Towards Interpretable Rule Learning

Johannes Fürnkranz
Johannes Kepler University, Linz
Institute for Applied Knowledge Processing
Computational Data Analytics Group

juffi@faw.jku.at

Joint Work with
Tomas Kliegr, Florian Beck, Van Quoc Phuong Hyunh, et al.

Understandability – State of Affairs

JOHANNES KEPLER UNIVERSITÄT LINZ

Data Mining often assumes

- 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. This implies that,* ***the more complex*** *or large a model is,* ***the more understandable*** *it is"*
>
> (Allahyari & Lavesson 2011)

Interpretable Rule Learning

JOHANNES KEPLER UNIVERSITÄT LINZ

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

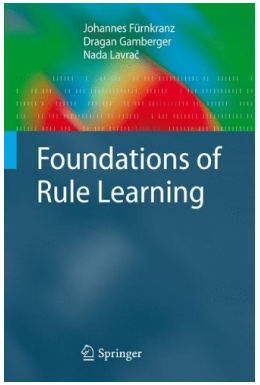
→ *we need to explicitly address interpretability*!

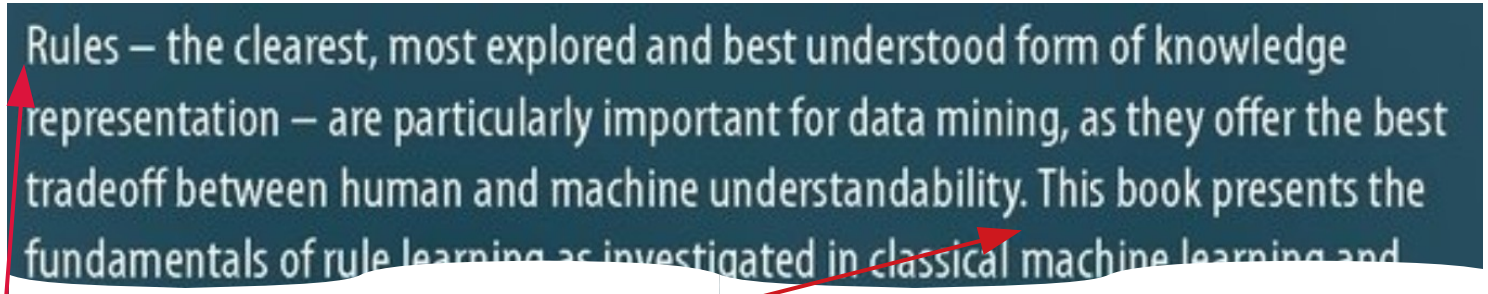
Interpretability and Rule Learning

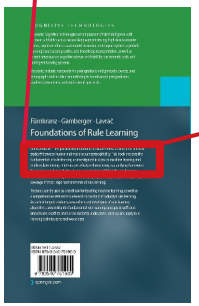
JOHANNES KEPLER UNIVERSITÄT LINZ

Rules (and decision trees) are often equated with interpretable concepts

- If we learn rules, then we are interpretable
- Shorter models are more interpretable than longer models







Note: The book has a 13-page index, which does not contain entries for understandability, interpretability, comprehensibility, or similar...

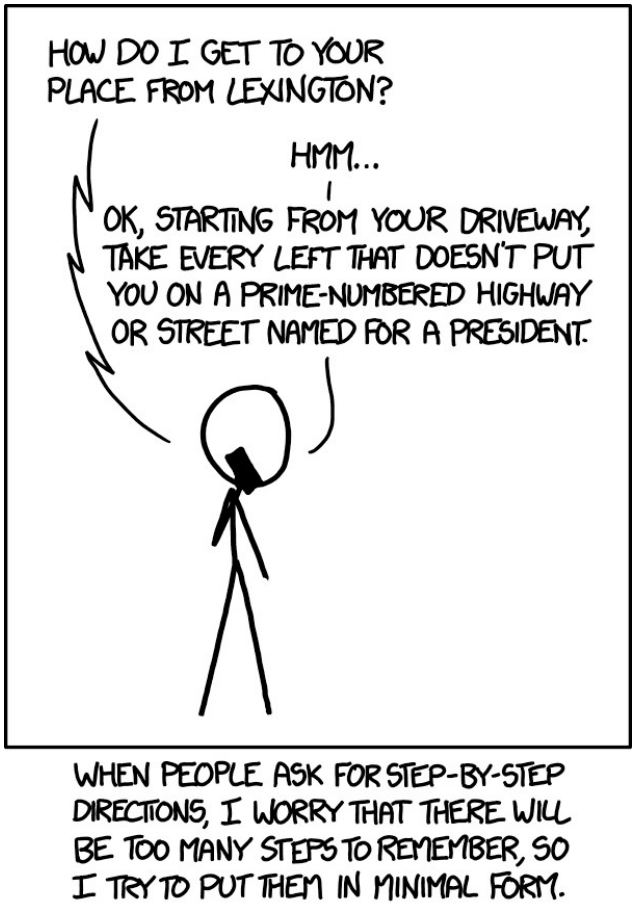
Are Shorter Explanations better?

- Complexity is often used as a simple surrogate for interpretability

Caveats

- Shorter explanations are often more predictive than longer ones
  - but do not necessarily need to be interpretable
- Focuses only on syntactic interpretability

Kolmogorov Directions



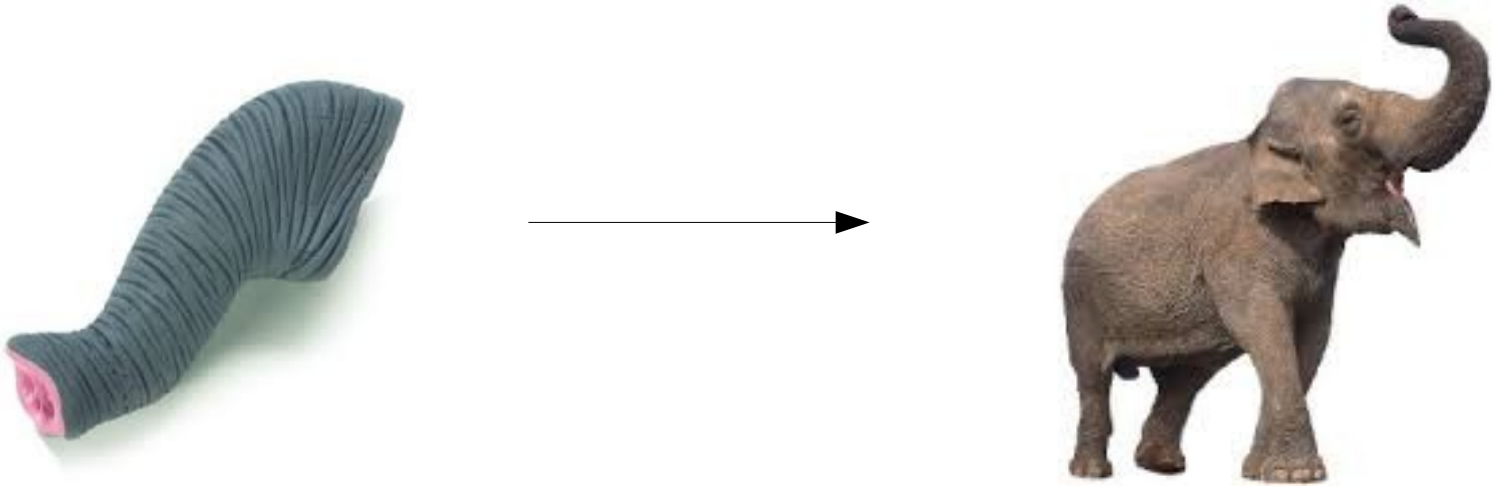
Source: https://www.xkcd.com/1155/
(Thanks to Jilles Vreeken for the pointer)

Discriminative Rules

JOHANNES KEPLER UNIVERSITÄT LINZ

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

- Example:

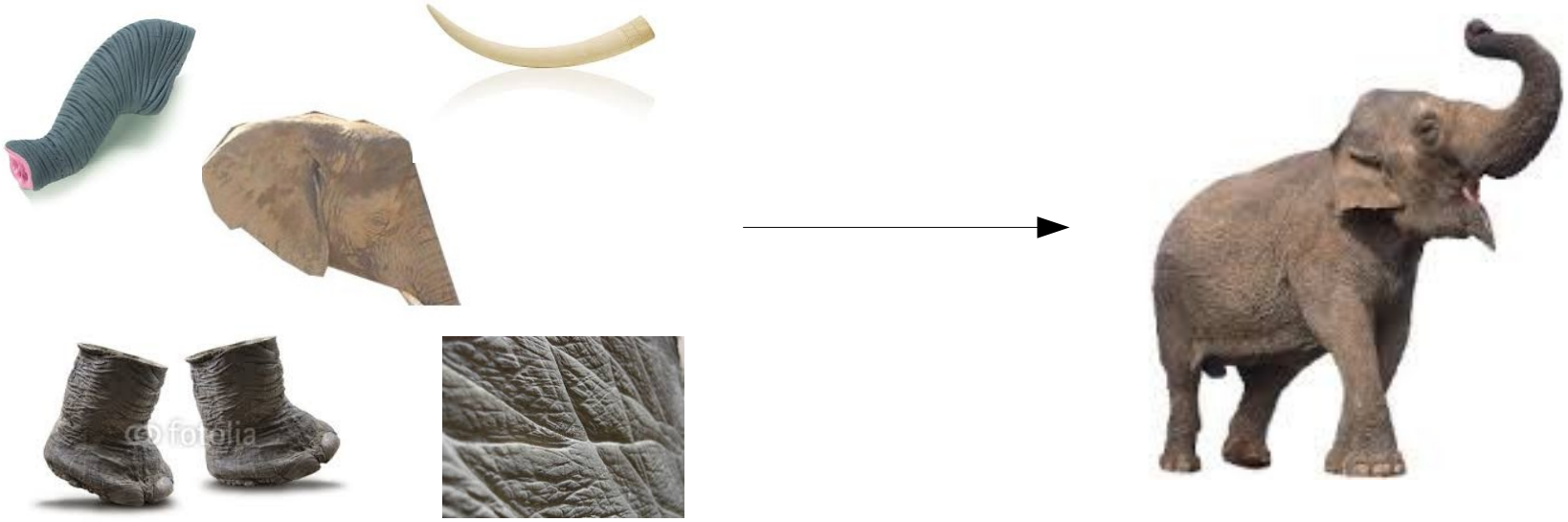


Characteristic Rules

JOHANNES KEPLER UNIVERSITÄT LINZ

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

- Example:



(Michalski 1983)

Discriminative Rules vs. Characteristic Rules

JOHANNES KEPLER UNIVERSITÄT LINZ

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

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

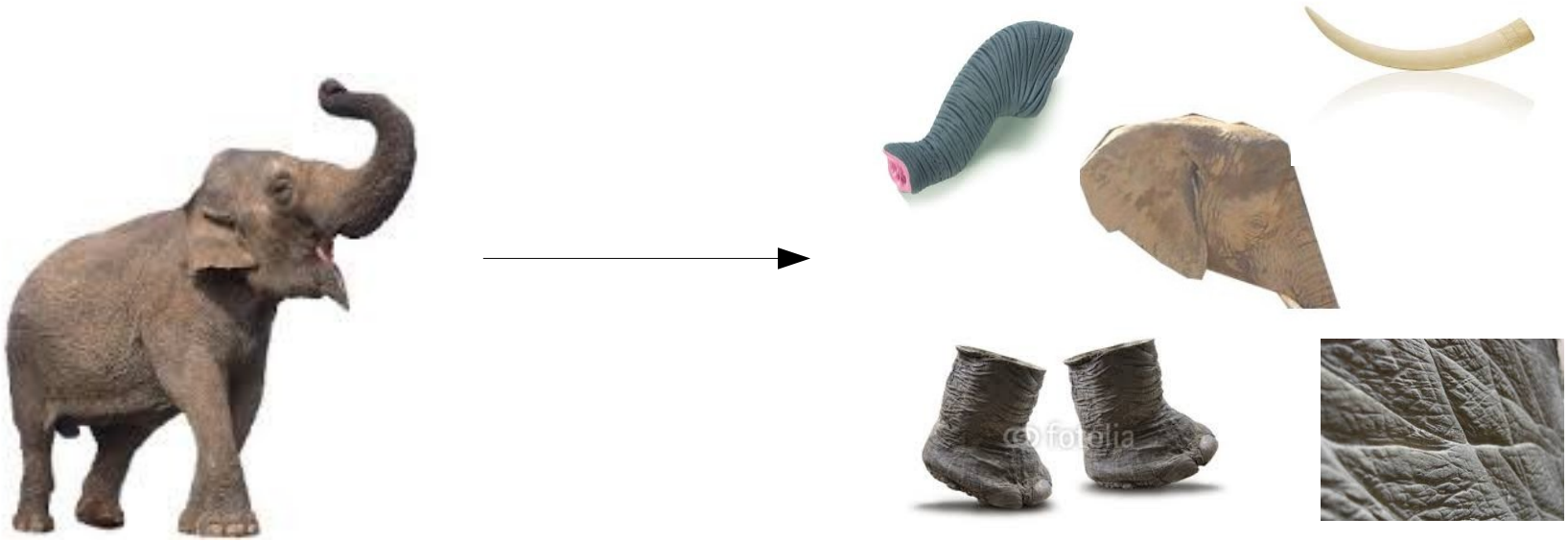
  - Most interesting are *maximal characteristic rules*.

Characteristic Rules

JOHANNES KEPLER UNIVERSITÄT LINZ

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

- Example:



Example Rules – Mushroom dataset

JOHANNES KEPLER UNIVERSITÄT LINZ

- 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)
```

(Stecher, Janssen, Fürnkranz 2016)

Example Rules – Coronary Heart Disease

JOHANNES KEPLER UNIVERSITÄT LINZ

```
[17| 0] class 1 :- holst < 0.0001, vkgq = 1, ergfr = 1, ergrt = 1.
[17| 0] class 1 :- ergg < 180.0001, ergst < 0.2001, vkgq = 1,
                   ehoef >= 68, ergfr = 1.
[14| 0] class 1 :- holst < 0.2001, vkgq = 1, ecgfr >= 70, ergd >= 100.
```

Longer rules with **higher coverage** (compared to h_{Lap})

```
[32|0] class 1 :- vkgq = 1, ergkp = 1, ergny = 1, ergrt = 1,
                  hight >= 154, ergfr = 1, holst < 0.3001, ecgpr = 1,
                  holfr = 1, ehoef >= 65.
[28|0] class 1 :- ergst < 0.3001, vkgq = 1, ergny = 1, hdl >= 0.72,
                  ergfr = 1, ecgrt = 1, ecgpr = 1, fib < 4.5001,
                  vkghl = 1, holst < 0.2001, ecgst = 1, holrt = 1, ldl < 4.7601.
[25|0] class 1 :- ergst < 0.2001, vkgq = 1, ergny = 1, ergrt = 1,
                  ergfr = 1, ecgpr = 1, ergkp = 1, ecgrt = 1, holfr = 1,
                  ehoef >= 64, ua < 308.0001.
```

(Stecher, Janssen, Fürnkranz 2016)

Example Rules – Brain Ischemia

JOHANNES KEPLER UNIVERSITÄT LINZ

```
[149| 0] ischemia :- b.i. < 60.0001.
[140| 0] ischemia :- b.i. < 70.0001, fibrin. >= 3.8.
[137| 0] ischemia :- b.i. < 75.0001, fibrin. >= 3.9.
```

Regular heuristics find **Barthel index** and **fibrinogen value** as relevant for a brain stroke.

Inverted heuristics in addition refer to
age, **diastolic blood pressure**, and **cholesterol**

```
[147| 0] ischemia :- rrrrdyast.. >= 70, fibrin. >= 2.8, b.i. < 60.0001.
[139| 0] ischemia :- age >= 58, rrrrdyast.. >= 80, b.i. < 60.0001.
[107| 0] ischemia :- rrrrdyast.. >= 80, fibrin. >= 3.5, b.i. < 65.0001, chol. >= 5.2.
```

Is Rule Length an Indicator for Interpretability?

(Fürnkranz, Kliegr, Paulheim 2020)

JOHANNES KEPLER UNIVERSITÄT LINZ

Crowdsourcing study on comparing the interpretability of rules:

- 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

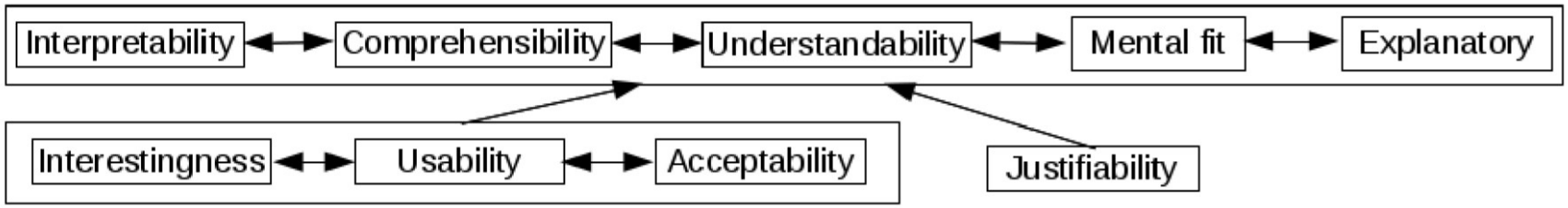
What is Interpretability?

JOHANNES KEPLER UNIVERSITÄT LINZ

Interpretability is an ill-defined concept

- with many intuitively well-understood connotations
  - understandability, interpretability, comprehensibility, plausibility, trustworthiness, justifiability, …
- but only a few formal definitions

Bibal & Frénay (2016) suggest the following clarification:



- note that interpretability depends on “interestingness / acceptance” and “justifiability”

Operational Definitions of Interpretability

JOHANNES KEPLER UNIVERSITÄT LINZ

Other definitions of interpretability mostly focus on whether the knowledge can be put to use.

- comprehensibility of a program (Schmid, Muggleton et al, 2017/18)

> **Definition 3** (*Comprehensibility, $C(S, P)$*) The comprehensibility of a definition (or program) P with respect to a human population S is the mean accuracy with which a human s from population S after brief study and without further sight can use P to classify new material sampled randomly from the definition's domain.

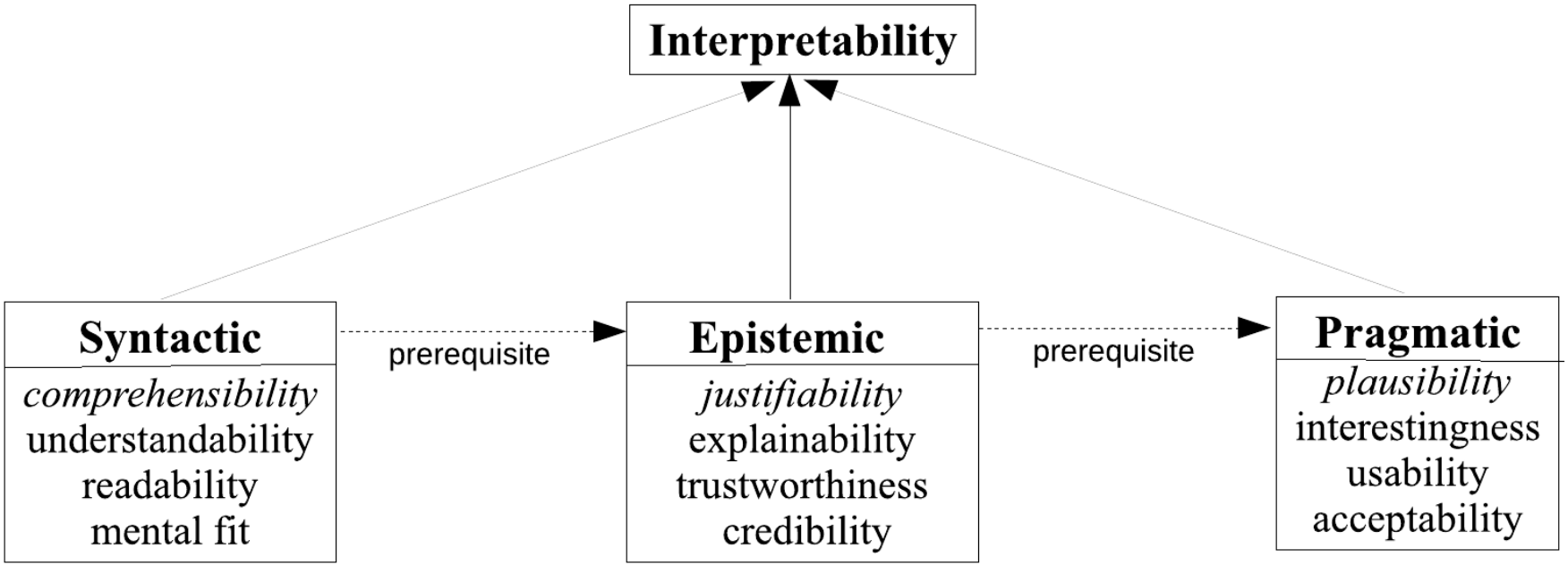
- interpretability with respect to a target model (Dhurandar et al. 2017)

> **Definition 2.2.** *CM-based δ-interpretability: Given a target model M_T belonging to a hypothesis class $\mathcal{H}$, a complex model M_C, and a target distribution D_T, the procedure P is δ-interpretable relative to the model pair (M_C, M_T), if it derives information I from M_C and produces a model $M_T(I) \in \mathcal{H}$ satisfying the following inequality: $e_{M_T(I)} \leq \delta \cdot e_{M_T}$, where $e_{\mathcal{M}}$ is the expected error of $\mathcal{M}$ relative to some loss function on D_T.*

(Fürnkranz, Kliegr, Paulheim, MLJ 2020)

Three Aspects of Interpretability

JOHANNES KEPLER UNIVERSITÄT LINZ



Can we read and understand the model?

Does the model connect to previous knowledge?

Does the model appear to be useful?

Cognitive Biases

JOHANNES KEPLER UNIVERSITÄT LINZ

- In order to understand interpretability, in particular pragmatic interpretability ("plausibility") it may be helpful to take a look at cognitive biases

> A **cognitive bias** is a systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments.
>
> (Tversky & Kahnemann,1974)

- **Hypothesis:**

  Cognitive biases may help to define interpretability biases.

(Tversky & Kahneman 1983)

Conjunctive Fallacy

JOHANNES KEPLER UNIVERSITÄT LINZ

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.

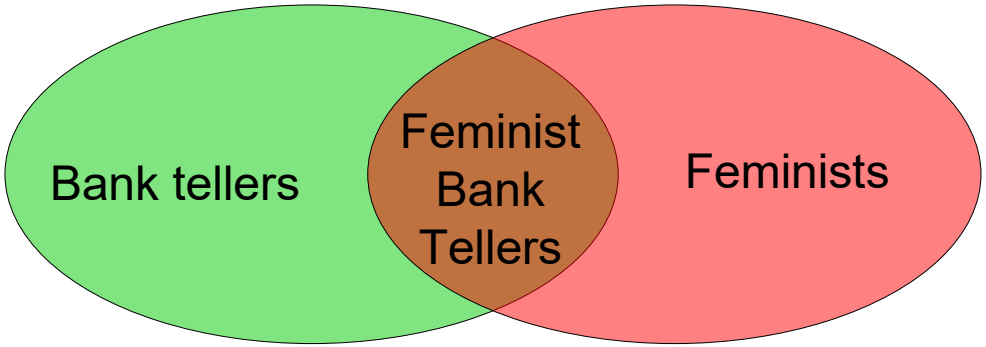
(Tversky & Kahneman 1983)

Conjunctive Fallacy

JOHANNES KEPLER UNIVERSITÄT LINZ

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

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



(Kahneman & Tversky 1972)

Representativeness Heuristics

JOHANNES KEPLER UNIVERSITÄT LINZ

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

Linda

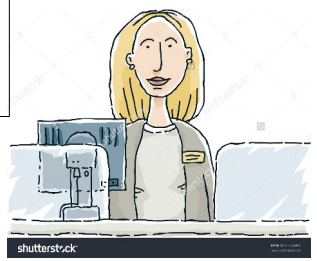
31 years old, single, outspoken, very bright. majored in philosophy, concerned with issues discrimination and social justice, anti-nuclear

less similar

more similar

Bank Teller

Bank Teller Feminist



(Kliegr, Bahnik, Fürnkranz, AIJ 2021)

Cognitive Biases Interact with Rule Interpretations

JOHANNES KEPLER UNIVERSITÄT LINZ

phenomenon	implications for rule-learning	debiasing technique
Representativeness Heuristic	Overestimate the probability of condition representative of consequent	Use natural frequencies instead of ratios or probabilities
Averaging Heuristic	Probability of antecedent as the average of probabilities of conditions	Reminder of probability theory
Disjunction Fallacy	Prefer more specific conditions over less specific	Inform on taxonomical relation between conditions; explain benefits of higher support
Base-rate Neglect	Emphasis on confidence, neglect for support	Express confidence and support in natural frequencies
Insensitivity to Sample Size	Analyst does not realize the increased reliability of confidence estimate with increasing value of support	Present support as absolute number rather than percentage; use support to compute confidence (reliability) intervals for the value of confidence
Availability Heuristic	Ease of recollection of instances matching the rule	Explain to analyst why instances matching the particular rule are (not) easily recalled
Reiteration Effect	Presentation of redundant rules or conditions increases plausibility	rule pruning; clustering; explaining overlap
Confirmation Bias	Rules confirming analyst's prior hypothesis are "cherry picked"	Explicit guidance to consider evidence for and against hypothesis; education about the bias; interfaces making users slow down
Mere Exposure Effect	Repeated exposure (even subconscious) results in increased preference	Changes to user interfaces that limit subliminal presentation of rules
Overconfidence and underconfidence	Rules with small support and high confidence are "overrated"	Present less information when not relevant via pruning, feature selection, limiting rule length; actively present conflicting rules/knowledge.
Recognition Heuristic	Recognition of attribute or its value increases preference	More time; knowledge of attribute/value
Information Bias	belief that more information (rules, conditions) will improve decision making even if it is irrelevant	Communicate attribute importance
Ambiguity Aversion	Prefer rules without unknown conditions	Increase user motivation; instruct users to provide textual justifications
Confusion of the Inverse	Confusing the difference between the confidence of the rule Pr(consequent|antecedent) with Pr(antecedent|consequent)	Training in probability theory; unambiguous wording
Misunderstanding of "and"	"and" is understood as disjunction	Unambiguous wording; visual representation
Context and Tradeoff Contrast	Preference for a rule is influenced by other rules	Removal of rules, especially of those that are strong, yet irrelevant
Negativity Bias	Words with negative valence in the rule make it appear more important	Review words with negative valence in data, and possibly replace with neutral alternatives
Primacy Effect	Information presented first has the highest impact	Education on the bias; resorting; rule annotation
Weak Evidence Effect	Condition only weakly perceived as predictive of target decreases plausibility	Numerical expression of strength of evidence; omission of weak predictors (conditions)
Unit Bias	Conditions are perceived to have same importance	Inform on discriminatory power of conditions

(Fürnkranz, Kliegr 2018)

The Need for Interpretability Biases

JOHANNES KEPLER UNIVERSITÄT LINZ

- 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, ...
- Additional aspects that could be considered in rule rule learning:
  - **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

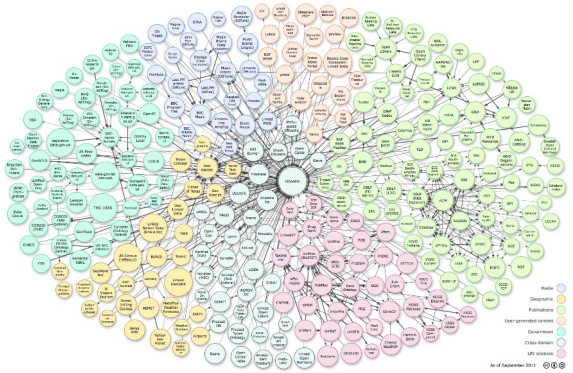
(Paulheim 2012)

Zero-Knowledge Data Mining

JOHANNES KEPLER UNIVERSITÄT LINZ

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
```

(Paulheim 2012)

Recognition Effects

JOHANNES KEPLER UNIVERSITÄT LINZ

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.`

Semantic Coherence

JOHANNES KEPLER UNIVERSITÄT LINZ

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".

→ for learning them we need a different bias

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

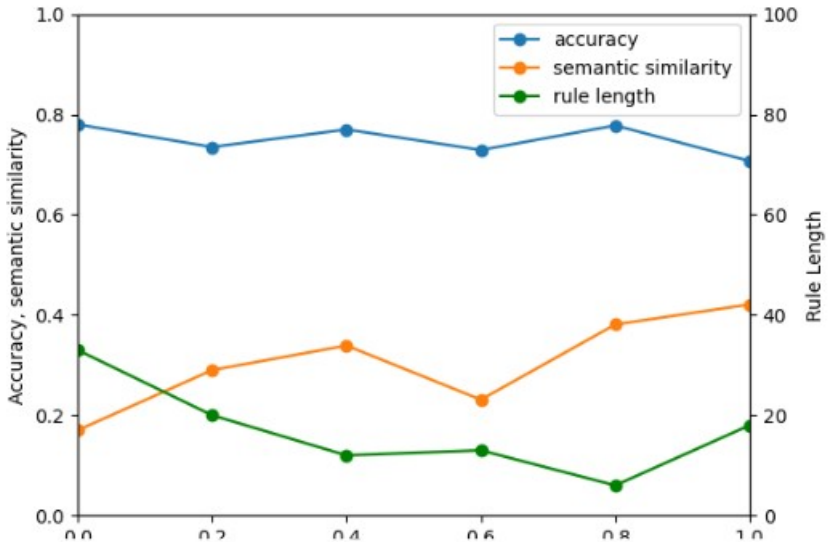
(Mahya & Fürnkranz, under review)

Improving Semantic Coherence

JOHANNES KEPLER UNIVERSITÄT LINZ

- CoRIFEE takes as input a set of rule sets (such as from a random forest), and returns a rule set with improved semantic coherence
  - semantic coherence is measured by the distance of the conditions in some reference ontology

Hepatitis Algorithm	α	Accuracy	Semantic Similarity	Number of Rules
Our Method	0	0.780	0.170	33
	0.2	0.735	0.290	28
	0.4	0.770	0.339	22
	0.6	0.729	0.231	18
	0.8	0.778	0.381	11
	1	0.707	0.421	9
Random Forest		0.787	0.162	50



Learning Deep (Strucutred) Rule Sets
Example: Parity / XOR

JOHANNES KEPLER UNIVERSITÄT LINZ

- Consider the parity / XOR problem
  - $n + r$ binary attributes sampled with an equal distribution of 0/1
  - n relevant binary attributes (the first n w.l.o.g.)
  - r irrelevant binary attributes

- Target concept:
  - is there an even number of 1's in the relevant attributes?

Encoding Parity with a Flat Rule Set

JOHANNES KEPLER UNIVERSITÄT LINZ

Most rule learning algorithms learn flat theories

- n-bit parity needs 2^{n-1} flat rules, no shorter encoding is possible
- each rule encoding one positive case in the truth table

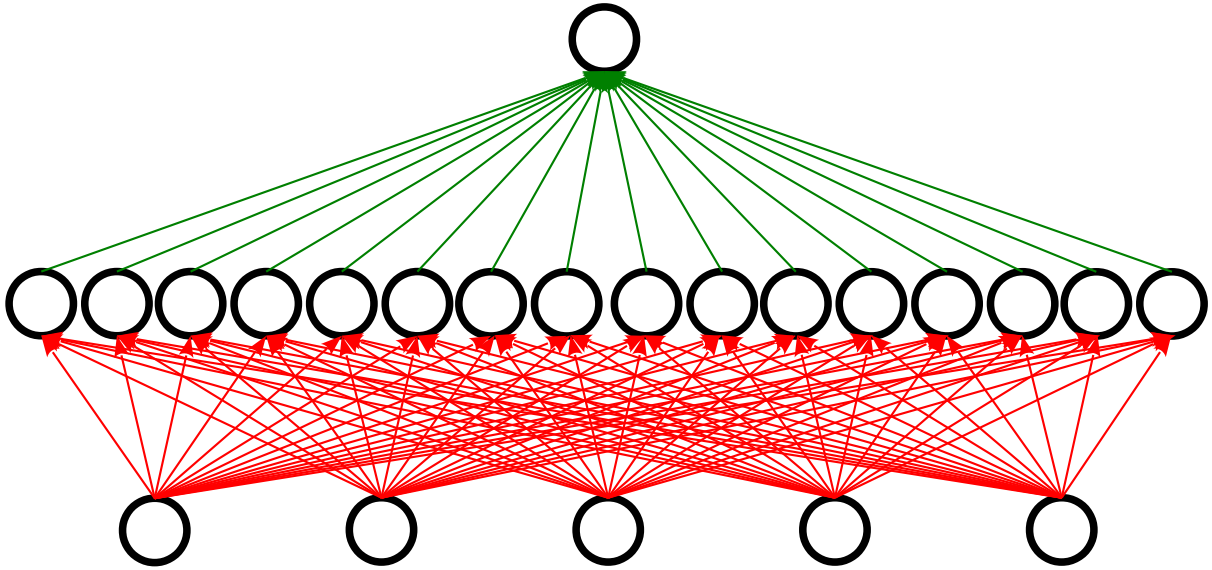
```
parity :-     x1,     x2,     x3,     x4, not x5.
parity :-     x1,     x2, not x3, not x4, not x5.
parity :-     x1, not x2,     x3, not x4, not x5.
parity :-     x1, not x2, not x3,     x4, not x5.
parity :- not x1,     x2, not x3,     x4, not x5.
parity :- not x1,     x2,     x3, not x4, not x5.
parity :- not x1, not x2,     x3,     x4, not x5.
parity :- not x1, not x2, not x2, not x4, not x5.
parity :-     x1,     x2,     x3, not x4,     x5.
parity :-     x1,     x2, not x3,     x4,     x5.
parity :-     x1, not x2,     x3,     x4,     x5.
parity :- not x1,     x2,     x3,     x4,     x5.
parity :- not x1, not x2, not x3,     x4,     x5.
parity :- not x1, not x2,     x3, not x4,     x5.
parity :- not x1,     x2, not x3, not x4,     x5.
parity :-     x1, not x2, not x2, not x4,     x5.
```

DNF formula with 2^{n-1} literals, each having n variables

Network View of a Flat Rule Set

JOHANNES KEPLER UNIVERSITÄT LINZ

- Flat Rule Sets can be converted into a network using a single AND and a single OR layer (analogous to Sum-Product Networks)



- Each node in the hidden layer corresponds to one rule
  - typically it is a local pattern, covering part of the target

Encoding Parity with a Structured Rule Base

JOHANNES KEPLER UNIVERSITÄT LINZ

But structured concepts are often more interpretable

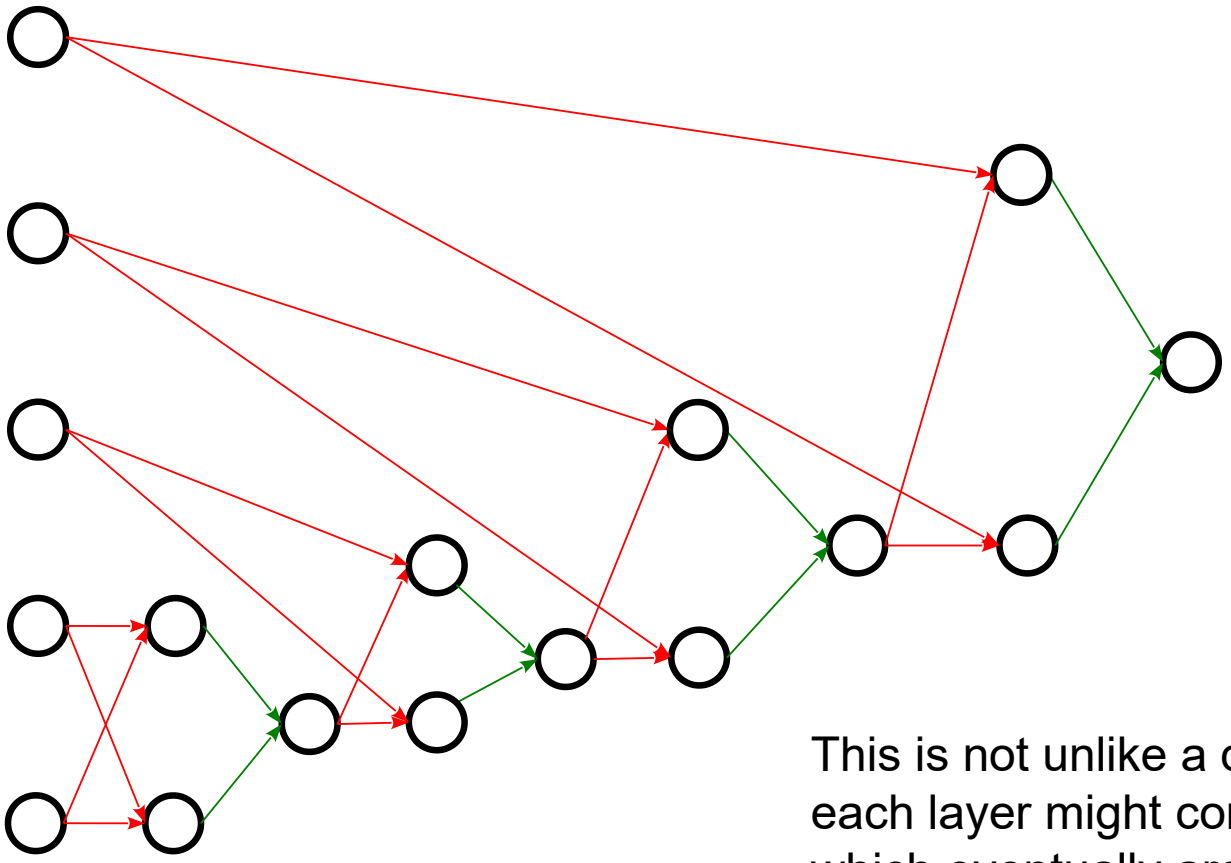
- in parity we need only $O(n)$ rules with intermediate concepts

```
parity45   :-     x4,     x5.
parity45   :- not x4, not x5.

parity345  :-     x3, not parity45.
parity345  :- not x3,     parity45.

parity2345 :-     x2, not parity345.
parity2345 :- not x2,     parity345.

parity     :-     x1, not parity2345.
parity     :- not x1,     parity2345.
```

Network View of a Structured Rule Base

JOHANNES KEPLER UNIVERSITÄT LINZ

- This is encodes a deep network structure



This is not unlike a deep network: each layer might contain more nodes, which eventually are not needed

Why is it good to learn deep rule sets?

JOHANNES KEPLER UNIVERSITÄT LINZ

- **Expressivity?** It does not necessarily increase expressivity
  - any structured rule base can be converted into an equivalent DNF expression, i.e., a flat set of rules
    - but this is also true for NNs → universal approximation theorem (one layer is sufficient; Hornik et al. 1989)
  - in both cases the number of terms (size of hidden layers, conjuncts in the DNF) is unbounded
    - Note that a disjunction of all examples is also a DNF expression

Why is it good to learn deep rule sets?

JOHANNES KEPLER UNIVERSITÄT LINZ

- **Interpretability?**
  - structured rule sets may be more compact
  - are they more interpretable?

  - **Example**: Why is $\mathbf{x} = (1,1,1,0,1,0,0,1,0,0,...)$ in parity?

```
parity :-     x1,     x2,     x3,     x4, not x5.
parity :-     x1,     x2, not x3, not x4, not x5.
parity :-     x1, not x2,     x3, not x4, not x5.
parity :-     x1, not x2, not x3,     x4, not x5.
parity :- not x1,     x2, not x3,     x4, not x5.
parity :- not x1,     x2,     x3, not x4, not x5.
parity :- not x1, not x2,     x3,     x4, not x5.
parity :- not x1, not x2, not x2, not x4, not x5.
parity :-     x1,     x2,     x3, not x4,     x5.
parity :-     x1,     x2, not x3,     x4,     x5.
parity :-     x1, not x2,     x3,     x4,     x5.
parity :- not x1,     x2,     x3,     x4,     x5.
parity :- not x1, not x2, not x3,     x4,     x5.
parity :- not x1, not x2,     x3, not x4,     x5.
parity :- not x1,     x2, not x3, not x4,     x5.
parity :-     x1, not x2, not x2, not x4,     x5.
```

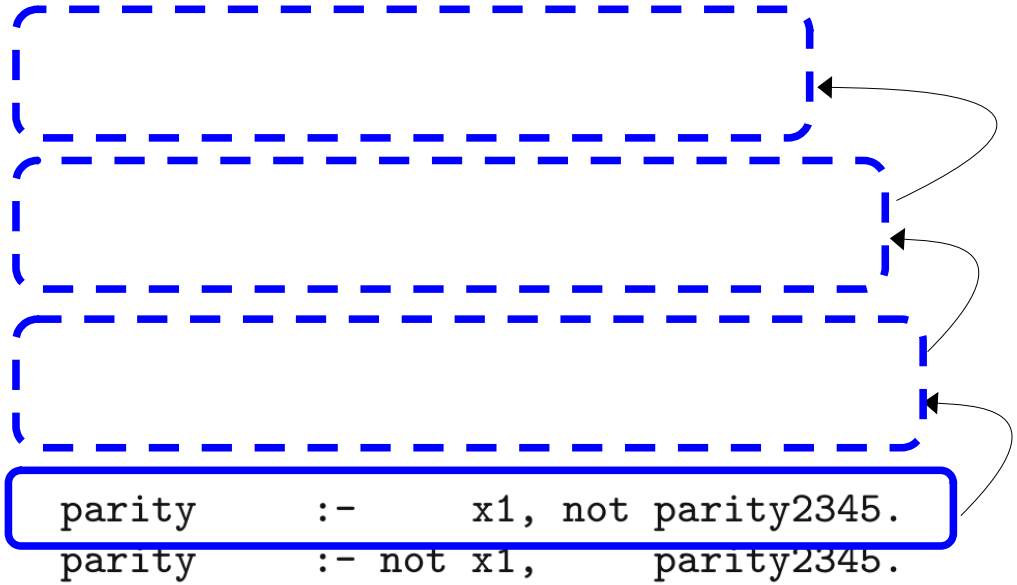
Even though the rule set is quite complex, we only need a single rule for giving a good explanation.

Why is it good to learn deep rule sets?

JOHANNES KEPLER UNIVERSITÄT LINZ

- **Interpretability?**
  - structured rule sets may be more compact
  - are they more interpretable?

    → Only if all subconcepts are easily interpretable!
  - **Example**: Why is $\mathbf{x} = (1,1,1,0,1,0,0,1,0,0,...)$ in parity?



Even though the rule set is more compact, we need to understand every subconcept in order to interpret the explanation.

(Fürnkranz et al. 2020)

Why is it good to learn structured rule bases?

JOHANNES KEPLER UNIVERSITÄT LINZ

- **Explicit representation of all aspects of the decision function**
  - rule sets are typically **not declarative**, require some sort of **tie breaking**
- two main approaches
  - **weighted rules** / probabilistic rules

$$\begin{array}{l} \boldsymbol{r}_1(0.8) : a \wedge b \rightarrow x \\ \boldsymbol{r}_2(0.9) : b \wedge c \rightarrow y \\ \boldsymbol{r}_3(0.7) : c \wedge d \rightarrow x \\ \boldsymbol{d} : \rightarrow z \end{array}$$

max: y (0.9)

sum: x (0.7+0.8 > 0.9)

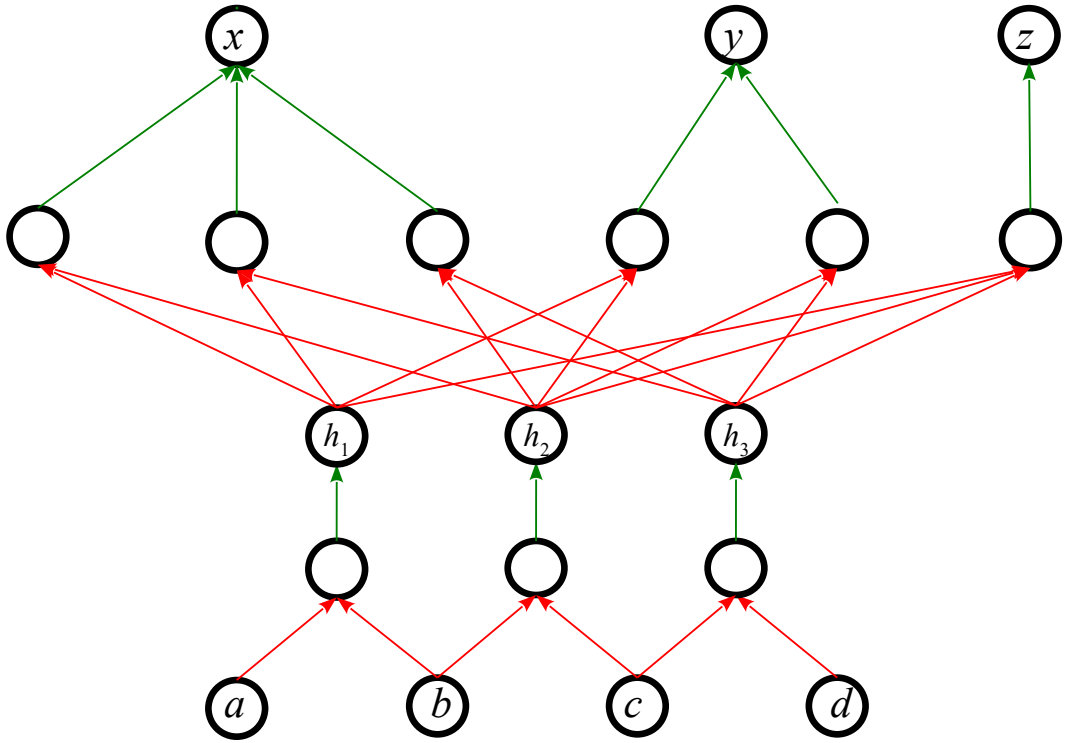
  - **decision lists** $\mathcal{D} = (\boldsymbol{r}_2, \boldsymbol{r}_1, \boldsymbol{r}_3, \boldsymbol{d})$
    - sort the rules according to some criterion
      - e.g., order in which they are learned
      - e.g., order according to weight (effectively equivalent to using weighted max)
    - use the first rule that fires

Declarative Version of Weighted Rule Sets

JOHANNES KEPLER UNIVERSITÄT LINZ

- Tie Breaking with Majority vote

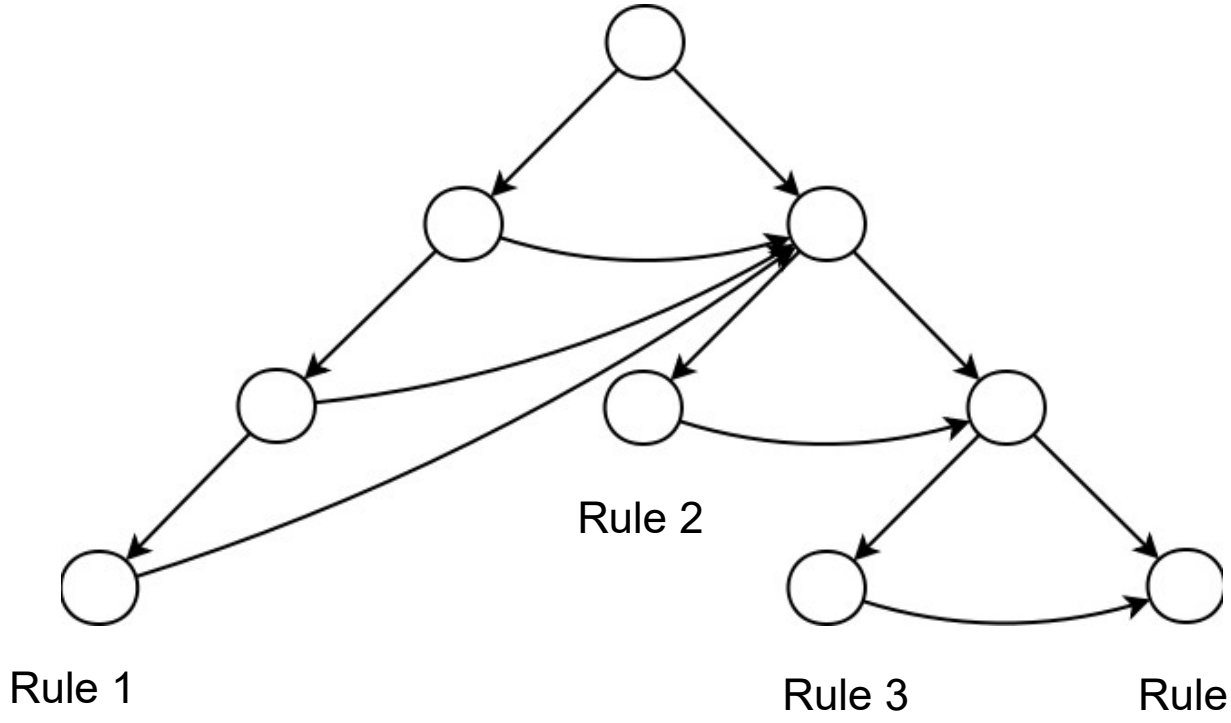
$$
\begin{aligned}
a \wedge b &\rightarrow h_1 \\
b \wedge c &\rightarrow h_2 \\
c \wedge d &\rightarrow h_3 \\
h_1 \wedge h_3 &\rightarrow x \\
h_1 \wedge \neg h_2 &\rightarrow x \\
h_3 \wedge \neg h_2 &\rightarrow x \\
h_2 \wedge \neg h_1 &\rightarrow y \\
h_2 \wedge \neg h_3 &\rightarrow y \\
\neg h_1 \wedge \neg h_2 \wedge \neg h_3 &\rightarrow z
\end{aligned}
$$



Declarative Version of Decision List

JOHANNES KEPLER UNIVERSITÄT LINZ

- A decision list is a decision graph, where not satisfied condition takes you to the start of the next rule
- Example of a decision list with 4 rules with 4, 2, 2, 1 conditions

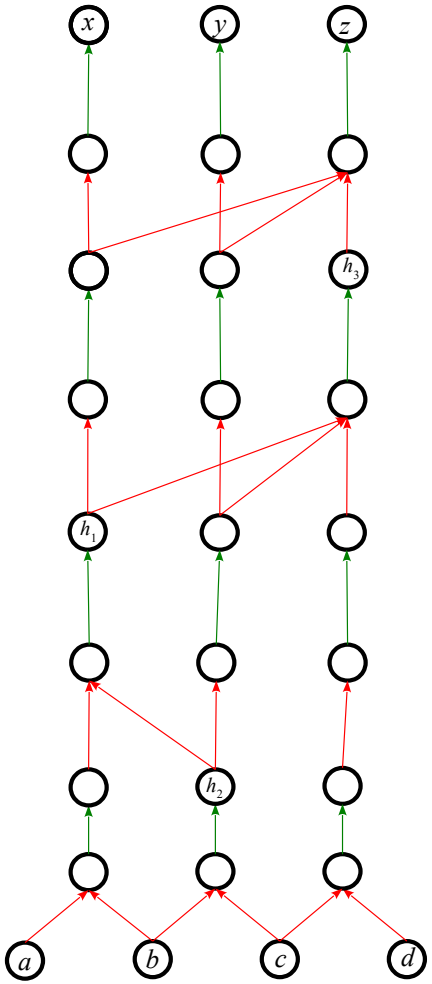


Declarative Version of Decision List

JOHANNES KEPLER UNIVERSITÄT LINZ

- In our example

$$
\begin{aligned}
b \wedge c &\rightarrow h_2 \\
h_2 &\rightarrow y \\
\neg h_2 \wedge a \wedge b &\rightarrow h_1 \\
h_1 &\rightarrow x \\
\neg h_1 \wedge \neg h_2 \wedge c \wedge d &\rightarrow h_3 \\
h_3 &\rightarrow x \\
\neg h_1 \wedge \neg h_2 \wedge \neg h_3 &\rightarrow z
\end{aligned}
$$



Why is it good to learn structured rule bases?

JOHANNES KEPLER UNIVERSITÄT LINZ

- **Learning Efficiency**
  - the hope is that deeper structures might be easier to learn
  - possibly contain fewer “parameters” that need to be found

How to Learn Deep Rule Sets

JOHANNES KEPLER UNIVERSITÄT LINZ

1. The Neural Network Approach
   - fix a network structure and optimize its parameters

   a) Binary/Ternary Neural Networks
   - most of the works focus on (memory) efficiency, not on logic interpretability

   b) Differentiable Logic
   - most of the works focus on first-order logic
   - diff-logic is an interesting exception

   c) Sum/Product Networks
   - focus on probabilities

→ We did a study in order to compare deep and shallow structure with a simple optimization algorithm (randomized hill-climbing)

(Beck & Fürnkranz 2020)

Does a Deep Structure help?

JOHANNES KEPLER UNIVERSITÄT LINZ

- To answer this empirically, we need to **compare** a powerful **shallow** rule learner with a powerful **deep** rule learner
  - But we do not have a powerful deep rule learner… (yet)
- Instead, we use a **simple optimization algorithm** to learn both, deep and shallow representations

1) Fix a network architecture
   - Shallow, single layer network RNC: [20]
   - Deep 3-layer network DRNC(3): [32, 8, 2]
   - Deep 5-layer network DRNC(5): [32, 16, 8, 4, 2]
2) Initialize Boolean weights probabilistically
3) Use stochastic local search to find best weight „flip“ on a mini-batch of data until convergence
4) Optimize finally on whole training set

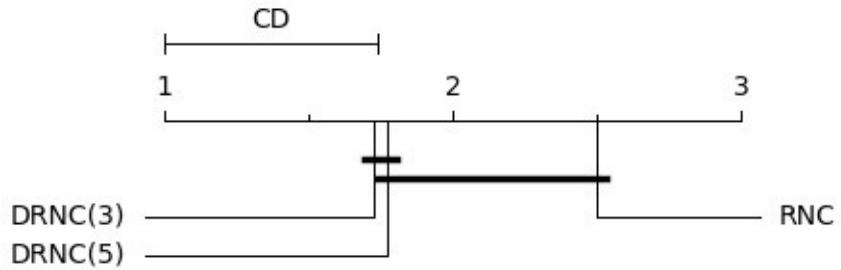
(Beck & Fürnkranz 2020)

Results on Artificial Datasets

JOHANNES KEPLER UNIVERSITÄT LINZ

- 20 artificial datasets with 10 Boolean inputs, 1 Boolean output
  - generated from a randomly initialized (deep) Boolean network

seed %(+)	DRNC(5)	DRNC(3)	RNC	RIPPER	CART
Ø Accuracy	0.9467	0.9502	0.9386	0.9591	*0.9644*
Ø Rank	1.775	1.725	2.5		

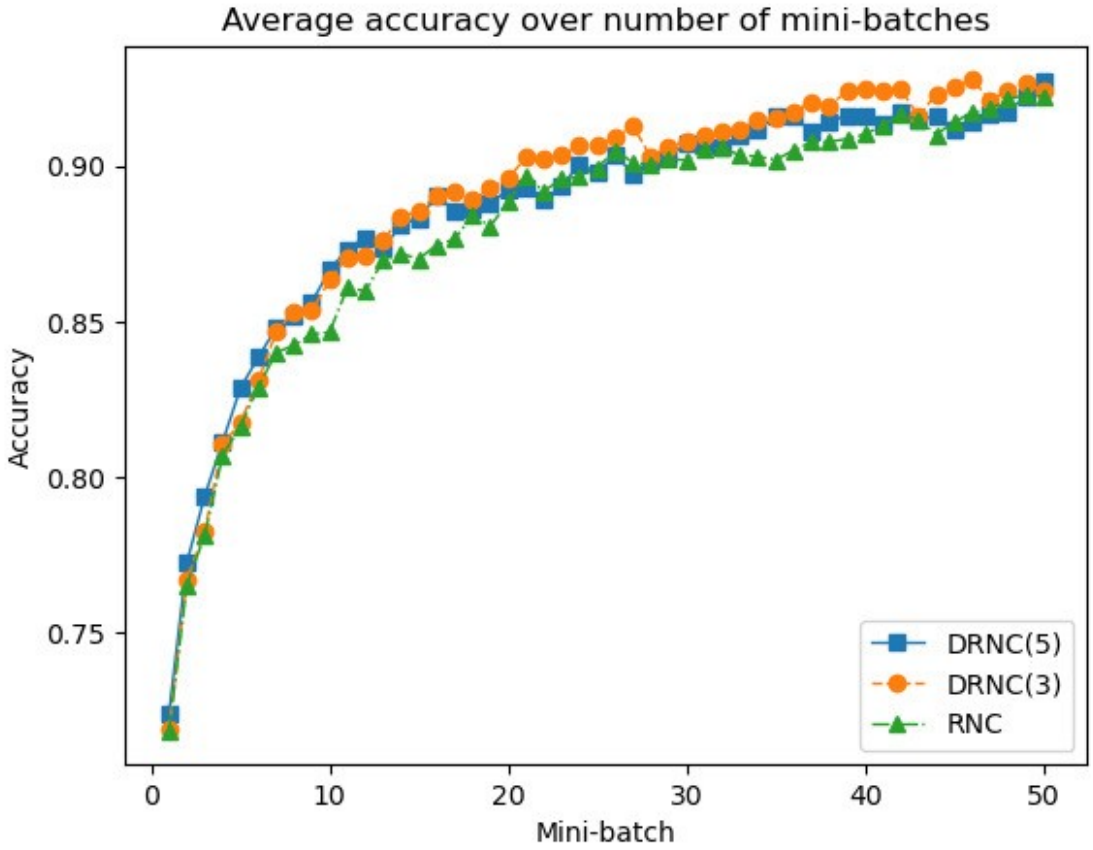


- DRNC(3) [DRNC(5)] outperforms RNC on a significance level of more than 95% [90%]

(Beck & Fürnkranz 2020)

Learning Curves (Artificial Datasets)

JOHANNES KEPLER UNIVERSITÄT LINZ



- DRNC(3) and DRNC(5) converge faster than RNC

(Beck & Fürnkranz 2020)

Results on Real-World (UCI) Datasets

JOHANNES KEPLER UNIVERSITÄT LINZ

dataset	%(+)	DRNC(5)	DRNC(3)	RNC	RIPPER	CART
car-evaluation	0.7002	0.8999	**0.9022**	0.8565	*0.9838*	0.9821
connect-4	0.6565	**0.7728**	0.7712	0.7597	0.7475	*0.8195*
kr-vs-kp	0.5222	0.9671	0.9643	**0.9725**	0.9837	*0.989*
monk-1	0.5000	***1***	0.9982	0.9910	0.9478	0.8939
monk-2	0.3428	0.7321	**0.7421**	0.7139	0.6872	*0.7869*
monk-3	0.5199	**0.9693**	0.9603	0.9567	0.9386	*0.9729*
mushroom	0.784	***1***	0.978	0.993	0.9992	*1*
tic-tac-toe	0.6534	0.8956	0.9196	**0.9541**	*1*	0.9217
vote	0.6138	***0.9655***	0.9288	0.9264	0.9011	0.9287
Ø Rank		1.556	2	2.444		

- DRNC(5) has the best performance on these real-world datasets, followed by DRNC(3)

How to Learn Deep Rule Sets

JOHANNES KEPLER UNIVERSITÄT LINZ

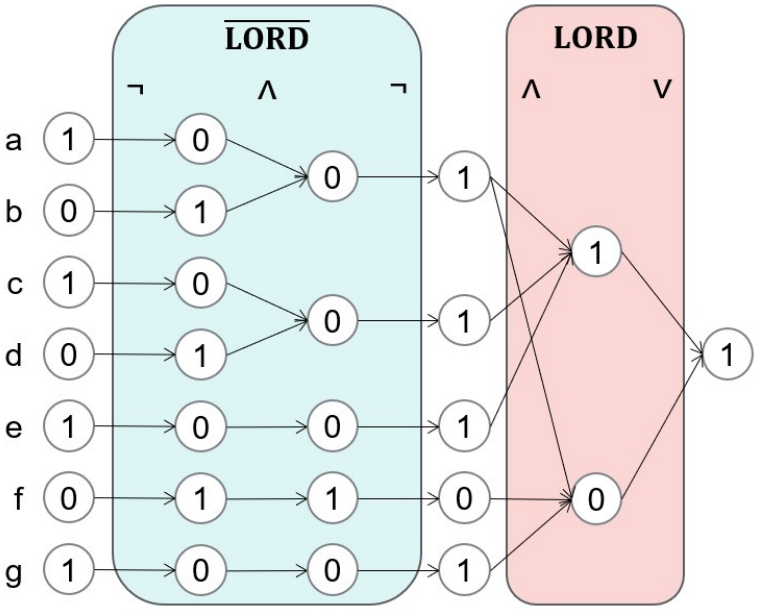
1. The Neural Network Approach
   - fix a network structure and optimize its parameters
2. The Rule Learning Approach
   - layerwise learning of multiple layers of conjunctive and disjunctive rules
     - use conjunctions as input features for CNF learner, and vice versa
   - DNF learners can be used for learning CNF layers

(Beck, Fürnkranz, Huynh 2023)

Learning Mixed Conjunctive and Disjunctive Rules

JOHANNES KEPLER UNIVERSITÄT LINZ

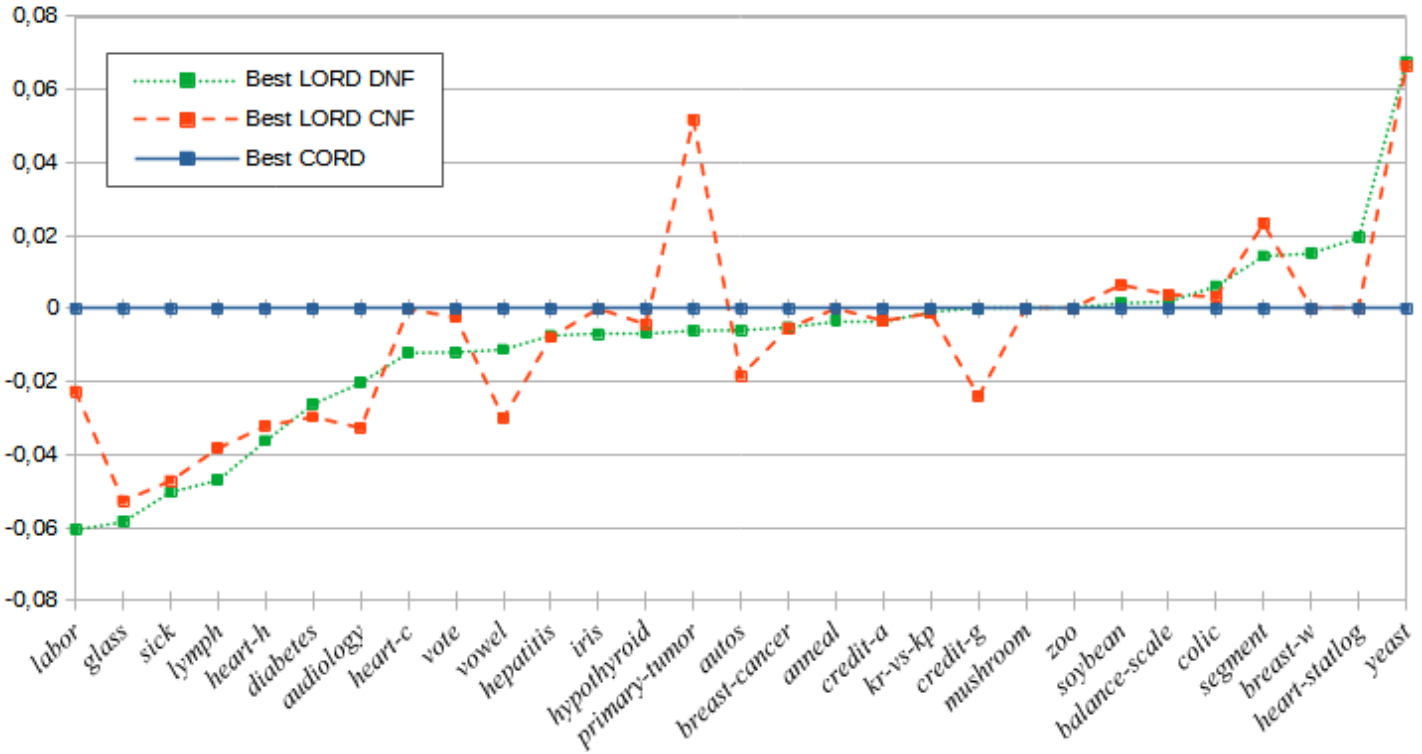
- LORD: A (powerful) conventional rule learner (i.e., DNF learner)
- NegLORD: Learn a CNF by inverting the problem to learn a DNF on the negated classes and negated inputs
- CORD: Allow a combination of conjunctive and disjunctive layers to potentially learn the best of both worlds



Results

JOHANNES KEPLER UNIVERSITÄT LINZ

- As known from previous works, some concepts can be better learned in CNF, some in DNF
- CORD is in most (but not all) cases better than either



Going Deeper

JOHANNES KEPLER UNIVERSITÄT LINZ

- CORD has 3 layers by default (disj./conj./disj.)
- More layers could be added with the same setup
- Results show modest but not consistent improvements for carefully tuned networks

FROM → TO	2 → 3	2 → 4	2 → 5	3 → 4	3 → 5	4 → 5
# IMPR.	6219	6189	6788	4407	4877	3189
# DET.	5274	5301	6057	4452	5007	3289
% IMPR.	24.75	24.63	27.01	17.54	19.41	12.69
% DET.	20.99	21.09	24.10	17.72	19.92	13.09
VALUES FOR BEST FIVE-LAYERED CORD:						
# IMPR.	126	139	144	86	97	40
# DET.	48	53	52	62	56	17
% IMPR.	43.45	47.93	49.66	29.66	33.45	13.79
% DET.	16.55	18.28	17.93	21.38	19.31	5.86

Analysis of Deeper Networks

JOHANNES KEPLER UNIVERSITÄT LINZ

- positive and negative correlation of various properties in the conjunctive and disjunctive layers of 5-layer networks with overall accuracy

	CORD				DORC			
	D_1	C_2	D_3	C_4	C_1	D_2	C_3	D_4
m	0.154	0.020	-0.101	-0.131	0.081	0.175	0.019	-0.098
# Rules	-0.189	-0.145	-0.092	-0.043	-0.084	-0.253	-0.134	-0.081
# Concepts	-	0.095	0.045	0.008	-	0.060	0.151	0.074
Avg. Depth	-	0.111	0.057	-0.018	-	0.117	0.159	0.107
Accuracy	0.203	0.520	0.690	-	-0.041	0.342	0.564	-

- e.g., higher values of the m-parameter (yielding more general rules) are good in early layers, wheras lower values are better in later layers
- accuracy increases in later layers

How to Learn Deep Rule Sets

JOHANNES KEPLER UNIVERSITÄT LINZ

1. The Neural Network Approach
   - fix a network structure and optimize its parameters
2. The Rule Learning Approach
   - layerwise learning of multiple layers of conjunctive and disjunctive rules
   - DNF learners can be used for learning CNF layers
3. Dedicated Search Algorithm
   - bidirectional search of multiple specializations (selecting conditions) and generalizations (pruning conditions) for learning individual rules did not bring much improvement in the LORD rule learner
     - one layer of specializations + one layer of generalizations is enough
   - ongoing work:
     - evaluate this for incremental constructions of AND/OR networks
       - similar to → (fuzzy) pattern trees (Hüllermeier 2015)

Conclusions

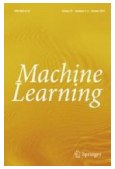
JOHANNES KEPLER UNIVERSITÄT LINZ

- Interpretability is a multi-faceted concept
  - complexity is only one aspect
- We need to develop techniques for biasing symbolic learning algorithms towards interpretability
  - semantic coherence, representativeness, …
- Learning deeply structured logical theories is an interesting and challenging problem
  - learning interpretable deep theories even more so...

References

JOHANNES KEPLER UNIVERSITÄT LINZ

- Huynh P. V. Q., Fürnkranz, J., Beck, F.: Efficient learning of large sets of locally optimal classification rules. *Machine Learning* 112(2): 571-610 (2023). doi:10.1007/s10994-022-06290-w.
- Beck F., Fürnkranz J.: An Empirical Investigation into Deep and Shallow Rule Learning. *Frontiers in Artificial Intelligence* 4, 2021. doi:10.3389/frai.2021.689398
- Fürnkranz J., Kliegr T., Paulheim H.: On cognitive preferences and the plausibility of rule-based models. *Machine Learning* 109(4): 853-898 (2020) doi:10.1007/s10994-019-05856-5
- Kliegr T., Bahník S., Fürnkranz: A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence* 295:103458 (2021) doi:10.3389/frai.2021.689398



frontiers in Artificial Intelligence

- Beck F., Fürnkranz J.: Beyond DNF: First Steps towards Deep Rule Learning, in *Proceedings of the 21st Conference Information Technologies -- Applications and Theory (ITAT)*, pp. 61--68, 2021.
- Beck F., Fürnkranz J.: An Investigation into Mini-Batch Rule Learning, in *Proceedings of the 2nd Workshop on Deep Continuous-Discrete Machine Learning (DeCoDeML)*, 2020.
- Fürnkranz J., Hüllermeier E., Loza Mencía E., Rapp M.: Learning Structured Declarative Rule Sets – A Challenge for Deep Discrete Learning, in *Proceedings of the 2nd Workshop on Deep Continuous-Discrete Machine Learning (DeCoDeML)*, 2020.
- Fürnkranz J., Kliegr T.: The Need for Interpretability Biases. Proc. IDA 2018: 15-27
- Stecher J., Janssen F., Fürnkranz J.: Shorter Rules Are Better, Aren't They? Proc. DS 2016: 279-294.
- Stecher J., Janssen F., Fürnkranz J.: Separating Rule Refinement and Rule Selection Heuristics in Inductive Rule Learning. Proc. ECML/PKDD (3) 2014: 114-129