Tutorial on Semantic Schema Discovery: principles, methods and future research directions
Part 1

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

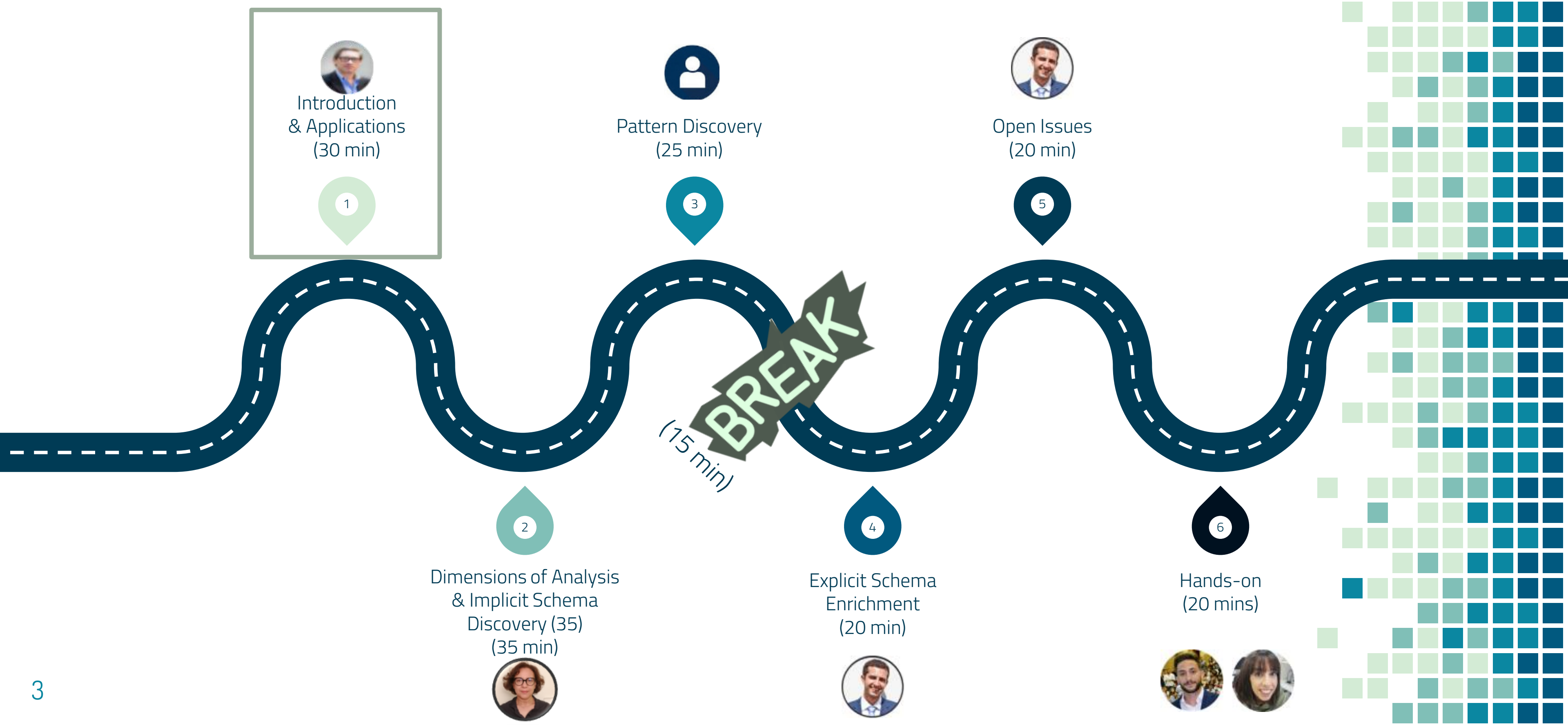
Georgia Troullinou
FORTH-ICS

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

Dimitris Plexousakis
FORTH-ICS

Haridimos Kondylakis
FORTH-ICS

ROADMAP



1

Introduction
& Applications
(30 min)

2

Dimensions of Analysis
& Implicit Schema
Discovery (35)
(35 min)

3

Pattern Discovery
(25 min)

BREAK
(15 min)

4

Explicit Schema
Enrichment
(20 min)

5

Open Issues
(20 min)

6

Hands-on
(20 mins)

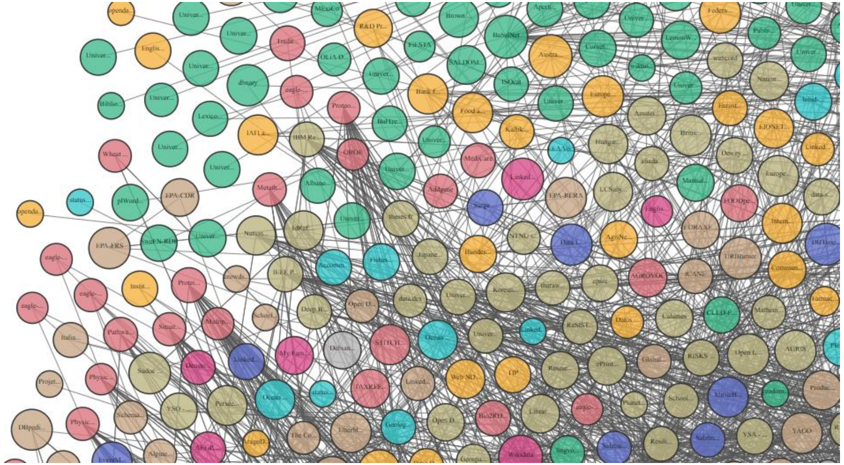
Introduction (15 minutes)

Big Data

- "*Big Data, Enormous Opportunity*" Ed Lazowska, Univ. of Washington
- The data avalanche: big data everywhere
  - Proliferation of sensors
  - Almost all information is nowadays produced in digital form
  - Dramatic reduction of cost for data storage
  - Fast-paced increases in network capacity
  - Great improvements in scalable computing infrastructures
  - Powerful models / digital twins
  - Algorithmic breakthroughs
- ...enabling a "big-data" revolution!

RDF Graph Discovery

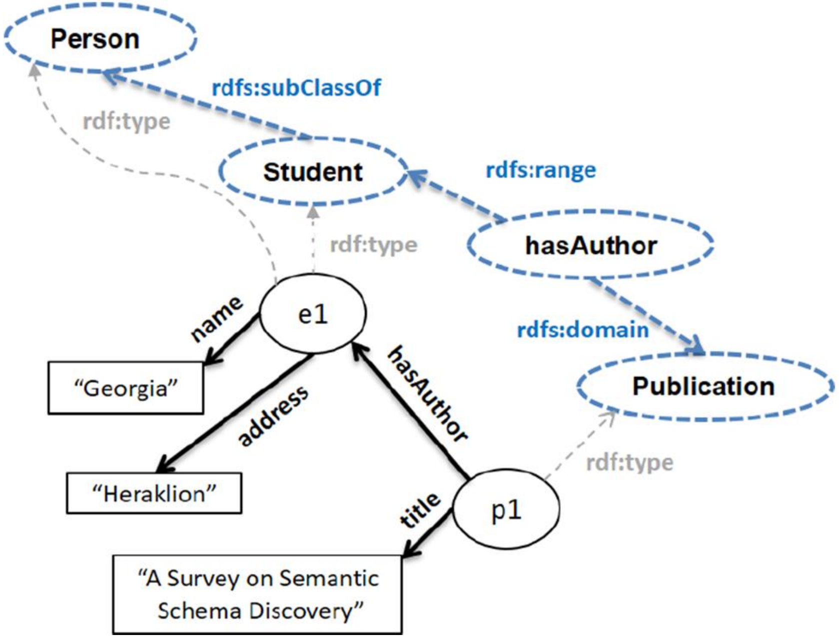
- An RDF graph can be large and complex, lack a fixed schema, include many heterogeneous values…



Proliferation of Weakly Structured Data

- Proliferation of weakly structured, irregular, incomplete and massive data sources
- Particularly the case of semantic web data
  - They do not follow a predefined schema
  - May include declarations on the schema
    - Incomplete Schema or Completely absent

Example



Person
rdfs:subClassOf
rdf:type
Student
rdfs:range
hasAuthor
rdf:type
e1
name
"Georgia"
address
"Heraklion"
hasAuthor
rdfs:domain
Publication
rdf:type
p1
title
"A Survey on Semantic Schema Discovery"

> “*A **schema of the data source** describes the types of data and the links between them provides a characterization of the content of this data source*

Scope

- We are interested in schema information retrieval / discovery approaches for data sources for which this schema is missing or partially defined
- Key research problem for data management with many approaches, algorithms and methods developed to cope with it.

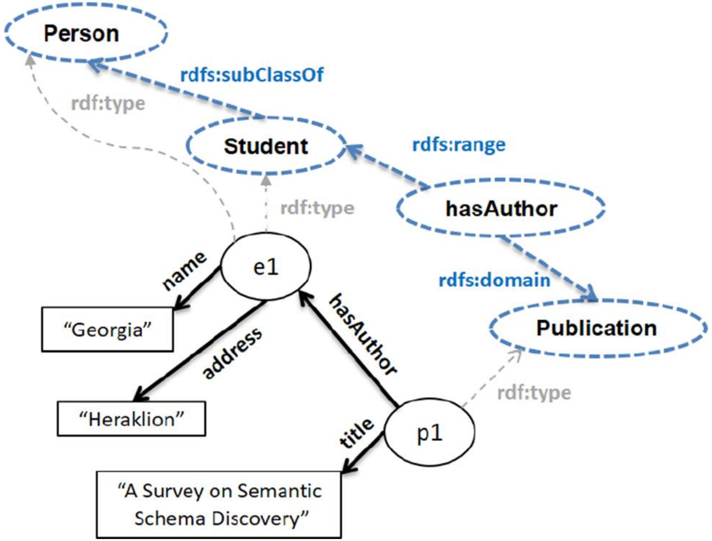
Target

- Improve understanding of this field
- Help students, researchers or practitioners identify the schema discovery algorithm, method or tool best suited for a specific problem
- Study, classify and compare the different schema discovery works, as well as provide a clarification of the terminology used

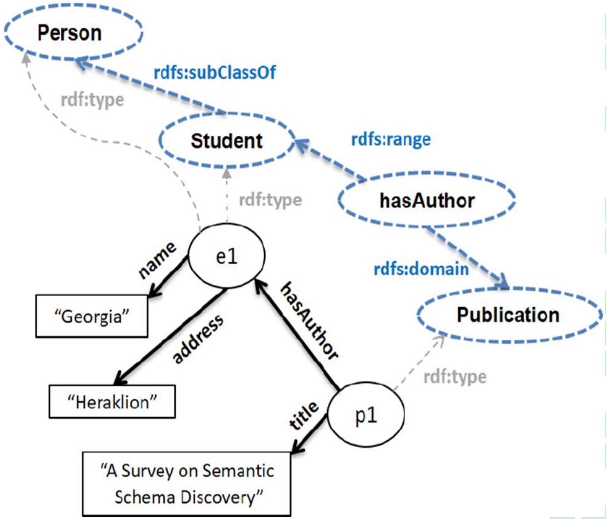
Preliminaries

The Resource Description Framework (RDF)

- RDF graph: set of triples



RDF Schema



- RDFS deductive constraints, stating connections between classes and properties

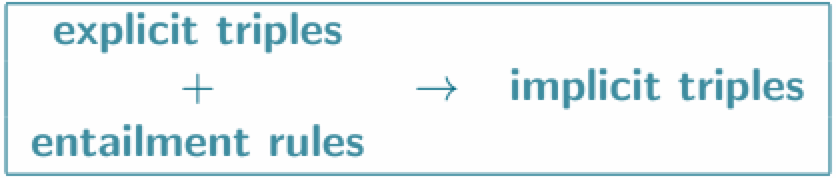
(Student,rdfs:subClassOf, Person)

(hasAuthor,rdfs:domain, Publication)

(hasAuthor,rdfs:range, Student)

Open-world assumption

- RDF data model based on the open-world assumption
- Deductive constraints lead to implicit triples: part of the graph even though not explicitly present



explicit triples
+ → implicit triples
entailment rules

- Exhaustive application of entailment leads to saturation (closure)

Applications (15 minutes)

Source Selection

- Before using a data source
  - Identifying the classes and the properties in the schema
  - Detect whether the source is likely to contain information that the user is looking

- e.g. LODatio automatically selects the relevant LOD data sources for a SPARQL query.

Query Formulation

- A schematic description of the content of a dataset is essential for formulating a query.
- The schema gives an overview of the content of a data source and the syntax of the different properties and classes

- e.g. Protégé provides schema auto-completion features to help the user when writing a query

Query Answering

- The schema of a data source is also useful for quickly determining the results for regular path queries.

- e.g. the schema could be used as a first filter
- to determine whether the dataset contains the answer to a query

Distributed Query Decomposition and Optimization

- When a query is issued over several data sources,
  - query decomposition is a key problem, as well as finding optimal execution plans

- the schema is essential for decomposing the query and sending the sub-queries only to the relevant sources

Data Indexing

- The schema of a data source could be used as an index.

- e.g. SchemEx uses schema-related information to build a three-layered index.
  - Each layer captures different types of schema information targeted at different types of queries,
  - Groups input data sources of the LOD cloud into nodes

Inference and Reasoning

- The inference rules and the semantic reasoning algorithms rely mainly on the available related schema declaration.
- They make it possible to generate new knowledge and to check the consistency of a data source.

Semantic Summarization

- Summarization works require a schema in order to summarize the contents of an RDF graph.
- The schema offers the first-level, natural way to abstract the contents of the graph.

Data integration and linking

- These tasks are often hindered by the lack of schema information on datasets.
- Tools proposed require schema-related information about the datasets to generate the appropriate links between the datasets

Data Quality Assessment

- Some works rely on the schema of a source to propose metrics to evaluate the quality of a data source, such as:
  - (i) the completeness of a dataset with respect to its schema
  - ii) the accuracy of a schema with respect to its dataset.

Data Partitioning

- Partitioning of the data aims at distributing effectively data over multiple nodes for improving query answering.
- In many cases, partitioning methods are based on an extracted schema of the dataset.

THANKS!

Any questions?

You can find us at

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