Estimation of individual privacy risk in data sharing using predictive models

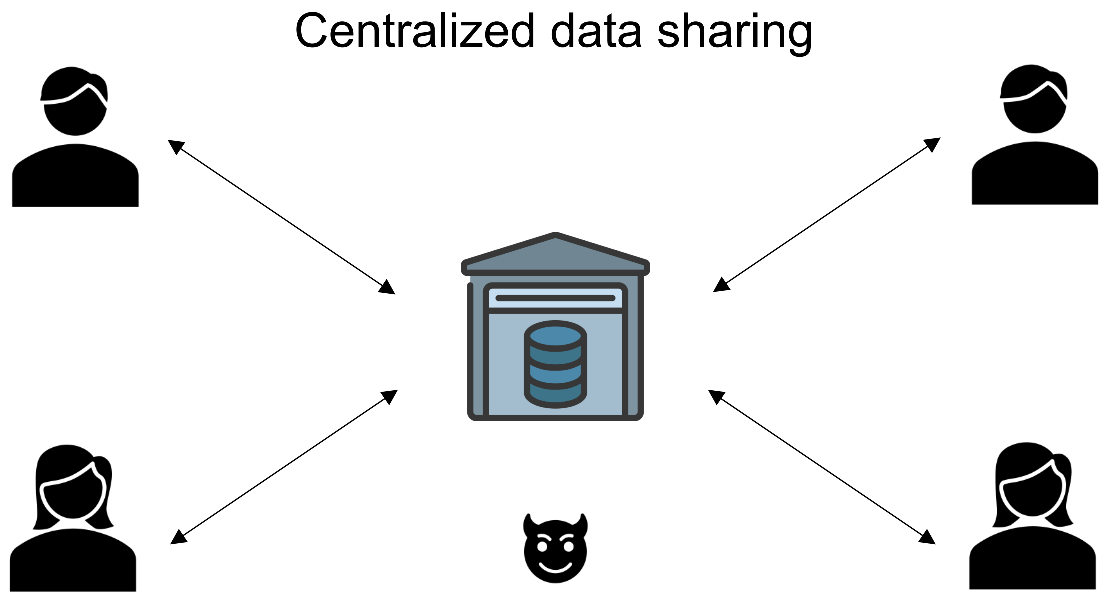
Master Thesis Proposal

Manjari Chaudhri

RWTH Aachen, Informatik 5, Information Systems
Prof. Dr. Stefan Decker

Advisors: Felix Hermsen and Mehdi Akbari Gurabi

Motivation for data spaces



Single point of failure

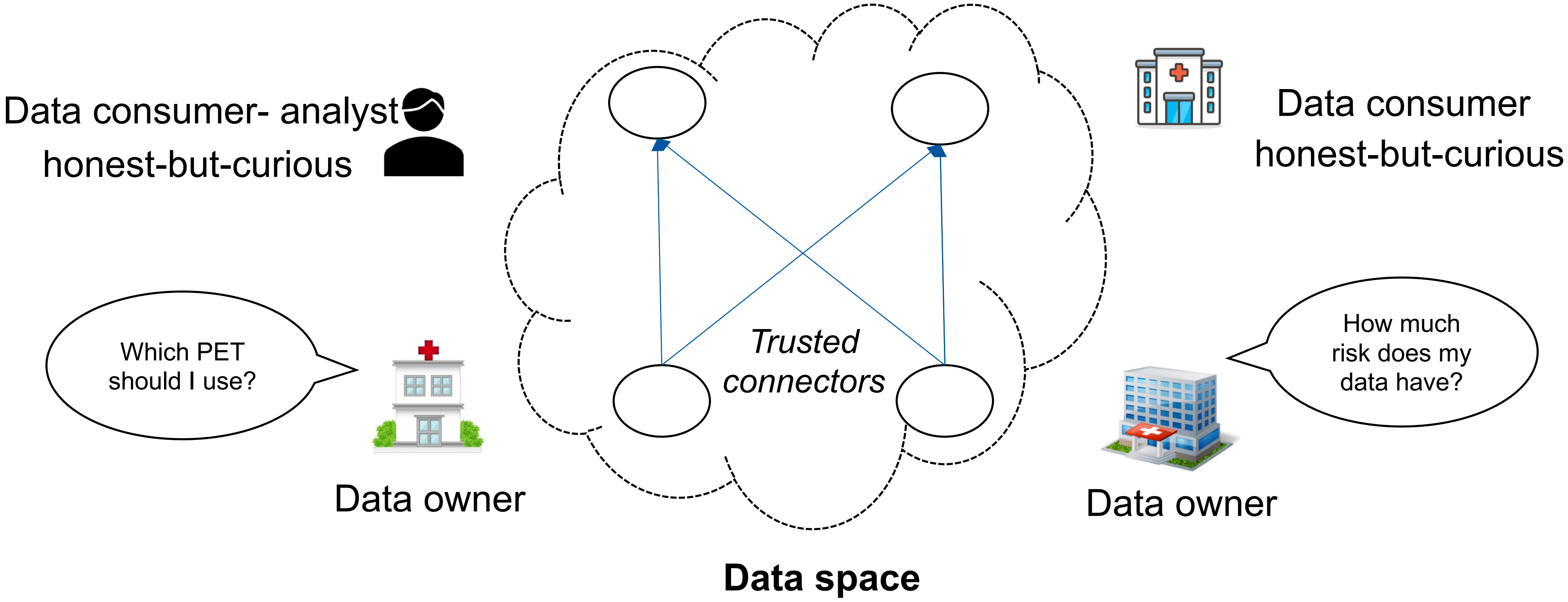
Inflexibility

Governance

Accountability

Ineffective

Data spaces – Medical data space



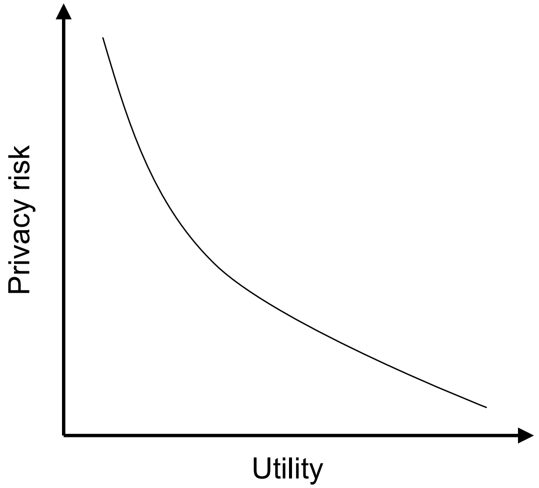
All data sharing must be GDPR complaint.

Data sharing and privacy risk

87% of Americans identified based on 5-digit ZIP, gender, date of birth – Sweeney (2000)

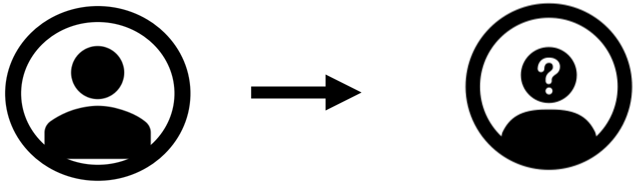
Solution – PETs: Syntactic anonymization, Differential Privacy, Synthetic data

How much data utility do we lose?



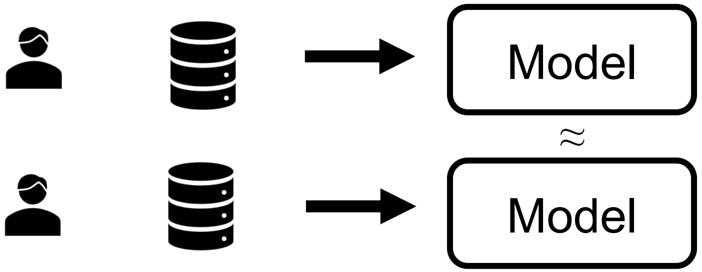
Privacy enhancing technologies

Anonymization



! Does not really work in practice

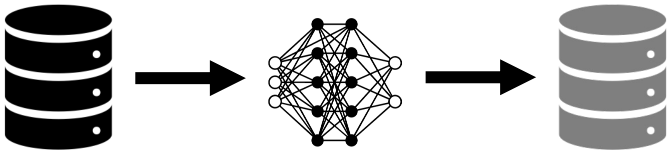
Differential Privacy



! Utility tradeoffs

Gives an upper bound on privacy risk

Synthetic Dataset Generation



! Utility tradeoffs

Residual risk remains

GDPR and privacy risk

Singling out

DOB	Gender	Attempts	Postcode
10-01-1955	Male	3	17329

→

Status
Accepted

Inference

Linkability

DOB	Gender	Attempts
13-03-1994	Male	4

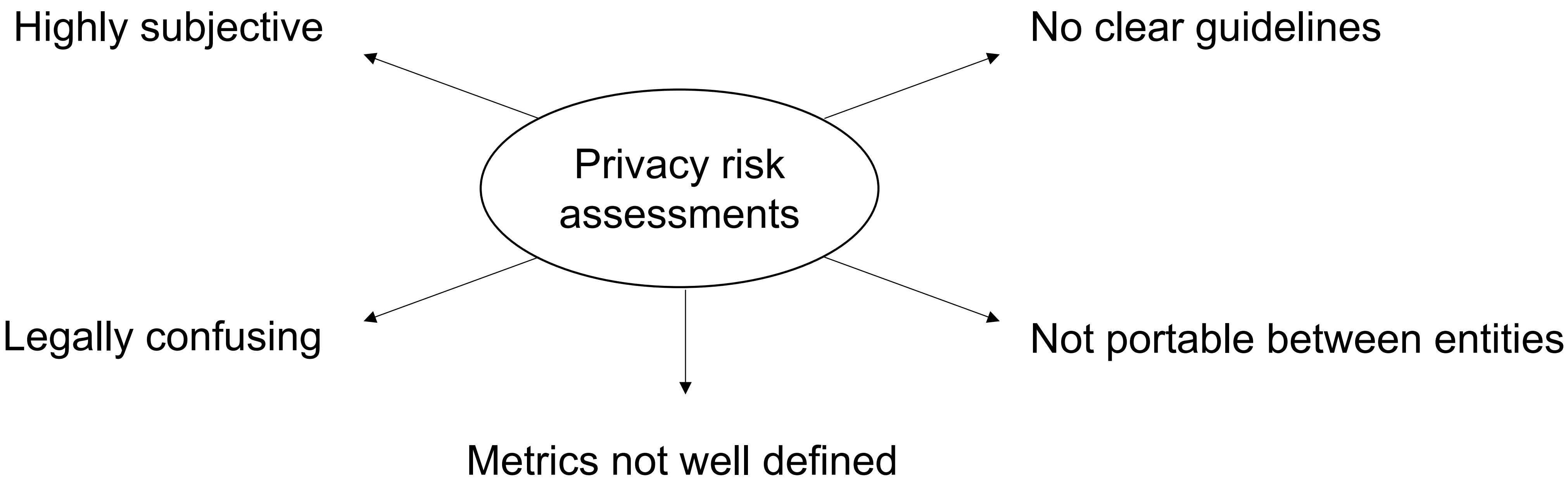
→ 👤

AND

Postcode	Status
52066	Waitlist

→ 👤

Problems with current methods



Privacy risk assessments

Highly subjective

No clear guidelines

Legally confusing

Not portable between entities

Metrics not well defined

Privacy metrics in research

In research, privacy is measured in many different ways

Two rough categories -

1. Inherent to data – entropy, information gain : statistical properties of the dataset
   - Quantifiable properties of a dataset – dispersion, skewness, correlation, outliers
2. Adversary based
   - Probability of success
   - Time to success
   - Accuracy

Privacy metrics in research

Information based	Attack based
Do not need an adversary.	Dependent of adversary capabilities
Which statistics are important for privacy?	How can we model every adversary?
How do they relate to practical privacy?	Can we make them computationally efficient?

Can we combine these two approaches to predict privacy risk in a computationally efficient manner?

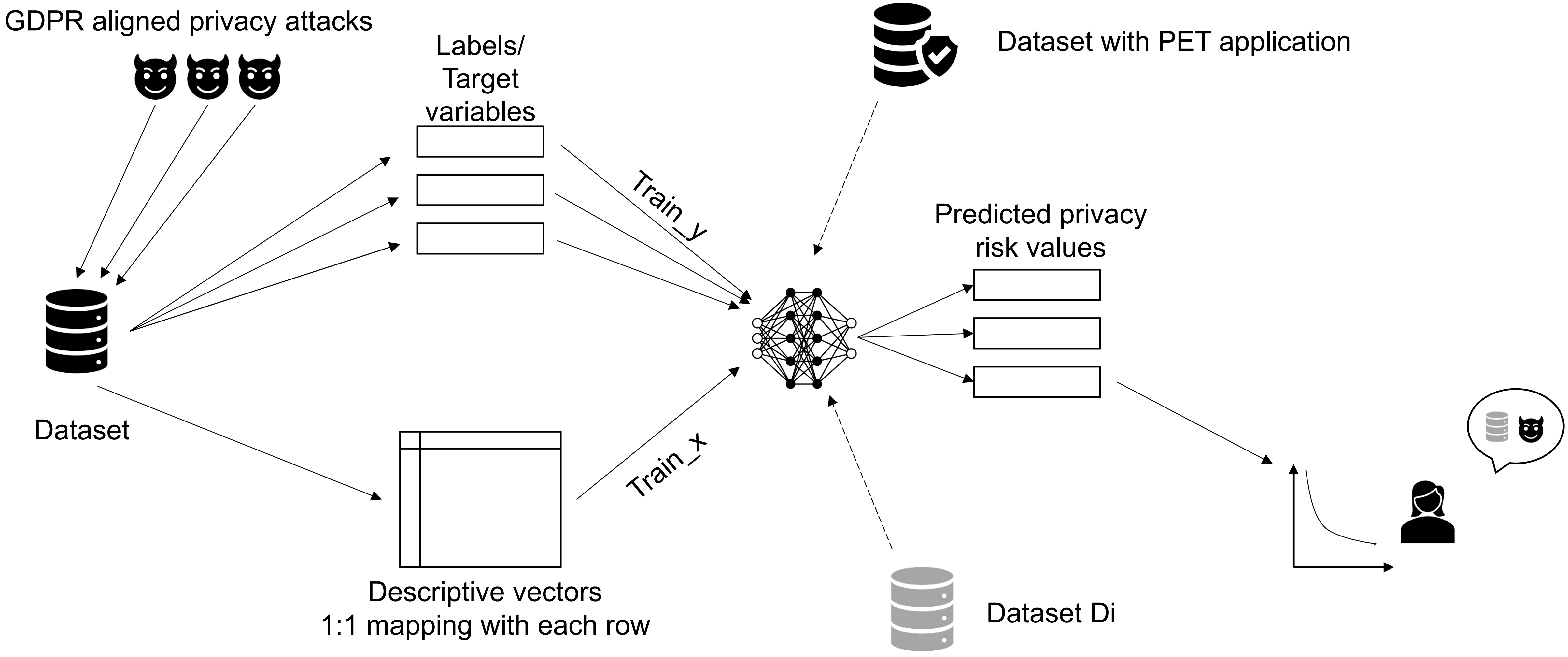
Does every record have the same level of risk?

Research question

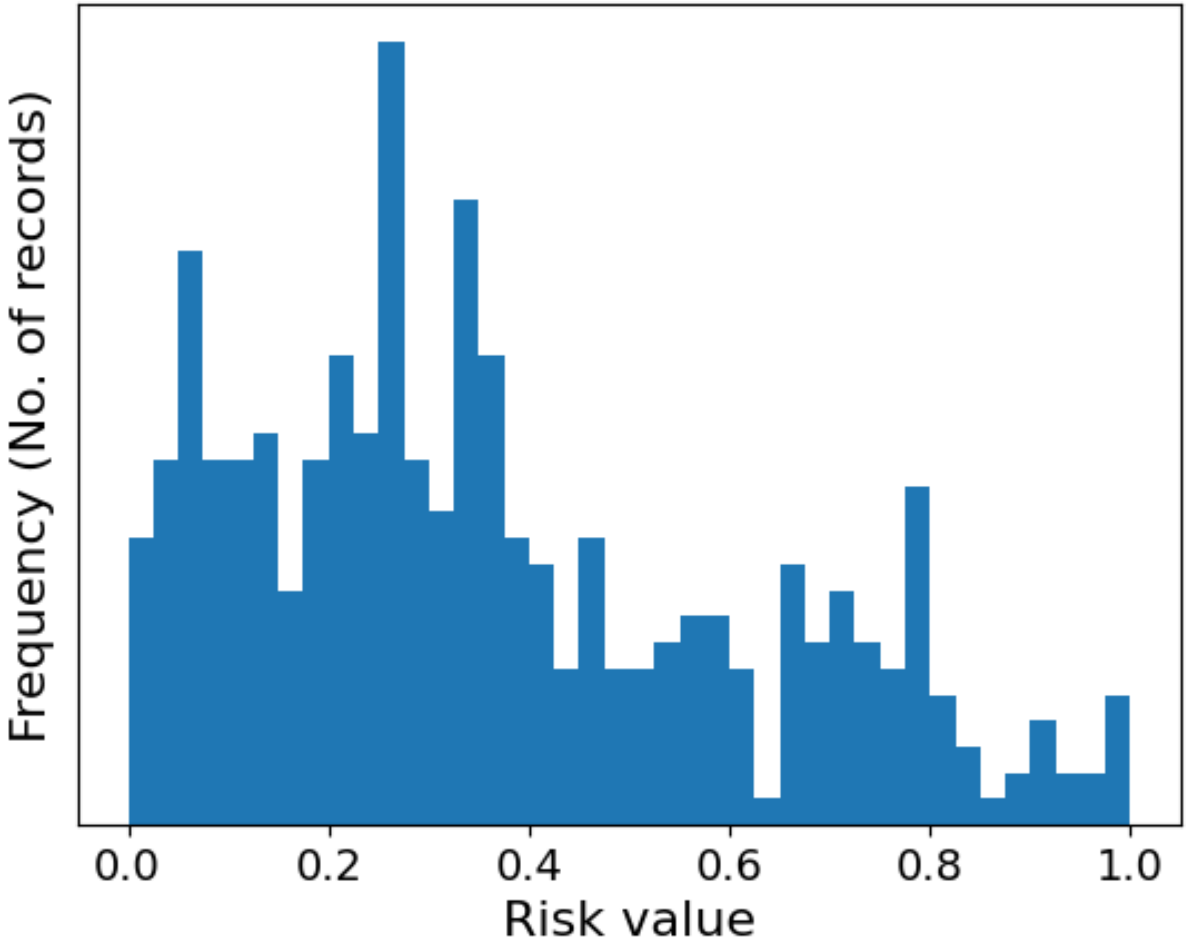
Research question – Can we use Machine Learning to predict privacy risk for each record individually? Can we make the privacy metric legally meaningful (GDPR aligned)?

Conceptual approach – Model learns from inherent characteristics of the dataset based on simulated attacks to predict the privacy risk score for each individual

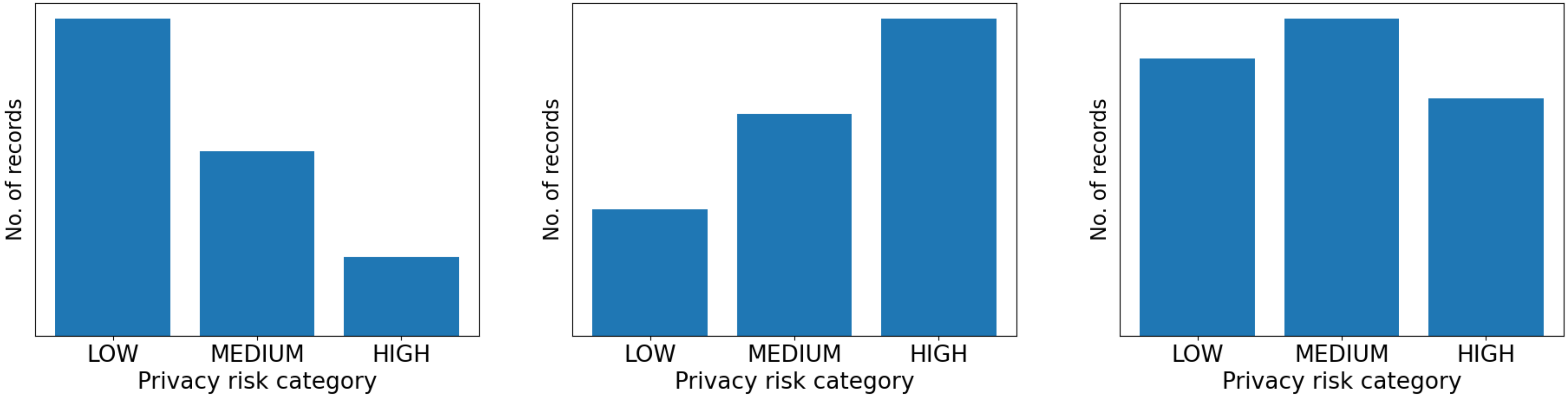
Conceptual approach



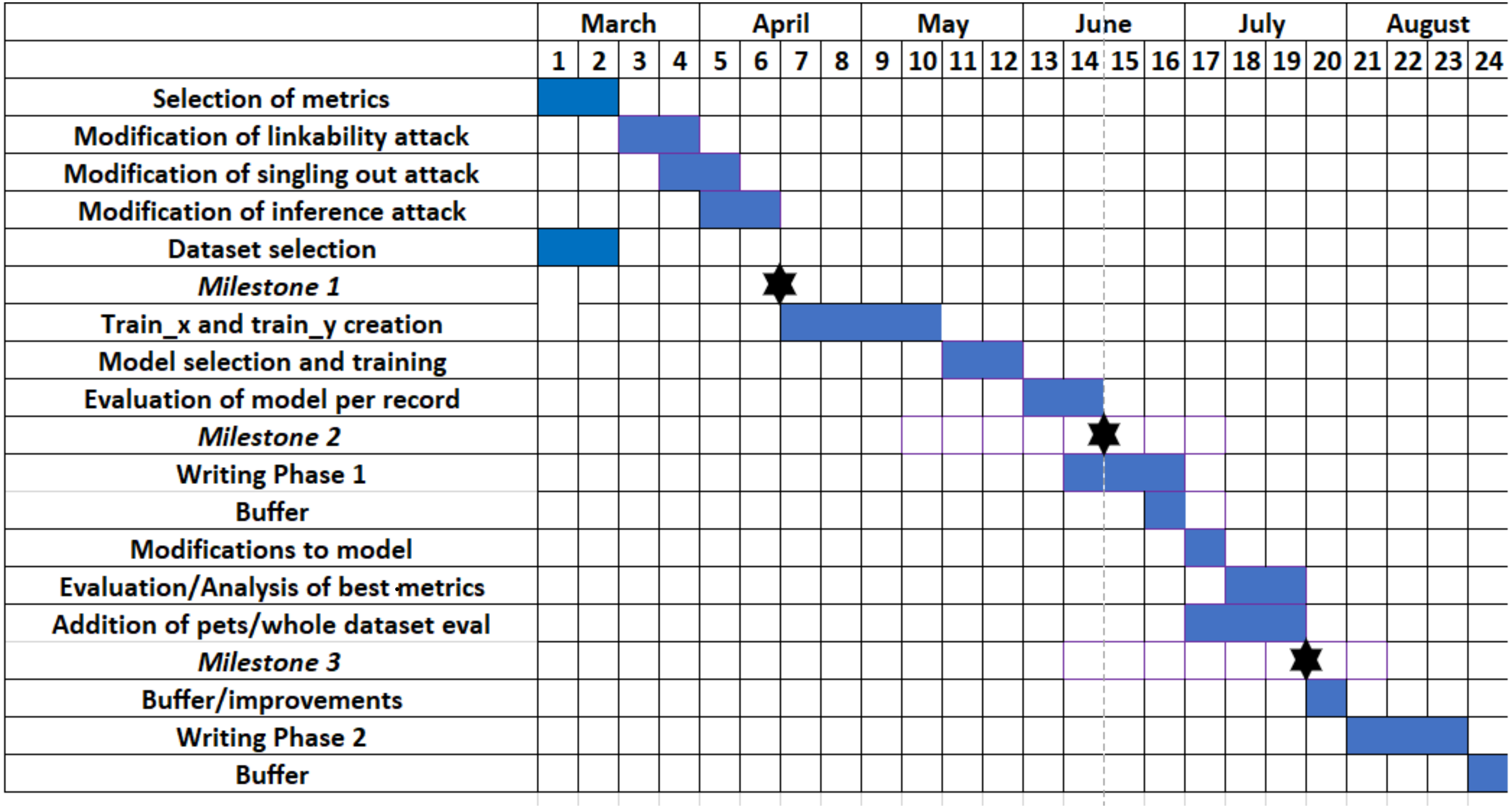
Risk distribution graph - Regression



Risk distribution graphs - Classification



Timeline



	March				April				May				June				July				August			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Selection of metrics	■	■																						
Modification of linkability attack			■	■																				
Modification of singling out attack				■	■																			
Modification of inference attack					■	■																		
Dataset selection	■	■																						
Milestone 1						✶																		
Train_x and train_y creation							■	■	■	■														
Model selection and training											■	■												
Evaluation of model per record													■	■										
Milestone 2														✶										
Writing Phase 1														■	■	■								
Buffer																■								
Modifications to model																	■							
Evaluation/Analysis of best metrics																		■	■					
Addition of pets/whole dataset eval																	■	■	■					
Milestone 3																			✶					
Buffer/improvements																				■				
Writing Phase 2																					■	■	■	
Buffer																								■

Summary

- Privacy risks in a dataset are currently a subjective measure.
- Statistical properties of a dataset can be used to quantify the privacy risk.
- Using attack based metrics (with a legal standing) can give practical meaning to privacy of an individual record.
- We will use machine learning to predict privacy risk by using a combination of the two approaches.

Thank you for your attention!