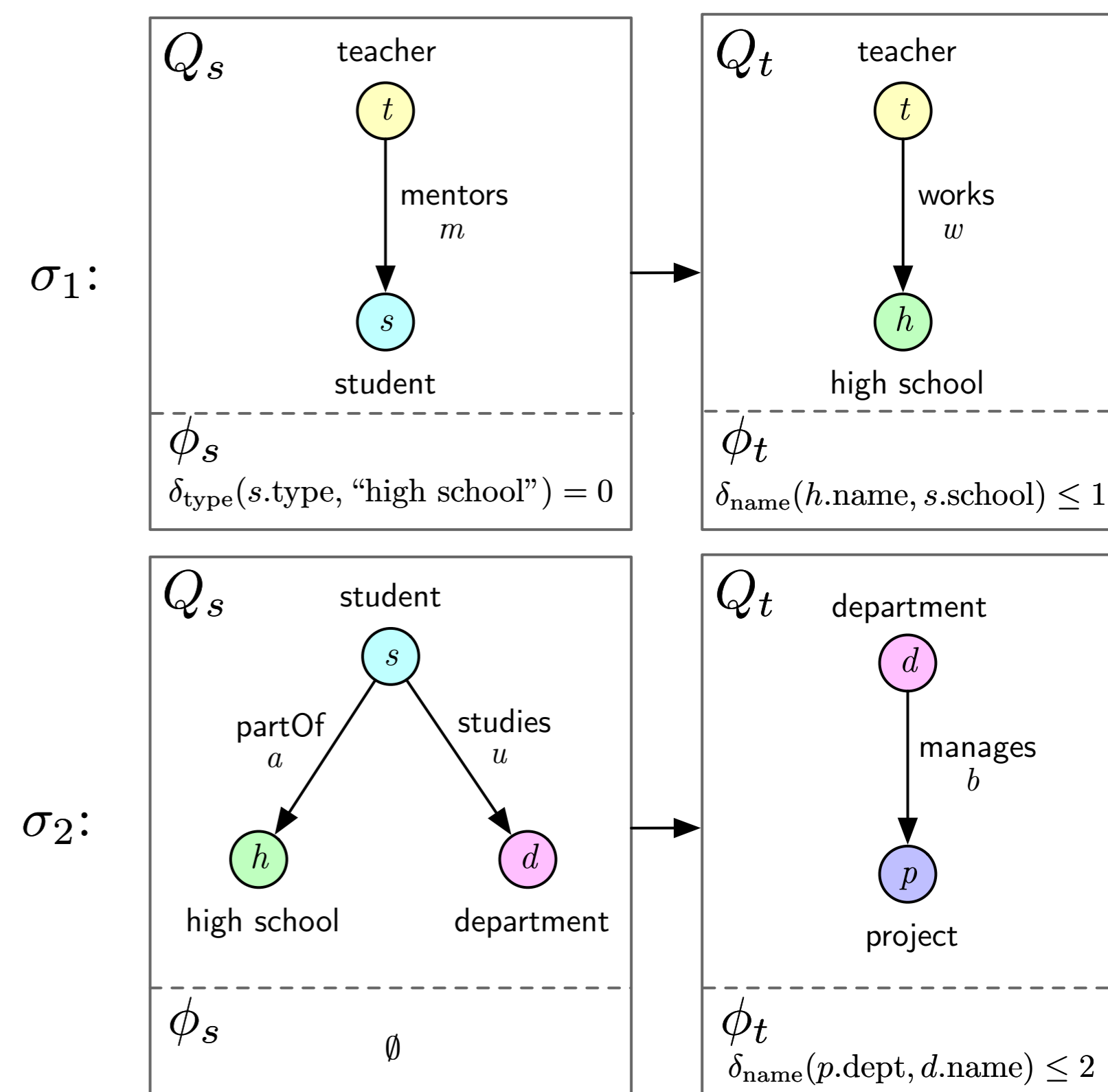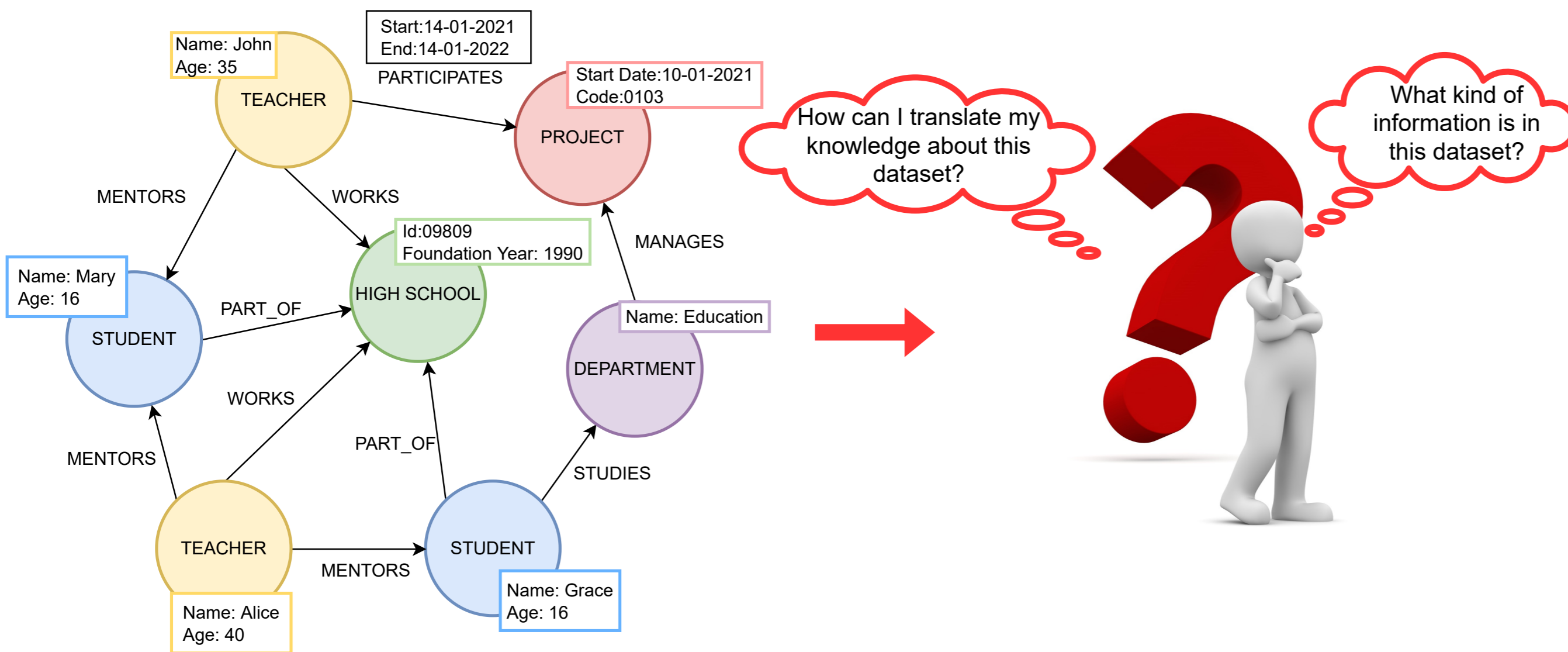# Data Profiling with Graph Generating Dependencies

Larissa C. Shimomura

l.capobianco.shimomura@tue.nl
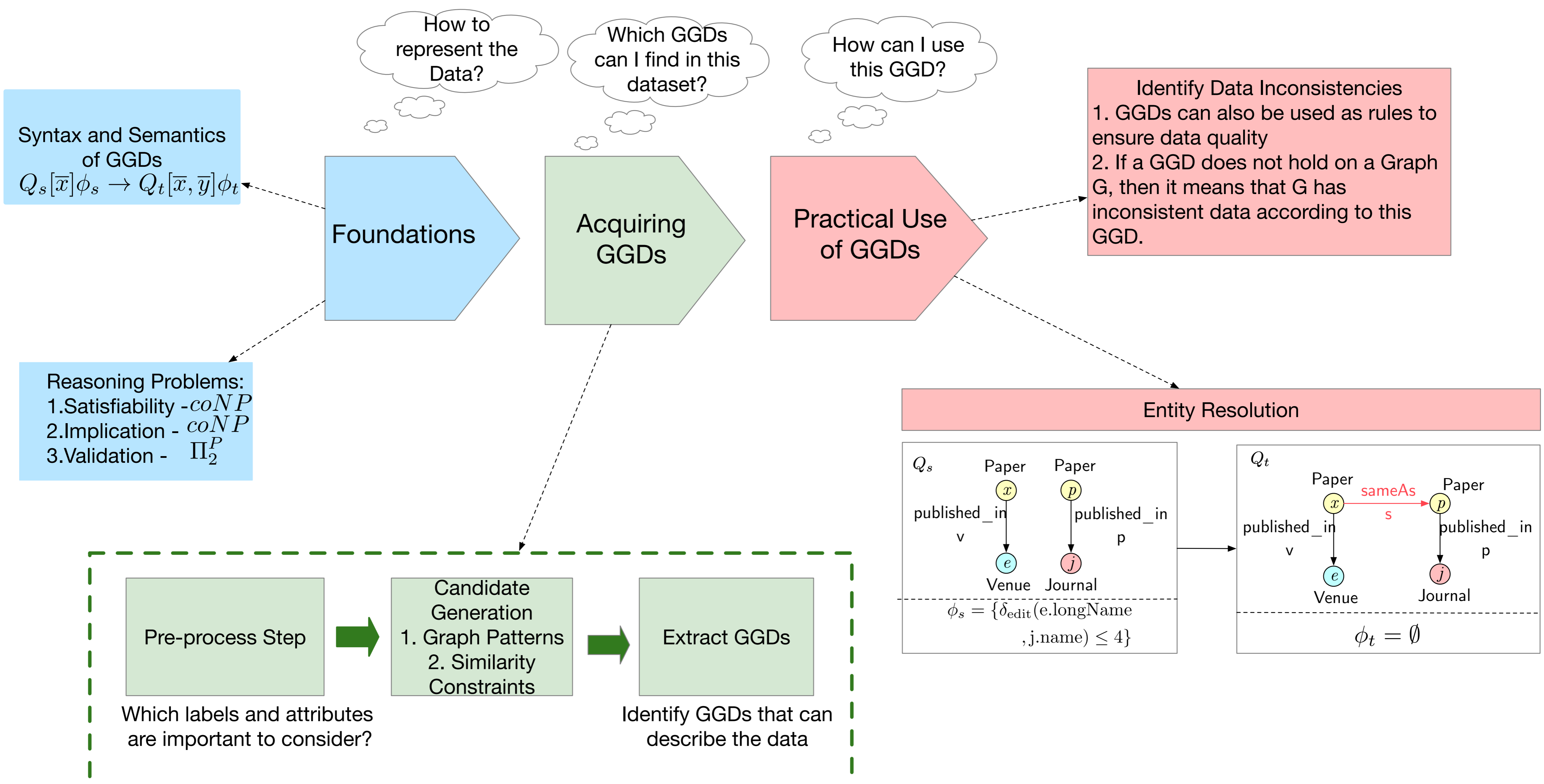
Supervised by: Nikolay Yakovets and George Fletcher

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Graph Generating Dependencies - GGDs

How can I translate my knowledge about this dataset?

What kind of information is in this dataset?

σ_1:

Q_s — teacher t, mentors m, student s

Q_t — teacher t, works w, high school h

ϕ_s: $\delta_{\text{type}}(s.\text{type}, \text{"high school"}) = 0$

ϕ_t: $\delta_{\text{name}}(h.\text{name}, s.\text{school}) \leq 1$

σ_2:

Q_s — student s, partOf a, high school h, studies u, department d

Q_t — department d, manages b, project p

ϕ_s: $\emptyset$

ϕ_t: $\delta_{\text{name}}(p.\text{dept}, d.\text{name}) \leq 2$

- Increasing use of graph-structured data sets in real-world applications, in which property graphs is an emerging standard.
- **Relationships in graph data are important!** - Interesting to capture dependencies which express whenever S happens then T should exist.

Project Overview

How to represent the Data?

Which GGDs can I find in this dataset?

How can I use this GGD?

Syntax and Semantics of GGDs
$Q_s[\overline{x}]\phi_s \rightarrow Q_t[\overline{x},\overline{y}]\phi_t$

Foundations → **Acquiring GGDs** → **Practical Use of GGDs**

Identify Data Inconsistencies
1. GGDs can also be used as rules to ensure data quality
2. If a GGD does not hold on a Graph G, then it means that G has inconsistent data according to this GGD.

Reasoning Problems:
1. Satisfiability - $coNP$
2. Implication - $coNP$
3. Validation - Π_2^P

Pre-process Step → **Candidate Generation** 1. Graph Patterns 2. Similarity Constraints → **Extract GGDs**

Which labels and attributes are important to consider?

Identify GGDs that can describe the data

Entity Resolution

Q_s — Paper x, published_in v, Venue e; Paper p, published_in p, Journal j

Q_t — Paper x —sameAs s→ Paper p; published_in v, Venue e; published_in p, Journal j

$\phi_s = \{\delta_{\text{edit}}(e.\text{longName}, j.\text{name}) \leq 4\}$

$\phi_t = \emptyset$

Conclusion

GGDs can express both topological constraints and also constraints over attributes as well as association between possibly different graph patterns which makes this type of dependency very expressive and fitting for a lot of use cases.

Smart Data Lake