Mitigating Bias in Cybersecurity Models

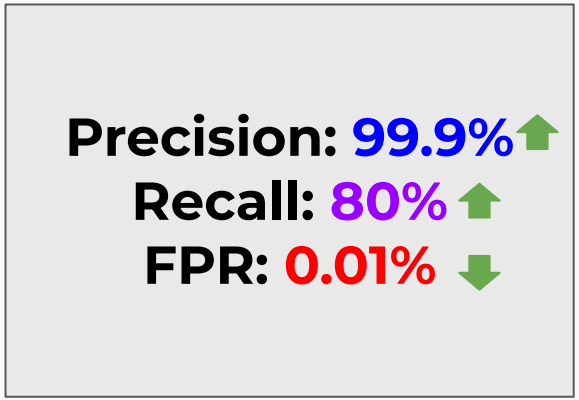
Presented by Mohamed Nabeel

08/29/2024

Gym

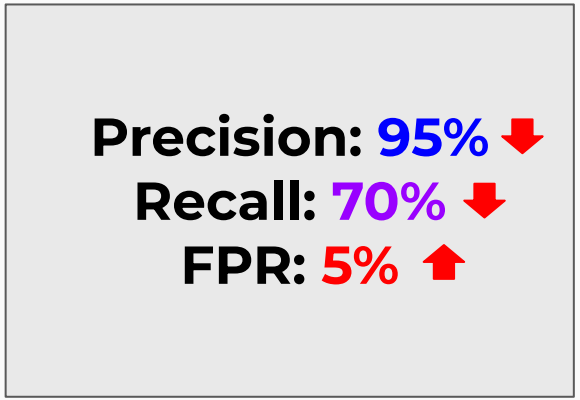
Marathon

Lab Environment

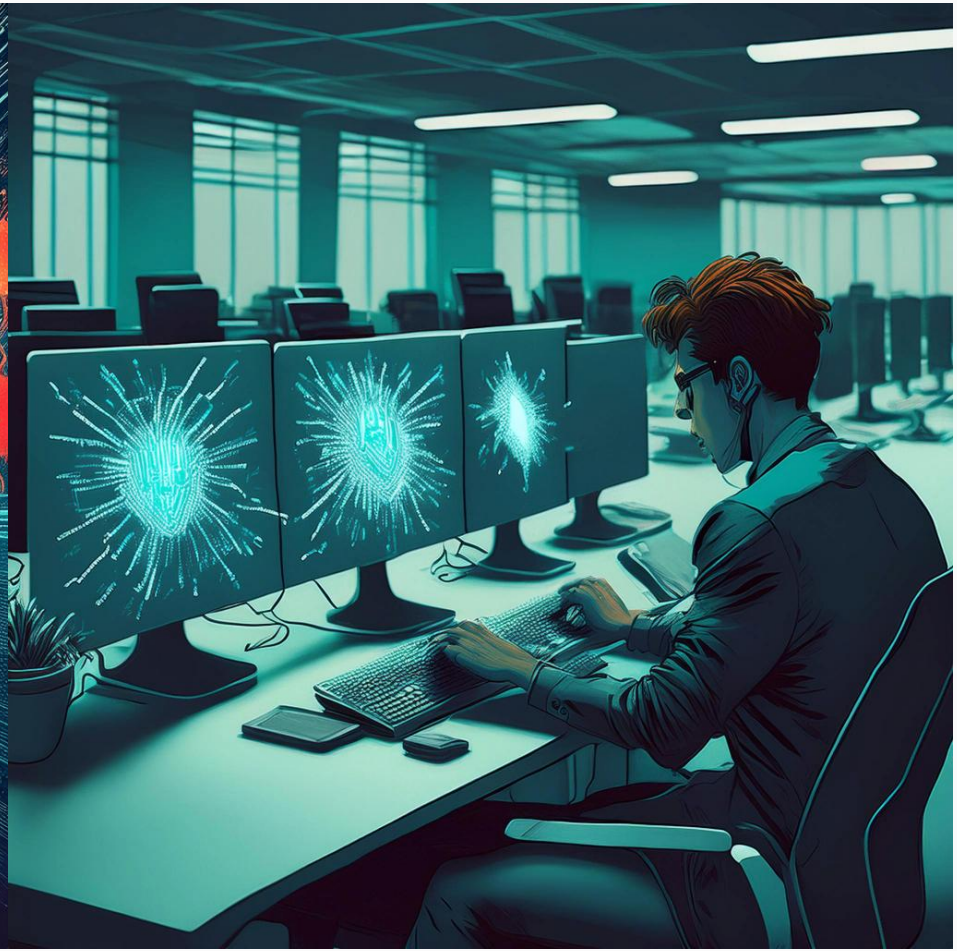


Model Testing Metrics

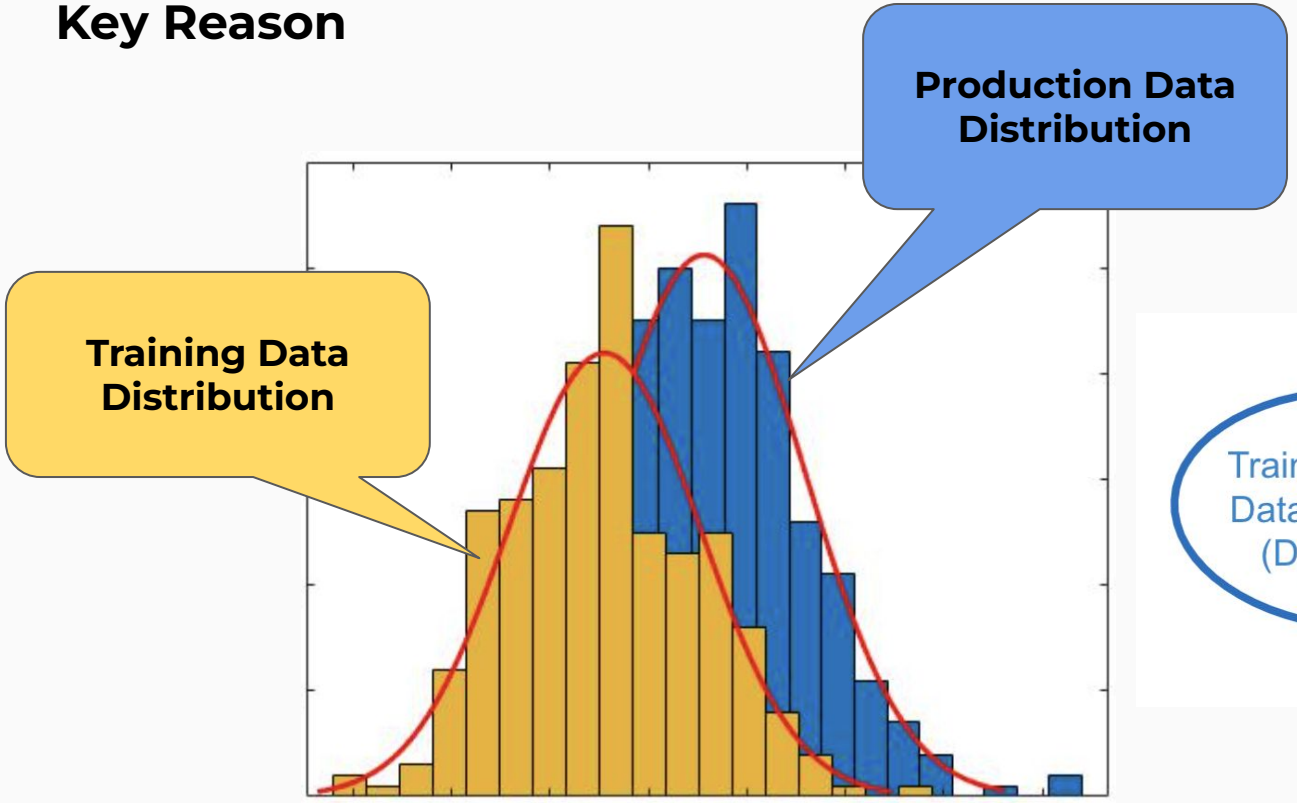
Prod Environment

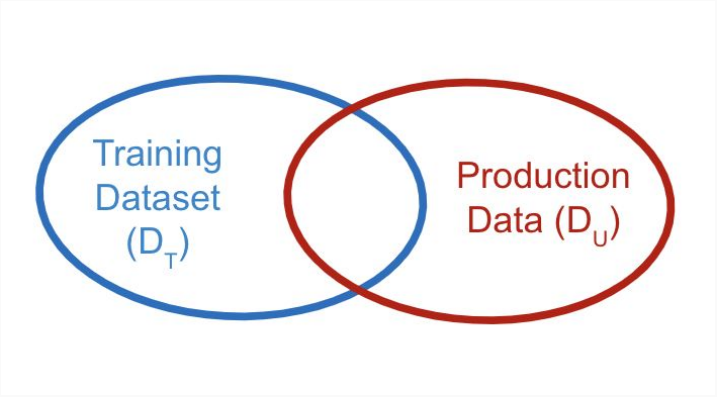


Model Deployment Metrics



Key Reason


Production Data Distribution
Training Data Distribution

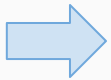

Training Dataset (D_T)
Production Data (D_U)

DATA

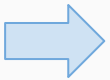
ML/DL ALGO

ML/DL MODEL

~~Garbage In~~
Biased Data



ML/DL Model



~~Garbage Out~~
Biased Prediction

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

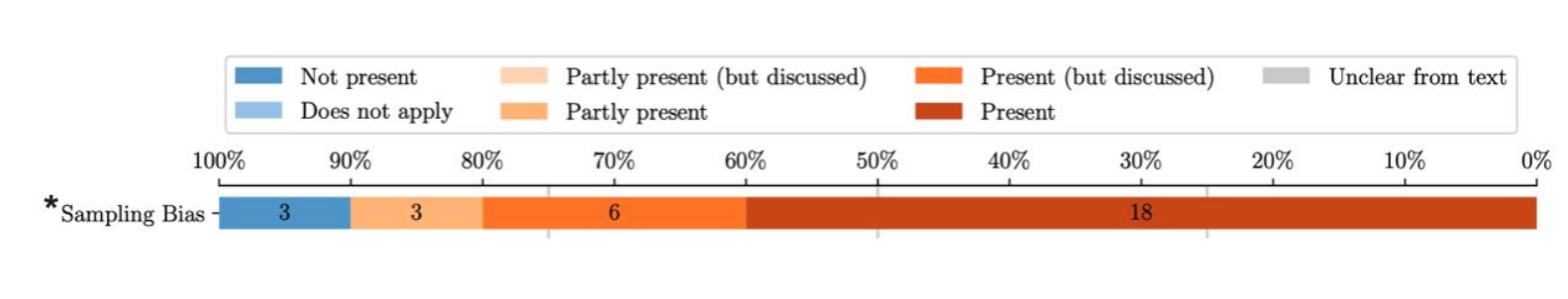
Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

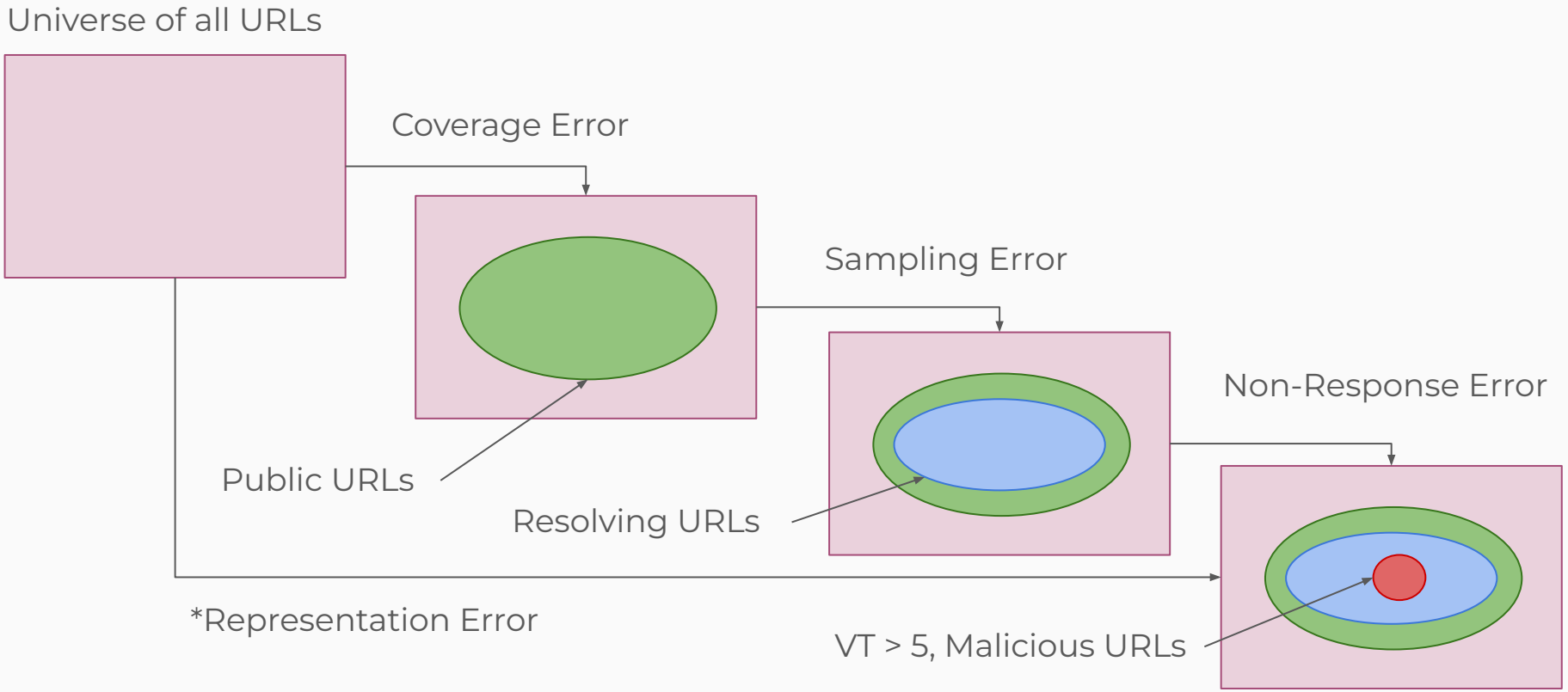
Sampling Bias in Cyber Security



90% has Sampling Bias

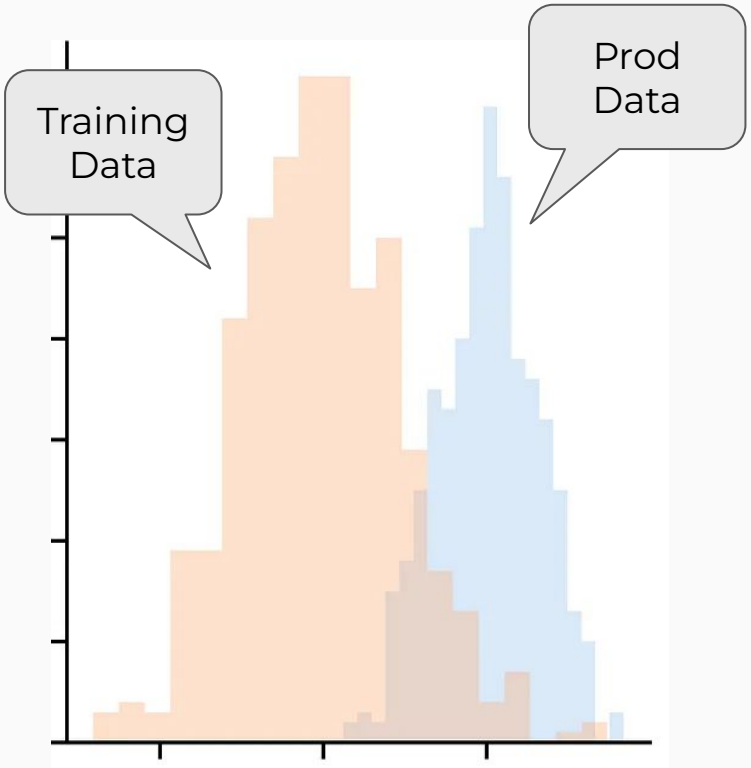
* Daniel, et al., Dos and Don'ts of ML in Computer Security, Usenix Security 2022

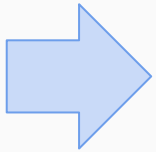
Sampling Bias - Example on URL Classification

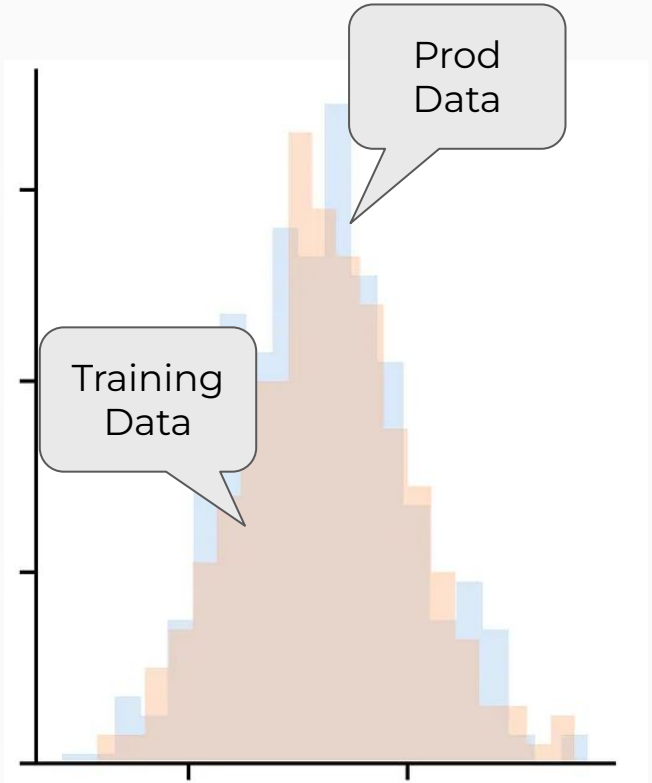


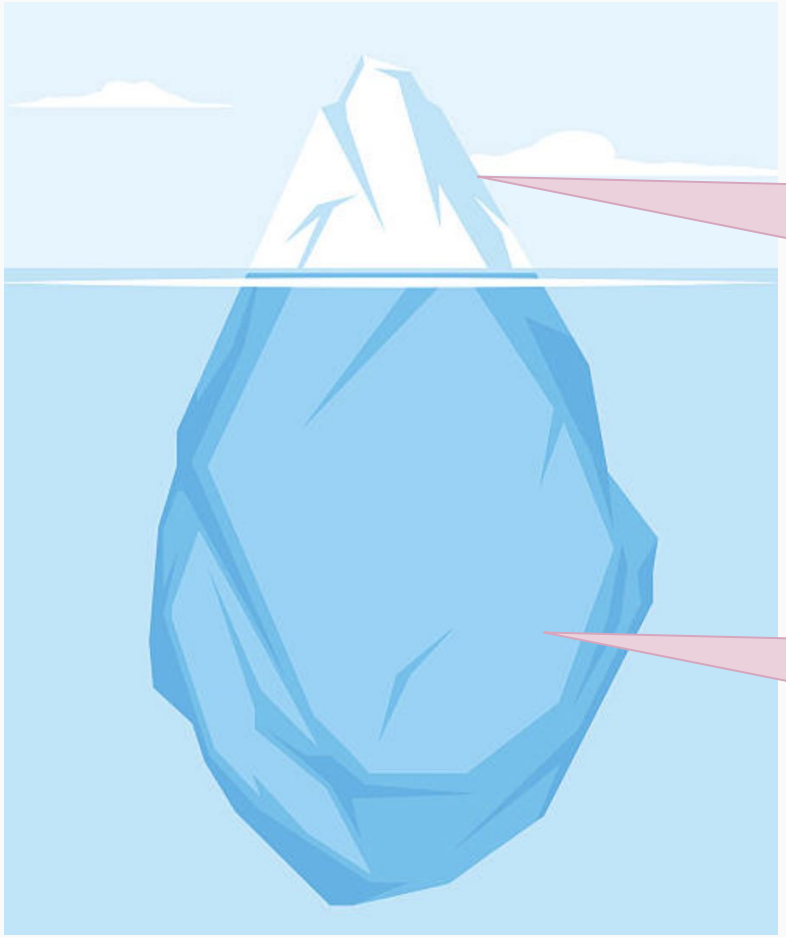
* Robert M Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public opinion quarterly*, 74(5):849–879, 2010.

How to Mitigate Sampling Bias?


Training Data
Prod Data




Prod Data
Training Data



Labeled
Data

Unlabeled
Data

Label production data to mitigate bias!

Cyber Labeling is Expensive!

Labeling Cats and Dogs

VS.

Labeling Malware

Labeled Data

What if we can use Unlabeled data?

Unlabeled Data

Self Supervised Learning

Labeled Data

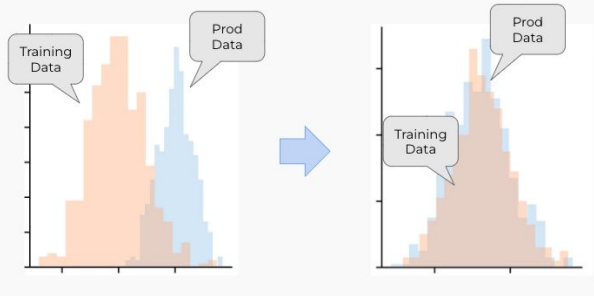
Unlabeled Data

A few labeled data

Many **Pseudo labels**

ML/DL Algorithm

Unbiased Classifier



Summary*

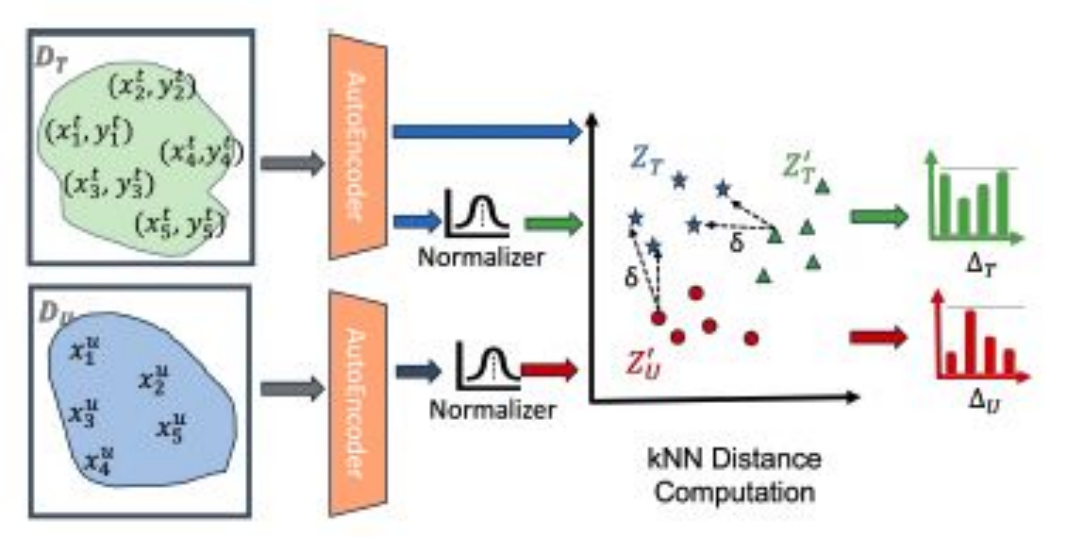
* More information in Detecting and Mitigating Sampling Bias in Cybersecurity with Unlabeled Data, Usenix Security 2024

Summary

- Next time when you evaluate a classifier, don’t forget to ask
  - What data is used to train the model?
  - How is sampling bias is addressed?
- We have a lot of unlabeled data
- Use this bias mitigation algorithm to debias the data to your classifiers
- More information in Detecting and Mitigating Sampling Bias in Cybersecurity with Unlabeled Data, Usenix Security 2024

Tackling Sample Bias

Bias Detection



Bias Mitigation