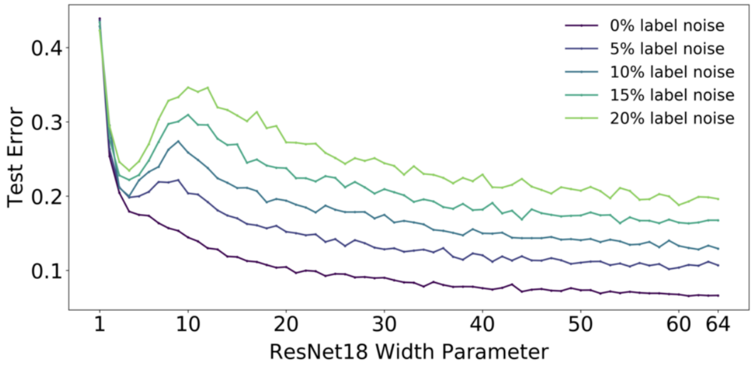
benign overfitting of label noise

If we care about trustworthiness:

This tutorial: Several examples of trustworthy learning algorithms that work well under label noise, missing data etc.

Figure sources: https://arxiv.org/pdf/1912.02292

Summary: Trustworthy ML under imperfect data



If accuracy alone is the goal:

benign overfitting of label noise

If we care about trustworthiness:

This tutorial: Several examples of trustworthy learning algorithms that work well under label noise, missing data etc.

Open questions

- What other data-related limitations do existing trustworthy algorithms suffer from?
- How to improve trustworthiness in other difficult problem settings?

Figure sources: https://arxiv.org/pdf/1912.02292

Summary: Trustworthy ML and unlabeled data

Summary: Trustworthy ML and unlabeled data

If accuracy alone is the goal:

SSL cannot be simultaneously better than both unsupervised and supervised learning

Can semi-supervised learning use all the data effectively? A lower bound perspective

Alexandru Ţifrea*
ETH Zurich
alexandru.tifrea@inf.ethz.ch

Gizem Yüce*
EPFL
gizem.yuce@epfl.ch

Amartya Sanyal
Max Planck Institute for Intelligent Systems, Tübingen
amsa@di.ku.dk

Fanny Yang
ETH Zurich
fan.yan@inf.ethz.ch

Summary: Trustworthy ML and unlabeled data

If accuracy alone is the goal:

SSL cannot be simultaneously better than both unsupervised and supervised learning

Can semi-supervised learning use all the data effectively? A lower bound perspective

Alexandru Ţifrea* ETH Zurich alexandru.tifrea@inf.ethz.ch

Gizem Yüce* EPFL gizem.yuce@epfl.ch

Amartya Sanyal Max Planck Institute for Intelligent Systems, Tübingen amsa@di.ku.dk

Fanny Yang ETH Zurich fan.yan@inf.ethz.ch

If we care about trustworthiness:

This tutorial: Several examples where unlabeled data can help to overcome limitations of supervised learning.

Summary: Trustworthy ML and unlabeled data

If accuracy alone is the goal:

SSL cannot be simultaneously better than both unsupervised and supervised learning

Can semi-supervised learning use all the data effectively? A lower bound perspective

Alexandru Ţifrea*
ETH Zurich
alexandru.tifrea@inf.ethz.ch

Gizem Yüce*
EPFL
gizem.yuce@epfl.ch

Amartya Sanyal
Max Planck Institute for Intelligent Systems, Tübingen
amsa@di.ku.dk

Fanny Yang
ETH Zurich
fan.yan@inf.ethz.ch

If we care about trustworthiness:

This tutorial: Several examples where unlabeled data can help to overcome limitations of supervised learning.

Open questions

- How fundamental are the improvements to trustworthiness due to unlabeled data?
- What other kinds of (potentially noisy) side information can be used to improve trustworthiness?

Thank you!