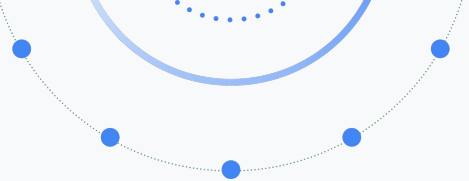


Fairness and Representation Learning

Moustapha Cisse & Sanmi Koyejo



Dear colleagues, the story you are about to hear is true. Only the names have been changed to protect innocent computer scientists...

A manager oversees several teams, all are using the same data to build predictive models for different products. The manager seeks to ensure both **fairness and accuracy** across the products.

Each team is solving a different prediction task.

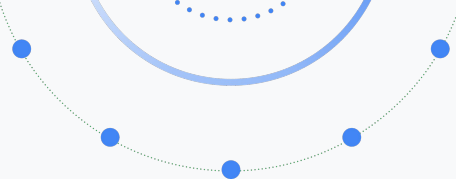
There is no company policy on fairness, thus no shared guidelines.

- Team **alpha** is fully focused on accuracy, but is oblivious (neighbors say they are apathetic) about fairness issues.
- Team **beta**, team **nu** and team **gamma** are all interested in fairness. Each team is really excited to implement this and has read the literature, but each team has selected different fairness definitions.
- Team **zeta** would like to improve the fairness of their predictions, but has no idea how to incorporate or measure fairness.
- The **manager** has decided to independently verify that all released products are fair

A manager oversees several teams, all are using the same data to build predictive models for different products. The manager seeks to ensure both **fairness and accuracy** across the products.

Challenges:

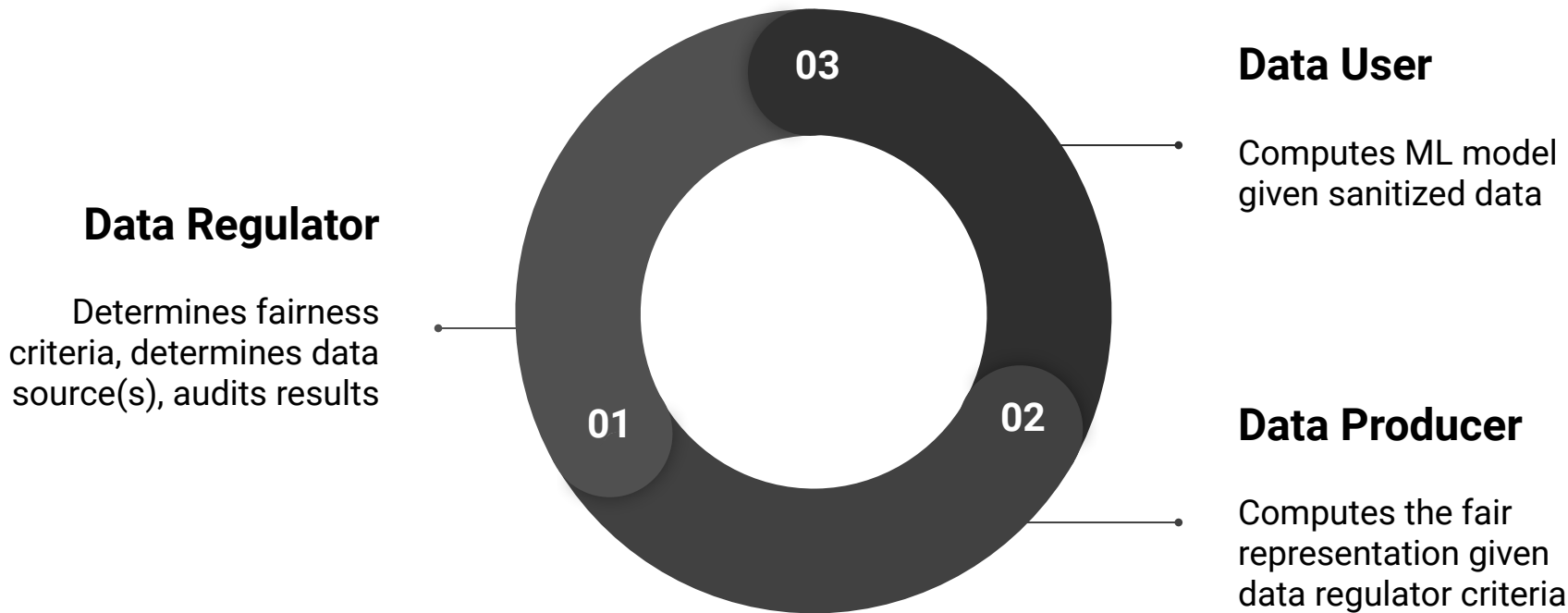
- Some teams do not have the expertise (or interest) to design fairer models.
- Different teams use different definitions of fairness.
- Incorporating fairness can have different impacts on the performance of the models across products.
- Auditing all the predictive models for fairness can be challenging when each team has its own recipe.



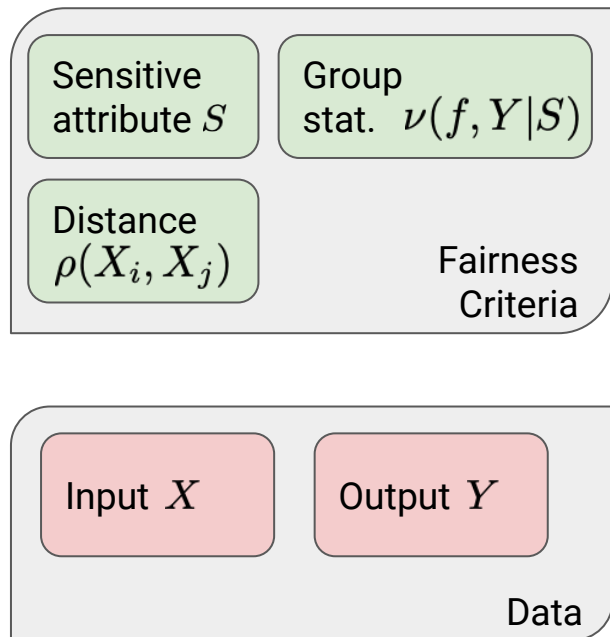
This tutorial will outline how representation learning can be used to address fairness problems, outline the (dis-)advantages of the representation learning approach, discuss existing algorithms and open problems.

A Framework for Fair Representation Learning

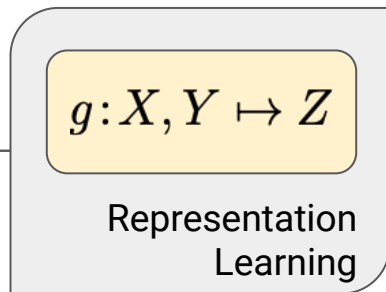




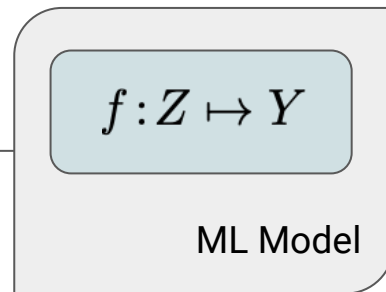
Data Regulator



Data Producer



Data User



Objectives

The data regulator determines which fairness criteria to use, and (optionally) audits the results.

When training:

- Input: interaction with users/experts/judges/policy to determine fairness criteria
- Output: fairness criteria

When auditing:

- Input (for auditing the **data producer**):
 - Learned representation
- Input (for auditing the **data user**):
 - Data and model predictions
- Output:
 - Are fairness criteria satisfied?

Data Regulator

Determines fairness criteria, determines data source(s), audits results

- INPUT: Data
- OUTPUT: Fairness criteria

AUDITING

- INPUT: Models
- OUTPUT: Satisfactory?



One of the key tasks of the **data regulator** is determining the **fairness criteria**

The most common algorithmic fairness criteria are **individual fairness** and **group fairness**...

Individual Fairness

Data Regulator

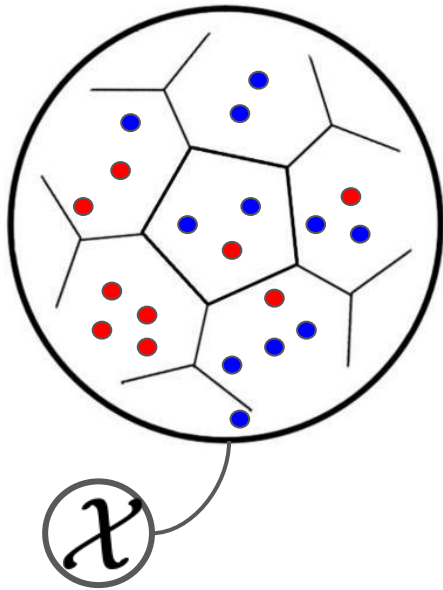


Individual Fairness: Similar individuals treated similarly



Pairs of **similar individuals playing the same sport** classified differently.

Individual Fairness: Similar individuals treated similarly

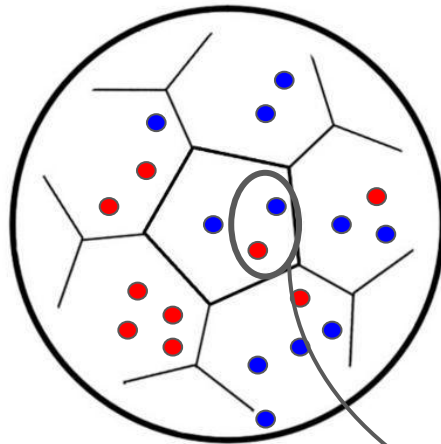


Data Regulator: Which individuals are similar? equiv., which individuals should be treated similarly?

One approach:

- Define a **partition** of the space into disjoint cells such that similar individuals are in the same cell.
- Individuals in the **same cell** should be **treated similarly** even if they are apparently different (e.g. dots with different colored attributes).

Individual Fairness: Similar individuals treated similarly



Data Regulator: Which individuals are similar?
quiv., which individuals should be treated similarly?

An algorithm $\mathcal{A}_{\mathcal{D}}$ is $(B, \epsilon(\mathcal{D}))$ -individually fair if \mathcal{X} can be partitioned into B disjoint subsets denoted $\{C_i\}_{i=1}^B$ such that $\forall x_1 \in \mathcal{X}$:

$$x_1, x_2 \in C_i \Rightarrow |l(\mathcal{A}_{\mathcal{D}}, x_1) - l(\mathcal{A}_{\mathcal{D}}, x_2)| \leq \epsilon(\mathcal{D})$$

Remark: Individual fairness implies algorithmic robustness (c.f. Xu & Mannor 2011).

Algorithmic Robustness Implies **Generalization**

⇒ Individual Fairness Implies **Generalization**

If a dataset \mathcal{D} consists of n i.i.d. samples and the algorithm $\mathcal{A}_{\mathcal{D}}$ is $(B, \epsilon(\mathcal{D}))$ -Individually Fair, then for any $\delta > 0$, with probability at least $1 - \delta$:

$$|l(\mathcal{A}_{\mathcal{D}}) - \hat{l}_{\mathcal{D}}(\mathcal{A}_{\mathcal{D}})| \leq \epsilon(\mathcal{D}) + M \cdot \sqrt{\frac{2B \cdot \ln 2 + 2 \ln(1/\delta)}{n}}$$

where $l(\mathcal{A}_{\mathcal{D}})$ (resp. $\hat{l}_{\mathcal{D}}(\mathcal{A}_{\mathcal{D}})$) is the risk (resp. empirical risk) of $\mathcal{A}_{\mathcal{D}}$.

Challenge: Individually fair models with **low training error + generalization**

Lipschitz Continuity implies Individual Fairness

If \mathcal{X} is compact w.r.t metric ρ and $l(\mathcal{A}_{\mathcal{D}})$ is $L(\mathcal{D})$ -Lipschitz continuous i.e.:

$$|l(\mathcal{A}_{\mathcal{D}}, x_1) - l(\mathcal{A}_{\mathcal{D}}, x_2)| \leq L(\mathcal{D}) \cdot \rho(x_1, x_2) \quad \forall x_1, x_2 \in \mathcal{X}$$

Then algorithm \mathcal{A} is $(\mathcal{N}(\gamma/2, \mathcal{X}, \rho), L(\mathcal{D})\gamma)$ -individually fair for all $\gamma > 0$ where $\mathcal{N}(\gamma/2, \mathcal{X}, \rho)$ is the covering number of \mathcal{X} .

Good news: One can achieve **fairness through Lipschitz regularization**.

Bad news: Data is non-Euclidean (e.g. images, graphs): $\rho \neq \|\cdot\|_2$.

Challenge: Can we learn a representation of the data such that $\rho = \|\cdot\|_2$ is a good metric to compare instances ?

Individual Fairness: **Advantages** and **Challenges**

Advantages:

- Intuitive and **easy to explain** to the data producer (and to non-experts)
- Individual fairness **implies generalization** (c.f. Xu & Mannor, 2012)
- Individual fairness **implies statistical parity** given regularity conditions (Dwork et al., 2012)

Challenges:

- Regulator **must provide a metric** or a set of examples to be treated similarly. Constructing a metric requires **significant domain expertise** and human insight.
- Fairness of the representation **heavily depends on the quality of the metric** chosen by the regulator.
- Optimizing and measuring individual fairness is generally **more computationally expensive** than other measures

Group Fairness

Data Regulator



Group Fairness: Similar Classifier Statistics Across Groups

Regulator: Which statistic $\nu(f, Y|S)$ should be equalized across the groups?

Commonly used measures are straightforward functions of classifier performance statistics, e.g.,

- Eq. of Opportunity (Hardt et. al. 2016)

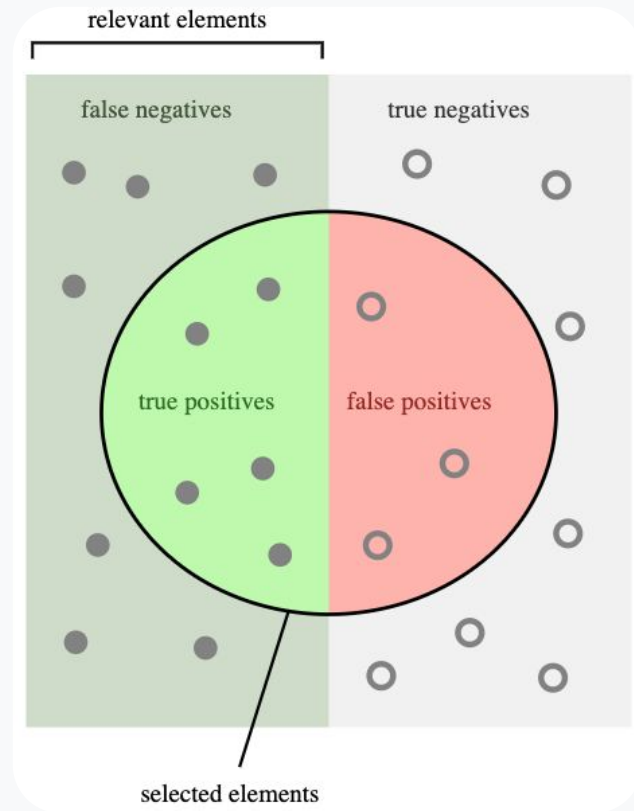
$$\mathbf{TP}_S = P(Y = 1, f = 1|S)$$

- Equalized Odds (Hardt et. al. 2016)

$$\{\mathbf{TP}_S; \mathbf{FP}_S\}$$

- Statistical parity (Dwork et. al. 2012)

$$\mathbf{TP}_S + \mathbf{FP}_S = P(f(Z) = 1|S)$$



(Im-)possibility Results for Group-Fair Classification

Classifier statistics are not arbitrarily flexible!

E.g. binary classification statistics have two **degrees of freedom**, thus can match at most two independent statistics across groups (c.f. Kleinberg et. al., 2017; Chouldechova, 2017)

Beyond binary classification, the **degrees of freedom** grow quadratically with number of classes

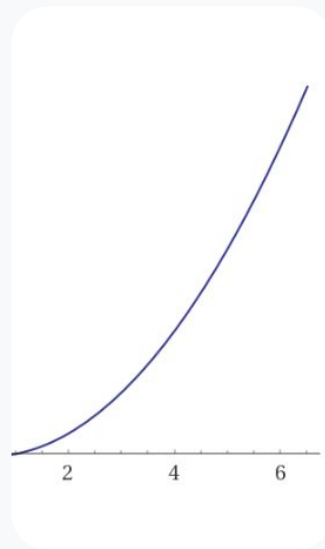


Figure showing number of classes vs. degrees of freedom
More independent constraints can be enforced when there are more classes.

Group Fairness: **Advantages** and **Challenges**

Advantages:

1. **Efficient to compute, measure and enforce** for the data producer and regulator.
2. Often **easier to explain** to policy-makers (as in terms of population behavior)
3. Much **more existing work**, strategies for representation learning

Challenges:

1. Data regulator **must determine which classifier statistic(s)** to equalize.
2. Fairness of the representation **depends on the quality of the fairness metric** chosen by the regulator.
3. Group fairness **can lead to (more) violated individual fairness**, e.g., intersectionality can lead to fairness gerrymandering (Kearns et. al., 2018), and other issues (McNamara et. al., 2019)

The Data Regulator: Measuring (Un-)fairness

- Regulator must choose how to measure (un-)fairness
 - For individual fairness: must choose the distance metric
 - For group fairness: must choose the classifier statistics to equalize
- However, remember that there are **no magic metrics** or measures;
Measurement 101: all measures have **blind spots**

“When a measure becomes a target, it ceases to be a good measure.”
- For ML, we generally specify all measures apriori and optimize them
- However, **all** metrics will have **failure cases**, i.e., unusual situations with non-ideal behavior
- One productive approach is to select measures that best capture tradeoffs relevant to the context

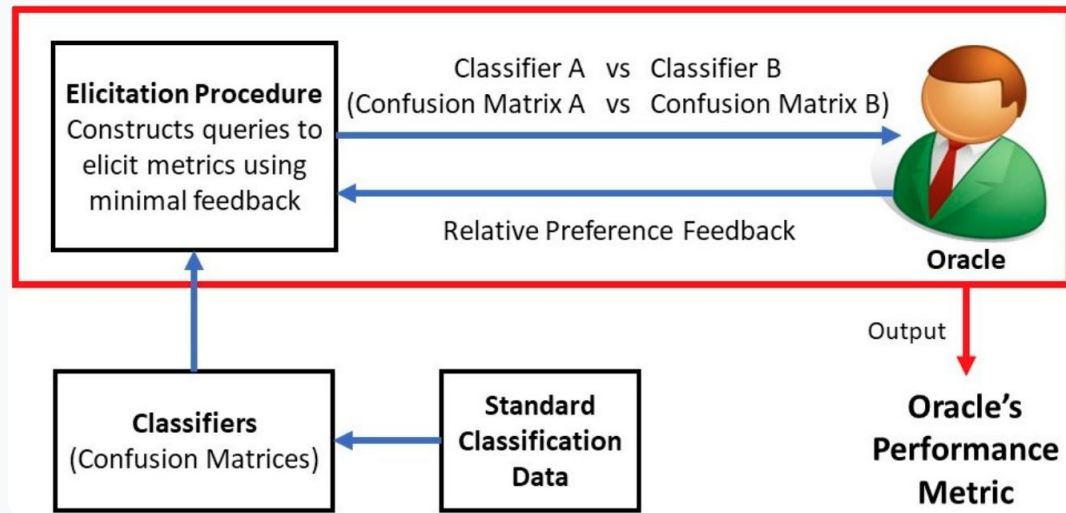
Metric Elicitation

Determine the ideal evaluation metric by interacting with users, experts (Hiranandani et. al., 2019).

Ongoing extension to eliciting group fairness metrics.

Complementary work on eliciting distance metrics for individual fairness (Ilvento, 2019; Jung et. al., 2019)

Figure from Hiranandani et. al (NeurIPS 2019)



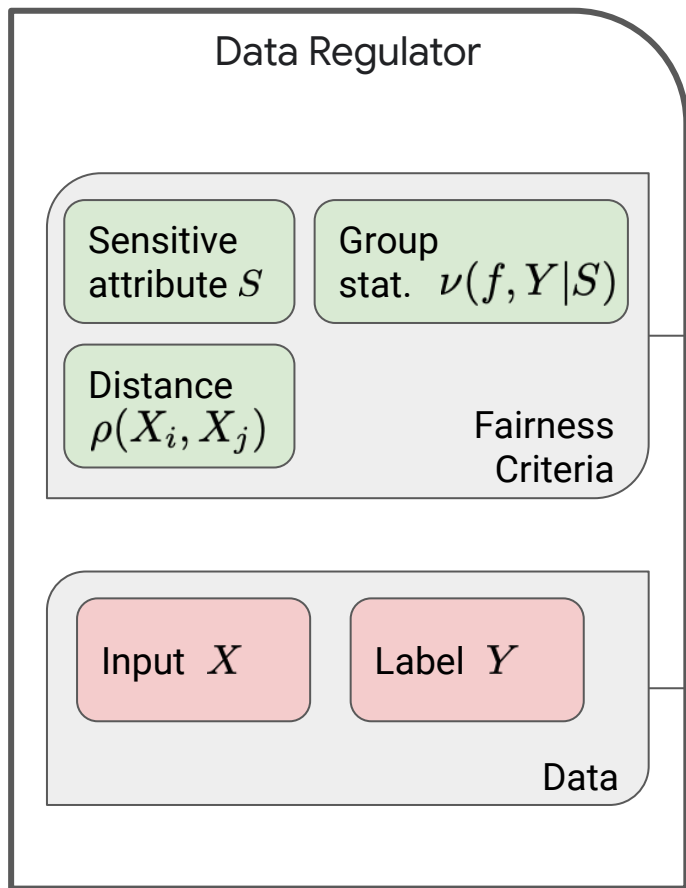
Poster # 226; Wednesday



Another key task of the **data regulator** is to **audit** the learning system (e.g., Madras et al., 2018)

The most efficient approach is to audit the learned representation, i.e., the **data producer**

For complex label-dependent settings, or for an adversarial **data user**, the **data regulator** must audit the final model, i.e., the **data user**



Objectives

The data producer computes the representation given the fairness criteria and input data.

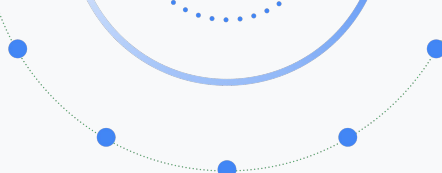
There are a variety of methods for representation learning with individual fairness or group fairness constraints, which, in turn, can be label (in-)dependent.

- Inputs: X, Y
 - Data
 - Fairness criteria $\nu(f, Y|S)$
Alternatively $\rho(X_i, X_j)$
- Output:
 - Learned representation
 $g: X, Y \mapsto Z$

Data Producer

Computes the fair representation given data regulator criteria

- INPUT: Fairness criteria
- OUTPUT: Representation



Representation learning is the task of estimating a concise and informative data summary, usually implemented as a low-dimensional data transformation.

$$g: X, Y \mapsto Z$$

Approaches in common use include PCA and non-linear autoencoders.

Individual Fairness

Data Producer

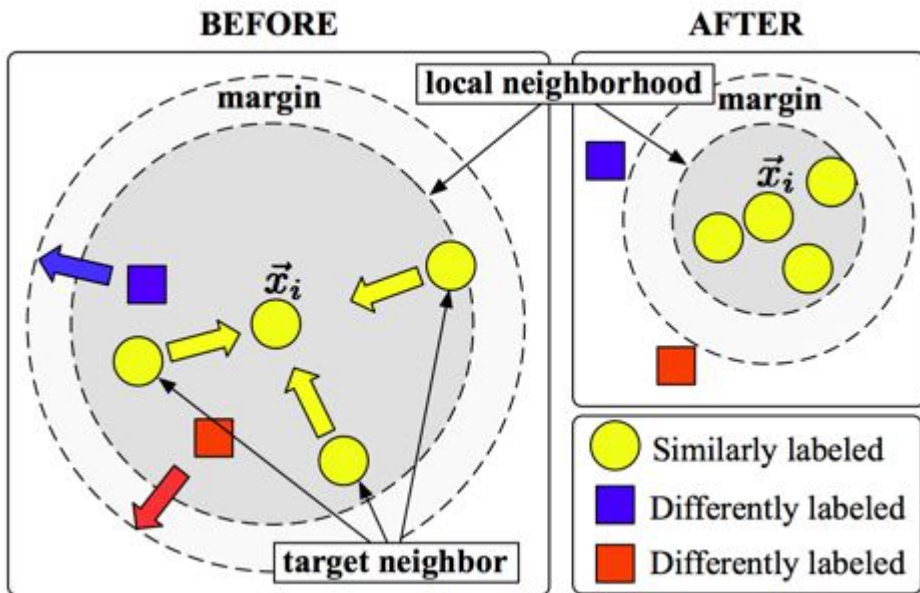


Individual Fairness: Metric Learning Approach

Regulator (to the data producer):
Provides sets of examples which should be treated similarly (e.g., similarly labeled points).

Producer: Learns a distance metric such that individuals which should be treated similarly are closer to each other.

Find a metric ρ such that $\forall(x_1, x_2, x_3)$:
 $x_1, x_2 \in C_i$ and $x_3 \in C_j (j \neq i)$
 $\Rightarrow \rho(x_1, x_2) \leq \rho(x_1, x_3)$



Individual Fairness: Metric Learning Approach

Regulator (to the data producer):
Provides sets of examples which should be treated similarly (e.g. similarly labeled points).

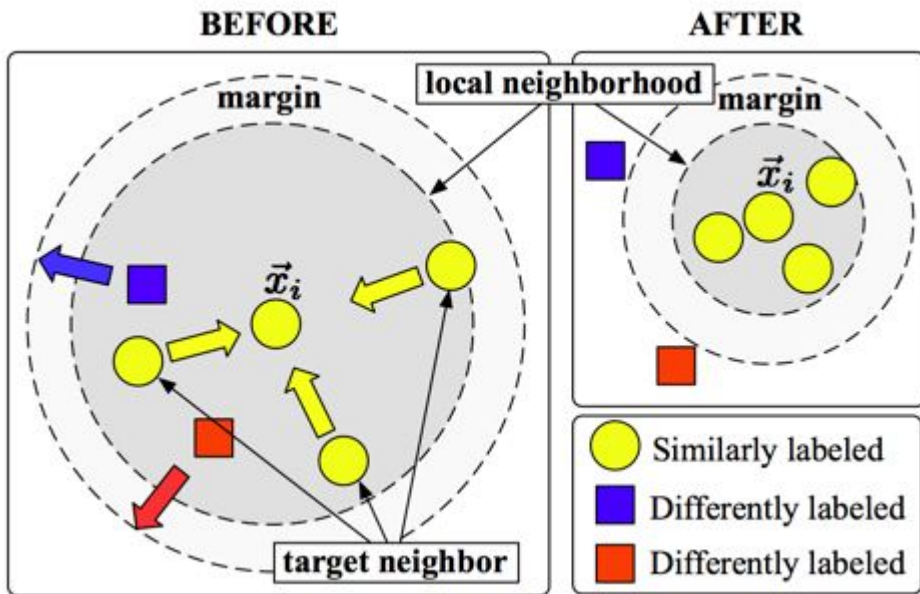
Producer: Equivalently, learn a representation such that individuals which should be treated similarly are closer to each other in the induced euclidean metric:

Find a metric ρ such that $\forall(x_1, x_2, x_3)$:

$x_1, x_2 \in C_i$ and $x_3 \in C_j (j \neq i)$

$$\Rightarrow \|z_1 - z_2\|_2 \leq \|z_1 - z_3\|_2$$

where $z_i = Lx_i$ and $\rho(x_i, x_j) = x_i^T L^T Lx_j$.



Group Fairness

Data Producer



Group fairness with representative prototypes

- **Representation:** $g: X, Y \mapsto Z$
Via prototypes, defined by parameterized mixture model that stochastically maps data to prototypes
- **Prediction:** $f: Z \mapsto Y$
Parameterized mixture model that stochastically maps prototypes to labels
- **Fairness Measure:** $\nu(f, Y|S)$
Statistical parity $\text{TP}_S + \text{FP}_S = P(f(Z) = 1|S)$
i.e., group averaged label probability across groups
Trained to minimize the weighted average of data approximation, prediction quality, and statistical parity

Group fairness, and individual fairness with ambient metric

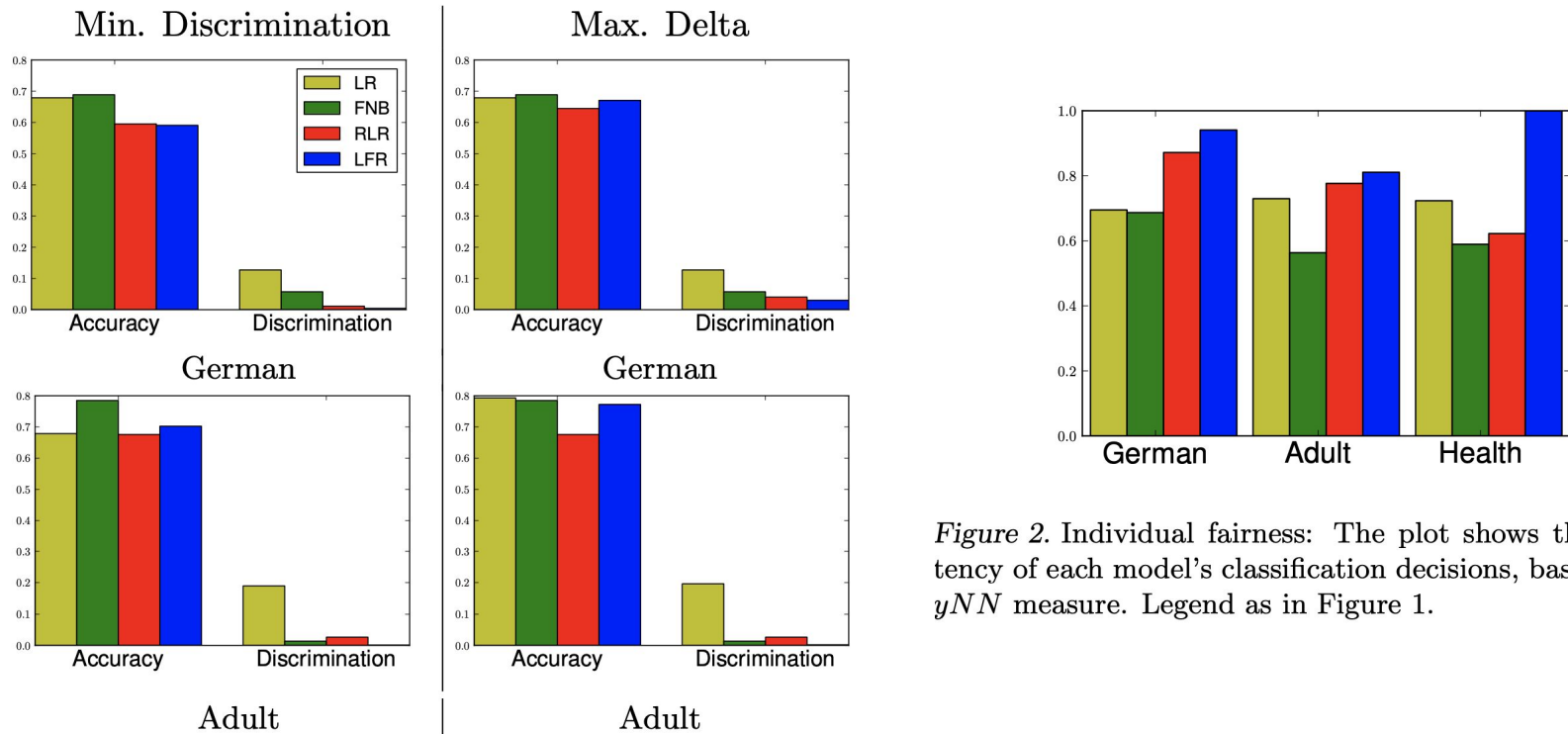
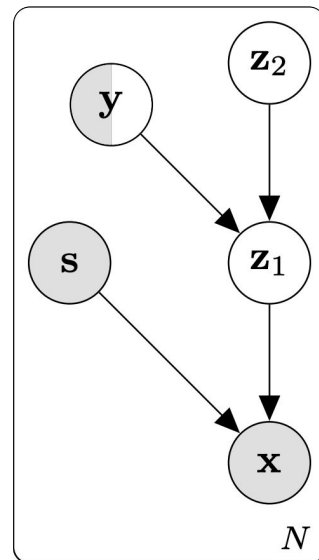


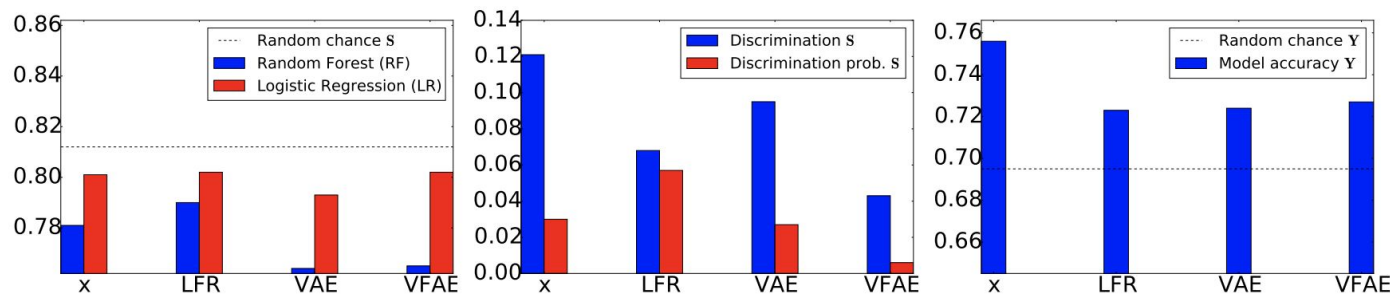
Figure 2. Individual fairness: The plot shows the consistency of each model's classification decisions, based on the yNN measure. Legend as in Figure 1.

Semi-supervised variational autoencoder + MMD fairness

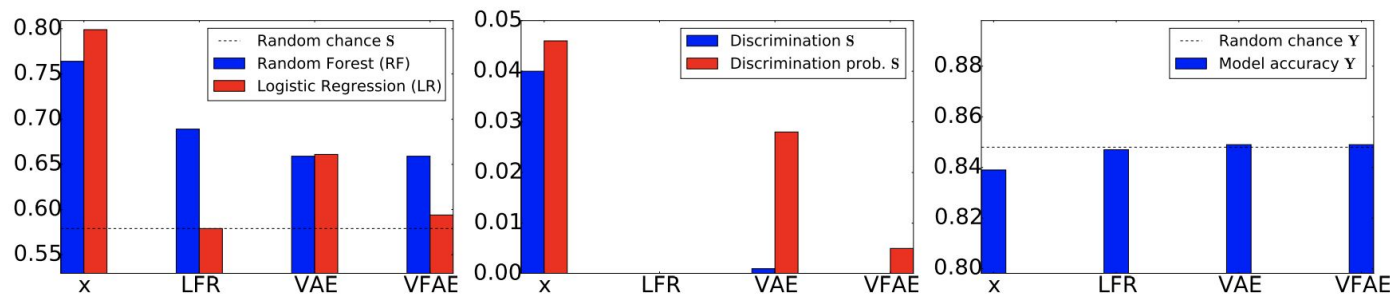
- Representation:** $g: X, Y \mapsto Z$
 Variational autoencoder
- Prediction:** $f: Z \mapsto Y$
 Logistic regression, Random forests
- Fairness Measure:** $\nu(f, Y|S)$
 Statistical parity $TP_S + FP_S = P(f(Z) = 1|S)$
 Implemented by penalizing MMD of stochastic embeddings across groups
- Trained using variational inference + MMD regularization.



Performance vs. group fairness



(b) German dataset



(c) Health dataset

Learning Controllable Fair Representations

- **Representation:** $g: X, Y \mapsto Z$
Variational autoencoder
- **Prediction:** $f: Z \mapsto Y$
Feedforward neural networks
- **Fairness Measure:** $\nu(f, Y|S)$
Statistical parity, equal opportunity, equalized odds
Information-theoretic approximations to the fairness metrics,
combined with variational approximations for efficient estimation
Recovers approximations to existing methods as special cases

Controllable Fair Representations

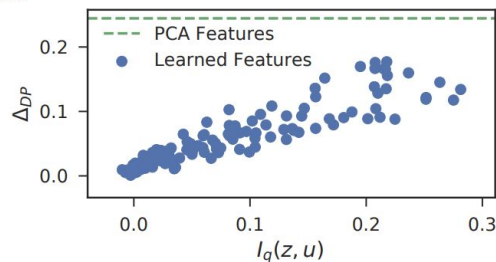
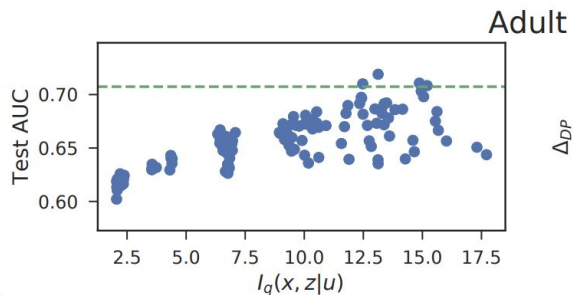
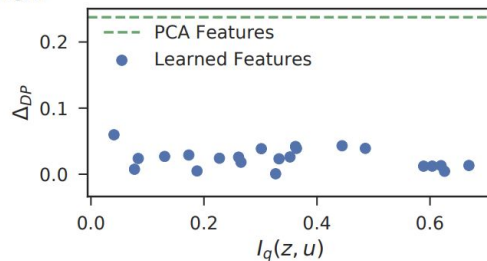
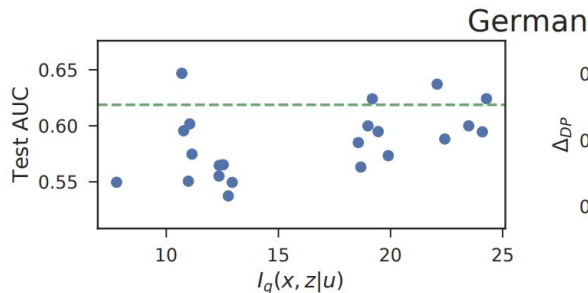
“Our method encourages representations that satisfy the fairness constraints while being more expressive, and that our method is able to balance the trade-off between multiple notions of fairness with a single representation and a significantly lower computational cost.”

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

	λ_1	λ_2
Zemel et al. (2013)	0	A_z/A_x
Edwards and Storkey (2015)	0	α/β
Madras et al. (2018)	0	γ/β
Louizos et al. (2015)	1	β

Mutual Information measures can approximate fairness quantities

The relationship between mutual information and fairness related quantities



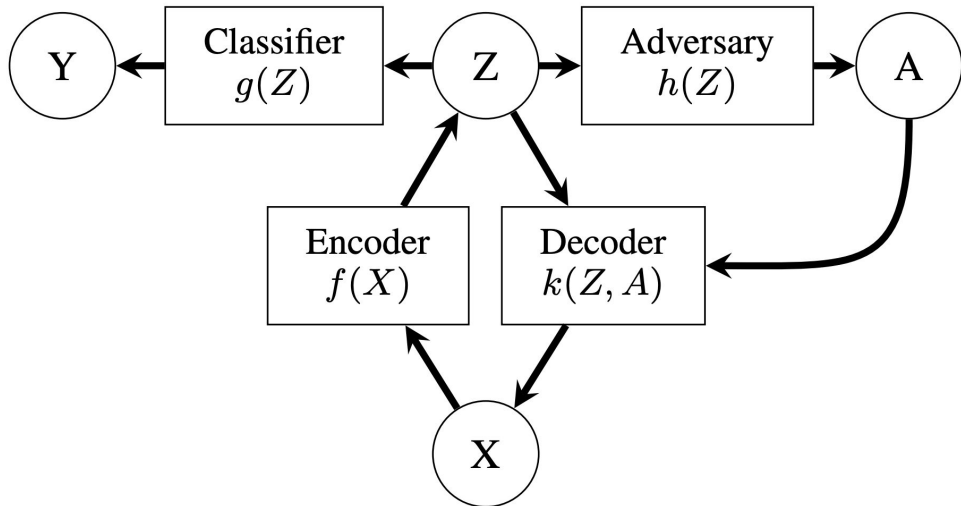
An Adversarial Approach for Learning Fair Representations

- **Representation:** $g: X, Y \mapsto Z$
Adversarially trained neural network autoencoder
- **Prediction:** $f: Z \mapsto Y$
Feedforward neural networks
- **Fairness Measure:** $\nu(f, Y|S)$
Statistical parity, equal opportunity, equalized odds
Specialized adversary loss functions for each fairness measure

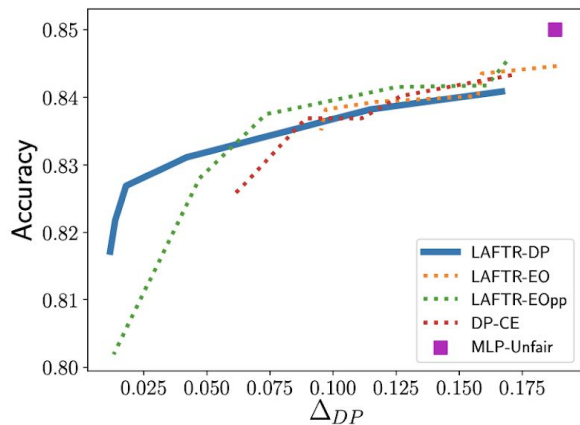
Provide bounds on the fairness violation of any subsequent classifier using the learned representation

Adversarially Learning Fair Representations

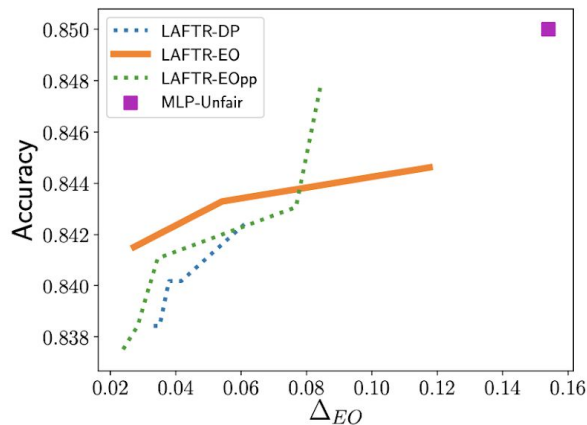
“We frame the data owner’s choice as a representation learning problem with an adversary criticizing potentially unfair solutions”



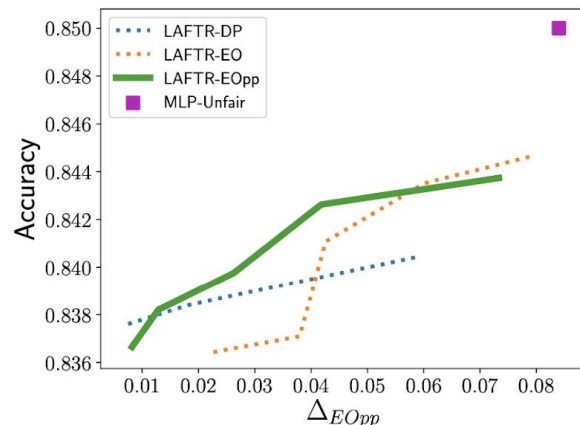
Group fairness/performance tradeoff on Adult dataset



(a) Tradeoff between accuracy and Δ_{DP}



(b) Tradeoff between accuracy and Δ_{EO}



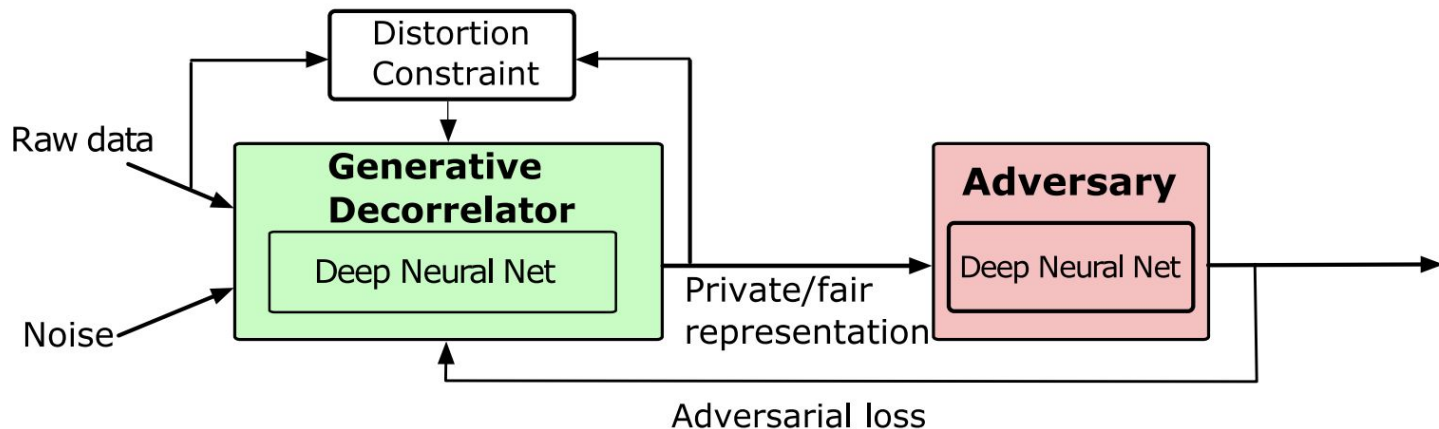
(c) Tradeoff between accuracy and Δ_{EOpp}

Generative Adversarial Representations

- **Representation:** $g: X, Y \mapsto Z$
Adversarially trained neural network autoencoder
- **Prediction:** $f: Z \mapsto Y$
Feedforward neural networks
- **Fairness Measure:** $\nu(f, Y|S)$
Statistical parity
Implemented by constructing representations that are robust against the optimal adversary

Generative Adversarial Representations

“GAP leverages recent advancements in adversarial learning to allow a data holder to learn universal representations that decouple a set of sensitive attributes from the rest of the dataset”



Distortion	0	0.003	0.0045	0.005	0.006	0.007	0.008	0.01
$\Delta_{\text{DemP}}(\text{white})$	0.061	0.055	0.04	0.03	0.03	0.02	0.02	0.01
$\Delta_{\text{DemP}}(\text{black})$	0.109	0.021	0.02	0.05	0.03	0.05	0.03	0.03
$\Delta_{\text{DemP}}(\text{Asian})$	0.14	0.082	0.07	0.07	0.06	0.07	0.06	0.03
$\Delta_{\text{DemP}}(\text{Indian})$	0.031	0.006	0.01	0	0.01	0	0.01	0.01

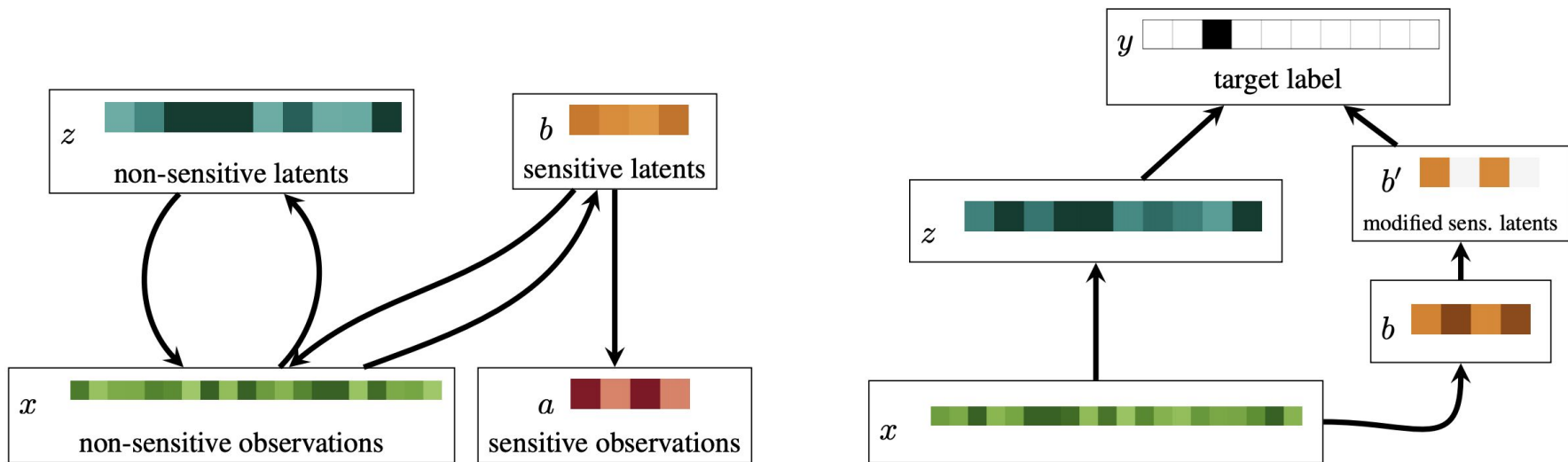
Table 6: The demographic parity fairness (indicated by $\Delta_{\text{DemP}}(\cdot)$) of ethnicity classification on the UTKFace dataset.

Fair Representations using Disentanglement

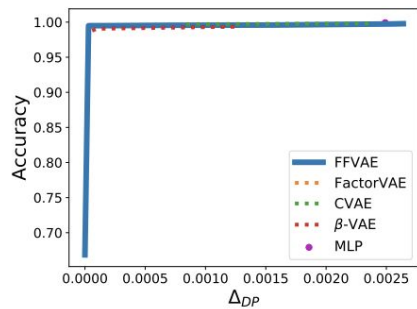
- **Representation:** $g: X, Y \mapsto Z$
Variational autoencoder
- **Prediction:** $f: Z \mapsto Y$
Feedforward neural networks
- **Fairness Measure:** $\nu(f, Y|S)$
Statistical parity $\text{TP}_S + \text{FP}_S = P(f(Z) = 1|S)$
Implemented by penalizing mutual information of sensitive and non-sensitive feature representations, which encourages disentanglement

Fair Representations using Disentanglement

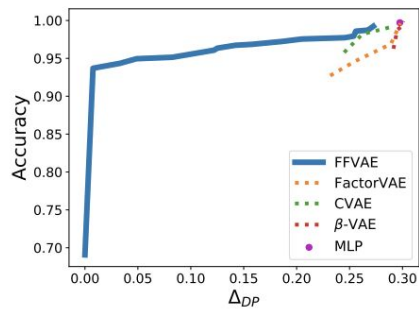
“Can we ... learn a flexibly fair representation that can be adapted, at test time, to be fair to a variety of protected groups and their intersections?”



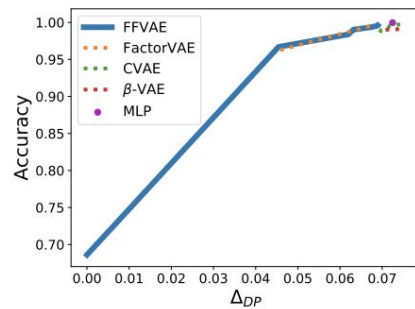
Group Fairness, Performance Tradeoffs



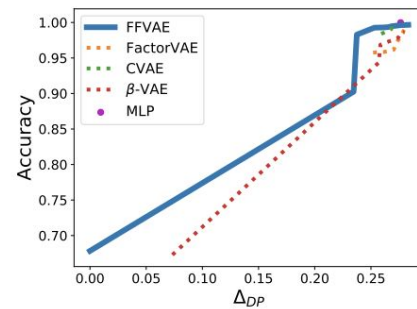
(a) $a = \text{Scale}$



(b) $a = \text{Shape}$



(c) $a = \text{Shape} \wedge \text{Scale}$



(d) $a = \text{Shape} \vee \text{Scale}$

Figure 2. Fairness-accuracy tradeoff curves, DSpritesUnfair dataset. We sweep a range of hyperparameters for each model and report Pareto fronts. Optimal point is the top left hand corner — this represents perfect accuracy and fairness. MLP is a baseline classifier trained directly on the input data. For each model, encoder outputs are modified to remove information about a . $y = \text{XPosition}$ for each plot.

Fairness of Disentangled Representations

- **Representation:** $g: X, Y \mapsto Z$
Disentangled representations (**independent of the fairness metric**)
- **Prediction:** $f: Z \mapsto Y$
Feedforward neural networks
- **Fairness Measure:** $\nu(f, Y|S)$
Statistical parity $\text{TP}_S + \text{FP}_S = P(f(Z) = 1|S)$
No explicit regularization for fairness measure.
Sensitive attributes are unknown during representation learning

Fairness of Disentangled Representations

“Analyzing the representations of more than 12 600 trained state-of-the-art disentangled models, we observe that several disentanglement scores are consistently correlated with increased fairness, suggesting that disentanglement may be a useful property to encourage fairness when sensitive variables are not observed.”

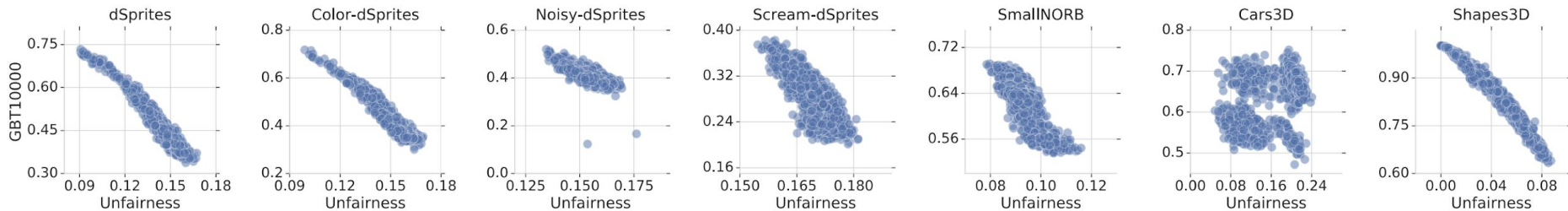


Figure 4: Unfairness of representations versus downstream accuracy on the different data sets.

When do disentangled representations imply fairness?

Thursday, Poster #34

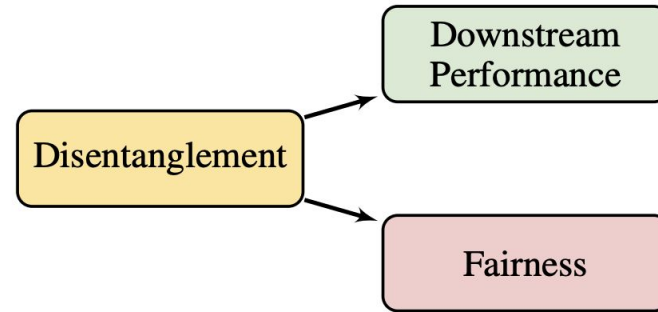
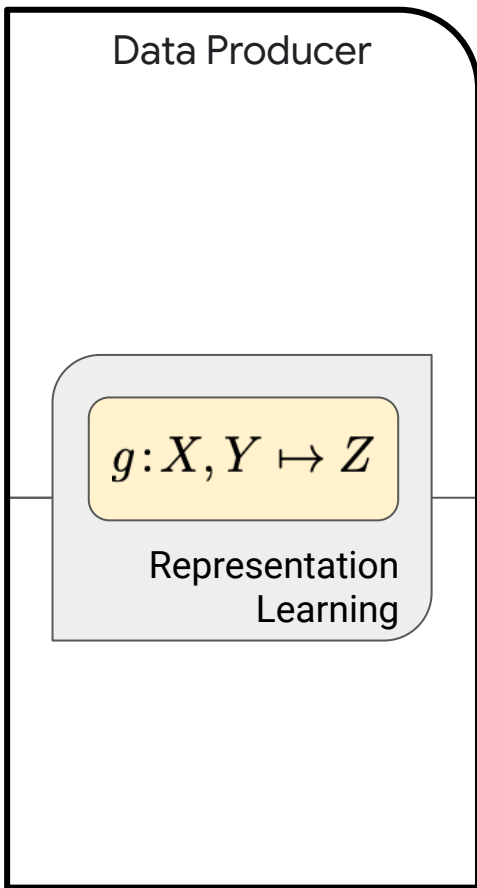


Figure 8: If disentanglement is a causal parent of downstream performance and fairness and there are no hidden confounders, then the former can be used as a proxy for the latter.

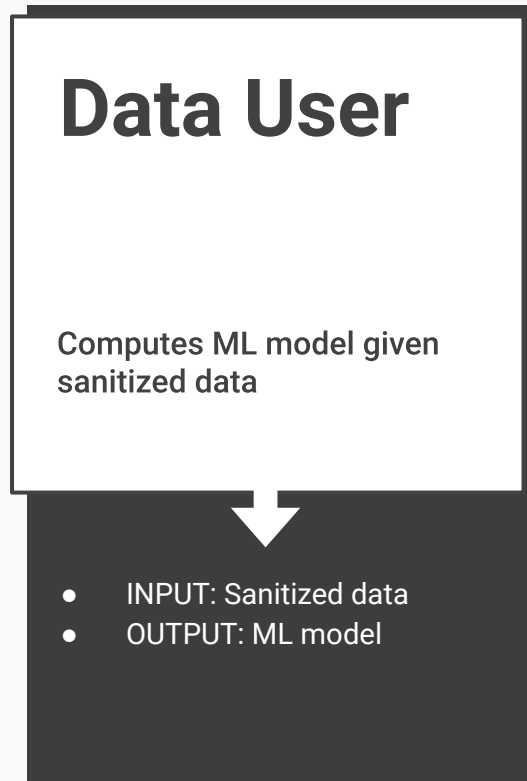


Objectives

Data producer computes the machine learning model given the sanitized representation.

Most of the fairness responsibility is with the data producer and regulator. The data user need only remain compliant with the pre-specified expectations, e.g., avoid adding new features that can result in fairness violations.

- Inputs:
 - Sanitized data Z, Y
- Output:
 - ML model $f : Z \mapsto Y$



Keep calm and carry ML on

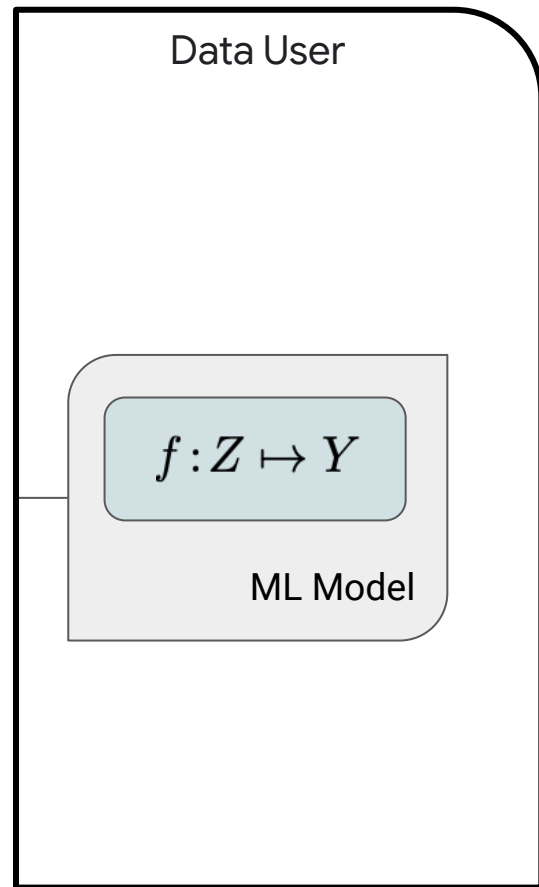
Data user may be ignorant of the fairness concerns in the system.

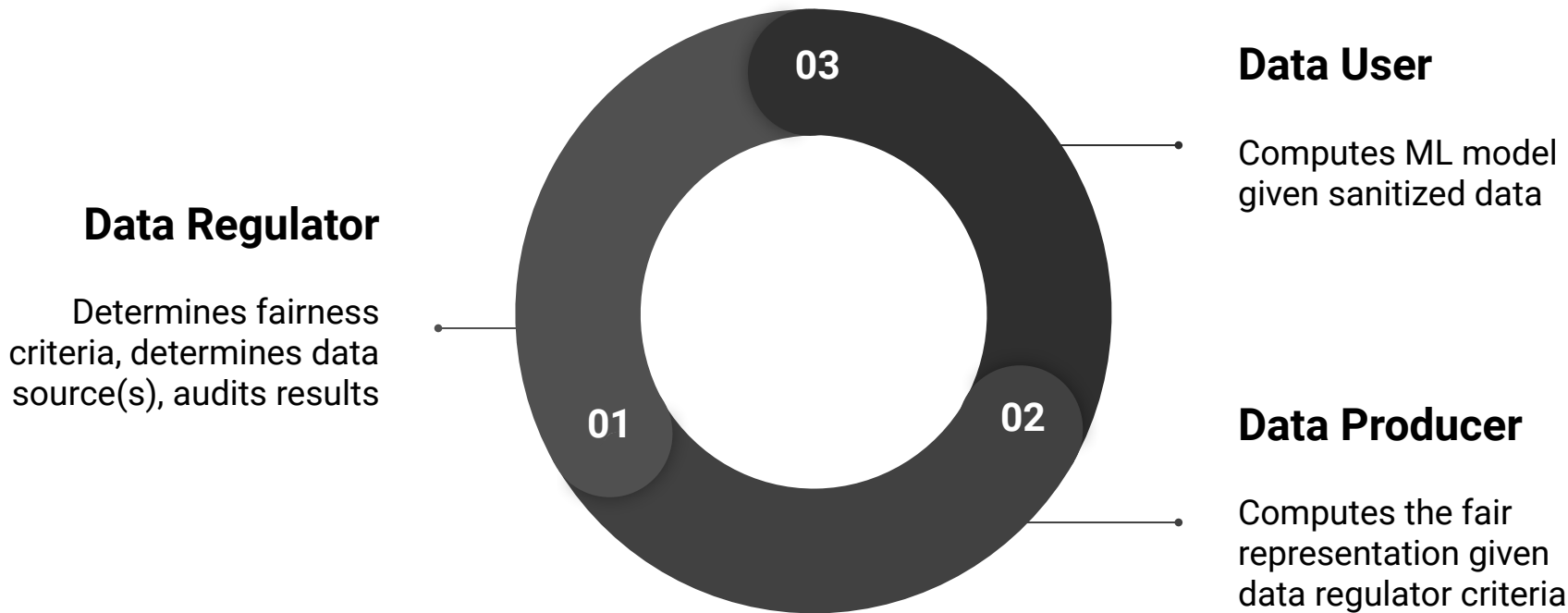
Data user trains ML models as usual

The PyTorch logo, a red flame-like symbol.

PyTorch







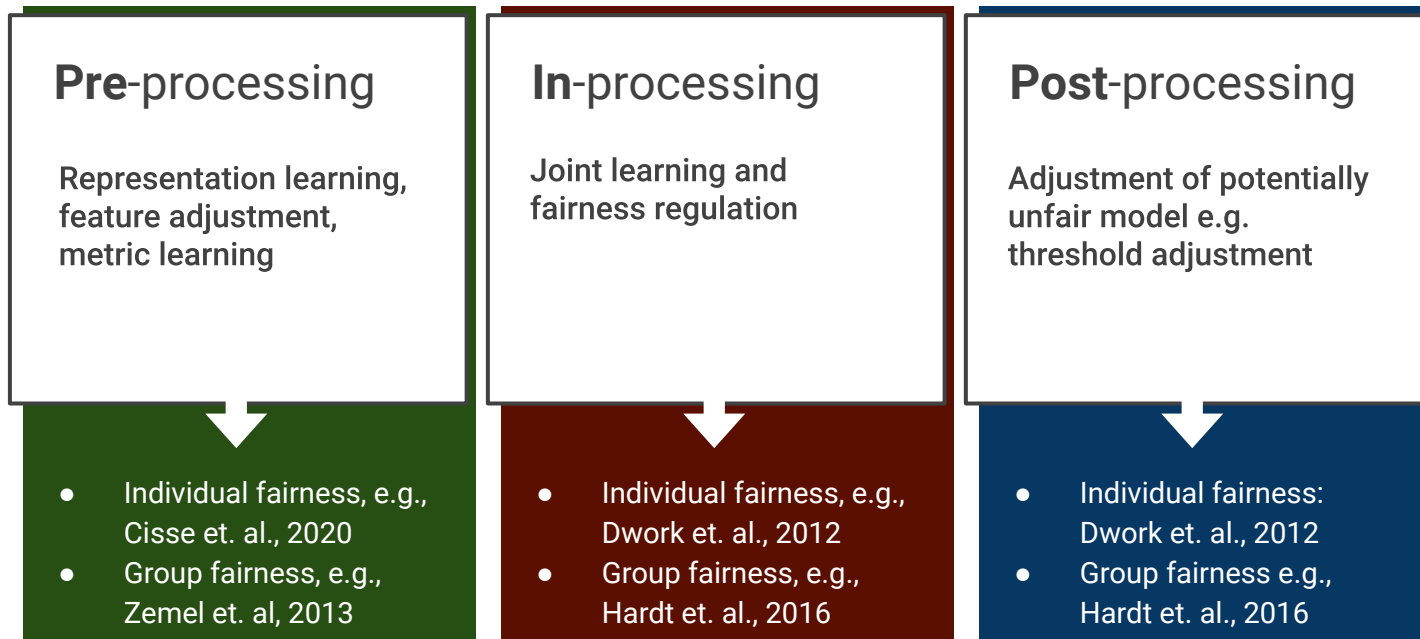
Pros of incorporating fairness using representation learning

- Often much more efficient than alternatives, especially with re-use
- Can be employed when the data user is untrusted, or apathetic about fairness
 - Data user is (mostly relieved) of the burden of directly reasoning about fairness
- **Inherits other good properties from representation learning**
 - Interpretability (in some cases)
 - Transportability (across datasets, institutions, ...)
 - Some robustness properties, some (weak) privacy properties
- **Audits can be much more efficient (especially when only auditing the representation)**

Cons of incorporating fairness using representation learning

- **Less precise control of fairness/performance tradeoff**
 - Should expect worse fairness/performance tradeoff than joint training
 - See the “cost of fairness in representation learning” (McNamara et. al., 2019)
- **May lead to fairness overconfidence**
 - Data user may act adversarially when optimizing for the performance metric of interest
 - Data user can still violate fairness, e.g., by violating data agreement
- **Startup costs can be high:**
 - Representation learning can be expensive, especially with multiple fairness constraints

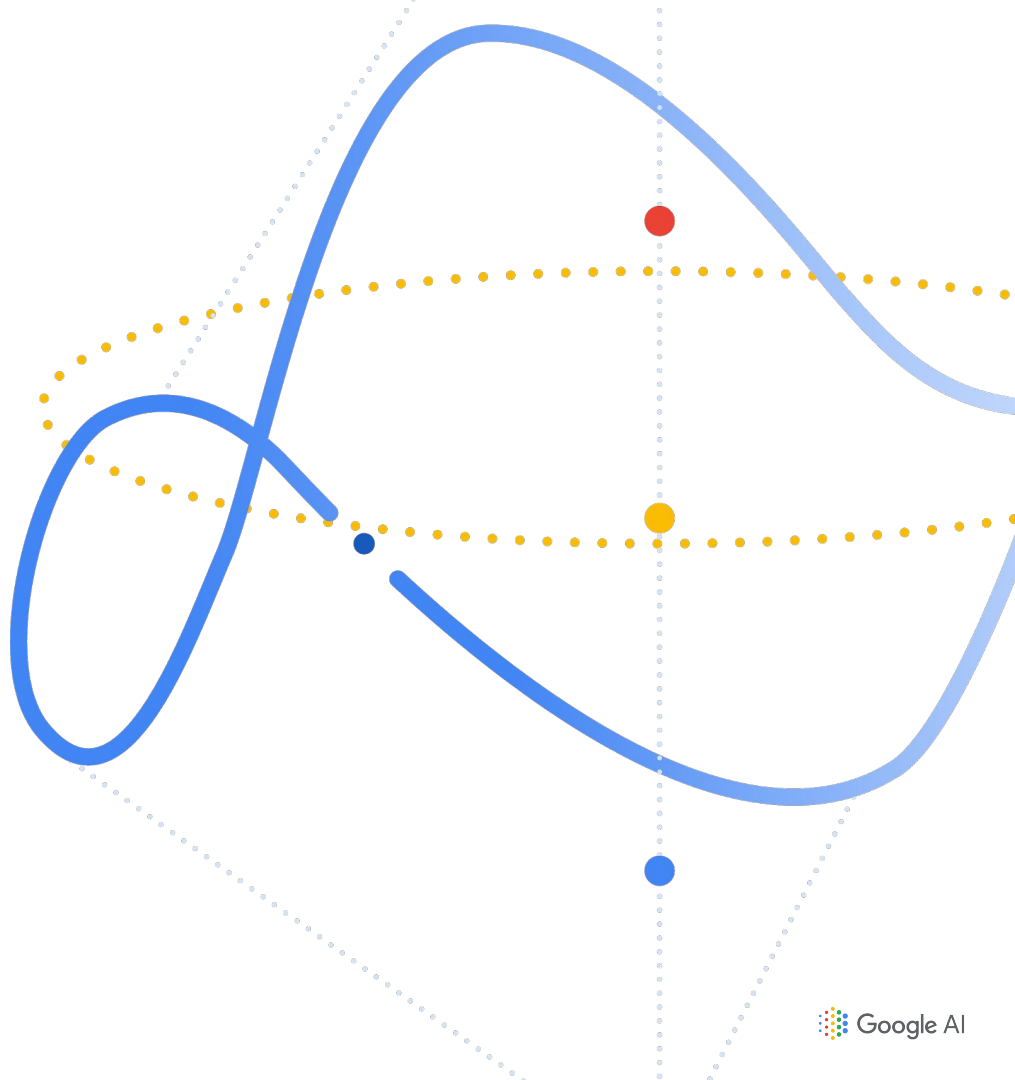
Alternative approaches for implementing algorithmic fairness



Some tradeoffs when comparing algorithmic fairness approaches

	Ease of implementation and (re-)use	Scalability	Ease of auditing	Fairness / Performance tradeoff	Generalization
Pre-processing , e.g., representation learning	✗	✗	✗		✗
In-processing , i.e., joint learning and fairness regulation			✗	✗	✗
Post-processing , e.g., threshold adjustment		✗	✗		

Back to the story ...



A manager oversees several teams, all are using the same data to build predictive models for different products. The manager seeks to ensure both **fairness and accuracy** across the products.

Each team is solving a different prediction task.

There is no company policy on fairness, thus no shared guidelines.

- Team **alpha** is fully focused on accuracy, but is oblivious (neighbors say they are apathetic) about fairness issues.
- Team **beta**, team **nu** and team **gamma** are all interested in fairness. Each team is really excited to implement this and has read the literature, but each team has selected different fairness definitions.
- Team **zeta** would like to improve the fairness of their predictions, but has no idea how to incorporate or measure fairness.
- The **manager** has decided to independently verify that all released products are fair

A manager oversees several teams, all are using the same data to build predictive models for different products. The manager seeks to ensure both **fairness and accuracy** across the products.

Challenges:

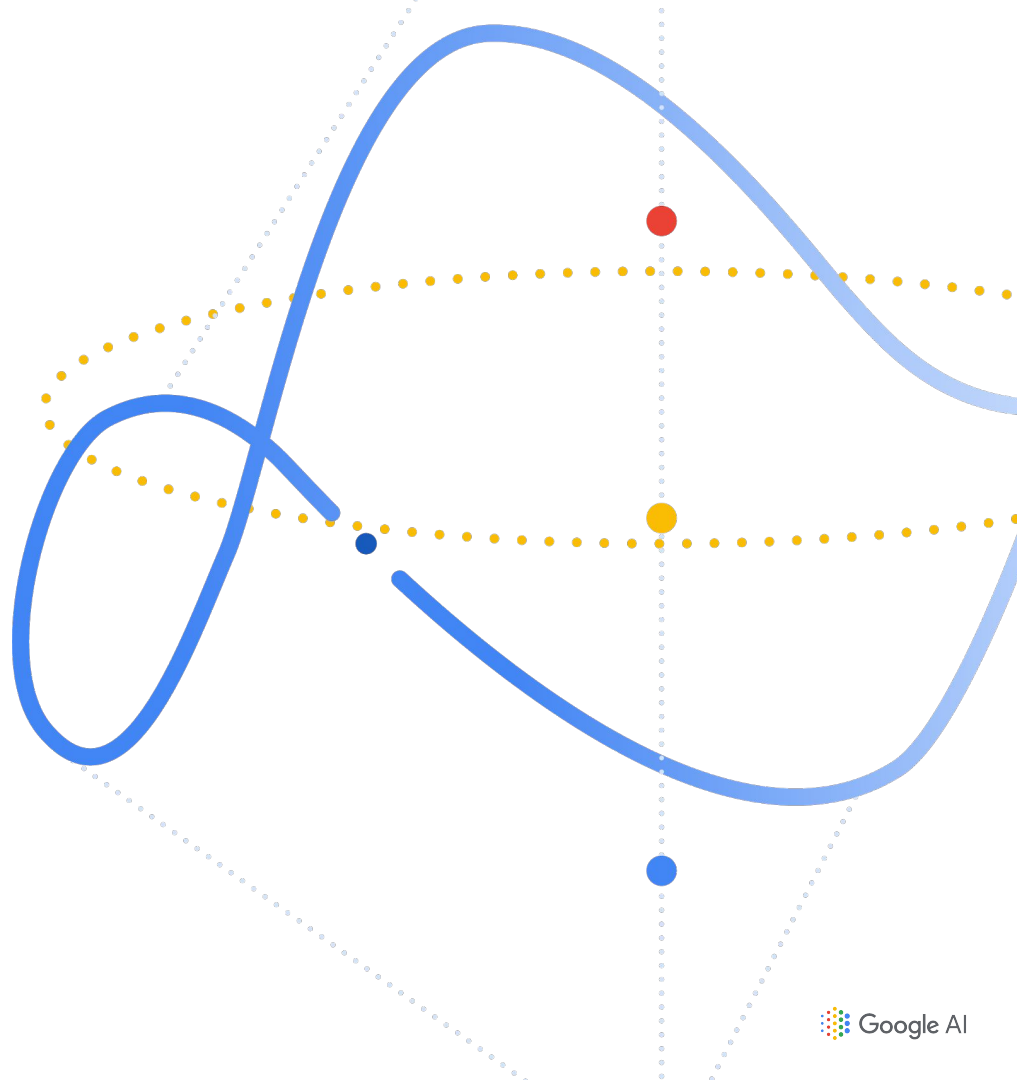
- Some teams do not have the expertise (or interest) to design fairer models.
- Different teams use different definitions of fairness.
- Incorporating fairness can have different impacts on the performance of the models across products.
- Auditing all the predictive models for fairness can be challenging when each team has its own recipe.

A manager oversees several teams, all are using the same data to build predictive models for different products. The manager seeks to ensure both **fairness and accuracy** across the products.

Representation learning to the rescue!

- Representation learning can be used to centralize fairness constraints, by moving the fairness responsibility from the data user to the data regulator
- Learned representation can simplify and centralize the task of fairness auditing
- Learned representations can be constructed to satisfy multiple fairness measures simultaneously
- Learned representations can simplify the task of evaluating the fairness/performance tradeoff, e.g., using performance bounds

Conclusion



Beyond algorithmic fairness

- Fairness is a nuanced and challenging issue with many **open problems**, e.g., incorporating user agency, metric selection, ...
- **Feedback loops** are common in deployed systems, i.e., predictions leading to (user) actions, which are collected as new data.
- **Inappropriate data** is often the source of bias, e.g., labels which are correlated with sensitive attributes due to sampling effects, non-causal data collection, systematic undersampling of sub-populations ...
- **Collecting additional data** may be the best way to improve **both** performance and fairness

Conclusion

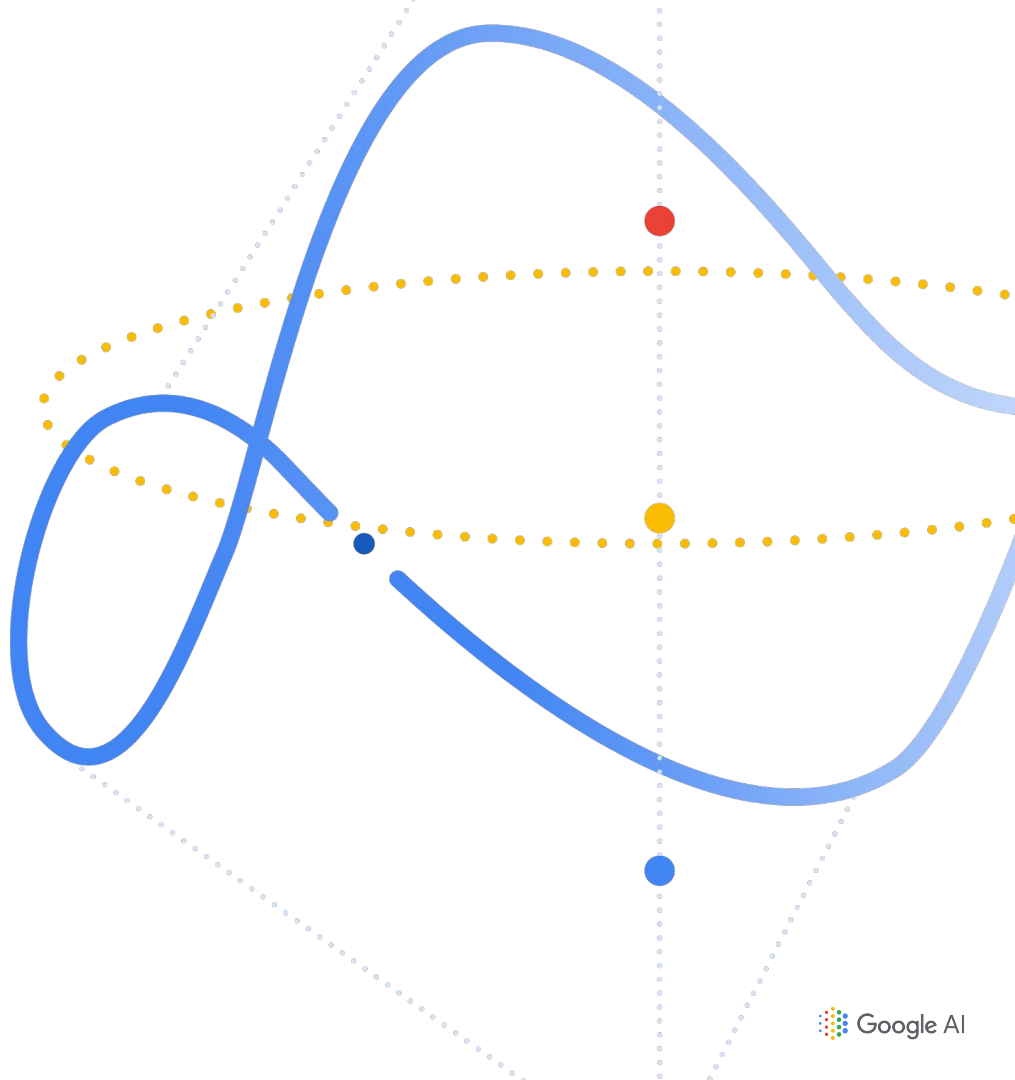
- Representation learning is a promising approach for implementing algorithmic fairness
- Fair representation learning can be implemented with modular separation between tasks/roles:
 - **Data regulator:** determines fairness measure(s), audits results
 - **Data producer:** learns the fair representation
 - **Data user:** agnostically learns the ML model
- **Some new-ish observations and results:**
 - Connections between individual fairness and robustness, generalization
 - Distance metric learning for individual fairness and representation learning
 - Elicitation for selecting fairness measures
 - Group fairness impossibility results depend on the number of classes

Lots of **open questions!**

- For the **data regulator**:
 - How does one pick appropriate fairness definitions, what is the role of metric elicitation?
 - What are some best practices for auditing the results?
- For the **data producer**:
 - Can we further improve algorithms for learning fair representations?
 - Can one construct algorithms for individually fair representation learning?
- For the **data user**:
 - What is the cost of fairness via representation learning?
 - What are some best practices for avoiding fairness leakage?
- And **many more algorithmic questions**:
 - How do these ideas apply beyond (binary) classification problems?
 - How do these ideas apply to continuous variables e.g. age?

Thank you for your attention!

Questions?



References

- Cisse M., Koyejo, S., 2020. Representation learning and fairness. *In Prep*.
- Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), pp.153-163.
- Creager, E., Madras, D., Jacobsen, J.H., Weis, M., Swersky, K., Pitassi, T. and Zemel, R., 2019, May. Flexibly Fair Representation Learning by Disentanglement. In *International Conference on Machine Learning* (pp. 1436-1445).
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., 2012, January. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226). ACM.
- Goodhart, C.A., 1984. Problems of monetary management: the UK experience. In *Monetary Theory and Practice* (pp. 91-121). Palgrave, London.
- Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).
- Hiranandani, G., Boodaghians, S., Mehta, R. and Koyejo, O.O., 2019. Multiclass Performance Metric Elicitation. In *Advances in Neural Information Processing Systems* (pp. 9351-9360).
- Ilvento, C., 2019. Metric Learning for Individual Fairness. *arXiv preprint arXiv:1906.00250*.
- Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L. and Wu, Z.S., 2019. Eliciting and Enforcing Subjective Individual Fairness. *arXiv preprint arXiv:1905.10660*.
- Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2018, July. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning* (pp. 2569-2577).
- Kleinberg, J., Mullainathan, S. and Raghavan, M., 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Vol. 67, p. 43). Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik.

References

- Liao, J., Huang, C., Kairouz, P. and Sankar, L., 2019. Learning Generative Adversarial RePresentations (GAP) under Fairness and Censoring Constraints. *arXiv preprint arXiv:1910.00411*.
- Louizos, C., Swersky, K., Li, Y., Welling, M. and Zemel, R., 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.
- Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B. and Bachem, O., 2019. On the Fairness of Disentangled Representations. In *Neural Information Processing Systems*
- Lum K, Johndrow J. A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077. 2016 Oct 25.
- Madras, D., Creager, E., Pitassi, T. and Zemel, R., 2018, July. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning* (pp. 3381-3390).
- McNamara, D., Ong, C.S. and Williamson, R.C., 2019, January. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 263-270). ACM.
- Quadrianto, N., Sharmanska, V. and Thomas, O., 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8227-8236).
- Song, J., Kalluri, P., Grover, A., Zhao, S. and Ermon, S., 2019, April. Learning Controllable Fair Representations. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 2164-2173).
- Stock, P. and Cisse, M., 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 498-512).
- Strathern, M., 1997. 'Improving ratings': audit in the British University system. *European review*, 5(3), pp.305-321.
- Xu, H. and Mannor, S., 2012. Robustness and generalization. *Machine learning*, 86(3), pp.391-423.
- Wang, X., Li, R., Yan, B. and Koyejo, O., 2019. Consistent Classification with Generalized Metrics. *arXiv preprint arXiv:1908.09057*.
- Weinberger, K.Q. and Saul, L.K., 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb), pp.207-244.
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C., 2013, February. Learning fair representations. In *International Conference on Machine Learning* (pp. 325-333).