Exploring Bias in AI & Machine Learning

Jim Box, SAS Institute

Framing the Problem

Let’s get calibrated

AI or ML?

What's the Difference

AI systems perform tasks that typically require human-level intelligence

- Understanding Language
- Recognizing images and patterns
- Making Decisions
- Learning from the past

Machine Learning uses data & algorithms to learn and make decisions

- ML may be part of the brains in an AI system, or it may be used in a stand-alone usage
- Generally, we think of predictive modelling

sas

AI in Its Own Words:

Artificial intelligence
A field that grows with each new day
Creating machines that think and learn
In ways that were once thought to be impossible

From self-driving cars to language translation
AI is changing the way we live our lives
Helping us to process vast amounts of data
And make decisions with speed and accuracy

But with great power comes great responsibility
We must ensure that AI is used ethically
And that its benefits are shared by all
For the sake of a better future for us all

Common Biases

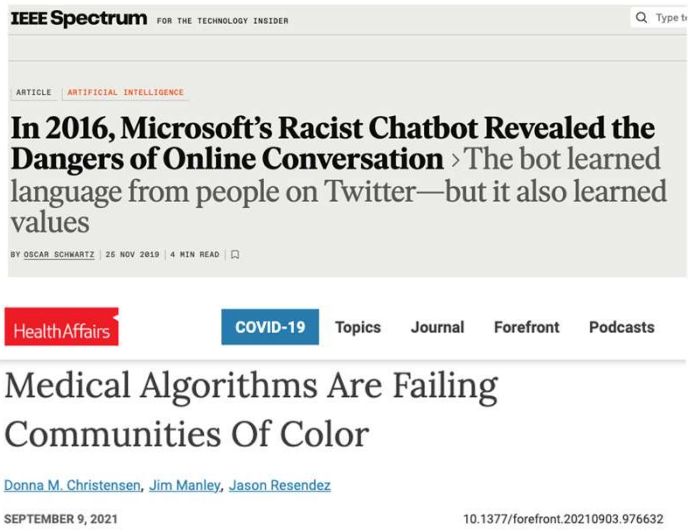
Too Many to Talk About

- Availability Bias
- Recall Bias
- Exclusion Bias
- Pre-processing Bias
- Measurement Bias
- Time-interval bias
- Historical Bias
- Selection Bias
- Confirmation Bias
- Cause/ Effect Bias
- Confounding Bias
- Collider Bias
- Prediction Bias
- Performance Bias
- Hindsight Bias
- Chronological Bias
- Funding Bias
- Automation Bias
- Deployment Bias
- Drift Bias
- Aggregation Bias
- Survivorship Bias
- Attrition Bias
- Reporting Bias
- Proxy Bias

Impact on Society

Financial Resources	Criminal Justice	Health Care	HR
			
Who gets approved for a loan?	Who gets a harsher sentence?	Who gets admitted to a care management program?	Who gets moved up to the next stage of the hiring process?

What Could Possibly go Wrong?



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

ARTICLE | ARTIFICIAL INTELLIGENCE

In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation › The bot learned language from people on Twitter—but it also learned values

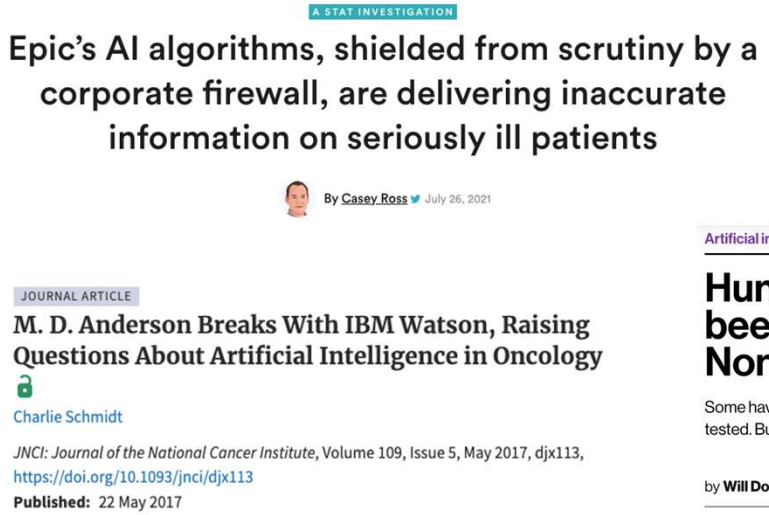
BY OSCAR SCHWARTZ | 25 NOV 2019 | 4 MIN READ

HealthAffairs COVID-19 Topics Journal Forefront Podcasts

Medical Algorithms Are Failing Communities Of Color

Donna M. Christensen, Jim Manley, Jason Resendez

SEPTEMBER 9, 2021 10.1377/forefront.20210903.976632



A STAT INVESTIGATION

Epic's AI algorithms, shielded from scrutiny by a corporate firewall, are delivering inaccurate information on seriously ill patients

By Casey Ross July 26, 2021

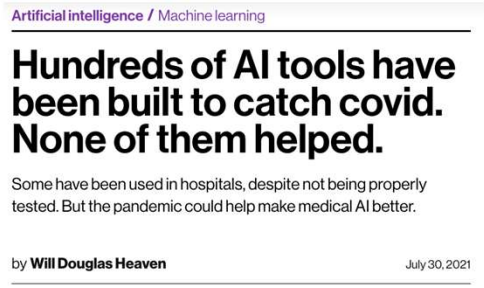
JOURNAL ARTICLE

M. D. Anderson Breaks With IBM Watson, Raising Questions About Artificial Intelligence in Oncology

Charlie Schmidt

JNCI: Journal of the National Cancer Institute, Volume 109, Issue 5, May 2017, djx113, https://doi.org/10.1093/jnci/djx113

Published: 22 May 2017



Artificial intelligence / Machine learning

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

by **Will Douglas Heaven** July 30, 2021

Musk's AI firm forced to delete posts praising Hitler from Grok chatbot

The popular bot on X began making antisemitic comments in response to user queries

Biases in ML Models

Just a few Examples

Machine Learning Process

Biased Training Data

Biased data gives biased results

Biased Training Data

Amazon Automates Resume Reviews

- Amazon receives hundreds of applications to open positions
- They are in the business of ranking things
- Created an AI system to comb through resumes and rank the applicants based on successful hires in the past
- Focused on interviewing the 5-star candidates

Biased Training Data

Amazon Automates Resume Reviews

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Biased Training Data

Example CV

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513

Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**
Smith College, August 2006-May2008

Biased Training Data

Keyword Exclusions

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513

Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**
Smith College, August 2006-May2008

Biased Training Data

Correlations

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513

Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**

Smith College, August 2006-May2008

Biased Training Data

Language Usage

ELENA SNAVELY

100 SAS Campus Drive
Cary, NC 27513

Elena.Snavely@sas.com

PROFESSIONAL OVERVIEW

Dynamic analyst with a proven record of achieving business objectives, committed to maximizing project potential. Experienced in critically evaluating information from disparate sources while proactively communicating results through presentations, reports, and discussions. Skilled at developing and optimizing models, conducting analysis utilizing statistical tools, and realizing business value from their output.

Key accomplishments and abilities:

- Capturing Customer Requirements
- Executing Project Plans
- Data Manipulation and Feature Engineering
- Executive Presentations
- Data Manipulation and Feature Engineering
- Managing and coaching groups
- Communicating effectively
- Design of User Interfaces

EXTRACURRICULAR

WOMEN'S CHESS CLUB CAPTAIN **Northampton, Massachusetts**
Smith College, August 2006-May2008

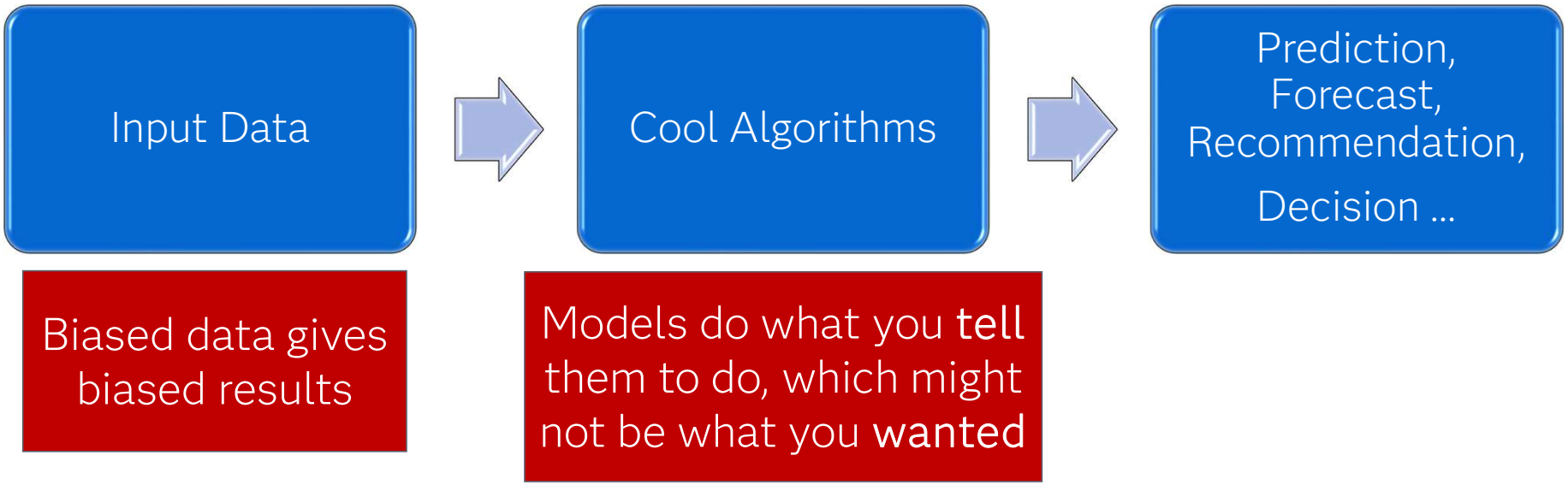
Biased Training Data

Takeaways

- Models trained on biased data will excel at applying that bias – even more efficiently than humans

- Your responsibility: Question the data
  - Where did it come from
  - How representative is it?
  - How did it get labeled?
  - Is there a feedback loop?

Models Do What you Tell Them



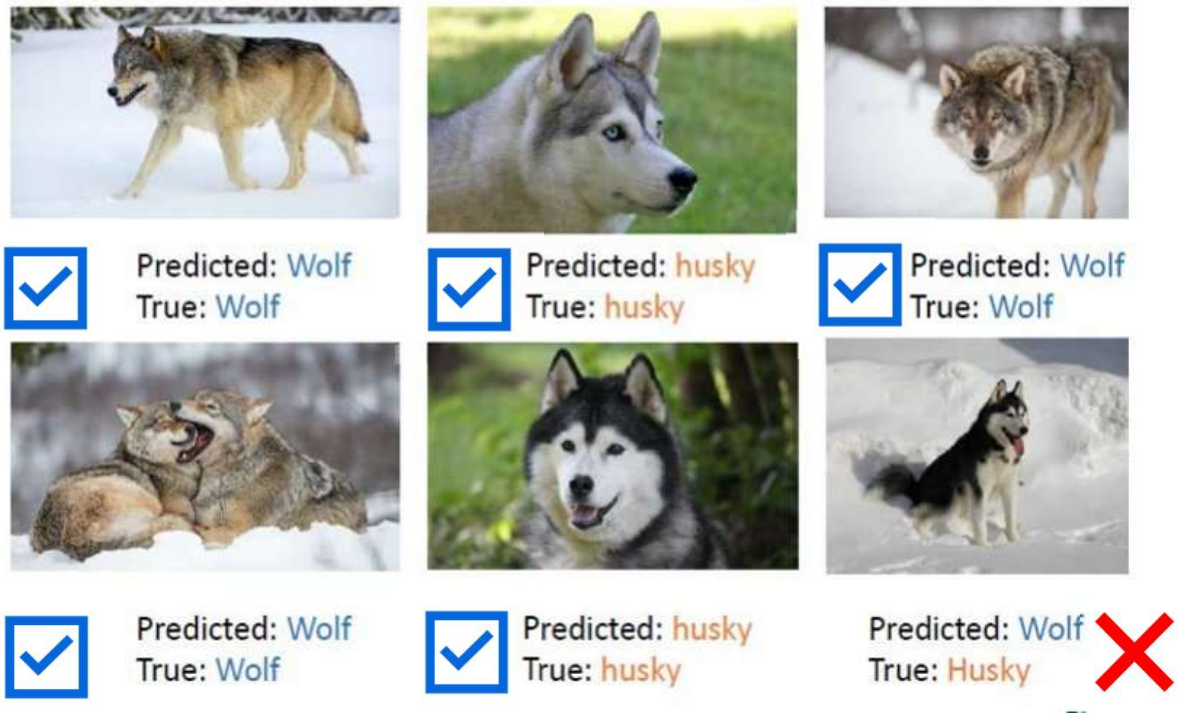
Models Do What you Tell Them

Husky or Wolf?

https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

Models Do What you Tell Them

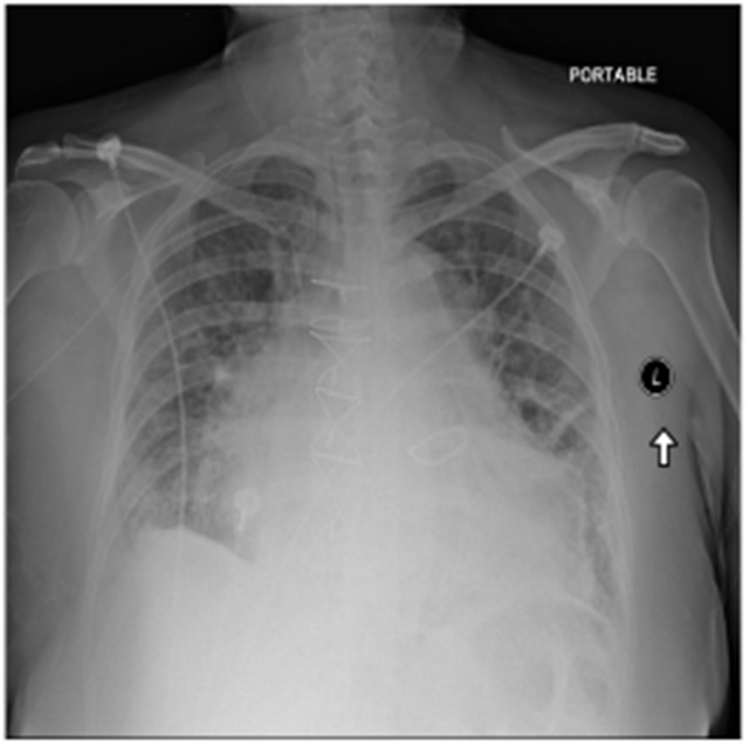
Husky or Wolf?



https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

Models Do What you Tell Them

Automating Radiology

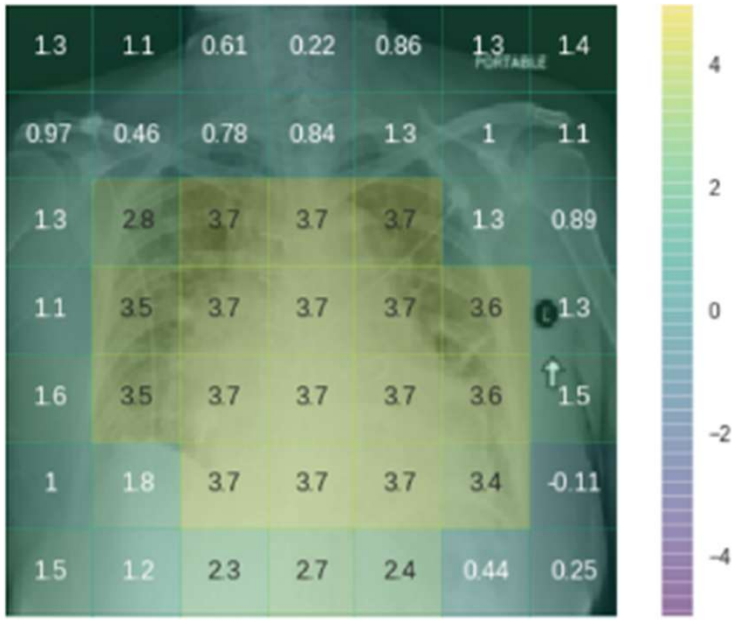


- Diagnosing Cardiomegaly (Enlarged Heart)
- Model trained on labeled images
- Apply the model to new images to test how it does
- This patient has the condition, and the model gave it a probability of 0.752, so it seems to have worked

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What you Tell Them

Automating Radiology

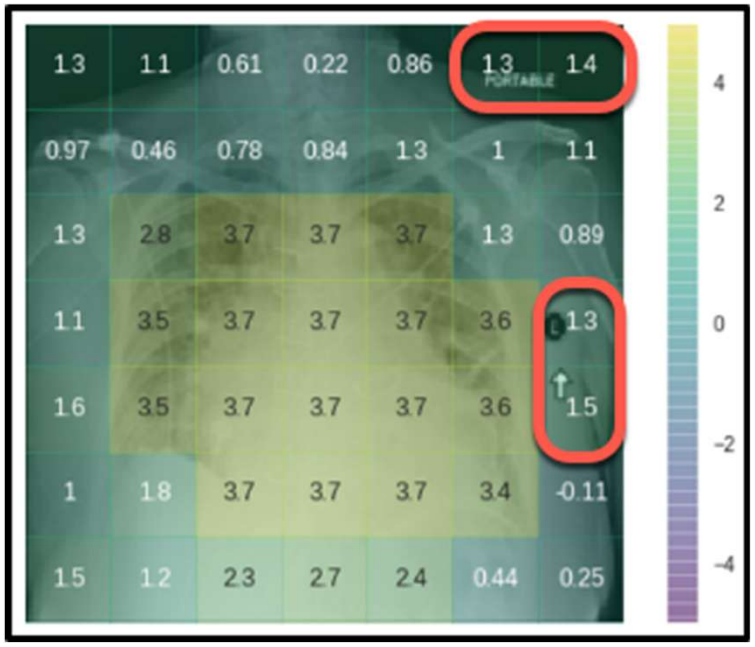


- Heatmap shows how different parts of the image contribute to the prediction
- Looks like the focus area is on the heart, which is good

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What you Tell Them

Automating Radiology

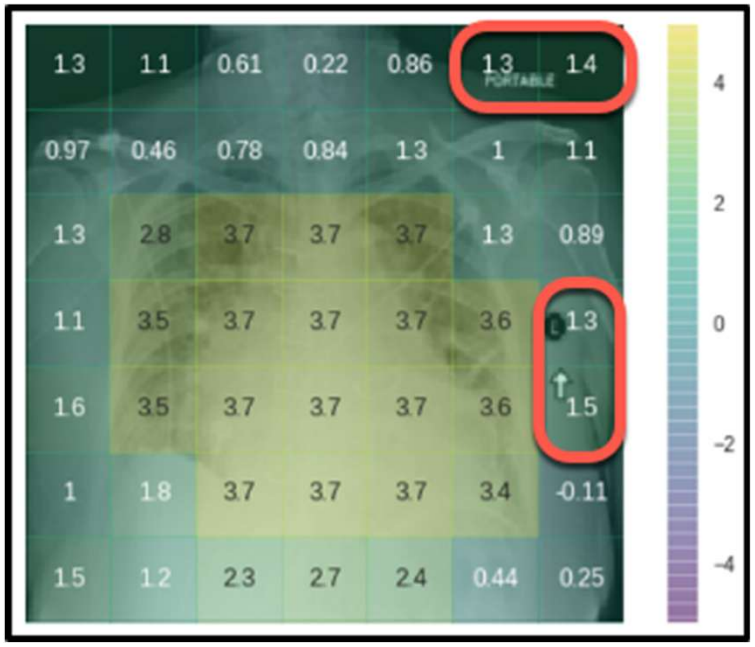


- Heatmap shows how different parts of the image contribute to the prediction
- Looks like the focus area is on the heart, which is good
- Some surprises, though

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What you Tell Them

Automating Radiology

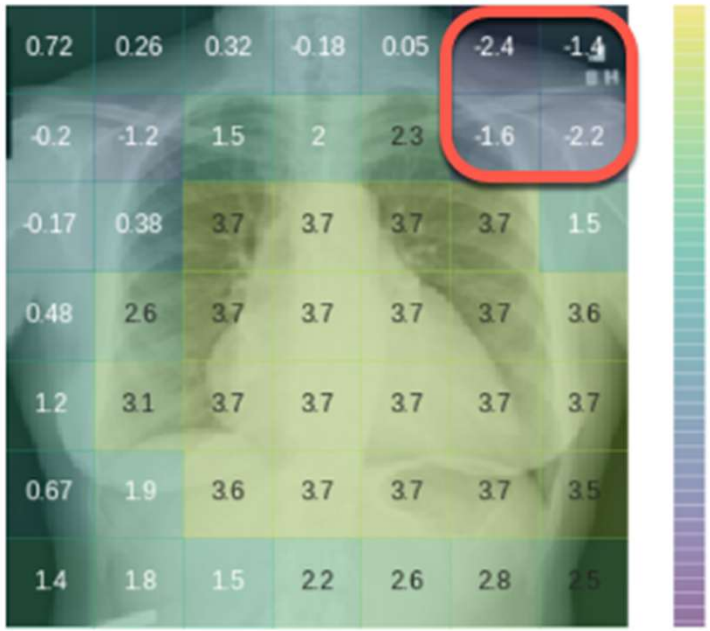


- Model is looking at markers of image metadata
- Model is using the fact that this image was taken with a portable x-ray machine, which is mainly used on sicker patients
- Model also considered the reviewing radiologist

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What you Tell Them

Automating Radiology



- Different image of a patient with the same condition
- Model downgraded the predication due to the lack of the portable stamp

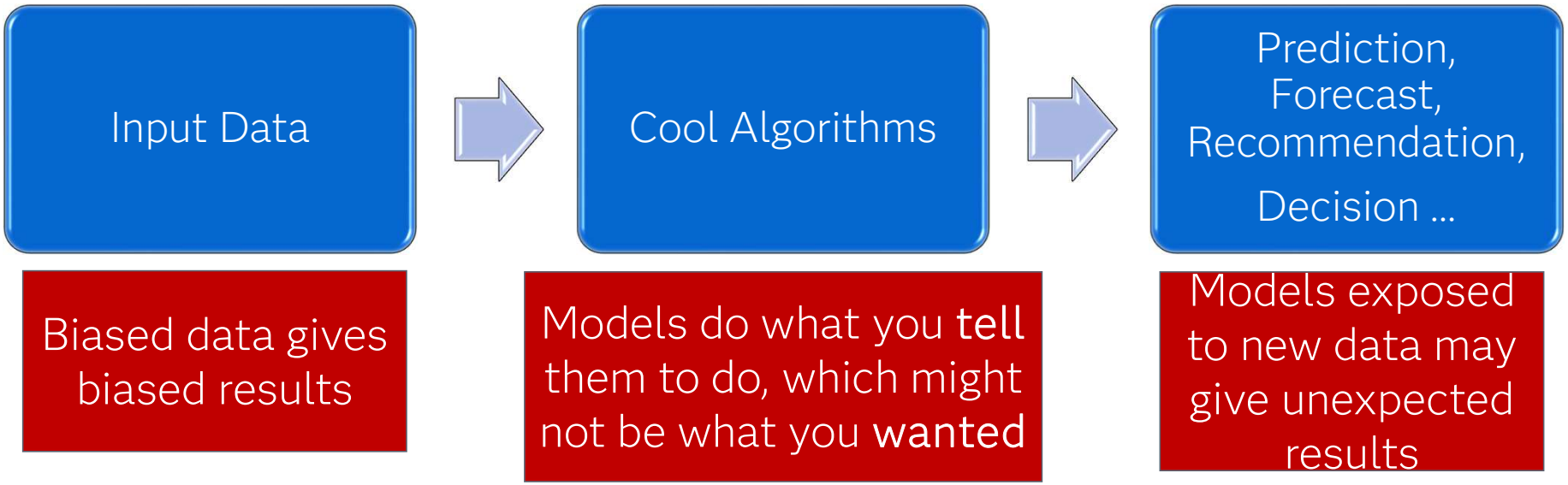
https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Models Do What you Tell Them

Takeaways

- Models are lazy, but effective
- Models do not care about context unless specifically instructed
- Your Responsibility: Question the Results
  - Why did the model make a specific prediction for this specific case?
  - What are the key inputs being used to make predictions?
  - What, exactly, was the model set up to do?

Models on New Data



Models on New Data

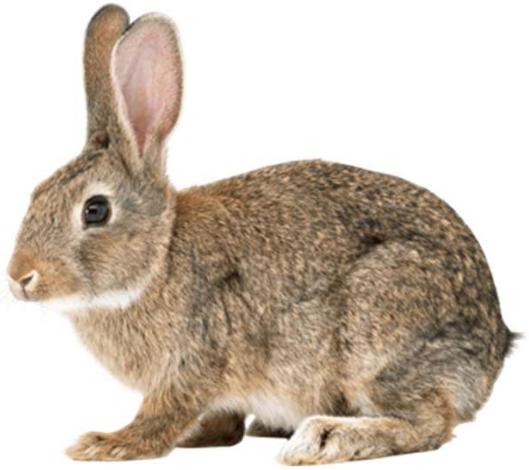
Surprises May be Very Harmful

- Models should only be applied to data that is similar to the data they were trained on

- Models will return a prediction on novel data, but it may not be trustworthy (although it will appear to be so)

Models on New Data

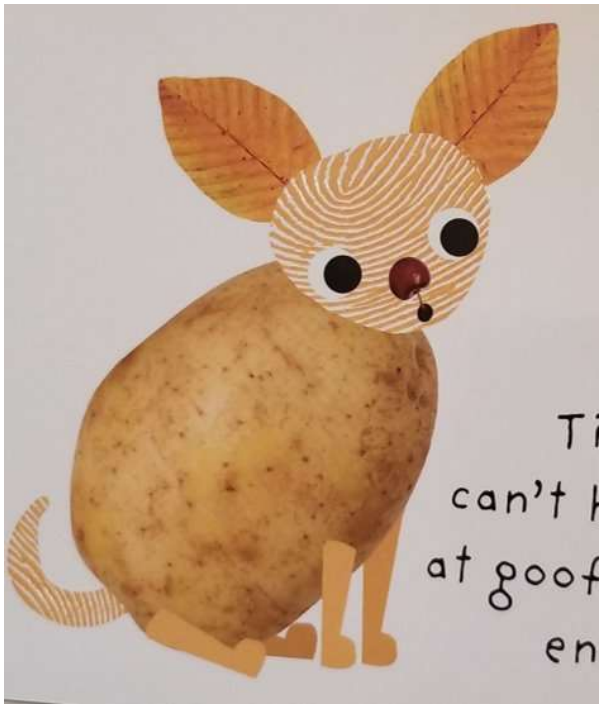
How Models Learn



https://www.pngarts.com

Models on New Data

Predictions



Happy Dog and Other Furry Friends. Written by Robert Newton. Illustrated by Ellie Boultwood

Models on New Data

Consequences can be Dire

Genetics research 'biased towards studying white Europeans'

Ethnic minorities set to miss out on medical benefits of research, scientist warns

People from minority ethnic backgrounds are set to lose out on medical benefits of genetics research due to an overwhelming bias towards studying white European populations, a leading scientist has warned.

> In a recent study, published in Psychiatric Genetics, Curtis found that a commonly used genetic test to predict schizophrenia risk gives scores that are 10 times higher in people with African ancestry than those with European ancestry. This is not because people with African ancestry actually have a higher risk of schizophrenia, but because the genetic markers used were derived almost entirely from studies of individuals of European ancestry.

https://www.theguardian.com/science/2018/oct/08/genetics-research-biased-towards-studying-white-

Models on New Data

Takeaways

- Your responsibility – Question the Application
  - Is this data similar to what the model was trained on
  - Do different groups (population subgroups) in my predictions get different results
  - Is there a human feedback loop that allows for retraining the model

Where do These Biases Come From?

Bias Comes from Us

- Like children, models can pick up patterns in the data that we are not explicitly trying to teach them

- There is a lack of awareness by Data Scientists/Statisticians about how historical/societal biases may be present in data modeling
  - How we collect data
  - The problems we decide to solve
  - The data we choose to train models on
  - How we assess accuracy
  - How we present the results

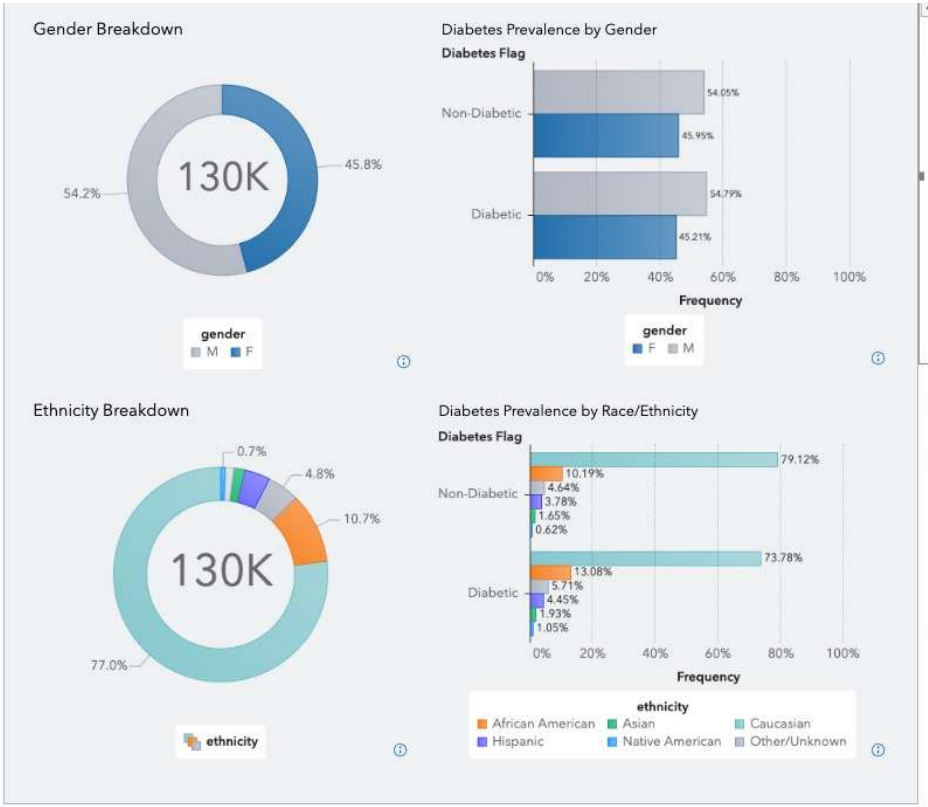
Biases in ML Models

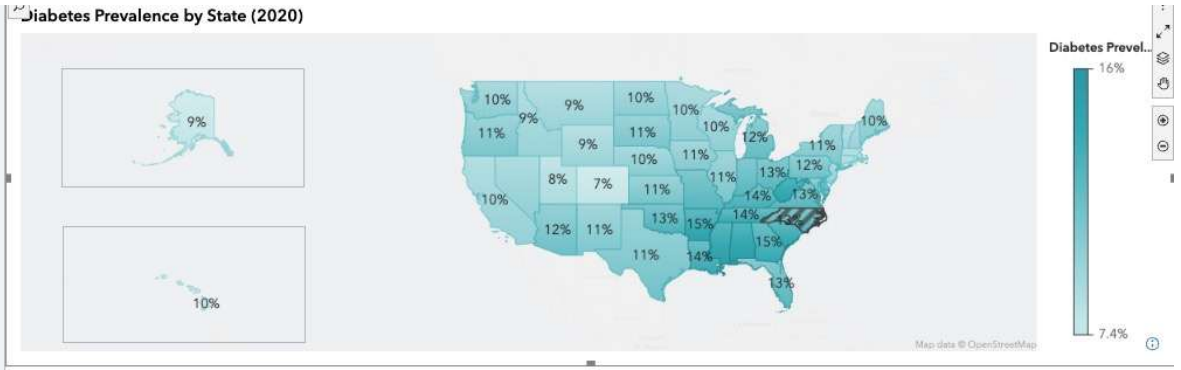
What do we do about it?

Know Your Variables

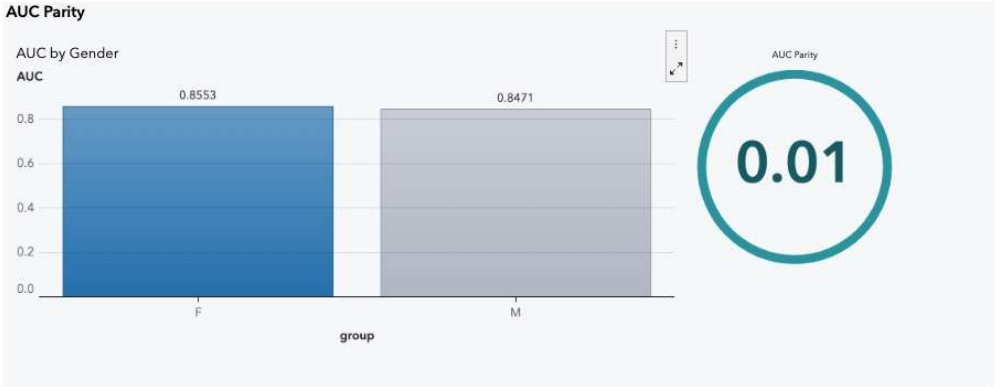
Name/Label	Length	Semantic Type	Information Privacy
Ethnic Group	43	ETHNICITY ▾	Sensitive
Gender	6	GENDER ▾	Private

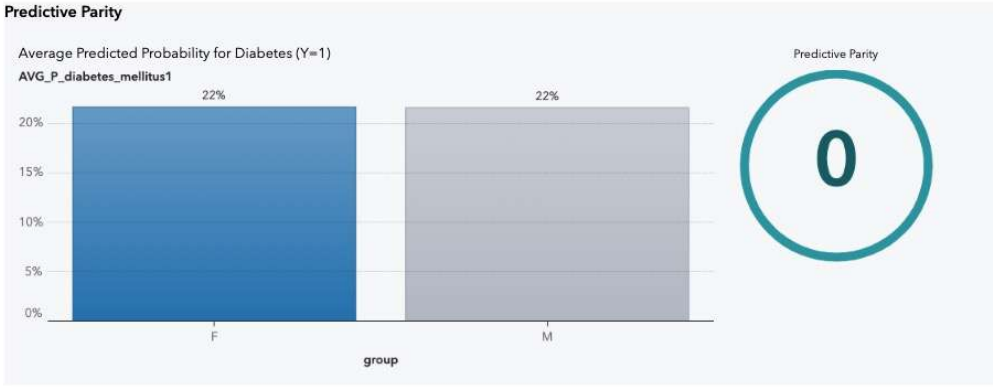
Review the Data

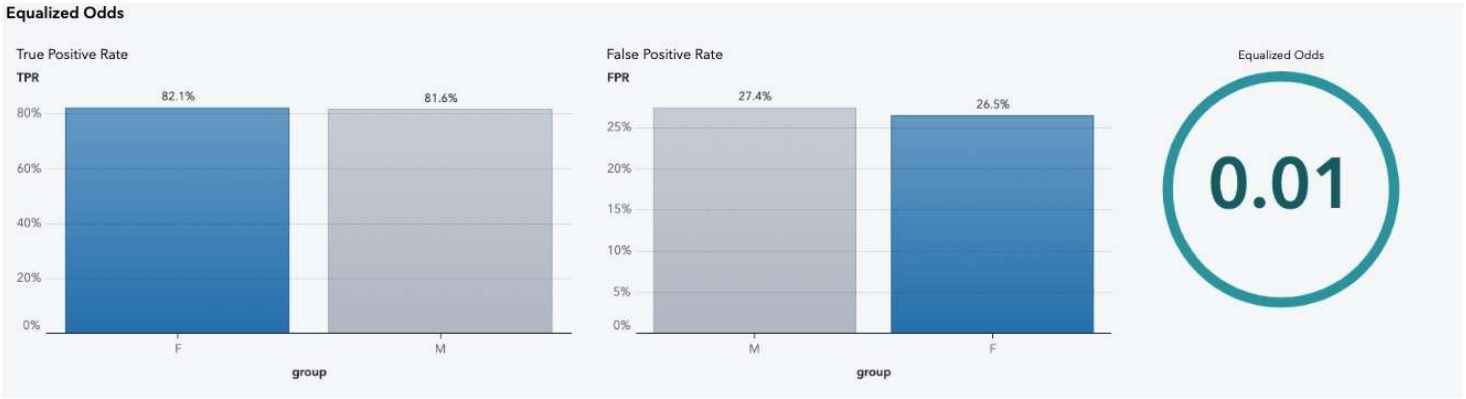




Assess Performance by Variable





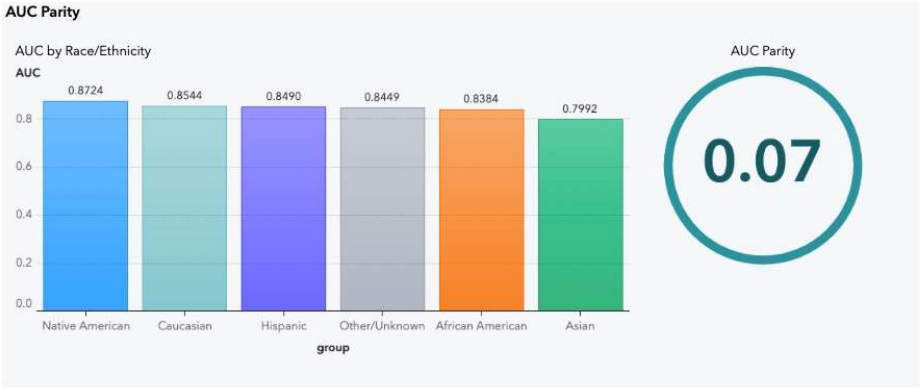


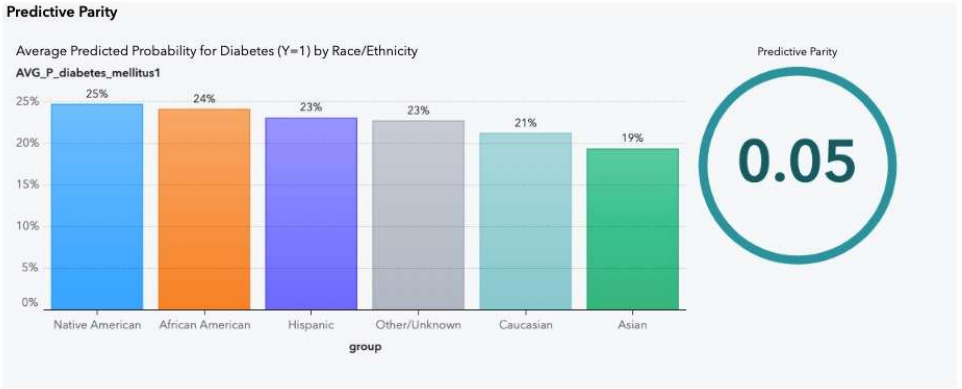
AUC Parity: is the requirement that the area under the receiver operating characteristic (ROC) curve is the same across groups. The value calculates the maximum difference in model accuracy between groups. The ideal model would be equaly accurate for all groups.

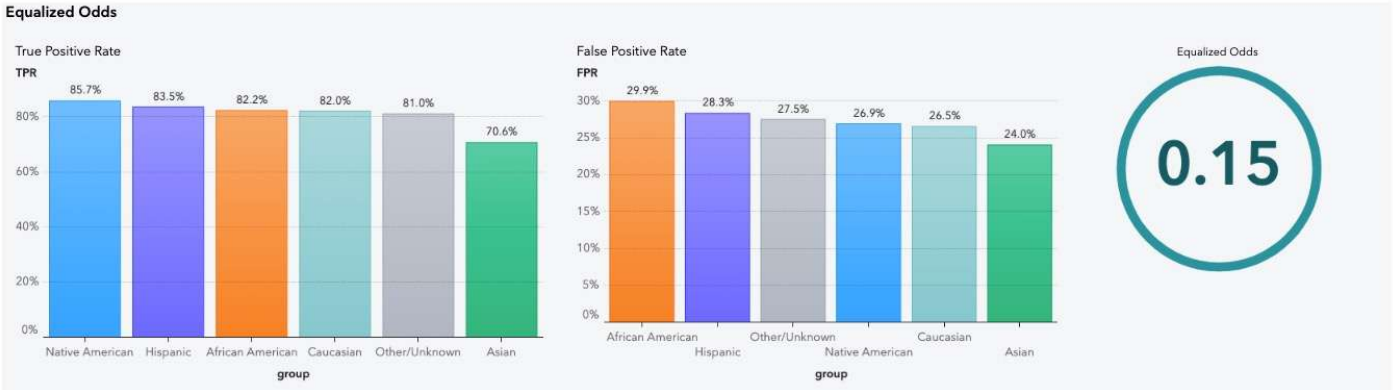
Predictive Parity: assesses the maximum difference in the average predicted probability produced by the model between groups. The ideal model would have no difference if the sensitive variable plays no role in the outcome of interest.

Equalized Odds: calculates the true positive (the rate at which the model identifies patients as diabetic when in fact they are) and false positive rates (the rate at which the model identifies patients as diabetic when in fact they are not) for each group (in this case race/ethnicity) and assesses the difference (max difference of the max difference for TPR and FPR). The ideal model will have the same false positive rate and true positive rates.

Assess Performance by Variable





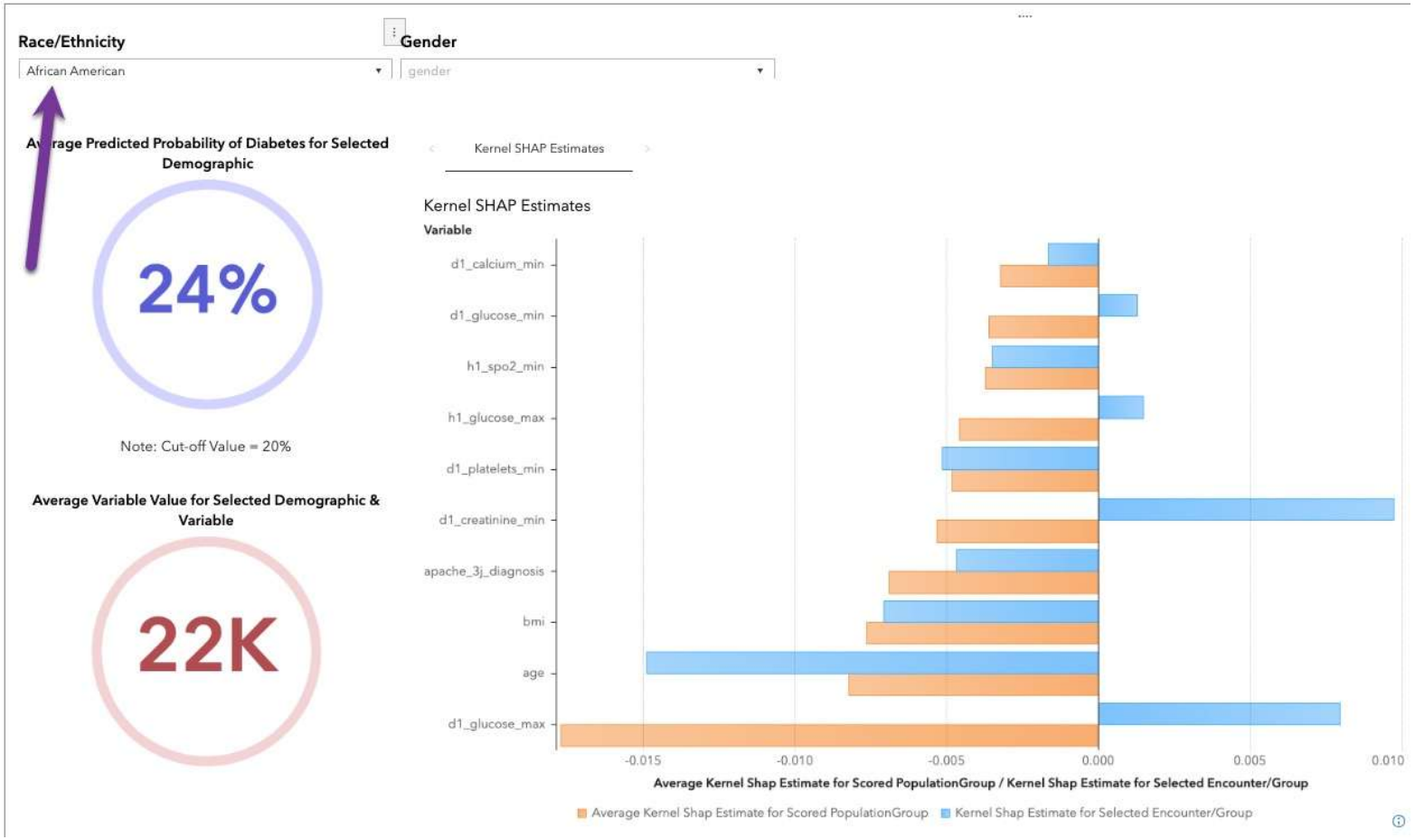


AUC Parity: is the requirement that the area under the receiver operating characteristic (ROC) curve is the same across groups. The value calculates the maximum difference in model accuracy between groups. The ideal model would be equaly accurate for all groups.

Predictive Parity: assesses the maximum difference in the average predicted probability produced by the model between groups. The ideal model would have no difference if the sensitive variable plays no role in the outcome of interest.

Equalized Odds: calculates the true positive (the rate at which the model identifies patients as diabetic when in fact they are) and false positive rates (the rate at which the model identifies patients as diabetic when in fact they are not) for each group (in this case race/ethnicity) and assesses the difference (max difference of the max difference for TPR and FPR). The ideal model will have the same false positive rate and true positive rates.

Evaluate Model Interpretability



Beware of Proxy Variables

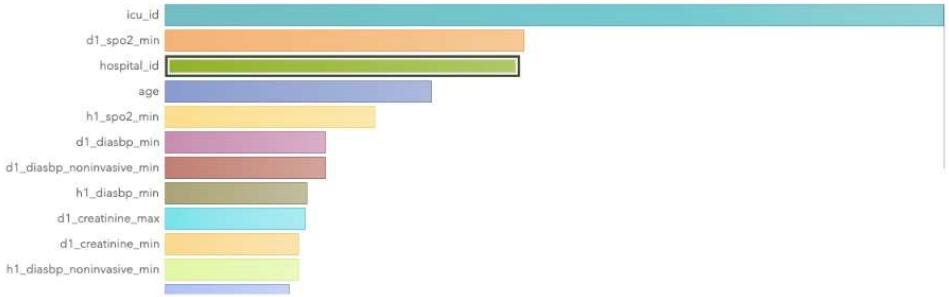
What are the characteristics of ethnicity?

African American

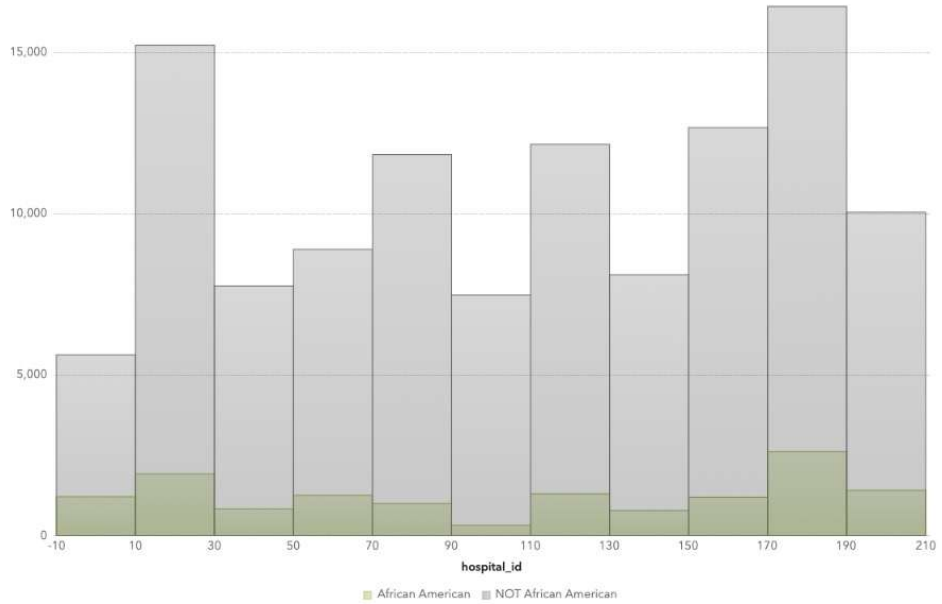
African American is the second most common value representing 10.69% (14K of 130K) of ethnicity. African American is less common than Caucasian (at 77.01%), but more common than Other/Unknown (at 4.81%). The three most related factors are icu_id, d1_spo2_min, and hospital_id.

Caucasian and African American are much more common than all other ethnicity values, together representing 87.70%.

What factors are most related to ethnicity?



What is the relationship between ethnicity and hospital_id?



What are the groups based on hospital_id by the chance of ethnicity being African American?

High Low

65.81% If hospital_id is between 175 and 176, h1_diasbp_min is greater than or equal to 83, then ethnicity has a 65.81% chance (231 of 351 cases) of being African American.

61.84% If d1_spo2_min is greater than or equal to 96, hospital_id is between 175 and 176, then ethnicity has a 61.84% chance (637 of 1K cases) of being African American.

60.79% If hospital_id is between 175 and 176, d1_creatinine_min is greater than or equal to 1.1, then ethnicity has a 60.79% chance (572 of 941 cases) of being African American.

Establish Guardrails

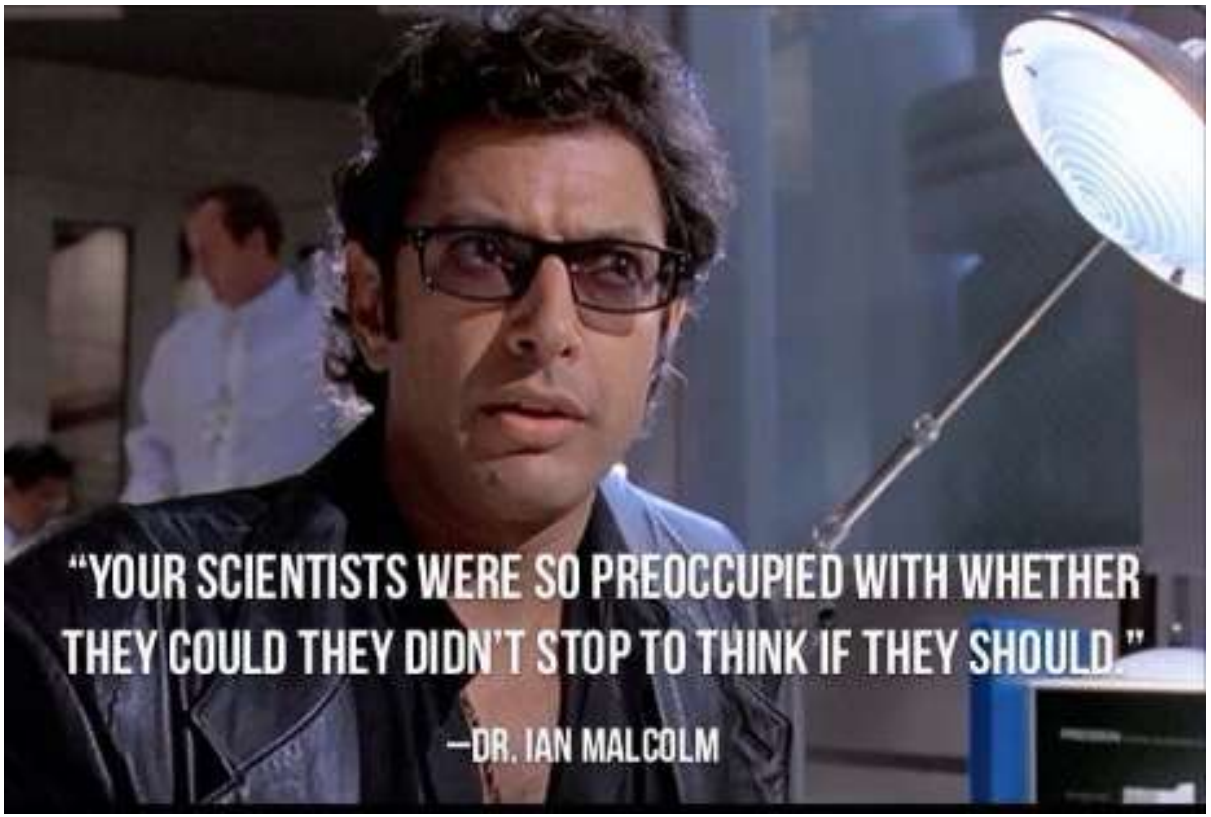
Enforce Responsible AI Best Practices: Trustworthy AI Life Cycle Workflow Available

Started: Tuesday | Modified: Tuesday | Views: 344

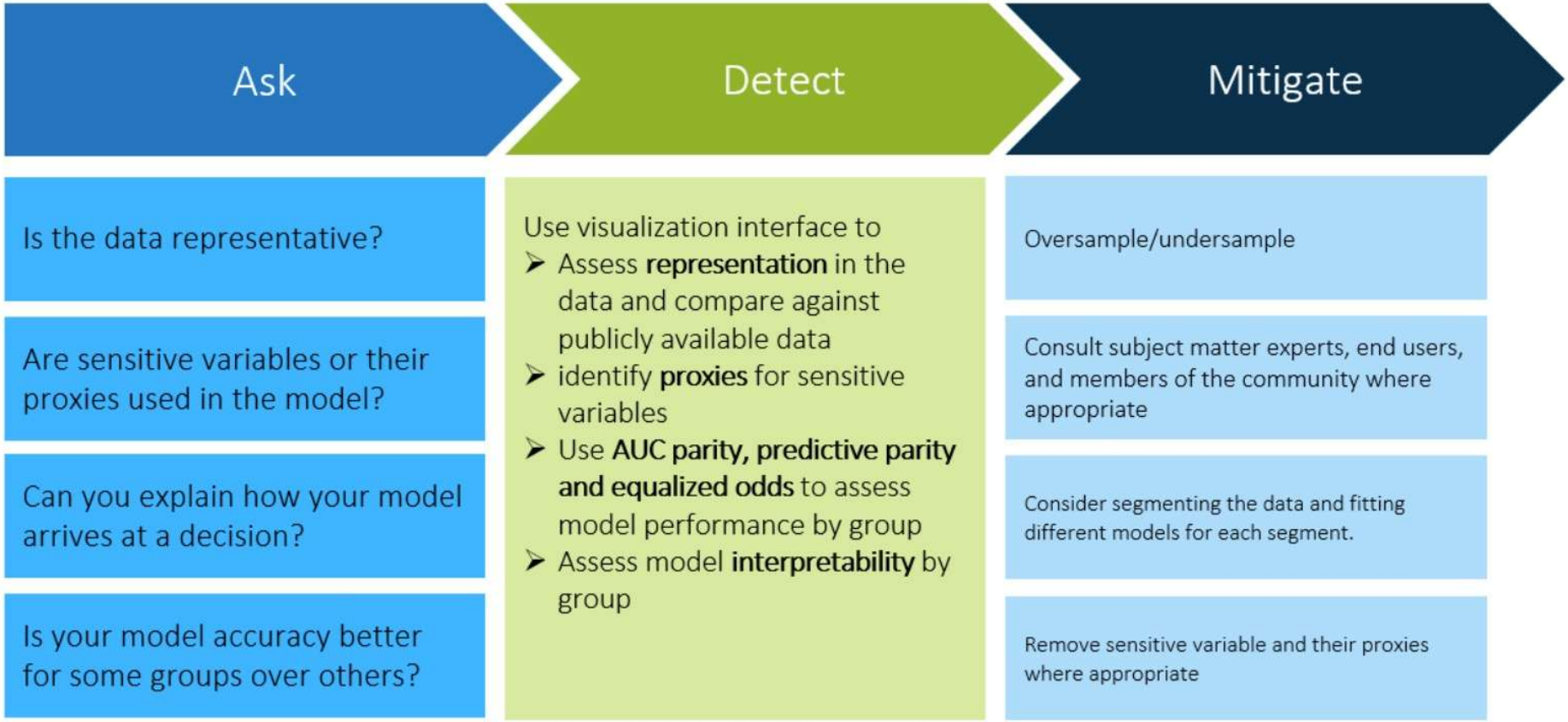
SAS has just released an experimental version of our Trustworthy AI Life Cycle Workflow for use with SAS® Model Manager and SAS® Workflow Manager on SAS Viya 2024.01 and later. Our Trustworthy AI Life Cycle workflow enforces standards and best practices set by the AI Risk Management Framework defined by the National Institute of Standards and Technology (NIST). The workflow allows organizations to document their considerations of AI systems' impact on human lives. Our workflow includes steps to ensure that the training data is representative of the population that is impacted, as well as that the model predictions and performance are similar across protected classes These steps help ensure that the model is not causing disparate impact or harm to a specific group. Furthermore, you can ensure that your model remains accurate over time by creating human-in-the-loop tasks to act when additional attention is needed.

https://communities.sas.com/t5/SAS-Communities-Library/Enforce-Responsible-AI-Best-Practices-Trustworthy-AI-Life-Cycle/ta-p/912717

Be Thoughtful



Summary

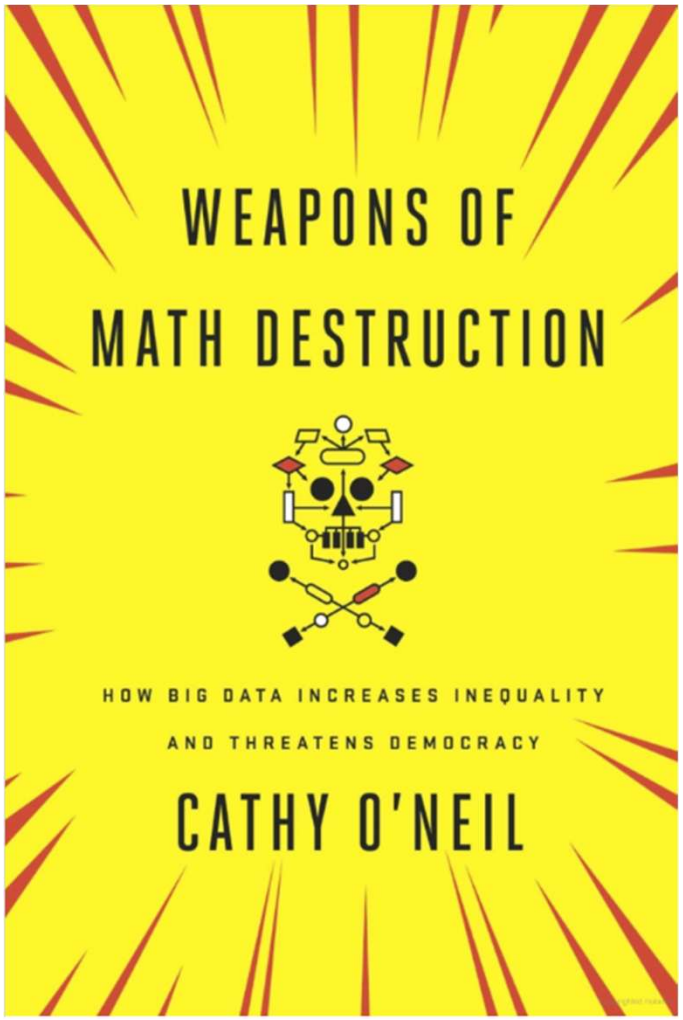


What's Next

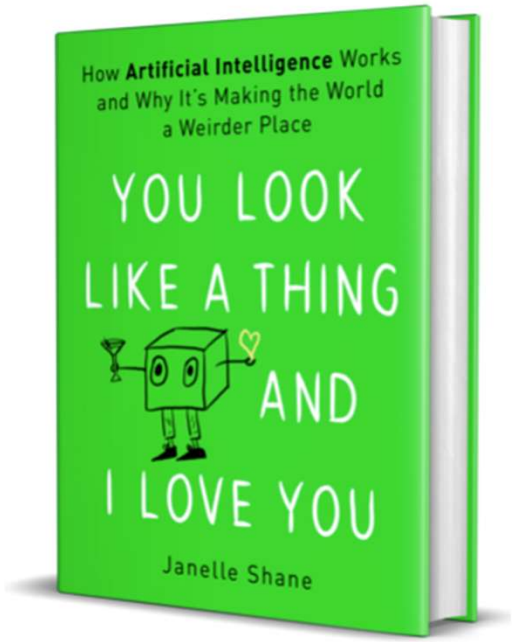
Where can you Learn More?

At the Movies

In a Book



On a Blog



AI Weirdness

AI WEIRDNESS: THE STRANGE SIDE OF MACHINE LEARNING

https://www.aiweirdness.com/

From Each Other

Elena Snavely (She/Her) · 1st
Problem Solver | Results Driven | Ethical AI Advocate

Hiwot Tesfaye (She/Her) · 1st
Technical Advisor | Office of Responsible AI

Allie DeLonay (She/Her) · 1st
Senior Data Scientist | Data Ethics Practice | Top 100 Brilliant Women in AI Ethics 2023

Thanks!

Jim Box
SAS Institute
Jim.Box@sas.com