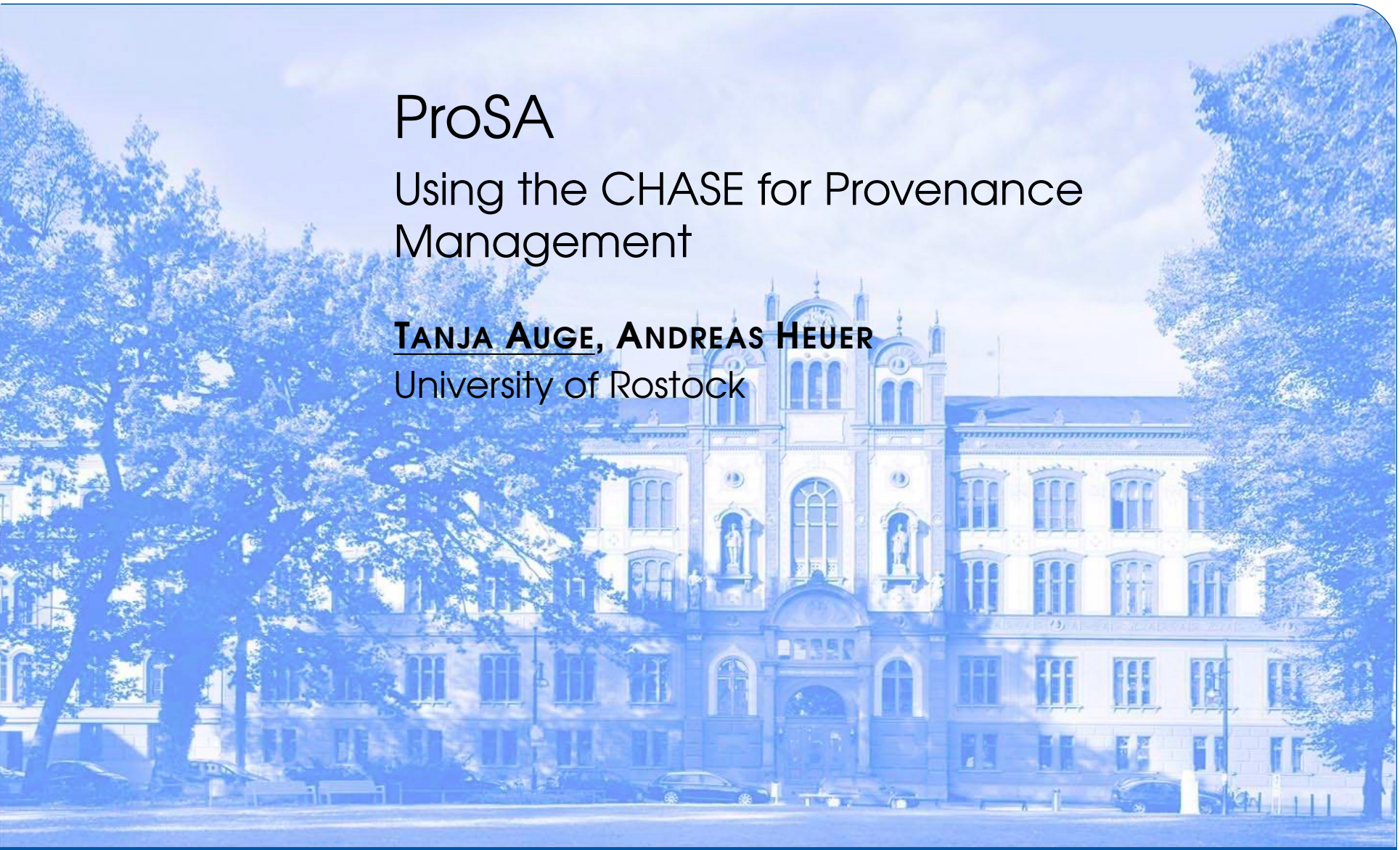
Universität Rostock Traditio et Innovatio



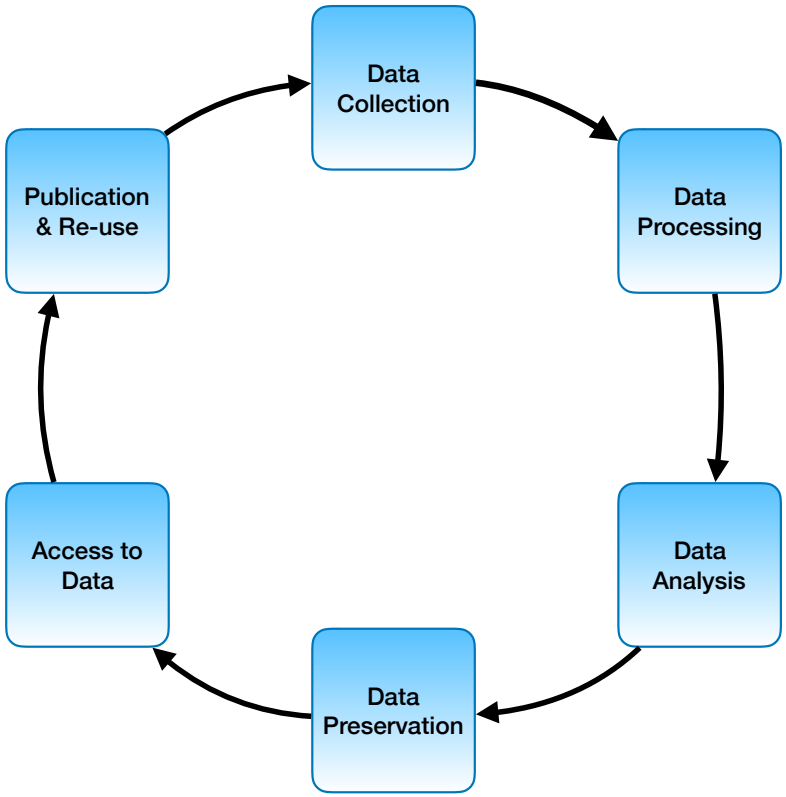
ProSA

Using the CHASE for Provenance Management

TANJA AUGE, ANDREAS HEUER
University of Rostock

Motivation

Research Data Management Lifecycle



Conditions:

- a long period of time
- a huge amount of data
- frequently change of data

Motivation

Provenance Management

- Traceability of a concrete result back to its possibly physical source
- Different provenance types
  - Data provenance
  - Metadata provenance
  - Workflow proveancne

Motivation

Data Provenance

1. How was a result actually achieved? $\Rightarrow$ provenance polynomials
2. Why was a certain result achieved? $\Rightarrow$ witness base
3. Where do the result tuples come from? $\Rightarrow$ table names

where-provenance (table name)	***why***-provenance (witness base)	***how***-provenance (polynomial)
R	$\{\{t_1\},\{t_2\},\{t_1,t_2\}\}$	$(t_1 \cdot (t_1 + t_2)) + (t_2 \cdot (t_1 + t_2))$

Motivation

Problem

Aim:

- Describing traceability, reconstructibility and replicability of result data
- Evaluation of provenance queries with changing data and schemas

⇒ Saving

- evaluation query
- result database
- minimal sub-database

ProSA

Using the CHASE for Provenance Management

(1) Calculation of a *minimal part of the original research database* to achieve replicable research

(2) Unification of provenance and evolution

Calculation of a minimal sub-database

additional information?

Constraints:

- Number of tuples of the original relation remains unchanged
- Sub-database can be mapped homomorphically to the original database
- Sub-database is an intensional description

$\Rightarrow$ Calculation of the provenance query Q_{prov} to determine the minimal sub-database

Unification of Provenance and Evolution

schema and data evolution

Aim:

- Evolution of provenance query Q_{prov}
- New database to be achieved, created after evolution

$\Rightarrow$ Calculation of a new combined provenance query Q'_{prov}

An Example

$I(S_1)$

	ID	Name	Subject	City
S_1	13	Paul	Math	Rostock
S_2	27	Tom	Math	Rostock
S_3	91	Lisa	Physics	Berlin
S_4	115	Maja	C.S.	Munich
S_5	317	Andrew	C.S.	Berlin

	ID	Module	Grade
G_1	13	Math	1.0
G_2	27	Math	2.3
G_3	27	Data	1.7
G_4	91	Theory	1.0
G_5	115	Math	4.0
G_6	115	Data	5.0
G_7	115	Law	1.3

An Example

Evaluation Query Q

$I(S_1)$

	ID	Name	Subject	City
S_1	13	Paul	Math	Rostock
S_2	27	Tom	Math	Rostock
S_3	91	Lisa	Physics	Berlin
S_4	115	Maja	C.S.	Munich
S_5	317	Andrew	C.S.	Berlin

	ID	Module	Grade
G_1	13	Math	1.0
G_2	27	Math	2.3
G_3	27	Data	1.7
G_4	91	Theory	1.0
G_5	115	Math	4.0
G_6	115	Data	5.0
G_7	115	Law	1.3

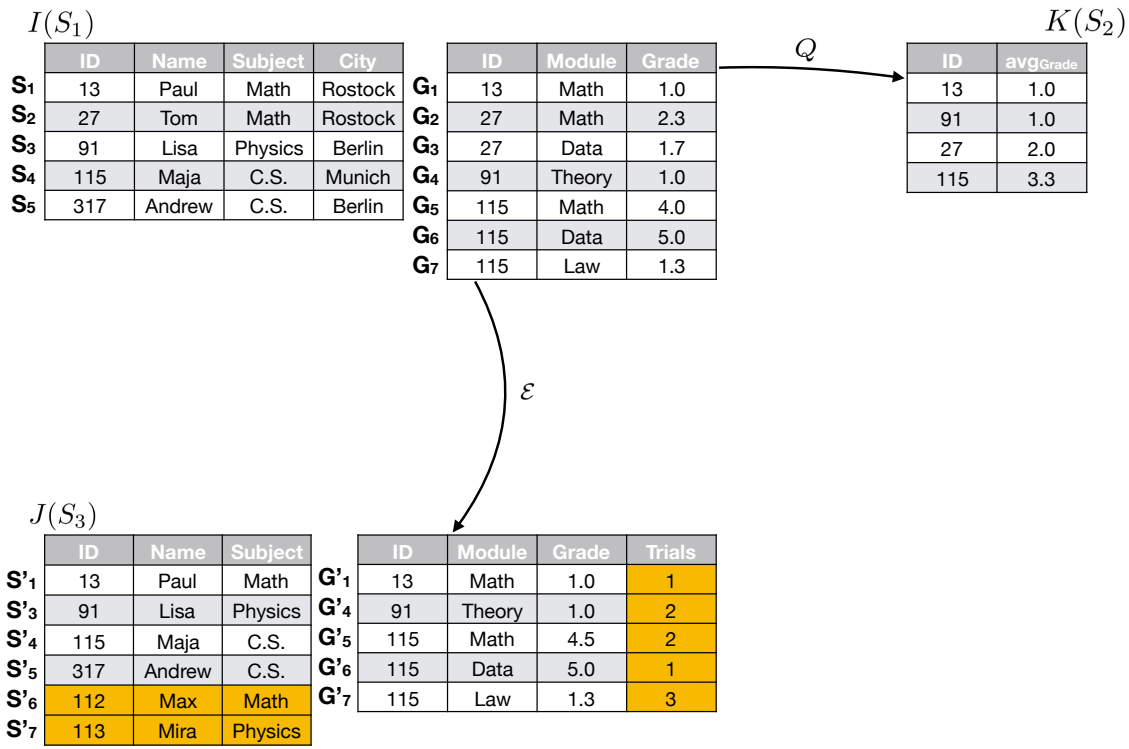
Q

$K(S_2)$

ID	$\text{avg}_{\text{Grade}}$
13	1.0
91	1.0
27	2.0
115	3.3

An Example

Evolution $\mathcal{E}$



$I(S_1)$

	ID	Name	Subject	City
S_1	13	Paul	Math	Rostock
S_2	27	Tom	Math	Rostock
S_3	91	Lisa	Physics	Berlin
S_4	115	Maja	C.S.	Munich
S_5	317	Andrew	C.S.	Berlin

	ID	Module	Grade
G_1	13	Math	1.0
G_2	27	Math	2.3
G_3	27	Data	1.7
G_4	91	Theory	1.0
G_5	115	Math	4.0
G_6	115	Data	5.0
G_7	115	Law	1.3

$K(S_2)$

ID	avg_{Grade}
13	1.0
91	1.0
27	2.0
115	3.3

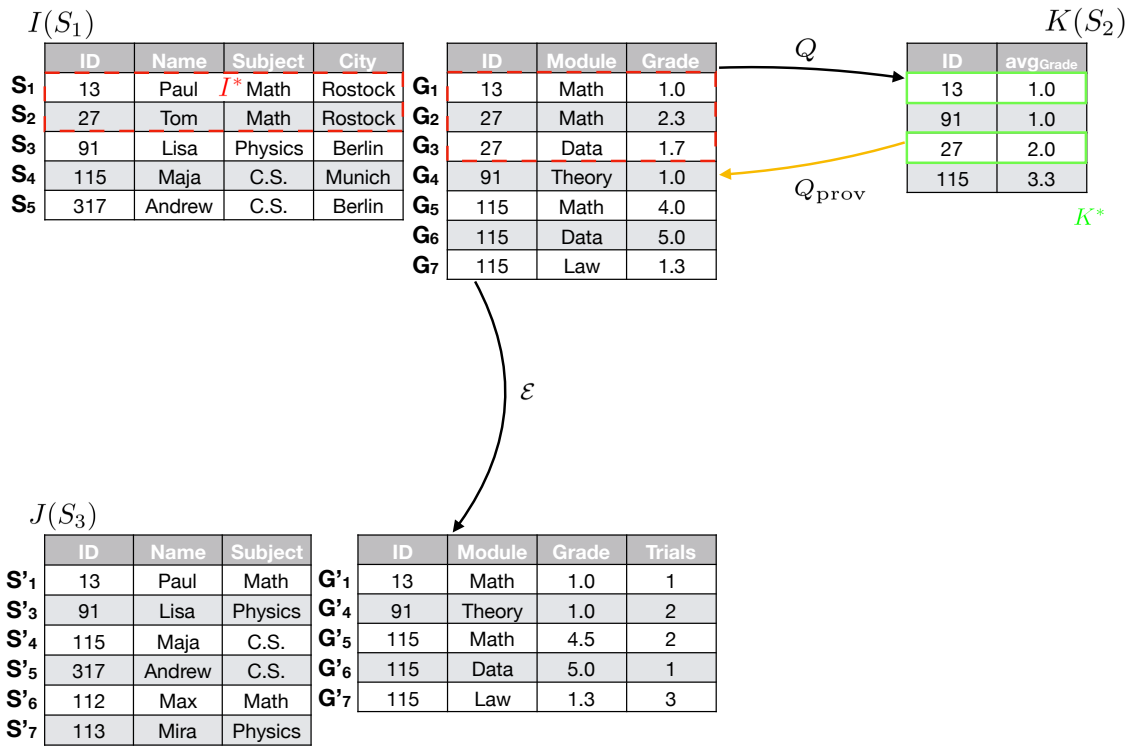
$J(S_3)$

	ID	Name	Subject
S'_1	13	Paul	Math
S'_3	91	Lisa	Physics
S'_4	115	Maja	C.S.
S'_5	317	Andrew	C.S.
S'_6	112	Max	Math
S'_7	113	Mira	Physics

	ID	Module	Grade	Trials
G'_1	13	Math	1.0	1
G'_4	91	Theory	1.0	2
G'_5	115	Math	4.5	2
G'_6	115	Data	5.0	1
G'_7	115	Law	1.3	3

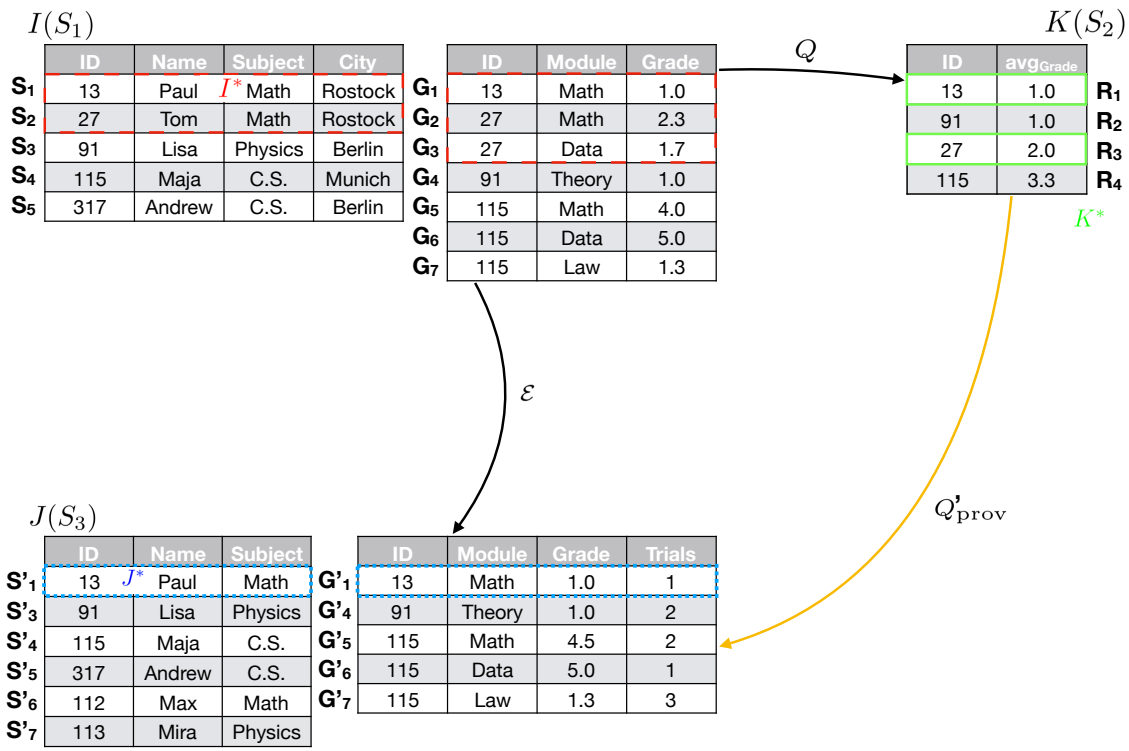
An Example

Provenance Query Q_{prov}



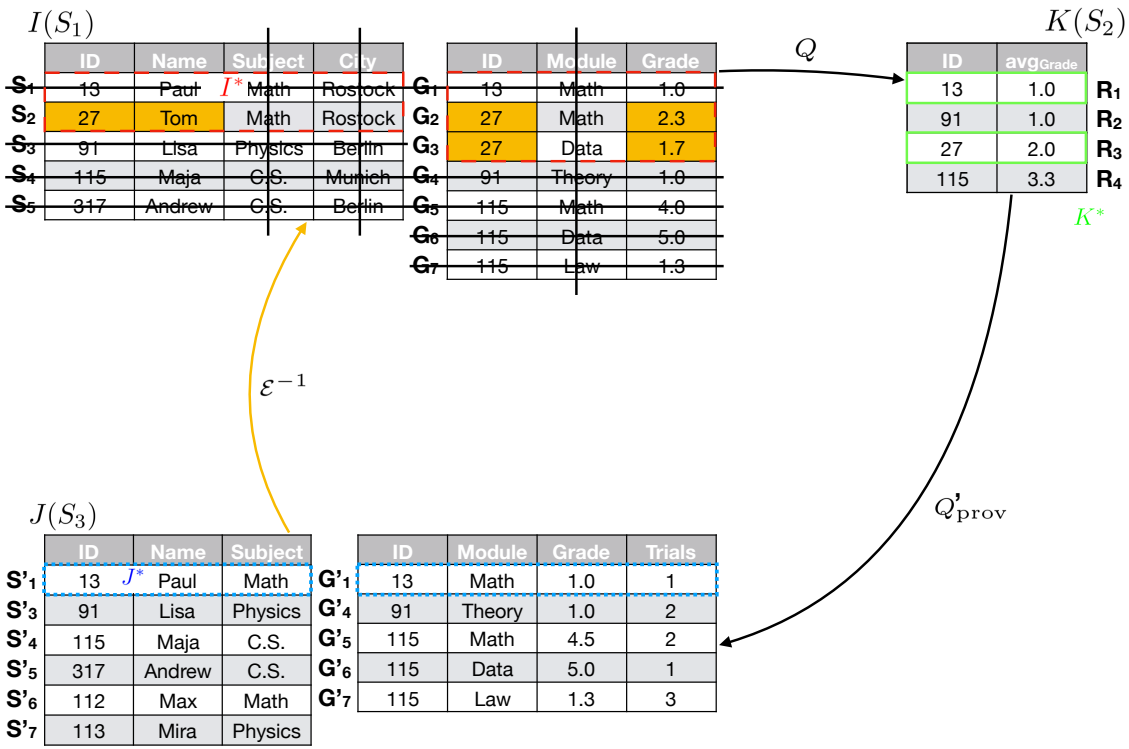
An Example

Updated Provenance Query Q'_{prov}

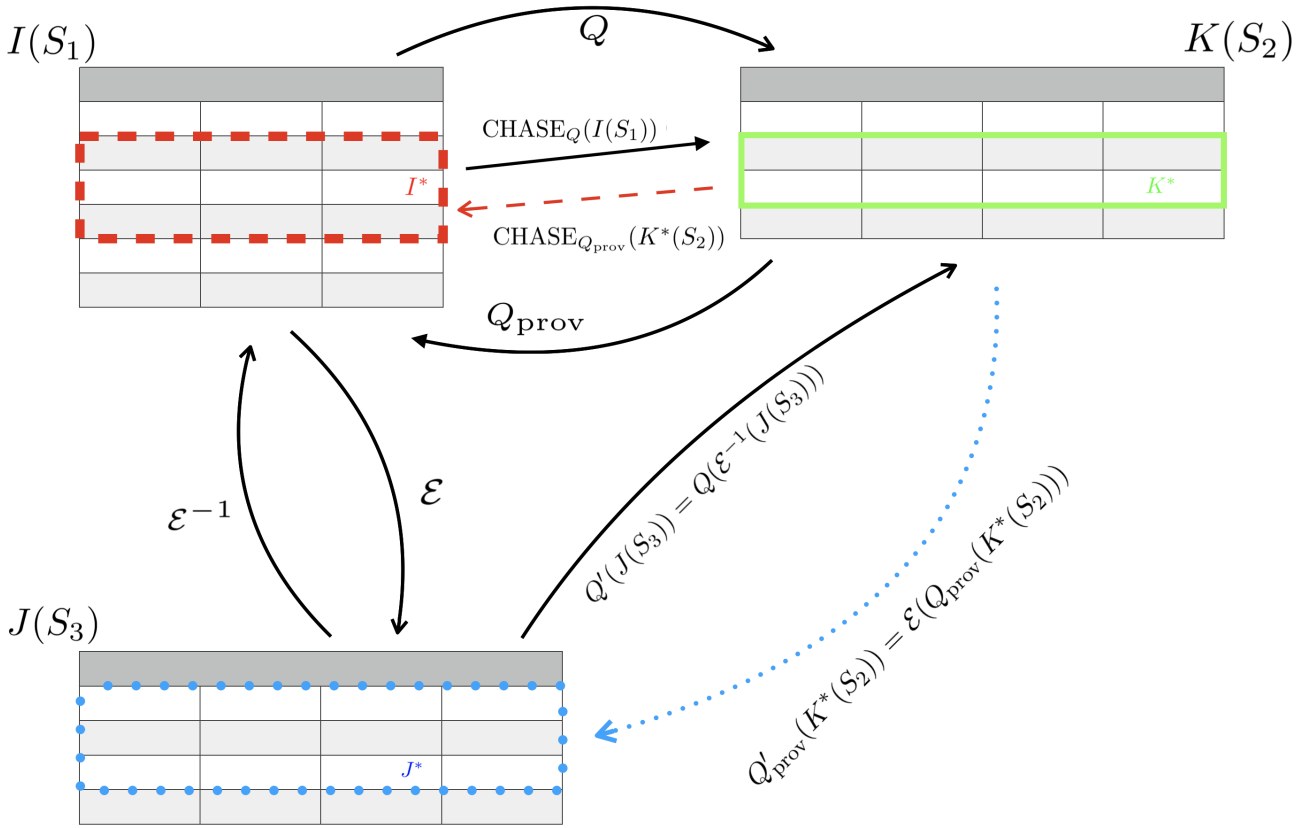


An Example

Additional Information



Using the CHASE for Provenance Management



CHASE algorithm

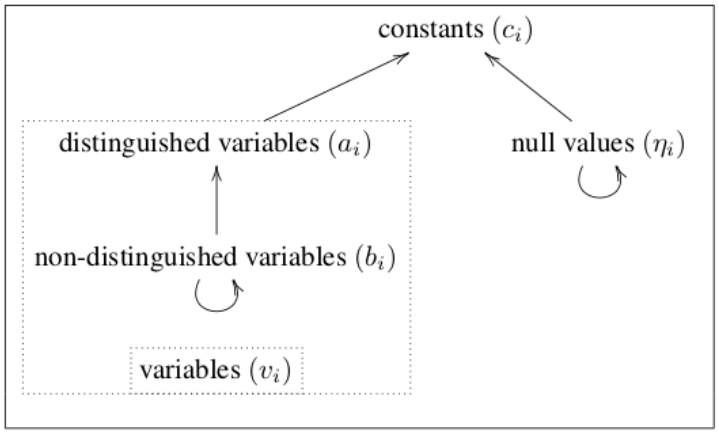
$$\text{chase}_\star(\bigcirc) = \circledast$$

- Different settings for CHASE object $\bigcirc$ and CHASE parameter $\star$
- Replaces data records by using null value or variable replacements
- Creates new tuple
- Existing theory of inverting the CHASE

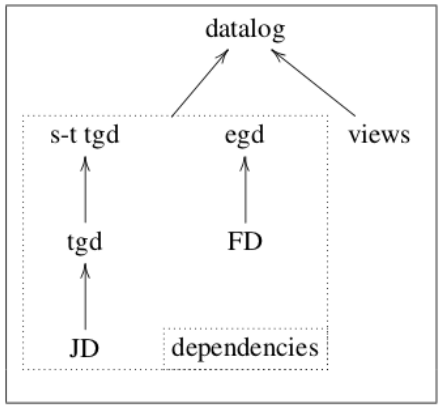
CHASE variants

	Parameter $\star$	Object $\bigcirc$	Result $\circledast$	Goal	Tool
0.	dependency	database schema	database schema with integrity constraint	optimized database design	
I.	dependency	query	query	semantic optimization	PDQ
II.	view	query	query using views	AQuV	ProvCB
II'.	operator	query	query using given operators	AQuO	
III.	s-t tgd, egd	source database	target database	data exchange, data integration	Llunatic, ChaseFUN
IV.	tgd, egd	database	modified database	cleaning	Llunatic, ChaseFUN
V.	tgd, egd	incomplete database	query result	certain answers	
VI.	s-t tgd, egd, tgd	database	query result	invertible query evaluation	

Unification of the CHASE

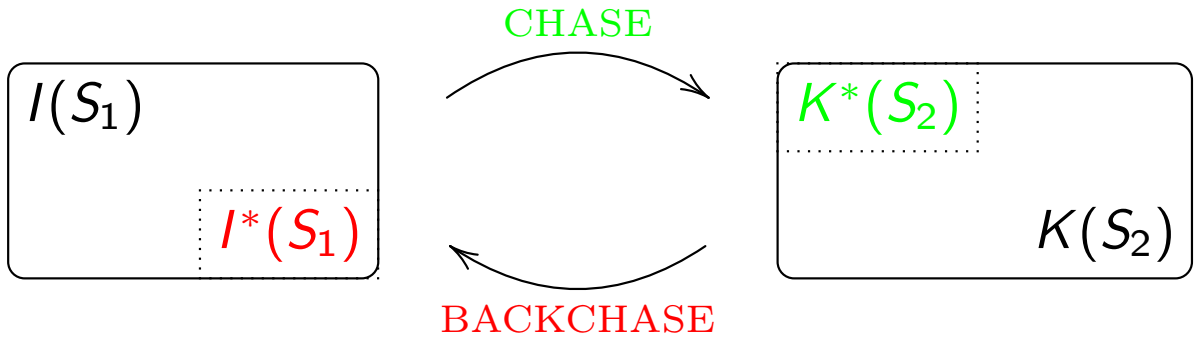


(a) CHASE object ◯



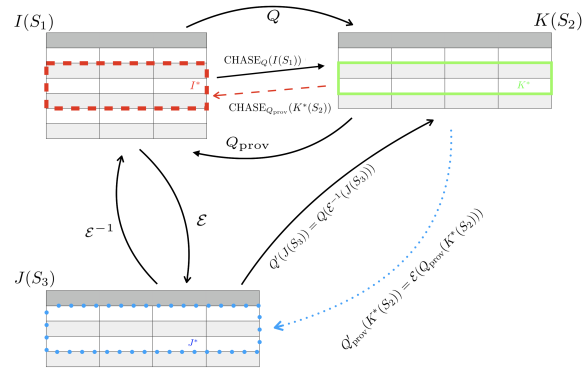
(b) CHASE parameter ⋆

CHASE&BACKCHASE



- CHASE phase: CHASE with $\bigcirc = I$ and $\star = Q$
- BACKCHASE phase: CHASE with $\bigcirc = K^\star$ and $\star = Q_{\text{prov}}$

Summary



- Research goals:
  1. Calculation of a minimal sub-database
  2. Unification of provenance and evolution
- Combination of CHASE&BACKCHASE
- Unification of the CHASE $\Rightarrow$ many application areas, but one tool