PyConKe 2022

3RD ANNUAL PYCON SUMMIT KE

Biggest Gathering For People & Organisations Interested In The Python Programming Language & Its Ecosystem

Detecting Systematic Deviations in Data and Models.

Adebayo Oshingbesan, Tanya Akumu

IBM Research | Africa

Twiga

United States International University-Africa

New Relic

python software foundation

Turing

IBM Research Africa – AI Sciences

Skyler Speakman

Girmaw Tadesse

Celia Cintas

Tanya Akumu

Adebayo Oshingbesan

Content

MOTIVATION

THEORY

EXAMPLES

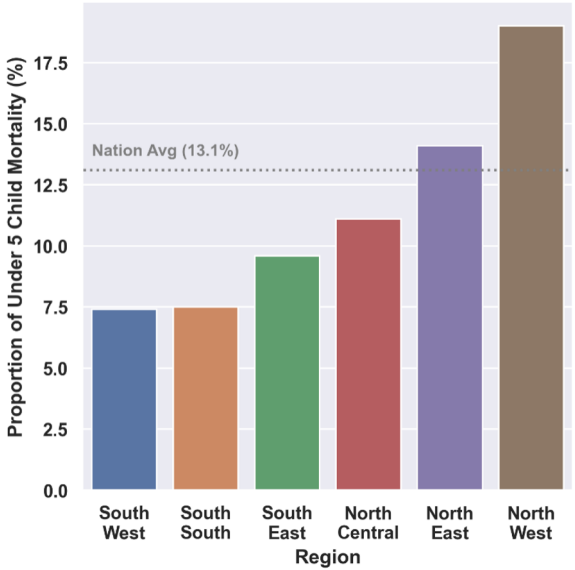
CODE

Motivation

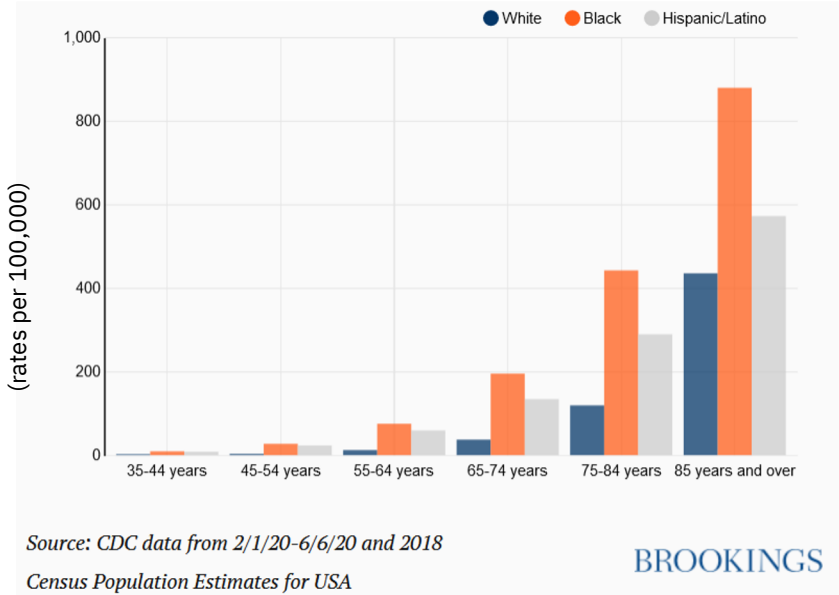
WHY DOES THIS EVEN MATTER?

- Understanding our data using exploratory data analysis
- Improving our data using data-centric AI
- Improving our models through AI fairness and explainability
- Feature selection
- Detecting concept drifts in deployed machine learning systems
- Inferring new classification rules
- Discovering novel treatment effects from an intervention

Under-5 Mortality in Nigeria by Region (2018)



COVID Deaths in U.S. by Age & Race



This is a common technique called stratification used to explore data by sorting into some distinct groups.

PROS

- Easy to interpret and communicate across range of technical backgrounds.
- Critical for understanding diverse populations.
- Applicable for almost any type of dataset.

CONS

- Limited to 1 or 2 Features at a time. Beyond that becomes obtuse.
  - "Which combination of features result in a sub-population with the most anomalous outcomes?"
- Relies on human intuition for choice of Features. No inherent 'Discovery'.

Stratification is essentially looking for systematic deviations in data manually.

We should be able to do in a more automated way for more combinations of features.

Systematic change to data

>>>>>>>>>>>>

Just stirring and hoping

Conventional model-centric approach:

AI = Code + Data

(algorithm/model)

Work on this

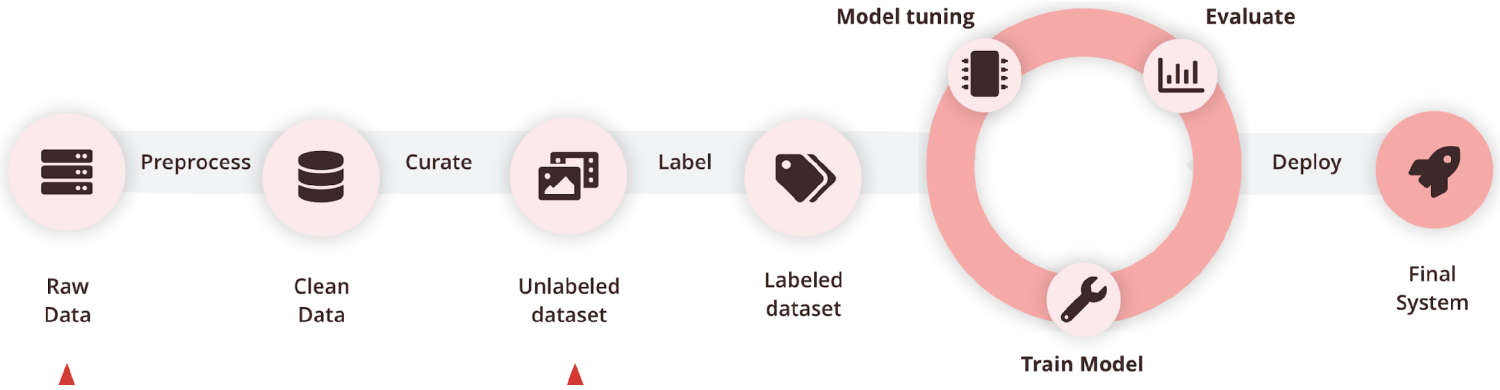
Data-centric approach:

AI = Code + Data

(algorithm/model)

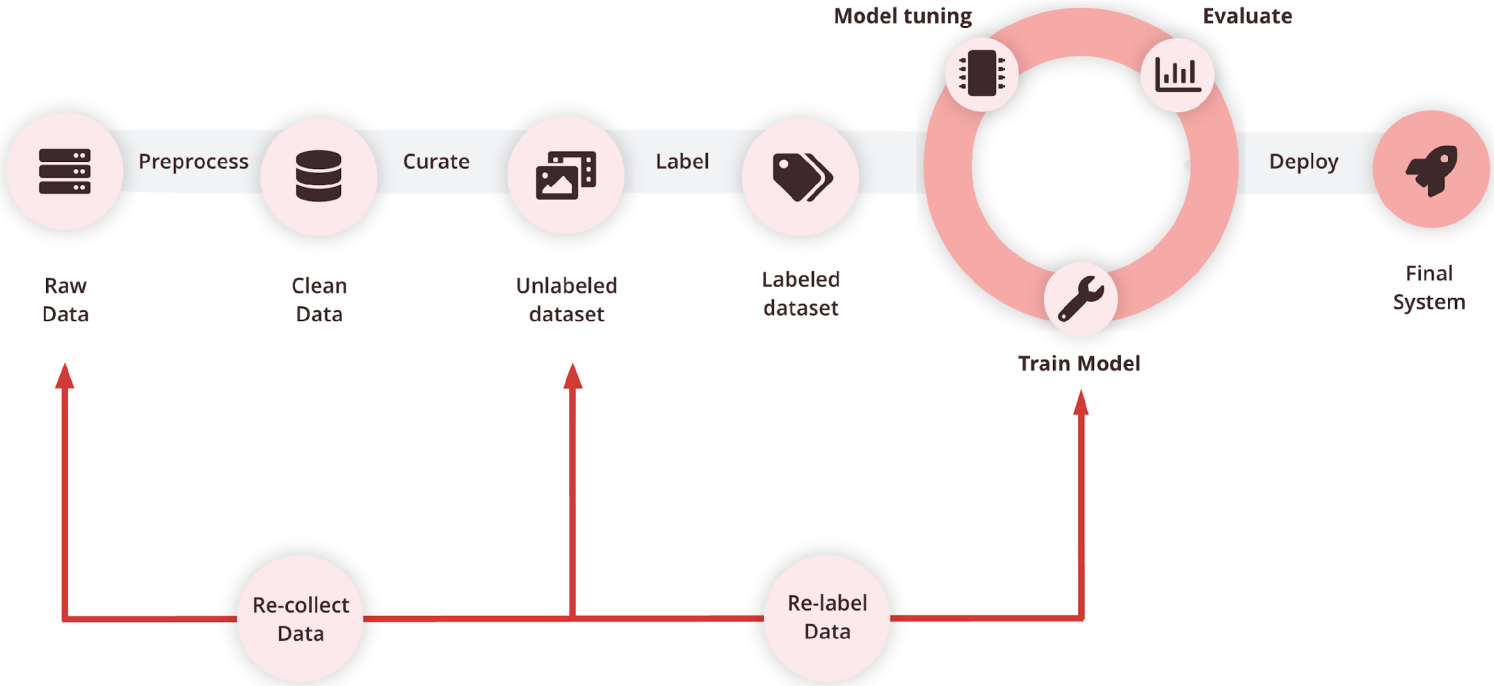
Work on this

Typical Model Centric-AI Flowchart



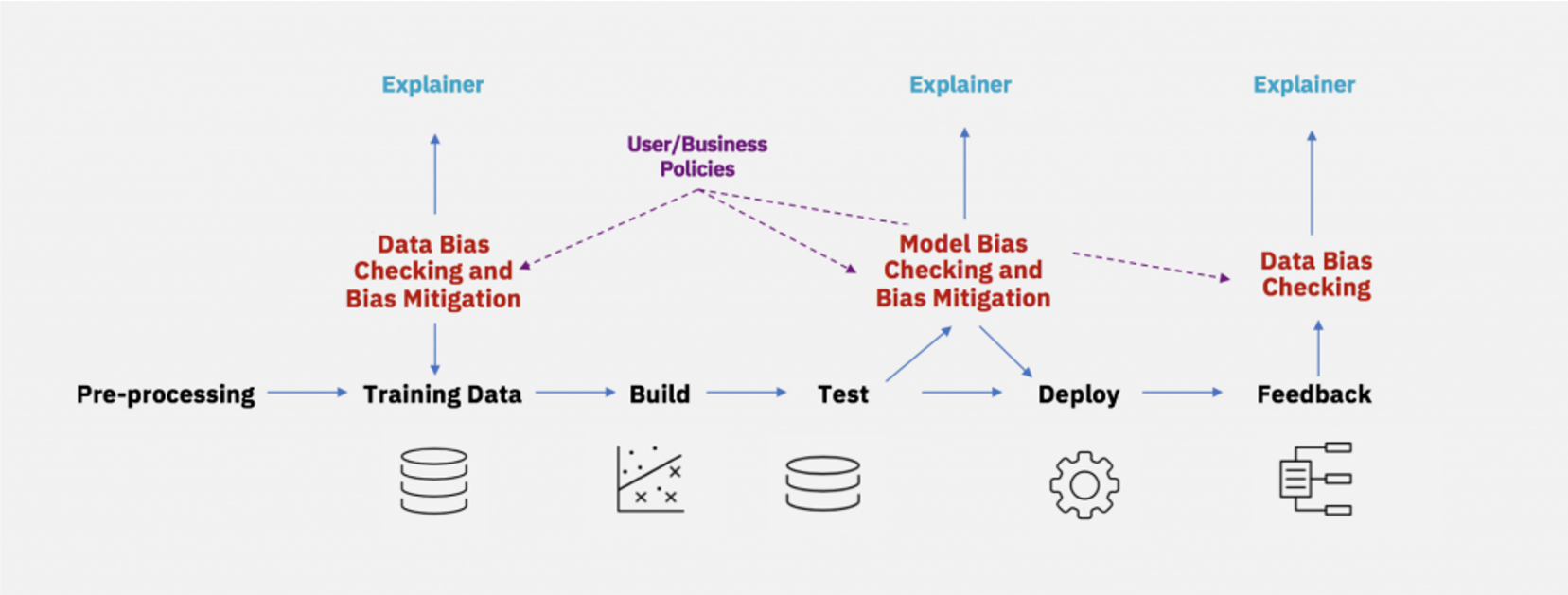
"Data Centric AI." *Mindkosh AI*, https://mindkosh.com/data-centric-ai/. Accessed 6 May 2022. s

Typical Data Centric-AI Flowchart

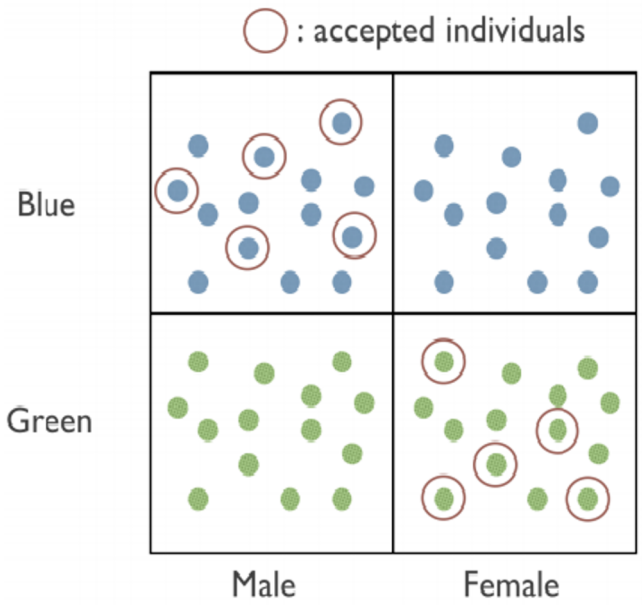


We need to know what data to recollect / relabel data?
This also involves detecting systematic deviations in data.

Mitigating bias throughout the AI lifecycle. Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." arXiv preprint arXiv:1810.01943 (2018).



Bias checking/detection is a critical component of ensuring AI fairness



Toy Example, Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. ***Michael Kearns, Seth Neel, Aaron Roth, Zhiwei Steven Wu***

Individual fairness ☑

Group fairness ☑

Subgroup fairness ✗

This phenomenon is called fairness gerrymandering.

This model is unfair to green males and blue females.

Easy to spot when you know it exists.

Extremely difficult to discover as you have exponentially many subgroups to look at.

For a dataset of 10 features with just 3 categories in each feature, you have about 300 million subgroups.

There needs to be an automated, systematic way to approach this.

Theory

Key idea is the use of anomalous detection to discover systematic deviations

Inspect individual pieces of hay at a time, ***O*(*n*).** Return **individual** anomalous records. **Computers are capable of this.**

Inspect ***O*(2^n)** combinations of hay pieces. Return the **subset** that includes fake hay. **Computers cannot do this.** This is the key limitation.

The key solution to this limitation is **linear time subset scanning (LTSS).**

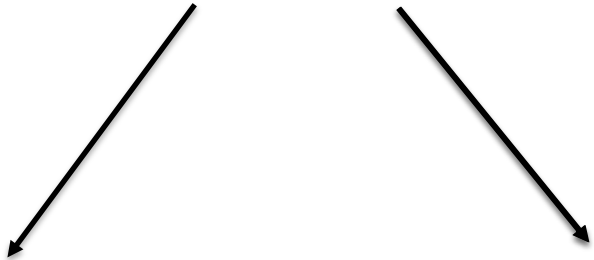
LTSS uses a scoring function and a priority function to reduce the computation time from exponential to linear.

This is the basis of **multi-dimensional subset scanning (MDSS).**

We define “bias” as the amount of divergence between the **predicted outcomes** and the **true outcomes** for a **subset** of records and efficiently maximize this function over **all self-similar subsets** of records.

The scanner returns a bias score and the discovered anomalous subset.

Detecting, Characterizing, and Leveraging Systematic Bias in Data & Models



AutoStrat

No model required

Bias Scan

Requires a predictive model

Very little difference in the code and theory

AutoStrat ≈ Bias Scan

- In AutoStrat, the predicted outcomes is constant for all data records
- AutoStrat can be viewed as the predictive bias of a simple predictor e.g., $P(Y_i) = \bar{Y}$
- AutoStrat is used for detecting systematic deviations in data
- Bias Scan replaces constant predicted outcomes with a black-box predictor $P(Y_i) = F(\boldsymbol{X})$
- Emphasis changes from exploring data to inspecting a model for mistakes
  - Predictive models have systematic deviations for a wide range of reasons
    - Data shifts between train and test
    - Protected features
    - Changes over time

- We have two types of scoring functions – parametric and non-parametric.
- Parametric scoring functions assumes the predicted and actual outcomes follow a specific distribution
- Non-parametric scoring functions makes no assumptions about the distribution of the predicted and actual outcomes
- MDSS is currently implemented on IBM's open-source AI Fairness Toolkit – AIF360.
- There is support for the following scoring functions:
  - Parametric:
    - Bernoulli (Binary and nominal classification tasks)
    - Gaussian (All regression tasks)
    - Poisson (Count regression tasks)
  - Non-Parametric:
    - BerkJones (All tasks but only in AutoStrat mode)

- The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

- The modules available in the toolkit includes:
  - Algorithms (for bias mitigation)
    - aif360.algorithms.preprocessing
    - aif360.algorithms.inprocessing
    - aif360.algorithms.postprocessing
  - Detectors (for bias detection)
    - aif360.detectors.mdss_detector
  - Explainers (for model explainabilty)
    - aif360.explainers
  - Fairness Metrics (for model fairness ranking)
    - aif360.metrics

Ongoing works on MDSS

- Addition of other parametric and non-parametric scoring functions
- Direct support for continuous features
  - Currently, continuous features need to be binned
- Support for non-parametric scoring functions in bias scan mode
- Application to new datasets and domains

Examples

Neonatal Mortality in Sub-Saharan Africa (AMANHI)

Mother's Age	Birth Weight (gms)	Mother's Education	Delivery Location	Birth Quarter	Delivery Person	Gestational Age (days)	Birth Year	Birth Weight for Age (Z score)	Mortality
25_to_29	3000_to_3500	0	Hospital	Q3	Midwife	LTE_270	2012	between_1_2	0
20_to_25	2500_to_3000	10_and_above	Home	Q1	Relative/Friend	between_280_290	2012	between_-2_-1	0
25_to_29	3000_to_3500	0	Home	Q4	Relative/Friend	between_270_280	2011	between_0_1	0
25_to_29	2500_to_3000	0	Home	Q2	Relative/Friend	LTE_270	2012	between_1_2	0
30_to_39	3000_to_3500	0	Home	Q4	Relative/Friend	LTE_270	2012	GT_2	0
40_and_above	2500_to_3000	0	Hospital	Q3	Midwife	LTE_270	2011	between_-1_0	0
LTE19	2500_to_3000	10_and_above	Home	Q4	Missing/Unknown	LTE_270	2011	GT_2	0
20_to_25	3000_to_3500	4_to_6	Home	Q2	Relative/Friend	between_270_280	2012	between_-1_0	0

20,000+ records and 1.5% of births have mortality event

Scanning wants to know which **subset of births** have anomalously high number of mortality events.

There's over 4 Trillion subsets to consider!

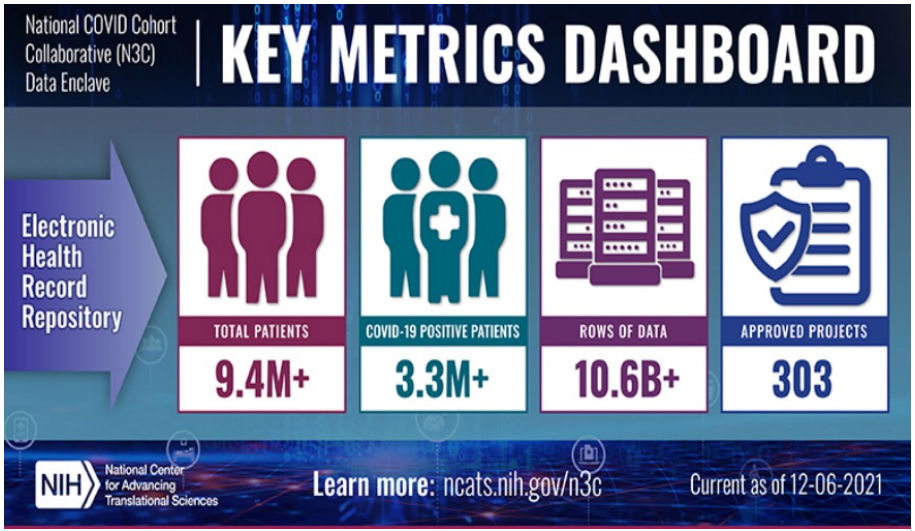
Births Subset:
Home Delivery
and
Medical Professional
(Doctor or Midwife present)

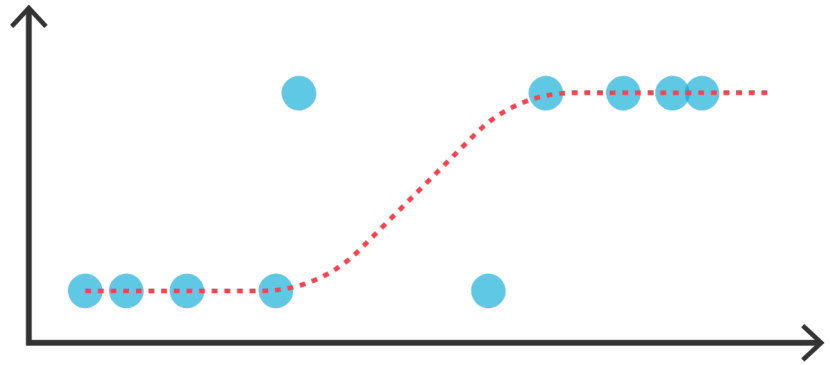
Neonatal Mortality Rate of this Subset
42.1%

Recall the average was 1.5% -- this group is very anomalous!

These circumstances account for **nearly half** (49.1%) of all neonatal deaths in the Ghana study.

Covid Mortality in the US from Electronic Medical Records





https://covid.cd2h.org/dashboard/cohort

Covid Mortality in the US from Electronic Medical Records

Where is our predictive model the most biased?

<u>Cancer Patients Under the Age of 50</u>

1468 patients

LR Model predicted 80 deaths

<<

Data shows 195 deaths

The predictive model has failed to capture a complex interaction between age, cancer, and mortality.

MADIVA

MULTIMORBIDITY IN AFRICA

Digital Innovation, Visualisation & Application Research Hub

NIH grant number (1U54TW012077-01)

NIH National Institutes of Health
Turning Discovery Into Health

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG 100 1922 2022

20 APHRC African Population and Health Research Center

NEWS RELEASES

Tuesday, October 26, 2021

NIH awards nearly $75M to catalyze data science research in Africa

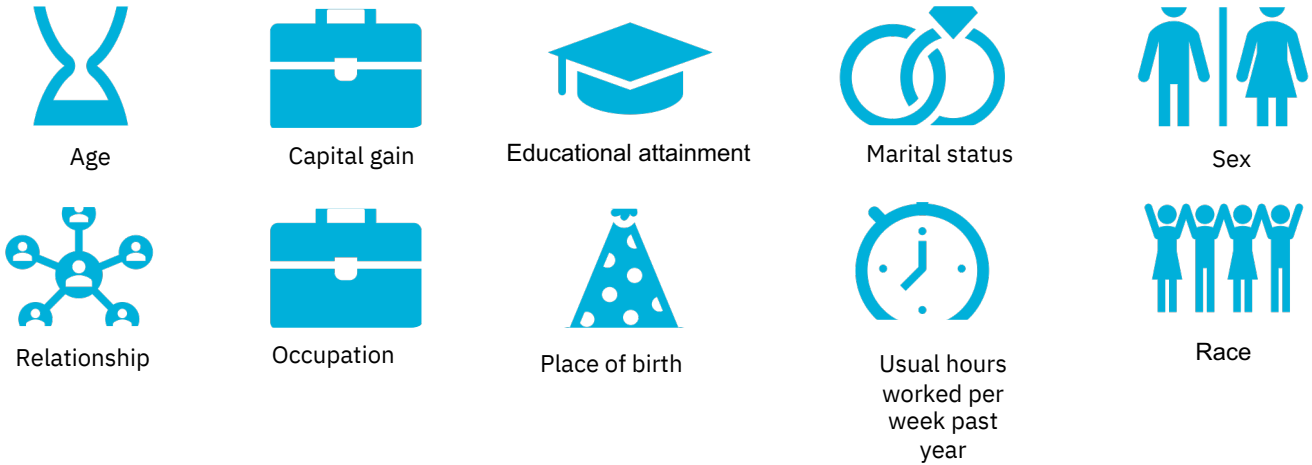
New program will establish data science research and training network across the continent.

CODE EXAMPLE: US CENSUS DATA

Dataset:

- 1994 US Census Data on income accessed from the UCI Adult Dataset Repository. This is a generic dataset that facilitates the benchmarking of machine learning algorithms.

- Contains ~16K samples with the following features:



CODE EXAMPLE: US CENSUS DATA

Key Question:

Can we identify **sub-populations** who, as a **subgroup**, have outcomes that significantly deviate from the **overall population?**

TARGET/OUTCOMES:

Y=1 → INCOME > 50K

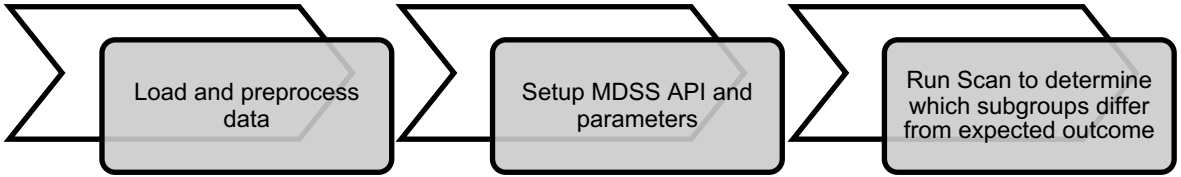
Y=0 → INCOME < 50K

Overall population = mean of the observed outcomes

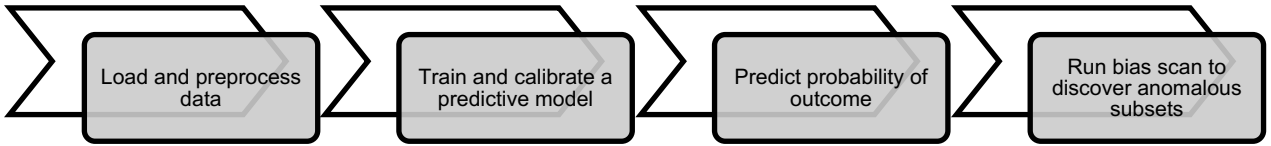
OR

Predictions from a trained model (bias scan)

CODE EXAMPLE: SCANNING PIPELINE



AUTOSTRAT MODE (no expectations provided)



BIAS SCAN (probabilities obtained from the model)

CODE EXAMPLE: US CENSUS DATA

Key Question:

Can we identify **sub-populations** who, as a subgroup, have outcomes that significantly deviate from the **overall population?**

TARGET/OUTCOMES:

Y=1 → INCOME > 50K

Y=0 → INCOME < 50K

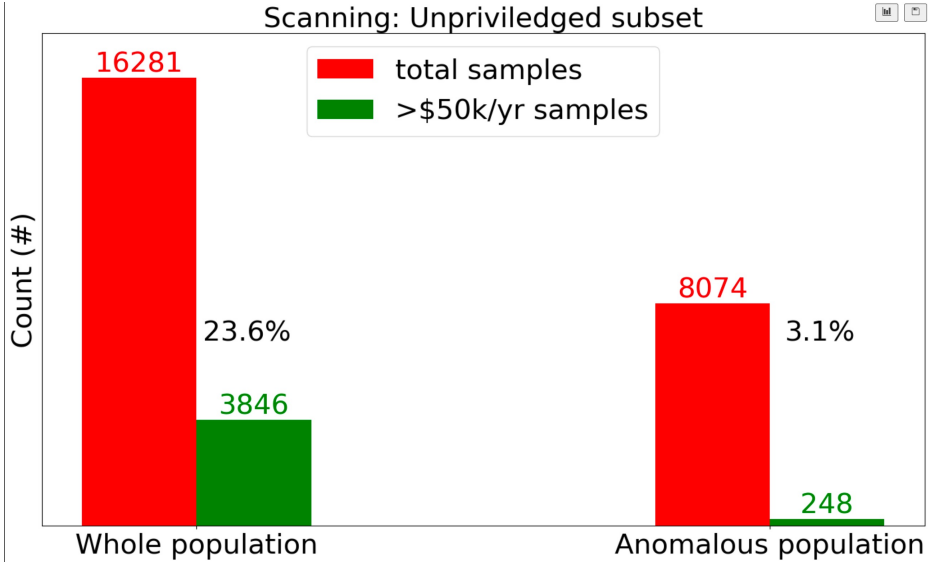
Overall population = mean of the observed outcomes

OR

Predictions from a trained model (bias scan)

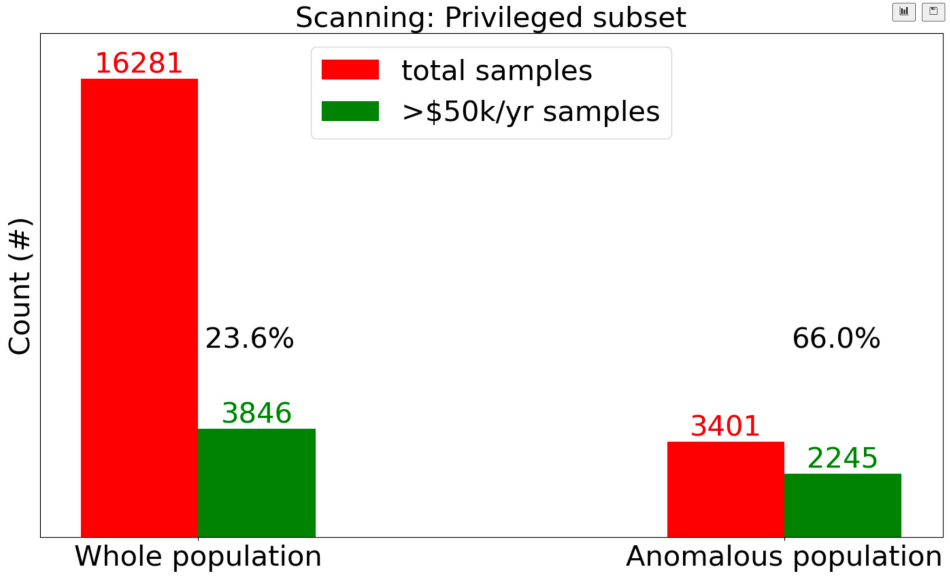
CODE EXAMPLE: OUTCOMES

- Detected subgroup has a size of 8074 out of 16K samples
- The observed probability of those earning >50K is 3.1% whereas the population mean is 24%
- This indicates a multiplicative decrease in the odds by 9.7



CODE EXAMPLE: OUTCOMES

- Detected subgroup has a size of 3401 out of 16K samples
- The observed probability of those earning >50K is66% whereas the population mean is 24%
- This indicates a multiplicative increase in the odds by 6.3



Code

Scan the QR to access the git repository

IBM Academic Initiative

The **IBM Academic Initiative** enables students and faculty members to access IBM education resources free of charge through a self-service portal.

Get for free:

- **IBM Cloud** and a feature code to upgrade your account
- Commercial grade IBM **software**
- e-learning courses & certification

Go to http://*ibm.com/academic* *and sign up using your university email address*

Our Contact details:

Adebayo Oshingbesan
Email: *adebayo.oshingbesan1@ibm.com*
LinkedIn: *Adebayo Oshingbesan*

Tanya Akumu
Email: *tanya.akumu@ibm.com*
LinkedIn: Tanya Akumu
Twitter: @*tanya-akumu*

https://ibm.biz/BdPxCR

Q&A