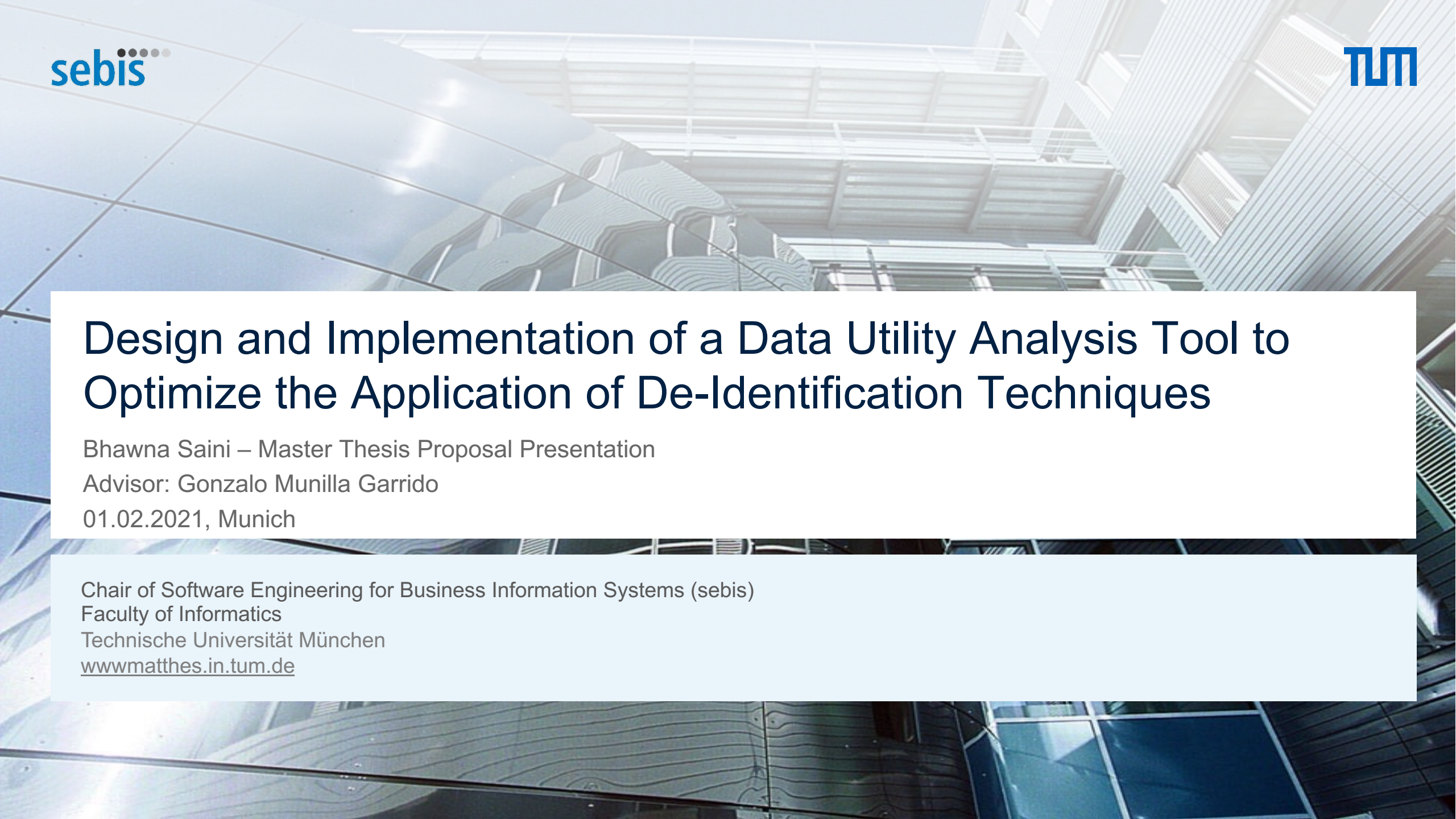


Design and Implementation of a Data Utility Analysis Tool to Optimize the Application of De-Identification Techniques

Bhawna Saini – Master Thesis Proposal Presentation

Advisor: Gonzalo Munilla Garrido

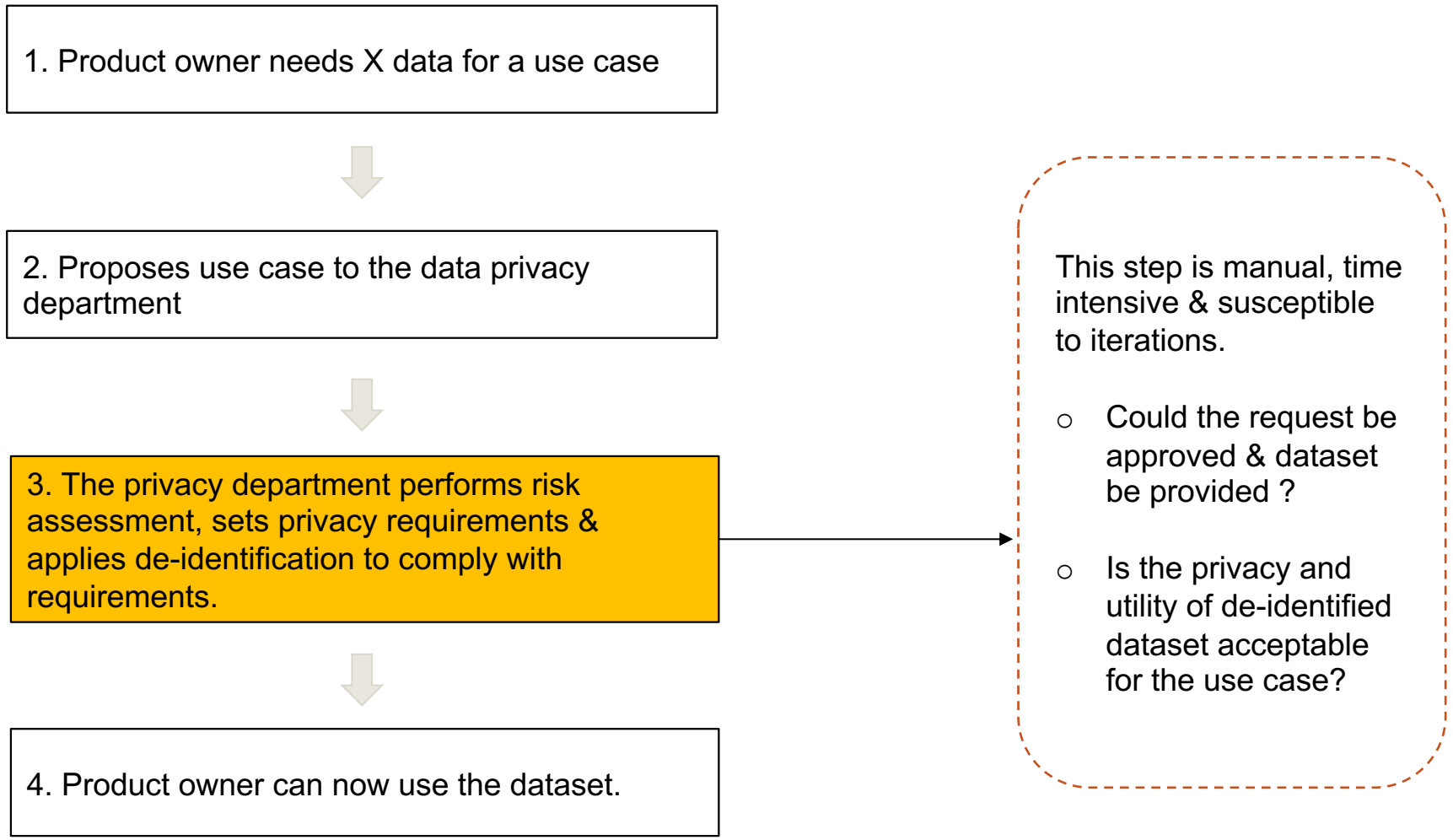
01.02.2021, Munich

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

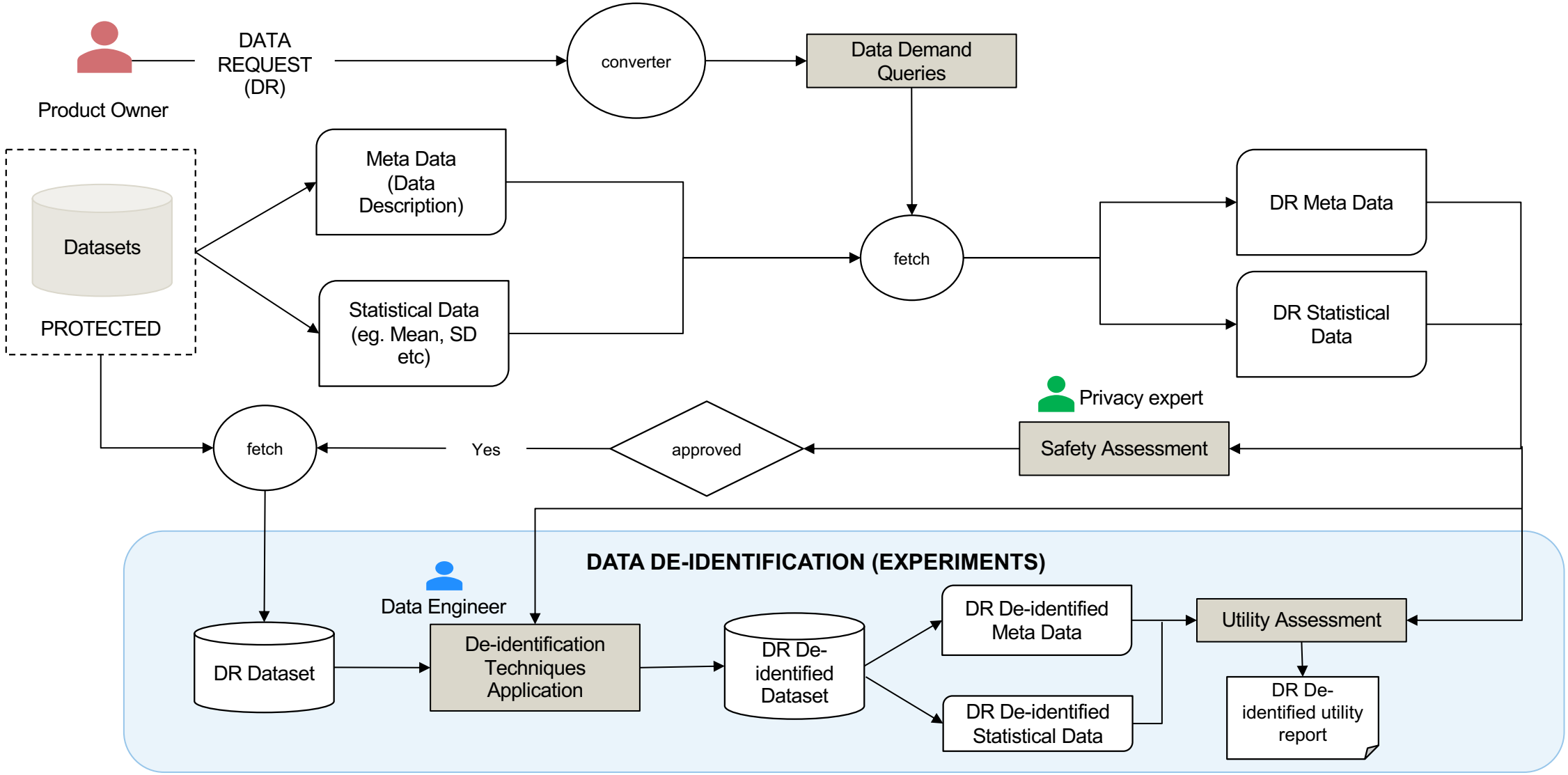
Outline

1. General Privacy Application Scenario in an Enterprise
2. Overview of a Data-De-identification Process
3. Current Status of the Data De-identification Process
4. Thesis Goal
5. Data Utility Metrics
6. Optimised Data De-Identification Process
7. Example
8. Research Questions
9. Master Thesis in a nutshell
10. Timeline

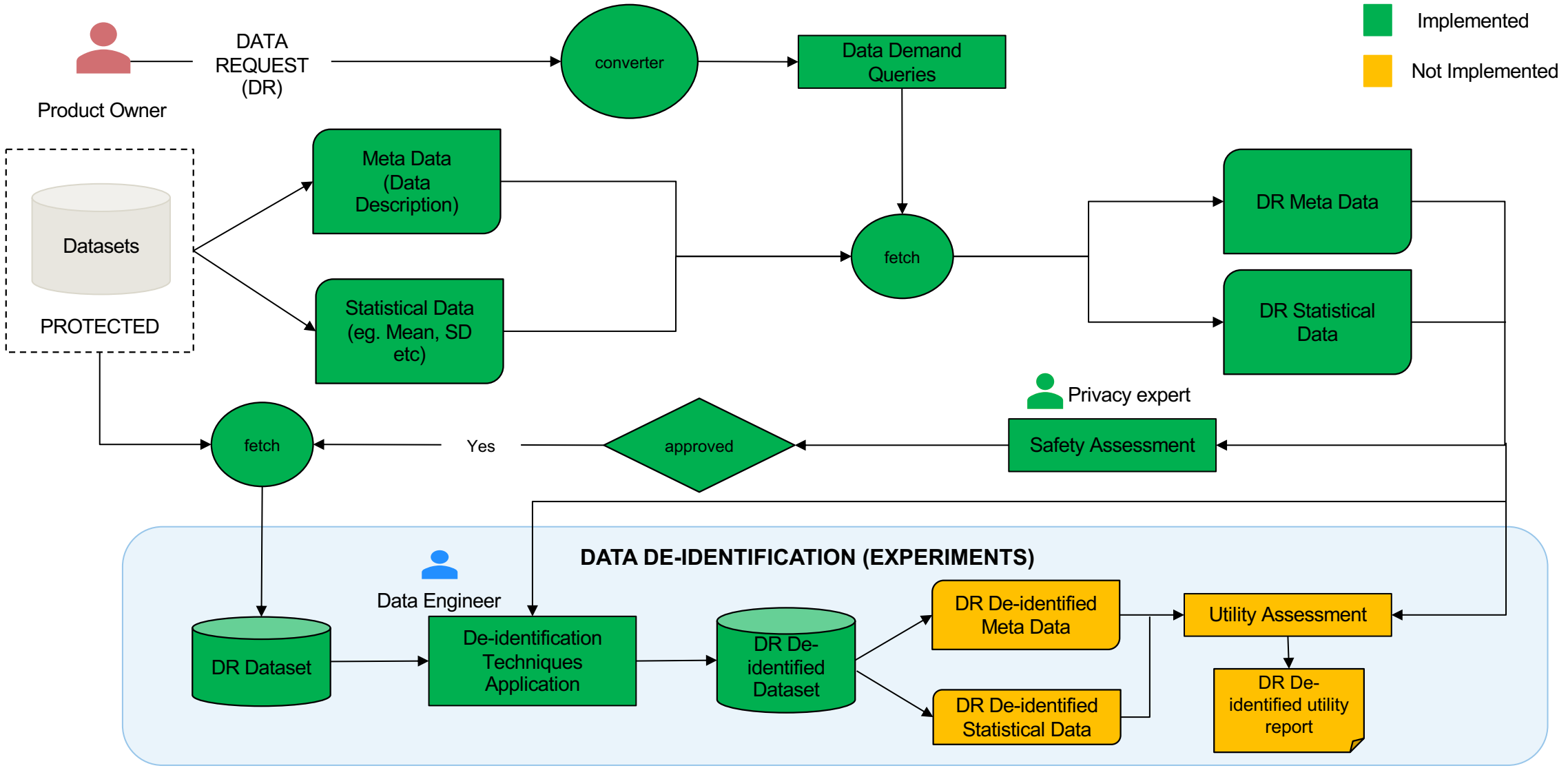
1. General Privacy Application Scenario in An Enterprise



2. Overview of a Data-De-identification Process: Workflow



2. Overview of a Data-De-identification Process: Status



3. Current Status of the Data De-identification Process

- Data De-identification is a manual step.
- De-identification is carried out based on the rules decided by privacy experts.
- These rules further depend on the dataset requirements and the dataset properties.
- **No data utility tool exists to measure the quality of the resulting de-identified data and the applicability of the de-identification techniques**.

4. Goal of the thesis

*Design and implement a **data utility analysis tool** that would help privacy experts to **quantify the performance** and **optimise the application** of de-identification techniques on datasets.*

This tool could further help in benchmarking the performance of various privacy models and their algorithms

5. Data Utility Metrics

- Quantifies the changes in utility/information-loss in de-identified data with regards to a specific goal

Example Goal	Suitable Utility Metric
Extract information of the distribution of values for specific attributes in a de-identified dataset.	Discernibility Metric

Discernibility Metric
Estimates data quality based on the size of the equivalence classes in the output dataset.

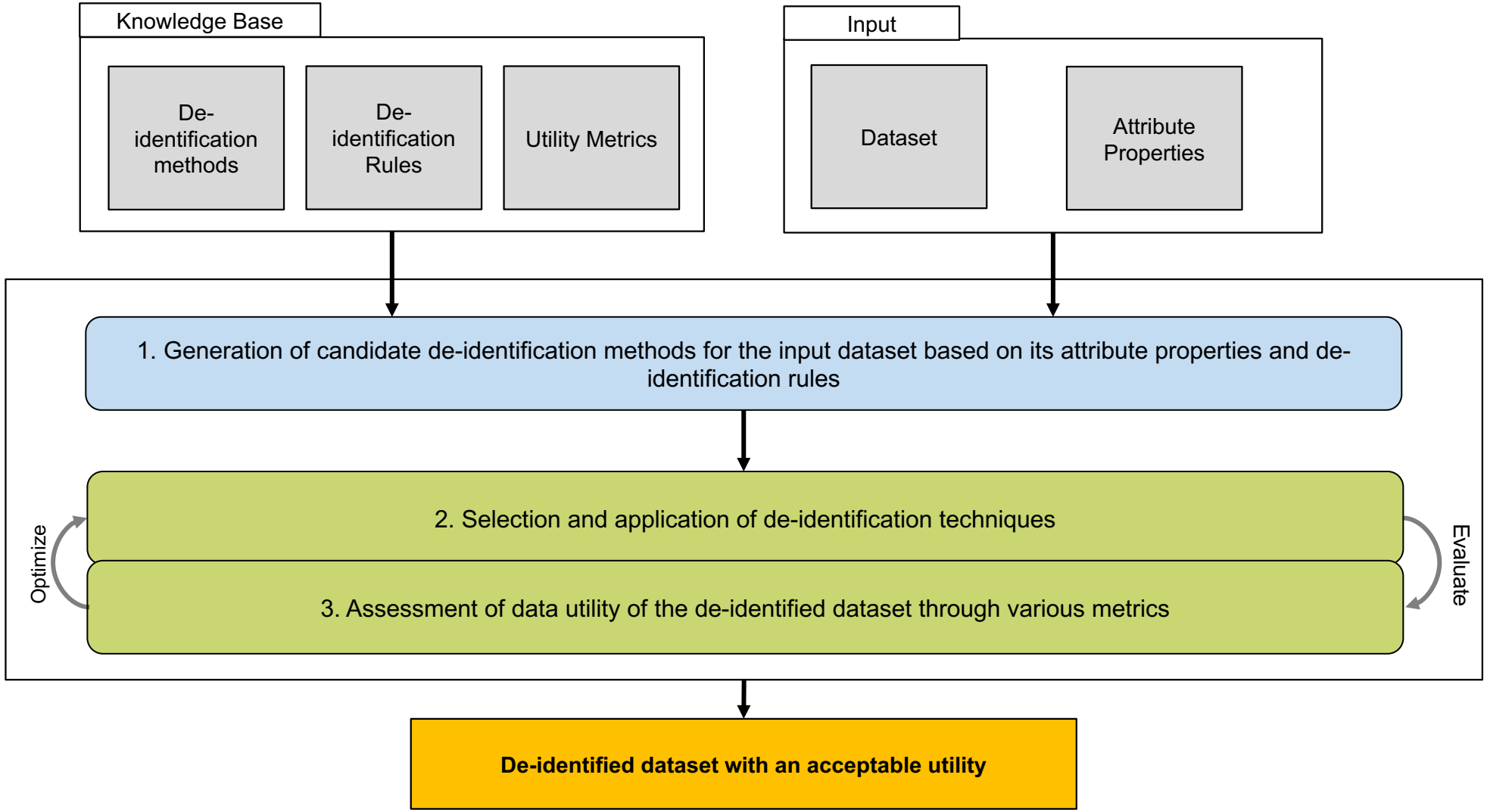
- As [1] highlights, a careful match of data utility metrics and de-identification methods is essential to extract meaningful insights

Utility Metric	Best suited de-identification Technique	Worst suited de-identification Technique
Normalized Averag Equivalence Class Size Metric	Aggregation, Rounding	Suppression, Pseudoanonymization

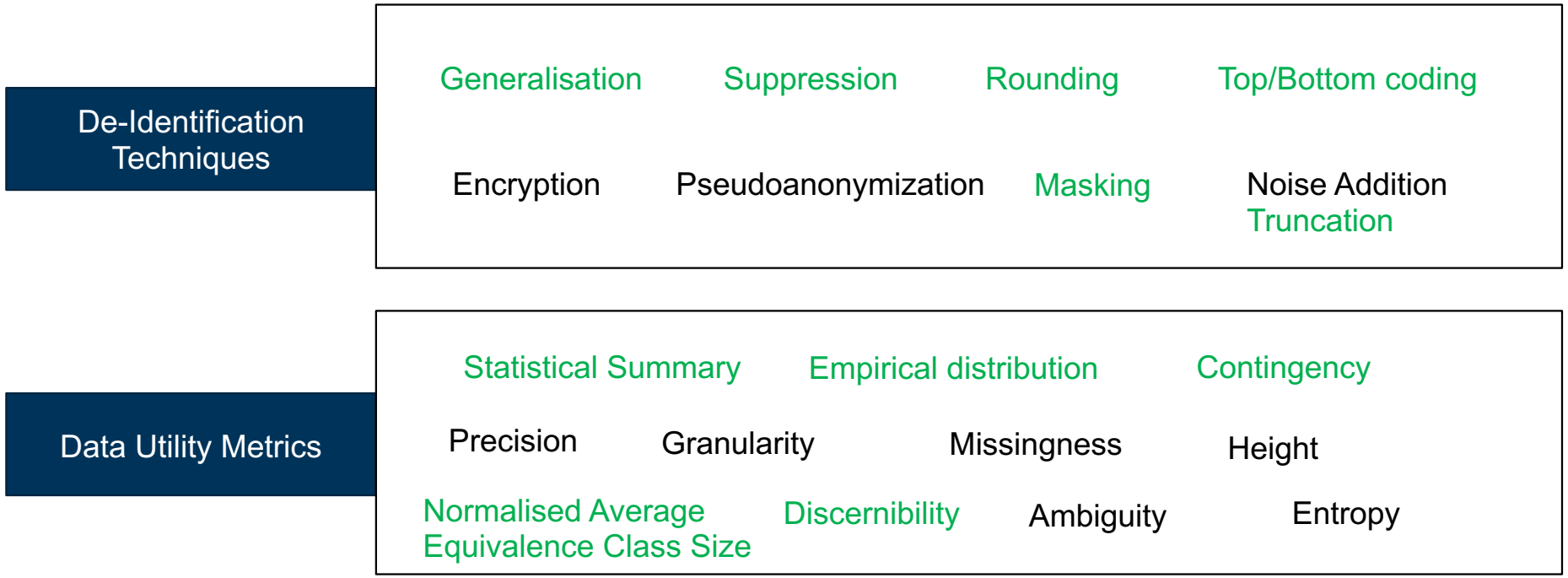
Normalized Average Equivalence Class Size Metric
Estimates data quality by calculating the average size of classes of indistinguishable records

[1] Tomashchuk, Oleksandr, et al. "A Data Utility-Driven Benchmark for De-identification Methods." *International Conference on Trust and Privacy in Digital Business*. Springer, Cham, 2019.

6. Optimised Data De-Identification Process



7. Example: Implemented De-identification Techniques and Utility Metrics



7. Example: Input

Mobility_Dataset

Vehicle_Number	Trip_ID	Timestamp	Speed	Latitude	Longitude
VN_101	TJ_456	2016-03-14 17:24:55	124	40.738563537597656	-73.964630126953125
VN_102	TJ_556	2016-03-15 07:34:32	82	43.864322246899875	-70.997653111345679
VN_101	TJ_457	2016-03-14 17:27:00	100	40.738563577898765	-73.964630167888987

Attribute Properties

name	identifier_type	value_type
Vehicle_Number	direct	Nominal
Trip_ID	indirect	Nominal
Timestamp	indirect	Continuous
Speed	indirect	discrete
Latitude	indirect	Continuous
Longitude	indirect	Continuous

7. Example: Candidate Method Generation

Attributes			
name	**identifier_type**	**value_type**	**Candidate De-Identification Methods**
Vehicle_Number	direct	Nominal	{Suppression, Pseudoanonymization}
Trip_ID	indirect	Nominal	{Encryption, Pseudoanonymization}
Timestamp	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}
Speed	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}
Latitude	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}
Longitude	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}

7. Example: De-Identification Method Selection

Attributes				
name	**identifier_type**	**value_type**	**Candidate De-Identification Methods**	**Parameter Tuning**
Vehicle_Number	direct	Nominal	{Suppression, Pseudoanonymization}	-
Trip_ID	indirect	Nominal	{Encryption, Pseudoanonymization}	-
Timestamp	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}	Direction= Right, No. of characters= 8
Speed	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}	Step Size = 10
Latitude	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}	No. of decimal points = 3
Longitude	indirect	Continuous	{Generalisation, Noise Addition, Rounding, Masking}	No. of decimal points = 3

7. Example: Application of De-Identification Techniques

Data De-Identification

Mobility_Dataset

Vehicle_ Number	Trip_ID	Timestamp	Speed	Latitude	Longitude
VN_101	TJ_456	2016-03-14 17:24:55	124	40.738563537597656	-73.964630126953125
VN_102	TJ_556	2016-03-15 07:34:32	82	43.864322246899875	-70.997653111345679
VN_101	TJ_457	2016-03-14 17:27:00	100	40.738563577898765	-73.964630167888987

1. De-identify

Mobility_Dataset_Deidentified

Trip_ID	Timestamp	Speed	Latitude	Longitude
T5GHI6_zOp956	2016-03-14	120-130	40.739	-73.960
NjuRzI678_9wer	2016-03-15	80-90	43.865	-70.998
T5GHI6_zOp956	2016-03-14	100-200	40.739	-73.965

2. Evaluate

3.Optimize

Utility Analysis

Attribute Oriented

	Vehicle_Number	Trip_ID	Timestamp	Speed	Latitude	Longitude
Precision	?					
Granularity						
Missingness						
Height						
Min						
Max						
Mean						
Standard Deviation						
…						

Dataset Oriented

Normalized Average Equivalence Class Size	?
Ambiguity	
Discernibility	
Entropy	

8. Research Questions

RQ1: What is the state-of-the-art of data utility metrics and data de-identification tools?

RQ2: How could the implementation of an enterprise level data utility analysis tool look like?

RQ3: Given the feedback during the user testing, in what ways could the tool be improved?

9. Master Thesis

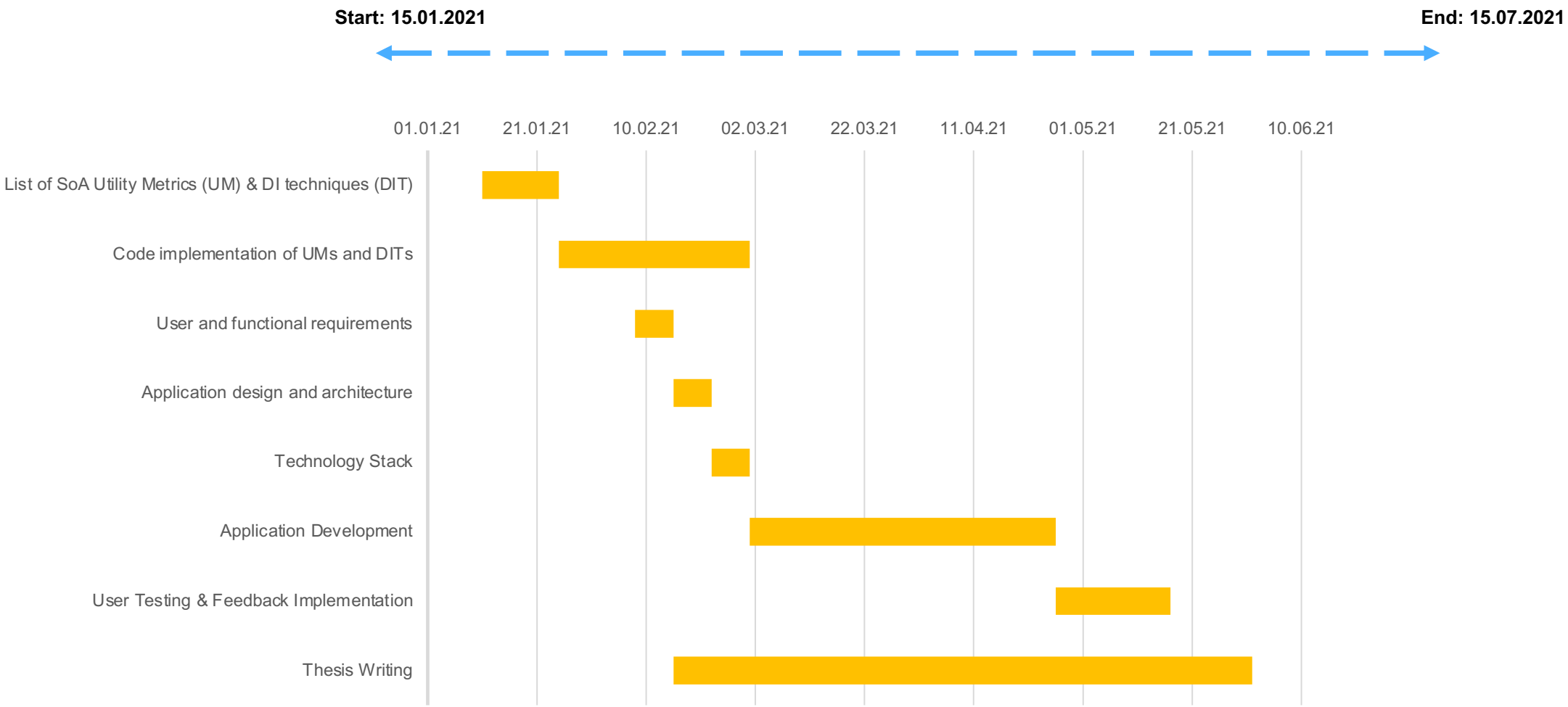
A. Exploration & Analysis of Utility Metrics (UMs) & De-Identification techniques (DITs)

a) List of state-of-the-art UMs & DITs
b) Code implementation of UMs and DITs

A. Development of Utility Analysis Tool

a) Application design and architecture
b) User and functional requirements
c) Technology stack
d) User feedback & tool improvements

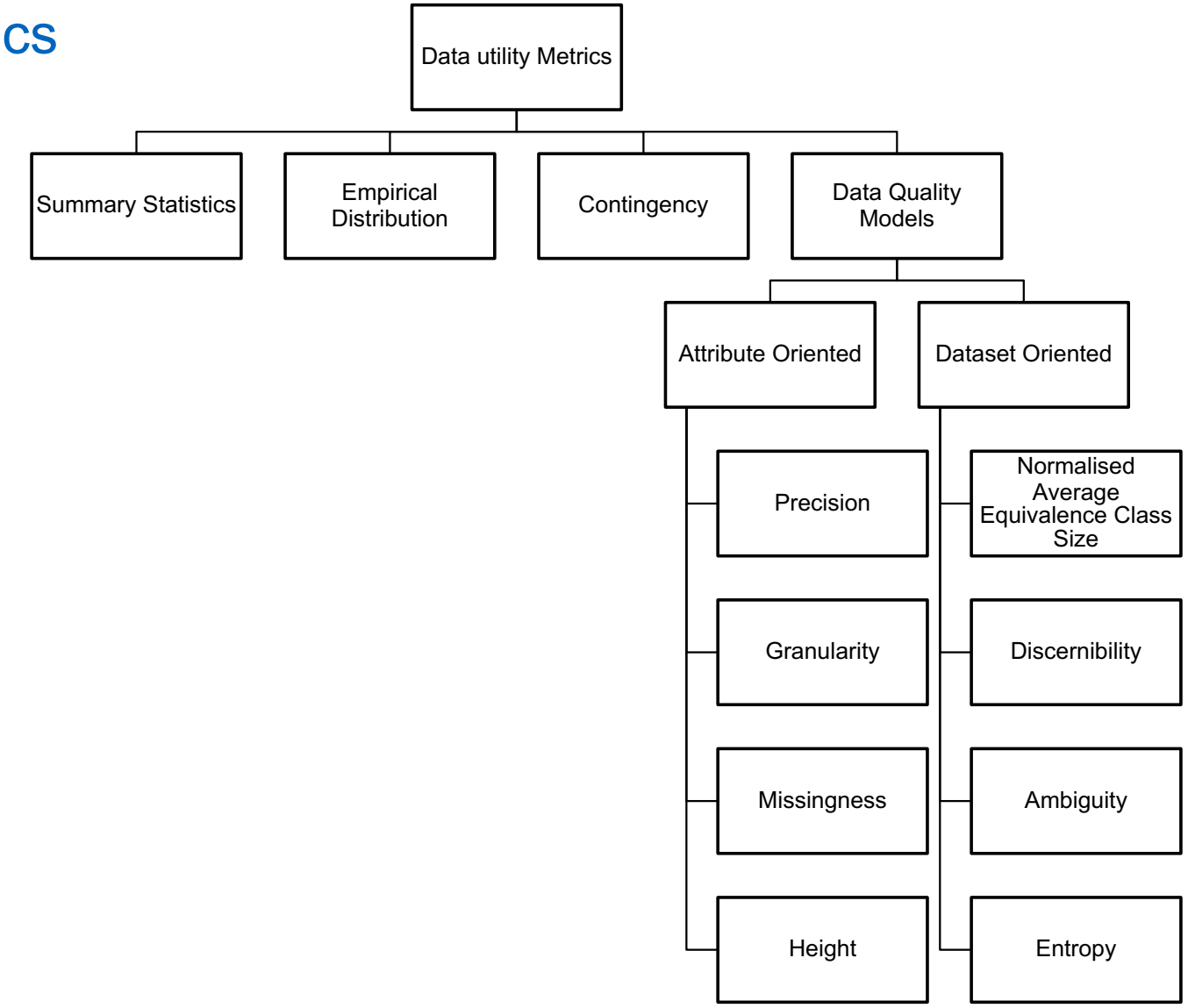
10.Timeline


Start: 15.01.2021
End: 15.07.2021
01.01.21
21.01.21
10.02.21
02.03.21
22.03.21
11.04.21
01.05.21
21.05.21
10.06.21
List of SoA Utility Metrics (UM) & DI techniques (DIT)
Code implementation of UMs and DITs
User and functional requirements
Application design and architecture
Technology Stack
Application Development
User Testing & Feedback Implementation
Thesis Writing

Thank You!

Back-Up

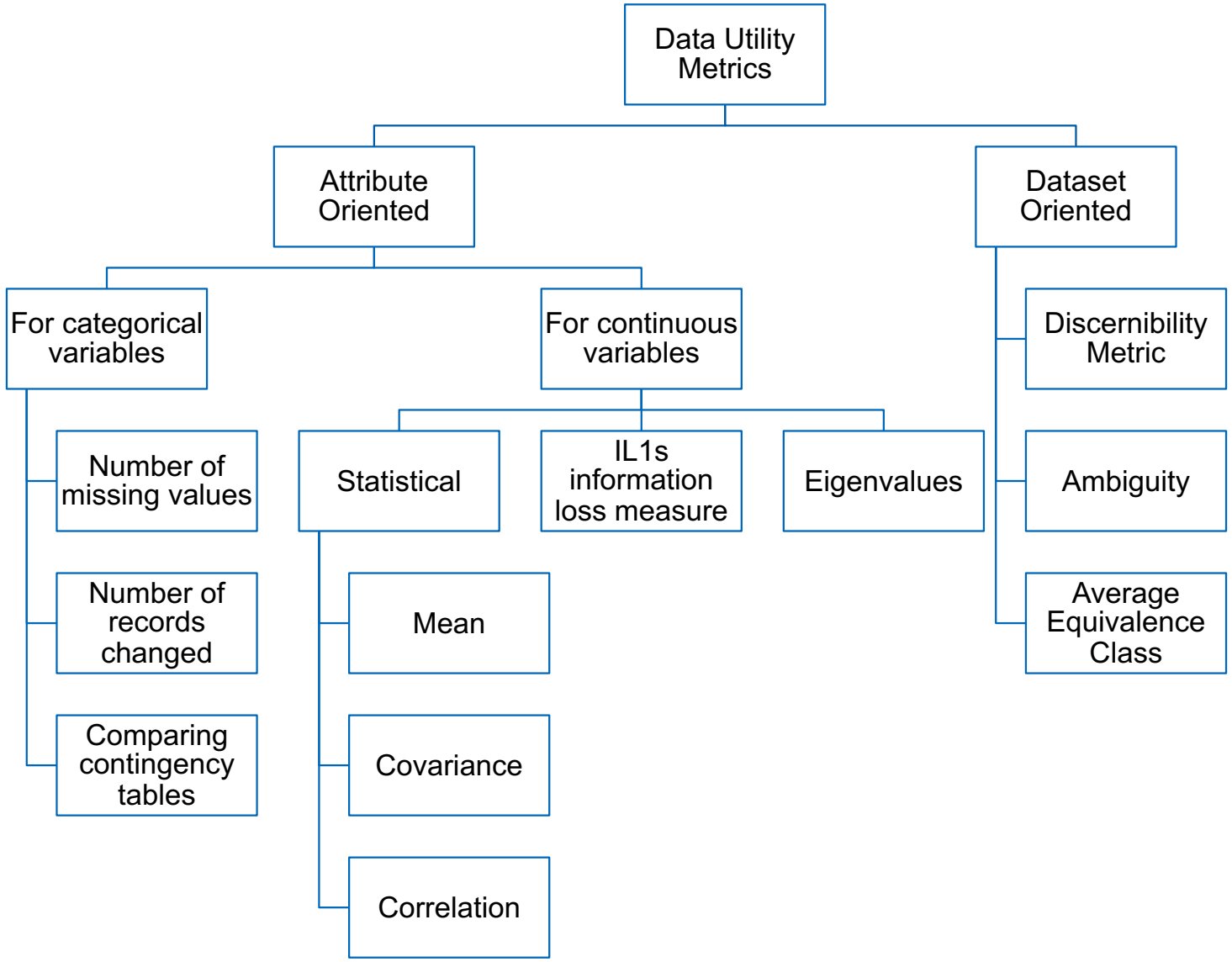
Data Utility Metrics



Utility Metrics

Utility Metric	Description
Summary Statistics	For a selected attribute, generate statistics (Range, Min, Max, Mean, Mode, Median, Variance, Standard Deviation, Geometric Mean etc.)
Empirical Distribution	For a selected attribute, visualize the frequency distribution of the values
Contingency	For two selected attributes, visualize the multivariate frequency distribution of the variables
Precision	Estimates data quality based on normalized generalization levels of transformed attribute values
Granularity	Measure summarizes the degree to which transformed attribute values cover the original domain of an attribute
Missingness	Percentage of number of missing values in the selected attribute
Height	Quantifies loss of information as the sum of the generalization levels applied to all attribute values
Normalised Average Equivalence Class Size	Estimates data quality by calculating the average size of classes of indistinguishable records
Discernibility	Estimates data quality based on the size of the equivalence classes in the output dataset.
Ambiguity	Quantifies the degree to which the records in the output dataset are ambiguous
Entropy	Measures differences in the distribution of attribute value

Data Utility Metrics



Data De-Identification: Properties [4][5]

Technique Name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of			Perturbative nature
				Singling out	Linking	Inference	
Statistical Tools							
Sampling	Yes	N/A	N/A	Partially	Partially	Partially	No
Aggregation	N/A	Continuous, discrete	All attributes	Yes	Yes	Yes	
Cryptographic Tools	Yes						
Deterministic encryption	Yes	All	All attributes	No	Partially	No	Yes
Order-preserving-encryption	Yes	All	All attributes	No	Partially	No	Yes
Homomorphic encryption	Yes	All	All attributes	No	No	No	Yes
Homomorphic secret sharing	Yes	All	All attributes	No	No	No	Yes

Data De-Identification: Properties

Technique Name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of			Perturbative nature
				Singling out	Linking	Inference	
Suppression	Yes						No
Masking	Yes	Categorical	Local Identifiers	Yes	Partially	No	No
Local Suppression	Yes	Categorical	Identifying attributes	Partially	Partially	Partially	No
Record Suppression	Yes	N/A	N/A	Partially	Partially	Partially	No
Pseudo anonymization	Yes	Categorical	Direct identifiers	No	Partially	No	
Generalisation	Yes	All, subject to meaning	Identifying attributes				No
Rounding	Yes	Continuous	Identifying attributes	No	Partially	Partially	No
Top/Bottom coding	Yes	Continuous, ordinal	Identifying attributes	No	Partially	Partially	No

Data De-Identification: Properties

Technique Name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of			Perturbative nature
				Singling out	Linking	Inference	
Randomization	No		Identifying attributes				Yes
Noise Addition	No	Continuous	Identifying attributes	Partially	Partially	Partially	Yes
Permutation	No	All	Identifying attributes	Partially	Partially	Partially	Yes
Micro aggregation	No	Continuous	Indirect Identifiers, and all other attributes	No	Partially	Partially	Yes

[4] Gloria Bondel et al. “Towards a Privacy-Enhancing Tool Based on De-Identification Methods.” In:PACIS.2020, p. 157.

[5] ISO/IEC 20889:2018.Privacy enhancing data de-identification terminology and classification of techniques. Standard. Geneva, CH: International Organization for Standardization, 2018.

10.Timeline

Start Date	End Date	Description	Duration (In days)
11.01.21	25.01.21	List of SoA Utility Metrics (UM) & DI techniques (DIT)	14
25.01.21	01.03.21	Code implementation of UMs and DITs	35
01.03.21	29.03.21	Experiments to match DITs & UMs	28
08.02.21	15.02.21	User and functional requirements	7
15.02.21	22.02.21	Application design and architecture	7
22.02.21	01.03.21	Technology Stack	7
01.03.21	26.04.21	Application Development	56
26.04.21	17.05.21	User Testing & Feedback Implementation	21
15.02.21	01.06.21	Thesis Writing	106