Jean-Marc Petit

INSA Lyon, Université de Lyon LIRIS CNRS (UMR 5205)

24th International Symposium on Methodologies for Intelligent Systems (ISMIS 2018) Limassol, Cyprus

・ロ> <同> <目> <目> <目> <同> <</li>

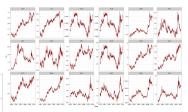
#### Data diversity

#### Data Diversity: not only a gender question !



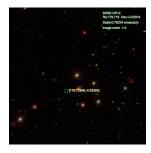


		OD WINTCOM OD O			
	000		where the set of the set		
			week is without		term interime (c21020)
thread Automod PPEDICT		Bud Artiston 0.0	IONING PREDER		090,0391.0
INCOMPANY 1 Metaland Market		windowskie bes to N'EGER O'G	USQUEERED is relationity to		
					3070,010070
lation, properties			( Norme		here property
tabon properties electronic nee a PETERS O C		M MT120	MATERIA D.O.G	Nerve an ACTION CO.0	
Warlonahip nga a PETERIN () (		Manager A Parager	AND AND ADDRESS		property, key shift control of
				But Manda Arrowaldas	property and and considered
a ayoners intoore		INDUSTRY, IGA AF N'ECER CO CO	Service VAD-ARIEL	AX. ANDYA 14YO MICH	
a ani ayemate islikatik		here is a lister of the	an 10/2000 00	Reader Very Routing	
			W. COLUMN PATIENT CO.C.	Brown, origon, it mittore is	
a where PERDIN		accolution.	N, MA PETERS 0.0		
a metadora neg hilihürin		Marraes 0.0.0 T	in which in forces in the		
		MALM NYCODY CO-CO		term.deferrise	sow_dowd
		pera product, of IVTICER. O. C.			
		10,00017207			
	MATERIA O.O.O.	attentiate (h1000)	midner	are served site and	USO, Elevery 14 shared min for defin
	MODIFICATION COL		M M70901 0 0 0	WERE NUCLAUSE	
pecket	NAME OF TAXABLE PARTY OF		ade VAD-ARE D	UNDERMINE	
HATCH OOG		econtation explifier	appropriate to ATTRACT COLOR		
NOLTON OFFICE		Marrage 0.0.0	Book of NY Stock CO. C. C.	And advert	
NO. 104. 6 PETERS		Mediate M 177027 0-0	MA NO VACUNCES	WHEN A PROOF OF CO	
inang allos [10]	HAINED DOG	No. 1 AUGO C.C.	and an every state of the second state	adapt is NUMER. C. C.	daard
personal transmission	Netw VARCHINGS	HER VARD RATES			
perior CODANDER	Meane 2000A00(080)		and desce allowed	gene product count	and shrane UVD-AU
	general of minical states		Read to Personal Co. Co.	008-1410-ARE	
	WHEN OF PERCENTION	pro-product	Red o Proces 0.0	American COD-0100	and sing to CARDIN
	phone of minical states of system UNICAR(310) of another UNICAR(310)		Beet, e Pristen 0.0	aminulanere LAID-ARIEL aminula Pritoch G	and sing to CARDIN
	ar syste UVCARDID ar syste UVCARDID ar ander VVCARDID ar ander vVCCARDID	MINUTER O O O O	1	American COD-0100	and sing to CARDIN
et velue initiate spin velue initiate encoder and schemelation encoders and schemelation encoders and schemelation	phone of minical states of system UNICAR(310) of another UNICAR(310)	Article C C C C C C C C C C C C C C C C C C C	( proclamation of a	aminulanere LAID-ARIEL aminula Pritoch G	and sing to CARDIN
ML Han Proofs spr. Jakar (2000) Benote Jaka M INCODER J. Sek. M	ar syste UVCARDID ar syste UVCARDID ar ander VVCARDID ar ander vVCCARDID	MINION CO.C.	1	aminulanere LAID-ARIEL aminula Pritoch G	and sing to CARDIN
ML Han Proofs spr. Jakar (2000) Benote Jaka M INCODER J. Sek. M	ar syste UVCARDID ar syste UVCARDID ar ander VVCARDID ar ander vVCCARDID	Article C C C C C C C C C C C C C C C C C C C	( proclamation of a	aminulanere LAID-ARIEL aminula Pritoch G	and sing to CARDIN
M. Han Process operation (2020) Received and Control (2020) (INCORE of the Add	genet of mirchingso of genet VVCAAData of genet VVCAAData of genet VVCAAData of genet vVCAAData biologic vVC	MINION CO.C.	gene product and gene product of the second	entrelinere UVO-M/30 entrel e NYEGEN (2) poblit.com NYEGEN (2)	and keyings 1910-1910 and almos 1970-09100 (Methodisma lang-and dis
All John P. TODA Spit, John P. TODA Benevic, John V.V.O.A.AUTO Infollition: Sea .10 Control Marcolan Marcolan Marcolan Marcolan Control Marcolan Control Marcolan Control	press of refrontion of gran UVCOADD of anyte UVCOADD	He Writes O O O     even Write O O O     even write Statement of	Burn Josefant ver Brin Josefant ver Brin Josefant Vitterin O.G.	American Antonia (MC-44)(1) American Antonia (MC-44)(1) Am	and sing to CARDIN
M AND PTOTS NO. 1997 AND TOTS NO. 1997 AND TOTS POSITIVE LINE AND POSITIVE LINE AND POSITIVE OF A DECISION AND A DECISION AND A DECISION AND A DECISION AND A DECISION AND A DECISION AND A DECIS	proc. or nerconcept or synar UNCOADD of anyon UNCOADD	Winner 0.000 setter WitchArtter 0 develue Mittaler 0.00 tages at Mittaler 0.00 tages at Mittaler 0.00	Burn Josefant ver Brin Josefant ver Brin Josefant Vitterin O.G.	initializer (VC-0.00)	and pine to Constraint Websitement and and
Al John NYICER ph John MYICER second John MYICER MYICER & MAINE MYICER & MAINE MYICER & O O O MYICER & O	press of refrontion of gran UVCOADD of anyte UVCOADD	di Articia dendi Vercontoni o dendi ver	gen product any en alternation of a set of transmission by product and transmission by product and transmission transmission	eministere UND-ANUL server, a referit polatiouri PTERE server and PTERE server and PTERE dest of PTERE dest of PTERE dest of PTERE	and pine to Constraint Websitement and and
et alles Process per una hillion inscrine, una Uniterative inscrine, una Uniterative inscrine de la Uniterative de la Un	Andrea, or restances to interprete transmission program (Control of the Distribution) and the Control of the Distribution and the Control of the and the Control of the and the Control of the Andread States of the Control and the Control of the Control of the Control and the Control of the Control of the Control and the Control of the Control of the Control of the Control and the Control of the Control of the Control of the Control and the Control of the Control of the Control of the Control and the Control of the Co	Dear Dearent Atmoska Con- Processing of the second second Processing of the second	Mana Danger Pada Ana Danger Pada Ana Pada Pada Pada Ana Pada Pada Pada Ana Pada Pada Pada Ana Pada Pada Pada Pada Ana Pada Pada Pada Pada Pada Ana Pada Pada Pada Pada Pada Pada Pada Pa	Americalization VALUE AND America de NYSOL Poblic Journ PYSOL America de NYSOL America de NYSOL America de NYSOL DUGUNa delano	and pine to Constraint Websitement and and
Al John PTOCH opti John HILLON Becolar, and United Apple Intellined: Sea Jil Becolar Michael Status Michael Al PTION Rep. press, press, or NOSE Rep. Press,	Annu Arten Methodski dr. gange McDraholski dr. gange McDraholski dr. gange McDraholski dr. gange McDraholski dr. gange McDraholski gang ganger Artikelik gang ganger Artikelik methodski Artikelik m	B AYERS 0.000 BER WOOK 0.000 BER WOOK 0.000 BER AYERS 0.0000 BER AYERS 0.00000 BER AYERS 0.00000 BER AY		Annualinan UND-MIDI Index Original Index Out PTER	and any information which any out of
M. Jako 170208 apis olara hilosofi apis olara hilosofi apis olara hilosofi apis olara hilosofi andre and apis olara hilosofi andre and apis olara hilosofi devel a 171201 devel a 171201 Statuta 1712	Press, of the control of the co			Exemplanes UNCANSE     Exemplanes UNCANSE     product ours Protocol     exemplanes     Exem	and alian instructions with the second dist
Al John PTOCH opti John HILLON Becolar, and United Apple Intellined: Sea Jil Becolar Michael Status Michael Al PTION Rep. press, press, or NOSE Rep. Press,	Press, of TRECONSTRUCT of speak VAROUNDER of parts VAROUNDER of parts VAROUNDER indefinitions pers press and parts pers press and parts pers press and parts person perso	B AYERS 0.000 BER WOOK 0.000 BER WOOK 0.000 BER AYERS 0.0000 BER AYERS 0.00000 BER AYERS 0.00000 BER AY		Annualinan UND-MIDI Index Original Index Out PTER	



#### Example from the astrophysics domain

#### The Sloan Digital Sky Survey (SDSS): Mapping the Universe !



Class	err	 errg	$\frac{r}{err_r}$	i erri	$\frac{z}{err_7}$
STAR	16.56	14.62	13.94	13.79	13.48
	0.01	0.00	0.01	0.01	0.00
Galaxie	19.79	17.77	16.59	16.07	15.63
	0.06	0.01	0.00	0.00	0.01
STAR	15.64	14.04	14.57	12.83	13.12
	0.01	0.00	0.01	0.00	0.01
Galaxie	21.61	20.81	19.87	19.30	19.03
	0.15	0.04	0.02	0.02	0.05
STAR	20.09	17.28	15.79	14.31	13.49
	0.04	0.00	0.00	0.00	0.00

5 magnitudes (u, g, r, i, and z) catalog database

### $\Rightarrow$ Require to deal with numerical interval data as first class citizen

See http://www.sdss.org/dr12/ for details

#### Data and metadata from SDSS

Look up	p common na	me					SCI	Look u	p common nar	THE					30
Type STAR		verun camo 301 2	x field 427	0bj 250	12	SDSS Object I 3764872069375		Тур				feld obj		SDSS Object	
31745	100	RA Dec	421	200	10	Galactic Coord		GALA	XY 756	301	2	427 259		37648720593	/00103
	Decimal	100,000	Sex	agesimal		/	b		Decimal	RA, Dec		wacesimal		Galactic Coon	tinates (I, 0)
179.7	13207494, -0	484210574	11:58:51.1	6, -00:29:03.15	276	5.198583367	59.630008868	179.7	18015380, -0.5	530204150		.32, -00:31:48.73	276	243764984	59.5897794
nagin	g							Imagin	a						
ags D		_AT_EDGE ST/	TIONARY E	BINNED1 INTER	6P					AT_EDGE STAT		BINNED1 INTER	P		
۶I	N			Ma	gnitud	les		5				м	agnitude		
			u	9	r	- I.	z				U	a	r		z
		w	20.09	17.28	15.79	14.31	13.49			w	19.79	17.77	16.59	16.07	15.6
				Magnitu	ie unce	ertainties						Magnitu	de unce	rtainties	
			err_u	err_g	err_r	err_i	err_z				err u	err g	err r	err i	orr
			0.04	0.00	0.00	0.00	0.00				0.06	0.01	0.00	0.00	0.0
		1 1							9	100					
Image MJD	mode	Other observations	P	arentID	nChild	extinction_r	PetroRad_r (arcsec)	Image MJD	mode	Other observations		parentID	nChild (	extinction_r	PetroRad (arcsec
51259	PRIMARY	2	1237648	720693756153	0	0.07	1.22 ± 0.008	51259	PRIMARY	5	123764	8720693756162	0	0.06	6.00 ± 0.6
Midul	Date	photoZ (K	D-tree meth	od)	Ga	alaxy Zoo 1 mo	rphology	Mjd-	Date	photoZ (K0	-tree met	hod)	Gal	axy Zoo 1 m	arphology
	/1999							03/22			± 0.0127			Uncerta	

SDSS	J11584	3.49-00	)3233.8

Look up com	in the second							
Type	run	rerun	cancol	feid	obj	SDSS Obje	t ID	
GALAXY	758	301	2	428	465	1237648720693	89083	NOT IOE
		RA,	Dec			Galactic Coor	dinates	((, b)
0	lecimal			Sexages	limal	1		b
179.6812169	944, -0.54	2727601	11:58	43.490	0:32:33.81	276.187065846	59.5	\$3578544

#### Imaging

Flags	DEBLENDED_AT_EDGE STATIONARY MOVED BINNED1

I			Magnitudes		
	u	9	r	1	z
6	22.40	20.10	18.97	18.46	18.01
		Magni	itude uncerta	inties	
	err_u	err_g	err_r	err_i	err_z
	0.49	0.03	0.02	0.02	0.04

Image MJD	mode	Other observations	parentiD	nChild	extinction_r	PetroRad_r (arcsec)	Þ	< ⊡ >	٠	2	Þ.	•	2	•	- 토	500
51259	PRIMARY	4	1237648720693690832	0	0.06	3.60 ± 0.304										

#### Data diversity

To cope with data diversity, key notions have be studied for years in computer science:

- data and metadata representation,
- data uncertainty,
- data inconsistency,
- data heterogeneity ...

Dealing with data diversity remains the hardest thing in practise ⇒ Require to understand *what's hidden behind the data*:

• Where do they come from ? How are they produced ?

 $\Rightarrow$  Be as close as possible of the available data sources and experts to better match their intended meaning

#### Data diversity

To cope with data diversity, key notions have be studied for years in computer science:

- data and metadata representation,
- data uncertainty,
- data inconsistency,
- data heterogeneity ...

Dealing with data diversity remains the hardest thing in practise  $\Rightarrow$  Require to understand *what's hidden behind the data*:

• Where do they come from ? How are they produced ?

 $\Rightarrow$  Be as close as possible of the available data sources and experts to better match their intended meaning

#### Data dependencies

# Classical example of data dependencies: functional dependencies

$$r \models X \rightarrow Y$$
 iff for all  $t_1, t_2 \in r$ 

If for all  $A \in X$ ,  $t_1[A] = t_2[A]$  then for all  $B \in Y$ ,  $t_1[B] = t_2[B]$ 

Turns out to be a very general notion, related to implications.

а	b	a  ightarrow b
0	0	1
0	1	1
1	0	0
1	1	1

Many connections with lattice theory, formal concept analysis (Galois connection) and logics (see for ex [11])

Crucial to understand relational database design

#### Beyond database design

New and timely applications require some forms of FD:

- Data quality: Analysing existing data to identify data quality problems [17, 9]
- Machine learning over relational databases: FD-aware optimization for in-database learning [19]
- Semantic query optimization: Query rewriting techniques based on data dependencies [12]

 $\Rightarrow$  Many extensions of FD have been proposed to take into account some forms of data diversity (e.g. see [10, 18] for a survey)

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

- Matching Dependencies, Denial constraints ... [17, 9, 15]
- Implications in Formal Concept Analysis (FCA) [7, 6]
- Association rules ... in Data mining [5]

#### Data diversity and data dependencies

### Questions and Contributions

How to take into account data diversity for data dependencies ? Does there exist unifying frameworks ?

Two contributions:

- RQL: a query language to express implications over relational databases (ISMIS 2005 [3], demo ICDM 2014 [13], TCS 2017 [14])
- Structural properties on attribute domains (ongoing work)

### Contents

#### RQL query language

- Preliminaries
- Main result underlying RQL
- The RQL language
- RQL implementation
- Summary
- 2 Structural properties on attribute domains
   Similarity map: a semilattice version
  - Data Dependencies with similarity maps

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

• Main results

3 Conclusion and perspective

#### Important known results for FD

Let F be a set of FD over a schema R  $CL(F) = \{X \subseteq R | X_F^+ = X\}$ : a closure system of F IRR(F) the set of irreducible elements of CL(F) by intersection

Reasoning on F is **equivalent** to reasoning on CL(F), for instance:

$$X_F^+ = \{A \in R \mid F \models X \to A\} = \cap \{Y \in CL(F) \mid X \subseteq Y\}$$

Let *r* be a relation over *R*. The **agree set** of *r* is  $ag(r) = \{ag(t_1, t_2) \mid t_1, t_2 \in r\}$  where  $ag(t_1, t_2) = \{A \in R \mid t_1[A] = t_2[A]\}$ 

r is an Armstrong relation for F iff  $IRR(F) \subseteq ag(r) \subseteq CL(F)$  [8]

RQL query language

Preliminaries

#### Example

	Bar(B)	Beer(Be)	Price(P)
$t_1$	Nota bene	Adelscott	2
t <sub>2</sub>	Montagne	1664	1.5
t <sub>3</sub>	Nota bene	1664	2
t <sub>4</sub>	Ritz	Adelscott	5
t <sub>5</sub>	Café Flore	Affligen	6

$$F = \{B \rightarrow P, P \rightarrow B\}$$
  

$$CL(F) = \{\emptyset, Be, BP, BBeP\}$$
  

$$IRR(F) = \{Be, BP\}$$

$$ag(r) = \{\emptyset, Be, BP\}$$
, often represented as:

#### Towards a rule query language

Focus on rules equivalent to implications (or FD)  $\Rightarrow$  Armstrong axioms (reflexivity, augmentation, transitivity) have to be sound and complete

**Idea**: Defining a rule query language (RQL) such that every RQL statement turns out to deliver **implications** 

・ロ> <同> <目> <目> <目> <同> <</li>

Require to identify syntactic constraints such that we remain within the reasoning of implications

#### Semantics of implications

Let  $b_0$  be a binary relation (given by a  $\{0, 1\}$ -relation)  $b_0 \models X \rightarrow Y \Leftrightarrow \forall t \in b_0$  $(\forall A \in X \ t.A = 1) \Rightarrow (\forall A \in Y \ t.A = 1)$ 

Let  $d = \{r_0, r_1, ..., r_n\}$  be a relational database  $r_0 \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in r_0$  $(\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$ 

 $d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_p}))$  $(\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$ 

$$d \models X \to Y \Leftrightarrow \forall t_1 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_n})), \\ \forall t_2 \in \pi_X(\sigma_{F'}(r_{j_0} \bowtie \ldots \bowtie r_{i_n})) \\ \text{such that } (t_1.rank = t_2.rank + 1) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A \in Y \ t_1$$

#### Semantics of implications

Let 
$$b_0$$
 be a binary relation (given by a  $\{0, 1\}$ -relation)  
 $b_0 \models X \rightarrow Y \Leftrightarrow \forall t \in b_0$   
 $(\forall A \in X \ t.A = 1) \Rightarrow (\forall A \in Y \ t.A = 1)$ 

Let 
$$d = \{r_0, r_1, ..., r_n\}$$
 be a relational database  
 $r_0 \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in r_0$   
 $(\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$ 

$$d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_{i_0} \bowtie \dots \bowtie r_{i_p})) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$$

$$d \models X \rightarrow Y \Leftrightarrow \forall t_1 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_n})), \\ \forall t_2 \in \pi_X(\sigma_{F'}(r_{j_0} \bowtie \ldots \bowtie r_{i_n})) \\ \text{such that } (t_1.rank = t_2.rank + 1) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A)$$

#### Semantics of implications

Let 
$$b_0$$
 be a binary relation (given by a  $\{0, 1\}$ -relation)  
 $b_0 \models X \rightarrow Y \Leftrightarrow \forall t \in b_0$   
 $(\forall A \in X \ t.A = 1) \Rightarrow (\forall A \in Y \ t.A = 1)$ 

Let 
$$d = \{r_0, r_1, ..., r_n\}$$
 be a relational database  
 $r_0 \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in r_0$   
 $(\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$ 

$$d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_p})) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$$

$$d \models X \to Y \Leftrightarrow \forall t_1 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_n})), \\ \forall t_2 \in \pi_X(\sigma_{F'}(r_{j_0} \bowtie \ldots \bowtie r_{i_n})) \\ \text{such that } (t_1.rank = t_2.rank + 1) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A \in Y \ t_1.A = t_2.A) \\ (\forall A \in Y \ t_1.A \in Y \$$

#### Semantics of implications

Let 
$$b_0$$
 be a binary relation (given by a  $\{0, 1\}$ -relation)  
 $b_0 \models X \rightarrow Y \Leftrightarrow \forall t \in b_0$   
 $(\forall A \in X \ t.A = 1) \Rightarrow (\forall A \in Y \ t.A = 1)$ 

Let 
$$d = \{r_0, r_1, ..., r_n\}$$
 be a relational database  
 $r_0 \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in r_0$   
 $(\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$ 

$$d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_p})) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$$

$$d \models X \to Y \Leftrightarrow \forall t_1 \in \pi_X(\sigma_F(r_{i_0} \bowtie \ldots \bowtie r_{i_n})), \\ \forall t_2 \in \pi_X(\sigma_{F'}(r_{j_0} \bowtie \ldots \bowtie r_{i_n})) \\ \text{such that } (t_1.rank = t_2.rank + 1) \\ (\forall A \in X \ t_1.A = t_2.A) \Rightarrow (\forall A \in Y \ t_1.A = t_2.A)$$

#### Semantics of implications (cont'ed)

$$d \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_0 \bowtie \dots \bowtie r_n)) (\forall A \in X(2 * ABS(t_1.A - t_2.A)/(t_1.A + t_2.A) < 0.1)) \Rightarrow (\forall A \in Y(2 * ABS(t_1.A - t_2.A)/(t_1.A + t_2.A) < 0.1))$$

 $d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_0 \bowtie \ldots \bowtie r_n)) \\ (\forall A \in X \ t_1.A \le t_2.A) \Rightarrow (\forall A \in Y \ t_1.A \le t_2.A)$ 

 $\begin{aligned} r_0 &\models X \to Y \Leftrightarrow \forall t_1, t_2, t_3 \in r_0 \\ & (\forall A \in X(t_1.A \leq t_2.A) \land (t_3.A \leq t_2.A)) \\ & \Rightarrow (\forall A \in Y(t_1.A \leq t_2.A) \land (t_3.A \leq t_2.A)) \end{aligned}$ 

#### Semantics of implications (cont'ed)

$$d \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_0 \bowtie \dots \bowtie r_n)) (\forall A \in X(2 * ABS(t_1.A - t_2.A)/(t_1.A + t_2.A) < 0.1)) \Rightarrow (\forall A \in Y(2 * ABS(t_1.A - t_2.A)/(t_1.A + t_2.A) < 0.1))$$

$$d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_0 \bowtie \ldots \bowtie r_n)) \\ (\forall A \in X \ t_1.A \leq t_2.A) \Rightarrow (\forall A \in Y \ t_1.A \leq t_2.A)$$

 $\begin{aligned} r_0 &\models X \rightarrow Y \Leftrightarrow \forall t_1, t_2, t_3 \in r_0 \\ & (\forall A \in X(t_1.A \leq t_2.A) \land (t_3.A \leq t_2.A)) \\ & \Rightarrow (\forall A \in Y(t_1.A \leq t_2.A) \land (t_3.A \leq t_2.A)) \end{aligned}$ 

#### Semantics of implications (cont'ed)

$$d \models X \rightarrow Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_0 \bowtie \dots \bowtie r_n)) (\forall A \in X(2 * ABS(t_1.A - t_2.A)/(t_1.A + t_2.A) < 0.1)) \Rightarrow (\forall A \in Y(2 * ABS(t_1.A - t_2.A)/(t_1.A + t_2.A) < 0.1))$$

$$d \models X \to Y \Leftrightarrow \forall t_1, t_2 \in \pi_X(\sigma_F(r_0 \bowtie \ldots \bowtie r_n)) \\ (\forall A \in X \ t_1.A \leq t_2.A) \Rightarrow (\forall A \in Y \ t_1.A \leq t_2.A)$$

$$\begin{aligned} r_0 &\models X \rightarrow Y \Leftrightarrow \forall t_1, t_2, t_3 \in r_0 \\ & (\forall A \in X(t_1.A \leq t_2.A) \land (t_3.A \leq t_2.A)) \\ & \Rightarrow (\forall A \in Y(t_1.A \leq t_2.A) \land (t_3.A \leq t_2.A)) \end{aligned}$$

Bridging the Gap between Data Diversity and Data Dependencies RQL query language Main result underlying RQL

#### Approach and contribution

Replaying part of the story underlying SQL and relational languages, especially through Tuple Relational Calculus (TRC)

What we did:

- Extend TRC to support rule expression (SafeRL logical language, see [14] for details)
- Propose a new syntactic practical language (RQL) from SafeRL

$$Q = \{ X \to Y \mid \forall t_1 \dots \forall t_n \left[ \psi(t_1, \dots, t_n) \to (\forall A \in X(\delta(A, t_1, \dots, t_n)) \to \forall A \in Y(\delta(A, t_1, \dots, t_n))) \right] \}$$

Bridging the Gap between Data Diversity and Data Dependencies RQL query language Main result underlying RQL

Main result.

#### THM

Let Q be a RQL query over a database d.

- 1. ans(Q, d) defines a closure system CL(Q) over sch(Q)
- There exists a SQL query Q' over d such that Q' computes a base B(Q) of CL(Q), i.e. IRR(Q) ⊆ B(Q) ⊆ CL(Q)
  - B(Q): agree sets for FD and binary relation for implications
  - Proof of 1. similar to the proof given for Functional Dependencies by Mannila and Raiha 1994 [21], Demetrovics and Thi 1995 [16].
  - Proof of 2. a bit more elaborated

Bridging the Gap between Data Diversity and Data Dependencies RQL query language The RQL language

```
RQL: a Practical Language
```

RQL has 5 clauses (with the "look and feel" of SQL):

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

```
FINDRULES
OVER A_1, ..., A_n
SCOPE t_1(SQL_1), ..., t_n(SQL_n)
WHERE condition(t_1, ..., t_n)
CONDITION ON A IS \delta(A, t_1, ..., t_n)
```

Bridging the Gap between Data Diversity and Data Dependencies RQL query language The RQL language

Examples

#### FINDRULES OVER Empno,Lastname,Workdept,Job,Sex,Bonus

```
SCOPE t1,t2 Emp
CONDITION ON A IS t1.A = t2.A;
```

FINDRULES OVER Empno, Lastname, Workdept, Job, Sex, Bonus, Mgrno SCOPE t1 Emp CONDITION ON A IS t1.A IS NULL

Bridging the Gap between Data Diversity and Data Dependencies RQL query language The RQL language

Examples

FINDRULES
OVER ...
SCOPE t1,t2,t3 sensors
WHERE t2.time = t1.time+interval 1 minute AND
t3.time = t2.time+interval 1 minute
CONDITION ON A IS t1.A < t2.A AND t2.A > t3.A;

Bridging the Gap between Data Diversity and Data Dependencies  $\mathsf{RQL}$  query language

RQL implementation

#### RQL query processing.

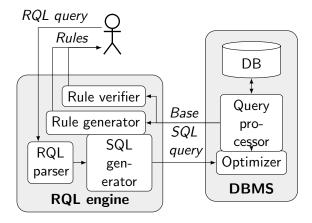


Figure: RQL queries processing overview

RQL query language RQL implementation

#### RQL Web Interface

bob@denar.com	≣ Query • ● Help • ● Log out ●About Sample DB	
	In Sample mode, you have a database filled with data on which you can try s are given below.	ome queries. Samples 🛛 🗙
	On the left you will find the list of the tables and views you have access to. F RQL and SQL queries on them!	eel free to try your own
	Note that the first queries are SQL and will give you informations about the d	lata in the database.
	Learn more about Sample mode	
		SQL examples:
TABLES	Submit your RQL or SQL	← SQL1 Content of Emp
I DEPT	query:	+ SQL 2 Schema of Emp
	FINDRULES OVER Educlevel, Sal, Bonus, Comm	RQL examples:
EWb	SCOPE t1, t2 Emp WHERE t1. Empro	+ RQL 1 Null values in Emp
VIEWS	CONDITION ON A IS t1.A >= t2.A	RQL 2 Functional dependencies on Emp
EMP_SUBSET		+ RQL3 Functional dependencies on a subset of Emp
		+ RQL4 Approximate functional dependencies on Emp
EMP_WITH_DEPTNAME		RQL5 Sequential dependencies on Emp
	Submit Query	RQL 6 Sequential dependencies for male employees
		RQL7 Sequential dependencies on manager and managees
		RQL 8 Null values in Dept

#### Figure: RQL Interface

Bridging the Gap between Data Diversity and Data Dependencies RQL query language RQL implementation

#### **RQL** Web Interface

The ru	le Sal Educlevel 🗕 Bo	nus is false							
Count	er-example:								
EMPN	O LASTNAME	WORKDEPT	JOB	EDUCLEVEL	SEX	SAL	BONUS	сомм	MGRNO
10	SPEN	C01	FINANCE	18	F	52750	500	4220	20
20	THOMP	null	MANAGER	18	М	41250	800	3300	null
Gener	ated query:								
	SELECT t1.*, t2.*								
	FROM Emp t1, Emp	t2							
	WHERE (t1.Sal >= )	t2.Sal AND t1.Edu	clevel >= t2.Ed	iuclevel)					
	AND CASE WHEN (t1	.Bonus >= t2.Bonu	(S) THEN 1 ELSE	0  END = 0					
	AND rownum <= 1								

#### Figure: Counter-example with RQL

#### Summary

- RQL: a practical language to express different semantics for implication
- Discovery of implications seen as a *query processing* problem
- Side effect: data analysts may interact with their data through counter-examples
- Advantages
  - Easy to learn for SQL-aware data analysts (especially CS students !)
  - http://rql.insa-lyon.fr

### Contents

#### RQL query language

- Preliminaries
- Main result underlying RQL
- The RQL language
- RQL implementation
- Summary

#### 2 Structural properties on attribute domains

- Similarity map: a semilattice version
- Data Dependencies with similarity maps

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

• Main results

#### **3** Conclusion and perspective

#### Come back to functional dependencies

$$r \models X \rightarrow Y$$
 iff for all  $t_1, t_2 \in r$   
If for all  $A \in X, t_1[A] = t_2[A]$  then for all  $B \in Y, t_1[B] = t_2[B]$ 

Let us focus on the equality  $t_1[A] = t_2[A]$  without defining new predicates on  $t_1[A]$  and  $t_2[A]$  values

◆□ ▶ ◆□ ▶ ◆ □ ▶ ◆ □ ▶ ● □ ● ● ●

### From equality to similarity

Two possibilities:

- Replace " $t_1[A] = t_2[A]$ " by " $t_1[A]$  is similar to  $t_2[A]$ "
  - $\Rightarrow$  Similarity seen as a *reflexive and symmetric binary relation*
- Replace "t<sub>1</sub>[A] = t<sub>2</sub>[A]" by "t<sub>1</sub>[A] and t<sub>2</sub>[A] are similar to some similarity value s"

 $\Rightarrow$  Similarity seen as an *idempotent and commutative map* 

 $\Rightarrow$  Focus on similarity map which appears to be less restrictive than similarity relation

#### Similarity relation

Let  $\mathcal{D}_A$  be the domain of attribute A and  $u, v \in \mathcal{D}_A$ Let S be a binary relation on  $\mathcal{D}_A$ 

#### Similarity

S is a similarity relation if S is reflexive (S(u, u) = 1) and symmetric (S(u, v) = S(v, u)).

S subsumes the equality operator Two meaningful values: true (1) and false (0) Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains

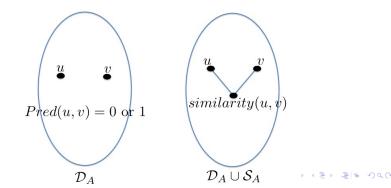
# Assumptions on similarity map

Notations:

• A is an attribute,  $\mathcal{D}_A$  its domain

•  $S_A$  new values denoting similarities for A (disjoint from  $D_A$ ) Assumption:

• For any subset of  $\mathcal{D}_A \cup \mathcal{S}_A$ , there is a **unique** similarity value.



#### Similarity map: a semilattice version

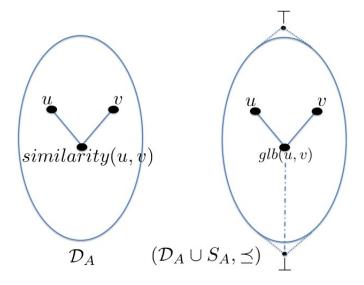
Let A be an attribute,  $S = D_A \cup S_A$  and  $m_A : S \times S \rightarrow S$  a similarity map that is:

- Idempotent  $(m_A(a, a) = a \text{ for all } a \in S)$ ,
- Commutative  $(m_A(a,a')=m_A(a',a)$  for all  $a,a'\in S)$ ,
- Associative  $(m_A(a, m_A(a', a'')) = m_A(m_A(a, a'), a'')$  for all  $a, a', a'' \in S$ ).

 $m_A$  induces a **partial order**  $\leq$  on S: for every  $a, a' \in S, a \leq a'$  whenever  $m_A(a, a') = a$ .

 $(S, \preceq)$  is a semilattice where  $glb(a, a') = m_A(a, a')$  for all  $a, a' \in S$ .

#### Illustration



## Example with numerical interval values

Consider an attribute A whose domain is intervals of integer, i.e.  $\mathcal{D}_A = \{[i,j] | i, j \in 1..n, i \leq j\}$ 

- What would be the similarity values S<sub>A</sub> ?
   ⇒ The set of closed sets of D<sub>A</sub> by intersection
- Let  $\{I_1, \ldots, I_m\} \subseteq \mathcal{D}_A \cup \mathcal{S}_A$ . Similarity value of  $\{I_1, \ldots, I_m\}$ ?  $\Rightarrow$  its intersection  $I = \bigcap \{I_1, \ldots, I_m\}$  $\Rightarrow I$  is clearly unique

#### Two examples of similarity map

Equality can be defined as:

$$m_A(x,y) = \begin{cases} x & \text{if } x = y \\ \bot & \text{otherwise} \end{cases}$$

 $\perp$  means " not similar" or 0 (false)

Similarity over intervals can be defined as:

$$m_{\mathcal{A}}(I_1, I_2) = \begin{cases} I_1 \cap I_2 & \text{if } I_1 \cap I_2 \neq \emptyset \\ \bot & \text{otherwise} \end{cases}$$

## Underlying assumption

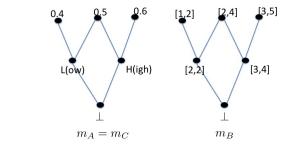
A dataset r has to be equipped with a semilattice structure for every attribute domain

 $\Rightarrow$  Allow to be as close as possible of data values to quantify their similarities and differences

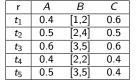
 $\Rightarrow$  Require an important data pre-processing task, that could be partially automated using data mining techniques

A different approach to address data diversity

## Running example



◆□ ▶ ◆□ ▶ ◆ □ ▶ ◆ □ ▶ ●□ ● ● ●



 $\Rightarrow$  Semantics for  $m_A$  and  $m_C$ 

- The values L and H qualify the different values
- $\perp$  otherwise, i.e. **not similar**.

# Application to functional dependencies

 $r \models X \rightarrow Y$  iff for all  $t_1, t_2 \in r$ for all  $A \in X, t_1[A] = t_2[A] \Rightarrow$  for all  $B \in Y, t_1[B] = t_2[B]$ 

can be reformulated as follows:

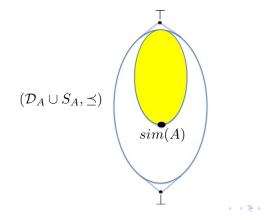
for all  $A \in X$ ,  $glb(t_1[A], t_2[A]) \neq \bot \Rightarrow$ for all  $B \in Y$ ,  $glb(t_1[B], t_2[B]) \neq \bot$ 

 $glb(t_1[A], t_2[A]) \neq \bot$  means there exists a similarity between the values of A on  $t_1, t_2$ 

# Minimal degree of similarities

Assume now an expert provides for each attribute A a minimal degree of similarity she expects.

Let  $sim : sch(r) \to (\mathcal{D}_A \cup \mathcal{S}_A) \setminus \{\bot\}$  be such a map.



#### Examples

$$r \models X \to Y \text{ iff for all } t_1, t_2 \in r$$
  
for all  $A \in X, t_1[A] = t_2[A] \Rightarrow$  for all  $B \in Y, t_1[B] = t_2[B]$ 

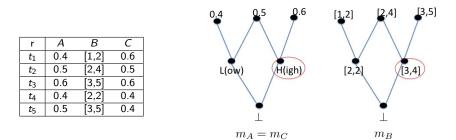
becomes

$$r \models_{sim} X \to Y \text{ iff for all } t_1, t_2 \in r$$
  
for all  $A \in X, sim(A) \preceq glb(t_1[A], t_2[A]) \Rightarrow$   
for all  $B \in Y, sim(B) \preceq glb(t_1[B], t_2[B])$ 

 $sim(A) \leq glb(t_1[A], t_2[A])$  means the similarity level between the values of A on  $t_1, t_2$  is above the mimimum

#### Example

Assume the expert tags those similarities: sim(A) = sim(C) = Hand sim(B) = [3,4]



< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

 $r \models_{sim} A \to B$  (or  $r \models_{sim} A$ ,  $High \to B$ , [3, 4])  $r \nvDash_{sim} C \to B \Rightarrow$  for ex. see counter-example  $t_1, t_2$  Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains Main results

#### Many results follow ...

Many well-known results on FD can be re-defined in this new setting

Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains Main results

Agree sets

Agree sets can be extended naturally: instead of getting a set of attributes (due to 0 and 1 interpretation values based on equality), we obtain a set of similarities

$$egin{aligned} & \mathsf{ag}(r) = \{ \mathsf{ag}(t_1, t_2) \mid t_1, t_2 \in r \} \ & \mathsf{ag}(t_1, t_2) = \{ \mathsf{ag}(t_1[A], t_2[A]) \mid A \in \mathsf{sch}(r) \} \ & \mathsf{ag}(t_1[A], t_2[A]) = \mathsf{glb}(t_1[A], t_2[A]) \end{aligned}$$

Example:  $ag(t_1, t_2) = < L, [2, 2], H >$ 

Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains

Main results

#### Example

r	A	В	С
$t_1$	0.4	[1,2]	0.6
t <sub>2</sub>	0.5	[2,4]	0.5
t <sub>3</sub>	0.6	[3,5]	0.6
t <sub>4</sub>	0.4	[2,2]	0.4
t <sub>5</sub>	0.5	[3,5]	0.4

ag(r)	A	В	С
$ag(t_1, t_2)$	L	[2,2]	Н
$ag(t_1, t_3)$		$\perp$	0.6
$ag(t_1, t_4)$	0.4	[2,2]	$\perp$
$ag(t_1, t_5)$	L	$\perp$	$\perp$
$ag(t_2,t_3)$	н	[3,4]	Н
$ag(t_2, t_4)$	L	[2,2]	L
$ag(t_2, t_5)$	0.5	[3,4]	L
$ag(t_3, t_4)$		$\perp$	$\perp$
$ag(t_3, t_5)$	Н	[3,5]	$\perp$
$ag(t_4, t_5)$	L	Ĺ	0.4

From ag(r), two interesting cases:

- replacing all values occurring in r by 1 and all other values by  $0 \Rightarrow$  classical FD with equality
- replacing ⊥ by 0 (or false) and all other values by 1 (true)
   ⇒ classical FD extended to similarities

Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains Main results

#### Closures and agree sets

From the agree set of *r*, the family  $\mathcal{F}_r$  of closed sets by the glb operation is:  $\mathcal{F}_r = \{g|b_{\prec_{sch(r)}}(T)|T \subseteq ag(r)\}$ 

#### Lemma

 $(\mathcal{F}_r, \preceq_{sch(r)})$  is a semilattice

Let  $\mathcal{M}(\mathcal{F}_r)$  be the meet irreducible elements of  $\mathcal{F}_r$ 

#### Theorem

$$\mathcal{M}(\mathcal{F}_r) \subseteq \mathsf{ag}(r) \subseteq \mathcal{F}_r$$

Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains Main results

#### Similarity, attribute closure and implications

Let  $\mathcal{F}$  be a family of closed sets,  $X \subseteq sch(r)$  and  $sim(X) = \{sim(A) | A \in X\}$ 

$$X^+_{sim(X)} = glb(\{Y \in \mathcal{F} \mid sim(X) \preceq_X Y\})$$

◆□ ▶ ◆◎ ▶ ◆目 ▶ ◆目 ● ● ● ● ●

#### Theorem

$$r \models_{sim} X \to Y \text{ iff } sim(Y) \preceq_Y X^+_{sim(X)}$$

Bridging the Gap between Data Diversity and Data Dependencies

Structural properties on attribute domains

Main results

# Example with $r \models_{sim} A \rightarrow B$ with sim(A) = H and sim(B) = [3, 4]

r	A	В	С
t <sub>1</sub>	0.4	[1,2]	0.4
t <sub>2</sub>	0.5	[2,4]	0.5
t <sub>3</sub>	0.6	[3,5]	0.6
t <sub>4</sub>	0.4	[2,2]	0.4
t5	0.5	[3,4]	0.4

ag(r)	A	В	С
$ag(t_1, t_2)$	L	[2,2]	Н
$ag(t_1, t_3)$	1	$\perp$	0.6
$ag(t_1, t_4)$	0.4	[2,2]	$\perp$
$ag(t_1, t_5)$	L	$\perp$	L
$ag(t_2, t_3)$	н	[3,4]	Н
$ag(t_2, t_4)$	L	[2,2]	L
$ag(t_2, t_5)$	0.5	[3,4]	L
$ag(t_3, t_4)$	1	T	$\perp$
$ag(t_3, t_5)$	н	[3,4]	$\perp$
$ag(t_4, t_5)$	L	$\perp$	0.4

◆□ → ◆□ → ◆三 → ◆三 → ●目 → ○○ ○

$$\begin{array}{l} A^+_{sim(A)} = glb_{\preceq_{ABC}} \{ < H, [3,4], H >, < 0.5, [3,4], L >, < \\ H, [3,4], \bot > \} = < H, [3,4], \bot > \end{array}$$

 $sim(B) \preceq_B < H, [3, 4], \bot >$ 

$$\Rightarrow$$
 *r*  $\models_{sim} A$ , *High*  $\rightarrow$  *B*, [3, 4]

Bridging the Gap between Data Diversity and Data Dependencies Structural properties on attribute domains Main results

#### Summary

- Using similarity maps on attribute domains allows to reconsider classical data dependencies
- Require to change our mind: most of the effort has to be done at the attribute domain level to define similarity map
- After this, the problem is embedded into a lattice structure allowing to revisit many known results

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

# Contents

#### RQL query language

- Preliminaries
- Main result underlying RQL
- The RQL language
- RQL implementation
- Summary

# 2 Structural properties on attribute domains

- Similarity map: a semilattice version
- Data Dependencies with similarity maps
- Main results

#### 3 Conclusion and perspective

# Conclusion

Two propositions to extend data dependencies

- First, through RQL, a query language devoted to implications (or FD)
- Second, through assumptions on attribute domains using semilattice structure induced by similarity maps

 $\Rightarrow$  Both are elegant formalisms to extend functional dependencies by taking into account data diversity

#### Perspective

#### Theoretical question

 $\Rightarrow$  Under which conditions the second approach leads to implications (Armstrong axioms) ?

#### Practical question

 $\Rightarrow$  Given a dataset D equipped with semilattice structures, how to discover implications satisfied in D ?

◆□ → ◆□ → ◆三 → ◆三 → ●目 → ○○ ○

## Acknowledgments

Joint work with Brice Chardin, Emmanuel Coquery, Marie Pailloux on RQL (partly funded by ANR, DAG project)



and Lhouari Nourine on structural properties on attribute domains (partly funded by the CNRS Mastodon program on *Data Quality*)



Bridging the Gap between Data Diversity and Data Dependencies Appendix

For Further Reading

## For Further Reading I



#### S. Abiteboul, R. Hull, and V. Vianu,

Foundations of Databases. Addison-Wesley, 1995.



B. Ganter and R. Wille,

Formal Concept Analysis. Springer, 1999.



Marie Agier-Pailloux, Jean-Marc Petit, Einoshin Suzuki: Towards Ad-Hoc Rule Semantics for Gene Expression Data. ISMIS 2005: 494-503



M. Agier-Pailloux, J.-M. Petit, and E. Suzuki, Unifying framework for rule semantics: Application to gene expression data, *Fundam. Inform.*, vol. 78, no. 4, pp. 543–559, 2007.



Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499



Jaume Baixeries and Victor Codocedo and Mehdi Kaytoue and Amedeo Napoli, Characterizing approximate-matching dependencies in formal concept analysis with pattern structures,

Discrete Applied Mathematics, 2018

Bridging the Gap between Data Diversity and Data Dependencies Appendix For Further Reading

For Further Reading II

F. Baklouti and G. Levy and R. Emilion A fast algorithm for general Galois lattices building. Elec. J. Symbolic data analysis, Vol. 2, N 1, 19-31, 2005



C. Beeri, M. Dowd, R. Fagin, and R. Statman. On the structure of Armstrong relations for functional dependencies. JACM, 31(1):30-46, 1984.



Bertossi, Leopoldo and Kolahi, Solmaz and Lakshmanan, Laks V. S., Data Cleaning and Query Answering with Matching Dependencies and Matching Functions. ICDT 2011, pp. 268-279



Loredana Caruccio ; Vincenzo Deufemia ; Giuseppe Polese Relaxed Functional Dependencies: A Survey of Approaches IEEE TKDE, vol 28, issue 1. 2016



#### N. Caspard and B. Monjardet,

The lattices of closure systems, closure operators, and implicational systems on a finite set: A survey.

Discrete Applied Mathematics, vol. 127, no. 2, pp. 241-269, 2003.

Bridging the Gap between Data Diversity and Data Dependencies Appendix

For Further Reading

## For Further Reading III



Upen S. Chakravarthy, John Grant and Jack Minker Logic-based approach to semantic query optimization ACM Transactions on Database Systems (TODS) TODS Homepage archive Volume 15 Issue 2, June 1990, pages 162-207 Brice Chardin, Emmanuel Coquery, Marie Pailloux, Jean-Marc Petit: RQL: A SQL-Like Query Language for Discovering Meaningful Rules. IEEE ICDM demo 2014: 1203-1206 Brice Chardin and Emmanuel Coquery and Marie Pailloux and JM Petit, RQL: A Query Language for Rule Discovery in Databases, Theor. Comput. Sci., vol. 658, pp. 357–374, 2017, Xu Chu, Ihab F. Ilyas, Paolo Papotti: Discovering Denial Constraints. PVLDB 6(13): 1498-1509 (2013) L Demetrovics and V D Thi Some remarks on generating Armstrong and inferring functional dependencies

relation.

Acta Cybernetica, vol. 12, no. 2, pp. 167-180, 1995.

Bridging the Gap between Data Diversity and Data Dependencies Appendix

For Further Reading

## For Further Reading IV



Wenfei Fan and Xibei Jia and Jianzhong Li and Shuai Ma, Reasoning about Record Matching Rules, *PVLDB*, vol 2, num 1, pp. 407–418, 2009



L. Jezková, Pablo Cordero, and Manuel Enciso. Fuzzy functional dependencies: A comparative survey. *Fuzzy Sets and Systems*, 317:88–120, 2017.

Abo Khamis, Mahmoud and Ngo, Hung Q. and Nguyen, XuanLong and Olteanu, Dan and Schleich, Maximilian: In-Database Learning with Sparse Tensors, PODS '18, pp. 325–340



S. Lopes, J.-M. Petit, and L. Lakhal,

Functional and approximate dependency mining: database and  $\ensuremath{\operatorname{FCA}}$  points of view,

(ロ) (同) (E) (E) (E) (O)

J. Exp. Theor. Artif. Intell., vol. 14, no. 2-3, pp. 93-114, 2002.

Heikki Mannila, Kari-Jouko Rih: Algorithms for Inferring Functional Dependencies from Relations. Data Knowl. Eng. 12(1): 83-99 (1994) Bridging the Gap between Data Diversity and Data Dependencies Appendix For Further Reading

#### Thank you

Merci

σας ευχαριστώ