Mining association patterns (Part 2)

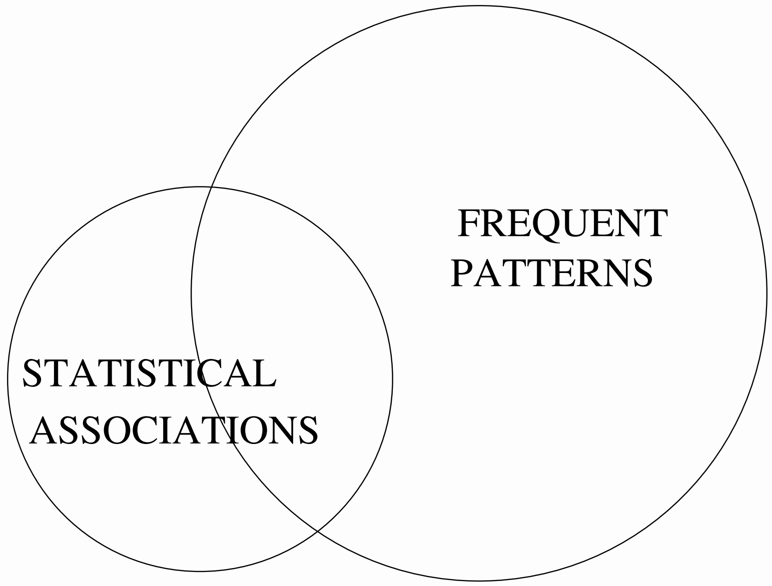
milk, cheese and bread
are often bought together

genes g1, g2, g3 and g4
are often over–expressed
in DLBC lymphomas

occurrence of certain insect species
makes it more likely to meet the
threatened white–backed woodpecker

How to find statistical association rules efficiently?

Problem: 2-step approach (frequents sets + postprocessing) very slow and often incomplete.



Recall experiments:

data	discovered %
Mushroom	100%
Chess	1%
T10I4D100K	100%
T40I10D100K	0%
Accidents	14%
Pumsb	27%
Retail	100%

Contents

1. Properties of statistical association rules
2. Pruning the search space
3. Algorithms

1. Properties of statistical association rules

Goal: Find association rules $\mathbf{X} \rightarrow C{=}c$ such that

1. rule expresses **statistical dependence**

$$P(\mathbf{X}, C{=}c) > P(\mathbf{X})P(C{=}c)$$

2. discovery is **statistically significant** (not spurious/due to chance) $\Rightarrow$ likely to hold in future data
3. pattern is specialized only if improvement is statistically significant $\Rightarrow$ **overfitting avoidance**
4. (optional) pattern is **not otherwise specious** (misleading)
   - prune out associations that definitely do not present any causal relationship

Recall: Negative dependence between $\mathbf{X}$ and C = positive dependence between $\mathbf{X}$ and $\neg C$

	C	$\neg C$	Σ
$\mathbf{X}$	$fr(\mathbf{X}C) =$ $fr(\mathbf{X})P(C)+n\delta$	$fr(\mathbf{X}\neg C) =$ $fr(\mathbf{X})P(\neg C)-n\delta$	$fr(\mathbf{X})$
$\neg\mathbf{X}$	$fr(\neg\mathbf{X}C) =$ $fr(\neg\mathbf{X})P(C)-n\delta$	$fr(\neg\mathbf{X}\neg C) =$ $fr(\neg\mathbf{X})P(\neg C)+n\delta$	$fr(\neg\mathbf{X})$
Σ	$fr(C)$	$fr(\neg C)$	n

$\Rightarrow$ search rules of form $\mathbf{X} \rightarrow C$ or $\mathbf{X} \rightarrow \neg C$ expressing positive dependence between condition and consequent

---

$\delta = \delta(\mathbf{X}, C) = P(\mathbf{X}, C) - P(\mathbf{X})P(C)$

1.1 Could we use δ or γ for search?

1. Should we use δ or γ?
   $\Rightarrow$ variable-based or value-based interpretation?
2. Is high δ or γ enough to guarantee good patterns?
3. How do we perform the search?
   - statistical dependence is **not a monotonic property**!
   - $AB \rightarrow C$ can express significant dependence, even if $A \rightarrow B$, $A \rightarrow C$ and $B \rightarrow C$ expressed independence [a]

[a] Voluntary home task: invent an example where this holds!

Value-based and variable-based interpretation of association $\mathbf{X} \rightarrow C = c$

Let $I_{\mathbf{X}}$ be an indicator variable:
$I_{\mathbf{X}} = 1$, if $\mathbf{X}$ holds, and $I_{\mathbf{X}} = 0$, otherwise.

> **Important**: Are we interested in association between **values** $I_{\mathbf{X}}{=}1$ and $C{=}c$ or between binary **variables** $I_{\mathbf{X}}$ and C?

$\Rightarrow$ **different goodness measures and different results!**

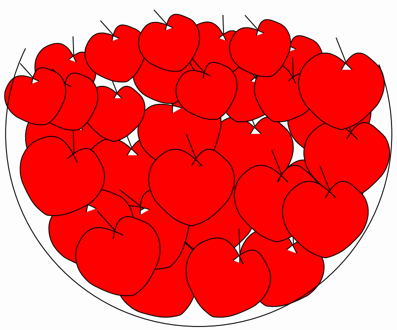
- **lift** γ measures strength of association between values
- **leverage** δ measures also strength of association between variables

Example: classifying apples by taste

Which rule should you choose? It depends whether you want a strong association between values or variables!

red $\rightarrow$ sweet

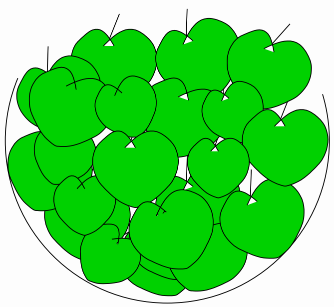
δ=0.22, γ=1.67



Basket 1

60 red apples

92% sweet



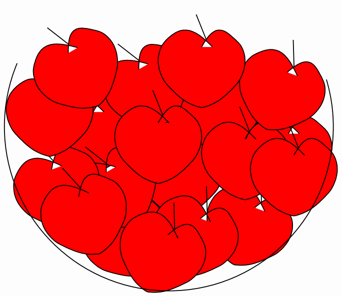
Basket 2

40 green apples

100% bitter

red$\wedge$big $\rightarrow$ sweet

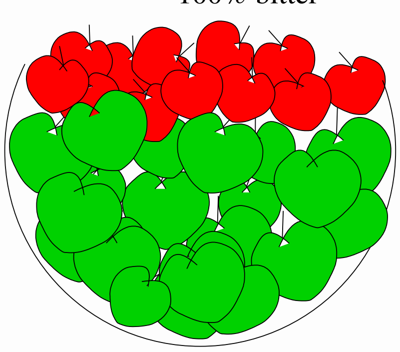
δ=0.18, γ=1.82



Basket 1

40 large red apples

100% sweet



Basket 2

40 green + 20 small red apples

75% bitter

$n = 100$

fr(sweet)=55

Is statistical dependence enough? Example

1000 students participated Mega Party. Data tells what each of them consumed in the party and what happened to them.

num	rule	fr_X	fr_C	fr_{XC}
1	*peppermint tea, sushi, chili sauce, sour cream* → *corona*	1	1	1
2	*vodka, sauerkraut, salmon* → *headache*	1	100	1
3	*cake* → *exam failure*	500	500	270
4	*magic mushrooms* → *intoxication*	20	20	20
5	*vodka* → *headache*	100	100	80
6	*vodka, salmon* → *headache*	40	100	30
7	*alcohol* → *exam failure*	333	500	300
8	*alcohol* → *cake*	333	500	200

Problem: High lift can be misleading

a) 1 student got corona ($=C$)! The student was the only one who had combination $\mathbf{X}$ = peppermint tea, sushi, chili sauce, sour cream. $\gamma(\mathbf{X}, C) = \frac{P(\mathbf{X}C)}{P(\mathbf{X})P(C)} = \frac{1}{P(C)} = n$.

b) 100 students had headache ($=C$), including one with unique combination $\mathbf{X}$ = vodka, sauerkraut, salmon. $\gamma(\mathbf{X}, C) = \frac{1}{P(C)} = \frac{1}{0.1} = 10$.

- Lift favors rare rules: $\gamma(\mathbf{X}, C) \leq \frac{1}{\max\{P(\mathbf{X}), P(C)\}}$

Problem: Leverage does not tell significance

Which rule is more significant?

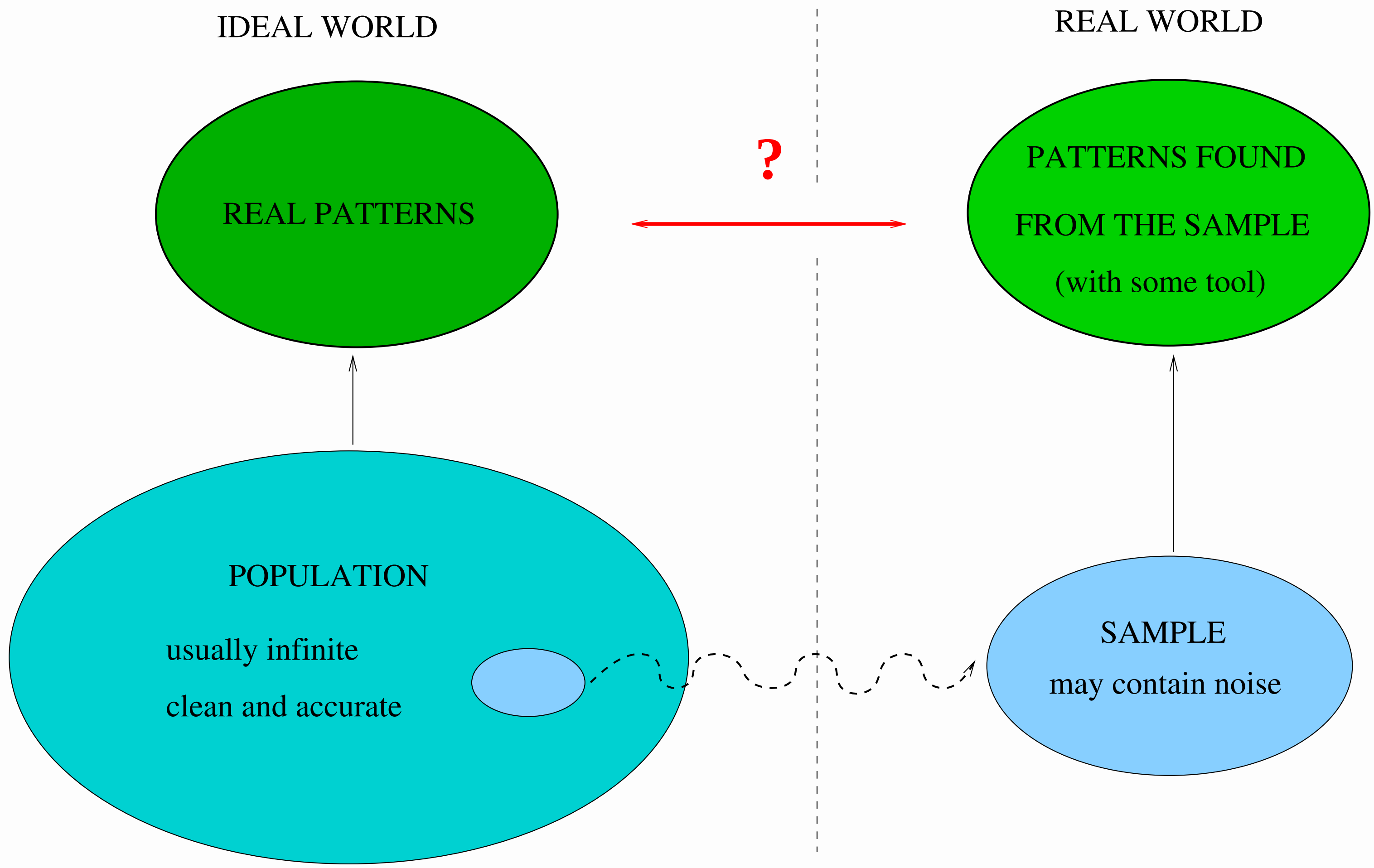
a) 500 students had chocolate cake ($\mathbf{X}$) and 500 failed the next day exam (C), including 270 cake eaters.
$\delta(\mathbf{X}, C) = \frac{1}{n^2}(n \cdot 270 - 500 \cdot 500) = 0.02$.

b) 20 students tried magic mushrooms ($\mathbf{X}$) and only them got a serious intoxication (C).
$\delta(\mathbf{X}, C) = \frac{1}{n^2}(n \cdot 20 - 20 \cdot 20) = 0.0196$.

- $\delta(\mathbf{X}, C) \leq \min\{P(\mathbf{X})P(\neg C), P(\neg \mathbf{X})P(C)\}$
- Leverage favors rules where $P(\mathbf{X}) \approx P(C) \approx 0.5$

$\Rightarrow$ What is the probability of observing such strong associations **by chance**?

1.2 Statistical significance: problem



Statistical significance

Idea: Given pattern, estimate probability of observing at least such a strong association, if the pattern actually expressed independence.

- if probability (p-value) is very small, the pattern is likely true
- many ways to estimate the probability
  - analytically (frequentist and Bayesian approaches)
  - empirically (randomization testing)

Further reading: Hämäläinen & Webb (2019): A tutorial on statistically sound pattern discovery, Section 2.3

Significance measures for rules $\mathbf{X} \rightarrow C{=}c$

1. **Fisher's exact p-value**:

$$p_F = \sum_{i=1}^{J} \frac{\binom{fr(\mathbf{X})}{fr(\mathbf{X}C=c)+i}\binom{fr(\neg\mathbf{X})}{fr(\neg\mathbf{X}C\neq c)+i}}{\binom{n}{fr(C=c)}},$$

where $J = \min\{fr(\mathbf{X}C\neq c), fr(\neg\mathbf{X}C=c)\}$

- very robust! use when possible
- often $ln(p_F)$ more convenient in programs
- remember: small values good (for p_F and $\ln(p)$)

cake → exam failure $p_F = 6.8e{-}3$ $(\ln(p) = -5.0)$
mushrooms → intoxication $p_F = 3.0e{-}42$ $(\ln(p) = -95.6)$

Significance measures for rules $\mathbf{X} \rightarrow C{=}c$

2. **Mutual information**:

$$MI = \log \frac{P(\mathbf{X}C)^{P(\mathbf{X}C)} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)} P(\neg\mathbf{X}C)^{P(\neg\mathbf{X}C)} P(\neg\mathbf{X}\neg C)^{P(\neg\mathbf{X}\neg C)}}{P(\mathbf{X})^{P(\mathbf{X})} P(\neg\mathbf{X})^{P(\neg\mathbf{X})} P(C)^{P(C)} P(\neg C)^{P(\neg C)}}$$

- now large values good
- you can get p-values for G statistic: $G = 2n \cdot MI$ (base e) or $G = 2n \cdot MI/log_2(e)$ (base 2)

3. χ^2**-measure**:

$$\chi^2 = \frac{n(P(\mathbf{X},C) - P(\mathbf{X})P(C))^2}{P(\mathbf{X})P(\neg\mathbf{X})P(C)P(\neg C)} = \frac{n\delta^2(\mathbf{X},C)}{P(\mathbf{X})P(\neg\mathbf{X})P(C)P(\neg C)}$$

- you can look p-values from the χ^2 distribution
- sensitive to accuracy of assumptions

Multiple hypothesis testing problem

Problem: The more patterns we test, the more will pass significance tests by chance.

E.g., if threshold $\alpha = 0.05$ and we test 10 000 spurious patterns, then about 500 patterns will incorrectly pass the test.

$\Rightarrow$ p-values need to be very small in DM!
$\Rightarrow$ Many strategies

Further reading: Hämäläinen & Webb (2019): A tutorial on statistically sound pattern discovery, Section 6

1.3 Problem: Overfitted rules can be misleading

- Since r5 vodka $\rightarrow$ headache strong, no surprise that r6 vodka, salmon $\rightarrow$ headache is strong!
- If you knew only r6, you might think that vodka alone is safe!
- Possible that salmon and headache are **conditionally independent** given vodka or salmon may even prevent vodka headache (=**negative conditional dependence**)

$\mathbf{Q}$ and C are **conditionally independent**, given $\mathbf{X}$ $\Leftrightarrow$

$$P(\mathbf{Q}, C|\mathbf{X}) = P(\mathbf{Q}|\mathbf{X})P(C|\mathbf{X}) \Leftrightarrow$$

$$P(\mathbf{X}, \mathbf{Q}, C) = P(\mathbf{X}, \mathbf{Q})P(C|\mathbf{X})$$

Problem: Specious associations are misleading

What about r3 cake $\rightarrow$ failure ($C \rightarrow F$)? ($fr = 270$)

- r7 alcohol $\rightarrow$ failure ($A \rightarrow F$) very strong, $P(F|A) = 0.9$ vs. $P(F|\neg A) = 0.3$
- If C independent of F given A, $P(F|CA) = 0.9$, so $CA \rightarrow F$ also strong (overfitted or redundant)
- If C independent of F given $\neg A$, $P(F|C\neg A) = 0.3$
- $fr(CA) = 200$ and $fr(C\neg A) = 300$
- Now expectation for $fr(CF) = fr(CAF) + fr(C\neg AF)$ is $fr(CA)P(F|A) + fr(C\neg A)P(F|\neg A) = 200 \cdot 0.9 + 300 \cdot 0.3 = 180 + 90 = 270$
- Rule cake $\rightarrow$ failure was just a side-product (specious)!

How to identify overfitted rules?

- $\mathbf{XQ} \rightarrow C$ can improve $\mathbf{X} \rightarrow C$ only, if $P(C|\mathbf{XQ}) > P(C|\mathbf{X})$
- If $P(C|\mathbf{XQ}) = P(C|\mathbf{X})$, then conditional independence
- What if $P(C|\mathbf{XQ})$ just slightly higher than $P(C|\mathbf{X})$? (may be due to chance)
- $\Rightarrow$ test **statistical significance of improvement!**
  - Assume conditional independence and estimate the probability of the observed improvement (or more)
  - $\Rightarrow$ measures M_C for evaluating conditional dependence

When improvement is significant?

- **Value-based** associations: evaluate $M_C(\mathbf{XQ} \rightarrow C \mid \mathbf{X} \rightarrow C)$
- **Variable-based** associations more tricky!
- **problem**: adding $\mathbf{Q}$ to $\mathbf{X} \rightarrow C$ may improve $P(C|X)$, but worsen $P(\neg C|\neg \mathbf{X})$
  - $P(Sweet|Red) = 0.92 < 1.0 = P(Sweet|Red, Big)$ but $P(\neg Sweet|\neg Red) = 1.0 > 0.75 = P(\neg Sweet|\neg(Red, Big))$
- evaluate $M_C(\mathbf{XQ} \rightarrow C \mid \mathbf{X} \rightarrow C)$ and $M_C(\neg\mathbf{X} \rightarrow \neg C \mid \neg(\mathbf{XQ}) \rightarrow \neg C)$

M = significance measure for unconditional dependence
M_C = corresponding measure for conditional dependence

Lessons to learn

1. Importance of **statistical significance**
   - significant discoveries are likely to hold in future data
   - remember multiple hypothesis testing problem
2. Decide whether you need **variable-based or value-based** definition
   - choose right measures and tools
3. Overfitted (too specialized) rules can be misleading (vodka, salmon → headache)
   - but sometimes other types of rules may also be specious (cake → exam failure)

2. Pruning the search space

Problem: Statistical association and significance are **not monotone** properties!

- given goodness measure M reflecting strength or significance of association, $M(\mathbf{X} \rightarrow C{=}c) = f(n, fr(\mathbf{X}C{=}c), fr(\mathbf{X}), fr(C{=}c))$
- M is **increasing by goodness** (ibg) if high values good (δ, γ, χ^2, MI) or **decreasing by goodness** (dbg) if small values good (p_F, $\ln(p_F)$)
- if M ibg, we may have $M(\mathbf{XQ} \rightarrow C{=}c) > M(\mathbf{X} \rightarrow C{=}c)$
- **How to prune exponential search space?**

Basic idea: branch and bound search using monotonic upper or lower bounds for M

Idea: In set X evaluate
monotonic upper or lower
bounds for M(XQ → C)

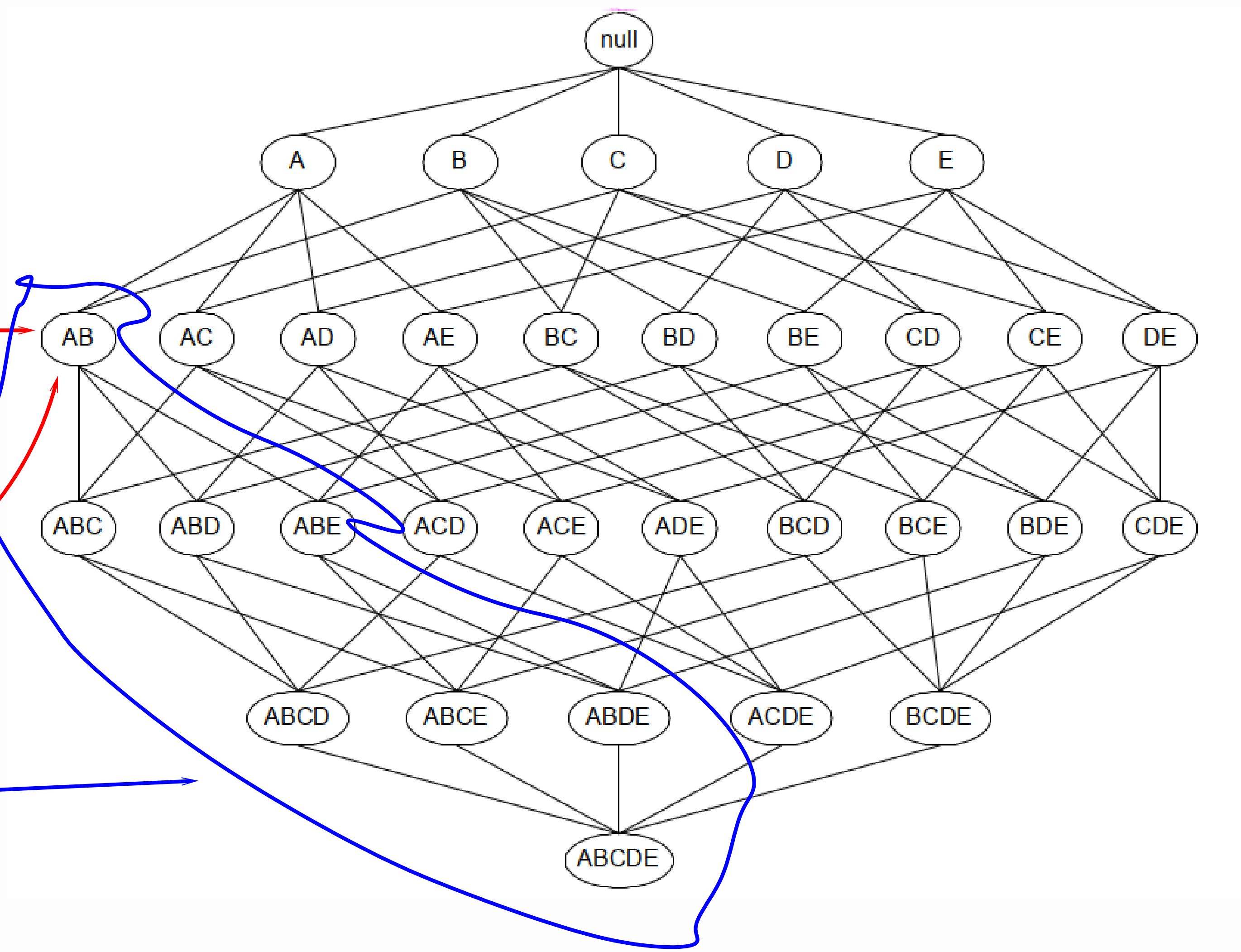
2 pruning rules:

1) find out here
that consequent C cannot yield
any good rules in AB or its
extensions ⇨

disable consequent C here

2) find our here
that no consequent can yield
good rules in AB or its
extensions ⇨

prune all sets here

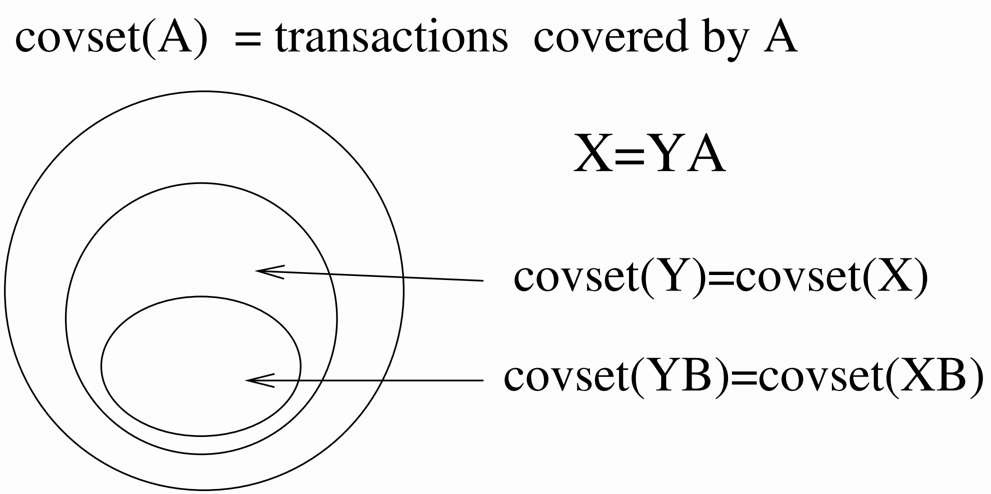


Pruning methods

- If M ibg, estimate upperbounds for $M(\mathbf{XQ} \rightarrow C{=}c)$ using information in $\mathbf{X}$ (possibly $\mathbf{X} = \emptyset$) ($P(C)$ known)
- e.g., upperbounds for δ:
  i) $\delta(\mathbf{X} \rightarrow C) \leq P(C)P(\neg C)$ for any $\mathbf{X}$
  ii) $\delta(\mathbf{XQ} \rightarrow C) \leq P(\mathbf{X})P(\neg C)$ for any $\mathbf{Q}$
  iii) $\delta(\mathbf{XQ} \rightarrow C) \leq P(\mathbf{X}C)P(\neg C)$ for any $\mathbf{Q}$
- If M is dbg, estimate lowerbounds
- Search for only top-Q rules $\Rightarrow$ update threshold $\max_M$ or $\min_M$ when new good rules found
- Estimate in advance which rules would be overfitted and can be ignored

Pruning methods

- Utilize **minimality condition**: If $\mathbf{X} = \mathbf{Y} \cup \{A\}$ and $P(A{=}a|\mathbf{Y}) = 1.0$ (i.e., $P(\mathbf{Y}A{=}a) = P(\mathbf{Y})$), then
  - i) A and $\neg A$ can be ignored in all $\mathbf{XQ}$
    rule $\mathbf{YQ} \rightarrow A{=}a$ redundant vs. $\mathbf{Y} \rightarrow A{=}a$
  - ii) all $B{=}b$, $B \notin \mathbf{X}$, can be ignored in all $\mathbf{XQ}$
    e.g., $\mathbf{X} \rightarrow B$ redundant vs. $\mathbf{Y} \rightarrow B$



In the figure, covset($\mathbf{Y}$) $\subseteq$ covset(A) since $P(A|\mathbf{Y}) = 1$

3. Algorithms for statistical association rules

- **Magnum Opus** [a]: search for sufficiently strong rules with γ or δ and test significance of improvement with a hypergeometric test
  - value-based: compares only $\mathbf{XQ} \rightarrow C$ vs. $\mathbf{X} \rightarrow C$
- **Kingfisher** [b]: search for the most significant rules with a significance measure (Fisher's p_F, χ^2, MI, etc.)
  - variable-based: compares both $\mathbf{XQ} \rightarrow C$ vs. $\mathbf{X} \rightarrow C$ and $\neg(\mathbf{XQ}) \rightarrow \neg C$ vs. $\neg\mathbf{X} \rightarrow \neg C$

---

[a] Webb 2005, Webb 2007
[b] Hämäläinen 2012

Algorithms for statistical classification rules

Now the consequence is fixed! Methods for finding best rules $\mathbf{X} \rightarrow C$ with

- χ^2 (Morishita and Sese, 2000 [a] and Nijssen and Kok, 2006)
- with any convex measure (like χ^2, MI), when $\mathbf{X}$ is closed (Nijssen et al., 2009)
- with measures of strength like δ and γ (Li, 2006)
- etc.

---

[a] also show that the problem of finding the best classification rule with χ^2 is NP-hard

Example: Kingfisher algorithm

Problem: Given a set of binary attributes $\mathbf{R} = \{A_1, \ldots, A_k\}$. Search for the most significant positive and negative dependency rules $\mathbf{X} \rightarrow A{=}a$, where $\mathbf{X} \subsetneq \mathbf{R}$, $A \in \mathbf{R} \setminus \mathbf{X}$, $a \in \{0, 1\}$!

- significance measures: p_F, $\ln(p_F)$, χ^2, MI, z-score
- in the following, let's use p_F
- rules are **non-redundant** ($\mathbf{X}$ contains no extra attributes which do not improve the dependency):
  $\nexists \mathbf{Y} \subsetneq \mathbf{X}$ such that $p_F(\mathbf{Y} \rightarrow A{=}a) \leq p_F(\mathbf{X} \rightarrow A{=}a)$
  - remember: smaller p-values better

Algorithm: the main idea

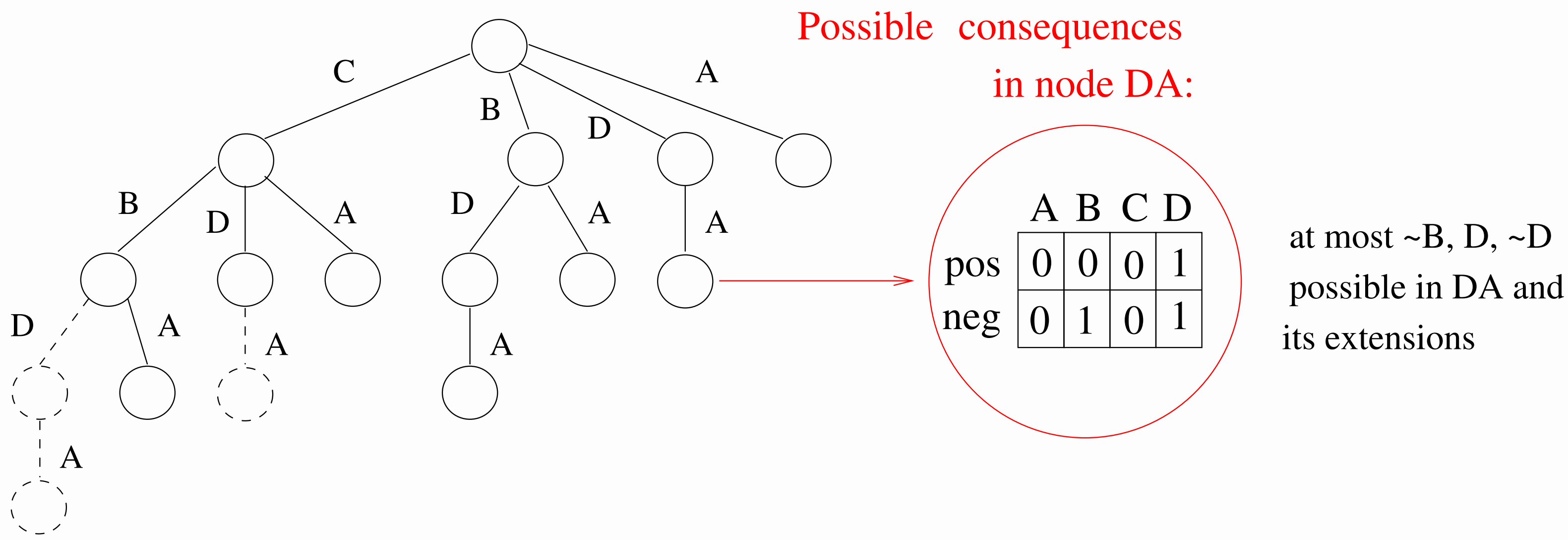
$\approx$ Boosted branch-and-bound

- Generate an enumeration tree (breadth-first) and keep record on possible consequences at each node
- Consequence $A_i{=}a_i$ is set **impossible** in set $\mathbf{X}$, if for all sets $\mathbf{Q}$, rule $\mathbf{X} \setminus \{A_i\}\mathbf{Q} \rightarrow A_i{=}a_i$ is insignificant or redundant
- Consequence can be possible **only if it was possible in all parents** (monotonicity!)
- A node is pruned when no possible consequences left

Algorithm: the main idea

In each node

- bitvectors for possible positive and negative consequences
- numeric vector for best p_F-values of simpler rules
  - for checking redundancy



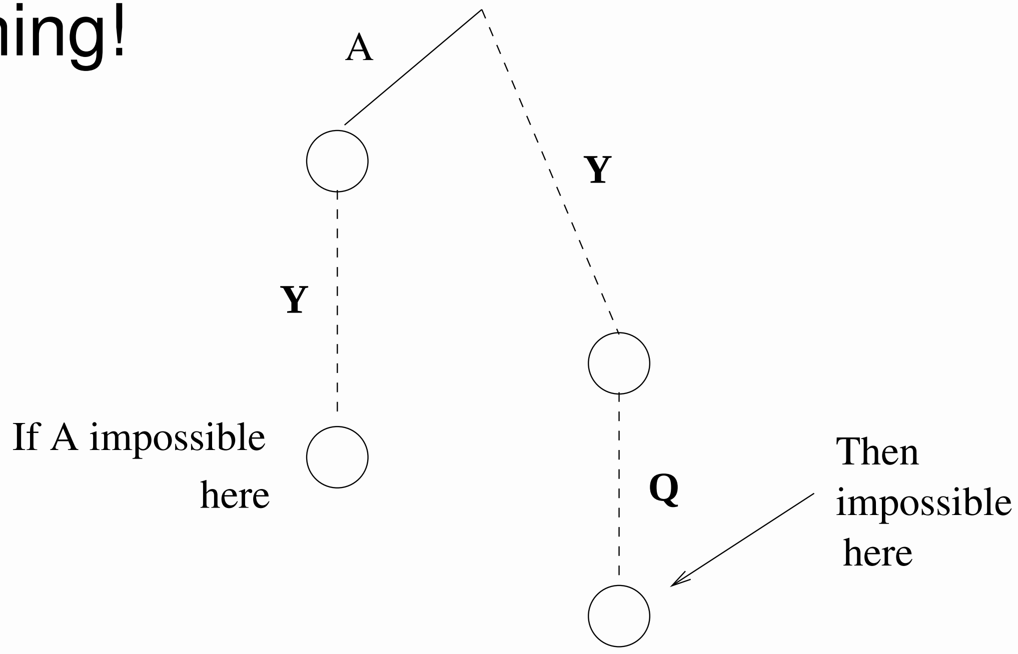
How to determine possible consequences?

At each node (set) $\mathbf{X}$

1. Initialization: combine parents' "possible" vectors (bit-and)
2. Updating: estimate lower bounds (LB) for $p_F(\mathbf{X} \setminus \{A_i\}\mathbf{Q} \rightarrow A_i{=}a_i)$ and decide if $A_i{=}a_i$ impossible in node $\mathbf{X}$
   - both for $A_i \in \mathbf{X}$ and $A_i \notin \mathbf{X}$
   - uses 3 different lower bounds $LB1$, $LB2$, $LB3$
     $LB1$: only $fr(A_i{=}a_i)$ known; $LB2$: $fr(A_i{=}a_i)$ and $fr(\mathbf{X})$ known, $A_i \notin \mathbf{X}$; $LB3$: $fr(A_i{=}a_i)$ and $fr(\mathbf{X})$ known, $A_i \in \mathbf{X}$
3. Utilize minimality condition

Algorithm: possible consequences

4. **Lapis philosophorum principle** (LP):
If $A{=}a$ impossible in $\mathbf{X}$, then set it impossible in parent $\mathbf{Y}$, $\mathbf{X} = \mathbf{Y} \cup \{A\}$, and all its descendants $\mathbf{YQ}$
   - Note: $\mathbf{Y}$ was already processed at previous level, but now updated again
   - Very efficient pruning!



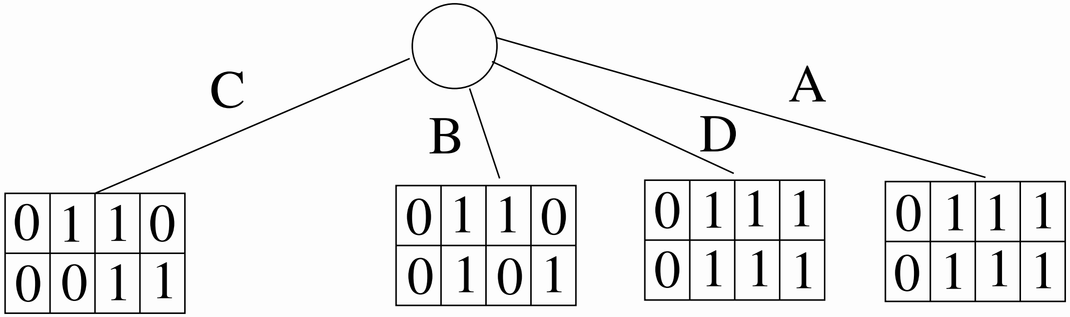
Extra: Simulation

Data
$\mathbf{R} = \{A, B, C, D\}$
$n = 100$

set	freq.
$ABC\neg D$	10
$A\neg B\neg CD$	85
$\neg AB\neg CD$	5

Search for the 10 best rules from the example data, when initial $p_{max} = 1.2 \cdot 10^{-8}$.

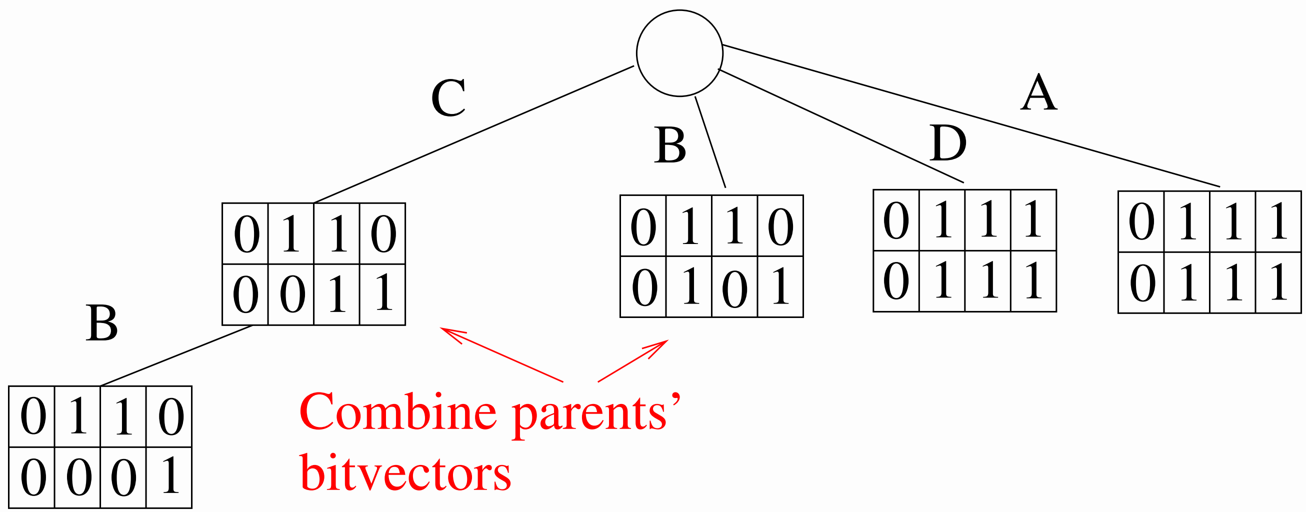
Simulation level 1: use LBs



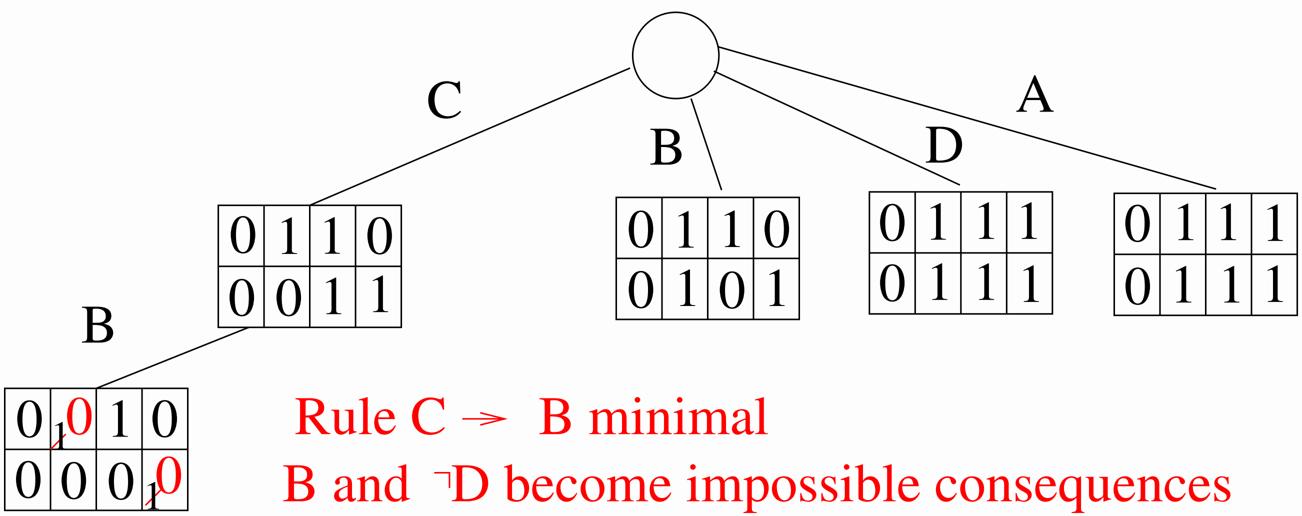
LB1: A and ¬A are impossible consequences

Otherwise, possible consequences are determined by LB2

Simulation level 2: intialize CB



Simulation level 2: evaluate CB



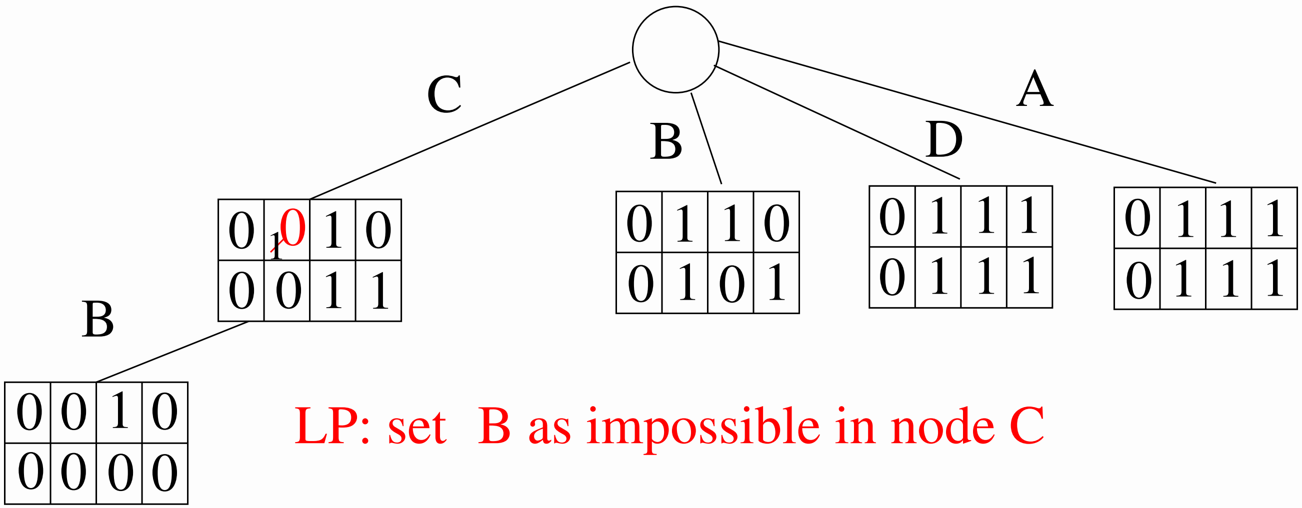
Rule C → B minimal

B and ¬D become impossible consequences

(more specific rules would be redundant)

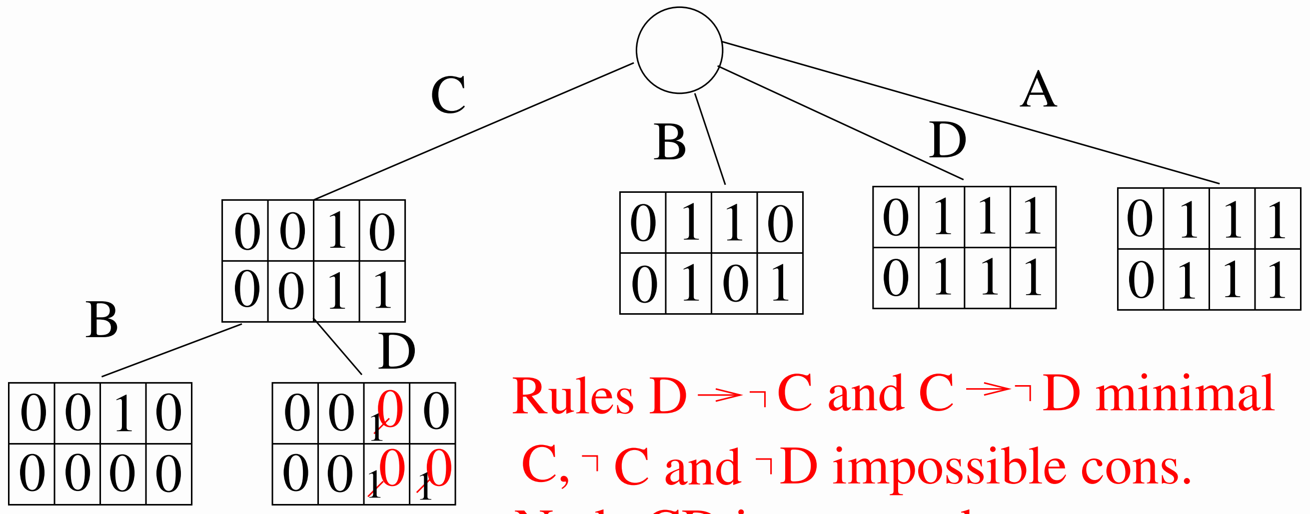
Rule	p
$C \rightarrow B$	$1.7 \cdot 10^{-10}$

Simulation level 2: utilize LP



Rule	p
$C \rightarrow B$	$1.7 \cdot 10^{-10}$

Simulation level 2: create and evaluate CD



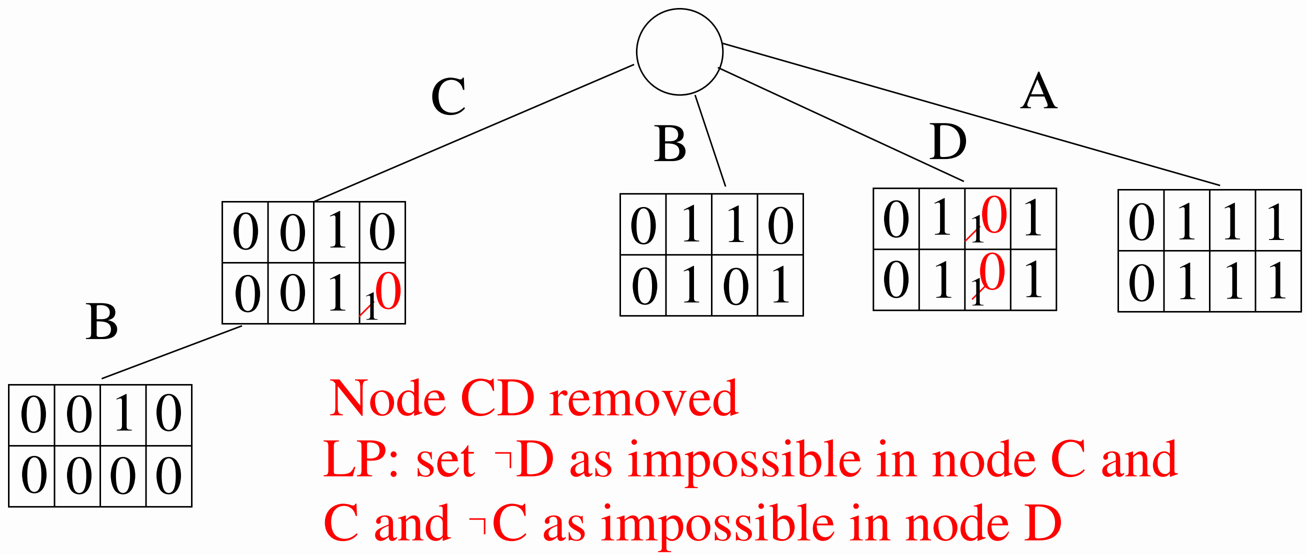
Rules D $\not\to \neg$C and C $\not\to \neg$D minimal

C, $\neg$ C and $\neg$D impossible cons.

Node CD is removed

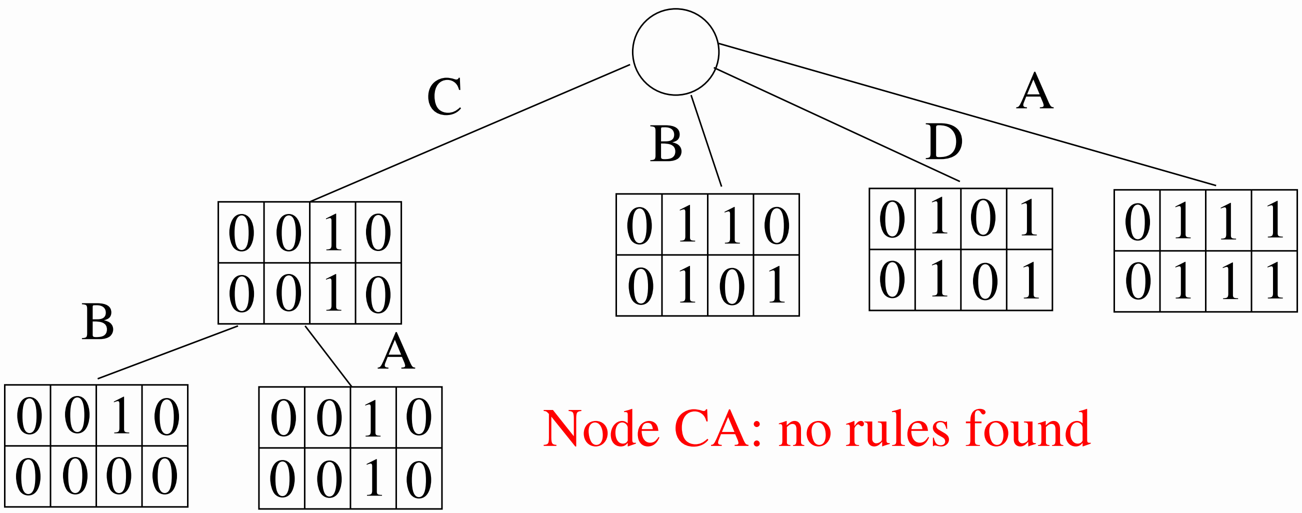
Rule	p
$D \to \neg C$	$5.8 \cdot 10^{-14}$
$C \to B$	$1.7 \cdot 10^{-10}$

Simulation level 2: utilize LP



Rule	p
$D \rightarrow \neg C$	$5.8 \cdot 10^{-14}$
$C \rightarrow B$	$1.7 \cdot 10^{-10}$

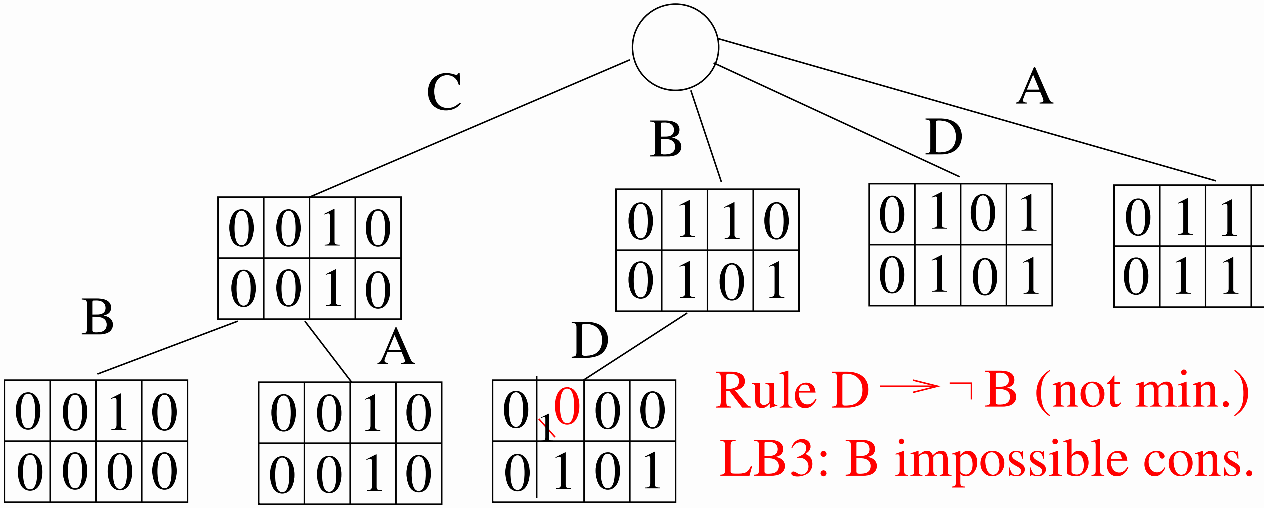
Simulation level 2: evaluate CA



Node CA: no rules found

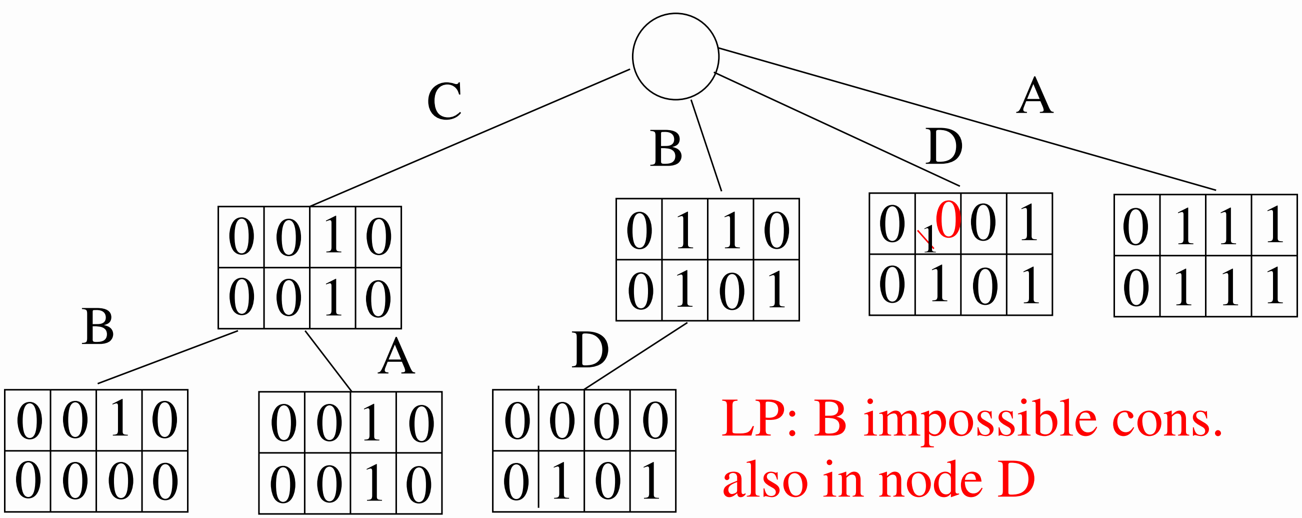
Rule	p
$D \to \neg C$	$5.8 \cdot 10^{-14}$
$C \to B$	$1.7 \cdot 10^{-10}$

Simulation level 2: evaluate BD



Rule	p
$D \rightarrow \neg C$	$5.8 \cdot 10^{-14}$
$C \rightarrow B$	$1.7 \cdot 10^{-10}$
$D \rightarrow \neg B$	$1.7 \cdot 10^{-10}$

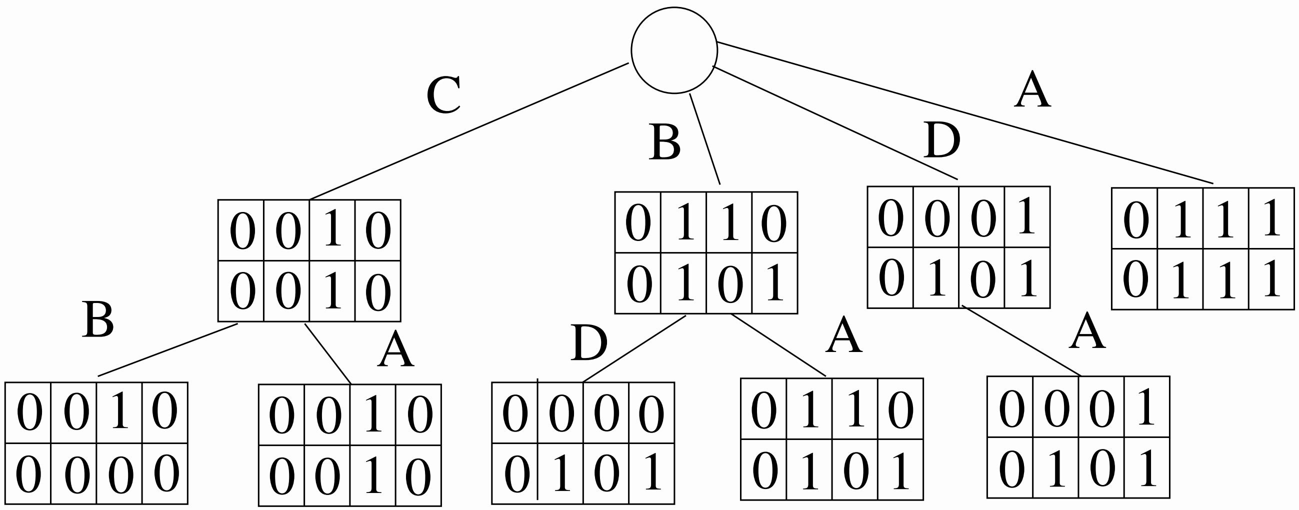
Simulation level 2: utilize LP



LP: B impossible cons.
also in node D

Rule	p
$D \rightarrow \neg C$	$5.8 \cdot 10^{-14}$
$C \rightarrow B$	$1.7 \cdot 10^{-10}$
$D \rightarrow \neg B$	$1.7 \cdot 10^{-10}$

Simulation level 2: evaluate BA *and* DA

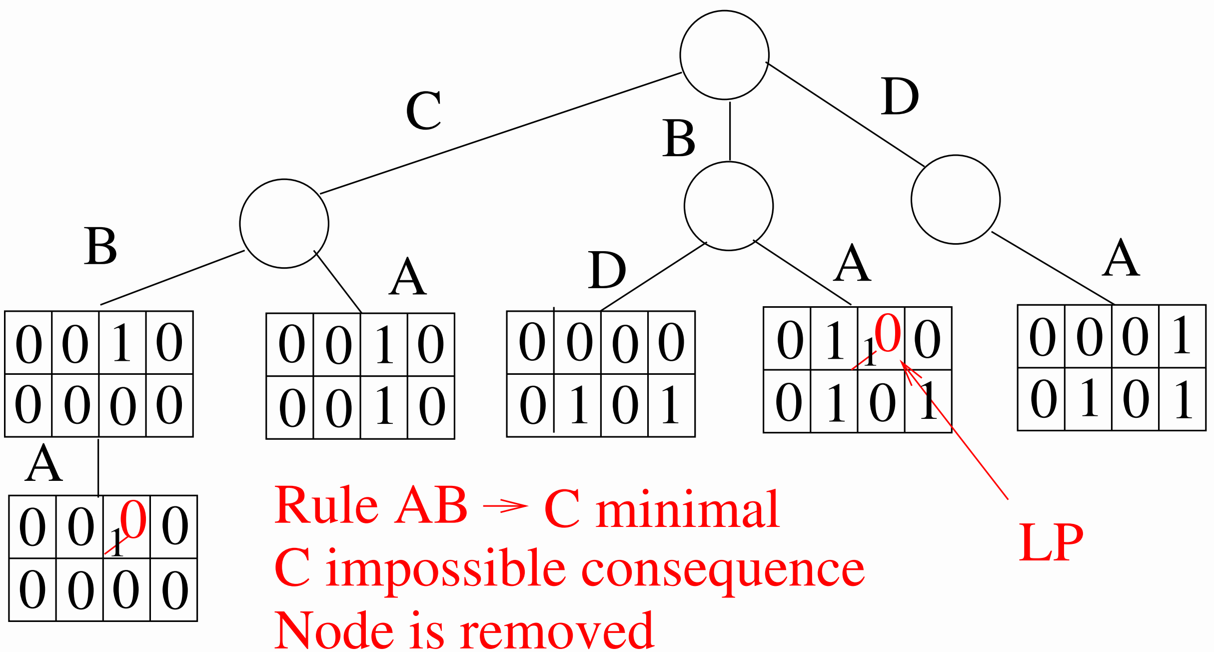


Rule	p
$D \rightarrow \neg C$	$5.8 \cdot 10^{-14}$
$C \rightarrow B$	$1.7 \cdot 10^{-10}$
$D \rightarrow \neg B$	$1.7 \cdot 10^{-10}$

BA and DA: no rules found

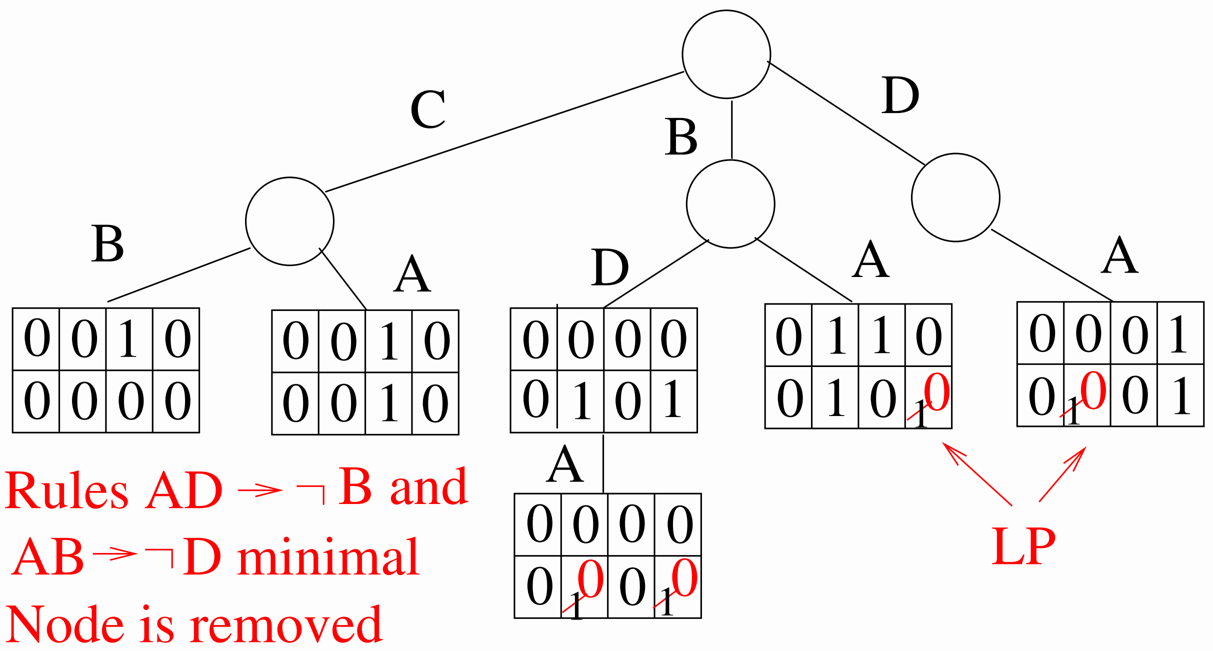
Note: Set A (1st level) no more needed.

Simulation level 3: evaluate CBA + use LP



Rule	p
$D \rightarrow \neg C$	$5.8 \cdot 10^{-14}$
$AB \rightarrow C$	$5.8 \cdot 10^{-14}$
$C \rightarrow B$	$1.7 \cdot 10^{-10}$
$D \rightarrow \neg B$	$1.7 \cdot 10^{-10}$

Simulation level 3: evaluate BDA + use LP



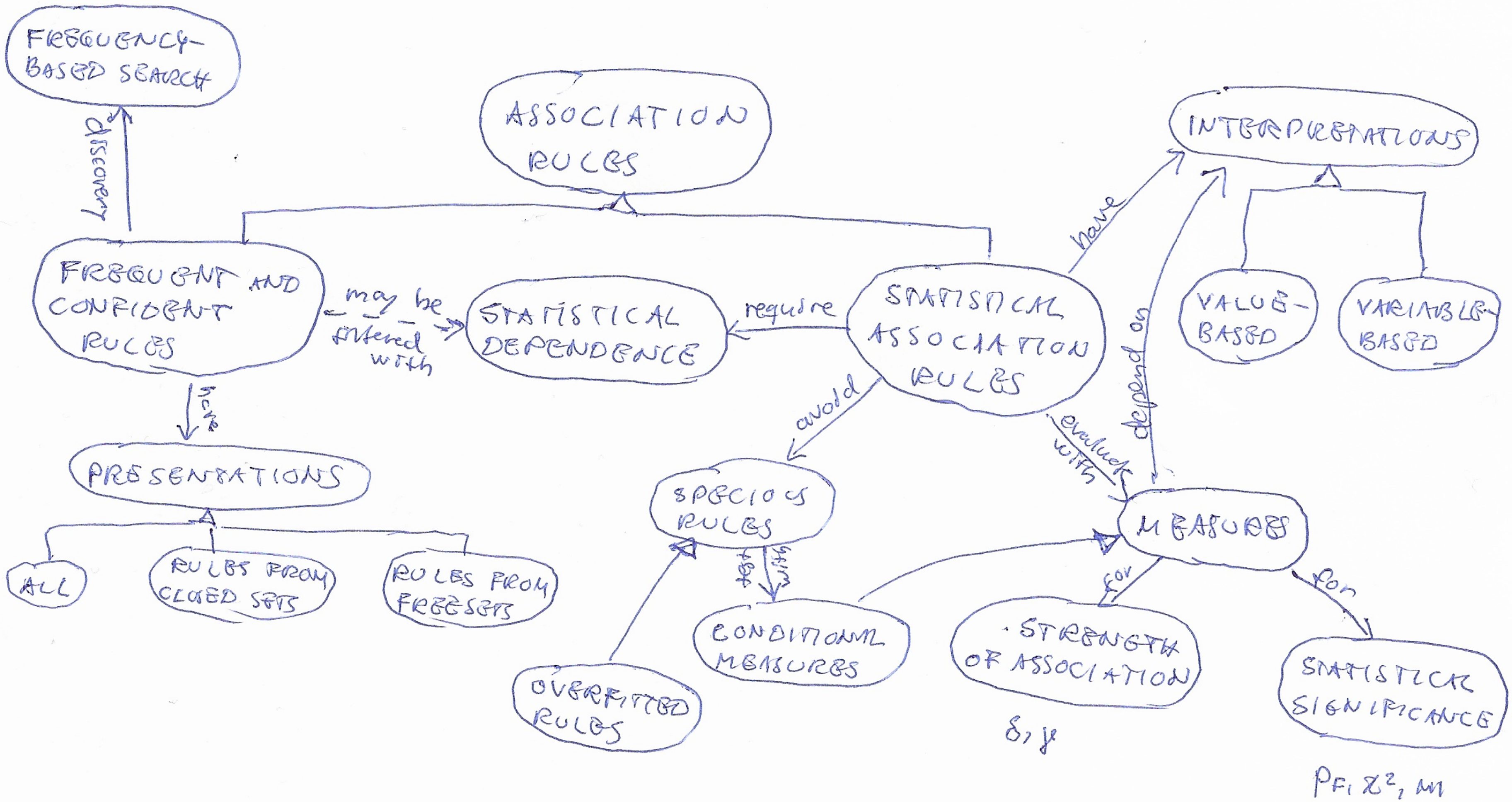
Rule	p
$AD \rightarrow \neg B$	$3.9 \cdot 10^{-18}$
$D \rightarrow \neg C$	$5.8 \cdot 10^{-14}$
$AB \rightarrow C$	$5.8 \cdot 10^{-14}$
$AB \rightarrow \neg D$	$5.8 \cdot 10^{-14}$
$C \rightarrow B$	$1.7 \cdot 10^{-10}$
$D \rightarrow \neg B$	$1.7 \cdot 10^{-10}$

DONE!

Summary

- If you want statistical associations, search them directly! (postprocessing frequent sets causes false positives & false negatives)
- statistical dependence **not monotonic** property (may be $AB \rightarrow C$ even if $A \perp\!\!\!\perp B$, $A \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C$)
- **Secret**: $UB(M)$ or $LB(M)$ may behave monotonically!
  - $\forall \mathbf{Q}$: if $UB(M(\mathbf{XQ} \rightarrow C)|fr(c)) < \min_M$, $UB(M(\mathbf{XQ} \rightarrow C)|fr(c), fr(\mathbf{X})) < \min_M$, or $UB(M(\mathbf{XQ} \rightarrow C)|fr(c), fr(\mathbf{X}), fr(\mathbf{X}C)) < \min_M$, $\mathbf{XQ} \rightarrow C$ can be pruned! (here M ibg)
- Remember overfitted and other misleading rules!

Concept map of Association rules



Reading

- Hämäläinen and Webb: A tutorial on statistically sound pattern discovery. Data Mining and Knowledge Discovery 33(2):325-377, 2019. **Sections 3.1, 4.1, 4.4.**
- Properties of statistical association rules explained with the Mega Party example (in MyCourses)
- (**Optional**) Hämäläinen: Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures, Knowledge and Information Systems 32: 383-414, 2012. **Sections 1-3, 4.1-4.2, the main idea from 5.1.**

Other references

- Li: On optimal rule discovery. IEEE Transactions on Knowledge and Data Engineering, 18(4):460-471, 2006.
- Lindley and Novick: The Role of Exchangeability in Inference, Annals of Statistics 9(1):45-58, 1981.
- Morishita and Sese: Traversing itemset lattices with statistical metric pruning. ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2000.
- Nijssen and Kok: Multi-class correlated pattern mining. 4th International Workshop on Knowledge Discovery in Inductive Databases, 2006.

Other references

- Nijssen, Guns, and De Raed. Correlated itemset mining in ROC space: a constraint programming approach. 15th ACM SIGKDD conf. Knowledge Discovery and Data Mining, 2009.
- Webb: Discovering significant patterns. Machine Learning, 68(1):1-33, 2007.
- Webb: Magnum Opus. Software, G. I. Webb & Associates, Melbourne, Australia.
- Webb and Zhang. K-optimal rule discovery. Data Mining and Knowledge Discovery, 10(1), 2005.