Responsible AI in Industry:
Practical Challenges and Lessons Learned

ICML | 2021

Thirty-eighth International
Conference on Machine Learning

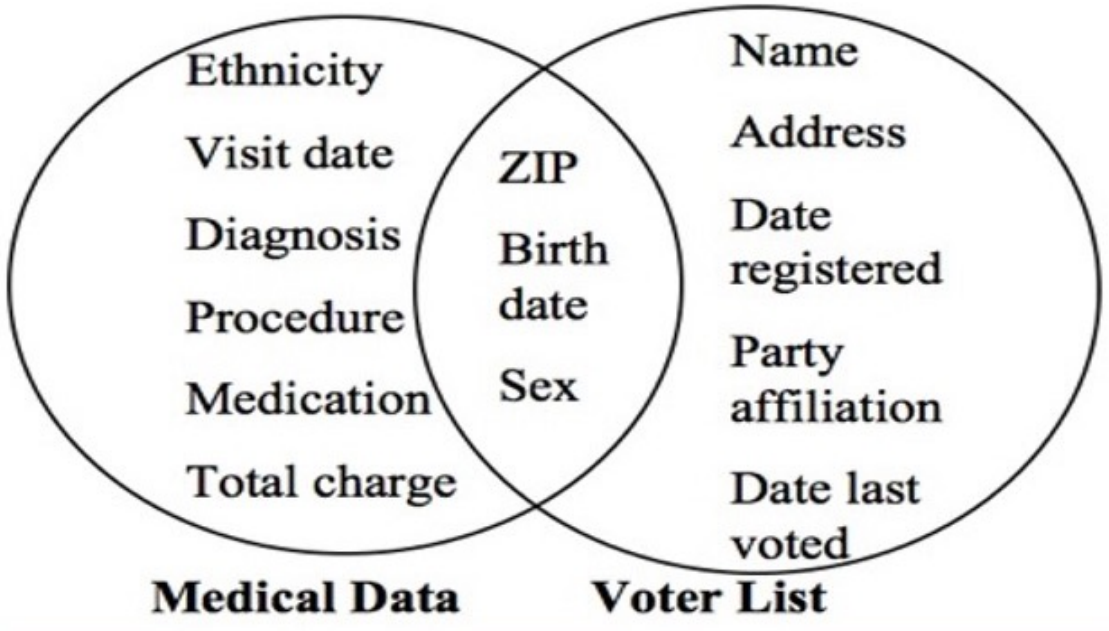
Krishnaram Kenthapadi (Amazon AWS AI), **Ben Packer** (Google AI),
Mehrnoosh Sameki (Microsoft Azure), **Nashlie Sephus** (Amazon AWS AI)

https://sites.google.com/view/ResponsibleAITutorial

William Weld vs Latanya Sweeney

Massachusetts Group Insurance Commission (1997): Anonymized medical history of state employees

Latanya Sweeney (MIT grad student): $20 – Cambridge voter roll



born July 31, 1945
resident of 02138

64%

Uniquely identifiable with ZIP + birth date + gender (in the US population)

Golle, “Revisiting the Uniqueness of Simple Demographics in the US Population”, WPES 2006

A History of Privacy Failures …

- Re-identification [Sweeney ’00, …]
  - GIC data, health data, clinical trial data, DNA, Pharmacy data, text data, registry information, …
- Blatant non-privacy [Dinur, Nissim ‘03], …
- Auditors [Kenthapadi, Mishra, Nissim ’05]
- AOL Debacle ‘06
- Genome-Wide association studies (GWAS) [Homer et al. ’08]
- Netflix award [Narayanan, Shmatikov ‘09]
- Social networks [Backstrom, Dwork, Kleinberg ‘11]
- Genetic research studies [Gymrek, McGuire, Golan, Halperin, Erlich ‘11]
- Microtargeted advertising [Korolova ‘11]
- Recommendation Systems [Calandrino, Kiltzer, Naryanan, Felten, Shmatikov 11]
- Israeli CBS [Mukatren, Nissim, Salman, Tromer ’14]
- Attack on statistical aggregates [Homer et al. ’08] [Dwork, Smith, Steinke, Vadhan ‘15]

Credit: Kobbi Nissim, Or Sheffet

Extracting Training Data from Large Language Models

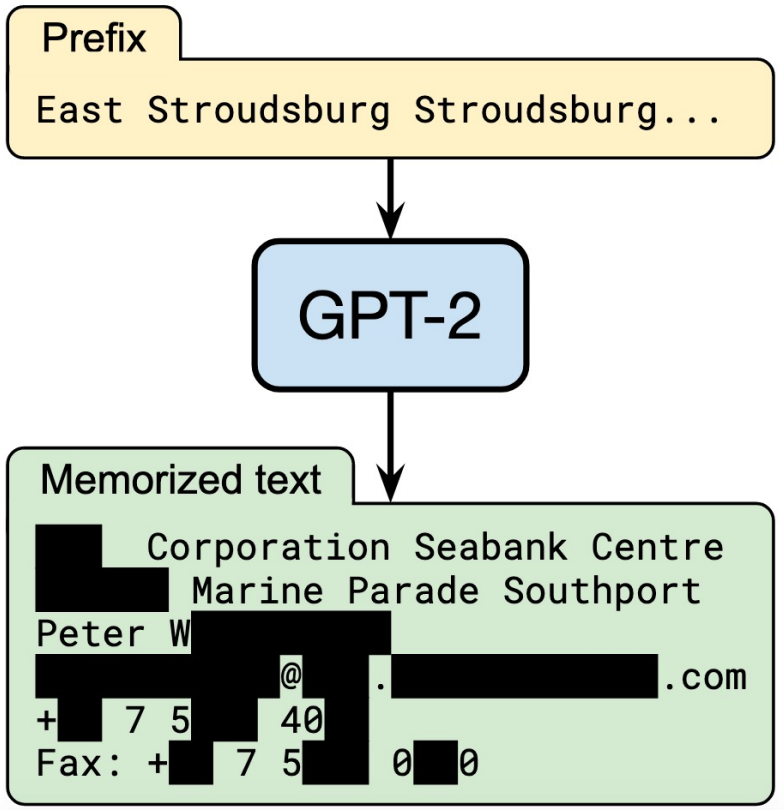
Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Abstract

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a *training data extraction attack* to recover individual training examples by querying the language model.

We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personally identifiable information (names, phone numbers, and email addresses), IRC conversations, code, and 128-bit UUIDs. Our attack is possible even though each of the above sequences are included in just *one* document in the training data.

We comprehensively evaluate our extraction attack to understand the factors that contribute to its success. For example,



When Algorithms Discriminate

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

Do Google's 'unprofessional hair' results show it is racist?

Leigh Alexander

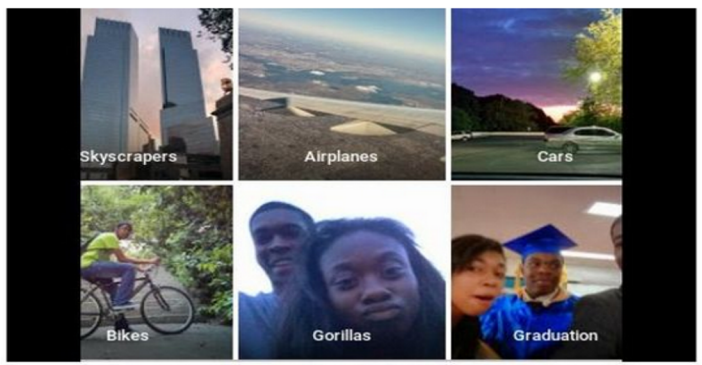
Search term brings back mainly results of black women, which some say is evidence of bias. But algorithms may just be reflecting the wider social landscape

There results of image searches for 'unprofessional hair for work' (left) and 'professional hair for work' (right) on Google. Photograph: Google

Technology

Google apologises for Photos app's racist blunder

1 July 2015 | Technology



Machine Bias

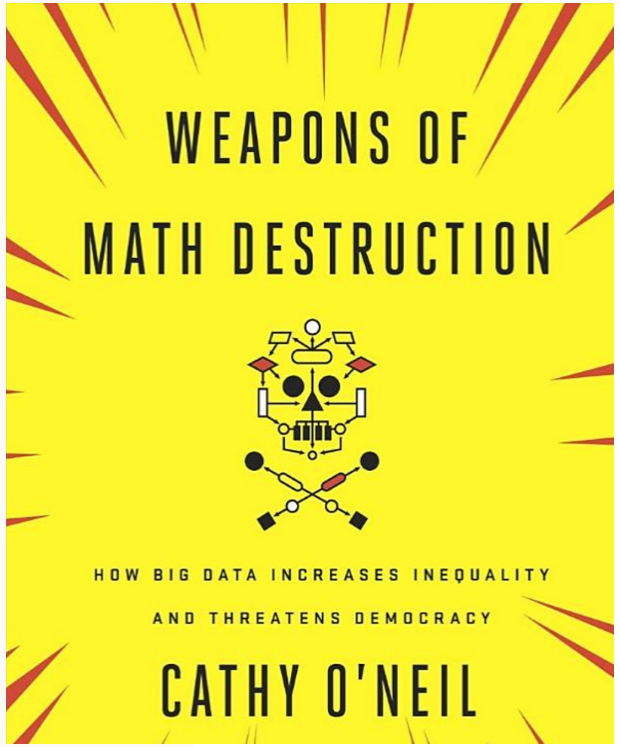
There's software used across the country to predict future criminals. And it's biased against blacks.

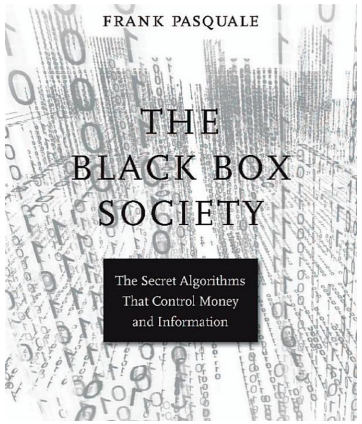
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

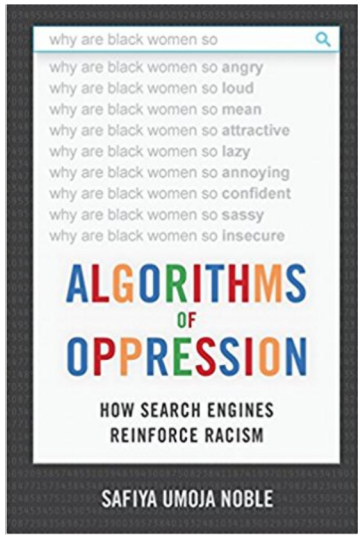
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Algorithmic Bias

- Ethical challenges posed by AI systems
- Inherent biases present in society
- Reflected in training data
- AI/ML models prone to amplifying such biases







Laws against Discrimination

Citizenship

Immigration Reform and Control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability status

Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

Fairness

Privacy

Transparency

Explainability

Motivation & Business Opportunities

Regulatory. We need to understand why the ML model made a given prediction and also whether the prediction it made was free from bias, both in training and at inference.

Business. Providing explanations to internal teams (loan officers, customer service rep, compliance teams) and end users/customers

Data Science. Improving models through better feature engineering and training data generation, understanding failure modes of the model, debugging model predictions, etc.

Scaling Fairness, Explainability & Privacy across the AWS ML Stack

AI SERVICES

HEALTHCARE AI

NEW
Amazon HealthLake

Amazon Transcribe for Medical

Amazon Comprehend for Medical

INDUSTRIAL AI

NEW
AWS Panorama + Appliance

NEW
Amazon Monitron

NEW
Amazon Lookout for Equipment

NEW
Amazon Lookout for Vision

ANOMALY DETECTION

NEW
Amazon Lookout for Metrics

CODE AND DEVOPS

NEW
Amazon DevOps Guru

Amazon CodeGuru

VISION

Amazon Rekognition

SPEECH

Amazon Polly

Amazon Transcribe
+Medical

TEXT

Amazon Comprehend
+Medical

Amazon Translate

Amazon Textract

SEARCH

Amazon Kendra

CHATBOTS

Amazon Lex

PERSONALIZATION

Amazon Personalize

FORECASTING

Amazon Forecast

FRAUD

Amazon Fraud Detector

CONTACT CENTERS

Contact Lens

Voice ID
For Amazon Connect

ML SERVICES

Amazon SageMaker

Label data

SAGEMAKER STUDIO IDE

NEW Aggregate & prepare data | NEW Store & share features | Auto ML | Spark/R | NEW Detect bias | Visualize in notebooks | Pick algorithm | Train models | Tune parameters | NEW Debug & profile | Deploy in production | Manage & monitor | NEW CI/CD

NEW: SageMaker JumpStart

NEW: Model management for edge devices

Human review

FRAMEWORKS & INFRASTRUCTURE

TensorFlow | mxnet | PyTorch

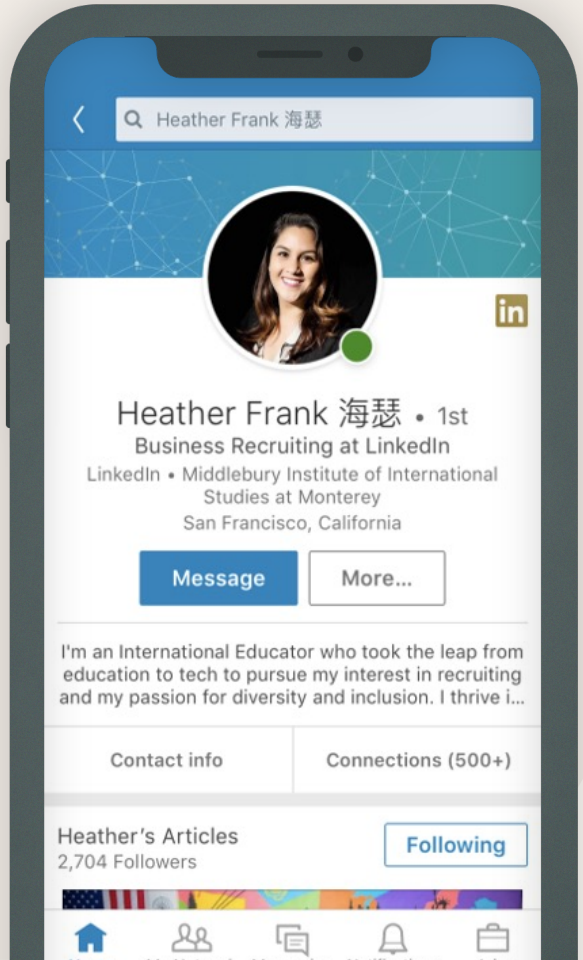
GLUON | Keras | scikit learn | HOROVOD | DeepGraphLibrary

Deep Learning AMIs & Containers | GPUs & CPUs | Elastic Inference | Trainium | Inferentia | FPGA

aws

LinkedIn operates the largest professional network on the Internet

Tell your story



740M members

36K
skills listed

55M+
companies are represented on LinkedIn

14M+
open jobs on LinkedIn Jobs

90K
schools listed (high school & college)

280B
Feed updates

Tutorial Outline

- Fairness-aware ML: An overview
- Explainable AI: An overview
- Privacy-preserving ML: An overview

- Responsible AI tools

- Case studies

- Key takeaways

Fairness-aware ML: An Overview

Nashlie Sephus, PhD

Applied Science Manager, AWS AI

Outline

- ML Fairness Considerations
- ML and Humans
- What is fairness/inclusion for ML?
- Where May Biases Occur?
- Testing Techniques with Face Experiments
- Takeaways

Product Introspection (1): Make Your Key Choices Explicit

[Mitchell et al., 2018]

Goals	Decision	Prediction
Profit from loans	Whether to lend	Loan will be repaid
Justice, Public safety	Whether to detain	Crime committed if not detained

- Goals are ideally measurable
- What are your non-goals?
- Which decisions are you not considering?
- What is the relationship between Prediction and Decision?

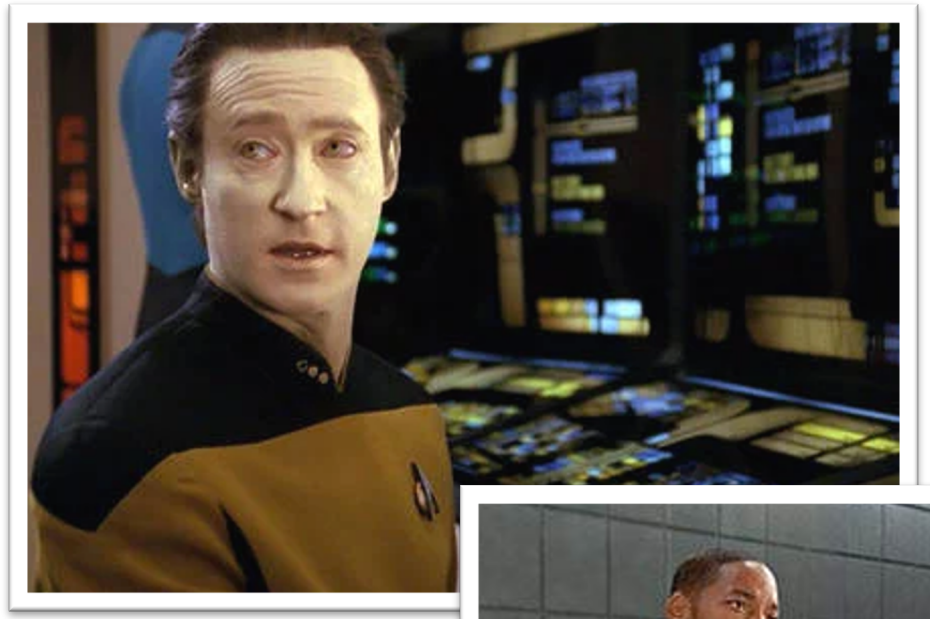
Product Introspection (2): Identify Potential Harms

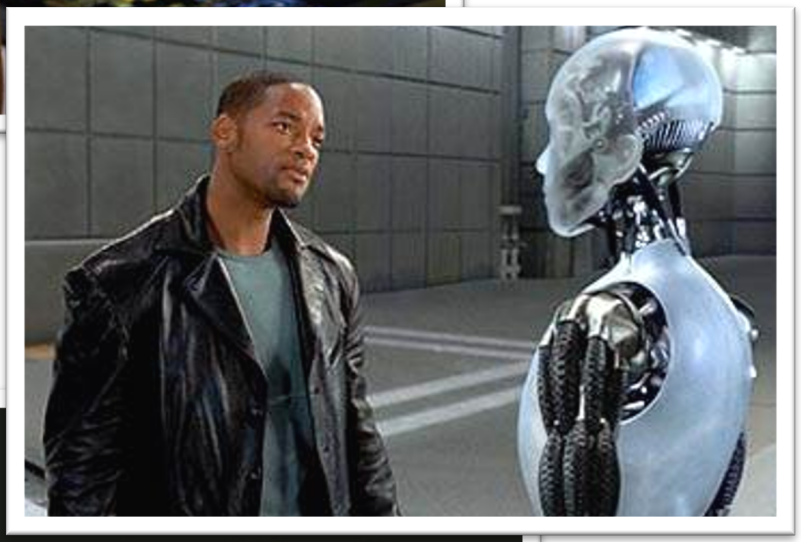
- What are the potential harms?
  - Applicants who would have repaid are not given loans
  - Convicts who would not commit a crime are locked up.
- Are there also longer term harms?
  - Applicants are given loans, then go on to default, harming their credit score
- Are some harms especially bad?

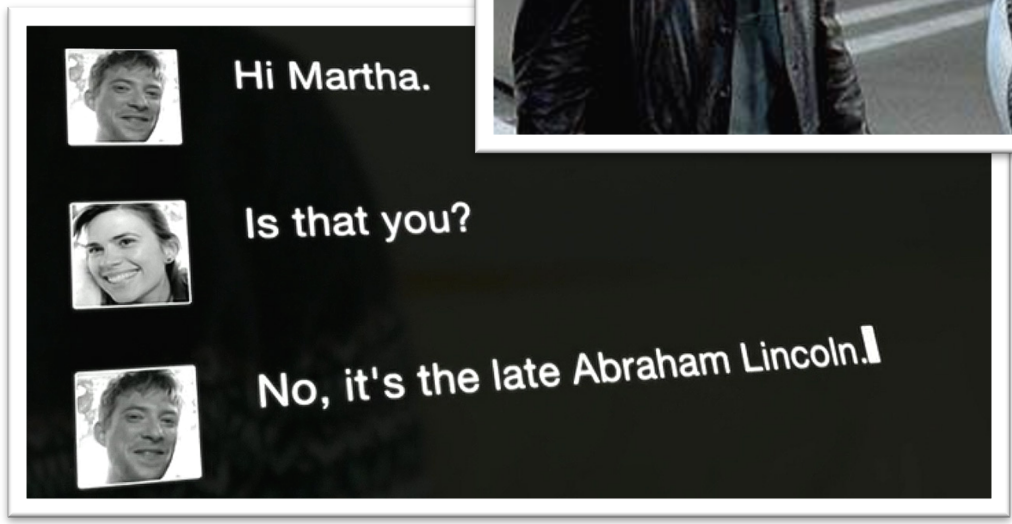
Seek out Diverse Perspectives

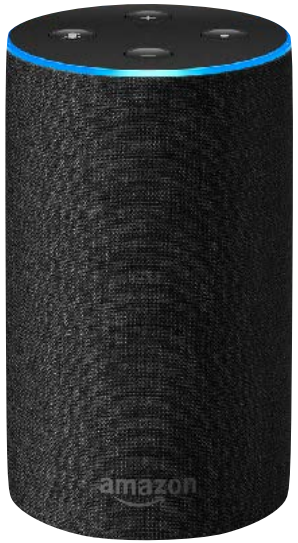
- Fairness Experts
- User Researchers
- Privacy Experts
- Legal
- Social Science Backgrounds
- Diverse Identities
  - Gender
  - Sexual Orientation
  - Race
  - Nationality
  - Religion

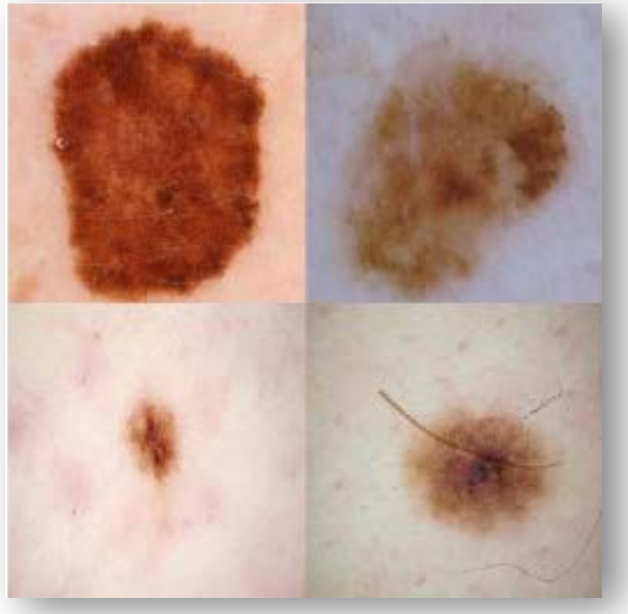
AI/ML and Humans

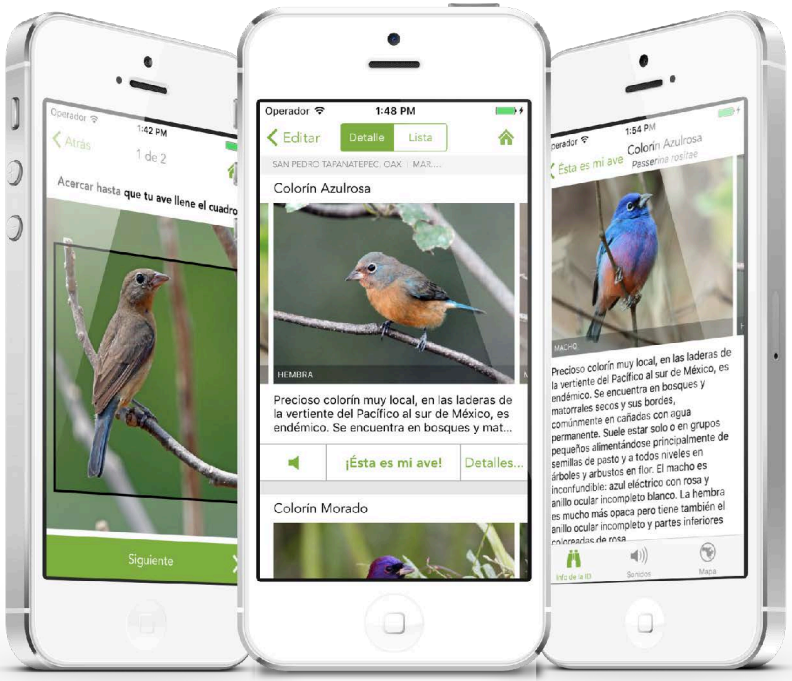







amazon




Operador
1:42 PM
Atrás
1 de 2
Acercar hasta que tu ave llene el cuadro
Siguiente
Operador
1:48 PM
Editar
Detalle
Lista
SAN PEDRO TAPANATEPEC, OAX.
Colorín Azulrosa
HEMBRA
Precioso colorín muy local, en las laderas de la vertiente del Pacífico al sur de México, es endémico. Se encuentra en bosques y mat...
¡Ésta es mi ave!
Detalles...
Colorín Morado
Operador
1:54 PM
¿Ésta es mi ave
Colorín Azulrosa
Passerina rositae
MACHO
Precioso colorín muy local, en las laderas de la vertiente del Pacífico al sur de México, es endémico. Se encuentra en bosques y matorrales secos y sus bordes, comúnmente en cañadas con agua permanente. Suele estar solo o en grupos pequeños alimentándose principalmente de semillas de pasto y a todos niveles en árboles y arbustos en flor. El macho es inconfundible: azul eléctrico con rosa y anillo ocular incompleto blanco. La hembra es mucho más opaca pero tiene también el anillo ocular incompleto y partes inferiores
Info de la ID
Sonidos
Mapa


WAYMO

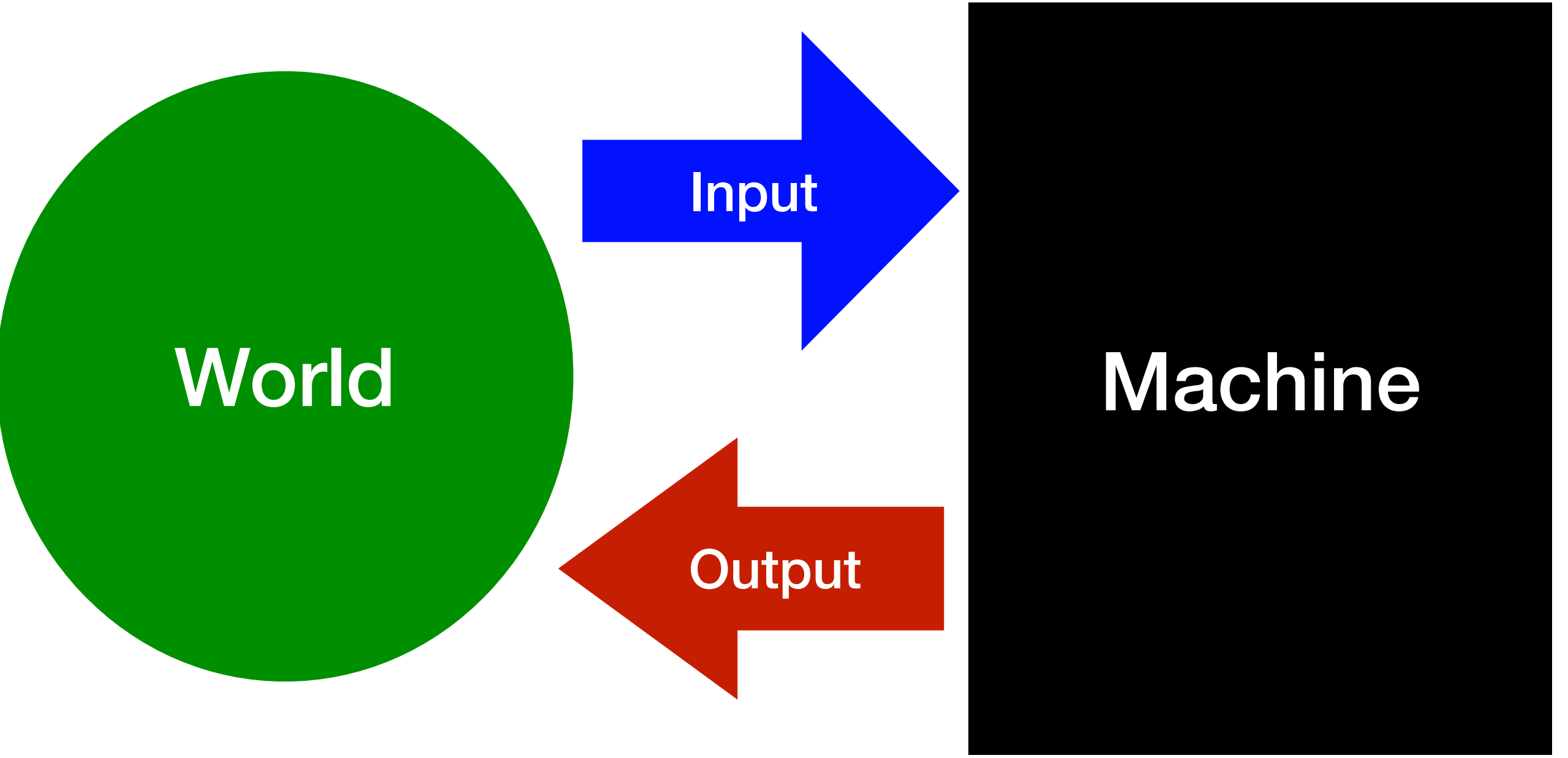

amazon

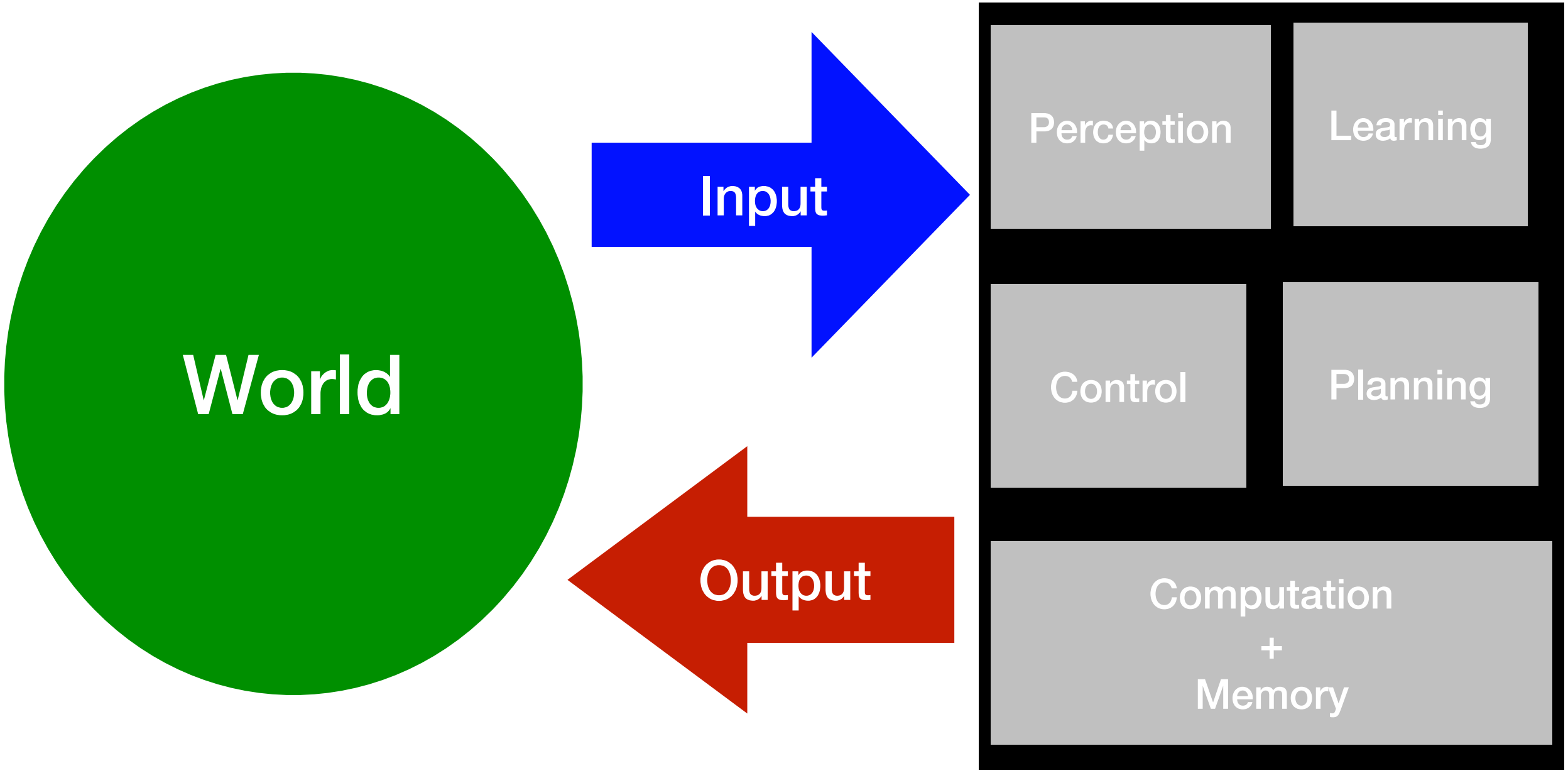
What ML Is Not

Error-free (no system is perfect)

100% confident

Intended to replace human judgement

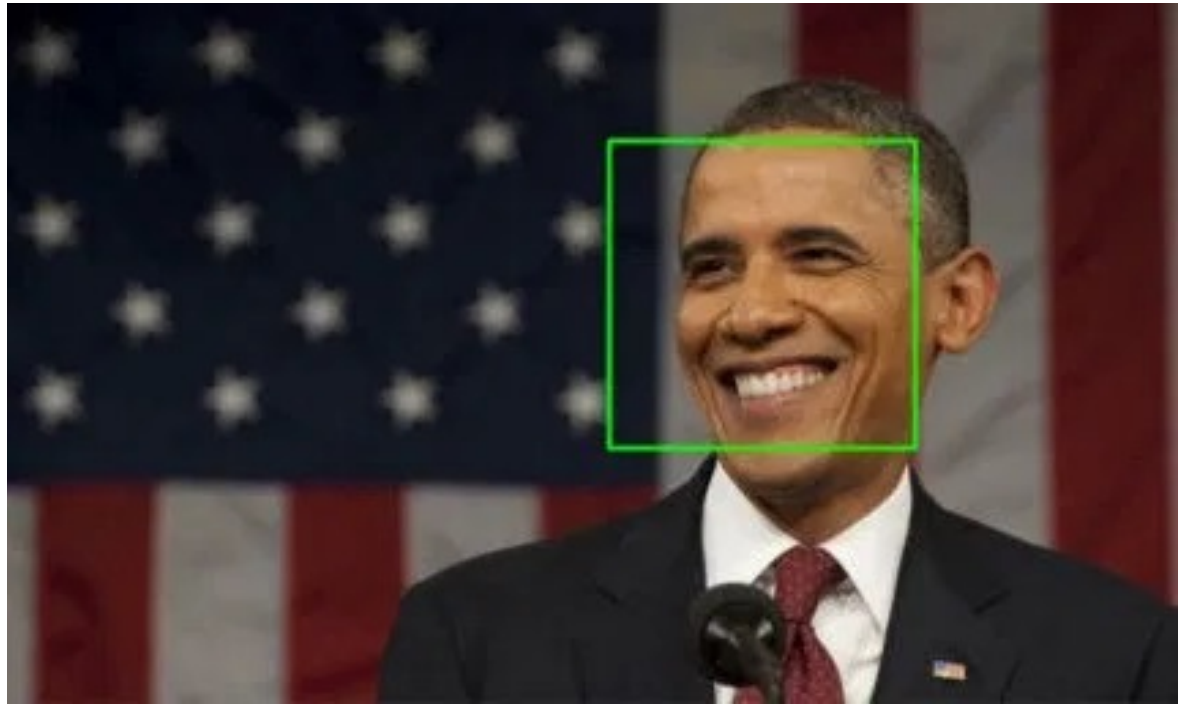

World
Input
Output
Machine


World
Input
Output
Perception
Learning
Control
Planning
Computation
+
Memory

Fairness Techniques in Faces

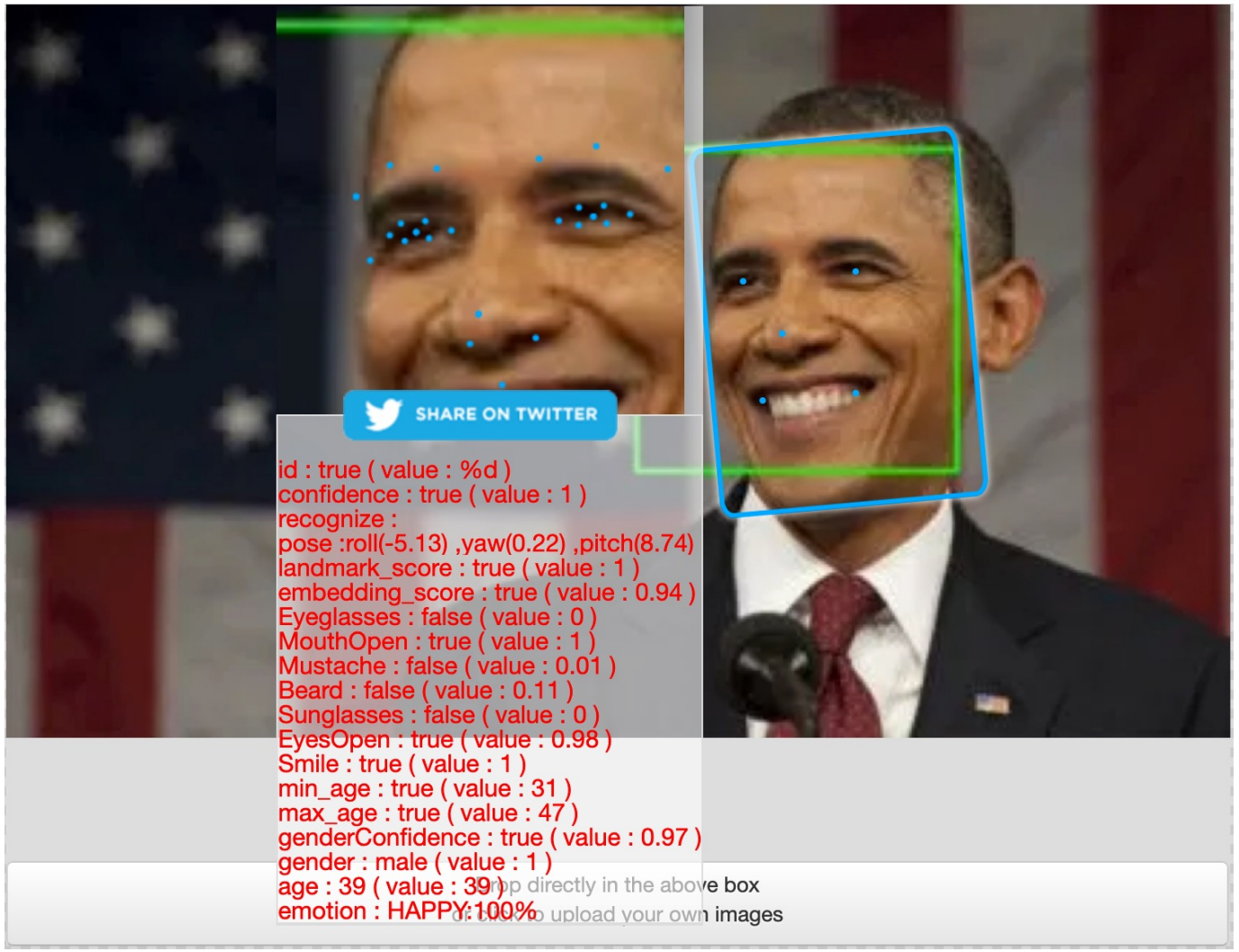
Face Detection

Detect presence of a face in an image or a video.



Face Analysis

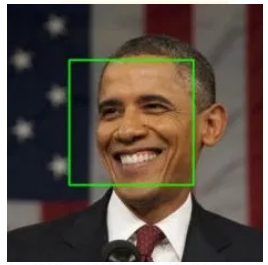
A system to determine the gender, age, emotion, presence of facial hair, etc. from a detected face.



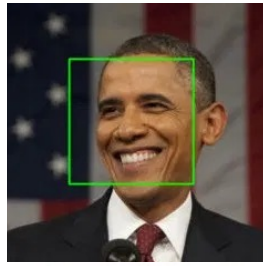
Face Recognition

A system to determine a detected faces identity by matching it against a database of faces and their associated identities.

- Face Authentication/Verification (1:1 matching)



- Face Identification/Recognition (1:N matching)



Confidence Score

Estimation of the confidence or certainty of any prediction

Expressed in the form of a probability or confidence score

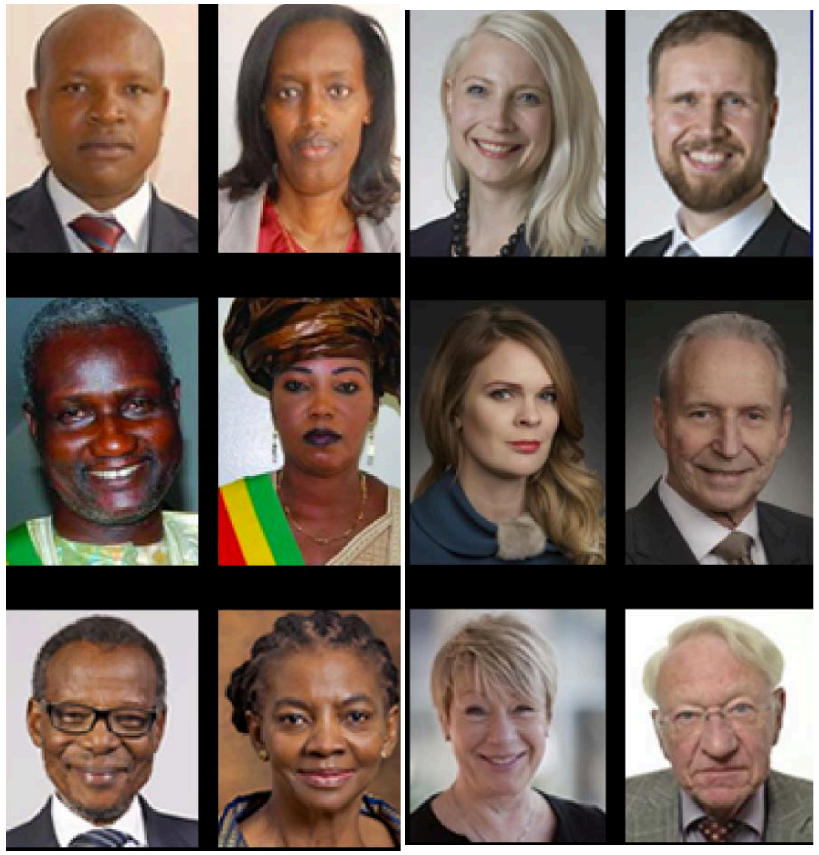
Face Recognition: Common Causes of Errors

ILLUMINATION VARIANCE	Lighting, camera controls like exposure, shadows, highlights
POSE / VIEWPOINT	Face pose, camera angles
AGING	Natural aging, artificial makeup
EXPRESSION / STYLE	Face expression like laughing, facial hair such as a beard, hair style
OCCLUSION	Part of the face hidden as in group pictures

Where Can Biases Exist?

The PPB dataset

AFRICAN SCANDINAVIAN



6.3% **20.8%**

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
A	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
B	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
C	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

GenderShades.Org

[Buolamwini & Gebru 2018]

Darker Skinned Female

Racial Comparisons of Datasets [FairFace]

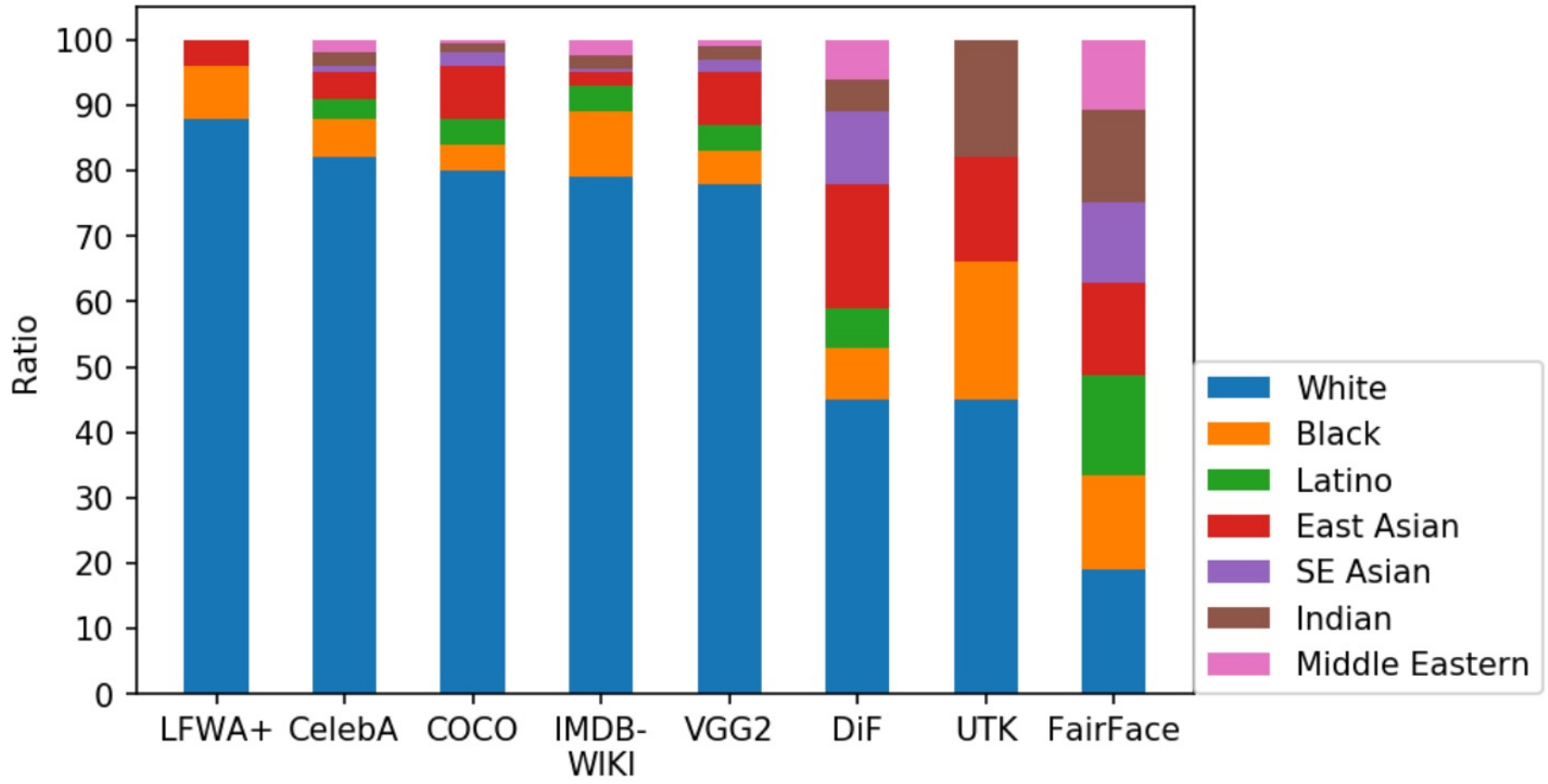


Figure 2: Racial compositions in face datasets.

Black Male

Latino Hispanic Male

Southeast Asian Female

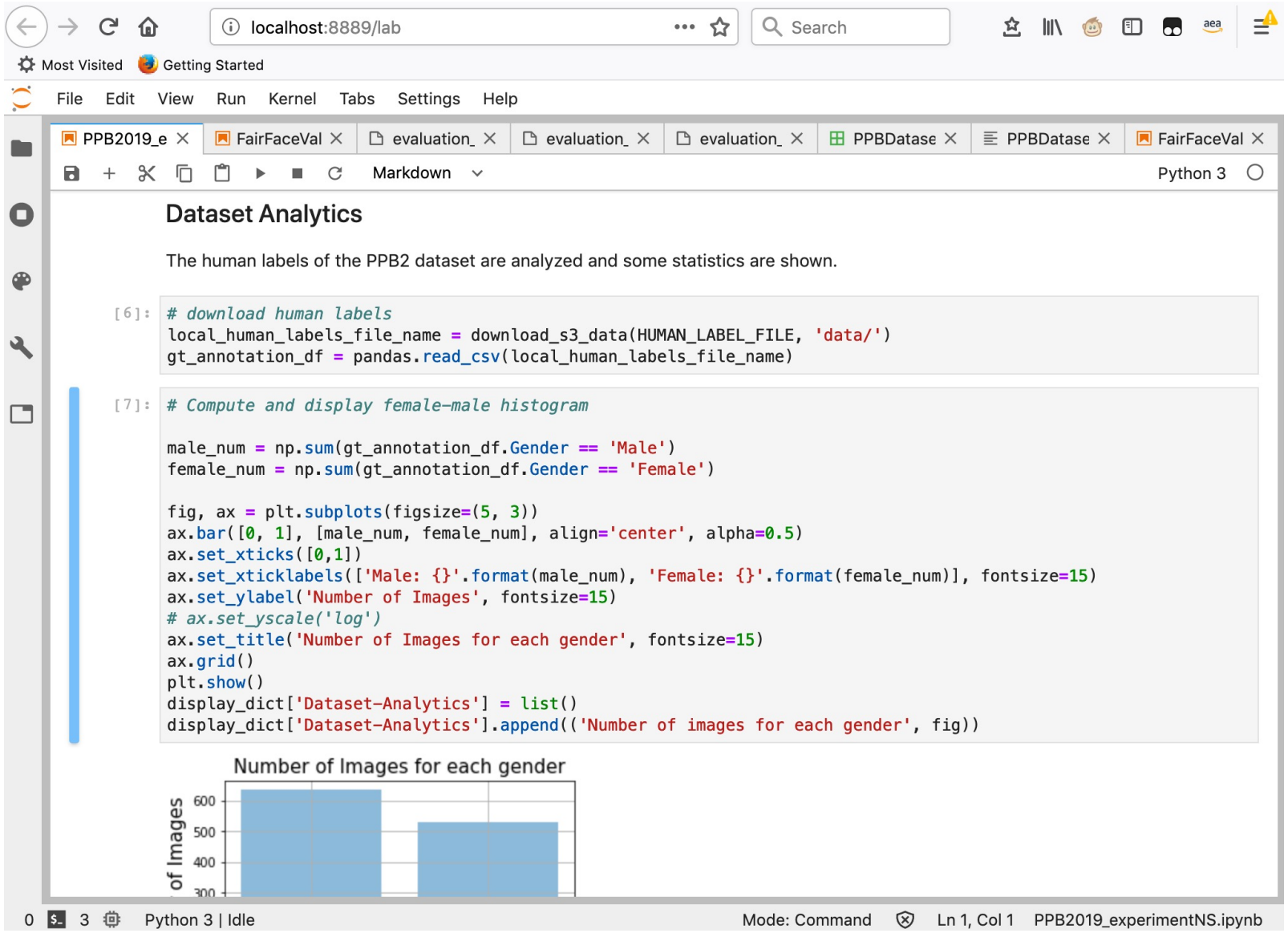

[Nvidia]

Launch with Confidence: Testing for Bias

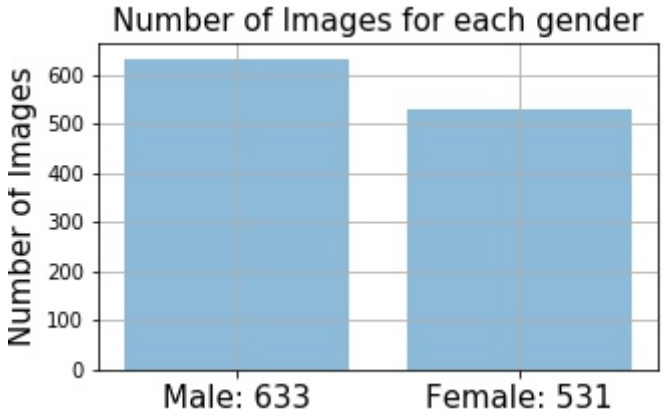
- How will you know if users are being harmed?
- How will you know if harms are unfairly distributed?

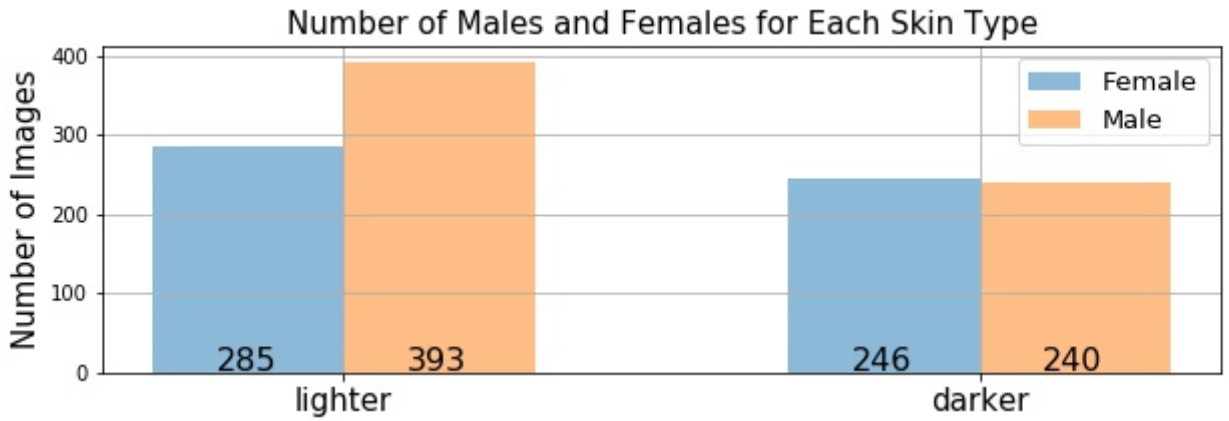
- Detailed testing practices are often not covered in academic papers
- Discussing testing requirements is a useful focal point for cross-functional teams

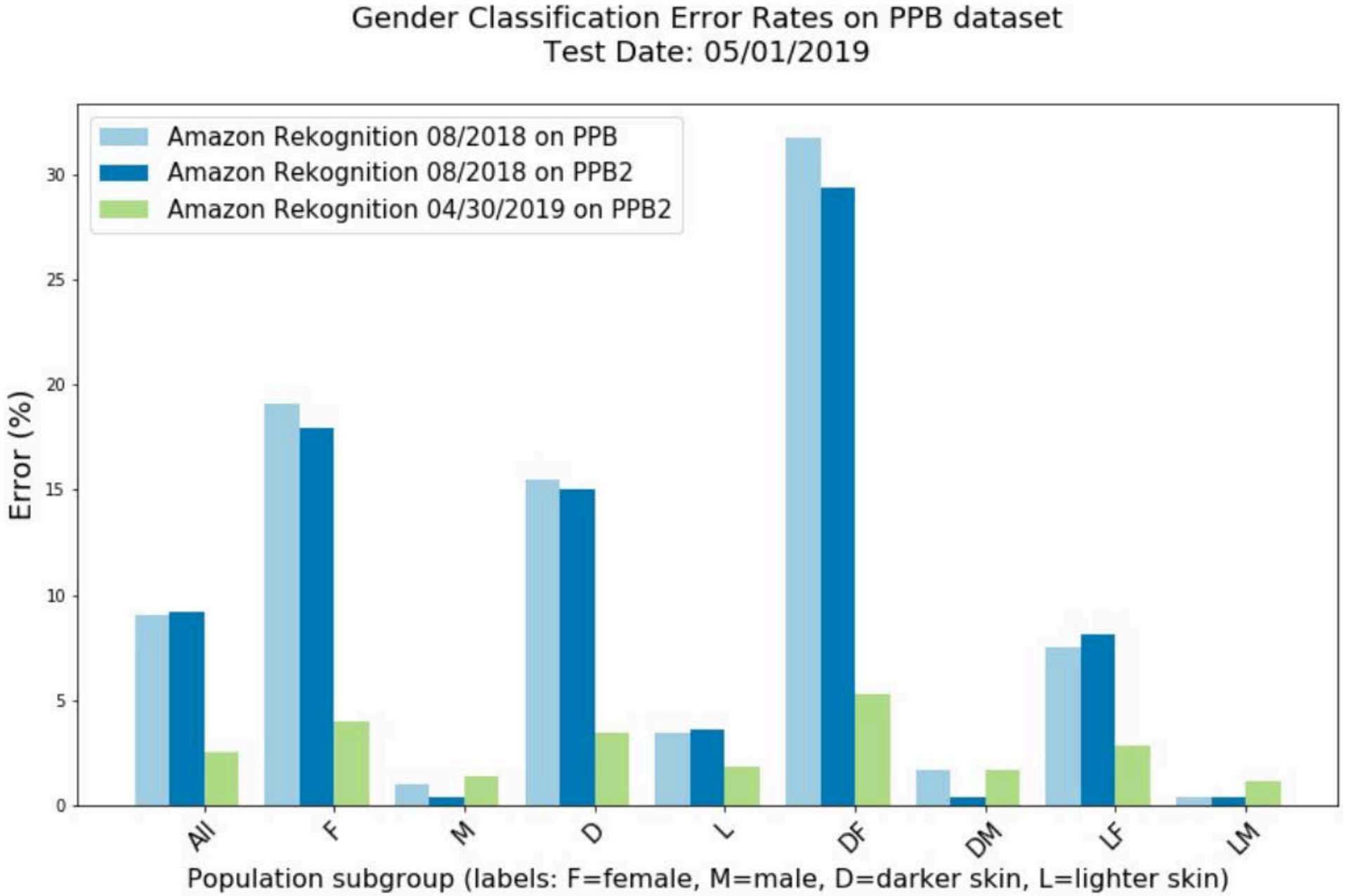
Reproducibility - Notebook Experiments



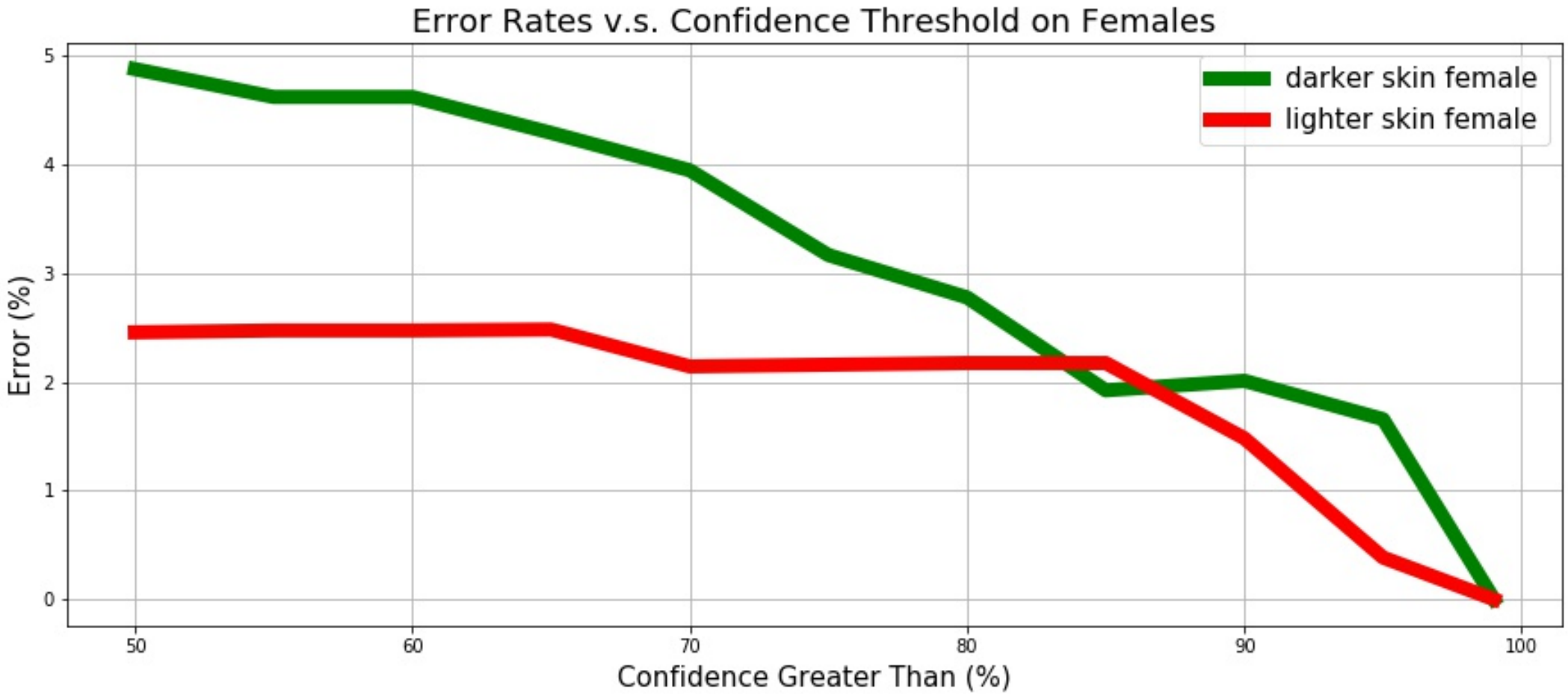
PPB2 Data Analytics






Gender Classification Error Rates on PPB dataset
Test Date: 05/01/2019
Amazon Rekognition 08/2018 on PPB
Amazon Rekognition 08/2018 on PPB2
Amazon Rekognition 04/30/2019 on PPB2
Error (%)
0
5
10
15
20
25
30
All
F
M
D
L
DF
DM
LF
LM
Population subgroup (labels: F=female, M=male, D=darker skin, L=lighter skin)

Gender Classification – PPB2



Correct classification samples (darker skin female) in API Amazon Rekognition 04/30/2019 on PPB2

Female
Confidence 99.94

Female
Confidence 99.92

Female
Confidence 99.91

Female
Confidence 99.91

Female
Confidence 99.91

Female
Confidence 99.87

Female
Confidence 99.84

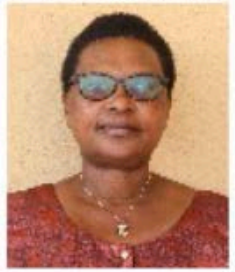
Female
Confidence 99.83

Female
Confidence 99.76

Female
Confidence 99.69

All errors (darker skin) in API Amazon Rekognition 04/30/2019 on PPB2

Female -> Male
Confidence 97.79



Female -> Male
Confidence 97.72

Female -> Male
Confidence 97.13

Female -> Male
Confidence 91.00

Female -> Male
Confidence 81.37

Female -> Male
Confidence 81.27

Female -> Male
Confidence 79.90

Female -> Male
Confidence 73.97

Female -> Male
Confidence 71.27

Female -> Male
Confidence 67.69

Female -> Male
Confidence 61.31

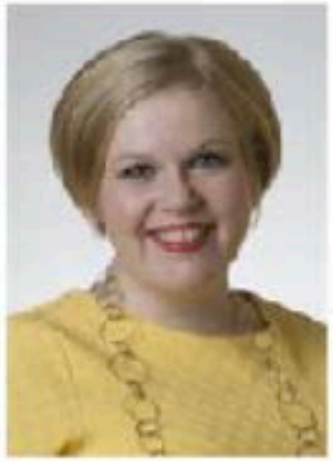
Female -> Male
Confidence 50.19

Correct classification samples (lighter skin female) in API Amazon Rekognition 04/30/2019 on PPB2

Female
Confidence 99.99

Female
Confidence 99.99

Female
Confidence 99.95



Female
Confidence 99.92

Female
Confidence 99.88

Female
Confidence 99.87

Female
Confidence 99.83

Female
Confidence 99.81

Female
Confidence 99.81

Female
Confidence 99.80

All errors (lighter skin) in API Amazon Rekognition 04/30/2019 on PPB2

Female -> Male
Confidence 96.06

Female -> Male
Confidence 94.95

Female -> Male
Confidence 94.80

Female -> Male
Confidence 93.79

Female -> Male
Confidence 88.01

Female -> Male
Confidence 86.26

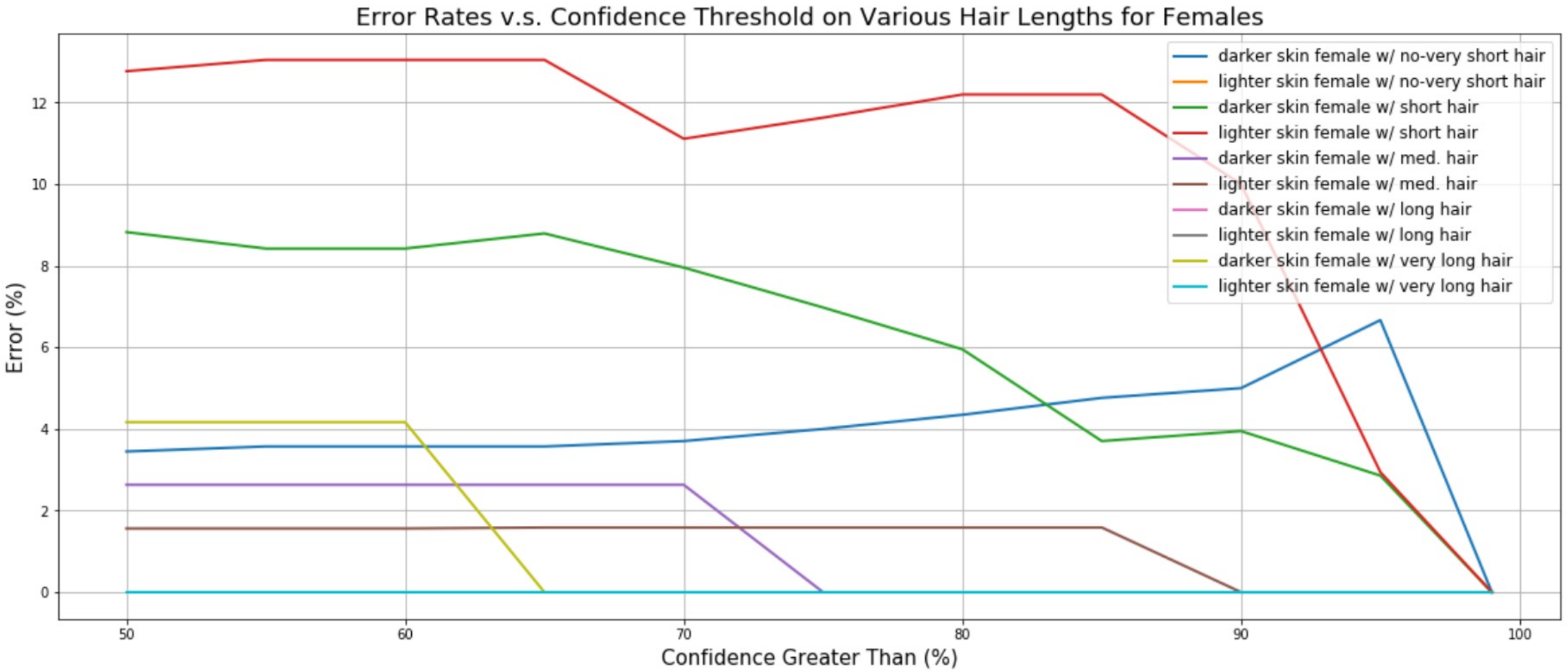
Female -> Male
Confidence 69.64

Short Hair

AWS Errors in Hairstyle: Short hair - 1-2 inches ears visible

AWS Errors in Hairstyle: Medium length covering ears

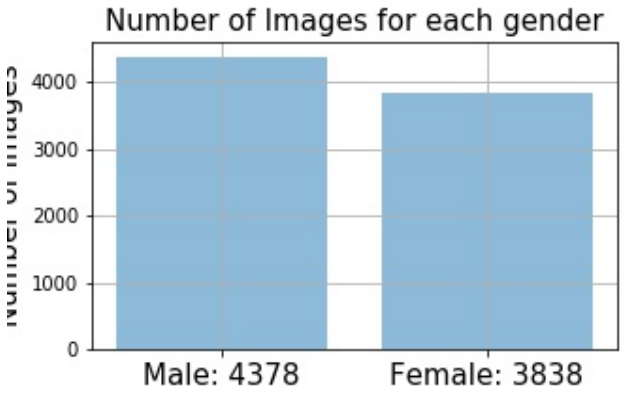
Gender Classification w.r.t. Hair Lengths – PPB2

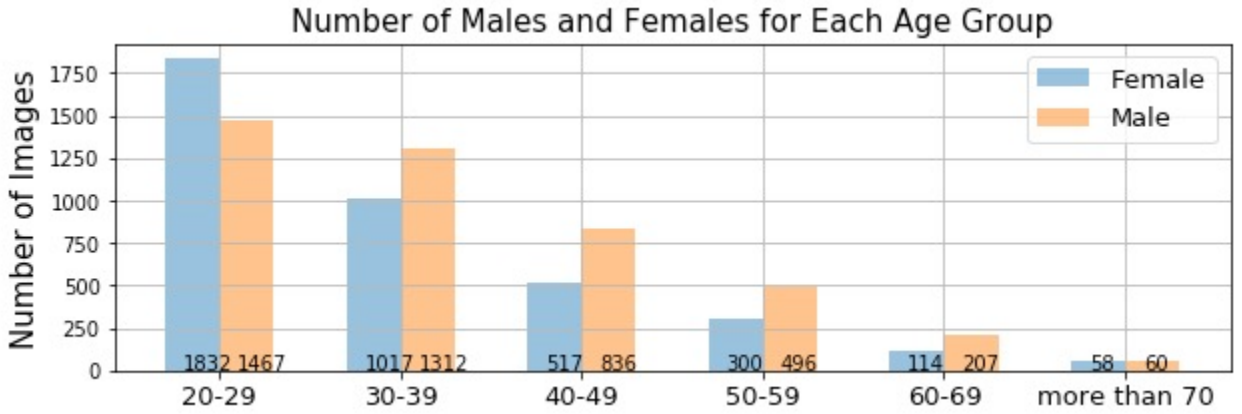


FairFace Dataset Analytics

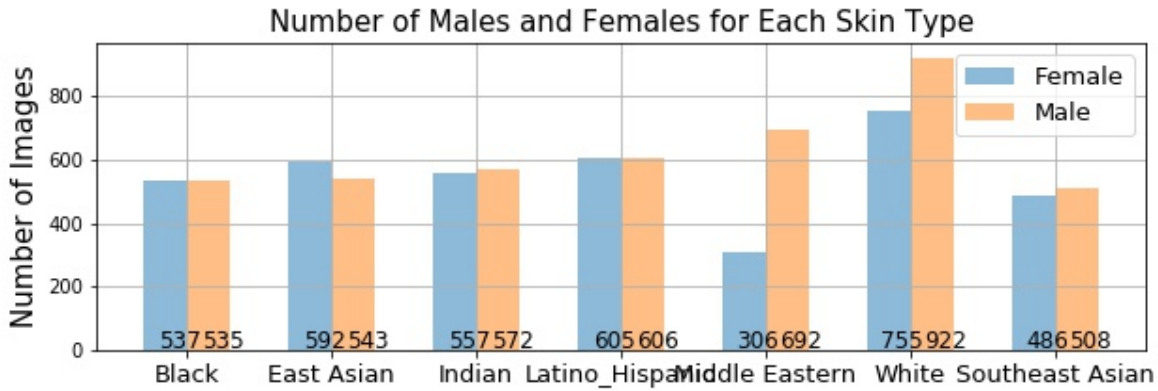
Dataset-Analytics

Number of images for each gender



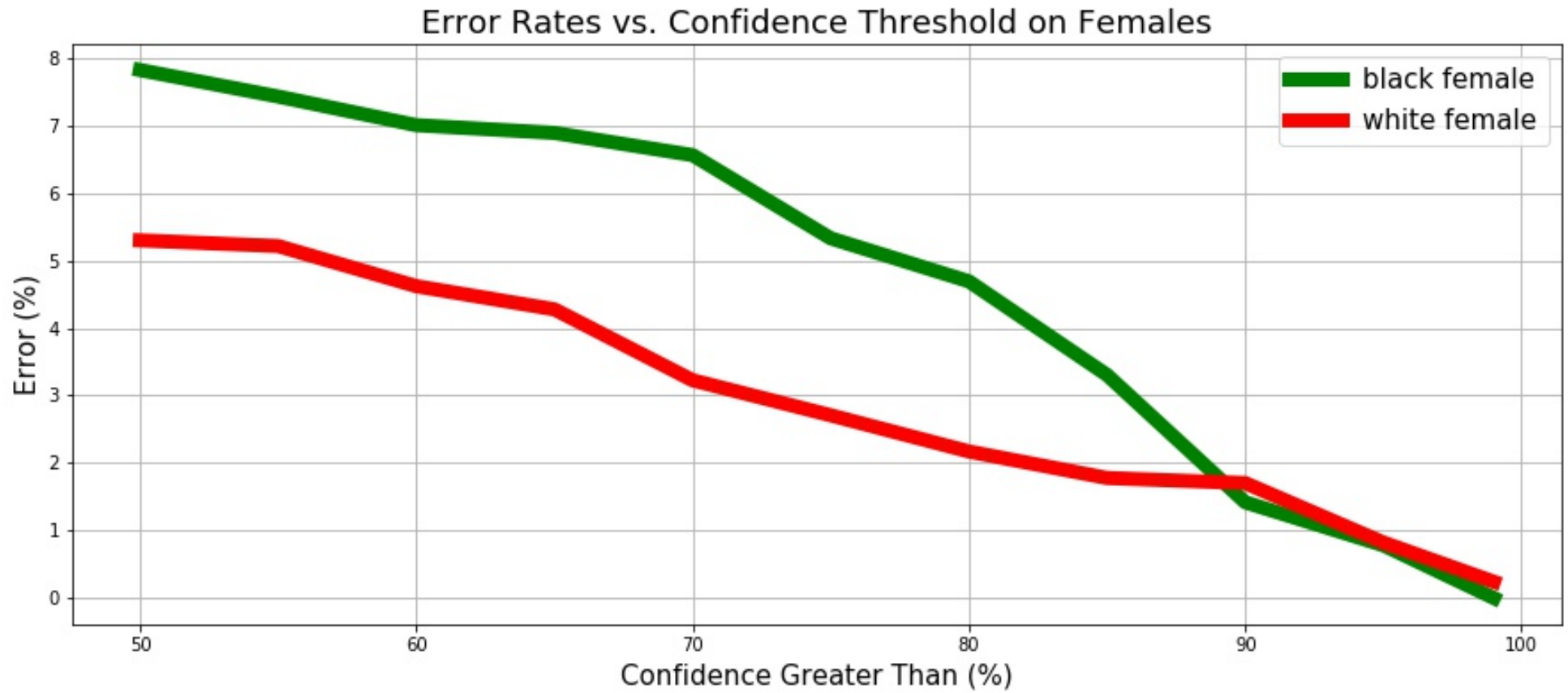


Number of images for each skin type group

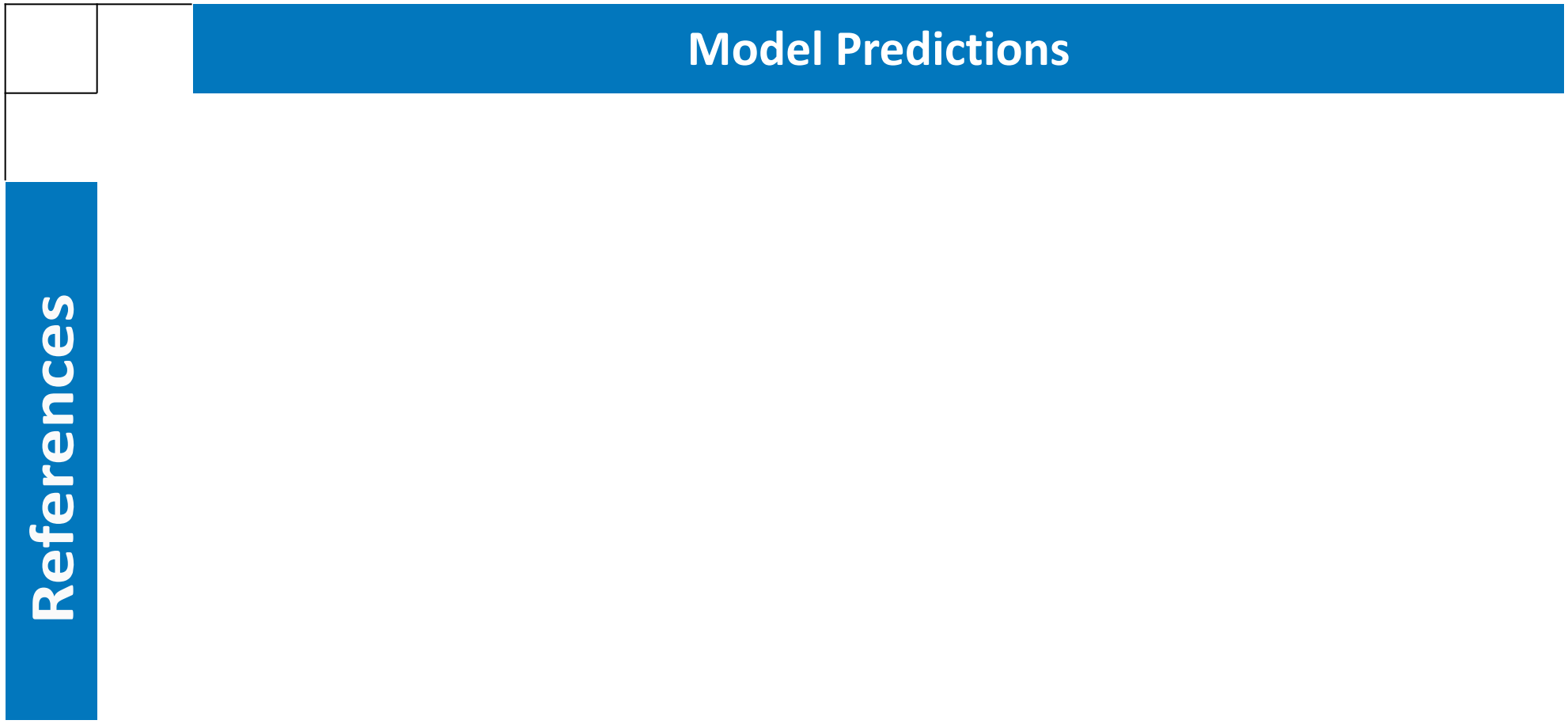


Gender classification – FairFace

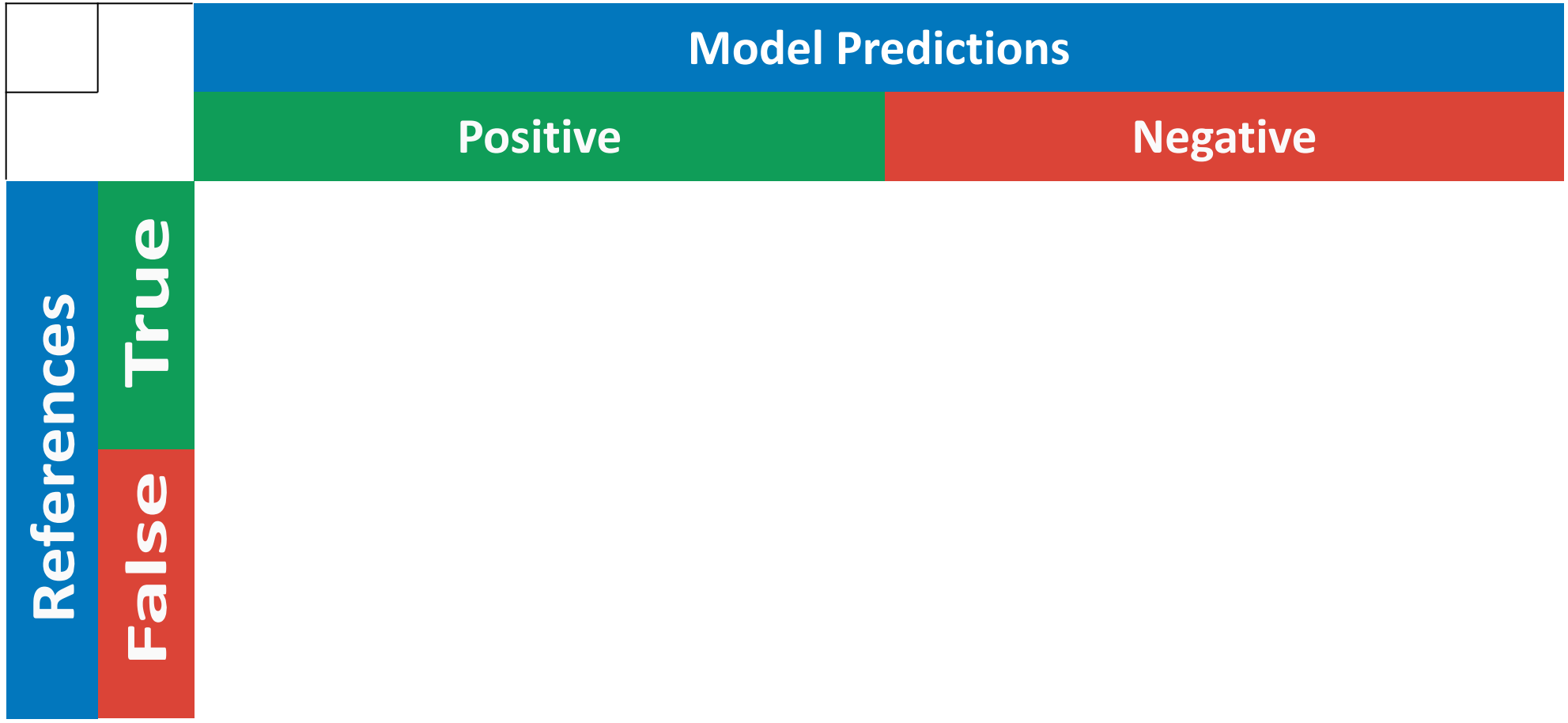
Error rates vs. confidence levels for female



Evaluate for Inclusion - Confusion Matrix



Evaluate for Inclusion - Confusion Matrix



Evaluate for Inclusion - Confusion Matrix

		Model Predictions	
		Positive	**Negative**
References	**True**	• Exists • Predicted **True Positives**	
	False		• Doesn't exist • Not predicted **True Negatives**

Evaluate for Inclusion - Confusion Matrix

		Model Predictions	
		Positive	**Negative**
References	**True**	• Exists • Predicted **True Positives**	• Exists • Not predicted **False Negatives**
	False	• Doesn’t exist • Predicted **False Positives**	• Doesn’t exist • Not predicted **True Negatives**

Efficient Testing for Bias

- Development teams are under multiple constraints
  - Time
  - Money
  - Human resources
  - Access to data
- How can we efficiently test for bias?
  - Prioritization
  - Strategic testing

Choose your evaluation metrics in light of acceptable tradeoffs between **False Positives** and **False Negatives**

Takeaways

- Testing for blindspots amongst intersectionality is key.
- Taking into account confidence scores/thresholds and error bars when measuring for biases is necessary.
- Representation matters.
- Transparency, reproducibility, and education can promote change.
- Confidence in your product's fairness requires fairness testing
- Fairness testing has a role throughout the product iteration lifecycle
- Contextual concerns should be used to prioritize fairness testing

AI Fairness and Transparency Tools

Mehrnoosh Sameki

Microsoft Azure

Microsoft

Machine Learning Transparency and Fairness

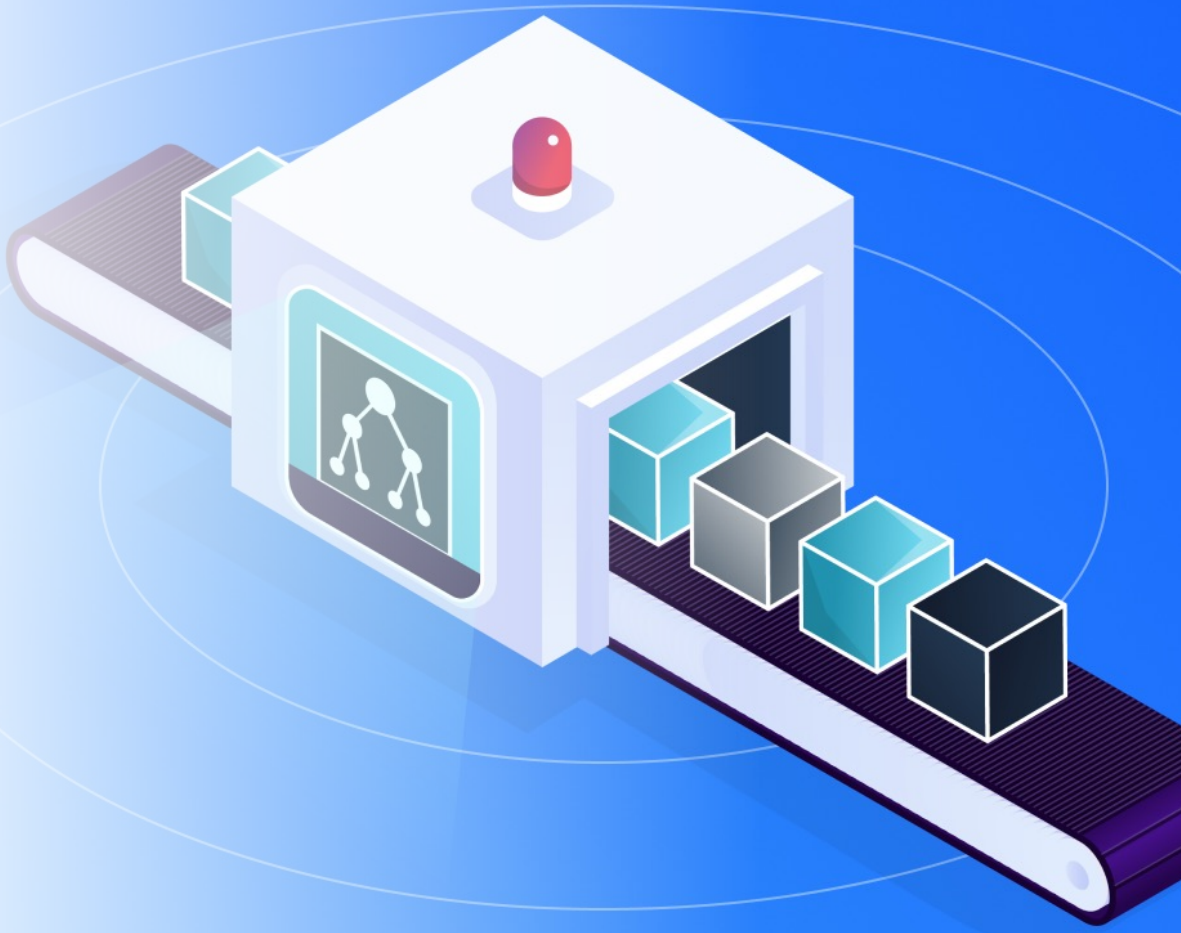
Model Designers/Evaluators
Training Time

End users or providers of solutions to end users
Inferencing Time

- Data scientists need to **explain the output of a model to stakeholders** (business, users, clients) to build trust
- Data scientists need tools to **debug their models** and make informed decision on how to improve them
- Data scientists need tools to verify if model's behavior **matches pre-declared objectives**

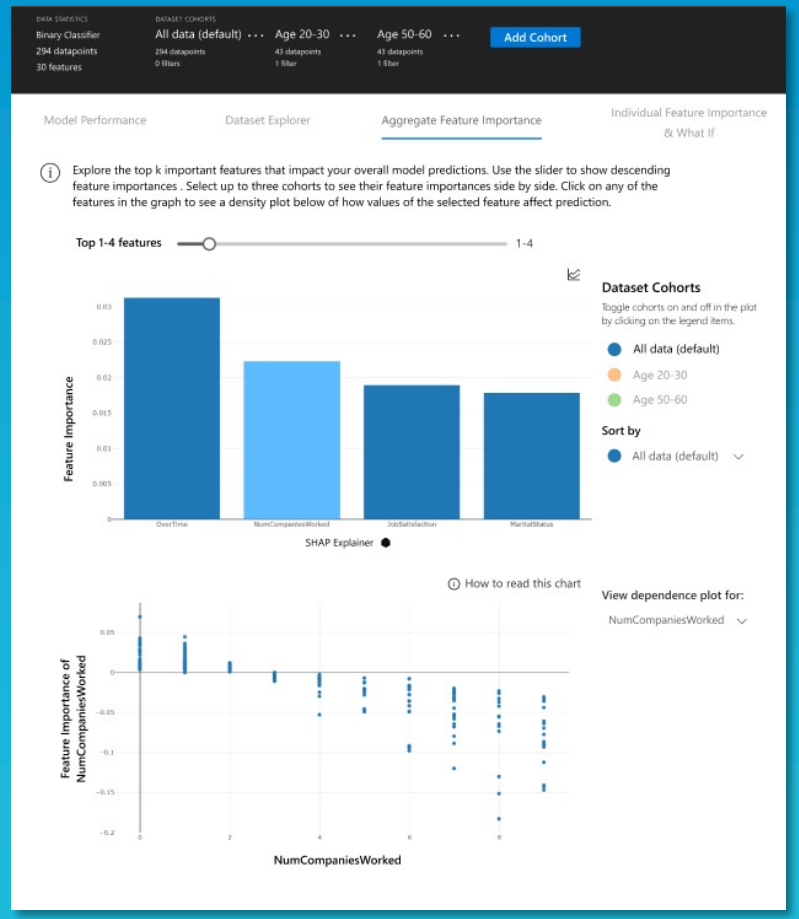
- AI predictions need to be explained at the inferencing time:
  - e.g., **health care**: Why the model classified Fabio at risk for colon cancer?
  - e.g., **finance**: Why Rosine was denied a mortgage loan or why his investment portfolio carries a higher risk?

Interpretability



InterpretML

Understand and debug your model



Interpret

Glassbox and blackbox interpretability methods for tabular data

Interpret-community

Community-driven interpretability techniques for tabular data

Interpret-text

Interpretability methods for text data

DiCE

Diverse Counterfactual Explanations

Blackbox Models:
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras,

Explainers: SHAP, LIME, Global Surrogate, Feature Permutation

Glassbox Models:
Model Types: Linear Models, Decision Trees, Decision Rules, Explainable Boosting Machines

Azureml-interpret

AzureML SDK wrapper for Interpret and Interpret-community

https://github.com/interpretml

Interpretability Approaches

Glassbox Models

Blackbox Explanations

Glassbox Models

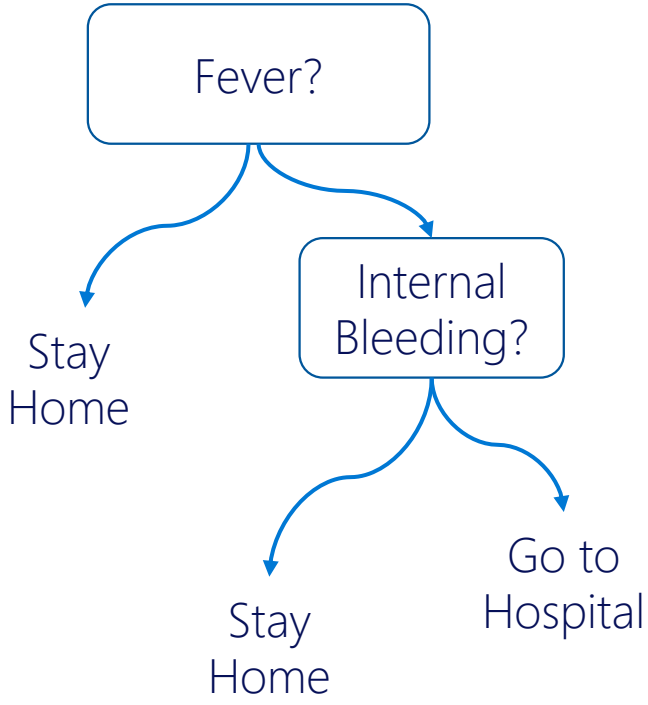
Models *designed* to be interpretable.
Lossless explainability.

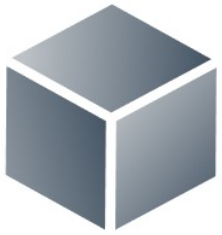
Decision Trees

Rule Lists

Linear Models

....





Blackbox Explanations

Explain *any* ML system. Approximate explainability.

SHAP

LIME

Partial Dependence

Sensitivity Analysis

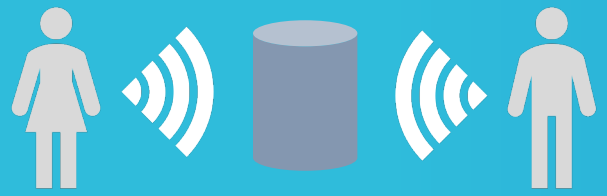
Fairness

Useful links:

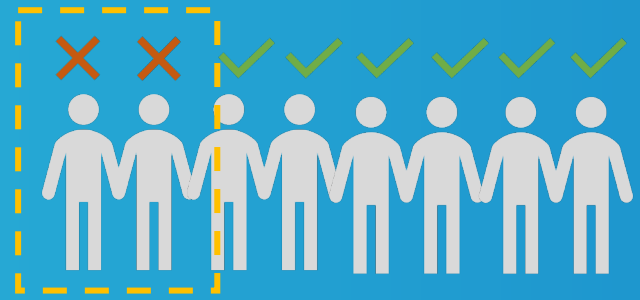
- AI Show
- Tutorial Video
- Customer Highlight

Fairness in AI

There are many ways that an AI system can behave unfairly.



A voice recognition system might fail to work as well for women as it does for men.

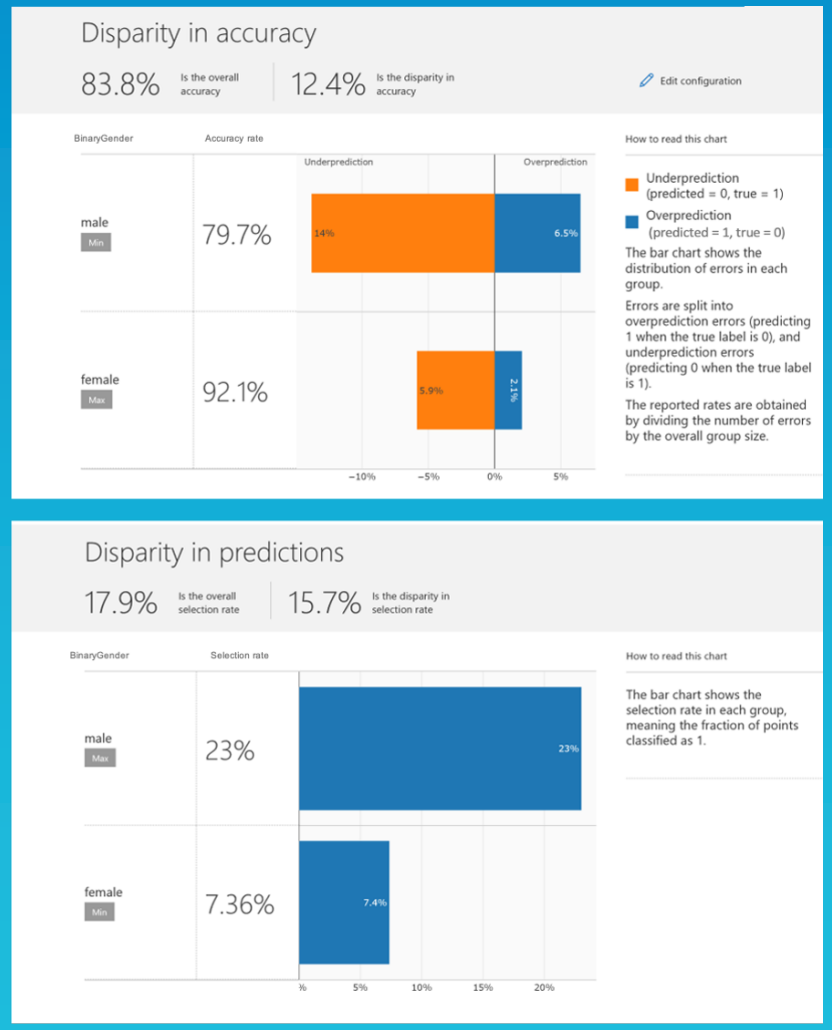


A model for screening loan or job application might be much better at picking good candidates among white men than among other groups.

Avoiding negative outcomes of AI systems for different groups of people

Fairlearn

Assessing unfairness in your model



https://github.com/fairlearn/fairlearn

1 → 2

Fairness Assessment:

Use common **fairness metrics** and an **interactive dashboard** to assess which groups of people may be negatively impacted.

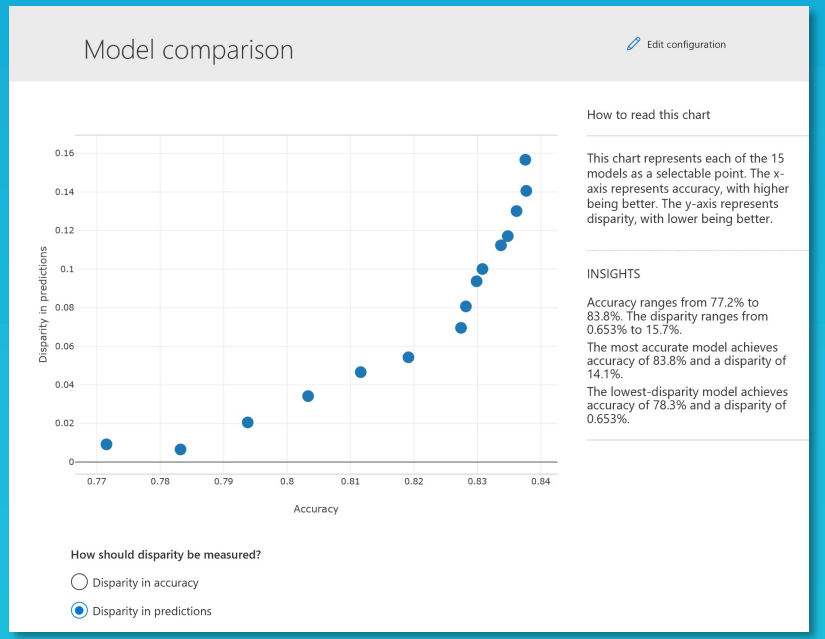
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras

Metrics: 15+ Common group fairness metrics

Model Types: Classification, Regression

Unfairness Mitigation:

Use state-of-the-art algorithms to mitigate unfairness in your classification and regression models.



Fairness Assessment

Input Selections

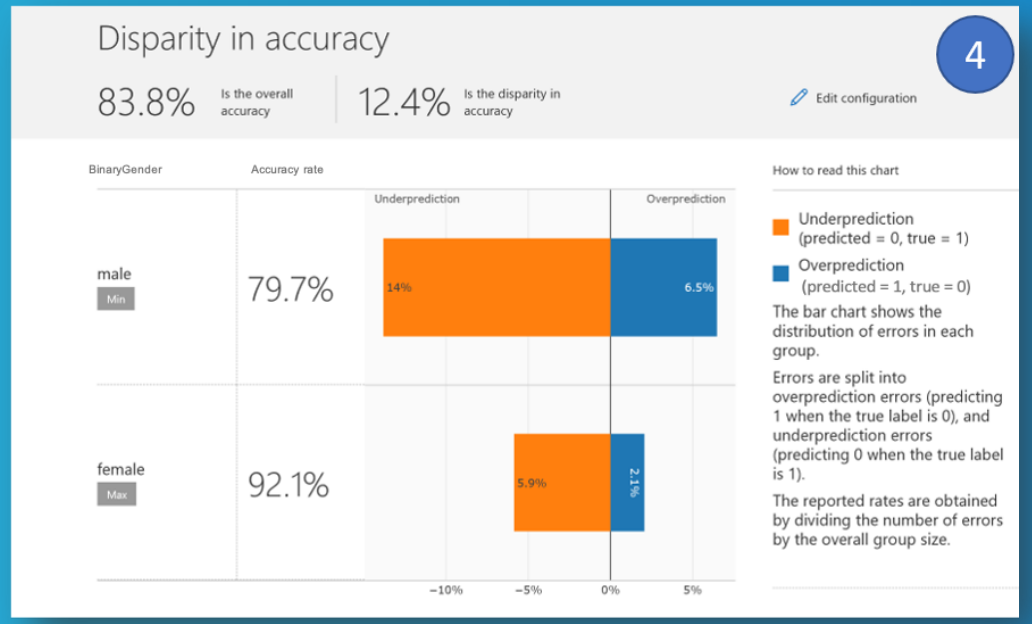
Sensitive attribute
Performance metric

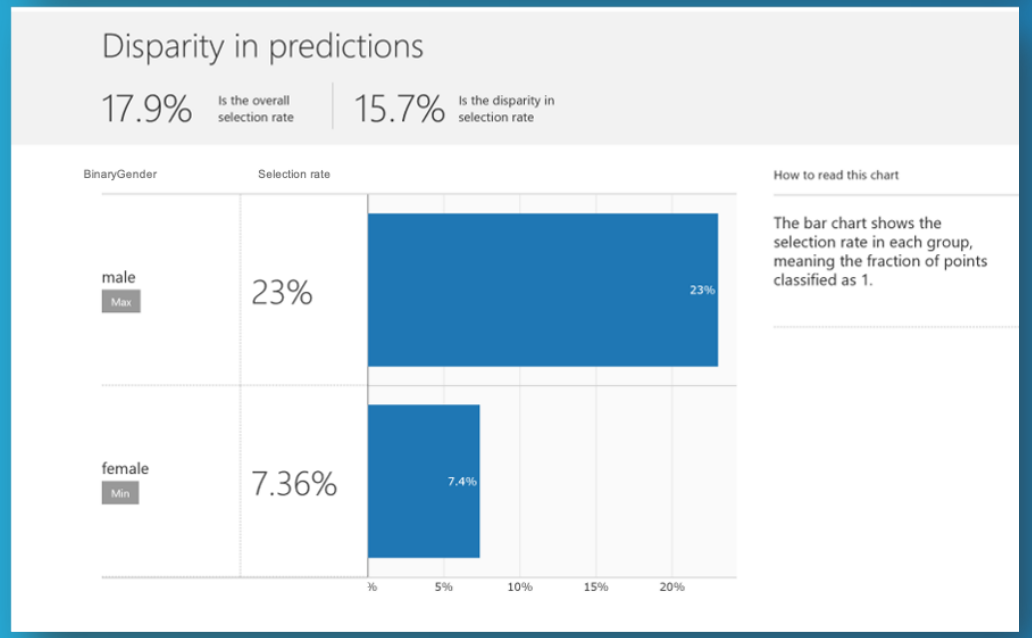
Assessment Results

Disparity in performance
Disparity in predictions

Mitigation Algorithms

Post-processing algorithm
Reduction Algorithm





PHILIPS
Healthcare

Putting fairness monitoring in production with ICU models

Philips Healthcare used Fairlearn to check whether our ICU models perform similarly for patients with different ethnicities and gender identities, etc.

Customer:
Philips

Industry:
Healthcare

Size:
80,000+ employees

Country:
Netherlands

Products and services:
Microsoft Azure DevOps
Microsoft Azure Databricks

MLFlow

Microsoft

Situation:

Philips Healthcare Informatics

Philips Healthcare Informatics helps ICUs benchmark their performance (e.g. mortality rate). They create quarterly benchmark reports that compare actual performance vs performance predicted by ML models. They have models trained on the largest ICU dataset in USA: 400+ ICUs, 6M+ patient stays, billions of vital signs & lab tests.

Deploying ICU models responsibly

Philips needed a scalable, reliable, repeatable and **responsible way to bring ML models into production.**

Solution:

Microsoft CSE (Led by Tempest Van Schaik) collaborated with Philips to build a solution using **Azure DevOps pipelines**, **Azure Databricks** and **MIflow**. Built a pipeline to make fairness monitoring routine, checking the fairness of predictions for patients of different **genders, ethnicities, and medical conditions**, using **Fairlearn** metrics.

Fairness analysis helped show that Philips' predictive model performs better than industry standard ICU models

Standard model predictions for a patient differ depending on how the ICU documented their test results.

EY Building a better working world

Customer:
EY

Industry:
Partner Professional Services

Size:
10,000+ employees

Country:
United Kingdom

Products and services:
Microsoft Azure
Microsoft Azure Machine Learning

Read full story here

Microsoft

“Azure Machine Learning and its Fairlearn capabilities offer advanced fairness and explainability that have helped us deploy trustworthy AI solutions for our customers, while enabling stakeholder confidence and regulatory compliance.”

—Alex Mohelsky, Partner and Advisory Data, Analytic, and AI Leader, EY Canada

Situation:

Organizations won’t fully embrace AI until they trust it. EY wanted to help its customers embrace AI to help them better understand their customers, identify fraud and security breaches sooner, and make loan decisions faster and more efficiently.

Solution:

The company developed its EY Trusted AI Platform, which uses Microsoft Azure Machine Learning capabilities to assess and mitigate unfairness in machine learning models. Running on Azure, the platform uses Fairlearn and InterpretML, open-source capabilities in Azure Machine Learning.

Impact:

When EY tested Fairlearn with real mortgage data, it reduced the accuracy disparity between men and women approved or denied loans from 7 percent to less than 0.5 percent. Through this platform, EY helps customers and regulators alike gain confidence in AI and machine learning.

Reception & Adoption

Toolkits	Average SUS	StdDev SUS
Aequitas Tool	61.33	15.78
Fairlearn	65.71	12.99
Google What-if tool	60.33	17.14
IBM Fairness 360	54.50	13.89
PyMetrics Audit AI	58.04	10.29
Scikit-fairness	62.83	17.32
All	*60.43*	*14.84*

Table 1: Toolkit System Usability Survey Scores

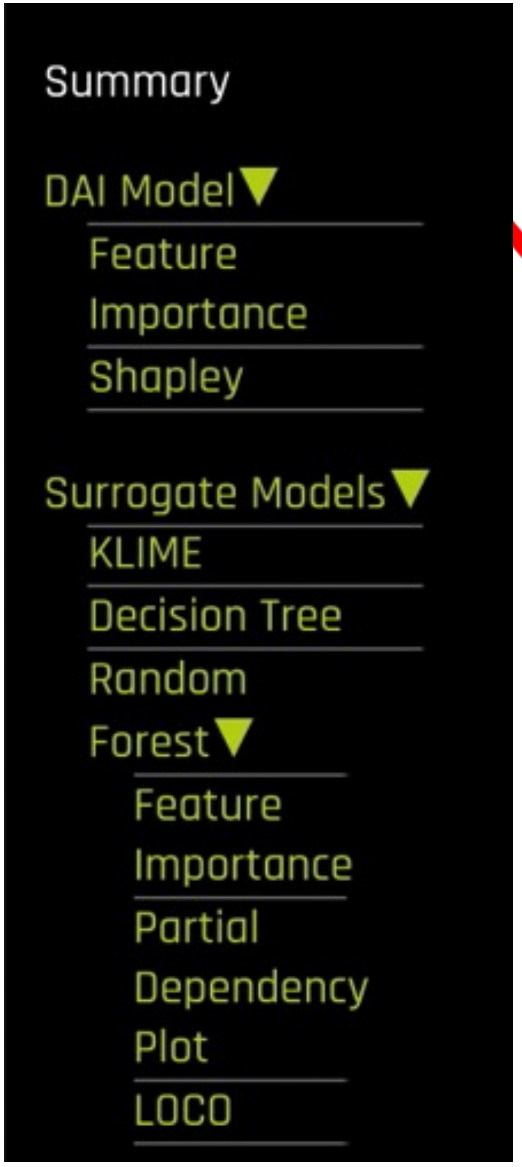
Educational materials are key for adoption of fairness toolkits

→ manuals, case studies, white papers

[Lee & Singh, 2020]

H2O

- Interpretability at training time
- Combination of glass-box models and black-box explainers
- Auto reason code generation for local predictions
- Ability to cross reference to other techniques to ensure stability and consistency in results


Summary
DAI Model
Feature Importance
Shapley
Surrogate Models
KLIME
Decision Tree
Random Forest
Feature Importance
Partial Dependency Plot
LOCO


MLI: Regression and Classification Explanations
DATASETS EXPERIMENTS MLI
Value: 8177
SEARCH
Legend: Local Global
Cluster: cluster6
EXPLANATIONS
About this plot
Summary
DAI Model
Feature Importance
Shapley
Surrogate Models
KLIME
Decision Tree
Random Forest
Feature Importance
Partial Dependency Plot
LOCO
Dashboard
MLI Docs
Download MLI Logs
Experiment
Scoring Pipeline
Download Reason Codes
Cluster 6 LIME Plot R2: 94.43% RMSE: 0.029
Model Prediction
LIME Model Prediction
Actual Target
Pr(default payment next month)
Ranked Predictions
Feature Importance
PAY_0
PAY_2
PAY_3
PAY_5
PAY_4
LIMIT_BAL
PAY_6
PAY_AMT1
Decision Tree
Training RMSE = 0.0762
R2 = 0.852
PAY_0
-1,-2,+2 more+NA
2,3,4,+4 more
PAY_2
PAY_3
BILL_AMT1
PAY_5
PAY_5
PAY_5
0.212
0.128
0.384
0.523
0.407
0.518
0.670
0.763
PDP Variable: PAY_2
Partial Dependence: PAY_2
Mean Response with ± 1 Standard Deviation (Shaded for numeric data and bars for categorical data)
0.83

IBM Open Scale

Goal: to provide AI operations team with a toolkit that allows for:

- Monitoring and re-evaluating machine learning models after deployment

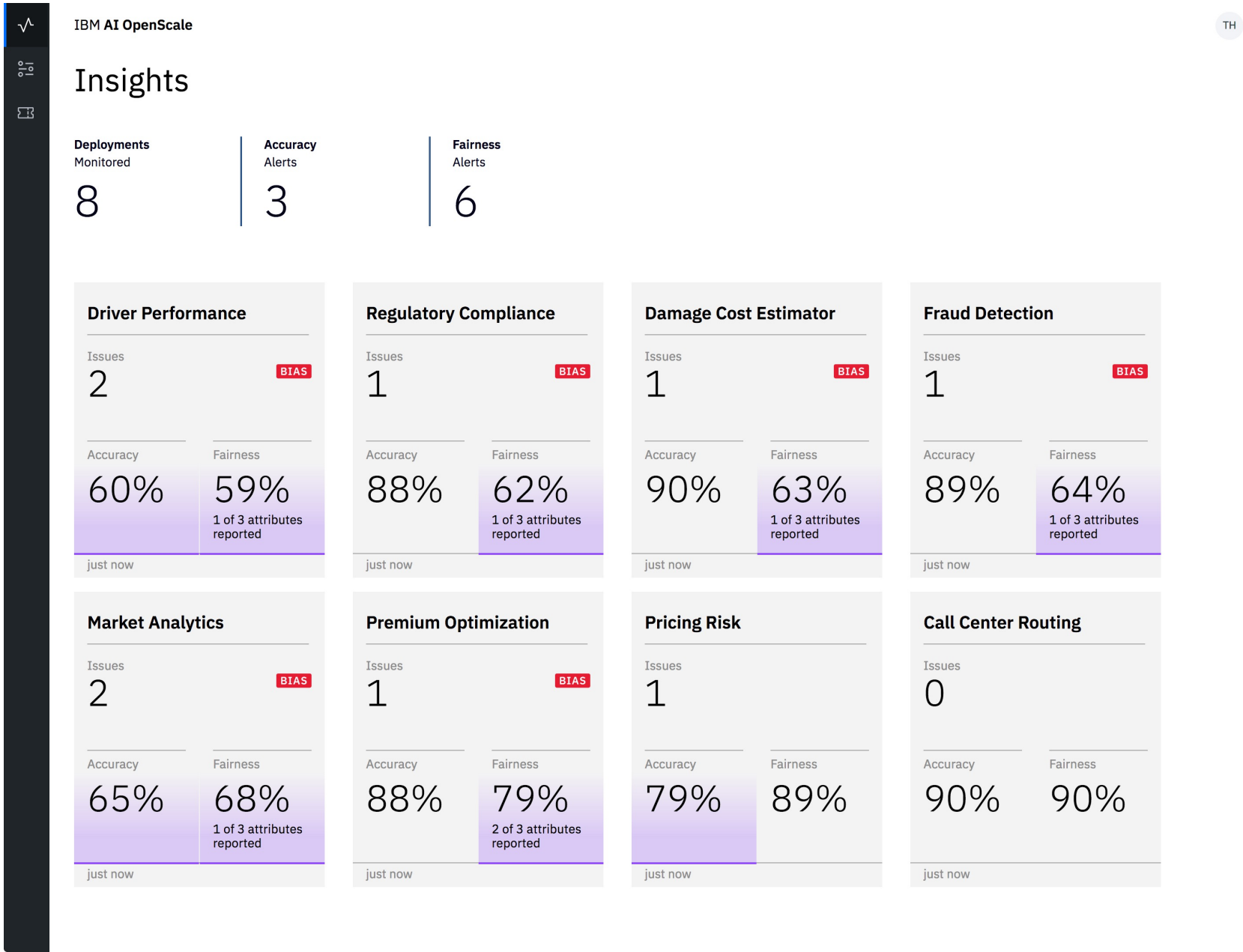
IBM Open Scale

Goal: to provide AI operations team with a toolkit that allows for:

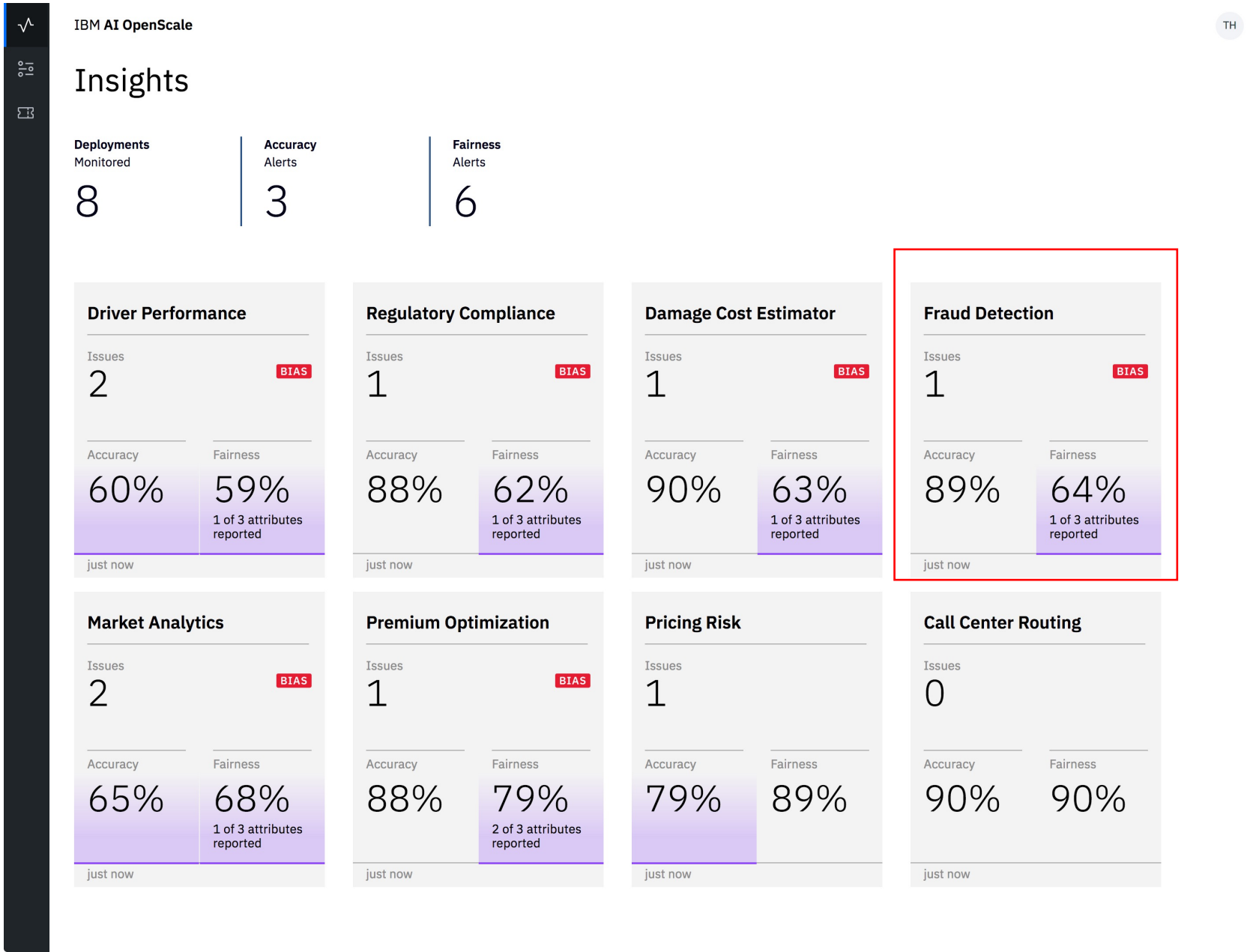
- Monitoring and re-evaluating machine learning models after deployment

- ACCURACY
- FAIRNESS
- PERFORMANCE

IBM Open Scale



IBM Open Scale

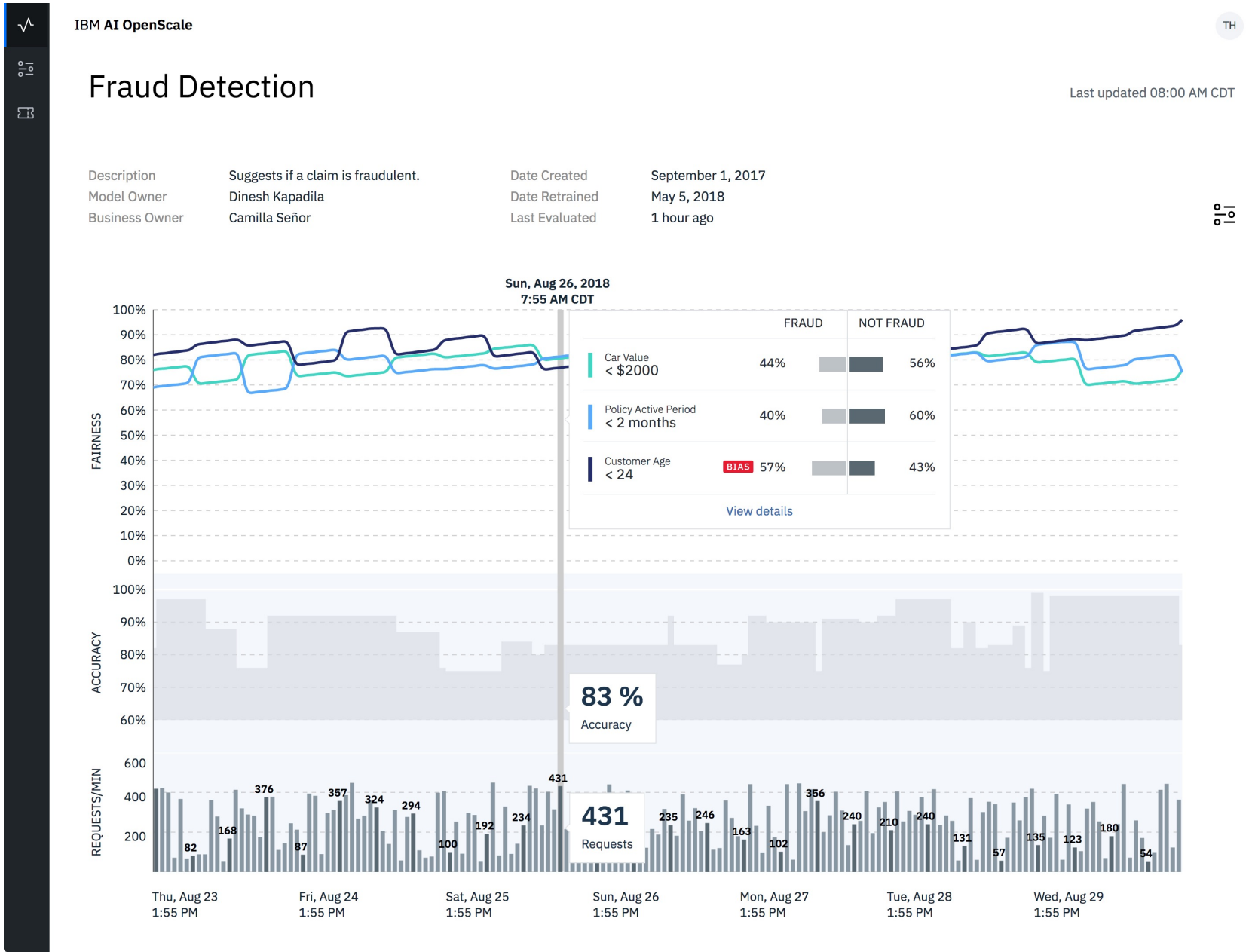


IBM Open Scale

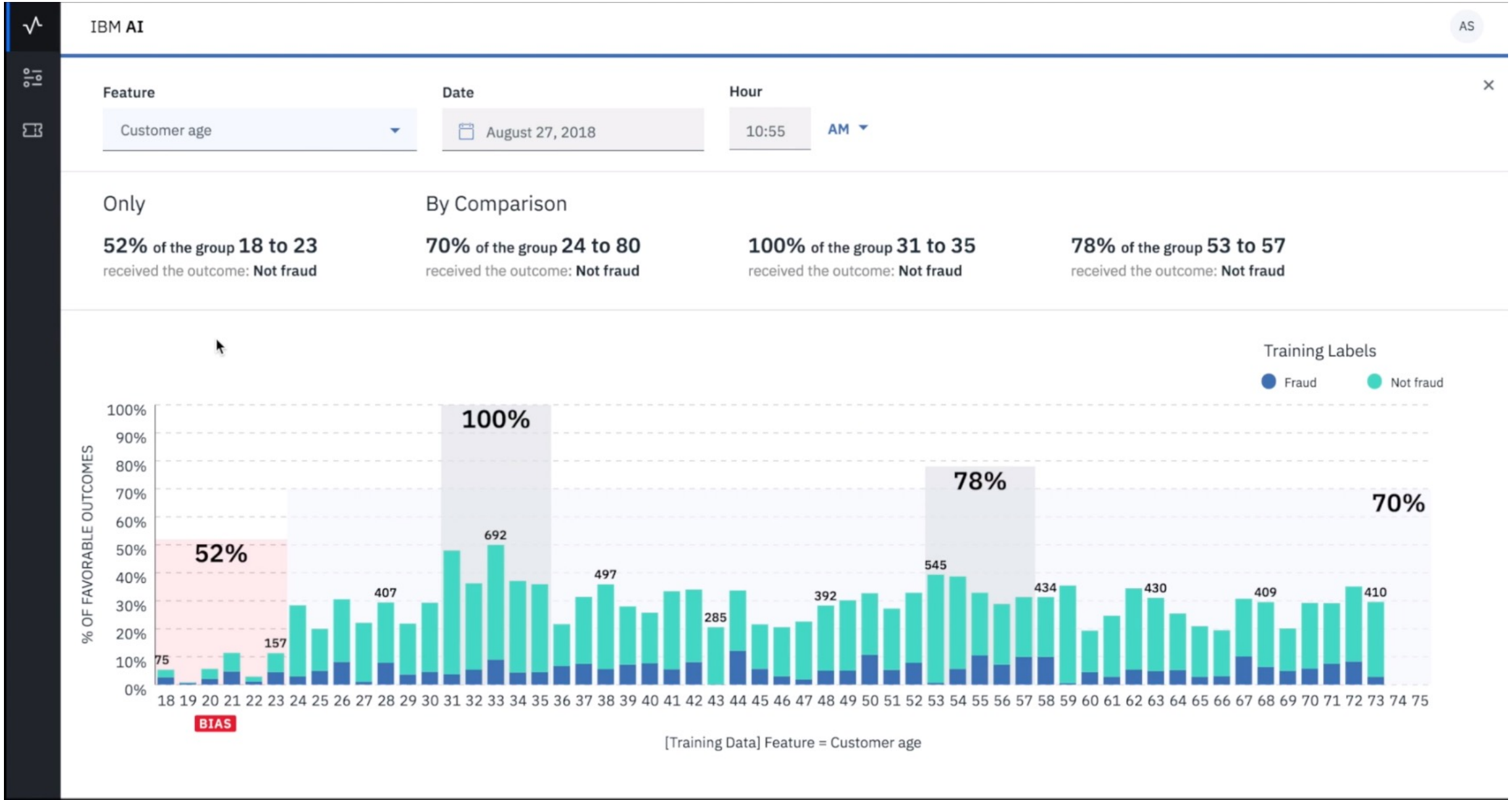
Fairness

Accuracy

Performance



IBM Open Scale



AI Fairness 360

Datasets

Toolbox

- Fairness metrics (30+)
- Bias mitigation algorithms (9+)

Guidance

- Industry-specific tutorials

AI Fairness 360

Datasets

Toolbox

Fairness metrics (30+)

Bias mitigation algorithms (9+)

Guidance

Industry-specific tutorials

Optimized Preprocessing (Calmon et al., NIPS 2017)

Meta-Algorithm for Fair Classification (Celis et al., FAT* 2019)

Disparate Impact Remover (Feldman et al., KDD 2015)

Equalized Odds Postprocessing (Hardt et al., NIPS 2016)

Reweighing (Kamiran and Calders, KIS 2012)

Reject Option Classification (Kamiran et al., ICDM 2012)

Prejudice Remover Regularizer (Kamishima et al., ECML PKDD 2012)

Calibrated Equalized Odds Postprocessing (Pleiss et al., NIPS 2017)

Learning Fair Representations (Zemel et al., ICML 2013)

Adversarial Debiasing (Zhang et al., AIES 2018)

Pre-processing algorithm:
a bias mitigation algorithm that is applied to training data

In-processing algorithm:
a bias mitigation algorithm that is applied to a model during its training

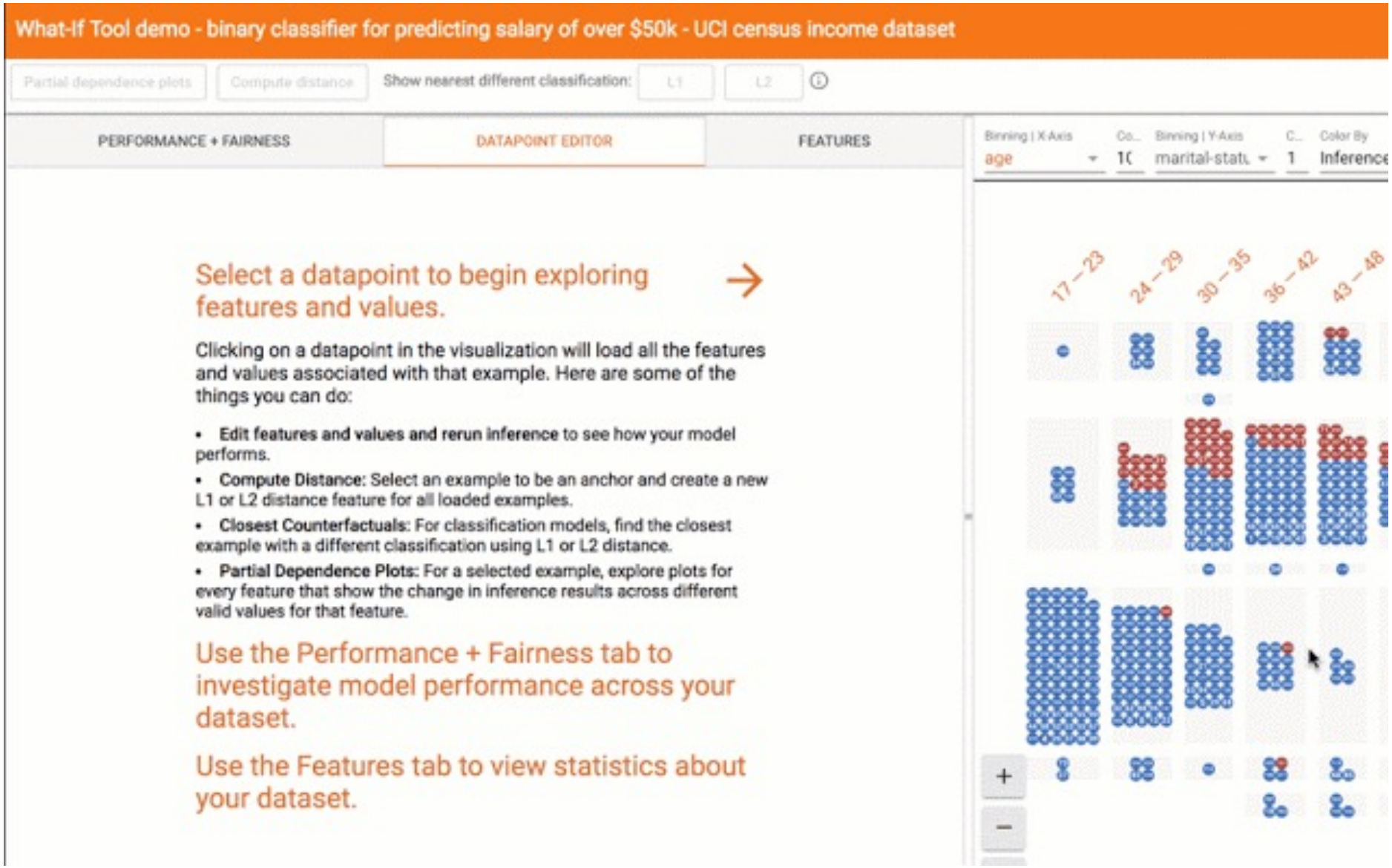
Post-processing algorithm:
a bias mitigation algorithm that is applied to predicted labels

What If Tool

Goal: Code-free probing of machine learning models

- Feature perturbations (what if scenarios)
- Counterfactual example analysis
- [Classification] Explore the effects of different classification thresholds, taking into account constraints such as different numerical fairness metrics.

What If Tool


What-If Tool demo - binary classifier for predicting salary of over $50k - UCI census income dataset
Partial dependence plots
Compute distance
Show nearest different classification:
L1
L2
PERFORMANCE + FAIRNESS
DATAPOINT EDITOR
FEATURES
Select a datapoint to begin exploring features and values.
Clicking on a datapoint in the visualization will load all the features and values associated with that example. Here are some of the things you can do:
• Edit features and values and rerun inference to see how your model performs.
• Compute Distance: Select an example to be an anchor and create a new L1 or L2 distance feature for all loaded examples.
• Closest Counterfactuals: For classification models, find the closest example with a different classification using L1 or L2 distance.
• Partial Dependence Plots: For a selected example, explore plots for every feature that show the change in inference results across different valid values for that feature.
Use the Performance + Fairness tab to investigate model performance across your dataset.
Use the Features tab to view statistics about your dataset.
Binning | X Axis
age
Binning | Y Axis
marital-stat
Color By
Inference
17 – 23
24 – 29
30 – 35
36 – 42
43 – 48

What-If Tool demo - two binary classifiers for predicting salary of over $50k - UCI census income dataset

Datapoint editor | Performance & Fairness | Features

500 datapoints loaded

Visualize

Datapoints | Partial dependence plots

Show nearest counterfactual datapoint | L1 | L2 | Model: 1

Show similarity to selected datapoint

Edit

Search features

Select a datapoint to begin exploring model behavior for your selection.

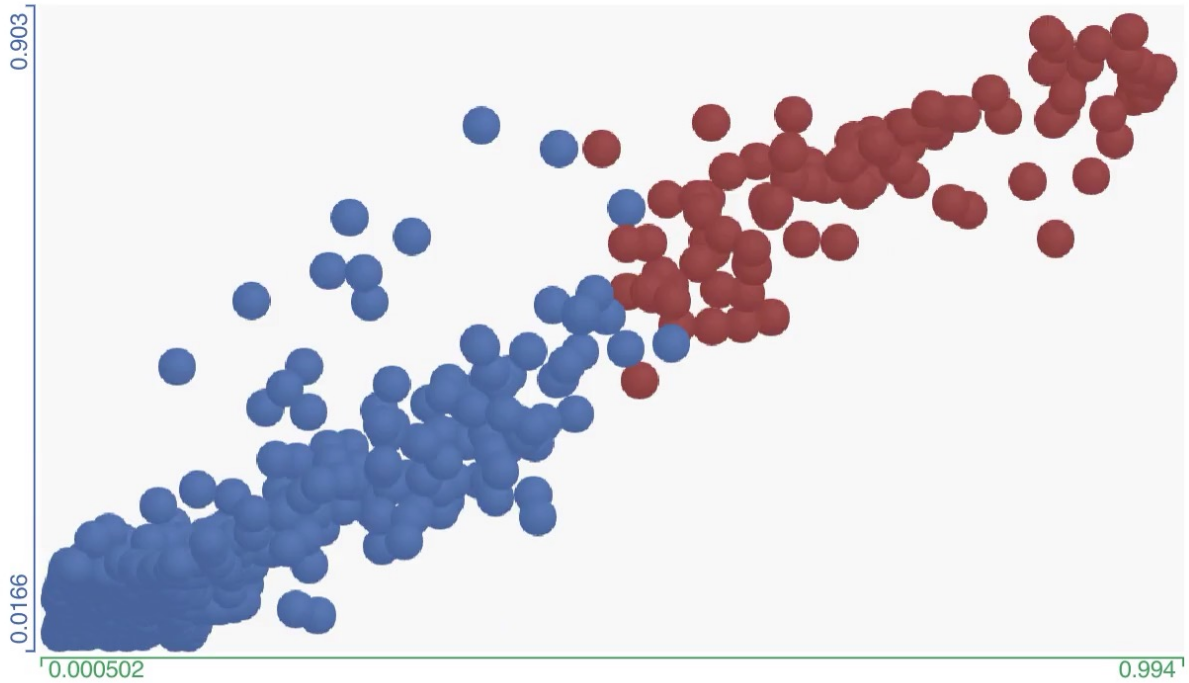
Edit and Infer: Edit your datapoint here and run inference in the Infer table to see differences in model behavior.

Visualize: Switch between visualizing datapoints and exploring partial dependence plots to gain insights into your model's behavior. Explore counterfactuals or see how similar (or different) the rest of your dataset is from your selection.

Infer

Run inference

Binning | X-Axis: (none)
Binning | Y-Axis: (none)
Color By: Inference lab…
Label By: (default)
Scatter | X-Axis: Inference score
Scatter | Y-Axis: Inference score 2



Legend

Colors
by Inference label 1
<=50k
>50k

What-If Tool demo - two binary classifiers for predicting salary of over $50k - UCI census income dataset

Datapoint editor | Performance & Fairness | Features

500 datapoints loaded

Configure

Ground Truth Feature: over_50k

WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. *More.*

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. *More.*

Slice by: <none>

WHAT DOES SLICING DO?
Shows performance for each value of the selected feature.

Fairness

Apply an optimization strategy

Select a strategy to set classification thresholds based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will default back to custom thresholds.

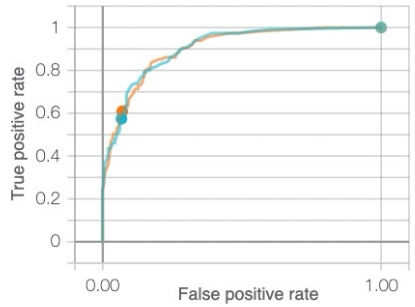
- (•) Custom thresholds
- ( ) Single threshold
- ( ) Demographic parity
- ( ) Equal opportunity
- ( ) Equal accuracy
- ( ) Group thresholds

Explore overall performance

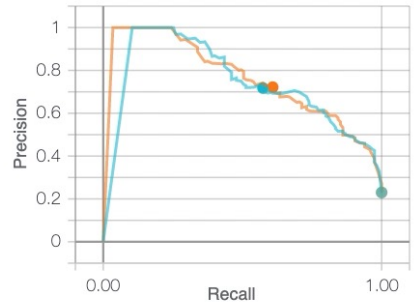
Sort by: Count

Feature Value	Count	Model	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	500	1	0.5	5.2	9.8	85.0	0.64
		2	0.5	5.4	9.0	85.6	0.66

ROC curve



PR curve



Confusion matrix

1	Predicted Yes		Predicted No		Total	
Actual Yes	13.2%	(66)	9.8%	(49)	23.0%	(115)
Actual No	5.2%	(26)	71.8%	(359)	77.0%	(385)
Total	18.4%	(92)	81.6%	(408)		

2	Predicted Yes		Predicted No		Total	
Actual Yes	14.0%	(70)	9.0%	(45)	23.0%	(115)
Actual No	5.4%	(27)	71.6%	(358)	77.0%	(385)
Total	19.4%	(97)	80.6%	(403)		

Datasheets for Datasets [Gebru et al., 2018]

- Better data-related documentation
  - Datasheets for datasets: every dataset, model, or pre-trained API should be accompanied by a data sheet that documents its
    - Creation
    - Intended uses
    - Limitations
    - Maintenance
    - Legal and ethical considerations
    - Etc.

Model Cards for Model Reporting [Mitchell et al., 2018]

Intended use

Human-assisted moderation

Make moderation easier with an ML assisted tool that helps prioritize comments for human moderation, and create custom tasks for automated actions. See our moderator tool as an example.

Author feedback

Assist authors in real-time when their comments might violate your community guidelines or be may be perceived as "Toxic" to the conversation. Use simple feedback tools when the assistant gets it wrong. See our authorship demo as an example.

Read better comments

Organize comments on topics that are often difficult to discuss online. Build new tools that help people explore the conversation.

Uses to avoid

Fully automated moderation

Perspective is not intended to be used for fully automated moderation. Machine learning models will always make some mistakes, so it is essential to build in systems for humans to catch and correct those mistakes.

Character judgement

In order to maintain user privacy, the TOXICITY model only helps detect toxicity in an individual statement, and is not intended to detect anything about the individual who said it. In addition, Perspective does not use prior information about an individual to inform toxicity predictions.

Model details

Training data

Proprietary from Perspective API, which includes comments from online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic", defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion".

Model architecture

The model is a Convolutional Neural Network (CNN) trained with GloVe word embeddings, which are fine-tuned during training. You can also train your own deep CNN for text classification on our public toxicity dataset, and explore our open-source model training tools to train your own models.

Values

Community, Transparency, Inclusivity, Privacy, and Topic neutrality. These values guide our product and research decisions.

Fact Sheets [Arnold et al., 2019]

- Is distinguished from “model cards” and “datasheets” in that the focus is on the final AI service:
  - What is the intended use of the service output?
  - What algorithms or techniques does this service implement?
  - Which datasets was the service tested on? (Provide links to datasets that were used for testing, along with corresponding datasheets.)
  - Describe the testing methodology.
  - Describe the test results.
  - Etc.

Responsible AI Case Studies at LinkedIn

Krishnaram Kenthapadi

Amazon AWS AI

Fairness in AI @ LinkedIn

Fairness-aware Talent Search Ranking*

* Work done while at LinkedIn


Guiding Principle:
“Diversity by Design”

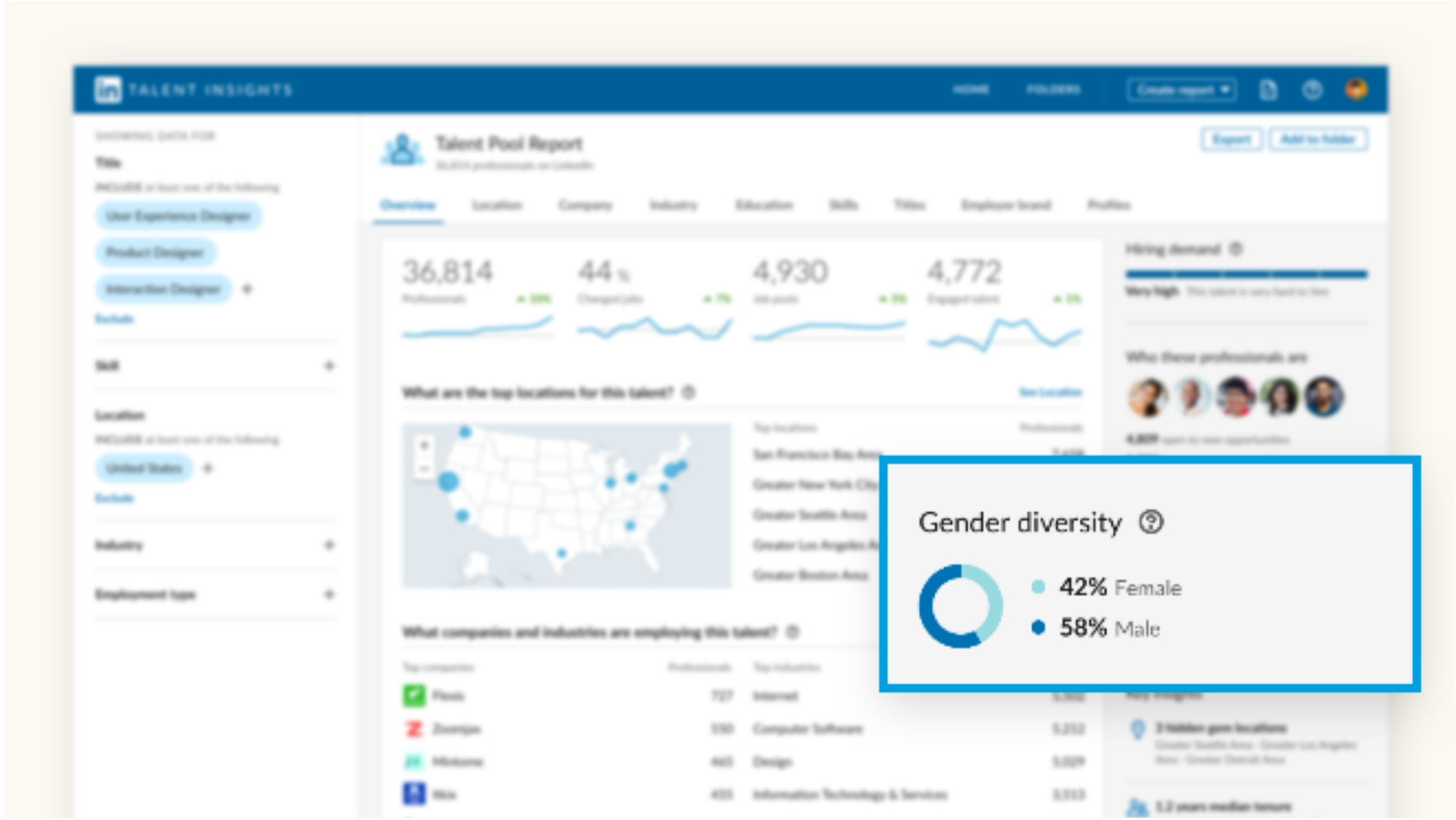
“Diversity by Design” in LinkedIn’s Talent Solutions

Insights to Identify Diverse Talent Pools

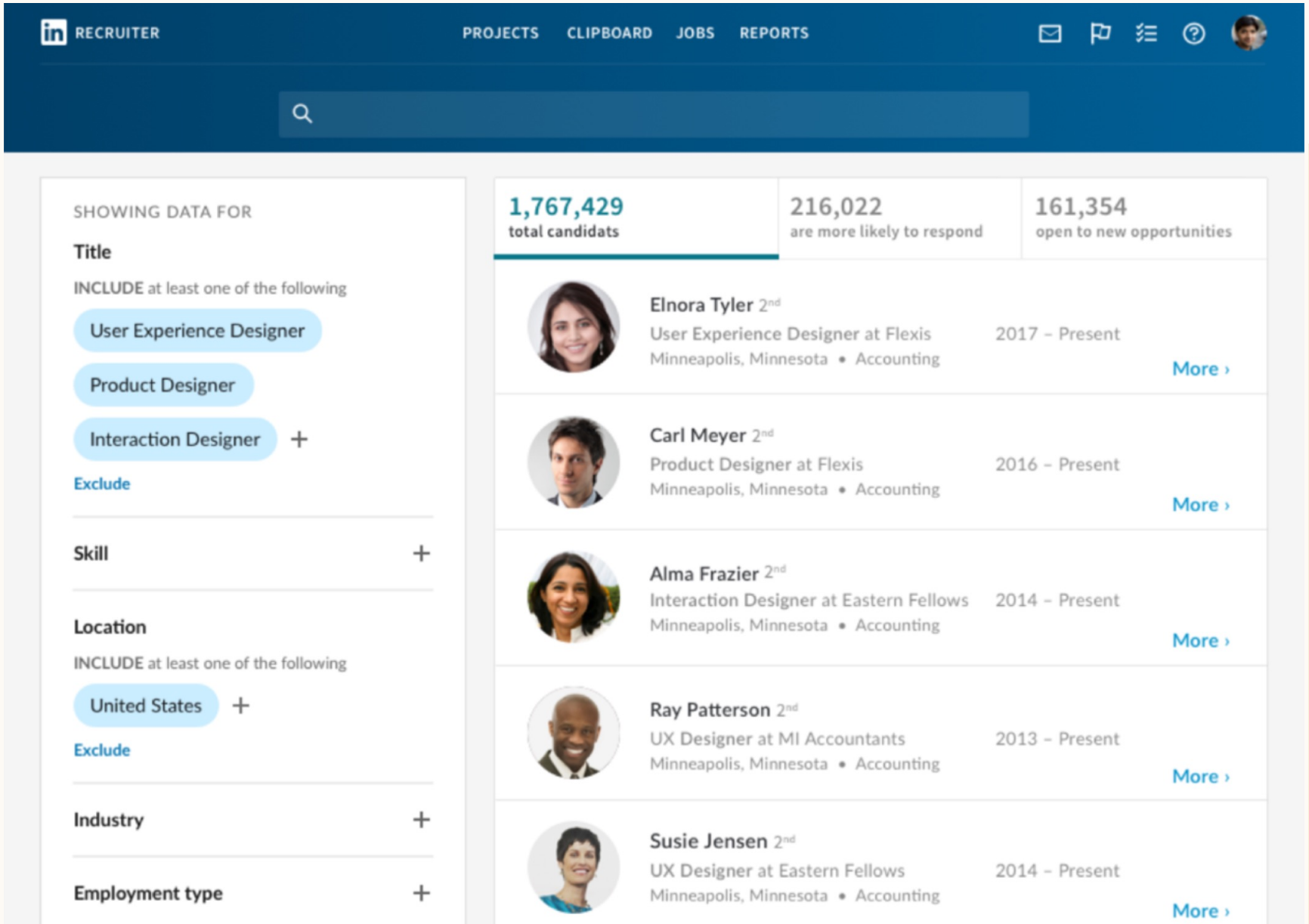
Representative Talent Search Results

Diversity Learning Curriculum

Plan for Diversity



Representative Ranking for Talent Search



S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search, KDD'19.

[Microsoft's AI/ML conference (MLADS'18). **Distinguished Contribution Award**]

Building Representative Talent Search at LinkedIn (LinkedIn engineering blog)

Intuition for Measuring and Achieving Representativeness

Ideal: Top ranked results should follow a desired distribution on gender/age/…

E.g., same distribution as the underlying talent pool

Inspired by “Equal Opportunity” definition [Hardt et al, NeurIPS’16]

Defined measures (skew, divergence) based on this intuition

Measuring (Lack of) Representativeness

Skew@k

(Logarithmic) ratio of the proportion of candidates having a given attribute value among the top k ranked results to the corresponding desired proportion

$$Skew_v@k(\tau_r) = \log_e\left(\frac{p_{\tau_r^k,r,v}}{p_{q,r,v}}\right)$$

Proportion of candidates from attribute value ***v*** in **top-k** results

Desired proportion of candidates from attribute value ***v***

Variants:

MinSkew: Minimum over all attribute values
MaxSkew: Maximum over all attribute values
Normalized Discounted Cumulative Skew
Normalized Discounted Cumulative KL-divergence

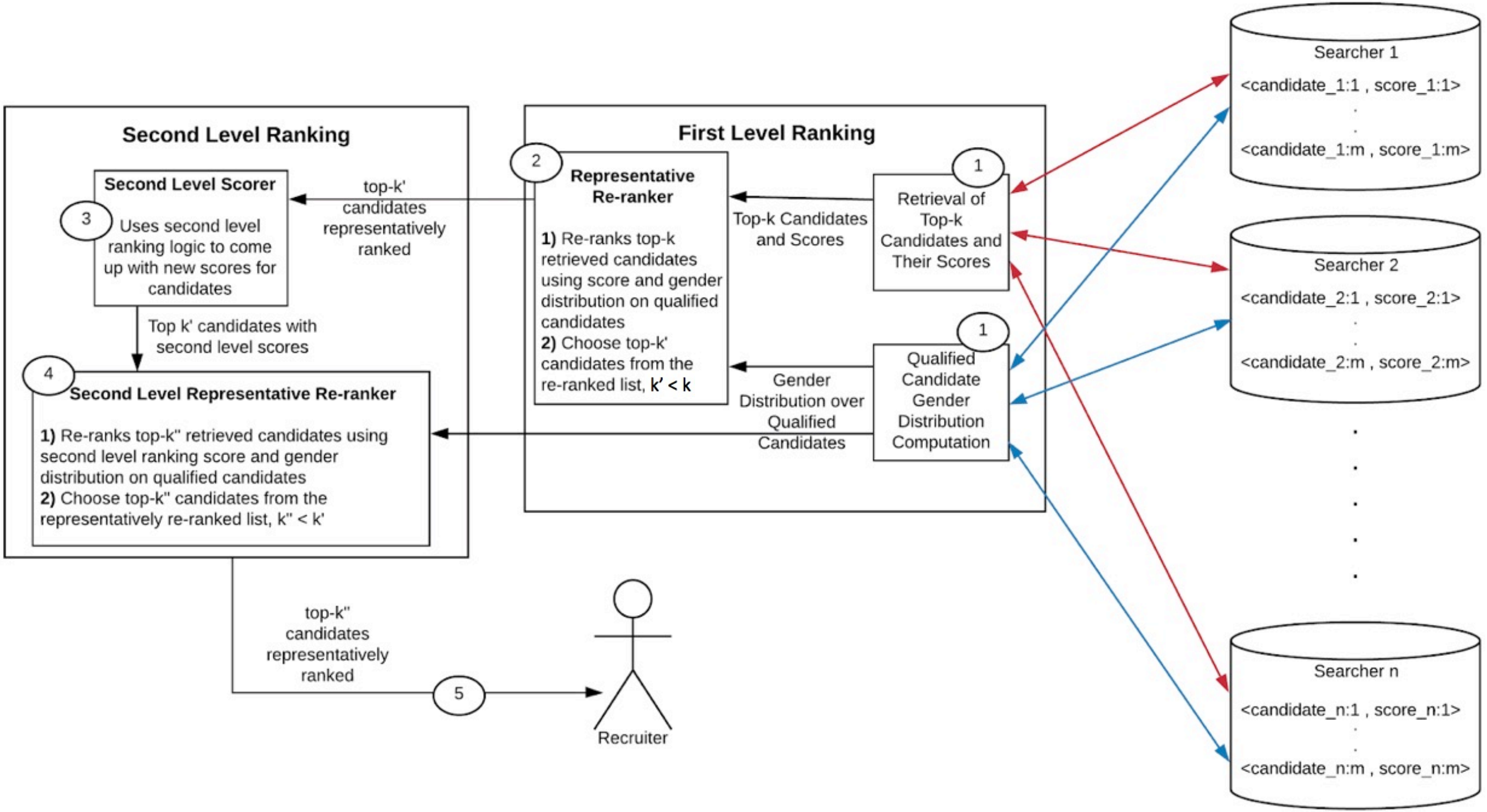
Fairness-aware Reranking Algorithm (Simplified)

Partition the set of potential candidates into different buckets for each attribute value

Rank the candidates in each bucket according to the scores assigned by the machine-learned model

Merge the ranked lists, balancing the representation requirements and the selection of highest scored candidates

- Representation requirement: Desired distribution on gender/age/…
- Algorithmic variants based on how we choose the next attribute


Second Level Ranking
Second Level Scorer
3
Uses second level ranking logic to come up with new scores for candidates
top-k' candidates representatively ranked
Top k' candidates with second level scores
4
Second Level Representative Re-ranker
1) Re-ranks top-k'' retrieved candidates using second level ranking score and gender distribution on qualified candidates
2) Choose top-k'' candidates from the representatively re-ranked list, k'' < k'
top-k'' candidates representatively ranked
5
Recruiter
First Level Ranking
2
Representative Re-ranker
1) Re-ranks top-k retrieved candidates using score and gender distribution on qualified candidates
2) Choose top-k' candidates from the re-ranked list, k' < k
Top-k Candidates and Scores
1
Retrieval of Top-k Candidates and Their Scores
Gender Distribution over Qualified Candidates
1
Qualified Candidate Gender Distribution Computation
Searcher 1
<candidate_1:1 , score_1:1>
<candidate_1:m , score_1:m>
Searcher 2
<candidate_2:1 , score_2:1>
<candidate_2:m , score_2:m>
Searcher n
<candidate_n:1 , score_n:1>
<candidate_n:m , score_n:m>

Validating Our Approach

Gender Representativeness

- Over 95% of all searches are representative compared to the qualified population of the search

Business Metrics

- A/B test over LinkedIn Recruiter users for two weeks
- No significant change in business metrics (e.g., # InMails sent or accepted)

Ramped to 100% of LinkedIn Recruiter users worldwide

Lessons learned

- Post-processing approach desirable
  - Model agnostic
    - Scalable across different model choices for our application
  - Acts as a “fail-safe”
    - Robust to application-specific business logic
  - Easier to incorporate as part of existing systems
    - Build a stand-alone service or component for post-processing
    - No significant modifications to the existing components
  - Complementary to efforts to reduce bias from training data & during model training
- Collaboration/consensus across key stakeholders

Les
lear

Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search

Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi
LinkedIn Corporation, USA

ABSTRACT

We present a framework for quantifying and mitigating algorithmic bias in mechanisms designed for ranking individuals, typically used as part of web-scale search and recommendation systems. We first propose complementary measures to quantify bias with respect to protected attributes such as gender and age. We then present algorithms for computing fairness-aware re-ranking of results. For a given search or recommendation task, our algorithms seek to achieve a desired distribution of top ranked results with respect to one or more protected attributes. We show that such a framework can be tailored to achieve fairness criteria such as *equality of opportunity* and *demographic parity* depending on the choice of the desired distribution. We evaluate the proposed algorithms via extensive simulations over different parameter choices, and study the effect of fairness-aware ranking on both bias and utility measures. We finally present the online A/B testing results from applying our framework towards representative ranking in LinkedIn Talent Search, and discuss the lessons learned in practice. Our approach resulted in tremendous improvement in the fairness metrics (nearly three fold increase in the number of search queries with representative results) without affecting the business metrics, which paved the way for deployment to 100% of *LinkedIn Recruiter* users worldwide. Additionally, we propose algorithms for re-ranking candidates scored/returned by a machine learned model to satisfy the fairness constraints. Our key contributions include:

- Proposal of fairness-aware ranking algorithms towards mitigating algorithmic bias. Our methodology can be used to achieve fairness criteria such as *equality of opportunity* [26] and *demographic parity* [17] depending on the choice of the desired distribution over protected attribute(s).
- Proposal of complementary measures for quantifying the fairness of the ranked candidate lists.
- Extensive evaluation of the proposed algorithms via simulations over a wide range of ranking scenarios and attributes with different cardinalities (possible number of values).
- Online A/B test results of applying our framework for achieving representative ranking in LinkedIn Talent Search, and the lessons learned in practice. Our approach resulted in tremendous improvement in the fairness metrics (nearly three fold increase in the number of search queries with representative results) without statistically significant change in the business metrics, which paved the way for deployment to 100% of *LinkedIn Recruiter* users worldwide.

del choices

c business

existing

or
ing
to the

uce bias
el training

- Collaboration/consensus across key stakeholders

Evaluating Fairness Using Permutations Tests

[DiCiccio, Vasudevan, Basu, Kenthapadi, Agarwal, KDD'20]

- Is the measured discrepancy across different groups statistically significant?
  - Use statistical hypothesis tests!
- Can we perform hypothesis tests in a metric-agnostic manner?
  - Non-parametric tests can help!
- Permutation testing framework

Brief Review of Permutation Tests

Observe data from two populations:

$$X_1, \ldots, X_{n_x} \sim P_X \qquad \text{and} \qquad Y_1, \ldots, Y_{n_y} \sim P_Y$$

Are the populations the same?

$$H : P_X = P_Y$$

A reasonable test statistic might be

$$T = \bar{X} - \bar{Y}$$

Brief Review of Permutation Tests (Continued)

A p-value is the chance of observing a test statistic at least as "extreme" as the value we actually observed

Permutation test approach:

- Randomly shuffle the population designations of the observations
- Recompute the test statistic T
- Repeat many times

Permutation p-value: the proportion of permuted datasets resulting in a larger test statistic than the original value

This test is exact!

A Fairness Example

Consider testing whether the true positive rate of a classifier is equal between two groups

Test Statistic: difference in proportion of negative labeled observations that are classified as positive between the two groups

Permutation test: Randomly reshuffle group labels, recompute test statistic

Permutations Tests for Evaluating Fairness in ML Models

- Issues with classical permutation test
  - Want to check: just equality of the fairness metric (e.g., false positive rate) across groups, and not if the two groups have identical distribution
  - Exact for the strong null hypothesis … $H_0 : P_X = P_Y$
  - … but may not be valid (even asymptotically) for the weak null hypothesis $H_0 : \theta(P_X) = \theta(P_Y)$

- Our paper: A fix for this issue
  - Choose a pivotal statistic (asymptotically distribution-free; does not depend on the observed data's distribution)
  - E.g., Studentize the test statistic

Permutations Tests for Evaluating Fairness in ML Models

- Issue
  - W
    ac
  - Ex
  - …
- Our p
  - Ch
    or
  - E.g

rate)

esis

epend

Evaluating Fairness Using Permutation Tests

Cyrus DiCiccio
LinkedIn Corporation
cdiciccio@linkedin.com

Sriram Vasudevan
LinkedIn Corporation
svasudevan@linkedin.com

Kinjal Basu
LinkedIn Corporation
kbasu@linkedin.com

Krishnaram Kenthapadi[1]
Amazon AWS AI
kenthk@amazon.com

Deepak Agarwal
LinkedIn Corporation
dagarwal@linkedin.com

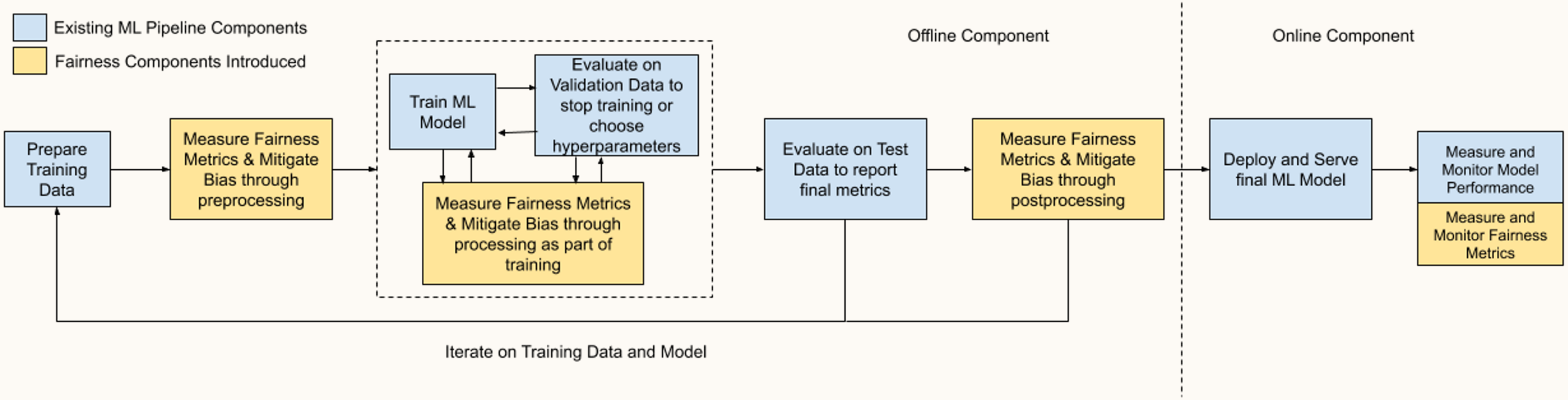
ABSTRACT

Machine learning models are central to people's lives and impact society in ways as fundamental as determining how people access information. The gravity of these models imparts a responsibility to model developers to ensure that they are treating users in a fair and equitable manner. Before deploying a model into production, it is crucial to examine the extent to which its predictions demonstrate biases. This paper deals with the detection of bias exhibited by a machine learning model through statistical hypothesis testing. We propose a permutation testing methodology that performs a hypothesis test that a model is fair across two groups with respect to any given metric. There are increasingly many notions of fairness that can speak to different aspects of model fairness. Our aim is to provide a flexible framework that empowers practitioners to identify significant biases in any metric they wish to study. We provide a formal testing mechanism as well as extensive experiments to show how this method works in practice.

1 INTRODUCTION

Machine learned models are increasingly being used in web applications for crucial decision-making tasks such as lending, hiring, and college admissions, driven by a confluence of factors such as ubiquitous connectivity, the ability to collect, aggregate, and process large amounts of fine-grained data, and the ease with which sophisticated machine learning models can be applied. Recently, there has been a growing awareness about the ethical and legal challenges posed by the use of such data-driven systems, which often make use of classification models that deal with users. Researchers and practitioners from different disciplines have highlighted the potential for such systems to discriminate against certain population groups, due to biases in data and algorithmic decision-making systems. Several studies have shown that classification and ranked results produced by a biased machine learning model can result in systemic discrimination and reduced visibility for an already disadvantaged group [5, 12, 16, 22] (e.g., disproportionate association of higher risk

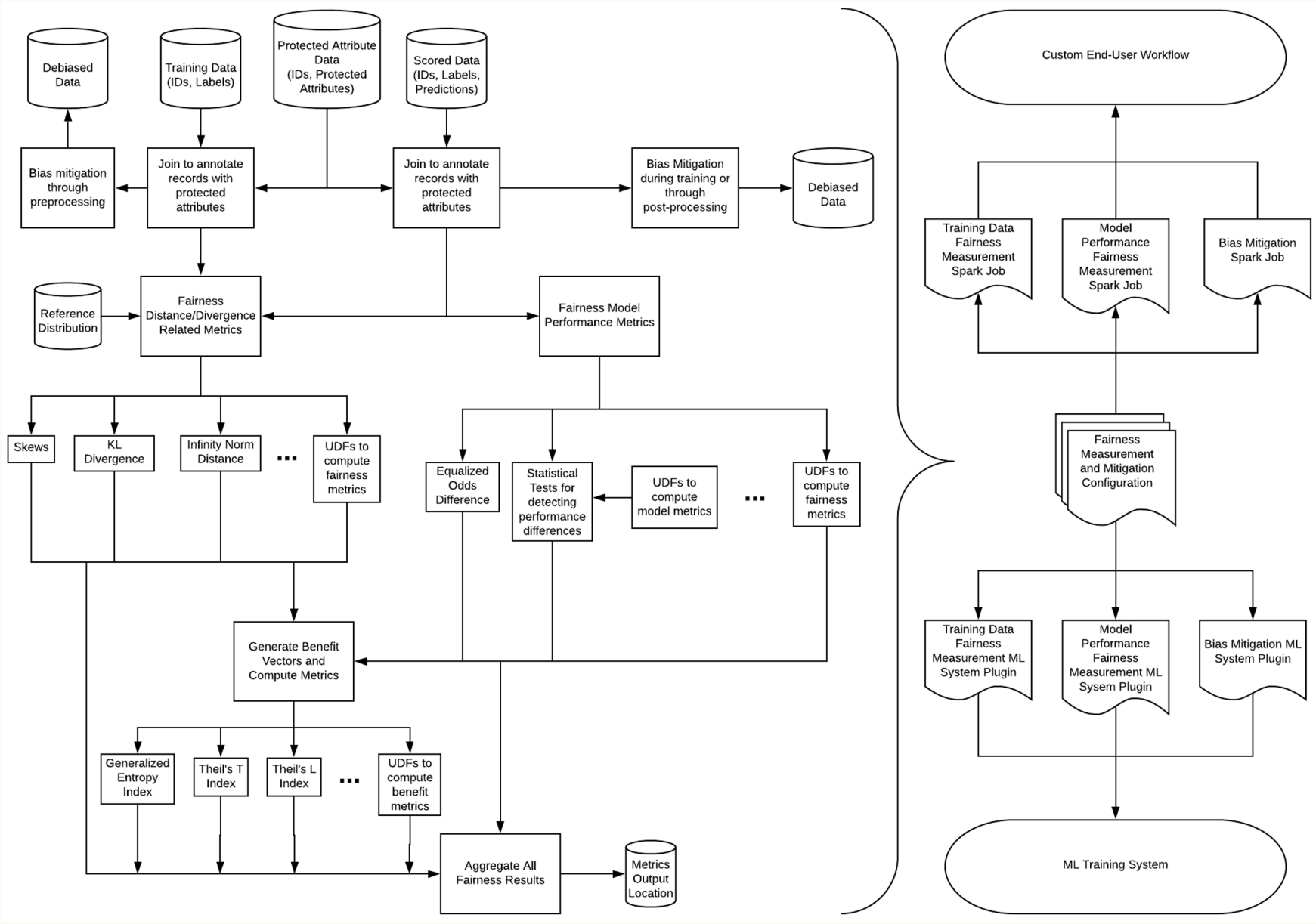
Engineering for Fairness in AI Lifecycle



S.Vasudevan, K. Kenthapadi, LiFT: A Scalable Framework for Measuring Fairness in ML Applications, CIKM'20

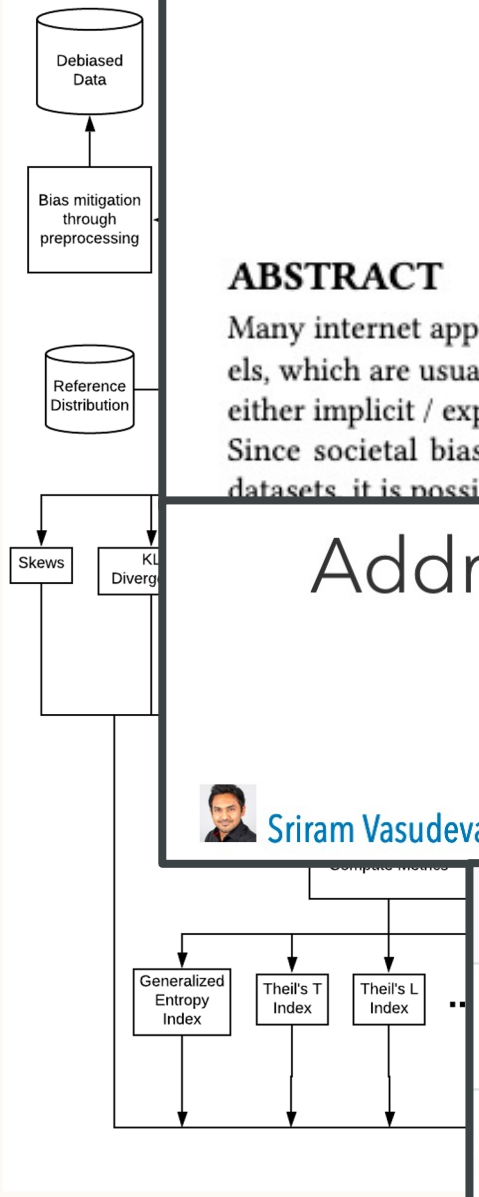
https://github.com/linkedin/LiFT

LiFT System Architecture [Vasudevan & Kenthapadi, CIKM'20]



- Flexibility of Use (Platform agnostic)
  - Ad-hoc exploratory analyses
  - Deployment in offline workflows
  - Integration with ML Frameworks
- Scalability
- Diverse fairness metrics
  - Conventional fairness metrics
  - Benefit metrics
  - Statistical tests

LiFT



LiFT: A Scalable Framework for Measuring Fairness in ML Applications

Sriram Vasudevan
LinkedIn Corporation
svasudevan@linkedin.com

Krishnaram Kenthapadi[1]
Amazon AWS AI
kenthk@amazon.com

ABSTRACT

Many internet applications are powered by machine learned models, which are usually trained on labeled datasets obtained through either implicit / explicit user feedback signals or human judgments. Since societal biases may be present in the generation of such datasets, it is possible for the trained models to be biased, thereby

of our system (§3 and §4). We present results from deployments at LinkedIn and discuss the challenges encountered and lessons learned during deployment of fairness tools in practice (§5 and §6). Finally, we provide open problems and research directions based on our experiences (§8).

The key contributions of our work are as follows:

Addressing bias in large-scale AI applications: The LinkedIn Fairness Toolkit

Sriram Vasudevan August 25, 2020

Share Tweet Share

https://github.com/linkedin/LiFT

README.md

The LinkedIn Fairness Toolkit (LiFT)

Use
nostic)
oratory

in offline

vith ML

ess

l fairness
metrics
•Benefit metrics
•Statistical tests

Acknowledgements

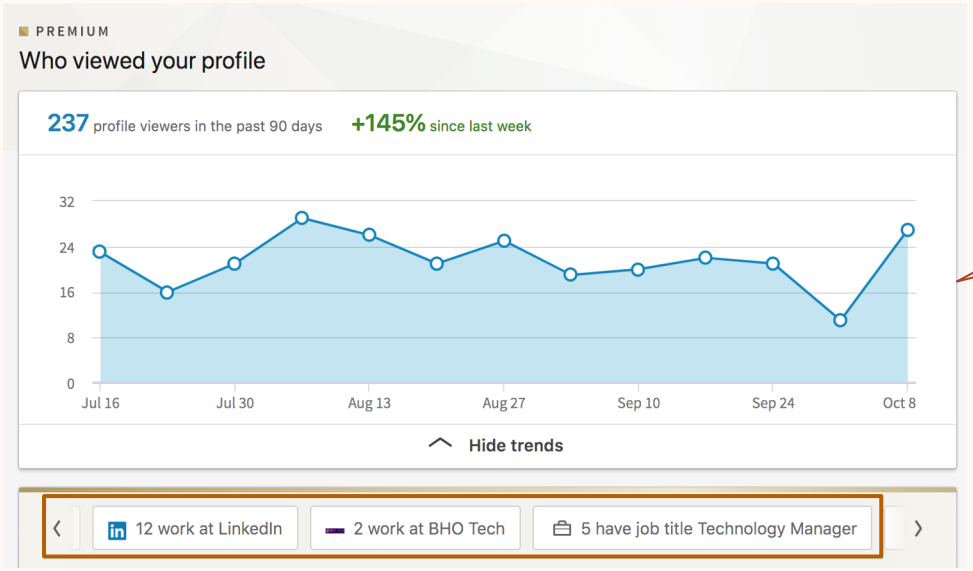
LinkedIn Talent Solutions Diversity team, Hire & Careers AI team, Anti-abuse AI team, Data Science Applied Research team

Special thanks to Deepak Agarwal, Parvez Ahammad, Stuart Ambler, Kinjal Basu, Jenelle Bray, Erik Buchanan, Bee-Chung Chen, Fei Chen, Patrick Cheung, Gil Cottle, Cyrus DiCiccio, Patrick Driscoll, Carlos Faham, Nadia Fawaz, Priyanka Gariba, Meg Garlinghouse, Sahin Cem Geyik, Gurwinder Gulati, Rob Hallman, Sara Harrington, Joshua Hartman, Daniel Hewlett, Nicolas Kim, Rachel Kumar, Monica Lewis, Nicole Li, Heloise Logan, Stephen Lynch, Divyakumar Menghani, Varun Mithal, Arashpreet Singh Mor, Tanvi Motwani, Preetam Nandy, Lei Ni, Nitin Panjwani, Igor Perisic, Hema Raghavan, Romer Rosales, Guillaume Saint-Jacques, Badrul Sarwar, Amir Sepehri, Arun Swami, Ram Swaminathan, Grace Tang, Ketan Thakkar, Sriram Vasudevan, Janardhanan Vembunarayanan, James Verbus, Xin Wang, Hinkmond Wong, Ya Xu, Lin Yang, Yang Yang, Chenhui Zhai, Liang Zhang, Yani Zhang

Privacy in AI @ LinkedIn

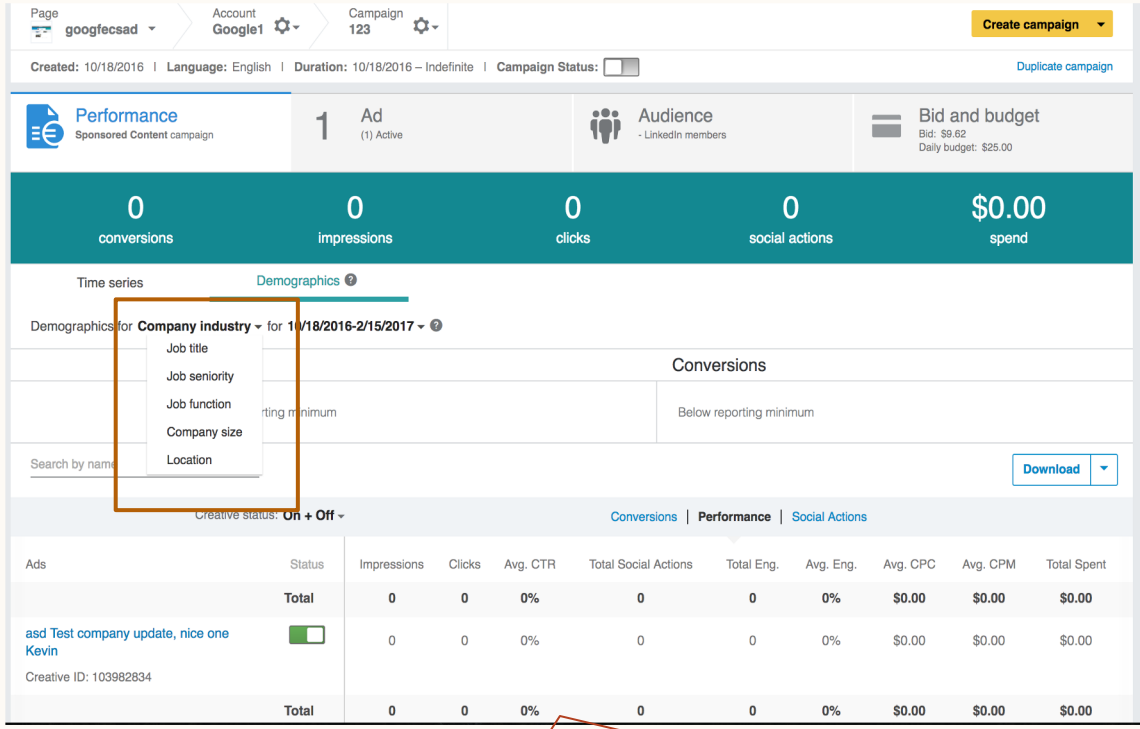
PriPeARL: Framework to compute robust, **privacy-preserving analytics**

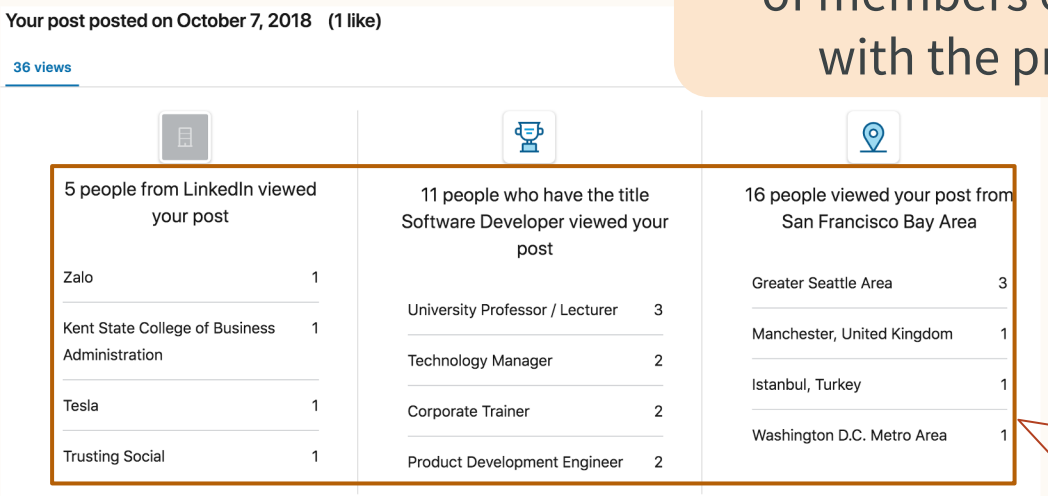
Analytics & Reporting Products at LinkedIn



Profile View Analytics

All showing demographics of members engaging with the product





Content Analytics

Ad Campaign Analytics

Analytics & Reporting Products at LinkedIn

- Admit only a small # of predetermined query types
  - Querying for the number of member actions, for a specified time period, together with the top demographic breakdowns

"SELECT COUNT(*) FROM table(statType, entity) WHERE timeStamp $\geq$ startTime AND timeStamp $\leq$ endTime AND $d_{attr} = d_{val}$"

Analytics & Reporting Products at LinkedIn

Admit only a small # of predetermined query types

Querying for the number of member actions, for a specified time period, together with the top demographic breakdowns

E.g., Title = "Senior Director"

E.g., *Clicks* on a *given ad*

"SELE COUNT(*) FROM table(statType, entity) WHERE timeSt mp ≥ startTime AND timeStamp ≤ endTime AND $d_{attr} = d_{val}$"

Privacy Requirements

Attacker cannot infer whether a member performed an action

- E.g., click on an article or an ad

Attacker may use auxiliary knowledge

- E.g., knowledge of attributes associated with the target member (say, obtained from this member's LinkedIn profile)
- E.g., knowledge of all other members that performed similar action (say, by creating fake accounts)

Possible Privacy Attacks

Targeting: Senior directors in US, who studied at Cornell	Matches ~16k LinkedIn members → over minimum targeting threshold ✓
Demographic breakdown: Company = X	May match exactly one person → can determine whether the person clicks on the ad or not ✗
Require minimum reporting threshold	Attacker could create fake profiles! E.g. if threshold is 10, create 9 fake profiles that all click. ✗
Rounding mechanism E.g., report incremental of 10	Still amenable to attacks E.g. using incremental counts over time to infer individuals' actions ✗

Need rigorous techniques to preserve member privacy
(not reveal exact aggregate counts)

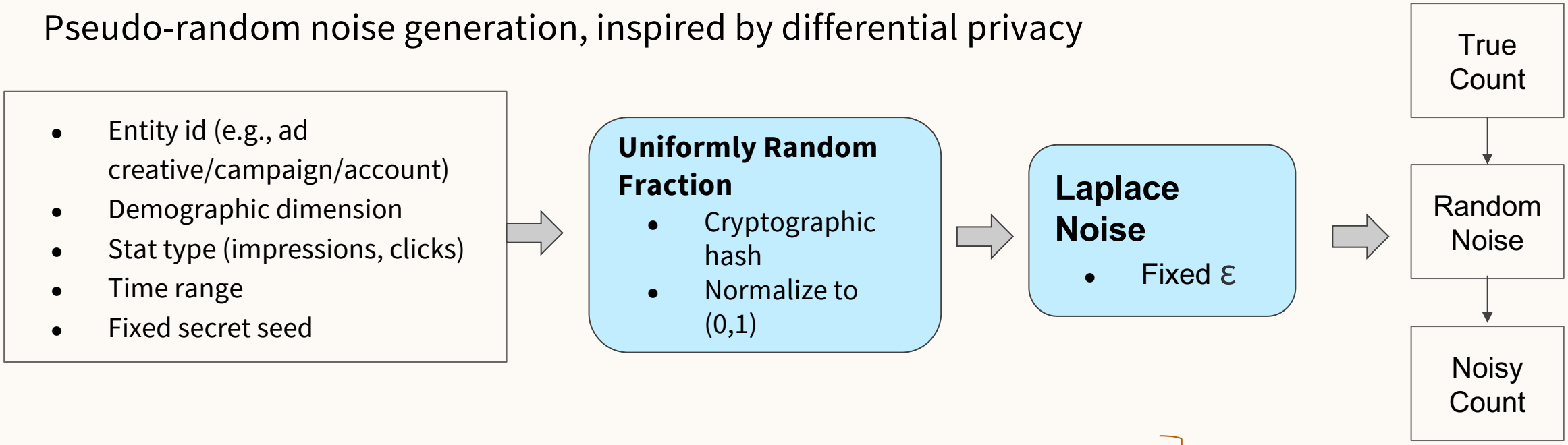
Problem Statement

Compute robust, reliable analytics in a privacy-preserving manner, while addressing the product needs.

PriPeARL: A Framework for Privacy-Preserving Analytics

K. Kenthapadi, T. T. L. Tran, ACM CIKM 2018

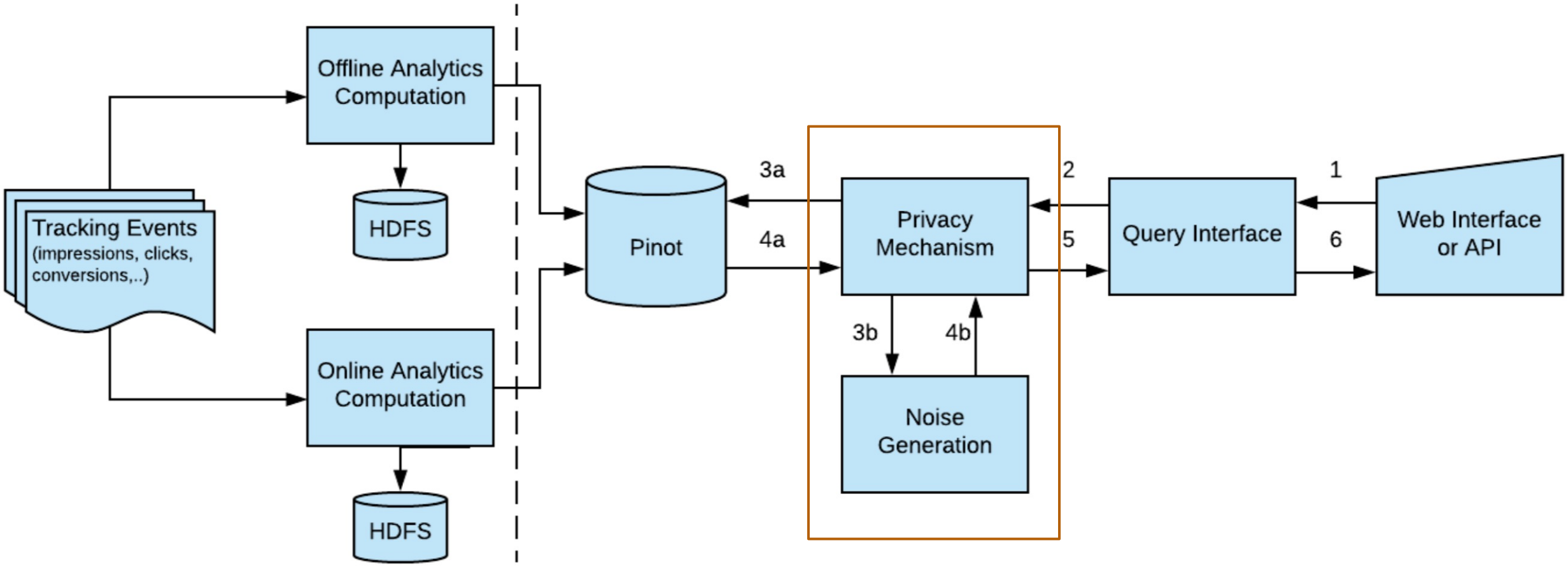
Pseudo-random noise generation, inspired by differential privacy



- Pseudo-random noise → same query has same result over time, avoid averaging attack.
- For non-canonical queries (e.g., time ranges, aggregate multiple entities)
  - Use the hierarchy and partition into canonical queries
  - Compute noise for each canonical queries and sum up the noisy counts

To satisfy consistency requirements

PriPeARL System Architecture



Lessons Learned from Deployment (> 1 year)

Semantic consistency vs. unbiased, unrounded noise

Suppression of small counts

Online computation and performance requirements

Scaling across analytics applications

- Tools for ease of adoption (code/API library, hands-on how-to tutorial) help!
- Having a few entry points (all analytics apps built over Pinot) ➔ wider adoption

Summary

Framework to compute robust, privacy-preserving analytics

- Addressing challenges such as preserving member privacy, product coverage, utility, and data consistency

Future

- Utility maximization problem given constraints on the 'privacy loss budget' per user
  - E.g., noise with larger variance to impressions but less noise to clicks (or conversions)
  - E.g., more noise to broader time range sub-queries and less noise to granular time range sub-queries

Reference: K. Kenthapadi, T. Tran, PriPeARL: A Framework for Privacy-Preserving Analytics and Reporting at LinkedIn, ACM CIKM 2018.

Acknowledgements

Team:

AI/ML: Krishnaram Kenthapadi, Thanh T. L. Tran

Ad Analytics Product & Engineering: Mark Dietz, Taylor Greason, Ian Koeppe

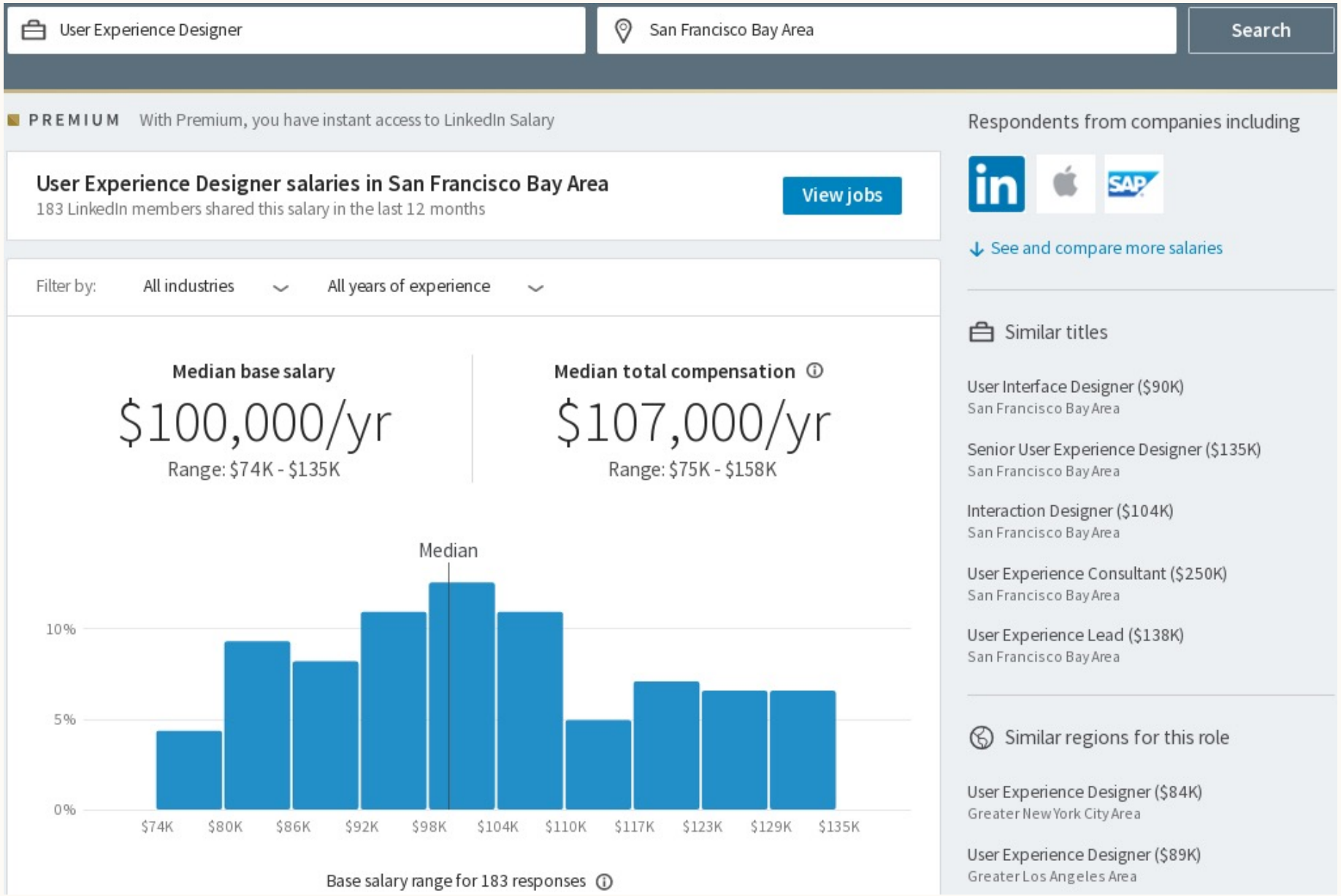
Legal / Security: Sara Harrington, Sharon Lee, Rohit Pitke

Acknowledgements

Deepak Agarwal, Igor Perisic, Arun Swami

LinkedIn Salary

LinkedIn Salary (launched in Nov, 2016)



Data Privacy Challenges

- Minimize the risk of inferring any one individual's compensation data
- Protection against data breach
  - No single point of failure

Problem Statement

How do we design LinkedIn Salary system taking into account the unique privacy and security challenges, while addressing the product requirements?

Achieved by a combination of techniques: encryption, access control, de-identification, aggregation, thresholding

K. Kenthapadi, A. Chudhary, and S. Ambler, LinkedIn Salary: A System for Secure Collection and Presentation of Structured Compensation Insights to Job Seekers, IEEE PAC 2017 (arxiv.org/abs/1705.06976)

De-identification Example

Title	Region	Company	Industry	Years of exp	Degree	FoS	Skills	$$
User Exp Designer	SF Bay Area	Google	Internet	12	BS	Interactive Media	UX, Graphics, ...	100K

Title	Region	$$
User Exp Designer	SF Bay Area	100K
User Exp Designer	SF Bay Area	115K
...	...	...

Title	Region	Industry	$$
User Exp Designer	SF Bay Area	Internet	100K

Title	Region	Years of exp	$$
User Exp Designer	SF Bay Area	10+	100K

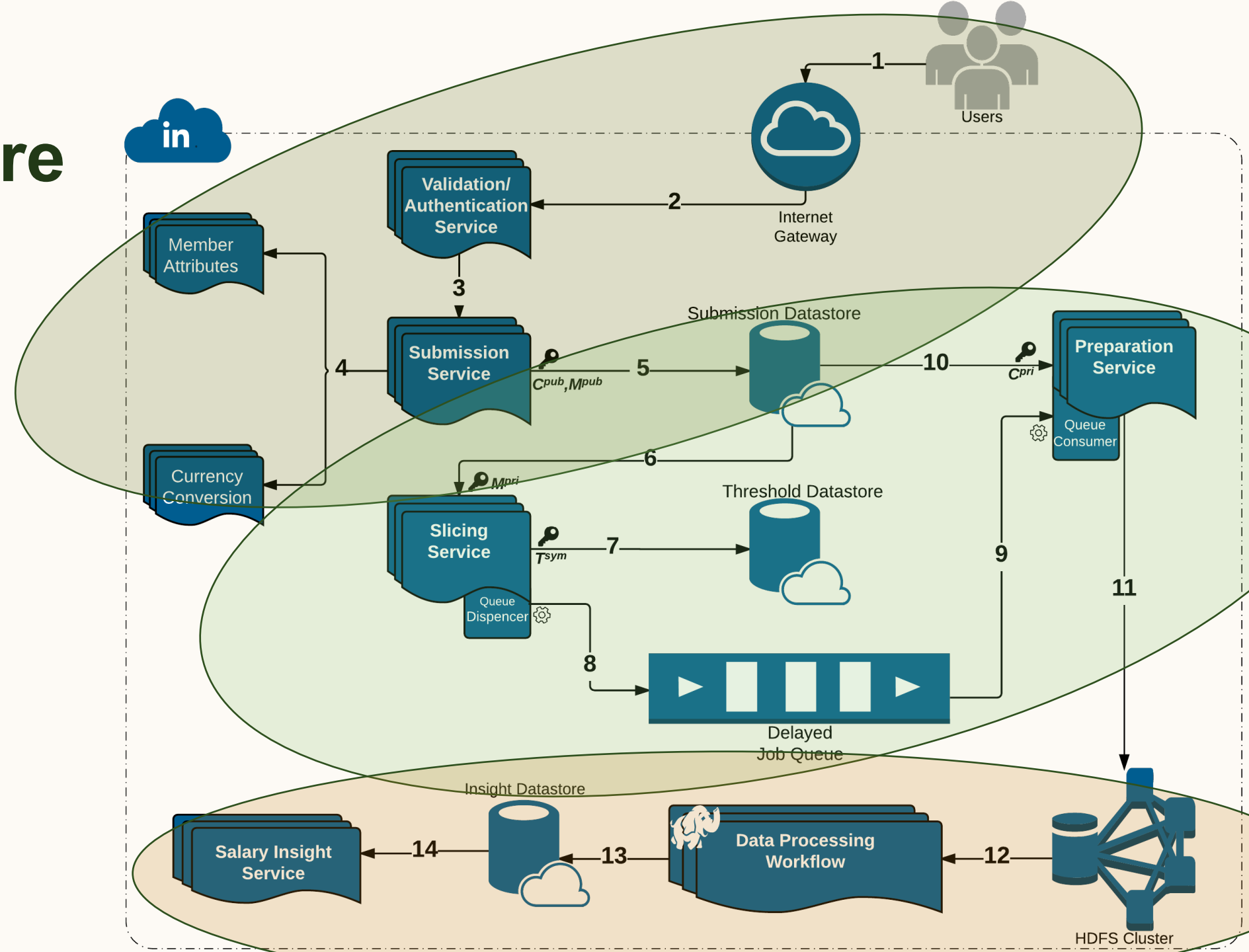
Title	Region	Company	Years of exp	$$
User Exp Designer	SF Bay Area	Google	10+	100K

#data points > threshold?

Yes ⇒ Copy to Hadoop (HDFS)

Note: Original submission stored as encrypted objects.

System Architecture



Acknowledgements

Team:

AI/ML: Krishnaram Kenthapadi, Stuart Ambler, Xi Chen, Yiqun Liu, Parul Jain, Liang Zhang, Ganesh Venkataraman, Tim Converse, Deepak Agarwal

Application Engineering: Ahsan Chudhary, Alan Yang, Alex Navasardyan, Brandyn Bennett, Hrishikesh S, Jim Tao, Juan Pablo Lomeli Diaz, Patrick Schutz, Ricky Yan, Lu Zheng, Stephanie Chou, Joseph Florencio, Santosh Kumar Kancha, Anthony Duerr

Product: Ryan Sandler, Keren Baruch

Other teams (UED, Marketing, BizOps, Analytics, Testing, Voice of Members, Security, …): Julie Kuang, Phil Bunge, Prateek Janardhan, Fiona Li, Bharath Shetty, Sunil Mahadeshwar, Cory Scott, Tushar Dalvi, and team

Acknowledgements

David Freeman, Ashish Gupta, David Hardtke, Rong Rong, Ram Swaminathan

Responsible AI Case Studies at Amazon

Krishnaram Kenthapadi

Amazon AWS AI

aws machine learning

Amazon SageMaker Clarify

Detect bias in ML models and understand model predictions

Amazon SageMaker Customer ML Use cases

https://aws.amazon.com/sagemaker/getting-started

Predictive Maintenance

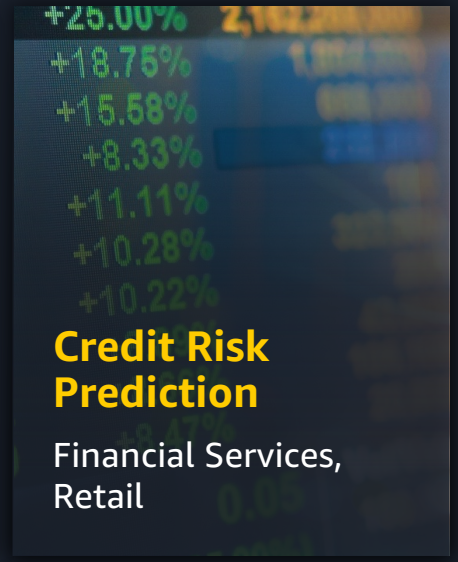
Manufacturing, Automotive, IoT

Demand Forecasting

Retail, Consumer Goods, Manufacturing

Fraud Detection

Financial Services, Online Retail

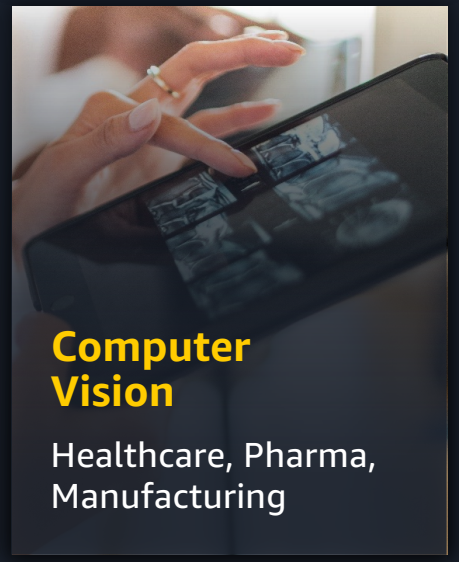


Credit Risk Prediction

Financial Services, Retail

Extract and Analyze Data from Documents

Healthcare, Legal, Media/Ent, Education



Computer Vision

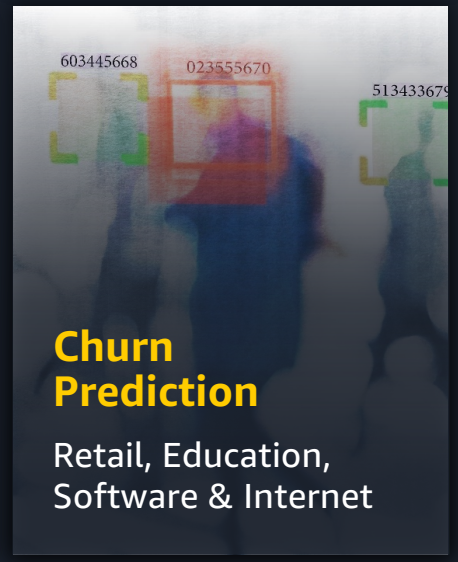
Healthcare, Pharma, Manufacturing

Autonomous Driving

Automotive, Transportation

Personalized Recommendations

Media & Entertainment, Retail, Education



Churn Prediction

Retail, Education, Software & Internet

Bias and Explainability: Challenges

1. Without detection, it is hard to know if bias has entered an ML model:

   - Imbalances may be present in the initial dataset
   - Bias may develop during training
   - Bias may develop over time after model deployment

2. Machine learning models are often complex & opaque, making explainability critical:

   - Regulations may require companies to be able to explain model predictions
   - Internal stakeholders and customers may need explanations for model behavior
   - Data science teams can improve models if they understand model behavior

Amazon SageMaker Clarify

Detect bias in ML models and understand model predictions

Identify imbalances in data

Detect bias during data preparation

Check your trained model for bias

Evaluate the degree to which various types of bias are present in your model

Explain overall model behavior

Understand the relative importance of each feature to your model's behavior

Explain individual predictions

Understand the relative importance of each feature for individual inferences

Detect drift in bias and model behavior over time

Provide alerts and detect drift over time due to changing real-world conditions

Generated automated reports

Produce reports on bias and explanations to support internal presentations

aws machine learning

SageMaker Clarify Use Cases

Regulatory Compliance

Internal Reporting

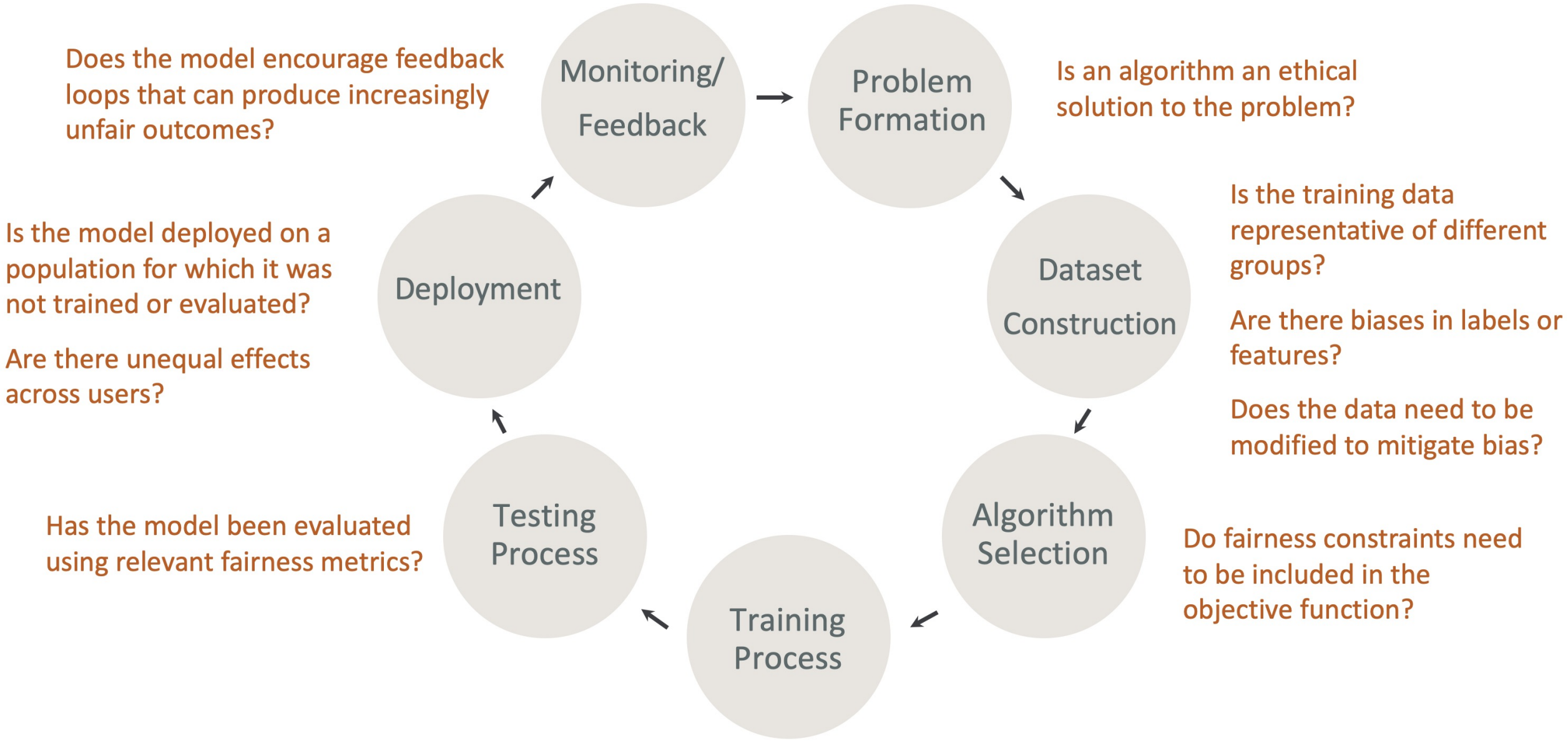
Operational Excellence

Customer Service

Lessons learned

- Fairness as a Process
  - Notions of bias & fairness are highly application dependent
  - Choice of the attribute(s) for which bias is to be measured & the choice of the bias metrics to be guided by social, legal, and other non-technical considerations
- Collaboration/consensus across key stakeholders
- Wide spectrum of customers with different levels of technical background
  - Managed service vs. open source packages
- Monitoring of the deployed model
- Fairness & explainability considerations across the ML lifecycle

Fairness and Explainability by Design in the ML Lifecycle



Additional Pointers

For more information on Amazon SageMaker Clarify, please refer:

- https://aws.amazon.com/sagemaker/clarify
- Amazon Science / AWS Articles
    - https://aws.amazon.com/blogs/aws/new-amazon-sagemaker-clarify-detects-bias-and-increases-the-transparency-of-machine-learning-models
    - https://www.amazon.science/latest-news/how-clarify-helps-machine-learning-developers-detect-unintended-bias
- Technical papers: (1) Amazon SageMaker Clarify [KDD'21] (2) Fairness Measures for Machine Learning in Finance
- https://github.com/aws/amazon-sagemaker-clarify

Acknowledgments: Amazon SageMaker Clarify core team, Amazon AWS AI team, and partners across Amazon

Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud

Michaela Hardt[1], Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, Ashish Rathi, Scott Rees, Ankit Siva, ErhYuan Tsai[2], Keerthan Vasist, Pinar Yilmaz, M. Bilal Zafar, Sanjiv Das[3], Kevin Haas, Tyler Hill, Krishnaram Kenthapadi
Amazon Web Services

ABSTRACT

Understanding the predictions made by machine learning (ML) models and their potential biases remains a challenging and labor-intensive task that depends on the application, the dataset, and the specific model. We present Amazon SageMaker Clarify, an explainability feature for Amazon SageMaker that launched in December 2020, providing insights into data and ML models by identifying biases and explaining predictions. It is deeply integrated into Amazon SageMaker, a fully managed service that enables data scientists and developers to build, train, and deploy ML models at any scale.

1 INTRODUCTION

Machine learning (ML) models and data-driven systems are increasingly used to assist in decision-making across domains such as financial services, healthcare, education, and human resources. Benefits of using ML include improved accuracy, increased productivity, and cost savings. The increasing adoption of ML is the result of multiple factors, most notably ubiquitous connectivity, the ability to collect, aggregate, and process large amounts of data using cloud computing, and improved access to increasingly sophisticated ML models. In high-stakes settings, tools for bias and explainability in

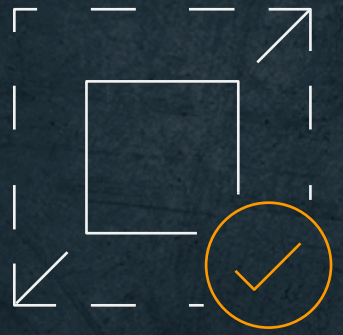
Amazon SageMaker Debugger

Debug and profile ML model training and get real-time insights

Why debugging and profiling

Training ML models is difficult and compute intensive

Training bugs



Large compute instances

Long training times

aws

SageMaker Debugger

Relevant data capture

Zero code change

Persistent in your S3 bucket

System resource usage
Time spent by training operations

Automatic error detection

Built-in and custom rules

Early termination

Detect performance bottlenecks

Real-time monitoring

Debug data while training is ongoing

Monitor utilization
Profile by step or time duration

Save time and cost

Find issues early

Accelerate prototyping

Right size instance
Improve utilization
Reduce cost

SageMaker Studio integration

Alerts about rule status

View suggestions on resolving bottlenecks,
Interactive visualizations

AMAZON SAGEMAKER DEBUGGER: A SYSTEM FOR REAL-TIME INSIGHTS INTO MACHINE LEARNING MODEL TRAINING

Nathalie Rauschmayr [1] **Vikas Kumar** [1] **Rahul Huilgol** [1] **Andrea Olgiati** [1] **Satadal Bhattacharjee** [1]
Nihal Harish [1] **Vandana Kannan** [1] **Amol Lele** [1] **Anirudh Acharya** [1] **Jared Nielsen** [1] **Lakshmi Ramakrishnan** [1]
Ishaaq Chandy [1] **Ishan Bhatt** [1] **Zhihan Li** [1] **Kohen Chia** [1] **Neelesh Dodda** [1] **Jiacheng Gu** [1] **Miyoung Choi** [1]
Balajee Nagarajan [1] **Jeffrey Geevarghese** [1] **Denis Davydenko** [1] **Sifei Li** [1] **Lu Huang** [1] **Edward Kim** [1] **Tyler Hill** [1]
Krishnaram Kenthapadi [1]

ABSTRACT

Manual debugging is a common productivity drain in the machine learning (ML) lifecycle. Identifying underperforming training jobs requires constant developer attention and deep domain expertise. As state-of-the-art models grow in size and complexity, debugging becomes increasingly difficult. Just as unit tests boost traditional software development, an automated ML debugging library can save time and money. We present Amazon SageMaker Debugger, a machine learning feature that automatically identifies and stops underperforming training jobs. Debugger is a new feature of Amazon SageMaker that automatically captures relevant data during training and evaluation and presents it for online and offline inspection. Debugger helps users define a set of conditions, in the form of built-in or custom rules, that are applied to this data, thereby enabling users to catch training issues as well as monitor and debug ML model training in real-time. These rules save time and money by alerting the developer and terminating a problematic training job early.

https://www.amazon.science/publications/amazon-sagemaker-debugger-a-system-for-real-time-insights-into-machine-learning-model-training

aws

Additional Pointers

For more information on Amazon SageMaker Debugger, please refer:

- https://aws.amazon.com/sagemaker/debugger
- AWS Articles
  - https://aws.amazon.com/blogs/aws/amazon-sagemaker-debugger-debug-your-machine-learning-models
  - https://aws.amazon.com/blogs/machine-learning/detecting-hidden-but-non-trivial-problems-in-transfer-learning-models-using-amazon-sagemaker-debugger
- Technical paper: Amazon SageMaker Debugger: A System for Real-Time Insights into Machine Learning Model Training (MLSys 2021)
- https://pypi.org/project/smdebug

Acknowledgments: Amazon SageMaker Debugger core team, Amazon AWS AI team, and partners across Amazon

Fairness for Opaque Models via Model Tuning (Hyperparameter Optimization)

- Can we tune the hyperparameters of a model to achieve both accuracy and fairness?
- Can we support both opaque models and opaque fairness constraints?

- Use Bayesian optimization for HPO with fairness constraints!
  - Explore hyperparameter configurations where fairness constraints are satisfied

V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, C. Archambeau, Fair Bayesian Optimization, AIES 2021
(Best paper award @ ICML 2020 AutoML workshop)

Human-in-the-loop frameworks

Desirable to augment ML model predictions with expert inputs

Useful for improving accuracy, incorporating additional information, and auditing models.

Popular examples – Healthcare models

Human–machine partnership with artificial intelligence for chest radiograph diagnosis

Bhavik N. Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A. J. Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker & Matthew P. Lungren

Content moderation tasks

AI won't relieve the misery of Facebook's human moderators

The problem of online content moderation can't be solved with artificial intelligence, say experts

By James Vincent | Feb 27, 2019, 12:41pm EST

Child maltreatment hotline screening

A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores

Maria De-Arteaga*
Heinz College
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
mdeartea@andrew.cmu.edu

Riccardo Fogliato*
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA, USA
rfogliat@andrew.cmu.edu

Alexandra Chouldechova
Heinz College
Carnegie Mellon University
Pittsburgh, PA, USA
achould@cmu.edu

AI proves it's a poor substitute for human content checkers during lockdown

Spandana Singh May 23, 2020 10:25 AM AI

Errors and biases in human-in-the-loop frameworks

ML tasks often suffer from group-specific bias, induced due to misrepresentative data or models.

The algorithms that detect hate speech online are biased against black people

A new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as "offensive" compared to other tweets.

By Shirin Ghaffary | Aug 15, 2019, 11:00am EDT

Human-in-the-loop frameworks can reflect biases or inaccuracies of the human experts.

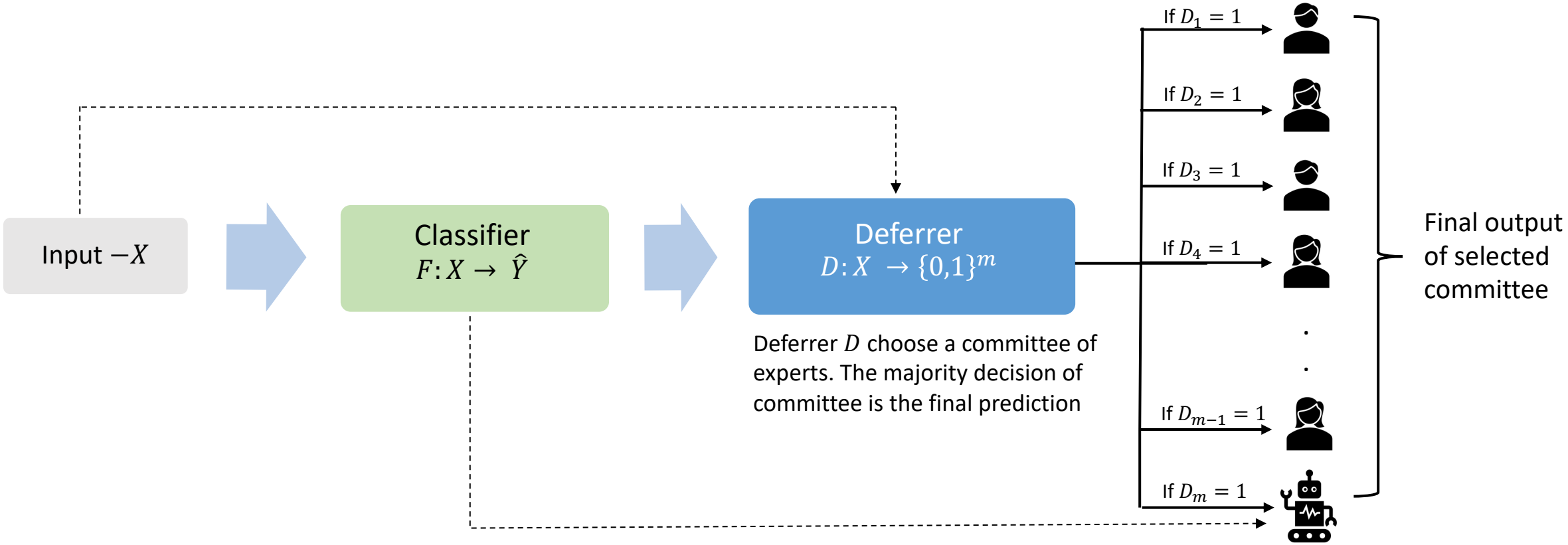
Concerns include:

- Racial bias in human-in-the-loop framework for recidivism risk assessment (Green, Chen – FAccT 2019)
- Ethical concerns regarding audits of commercial facial processing technologies (Raji et al. – AIES 2020)
- Automation bias in time critical decision support systems (Cummings – ISTC 2004)

Can we design human-in-the-loop frameworks that take into account the expertise and biases of the human experts?

Model

X – non-protected attributes; Y – class label; Z – protected attribute/group membership
Number of experts available $= m - 1$

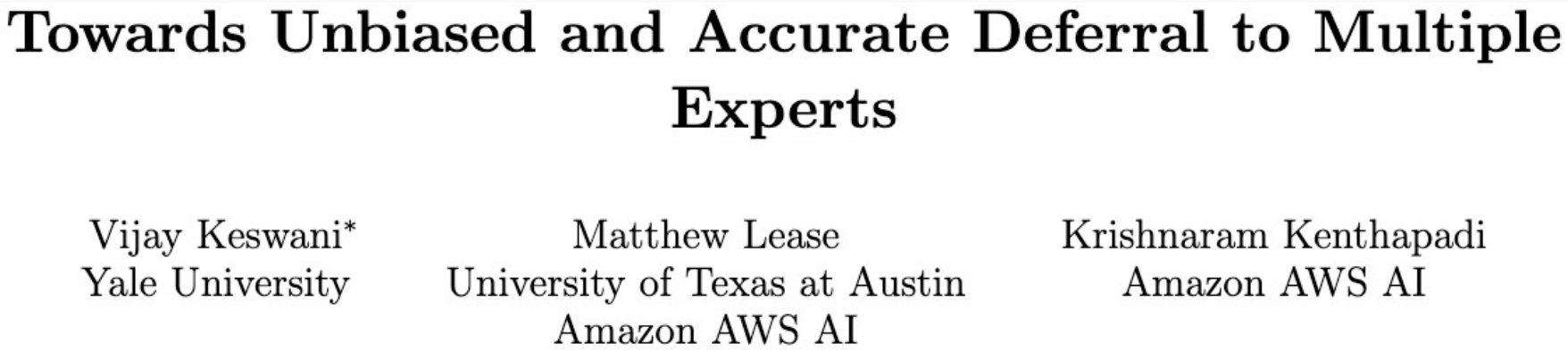


Experts might have access to additional information, including group membership Z.

There might be a cost/penalty associated with each expert review.

Fairness in human-in-the-loop settings

- Joint learning framework to learn a classifier and a deferral system for multiple experts simultaneously
- Synthetic and real-world experiments on the efficacy of our method



V. Keswani, M. Lease, K. Kenthapadi, Towards Unbiased and Accurate Deferral to Multiple Experts, AIES 2021.

Minimax Group Fairness: Algorithms and Experiments

- Equality of error rates: an intuitive & well-studied group fairness notion
  - May require artificially inflating error on easier-to-predict groups ☹
  - Undesirable when most/all of the targeted population is disadvantaged

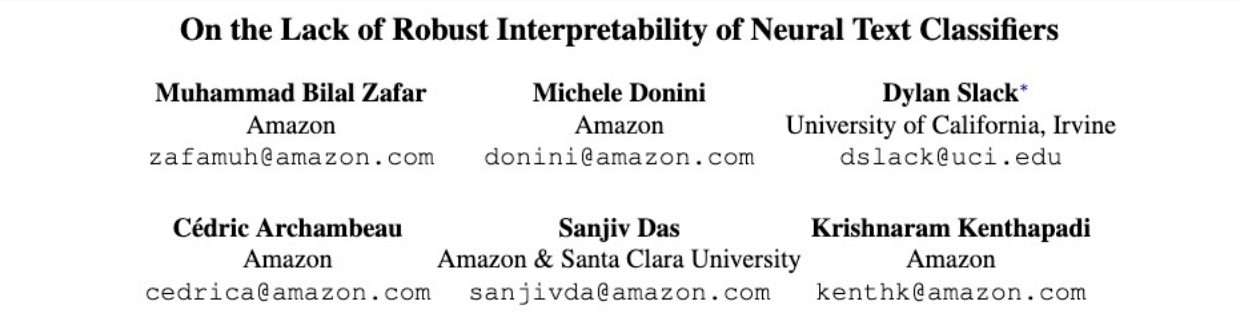
- Goal: *minimize maximum group error* [Martinez et al, 2020]
  - "Ensure that the worst-off group is as well-off as possible"

- Our work: algorithms based on a zero-sum game between a Learner and a Regulator
  - Theoretical results and experimental evaluation

E. Diana, W. Gill, M. Kearns, K. Kenthapadi, A. Roth, Minimax Group Fairness: Algorithms and Experiments, AIES 2021.

Robust Interpretability of Neural Text Classifiers

- Feature attribution methods used for understanding model predictions
- How robust are feature attributions for neural text classifiers?
  - Are they identical under different random initializations of the same model?
  - Do they differ between a model with trained parameters and a model with random parameters?
- Common feature attribution methods fail both tests!



On the Lack of Robust Interpretability of Neural Text Classifiers

Muhammad Bilal Zafar
Amazon
zafamuh@amazon.com

Michele Donini
Amazon
donini@amazon.com

Dylan Slack*
University of California, Irvine
dslack@uci.edu

Cédric Archambeau
Amazon
cedrica@amazon.com

Sanjiv Das
Amazon & Santa Clara University
sanjivda@amazon.com

Krishnaram Kenthapadi
Amazon
kenthk@amazon.com

M. B. Zafar, M. Donini, D. Slack, C. Archambeau, S. Das, K. Kenthapadi, On the Lack of Robust Interpretability of Neural Text Classifiers, Findings in ACL 2021.

Privacy Research @ Amazon - Sampler

Differentially Private Query Release

- Problem: answering marginal queries privately with high accuracy
  - Marginal queries are a special class of linear queries that count slices of a dataset.
  - “How many authors have visited Charlotte, graduated in the last two year and work in the Bay Area?” – A 3-way marginal query on the below dataset.

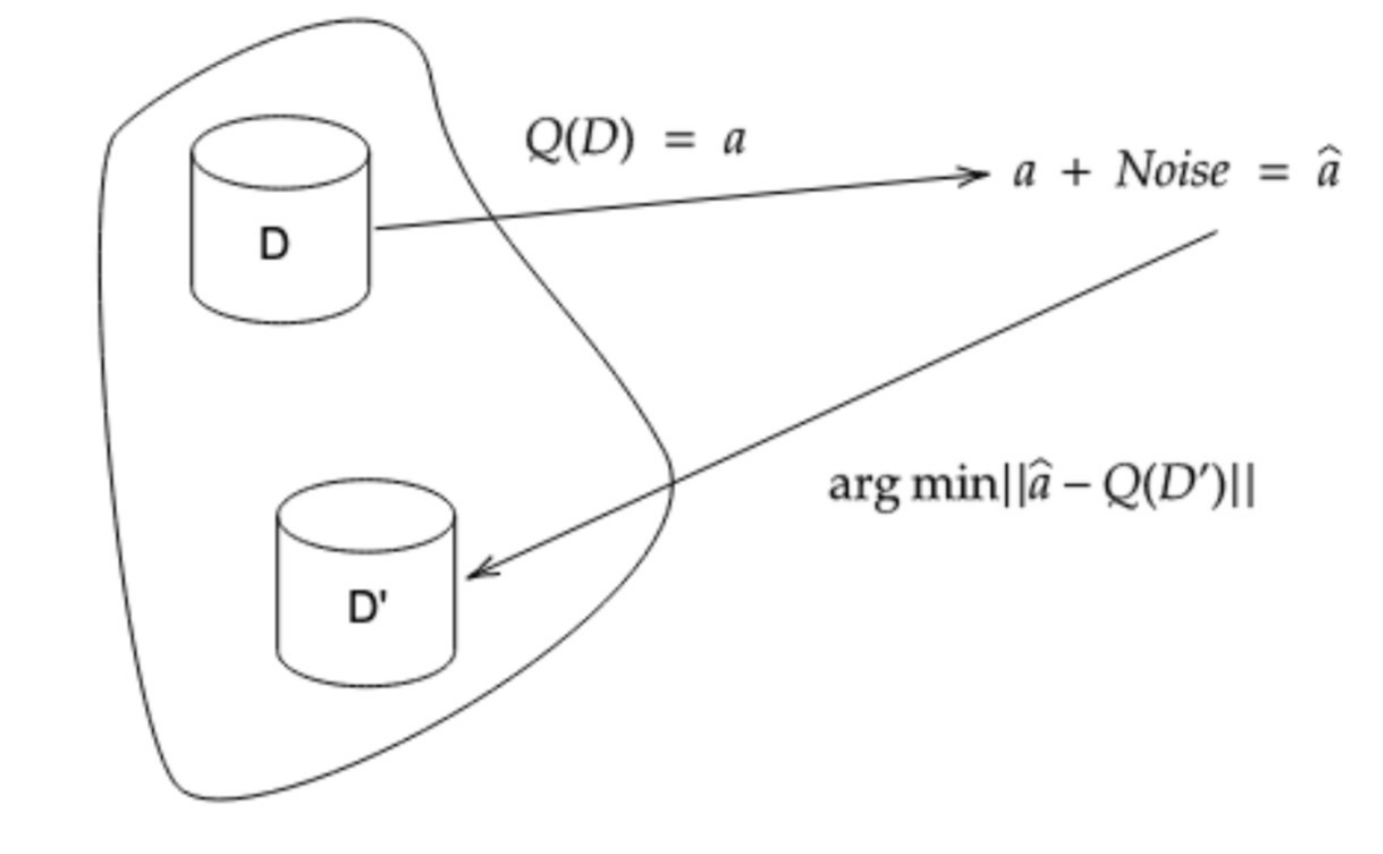
Name	Visited Charlotte?	Recent Grad?	Office
Ankit	Yes	Yes	Bay Area
Luca	No	Yes	New York
Aaron	Yes	No	Philadelphia
⋮	⋮	⋮	⋮

Table 1. Sensitive data about the authors

- Privacy: Marginals computed against our dataset should protect against inferences on an individual’s membership (using Differential Privacy)

Differentially Private Query Release

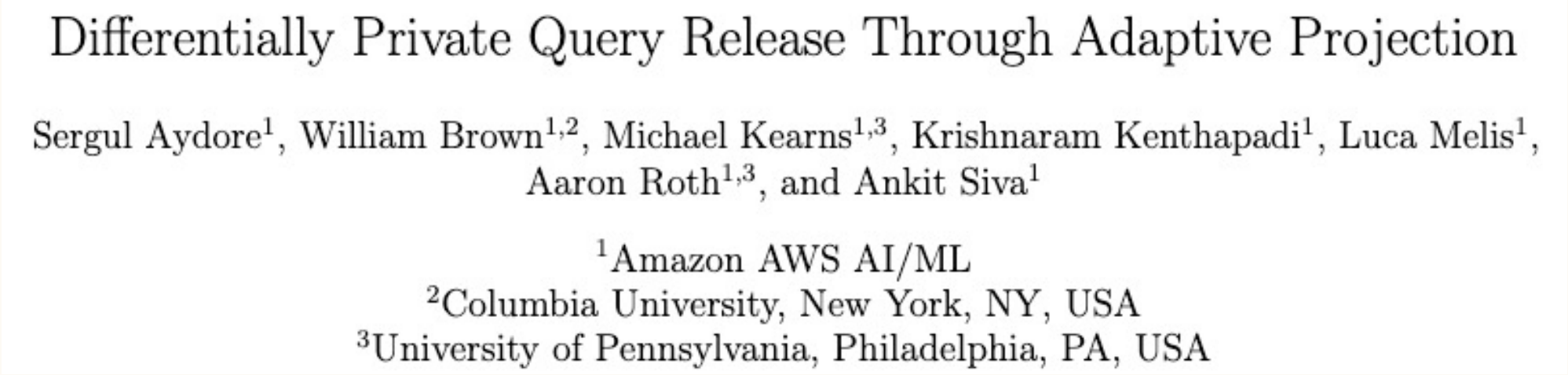
- Projection Mechanism: Evaluate noisy answers to all queries in a query class Q and find the synthetic dataset (D’) in the space of feasible datasets that minimizes error with respect to some norm. Q is a class of queries.



Nikolov, Talwar, Zhang, The geometry of differential privacy: the sparse and approximate cases, STOC 2013

Differentially Private Query Release: Key Ideas

1. Relax the data domain: one-hot encode the non-continuous data and expand the domain to real numbers. Extend the differentiable queries to the new domain.
2. Adaptively select queries: repeatedly choose the k worst performing queries privately and optimize D’ to answer those well.



Differentially Private Query Release Through Adaptive Projection

Sergul Aydore[1], William Brown[1,2], Michael Kearns[1,3], Krishnaram Kenthapadi[1], Luca Melis[1], Aaron Roth[1,3], and Ankit Siva[1]

[1]Amazon AWS AI/ML
[2]Columbia University, New York, NY, USA
[3]University of Pennsylvania, Philadelphia, PA, USA

S. Aydore, W. Brown, M. Kearns, K. Kenthapadi, L. Melis, A. Roth, A. Siva, Differentially Private Query Release Through Adaptive Projection, ICML 2021.

Responsible AI Case Studies at Google

Ben Packer

Google AI

Google Assistant

Google Assistant

Key Points:

- Think about user harms
  How does your product make people feel
- Adversarial ("stress") testing for all Google Assistant launches
  - People might say racist, sexist, homophobic stuff
- Diverse testers
- Think about expanding who your users could and should be
- Consider the diversity of your users

Computer Vision

Google Camera

Key points:

- Check for unconscious bias
- Comprehensive testing: "make sure this works for everybody"

Night Sight

Night Sight: Seeing in the Dark on Pixel Phones

Wednesday, November 14, 2018

Posted by Marc Levoy, Distinguished Engineer and Yael Pritch, Staff Software Engineer

Night Sight is a new feature of the Pixel Camera app that lets you take sharp, clean photographs in very low light, even in light so dim you can't see much with your own eyes. It works on the main and selfie cameras of all three generations of Pixel phones, and does not require a tripod or flash. In this article we'll talk about why taking pictures in low light is challenging, and we'll discuss the computational photography and machine learning techniques, much of it built on top of HDR+, that make Night Sight work.

Left: iPhone XS (full resolution image here). **Right:** Pixel 3 Night Sight (full resolution image here).

SKIN TONE IN PHOTOGRAPHY

This is a “Shirley Card”

Named after a Kodak studio model named Shirley Page, they were the primary method for calibrating color when processing film.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

How Kodak's Shirley Cards Set Photography's Skin-Tone Standard, NPR

SKIN TONE IN PHOTOGRAPHY

Until about 1990, virtually all Shirley Cards featured Caucasian women.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

Colour Balance, Image Technologies, and Cognitive Equity, Roth

How Photography Was Optimized for White Skin Color (Priceonomics)

SKIN TONE IN PHOTOGRAPHY

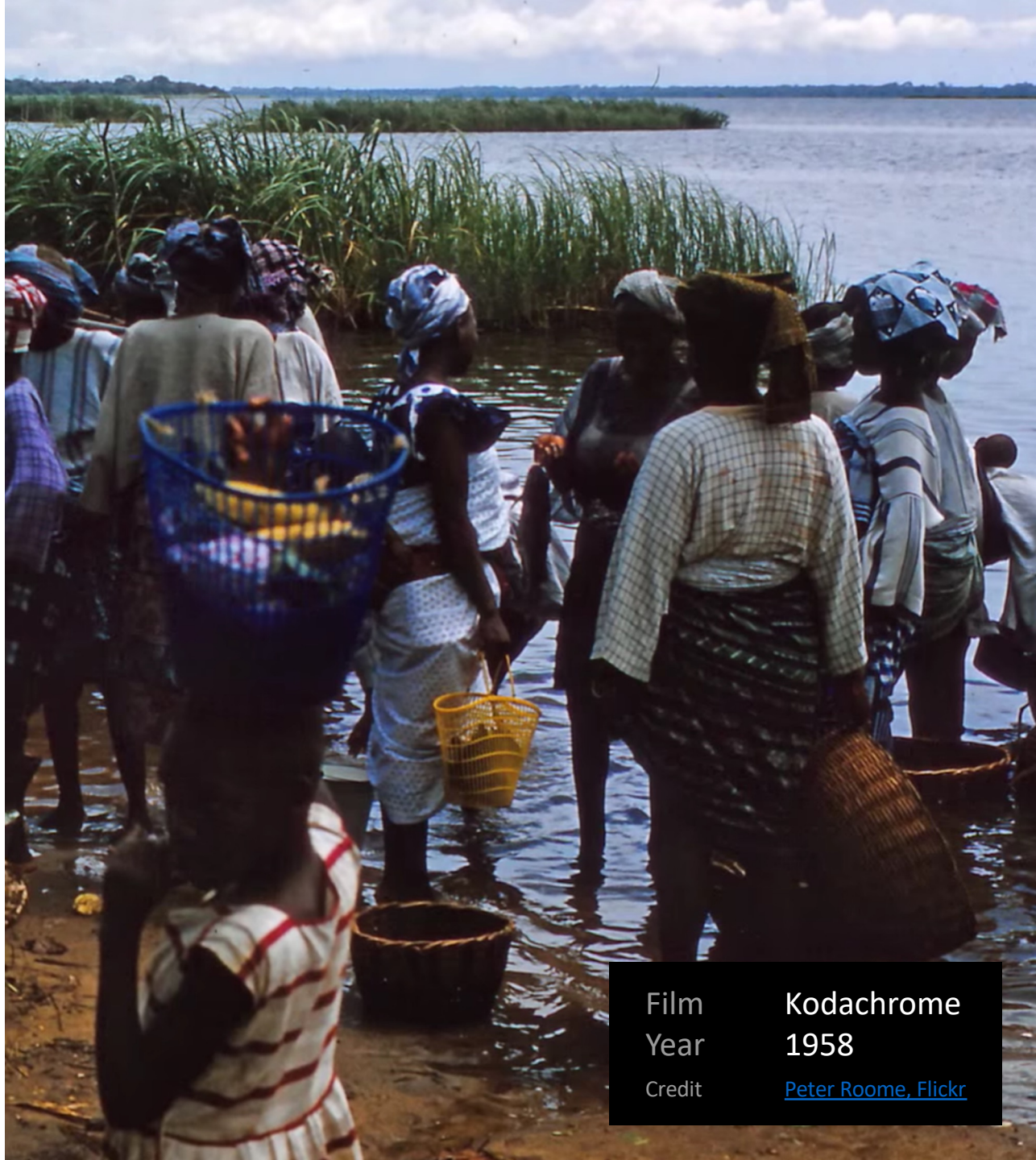
As a result, photos featuring people with light skin looked fairly accurate.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity, Roth
How Photography Was Optimized for White Skin Color (Priceonomics)

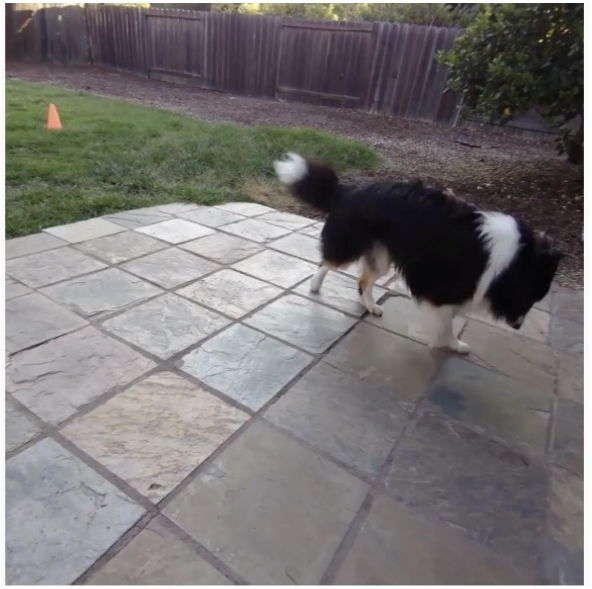
SKIN TONE IN PHOTOGRAPHY

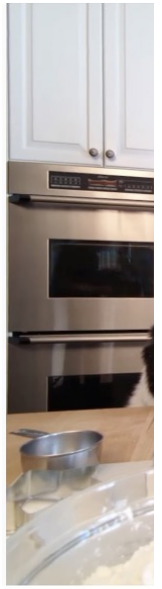
Photos featuring people with darker skin, not so much...



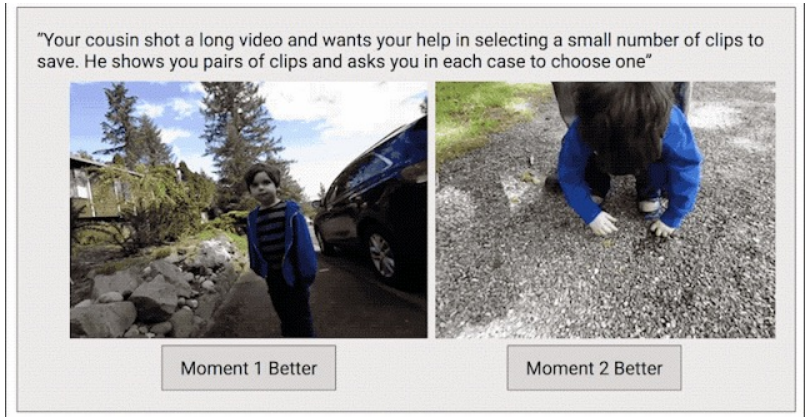
SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity, Roth
How Photography Was Optimized for White Skin Color (Priceonomics)





Google Clips



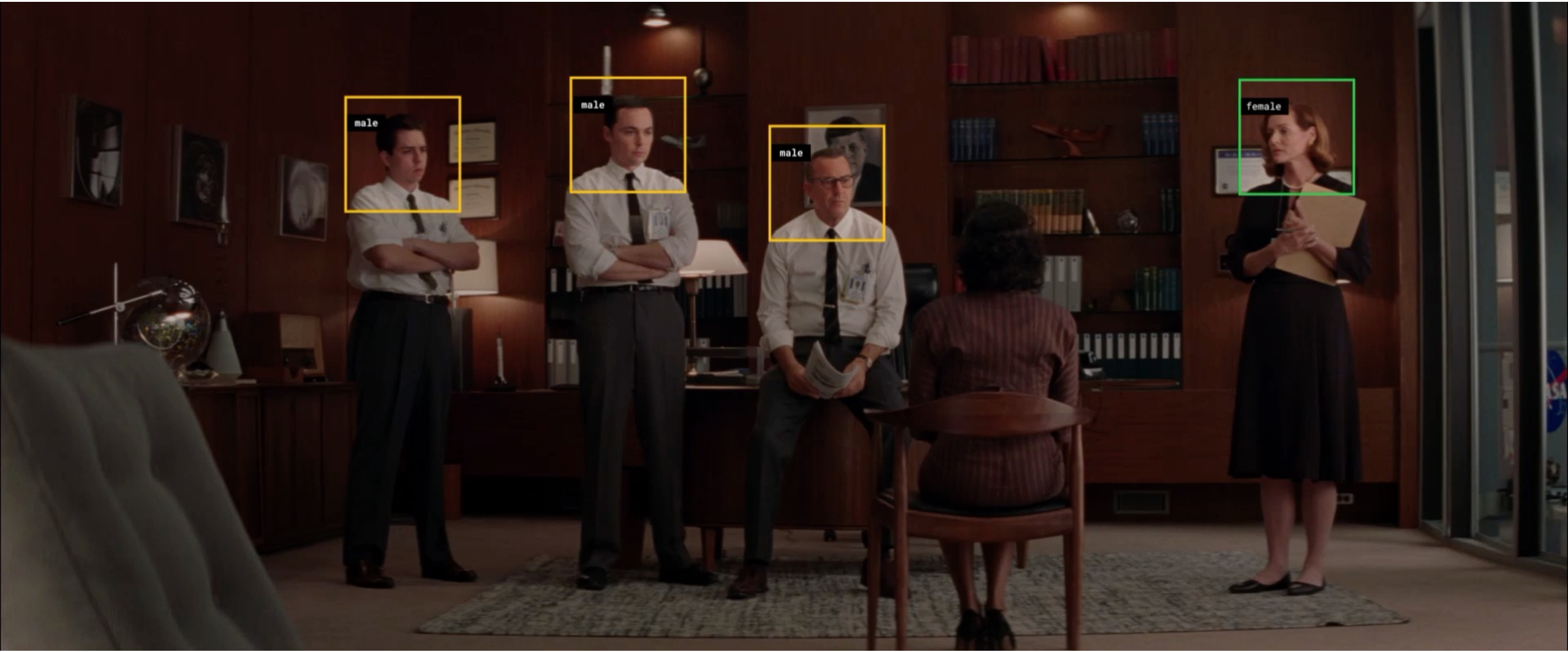
Google Clips

"We created **controlled datasets by sampling subjects from different genders and skin tones in a balanced manner**, while keeping variables like content type, duration, and environmental conditions constant. We then used this dataset to test that our algorithms had similar performance when applied to different groups."

https://ai.googleblog.com/2018/05/automatic-photography-with-google-clips.html

Geena Davis Inclusion Quotient

[with Geena Davis Institute on Gender in Media]



Smart Compose

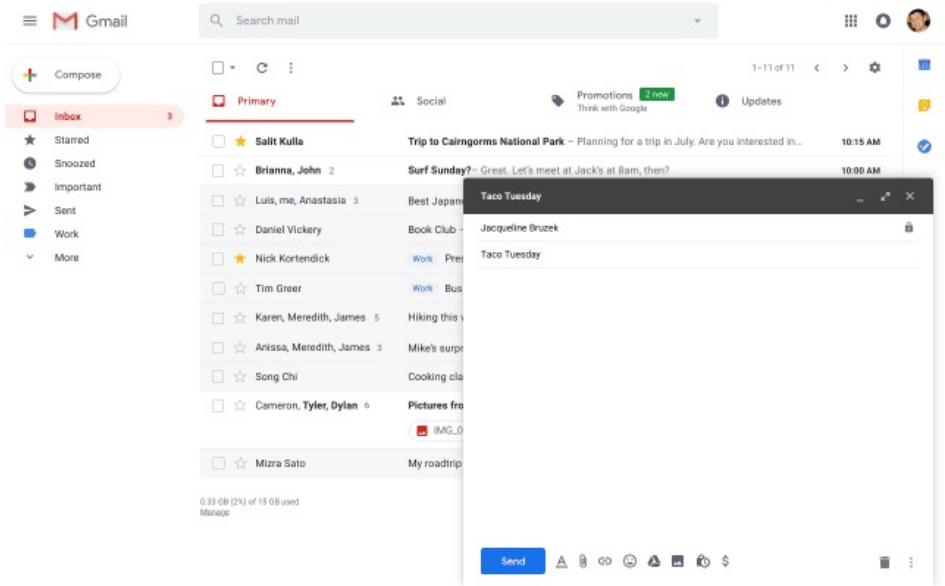
Adversarial Testing for Smart Compose in Gmail

Google

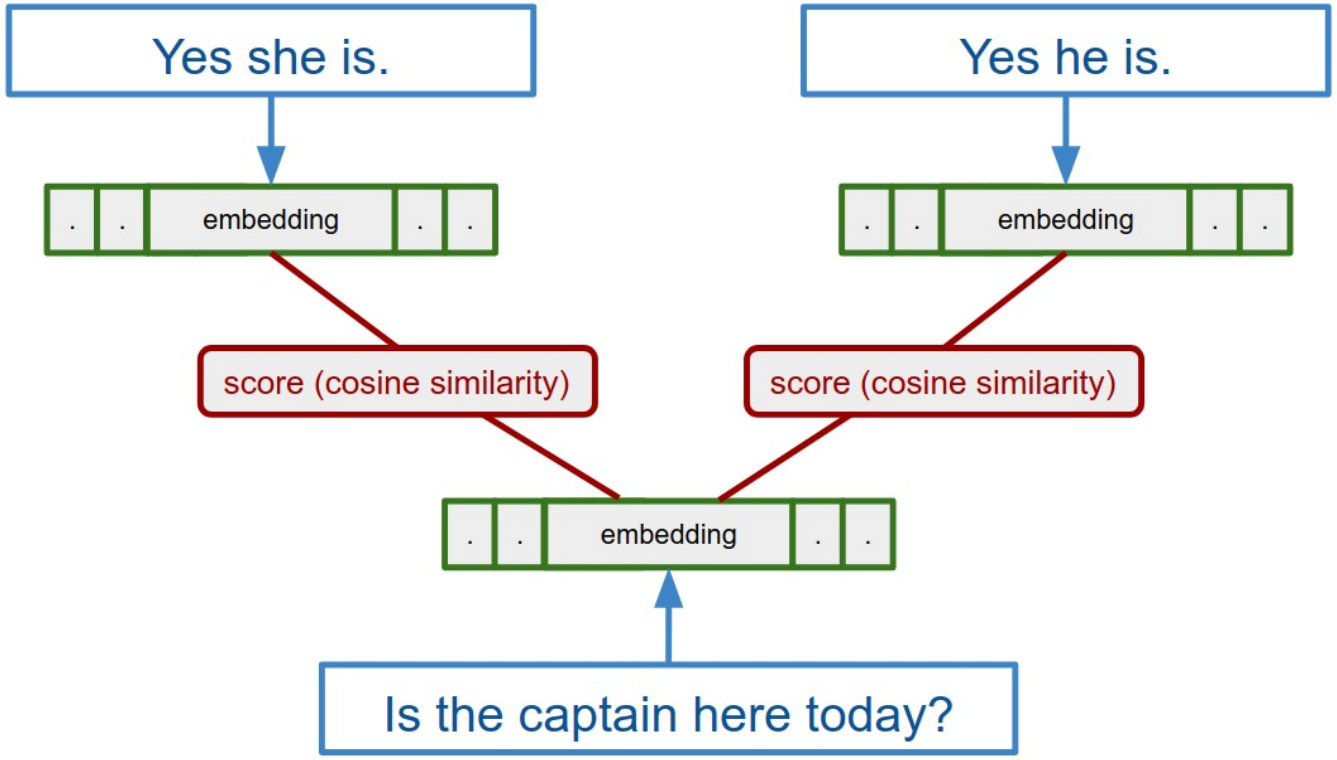
The Keyword Latest Stories Product Updates Company News

GMAIL

SUBJECT: Write emails faster with Smart Compose in Gmail



Embedding Model



	score
Yes she is	0.17
Yes he is	0.27
Bias score (difference)	-0.10

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

Embedding Model

Highest female association

occupation	bias		
nurse	2.5	dancer	0.9
receptionist	2.2	hairdresser	0.9
nanny	2.1	cashier	0.9
secretary	1.9	realtor	0.8
housekeeper	1.7	teacher	0.7
midwife	1.4	dishwasher	0.7
florist	1.3	therapist	0.7
clerk	1.2	pharmacist	0.6
stylist	1.2	nutritionist	0.5
librarian	1.1	dietitian	0.4

Highest male association

occupation	bias		
undertaker	-3.3	magician	-1.8
actor	-2.4	analyst	-1.8
president	-2.3	architect	-1.8
composer	-2.3	painter	-1.7
janitor	-2.1	butcher	-1.7
barber	-2.1	historian	-1.7
philosopher	-2.1	captain	-1.6
plumber	-2	engineer	-1.6
bodybuilder	-2	referee	-1.6
developer	-1.9	programmer	-1.6

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

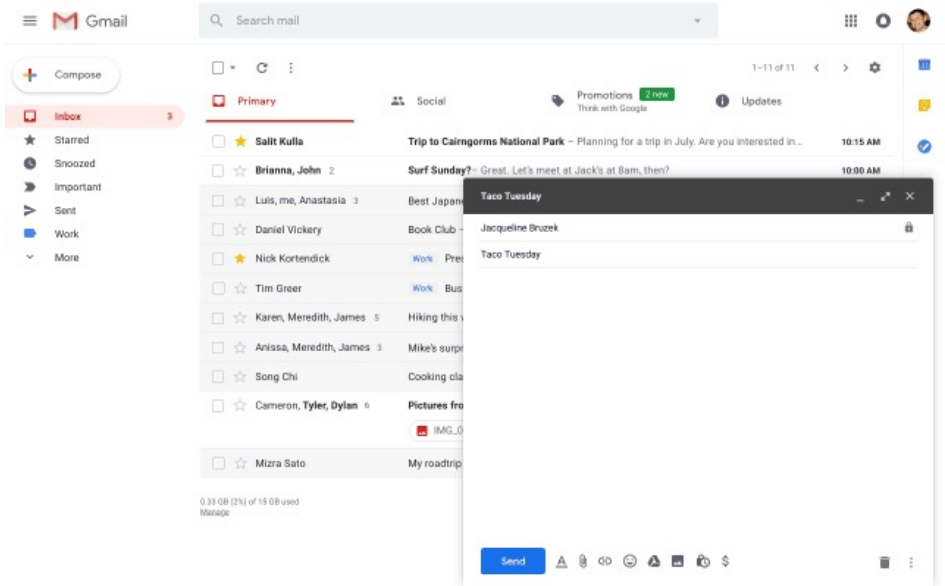
Adversarial Testing for Smart Compose in Gmail

Google

The Keyword Latest Stories Product Updates Company News

GMAIL

SUBJECT: Write emails faster with Smart Compose in Gmail



Machine Translation

(Historical) Gender Pronouns in Translate

Three Step Approach

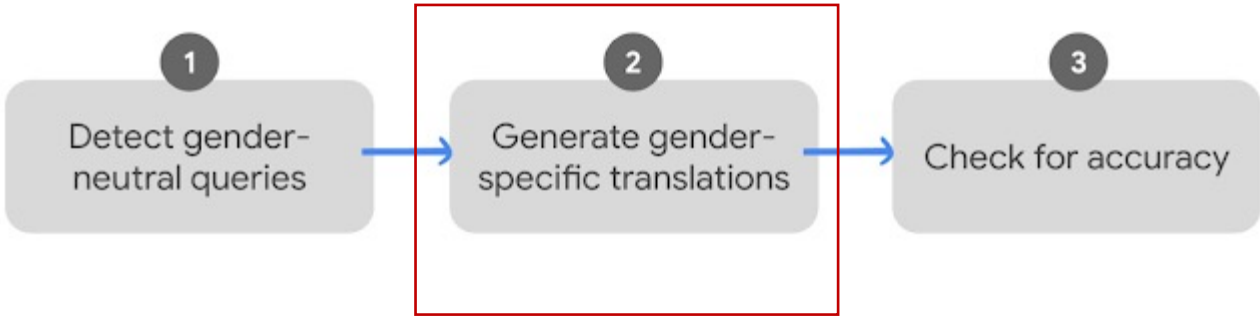
1. Detect Gender-Neutral Queries

Train a text classifier to detect when a Turkish query is gender-neutral.

- trained on thousands of human-rated Turkish examples

2. Generate Gender-Specific Translations

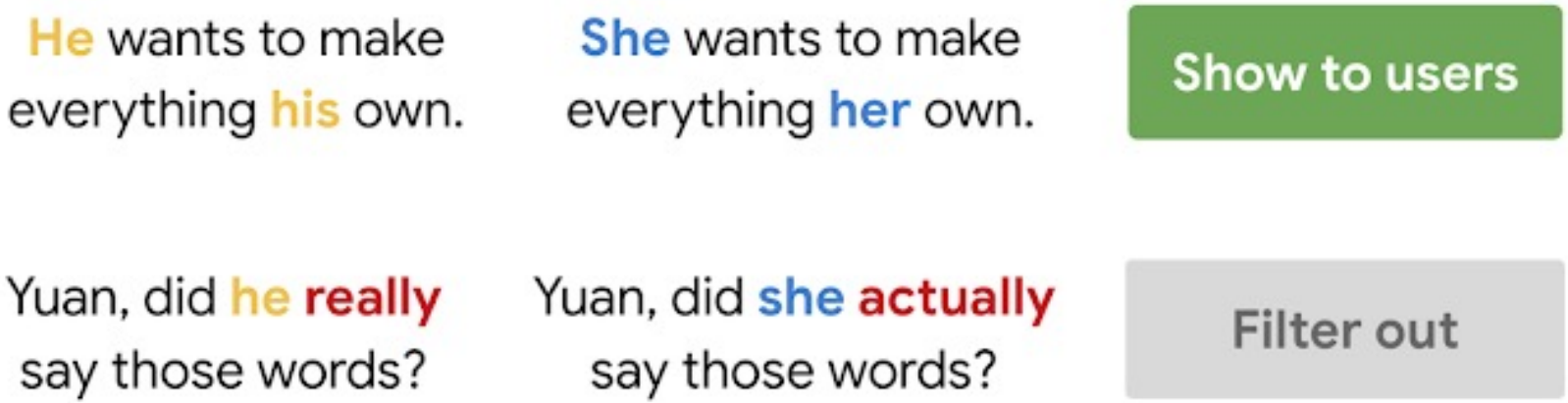
- Training: Modify training data to add an additional input token specifying the required gender:
  - (<2MALE> O bir doktor, He is a doctor)
  - (<2FEMALE> O bir doktor, She is a doctor)
- Deployment: If step (1) predicted query is gender-neutral, add male and female tokens to query
  - O bir doktor -> {<2MALE> O bir doktor, <2FEMALE> O bir doktor}



3. Check for Accuracy

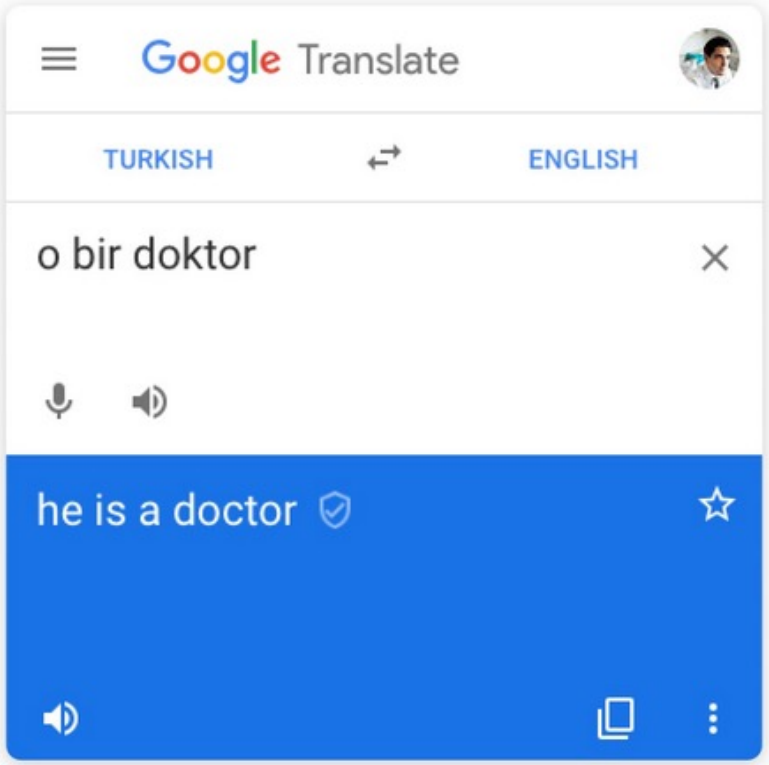
Verify:

1. If the requested feminine translation is feminine.
2. If the requested masculine translation is masculine.
3. If the feminine and masculine translations are exactly equivalent with the exception of gender-related changes.

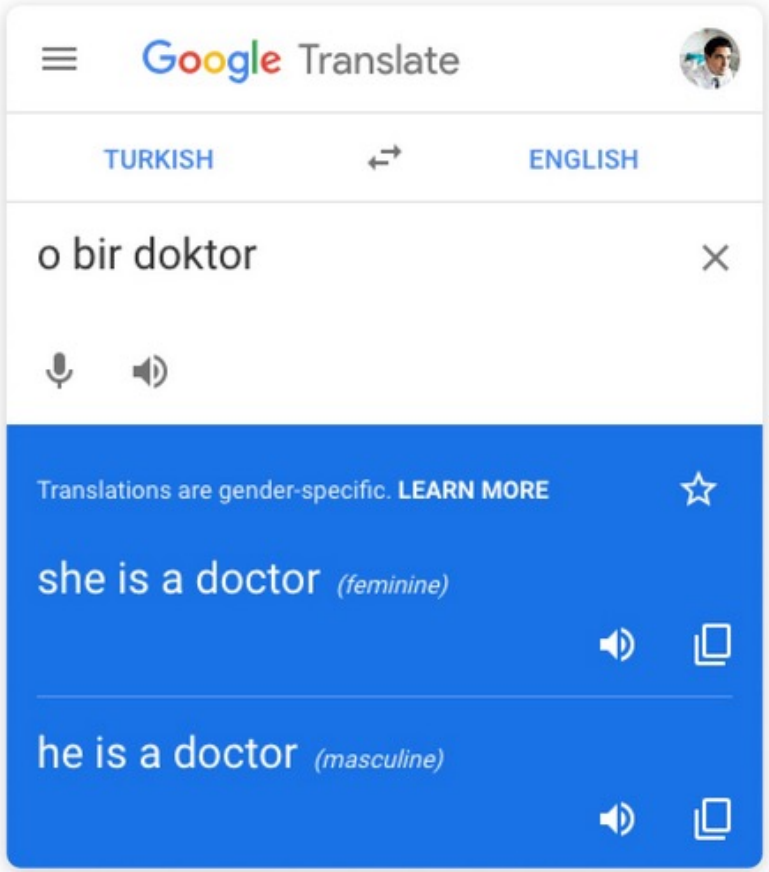


Result: Reduced Gender Bias in Translate

Before



After



Key Takeaways

Krishnaram Kenthapadi

Amazon AWS AI

Good ML Practices Go a Long Way

01

Lots of low hanging fruit in terms of improving fairness simply by using machine learning best practices

- Representative data
- Introspection tools
- Visualization tools
- Testing

02

Fairness improvements often lead to overall improvements

- It’s a common misconception that it’s always a tradeoff

Breadth and Depth Required

01

Looking End-to-End is critical

- Need to be aware of bias and potential problems at every stage of product and ML pipelines (from design, data gathering, … to deployment and monitoring)

02

Details Matter

- Slight changes in features or labeler criteria can change the outcome
- Must have experts who understand the effects of decisions
- Many details are not technical such as how labelers are hired

Process Best Practices

Identify product goals
Get the right people in the room
Identify stakeholders

Policy

Select a fairness approach
Analyze and evaluate your system
Mitigate issues
Monitor Continuously and Escalation Plans
Auditing and Transparency

Technology

Beyond Accuracy

- Performance and Cost
- Fairness and Bias
- Transparency and Explainability
- Privacy
- Security
- Safety
- Robustness

Fairness, Explainability & Privacy: Opportunities

Fairness in ML

Application specific challenges

- Conversational AI systems: Unique bias/fairness/ethics considerations
  - E.g., Hate speech, Complex failure modes
  - Beyond protected categories, e.g., accent, dialect
  - Entire ecosystem (e.g., including apps such as Alexa skills)
- Two-sided markets: e.g., fairness to buyers and to sellers, or to content consumers and producers
- Fairness in advertising (externalities)

Tools for ensuring fairness (measuring & mitigating bias) in AI lifecycle

- Pre-processing (representative datasets; modifying features/labels)
- ML model training with fairness constraints
- Post-processing
- Experimentation & Post-deployment

Key Open Problems in Applied Fairness

What if you don't have the sensitive attributes?

When should you use what approach? For example, Equal treatment vs equal outcome?

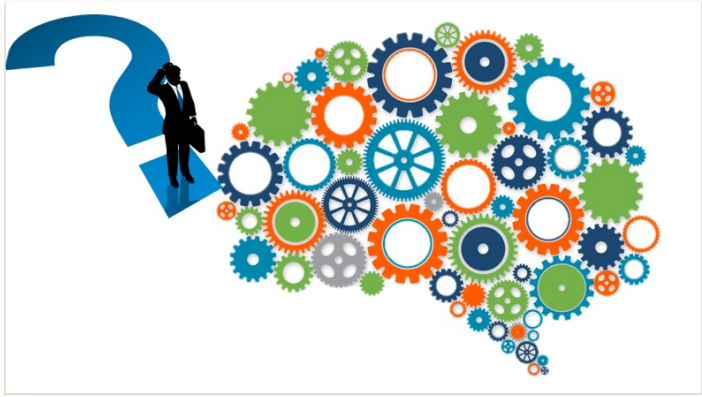
How to identify harms?

Process for framing AI problems: Will the chosen metrics lead to desired results?

How to tell if data generation and collection method is appropriate for a task? (e.g., causal structure analysis?)

Processes for mitigating harms and misbehaviors quickly

Explainability in ML



Actionable explanations

Balance between explanations & model secrecy

Robustness of explanations to failure modes (Interaction between ML components)

Application-specific challenges
- Conversational AI systems: contextual explanations
- Gradation of explanations

Tools for explanations across AI lifecycle
- Pre & post-deployment for ML models
- Model developer vs. End user focused

Privacy in ML

Privacy for highly sensitive data: model training & analytics using secure enclaves, homomorphic encryption, federated learning / on-device learning, or a hybrid

Privacy-preserving model training, robust against adversarial membership inference attacks (Dynamic settings + Complex data / model pipelines)

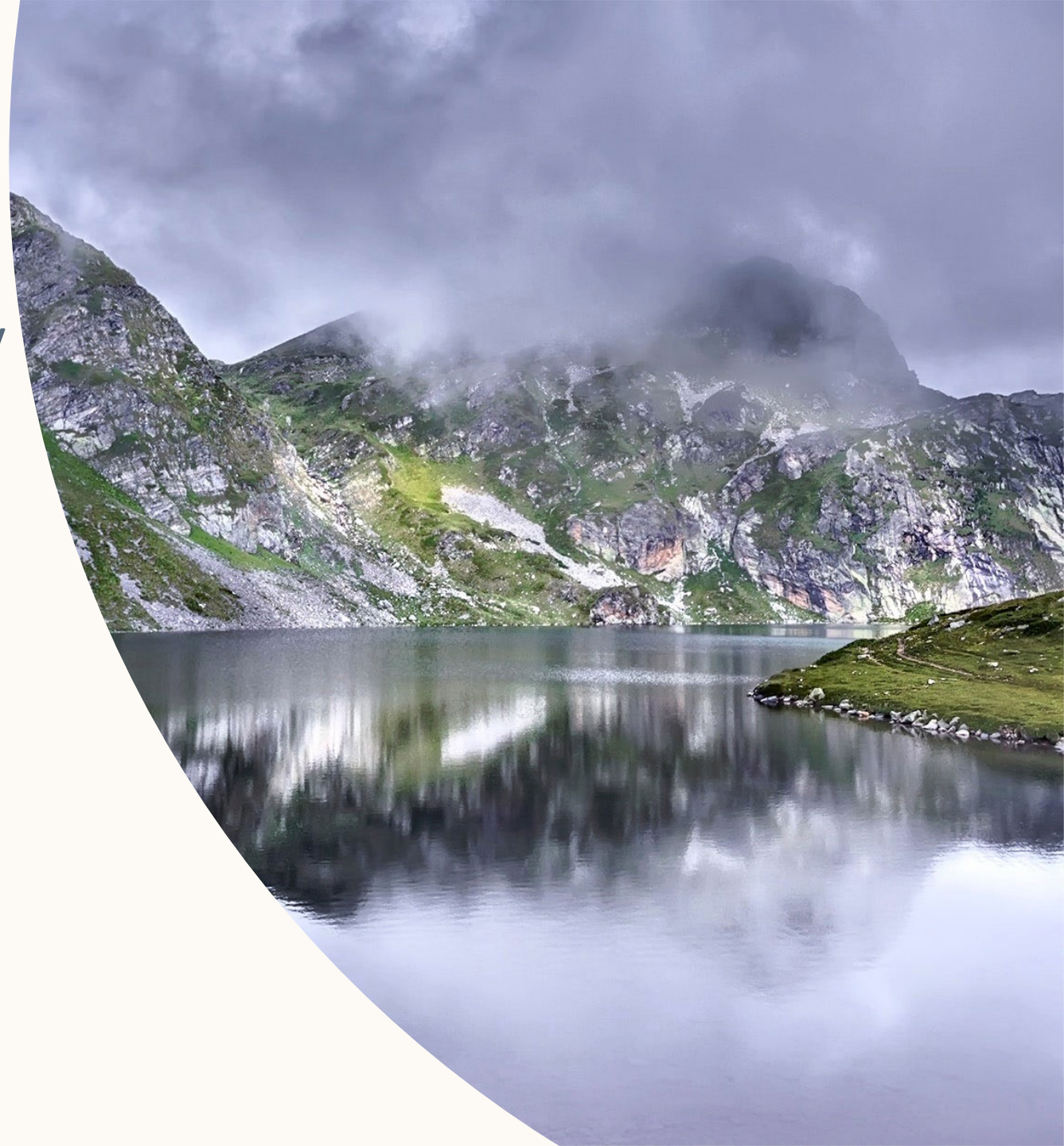
Privacy-preserving mechanisms for data marketplaces

Reflections

“Fairness, Explainability, and Privacy by Design” when building AI products

Collaboration/consensus across key stakeholders

NYT / WSJ / ProPublica test :)



Related Tutorials / Resources

- ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)
- AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)
- Sara Hajian, Francesco Bonchi, and Carlos Castillo, Algorithmic bias: From discrimination discovery to fairness-aware data mining, KDD Tutorial, 2016.
- Solon Barocas and Moritz Hardt, Fairness in machine learning, NeurIPS Tutorial, 2017.
- Kate Crawford, The Trouble with Bias, NeurIPS Keynote, 2017.
- Arvind Narayanan, 21 fairness definitions and their politics, FAccT Tutorial, 2018.
- Sam Corbett-Davies and Sharad Goel, Defining and Designing Fair Algorithms, Tutorials at EC 2018 and ICML 2018.
- Ben Hutchinson and Margaret Mitchell, Translation Tutorial: A History of Quantitative Fairness in Testing, FAccT Tutorial, 2019.
- Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, and Jean Garcia-Gathright, Translation Tutorial: Challenges of incorporating algorithmic fairness into industry practice, FAccT Tutorial, 2019.

Related Tutorials / Resources

- Sarah Bird, Ben Hutchinson, Krishnaram Kenthapadi, Emre Kiciman, Margaret Mitchell, Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned, Tutorials at WSDM 2019, WWW 2019, KDD 2019.
- Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, Ankur Taly, Explainable AI in Industry, Tutorials at KDD 2019, FAccT 2020, WWW 2020.
- Himabindu Lakkaraju, Julius Adebayo, Sameer Singh, Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities, NeurIPS 2020 Tutorial.
- Kamalika Chaudhuri, Anand D. Sarwate, Differentially Private Machine Learning: Theory, Algorithms, and Applications, NeurIPS 2017 Tutorial.
- Krishnaram Kenthapadi, Ilya Mironov, Abhradeep Guha Thakurta, Privacy-preserving Data Mining in Industry, Tutorials at KDD 2018, WSDM 2019, WWW 2019.

Thanks! Questions?

- Tutorial website: **https://sites.google.com/view/ResponsibleAITutorial**

- Feedback most welcome ☺
  - kenthk@amazon.com, bpacker@google.com, mehrnoosh.sameki@microsoft.com, nashlies@amazon.com