Fairness in Machine Learning

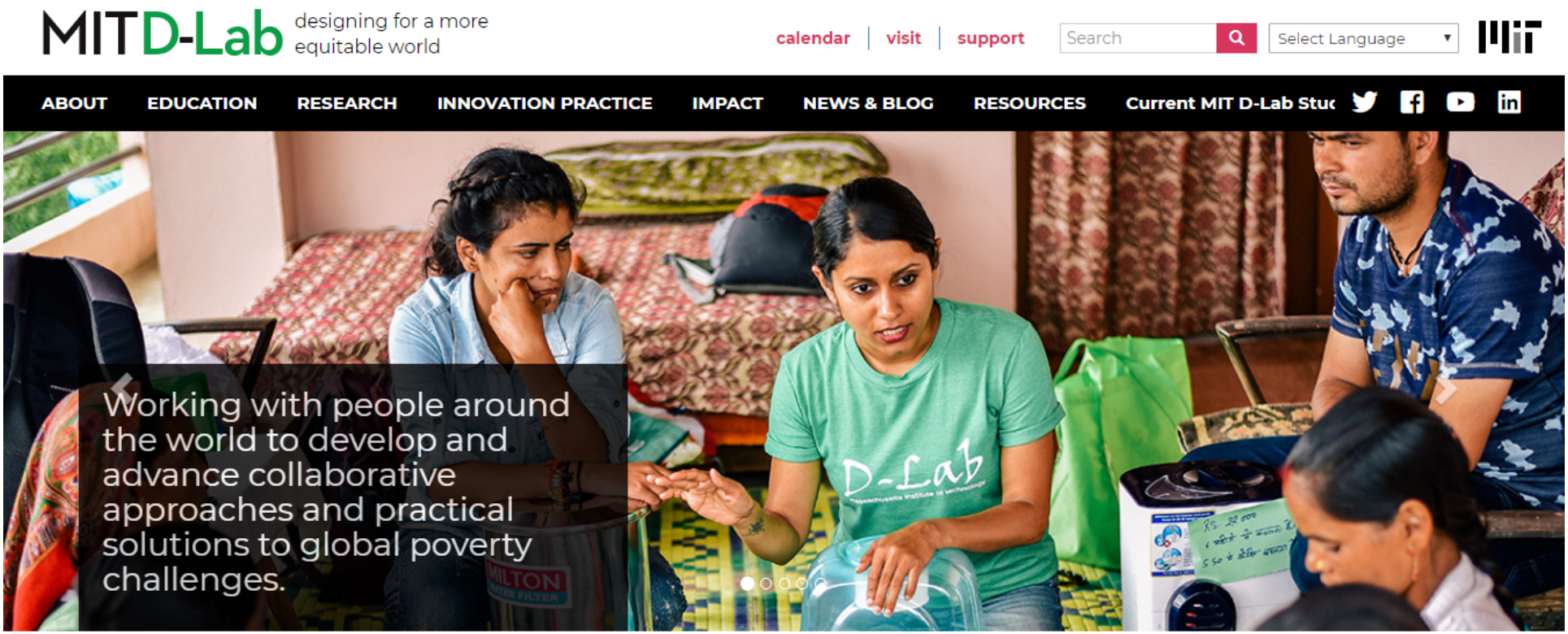
Mike Teodorescu
Assistant Professor, Information Systems Department, Boston College
D-Lab, Massachusetts Institute of Technology

Hello

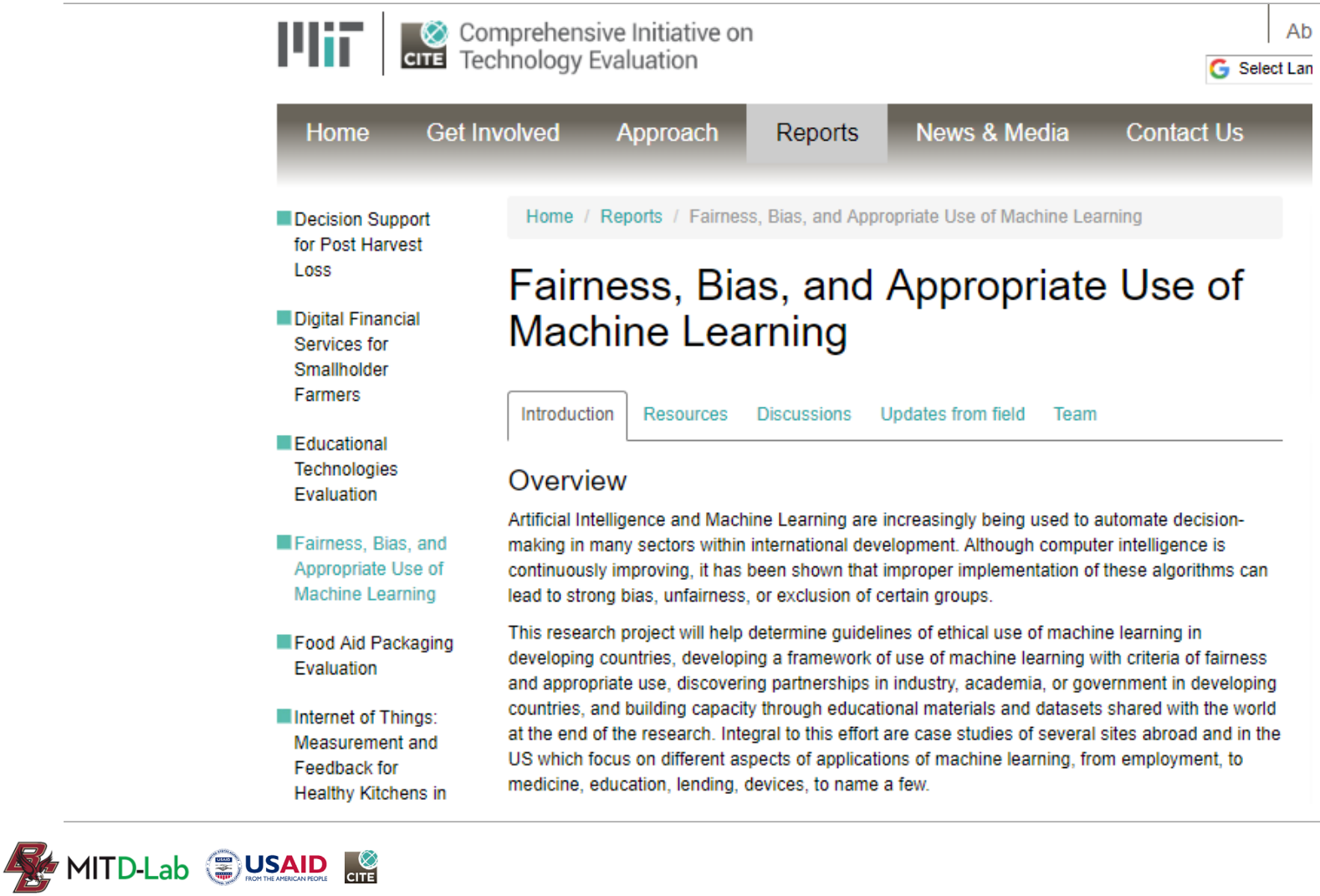
Boston College Information Systems Department

- We study Information Technology and its effects on firms and society
- Mix of computer science and economics coursework
- We train the data scientists who go to banks, government agencies, credit bureaus, chief data officers, chief information officers, chief marketing officers, etc.
- Our department is a mix of computer scientists, economists, and ethicists
- Some of our students become entrepreneurs

The Lab



MITD-Lab USAID FROM THE AMERICAN PEOPLE CITE



MITD-Lab USAID CITE

The Grant

■ Publications

Exploring Fairness in Machine Learning for International Development

Publication | Mar 02, 2020 | Yazeed Awwad, Richard Fletcher, Daniel Frey, Amit Gandhi, Maryam Najafian, Mike Teodorescu

Exploring Fairness in Machine Learning for International Development

https://d-lab.mit.edu/resources/publications/exploring-fairness-machine-learning-international-development?fbclid=IwAR1w2VWmaFlnHyXRsmk6zMxyj153-QGmKzspJ1yk-nkLXi4xhRdZukcc4sw

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

170

WASHINGTON — The Department of Housing and Urban Development sued Facebook on Thursday for engaging in housing discrimination by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.

In addition to targeting Facebook's advertising practices, the housing department, known as HUD, claims in its lawsuit that the company uses its data-mining practices to determine which of its users are able to view housing-related ads. On both counts, the agency said, Facebook is in violation of the federal Fair Housing Act.

https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html

MPW • AMAZON

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

By DAVID MEYER October 10, 2018

Machine learning, one of the core techniques in the field of artificial intelligence, involves teaching automated systems to devise new ways of doing things, by feeding them reams of data about the subject at hand. One of the big fears here is that biases in that data will simply be reinforced in the AI systems—and Amazon seems to have just provided an excellent example of that phenomenon.

https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

MIT Technology Review

Sign in | Subscribe

Topics | Magazine | Newsletters | Events

IAN WALDIE/GETTY IMAGES

Tech Policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** Jan 21, 2019

https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai/

MITD-Lab | USAID | CITE

Why should organizations care about fair ML?

- Well, because it's the right thing to do
- Machine Learning models are **susceptible to bias and discrimination** and can cause harm to both job applicants and organizations alike
- Organizations can expose themselves to legal liabilities, loss of customers, reputational damages (Anjuwa, 2020)
- Certain algorithm categories are **inherently difficult to follow intuitively** (for example, neural networks) and processes need to be established during the implementation as well as after (Serra, 2018)
- **ML algorithms can pick up societal biases**

Current state of the art

- The use of machine learning in organizations presents a double-edged sword: machine learning tools **reduce costs on otherwise repetitive, time-consuming tasks**, yet run the risks of introducing **systematic unfairness** in organizational processes.
- Issues of behavioral ethics in machine learning implementations in organizations have not been thoroughly addressed in prior literature, as many of the necessary concepts are disparate across three literatures – **ethics, machine learning, and management**.

Protected Attributes

Algorithms should make clear choices to avoid discrimination based on individual protected attributes, such as:

- **race;**
- **religion;**
- **national origin;**
- **gender;**
- **marital status;**
- **age;**
- **socioeconomic status.**

Example of Laws In the US

- Penalties for discriminating in housing (US Fair Housing Act)
- Hiring (the collection of laws also known as Federal Equal Employment Opportunity – Civil Rights Act Title VII 1964, EPA 1963, ADEA 1967, ADA 1990, Rehabilitation Act 1973, Civil Rights Act 1991, GINA 2008).
- **The most recent iteration of ECOA requires firms to test algorithms for unfair outcomes and has penalties for failures in testing**
- Many countries have their own varieties of laws

ECOA prevents discrimination in lending

- The Federal Trade Commission (FTC), the nation's consumer protection agency, enforces the Equal Credit Opportunity Act (ECOA), which prohibits credit discrimination on the basis of **race**, **color**, **religion**, national origin, sex, marital status, **age**, or because you get public assistance.

- https://www.consumer.ftc.gov/articles/0347-your-equal-credit-opportunity-rights

ML Risks

- Biased sampling
- Small sample sizes for certain subgroups of the data
- Hidden correlations in input data
- Data not representative for all groups (Tommasi et al., 2017)
- Data may include protected attributes (Datta, 2017)
- Data with a large degree of noise (Domingos, 2012)
- Due to correlations between protected variables, resulting unfairness levels dramatically differ not only across protected categories, but also within them, i.e. across protected subgroups (Kearns et al, 2018)

It all starts with the training set

- **Bad training data => bad prediction**
- The training set can carry the biases of the **people labeling the data**
- The training data may not be representative of all the groups
- Past data may not predict current events (or individuals may misremember past situations **selective perception**, Dearborn & Simon, 1958)

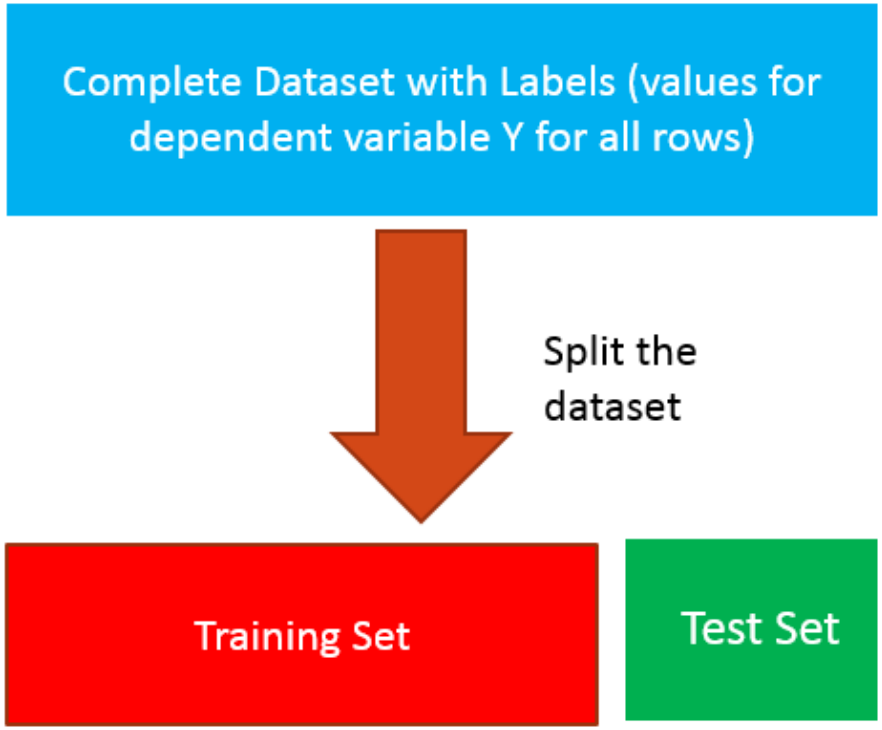
What is “Learning”?

- Data component:
- Training set
- Test set

Supervision:

- Supervised learning – human checks each prediction; algorithm adjusts based on corrections
- Unsupervised learning – algorithm runs according to a preprogrammed optimization criterion

Train, Test

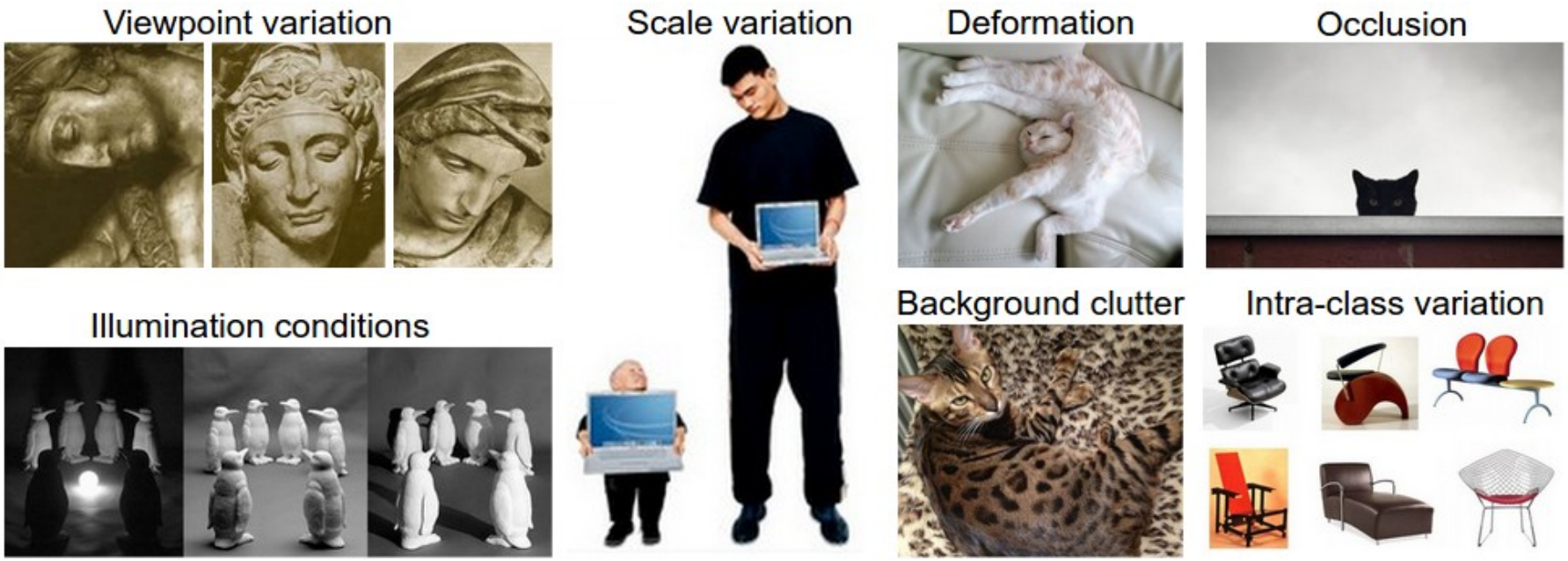


Key point: you should randomly sample the training set and the test set

What is “Machine Learning”?

- A subfield of Computer Science
- A subfield of Statistics
- The study of methods that enable computers to find patterns in data and use those patterns to construct predictions (aka “data science”)
- A collaboration across fields as diverse as neurobiology, linguistics, mathematics, art, and engineering
- The pursuit of a machine that will think and speak like a human (aka “the Turing Test”)

Classification is often an ML Task

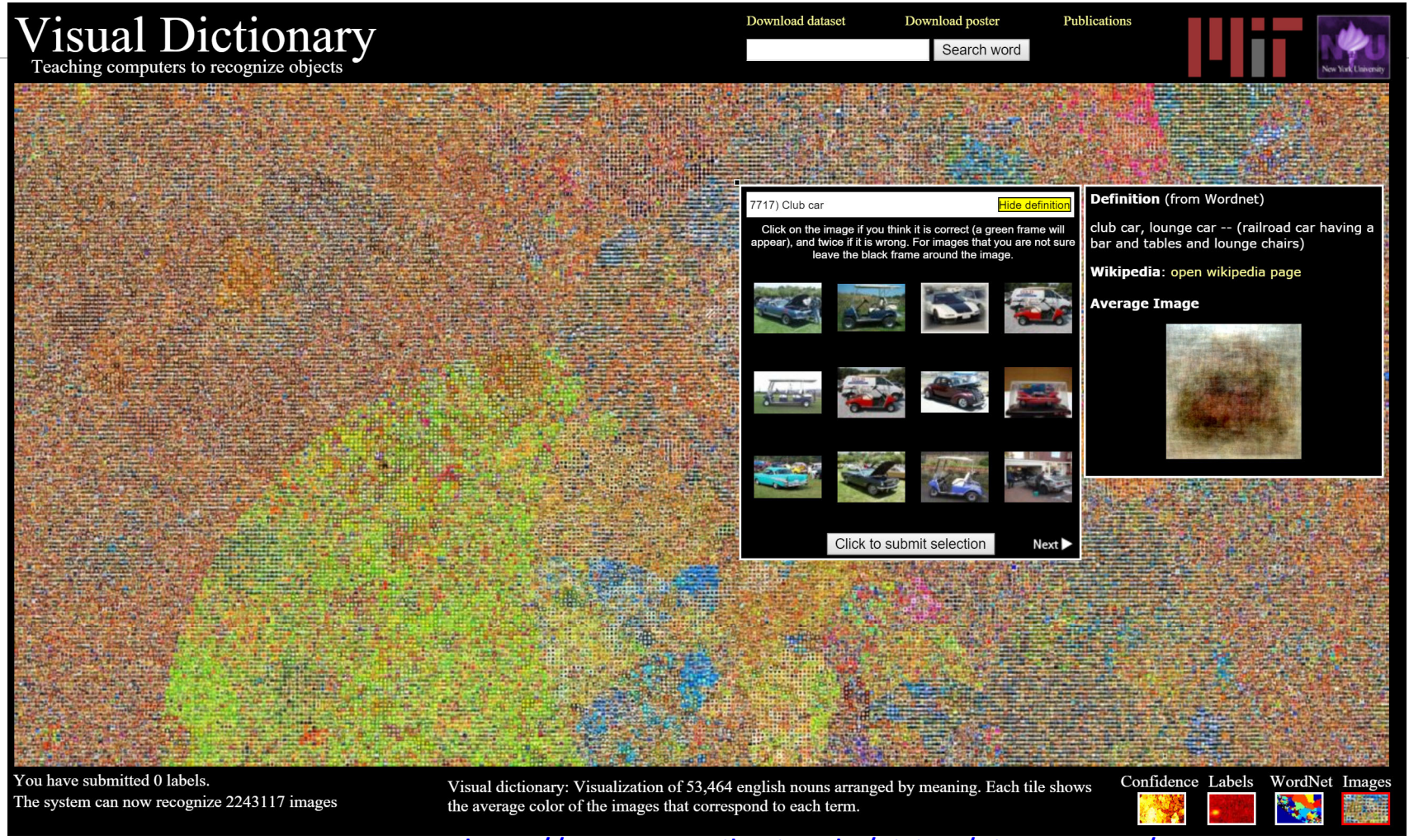


http://cs231n.github.io/assets/challenges.jpeg

Image recognition is particularly hard

https://www.freecodecamp.org/news/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d/

https://www.pinterest.cl/pin/751467887795716475/



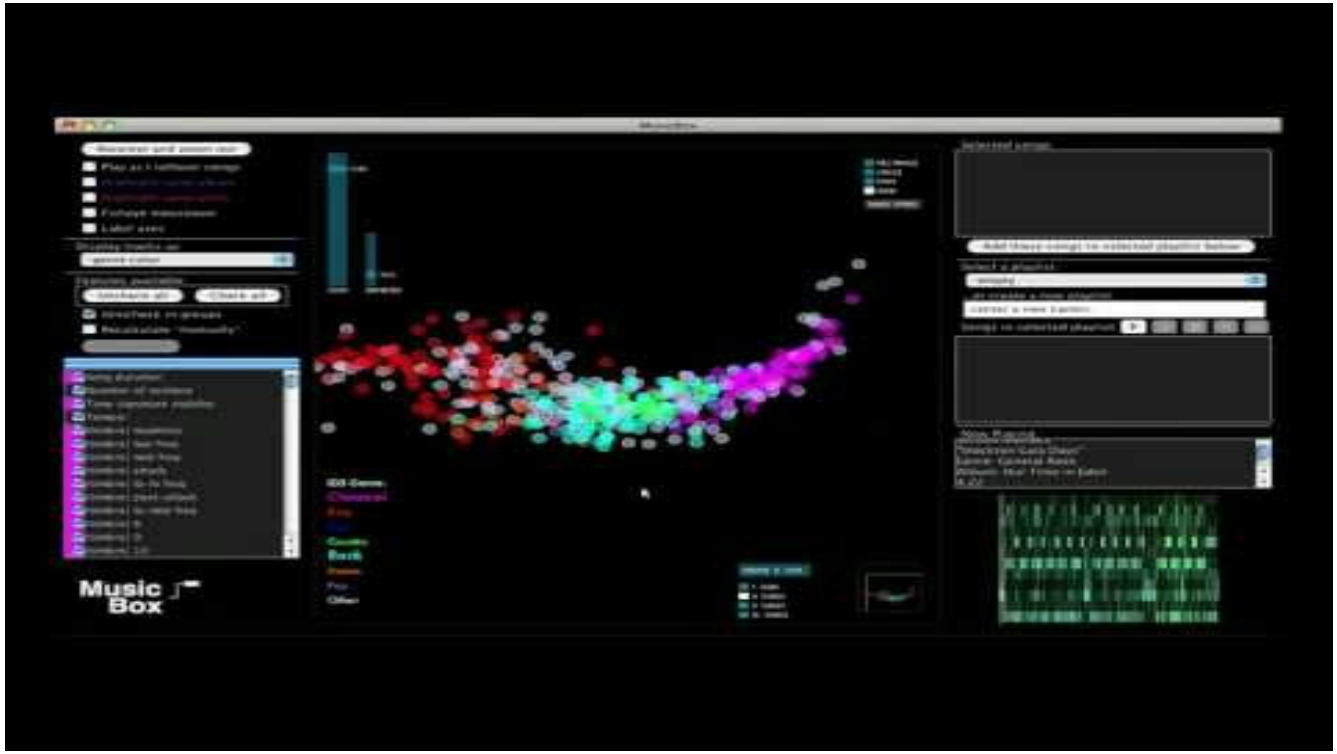
http://groups.csail.mit.edu/vision/TinyImages/

MITD-Lab USAID FROM THE AMERICAN PEOPLE CITE

Gesture recognition as classification task

https://www.youtube.com/watch?v=MwZMNMmODJA

Music genre recognition



http://thesis.flyingpudding.com/videos/demo/index.html

All these great applications… when do we get in trouble?

→Whenever the classifier discriminates based on a protected attribute

Facial recognition: Not Identical By Race

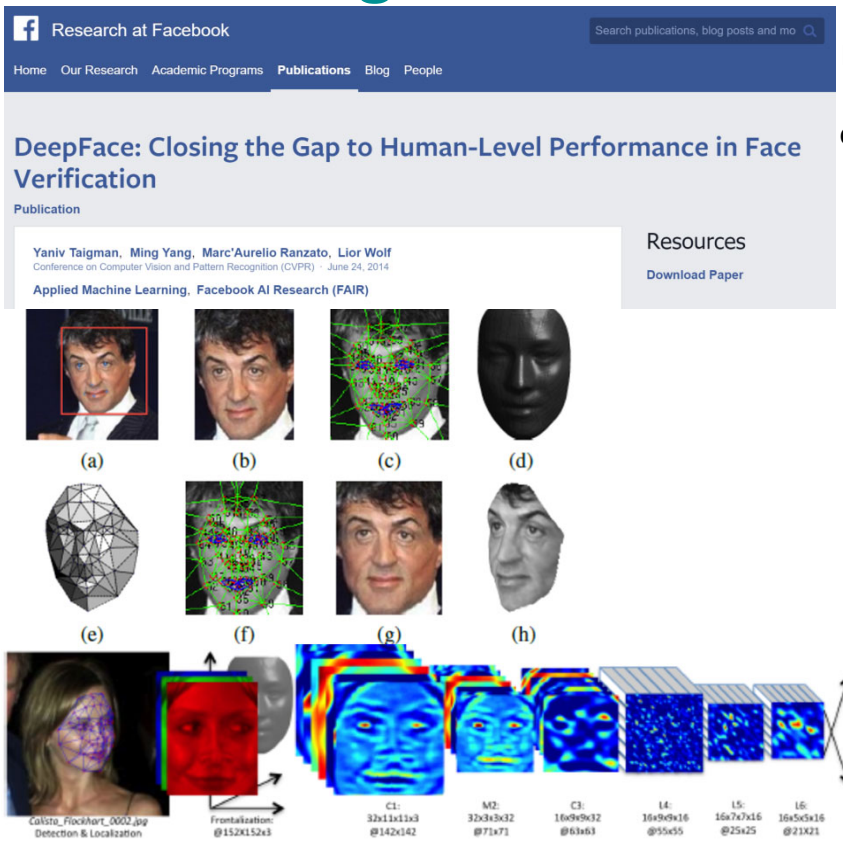
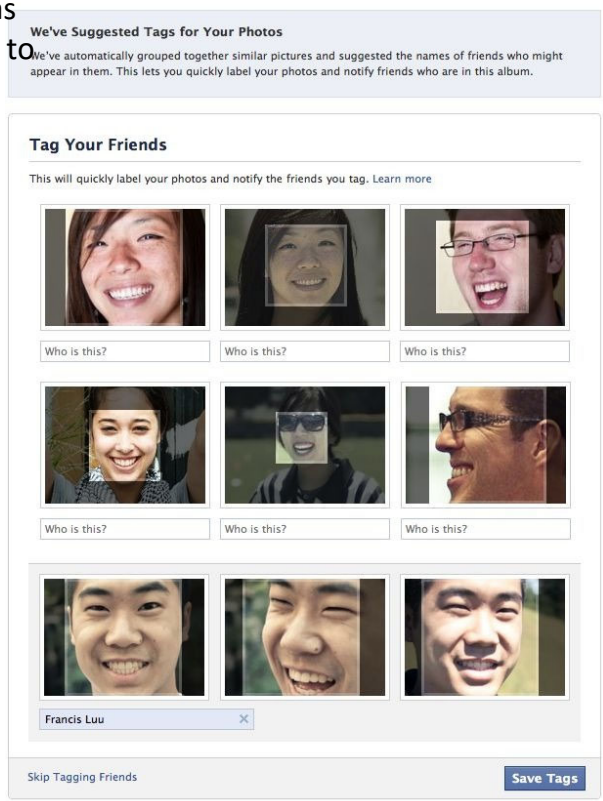


Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

https://research.facebook.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/

Facebook's DeepFace facial recognition software has over 97% accuracy, close to a human (~97.5%)



http://mashable.com/2010/12/15/facebook-photo-tag-suggestions/#dBJbfcc3LEqW

Base Case: Fairness Through Unawareness

- The default fairness method in machine learning is **fairness-through-unawareness** (Kusner *et al,* 2017; Chen *et al*, 2019)
- Fairness-through-unawareness refers to leaving out of the model protected social attributes such as gender, race, and other characteristics deemed sensitive. In ethics, the **color-blind perspective** is characterized by a belief that people should be treated equally no matter their cultural background, thus prejudice can be eliminated by ignoring or avoiding sensitive attributes (Podsiadlowski et al, 2013; Wolsko et al, 2000). By downplaying potential differences in data due to different backgrounds, the potential for bias is thought to be limited. This ends up perpetuating bias

Failures of Fairness through Unawareness

- Researchers at Carnegie Mellon University revealed that **gender**, a protected attribute, caused an unintentional change in Google's **advertising system** such that ad listings targeted for users seeking high-income jobs were presented to men at nearly six times the rate they were presented to women (Datta *et al.*, 2015). In this scenario, it is possible Google's fairness-through-unawareness approach failed due to oversampling of majority category members (i.e., males) in the data.

Failures of fairness through Unawareness (2)

- Xerox similarly removed race from its hiring algorithm without realizing that commuting distance, which remained in the dataset, was highly correlated with poverty level (O'Neil, 2016: 118-119). The company did self-correct this error eventually.

- In a well publicized recent example of an automated resume parsing tool by Amazon (Medhora, 2018), Amazon found certain keywords to be discriminatory against female applicants, even without any codification of gender. ➔ **Tainted training data will give you a tainted outcome**

- **This is also found in the popular word embeddings method (for example, woman is to homemaker what man is to programmer, which is inherently unfair – see Bolukbasi *et al*, 2016)**

Failures of fairness through Unawareness (3)

- One common problem with the fairness-through-unawareness approach is that the **attributes supposedly kept out of the algorithm's knowledge are in fact encoded into the other unprotected features, leading to discrimination.**

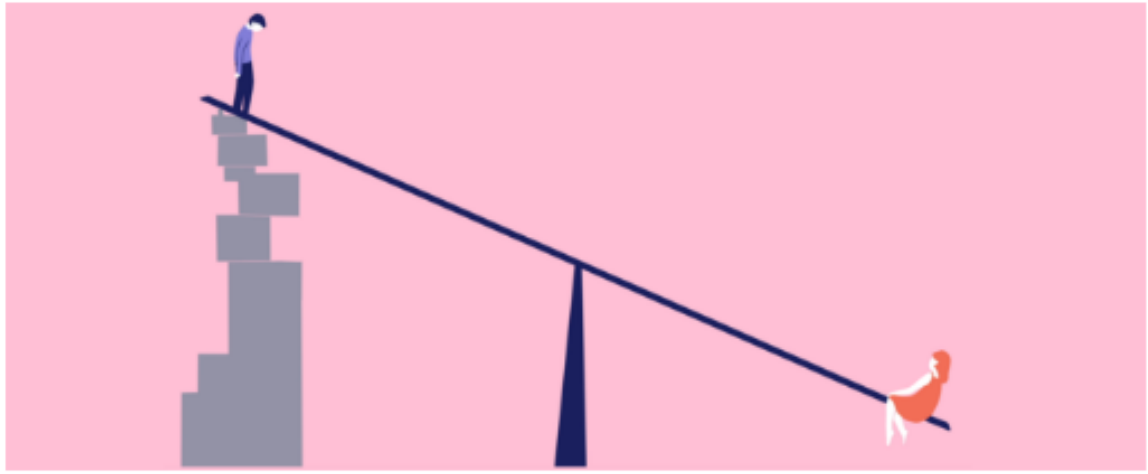
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Man is to Doctor as Woman is to Nurse: the Gender Bias of Word Embeddings

Why we should worry about gender inequality in Natural Language Processing techniques

Tommaso Buonocore Follow
Mar 8 · 9 min read ★



https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17

Demographic Parity

- Demographic parity is what we call a "group level fairness" criterion, where the decision – for example hiring an applicant – is independent of the protected attribute

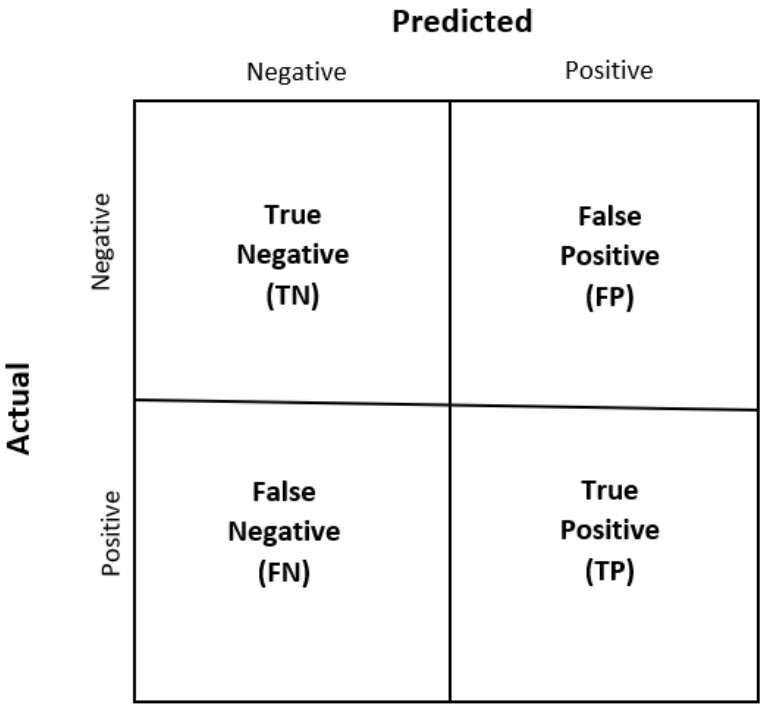
$$P(\hat{Y}|A = a) = P(\hat{Y}|A = a'), \hat{Y} \perp A$$

- Example: probability of being hired is independent of gender

(Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. doi:10.31235/osf.io/ustxg)

- However, while enforcing group level fairness (say, same hiring rate for females and males), this can be unfair to the individual: it could force the algorithm to drop otherwise qualified individuals just to achieve equal hiring rates across the two groups.

A 2 by 2: The Confusion Matrix



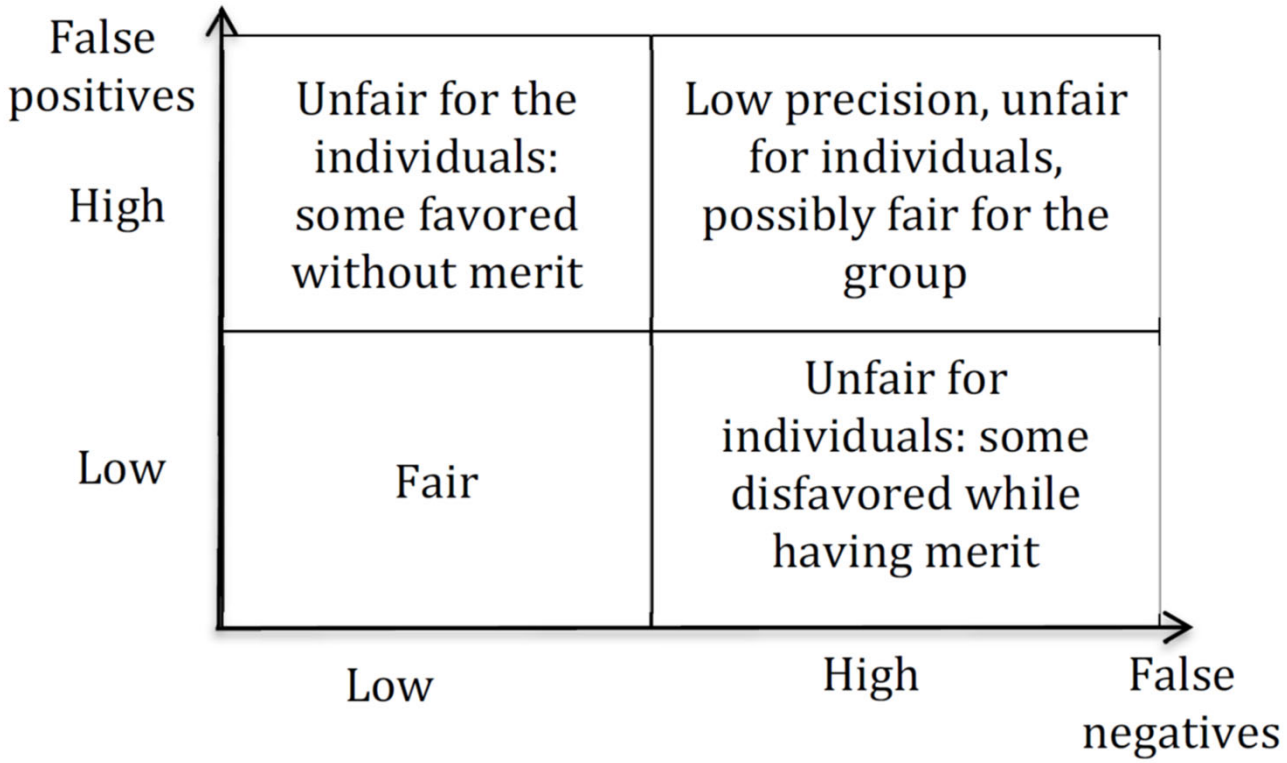
$$Recall = TPR = \frac{TP}{TP+FN}$$ *(how many relevant positives were missed? Important if FN cost is high)*

$$Specificity = TNR = \frac{TN}{TN+FP}$$ *(how many of the predicted false are actually false?)*

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Fairness at the individual or group level



Equalized Odds

- Equalizing the odds = matching the True Positive Rate and False Positive Rate for different values of the protected attribute (Hardt *et al,* 2016)

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = y\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = y\right\}, \quad y \in \{0, 1\}$$

(Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems.)

- This is hard to do but if achieved is one of the highest levels of algorithmic fairness
- And there are many different criteria which an organization has to choose from

Equalized Opportunity

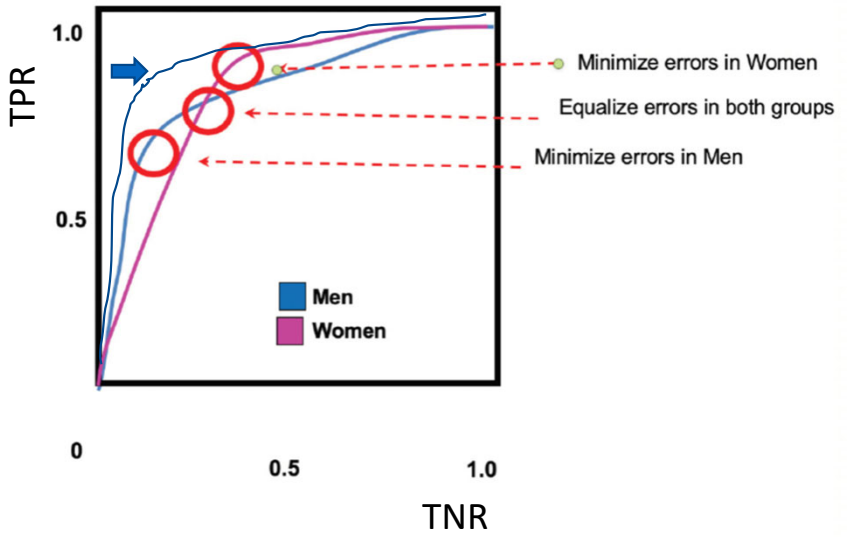
- A more restricted case of equality of odds looks only to equalize one side of the outcome—only those given a positive answer are a fairness concern (Hardt *et al,* 2016)

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=1\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=1\right\}$$

(Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems.)

- Example: in an admission process we ensure admitted students are treated equally across demographics but not concerned with equality in rejections. Because it is less restrictive than equality of odds, it should be preferred in cases where one is indifferent to equality in the negatively classified set.

Automation isn't a panacea for fairness in ML – Oftentimes we tradeoff accuracy for fairness



Prior work in USAID ML grant (which also supported this paper): Awwad, Y.; Fletcher, R.; Frey, D.; Gandhi, A.; Najafian, M.; Teodorescu, M. (alphabetical) 2020. Exploring Fairness in Machine Learning for International Development. MIT D-Lab | CITE Report. Cambridge: MIT D-Lab.

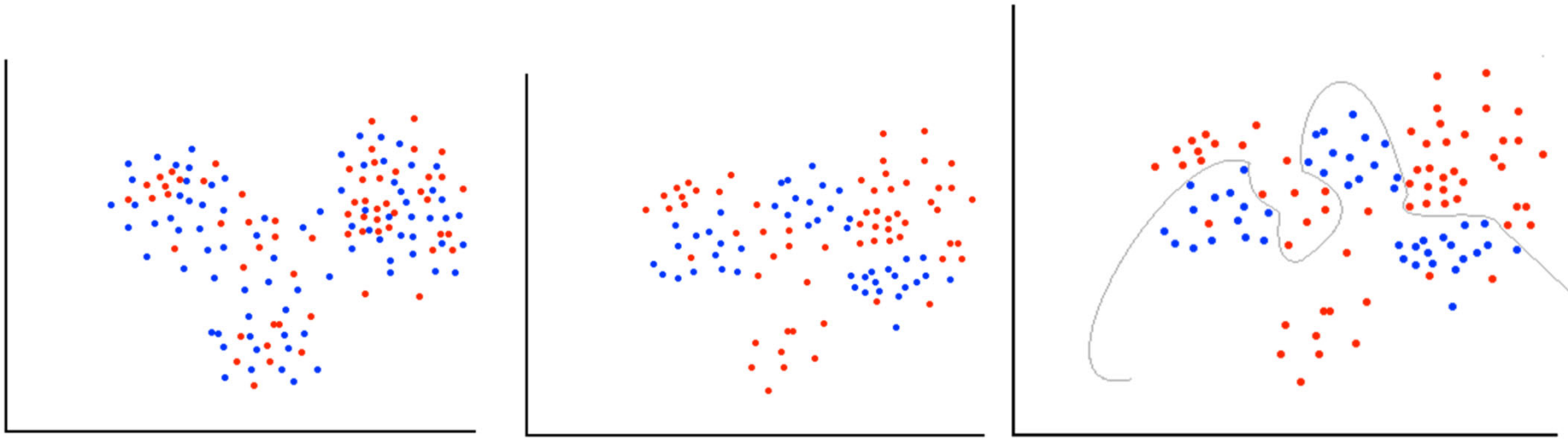
Funding: USAID-MIT Grant AID-OAA-A-12-00095 "Appropriate Use of Machine Learning in Developing Country Contexts"

MITD-Lab USAID CITE

Problem Gets Worse With Multiple Groups

- Satisfying a criterion for one protected attribute may be mutually exclusive with the satisfaction of a criterion for another protected attribute, leaving the algorithm designer to choose which attribute to optimize on as there often is no “globally fair” solution to satisfy equality of odds on all protected attributes
- Auditing fairness can be computationally infeasible if looking at combinations of groups (“the computational problem of auditing subgroup fairness for both equality of false positive rates and statistical parity […] **is computationally hard in the worst case, even for simple structured subclasses**”) – Kearns et al 2017
- Group fairness fails under composition (“ if all advertisers in an advertising system **independently satisfy Equalized Odds** (Hardt et al., 2016), **does the entire advertising system have the same guarantee?** Our first result is that naıve composition of group-fair classifiers **will not in general yield a group-fair system**" - Dwork & Ivento 2018)

Visualizing the Attributes

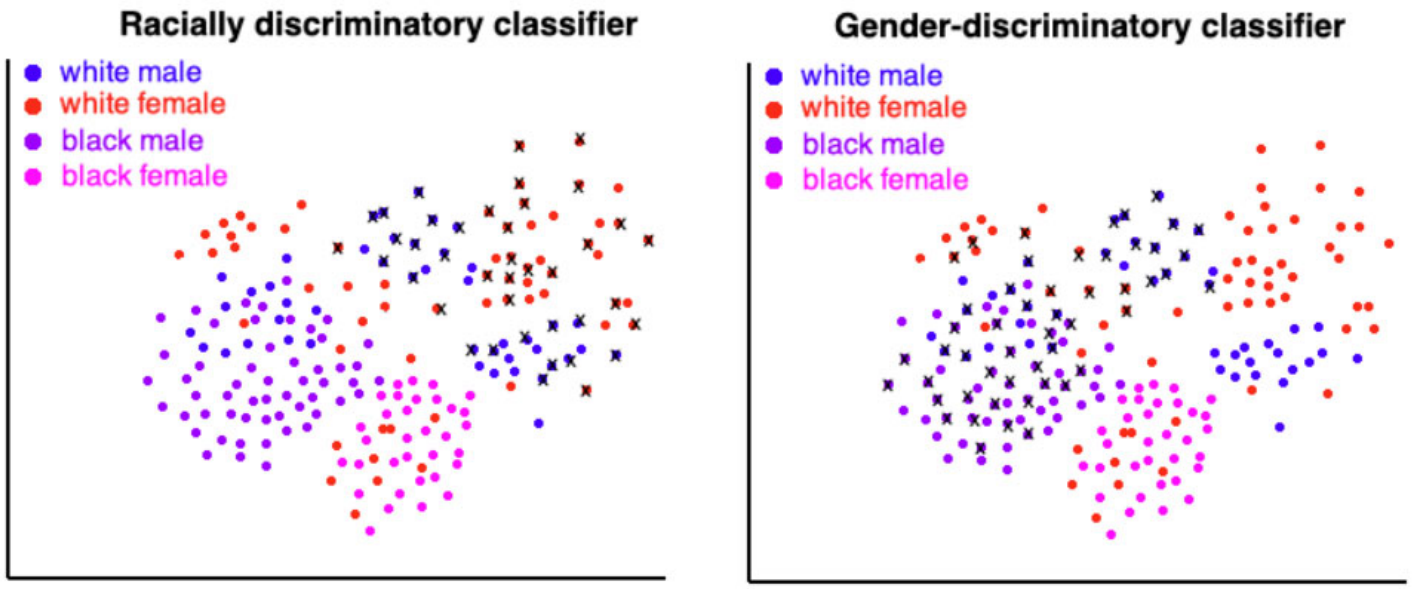


Hard to discriminate by gender:
No Gender Disparity

Problematic: Can discriminate

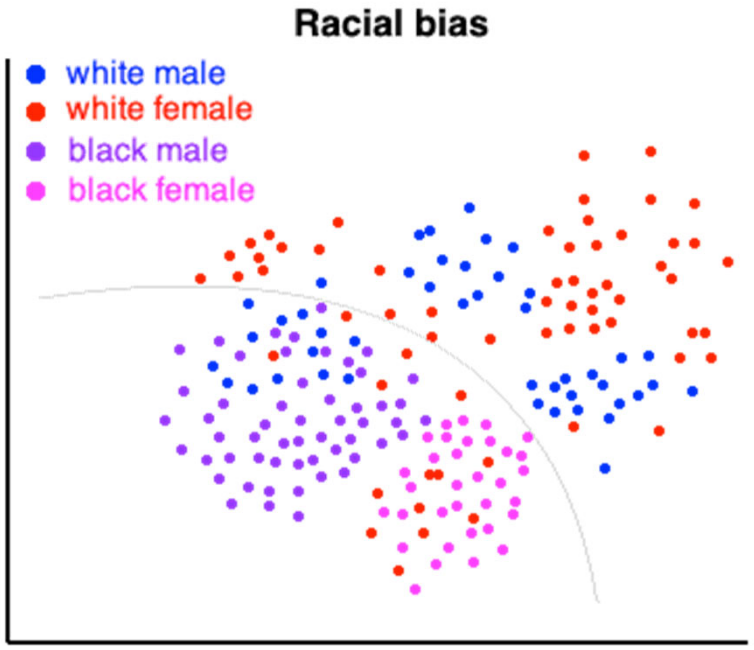
Problematic: Can discriminate

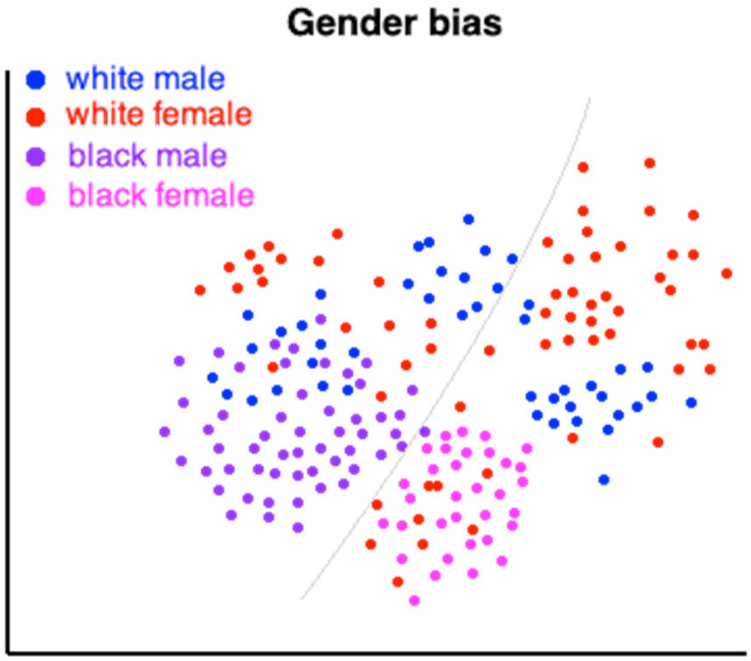
Unfair Classifiers



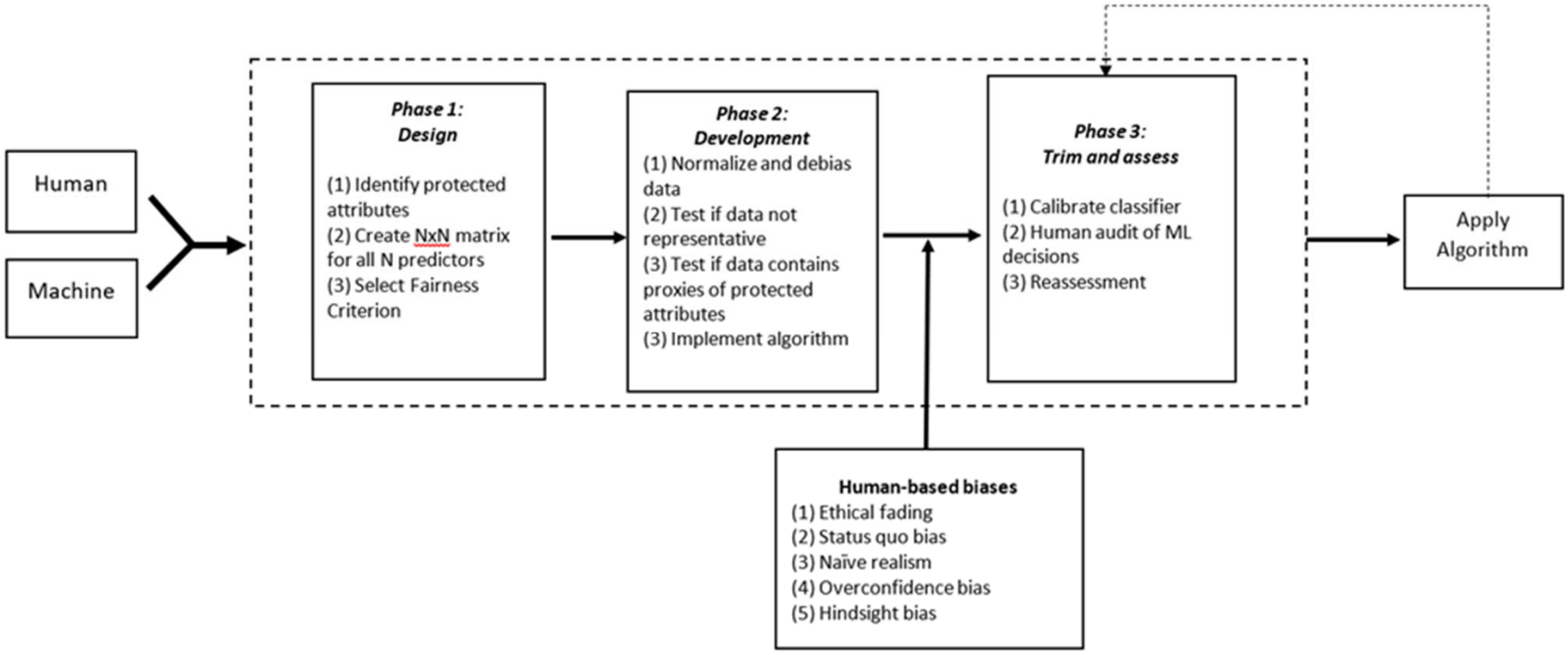
A non-discriminatory classifier avoids dividing the attribute network, instead showing a more random scatter

Unfair Classifiers





“A Framework for Fairer Machine Learning in Organizations” – M. Teodorescu, L. Morse, Y. Awwad, G. Kane, R&R MISQ, 2019/2020



Phase I: Design

Identify Protected Attributes

- The first step is to determine **which variables in an organization's hiring dataset constitute protected attributes** and whether they are the same as the legally-protected demographic features (i.e., race, gender, age, sexual orientation, disability, marital status, ethnicity), or whether these protected attributes are codified in the regulations for a specific business sector. It is the organization's responsibility to establish the baseline protected attributes.

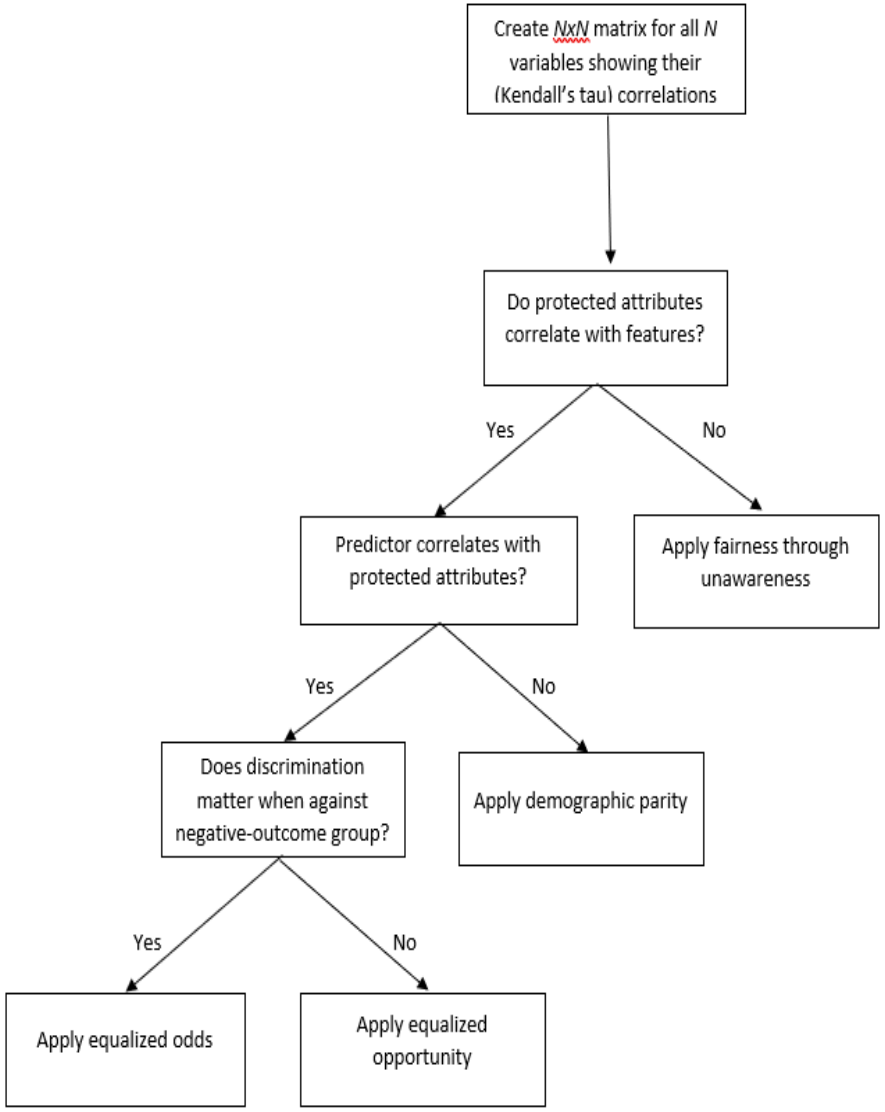
Phase I: Design (continued)

Create NxN Matrix of Predictors

- A given predictor in the dataset may correlate with one or several protected attributes, which may bias a machine learning model toward negative hiring outcomes for a protected demographic group. If a protected attribute strongly correlates with one or more predictors, efforts should be undertaken by the organization to remove multicollinearity issues. However, there is a tradeoff between eliminating *all* predictors correlated with protected attributes versus optimizing model performance.
- If there are high correlation values between the rows/columns representing protected attributes and those representing features, then those features should be treated with caution at the very least, or simply treated as protected attributes.

Phase I: Design (continued)

Select Algorithmic Fairness Criterion



Phase II: Development

Normalize and Debias Data

- A first step in developing a machine learning approach is to analyze the data to ensure **likelihood of sampling bias is minimized**, as well as normalizing the data. The data **may not be representative of all groups**, which can affect the proper use of any fairness criterion. As such, ensuring there is no sampling bias in the data is an essential step.

Implement Algorithm

- This step is left to the developers of the firm and involves use of either the firm's in-house capabilities for custom solutions or reuse of existing implementations (APIs).

Phase III: Post-Model Assessment

Calibrate Classifier: A Posteriori Tests for Fairness

- In the organizational hiring setting, **a posteriori "false positives", such as employees hired but who were not high performing, are easy to determine, but not "false negatives",** such as missed hiring opportunities, as employees who were misclassified as low quality and not hired. False negatives could be determined by looking at denied applicants and their employment trajectory post-denial at competitors (de Cuerto, 2012: 33).

Phase III: Post-Model Assessment (2)

Human Audit Committee

- We propose that a select organizational members come together to evaluate, analyze, and prepare advice regarding the quality and appropriateness of the machine learning algorithm's output by forming a **fairness "audit committee".**

- Similar to existing structures of external audit committees (see SEC, 2016), the fairness committee should consist of three or more qualified employees who have adequate expertise in employment/hiring (measured by previous experience in hiring) as well as a strong grasp of machine learning fairness principles and processes.

Phase III: Post-Model Assessment (3)

Reassessment

- The accuracy of an algorithm can change over time with new data, including in the case of hiring changes in skills of the applicant population, changes in needs of the company, and broadly socio-economic changes in the applicant pool which can affect the associations an algorithm may make.
- **A periodic reassessment of the fairness of the algorithm is therefore recommended** and is illustrated by a recursive step from applying the algorithm in production back to auditing the algorithm (phase 3 of our framework).

Why does unfairness happen? Ethical Fading

- Why do organizations frequently end up making unfair hiring decisions without realizing they have deviated from their values? To answer this question, we draw on a key concept in the behavioral ethics literature known as **ethical fading.**

- Ethical fading is defined as the "process when the moral colors of an ethical decision fade into leached hues that are void of moral implications" (Tenbrunsel and Messick, 2003: p. 224).

- Ethical fading occurs in the hiring process when organizational members simply do not "see" ethics as being a part of the relevant decision criteria.

- **Machine learning may fade the ethical implications of human actions**, leading individuals to be driven **not as much by values but by pragmatic concerns**.

Acknowledgements

- Thank you to the organizers – Professors Rob Seamans and Raj Choudhury, as well as to NYU and Microsoft for funding this workshop!
- This is joint work with Professors Lily Morse and Gerald Kane (Boston College) and Yazeed Awwad (MIT).
- Separate study with colleagues Marios Kokkodis and Naylia Ordabayeva (Boston College)
- USAID-MIT Grant AID-OAA-A-12-00095 “Appropriate Use of Machine Learning in Developing Country Contexts” and the Carroll School of Management at Boston College for funding
- Currently paper in R&R at MIS Quarterly – my thanks to the co-editors and reviewers for feedback
- **Happy to collaborate with other participants on these topics!**

References

- Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pp. 3315-3323.
- Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D. and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Advances in Neural Information Processing Systems, pp. 656-666.
- Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems, pp. 3315-3323.
- Piech, Chris. Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction. 2018
- Pleiss, G., et al. On Fairness and Calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30 , pp. 5684–5693.
- Verma, S. and Rubin, J., 2018, May. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1-7.

References

- Abdi, H. (2007). The Kendall rank correlation coefficient. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, pp.508-510.
- Agrawal, A., Gans, J. and Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Press, 195-206.
- Ajunwa, Ifeoma, The Paradox of Automation as Anti-Bias Intervention (forthcoming). Cardozo Law Review.
- Angst, C., Agarwal, R. (2009). Adoption of Electronic Health Records in the Presence of Privacy Concerns: The Elaboration of Likelihood Model and Individual Persuasion. *MIS Quarterly*, *33*, 339-370.
- Angst, C. (2009). Protect My Privacy or Support the Common-Good? Ethical questions about electronic health information exchanges. *Journal of Business Ethics, 90*, 169-178.
- Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In Blind Pursuit of Racial Equality? Psychological Science, 21(11), 1587-1592.
- Ashcraft, C., McLain, B. and Eger, E. (2016). Women in tech: The facts. National Center for Women & Technology (NCWIT).

References

- Bazerman, M. H., & Tenbrunsel, A. E. (2011). Blind Spots: Why We Fail to Do What's Right and What to Do about It. Princeton, N.J.: Princeton University Press.
- Byrne, E. F. (1995). The 2-Tiered Ethics of EDP. *Journal of Business Ethics*, *14*(1), 53-61.
- Benda, B.B., Corwyn, Robert F., Toombs, N.J. (2001). Recidivism among adolescent serious offenders Prediction of Entry into the Correctional System for Adults, Criminal Justice and Behavior, Vol. 28 No. 5, October 2001: 588-613.
- Bella, A., Ferri, C., Hernández-Orallo, J. and Ramírez-Quintana, M.J., 2010. Calibration of machine learning models. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques (pp. 128-146). IGI Global.
- Bradfield, A., & Wells, G. L. (2005). Not the same old hindsight bias: Outcome information distorts a broad range of retrospective judgments. Memory Cognition, 33, 120-130.
- Bonta, James, Law, Moira, Hanson, Karl (1988). The Prediction of Criminal and Violent Recidivism Among Mentally Disordered Offenders: A Meta-Analysis, Psychological Bulletin, Vol 123, No. 2, 1998: 123-142.
- Bonta, James (2002). Offender risk assessment Guidelines for Selection and Use, Criminal Justice and Behavior, Vol. 29 No. 4, Aug 2002: 355-379.

References

- Brennan, T., Dieterich, W., Ehret, B. (2009). Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System", Criminal Justice And Behavior, Vol. 36 No. 1, January 2009: 21-40.
- Chen I., Johansson F.D., and Sontag D. (2018). Why is my classifier discriminatory? arXiv preprint arXiv:1805.12002, 2018. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.
- Chen, J., Kallus, N., Mao, X., Svacha, G. and Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 339-348). ACM.
- Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NIPS 2017), 2017: 5036–5044.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A. (2017). Algorithmic decision making and the cost of fairness, Proceedings of KDD '17, Halifax, NS, Canada, Aug 13-17 2017.

References

- Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample selection bias correction theory. In: Proceedings of the 19th annual conference on algorithmic learning theory. Oct 13 2008, Budapest, Hungary, pp. 38-53.
- David, H., Katz, L.F. and Kearney, M.S. (2006). The polarization of the US labor market. American Economic Review, 96(2), pp.189-194.
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Business News, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed May 21, 2019)
- Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies, 1, 92-112.
- Datta, A., Fredrikson, M., Ko, G., Mardziel, P., Sen, S. 2017. Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120.
- de Cuerto, F. J. (2012). Relative Importance of False Positives in the Selection Process. FIU Electronic Theses and Dissertations. 569. http://dx.doi.org/ 10.25148/etd.FI12041905. (Accessed August 8, 2019).

References

- Dearborn, D.C., & Simon, H.A. (1958). Selective perception: A note on the departmental identifications of executives. Sociometry, 21, 140-144.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. ITCS ’12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226. Cambridge, Massachusetts — Jan 08 - 10, 2012
- Ely, R.J., & Thomas, D.A. (2001). Cultural Diversity at Work: The Effects of Diversity Perspectives on Work Group Processes and Outcomes. Administrative Science Quarterly, 46(2), 229-273.
- Gao, J., Burnicki, A.C. & Burt, J.E. (2016). Bias-variance decomposition of errors in data-driven land cover change modeling. Landscape Ecol (2016) 31: 2397, December 2016, Vol 31, Issue 10, pp 2397–2413.
- Ghassami, A. (2018). Fairness in Supervised Learning: An Information Theoretic Approach. IEEE International Symposium on Information Theory (ISIT 2018). doi:10.1109/isit.2018.8437807.
- Mann, G., O'Neil, C. (2016). Hiring Algorithms Are Not Neutral. Harvard Business Review, December 09, 2016. https://hbr.org/2016/12/hiring-algorithms-are-not-neutral

References

- Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. WWW, 2018.
- Friedman, J., Hastie, T. and Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York, NY, USA: Springer Series in Statistics.
- Hardt M., Price E., and Srebro N. (2016). Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems 29: NIPS 2016.
- Henwood, F., Plumeridge, S. and Stepulevage, L. (2002). A tale of two cultures? Gender and inequality in computer education. In Technology and In/equality (pp. 123-140). Routledge.
- Jansen, W., Vos, M., Otten, S., Podsiadlowski, A., van der Zee, K. (2015). Colorblind or colorful? How diversity approaches affect cultural majority and minority employees. Journal of Applied Social Psychology, 46, 81-93.
- Kearns, M. J., Neel, S. Roth, A., and Wu Z.S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Proceedings of the 35th International Conference on Machine Learning, ICML ’18. PMLR, 2018.

References

- Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D. and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In Advances in Neural Information Processing Systems, pp. 656-666.
- Kusner, M.J., Loftus, J.R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Annual Conference on Neural Information Processing Systems: NIPS 2017.
- Ladin, K., & Hanto, D. W. (2011). Rational Rationing or Discrimination: Balancing Equity and Efficiency Considerations in Kidney Allocation. American Journal of Transplantation, 11, 2317-2321.
- Markus, H. R., Steele, C. M., & Steele, D. M. (2000). Colorblindness as a barrier to inclusion: Assimilation and non-immigrant minorities. Daedalus, 129, 233-259.
- Marsland, S. (2014). Machine learning: an algorithmic perspective. Chapman and Hall/CRC.
- Medeiros K., Mecca J., Gibson C., Giorgini V., Mumford, M., Devenport L., Connelly, S. (2014). Biases in Ethical Decision Making among University Faculty, Accountability in Research, Vol 21(4): 218-240.

References

- Moore, D. A., & Healy, P. J. (2008). The Trouble with Overconfidence. Psychological Review, 115(2), 502-517.
- Molinara, M., Ricamato, M.T. and Tortorella, F. (2007). Facing imbalanced classes through aggregation of classifiers. In 14th international conference on image analysis and processing (ICIAP 2007), IEEE Proceedings pp. 43-48.
- Neuilly, M.-A., Zgoba, K.M., Tita, G.E., Lee, S (2011). Predicting Recidivism in Homicide Offenders Using Classification Tree Analysis, Homicide Studies 15(2), 2011: 154–176.
- O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.
- Podsiadlowski, A., Groschke, D., Kogler, M., Springer, C., and van der Zee, K. (2013). Managing a culturally diverse workforce: Diversity perspectives in organizations. International Journal of Intercultural Relations, 37(2), 159-175.
- Podsiadlowski, A., Otten, S., and van der Zee, K. (2009). Diversity perspectives. In Symposium on workplace diversity in Groningen, The Netherlands.

References

- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. and Weinberger, K.Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems, pp. 5680-5689.
- Reisig, M. D., Holtfreter, K., Morash, M. (2006), Assessing Recidivism Risk Across Female Pathways to Crime, Justice Quarterly, Vol. 23, No. 3, Sep 2006: 384-405.
- Richeson, J.A., Nussbaum, R.J. (2004). The impact of multiculturalism versus colorblindness on racial bias. Journal of Experimental Social Psychology, 40, 417-423.
- Rohinton P. M. (2018). AI & Global Governance: Three Paths Towards a Global Governance of Artificial Intelligence. UN University, UNU Office of Communications, Articles & Insights, October 28, 2018. https://cpr.unu.edu/ai-global-governance-three-paths-towards-a-global-governance-of-artificial-intelligence.html (Accessed June 10, 2019)
- Roese, N. J., Vohs, K. D. (2012). Hindsight Bias. Perspectives on Psychological Science, 7(5), 411-426.
- Rosenthal, L., Levy, S. R. (2010). The Colorblind, Multicultural, and Polycultural Ideological Approaches to Improving Intergroup Attitudes and Relations. *Social Issues and Policy Review*, 4(1), 215-246.

References

- Ryan, C. S., Hunt, J. S., Weible, J. A., Peterson, C. R., and Casas, J. F. (2007). Multicultural and Colorblind Ideology, Stereotypes, and Ethnocentrism among Black and White Americans. Group Processes & Intergroup Relations, 10(4), 617-637.
- Samuelson, W., Zeckhauser, R. (1988). Status Quo Bias in Decision Making. Journal of Risk and Uncertainty, 1, 7-59.
- Serra, J. (2018) “Unintuitive Properties of Deep Learning Networks.” Telefonica Research, Barcelona. (March 2018).
- Simonite, T. (2017). How to Upgrade Judges with Machine Learning, MIT Technology Review.
- Sommers, S.R. (2006). On racial diversity and group decision making: Identifying multiple effects of racial composition on jury deliberations. Journal of Personality and Social Psychology, 90, 597-612.
- Tommasi, T., Patricia, N., Caputo B. (2017). "A Deeper Look at Dataset Bias". Domain Adaptation in Computer vision Applications. Springer, Cham, 2017. 37-55
- Tucker, I. “'A white mask worked better': why algorithms are not colour blind.” The Guardian. May 28, 2017

References

- Van Boven, L. (2007). Naïve Realism. In R. F. Baumeister & K. D. Vohs (Eds.), Encyclopedia of Social Psychology (pp. 603-603). Thousand Oaks, CA: Sage.
- Van Oudenhoven, J. P., Prins, K. S., & Buunk, B. P. (1998). Attitudes of minority and majority members towards adaptation of immigrants. European Journal of Social Psychology, 28(6), 995-1013.
- Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. doi:10.31235/osf.io/ustxg
- Wadsworth C., Vera F., Piech C. (2018). Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction. CoRR abs/1807.00199.
- Wolsko, C., Park, B., Judd, C. M., & Wittenbrink, B. (2000). Framing interethnic ideology: Effects of multicultural and color-blind perspectives on judgments of groups and individuals. Journal of Personality and Social Psychology, 78, 635-654.
- CFPB Consumer Laws and Regulations. ECOA. June 2013 p. 6 https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf

Thank you

Dr Mike Teodorescu
Information Systems, BC
D-Lab, MIT

teodores@bc.edu
hmteodor@mit.edu