RAISE
RESPONSIBLE AI SYSTEMS & EXPERIENCES

Responsible AI / Trustworthy ML

An Overview

Sahil Verma

Broad wings of Trustworthy ML

- Fairness
- Explainability / Interpretability
- Robustness
- Privacy-preservation
- Model-Drift

Fairness

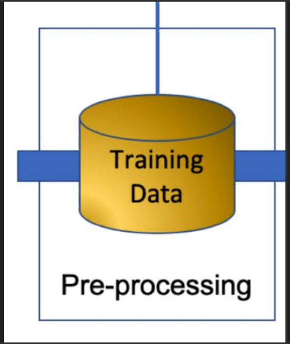
This is the subfield dealing with the problems related to defining unfairness, detecting it, and developing algorithms to counteract it.

There are more than 20 definitions of unfairness [1], and researchers choose a subset of these (mostly one) and develop a counteracting algorithm for that.

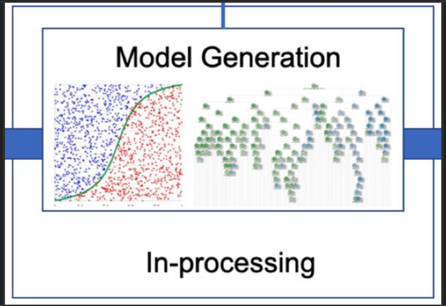
- Demographic Parity: Equal probability of getting the favorable outcome for each group.
- Equal Opportunity: Given equal qualification, each group should have equal chance of getting the favorable outcome.
- Individual Fairness: Similar individuals should get the same outcome.

[1] Fairness Definitions Explained. S. Verma and J. Rubin. FairWare, 2018.

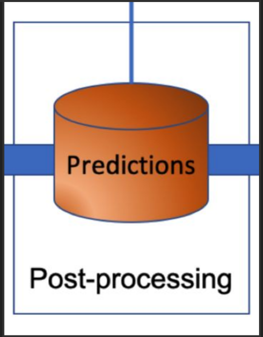
Unfairness Counteracting Interventions



Unfairness Counteracting Interventions

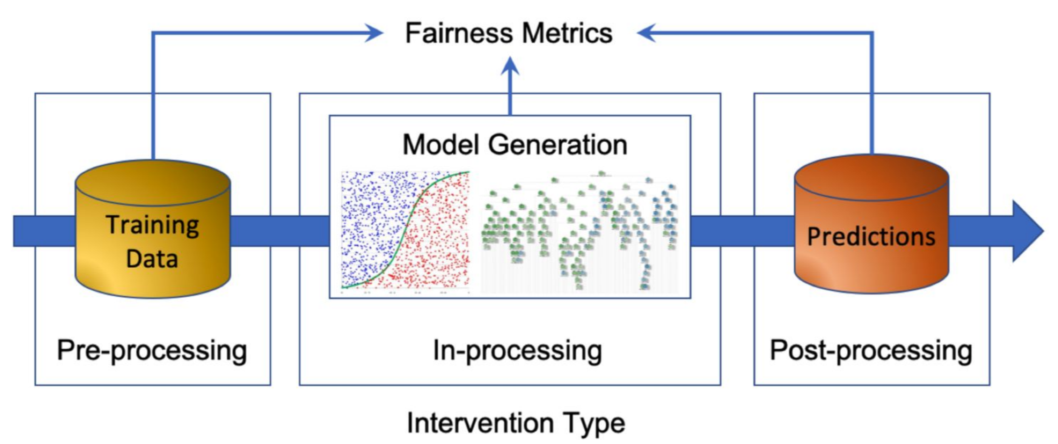


Unfairness Counteracting Interventions



Unfairness Counteracting Interventions

Each of these intervention methods prioritize a specific definition of unfairness they defend against.



Picture Credit: Fairness in Machine Learning: A Survey. S. Caton C. Haas

Open problems in Fairness

- Measuring statistical metrics of fairness on datapoints without labels.
- Defining a usage similarity metric for individual fairness.
- Extending the fairness metrics to multi-class classification scenarios.
- Establishing advantages and disadvantages comparisons across intervention algorithms (along ML pipeline and fairness definition dimensions).
- Demonstrating large-scale impact of deployment of intervention algorithms.

Resources for Fairness Enthusiasts

- Fairness Definitions Explained. S. Verma and J. Rubin. FairWare’ 18.
- Fairness-Aware Machine Learning. J. Dunkelau and M. Leuschel. ArXiv ‘20.
- Fairness through Awareness. C. Dwork et al. ITCS ‘12.
- Certifying and Removing Disparate Impact. M. Feldman et al. KDD’ 15.
- On the (im)possibility of fairness. S. Friedler et al. ArXiv’ 16.
- CS 294: Fairness in Machine Learning. Moritz Hardt. UCB course.

Explainability / Interpretability

Interpretability is the degree to which a human can understand the cause of a decision

Interpretability is the degree to which a human can consistently predict a model’s result. - Tim Miller, 2017.

Decisions made by a model can be comprehensible to a human if:

a. Model is simple enough to understand and follow along its computation, or
b. A tool intuitively explains why a model is making a decision.

The first kind of models are termed as interpretable models and a tool that generates intuitive explanations for opaque models are termed as explainability tools.

Interpretable and Opaque Models

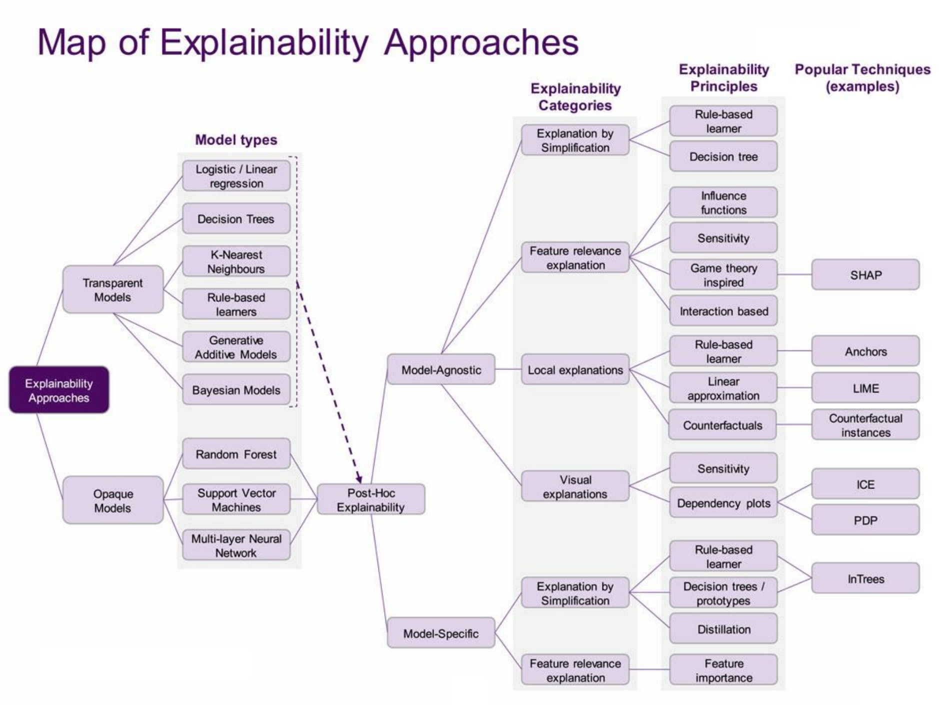
Examples of interpretable models: Linear/Logistic regression, decision trees, K-Nearest Neighbors, rule-based systems. Decisions made by such models are comprehensible to humans as they can easily follow along the computation.

Exception: 1000 branch deep decision tree.

Examples of opaque models: SVMs, Random Forest, and Multi-layer NNs.

Decisions made by such models are not easily comprehensible to humans as they involve humongous number of parameters. Plethora of explainability tools have been developed to generate explanations for them.

Taxonomy of Explainability Methods



Credit: Principles and Practice of Explainable Machine Learning. V. Belle and I. Papantonis. Frontiers in Big Data

Open problems in Explainability / Interpretability [1]

- Assimilating knowledge from human needs for explainability (HCI community) and employing it to develop better explainability methods.
- Adapting current explainability methods to black-box model access situation.
- Developing quantitative evaluation metrics. Establish standardized benchmarks.
- Developing scalable implementations of existing methods.
- Adding actionability dimension to explainability methods.

[1] Pitfalls of Explainable ML: An Industry Perspective. S. Verma et al. ArXiv’ 21.

Resources for Explainability Enthusiasts

- Interpretable Machine Learning. C. Molnar. Online Book’ 21.
- Towards a Rigorous Science of Interpretable Machine Learning. F. Doshi-Velez and B. Kim. ArXiv’ 17.
- Why Should I Trust You? Explaining the Predictions of Any Classifier. M. Ribeiro, S. Singh, and C. Guestrin. KDD’ 16.
- Stop Explaining Black Box Machine Learning Models for High Stakes Decisions. C. Rudin. Nature Machine Intelligence’ 19.
- A Unified Approach to Interpreting Model Predictions. S. Lundberg and S. Lee. NeurIPS’ 17.
- Principles and Practice of Explainable Machine Learning. V. Belle and I. Papantonis. Frontiers in Big Data’ 21.

Robustness

Robustness of a ML pipeline is the ability to perform consistently in presence of “small” changes/corruptions in the model or the datapoints. Broadly, it is these types:

- Learning in presence of corrupted training data, also termed as robust statistics.
- Defending against attacks on ML models, also termed as adversarial attacks.

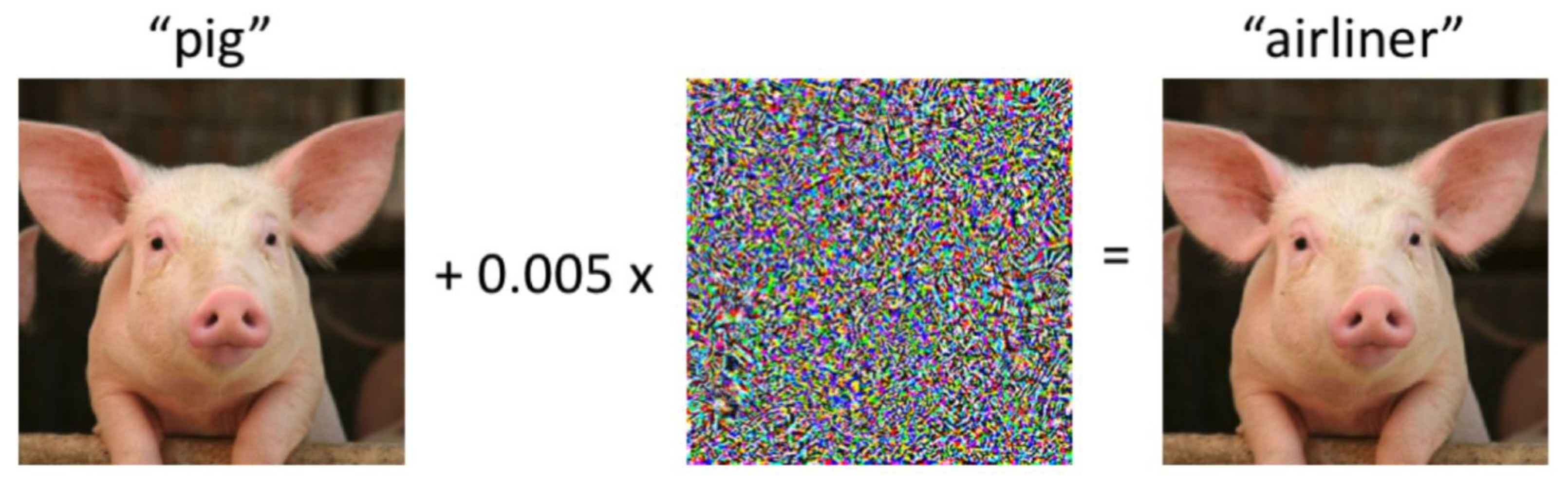
The first task involves learning basic statistics from the data (mean and covariance estimation) or learning supervised ML models.

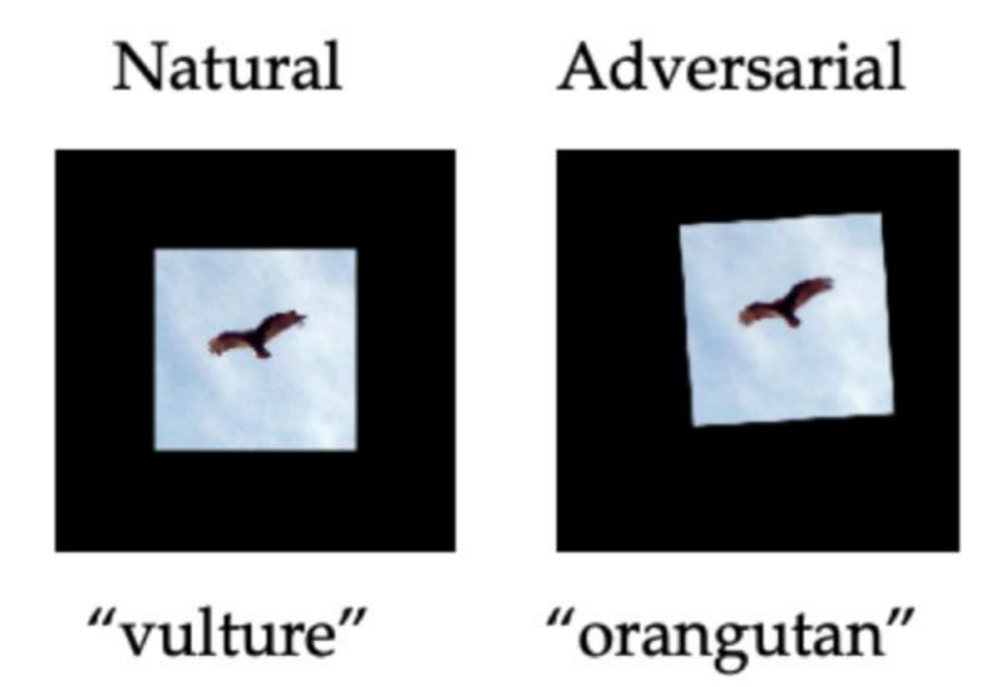
The second task involves developing empirical and certified defenses against adversarial attacks for models deployed in high-stakes environments.

Adversarial Attacks

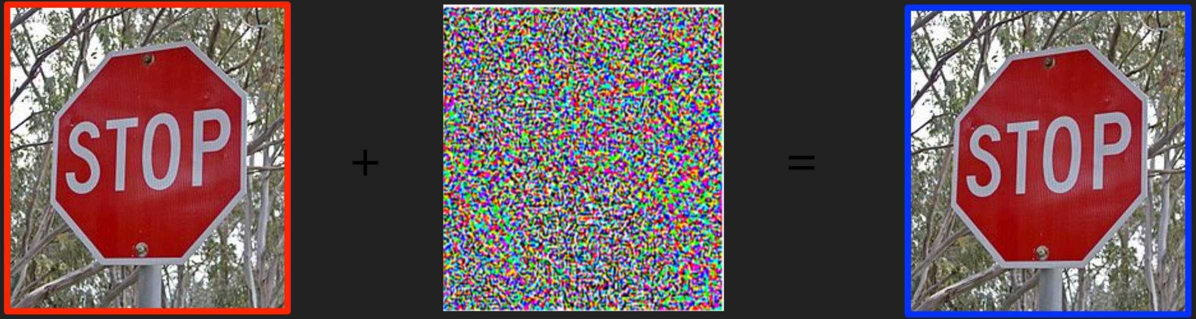
This is about data corruption at test or deployment time. ML models can generate surprising results with imperceptible changes to its inputs.

Let’s take a look.


“pig”
+ 0.005 x
=
“airliner”

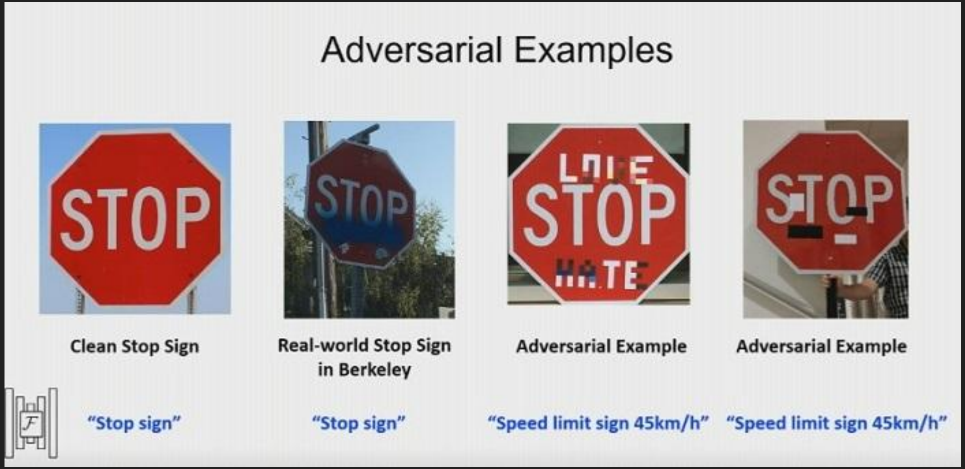


Natural
Adversarial
“vulture”
“orangutan”



classified as **Stop Sign**

classified as **Max Speed 100**



Adversarial Examples

Clean Stop Sign	Real-world Stop Sign in Berkeley	Adversarial Example	Adversarial Example
"Stop sign"	*"Stop sign"*	*"Speed limit sign 45km/h"*	*"Speed limit sign 45km/h"*

Open problems in Robustness

- Developing methods that get better accuracy when adversarially trained. Current models get much lower accuracy than their non-robust counterparts.
- Developing less expensive (in terms of runtime) adversarial training methods.
- ...

Resources for Robustness Enthusiasts

- Adversarial Robustness: Theory and Practice. Z. Kolter and A. Madry. NeurIPS’ 18.
- Adversarial Machine Learning Reading List. N. Carlini.
- Robustness in Machine Learning. J. Li. UW CSE course.
- Certified Defenses Against Adversarial Examples. A. Raghunathan, J. Steinhardt, P. Liang. ICLR’ 18
- Robust Machine Learning Systems: Challenges, Current Trends, Perspectives, and the Road Ahead. M. Shafique et al. ArXiv’ 21.

Privacy-preservation

The two fundamental goals of privacy in ML are:

1. User-control: Provide the user with the right to what information is being collected, for what purpose, and for how long.
2. Data-protection: Protection of the collected data from malicious users, and removal of identifiable information form the data.

We all understand why is user-control an important right of any user, so here will we discuss more about the second aspect of privacy, and learn about what endangers it and why it is necessary.

Privacy Attacks in ML

ML models trained using sensitive data can be subjects to attacks like:

- Membership Inference Attack: Given a datapoint and black-box access to a model, the adversary can detect if that datapoint was used to train the model.
- Model Inversion: Given black-box access to a model, the adversary can infer the sensitive attributes of the input datapoints.
- Model Extraction: Given black-box access to a model, the adversary can infer the model’s parameters, which defeats the goal of a fraud detection model, for example. They can also extract the model architecture in some cases.
- Property Inference: Given black-box access to a model, the adversary can infer properties about some subset of the training data.

Defenses Against Privacy Attacks

- Cryptography: Clients encrypt data before sending it to servers, which hold ML models. This keeps both the client’s data and model parameters private.
- Differential Privacy: DP is the technique about learning nothing about an individual, but about the overall population. This requires adding noise during the training procedure. Noisy SGD and PATE are two popular approaches.
- Trusted Environments: Developing specialized hardware and software for secure execution of sensitive models.
- ML approaches: dropout, weight normalization, dimensionality reduction, etc.

Open problems in Privacy-preservation

- Developing evaluation methods for privacy-preserving algorithms.
- Developing differentially private ML algorithms that offer an acceptable privacy-accuracy trade-off.
- Developing effective defenses against the privacy attacks on ML models.

Resources for Privacy-preservation Enthusiasts

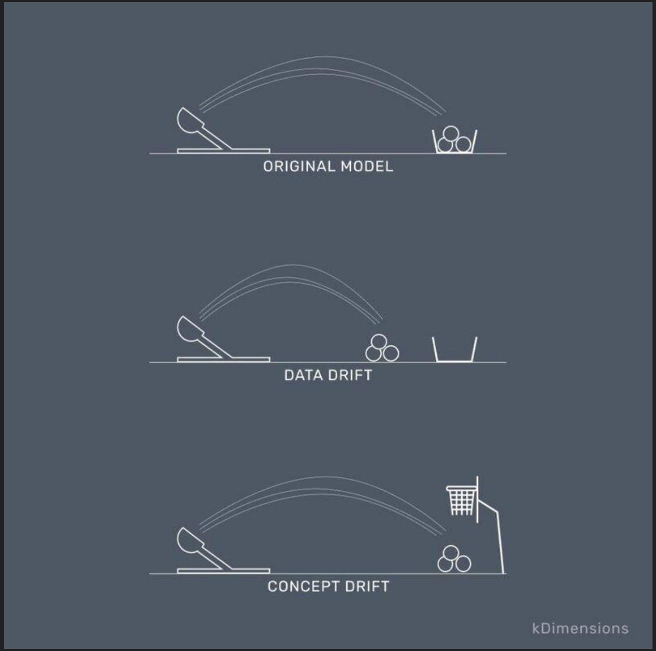
- Deep learning with differential privacy. M. Abadi et al. CCS’ 16
- On the protection of private information in machine learning systems: Two recent approaches. M. Abadi et al. CSF ‘17
- Membership inference attacks against machine learning models. R. Shokri et al. SP ‘17
- Differential privacy: A survey of results. C. Dwork. TAMC’ 08
- Scalable private learning with pate. N. Papernot et al. ICLR ’18.

Model-Drift

ML assumes that the data input to it is from the its training distribution. This is a static environment, which seldom exists in real-world.

When the distribution of the data changes, the predictive performance of the model can detriment. The two main causes are data drift and concept drift.

Data-drift and concept-drift



Data-drift and concept-drift

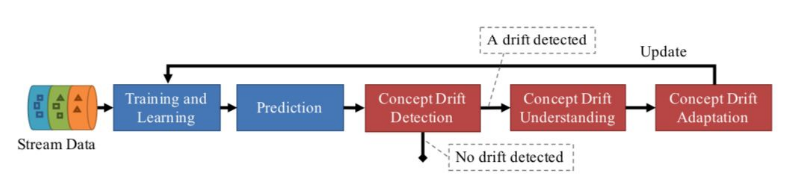
Real-world example of these drifts:

- Spam detection: The nature of spams sent to users of a platform can change as the current spam detection tools catch the spammers, and hence the tools also need to evolve with time.
- Medical data: A model trained on medical images from one hospital might not perform well on similar medical images from a different hospital due to data drift.

Handling Model Drift

Handling model drift consists of two steps:

- Drift Detection: Detecting drift is a non-trivial problem, because the model is only provided with streaming data, not the sampling distribution.
- Drift Adaptation: If detected, ML models needs to be adapted. This can involve retraining, fine-tuning, and using model-ensembles.



Open problems in Model-Drift Detection

- Developing unsupervised drift detection methods that do not assume the existence of ground-truth labels (unlike most current methods).
- Developing benchmarks with real-world datasets which have ground-truth knowledge of the data-drift type and time of introduction.
- Developing data-drift detection techniques for unexplored data modalities like text and graphs.
- Developing drift detection techniques, which along with the timestamp, also inform about the regions of data drift.

Resources for Model-Drift Enthusiasts

- Learning under Concept Drift: an Overview. I. Zliobaite. ArXiv’ 10.
- Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. S. Rabanser et al. NeurIPS’ 19.
- Learning under Concept Drift: A Review. J. Lu et al. IEEE Transactions on Knowledge and Data Engineering’ 19.

Closing remarks

- This field has thrown light on the human-affecting side of ML applications, and therefore has garnered a lot of attention.
- Due to its unripened state, there is a substantial amount of open problems that need exploration and RAISE series should hopefully give in-depth details.
- We are planning to bring on industry practitioners to help talk about the problems they face when ML applications are deployed in real-world. We would also bring academia partners for their perspective on the problems important to them.

Come, join, and help us build ML which is transparent, secure, privacy-preserving, and safe for everyone.

Thank you for your attention!
Questions?

Sahil Verma

vsahil@uw.edu

@Sahil1V