Algorithmic Fairness: from Representation Learning to Robustness

Elliot Creager

creager@uwaterloo.ca

VECTOR INSTITUTE

UNIVERSITY OF WATERLOO

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

NOTE: This talk will be a high-level overview of many research topics
Please interrupt me if you get lost or have questions

Algorithmic Fairness: from Representation Learning to Robustness

Elliot Creager

creager@uwaterloo.ca

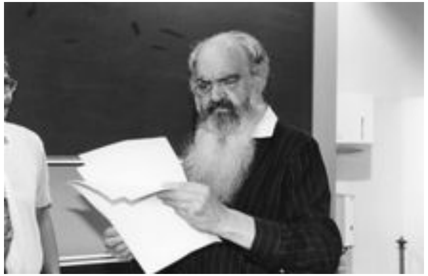
VECTOR INSTITUTE

UNIVERSITY OF WATERLOO

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

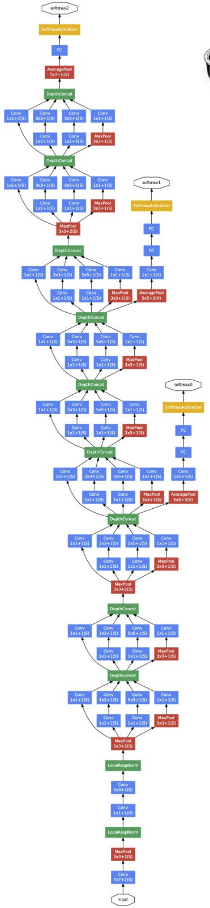
Shankar, S., et al. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.* Neurips Workshops 2017

“The purpose of a system is what it does”

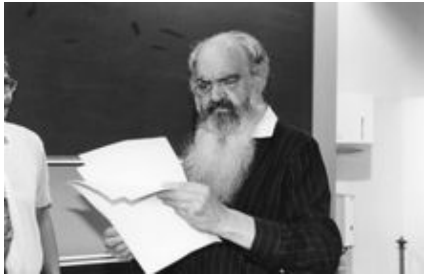


—Stafford Beer

Shankar, S., et al. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.* Neurips Workshops 2017
Beer, S. *What is Cybernetics?,* Kybernetes 2002.



"The purpose of a system is what it does"



—Stafford Beer

Shankar, S., et al. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.* Neurips Workshops 2017
Beer, S. *What is Cybernetics?,* Kybernetes 2002.
Szegedy, C., et al. *Going Deeper with Convolutions.* CVPR 2015

Research agenda…

goal *To build robust and adaptable machine learning algorithms, and apply them responsibly*

methods *Study model failures*

Socially beneficial learning objectives

Scope

Algorithmic fairness:

technical approaches to mitigating algorithmic discrimination

Other approaches:

Investigative journalism, auditing

Policy making and advocacy

Community organizing

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J., 2019. *Fairness and Abstraction in Sociotechnical Systems*
Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., Robinson, D.G., 2020. *Roles for Computing in Social Change.*

Scope

Algorithmic fairness:

technical approaches to mitigating algorithmic discrimination

Other approaches:

Investigative journalism, auditing

Policy making and advocacy

Community organizing

Not a problem to be “solved” by Comp. Sci. alone

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J., 2019. *Fairness and Abstraction in Sociotechnical Systems*
Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., Robinson, D.G., 2020. *Roles for Computing in Social Change.*
Gebru, T., Denton, E. 2021 *NeurIPS Tutorial: Beyond Fairness in Machine Learning*
Ndebele, L., 2022 *Social media companies urged to block hate speech linked to Tigray conflict.*
Mahoozi, S., 2022. *Mahsa Amini death: facial recognition to hunt hijab rebels in Iran*
Barocas, S., Biega, A.J., Fish, B., Niklas, J., Stark, L., 2020. When not to design, build, or deploy

Why is algorithmic fairness challenging?

Subjective

- Many formulations, which may not be compatible

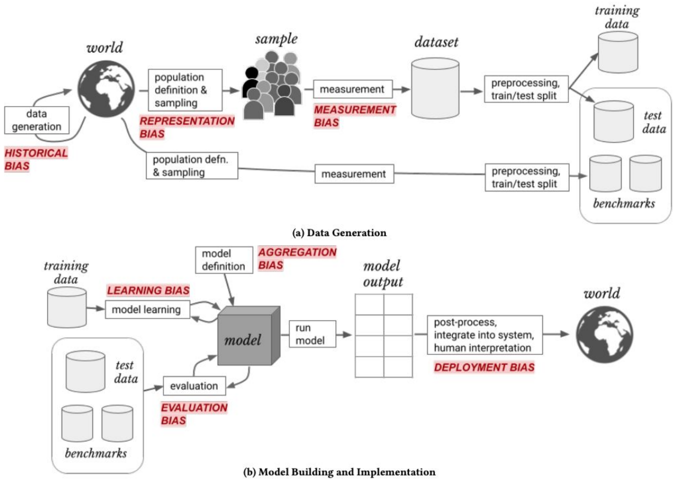
Why is algorithmic fairness challenging?

Subjective

- Many formulations, which may not be compatible

Context-specific

- No one-size-fits-all solution
- Many components in ML pipeline
- “Spurious” associations due to historical inequities



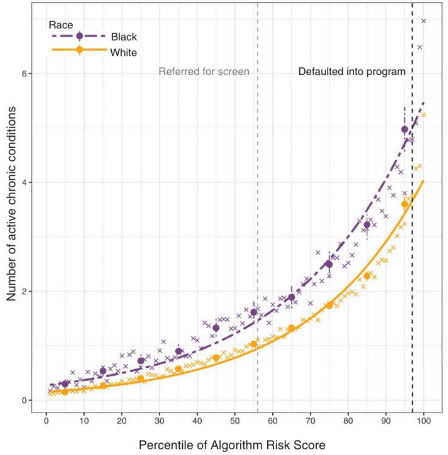
Suresh, H., Guttag, J.V., 2021. *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*

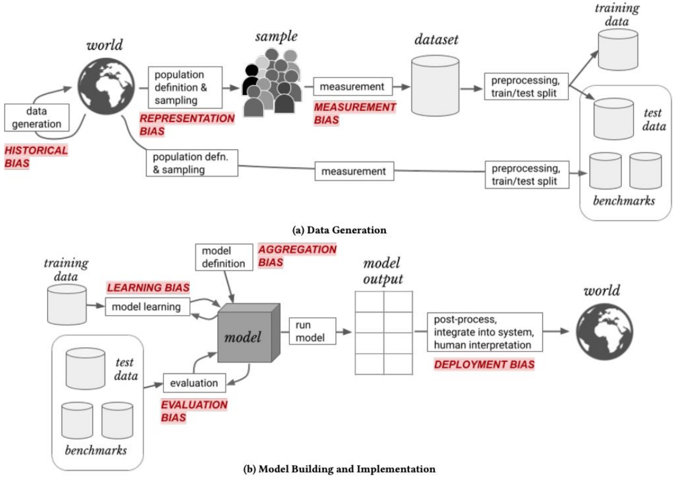
Why is algorithmic fairness challenging?

Subjective
- Many formulations, which may not be compatible

Context-specific
- No one-size-fits-all solution
- Many components in ML pipeline
- “Spurious” associations due to historical inequities

Limited data
- Demographic information often unavailable
- Available data not representative
- Available “targets” may not tell whole story





Suresh, H., Guttag, J.V., 2021. *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*
Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. *Dissecting racial bias in an algorithm used to manage the health of populations.*

Fair Representation Learning

Fair representation learning - intro

Learning Adversarially Fair and Transferable Representations

David Madras [1 2] Elliot Creager [1 2] Toniann Pitassi [1 2] Richard Zemel [1 2]

- Classification: a tale of two parties
- Example: **targeted advertising**: owner $\rightarrow$ vendor $\rightarrow$ prediction

Data owner

Prediction vendor

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Why fairness?

- Want to minimize unfair targeting of disadvantaged groups by vendors
  - e.g. showing ads for worse lines of credit, lower paying jobs
- We want **fair predictions**

Data owner

Prediction vendor

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Why fair representations?

- Previous work emphasized the role of the vendor
- Can we trust the vendor?
- How can the **owner** ensure fairness?

Data owner

Prediction vendor

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Why fair representations?

- How should the data be represented?
  - Feature selection? Measurement?
- How can we choose a data representation that ensures fair classifications downstream?
- Let's *learn* a fair representation!

Data owner→*Representation learner*

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Machine “learning” as fitting probability distributions

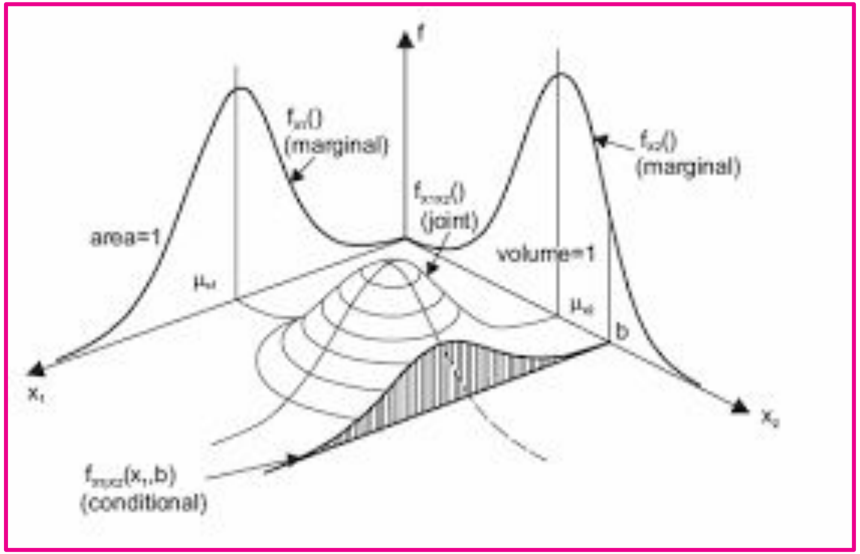
Probability distribution p(X) scores likelihood of X being observed at value x

p(X=x) is a number between 0 and 1

sum of p(X=x) over all possible x is 1

Joint distribution p(X, Y) - likelihood of X=x and Y=y

Conditional distribution p(Y|X) - likelihood of Y=y given X=x



[Melchers 1999]

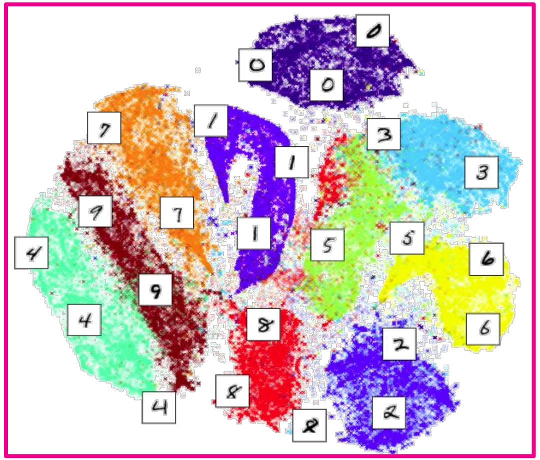
Machine "learning" as fitting probability distributions

Supervised learning - think image recognition (CV)

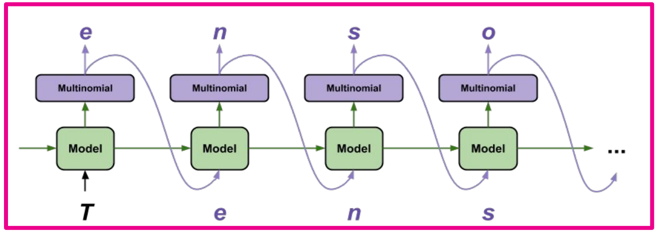
- Conditional distribution fitting
- Use labeled dataset $D=\{(x_i, y_i)\}$
- Train parameterized function f_θ to fit $p(Y|X)$
  - $f_\theta(X=x)[Y] \approx p(Y=y|X=x)$

Unsupervised learning - think language modeling (NLP)

- Marginal/unconditional distribution fitting
- Use unlabeled dataset $D=\{x_i\}$
- Train parameterized function f_θ to fit $p(X)$ or sample from $p(X)$
  - $f_\theta(X=x) \approx p(X=x)$
  - or
  - $X \sim f_\theta()$ with prob. $p(X=x)$



Source: https://nlml.github.io/in-raw-numpy/in-raw-numpy-t-sne/



Source: https://www.tensorflow.org/tutorials/text/text_generation

Fitting probability distributions as optimization

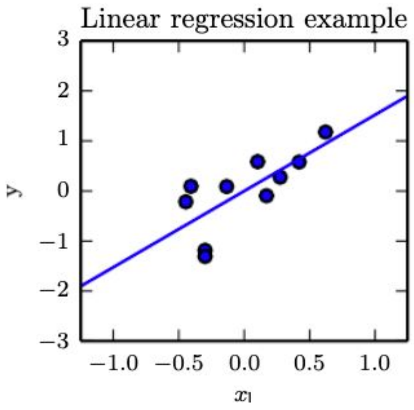
To fit a model to data, write a "loss function" in terms of the model parameters, then minimize it!

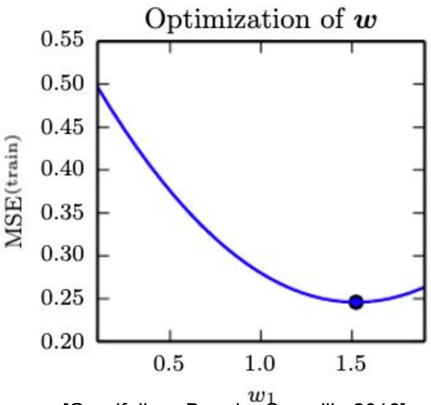
E.g. linear regression: we want *P(y|x)*

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + b,$$

$$\text{Loss}(\boldsymbol{w}) = \frac{1}{m^{(\text{train})}} ||\boldsymbol{X}^{(\text{train})}\boldsymbol{w} - \boldsymbol{y}^{(\text{train})}||_2^2$$

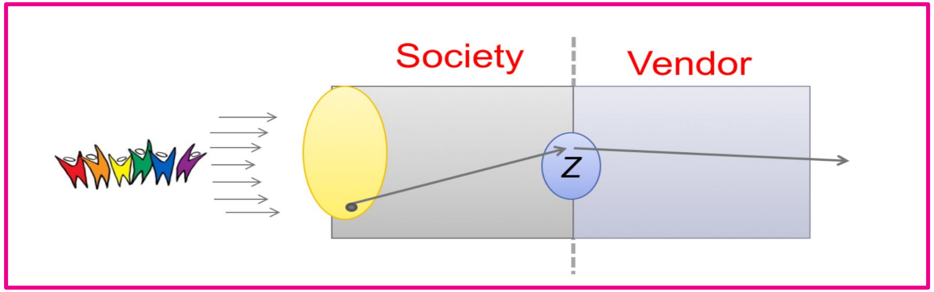
$$\text{Minimize}_{\boldsymbol{w}}\ \text{Loss}(\boldsymbol{w})$$





[Goodfellow, Bengio, Courville 2016]

Fair Representation Learning



Assume: data $X \in \mathbb{R}^d$, label $Y \in 0,1$, sensitive attribute $A \in 0,1$
Goal: predict $\hat{Y}$ fairly with respect to A

- Demographic parity

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

- Equalized odds

$$P(\hat{Y} \neq Y|A = 0, Y = y) = P(\hat{Y} \neq Y|A = 1, Y = y) \; \forall y \in \{0,1\}$$

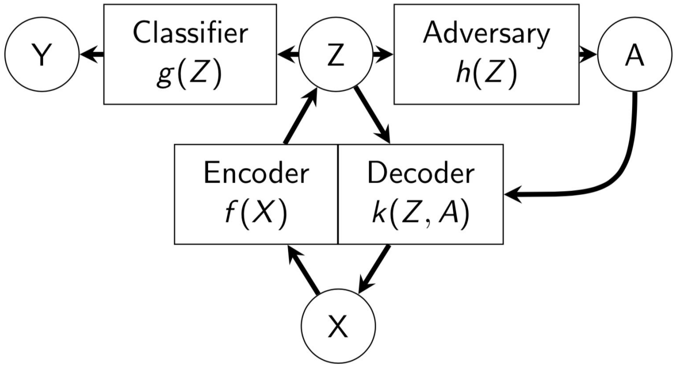
- Equal opportunity: equalized odds with only $Y = 1$

$$P(\hat{Y} \neq Y|A = 0, Y = 1) = P(\hat{Y} \neq Y|A = 1, Y = 1)$$

- Fair classification: learn $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$
  - encoder f, classifier g
- Fair representation: learn $X \xrightarrow{f} Z \xrightarrow{g} \hat{Y}$
- $Z = f(X)$ should:
  - Maintain **useful information** in X
  - **Yield fair downstream classification** for vendors g

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Fair Representation Learning



- Consider two types of unfair vendors
  - The **indifferent** vendor: doesn't care about fairness, only maximizes utility
  - The **malicious** vendor: doesn't care about utility, discriminates maximally
- This suggests an adversarial learning scheme

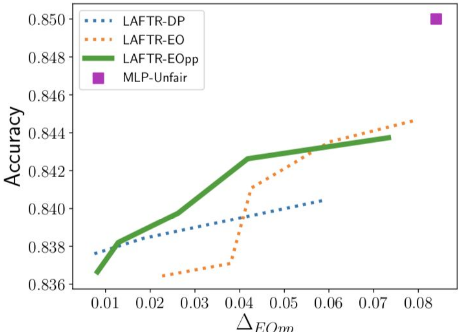
- The classifier is the indifferent vendor, forcing the encoder to make the representations useful
- The adversary is the malicious vendor, forcing the encoder to hide the
- Our game: encoder-decoder-classifier vs. adversary
- Goal: learn a fair encoder

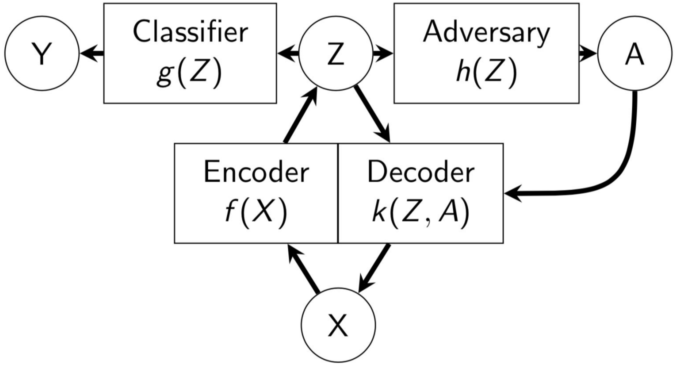
$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[\mathcal{L}(f,g,h,k)\right].$$

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

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Fair Representation Learning





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

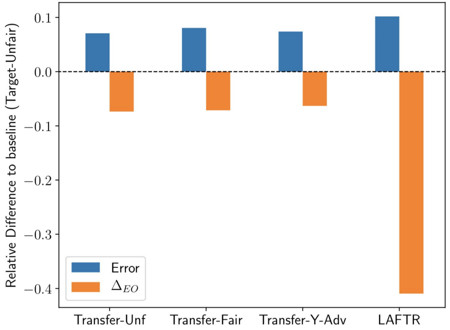
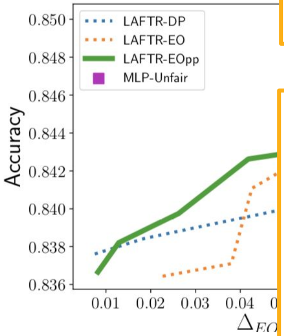


Figure 2: Fair transfer learning on Health dataset. Down is better in both metrics.

Madras, D., **Creager, E.**.,Pitassi, T., and Zemel, R. *Learning Adversarially Fair and Transferable Representations.* ICML 2018

Fair Representation Learning



Task transfer - flexibility in the target label

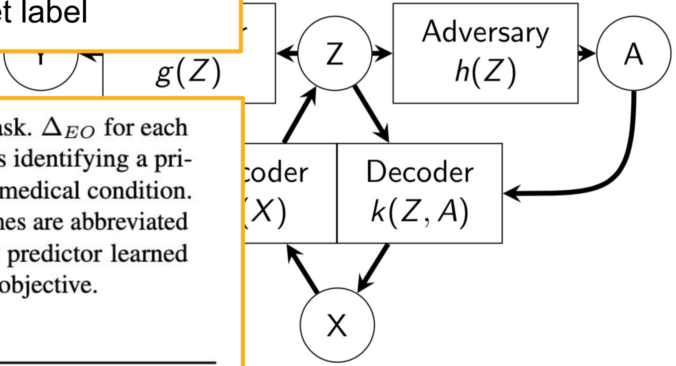
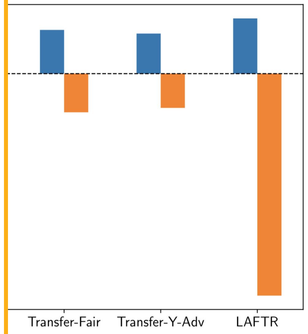


Table 1. Results from Figure 3 broken out by task. Δ_{EO} for each of the 10 transfer tasks is shown, which entails identifying a primary condition code that refers to a particular medical condition. Most fair on each task is bolded. All model names are abbreviated from Figure 3; "TarUnf" is a baseline, unfair predictor learned directly from the target data without a fairness objective.

TRA. TASK	TARUNF	TRAUNF	TRAFAIR	TRAY-AF	LAFTR
MSC2A3	0.362	0.370	0.381	0.378	**0.281**
METAB3	0.510	0.579	**0.436**	0.478	0.439
ARTHSPIN	0.280	0.323	0.373	0.337	**0.188**
NEUMENT	0.419	0.419	0.332	0.450	**0.199**
RESPR4	0.181	0.160	0.223	0.091	**0.051**
MISCHRT	0.217	0.213	0.171	0.206	**0.095**
SKNAUT	0.324	**0.125**	0.205	0.315	0.155
GIBLEED	0.189	0.176	0.141	0.187	**0.110**
INFEC4	0.106	0.042	0.026	**0.012**	0.044
TRAUMA	0.020	0.028	0.032	0.032	**0.019**

- Downstream vendors will have unk
- Does **fairness** transfer?
- We test this as follows:
  1. Train encoder f on data X, with
  2. Freeze encoder f
  3. On new data X', train classifier
  4. Observe fairness and accuracy o
- Compare LAFTR encoder f to oth
- We use Heritage Health dataset
  - Y is Charlson comorbidity index
  - Y' is whether or not a certain ty
  - Check for fairness w.r.t. age



on Health dataset. Down is better in both metrics.

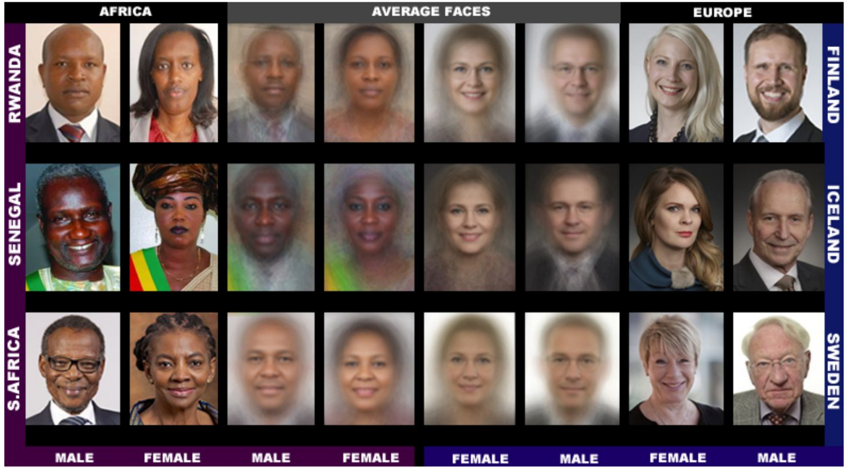
...but no flexibility in the sensitive attribute...

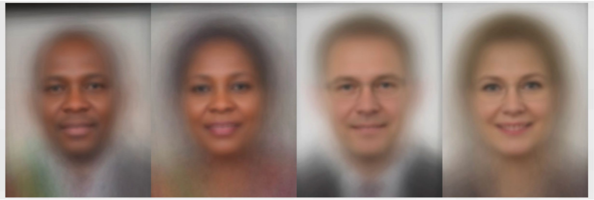
Subgroup Fairness

Subgroup fair representation learning?

Subgroup discrimination

- We would like to handle the case where $\mathbf{a} \in \{0,1\}^{N_a}$ is a vector of sensitive attributes
- ML systems can discriminate against subgroups defined via conjunctions of sensitive attributes (Buolamwini & Gebru, 2018)





99% 60.2% 100% 94.8%

[Adapted from slide by Amirata Ghorbani]

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

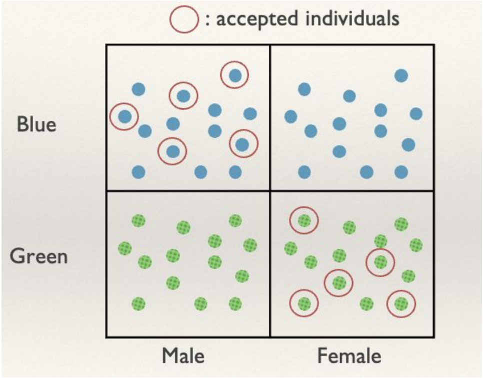
Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

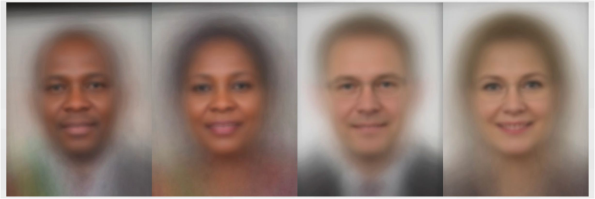
Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Fairness Gerrymandering and Multicalibration/Multiaccuracy

A classifier that is fair w.r.t. groups *A* and *B* can be unfair to their intersection *A* U *B*

[Adapted from slide by Seth Neel]

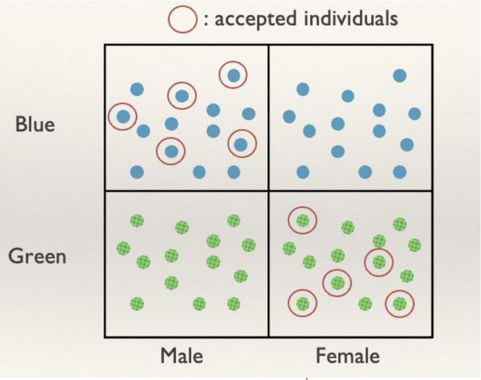




[Adapted from slide by Amirata Ghorbani]

Kearns et al, *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*, ICML 2018.
Hébert-Johnson, et al. *Multicalibration: Calibration for the (computationally-identifiable) masses,* ICML 2018.
Kim, Ghorbani, and Zou, *Multiaccuracy: Black-Box Post-Processing for Fairness in Classification*, AEIS 2019.

[Adapted from slide by Seth Neel]



Fairness Gerrymandering and Multicalibration/Multiaccuracy

A classifier that is fair w.r.t. groups *A* and *B* can be unfair to their intersection *A* U *B*

Possible approach: adaptively choose new groups as training progresses

a) Intersections of existing groups
   e.g. *A* U *C* or *B* U *C* U *D*
b) Infer new ("computationally identifiable") groups directly from data

…a lot like boosting!

Algorithm 1: Multiaccuracy Boost

Given:

- initial hypothesis $f_0 : \mathcal{X} \to [0,1]$
- auditing algorithm $\mathcal{A} : (\mathcal{X} \times [-1,1])^m \to [-1,1]^{\mathcal{X}}$
- accuracy parameter $\alpha > 0$
- validation data $D = D_0, \ldots, D_T \sim \mathcal{D}^m$

Let:

- $\mathcal{X}_0 \leftarrow \{x \in \mathcal{X} : f_0(x) \leq 1/2\}$
- $\mathcal{X}_1 \leftarrow \{x \in \mathcal{X} : f_0(x) > 1/2\}$ `// partition X according to f0`
- $\mathcal{S} \leftarrow \{\mathcal{X}, \mathcal{X}_0, \mathcal{X}_1\}$

Repeat: from $t = 0, 1, \ldots, T$

- For $S \in \mathcal{S}$: `// audit ft on X,X0,X1 with fresh data`
  $h_{t,S} \leftarrow \mathcal{A}(D_t; (f_t - y)_S)$
- $S^* \leftarrow \mathrm{argmax}_{S \in \mathcal{S}} \mathbf{E}_{x \sim D_t}[h_{t,S}(x) \cdot (f_t(x) - y(x))]$ `// take largest residual`
- if $\mathbf{E}_{x \sim D_t}[h_{t,S^*}(x) \cdot (f_t(x) - y(x))] \leq \alpha$:
  return f_t `// terminate when at most alpha`
- $f_{t+1}(x) \propto e^{-\eta h_{t,S^*}(x)} \cdot f_t(x) \quad \forall x \in S^*$ `// multiplicative weights update`

Kearns et al, *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*, ICML 2018.
Hébert-Johnson, et al. *Multicalibration: Calibration for the (computationally-identifiable) masses,* ICML 2018.
Kim, Ghorbani, and Zou, *Multiaccuracy: Black-Box Post-Processing for Fairness in Classification*, AEIS 2019.

Adversarially Reweighted Learning

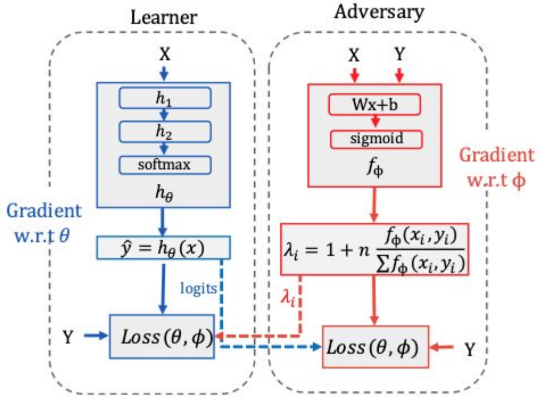


Figure 2: ARL Computational Graph

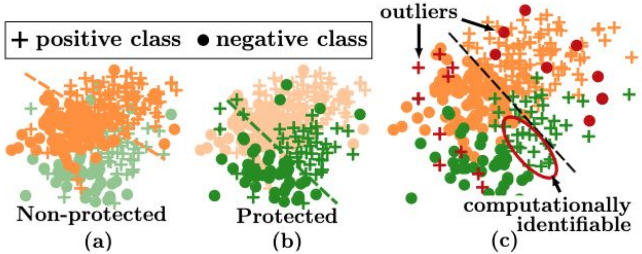


Figure 1: Computational-identifiability example

Table 1: Main results: ARL vs DRO

dataset	method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	Baseline	0.898	0.891	0.867	0.875
Adult	DRO	0.874	0.882	0.843	0.891
Adult	DRO (auc)	0.899	0.908	0.869	0.933
Adult	ARL	**0.907**	**0.915**	**0.881**	**0.942**

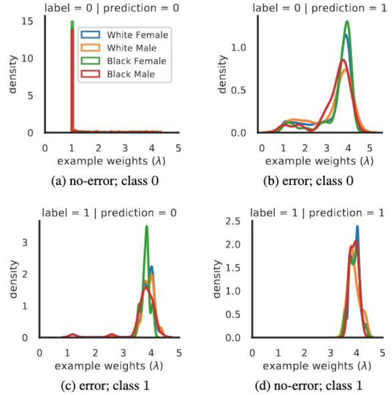


Figure 5: Example weights learnt by ARL.

Adversarial training can also be used to reweight training points

Implicitly this looks for worst-case subgroups

Lahoti, et al, *Fairness without Demographics through Adversarially Reweighted Learning*, NeurIPS 2020

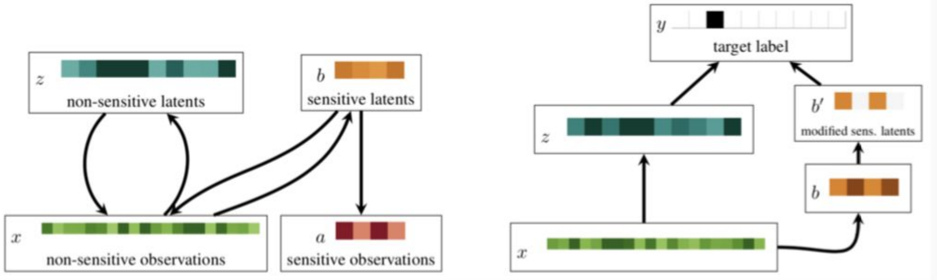
Flexibly fair VAE

We want *flexible fairness*

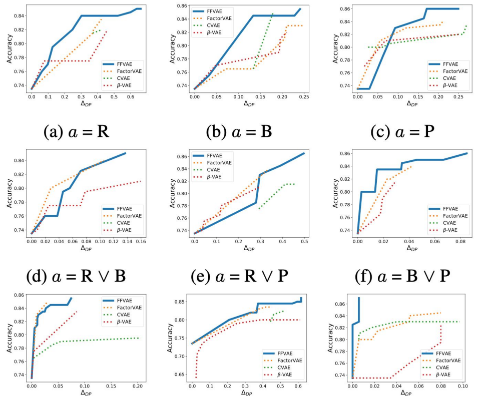
I.e. a single representation that adapts to many distinct downstream fair classification tasks

“Sensitive latents” absorb sensitive observations *and* are disentangled

At task time, noise/zero out desired dimensions of the representation



Data flow at train time (left) and test time (right) for FFVAE



(a) $a = \mathrm{R}$ (b) $a = \mathrm{B}$ (c) $a = \mathrm{P}$

(d) $a = \mathrm{R} \vee \mathrm{B}$ (e) $a = \mathrm{R} \vee \mathrm{P}$ (f) $a = \mathrm{B} \vee \mathrm{P}$

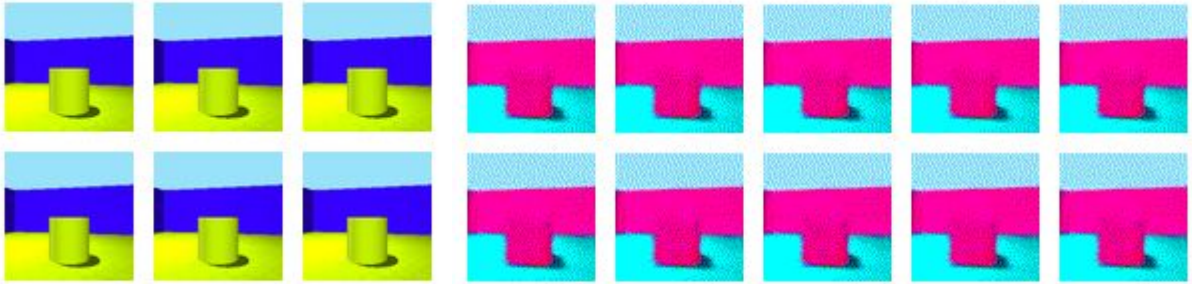
(g) $a = \mathrm{R} \wedge \mathrm{B}$ (h) $a = \mathrm{R} \wedge \mathrm{P}$ (i) $a = \mathrm{B} \wedge \mathrm{P}$

Creager, E.., et al. *Flexibly Fair Representation Learning by Disentanglement.* ICML 2019
Traüble, F.**, Creager, E.**., et al. *On disentangled representations learned from correlated data.* ICML 2021

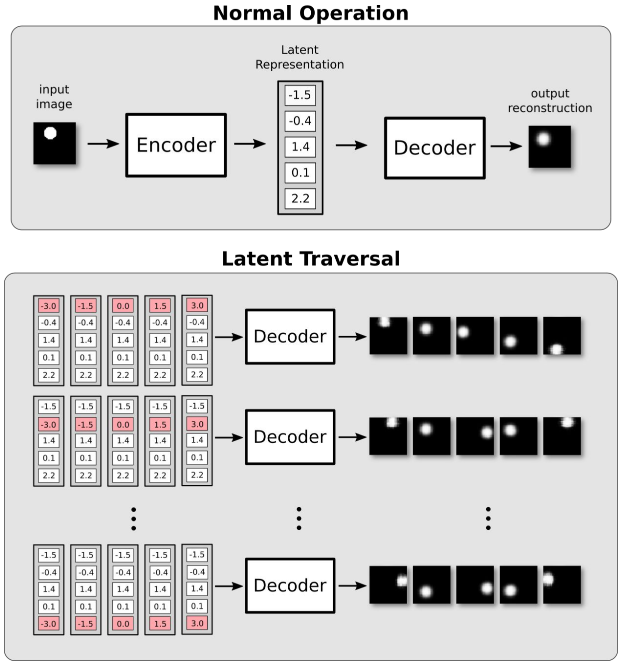
Disentangled representations

“Disentangled” - each dimension of the learned representation has corresponds to no more than one underlying Factor of Variation (FoV)

Observed data *Learned representation*



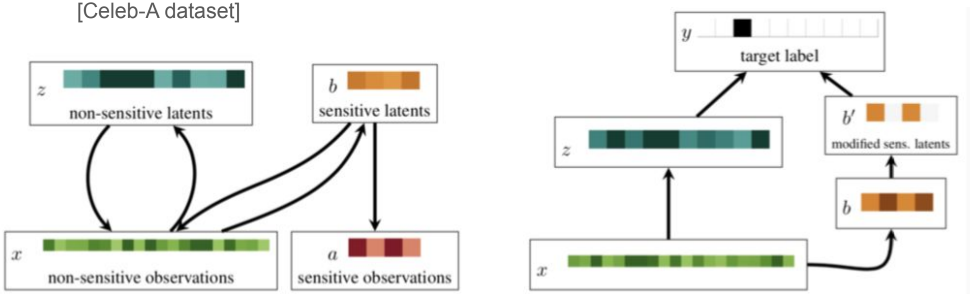
[Source: https://github.com/google-research/disentanglement_lib]



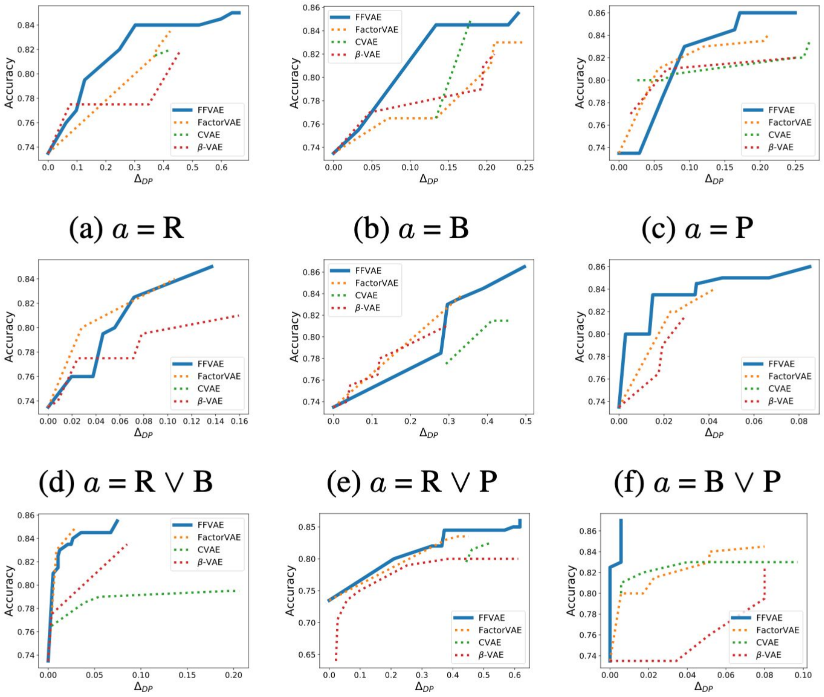
[Source: https://medium.com/@davidlmorton/learning-disentangled-representations-part-1-simple-dots-c5553ecc995b]

Flexibly fair VAE - results

[Celeb-A dataset]



Data flow at train time (left) and test time (right) for FFVAE



(a) $a = \mathrm{R}$ (b) $a = \mathrm{B}$ (c) $a = \mathrm{P}$

(d) $a = \mathrm{R} \vee \mathrm{B}$ (e) $a = \mathrm{R} \vee \mathrm{P}$ (f) $a = \mathrm{B} \vee \mathrm{P}$

(g) $a = \mathrm{R} \wedge \mathrm{B}$ (h) $a = \mathrm{R} \wedge \mathrm{P}$ (i) $a = \mathrm{B} \wedge \mathrm{P}$

Creager, E.., et al. *Flexibly Fair Representation Learning by Disentanglement.* ICML 2019
Traüble, F.**, Creager, E.**., et al. *On disentangled representations learned from correlated data.* ICML 2021

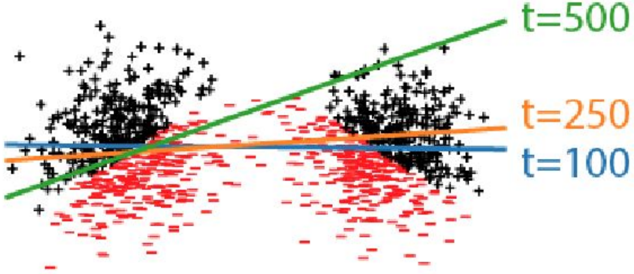
Dynamic Fairness

Short-term Decisions have Long-term Consequences

When ML is used for *decision making*, we have to model long-term effects

ML predictions influence the outside world!

What looks fair today could create future unfairness…



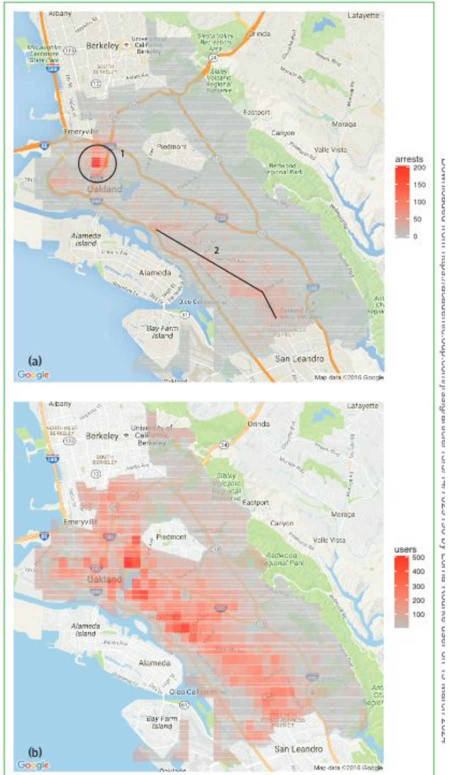
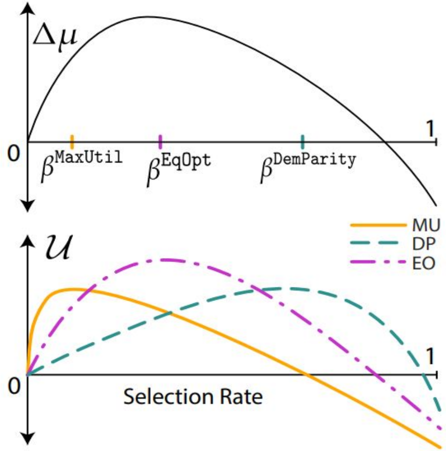


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

Lum, K., and William I. *To predict and serve?* Significance 13.5 (2016): 14-19.
Hashimoto, T., et al. *Fairness without demographics in repeated loss minimization.* ICML 2018.

The Dynamics of Fair Lending



^ Per-group score change for various bank policies

Dynamics in individual credit scores

- X: represents credit score
- A: represents demographic group
- T: represents loan
- Y: represents potential repayment

Treat bank policy (loan predictor) as supervised problem

Evaluated one-step fairness of various constrained classifiers

Structural eqns:

Bank policy *T = f_T(U_T, A, X)*

Potential outcome *Y = f_Y(U_Y, X, A)*

Next-step score *\tilde X = f_{\tilde X}(Y, T, X)*

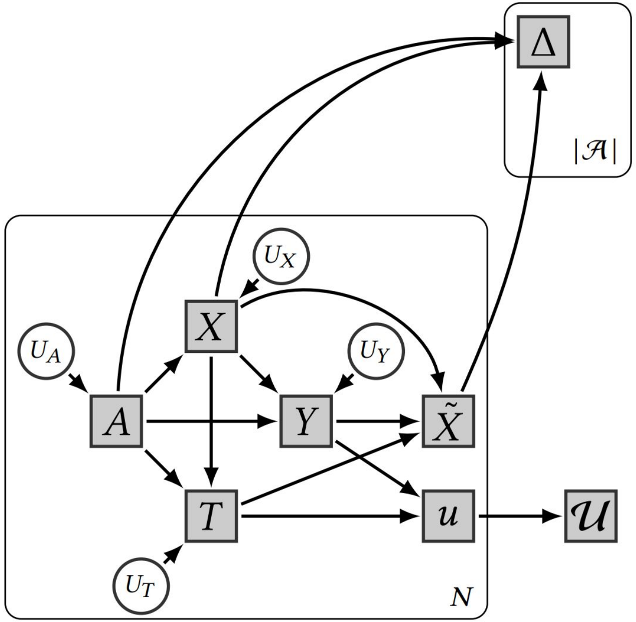
j-th Group avg score improvement *\Delta_j*

- *Computed as avg(\tilde X - X) for group j*

Liu, L. T., et al. *Delayed impact of fair machine learning,* ICML 2018.

The Dynamics of Fair Lending

Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j



Structural eqns:

Bank policy *T = f_T(U_T, A, X)*

Potential outcome *Y = f_Y(U_Y, X, A)*

Next-step score *\tilde X = f_{\tilde X}(Y, T, X)*

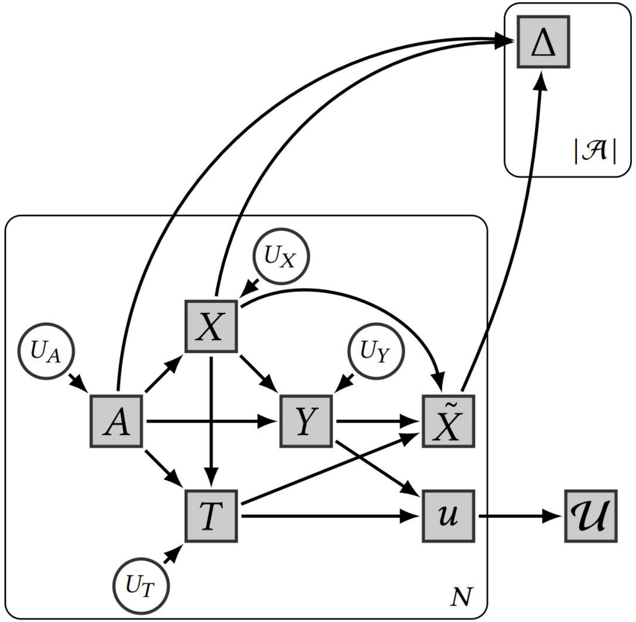
j-th Group avg score improvement *\Delta_j*

- *Computed as avg(\tilde X - X) for group j*

Liu, L. T., et al. *Delayed impact of fair machine learning,* ICML 2018.
Creager, E. et al *Causal modeling for fairness in dynamical systems,* ICML 2020.

The Dynamics of Fair Lending

Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j



Structural eqns:

Bank

Pote

Next

j-th C

- *Computed as avg(\tilde X - X) for group j*

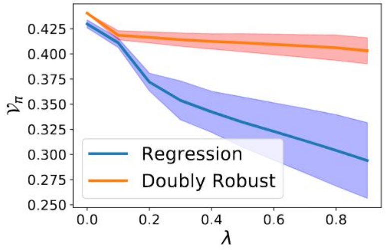


Figure 7: Test set value of a fairness-utility objective using the two off-policy estimators. Hyperparameter λ governs the tradeoff. Higher values of the objective $\mathcal{V}_\pi$ are better.

Liu, L. T., et al. *Delayed impact of fair machine learning,* ICML 2018.
Creager, E. et al *Causal modeling for fairness in dynamical systems,* ICML 2020.

Dynamic fairness: challenges and open questions

How to model the dynamics of social environments

How to balance short- and long-term fairness

Exploration vs exploitation problem: how to learn fair decision making without making too many (unfair) mistakes

Robust Fairness

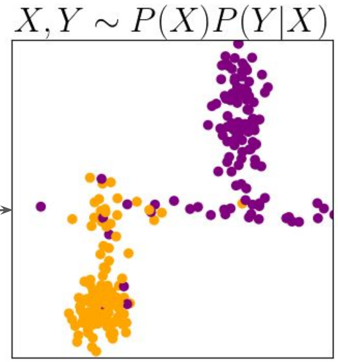
What does it mean to be “robust”?

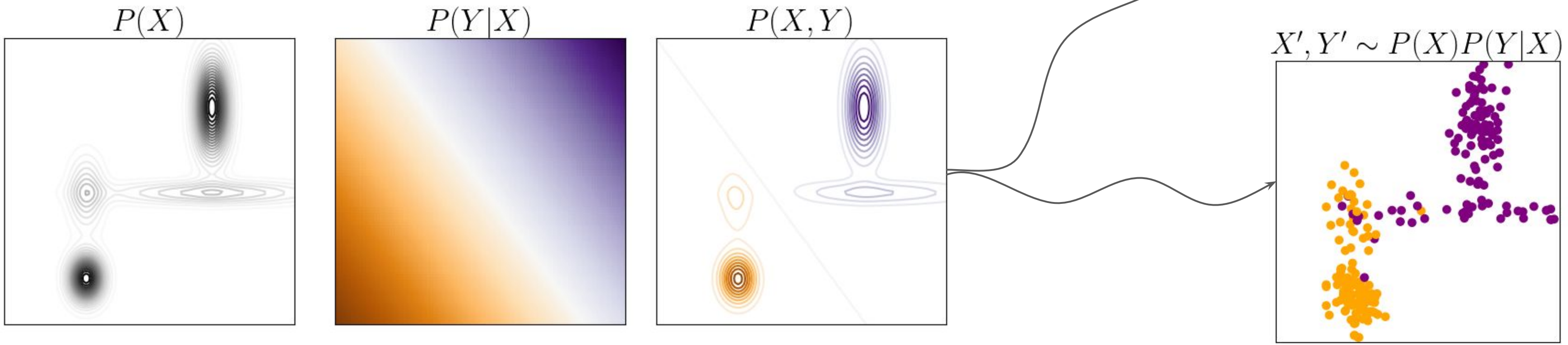
Robustness can have different meanings in different contexts

Recall learning theory: models have bounded error *when data are i.i.d.*

i.i.d. = independent and identically distributed

For “robust” performance, go *beyond* in-distribution generalization





Taxonomy of model failures

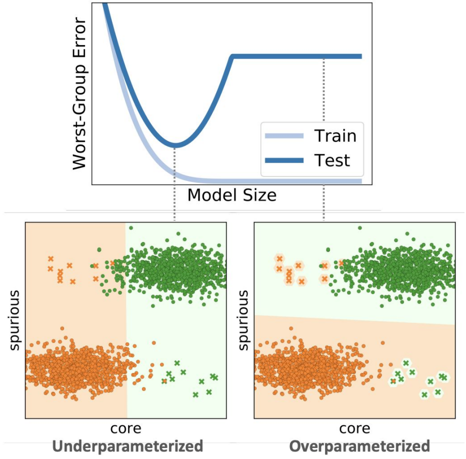
To understand “robustness”, contrast with brittleness of models in practice

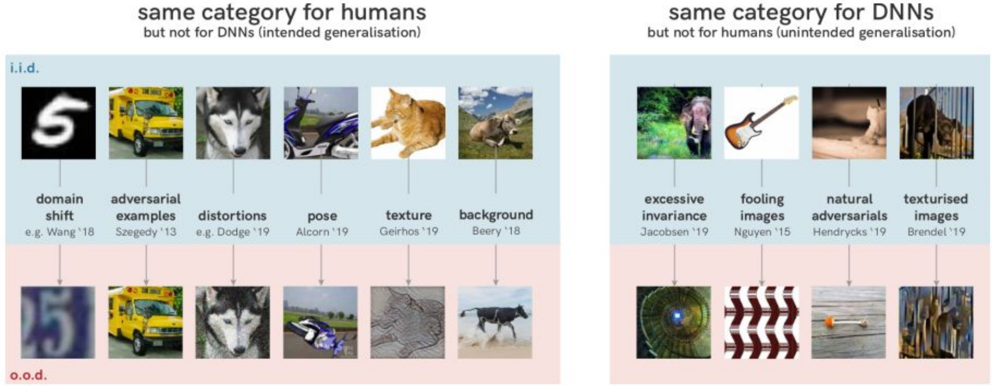
~~Overfitting/underfitting~~ (handled by standard learning theory)

Adversarial examples & security threats

Shortcut learning

Algorithmic discrimination…?





Shah, H., Tamuly, K., Raghunathan, A., Jain, P., Netrapalli, P., 2020. *The Pitfalls of Simplicity Bias in Neural Networks.*
Sagawa, S., Raghunathan, A., Koh, P.W., Liang, P., 2020. *An Investigation of Why Overparameterization Exacerbates Spurious Correlations*
Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut Learning in Deep Neural Networks
D’Amour, A., Heller, K., et al., 2020. *Underspecification Presents Challenges for Credibility in Modern Machine Learning.*

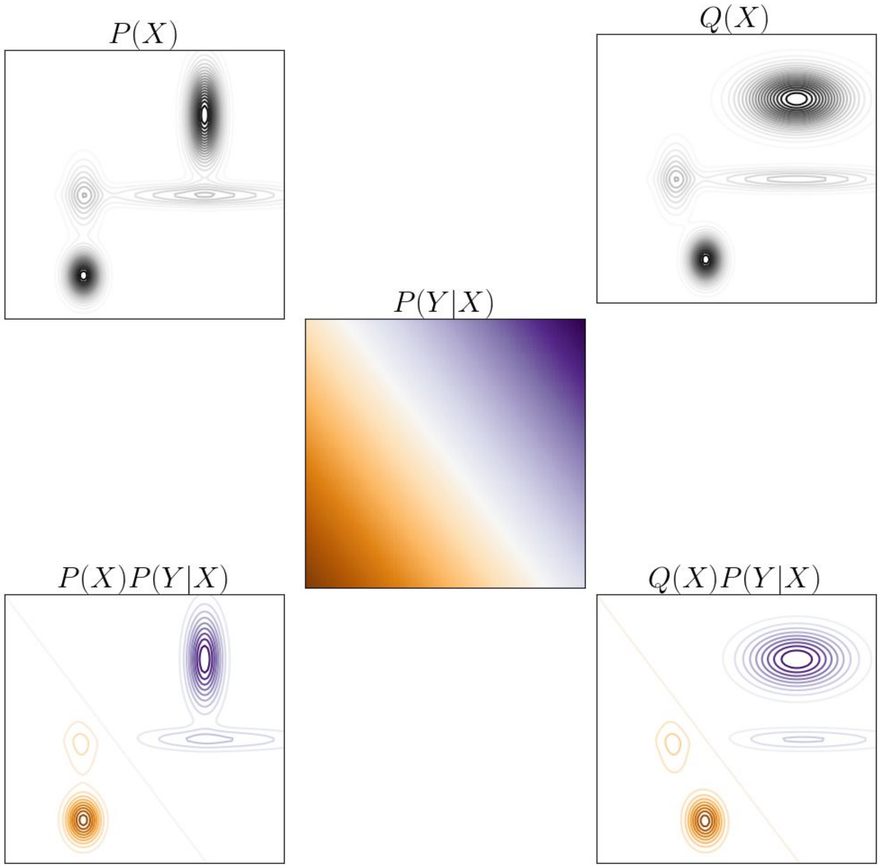
Incorporating “robustness” into learning algorithms

Learning theory provides a “spec” for the model: in-distribution generalization

To learn a “robust” model, we need to define a new spec

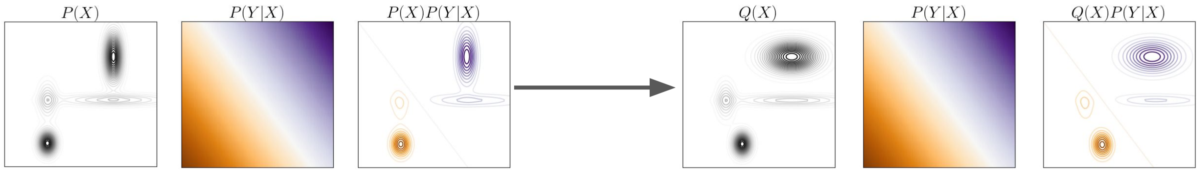
Out-of-distribution (OOD) generalization

What family of distributions should my model handle?

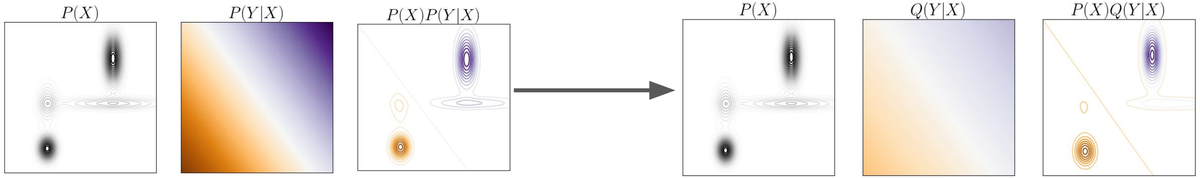


Characterizing distribution shift

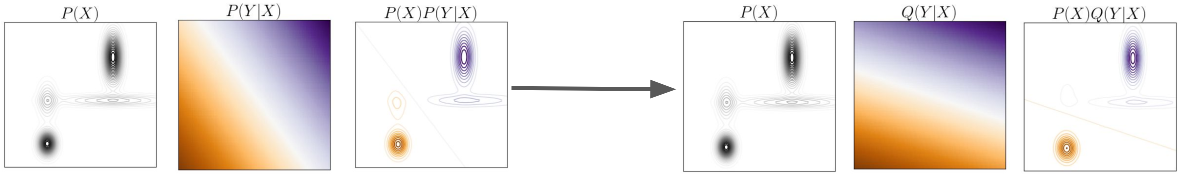
Covariate shift



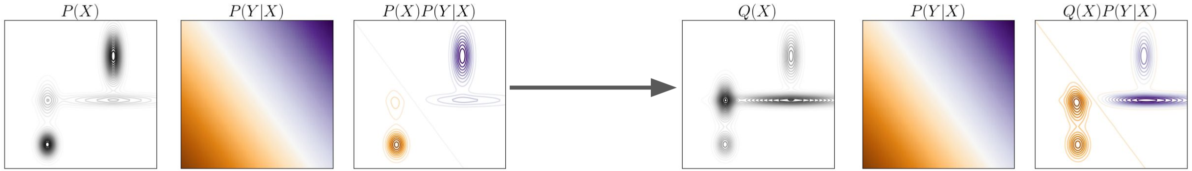
Label noise



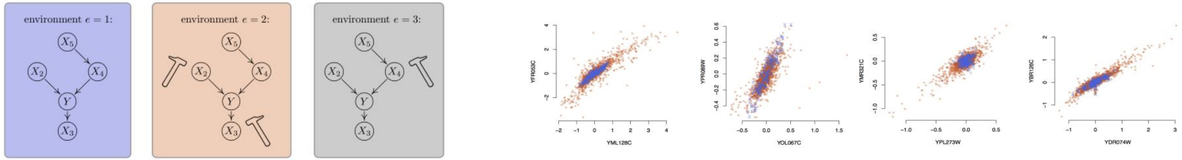
Concept shift



Subpopulation shift



Intervention (on causal graph)



Peters, J., Bühlmann, P., Meinshausen, N., 2015. *Causal inference using invariant prediction: identification and confidence intervals.*

Adversarial Robustness

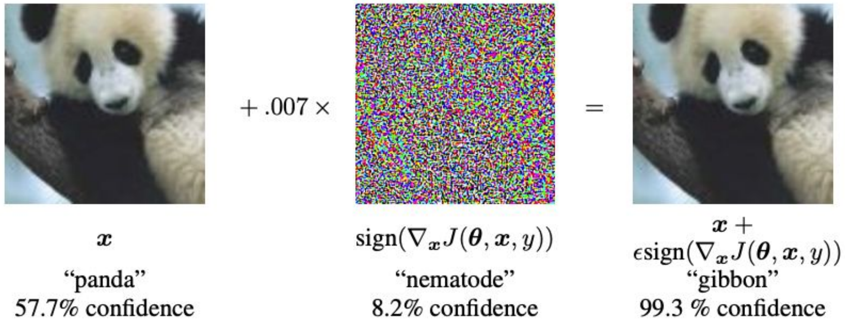
Adversarial examples - small worst-case perturbations in feature space

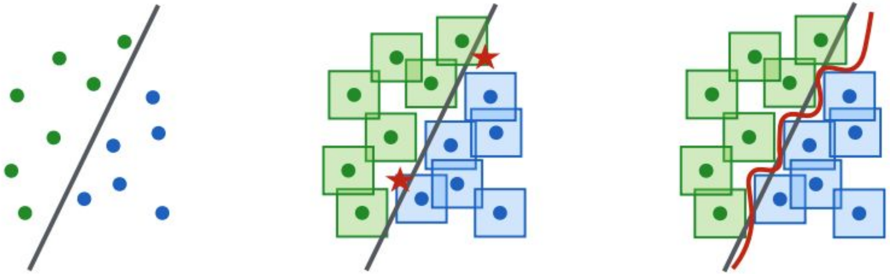
Attacks - white box, black box, …

Adversarial training - train w/ adv. Examples

I.e. train under family of nearby distributions

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y) \right]$$





Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. *Explaining and Harnessing Adversarial Examples.*
Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2019. *Towards Deep Learning Models Resistant to Adversarial Attacks.*

Adversaries “in the wild”

Adversarial examples can be used for *model evasion*

Other security concerns

- Model inversion/data extraction
- Data poisoning

Robustness w.r.t. a specific *threat model*

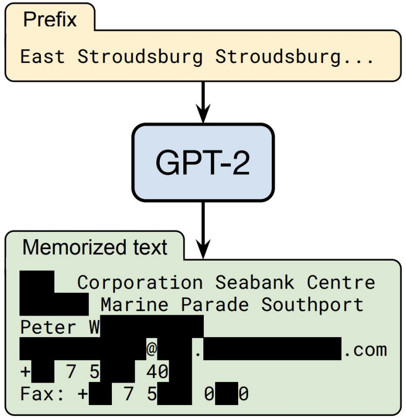
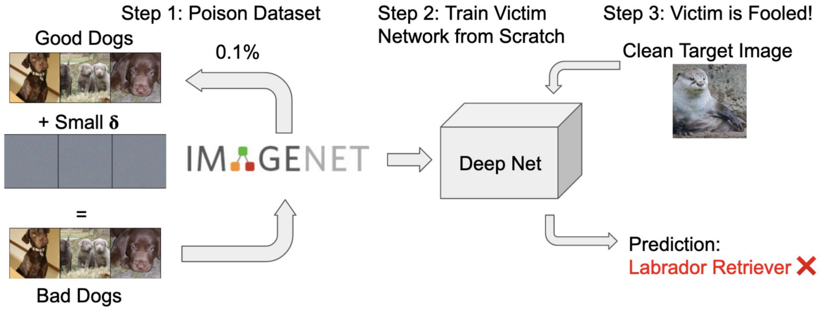


Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person’s name and access to a facial recognition system that returns a class confidence score.



Fredrikson, M., Jha, S., Ristenpart, T., 2015. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*
Geiping, J., Fowl, L., Huang, W.R., Czaja, W., Taylor, G., Moeller, M., Goldstein, T., 2021. *Witches’ Brew: Industrial Scale Data Poisoning via Gradient Matching.*
Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., Raffel, C., 2021. *Extracting Training Data from Large Language Models.*

Distributionally Robust Optimization

Minimize a *worst-case loss* over "nearby" distributions

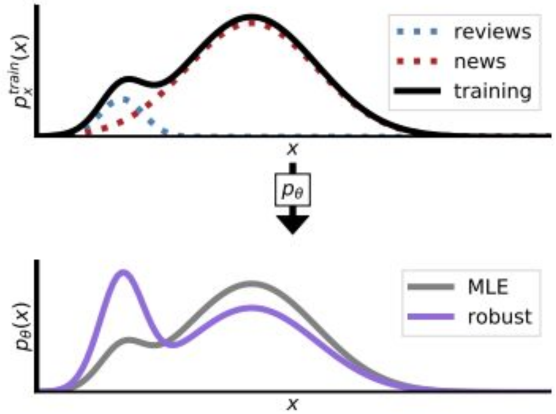
$$\min_{\theta} \max_{Q} \mathbb{E}_Q[\mathcal{L}(X, Y; \theta)] \text{ such that } \; Q \text{ close to } P$$

How to optimize for *Q* when we have samples from *P*?

Importance weighting

$$\mathbb{E}_Q[\mathcal{L}(X, Y; \theta)] = \mathbb{E}_P[\frac{Q(X,Y)}{P(X,Y)}\mathcal{L}(X, Y; \theta)]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \underbrace{\frac{Q(X_i,Y_i)}{P(X_i,Y_i)}}_{\lambda_i \text{ “imp. weight”}} \mathcal{L}(X_i, Y_i; \theta)$$



Group DRO learns just a few importance weights shared by example belonging to the same *group*

Duchi, J., Glynn, P., Namkoong, H., 2018. *Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach.*
Oren, Y., Sagawa, S., Hashimoto, T.B., Liang, P., 2019. *Distributionally Robust Language Modeling*
Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P., 2020. *Distributionally Robust Neural Networks for Group Shifts*

Domain Generalization

Train on data that varies $p(x,y|e)$ across "domains" (a.k.a "environments") e

Learn "core" or "invariant" features

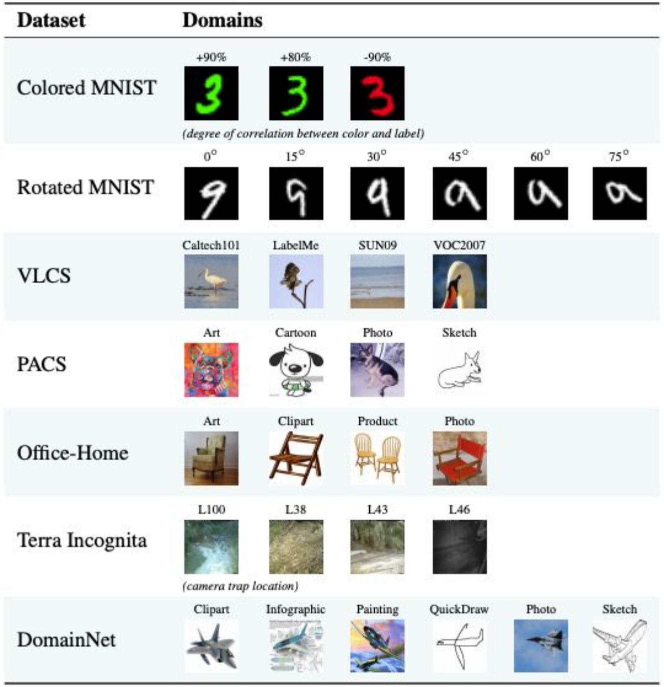
Requires *known* training set partitions, i.e. environment labels

Require OOD generalization to never-before-seen test environment

Typically assume P(Y|X) fixed…P(Y), P(X) may change

Train: cows on grass

Test: cows on beaches



Dataset	Domains					
Colored MNIST	+90%	+80%	-90%			
	(degree of correlation between color and label)					
Rotated MNIST	0°	15°	30°	45°	60°	75°
VLCS	Caltech101	LabelMe	SUN09	VOC2007		
PACS	Art	Cartoon	Photo	Sketch		
Office-Home	Art	Clipart	Product	Photo		
Terra Incognita	L100	L38	L43	L46		
	(camera trap location)					
DomainNet	Clipart	Infographic	Painting	QuickDraw	Photo	Sketch

Beery, Van Horn, and Perona, *Recognition in terra incognita*, ECCV 2018
Gulrajani and Lopez-Paz, *In search of lost domain generalization*, ICLR 2021
Robert Geirhos, et al., *Shortcut Learning in Deep Neural Networks*, Nature Machine Intelligence vol. 2, 2021

Practical Concerns

i.i.d assumption

$(X^{train}, Y^{train}) \sim P$ and $(X^{test}, Y^{test}) \sim P$

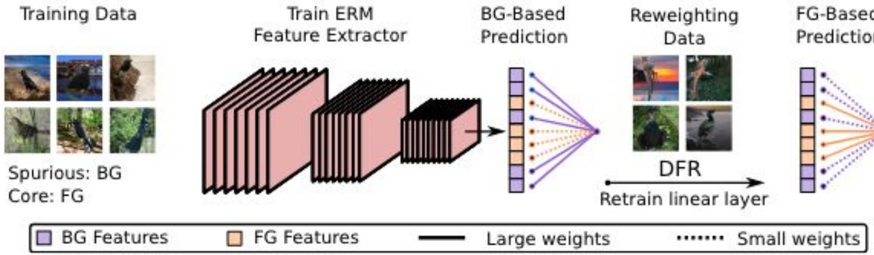
justifies train/validation/test splits

By relaxing the i.i.d. assumption, we break model selection/hyperparameter tuning!

Under fair model selection criteria, ERM (standard training) is hard to beat

If OOD/target data available, adapting ERM features may suffice

Dataset / algorithm	**Out-of-distribution accuracy (by domain)**						
Rotated MNIST	0°	15°	30°	45°	60°	75°	Average
Ilse et al. [2019]	93.5	99.3	99.1	99.2	99.3	93.0	97.2
Our ERM	95.6	99.0	98.9	99.1	99.0	96.7	**98.0**
PACS	A	C	P	S			Average
Asadi et al. [2019]	83.0	79.4	96.8	78.6			84.5
Our ERM	88.1	78.0	97.8	79.1			**85.7**
VLCS	C	L	S	V			Average
Albuquerque et al. [2019]	95.5	67.6	69.4	71.1			75.9
Our ERM	97.6	63.3	72.2	76.4			**77.4**
Office-Home	A	C	P	R			Average
Zhou et al. [2020]	59.2	52.3	74.6	76.0			65.5
Our ERM	62.7	53.4	76.5	77.3			**67.5**



Gulrajani, I., Lopez-Paz, D., 2020. *In Search of Lost Domain Generalization.*
Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S., 2021. *Long-tail Learning via Logit Adjustment*
Kirichenko, P., Izmailov, P., Wilson, A.G., 2022. *Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations.*

Fairness & Robustness: Learning Objectives

Under what settings are fair learning and robust learning equivalent?

What lessons can be exchanged between the research areas?

- Methods
- Data
- Articulating assumptions + limitations

Statistic to match/optimize	e known?	DG method	Fairness method
match $\mathbb{E}[\ell \mid e]\ \forall e$	yes	REx (Krueger et al., 2021),	CVaR Fairness (Williamson & Menon, 2019)
$\min \max_e \mathbb{E}[\ell \mid e]$	yes	Group DRO (Sagawa et al., 2020)	
$\min \max_q \mathbb{E}_q[\ell]$	no	DRO (Duchi et al., 2021)	Fairness without Demographics (Hashimoto et al., 2018; Lahoti et al., 2020)
match $\mathbb{E}[y \mid \Phi(x), e]\ \forall e$	yes	IRM (Arjovsky et al., 2019)	Group Sufficiency (Chouldechova, 2017; Liu et al., 2019)
match $\mathbb{E}[y \mid \Phi(x), e]\ \forall e$	no	**EIIL (ours)**	**EIIL (ours)**
match $\mathbb{E}[\hat{y} \mid \Phi(x), e, y = y']\ \forall e$	yes	C-DANN (Li et al., 2018) PGI (Ahmed et al., 2021)	Equalized Odds (Hardt et al., 2016)
match $\left\lvert \mathbb{E}[y \mid S(x), e] - \mathbb{E}[\hat{y}(x) \mid S(x), e] \right\rvert\ \forall e$	no		Multicalibration (Hébert-Johnson et al., 2018)
match $\left\lvert \mathbb{E}[y \mid e] - \mathbb{E}[\hat{y}(x) \mid e] \right\rvert\ \forall e$	no		Multiaccuracy (Kim et al., 2019)
match $\left\lvert \mathbb{E}[y \neq \hat{y}(x) \mid y = 1, e] \right\rvert\ \forall e$	no		Fairness Gerrymandering (Kearns et al., 2018)

Table 1. Domain Generalization (DG) and Fairness methods can be understood as matching or optimizing some statistic across conditioning variable e, representing "environment" or "domains" in DG and "sensitive" group membership in the Fairness. Φ and S are learned vector and scalar functions of the inputs, respectively.

Creager, E., Jacobsen, J.-H., Zemel, R., 2021. *Environment Inference for Invariant Learning*

Lessons from robustness to fairness

Formal framework for characterizing distribution shift and model failure

“My data is biased; let’s collect more”

↓

“My model needs to handle covariate shift; assuming fixed P(Y|X), let’s improve coverage over P(X)”

Methods for improving OOD generalization

Algorithmic fairness as OOD generalization

Some unfairness comes from failure to generalize “out of distribution” (OOD)

Recall: subpopulation shift

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D., 2017. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.*

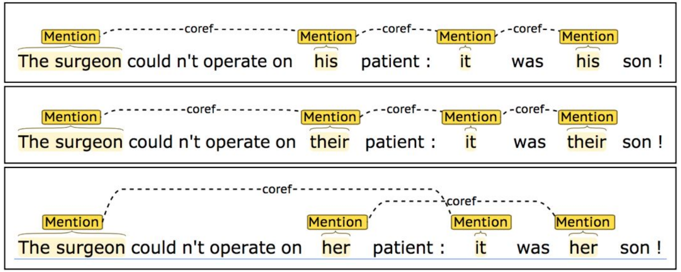
Algorithmic fairness as OOD generalization

Some unfairness comes from failure to generalize "out of distribution" (OOD)

Recall: subpopulation shift

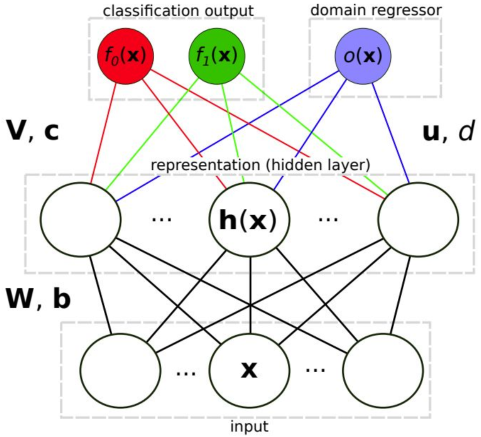
Some "shifts" in data are extremely subtle

E.g. bias in coreference resolution



Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D., 2017. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.*

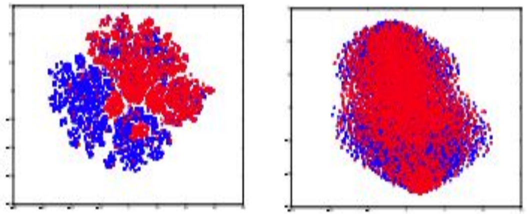
Representation learning approaches

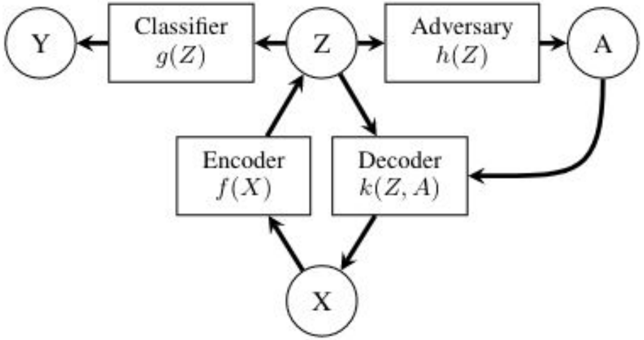


Neural net approaches to statistical fairness influenced by domain adaptation

E.g. adversarial training with auxiliary labels

“Fair” representations can transfer to new tasks





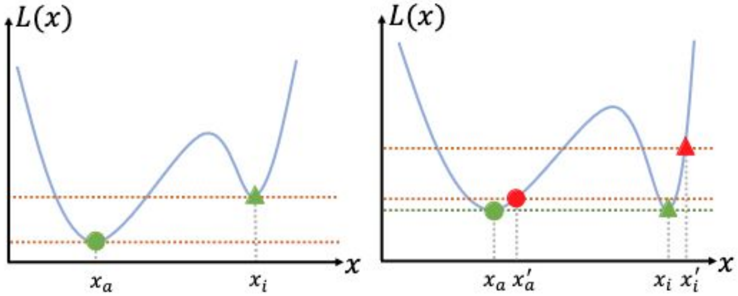
TRA. TASK	TARUNF	TRAUNF	TRAFAIR	TRAY-AF	LAFTR
MSC2A3	0.362	0.370	0.381	0.378	**0.281**
METAB3	0.510	0.579	**0.436**	0.478	0.439
ARTHSPIN	0.280	0.323	0.373	0.337	**0.188**
NEUMENT	0.419	0.419	0.332	0.450	**0.199**
RESPR4	0.181	0.160	0.223	0.091	**0.051**
MISCHRT	0.217	0.213	0.171	0.206	**0.095**
SKNAUT	0.324	**0.125**	0.205	0.315	0.155
GIBLEED	0.189	0.176	0.141	0.187	**0.110**
INFEC4	0.106	0.042	0.026	**0.012**	0.044
TRAUMA	0.020	0.028	0.032	0.032	**0.019**

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., 2015. *Domain-Adversarial Neural Networks.*
Edwards, H., Storkey, A., 2016. *Censoring Representations with an Adversary.*
Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R., 2017. The Variational Fair Autoencoder.
Madras, D., **Creager, E.**, Pitassi, T., Zemel, R., 2018. *Learning Adversarially Fair and Transferable Representations.*

Limitations of Representation Learning

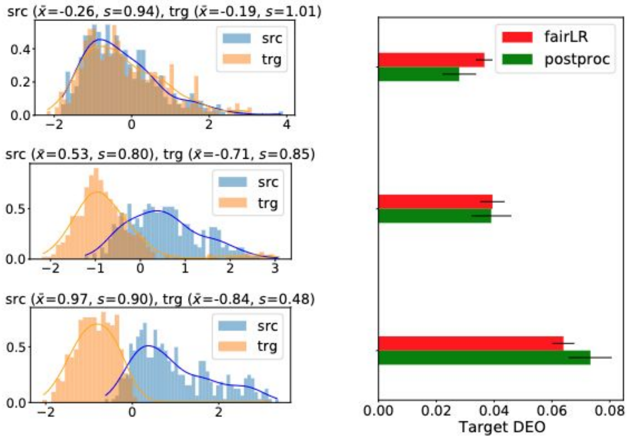
Just like standard ML, fair predictors can fail under distribution shift

Theory shows that even “transferable” representations can fail under dramatic distribution shifts



(a) Normal learning (Unfair)

(b) Traditional fairness learning (In-distribution fairness)



Wang, H. et al, *How Robust is Your Fairness? Evaluating and Sustaining Fairness under Unseen Distribution Shifts*, TMLR 2023
Rezaei, A. et al, *Robust Fairness under Covariate Shift*, AAAI 2021
Lechner, T. et al *Impossibility Results for Fair Representations*

Fair *and* robust learning

Fair representations can fail under distribution shifts

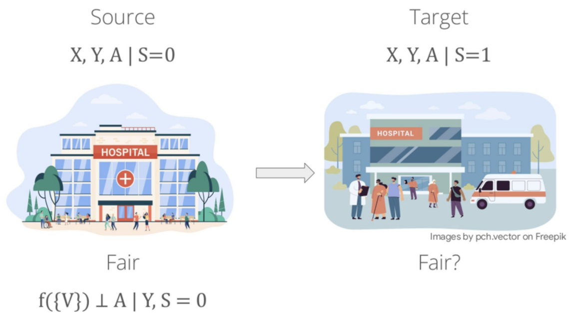
Fair learning + DRO helps

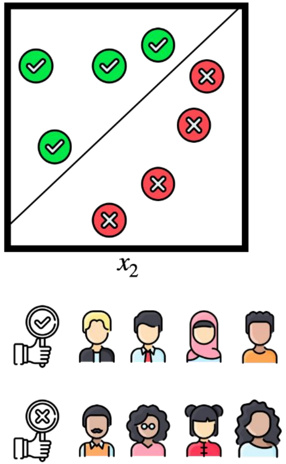
Mostly simulated studies

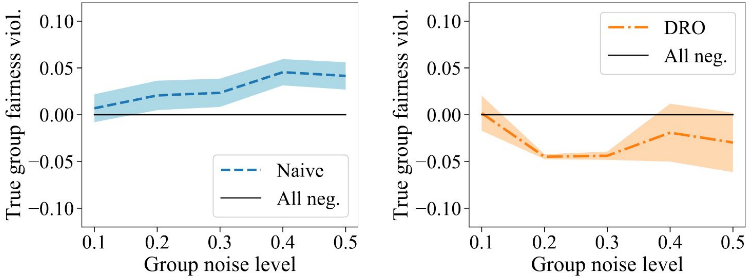
Noisy observations

Sensitive attributes

Targets (esp. in risk assessment)







Lechner, T., Ben-David, S., Agarwal, S., Ananthakrishnan, N., 2021. *Impossibility results for fair representations.*
Rezaei, A., Liu, A., Memarrast, O., Ziebart, B., 2021. *Robust Fairness under Covariate Shift.*
Singh, H., Singh, R., Mhasawade, V., Chunara, R., 2021. *Fairness Violations and Mitigation under Covariate Shift*
Fogliato, R., Chouldechova, A., G'Sell, M., 2020. *Fairness Evaluation in Presence of Biased Noisy Labels*
Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., Jordan, M., 2020. *Robust Optimization for Fairness with Noisy Protected Groups*
Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., Dieng, A., Liu, Y., Natarajan, V., Karthikesalingam, A., Heller, K., Chiappa, S., D'Amour, A., 2022 .*Diagnosing failures of fairness transfer across distribution shift in real-world medical settings*

Fairness/robustness: challenges and open questions

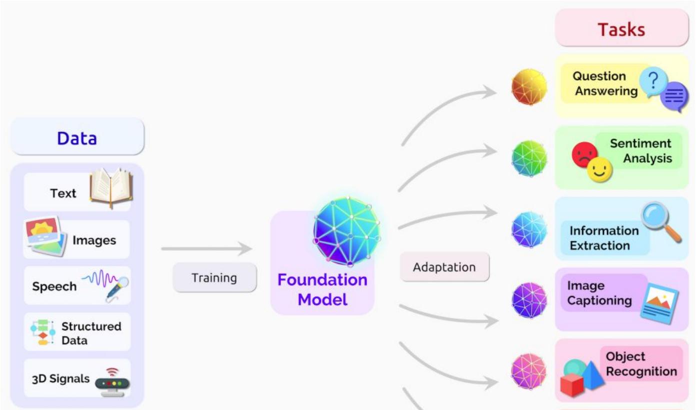
How to characterize and measure distribution shifts relevant to algorithmic discrimination?

Can we formulate causal models for data bias in practical settings?

How to ensure statistically fair models are robust to distribution shift?

What's next?

Improving fairness and robustness of **foundation models**



Modern representation learning looks different…

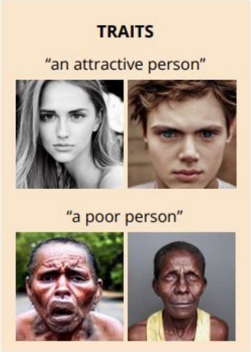
> Train across web-scale data

> No labels

> Multiple data modalities (image, text, …)

…these **foundation models** are adapted for many tasks

Internal representations of these models contain problematic stereotypes



Bommasani, R., et al. *On the Opportunities and Risks of Foundation Models.* Technical Report 2022
Beer, S. *What is Cybernetics?,* Kybernetes 2002.
Bianchi et al, *Easily accessible text-to-image generation amplifies demographic stereotypes at large scale.* FAccT 2023.

Summary

My lab is focused on machine learning and its the societal implications

Within this research agenda, a key area is *Algorithmic Fairness*

- Fair Representation Learning
- Subgroup Fairness
- Dynamic Fairness
- Robust Fairness

creager@uwaterloo.ca
ecreager.github.io

VECTOR INSTITUTE

UNIVERSITY OF WATERLOO

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY