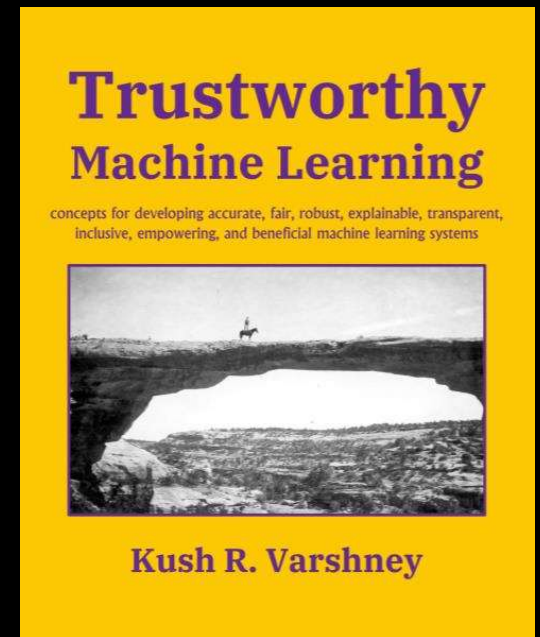


Carative AI

—

Kush R. Varshney
Distinguished Research Scientist and Manager

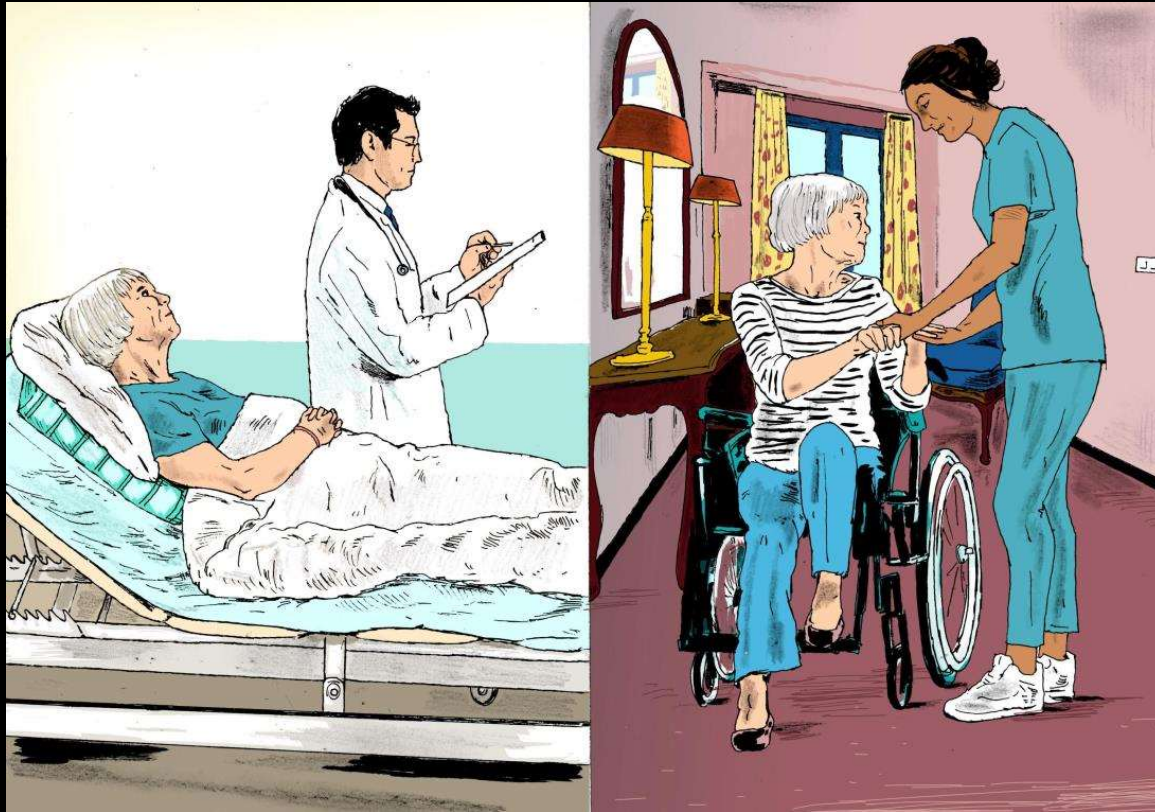
@krvarshney
krvarshn@us.ibm.com



Research



Curing and caring

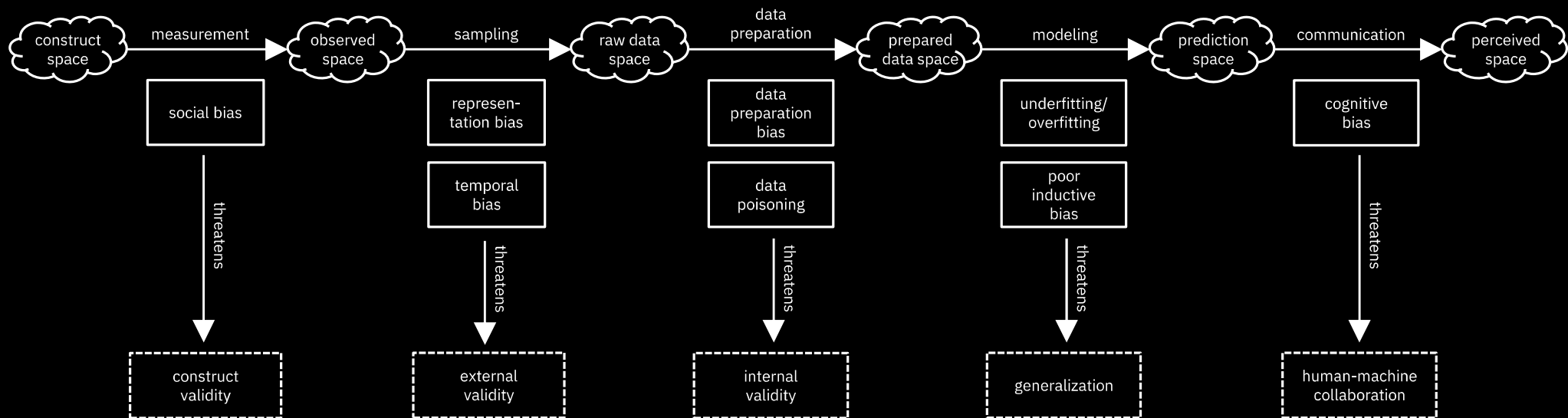


<https://www.wsj.com/articles/treating-disease-is-no-substitute-for-caring-for-the-ill-11575047264>

Determining an appropriate fairness metric

Kush R. Varshney. *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

A basic model of the world (the doctor's approach)



Determining an appropriate fairness metric

Kush R. Varshney. *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

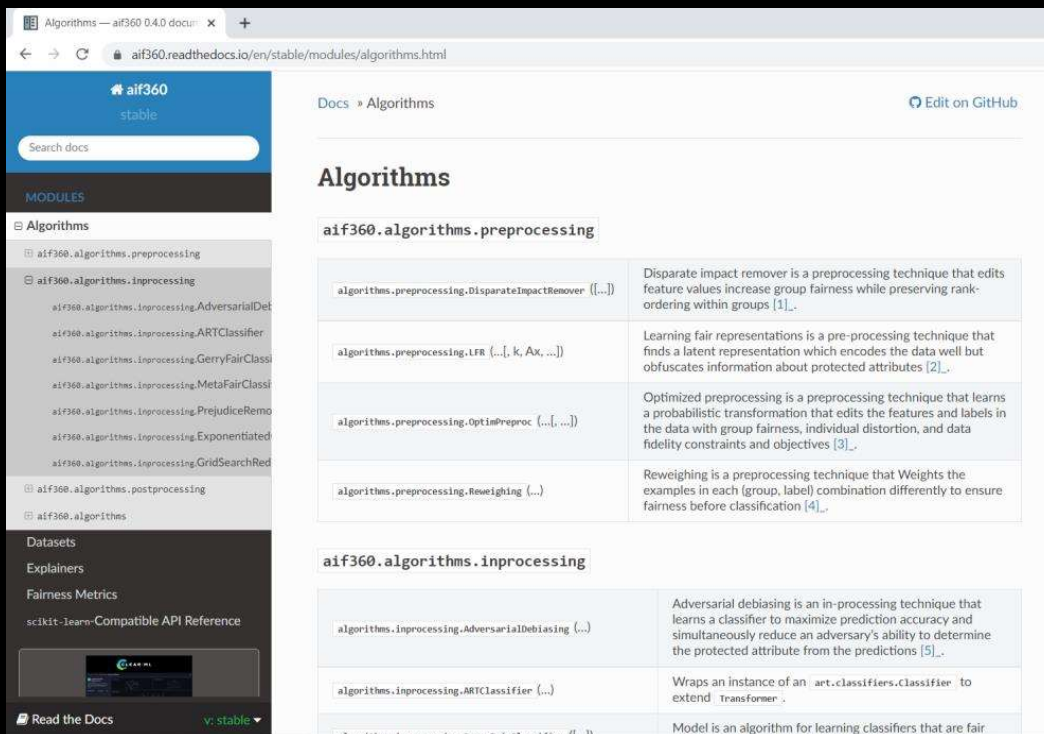
Guidelines that 'make sense' (the doctor's approach)

Type	Statistical Relationship	Fairness Metric	Social Bias in Measurement	Favorable Label
independence	$\hat{Y} \perp\!\!\!\perp Z$	statistical parity difference	yes	assistive or non-punitive
separation	$\hat{Y} \perp\!\!\!\perp Z \mid Y$	average odds difference	no	assistive
sufficiency (calibration)	$Y \perp\!\!\!\perp Z \mid \hat{Y}$	average predictive value difference	no	non-punitive

Mitigating unwanted biases

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias." *IBM Journal of Research and Development* 63.4/5 (Jul./Sep. 2019), p. 4.

Tools that are 'blessed' (the doctor's approach)



The screenshot shows the aif360 documentation website. The left sidebar contains a search bar and a navigation menu with categories: Algorithms, Datasets, Explainers, and Fairness Metrics. The main content area is titled "Algorithms" and lists several preprocessing and in-processing techniques with their descriptions.

aif360.algorithms.preprocessing	
<code>algorithms.preprocessing.DisparateImpactRemover (...)</code>	Disparate impact remover is a preprocessing technique that edits feature values increase group fairness while preserving rank-ordering within groups [1].
<code>algorithms.preprocessing.LFR (...)</code>	Learning fair representations is a pre-processing technique that finds a latent representation which encodes the data well but obfuscates information about protected attributes [2].
<code>algorithms.preprocessing.OptimPreproc (...)</code>	Optimized preprocessing is a preprocessing technique that learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives [3].
<code>algorithms.preprocessing.Reweighting (...)</code>	Reweighting is a preprocessing technique that Weights the examples in each (group, label) combination differently to ensure fairness before classification [4].

aif360.algorithms.inprocessing	
<code>algorithms.inprocessing.AdversarialDebiasing (...)</code>	Adversarial debiasing is an in-processing technique that learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions [5].
<code>algorithms.inprocessing.ARTClassifier (...)</code>	Wraps an instance of an <code>art.classifiers.Classifier</code> to extend <code>Transformer</code> .
<code>algorithms.inprocessing.GerryFairClassifier (...)</code>	Model is an algorithm for learning classifiers that are fair

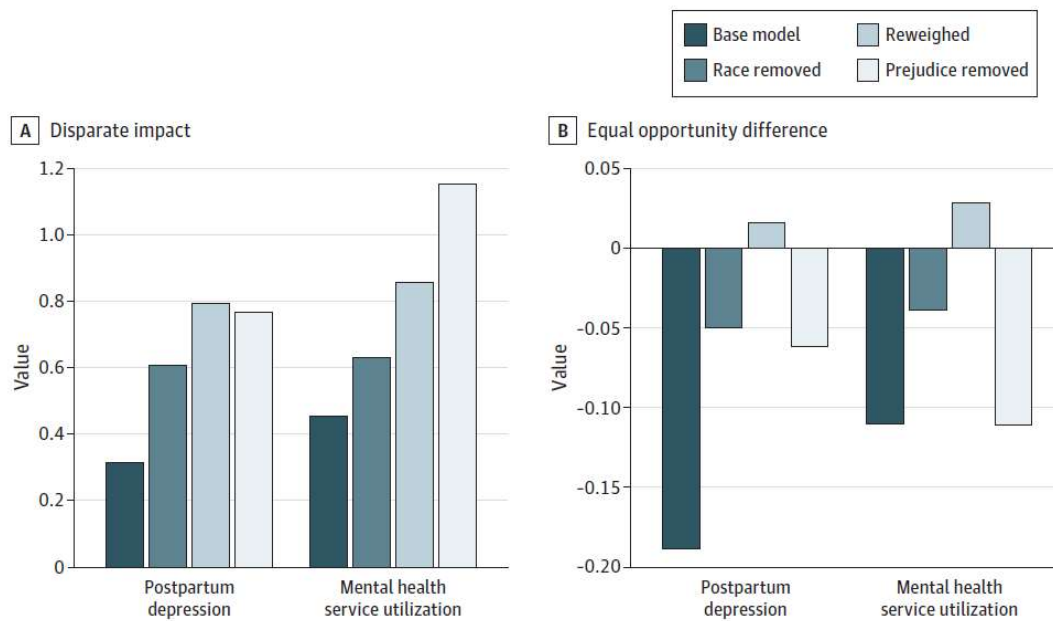
Something's missing

**What about
looking into the
patient's eyes?**

Clinical prediction of postpartum depression

Yoonyoung Park, Jianying Hu, Moninder Singh, Issa Sylla, Irene Dankwa-Mullan, Eileen Koski, and Amar K. Das. "Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression." *JAMA Network Open* 4.4 (Apr. 2021), p. e213909.

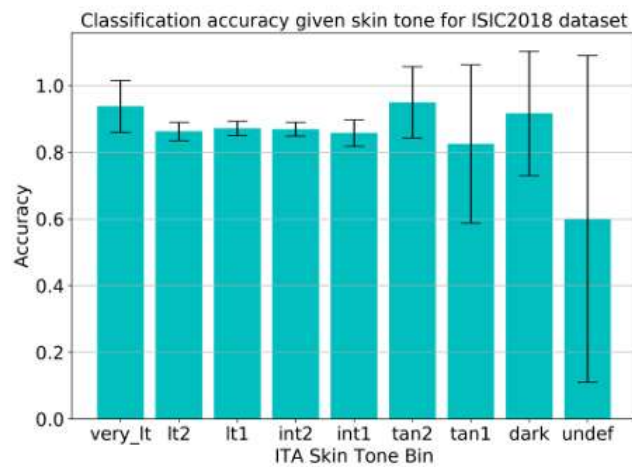
Figure 1. Comparison of Bias Metrics Before and After Debiasing



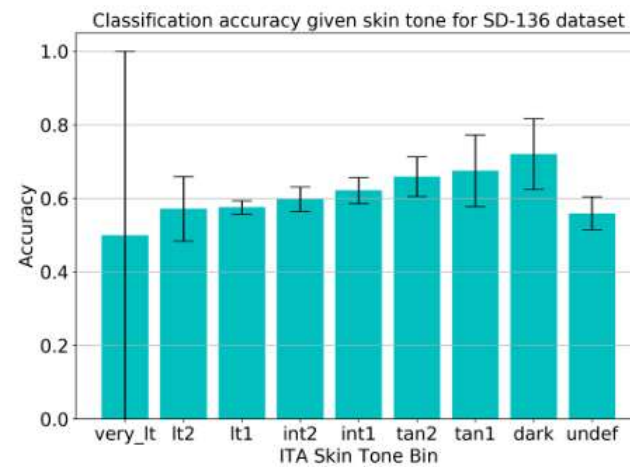
Comparison of bias metrics in the test data set using base model, model without race variable, debiased model through reweighing, and debiased model through Prejudice Remover (logistic regression). The reference value for unbiasedness is 1.0 for disparate impact (A) and 0 for equal opportunity difference (B).

Skin disease diagnosis

Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney, "Fairness of Classifiers Across Skin Tones in Dermatology." In: *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2020, p. 320-329.



(a)

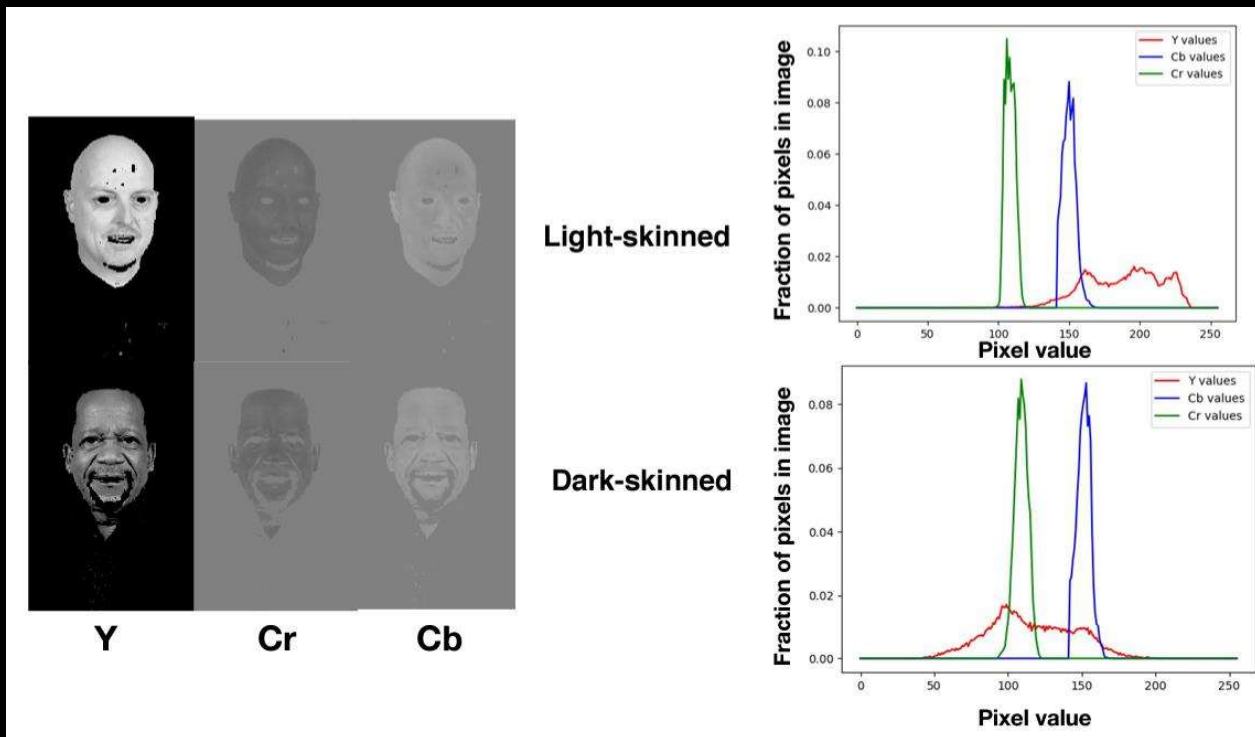


(b)

Fig. 4. Accuracy versus ITA for (a) ISIC2018, and (b) SD-136 validation sets.

Face attribute classification

Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilović, and Kush R. Varshney. "Understanding Unequal Gender Classification Accuracy from Face Images." arXiv:1812.00099, 2018.



Legal financial obligations in Jefferson County, Alabama

<https://www.ibm.com/blogs/journey-to-ai/2022/03/ibm-teams-up-with-organizations-on-ai-incubator-for-social-impact/>



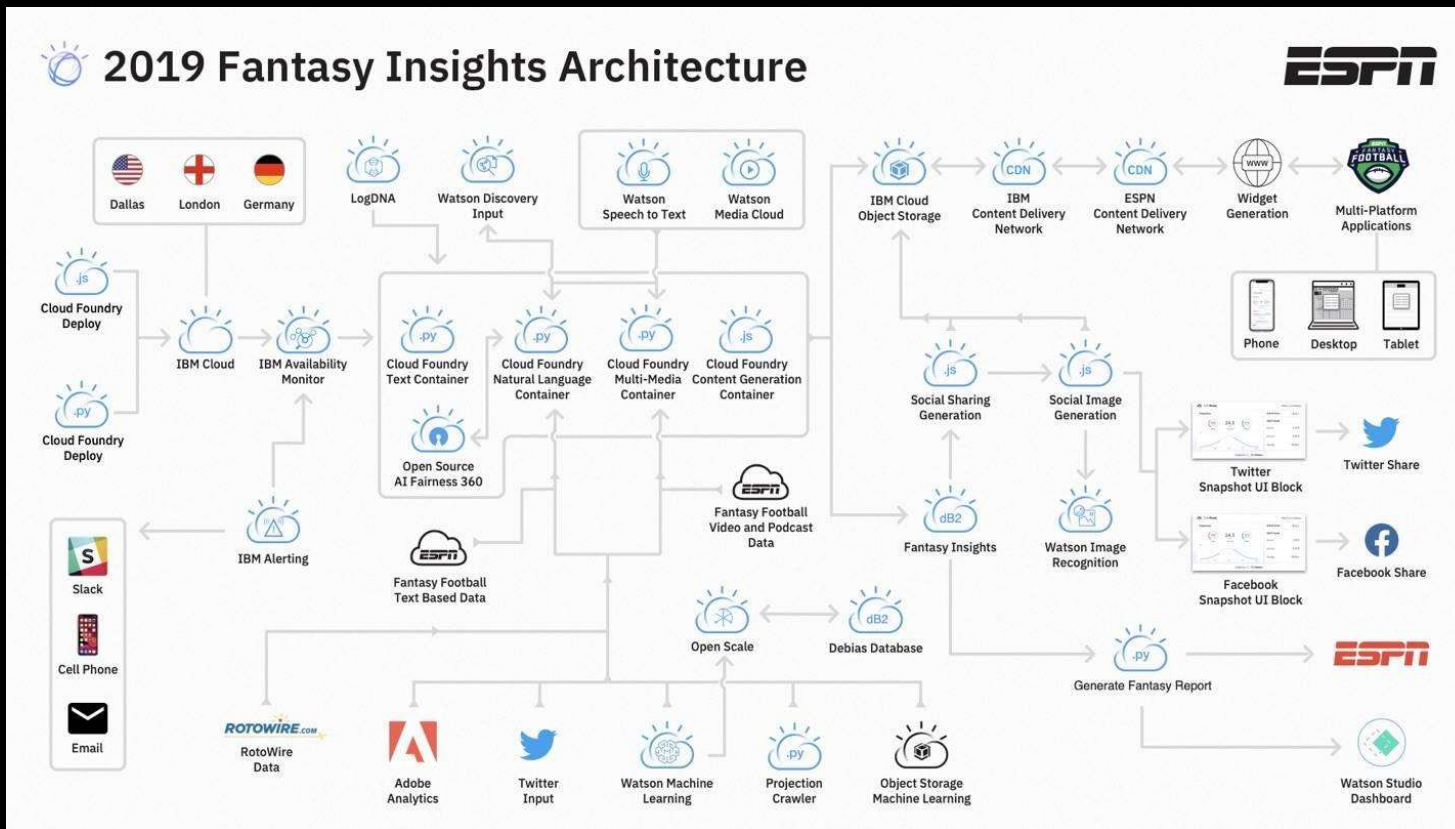
The Ad Council's "It's Up to You" campaign for covid-19 vaccine awareness

<https://www.ibm.com/downloads/cas/BVDAL4LK>

The screenshot shows a web browser window with the URL `developer.ibm.com/data/bias-in-advertising/`. The page is titled "Bias in Advertising Data" and is categorized as a "CSV" dataset. The main content area includes a description: "A synthetic dataset of user records that can be used to demonstrate discovery, measurement, and mitigation of bias in advertising." Below the description are "Save" and "Like" icons. To the right, there are two buttons: "Get this dataset" and "View dataset notebooks". The authors listed are Ketan Barve, Karthikeyan Natesan Ramamurthy, Josh Price, Vishnupriya Pradeep, and Skyler Speakman, with update and publish dates of June 16, 2022. The page also features social media icons for Facebook, Twitter, and LinkedIn. A "Categories" section lists "Artificial Intelligence". The left sidebar contains navigation links for "Artificial Intelligence", "Getting started with Artificial Intelligence", "APIs", "Articles", "Blogs", "Courses", "Dataset", "Learning Paths", "Models", "Open Projects", "Code Patterns", "Podcasts", "Series", "Tutorials", "Videos", and "Community".

Fantasy football “bust or boom”

<https://developer.ibm.com/articles/espn-fantasy-football-watson-insights/>



Health cost as a proxy for care management

Moninder Singh and Karthikeyan Natesan Ramamurthy.
 "Understanding Racial Bias in Health Using the Medical
 Expenditure Panel Survey Data." In: *Proceedings of the
 NeurIPS Workshop on Fair Machine Learning for Health*, 2019.

Table 1: Second year healthcare expenditure and utilization (outcome) metrics for panel 20, 2015-2016 MEPS cohort.

Metric	Entire Population		Top Decile (second year expenditure)	
	Race		Race	
	White	Black	White	Black
Average expense (both races)	\$5.6K		\$34.9K	
Top decile expense (both races)	\$13.6K			
Average expense	\$5.9K	\$4K	\$34.7K	\$36.2K
% of race in top decile			10.7%	7.1%
Average number of ER visits	0.18	0.21	0.62	0.83
Average number of IP nights	0.33	0.45	2.61	4.91
% with ER visits	12.9%	15.5%	40.4%	48%
% with IP nights	6.8%	6.8%	44.7%	54.3%

Child mortality prediction in sub-Saharan Africa

Ifrah Idrees, Skyler Speakman, William Ogallo, and Victor Akinwande. "Successes and Misses of Global Health Development: Detecting Temporal Concept Drift of Under-5 Mortality Prediction Models with Bias Scan." In: *Proceedings of the AMIA Joint Summits on Translational Science*, 2021, p. 286-295.

Table 5: Anomalous sub-populations with higher-than-expected under-5 mortality rates for each country.

Country Years, Mortality	Sub-Population	Size		Mortality %		Mean Model Predictions %	
		T0	T1	T0	T1	T1 Raw	T1 Shifted
Tanzania 2004, 2015 11.6%, 6.8%	Respondent's Age = Below 20	395	550	7.6	7.6	7.3	4.6
Burkina Faso 2003, 2010 15.9%, 11.7%	Visited Health Facility in past year = No Relationship structure = 3 or more adults	2957	1476	16.0	16.0	14.1	11.4
Kenya 2003, 2014 11.0%, 5.5%	Number of Births = 2, Gender of Head of Household = Male, Household size = 1-3	60	126	56.7	57.1	29.9	18.3
Ethiopia 2000, 2016 16.1%, 8.1%	Null	7245	7193	16.1	8.1	15.2	8.6
Nigeria 2003, 2018 19.6%, 13.1%	Relationship structure = Two adults, Number of Births = 2, Household size = 1-3	68	328	73.5	76.8	56.5	46.1

Not enough caring

Scott Barry Kaufman. "Why Don't People Care That More Men Don't Choose Caregiving Professions?." In: *Scientific American*, Feb. 2020.



**We need both,
but we often
devalue the
second**

Reason 1: Moral imperative

Manish Bhardwaj. "Accompaniment as a Path to Justice."
<https://thecenter.mit.edu/events/iap-seminar-series-innovation-and-social-justice/>.

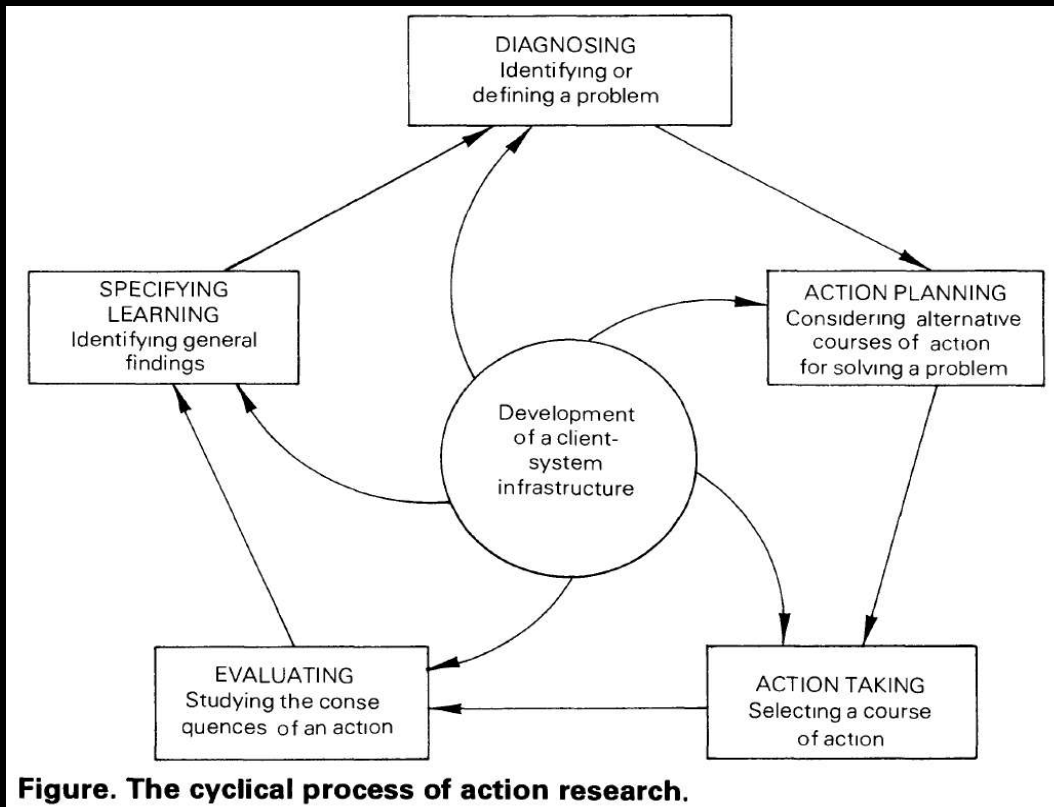
Jean Watson. *Nursing: The Philosophy and Science of Care*.
Boulder, Colorado: University of Colorado Press, 2008.

Accompaniment: journeying with the marginalized, literally and figuratively, until justice is achieved

Nursing theory: kindness and equanimity by treating all patients as they are and respecting their values, even if they are different from your own

Reason 2: Context and action research

Gerald I. Susman and Roger D. Evered. "An Assessment of the Scientific Merits of Action Research." In: *Administrative Science Quarterly* 23.4 (Dec. 1978), pp. 582–603.



Learn best practices from working on real problems

Uncountably infinite variety of AI contexts, use cases, and problems

Pick a few prototypical examples across industries and sectors, and generalize from them

Carative approach

Start with the real-world problem as experienced by the most vulnerable

Listen to them and understand their values

Meet them where they are and work toward a solution to their problem all the way to the end

Conduct a qualitative assessment of the entire solution by interviewing the affected communities

Simpa Networks



Basic idea of pay-as-you-go solar power

Hugo Gerard, Kamalesh Rao, Mark Simithraaratchy, Kush R. Varshney, Kunal Kabra, and G. Paul Needham, "Predictive Modeling of Customer Repayment for Sustainable Pay-As-You-Go Solar Power in Rural India," In: *Data for Good Exchange Conference*. Sep. 2015.

- Allows poor people who live in areas with unreliable grid power to obtain solar panels for their homes
- Progressive purchase financing scheme
 - Small down payment, incremental payments over 2-3 years
- Physical device can cut off transformer for lack of payment
- “Unlocked” after system is fully paid for
 - System is repossessed for too much non-payment
- **Machine learning task:** predict repayment
 - **Fairness issue:** caste and religion highly correlated with surname

Logistic regression model and results

Hugo Gerard, Kamallesh Rao, Mark Simithraaratchy, Kush R. Varshney, Kunal Kabra, and G. Paul Needham, "Predictive Modeling of Customer Repayment for Sustainable Pay-As-You-Go Solar Power in Rural India," In: *Data for Good Exchange Conference*. Sep. 2015.

Rank	Feature	Cat.	Imp.
1	*day rate	exog.	-
2	*gender = male	demo.	-
3	*branch = Mathura	exog.	-
4	*arable land	agri.	-
4	*branch = Bareilly-1	exog.	+
4	*down payment = 2500	exog.	-
4	*nature of business = skilled labour/driver	bus.	-
4	*spoken languages	demo.	+
4	*(value of inventory)/(family size)	bus.	-
10	*understand languages	demo.	+
10	*written languages	demo.	+
12	*nature of business = grains/fruits	bus.	-
13	agriculture = true	agri.	-
13	commodity 2 = potato	agri.	-
13	(candle expense)/(family size)	expense	-
13	price per quintal	agri.	-
13	salaried company = milk	salaried	-
18	age	demo.	+
19	battery expense	expense	-
19	distance to first recharge agent	demo.	-
19	monthly labour income	labour	+
19	address type = owned	demo.	+
23	number of cows \geq 3	asset	+
23	family size	demo.	+
23	nature of business = beauty	bus.	-
23	(loan repayment)/(total expense)	expense	+

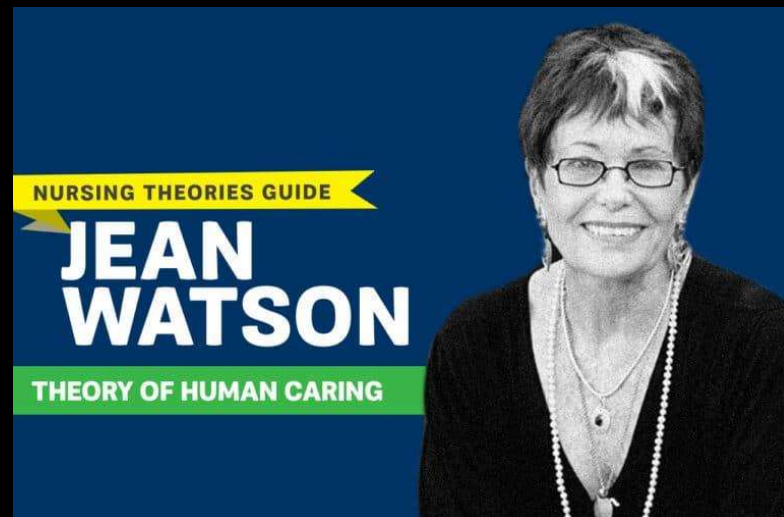
Model	Cross-Validation AUC	Future Test AUC
Simpa's prior model	0.546	0.477
Prior model + exogenous variables	0.642	0.537
Application form data	0.722	0.563
Application form data + exogenous variables	0.699	0.628



the result that matters

Watson and Watson

Kush R. Varshney. "The Watsons Meet Watson: A Call for Carative AI." In: *Montreal AI Ethics Blog*, Mar. 2022.
<https://montrealethics.ai/the-watsons-meet-watson-a-call-for-carative-ai/>



Thank you

Kush R. Varshney
Distinguished Research Scientist

—
krvarshn@us.ibm.com

<http://www.trustworthymachinelearning.com>

© Copyright IBM Corporation 2022. All rights reserved. The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. Any statement of direction represents IBM's current intent, is subject to change or withdrawal, and represent only goals and objectives. IBM, the IBM logo, and ibm.com are trademarks of IBM Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available at [Copyright and trademark information](#).

