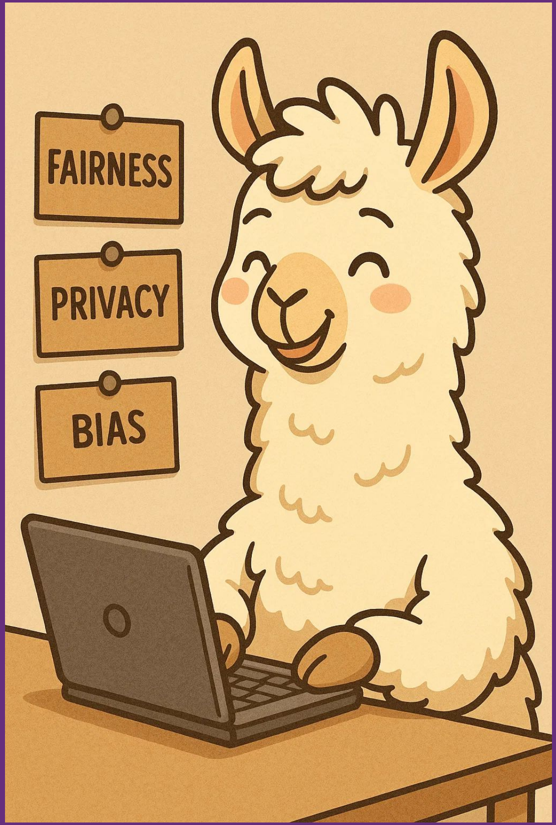


Image generated using Sora

Mila

27-05-2025

Technical Mitigation Strategies

Khaoula Chehbouni and Prakhar Ganesh

About me

Prakhar Ganesh (he/him)

PhD student in Computer Science
at McGill University / Mila

Research in Fairness and Privacy in AI
& Multiplicity in AI

Why Responsible AI?

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

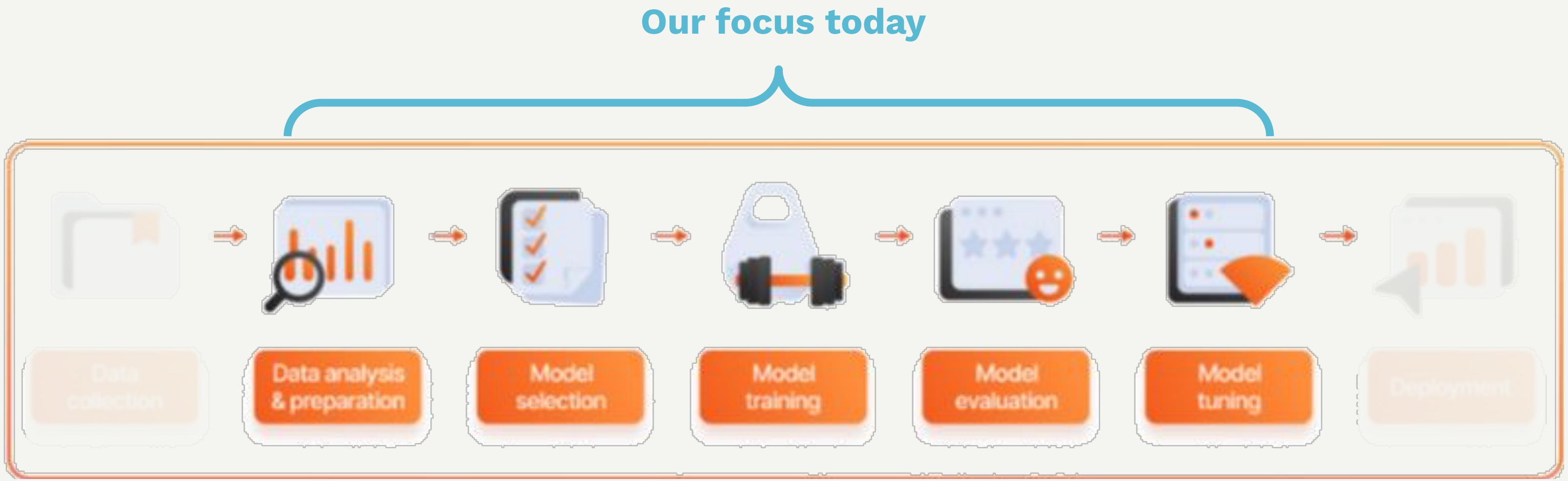
US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

https://medium.com/data-science/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

AI Pipelines

https://intelliarts.com/blog/the-role-of-pipelines-in-the-successful-machine-learning-project/

AI Pipelines



https://intelliarts.com/blog/the-role-of-pipelines-in-the-successful-machine-learning-project/

Outline of the Talk

Technical Mitigation Strategies in ML

- A Broader Perspective: Rashomon Effect and Multiplicity
- Fairness/Bias in AI
- Robustness in AI
- Privacy in AI
- Explainability/Interpretability in AI

Technical Mitigation Strategies in LLMs

- Safety Challenges in LLMs
- The Safety Mitigation Pipeline
- Are Safety Safeguards Robust?
- Explainability in LLMs

Technical Mitigation Strategies in ML

A Broader Perspective: ***Rashomon Effect and Multiplicity***

Rashomon Effect

Based on *Rashomon (1950)*
by Akira Kurosawa

Mila

Rashomon Effect

Rashomon effect is "*a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*"

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect

Rashomon effect is "*a combination of a **difference of perspective** and **equally plausible accounts**, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*"

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect

Rashomon effect is "*a combination of a difference of perspective and equally plausible accounts, with the* ***absence of evidence to elevate one above others****, with the* ***inability to disqualify any particular version*** *of the truth…*"

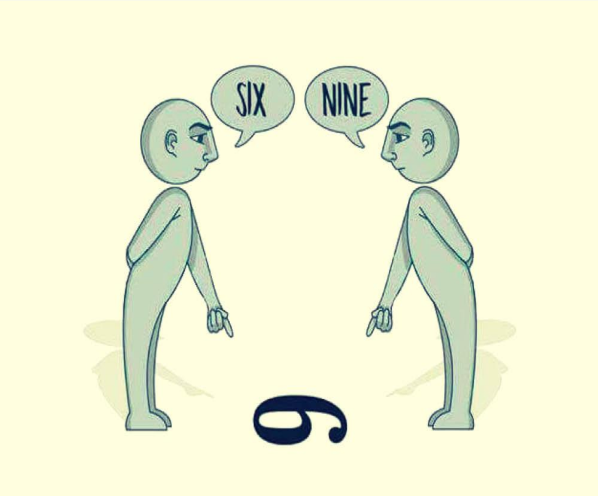
Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

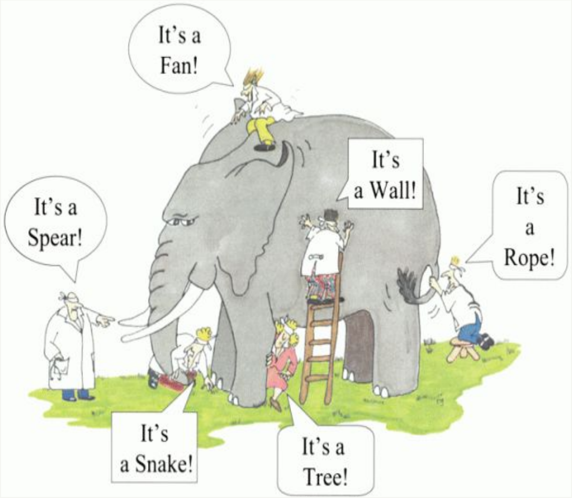
Rashomon Effect

Rashomon effect is “*a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others, with the inability to disqualify any particular version of the truth…*”

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect





https://classicallyeducated.wordpress.com/2020/05/19/ambrose-bierce-by-way-of-the-rashomon-effect/

https://medium.com/stotle-inc/rashomon-effect-lessons-for-building-effective-bi-dashboards-1b484b3137e9

Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect in AI

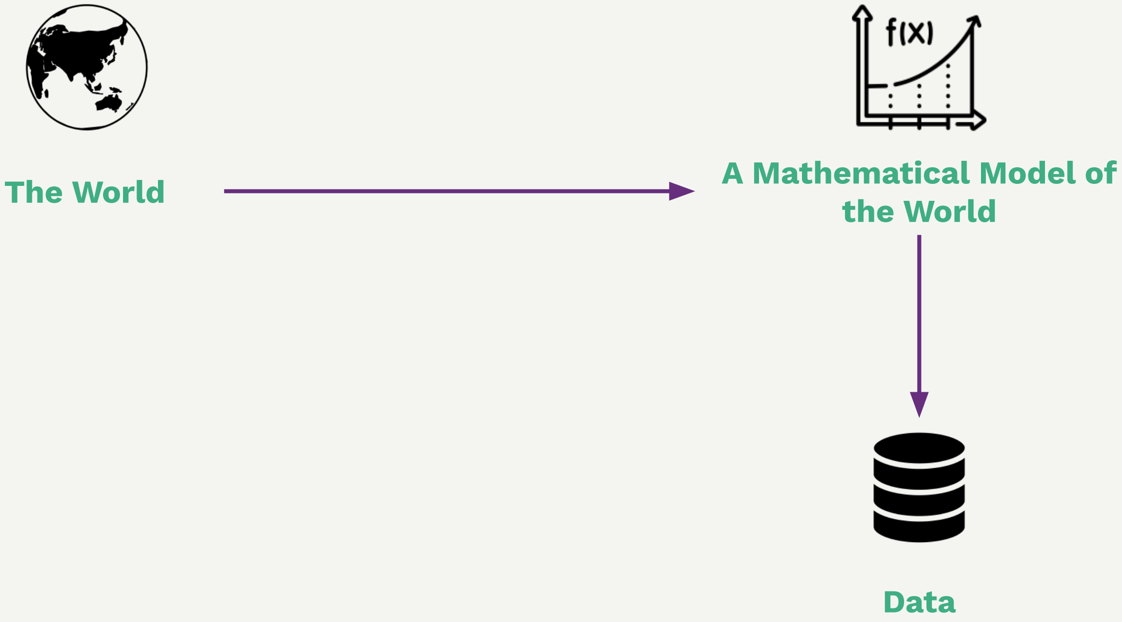
Rashomon Effect in AI

The World

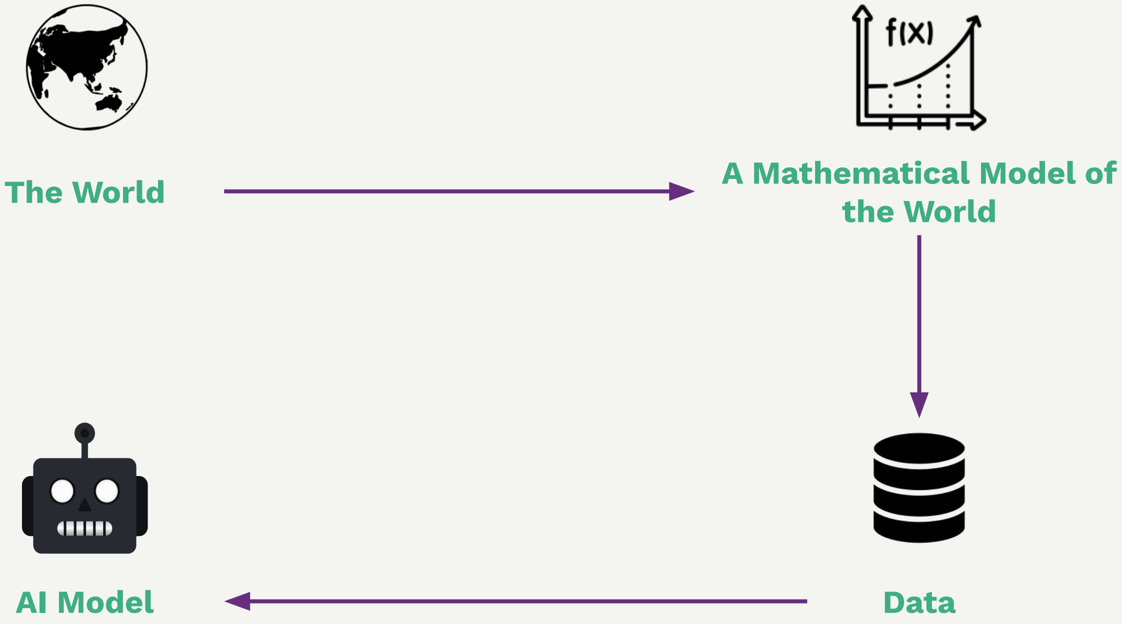
Rashomon Effect in AI

The World → A Mathematical Model of the World

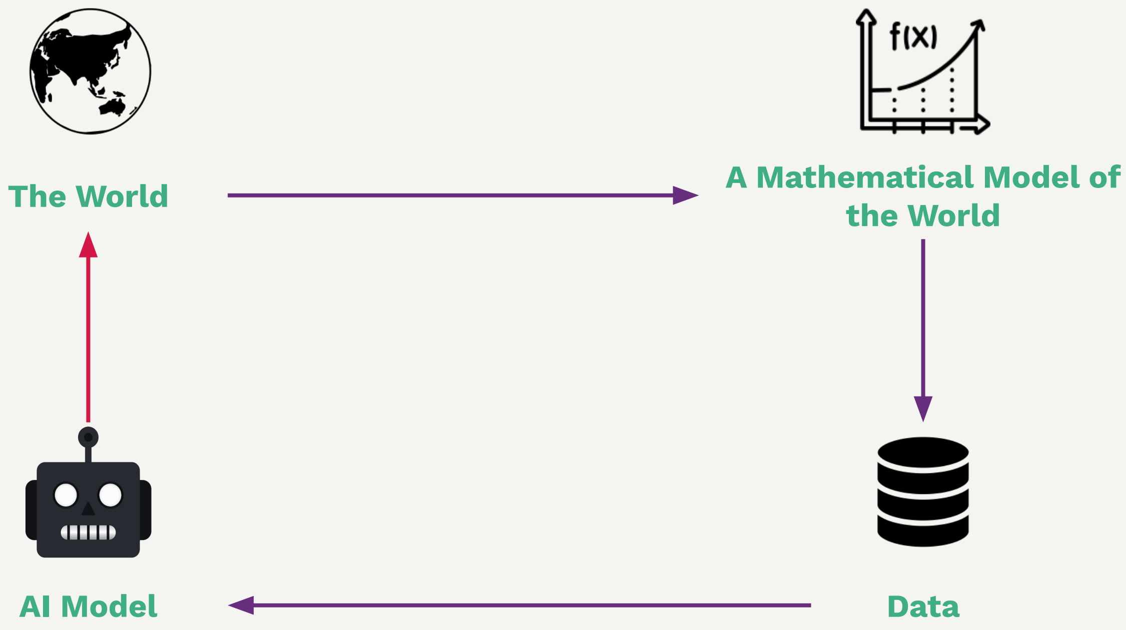
Rashomon Effect in AI



Rashomon Effect in AI



Rashomon Effect in AI



Rashomon Effect in AI

Loan Applications

The World

Rashomon Effect in AI

Loan Applications

The World

A Mathematical Model of the World

Chances of Repaying Loan depends on Annual Income and Credit Score

Rashomon Effect in AI

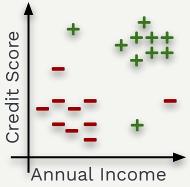
Loan Applications

The World

A Mathematical Model of the World

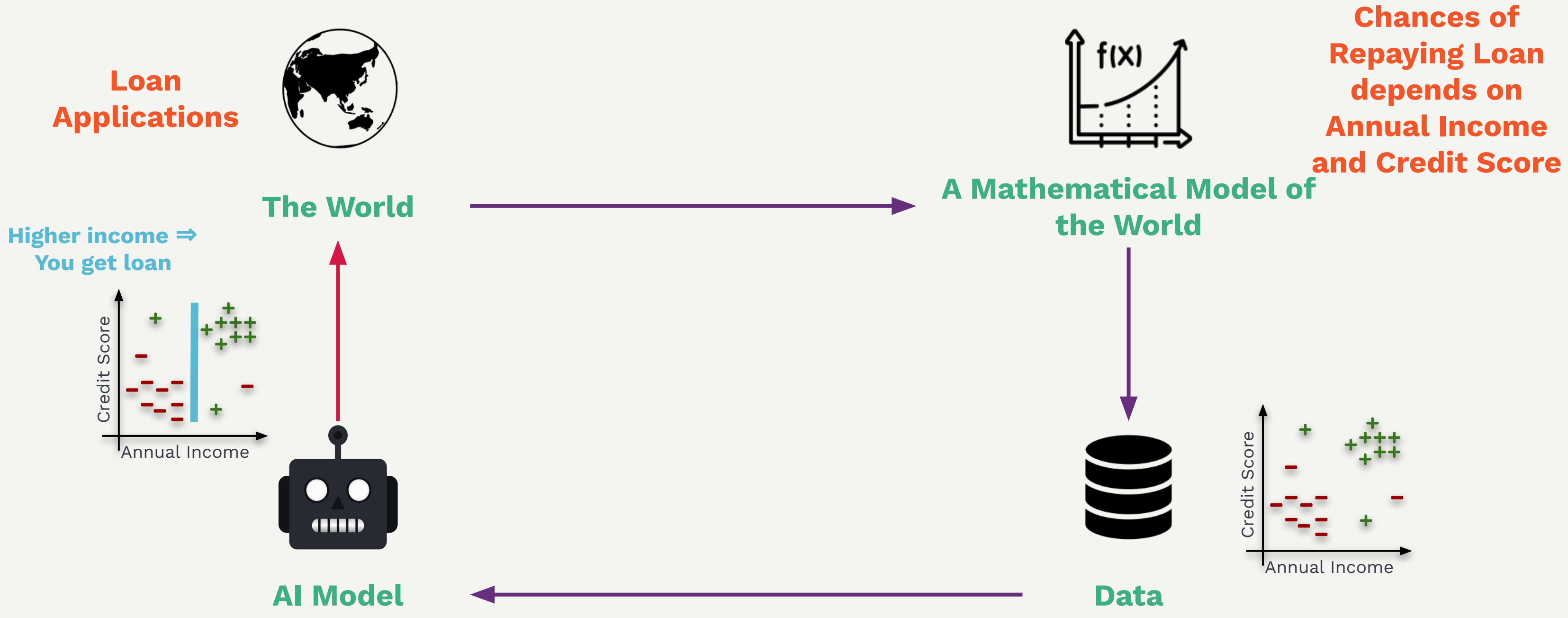
Chances of Repaying Loan depends on Annual Income and Credit Score

Data

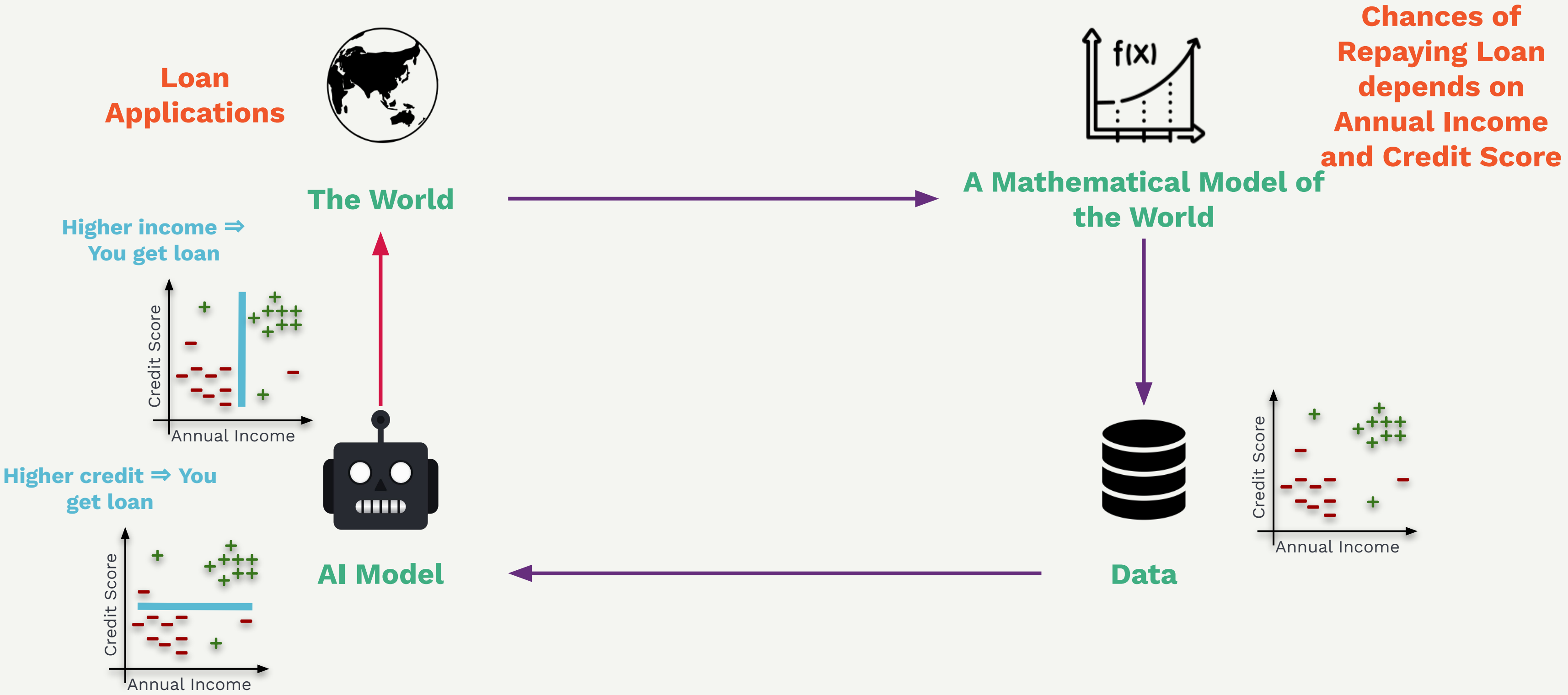


Mila

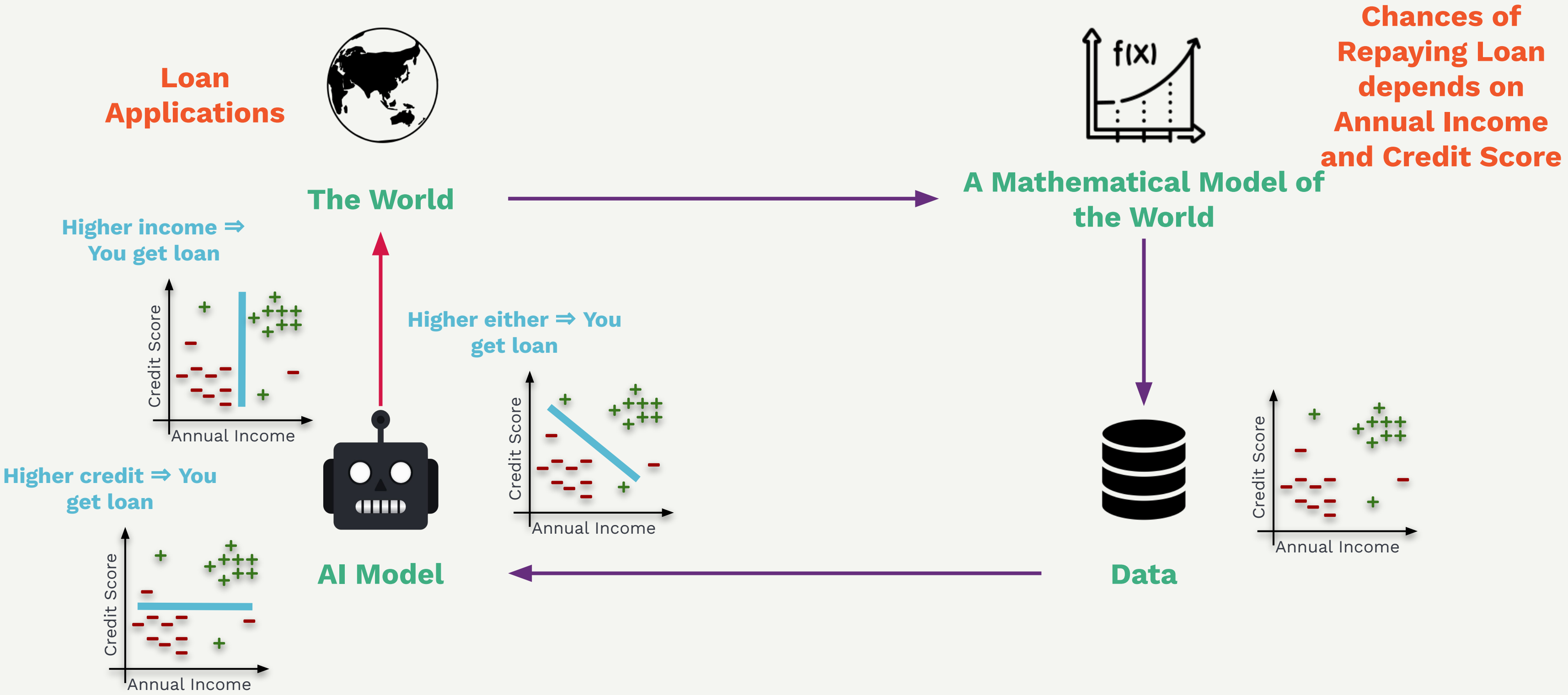
Rashomon Effect in AI



Rashomon Effect in AI



Rashomon Effect in AI



Rashomon Effect in AI

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

“What I call the Rashomon Effect is that there is often a multitude of different descriptions (equations f(x)) in a class of functions giving about the same minimum error rate.”

Breiman, L. (2001). *Statistical modeling: The two cultures*. Statistical science, 16(3), 199-231.

Multiplicity

There are many different models which can achieve the same error on the given data

Multiplicity

There are many different models which can achieve the same error on the given data

- Some of them might be fairer than others

Multiplicity

There are many different models which can achieve the same error on the given data

- Some of them might be fairer than others
- Some of them might be more robust than others

Multiplicity

There are many different models which can achieve the same error on the given data

- Some of them might be fairer than others
- Some of them might be more robust than others
- Some of them might be more interpretable than others

Multiplicity

There are many different models which can achieve the same error on the given data

- Some of them might be fairer than others
- Some of them might be more robust than others
- Some of them might be more interpretable than others
- Some of them might protect privacy better than others

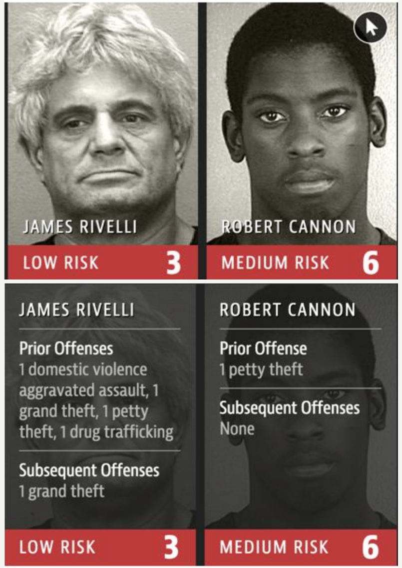
Multiplicity

There are many different models which can achieve the same error on the given data

- Some of them might be fairer than others
- Some of them might be more robust than others
- Some of them might be more interpretable than others
- Some of them might protect privacy better than others
- ...

Fairness/Bias in AI

Bias in AI



https://www.propublica.org/article/machine-bias-risk
-assessments-in-criminal-sentencing

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

http://gendershades.org/

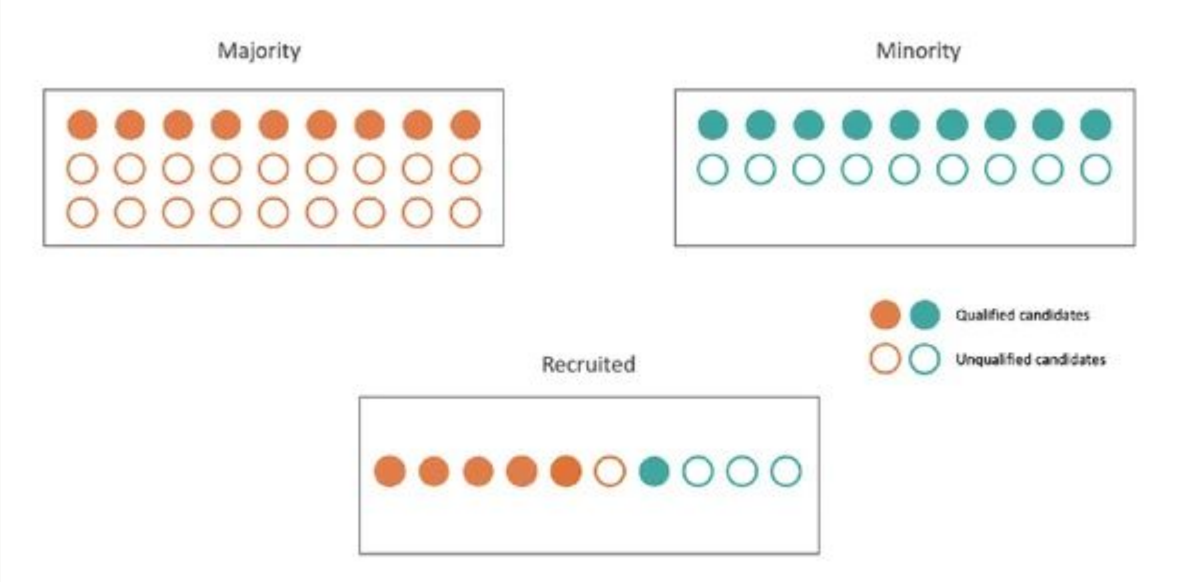
Mila

The Three Definitions of Fairness

Independence

60% of recruited individuals

The Three Definitions of Fairness

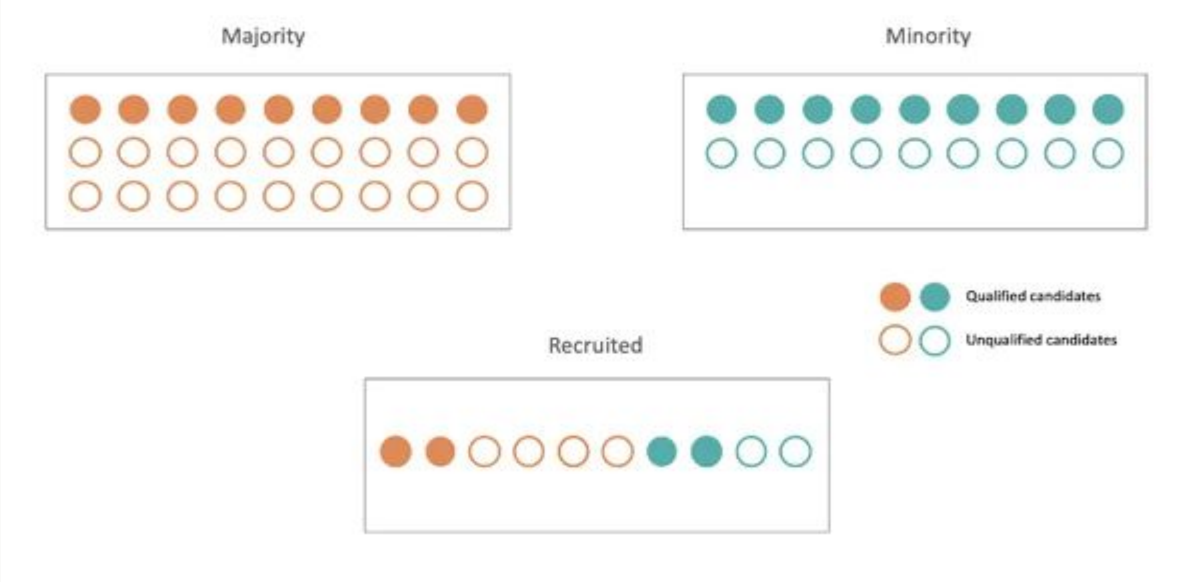


https://blog.dataiku.com/measuring-fairness-in-machine-learning-models

Separation

2 out of 9 qualified candidates were recruited

The Three Definitions of Fairness



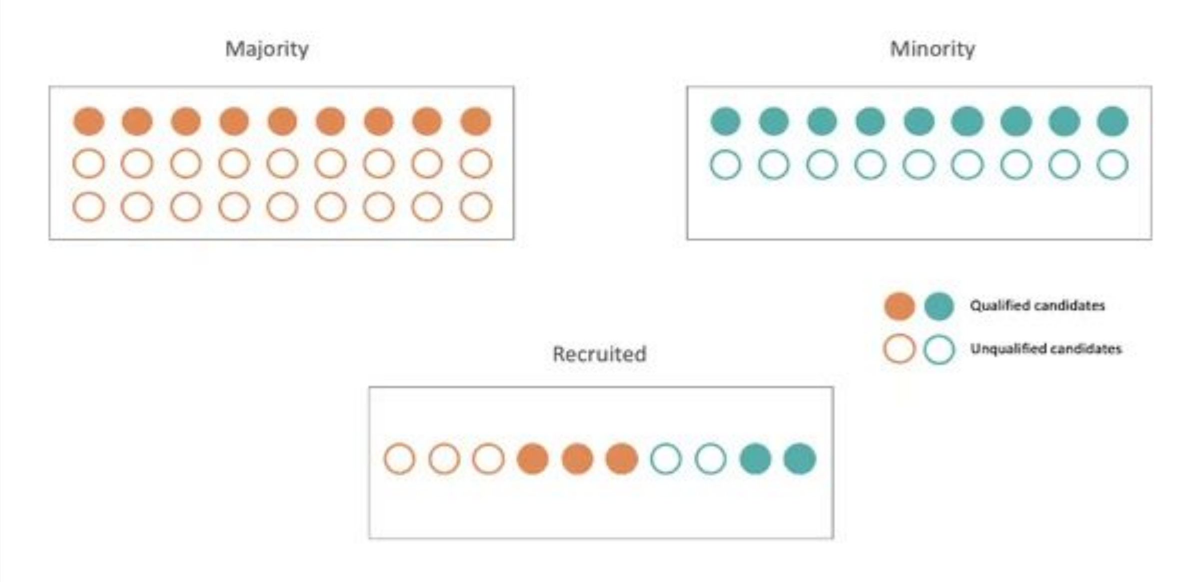
https://blog.dataiku.com/measuring-fairness-in-machine-learning-models

Mila

Out of all recruited, 50% are qualified

Sufficiency

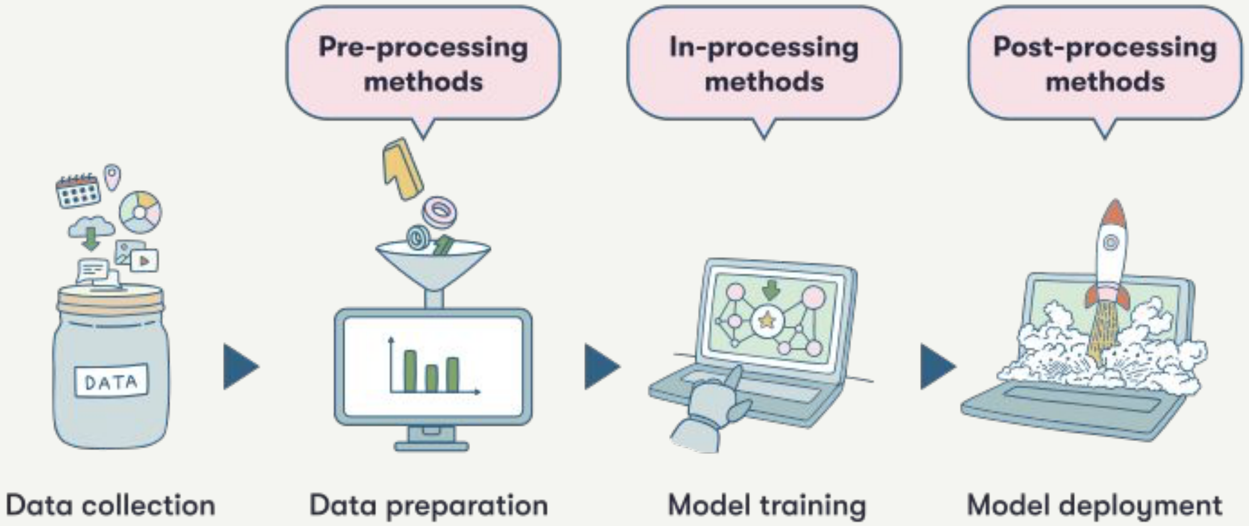
The Three Definitions of Fairness



https://blog.dataiku.com/measuring-fairness-in-machine-learning-models

Mila

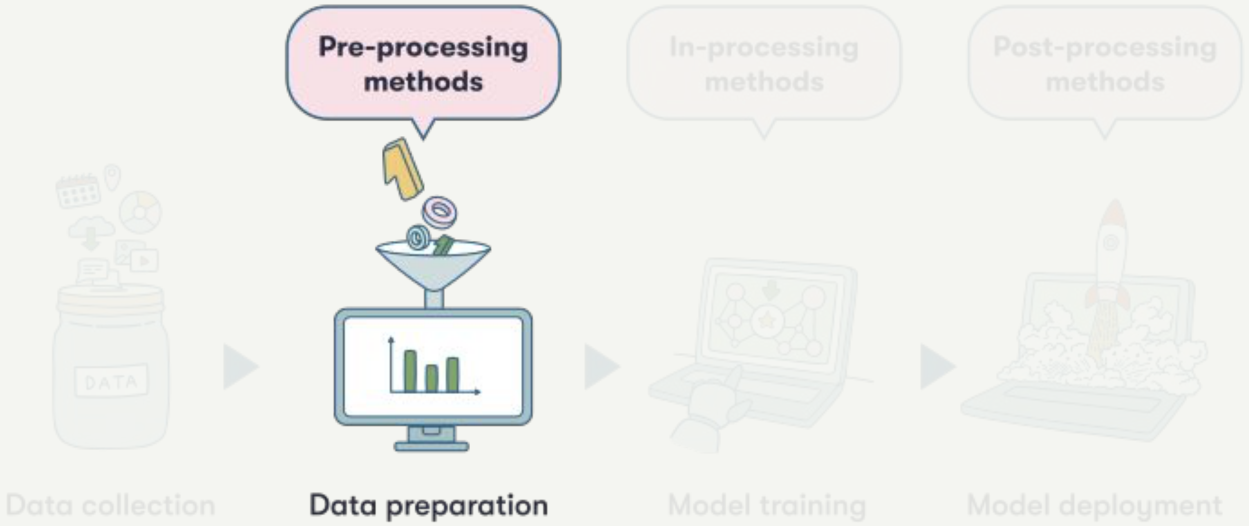
Bias Mitigation: Pre/In/Post-Processing



https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/fairness-and-accountability/detecting-and-mitigating-bias-and-unfairness/

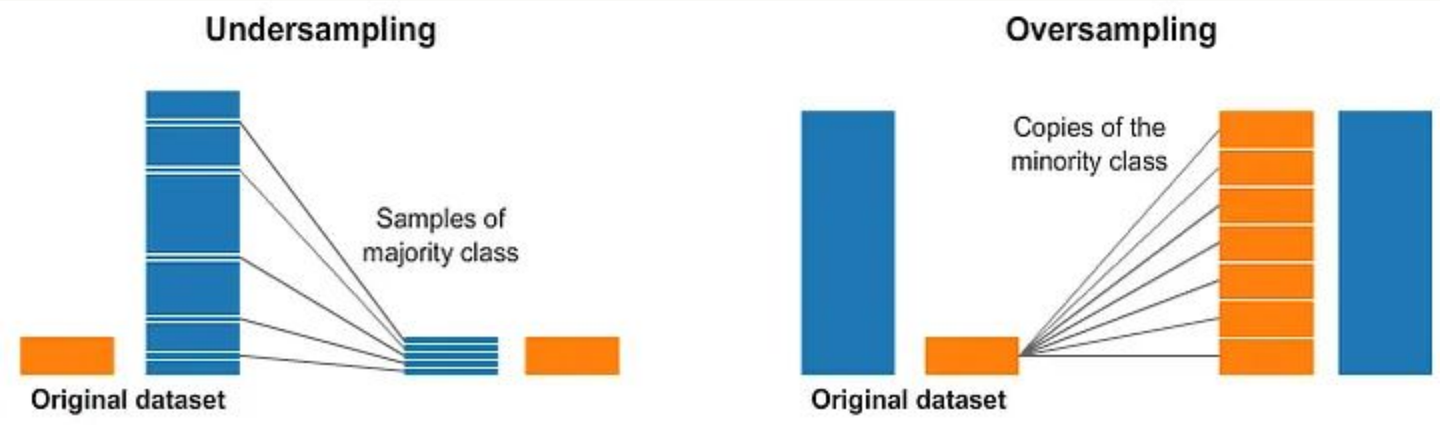
Balance the Data

Bias Mitigation: Pre-Processing



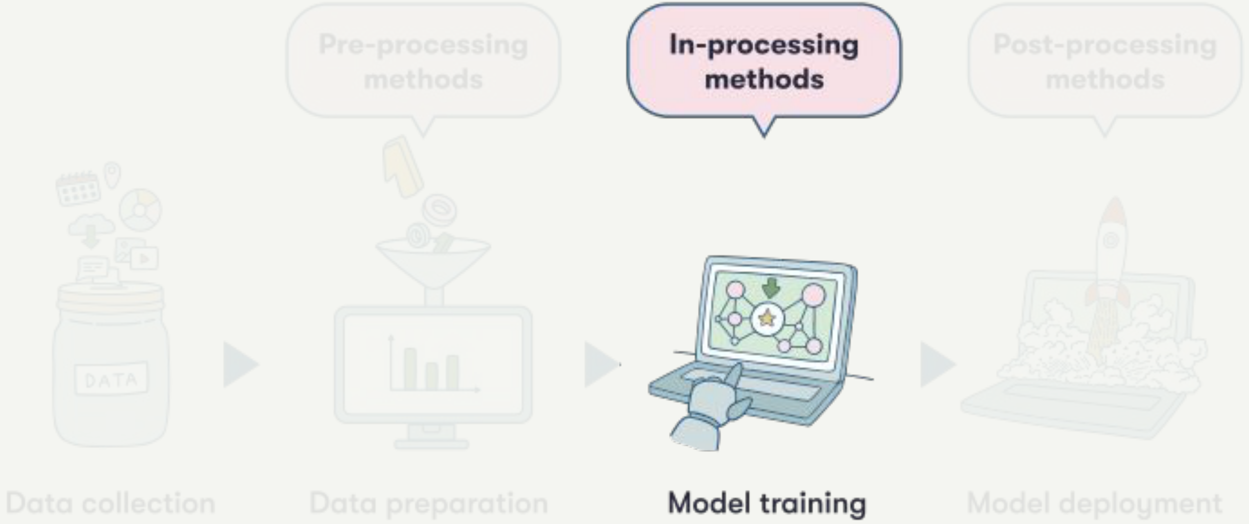
https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/fairness-and-accountability/detecting-and-mitigating-bias-and-unfairness/

Bias Mitigation: Pre-Processing



https://medium.com/@gaikwads070/oversampling-and-undersampling-are-techniques-used-in-data-preprocessing-31bc153fb673

Bias Mitigation: In-Processing



https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/fairness-and-accountability/detecting-and-mitigating-bias-and-unfairness/

Main Takeaway: Learn to be Fair

Bias Mitigation: In-Processing

Objective

Minimizing the loss to train the model on the train set:

$M^* = \text{argmin } Loss(M, D_{TRAIN})$

Constraints

subject to given fairness constraints F, ε:

$F(M, D_{TRAIN}) \leq \varepsilon$

Objective w/ Regularization

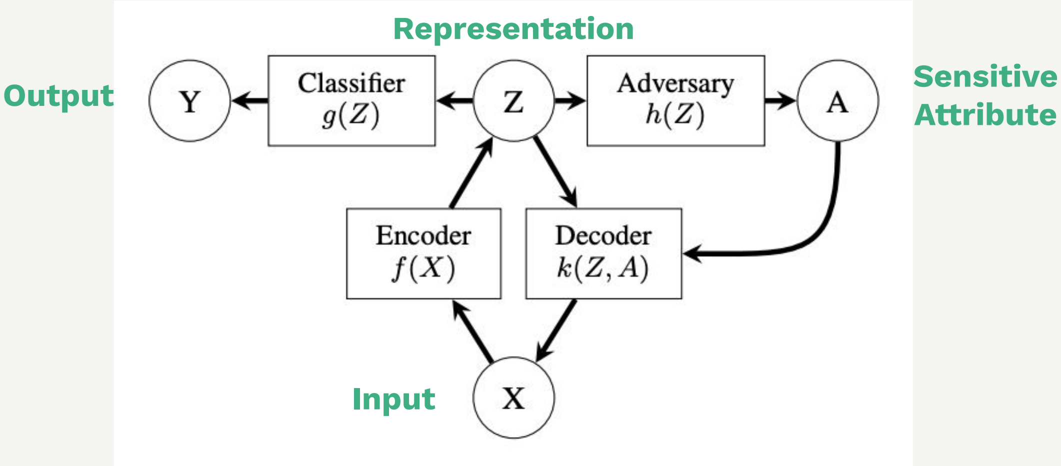
Minimizing the loss to train the model on the train set:

$M^* = \text{argmin } Loss(M, D_{TRAIN}) + R_F(M, D_{TRAIN})$

Ganesh, P., Taik, A. and Farnadi, G., 2025. The Curious Case of Arbitrariness in Machine Learning. arXiv preprint arXiv:2501.14959.

Main Takeaway: Learn to be Fair

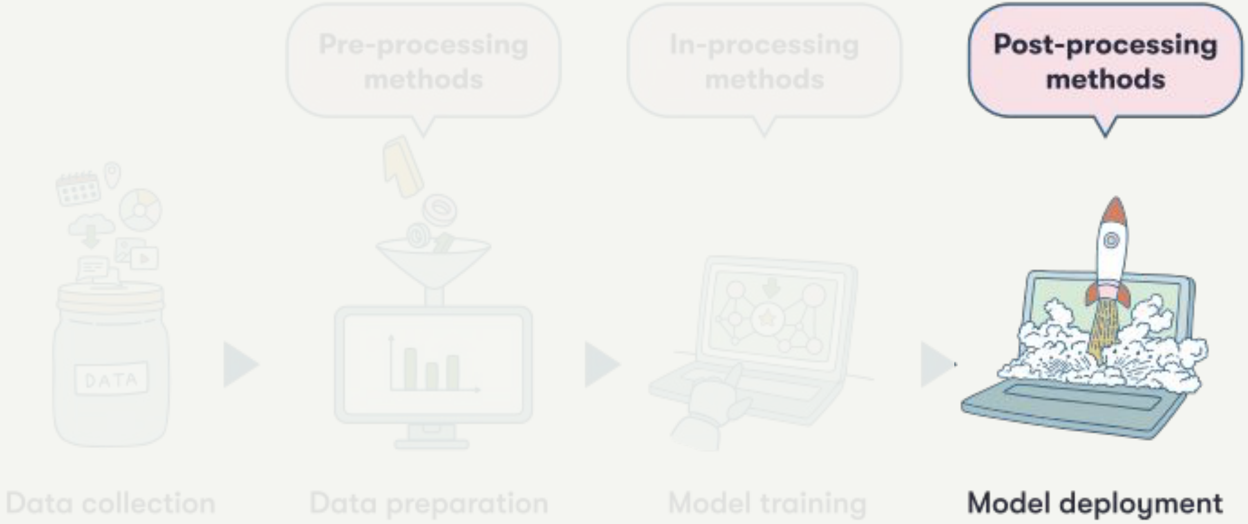
Bias Mitigation: In-Processing



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

Adjust Outputs when Cannot Retrain

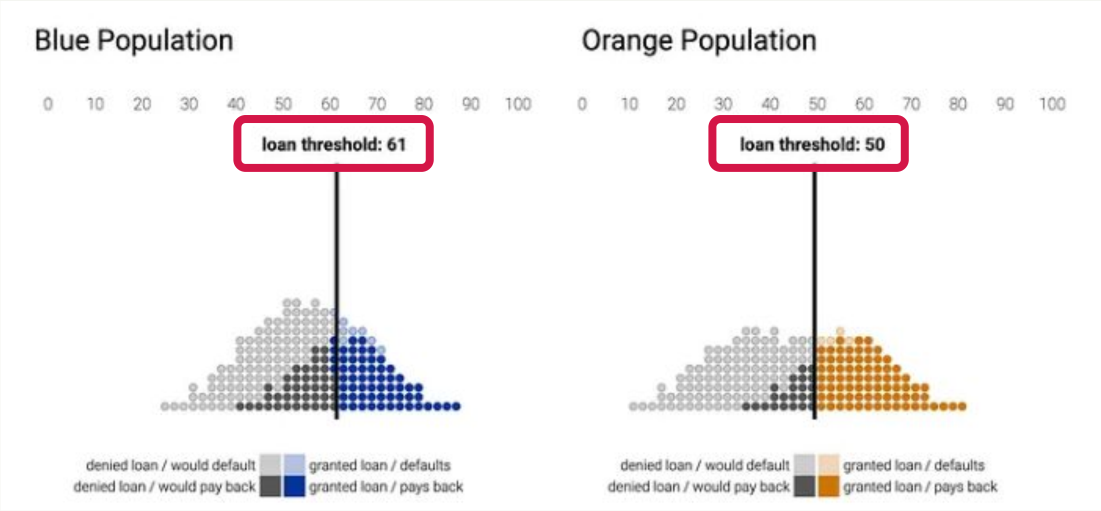
Bias Mitigation: Post-Processing



https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/fairness-and-accountability/detecting-and-mitigating-bias-and-unfairness/

Main Takeaway: Adjust Outputs when you Cannot Retrain

Bias Mitigation: Post-Processing



https://jonathan-hui.medium.com/address-ai-bias-with-fairness-criteria-tools-9af1ab8e4289

Discrimination Hacking or “D-Hacking”

Multiplicity means we can get,

- An unfair model
- Or a fair model

Discrimination Hacking or “D-Hacking”

Multiplicity means we can get,

- An unfair model
- Or a fair model
- **Or a model which is fair on some dataset but unfair when deployed!**

Black, E., Gillis, T. and Hall, Z.Y., 2024, June. D-hacking. In Proceedings of the 2024 ACM FAccT (pp. 602-615).

Discrimination Hacking or “D-Hacking”

Multiplicity means we can get,

- An unfair model
- Or a fair model
- **Or a model which is fair on some dataset but unfair when deployed!**

Reducing Regulatory Requirements to a Single Dataset/Evaluation can result in D-Hacking!

Black, E., Gillis, T. and Hall, Z.Y., 2024, June. D-hacking. In Proceedings of the 2024 ACM FAccT (pp. 602-615).

Mila

Robustness in AI

Two Types of Robustness

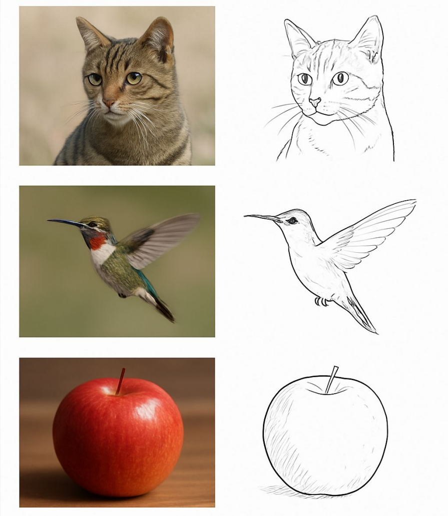
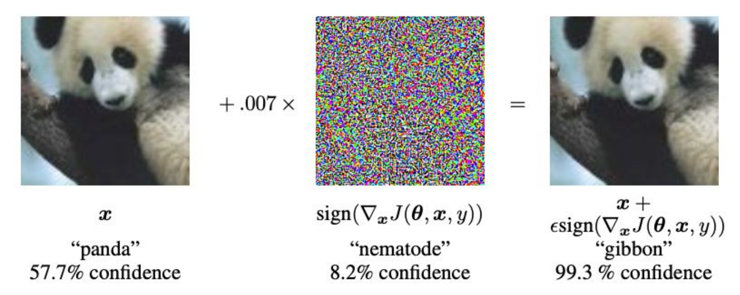


Image generated using Sora



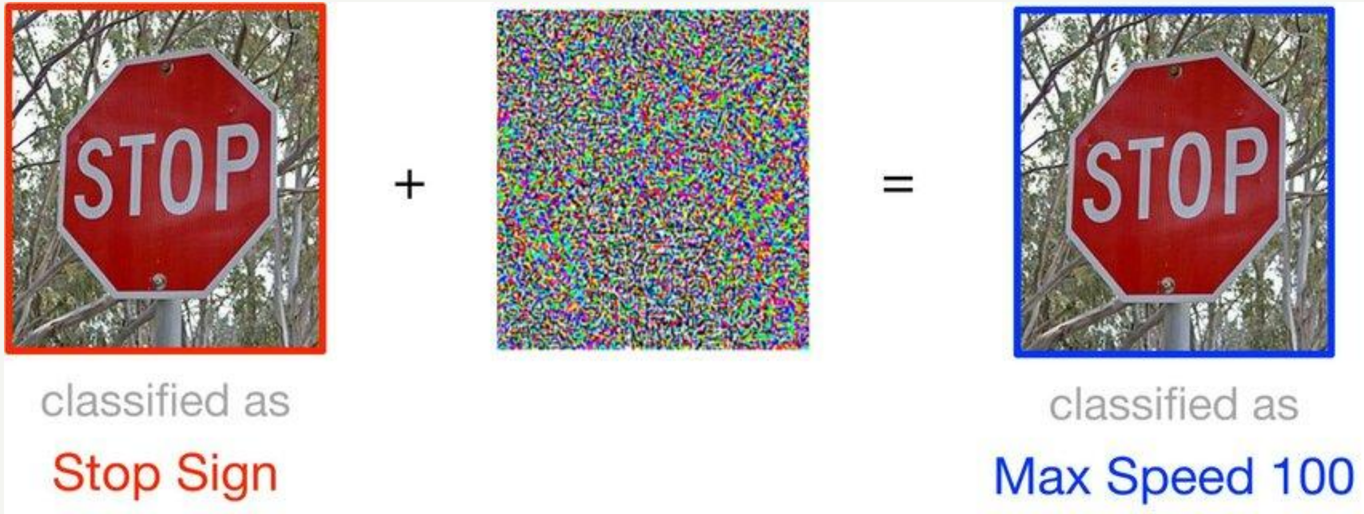
Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Adversarial Examples

AI models can be extremely brittle!

Adversarial Examples

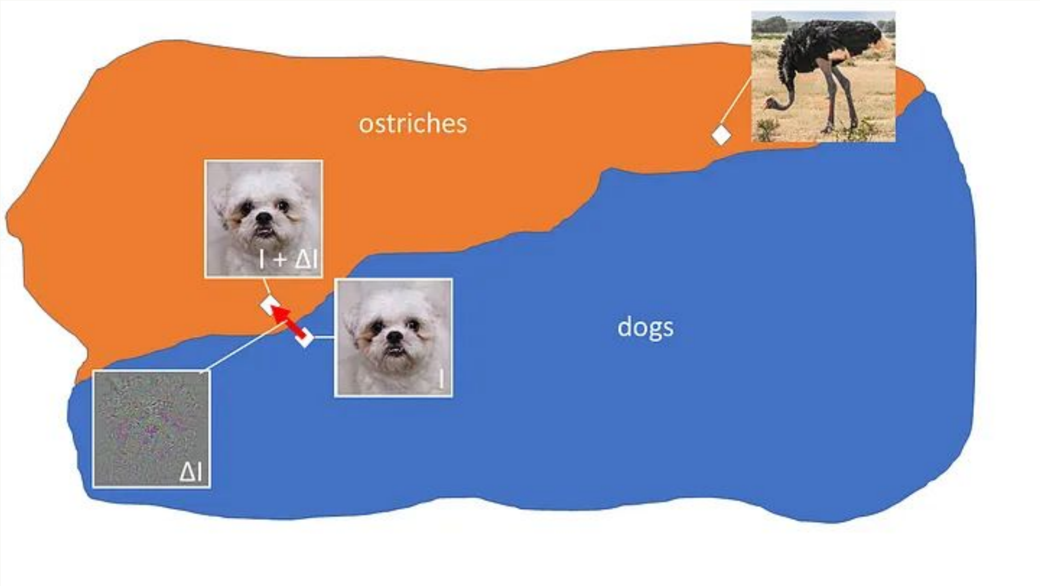
AI models can be extremely brittle!



Mila

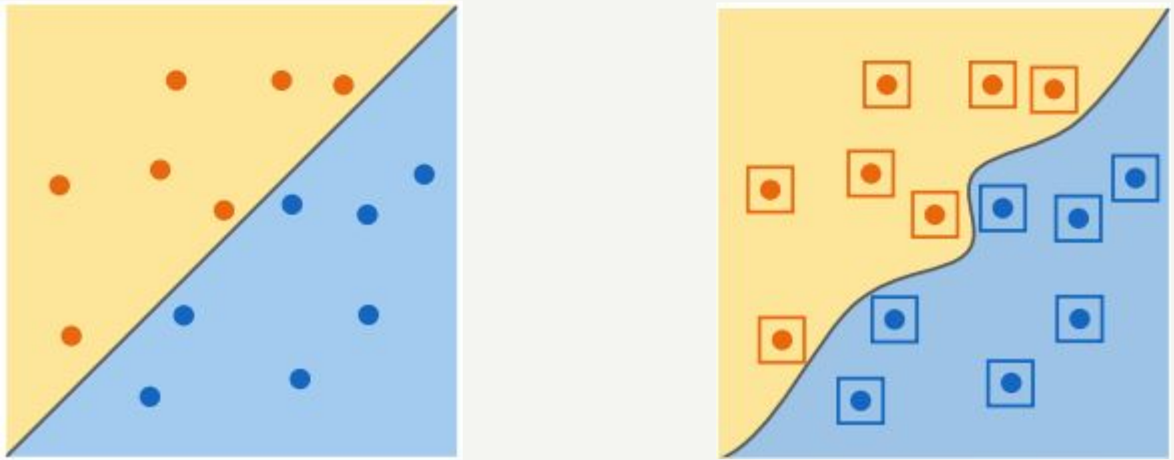
Adversarial Examples

AI models can be extremely brittle!



https://medium.com/fiveai/a-simple-but-pretty-good-understanding-of-adversarial-examples-8ab0cb7d62b0

Adversarial Training



Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H. and Zhu, J., 2022, Exploring Memorization in Adversarial Training. In ICLR.

Robustness Guarantees

No one has control over the actual input to the model – can provide robustness guarantees!!

Original Input → **Random Noise** → Noisy Input → **Our Model** →

Privacy in AI

What does privacy mean to you?

Control what information about you is collected, used, or shared

Protection of people's physical selves against invasive procedures

Protection against unwarranted intrusion

Protection of personal communication

What does privacy mean to you?

Control what information about you is collected, used, or shared

The Right to be Left Alone

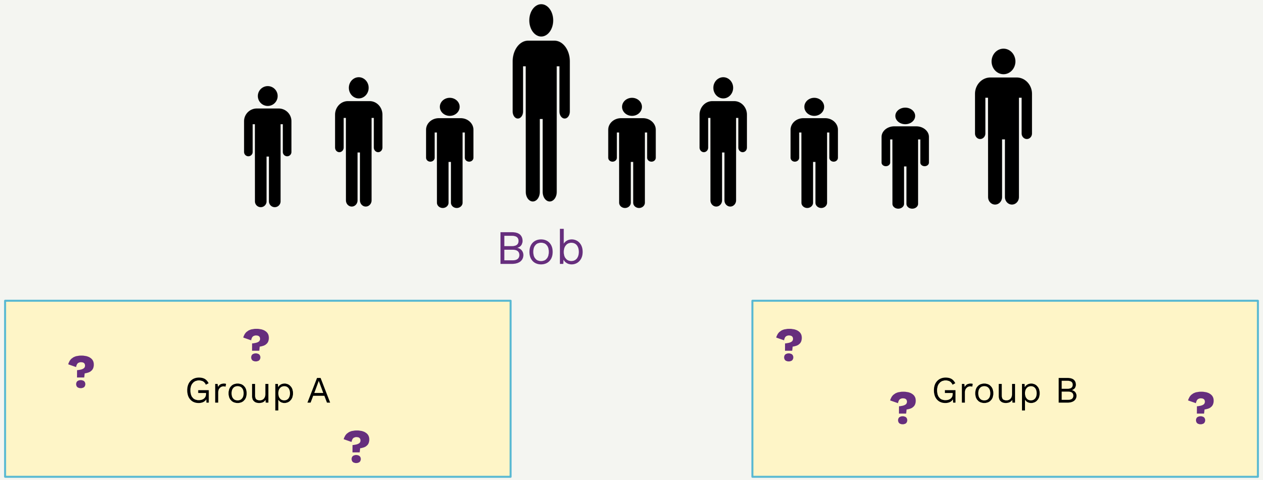
Protection against unwarranted intrusion

Protection of personal communication

Privacy as Membership

Privacy as Membership

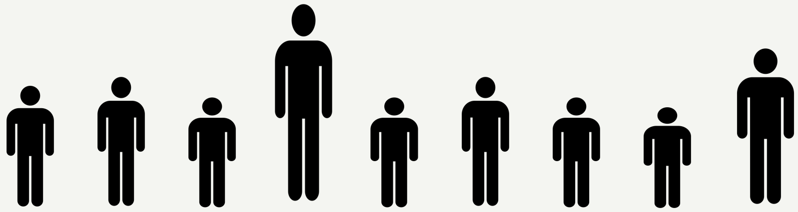
Consider the following example



Mila

Privacy as Membership

Consider the following example



Bob

Group A

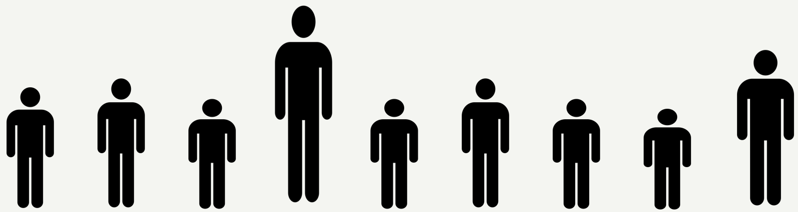
Average Height = 5’10

Group B

Average Height = 5’6

Privacy as Membership

Consider the following example



Bob

? ? Group A ?

Average Height = 5'10

? ? Group B ?

Average Height = 5'6

Can you guess which group Bob belongs to?

The Promise of Anonymization

Name	Zipcode	Age	Gender	Genetic Marker for Cancer
Alice	117068	27	F	Yes
Bob	167056	64	M	No
Charlie	118567	32	M	No
David	191504	81	M	No

Insurance companies: Who has genetic markers for cancer? I would like to raise their premium and get more money!

Alice has genetic markers for cancer.

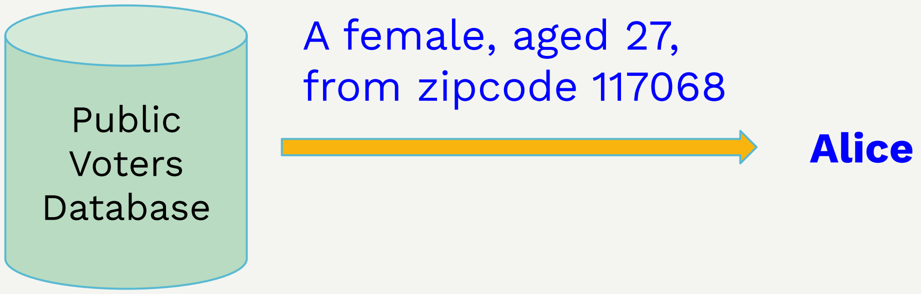
The Promise of Anonymization

Name	Zipcode	Age	Gender	Genetic Marker for Cancer
	117068	27	F	Yes
	167056	64	M	No
	118567	32	M	No
	191504	81	M	No

Insurance companies: Who has genetic markers for cancer? I would like to raise their premium and get more money!

A female, aged 27, from zipcode 117068, has genetic markers for cancer.

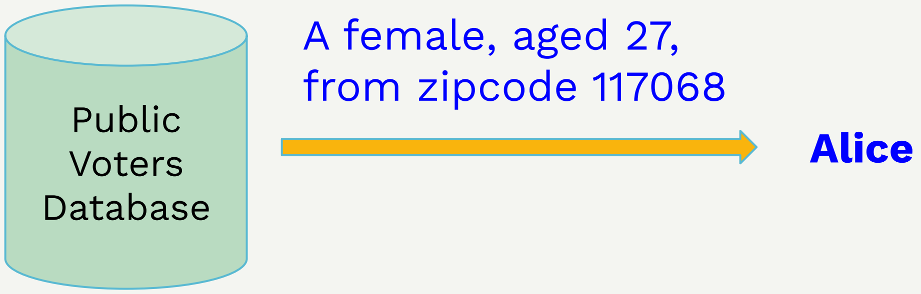
The Promise of Anonymization



Insurance companies: Who has genetic markers for cancer? I would like to raise their premium and get more money!

A female, aged 27, from zipcode 117068, has genetic markers for cancer.

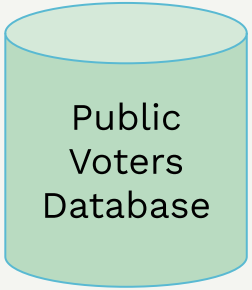
The Promise of Anonymization



Insurance companies: Who has genetic markers for cancer? I would like to raise their premium and get more money!

Alice has genetic markers for cancer.

The Promise of Anonymization



A female, aged 27,
from zipcode 117068

Alice

Linkage Attacks

Insurance companies: Who has genetic markers for cancer? I would like to raise their premium and get more money!

Alice has genetic markers for cancer.

Mila

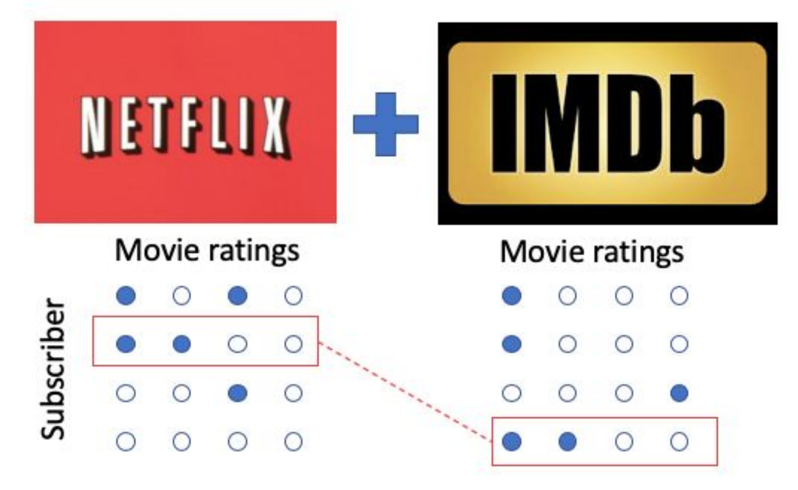
The Promise of Anonymization

Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

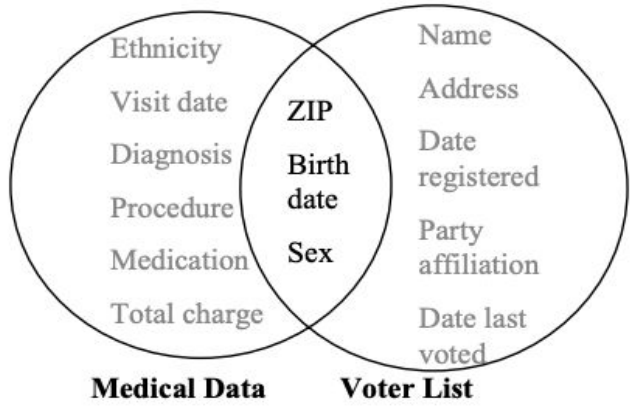
February 5, 2008



The Promise of Anonymization

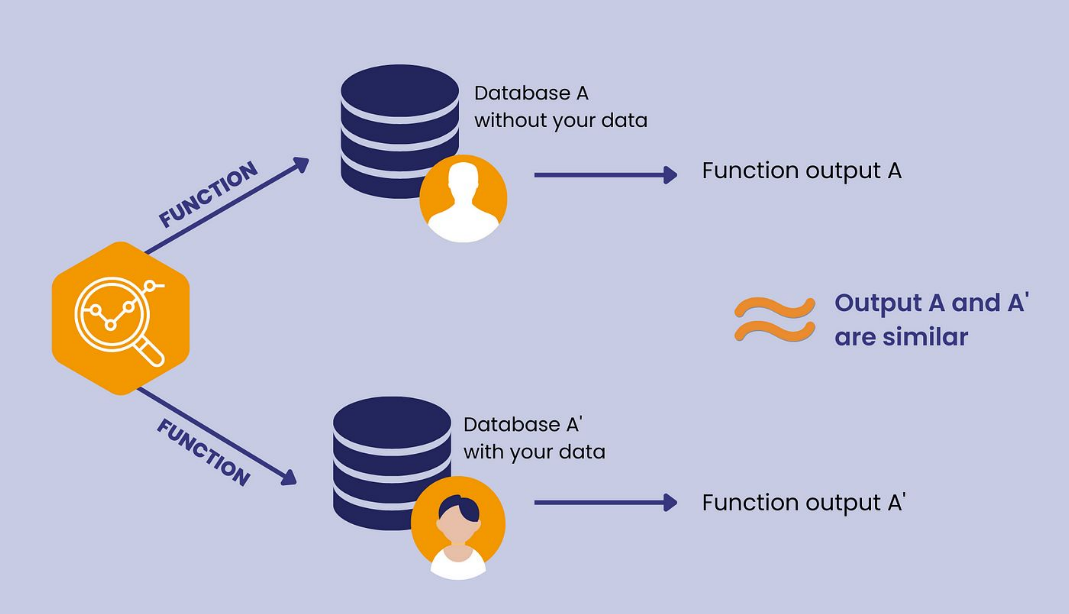
Simple Demographics Often Identify People Uniquely

Latanya Sweeney
Carnegie Mellon University
latanya@andrew.cmu.edu



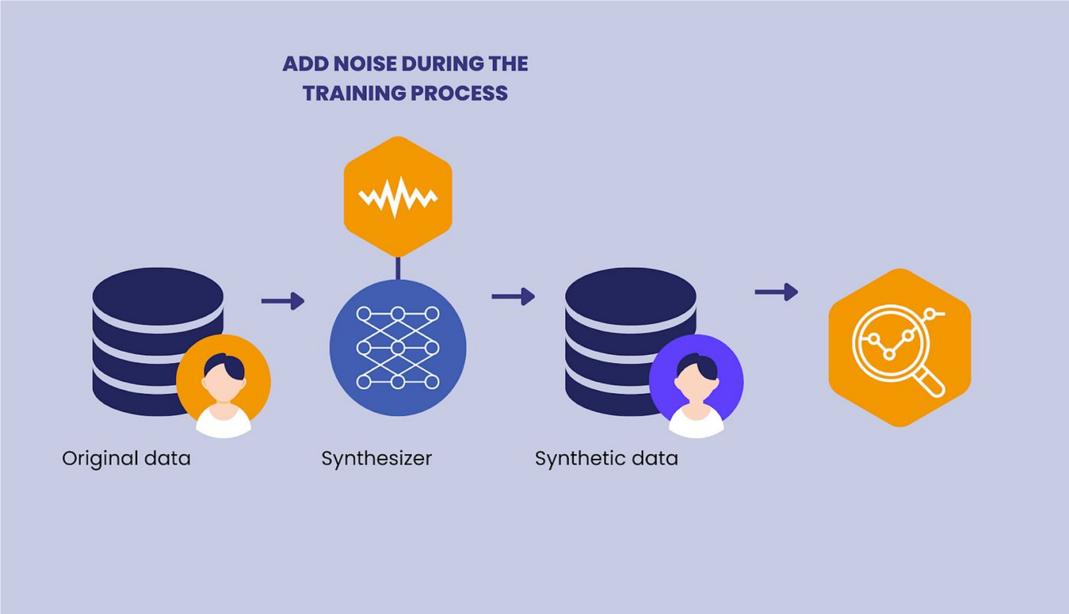
Mila

Differential Privacy



https://medium.com/statice/what-is-differential-privacy-definition-mechanisms-and-examples-7855bdc318d1

Adding Noise



https://medium.com/statice/what-is-differential-privacy-definition-mechanisms-and-examples-7855bdc318d1

Other Concepts in Privacy

Other Concepts in Privacy

Federated Learning: *multiple entities coming together to collaboratively train models while ensuring that their data remains decentralized.*

Other Concepts in Privacy

Federated Learning: *multiple entities coming together to collaboratively train models while ensuring that their data remains decentralized.*

Homomorphic Encryption: performing *complex mathematical operations on encrypted data without compromising the encryption.*

Other Concepts in Privacy

Federated Learning: *multiple entities coming together to collaboratively train models while ensuring that their data remains decentralized.*

Homomorphic Encryption: performing *complex mathematical operations on encrypted data without compromising the encryption.*

Privacy by Design Principles: *proactively embedding privacy in ML systems, to anticipate and prevent privacy invasive events before they occur.*

Other Concepts in Privacy

Federated Learning: *multiple entities coming together to collaboratively train models while ensuring that their data remains decentralized.*

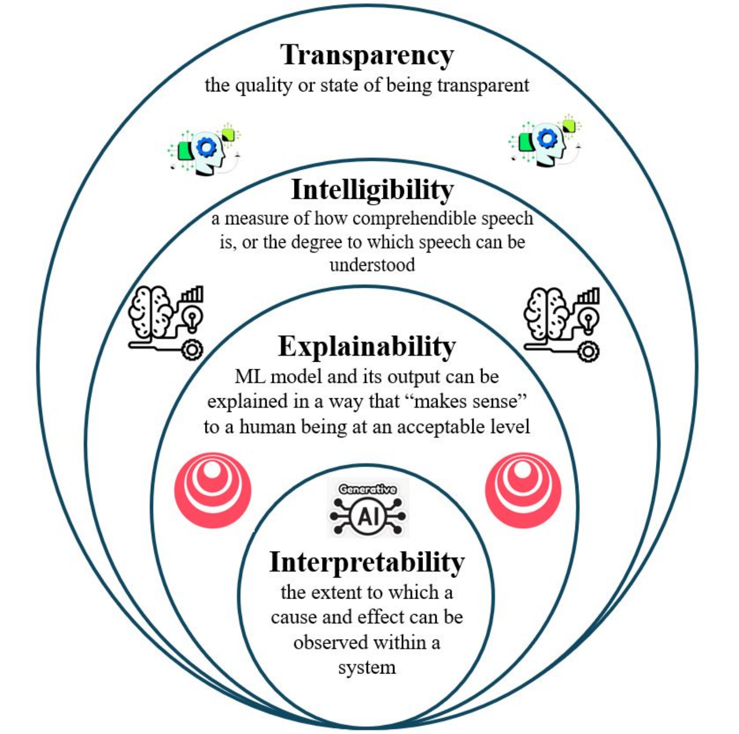
Homomorphic Encryption: performing *complex mathematical operations on encrypted data without compromising the encryption.*

Privacy by Design Principles: *proactively embedding privacy in ML systems, to anticipate and prevent privacy invasive events before they occur.*

Secure Multi-Party Computations, Zero Knowledge Proofs, ...

Explainability/Interpretability in AI

On The Road to Transparency



Shafik, W., Hidayatullah, A.F., Kalinaki, K., Gul, H., Zakari, R.Y. and Tufail, A., 2024. A Systematic Literature Review on Transparencyand Interpretability of AI models in Healthcare: Taxonomies, Tools, Techniques, Datasets, OpenResearch Challenges, and Future Trends.

Interpretability by Design

Interpretability by Design

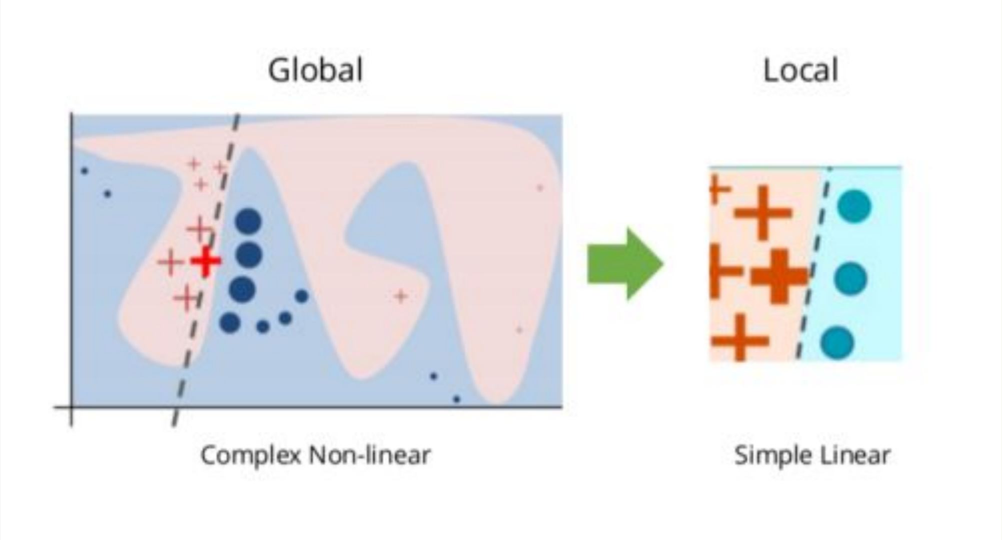
“Rashomon sets constructed with noisy data tend to contain simpler models than corresponding Rashomon sets with non noisy data”

The more inherently noisy a task is, the easier it is to find a good and interpretable model!!

Boner, Z., Chen, H., Semenova, L., Parr, R. and Rudin, C., 2024. Using Noise to Infer Aspects of Simplicity Without Learning. Advances in Neural Information Processing Systems, 37, pp.131824-131858.

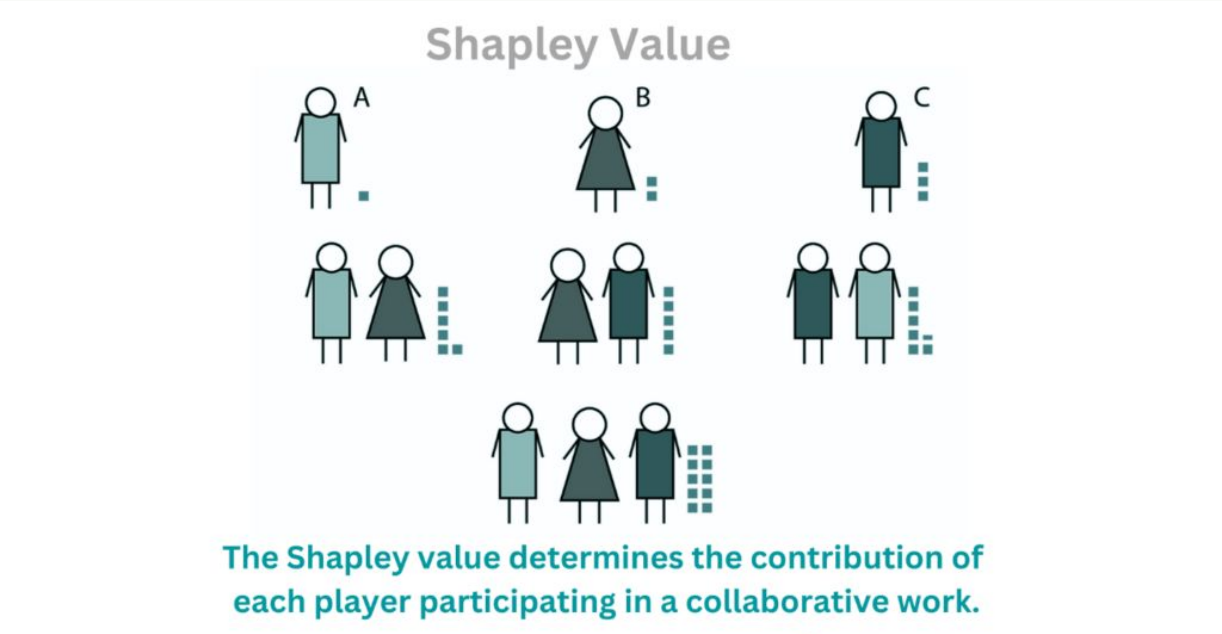
Explaining Complex Models

Explaining Complex Models: LIME



Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Explaining Complex Models: SHAP



https://www.wallstreetmojo.com/shapley-value/

X-Hacking and Fairwashing

Explanations are approximations of how the decisions are being made.

Multiplicity means we can get a model,

- Whose explanations are acceptable
- **But the actual prediction mechanism is not!**

Sharma, R., Redyuk, S., Mukherjee, S., Sipka, A., Vollmer, S. and Selby, D., 2024. X Hacking: The Threat of Misguided AutoML. arXiv preprint arXiv:2401.08513.
Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S. and Tapp, A., 2019, May. Fairwashing: the risk of rationalization. In International Conference on Machine Learning (pp. 161-170). PMLR.
Shahin Shamsabadi, A., Yaghini, M., Dullerud, N., Wyllie, S., Aïvodji, U., Alaagib, A., Gambs, S. and Papernot, N., 2022. Washing the unwashable: On the (im) possibility of fairwashing detection. Advances in Neural Information Processing Systems, 35, pp.14170-14182.