Relational Concept Analysis: Mining Multi-relational Datasets for Assisted Class Model

Marianne Huchard

▶ **To cite this version:**

Marianne Huchard. Relational Concept Analysis: Mining Multi-relational Datasets for Assisted Class Model. 2014. lirmm-01075525

HAL Id: lirmm-01075525
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01075525

Submitted on 17 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Relational Concept Analysis (RCA)

Mining multi-relational datasets
Applied to class model evolution

SATToSE 2014

Marianne Huchard

July 11, 2014

An introduction to RCA

Brief presentation of FCA – Formal Concept Analysis

A methodology for:

- data analysis, data mining
- knowledge representation
- unsupervised learning

Roots:

- lattice theory, Galois correspondences (Birkhoff, 1940; Barbut & Monjardet, 1970)
- concept lattices (Wille, 1982)

Brief presentation of FCA – Formal Concept Analysis

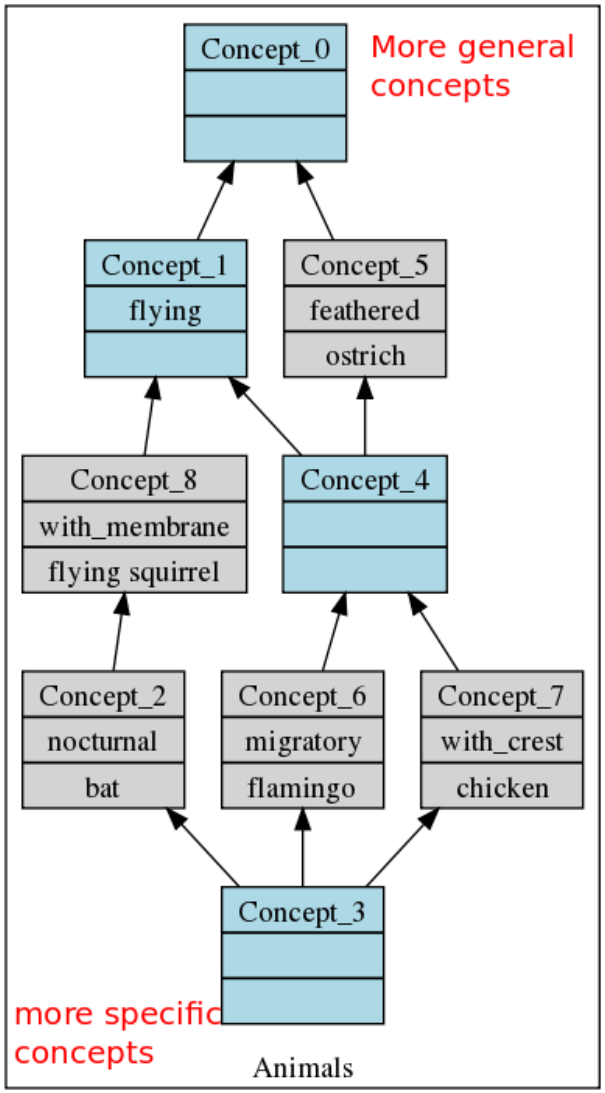
Contexts and concepts

- Handled data
  - entities with characteristics
  - provided with a Formal Context (a binary table)

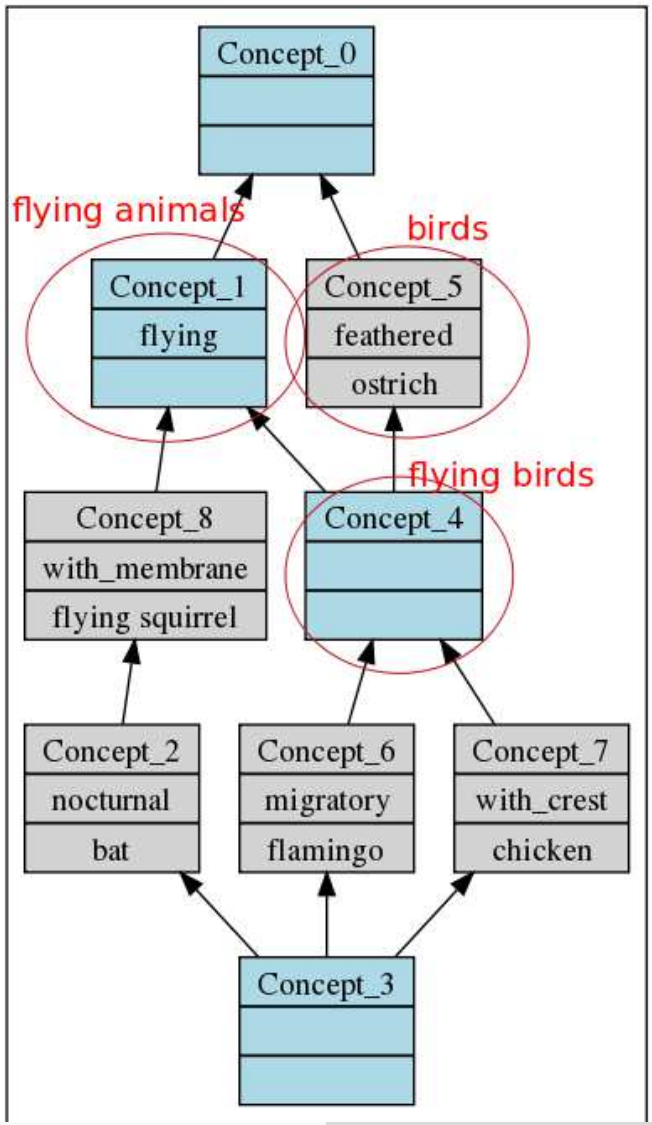
	flying	nocturnal	feathered	migratory	with_crest	with_membrane
 flying squirrel	X					X
 bat	X	X				X
 ostrich			X			
 flamingo	X		X	X		
 chicken	X		X		X	

- Concept : maximal group of entities sharing characteristics
- Concept lattice : concepts with a partial order relation

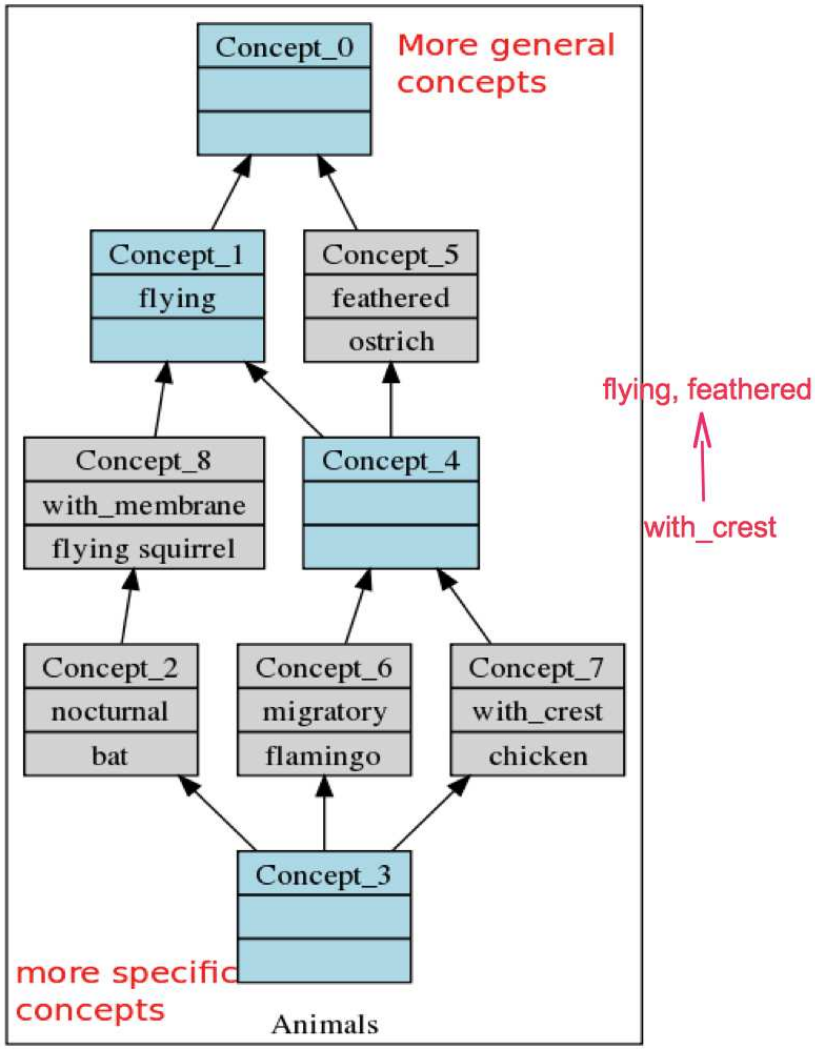
Brief presentation of FCA – Formal Concept Analysis



Brief presentation of FCA – Formal Concept Analysis



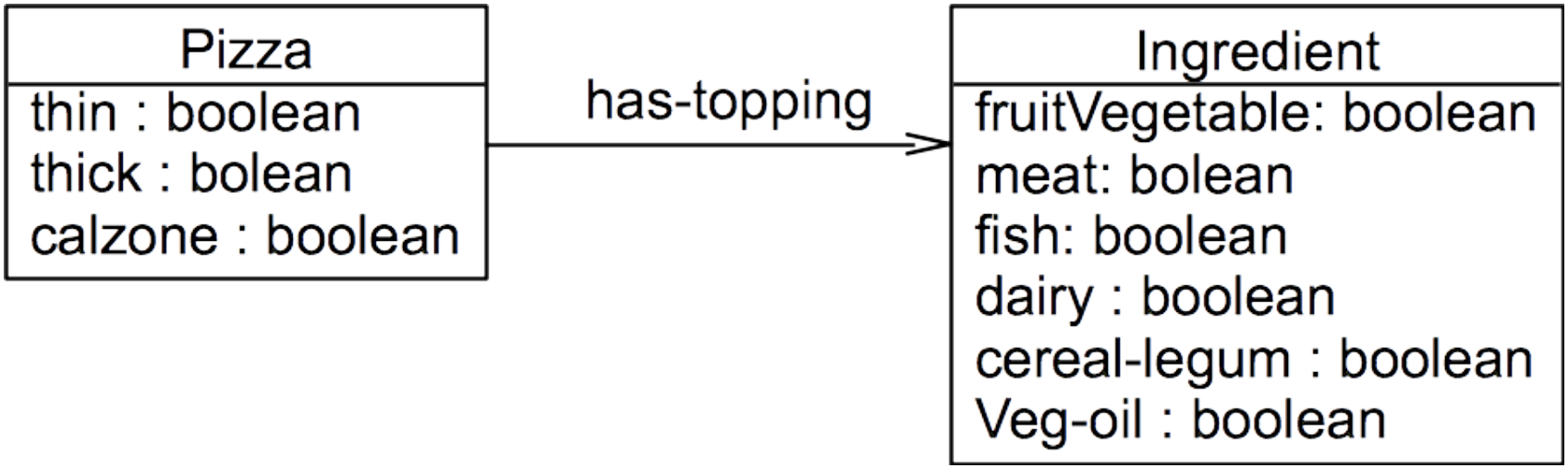
Brief presentation of FCA – Formal Concept Analysis



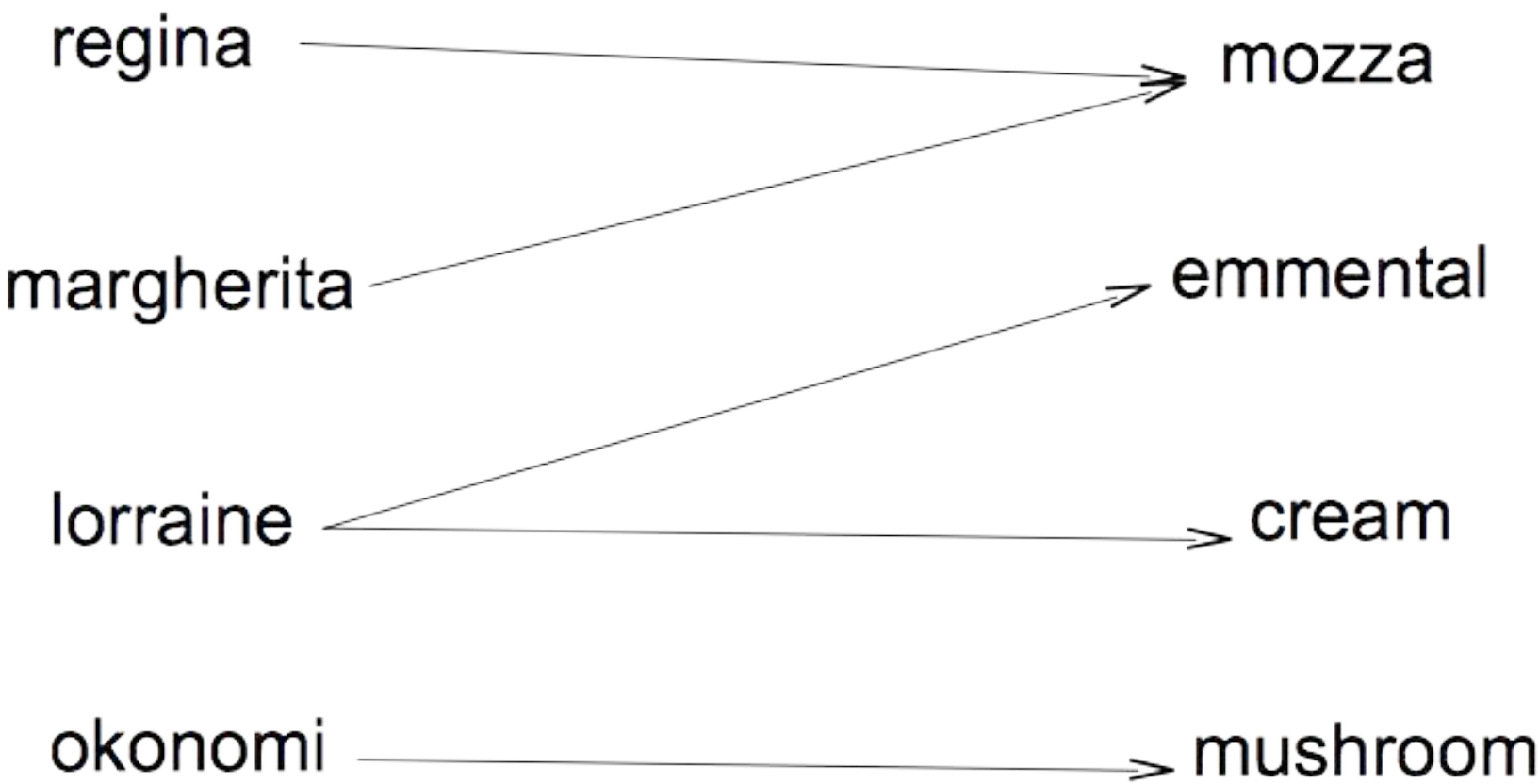
FCA and complex data

- many-valued contexts (integers, floats, terms, structures, symbolic objects, intervals, etc.)
  (Ganter/Wille, Polaillon, ...)
- fuzzy descriptions (Yahia et al., Belohlavek, ...)
- hierarchies on values (Godin et al., Carpineto/Romano, ...)
- logical description (Chaudron et al., Ferré et al., ...)
- graphs (Liquière, Prediger/Wille, Ganter/Kuznetsov, ...)
- **Multi-relational data** (Priss, Hacène-Rouane et al., ...)
- etc.

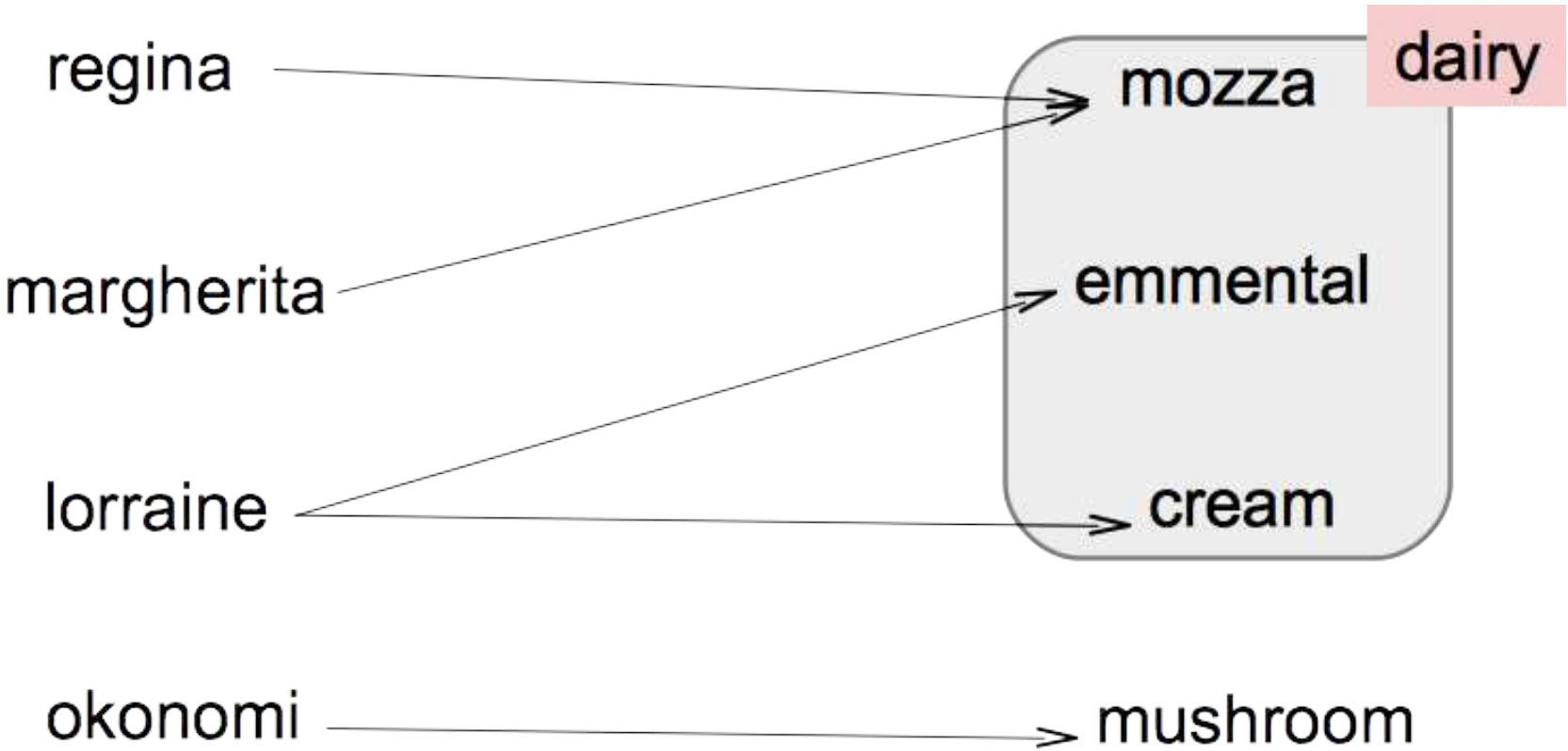
A flavor of Relational Concept Analysis



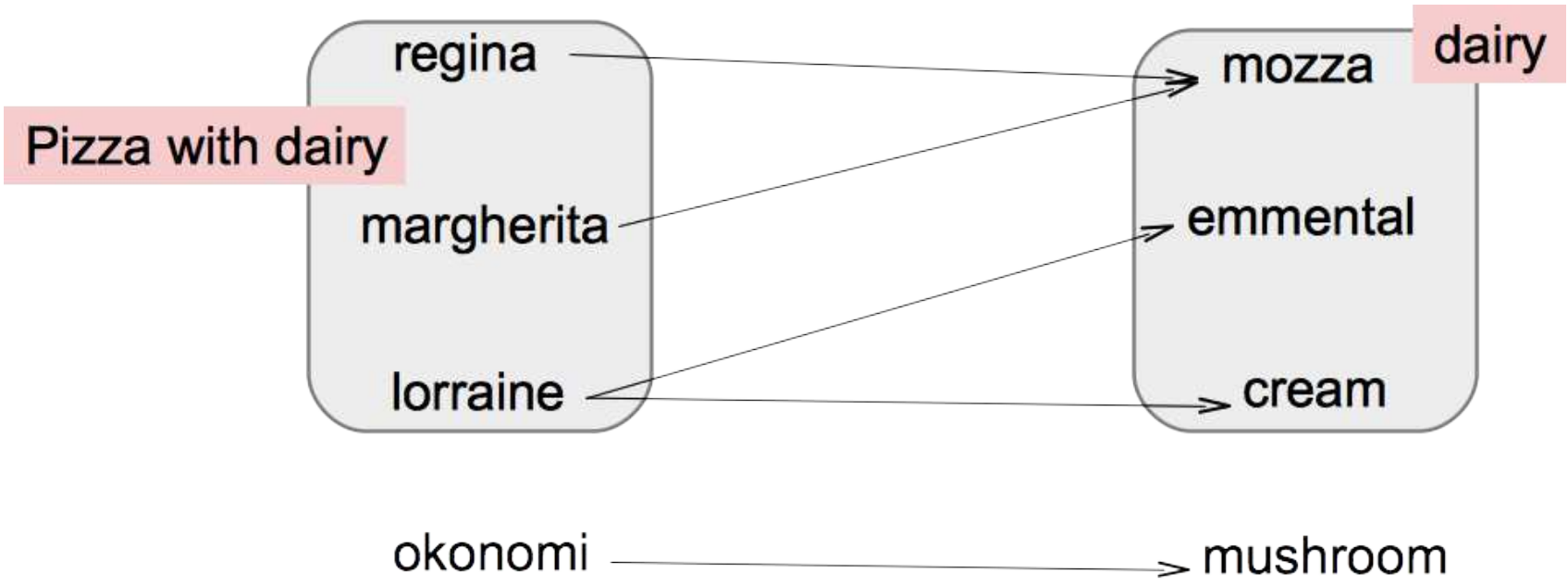
A flavor of Relational Concept Analysis



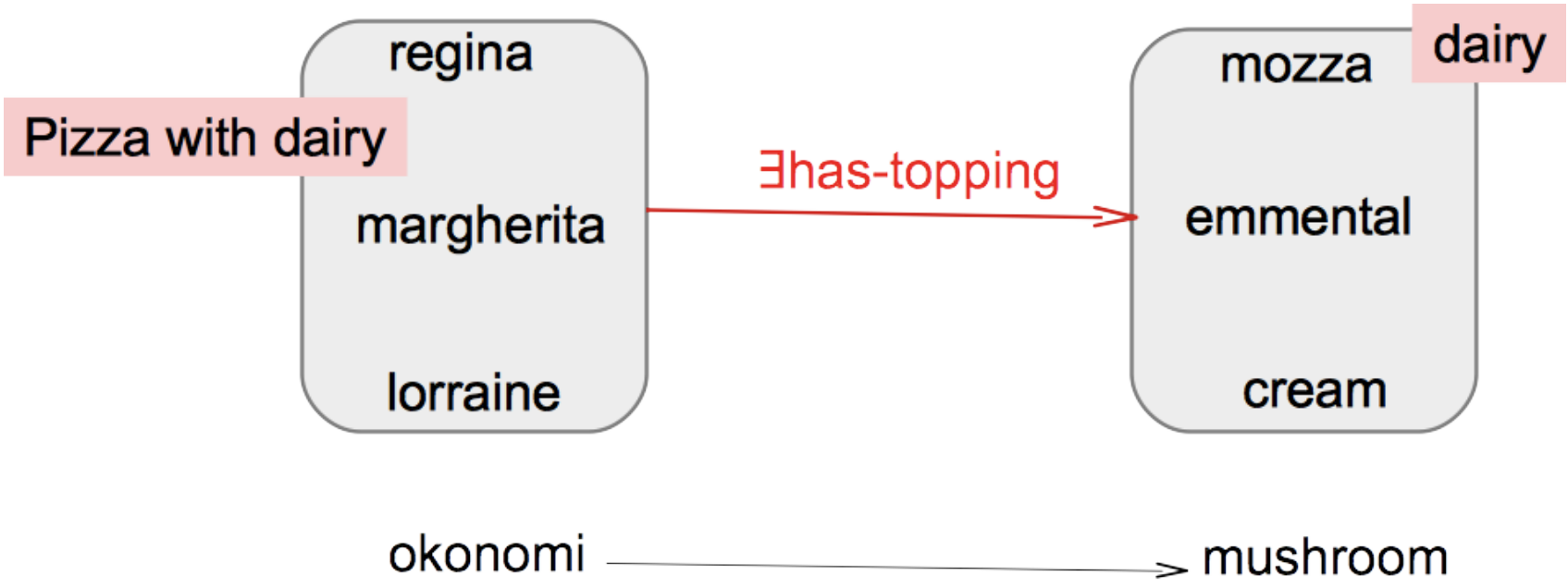
A flavor of Relational Concept Analysis



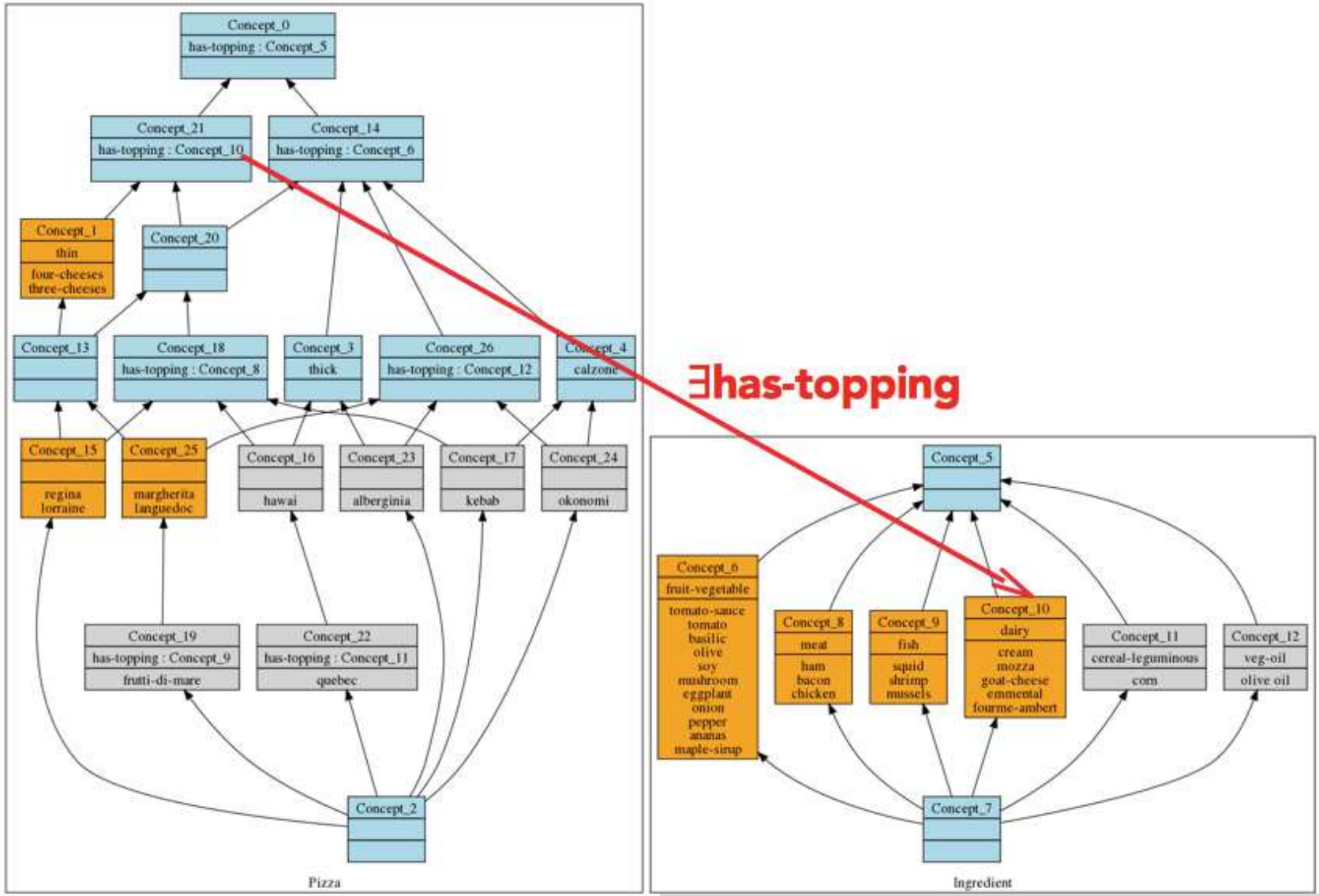
A flavor of Relational Concept Analysis



A flavor of Relational Concept Analysis



A flavor of Relational Concept Analysis



Relational Concept Analysis (RCA) [HHNV13]

- Extends the purpose of FCA for taking into account object categories and links between objects
- Main principles:
  - a relational model based on the entity-relationship model
  - integrate relations between objects as *relational* attributes
  - iterative process
- RCA provides a set of interconnected lattices
- Produced structures can be represented as ontology concepts within a knowledge representation formalism such as description logics (DLs).

Joint work with:
A. Napoli, C. Roume, M. Rouane-Hacène, P. Valtchev

Relational Context Family (RCF)

A simple entity-relationship model to introduce RCA

Relational Context Family

- **object-attribute contexts**
  - Pizza
  - Ingredient
- **object-object context**
  - has-topping $\subseteq$ Pizza $\times$ Ingredient

Relational Context Family (RCF) / object-attributes contexts

Pizza	thin	thick	calzone
okonomi			×
alberginia		×	
margherita	×		
languedoc	×		
four-cheeses	×		
three-cheeses	×		
frutti-di-mare	×		
quebec		×	
regina	×		
hawai		×	
lorraine	×		
kebab			×

Ingredient	fruit-vegetable	meat	fish	dairy	cereal-leguminous	veg-oil
tomato-sauce	×					
cream				×		
tomato	×					
basilic	×					
olive	×					
olive oil						×
soy	×					
mushroom	×					
eggplant	×					
onion	×					
pepper	×					
ananas	×					
mozza				×		
goat-cheese				×		
emmental				×		
fourme-ambert				×		
squid			×			
shrimp			×			
mussels			×			

Relational Context Family (RCF) / object-object context / part 1

has-topping	tomato-sauce	cream	tomato	basilic	olive	olive oil	soy	mushroom	eggplant	onion	pepper	ananas
okonomi	×					×	×	×				
alberginia	×					×	×		×	×		
margherita	×		×	×	×	×						
languedoc	×		×	×	×	×				×	×	
four-cheeses		×										
three-cheeses		×										
frutti-di-mare	×				×	×						
quebec	×											
regina	×							×				
hawai	×											×
lorraine		×								×		
kebab	×		×		×					×		

Relational Context Family (RCF) / object-object context / part 2

has-topping	mozza	goat-cheese	emmental	fourme-ambert	squid	shrimp	mussels	ham	bacon	chicken	maple-sirup	corn
okonomi												
alberginia												
margherita	×											
languedoc	×											
four-cheeses	×	×	×	×								
three-cheeses	×	×	×									
frutti-di-mare	×				×	×	×					
quebec	×							×			×	×
regina	×								×			
hawai	×							×				
lorraine			×						×			
kebab			×							×		

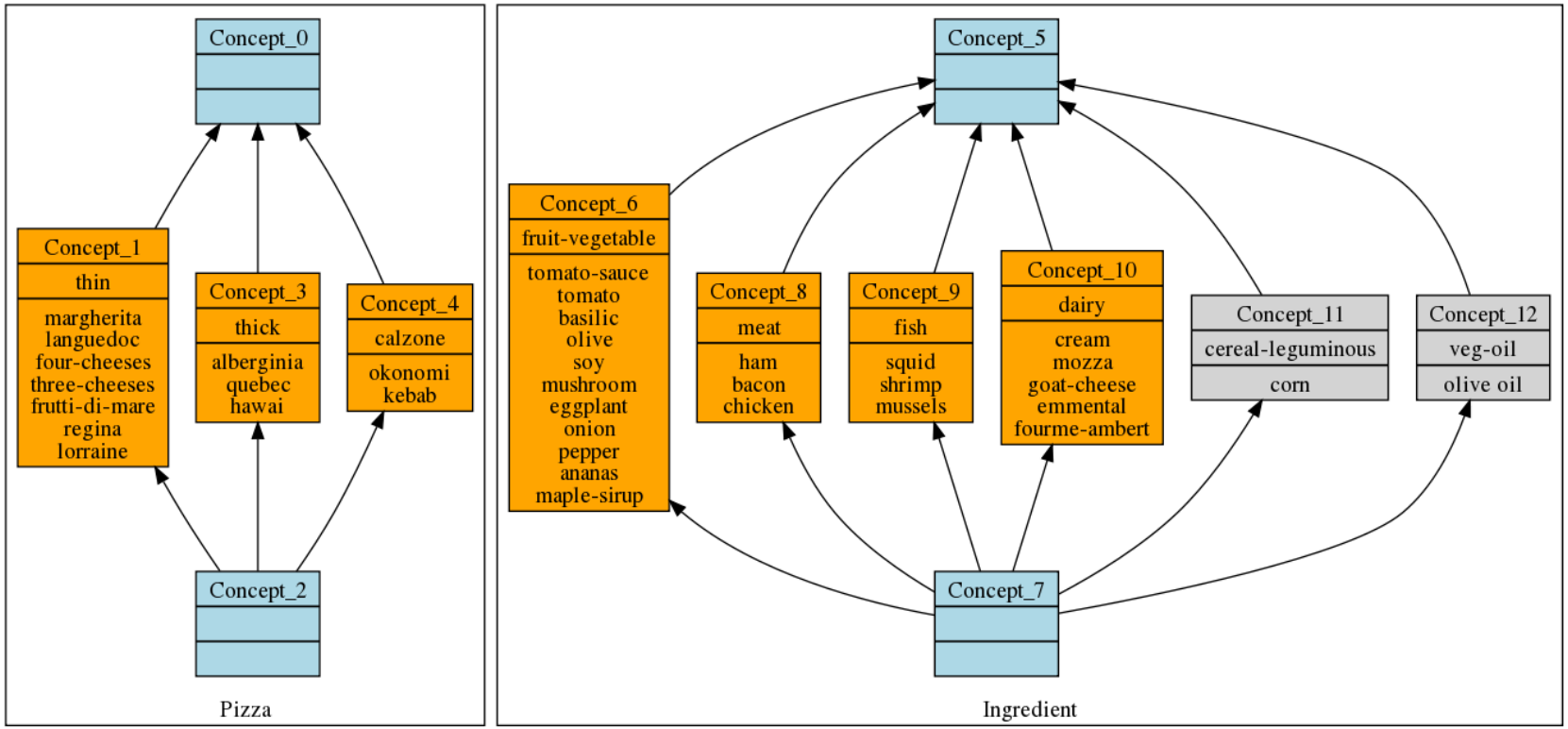
Data patterns we would like to extract

Using a classification on ingredients by their categories of topping (fruit-vegetable, dairy, etc.)

- create groups
  - The group of pizzas that contain at least one topping which is a vegetable
  - The group of pizzas (four-cheese and three-cheese) that have all their topping in dairy ingredients
- find implications
  - For pizzas: have meat $\Rightarrow$ have dairy
  - For pizzas: being thin $\Rightarrow$ have at least dairy
  - For pizzas: have only dairy $\Rightarrow$ being thin

RCA - Initial Lattice building

At the beginning, only the object-attribute contexts are used to build the foundation of the concept lattice family



RCA - Introducing relations as relational attributes

Given an object-object context $R_j = (O_k, O_l, I_j)$,
There are different possible schemas between an object of domain O_k and concepts formed on O_l.

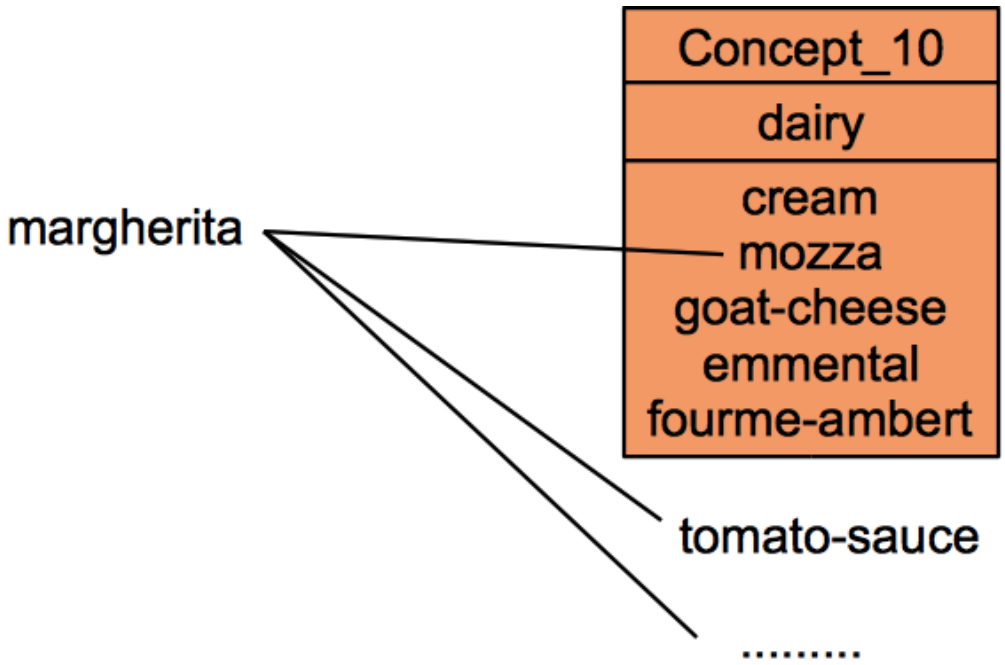
E. g.

- **Existential**: an object is linked (by R_j) to at least one object of the extent of a concept
- **Universal**: an object is linked (by R_j) only to objects of the extent of a concept

$\exists$ and $\forall$ are **scaling operators**

RCA - Existential relational attributes

margherita has one topping in Concept_10 extent: **mozza**.
It has other links to other concept extents.



∃**has-topping.Concept_10** is assigned to **margherita**

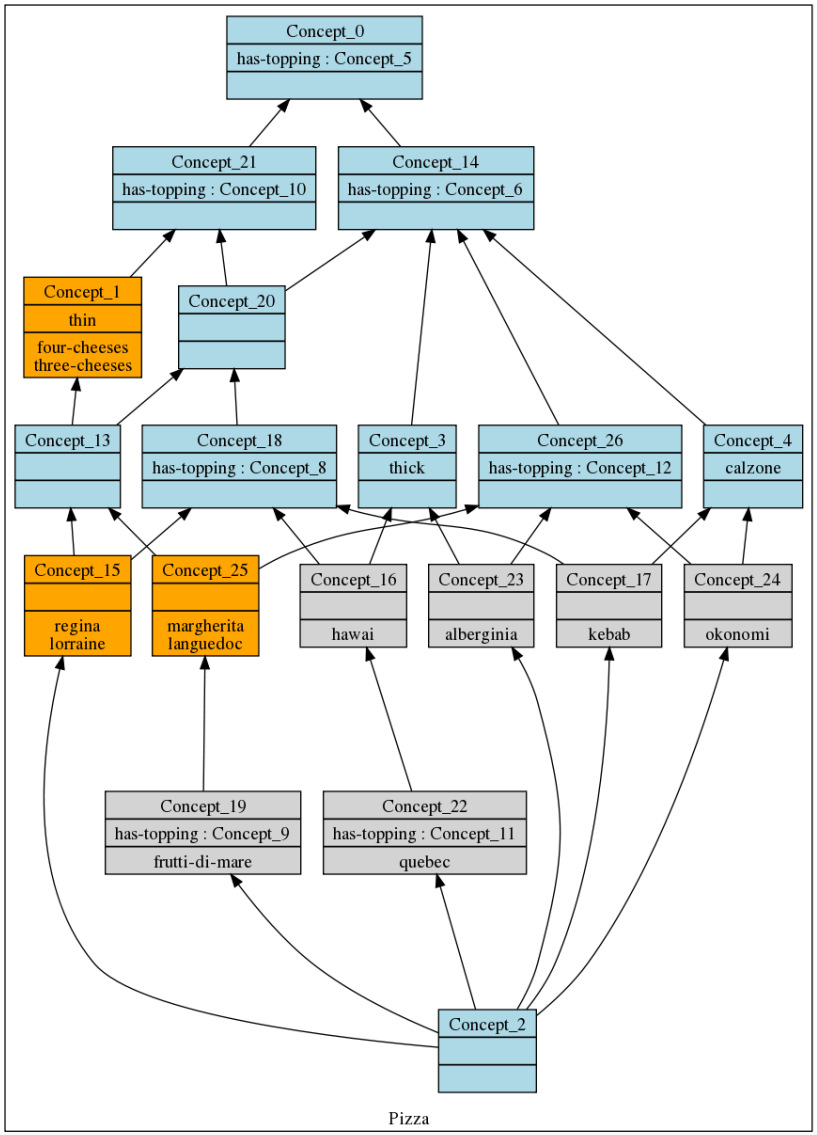
RCA - Relational extension

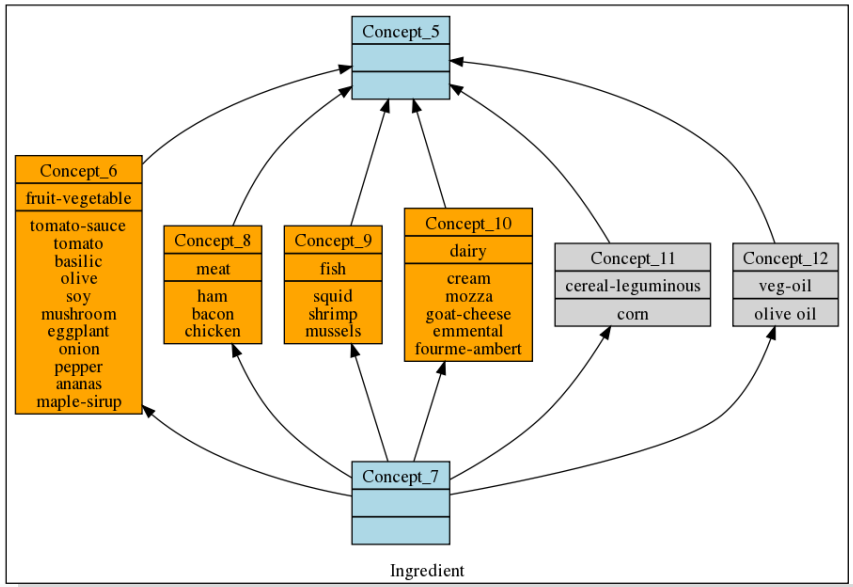
Scaled relations with domain O_i are concatenated to K_i, the object-attribute context on O_i

Pizza	thin	thick	calzone
okonomi			×
alberginia		×	
margherita	×		
languedoc	×		
four-cheeses	×		
three-cheeses	×		
frutti-di-mare	×		
quebec		×	
regina	×		
hawai		×	
lorraine	×		
kebab			×

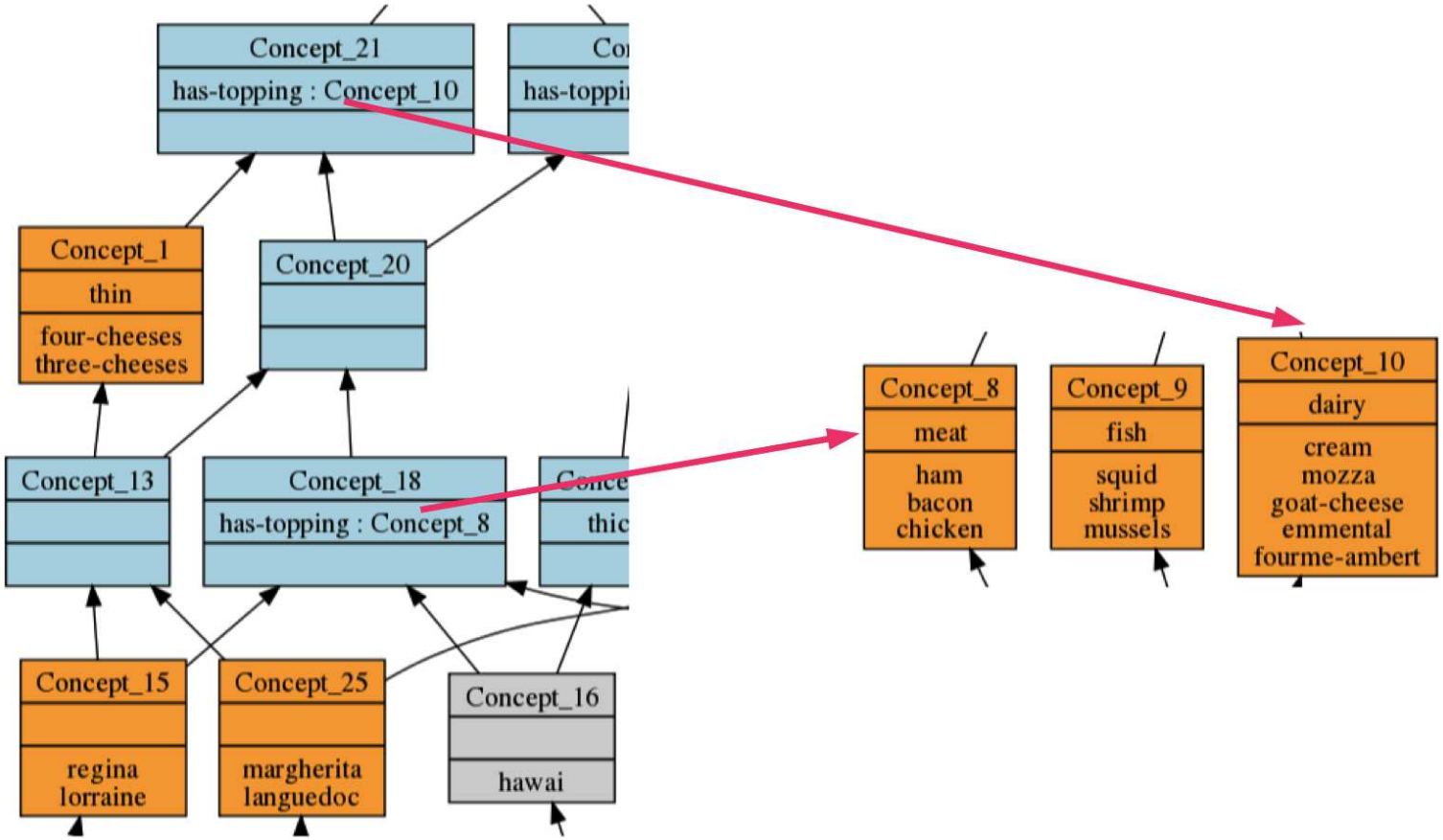
has-topping	∃has-topping. Concept_7	∃has-topping. Concept_5	∃has-topping. Concept_6	∃has-topping. Concept_8	∃has-topping. Concept_9	∃has-topping. Concept_10	∃has-topping. Concept_11	∃has-topping. Concept_12
okonomi		×	×					×
alberginia		×	×					×
margherita		×	×			×		×
languedoc		×	×			×		×
four-cheeses		×				×		
three-cheeses		×				×		
frutti-di-mare		×	×		×	×		×
quebec		×	×	×		×	×	
regina		×	×	×		×		
hawai		×	×	×		×		
lorraine		×	×	×		×		
kebab		×	×	×		×		

Relational Concept Family / exists





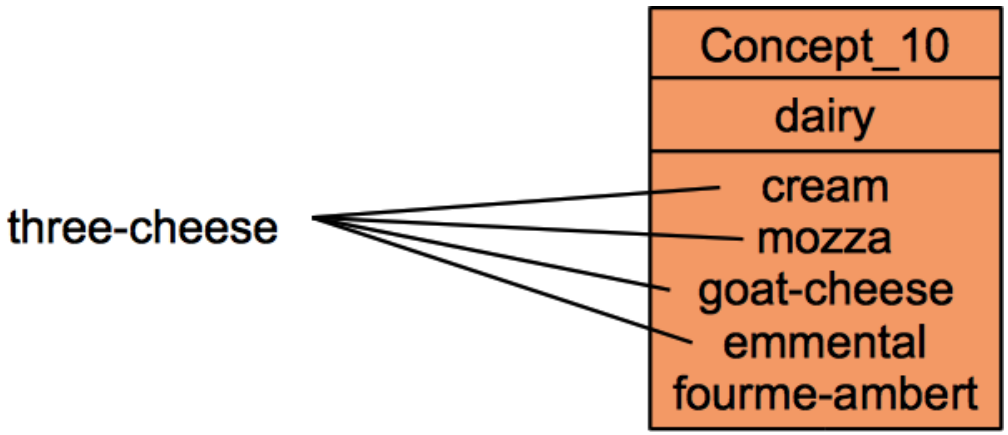
Relational Concept Family / exists



Concept_21: pizzas with at least one topping in dairy
Concept_18: pizzas with at least one topping in meat
have at least one meat topping $\Rightarrow$ have at least one dairy topping

RCA - Universal relational attributes

three-cheese has topping in and only in Concept_10 extent.



$\forall\exists$**has-topping.Concept_10** is assigned to **three-cheese**

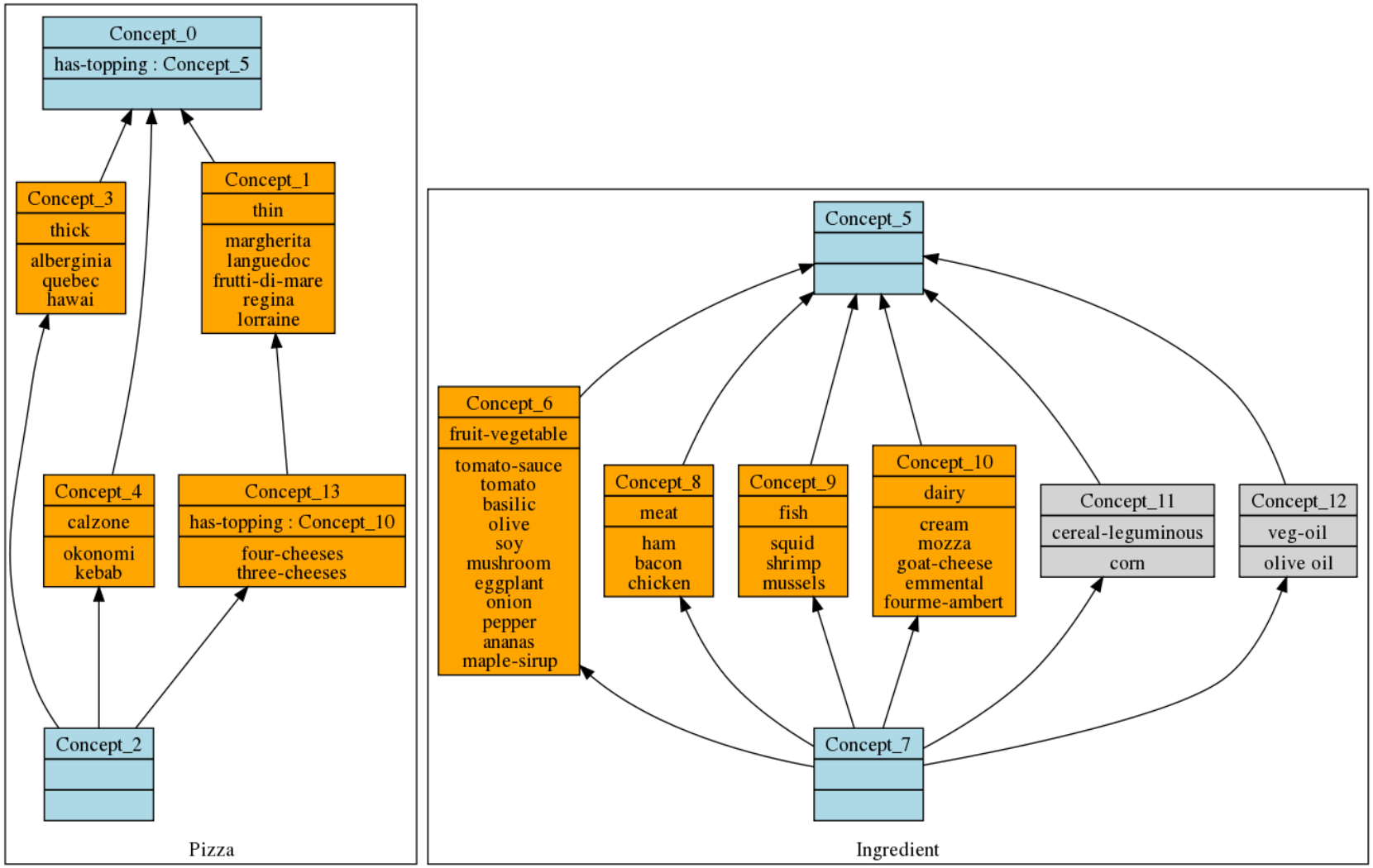
RCA - Relational extension

Scaled relations with domain O_i are concatenated to K_i, the object-attribute context on O_i

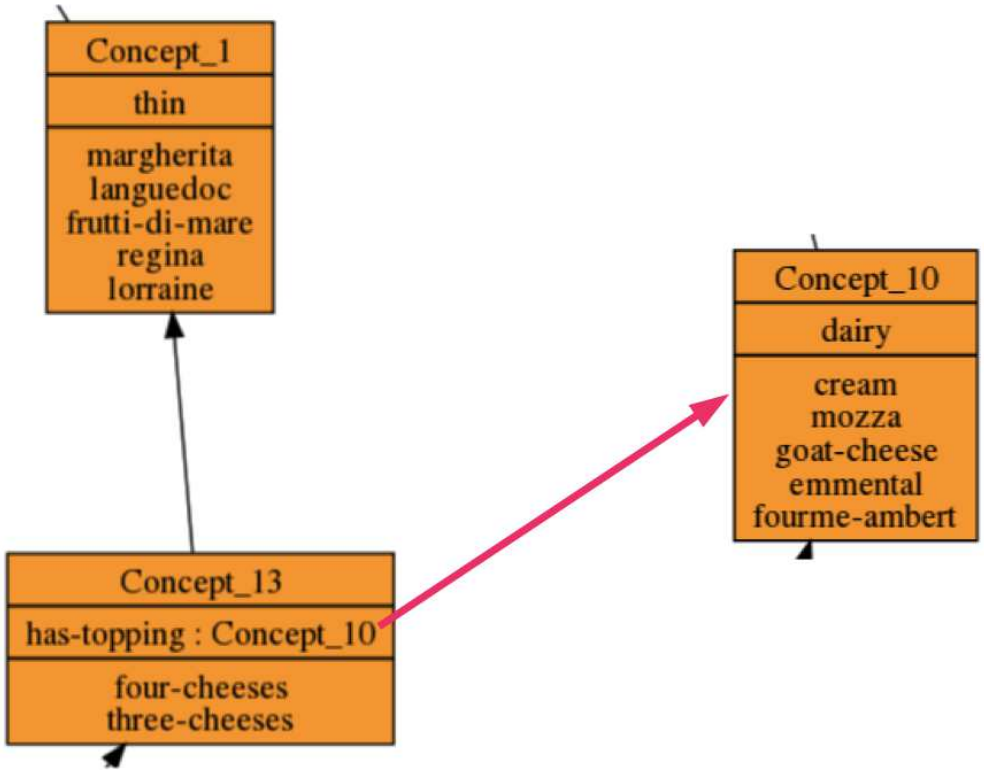
Pizza	thin	thick	calzone
okonomi			×
alberginia		×	
margherita	×		
languedoc	×		
four-cheeses	×		
three-cheeses	×		
frutti-di-mare	×		
quebec		×	
regina	×		
hawai		×	
lorraine	×		
kebab			×

has-topping	∃A has-topping. Concept_7	∃A has-topping. Concept_5	∃A has-topping. Concept_6	∃A has-topping. Concept_8	∃A has-topping. Concept_9	∃A has-topping. Concept_10	∃A has-topping. Concept_11	∃A has-topping. Concept_12
okonomi		×						
alberginia		×						
margherita		×						
languedoc		×						
four-cheeses		×				×		
three-cheeses		×				×		
frutti-di-mare		×						
quebec		×						
regina		×						
hawai		×						
lorraine		×						
kebab		×						

Relational Concept Family / forall

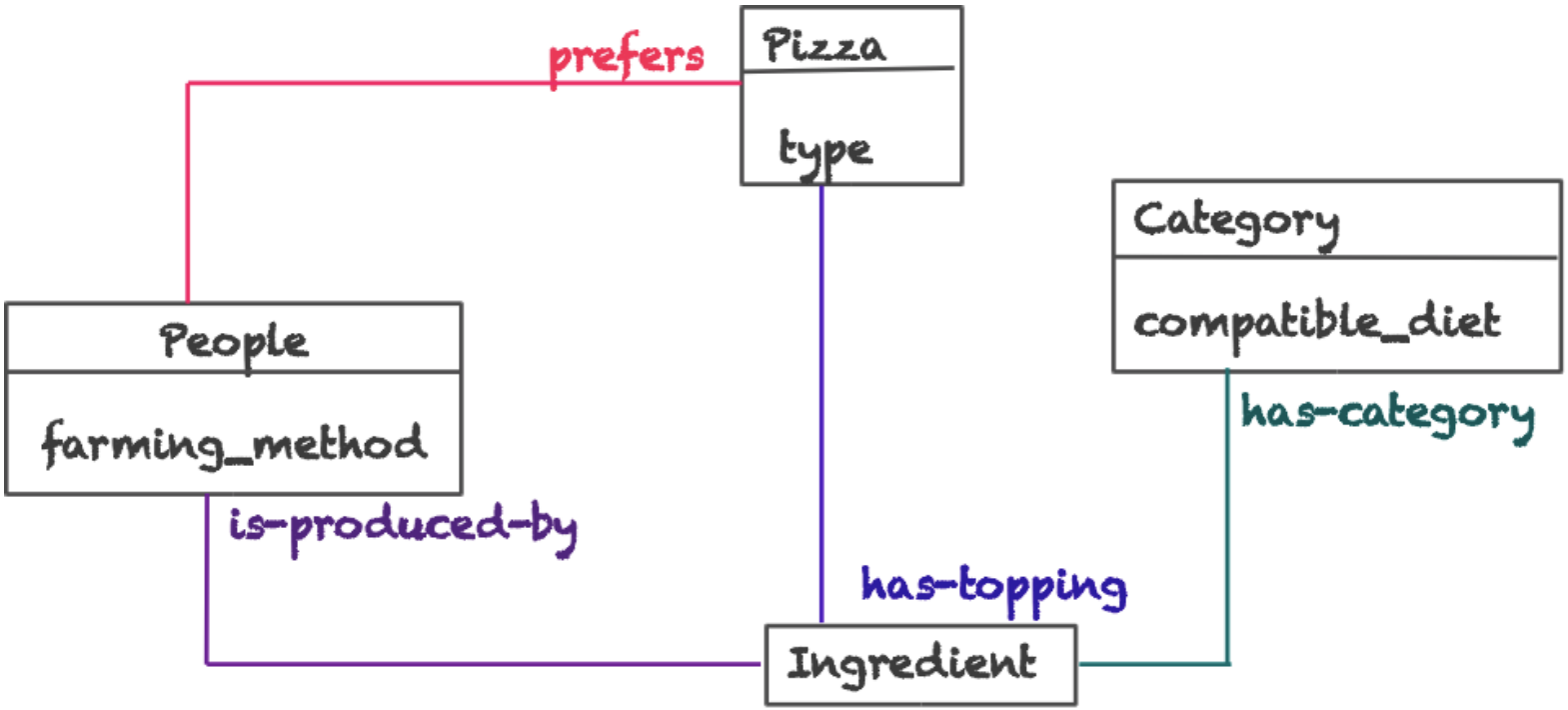


Relational Concept Family / forall



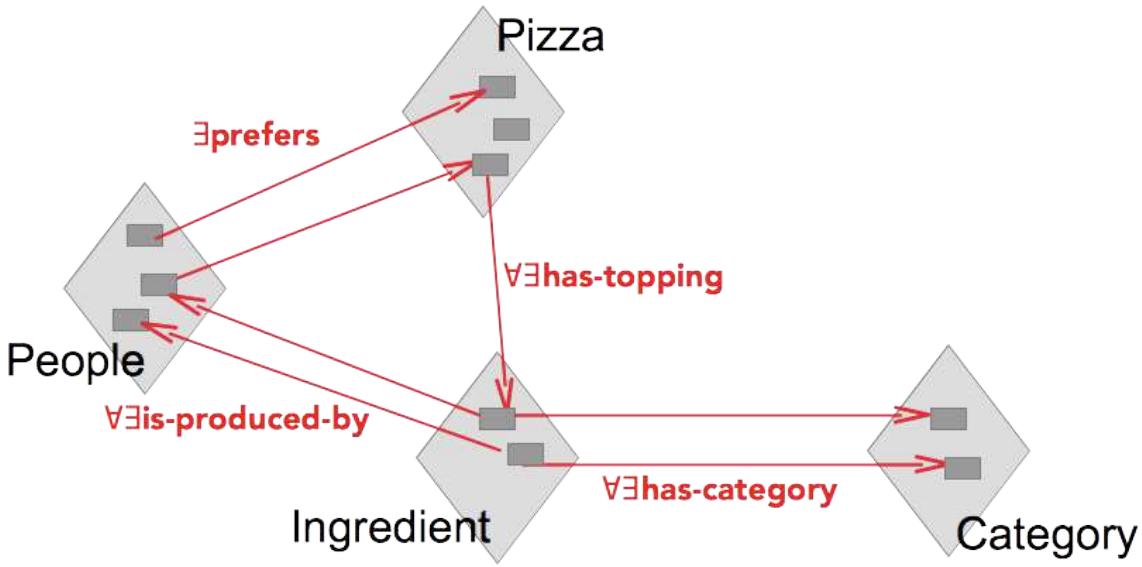
Concept_13: pizzas with only dairy topping
Concept_1: thin pizzas
have only dairy topping $\Rightarrow$ thin

General Entity-Relationship diagram may have circuits



$\exists$ prefers $\forall\exists$ has-topping $\forall\exists$ has-category $\forall\exists$ is-produced-by

General Entity-Relationship diagram may have circuits



Example of possible learned knowledge

- $\forall\exists$has-category.Vegetable $\Leftrightarrow$ $\forall\exists$is-produced-by.Organic farmers
- A subgroup of organic farmers prefer at least one pizza with only vegan topping ingredients and produced only by organic farmers

The RCA schema

Input

RCF: n object-attribute contexts, m object-object contexts

Initialization step

Build the concept lattice for each object-attribute context

Step p

- ▷ Apply relational scaling to all object-object contexts
- ▷ Build relational extension of each object-attribute context:
  object-attribute context + scaled object-object contexts
- ▷ Build the concept lattice for each relational extension

Output (fix point)

The concept lattice family obtained when no new concepts are added

A synthesis on RCA

- an iterative method to produce interconnected classifications
- converges after a number of iterations that depends on the structure
- a variety of scaling operators
- reduced structures can be used instead lattices: AOC-posets, iceberg lattices

Tools

- Galicia: http://galicia.sourceforge.net/
- eRCA: http://code.google.com/p/erca/
- **RCAexplore: http://dolques.free.fr/rcaexplore/site_web/**

An introduction to RCA

RCA for model evolution

- In follow-up of model evolution
- In assisting model evolution

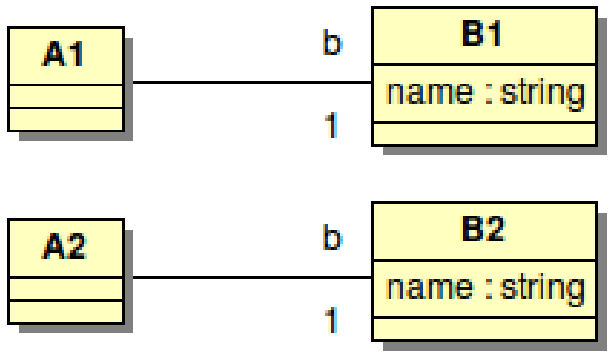
Context and Problematic

Environment and Territory domains

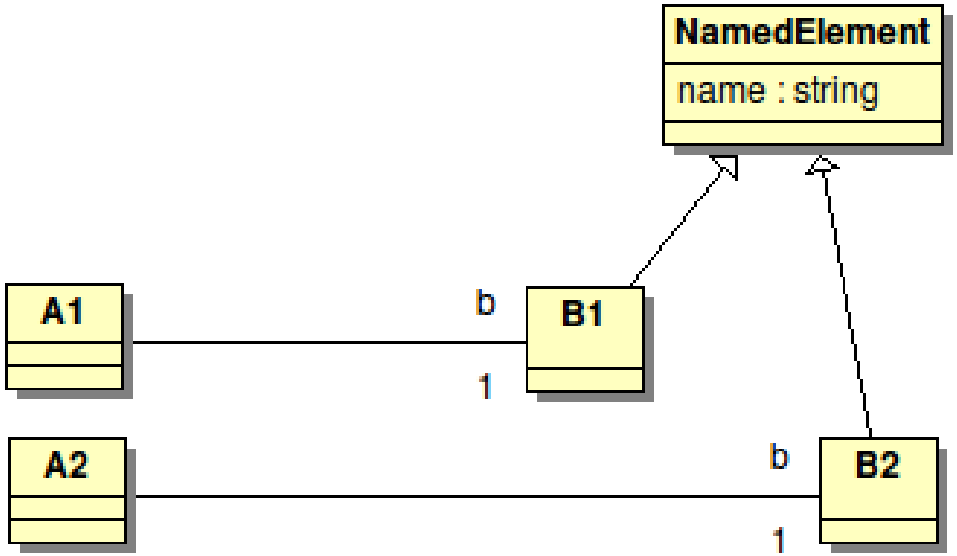
- Development of Information System involves many actors and scientists: *EIS-Pesticides*
- Meeting after meeting, the designer has to merge various viewpoints in a global UML that evolves progressively
- During the analysis phase, models are archived after each major change

Joint work with B. Amar, X. Dolques, F. Le Ber, T. Libourel, A. Miralles, C. Nebut, A. Osman-Guédi

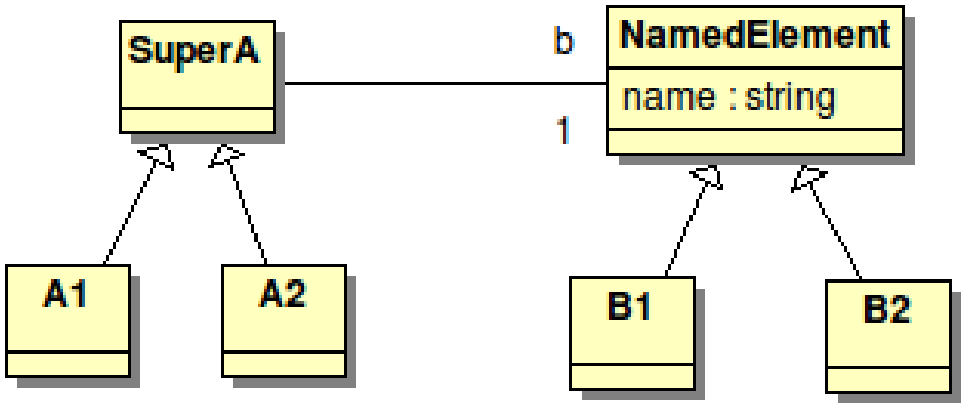
RCA for class model normalization



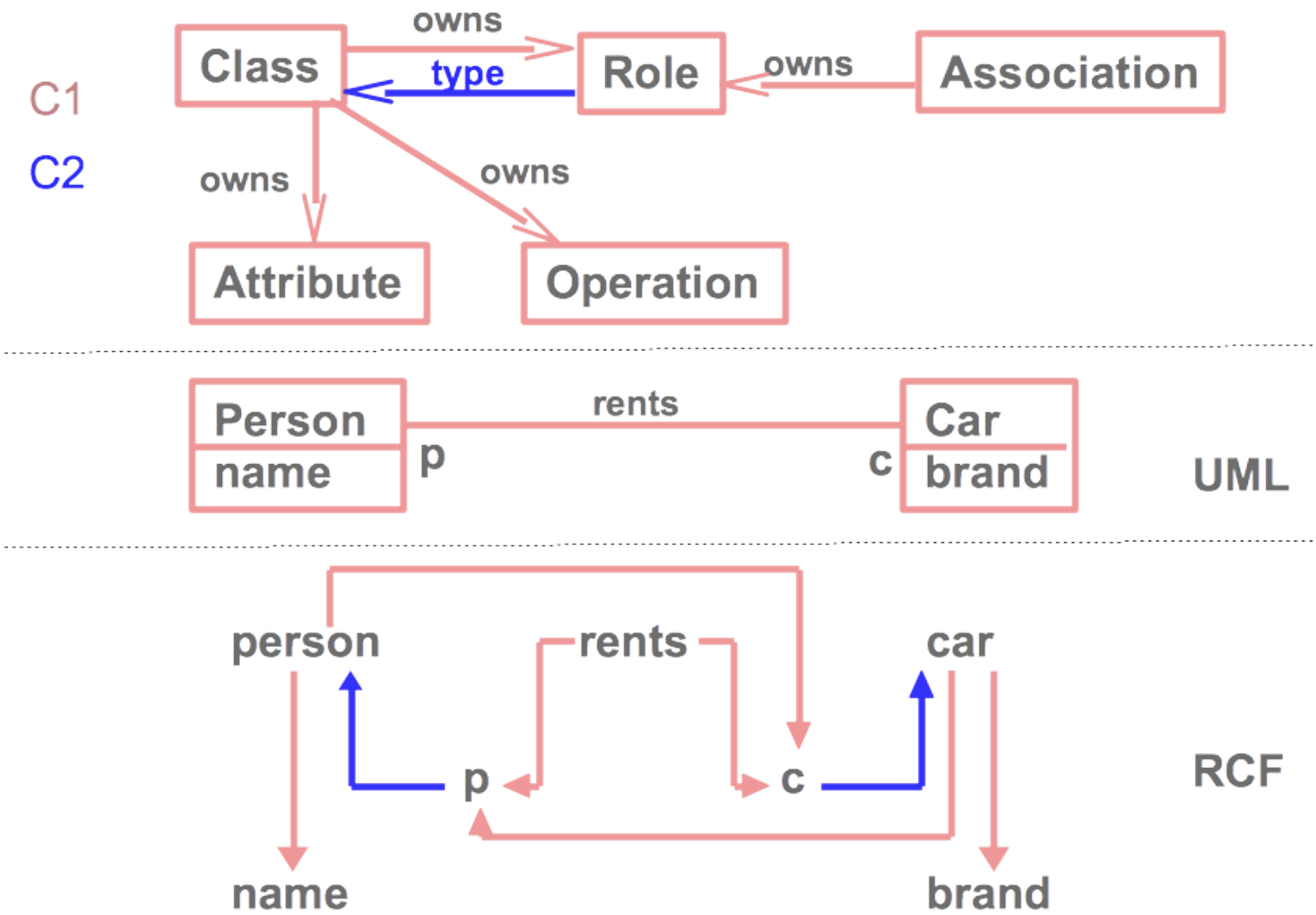
RCA for class model normalization



RCA for class model normalization



RCA for class model normalization



RCA for class model normalization

Strong properties of the resulting class model

- No redundancy
- All abstractions are created
- All specialization links are present

Approach

Develop methods using the class model normal form obtained with RCA for class model construction and evolution:

- monitoring
- assisting

Model evolution monitoring

Classical model indicators

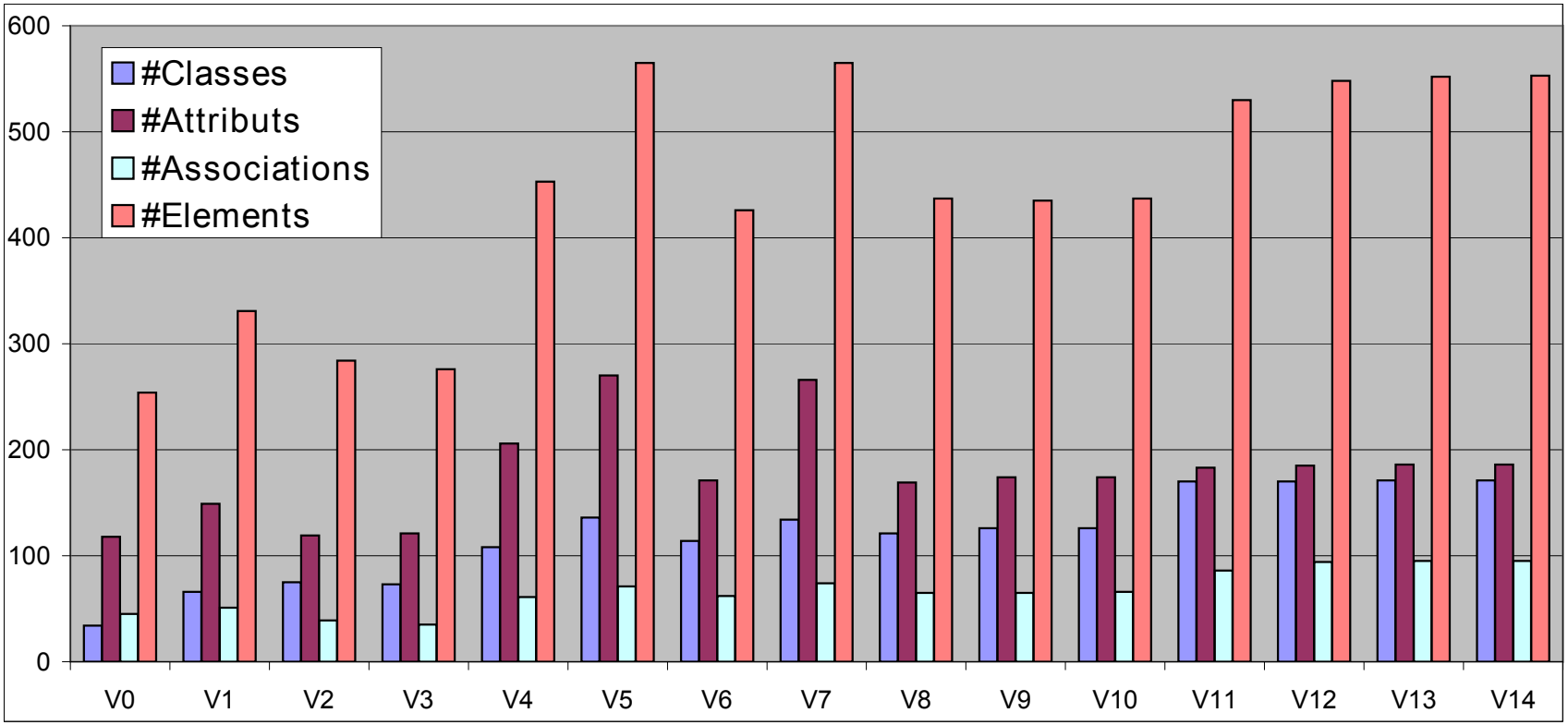
The domain experts mainly used the number of elements of various kinds (classes, methods. . .)

- Do not reveal complex evolution :
  - precision in the description of model elements
  - level of abstraction and factorization

Proposal

Develop indicators based on the application of RCA
As RCA produces a unique normal form, our metrics are based on the comparison of these normal forms (here with configuration C1)

Evolution of the different model elements



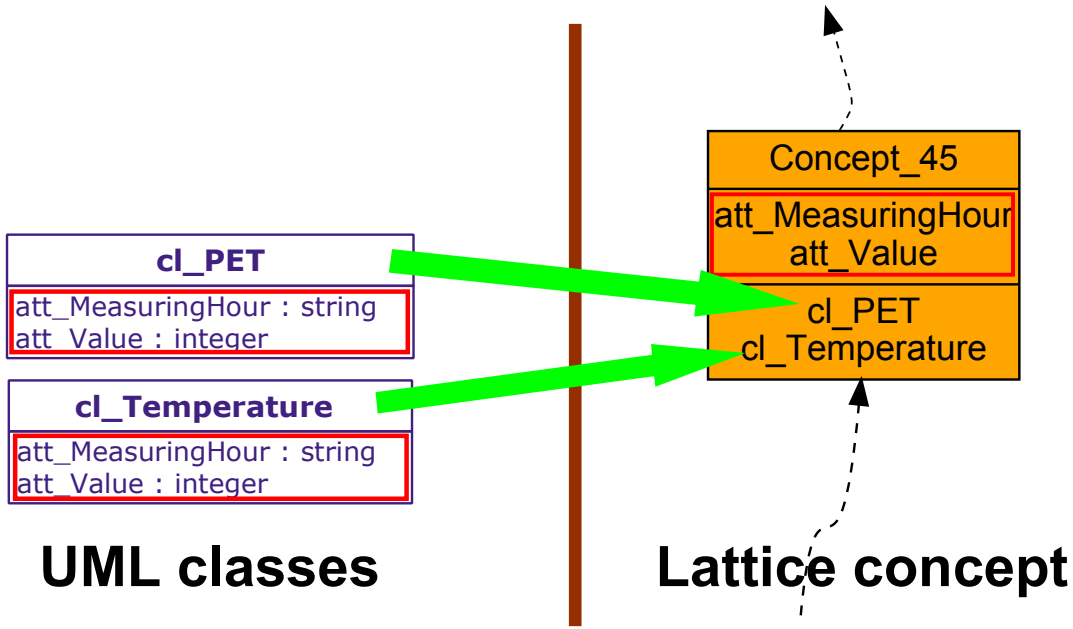
Lattice indicators evolution: #Merge/#Model Elements

The metrics based on the ratio of merged concepts:

#Merge / #Model Elements

- Merged Concepts have a proper extent that contains more than one element
- They merge several formal objects with the same description

Example of merged concept



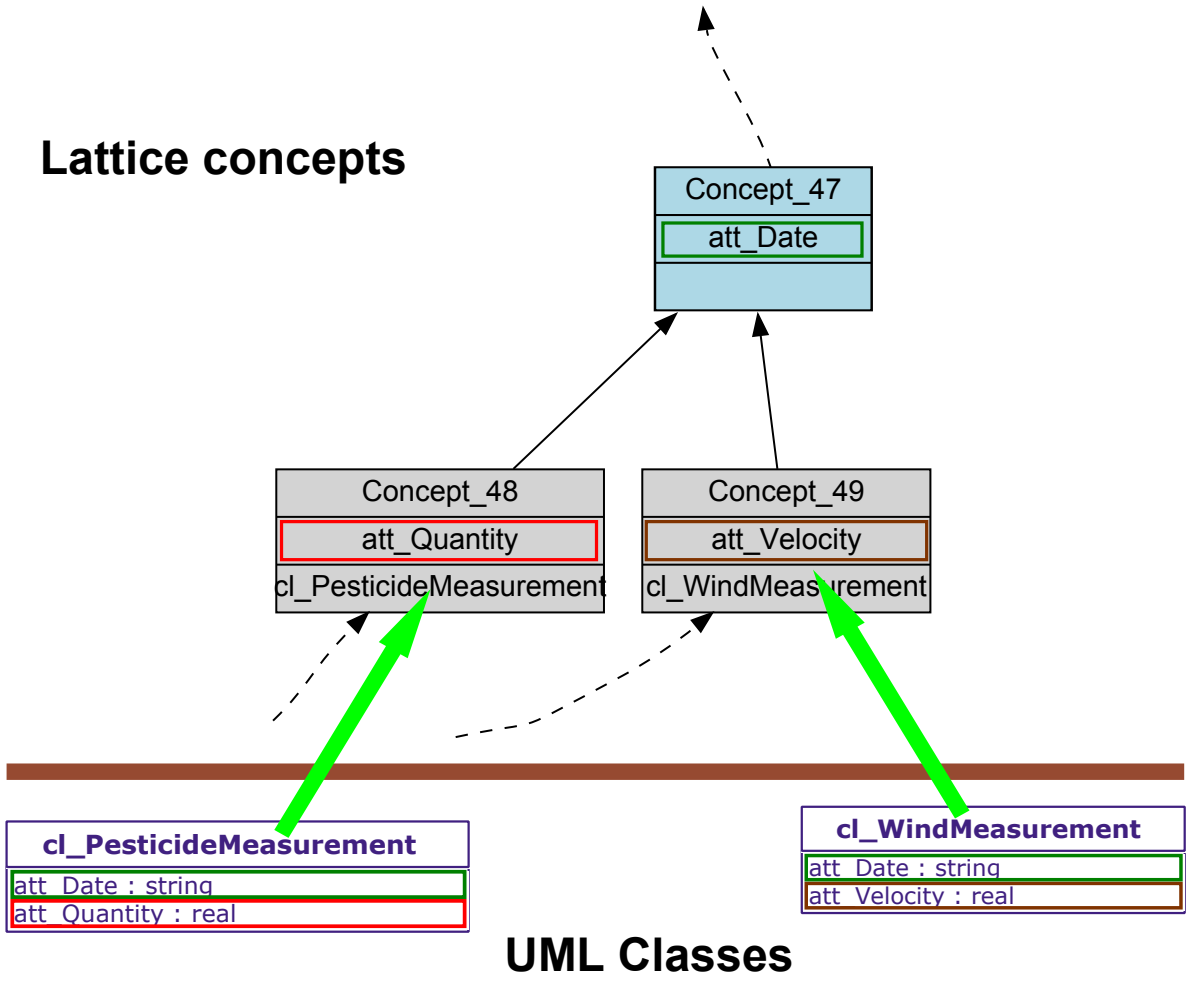
Lattice indicators evolution: #New/#Model Elements

The metrics based on the ratio of new concepts:

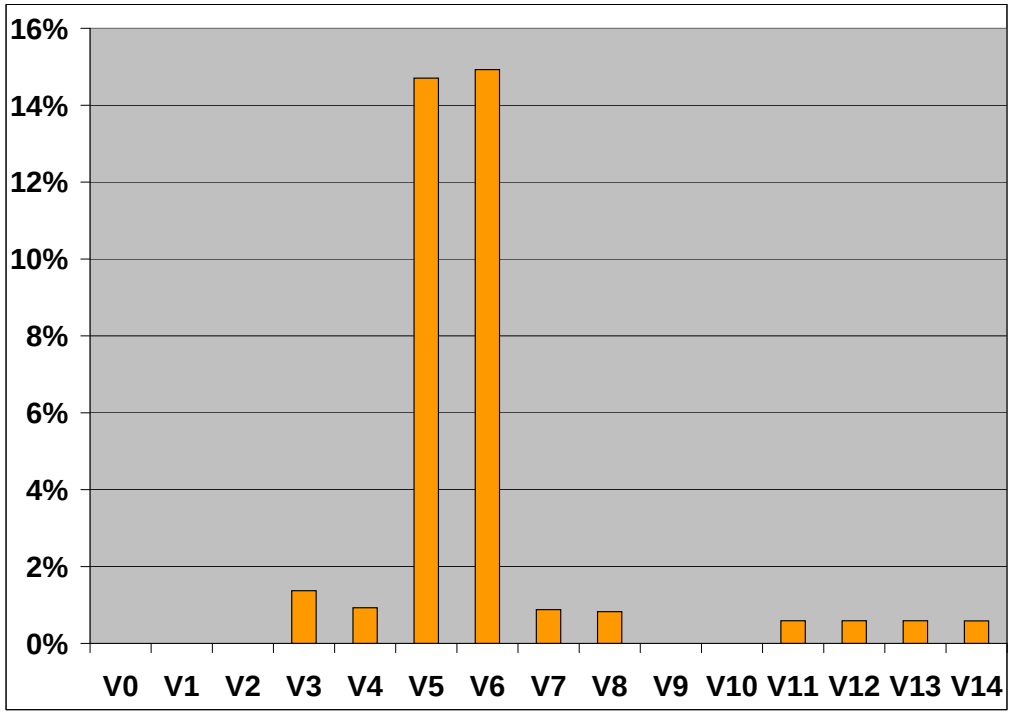
#New / #Model Elements

- New Concepts have an empty proper extent
- They factorize formal attributes

Example of new concept

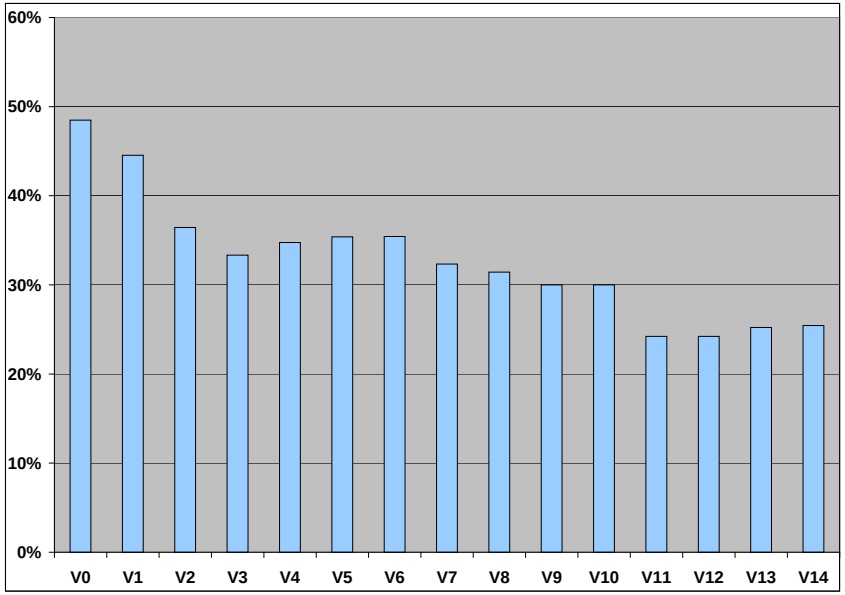


Indicators on Classes : Merged Classes



- V5, V6 : Package duplication

Indicators on Classes : New Classes



- Progressive decrease even if the number of classes increases
- The abstraction level of the model improves
- V5, V6 : the package duplication degrades the abstraction level

Discussion

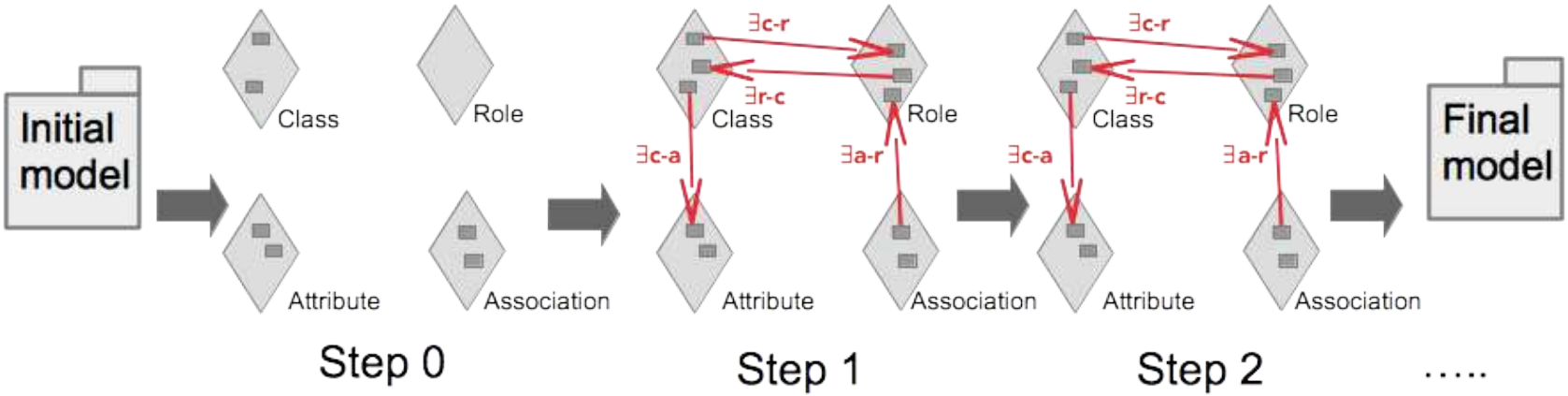
Classical metrics to analyze

- Evolution of data encapsulation ($\simeq$ number of classes)
- Evolution of the completion of the model ($\simeq$ number of attributes)
- Evolution of the relational aspect ($\simeq$ number of roles / associations)

RCA-based metrics complete the analysis

- Evolution of the merged ratio indicates if identical or badly described model elements are introduced
- Evolution of the new ratio indicates the level of abstraction

Traditional RCA approach

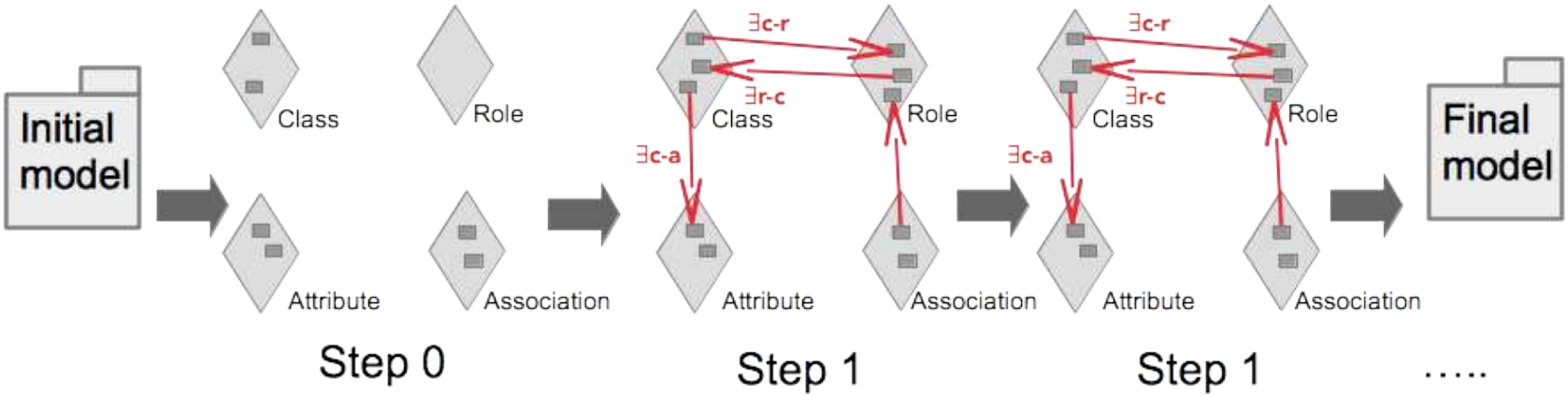


Issue

The final model contains many merged or new elements, this is difficult to analyze to keep the relevant part

Exploration path

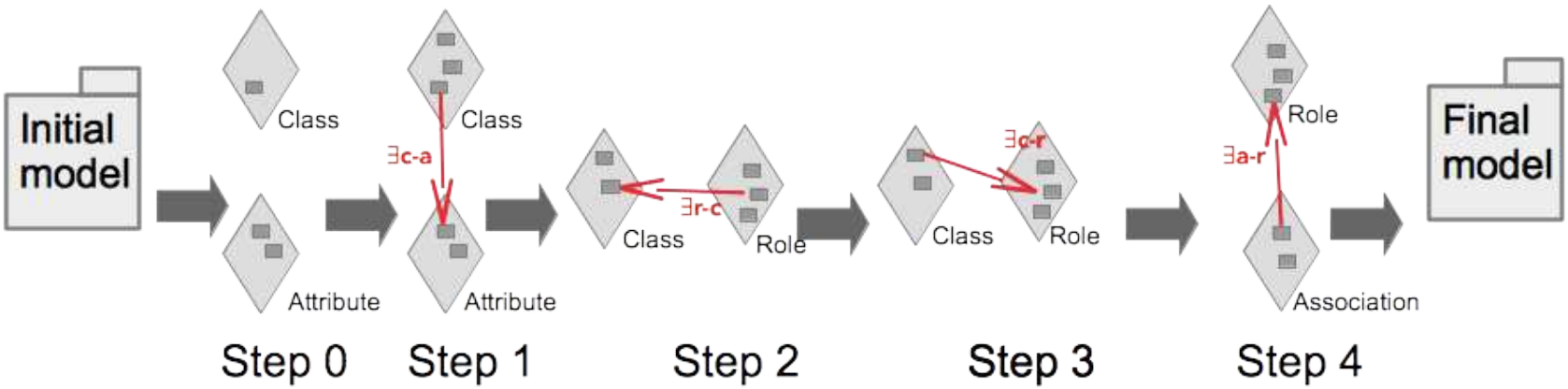
Fighting against possible high number of concepts to be analyzed
by choosing good configurations
by bringing concepts step by step



Auto path: all contexts are considered, but the process stops at each step and presents the concepts to the designer

Exploration path

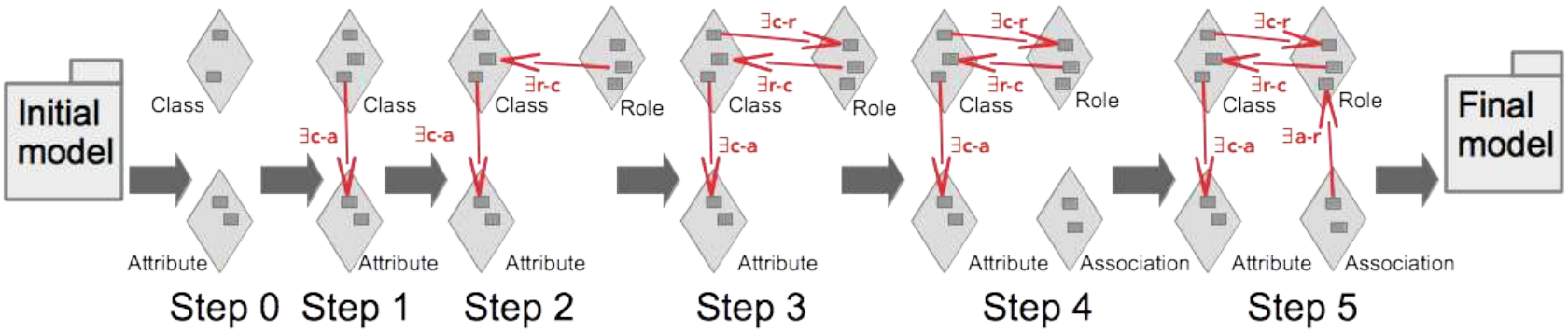
Fighting against possible high number of concepts to be analyzed
by using parts of the RCF



Path 1: each step considers a specific part of the RCF

Exploration path

Fighting against possible high number of concepts to be analyzed
by using parts of the RCF - cumulative



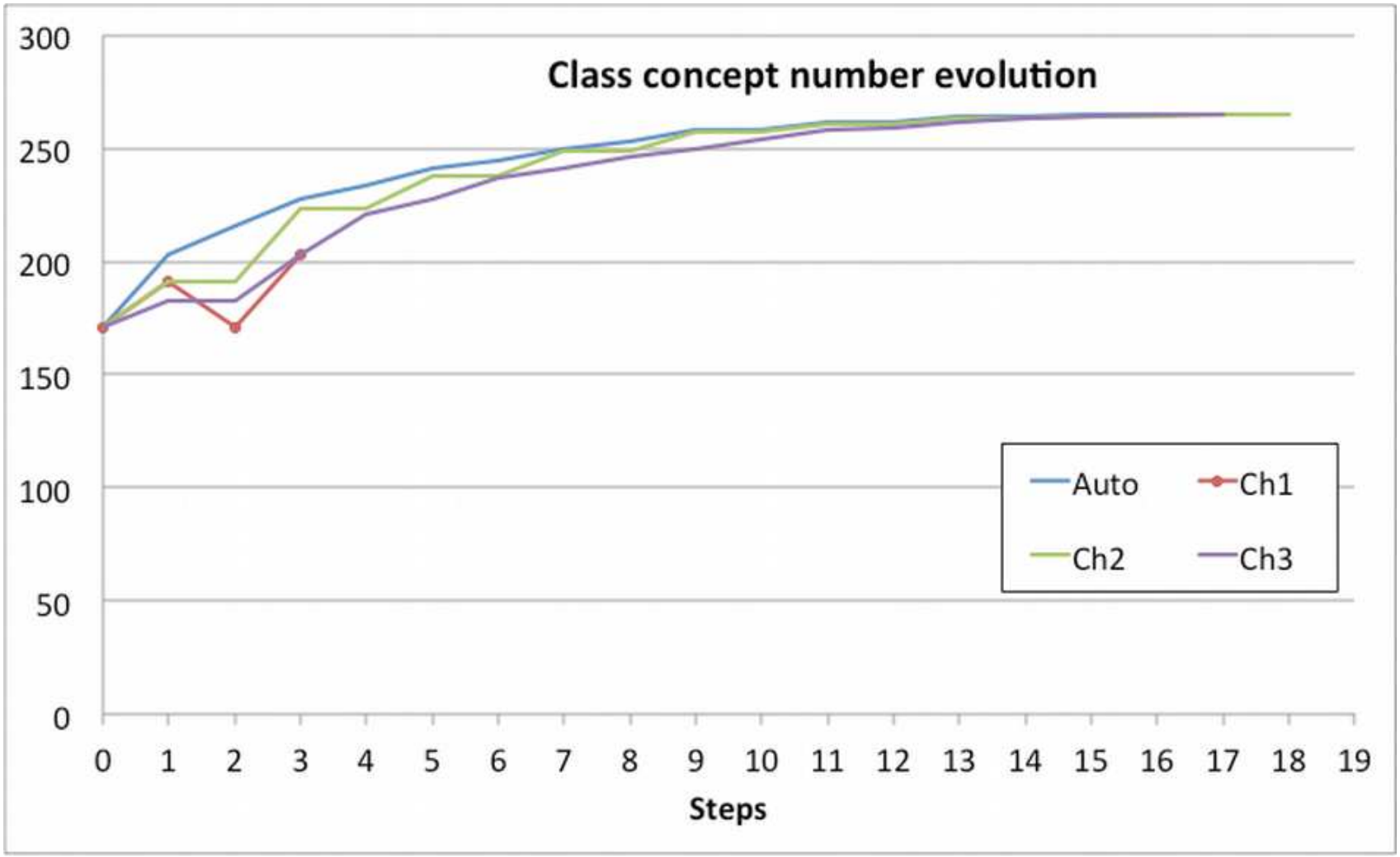
Path 2: Begin by class/attributes, add roles, add associations
Path 3: A variant that begins by class/roles

Quantitative analysis: ex. with class concepts to be analyzed at each step

RCA application on Pesticides: 171 classes before, 265 concepts

step tr.	Auto	Path 1	Path 2	Path 3	step tr.	Auto	Path 1	Path 2	Path 3
0 →1	32	20	20	12	10 →11	4		4	4
1 →2	13	-20	0	0	11 →11	0		0	1
2 →3	12	32	32	20	12 →13	2		2	3
3 →4	6		0	18	13 →14	0		0	1
4 →5	7		15	7	14 →15	1		1	1
5 →6	4		0	9	15 →16	0		0	1
6 →7	5		11	4	16 →17	Auto		1	0
7 →8	3		0	5	17 →18	Auto		0	
8 →9	5		8	4					
9 →10	0		0	4					

Class concept number evolution



Discussion

- Exploration divides the burden of the analysis
- The process is controlled by the expert
- Paths cannot be chosen by chance, cumulative paths ensure completeness
- Perspectives: define a complete methodology and tools

General Conclusion

- RCA: an opportunity for analyzing more deeply dataset composed of objects and relations
- Can be mixed with other FCA extension (to numerical data for example)
- Exploratory RCA allows us step-by-step analysis, considering a subset of the dataset and changing structures (lattices, AOC-posets, iceberg)

Perspectives

- A querying mechanism and navigation tools
- Comparing AOC-poset and lattice in the applications
- Studying effect of exploration on the method convergence

Class concept number evolution

Questions?


Michel Dao, Marianne Huchard, Mohamed Rouane Hacene, Cyril Roume, and Petko Valtchev.
Improving Generalization Level in UML Models Iterative Cross Generalization in Practice.
In *ICCS 2004*, pages 346–360, 2004.


Jean-Rémy Falleri.
Contributions à l'IDM : reconstruction et alignement de modèles de classes.
PhD thesis, Université Montpellier 2, 2009.


Jean-Rémy Falleri, Marianne Huchard, and Clémentine Nebut.
A generic approach for class model normalization.
In *ASE 2008*, pages 431–434, 2008.


Mohamed Rouane Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev.
Relational concept analysis: mining concept lattices from multi-relational data.
Ann. Math. Artif. Intell., 67(1):81–108, 2013.


Cyril Roume.
Analyse et restructuration de hiérarchies de classes.
PhD thesis, Université Montpellier 2, 2004.