Semantic Relational Learning

Nada Lavrač
Jožef Stefan Institute,
Ljubljana, Slovenia

Novi Sad
October 4, 2019

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Jožef Stefan Institute, Ljubljana, Slovenia

- **Jožef Stefan Institute (JSI, founded in 1949)**
  - named after a distinguished physicist Jožef Stefan (1835-1893)

$$j = \sigma T^4$$

  - leading national research organization in natural sciences and technology (~700 researchers and students)

- **Jožef Stefan International Postgraduate School (founded in 2004)**
  - Offers four MSc and PhD programs (in English): ICT, nanotechnologies, ecotechnologies and sensor technologies

Department of Knowledge Technologies

- **Head:** Nada Lavrač, **Staff:** 45 researchers
- **Knowledge Technologies**
  - Making AI techniques operational for practical problems

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Department of Knowledge Technologies

- **Head:** Nada Lavrač, **Staff:** 45 researchers
- **Knowledge Technologies**
  - Making AI techniques operational for practical problems
- **Main research areas**
  - Data Mining and Machine Learning
  - Text Mining and Human Language Technologies
  - Web Services and Semantic Web
  - Ontologies and Knowledge Management
  - Decision Support Systems
- **Applications**
  - Medicine, Bioinformatics, Public Health
  - Ecology, Finance, …

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Department of Knowledge Technologies

- **My research preferences**
  - Data Mining
  - Text Mining
  - Web Services and Workflows
  - Knowledge Management

- **Applications**
  - Medicine, Bioinformatics
  - Public Health

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Talk outline

- **First Generation Data Mining**
  – Basics of Machine Learning and Data Mining
- **Second Generation Data Mining**
  – Selected Algorithms and Biomedical Applications
- **Third Generation DM Techniques and Platforms**
  – Relational Data Mining
  – Semantic Relational Learning: Using ontologies in DM
- **Current Work and Conclusions**

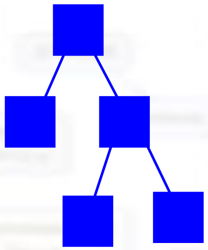
Machine Learning and Data Mining

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

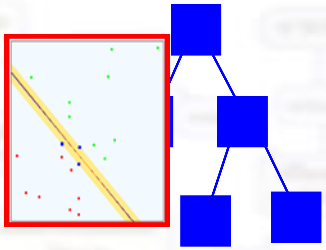
Machine Learning and Data Mining

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

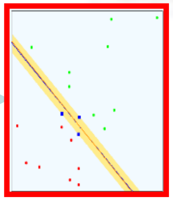
Machine Learning
Data Mining



model, patterns, …

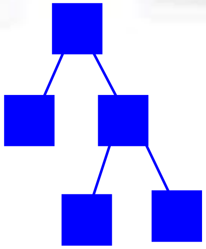
Given: class labeled data
Find: classification model or
set of interesting patterns in the data

new unclassified instance



classified instance

black box classifier
no explanation



symbolic model
symbolic patterns

explanation

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

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

Pattern discovery 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

PATTERN

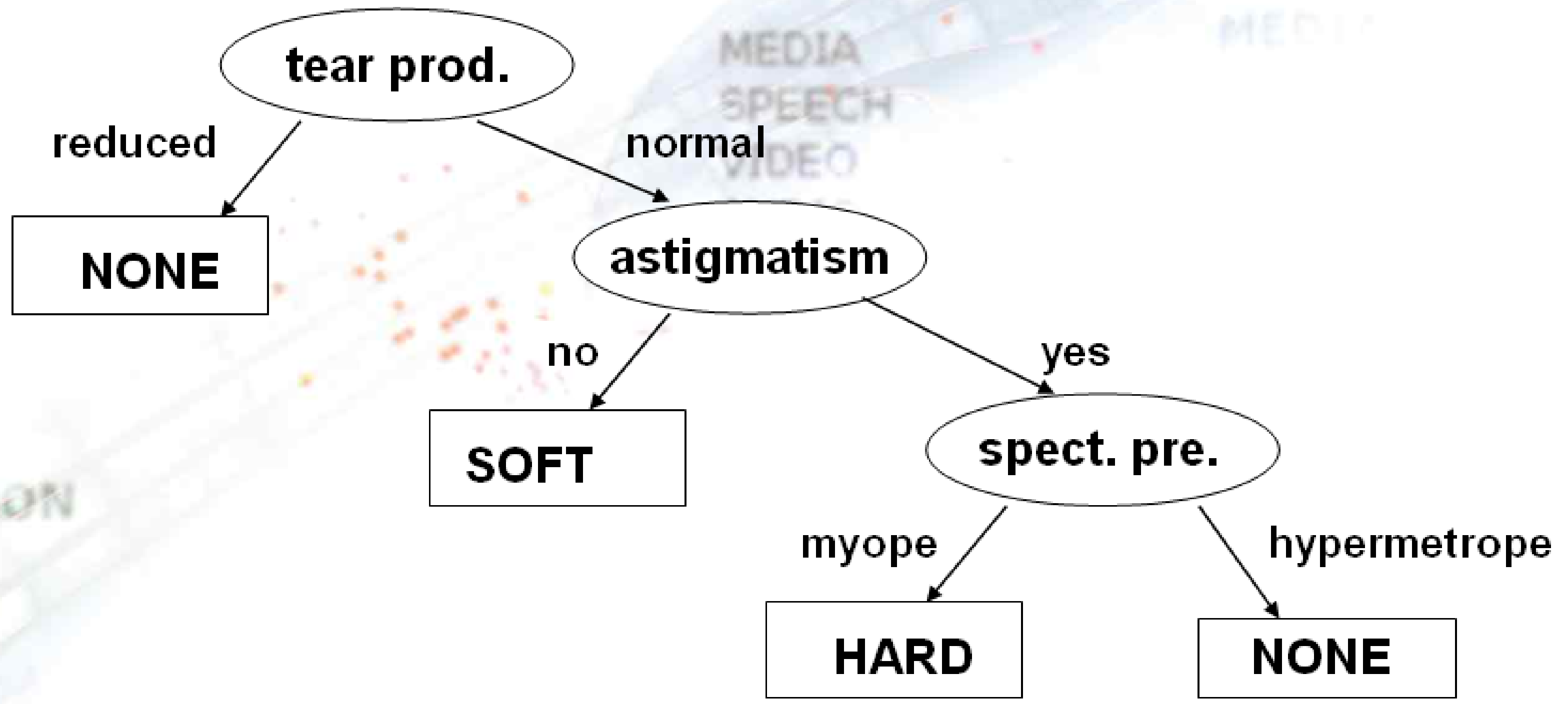
Rule:

IF
Tear prod. =
reduced

THEN
Lenses =
NONE

Classical machine learning techniques for knowledge discovery in data

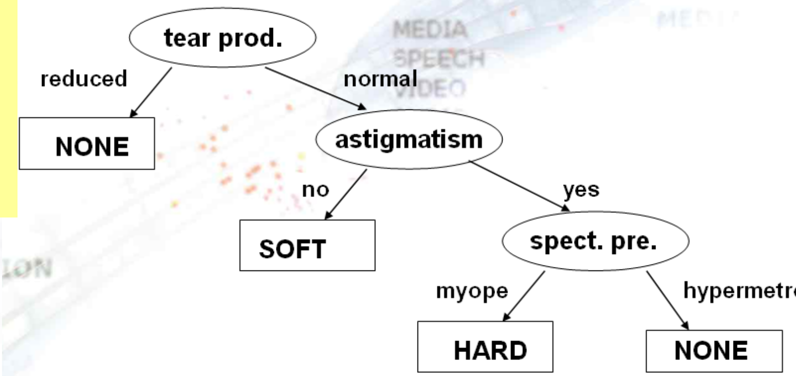
KNOWLEDGE = a model whose validity is confirmed by the domain expert



USE of AUTOMATICALLY INDUCED KNOWLEDGE as additional expert opinion for decision support

Example: Learning a classification model 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



$$Gain(S,A) = E(S) - \sum_{v \in Values(A)} p_v \cdot E(S_v)$$

Heuristic for determining the most informative attributa:

Gain(S,A) estimate of reduced entropy of dataset S after splitting the date based on values of attribute A

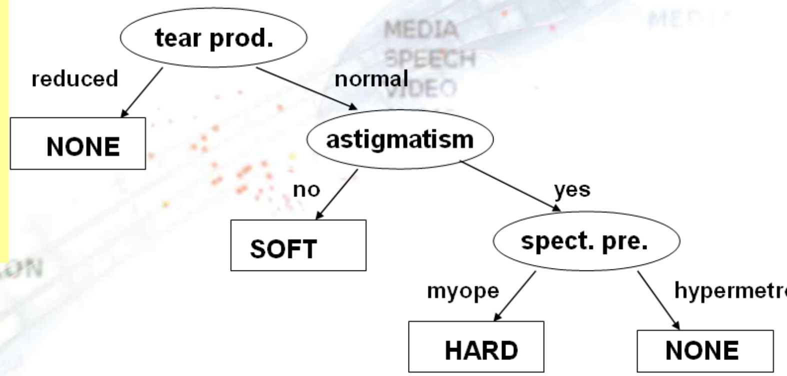
Entropy measure of impurity of training set S: $E(S) = - p_+ \log_2 p_+ - p_- \log_2 p_-$

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Example: Learning a classification model 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 AND astigmatism=yes AND spect. pre.=hypermetrope

lenses=SOFT ← tear production=normal AND astigmatism=no

lenses=HARD ← tear production=normal AND astigmatism=yes AND spect. pre.=myope

lenses=NONE ←

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

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)

- simplified single concept learning
- “one vs. all” multi-class learning

Other tasks: Learning from Numeric Class Data

Person	Age	Spect. presc.	Astigm.	Tear prod.	LensPrice
O1	17	myope	no	reduced	0
O2	23	myope	no	normal	8
O3	22	myope	yes	reduced	0
O4	27	myope	yes	normal	5
O5	19	hypermetrope	no	reduced	0
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	5
O15	43	hypermetrope	yes	reduced	0
O16	39	hypermetrope	yes	normal	0
O17	54	myope	no	reduced	0
O18	62	myope	no	normal	0
O19-O23	...	...	...	...	...
O24	56	hypermetrope	yes	normal	0

Numeric class values – regression analysis

Other tasks: Learning from Unlabeled 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

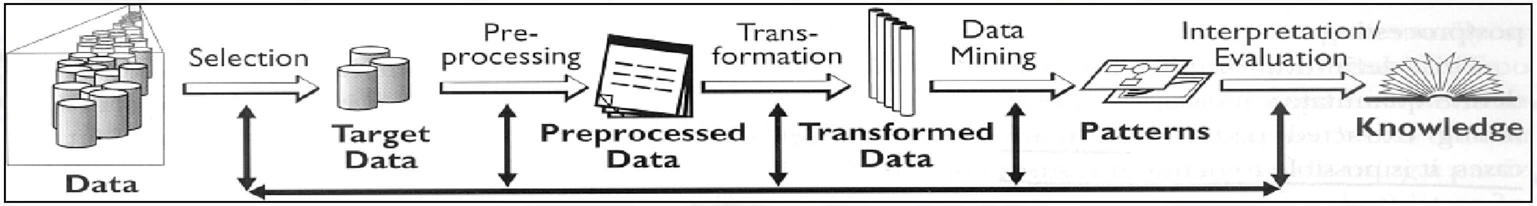
Unlabeled data - clustering: grouping of similar instances
(similar instances – many common values)

First Generation Data Mining

- **First machine learning algorithms for**
  - Decision tree and rule learning in 1970s and early 1980s by Quinlan, Michalski et al., Breiman et al., …
- **Characterized by**
  - Learning from data stored in a single data table
  - Relatively small set of instances and attributes
- **Lots of ML research followed in 1980s**
  - Numerous conferences ICML, ECML, … and ML sessions at AI conferences IJCAI, ECAI, AAAI, …
  - Extended set of learning tasks and algorithms addressed

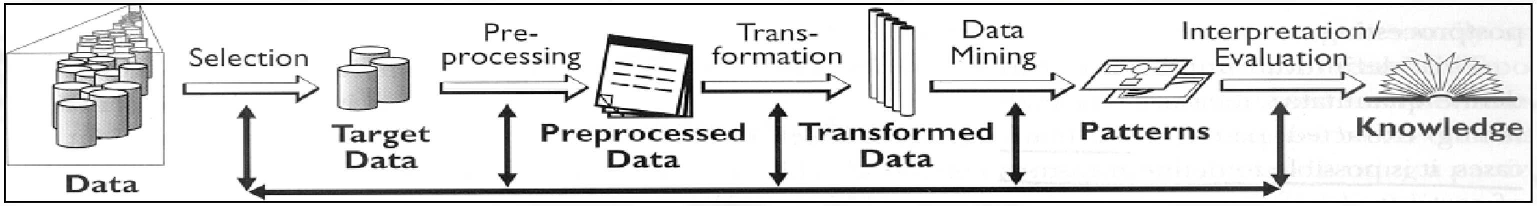
Second Generation Data Mining

- **Developed since 1990s:**
  - Focused on data mining tasks characterized by large datasets described by large numbers of attributes
  - Industrial standard: CRISP-DM methodology (1997)



Second Generation Data Mining

- **Developed since 1990s:**
  - Focused on data mining tasks characterized by large datasets described by large numbers of attributes
  - Industrial standard: CRISP-DM methodology (1997)



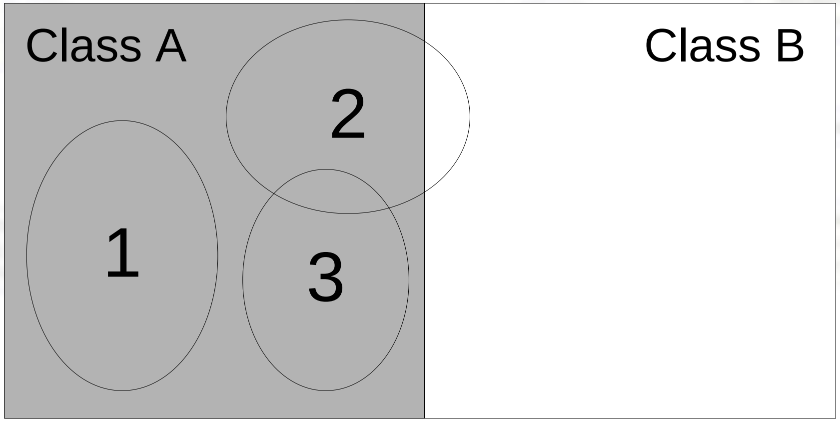
  - New conferences on practical aspects of data mining and knowledge discovery: KDD, PKDD, …
  - New learning tasks and efficient learning algorithms:
    - Learning predictive models: Bayesian network learning,, relational data mining, statistical relational learning, SVMs, …
    - Learning descriptive patterns: association rule learning, subgroup discovery, …

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Subgroup Discovery

- Data transformation:
  - binary class values (positive vs. negative examples of Target class)
- Subgroup discovery:
  - a task in which individual interpretable patterns in the form of rules are induced from data, labeled by a predefined property of interest.
- SD algorithms learn several independent rules that describe groups of target class examples
  - subgroups must be large and significant

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



Subgroup discovery in 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 enough, distributionaly unusual)

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)

Subgroup discovery in functional genomics

- Functional genomics is a typical scientific discovery domain, studying genes and their functions
- Very large number of attributes (genes)
- Interesting subgroup describing patterns discovered by SD algorithm

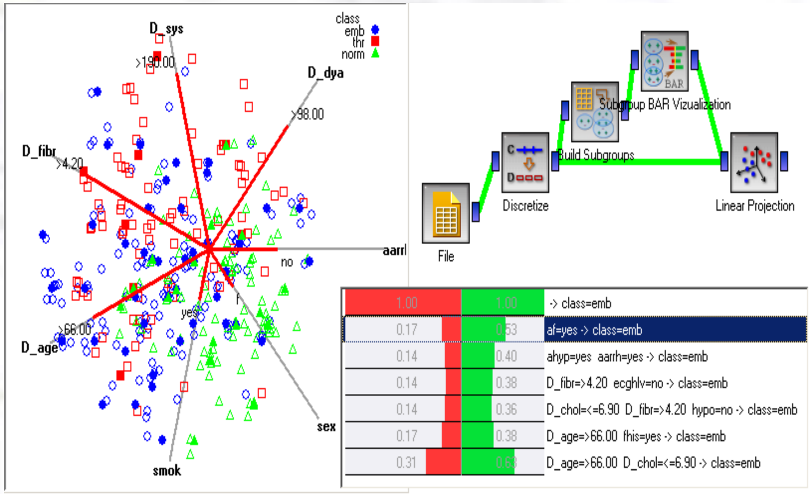
CancerType = Leukemia	
IF	KIAA0128 = DIFF. EXPRESSED
AND	prostoglandin d2 synthase = NOT_ DIFF. EXPRESSED

- Interpretable by biologists

D. Gamberger, N. Lavrač, F. Železný, J. Tolar
Journal of Biomedical Informatics 37(5):269-284, 2004

SD algorithms in the Orange DM Platform

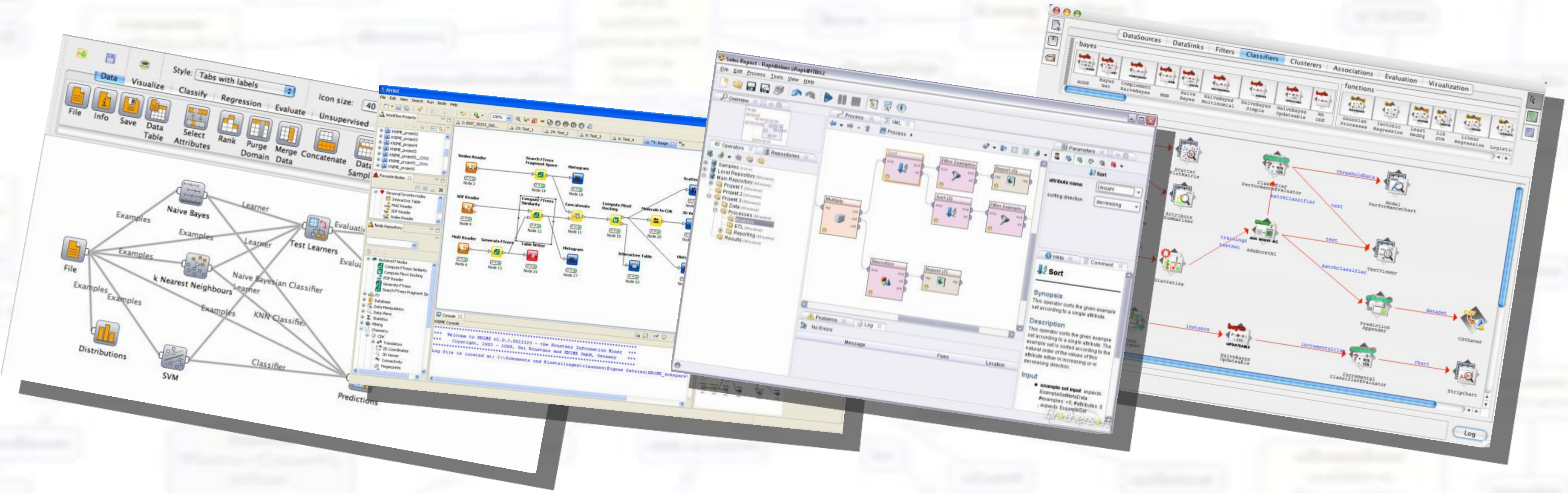
- **Orange** data mining toolkit
  - classification and subgroup discovery algorithms
  - data mining workflows
  - visualization



- **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

- **First Generation Data Mining**
  - Basics of Machine Learning and Data Mining
- **Second Generation Data Mining**
  - Selected Algorithms and Biomedical Applications
- **Third Generation DM Techniques and Platforms**
  - Relational Data Mining
  - Semantic Relational Learning: Using ontologies in DM
- **Current Work and Conclusions**

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
...	...	...	...	...	...	...	...

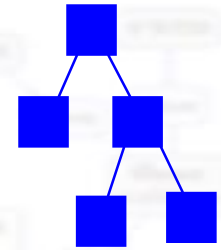
order				
Customer ID	Order ID	Store ID	Delivery Mode	Paymt Mode
...	...	...	...	...
3478	2140267	12	regular	cash
3478	3446778	12	express	check
3478	4728386	17	regular	check
3479	3233444	17	express	credit
3479	3475886	12	regular	credit
...	...	...	...	...

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



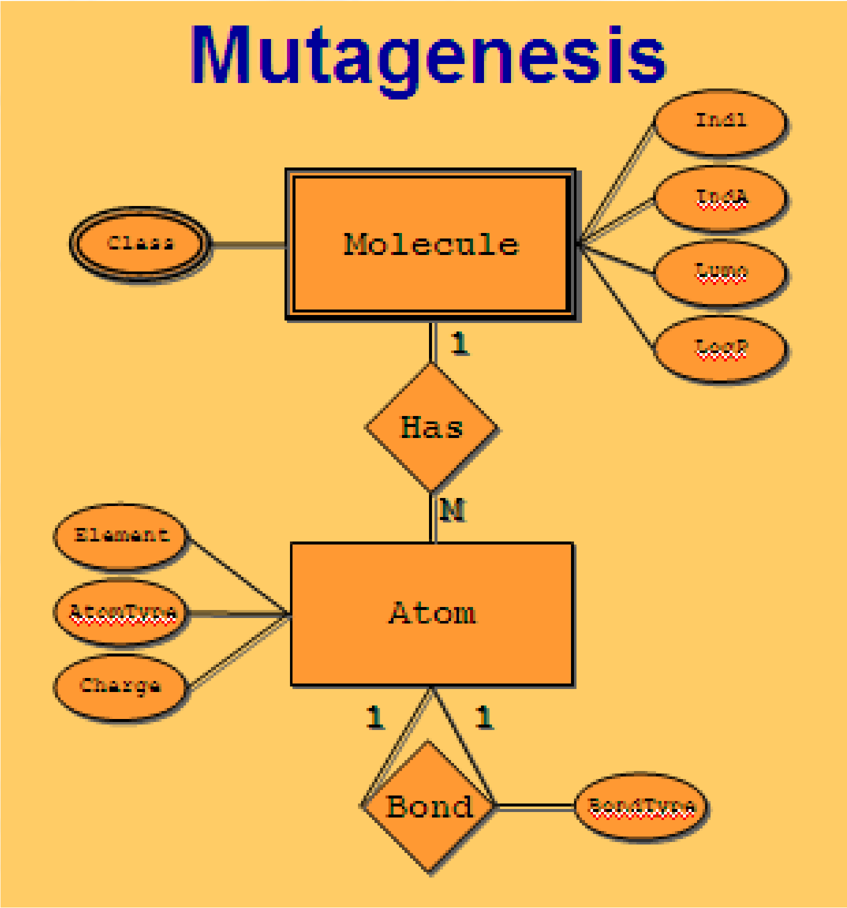
model, patterns,
…

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

Find: a classification model, a set of patterns

Relational Data Mining

- Learning from multiple tables
  - patient records connected with other patient and demographic information
- Complex relational problems:
  - temporal data: time series in medicine, ...
  - structured data: representation of molecules and their properties in protein engineering, biochemistry, ...



Relational Data Mining through Propositionalization

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.

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

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Relational Data Mining through Propositionalization

customer							
ID	Zip	Sex	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.

Step 1

Propositionalization

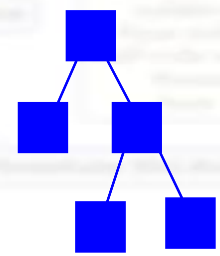
1. constructing relational features
2. constructing a propositional table

	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

	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



Classification model

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

1. constructing relational features
2. constructing a propositional table

	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

	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

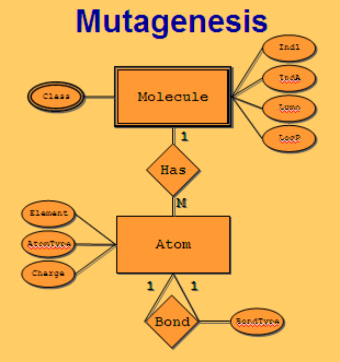
Subgroup discovery

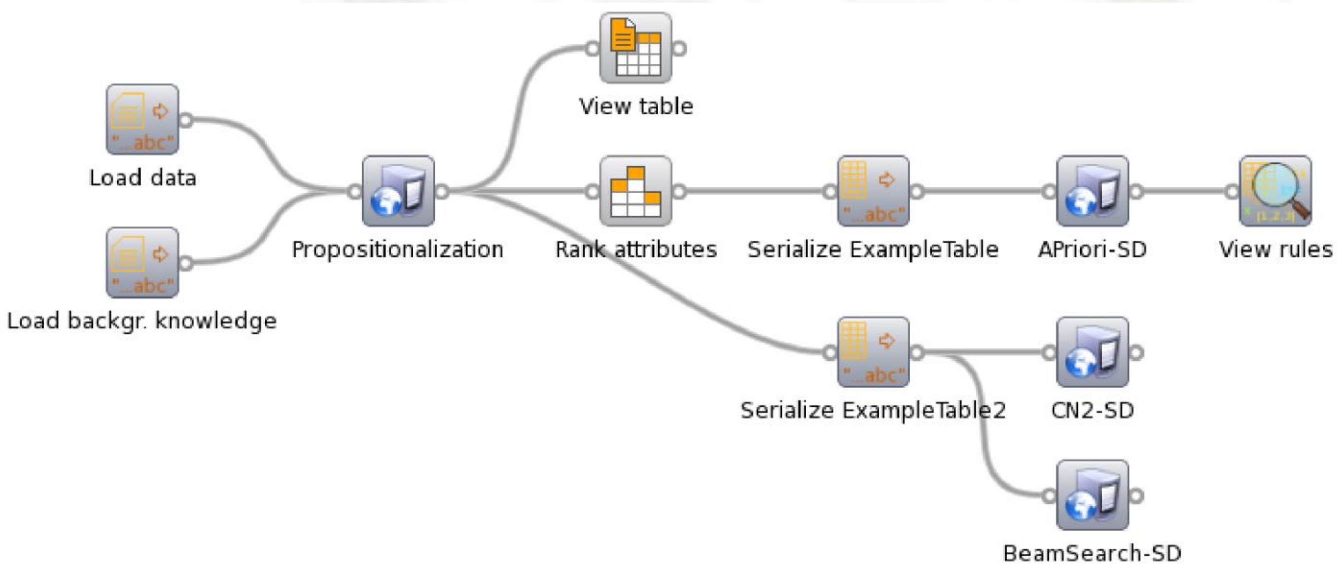
```
target(A) :-
   'Doctor'(A), 'Italy'(A).

target(A) :-
   'Public'(A), 'Gold'(A).

target(A) :-
   'Poland'(A), 'Deposit'(A), 'Gold'(A).

target(A) :-
   'Germany'(A), 'Insurance'(A).

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

patterns (set of rules)

Relational Data Mining in Orange4WS

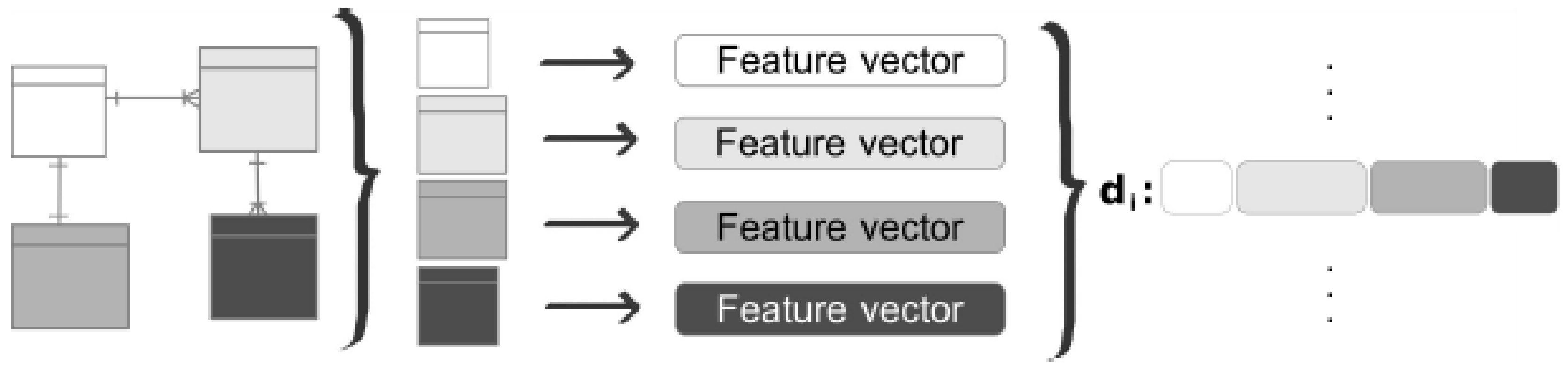
- service for propositionalization through efficient first-order feature construction (Železny and Lavrač, MLJ 2006)
  - f121(M):- hasAtom(M,A), atomType(A,21)
  - f235(M):- lumo(M,Lu), lessThr(Lu,1.21)
- subgroup discovery using CN2-SD
  - mutagenic(M) ← feature121(M), feature235(M)





Wordification approach to RDM

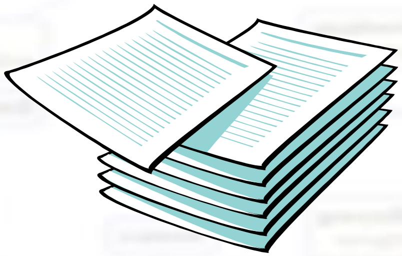
- Transform a relational database into a document corpus
  - For each individual (row) in the main table, concatenate the "words" generated for the main table with the "words" generated for the other tables, linked through external keys



Text mining: Words/terms as binary features

Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

Text mining

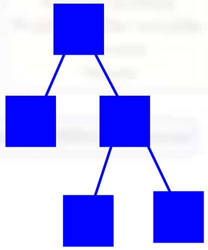


BoW vector construction

Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

Document	Word1	Word2	…	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13	…	…	…	…	…
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23	…	…	…	…	…
d24	0	0	1	0	NO

Data Mining



model, patterns, clusters,

…

Wordification Methodology

- One individual of the main data table in the relational database ~ one text document
- Features (attribute values) ~ the words of this document
- Individual words (called **word-items** or **witems**) are constructed as combinations of:

$$[table\ name]\_[attribute\ name]\_[value]$$

- **n-grams** are constructed to model feature dependencies:

$$[witem_1]\_[witem_2]\_ \ldots \_[witem_n]$$

Wordification Methodology

- Transform a relational database to a document corpus
  - For each individual (row) in the main table, concatenate words generated for the main table with words generated for the other tables, linked through external keys
- Construct BoW vectors with TF-IDF weights on words
  (optional: Perform feature selection)
- Apply text mining or propositional learning on BoW table

Wordification on simplified trains problem

TRAIN

trainID	eastbound
t1	east
…	…
t5	west
…	…

CAR

carID	shape	roof	wheels	train
c11	rectangle	none	2	t1
c12	rectangle	peaked	3	t1
…	…	…	…	…
c51	rectangle	none	2	t5
c52	hexagon	flat	2	t5
…	…	…	…	…

t1: [car_roof_none, car_shape_rectangle, car_wheels_2, car_roof_none__car_shape_rectangle, car_roof_none__car_wheels_2, car_shape_rectangle__car_wheels_2, car_roof_peaked, car_shape_rectangle, car_wheels_3, car_roof_peaked__car_shape_rectangle, car_roof_peaked__car_wheels_3, car_shape_rectangle__car_wheels_3], **east**

Wordification

t1: [car_roof_none, car_shape_rectangle, car_wheels_2, car_roof_none__car_shape_rectangle, car_roof_none__car_wheels_2, car_shape_rectangle__car_wheels_2, car_roof_peaked, car_shape_rectangle, car_wheels_3, car_roof_peaked__car_shape_rectangle, car_roof_peaked__car_wheels_3, car_shape_rectangle__car_wheels_3], **east**

t5: [car_roof_none, car_shape_rectangle, car_wheels_2, car_roof_none__car_shape_rectangle, car_roof_none__car_wheels_2, car_shape_rectangle__car_wheels_2, car_roof_flat, car_shape_hexagon, car_wheels_2, car_roof_flat__car_shape_hexagon, car_roof_flat__car_wheels_2, car_shape_hexagon__car_wheels_2], **west**

TF-IDF calculation for BoW vector construction:

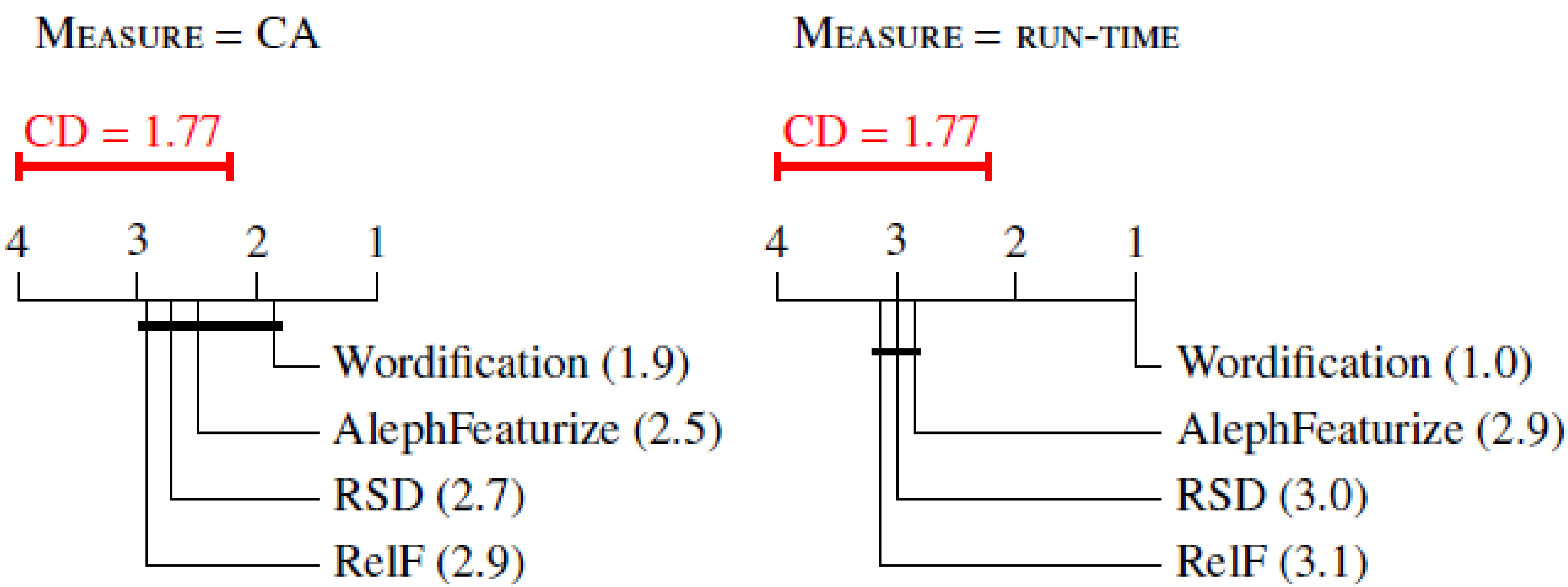
	car_shape _rectangle	car_roof _peaked	car_wheels_3	car_roof_peaked__ car_shape_rectangle	car_shape_rectangle __car_wheels_3	...	class
t1	0.000	0.693	0.693	0.693	0.693	...	east
...	...	...	...	...	...	...	...
t5	0.000	0.000	0.000	0.000	0.000	...	west
...	...	...	...	...	...	...	...

TF-IDF weights

- No explicit use of existential variables in features, TF-IDF instead
- The weight of a word indicates how relevant is the feature for the given individual
- The TF-IDF weights can then be used either for filtering words with low importance or for using them directly by a propositional learner (e.g. J48)

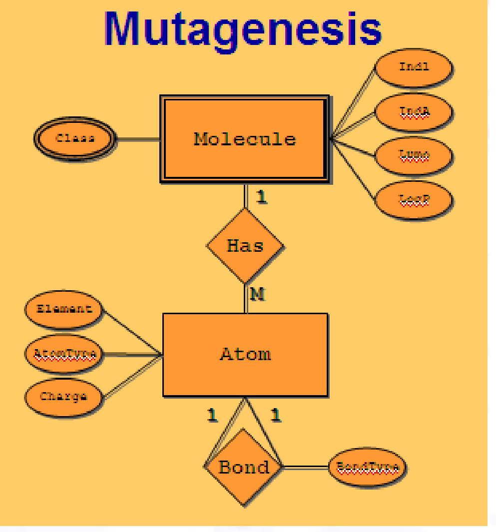
Evaluation of propositionalization approaches in relational classification tasks

- Cross-validation experiments on 8 relational datasets: Trains (in two variants), Carcinogenesis, Mutagenensis with 42 and 188 examples, IMDB, and Financial.
- Results (using J48 for propositional learning)



Semantic Relational Learning

- **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
...	...	...	...	...	...	...	...

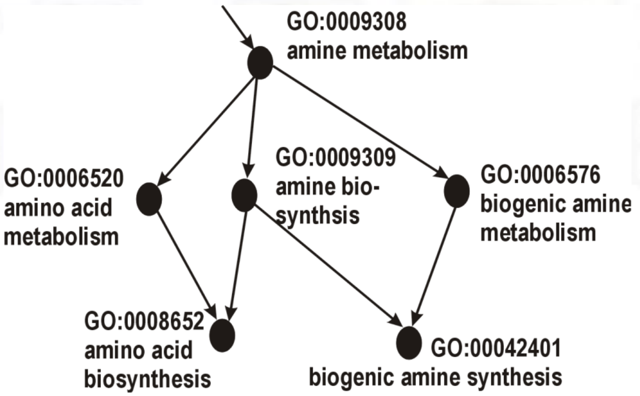
order

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.



Talk outline

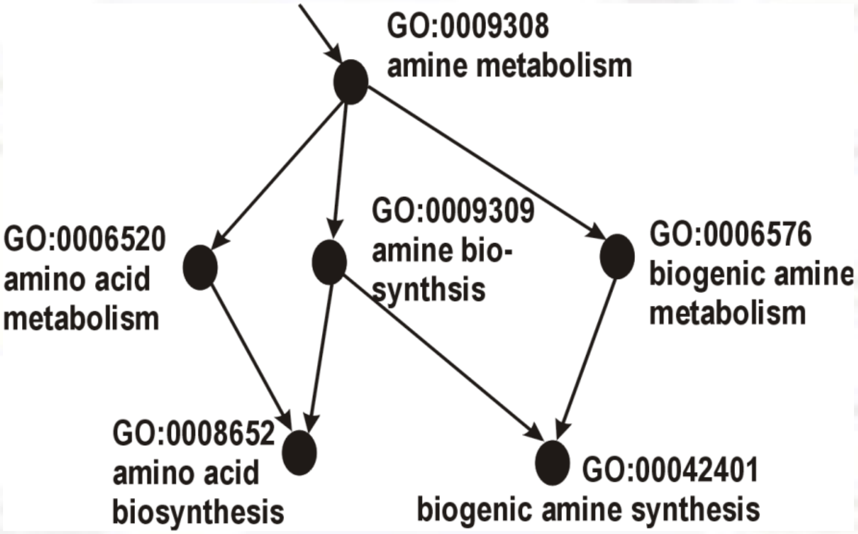
- **First Generation Data Mining**
  - Basics of Machine Learning and Data Mining
- **Second Generation Data Mining**
  - Selected Algorithms and Biomedical Applications
- **Third Generation DM Techniques and Platforms**
  - Relational Data Mining
  - Semantic Relational Learning: Using ontologies in DM
- **Current Work and Conclusions**

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Semantic Relational Learning: Using domain ontologies in DM

Using domain ontologies as background knowledge, e.g., using the Gene Ontology (GO)

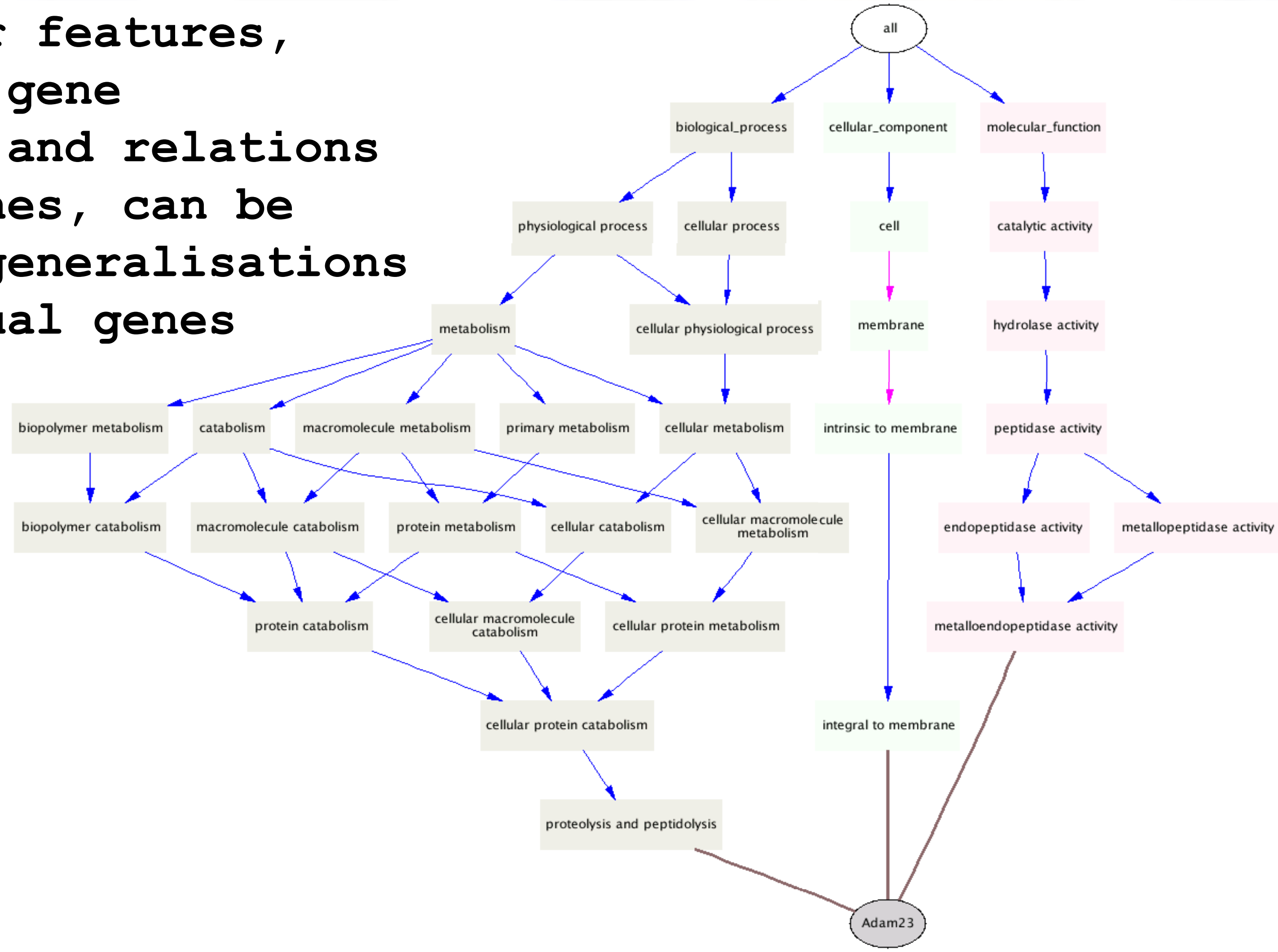
- GO is a database of terms, describing gene sets in terms of their
  - functions (12,093)
  - processes (1,812)
  - components (7,459)
- Genes are annotated to GO terms
- Terms are connected (is_a, part_of)
- Levels represent terms generality



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Using GO as background knowledge in DNA microarray data analysis

First-order features, describing gene properties and relations between genes, can be viewed as generalisations of individual genes



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Propositionlization approach to Semantic data mining

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

Step 2: Automatically generate generalized relational features

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

Step 3: Propositionalization - determine truth values of features

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

Step 4: Learn rules by a subgroup discovery algorithm 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

Over-expressed

IF

f2 and f3

[4,0]

Semantic Relational Learning in Orange4WS

- A special purpose Semantic Data Mining algorithm SEGS
  - discovers interesting gene group descriptions as conjunctions of ontology concepts from GO, KEGG and Entrez
  - integrates public gene annotation data through relational features
  - SEGS algorithm (Trajkovski, Železny, Lavrač and Tolar, JBI 2008) is available in Orange4WS

Third Generation Data Mining Platforms

Should be …

- cloud-based
- service oriented (DM algorithms as web services)
- web-based
- enable simple construction of web services from available algorithms

New platform: **ClowdFlows** (Kranjc et al.,2012)

- **Is cloud-based, service-oriented, on the web**

http://clowdflows.org

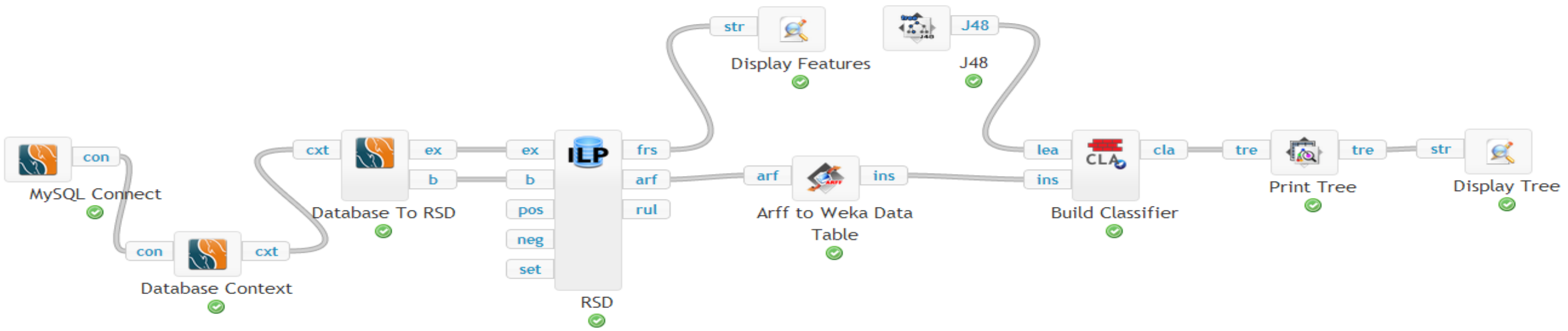
ClowdFlows platform

- Allows for simple creation and execution of complex DM procedures
  - Algorithms are web services (in the “cloud”)
  - No installation of the platform is needed
  - Workflows are available to everyone

  through any browser with a simple web click

  e.g. propositionalization at

  http://clowdflows.org/workflow/611/

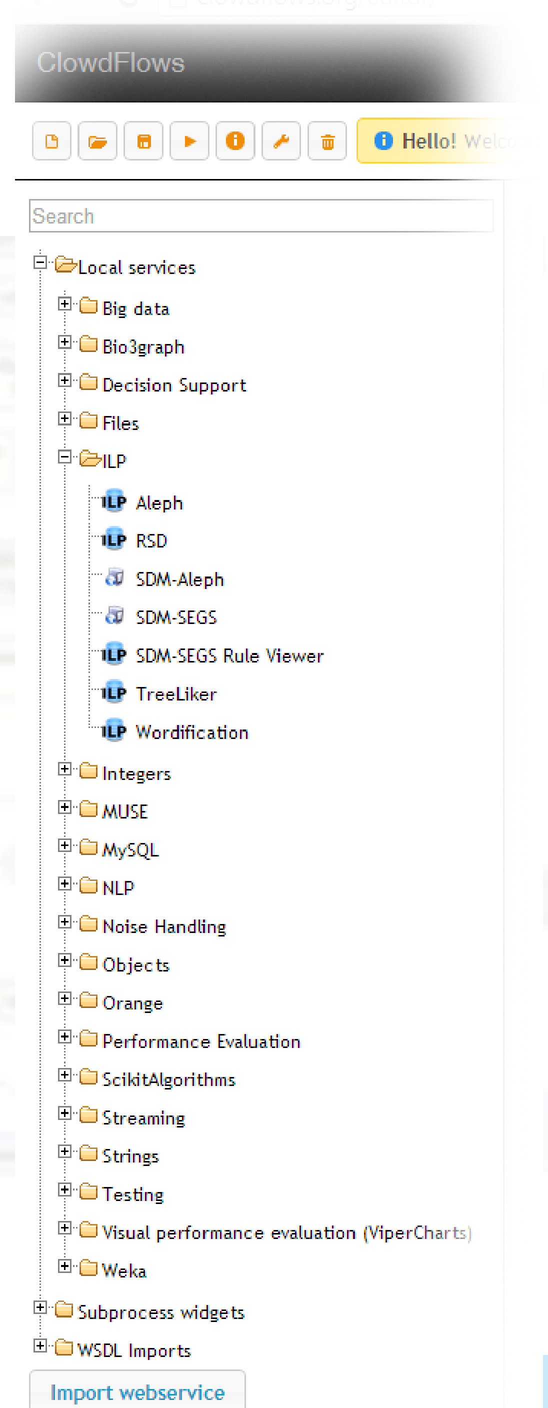


DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

ClowdFlows platform

- **Large repository of algorithms**
  - Includes relational and semantic data mining algorithms
  - All algorithms from the Orange platform
  - WEKA algorithms as web services
  - Big data analysis

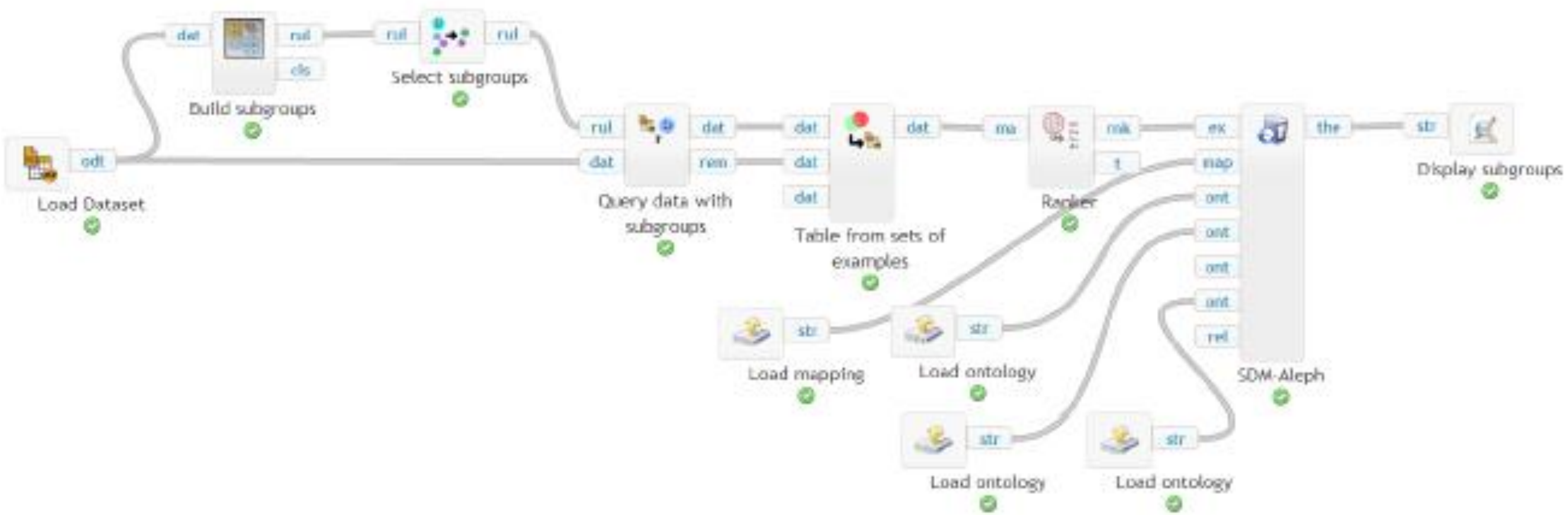
    Kranjc et al., Inf. Proc. and Manag. 2014
  - Text analysis
  - Social network analysis
  - Analysis of data streams
- **Large repository of workflows**
  - access to our JSI technology heritage



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Example: Semantic data mining

- Discovering interesting subgroups in data and their biological explanation with ontologies (Vavpetič et al., JIIS 2014)

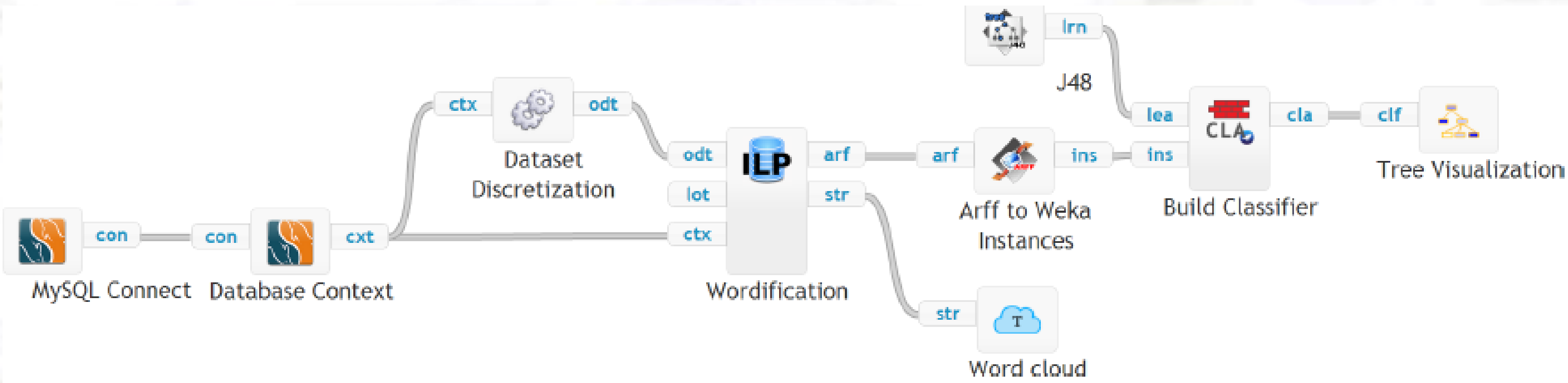


http://clowdflows.org/workflow/910/

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Wordification implemented in ClowdFlows

- Propositionalization through wordification, available at http://clowdflows.org/workflow/1455/



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Challenge addressed in recent work

The challenge is to fill the current gap between semantic web and data science: Which part of the semantic web is most important to my current interests?

Semantic Data Mining

+ Finds complex rules
+ Higly informative
- Computationally demanding
- Complexity grows exponentially

Fast
Scalable
Informative

Network analysis

+ Can process massive data
+ Fast, easy to calculate
- Less informative results

Challenge addressed in recent work

New challenge and methodology

- Take a large knowledge graph such as BioMine, or a Linked Open Data resource, such as Bio2RDF
  (to be addressed in our joint work with Dumontier)

- Use Semantic data mining (SDM) to mine experimental data with ontologies as background knowledge to get explanations for groups of TargetClass objects, e.g.
  BreastCancer ← chromosome AND cell cycle

- Reduce the complexity of learning in a huge search space of ontology terms by network analysis-based node filtering

(Kralj et al., LPNMR 2018, JMLR 2019)

Network analysis for feature selection

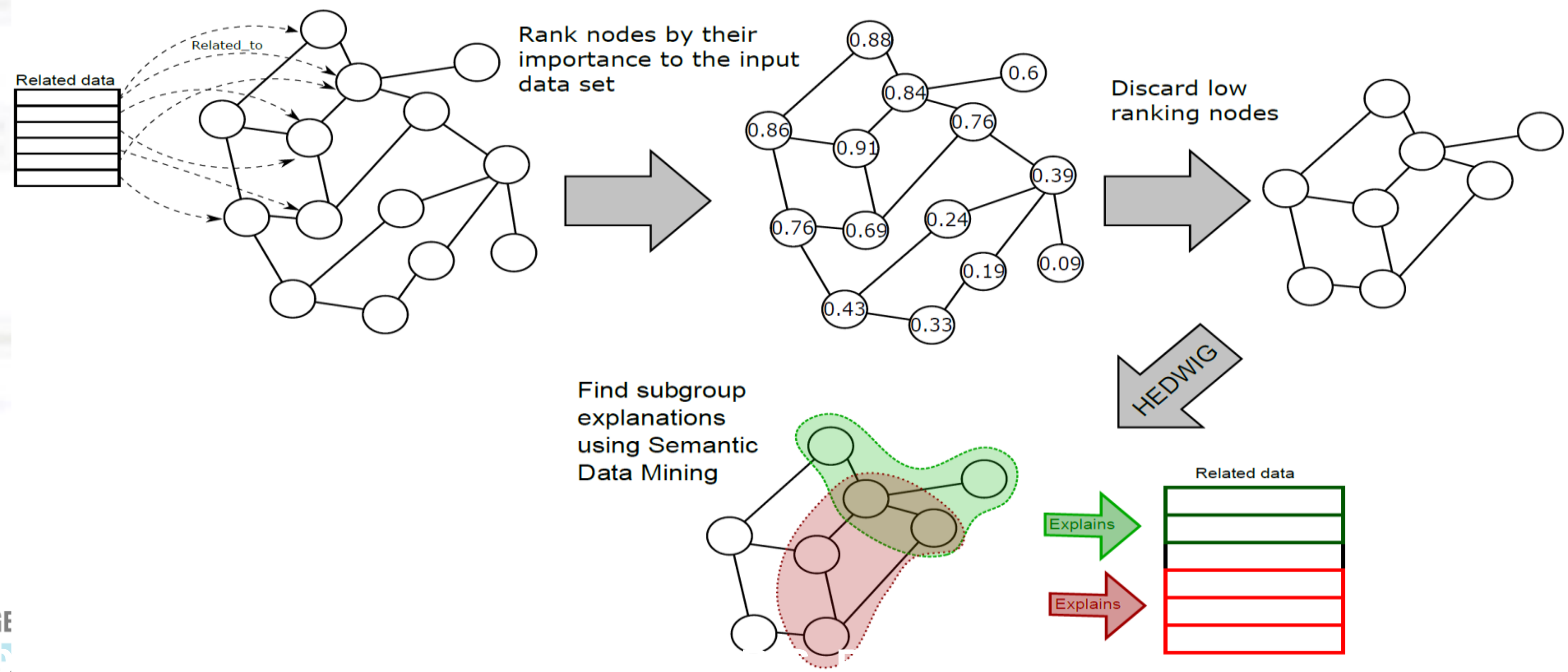
Which part of a given knowledge graph is most relevant for the given experimental data ?

Semantic relational learning	?	Network analysis
SDM algorithm Hedwig + Finds complex rules + Higly informative - Computationally demanding - Complexity grows exponentially	Fast Scalable Informative	+ Can process massive data + Fast, easy to calculate - Less informative results

(Kralj et al., LPNMR 2016, JMLR 2019)

Network analysis for feature reduction: NetSDM (Kralj et al. 2019)

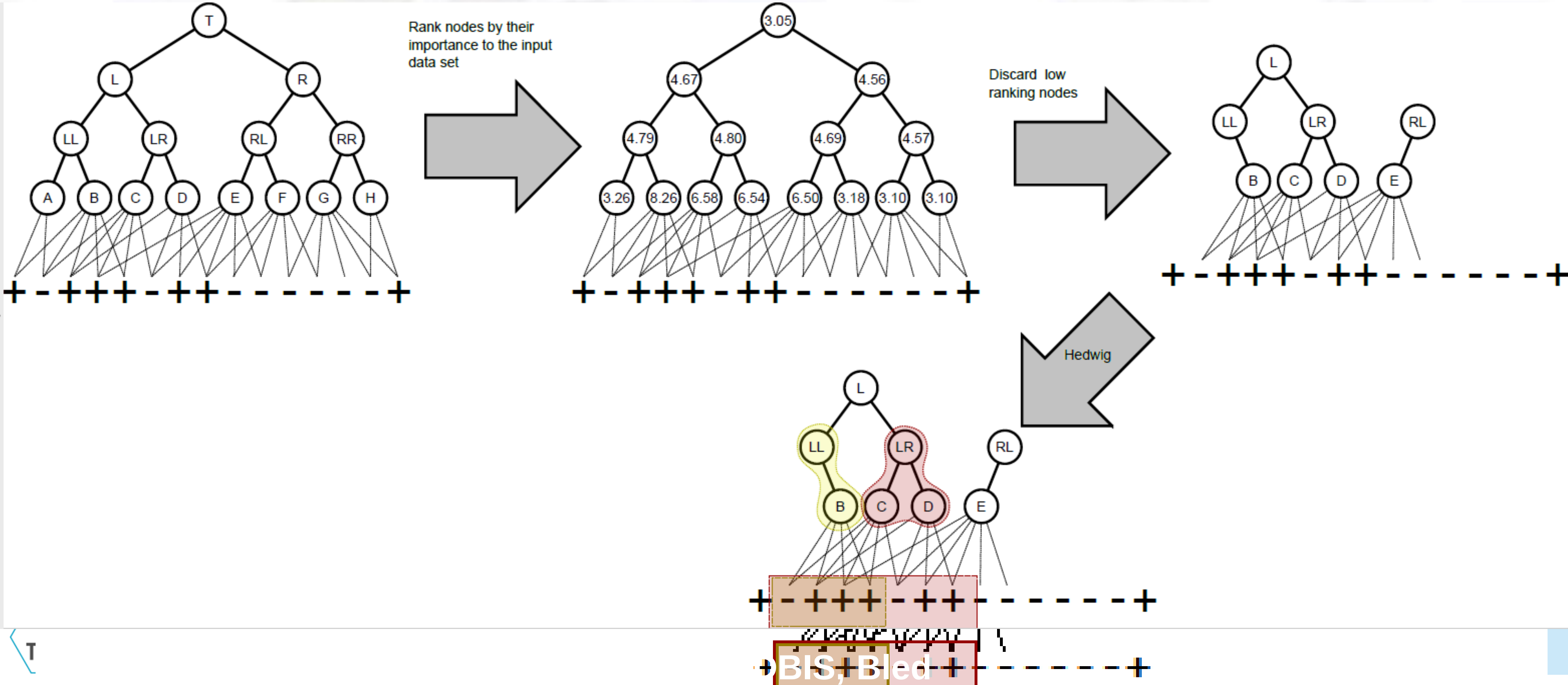
- Use network analysis (Personalized PageRank) to estimate the importance of features (e.g. ontology terms)
- Reduce the complexity of the huge search space of ontology terms by network analysis based term filtering
- Same accuracy, up to 100% speed up



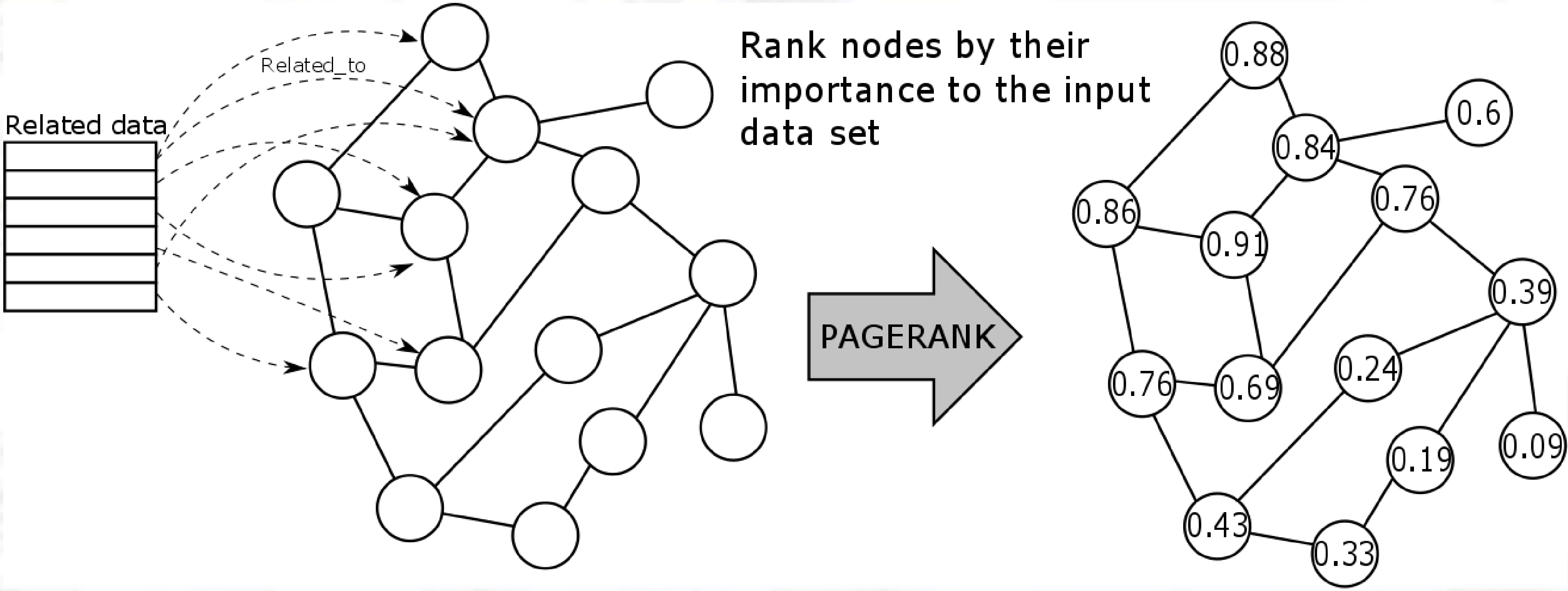
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

NetSDM algorithm outline

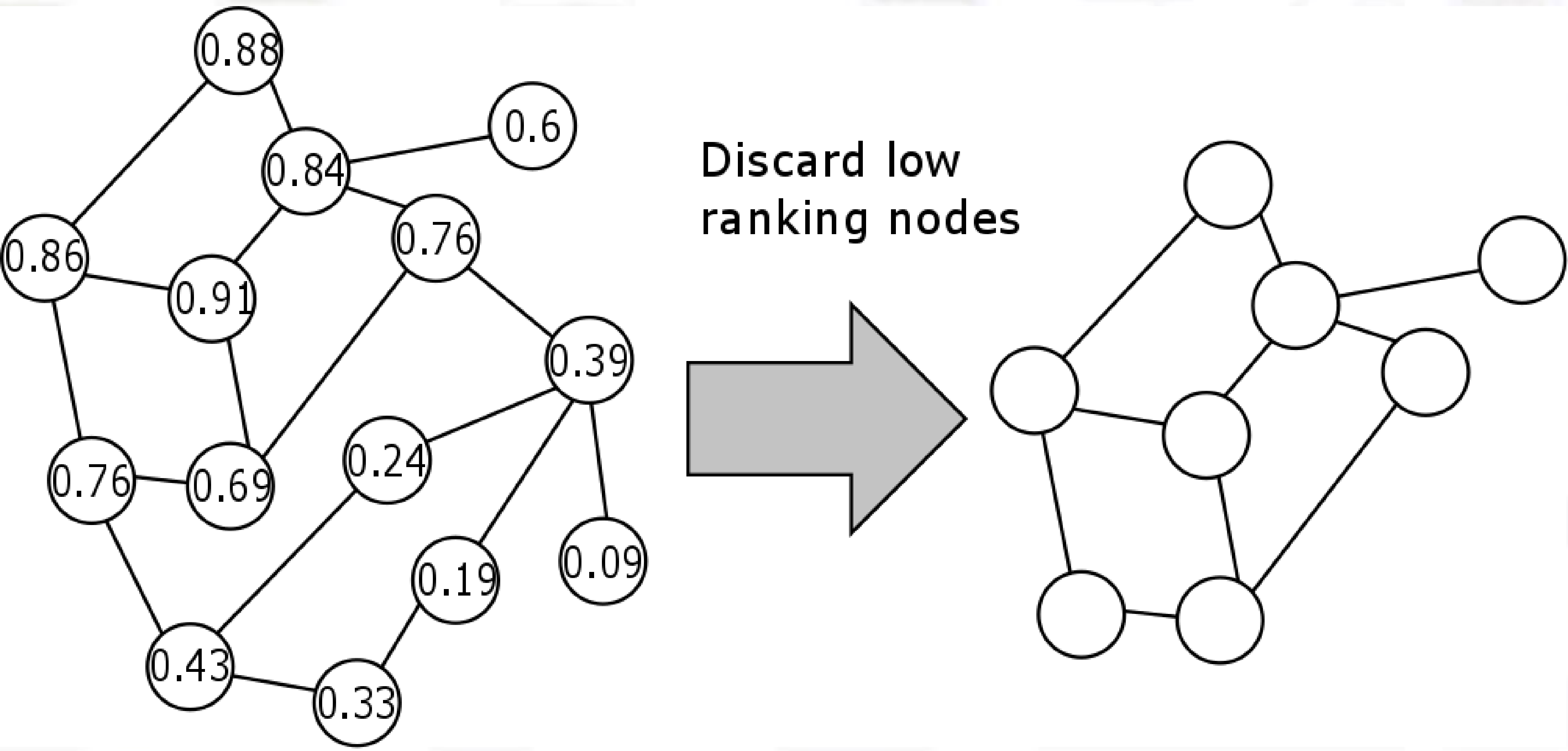
1. Estimate ontology term relevance
2. Delete terms with low relevance
3. Run semantic relational learning algorithm Hedwig on pruned ontolgy



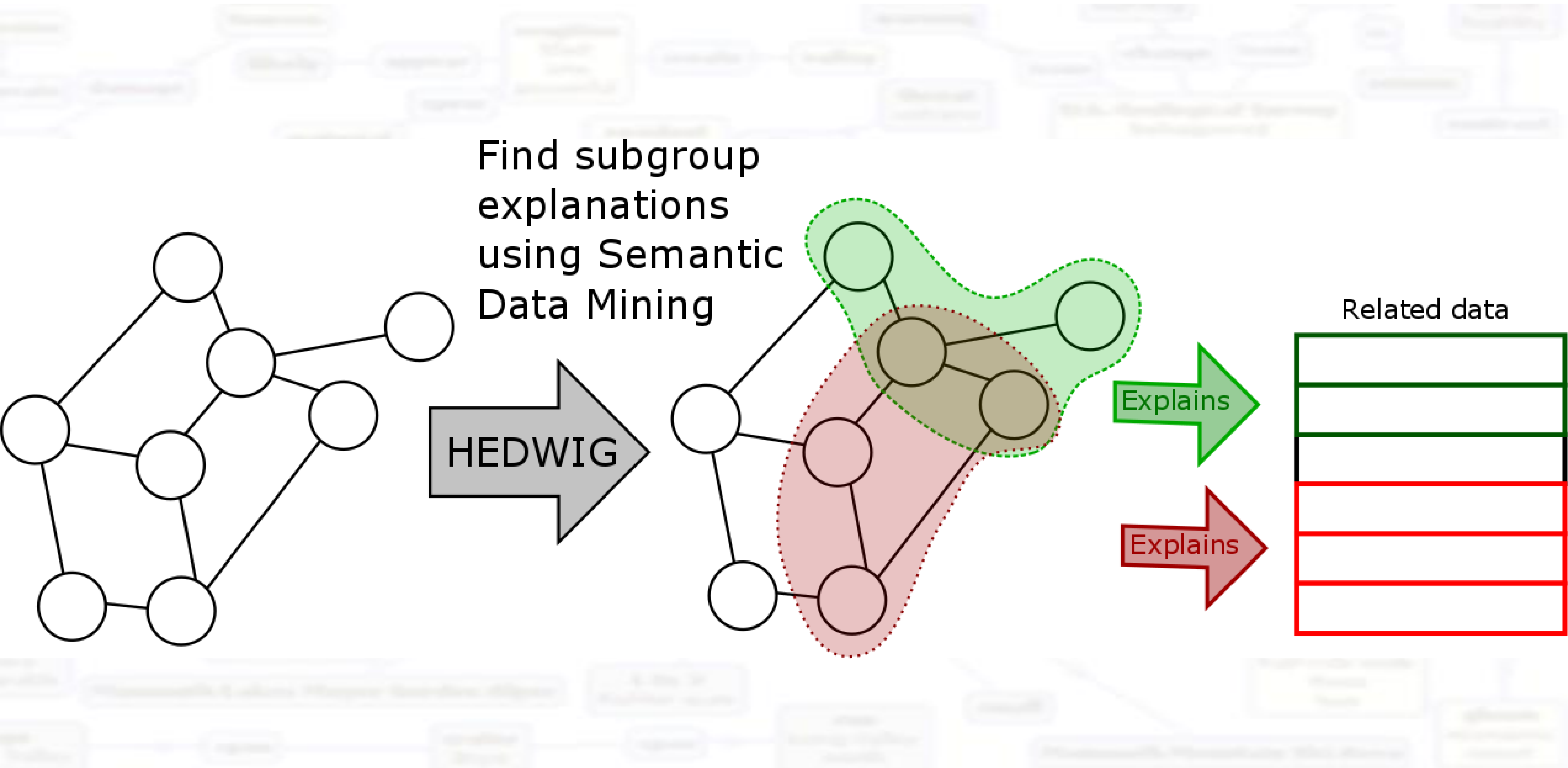
Step 1


Related data
Related_to
Rank nodes by their importance to the input data set
PAGERANK
0.88
0.6
0.84
0.86
0.76
0.91
0.39
0.76
0.69
0.24
0.09
0.19
0.43
0.33

Step 2

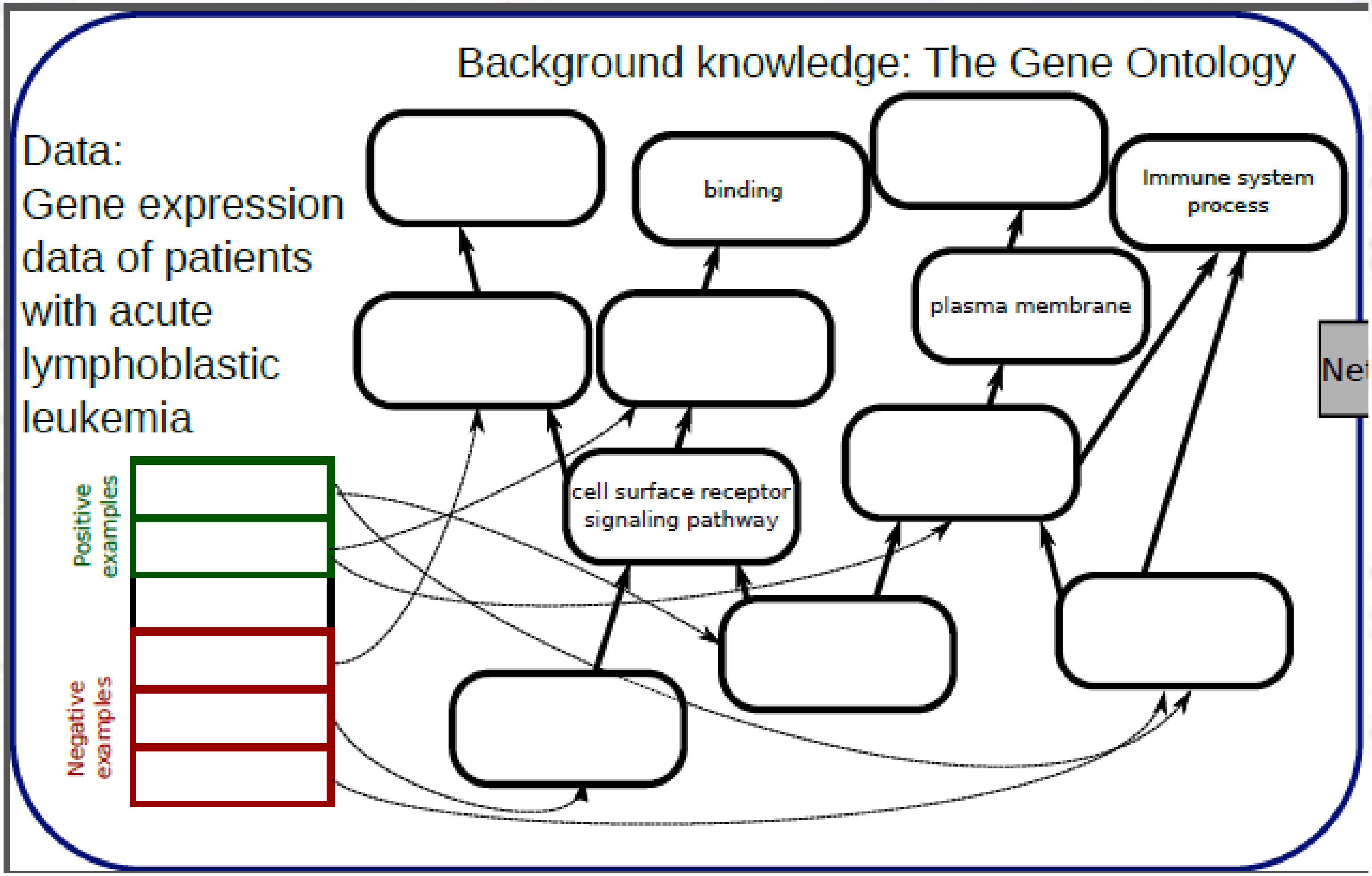

0.88
0.6
0.84
0.76
0.86
0.91
0.39
0.24
0.76
0.69
0.19
0.09
0.43
0.33
Discard low
ranking nodes

Step 3


Find subgroup explanations using Semantic Data Mining
HEDWIG
Explains
Explains
Related data

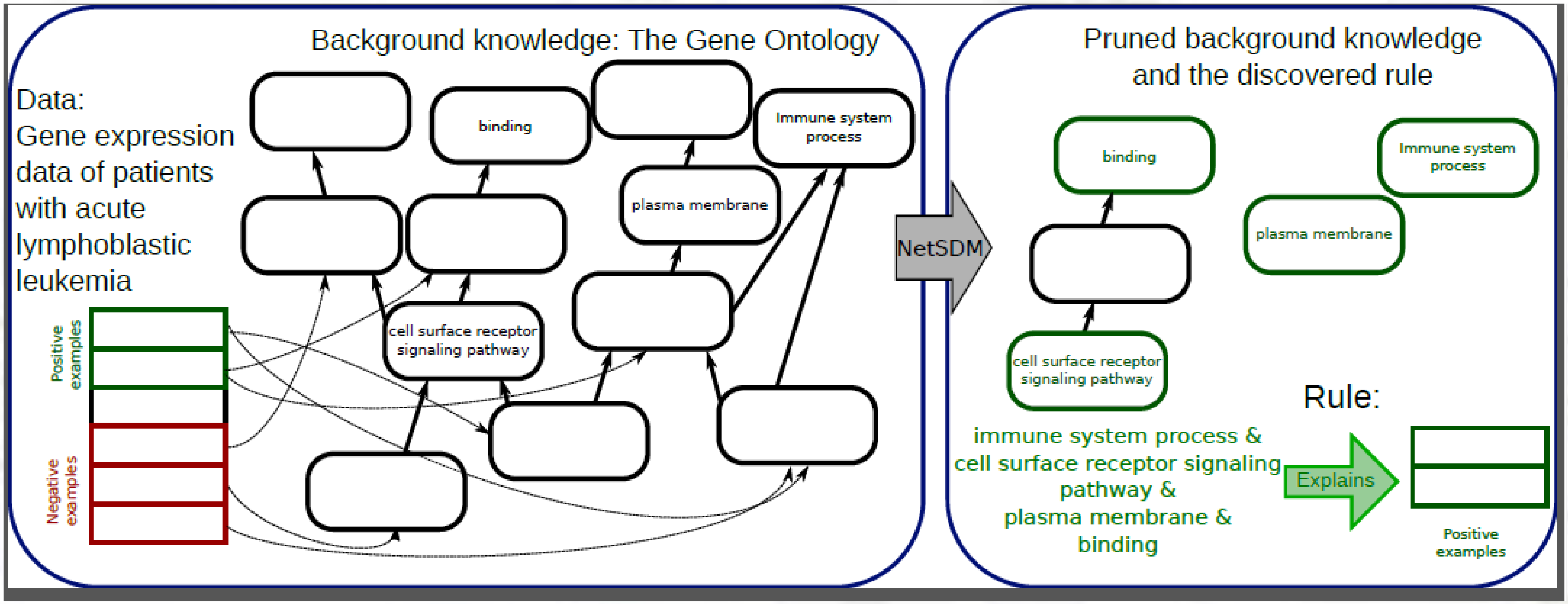
Example: Analysis of ALL data using Gene Ontology

Input to NetSDM:



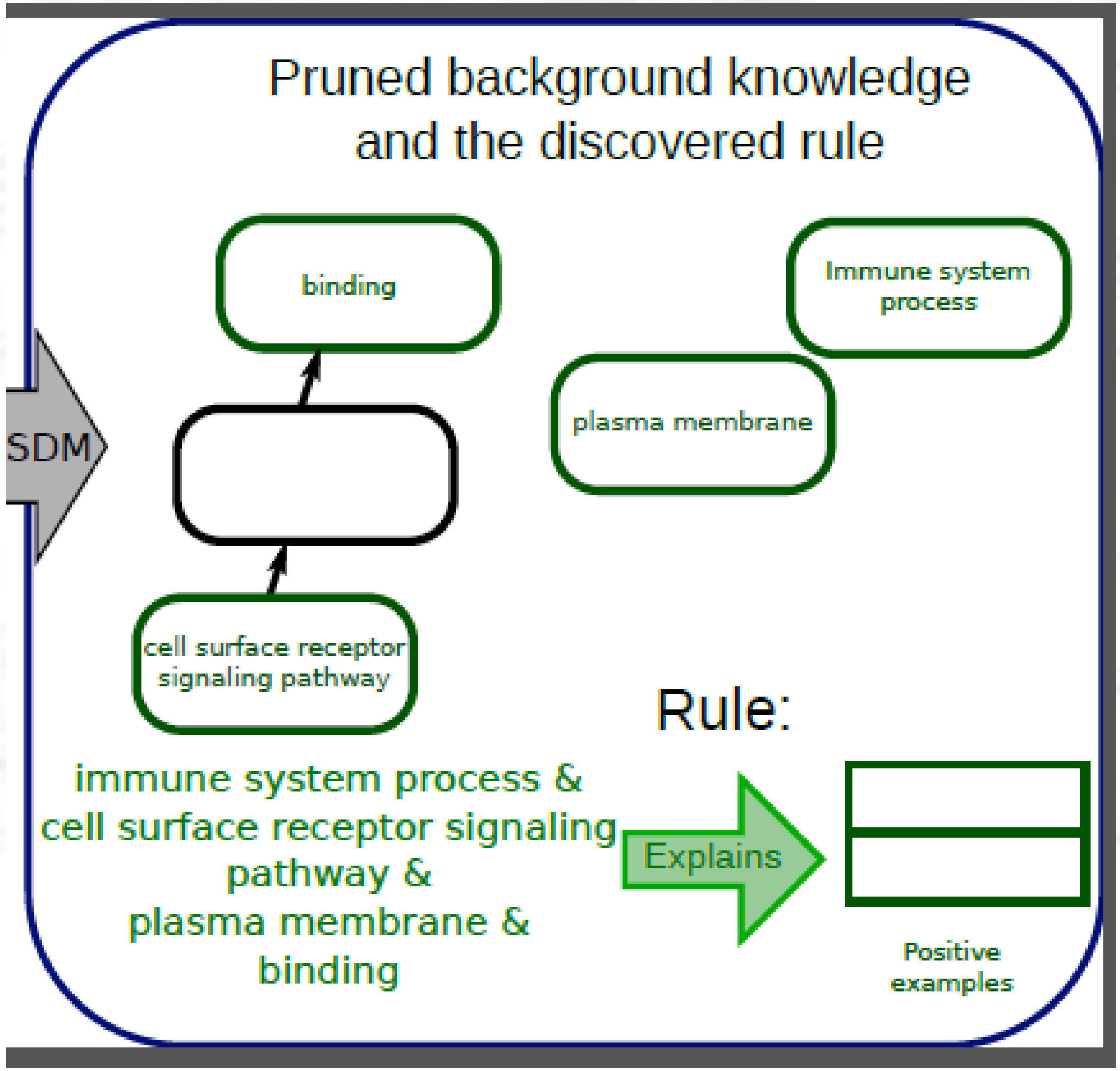
Example: Analysis of ALL data using Gene Ontology

NetSDM:



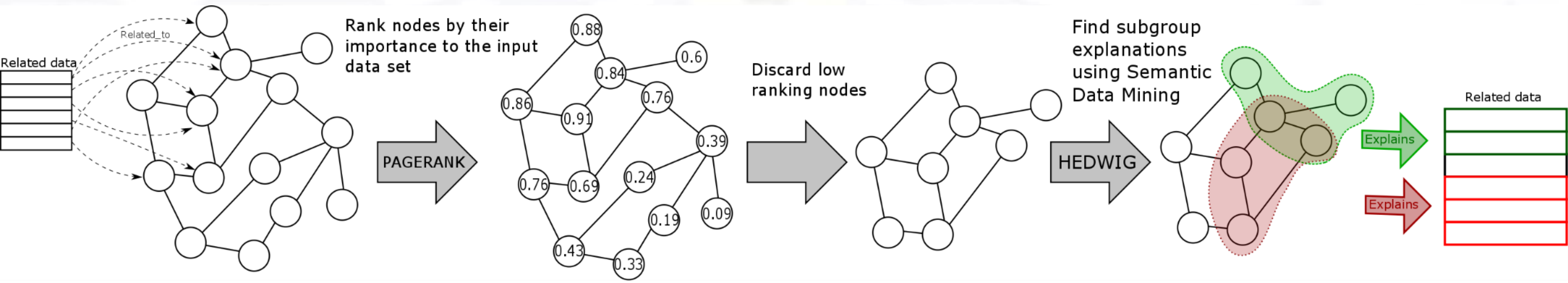
Example: Analysis of ALL data using Gene Ontology

Output of NetSDM:



Results

- Personalized PageRank can be effectively used to decrease the size of the search space of Semantic Relational Learning algorithms
- Accuracy did not decrease even when significantly decreasing the size of the background knowledge to less than 5%.
- Time, taken to discover rules on pruned background knowledge, is shorted by a factor of 100



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Summary and conclusions

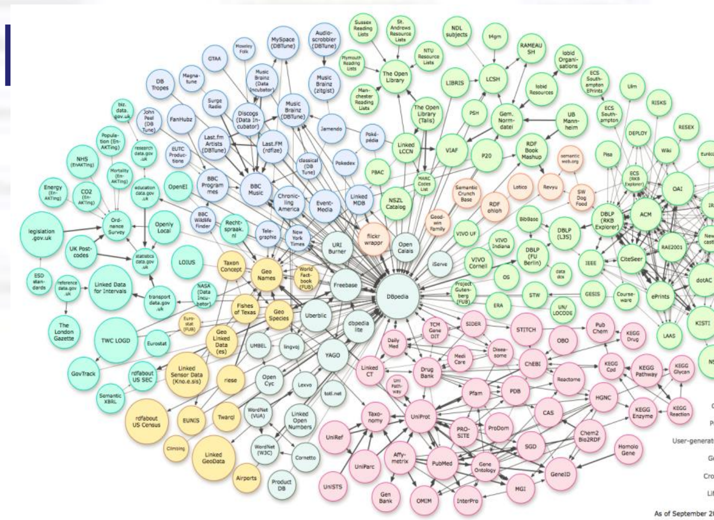
- The presented approaches
  - Can be effectively used for relational and semantic data mining, but are only applicable to individual centered representations (1-to-many, not many-to-many relations)
  - Can be used for **structured data flattening**, as **data preprocessing** step for modern DM, e.g. deep learning
  - A **wordification approach** to propositionalization is especially powerful (Perovšek et al. 2016), can be used as a data fusion mechanism when mining **heterogeneous information networks** (Grčar et al. 2014)
  - **Network analysis** can be used as a mechanism for **feature reduction**

…. all these being implemented and made publicly reusable as

complex workflows in ClowdFlows

Paradigm shift in Semantic Data Mining: Mining Linked Open Data

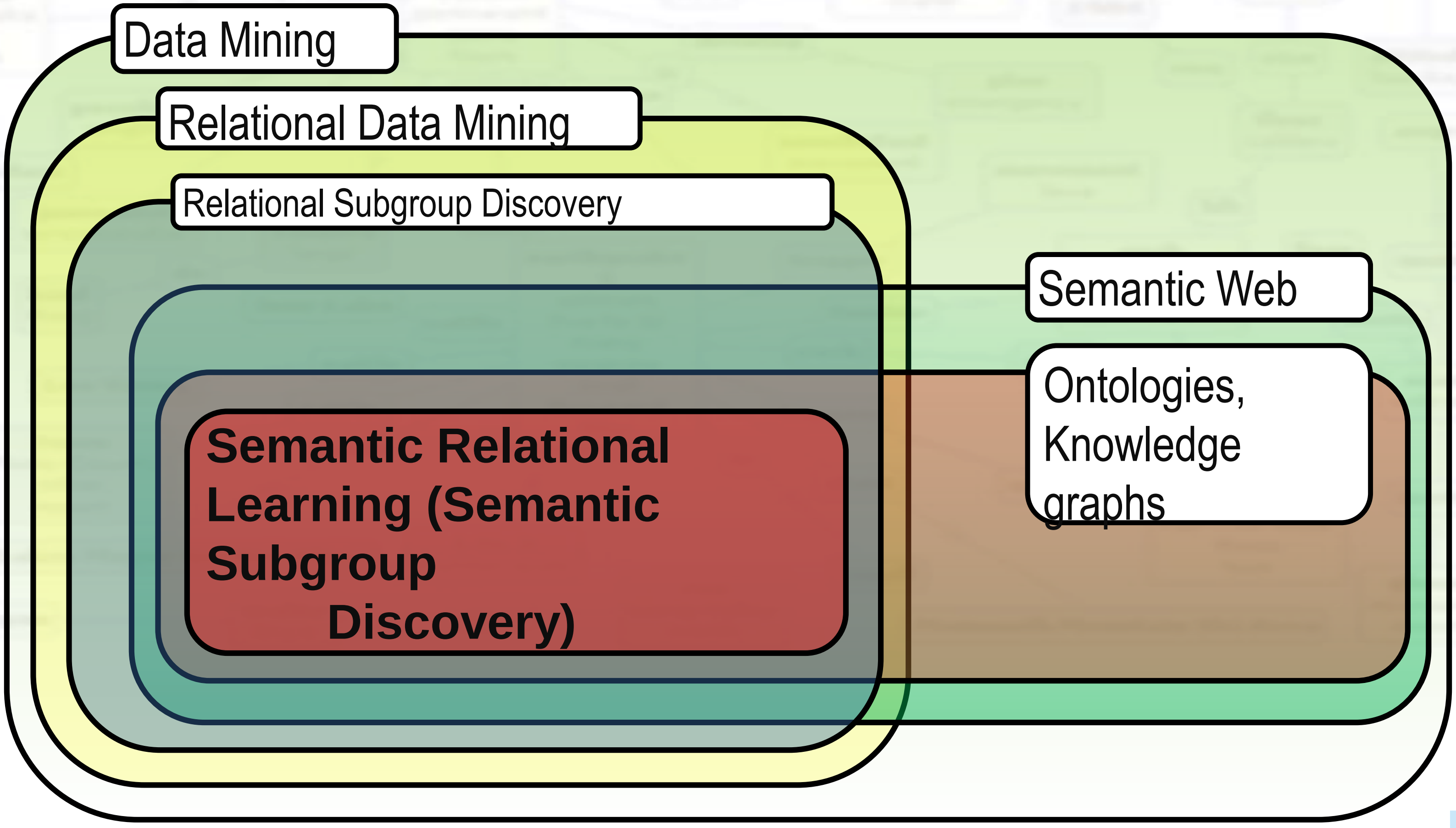
- We envision a paradigm shift from data mining (mining of empirical data) in standard data mining platforms to **knowledge mining on the web**
  - mining knowledge encoded in knowledge graphs,
  - constrained by annotated (empirical) data collections
- Results of Kralj et al. show to be p
  Linked Open Data



- future work is planned, using
  Bio2RDF (with M. Dumontier)
- Combining with embedding
  technology (project EMBEDDIA)

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Summary and conclusion: Semantic Relational Learning in context



DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute