Tutorial on Semantic Schema Discovery: principles, methods and future research directions

Part 2

Kenza Kellou-Menouer, Nikolaos Kardoulakis, Georgia Troullinou, Zoubida Kedad, Dimitris Plexousakis, Haridimos Kondylakis

ETIS Équipes Traitement de l'Information et Systèmes
ENSEA
FORTH Institute of Computer Science
david données et algorithmes pour une ville intelligente et durable
UVSQ université PARIS-SACLAY
CY CERGY PARIS UNIVERSITÉ

TEAM PRESENTATION

Kenza Kellou-Menouer
ETIS Lab, ENSEA Cergy, France

Nikolaos Kardoulakis
FORTH-ICS

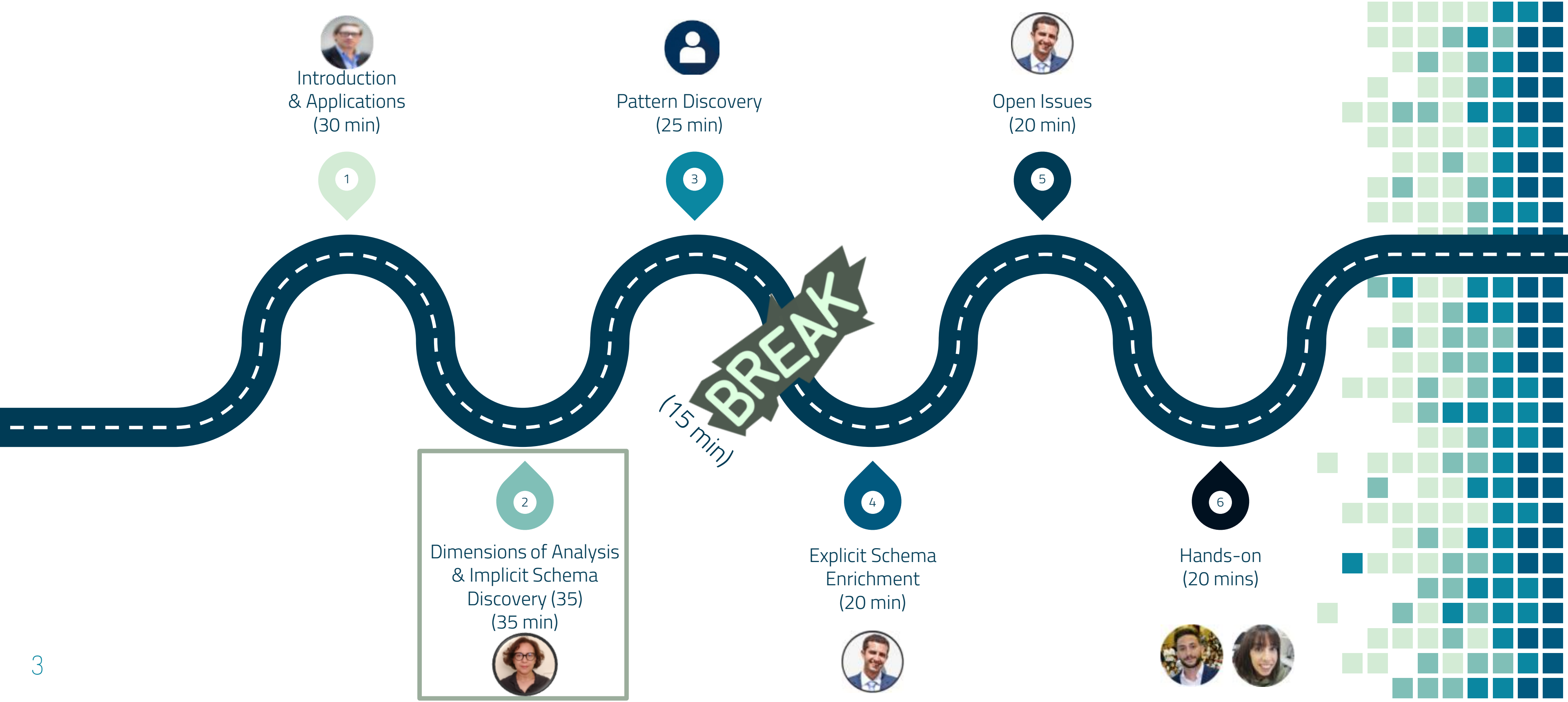
Georgia Troullinou
FORTH-ICS

Zoubida Kedad
University of Versailles St.-Quentin-en-Yvelines

Dimitris Plexousakis
FORTH-ICS

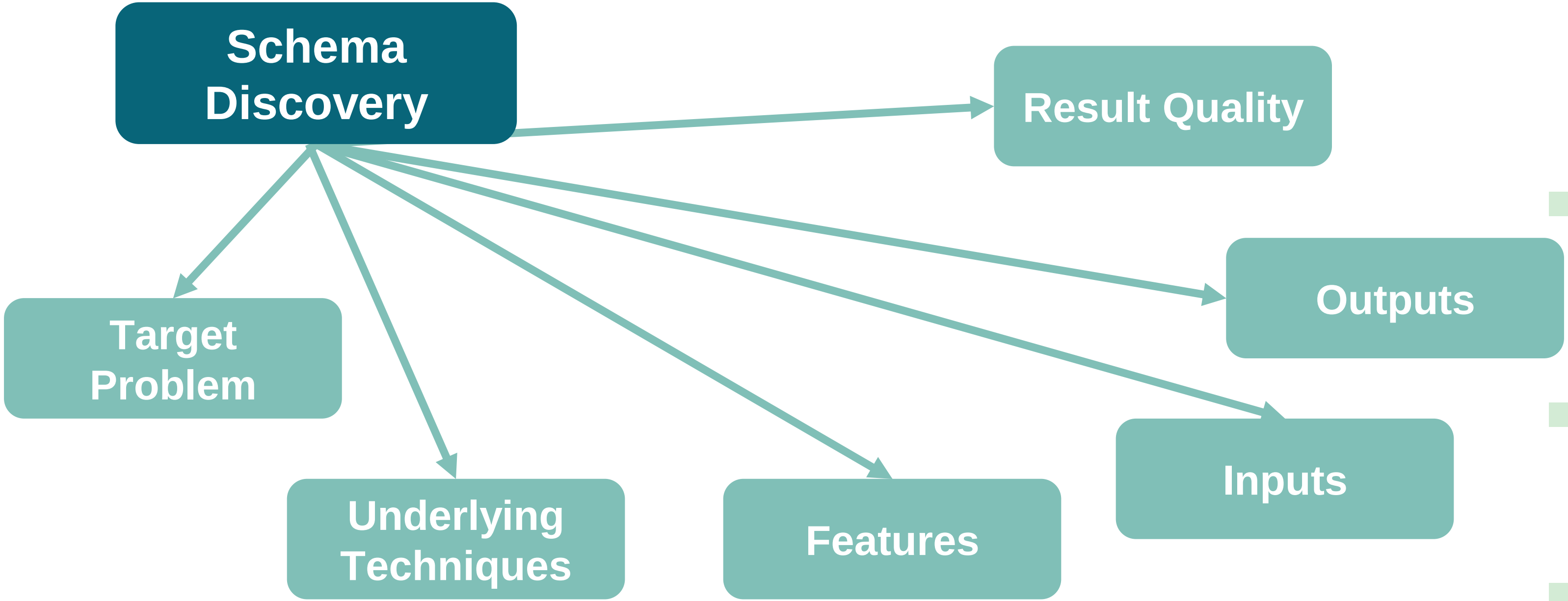
Haridimos Kondylakis
FORTH-ICS

ROADMAP


Introduction
& Applications
(30 min)
1
Pattern Discovery
(25 min)
3
Open Issues
(20 min)
5
BREAK
(15 min)
2
Dimensions of Analysis
& Implicit Schema
Discovery (35)
(35 min)
4
Explicit Schema
Enrichment
(20 min)
6
Hands-on
(20 mins)

Analysis Dimensions

Analysis Dimensions for Schema Discovery



Target Problem

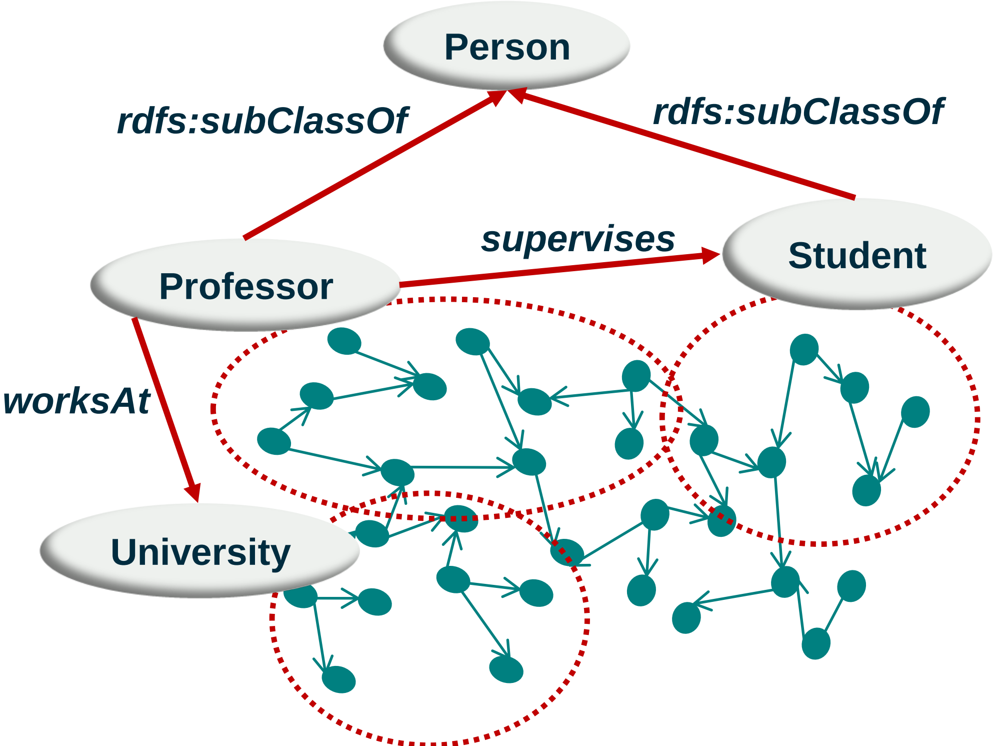
- Implicit Schema Discovery
- Explicit Schema Enrichment
- Pattern Discovery

Implicit Schema Discovery

- Schema discovery from the instances of the dataset
  - No additional information required
  - Based on grouping instances / paths

Implicit Schema Discovery

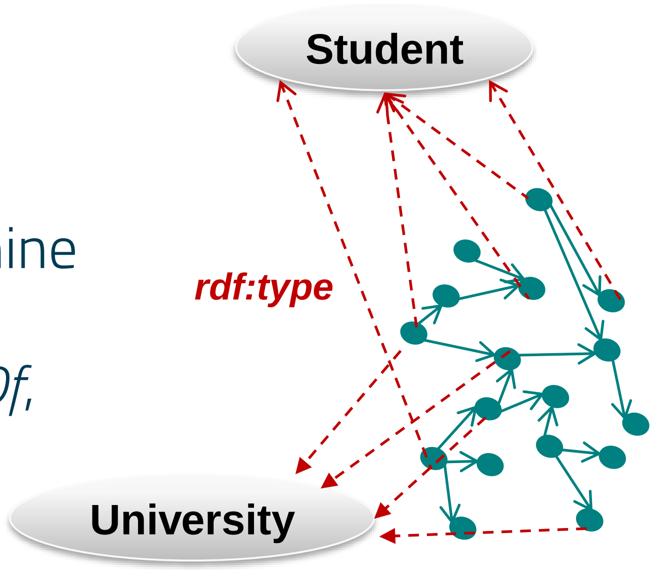
- Resulting schema
  - Classes / types : subsets of similar instances
  - Links between the classes



Explicit Schema Enrichment

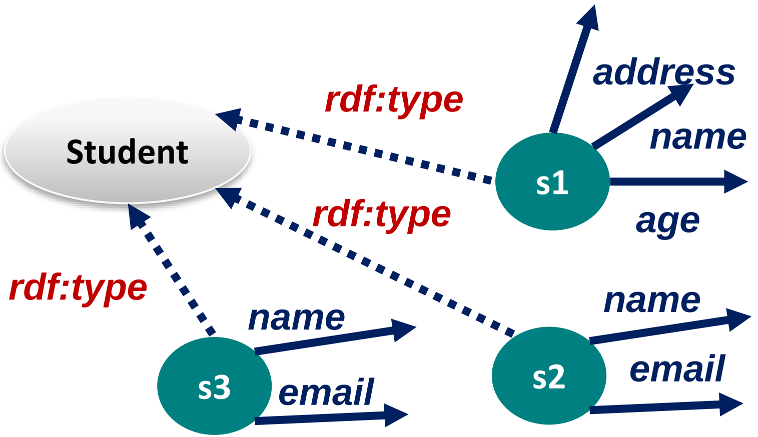
- Enriching the existing schema using the declarations provided in the dataset
  - *rdf:type, rdfs:domain, rdfs:range*

- Inference of new statements using machine learning or statistical approaches
  - *rdf:type, rdfs:subclassOf, rdfs:subPropertyOf, owl:SymetricProperty*



Structural Pattern Discovery

- Identifying all the existing patterns (versions) of the entities in a dataset / type



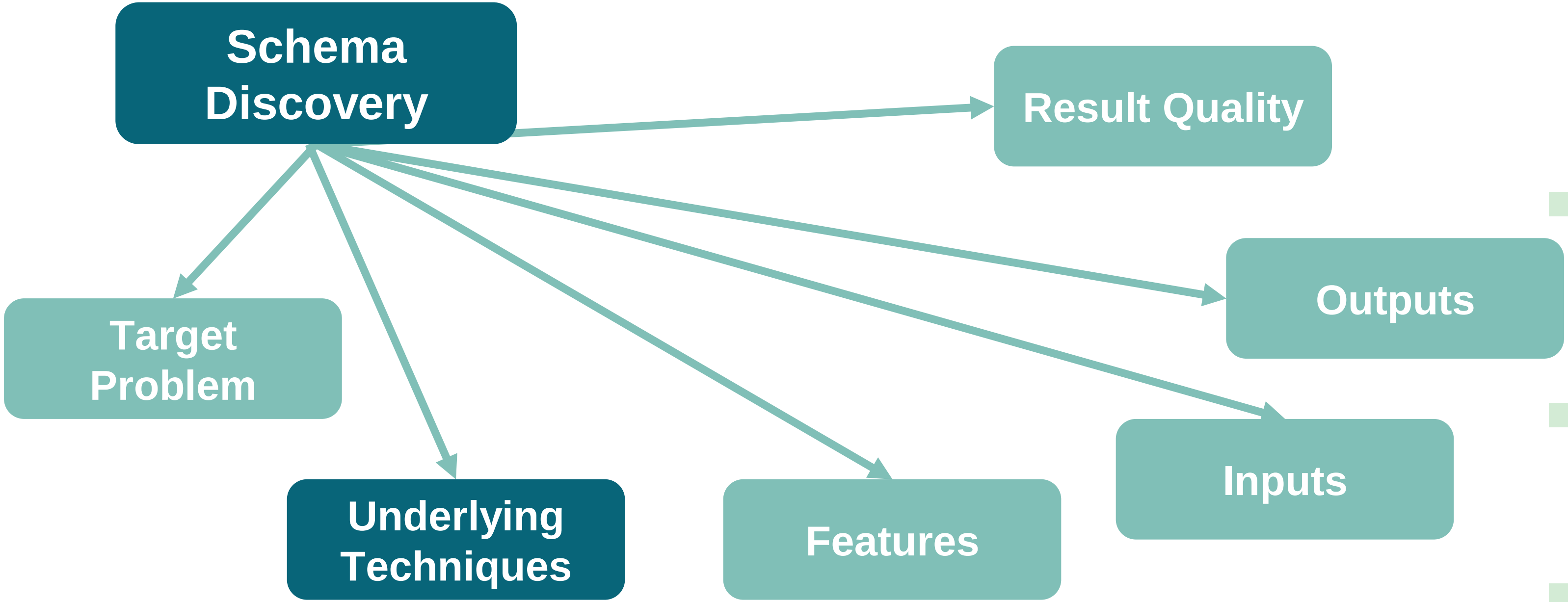
P1={Age, Name, Address}

P2={Name, Email}

Structural Pattern Discovery

- Characterizing the co-occurrence relationships among the properties of the dataset
- Output: Exact or Approximate patterns

Analysis Dimensions for Schema Discovery



Underlying Techniques for Schema Discovery

- Machine learning
  - Supervised learning algorithms (classification)
  - Unsupervised learning algorithms (clustering, frequent pattern mining)
- Formal methods
  - Formal Concept Analysis, Bisimulation
- Statistical techniques
  - Frequency or distribution of the properties

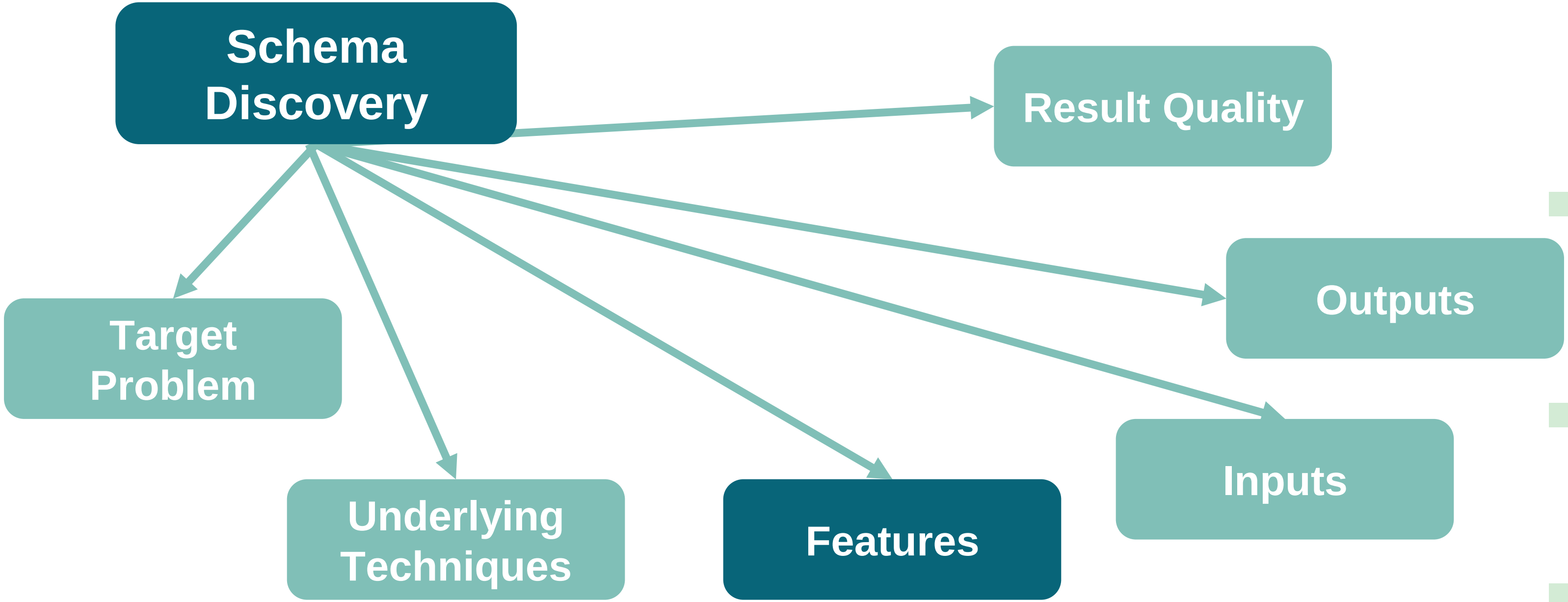
Machine Learning Algorithms

- Classification algorithms → **Explicit schema enrichment using existing type definitions**
  - K-NN
- Clustering algorithms → **Implicit type discovery by grouping similar instances**
  - K-means, Dbscan, H clustering
- Frequent pattern mining → **Discovering association rules or structural patterns**
  - Apriori

Other Techniques

- Bisimulation → **Grouping similar paths**
- Statistical techniques → **Analysing property distribution to infer new type declarations**
- Formal Concept Analysis → **Implicit type discovery**

Analysis Dimensions for Schema Discovery



Scalability

- Ability of the existing approaches to deal with massive datasets
- Highly depens on the underlying technique and computational complexity of the algorithm

Stability

- Providing the same schema for different executions of the schema discovery algorithm on the same dataset
- Dependent on the sensitivity of the underlying algorithm to the exploration order of the dataset

Incrementality

- Dealing with the changes occurring in the dataset and propagating these changes into the schema
- Ability to incrementally adapt the existing schema instead of generating a new one

Hybrid Approaches

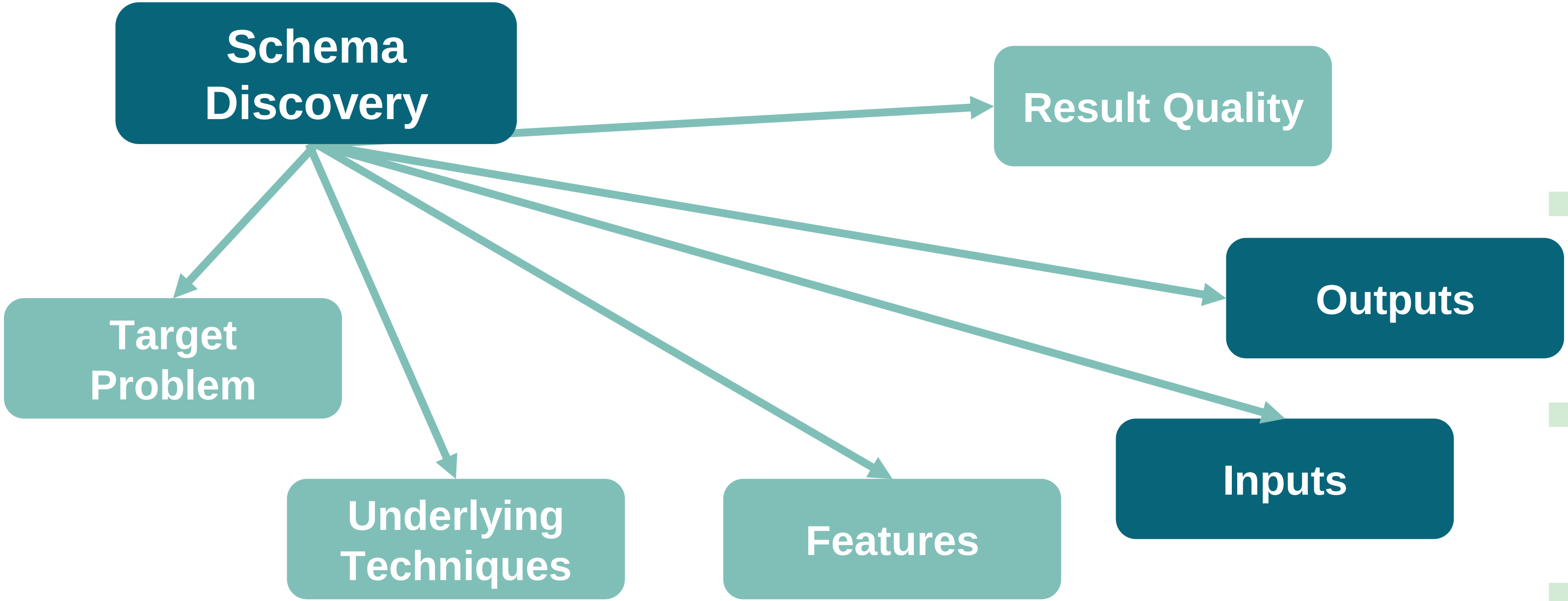
- Ability to exploit both the instances and the schema related information when provided

- Taking into account the existing schema related statements during schema discovery

Online Schema Discovery

- Ability to process remote datasets that can not be copied locally
- Coping with access restrictions enforced by the server
  - Number of issued queries, size of the result, etc.

Analysis Dimensions for Schema Discovery



Inputs

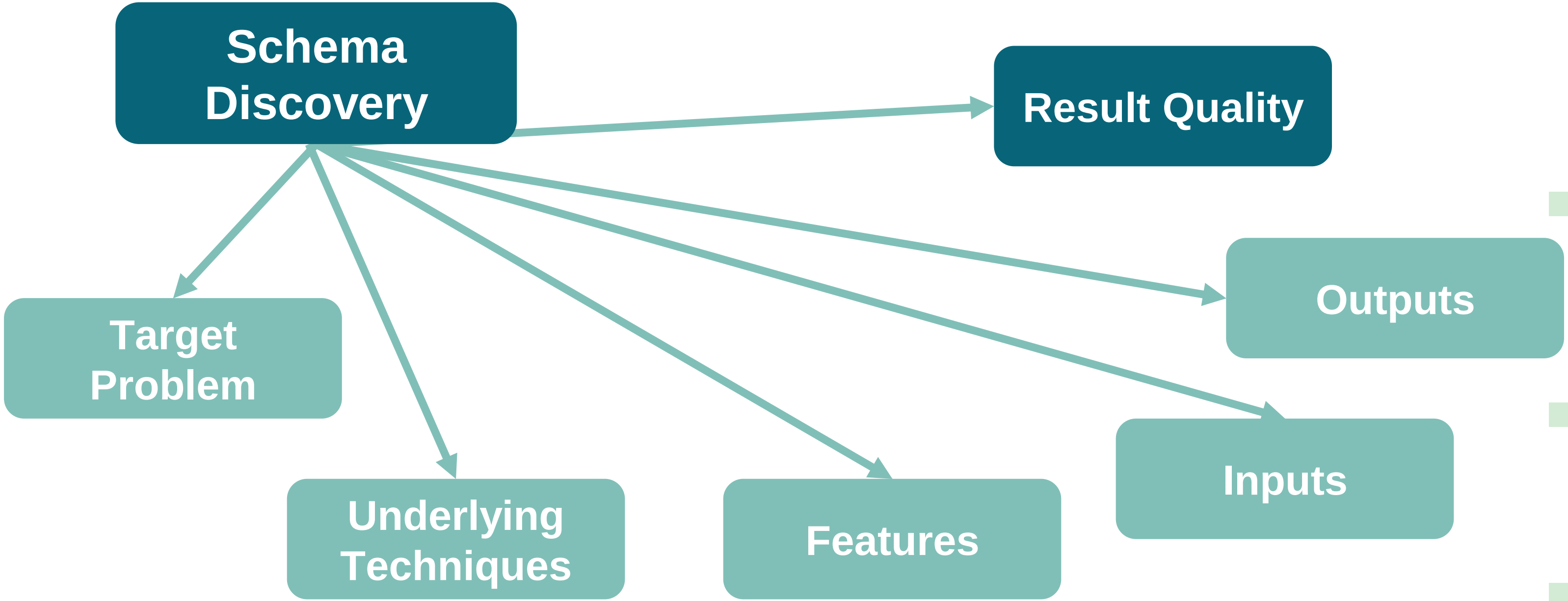
- User Defined Parameters
  - Required by the algorithms used for schema discovery
    - Similarity thresholds, number of clusters, etc.

- Dataset-Related Inputs
  - Schema declarations
    - RDF Type definitions, RDFS / OWL classes and sub-classes, OWL ontologies

Outputs

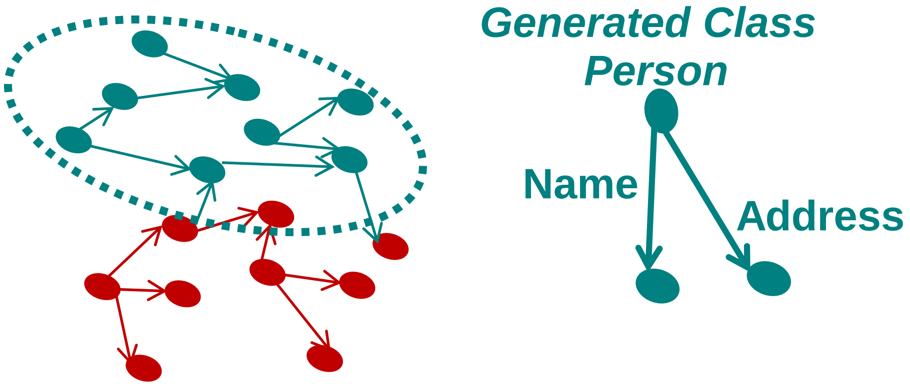
- Types
  - *rdf:type* statements
- Semantic links
  - Ex: *rdfs:domain*, *rdfs:range* statements
- Hierarchical links
  - Ex: *rdfs:subClassOf*, *rdfs:subPropertyOf*
- Patterns / co-occurrence of properties
- Path plans

Analysis Dimensions for Schema Discovery



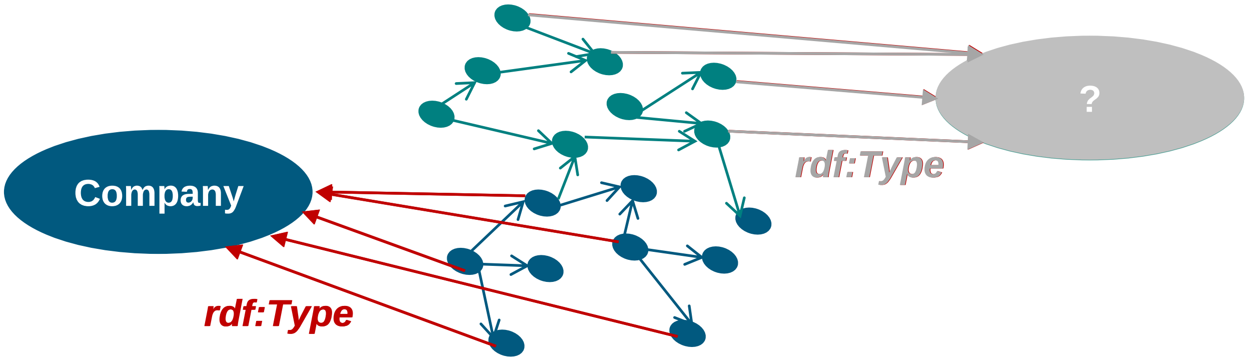
Schema Completeness

- Implicit schema discovery approaches
  - Comparing the generated classes to the actual classes of the instances: have all the classes been extracted ?



Schema Completeness

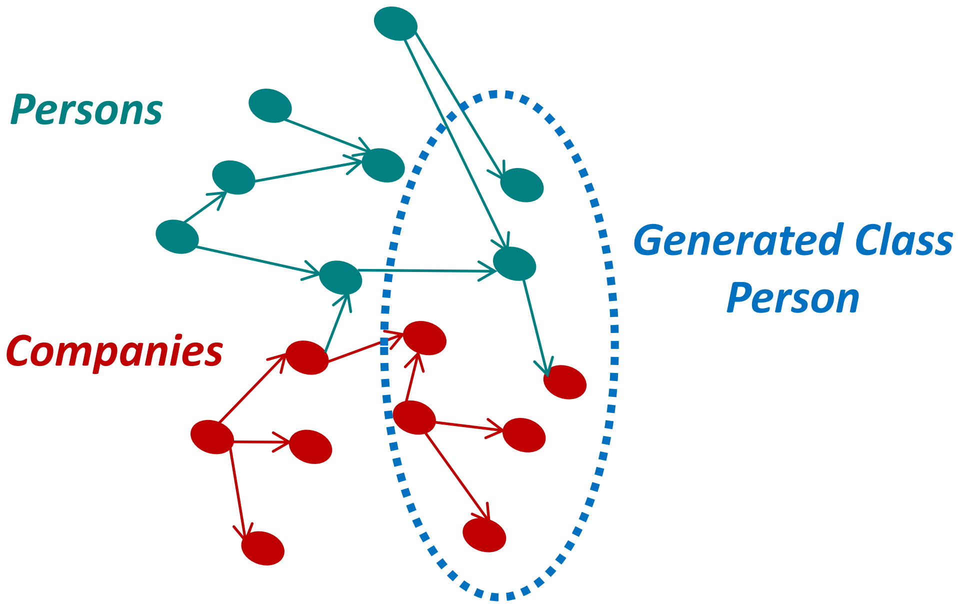
- Explicit schema enrichment approaches
  - The completeness of the generated declarations depends on the completeness of the existing ones



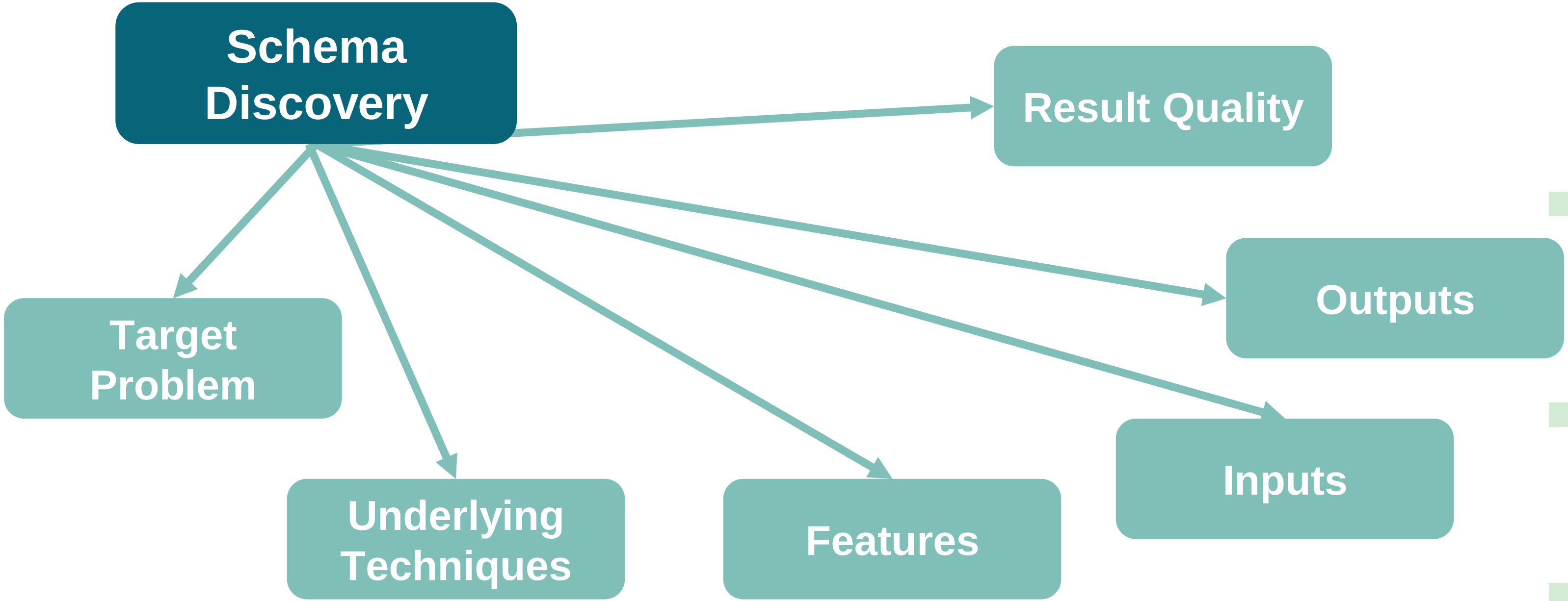
Class Accuracy

- Implicit schema discovery
  - Are the instances grouped in a generated class actually instances of this class?
- Explicit schema enrichment
  - Are the instances assigned to an existing class class actually instances of this class?

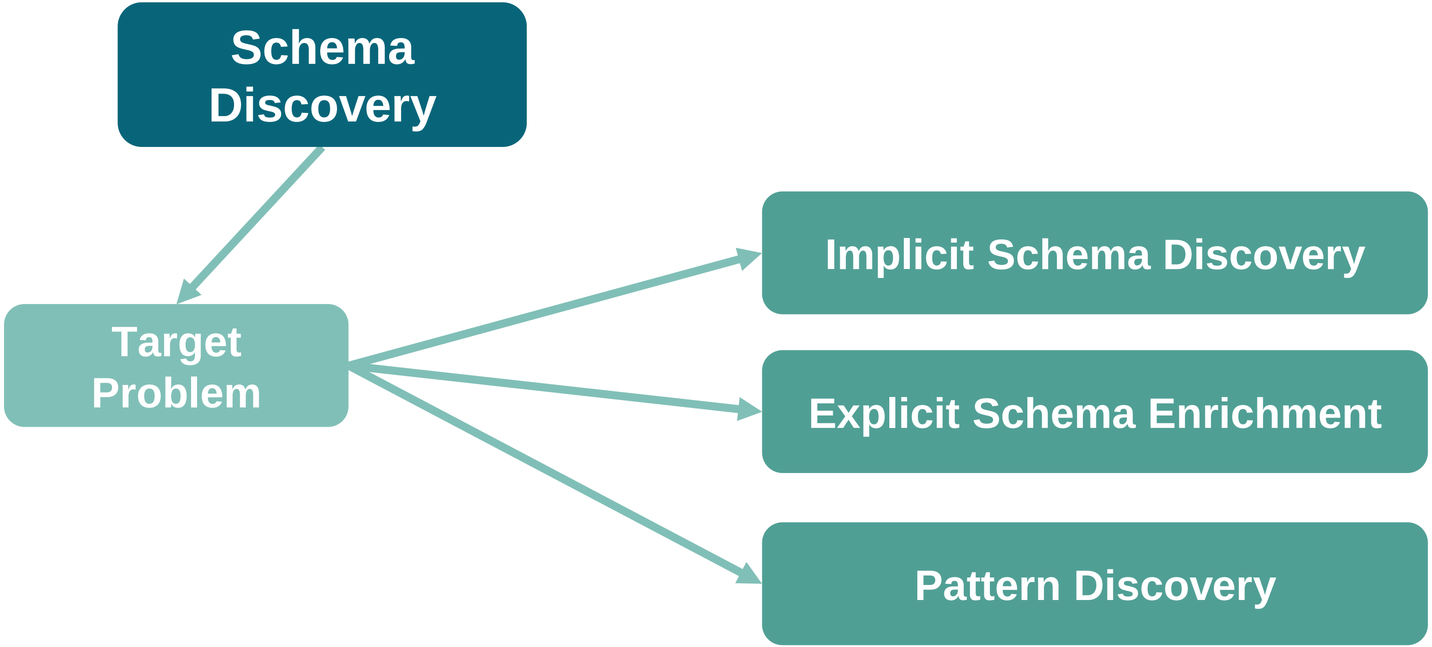
Class Accuracy



Analysis Dimensions for Schema Discovery



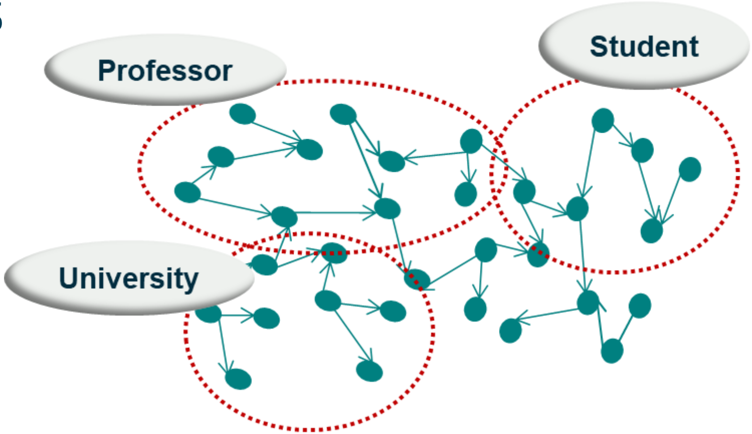
Analysis Dimensions for Schema Discovery



Implicit Schema Discovery

Implicit Schema Discovery

- Inferring the schema of a dataset from its instances
  - Classes, properties, relationships
  - Path-based summary



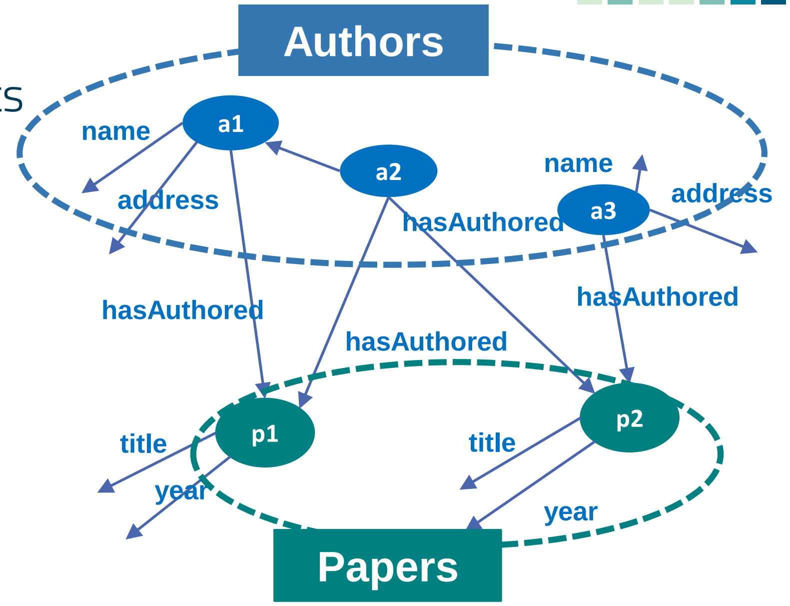
Implicit Schema Discovery

- Two alternative approaches
  - Grouping the instances of the dataset
  - Grouping the paths in the dataset

Implicit Schema Discovery by Grouping Instances

- The classes of the schema are defined as clusters of similar instances
  - Instances having similar property sets

- Underlying techniques
  - Clustering algorithms
  - Formal Concept Analysis
  - Indexing
- Most of the approaches deal with RDF datasets



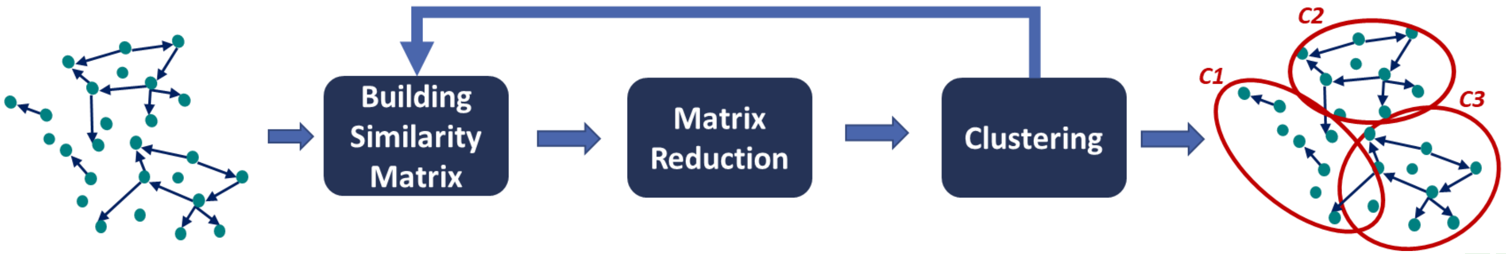
Implicit Schema Discovery Approaches Based on Instance Grouping

Approach	Underlying Technique	Input Data Graph	Output
StaTIX *[Lutov et al. IEEE Big Data 2018]*	Louvain Hierarchical Clustering	RDF Graphs	Types
HC *[Christodoulou et al. TLDKS 2015]*	Hierarchical Clustering	RD Graphs	Types, Hierarchical Link, Semantic Links
DiscoPG *[Bonifati et al. VLDB 2022]*	Hierachical Clustering / Gaussian mixture Models	Property Graphs	Graph Schema
SDA *[Menouer & Kedad TLDKS 2016]*	Density-Based Clustering	RDF Graphs	Types, Hierarchical Link, Semantic Links
SC-DBScan *[Bouhamoum et al. ESWC 2021]*	Density-Based Clustering	RDF Graphs	Types
HInT [*Kardoulakis et al. SSDBM 2021*]	Locality Sensitive Hashing	RDF Graphs	Types
FCD *[Kirchberg et al. FCA 2012]*	Formal Concept Analysis	RDF Graphs	Lattice of concepts (types)

StaTIX – Statistical Type Inference

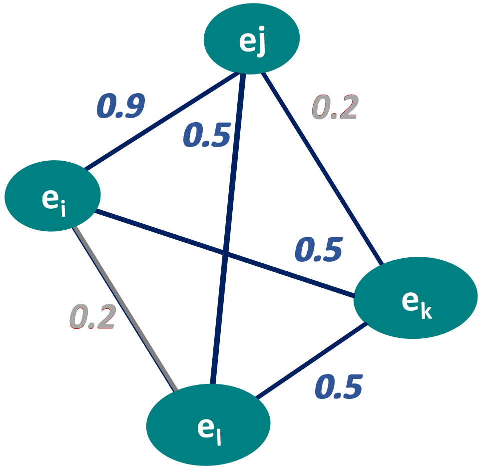
[Lutov et al. Big Data 2018]

- Input: RDF data graph
- Output: a set of overlapping types for the instances
- Using an enhanced hierarchical clustering algorithm



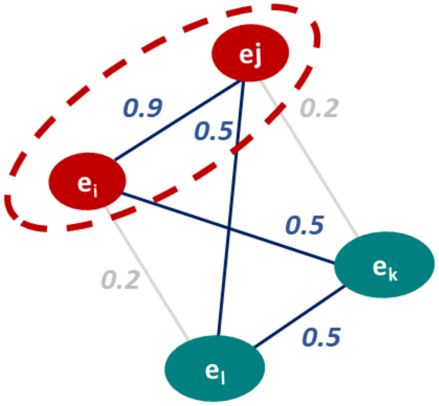
StaTIX Type Inference Principle

- Similarity Matrix
  - Property vectors of weighted properties
    - For each p_i, $w_i = 1/\sqrt{freq_i}$
  - Cosine similarity
- Matrix Reduction
  - Identifying insignificant links
    - among the ones having insignificant weights
  - Up to a maximal number of reducible links for each node



StaTIX Type Inference Principle

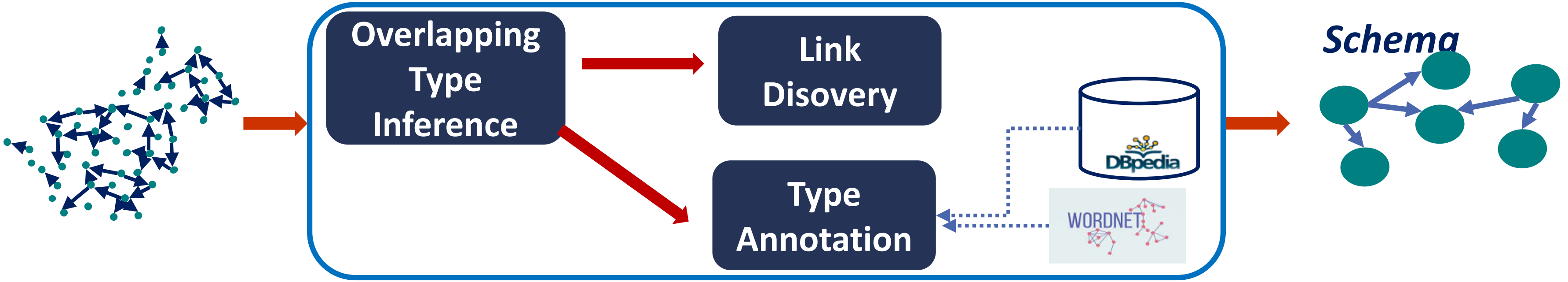
- Louvain clustering algorithm
  - Hierarchical, extended for overlap detection
  - Iterative optimization of the modularity gain $\Delta Q i, j$



Merge (ei, ej) if it maximizes ΔQij

SDA - Schema Discovery for RDF Datasets *[Kellou-Menouer & Kedad ER 2016]*

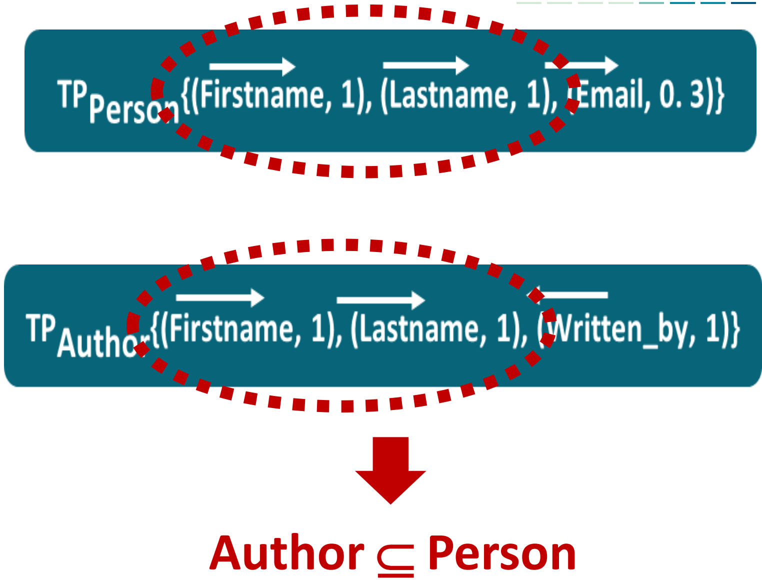
- Input: RDF data graph
- Output: Overlapping types, Hierarchical and semantic links



Type and Link Inference Principles

- Density based clustering (DBScan)
  - Entities described by their set of incoming/outgoing properties
  - Jaccard similarity
  - Probabilistic type profiles

- Overlapping types
  - Analysis of the shared properties between type profiles



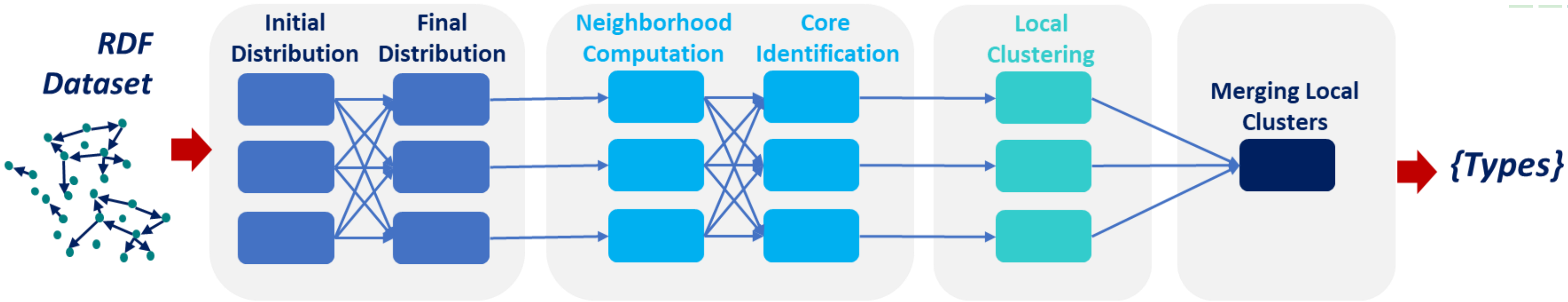
Type and Link Inference Principles

- Semantic links
  - Analysis of incoming/outgoing properties in type profiles

- Hiearchical links (*rdfs:subClassOf*)
  - Hierarchical clustering over the type profiles

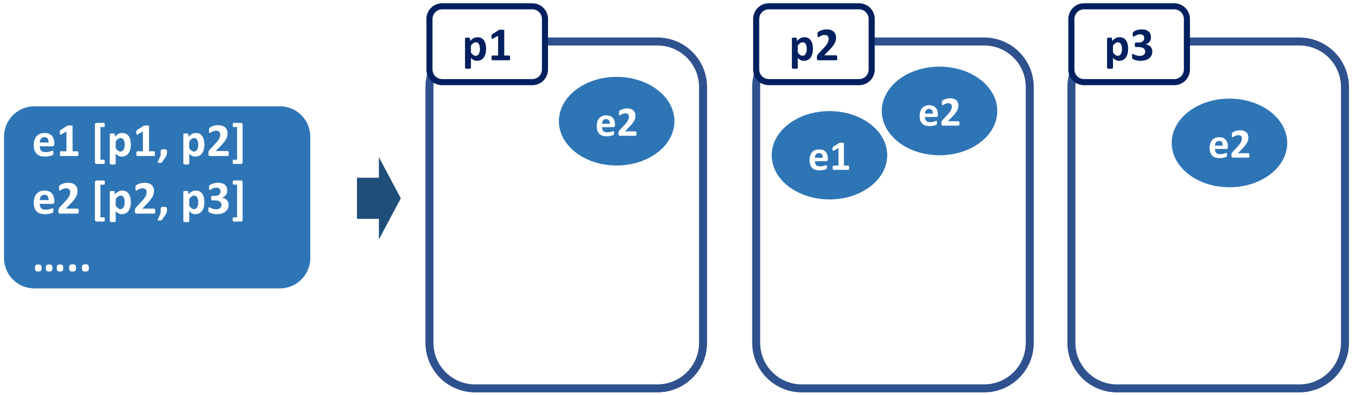
SC-DBScan: Scalable Density Based Schema Discovery *[Bouhamoum et al. ESCW 2021]*

- Distributed density-based clustering algorithm, implemented on Spark



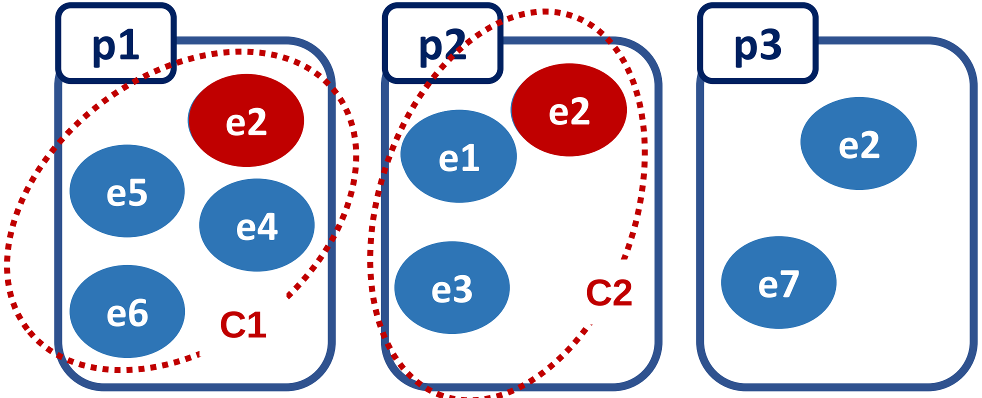
SC-DBSCAN Type Discovery Principle

- Entity Distribution:
  - A data chunk is created for each property p_i and contains entities described by p_i



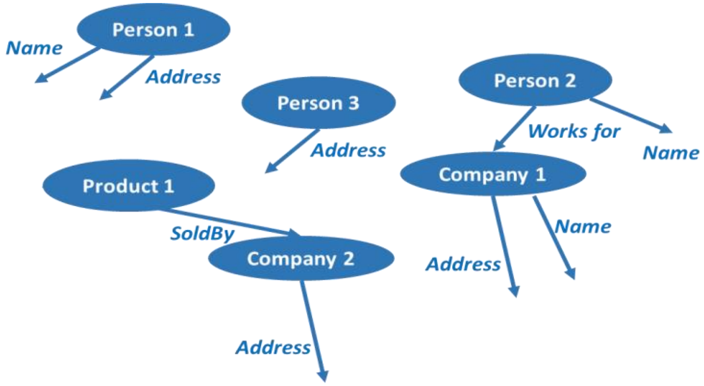
SC-DBSCAN Type Discovery Principle

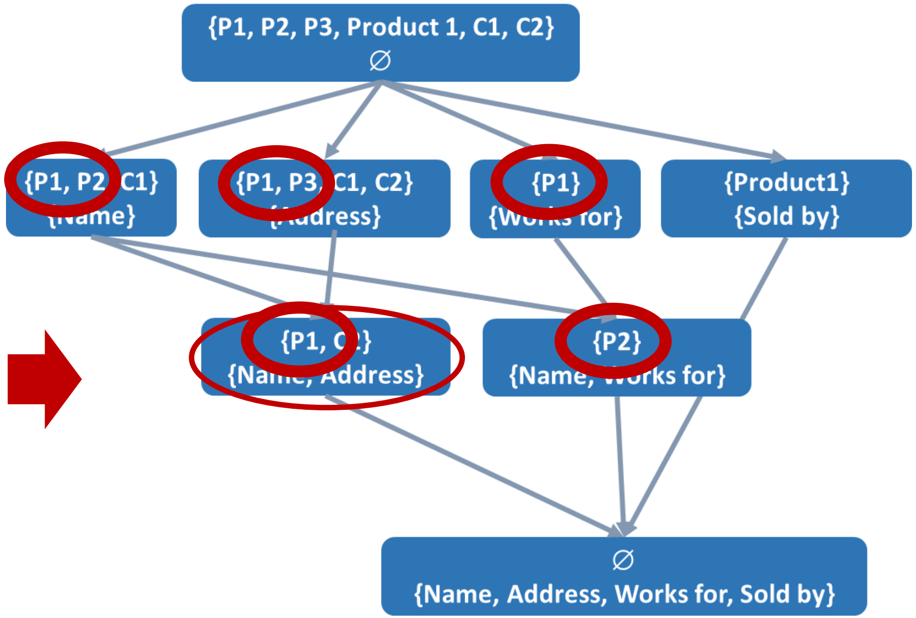
- Local clustering on each computing node using DBScan
- Merging local clusters if they share a core entity



FCD – Formal Concept Discovery in Semantic Web Data *[Kirchberg et al. FCA 2012]*

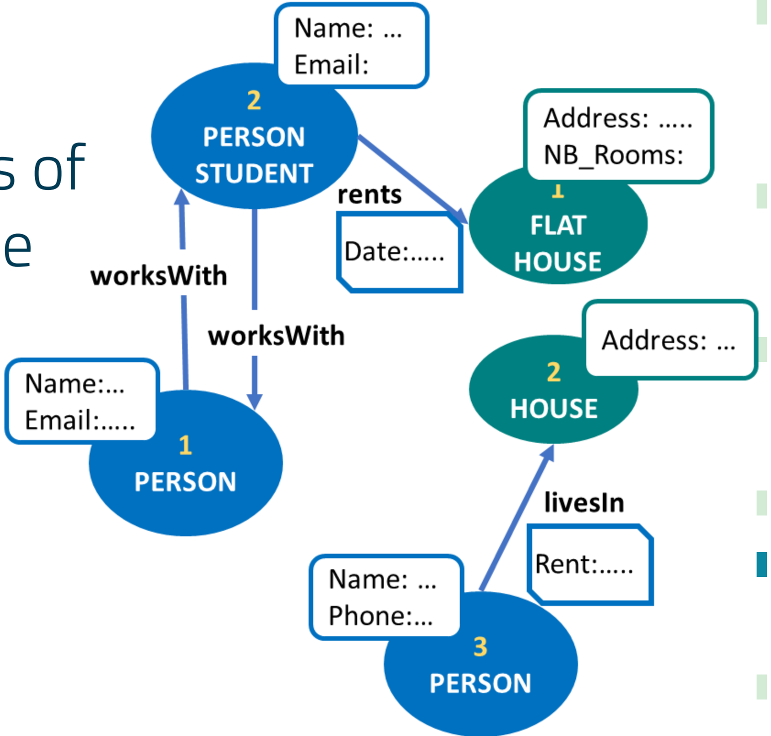
- Input: An RDF data Graph
- Output: A lattice of concepts
- Using Formal Concept Analysis





Disco PG – Property Graph Schema Discovery *[Bonifati et al. VLDB 2022]*

- Discovery principle: Computing the subtypes of a set of nodes having the same label



Disco PG – Property Graph Schema Discovery

- Compute the subtypes of a set of node C labelled L
    - Hierarchical clustering
        - Each cluster corresponds to a subtype
        - Nodes in a cluster are characterized by a unique combination of labels and properties
    - Dice similarity

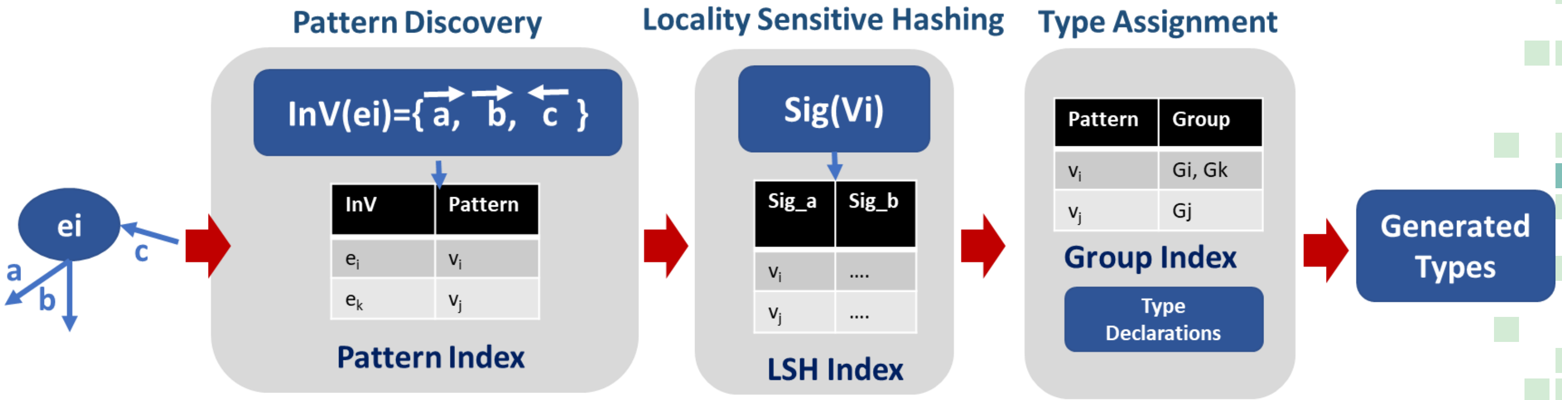
HInT – Hybrid and Incremental Schema Discovery *[Kardoulakis et al. SSDBM 2021]*

- Input: RDF data graph
- Output: a set of types

- Discovery principle : processing instances independently using Locality-Sensitive Hashing
- No pairwise comparison required

HInt – Hybrid and Incremental Schema Discovery

- Locality Sensitive Hashing: Two similar instances have a high probability of having the same signature



Implicit Schema Discovery

- Two alternative approaches
  - Grouping the instances of the dataset
  - Grouping the paths in the dataset

Implicit Schema Discovery by Grouping Paths

- Providing a representation of the data graph where identical paths are grouped

- Underlying techniques
  - Bisimulation
  - Path merging
  - Clustering algorithms

- RDF or OEM Data graphs

Implicit Schema Discovery Approaches Based on Path Grouping

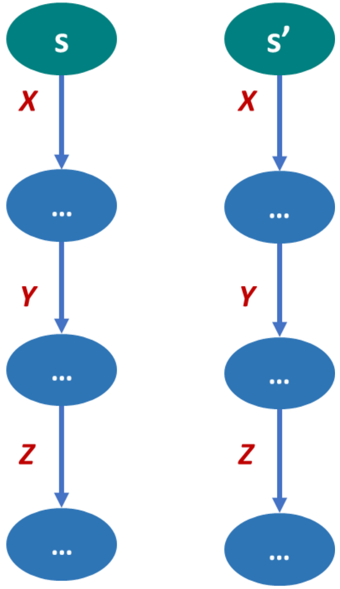
Approach	Underlying Technique	Input Data Graph	Output
Bisimulation of RDF Graphs *[Schatzle et al. SWIM13]*	Bisimulation	RDF Graphs	Path Plans
Dataguides *[Goldman et al. VLDB 1997]*	Path merging	Semi-structured data (OEM)	Path Plans
Approximate Dataguides *[Wang et al. EDBT 2000]*	Clustering (COBWEB)	Semi-structured data (OEM)	Path Plans / Types

Bisimulation of RDF Graphs

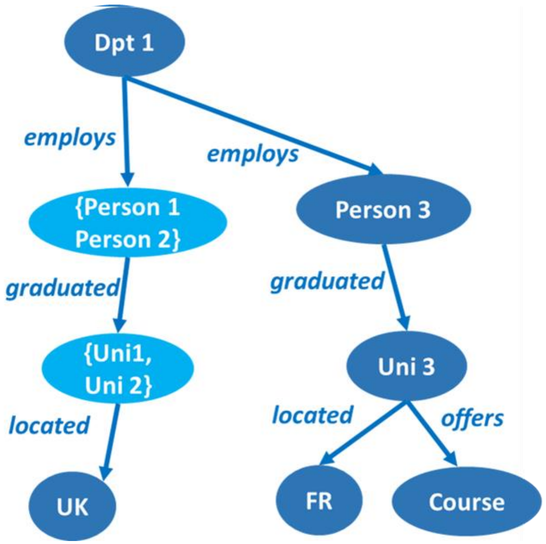
[Schatzle et al. SWIM 2013]

- Input: an RDF graph G
- Output: a bisimulation reduction of G

- Building a bisimulation partition
  - Grouping nodes s and s' if for each path starting from s, there is a path starting from s' with the same lenght and same sequence of predicates



Building a Bisimulation Partition



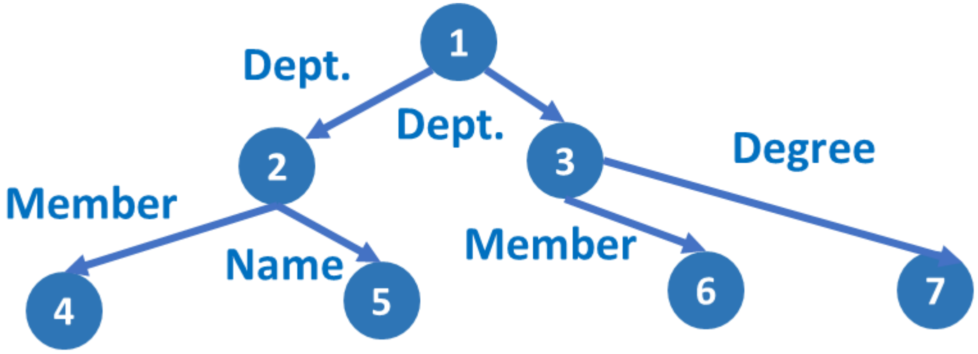
Dataguides *[Goldman et al. VLDB 1997]*

- Input: Semi-structured data described in OEM
- Output: Path Plans

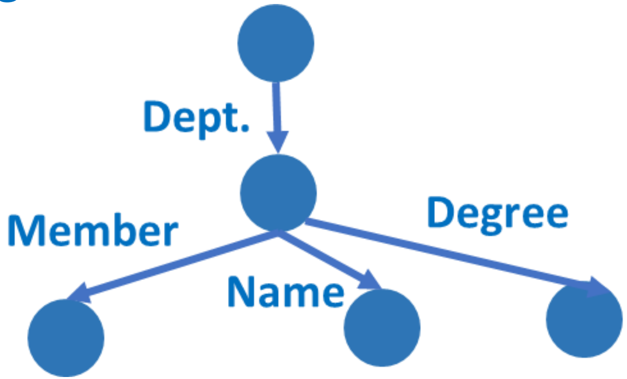
- A DataGuide *D* for an OEM graph *G* is a graph such that:
  - Every label path of G has exactly one data path instance in *D*
  - Every label path of *D* is a label path of G

Example

Initial Graph G



Exact Dataguide of G



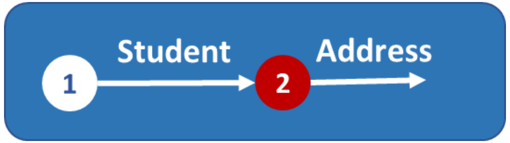
Approximate Dataguides *[Wang et al. EDBT 2000]*

- Input: Semi-structured data described in OEM
- Output: Path Plans

- Nodes in the input graph are grouped according to the similarity of their incoming/outgoing edges
  - COBWEB clustering algorithm

Example

More compact but less accurate :



Comparison of Implicit Schema Discovery Approaches

Comparing Path-Based Approaches

Approach	Result Size wrt. Initial Graph	Scalability	Stability	Incrementality
Bisimilation of RDF Graphs *[Schatzle et al. SWIM 2013]*	Smaller	Scalable (Map/Reduce)	Stable	-
Dataguides *[Goldman et al. VLDB 1997]*	May be larger	-	Stable	-
Approximate Dataguides *[Wang et al. EDBT 2000]*	Smaller that a dataguide	-	Not stable (COBWEB)	Incremental

- Query/Indexing Oriented
- Not always accurate, may be larger than the initial graph

Comparing Instance-Based Approaches

Approach	Scalability	Incrementality	Stability	Hybrid	Multiple Typing	Type Labels
StaTIX	-	-	Yes	-	Yes	-
HC	-	-	Yes	-	-	Yes
DiscoPG	-	Yes	Yes	-	Yes	-
SDA	-	Yes (typing new entities)	Yes	-	Yes	Yes
SC-DBScan	Yes	Yes	Yes	-	-	-
HInT	Yes	Natively Incremental	Yes	Yes	Yes	-
FCA	-	-	Yes	-	Yes (?)	

Comparing Instance-Based Approaches

- Clustering-based approaches
  - Require the computation of a similarity matrix and/or input parameters
  - Clusters with arbitrary shapes are more suited to very heterogeneous datasets

- Formal Concept Analysis
  - Concepts vs. Types
  - The generated lattice can be very large

Open Issues

- Most of the approaches generate Types/Classes but not links

- Annotation of the resulting types is not always supported

Open Issues

- Most of the approaches do not make use of schema related declarations if provided

- Dealing with online remote sources and coping with access restrictions has not been addressed yet

THANKS!

Any questions?

You can find us at

https://users.ics.forth.gr/~kondylak/iswc_2022_tutorial/