Research in Logic & Data Management

Wim Martens

University of Bayreuth

Logic Mentoring Workshop @ LICS 2020

UNIVERSITÄT BAYREUTH

Why Data Management?

(1) It is an incredibly relevant field
(2) The Logic Force is strong in Data Management
(3)

[Image removed]

(4) I chose to go into Data Management 15 years ago
and I never regretted it

Working in data management and database theory
has significantly helped me in getting a tenured position

Logic & Data Management?

$$\mathrm{FO} \equiv \mathrm{SQL}$$

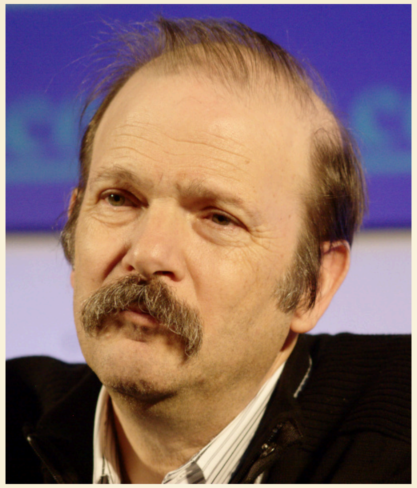
-- E.F. Codd, paraphrased

Logic & Data Management?

Many people with outstanding logic skills work in database theory

Kolaitis

Muscholl



Vardi

Grohe

Fagin

Libkin

Schweikardt

did not
find
picture

You

...and many, many more!

Logic & Data Management?

Have a look at...

...the Gems of PODS!

databasetheory.org/gems

Formal Languages & Data Management?

My own background was more from formal languages...

- But still, I felt more than welcome in PODS & ICDT

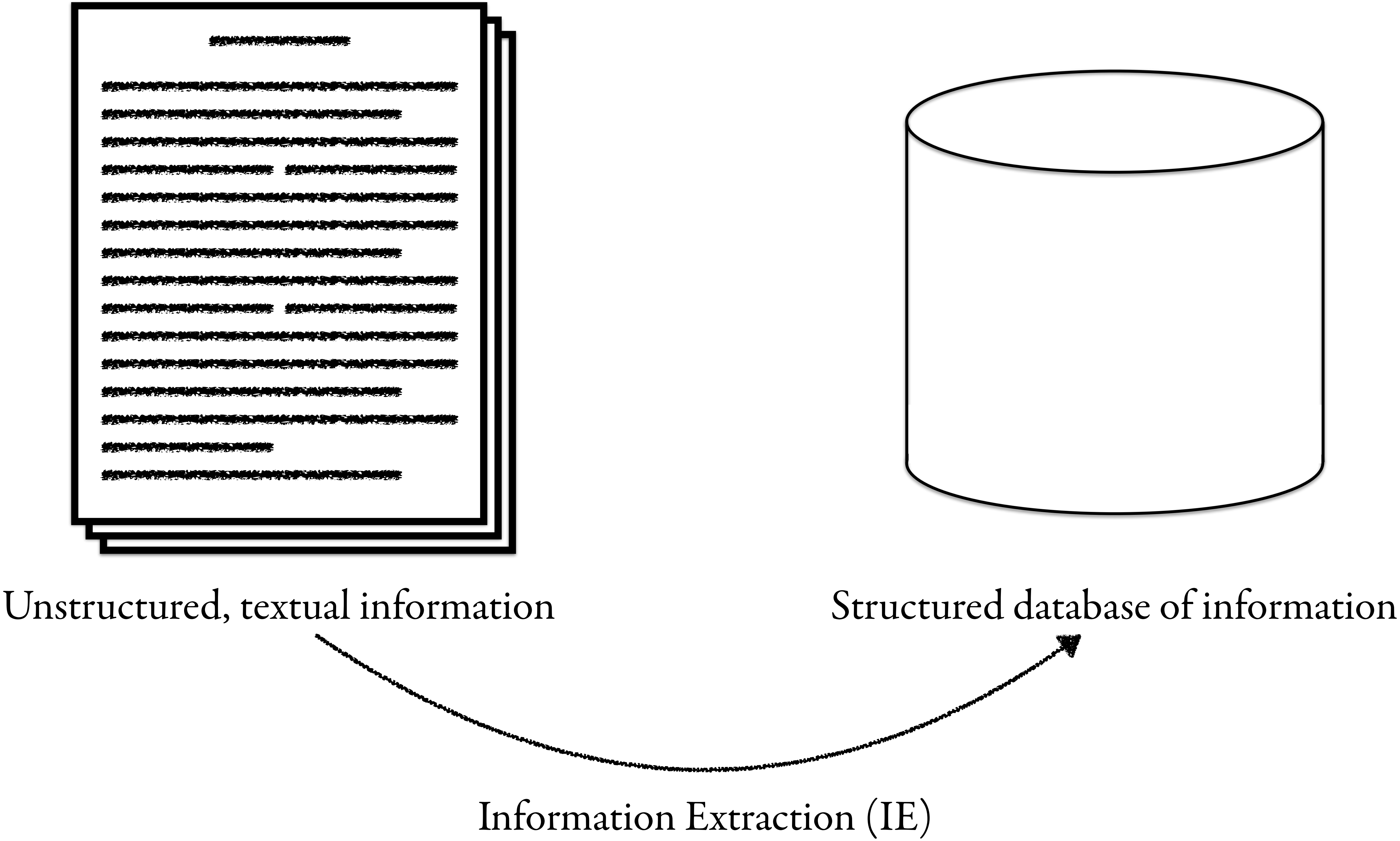
Lately, I've been doing some work in...

Information Extraction

Graph Databases

Information Extraction

General Idea



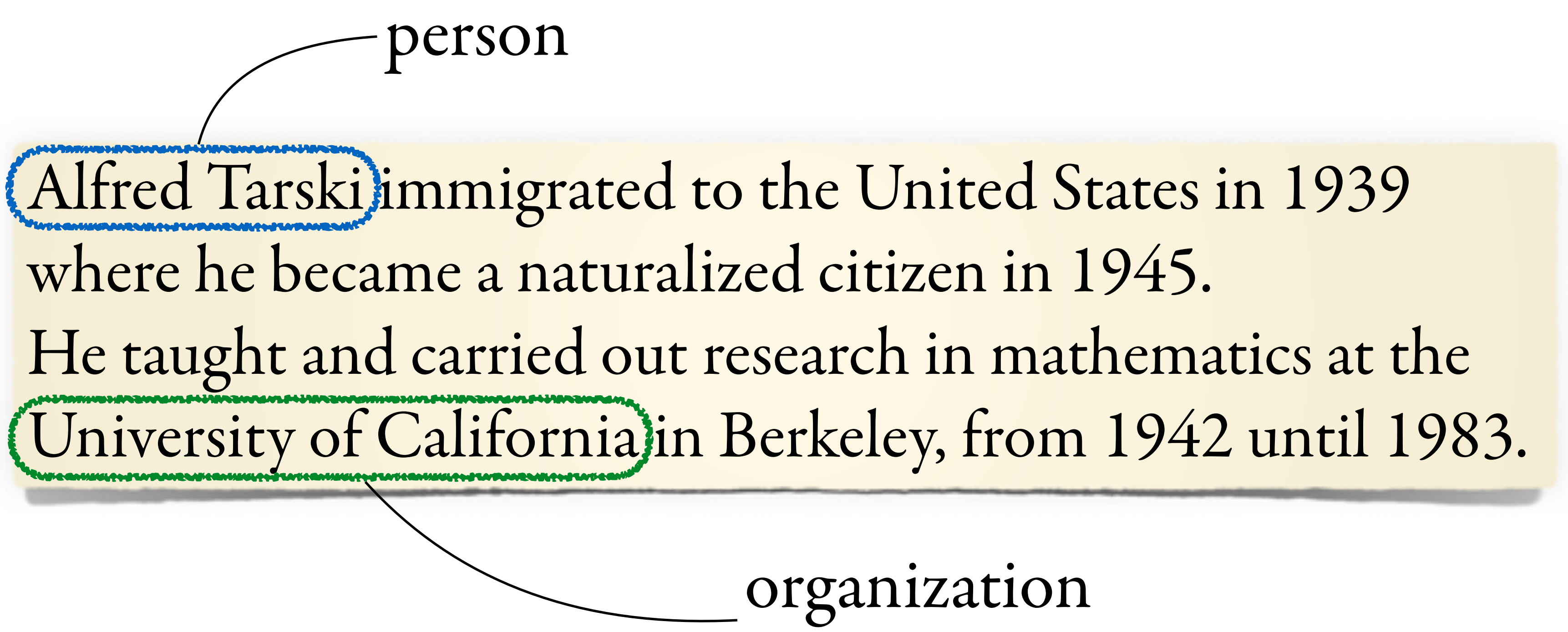
Unstructured, textual information

Structured database of information

Information Extraction (IE)

IE Tasks

[Kimelfeld, EDBTSS'19]



person

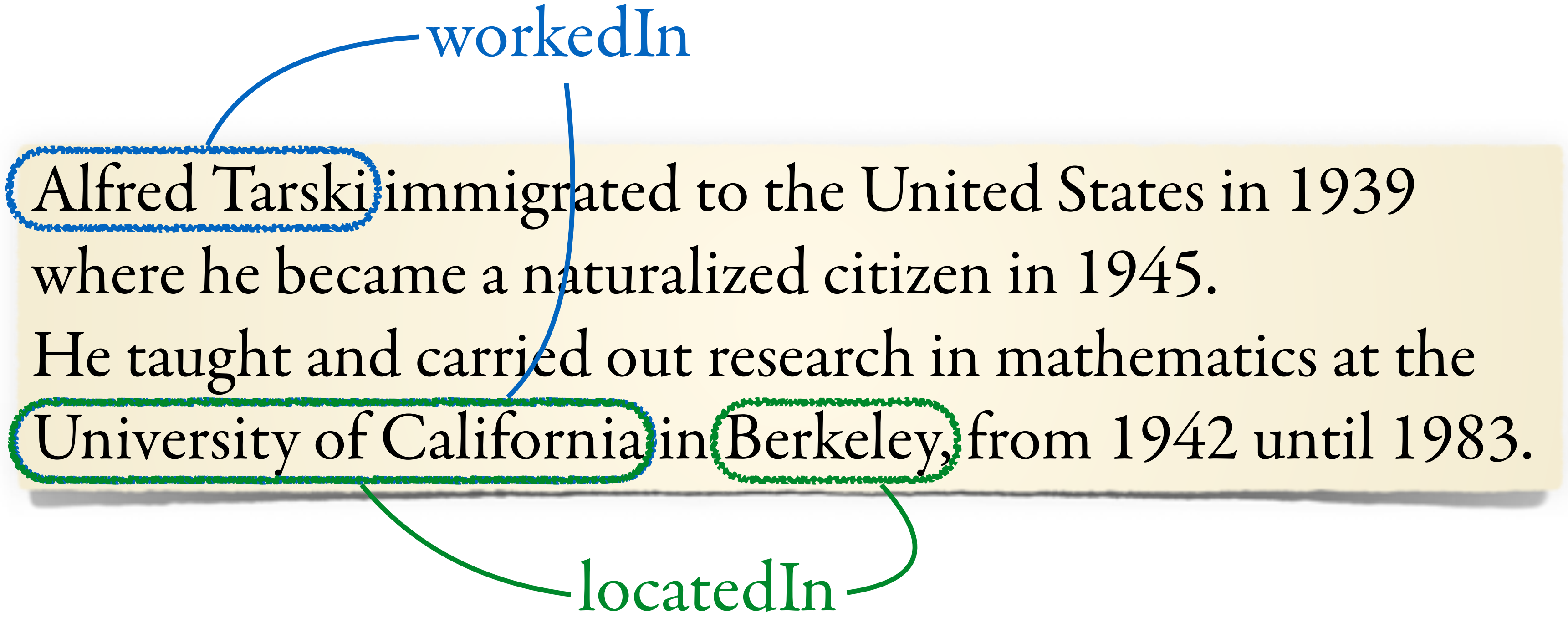
Alfred Tarski immigrated to the United States in 1939
where he became a naturalized citizen in 1945.
He taught and carried out research in mathematics at the
University of California in Berkeley, from 1942 until 1983.

organization

- Named Entity Recognition

IE Tasks

[Kimelfeld, EDBTSS'19]

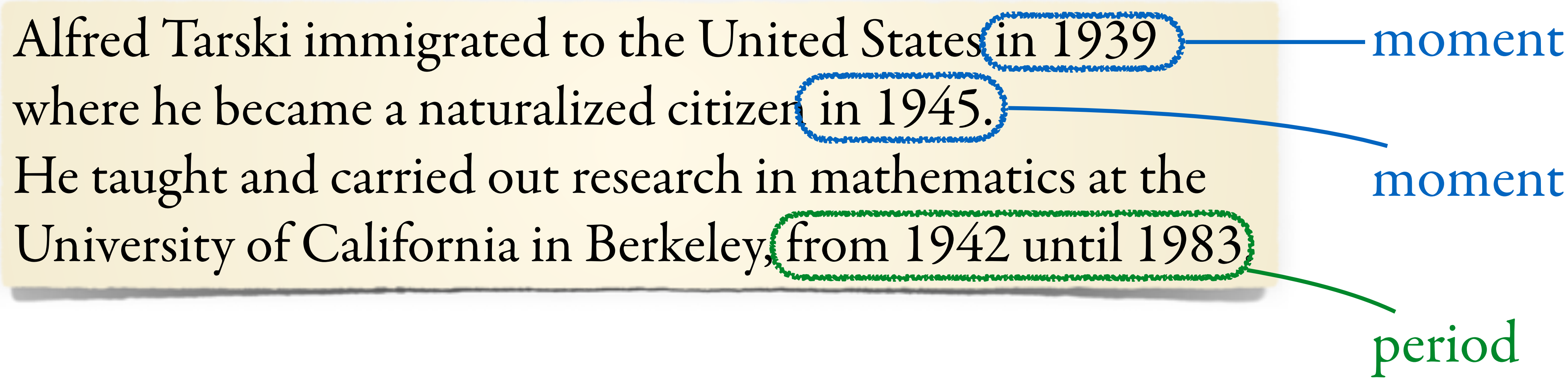


Alfred Tarski immigrated to the United States in 1939 where he became a naturalized citizen in 1945.
He taught and carried out research in mathematics at the University of California in Berkeley, from 1942 until 1983.

- Named Entity Recognition
- Relation Extraction

IE Tasks

[Kimelfeld, EDBTSS'19]

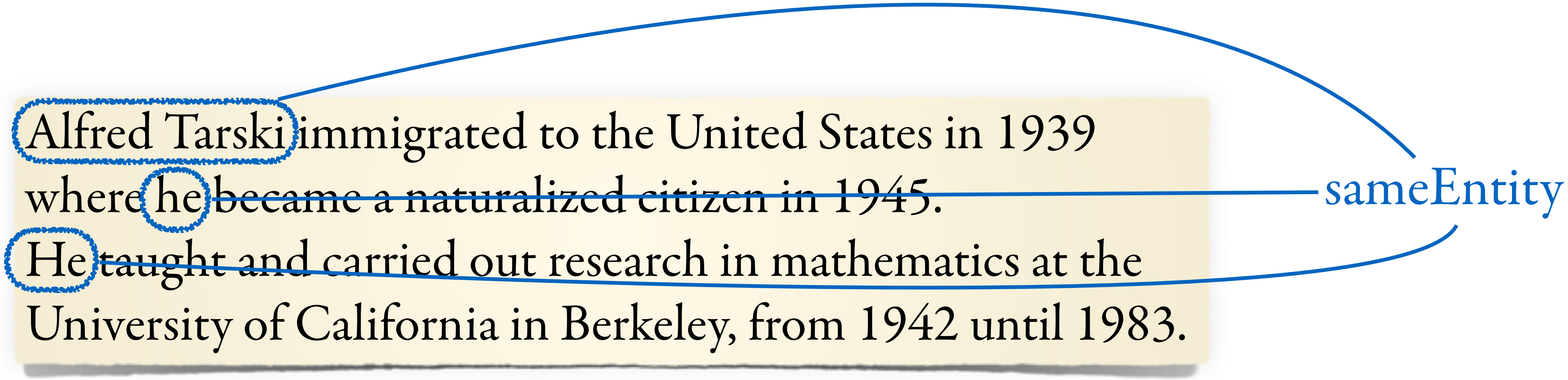


Alfred Tarski immigrated to the United States in 1939 where he became a naturalized citizen in 1945. He taught and carried out research in mathematics at the University of California in Berkeley, from 1942 until 1983

- Named Entity Recognition
- Relation Extraction
- Temporal IE

IE Tasks

[Kimelfeld, EDBTSS'19]

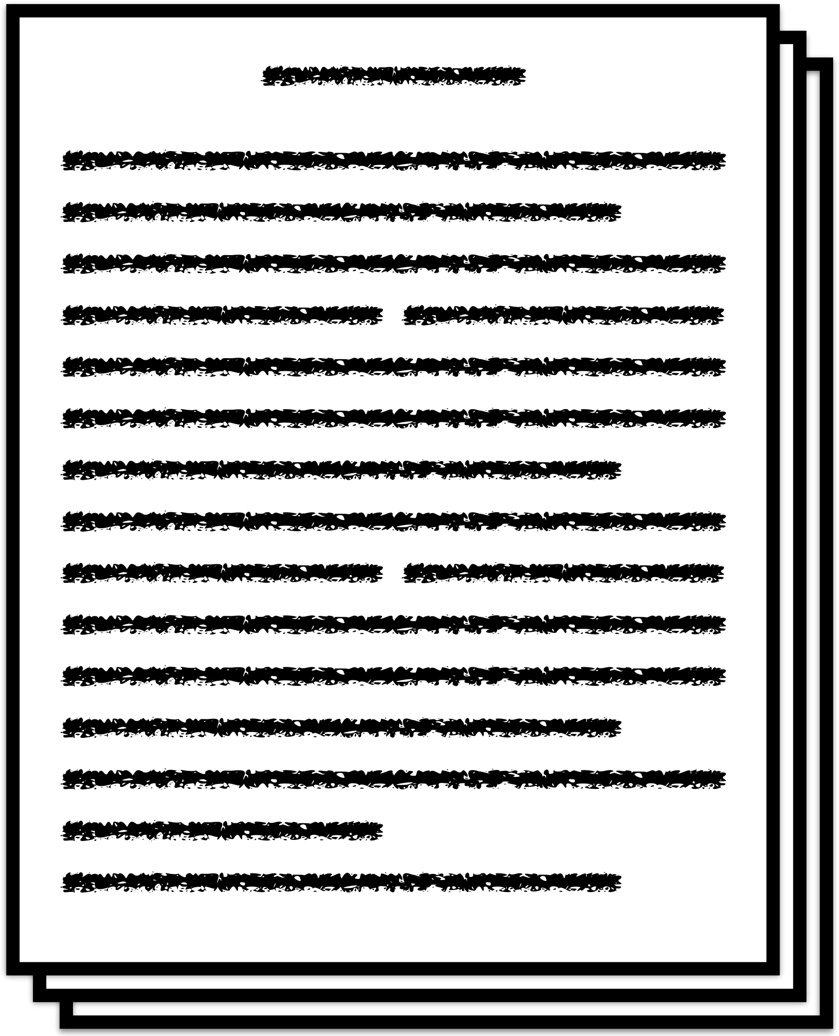


Alfred Tarski immigrated to the United States in 1939 where he became a naturalized citizen in 1945. He taught and carried out research in mathematics at the University of California in Berkeley, from 1942 until 1983.

- Named Entity Recognition
- Relation Extraction
- Temporal IE
- Coreference Resolution
- ...

Document Spanner Framework

[Fagin et al., PODS 2013]



Unstructured, textual information

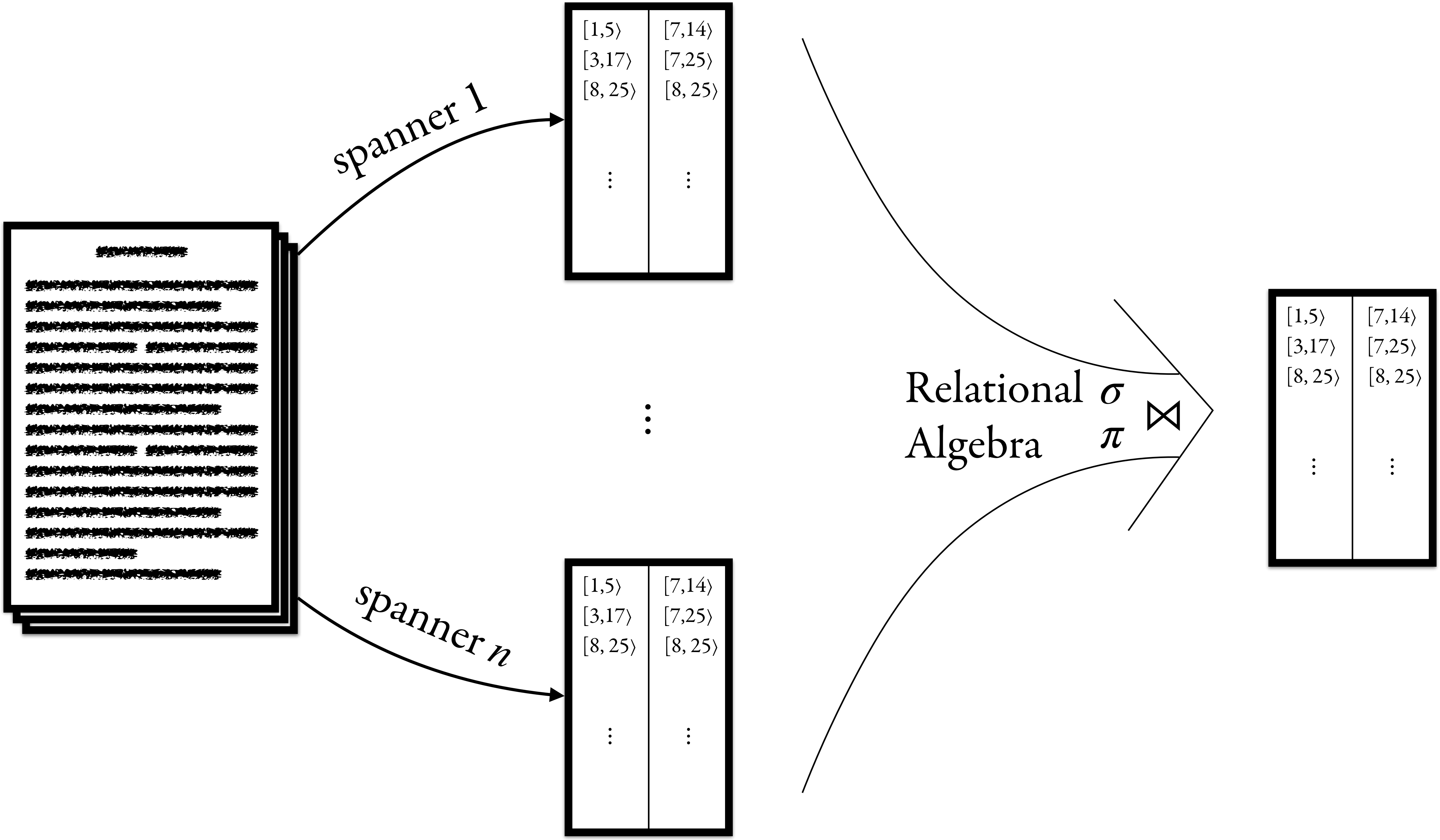
[1,5⟩	[7,14⟩
[3,17⟩	[7,25⟩
[8, 25⟩	[8, 25⟩
⋮	⋮

A relation of "intervals", i.e. start/end positions in the text

Document Spanner: automata, regular expressions, logic, datalog, ...

Document Spanner Framework

[Fagin et al., PODS 2013]



spanner 1

[1,5⟩	[7,14⟩
[3,17⟩	[7,25⟩
[8, 25⟩	[8, 25⟩
⋮	⋮

⋮

spanner n

[1,5⟩	[7,14⟩
[3,17⟩	[7,25⟩
[8, 25⟩	[8, 25⟩
⋮	⋮

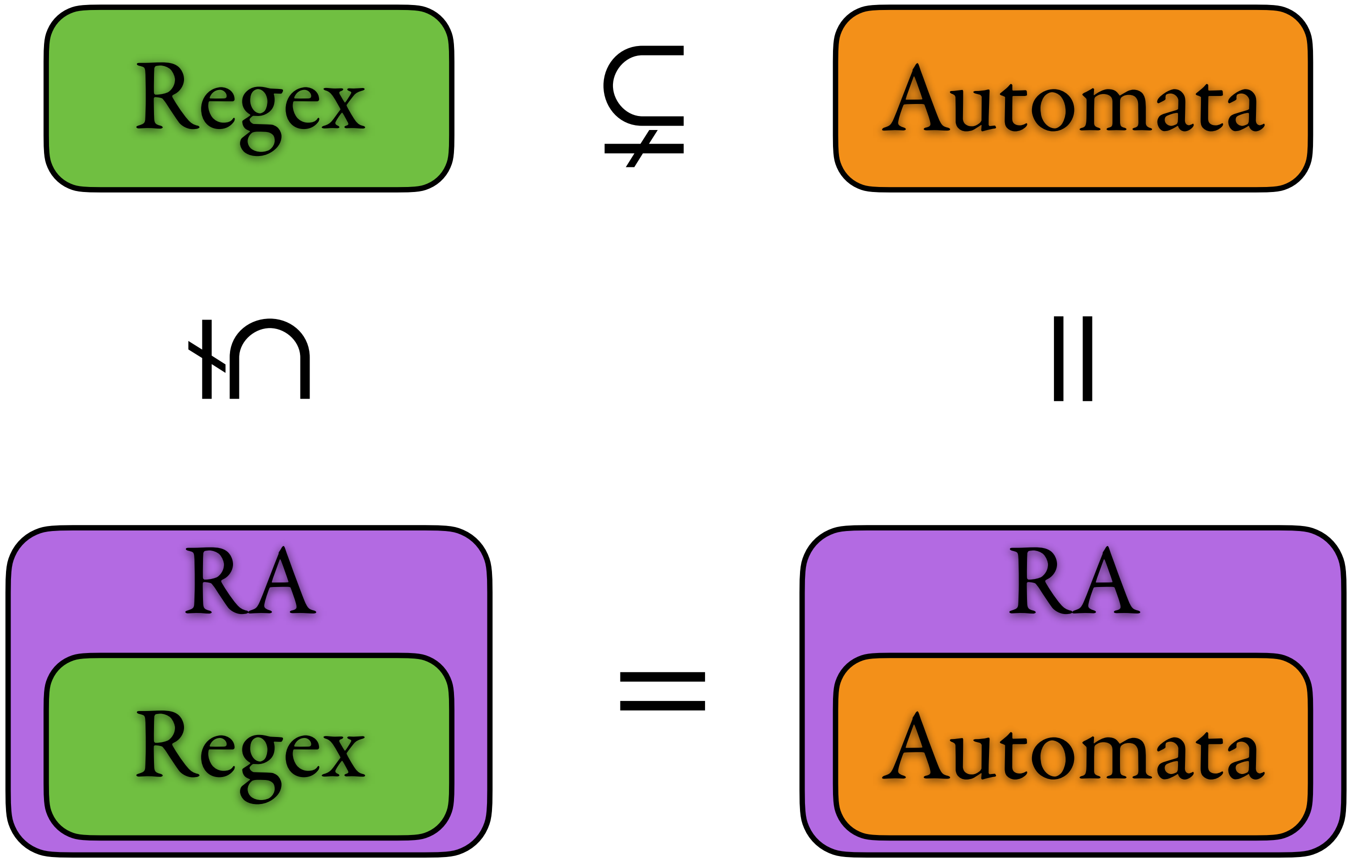
Relational Algebra σ π $\bowtie$

[1,5⟩	[7,14⟩
[3,17⟩	[7,25⟩
[8, 25⟩	[8, 25⟩
⋮	⋮

Research Questions
in
Information Extraction

Spanners: Research Questions

Expressiveness



Regex ⊊ Automata

RA(Regex) = RA(Automata)

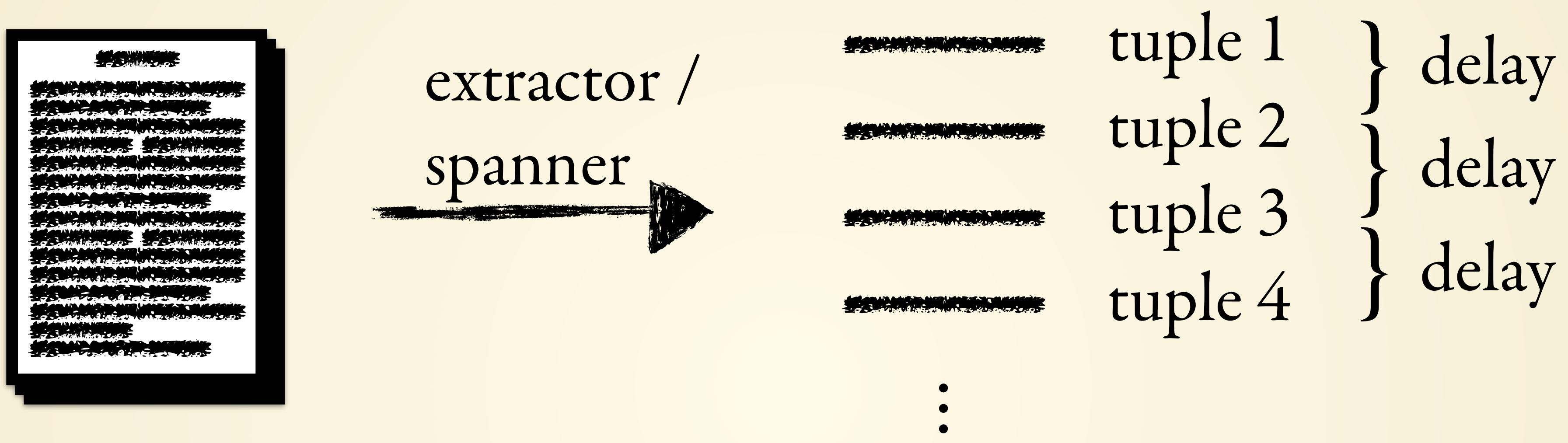
Expressiveness of Regular Spanners

⇝ [Fagin, Kimelfeld, Reiss, Vansummeren '15]

Spanners: Research Questions

Evaluation

Computing the Output of a Document Spanner



Which spanners can you evaluate using guarantees on

- time until the first answer and
- time delay between answers

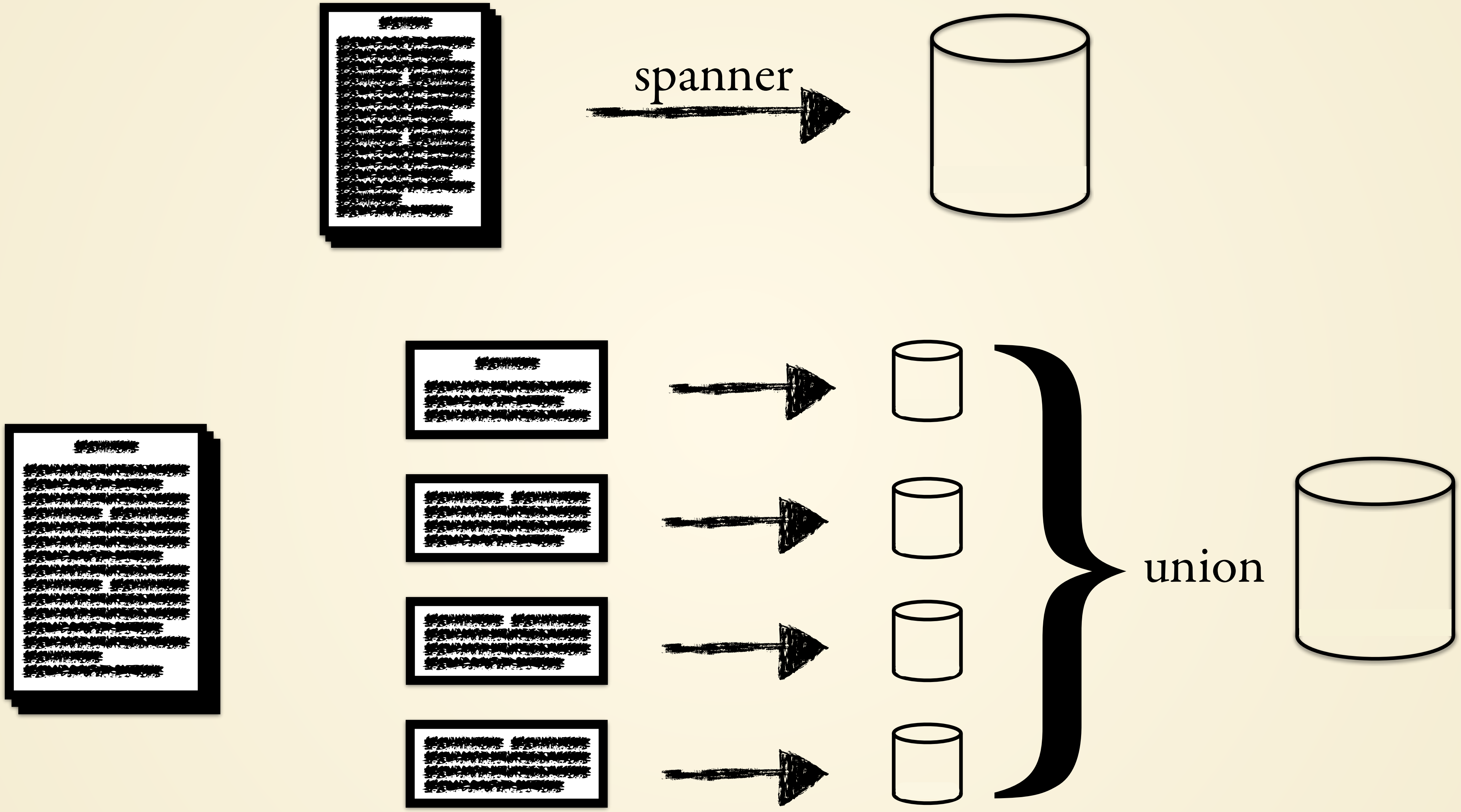
Enumeration Complexity of Document Spanners

⇝ [Arenas et al. PODS'19, Amarilli et al. ICDT'19,Florenzano et al. PODS'17]

Spanners: Research Questions

Static Analysis

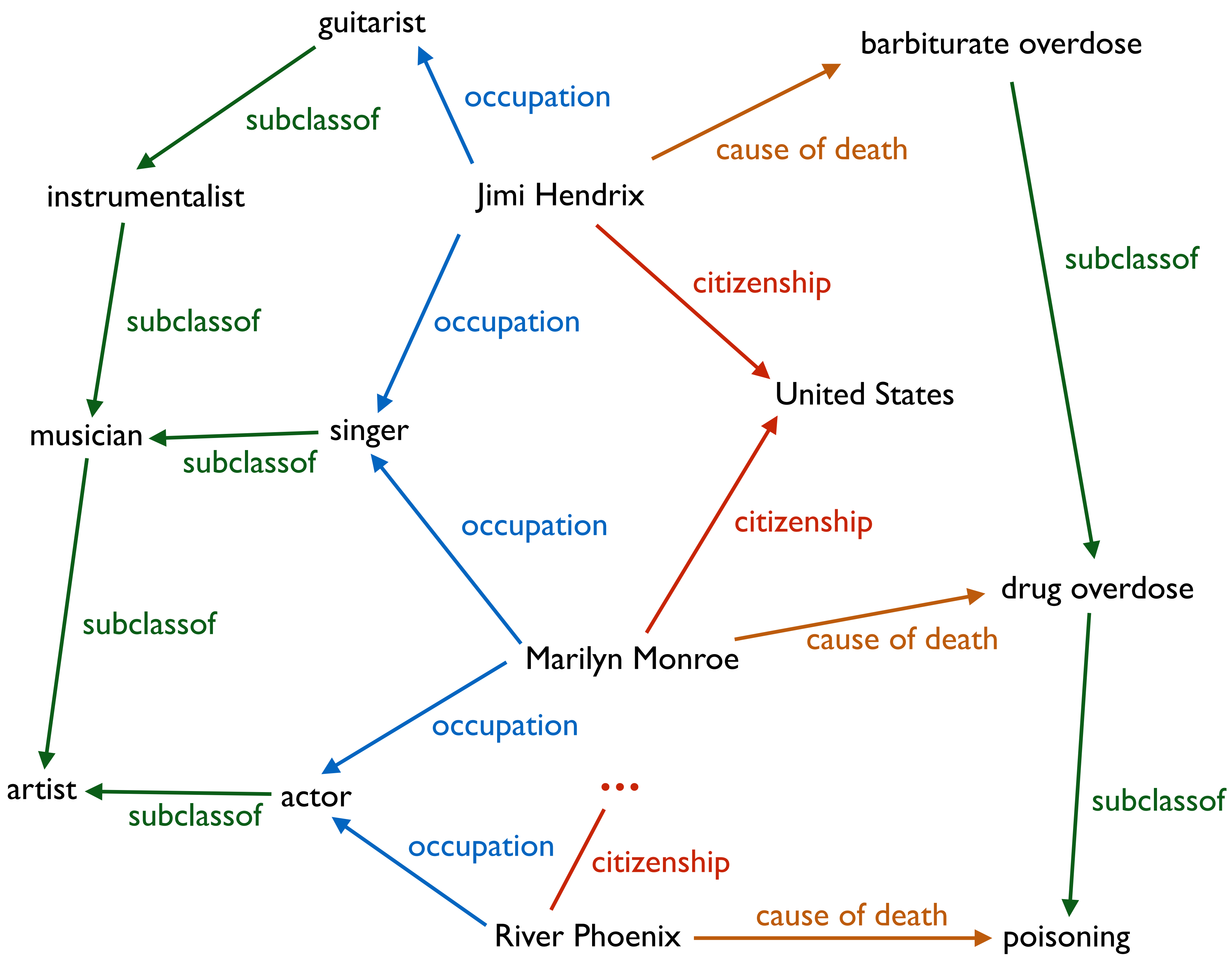
Parallelizability

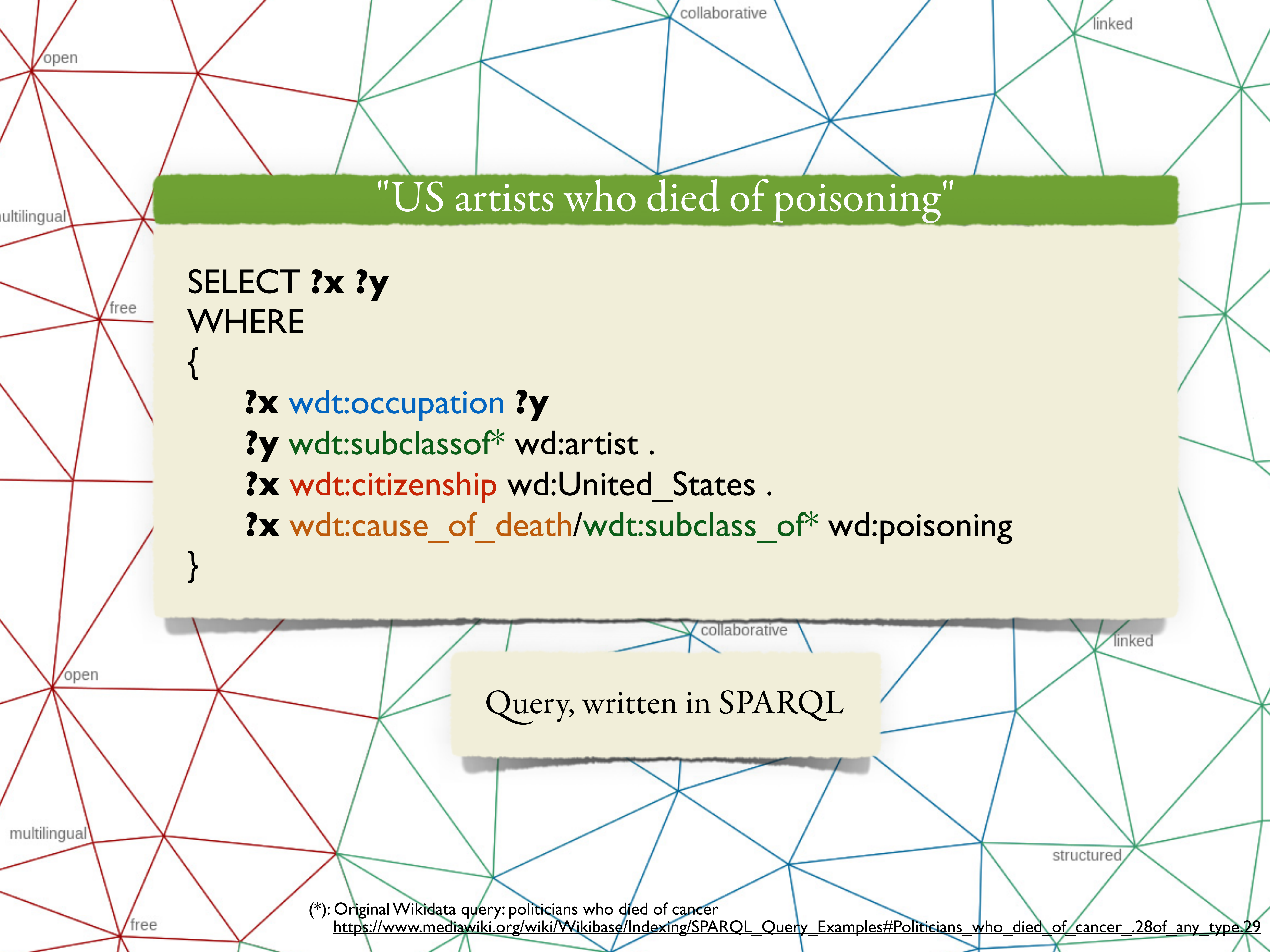


Splittability of Document Spanners ⇝ [Doleschal et al. PODS '19]

Graph Databases

What is a Graph Database?





"US artists who died of poisoning"

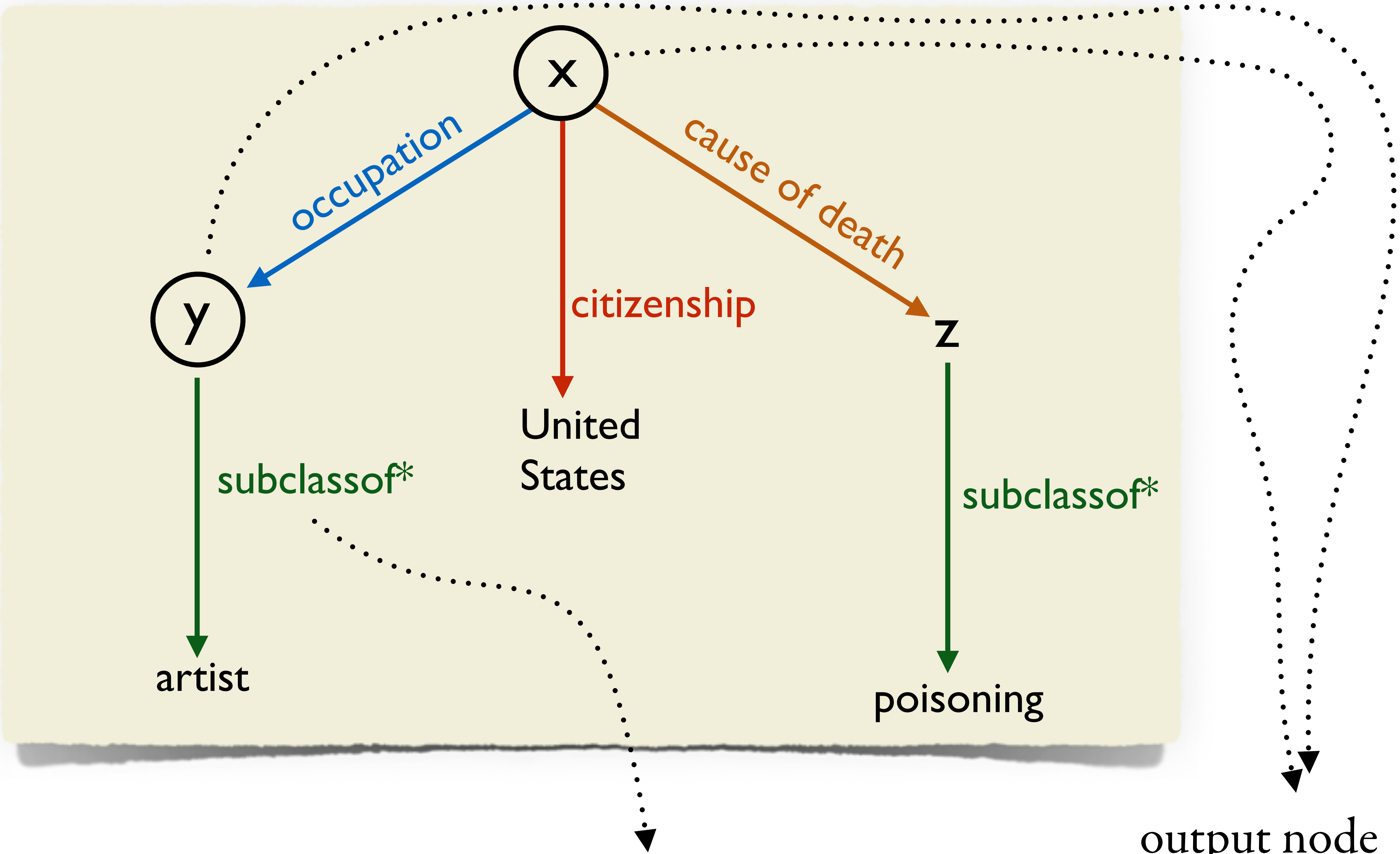
```
SELECT ?x ?y
WHERE
{
      ?x wdt:occupation ?y
      ?y wdt:subclassof* wd:artist .
      ?x wdt:citizenship wd:United_States .
      ?x wdt:cause_of_death/wdt:subclass_of* wd:poisoning
}
```

Query, written in SPARQL

(*): Original Wikidata query: politicians who died of cancer
https://www.mediawiki.org/wiki/Wikibase/Indexing/SPARQL_Query_Examples#Politicians_who_died_of_cancer_.28of_any_type.29

The Query, Visualized

"US artists who died of poisoning"

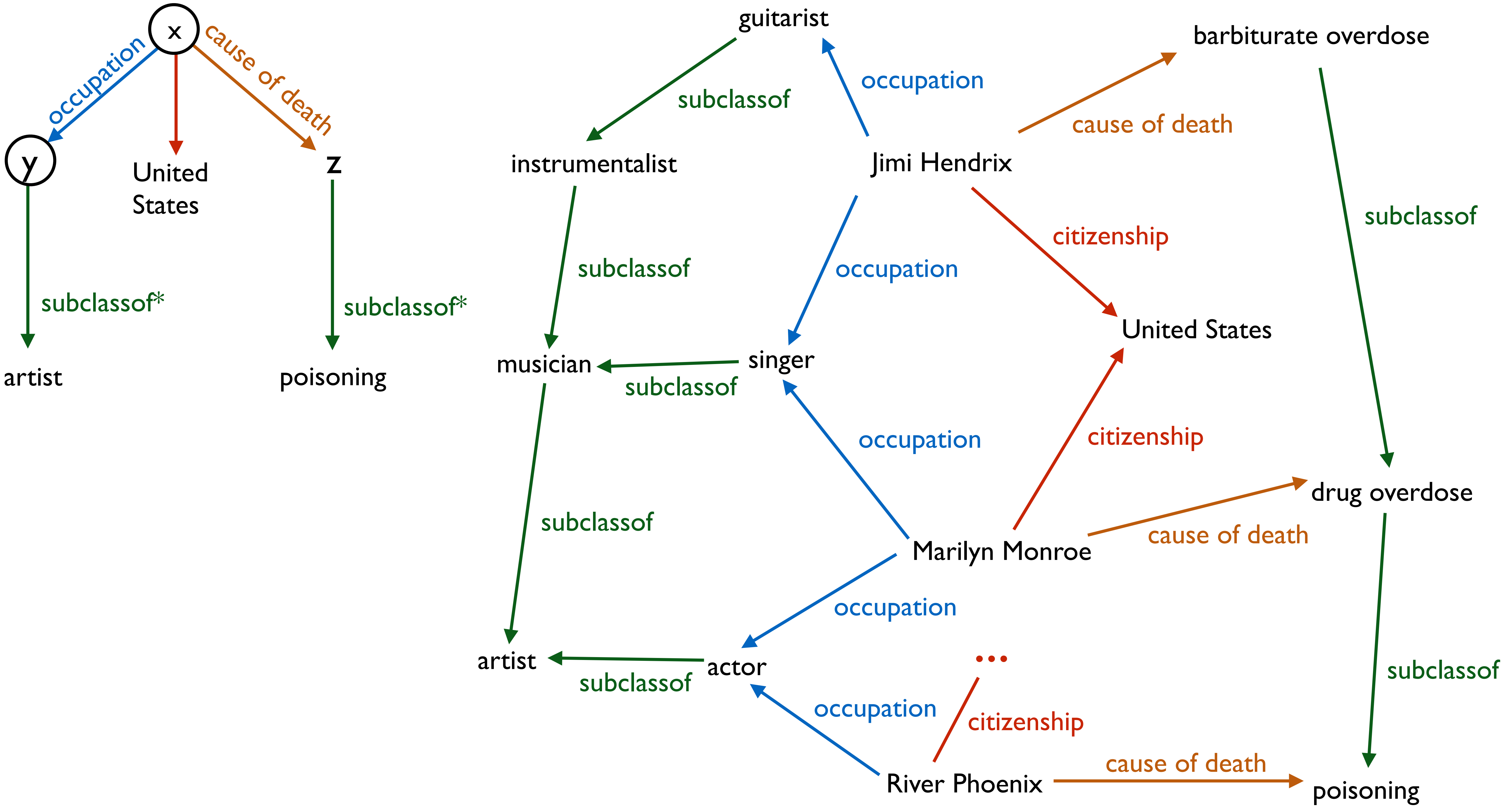


output node

Regular Expressions on edges
Regular Path Queries (RPQs)

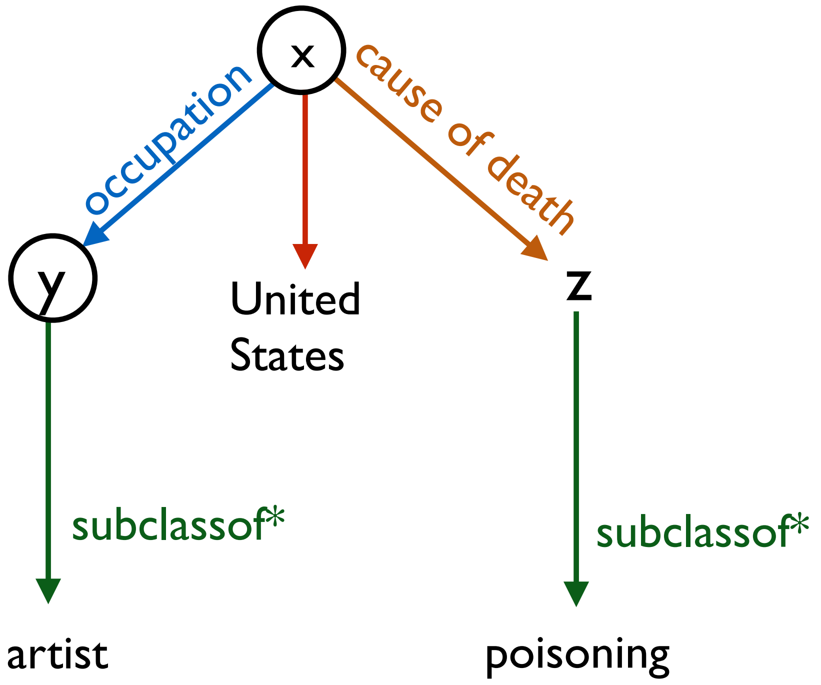
Graph Queries By Example

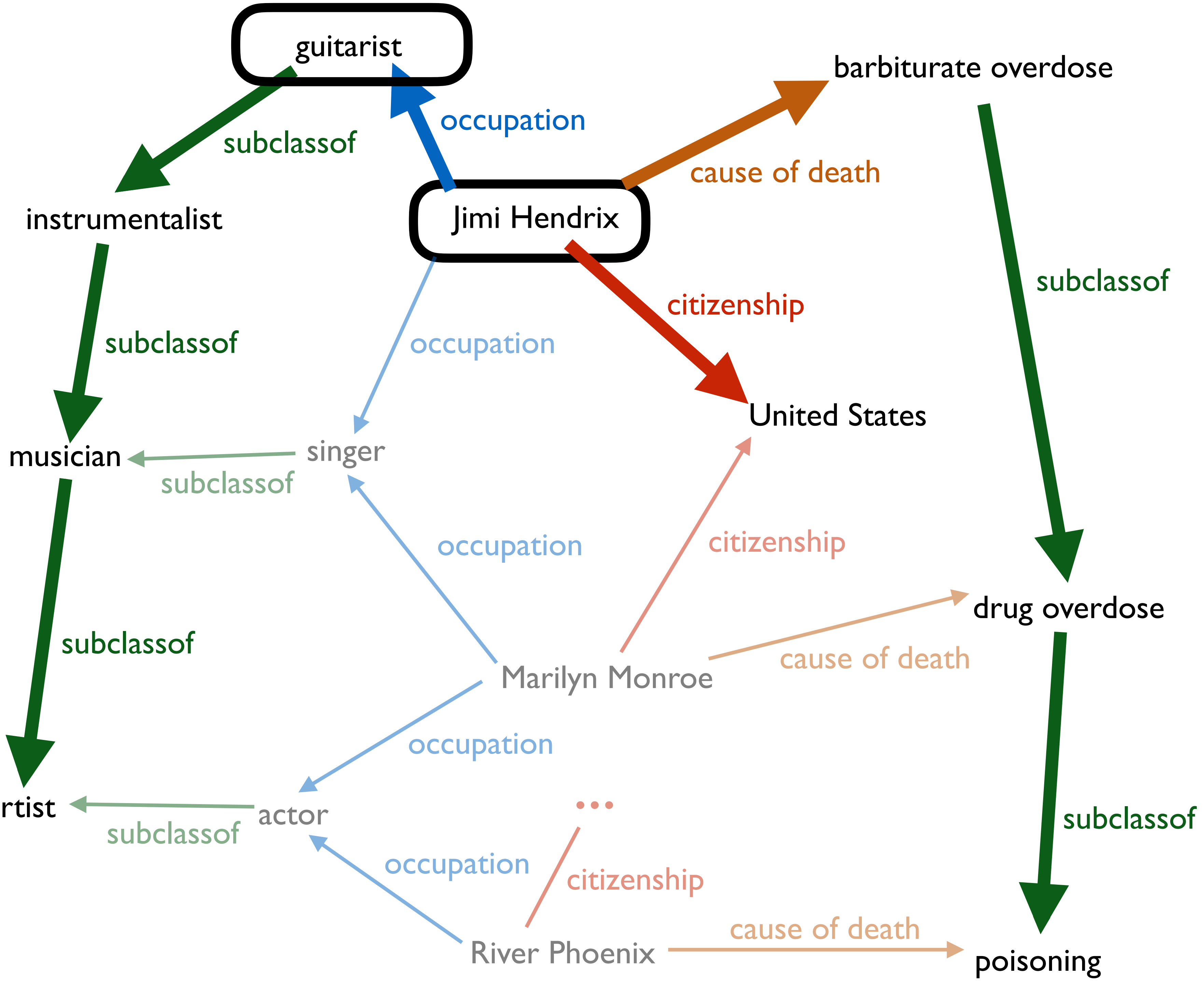
"US artists who died of poisoning"



Graph Queries By Example

"US artists who died of poisoning"





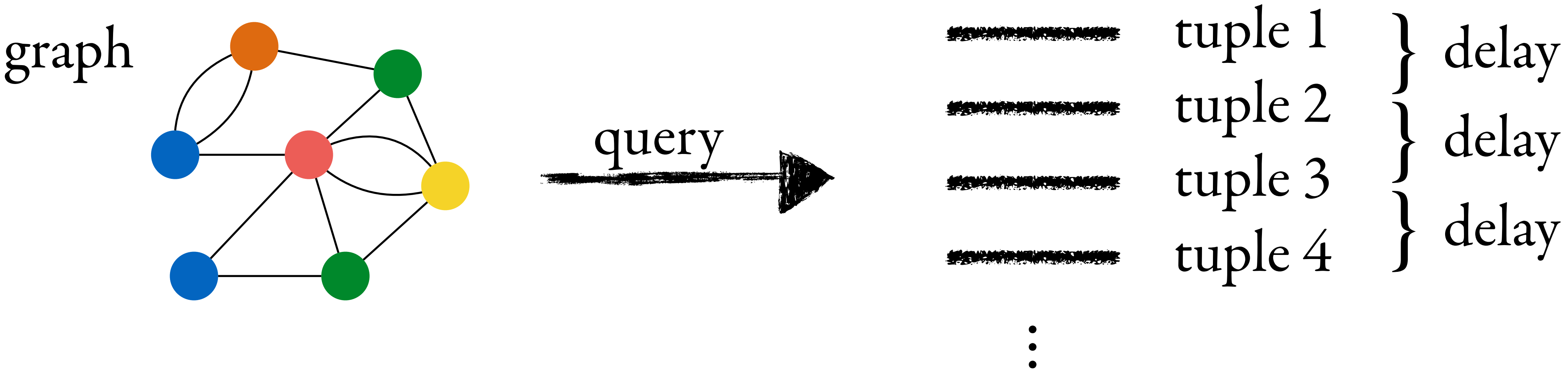
Answer:
(Jimi Hendrix, guitarist)
...

Graph Queries By Example

Such queries are called **Conjunctive Regular Path Queries (CRPQs)**
They are at the core of modern graph database query languages

Research Questions
in
Graph Databases

Classic Types of Research Questions



Enumerating answers with small delay

⇝ [M., Trautner ICDT'18, Arenas et al., PODS'19]

Answer testing, counting number of answers

⇝ [Arenas et al. WWW'12, Losemann, M. PODS'12]

Classic Types of Research Questions

$$\text{Query 1} \overset{?}{\subseteq} \text{Query 2}$$

important task in

- query optimization
- reasoning about queries in knowledge bases

Containment of Conjunctive Regular Path Queries is EXPSPACE-complete

⇝ [Calvanese et al., KR'00]

Classic Types of Research Questions

There is MUCH more!

Just check the
SIGMOD / PODS / VLDB / ICDT / EDBT / ICDE
proceedings for papers on graph databases

Nice overview on theory aspects:
[Barceló PODS'13]

Why Are We Not Done?

Three New Aspects to Stir The Pot

(1)

There are different semantics of regular path queries in the literature and in graph database systems!

every path

trail

simple path

shortest path

The differences between these are significant

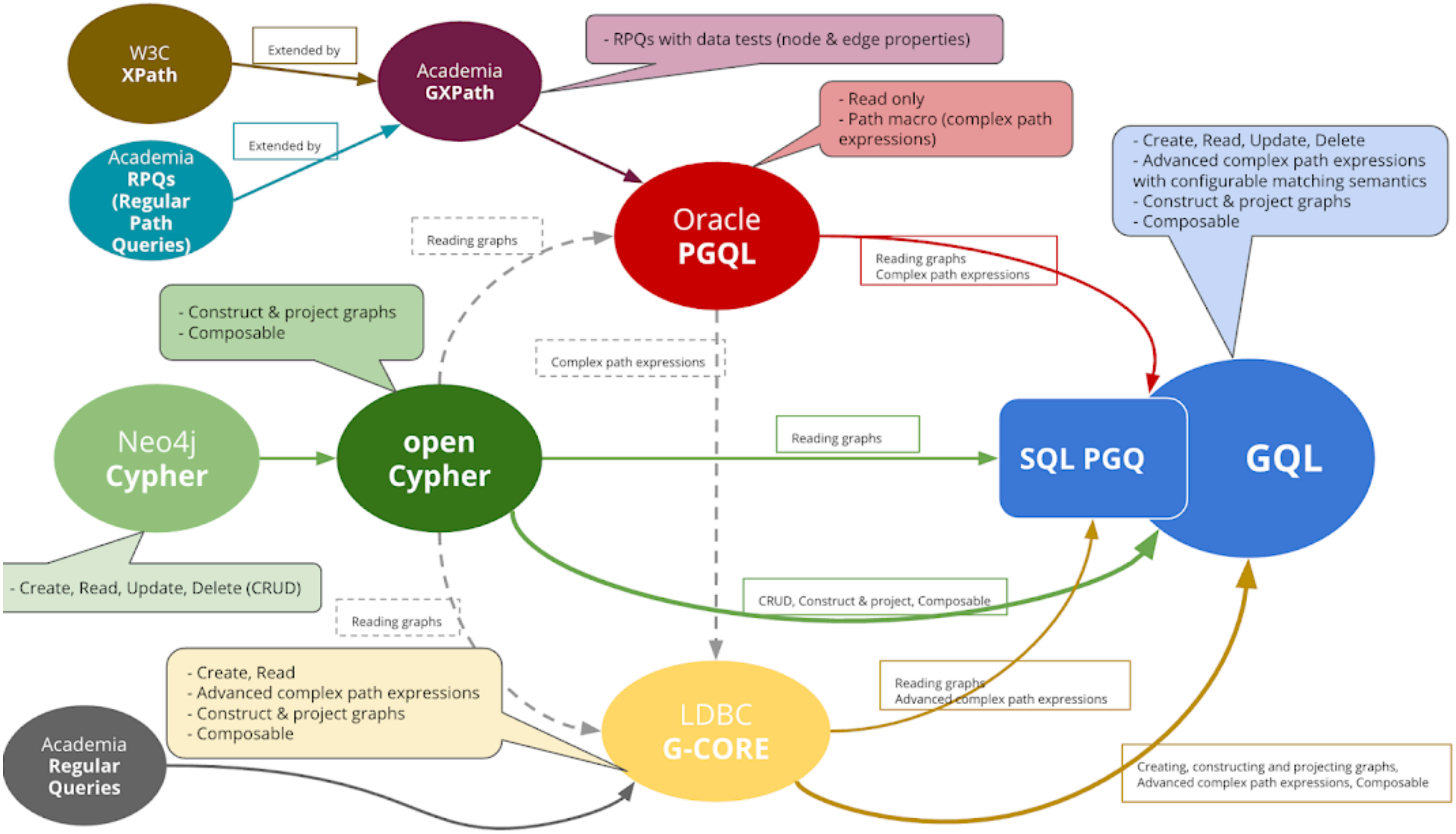
(2)

We now have data about which kinds of queries are used in practice

(3)

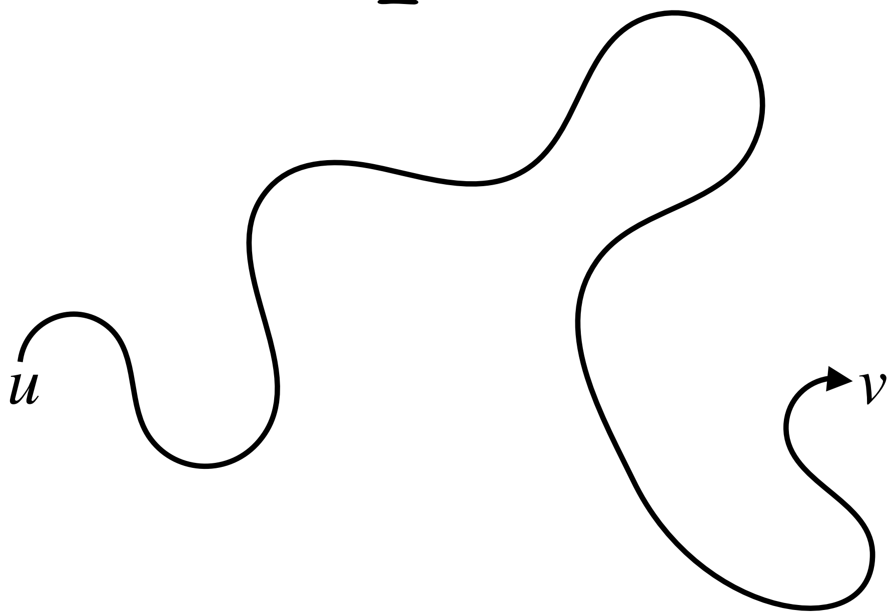
There is a new standardization effort for graph-structured data
(which brings up many new questions)

(3): GQL Influence Graph

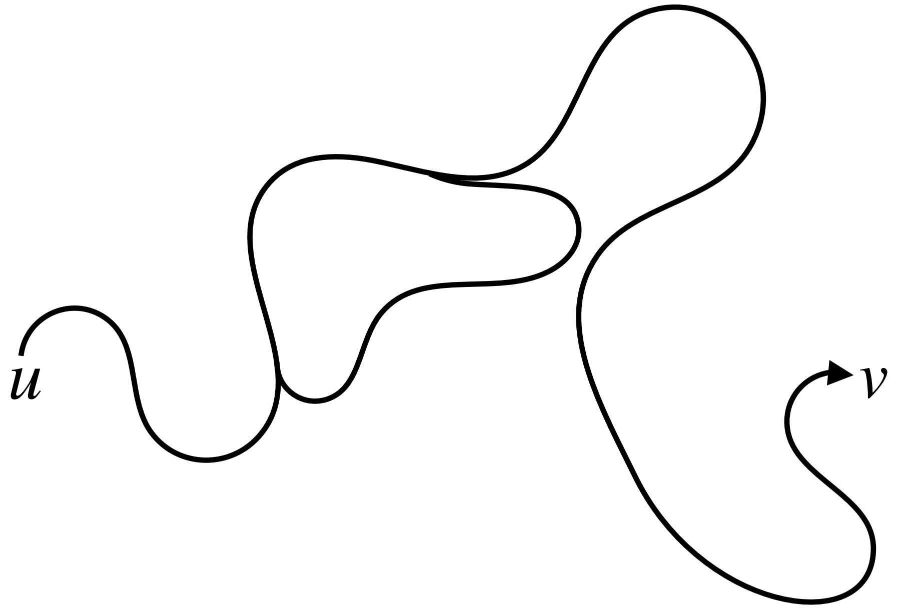

W3C XPath
Extended by
Academia GXPath
- RPQs with data tests (node & edge properties)
Academia RPQs (Regular Path Queries)
Extended by
- Read only
- Path macro (complex path expressions)
Oracle PGQL
- Create, Read, Update, Delete
- Advanced complex path expressions with configurable matching semantics
- Construct & project graphs
- Composable
Reading graphs
Reading graphs
Complex path expressions
- Construct & project graphs
- Composable
Complex path expressions
Neo4j Cypher
open Cypher
Reading graphs
SQL PGQ
GQL
- Create, Read, Update, Delete (CRUD)
CRUD, Construct & project, Composable
Reading graphs
- Create, Read
- Advanced complex path expressions
- Construct & project graphs
- Composable
LDBC G-CORE
Reading graphs
Advanced complex path expressions
Academia Regular Queries
Creating, constructing and projecting graphs, Advanced complex path expressions, Composable

[https://www.gqlstandards.org/existing-languages]

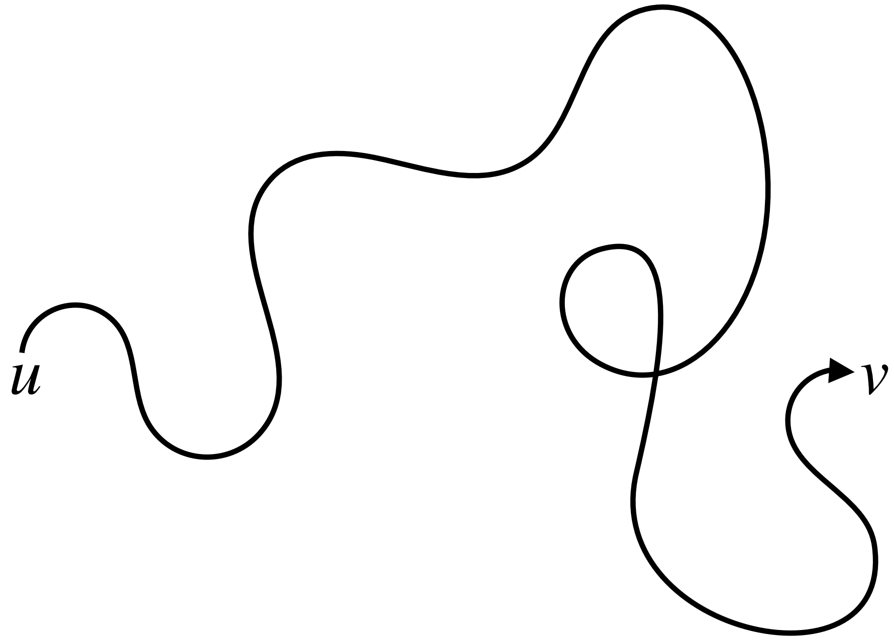
(1): Simple Paths and Trails



Path	✓
Trail	✓
Simple path	✓



Path	✓
Trail	✗
Simple path	✗



Path	✓
Trail	✓
Simple path	✗

(1): Impact of Simple Paths / Trails

The complexity of answer testing / query evaluation changes drastically!

Reason:

- Reachability is easy
- Finding long simple paths is hard

Some papers on simple paths / trails:
[Cruz et al. SIGMOD'87, Mendelzon, Wood SICOMP'95, Bagan et al. PODS'13, M., Trautner ICDT'18, M., Niewerth, Trautner STACS'20]

(2): Expressions Used in Practice

Expression Type	Relative	Expression Type	Relative
A^*	48.76%	$a^*b?$	<0.01%
A	32.10%	abc^*	<0.01%
$a_1 \ldots a_k$	8.66%	$A_1 \ldots A_k$	<0.01%
a^*b	7.73%	ab^*+c	<0.01%
A^+	1.54%	a^*+b	<0.01%
$a_1? \ldots a_k?$	1.15%	$a + b^+$	<0.01%
$aA?$	0.01%	$a^+ + b^+$	<0.01%
$a_1\, a_2? \ldots a_k?$	0.01%	$(ab)^*$	<0.01%
$A?$	<0.01%		

$k \leq 6$

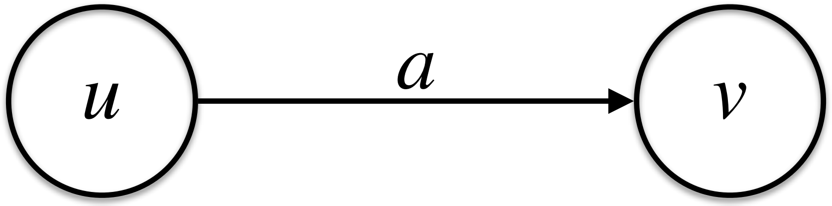
Disjunction of symbols: $A,\ A_1,\ \ldots$

Single symbols: $a,\ b,\ c, a_1, \ldots$

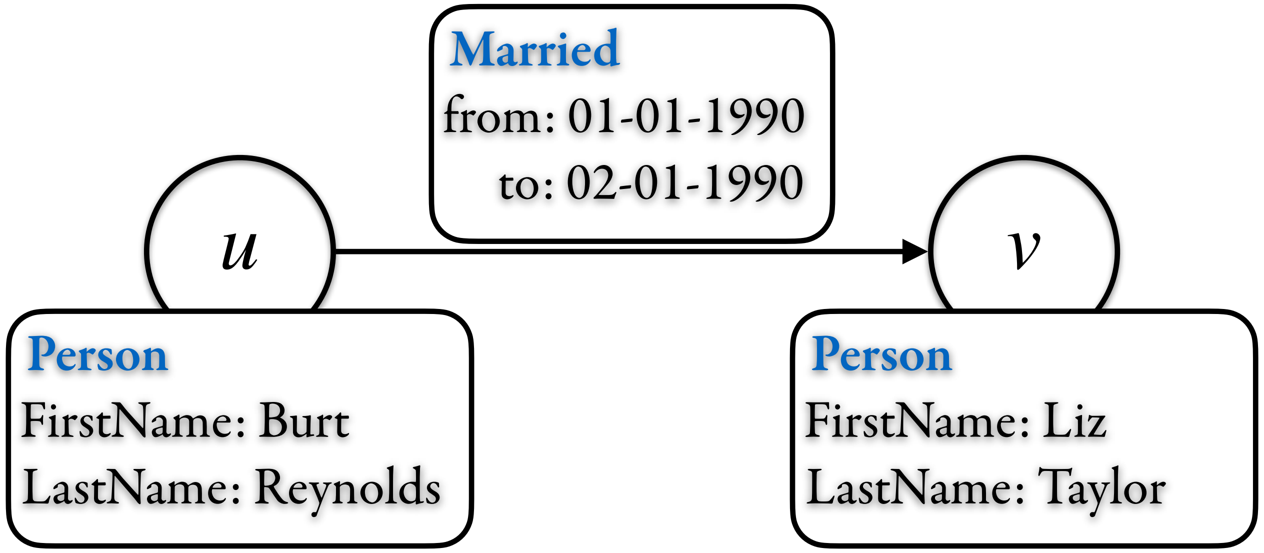
[Bonifati, M., Timm PVLDB'17, WWW'18, WWW'19, SIGMOD'20]

(3): Standardization Effort

Graph:



Property graph:



Married
from: 01-01-1990
to: 02-01-1990

Person
FirstName: Burt
LastName: Reynolds

Person
FirstName: Liz
LastName: Taylor

(3): Standardization Effort

Currently under development:

- Query language (GQL)
- Update language
- Schema language
  - Type system
  - Key / cardinality constraints
- **Data model!**

A lot of theory / practice interaction
is taking place here

Keep an eye on gqlstandards.org!

To Conclude

Logic and FL Topics

There are plenty of nice topics in database theory that connect to logic!

- Information Extraction
- Graph Databases
- Tree-Structured Data (e.g., JSON)
- Tabular Data (e.g., CSV-like data)
- Query (i.e., formula) evaluation
- Query optimization
- Data exchange
- Schema languages
- Probabilistic data
- Incomplete data
- Data management & AI

. . .

Moreover,
(1) the field nourishes connections to practice
(2) database theory has a very nice community
(3) you can find some really nice problems to work on

Thank You!

References

[Amarilli et al. ICDT'19]
Antoine Amarilli, Pierre Bourhis, Stefan Mengel, Matthias Niewerth:
Constant-Delay Enumeration for Nondeterministic Document Spanners.
ICDT 2019: 22:1-22:19

[Arenas et al., PODS'19]
Marcelo Arenas, Luis Alberto Croquevielle, Rajesh Jayaram, Cristian Riveros:
Efficient Logspace Classes for Enumeration, Counting, and Uniform Generation.
PODS 2019: 59-73

[Arenas et al., WWW'12]
Marcelo Arenas, Sebastián Conca, Jorge Pérez:
Counting beyond a Yottabyte, or how SPARQL 1.1 property paths will prevent adoption of the standard.
WWW 2012: 629-638

[Bagan et al. PODS'13]
Guillaume Bagan, Angela Bonifati, Benoît Groz:
A trichotomy for regular simple path queries on graphs.
PODS 2013: 261-272

[Barceló PODS'13]
Pablo Barceló Baeza:
Querying graph databases.
PODS 2013: 175-188

References

[Bonifati et al. PVLDB 2017]
Angela Bonifati, Thomas Timm, and Wim Martens.
An Analytical Study of Large SPARQL Query Logs.
PVLDB 11(2): 149-161 (2017)

[Bonifati et al. WWW 2019]
Angela Bonifati, Thomas Timm, and Wim Martens.
Navigating the Maze of Wikidata Query Logs.
The Web Conference 2019

[Calvanese et al. KR 2000]
Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Moshe Y. Vardi:
Containment of Conjunctive Regular Path Queries with Inverse.
KR 2000: 176-185

[Cruz et al. SIGMOD'87]
Isabel F. Cruz, Alberto O. Mendelzon, Peter T. Wood:
A Graphical Query Language Supporting Recursion.
SIGMOD Conference 1987: 323-330

References

[Doleschal et al. PODS'19]
Johannes Doleschal, Benny Kimelfeld, Wim Martens, Yoav Nahshon, Frank Neven:
Split-Correctness in Information Extraction.
PODS 2019: 149-163

[Fagin et al. PODS'13 / JACM'15]
Ronald Fagin, Benny Kimelfeld, Frederick Reiss, Stijn Vansummeren:
Spanners: a formal framework for information extraction.
PODS 2013: 37-48, full version in J. ACM 62(2): 12:1-12:51, 2015

[Fagin et al. TODS'16]
Ronald Fagin, Benny Kimelfeld, Frederick Reiss, Stijn Vansummeren:
Declarative Cleaning of Inconsistencies in Information Extraction.
ACM Trans. Database Syst. 41(1): 6:1-6:44 (2016)

[Florenzano et al. PODS'17]
Fernando Florenzano, Cristian Riveros, Martín Ugarte, Stijn Vansummeren, Domagoj Vrgoc:
Constant Delay Algorithms for Regular Document Spanners.
PODS 2018: 165-177

References

[Kimelfeld EDBTSS'19]
Benny Kimelfeld.
Information Extraction with Document Spanners & Big Data Analytics with Logical Formalisms.
EDBT 2019 Summer School, https://edbtschool2019.liris.cnrs.fr/

[Losemann, Martens PODS'12]
Katja Losemann, Wim Martens:
The complexity of evaluating path expressions in SPARQL.
PODS 2012: 101-112

[Martens, Trautner ICDT'18]
Wim Martens, Tina Trautner:
Evaluation and Enumeration Problems for Regular Path Queries.
ICDT 2018: 19:1-19:21

[Martens, Niewerth, Trautner STACS'20]
Wim Martens, Matthias Niewerth, Tina Trautner:
A Trichotomy for Regular Trail Queries.
STACS 2020: 7:1-7:16

[Mendelzon, Wood SICOMP'95]
Alberto O. Mendelzon, Peter T. Wood:
Finding Regular Simple Paths in Graph Databases.
SIAM J. Comput. 24(6): 1235-1258 (1995)