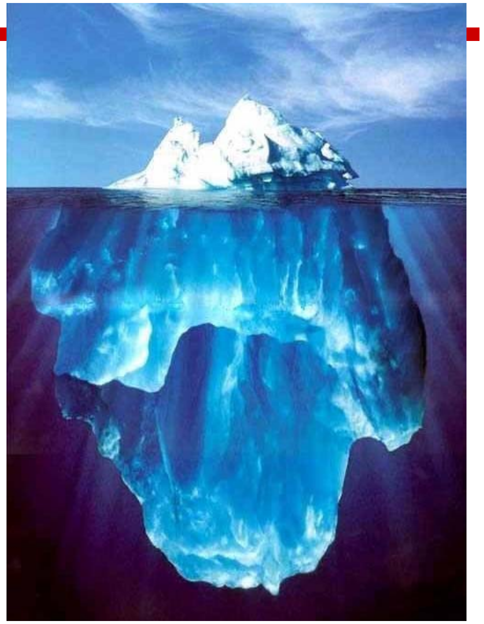
Formal Concept Analysis

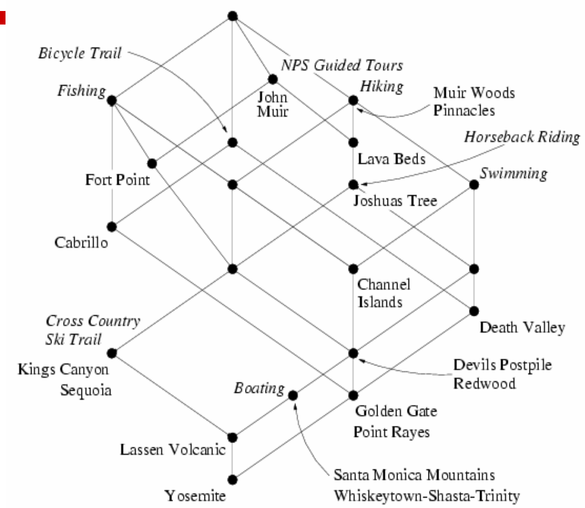


2 Closure Systems and Implications

5 Implications

Implications

Def.: An **implication** $X \rightarrow Y$ holds in a context, if every object having all attributes in X also has all attributes in Y.



- **Examples:**

{ Swimming } → { Hiking }

{ Boating } → { Swimming, Hiking, NPS Guided Tours, Fishing }

{ Bicycle Trail, NPS Guided Tours } → { Swimming, Hiking }

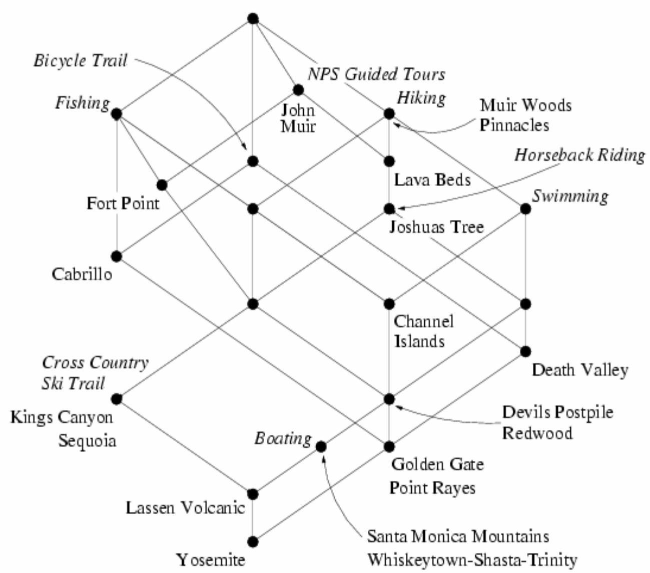
Independency

Def.: Let $X \subseteq M$. The attributes in X are **independent**, if there are no trivial dependencies between them.

Example:

- Fishing
- Bicycle Trail
- Swimming

are independent attributes.



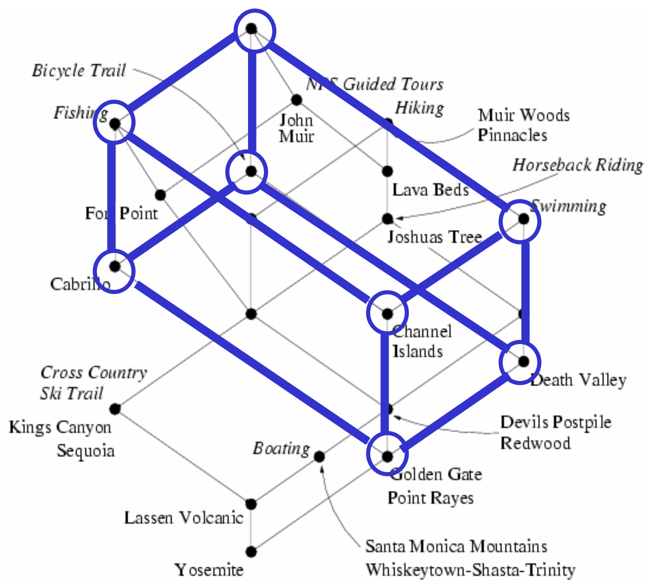
Independency

Lemma: Attributes are independent if they span a hyper-cube.

Example:

- Fishing
- Bicycle Trail
- Swimming

are independent attributes.



Concept Intents and Implications

Def.: A subset $T \subseteq M$ **respects** an implication $A \to B$, if $A \not\subseteq T$ or $B \subseteq T$.

T **respects a set** $\mathcal{L}$ of implications, if T respects every single implication in $\mathcal{L}$.

Lemma: An implication $A \to B$ holds in a context iff $B \subseteq A''$. It is then respected by all concept intents.

Lemma: Is $\mathcal{L}$ a set of implications in M, then

$$\mathcal{H}(\mathcal{L}) := \{ X \subseteq M \mid X \text{ respects } \mathcal{L} \}$$

is a closure system.

The related closure operator is constructed as follows:
For a set $X \subseteq M$ let

$$X^{\mathcal{L}} := X \cup \bigcup \{ B \mid A \to B \in \mathcal{L}, A \subseteq X \}.$$

Compute $X^{\mathcal{L}}, X^{\mathcal{L}\mathcal{L}}, X^{\mathcal{L}\mathcal{L}\mathcal{L}}$,..., until a set

$$\mathcal{L}(X) := X^{\mathcal{L}\ldots\mathcal{L}}$$

with $\mathcal{L}(X)^{\mathcal{L}} = \mathcal{L}(X)$ (i.e., a fix point) is reached. (for infinite contexts this may be an infinite process). $\mathcal{L}(X)$ ist then the closure of X with respect to the closure system $\mathcal{H}(\mathcal{L})$.

Bem.: Dies ist der Algorithmus AttrHülle der Datenbankvorlesung!

Def.: An implication $A \to B$ **is (semantically) entailed** from a set $\mathcal{L}$ of implications, if every subset of M respecting $\mathcal{L}$ also respects $A \to B$.
A family $\mathcal{L}$ of implications ist called **closed** if every implication entailed from $\mathcal{L}$ is already contained in $\mathcal{L}$.

Lemma: A set $\mathcal{L}$ of implications on M is closed iff the following conditions (Amstrong rules) are fulfilled for all $W, X, Y, Z \subseteq M$:

1. $X \to X \in \mathcal{L}$,
2. If $X \to Y \in \mathcal{L}$, then $X \cup Z \to Y \in \mathcal{L}$,
3. If $X \to Y \in \mathcal{L}$ and $Y \cup Z \to W \in \mathcal{L}$, then $X \cup Z \to W \in \mathcal{L}$.

Bem.: Auch diese Regeln sollten einem aus der Datenbankvorlesung bekannt vorkommen!

Def.: A set $\mathcal{L}$ of implications of a context (G, M, I) is called **complete**, if every implication of (G, M, I) is entailed from $\mathcal{L}$.
A set $\mathcal{L}$ of implications is called **non-redundant**, if no implication is entailed from the others.

Def.: $P \subseteq M$ is called **pseudo intent** of (G, M, I) if $P \neq P''$ and for every pseudo intent $Q \subseteq P$ with $Q \neq P$ holds $Q'' \subseteq P$.

Theorem: The set of implications

$$\mathcal{L} := \{ P \to P'' \mid P \text{ Pseudoinhalt} \}$$

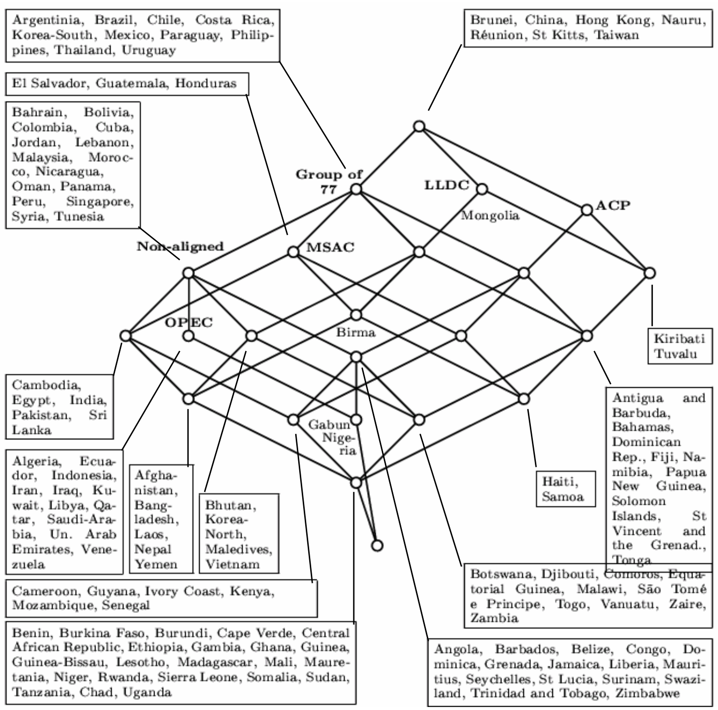
is non-redundant and complete. We call $\mathcal{L}$ **stem basis.**

Example: Membership of developing countries in supranational groups
(Source: Lexikon Dritte Welt. Rowohlt-Verlag, Reinbek 1993)

Taken from: B. Ganter, R. Wille: Formal Concept Analysis - Mathematical Foundations. Springer, Heidelberg 1999

	Group of 77	Non-aligned	LLDC	MSAC	OPEC	ACP
Afghanistan	×	×	×	×		
Algeria	×	×			×	
Angola	×	×				×
Antigua and Barbuda	×					×
Argentina	×					
Bahamas	×					×
Bahrain	×	×				
Bangladesh	×	×	×	×		
Barbados	×	×				×
Belize	×	×				×
Benin	×	×	×	×		×
Bhutan	×	×	×			
Bolivia	×	×				
Botswana	×	×	×			×
Brazil	×					
Brunei						
Burkina Faso	×	×	×	×		×
Burundi	×	×	×	×		×
Cambodia	×	×		×		
Cameroon	×	×		×		×
Cape Verde	×	×	×	×		×
Central African Rep.	×	×	×	×		×
Chad	×	×	×	×		×
Chile	×					
China						
Colombia	×	×				
Comoros	×	×	×			×
Congo	×	×				×
Costa Rica	×					
Cuba	×	×				
Djibouti	×	×	×			×
Dominica	×	×				×
Dominican Rep.	×					×
Ecuador	×	×			×	
Egypt	×	×		×		
El Salvador	×			×		
Equatorial Guinea	×	×	×			×
Ethiopia	×	×	×	×		×
Fiji	×					×
Gabon	×	×			×	×
Gambia	×	×	×	×		×
Ghana	×	×	×	×		×
Grenada	×	×				×
Guatemala	×			×		
Guinea	×	×	×	×		×
Guinea-Bissau	×	×	×	×		×
Guyana	×	×		×		×
Haiti	×		×	×		×
Honduras	×			×		
Hong Kong						
India	×	×		×		
Indonesia	×	×			×	
Iran	×	×			×	
Iraq	×	×			×	
Ivory Coast	×	×		×		×
Jamaica	×	×				×
Jordan	×	×				
Kenya	×	×		×		×
Kiribati			×			×
Korea-North	×	×	×			
Korea-South	×					
Kuwait	×	×			×	
Laos	×	×	×	×		
Lebanon	×	×				
Lesotho	×	×	×	×		×
Liberia	×	×				×
Libya	×	×			×	
Madagascar	×	×	×	×		×
Malawi	×	×	×			×
Malaysia	×	×				
Maledives	×	×	×			
Mali	×	×	×	×		×
Mauretania	×	×	×	×		×
Mauritius	×	×				×
Mexico	×					
Mongolia			×			
Morocco	×	×				
Mozambique	×	×		×		×
Myanmar	×		×	×		
Namibia	×					×
Nauru						
Nepal	×	×	×	×		
Nicaragua	×	×				
Niger	×	×	×	×		×
Nigeria	×	×			×	×
Oman	×	×				
Pakistan	×	×		×		
Panama	×	×				
Papua New Guinea	×					×
Paraguay	×					
Peru	×	×				
Philippines	×					
Qatar	×	×			×	
Réunion						
Rwanda	×	×	×	×		×
Samoa	×		×	×		×
São Tomé e Principe	×	×	×			×
Saudi Arabia	×	×			×	
Senegal	×	×		×		×
Seychelles	×	×				×
Sierra Leone	×	×	×	×		×
Singapore	×	×				
Solomon Islands	×					×
Somalia	×	×	×	×		×
Sri Lanka	×	×		×		
St Kitts						
St Lucia	×	×				×
St Vincent& Grenad.	×					×
Sudan	×	×	×	×		×
Surinam	×	×				×
Swaziland	×	×				×
Syria	×	×				
Taiwan						
Tanzania	×	×	×	×		×
Thailand	×					
Togo	×	×	×			×
Tonga	×					×
Trinidad and Tobago	×	×				×
Tunisia	×	×				
Tuvalu			×			×
Uganda	×	×	×	×		×
United Arab Emirates	×	×			×	
Uruguay	×					
Vanuatu	×	×	×			×
Venezuela	×	×			×	
Vietnam	×	×	×			
Yemen	×	×	×	×		
Zaire	×	×	×			×
Zambia	×	×	×			×
Zimbabwe	×	×				×

The abbreviations stand for: LLDC := Least Developed Countries, MSAC := Most Seriously Affected Countries, OPEC := Organization of Petrol Exporting Countries, ACP := African, Caribbean and Pacific Countries.



Stem basis of the 3rd World context:

{ OPEC } → { Group of 77, Non-Alligned }
{ MSAC } → { Group of 77 }
{ Non-Alligned } → { Group of 77 }
{ Group of 77, Non-Alligned, MSAC, OPEC } → { LLDC, AKP }
{ Group of 77, Non-Alligned, LLDC, OPEC } → { MSAC, AKP }

Determining the stem basis with Next-Closure

is based on the following theorem:

Theorem: The set of all intents and pseudo-intents is a closure system. Its corresponding closure operator is given as follows:

Starting from set *X* we compute successively

$$X^{\mathcal{L}^*} := X \cup \bigcup \{ B \mid A \to B \in \mathcal{L},\ A \subseteq X,\ A \neq X \}$$

$$X^{\mathcal{L}^*\mathcal{L}^*} := X^{\mathcal{L}^*} \cup \bigcup \{ B \mid A \to B \in \mathcal{L},\ A \subseteq X^{\mathcal{L}^*},\ A \neq X^{\mathcal{L}^*} \}$$

etc, until a set $\mathcal{L}^*(X)$ with $\mathcal{L}^*(X) = \mathcal{L}^*(\mathcal{L}^*(X))$ is reached. This is then the desired intent or pseudo-intent.

The first part of the theorem is proven by using the following lemma:

Lemma: If P and Q are concept intents or pseudo-intents with $P \neq Q$, $P \not\subset Q$, and $Q \not\subset P$, then $P \cap Q$ is a concept intent.

Proof: P and Q, and therefore also $P \cap Q$, respect all implications in $\mathcal{L} \setminus \{ P \to P'', Q \to Q'' \}$. If $P \neq P \cap Q \neq Q$, then $P \cap Q$ respects these implications, too, and is hence a concept intent.

Algorithm **Next-Closure** for computing all concept intents and the stem basis:

0) The set $\mathcal{L}$ of all implications is set to the empty set.

1) The lectically first intent or pseudo-intent is $\varnothing$.

2) Is *A* determined to be intent or pseudo-intent, then the lectically next intent/pseudo-intent is computed by checking all $i \in M \setminus A$ in decreasing order until $A <_i \mathcal{L}^*(A \bullet i)$ holds.
$\mathcal{L}^*(A \bullet i)$ is then the next intent or pseudo-intent.

3) If $\mathcal{L}^*(A \oplus i) = (\mathcal{L}^*(A \bullet i))''$, then $\mathcal{L}^*(A \bullet i)$ is a concept intent, else it is a pseudo-intent, and the implication $\mathcal{L}^*(A \bullet i) \to (\mathcal{L}^*(A \bullet i))''$ is added to $\mathcal{L}$.

4) If $\mathcal{L}^*(A \bullet i) = M$, then stop, else $A \leftarrow \mathcal{L}^*(A \bullet i)$ and continue at 2).

Example: on blackboard

	Mobil (1)	Telephone (2)	Fax (3)	Fax w. N.-Paper (4)
Sinus 44		X		
Nokia 6110	X	X		
T-Fax 301			X	X
T-Fax 360 PC			X	

A	i	$A \bullet i$	$\mathcal{L}^*(A \bullet i)$	$A <_i \mathcal{L}^*(A \bullet i)$?	$(\mathcal{L}^*(A \bullet i))''$	$\mathcal{L}$	intents

Association Rules

{ veil color: white, gill spacing: close } $\rightarrow$ { gill attachment: free }

Support: 78,52 % Confidence: 99,6 %

The input data of association rules algorithms can be written as a formal context (*G*,*M*,*I*):

- *M* is a set of items,
- *G* consists of the transaction IDs,
- and the relation *I* is the list of transactions.

Association Rules

{ veil color: white, gill spacing: close } $\rightarrow$ { gill attachment: free }

Support: 78,52 % Confidence: 99,6 %

The **support** is the percentage of all objects having all attributes in premise and conclusion:

Def.: The support of an attribute set $X \subseteq M$ is given by $$\text{supp}(X) = \frac{|X'|}{|G|}$$

The support of an association rule $X \rightarrow Y$ is given by supp ($X \rightarrow Y$) := supp ($X \cup Y$).

The **confidence** is the percentage of all objects fulfilling the premise among all objects fulfilling both premise and conclusion.

Def.: The confidence of a rule $X \rightarrow Y$ is given by $$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

Bases of Association Rules

{ veil color: white, gill spacing: close } $\rightarrow$ { gill attachment: free }

Support: 78,52 % Confidence: 99,6 %

Classical Data Mining Task: Find, for given minsupp, minconf $\in$ [0,1], all rules with support and confidence above these thresholds

Our task: Find a **basis** of rules, i.e., a minimal set of rules out of which all other rules can be derived.

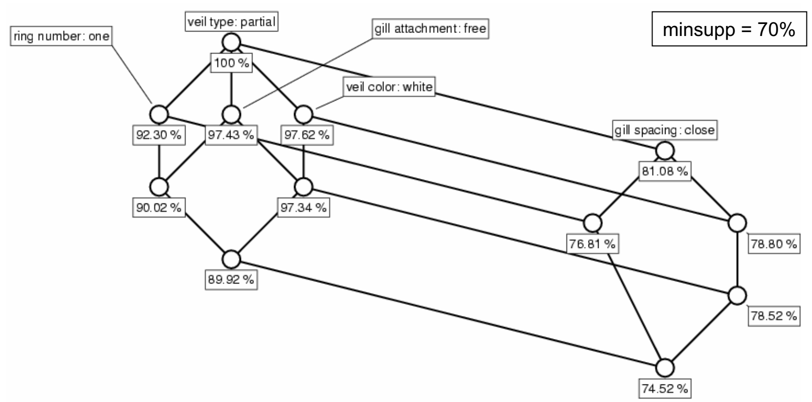
- From $B' = B'''$ follows

$$\text{supp}(B) = \frac{|B'|}{|G|} = \frac{|B'''|}{|G|} = \text{supp}(B'')$$

Theorem: $X \rightarrow Y$ and $X'' \rightarrow Y''$ have the same support and the same confidence.

Hence for computing association rules, it is sufficient to compute the supports of all frequent sets with $B = B''$ (i.e., the intents of the iceberg concept lattice).

Advantage of the use of iceberg concept lattices (compared to frequent itemsets)

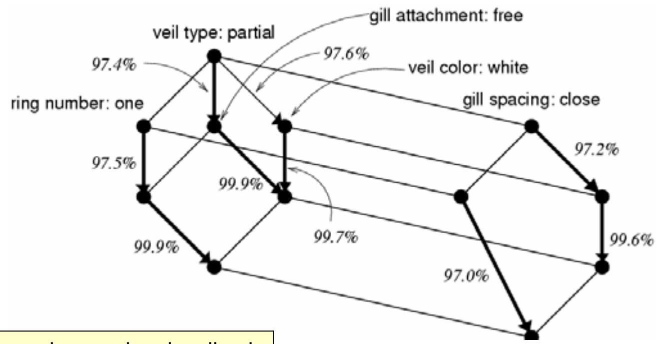


32 frequent itemsets are represented by 12 frequent concept intents

→ **more efficient computation (e.g. TITANIC)**

→ **fewer rules (without information loss!)**

Advantage of the use of iceberg concept lattices (compared to frequent itemsets)



Association rules can be visualized in the iceberg concept lattice:

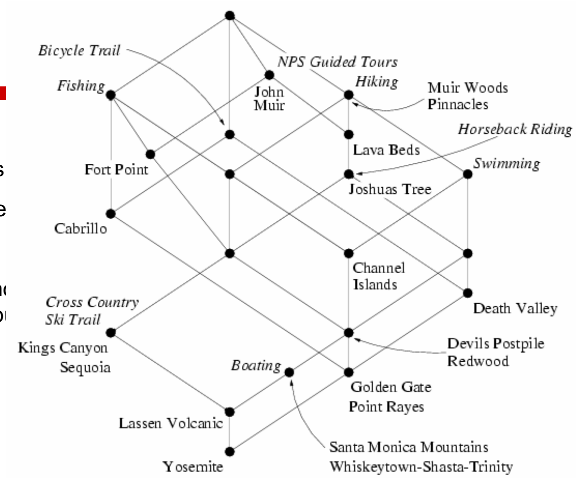
- **exact rules** ← conf = 100 %
- **approximate rules** ← conf < 100 %

Exact Association Rules

can be derived from the stem basis

In concept lattices, they can be dire

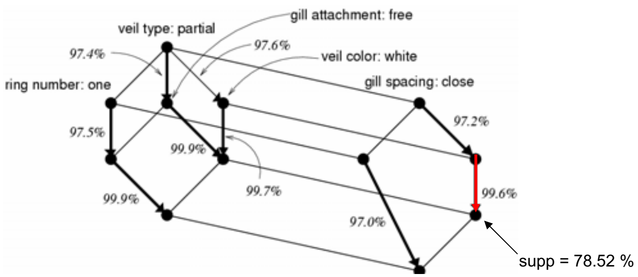
- **Lemma:** An implication $X \rightarrow Y$ h
all concepts generated by the attrib
by attributes in Y.



- **Examples:**
  - { Swimming } → { Hiking }
    (supp=10/19 ≈ 52.6%, conf = 100%)
  - { Boating } → { Swimming, Hiking, NPS Guided Tours, Fishing }
    (supp=4/19 ≈ 21.0%, conf = 100%)
  - { Bicycle Trail, NPS Guided Tours } → { Swimming, Hiking }
    (supp=4/19 ≈ 21.0%, conf = 100%)

Approximate Association Rules

Def.: The **Luxenburger basis** consists of all valid association rules $X \rightarrow Y$ such that there are concepts (A_1, B_1) and (A_2, B_2) where (A_1, B_1) is a direct upper neighbor of (A_2, B_2), $X = B_1$, and $X \cup Y = B_2$.

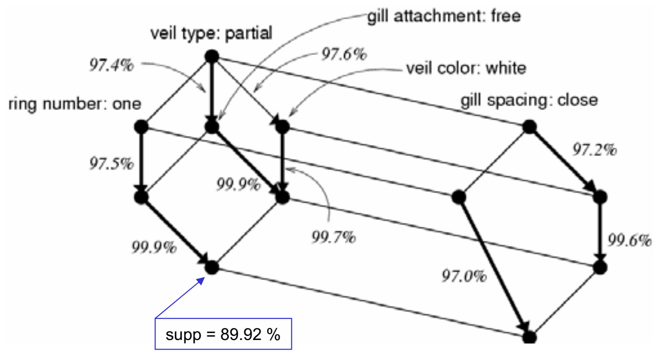


Each arrow indicates a rule of the basis, e.g. the rightmost arrow stands for { veil type: partial, gill spacing: close, veil color: white } → { gill attachment: free } (conf = 99.6 %, supp = 78.52 %)

Theorem: From the Luxenburger-Basis all approximate rules (incl. support und confidence) can be derived with the following rules:

- $\phi(X \to Y) = (X \to Y \setminus Z)$, für $\phi \in \{ \text{conf}, \text{supp} \}$, $Z \subseteq X$
- $\phi(X'' \to Y'') = \phi(X \to Y)$
- $\text{conf}(X \to X) = 1$
- $\text{conf}(X \to Y) = p$, $\text{conf}(Y \to Z) = q \Rightarrow \text{conf}(X \to Z) = p \cdot q$
  for all frequent concept intents $X \subset Y \subset Z$.
- $\text{supp}(X \to Z) = \text{supp}(Y \to Z)$, for all $X, Y \subseteq Z$.

The basis is minimal with this property.



Example:

{ ring number: one } → { veil color: white }
- has support 89.92 % (the support of the largest concept having both attributes in its intent)
- and confidence 97.5 % × 99.9 % ≈ 97.4 %.

Name	Number of objects	Average size of objects	Number of items
T10I4D100K	100,000	10	1,000
Mushrooms	8,416	23	127
C20D10K	10,000	20	386
C73D10K	10,000	73	2,177

Some experimental results

Dataset (Minsupp)	Exact rules	D.-G. basis	Minconf	Approximate rules	Luxenburger basis
T10I4D100K (0.5%)	0	0	90%	16,269	3,511
			70%	20,419	4,004
			50%	21,686	4,191
			30%	22,952	4,519
Mushrooms (30%)	7,476	69	90%	12,911	563
			70%	37,671	968
			50%	56,703	1,169
			30%	71,412	1,260
C20D10K (50%)	2,277	11	90%	36,012	1,379
			70%	89,601	1,948
			50%	116,791	1,948
			30%	116,791	1,948
C73D10K (90%)	52,035	15	95%	1,606,726	4,052
			90%	2,053,896	4,089
			85%	2,053,936	4,089
			80%	2,053,936	4,089