Relational and Semantic Data Mining

Nada Lavrač
Anže Vavpetič

Jožef Stefan Institute
Ljubljana, Slovenia

LPNMR-2015, Lexington, September 2015

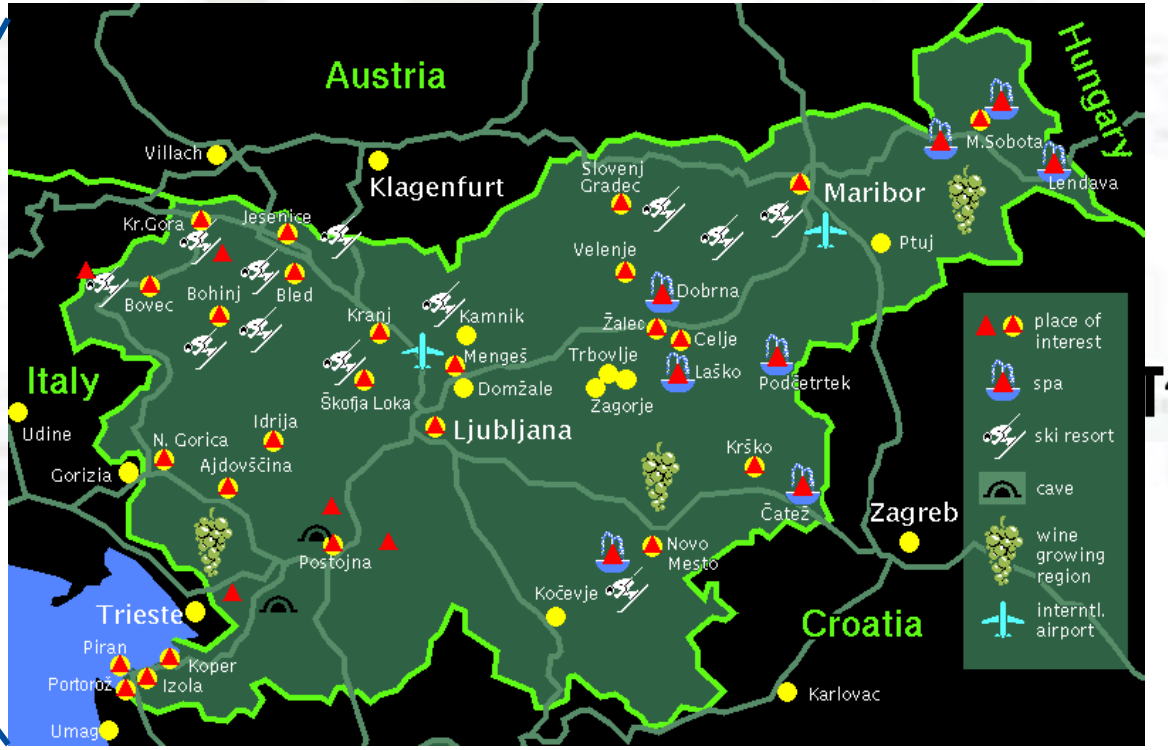
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Jožef Stefan Institute, Ljubljana, Slovenia

- **Jožef Stefan Institute**
  - leading national research organization in natural sciences and technology (~900 staff, researchers and students, ~500 PhDs)

Jožef Stefan Institute, Ljubljana, Slovenia

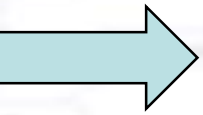
- **Jožef Stefan Institute**
  - leading national research organization in natural sciences and technology (~900 staff, researchers and students, ~500 PhDs)



Department of Knowledge Technologies

~ 40 researchers (~ 10 PhD students) + NAO robot

Talk outline



- Intro. to Machine Learning and Data Mining
- Inductive Logic Programming (ILP) and Relational Data Mining (RDM)
- Propositionalization approach to RDM
- Semantic data mining
- Summary and future work

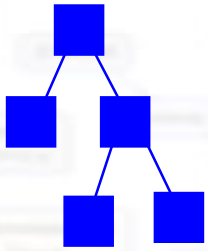
Machine Learning and Data Mining

data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

knowledge discovery from data

Machine Learning
Data Mining



model, patterns, …

Given: class labeled data
Find: classification model or
set of interesting patterns in the data

Example: Contact lens data

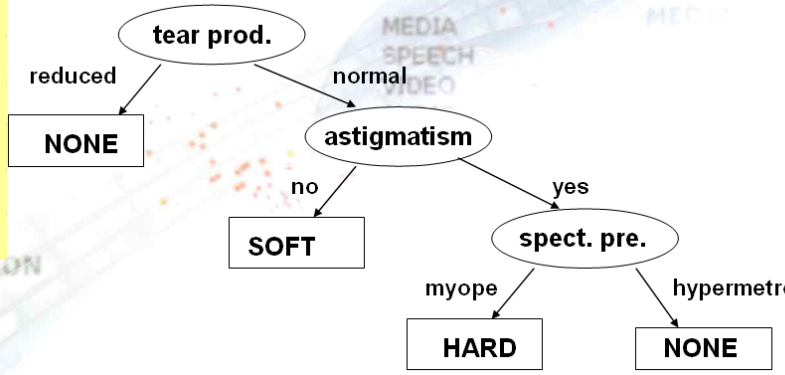
DATA

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

Learning models from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

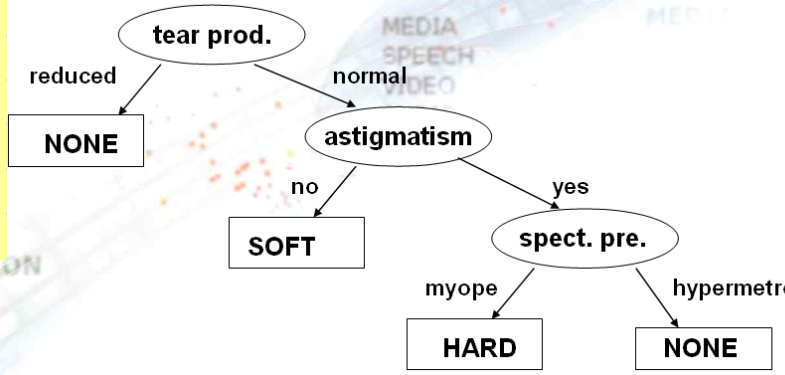
Data Mining



Learning models from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

Data Mining



lenses=NONE ← tear production=red
lenses=NONE ← tear production=normal ∧ astigmatism=yes ∧ spect. pre.=hypermetrope
lenses=SOFT ← tear production=normal ∧ astigmatism=no
lenses=HARD ← tear production=normal ∧ astigmatism=yes ∧ spect. pre.=myope
lenses=NONE ←

Finding patterns in Contact lens data

DATA

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NONE

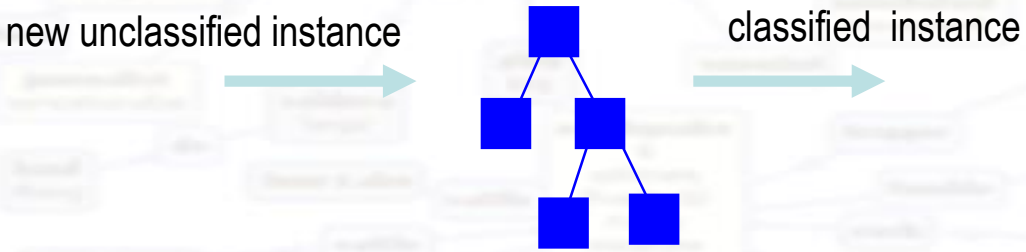
PATTERNs: rules describing subgroups of instances

Lenses = NONE ← Tear prod. = reduced

Why learn symbolic models and patterns

Use learned models

to classify and explain classifications of new instances



Use discovered patterns

for data exploration

Task reformulation: Binary class values

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NO
O2	23	myope	no	normal	YES
O3	22	myope	yes	reduced	NO
O4	27	myope	yes	normal	YES
O5	19	hypermetrope	no	reduced	NO
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	YES
O15	43	hypermetrope	yes	reduced	NO
O16	39	hypermetrope	yes	normal	NO
O17	54	myope	no	reduced	NO
O18	62	myope	no	normal	NO
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	NO

Binary classes (positive vs. negative examples of Target class)

- simplifies single concept learning
- is used in “one class vs. all” multi-class learning methods

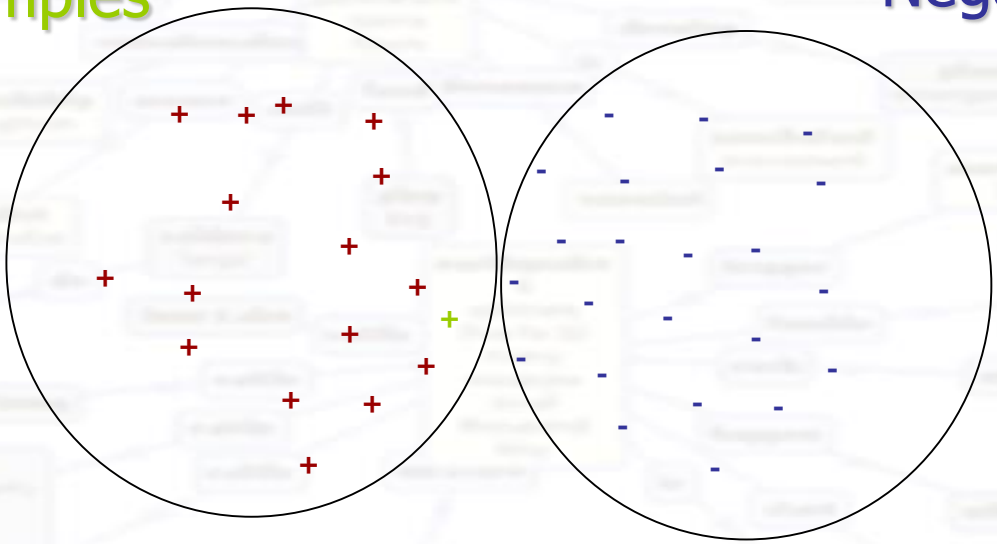
Classification rules learned from contact lens data with binary classes

lenses=YES ← tear production=normal ∧
astigmatism=no
lenses=YES ← tear production=normal ∧
astigmatism=yes ∧
spect. pre.=myope
lenses=NO ← tear production=reduced
lenses=NO ← tear production=normal ∧
astigmatism=yes ∧
spect. pre.=hypermetrope
lenses=NO ←

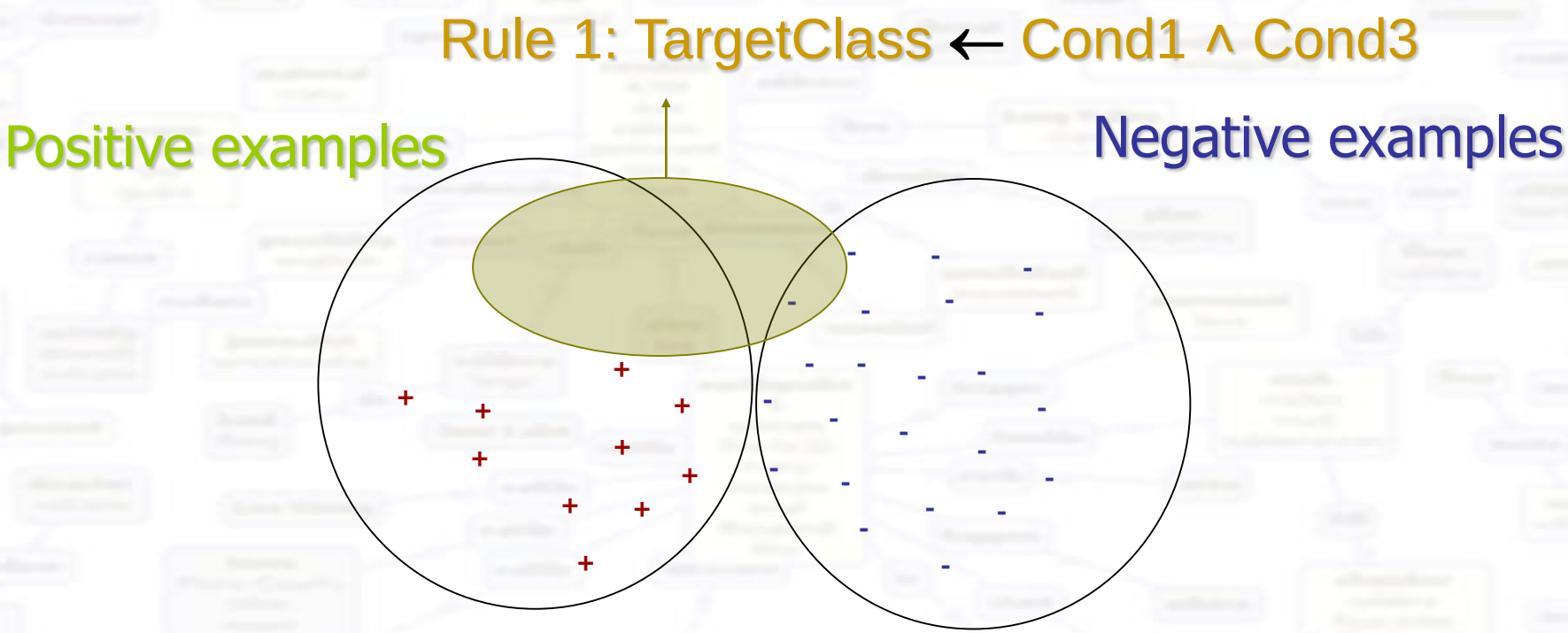
Covering algorithm used in rule learning

Positive examples

Negative examples



Covering algorithm used in rule learning

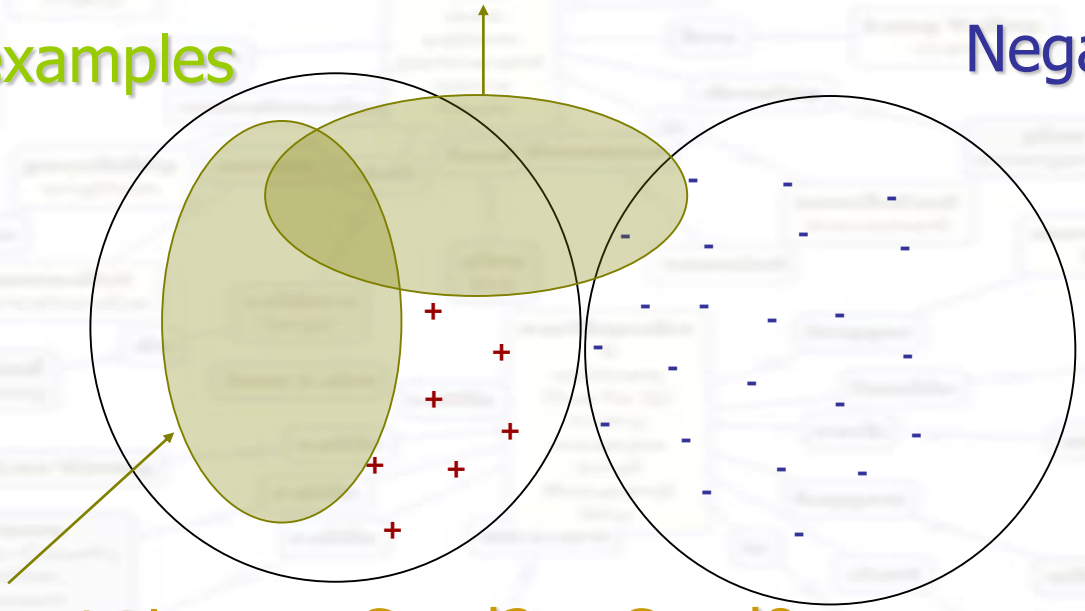


Covering algorithm used in rule learning

Rule 1: TargetClass ← Cond1 ∧ Cond3

Positive examples

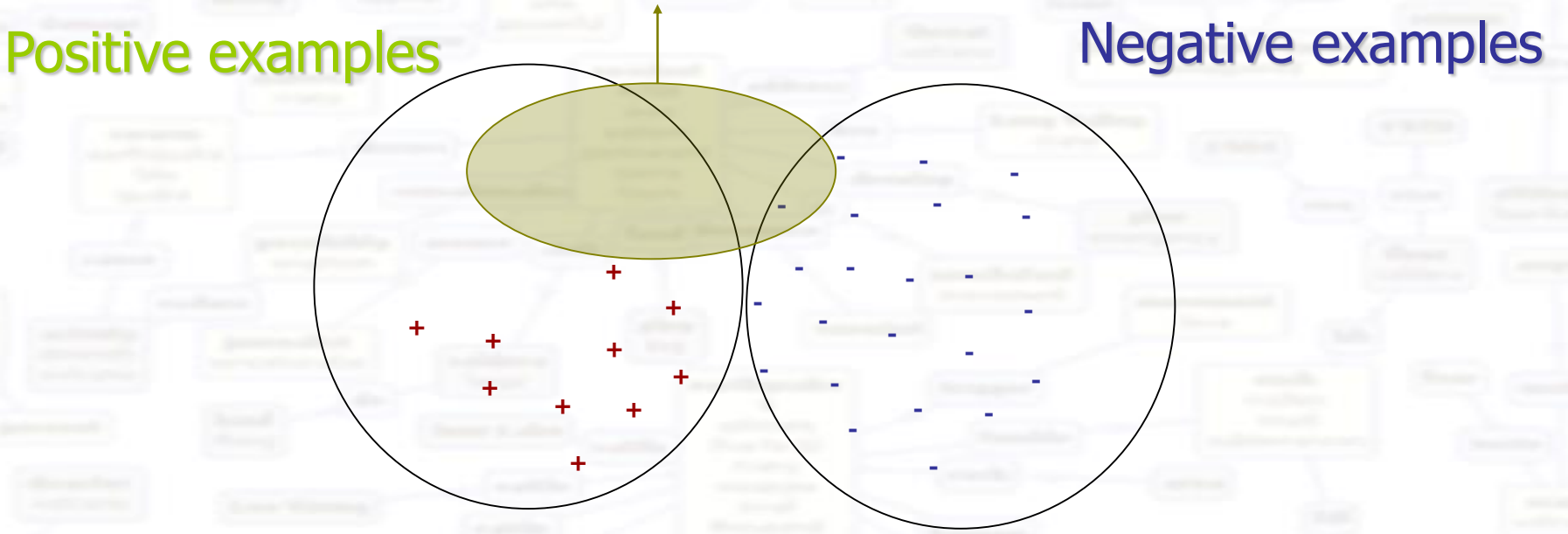
Negative examples



Rule 2: TargetClass ← Cond2 ∧ Cond8

Heuristics used in classification rule learning

Rule: Class ← Conditions



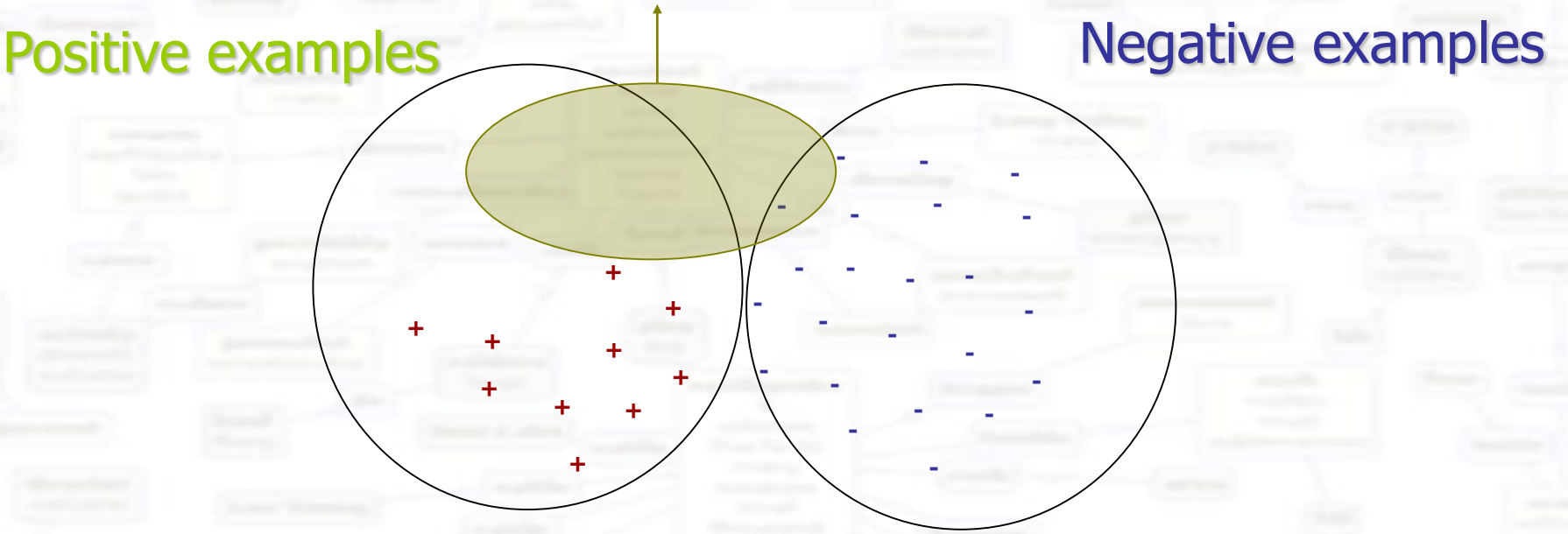
Evaluating quality of classification rules: e.g., using accuracy or precision

$$q(\text{Rule}) = p(\text{Class}|\text{Conds})$$

Precision (rel. freq. of correctly covered pos. examples)

Heuristics used in subgroup discovery

Rule: Class ← Conditions



Evaluating quality of pattern describing rules in subgroup discovery:

e.g., trading off coverage and precision gain: Coverage x (Precision – Default)

$$q(Rule) = p(Conds) \times (p(Class|Conds) - p(Class))$$

Coverage (rel. freq. of covered examples)

Precision (rel. freq. of correctly covered pos. examples)

Subgroup discovery example: High CHD Risk Group Detection

Input: Patient records described by anamnestic, laboratory and ECG attributes

Task: Find and characterize population subgroups with high CHD risk (large, distributionally unusual subgroups)

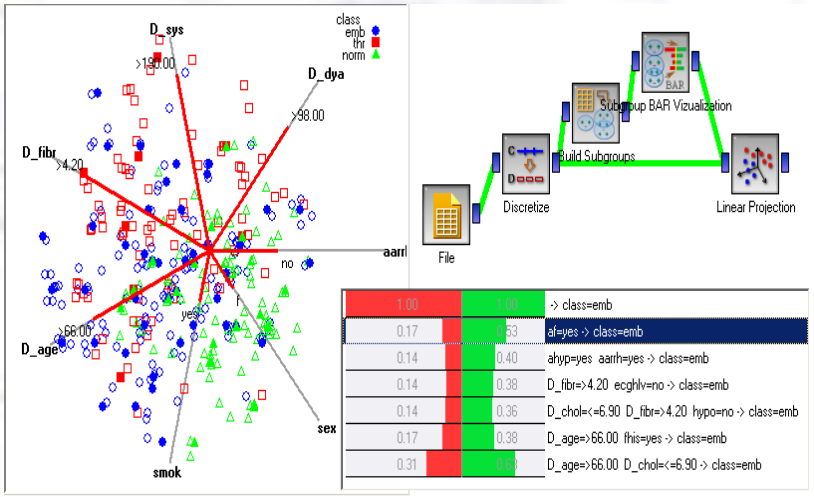
From **best induced descriptions**, five were selected by the expert as **most actionable** for CHD risk screening (by GPs):

high-CHD-risk ← male ∧ pos. fam. history ∧ age > 46
high-CHD-risk ← female ∧ bodymassIndex > 25 ∧ age > 63
high-CHD-risk ← ...
high-CHD-risk ← ...
high-CHD-risk ← ...

(Gamberger & Lavrač, JAIR 2002)

SD algorithms in the Orange DM Platform

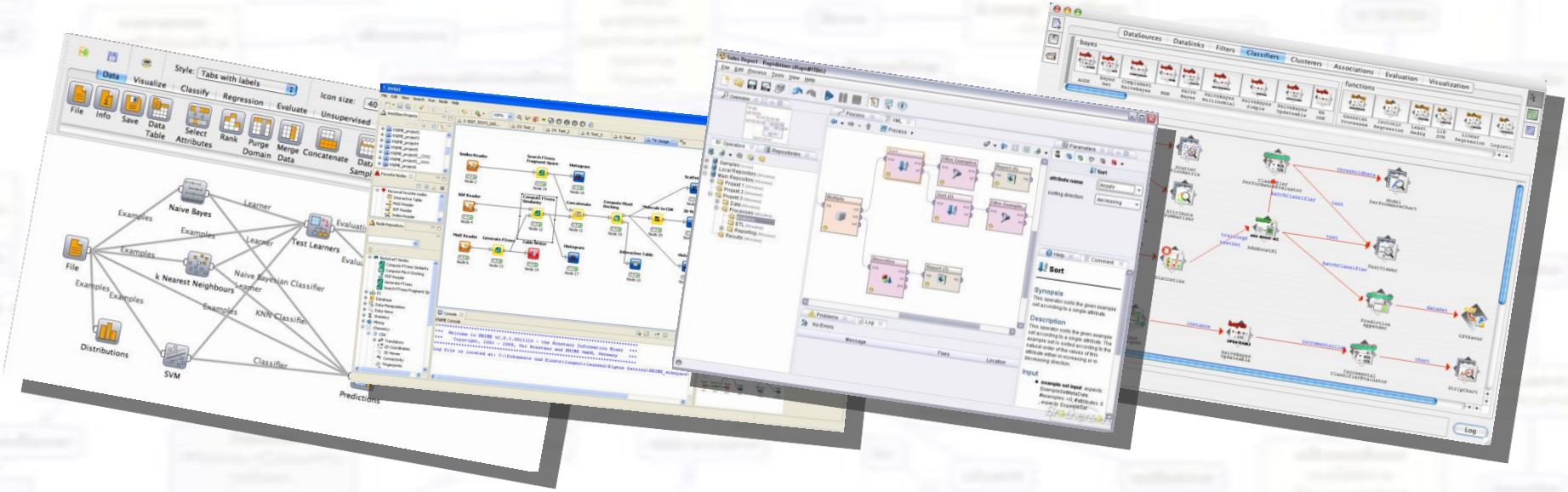
- **Orange** data mining toolkit
  - classification and subgroup discovery algorithms
  - data mining workflows
  - visualization
  - developed at FRI, Ljubljana



- **SD Algorithms in Orange**
  - SD (Gamberger & Lavrač, JAIR 2002
  - APRIORI-SD (Kavšek & Lavrač, AAI 2006
  - CN2-SD (Lavrač et al., JMLR 2004): Adapting CN2 classification rule learner to Subgroup Discovery

Other Data Mining Platforms

WEKA, KNIME, RapidMiner, Orange4WS, …



– include numerous data mining algorithms
– enable data and model visualization
– enable complex **workflow** construction

Talk outline

- Intro. to Machine Learning and Data Mining
- Inductive Logic Programming (ILP) and Relational Data Mining (RDM)
- Propositionalization approach to RDM
- Semantic data mining
- Summary and future work

ILP and Relational Data Mining

customer

ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

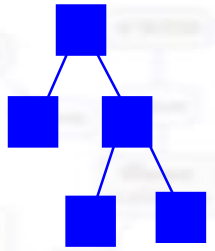
store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

knowledge discovery from data

Relational Data Mining



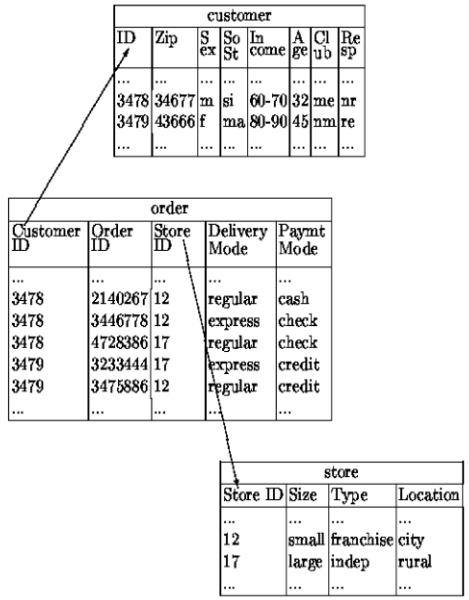
model, patterns, …

Given: a relational database, a set of tables, sets of logical facts, a graph, …

Find: a classification model, a set of patterns

ILP and Relational Data Mining

- **ILP, relational learning, relational data mining**
  - Learning from complex relational databases



customer

ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

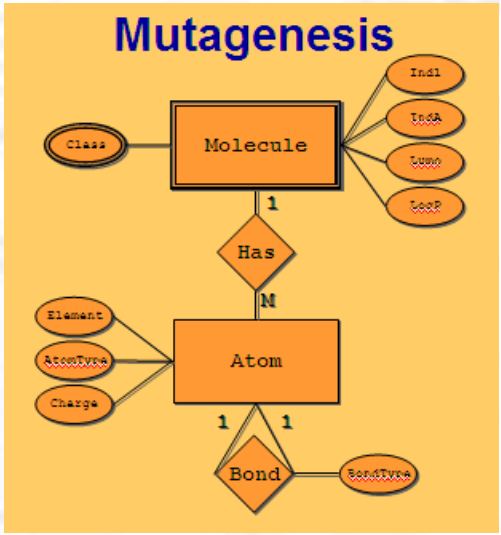
store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

ILP and Relational Data Mining

- **ILP, relational learning, relational data mining**
  - Learning from complex relational databases
  - Learning from complex structured data, e.g. molecules and their biochemical properties



customer

ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

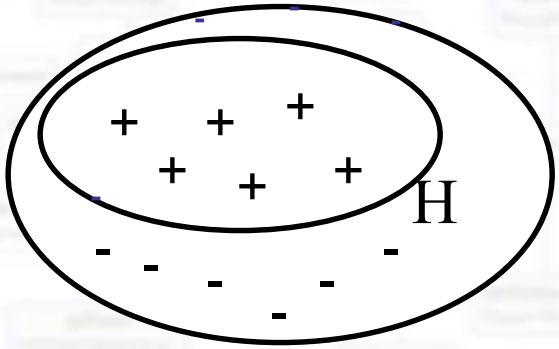
store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

ILP for Logic Programming

- **Given:**
  - A set of observations (ground facts)
    - positive examples E^+
    - negative examples E^-
  - background knowledge B (definite clauses)
  - hypothesis language (definite clauses) L_H
  - covers relation (logical entailment)



- **Find:**
  A hypothesis (a theory) $H \in L_H$, such that (given B) H covers all positive and no negative examples

- In logic, **find** H such that
  - $\forall e \in E^+ : B \cup H \models e$ (H is complete)
  - $\forall e \in E^- : B \cup H \not\models e$ (H is consistent)

Inductive Logic Programming Example

E^+ = {sort([2,1,3],[1,2,3])}
E^- = {sort([2,1],[1]),sort([3,1,2],[2,1,3])}

B : definitions of `permutation/2` and `sorted/1`

- **Predictive ILP: Learning a theory H**

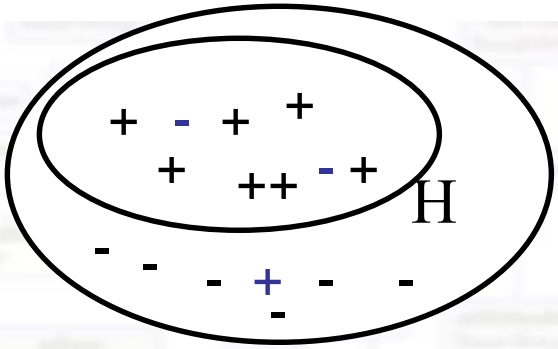
```
sort(X,Y) ← permutation(X,Y), sorted(Y).
```

- **Descriptive ILP: Finding individual patterns**

```
sorted(Y) ← sort(X,Y).
permutation(X,Y) ← sort(X,Y).
sorted(X) ← sort(X,X).
```

ILP for relational learning

- **Given:**
  - A set of observations (ground facts)
    - positive examples E^+
    - negative examples E^-
  - background knowledge B (definite clauses)
  - hypothesis language (definite clauses) L_H
  - covers relation (theta-subsumption)
  - quality criterion, e.g., predictive accuracy A(H)



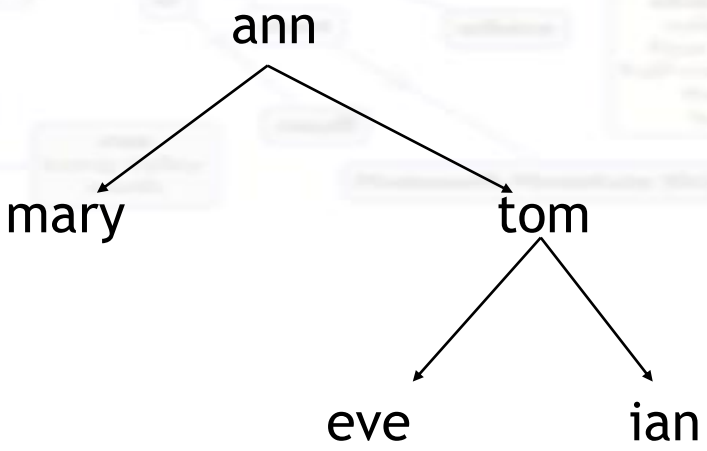
- **Find:**
  A hypothesis (a set of clauses) $H \in L_H$, such that (given B) H is optimal w.r.t. given quality criterion

  (relaxing the request for finding a hypothesis H $\in L_H$, such that (given B) H covers all positive and no negative examples)

Relational Learning Example

E^+ = {daughter(mary,ann),daughter(eve,tom)}
E^- = {daughter(tom,ann),daughter(eve,ann)}

B = **Facts:** {mother(ann,mary), mother(ann,tom),
father(tom,eve), father(tom,ian),
female(ann), female(mary),
female(eve), male(pat),male(tom)}
Rules: {parent(X,Y) ← mother(X,Y),
parent(X,Y) ← father(X,Y)}



Relational Learning Example

- **Learning models:** Induce a theory, e.g., as a single definite clause

```
daughter(X,Y) ← female(X), parent(Y,X).
```

or as a set of definite clauses

```
daughter(X,Y) ← female(X), mother(Y,X).
daughter(X,Y) ← female(X), father(Y,X).
```

- **Finding patterns:** Induce individual rules (individual general clauses)

```
← daughter(X,Y), mother(X,Y).
female(X)← daughter(X,Y).
mother(X,Y); father(X,Y) ← parent(X,Y).
```

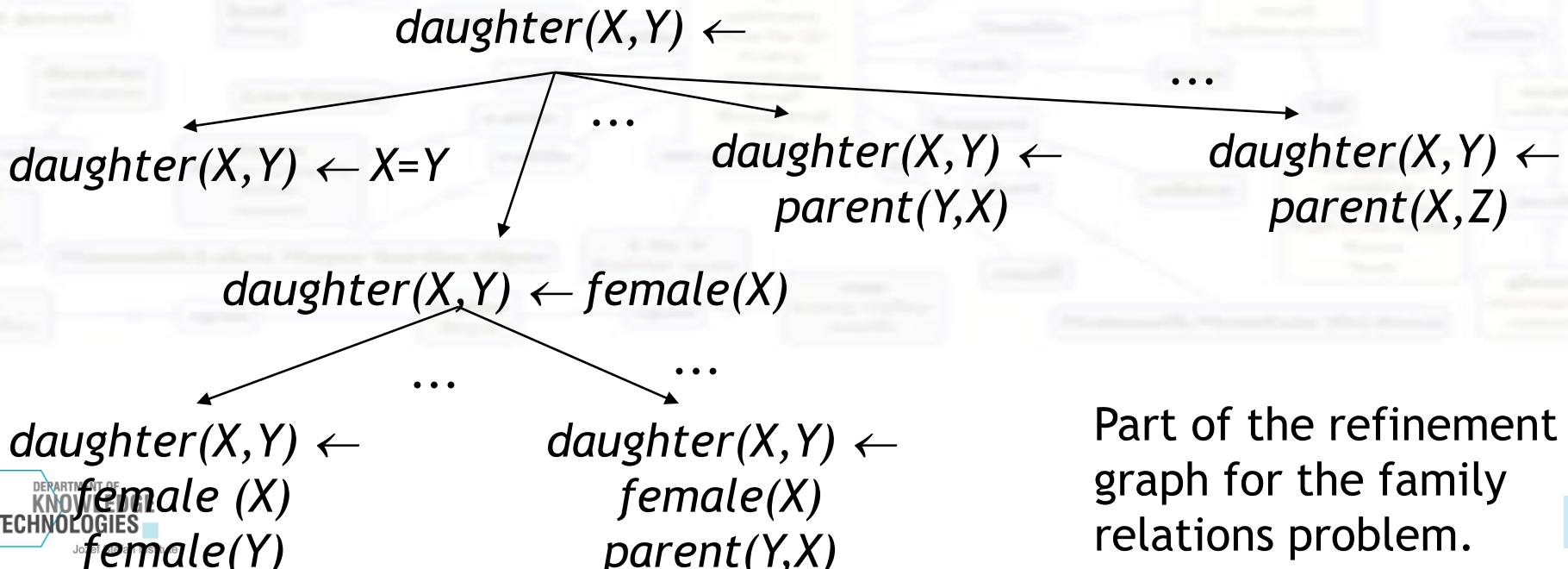
ILP as search of program clauses

- **ILP systems structure the hypothesis space based on syntactic generality relation (θ-subsumption)**
  - Clause c_1 θ-subsumes c_2 ($c_1 \geq_\theta c_2$) iff $\exists\theta : c_1\theta \subseteq c_2$
  - Hypothesis $H_1 \geq \theta H_2$ iff $\forall c_2 \in H_2$ exists $c_1 \in H_1$ such that $c_1 \geq \theta c_2$
- **Example**

  c1 = daughter(X,Y) ← parent(Y,X)
  c2 = daughter(mary,ann) ← female(mary), parent(ann,mary),
  c1 θ-subsumes c_2 under θ = {X/mary,Y/ann}
- **Learning strategies**
  - Top-down search of refinement graphs (FOIL)
  - Bottom-up search (building least general generalizations, inverting resolution (CIGOL), inverting entailment (PROGOL))
  - Mixed strategy (Aleph)

Generality ordering of clauses

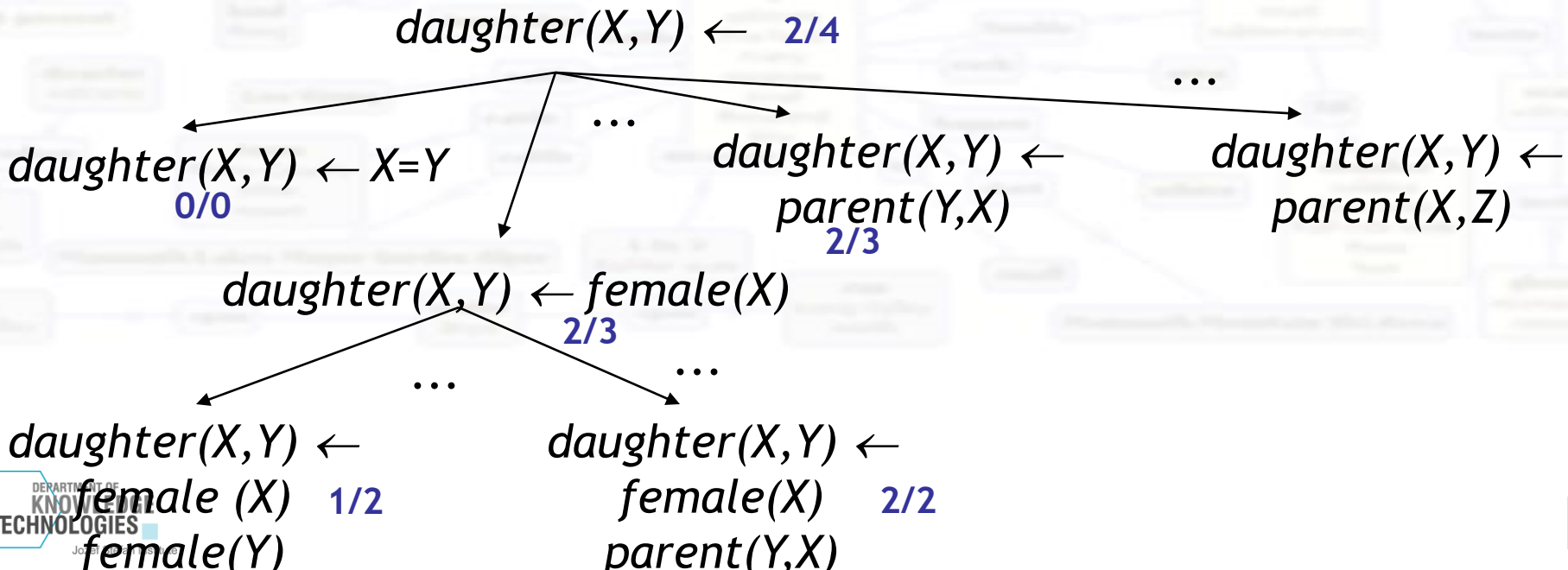
Training examples		Background knowledge	
daughter(mary,ann).	⊕	parent(ann,mary).	female(ann).
daughter(eve,tom).	⊕	parent(ann,tom).	female(mary).
daughter(tom,ann).	⊖	parent(tom,eve).	female(eve).
daughter(eve,ann).	⊖	parent(tom,ian).	



Part of the refinement graph for the family relations problem.

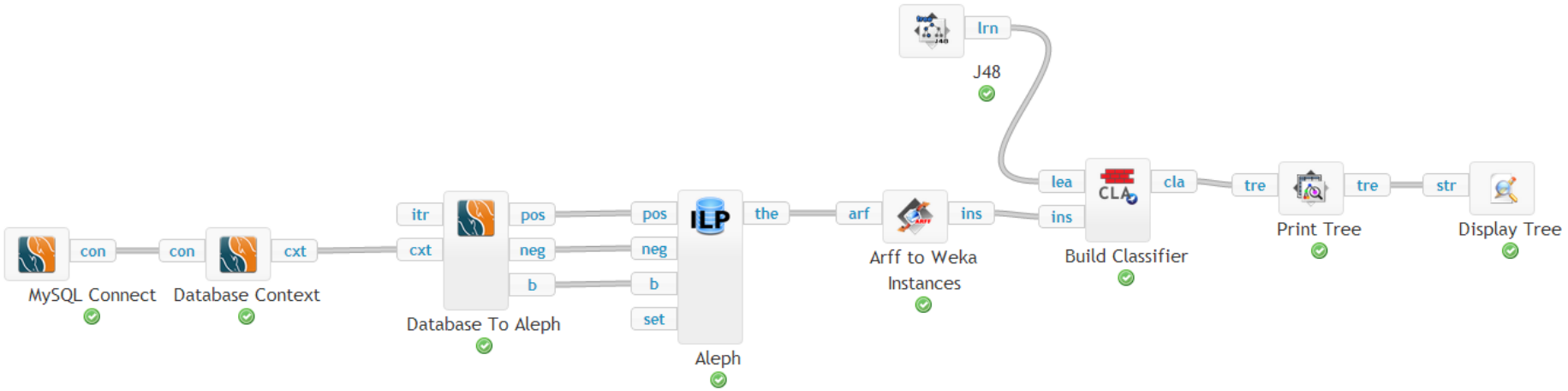
Top-down search of refinement graphs

Training examples		Background knowledge	
daughter(mary,ann).	⊕	parent(ann,mary).	female(ann.).
daughter(eve,tom).	⊕	parent(ann,tom).	female(mary).
daughter(tom,ann).	⊖	parent(tom,eve).	female(eve).
daughter(eve,ann).	⊖	parent(tom,ian).	



Selected ILP algorithms are available online in the ClowdFlows platform

- **Example: ILP system Aleph in ClowdFlows available at** http://clowdflows.org/workflow/2224/



Talk outline

- Intro. to Machine Learning and Data Mining
- Inductive Logic Programming (ILP) and Relational Data Mining (RDM)
- Propositionalization approach to RDM
- Semantic data mining
- Summary and future work

Relational Data Mining through Propositionalization

customer

ID	Zip	Sex	So St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Step 1

Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Relational Data Mining through Propositionalization

customer

ID	Zip	Sex	So St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

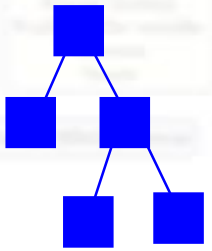
Step 1

Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Step 2

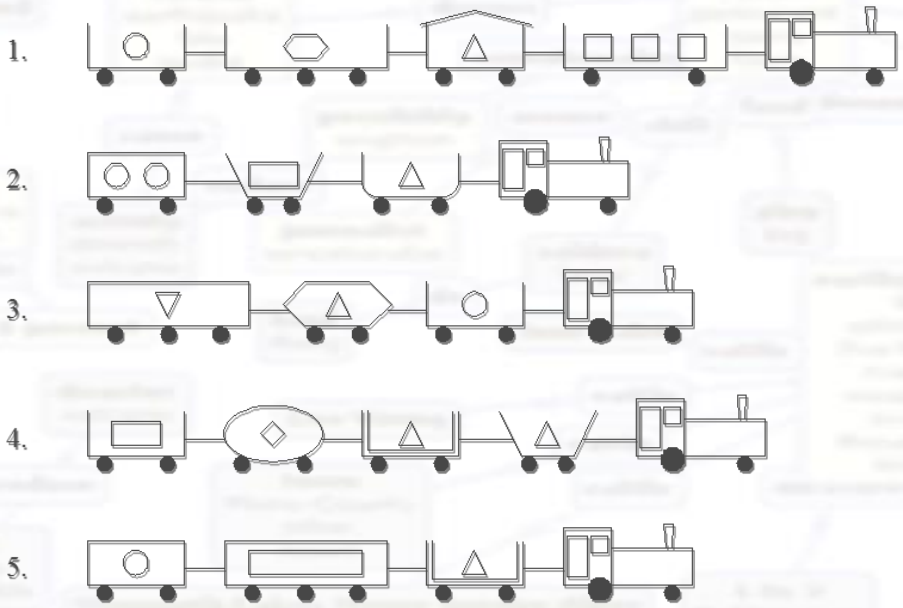
Data Mining



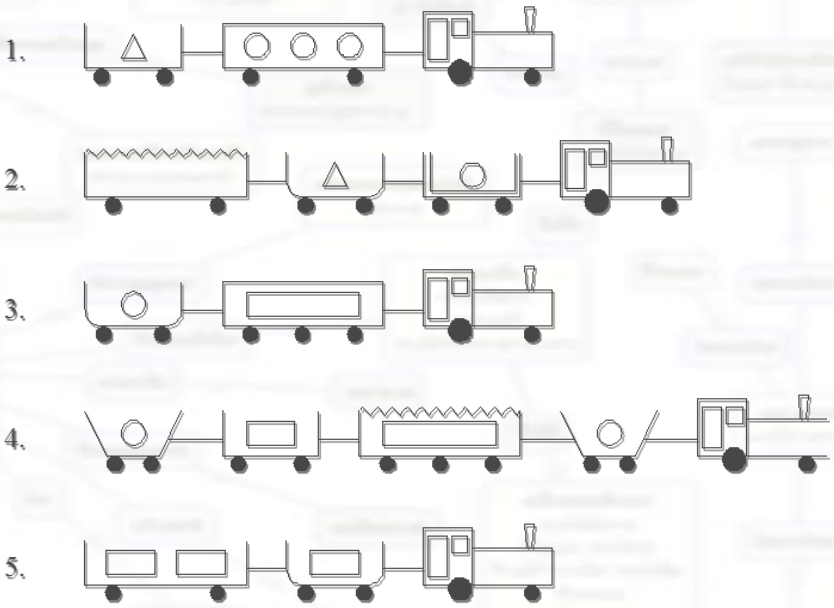
model, patterns, …

Sample ILP problem: East-West trains

1. TRAINS GOING EAST



2. TRAINS GOING WEST



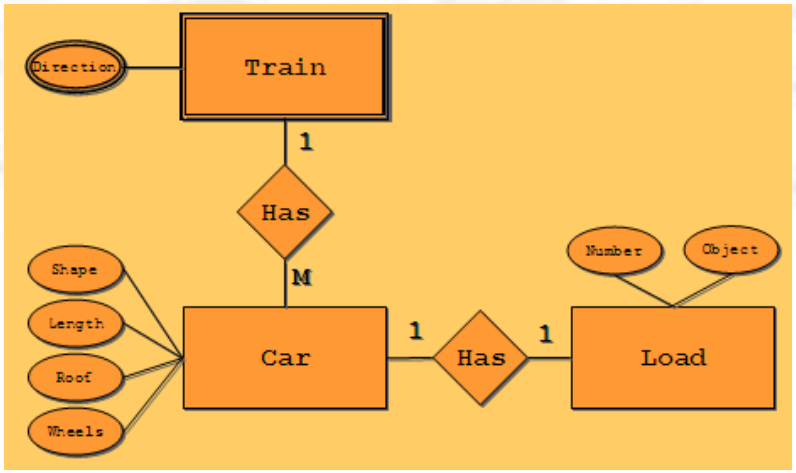
Relational representation of East-West trains

TRAIN_TABLE

TRAIN	EASTBOUND
t1	TRUE
t2	TRUE
…	…
t6	FALSE
…	…

LOAD	CAR	OBJECT	NUMBER
l1	c1	circle	1
l2	c2	hexagon	1
l3	c3	triangle	1
l4	c4	rectangle	3
…	…	…	

CAR	TRAIN	SHAPE	LENGTH	ROOF	WHEELS
c1	t1	rectangle	short	none	2
c2	t1	rectangle	long	none	3
c3	t1	rectangle	short	peaked	2
c4	t1	rectangle	long	none	2
…	…	…			…



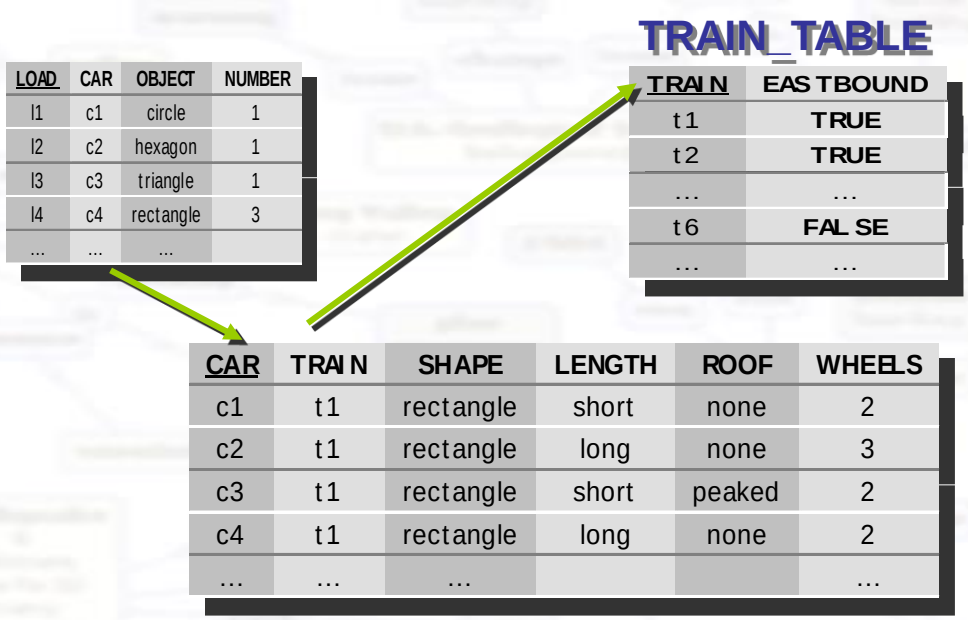
Propositionalization approach

Propositionalization through first-order feature construction (1BC, RSD, …):

f1(T):-hasCar(T,C),clength(C,short).

f2(T):-hasCar(T,C), hasLoad(C,L),
loadShape(L,circle)

f3(T) :- ….



Propositionalization approach

Standard propositionalization through first-order feature construction (1BC, RSD, …):

f1(T):-hasCar(T,C),clength(C,short).

f2(T):-hasCar(T,C), hasLoad(C,L),
loadShape(L,circle)

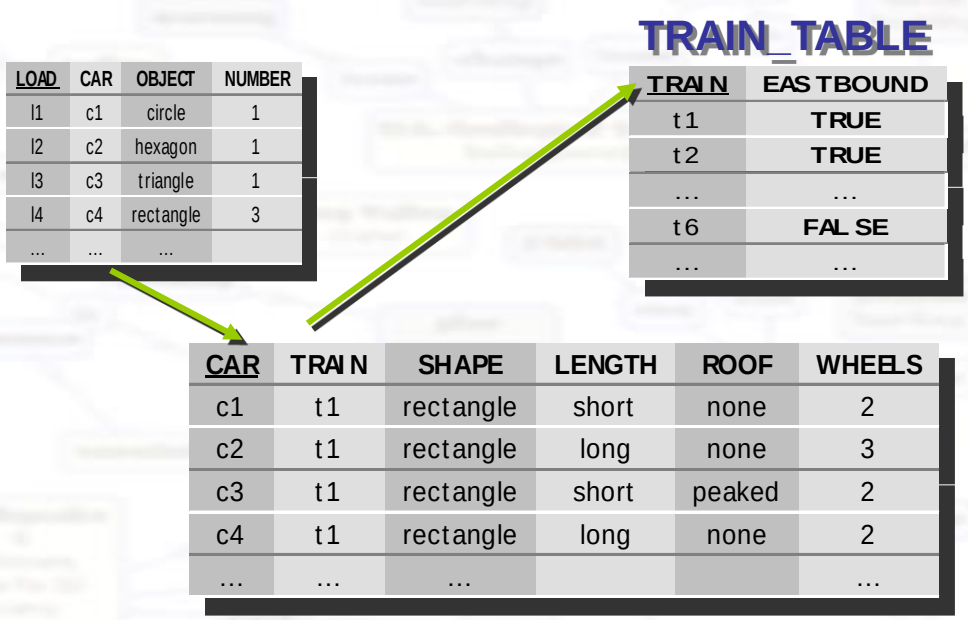
f3(T) :- ….

Propositional learning:

t(T) ← f1(T), f4(T)

Relational interpretation:

eastbound(T) ←
hasShortCar(T),hasClosedCar(T).



LOAD

LOAD	CAR	OBJECT	NUMBER
l1	c1	circle	1
l2	c2	hexagon	1
l3	c3	triangle	1
l4	c4	rectangle	3
…	…	…	

TRAIN_TABLE

TRAIN	EASTBOUND
t1	TRUE
t2	TRUE
…	…
t6	FALSE
…	…

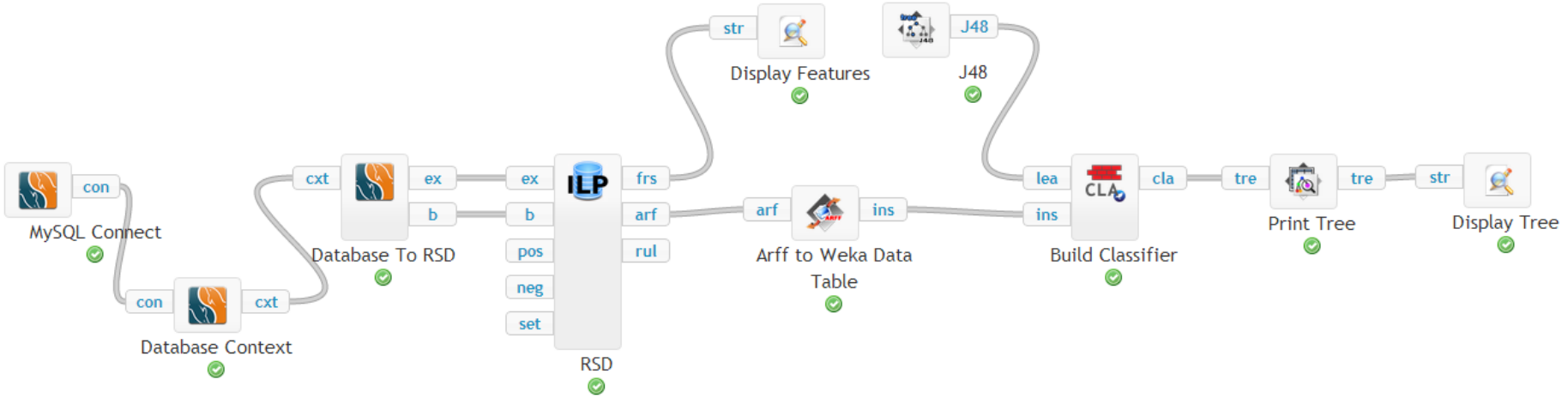
CAR	TRAIN	SHAPE	LENGTH	ROOF	WHEELS
c1	t1	rectangle	short	none	2
c2	t1	rectangle	long	none	3
c3	t1	rectangle	short	peaked	2
c4	t1	rectangle	long	none	2
…	…	…			…

PROPOSITIONALIZED TRAIN_TABLE

train(T)	f1(T)	f2(T)	f3(T)	f4(T)	f5(T)
t1	t	t	f	t	t
t2	t	t	t	t	t
t3	f	f	t	f	f
t4	t	f	t	f	f
…	…	…			…

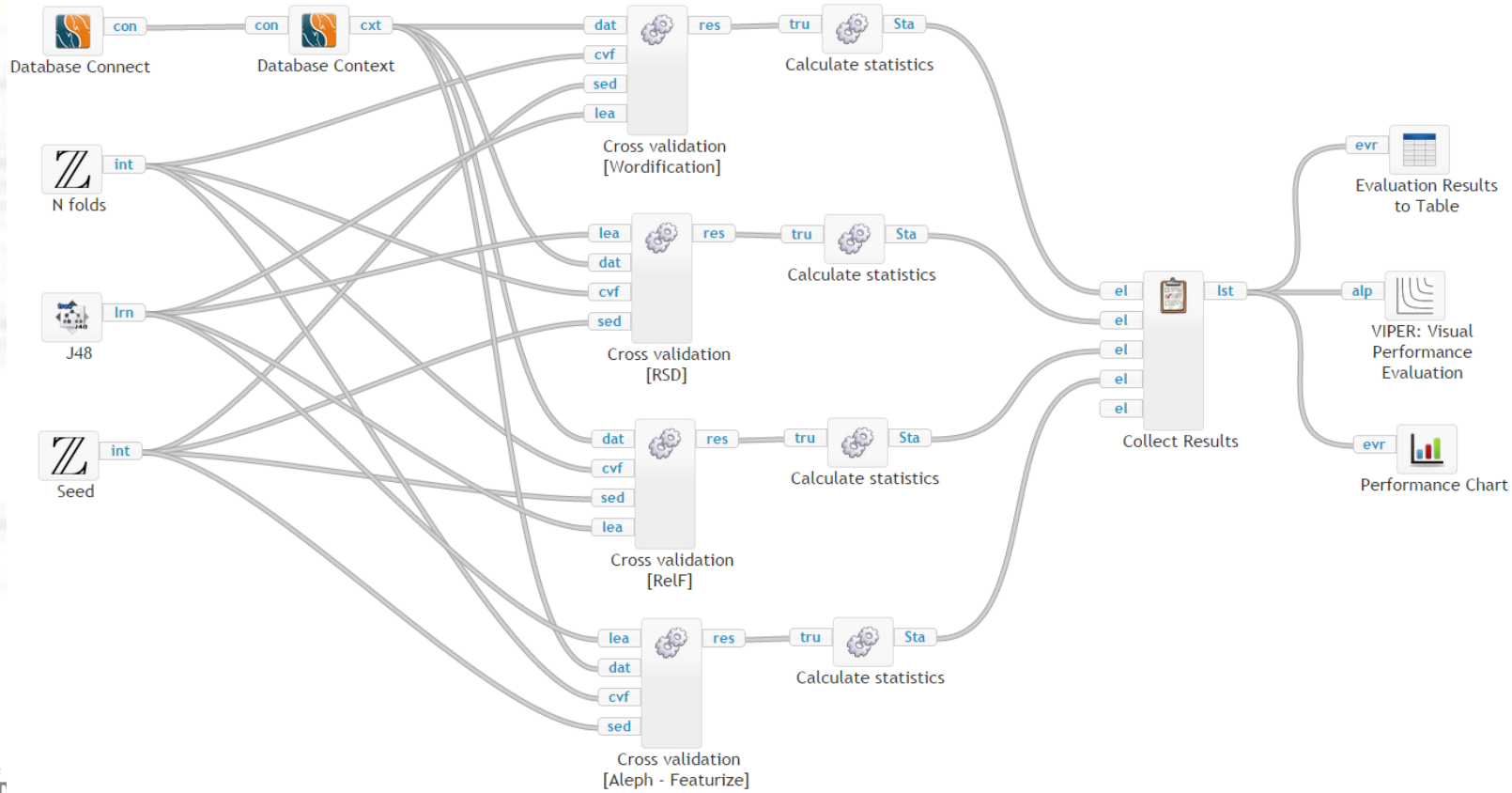
Propositionalization algorithms are available online in the ClowdFlows platform

- **ClowdFlows** - browsed-based DM platform for data mining in the cloud and workflow sharing on the web (Kranjc et al. 2012)
- **Example workflow**: Propositionalization with RSD available in ClowdFlows at http://clowdflows.org/workflow/611/



Propositionalization in ClowdFlows

- **Example workflow:** Comparison of propositionalization algorithms (RSD, RelF, …), available in ClowdFlows at http://clowdflows.org/workflow/4018/

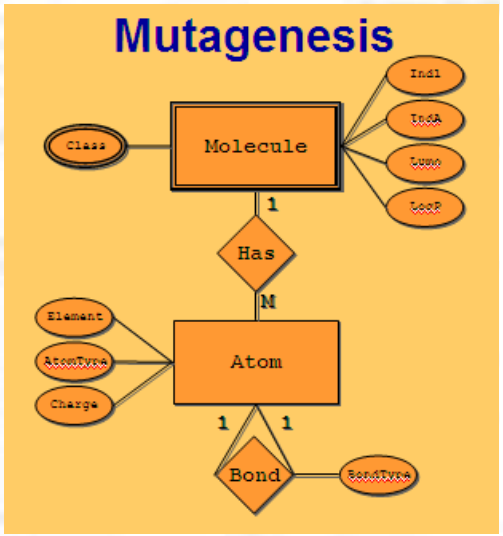


Talk outline

- Intro, to Machine Learning and Data Mining
- Inductive Logic Programming (ILP) and Relational Data Mining (RDM)
- Propositionalization approach to RDM
- Semantic data mining
- Summary and future work

Relational and semantic data mining

- **ILP, relational learning, relational data mining**
  - Learning from complex relational databases
  - Learning from complex structured data, e.g. molecules and their biochemical properties
  - Learning by using domain knowledge in the form of ontologies = **semantic data mining**



customer

ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

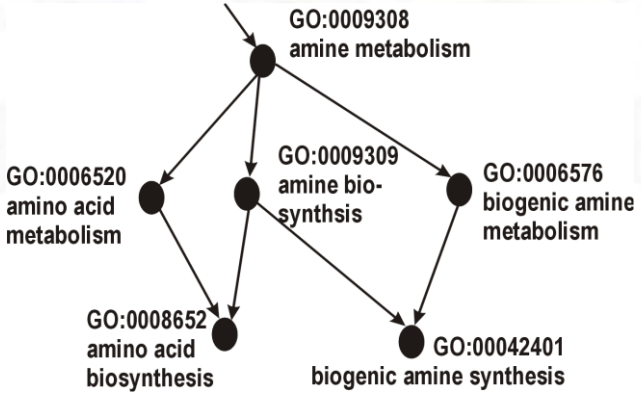
order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

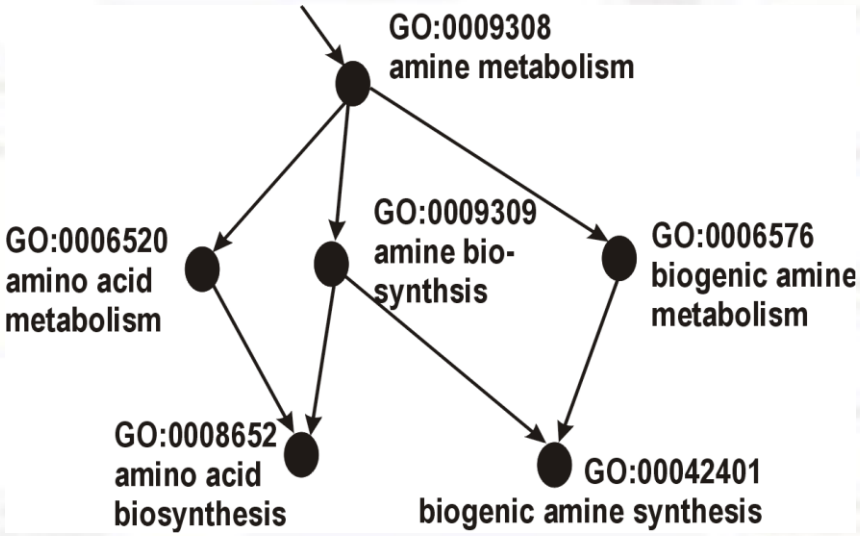
Relational representation of customers, orders and stores.



Using domain ontologies

Using domain ontologies as background knowledge

- E.g., the Gene Ontology (GO)
- GO is a database of terms, describing gene sets in terms of their
  - functions (12,093)
  - processes (1,812)
  - components (7,459)
- Genes are annotated to GO terms
- Terms are connected (is_a, part_of)
- Levels represent terms generality



Using domain ontologies

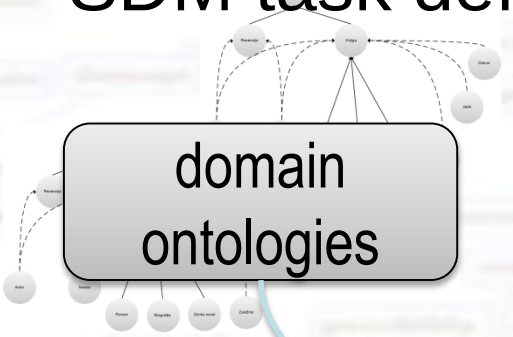
- Using background knowledge in data mining has been a topic of extensive research
  - Hierarchical attribute values, hierarchy/taxonomy of attributes, since 1986
  - ILP, relational data mining, propositionalization, since 1991
  - Ontologies (Tim Berners-Lee), since 1989
    - accepted formalism for consensual knowledge representation for Semantic Web applications, a basic for the Semantic Web
    - Description logic, OWL, Protégé ontology editor
  - Using ontologies in data mining, since 2004

Semantic Data Mining

- Ontology-driven (semantic) data mining is an emerging research topic
- Semantic Data Mining (SDM) - a term denoting:
  - the new challenge of mining semantically annotated resources, with ontologies used as background knowledge in mining experimental data
  - approaches with which semantic data are mined

Semantic Data Mining

SDM task definition used in our work



domain ontologies

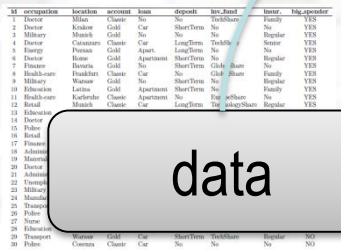
annotations, mappings

Semantic data mining

model, patterns

```
target(A) :-
   'Doctor'(A), 'Italy'(A).

target(A) :-
   'Service'(A), 'Germany'(A).
```



data

Given:

- transaction data table, relational database, text documents, Web pages, …
- one or more domain ontologies
- a mapping from examples to ontology concepts

Find: a classification model, a set of patterns

Our early work: Semantic subgroup discovery

- The approach: Using relational subgroup discovery in the SDM context
  - General purpose system **RSD** for **Relational Subgroup Discovery**, using a propositionalization approach to relational data mining (Železny and Lavrac, MLJ 2006)
  - Applied to semantic data mining in a biomedical application by using the Gene Ontology as background knowledge in analyzing microarray data

RSD: Propositionalization approach to RDM and SDM

customer

ID	Zip	Sex	So St	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

Step 1

Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

1. constructing relational features
2. constructing a propositional table

RSD: Propositionalization approach to RDM and SDM

customer

ID	Zip	S ex	So St	In come	A ge	Cl ub	Re sp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

order

Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

store

Store ID	Size	Type	Location
...	...	...	...
12	small	franchise	city
17	large	indep	rural
...	...	...	...

Relational representation of customers, orders and stores.

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

Step 1

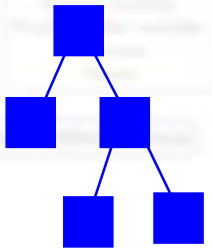
Propositionalization

	f1	f2	f3	f4	f5	f6	...				...	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

1. constructing relational features
2. constructing a propositional table

Step 2

Data Mining



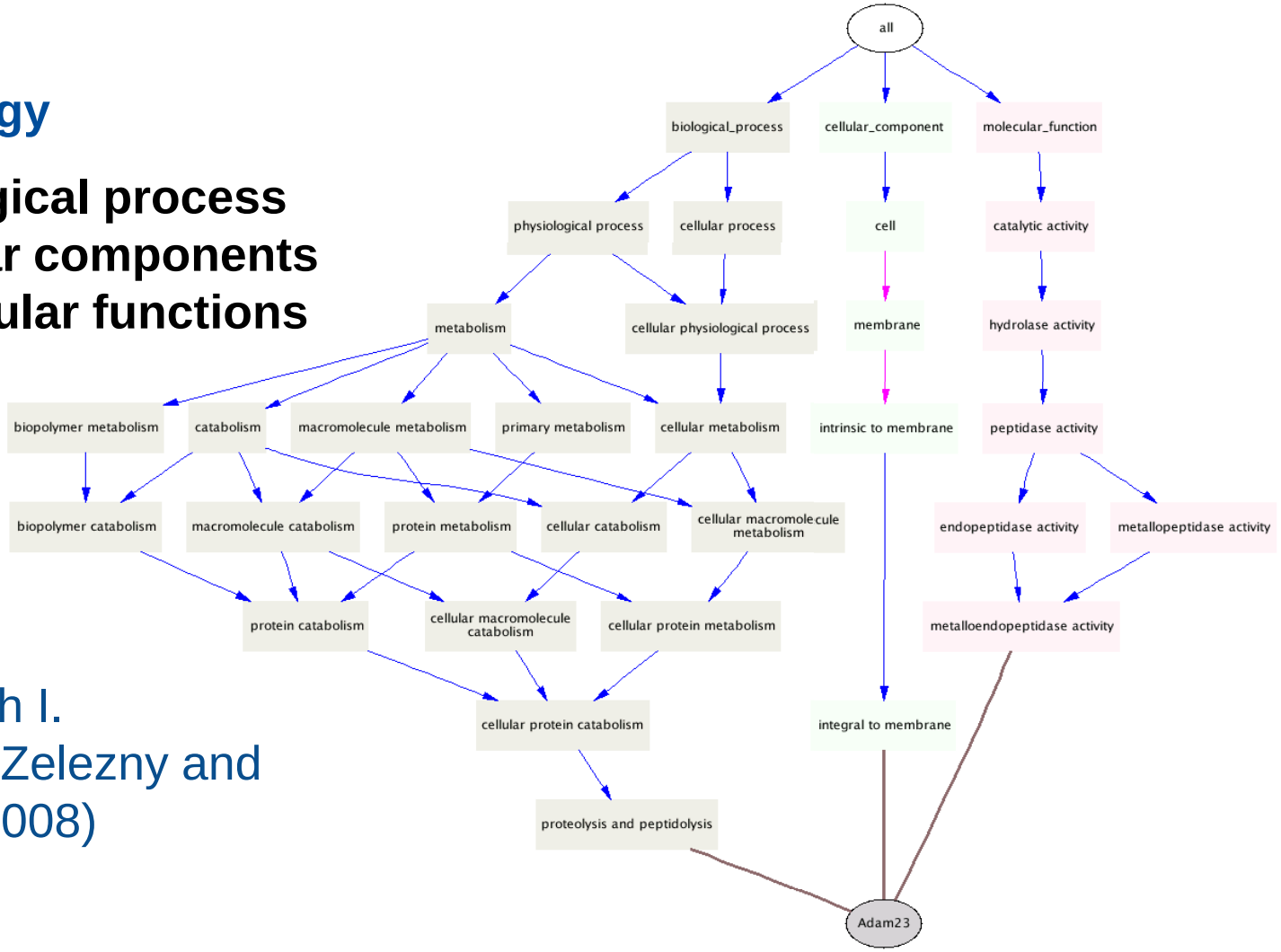
model, patterns, …

Semantic subgroup discovery with RSD

Using GO as background knowledge in DNA microarray data analysis with relational subgroup discovery system RSD

Gene Ontology

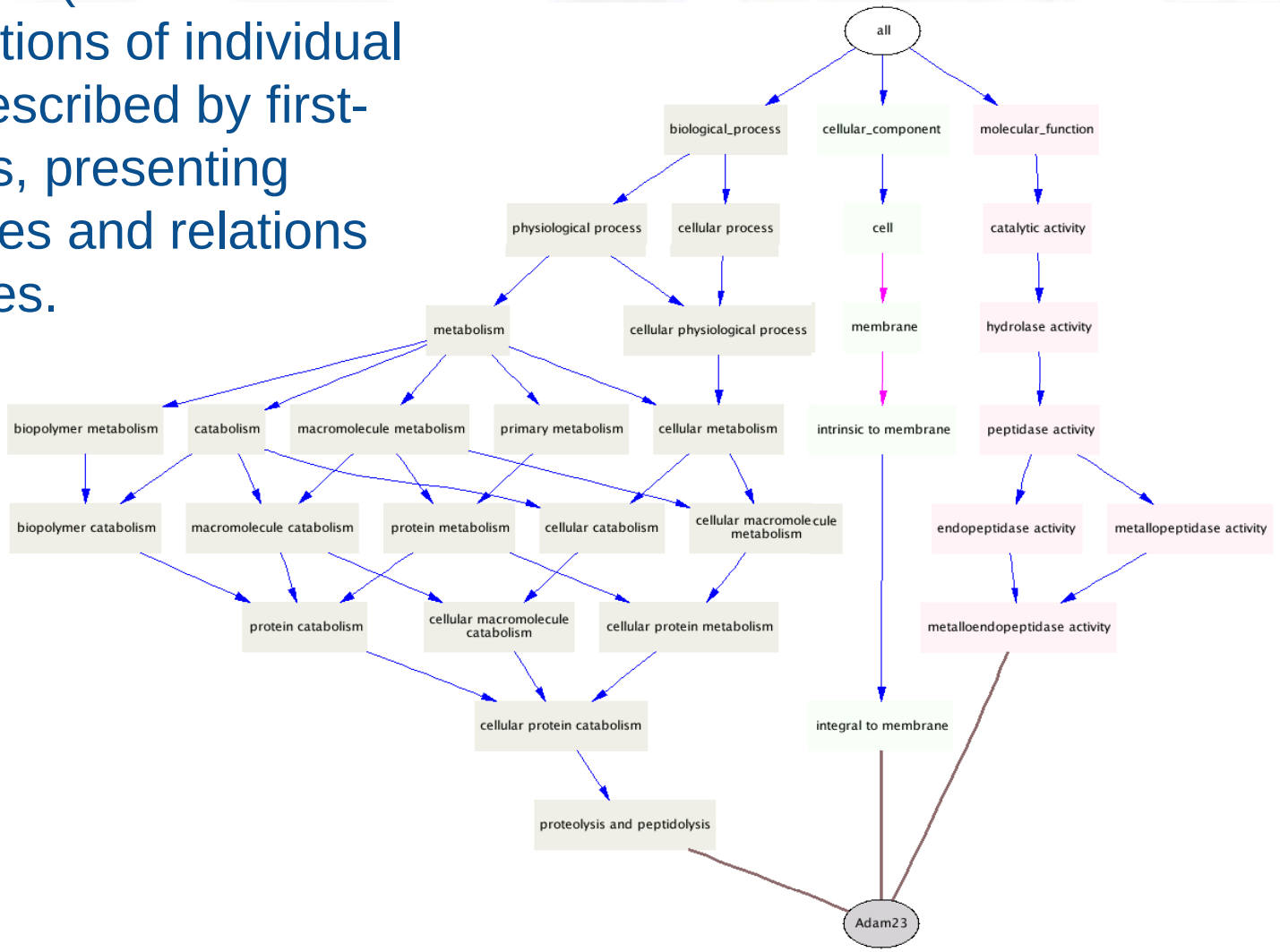
12,093 biological process
1,812 cellular components
7,459 molecular functions



Joint work with I. Trajkovski, F. Zelezny and J. Tolar (JBI 2008)

Semantic subgroup discovery with RSD

Ontology terms (can be viewed as generalisations of individual genes) are described by first-order features, presenting gene properties and relations between genes.



Semantic subgroup discovery with RSD

Application of RSD to microarray data analysis using GO as background knowledge (Zelezny et al. 2006, Trajkovski et al. 2008)

1. Take ontology terms represented as logical facts in Prolog, e.g.

```
component(gene2532,'GO:0016020').
function(gene2534,'GO:0030554').
process(gene2534,'GO:0007243').
interaction(gene2534,gene4803).
```

2. Automatically generate generalized relational features:

```
f(2,A):-component(A,'GO:0016020').
f(7,A):-function(A,'GO:0030554').
f(11,A):-process(A,'GO:0007243').
f(224,A):- interaction(A,B), function(B,'GO:0016787'),
                    component(B,'GO:0043231').
```

3. Propositionalization: Determine truth values of features

4. Learn rules by a subgroup discovery algorithm CN2-SD

Semantic subgroup discovery with RSD

Step 2: Construction of first order features with supp. > *min_supp.*

```
f(7,A):-function(A,'GO:0046872').
f(8,A):-function(A,'GO:0004871').
f(11,A):-process(A,'GO:0007165').
f(14,A):-process(A,'GO:0044267').
f(15,A):-process(A,'GO:0050874').
f(20,A):-function(A,'GO:0004871'), process(A,'GO:0050874').
f(26,A):-component(A,'GO:0016021').
f(29,A):- function(A,'GO:0046872'), component(A,'GO:0016020').
f(122,A):-interaction(A,B),function(B,'GO:0004872').
f(223,A):-interaction(A,B),function(B,'GO:0004871'),
    process(B,'GO:0009613').
f(224,A):-interaction(A,B),function(B,'GO:0016787'),
    component(B,'GO:0043231').
```

existential

RSD propositionalization step

diffexp g1 (gene64499)
diffexp g2 (gene2534)
diffexp g3 (gene5199)
diffexp g4 (gene1052)
diffexp g5 (gene6036)
....

random g1 (gene7443)
random g2 (gene9221)
random g3 (gene2339)
random g4 (gene9657)
random g5 (gene19679)
....

	f1	f2	f3	f4	f5	f6	…				…	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

RSD: Rule construction with CN2-SD

	f1	f2	f3	f4	f5	f6	…			…	fn
g1	1	0	0	1	1	1	0	0	1	0	1	1
g2	0	1	1	0	1	1	0	0	0	1	1	0
g3	0	1	1	1	0	0	1	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1	0
g1	0	0	1	1	0	0	0	1	0	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1	1
g3	0	0	0	0	1	0	0	1	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0	1

diffexp (Gene) ← f2 ∧ f3 [4,0]

RSD subgroup discovery in two steps

- **Step 1:** **Construct relational logic features** of genes such as

  interaction(A, G) & function(G, protein_binding)

  (*gene A interacts with another gene whose functions include protein binding*)

  and **propositional table construction** with features as attributes

- **Step 2:** Using these features to **discover and describe subgroups of genes** that are differentially expressed (e.g., belong to class DIFF.EXP. of top 300 most differentially expressed genes) in contrast with RANDOM genes (randomly selected genes with low differential expression).

- Sample subgroup description:

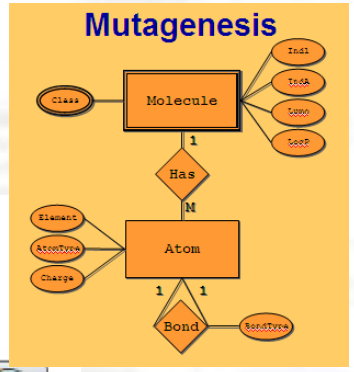
```
diffexp(A) :- interaction(A,B) AND
              function(B,'GO:0004871') AND
              process(B,'GO:0009613')
```

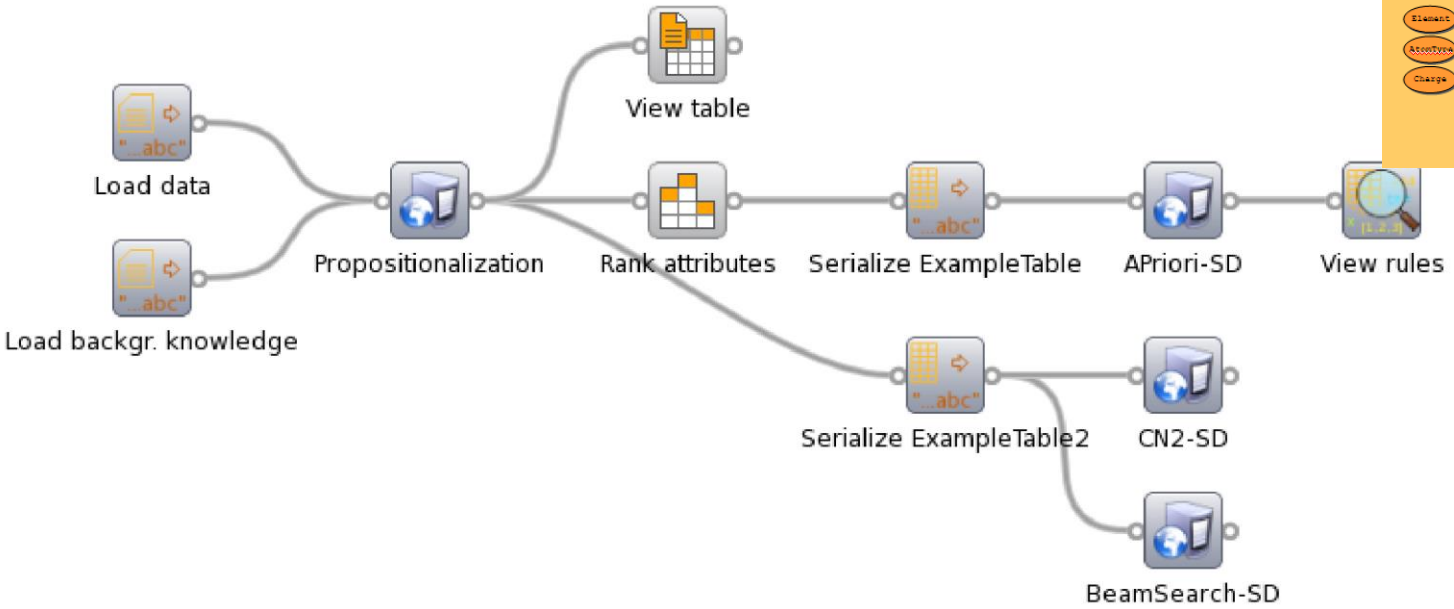
RSD implementation in Orange4WS

RSD implemented as a workflow in Orange4WS

- propositionalization
- subgroup discovery algorithms: SD, Apriori-SD, CN2-SD
- applied also to standard ILP problems, e.g. mutagenicity prediction

mutagenic(M) ← feature121(M), feature235(M)





Semantic subgroup discovery with SEGS

- More recent work: semantic subgroup discovery with SEGS
- **Gene set enrichment**: moving from single gene to gene set analysis
  - A gene set is enriched if the genes in the set are statistically significantly differentially expressed compared to the rest of the genes.
  - Observation: e.g., an 20% increase in all genes members of a biological pathway may alter the execution of this pathway … and its impact on other processes … significantly more then a 10-fold increase in a single gene.

Semantic subgroup discovery with SEGS

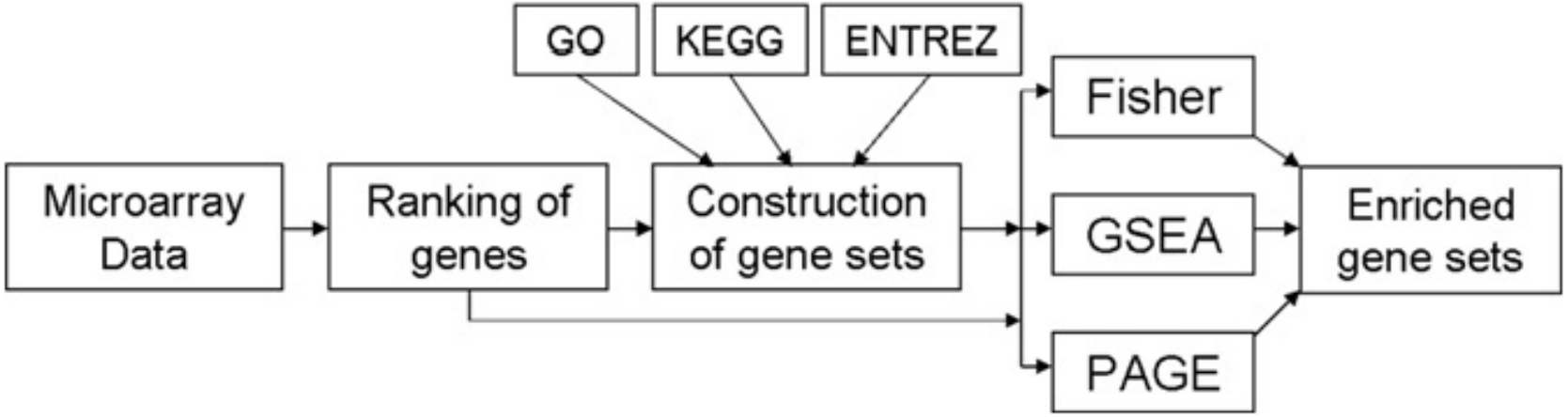
- Gene set enrichment methods:
  - Single GO terms:
    - Gene Set Enrichment Analysis (GSEA)
    - Parametric Analysis of Gene Set Enrichment (PAGE)
  - Conjunctions of GO terms: **SEGS** - special purpose semantic subgroup discovery system for **Searching for Enriched Gene Sets**

Semantic subgroup discovery with SEGS

- **The SEGS approach:**
  - fuse information from GO, KEGG and ENTREZ
    - **Gene Ontology (GO):** standardized biological terms used to annotate gene products: Molecular Functions, Biological Processes, Cellular Components
    - **Kyoto Encyclopedia of Genes and Genomes (KEGG):** manually drawn pathway maps representing the knowledge on the molecular interaction and reaction networks
    - **ENTREZ:** gene annotations with GO and KO terms and gene-gene interaction data
  - **generate gene set candidates** by performing top-down search of rules, formed as conjunctions of ontology terms as conjunctions of GO, KEGG and ENTREZ terms
  - **combine Fisher, GSEA and PAGE enrichment tests** to select most interesting groups of differentially expressed genes

Semantic subgroup discovery with SEGS

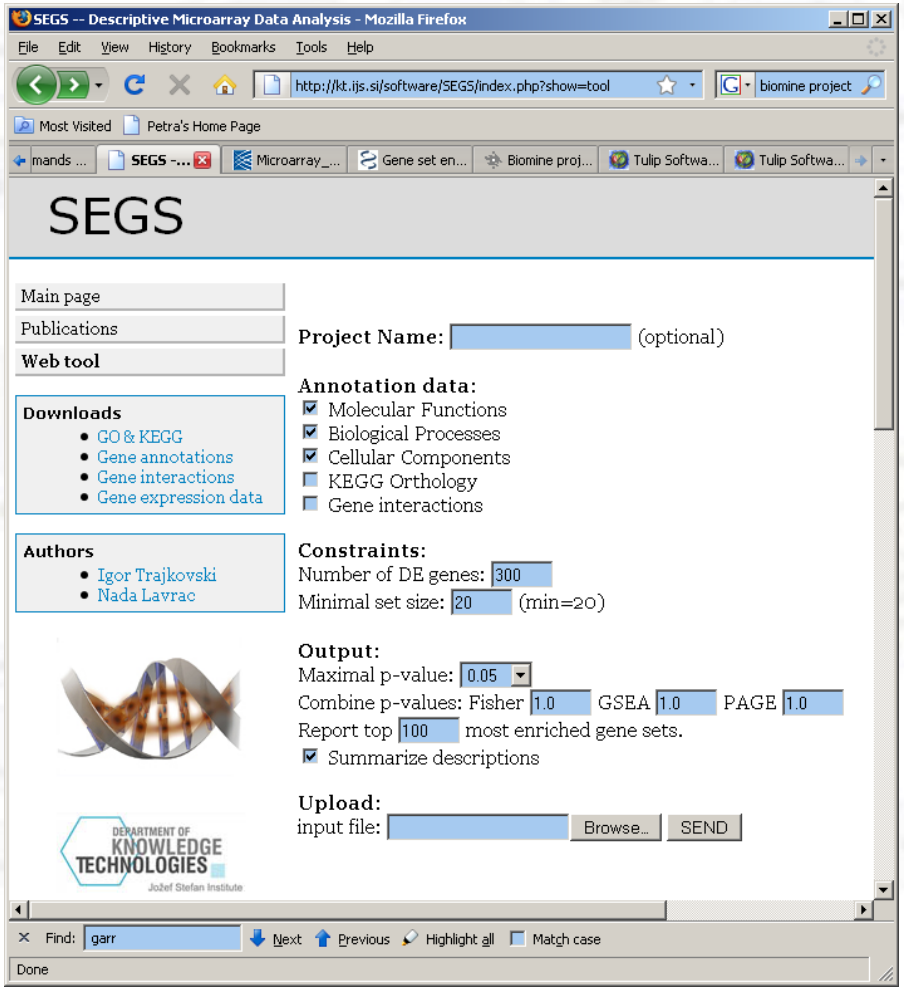
- SEGS workflow is implemented in the Orange4WS data mining environment

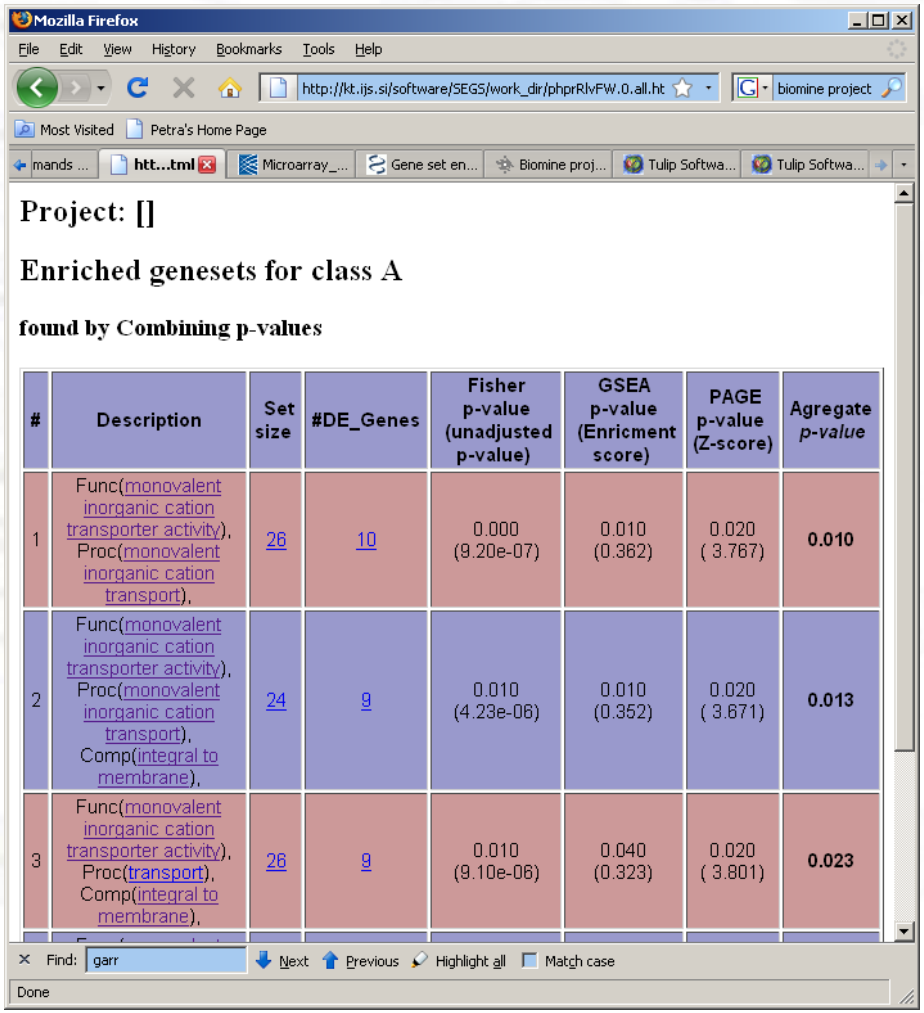


- SEGS is also implemented also as a Web applications

(Trajkovski et al., IEEE TSMC 2008, Trajkovski et al., JBI 2008)

Semantic subgroup discovery with SEGS





#	Description	Set size	#DE_Genes	Fisher p-value (unadjusted p-value)	GSEA p-value (Enricment score)	PAGE p-value (Z-score)	Agregate *p-value*
1	Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport),	26	10	0.000 (9.20e-07)	0.010 (0.362)	0.020 (3.767)	**0.010**
2	Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport), Comp(integral to membrane),	24	9	0.010 (4.23e-06)	0.010 (0.352)	0.020 (3.671)	**0.013**
3	Func(monovalent inorganic cation transporter activity), Proc(transport), Comp(integral to membrane),	26	9	0.010 (9.10e-06)	0.040 (0.323)	0.020 (3.801)	**0.023**

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Biomine graph exploration (Toivonnen et al.)

- SEGS can be combined with other biomedical resources, such as BioMine
- **BioMine graph** contains information from public databases, including annotated sequences, proteins, orthology groups, genes and gene expressions, gene and protein interactions, PubMed articles, and different ontologies.
  - **nodes (~1 mio)** correspond to different concepts (such as gene, protein, domain, phenotype, biological process, tissue)
  - **semantically labeled edges (~7 mio)** connect related concepts
- **BioMine query engine** answers queries to potentially discover new links between entities by sophisticated graph exploration algorithms

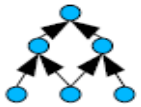
Complex data mining methodology
SegMine = SEGS + BioMine

SegMine overview

1donor1-P2	1donor2-P2	1donor3-P2	2donor1-P11	2donor2-P7	2donor3-P8
25.71	41.29	33.11	49.53	54.89	36.59
8.15	11.84	12.85	6.7	7.61	9.82
7.69	108.73	291.82	9.71	105.98	84.38
95.46	86.82	110.13	118.57	92.53	118.26
1.53	1.11	15.98	1.41	1.25	5.03
50.94	53.07	36.16	43.25	73.51	32.19
2.89	0.64	4.24	1.63	6.91	4.41
184.58	150.62	119.35	141.87	155.45	157.76
5.45	1.51	0.72	0.34	2.83	0.65
292.55	359.93	465.48	289.12	344.66	291.91
9.34	12.14	9.67	7.82	5.39	8.37
7.04	52.98	47.63	89.49	55.46	40.43
4.41	39.9	17.72	26.42	19.17	12.15
0.35	0.65	2.2	0.34	0.41	1.95
31.09	43.62	151.49	25.51	101.89	26.77

raw data from a microarray experiment (expression of genes)

SEGS



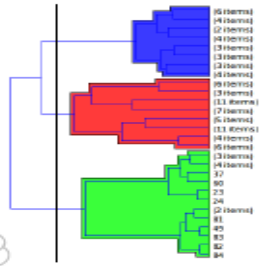
knowledge from ontologies

RULE 1 := organelle organization
AND intracellular non-membrane-bounded organelle
AND INTERACT: transcription coactivator activity

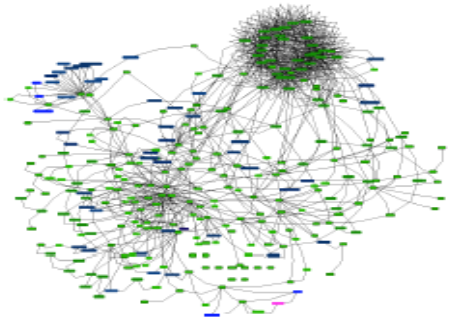
RULE 2 := cellular macromolecule metabolic process
AND nuclear part
AND INTERACT: chromatin binding

RULE 3 := cellular response to stimulus
AND intracellular organelle part
AND INTERACT: RNA binding

expert analysis



interpretation of gene expression data: **rules**, **clusters**, **genesets**



Biomine

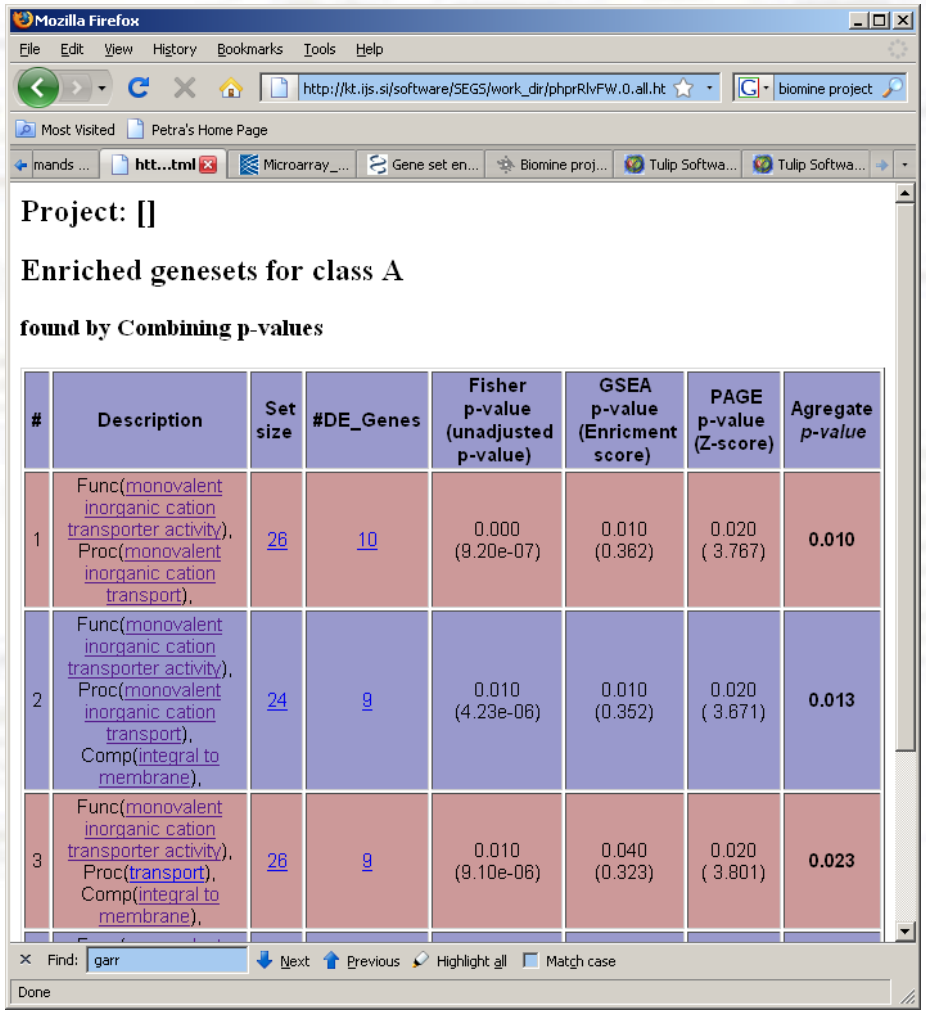
public databases

expert analysis

Podpečan et al., BMC Bioinformatics 2011

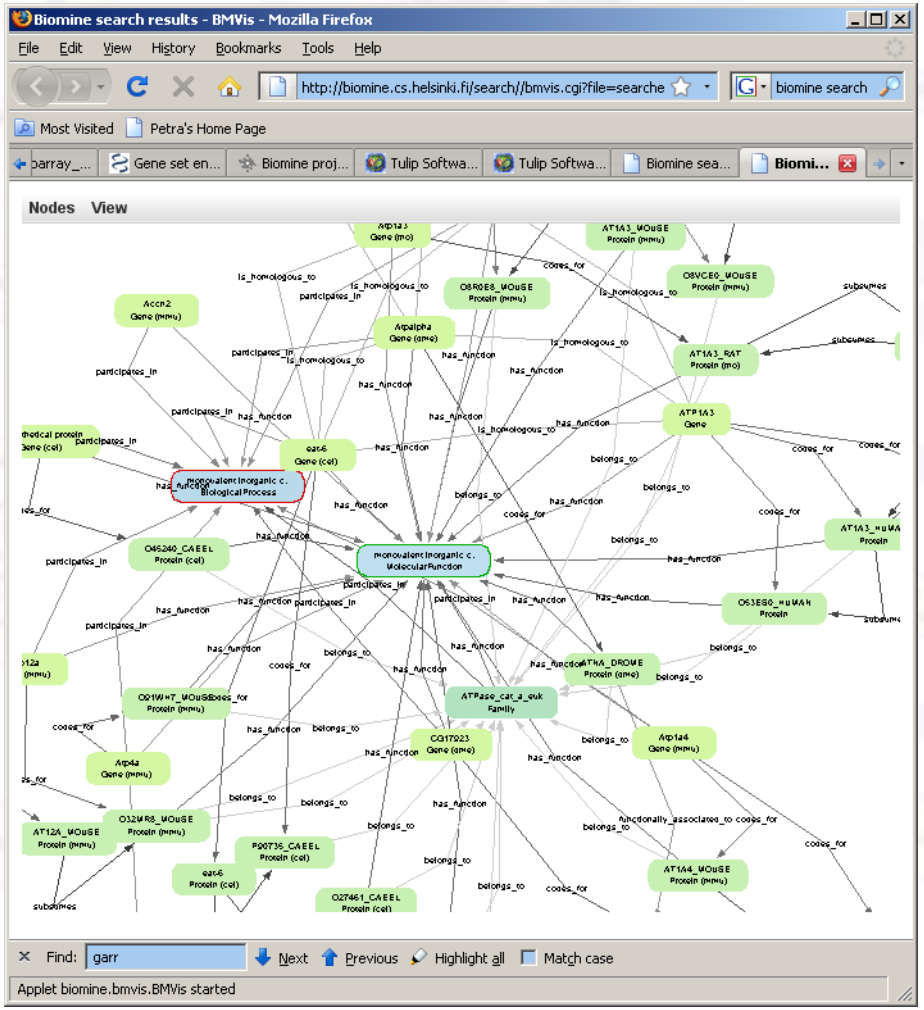
SEGS + BioMine outputs

SEGS output:

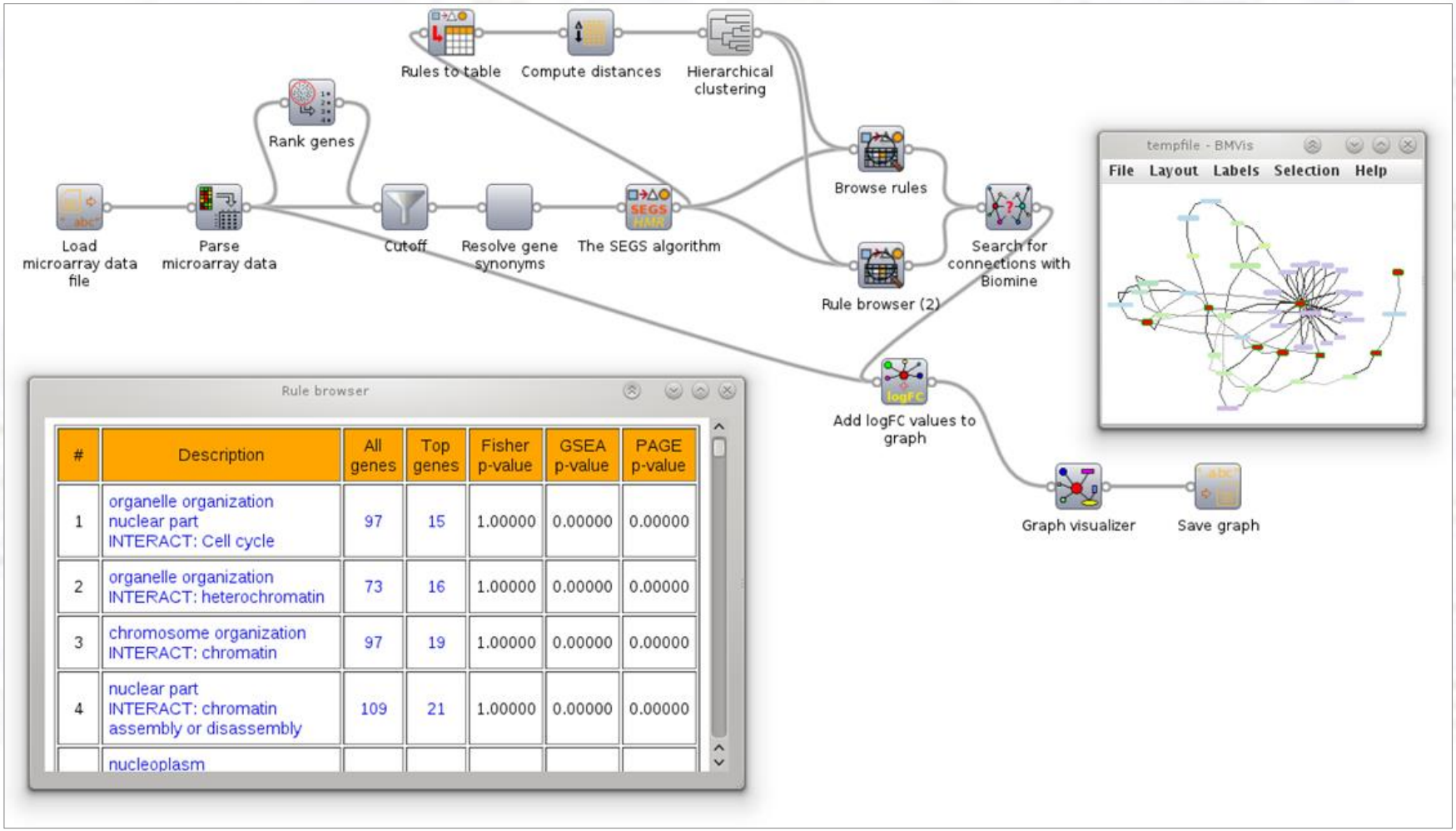


#	Description	Set size	#DE_Genes	Fisher p-value (unadjusted p-value)	GSEA p-value (Enricment score)	PAGE p-value (Z-score)	Agregate *p-value*
1	Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport),	26	10	0.000 (9.20e-07)	0.010 (0.362)	0.020 (3.767)	**0.010**
2	Func(monovalent inorganic cation transporter activity), Proc(monovalent inorganic cation transport), Comp(integral to membrane),	24	9	0.010 (4.23e-06)	0.010 (0.352)	0.020 (3.671)	**0.013**
3	Func(monovalent inorganic cation transporter activity), Proc(transport), Comp(integral to membrane),	26	9	0.010 (9.10e-06)	0.040 (0.323)	0.020 (3.801)	**0.023**

BioMine query output:



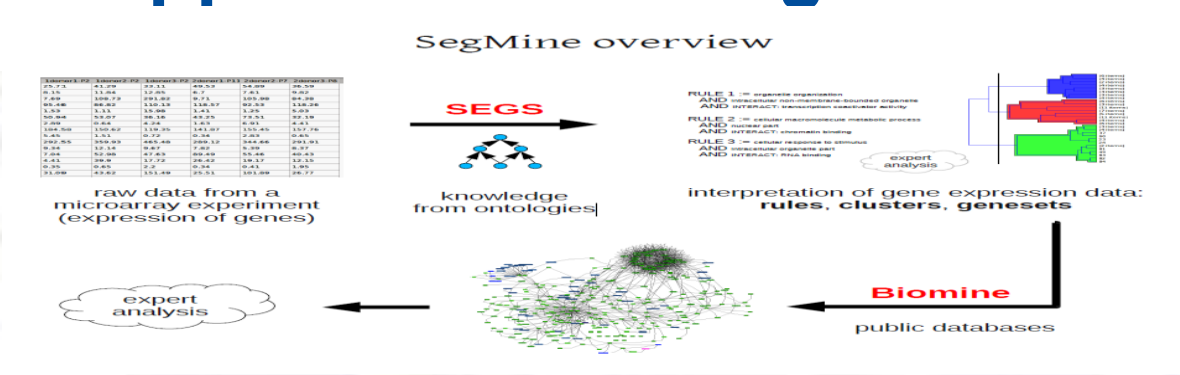
SegMine methodology implemented in Orange4WS



#	Description	All genes	Top genes	Fisher p-value	GSEA p-value	PAGE p-value
1	organelle organization nuclear part INTERACT: Cell cycle	97	15	1.00000	0.00000	0.00000
2	organelle organization INTERACT: heterochromatin	73	16	1.00000	0.00000	0.00000
3	chromosome organization INTERACT: chromatin	97	19	1.00000	0.00000	0.00000
4	nuclear part INTERACT: chromatin assembly or disassembly	109	21	1.00000	0.00000	0.00000
	nucleoplasm					

http://segmine.ijs.si/

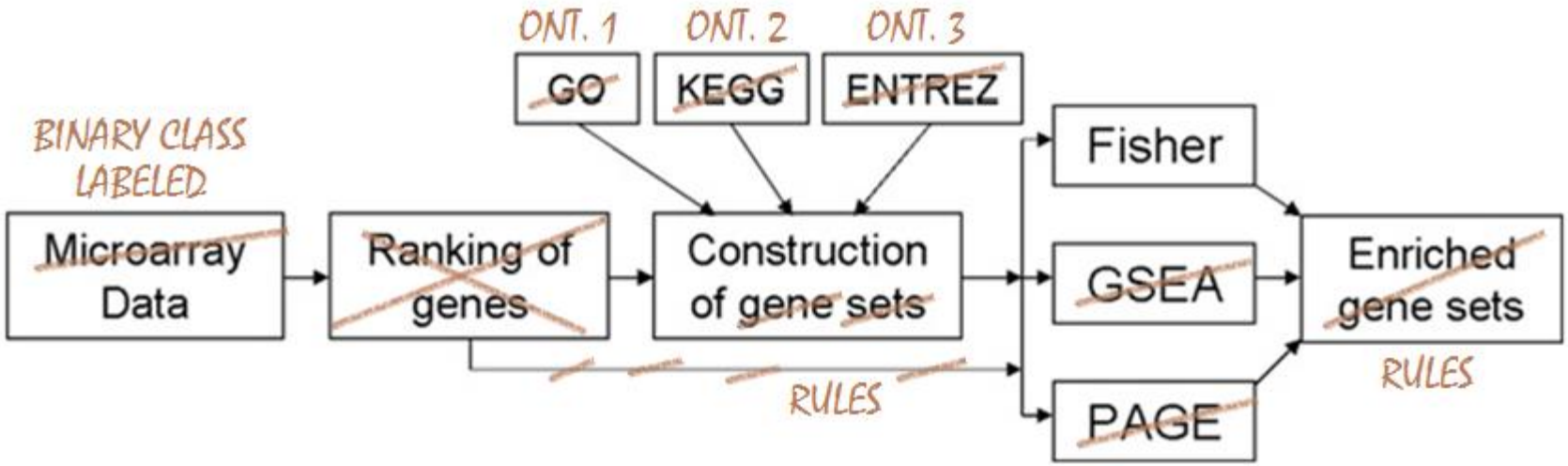
Biomedical applications of SegMine methodology



- Combination of advanced data processing and mining algorithms
- Enables semantic analysis of gene expression using background knowledge in the form of ontologies
- SegMine tool is actively used at the National Institute for Biology
- Successful application in human stem cell data analysis: new hypotheses, enabling better understanding of cell senescence mechanisms
- General purpose Semantic Data Mining algorithm g-SEGS is also available in Orange4WS

From SEGS to SDM-SEGS: Generalizing SEGS

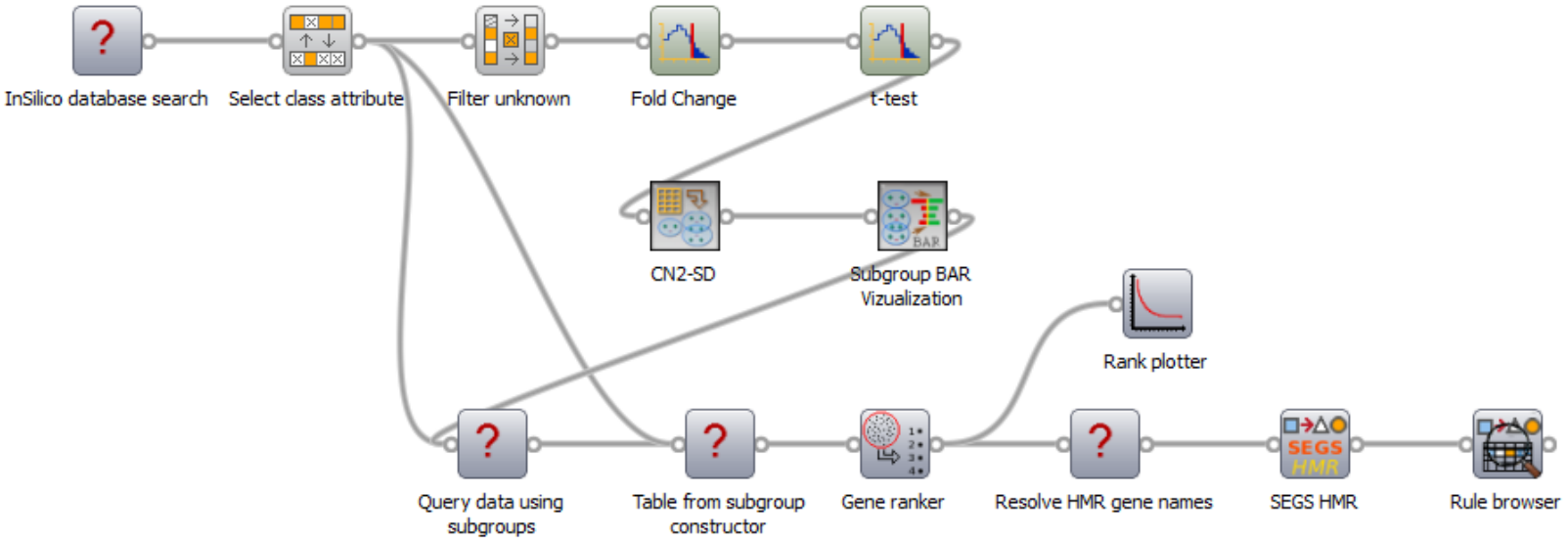
- SDM-SEGS: a semantic data mining system generalizing SEGS



- Discovers subgroups both for ranked and labeled data
- Exploits input ontologies in OWL format
- Implemented as a web service in Orange4WS
  - Can also be used e.g. in Taverna

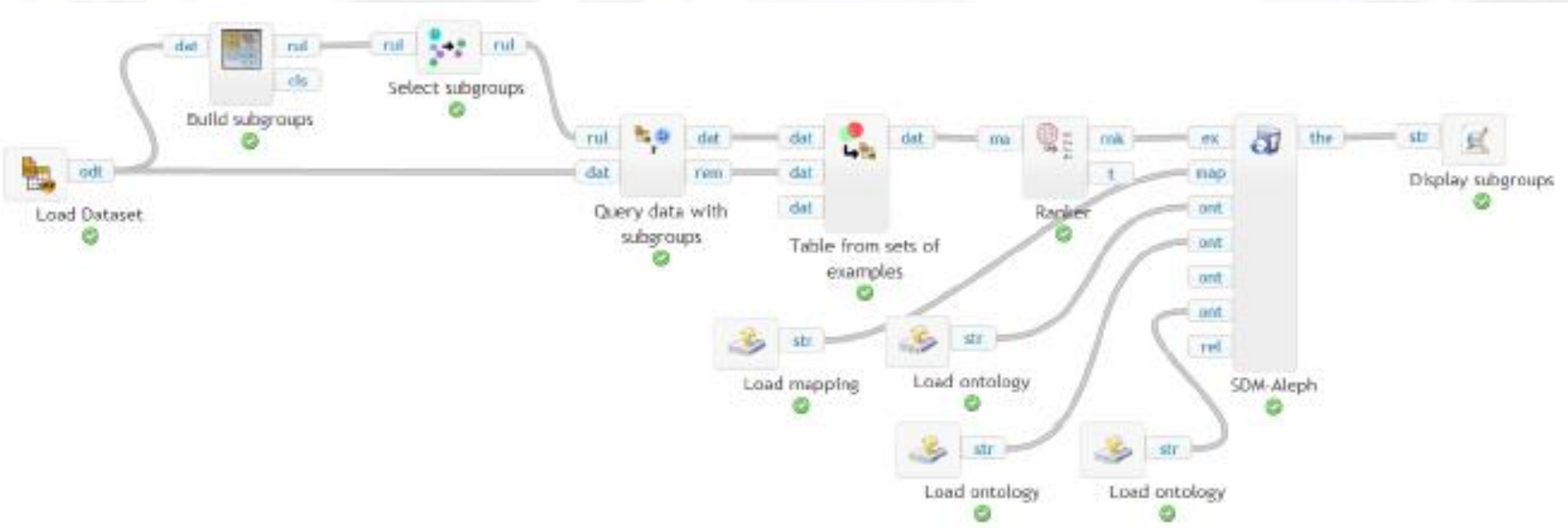
Recent work

- Semantic Subgroup Discovery workflows in Orange4WS and ClowdFlows (Vavpetič et al., 2012)



Recent biomedical applications

- Subgroup discovery and semantic explanation of subgroups on breast cancer data (Vavpetič et al., JIIS 2014)



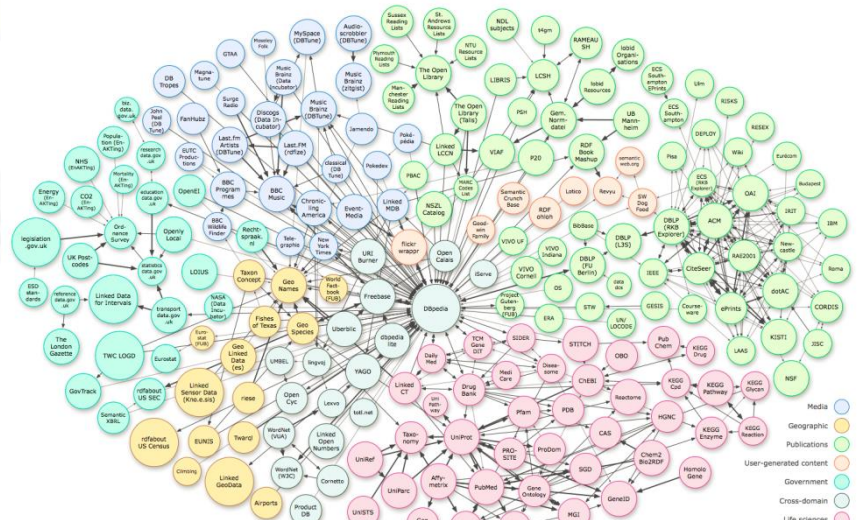
- The workflow, implemented in ClowdFlows, is available for sharing at http://clowdflows.org/workflow/1283/

Future challenges for Semantic Data Mining

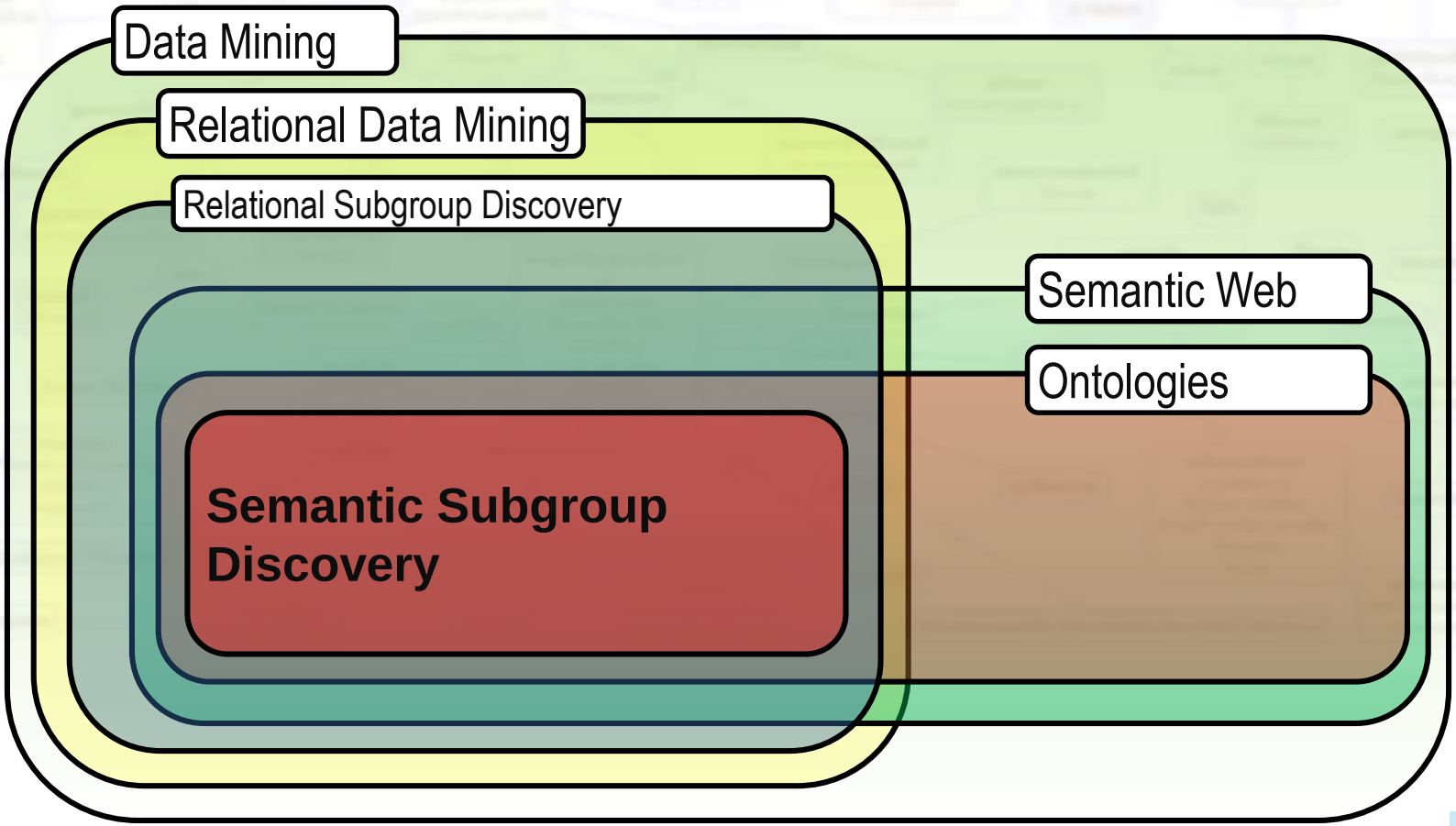
- Current SDM scenario: Mining empirical data with ontologies as background knowledge
  - abundant empirical data, but
  - scarce background knowledge
- Future SDM scenarios:
  - envisioning a growing amount of semantic data
    - abundance of ontologies and semantically anotated data collections
    - e.g. Linked Data
      - over 6 billion RDF triples
      - over 148 million links

Future work

- We may envision a paradigm shift from data mining to knowledge mining
- The envisioned future Semantic data mining scenario in mining the Semantic Web:
  - mining knowledge encoded in domain ontologies,
  - constrained by annotated (empirical) data collections.



Summary: RDM and SDM in Context



Acknowledgements

- Work on subgroup discovery was done jointly with D. Gamberger (RBI) and P. Kralj Novak (JSI)

- Work on relational data mining and semantic data mining was done jointy with I.Trajkovski (Skopje Uni.), F. Železny (CTU, Prague), J. Tolar (Univ. of Minnesota), I. Mozetič and A. Vavpetič (JSI), and colleagues from the National Institute of Biology

- Work on new data mining platforms and advanced workflows was done jointly with A. Vavpetič, V. Podpečan and J. Kranjc (JSI)

Acknowledgements