Semantic Relational Learning

Nada Lavrač

Jožef Stefan Institute
Ljubljana, Slovenia

Talk overview

- Advances in Relational Learning
  - Background: Machine Learning (ML)
  - Relational Learning (RL)
  - Semantic Relational Learning (SRL)

- Advances in Network Analysis for SRL

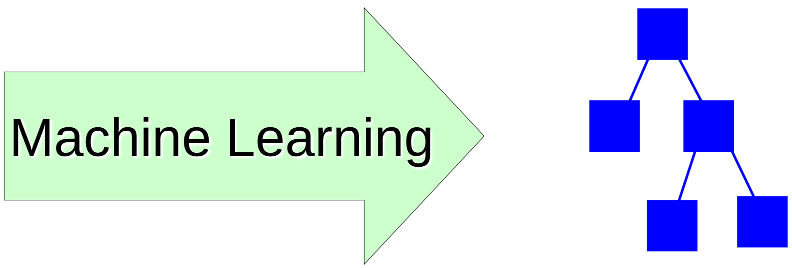
- Conclusions and future work

Background: Machine Learning

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

knowledge discovery
from data



model, patterns, …

Given: transaction data table, a relational database, …

Find: a classification model, a set of interesting patterns

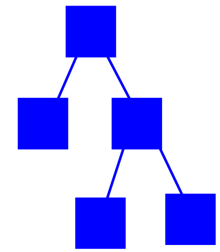
Background: Machine Learning

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

knowledge discovery
from data

Data Mining

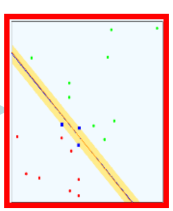


model, patterns, …

Given: transaction data table, a set of text documents, …

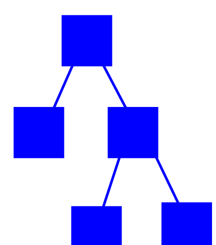
Find: a classification model, a set of interesting patterns

new unclassified instance



classified instance

black box classifier
no explanation



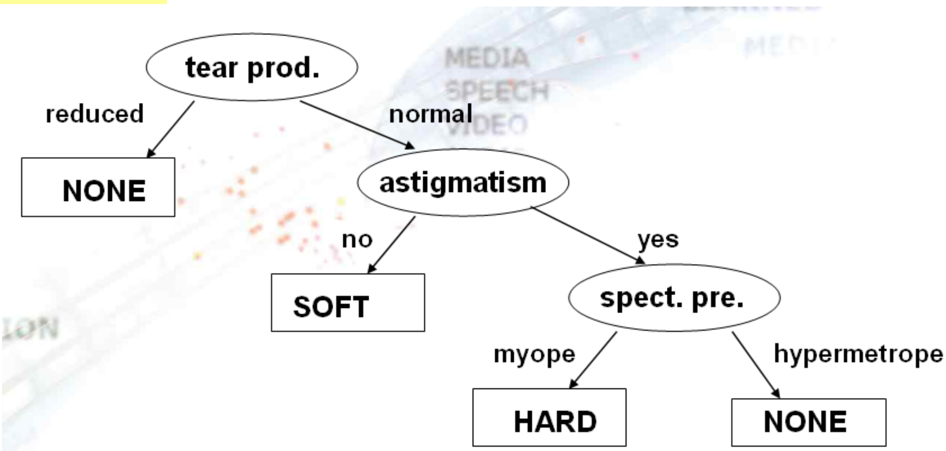
symbolic model
symbolic patterns

explanation

Learning a decision tree classifier

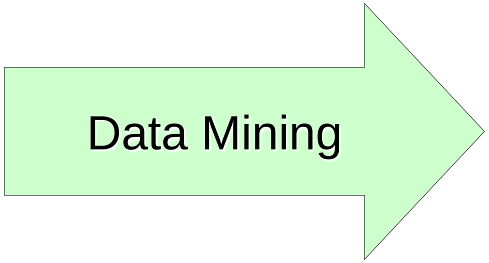
Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13	...	...	...	...	...
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	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 classification rules

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



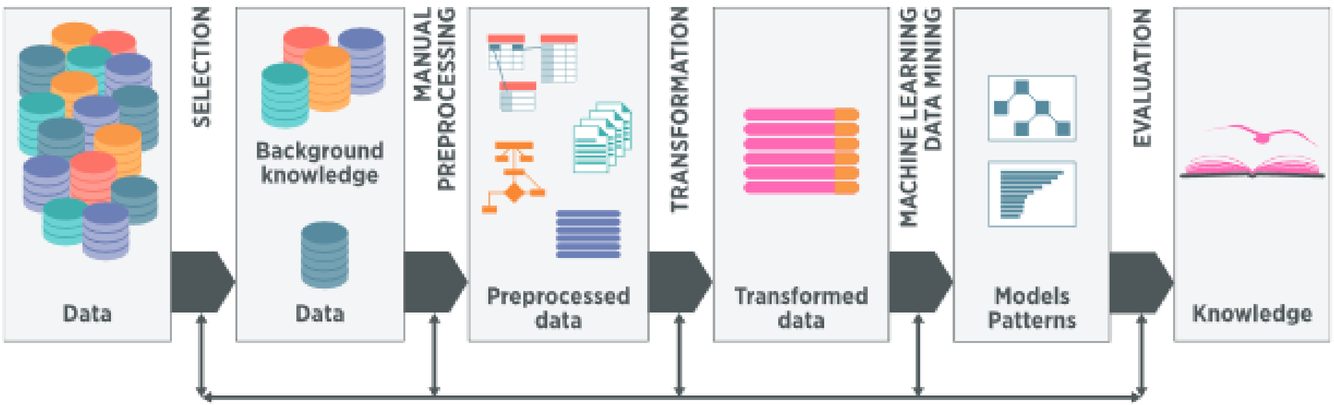
lenses=NONE ← tear production=reduced
lenses=NONE ← tear production=normal AND astigmatism=yes AND spect. pre.=hypermetrope
lenses=SOFT ← tear production=normal AND astigmatism=no
lenses=HARD ← tear production=normal AND astigmatism=yes AND spect. pre.=myope
lenses=NONE ←

First Generation Machine Learning

- **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

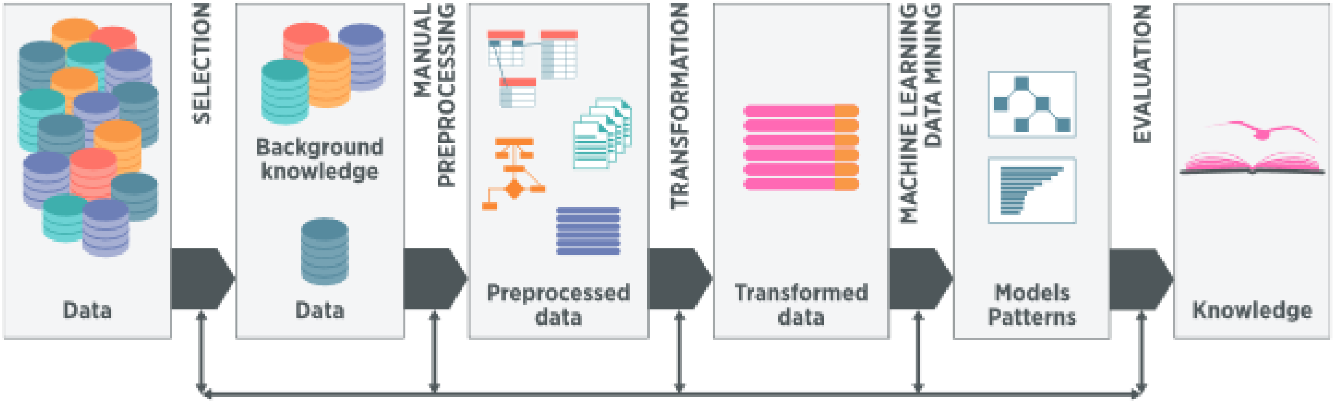
Second Generation Machine Learning

- **Developed since 1990s:**
  - Focused on data mining tasks characterized by large datasets described by large numbers of attributes
  - KDD process:



Second Generation Machine Learning

- **Developed since 1990s:**
  - Focused on data mining tasks characterized by large datasets described by large numbers of attributes
  - KDD process:

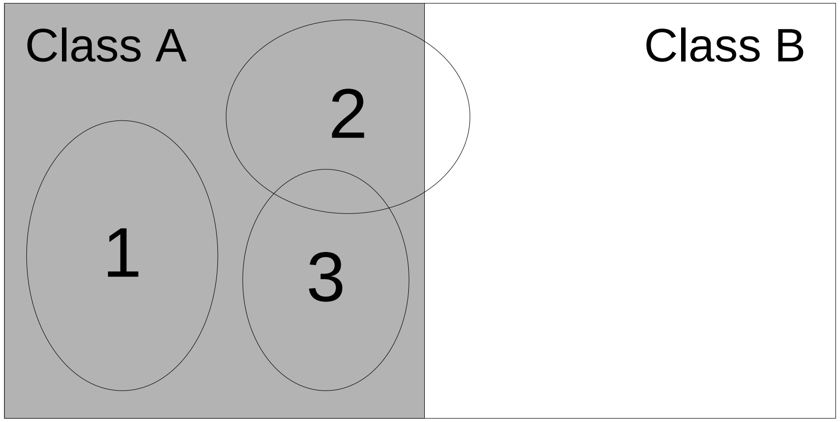


  - New learning tasks and efficient learning algorithms:
    - Learning predictive models: Bayesian network learning, SVMs, relational learning, …
    - Learning descriptive patterns: association rule learning, subgroup discovery, …

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)

Induced subgroups and their statistical chatacterization

Subgroup A2 for female patients:

high-CHD-risk ← female AND bodymassIndex > 25
AND age > 63

Supporting characteristics (computed using ℵ2 statistical significance test): positive family history and hypertension. Women in this risk group typically have slightly increased LDL cholesterol values and normal but decreased HDL cholesterol values.

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

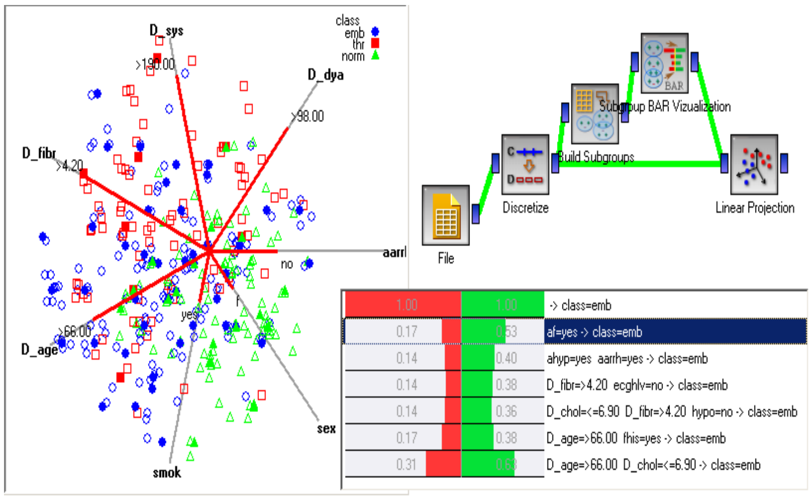
- Interpretable by biologists

D. Gamberger, N. Lavrač, F. Železný, J. Tolar

Journal of Biomedical Informatics 37(5):269-284, 2004

SD algorithms in Orange DM Platform

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

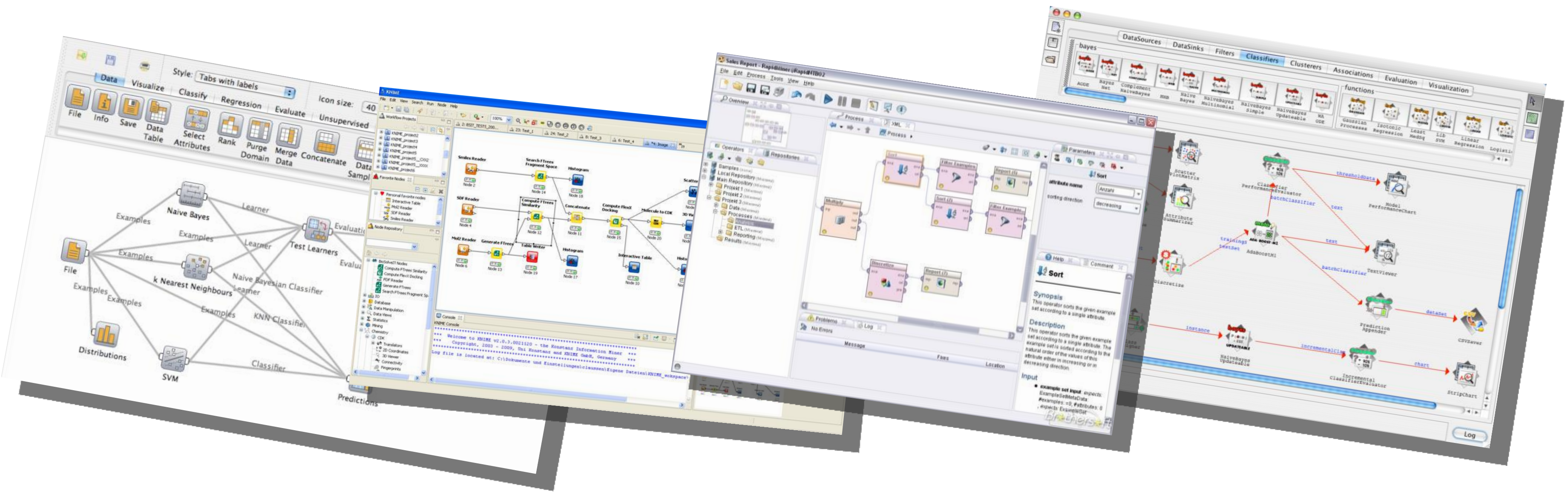


- SD Algorithms in Orange
  - SD (Gamberger & Lavrač, JAIR 2002)
  - Apriori-SD (Kavšek & Lavrač, AAI 2006)
  - CN2-SD (Lavrač et al., JMLR 2004)

Second Generation Data Mining Platforms

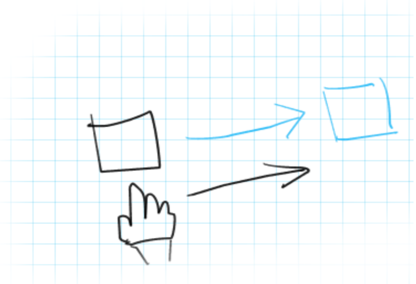
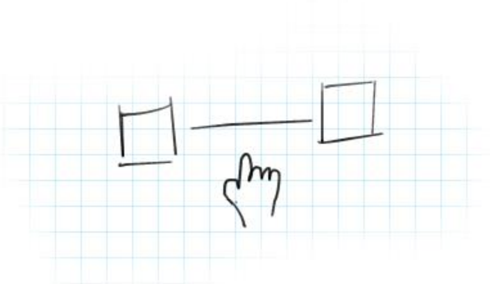
Orange, WEKA, KNIME, RapidMiner, Orange4WS, …

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



Data Mining Workflows for Open Data Science

- Workflows are executable visual representations of procedures
  - divided into smaller chunks of code (components)
  - organized as sequences of connected components.
- Suitable for representing complex scientific pipelines
  - by explicitly modeling dependencies of components
- Building scientific workflows consists of simple operations on workflow elements (drag, drop, connect), suitable for non-experts

Second Generation Data Mining Platforms

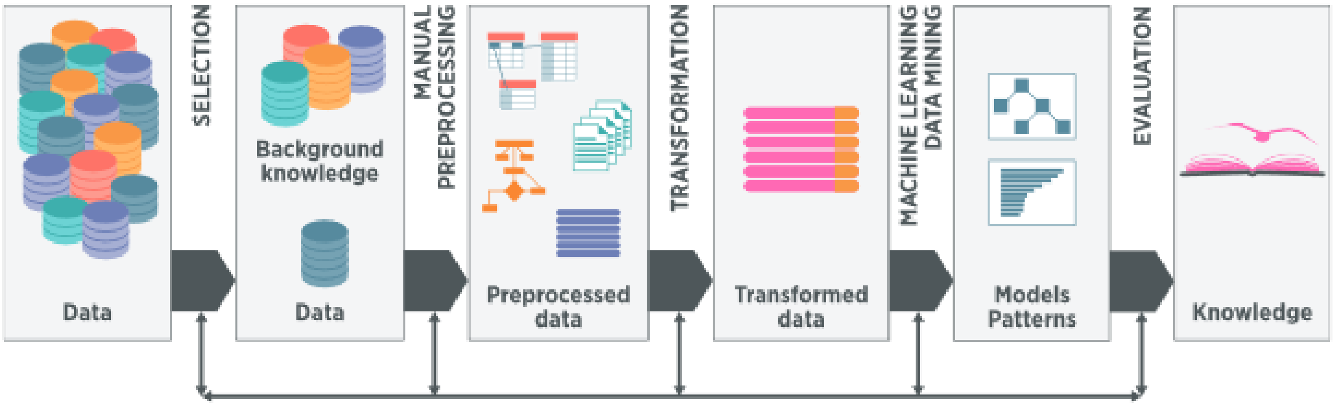
Orange, WEKA, KNIME, RapidMiner, Orange4WS, …

- include numerous data mining algorithms
- enable data and model visualization
- enable complex **workflow** construction
- … but do not include algorithm for mining complex structured data

… developing efficient relational data mining algorithms and making them reusable is still a great challenge

Representation learning: A step in KDD process

- KDD process:



- Important steps:
  - Manual data preprocessing
  - Automated data transformation
- Representation learning = Automated data transformation, performed on manually preprocessed data
- Transformation requires handling heterogeneous data types
  - Data (feature vectors, documents, pictures, data streams, …)
  - Background knowledge (multi-relational data tables, networks, text corpora, …)

Talk overview

- Advances in Relational Learning
  - Background: Machine Learning (ML)
  - Relational Learning (RL)
  - Semantic Relational Learning (SRL)

- Advances in Network Analysis for SRL

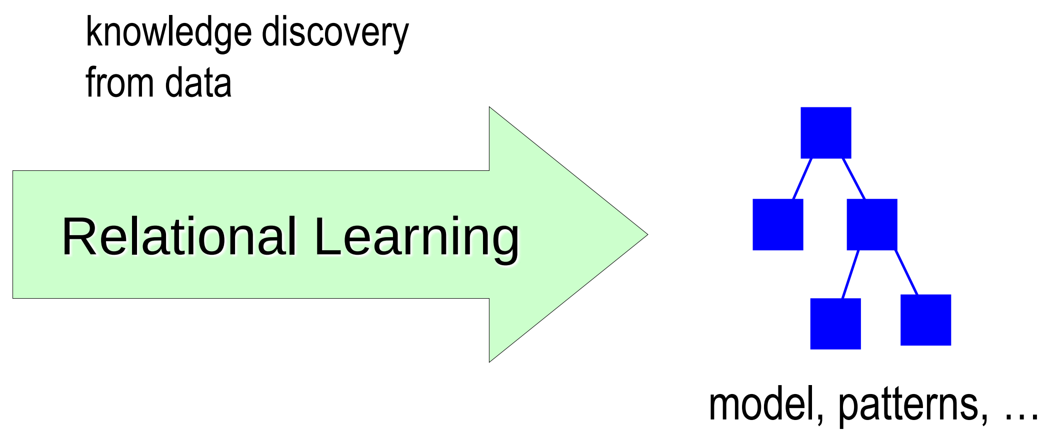
- Conclusions and future work

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.



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

Find: a classification model, a set of patterns

Relational Learning

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

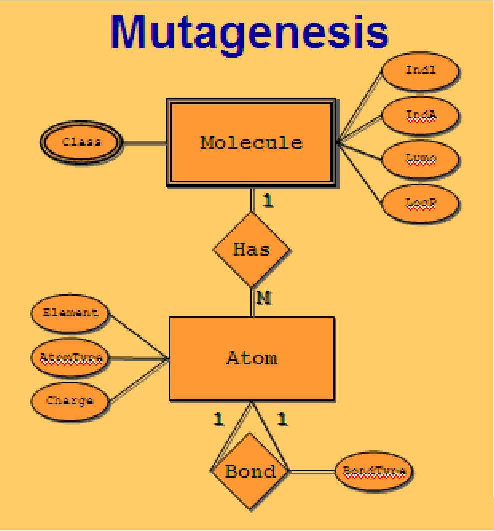
customer							
ID	Zip	Sex	SoSt	Income	Age	Club	Resp
...	...	...	...	...	...	...	...
3478	34677	m	si	60-70	32	me	nr
3479	43666	f	ma	80-90	45	nm	re
...	...	...	...	...	...	...	...

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, relational learning, relational data mining**
  - Learning from complex multi-relational data
  - 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
...	...	...	...	...

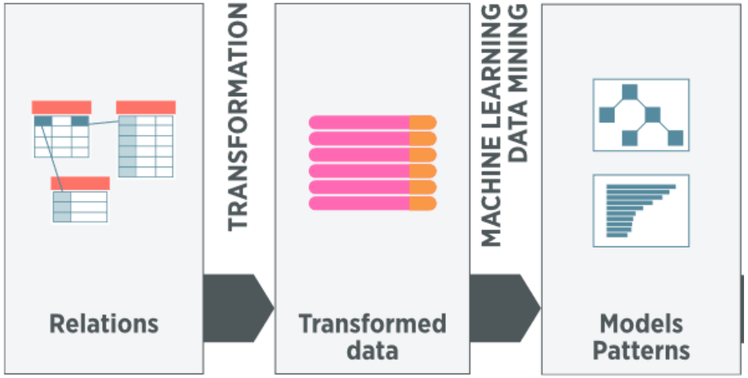
store

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

Relational representation of customers, orders and stores.

Representation Learning in Relation Learning setting

- Relational learning is characterized by using background knowledge (domain knowledge) in the data mining process
- Representation learning = automated transformation of multi-relational data into tabular data format



- Two approaches:
  - Propositionalization: data transformation into symbolic feature vectors
  - Embeddings: data transformation into numeric feature vectors (out of scope of this talk)

Propositionalization approach to Relational Learning

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

1. construct relational features
2. construct a propositional table

Propositionalization approach to Relational Learning

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

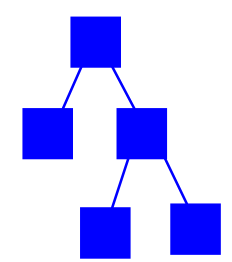
1. construct relational features
2. construct a propositional table

	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



model, patterns, …

Propositionalization approach to Relational Learning: Relational subgroup discovery (RSD)

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. construct relational features
2. construct 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).
```

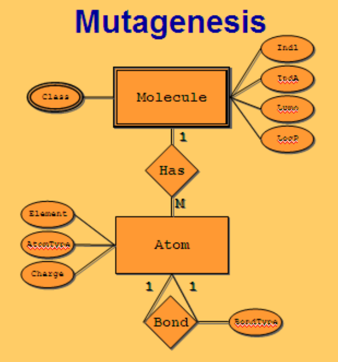
patterns (set of rules)

Relational Subgroup Discovery (RSD) (Železny and Lavrač, MLJ 2006)

- Propositionalization through efficient first-order feature construction

  f121(M):- hasAtom(M,A), atomType(A,21)

  f235(M):- lumo(M,Lu), lessThr(Lu,1.21)

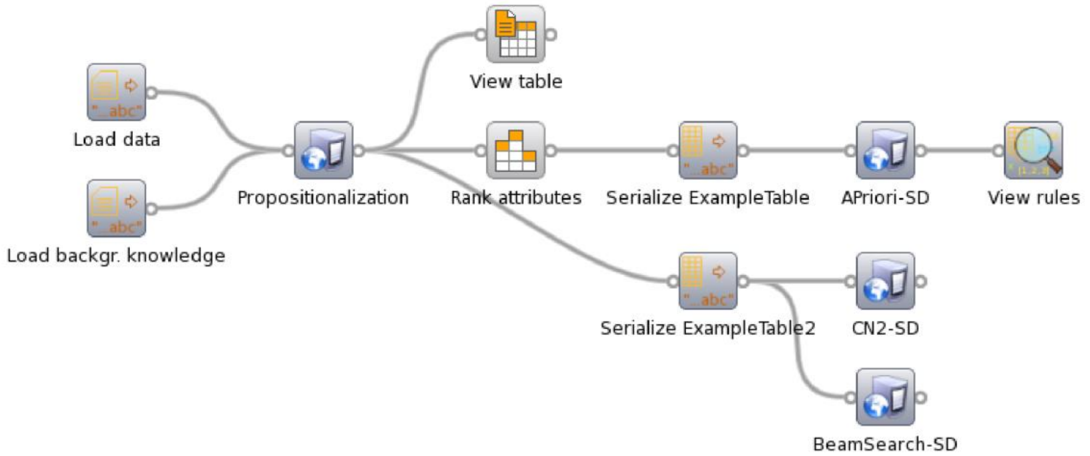


- Transformation into tabular data form

  i.e. binary valued feature vectors

- Subgroup discovery using CN2-SD

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

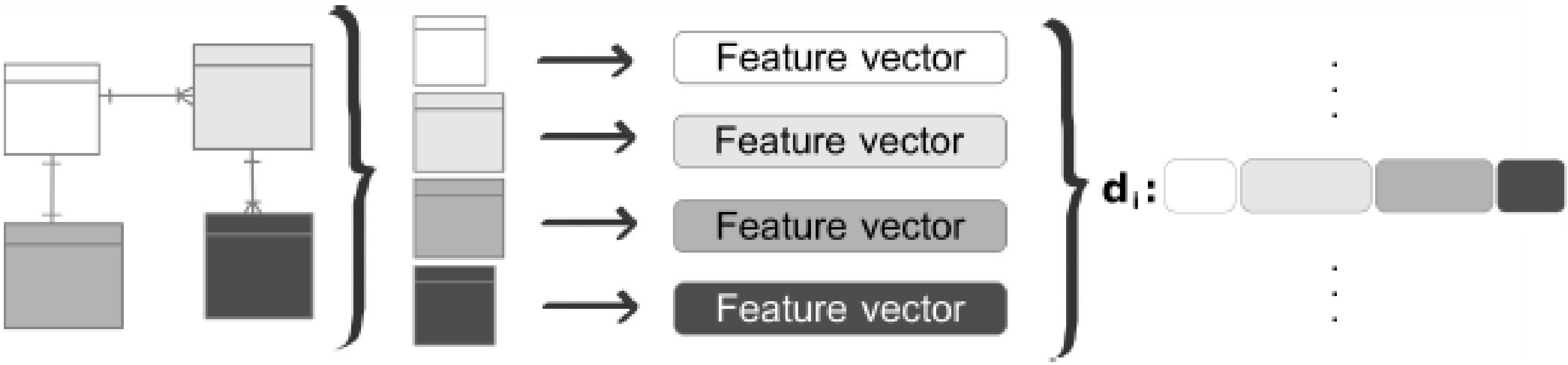


Other propositionalization approaches

- Propositionalization algorithms
  - RSD
  - 1BC
  - RelF
  - …
  - Aleph ILP learner, with its featurize functionality
  - Wordification
    - Our work (Perovsek et al., Wordification: Propositionalization by unfolding relational data into bags of words. Expert Syst. Appl., 2015
    - Recent work of Zacerucha (ILP-2019)

Wordification: Generate simplified relational features

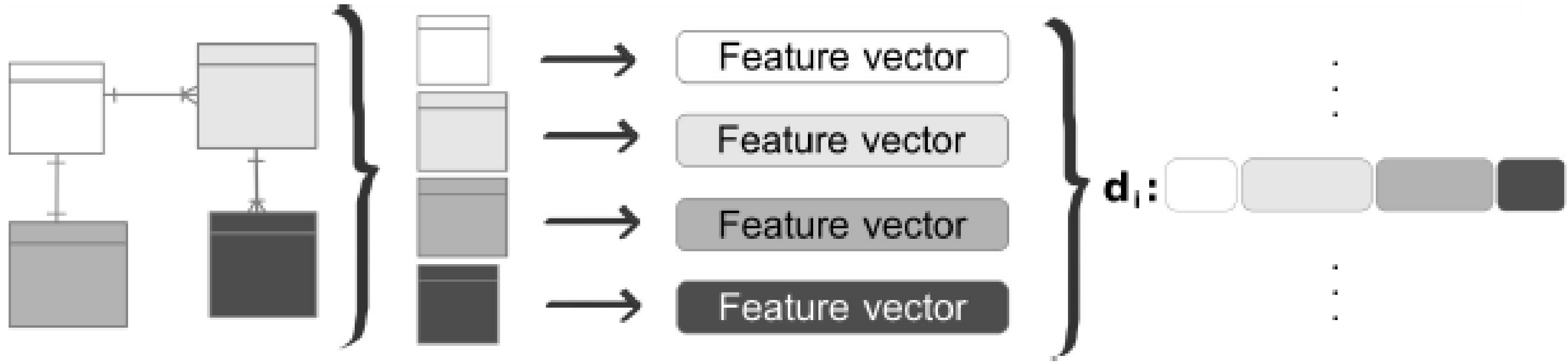
- Transform a relational database into a “document corpus”: For each row in main table, concatenate its “words” with “words” generated for the other tables, linked through external keys



Perovšek et al. Wordification: Propositionalization by unfolding relational data into bags of words. Expert Syst. Appl., 2016

Generate simplified relational features: Wordification

- Transform a relational database into a document corpus: For each row in main table, concatenate its “words” with “words” generated for the other tables



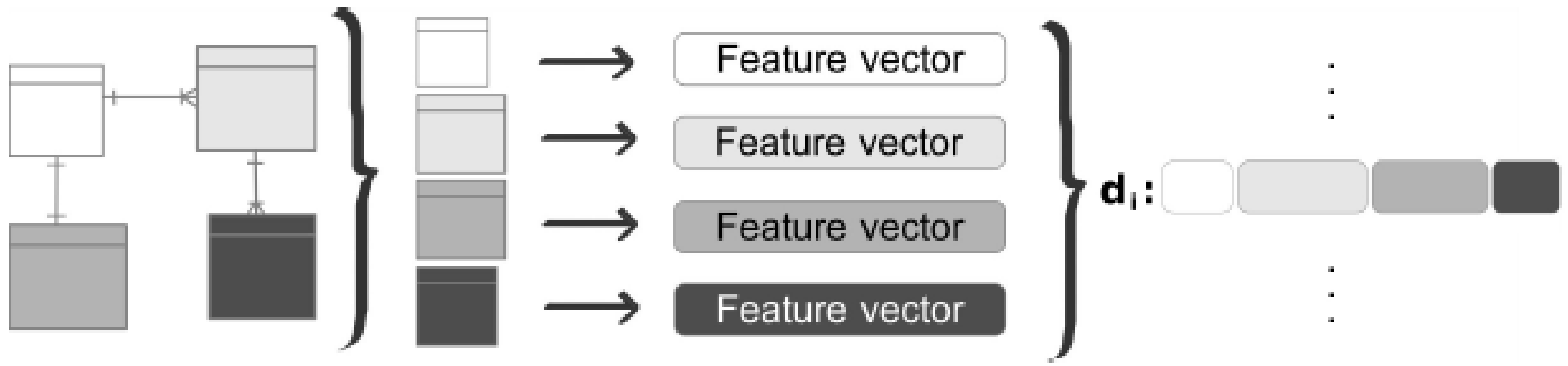
- Individual words (called **word-items**) are constructed as combinations of:

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

- **n-grams** (conjuncts) are constructed to model feature dependencies

Generate simplified relational features: Wordification

- Transform a relational database into a document corpus: For each row in main table, concatenate its “words” with “words” generated for the other tables

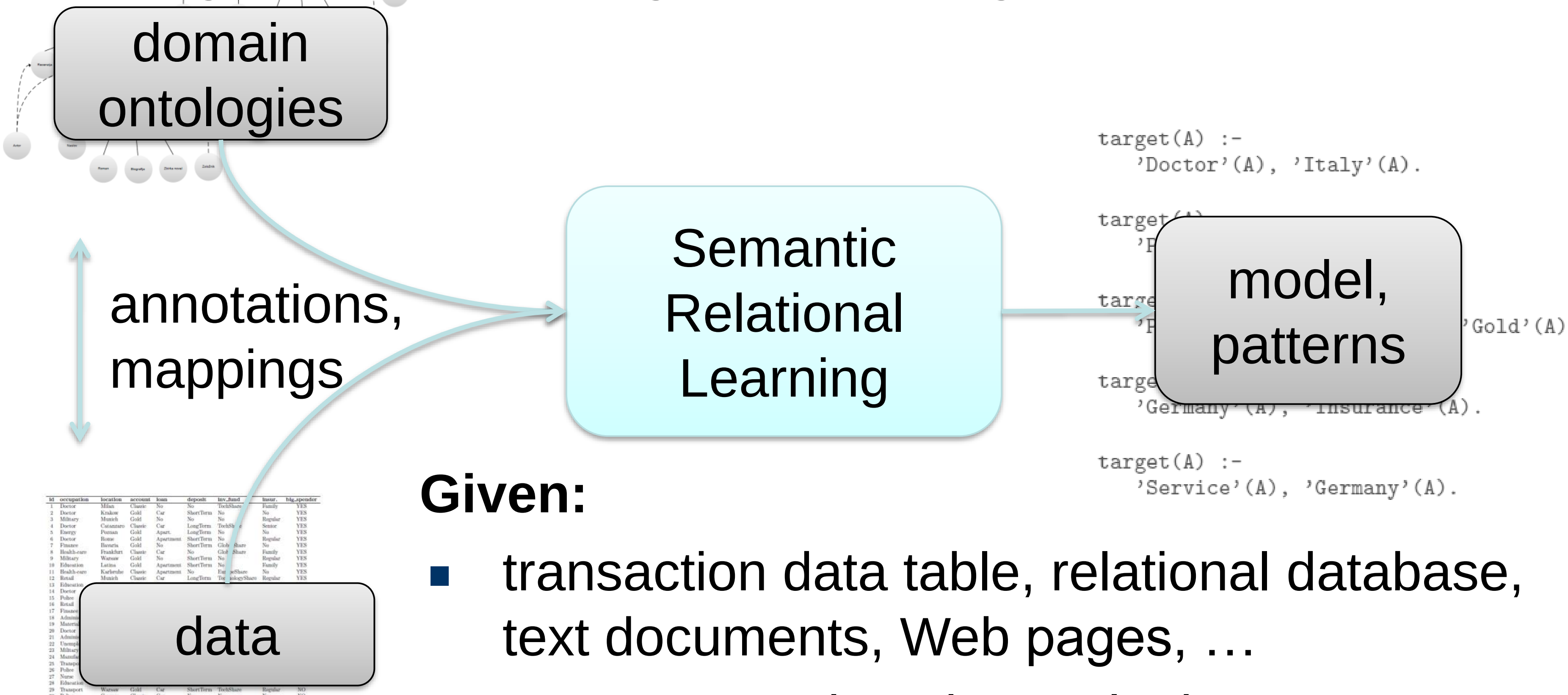


- Outperforms all other propositionalization algorithms (RSD, …)
  - Same or better accuracy
  - Significant speed up (10-100%)
- Further advances by Zaverucha (ILP-2019)

Talk overview

- Advances in Relational Learning
  - Background: Machine Learning (ML)
  - Relational Learning (RL)
  - Semantic Relational Learning (SRL)
- Advances in Network Analysis for SRL
- Conclusions and future work

Semantic Relational Learning: Using ontologies as background knowledge in learning



Given:

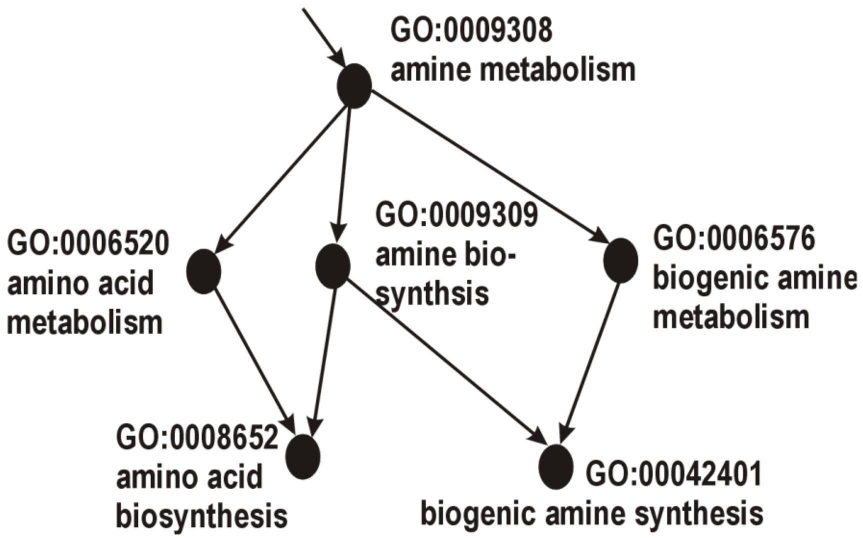
- transaction data table, relational database, text documents, Web pages, …
- one or more domain ontologies

Find: a classification model, a set of patterns

Using domain ontologies in Machine Learning

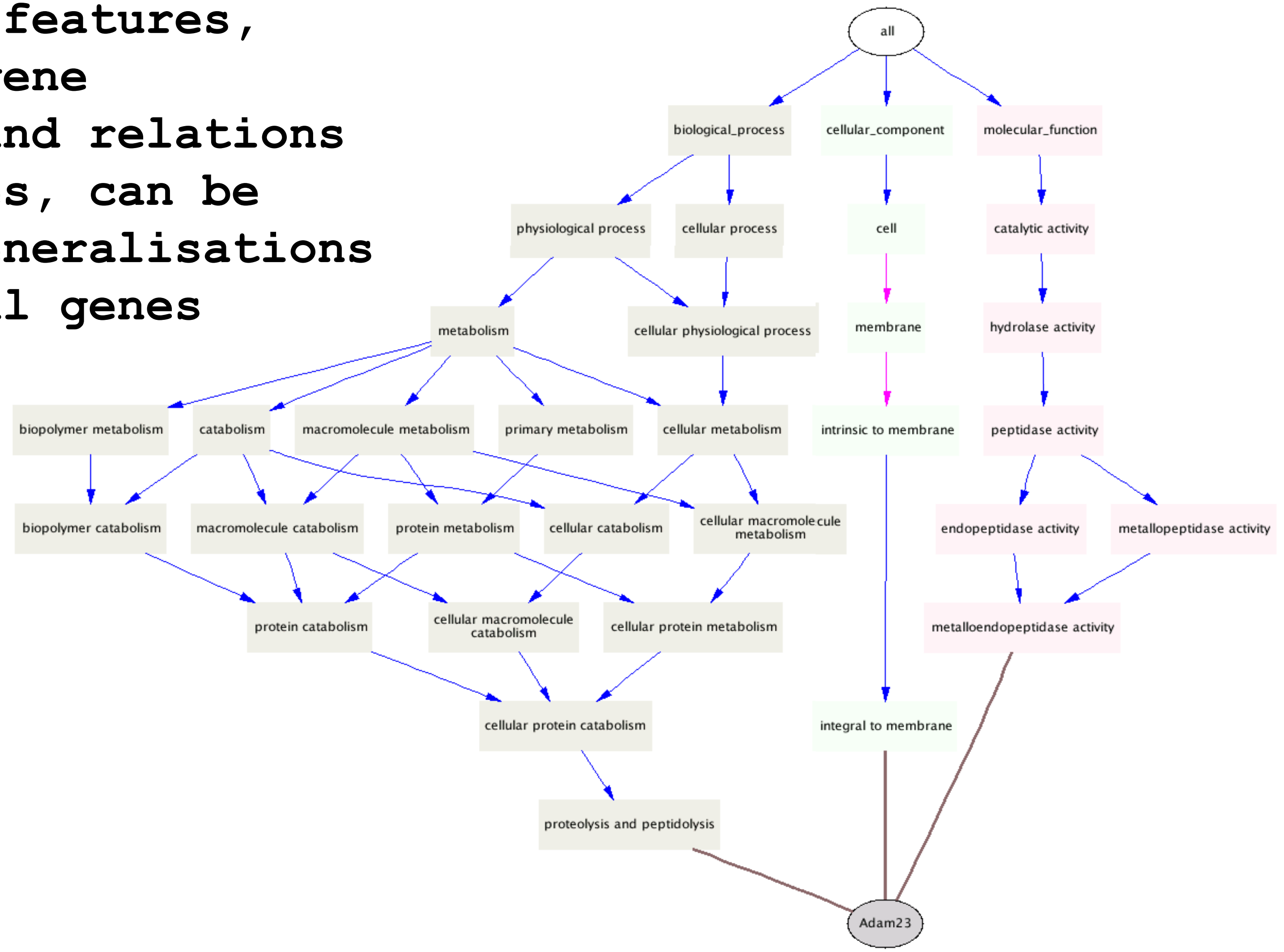
Using domain ontologies as background knowledge, e.g., using the **Gene Ontology** (GO)

- GO is a database of terms, describing gene sets in terms of their
  - functions (over 12,000)
  - processes (over 2,000)
  - components (over 7,500)
- Genes are annotated to GO terms
- Terms are connected (is_a, part_of)
- Levels represent terms generality



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



RSD: Propositionalization approach to Semantic Relational Learning

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: RSD feature construction

Construction of first order features, with support > *min_support*

```
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: RSD Propositionalization

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: 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

Over-expressed

IF

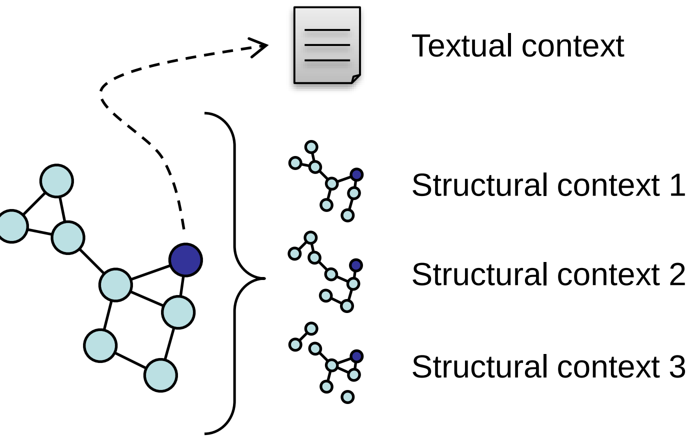
f2 and f3

[4,0]

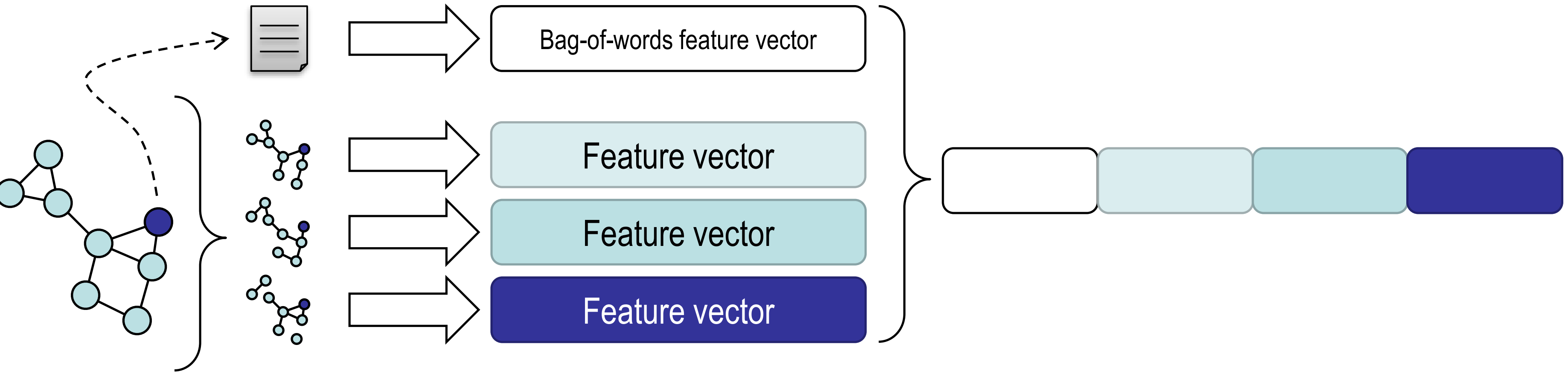
Other propositionalization approaches

- Propositionalization approaches for semantic data mining and heterogeneous information network (knowledge graph) analysis:
  - SDM-Aleph, Hedwig (Vavpetič et al.)
  - HinMine (Grčar et al. 2014, Kralj et al.)
  - NetSDM (Kralj et al.)
  - …

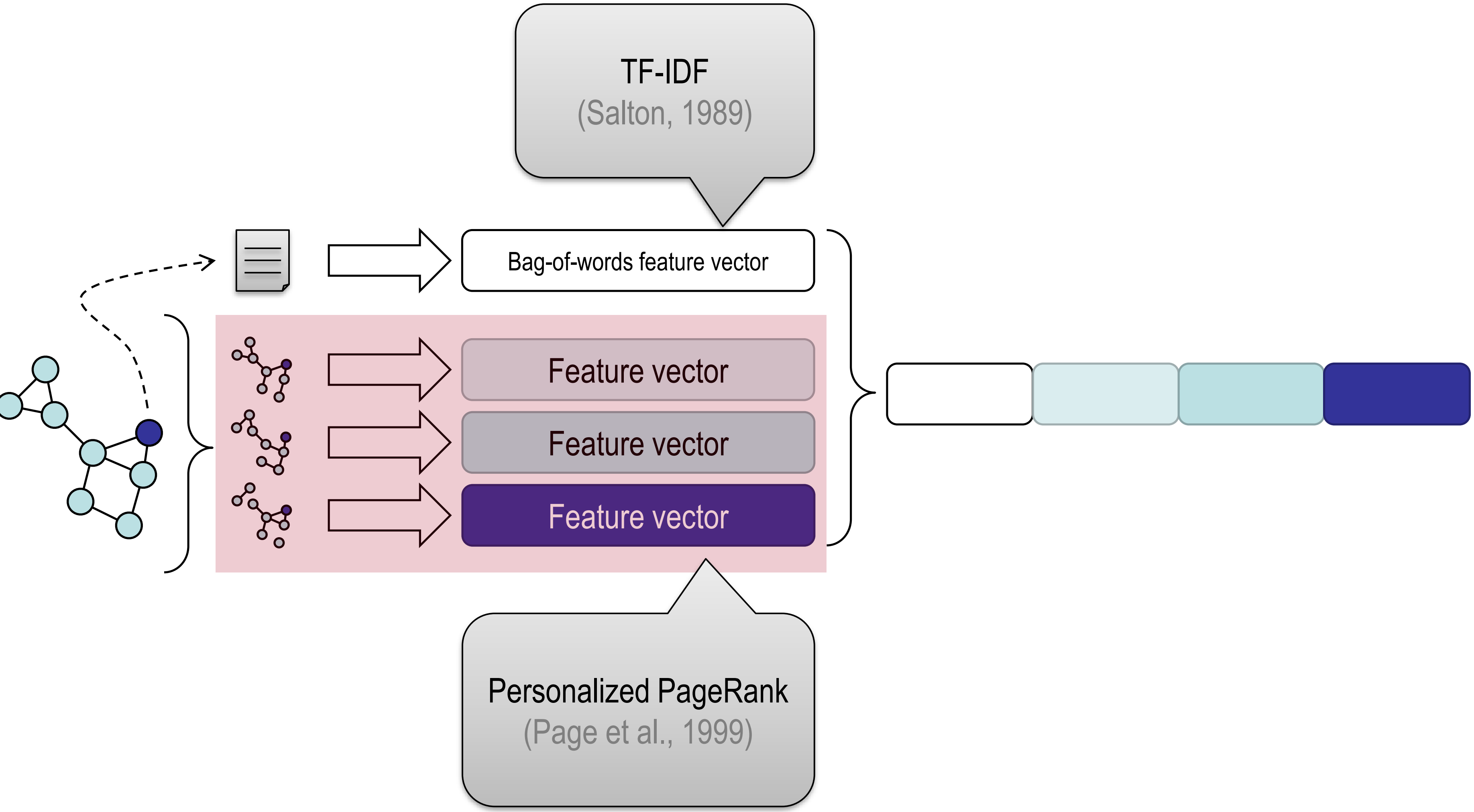
Propositionalization of heterogeneous knowledge graphs



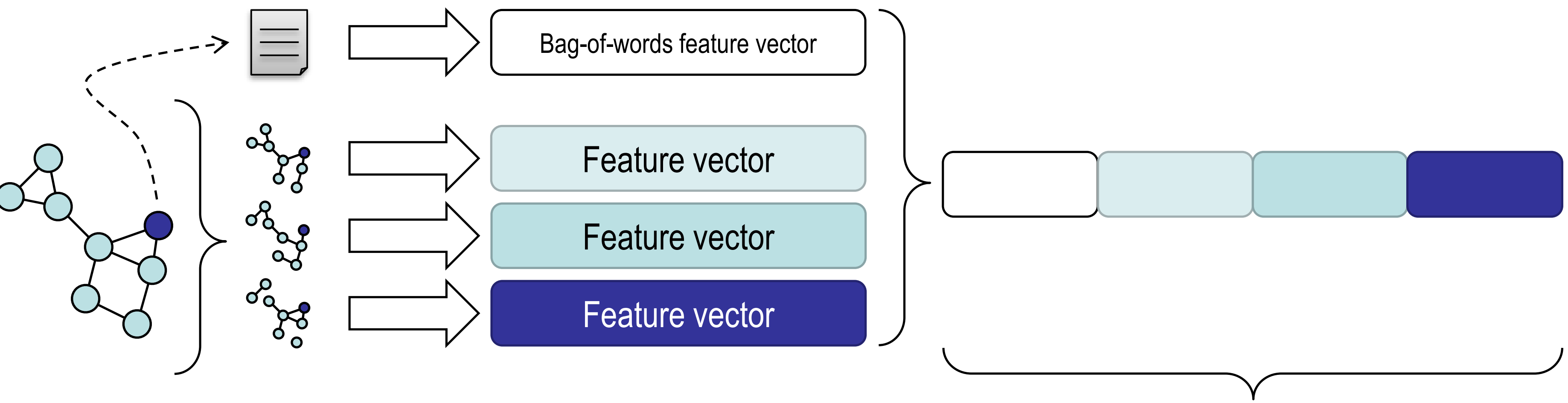
Propositionalization of heterogeneous knowledge graphs



Propositionalization of heterogeneous knowledge graphs



Propositionalization of heterogeneous knowledge graphs



- Concatenate and normalize concatenated weighted feature vectors
- Can be used directly by a text mining algorithm
- Can be used as input layer to an embedding algorithm

Talk overview

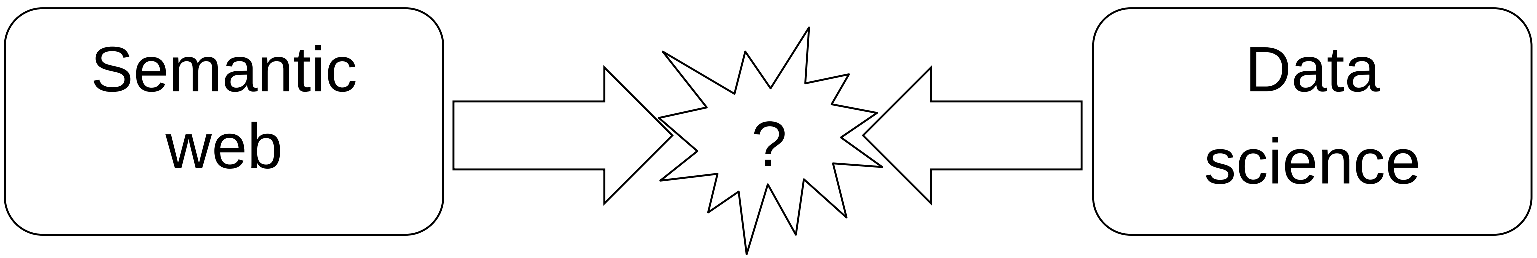
- Advances in Relational Learning
  - Background: Machine Learning (ML)
  - Relational Learning (RL)
  - Semantic Relational Learning (SRL)

Advances in Network Analysis for SRL

- Conclusions and future work

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
- Use Semantic Relational Learning algorithm to mine experimental data with ontologies as background knowledge to get explanations for groups of TargetClass objects, e.g.

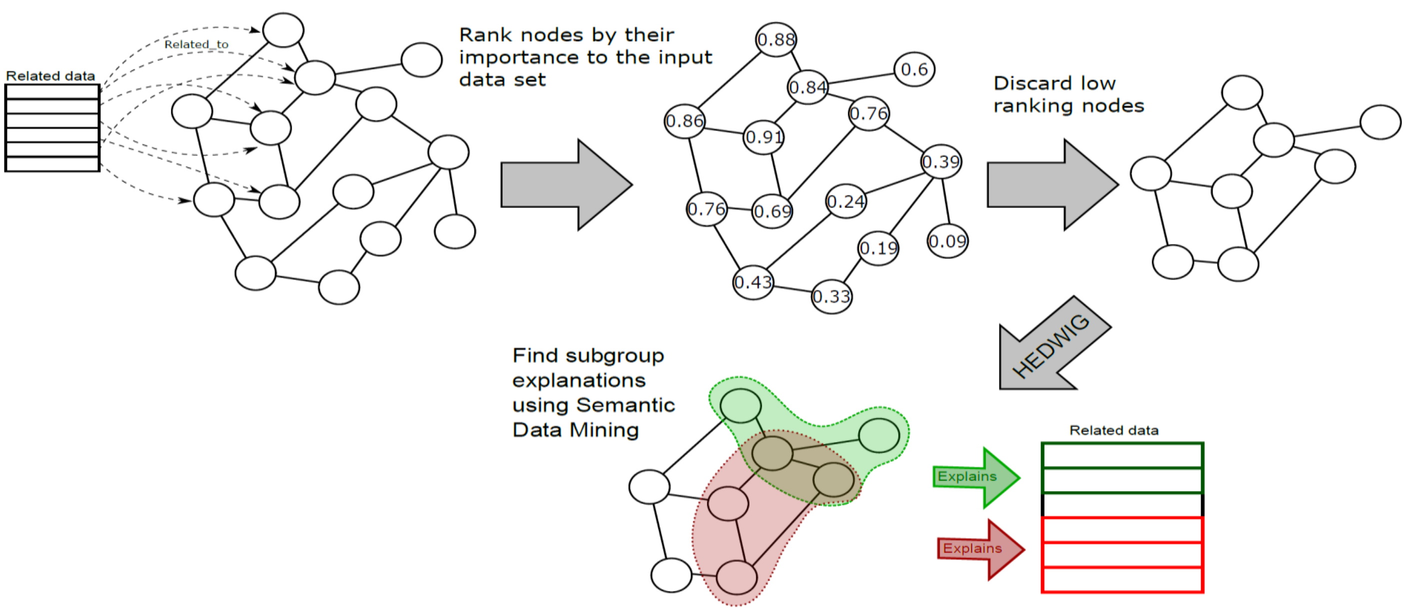
  BreastCancer ← chromosome AND cell cycle

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

(Kralj et al., JMLR 2019)

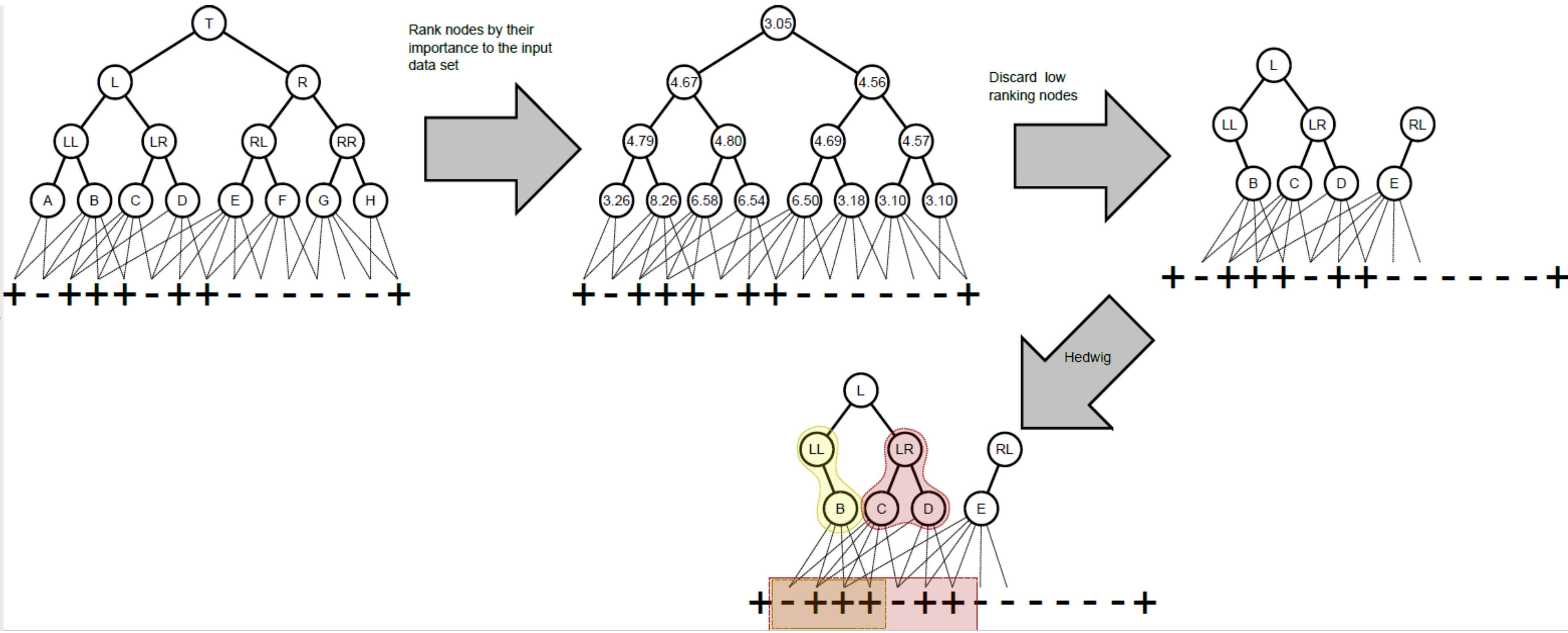
NetSDM Network analysis for feature reduction (Kralj et al. 2019)

- 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

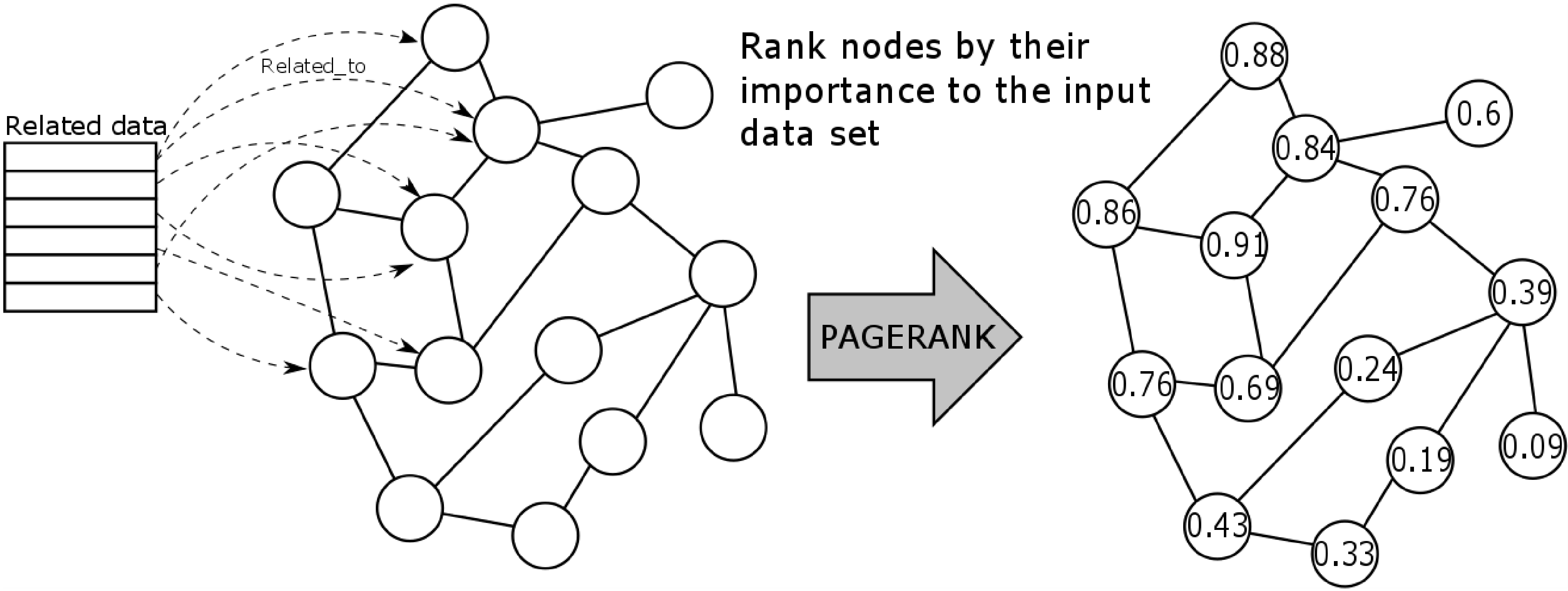


NetSDM algorithm outline

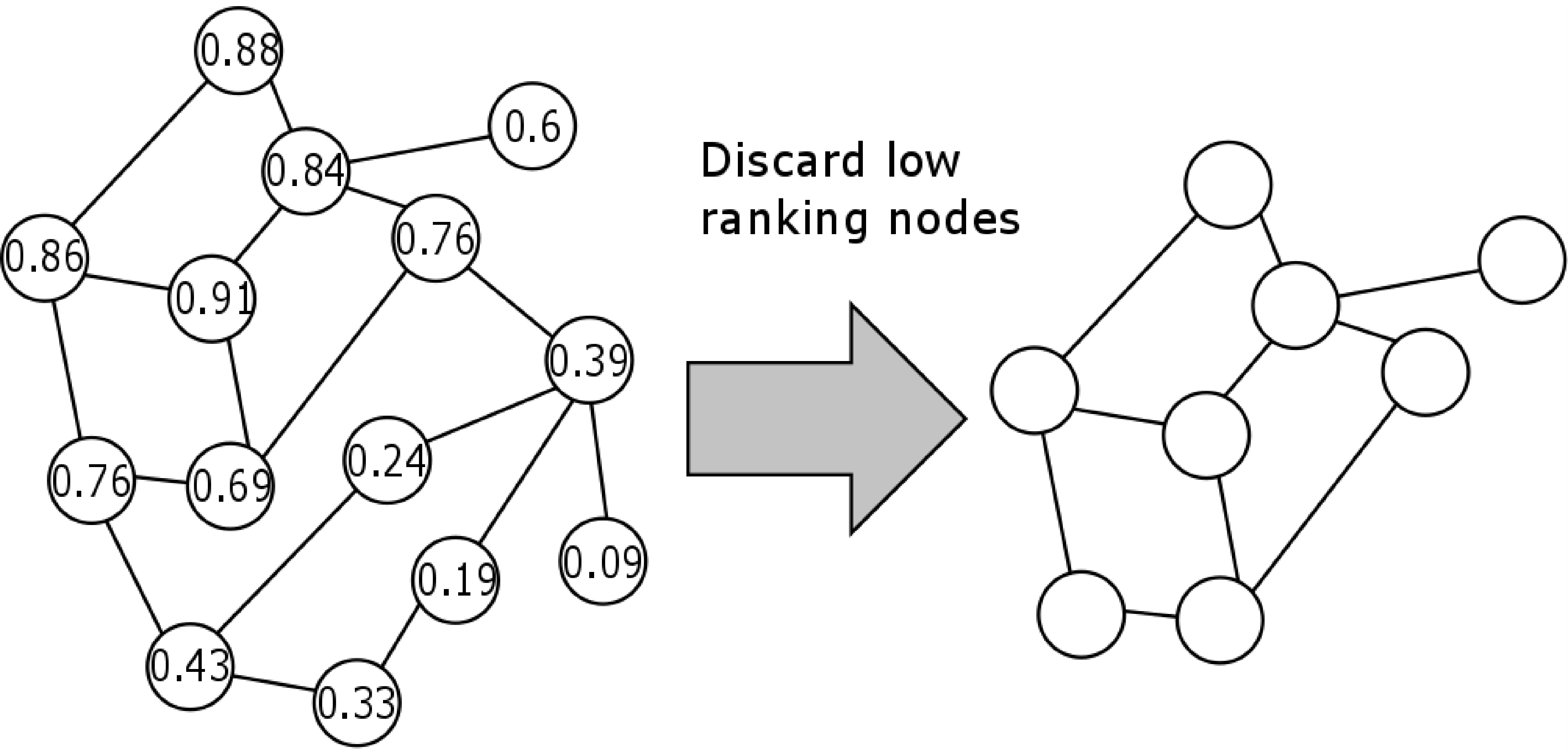
1. Estimate ontology term relevance
2. Delete terms with low relevance
3. Run semantic relational learning algorithm Hedwig on pruned ontolgy



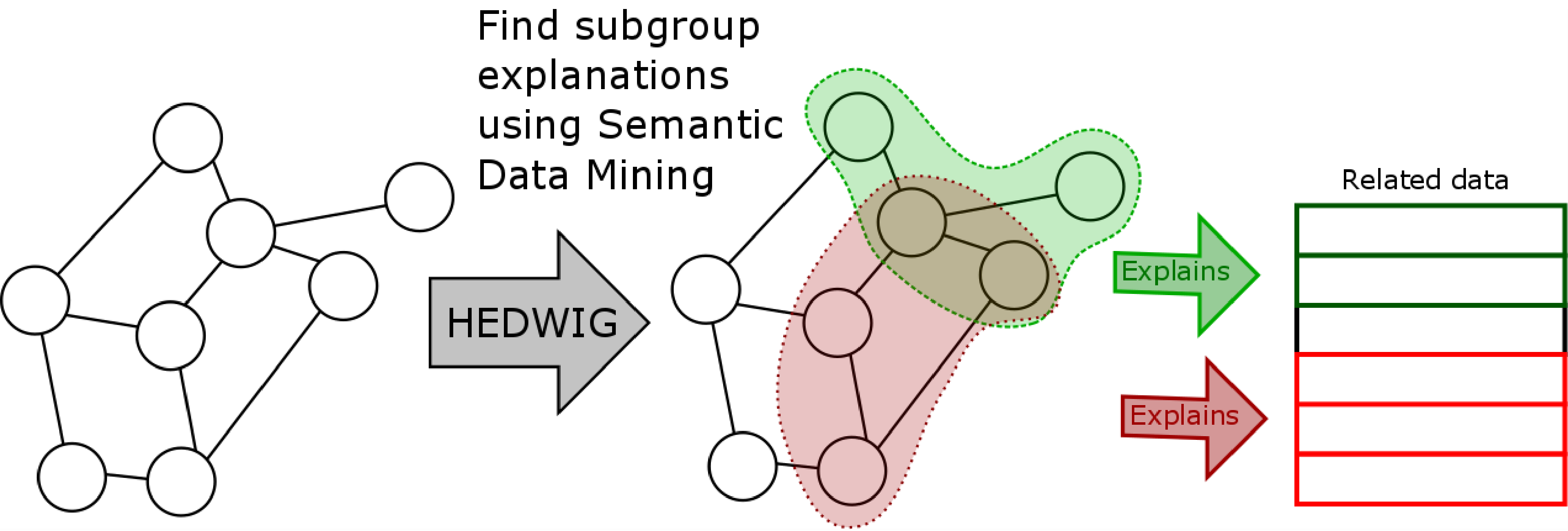
Methodology: Step 1



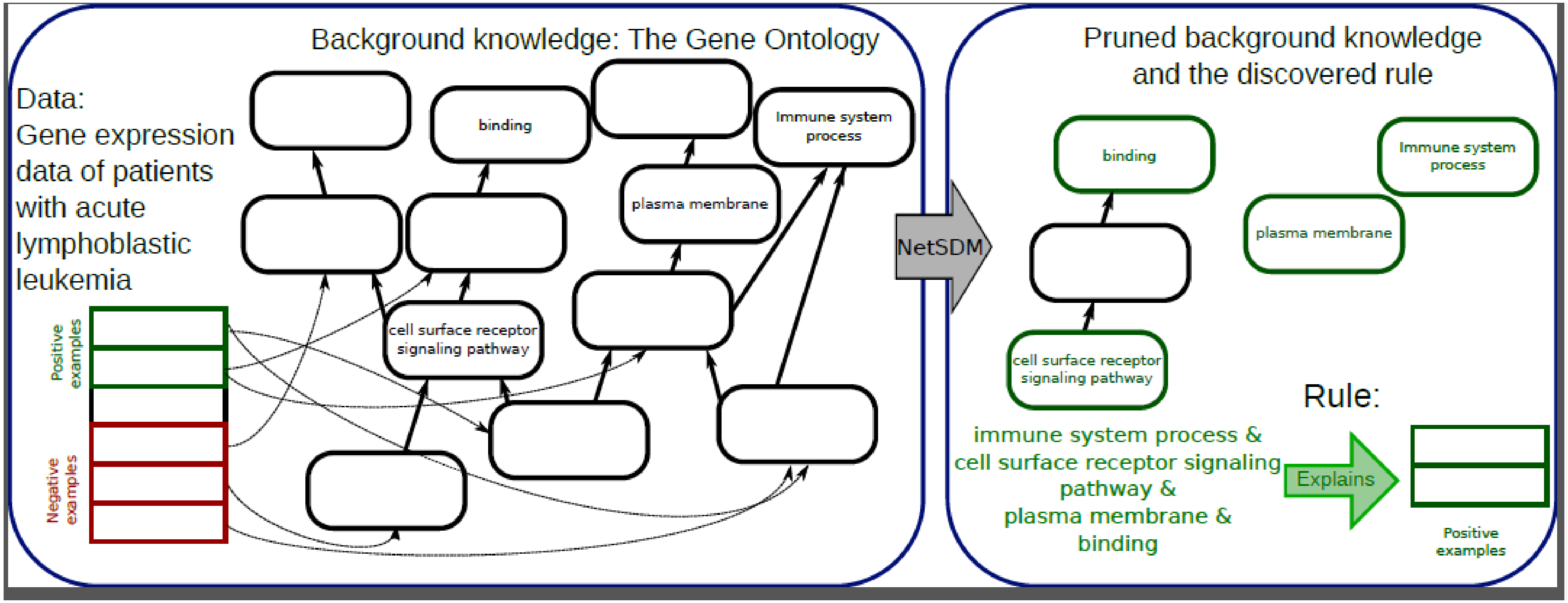
Methodology: Step 2



Methodology: Step 3

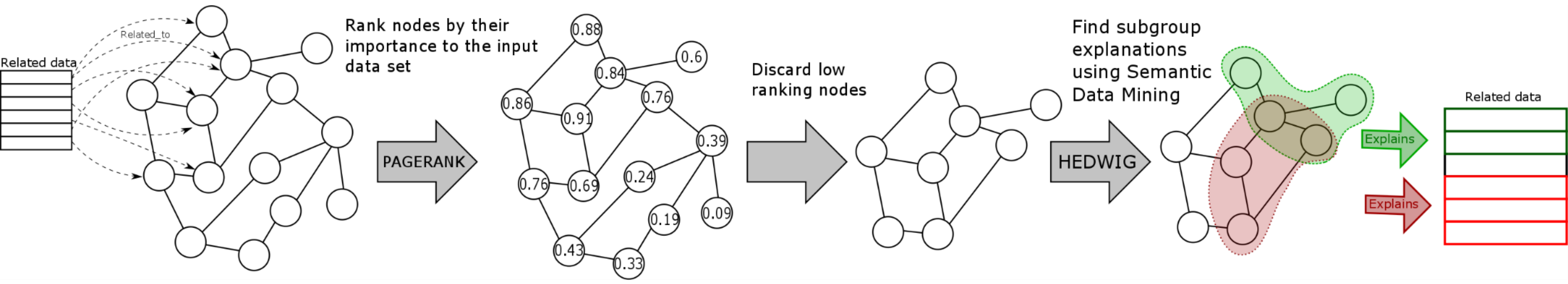


NetSDM:


Background knowledge: The Gene Ontology
Data:
Gene expression
data of patients
with acute
lymphoblastic
leukemia
Positive examples
Negative examples
binding
Immune system process
plasma membrane
cell surface receptor signaling pathway
NetSDM
Pruned background knowledge
and the discovered rule
binding
Immune system process
plasma membrane
cell surface receptor signaling pathway
Rule:
immune system process &
cell surface receptor signaling
pathway &
plasma membrane &
binding
Explains
Positive examples

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



Talk overview

- Advances in Relational Learning
  - Background: Machine Learning (ML)
  - Relational Learning (RL)
  - Semantic Relational Learning (SRL)

- Advances in Network Analysis for SRL

Conclusions and future work

Summary and conclusions

- Our propositionalization 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
  - Can be used as a data fusion mechanism when mining **heterogeneous information networks** (Grčar et al. 2014)
  - A **wordification approach** to propositionalization is especially powerful (Perovšek et al. 2016), including visualization of relational data with word clouds

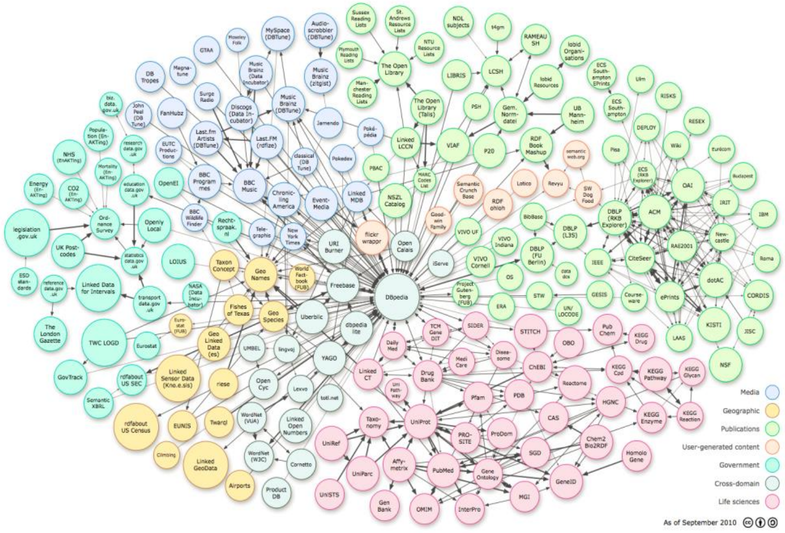
…. all these being implemented and made publicly reusable as complex **workflows in ClowdFlows**

Summary and conclusion: Future work

- Current Semantic data mining scenario (addressed in this lecture): Mining empirical data with ontologies as background knowledge
    - abundant empirical data, but
    - relatively scarce background knowledge
- Future Semantic data mining scenario:
  - Given abundance of ontologies and semantically anotated data collections
    - e.g. Linked Open Data and large knowledge graphs, consisting of
      - billions of RDF triples
      - millions of links

Paradigm shift in Semantic Data Mining: Mining Linked Open

- 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 promising for mining Linked Open Data
- Current work in mining knowledge graphs by Škrlj et al.



Summary: Semantic Relational Learning in context