# KNOWLEDGE DISCOVERY IN GRAPH DATA

FATIHA SAÏS

UNIVERSITÉ PARIS SACLAY
MASTER 2 OF COMPUTER SCIENCE – DATA SCIENCE

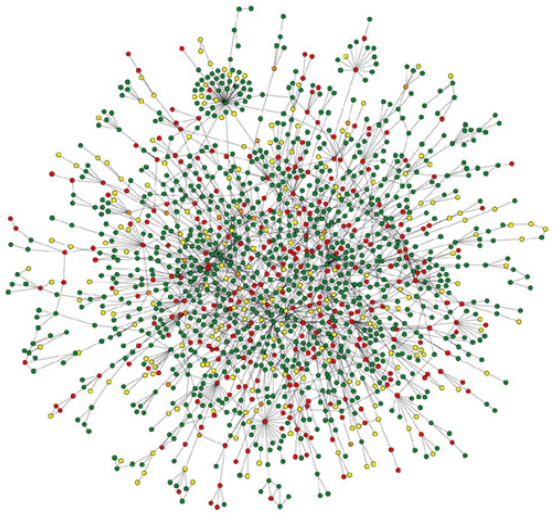
LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

cnrs
dépasser les frontières

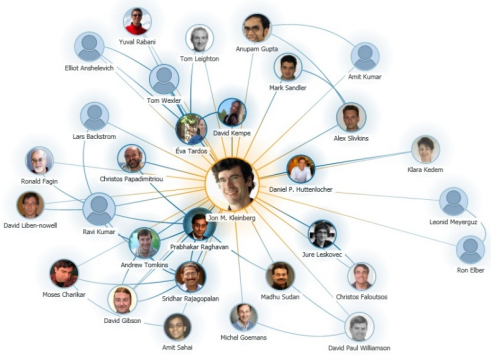
université PARIS-SACLAY

# WHERE THERE IS INFORMATION, THERE ARE GRAPHS!

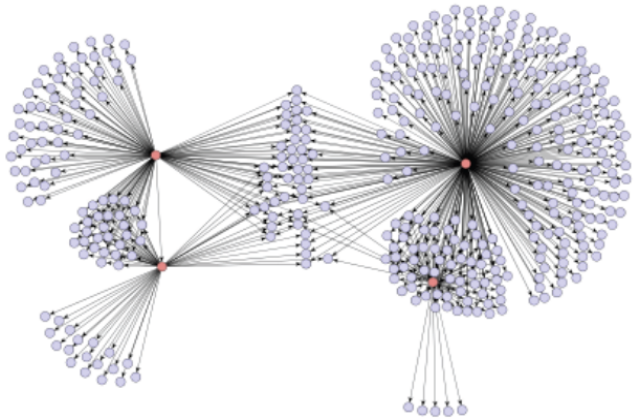
Global Flight Graph



Biological Graph: Protein Interaction



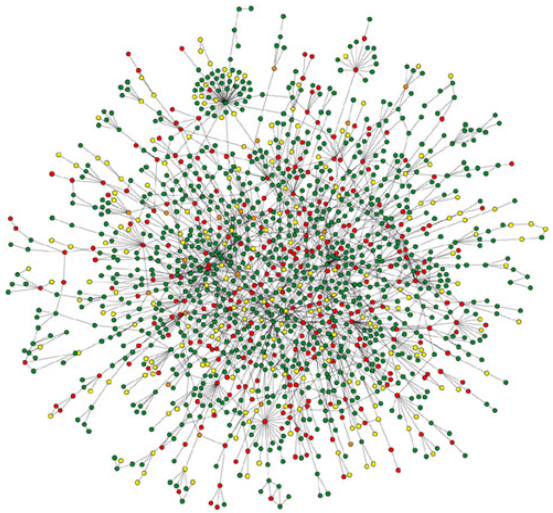

Research Collaboration Network



Product Recommendation Network via Emails

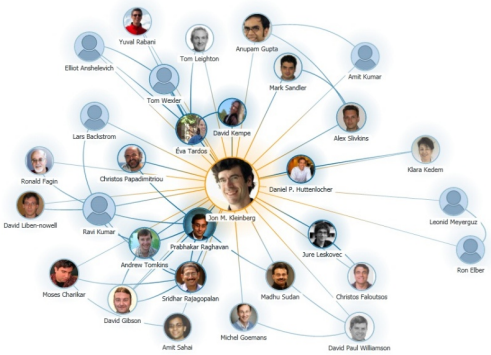
# WHERE THERE IS INFORMATION, THERE ARE GRAPHS!

Global Flight Graph

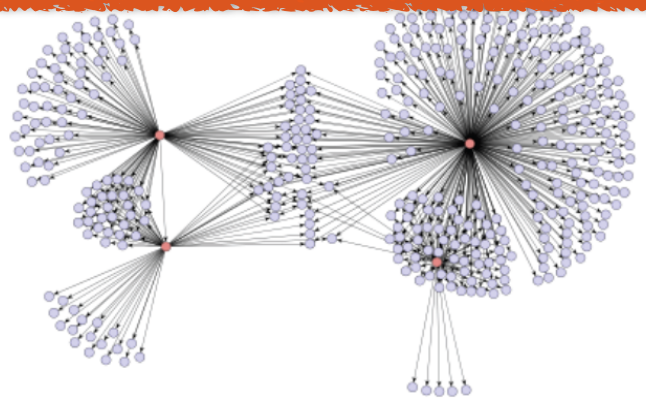


Biological Graph: Protein Interaction

Most of the time they are treated as Homogeneous graphs:
single-typed nodes, single-typed links!




Research Collaboration Network



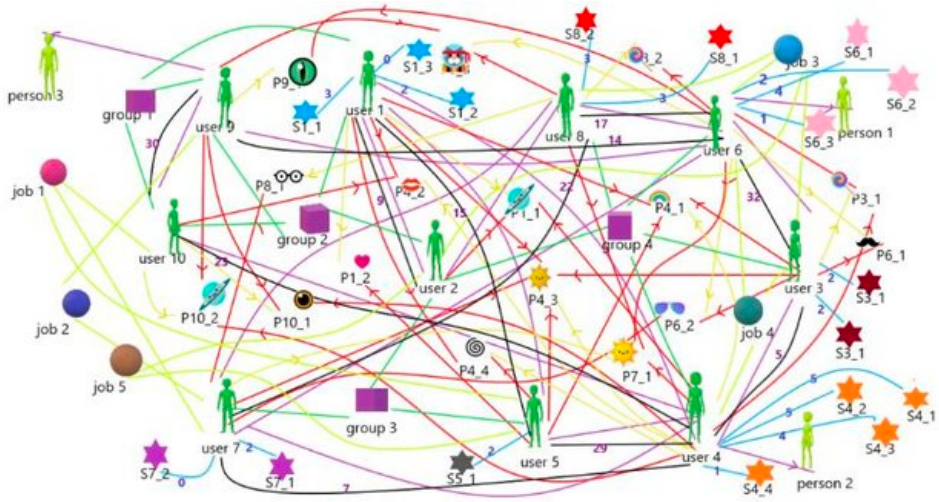
Product Recommendation Network via Emails

# THE REAL WORLD: HETEROGENEOUS GRAPHS

Bibliographic Graph


**Conference, Paper, Author,**
**Publishing, writing, co-author, …**

Social Graph



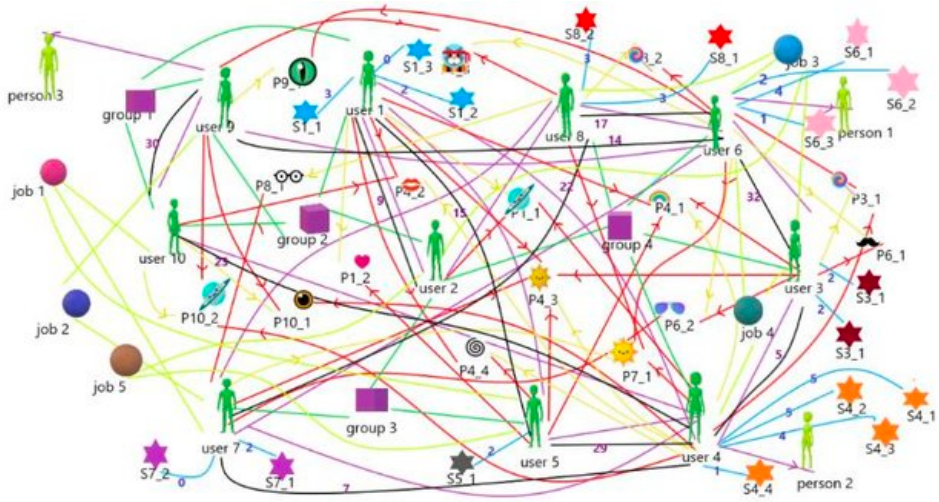

**Groups, Users, Posts,**
**Friendship, Reactions, …**

# THE REAL WORLD: HETEROGENEOUS GRAPHS

Bibliographic Graph


**Conference, Paper, Author,**
**Publishing, writing, co-author, …**

Social Graph




**Groups, Users, Posts,**
**Friendship, Reactions, …**

Homogeneous graphs are *information loss* projection of heterogeneous graphs!

**Directly Mining information-richer heterogeneous networks**

# HOMOGENEOUS VS. HETEROGENEOUS GRAPH MINING

**Homogeneous graphs can often be derived from their original heterogeneous graphs**

- E.g., coauthor networks can be derived from author-paper-conference networks by projection on authors only
- Paper citation networks can be derived from a complete bibliographic network with papers and citations projected

**Heterogeneous graphs carry richer information than its corresponding projected homogeneous graphs**

**Our emphasis: Discovering of knowledge from heterogeneous graphs**

# WHAT CAN BE DISCOVERED FROM HETEROGENEOUS GRAPHS?


## DBLP: A Computer Science bibliographic database

**A sample of publication record in DBLP* ( millions of papers, authors, venues), …**

| Knowledge hidden in DBLP Network | Mining Functions |
|---|---|
| How are CS research areas **structured**? | Clustering |
| Who are the **leading** researchers on Web search? | Ranking |
| What are the most essential **terms, venues, authors** in **AI**? | Classification + Ranking |
| Who are the **peer** researchers of Jure Leskovec? | Similarity Search |
| Whom **will** Christos Faloutsos **collaborate with**? | Relationship Prediction |
| Which types of **relationships** are most **influential** for an author to decide her topics? | Relation Strength Learning |
| How was the field of Data Mining **emerged** or **evolving**? | Network Evolution |
| Which authors are **rather different** from his/her peers in IR? | Outlier/anomaly detection |
| How many supervisors one can have for his/her PhD? | Axiom mining |

*https://dblp.uni-trier.de/

# OUR HETEROGENEOUS GRAPHS ARE

# KNOWLEDGE GRAPHS

# COURSE AGENDA

## Part 1 - Knowledge Discovery for linking knowledge graphs

- Keys and conditional keys discovery for data linking
- Referring Expressions Discovery for Data Linking

## Part 2 - Rule Discovery for knowledge graph completion

- Horn rule mining: survey of existing approaches

## Part 3 - Knowledge Graph Embedding techniques

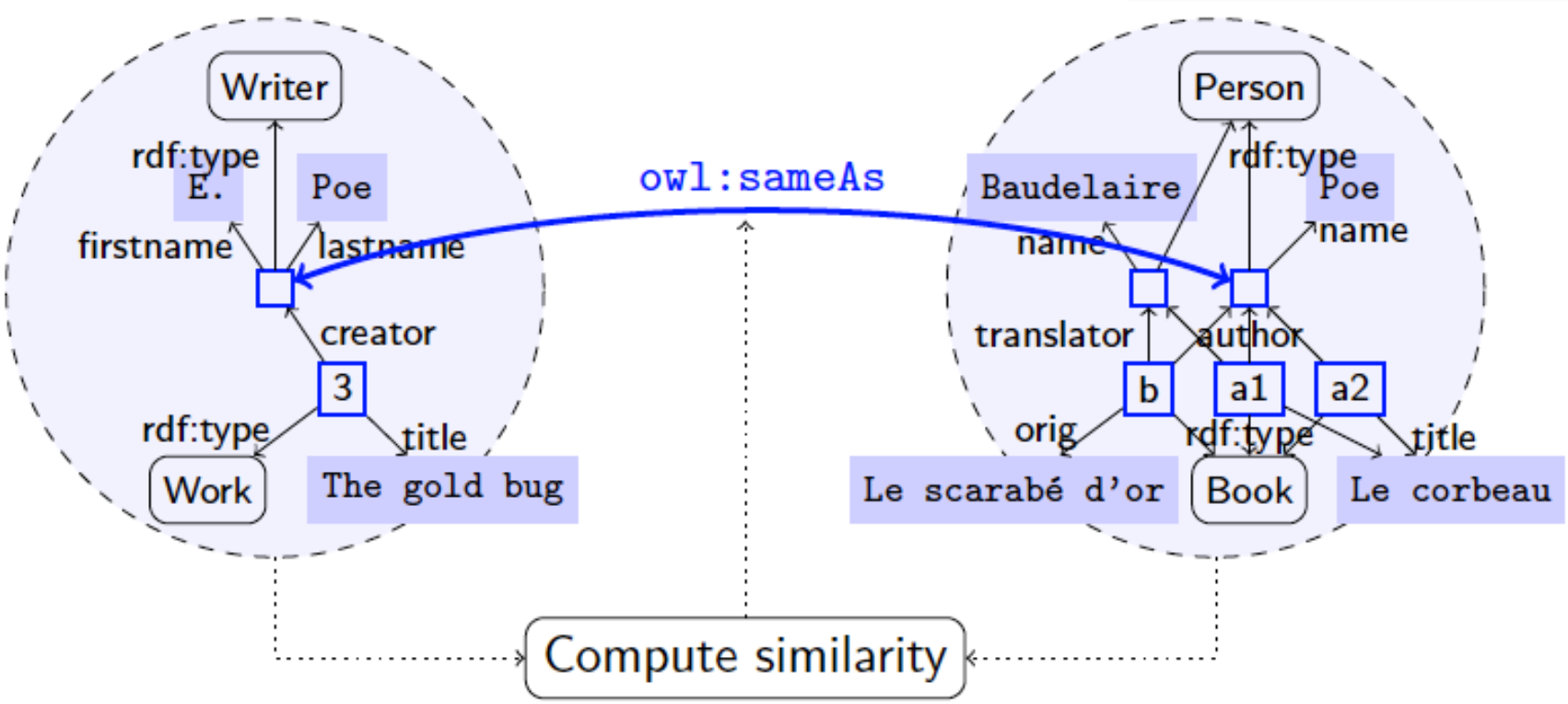
# KEY DISCOVERY IN GRAPH DATA FOR DATA LINKING

FATIHA SAÏS

UNIVERSITÉ PARIS SACLAY
MASTER 2 OF COMPUTER SCIENCE – DATA SCIENCE



# DATA LINKING PROBLEM




**Different kinds of approaches:**

- Similarity based,
- ML based
- Rule based

# KEY DISCOVERY FOR DATA LINKING

- **Rule-based data linking approaches** [Saïs et al 2007, 2009]: need for knowledge to be declared in an ontology language or other languages.

  homepage(X, Y) ∧ homepage(Z, Y) ➔ sameAs(X, Z)

**A key: is a set of properties that uniquely identifies every instance of a class**

| | … | homepage |
|---|---|---|
| **museum11** | | www.louvre.com |
| **museum12** | | www.musee-orsay.fr |
| **museum13** | | www.quai-branly.fr |
| **museum14** | | … |

| homepage | … | |
|---|---|---|
| www.louvre.com | | **museum21** |
| www.musee-orsay.fr | | **museum22** |
| www.quai-branly.fr | | **museum23** |
| … | | **museum24** |

# KEY DISCOVERY FOR DATA LINKING

- **Rule-based data linking approaches** [Saïs et al 2007, 2009]: need for knowledge to be declared in an ontology language or other languages.

  homepage(X, Y) ∧ homepage(Z, Y) ➔ sameAs(X, Z)

- **Then we may infer:**

  *sameAs(museum11, museum21)*
  *sameAs(museum12, museum22)*
  *sameAs(museum13, museum23)*

**A key: is a set of properties that uniquely identifies every instance of a class**

| | … | homepage |
|---|---|---|
| **museum11** | | www.louvre.com |
| **museum12** | | www.musee-orsay.fr |
| **museum13** | | www.quai-branly.fr |
| **museum14** | | … |

SameAs
SameAs
SameAs

| homepage | … | |
|---|---|---|
| www.louvre.com | | **museum21** |
| www.musee-orsay.fr | | **museum22** |
| www.quai-branly.fr | | **museum23** |
| … | | **museum24** |

# KEY DISCOVERY FOR DATA LINKING

- **Rule-based data linking approaches** [Saïs et al 2007, 2009]: need for knowledge to be declared in an ontology language or other languages.

  homepage(X, Y) ∧ homepage(Z, Y) ➔ sameAs(X, Z)

- **Then we may infer:**

  *sameAs(museum11, museum21)*
  *sameAs(museum12, museum22)*
  *sameAs(museum13, museum23)*

**A key: is a set of properties that uniquely identifies every instance of a class**

| | … | homepage | | homepage | … | |
|---|---|---|---|---|---|---|
| **museum11** | | www.louvre.com | SameAs | www.louvre.com | | **museum21** |
| **museum12** | | www.musee-orsay.fr | SameAs | www.musee-orsay.fr | | **museum22** |
| **museum13** | | www.quai-branly.fr | SameAs | www.quai-branly.fr | | **museum23** |
| **museum14** | | … | | … | | **museum24** |

How to automatically discover keys from KGs?

# KEYS IN RELATIONAL DATABASES

- **Key**: A set of properties that uniquely identifies every instance in the data

| | FirstName | LastName | Birthdate | Profession |
|---|---|---|---|---|
| **Person1** | Anne | Tompson | 15/02/88 | Actor |
| **Person2** | Marie | Tompson | 02/09/75 | Researcher |
| **Person3** | Marie | David | 15/02/85 | Teacher |
| **Person4** | Vincent | Solgar | 06/12/90 | Teacher |

*Is [FirstName] a key?*

*Is [LastName] a key?*

# KEYS IN RELATIONAL DATABASES

- **Key**: A set of properties that uniquely identifies every instance in the data

| | FirstName | LastName | Birthdate | Profession |
|---|---|---|---|---|
| **Person1** | Anne | Tompson | 15/02/88 | Actor |
| **Person2** | Marie | Tompson | 02/09/75 | Researcher |
| **Person3** | Marie | David | 15/02/85 | Teacher |
| **Person4** | Vincent | Solgar | 06/12/90 | Teacher |

*Is [FirstName] a key?* ✖

*Is [LastName] a key?* ✖

# KEYS IN RELATIONAL DATABASES

- **Key**: A set of properties that uniquely identifies every instance in the data

| | FirstName | LastName | Birthdate | Profession |
|---|---|---|---|---|
| **Person1** | Anne | Tompson | 15/02/88 | Actor |
| **Person2** | Marie | Tompson | 02/09/75 | Researcher |
| **Person3** | Marie | David | 15/02/85 | Teacher |
| **Person4** | Vincent | Solgar | 06/12/90 | Teacher |

*Is [FirstName] a key?* ✖

*Is [LastName] a key?* ✖

*Is [FirstName,LastName] a key?* ✔

# KEYS: DATABASES VS. KNOWLEDGE GRAPHS

- **RDF data conform to ontologies**
  - Key discovery **on a given class**
  - Inference => obtain a more complete information about entities

  - Knowledge based pruning (e.g. key heritage)
    - {SSN}: key for all the instances of the class Person
    - {SSN}: key for all subclasses of the class Person (ex. Researcher, Professor etc.)

-

# KEYS: DATABASES VS. KNOWLEDGE GRAPHS

- **RDF data conform to ontologies**
  - Key discovery **on a given class**
  - Inference => obtain a more complete information about entities

  - Knowledge based pruning (e.g. key heritage)
    - {SSN}: key for all the instances of the class Person
    - {SSN}: key for all subclasses of the class Person (ex. Researcher, Professor etc.)

- **RDF data completeness**
  - Interpretation of no values

- **RDF data quality**
  - Deal with erroneous data

- **Volume of RDF datasets**

# KEYS DECLARED BY EXPERTS FOR DATA LINKING

- **Not an easy task:**
  - Experts are not aware of all the keys

    **Ex. {SSN}, {ISBN} easy to declare**
    **Ex. {Name, DateOfBirth, BornIn} is it a key for the class Person?**

  - Erroneous keys can be given by experts

  - As many keys as possible
    - More keys => More linking rules

**Goal: Discover keys automatically**

# KEYS IN KNOWLEDGE GRAPHS

# KEY PROPERTIES - KEY MONOTONICITY

- **Key monotonicity:** When a set of properties is a key, all its supersets are also keys
- **Minimal Key:** A key that by removing one property stops being a key

| | FirstName | LastName | SSN | DateOfBirth | StudiedIn | HasSibling |
|---|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | – | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | 05/03/85 | – | p1, p4 |
| **p4** | Helen | Roger | 767960154 | 10/08/79 | UCC, UCD | – |
| **p4** | Marc | Brown | – | – | Yale | p1, p2 |
| **p5** | Helen | Roger | 767960154 | – | – | – |

***Minimal key: [FirstName,LastName]***

***Not a minimal key: [FirstName,LastName, dateOfBirth]***

# HASKEY AXIOM IN OWL2

**Satisfaction of Keys**

## Axiom

$$\texttt{HasKey(}\ \texttt{CE}\ \texttt{(}\ \texttt{OPE}_1\ \texttt{...}\ \texttt{OPE}_m\ \texttt{)}\ \texttt{(}\ \texttt{DPE}_1\ \texttt{...}\ \texttt{DPE}_n\ \texttt{)}\ \texttt{)}$$

## Condition

$\forall\ x, y, z_1, \ldots, z_m, w_1, \ldots, w_n$ :

if $x \in (CE)^C$ and $x \in NAMED$ and

$y \in (CE)^C$ and $y \in NAMED$ and

$(x, z_i) \in (OPE_i)^{OP}$ and $(y, z_i) \in (OPE_i)^{OP}$ and $z_i \in NAMED$ for each $1 \leq i \leq m$ and

$(x, w_j) \in (DPE_j)^{DP}$ and $(y, w_j) \in (DPE_j)^{DP}$ for each $1 \leq j \leq n$

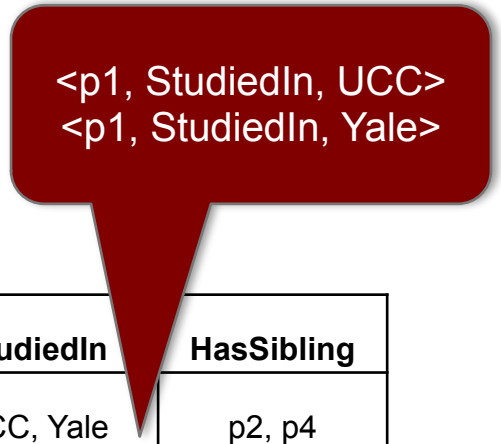
then x = y

# KEYS IN KNOWLEDGE GRAPHS

| | FirstName | LastName | SSN | DateOfBirth | StudiedIn | HasSibling |
|---|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | – | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | 05/03/85 | – | p1, p4 |
| **p4** | Helen | Roger | 767960154 | 10/08/79 | UCC, UCD | – |
| **p4** | Marc | Brown | – | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 767960154 | – | – | – |

# KEYS IN KNOWLEDGE GRAPHS

- **Multi-valued properties:** **value** difference or **set** difference




| | FirstName | LastName | SSN | DateOfBirth | StudiedIn | HasSibling |
|---|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | – | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | 05/03/85 | – | p1, p4 |
| **p4** | Helen | Roger | 767960154 | 10/08/79 | UCC, UCD | – |
| **p4** | Marc | Brown | – | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 767960154 | – | – | – |

# KEYS IN KNOWLEDGE GRAPHS

- **Multi-valued properties: value** difference or **set** difference
- **Incomplete descriptions:**
  - **Optimistic**: not given value is considered as being different from the existing ones?
  - **Pessimistic**: not given value is considered as likely being identical to the existing ones?

| | FirstName | LastName | SSN | DateOfBirth | StudiedIn | HasSibling |
|---|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | – | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | 05/03/85 | – | p1, p4 |
| **p4** | Helen | Roger | 767960154 | 10/08/79 | UCC, UCD | – |
| **p4** | Marc | Brown | – | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 767960154 | – | – | – |

No triple containing the birthdate of p1

# KEYS IN KNOWLEDGE GRAPHS

- **Three main types of keys in the KGs:**
  - **S-keys (conforming to OWL2)**
  - **SF-keys**
  - **F-keys**

| | **Multivaluation** | **Incompletness** |
|---|---|---|
| **S-Keys** | Value difference | Optimistic |
| **SF-Keys** | Set difference | Optimistic |
| **F-Keys** | Set difference | Pessimistic |

- **Optimistic**: not given value is considered as being different from the existing ones
- **Pessimistic**: not given value is considered as likely being identical to the existing ones

# KEYS IN KNOWLEDGE GRAPHS

- **Three main types of keys in the KGs:**
  - **S-keys (conforming to OWL2)**

| | FirstName | LastName | SSN | StudiedIn | HasSibling |
|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | – | p1, p4 |
| **p4** | Marie | Roger | 767960154 | UCC, UCD | – |
| **p4** | Marc | Brown | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 967960158 | – | – |

**S-keys**
{FirstName, LastName}
{SSN}
{LastName, StudiedIn}
{FirstName, HasSibling}

# KEYS IN KNOWLEDGE GRAPHS

- **Three main types of keys in the KGs:**
  - **S-keys (conforming to OWL2)**

| | FirstName | LastName | SSN | StudiedIn | HasSibling |
|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | – | p1, p4 |
| **p4** | Marie | Roger | 767960154 | UCC, UCD | – |
| **p4** | Marc | Brown | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 967960158 | – | – |

**S-keys**
{FirstName, LastName}
{SSN}
**{LastName, StudiedIn}**
{FirstName, HasSibling}

# KEYS IN KNOWLEDGE GRAPHS

- **Three main types of keys in the KGs:**
  - **S-keys (conforming to OWL2)**
  - **SF-keys**

| | FirstName | LastName | SSN | StudiedIn | HasSibling |
|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | – | p1, p4 |
| **p4** | Marie | Roger | 767960154 | UCC, UCD | – |
| **p4** | Marc | Brown | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 967960158 | – | – |

**S-keys**
{FirstName, LastName}
{SSN}
**{LastName, StudiedIn}**
**{FirstName, HasSibling}**

**SF-keys**
{FirstName, LastName}
{SSN}
**{StudiedIn}**
**{HasSibling}**

# KEYS IN KNOWLEDGE GRAPHS

- **Three main types of keys in the KGs:**
  - **S-keys (conforming to OWL2)**
  - **SF-keys**
  - **F-keys**

| | FirstName | LastName | SSN | StudiedIn | HasSibling |
|---|---|---|---|---|---|
| **p1** | Marie | Brown | 121558745 | UCC, Yale | p2, p4 |
| **p2** | John | Brown | 232351234 | – | p1, p4 |
| **p4** | Marie | Roger | 767960154 | UCC, UCD | – |
| **p4** | Marc | Brown | – | UCD | p1, p2 |
| **p5** | Helen | Roger | 967960158 | – | – |

**S-keys**
{FirstName, LastName}
{SSN}
**{LastName, StudiedIn}**
**{FirstName, HasSibling}**

**SF-keys**
{FirstName, LastName}
{SSN}
**{StudiedIn}**
**{HasSibling}**

**F-keys**
{FirstName, LastName}
**{LastName,StudiedIn}**
...

# KEY DISCOVERY APPROACHES

- **SF-Keys**
  - Keys and Pseudo-Keys Detection for Web Datasets Cleaning and Interlinking [Atencia et al.12]
- **F-Keys**
  - ROCKER: A Refinement Operator for Key Discovery [Soru et al. 2015]
- **S-Keys**
  - An automatic key discovery approach for data linking [Pernelle et al. 13]
  - SAKey: Scalable almost key discovery in RDF data [Symeonidou et al. 14]
  - VICKEY: Conditional key discovery [Symeonidou et al. 17]
  - Linkkey: Data interlinking through robust Linkkey extraction [Atencia et al.14]

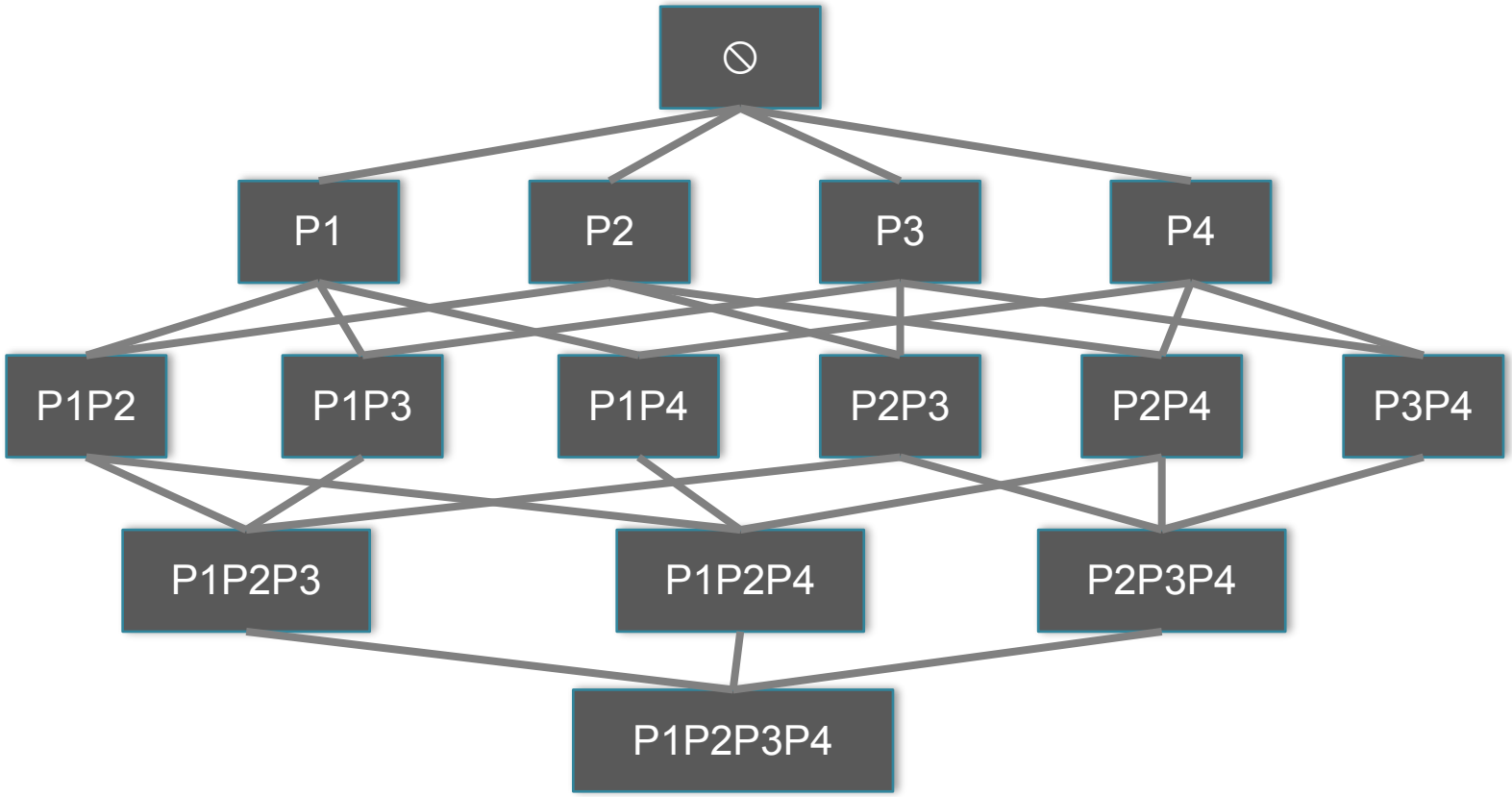
# KEY DISCOVERY APPROACHES

- **SF-Keys**
  - **Keys and Pseudo-Keys Detection for Web Datasets Cleaning and Interlinking** [Atencia et al.12]
- **F-Keys**
  - ROCKER: A Refinement Operator for Key Discovery [Soru et al. 2015]
- **S-Keys**
  - An automatic key discovery approach for data linking [Pernelle et al. 13]
  - SAKey: Scalable almost key discovery in RDF data [Symeonidou et al. 14]
  - VICKEY: Conditional key discovery [Symeonidou et al. 17]
  - Linkkey: Data interlinking through robust Linkkey extraction [Atencia et al.14]

# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

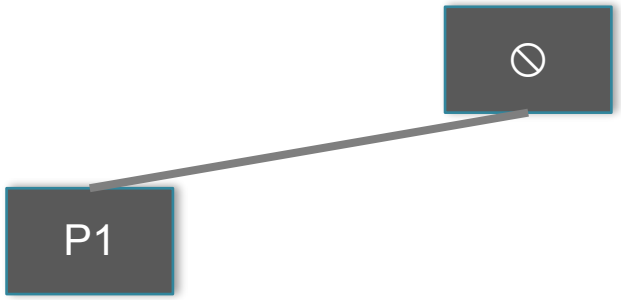
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions



# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

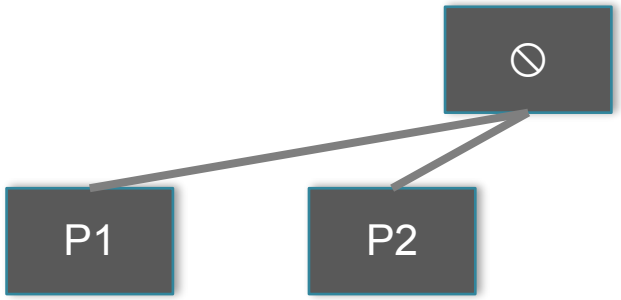
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions



# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

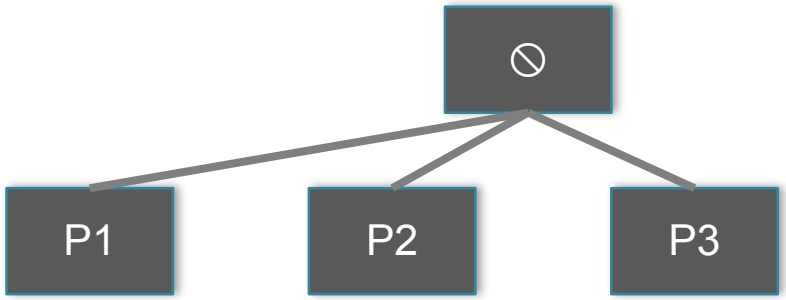
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions



# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

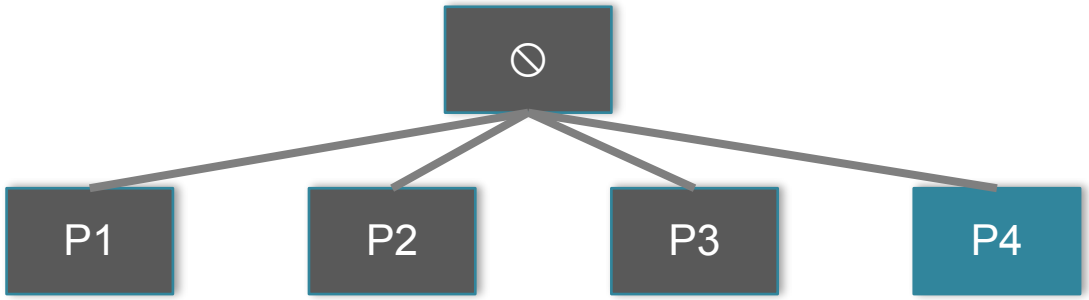
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions



# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

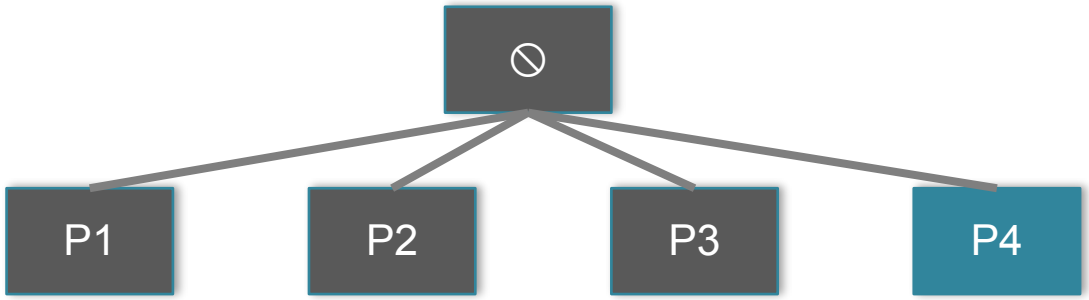
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions



# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions

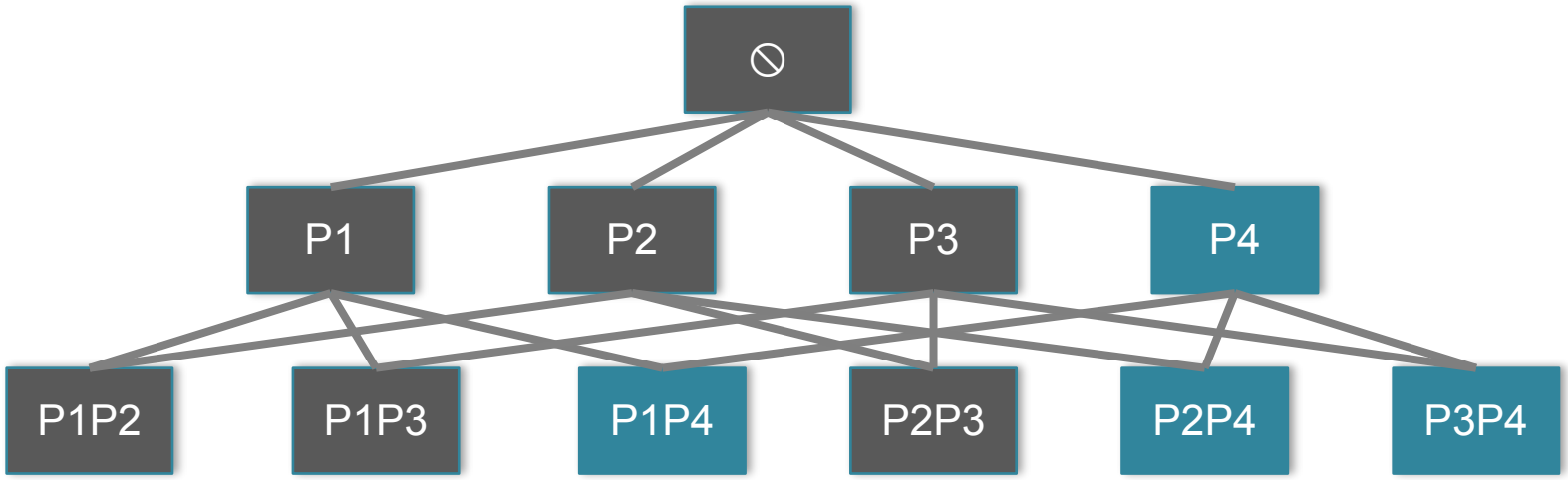



**If P4 is a key => P1P4, P2P4,.., P1P2P3P4 are also keys**

# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions

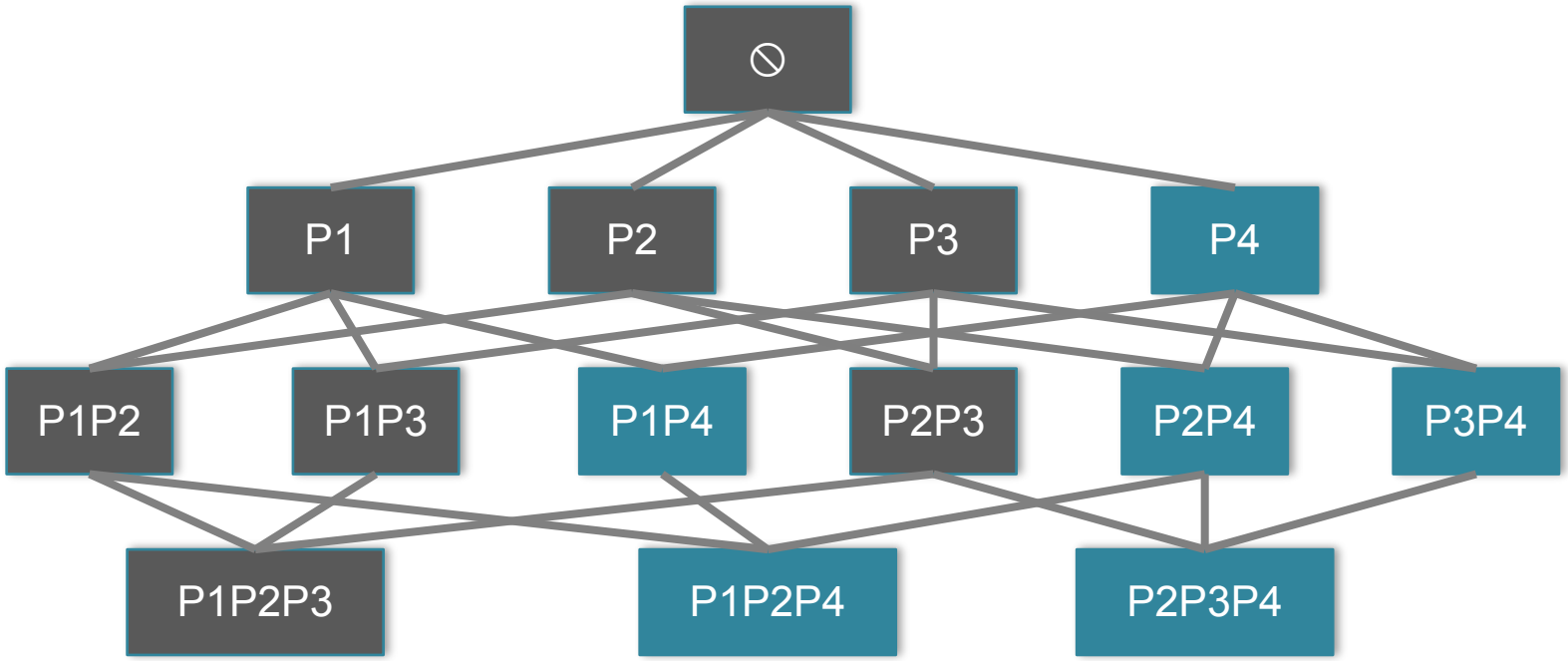



**If P4 is a key => P1P4, P2P4,.., P1P2P3P4 are also keys**

# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions

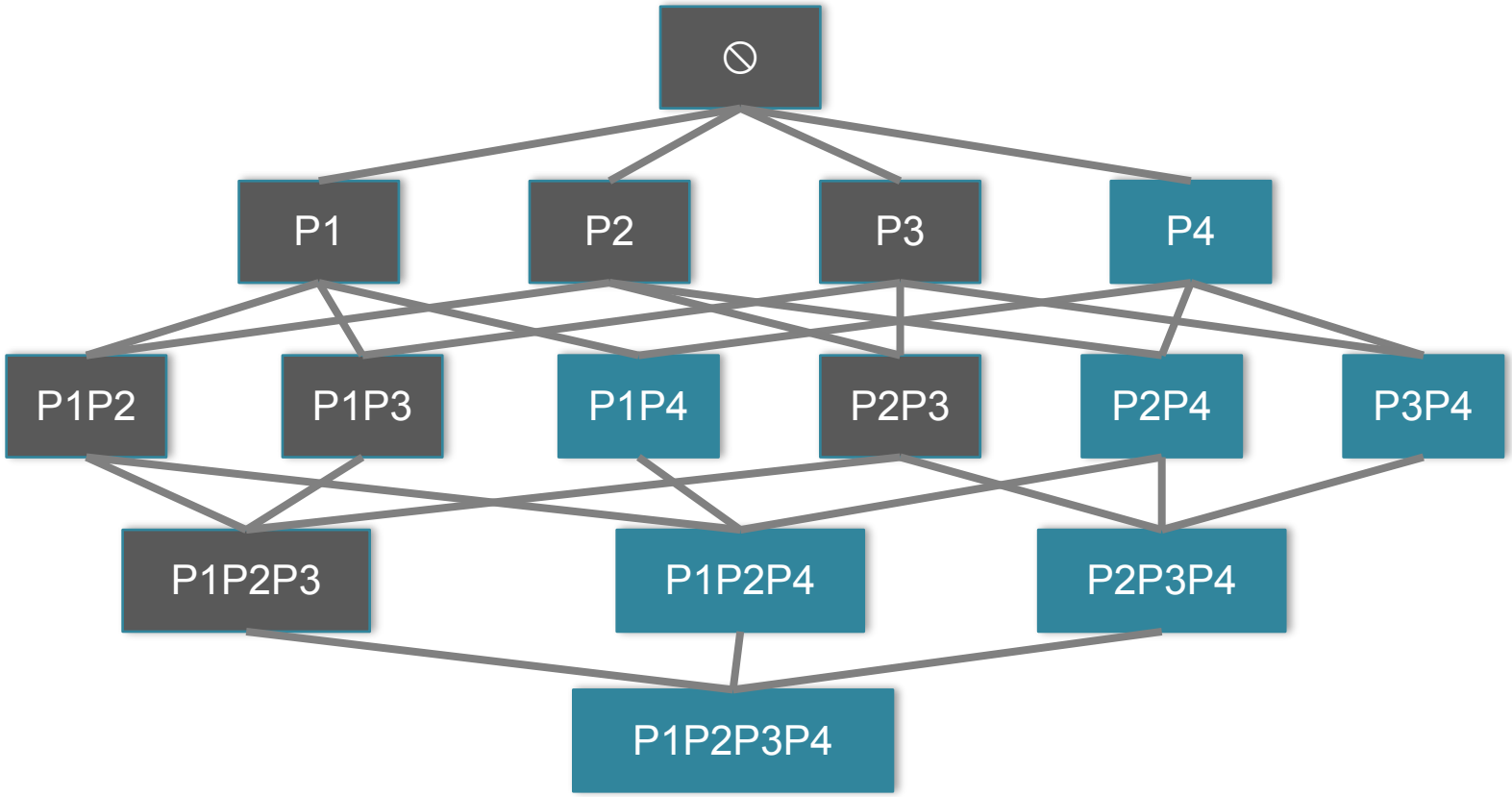



**If P4 is a key => P1P4, P2P4,.., P1P2P3P4 are also keys**

# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions




**If P4 is a key => P1P4, P2P4,.., P1P2P3P4 are also keys**

# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

- **To verify if a set of properties is a key**
  - ***Partition*** instances according to their sharing values
  - If each partition contains only one instances => **Key**

- Key quality measures
  - **Support** of a set of properties *P*:

$$\text{support}(P) = \frac{\#\ instances\ described\ by\ P}{\#\ all\ instances}$$

  - **Discriminability** of a set of properties *P* (pseudo-keys):

$$dis(P) = \frac{\#\ singleton\ partitions}{\#\ partitions}$$

# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING

[Atencia et al.12]

| | Name | Actor | Director | ReleaseDate | Website | Language |
|---|---|---|---|---|---|---|
| **film1** | Ocean's 11 | **B. Pitt**<br>**J. Roberts** | S. Soderbergh | 3/4/01 | www.oceans11.com | --- |
| **film2** | Ocean's 12 | **B. Pitt**<br>**J. Roberts** | S. Soderbergh<br>R. Howard | 2/5/04 | www.oceans12.com | english |
| **film3** | Ocean's 13 | B. Pitt<br>G. Clooney | S. Soderbergh<br>R. Howard | 30/6/07 | www.oceans13.com | english |
| **film4** | The descendants | N. Krause<br>G. Clooney | A. Payne | 15/9/11 | --- | english |
| **film5** | Bourne Identity | D. Liman | --- | 12/6/12 | www.bourneIdentity.com | english |

**Partitions using [Actor]**


→ **[Actor]**: pseudokey
Support = 5/5 =1
Discriminability = 3/4=0.75

# KEY DISCOVERY APPROACHES

- **SF-Keys**
  - Keys and Pseudo-Keys Detection for Web Datasets Cleaning and Interlinking
- **F-Keys**
  - **ROCKER: A Refinement Operator for Key Discovery**
- **S-Keys**
  - An automatic key discovery approach for data linking
  - SAKey: Scalable almost key discovery in RDF data
  - VICKEY: Conditional key discovery
  - Linkkey: Data interlinking through robust Linkkey extraction

# ROCKER: A REFINEMENT OPERATOR FOR KEY DISCOVERY

[Soru et al. 2015]

- **Bottom-up approach that discovers in an efficient way**
  - **F-keys for a given class**
  - **F-pseudo keys for a given class**
- **Key quality measures**
  - ***Discriminability(P):*** # of distinguished instances using the set *P*
  - ***Score(P)*** = *discriminability(P)/# instances* Score: [0,1]
    - Key => **score = 1**
    - Pseudo key => **score < 1**

| | Name | Director | ReleaseDate | Website | Language |
|---|---|---|---|---|---|
| **film1** | Ocean's 11 | S. Soderbergh | 3/4/01 | www.oceans11.com | --- |
| **film2** | Ocean's 12 | S. Soderbergh<br>R. Howard | 2/5/04 | www.oceans12.com | english |
| **film4** | The descendants | A. Payne | 15/9/11 | --- | english |
| **film5** | Bourne Identity | D. Liman | 12/6/12 | --- | english |

# ROCKER: A REFINEMENT OPERATOR FOR KEY DISCOVERY

[Soru et al. 2015]

- **Bottom-up approach that discovers in an efficient way**
  - **F-keys for a given class**
  - **F-pseudo keys for a given class**
- **Key quality measures**
  - ***Discriminability(P):*** # of distinguished instances using the set *P*
  - ***Score(P)*** = *discriminability(P)/# instances* Score: [0,1]
    - Key => **score = 1**
    - Pseudo key => **score < 1**

**Not a key**

| | Name | Director | ReleaseDate | Website | Language |
|---|---|---|---|---|---|
| **film1** | Ocean's 11 | S. Soderbergh | 3/4/01 | www.oceans11.com | --- |
| **film2** | Ocean's 12 | S. Soderbergh<br>R. Howard | 2/5/04 | www.oceans12.com | english |
| **film4** | The descendants | A. Payne | 15/9/11 | --- | english |
| **film5** | Bourne Identity | D. Liman | 12/6/12 | --- | english |

# ROCKER: A REFINEMENT OPERATOR FOR KEY DISCOVERY

[Soru et al. 2015]

- **Bottom-up approach that discovers in an efficient way**
  - **F-keys for a given class**
  - **F-pseudo keys for a given class**
- **Key quality measures**
  - ***Discriminability(P):*** # of distinguished instances using the set *P*
  - ***Score(P)*** = *discriminability(P)/# instances* Score: [0,1]
    - Key => **score = 1**
    - Pseudo key => **score < 1**

Key

| | Name | Director | ReleaseDate | Website | Language |
|---|---|---|---|---|---|
| **film1** | Ocean's 11 | S. Soderbergh | 3/4/01 | www.oceans11.com | --- |
| **film2** | Ocean's 12 | S. Soderbergh<br>R. Howard | 2/5/04 | www.oceans12.com | english |
| **film4** | The descendants | A. Payne | 15/9/11 | --- | english |
| **film5** | Bourne Identity | D. Liman | 12/6/12 | --- | english |

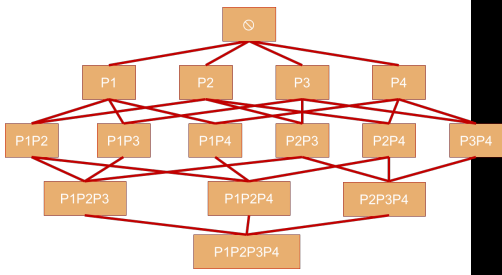
*Tool available in https://github.com/AKSW/rocker*

# KEY DISCOVERY APPROACHES

- **SF-Keys**
  - Keys and Pseudo-Keys Detection for Web Datasets Cleaning and Interlinking
- **F-Keys**
  - ROCKER: A Refinement Operator for Key Discovery
- **S-Keys**
  - **An automatic key discovery approach for data linking**
  - SAKey: Scalable almost key discovery in RDF data
  - VICKEY: Conditional key discovery
  - Linkkey: Data interlinking through robust Linkkey extraction

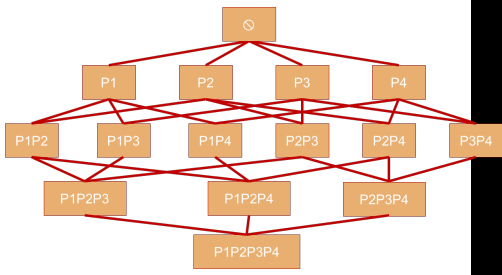
# KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - ➢ need of efficient filtering and prunings



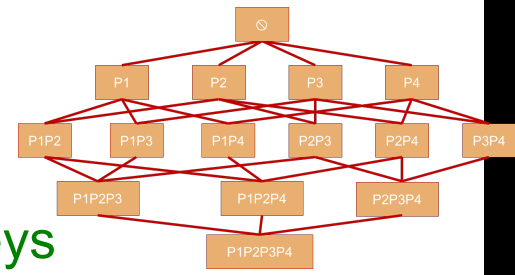
# KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - need of efficient filtering and prunings
- For each combination scan all the instances



# KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - need of efficient filtering and prunings
- For each combination scan all the instances
  - maximal non-keys → (derive) → minimal keys




| | FirstName | LastName | Phone | Profession |
|---|---|---|---|---|
| **Person1** | Anne | Tompson | 0169154259 | Actor, Director |
| **Person2** | Marie | Tompson | 0169154226 | Actor |
| **Person3** | Marie | David | 0425154012 | Actor |
| **Person4** | Vincent | Solgar | 0425154009 | Actor, Director |
| **Person5** | Simon | Roche | 0321455823 | Teacher |
| **Person6** | Jane | Ser | 0425462914 | Teacher, Researcher |
| **Person7** | Sara | Khan | 0425462915 | Teacher |
| **Person8** | Theo | Martin | 0321455823 | Teacher, Researcher |
| **Person9** | Marc | Blanc | 0169154228 | Teacher |

*Is [LastName] a non-key?* ➔ scan only a part of the data

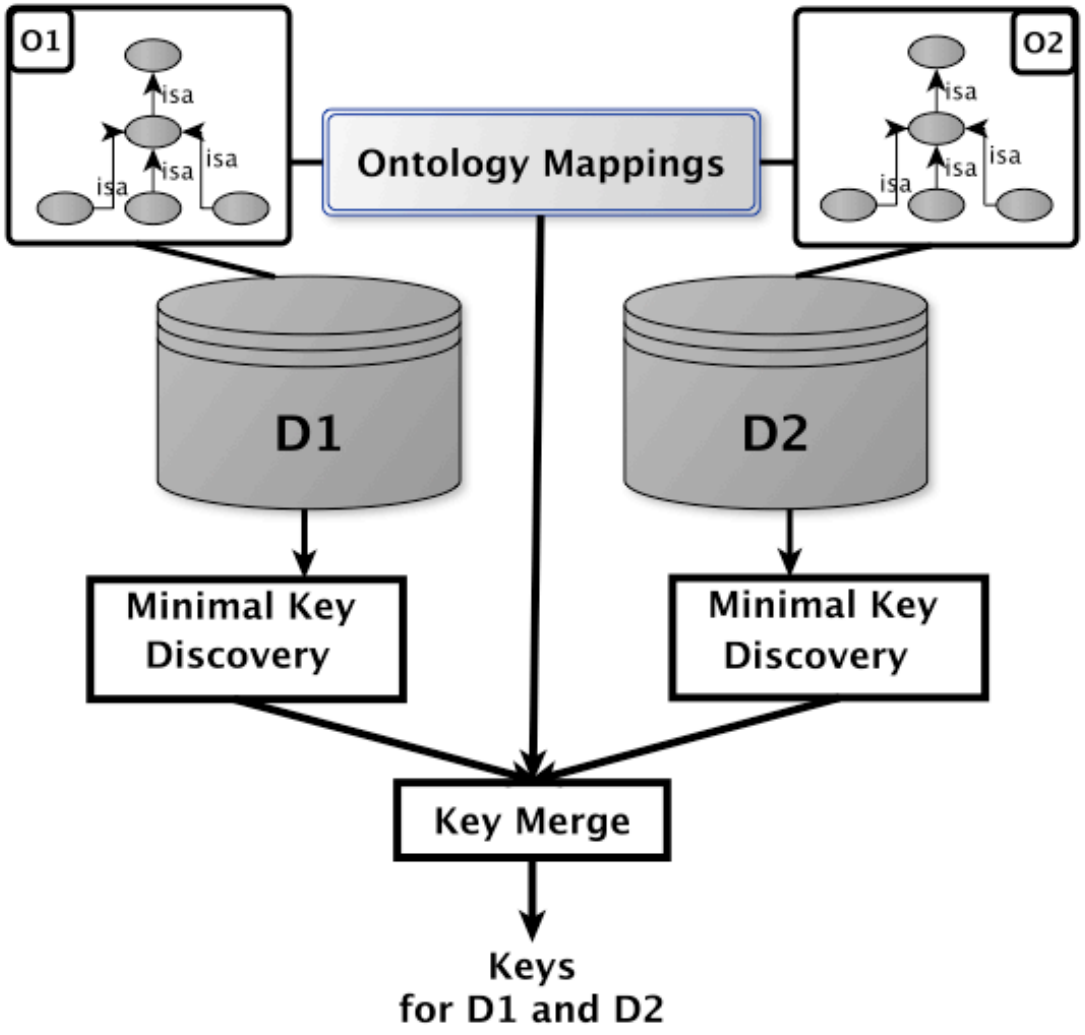
KD2R

SAKEY

VICKEY

# AN AUTOMATIC KEY DISCOVERY APPROACH FOR DATA LINKING (KD2R)

[Pernelle et al. 2013]



# SAKEY: SCALABLE ALMOST KEY DISCOVERY IN RDF DATA

[Symeonidou et al.14]

- **SAKey: Scalable Almost Key discovery approach for:**
  - Incomplete and erroneous data
  - Large datasets

- **Discovers *almost keys***
  - Sets of properties that are not keys due to *n* exceptions

# SAKEY: SCALABLE ALMOST KEY DISCOVERY IN RDF DATA

[Symeonidou et al.14]

- **SAKey: Scalable Almost Key discovery approach for:**
  - Incomplete and erroneous data
  - Large datasets

- **Discovers *almost keys***
  - Sets of properties that are not keys due to *n* exceptions

| | Region | Producer | Colour |
|---|---|---|---|
| **Wine1** | Bordeaux | Dupont | White |
| **Wine2** | Bordeaux | Baudin | Rose |
| **Wine3** | Languedoc | Dupont | Red |
| **Wine4** | Languedoc | Faure | Red |

- Examples of keys
  **{Region, Producer}: 0-almost key**

# SAKEY: SCALABLE ALMOST KEY DISCOVERY IN RDF DATA

[Symeonidou et al.14]

- **SAKey: Scalable Almost Key discovery approach for:**
  - Incomplete and erroneous data
  - Large datasets

- **Discovers *almost keys***
  - Sets of properties that are not keys due to *n* exceptions

| | **Region** | **Producer** | **Colour** |
|---|---|---|---|
| **Wine1** | Bordeaux | **Dupont** | White |
| **Wine2** | Bordeaux | Baudin | Rose |
| **Wine3** | Languedoc | **Dupont** | Red |
| **Wine4** | Languedoc | Faure | Red |

- Examples of keys

  **{Region, Producer}: 0-almost key**

  **{Producer}: 2-almost key**

# SAKEY: SCALABLE ALMOST KEY DISCOVERY IN RDF DATA

[Symeonidou et al.14]

- **Non-key discovery first**
  - Set of properties that is not a key

| | museumName | museumAddress | inCountry |
|---|---|---|---|
| **Museum1** | Archaeological Museum | 44 Patission Street | Greece |
| **Museum2** | Pompidou | - | France |
| **Museum3** | Musée d'Orsay | 62, rue de Lille | France |
| **Museum4** | Madame Tussauds | Marylebone Road | England |
| **Museum5** | Vatican Museums | Piazza San Giovanni | Italy |
| **Museum6** | Deutsches Museum | Museumsinsel 1 | Germany |
| **Museum7** | Olympia Museum | Archea Olympia | Greece |
| **Museum8** | Dalí museum | 1, Dali Boulevard | Spain |

# SAKEY: SCALABLE ALMOST KEY DISCOVERY IN RDF DATA

**[Symeonidou et al.14]**

- **Non-key discovery first**
  - Set of properties that is not a key

key

| | **museumName** | **museumAddress** | **inCountry** |
|---|---|---|---|
| **Museum1** | Archaeological Museum | 44 Patission Street | Greece |
| **Museum2** | Pompidou | - | France |
| **Museum3** | Musée d'Orsay | 62, rue de Lille | France |
| **Museum4** | Madame Tussauds | Marylebone Road | England |
| **Museum5** | Vatican Museums | Piazza San Giovanni | Italy |
| **Museum6** | Deutsches Museum | Museumsinsel 1 | Germany |
| **Museum7** | Olympia Museum | Archea Olympia | Greece |
| **Museum8** | Dalí museum | 1, Dali Boulevard | Spain |

# SAKEY: SCALABLE ALMOST KEY DISCOVERY IN RDF DATA

**[Symeonidou et al.14]**

- **Non-key discovery first**
  - Set of properties that is not a key

key | Non-key

| | museumName | museumAddress | inCountry |
|---|---|---|---|
| **Museum1** | Archaeological Museum | 44 Patission Street | Greece |
| **Museum2** | Pompidou | - | **France** |
| **Museum3** | Musée d'Orsay | 62, rue de Lille | **France** |
| **Museum4** | Madame Tussauds | Marylebone Road | England |
| **Museum5** | Vatican Museums | Piazza San Giovanni | Italy |
| **Museum6** | Deutsches Museum | Museumsinsel 1 | Germany |
| **Museum7** | Olympia Museum | Archea Olympia | Greece |
| **Museum8** | Dalí museum | 1, Dali Boulevard | Spain |

# N-ALMOST KEYS

[Symeonidou et al. 14]

**Exception of a key: an instance that shares values with another instance for a given set of properties *P***

| Films | HasName | HasActor | HasDirector | ReleaseDate | HasWebsite | HasLanguage |
|---|---|---|---|---|---|---|
| **f1** | "Ocean's 11" | "B. Pitt"<br>"J. Roberts" | "S. Soderbergh" | "3/4/01" | www.oceans11.com | --- |
| **f2** | "Ocean's 12" | "B. Pitt"<br>"G. Clooney"<br>"J. Roberts" | "S. Soderbergh"<br>"R. Howard" | "2/5/04" | www.oceans12.com | --- |
| **f3** | "Ocean's 13" | "B. Pitt"<br>"G. Clooney" | "S. Soderbergh"<br>"R. Howard" | "30/6/07" | www.oceans13.com | --- |
| **f4** | "The descendants" | "N. Krause"<br>"G. Clooney" | "A. Payne" | "15/9/11" | www.descendants.com | "english" |
| **f5** | "Bourne Identity" | "D. Liman" | --- | "12/6/12" | www.bourneIdentity.com | "english" |
| **f6** | "Ocean's 12" | --- | "R. Howard" | "2/5/04" | --- | --- |

# N-ALMOST KEYS

[Symeonidou et al. 14]

**Exception of a key:** an instance that shares values with another instance for a given set of properties *P*

- *f1, f2 and f3* are three exceptions for the property set {HasActor}

| Films | HasName | HasActor | HasDirector | ReleaseDate | HasWebsite | HasLanguage |
|---|---|---|---|---|---|---|
| **f1** | "Ocean's 11" | **"B. Pitt"**<br>"J. Roberts" | "S. Soderbergh" | "3/4/01" | www.oceans11.com | --- |
| **f2** | "Ocean's 12" | **"B. Pitt"**<br>"G. Clooney"<br>"J. Roberts" | "S. Soderbergh"<br>"R. Howard" | "2/5/04" | www.oceans12.com | --- |
| **f3** | "Ocean's 13" | **"B. Pitt"**<br>"G. Clooney" | "S. Soderbergh"<br>"R. Howard" | "30/6/07" | www.oceans13.com | --- |
| **f4** | "The descendants" | "N. Krause"<br>"G. Clooney" | "A. Payne" | "15/9/11" | www.descendants.com | "english" |
| **f5** | "Bourne Identity" | "D. Liman" | --- | "12/6/12" | www.bourneIdentity.com | "english" |
| **f6** | "Ocean's 12" | --- | "R. Howard" | "2/5/04" | --- | --- |

# N-ALMOST KEYS

[Symeonidou et al. 14]

**Exception of a key:** an instance that shares values with another instance for a given set of properties *P*

- *f1, f2 and f3* are three exceptions for the property set {HasActor}

**Exception Set $E_P$: set of exceptions for *P***

- $E_P$ = {f1, f2, f3} ∪ {f2, f3, f4} = {f1, f2, f3, f4} for {HasActor}

| Films | HasName | HasActor | HasDirector | ReleaseDate | HasWebsite | HasLanguage |
|---|---|---|---|---|---|---|
| **f1** | "Ocean's 11" | **"B. Pitt"** "J. Roberts" | "S. Soderbergh" | "3/4/01" | www.oceans11.com | --- |
| **f2** | "Ocean's 12" | **"B. Pitt"** **"G. Clooney"** "J. Roberts" | "S. Soderbergh" "R. Howard" | "2/5/04" | www.oceans12.com | --- |
| **f3** | "Ocean's 13" | **"B. Pitt"** **"G. Clooney"** | "S. Soderbergh" "R. Howard" | "30/6/07" | www.oceans13.com | --- |
| **f4** | "The descendants" | "N. Krause" **"G. Clooney"** | "A. Payne" | "15/9/11" | www.descendants.com | "english" |
| **f5** | "Bourne Identity" | "D. Liman" | --- | "12/6/12" | www.bourneIdentity.com | "english" |
| **f6** | "Ocean's 12" | --- | "R. Howard" | "2/5/04" | --- | --- |

# N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

- {HasActor} is a 4-almost key

# N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

- {HasActor} is a 4-almost key

***n*-non key: a set of properties where $|E_P| \geq n$**

- Using all the maximal *n*-non keys we can derive all the minimal (*n*-1)-almost keys

# N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

- {HasActor} is a 4-almost key

***n*-non key: a set of properties where $|E_P| \geq n$**

- Using all the maximal *n*-non keys we can derive all the minimal (*n*-1)-almost keys

All combinations of properties

# N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

- {HasActor} is a 4-almost key

***n*-non key: a set of properties where $|E_P| \geq n$**

- Using all the maximal *n*-non keys we can derive all the minimal (*n*-1)-almost keys


All combinations of properties

All sets of properties that contain at least 5 exceptions

# N-ALMOST KEYS

[Symeonidou et al. 14]

***n*-almost key: a set of properties where $|E_P| \leq n$**

- {HasActor} is a 4-almost key

***n*-non key: a set of properties where $|E_P| \geq n$**

- Using all the maximal *n*-non keys we can derive all the minimal (*n*-1)-almost keys

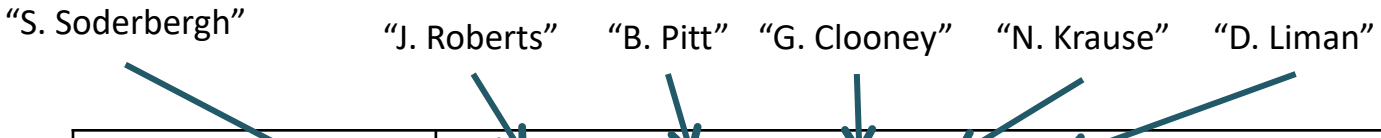

All combinations of properties

All sets of properties that contain at least 5 exceptions

All sets of properties that contain less than 5 exceptions

# N-NON KEY DISCOVERY: INITIAL MAP

[Symeonidou et al. 14]

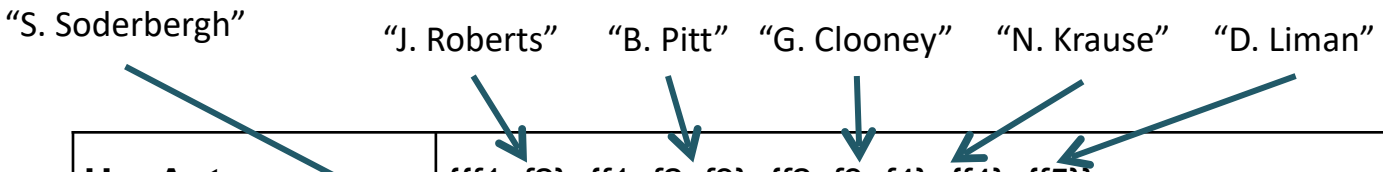



| | |
|---|---|
| **HasActor** | {{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}} |
| **HasDirector** | {{f1,f2,f3}, {f2, f3, f6}, {f4}} |
| **ReleaseDate** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasName** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasLanguage** | {{f4, f5}} |
| **HasWebsite** | {{f1}, {f2}, {f3}, {f4}, {f5}, {f6}} |

# N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

**Singleton filtering**

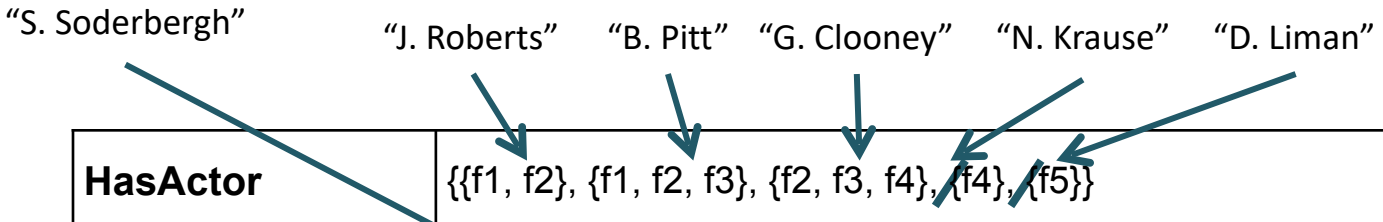



| | |
|---|---|
| **HasActor** | {{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}} |
| **HasDirector** | {{f1,f2,f3}, {f2, f3, f6}, {f4}} |
| **ReleaseDate** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasName** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasLanguage** | {{f4, f5}} |
| **HasWebsite** | {{f1}, {f2}, {f3}, {f4}, {f5}, {f6}} |

# N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

**Singleton filtering**

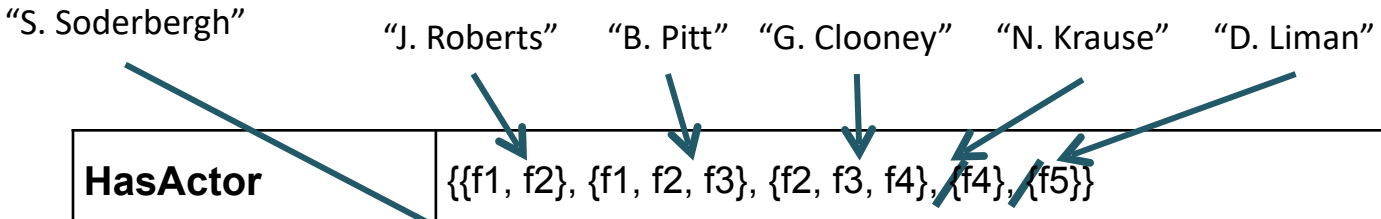



| | |
|---|---|
| **HasActor** | {{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}} |
| **HasDirector** | {{f1,f2,f3}, {f2, f3, f6}, {f4}} |
| **ReleaseDate** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasName** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasLanguage** | {{f4, f5}} |
| **HasWebsite** | {{f1}, {f2}, {f3}, {f4}, {f5}, {f6}} |

# N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

**Singleton filtering**




| | |
|---|---|
| **HasActor** | {{f1, f2}, {f1, f2, f3}, {f2, f3, f4}, {f4}, {f5}} |
| **HasDirector** | {{f1,f2,f3}, {f2, f3, f6}, {f4}} |
| **ReleaseDate** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasName** | {{f1}, {f2, f6}, {f3}, {f4}, {f5}} |
| **HasLanguage** | {{f4, f5}} |
| **HasWebsite** | {{f1}, {f2}, {f3}, {f4}, {f5}, {f6}} |

Single key

# N-NON KEY DISCOVERY: DATA FILTERING

[Symeonidou et al. 14]

**Singleton filtering**

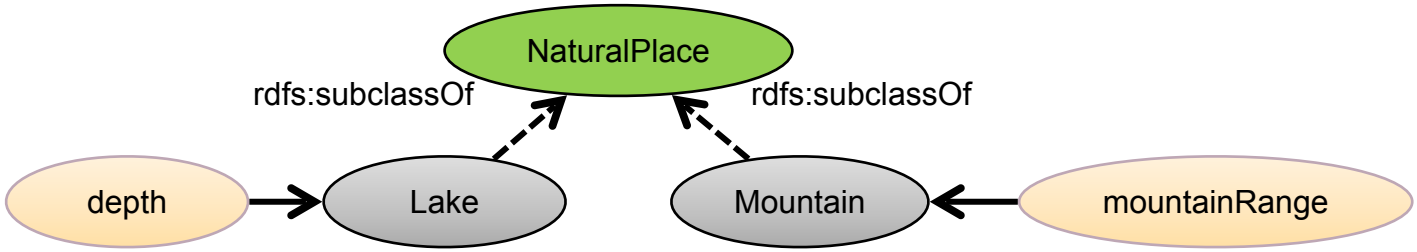
| | |
|---|---|
| **HasActor** | {{f1, f2, f3}, {f2, f3, f4}} |
| **HasDirector** | {{f1,f2,f3}, {f2, f3, f6}} |
| **ReleaseDate** | {{f2, f6}} |
| **HasName** | {{f2, f6}} |
| **HasLanguage** | {{f4, f5}} |

# N-NON KEY DISCOVERY: POTENTIAL N-NON KEYS

[Symeonidou et al. 14]

**Combinations of properties not needed be explored**

- Incomplete data
- Properties referring to different classes

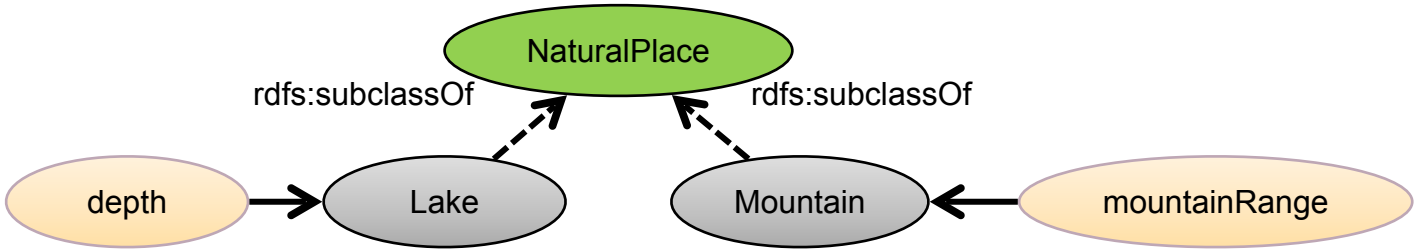


# N-NON KEY DISCOVERY: POTENTIAL N-NON KEYS

[Symeonidou et al. 14]

**Combinations of properties not needed be explored**

- Incomplete data
- Properties referring to different classes




- **Potential *n*-non keys**: Sets of properties that possibly refer to *n*-non keys

# N-NON KEY DISCOVERY

[Symeonidou et al. 14]

**HasActor**

- {f1, f2, f3} ∪ {f2, f3, f4} = {f1, f2, f3, f4} => 4-non key

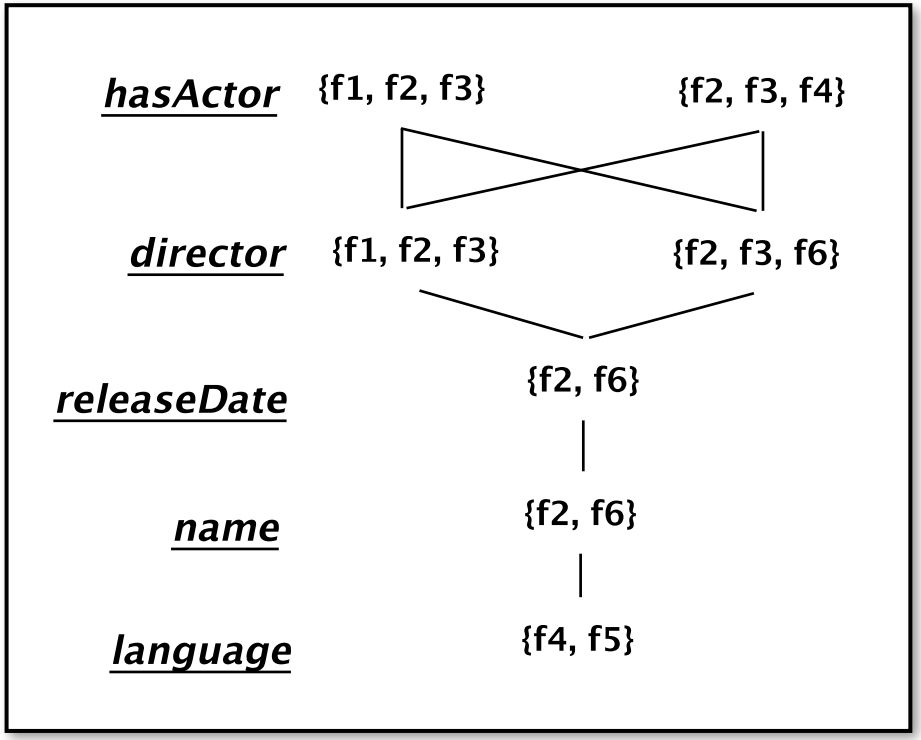
**Composite *n*-non keys**

- Intersections between sets of different properties

| | |
|---|---|
| **HasActor** | {{f1, f2, f3}, {f2, f3, f4}} |
| **HasDirector** | {{f1,f2,f3}, {f2, f3, f6}} |
| **ReleaseDate** | {{f2, f6}} |
| **HasName** | {{f2, f6}} |
| **HasLanguage** | {{f4, f5}} |

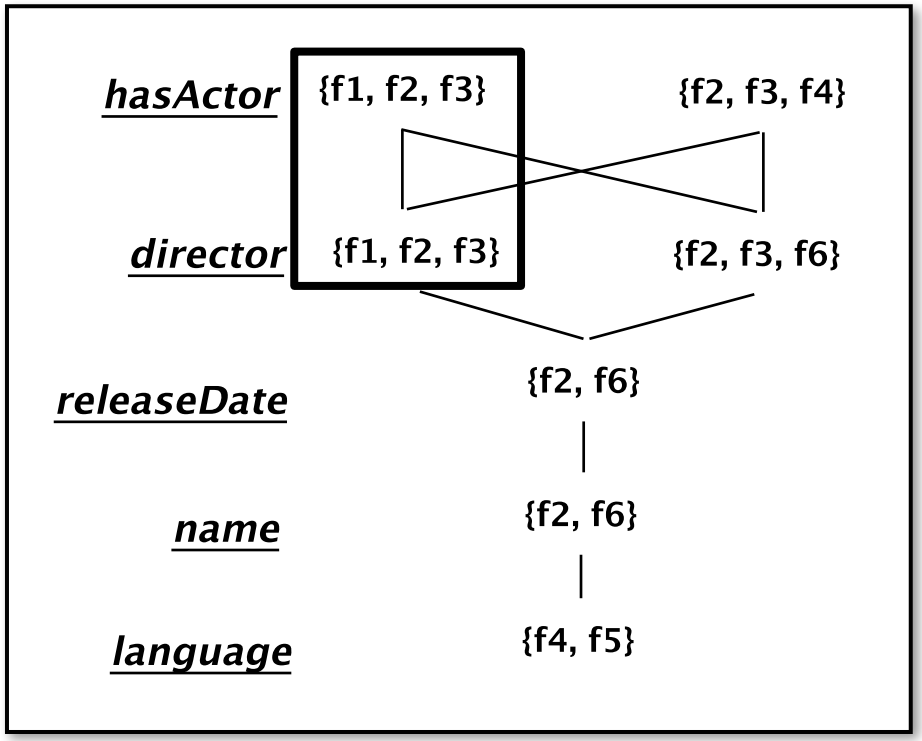
# N-NON KEY DISCOVERY

[Symeonidou et al. 14]



# N-NON KEY DISCOVERY

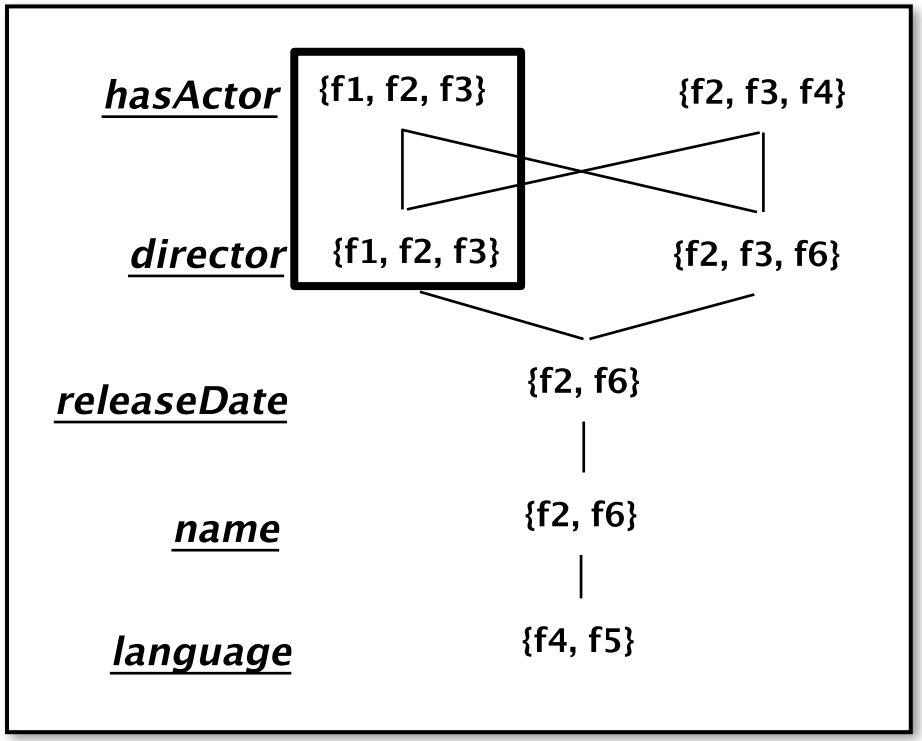
[Symeonidou et al. 14]



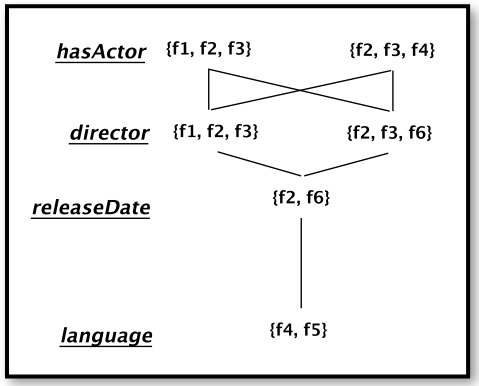

{hasActor, director} ➔ 3-non key

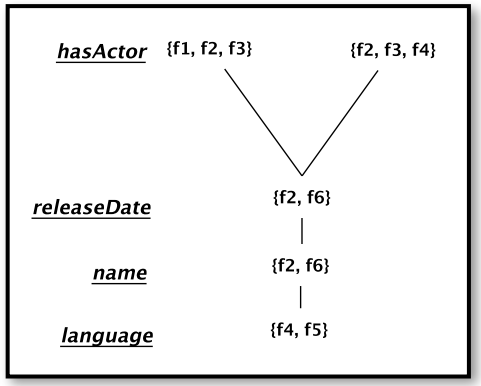
# N-NON KEY DISCOVERY

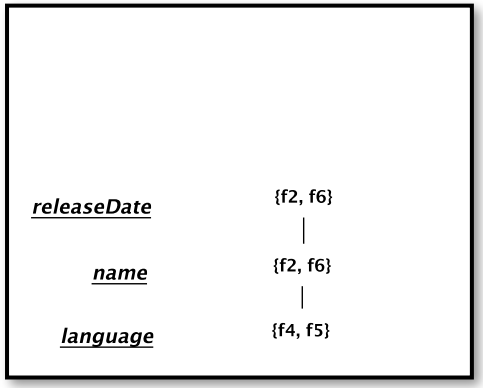
[Symeonidou et al. 14]




{hasActor, director} ➔ 3-non key



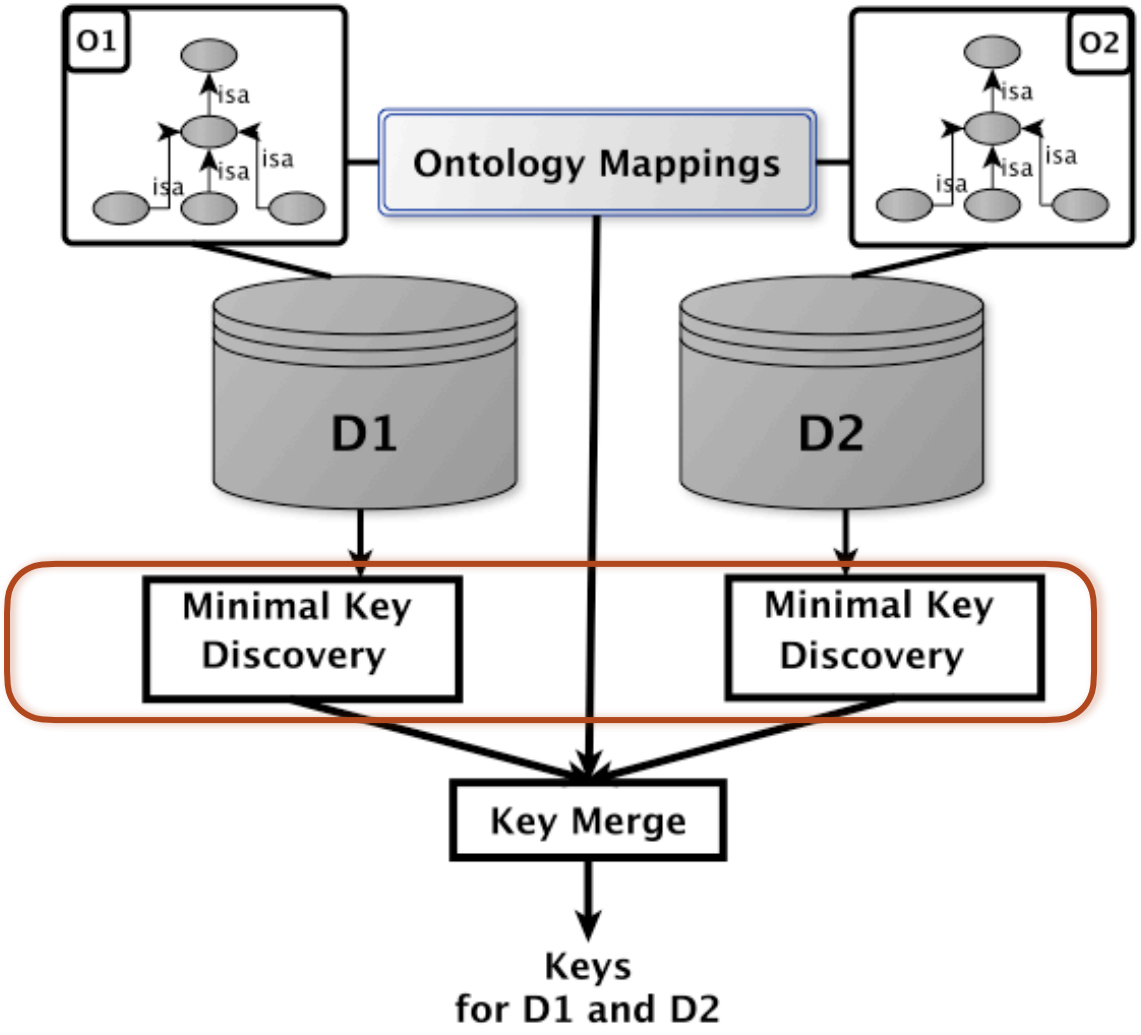




...

# AN AUTOMATIC KEY DISCOVERY APPROACH FOR DATA LINKING (KD2R)

[Pernelle et al. 2013]



# N-ALMOST KEY DISCOVERY

**Principle iterating two steps:**

(1) computing the Cartesian product of complement sets of the discovered non keys

(2) selecting only the minimal sets.

**Algorithm 2:** keyDerivation

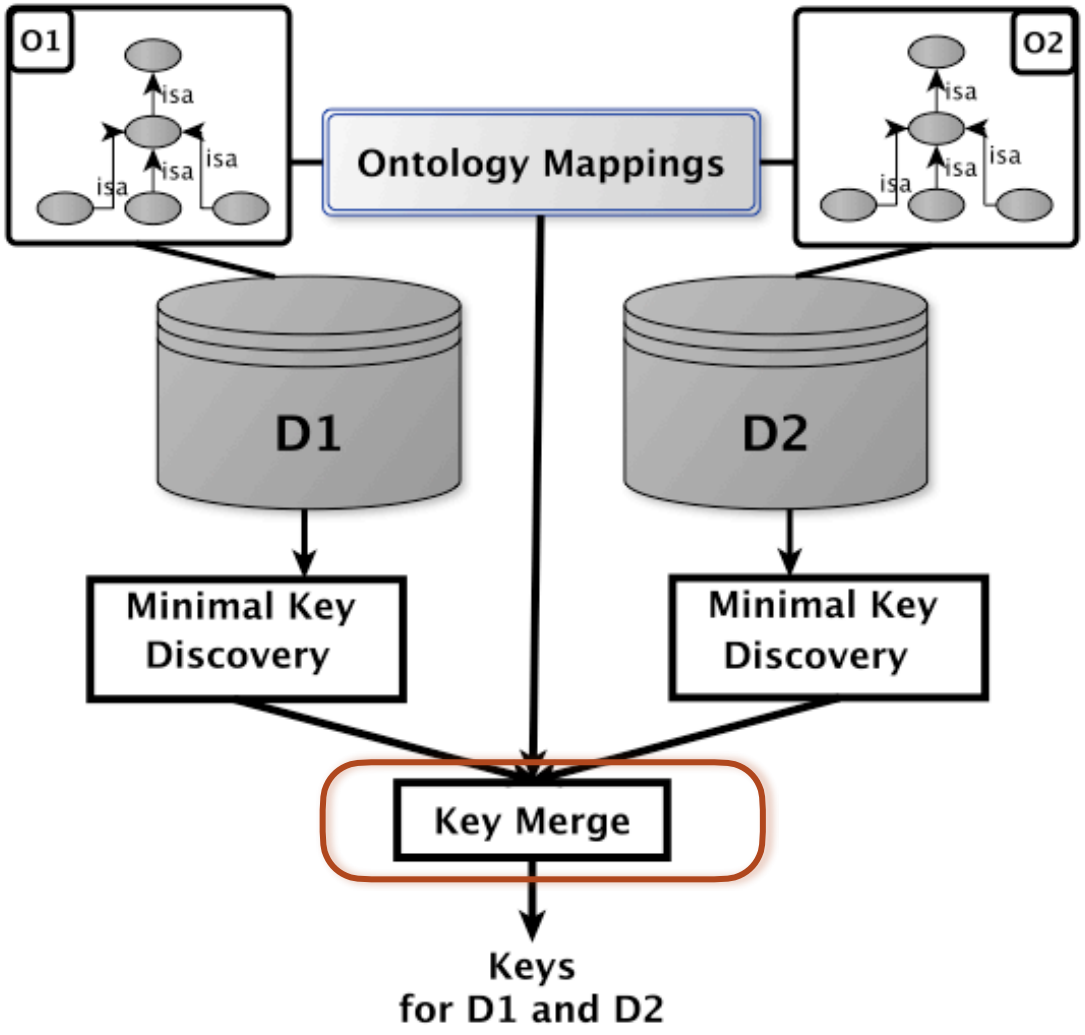
```
Input: compSets: set of complement sets
Output: KeySet: set of n-almost keys
1  KeySet ← ∅
2  orderedProperties = getOrderedProperties(compSets)
3  for each p_i ∈ orderedProperties do
4      selectedCompSets ← selectSets(p_i, compSets)
5      if (selectedCompSets == ∅) then
6          KeySet = KeySet ∪ {{p_i}}
7      else
8          KeySet = KeySet ∪ {p_i × keyDerivation(selectedCompSets)}
9      compSets = remove(compSets, p_i)
10     if ( ∃ set ∈ compSet s.t. set == ∅ ) then
11         break
12 return KeySet
```

**Optimization:**
**Ordering of the properties by frequency** in the complement sets.

At each iteration, **the most frequent property is selected**

# AN AUTOMATIC KEY DISCOVERY APPROACH FOR DATA LINKING (KD2R)

[Pernelle et al. 2013]



# KEY MERGE: SEVERAL DATASETS

[Symeonidou et al. 14]

**Goal: Keys valid in both datasets**

- More sure keys

**Intuition: Computation of Cartesian product of sets of keys**

- Keep only minimal keys

# KEY MERGE: SEVERAL DATASETS

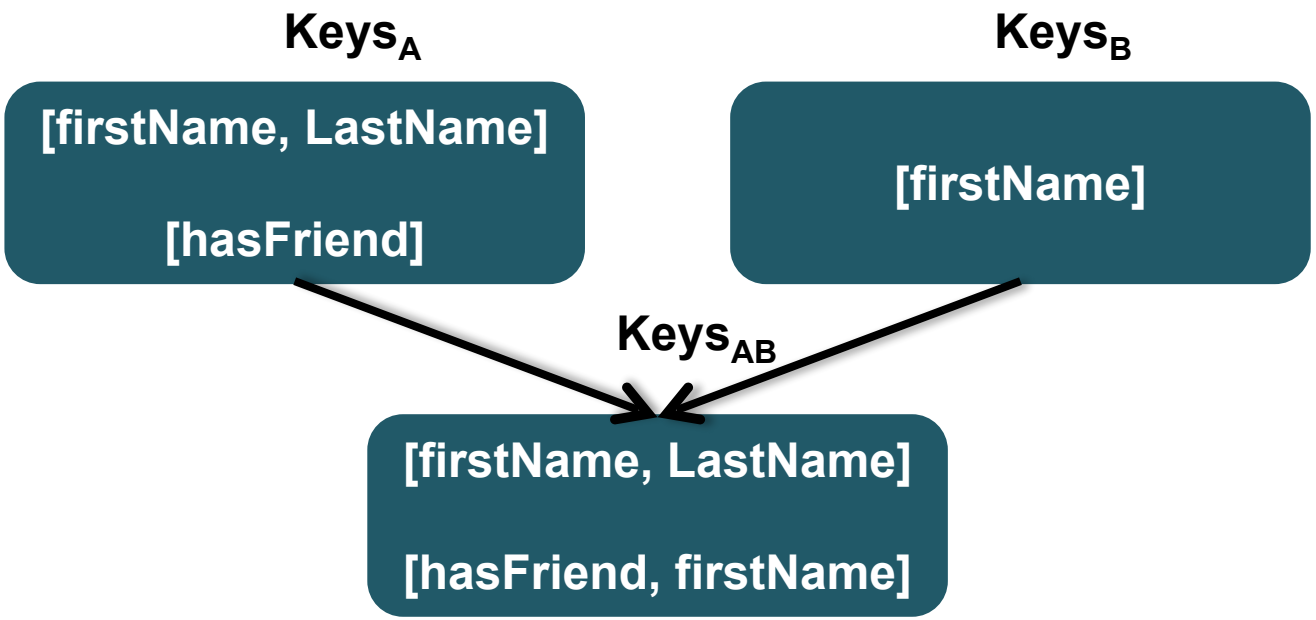
[Symeonidou et al. 14]

**Goal: Keys valid in both datasets**

- More sure keys

**Intuition: Computation of Cartesian product of sets of keys**

- Keep only minimal keys



# DATA LINKING USING ALMOST KEYS

[Symeonidou et al.14]

- **Goal: Compare linking results using almost keys with different *n***

- **Evaluation of linking using**
  - Recall
  - Precision
  - F-Measure

- **Datasets**
  - OAEI 2010
  - OAEI 2013

# EXAMPLE: DATA LINKING USING ALMOST KEYS

**[Symeonidou et al.14]**

- **OAEI 2013 - Person**

| | Almost keys | Recall | Precision | F-Measure |
|---|---|---|---|---|
| **0-almost key** | {BirthDate, award} | 9.3% | 100% | 17% |
| **2-almost key** | {BirthDate} | 32.5% | 98.6% | 49% |

| # exceptions | Recall | Precision | F-measure |
|---|---|---|---|
| **0, 1** | 25.6% | 100% | 41% |
| **2, 3** | 47.6% | 98.1% | 64.2% |
| **4, 5** | 47.9% | 96.3% | 63.9% |
| **6, ..., 16** | 48.1% | 96.3% | 64.1% |
| **17** | 49.3% | 82.8% | 61.8% |

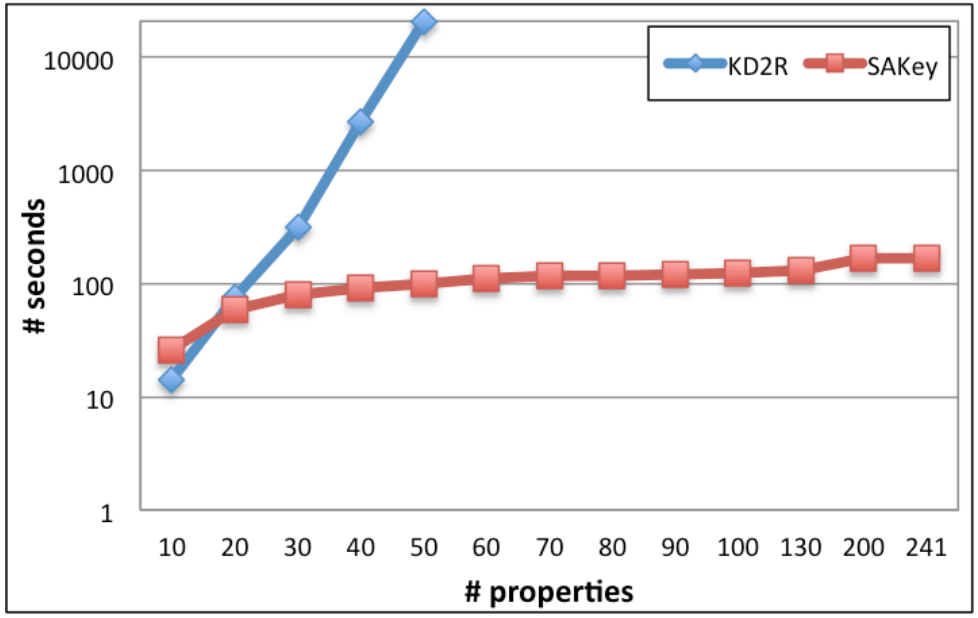
*Tool available in https://www.lri.fr/sakey*

# KD2R VS. SAKEY - NON KEY DISCOVERY

[Symeonidou et al. 14]

| Class | # triples | # Instances | #Properties | KD2R Runtime | SAKey Runtime (*n*=0) |
|---|---|---|---|---|---|
| **DB:Website** | 8506 | 2870 | 66 | 13min | 1s |
| **YA:Building** | 114783 | 54384 | 17 | 26s | 9s |
| **DB:BodyOfWater** | 1068428 | 34000 | 200 | outOfMem. | 37s |
| **DB:NaturalPlace** | 1604348 | 49913 | 243 | outOfMem. | 1min10s |

**Dbpedia class=**
**DB:NaturalPlace**

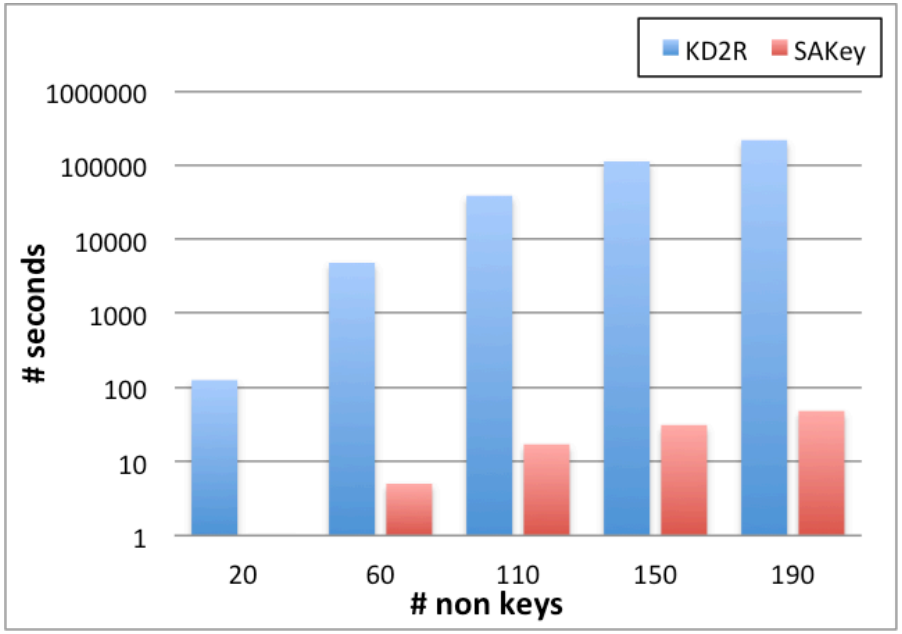


# KD2R VS. SAKEY - KEY DERIVATION

[Symeonidou et al. 14]

| Class | # non keys | # keys | KD2R | SAKey (*n*=0) |
|---|---|---|---|---|
| **DB:Lake** | 50 | 480 | 1min10s | 1s |
| **DB:Mountain** | 49 | 821 | 8min | 1s |
| **DB:BodyOfWater** | 220 | 3846 | > 1 day | 66s |
| **DB:NaturalPlace** | 302 | 7011 | > 2 days | 5min |

**Dbpedia class= DB:BodyOfWater**



# KEY ISSUES

- **Key limitations**
  - Cases of datasets having no keys
  - Keys are generic, i.e., they are true for every instance of a class in a dataset

- **Motivating example**
  - In many countries, a student can be supervised by multiple supervisors
  - In German Universities, a student can be supervised by only one professor
    - {supervises} key for the instances of the class Professor with condition that they are in a German university

- **Almost keys express keys that <u>do not uniquely identify</u> <u>every instance</u> of a class in a dataset**

Approximate keys do not express the conditions under which they are true

# KEY DISCOVERY APPROACHES

- **SF-Keys**
  - Keys and Pseudo-Keys Detection for Web Datasets Cleaning and Interlinking
- **F-Keys**
  - ROCKER: A Refinement Operator for Key Discovery
- **S-Keys**
  - An automatic key discovery approach for data linking
  - SAKey: Scalable almost key discovery in RDF data
  - **VICKEY: Conditional key discovery**
  - Linkkey: Data interlinking through robust Linkkey extraction

# VICKEY: MINING CONDITIONAL KEYS ON RDF DATASETS

**[Symeonidou et al.17]**

- **Conditional key:** a key, that is valid for the instances of a class satisfying a specific condition
  - **Condition part:** pairs of property and value
    - Eg. {Lab=INRA}, {Gender=Male}, {Gender=Female ^ Lab=INRA} etc.
  - **Key part:** a set of properties
    - Eg. {FirstName}, {LastName}, {FirstName, LastName} etc.

# VICKEY: MINING CONDITIONAL KEYS ON RDF DATASETS

**[Symeonidou et al.17]**

- **Conditional key:** a key, that is valid for the instances of a class satisfying a specific condition
  - **Condition part:** pairs of property and value
    - Eg. {Lab=INRA}, {Gender=Male}, {Gender=Female ^ Lab=INRA} etc.
  - **Key part:** a set of properties
    - Eg. {FirstName}, {LastName}, {FirstName, LastName} etc.

*Instances of the class Person*

| | FirstName | LastName | Gender | Lab | Nationality |
|---|---|---|---|---|---|
| instance1 | Claude | Dupont | Female | Paris-Sud | France |
| instance2 | Claude | Dupont | Male | Paris-Sud | Belgium |
| **instance3** | Juan | Rodríguez | Male | **INRA** | Spain, Italy |
| **instance4** | Juan | Salvez | Male | **INRA** | Spain |
| **instance5** | Anna | Georgiou | Female | **INRA** | Greece, France |
| instance6 | Pavlos | Markou | Male | Paris-Sud | Greece |
| **instance7** | Marie | Legendre | Female | **INRA** | France |

**{LastName}** is a *key* under the *condition* **{Lab=INRA}**

# CONDITIONAL KEY QUALITY MEASURES

**[Symeonidou et al.17]**

- **Support:** #instances both satisfying condition part and instantiating key part
- **Coverage:** Support/#all_Instances

# CONDITIONAL KEY QUALITY MEASURES

**[Symeonidou et al.17]**

- **Support:** #instances both satisfying condition part and instantiating key part
- **Coverage:** Support/#all_Instances

*Instances of the class Person*

| | FirstName | LastName | Gender | Lab | Nationality |
|---|---|---|---|---|---|
| instance1 | Claude | Dupont | Female | Paris-Sud | France |
| instance2 | Claude | Dupont | Male | Paris-Sud | Belgium |
| **instance3** | Juan | Rodríguez | Male | **INRA** | Spain, Italy |
| **instance4** | Juan | Salvez | Male | **INRA** | Spain |
| **instance5** | Anna | Georgiou | Female | **INRA** | Greece, France |
| instance6 | Pavlos | Markou | Male | Paris-Sud | Greece |
| **instance7** | Marie | Legendre | Female | **INRA** | France |

**Support:** 4
**Coverage:** 4/7 = 0.57

**{LastName}** is a *key* under the *condition* **{Lab=INRA}**

# VICKEY: MINING EFFICIENTLY CONDITIONAL KEYS

Symeonidou et al. 2017

**Goal of VICKEY**

- Given all instances of a class in a dataset, a min_support, a min_coverage, mine all minimal conditional keys with
  - support ≥ min_support
  - coverage ≥ min_coverage

**Exponential search space $(O(|V|^{|P|})$ with**

- V, the set of objects of a given class in the dataset
- P, the set of properties of a given class in the dataset

# VICKEY: MINING EFFICIENTLY CONDITIONAL KEYS

Symeonidou et al. 2017

**Goal of VICKEY**

- Given all instances of a class in a dataset, a min_support, a min_coverage, mine all minimal conditional keys with
  - support ≥ min_support
  - coverage ≥ min_coverage

**Exponential search space $(O(|V|^{|P|})$ with**

- V, the set of objects of a given class in the dataset
- P, the set of properties of a given class in the dataset

**Conditional keys can be efficiently discovered from the non-keys**

# VICKEY: MINING EFFICIENTLY CONDITIONAL KEYS

Symeonidou et al. 2017

**Non-key:** a set of properties where two instances share at least one value for each property in the non-key (SAKey)

*Instances of the class Person*

| | FirstName | LastName | Gender | Lab | Nationality |
|---|---|---|---|---|---|
| **instance1** | Claude | Dupont | Female | Paris-Sud | France |
| **instance2** | Claude | Dupont | Male | Paris-Sud | Belgium |
| **instance3** | **Juan** | Rodríguez | **Male** | **INRA** | **Spain**, Italy |
| **instance4** | **Juan** | Salvez | **Male** | **INRA** | **Spain** |
| **instance5** | Anna | Georgiou | Female | INRA | Greece, France |
| **instance6** | Pavlos | Markou | Male | Paris-Sud | Greece |
| **instance7** | Marie | Legendre | Female | INRA | France |

**Non-key**
**{FirstName, Gender, Lab, Nationality}**

# VICKEY: MINING EFFICIENTLY CONDITIONAL KEYS

Symeonidou et al. 2017

**Non-key:** a set of properties where two instances share at least one value for each property in the non-key (SAKey)

*Instances of the class Person*

| | FirstName | LastName | Gender | Lab | Nationality |
|---|---|---|---|---|---|
| **instance1** | Claude | Dupont | Female | Paris-Sud | France |
| **instance2** | Claude | Dupont | Male | Paris-Sud | Belgium |
| **instance3** | **Juan** | Rodríguez | **Male** | **INRA** | **Spain**, Italy |
| **instance4** | **Juan** | Salvez | **Male** | **INRA** | **Spain** |
| **instance5** | Anna | Georgiou | Female | INRA | Greece, France |
| **instance6** | Pavlos | Markou | Male | Paris-Sud | Greece |
| **instance7** | Marie | Legendre | Female | INRA | France |

**Non-key**
**{FirstName, Gender, Lab, Nationality}**

**Given a maximal non-key**

- **Condition part:** subset of properties of the non-key
- **Key part:** subset of the remaining properties of the non-key

# CONDITIONAL KEY GRAPH EXPLORATION

**[Symeonidou et al.17]**

- **Step 1: Discover all minimal conditional keys with condition {p=a}**
- Step 2: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge p_2=a_2\}$
- …
- Step n: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$

# CONDITIONAL KEY GRAPH EXPLORATION

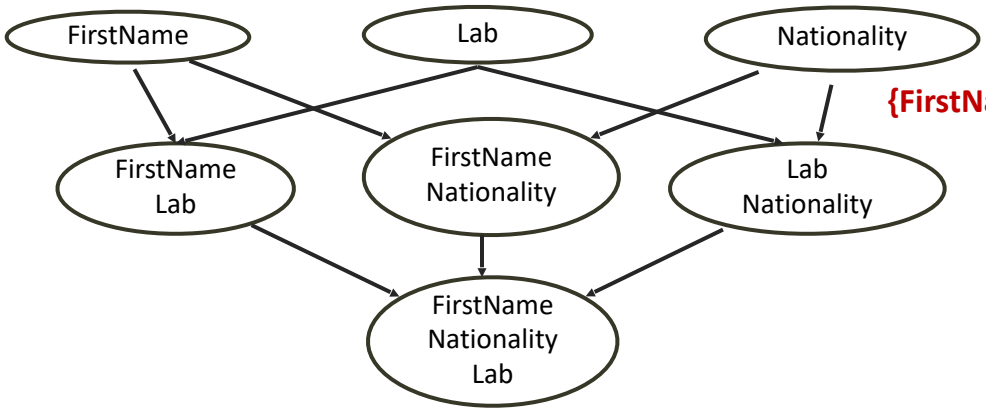
[Symeonidou et al.17]

- **Step 1: Discover all minimal conditional keys with condition {p=a}**
- Step 2: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge p_2=a_2\}$
- …
- Step n: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$

*Condition*

**Gender=Female**

*Conditional key graph*




**Non-key**

**{FirstName, Gender, Lab, Nationality}**

# CONDITIONAL KEY GRAPH EXPLORATION

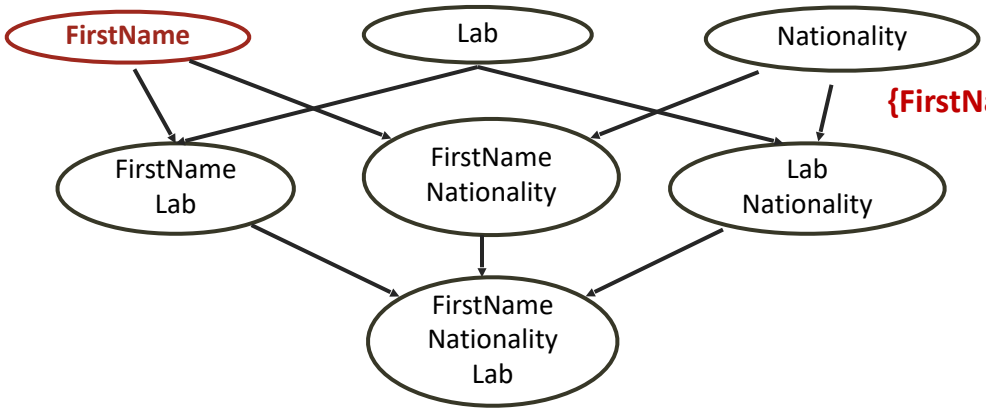
[Symeonidou et al.17]

- **Step 1: Discover all minimal conditional keys with condition {p=a}**
- Step 2: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge p_2=a_2\}$
- …
- Step n: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$

*Condition*

**Gender=Female**

*Conditional key graph*




**Non-key**
**{FirstName, Gender, Lab, Nationality}**

**Key {FirstName} for cond {Gender=Female}**

# CONDITIONAL KEY GRAPH EXPLORATION

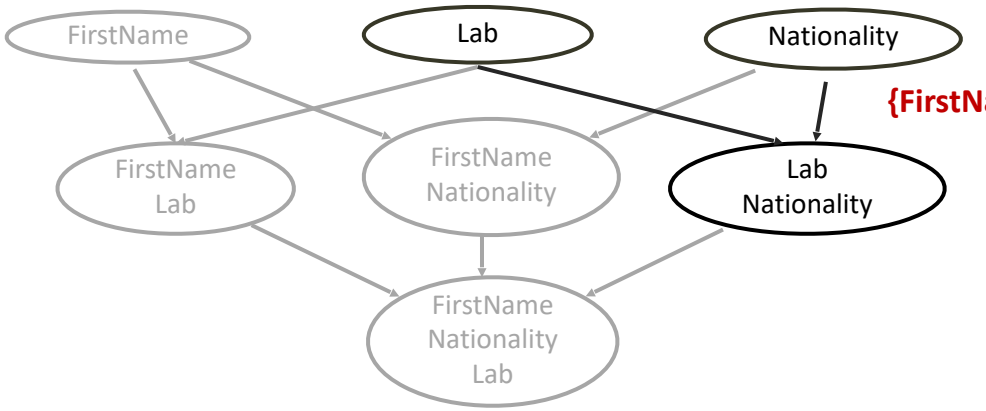
**[Symeonidou et al.17]**

- **Step 1: Discover all minimal conditional keys with condition {p=a}**
- Step 2: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge p_2=a_2\}$
- …
- Step n: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$

*Condition*

**Gender=Female**

*Conditional key graph*




**Non-key**
**{FirstName, Gender, Lab, Nationality}**

**Key {FirstName} for cond {Gender=Female}**

# CONDITIONAL KEY GRAPH EXPLORATION

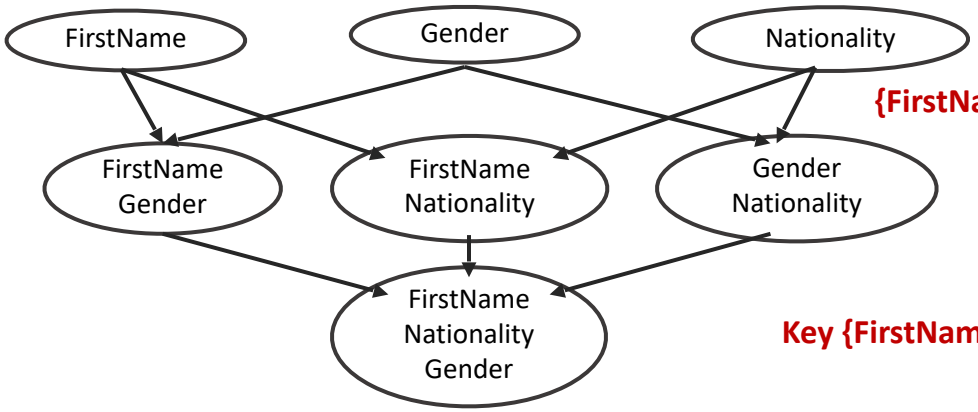
Symeonidou et al. 2017

- Given a non-key
  - **Step 1: Discover all minimal conditional keys with condition of size 1 {p=a}**
  - Step 2: Discover all minimal conditional keys with condition of size 2 {$p_1=a_1 \wedge p_2=a_2$}
  - …
  - Step n: Discover all minimal conditional keys with condition of size n {$p_1=a_1 \wedge \ldots \wedge p_n=a_n$}

*Condition*

**Lab=INRA**

*Conditional key graph*




**Non-key**
**{FirstName, Gender, Lab, Nationality}**

**Key {FirstName} for cond {Gender=Female}**

# CONDITIONAL KEY GRAPH EXPLORATION

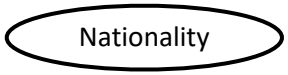
Symeonidou et al. 2017

- Given a non-key
  - Step 1: Discover all minimal conditional keys with condition of size 1 {p=a}
  - **Step 2: Discover all minimal conditional keys with condition of size 2 {$p_1=a_1 \wedge p_2=a_2$}**
  - …
  - Step n: Discover all minimal conditional keys with condition of size n {$p_1=a_1 \wedge \ldots \wedge p_n=a_n$}

***Condition***

**Gender=Female**
**^**
**Lab=INRA**

***Conditional key graph***



# EXPERIMENTAL EVALUATION

Symeonidou et al. 2017

**Evaluation of VICKEY's scalability in large datasets**

**Evaluation of conditional keys in data linking**

# VICKEY'S SCALABILITY

Symeonidou et al. 2017

**Run VICKEY in large datasets**

**Compare the runtime results with *AMIE - a generic rule mining approach adapted to mine conditional keys***

- Both approaches take as input non-keys mined by SAKey to reduce the search space of conditional key mining

**Coverage 1%**

# RUNTIME RESULTS - VICKEY VS. AMIE[2]

Symeonidou et al. 2017

| Class* | #Triples | #Instances | #Properties | #NonKeys | VICKEY | AMIE | #ConditionalKeys |
|---|---|---|---|---|---|---|---|
| Actor | 57.2k | 5.8k | 71 | 137 | 4.52m | 12.58h | 311 |
| Album | 786.1k | 85.3k | 39 | 68 | 1.53h | 3.90h | 304 |
| Book | 258.4k | 30.0k | 51 | 95 | 11.84h | >1d | 419 |
| Film | 832.1k | 82.6k | 74 | 132 | 1.37h | 3.64h | 185 |
| Mountain | 127.8k | 16.4k | 58 | 47 | 2.86m | 23.57m | 257 |
| Museum | 12.9k | 1.9k | 65 | 17 | 1.46s | 6.45s | 58 |
| Organization | 1.82M | 178.7k | 553 | 3221 | 26.32h | > 36h | 28 |
| Scientist | 258.5k | 19.7k | 73 | 309 | 27.67m | > 1d | 582 |
| University | 85.5k | 8.7k | 89 | 140 | 14.45h | >1d | 941 |

*All used classes are obtained from DBpedia*

# DATA LINKING - VICKEY VS. SAKEY[1]

Symeonidou et al. 2017

**Goal: link two datasets using**

- Classical keys discovered by SAKey[1]
- Conditional keys discovered by VICKEY
    - Supervises(Prof1, stud1)^Supervises(Prof2, stud1)^teachesIn(Prof1, “Gemany”)^teachesIn(Prof2, “Gemany”)=> sameAs(Prof1, Prof2)
- Both classical keys and conditional keys

**Evaluate obtained links using the existing goal-standard with**

- Recall
- Precision
- F-Measure

# DATA LINKING - VICKEY VS. SAKEY

[Symeonidou et al.17]

- **Goal: link two datasets using**
  - Classical keys discovered by SAKey
  - Conditional keys discovered by VICKEY
  - Both classical keys and conditional keys

- **Evaluate obtained links using the existing goal-standard with**
  - Recall
  - Precision
  - F-Measure

# DATA LINKING - VICKEY VS. SAKEY

[Symeonidou et al.17]

- **Link two knowledge bases containing information of Wikipedia**
  - Yago[3]
  - DBpedia[4]

- **Used classes of DBpedia and Yago**
  - Actor
  - Album
  - Book
  - Film
  - Mountain
  - Museum
  - Organization
  - Scientist
  - University

# DATA LINKING - VICKEY VS. SAKEY

**[Symeonidou et al.17]**

| Class | | Recall | Precision | F-Measure | |
|---|---|---|---|---|---|
| **Actor** | Keys[1]* | 0.27 | 0.99 | **0.43** | |
| | Conditional keys** | 0.57 | 0.99 | 0.73 | **x 1.75** |
| | **Keys[1]+Conditional keys** | 0.6 | 0.99 | **0.75** | |
| **Album** | Keys[1] | 0 | 1 | **0.00** | |
| | Conditional keys | 0.15 | 0.99 | 0.26 | **x 869** |
| | **Keys[1]+Conditional keys** | 0.15 | 0.99 | **0.26** | |
| **Film** | Keys[1] | 0.04 | 0.99 | **0.08** | |
| | Conditional keys | 0.38 | 0.96 | 0.54 | **x 7.1** |
| | **Keys[1]+Conditional keys** | 0.39 | 0.98 | **0.55** | |
| **Museum** | Keys[1] | 0.12 | 1 | **0.21** | |
| | Conditional keys | 0.25 | 1 | 0.40 | **x 2.19** |
| | **Keys[1]+Conditional keys** | 0.31 | 1 | **0.47** | |

**Keys[1] from SAKey*

***Conditional keys from VICKEY*

*Tool available* [https://github.com/lgalarra/vickey](https://github.com/lgalarra/vickey)

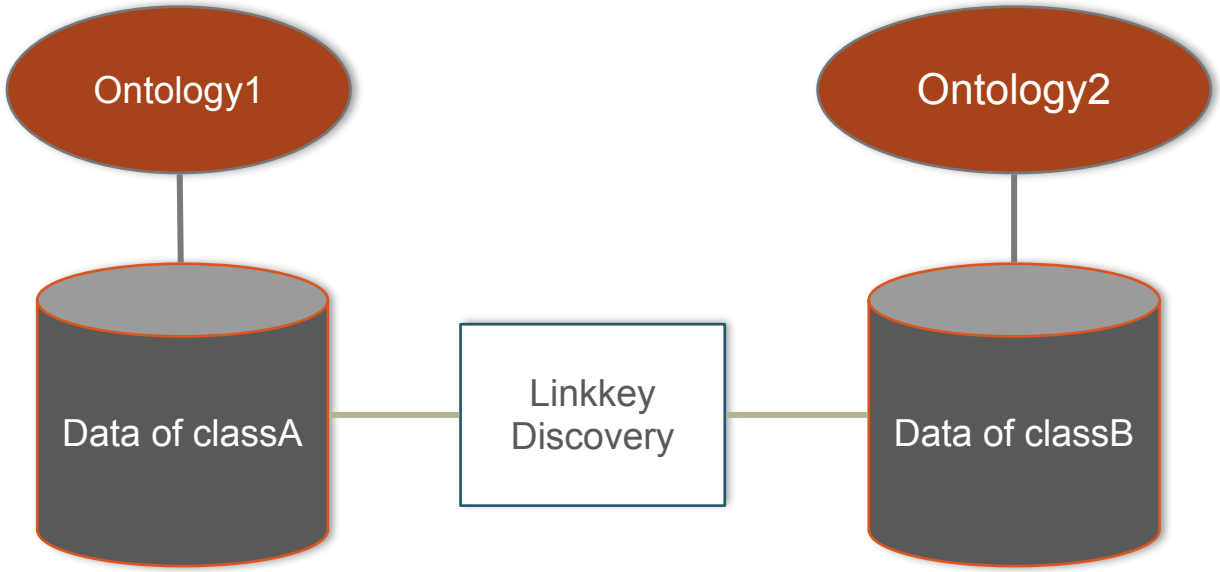
# KEY DISCOVERY APPROACHES

- **SF-Keys**
  - Keys and Pseudo-Keys Detection for Web Datasets Cleaning and Interlinking
- **F-Keys**
  - ROCKER: A Refinement Operator for Key Discovery
- **S-Keys**
  - An automatic key discovery approach for data linking
  - SAKey: Scalable almost key discovery in RDF data
  - VICKEY: Conditional key discovery
  - **Linkkey: Data interlinking through robust Linkkey extraction**

# DATA INTERLINKING THROUGH ROBUST LINKKEY EXTRACTION

[Atencia et al.14]

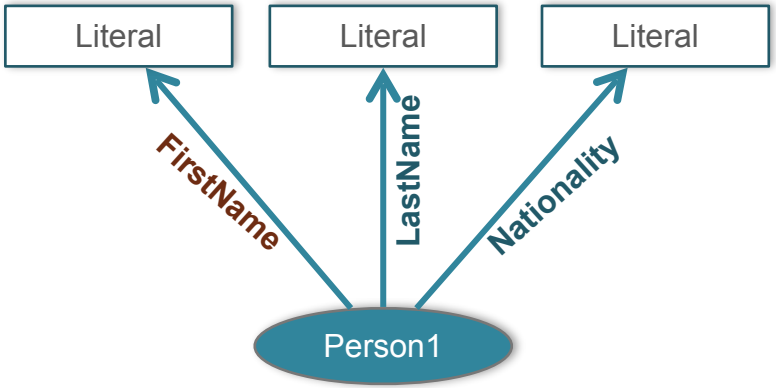
- **Given a pair of classes in two datasets conforming to two different ontologies:**
  - Discover ***Linkkeys*** – maximal sets of property pairs that can link instances of two different classes

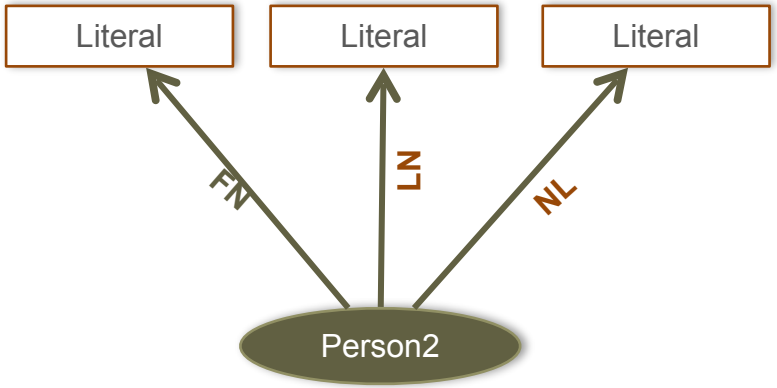


# DATA INTERLINKING THROUGH ROBUST LINKKEY EXTRACTION

[Atencia et al.14]

- **Given a pair of classes in two datasets conforming to two different ontologies:**
  - Discover ***Linkkeys*** – maximal sets of property pairs that can link instances of two different classes

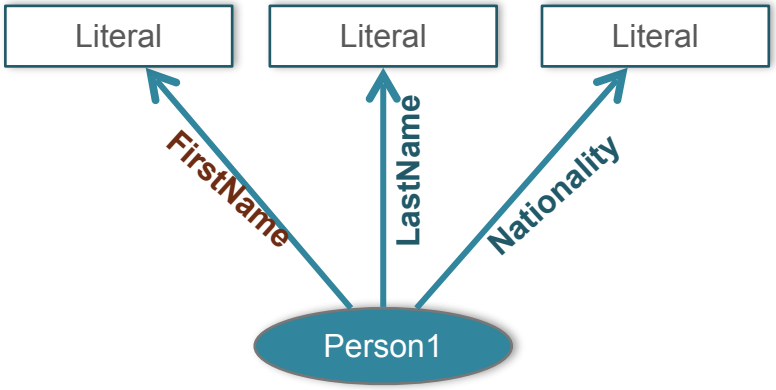


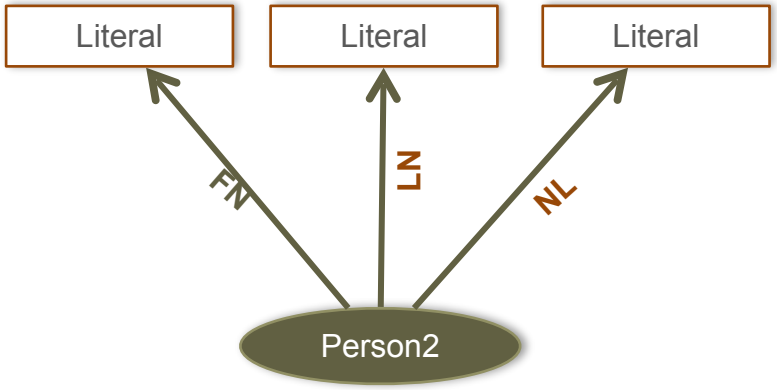



# DATA INTERLINKING THROUGH ROBUST LINKKEY EXTRACTION

[Atencia et al.14]

- **Given a pair of classes in two datasets conforming to two different ontologies:**
  - Discover ***Linkkeys*** – maximal sets of property pairs that can link instances of two different classes




| | LastName | Nationality | Profession |
|---|---|---|---|
| **P11** | Tompson | Greek | |
| **P12** | Dupont | French | Researcher |

| | LN | NL | PR |
|---|---|---|---|
| **P21** | Tompson | Greek | |
| **P22** | Tompson | Greek, French | Reseacher |

Ex. {<LastName,LN>,<Nationality,NL>}, {<Profession,PR>,<Nationality,NL>}

# SUMMARY

- **S-keys, F-Keys, SF-keys**
  - F-keys and SF-keys would better work in more complete data
  - S-keys are more resistant to incompleteness
- **Keys and almost/pseudo keys:**
  - Better linking results in terms of recall with n-almost keys/pseudo keys than with sure keys
- **Conditional keys:**
  - Better linking results in terms of recall when conditional keys are used

- **Improvements:**
  - Define the number of exceptions
  - Handle numerical values (one approach already exists)
  - Handle data heterogeneity in a dataset
  - Choose right semantic using the data (data completeness)
  - …

# REFERENCES


[Atencia et al. 2012] Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking.Manuel Atencia, Jérôme David, François Scharffe. In EKAW 2012

[Atencia et al. 2014] Data interlinking through robust Linkkey extraction. Atencia, Manuel, Jérôme David, and Jérôme Euzenat. ECAI, 2014.

[Soru et al. 2015] ROCKER: a refinement operator for key discovery. Soru, Tommaso, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. In WWW, 2015.

[Pernelle et al. 2013] An Automatic Key Discovery Approach for Data Linking. Nathalie Pernelle, Fatiha Saïs. and Danai Symeounidou. In Journal of Web Semantics

[Symeonidou et al. 2014] SAKey: Scalable almost key discovery in RDF data. Symeonidou, Danai, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. In ISWC 2014.

[Symeonidou et al. 2017] VICKEY: Mining Conditional Keys on RDF datasets. Danai Symeonidou, Luis Galarraga, Nathalie Pernelle, Fatiha Saïs and Fabian Suchanek. In ISWC 2017.