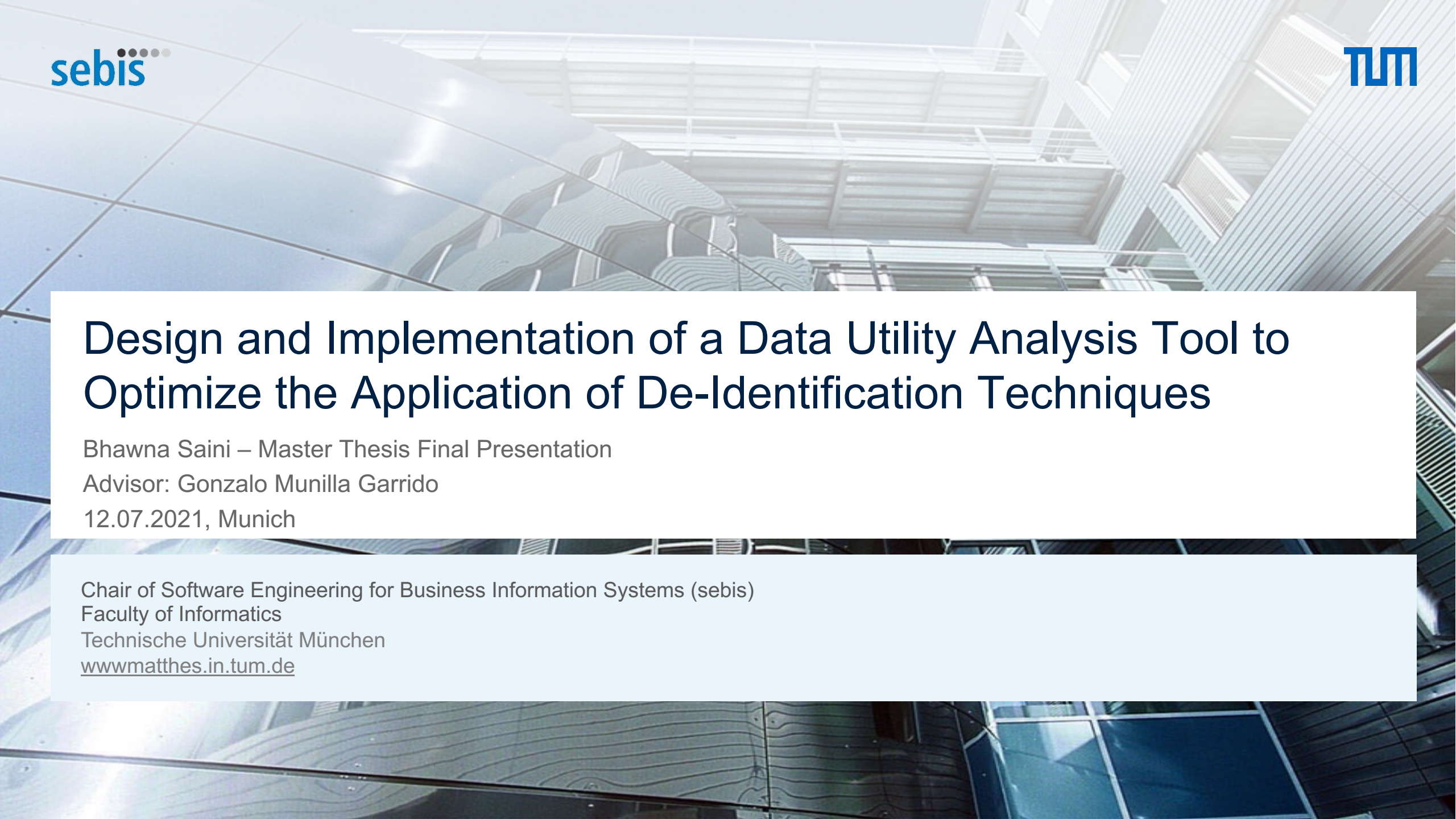


Design and Implementation of a Data Utility Analysis Tool to Optimize the Application of De-Identification Techniques

Bhawna Saini – Master Thesis Final Presentation
Advisor: Gonzalo Munilla Garrido
12.07.2021, 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 A Large Enterprise Setting
2. Overview of a Data-De-identification Process: Workflow & Status
3. Need of A Data Utility Assessment Process
4. Goal of the thesis
5. Design Of Data Utility Analysis Tool
6. Demonstration
7. Requirement Analysis
8. Component Diagram
9. Technology Stack
10. Evaluation
11. Conclusion, Limitations and Future Work

1. General Privacy Application Scenario In A Large Enterprise Setting

Product owner needs X data for a use case

Proposes use case to the data privacy department

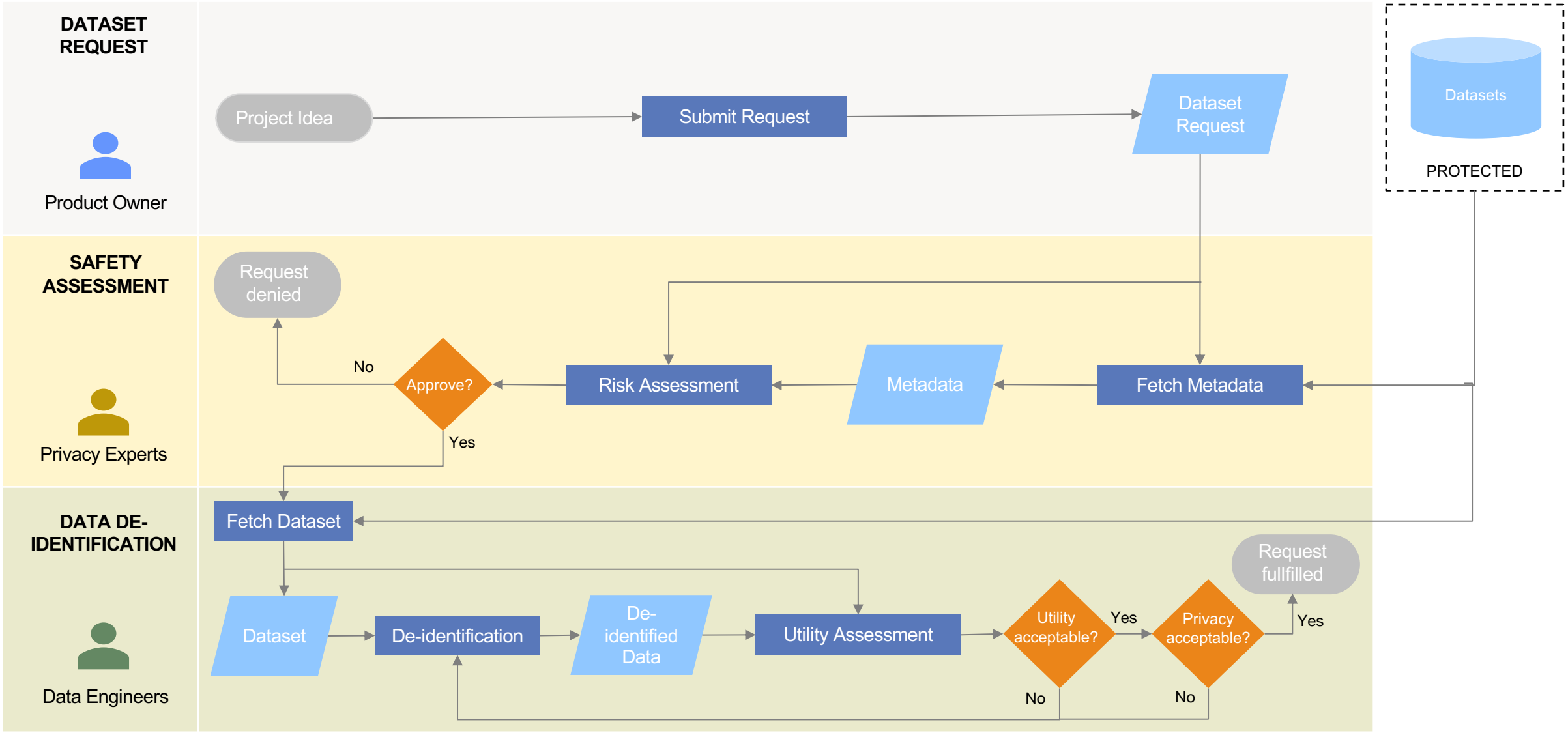
Privacy department performs risk assessment, sets privacy requirements & applies de-identification to comply with requirements.

Product owner can now use the dataset.

This step is manual, time intensive & susceptible to iterations.

- Could the request be approved & dataset be provided ?
- Is the privacy and utility of de-identified dataset acceptable for the use case?

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

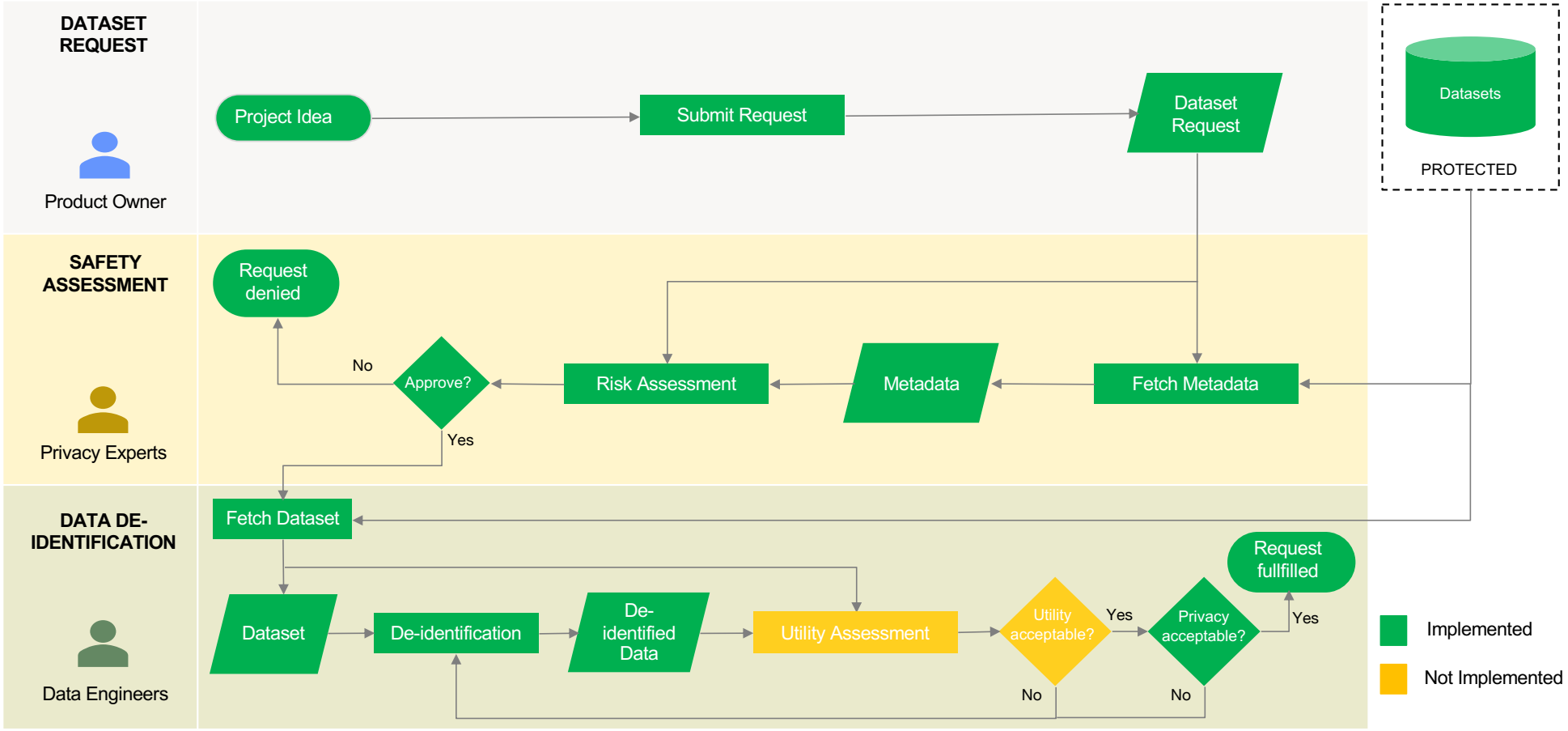


2. Overview of a Data-De-identification Process: Status

Manual Process

Rule Based

No Data Utility Assessment



3. Need of A Data Utility Assessment Process

1

How does one decide whether the resulting utility of dataset after de-identification is adequate for a given use case?

- Utility Metrics provide an overview of change in data utility due to the de-identification process [1]

2

How does one decide which set of de-identification techniques is suitable for the requested datasets in a way that the data utility is not significantly impacted?

- Different de-identification techniques affect the data differently with varying level of information loss [2]
- Utility Metrics help in understanding the implications of the de-identification techniques [1]

[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.

[2] Garfinkel, Simson L. "De-identification of personal information." *National institute of standards and technology* (2015).

4. Goal of the thesis

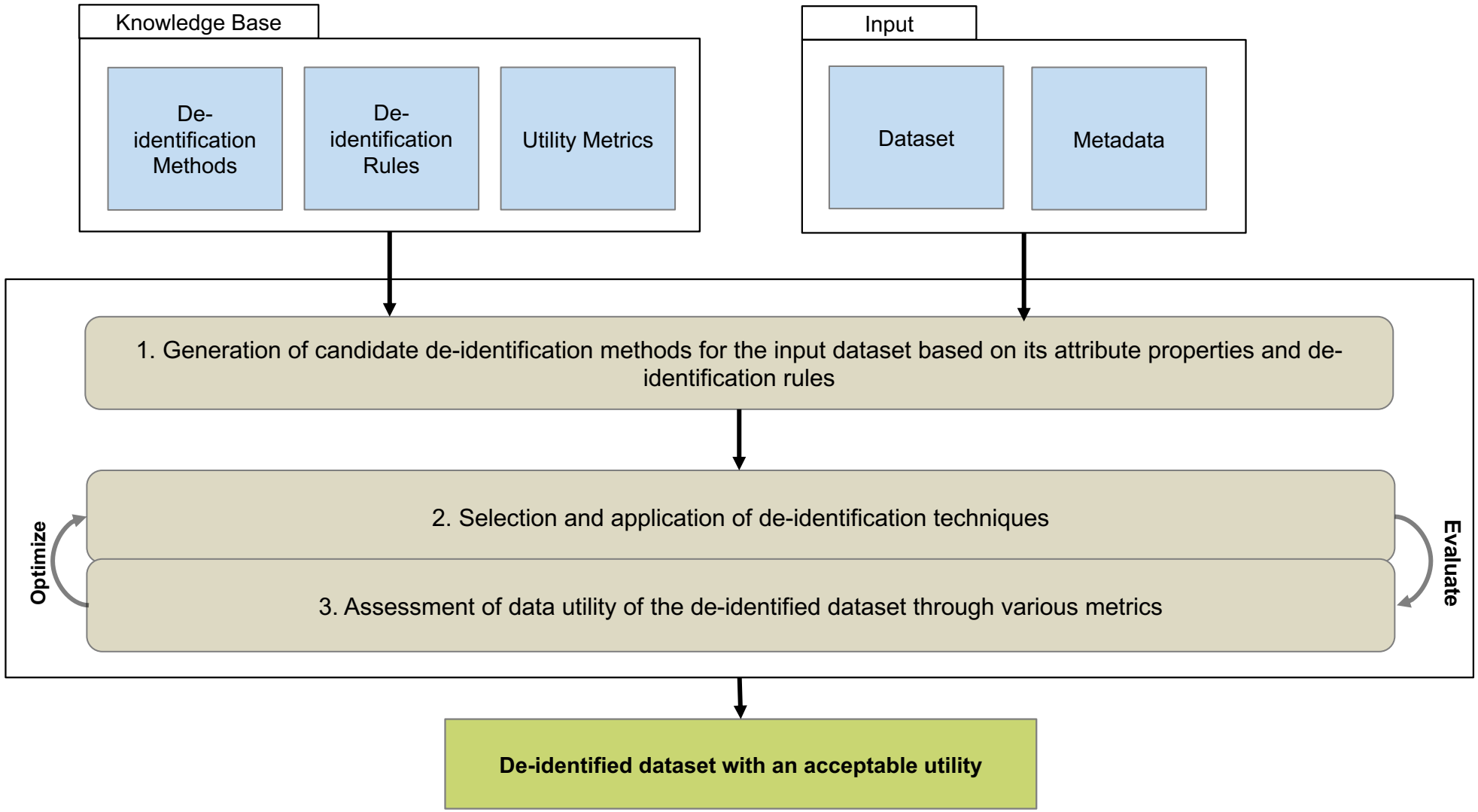
*Design and implement a **data utility analysis tool** that would help privacy experts to **quantify the performance** and **optimise the application** of de-identification techniques on datasets.*

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 application demonstration, in what ways could the tool be improved?

5. Design Of Data Utility Analysis Tool



6. Demonstration

7. Requirement Analysis

The tool should be able to process large amounts of data

Data de-identification and utility analysis should be done on server-side

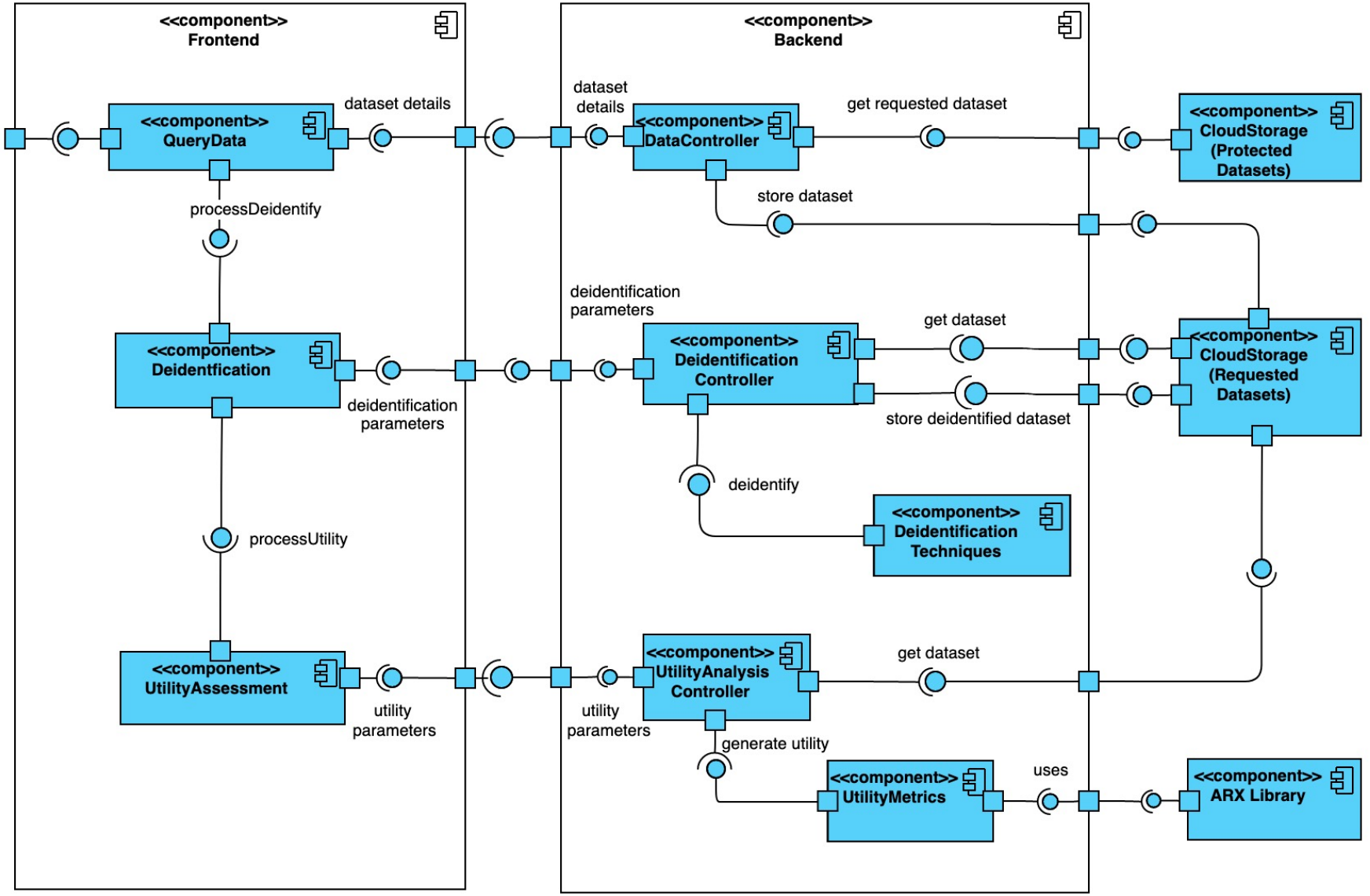
The application should support data to be uploaded from cloud storage

The data should not be allowed to be downloaded

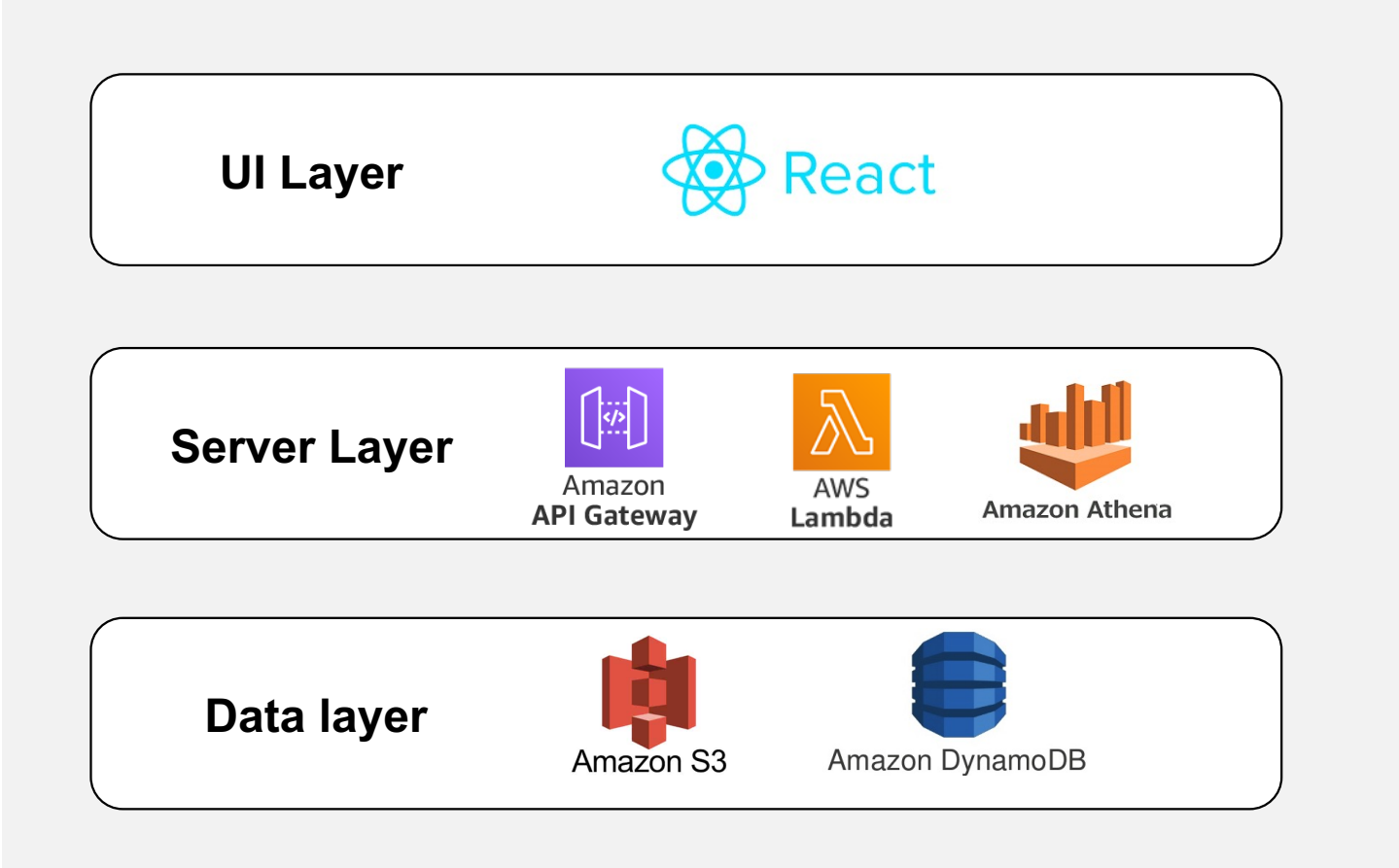
The de-identified data should be stored in the cloud

The tool offer easy to use interface

8. Component Diagram



9. Technology Stack



10. Evaluation

Evaluation Based on Requirements

The de-identified data should be stored in the cloud	
Data de-identification and utility analysis should be done on server-side	
The application should support data to be uploaded from cloud storage	
The data should not be allowed to be downloaded	
The tool should be able to process large amounts of data	
The tool offer easy to use interface	

Evaluation Based on Test with Dataset from Industry Partner

- Successful integration with industry partner's development environment
- Deployment using one command
- Application tested with automotive dataset

Size of dataset	*200MB*
Total number of rows	*1,458,644*
Total time to fetch data from cloud storage	*13 seconds*
Total Time to de-identify	*50 seconds*

Evaluation Based on Feedback

- Discovered use cases of the application
- Suggested application enhancements
- Identified potential research topics
- Tested application usability

10. Evaluation: Based on Feedback

Use Cases Discovered

1. The tool can help to define de-identification thresholds and standards
2. The tool can be used to build a public database of de-identified datasets

Application Enhancements

1. The application should support generalisation for the date/time data type.
2. The application should display description and units of attributes of a dataset
3. The application should support caching of dataset to reduce latency and costs

Research Topics

1. Determination of acceptable levels for data utility metrics
2. Automatic classification of attributes in a dataset based on their identifier types

Application Usability

1. Intuitive design
2. Ease of learning
3. Efficiency of use
4. Memorability
5. Error frequency
6. Subjective satisfaction

11. Conclusion, Limitations and Future Work

Conclusion

- Investigated the data de-identification process in a large enterprise setting.
- Identified and compliled a comprehensive list of data de-identification techniques and utility metrics.
- Designed and developed an enterprise level data utility analysis tool that helps to understand the implications of de-identification techniques
- Conducted a formal and in-depth evaluation of tool with experts in privacy and software domain

Limitations

- Not all de-identification techniques and utility metrics are implemented
- Every change in de-identification technique results in de-identification of the entire dataset

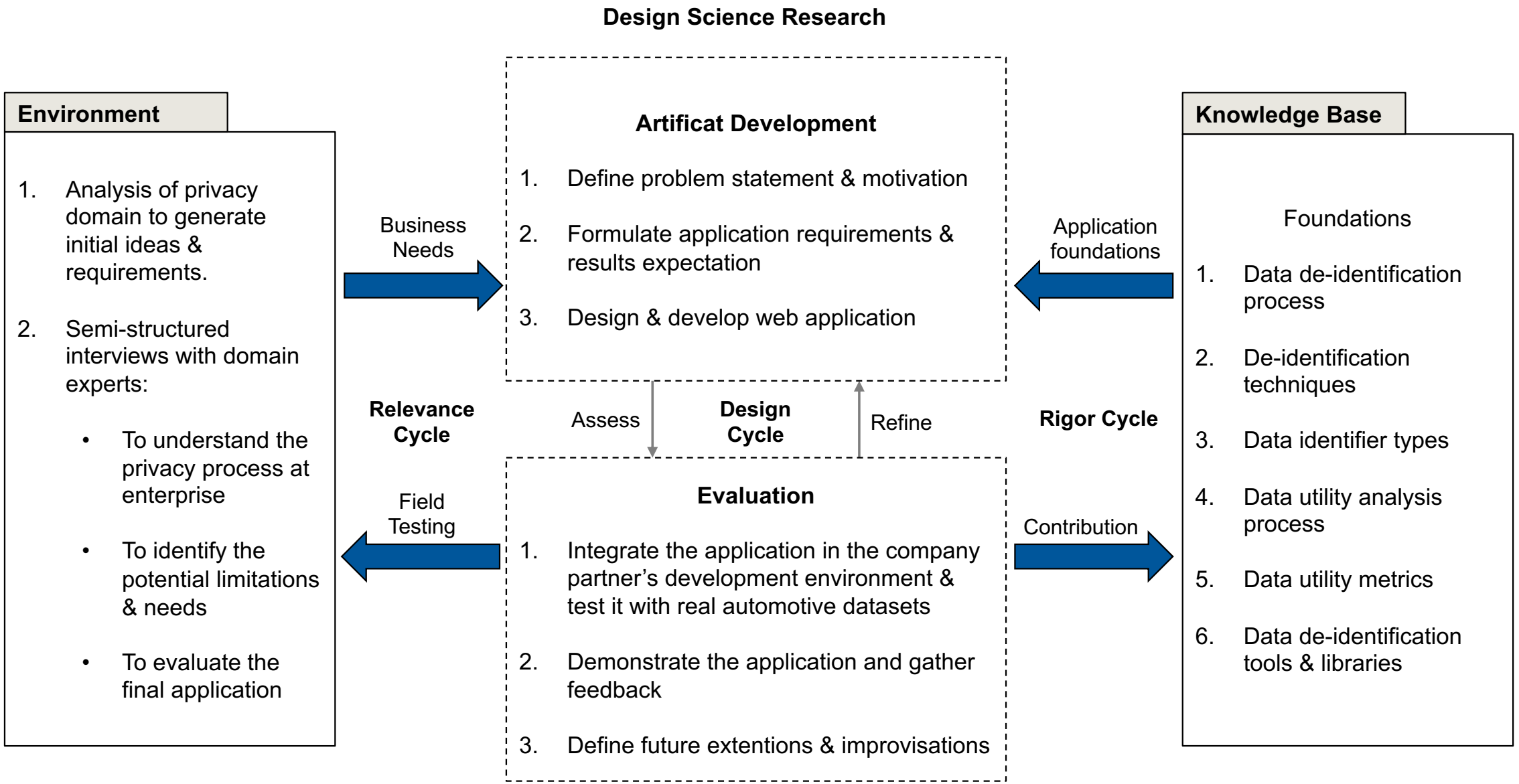
Future Work

- Support of the concept of domain generalisation hierarchy
- Automatic classification of direct identifiers
- Caching of datasets
- Display description of the attributes of datasets
- Define scientific scale that helps to determine acceptable level of utility metrics
- Integration of k-anonymity algorithm developed by Sharada Sowmya in her master thesis *“Implementaion of K-Anonymity for the Use Case of Automotive Industry in Big Data Context”*

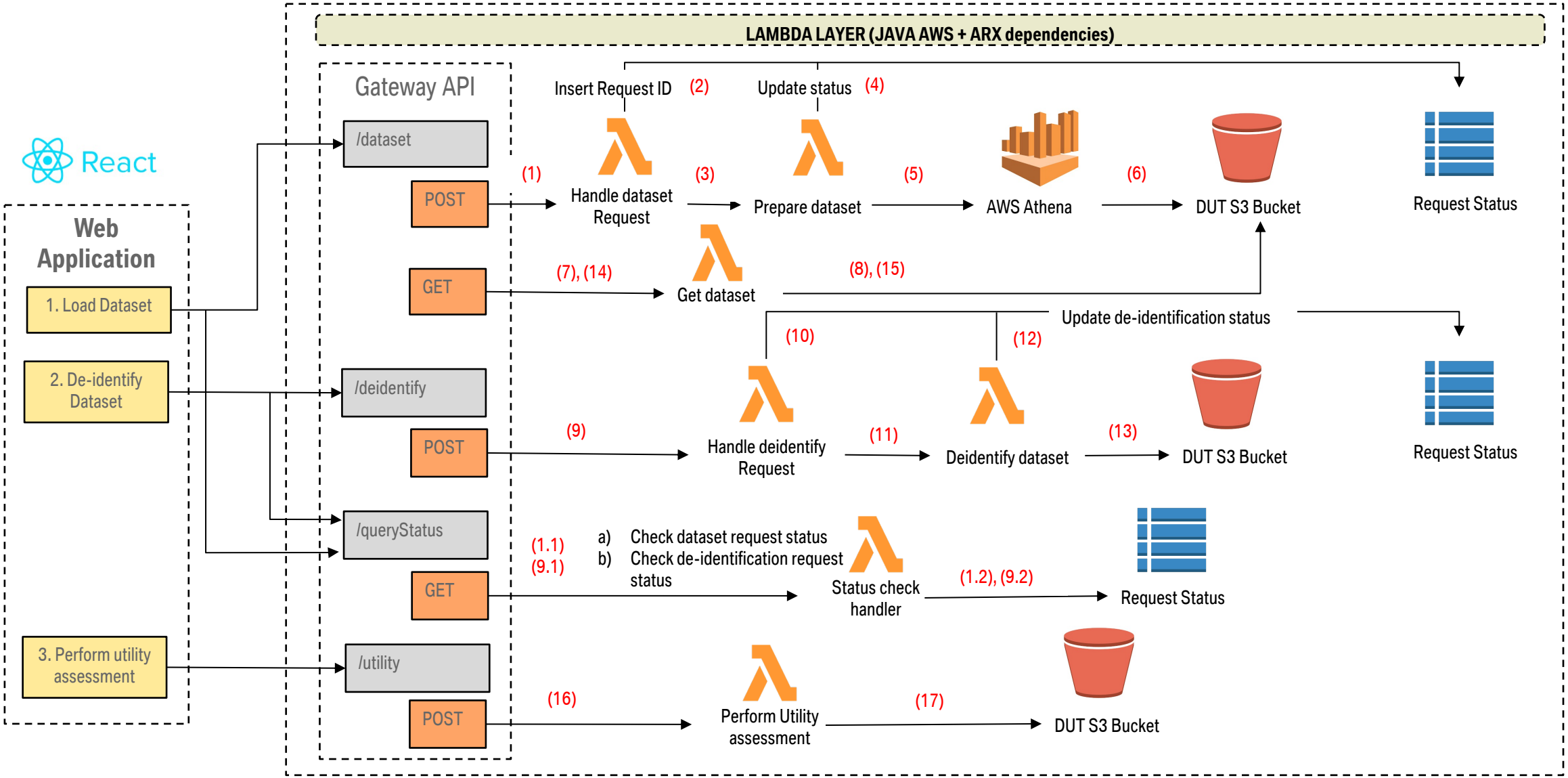
Thank you for your time!

Back-up Slides

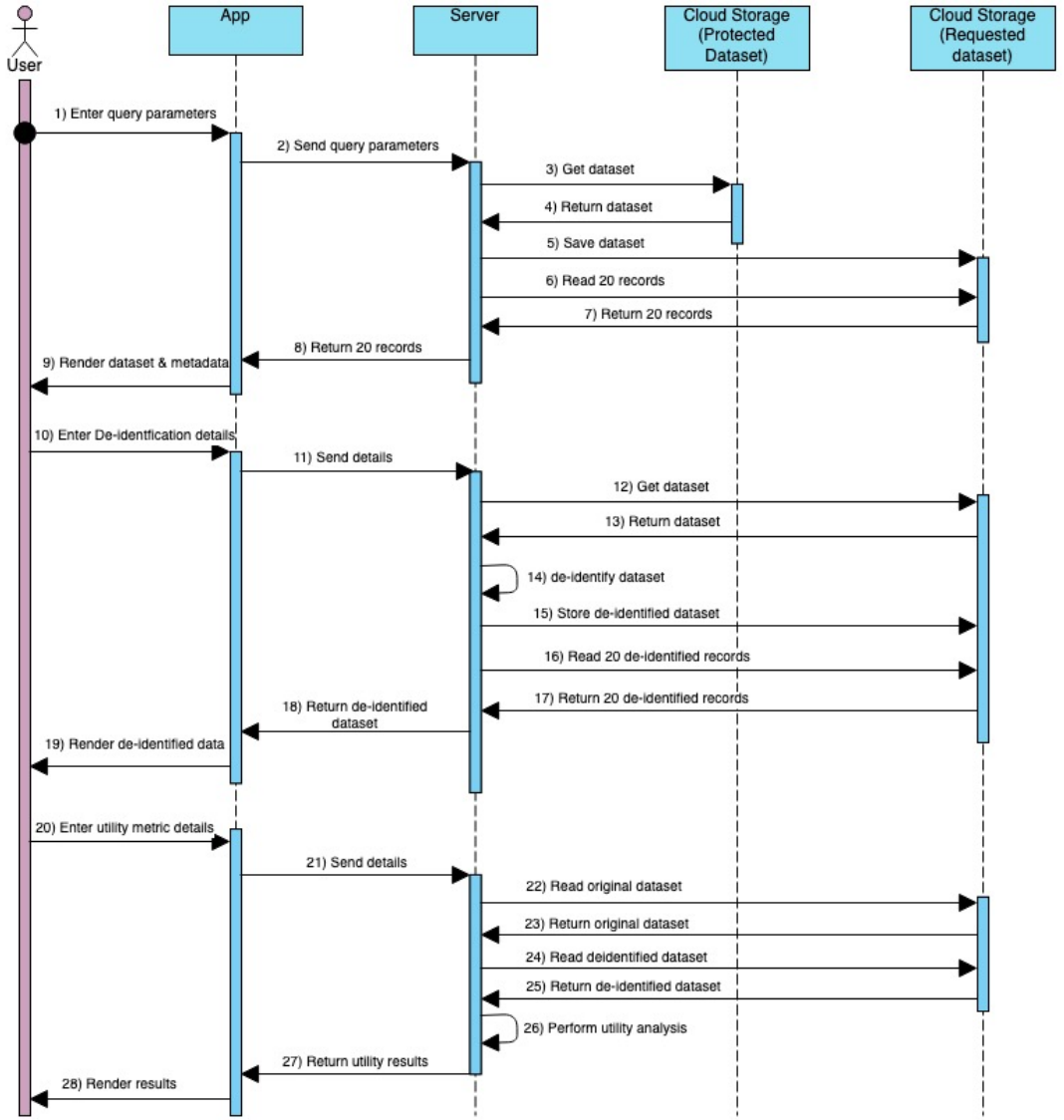
Research Methodology



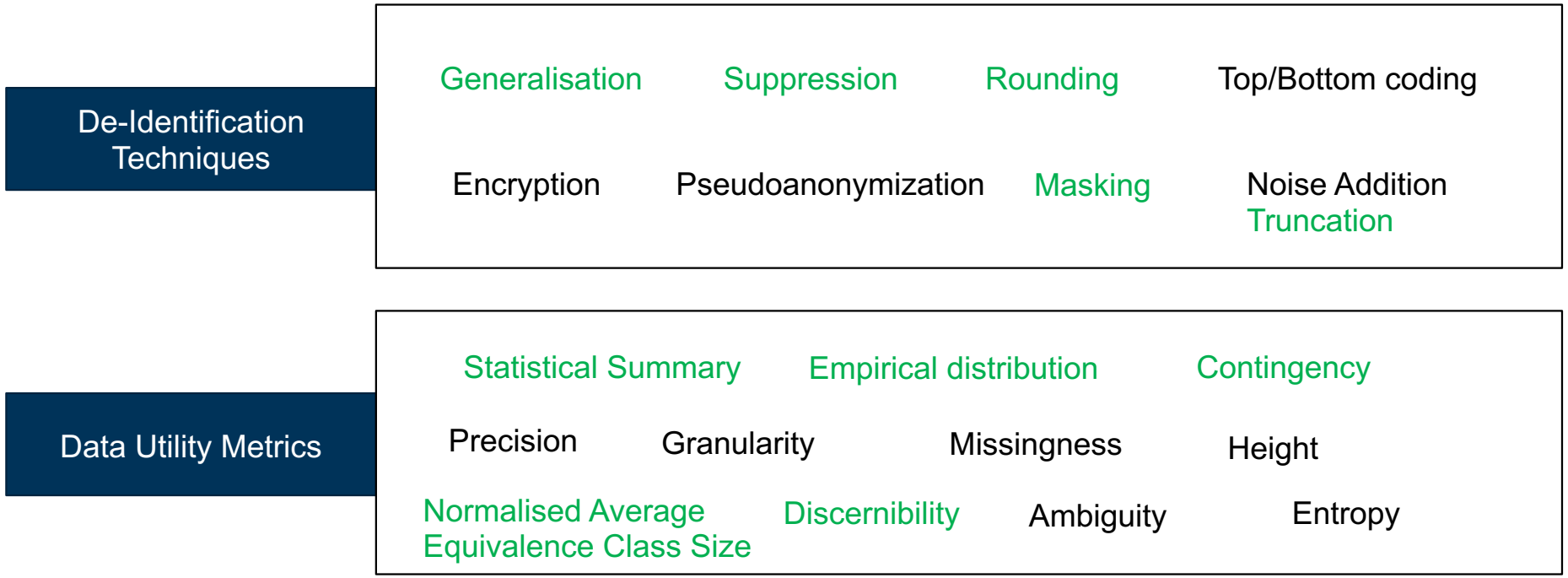
Technical Architecture



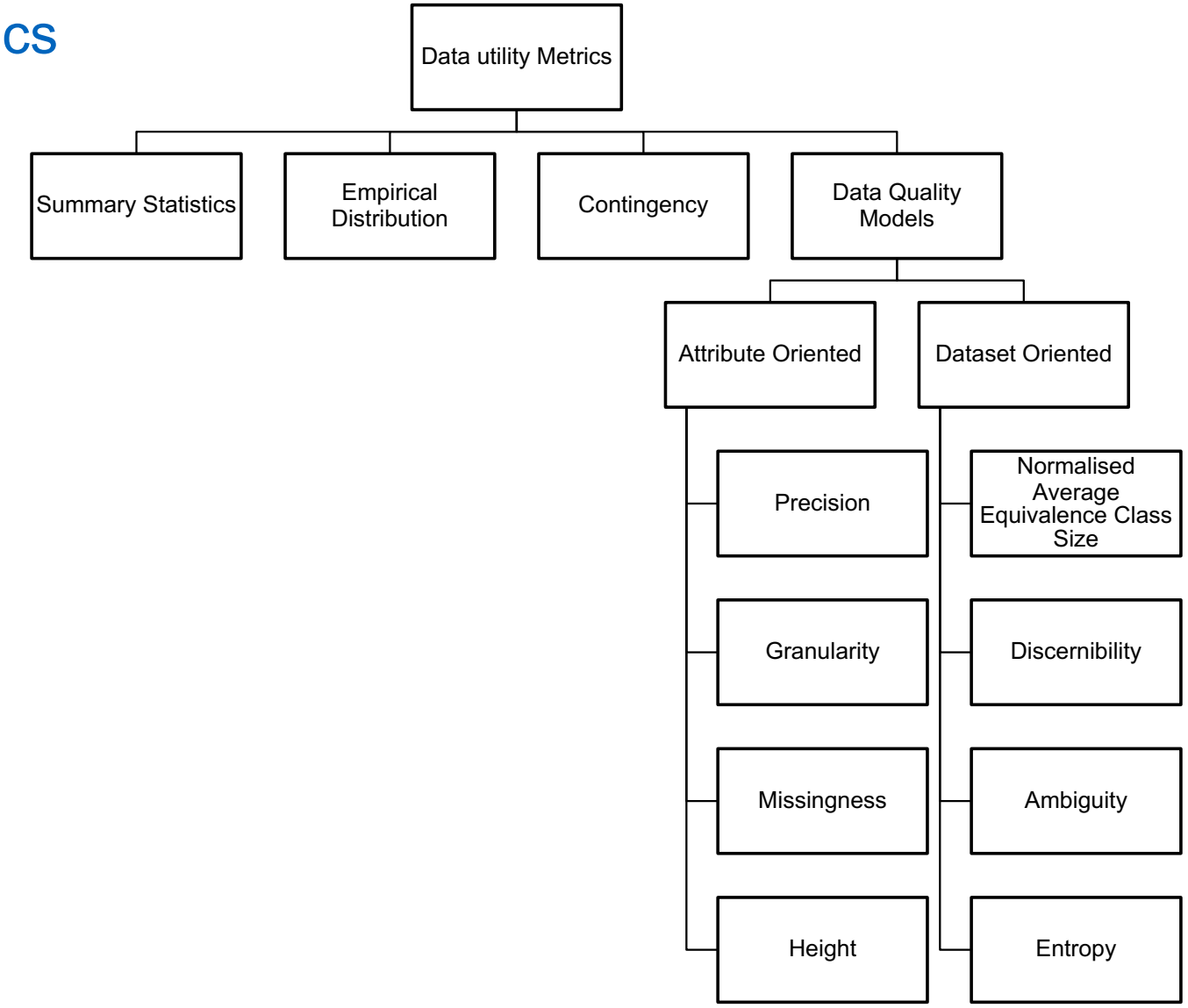
Sequence Diagram



Implemented De-identification Techniques and Utility Metrics



Data Utility Metrics



Discernibility Metric

- Penalizes the records in a dataset if they become **indistinguishable** from other records or suppressed

- Record is assigned penalty ‘e’ if it is indistinguishable from ‘e’ other records. The penalty of the entire group becomes e^2. If the record is suppressed, then it is assigned the penalty ‘D’, which is the size of the dataset.

- The total discernibility score is calculated as:

$$C_{DM} = \sum_{e \in E} |e|^2 + |S||D|$$

E: set of equivalence classes
D: dataset
S: set of suppressed records

- Assigns penalty based on equivalence class and not on actual information loss

- Disregards the distribution in the original dataset

Average Equivalence Class Size Metric

- Measures the average size of groups of indistinguishable records

- The metric is calculated as:

$$C_{AVG} = \left(\frac{total\ number\ of\ records}{total\ number\ of\ equivalence\ classes}\right)/(\mathrm{k})$$

k: minimum size of equivalence class

Like discernibility, does not ignore pre-existing equivalence classes

Utility Metrics

Implemented

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

Utility Metrics help to make a trade-off between minimizing re-identification risk & maximizing utility [3]

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

[3] Templ, Matthias. "Statistical disclosure control for microdata." *Cham: Springer International Publishing* (2017).

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.