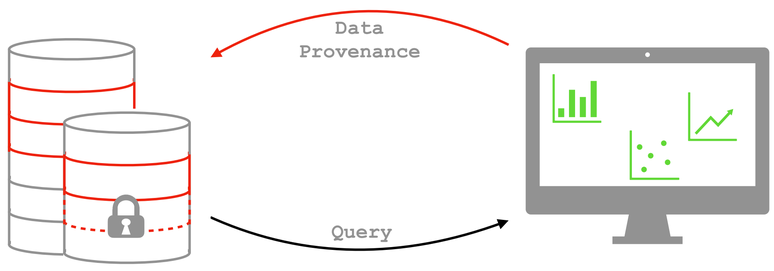
Universität Rostock

Traditio et Innovatio

Privacy Aspects of Provenance Queries

Motivation

Privacy and Data Provenance:



Privacy:

- protection of personal data against unauthorized collection, storage and publication
- possibility of not re-identifying single persons in a bunch of data

- big data: amount of data ↑, transparency ↓
- GDPR: data protection more important than ever before

Data Provenance:

- lineage of data
- ***where***: Where does the data come from?
  ⇒ names of the source relations like `Grades`
- ***why***: Why was this result achieved?
  ⇒ witness bases like $\{\{t_4, t_5, t_6\}\}$
- ***how***: How was the result calculated?
  ⇒ provenance polynomials like $\frac{4.0\ t_4 + 5.0\ t_5 + 1.3\ t_6}{t_4 + t_5 + t_6}$

Possible provenance-based Database Reconstructions:

Grades

	ID	Module	Grade
t_1	13	Mathematics	~~1.0~~ A
t_2	27	Data Science	~~2.3~~ B
t_3	27	Theory	~~1.7~~ B
t_4	115	Mathematics	~~4.0~~ D
t_5	115	Data Science	~~5.0~~ E
t_6	115	Law	~~1.3~~ A

ID	avg_{Grade}
13	~~1.0~~ A
27	~~2.0~~ B
115	~~3.3~~ C

where Grades

why $\{\{t_4, t_5, t_6\}\}$

how $\frac{4.0\ t_4 + 5.0\ t_5 + 1.3\ t_6}{t_4 + t_5 + t_6}$

where

ID	Module	Grade
13	η_1	1.0
27	η_2	2.0
115	η_3	3.3

why

ID	Module	Grade
13	η_1	1.0
27	η_2	2.0
27	η_3	2.0
115	η_4	3.3
115	η_5	3.3
115	η_6	3.3

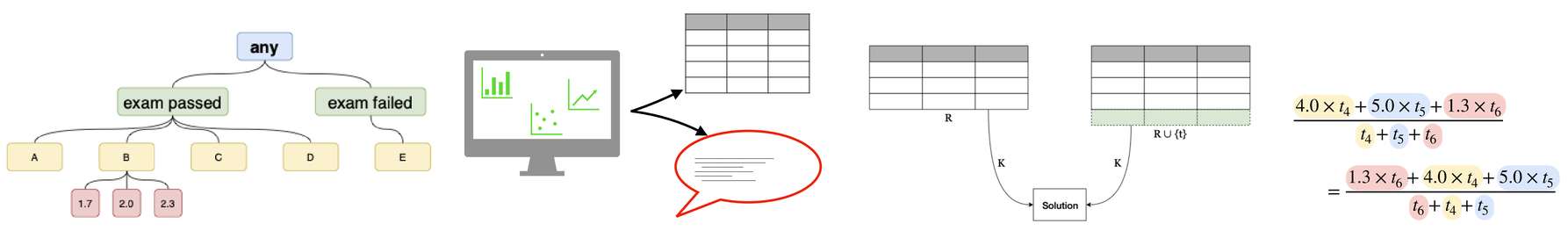
how

ID	Module	Grade
13	η_1	~~1.0~~ A
27	η_2	~~2.3~~ B
27	η_3	~~1.7~~ B
115	η_4	~~4.0~~ D
115	η_5	~~5.0~~ E
115	η_6	~~1.3~~ A

Data Protection Problems with *where*, *why* and *how*

- ***where***: (1) no data worth protecting available or (2) save the tuple itself ⇒ privacy aspects negligible or a hughe problem
- ***why***: if distribution of data is known and data is equal ⇒ privacy aspects could be a problem
- ***how***: too much information recoverable ⇒ privacy aspects are in all probability a problem

Solution Approaches:



No. 1: Generalization | No. 2: Intensional answers | No. 3: Differential privacy | No. 4: Permutation

Tanja Auge , Nic Scharlau , Andreas Heuer

INSTITUTE OF COMPUTER SCIENCE I DATABASE RESEARCH GROUP I
University of Rostock, Albert-Einstein-Straße 22 I 18059 Rostock, Germany