

Mitigating Algorithmic Bias

Agnes Xu

May 6, 2024

Bail Decision and COMPAS

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is one of the most popular scores used nationwide and is increasingly being used in pretrial and sentencing, the so-called “front-end” of the criminal justice system.
- Studies have found that, when using COMPAS, black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk. (unequal false positive rate)
- Discussion has been going on about algorithmic bias problem like this, including what is fairness, why the algorithm is biased, and how to mitigate such bias.

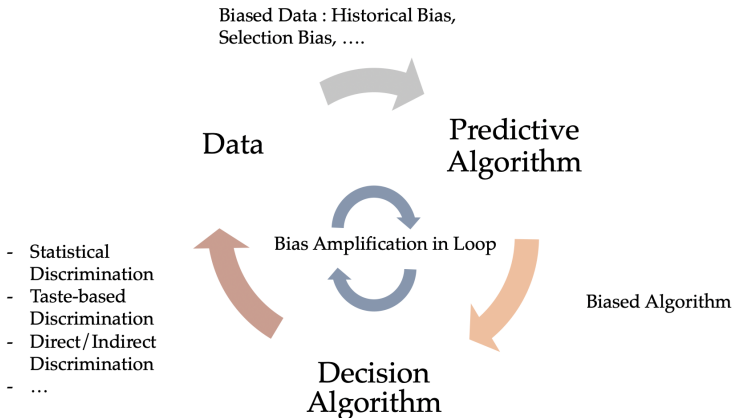
This Paper

- The difference between including sensitive variable into predictive model or not, how much sacrifice in predictive accuracy would be caused by removing those variables?
- Can simply removing sensitive variables really solve the bias? Is there any chance of reconstructing those variables from remaining variables?
- Which measurement of fairness makes more sense? Is there a "hyper" decision rule for the fairness criteria?
- How can we improve the model without directly involving sensitive variable? For example, using embedding models like PCA and auto-encoders. Or including fairness measures in penalty terms in regularization models

Research Design

- Using the COMPAS data, we're going to first examine the fairness measures widely used, and then move on to improving predictive algorithm and comparing the result of models with and without sensitive variables.
- Representation Learning Approach: We trained PCA and auto-encoders to prevent directly involving any demographic variables and use pre-training methods to balance the represented variables.
- Regularization Approach: Suppose the statistical model is $f(x)$ and the average predicted difference between two models is $p(x)$, then regularize $g(x) = f(x) + ap(x)$, where a is a hyperparameter to be tuned. In this way the final model is trained to be including the knowledge from theoretical model to produce more robust result against model misspecification. We can apply the similar logic here by setting $p(x)$ to be some fairness measures, and tune the algorithm in the direction that not only reduces variance and bias, but also the deviation from fair status.

Bias and Discrimination Loop



Bias

- *Bias-in-Bias-out* (data \rightarrow predictive algorithm): Training data is corrupted by historical bias and bias happening in the measuring, collecting and labelling process. (\rightarrow pre-process techniques)
- *Algorithmic Bias* (predictive algorithm \rightarrow decision algorithm) : Predictive algorithm performing biased towards some unfavoured group. (\rightarrow in- and post- process techniques)
- *Feedback Loop* : Enhance and spread bias. (\rightarrow break the loop: collect random data / reinforcement learning)

Discrimination

- *Direct and Indirect discrimination*: sensitive attributes included explicitly (direct, or taste-based) or reconstructed through other variables (indirect) in objective function.
- *Statistical Discrimination*: In lack of perfect information, decision-makers use average group statistics to judge a person from that group.

Predictive Fairness

Different notions of predictive fairness fall into different level of population:

- *Individual Fairness* : Give similar predictions to similar people
- *Group Fairness* : Treat different groups equally
- *Subgroup Fairness* : Balancing fairness at group and individual level by picking a group level fairness constraint and test it over a large collection of subgroups.

Predictive Fairness - Group Level

group fairness is often framed in terms of protected attributes while allowing for differing treatment based on a set of qualifications.

Notion	Formula
Demographic Parity	$P(\hat{Y} A = 0) = P(\hat{Y} A = 1)$
Conditional Parity ¹	$\Pr(\hat{Y} = 1 A = a, \mathbf{z} = \mathbf{z}) = \Pr(\hat{Y} = 1 A = a', \mathbf{z} = \mathbf{z})$
Equalized Odds ²	$P(\hat{Y} = 1 A = 0, Y = 1) = P(\hat{Y} = 1 A = 1, Y = 0)$
Calibration	$P(\hat{Y} = 1 A = a, Y = y) = P(Y = 1 A = a, Y = y)$

- **Incompatibility** : Kleinberg et al. (2016), Pleiss et al. (2017)
- **Motivation** : Proportional representation is often desirable its own right; alternatively, the absence of proportional allocation of goods can signal discrimination in the allocation process, typically against historically mistreated or under-represented groups. (Dwork and Ilvento, 2018)
- **Critics** : One can maintain group level fairness while discriminate against individuals through redlining and other strategic design.

¹Ritov et al. (2017)

²and a bunch of balance conditions for terms in confusion matrix

Predictive Fairness - Individual Level

Notion	Formula
FTA	$\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$
FTU	$\hat{Y} : X \rightarrow Y$
CF	$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$

- **Motivation** : Makes more sense ethically.
- **Critics**: Individual level fairness is hard to achieve due to strong heterogeneity. And the assumptions casual approach imposed are too strong in practice. For the statistical approach (FTA), though similarity metric is assumed known and computable, in reality historical bias, computational and ethical& legal problems exists with it (Kim et al. 2018).

Composition Problem of Fairness Notions

from fair classifier to fair system

Dwork and Ilvento (2018) :

- Naive composition of group-fair classifiers will not in general yield a group-fair system, paralleling an analogous result for individually fair classifiers (Dwork and Ilvento 2018b) and such composition requires additional considerations for subgroup fairness.
- A promising direction for future work is the augmentation of classifiers group fairness for large, intersecting, groups (Kearns et al., 2017; Hebert-Johnson et al., 2017), as well as classifiers with Individual Fairness for large subgroups) (Kim et al., 2018), to incorporate contextual information, with the goal of improving composition.

Balancing Group and Individual Fairness

Berk et al. (2017) : fair regression with hybrid penalty term involving group and individual level unfairness measure weighed according to relative importance.

Hu and Chen (2017) : resolve both within the particular domain of employment discrimination by constructing a dual labor market model composed of a Temporary Labor Market, in which firm strategies are constrained to ensure group-level fairness, and a Permanent Labor Market, in which individual worker fairness is guaranteed. Fairness restrictions on hiring practices induces an equilibrium that Pareto-dominates those arising from strategies that employ statistical discrimination or a “group-blind” criterion. Individual worker reputations produce externalities for collective reputation, generating a feedback loop termed a “self-fulfilling prophecy.”

Decision Fairness

Davies et al. (2017) :

- 1 **Statistical Parity** : an equal proportion of defendants are detained in each race group

$$\mathbb{E}[d(X) \mid g(X)] = \mathbb{E}[d(X)]$$

- 2 **Conditional Statistical Parity** : controlling for a limited set of “legitimate” risk factors, an equal proportion of defendants are detained within each race group

$$\mathbb{E}[d(X) \mid \ell(X), g(X)] = \mathbb{E}[d(X) \mid \ell(X)]$$

- 3 **Predictive Equality** : the accuracy of decisions is equal across race groups, as measured by false positive rate (FPR)

$$\mathbb{E}[d(X) \mid Y = 0, g(X)] = \mathbb{E}[d(X) \mid Y = 0]$$

Trade-offs

Imposing fairness constraint to algorithms means deviating from the most use of information. Also the nature of empirical problems suggest incompatibility of fairness notions. Any fairness enhancement techniques faces the two fundamental trade-offs:

- ① *Fairness v.s Accuracy*
- ② *Fairness Notion v.s Fairness Notion* : Incompatibility of fairness notions (Kleinberg et al. 2016)

Categories

Existing methods to obtain fair predictive models can be classified by the stage of training where they try to intervene:

- **Pre-process**
- **In-process**
- **Post-process**

Or the class of algorithms:

- **Regularization**
- **Adversarial Learning**
- **Representation Learning**
- **Causal Modeling**

Pre-process

Methods in this category include changing the training data before feeding it into a machine learning algorithm and simulating debiased data:

- **Re-weight or adjust labels** : Pedreschi et al. (2008), Kamiran and Calders (2009), Luong et al. (2011), Kamiran and Calders (2012)
- **Modify feature representations** : Zemel et al. (2013), Feldman et al. (2015), Louizos et al. (2016), Calmon et al. (2017), Samadi et al. (2018)
- **Generate fair synthetic data** : Xu et al. (2018)

In-process

Main works on algorithmic fairness focus on the training process, where fairness notions are used as constraint in optimization problem or regularization term

- **Fairness as Optimization Constraint** : Zafar et al. (2017), Bechavod and Ligett (2010), Woodworth et al. (2017)
- **Fairness as Regularizer** : Kamishima et al. (2011), Bechavod and Ligett (2017), Quadrianto and Sharmanska.(2017)
- **Group-specific Classifier** : Dwork et al. (2018)
- **Tree-based Method** : Kamiran et al. (2010), Kamiran et al. (2013), Valdivia et al. (2020), Raff et al. (2017), Fantin (2020)
- **Regression**: Berk et al. (2013), Agarwal et al. (2019)

Post-process/Decision

These methods perform post-processing of the output scores of the classifier to make decisions fairer. This is performed after training by accessing a holdout set which was not involved during the training of the model

- **Flipping decisions** : Hardt et al. (2016), Woodworth et al. (2017)
- **Multiple Threshold Rule** : Davies et al. (2017), Kleinberg et al. (2018)

Adjusting Labels and Reweighing Data

pre-processing historically biased data

Kamiran and Calders (2012) : four ways to preprocess the dataset

- ① **Suppression** : remove protected attributes A and attributes most correlated with A .
- ② **Massaging the dataset** : change the labels of the most likely victims(discriminated ones) and profitters (favored ones) in the dataset according to rank (Kamiran and Calders, 2009a).
- ③ **Reweighing** : assign weights to tuples in the training dataset. (Calders, Kamiran and Pechenizkiy , 2009)
- ④ **Sampling** : two techniques to select which objects to duplicate, and which to remove
 - ① Uniform Sampling (with replacement) : every object has a uniform probability to be duplicated to increase the size or to be skipped to decrease the size of a group
 - ② Preferential Sampling : borderline objects get high priority for being duplicated or being skipped. A ranker is used to decide which objects are at the border.

Adjusting Labels and Reweighing Data

pre-processing historically biased data

Luong et al. (2011) : a variant of KNN classification for the discovery of discriminated objects : data object is discriminated if there exist a significant difference of treatment among its neighbors belonging to a protected- by-law group (i.e., the deprived community) and its neighbors not belonging to it (i.e., the favored community). Then change the changing the decision value for tuples labeled as discriminated before training a classifier. ³

³compared to messaging approach which only change the minimal number of objects, this approach continues relabeling until all labels are consistent.

Fair Representation Learning

dimension reduction methods, protected attributes as sensitive variable correlated with objective

Zemel et al. (2013) : fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group.

Samadi et al. (2018) : fair PCA, to maintain similar fidelity for different groups and populations in the dataset in PCA pre-process.

Fair VAE

protected attributes as nuisance variables

Louizos et al. (2016) : introduce an additional penalty term based on the “Maximum Mean Discrepancy” (MMD) penalizing difference between variational posterior $q(\mathbf{z} \mid \mathbf{x}, A = 0)$ and $q(\mathbf{z} \mid \mathbf{x}, A = 1)$.

Liu and Burke (2018) : representation-learning task as an optimization objective that minimize the loss of the mutual information between the encoding and the sensitive variable.

Davies and Goel (2018) : flexibly fair representation learning by disentangling information from multiple sensitive attributes.

Fairness as Constraint Optimization

Zafar et al. (2017) : introduced measure of decision boundary fairness and mechanism to maximize accuracy of convex margin-based fair classifiers (logistic regression and SVM) with disparate treatment and disparate impact constraint.

Woodworth et al. (2017) : estimate an empirical risk minimizing predictor \hat{Y} subject to approximate non-discrimination constraint (relaxation of equalized odds based only on a second-moment condition instead of full conditional independence)

Regression with Fairness Regularizer

In-process, regularization

Berk et al. (2017) : Three fairness regularizer for (linear and logistic) regression and by varying the weight on the fairness regularizer, we can compute the efficient frontier of the accuracy-fairness trade-off which is measured by *Cost of Fairness (PoF)*

- **Individual Fairness** : "for every cross pair $(x, y) \in A_1, (x', y') \in A_2$, the model w is penalized for how differently it treats x and x' "

$$f_1(w, A) = \frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in A_1 \\ (x_j, y_j) \in A_2}} d(y_i, y_j) (w \cdot x_i - w \cdot x_j)^2$$

- **Group Fairness** : "On average, the two groups' instances should have similar labels (weighted by the nearness of the labels of the instances)"

$$f_2(w, A) = \left(\frac{1}{n_1 n_2} \sum_{(x_i, y_i) \in A_1} d(y_i, y_j) (w \cdot x_i - w \cdot x_j) \right)^2$$

- **Hybrid Fairness**: "Hybrid fairness requires both positive and both negatively labeled cross pairs to be treated similarly in an average over the two groups"

$$f_3(w, A) = \left(\sum_{\substack{(x_i, y_i) \in A_1 \\ (x_j, y_j) \in A_2 \\ y_i = y_j = 1}} \frac{d(y_i, y_j) (w \cdot x_i - w \cdot x_j)}{n_{1,1} n_{2,1}} \right)^2 + \left(\sum_{\substack{(x_i, y_i) \in A_1 \\ (x_j, y_j) \in A_2 \\ y_i = y_j = -1}} \frac{d(y_i, y_j) (w \cdot x_i - w \cdot x_j)}{n_{1,-1} n_{2,-1}} \right)^2$$

Fair Decision Tree

Most fair decision tree techniques consist of a modification of the information gain metric when selecting attributes to split leaf nodes. The general idea is to add a new fairness information gain that measures how closely the selected attribute correlates with the protected features. Often this criteria is subtracted or multiplied by the standard information gain criteria to split based on accurate but fair attributes

Kamiran et al. (2010) : Define a good split to be the one that achieves a high purity with respect to the class label, but a low purity with respect to the sensitive attribute. Then guide the iterative tree refinement procedure, disallowing steps that would increase discrimination in the predictions or explicitly adding a penalty term for increasing discrimination into the quality scores of the splits.

Valdivia et al. (2020) : Use meta-learning to guide an tree classifier to obtain the best tradeoffs between accuracy and fairness (Pareto optimal), by learning the best combination of hyperparameters given a dataset and its protected attribute that identifies a protected group.

Fair Forest

Raff et al. (2017) : modify the information gain similarly as in tree models and modify the Gini impurity score to measure both the class label and the protected feature. The model chooses features at nodes in the decision trees which are correlated with the label but not with the protected attribute. The trees are combined into forest using standard majority voting.⁵

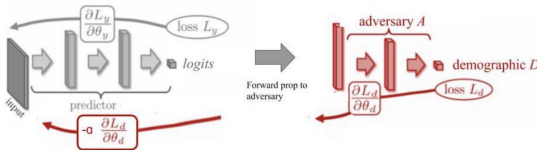
Fantin (2020) : use the extra random decision tree algorithm (Geurts et al. 2016) for the base trees and add the constraint that only one feature is randomly selected at each split. This added constraint amounts to building a completely random decision tree. Trees are combined linearly using a weighted voting mechanism learnt from a ridge regression to avoid trivial decision trees for improved accuracy.

⁵if a feature is highly correlated with the label and the protected attribute, but no better feature exists to split the node on then this feature will be chosen. So in certain data sets, this algorithm can still lead to biased results.

Debiasing Adversarial Learning

In-process based on feedback structure

Wadsworth et al. (2017) : use the feedback structure to check whether a trained classifier is fair or not and then update the model accordingly. The goal is to the predictor to be good at predicting \hat{Y} and bad at predicting a logit that is highly correlated with demographic.



The output of predictor is passed to an adversary, which tries to predict demographic D . The adversary may have different inputs depending on the fairness definition needing to be achieved. For instance, i to achieve Equality of Odds, the adversary would get the true label Y in addition to the predicted label \hat{Y} .

FairGAN

Fair GAN models are often constructed as minimax optimization problems that aim at maximizing the predictor's capability to accurately predict the outcomes while minimizing the adversary's capability to predict the sensitive feature.

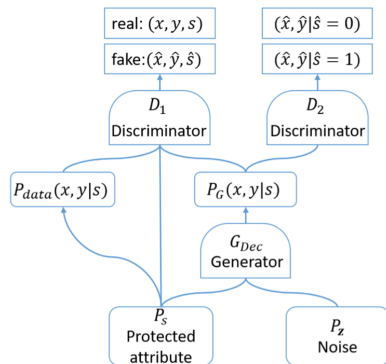
- **generating fair synthetic data** : Xu et al. (2018), Abusitta et al. (2019)
- **learn fair representations** : Edwards and Storkey (2015), Beutel et al. (2017) , Madras et al. (2018)

FairGAN

pre-process, adversarial learning to generate synthetic data

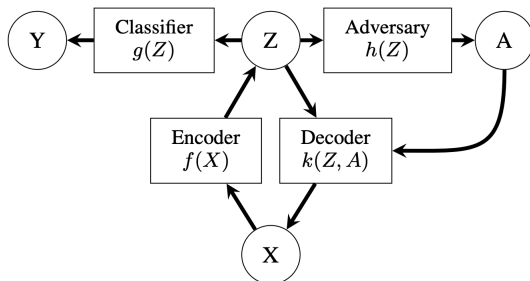
Xu et al. (2018) : generated synthetic debiased data (using *statistical parity* as fairness condition) from FairGAN for training and testing.

- G_{Dec} : To generate the dataset with discrete attributes, G_{Dec} combine a regular generator G and a decoder Dec . $(\hat{x}, \hat{y}) = G_{Dec}(z, s) = Dec(G(z, s))$.
- D_1 : Regular discriminator.
- D_2 : Fairness discriminator.
- The minimax game:
$$\min_{G_{Dec}} \max_{D_1, D_2} V(G_{Dec}, D_1, D_2) = V_1(G_{Dec}, D_1) + \lambda V_2(G_{Dec}, D_2),$$
where λ specifies a trade off between utility and fairness of data generation.



Learning Fair Representation

Madras et al. (2018) : Choice of data representation Z as a representation learning problem with an adversary criticizing potentially unfair solutions. Common group fairness metrics (demographic parity, equalize odds, and equal opportunity) are connected to adversarial learning by proper adversarial objective functions for each metric that upper bounds the unfairness of arbitrary downstream classifiers in the limit of adversarial training.



Fair Causal Inference

- **Counterfactual Fairness(CF):** Kusner et al. (2017)
- **Multi-World Fairness(MWF):** Russell et al. (2017)
- **Non-causal perspective of CF:** Johnson et al. (2016)
- **Interventional CF:** Kilbertus et al. (2017)
- **CF without explicit structural equations:** Nabi and Shpitser (2018)
- **Path-specific CF (PSCF):** Chiappa and Gillam (2018)

Flipping Predictions Towards Fairness

post-process

Hardt et al. (2016) : first learn an accurate but possibly discriminatory predictor than “correct” it by taking into account protected attributes. When Outcome Y is binary and the predictors \hat{Y} are real-valued, unconstrained Bayes optimal least-square regressor can be post hoc corrected to the optimal predictor with respect to the 0-1 loss.

Woodworth et al. (2017) : show that post-hoc correction approach (Hardt et al. 2016) can be highly suboptimal⁷ and that it is necessary to directly incorporate non-discrimination into the learning process. Then with additional samples from (\hat{Y}, Y, A) , derive a randomized predictor to further reduce discrimination.

⁷First, for general distributions, it is impossible to learn the Bayes optimal predictor from finite samples of data. Also, the post hoc correction of even the optimal unconstrained predictor with respect to the 0-1 (non-convex) or even hinge (non-strict but convex) losses can have much worse performance than the best non-discriminatory predictor. Moreover, if the hypothesis class is restricted there can also be a gap between the post hoc correction of the optimal predictor in the hypothesis class and the best non-discriminatory predictor, even when optimizing a strictly convex loss function.

Fair Decision Rule

Davies et al. (2017) : Formulate the trending decision fairness conditions as constraints in a utility maximizing framework and prove that threshold rules are always the optimum. And discuss the limitation of fairness definitions and the relationship between fairness of decision and prediction.

Kleinberg et al. (2018) : The use of race will always be strictly improving for the equitable planner's objective function as long as it is useful for predicting outcome. And a preference for fairness should not change the choice of predictor.

Evaluating Algorithms

Lakkaraju et al. (2017) : compare the performance of predictive models and human decision makers without resorting to counterfactual inference.

Hamilton (2017) : compare four fairness metrics using four algorithms across three datasets.

Evaluating Impact

- Statistical paradigm:
 - **two-stage model of delayed impact:** Liu et al. (2018)., Kannan et al. (2018)
 - **simulated long-term impact:** Ensign et al. (2018), D'Amour et al. (2020)
- Causal paradigm:
 - **discrimination mechanism:** Kusner et al. (2019), Heidari et al. (2019), Nabi et al. (2019)
 - **fair policy intervention:** Kusner et al. (2018)

Statistical Methods to Evaluate Impact

Liu et al. (2018) : an one-step feedback model showing that common fairness criteria in general do not promote improvement over time, and may in fact cause harm in cases where an unconstrained objective would not.

Ensign et al. (2018) : a mathematical model of predictive policing that proves why this feedback loop occurs, and how to change the inputs to the predictive policing system so the runaway feedback loop does not occur, allowing the true crime rate to be learned. (with empirical result on labour market data)

D'Amour et al. (2020) : Adapt the Markov Decision Processes (MDPs) framework for fairness-oriented simulations to model feedback dynamics and analyse long term fairness consequences.

Datasets

Dataset Name	Domain	# Records	Sensitive Attributes	Target Attributes
ProPublica	Criminal risk assessment	6,167	Race; Gender	Whether an inmate has recidivated (was arrested again) in less than two years after release from prison
Adult	Income	48,842	Age; Gender	Whether an individual earns more or less than 50,000\$ per year
German	Credit	1,000	Gender; Age	Whether an individual should receive a good or bad credit risk score
Ricci	Promotion	118	Race	Whether an individual receives a promotion
Mexican poverty	Poverty	183	Young and old families; Urban and rural areas	Poverty level of households
Diabetes	Health	100,000	Race	Whether a patient will be readmitted
Heritage health	Health	147,473	Age	Whether an individual will spend any days in the hospital in the next year
College Admissions	College Admissions	20,000	Gender; Race	Whether a law student will pass the bar exam
Bank Marketing	Marketing	41,188	Age	Whether the client subscribed to a term deposit service
Loans Default	Loans	30,000	Gender	Whether a customer will default on payments
Dutch Census	Census	189,725	Gender	Whether an individual holds a highly prestigious occupation
Communities and Crimes	Crime	1,994	Percentage of African-American population	For each community, the number of violent crimes per 100,000 individuals

COMPAS Data

- Source: Broward County, Florida originally compiled by ProPublica, available in R package 'fairness'
- Size: 6172 defendants
- Variables: race, age, sex, number of prior convictions, and COMPAS violent crime risk score (a discrete score between 1 and 10)

Introduction
○○○○○○

Fairness
○○○○○○○

Predictive Algorithm
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

Decision
○






Evaluation
○○○

Data
○○●

GitHUb Repository

GitHub Link

Reference

-  Pessach, Dana, and Erez Shmueli. "Algorithmic fairness." arXiv preprint arXiv:2001.09784 (2020).
-  Mehrabi et al. (2019) - A Survey on Bias and Fairness in Machine Learning
-  Chen et al. (2020) - Bias and Debias in Recommender System- A Survey and Future Directions
-  Berk, Richard, et al. "A convex framework for fair regression." arXiv preprint arXiv:1706.02409 (2017).
-  Xu, Depeng, et al. "Fairgan: Fairness-aware generative adversarial networks." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.