Data Mining

Prof. Dr. Stefan Kramer
Johannes Gutenberg-Universität Mainz

Outline

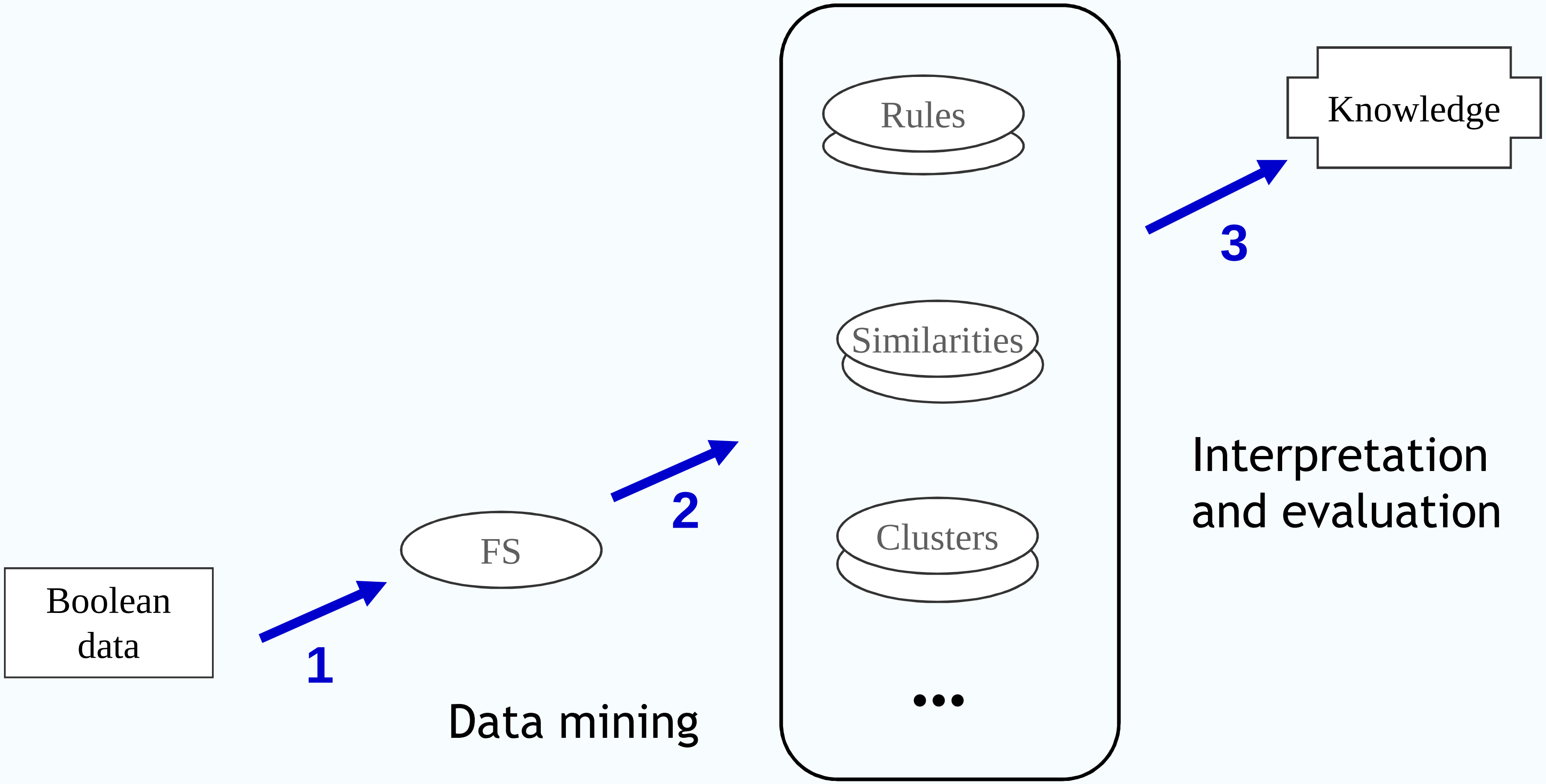
- Condensed representations: closed and free sets

Condensed Representations: Closed and Free Sets

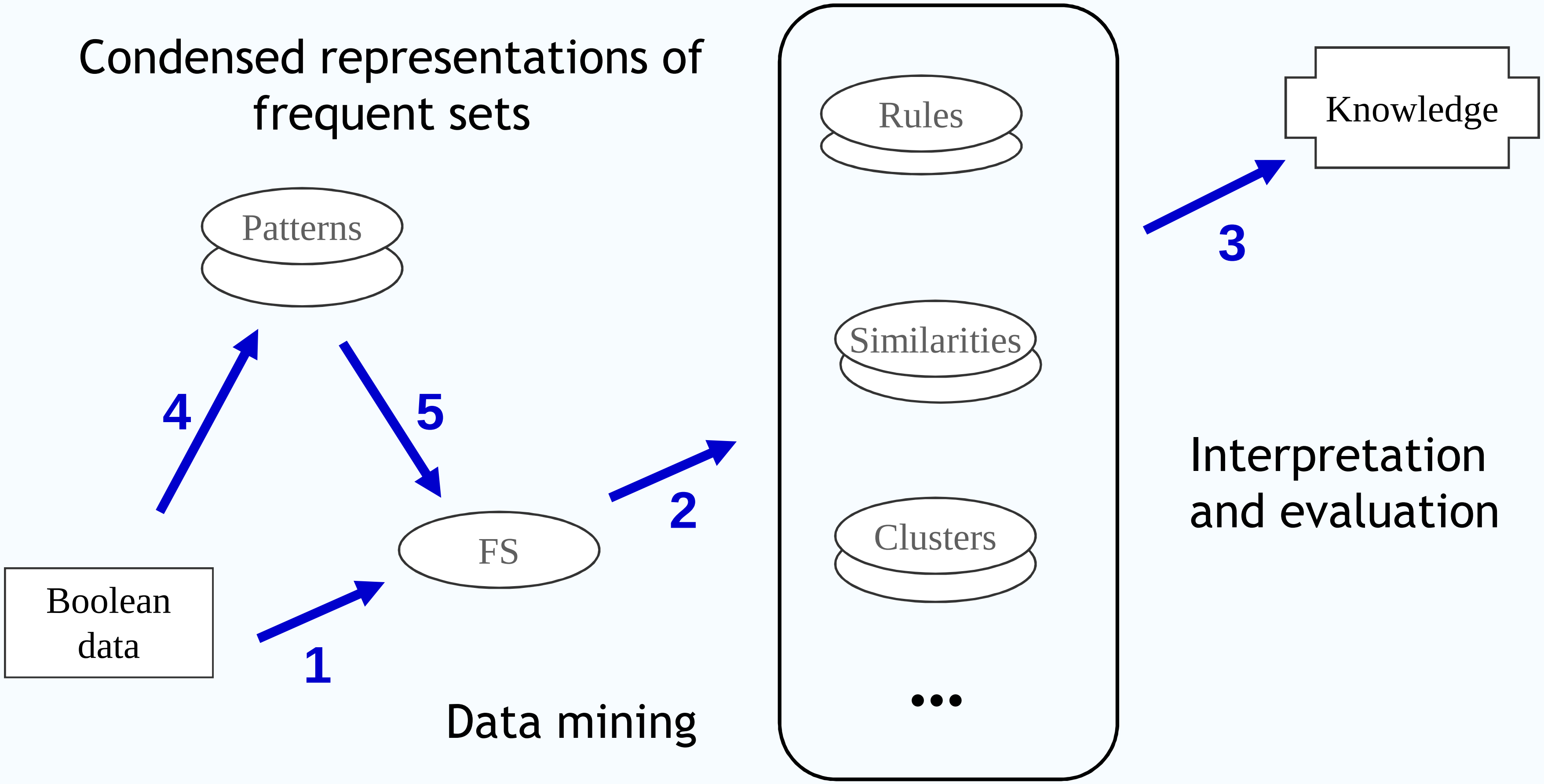
Condensed Representations: Motivation

- Problem of APriori-like approaches: computing frequent itemsets intractable in *dense* and *highly correlated Boolean* data (remember: exponential in the worst-case)
- Distinction: *sparse* and *dense* dataset
- Condensed representations: remove redundancy and provide more interesting patterns to the end-user

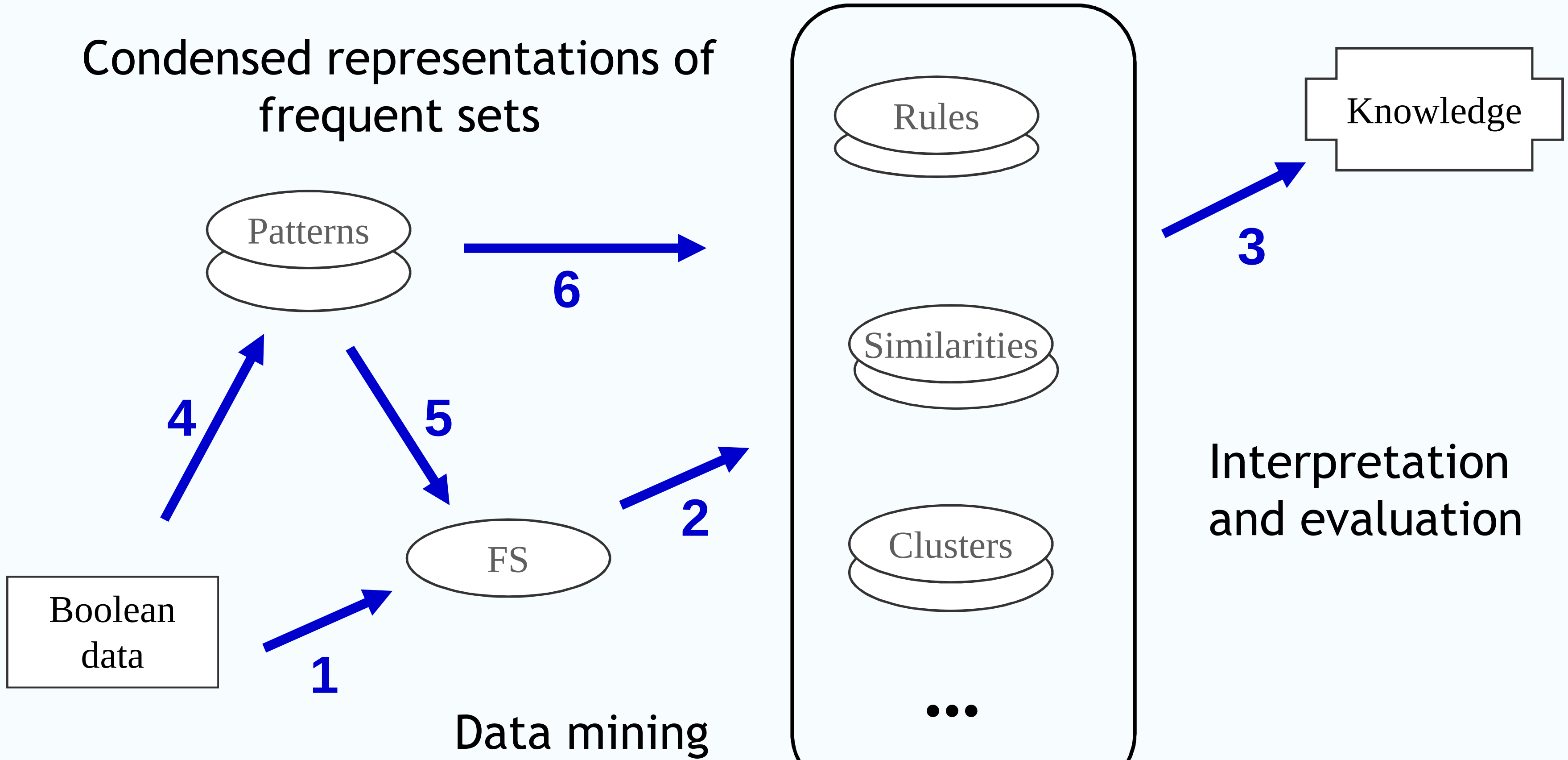
Multiples Uses of Frequent Itemsets



… Based on Condensed Representations



… Based on Condensed Representations



The "Closure" Evaluation Function

- The closure of X is the *maximal* superset of X that has exactly the same frequency as X (!)

closure(X, r) = items(objects(X, r), r)

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

closure({A}, r) = {A,C}

Note:
$A \Rightarrow C$ has confidence 1.0

Closed Sets

- X is a closed set iff X = closure(X, r). It is a maximal set of items that *support exactly the same transactions*.

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

{A,C} is closed {A,B} is not closed

$C_{Close}(S)$

- How about the empty set?
- *Closedness is* not *an anti-monotonic property!*

Closed Sets

- X is a closed set iff X = closure(X, r). It is a maximal set of items that *support exactly the same transactions*.

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

Frequent (MinSupport = 2)
A:3, B:4, C:4, D:2,
AB:2, AC:3, BC:3, BD:2,
ABC:2
Frequent closed:
B:4, C:4,
AC:3, BC:3, BD:2, ABC:2

Closed Sets

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

Frequent:
A:3, B:4, C:4, D:2,
AB:2, AC:3, BC:3, BD:2,
ABC:2
Frequent closed:
B:4, C:4,
AC:3, BC:3, BD:2, ABC:2

A	B
1	0
1	1
0	1
0	1
1	1

?

B	D
0	0
1	0
1	1
1	1
1	0

?

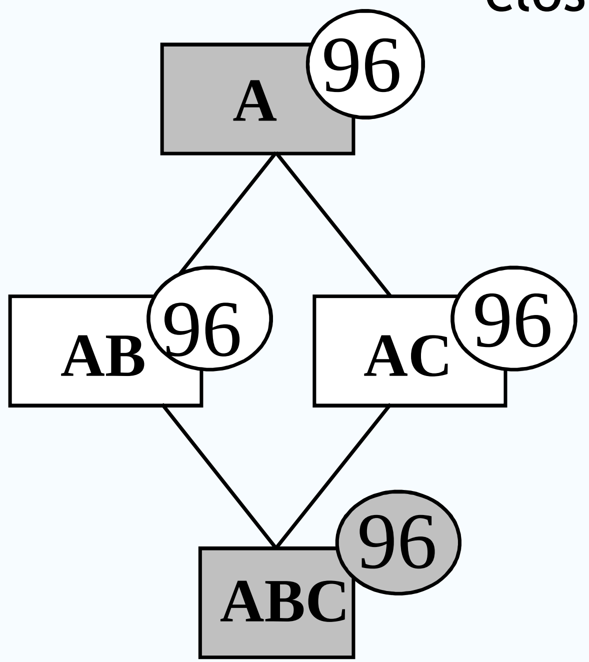
Possible: *confidence 1.0 (logical) rules:*
$A \rightarrow C$, $D \rightarrow B$, $AB \rightarrow C$

Properties of the Closure

- $X \subseteq closure(X)$
- $closure(closure(X)) = closure(X)$
- $Y \subseteq X \Rightarrow closure(Y) \subseteq closure(X)$

Using Closed Sets

closure(A)={ABC}



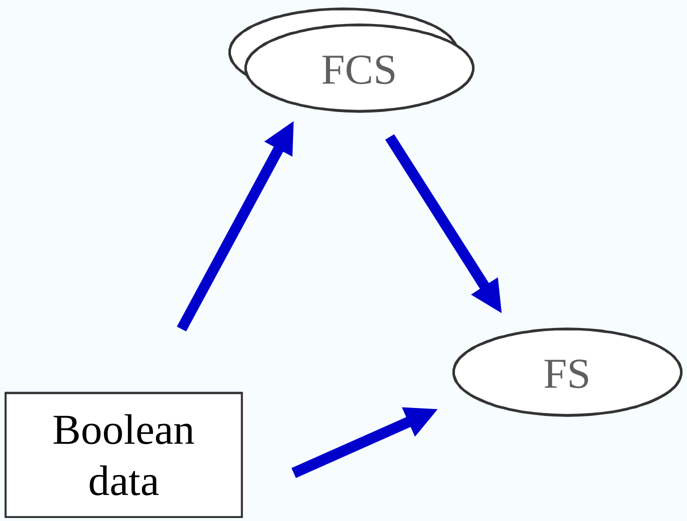
closure({ABC})={ABC}

Comparison with Maximally Specific Itemsets and Borders

- With borders/version spaces:
  possible to generate *all solution patterns*
- With frequent closed sets:
  possible to generate *all solution patterns along with their frequencies*

Closed Sets and How to Use Them

When S is frequent, choose the frequent closed set X s.t. $S \subseteq X$ that has the maximal support and return freq(S,r) = freq(X,r)



Example Frequent Closed Sets

1	ABCD
2	AC
3	AC
4	ABCD
5	BC
6	ABC

16 frequent sets

1 maximal frequent set

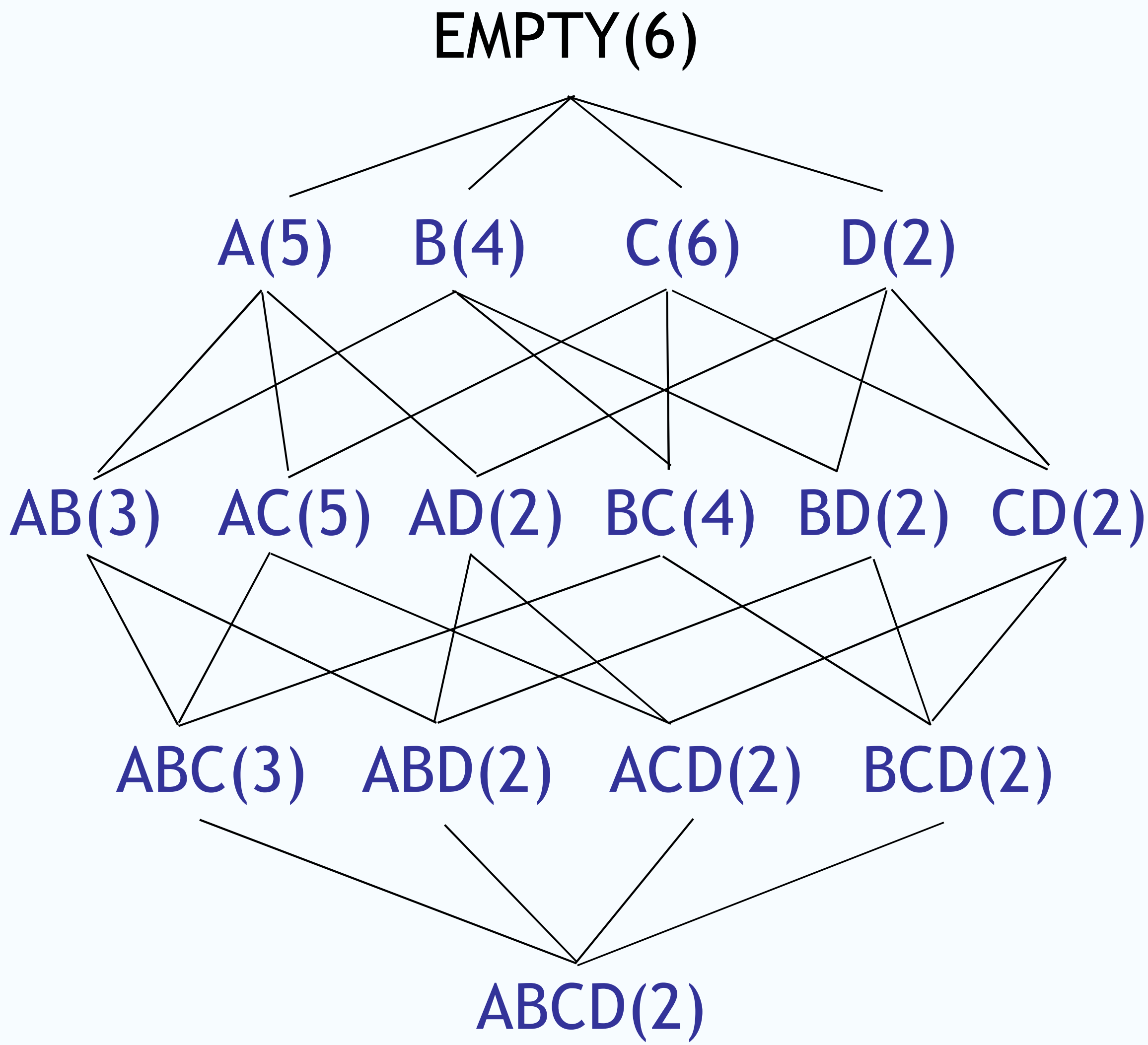
5 frequent closed sets

C, AC, BC, ABC, ABCD

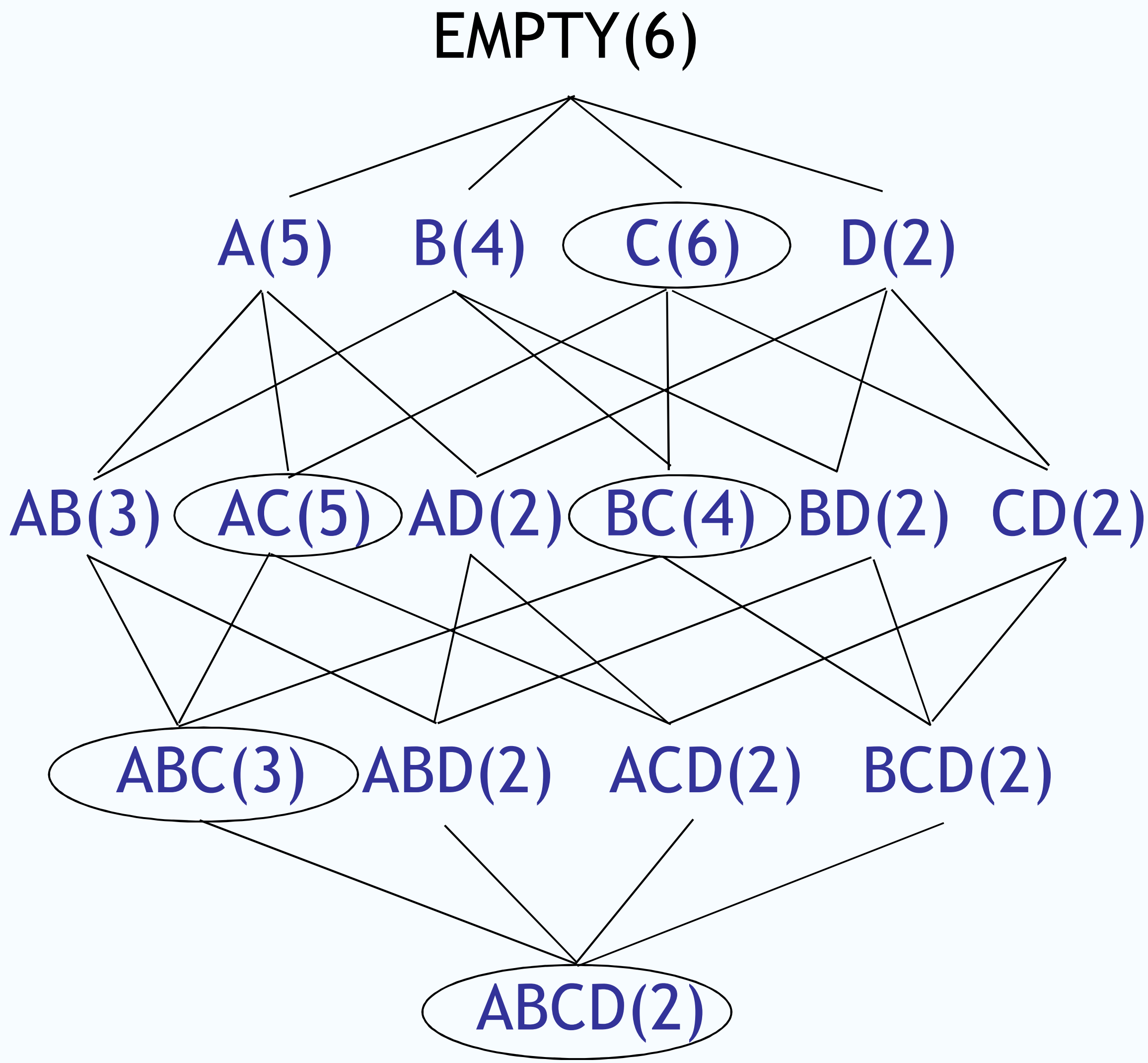
$A \rightarrow C$, $B \rightarrow C$, $AB \rightarrow C$, $ABD \rightarrow C$, etc.

Minimum frequency threshold = 2

Example: **Closed Sets?**



Example: Closed Sets!



Free Sets

- An itemset X is called *free* if *every proper subset* of X has a frequency *strictly greater* than that of X.
- In other words, X is a free set iff there is no logical (confidence 1.0) rule that holds between any of its subsets
- Free sets are special cases of δ-free sets (see next slides), *closed sets are the closures of free sets*

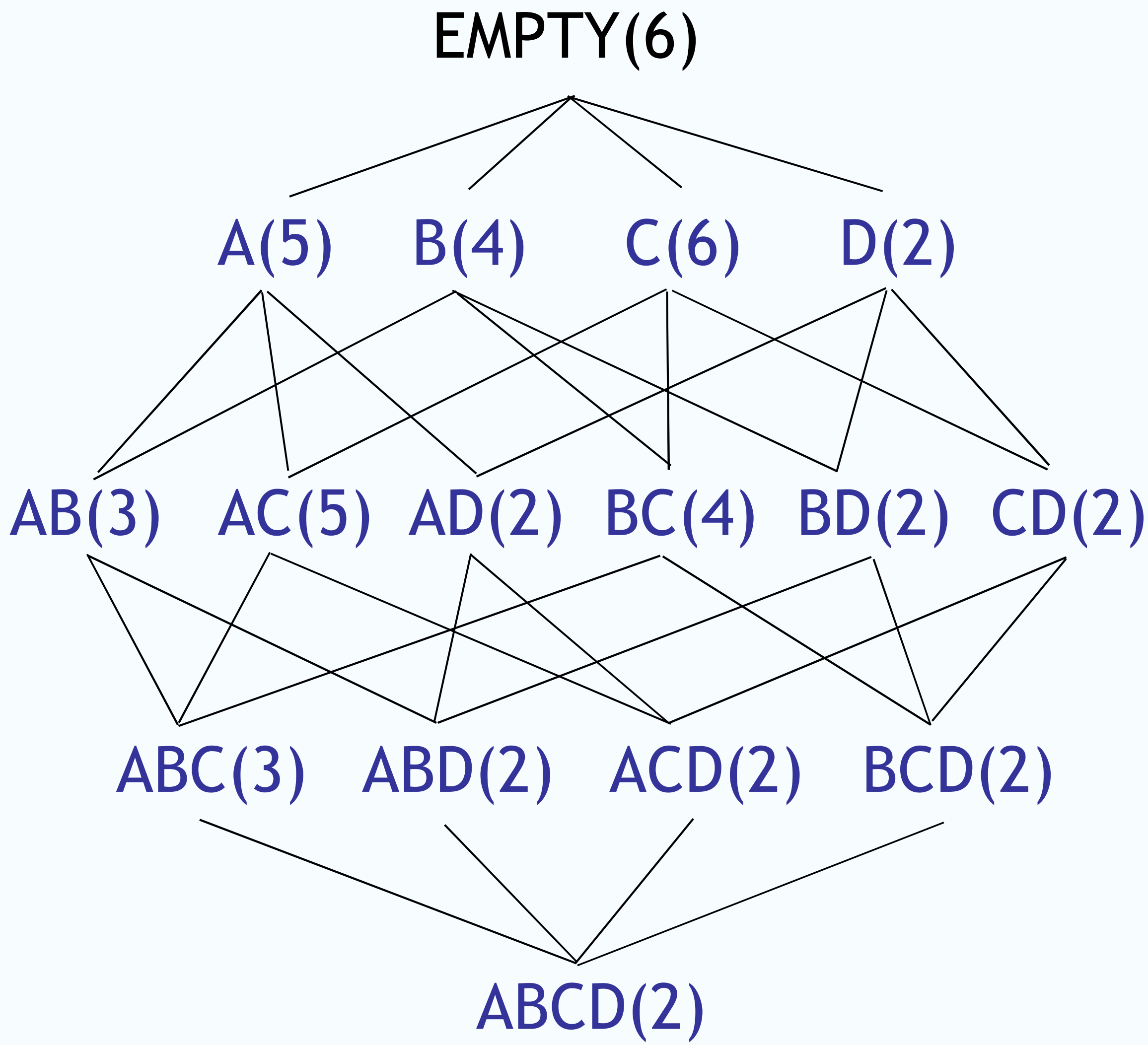
(how about the empty set?)

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

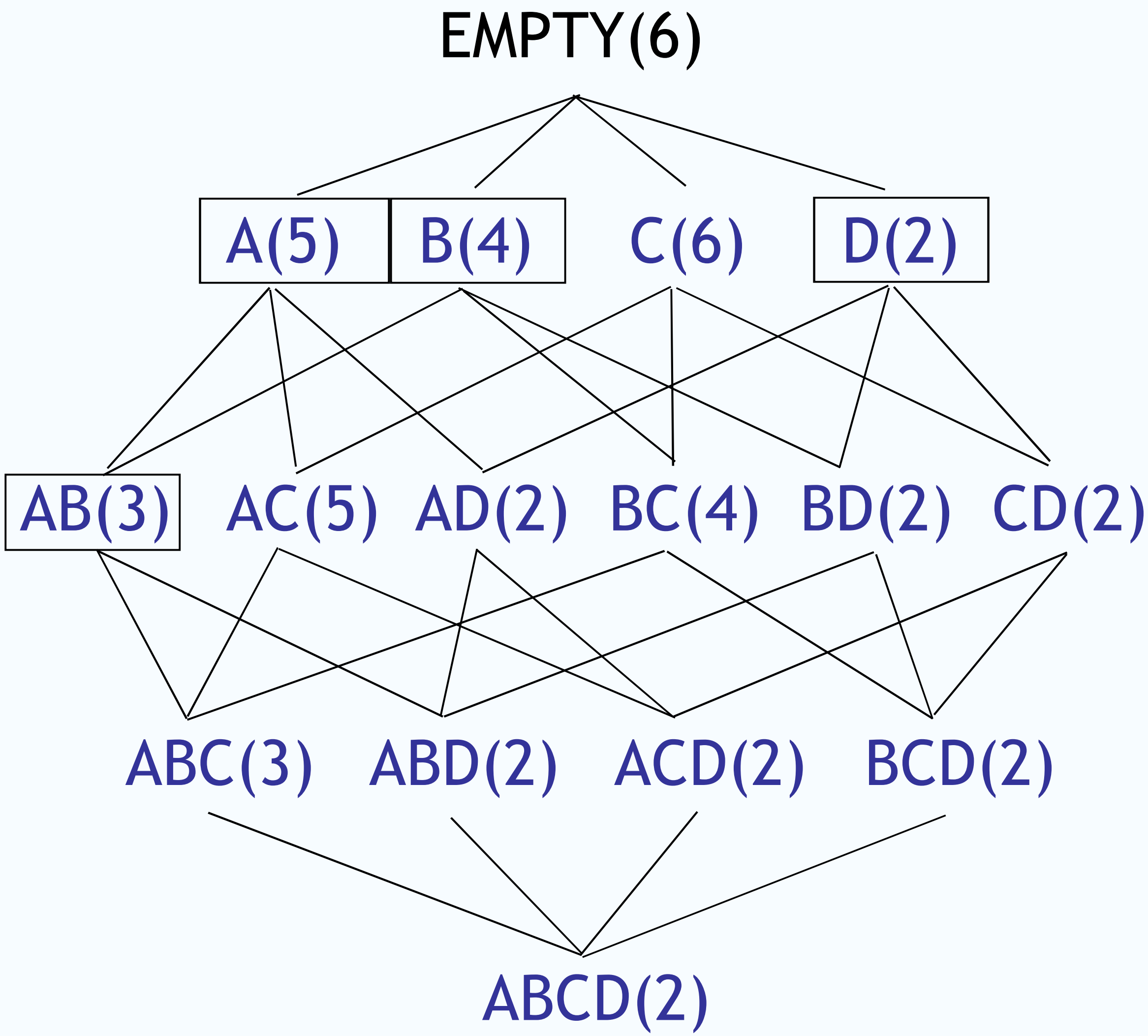
{A,B} is free {A,C} is not free

$C_{Free}(S)$ ***(anti-monotonic!)***

Example: Free Sets?

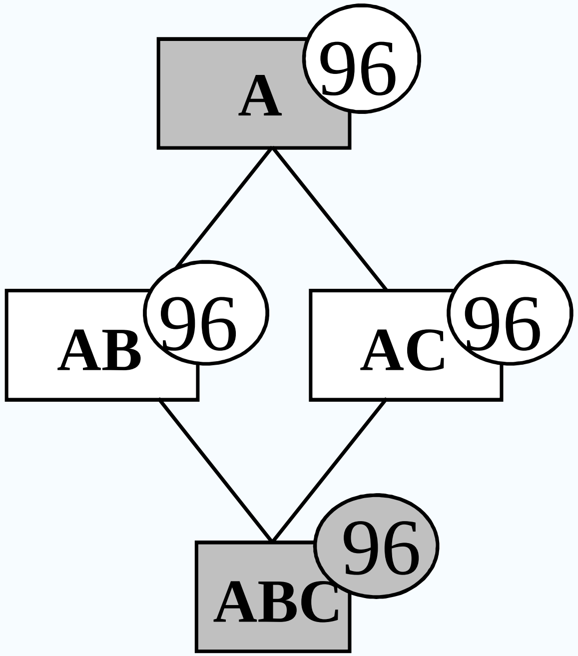


Example: Free Sets



Closed and Free Sets

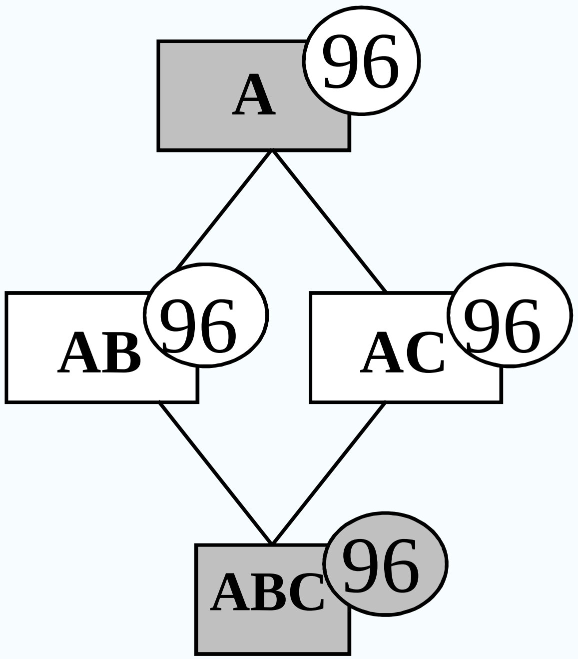
closure(A)={ABC}



closure({ABC})={ABC}

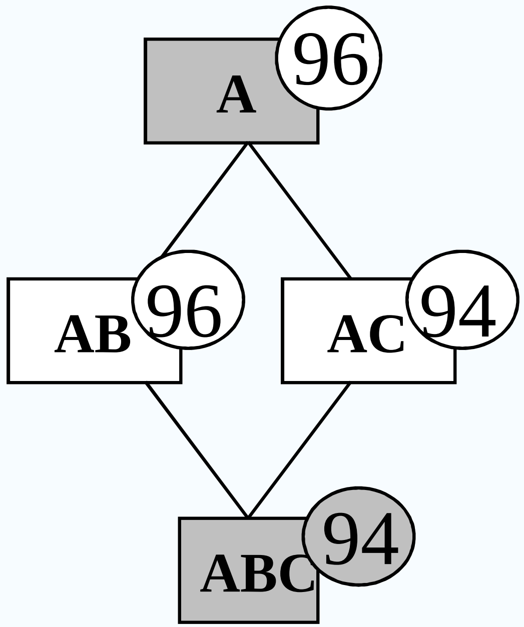
Closures and δ-Closures

closure(A)={ABC}



closure({ABC})={ABC}

B,C $\in$ closure$_\delta$(A)



δ-Freeness 1

- A δ-free-set is such that there is no δ-strong rule that holds between any of its subsets
- $X \Rightarrow_\delta Y$ is δ-strong if it has at most δ exceptions

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

{A,B} is free, but not 1-free

$C_{\delta\text{-free}}(S)$ (anti-monotonic)

δ-Freeness 2

- ...is (as, e.g., the minimum frequency constraint) *anti-monotonic*! Any subset of a delta-free itemset is also delta-free
- Any superset of a non-delta-free itemset is also non-delta-free
- ...provides a *condensed representation*: frequent free itemsets are less numerous than frequent itemsets while providing almost the same information

δ-Freeness 3

- If X is a frequent free itemset then possible to derive frequencies of supersets of X without having *to count them*
- Compute closure F of X = maximal superset such that frequency is that of X; every set between F and X has frequency of X

APriori Can Be Used to Solve Any Anti-Monotonic Constraint

Most important modification here from:

$$\mathcal{F}_l(r) := \{X \in \mathcal{C}_l \mid \mathit{fr}(X, r) \geq \mathit{min\_fr}\};$$

to:

$$\mathcal{F}_l(r) := \{X \in \mathcal{C}_l \mid \mathit{fr}(X, r) \geq \mathit{min\_fr} \text{ and}$$

$$\mathit{fr}(X, r) \neq \mathit{fr}(Y, r) \text{ for all } Y \subset X\};$$

Discovery of All Frequent Closed Sets

- Find all frequent free sets in the described manner
- Compute closures of frequent free sets from the database
  - determine transactions, where they occur, and intersect them

Examples of Condensed Representations

1	ABCD
2	AC
3	AC
4	ABCD
5	BC
6	ABC

16 frequent sets

1 maximal frequent set

Frequent closed sets

C, AC, BC, ABC, ABCD

Frequent free sets

$\emptyset$, A, B, D, AB

Frequent 1-free sets

$\emptyset$, B, D

Minimum frequency threshold = 2

Example Closed and Free Sets