# KNOWLEDGE GRAPH COMPLETION
# PART 4: KEY DISCOVERY


FATIHA SAÏS [1]

NATHALIE PERNELLE[1]

**DANAI SYMEONIDOU[2]**

[1] LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

[2] INRA, GAMMA TEAM

# OUTLINE

- Key discovery in relational databases
- Key discovery in the Semantic Web
- Different approaches for key discovery in the Semantic Web
- Conclusions

# 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. SEMANTIC WEB

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

  - 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 THE SEMANTIC WEB

# KEYS - 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]***

# KEYS IN THE SEMANTIC WEB

| | 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 THE SEMANTIC WEB

<p1, StudiedIn, UCC>
<p1, StudiedIn, Yale>

| | 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 THE SEMANTIC WEB

No triple containing the birthdate of p1

| | 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 THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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 THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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 THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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 THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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 THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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}
{SSN}
{LastName,StudiedIn}
...

# KEYS IN THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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}
**{SSN}**
{LastName,StudiedIn}
...

# KEYS IN THE SEMANTIC WEB

- **Three main types of keys in the SW:**
  - **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}
{SSN}
**{LastName,StudiedIn}**

# 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

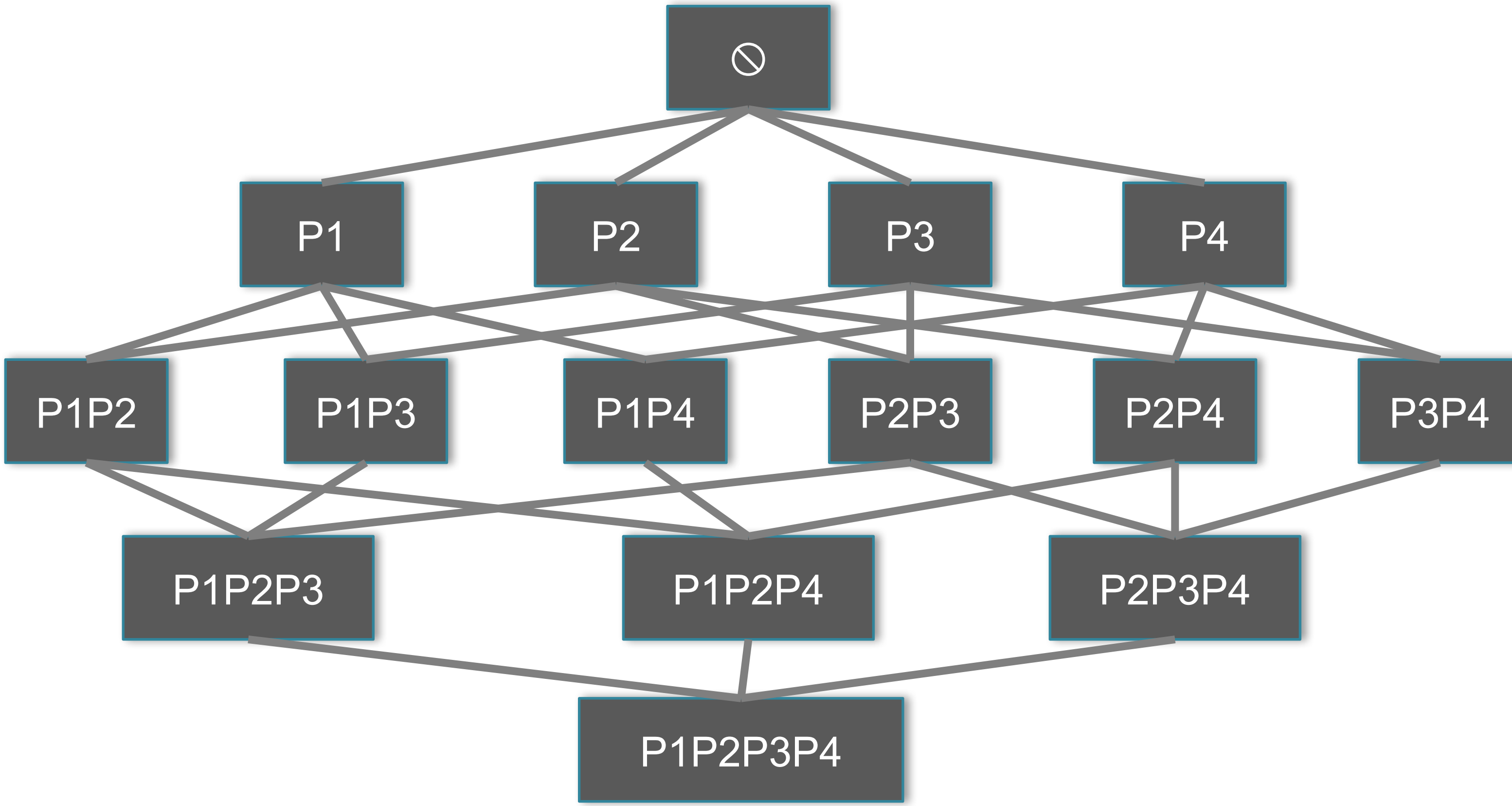
- **S-Keys**
  - An automatic key discovery approach for data linking
  - SAKey: Scalable almost key discovery in RDF data
  - VICKEY: Conditional key discovery
  - Linkkey: Data interlinking through robust Linkkey extraction

# KEY DISCOVERY 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

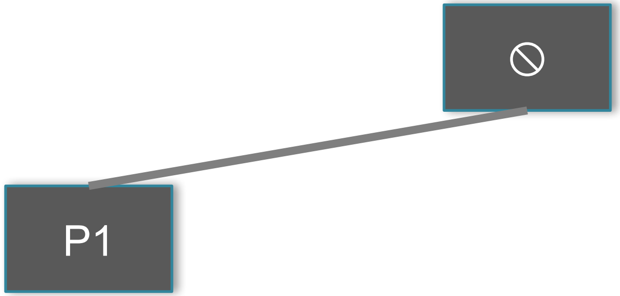
# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING [Atencia et al.12]

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



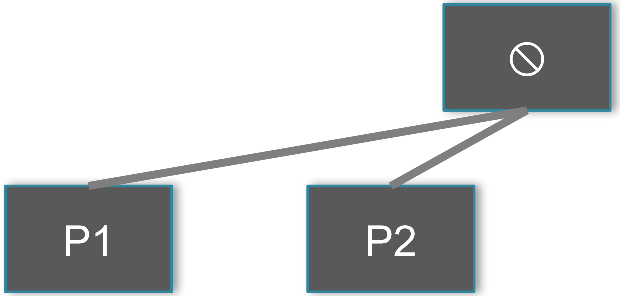
# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING [Atencia et al.12]

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



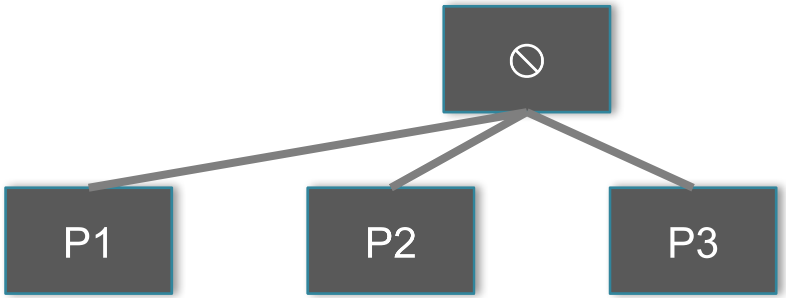
# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING [Atencia et al.12]

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



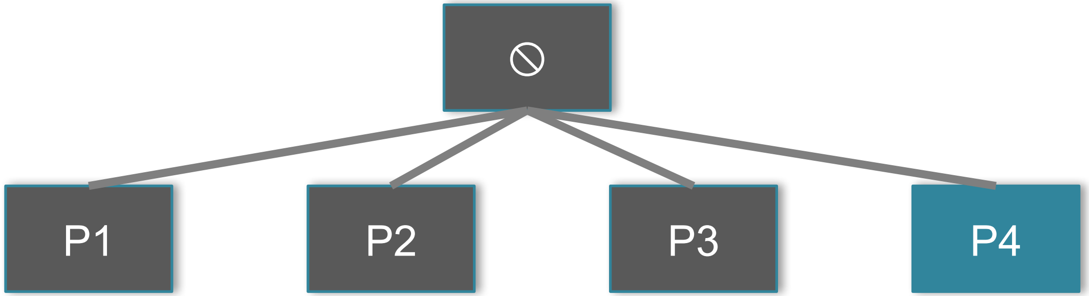
# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING [Atencia et al.12]

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



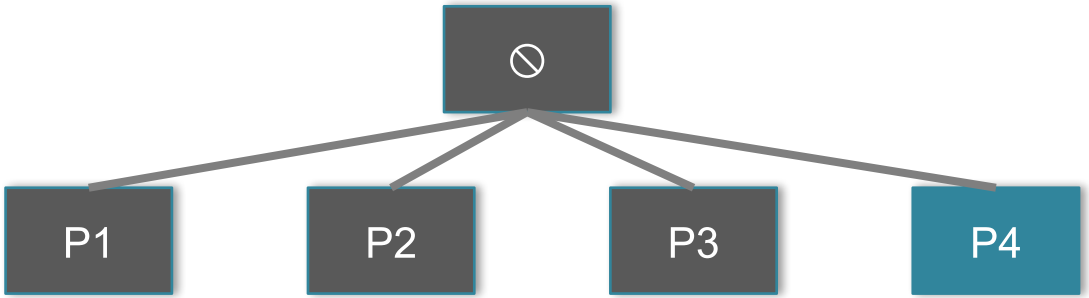
# KEYS AND PSEUDO-KEYS DETECTION FOR WEB DATASETS CLEANING AND INTERLINKING [Atencia et al.12]

- **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]

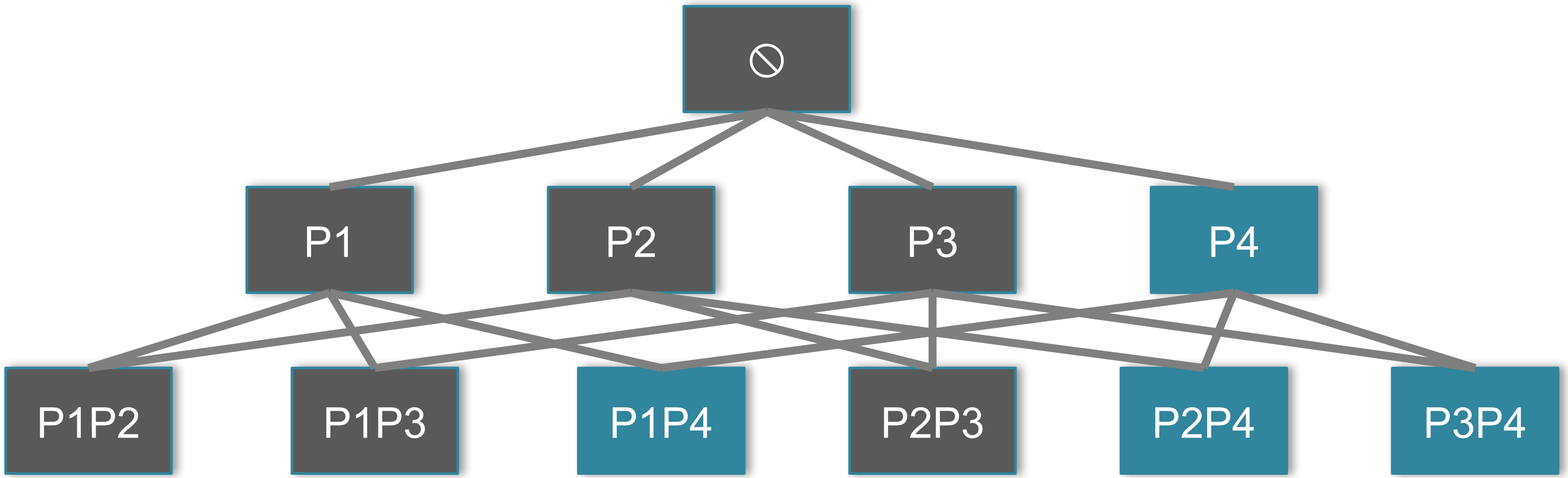
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions




**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]

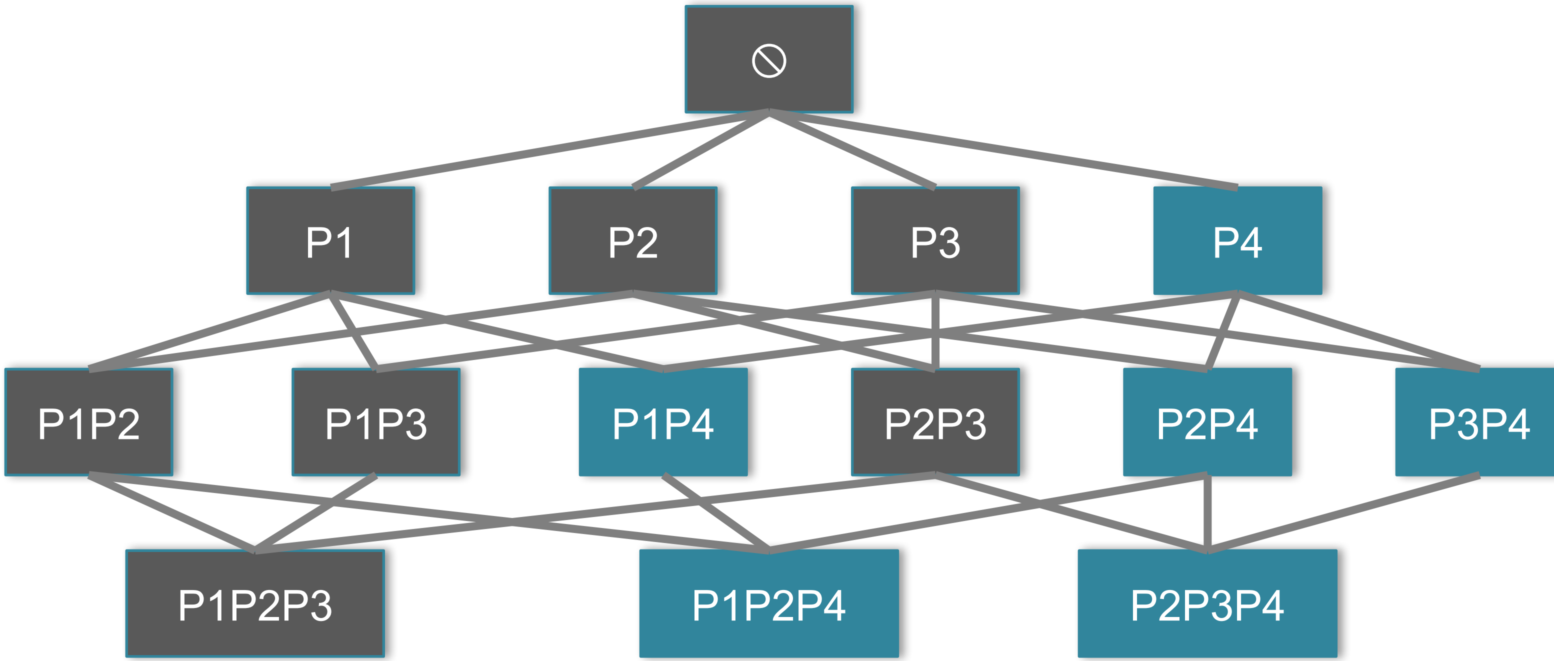
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions




**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]

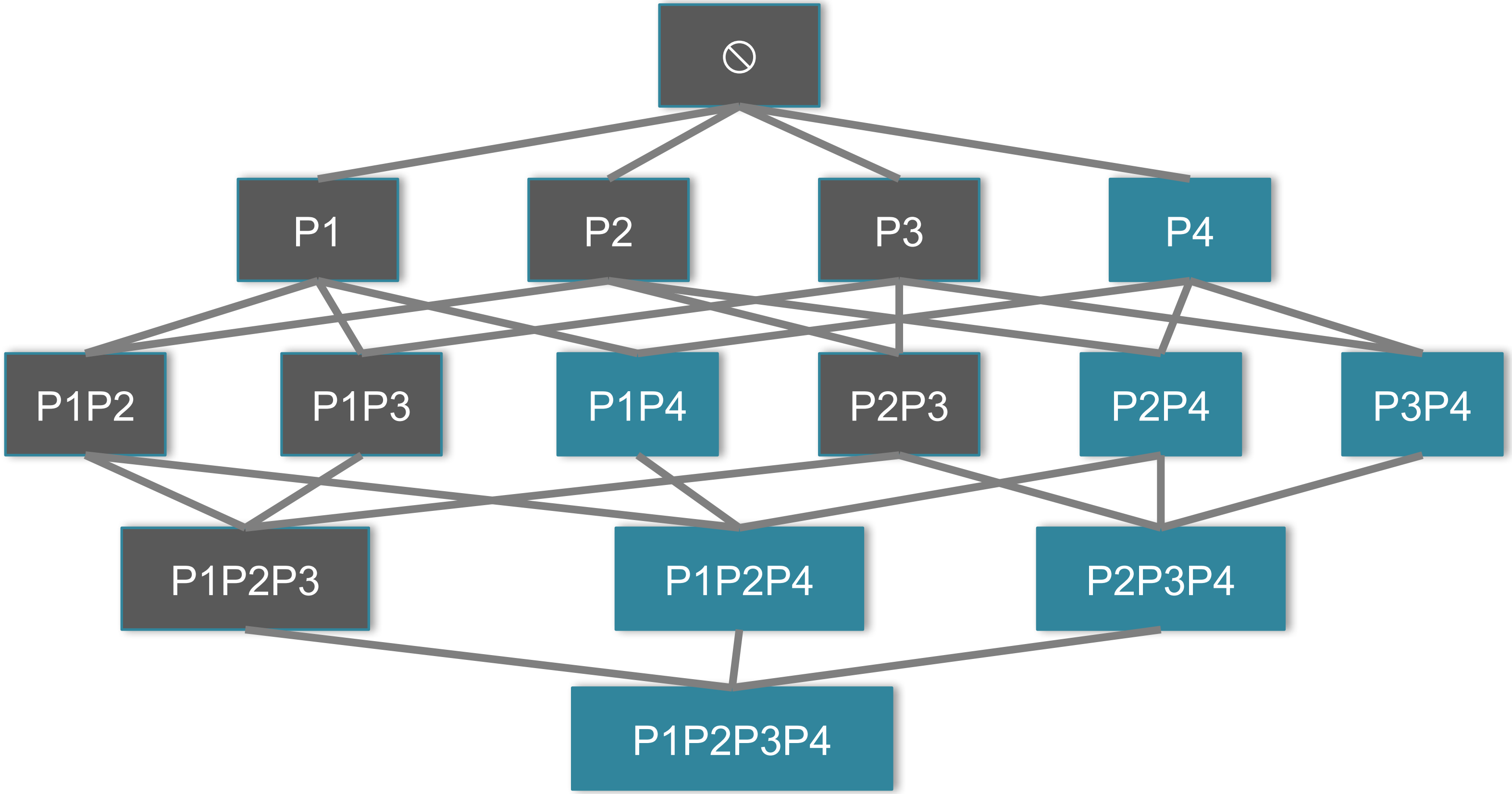
- **Bottom-up approach that discovers:**
  - **SF-keys for a given class**
  - **SF-pseudo keys for a given class:** sf-keys that tolerate exceptions




**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*:

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

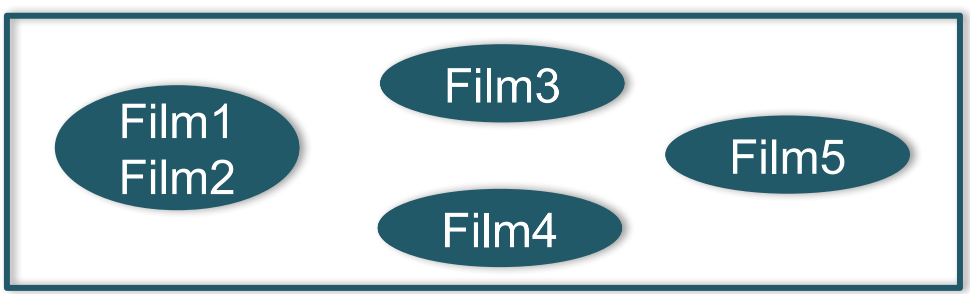
  - **Discriminability** of a set of properties *P* (pseudo-keys):

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

# 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 | --- | 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 | --- | 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 | --- | 12/6/12 | --- | english |

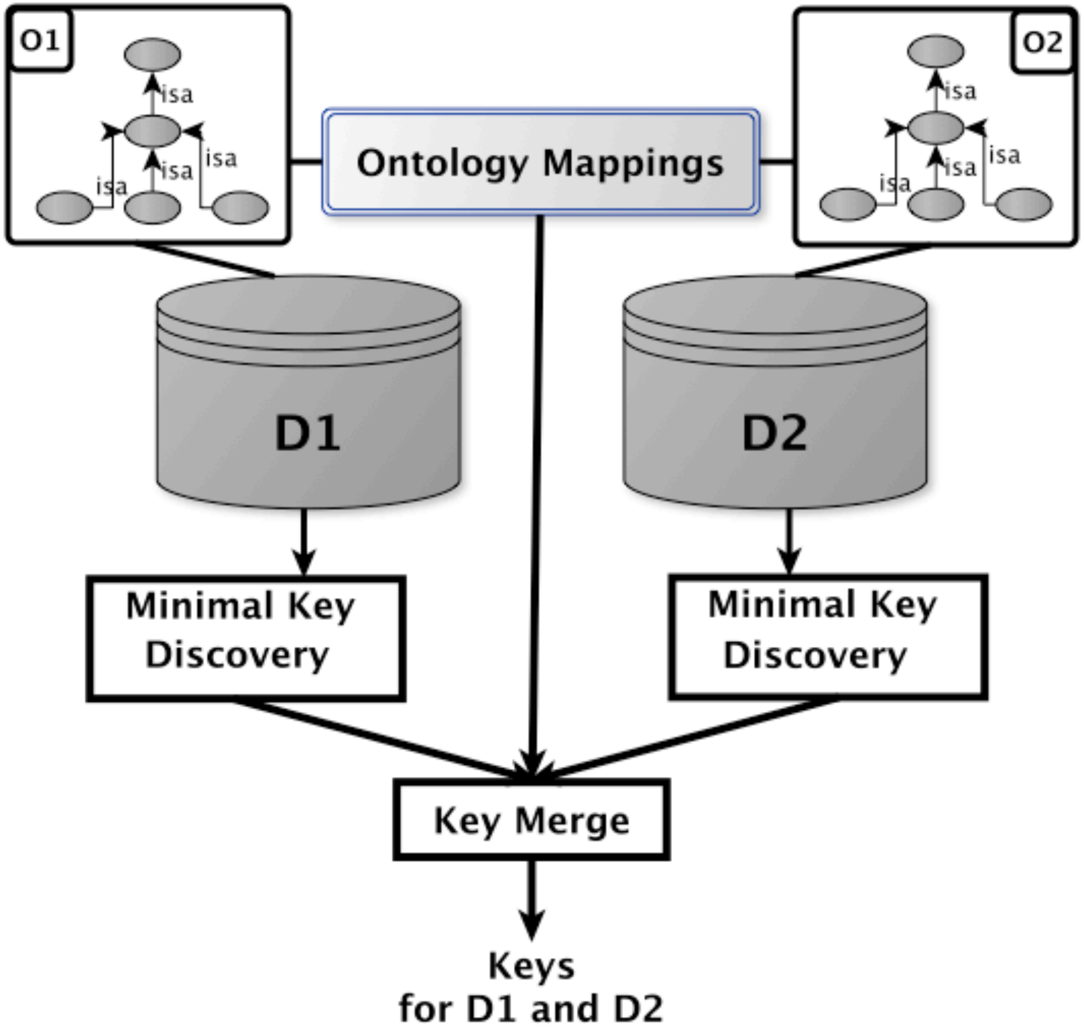
*Tool available in https://github.com/AKSW/rocker*

# 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

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

[Pernelle et al. 2013]



# 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

# 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 Set $E_P$:** set of instances that share values for the set of properties *P*
- **n-almost key: a set of properties where $|E_P| \leq n$**
- **n-non key: a set of properties where $|E_P| \geq n$**

All combinations of properties

# *N*-ALMOST KEYS

[Symeonidou et al.14]

- **Exception Set $E_P$:** set of instances that share values for the set of properties *P*
- **n-almost key: a set of properties where $|E_P| \leq n$**
- **n-non key: a set of properties where $|E_P| \geq n$**


All combinations of properties

All sets of properties that contain at least 5 exceptions

# *N*-ALMOST KEYS

[Symeonidou et al.14]

- **Exception Set $E_P$:** set of instances that share values for the set of properties *P*
- **n-almost key: a set of properties where $|E_P| \leq n$**
- **n-non key: a set of properties where $|E_P| \geq n$**

5-non keys

4-almost keys

All combinations of properties

All sets of properties that contain at least 5 exceptions

All sets of properties that contain at most 4 exceptions

# 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](https://www.lri.fr/sakey)*

# 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, valid for 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

*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.17]

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

# CONDITIONAL KEY GRAPH EXPLORATION

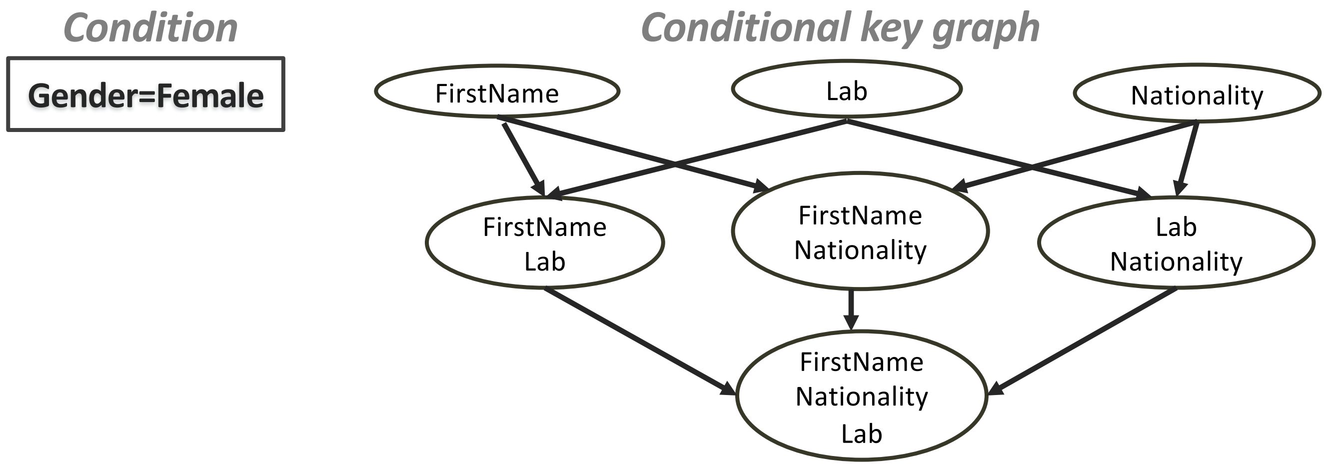
**[Symeonidou et al.17]**

- **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]

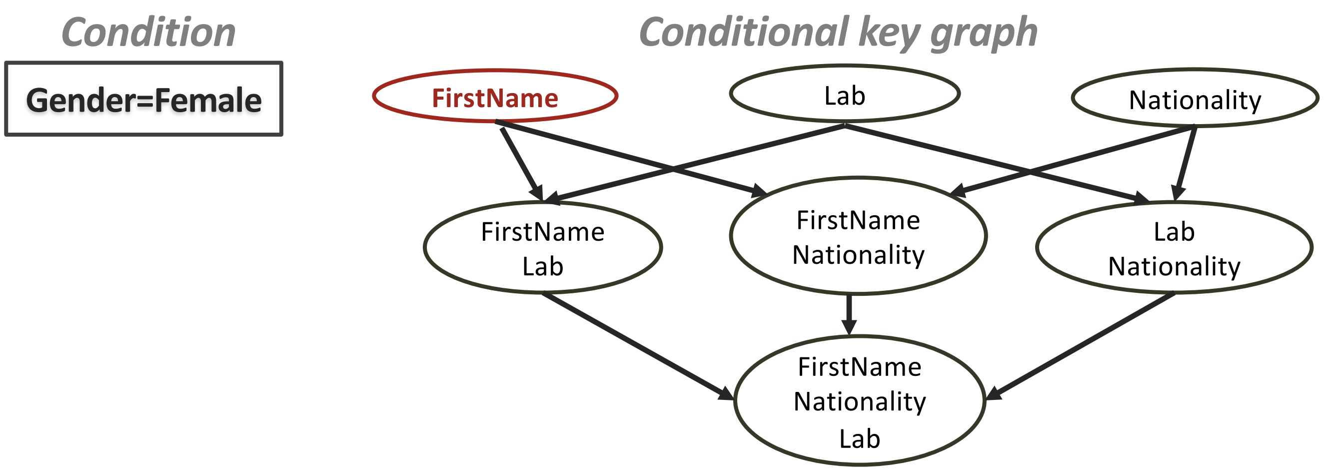
- **Step 1: Discover all minimal conditional keys with condition {p=a}**
- Step 2: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge p_2=a_2\}$
- …
- Step n: Discover all minimal conditional keys with condition $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$



# 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 $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$

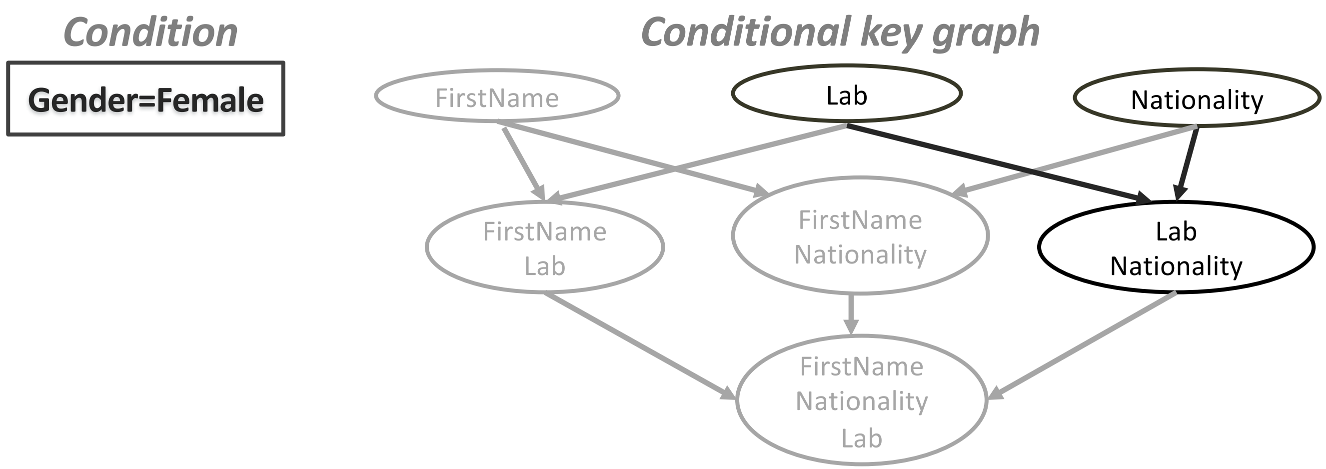



**Key {FirstName} for cond {Gender=Female}**

# 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 $\{p_1=a_1 \wedge \ldots \wedge p_n=a_n\}$




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

# 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** | **x 1.75** |
| | Conditional keys** | 0.57 | 0.99 | 0.73 | |
| | **Keys[1]+Conditional keys** | 0.6 | 0.99 | **0.75** | |
| **Album** | Keys[1] | 0 | 1 | **0.00** | **x 869** |
| | Conditional keys | 0.15 | 0.99 | 0.26 | |
| | **Keys[1]+Conditional keys** | 0.15 | 0.99 | **0.26** | |
| **Film** | Keys[1] | 0.04 | 0.99 | **0.08** | **x 7.1** |
| | Conditional keys | 0.38 | 0.96 | 0.54 | |
| | **Keys[1]+Conditional keys** | 0.39 | 0.98 | **0.55** | |
| **Museum** | Keys[1] | 0.12 | 1 | **0.21** | **x 2.19** |
| | Conditional keys | 0.25 | 1 | 0.40 | |
| | **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

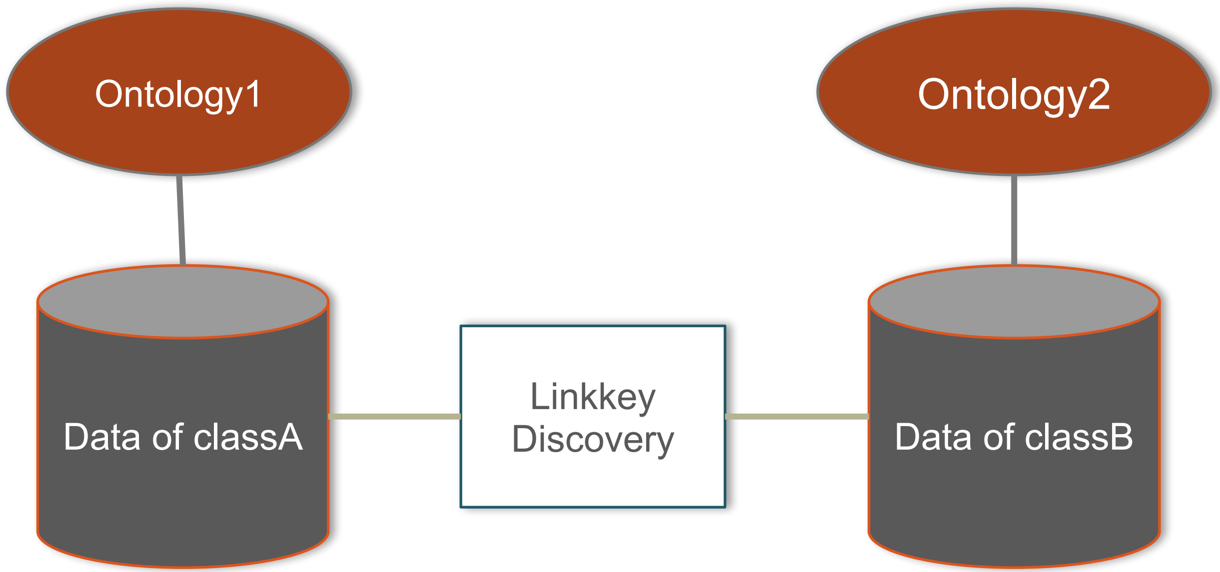
- **F-Keys**
  - ROCKER: A Refinement Operator for Key Discovery

- **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

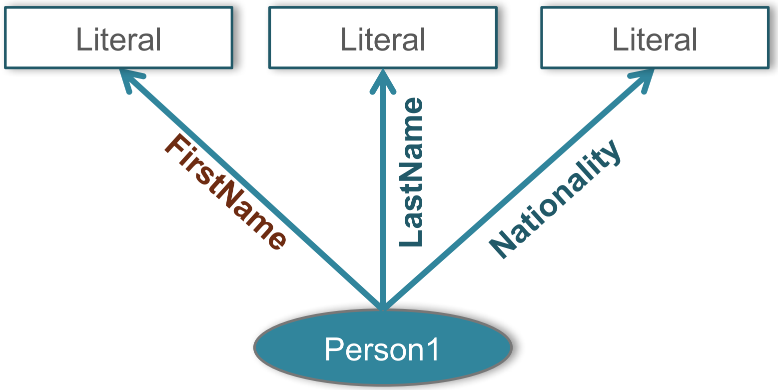
[Atencia et al.14]

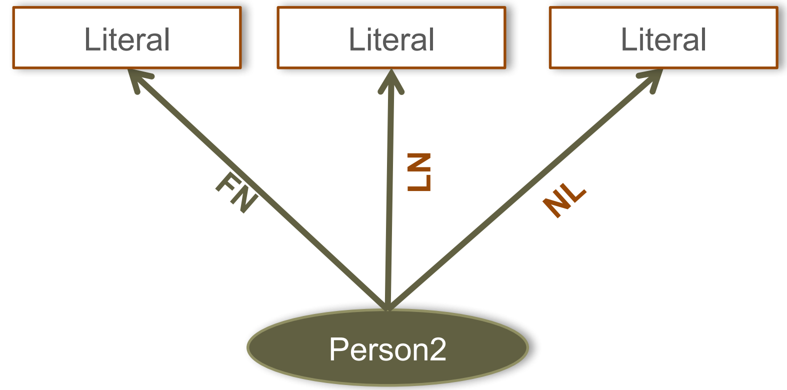
- **Given a pair of classes in two datasets conforming to two different ontologies:**
  - Discover ***Linkkeys*** – maximal sets of property pairs that can link instances of two different classes

# 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.