FR-Train:
A Mutual Information-based Fair and Robust Training

Yuji Roh, Kangwook Lee, Steven E. Whang, Changho Suh

Yuji Roh, Data Intelligence Lab, KAIST

Trustworthy AI

"AI has significant potential to help solve challenging problems,
including by advancing medicine, understanding language, and fueling scientific discovery.
To realize that potential, it's critical that AI is used and developed **responsibly**. "

- Google AI, 2020

"Moving forward, "build for performance" will not suffice as an AI design paradigm.
We must learn how to build, evaluate and monitor for **trust**."

- IBM Trusting AI, 2020

Trustworthy AI

Fairness

Robustness

Explainability

Value Alignment

Transparency & Accountability

Trustworthy AI

Data-related

Fairness

Robustness

Explainability

Value Alignment

Transparency & Accountability

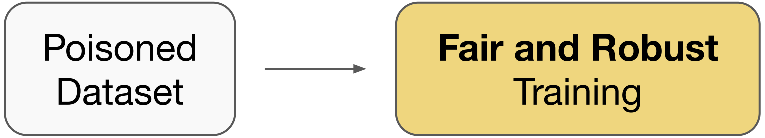
Two approaches

- **Two-step** approach: Sanitize data -> Fair training
  Downside: very difficult to “decouple” poisoning and bias

Two approaches

- **Two-step** approach: Sanitize data -> Fair training
  Downside: very difficult to “decouple” poisoning and bias

- **Holistic** approach: Fair & Robust training
  Performing the two operations along with model training results in much better performance



Two approaches

- **Two-step** approach: Sanitize data -> Fair training
  Downside: very difficult to "decouple" poisoning and bias

Poisoned Dataset → Sanitization → Fair Training

FR-Train

- **Holistic** approach: Fair & Robust training
  Performing the two operations along with model training results in much better performance

Poisoned Dataset → **Fair and Robust** Training

Trustworthy AI

Data-related

Fairness

Robustness

Explainability

Value Alignment

Transparency & Accountability

Fairness

X	Feature
Y	Label
Z	Group attribute
$\hat{Y}$	Predicted label

- A machine learning model learns bias and discriminations in the data
- The fairness of a (binary) classifier can be defined in various ways:

Demographic Parity

(⇔ Disparate Impact)

$$\mathbb{P}(\hat{Y}=1|Z=0) \approx \mathbb{P}(\hat{Y}=1|Z=1)$$

Equalized Odds

$$\mathbb{P}(\hat{Y}=1|Z=0,Y=1) \approx \mathbb{P}(\hat{Y}=1|Z=1,Y=1)$$
$$\mathbb{P}(\hat{Y}=1|Z=0,Y=0) \approx \mathbb{P}(\hat{Y}=1|Z=1,Y=0)$$

- The level of fairness can be measured as a ratio or difference

Fairness

X	Feature
Y	Label
Z	Group attribute
$\hat{Y}$	Predicted label

- A machine learning model learns bias and discriminations in the data
- The fairness of a (binary) classifier can be defined in various ways:

Demographic Parity
(⇔ **Disparate Impact**)

$$\mathbb{P}(\hat{Y}=1|Z=0) \approx \mathbb{P}(\hat{Y}=1|Z=1)$$

Equalized Odds

$$\mathbb{P}(\hat{Y}=1|Z=0,Y=1) \approx \mathbb{P}(\hat{Y}=1|Z=1,Y=1)$$
$$\mathbb{P}(\hat{Y}=1|Z=0,Y=0) \approx \mathbb{P}(\hat{Y}=1|Z=1,Y=0)$$

In this talk

- The level of fairness can be measured as a ratio or difference

$$DI := min(\frac{\mathbb{P}(\hat{Y}=1|Z=0)}{\mathbb{P}(\hat{Y}=1|Z=1)}, \frac{\mathbb{P}(\hat{Y}=1|Z=1)}{\mathbb{P}(\hat{Y}=1|Z=0)})$$

Robustness

- Datasets are **easy to publish** nowadays, but as a result **easy to “poison”** as well
  - Poison = noisy, subjective, or even adversarial

kaggle

Google Dataset Search

  - Attacker’s goal : Increase the test loss by poisoning data
  - Defender’s goal : Train a classifier with small test loss
- Already a serious issue in federated learning

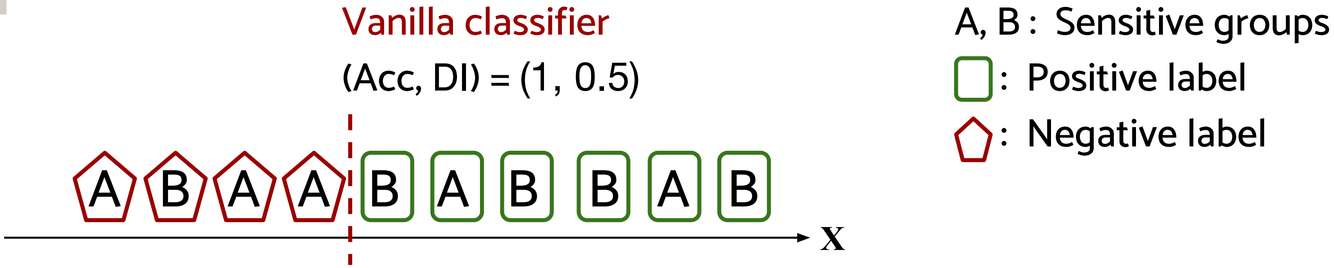
Fairness + Robustness

What happens if we just apply a fairness-aware algorithm on a poisoned dataset?

- May result in a strictly **suboptimal** (accuracy, fairness) than vanilla training

Motivating example

Clean

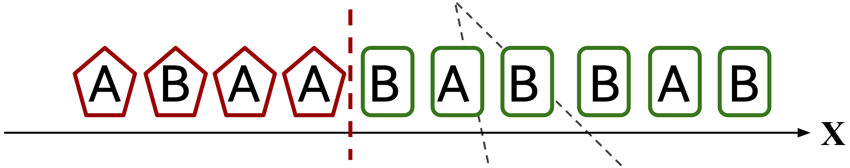


Motivating example

Clean

Vanilla classifier

(Acc, DI) = (1, 0.5)



$$\mathbb{P}(\hat{Y} = 1 | Z = A) = \frac{2}{5} = 0.4$$

$$\mathbb{P}(\hat{Y} = 1 | Z = B) = \frac{4}{5} = 0.8$$

$$DI = \min\left(\frac{0.4}{0.8}, \frac{0.8}{0.4}\right) = 0.5$$

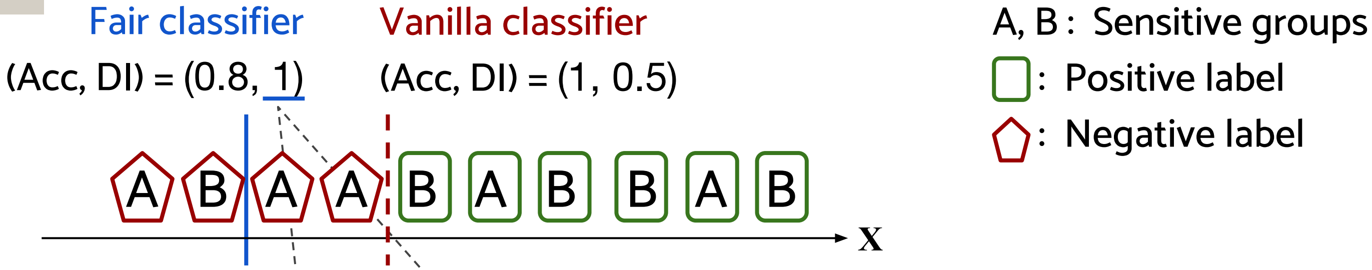
A, B : Sensitive groups

□ : Positive label

⬠ : Negative label

Motivating example

Clean

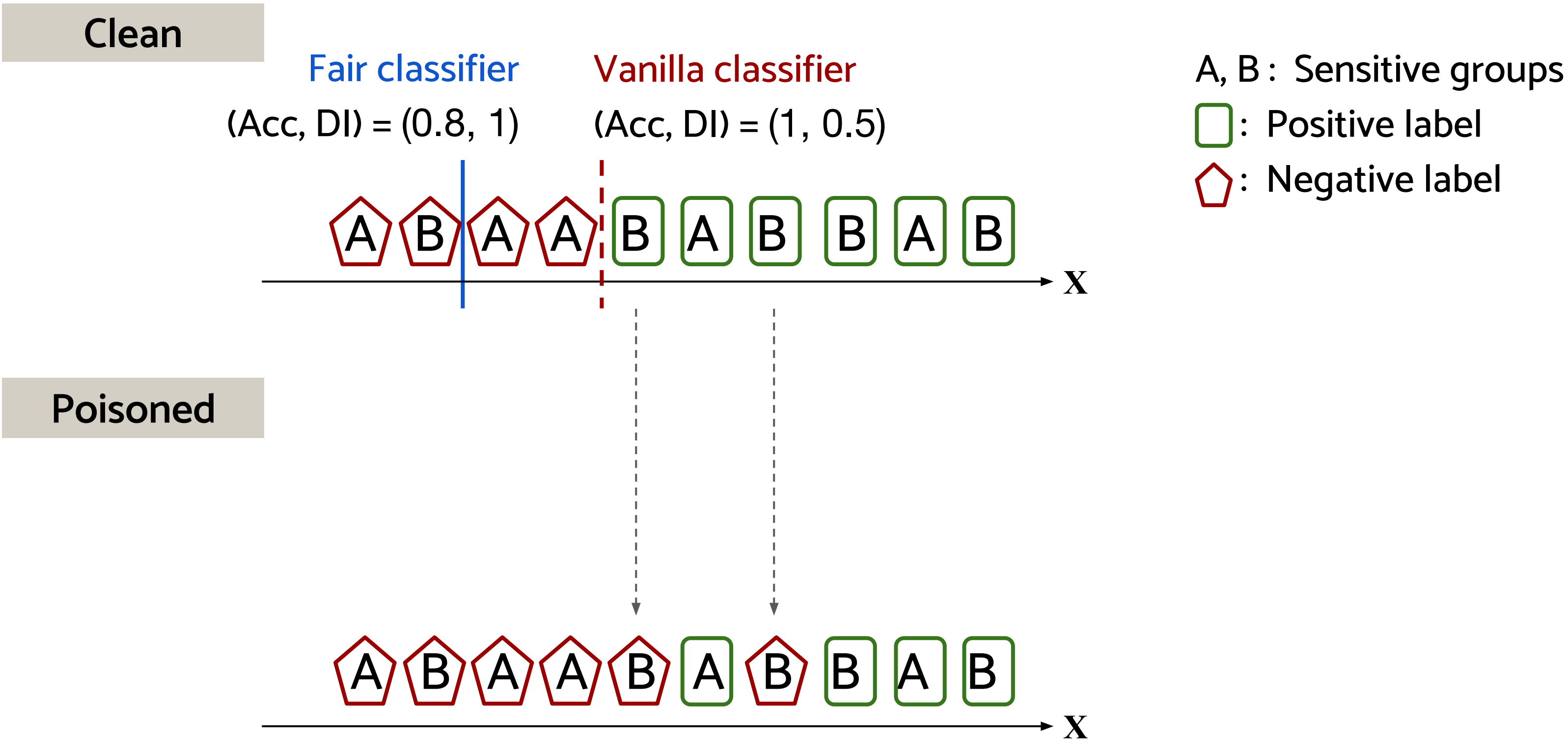


$$\mathbb{P}(\hat{Y} = 1 | Z = A) = \frac{4}{5} = 0.8$$

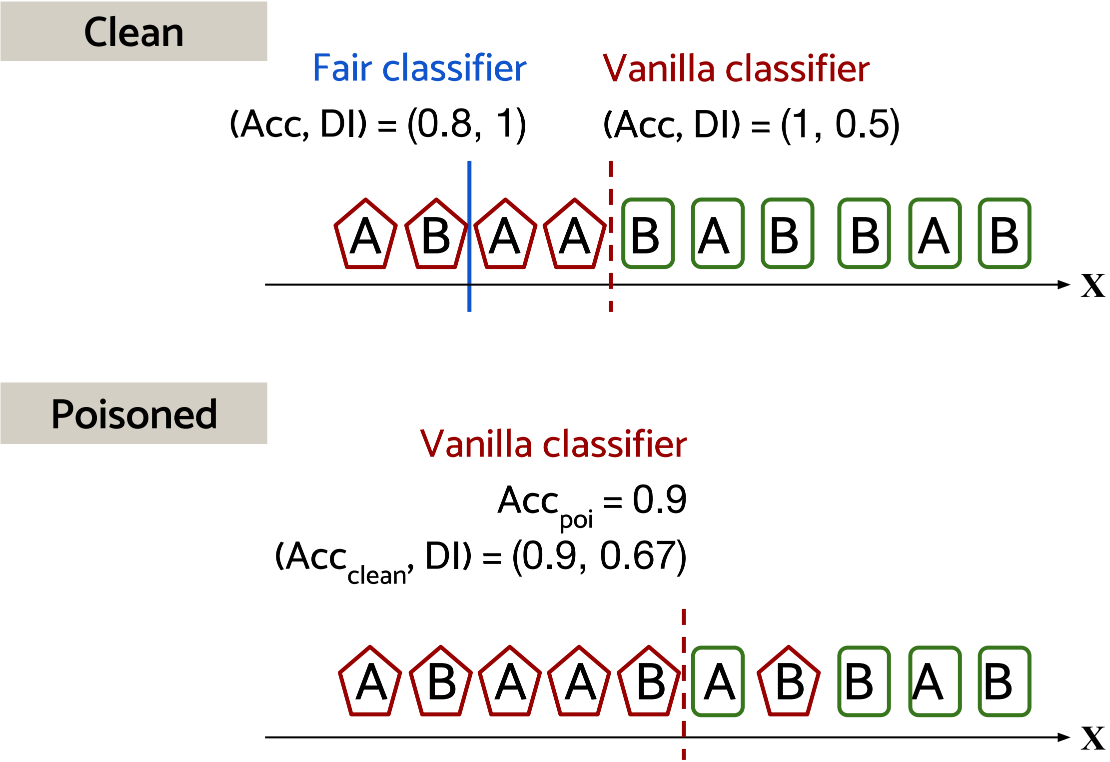
$$\mathbb{P}(\hat{Y} = 1 | Z = B) = \frac{4}{5} = 0.8$$

$$DI = \min(\frac{0.8}{0.8}, \frac{0.8}{0.8}) = 1$$

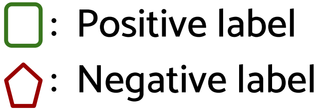
Motivating example



Motivating example



A, B : Sensitive groups



Acc : ↓

DI : ↑

Motivating example

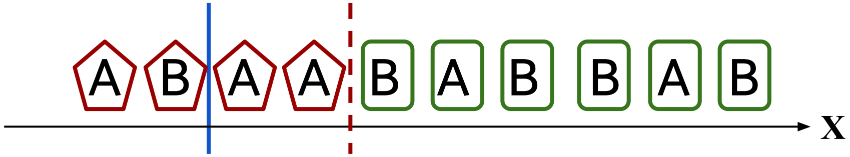
Clean

Fair classifier

(Acc, DI) = (0.8, 1)

Vanilla classifier

(Acc, DI) = (1, 0.5)



A, B : Sensitive groups

: Positive label

: Negative label

Poisoned

Vanilla classifier

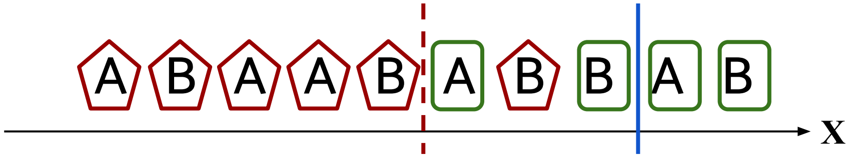
$Acc_{poi} = 0.9$

$(Acc_{clean}, DI) = (0.9, 0.67)$

Fair classifier

$Acc_{poi} = 0.8$

$(Acc_{clean}, DI) = (0.6, 1)$



Acc : ↓

DI : —

Suboptimal

Fairness + Robustness

What happens if we just apply a fairness-aware algorithm on a poisoned dataset?

- May result in a strictly **suboptimal** (accuracy, fairness) than vanilla training

We need a holistic approach to fair and robust training. **FR-Train!**

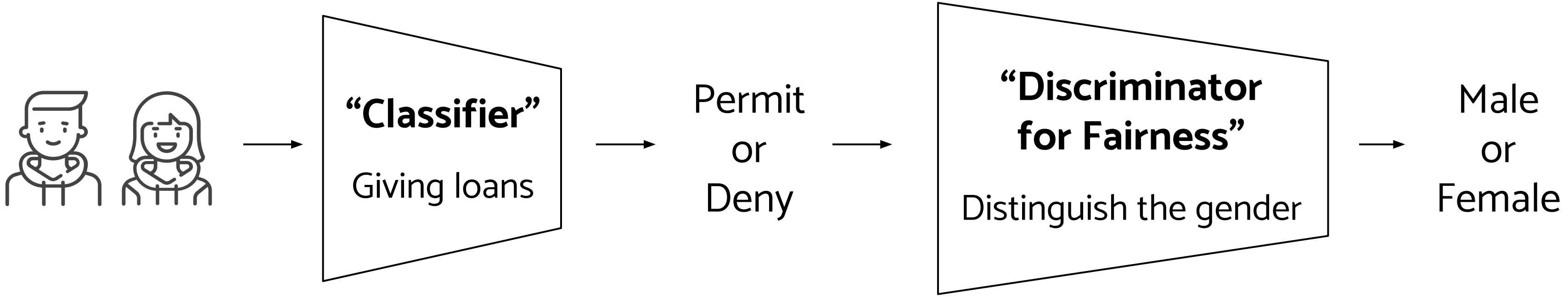
FR-Train - Main contributions

- FR-Train is a **holistic framework** for fair and robust training

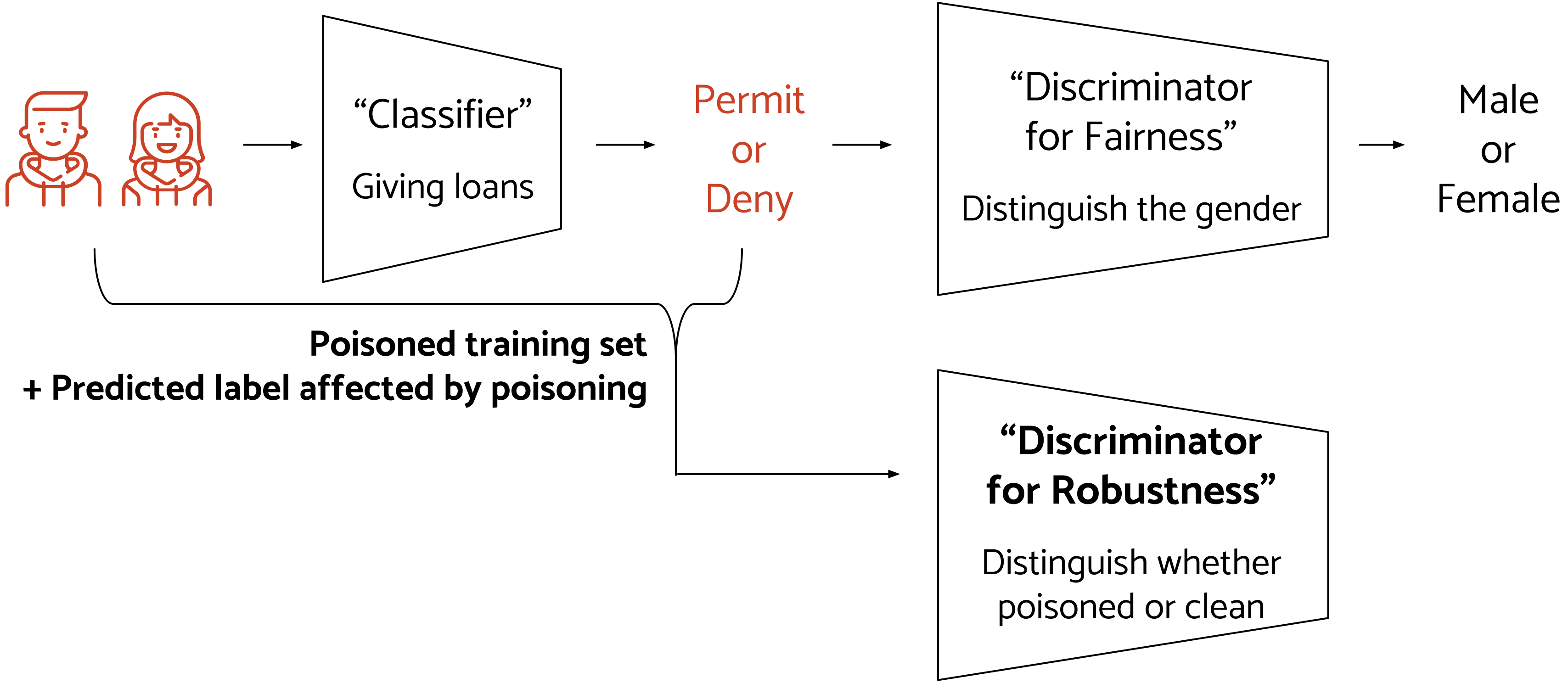
- Extends a state-of-the-art fairness-only method called Adversarial Debiasing
  - Provides a novel mutual information (MI)-based interpretation of adversarial learning
  - Adds a robust discriminator that uses a small clean validation set for data sanitization

- We also propose crowdsourcing methods for constructing a clean validation set

FR-Train

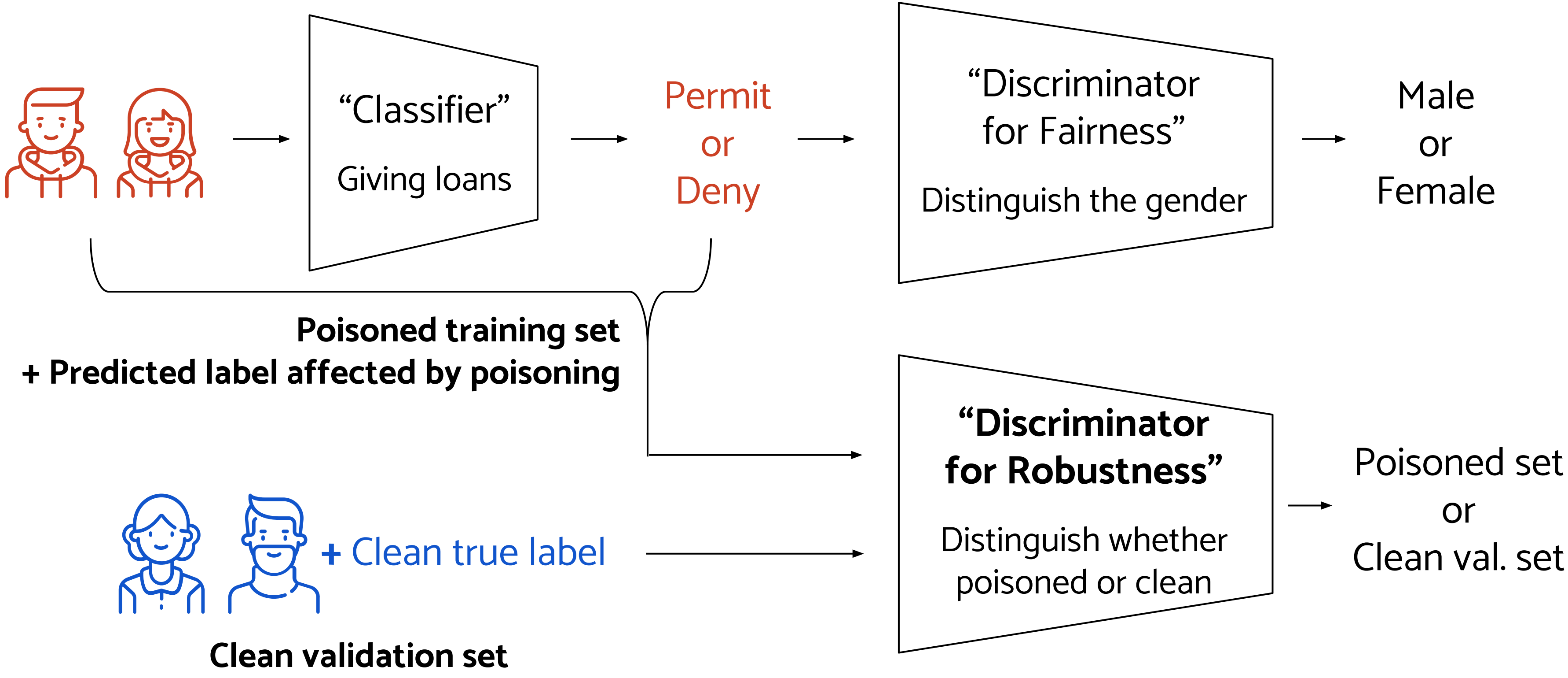

“Classifier”
Giving loans
Permit
or
Deny
“Discriminator
for Fairness”
Distinguish the gender
Male
or
Female

FR-Train

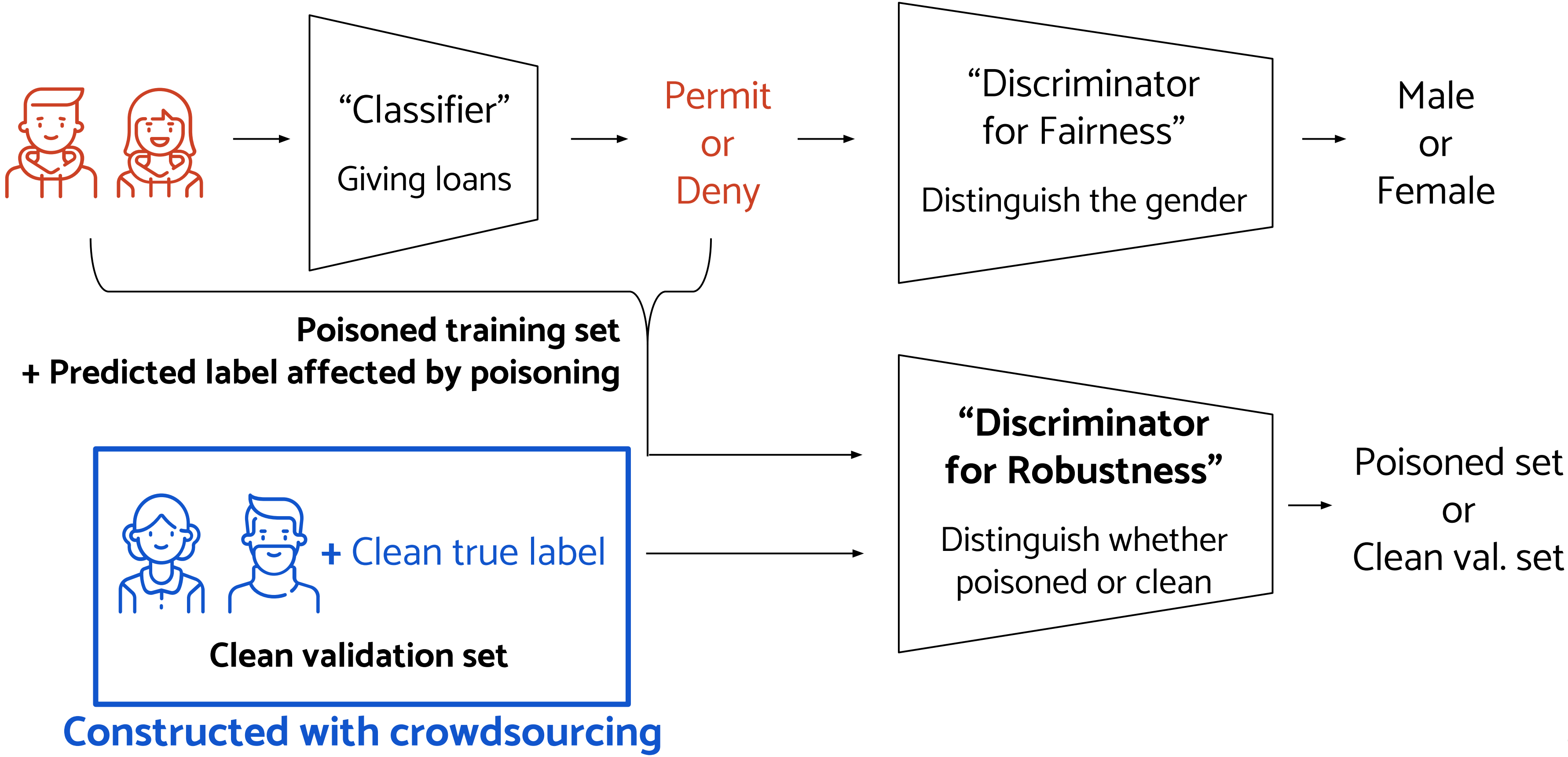


“Classifier”
Giving loans
Permit
or
Deny
“Discriminator
for Fairness”
Distinguish the gender
Male
or
Female
Poisoned training set
+ Predicted label affected by poisoning
“Discriminator
for Robustness”
Distinguish whether
poisoned or clean

FR-Train

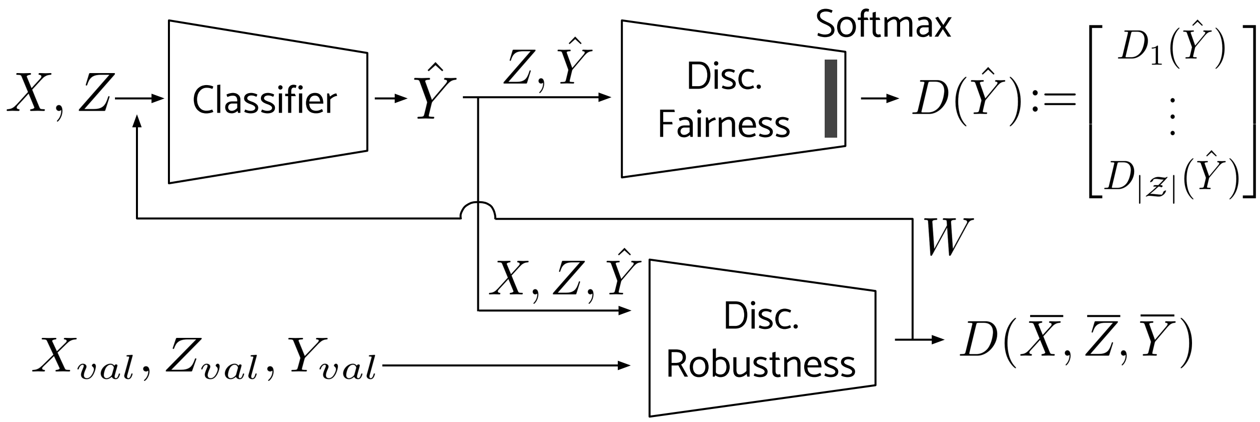

“Classifier”
Giving loans
Permit
or
Deny
“Discriminator
for Fairness”
Distinguish the gender
Male
or
Female
Poisoned training set
+ Predicted label affected by poisoning
“Discriminator
for Robustness”
Distinguish whether
poisoned or clean
Poisoned set
or
Clean val. set
+ Clean true label
Clean validation set

FR-Train



“Classifier”
Giving loans
Permit
or
Deny
“Discriminator
for Fairness”
Distinguish the gender
Male
or
Female
Poisoned training set
+ Predicted label affected by poisoning
+ Clean true label
Clean validation set
Constructed with crowdsourcing
“Discriminator
for Robustness”
Distinguish whether
poisoned or clean
Poisoned set
or
Clean val. set

Mutual information-based interpretation



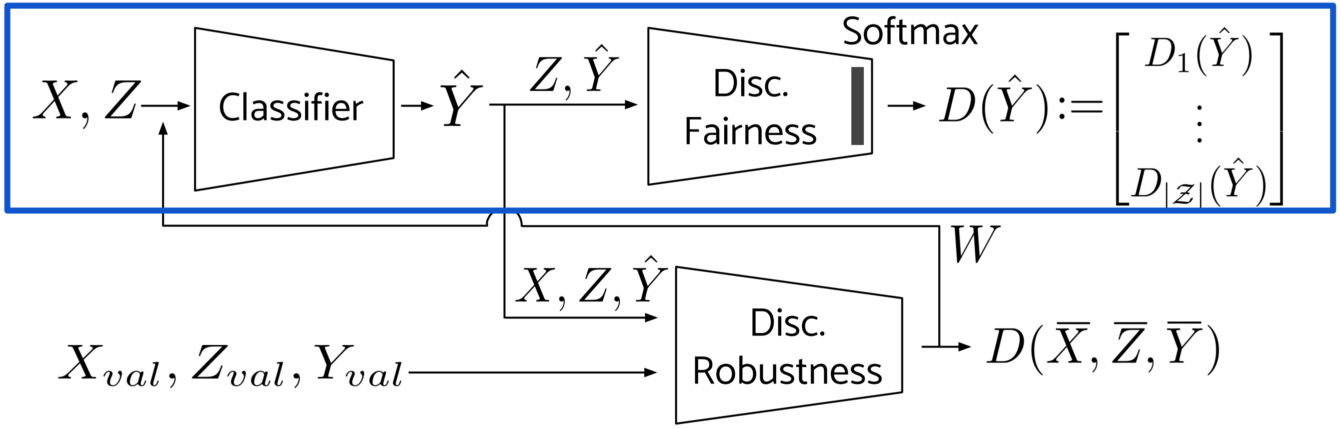
Theorem 1 - Fairness

$$I(Z; \hat{Y}) = \max_{D_z(\hat{y}): \sum_z D_z(\hat{y})=1,\ \forall \hat{y}} \sum_{z \in \mathcal{Z}} P_Z(z) \mathbb{E}_{P_{\hat{Y}|z}} \left[\log D_z(\hat{Y}) \right] + H(Z)$$

Theorem 2 - Robustness

$$I(V; \overline{X}, \overline{Z}, \overline{Y}) = \max_{D_v(x,z,y): \sum_v D_v(x,z,y)=1,\ \forall (x,z,y)} \sum_{v \in \mathcal{V}} P_V(v) \mathbb{E}_{P_{\overline{X}, \overline{Z}, \overline{Y}|v}} \left[\log D_v(\overline{X}, \overline{Z}, \overline{Y}) \right] + H(V)$$

Mutual information-based interpretation



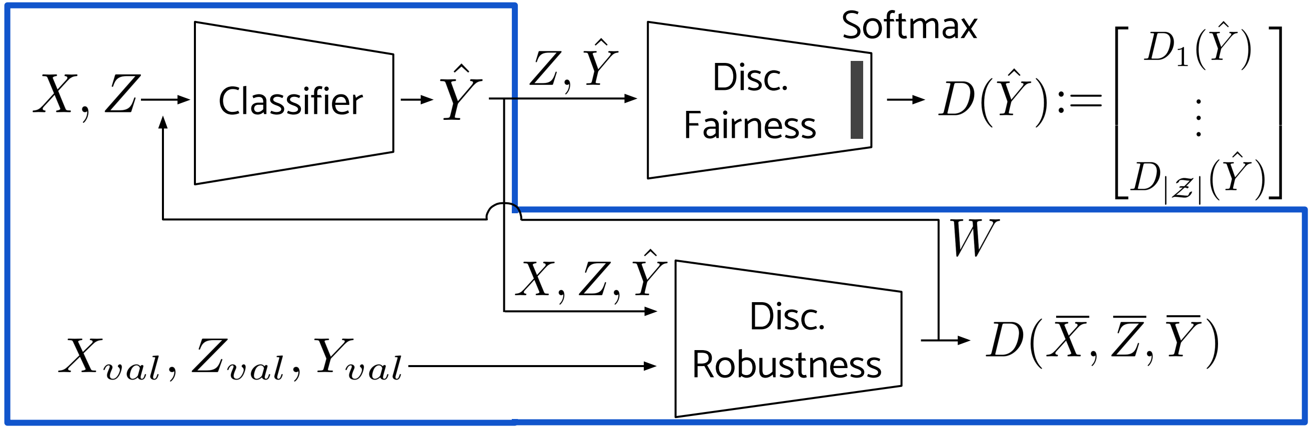
Theorem 1 - Fairness

$$I(Z;\hat{Y}) = \max_{D_z(\hat{y}):\sum_z D_z(\hat{y})=1,\ \forall \hat{y}} \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log D_z(\hat{Y})\right] + H(Z)$$

Theorem 2 - Robustness

$$I(V;\overline{X},\overline{Z},\overline{Y}) = \max_{D_v(x,z,y):\sum_v D_v(x,z,y)=1,\ \forall (x,z,y)} \sum_{v\in\mathcal{V}} P_V(v)\mathbb{E}_{P_{\overline{X},\overline{Z},\overline{Y}|v}}\left[\log D_v(\overline{X},\overline{Z},\overline{Y})\right] + H(V)$$

Mutual information-based interpretation



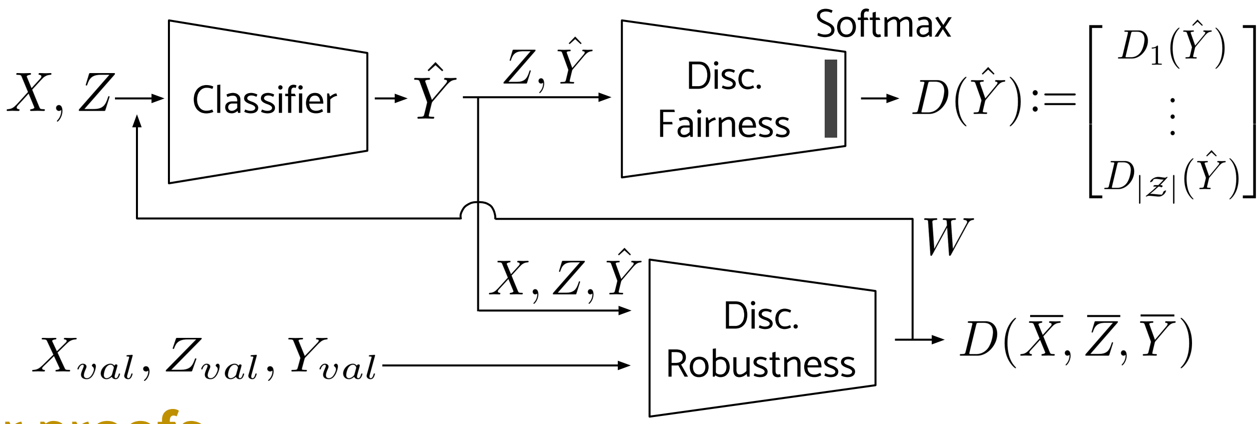
Theorem 1 - Fairness

$$I(Z;\hat{Y}) = \max_{D_z(\hat{y}):\sum_z D_z(\hat{y})=1,\ \forall \hat{y}} \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log D_z(\hat{Y})\right] + H(Z)$$

Theorem 2 - Robustness

$$I(V;\overline{X},\overline{Z},\overline{Y}) = \max_{D_v(x,z,y):\sum_v D_v(x,z,y)=1,\ \forall (x,z,y)} \sum_{v\in\mathcal{V}} P_V(v)\mathbb{E}_{P_{\overline{X},\overline{Z},\overline{Y}|v}}\left[\log D_v(\overline{X},\overline{Z},\overline{Y})\right] + H(V)$$

Mutual information-based interpretation



See paper for proofs

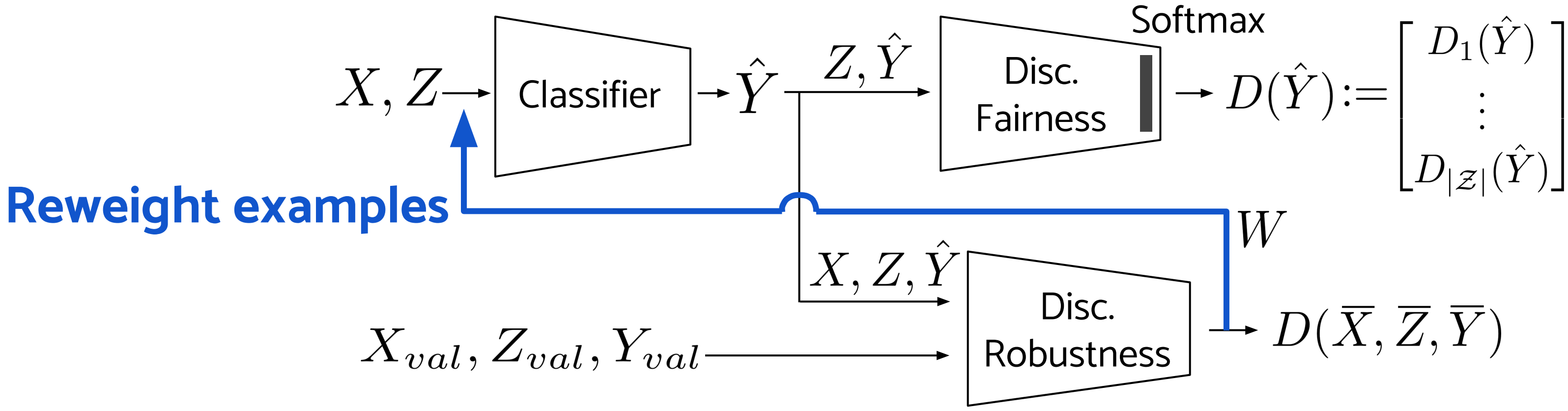
Theorem 1 - Fairness

$$I(Z;\hat{Y}) = \max_{D_z(\hat{y}):\sum_z D_z(\hat{y})=1,\ \forall \hat{y}} \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log D_z(\hat{Y})\right] + H(Z)$$

Theorem 2 - Robustness

$$I(V;\overline{X},\overline{Z},\overline{Y}) = \max_{D_v(x,z,y):\sum_v D_v(x,z,y)=1,\ \forall (x,z,y)} \sum_{v\in\mathcal{V}} P_V(v)\mathbb{E}_{P_{\overline{X},\overline{Z},\overline{Y}|v}}\left[\log D_v(\overline{X},\overline{Z},\overline{Y})\right] + H(V)$$

Mutual information-based interpretation



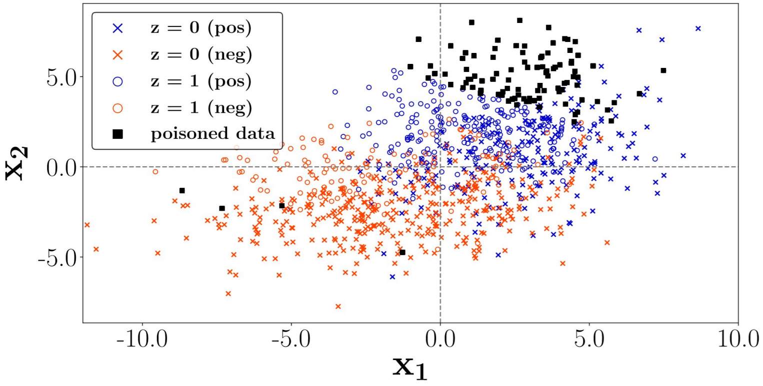
Theorem 1 - Fairness

$$I(Z;\hat{Y}) = \max_{D_z(\hat{y}):\sum_z D_z(\hat{y})=1,\ \forall \hat{y}} \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log D_z(\hat{Y})\right] + H(Z)$$

Theorem 2 - Robustness

$$I(V;\overline{X},\overline{Z},\overline{Y}) = \max_{D_v(x,z,y):\sum_v D_v(x,z,y)=1,\ \forall (x,z,y)} \sum_{v\in\mathcal{V}} P_V(v)\mathbb{E}_{P_{\overline{X},\overline{Z},\overline{Y}|v}}\left[\log D_v(\overline{X},\overline{Z},\overline{Y})\right] + H(V)$$

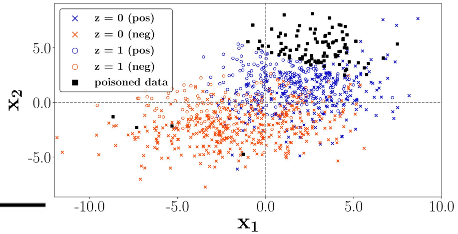
Experimental setting



- Synthetic data
  - Poisoning (label flipping): 10% of training data
  - Validation set: 10% of training data

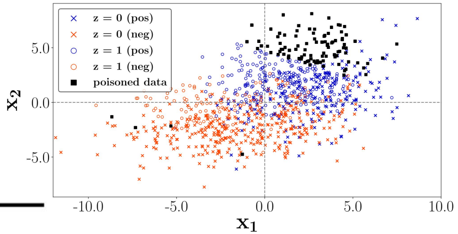
- Real data (results in paper)
  - **COMPAS**: Predict recidivism in two years for criminals
  - **AdultCensus**: Predict whether annual income > $50K or not
  - Poisoning: 10% of training data
  - Validation set: 5% of training data

Synthetic data results



	Method	Clean data DI	Clean data Acc.	Poisoned data DI	Poisoned data Acc.
Fair-only algorithms	FC	.822	.806	.831 (1.1% $\uparrow$)	.760 (5.7% $\downarrow$)
	LBC	.819	.760	.827 (1.0% $\uparrow$)	.715 (5.9% $\downarrow$)
	AD	.807	.811	.834 (3.4% $\uparrow$)	.769 (5.2% $\downarrow$)
Two-step approach : Data sanitization + Fair training	RML+FC	.822	.806	.802 (2.4% $\downarrow$)	.529 (34.% $\downarrow$)
	RML+LBC	.819	.760	.810 (1.1% $\downarrow$)	.752 (1.1% $\downarrow$)
	RML+AD	.807	.811	.808 (0.1% $\uparrow$)	.756 (6.8% $\downarrow$)
Logistic regression —	LR	.409	.885	.446 (9.1% $\uparrow$)	.819 (7.5% $\downarrow$)
Data sanitization — using clean val. set	RML	.471	.876	.395 (16.% $\downarrow$)	.869 (0.8% $\downarrow$)
	FR-Train	**.818**	**.807**	**.827 (1.1% $\uparrow$)**	**.814 (0.9% $\uparrow$)**

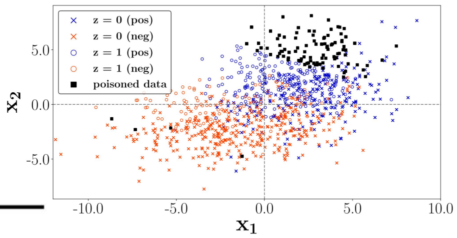
Synthetic data results



	Method	Clean data DI	Clean data Acc.	Poisoned data DI	Poisoned data Acc.
Fair-only algorithms	FC	.822	.806	.831 (1.1% $\uparrow$)	.760 (5.7% $\downarrow$)
	LBC	.819	.760	.827 (1.0% $\uparrow$)	.715 (5.9% $\downarrow$)
	AD	.807	.811	.834 (3.4% $\uparrow$)	.769 (5.2% $\downarrow$)
Two-step approach : Data sanitization + Fair training	RML+FC	.822	.806	.802 (2.4% $\downarrow$)	.529 (34.% $\downarrow$)
	RML+LBC	.819	.760	.810 (1.1% $\downarrow$)	.752 (1.1% $\downarrow$)
	RML+AD	.807	.811	.808 (0.1% $\uparrow$)	.756 (6.8% $\downarrow$)
Logistic regression —	LR	.409	.885	.446 (9.1% $\uparrow$)	.819 (7.5% $\downarrow$)
Data sanitization using clean val. set —	RML	.471	.876	.395 (16.% $\downarrow$)	.869 (0.8% $\downarrow$)
	FR-Train	**.818**	**.807**	**.827 (1.1% $\uparrow$)**	**.814 (0.9% $\uparrow$)**

Low accuracy

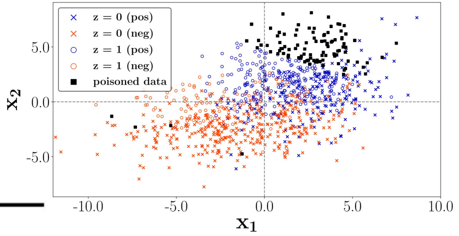
Synthetic data results



	Method	Clean data DI	Clean data Acc.	Poisoned data DI	Poisoned data Acc.
Fair-only algorithms	FC	.822	.806	.831 (1.1% $\uparrow$)	.760 (5.7% $\downarrow$)
	LBC	.819	.760	.827 (1.0% $\uparrow$)	.715 (5.9% $\downarrow$)
	AD	.807	.811	.834 (3.4% $\uparrow$)	.769 (5.2% $\downarrow$)
Two-step approach : Data sanitization + Fair training	RML+FC	.822	.806	.802 (2.4% $\downarrow$)	.529 (34.% $\downarrow$)
	RML+LBC	.819	.760	.810 (1.1% $\downarrow$)	.752 (1.1% $\downarrow$)
	RML+AD	.807	.811	.808 (0.1% $\uparrow$)	.756 (6.8% $\downarrow$)
Logistic regression	LR	.409	.885	.446 (9.1% $\uparrow$)	.819 (7.5% $\downarrow$)
Data sanitization using clean val. set	RML	.471	.876	.395 (16.% $\downarrow$)	.869 (0.8% $\downarrow$)
	FR-Train	**.818**	**.807**	**.827 (1.1% $\uparrow$)**	**.814 (0.9% $\uparrow$)**

Low fairness

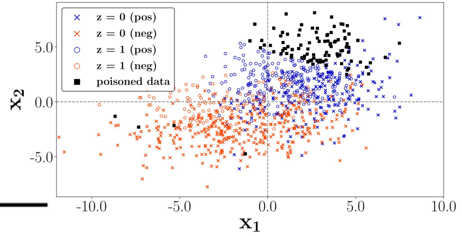
Synthetic data results



	Method	Clean data		Poisoned data	
		DI	Acc.	DI	Acc.
Fair-only algorithms	FC	.822	.806	.831 (1.1% ↑)	.760 (5.7% ↓)
	LBC	.819	.760	.827 (1.0% ↑)	.715 (5.9% ↓)
	AD	.807	.811	.834 (3.4% ↑)	.769 (5.2% ↓)
Two-step approach : Data sanitization + Fair training	RML+FC	.822	.806	.802 (2.4% ↓)	.529 (34.% ↓)
	RML+LBC	.819	.760	.810 (1.1% ↓)	.752 (1.1% ↓)
	RML+AD	.807	.811	.808 (0.1% ↑)	.756 (6.8% ↓)
Logistic regression	LR	.409	.885	.446 (9.1% ↑)	.819 (7.5% ↓)
Data sanitization using clean val. set	RML	.471	.876	.395 (16.% ↓)	.869 (0.8% ↓)
	FR-Train	**.818**	**.807**	**.827 (1.1% ↑)**	**.814 (0.9% ↑)**

Also low accuracy

Synthetic data results



	Method	Clean data DI	Clean data Acc.	Poisoned data DI	Poisoned data Acc.
Fair-only algorithms	FC	.822	.806	.831 (1.1% ↑)	.760 (5.7% ↓)
	LBC	.819	.760	.827 (1.0% ↑)	.715 (5.9% ↓)
	AD	.807	.811	.834 (3.4% ↑)	.769 (5.2% ↓)
Two-step approach : Data sanitization + Fair training	RML+FC	.822	.806	.802 (2.4% ↓)	.529 (34.% ↓)
	RML+LBC	.819	.760	.810 (1.1% ↓)	.752 (1.1% ↓)
	RML+AD	.807	.811	.808 (0.1% ↑)	.756 (6.8% ↓)
Logistic regression —	LR	.409	.885	.446 (9.1% ↑)	.819 (7.5% ↓)
Data sanitization — using clean val. set	RML	.471	.876	.395 (16.% ↓)	.869 (0.8% ↓)
	FR-Train	**.818**	**.807**	**.827 (1.1% ↑)**	**.814 (0.9% ↑)**

Holistic approach = high fairness & accuracy

Takeaways

- Trustworthy AI needs both fair and robust training

- However, addressing fairness and robustness separately is suboptimal

- FR-Train is a **holistic framework for trustworthy AI** performing fair and robust training
  - Mutual information-based interpretation of adversarial learning
  - Novel architecture that enjoys the synergistic effect of fair and robust discriminators
  - Requires a small clean validation set, which can be constructed using crowdsourcing

- Lots of open problems
  - Without clean validation set
  - Other poisoning
  - Algorithm stability

Takeaways

- Trustworthy AI needs both fair and robust training
- However, addressing fairness and robustness separately is suboptimal
- FR-Train is a **holistic framework for trustworthy AI** performing fair and robust training
  - Mutual information-based interpretation of adversarial learning
  - Novel architecture that enjoys the synergistic effect of fair and robust discriminators
  - Requires a small clean validation set, which can be constructed using crowdsourcing
- Lots of open problems
  - Without clean validation set
  - Other poisoning
  - Algorithm stability

Thank you :)