



INTRODUCTION TO SEMANTIC DATA MINING

Chiara Renso

KDD-LAB, ISTI-CNR, Pisa, Italy

www-kdd.isti.cnr.it/chiara/

OUTLINE OF THE SEMINAR

- Introduction
 - the need for semantics, the knowledge discovery process, ontologies.
- Overview of some methods for semantic data mining:
 - Semantic enhancement in the mining process/tasks
 - Data transformation based on domain knowledge
 - Navigate extracted models
 - Languages for Data Mining (DMQL)

SEMANTICS

Semantics is the study of meaning*.

Semantic Data Mining means to get a meaning from the data mining task



WHY SEMANTIC DATA MINING?

- Data Mining: the extraction of **useful and interesting knowledge** from large masses of data
- However, Data Mining research put most of the effort on the algorithms development, efficiency, preprocessing, visualization...
- The evaluation of the results is done with statistical quantitative measures (precision, recall, similarity, coverage, confidence, etc)

... BUT

Which is the “real value” of the knowledge for the final users?

THE DATA MINING USER

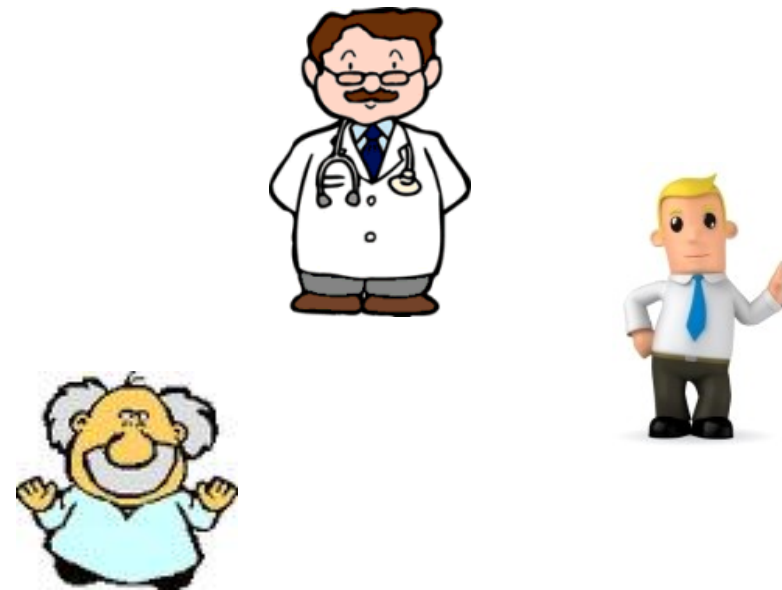
Which user... ?

Analyst User



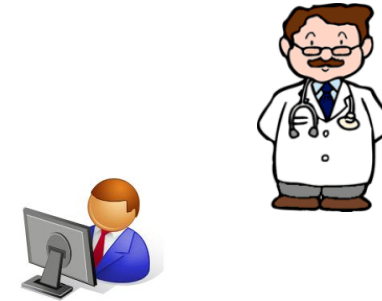
The **DM analyst** knows how to run DM tools, but usually lacks domain knowledge

Final User/Domain Expert



The **Domain Expert** is usually a professional in a specific domain, not necessarily knows how to run a DM tool/process

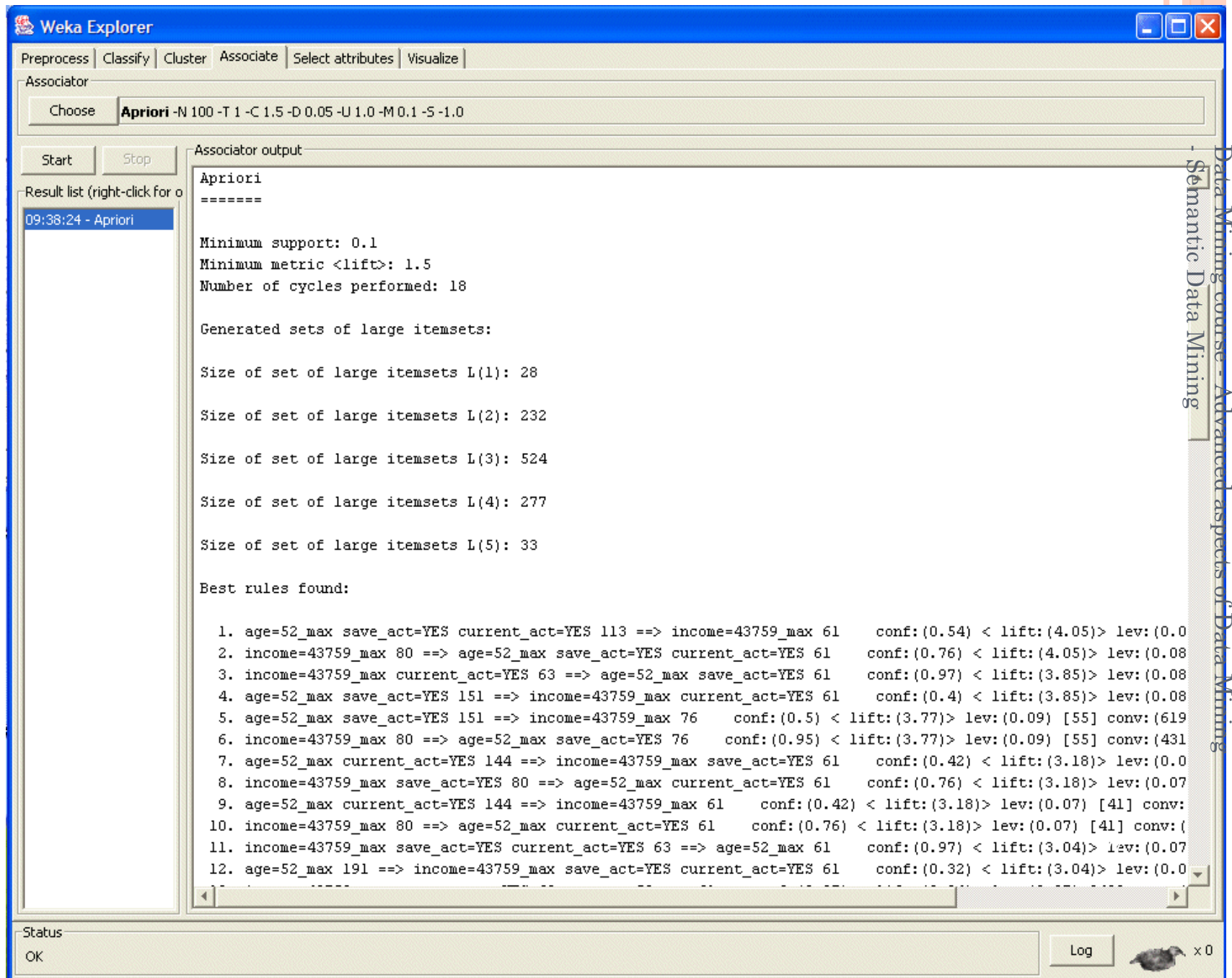
THE DATA MINING USER



- Most of current DM evaluation techniques are oriented **towards the DM analyst**. Not necessarily useful for the domain expert.

There is the need to take into account the knowledge coming from the domain experts

EXAMPLES



The screenshot shows the Weka Explorer interface with the Apriori algorithm selected. The main window displays the following output:

```
Apriori
=====
Minimum support: 0.1
Minimum metric <lift>: 1.5
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 28
Size of set of large itemsets L(2): 232
Size of set of large itemsets L(3): 524
Size of set of large itemsets L(4): 277
Size of set of large itemsets L(5): 33

Best rules found:

1. age=52_max save_act=YES current_act=YES 113 ==> income=43759_max 61    conf:(0.54) < lift:(4.05)> lev:(0.0
2. income=43759_max 80 ==> age=52_max save_act=YES current_act=YES 61    conf:(0.76) < lift:(4.05)> lev:(0.08
3. income=43759_max current_act=YES 63 ==> age=52_max save_act=YES 61    conf:(0.97) < lift:(3.85)> lev:(0.08
4. age=52_max save_act=YES 151 ==> income=43759_max current_act=YES 61    conf:(0.4) < lift:(3.85)> lev:(0.08
5. age=52_max save_act=YES 151 ==> income=43759_max 76    conf:(0.5) < lift:(3.77)> lev:(0.09) [55] conv:(619
6. income=43759_max 80 ==> age=52_max save_act=YES 76    conf:(0.95) < lift:(3.77)> lev:(0.09) [55] conv:(431
7. age=52_max current_act=YES 144 ==> income=43759_max save_act=YES 61    conf:(0.42) < lift:(3.18)> lev:(0.0
8. income=43759_max save_act=YES 80 ==> age=52_max current_act=YES 61    conf:(0.76) < lift:(3.18)> lev:(0.07
9. age=52_max current_act=YES 144 ==> income=43759_max 61    conf:(0.42) < lift:(3.18)> lev:(0.07) [41] conv:
10. income=43759_max 80 ==> age=52_max current_act=YES 61    conf:(0.76) < lift:(3.18)> lev:(0.07) [41] conv:(
11. income=43759_max save_act=YES current_act=YES 63 ==> age=52_max 61    conf:(0.97) < lift:(3.04)> lev:(0.07
12. age=52_max 191 ==> income=43759_max save_act=YES current_act=YES 61    conf:(0.32) < lift:(3.04)> lev:(0.0
```

The interface includes a menu bar (Preprocess, Classify, Cluster, Associate, Select attributes, Visualize), an Associator section with a 'Choose' button and the selected algorithm 'Apriori -N 100 -T 1 -C 1.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0', and 'Start' and 'Stop' buttons. A 'Result list' on the left shows '09:38:24 - Apriori' selected. The status bar at the bottom shows 'Status OK' and a 'Log' button.

kMeans
=====

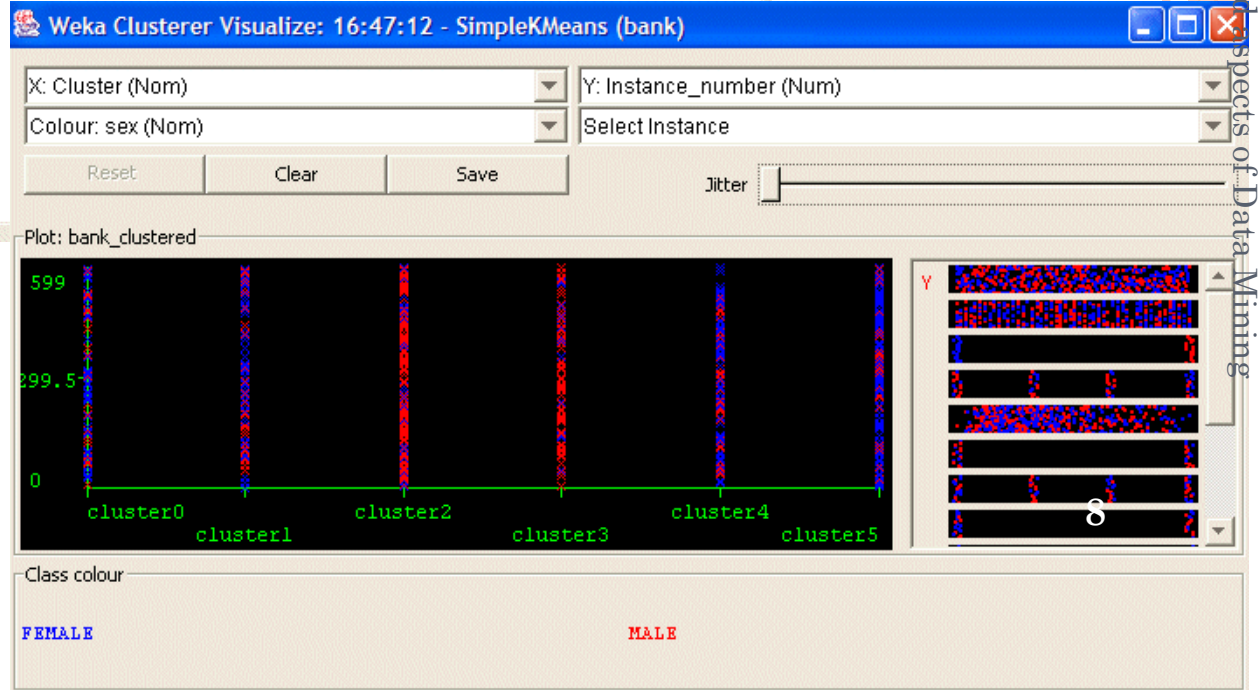
Number of iterations: 9

Cluster centroids:

Cluster	Mean/Mode:	36.6061	FEMALE	RURAL	23215.9002	NO	3	NO	YES	YES	NO	NO
	Std Devs:	14.4317	N/A	N/A	12378.3336	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cluster 1	Mean/Mode:	38.1176	FEMALE	INNER_CITY	24775.7982	YES	1	NO	YES	YES	YES	YES
	Std Devs:	13.793	N/A	N/A	12444.5713	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cluster 2	Mean/Mode:	44.0479	MALE	INNER_CITY	28547.224	YES	0	YES	YES	YES	NO	NO
	Std Devs:	14.2211	N/A	N/A	12696.4468	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cluster 3	Mean/Mode:	40.5068	MALE	TOWN	25975.293	YES	0	YES	NO	YES	YES	YES
	Std Devs:	13.6353	N/A	N/A	11111.66	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cluster 4	Mean/Mode:	49.7843	FEMALE	INNER_CITY	33917.4538	NO	0	YES	YES	YES	NO	YES
	Std Devs:	13.6872	N/A	N/A	14195.1688	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Cluster 5	Mean/Mode:	41.5234	FEMALE	TOWN	26191.8366	YES	0	NO	YES	YES	NO	NO
	Std Devs:	13.5728	N/A	N/A	11737.3135	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Clustered Instances

0	66 (11%)
1	85 (14%)
2	146 (24%)
3	73 (12%)
4	102 (17%)
5	128 (21%)





HOW TO ENRICH DATA MINING WITH SEMANTICS?



..... SEMANTIC ENRICHMENT!!!!!!

Integrating semantics in the knowledge
discovery process

WHICH SEMANTICS?

Semantics...

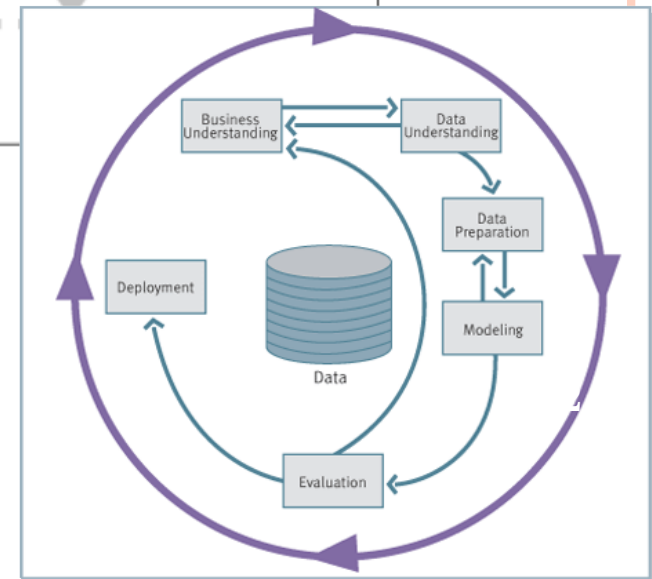
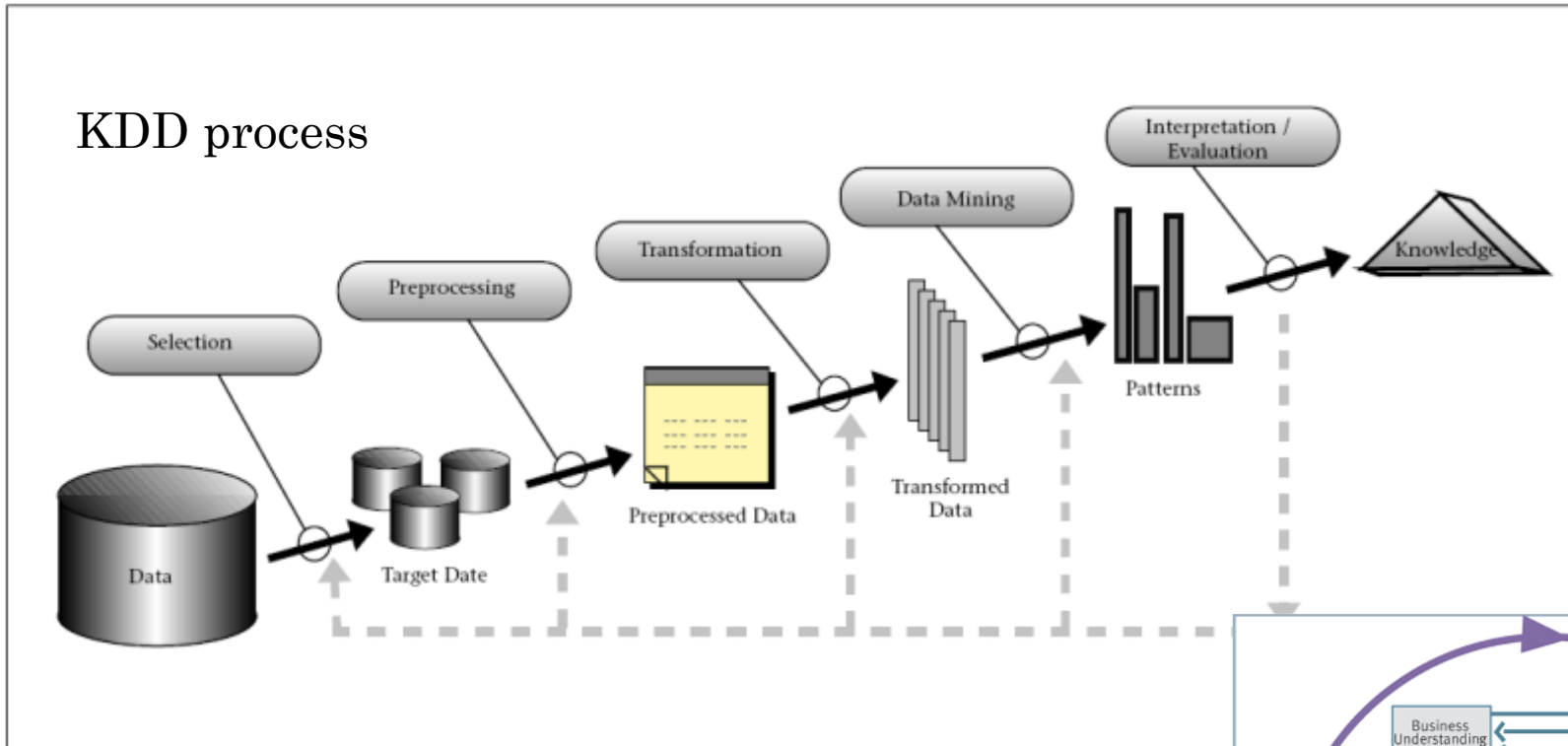
- May come explicitly from the **domain expert**
 - Example: the knowledge about symptoms and disease



- May be obtained from the **context**
 - Example: items that appears together with the pattern or the geographical area where an event happened, or specific attributes of the mined objects.



THE KNOWLEDGE DISCOVERY PROCESS



CRISP-DM process

DOMAIN KNOWLEDGE

- Domain knowledge - or background knowledge – represents contextual information
- This knowledge usually comes from the domain expert. Therefore is difficult to obtain!
- For example, the knowledge about the correlation between given symptoms and a disease, the semantic relationships between items sold in a supermarket, etc.

Some patterns are well known (thus useless), others are unknown but useless, others are UNKNOWN and USEFUL

Milk sold together with cookies may be a well known correlation, while diapers with beer is unknown and interesting

SEMANTIC ENRICHMENT: WHEN, HOW AND WHERE?

WHEN enriching: one step or the entire data mining process?

Some approaches in the literature work on single steps others on the entire process, some apply semantic enrichment to pre or post preprocessing, etc.

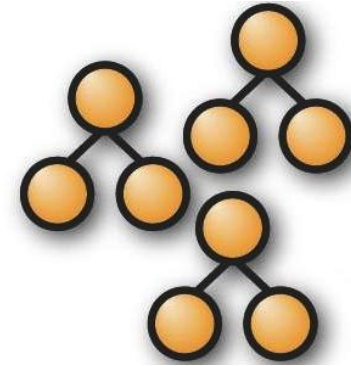
HOW to enrich the mining process/task?

transforming the datasets or modifying the algorithms or improving the postprocessing...

WHERE to store/represent semantics?

user may interact, but usually it is represented in ontologies or taxonomies

WHAT IS AN ONTOLOGY?



An ontology is a **shared conceptualization of a domain**

- A **formal** ontology is a set of definitions in a formal language for terms describing the world
- Ontologies are typically designed by an ontology engineer or a domain expert to formally represent some knowledge

ELEMENTS OF AN ONTOLOGY

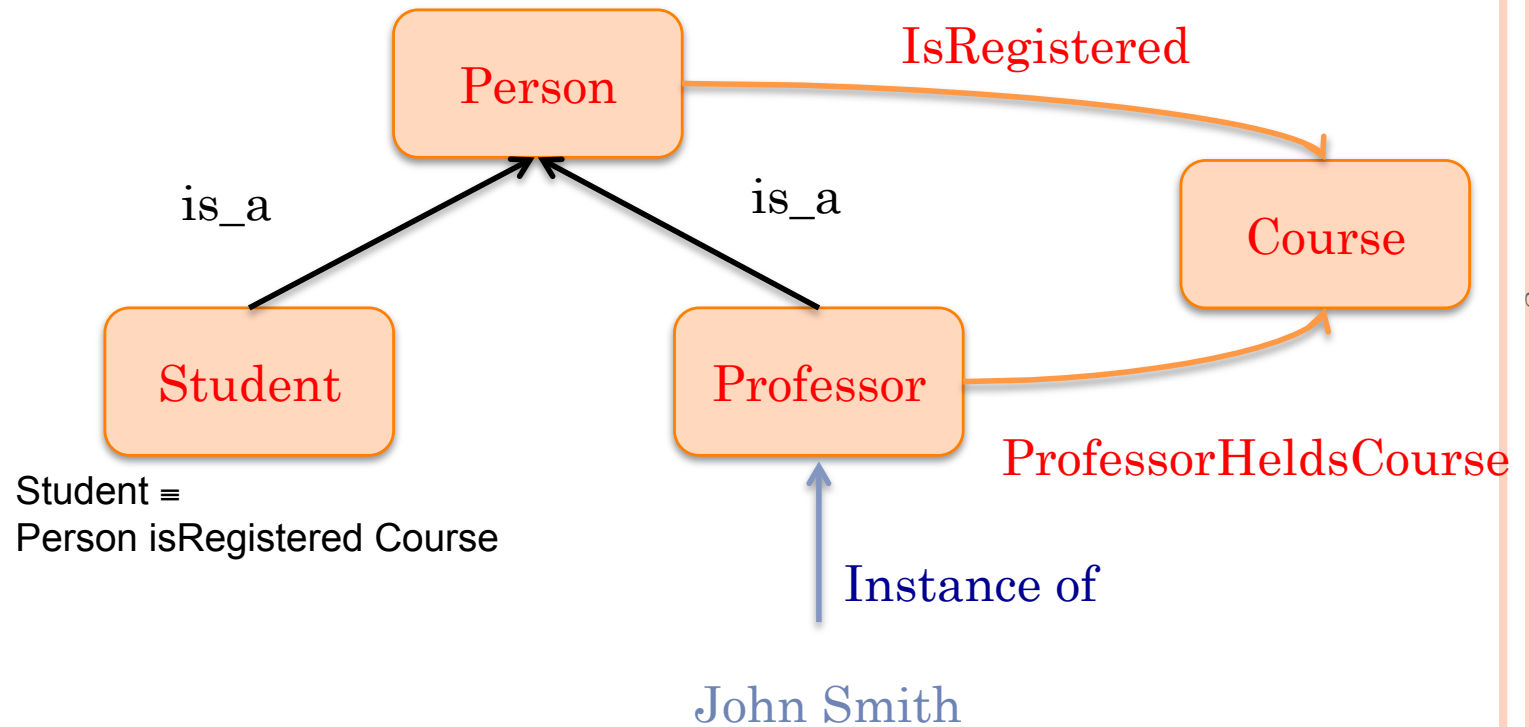
Main components of a formal ontology are:

- **Concepts (or classes)**: concepts of the domain.
 - Student, Course and Professor are three classes in a University ontology
- **Relationships** between concepts.
 - *professorHeldCourse* may connect the classes Professor and Course.
- **Is_a hierarchy**: represents the *kind of* relationship, or father-child, or subclass
 - A Student *isa* Person

ELEMENTS OF AN ONTOLOGY

- **Instances:** are specific elements of the domain.
 - a professor called John Smith is an instance of the Professor class
- **Axioms:** represent formal sentences that are always true. Axioms are associated to classes thus defining the instances belonging to that class
 - A student is a person registered to a course

AN EXAMPLE OF AN ONTOLOGY



ONTOLOGY LANGUAGES



```
<owl:DatatypeProperty rdf:about="#ResponseTime">
  <rdfs:domain rdf:resource="#ResponseTimeAssertion"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#int"/>
  <ns_1:assertionPropertyComparator rdf:datatype="http://www.w3.org/
2001/XMLSchema#string">
    LessThanComparator
  </ns_1:assertionPropertyComparator>
  <ns_1:controlWidget rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    integer
  </ns_1:controlWidget>
</owl:DatatypeProperty>

<owl:Class rdf:about="#ResponseTimeAssertion">
  <rdfs:subClassOf rdf:resource="http://www.webifysolutions.com/
2005/10/catalog/assertion#InteroperabilityAssertion"/>
</owl:Class>

<owl:DatatypeProperty rdf:about="#VisaType">
  <rdfs:domain rdf:resource="#VisaTypeAssertion"/>
  <rdfs:range>
    <owl:DataRange>
      <owl:oneOf rdf:parseType="Resource">
        <rdf:first rdf:datatype="http://www.w3.org/
2001/XMLSchema#string">
          Tourist
        </rdf:first>
        <rdf:rest rdf:parseType="Resource">
          <rdf:first rdf:datatype="http://www.w3.org/
2001/XMLSchema#string">
            Business
          </rdf:first>
          <rdf:rest rdf:parseType="Resource">
            <rdf:first rdf:datatype="http://www.w3.org/
2001/XMLSchema#string">
              Student
            </rdf:first>
            <rdf:rest rdf:resource="http://www.w3.org/
1999/02/22-rdf-syntax-ns#nil"/>
          </rdf:rest>
        </rdf:rest>
      </owl:oneOf>
    </owl:DataRange>
  </rdfs:range>
  <ns_1:assertionPropertyComparator rdf:datatype="http://www.w3.org/
2001/XMLSchema#string">
    EqualsComparator
  </ns_1:assertionPropertyComparator>
  <ns_1:controlWidget rdf:datatype="http://www.w3.org/
2001/XMLSchema#string">
    dataRange
  </ns_1:controlWidget>
</owl:DatatypeProperty>

<owl:Class rdf:about="#VisaTypeAssertion">
  <rdfs:subClassOf rdf:resource="http://www.webifysolutions.com/
2005/10/catalog/assertion#ContentBasedAssertion"/>
</owl:Class>
```

- There are several languages used to define ontologies.
- A well known standard is OWL (Semantic web)
- Based on Description Logics
- An OWL file is a text file written in XML

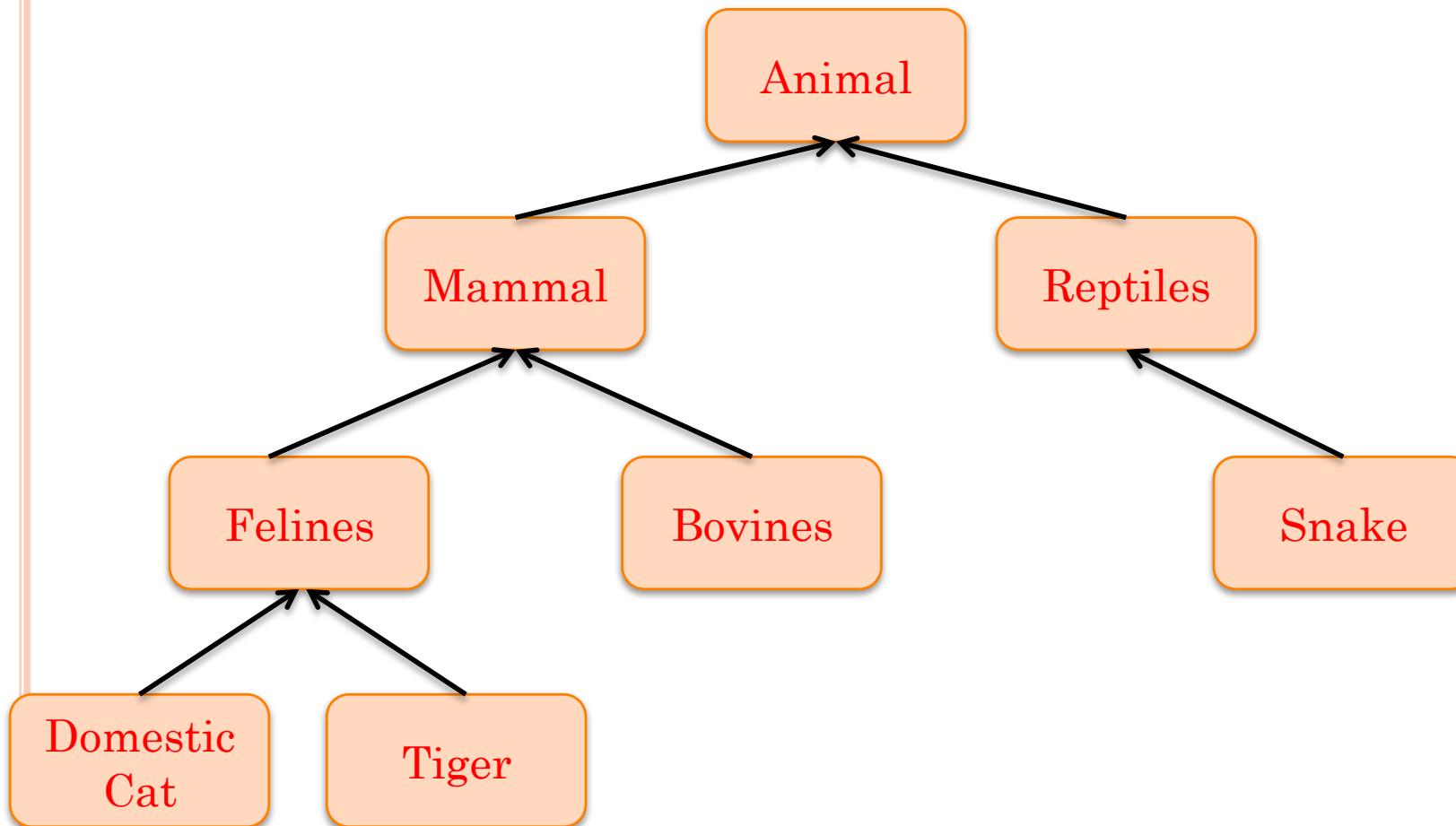
ONTOLOGY REASONERS



- Ontology languages are typically accompanied by an inference engine called **reasoner**
- It can perform some reasoning task:
 - subsumption – check the subclass relationship and check if the ontology is consistent
 - instance checking – checking if an instance belongs to a class.
- Usually in semantic data mining reasoners are not used and the ontology is used as a conceptual map describing domain knowledge

TAXONOMY

An ontology with only *is_a* relationships is called a **taxonomy**

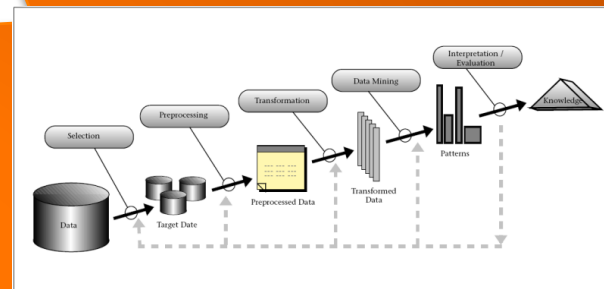


ORGANIZATION OF THE SURVEY

- Semantic enrichment applies to the whole knowledge discovery process or specific steps;
- Data preprocessing and postprocessing based on domain knowledge;
- Navigate the extracted pattern
 - Data Mining Query Languages

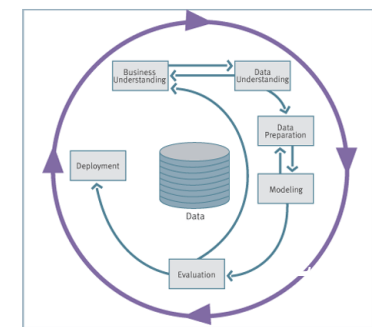
SEMANTIC ENRICHMENT OF THE PROCESS

Objective: integrate domain knowledge along the
KDD/CRISP- DM process



ROLES OF MEDICAL ONTOLOGY IN ASSOCIATION MINING CRISP-DM CYCLE

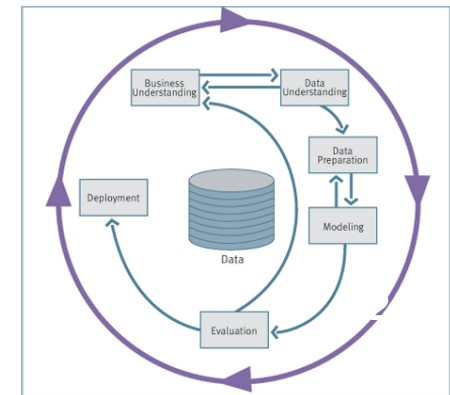
- Paper [1] describes domain knowledge in an ontology to be integrated to each step of the CRISP-DM process.
 - In **Business Understanding** ontologies help to get a better understanding of the domain
 - In **Data Understanding** map ontology elements to data. Discover missing/redundant attributes
 - In **Data Preparation** ontologies may help in selecting groups of attributes for DM tasks



[1] Čeřpivov´a, H., Rauch, J., Sv´atek V., Kejkula M., Tomeĉkov´a M.: Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa 2004.

ROLES OF MEDICAL ONTOLOGY IN ASSOCIATION MINING CRISP-DM CYCLE

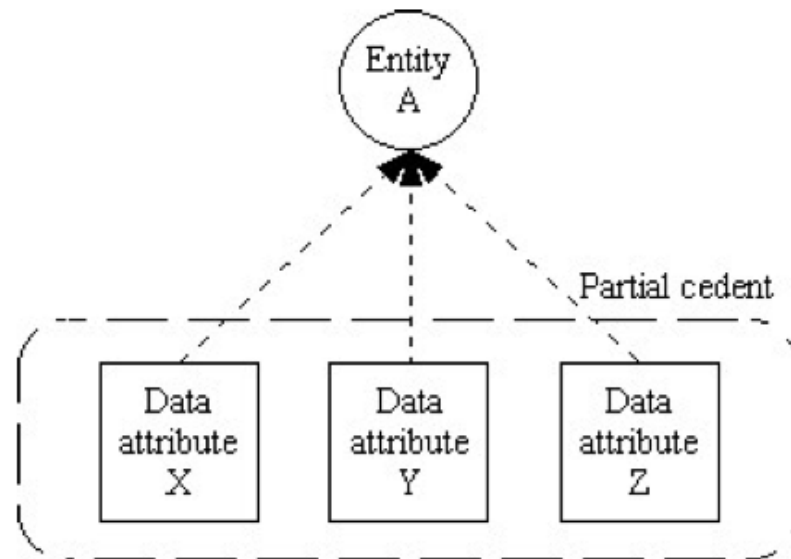
- In the **Modelling** phase helps in designing the mining sessions
- In the **Evaluation** phase patterns may be interpreted in terms of background knowledge
- In the **Deployment** phase mining results are mapped back the ontology for the easily distribution of results



ROLES OF MEDICAL ONTOLOGY IN ASSOCIATION MINING CRISP-DM CYCLE

Experiment on a medical dataset

- **Data Understanding:** They mapped the STULONG* attributes to the UMLS^ ontology → find redundant attributes
- **Data Preparation:** Grouped attributes into groups based on the ontology:

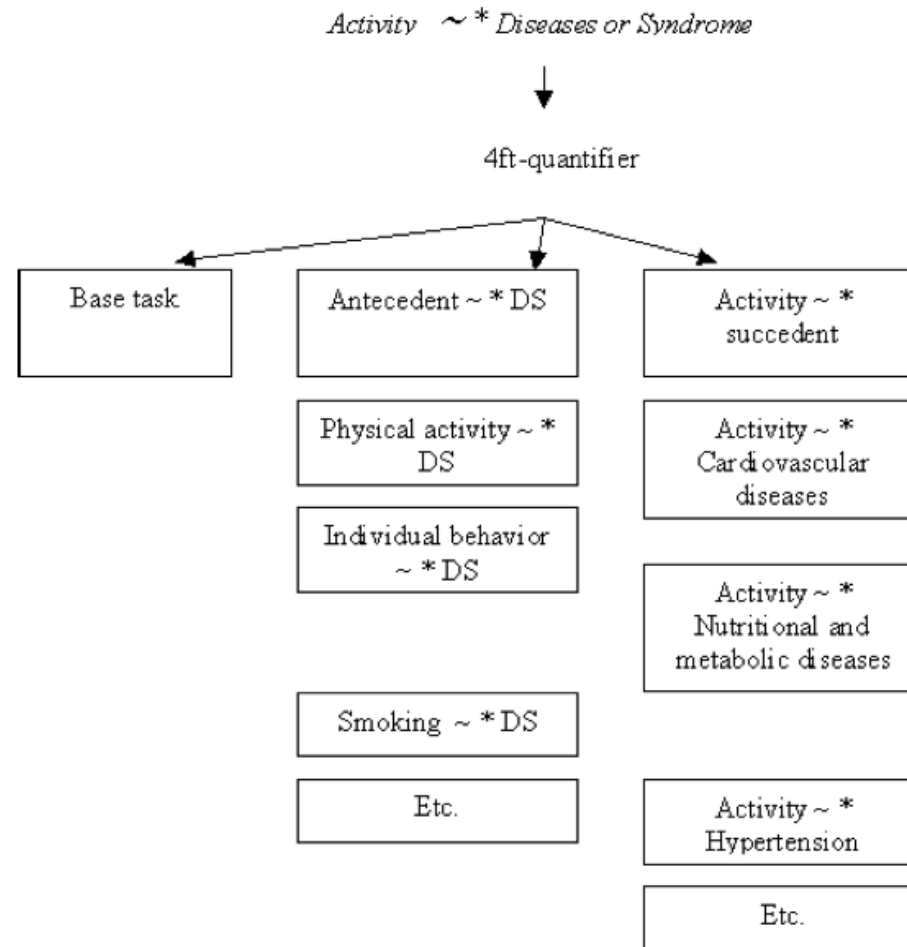


(*) Publicly available dataset on cardiovascular disease <http://euromise.vse.cz/challenge2004/data/index.html>

(^) A Medical Ontology

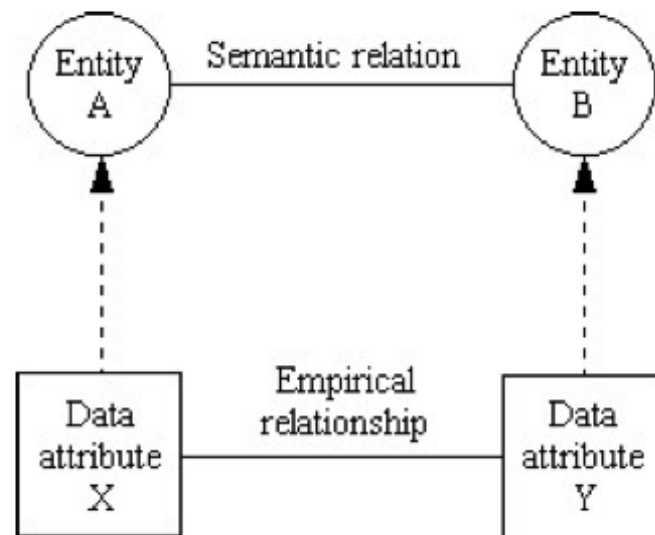
ROLES OF MEDICAL ONTOLOGY IN ASSOCIATION MINING CRISP-DM CYCLE

- Mining phase: Use ontologies to design the individual sessions of the mining tasks, conceptually more homogeneous



ROLES OF MEDICAL ONTOLOGY IN ASSOCIATION MINING CRISP-DM CYCLE

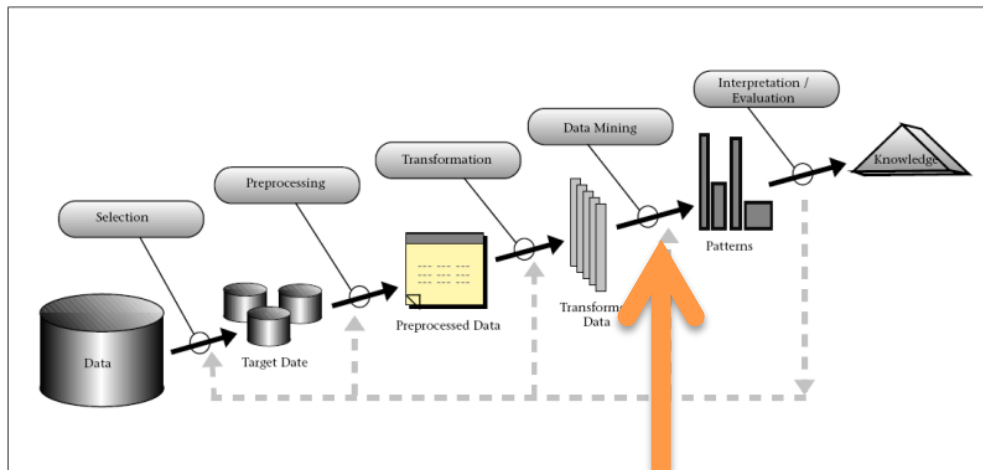
- **Result Evaluation:** Association rule results are matched to semantic relations that represents the explanation of the discovered association.



- Confirmation of prior knowledge, new knowledge compatible with prior knowledge, or conflict.



SEMANTIC ENRICHMENT OF DATA MINING TASKS



SEMANTIC ENRICHMENT OF DATA MINING TASKS

- Approaches that adapt DM algorithms to integrate background knowledge
- Generally use ontologies and/or interactions with human experts
- Paper [1] proposes to use background knowledge introduced manually by the user as rules, to extract association rules that are consistent with the background knowledge.
- Association Rules are Confirmations, Exceptions or New Knowledge.

MODIFICATION OF DATA MINING ALGORITHMS

- In these approaches the data mining algorithms are modified to take into account semantics.
- Use of concept hierarchy or constraints
- Onto4AR [2] is an example of this kind of approaches

[2] Cláudia Antunes. Onto4AR: a framework for mining association rules , in Workshop on Constraint-Based Mining and Learning (CMILE - ECML/PKDD 2007), Warsaw, September 2007.

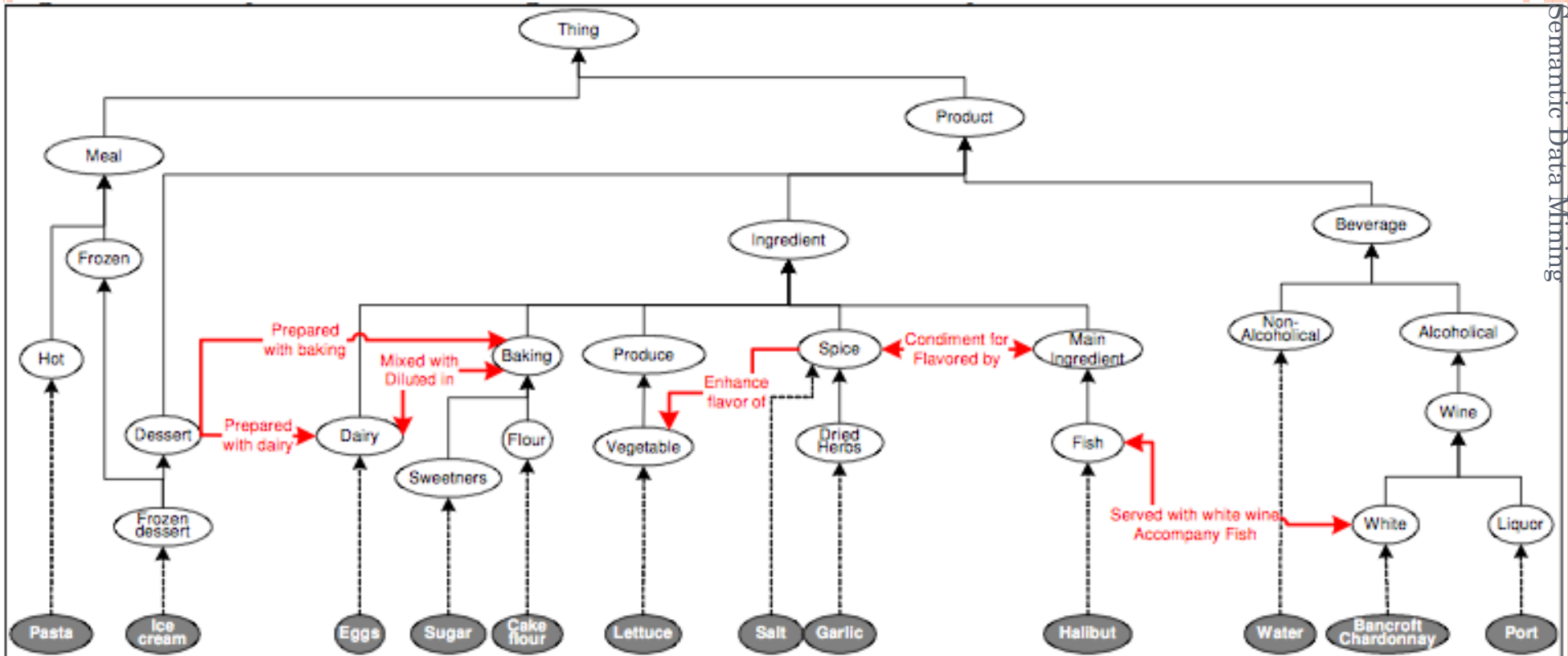
ONTO4AR: A FRAMEWORK FOR MINING ASSOCIATION RULES

This approach **applies to association rules**

- Aside the interestingness measures (confidence, etc.) they propose content constraints, both based on ontologies.

Given a transactions dataset D and an ontology O with instances I , find all association rules in the form $A \rightarrow B$ where A and B are disjoint itemsets of I that may occur on D , and the itemsets A and B satisfy **a set of constraints C defined over O** .

ONTO4AR: A FRAMEWORK FOR MINING ASSOCIATION RULES



ONTO4AR

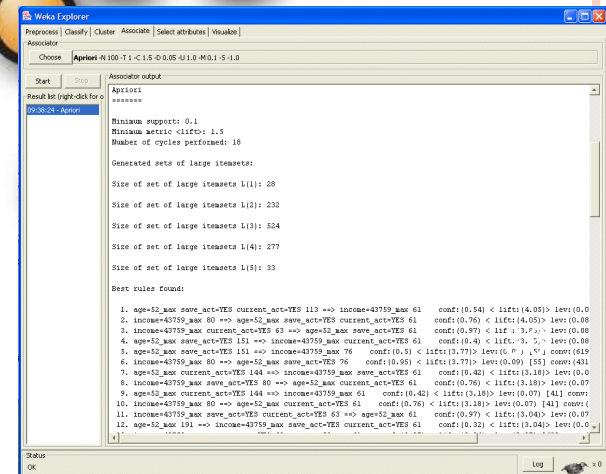
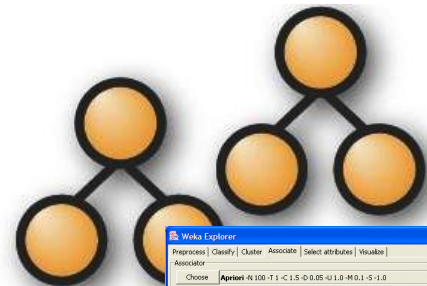
Constraints are used to prune the candidate generation. Two kinds of constraints:

- Taxonomical constraints – based on the child-parent relationship between classes
 - {Cake flour, Sugar} satisfies the **same-family** constraint since Baking is the common father.
- Non-taxonomical constraints – based on the relations between classes
 - {Halibut, Bancroft Chardonnay} is **strongly connected** since there is a relation in the ontology between their parents

ONTO4AR

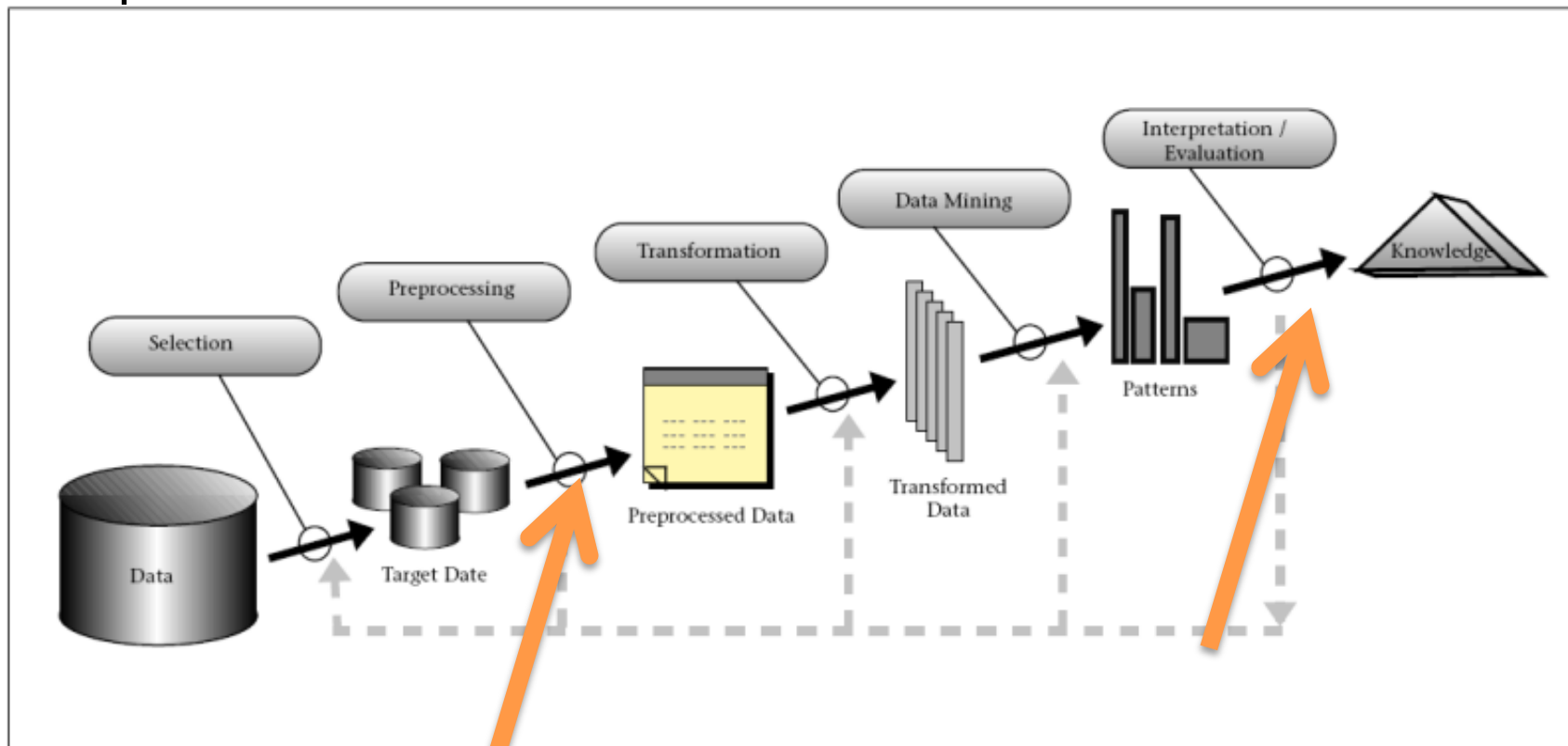
Frequent pattern mining algorithm is modified to add a **pruning step**: candidates that does not satisfy the constraints defined in the ontology are disregarded

REFINE INTERMEDIATE RESULTS WITH DOMAIN KNOWLEDGE



REFINE INTERMEDIATE RESULTS WITH DOMAIN KNOWLEDGE

These approaches consider semantics integration into the preprocessing and/or post processing

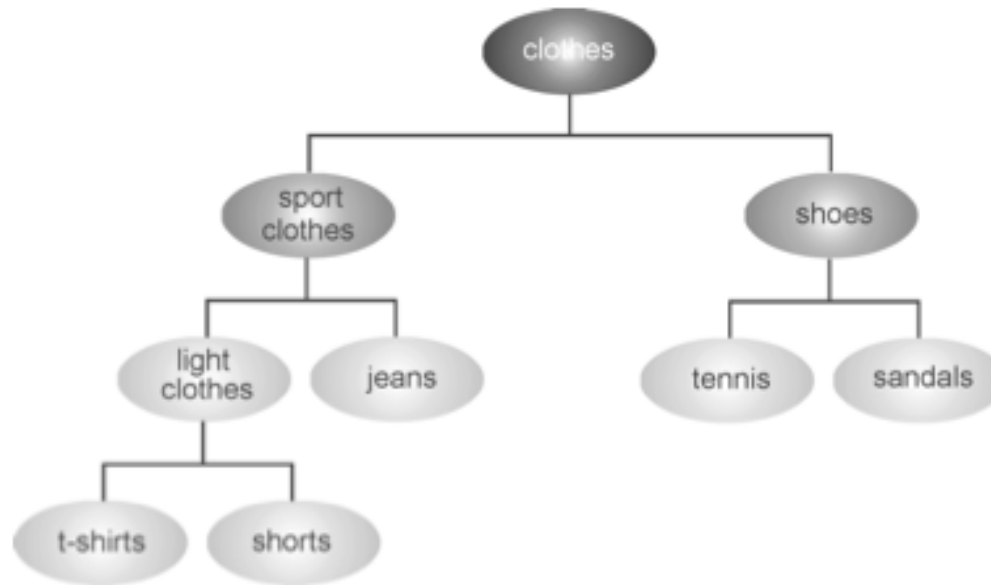


REFINE INTERMEDIATE RESULTS WITH DOMAIN KNOWLEDGE: DATA PREPROCESSING

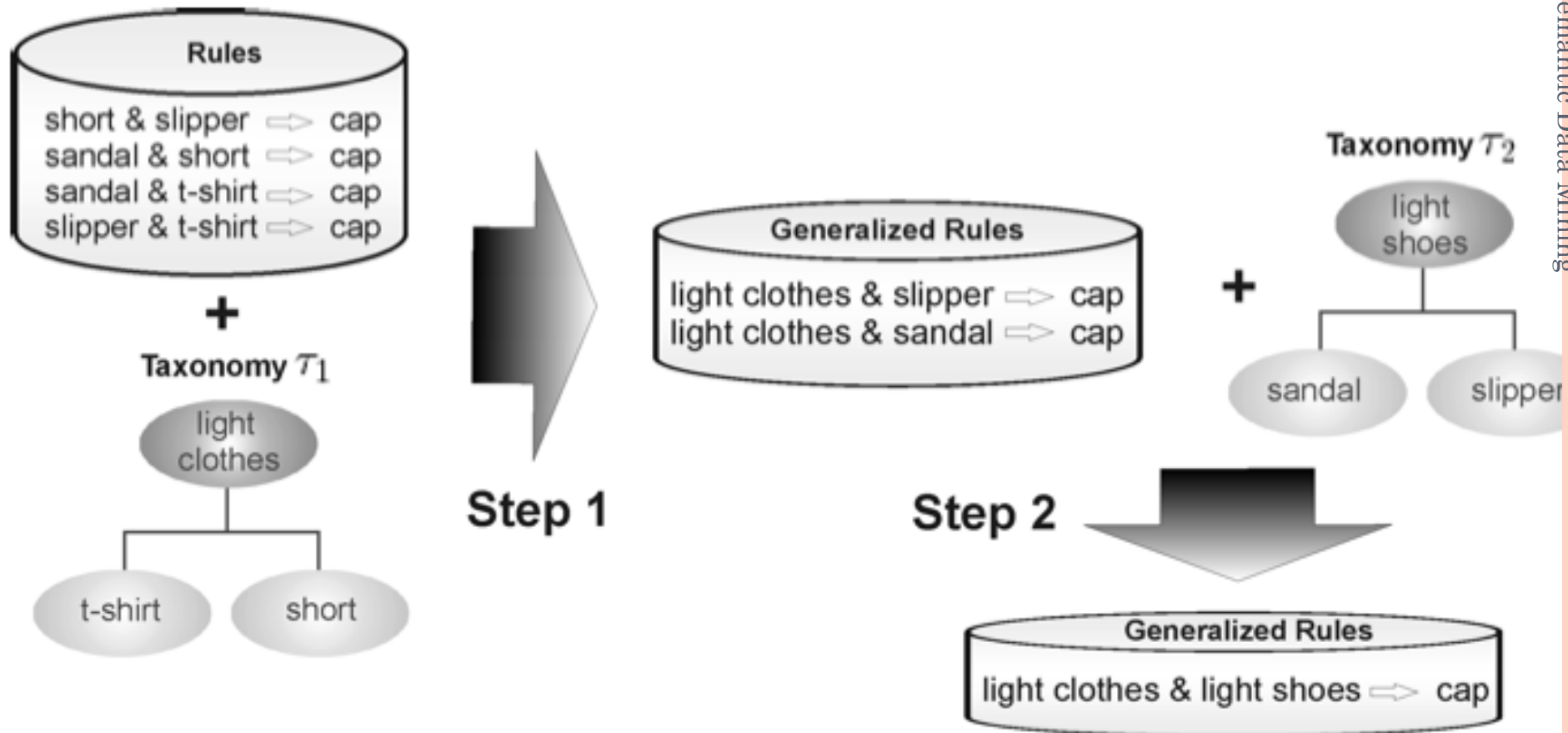
- Paper [1] proposed to match the source attributes with the corresponding ontology concepts
- Therefore the initial dataset is reduced incorporating some semantics coming from the ontology

REFINE INTERMEDIATE RESULTS WITH DOMAIN KNOWLEDGE: DATA POSTPROCESSING

- The GART approach: generalize association rules using taxonomies → reduce the number of rules

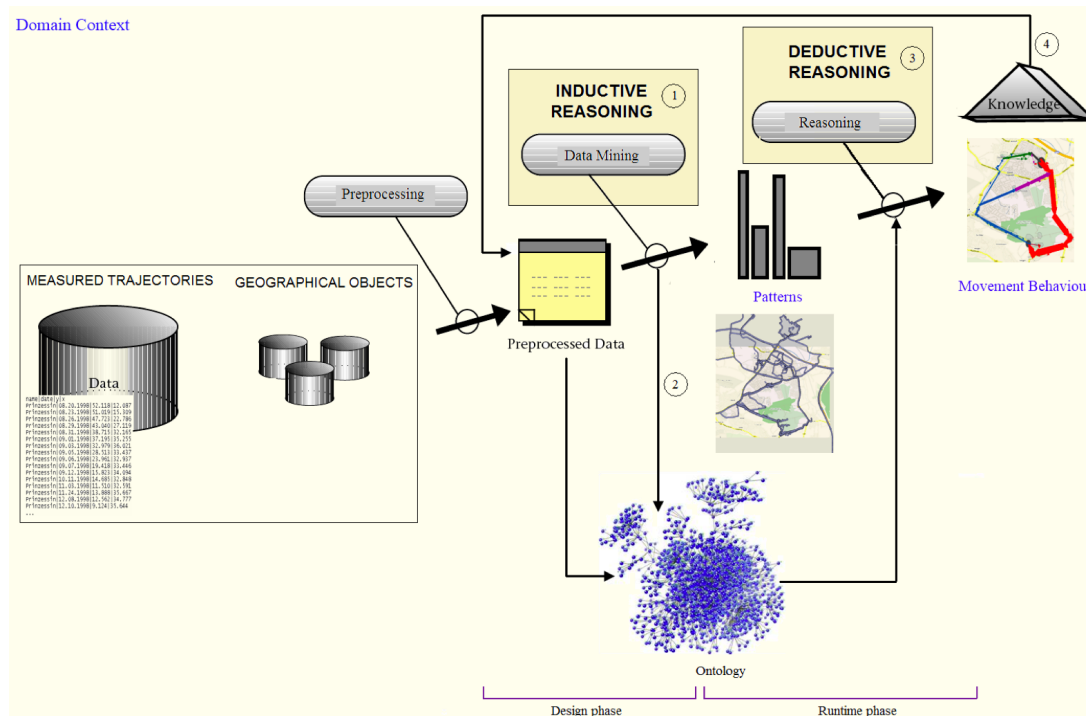


USING ONTOLOGIES TO FACILITATE THE ANALYSIS OF ASSOCIATION RULES

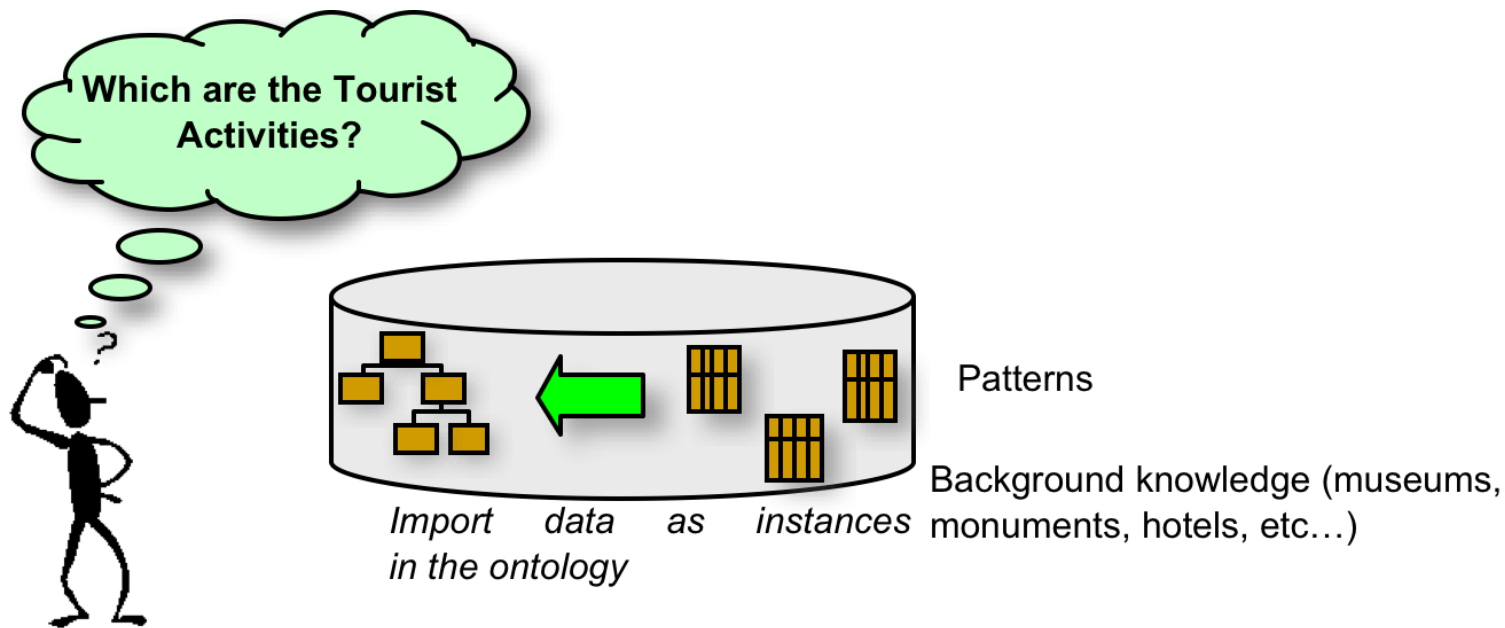


REFINE INTERMEDIATE RESULTS WITH DOMAIN KNOWLEDGE: DATA POSTPROCESSING

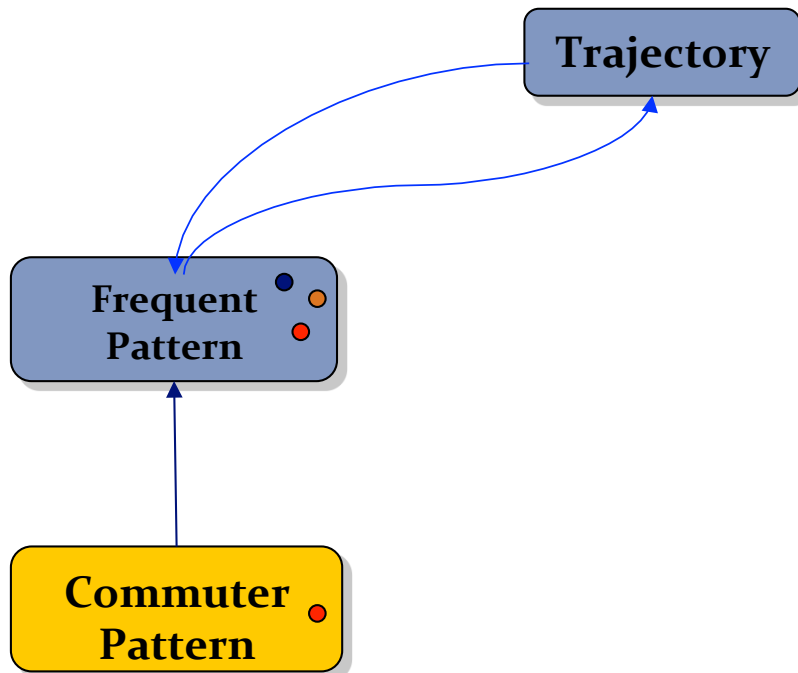
This approach uses ontologies to classify computed patterns into predefined ontology classes



TOWARDS SEMANTIC INTERPRETATION OF MOVEMENT BEHAVIOR



TOWARDS SEMANTIC INTERPRETATION OF MOVEMENT BEHAVIOR



Trajectory data populates a domain ontology and the **reasoning engine** classifies trajectories into the class satisfying the concept definition (axiom).

Commuter Pattern ≡ a pattern frequently starting outside the city, stopping inside the city for a long time and going back outside the city

Ontology Axiom

PATTERN INTERPRETATION FRAMEWORK FOR MOVEMENT DATA

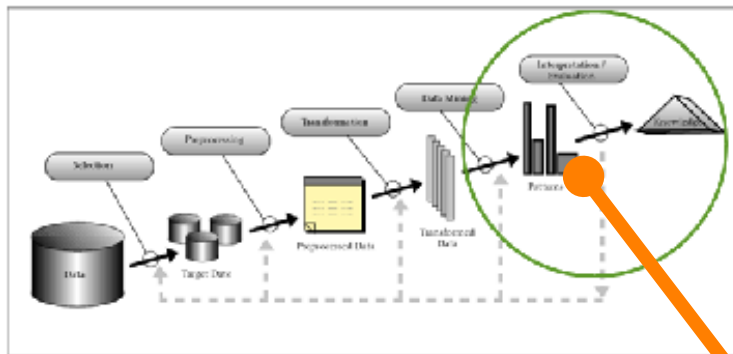
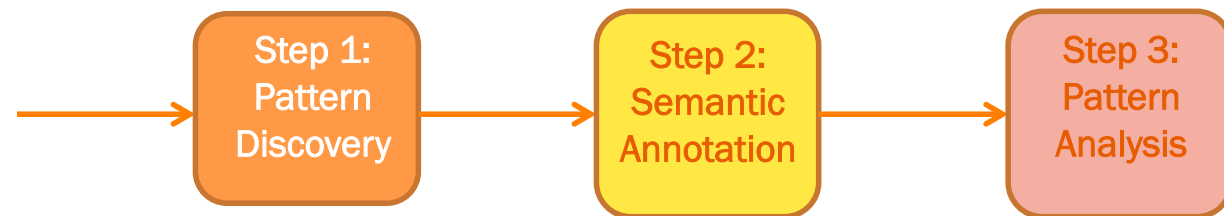


Figure 2: Stages that compose the KDD process - Fayyad 1996.

This method can be seen as a post processing but also a preprocessing.

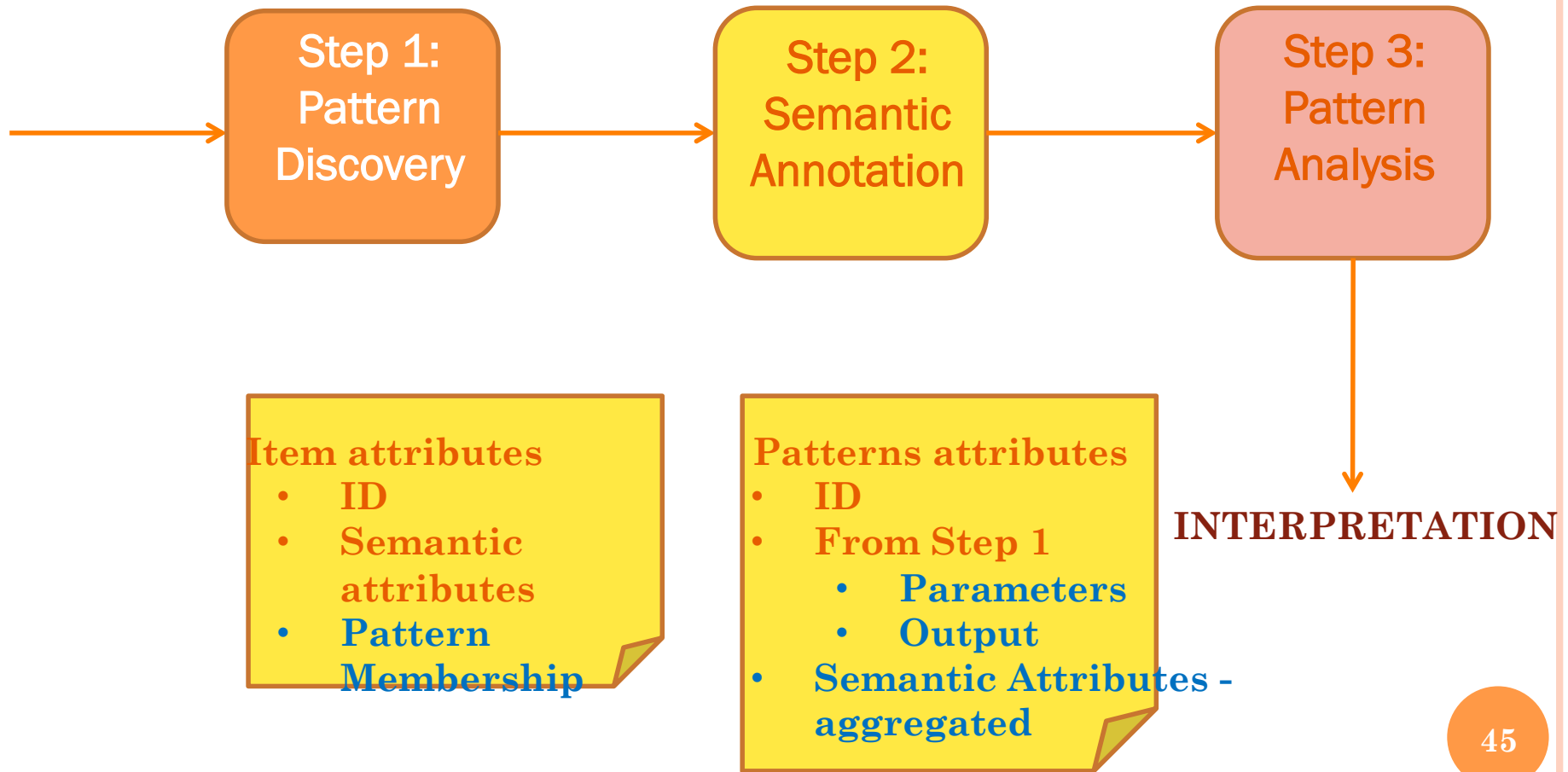
Patterns are annotated with semantic information and then mined again



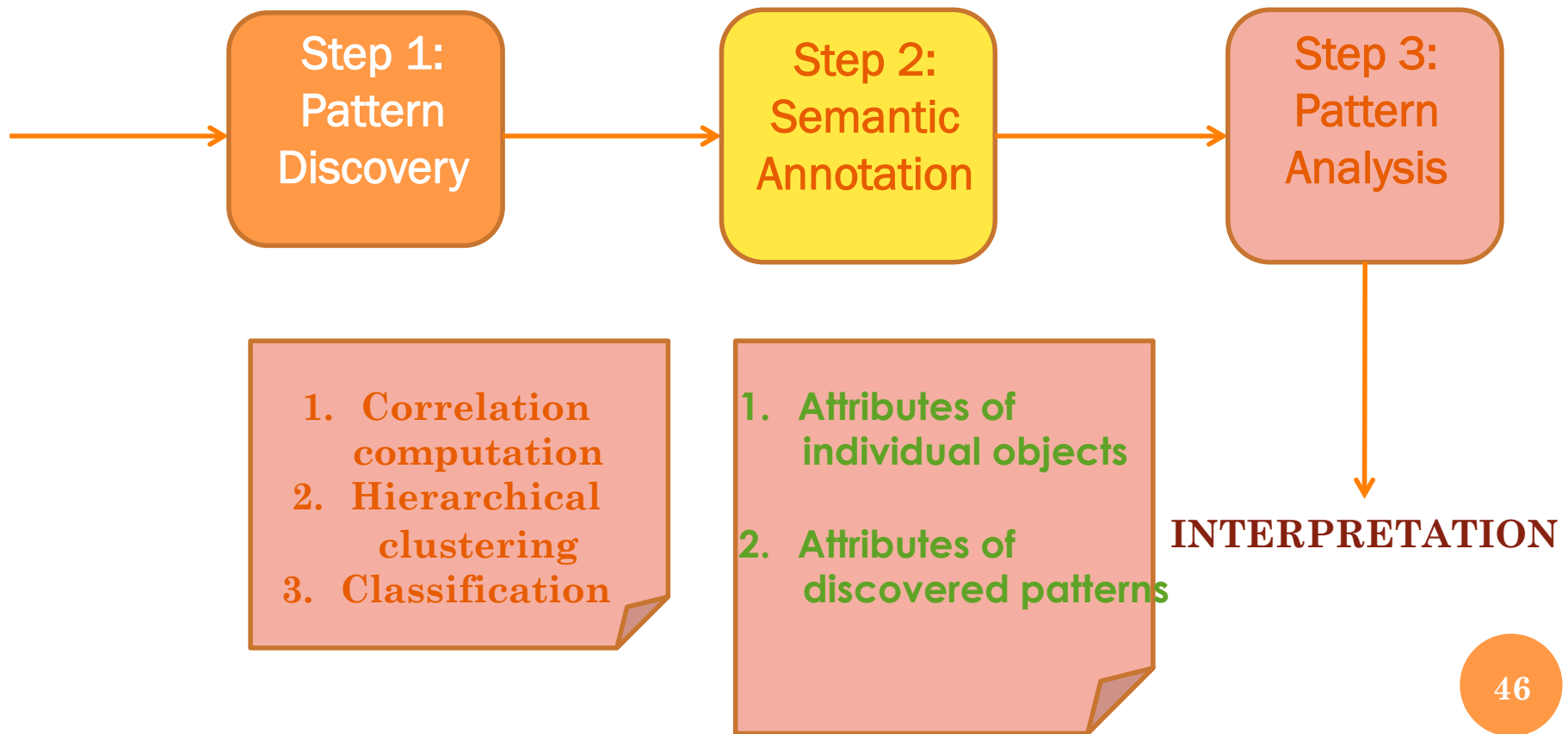
44

INTERPRETATION

STEP 2: SEMANTIC ANNOTATION

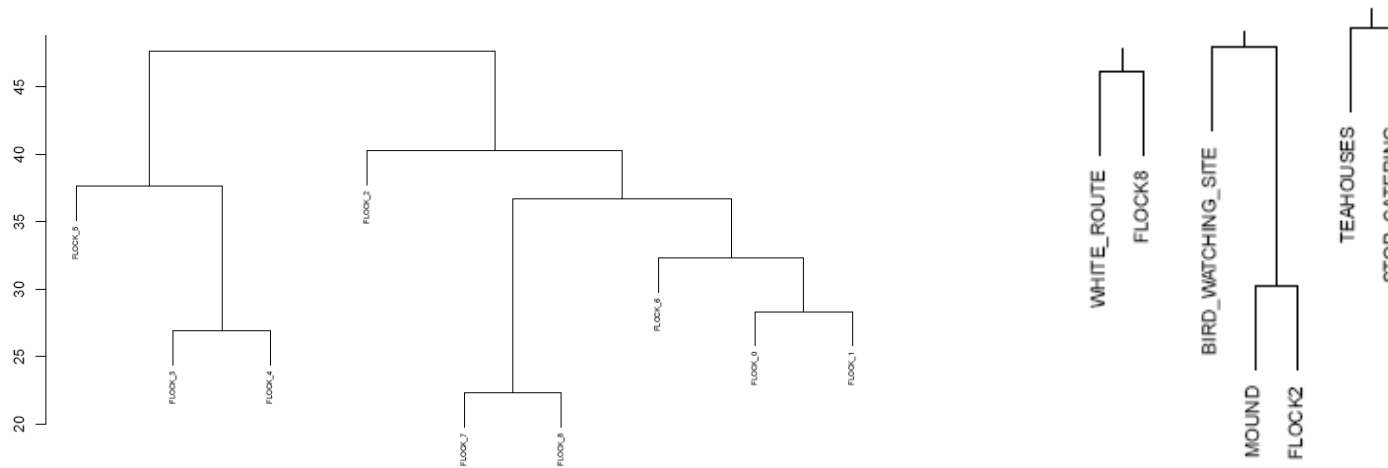


STEP 3: PATTERN ANALYSIS



PATTERN ANALYSIS

- This framework has been applied to mobility data from a dataset of pedestrians moving in a park.
- Semantics (annotations) comes from questionnaires filled by users.
- The objective is to mine patterns enriched with semantic information.





NAVIGATE THE EXTRACTED PATTERNS

The idea is to provide the user with a tool to navigate the extracted patterns in a meaningful way

USING ONTOLOGIES TO FACILITATE THE ANALYSIS OF ASSOCIATION RULES – GART [3]

The screenshot shows the RuIEE-GAR web application interface. The main content area is titled 'Analyze Generalized Rules' and includes the following controls:

- Columns:** A text input field containing 'Rule, Sup, Cov', a 'Column' dropdown menu, and an 'Add' button.
- Rule Base:** A dropdown menu showing 'ASSOCIATION 72 - Generalized Association Rule mod' and a 'Remove' button.
- Restrictions:** A text input field, an 'AND' dropdown menu, a 'Column' dropdown menu, an 'Operator' dropdown menu, and an 'Add' button.
- Sorting:** A text input field, a 'Column' dropdown menu, and an 'Add' button.
- A 'List Rules' button.
- A 'Save Information' button.

Below the controls, the interface displays the following information:

Rules before generalization: **8** and Rules after generalization: **4** Downloads: [Data Set](#) | [Rule Set](#) | [Generalized Rule Set](#) | [Taxonomy Set](#)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Rule	Sup	Cov
<input type="checkbox"/>	E S		IF (calcados_abertos) & (roupas_leves) THEN bone	0.6667	0.8334
<input type="checkbox"/>			IF camiseta & sandalia THEN short	0.1667	0.3334
<input type="checkbox"/>		M	IF chinelo & sandalia THEN short	0.1667	0.1667
<input type="checkbox"/>	E S		IF (roupas_leves) THEN tenis	0.8333	0.8333

E = Expanded Rule, S = Source Rules, M = Measures.

NAVIGATE PATTERNS - SEMANTIC ENHANCEMENT OF PATTERN STORAGE AND QUERYING

- Inductive database paradigm [6]
- Databases storing data and the inductively inferred patterns.
- Patterns are stored in the database and can be queried

- Approaches for a Data Mining Query Language

INDUCTIVE DATABASES

- In his pioneering work Mannila formalized the notion of Inductive Database as a relational database with inductive rules.
- The inductive database can be queried. For example association rule discovery can be expressed by a query

alarm type	alarming element	element type	date	time	week	alarm severity	alarm text
1111	E1.1	ABC	980119	233605	4	1	LINK FAILURE
2222	E2	CDE	980119	233611	4	3	HIGH ERROR RATE
3333	A	EFG	980119	233627	4	1	CONNECTION NOT ESTABLISHED
4444	B2.1	GHI	980119	233628	4	2	LINK FAILURE

s_0	$e(r_0).f$	$e(r_0).c$
alarm_type=1111 \Rightarrow element_type=ABC	0.25	1.00
alarm_type=222 \Rightarrow alarming_element=E2, element_type=CDE	0.25	1.00
alarm_type=1111, element_type=ABC \Rightarrow alarm_text=LINK_FAILURE	0.25	1.00
alarm_type=5555 \Rightarrow alarm_severity=1	0.00	0.00

DATA MINING QUERY LANGUAGES - DMQL

- A data mining task can be specified in the form of a data mining query
- Create and manipulate data mining models through a SQL-based interface
- Abstract away the data mining particulars
- Data mining can be performed on data in the database
- Approaches differ on what kinds of models should be created, and what operations we should be able to perform

A DATA MINING QUERY:

A data mining query is defined in terms of data mining task primitives:

- The set of task-relevant **data to be mined**
- The kind of **knowledge** to be mined
- The **background knowledge** to be used in the discovery process
- The **interestingness measures** and thresholds for pattern evaluation
- The expected representation for **visualizing the discovered patterns**

DATA MINING QUERY LANGUAGES

The so-called **closure property** means that the results of data mining tasks can be stored and possibly mined again

It has been inspired by Inductive databases and enables to combine queries in sequences and scripts.

DMQL – TWO APPROACHES

- The first one assumes that data and pattern storage systems and solvers are **already embedded into a common system.**
 - DMQL and MINE RULE are representative of this approach.
- A second approach assumes that storage systems are loosely coupled with solvers
 - OLE DB for DM (Microsoft). It is an API between different components that also provides a language for creating and filling extraction contexts, and then access them for manipulations and tests

DMQL - HAN ET AL.

- It has been designed to support various rule mining extractions: classification rules, association rules.
- Definition of meta-patterns, to restrict the syntactic aspect of the extracted rule
 - $\text{buy}^+(X, Y) \wedge \text{town}(X, \text{Berlin}) \Rightarrow \text{buy}(X, Z)$
 - restricts the search to association rules with implication between bought products for customers living in Berlin. Symbol + denotes that the predicate “buy” can appear several times in the left part of the rule.
 - also enables to define thresholds on the noise or novelty of extracted rules.
 - enables to define a hierarchy on attributes such that generalized association rules can be extracted.

[7] Jiawei Han , Yongjian Fu , Wei Wang , Krzysztof Koperski , Osmar Zaiane DMQL: A Data Mining Query Language for Relational Databases (1996)

[8] Jiawei Han, Yongjian Fu, Wei Wang, Jenny Chiang, Wan Gong, Krzysztof Koperski, Deyi Li, Yijun Lu, Amynmohamed Rajan, Nebojsa Stefanovic, Betty Xia, Osmar R. Zaiane: DBMiner: A System for Mining Knowledge in Large Relational Databases. KDD 1996: 250-255

DMQL – HAN ET AL.

Syntax

Specify background knowledge	→	use database <database_name> {use hierarchy <hierarchy_name> for <attribute>}
Specify rules to be discovered	→	<rule_spec>
Relevant attributes or aggregations	→	related to <attr_or_agg_list> from <relation(s)>
Collect the set of relevant data to mine	→	[where <conditions>] [order by <order list>]
Specify threshold parameters	→	{with [<kinds of>] threshold = <threshold_value> [for <attribute(s)>]}

DMQL - EXAMPLE

```
use database Hospital
find association rules as Heart_Health
related to Salary, Age, Smoker,
    Heart_Disease
from Patient_Financial f,
    Patient_Medical m
where f.ID = m.ID and m.age >= 18
with support threshold = .05
with confidence threshold = .7
```

MINE-RULE

Meo et al proposed an SQL operator called MINE RULE for mining association rules

```
MINE RULE SimpleAssociations AS
```

```
SELECT DISTINCT 1..n item AS BODY
```

```
1..1 item AS HEAD
```

```
SUPPORT, CONFIDENCE
```

```
FROM Purchase
```

```
GROUP BY transaction
```

```
EXTRACTING RULES WITH SUPPORT: 0.1
```

```
CONFIDENCE: 0.2
```

MINE-RULE

tr.	customer	item	date	price	q.ty
1	customer ₁	ski_pants	12/17/95	140	1
	customer ₁	hiking_boots	12/17/95	180	1
2	customer ₂	col_shirts	12/18/95	25	2
	customer ₂	brown_boots	12/18/95	150	1
	customer ₂	jackets	12/18/95	300	1
3	customer ₁	jackets	12/18/95	300	1
4	customer ₂	col_shirts	12/19/95	25	3
	customer ₂	jackets	12/19/95	300	2

Purchase



```

MINE RULE SimpleAssociations AS
  SELECT DISTINCT 1..n item AS BODY
    1..1 item AS HEAD
      SUPPORT, CONFIDENCE
FROM Purchase
  GROUP BY transaction
  EXTRACTING RULES WITH SUPPORT: 0.1
  CONFIDENCE: 0.2
    
```

BODY	HEAD	S.	C.
{ski_pants}	{hiking_boots}	0.25	1
{hiking_boots}	{ski_pants}	0.25	1
{col_shirts}	{brown_boots}	0.25	0.5
{col_shirts}	{jackets}	0.5	1
{brown_boots}	{col_shirts}	0.25	0.5
{brown_boots}	{jackets}	0.25	1
{jackets}	{col_shirts}	0.5	0.66
{jackets}	{brown_boots}	0.25	0.33
{col_shirts,brown_boots}	{jackets}	0.25	1
{col_shirts,jackets}	{brown_boots}	0.25	0.5
{brown_boots,jackets}	{col_shirts}	0.25	?

Simple Associations

MINE-RULE

MINE-RULE can use taxonomies (as GART) for reducing the number of association rules

```

MINE RULE BootsPantsRule AS
    SELECT DISTINCT item AS BODY,
           item AS HEAD
    SUPPORT, CONFIDENCE

WHERE HEAD.item IN (SELECT node FROM
ItemHierarchy WHERE ancestor=pants)
    AND BODY.item IN (SELECT node FROM
ItemHierarchy WHERE ancestor=boots)

FROM Purchase
    GROUP BY transaction
    EXTRACTING RULES WITH SUPPORT: 0.1
CONFIDENCE: 0.2
    
```



node	ancestor	level
hiking_boots	hiking_boots	0
hiking_boots	boots	1
hiking_boots	shoes	2
brown_boots	brown_boots	0
brown_boots	normal_boots	1
brown_boots	boots	2
brown_boots	shoes	3
...

SUMMARY AND CONCLUSIONS

- Semantic Data Mining develops techniques to embed semantics into the knowledge discovery process

There is no a standard method to exploit semantics in KDD

- Some approaches (1) enrich the KDD process, others (2) preprocess semantics transforming the dataset or the postprocessing or (3) modify the DM algorithms to take into account semantics or (4) navigate the extracted models

SUMMARY AND CONCLUSIONS



Semantic Data Mining is still in its infancy and the approaches are sometimes preliminary.

There are several open issues:

1. How to represent/embed semantics?
 - Ontology is the most used formalism, but how to build ontologies and to find a consensus is a drawback of this approach
 - Domain expert users are usually difficult to involve in the mining process – they are not DM experts
2. How to evaluate/validate the results?
 - Having considered semantics in the knowledge discovery should guarantee that the patterns are “semantic-enriched”.... but how to validate them?

REFERENCES USED IN THE SLIDES

- [1] Cespivova, H., Rauch, J., Svátek V., Kejkula M., Tomečková M.: Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa 2004.
- [2] Cláudia Antunes. Onto4AR: a framework for mining association rules , in Workshop on Constraint-Based Mining and Learning (CMILE - ECML/PKDD 2007), Warsaw, September 2007.
- [3] Domingues, Rezende, Using ontologies to facilitate the analysis of association rules workshop of knowledge discovery and ontologies PKDD2005
- [4] Baglioni, Macedo, Renso, Trasarti, Wachowicz Towards Semantic Interpretation of Movement Behavior, AGILE 2009
- [5] Rebecca Ong, Monica Wachowicz, Mirco Nanni, Chiara Renso: From Pattern Discovery to Pattern Interpretation in Movement Data. ICDM SADM Workshop 2010: 527-534
- [6] Mannila Inductive Databases and Condensed Representations for Data Mining, International Loginc Programming Symposium, 1997
- [7] Jiawei Han , Yongjian Fu , Wei Wang , Krzysztof Koperski , Osmar Zaiane DMQL: A Data Mining Query Language for Relational Databases (1996)
- [8] Jiawei Han, Yongjian Fu, Wei Wang, Jenny Chiang, Wan Gong, Krzysztof Koperski, Deyi Li, Yijun Lu, Aymn Mohamed Rajan, Nebojsa Stefanovic, Betty Xia, Osmar R. Zaiane: DBMiner: A System for Mining Knowledge in Large Relational Databases. KDD 1996: 250-255
- [9] Meo, Psaila, Ceri, A New SQL-like Operator for Mining Association Rules VLDB 1996

OTHER REFERENCES ON SEMANTIC DATA MINING

This survey is far to be complete! Several other approaches that adds semantics to knowledge discovery have been proposed in the last decade. Some of them are listed below:

- Ahmed Sultan Al-Hegami. Pruning based interestingness of mined classification patterns. *Int. Arab J. Inf. Technol.*, 6(4):336–343, 2009.4.
- Aijun An, Shakil M. Khan, and Xiangji Huang. Objective and subjective algorithms for grouping association rules. In *ICDM*, pages 477–480, 2003.5.
- Sarabjot S. Anand, David A. Bell, and John G. Hughes. The role of domain knowledge in data mining. In *CIKM '95: Proceedings of the fourth international Conference on Information and Knowledge Management*, pages 37–43, Baltimore, Maryland, United States, 1995. ACM.
- John M. Aronis and Foster J. Provost. Efficiently constructing relational features from background knowledge for inductive machine learning. In *KDD '94: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 347–358, 1994.
- John M. Aronis, Foster J. Provost, and Bruce G. Buchanan. Exploiting background knowledge in automated discovery. In *KDD '96: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 355–358, 1996.
- Abraham Bernstein, Foster Provost, and Shawndra Hill. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):503– 518, 2005.
- Vania Bogorny, Paulo Engel, and Luis Otavio Alvares. *Enhancing the Process of Knowledge Discovery in Geographic Databases using Geo-Ontologies*. Idea Group, 2007.
- Vania Bogorny, Joao Valiati, and Luis Alvares. *Semantic-based pruning of redundant and uninteresting frequent geographic patterns*. *GeoInformatica*, 2008.
- Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, New York, NY, USA, 1997. ACM.
- Laurent Brisson. Knowledge extraction using a conceptual information system (Ex- CIS), volume 4623 of *Lecture Notes in Computer Science*, pages 119–134. Springer Berlin / Heidelberg, 2007.
- Laurent Brisson and Martine Collard. *How to Semantically Enhance a Data Mining Process?*, volume 19 of *Lecture Notes in Business Information Processing*, chapter 3, pages 103–116. Springer Berlin Heidelberg, 2009.
- Mario Cannataro and Carmela Comito. A data mining ontology for grid programming. In *Proc. 1st Int. Workshop on Semantics in Peer-to-Peer and Grid Computing*, in conjunction with WWW2003, pages 113–134, 2003.

OTHER REFERENCES OF PAPERS ON SEMANTIC DATA MINING

- Mario Cannataro, Pietro Hiram Guzzi, Tommaso Mazza, and Pierangelo Veltri. Using ontologies in Proteus for modeling data mining analysis of proteomics experiments. In *Studies in Health Technologies and Informatics*, volume 112, pages 17 – 26,
- Longbing Cao. Behavior informatics and analytics: Let behavior talk. *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on Data Mining*, pages 87–96, 2008.
- Longbing Cao, Philip S. Yu, Chengqi Zhang, and Yanchang Zhao. *Data Mining for Business Applications*. Springer US, 2009.20..
- Longbing Cao and Chengqi Zhang. Domain-Driven Actionable Knowledge Discovery in the Real World, volume 3918/2006 of *Lecture Notes in Computer Science*, pages 821–830. Springer Berlin / Heidelberg, 2006.
- Doina Caragea, Jun Zhang, Jyotishman Pathak, and Vasant Honavar. Learning Classifiers from Distributed, Ontology-Extended Data Sources, volume 4081 of *Lecture Notes in Computer Science*, pages 363–373. Springer Berlin/ Heidelberg, 2006.
- Deborah Carvalho, Alex A. Freitas, and Nelson Ebecken. Evaluating the Correlation Between Objective Rule Interestingness Measures and Real Human Interest, volume 3721 of *Lecture Notes in Computer Science*, pages 453–461. Springer Berlin/Heidelberg, 2005.
- Barbara Catania. Towards effective solutions for pattern management. *IJCSA*, 5(3):36–45, 2008.
- M. Charest, S. Delisle, O. Cervantes, and Y. Shen. Intelligent data mining assistance via CBR and ontologies. *DEXA'06. 17th International Conference on Database and Expert Systems Applications*. Invited paper., pages 593–597, 2006.
- Xiaoming Chen, Xuan Zhou, Richard B. Scherl, and James Geller. Using an Intelligent Ontology for Improved Support in Rule Mining, volume 2737 of *Lecture Notes in Computer Science*, pages 320 – 329. Springer, 2003.
- Claudia Diamantini and Domenico Potena. Semantic annotation and services for KDD tools sharing and reuse. In *ICDMW '08: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 761–770, Washington, DC, USA, 2008. IEEE Computer Society.
- Robert Engels. Planning tasks for knowledge discovery in databases; performing task-oriented user-guidance, 1996.

OTHER REFERENCES OF PAPERS ON SEMANTIC DATA MINING

- Inhaunima Neves Ferraz and Ana Cristina Bicharra Garcia. Ontology in association rules pre-processing and post-processing. In IADIS European Conf. Data Mining, pages 87–91, 2008.
- Samah Jamal Fodeh and Pang-Ning Tan. Incorporating background knowledge for subjective rule evaluation. In 19th International Conference on Tools with Artificial Intelligence, pages 148–155, 2007.
- Dominique Francisci and Martine Collard. Multi-criteria evaluation of interesting dependencies according to a data mining approach. In Congress on Evolutionary Computation, pages 1568–1574. IEEE Press, 2003.
- Minos N. Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Spirit: Sequential pattern mining with regular expression constraints. In VLDB, pages 223–234, 1999.38. Jiawei Han. Mining knowledge at multiple concept levels. In CIKM, pages 19–24, 1995.
- Jiawei Han and Yongjian Fu. Discovery of multiple-level association rules from large databases. In VLDB, pages 420–431, 1995.
- Robert J. Hilderman and Howard J. Hamilton. Applying objective interestingness measures in data mining systems. In PKDD, pages 432–439, 2000.42. Robert J. Hilderman and Howard J. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In Lecture Notes in Computer Science, pages 247–259. Springer-Verlag, 2001.
- Nguyen Sinh Hoa and Nguyen Hung Son. Improving Rough Classifiers Using Concept Ontology, volume 3518/2005 of Lecture Notes in Computer Science, pages 312–322. Springer Berlin/Heidelberg, 2005.
- Szymon Jaroszewicz and Dan A. Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In KDD '04: Proceedings of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pages 178–186, Seattle, WA, USA, 2004. ACM
- Roberto J. Bayardo Jr., Rakesh Agrawal, and Dimitrios Gunopulos. Constraint- based rule mining in large, dense databases. In ICDE, pages 188–197, 1999.
- Alexandros Kalousis, Abraham Bernstein, and Melanie Hilario. Meta-learning with kernels and similarity functions for planning data mining workflows. In ICML/COLT/UAI 2008, Planing to Learn Workshop (PlanLearn), 2008.
- Nittaya Kerdpraso and Kittisak Kerdpraso. Semantic Knowledge Integration to Support Inductive Query Optimization, volume 4654/2007 of Lecture Notes in Computer Science, pages 157–169. Springer Berlin / Heidelberg, 2007.
- Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discovered association rules. In CIKM '94: Proceedings of the third international Conference on Information and Knowledge Management, pages 401–407, Gaithersburg, Maryland, United States, 1994. ACM Press.
- Evangelos Kotsifakos, Gerasimos Marketos, and Yannis Theodoridis. A Framework for Integrating Ontologies and Pattern-Bases, volume Data Mining with Ontologies: Implementations, Findings and Frameworks, pages 237–255. Information Science Reference, 2008.

OTHER REFERENCES OF PAPERS ON SEMANTIC DATA MINING

- Philippe Lenca, Patrick Meyer, Benoît Vaillant, and Stéphane Lallich. On selecting interestingness measures for association rules : user oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610– 626, 2008.
- Israel-Cesar Lerman and Jérôme Aze. A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. In *Quality Measures in Data Mining*, pages 207–236. 2007.
- Jiye Li, Nick Cercone, Serene W. H. Wong, and Lisa Jing Yan. Enhancing rule importance measure using concept hierarchy. In Philippe Lenca and Stéphane Lallich, editors, *The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), workshop on Quality issues, measures of interestingness and evaluation of data mining models (QIMIE)*, 2009.
- Bing Liu, Wynne Hsu, and Shu Chen. Using general impressions to analyze discovered classification rules. In *Knowledge Discovery and Data Mining*, pages 31–36, 1997.
- Bing Liu, Wynne Hsu, Lai-Fun Mun, and Hing-Yan Lee. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):817–832, 1999.
- Claudia Marinica and Fabrice Guillet. Knowledge-based interactive postmining of association rules using ontologies. *IEEE Trans. Knowl. Data Eng.*, 22(6):784–797, 2010.
- Mei, Qiaozhu and Xin, Dong and Cheng, Hong and Han, Jiawei and Zhai, Chengxian, Semantic annotation of frequent patterns}, *ACM Trans. Knowl. Discov. Data*, December 2007, Vol 1, issue 3
- Srujana Merugu, Saharon Rosset, and Claudia Perlich. A new multi-view regression approach with an application to customer wallet estimation. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 656–661, Philadelphia, PA, USA, 2006. ACM.
- Patrick Meulstee and Mykola Pechenizkiy. Food sales prediction: "If Only It Knew What We Know". *Data Mining Workshops, 2008. International Conference on Data Mining ICDM'08. IEEE International Conference on*, 0:134–143, 2008.
- Ryszard S. Michalski and Kenneth A. Kaufman. Building knowledge scouts using KGL metalanguage. *Fundamenta Informaticae*, 40:433–447, 2000.
- Katharina Morik and Martin Scholz. The Mining Mart approach to knowledge discovery in databases. In In Ning Zhong and Jiming Liu, editors, *Intelligent Technologies for Information Analysis*, pages 47–65. Springer, 2003.6
- Miho Ohsaki, Hidenao Abe, Shusaku Tsumoto, Hideto Yokoi, and Takahira Yamaguchi. Proposal of medical KDD support user interface utilizing rule interestingness measures. In *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 759–764, Washington, DC, USA, 2006. IEEE Computer Society.
- Pance Panov, Sašo Džeroski, and Larisa Soldatova. OntoDM: An ontology of data mining. In *ICDMW '08: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 752–760, Washington, DC, USA, 2008. IEEE Computer Society.

OTHER REFERENCES OF PAPERS ON SEMANTIC DATA MINING

- Joseph Phillips and Bruce G. Buchanan. Ontology-guided knowledge discovery in databases. In K-CAP '01: Proceedings of the 1st international conference on Knowledge Capture, pages 123–130, Victoria, British Columbia, Canada, 2001. ACM.
- Gregory Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In Knowledge Discovery in Databases, pages 229–248. AAAI/MIT Press, 1991.
- Giuseppe Psaila and Pier Luca Lanzi. Hierarchy-based mining of association rules in data warehouses. In SAC '00: Proceedings of the 2000 ACM symposium on Applied computing, pages 307–312, New York, NY, USA, 2000. ACM.
- Stefano Rizzi, Elisa Bertino, Barbara Catania, Matteo Golfarelli, Maria Halkidi, Manolis Terrovitis, Panos Vassiliadis, Michalis Vazirgiannis, and Euripides Vrachnos. Towards a logical model for patterns. In Internal Conference on Conceptual Modeling - ER 2003, pages 77–90, 2003.
- Saharon Rosset, Claudia Perlich, Bianca Zadrozny, Srujana Merugu, Sholom M. Weiss, and R. Lawrence. Wallet estimation models. In Proceedings of the International Workshop on CRM: Data Mining Meets Marketing, 2005.
- Giovanni Maria Sacco. DT-miner: Data mining for the people. In DEXA '08: Proceedings of the 2008 19th International Conference on Database and Expert Systems Application, pages 387–391, Washington, DC, USA, 2008. IEEE Computer Society.
- A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, 8:970–974, 1996.74. Abraham Silberschatz and Alexander Tuzhilin. On subjective measures of interest- ingness in knowledge discovery. In Knowledge Discovery and Data Mining, pages 275–281, 1995.
- P.Smyth and R.M.Goodman. An information theoretic approach to rule induction from databases. IEEE Trans. on Knowl. and Data Eng., 4(4):301–316, 1992.
- Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. In VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pages 407–419, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
- Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal. Mining association rules with item constraints. In KDD, pages 67–73. AAAI Press, 1997.
- Vojtech Svatek, Jan Rauch, and Martin Ralbovsky . Ontology-Enhanced Association Mining, volume 4289/2006 of Lecture Notes in Computer Science, pages 163–179. Springer Berlin / Heidelberg, 2006.
- Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 32–41, New York, NY, USA, 2002. ACM.

OTHER REFERENCES OF PAPERS ON SEMANTIC DATA MINING

- Rudiger Wirth, Colin Shearer, Udo Grimmer, Thomas Reinartz, Jorg Schlosser, Christoph Breitner, Robert Engels, and Guido Lindner. Towards process-oriented tool support for knowledge discovery in databases. Lecture Notes in Computer Science, Principles of Data Mining and Knowledge Discovery, 1263:243–253, 1997.
- Dong Xin, Xuehua Shen, Qiaozhu Mei, and Jiawei Han. Discovering interesting patterns through user's interactive feedback. In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pages 773–778, Philadelphia, PA, USA, 2006. ACM.
- M Z'akova, P Kremen, F Z'elezny, and N Lavra'c. Using ontological reasoning and planning for data mining workflow composition. In ECML PKDD '08: Proceedings of the European Conference on Machine Learning and Knowledge Discovery Workshop, 2008.
- Sai Zeng, Ioana Boier-Martin, Prem Melville, Conrad Murphy, and Christian A. Lang. Predictive modeling for collections of accounts receivable. In DDDM '07: Proceedings of the 2007 international workshop on Domain Driven Data Mining, pages 43–48, San Jose, California, 2007. ACM.
- Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Chengqi Zhang, and Hans Bohlscheid. Combined Pattern Mining: From Learned Rules to Actionable Knowledge, volume 5360/2008 of Lecture Notes in Computer Science, pages 393–403. Springer Berlin / Heidelberg, 2008.
- Ning Zhong, Chunnian Liu, and Setsuo Ohsuga. A way of increasing both autonomy and versatility of a KDD system. In ISMIS '97: Proceedings of the 10th International Symposium on Foundations of Intelligent Systems, pages 94–105, London, UK, 1997. Springer-Verlag.
- Zonglin Zhou, Huan Liu, Stan Z. Li, and Chin-Seng Chua. Rule mining with prior knowledge - a belief networks approach. *Intell. Data Anal.*, 5(2):95–110, 2001.