VICKEY: Mining conditional keys on Knowledge Bases

Danai Symeonidou, Luis Galárraga, Nathalie Pernelle, Fatiha Saïs, Fabian Suchanek

ISWC 2017 – October 24th 2017
Vienna

Keys in Knowledge bases

- What is a **Key**
  - Set of properties that uniquely identifies every instance of a class
    - Given the key **{FirstName, LastName, AuthorOf}** and two people x1, x2
      - $\text{FirstName}(x1,y) \wedge \text{FirstName}(x2,y) \wedge \text{LastName}(x1,z) \wedge \text{LastName}(x1,z) \wedge \text{AuthorOf}(x1,k) \wedge \text{AuthorOf}(x2,k) \rightarrow$ **sameAs(x1, x2)**

- **Use of keys**
  - Source of knowledge for the data
  - Ontology enrichment (hasKey Axiom in OWL2)
  - Data Linking

Keys in Knowledge bases

- To obtain keys found in Knowledge Bases (KBs)
  - Different automatic key discovery approaches (eg. SAKey[1], KD2R[5])

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

Key: {LastName, Gender}

- **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

Problem statement

- **Motivating example**
  - In many countries of the world, 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

- **Approximate keys**[1,6] express keys that do not uniquely identify every instance of a class in a dataset

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

Conditional keys by VICKEY

- **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

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

- **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

- **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

- **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

- **A key is also a conditional key under any condition**
  - {LastName, Gender} is a *key*

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

VICKEY: Mining efficiently conditional keys

- **A key is also a conditional key under any condition**
  - {LastName, Gender} is a *key*
  - {LastName, Gender} is a *key* under the *condition* {Lab=INRA}

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

VICKEY: Mining efficiently conditional keys

- **A key is also a conditional key under any condition**
  - {LastName, Gender} is a *key*
  - {LastName, Gender} is a *key* under the *condition* {Lab=INRA}
  - {LastName, Gender} is a *key* under the *condition* {Lab=INRA^Nationality=Greece}

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

VICKEY: Mining efficiently conditional keys

- **A key is also a conditional key under any condition**
  - {LastName, Gender} is a *key*
  - {LastName, Gender} is a *key* under the *condition* {Lab=INRA}
  - {LastName, Gender} is a *key* under the *condition* {Lab=INRA^Nationality=Greece}

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

New conditional keys that cannot be inferred from keys can be found in non-keys

VICKEY: Mining efficiently conditional keys

- **Non-key:** a set of properties where two instances share at least one value for each property in the non-key (SAKey[1])

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

- Given a non-key
  - Step 1: Discover all minimal conditional keys with condition of size 1
  - Step 2: Discover all minimal conditional keys with condition of size 2
  - …
  - Step n: Discover all minimal conditional keys with condition of size n

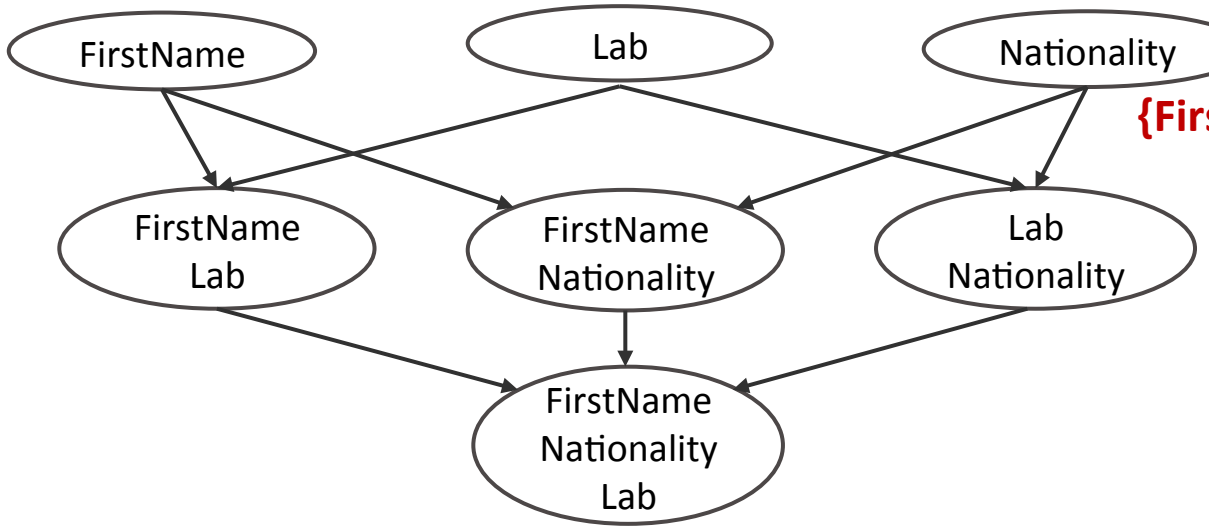
Conditional key graph exploration

- 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

Conditional key graph



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

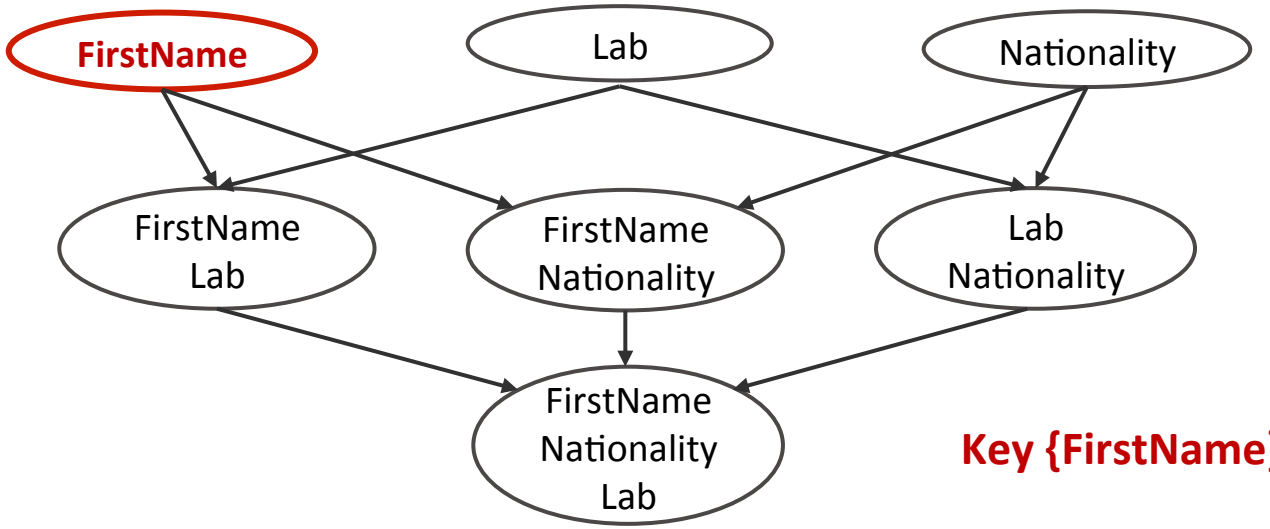
Conditional key graph exploration

- 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

Conditional key graph



Key {FirstName} for cond {Gender=Female}

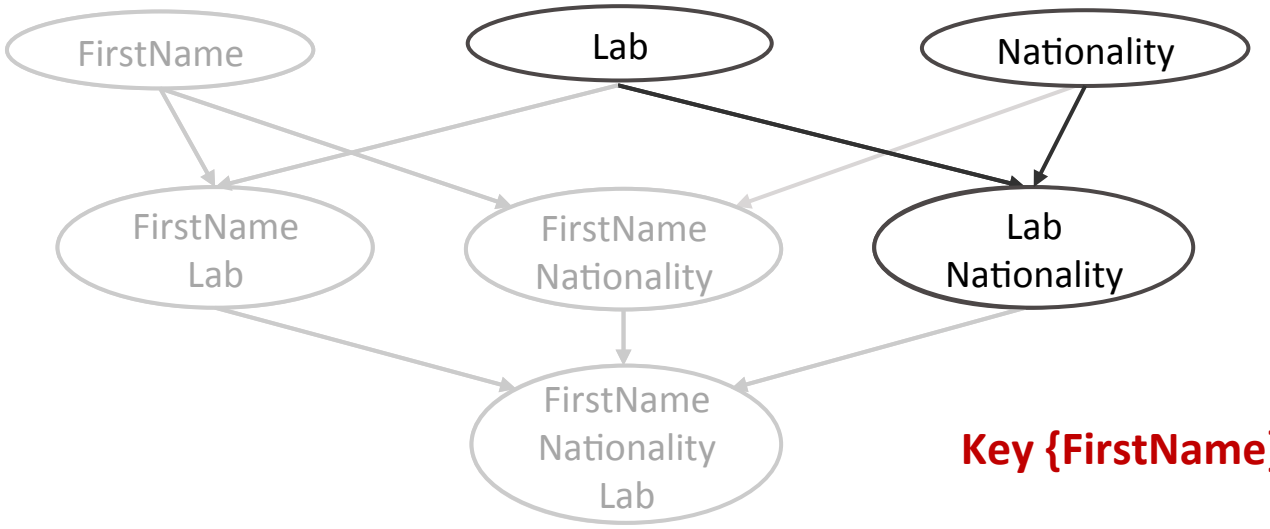
Conditional key graph exploration

- 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

Conditional key graph



Key {FirstName} for cond {Gender=Female}

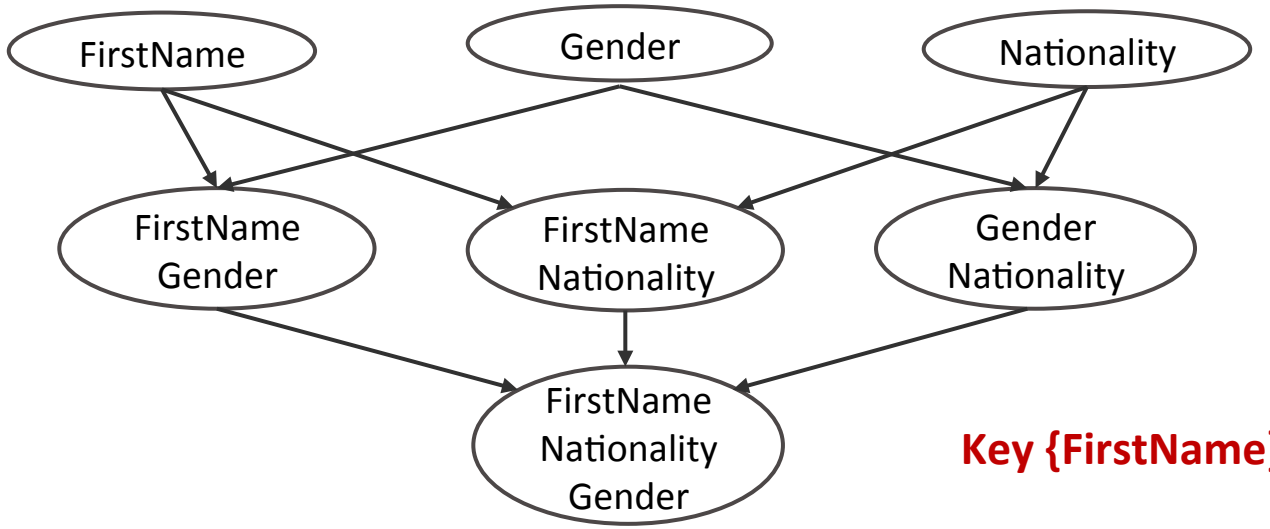
Conditional key graph exploration

- 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



Key {FirstName} for cond {Gender=Female}

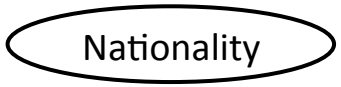
Conditional key graph exploration

- 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

- Evaluation of VICKEY's **scalability** in large datasets
- Evaluation of conditional keys **in data linking**

VICKEY's scalability

- 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[1] to reduce the search space of conditional key mining

- Coverage 1%

Runtime results - VICKEY vs. AMIE[2]

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]

- **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[1]

- 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[1]

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*

Conclusion

- VICKEY: a conditional key mining approach
  - Providing new knowledge about the data
  - Able to treat large datasets in a scalable way
  - Able to improve significantly the data linking results

- **VICKEY is available here: https://github.com/lgalarra/vickey**

Conclusion

- VICKEY: a conditional key mining approach
  - Providing new knowledge about the data
  - Able to treat large datasets in a scalable way
  - Able to improve significantly the data linking results

- **VICKEY is available here: https://github.com/lgalarra/vickey**

Thanks for your attention!

References

- [1] D.Symeonidou, V.Armant, N.Pernelle, F.Saïs. SAKey: Scalable Almost Key discovery in RDF data. In ISWC, 2014.
- [2] L.Galárraga, C.Teflioudi, K.Hose, F.Suchanek. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In WWW, 2013.
- [3] J.Lehmann, R.Isele, M.Jakob, A.Jentzsch, D.Kontokostas, P.Mendes, S.Hellmann, M.Morsey, P.Kleef, S.Auer, C.Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web J., 6(2), 2015.
- [4] F.Suchanek, G.Kasneci, G.Weikum. Yago: a core of semantic knowledge. In WWW, 2007.
- [5] N. Pernelle, F. Saïs, and D. Symeonidou. An automatic key discovery approach for data linking. J. of Web Semantics, 23, 2013.
- [6] M.Atencia, J. David, and F. Scharffe. Keys and pseudo-keys detection for web datasets cleansing and interlinking. In EKAW, 2012.

Minimal Conditional keys

- A conditional key with condition part CD and properties P is minimal, if
  - the removal of a condition in CD results in a non conditional key
    - **P={FirstName, LastName} CD=~~worksAt(x, INRA)~~^livesIn(x, Paris)<**
  - the removal of a property in P results in a non conditional key
    - **P={FirstName, ~~LastName~~} CD=worksAt(x, INRA)^livesIn(x, Paris)**
  - the transfer of a property p from CD to P (with the corresponding removal of the condition) results in a non conditional key
    - **P={FirstName, LastName} CD=worksAt(x, INRA)^livesIn(x, Paris)**
    - **P={FirstName, LastName,worksAt} CD=livesIn(x, Paris)**

Minimal Conditional keys

- A conditional key with condition part CD and properties P is minimal, if
  - the removal of a condition in CD results in a non conditional key
    - **P={FirstName, LastName} CD=~~worksAt(x, INRA)~~^livesIn(x, Paris)<**
  - the removal of a property in P results in a non conditional key
    - **P={FirstName, ~~LastName~~} CD=worksAt(x, INRA)^livesIn(x, Paris)**
  - the transfer of a property p from CD to P (with the corresponding removal of the condition) results in a non conditional key
    - **P={FirstName, LastName} CD=worksAt(x, INRA)^livesIn(x, Paris)**
    - **P={FirstName, LastName,worksAt} CD=livesIn(x, Paris)**

Data Linking

Class	\#Pro	\#Ks	\#NKs	\#CKs
Actor	16	93	22	748
Album	5	1	2	5864
Book	7	5	2	538
Film	9	14	13	26750
Mount.	5	3	2	775
Museum	7	14	5	80
Organiz.	17	149	3	9737
Scientist	10	22	8	407
Univ.	9	5	5	449

Discovered conditional key examples:

motto is a key for universities in Italy and some other countries – but not in all countries;

name is a key for organizations in certain places – but not all places.