Schema Evolution in Research Data

Tanja Auge

University of Regensburg, Germany

11th March 2024

Universität Regensburg

Introduction

A quick internet search:

- schema evolution in databases
- research data management
- replicability and reproducibility
- FAIR principles and various additional guidelines (e.g. from the NFDI, DFG, RDA, ...)

Universität Regensburg

Introduction

A quick internet search:

- schema evolution in databases
- research data management
- replicability and reproducibility
- FAIR principles and various additional guidelines (e.g. from the NFDI, DFG, RDA, ...)

My background:

- Postdoc at the Chair of Data Engineering, University of Regensburg
- PhD topic: provenance management using schema mappings with annotations
- Research topics: provenance, research data management, schema evolution

Universität Regensburg

Research Data Management

Research data[1] is an essential foundation for scientific work. [...] Research data might include *measurement data*, *laboratory values*, *audiovisual information*, *texts*, *survey data*, *objects from collections*, or *samples* that were created, developed or evaluated during scientific work.

[1] https://www.dfg.de

[2] https://www.go-fair.org

Universität Regensburg

Research Data Management

Research data[1] is an essential foundation for scientific work. [...] Research data might include *measurement data*, *laboratory values*, *audiovisual information*, *texts*, *survey data*, *objects from collections*, or *samples* that were created, developed or evaluated during scientific work.

The **FAIR Principles**[2] provide guidelines to improve the **F***indability*, **A***ccessibility*, **I***nteroperability*, and **R***euse* of digital assets.

[1] https://www.dfg.de

[2] https://www.go-fair.org

Universität Regensburg

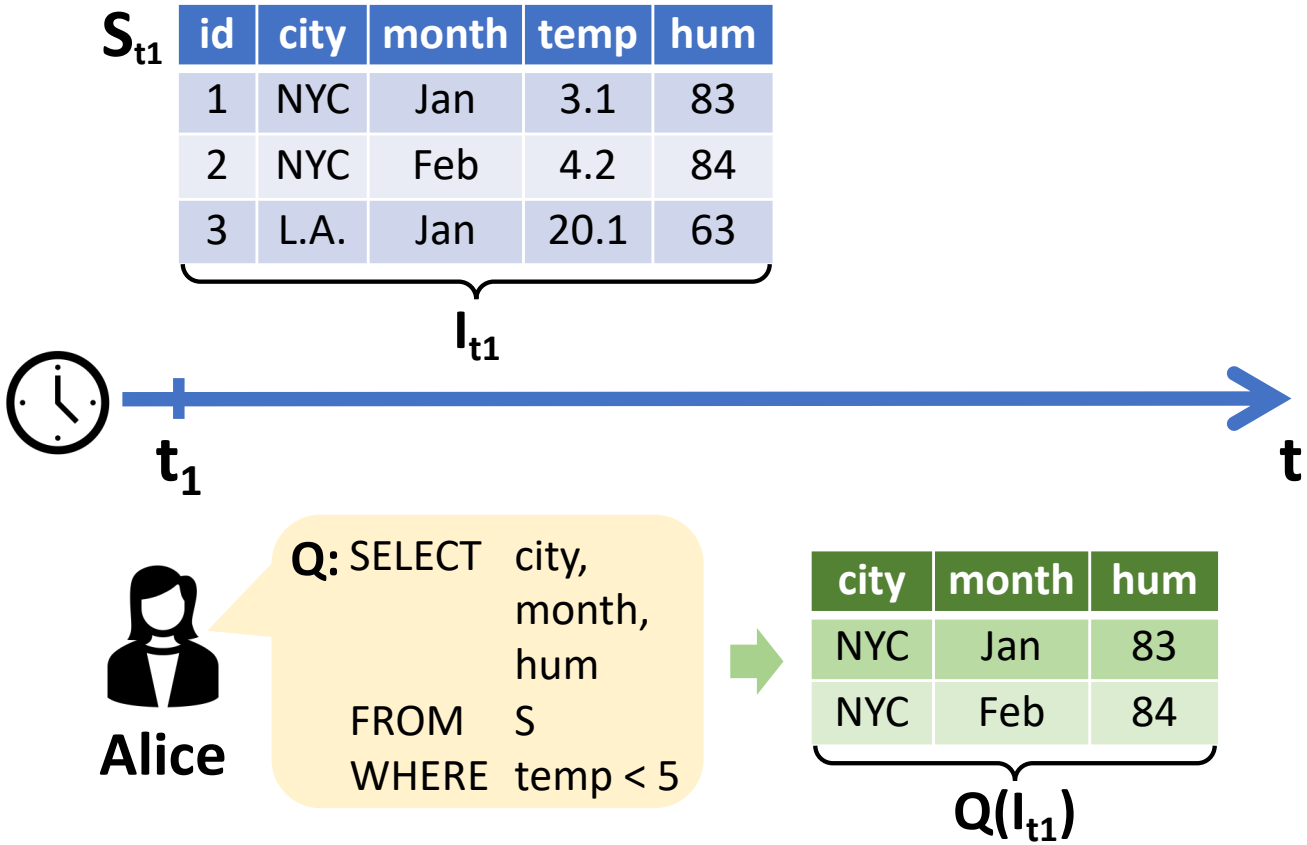
Research Data Lifecycle

Universität Regensburg

Research Data Lifecycle

Universität Regensburg

A sample Evaluation



Universität Regensburg

Challenges, Conditions and Goals

Challenges:

- schema changes and data updates over time
- ensuring reproducibility of data and query results (according to FAIR)

Universität Regensburg

Challenges, Conditions and Goals

Challenges:

- schema changes and data updates over time
- ensuring reproducibility of data and query results (according to FAIR)

Conditions in the context of research data:

- updating data means adding a new modified data record
- even in long-term studies, schema changes are rare
- privacy aspects in data and data handling

Universität Regensburg

Challenges, Conditions and Goals

Challenges:

- schema changes and data updates over time
- ensuring reproducibility of data and query results (according to FAIR)

Conditions in the context of research data:

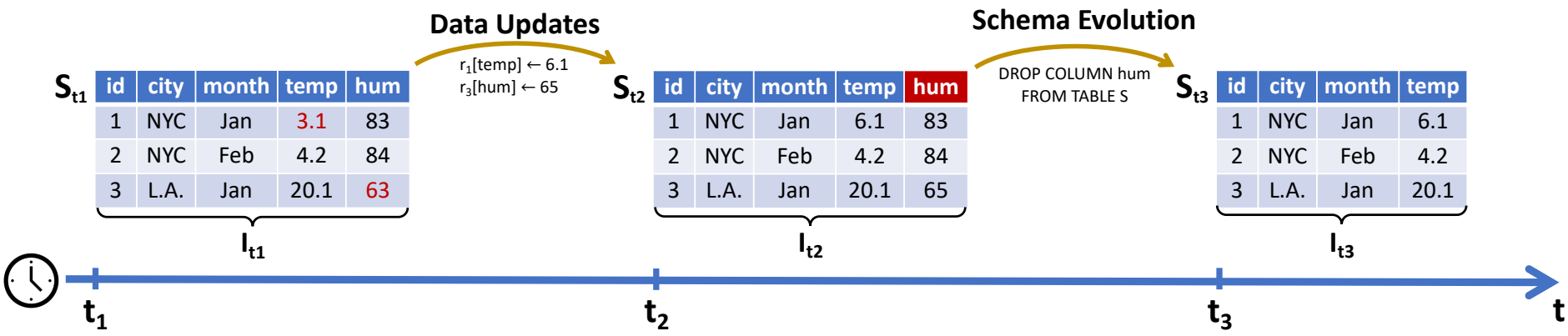
- updating data means adding a new modified data record
- even in long-term studies, schema changes are rare
- privacy aspects in data and data handling

Goals:

- improve the reproducibility of previous query results
- predict information loss

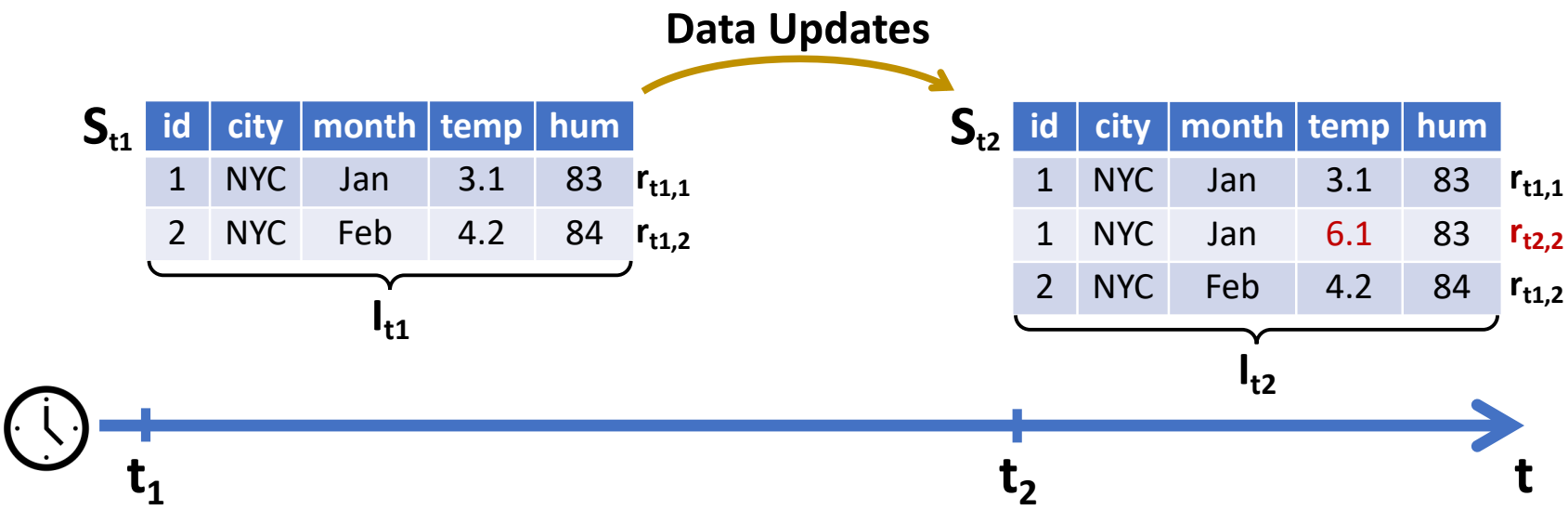
Universität Regensburg

A sample Evolution



Universität Regensburg

Updating Data



Universität Regensburg

Schema Evolution

Schema evolution[8] describes the *ability for a database schema to evolve* without the loss of existing information.

[8] J. F. Roddick: A survey of schema versioning issues for database systems. *Inf. Softw. Technol. 37(7)*, 1995

[9] H. J. Moon, C. Curino, A. Deutsch, C.-Y. Hou, C. Zaniolo: Managing and querying transaction-time databases under schema evolution. *Proc. VLDB Endow. 1(1)*, 2008

Universität Regensburg

Schema Evolution

Schema evolution[8] describes the *ability for a database schema to evolve* without the loss of existing information.

Schema Modification Operators (SMO)[9] are *representations of the mappings* between successive schema versions.

[8] J. F. Roddick: A survey of schema versioning issues for database systems. *Inf. Softw. Technol. 37(7)*, 1995

[9] H. J. Moon, C. Curino, A. Deutsch, C.-Y. Hou, C. Zaniolo: Managing and querying transaction-time databases under schema evolution. *Proc. VLDB Endow. 1(1)*, 2008

Universität Regensburg

Most common Schema Changes

	(a)	(b_1)	(b_2)	(c)
CREATE Table	2.3%	20.4%	29.1%	8.9%
DROP Table	0%	31.5%	23.6%	3.3%
MERGE Table	–	–	–	1.5%
ADD Column	79.5%	25.9%	25.5%	38.7%
DROP Column	0%	18.5%	18.2%	26.4%
RENAME Column	2.3%	–	–	16.0%
MERGE Column	9.2%	–	–	–
SPLIT Column	6.8%	–	–	–

(a) research database[3]

(b) open source programs[4,5]

(c) Wikipedia[6]

[3] T. Auge, E. Manthey, S. Jürgensmann, S. Feistel, A. Heuer: Schema evolution and reproducibility of long-term hydrographic data sets at the IOW. *LWDA*, 2020

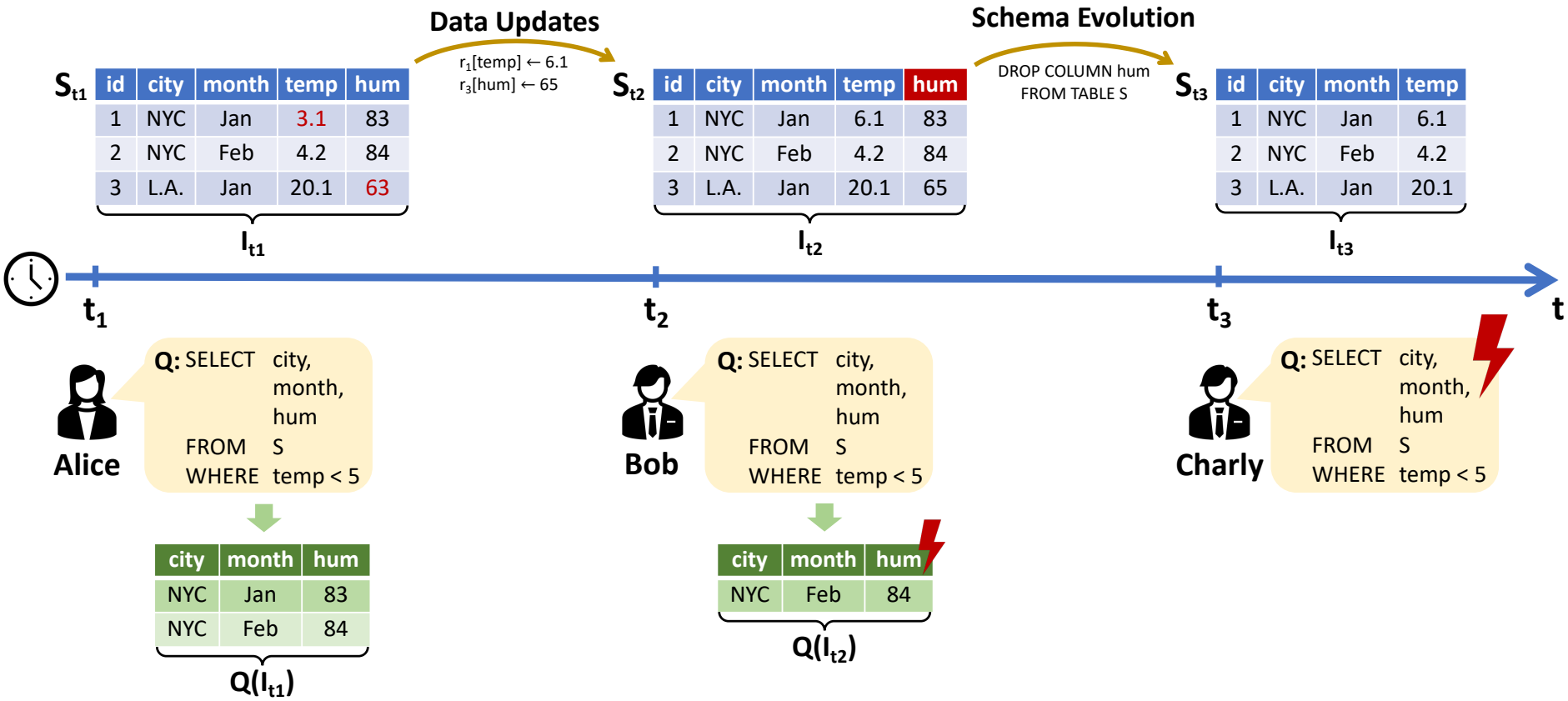
[4] S. Wu, I. Neamtiu: Schema evolution analysis for embedded databases. *ICDE Workshops*, 2011

[5] D. Brainінger, W. Mauerer, S. Scherzinger: Replicability and Reproducibility of a Schema Evolution Study in Embedded Databases. *ER (Workshops)*, 2020

[6] C. Curino, H. J. Moon, L. Tanca, C. Zaniolo: Schema evolution in wikipedia. *ICEIS(1)*, 2008

Universität Regensburg

A sample Evolution without Reproducibility



Universität Regensburg

Reproducibility by Provenance

Provenance[6] generally refers to any information that describes the *production process of an end product*, which can be anything from a piece of data to a physical object.

[6] M. Herschel, R. Diestelkämper, H. Ben Lahmar: A survey on provenance: What for? What form? What from? *VLDB J. 26(6)*, 2017

Universität Regensburg

Reproducibility by Provenance

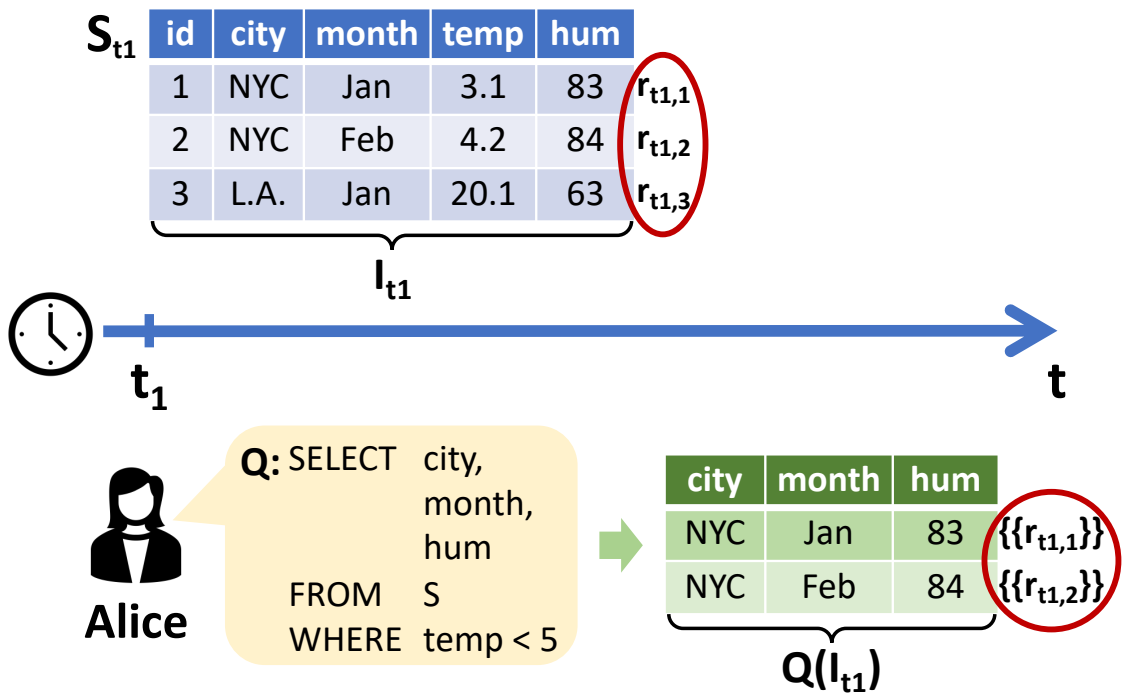
Provenance[6] generally refers to any information that describes the *production process of an end product*, which can be anything from a piece of data to a physical object.

Data provenance[6] allows to track the *processing of individual data items* at the level of individual data items (and the operations they undergo).

[6] M. Herschel, R. Diestelkämper, H. Ben Lahmar: A survey on provenance: What for? What form? What from? *VLDB J. 26(6)*, 2017

Universität Regensburg

A sample for a reproducible Evaluation



Universität Regensburg

Our conceptual Approach

Conceptual Idea:

- combine query evaluation, schema evolution and provenance information with one technique
- provenance information: data provenance and side tables

Universität Regensburg

Our conceptual Approach

Conceptual Idea:

- combine query evaluation, schema evolution and provenance information with one technique
- provenance information: data provenance and side tables

Goals:

- support plausibility checks
- verify reproducibility and replicability

Universität Regensburg

Our conceptual Approach

Conceptual Idea:

- combine query evaluation, schema evolution and provenance information with one technique
- provenance information: data provenance and side tables

Goals:

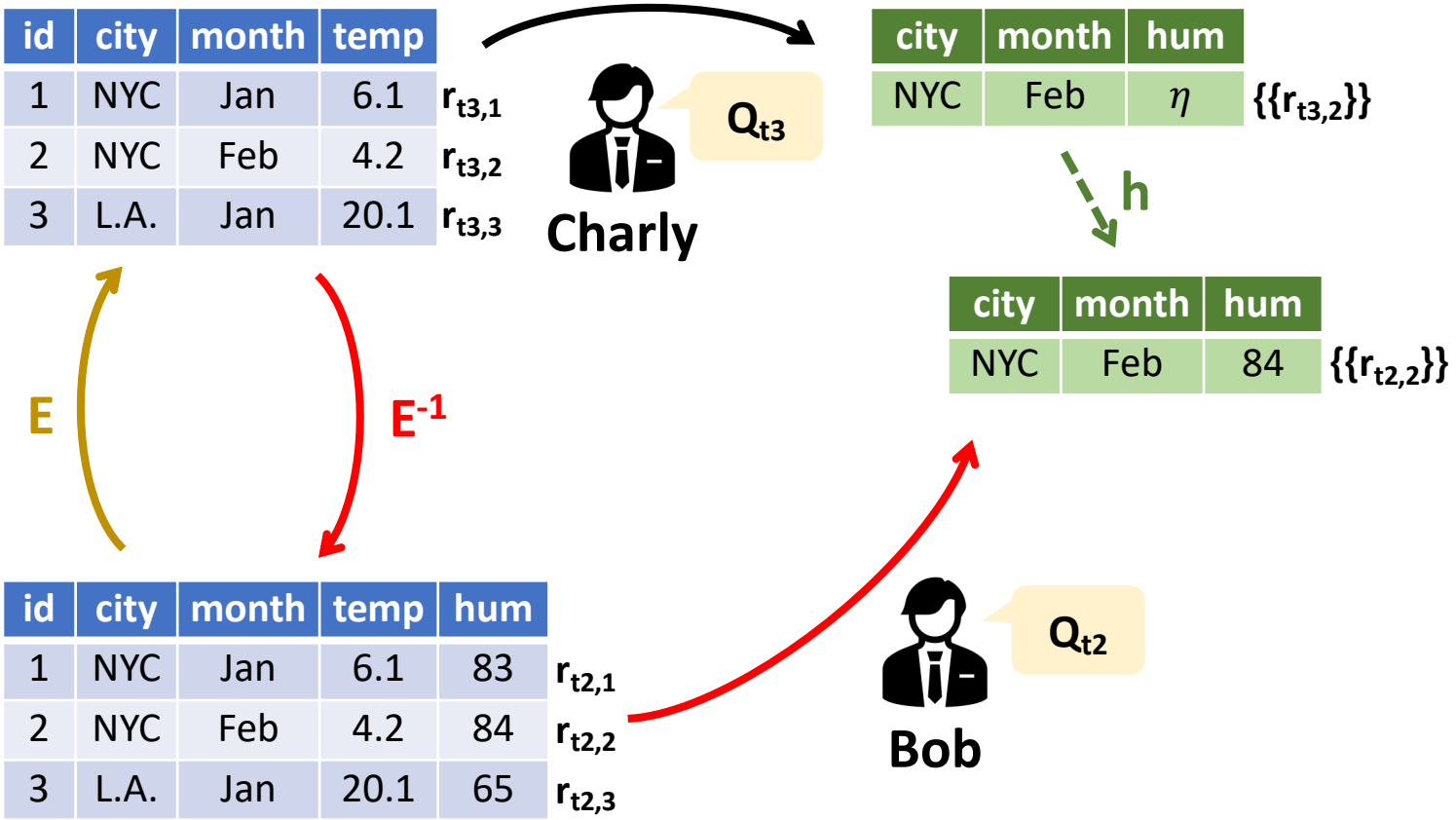
- support plausibility checks
- verify reproducibility and replicability

Furhter Steps:

- include data updates by extended time stamps
- extend the supported evaluation language

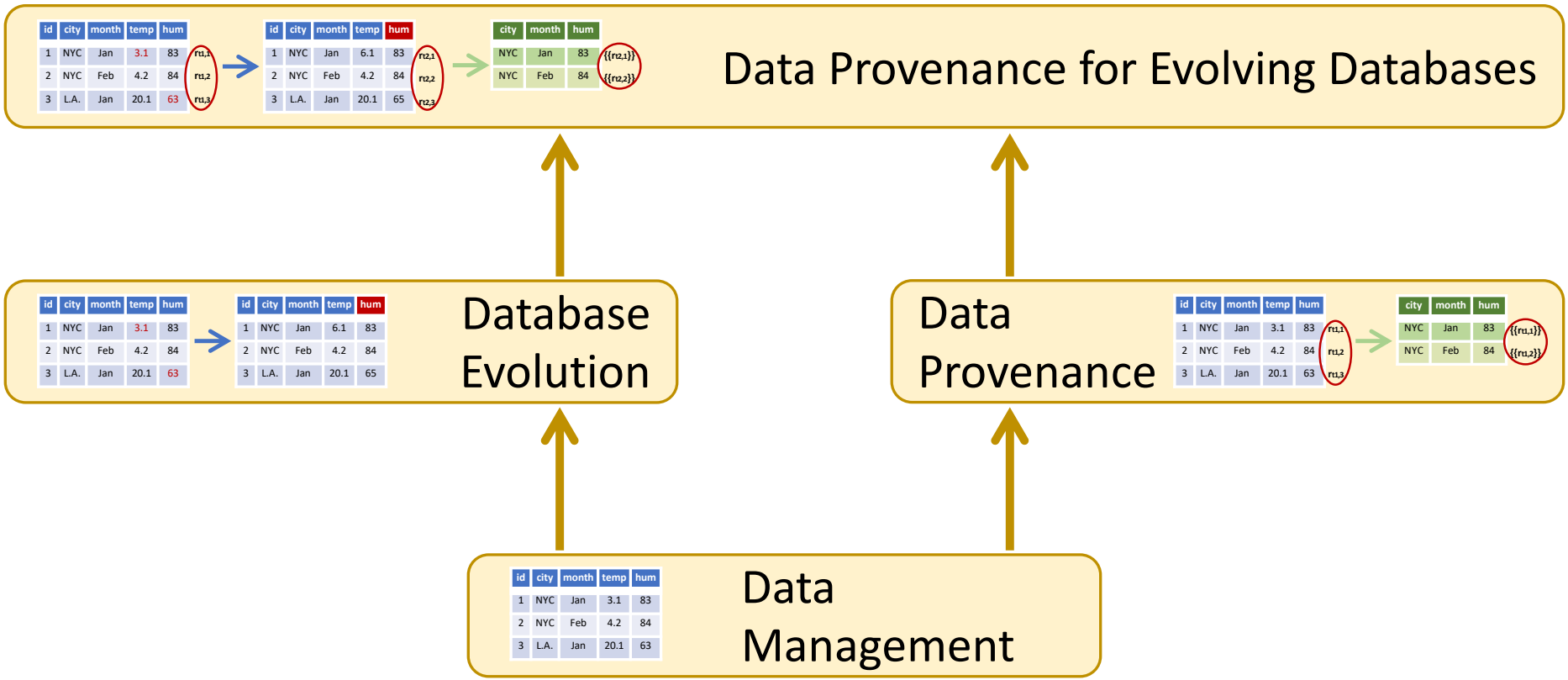
Universität Regensburg

A sample for a reproducible Evaluation including Schema Evolution



Universität Regensburg

Schema Evolution in Research Data



Universität Regensburg