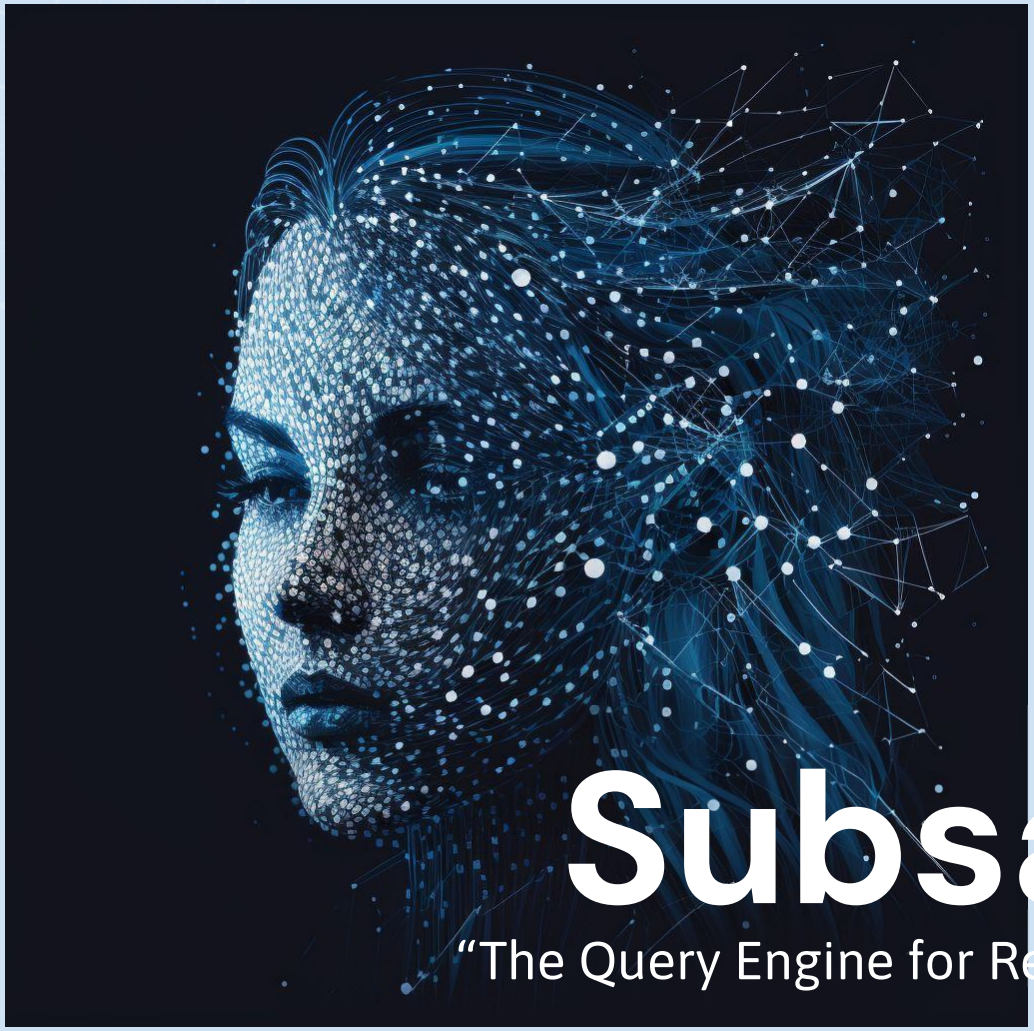
Simrun Sharma
Antara Bhide
Minling Zhou



Subsalt

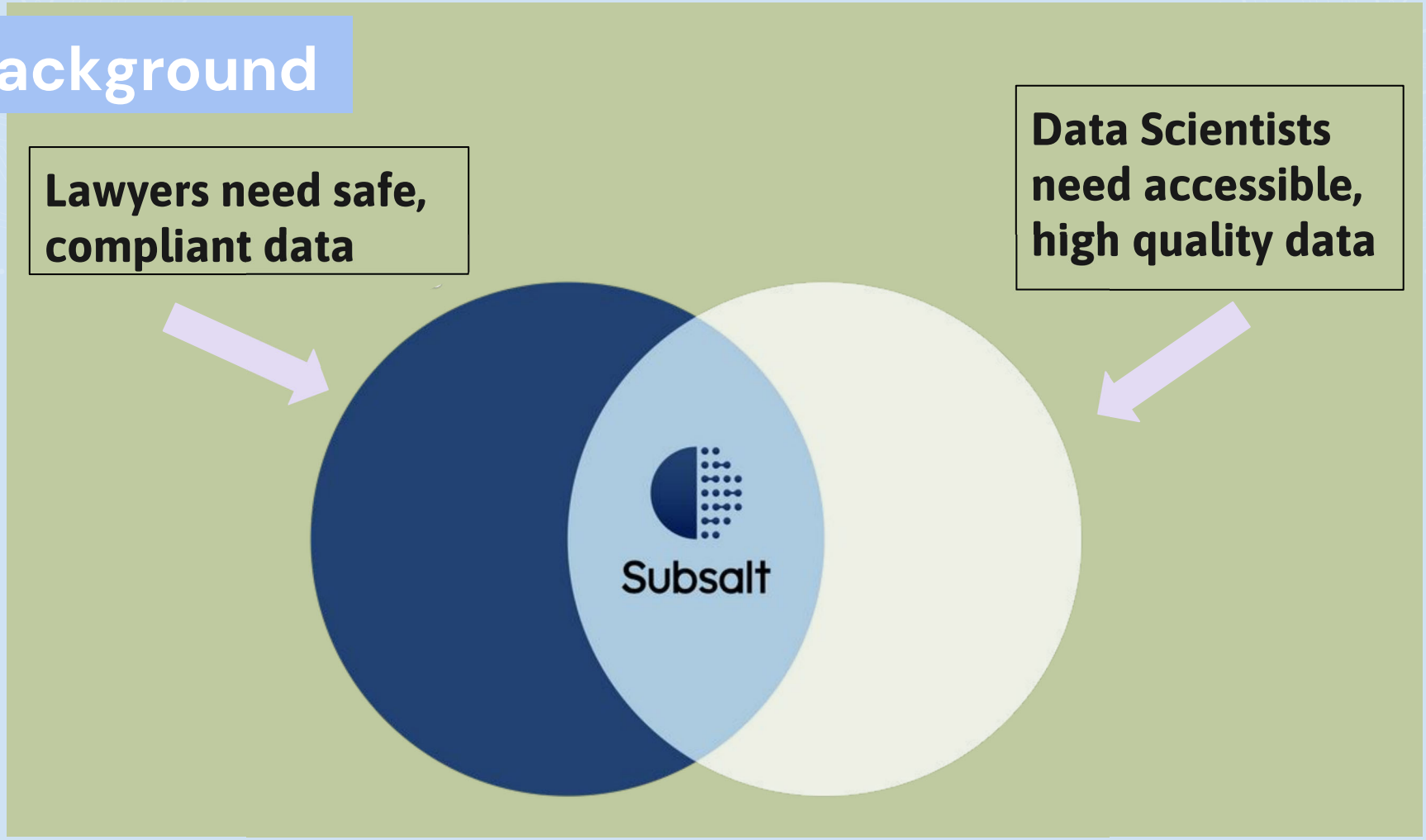
"The Query Engine for Regulated Data"

What is Subsalt?

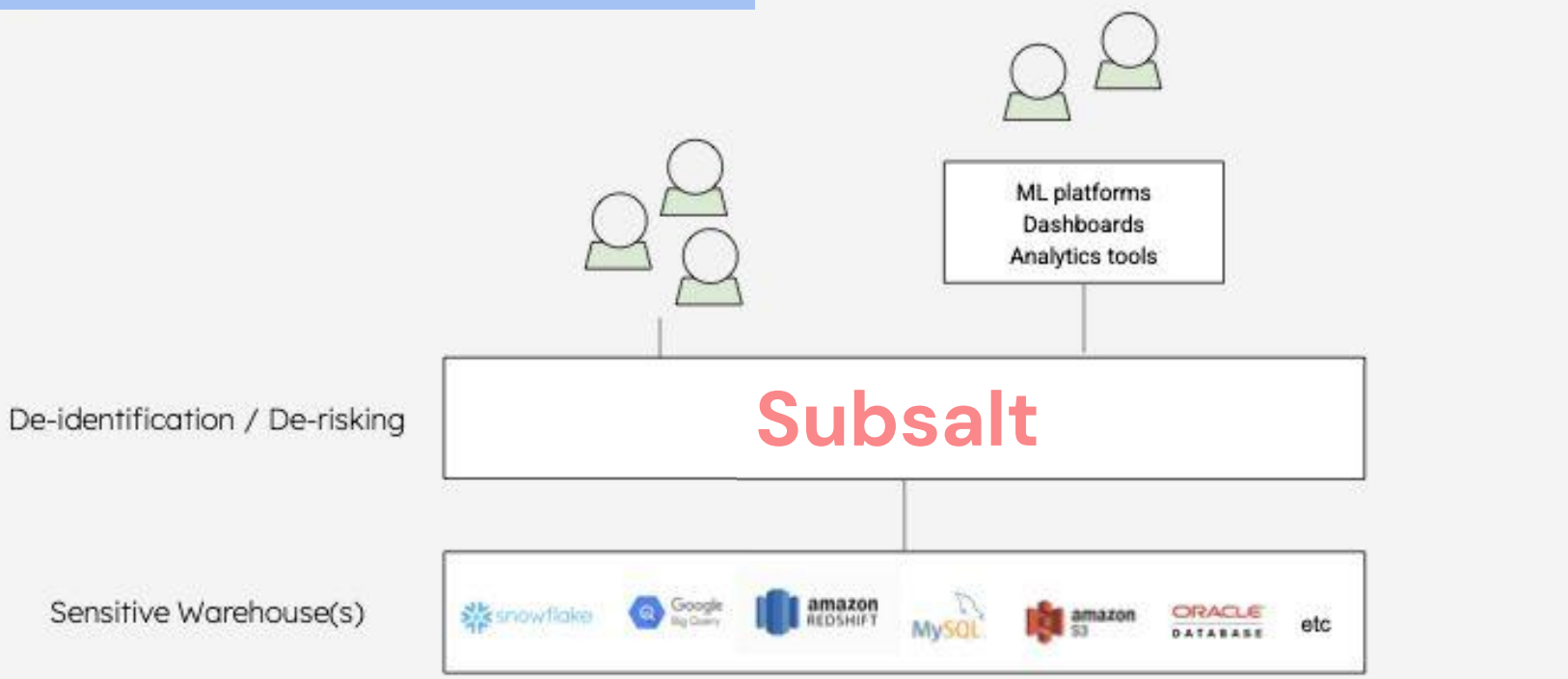
SIMRUN

Data is the new oil. It’s valuable, but if unrefined it cannot really be used.To create a valuable entity data be broken down.”
— Clive Humby, 2006

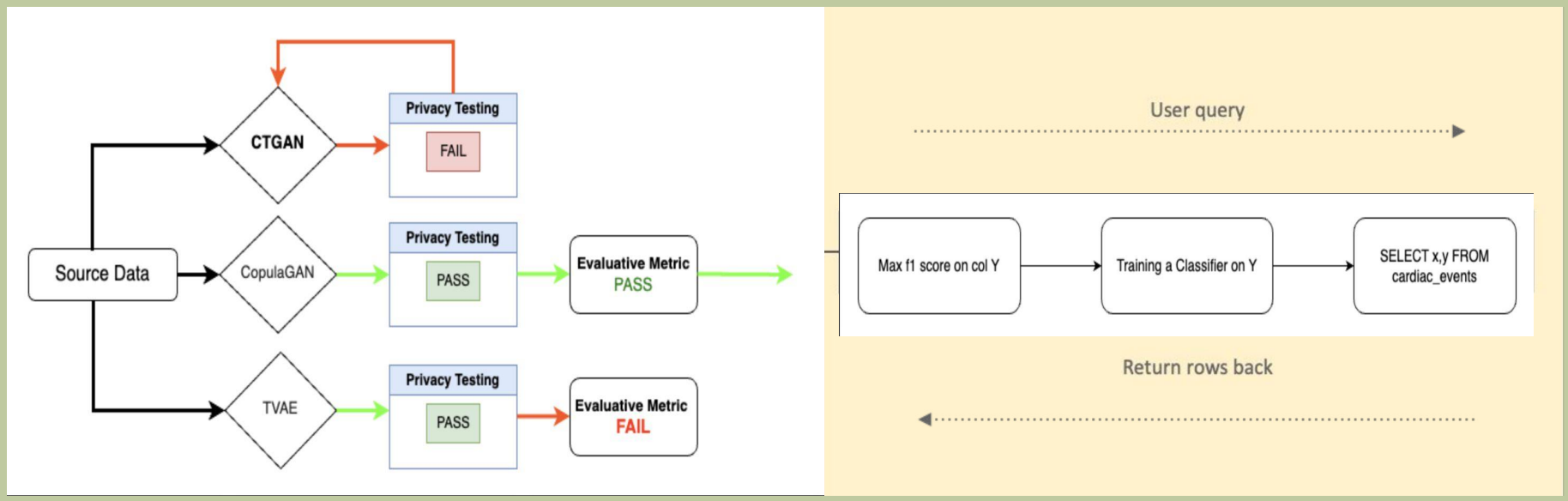
Background


Lawyers need safe, compliant data
Data Scientists need accessible, high quality data
Subsalt

Production Process



Developmental Cycle



The Problem

“High computational costs and repetitive privacy testing make large-scale data training slow and costly, reducing the product's value by hindering fast, efficient access to sensitive data.”

Privacy Testing — Risky Rows and Row Memorization

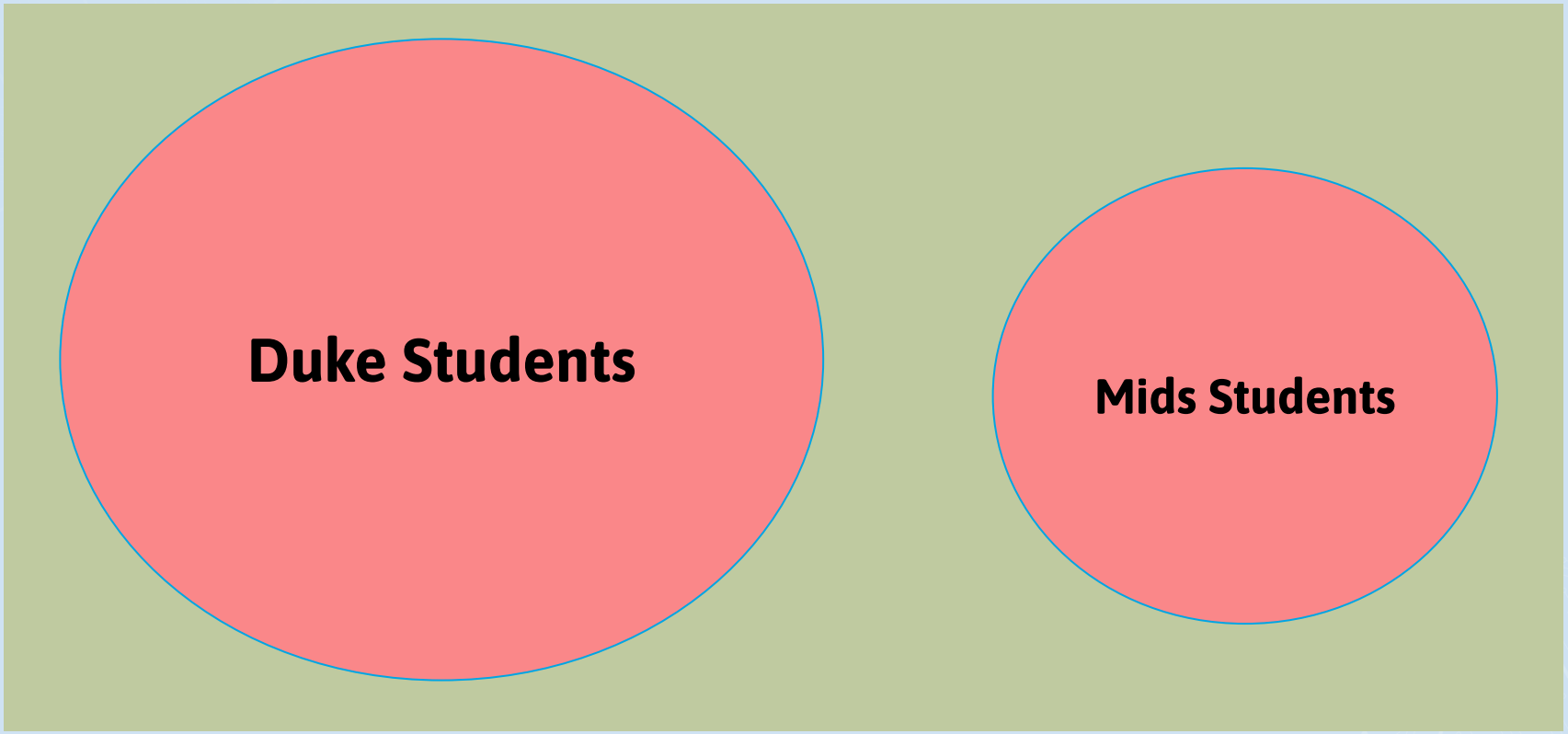
Source Data

Attendance	Internet Access	Disability	Gender
73	Yes	No	Male
98	No	No	Female
72	Yes	Yes	Male

Synthetic Data

Attendance	Internet Access	Disability	Gender
75	No	Yes	Male
92	Yes	No	Female
72	Yes	Yes	Male

Equivalence Classes



Indirect Identifiers

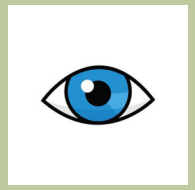
Area Of Residence

Health Conditions

Age

Occupation

Eye Color



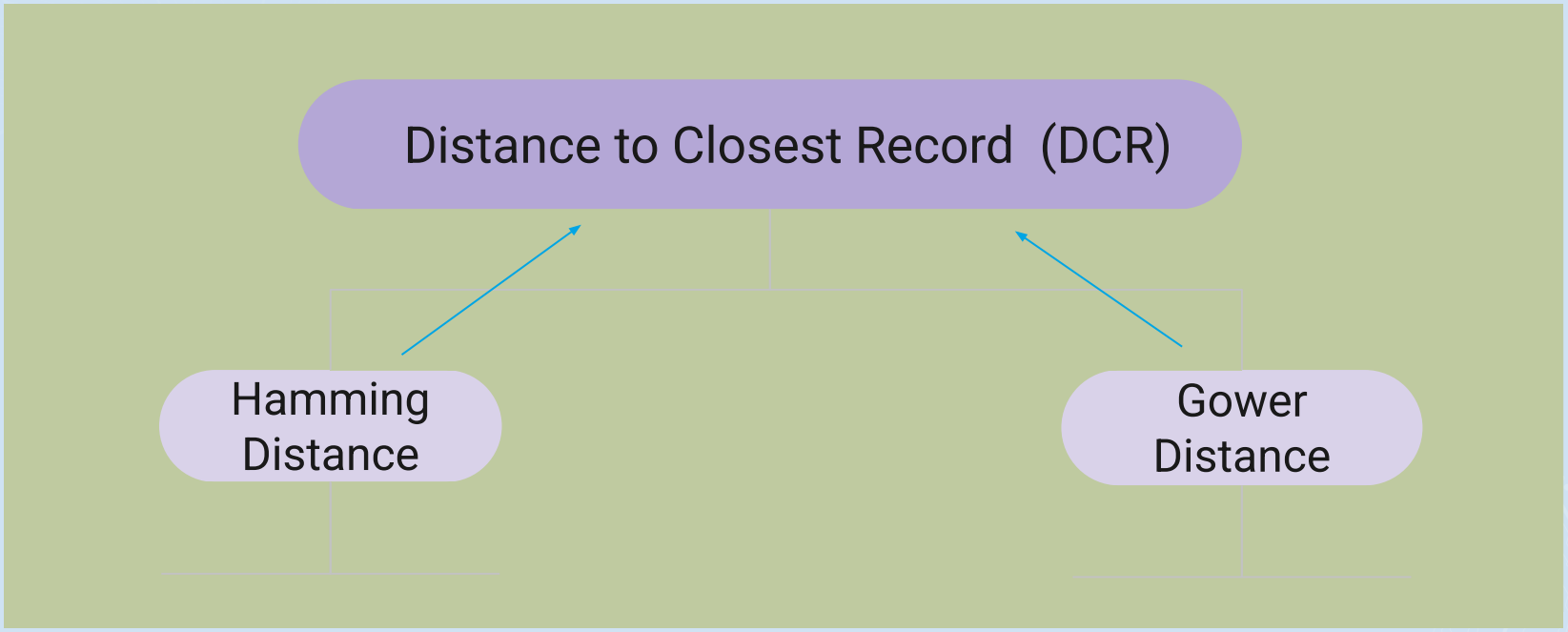
Race

Current Pitfalls

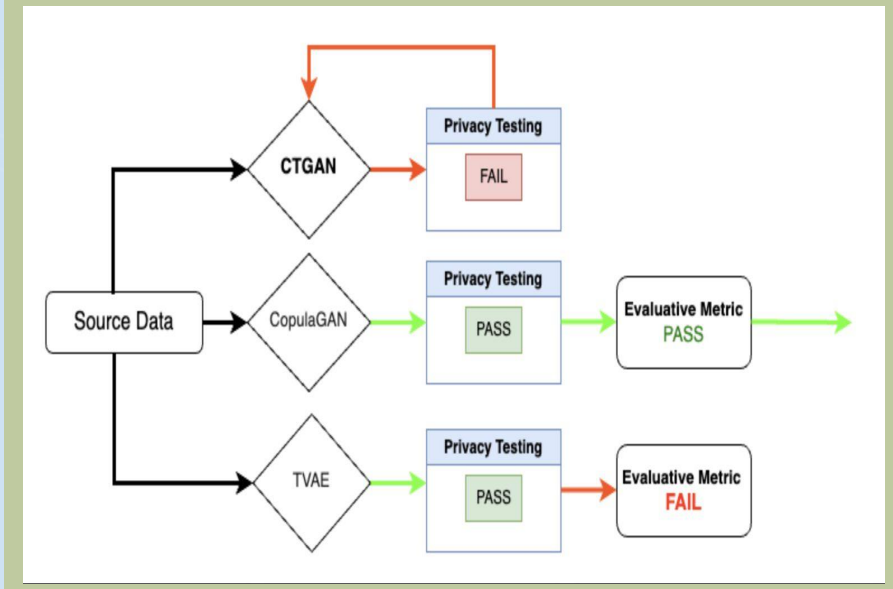
- Timings for training and privacy
- Computational costs
- Privacy tests still running post failure
- No filtering process before training
- Have to train repetitively to check privacy

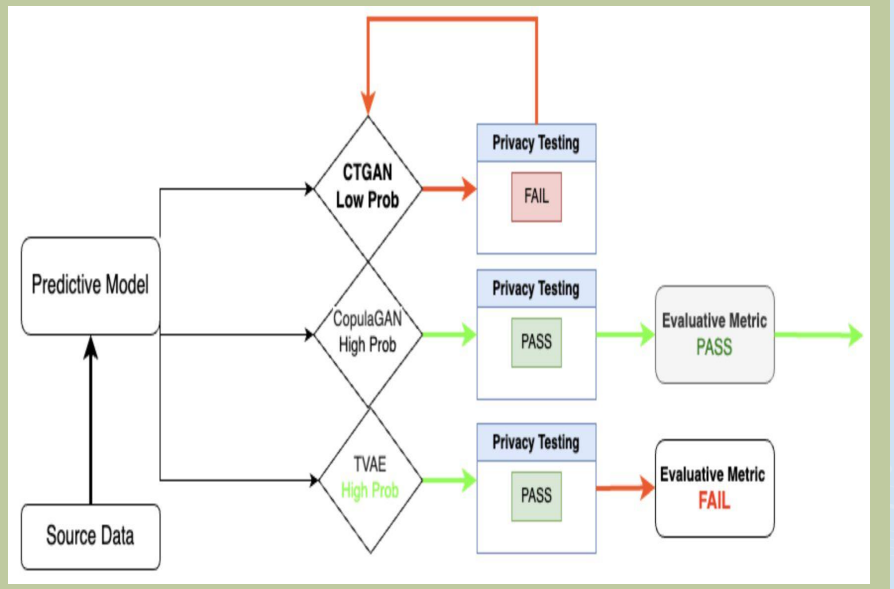
PRIVACY LAWS

Re-identification

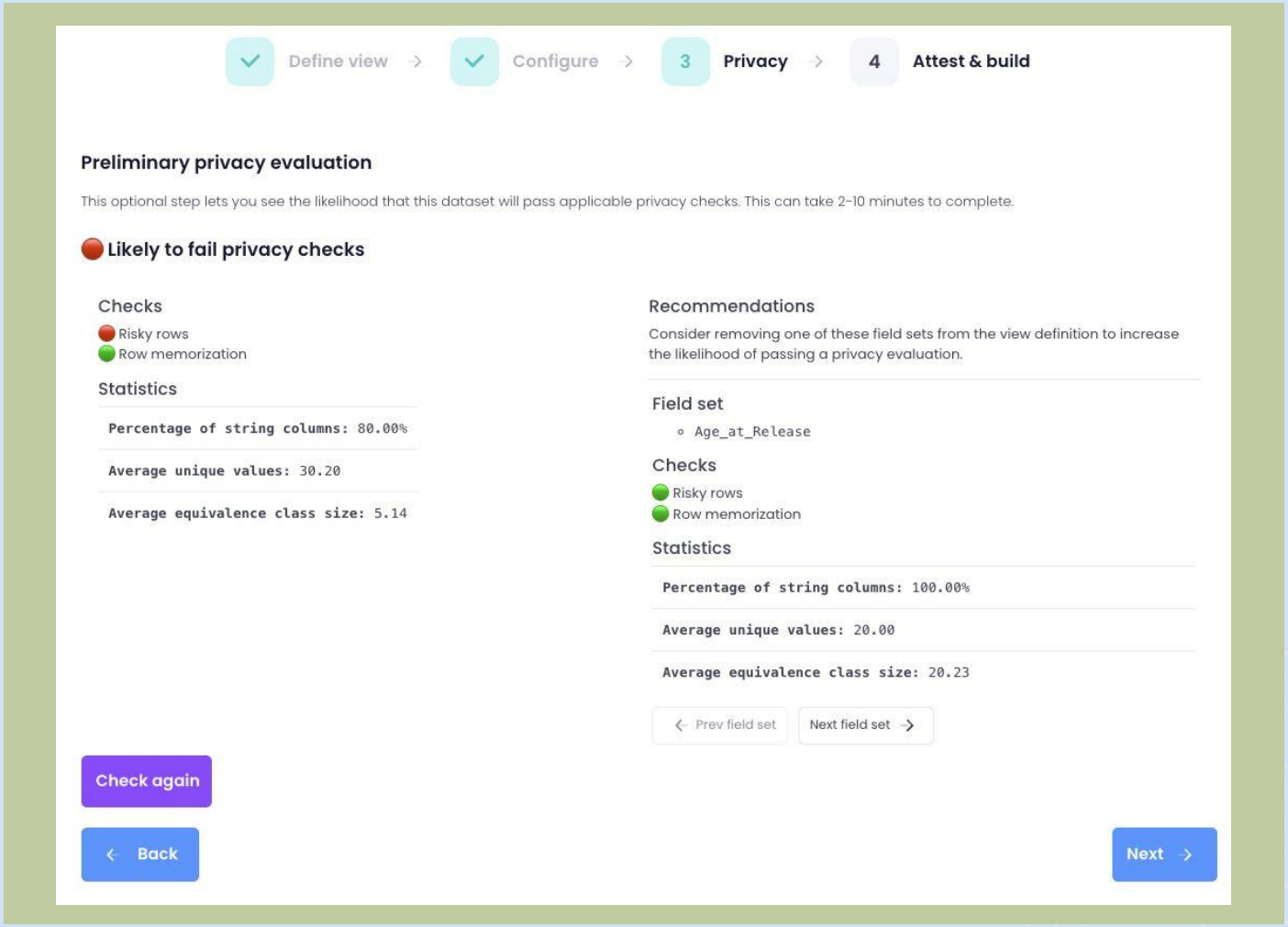


Proposed Solutions

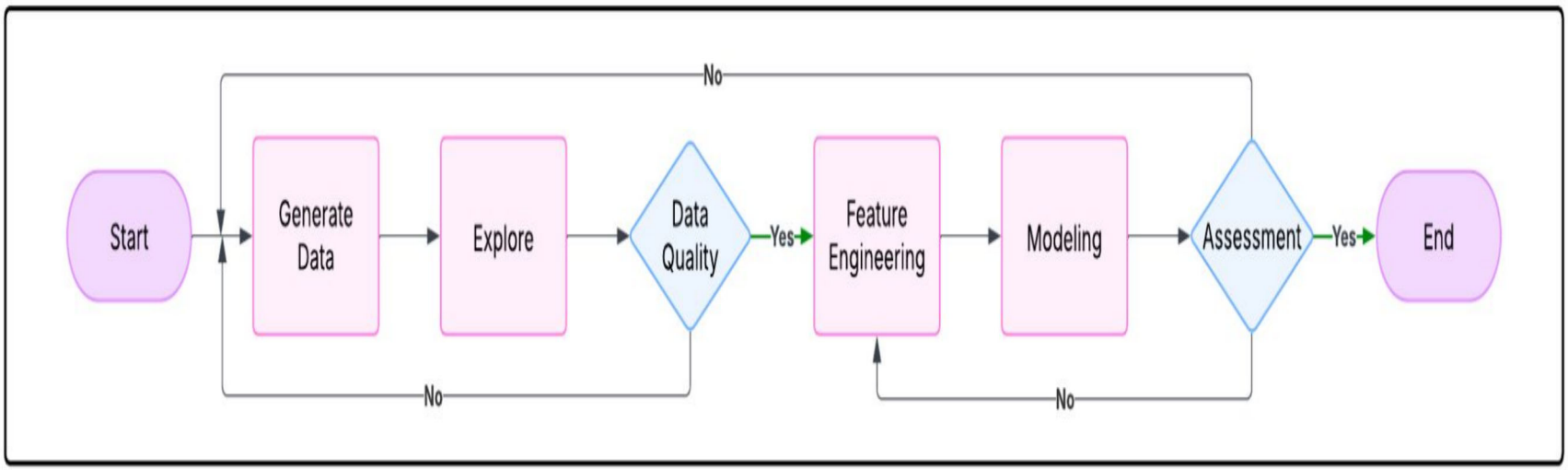




Predictive Models



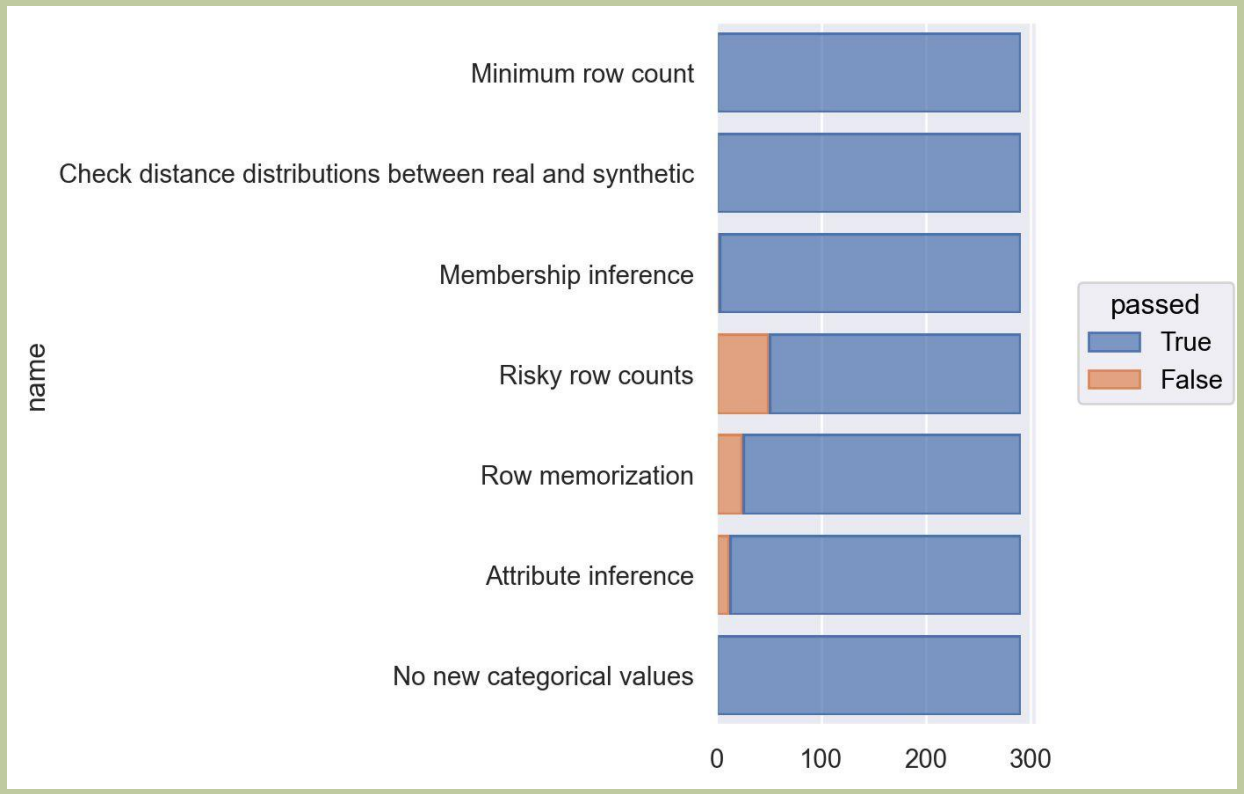
Methodology



Methodology

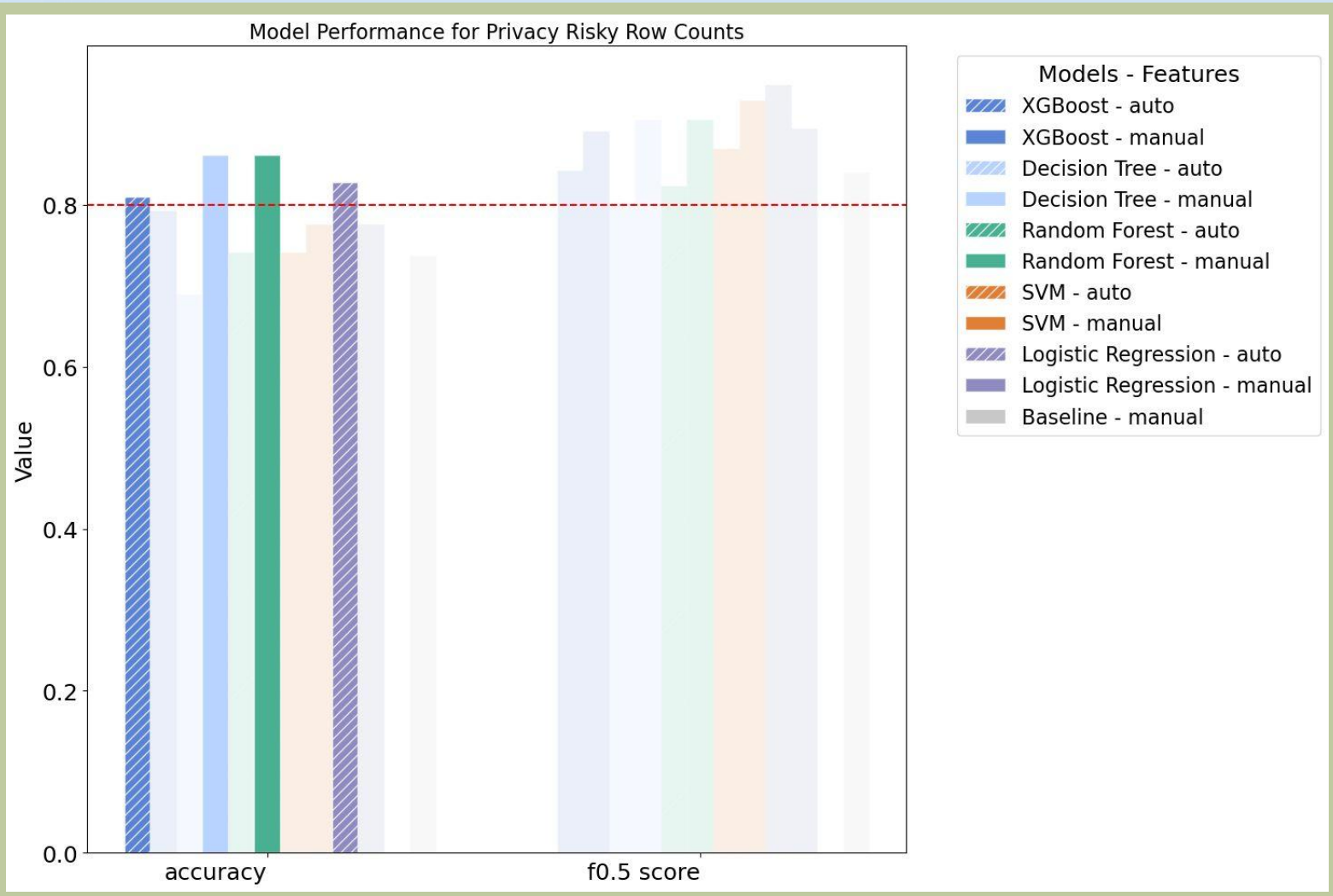
Define Problem	Feature Engineering	Model Training	Metric Selection
Binary Classification	❏ Table View Statistics ❏ Generative Model Config ❏ Equivalence Class Metrics ❏ Manual vs. Auto Feature Selection	❏ Decision Tree ❏ Random Forest ❏ XGBoost ❏ Logistic Regression ❏ SVM	❏ Balanced Accuracy ❏ Accuracy ❏ F0.5 score ❏ Specificity ❏ Sensitivity

Result – Overview

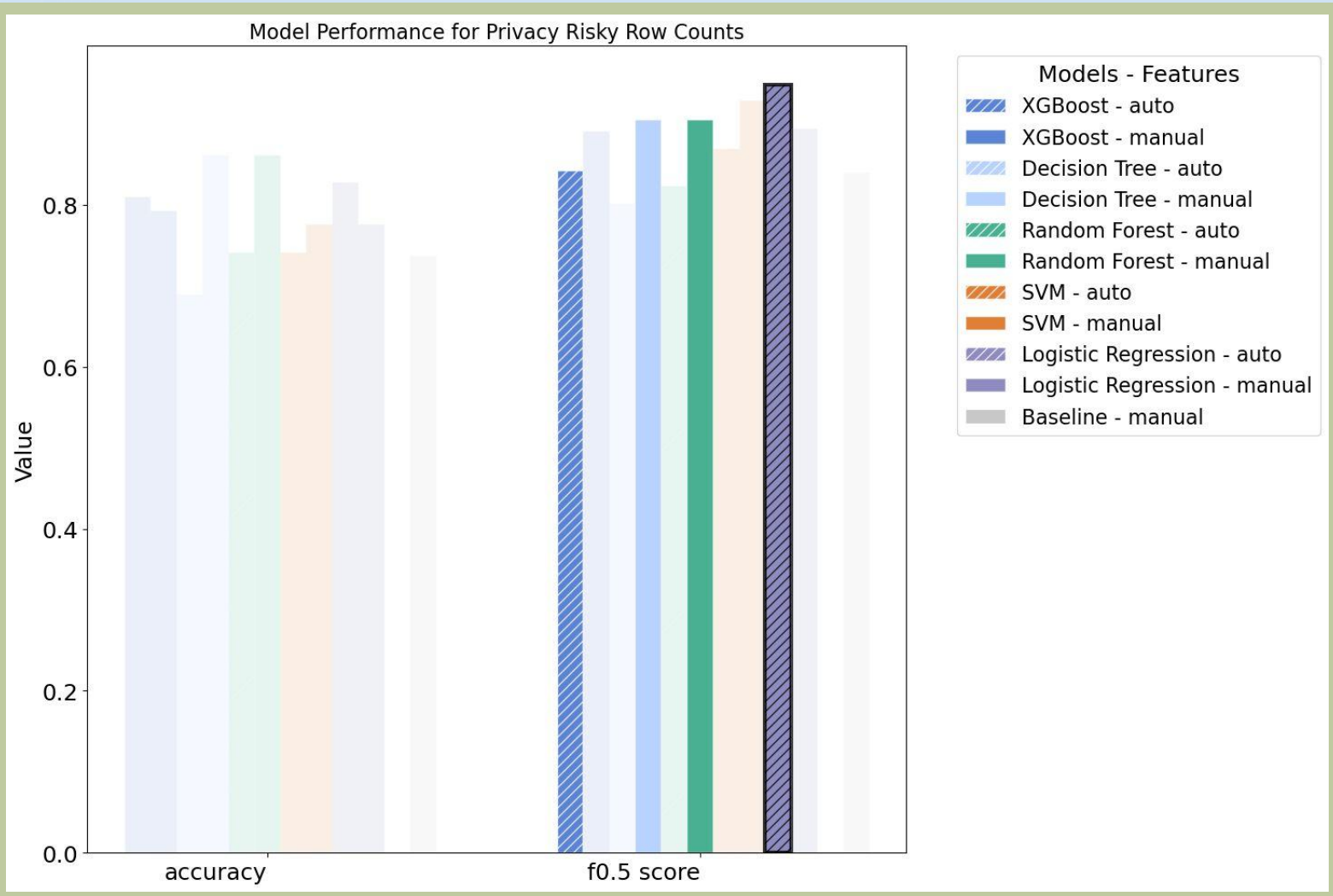


- ❏ 291 data entries
- ❏ 7 privacy tests
- ❏ 4 types pass/fail
  - ❏ Membership Inference 289/2
  - ❏ Attribute Inference 279/12
  - ❏ **Row Memorization** 266/25
  - ❏ **Risky Row Count** 241/50

Result – Risky Rows



Result – Risky Rows



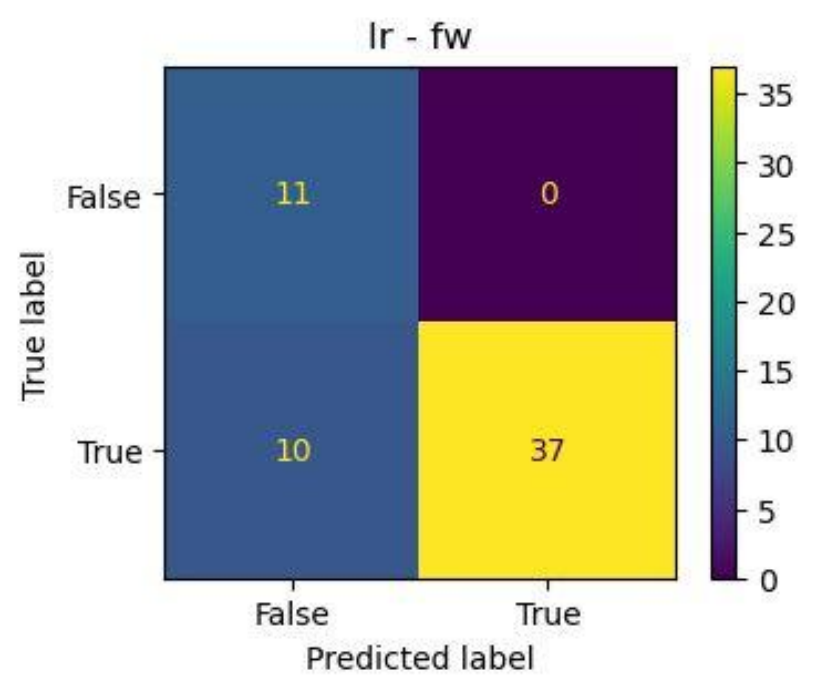
Result – Summary

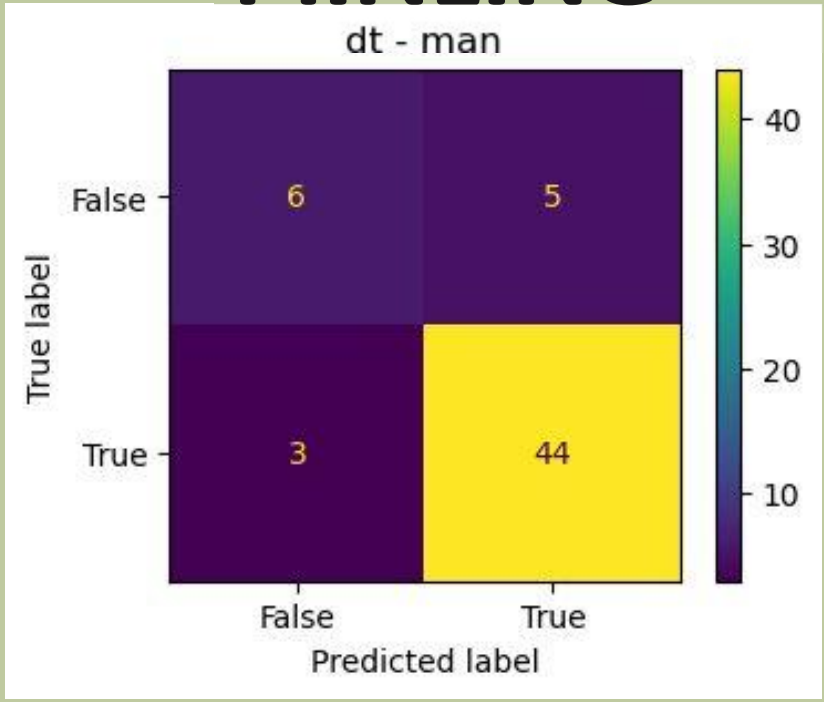
Risky Row Count	Model type	Accuracy	F0.5 Score	Specificity (TNR)	Sensitivity (TPR)
Best model	Logistic Regression	**0.83**	**0.95**	**1**	0.79
Base model	Decision Tree	0.74	0.84	0.42	**0.82**

- The **Base model** (Decision Tree) is slightly better at identifying truly passed test (higher sensitivity = 0.82).
- The **Best mode**l (Logistic Regression) is more accurate choice for a Risky Row privacy test because it perfectly identifies failed cases (specificity = 1.0) and has strong overall performance.

Result – Risky Row Count

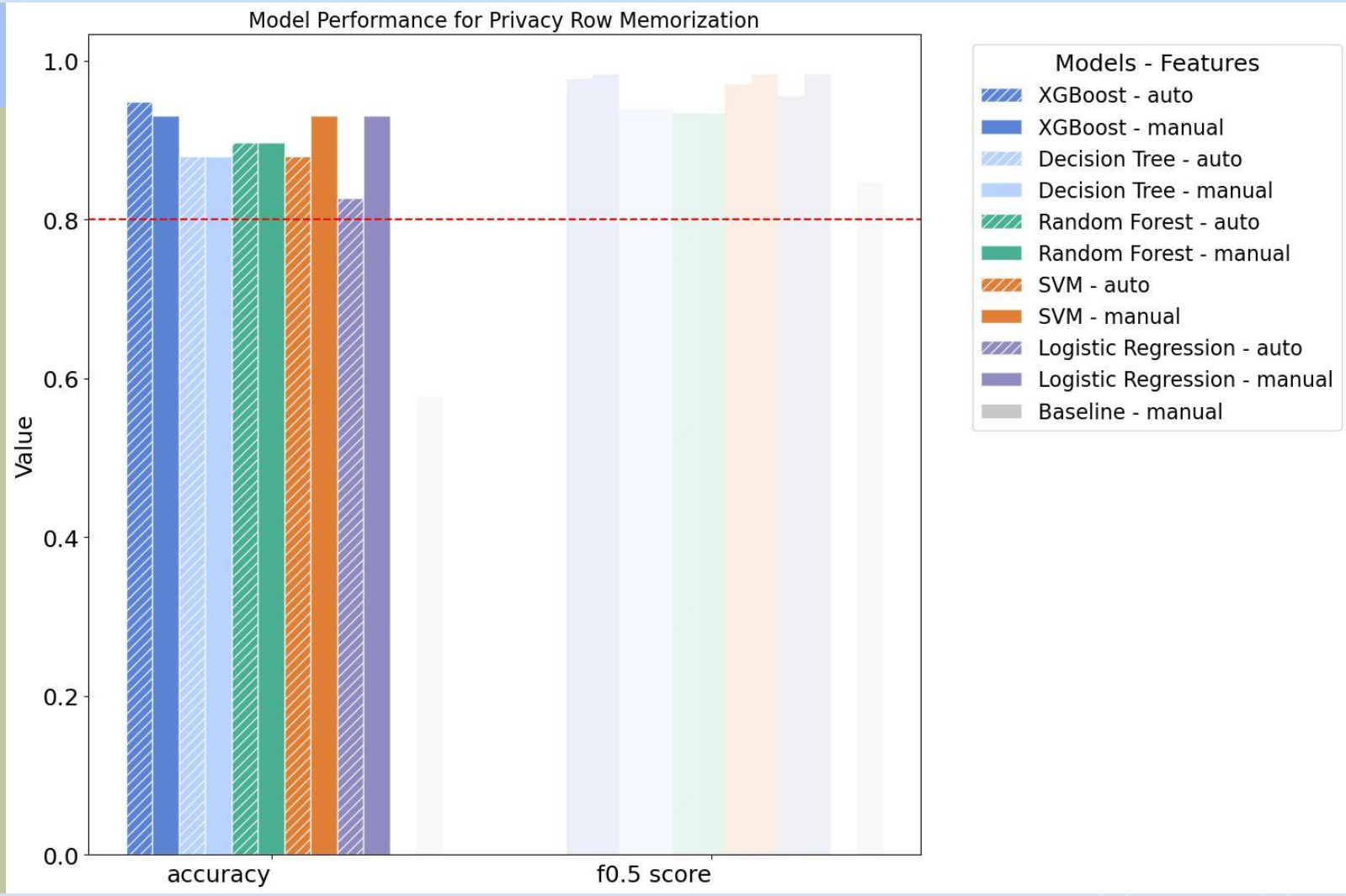
MINLING



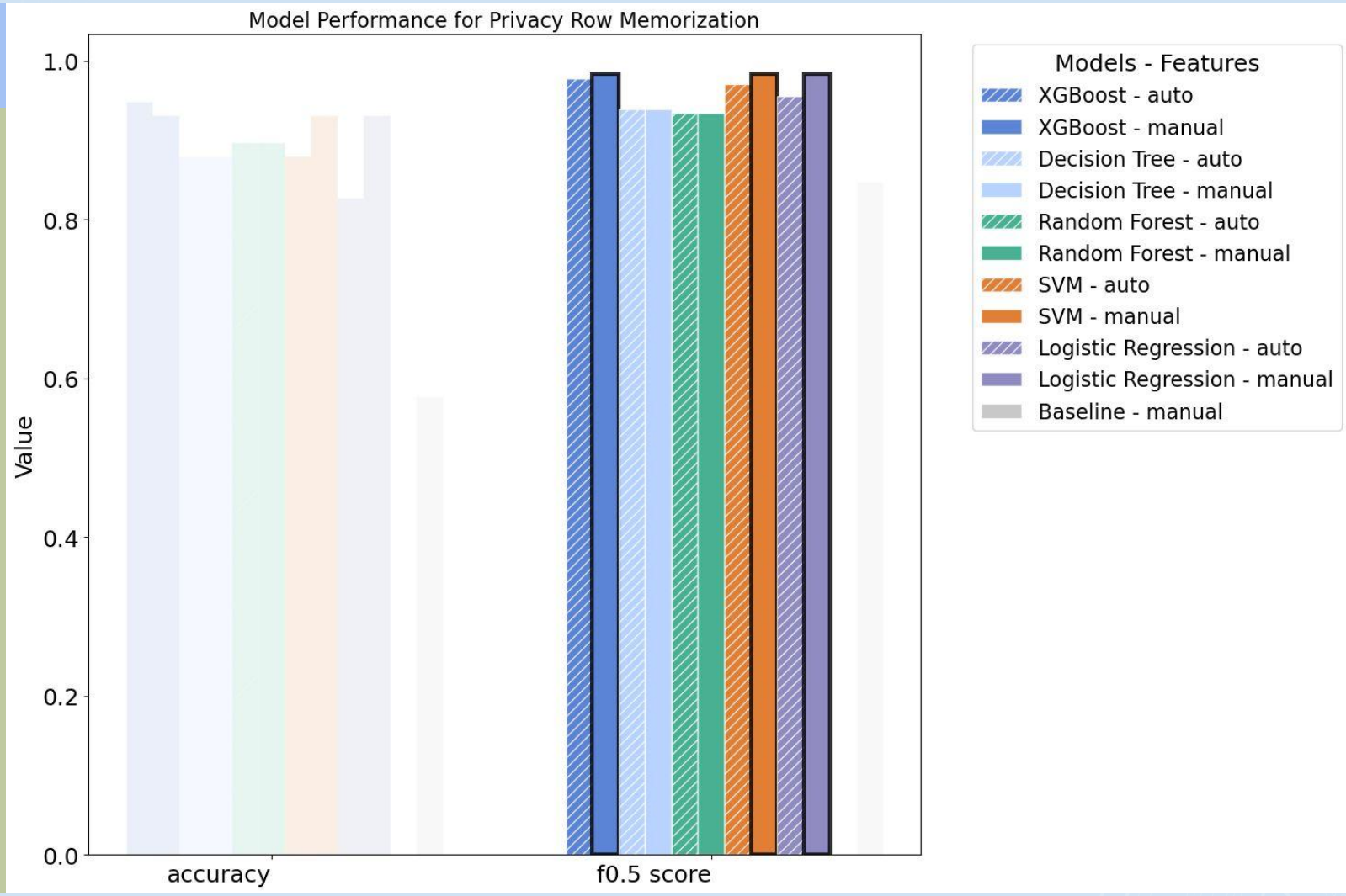


Accuracy: 0.83 F0.5 score: **0.95**	Accuracy: **0.86** F0.5 score: 0.91

Result – Row Memorization

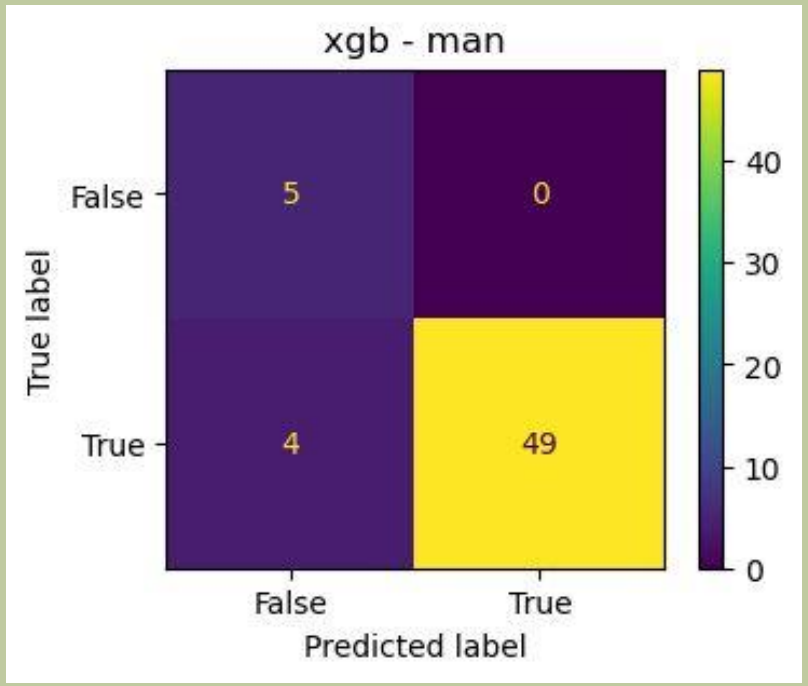


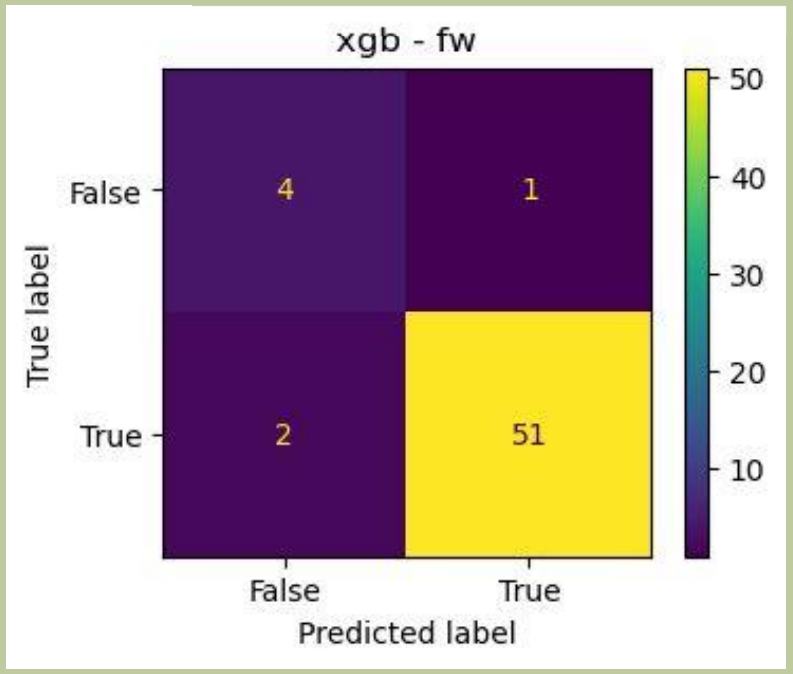
Result – Row Memorization



Result – Row Memorization

MINLING





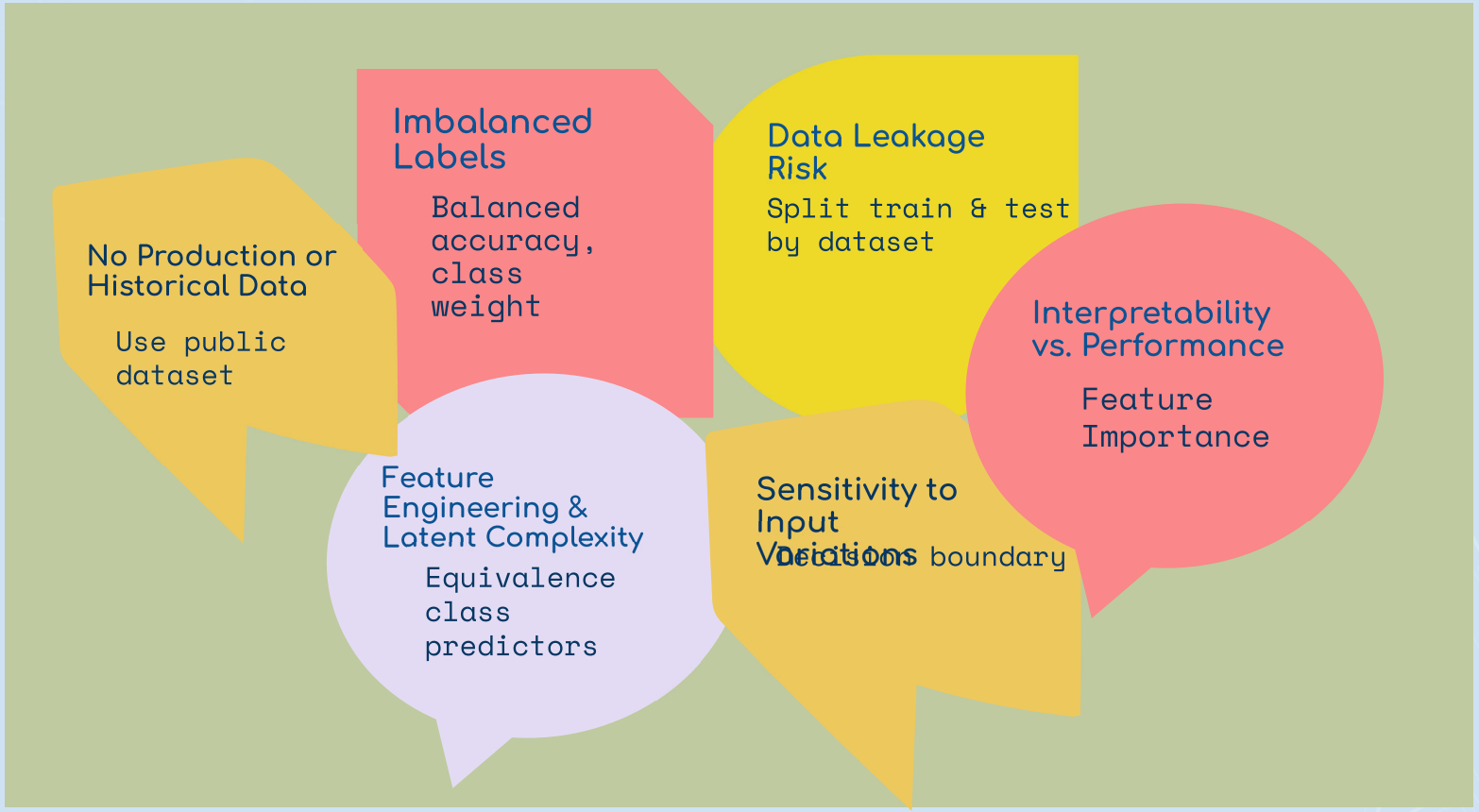
Accuracy: 0.93 F0.5 score: **0.984**	Accuracy: **0.95** F0.5 score: 0.977

Result – Summary

Row Memorization	Model type	Accuracy	F0.5 Score	Specificity (TNR)	Sensitivity (TPR)
Best model	XGBoost / SVM / Logistic Regression	**0.93**	**0.98**	**1**	**0.92**
Base model	Decision Tree	0.58	0.85	0.75	0.56

- The **Best Model** (XGBoost / SVM / Logistic Regression) is the clear winner for the Row Memorization privacy test, given its high accuracy, perfect detection of memorized rows, and excellent overall performance.

Challenges and Limitations



Future Plan

Expand Dataset & Improve Generalization

Leverage decision boundary generating more privacy test failure, incorporate more diverse dataset.

Feature Engineering & Interpretability

Use tools like SHAP to explain black-box models.

Privacy Test Expansion

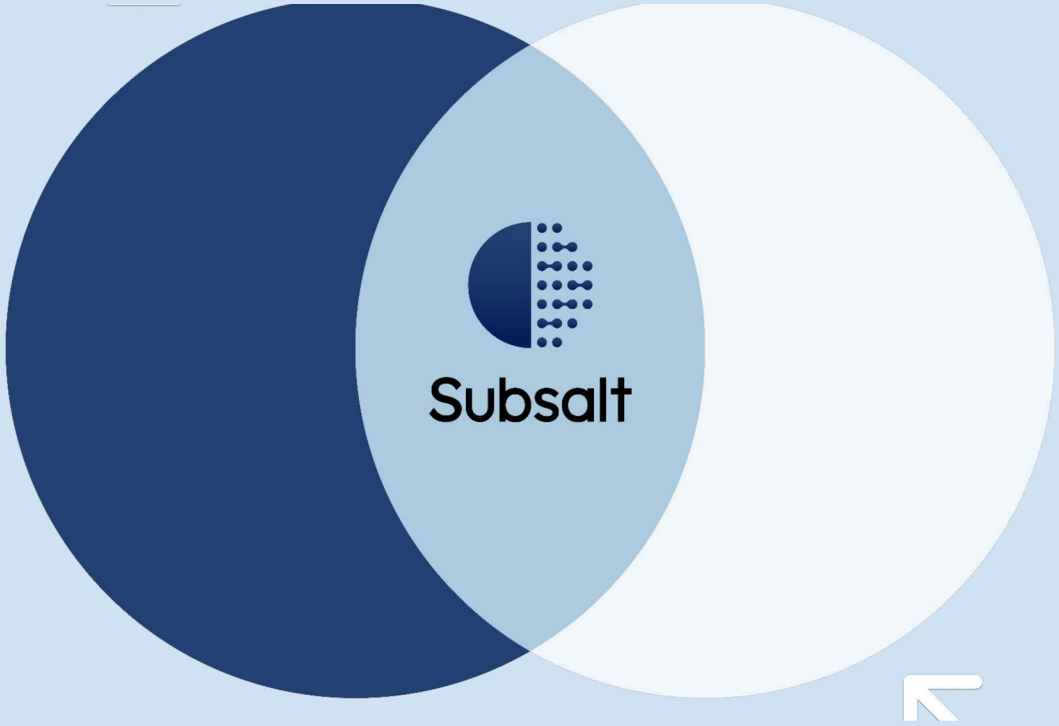
Investigate more privacy tests as failure patterns evolve.

Model Deployment & Validation

Deploy models in a controlled environment to evaluate performance on pre-production data.

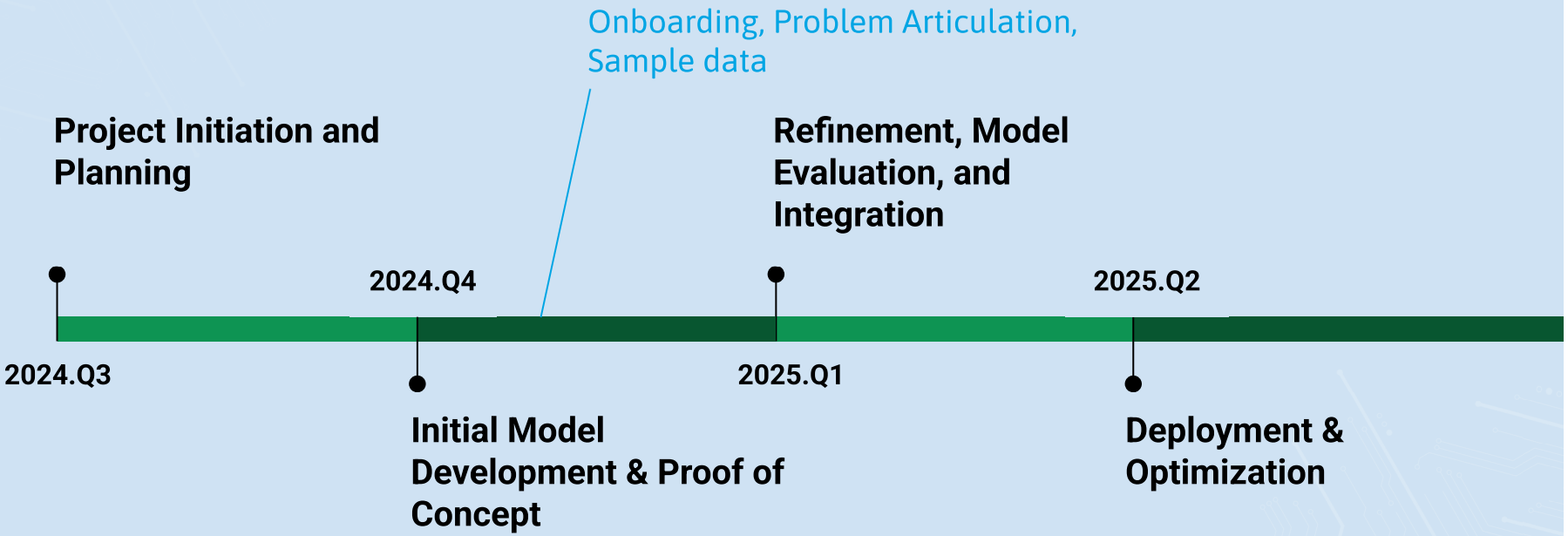
Thank You

A heartfelt thank you to **Luke Segars**, **Dylan Moradpour**, **Ethan Woodward**, and **Vijay Keswani** for your tremendous support, guidance, and collaboration throughout the year on SubSalt.



CONCLUSION

Project Timeline

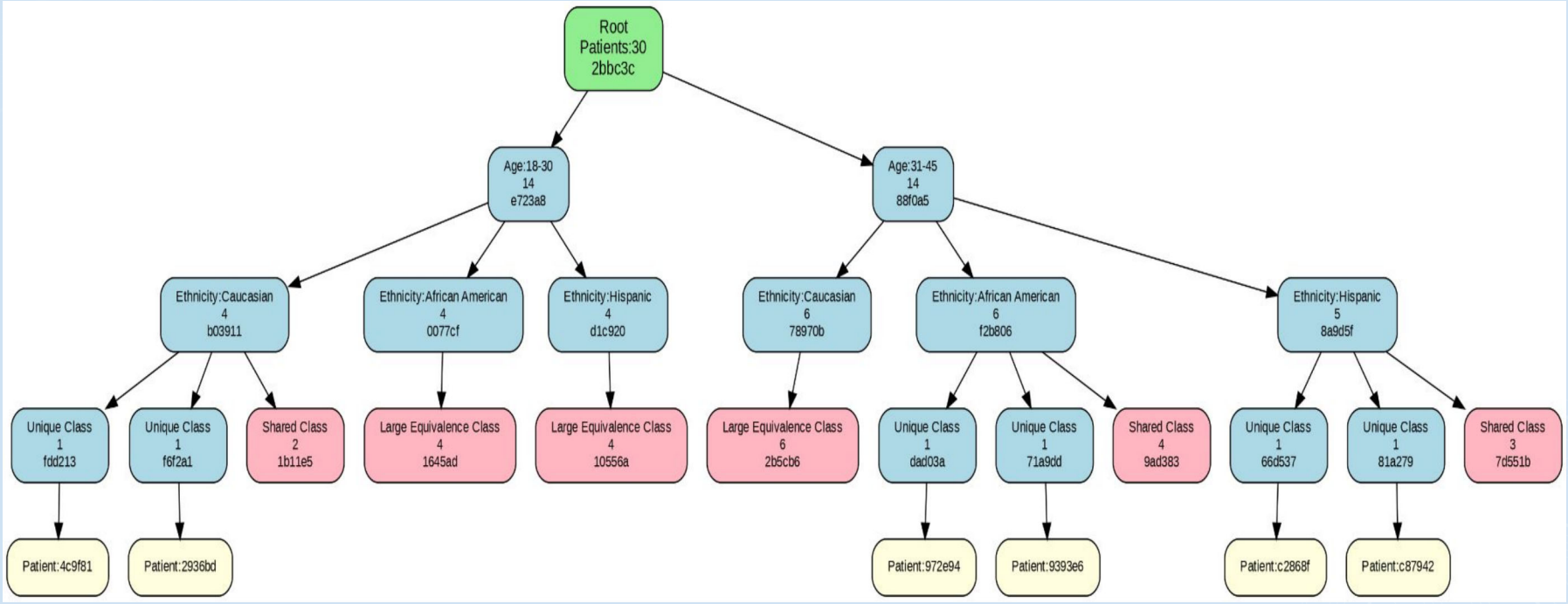


Potential Solution

1. **Dedicated MetaModel for Each Training Model**

2. **Unified MetaModel with Specialized Expertise**

Potential Solution



WHAT SUBSALT DOES

-
-

-
-
-
-
-

Questions for Luke -

Industries being targeted

Background

Potential Risk

01	Usage of Public Datasets	• Use business data • Use synthetic data generated from business data • Use historical report
02	Gap of Documentation vs. Sample Data	• Dual diligence • Agile
03	**Schedule Management**	• High Availability • Flexibility

HOW THEY DO IT

This is where we explain the workflow and how they’re doing it right now.

THE PROBLEM

- Waste of time and computational resources

LAWS AND THE MATH PAPER

This wont be the main focus but we should mention a few words

What subsalts solution is and what we are trying to build

- A way to catch datasets that are likelier to fail
-

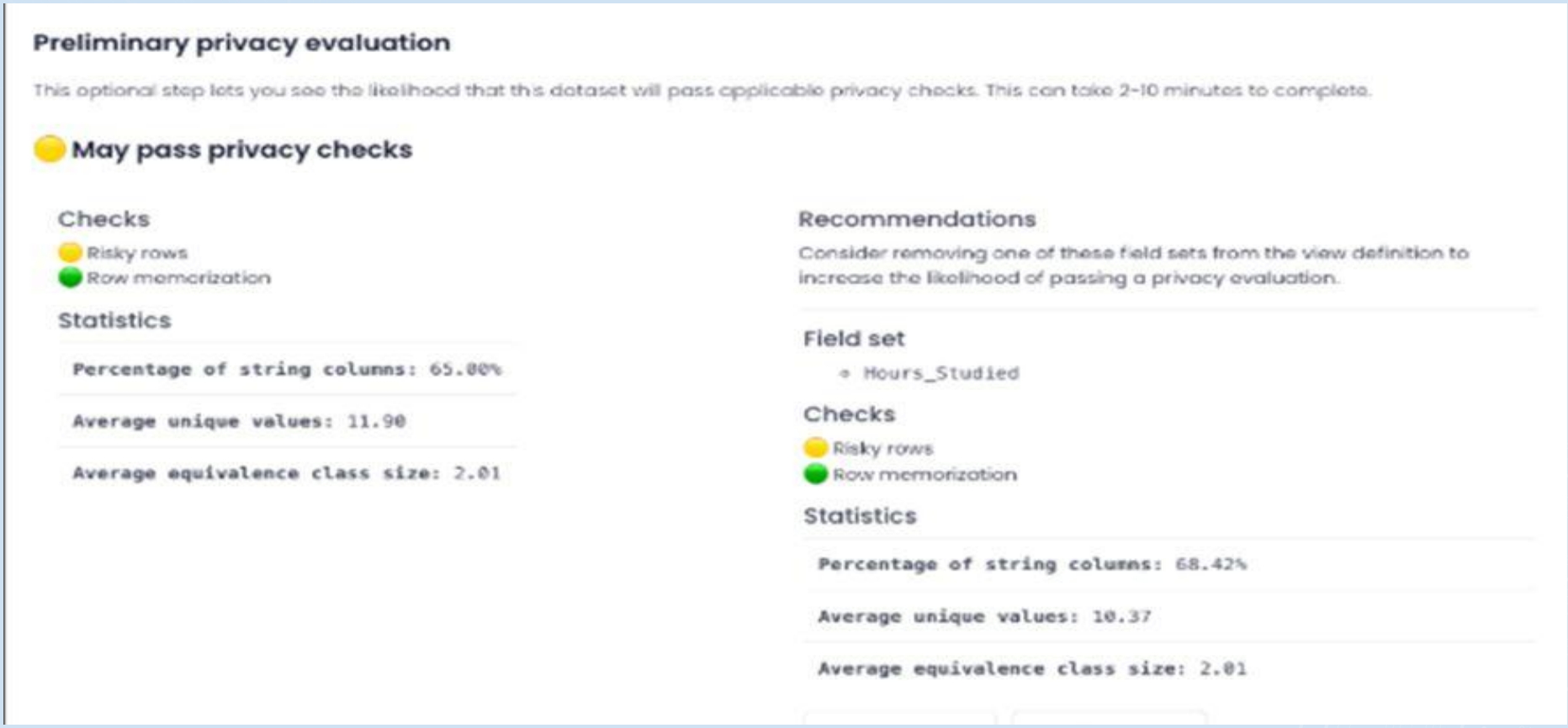
How we plan to do it

We do not need to have a solution yet -

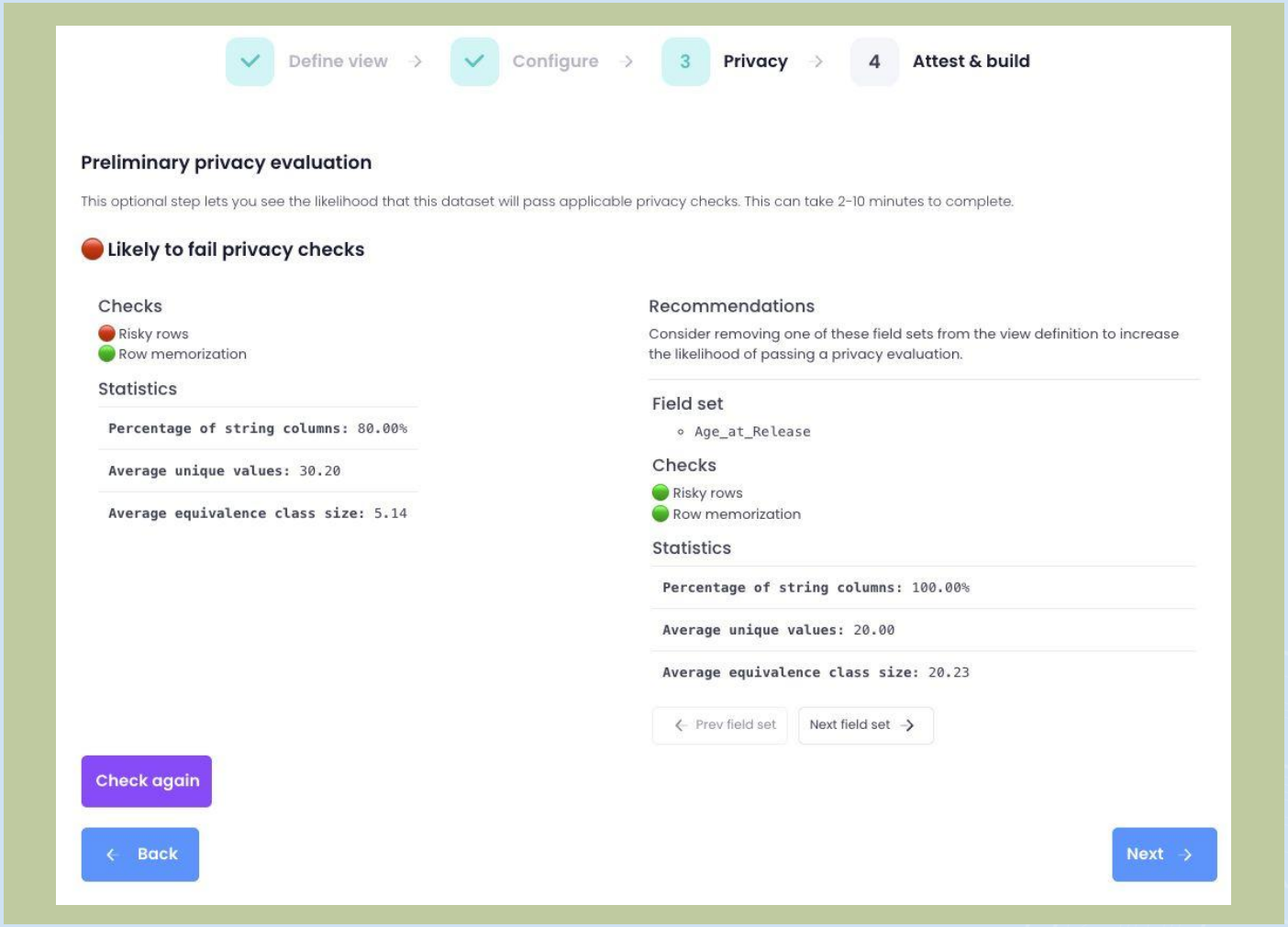
A few ideas will be good though

QUESTION FOR LUKE - Confirm workflow what happens when attributes are removed/ flagged as indirect by admin

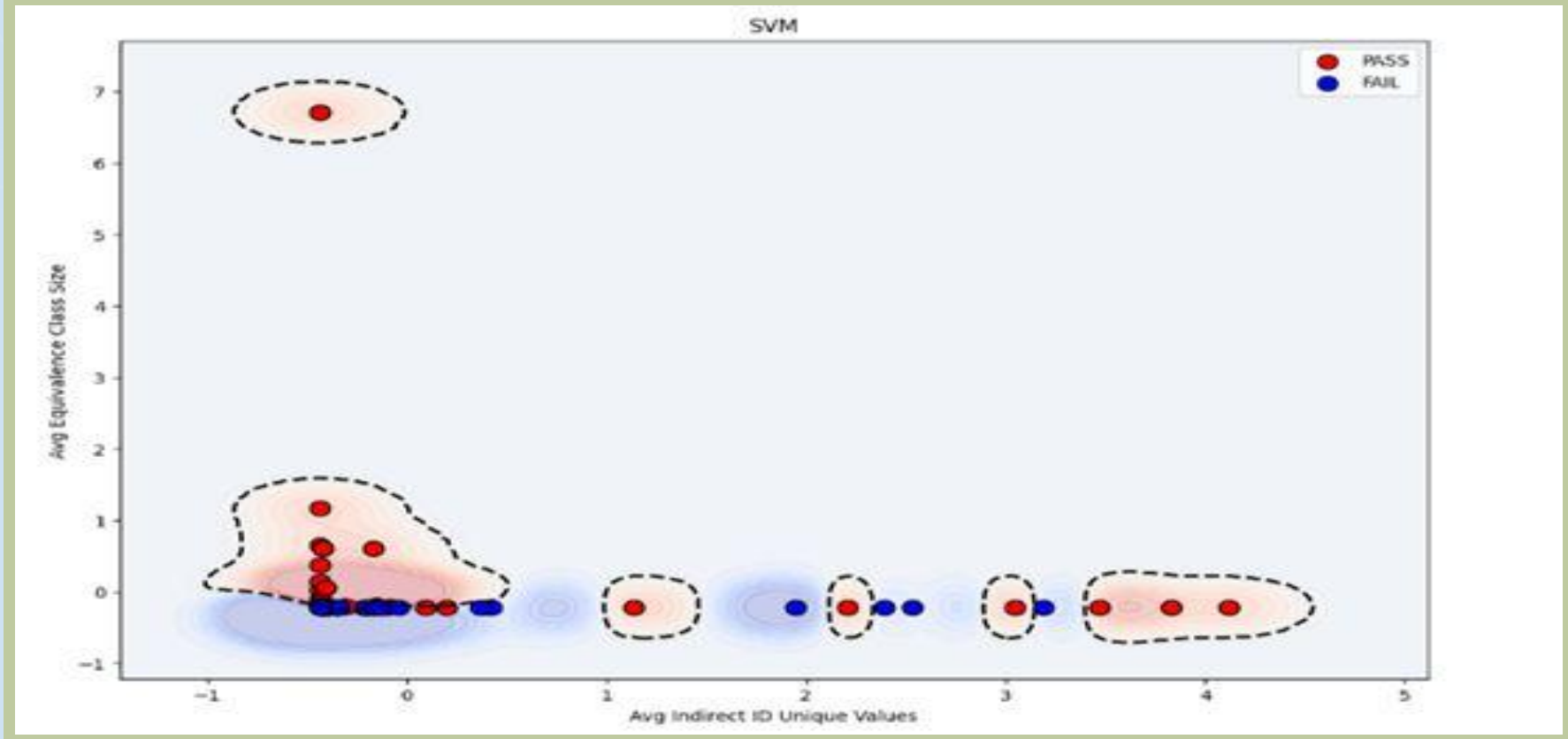
RECOMMENDATION SYSTEM



Recommendation System



SVM (Kernel: RBF)



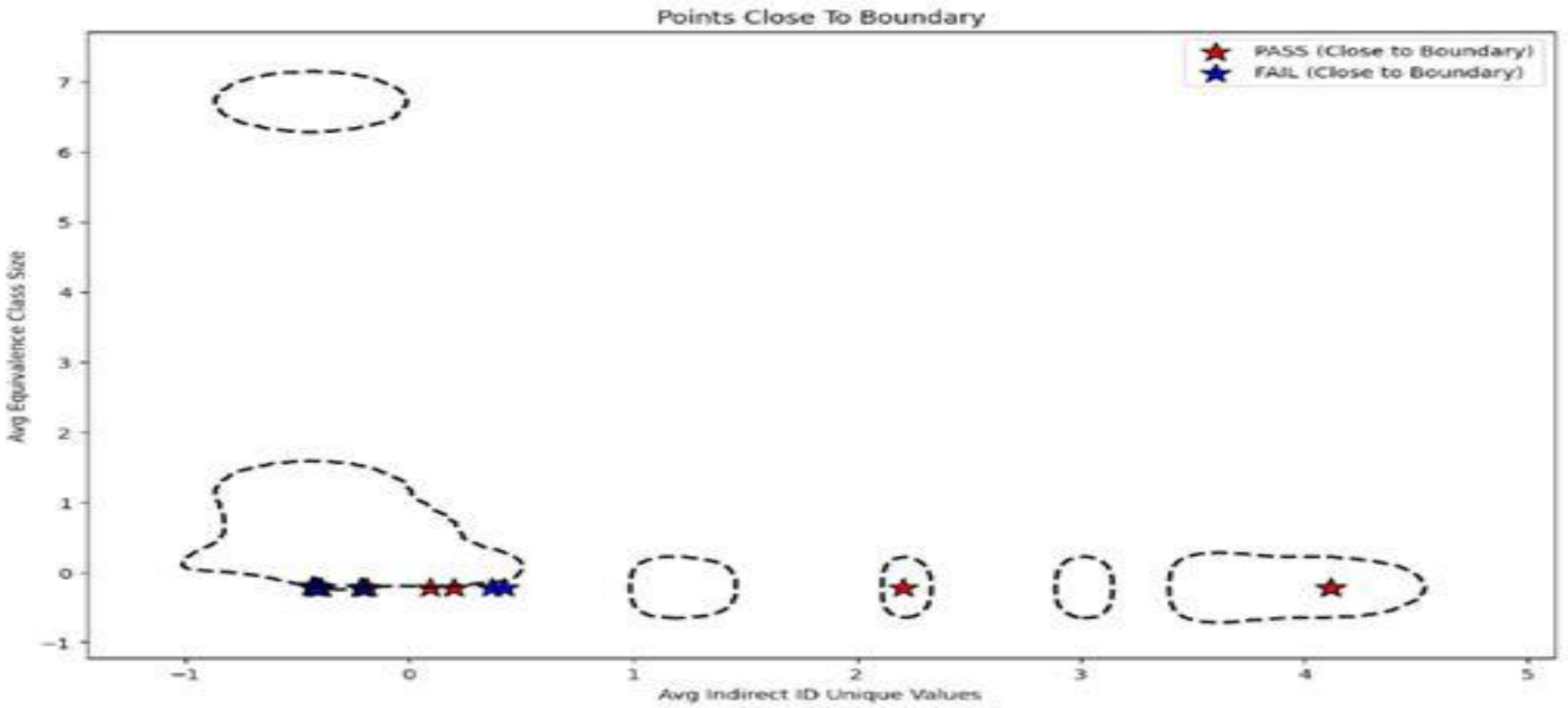
Decision Boundary

Distance From Boundary

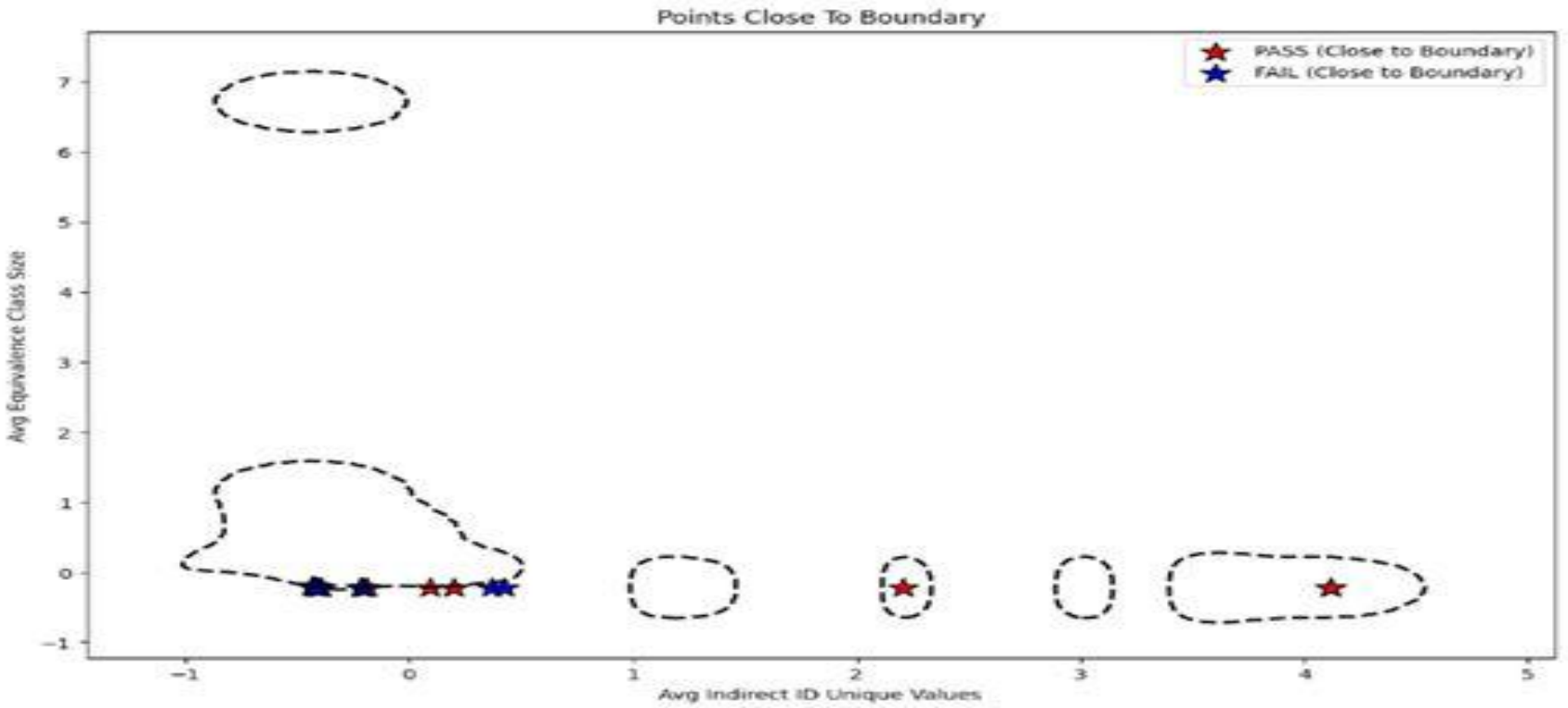
SIGNED FUNCTIONAL DISTANCE (SVM Decision Function)

- Average Distance to Boundary for PASS points: 2.22
- Average Distance to Boundary for FAIL points: 1.77
- Threshold for points close to boundary - 1 unit.

SVM (Kernel: RBF)



SVM (Kernel: RBF)



CONCLUSION

- Predicting Test Outcomes
- Helping make more tests pass with minimal removal of data