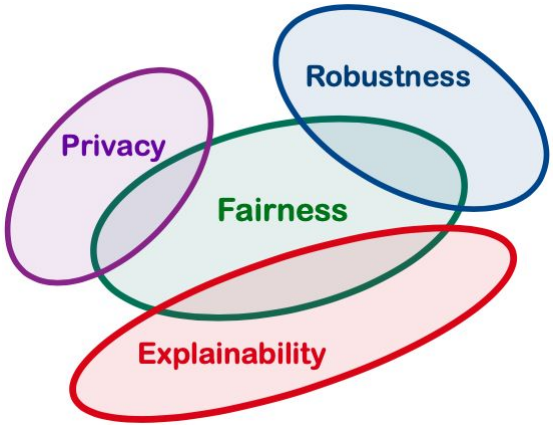
Algorithmic Fairness: at the Intersections

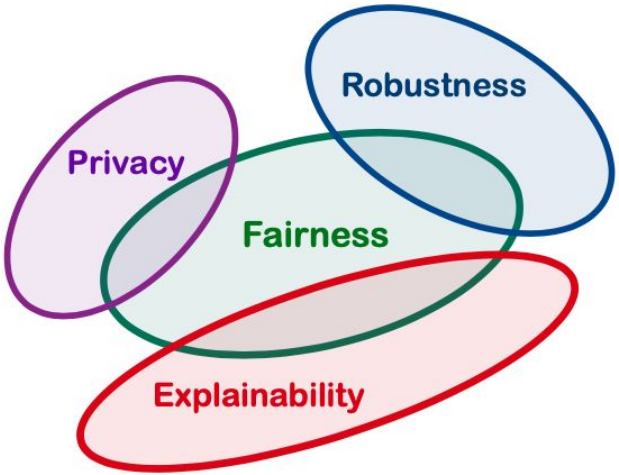


Golnoosh Farnadi[1], Q. Vera Liao[2], Elliot Creager[3]

https://sites.google.com/mila.quebec/fairnesstutorial/
NeurIPS 2022 | New Orleans, LA | December 5, 2022
[1]HEC Montréal/Université de Montréal and MILA [2]Microsoft Research [3]University of Toronto and Vector Institute

Objectives

- Motivate study of “fairness” in machine learning
- Overview of four topics
  - **fairness**, privacy, robustness and explainability
- Highlight their intersections
  - Why think about these topics together?
    - What new opportunities and challenges arise?
  - Open research questions
- Learning in context
  - Each method has its own scope and limitations
  - “Fair” ML work not only approach to mitigating AI harm



Participation

Join us in the Zoom meeting

Type your questions in the chat; we will address them at the end of the talk

We will attempt to monitor the RocketChat as well :)

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

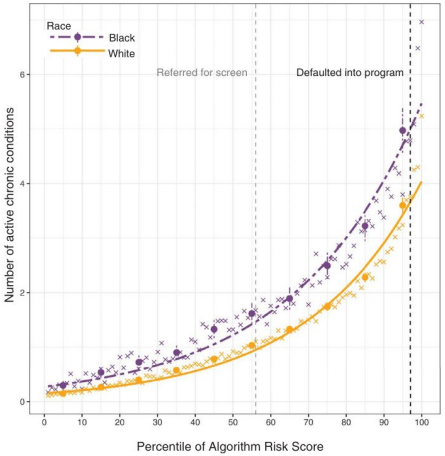
- Many formulations, which may not be compatible

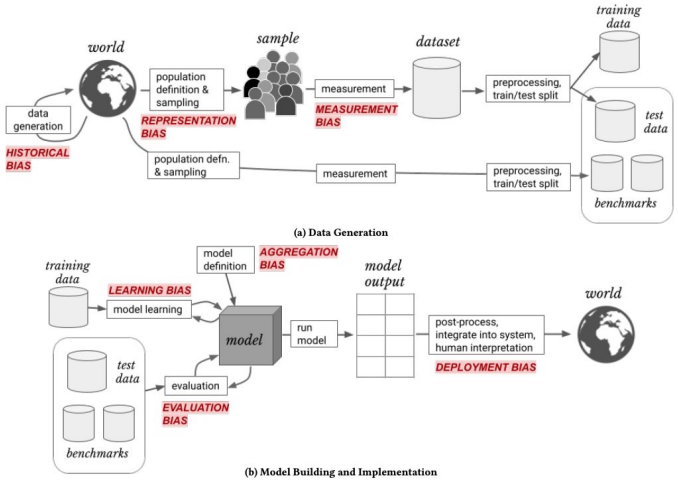
Context-specific

- 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

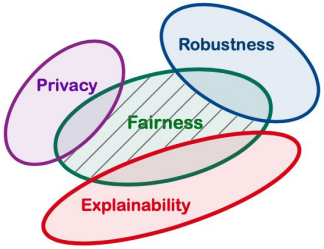




Suresh, H., Guttag, J.V., 2021. *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*

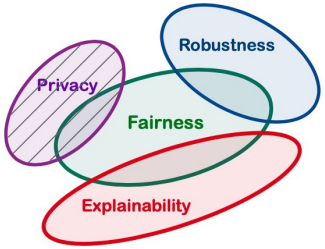
Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. *Dissecting racial bias in an algorithm used to manage the health of populations.*

Overview



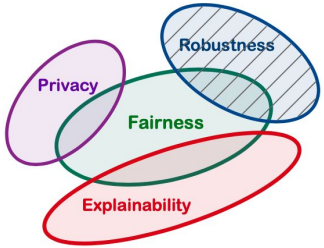
[15m] **Fairness overview**

Golnoosh Farnadi



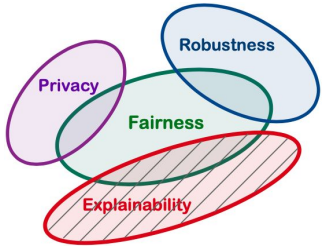
[30m] **Privacy overview + Fairness/Privacy**

Golnoosh Farnadi



[30m] **Robustness overview + Fairness/Robustness**

Elliot Creager



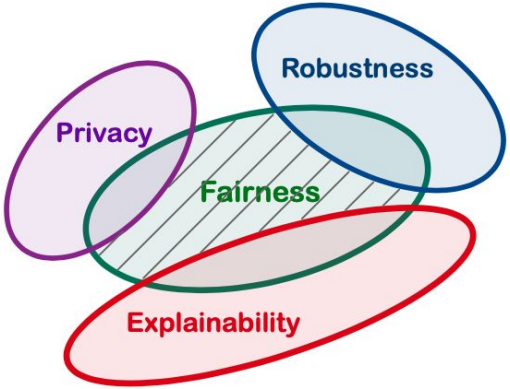
[30m] **Explainability overview + Fairness/Explainability**

Q. Vera Liao

[15m] **Q&A**

[30m] **Panel**

```
Moderator: Su Lin Blodgett
Panelists: Reza Shokri
           Ferdinando Fioretto
           Amir-Hossein Karimi
           Pratyusha Kalluri
           Elizabeth Anne Watkins
```



Introduction to Algorithmic Fairness

Why algorithmic discmrination matters?

- **Allocation harm**: E.g., Amazon Hiring system, COMPAS risk assessment
- **Quality of service harm**: E.g., gender shades, VMS make women sick
- **Stereotyping harm**, e.g., Black criminality in predictive policing, gender issues in NLP (in translation)
- **Denigration harm**, e.g., mislabeling images of Black women as Gorillas, Chatbot Tay for hate speech
- **Over and under-representation harm**, e.g., images of men in image search results

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning.Special Interest Group for Computing, Information and Society (SIGCIS)(2017)
Kate Crawford at NeurIPS 2017 Tutorial

Laws against Discrimination

Legally recognized 'protected classes'

Race (Civil Rights Act of 1964)
Color (Civil Rights Act of 1964)
Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
Religion (Civil Rights Act of 1964)
National origin (Civil Rights Act of 1964)
Citizenship (Immigration Reform and Control Act)
Age (Age Discrimination in Employment Act of 1967)
Pregnancy (Pregnancy Discrimination Act)
Familial status (Civil Rights Act of 1968)
Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

sensitive attributes

Regulated domains

Credit (Equal Credit Opportunity Act)
Education (Civil Rights Act of 1964; Education Amendments of 1972)
Employment (Civil Rights Act of 1964)
Housing (Fair Housing Act)
Public Accommodation (Civil Rights Act of 1964)
Extends to marketing and advertising; not limited to final decision
This list sets aside complex web of laws that regulates the government

Canadians have the right to be treated fairly in workplaces free from discrimination, and our country has laws and programs to protect this right. **The Canadian Human Rights Act** is a broad-reaching piece of legislation that prohibits discrimination on the basis of gender, race, ethnicity and other grounds. May 30, 2022

https://laws-lois.justice.gc.ca/eng/acts/h-6/fulltext.html

Initiating an Anti-Discrimination Regime in China

The 1982 Constitution has enshrined the principle of equality of all citizens before the law (Article 33). Articles 4, 36, 48, and 89 also guarantee the rights of ethnic minorities, religious freedom and gender equality and prohibits discrimination on those grounds.

Article 14. Equality before law. -The State shall not deny to any person equality before the law or the equal protection of the laws within the territory of India. (1) **The State shall not discriminate against any citizen on grounds only of religion, race, caste, sex, place of birth or any of them**.

EU Charter of Fundamental Rights

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited. 2.

https://www.refworld.org/pdfid/4d886bf02.pdf

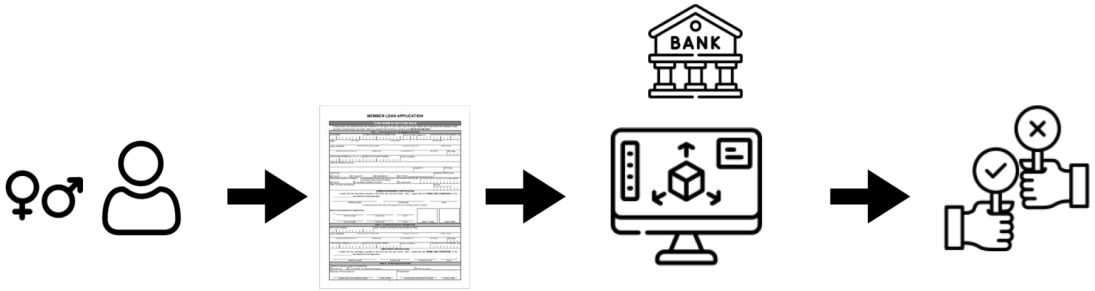
The basis for progressively redressing these conditions lies in the Constitution which, amongst others, upholds the values of human dignity, equality, freedom and social justice in a united, non-racial and non-sexist society where all may flourish;

South Africa also has international obligations under binding treaties and customary international law in the field of human rights which promote equality and prohibit unfair discrimination. Among these obligations are those specified in the Convention on the Elimination of All Forms of Discrimination Against Women and the Convention on the Elimination of All Forms of Racial Discrimination;

Running Example

Confusion Matrix

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	TP	FP
$\hat{Y} = 0$	FN	TN



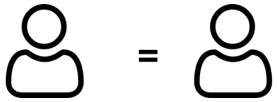
S	X	$\hat{Y}$	Y
sensitive attribute	non-sensitive attributes	Prediction decision	Actual Outcome

Note gender is assumed to be binary for the sake of simplicity

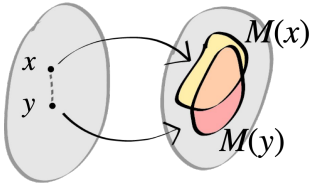
Scheuerman, M.K., Brubaker, J.R., 2018 *Gender is not a Boolean: Towards Designing Algorithms to Understand Complex Human Identities.*
Hu, L., Kohler-Hausmann, I., 2020. *What's Sex Got To Do With Fair Machine Learning?*
Lu, C., Kay, J., McKee, K., 2022. *Subverting machines, fluctuating identities: Re-learning human categorization.*

What does fairness in ML mean?

Individual: measures the impact that discrimination has on the individuals



E.g., similar applicants, should have similar probability of receiving positive loan approval

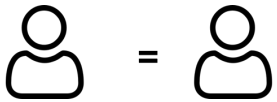


The Lipschitz condition requires that any two individuals x, y that are at distance $d(x, y) \in [0, 1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x, y)$. In other words, the distributions over outcomes observed by x and y are indistinguishable up to their distance $d(x, y)$.

Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 2012.

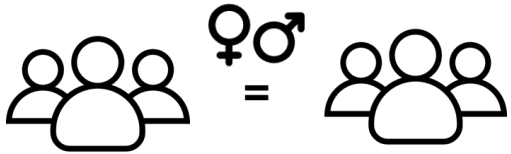
What does fairness in ML mean?

Individual: measures the impact that discrimination has on the individuals



E.g., similar applicants, should have similar probability of receiving positive loan approval

Group: measures the impact that the discmrination has on the groups of individuals



E.g., The probability of receiving positive loan approval should be similar among female and male applicants

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1-7). IEEE.

Statistical Fairness Notions

Demographic Parity

$$P(\hat{Y} = 1 \mid S = 1) = P(\hat{Y} = 1 \mid S = 0)$$

equal probability of receiving a positive loan approval for female and male applicants

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009, December). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 13-18). IEEE.

Statistical Fairness Notions

Demographic Parity

$$P(\hat{Y} = 1 \mid S = 1) = P(\hat{Y} = 1 \mid S = 0)$$

equal probability of receiving a positive loan approval for female and male applicants

Equal opportunity

$$P(\hat{Y} = 1 \mid Y = 1, S = 1) = P(\hat{Y} = 1 \mid Y = 1, S = 0)$$

classifier should give similar results to applicants of both genders with actual positive loan approval.

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Statistical Fairness Notions

Demographic Parity

$$P(\hat{Y} = 1 \mid S = 1) = P(\hat{Y} = 1 \mid S = 0)$$

equal probability of receiving a positive loan approval for female and male applicants

Equal opportunity

$$P(\hat{Y} = 1 \mid Y = 1, S = 1) = P(\hat{Y} = 1 \mid Y = 1, S = 0)$$

classifier should give similar results to applicants of both genders with actual positive loan approval.

Equalized odds

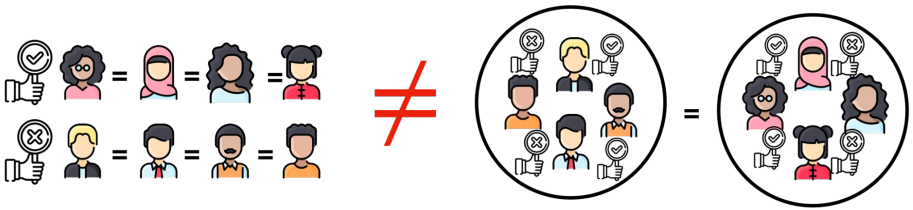
$$P(\hat{Y} = 1 \mid Y, S = 1) = P(\hat{Y} = 1 \mid Y, S = 0)$$

applicants with a rejected loan application and applicants with an accepted loan application should have a similar classification, regardless of their gender.

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Impossibility of Fairness

Impossibility wrt group and individual notions



Impossibility wrt various group fairness notions

- Independence
- Separation
- Sufficiency

You can only achieve one of these measures: demographic parity, equality of odds, and equality of opportunity

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." *arXiv preprint arXiv:1609.07236* (2016).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Causal Notions of Fairness

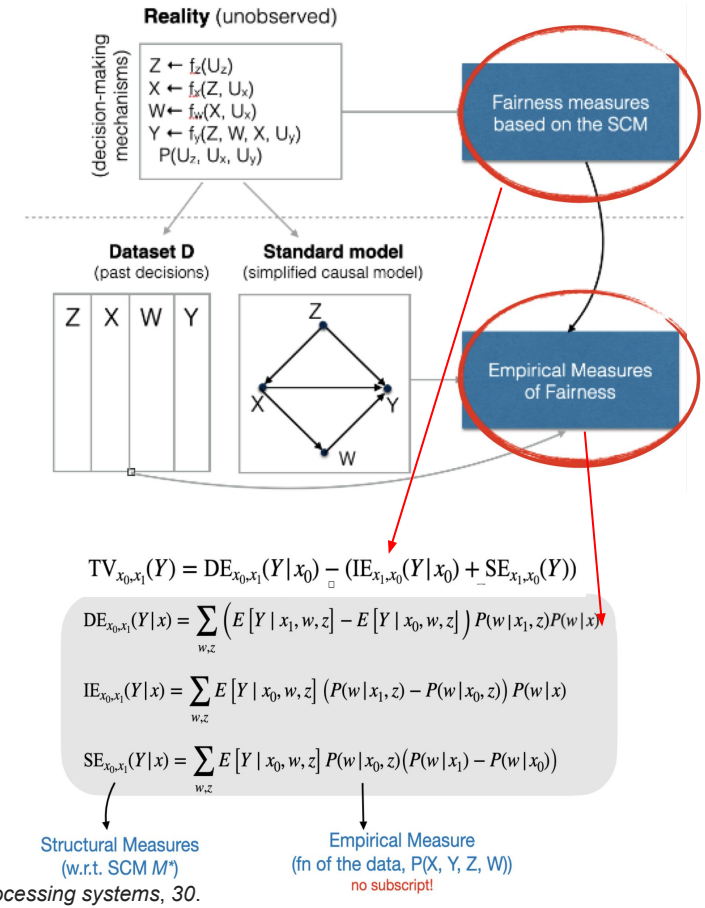
- Causal fairness notions are based on social-legal requirements, e.g.,

US Supreme Court, 2015

A disparate-impact claim relying on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity.

- Based on existence of causal mechanisms, which are almost never observed.

- Construction of causal graph to encode assumptions about underlying SCM is required

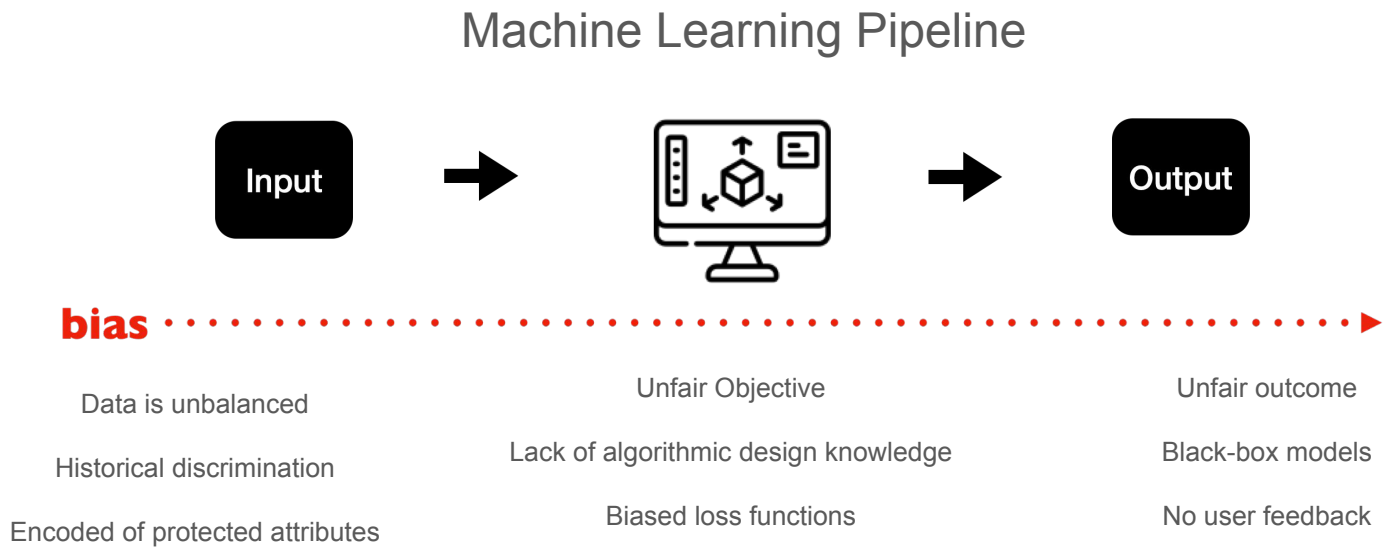


Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

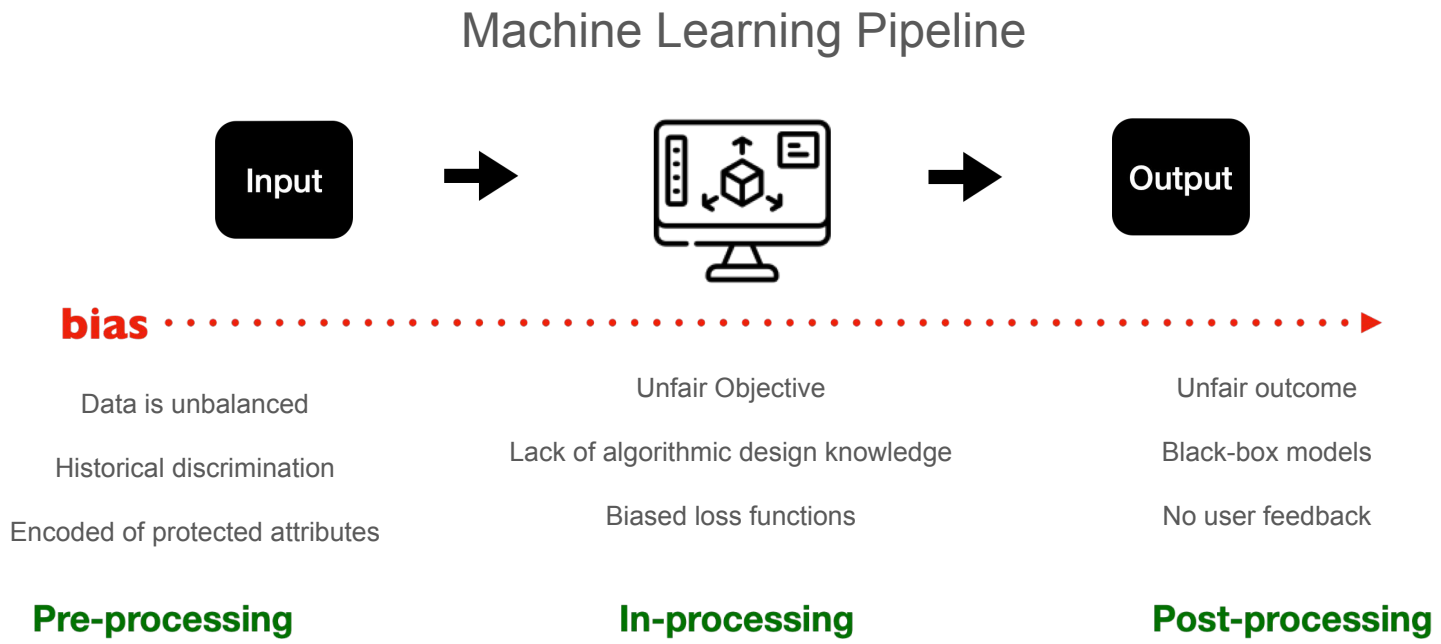
Zhang & Bareinboim. “Fairness in Decision-Making - The Causal Explanation Formula, AAAI, 2018

Nilforoshan, Hamed, et al. "Causal conceptions of fairness and their consequences." *International Conference on Machine Learning*. PMLR, 2022.

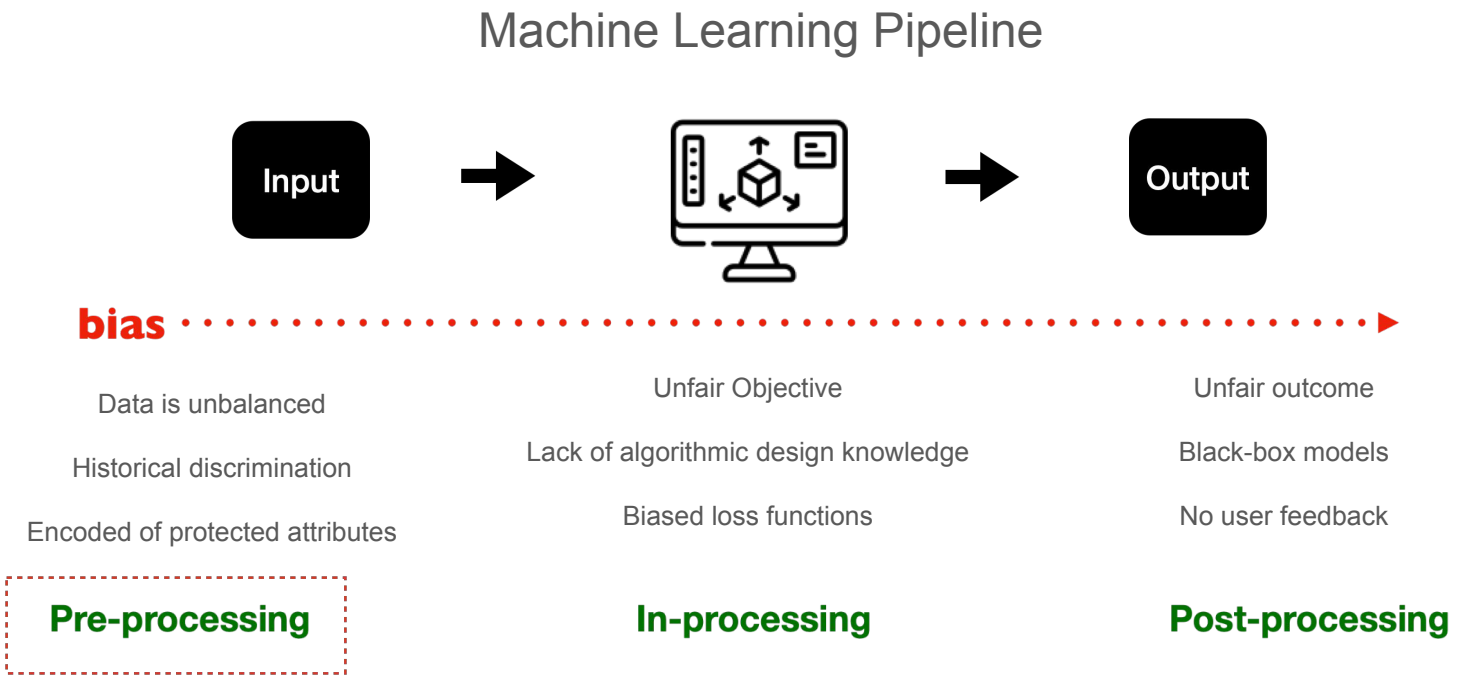
How to mitigate algorithmic discrimination in machine learning?



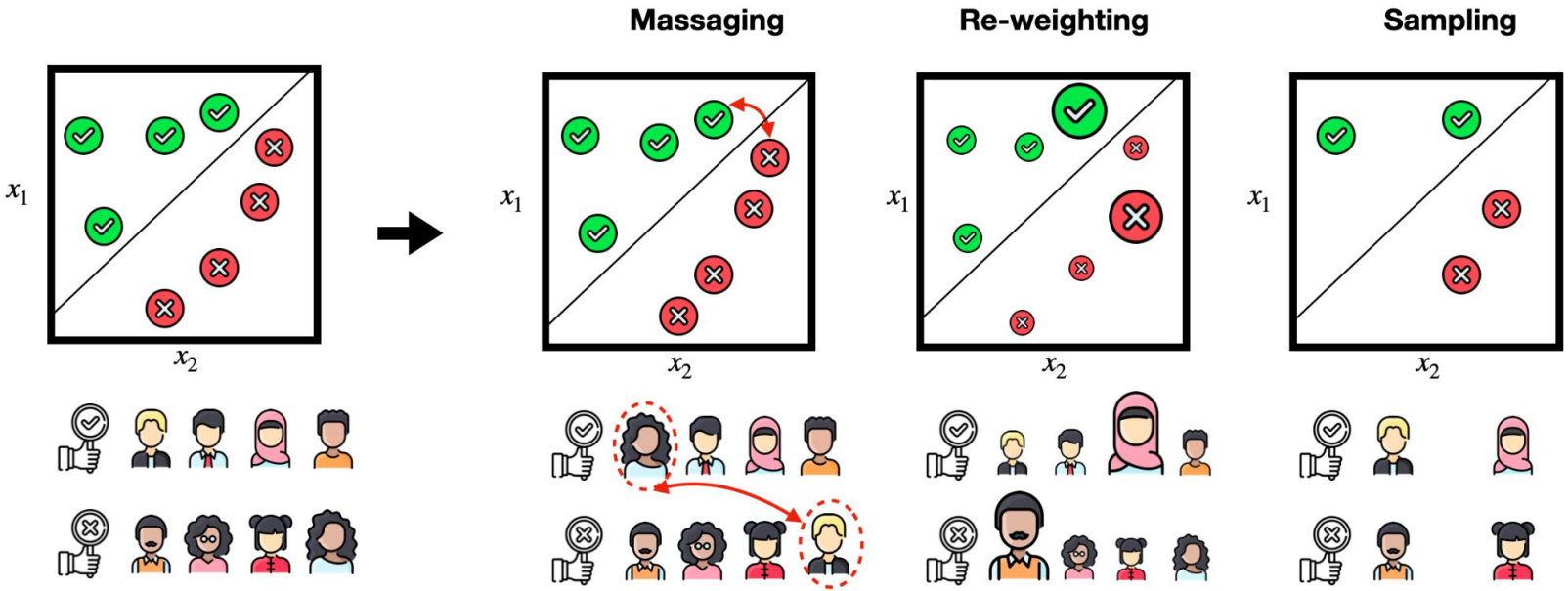
How to mitigate algorithmic discrimination in machine learning?



How to mitigate algorithmic discrimination in machine learning?



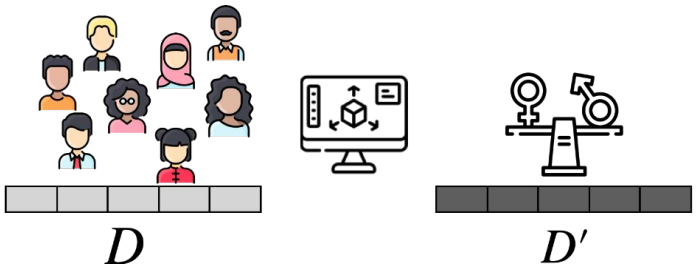
Fairness in Pre-Processing: Data De-Biasing



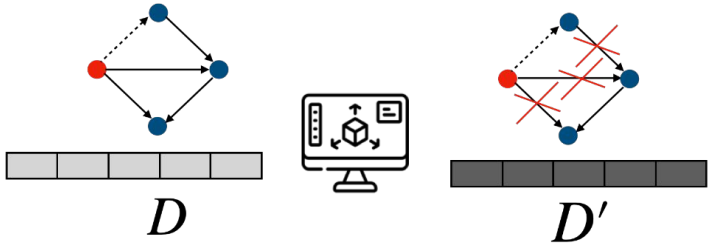
Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, *33*(1), 1-33.

Alabdulmohsin, I., Schrouff, J., & Koyejo, O. (2022). A Reduction to Binary Approach for Debiasing Multiclass Datasets. *arXiv preprint arXiv:2205.15860*.

Fairness in Pre-Processing: Data Generative Models



E.g., Using Generative Adversarial Networks (GANs), Variational Autoencoders, etc.

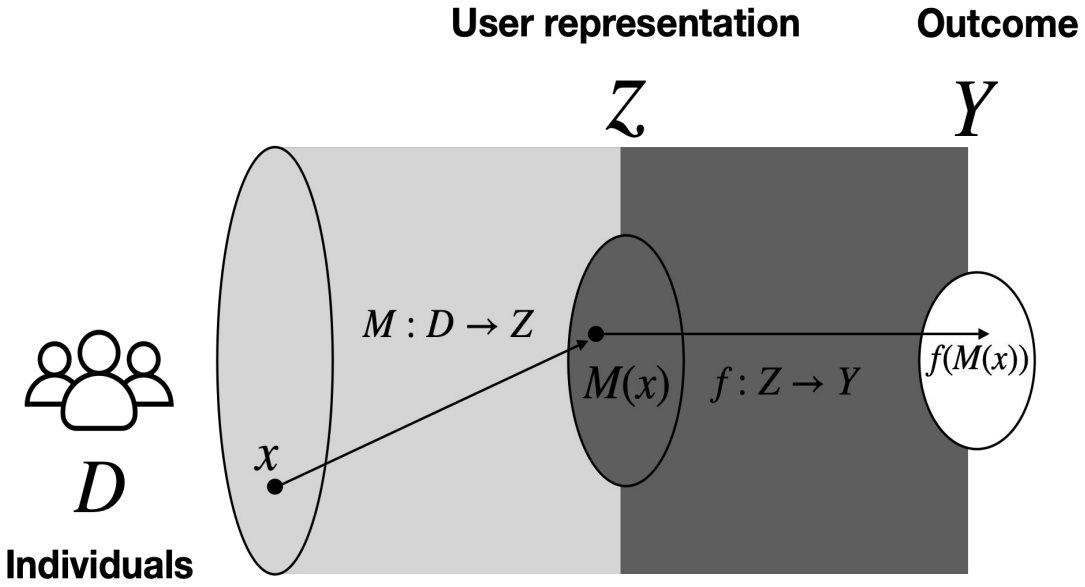


E.g., Using SCMs (by removing paths from sensitive attributes)

Sattigeri, Prasanna, et al. "Fairness GAN: Generating datasets with fairness properties using a generative adversarial network." *IBM Journal of Research and Development* 63.4/5 (2019): 3-1.

van Breugel, Boris, et al. "Decaf: Generating fair synthetic data using causally-aware generative networks." *Advances in Neural Information Processing Systems* 34 (2021): 22221-22233.

Fair Representation Learning



Goal of Representation learning

Preserve Performance:

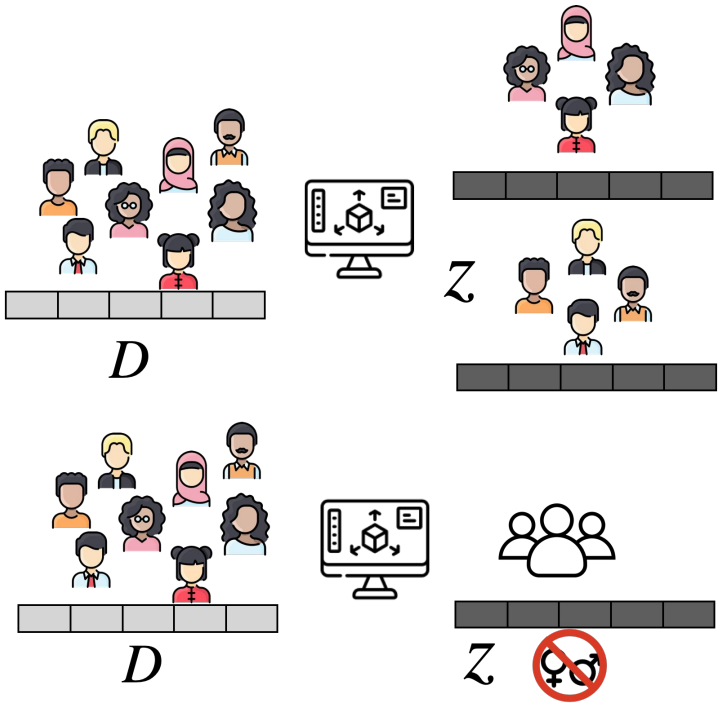
Reconstruction term: the learned representation should resemble the original data

Utility terms: the learned representation should predict target variable

+ Fairness

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In International conference on machine learning (pp. 325-333). PMLR

Fair Representation Learning: Group Fairness



Fairness

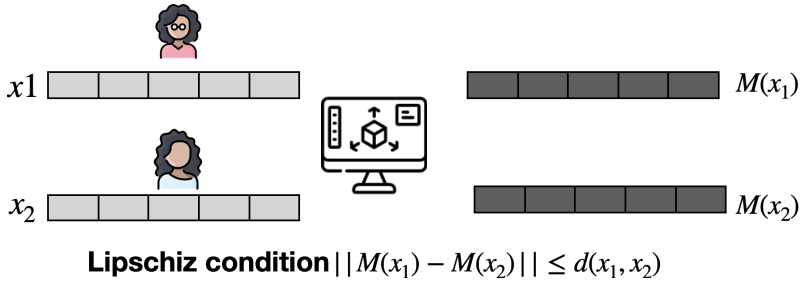
- **Balancing the distribution** among various groups
- **Remove sensitive attributes** (common approach is to use deep learning: VAE, adversarial learning, or disentangled learning)

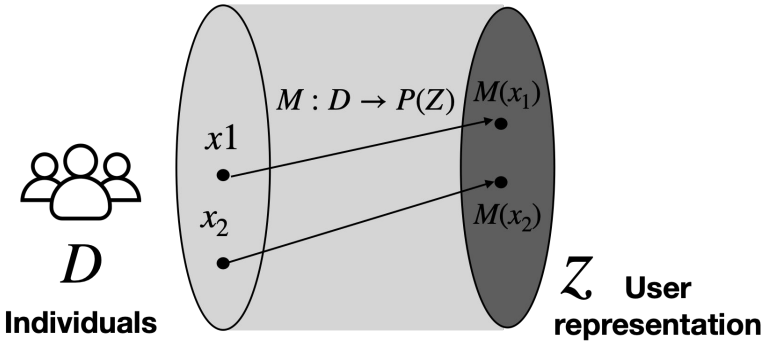
Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. arXiv preprint arXiv:1511.00830.

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018, July). Learning adversarially fair and transferable representations. In International Conference on Machine Learning (pp. 3384-3393). PMLR.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., & Bachem, O. (2019). On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, *32*.

Fair Representation Learning: Individual Fairness





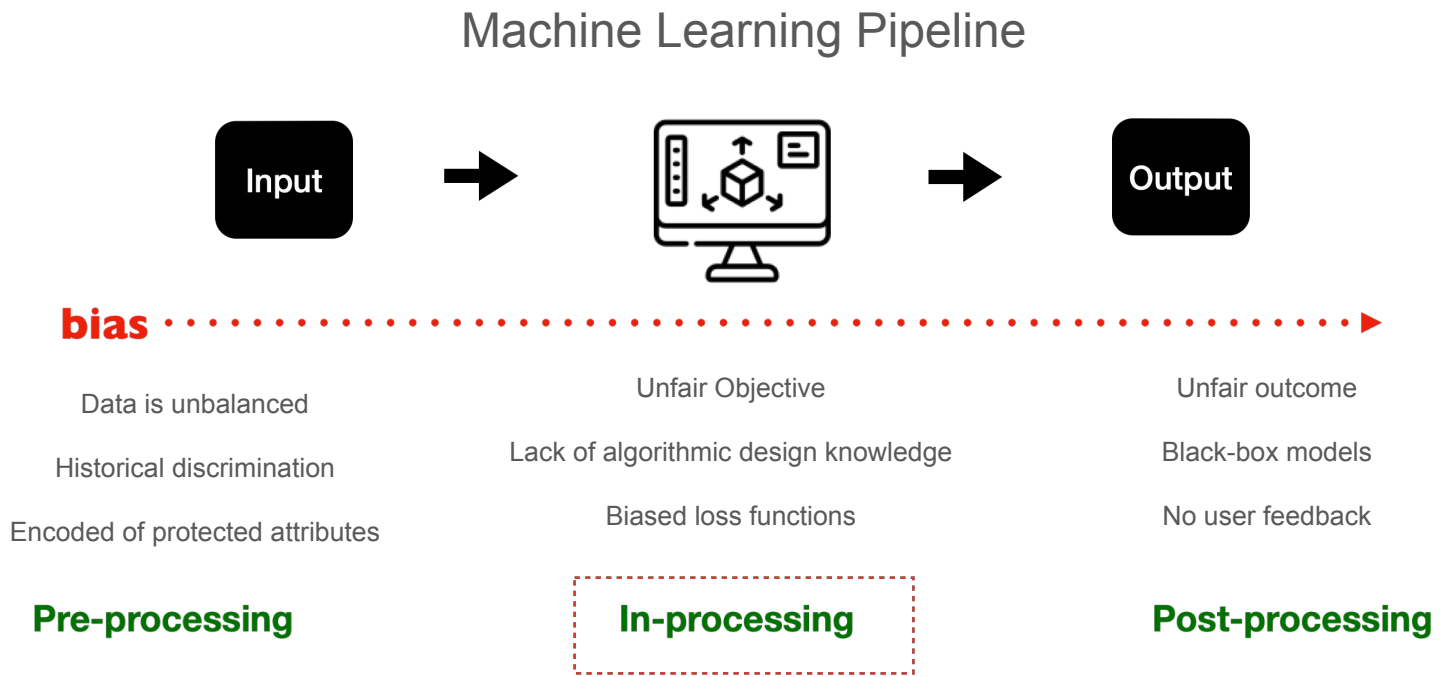
Fairness

- **Similar individuals** should map to **similar distributions.**
- **Task-specific** similarity metric. Ideally captures ground truth or society's best approximation
- Many applications: ranking in recommender systems, financial risk metrics, health metric for treating patients, etc.

Dong, Yushun, et al. "Individual fairness for graph neural networks: A ranking based approach." Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021.

Salganik, Rebecca, Fernando Diaz, and Golnoosh Farnadi. "Analyzing the Effect of Sampling in GNNs on Individual Fairness." *arXiv preprint arXiv:2209.03904* (2022).

How to mitigate algorithmic discrimination in machine learning?



In-processing techniques

- Supervised learning tasks are often expressed as optimization problems

$$\underset{\theta}{\textbf{minimize}} \qquad f(X, Y; \theta)$$

- The optimization problem: finding the parameters that give the best model w.r.t the desired properties

Fairness in another desired property of the learned models

$$g(X, Y; \theta)$$

In-processing techniques

- **Not all optimization problems are the same!**
- Some problems are **computational easy**
- Some problems are **hard**, but **behave well** (approximation methods work well)
- Some problems are **hard**, but have **structure**. And we can exploit this structure.

Adding fairness can change these properties!

In-processing techniques

Fairness as Constrained Optimization

$$\underset{\theta}{\textbf{minimize}} \; f(X, Y; \theta) \quad \textbf{subject to} \; g(X, Y; \theta)$$

Fairness as Regularizer

$$\underset{\theta}{\textbf{minimize}} \; f(X, Y; \theta) + \lambda g(X, Y; \theta)$$

Fairness as Multi-objective Optimization

$$\underset{\theta}{\textbf{minimize}} \; f(X, Y; \theta) \times g(X, Y; \theta)$$

Choi, Y., Farnadi, G., Babaki, B., & Van den Broeck, G. (2020, April). Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 06, pp. 10077-10084).

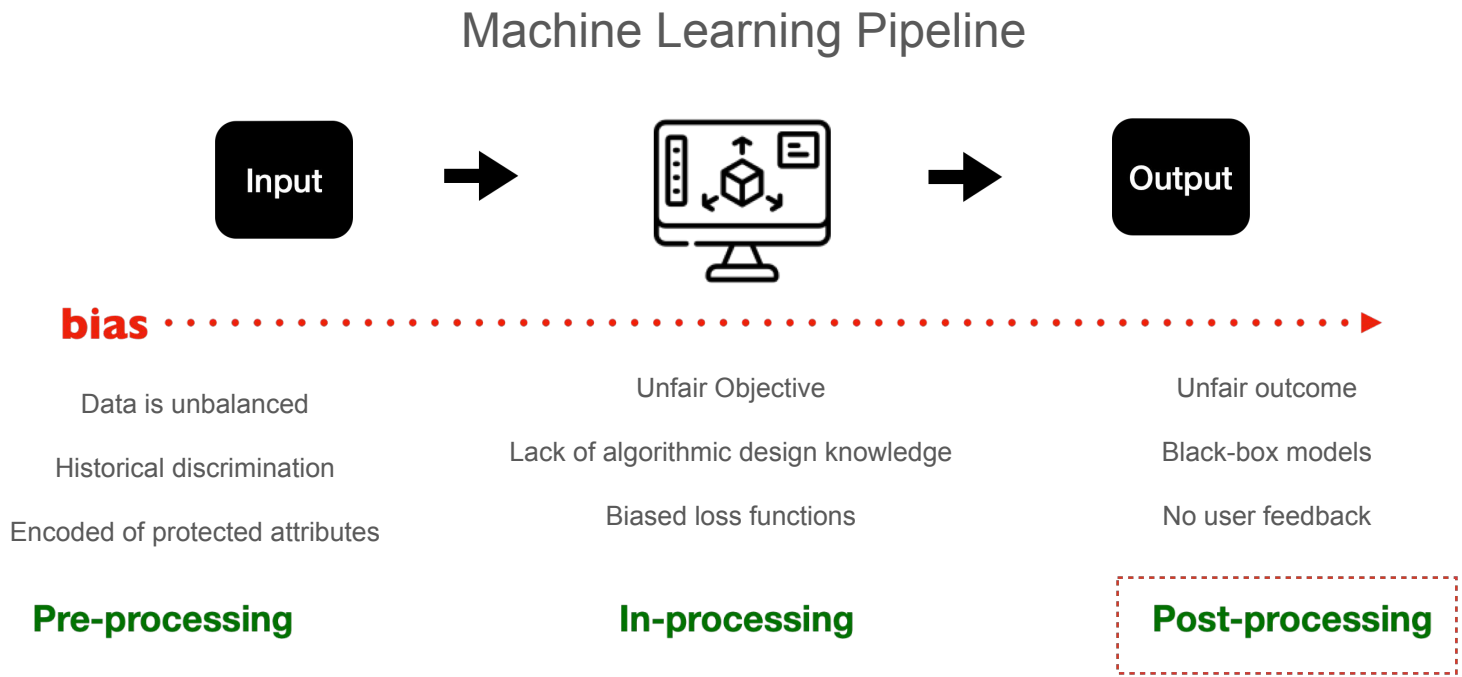
Mohammadi, K., Sivaraman, A., & Farnadi, G. (2022). FETA: Fairness Enforced Verifying, Training, and Predicting Algorithms for Neural Networks. *arXiv preprint arXiv:2206.00553*.

Kamishima, Toshihiro, Shotaro Akaho, and Jun Sakuma. "Fairness-aware learning through regularization approach." 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 2011.

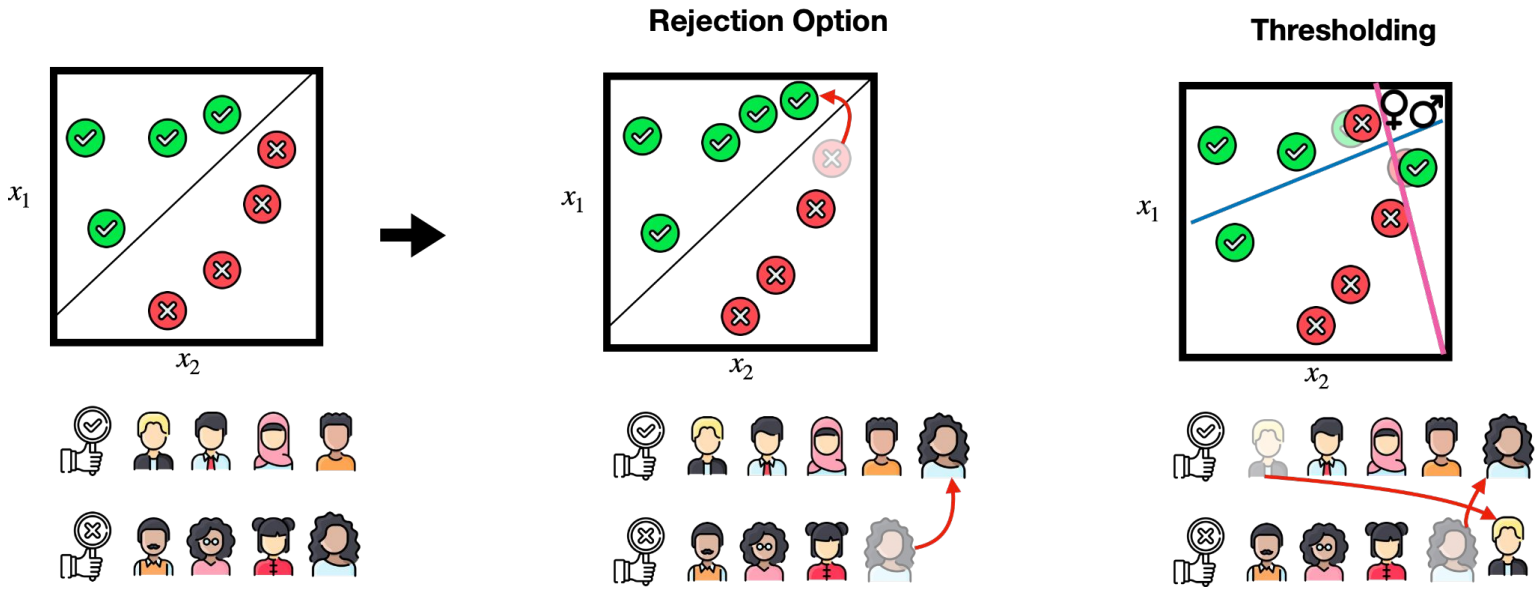
A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, “A Reductions Approach to Fair Classification,” arXiv.org, 16-Jul-2018. [Online]. Available: https://arxiv.org/abs/1803.02453.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning* (pp. 2564-2572). PMLR.

How to mitigate algorithmic discrimination in machine learning?



Fairness in Post-Processing



Kamiran, F., Karim, A., & Zhang, X. (2012, December). Decision theory for discrimination-aware classification. In 2012 IEEE 12th International Conference on Data Mining (pp. 924-929). IEEE.

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

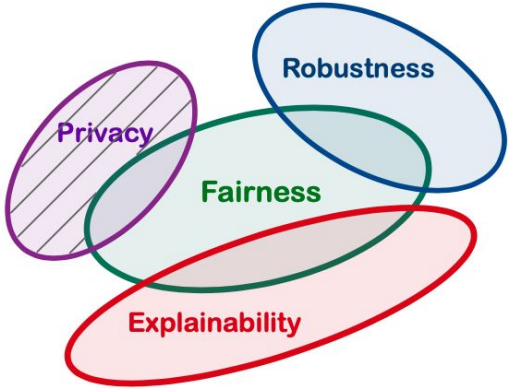
Trade-offs

	Ease of implementation and (re)-use	Scalability	Ease of auditing	Fairness/ Performance tradeoff	Generalization
Pre-processing, e.g., representation learning	✓	✓	✓		✓
In-processing, e.g., fairness regularizer			✓	✓	✓
Post-processing, e.g., thresholding		✓	✓		

Inspired by Sanmi Koyejo's talk on fair representation learning tutorial at NeurIPS 2019

Summary

- No free lunch: Fairness is a **socio-technical challenge**
- Many aspects of fairness are **NOT** captured by the statistical measures
- One notion **cannot** simultaneously satisfy all metrics
- Algorithmic fairness is **highly dependent** on the fairness notion, and the result change by changing the notion of fairness
- We may need to make a **trade-off** in different contexts



Introduction to Private Learning

Why Privacy matters?

Internet, social media, emails

drones

IOT

Personal Data: Increasingly more and more devices collect and stream data

Warning: a few corporations own the data and they might abuse them

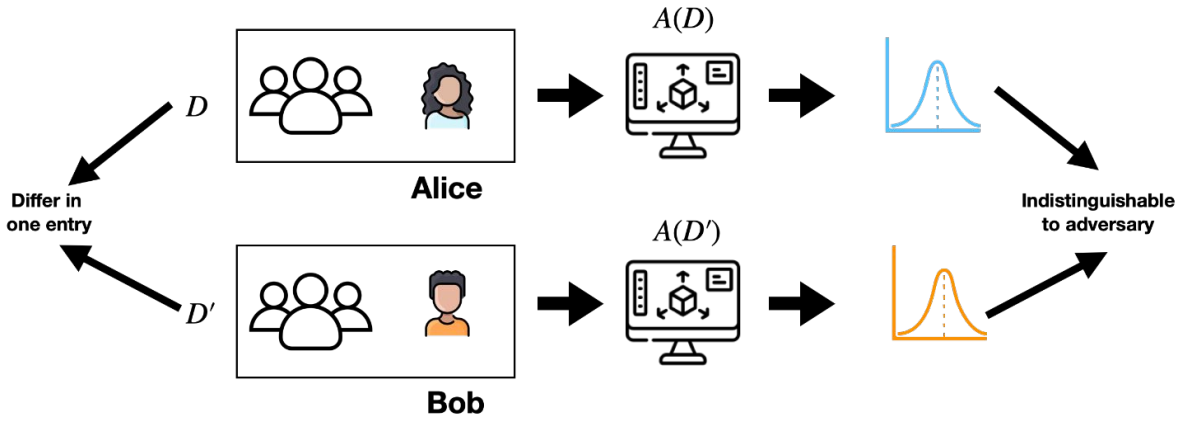
GDPR — Data Protection Impact Assessment

https://formiti.com/global-data-privacy-compliance-staying-ahead-of-the-curve/

https://www.i-sight.com/resources/a-practical-guide-to-data-privacy-laws-by-country/

Privacy Regulations

Differential Privacy



$$\sup_{y \in Y} \log \frac{P(A(D) = y)}{p(A(D') = y)} \leq \epsilon$$

$$\text{noise} \propto \frac{\text{sensitivity}}{\text{privacy loss}}$$

$$P(A(D) = y) \leq e^{\epsilon} p(A(D') = y)$$

$(\epsilon, 0)$ – **Differential Privacy**

Laplace mechanism

$$P(A(D) = y) \leq e^{\epsilon} p(A(D') = y) + \delta$$

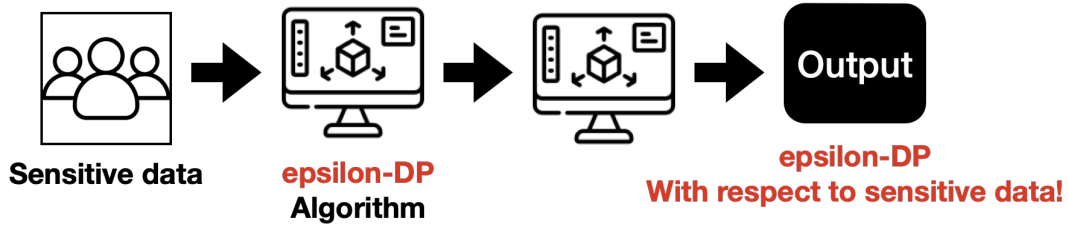
(ϵ, δ) – **Differential Privacy**

Gaussian mechanism

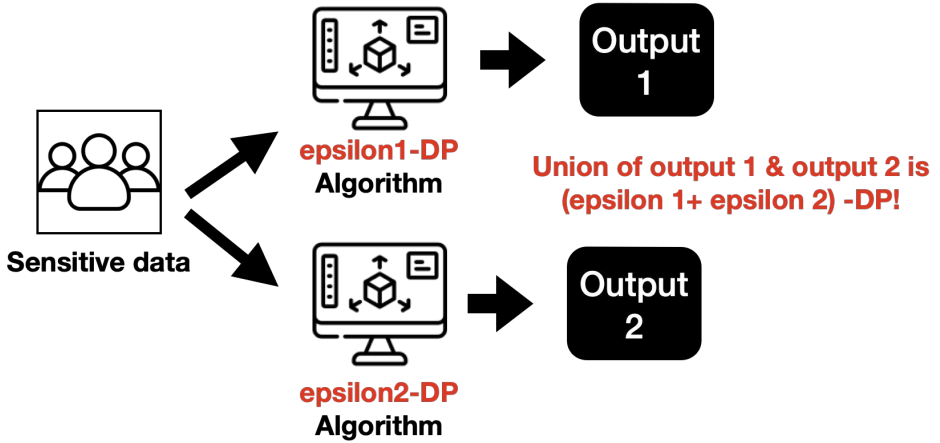
Dwork, Cynthia. "Differential privacy: A survey of results." *International conference on theory and applications of models of computation*. Springer, Berlin, Heidelberg, 2008.

Properties of DP

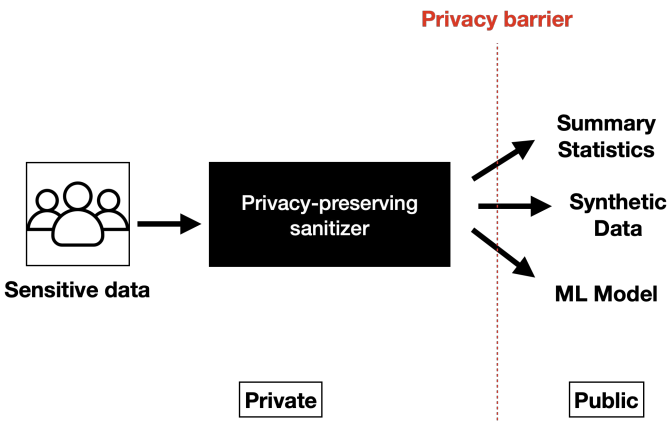
Post-processing invariance



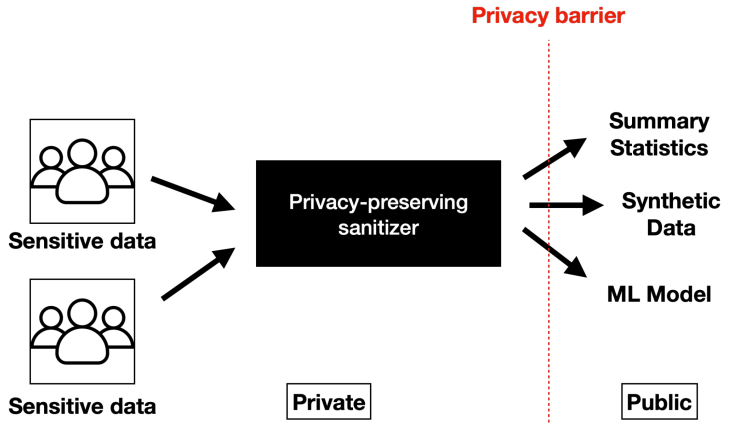
Composability



Privacy-preserving machine learning



Privacy settings in ML (single data source)



Privacy settings in ML (multiple data sources)

Empirical Risk Minimization

Empirical Risk Minimization (ERM) is a common paradigm for prediction problems

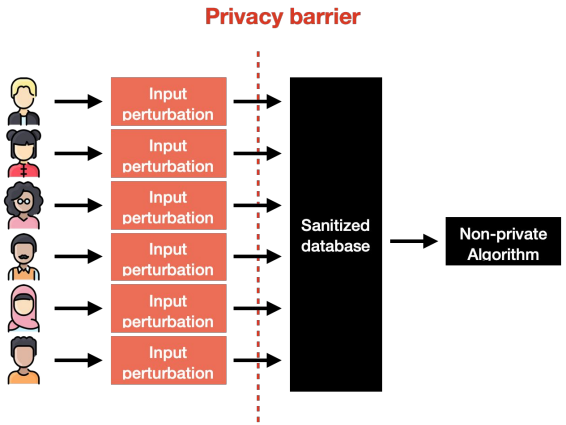
$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i)) + \lambda R(\mathbf{w})$$

Empirical Risk Minimization

Empirical Risk Minimization (ERM) is a common paradigm for prediction problems

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i)) + \lambda R(\mathbf{w})$$

Input Perturbation

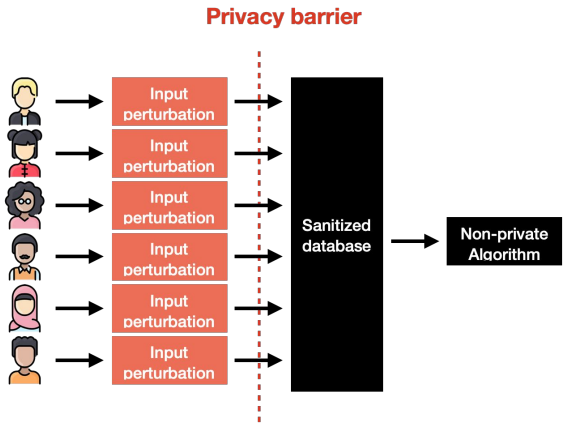


Empirical Risk Minimization

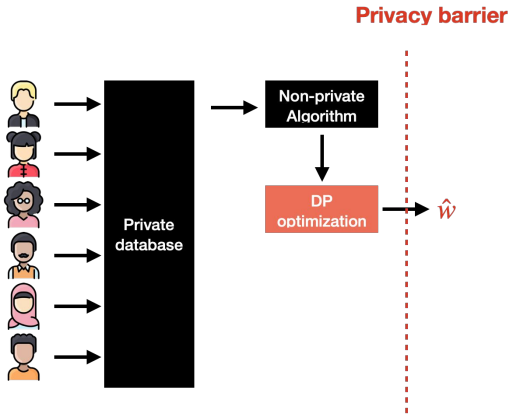
Empirical Risk Minimization (ERM) is a common paradigm for prediction problems

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i)) + \lambda R(\mathbf{w})$$

Input Perturbation



Objective Perturbation

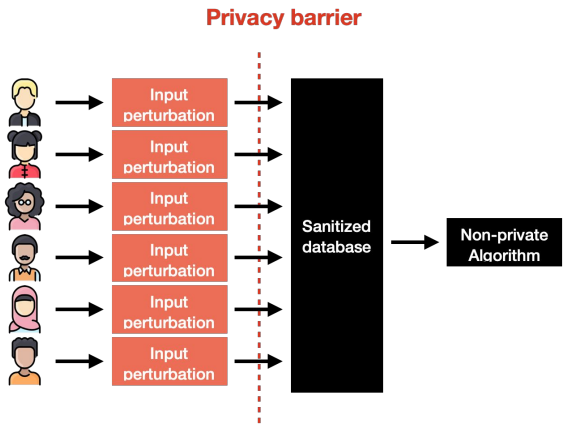


Empirical Risk Minimization

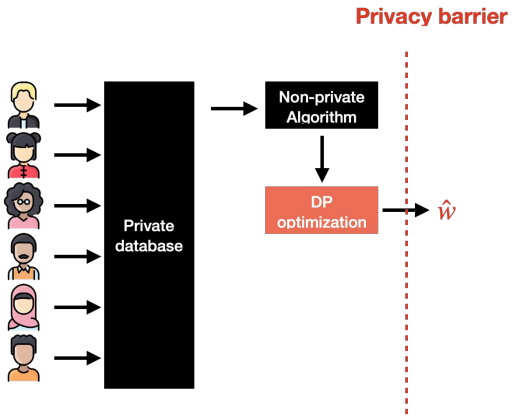
Empirical Risk Minimization (ERM) is a common paradigm for prediction problems

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, (\mathbf{x}_i, y_i)) + \lambda R(\mathbf{w})$$

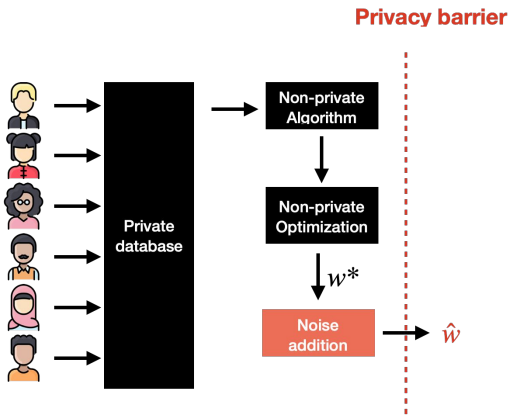
Input Perturbation



Objective Perturbation



Output Perturbation

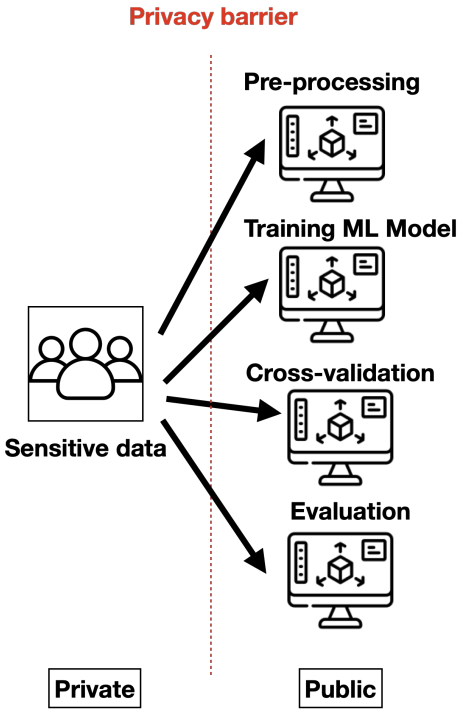


What Composition says about Multistage ML Methods?

- How to allocate privacy risk across different stages of the machine learning pipeline?
- Basic composition: privacy is **additive**

- Having k algorithms with $(\epsilon_i, \delta_i) : i = 1,2,\ldots,k$
- Total privacy loss:

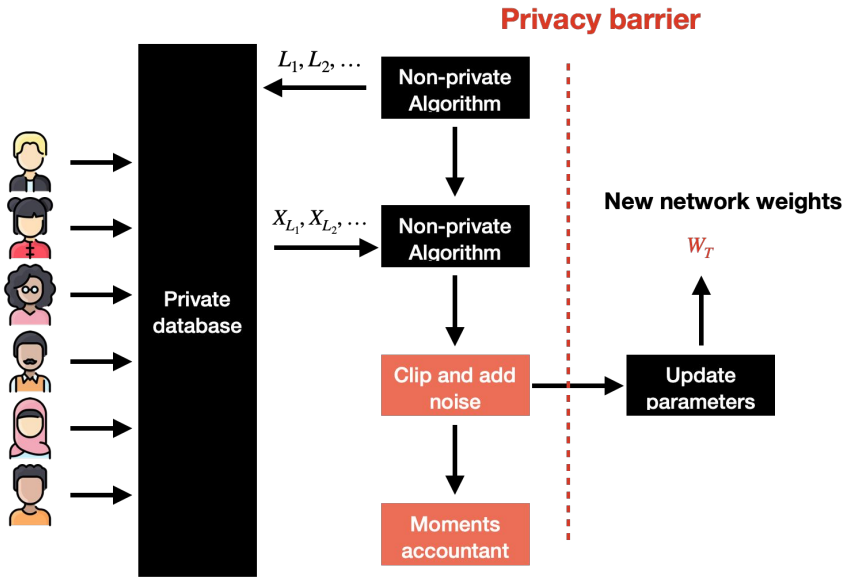
$$\left(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i\right)$$



Bassily, Raef, Adam Smith, and Abhradeep Thakurta. "Private empirical risk minimization: Efficient algorithms and tight error bounds." *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE, 2014.

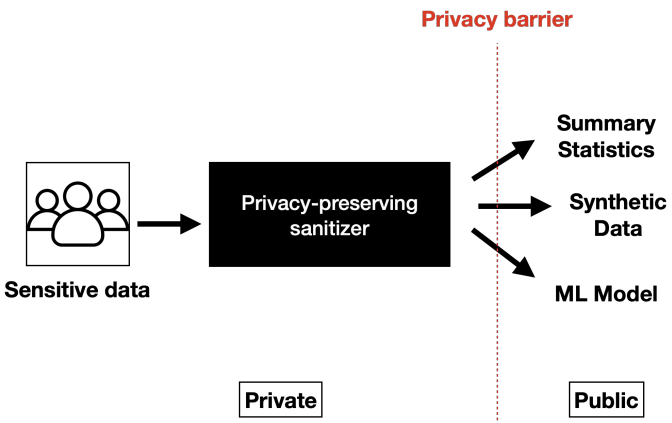
DP in Deep Learning

- Stochastic Gradient Descent (SGD) is a popular method for optimization
- Main idea of DP-SGD: use moments accountant to track privacy loss
- Additional components: **Gradient clipping**, **Noise addition**, **data augmentation**, **mini-batching**, etc.

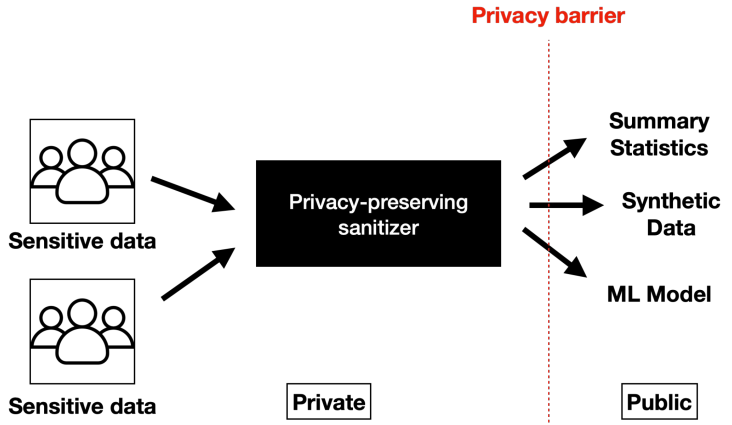


Abadi, Martin, et al. "Deep learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016.

Privacy-preserving machine learning



Privacy settings in ML
(single data source)



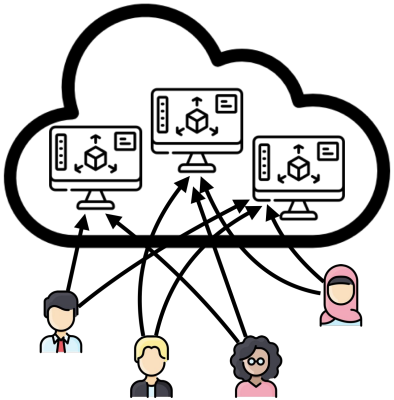
Privacy settings in ML
(multiple data sources)

Secure Multi Party Computation (MPC)

Multiple parties jointly compute

- A **function** or **output**
- While their **input remains private**
- In a **distributed** fashion

No central entity- distributed setup to preserved privacy



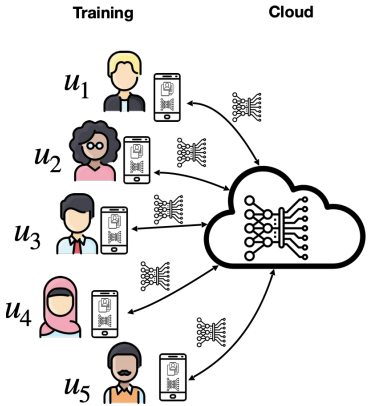
Goldreich, O. (1998). Secure multi-party computation. *Manuscript. Preliminary version*, *78*, 110.

Federated Learning (FL) or Split learning

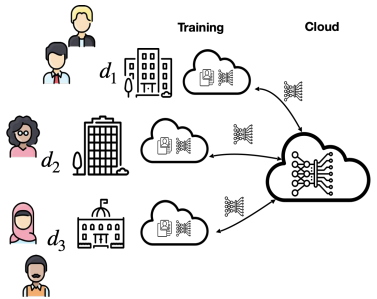
- FL is a machine learning setting where **multiple entities** (clients) collaborate in solving a machine learning problem, under the coordination of a **central server** or service provider.
- Each client's **raw data is stored locally** and not exchanged or transferred; instead, focused updates intended for immediate **aggregation** are used to achieve the learning objective.

FL/Split Learning are not private!

Cross-device FL



Cross-silo FL

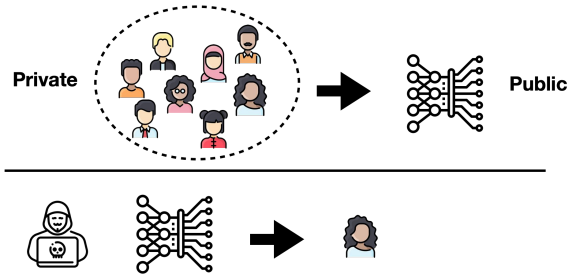


McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

Gupta, Otkrist, and Ramesh Raskar. "Distributed learning of deep neural network over multiple agents." *Journal of Network and Computer Applications* 116 (2018): 1-8.

Memorization and Overlearning issues of Deep Learning Models

- Unintended memorization in DL is:
  - Persistent
  - Hard-to-avoid issue that can have serious consequences
- Overlearning is not a result of overfitting
- Memorization does not only happen in large models
- Memorization can happen in various context, text, vision, etc.

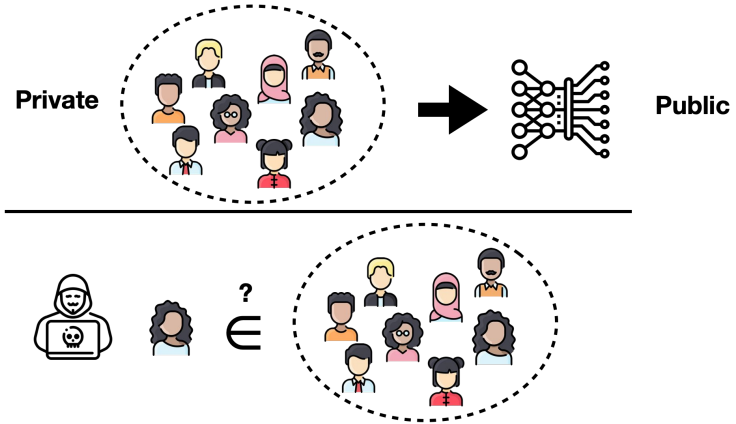


Reconstruction attack

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)* (pp. 267-284).

Membership attack

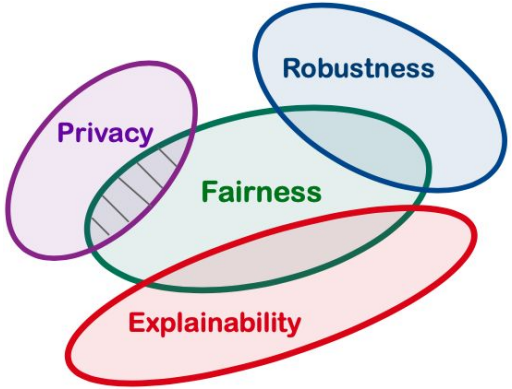
The basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset.



Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017.

Summary

- **Training** does not on its own **guarantee privacy**
- Deep learning models **memorize** sensitive information
- There are various privacy enhancing technologies (PETs)
- Output privacy **does not guarantee** input privacy
- DP in ML pipeline is challenging due to its **iterative** nature
- Good **DP** algorithms should **generalize** since they learn about populations, not individuals.
- In Federated learning/split learning, raw data never leave clients devices but it is not necessarily make these algorithm private
- Privacy budget is relative to the task



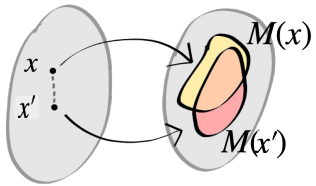
At the Intersections: Fairness & Privacy

Fairness and Privacy: aligned goals

- DP aims at rendering the participation of individuals indistinguishable to an observer who accesses the outputs of a computation
- Fairness attempts at equalizing properties of outputs across different individuals.

Privacy and fairness can be viewed as aligned objectives, e.g., Dwork et al, 2021 shows **individual** fairness is a generalization of DP.

$$d_x(x, x') = \epsilon \, |x \Delta x'|$$

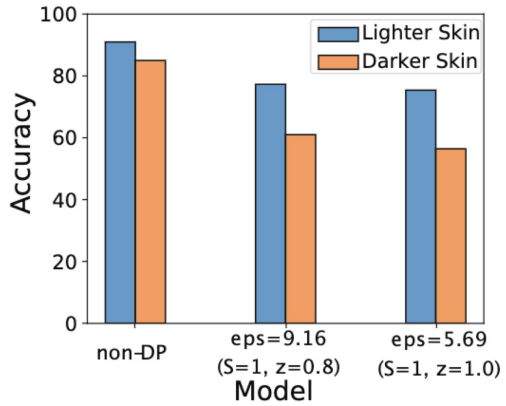


$$d_y(M(x), M(x')) = \sup_{y \in Y} \log \frac{P(A(D) = y)}{p(A(D') = y)}$$

Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 2012.

Fairness and Privacy: contrastive goals

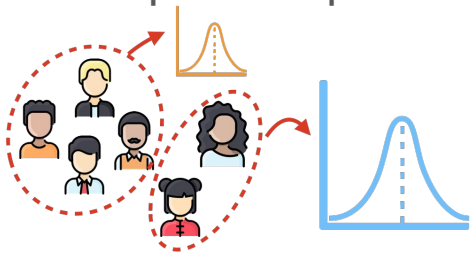
Privacy and fairness can be viewed as contrastive objectives, e.g., it has been observed that the outputs of DP classifiers may create or exacerbate disparate impacts among **groups** of individuals



Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov. "Differential privacy has disparate impact on model accuracy." *Advances in neural information processing systems* 32 (2019).

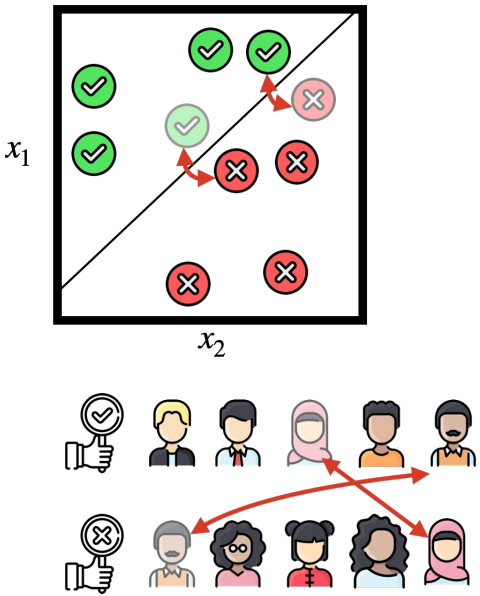
DP and Fairness

- The accuracy of the minority group was disproportionately impacted by the private training.
- These observations were validated on several **vision** and **natural language processing** tasks and in both a **centralized** and **federated** setting.
- The **size of a protected group** would play a crucial role to the exacerbation of the disparate impacts in private training



Abadi, Martin, et al. "Deep learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016.

DP in EMR and Fairness

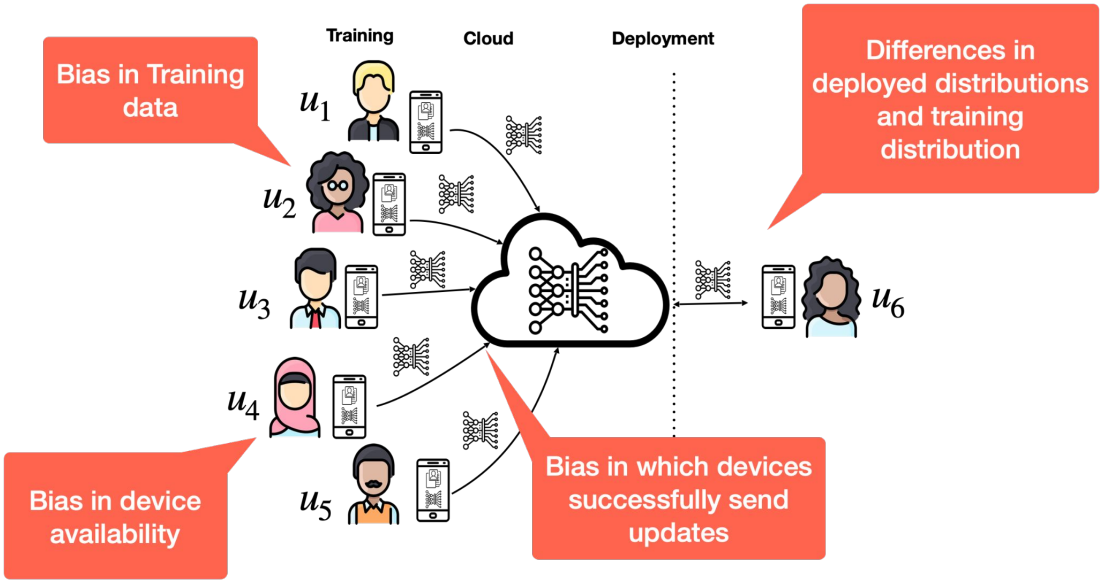


- **Output perturbation: Input norms** and **distance to decision boundary** are two key characteristics of the **data** connected with exacerbating the disparate impacts of private learning tasks.
- **DP-SGD:** The two key characteristics of DP-SGD are **clipping the gradients** whose L2 norm exceeds a given bound C and **perturbing** the averaged clipped gradients with **Gaussian noise**.

Tran, Cuong, My Dinh, and Ferdinando Fioretto. "Differentially private empirical risk minimization under the fairness lens." *Advances in Neural Information Processing Systems* 34 (2021): 27555-27565.

Federated Learning and Fairness

- Fairness: Resource allocation, Quality of service, Client selection (infrastructure), Incentive, etc.
- Personalization with Multi-task learning, fine-tuning, and Meta-learning.



Pentyala, S., Neophytou, N., Nascimento, A., De Cock, M., & Farnadi, G. (2022). PrivFairFL: Privacy-Preserving Group Fairness in Federated Learning. *arXiv preprint arXiv:2205.11584*.

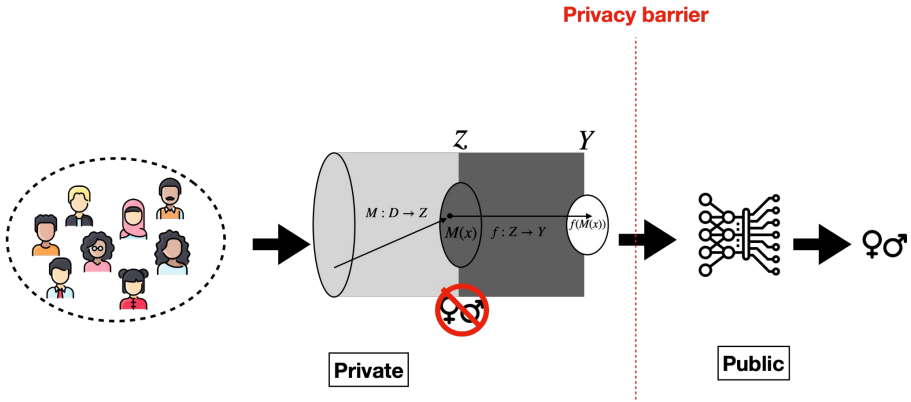
Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020). Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*.

Li, T., Sanjabi, M., Beirami, A., & Smith, V. (2019). Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*.

Li, T., Hu, S., Beirami, A., & Smith, V. (2021, July). Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning* (pp. 6357-6368). PMLR.

Fairness and Overlearning issues of DL

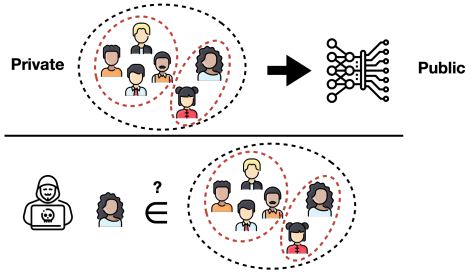
- "Overlearning" means that a model trained for a seemingly simple objective implicitly learns to recognize attributes and concepts that are
  - (1) not part of the learning objective
  - (2) sensitive from a privacy or bias perspective.



Song, Congzheng, and Vitaly Shmatikov. "Overlearning reveals sensitive attributes." *arXiv preprint arXiv:1905.11742* (2019).

Membership Attacks and Fairness

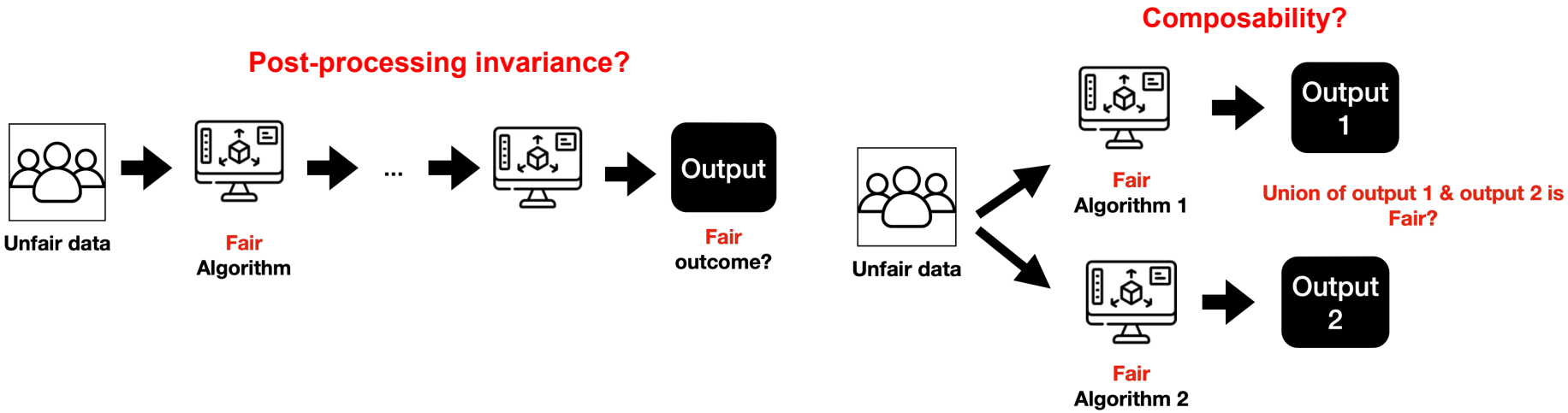
- Fairness comes at the cost of privacy, and this cost is not distributed equally
- The information leakage of fair models increases significantly on the unprivileged subgroups, which are the ones for whom we need fair learning.
- The more biased the training data is, the higher the privacy cost of achieving fairness for the unprivileged subgroups will be.



Chang, H., & Shokri, R. (2021, September). On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 292-303). IEEE.

Kulynych, Bogdan, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. "Disparate vulnerability to membership inference attacks." *arXiv preprint arXiv:1906.00389* (2019).

Fairness & Privacy Properties

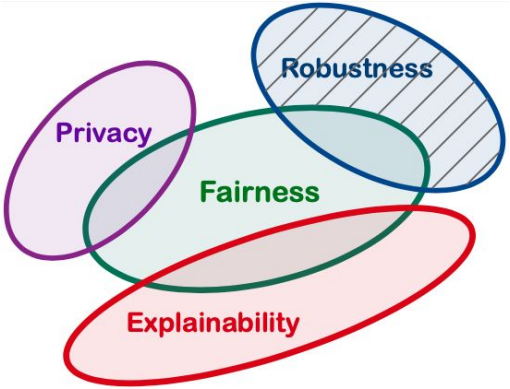


Dwork, C., & Ilvento, C. (2018). Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*.

Dwork, Cynthia, and Christina Ilvento. "Individual fairness under composition." *Proceedings of Fairness, Accountability, Transparency in Machine Learning* (2018).

Open challenges

- Explore various group fairness techniques and their relations to DP and other PPTs
- Explore data pre-processing techniques that combine privacy and fairness
- Combine various PPTs, e.g., MPC+FL+DP, may help performance, privacy, fairness tradeoffs
- Explore new application domains at the intersection of privacy and fairness
- Analyze fairness and privacy in sequential decision making models
- Analyze fairness and privacy under composition
- …



Introduction to Robust Learning

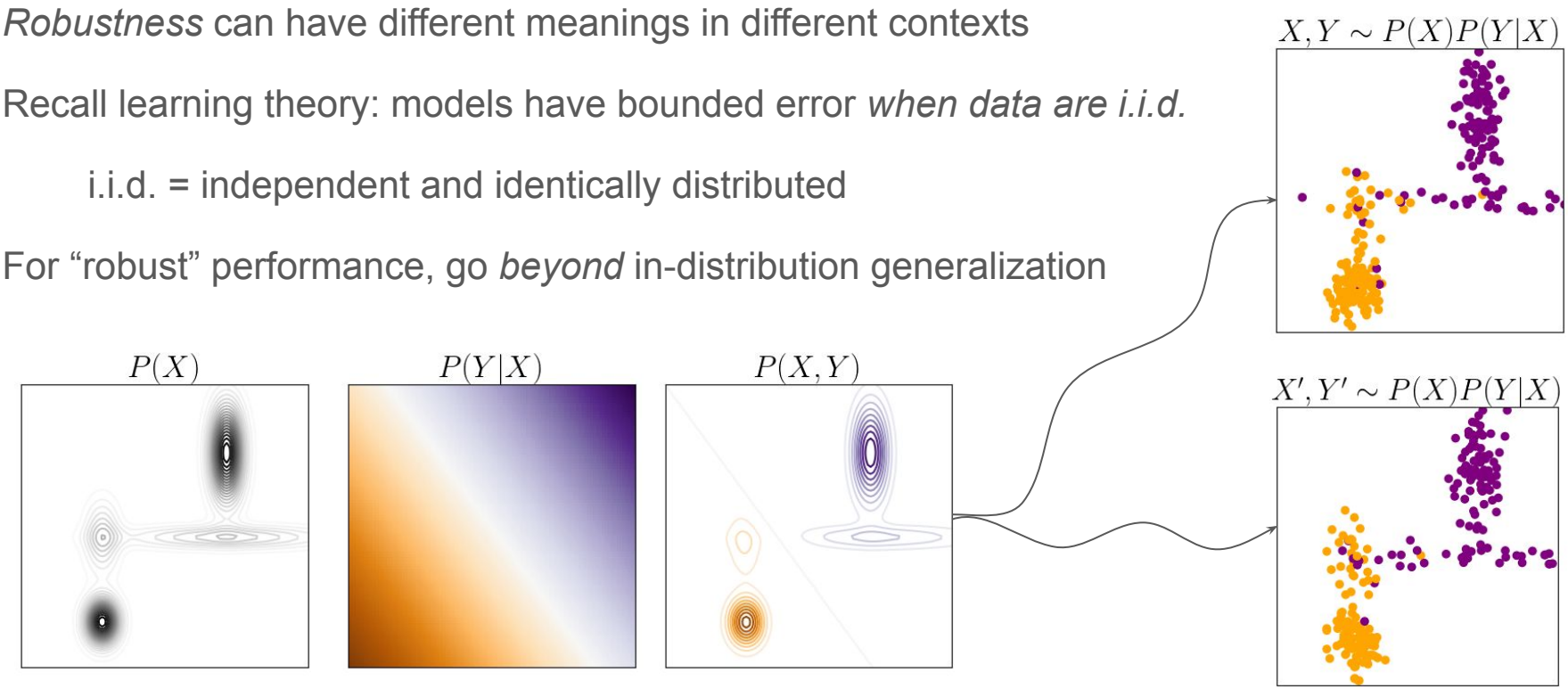
What does it mean to be “robust”?

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

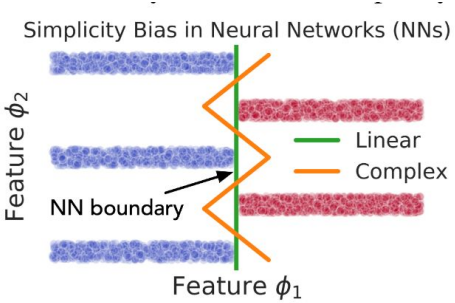
~~Overfitting/underfitting~~ (handled by standard learning theory)

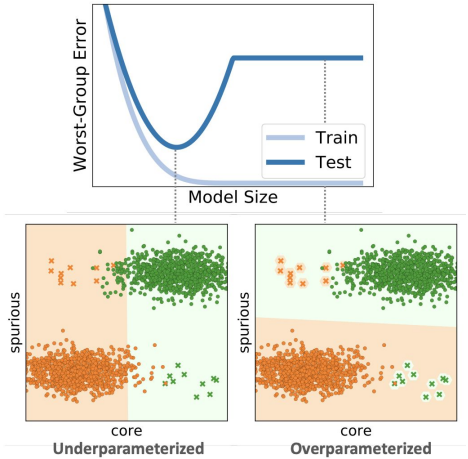
Adversarial examples & security threats

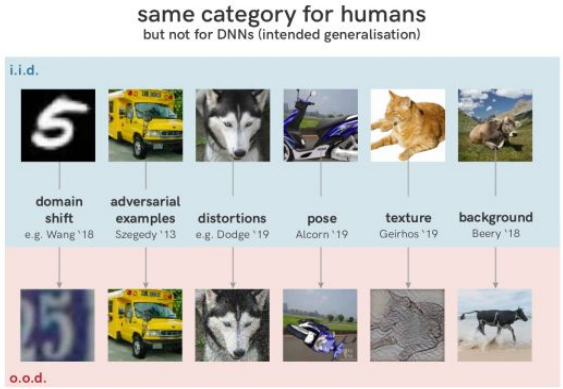
Shortcut learning

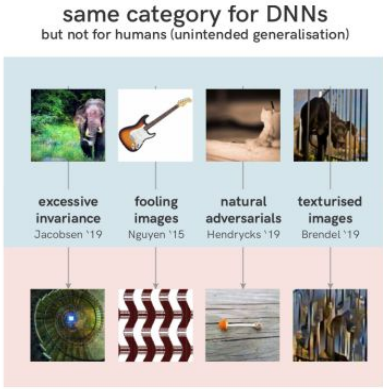
Simplicity bias

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.*

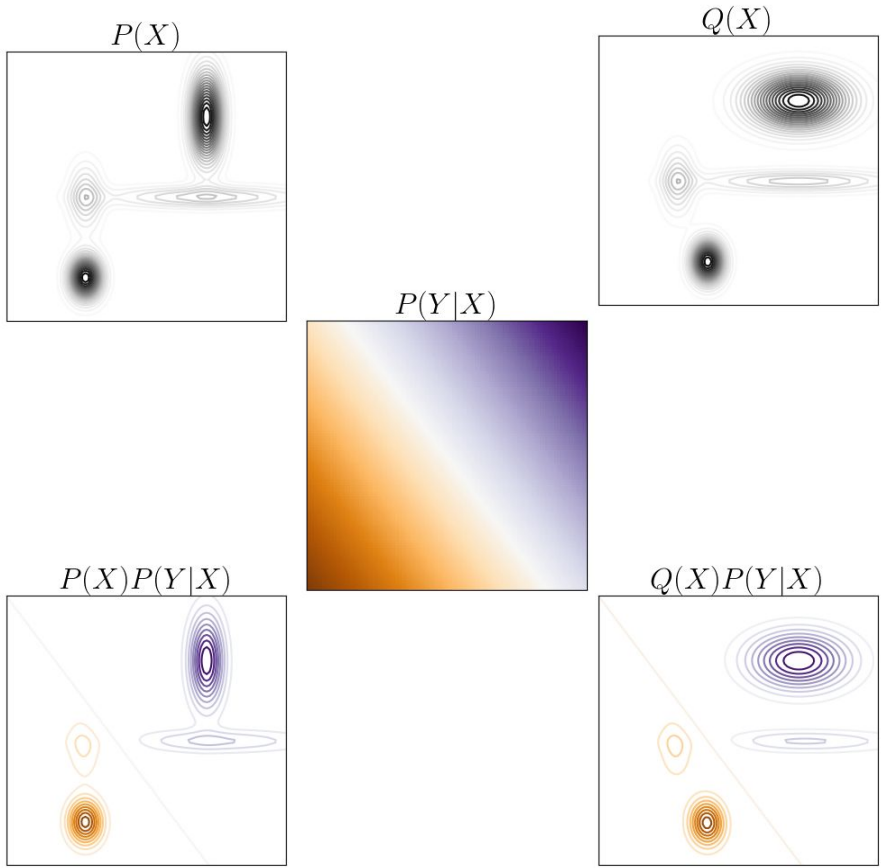
Incorporating “robustness” into learning algorithms

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

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

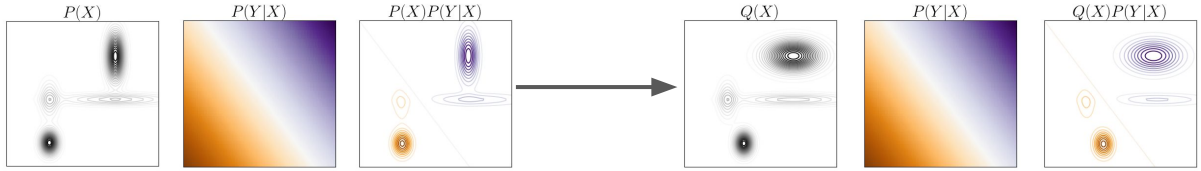
Out-of-distribution (OOD) generalization

What family of distributions should my model handle?

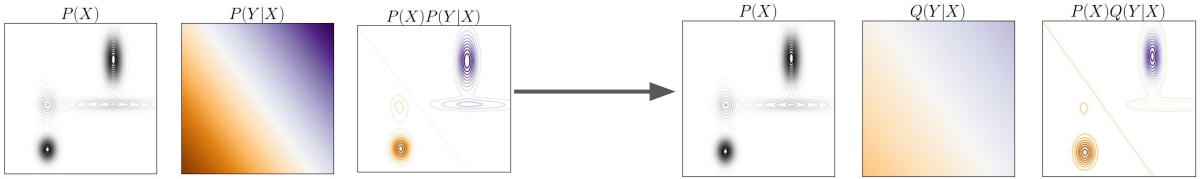


Characterizing distribution shift

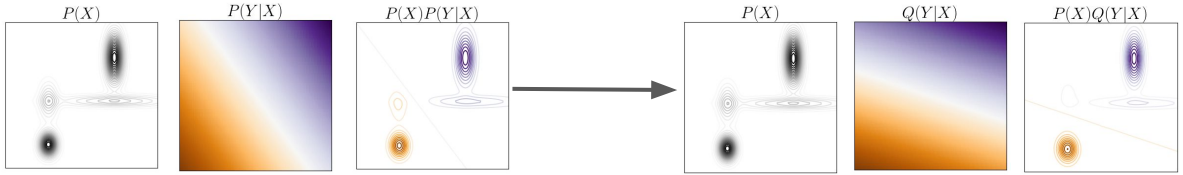
Covariate shift



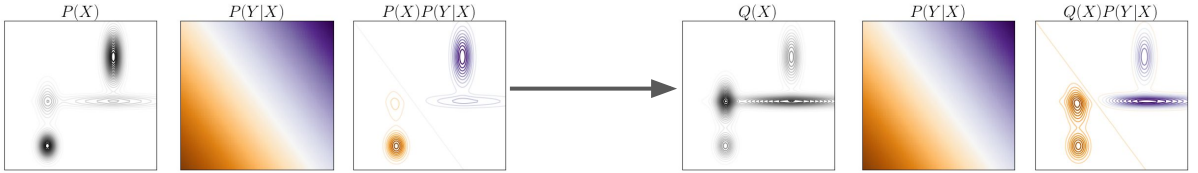
Label noise



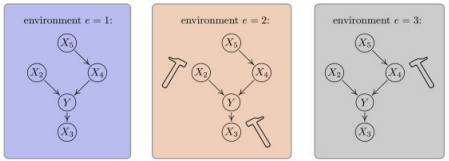
Concept shift

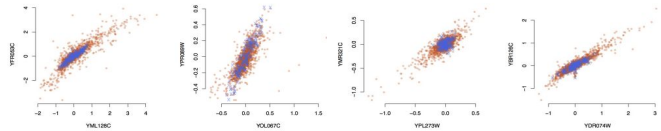


Subpopulation shift



Intervention (on causal graph)





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

Adversarial Robustness

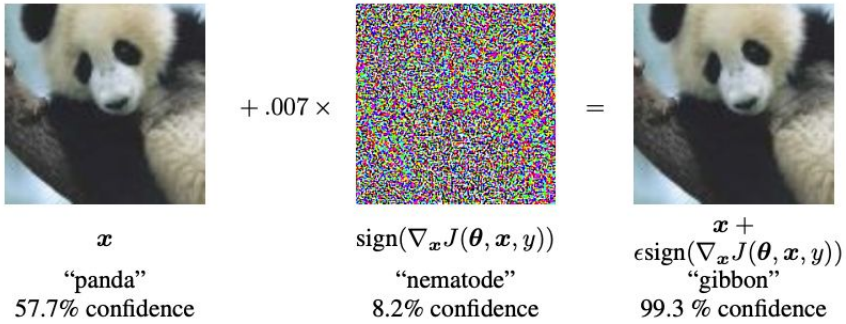
Adversarial examples - small worst-case perturbations in feature space

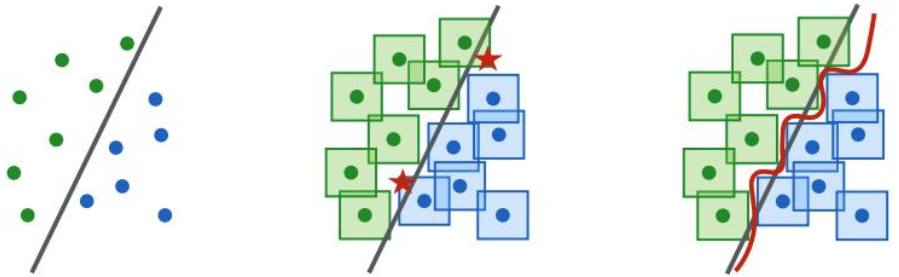
Attacks - white box, black box, …

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*

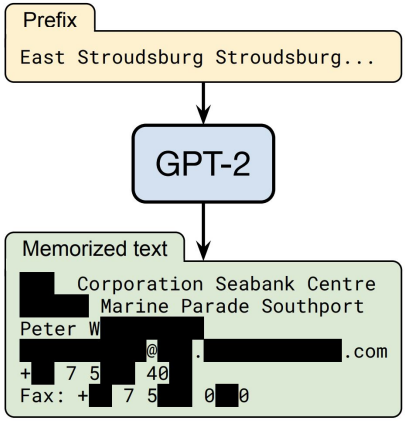
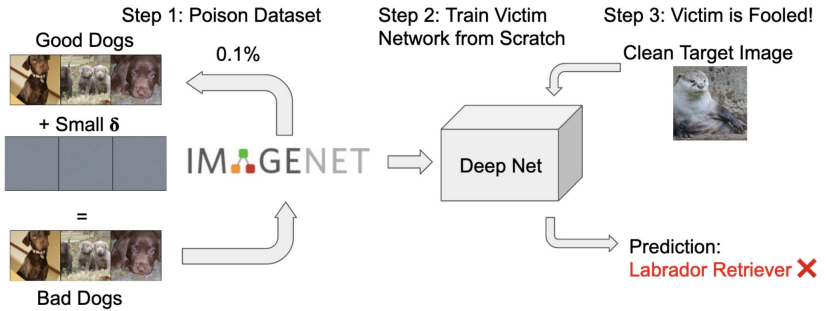


Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person’s name and access to a facial recognition system that returns a class confidence score.



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

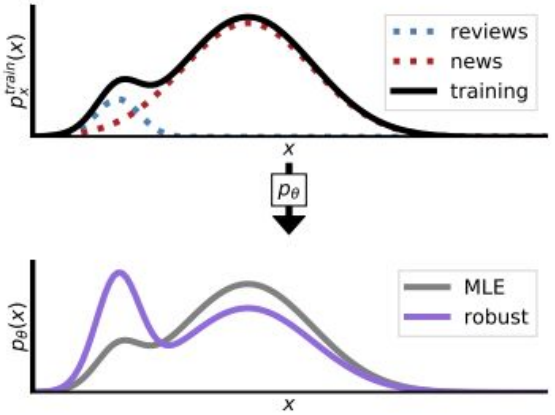
$$\min_{\theta} \max_{Q} \mathbb{E}_Q[\mathcal{L}(X, Y; \theta)] \text{ such that } Q \text{ close to } P$$

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*

Transfer Learning and Domain Adaptation

Can *knowledge* from a related task be leveraged for the task at hand?

Source task => Target task

Many flavours: multi-task learning, meta learning, few shot…

Domain adaptation: train using $(X_s, Y_s) \sim P_{\text{source}}$ and $X_t \sim P_{\text{target}}$

Methods: domain-invariant representation learning

min_f E[Loss(f(X_s), Y_s)] s.t. f(X_s) ≈ f(X_t)

Enforce f(X_s) ≈ f(X_t) using kernels (e.g. MMD) or adversarial training

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B., 2008. *Covariate Shift by Kernel Mean Matching*
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.

Domain Generalization

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

Learn "core" or "invariant" features

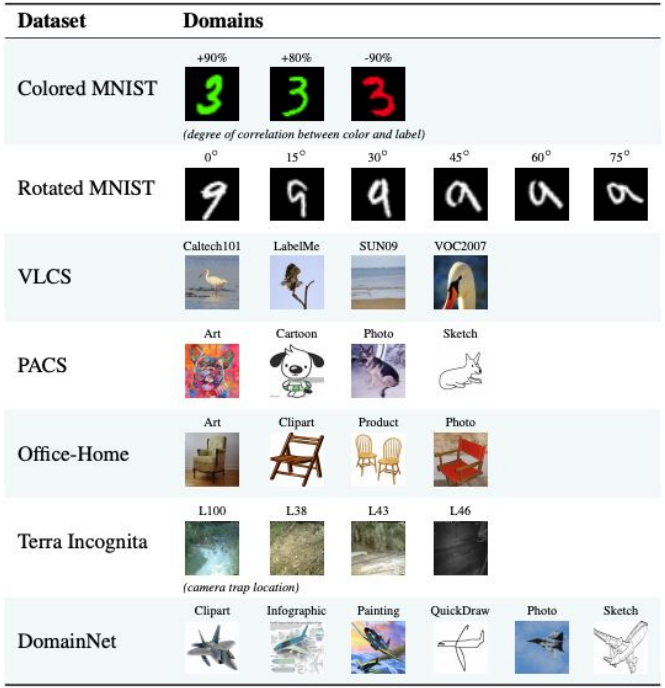
Requires *known* training set partitions, i.e. environment labels

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



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

Invariant Risk Minimization

ERM $\Phi^* = \arg\min_{\Phi} \sum_e R^e(\Phi)$

Per-environment risk

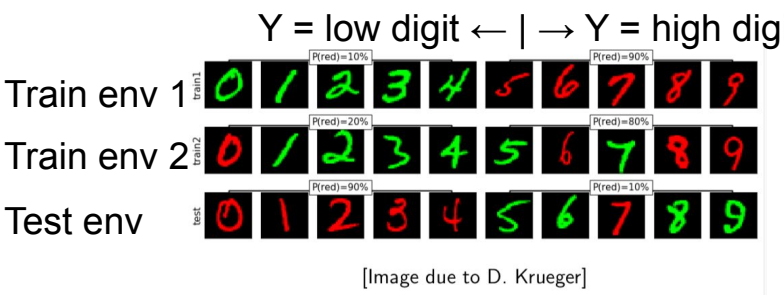
$$R^e(\Phi) = \mathbb{E}[\ell(\Phi(x), y)|e]$$

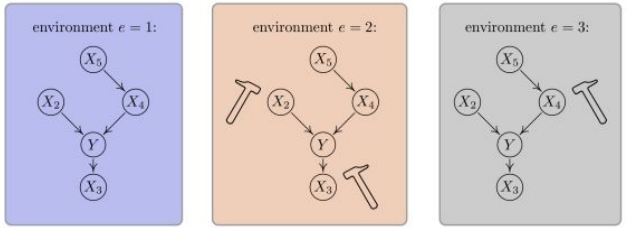
IRM: Learn representation that yields *Bayes optimal* classifier in every training environment

I.e. minimize risk subject to the **Environment Invariance Constraint** (EIC)

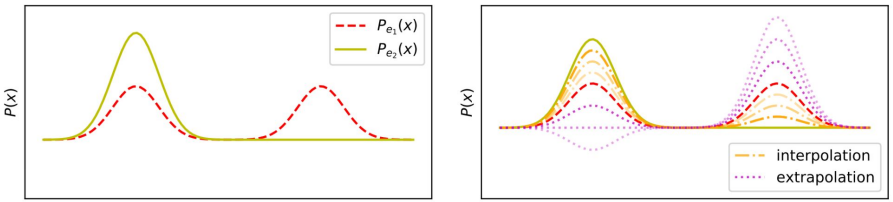
$$\mathbb{E}[y|\Phi(x) = h, e_1] = \mathbb{E}[y|\Phi(x) = h, e_2]$$
$$\forall h \in \mathcal{H} \quad \forall e_1, e_2 \in \mathcal{E}^{obs}.$$

IRMv1 regularizer is a differentiable proxy for EIC





Similar to (partial) causal discovery



REx: Optimize empirical risk s.t. low variance on per-environment risks

Peters, J., Bühlmann, P., Meinshausen, N., 2015. *Causal inference using invariant prediction*
Arjovsky et al 2019. *Invariant Risk Minimization*.
Krueger, D., et al 2021. *Out-of-Distribution Generalization via Risk Extrapolation (REx)*.

Practical Concerns

i.i.d assumption

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

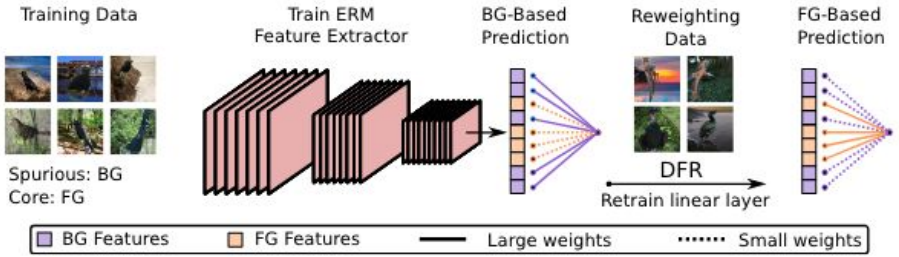
justifies train/validation/test splits

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

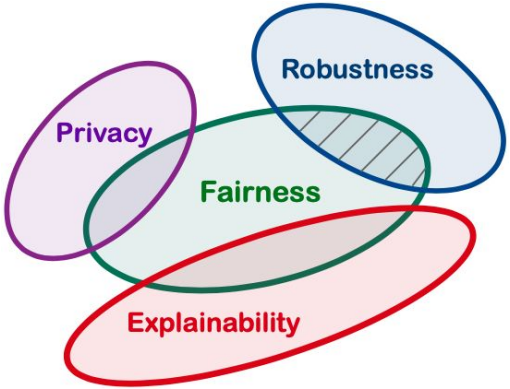
Under fair model selection criteria, ERM (standard training) is hard to beat

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.*



At the Intersections: Fairness & Robustness

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 $\mathbb{E}[\ell\|e]\ \forall e$	yes	REx (Krueger et al., 2021),	CVaR Fairness (Williamson & Menon, 2019)
min max$_e$ $\mathbb{E}[\ell\|e]$	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 $\mathbb{E}[y\|\Phi(x), e]\ \forall e$	yes	IRM (Arjovsky et al., 2019)	Group Sufficiency (Chouldechova, 2017; Liu et al., 2019)
match $\mathbb{E}[y\|\Phi(x), e]\ \forall e$	no	**EIIL (ours)**	**EIIL (ours)**
match $\mathbb{E}[\hat{y}\|\Phi(x), e, y = y']\ \forall e$	yes	C-DANN (Li et al., 2018) PGI (Ahmed et al., 2021)	Equalized Odds (Hardt et al., 2016)
match $\left\|\mathbb{E}[y\|S(x), e] - \mathbb{E}[\hat{y}(x)\|S(x), e]\right\|\ \forall e$	no		Multicalibration (Hébert-Johnson et al., 2018)
match $\left\|\mathbb{E}[y\|e] - \mathbb{E}[\hat{y}(x)\|e]\right\|\ \forall e$	no		Multiaccuracy (Kim et al., 2019)
match $\left\|\mathbb{E}[y \neq \hat{y}(x)\|y = 1, e]\right\|\ \forall e$	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

Caveat: not the whole story!

- Technical fairness approaches limited in scope
- Task and target variable definition matter *a lot*

However, some unfairness comes from failure to generalize OOD

- Recall: subpopulation shift

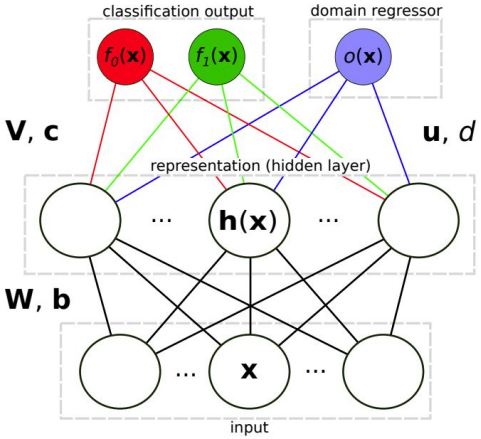
Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D., 2017. *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.*

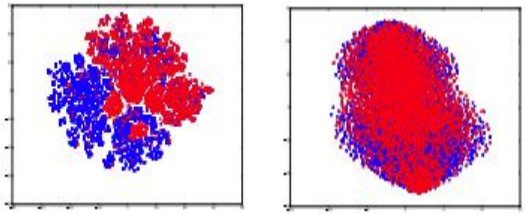
Representation learning approaches

Neural net approaches to statistical fairness influenced by domain adaptation

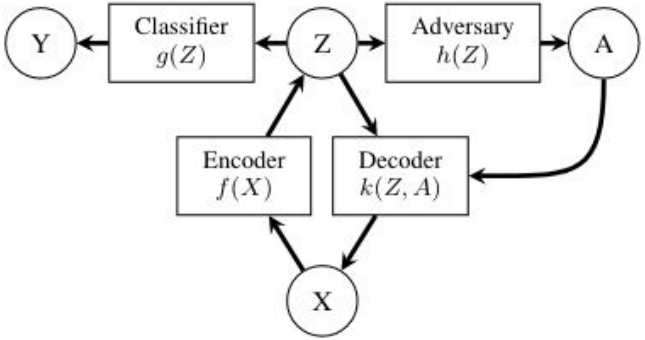
E.g. adversarial training with auxiliary labels

“Fair” representations can transfer to new tasks



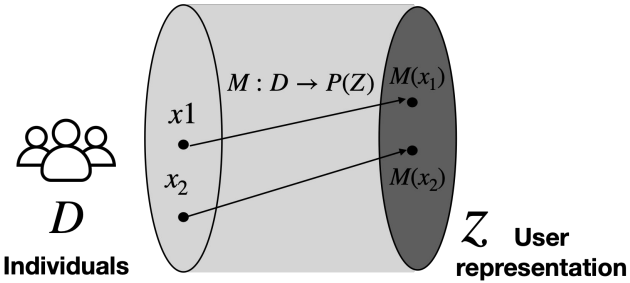


Tra. Task	TarUnf	TraUnf	TraFair	TraY-AF	LAFTR
MSC2a3	0.362	0.370	0.381	0.378	**0.281**
METAB3	0.510	0.579	**0.436**	0.478	0.439
ARTHSPIN	0.280	0.323	0.373	0.337	**0.188**
NEUMENT	0.419	0.419	0.332	0.450	**0.199**
RESPR4	0.181	0.160	0.223	0.091	**0.051**
MISCHRT	0.217	0.213	0.171	0.206	**0.095**
SKNAUT	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.*

Fairness via robustness to perturbations (i.e. smoothness)



Robustness methods can encourage *smoothness* of model's predictive fn w.r.t.

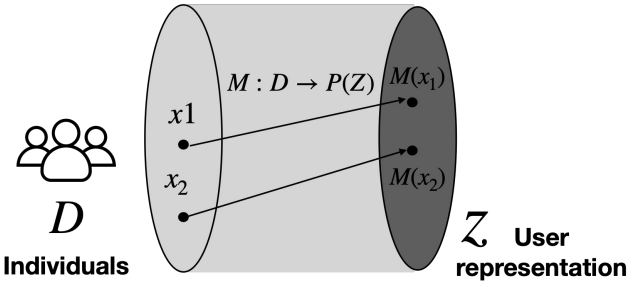
- Pairwise similarity metric for *individual fairness*
  - e.g. distributionally robust optimization or adversarial robustness
- Subpopulation shift for *group fairness*
  - e.g. (group) distributionally robust optimization

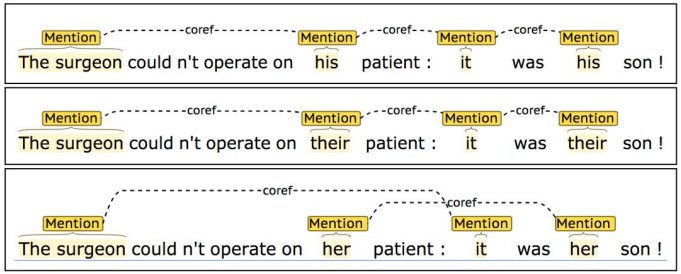
Yurochkin, M., Bower, A., Sun, Y., 2020. *Training individually fair ML models with Sensitive Subspace Robustness.*

Yeom, S., Fredrikson, M., 2020. *Individual Fairness Revisited: Transferring Techniques from Adversarial Robustness*

Hashimoto, T.B., Srivastava, M., Namkoong, H., Liang, P., 2018. *Fairness Without Demographics in Repeated Loss Minimization.*

Fairness via robustness to perturbations (i.e. smoothness)





Robustness methods can encourage *smoothness* of model's predictive fn w.r.t.

- Pairwise similarity metric for *individual fairness*
  - e.g. distributionally robust optimization or adversarial robustness
- Subpopulation shift for *group fairness*
  - e.g. (group) distributionally robust optimization
- Feature-level perturbation known to reveal model sensitivity (e.g. gendered pronoun swap in text)
  - e.g. "counterfactual" data augmentation

Yurochkin, M., Bower, A., Sun, Y., 2020. *Training individually fair ML models with Sensitive Subspace Robustness.*

Yeom, S., Fredrikson, M., 2020. *Individual Fairness Revisited: Transferring Techniques from Adversarial Robustness*

Hashimoto, T.B., Srivastava, M., Namkoong, H., Liang, P., 2018. *Fairness Without Demographics in Repeated Loss Minimization.*

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E.H., Beutel, A., 2019. *Counterfactual Fairness in Text Classification through Robustness*

Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B., 2018. *Gender Bias in Coreference Resolution.*

Min-max fairness

Recall tradeoff: matching performance across groups vs overall accuracy

$\min_f$ E[Loss(f(X),Y)] s.t. E[Loss(f(X),Y)|A=0] = E[Loss(f(X),Y)|A=1]

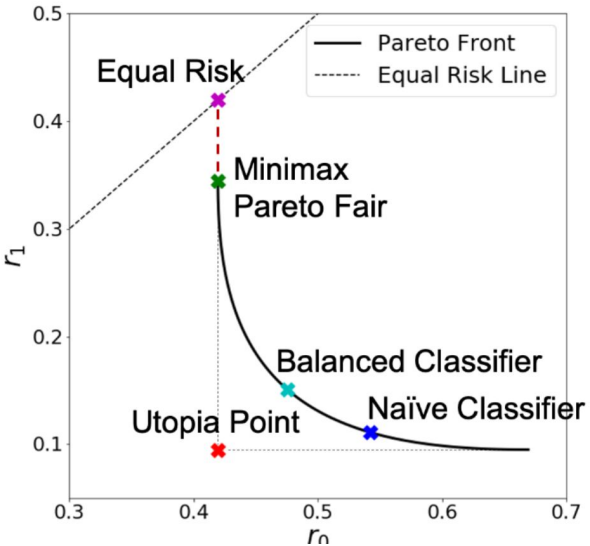
may increase loss for non-worst-off groups…"unnecessary harm"?

Alternative fairness notion:

$\min_f \max_a$ E[Loss(f(X),Y)|A=a]

(compatible with distributionally robust optimization)

Can also consider pareto front over {E[Loss(f(X),Y)|A=a]}



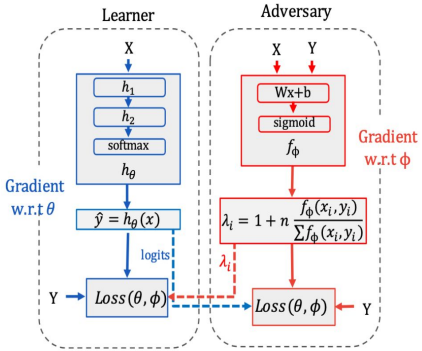
Martinez, N., Bertran, M., Sapiro, G., 2020. *Minimax Pareto Fairness: A Multi Objective Perspective.*
Diana, E., Gill, W., Kearns, M., Kenthapadi, K., Roth, A., 2021. *Minimax Group Fairness: Algorithms and Experiments.*

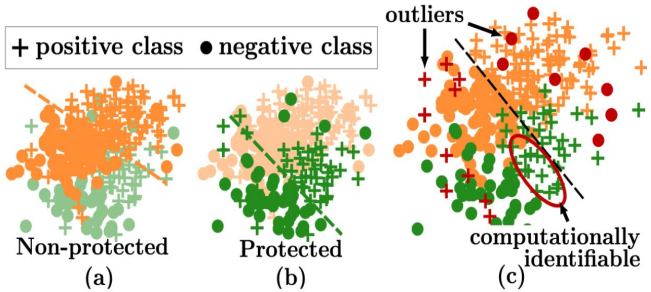
Lessons from fairness to robustness

Access to auxiliary labels limited in practice

Fairness without demographics

- Multiaccuracy/multicalibration
  - consider “computationally identifiable groups”
- Adversarially reweighted learning
- Environment Inference for Invariant Learning





(a) **Inferred environment 1** *(mostly) landbirds on land, and waterbirds on water*

(b) **Inferred environment 2** *(mostly) landbirds on water, and waterbirds on land*

Hébert-Johnson, Ú., Kim, M.P., Reingold, O., Rothblum, G.N., 2018. *Calibration for the (Computationally-Identifiable) Masses.*
Kim, M.P., Ghorbani, A., Zou, J., 2018. *Multiaccuracy: Black-Box Post-Processing for Fairness in Classification.*
Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., Chi, E.H., 2020. *Fairness without Demographics through Adversarially Reweighted Learning.*
Creager, E., Jacobsen, J.-H., Zemel, R., 2021. *Environment Inference for Invariant Learning*

Causality-inspired methods

Graphs encode assumptions about dist'ns

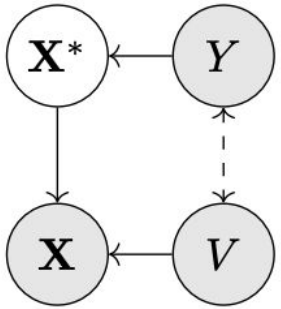
Fairness on confounded data

⇔

Independence on unconfounded data

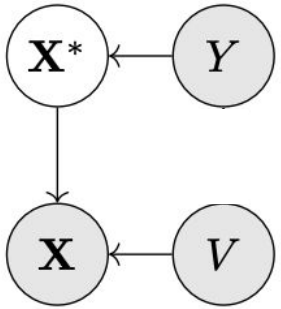
Unconfounded data not available

emulate via importance weights



Confounded (training) data: The main label Y and auxiliary label V generate input $\boldsymbol{X}$, but Y only affects $\boldsymbol{X}$ through $\boldsymbol{X}^*$

$P_C = P(\boldsymbol{X}|\boldsymbol{X}^*, V)P(\boldsymbol{X}^*|Y)P(Y)P(V|Y)$



Unconfounded data: Spurious correlation between Y and V is removed

$P_U = P(\boldsymbol{X}|\boldsymbol{X}^*, V)P(\boldsymbol{X}^*|Y)P(Y)\mathbf{P(V)}$

Schoelkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J., 2012. *On Causal and Anticausal Learning.*
Veitch, V., D'Amour, A., Yadlowsky, S., Eisenstein, J., 2021. *Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests.*
Makar, M., D'Amour, A., 2022. *Fairness and robustness in anti-causal prediction.*

Fair *and* robust learning

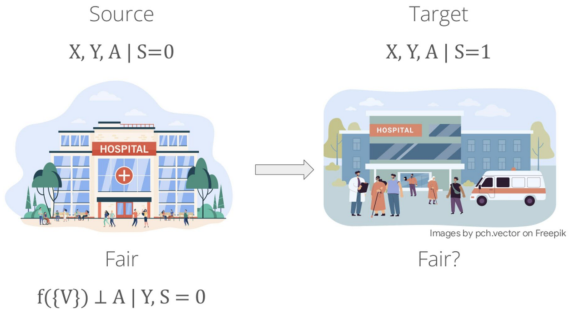
Fair representations can fail under distribution shifts

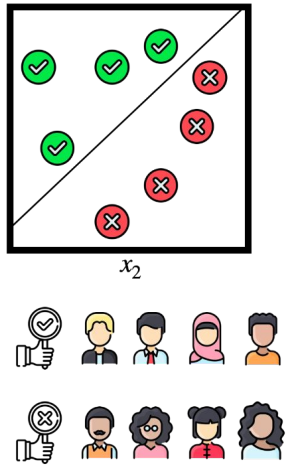
Fair learning + DRO helps

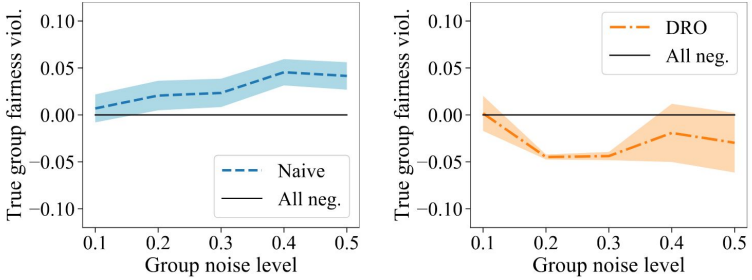
- Mostly simulated studies

Noisy observations

- Sensitive attributes
- Targets (esp. in risk assessment)







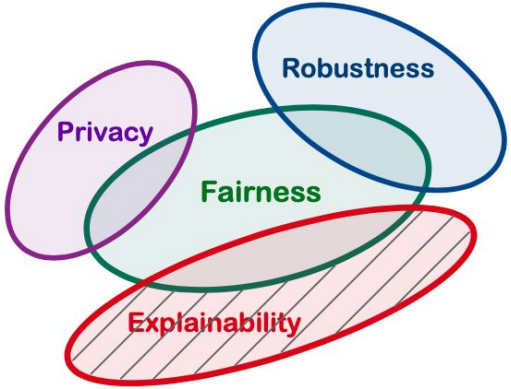
Lechner, T., Ben-David, S., Agarwal, S., Ananthakrishnan, N., 2021. *Impossibility results for fair representations.*
Rezaei, A., Liu, A., Memarrast, O., Ziebart, B., 2021. *Robust Fairness under Covariate Shift.*
Singh, H., Singh, R., Mhasawade, V., Chunara, R., 2021. *Fairness Violations and Mitigation 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?

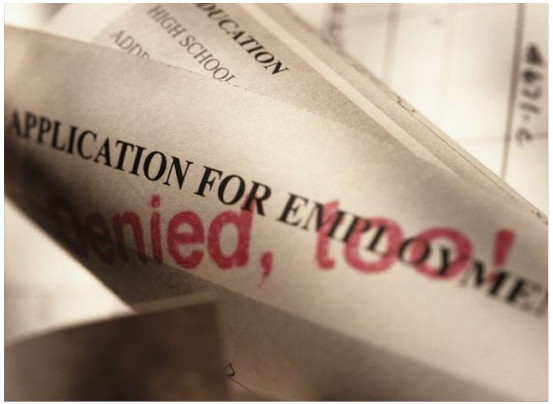


Introduction to AI Explainability

Why explainable AI (XAI)?

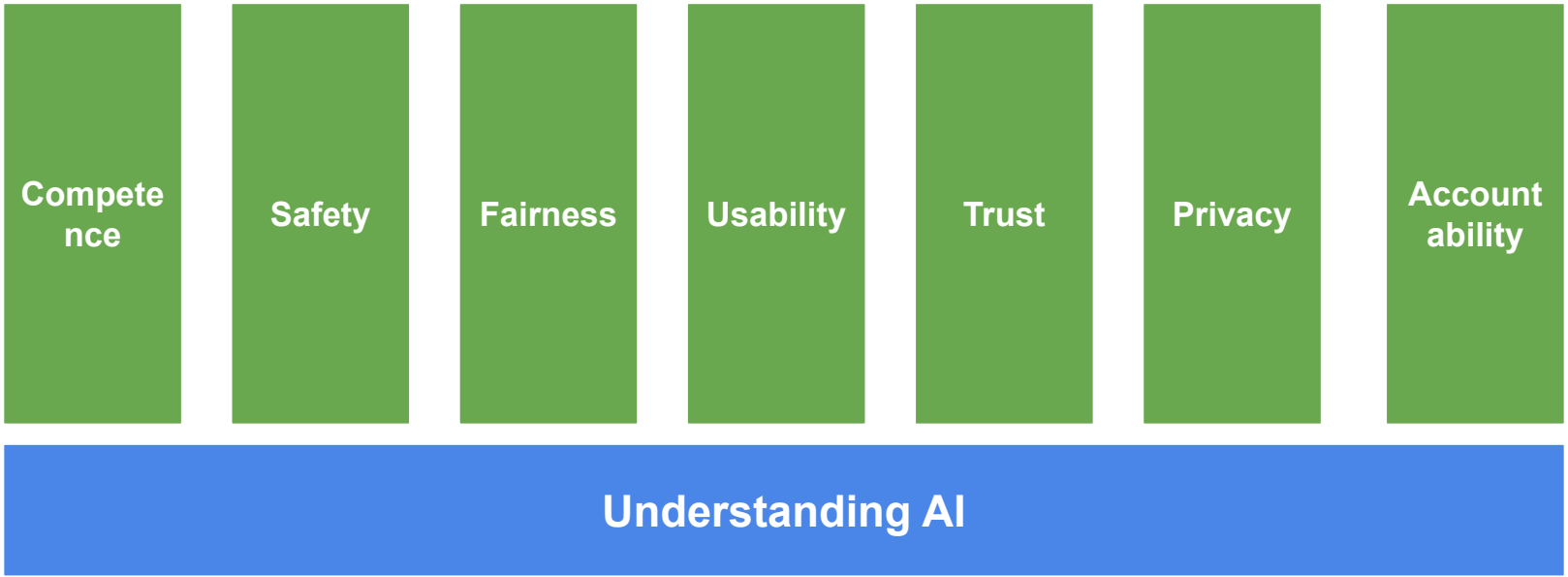
explainability/interpretability/intelligibility/…
= making AI **understandable by people**

AI is increasingly used to assist humans and impact many aspects of human lives



Human scrutiny and interventions are critical

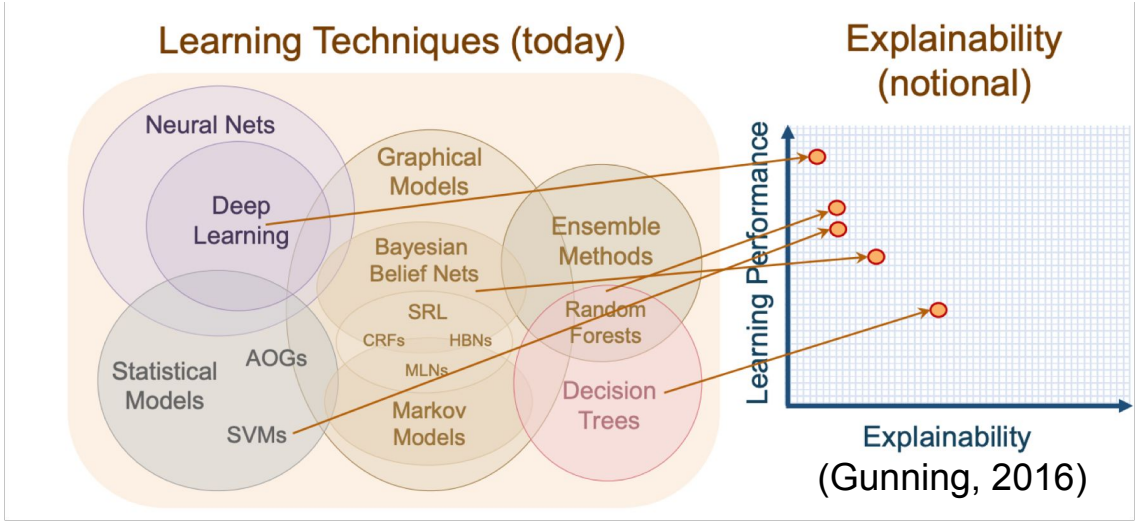
Explainability as means to many ends



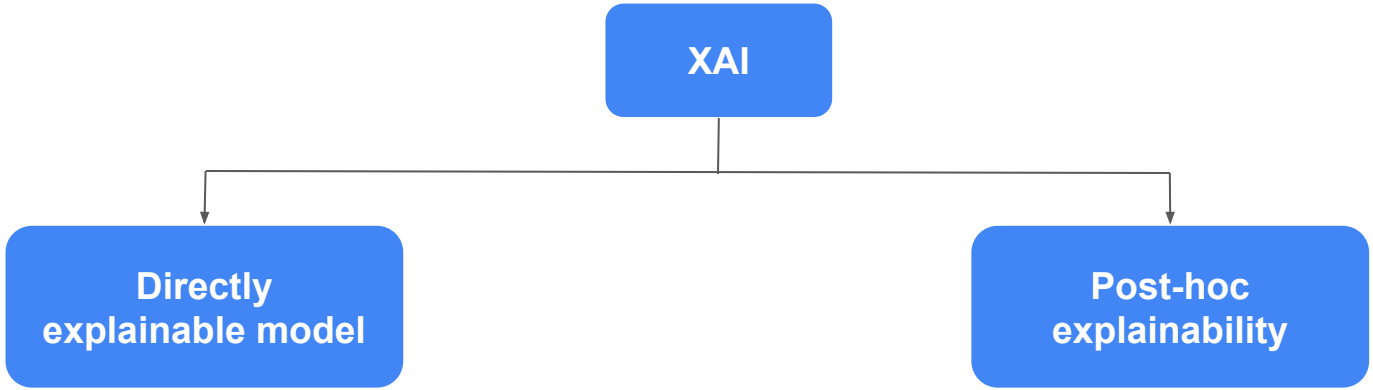
Why is explainable AI challenging?

First, not all algorithms are directly explainable

Explainability-performance tradeoff?
Only in *some* settings



Rudin. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence

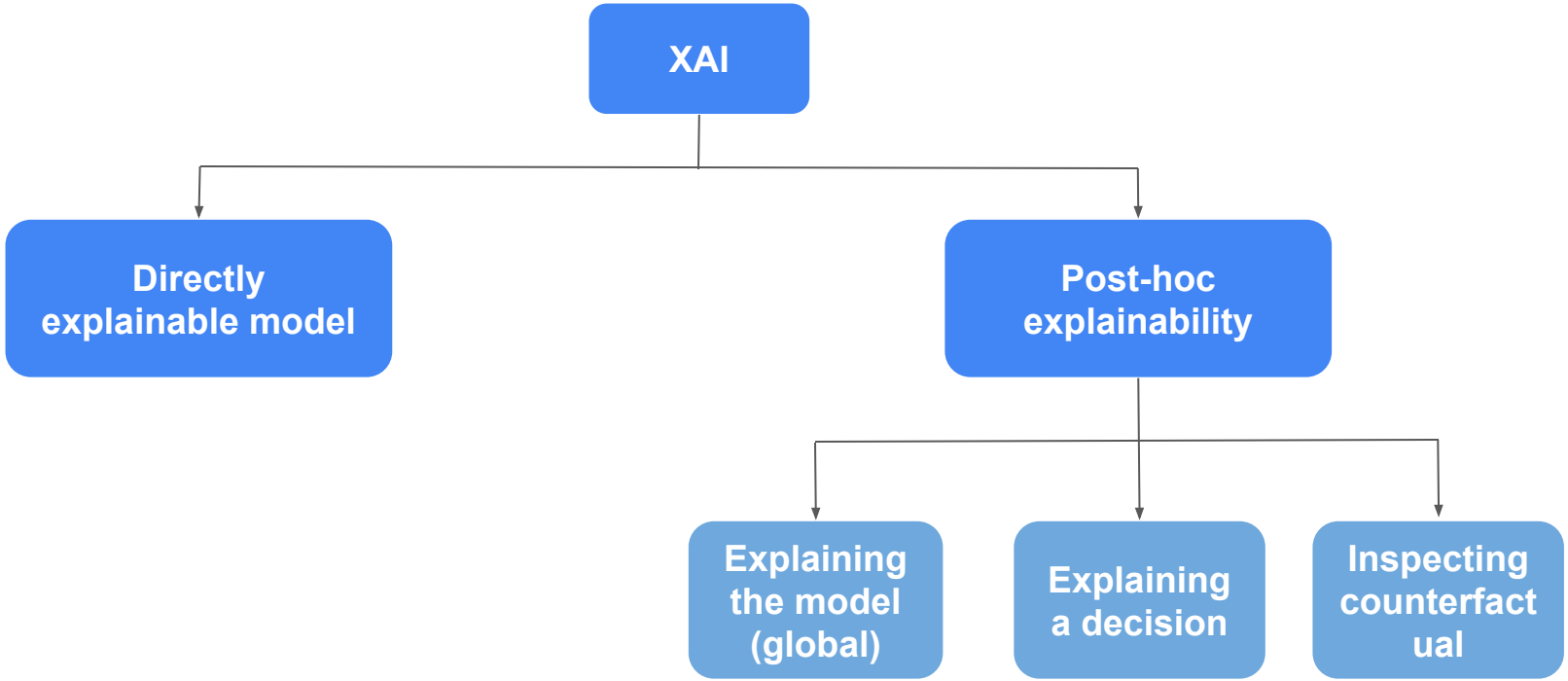


- Linear model
- Rule-based model
- Decision-tree

Breaking the "explainability-performance trade-off"

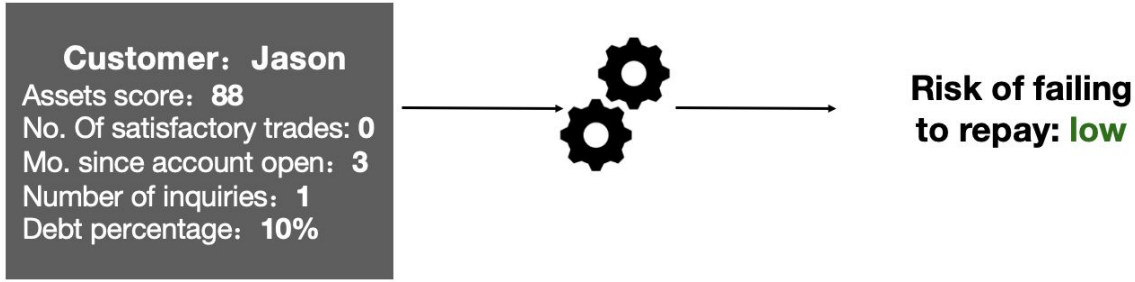
- General additive models
- General rule models
- …

Caruana et al. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015
Wei et al. Generalized Linear Rule Models. ICML 2019



Guidotti et al. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* .

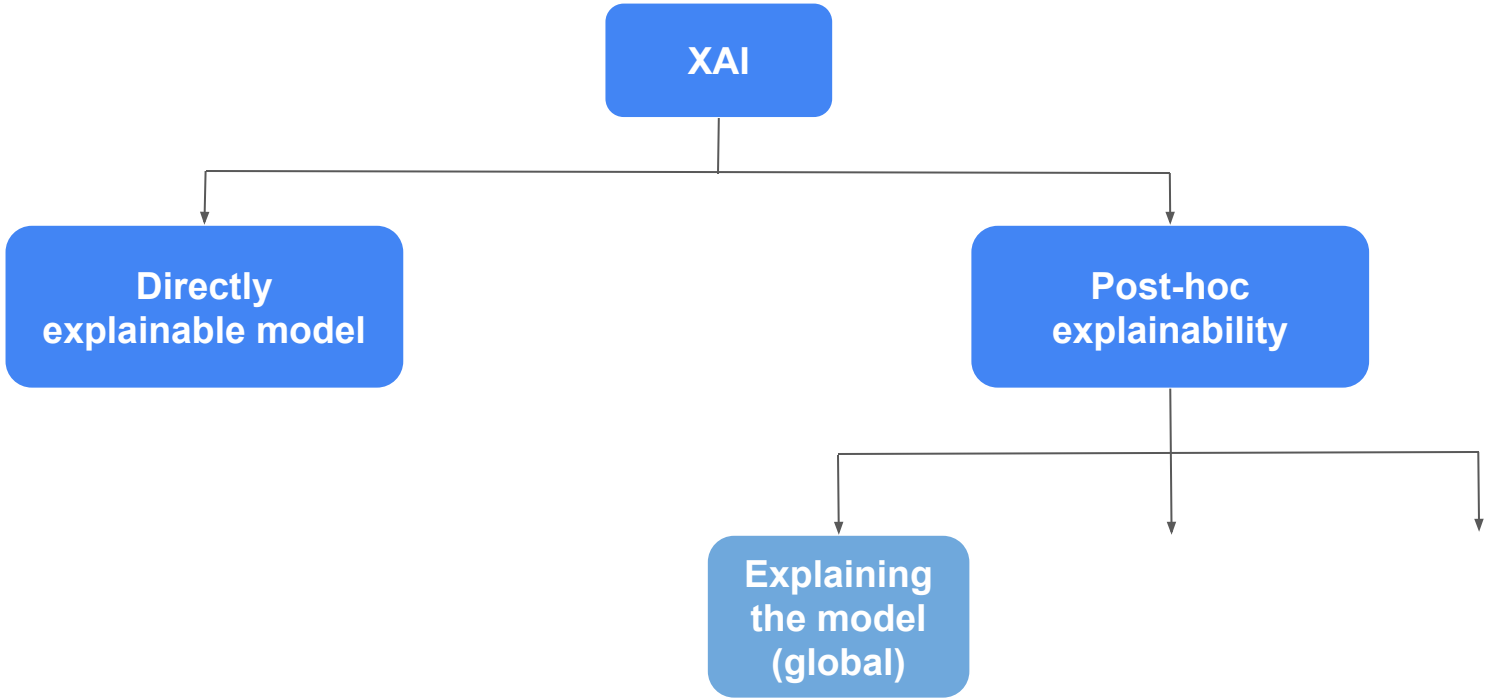
Use case: a decision-support ML for loan application approval



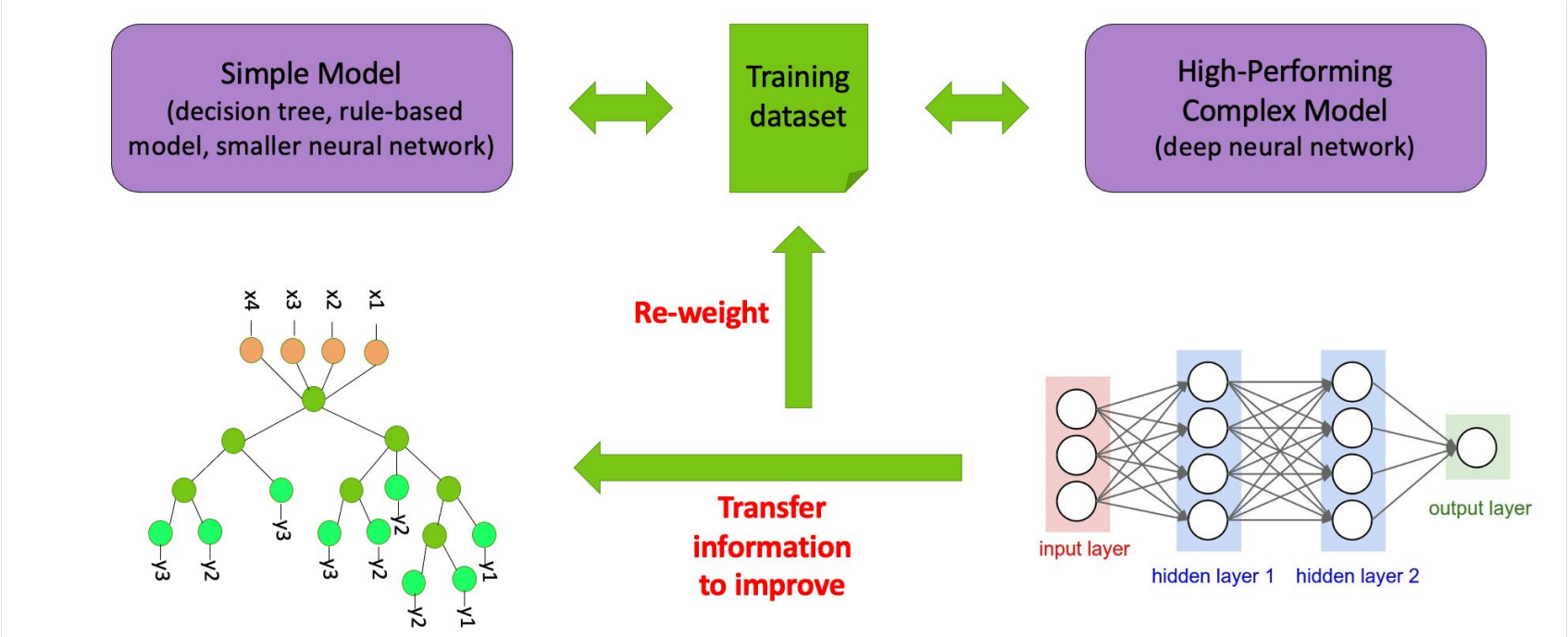
Data scientist
Must ensure the model works appropriately before deployment

Loan officer
Needs to assess the model's prediction and make the final judgment

Bank customer
Wants to understand the reason for the application result


XAI
Directly explainable model
Post-hoc explainability
Explaining the model (global)

Post-hoc global explanation: knowledge distillation/ approximation



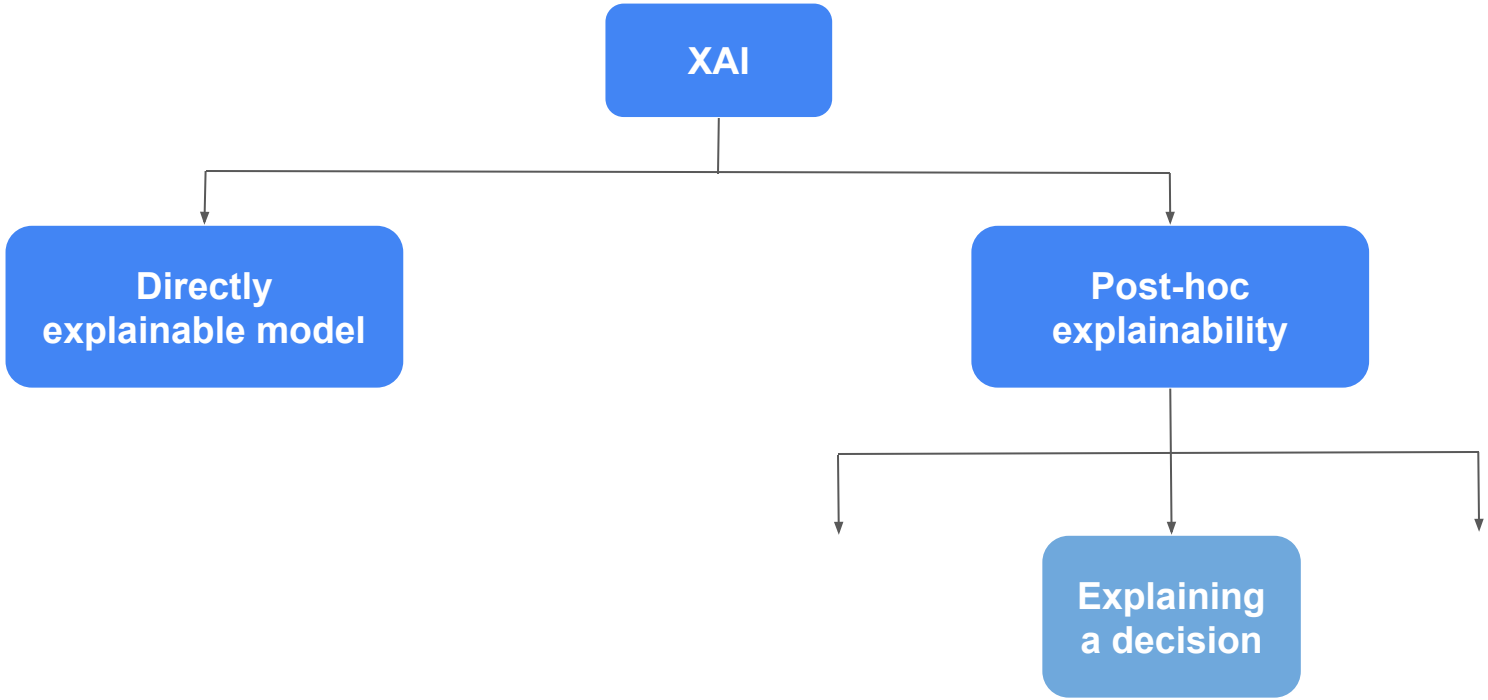
Example global explanation: rule sets

- If {**assets score> 90, Mo. since account opening>6**}:**Low** risk
- Else if {**Debt percentage< 15**}:**Low** risk

Loan officer

What kind of customers does the model consider as low risk?

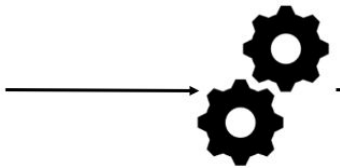
Lakkaraju et al.,Faithful and customizable explanations of black box models. AIES 2019.



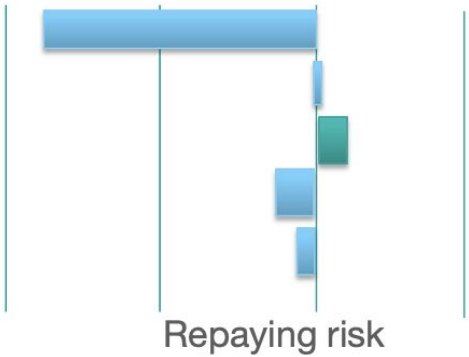
Guidotti et al. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* .

Explaining a decision by feature: feature contribution

Customer: Jason
Assets score: **88**
No. Of satisfactory trades: **0**
Mo. since account open: **3**
No. of inquiries: **1**
Debt percentage: **10%**



Risk of failing to repay: low

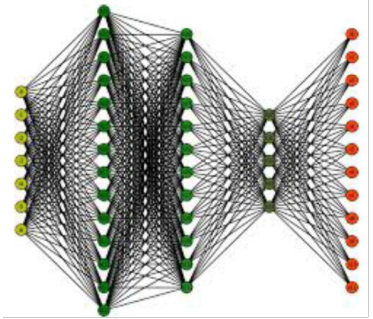


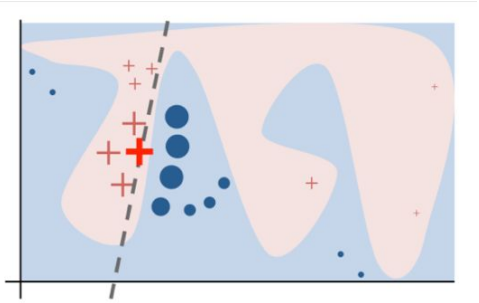
Assets score
No. Of satisfactory trades
Mo. since account open
No. Of inquiries
Debt percentage

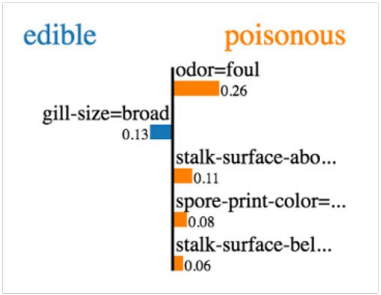
Loan officer

Why is Jason predicted of low risk?

Example post-hoc local explanation: LIME

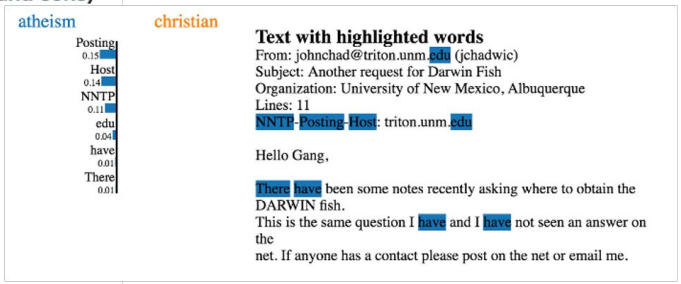






tabular

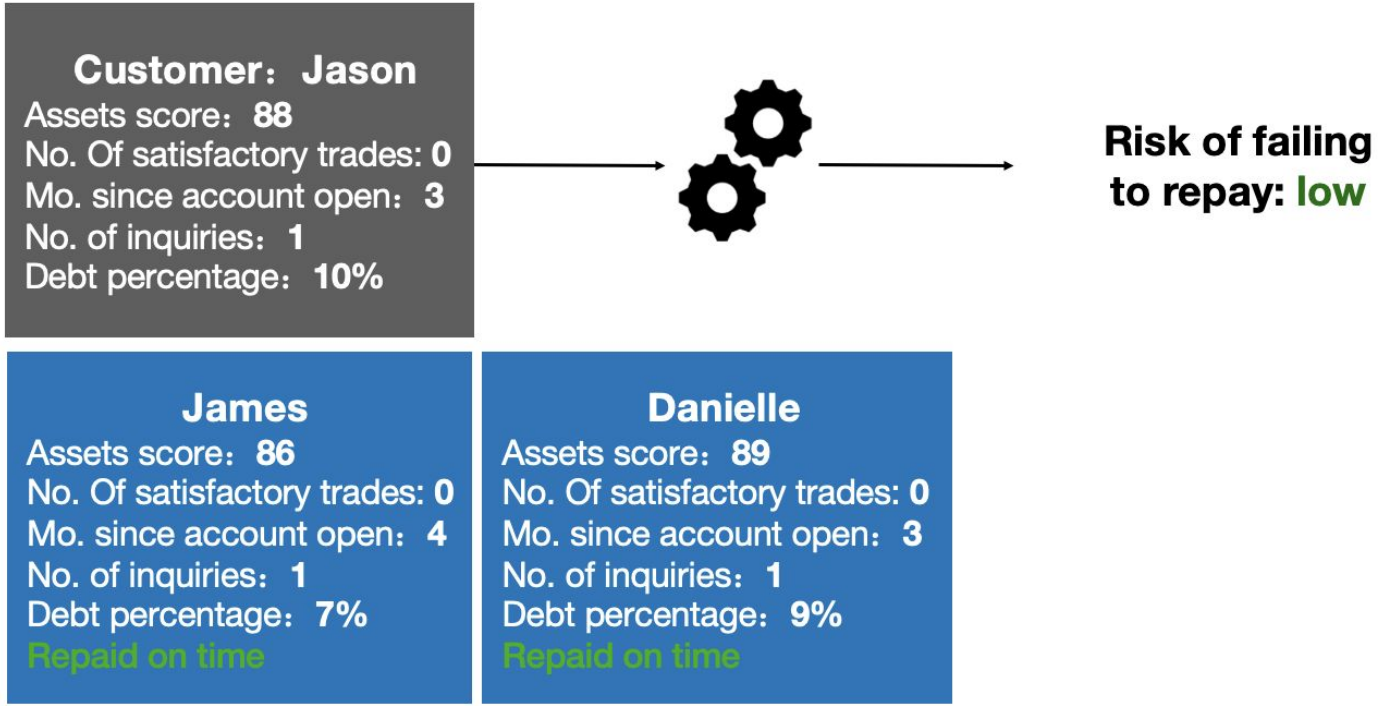
image



texts

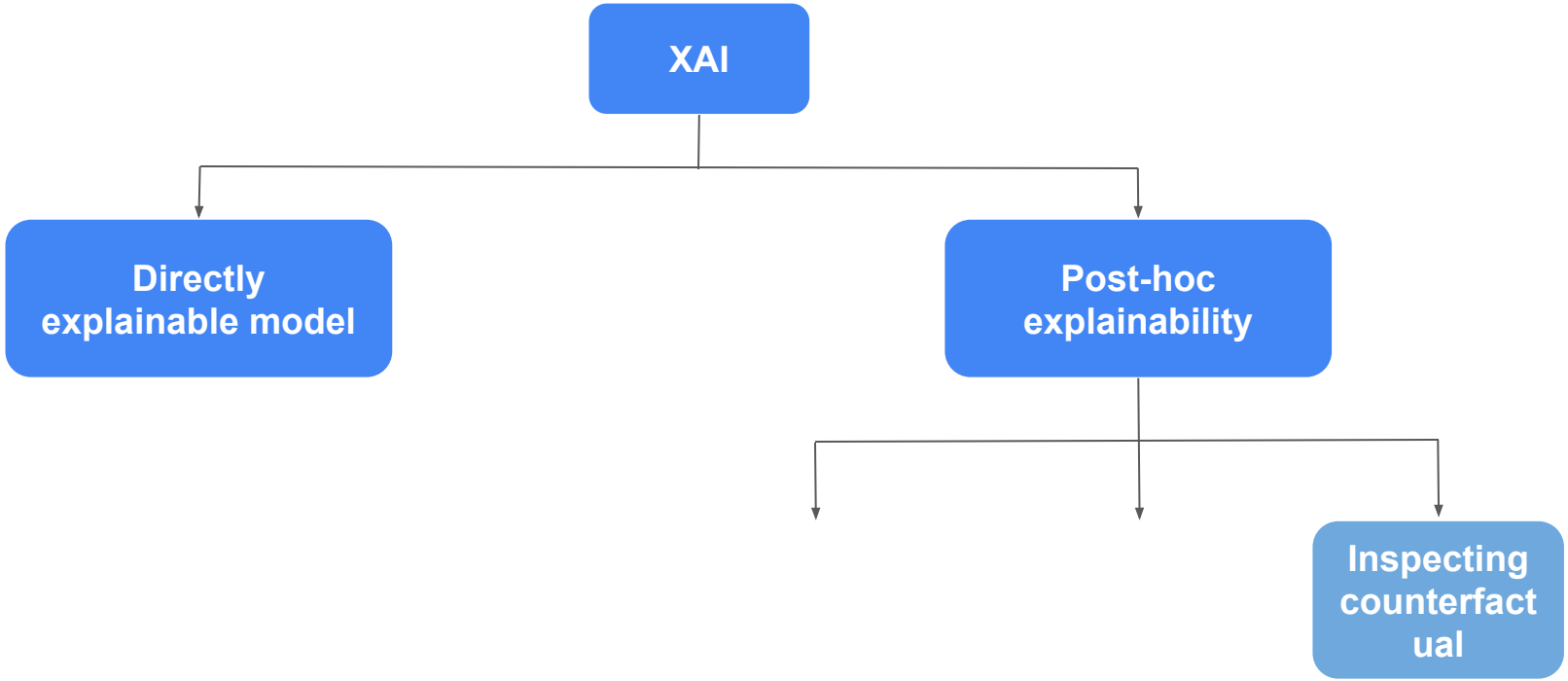
Ribeiro et al. Why should i trust you?" Explaining the predictions of any classifier. KDD 2016

Explaining a decision by examples

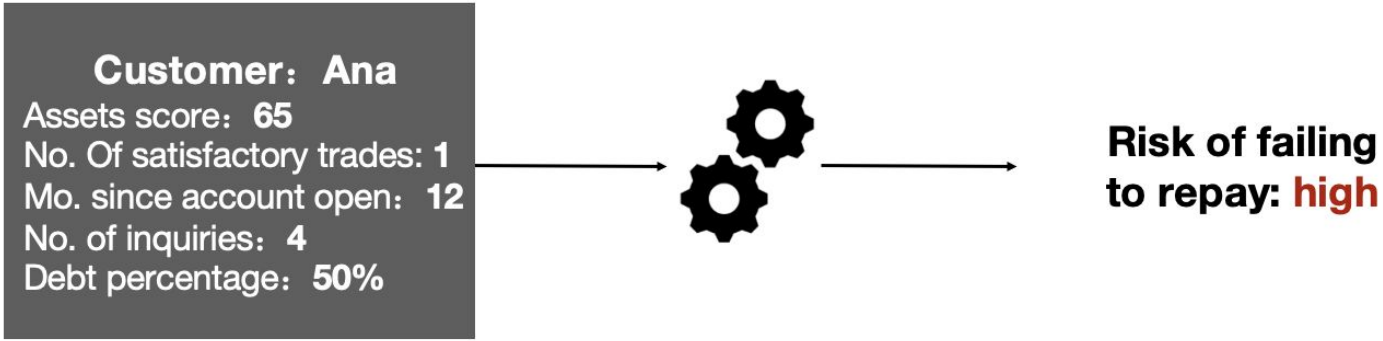


Chen et al. This looks like that: deep learning for interpretable image recognition. NeurIPS 2019

Gurumoorthy et al. Efficient Data Representation by Selecting Prototypes with Importance Weights. ICDM 2019


XAI
Directly explainable model
Post-hoc explainability
Inspecting counterfactual

Inspecting counterfactual: contrastive features



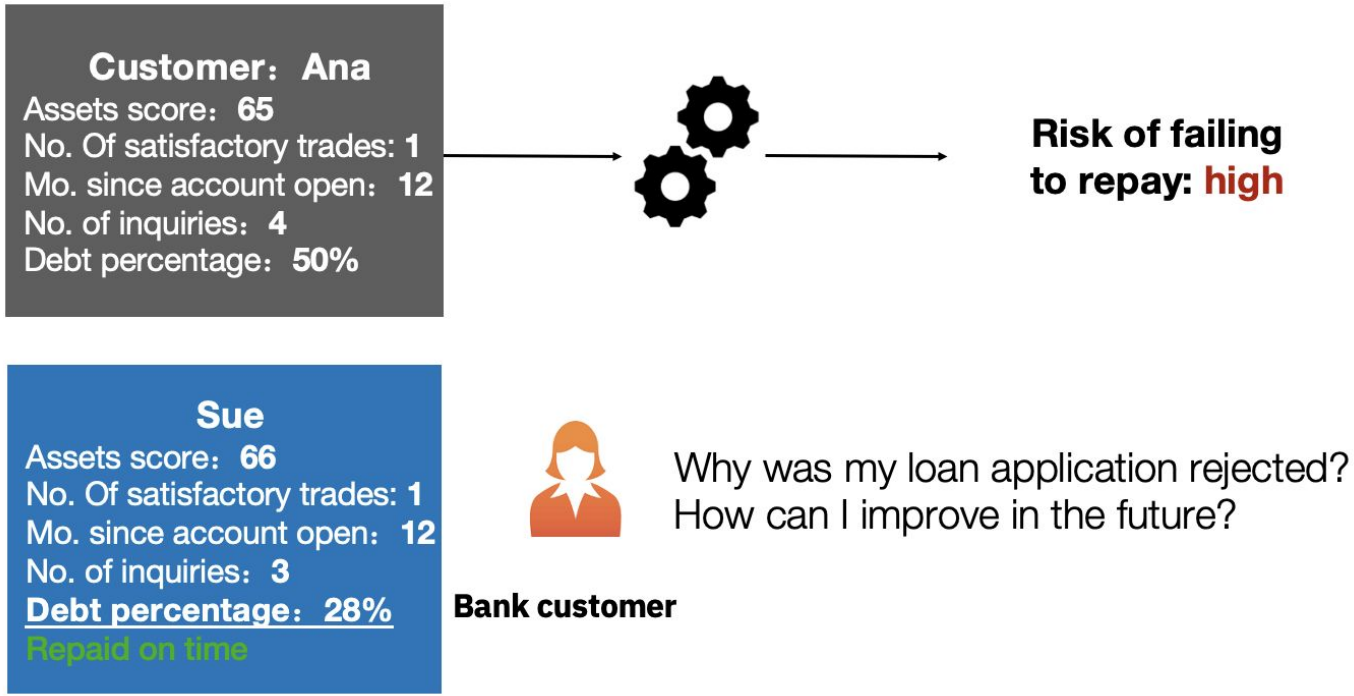
•If **{debt percentage under 30%}** ,
you will no longer be predicted of high risk

Why was my loan application rejected? How can I improve in the future?

Bank customer

Dhurandhar, et al. Explanations based on the missing: Towards contrastive explanations with pertinent negatives.NeurIPS 2018

Inspecting counterfactual: counterfactual examples



Mothilal et al. Explaining machine learning classifiers through diverse counterfactual explanations. FAccT 2020

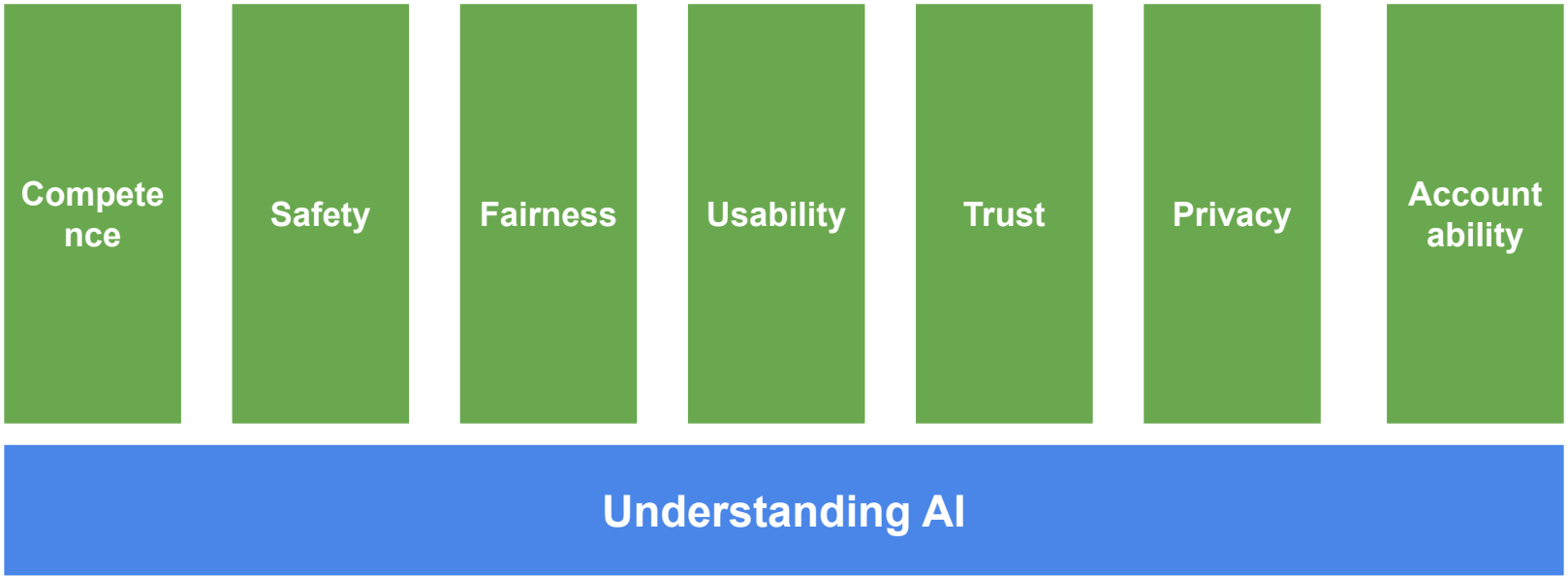
Why is explainable AI challenging?

HCXAI
HUMAN-CENTERED EXPLAINABLE AI

- Exposing algorithmic processes does not guarantee human understanding
  - Understanding is multi-faceted
  - Understanding may require information beyond algorithmic processes
- Challenges to support many ends of explainability

Liao & Varshney, (2021). Human-centered Explainable AI (XAI): From Algorithms to User Experiences.
Liao et al. (2020). Questioning the AI: informing design practices for explainable AI user experiences. CHI 2020
Ehsan et al. (2021). Operationalizing human-centered perspectives in explainable AI. CHI 2021 EA

Explainability as means to many ends

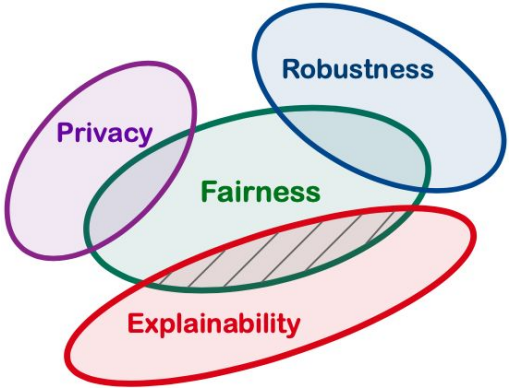


Why is explainable AI challenging?

HCXAI
HUMAN-CENTERED EXPLAINABLE AI

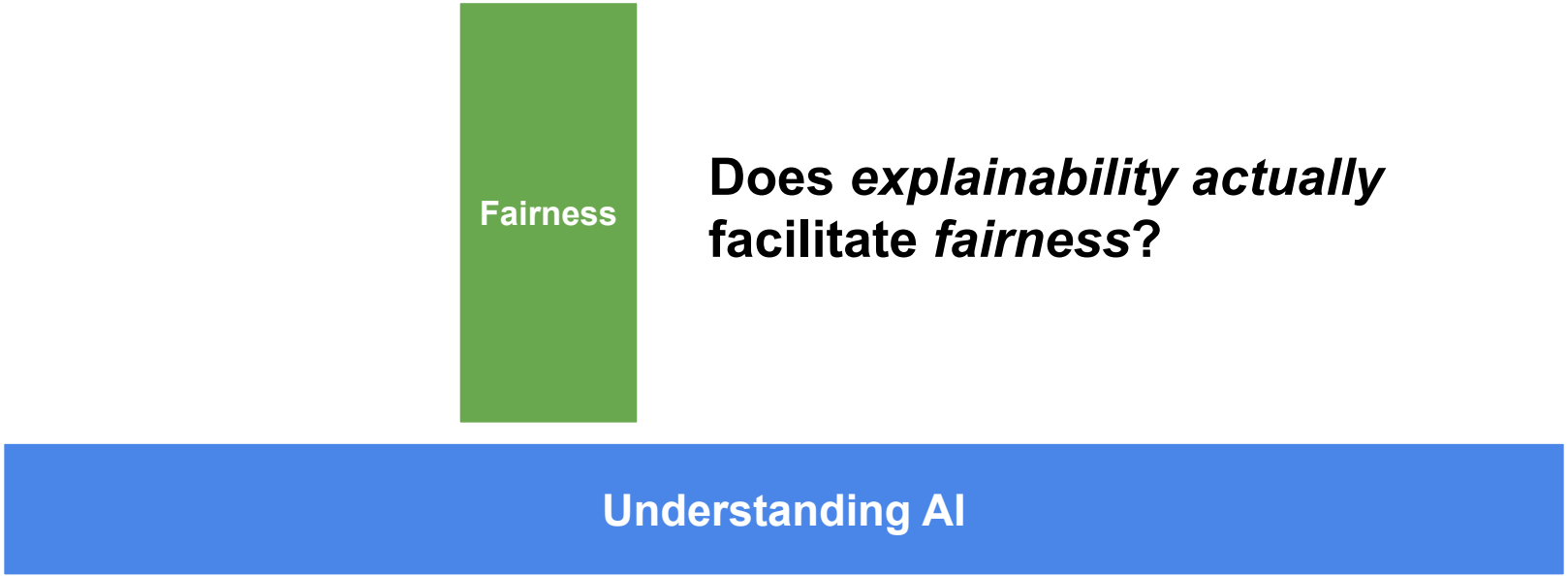
- Exposing algorithmic processes does not guarantee human understanding
  - Understanding is multi-faceted
  - Understanding may require information beyond algorithmic processes
- Challenges to support many ends of explainability
  - No one-fits-all solutions
  - Empirical results are still inconclusive in many cases
  - Different end-goals/use cases are not accounted for when developing algorithms
- Current XAI paradigms may not be all compatible with human cognitive processes to seek and consume explanations

Liao & Varshney, (2021). Human-centered Explainable AI (XAI): From Algorithms to User Experiences.



At the Intersections: Fairness & Explainability

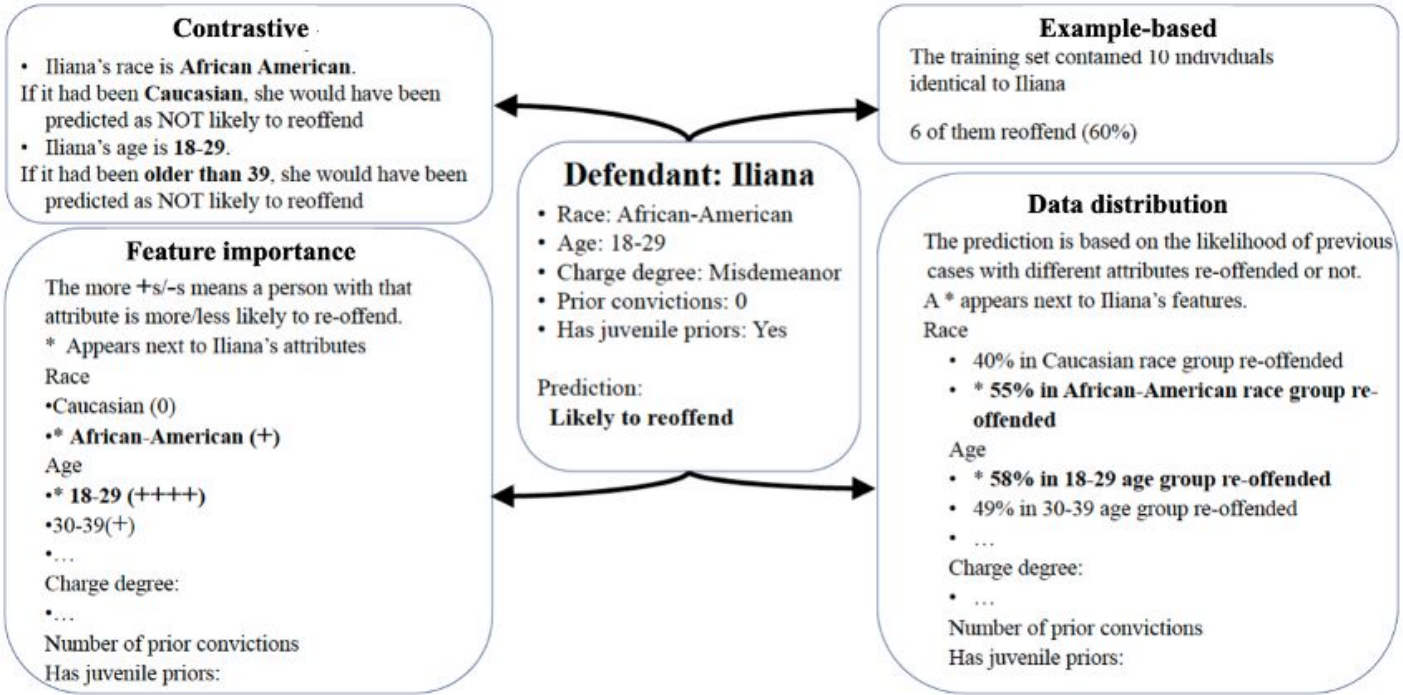
What happens at the intersection?



Why explainability as human interface for fairness?

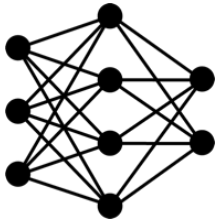
- The decisions to apply fairness metrics and bias mitigation may need to be human-in-the-loop
- When metrics are not available: e.g., end-users, auditing & governance

Which explanation supports human fairness judgment?



Dodge et al.. Explaining models: an empirical study of how explanations impact fairness judgment. IUI 2019

Evaluation construct: fairness *calibration*



Statistically unfair model

Condition 1

How the AI made the prediction was fair

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Statistically fairer model

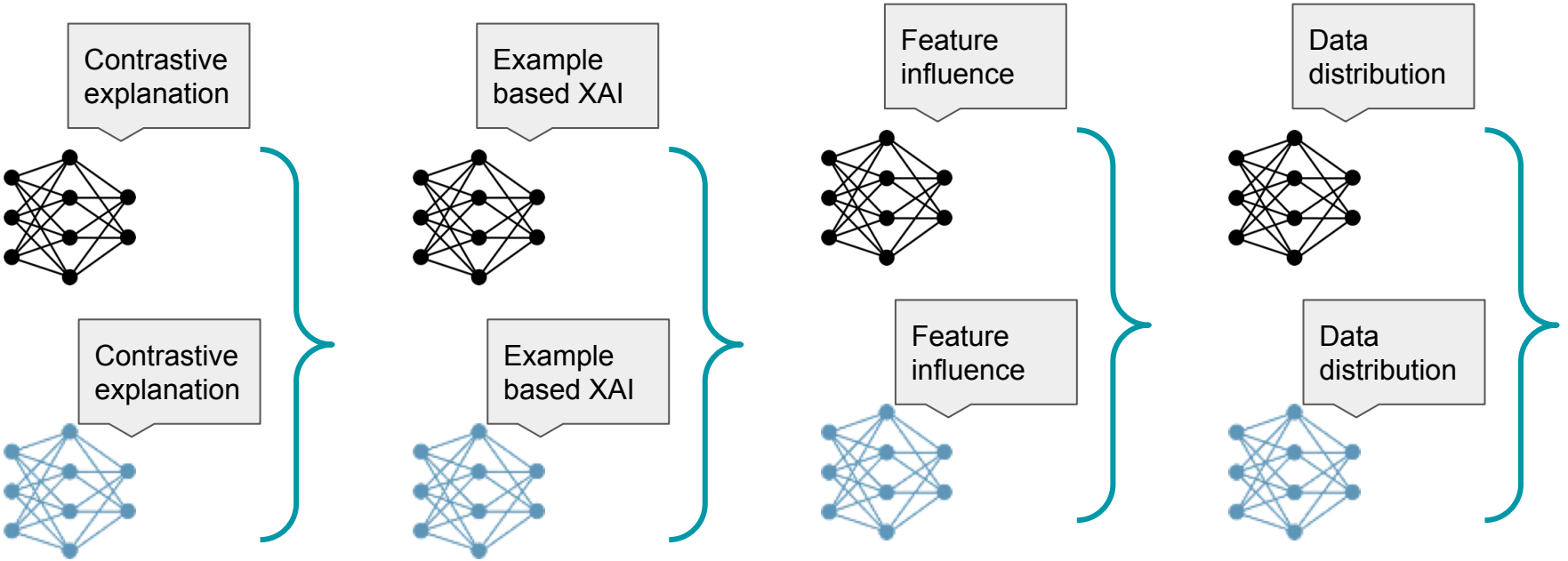
Condition 2

How the AI made the prediction was fair

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Measurement: **fairness rating distance** between fairer and unfair models

Evaluation construct: fairness *calibration*



Which explanation supports human fairness judgment?

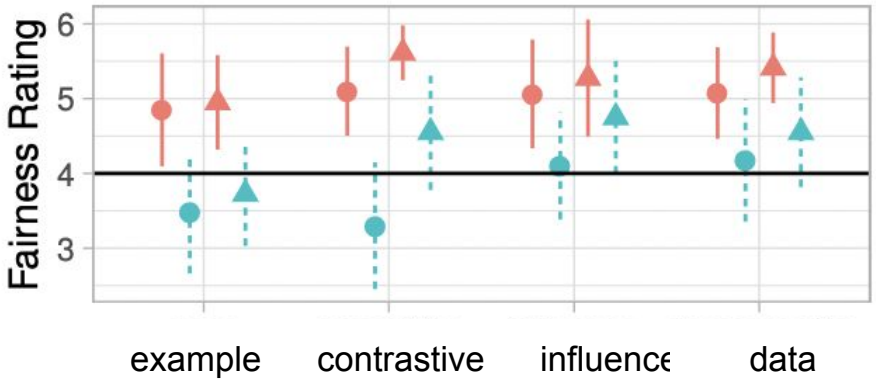


Figure 2: Overall mean ratings of fairness, per explanation type, data process treatment *(raw=●, processed=▲)*, and sample group *(impacted=blue dashed lines, non-impacted=red solid lines)*. The lines indicate the 95% confidence intervals.

- All explanations helped people distinguish between fairer and unfair models
- Local explanations are slightly more effective
- Especially effective when contrastive explanation reveals issues of individual unfairness

Contrastive

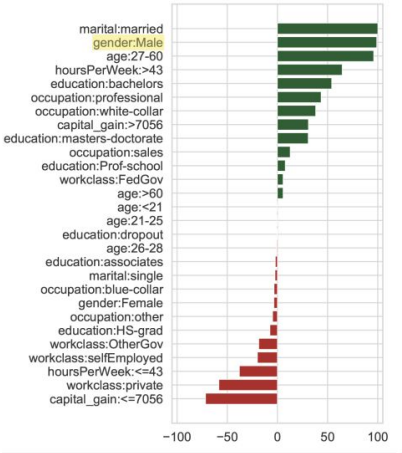
- Iliana's race is **African American**.

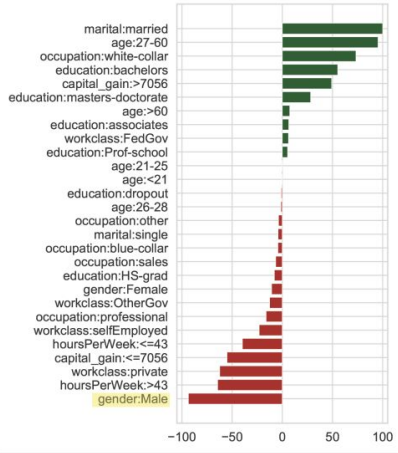
If it had been **Caucasian**, she would have been predicted as NOT likely to reoffend

- Iliana's age is **18-29**.

If it had been **older than 39**, she would have been predicted as NOT likely to reoffend

Open question: Explanation and fairwashing





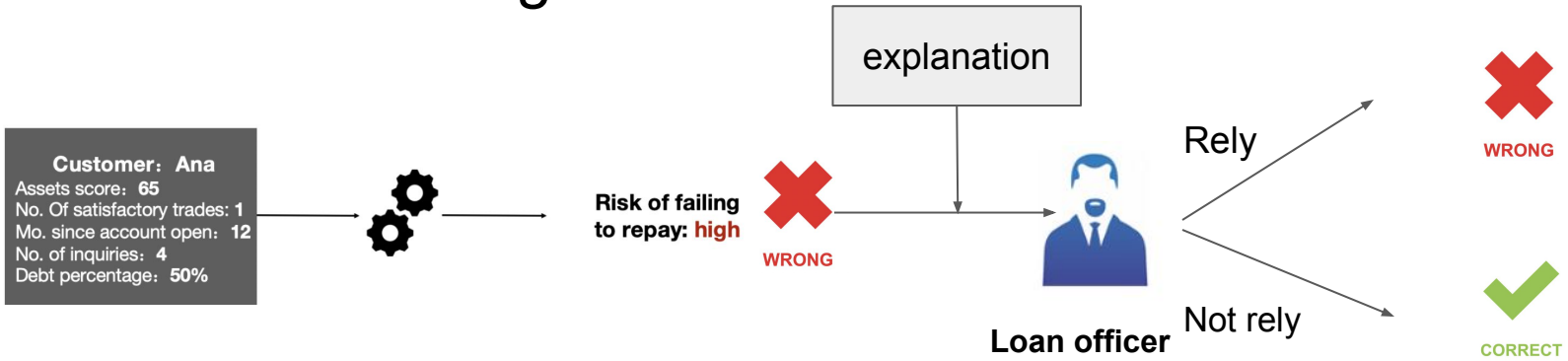
if capital_gain:>7056 **then**
  income≥$50k$
else if marital:single **then**
  income<$50k$
else if education:HS-grad **then**
  income<$50k$
else if occupation:other **then**
  income<$50k$
else if occupation:white-collar **then**
  income≥$50k$
else
  income:<$50k$

How precise is explanation in revealing model biases, especially with post-hoc explanations?

Fairwashing: It is possible to create explanations that are highly faithful but disguise model biases.

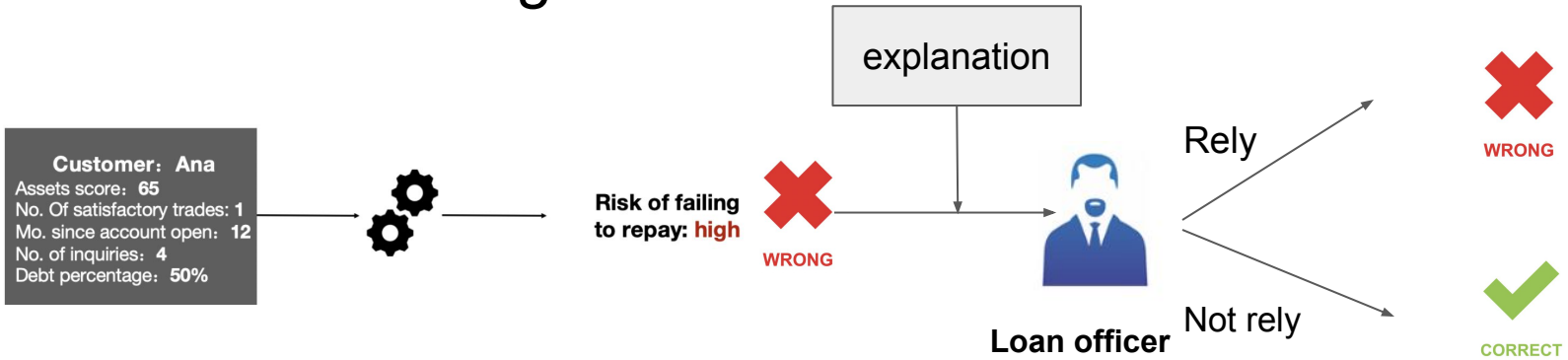
Aïvodji et al. Fairwashing: the risk of rationalization. *ICML 2019*
Anders et al. Fairwashing explanations with off-manifold detergent. *ICML 2020*

Open question: Explanation for fair outcome of human-AI joint decision-making



Fair outcomes for male v.s. Female customers?

Open question: Explanation for fair outcome of human-AI joint decision-making



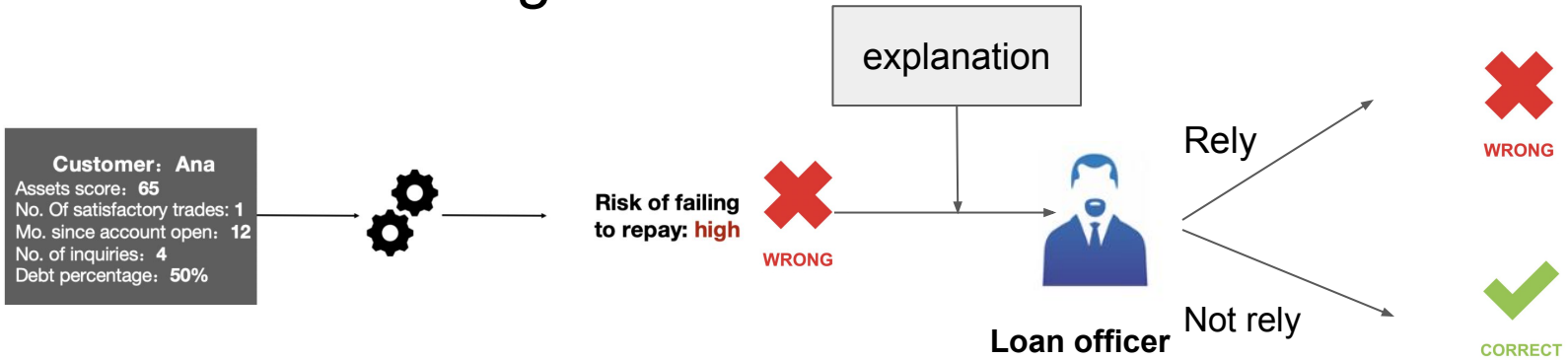
Fair outcomes for male v.s. Female customers?

How does adding explanations impact human reliance and joint outcomes?

How does perceived unfairness impact human reliance and joint outcomes?

What if human has their own biases?

Open question: Explanation for fair outcome of human-AI joint decision-making



Fair outcomes for male v.s. Female customers?

Some known empirical results:

- Presenting explanations can lead to higher (over) reliance
- Perception of AI unfairness leads to lower reliance, regardless of model correctness
- In some settings, AI support can exacerbate existing human biases

Schoeffer et al. (2022). On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *arXiv*
Bansal et al. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *CHI2021*
Zhang et al. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making.FAccT 2020
Green& Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *CSCW2021*

Open question: Explanation for fair recourse

•If **{debt percentage under 30%}** ,
you will no longer be predicted of high risk

Why was my loan application rejected? How can I improve in the future?

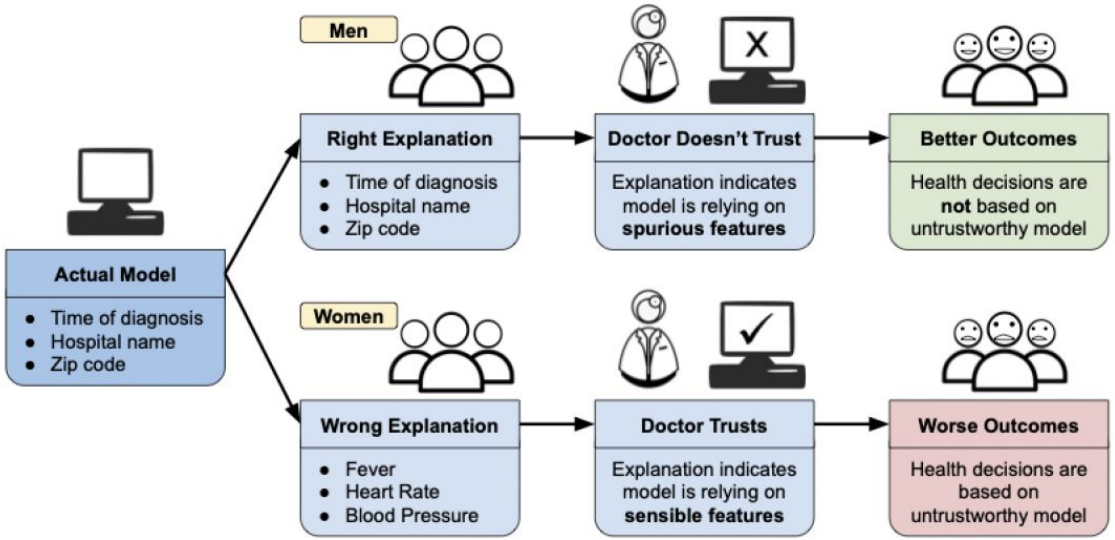
Bank customer

Recourse: an actionable set of changes for a person to obtain a desired outcome from a model

How to define actionability?
Is there actionability disparity for different groups?

Ustun et al. Actionable recourse in linear classification. *FAccT 2019*
Barocas et al. The hidden assumptions behind counterfactual explanations and principal reasons. FAccT2020
Karimi et al (2021). A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*.

Fairness of explanations: Disparities in explanation quality



		LIME	SHAP	SmoothGrad	IntGrad	VanillaGrad	Maple
German Credit	LR	0.424	**0.008**	1.0	**0.008**	1.0	0.905
Student Performance	LR	1.0	0.421	1.0	0.421	1.0	1.0
COMPAS	LR	0.841	0.151	1.0	0.131	1.0	0.401

(a) Ground truth — 2/18 significant

		LIME	SHAP	SmoothGrad	IntGrad	VanillaGrad	Maple
German Credit	LR	**0.032**	0.056	**0.032**	0.056	**0.032**	0.421
	NN	0.421	0.421	0.690	0.421	0.310	0.548
Student Performance	LR	0.691	0.548	0.690	0.549	0.690	0.690
	NN	0.056	**0.016**	0.056	**0.016**	0.056	**0.031**
COMPAS	LR	0.222	**0.008**	0.151	0.310	0.151	0.548
	NN	0.095	**0.016**	**0.008**	**0.016**	**0.016**	0.222

(b) Prediction Gap — 11/36 significant

		LIME	SHAP	SmoothGrad	IntGrad	VanillaGrad	Maple
German Credit	LR	0.100	**0.008**	1.0	**0.008**	0.690	0.690
	NN	0.421	0.222	**0.016**	**0.008**	**0.016**	0.675
Student Performance	LR	0.690	0.016	1.0	**0.008**	0.841	1.0
	NN	0.690	**0.016**	0.917	**0.008**	0.100	1.0
COMPAS	LR	**0.007**	**0.008**	1.0	**0.008**	0.158	0.690
	NN	0.310	0.151	1.0	0.222	0.310	0.548

(c) Sparsity — 11/36 significant

		LIME	SHAP	SmoothGrad	IntGrad	VanillaGrad	Maple
German Credit	LR	0.222	0.222	0.548	1.0	**0.016**	1.0
	NN	0.690	0.100	0.056	0.310	0.100	0.841
Student Performance	LR	0.690	0.690	0.548	0.690	0.310	1.0
	NN	0.310	0.310	0.690	0.056	0.056	0.841
COMPAS	LR	0.421	0.222	0.222	0.222	**0.008**	0.841
	NN	0.310	**0.008**	0.100	**0.008**	**0.008**	0.690

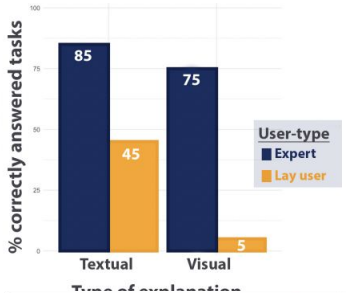
(d) Stability — 5/36 significant

		LIME	SHAP	SmoothGrad	IntGrad	VanillaGrad	Maple
German Credit	LR	**0.016**	1.0	1.0	1.0	1.0	0.690
	NN	0.548	1.0	1.0	0.841	1.0	1.0
Student Performance	LR	0.421	1.0	1.0	1.0	1.0	1.0
	NN	0.690	0.548	0.841	0.222	0.690	1.0
COMPAS	LR	0.310	0.672	1.0	1.0	1.0	0.841
	NN	0.151	1.0	1.0	0.690	1.0	0.548

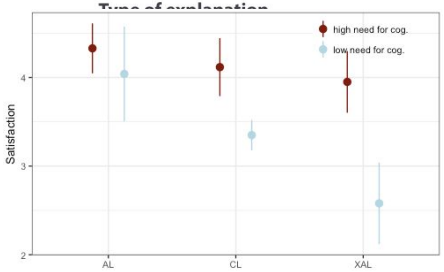
(e) Consistency — 1/36 significant

Dai, et al.. Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations. AIES 2022.
Balagopalan, at al. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. FAccT 2022.

Fairness of explanations: Disparity of experience



AI novices have less performance gain but more illusory satisfaction



Decrease task satisfaction for people with personality trait of low Need for Cognition

People may benefit less when they lack either the ability or motivation to cognitively engage with XAI

Szymanski et al. Visual, textual or hybrid: the effect of user expertise on different explanations. IUI 2021
Ghai et al. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. CSCW 2021
Liao & Varshney, (2021). Human-centered explainable ai (xai): From algorithms to user experiences.

Open questions: explainability and fairness

- How to ensure explanation faithfulness for fairness?
- How to ensure fair explainability?
- What are the implications for fair human-AI joint work and what are the best practices?
- How to cope with disparities created by explainability?