Developing an Architecture for Data Quality Measurement
to achieve utility-driven Data Suppression*

Presenter: Sanchit Alekh (Matrikelnr: 359831)

Supervisors: Prof. Dr. Christoph Quix (RWTH Aachen), Prof. Dr. Stefan Decker (RWTH Aachen)

Advisors: Christoph Kreibich (AUDI AG), Dr. Sandra Geisler (Fraunhofer FIT)

*In co-operation with the AUDI AG

RWTH 5 | RWTH AACHEN UNIVERSITY

Agenda

Problem Statement

Typical Data Flow

Case Study: Process Data at Audi

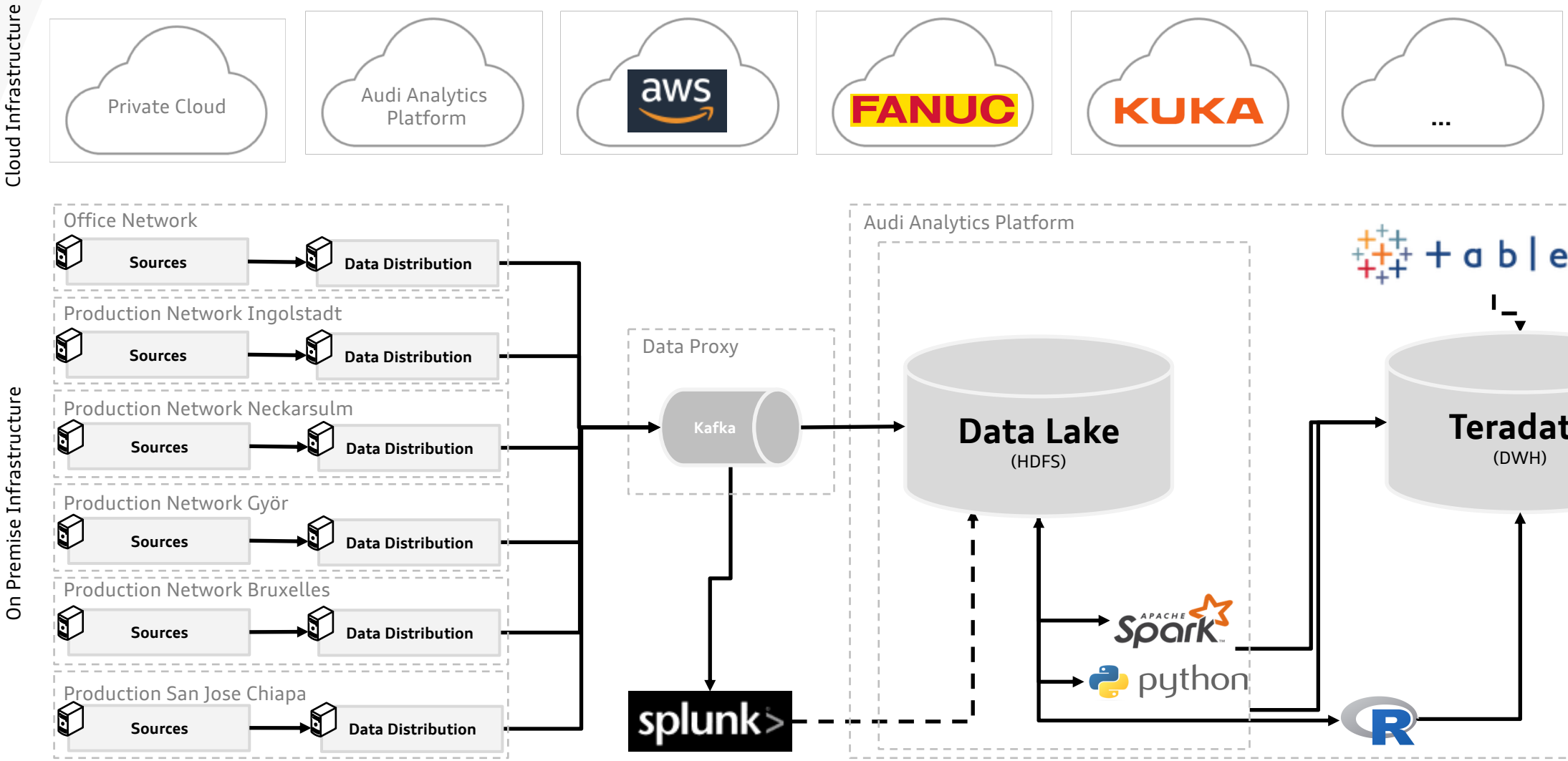
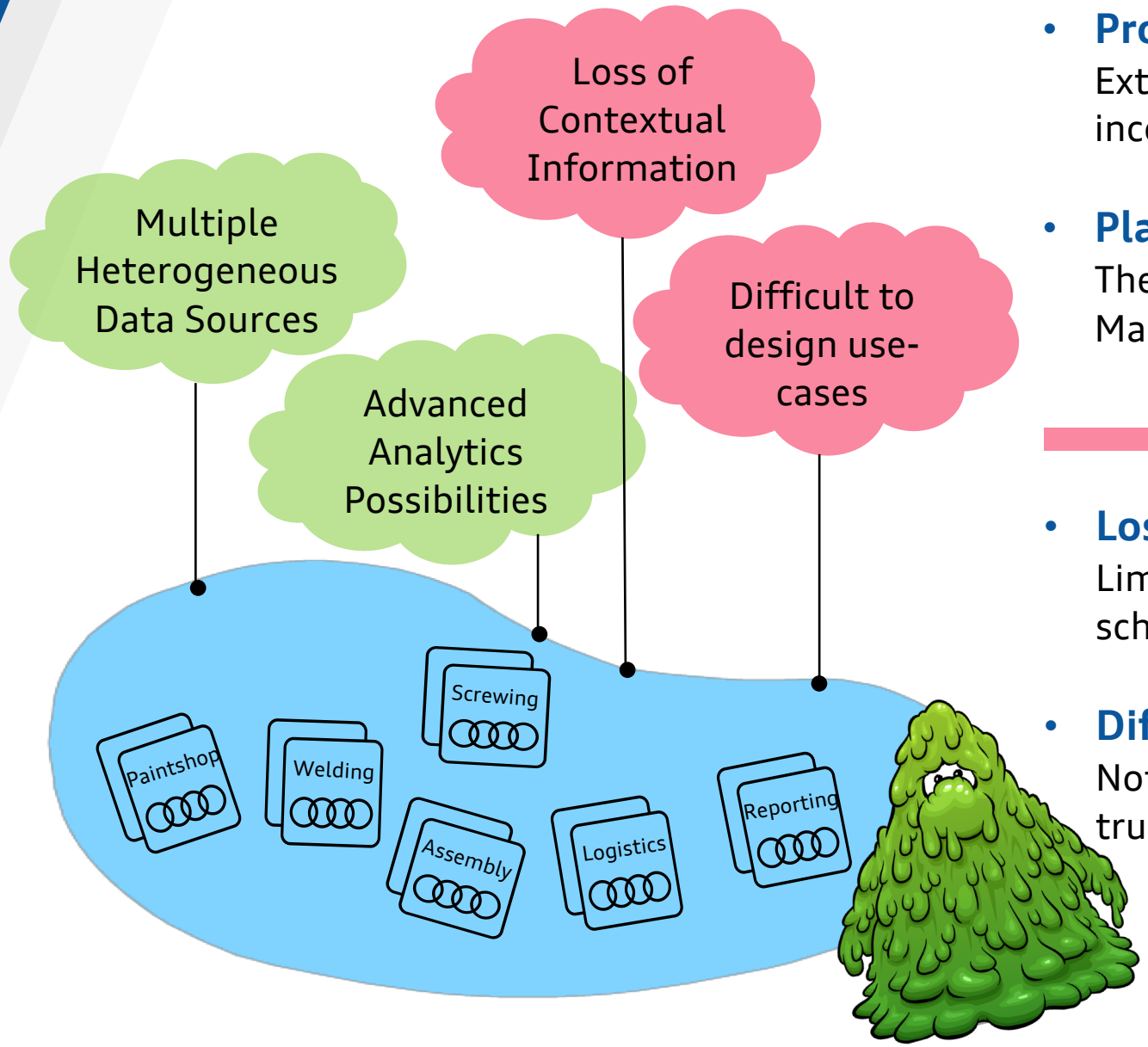


Image Source: AUDI AG

Data Lake
The Idea and the Conundrum

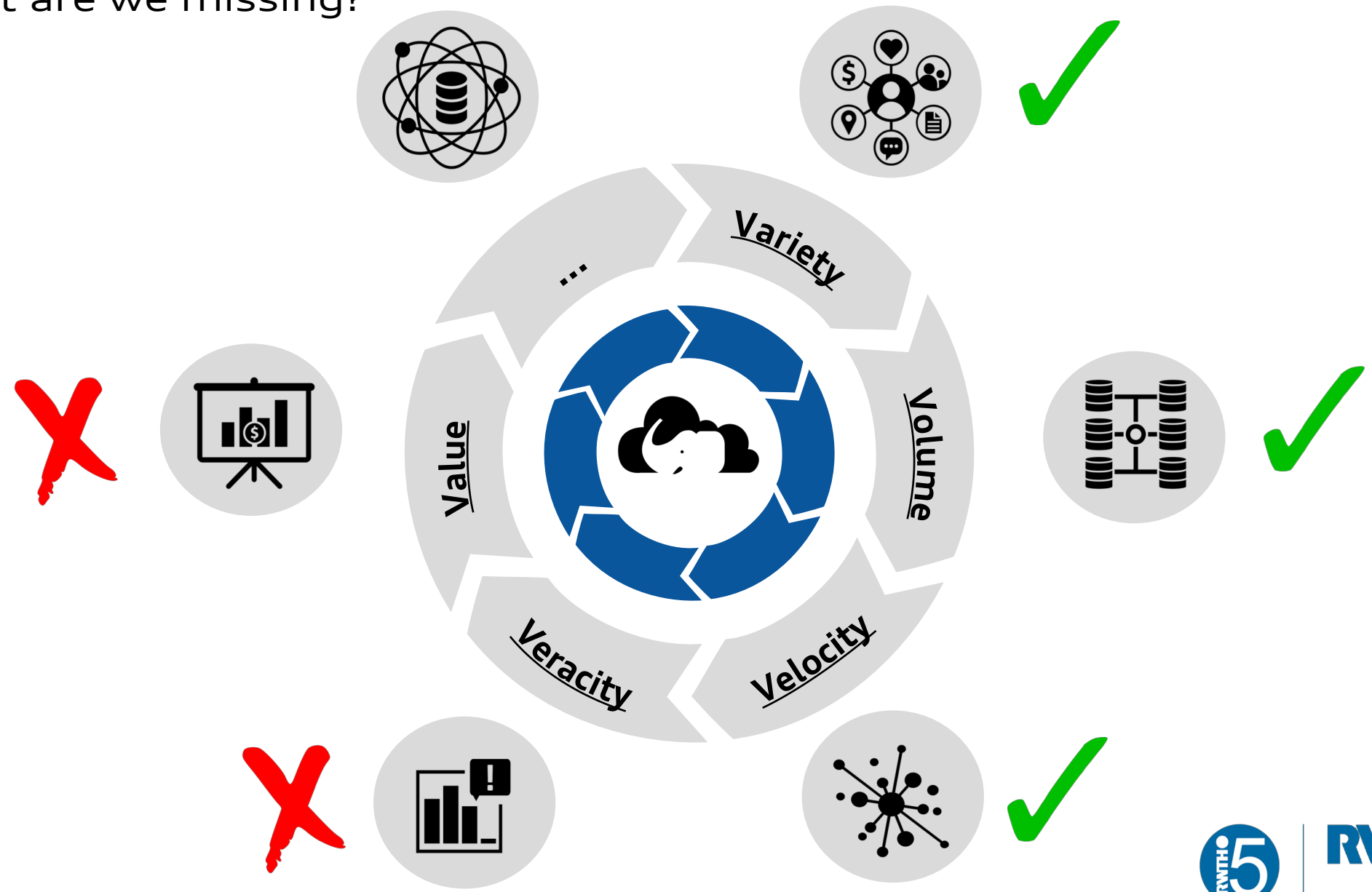


- **Provides a common sink for heterogeneous data**
  Extremely beneficial to bring together data from mostly incompatible data sources and make them available
- **Platform to run complex learning algorithms**
  The Hadoop ecosystem, with components such as Spark and MapReduce, is extremely powerful for advanced analytics

- **Loss of contextual information**
  Limited metadata available for the attributes, frequent schema changes and overcrowding
- **Difficult to design use-cases around the data**
  Not possible to determine the suitability and trustworthiness for a dataset for a particular use-case

The Vs of Big Data

What are we missing?



Data Sharing Scenario

Example Data Flow

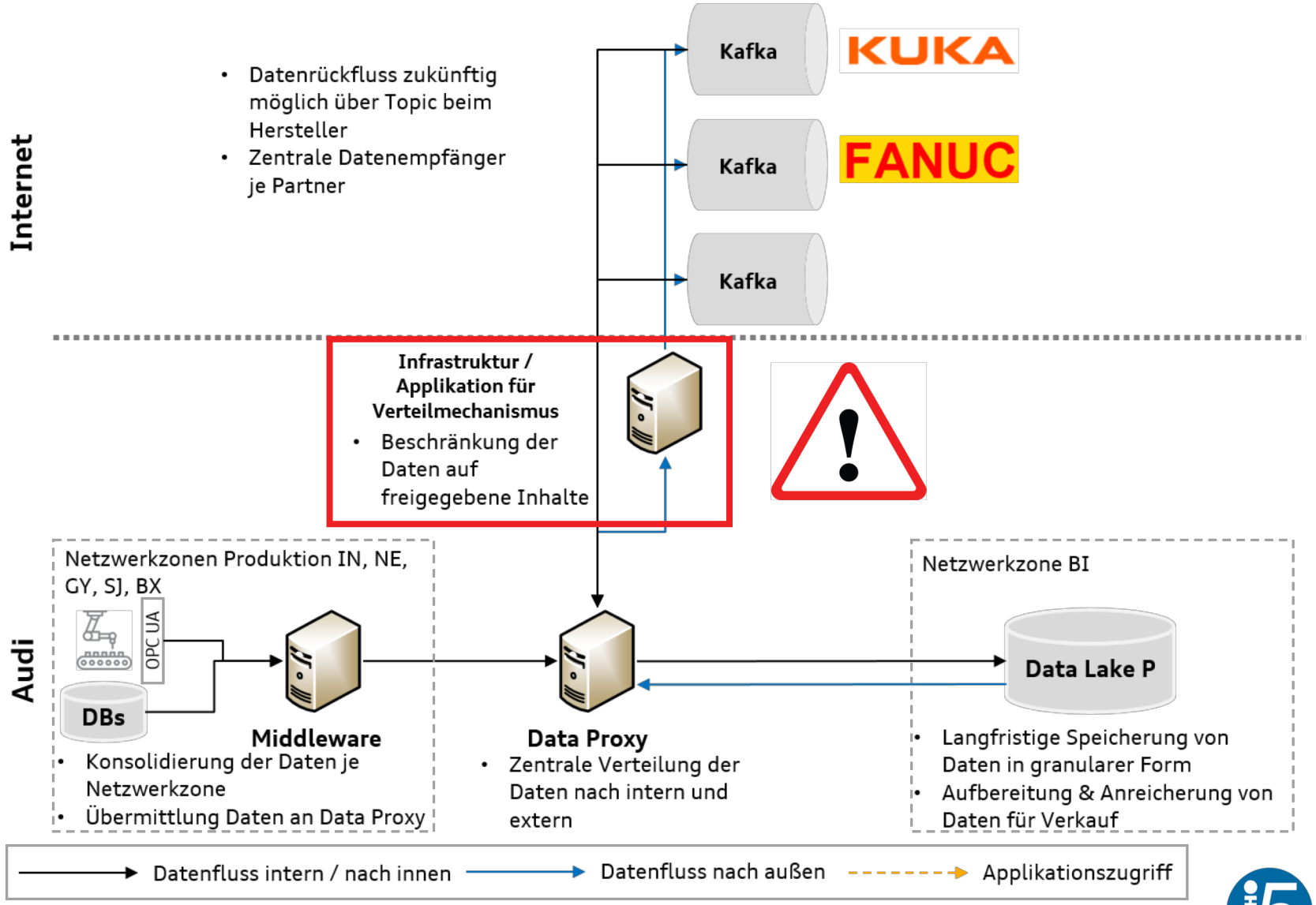
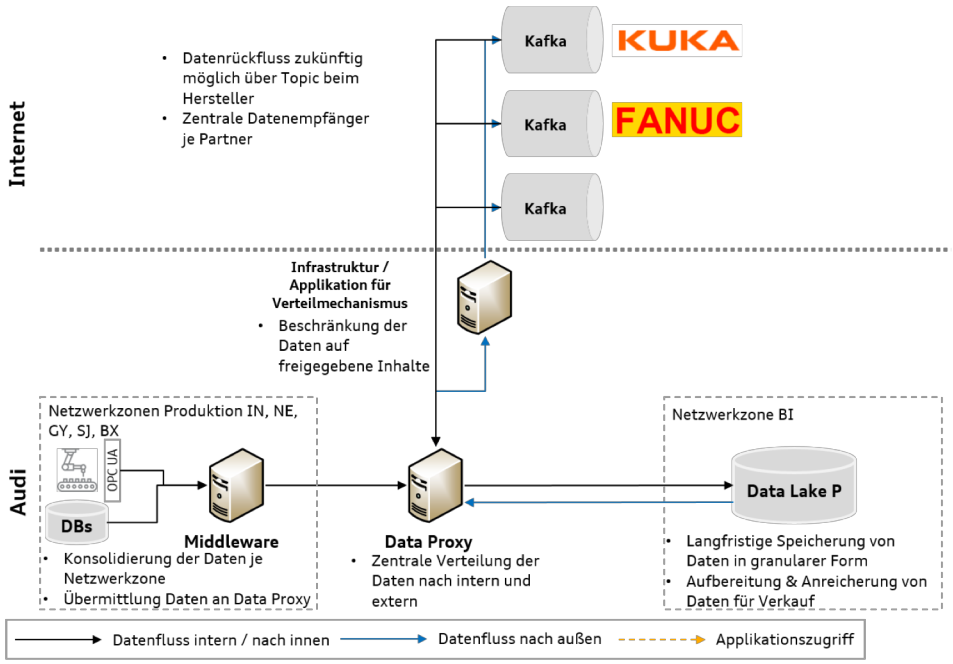


Image Source: AUDI AG

Data Sharing Scenario
Issues with Example Data Flow



- **No process to determine the quality of data**
  What value does the data contain for data mining? Is it *gold* or *garbage*?
- **No estimate of the privacy of the data**
  Is the data *safe* to be shared to third-parties? Does it reveal any *organizational secrets*?
- **0/1 choice between data usability and privacy**
  You either get access to a particular attribute, or you don't.
- **Lack of Transparency**
  The decision for sharing/not sharing an attribute is ad-hoc. This leads to mistrust and non-transparency

Image Source: AUDI AG

Goals of this Thesis

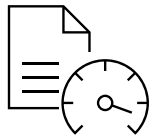
- An Approach for measuring Data Quality and Privacy
- Gain granular control over Privacy by using PETs
- Study the Relationship between Data Quality and Privacy
- Provide an effective mechanism for Metadata Management

Image Source: revo250 for The Noun Project

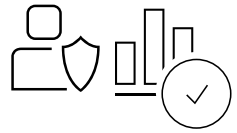
Methodology

Related Work

How has this problem previously been addressed?



- **Total Data Quality Management** [1]
  - › First general methodology for DQ Measurement
  - › Emphasized the importance of quality dimensions
  - › Vague, no guidelines for improvement
- **DQ Management for Data Lake Systems** [2]
  - › Developed for the Data Lake Architecture
  - › Focus on time-based analysis and visualization
  - › Need for a Metadata Repository is identified
  - › Usability aspects of data not taken into account
  - › Strictly a measurement and monitoring approach

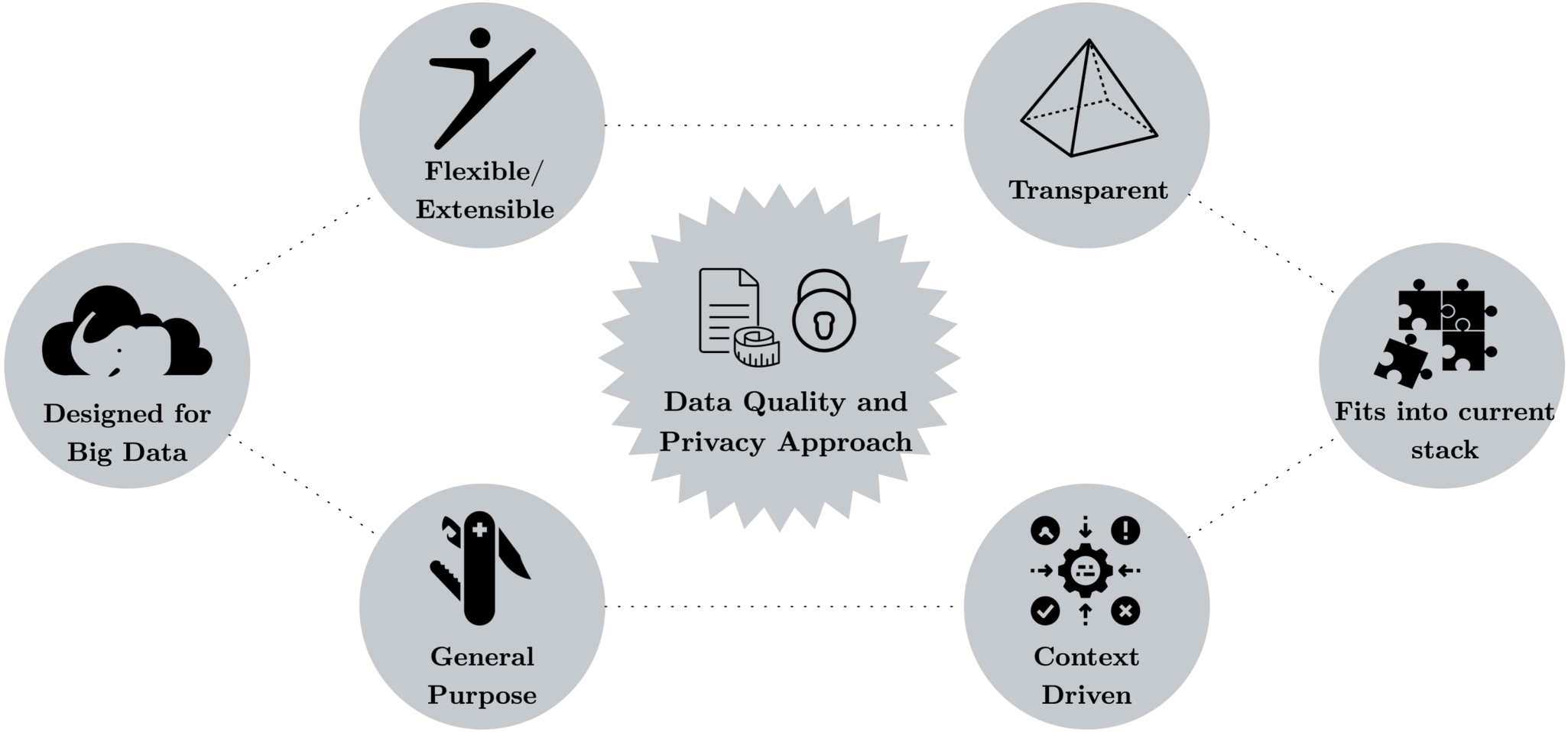


- **Anonymization, e.g., k-Anonymity** [3]
  - › Very popular and easily understandable
  - › Can't protect against background knowledge or homogeneity attacks
- **Differential Privacy** [4]
  - › Ensures plausible deniability, effective for PPDM
  - › Being used widely by companies such as Apple
  - › Determining global sensitivity and the corresponding ε is not trivial
- **Data Mining from Anonymized Data** [5]
  - › Possible to train classification algorithms on anonymized data with good accuracy

[1] **Richard Y Wang and Diane M Strong**. Beyond accuracy: What data quality means to data consumers
[2] **Alexandra Dalevskaya.** Data Quality Management for Data Lake Systems
[3] **Latanya Sweeney.** K-Anonymity: A Model for Protecting Privacy
[4] **Cynthia Dwork.** Differential Privacy: A Survey of Results
[5] **Ali Inan, Murat Kantarcioglu, Elisa Bertino.** Using Anonymized Data for Classification

Design Considerations

What qualities should an ideal solution possess?



Data Quality
How do we define it?

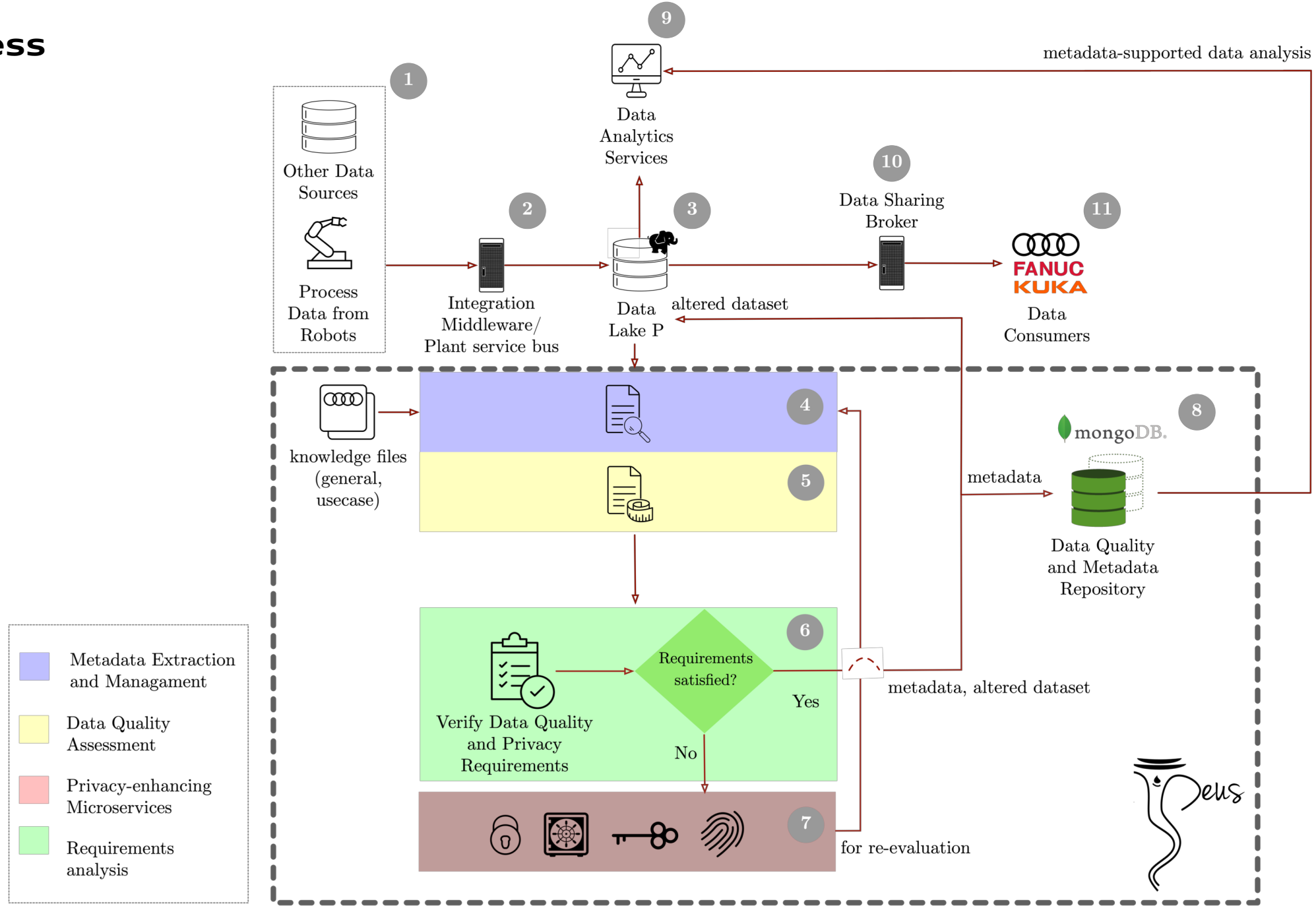
Quality = "Fitness for Use" [1,2]

- Automatically measurable attributes
- No or little pre-defined knowledge
- Objective, interpreted uniformly across use-cases

- Needs some background knowledge
- More subjective, can be interpreted in different ways
- Varies across use cases

[1] **Sandra Geisler.** A systematic evaluation approach for data-stream based applications. Retrieved from publica.fraunhofer.de
[2] **Richard Y Wang and Diane M Strong.** Beyond accuracy: What data quality means to data consumers

Process


9
metadata-supported data analysis
1
Other Data Sources
Process Data from Robots
Data Analytics Services
2
3
10
Data Sharing Broker
11
Integration Middleware/ Plant service bus
Data Lake P
altered dataset
FANUC
KUKA
Data Consumers
knowledge files (general, usecase)
4
5
8
mongoDB.
metadata
Data Quality and Metadata Repository
6
Requirements satisfied?
Yes
No
Verify Data Quality and Privacy Requirements
metadata, altered dataset
7
for re-evaluation
Deus
Metadata Extraction and Managament
Data Quality Assessment
Privacy-enhancing Microservices
Requirements analysis

Process

Metadata Extraction and Management

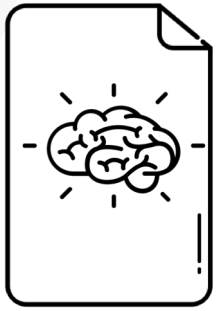
Knowledge Files

- **What?**
  Documents that capture essential information about:
  - The dataset
  - The use-case which will employ the dataset
- **Who?**
  Provided by the managers/owners of the dataset and the use-case
- **Why?**
  - To provide contextual information
  - Enable metadata-driven data analytics
  - Calculate quality and privacy scores
- **How?**
  Managers fill-in data via a form interface or Excel
- This is parsed by our *Excel Parsers* and converted to *.know* files (JSON format)


```
{
  "headers": {
    "datetime": "2018-09-20T07:48:05.724",
    "dataset-ID": "boschwps@datalakep1",
    "type": "general.know"
  },
  "catalogue": {
    "accessibility": "[FP/45,PN/62]",
    "responsibility": "Hr. Max Mustermann (N/FP-45)",
    "data-source": "Bosch WPS",
    "attributes": "[dateTime, timestamp, machine, current, energy, thickness, wear]",
    "data-description": "Data collected from the welding process in body shop",
    "historical-use": "[Data Lake P, Predictive Maintenance, Reporting]",
    "date-of-deletion": "2028-12-31T12:00:00.000",
    "data-storage": "Data Lake P"
  },
  "global": {
    "measurement-accuracy": "0.8"
  },
  "attributeProperties": {
    "dateTime": {
      "default": "0000-00-00T00:00:00.000",
      "unit": "n.a.",
      "privacy-sensitivity": "non-sensitive",
      "time-type": "collection",
      "maximum": "n.a.",
      "unit-shorthand": "n.a.",
      "data-type": "timestamp",
      "minimum": "n.a."
    },
    "current": {
      "default": "0.0",
      "unit": "ampere",
      "privacy-sensitivity": "sensitive",
      "time-type": "n.a.",
      "maximum": "10.0",
      "unit-shorthand": "A",
      "data-type": "numeric",
      "minimum": "0.0"
    },
```

Metadata Extraction and Management

Types of Knowledge Files



General Knowledge

- Provided by the *data owner*
- Helps to know the dataset, its attributes and properties closer
- Useful in calculating *fitness* metrics

Accuracy of the measuring device (between 0-1)

Department(s) with access to this data

Where is the data originating from?

Who owns/maintains this data?

Unique ID for the dataset

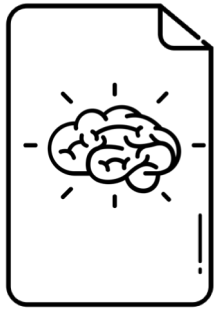
What is this data about?

Intended storage destination

What projects or use-cases has this data been used for?

Attribute properties: range, data type, default value, privacy

ISO Timestamp when the data should be auto-deleted



Usecase Knowledge

- Provided by the *use-case owner*
- Helps to understand the use-case, its properties and requirements
- Useful in calculating *usability* metrics

Dataset ID for which the use-case information is being recorded

Attributes used for the data analytics task

Attributes desirable for the use-case, but not present

Unique ID of the use-case

Reputation of the data source

Quality and Privacy requirements of the use-case

Who owns/maintains or is responsible for the use-case?

Volatility of the data for the use-case

Short text description of the use-case

Data Quality and Privacy Assessment
Fitness Dimensions

DQ Dimension	Weight	Metrics Definition
Accuracy	Between 0 and 1	accuracy of the measuring device at the source or probability of defects in measurement
Completeness	---do---	number of values that are not null/total number of values of that attribute
Consistency	---do---	measured by the consistency criterion, e.g. a range of values. For attributes with no consistency criterion, a check of the data type
Timeliness	---do---	latency between the time of data collection and reporting (arrival in the Data Lake P or middleware etc.)
Uniqueness	---do---	proportion of duplicates in the dataset
Volume	---do---	size of the dataset.
Interpretability	---do---	number of measurable attributes with units/total number of measurable attributes
Credibility	---do---	number of elements with default values/total number of elements

Data Quality and Privacy Assessment

Usability Dimensions

DQ Dimension	Weight	Metrics Definition
Volatility	Between 0 and 1	time length for which the data remains valid.
Reputation	---do---	what is the trustworthiness of the source that generated the data? (discrete values)
Relevance/Utility	---do---	% of the total attributes which will actually be used for the analytics task
Desirability	---do---	Does the analytics task desire any extra attributes which are not yet present in the dataset? What % of the desirable attributes are not present?

0 (lowest quality)

1 (highest quality)

Data Quality and Privacy Assessment

Privacy Dimensions

DQ Dimension	Weight	Metrics Definition
(Non-) Sensitivity	Between 0 and 1	% of attributes which are not key attributes/quasi-identifiers
Distinguishability	---do---	% of distinct pairs of key attributes/quasi-identifiers.
(Non-) Linkability	---do---	% of key and quasi attributes which are not tokenized.

0 (lowest privacy)

1 (highest privacy)

Data Quality and Privacy Assessment
Cataloguing Information

Information Attribute	Description
Accessibility	List of departments which have access to the information
Responsibility	Contact person who is responsible for maintaining and updating the metrics and knowledge files for the dataset
Historical Use	For what use-cases has the dataset been used in past
Date of Deletion	The date by which the data is required to be deleted

Data Quality and Privacy Assessment

How it all comes together

Process Knowledge

- Knowledge Files are parsed and materialized into objects by the *Knowledge Parsers*
- Pre-Evaluation:
  Using these objects, dimensions which don't need interaction with dataset are pre-evaluated

Evaluate

- *Fitness, Usability* and *Privacy* dimensions are evaluated from the dataset
- The metrics are calculated in Spark and Spark SQL and normalized to lie between 0 and 1

Combine

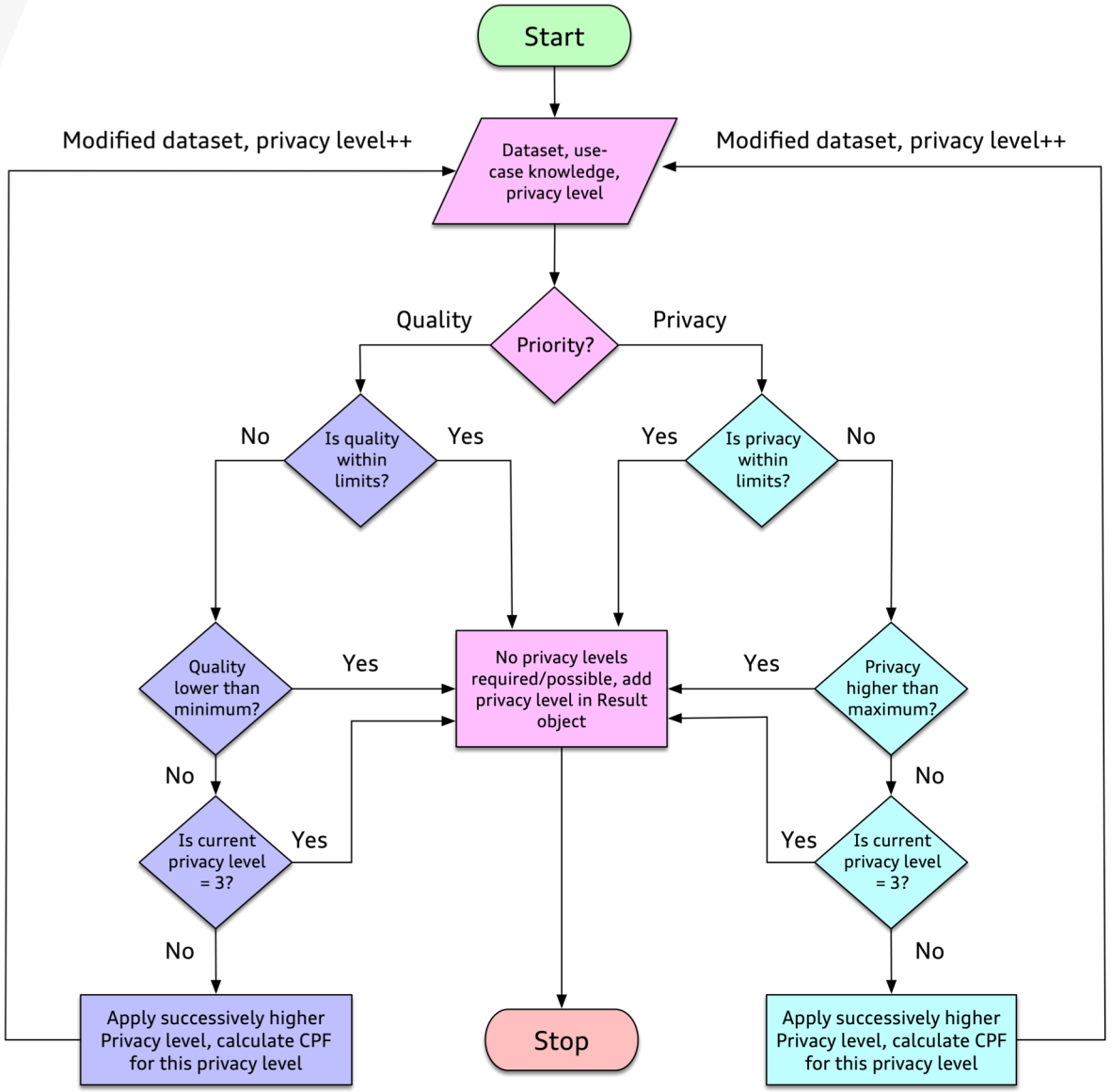
- Pre-determined weights are used to calculate the final *fitness*, *usability* and *privacy* scores
- The final data quality score is calculated based on the weights assigned to *fitness* and *usability*

$$fitness = \frac{\sum_{dim \in fitness} dim}{\sum_{dim \in fitness} 1} \qquad usability = \frac{\sum_{dim \in use} dim}{\sum_{dim \in use} 1} \qquad privacy = \frac{\sum_{dim \in priv} dim}{\sum_{dim \in priv} 1}$$

$$combinedscore = \frac{(fitness \times weight_{fitness}) + (usability \times weight_{usability})}{weight_{fitness} + weight_{usability}}$$

Requirements Analysis

Determining whether the Metrics are Adequate



- **Function**
  To determine whether the evaluated data quality and privacy scores are sufficient, as per the requirements defined by the use-case

- **Cases**
  - If the requirements are satisfied, the final *Result* file is produced and metadata is written to the MongoDB database
  - If requirements are not satisfied, we go to the next phase, where privacy-enhancing microservices are applied

Privacy-Enhancing Microservices
Privacy Levels

1

- Tokenize key attributes
- 5-anonymize quasi attributes

2

- Tokenize key attributes and 50% of quasi attributes
- Round-off sensitive attributes to integers
- 10-anonymize quasi attributes

3

- Tokenize all key attributes and quasi attributes
- Round-off sensitive attributes to integers, smoothen curves with a Gaussian function, randomly suppress String data

Privacy-Enhancing Microservices

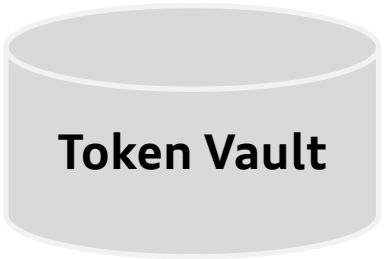
Tokenization via Secure Tokens

Tokens

Secure Token = 32 char md5 checksum + 4 char salt

Insert token/
Batch insert token

Retrieve token value



SecureFileServices

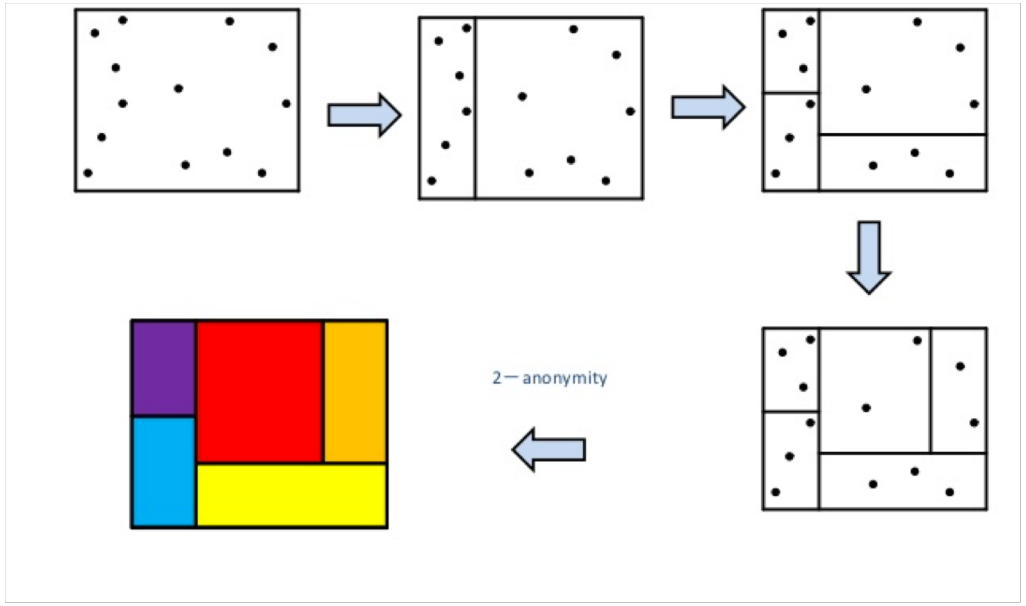
- AES/symmetric encryption
- Decryption with a passkey
- Always exists in encrypted format
- Temporary decrypted files are immediately destroyed

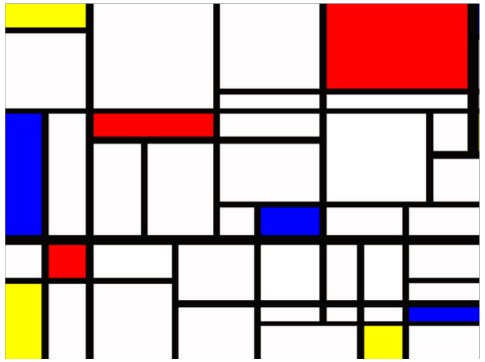
Privacy-Enhancing Microservices
Mondrian k-Anonymity

Mondrian Multidimensional Partitioning :
Recursive greedy partitioning of the vector space

Partition(region, k)

1. Choose the dimension that results in the most even k-anonymous partition
2. If possible, partition the region according to that dimension into R1 and R2
3. Return Partition(R1, k) U Partition(R2, k)
4. If not possible, Return.





K-Anonymity Illustration from the slides of Hiroshi Nakagawa, The University of Tokyo
Mondrian Artwork from: https://musartboutique.com/mondrian-art-hundred-year-legacy/

Privacy-Enhancing Microservices

Cumulative Penalty Factor: Adjusting usability based on PETs

$$\textbf{Usability}_{\textbf{adjusted}} = \text{Usability x CPF}$$

PET used	Usability Loss
Tokenization	100
5-anonymization	20
10-anonymization	30
Rounding-off	5
Partial Suppression (100)	10
Rounding-off + Partial Suppression(100)	20

$$\textbf{Usability Loss} = \frac{\sum_{i=0}^{n(PETs)} numAttributespeti \times usabilityLoss_{peti}}{numPrivacyAttributes \times 100}$$

$$\textbf{Privacy Ratio} = \frac{numPrivacyAttributes}{numTotalAttributes}$$

$$\textbf{Normalized Usability Loss} = Usability\ Loss \times Privacy\ Ratio$$

$$\textbf{CPF} = 1 - Normalized\ Usability\ Loss$$

For the Böllhoff Riveting Dataset

$$CPF_{level1} = 0.80$$

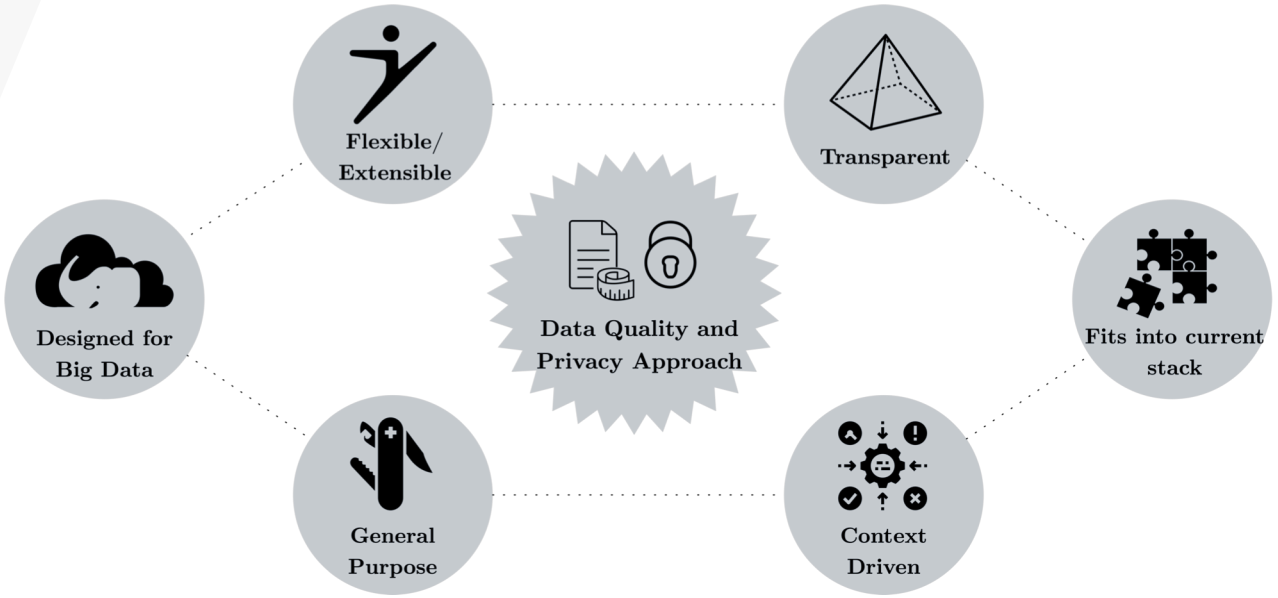
$$CPF_{level2} = 0.6428$$

$$CPF_{level3} = 0.4571$$

Results and Evaluation

Design Considerations Revisited

How our implementation fulfils these criteria



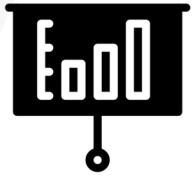
- **Flexibility**
  Most processes are fully *extensible*, and can be modified according to requirements
- **Designed for Big Data**
  Written using the *Spark API* to leverage the power of distributed computing
- **General Purpose**
  Applicable to data from different *domains*
- **Context Driven**
  Considers the *usability* aspect of data
- **Fits into current infrastructure**
  Uses widely used, *open source* technologies
- **Transparent**
  Highly *granular* analysis of metrics without any black boxes, CPF measures the usability loss

Survey Results

SNo.	Question	Aim
1	*Which is your role within in the organization?*	Determine the role of the respondent *(data owner, user or data quality and privacy expert)*
2	*What does 'good quality data' mean to you?*	Determine the most well-understood and favoured understanding of *good data quality*
3	*For the data quality dimension 'consistency', which of the following aspects do you consider as important?*	Determine whether the definition used in our approach closely matches the commonly understood meaning of the term
4	*In your opinion, what qualities should a dataset have, to be labelled as 'useful' for an internal use-case?*	Determine the most well-understood and favoured understanding of the term *usefulness*
5	*Do you agree that a dataset with high uniqueness, i.e. less number of repetitive or redundant values, has a better utility for analytics?*	Determine whether high uniqueness of values within a dataset is understood as an indicator of high data quality
6	*What are the preferable qualities of PETs?*	Understand the expectations that the respondents have, from a privacy-enhanced dataset, and whether our approach is able to ensure this
7	*How can adequate trust be established between parties in data sharing scenarios in the automotive sector?*	Understand critical motivating factors that drive trust in data-sharing
8	*According to you, what are the most important features of a highly 'interpretable' dataset?*	Determine the most commonly understood meaning of *interpretability*
9	*What would be an acceptable way of tokenizing a dataset?*	Evaluate the respondents' opinion about whether tokenization is acceptable in data-sharing, and what kinds of attributes should be tokenized
10	*What steps should be taken towards achieving 'metadata management' effectively?*	Understand whether the respondents opine that metadata management is important, and the most important steps that should be taken to ensure it

- **Widespread understanding of dimensions**
  High degree of *conformity* of the definitions of our dimensions, with the respondents' expectations
- **Support for PETs in Data Sharing**
  95% of our respondents agreed that sensitive information should be *anonymized* before being shared
- **Importance of Metadata Management**
  83% people responded that in-spite of high effort, *metadata* should be collected and continually maintained

Evaluation
Applying our Methodology to Real Datasets



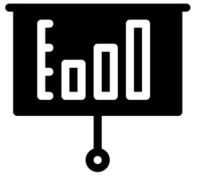
Use-case #1

Based on Bosch WPS Weldlogs

Predicting thickness of welding pt. based on

- Current
- Energy
- Machine Wear

- Put data quality into perspective without applying any PETs on the dataset.
- Analyse the results, and evaluate how they aid in data mining processes such as data pre-processing and transformation



Use-case #2

Based on Böllhoff Riveting Dataset

- Determine most common rivet types
- Based on the process counter
- Also find out the average compression force and compression time

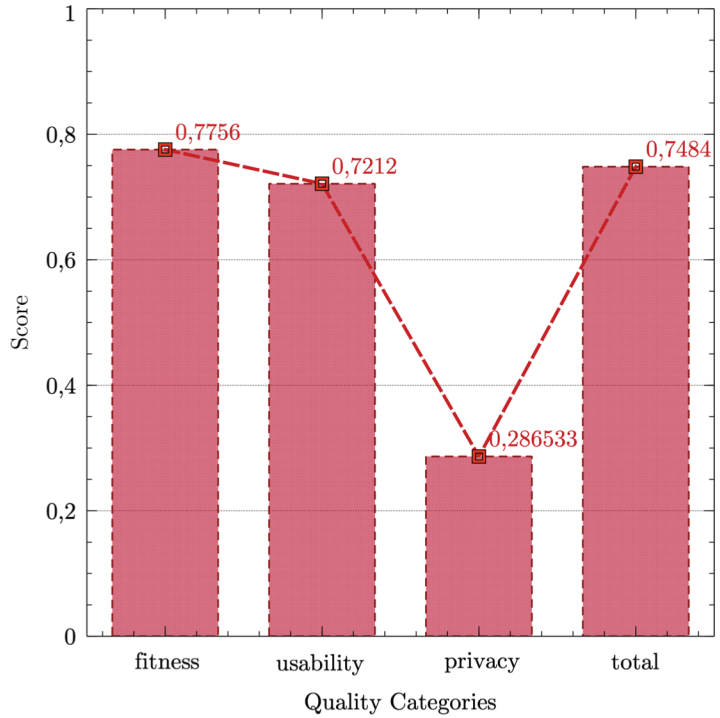
- Apply the three privacy levels
- Study the relationship between privacy and data quality
- Determine the usability of various Privacy Levels by querying the dataset via a *'Bag of Queries'*

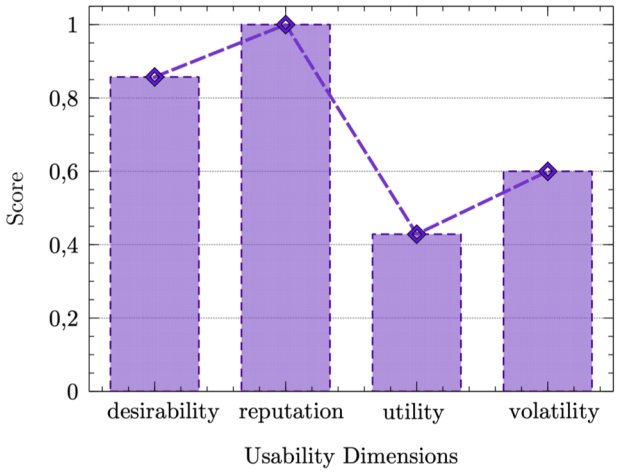
Evaluation

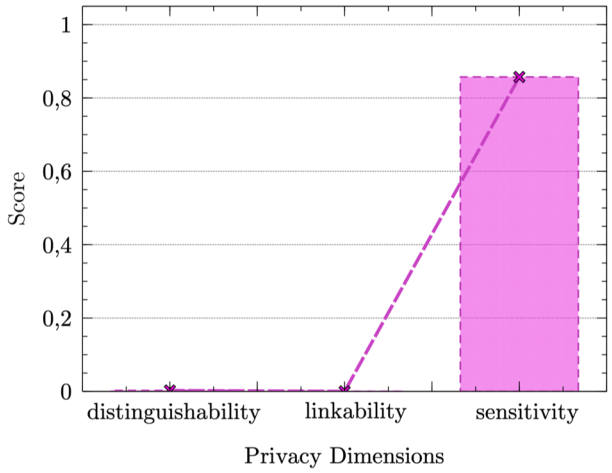
Use-case 1: Bosch Weldlog Dataset

- **Dataset Description**
  15,000 records of welding logs
  Attributes contained: *(datetime, timestamp, machine, current, energy, thickness, wear)*

- **Insights**
  - *Attribute-shift problem* – identified due to the unexpectedly low completeness score of *current*
  - Ripple-effect on other attributes due to the attribute-shift, leading to a low consistency score
  - Identification of sensor malfunction due to low uniqueness and credibility of *current*
  - Observation that data is unsuitable for sharing in its current form

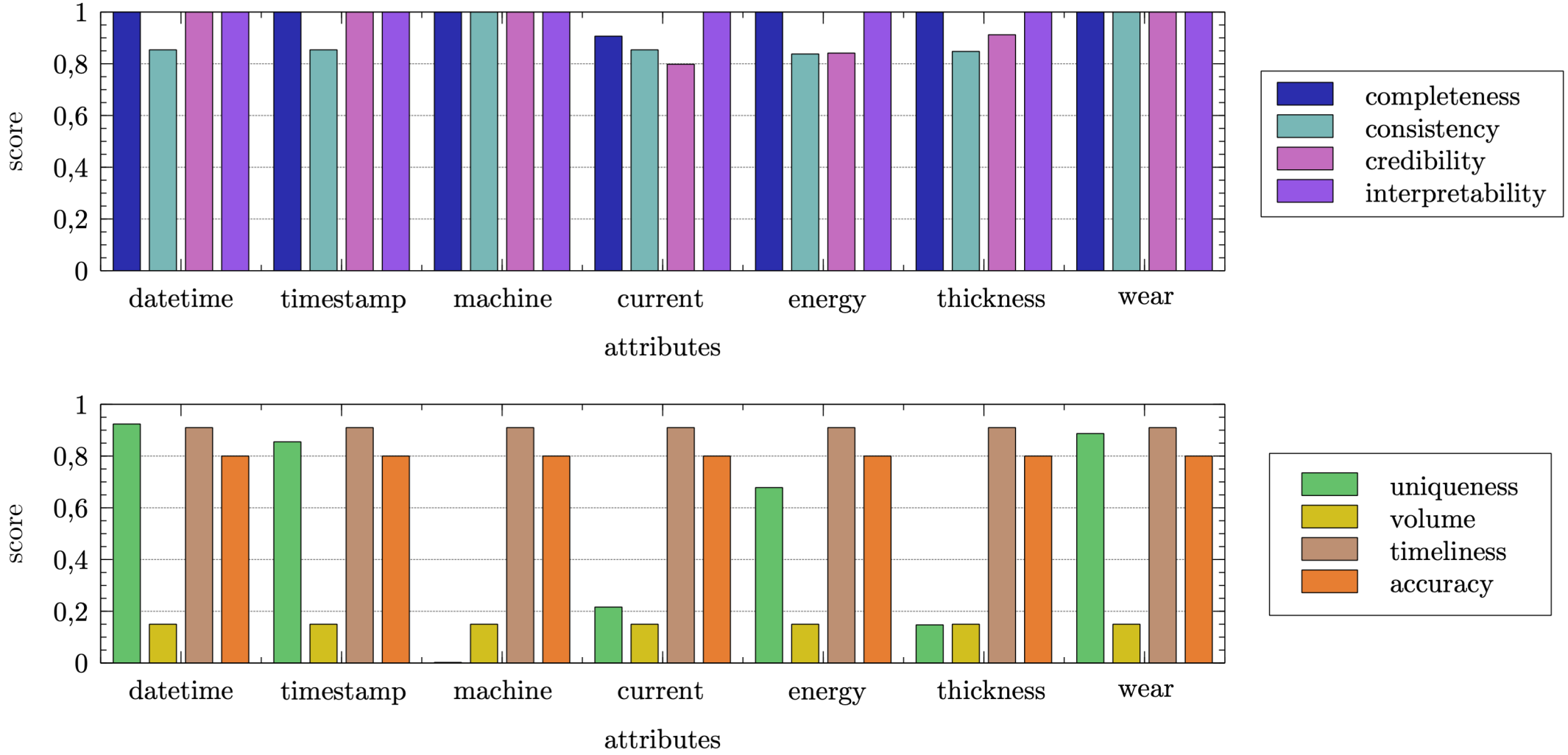






Evaluation

Use-case 1: Bosch Weldlog Dataset



Evaluation

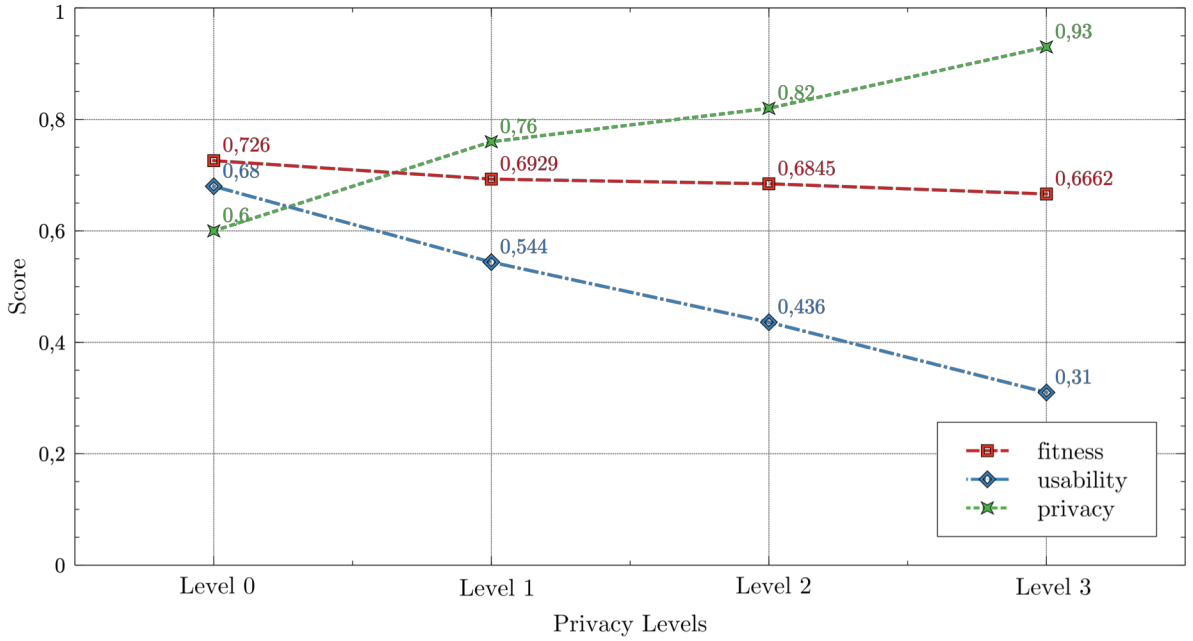
Use-case 2: Böllhoff Riveting Dataset

- **Dataset Description**
  34,118 records of riveting logs
  15 Attributes selected out of the original 104

- **Observations**
  - Higher privacy levels yield a higher privacy score, justifying their strictness ranking
  - There is a consistent dip in the usability score of the dataset, effected by the CPF
  - Fitness scores also witness a small but noticeable dip

- **Miscellaneous**
  - Timestamp error
  - Float error

Attribute	Data Type	Privacy Sensitivity	Minimum Value	Maximum Value	Unit
sourceTimestamp	timestamp	key-attribute	N.A.	N.A.	N.A.
serverTimestamp	timestamp	key-attribute	N.A.	N.A.	N.A.
processCounter	string	quasi-identifier	N.A.	N.A.	N.A.
joiningPoint Description	string	quasi-identifier	N.A.	N.A.	N.A.
joiningPointName	string	quasi-identifier	N.A.	N.A.	N.A.
rivetType	numeric	quasi-identifier	0	10	unitless
rivetLength	numeric	sensitive	0.0	10.0	millimetre
contactForce	numeric	sensitive	0.0	10.0	newton
contactTime	numeric	sensitive	0.0	10.0	nanoseconds
preClampingForce	numeric	sensitive	0.0	10.0	newton
preClampingTime	numeric	sensitive	0.0	10.0	nanoseconds
compressionForce	numeric	sensitive	0.0	10.0	newton
compressionTime	numeric	sensitive	0.0	10.0	nanoseconds
processControl Relative	boolean	non-sensitive	N.A.	N.A.	N.A.
returnDistance	numeric	sensitive	0.0	10.0	millimetre



Evaluation

Use-case 2: Bag of Queries

1. How many distinct *joining points* are present in the dataset?
2. What is the average *compression force*?
3. What is the average *compression time*?
4. What is the most common *rivet type* in the dataset?
5. What is the most common *rivet type* for the most frequent process counter?
6. What percentage of riveting operations were performed after noon?
7. How many distinct counts of *process counters* are present in the dataset?
8. What is the difference between *source time* and *server time*?

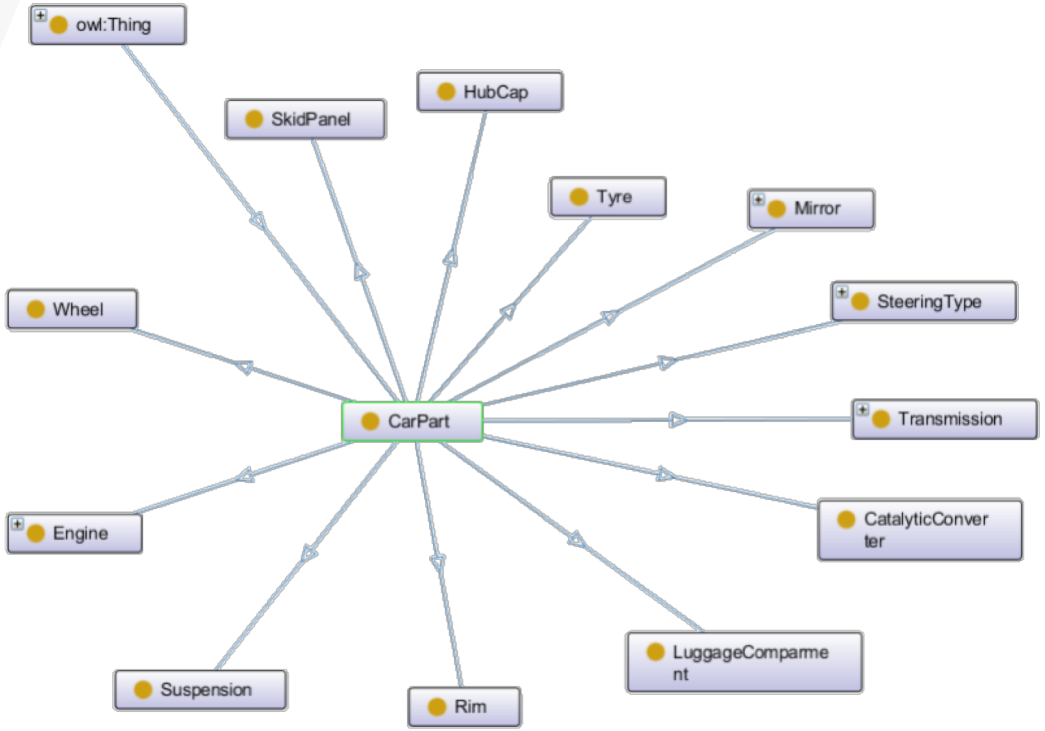
QNo.	Query Type	Response (Raw Dataset)	Response *(Deviation)* Level 1	Response *(Deviation)* Level 2	Response *(Deviation)* Level 3
1	Data Analytics	13	12 *(8%)*	10 *(26.08 %)*	- *(-)*
2	Data Analytics	4.71	4.71 *(0%)*	4.59 *(2.58%)*	4.16 *(12.4%)*
3	Data Analytics	2.50	2.50 *(0%)*	2.42 *(3.25%)*	2.28 *(9.21%)*
4	Data Analytics	2	2 *(0%)*	2 *(0%)*	- *(-)*
5	Data Analytics	2	2 *(0%)*	2 *(0%)*	- *(-)*
6	Private	35%	- *(-)*	- *(-)*	- *(-)*
7	Private	21,586	4214 *(134.6%)*	- *(-)*	- *(-)*
8	Private	0	- *(-)*	- *(-)*	- *(-)*

Query Type	Cumulative Deviation (Level 1)	Cumulative Deviation (Level 2)	Cumulative Deviation (Level 3)
Data Analytics	1.6%	6.382%	10.805%
Private	134.6%	-	-

Conclusion
Achievements of this thesis

- **A flexible data quality metric for organizational big data**
  Current approaches required very high human effort, were too general or rigid, or weren't suitable for big data
- **An approach for managing data privacy, and evaluate its corresponding impact on data quality**
  On-demand privacy enforcement based on privacy requirements of the use-case, while still maintaining utility for analytics
- **Metadata management for metadata-driven analytics**
  - Enables us to get a better understanding of the collected data.
  - Data scientists have wide-ranging access to essential metadata

Conclusion
Scope for Future Work



- **Ontology-Driven Knowledge Generation**
  Organizations often collect the same kind of data. Creating a concept hierarchy can help us automatically tag metadata
- **Inbuilt Dashboarding/Visualization Capabilities**
  The metrics calculated by our approach will be much more useful if there is an inbuilt tool for monitoring and visualization
- **Use-cases in different Domains**
  Adapting this approach to use-cases in other domains will provide a better idea about its versatility

Automotive Ontology Image from: https://www.w3.org/community/gao/wiki/Main_Page

Questions?

Appendix

Data Quality
Creating value out of Big Data

Measurement

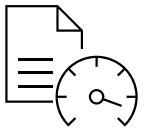
- Effectively measure the amount of value contained in data
- To be able to quantify data fitness as well as usability
- Assess the effect of anonymisation and perturbation towards data quality and utility

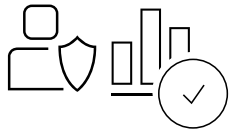
Management

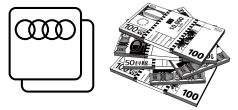
- Adjustable levels of data quality
- Anonymisation to protect business secrets
- Balance between privacy enforcement and data utility
- Metadata management for internal data governance

‘High-value Data’

- Get-what-you-pay-for model of data sharing
- Analytics algorithms can train using only high-quality data
- More knowledge about collected data within the organization
- No unintended revelation of sensitive data



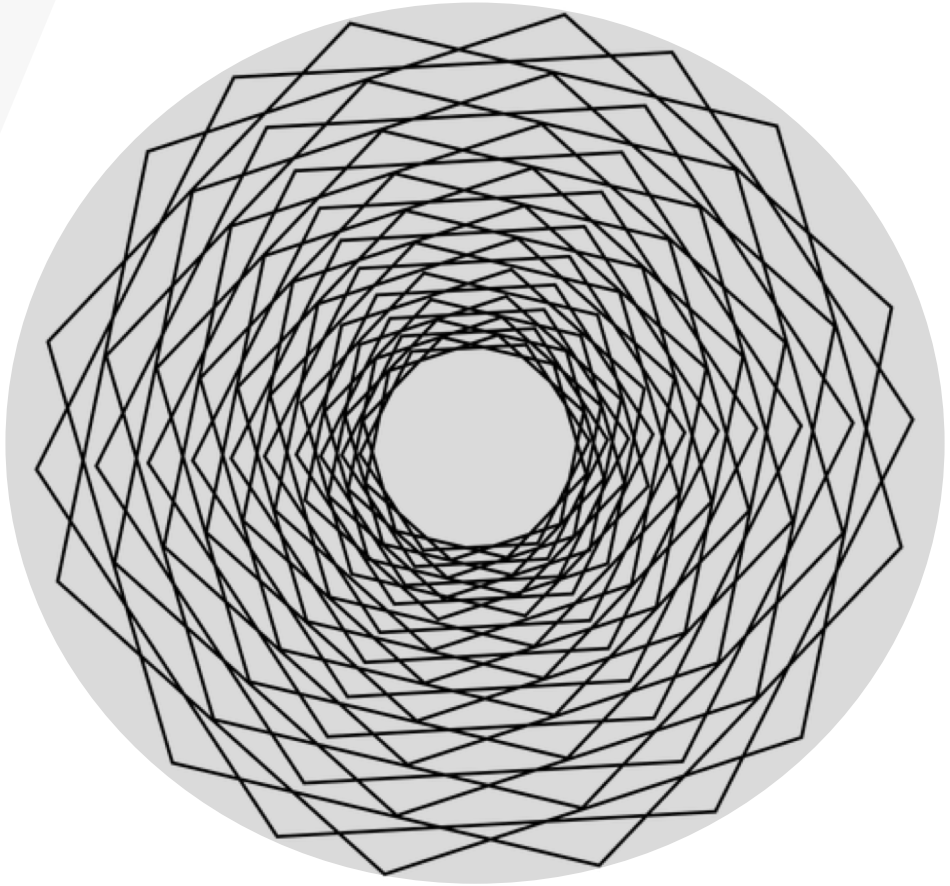




Quality Measurement + Quality Management = HIGH VALUE DATA

Data Lake without Data Management

Coming to a full circle



- **Enterprise Data Lake can, instead become a swamp** [2]
  A Data Lake guarantees efficient storage of heterogeneous data, but no data management. Just dumping data inside it will counteract its advantages
- **Data Lake provides the right environment for analytics**
  But selecting the right dataset and cleaning it can still be a mammoth task
- **Data Lake doesn't ensure superior data quality** [2]
  There must be a method to quantify the quality of data contained in the data lake, in a manner which suits data users
- **Metadata and Knowledge Management is key** [3]
  If no metadata is associated with the data, it loses context. This creates a silo which hinders its usage for analytics

Fractal Image: https://www.pngarts.com/explore/13851

[2] **Rihan Hai, Sandra Geisler, and Christoph Quix.** Constance: An intelligent data lake system
[3] **Data Lake Governance**. https://searchcio.techtarget.com/feature/Data-lake-governance-A-big-data-do-or-die

Data Quality
What can we do with it?

- **Data Cataloguing and Governance**
  Knowing more about the data we collect, enabling analytics using enterprise-wide data
- **Data Selection**
  Selecting the correct data for essential tasks such as predictive maintenance and fault detection
- **Data Migration**
  Selecting high-value data to be migrated to the data warehouse or the enterprise cloud
- **Selling and Data Economy**
  Knowing the worth of the data to create maximum value by selling it to third parties, e.g. robot manufacturers

Image: https://www.keylink.net.au/product/trillium-software-system

Data Privacy
Why do we need it?

- **Policy Requirement**
  Mandatory legal requirements regarding what kind of data must (or must not) be stored or disclosed. E.g. GDPR

- **Data Disclosure**

  - **For non-business purposes**
    Data disclosure for use-cases where critical business decisions are not involved. E.g. hackathons

  - **For business purposes**
    Data disclosure scenarios where the mining task has high stakes, as they are likely to affect business decisions

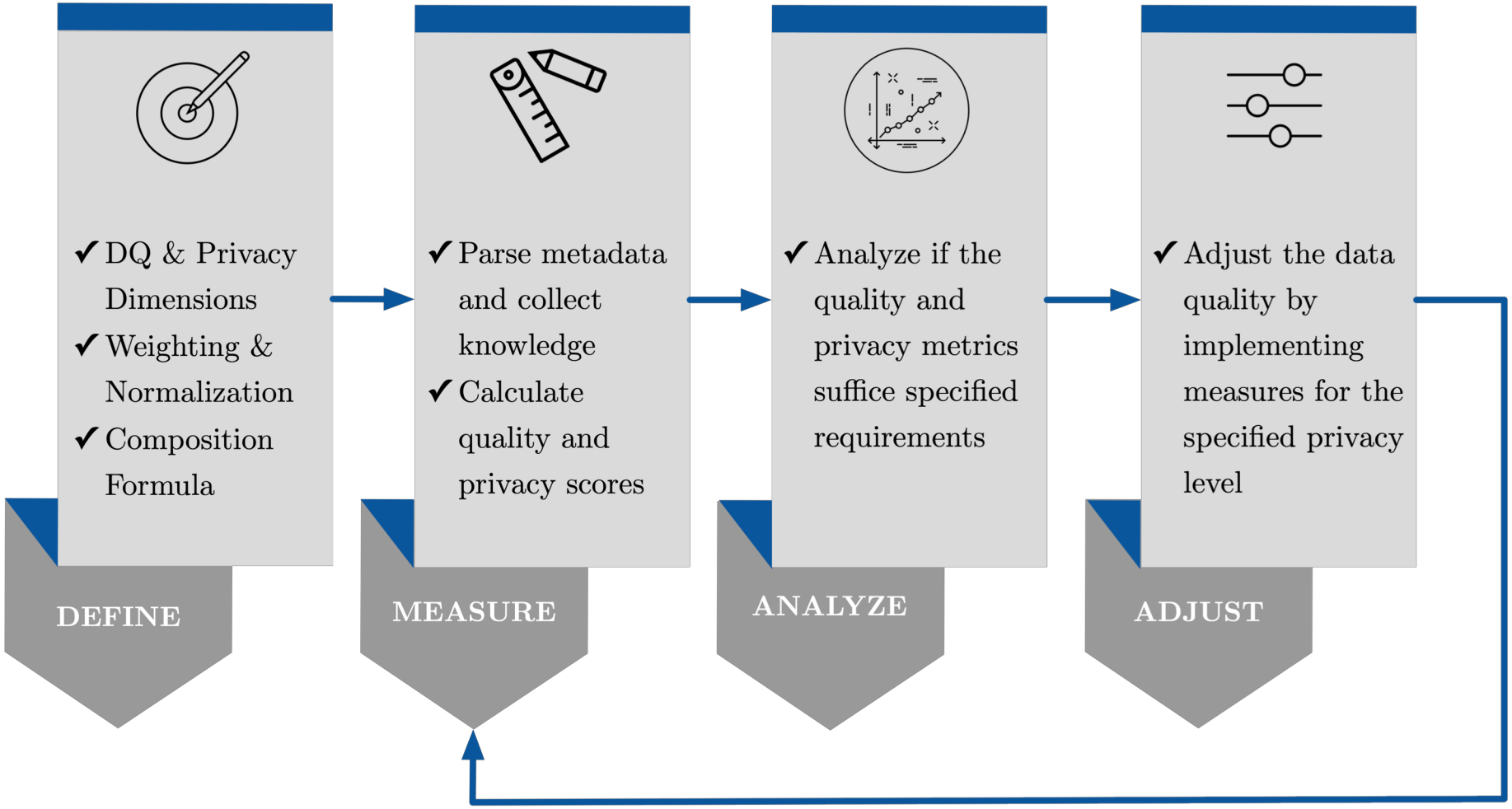
In all disclosure scenarios, the privacy as well as confidentiality must be maintained

Privacy Image: https://oit.ncsu.edu/it-security/data-privacy-month-2016-guard-your-privacy/

Data Quality Management
Within the scope of Data Governance

Adapted from **Andreas Kirmse, Max Hoffmann, Johannes Lipp, Max Haberstroh:** Audi Data Lake P Workshop, 09.07.2018

Implementation Plan

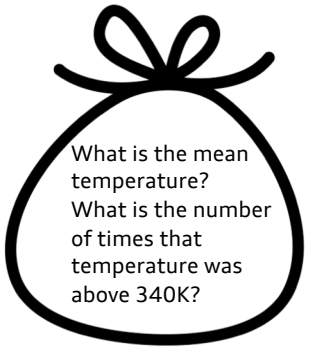


Evaluation Strategy

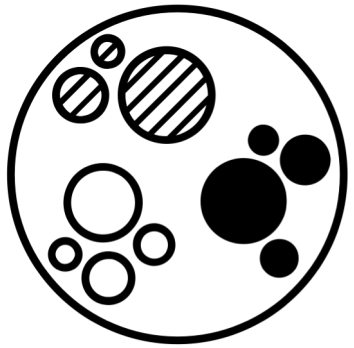
Q1 How useful is the quality measure?

- Expert Questionnaire
- User Questionnaire

Q2 What is the relationship between data quality and utility after applying PETs?



Bag of Queries



Learning Modelling

Survey Responses - 1

1

1	Data User / Use-case Owner	31 / **74%**
2	Data Quality or Privacy Expert	7 / **17%**
3	Data Owner / Manager	4 / **10%**

2

1	Data which follows the expected schema	36 / **86%**
2	Data which doesn't have many null or incomplete values	30 / **71%**
3	Data which is reported on time	21 / **50%**
4	Data which is about something exciting	8 / **19%**
5	Data which hides personal details or business secrets	6 / **14%**
6	Data which other colleagues have found useful	5 / **12%**

3

1	Values have the correct data type	39 / **93%**
2	Values are within the expected range	27 / **64%**
3	Values satisfy the business constraints	17 / **40%**
4	Values are easy to read	7 / **17%**
5	Values have a compact representation	4 / **10%**

4

1	It contains all the attributes required for the use-case	41 / **98%**
2	It is time-relevant for the use-case, i.e. it has not expired	32 / **76%**
3	It has been released by a trusted authority	18 / **43%**
4	It has been used in the past for similar use-cases	9 / **21%**
5	It is strictly confidential, and very few people have access to it	3 / **7%**

5

Average: 5.60

0	1	2	3	4	5	6	7	8	9	10

6

1	They should anonymize information about individuals or company secrets	40 / **95%**
2	The resulting dataset should still be usable for analytics purposes	36 / **86%**
3	They should protect the dataset against attackers who try to link datasets to gather more information (linkage attacks)	31 / **74%**
4	The effects of applying these PETs should be quantifiable and transparent	28 / **67%**
5	They should make the data analytics tasks more complicated	1 / **2%**

Survey Responses - 2

7

1	By maintaining a delicate balance between data privacy and usability	30 / **71%**
2	By providing measurable quality guarantees for the released data	30 / **71%**
3	By being transparent about the pre-processing applied to the data before releasing	28 / **67%**
4	By releasing artificially-simulated data instead of the real data	5 / **12%**
5	By intentionally releasing fake data to misguide other organizations	0 / **0%**

8

1	Numerical values (e.g. current, voltage) must be associated with units	33 / **79%**
2	Values must not be rendered useless by privacy-enhancing technologies	28 / **67%**
3	Values must be human-readable	26 / **62%**
4	Values must not be too long	5 / **12%**
5	Values should be rounded-off	2 / **5%**

9

1	Only the most privacy-critical attributes should be tokenized	33 / **79%**
2	Only those attributes should be tokenized, which are not critical for the analytics task	15 / **36%**
3	Users should have the option of choosing the non-tokenized dataset (e.g. at a higher price)	12 / **29%**
4	No attribute should be tokenized	1 / **2%**
5	All attributes should be tokenized	0 / **0%**

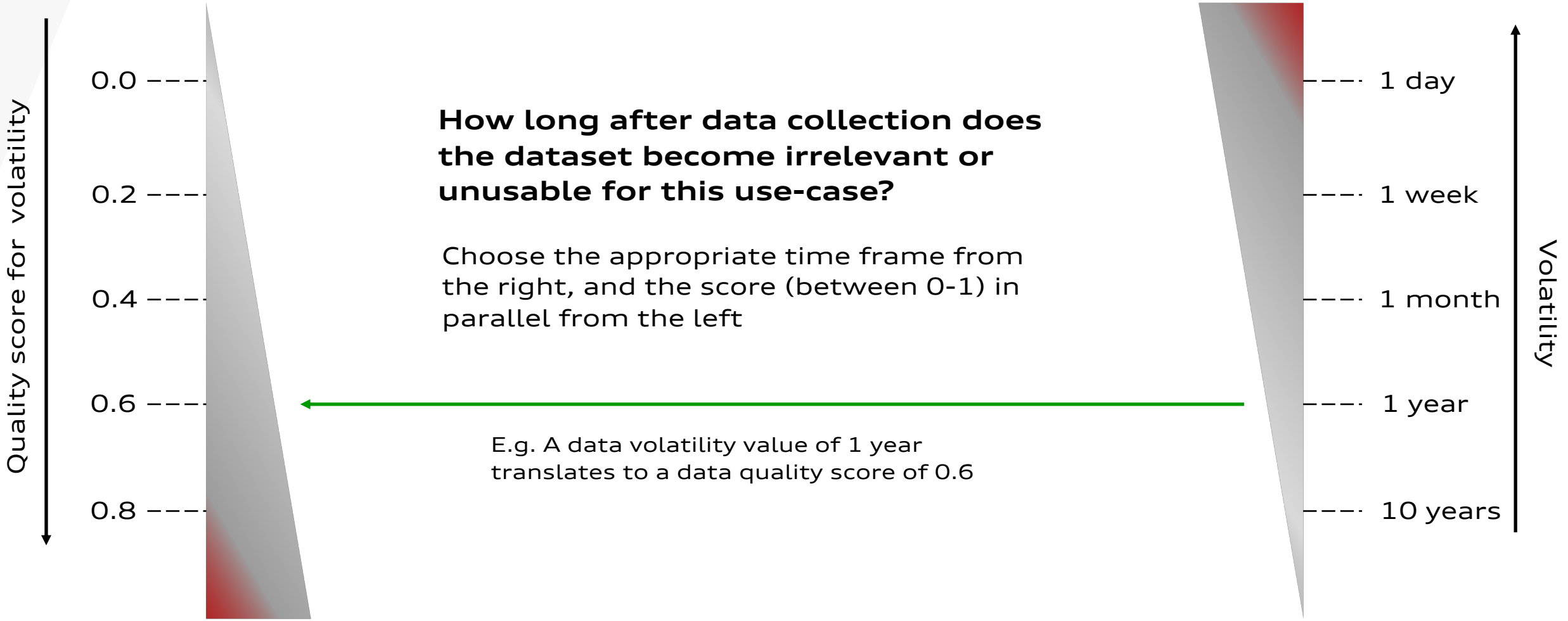
10

1	Metadata should be continually collected and updated	35 / **83%**
2	An organization-wide data catalog (like Yellow Pages) or centralized metadata repository should be maintained	29 / **69%**
3	A Data Marketplace should be set up, where parties can provide and consume datasets	22 / **52%**
4	Nothing should be done, as metadata is not important	1 / **2%**
5	Metadata is important, but collecting it is very tedious, hence spending time and effort on it is not worthwhile	0 / **0%**

Data Volatility

Premise

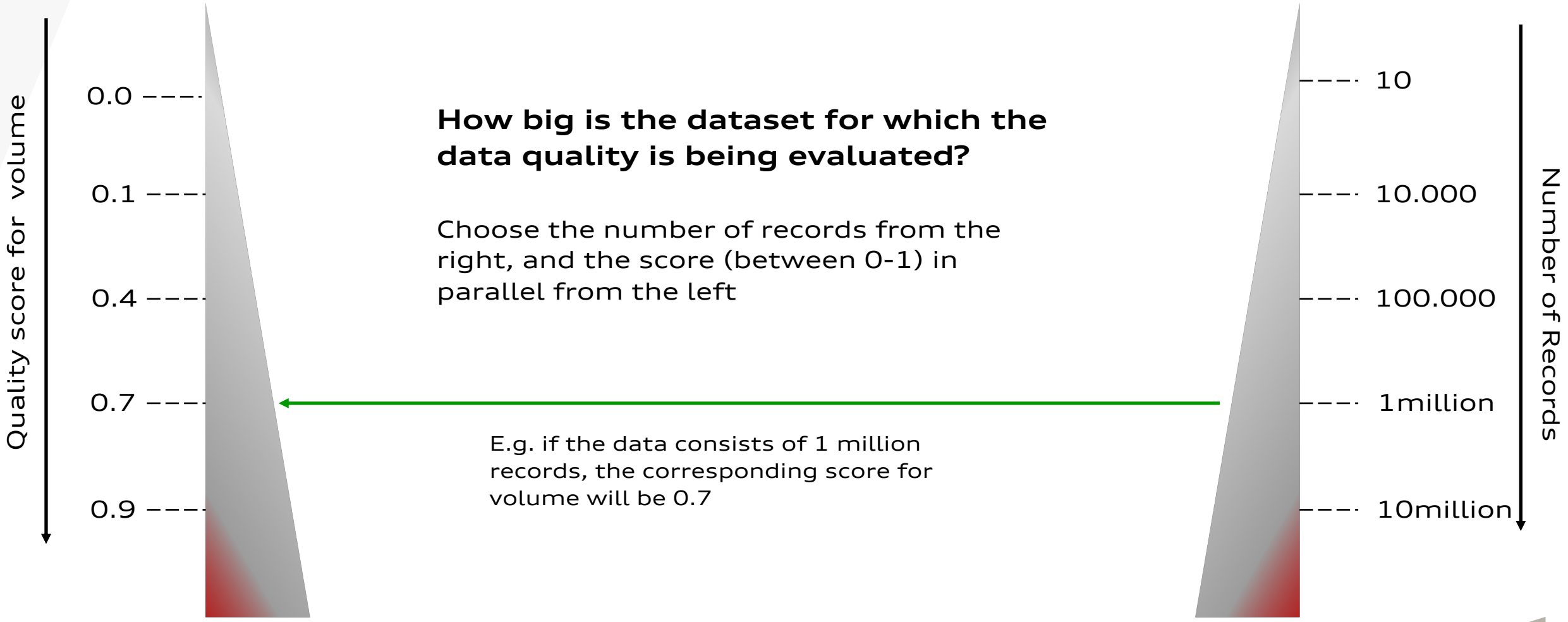
- Volatile data has lesser utility in long-term decision making*
- Volatile data should not be used to access risk or for use-cases involving risk
- Volatile data should not be used to confirm a hypothesis



*Al-Debei, Mutaz M. "Data warehouse as a backbone for business intelligence: Issues and challenges." *European Journal of Economics, Finance and Administrative Sciences* 33, no. 1 (2011): 153-166.

Volume
Premise

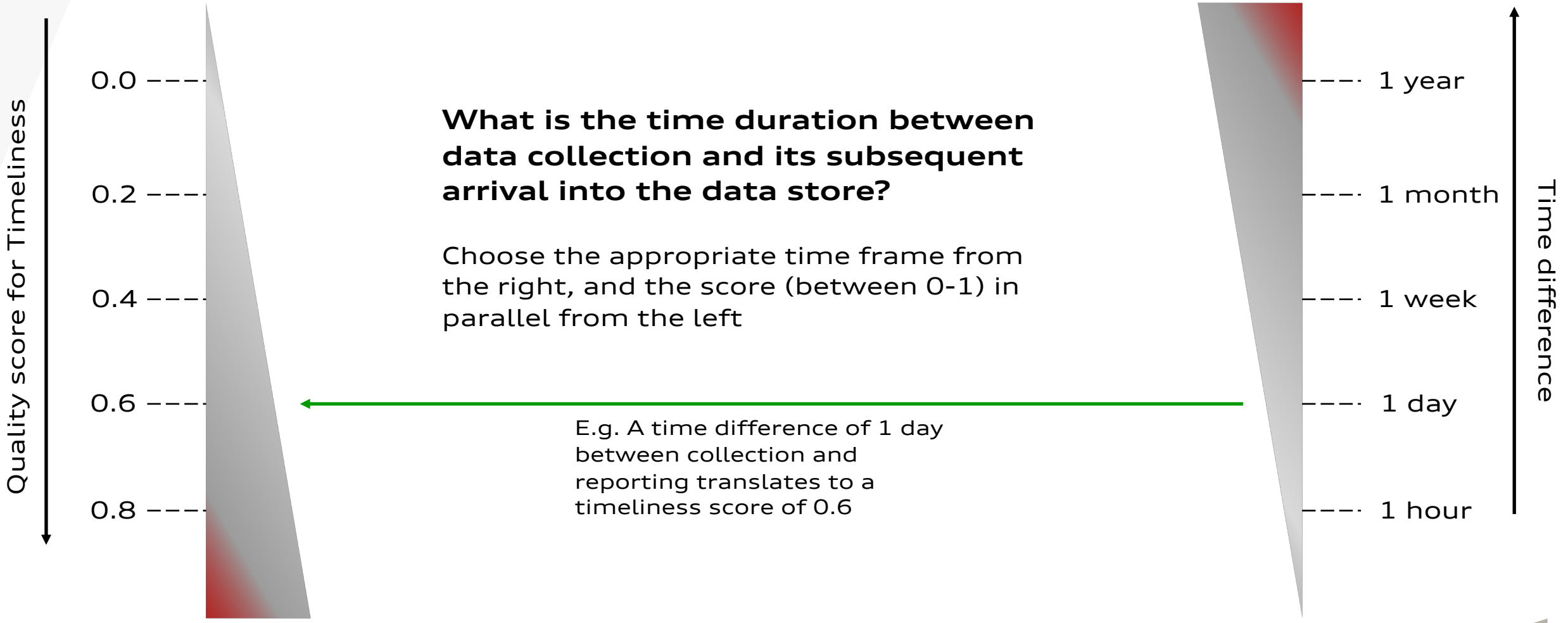
- More data is better than a better-performing algorithm



*Al-Debei, Mutaz M. "Data warehouse as a backbone for business intelligence: Issues and challenges." *European Journal of Economics, Finance and Administrative Sciences* 33, no. 1 (2011): 153-166.

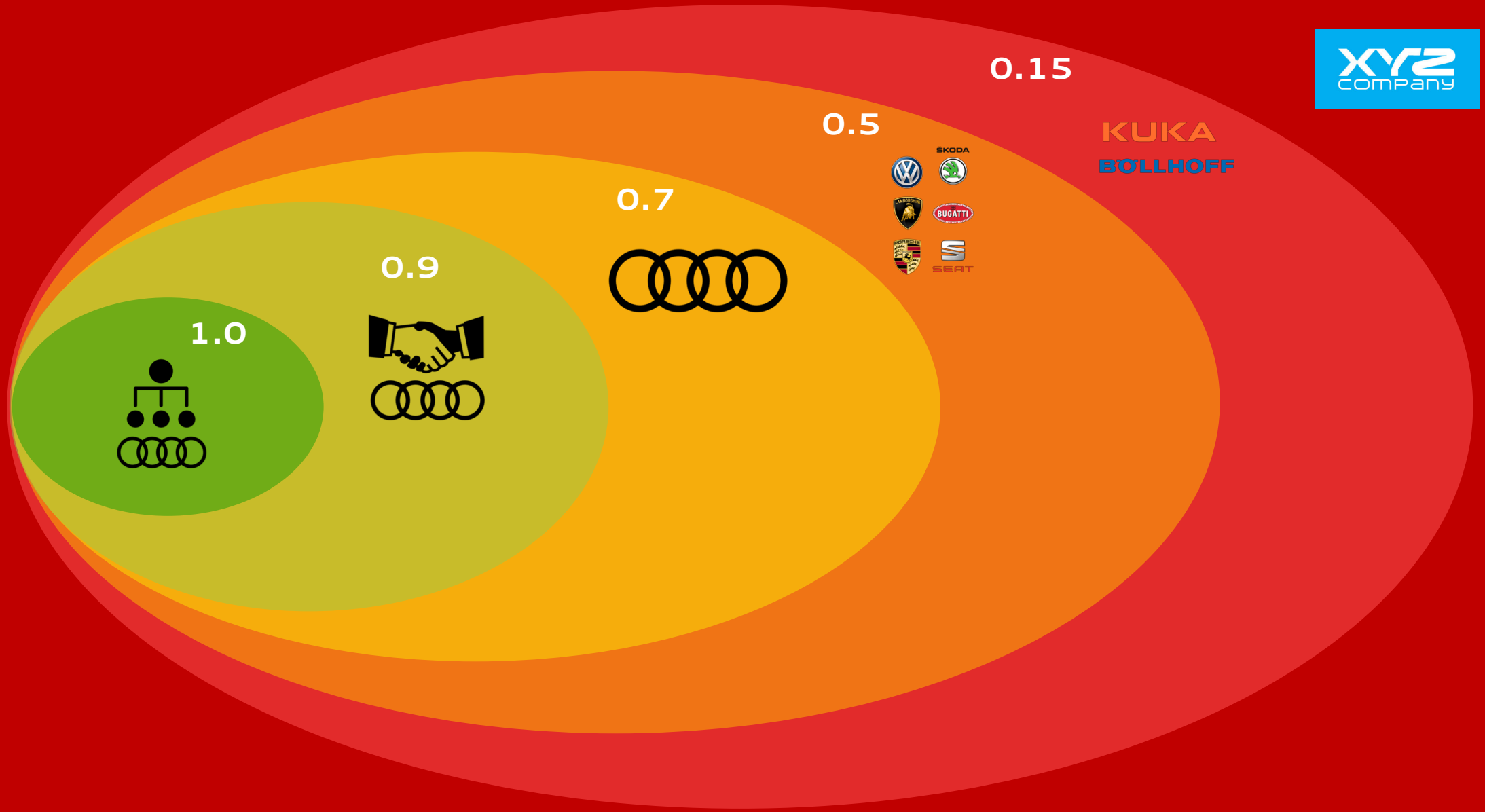
Timeliness
Premise

- Data processing and analysis based on untimely or expired data will likely produce useless or misleading conclusions, leading to decision-making mistakes.



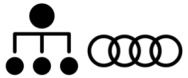
*Al-Debei, Mutaz M. "Data warehouse as a backbone for business intelligence: Issues and challenges." *European Journal of Economics, Finance and Administrative Sciences* 33, no. 1 (2011): 153-166.

Trust Networks affect Reputation



Trust Levels

1. **Complete** Trust



2. **Reasonable** Trust

3. **Considerable** Trust

4. **Plausible** Trust

5. **Sufficient** Trust

6. **Insufficient** Trust