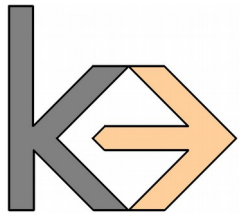
On the Understandability of Rule Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Johannes Fürnkranz
TU Darmstadt
Knowledge Engineering Group
Hochschulstrasse 10
D-64289 Darmstadt
06151/166238

juffi@ke.tu-darmstadt.de



Joint Work with T. Kliegr, E. Loza, J. Zilke, F. Janssen, H. Paulheim, and J. Stecher

Thanks to J. Vreeken, K. Kersting

Data Mining

statistics

background knowlege

utility theory
cost models

Data Mining is the non-trivial *process* of identifying

- valid
- novel
- potentially useful
- ultimately understandable

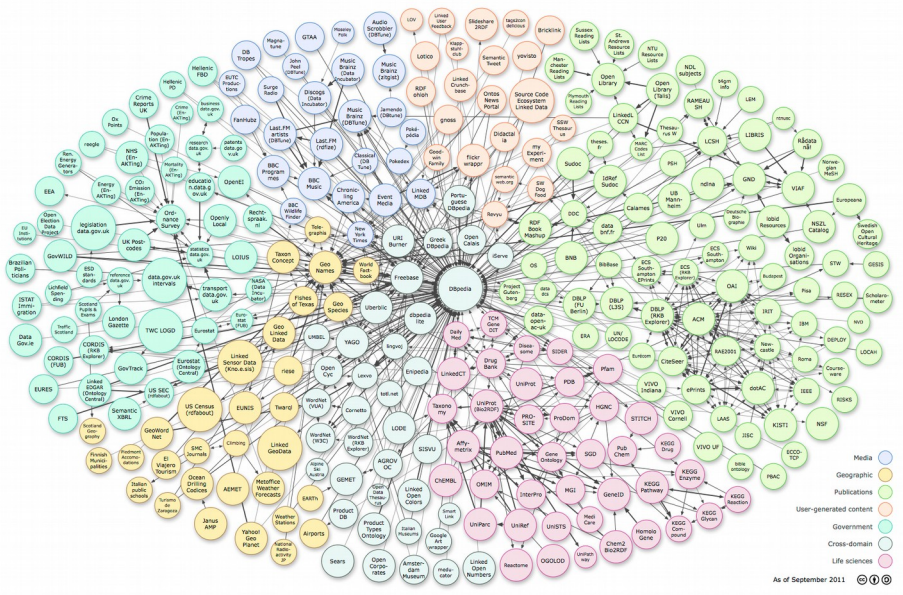
patterns in data.

(Fayyad et al. 1996)

Understandabiity = Rules?

- Rules provide a good (the best?) trade-off between
  - human understandability
  - machine executability
- Used in many applications which will gain importance in the near future
  - Security
  - Spam Mail Filters
  - Semantic Web
- But they are not a universal tool
  - e.g., learned rules sometimes lack in predictive accuracy

    → challenge to close or narrow this gap



Understandability – State of Affairs

Data Mining essentially assume

- Rules are inherently understandable
- Shorter rules are more understandable than longer rules
- Good explanations = Good fit to the data
- No additional criteria or algorithms are needed to address understandability

But there has been some evidence that these assumptions are not always correct, e.g.

> *"The results also suggest that, at least in some cases, understandability is negatively correlated with the complexity, or the size, of a model."*
>
> (Allahyari & Lavesson 2011)

Overview

- Motivation
  - Understandability has not received much attention
- Understandability
  - Conjunctive Fallacy
  - Gambler's Fallacy
  - Representativeness Heuristic
- Different Types of Rules
  - Discriminative vs. Characteristic Rules
  - Formal Concepts
  - Closed Itemsets
  - Occam's Razor & MDL
- Heuristic Rule Learning
  - Concept Learning
  - Coverage Spaces
  - Rule Learning Heuristics
  - Inverted Heuristics
- Understandability of Rules
  - Rule Length
  - Semantic Coherence
  - Recognition Heuristic
  - Relevance
  - Structure
- Conclusions

Conjunctive Fallacy

(Tversky & Kahneman 1983)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

A) Linda is a bank teller.

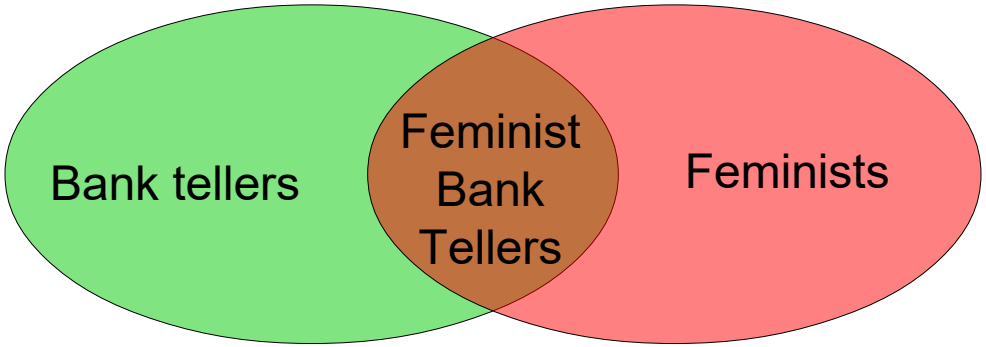
B) Linda is a bank teller and is active in the feminist movement.

Conjunctive Fallacy

(Tversky & Kahneman 1983)

- The majority of people (85%) preferred B)
- However, B) is a specialization of A), so that A) cannot be less probable than B)

$$\Pr(bank \wedge feminist) = \Pr(feminist | bank) \cdot \Pr(bank) \leq \Pr(bank)$$



Representativeness Heuristics

(Kahneman & Tversky 1972)

Humans tend to judge probability of a subgroup according to how similar it is to a prototype of the base group.

Linda

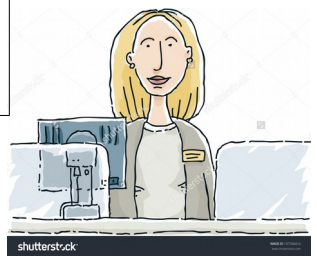
31 years old, single, outspoken, very bright. majored in philosophy, concerned with issues discrimination and social justice, anti-nuclear

less similar

more similar

Bank Teller

Bank Teller
Feminist



Gambler's Fallacy

Which sequence of outcomes on the roulette table is more likely?

A) 0 0 0 0 0 0

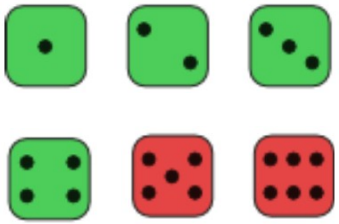
B) 27 18 4 23 8 17

People tend to think the 2nd sequence is more likely because it is *more representative of a random sequence*.

Gambler's Fallacy

(Tversky & Kahneman 1983)

Consider a regular six-sided die with four green faces and two red faces. The die will be rolled 20 times and the sequence of greens (G) and reds (R) will be recorded. You are asked to select one sequence, from a set of three, and you will win $25 if the sequence you choose appears on successive rolls of the die.



A) 

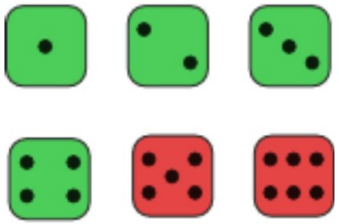
B) 

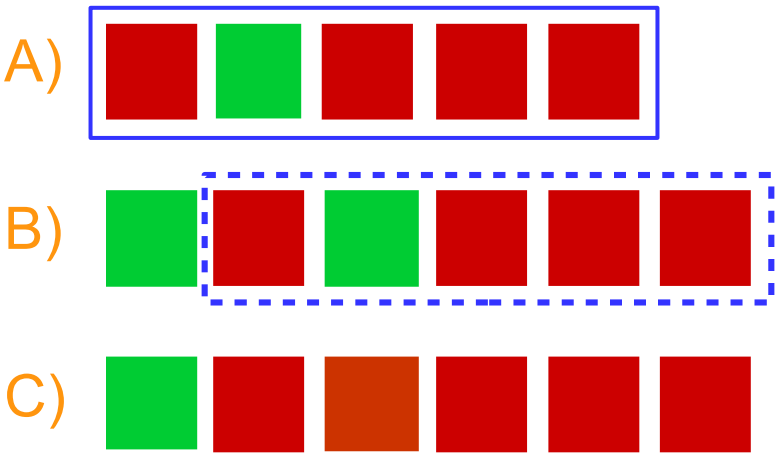
C) 

Gambler's Fallacy

(Tversky & Kahneman 1983)

Consider a regular six-sided die with four green faces and two red faces. The die will be rolled 20 times and the sequence of greens (G) and reds (R) will be recorded. You are asked to select one sequence, from a set of three, and you will win $25 if the sequence you choose appears on successive rolls of the die.





65% bet on B) even though A) is a subsequence of B) and will thus appear more frequently

Understandability vs. Rule Length

Conventional Rule learning algorithms tend to learn short rules

- They favor to add conditions that exclude many negative examples

Typical intuition: Short rules are better

- long rules are less understandable, therefore short rules are preferable
- short rules are more general, therefore (statistically) more reliable and would have been easier to falsify on the training data

Claim: Shorter rules are not always better

- **Predictive Performance:** Longer rules often cover the same number of examples than shorter rules so that (statistically) there is no preference for choosing one over the other
- **Understandability:** In many cases, longer rules may be much more intuitive than shorter rules

→ *we need to understand understandability*!

Overview

- Motivation
  - Understandability has not received much attention
- Understandability
  - Conjunctive Fallacy
  - Gambler's Fallacy
  - Representativeness Heuristic
- Different Types of Rules
  - Discriminative vs. Characteristic Rules
  - Formal Concepts
  - Closed Itemsets
  - Occam's Razor & MDL
- Heuristic Rule Learning
  - Concept Learning
  - Coverage Spaces
  - Rule Learning Heuristics
  - Inverted Heuristics
- Understandability of Rules
  - Rule Length
  - Semantic Coherence
  - Recognition Heuristic
  - Relevance
  - Structure
- Conclusions

Discriminative Rules

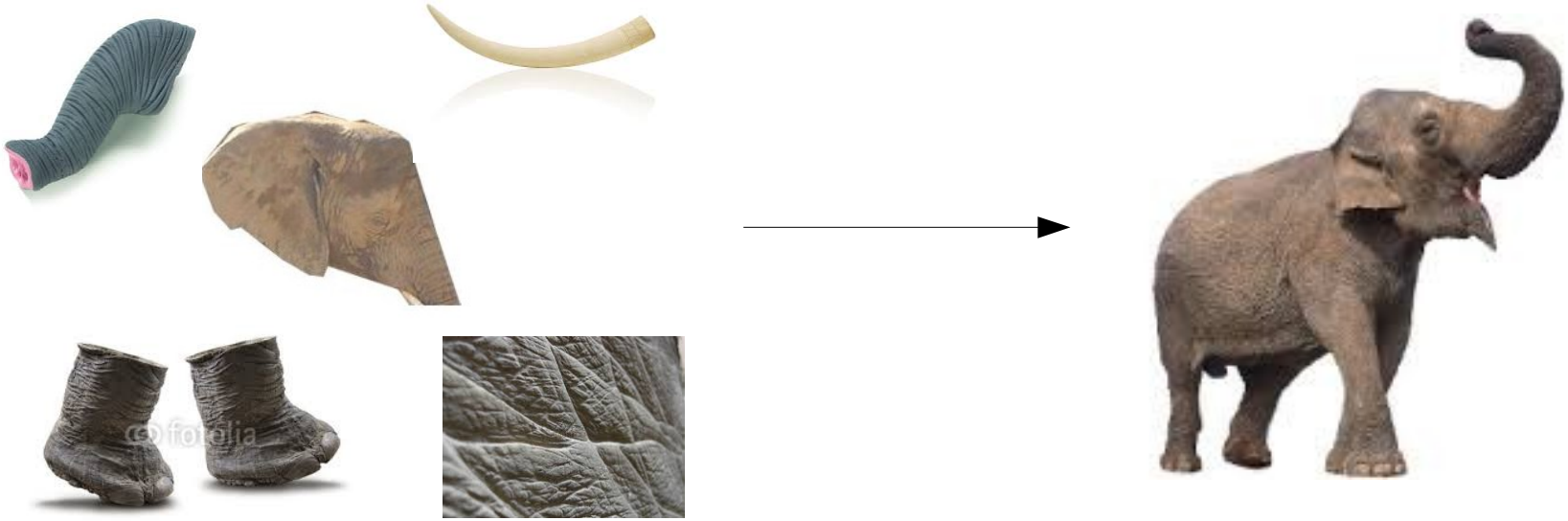
- Allow to quickly **discriminate an object** of one category from objects of other categories
- Typically a few properties suffice

- Example:

Characteristic Rules

- Allow to characterize an object of a category
- Focus is on all properties that are **representative** for objects of that category

- Example:



Discriminative Rules vs. Characteristic Rules

(Michalski 1983)

Michalski (1983) discerns two kinds of classification rules:

- **Discriminative Rules:**
  - A way to distinguish the given class from other classes

$$\texttt{Features} \rightarrow \texttt{Class}$$

  - Most interesting are *minimal discriminative rules*.
- **Characteristic Rules:**
  - A conjunction of all properties that are common to all objects in the class

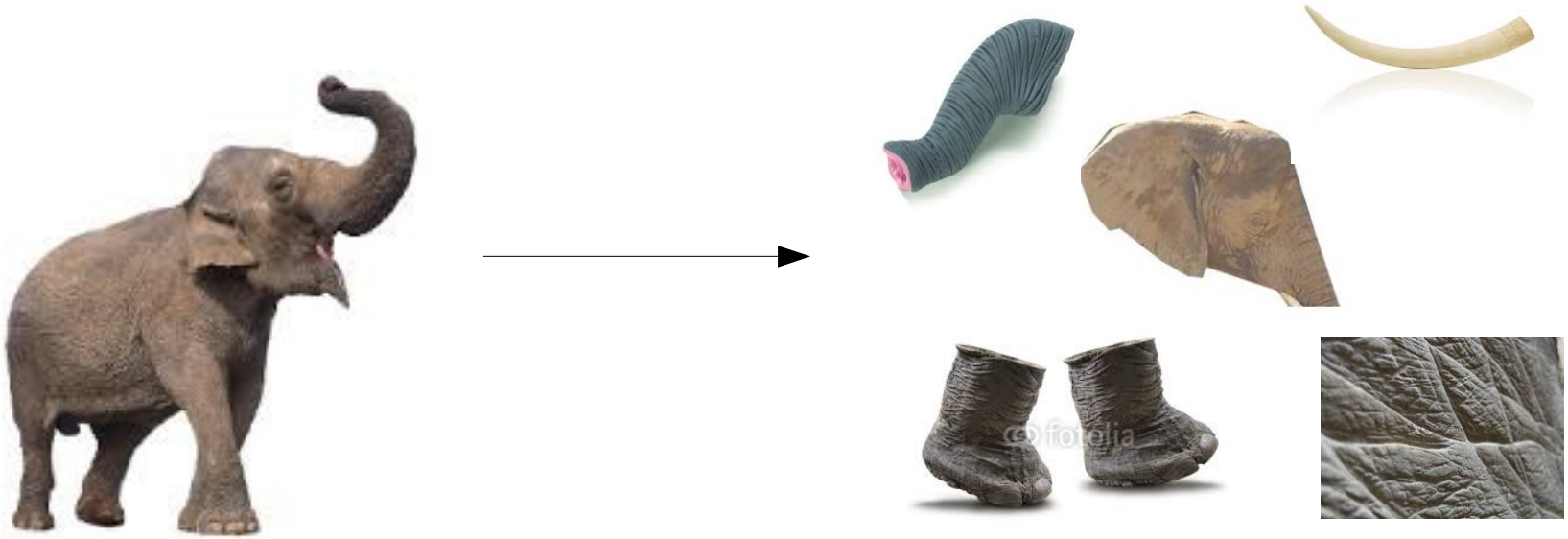
$$\texttt{Class} \rightarrow \texttt{Features}$$

  - Most interesting are *maximal characteristic rules*.

Characteristic Rules

- An alternative view of characteristic rules is to invert the implication sign
- All properties that are implied by the category

- Example:



(Informal) Formal Concept Analysis

(Wille 1982)

Intent of a Concept (Rule)

- Conjunction of Features

Extent of a Concept (Coverage)

- All objects (examples) that are covered by a rule

Formal Concept:

- A rule that cannot be further extended without losing coverage of one of its covered examples (*maximal intent*)
- Along with *all* covered examples (*maximal extent*)

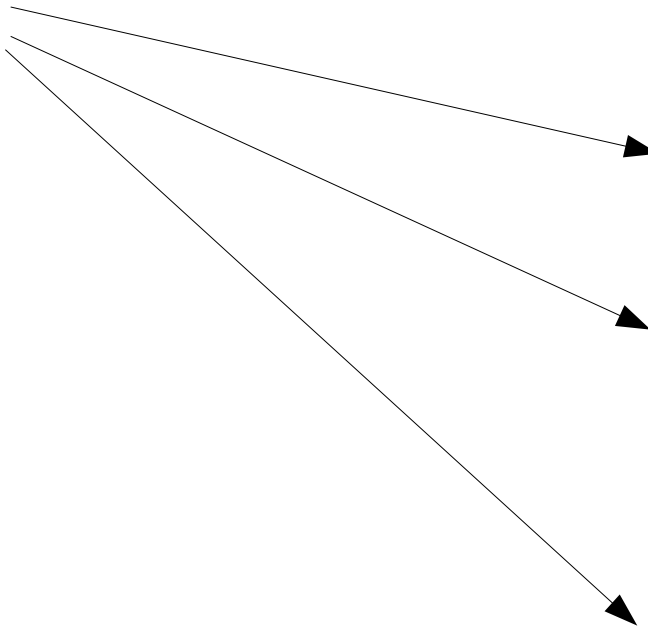
- Essentially, a formal concept is a maximal discriminative / characteristic rule (i.e., an equivalence)

Features ↔ **Class**

FCA Example

Concept "`education = university`"

- Maximal extent:



Education	Marital S.	Income
Primary	Single	Low
Primary	Single	Low
Primary	Married	Low
University	Divorced	High
University	Married	High
Secondary	Single	Low
University	Single	High
Secondary	Divorced	High
Secondary	Single	High
Secondary	Married	Low
Secondary	Divorced	Low
University	Divorced	High
Secondary	Divorced	Low

FCA Example

Concept "`education = university`"

- Maximal extent
  - All covered examples
- Maximal intent
  - All conditions common to the covered examples

→ Formal Concept

"`Education = university` **AND** `Income = high`"

Education	Marital S.	Income
Primary	Single	Low
Primary	Single	Low
Primary	Married	Low
University	Divorced	High
University	Married	High
Secondary	Single	Low
University	Single	High
Secondary	Divorced	High
Secondary	Single	High
Secondary	Married	Low
Secondary	Divorced	Low
University	Divorced	High
Secondary	Divorced	Low

Closed Itemsets

In association rule discovery, formal concepts are called **closed itemsets**

- Although there is no statistical difference between an itemset and its closure (except for #items), their interestingness may change

Shopping Basket of a young family:

Itemset

Closed Itemset

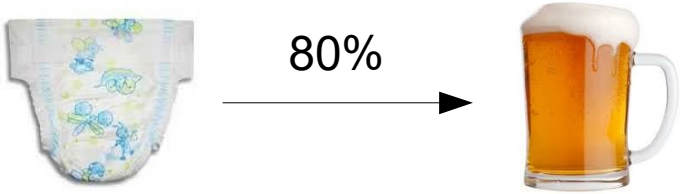
Rule Pruning

Rules are often pruned in order to get the shortest rule

- Remove conditions from the rule as long as the evaluation measure does not significantly change

This may also significantly change the semantics without changing the statistics

Rule



80% of customers who buy diapers also buy beer

Pruned Rule

`true` 80%

80% of all customers buy beer

Overview

- Motivation
  - Understandability has not received much attention
- Understandability
  - Conjunctive Fallacy
  - Gambler's Fallacy
  - Representativeness Heuristic
- Different Types of Rules
  - Discriminative vs. Characteristic Rules
  - Formal Concepts
  - Closed Itemsets
  - Occam's Razor & MDL
- Heuristic Rule Learning
  - Concept Learning
  - Coverage Spaces
  - Rule Learning Heuristics
  - Inverted Heuristics
- Understandability of Rules
  - Rule Length
  - Semantic Coherence
  - Recognition Heuristic
  - Relevance
  - Structure
- Conclusions

Conjunctive Rule

```
+  :-  feature1, feature2.
```

Head of the rule (THEN-part)

- contains a prediction
- typically + if object belongs to concept, – otherwise

Body of the rule (IF-part)

- contains a conjunction of conditions
- a condition is a binary feature

- Coverage
  - A rule is said to ***cover*** an example if the example satisfies the conditions of the rule.
- Prediction
  - If a rule covers an example, the rule's head is predicted for this example.

A Sample Database

No.	Education	Marital S.	Income	Children?	**Approved?**
1	Primary	Single	Low	N	-
2	Primary	Single	Low	Y	-
3	Primary	Married	Low	N	+
4	University	Divorced	High	N	+
5	University	Married	High	Y	+
6	Secondary	Single	Low	N	-
7	University	Single	HIgh	N	+
8	Secondary	Divorced	High	N	+
9	Secondary	Single	High	Y	+
10	Secondary	Married	Low	Y	+
11	Primary	Married	High	N	+
12	Secondary	Divorced	Low	Y	-
13	University	Divorced	High	Y	-
14	Secondary	Divorced	Low	N	+

Property of Interest
("class variable")

A Possible Solution

```
+ :- E=primary,    I=low,    M=married,  C=no.
+ :- E=university, I=high,   M=divorced, C=no.
+ :- E=university, I=high,   M=married,  C=no.
+ :- E=university, I=high,   M=single,   C=no.
+ :- E=secondary,  I=high,   M=divorced, C=no.
+ :- E=secondary,  I=high,   M=single,   C=yes.
+ :- E=secondary,  I=low,    M=married,  C=yes.
+ :- E=primary,    I=high,   M=married,  C=no.
+ :- E=secondary,  I=low,    M=divorced, C=no.
```

The solution is

- a set of rules
- that is complete and consistent on the training examples

- but it does not generalize to new examples
- and is not easily understandable

A Better Solution

```
+ :- Marital = married.
+ :- Marital = single,   Income = high.
+ :- Marital = divorced, Children = no.
```

This solution is also

- a set of rules
- that is complete and consistent on the training examples

- but it does ~~not~~ generalize to new examples
- and is ~~not~~ easily understandable

Occam's Razor

> Entia non sunt multiplicanda sine necessitate.
>
> *William of Ockham (1285 - 1349)*

- Machine Learning Interpretation:
  - Simple concepts are better
- (Debatable) Justifications:
  - There are more complex theories than simple theories, so that a simple theory is less likely to explain the data
  - Simpler theories are easier to falsify
- Empirically, we know that simpler theories perform better (overfitting)

Kolmogorov Complexity

(Kolmogorov 1963)
(Li and Vitanyi 1997)

Kolmogorov Complexity of an object X is the length of the shortest program that produces X as its output

- measure for information contained in a bit string, but – unlike Shannon's information content – takes patterns into account

M_1: `0 0 0 0 0 0 0 0 0 0 0 0`

$IC(M_1) = -1 \cdot \log(1) = 0$

M_2: `0 1 1 1 0 1 0 0 0 1 1 0`

$IC(M_2) = -0.5 \cdot \log(0.5) = 1$

M_3: `0 1 0 1 0 1 0 1 0 1 0 1`

$IC(M_3) = -0.5 \cdot \log(0.5) = 1$

M_2 and M_3 have the same information content, but M_3 has a much lower Kolmogorov complexity (but exact value is hard to compute)

Minimum Description/Message Length Principle

The best hypothesis is the one that compresses the data the most

- Length of data is relative to a hypothesis (program) plus the length of this hypothesis

$$IC(M,H) = -\log(p(M,H)) = -\log(p(M|H)) - \log(p(H))$$

description length of the message given the hypothesis

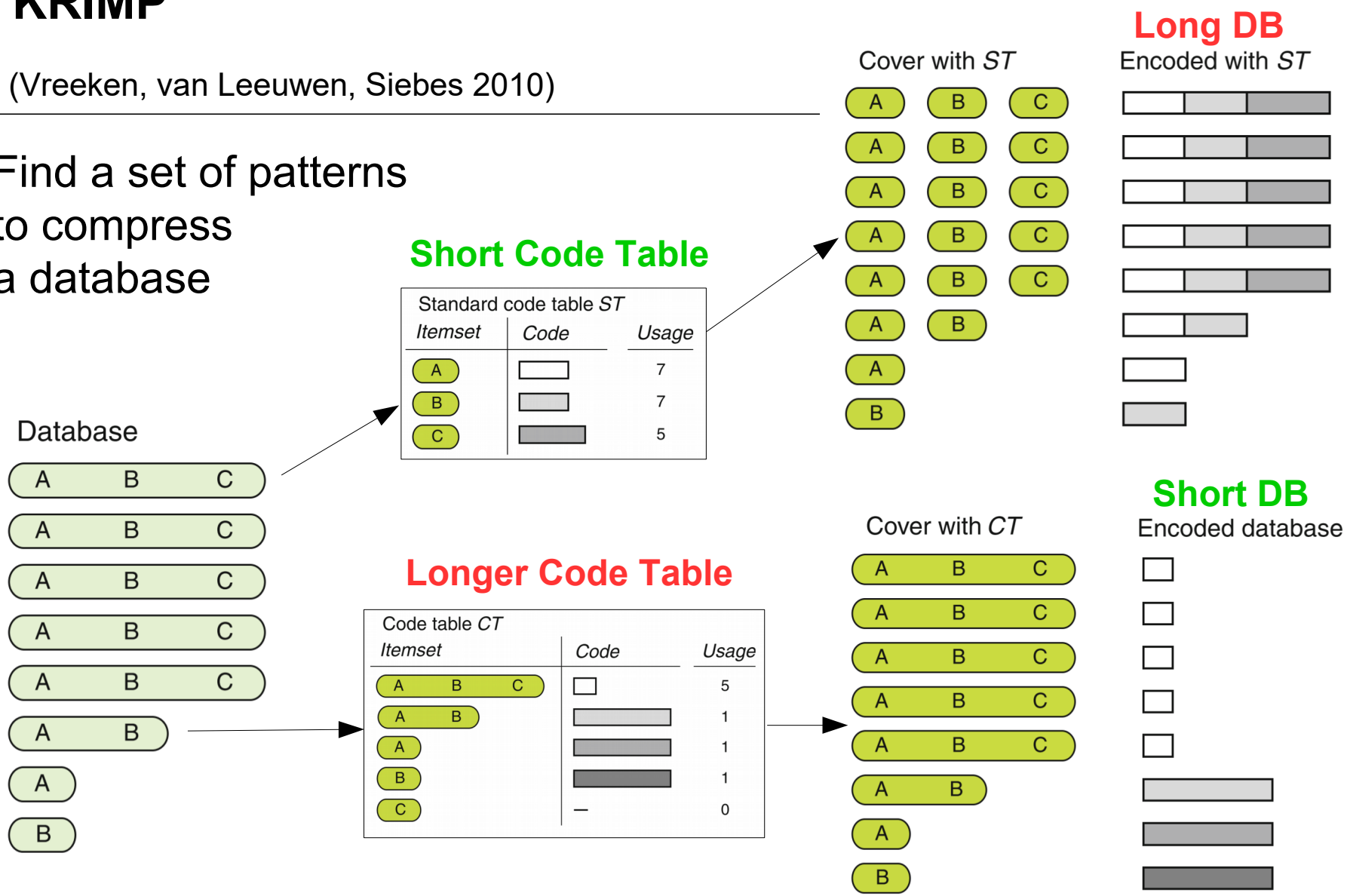
description length of the hypothesis

- may be viewed as a formal generalization of Occam's Razor to hypotheses that do not make the same number of mistakes
- frequently used as selection / pruning criterion in rule learning

KRIMP

(Vreeken, van Leeuwen, Siebes 2010)

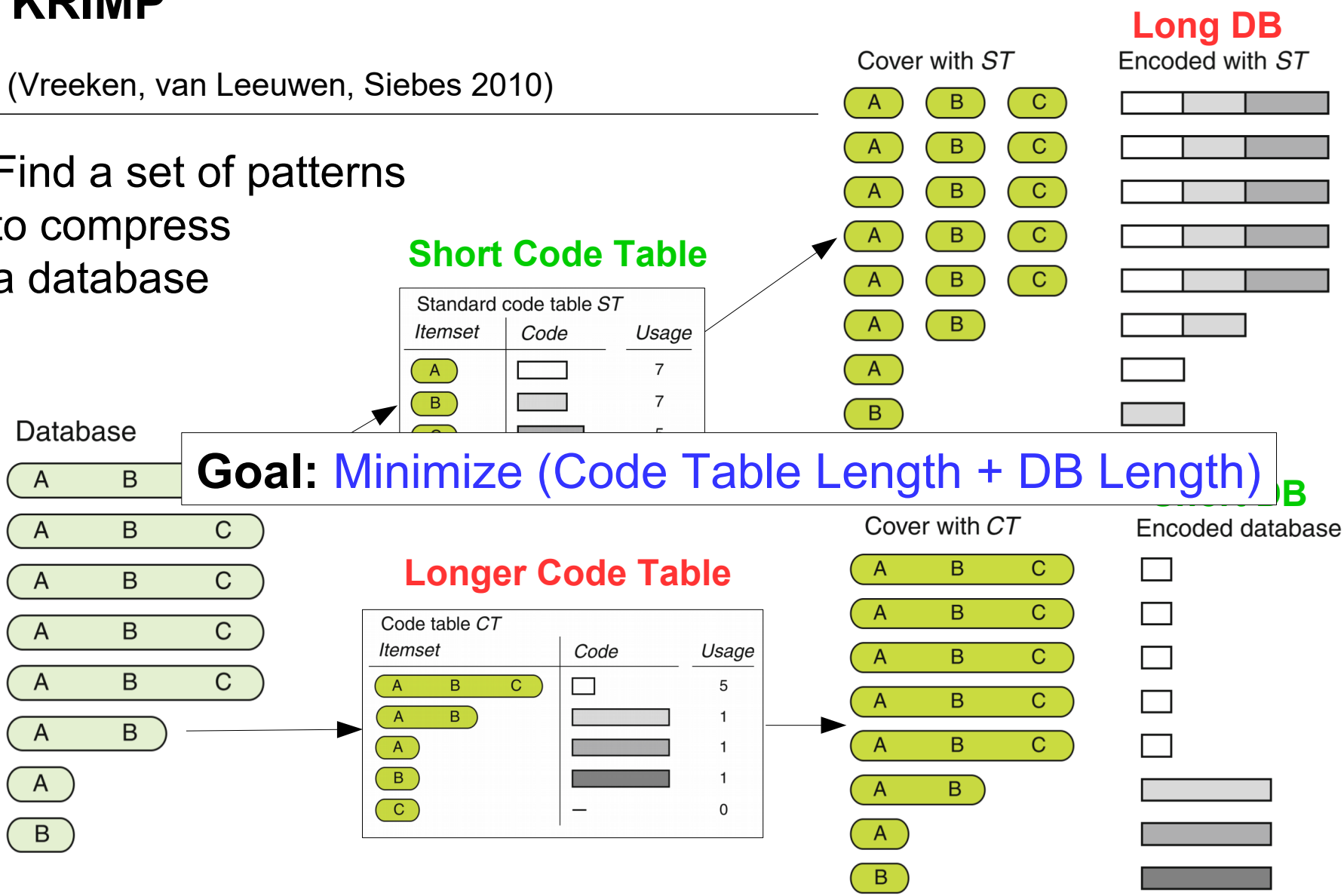
Find a set of patterns
to compress
a database



KRIMP

(Vreeken, van Leeuwen, Siebes 2010)

Find a set of patterns
to compress
a database



MDL and Understandability

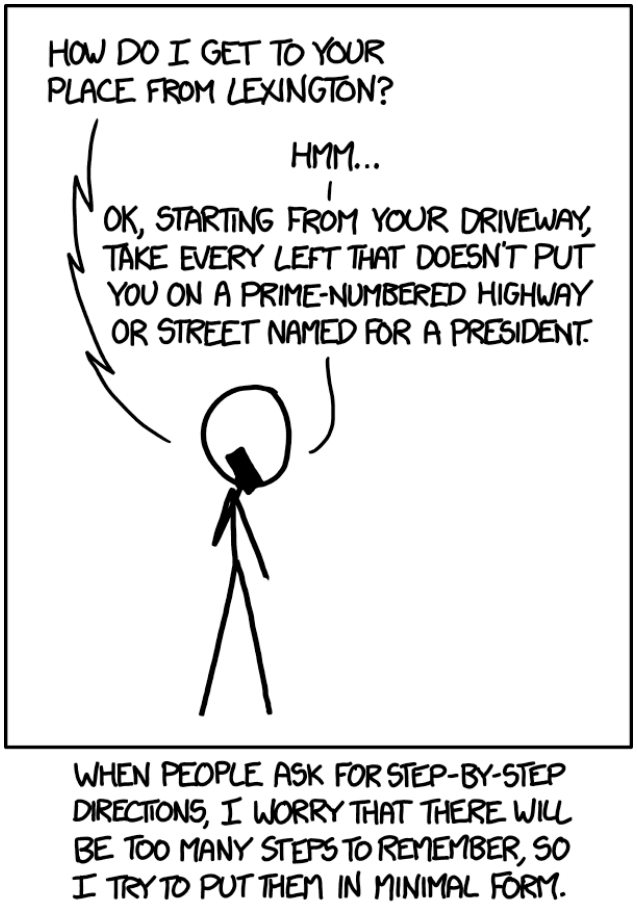
Source: https://www.xkcd.com/1155/
(Thanks to Jilles Vreeken for the pointer)

- Minimum Description Length explanations may be predictive
- but do not need to be interpretable

Other dimensions:

- Representativeness
- Redundancy
- Coherence
- Structure
- ...

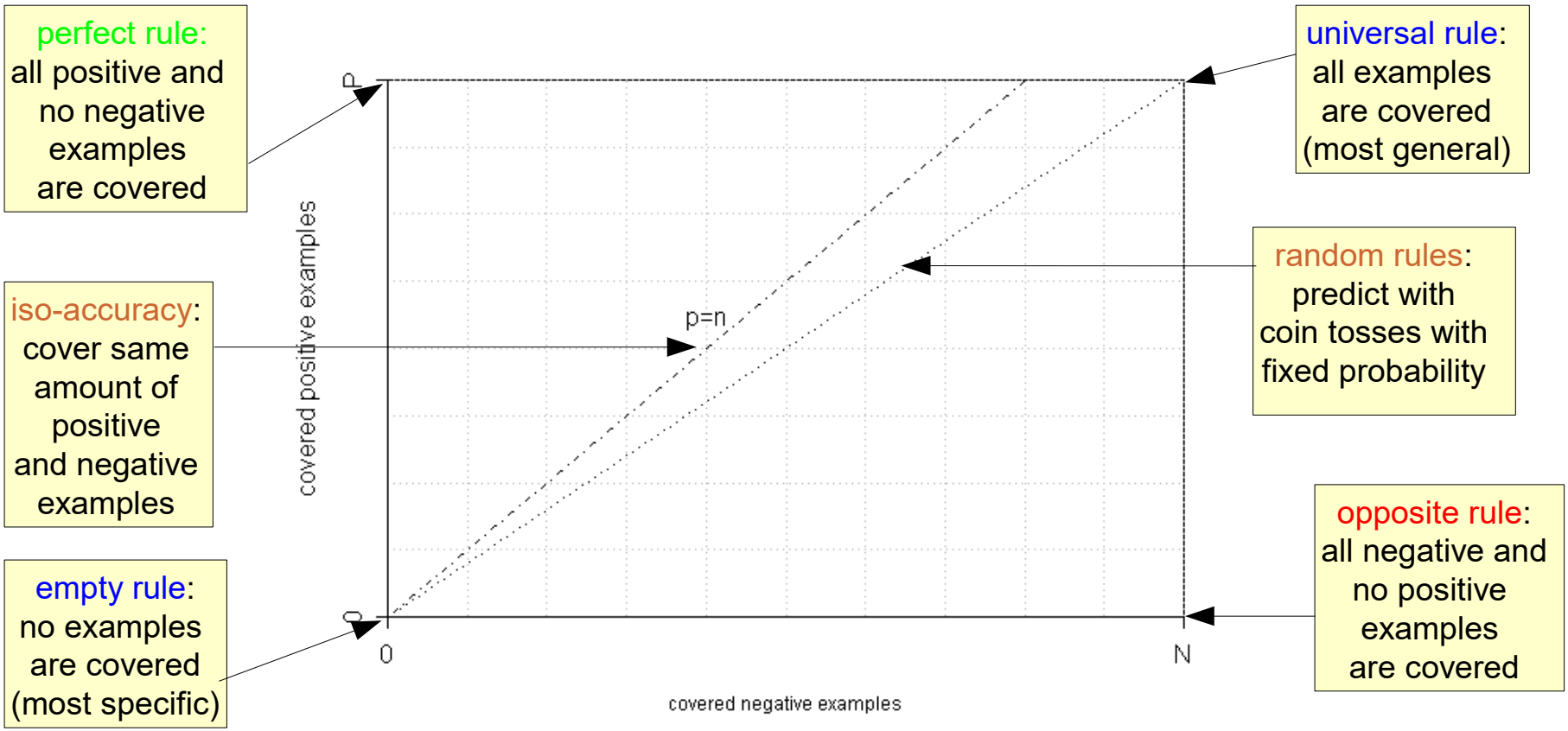
Kolmogorov Directions



Coverage Spaces

(Fürnkranz & Flach 2005)

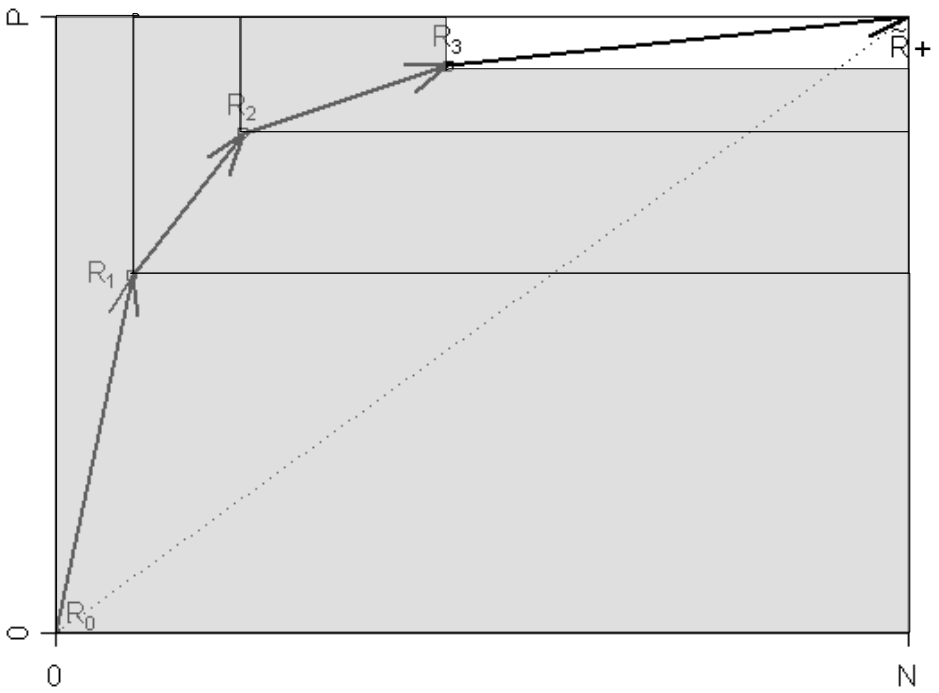
- good tool for visualizing properties of rule evaluation heuristics
  - each point is a rule covering p positive and n negative examples



Rule Selection: Covering Strategy

(survey → Fürnkranz 1999)

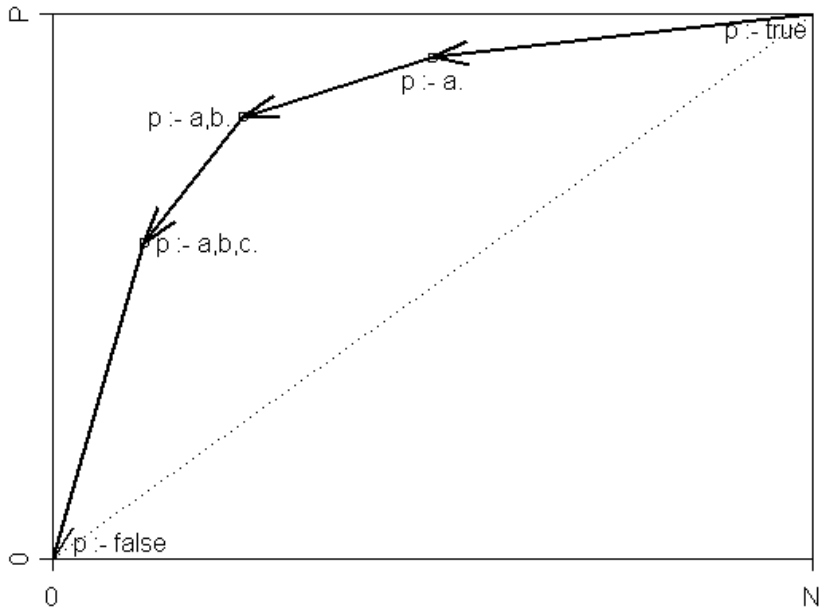
- Covering or Separate-and-Conquer rule learning learning algorithms learn one rule at a time
  - and then removes the examples covered by this rule
- This corresponds to a path in coverage space:
  - The empty theory R_0 (no rules) corresponds to (0,0)
  - Adding one rule never decreases p or n because adding a rule covers *more* examples (generalization)
  - The universal theory R+ (all examples are positive) corresponds to (N,P)



Rule Refinement: Top-Down Hill-Climbing

- successively extends a rule by adding conditions

- This corresponds to a path in coverage space:
  - The rule `p:-true` covers all examples (universal theory)
  - Adding a condition never increases p or n (specialization)
  - The rule `p:-false` covers no examples (empty theory)



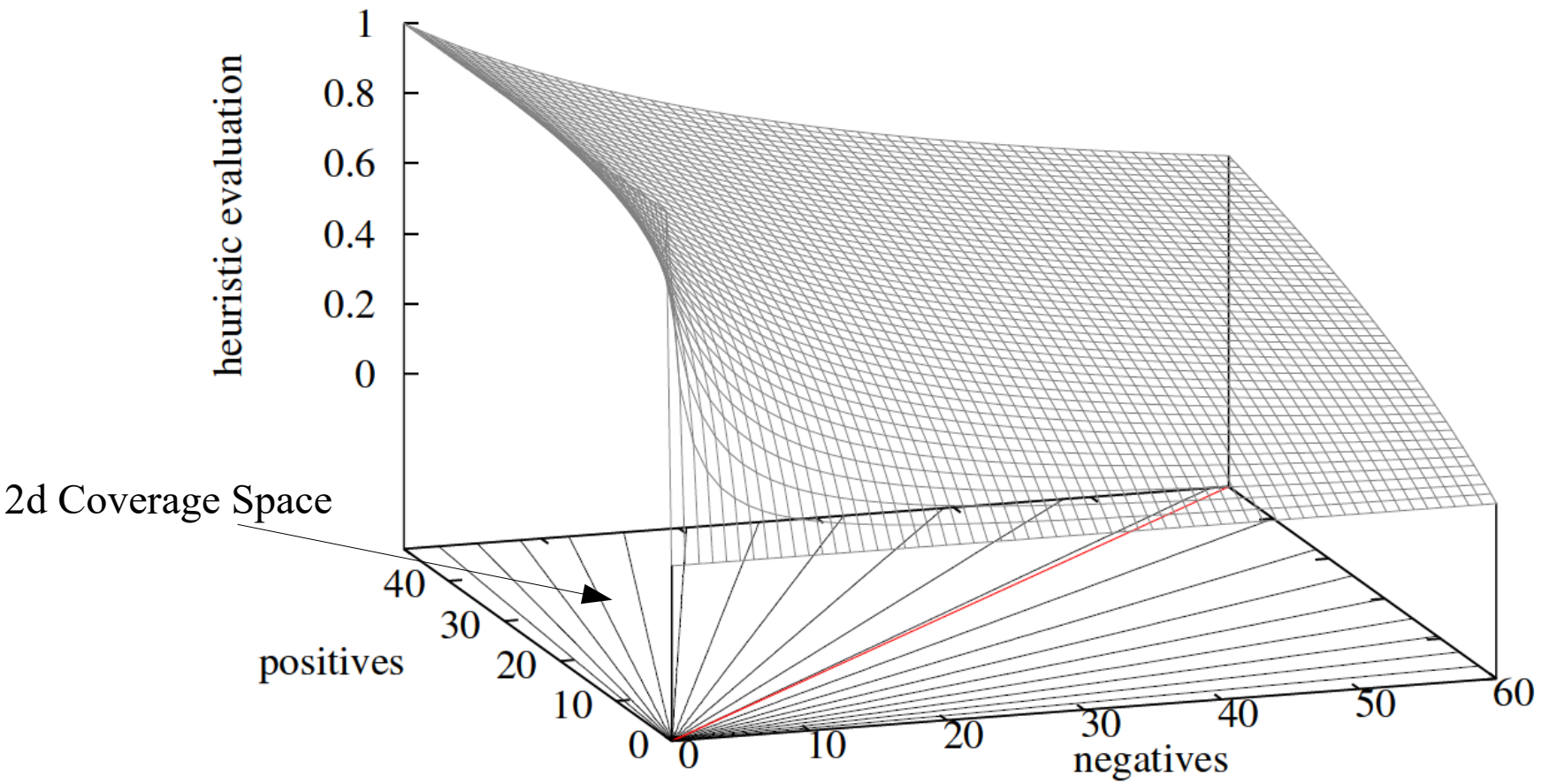
- which conditions are selected depends on a *heuristic function* that estimates the quality of the rule

Rule Learning Heuristics

- How can we measure the quality of a rule?
  - should cover as few negative examples as possible (*consistency*)
  - should cover as many positive examples as possible (*completeness*)

- An evaluation heuristic should therefore trade off these two properties
  - Example: Laplace heuristic $h_{Lap} = \frac{p+1}{p+n+2}$
    - grows with $p \to \infty$
    - grows with $n \to 0$
  - Example: Precision $h_{Prec} = \frac{p}{p+n}$
    - is not a good heuristic. Why?

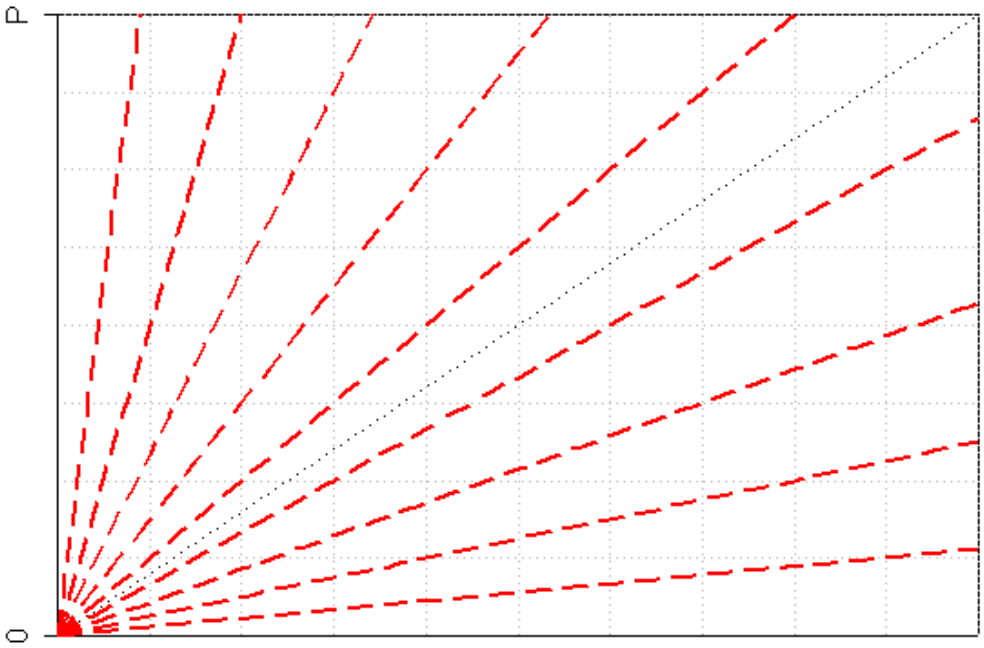
3d-Visualization of Precision



Precision

- *basic idea:*
  - percentage of positive examples among covered examples
- *effects:*
  - rotation around origin (0,0)
  - all rules with same angle equivalent
  - in particular, all rules on P/N axes are equivalent
- typically **overfits**

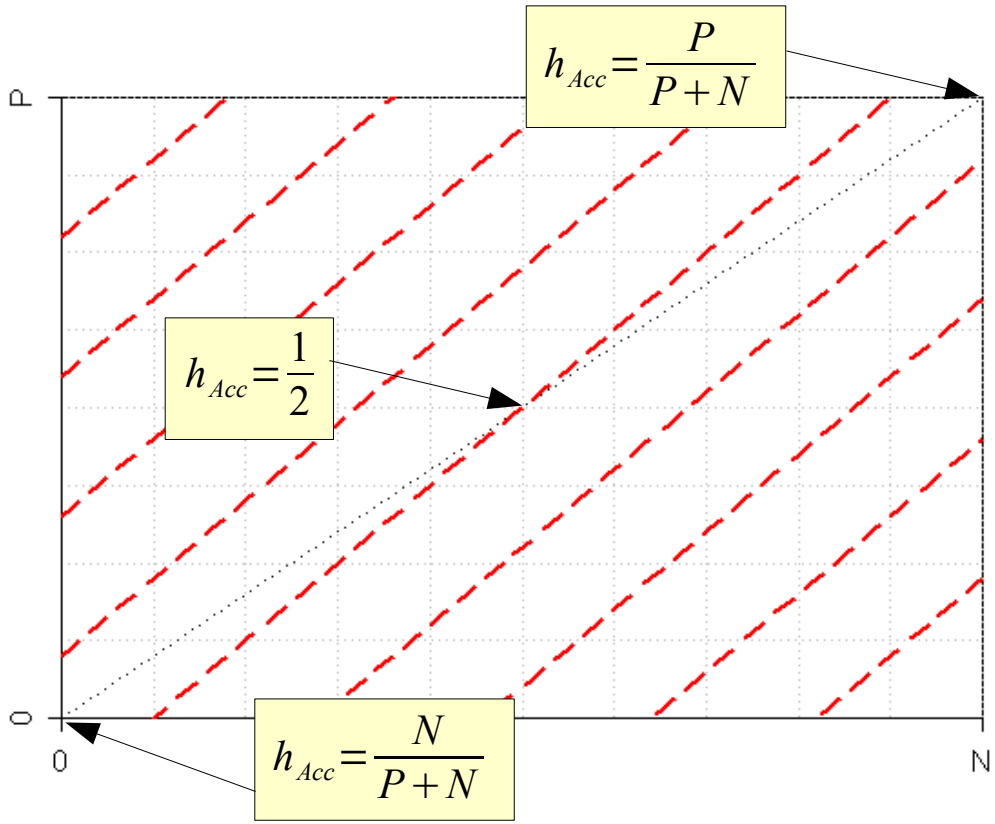
$$h_{Prec} = \frac{p}{p+n}$$



Accuracy

$$h_{Acc} = \frac{p + (N - n)}{P + N} \simeq p - n$$

- *basic idea:*
  percentage of correct classifications (*covered positives* plus *uncovered negatives*)
- *effects:*
  - isometrics are parallel to 45° line
  - covering one positive example is as good as not covering one negative example



Weighted Relative Accuracy

- Two Basic ideas:
  - **Precision Gain:** compare precision to precision of a rule that classifies all examples as positive

$$\frac{p}{p+n} - \frac{P}{P+N}$$

  - **Coverage:** Multiply with the percentage of covered examples

$$\frac{p+n}{P+N}$$

- Resulting formula:

$$h_{WRA} = \frac{p+n}{P+N} \cdot \left(\frac{p}{p+n} - \frac{P}{P+N} \right)$$

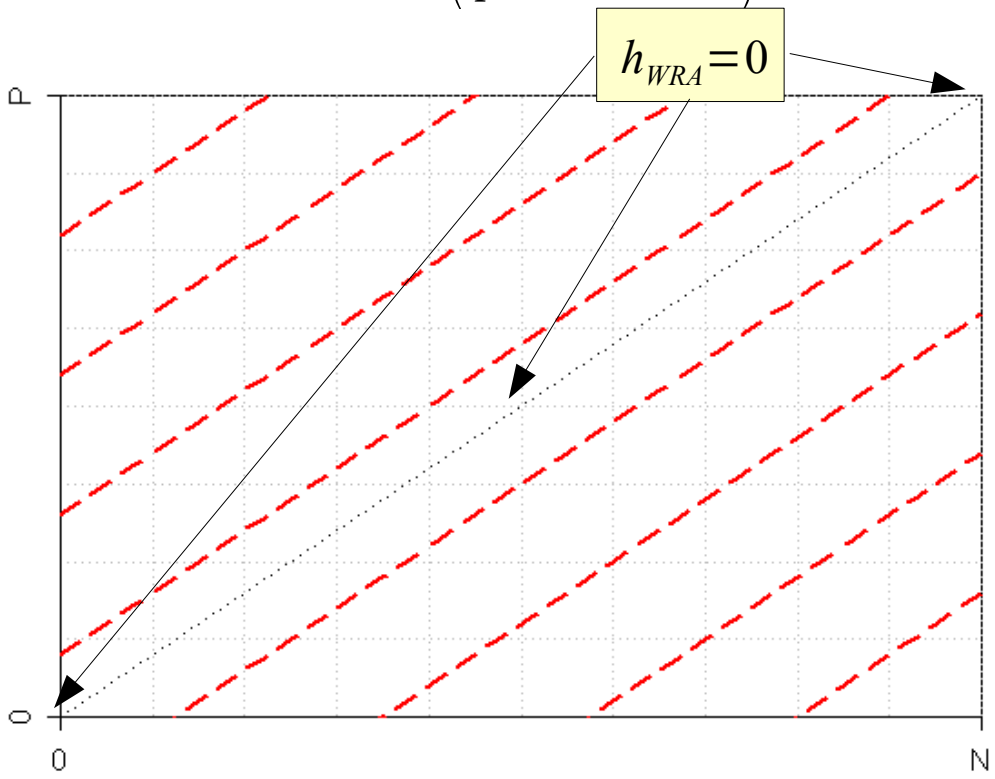
  - one can show that this sorts rules in exactly the same way as

$$h_{WRA}' = \frac{p}{P} - \frac{n}{N}$$

Weighted relative accuracy

- *basic idea:*
  compte the distance from the diagonal (i.e., from random rules)
- *effects:*
  - isometrics are parallel to diagonal
  - covering $x\%$ of the positive examples is considered to be as good as not covering $x\%$ of the negative examples
- typically **over-generalizes**

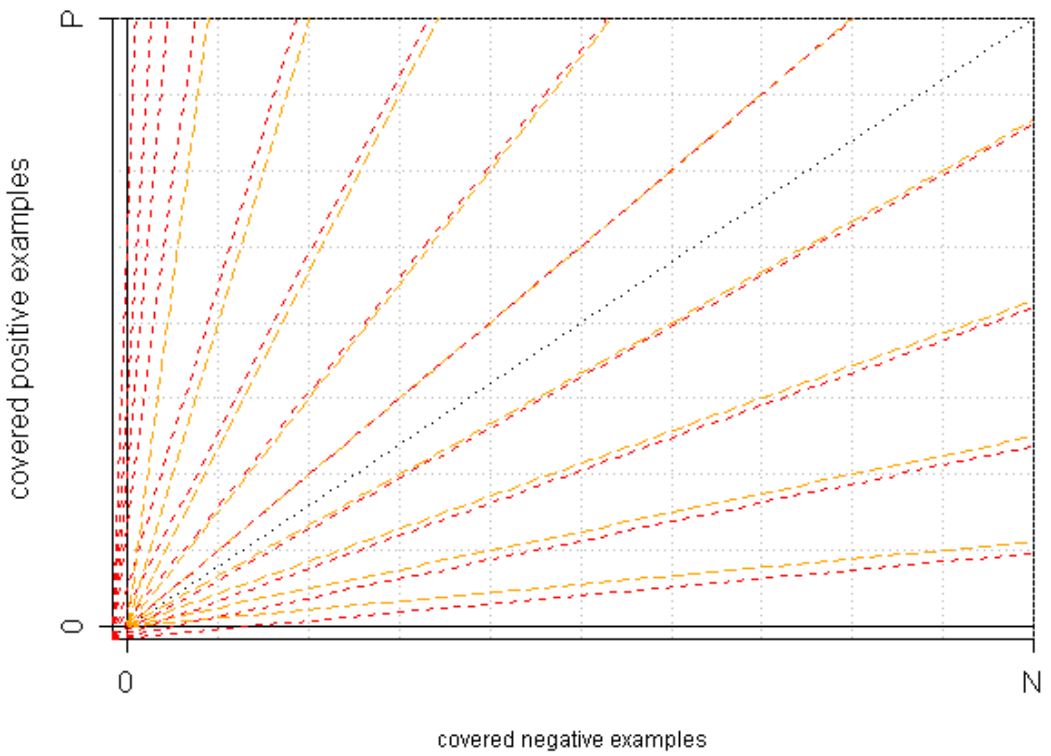
$$h_{WRA} = \frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right) \simeq \frac{p}{P} - \frac{n}{N}$$



Laplace-Estimate

- *basic idea:*
  precision, but count coverage for positive and negative examples starting with 1 instead of 0
- *effects:*
  - origin at $(-1,-1)$
  - different values on $p=0$ or $n=0$ axes
  - not equivalent to precision

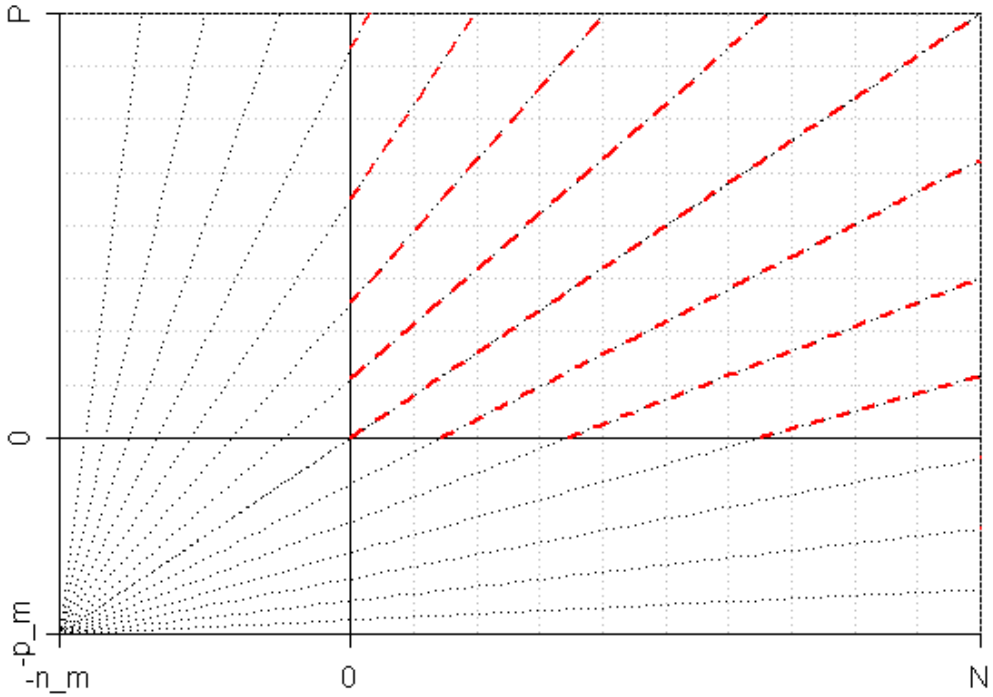
$$h_{Lap} = \frac{p+1}{(p+1)+(n+1)} = \frac{p+1}{p+n+2}$$



m-estimate

- *basic idea:*
  initialize the counts with m examples in total, distributed according to the prior distribution $P/(P+N)$ of p and n.
- *effects:*
  - origin shifts to $(-mP/(P+N), -mN/(P+N))$
  - with increasing m, the lines become more and more parallel
- can be re-interpreted as a **trade-off between WRA and precision/confidence**

$$h_m = \frac{p + m\frac{P}{P+N}}{\left(p + m\frac{P}{P+N}\right) + \left(n + m\frac{N}{P+N}\right)} = \frac{p + m\frac{P}{P+N}}{p + n + m}$$



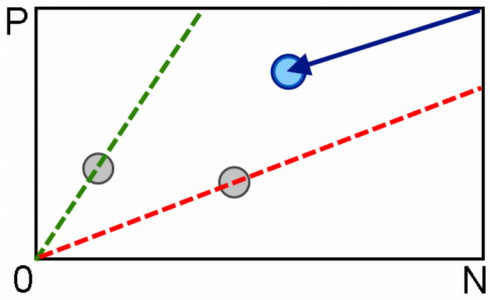
Overview

- Motivation
  - Understandability has not received much attention
- Understandability
  - Conjunctive Fallacy
  - Gambler's Fallacy
  - Representativeness Heuristic
- Different Types of Rules
  - Discriminative vs. Characteristic Rules
  - Formal Concepts
  - Closed Itemsets
- Heuristic Rule Learning
  - Concept Learning
  - Coverage Spaces
  - Rule Learning Heuristics
- Inverted Heuristics
- Explain-A-LOD
  - Semantic Coherence
  - Representation Heuristics
- Algorithmic Enhancements
  - Structured theories
  - More complex problems
- Conclusions

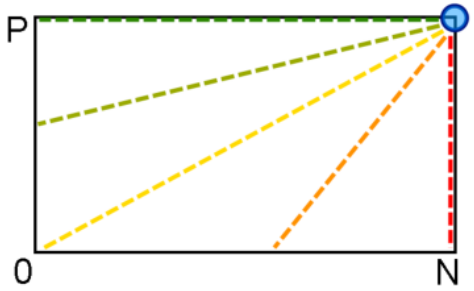
Inverted Heuristics – Motivation

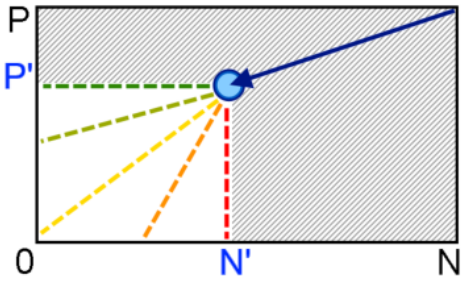
(Stecher, Janssen, Fürnkranz 2014)

- While the search proceeds top-down
- the evaluation of refinements happens from the point of view of the origin (bottom-up)



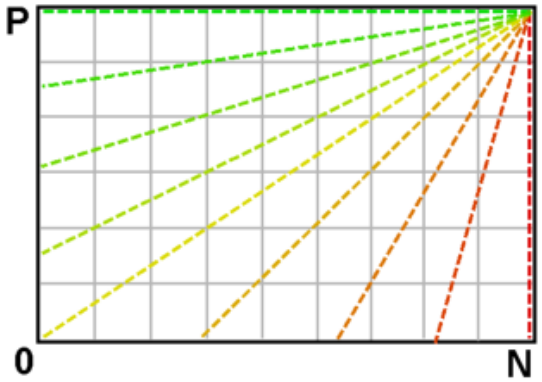
- Instead, we want to evaluate the refinement from the point of view of the predecessor

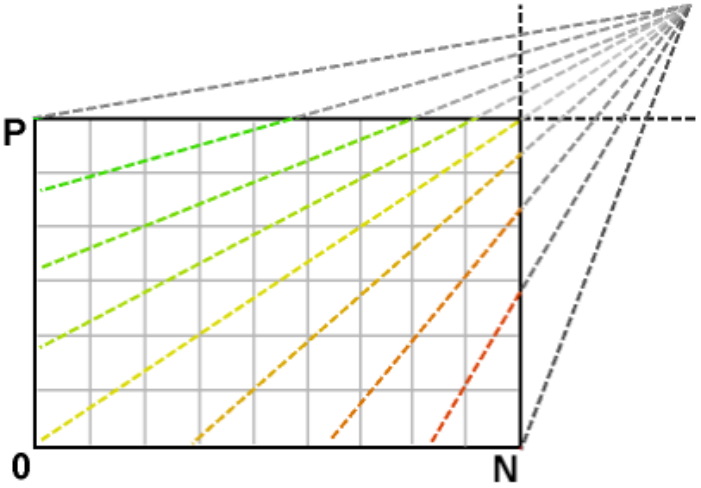




Inverted Heuristics

- Many heuristics can be “inverted” by replacing changing their angle point from the origin to the current rule





$$h'_{precision}(p,n,P,N) = \frac{N-n}{(P+N)-(p+n)}$$

$$h'_{m-Estimate}(p,n,P,N) = \frac{N-n+m\cdot\frac{P}{P+N}}{(P+N)-(p+n-m)}$$

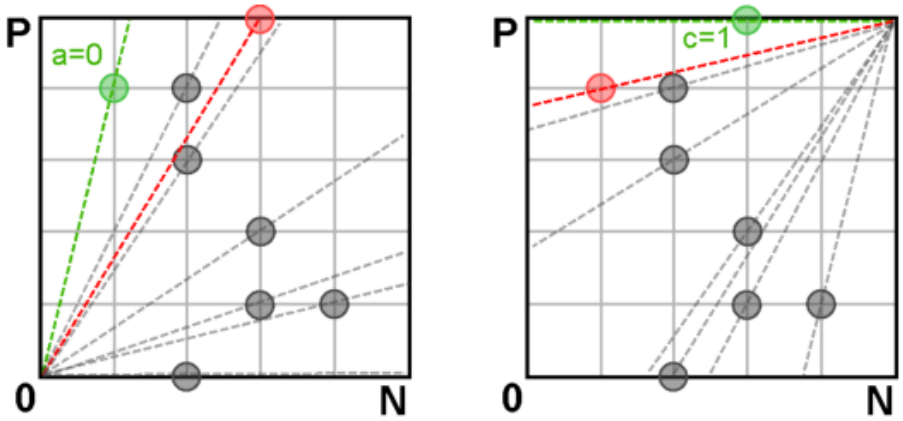
- **Note:** not all heuristics can be inverted
  - e.g. WRA is invariant w.r.t. inversion (because of symmetry)

Inverted Heuristics – Example

First refinement step in small example dataset

- 4 Attributes, 10 data points, binary-class

a	b	c	d	*C*
0	1	1	1	+
0	1	1	1	+
0	0	1	0	-
1	1	1	0	-
1	0	0	1	-
0	1	1	0	+
0	0	1	1	+
1	1	1	0	-
1	0	1	1	+
1	0	0	1	-

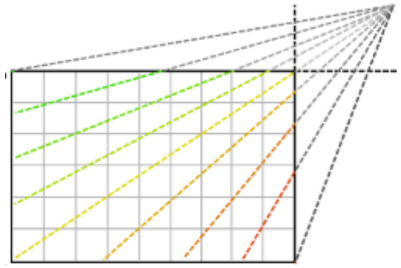


Inverted heuristic function (right image) selects preferable refinement condition $c=1$ with coverage of $(p,n)=(5,3)$

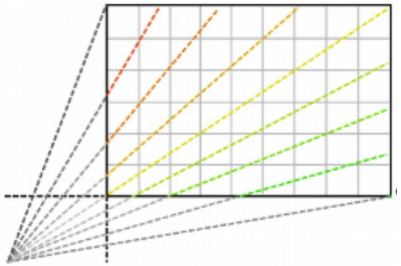
Implementation

- Modification of a conventional covering algorithm
  - CN2-like
  - No pruning, no significance test

- Rule refinement proceeds with inverted heuristics
  - In each iteration, the best condition is added to the rule until the rule covers no more examples

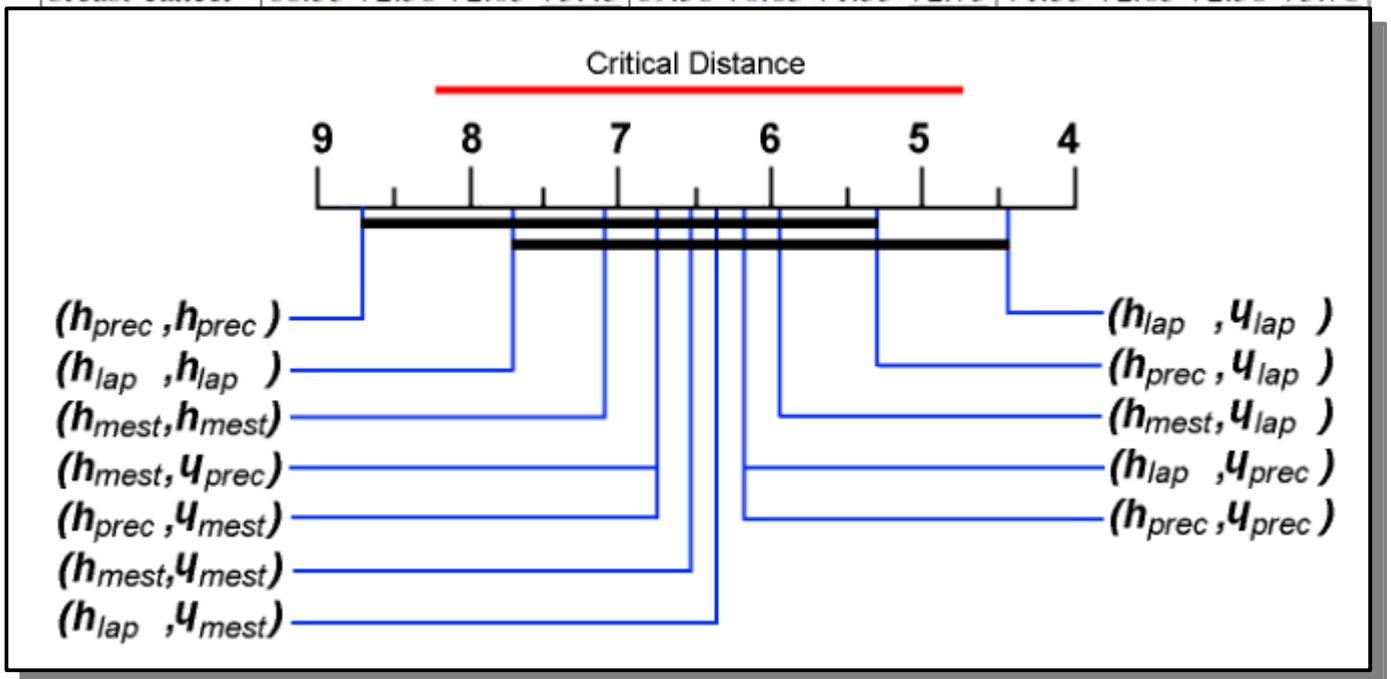


- Rule selection proceeds with regular heuristics
  - Among all refinements on the path, the best rule is selected using a regular heuristic



Results:
Inverted heuristics tend to work better

Dataset	$(h_{prec}, .)$				$(h_{lap}, .)$				$(h_{mest}, .)$			
	h_{prec}	u_{prec}	u_{lap}	u_{mest}	h_{lap}	u_{prec}	u_{lap}	u_{mest}	h_{mest}	u_{prec}	u_{lap}	u_{mest}
breast-cancer	68.53	72.38	72.03	73.43	69.58	70.63	71.33	72.73	71.33	72.03	72.38	73.78
tic-tac-toe	97.39	98.02	97.60	97.81	97.60	98.02	97.60	97.91	98.12	98.02	97.60	97.81
vote	94.94	93.56	94.25	94.48	95.40	94.25	94.25	94.94	93.33	93.56	94.71	96.09
zoo	84.16	88.12	92.08	90.01	86.14	88.12	92.08	90.10	89.11	88.12	92.08	90.10
average rank	3.075	2.400	1.975	2.550	3.000	2.500	1.975	2.525	2.700	2.625	2.225	2.450



Inverted Heuristics – Rule Length

- Inverted Heuristics tend to learn longer rules
  - If there are conditions that can be added without decreasing coverage on the positive examples, inverted heuristics will add them first (before adding discriminative conditions)

Dataset	(h_{lap}, h_{lap}) R	(h_{lap}, h_{lap}) L	(h_{lap}, h'_{lap}) R	(h_{lap}, h'_{lap}) L	Dataset	(h_{lap}, h_{lap}) R	(h_{lap}, h_{lap}) L	(h_{lap}, h'_{lap}) R	(h_{lap}, h'_{lap}) L
breast-cancer	25	67	38	173	ionosphere	17	25	8	42
car	107	495	107	506	labor	5	7	3	12
contact-lenses	5	14	5	15	lymphography	18	42	11	47
futebol	4	7	2	5	monk3	13	38	11	32
glass	50	103	14	83	mushroom	11	13	7	35
hepatitis	13	26	7	46	primary-tumor	80	319	72	518
horse-colic	44	114	19	111	soybean	62	134	45	195
hypothyroid	27	65	9	69	tic-tac-toe	22	84	16	69
iris	7	15	5	17	vote	13	48	12	58
idh	4	5	2	5	zoo	19	19	6	14
averages						28.2	85.6	20.6	106.2

Example: Mushroom dataset

- The best three rules learned with conventional heuristics

```
poisonous :- odor = foul.                (2160,0)
poisonous :- gill-color = buff.          (1152,0)
poisonous :- odor = pungent.              (256,0)
```

- The best three rules learned with inverted heuristics

```
poisonous :- veil-color = white, gill-spacing = close,
             no bruises, ring-number = one,
             stalk-surface-above-ring = silky.   (2192,0)
poisonous :- veil-color = white, gill-spacing = close,
             gill-size = narrow, population = several,
             stalk-shape = tapering.               (864,0)
poisonous :- stalk-color-below-ring = white,
             ring-type = pendant, ring-number = one,
             stalk-color-above-ring = white,
             cap-surface = smooth, stalk-root = bulbuous,
             gill-spacing = close.                 (336,0)
```

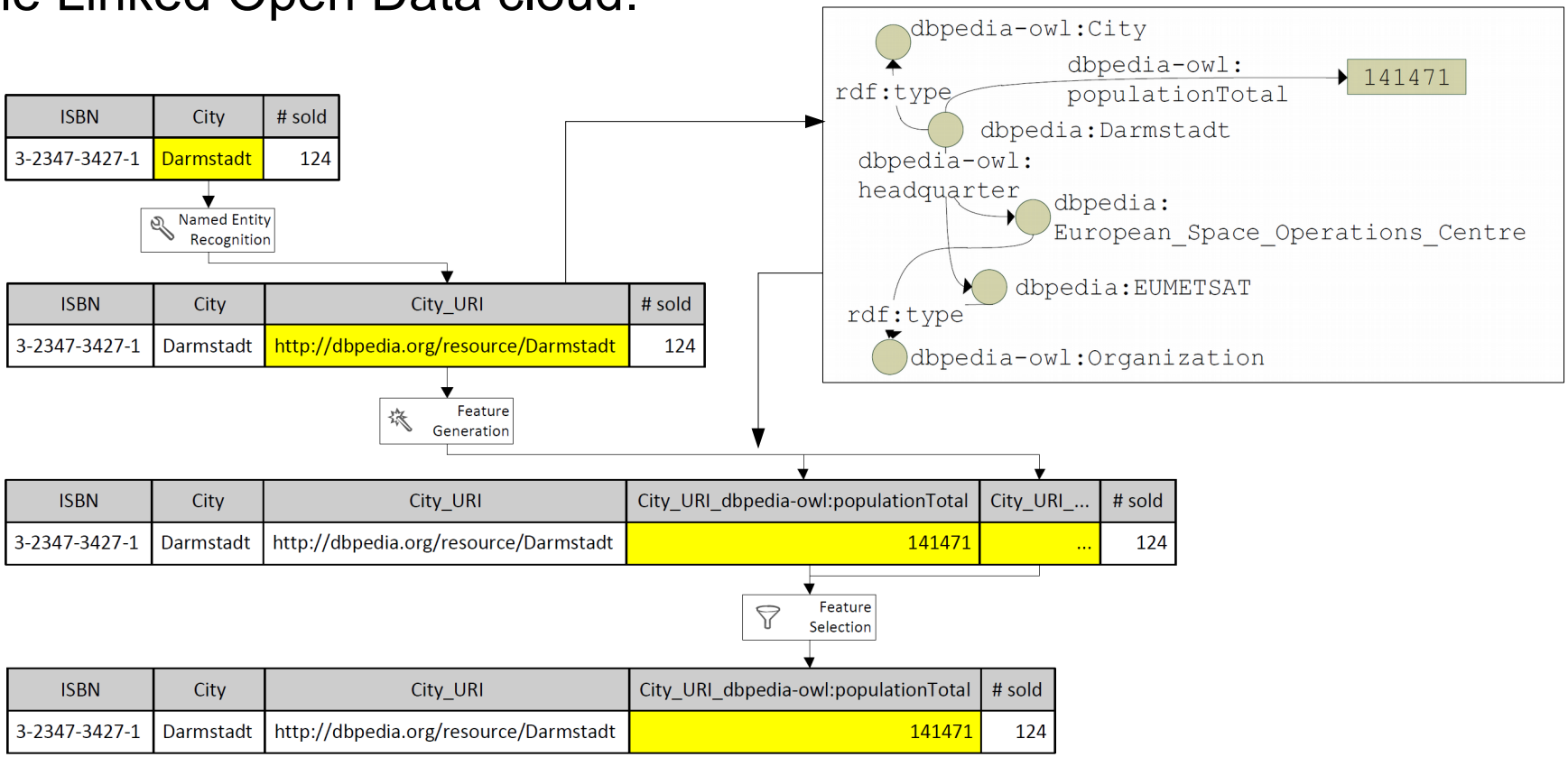
Overview

- Motivation
  - Understandability has not received much attention
- Understandability
  - Conjunctive Fallacy
  - Gambler's Fallacy
  - Representativeness Heuristic
- Different Types of Rules
  - Discriminative vs. Characteristic Rules
  - Formal Concepts
  - Closed Itemsets
  - Occam's Razor & MDL
- Heuristic Rule Learning
  - Concept Learning
  - Coverage Spaces
  - Rule Learning Heuristics
  - Inverted Heuristics
- Understandability of Rules
  - Rule Length
  - Semantic Coherence
  - Recognition Heuristic
  - Relevance
  - Structure
- Conclusions

Explain-A-LOD

(Paulheim & Fürnkranz 2012)

- Generates features for data mining using features derived from the Linked Open Data cloud.

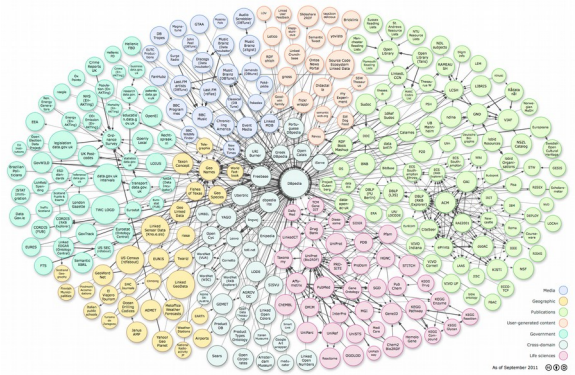


Zero-Knowledge Data Mining

(Paulheim 2012)

Mine a database without explicit background knowledge

City	Country	Index 2010
Vienna	Austria	108.6
Zürich	Switzerland	108.0
Auckland	New Zealand	107.4
Munich	Germany	107.0
Vancouver	Canada	107.4
Düsseldorf	Germany	107.2
Frankfurt	Germany	107.0
Geneva	Switzerland	107.9
Copenhagen	Denmark	106.2
Sydney	Australia	106.3



LOD

Quality-of-living Index

```
QOL = High :-
  European capital of culture
```

Some more rules

(Paulheim 2012)

Good discriminative rules, highly rated by users:

- `QOL = High :- Many events take place.`
- `QOL = High :- Host City of Olympic Summer Games.`
- `QOL = Low  :- African Capital.`

Good discriminative rules, but lowly rated by users:

- ```
  QOL = High :- # Records Made >= 1,
                # Companies/Organisations >= 22.
  ```
- ```
  QOL = High :- # Bands >= 18,
                # Airlines founded in 2000 > 1.
  ```
- ```
  QOL = Low  :- # Records Made = 0,
                Average January Temp <= 16.
  ```
  ```

Is Rule Length an Indicator for Interpretability?

(Kliegr & Fürnkranz 2017)

- Crowd-Sourcing experiment with rules in 4 domains

Quality of Living

if the country falls into all of the following groups simultaneously

* European Union Member Economies **and**

* Imperial free cities

then the quality of living is **highest**

Movie Ratings

if the movie falls into all of the following group(s) (simultaneously)

* Films Released in 2005 **and**

* Englishlanguage Films

then the movie is rated as **good**

Traffic Accidents

if the country falls into all of the following groups simultaneously

* States And Territories Established In 2006 **and**

* Serbo-Croatian-Speaking Countries

then the risk of traffic accidents is **low**

Mushrooms

if the mushroom falls into all of the following groups simultaneously

veil color is white ***and***

stalk surface below ring is *silky*

then the mushroom is **poisonous**

Is Rule Length an Indicator for Interpretability?

(Kliegr & Fürnkranz 2017)

```
Rule 1: if the movie falls into all of the following group(s)
  (simultaneously)

    Englishlanguage Films

then the movie is rated as bad

Rule 2: if the movie falls into all of the following group(s)
 (simultaneously)

    Englishlanguage Films and
    Films Released In 2005

then the movie is rated as bad

Which of the rules do you find as more plausible?
```

Is Rule Length an Indicator for Interpretability?

(Kliegr & Fürnkranz 2017)

Result:

- in two out of four domains there was no correlation
- in the other two longer rules were considered to be more plausible

dataset	units	judg	qfr [%]	Kendall's τ		Spearman's ρ	
Traffic	80	412	12	0.05	(0.226)	0.06	(0.230)
Quality	36	184	11	**0.20**	(0.002)	**0.23**	(0.002)
Movies	32	156	14	−0.01	(0.837)	−0.02	(0.828)
Mushroom	10	250	14	**0.37**	(0.000)	**0.45**	(0.000)
total	158	962	13				

→ no evidence that shorter rules are better understood

Semantic Coherence

Rule discovery algorithms only check the discriminative power of a condition to be added

- First world / Third world would be a plausible distinction
- A distinction based on latitude is less plausible
- A distinction based on number of records made even less plausible

→ conditions that may cover the same examples may have a different “degree of understandability”.

Similarly, combinations of conditions that are semantically far, do not appear to be plausible.

- Number of records made and number of companies are coherent
- Number of companies and average temperature are not coherent

Recognition Heuristic

(Gigerenzer & Todd 1999)

Which of the two cities is larger?

Chongqing

Chengdu

Recognition Heuristic

(Gigerenzer & Todd 1999)

Which of the two cities is larger?

Hongkong

Chengdu

Recognition Heuristic

(Gigerenzer & Todd 1999)

"if one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion"

Hongkong		ca. 7,000,000
Chengdu		ca. 15,000,000
Chongqing		ca. 30,000,000

Recognition Results

(Kliegr & Fürnkranz 2017)

Measuring Recognition:

- **Idea:** The more central a concept is in a knowledge graph, the more likely it is to be recognized → use page rank

Results: (correlation of page rank with plausibility)

Dataset	Min		Avg		Max	
Quality	0.11	(0.048)	0.01	(0.882)	0.07	(0.213)
Movies	**0.22**	(0.000)	-0.12	(0.051)	-0.07	(0.275)
Traffic	-0.03	(0.471)	0.03	(0.533)	0.05	(0.195)

- weak relevance of minimum page rank among all conditions of the rule in at least one dataset

→ it is good if all conditions are well recognized

Relevance

- Obtained additional information about relevance of literals and attributes

We kindly ask you to assist us in an experiment that will help researchers understand which properties influence mushroom being considered as poisonous/edible.

Example task follows:

Property: Cap shape

Possible values: bell, conical, convex, flat, knobbed, sunken

What is the relevance of the property given above for determining whether a mushroom is edible or poisonous?

Give a judgement on a 10 point scale, where:

1 = Completely irrelevant
10 = Very relevant

Obtaining further information If the meaning of one of the properties is not clear, you can try looking it up in Wikipedia.

We kindly ask you to assist us in an experiment that will help researchers understand which factors can influence movie ratings.

Example task follows:

Condition: Academy Award Winner or Nominee

The condition listed above will contribute to a movie being rated as:

Good (Strong influence)
Good (Weak influence)
No influence
Bad (Weak influence)
Bad (Strong influence)

Select one option.

Obtaining further information

If the meaning of one of the conditions is not clear, you can click on the condition to see explanation in Wikipedia.

For example, consider condition "Obtaining XYZ award."
If you are not sure what exactly award XYZ is, you should click on the link to consult the Wikipedia article.

Thank you for your assistance !

Relevance Results

- Relevance is relevant
  - but quite subjective and domain-dependent

Dataset	Min		Avg		Max	
			Literal Relevance			
Quality	−**0.24**	(0.000)	**0.29**	(0.000)	**0.31**	(0.000)
Movies	−0.11	(0.072)	0.15	(0.012)	**0.22**	(0.000)
Traffic	−0.04	(0.377)	0.04	(0.311)	0.01	(0.797)
Mushroom	**0.22**	(0.000)	−**0.19**	(0.000)	0.11	(0.037)
			Attribute Relevance			
Traffic	−0.01	(0.745)	0.01	(0.757)	0.00	(0.983)
Mushroom	**0.30**	(0.000)	−0.11	(0.018)	**0.27**	(0.000)

Overview

- Motivation
  - Understandability has not received much attention
- Understandability
  - Conjunctive Fallacy
  - Gambler's Fallacy
  - Representativeness Heuristic
- Different Types of Rules
  - Discriminative vs. Characteristic Rules
  - Formal Concepts
  - Closed Itemsets
- Heuristic Rule Learning
  - Concept Learning
  - Coverage Spaces
  - Rule Learning Heuristics
- Inverted Heuristics
- Explain-A-LOD
  - Semantic Coherence
  - Representation Heuristics
- Algorithmic Enhancements
  - Structured theories
  - More complex problems
- Conclusions

Structured Concepts

Most rule learning algorithms learn flat theories

- e.g., n-bit parity needs 2^n flat rules

```
+ :-     x1,     x2,     x3,     x4.
+ :-     x1,     x2, not x3, not x4.
+ :-     x1, not x2,     x3, not x4.
+ :-     x1, not x2, not x3,     x4.
+ :- not x1,     x2, not x3,     x4.
+ :- not x1,     x2,     x3, not x4.
+ :- not x1, not x2,     x3,     x4.
+ :- not x1, not x2, not x2, not x4.
```

But structured concepts are often more interpretable

- e.g. only $O(n)$ rules with intermediate concepts

```
+          :- x1, not parity234.
+          :- not x1, parity234.

parity234 :- x2, not parity34.
parity234 :- not x2, parity34.

parity34  :- x3, x4.
parity34  :- not x3, not x4.
```

Previous work in the 90s in inductive logic programming (ILP) and restructuring knowledge bases was not successful

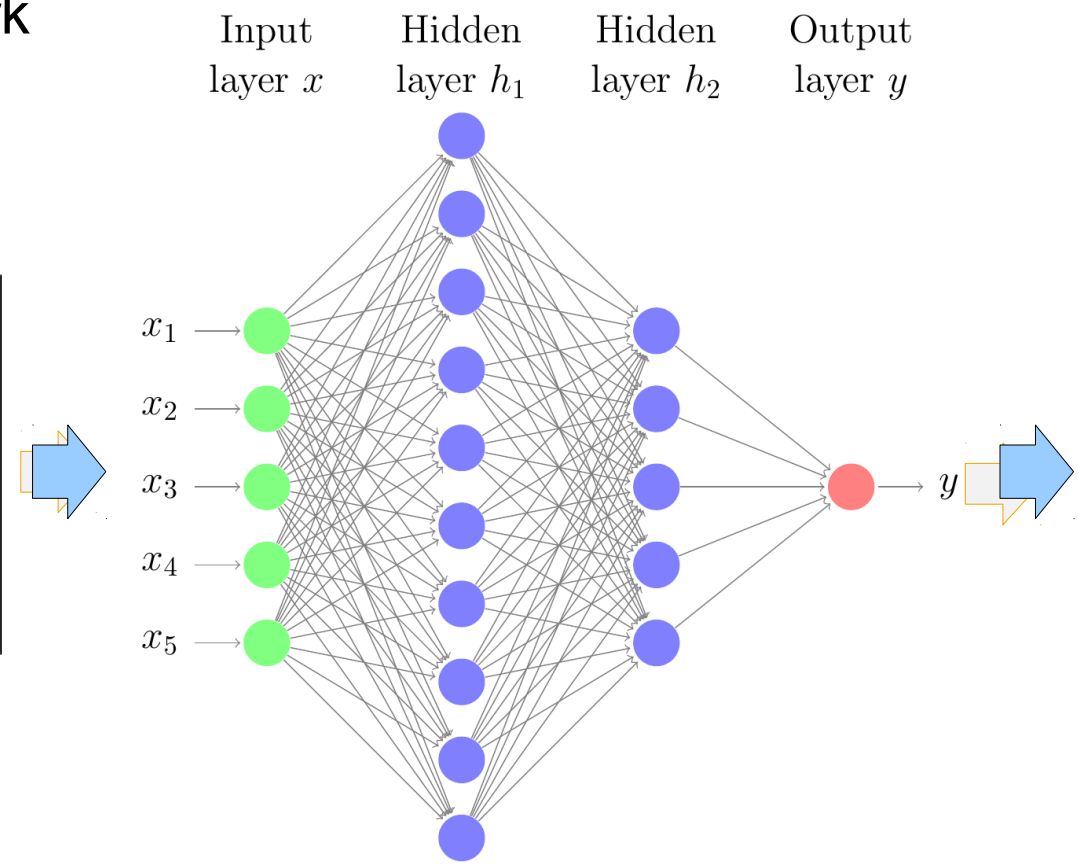
- new approaches could borrow ideas from Deep Learning

Rule Extraction from Neural Networks

Pedagogical Strategy:

- Train a (deep) network

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…



o
1
0
1
1
0
1
1
1
…

DeepRED:
Rule Extraction from Deep Networks

(Zilke, Loza, Janssen 2016)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$	$h_{2,2}$	…	$h_{2,5}$
0.034	0.635	…	0.928
0.089	0.049	…	0.435
0.057	0.369	…	0.233
0.346	0.462	…	0.181
0.834	0.945	…	0.354
0.443	0.644	…	0.892
0.778	0.691	…	0.708
…	…	…	…

o
1
0
1
0
1
1
1
…

Step 1: Propagate activation through network

DeepRED:
Rule Extraction from Deep Networks

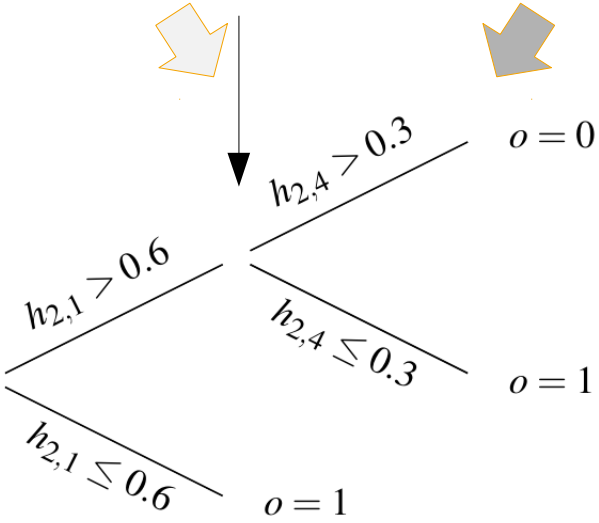
(Zilke, Loza, Janssen 2016)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$	$h_{2,2}$	…	$h_{2,5}$	o
0.034	0.635	…	0.928	1
0.089	0.049	…	0.435	0
0.057	0.369	…	0.233	1
0.346	0.462	…	0.181	0
0.834	0.945	…	0.354	1
0.443	0.644	…	0.892	1
0.778	0.691	…	0.708	1
…	…	…	…	…

Step 2: Find a decision tree that describes an output node using activation values of the previous hidden layer h_i



DeepRED:
Rule Extraction from Deep Networks

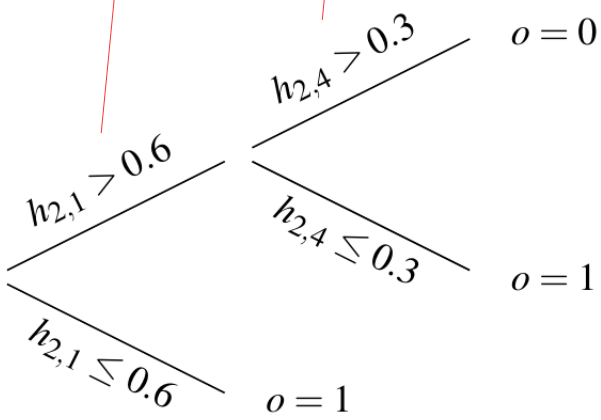
(Zilke, Loza, Janssen 2016)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}>0.3$	$h_{2,1}>0.6$	…	$h_{2,4}>0.3$	o
0	0	…	1	1
0	0	…	1	0
0	0	…	0	1
1	0	…	0	0
1	1	…	1	1
1	0	…	1	1
1	1	…	1	1
…	…	…	…	…

Step 3: Replace target activations h_i by split points on h_i using in prediction model $h_i \rightarrow h_{i+1}$



DeepRED:
Rule Extraction from Deep Networks

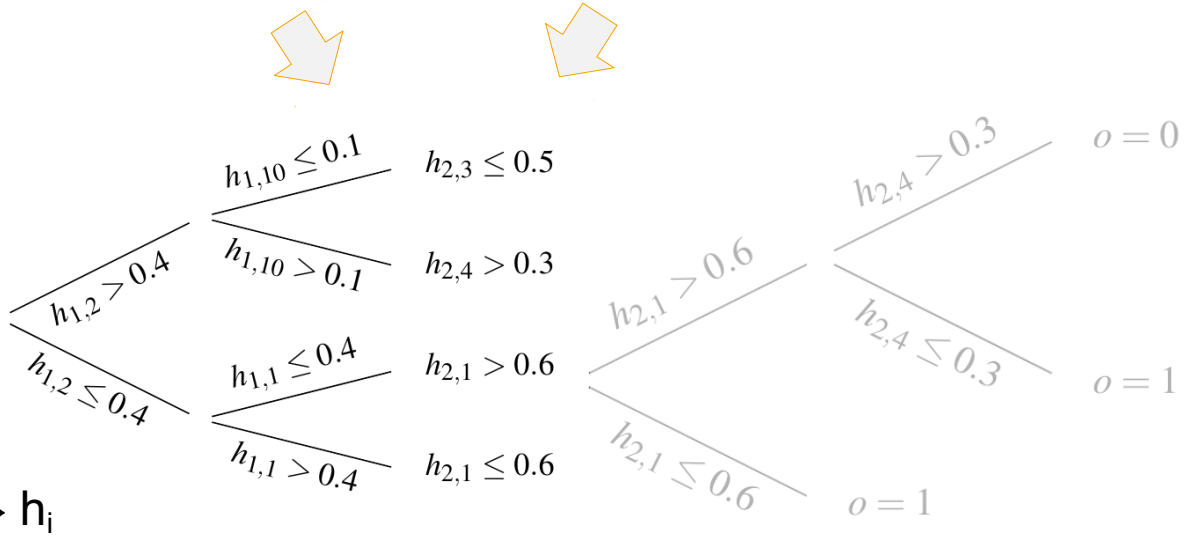
(Zilke, Loza, Janssen 2016)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$>0.3	$h_{2,1}$>0.6	…	$h_{2,4}$>0.3
0	0	…	1
0	0	…	1
0	0	…	0
1	0	…	0
1	1	…	1
1	0	…	1
1	1	…	1
…	…	…	…

o
1
0
1
0
1
1
1
…



Step 4: induce model $h_{i-1} \rightarrow h_i$

DeepRED:
Rule Extraction from Deep Networks

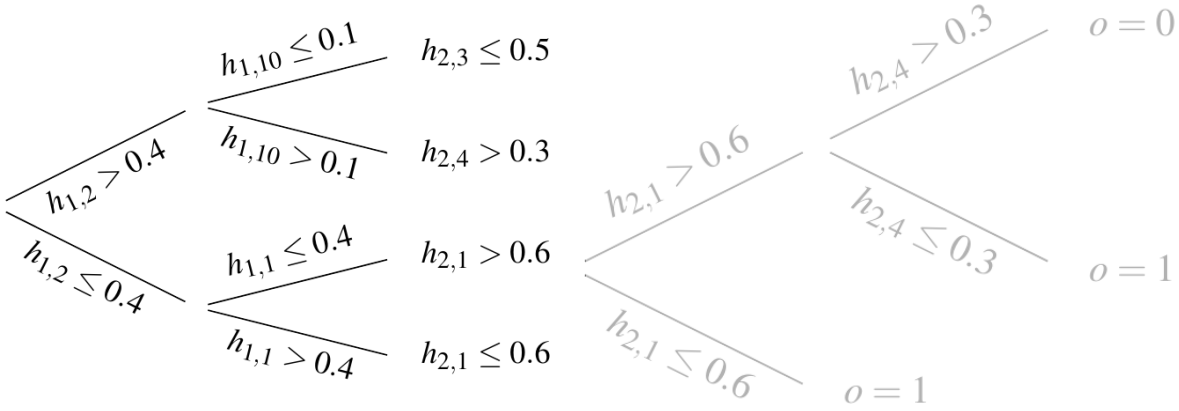
(Zilke, Loza, Janssen 2016)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}$	$h_{1,2}$	…	$h_{1,10}$
0.865	0.079	…	0.818
0.050	0.675	…	0.613
0.767	0.485	…	0.020
0.388	0.160	…	0.491
0.555	0.767	…	0.606
0.312	0.231	…	0.376
0.770	0.211	…	0.805
…	…	…	…

$h_{2,1}$>0.3	$h_{2,1}$>0.6	…	$h_{2,4}$>0.3
0	0	…	1
0	0	…	1
0	0	…	0
1	0	…	0
1	1	…	1
1	0	…	1
1	1	…	1
…	…	…	…

o
1
0
1
0
1
1
1
…



Repeat for all layers until input layer is reached

DeepRED:
Rule Extraction from Deep Networks

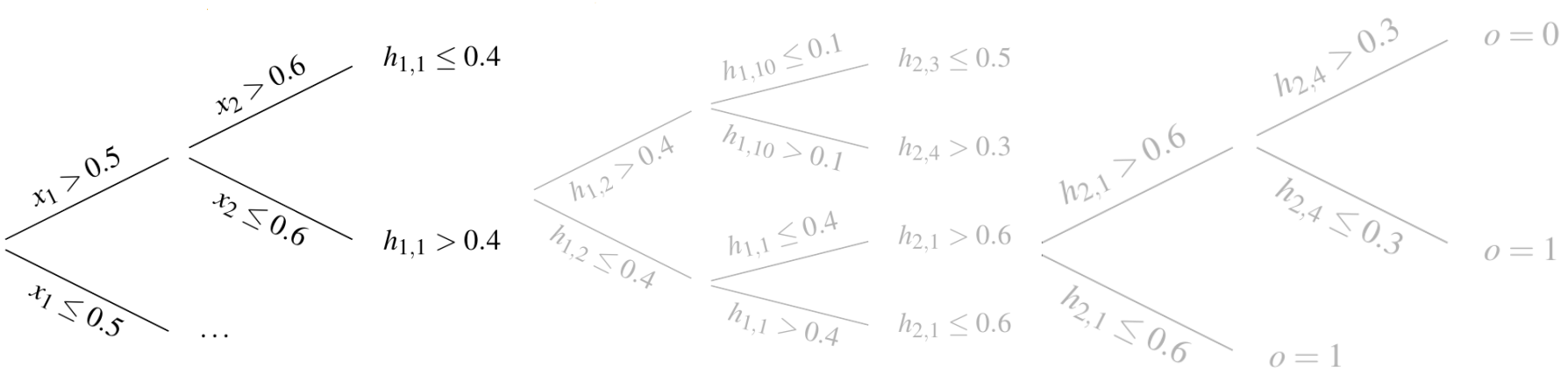
(Zilke, Loza, Janssen 2016)

x_1	x_2	x_3	x_4	x_5
0.5	1	0.200	0.648	0.875
0.5	1	0.197	0.889	0.487
0.5	0.25	0.972	0.754	0.711
0	0.75	0.884	0.580	0.213
0.5	0	0.860	0.795	0.475
1	0.75	0.505	0.905	0.692
1	0.75	0.731	0.084	0.409
…	…	…	…	…

$h_{1,1}>0.4$	$h_{1,2}>0.4$	…	$h_{1,10}>0.1$
1	0	…	1
0	1	…	1
1	1	…	0
0	0	…	1
1	1	…	1
0	0	…	1
1	0	…	1
…	…	…	…

$h_{2,1}>0.3$	$h_{2,1}>0.6$	…	$h_{2,4}>0.3$
0	0	…	1
0	0	…	1
0	0	…	0
1	0	…	0
1	1	…	1
1	0	…	1
1	1	…	1
…	…	…	…

o
1
0
1
0
1
1
1
…

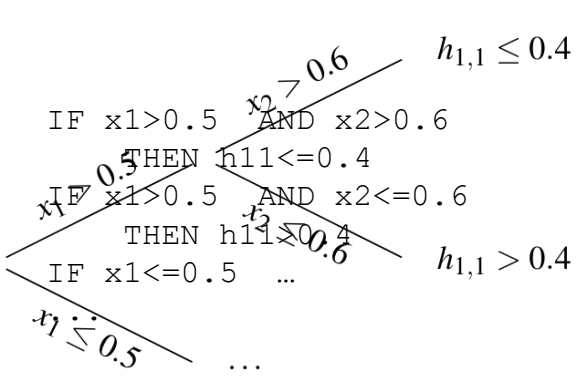


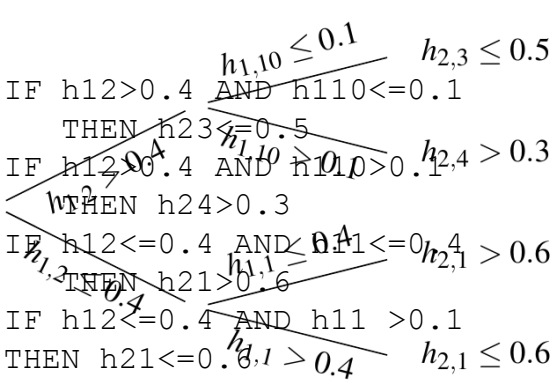
DeepRED: Rule Extraction from Deep Networks

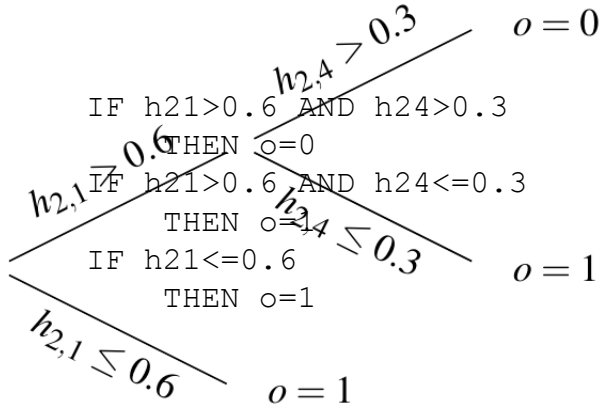
(Zilke, Loza, Janssen 2016)

Represent output as a function of inputs

- Extract rule sets $R(h_{i-1} \rightarrow h_i)$ from decision trees
- Optional: Combine rules into a single rule set
  - Advance layerwise
    - put $R(h_{i-1} \rightarrow h_i)$ into $R(h_i \rightarrow h_o)$ to get $R(h_{i-1} \rightarrow h_o)$
    - delete unsatisfiable and redundant terms







Can DeepRED make use of complex concepts hidden in NNs?

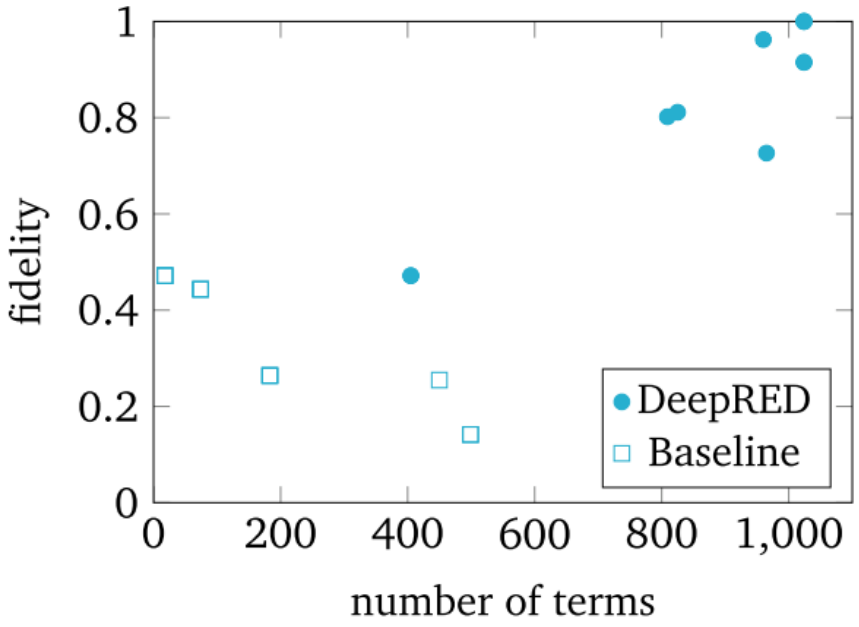
(Zilke, Loza, Janssen 2016)

XOR

- parity function: $x \in \{0,1\}^8 \rightarrow XOR(x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8)$
- 2^8 examples split into 150 training and 106 test examples
- top-down approaches (e.g. C4.5) usually need all examples to learn consistent model

Results

- as expected, baseline fails
- DeepRED is able to extract rules that classify all or almost all test examples correctly



Open Question

- Understandability of intermediate concepts?

Steps Towards More Understandable Rule Learning Algorithms

Understandability of the learned rules should be explicitly considered in rule learning algorithms

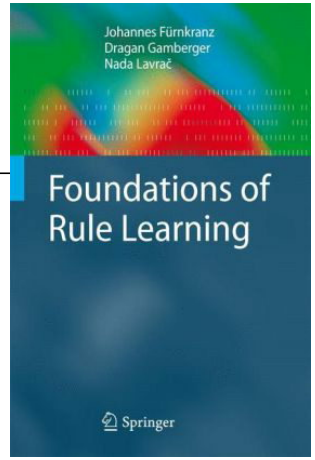
1. Understand Understandability
   - Take a closer look at results from cognitive science
2. Develop heuristics that include understandability
   - Of course, discriminative power should not be ignored
3. Integrate them in Rule Learning Algorithms
   - Possibly also as a post-processor (“rule beautification”)
4. Develop better algorithms
   - E.g., for learning structured concepts
5. Evaluate in user studies
   - Automatic evaluation would not be convincing

Conclusions

- Understandability is currently mostly defined via rule length
  - Occam's Razor: Shorter rules are better
- On the other hand, longer rules are often more convincing
  - Characteristic rules, closed itemsets, formal concepts, rules learned with inverted heuristics, ...
- Understandability is more than short rules, e.g.
  - **Representativeness**: a rule that is more typical to what we expect is more convincing
  - **Semantic coherence**: rules that have semantically similar conditions are better
  - **Recognition**: rules with well-recognized conditions are better
  - **Structure:** flat rules are not very natural

→ these should be considered when evaluating understandability!

References



- Allahyari H., Lavesson N.: User-oriented Assessment of Classification Model Understandability. SCAI 2011: 11-19
- Fürnkranz J., Flach, P.: ROC 'n' Rule Learning - Towards a Better Understanding of Covering Algorithms. Machine Learning 58(1): 39-77 (2005)
- Fürnkranz J., Gamberger D., Lavrac N.: Foundations of Rule Learning. Springer (2012)
- Fürnkranz J., Kliegr T.: A Brief Overview of Rule Learning. Proceedings RuleML 2015: 54-69 (2015)
- Gigerenzer G., Todd. P.M.: Simple Heuristics that make us smart. Oxford University Press (1999)
- Kahneman, D., Tversky, A. Subjective probability: A judgment of representativeness, Cogn. Psych. 3(3):430–454 (1972)
- Kliegr, T., Fürnkranz J.: On the Interpretability of Rule-Based Models. 2017, in preparation.
- Michalski, R.S.: A Theory and Methodology of Inductive Learning. Artificial Intelligence 20(2): 111-161 (1983)
- Paulheim, H.: Generating Possible Explanations for Statistics from Linked Open Data. Proc. ESWC-12, (2012)
- Paulheim, H., Fürnkranz, J.: Unsupervised Feature Generation from Linked Open Data. Proc. WIMS'12. (2012)
- Stecher J., Janssen F., Fürnkranz J.: Separating Rule Refinement and Rule Selection Heuristics in Inductive Rule Learning. Proceedings ECML/PKDD (3) 2014: 114-129 (2014)
- Stecher J., Janssen F., Fürnkranz J.: Shorter Rules Are Better, Aren't They? Proceedings DS 2016: 279-294
- Tversky A. and Kahneman, D.: "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement" in Psychological Review, 91, pp. 293-315, (1984)
- Vreeken, J., van Leeuwen M., Siebes, A: Krimp: mining itemsets that compress. DMKD. 23(1): 169-214 (2011)
- Wille R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival (Ed.): Ordered Sets, 445–470, Reidel, Dordrecht-Boston (1982)
- Zilke J., Loza Mencía E., Janssen F: DeepRED - Rule Extraction from Deep Neural Networks. DS 2016: 457-473