MONASH University
Information Technology

Scoperta K-ottimale del modello: Un metodo efficiente ed efficace ad estrazione mineraria esplorativa di dati

Geoff Webb

Monash University

http://www.csse.monash.edu.au/~webb

MONASH University
Information Technology

K-optimal pattern discovery: An Efficient and Effective Approach to Exploratory Data Mining

Geoff Webb

Monash University

http://www.csse.monash.edu.au/~webb

Outline

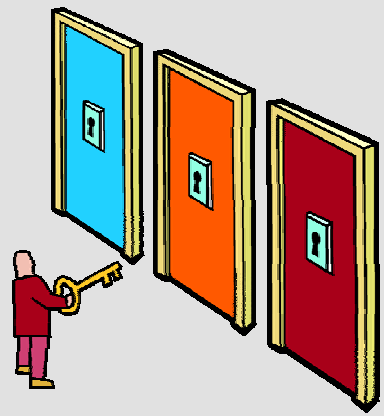
- Association rules are undervalued
- Minimum support is not always an appropriate constraint
- K-optimal techniques provide an efficient and effective alternative

Evils of model selection

- Many data mining techniques seek to identify a single model that best fits the observed data.
- In many applications many models will (almost) equally fit the data
- Data mining systems often make arbitrary choices
- A system may have no basis on which to select models, but an expert often will
  - ease / cost of operatalisation
  - comprehensibility / compatibility with existing knowledge and beliefs
  - social / legal / ethical / political acceptability

Exploratory pattern discovery

- Exploratory pattern discovery seeks all patterns that satisfy user-defined constraints
  - Pattern = rule, itemset, …
- The user can select from these patterns
  - can use criteria that might be infeasible to quantify



Association rule discovery

- Utilizes *minimum support constraint*
- Finds all rules that satisfy minimum support together with other user specified constraints such as *minimum confidence*

Limitations of minimum support

- Discontinuity in interestingness function
- The vodka and caviar problem
  - some high value associations are infrequent
- Minimum support may not be relevant
  - cannot be sufficiently low to capture all valid rules
  - cannot be sufficiently high to exclude all spurious rules
- Feast or famine
  - minimum support is a crude control mechanism
- Cannot handle dense data
- Cannot prune search space using constraints on relationship between antecedent and consequent
  - eg confidence

Very low support rules can be significant

Data file: Brijs retail.itl [50% sample]

44081 cases / 44081 holdout cases / 16470 items

The following 5 rules passed holdout evaluation

168 & 4685 → 1 [Coverage=0.000 (3); Support=0.000 (3); Strength estimate=0.601; Lift estimate=192.06]

168 & 3021 → 1 [Coverage=0.000 (3); Support=0.000 (3); Strength estimate=0.601; Lift estimate=192.06]

1476 & 4685 → 1 [Coverage=0.000 (2); Support=0.000 (2); Strength estimate=0.502; Lift estimate=160.21]

168 & 783 → 1 [Coverage=0.000 (4); Support=0.000 (3); Strength estimate=0.501; Lift estimate=160.05]

3021 & 4685 → 1 [Coverage=0.000 (4); Support=0.000 (3); Strength estimate=0.501; Lift estimate=160.05]

Very high support rules can be spurious

Data file: mush.data [50% sample]

4062 cases / 4062 holdout cases / 127 values

100 productive rules with highest support

The following 8 rules failed holdout evaluation

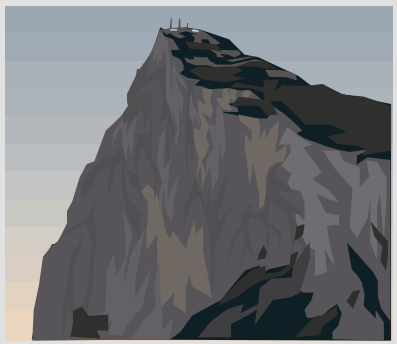
gill-attachment=f -> ring-number=o
[Coverage=0.976 (3965); Support=0.900 (3656); Strength=0.922; Lift=1.00]
Holdout coverage = 3949, holdout support = 3640
Fails positive correlation, p = 0.394850

ring-number=o -> bruises?=f
[Coverage=0.922 (3745); Support=0.541 (2197); Strength=0.587; Lift=1.00]
Holdout coverage = 3743, holdout support = 2211
Fails positive correlation, p = 0.004757

…..

Roles of constraints

1. Select most relevant patterns
   - patterns that are likely to be interesting
2. Control the number of patterns that the user must consider
3. Make computation feasible



Minimum support can get overloaded

K-optimal pattern discovery

- Find *k* patterns that optimise a measure of interest within other constraints that the user may specify
  - user empowered to use relevant measure of interest
  - user can specify the number of patterns to be returned
- Efficiency derived from use of measure of interest to prune the search space.

Previous *k*-optimal techniques

- ***k*-optimal classification rule discovery (Webb, 1995)**
- ***k*-optimal subgroup discovery (Wrobel, 1997)**
- **finding *k* most interesting patterns using sequential sampling (Scheffer & Wrobel, 2002)**
- **OPUS-AR (Webb, 2002)**
- **mining top.*k* frequent closed patterns without minimum support (Han, Wang, Lu, Tzvetkov, 2002)**

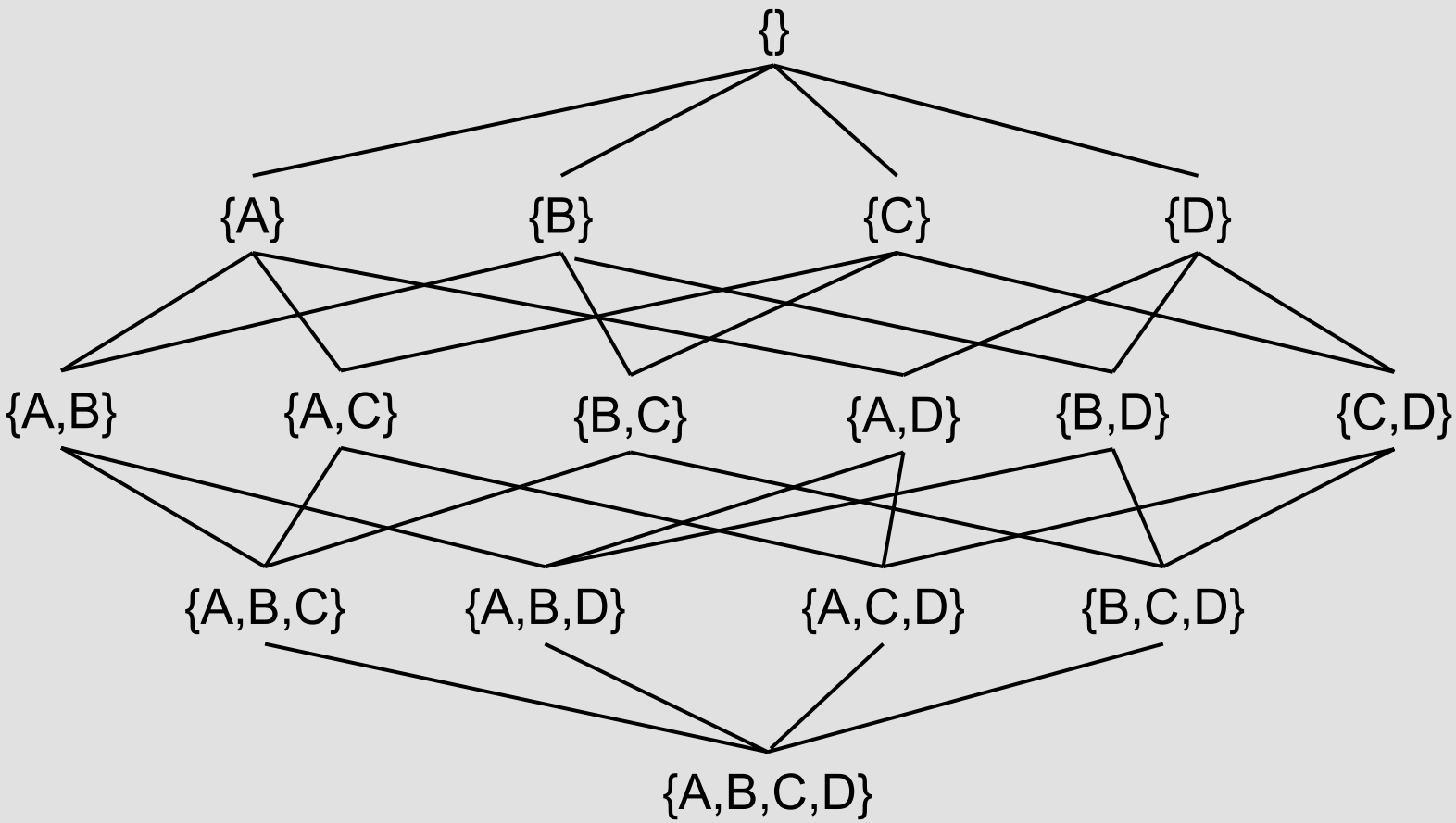
Quantifying interest

- Many different measures of interest
- Most relate to degree of interdependence between antecedent and consequent
- lift ($A \rightarrow C$) = strength($A \rightarrow C$) / F(C)
  - proportional increase in strength in context of antecedent
- leverage($A \rightarrow C$) = support($A \rightarrow C$) – F(A) $\times$ F(C)
  - difference between observed and expected frequency
  - also known as *interest*

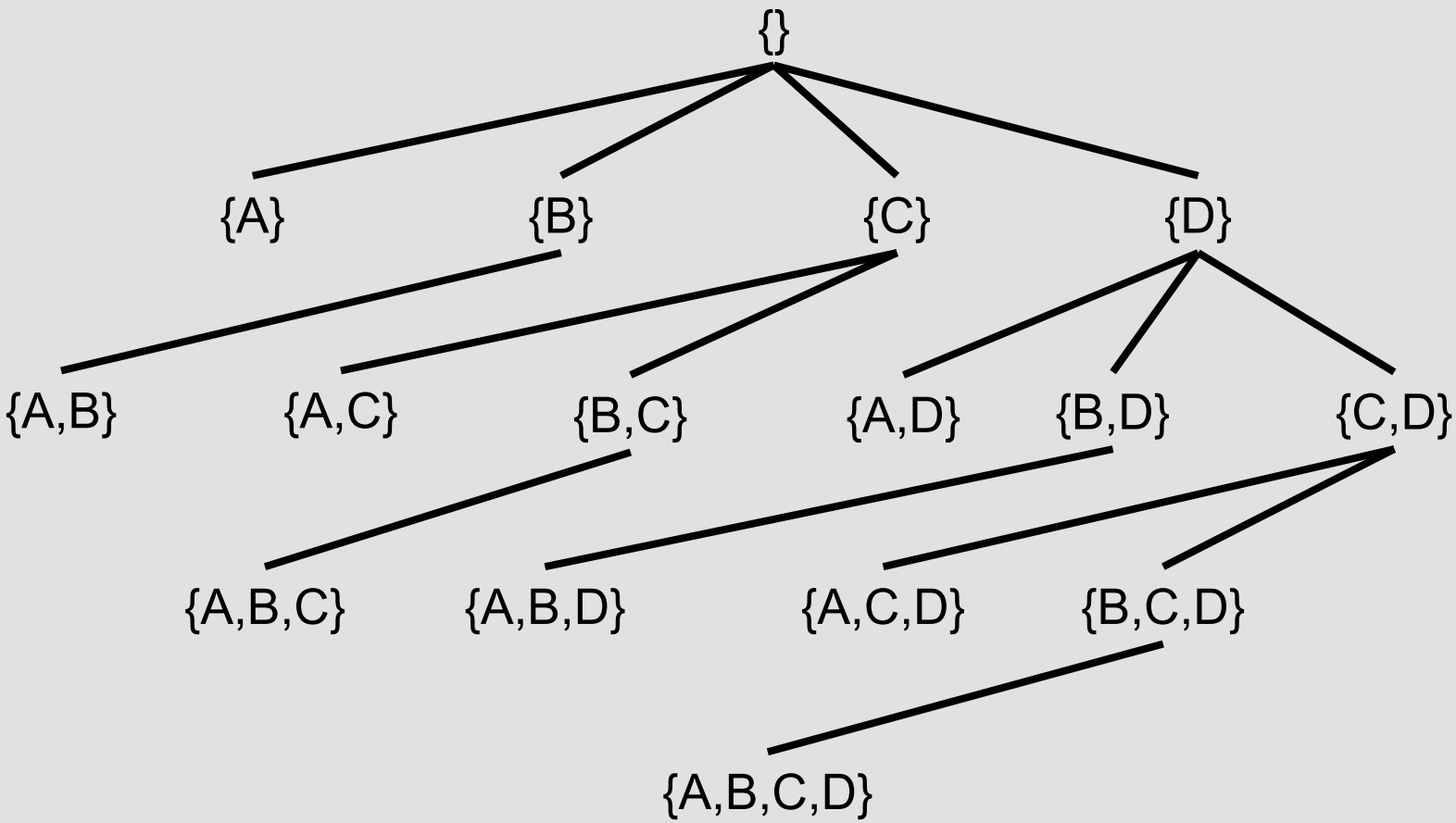
Techniques

- Restrict each consequent to any single condition
- Perform OPUS branch and bound search over antecedents
- Propagate set of conditions available for consequent through the search space
- Can benefit from constraints
  - on relationship between antecedent and consequent
  - that are monotone, anti-monotone or neither.
    - eg confidence

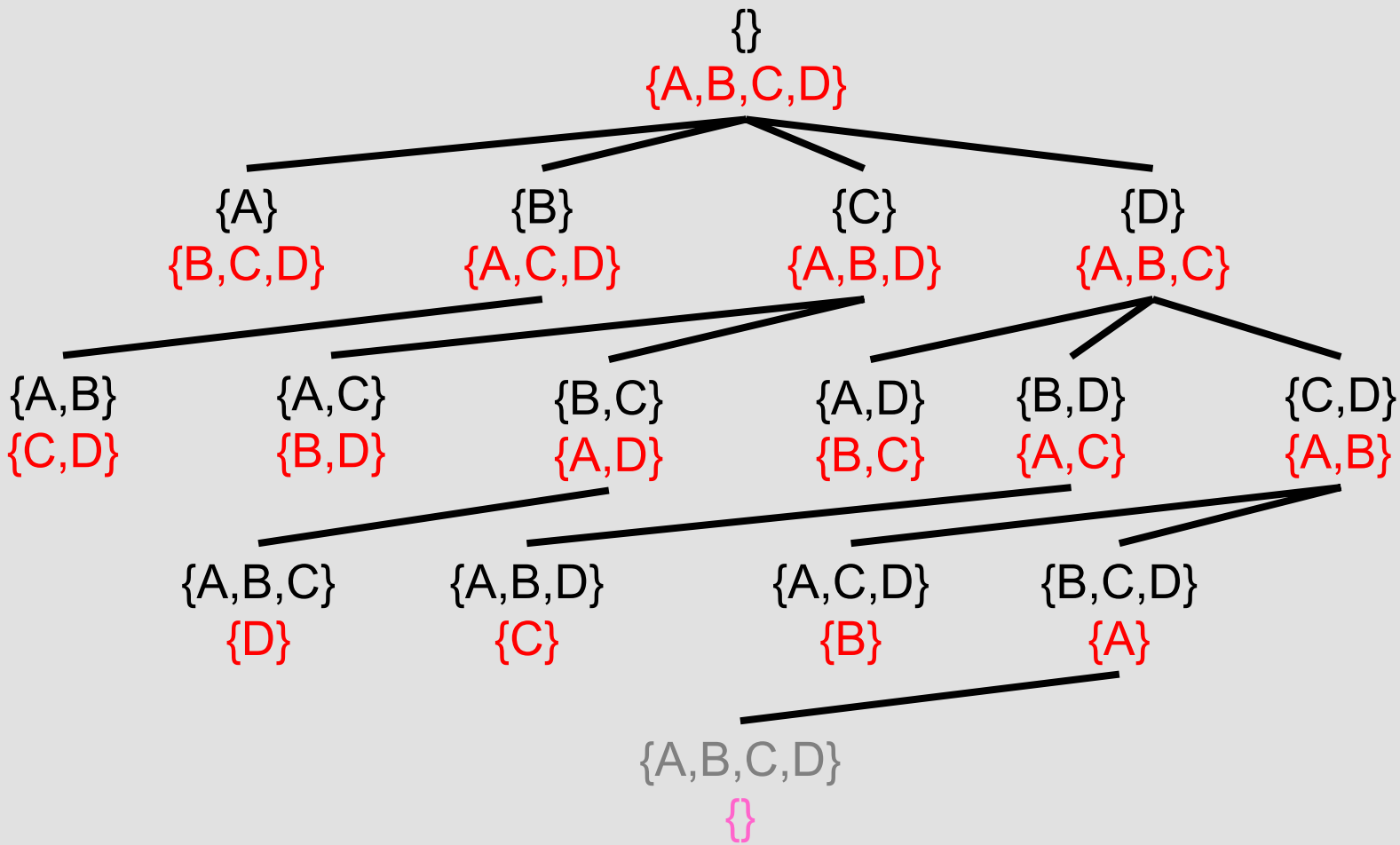
Generalization lattice for antecedents



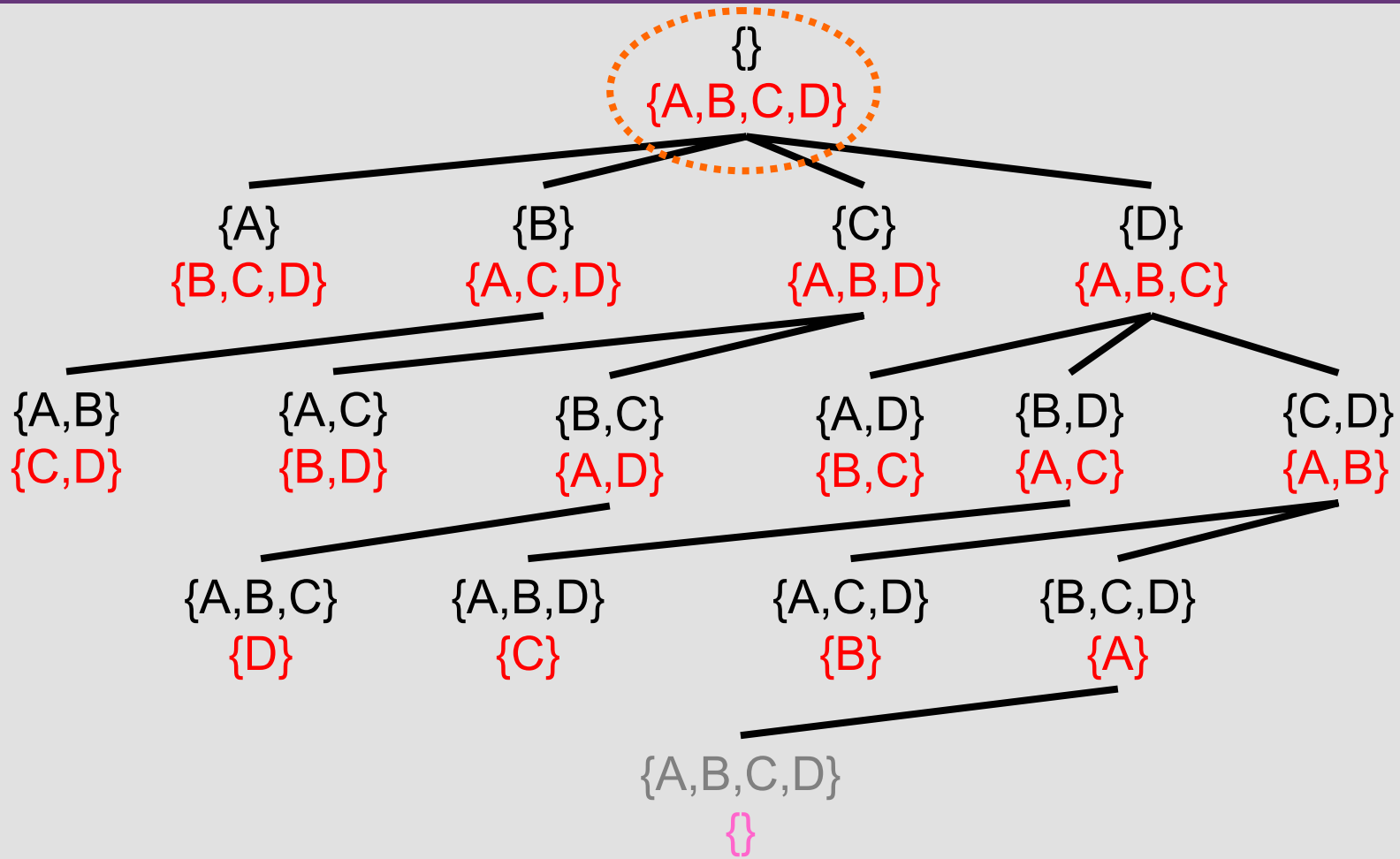
Search tree for antecedents



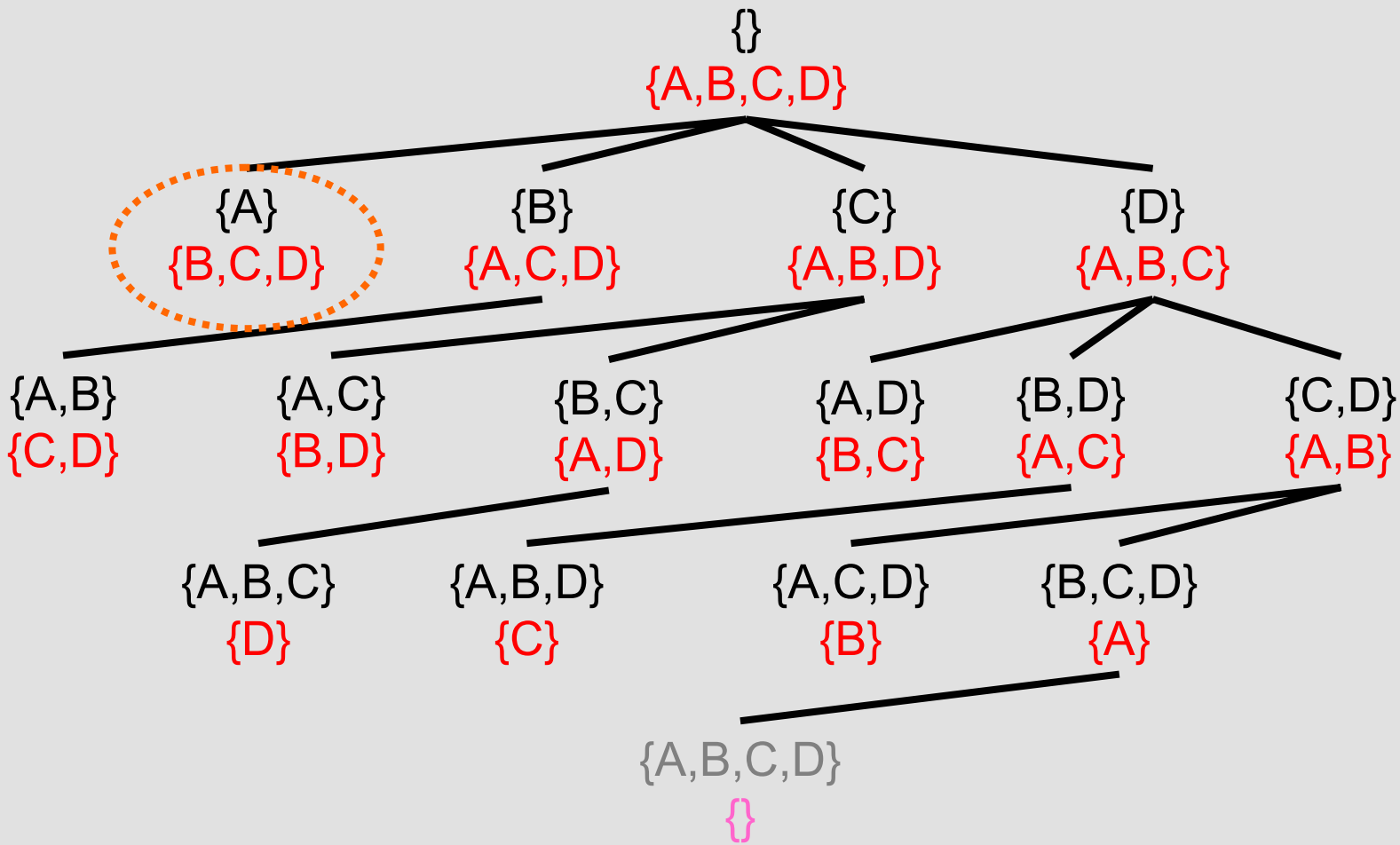
Search tree with consequent propagation



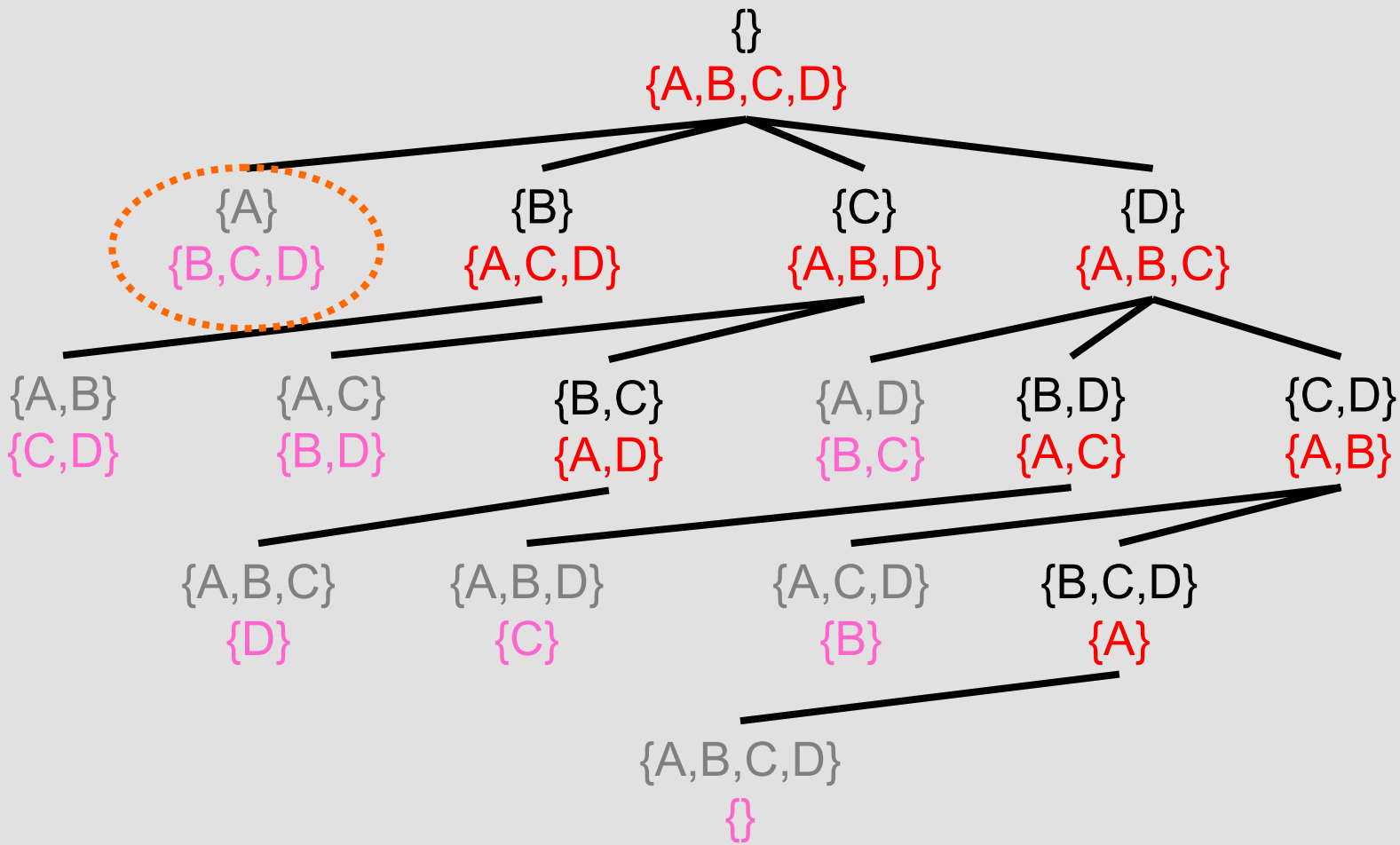
Step through tree maintaining *k*-optimal so far



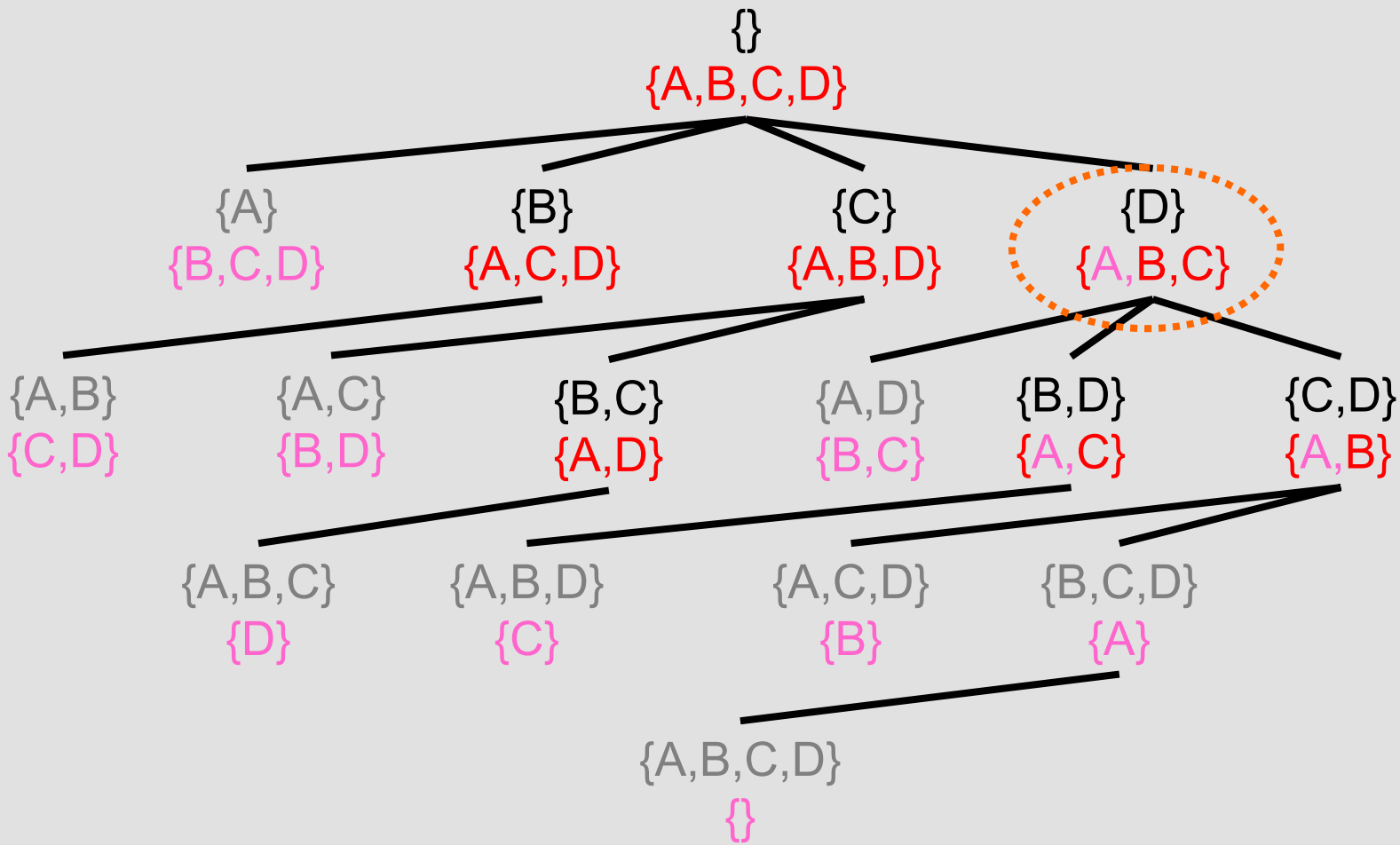
Step through tree maintaining *k*-optimal so far



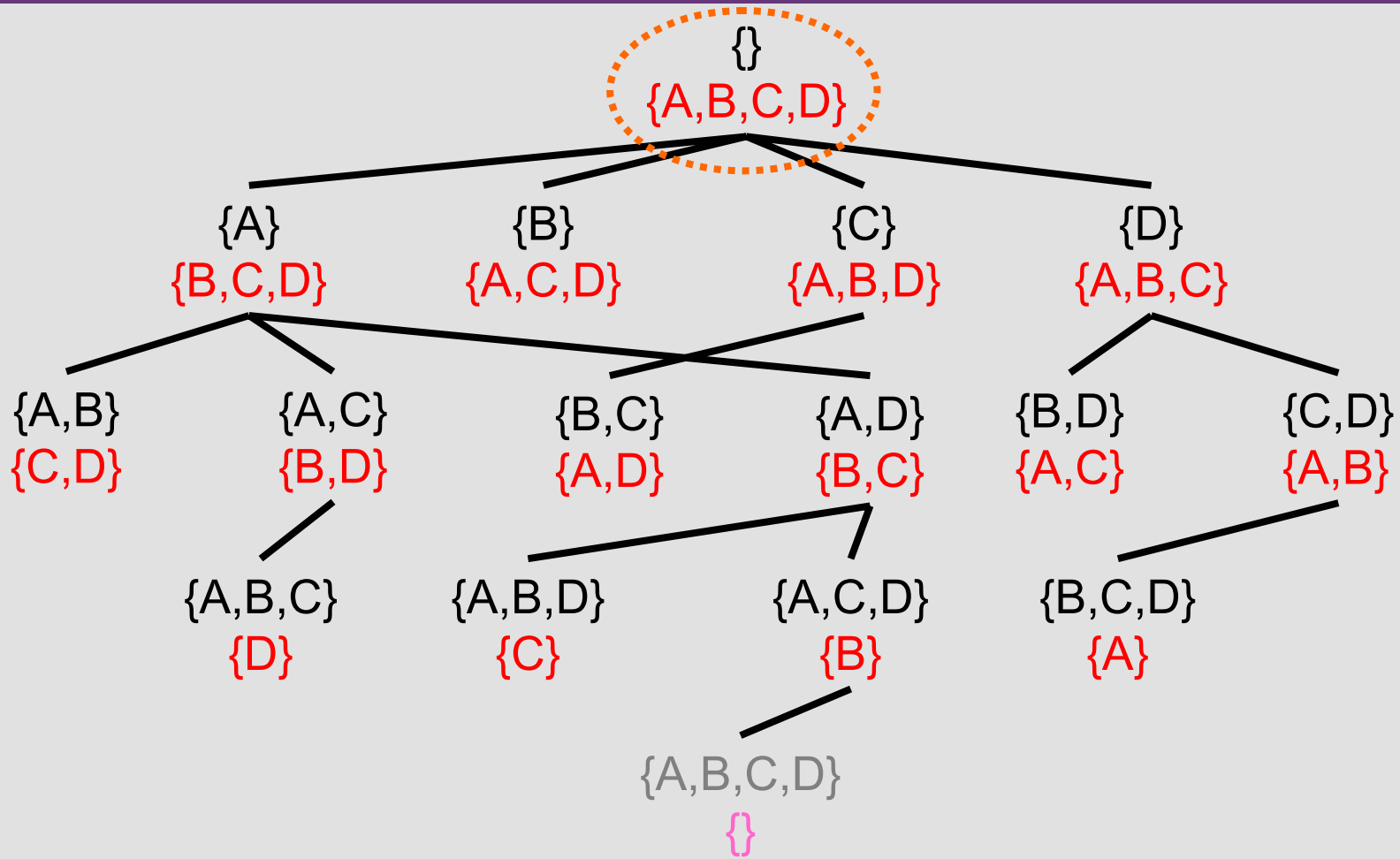
Antecedent pruning is propagated both downward and through siblings



Consequent pruning is propagated downward



Search space reordering



Efficiency

- The *k*-optimal constraint is often sufficient to enable efficient search
- Where minimum support is not a primary metric, OPUS-AR is often more efficient than frequent itemset approaches.

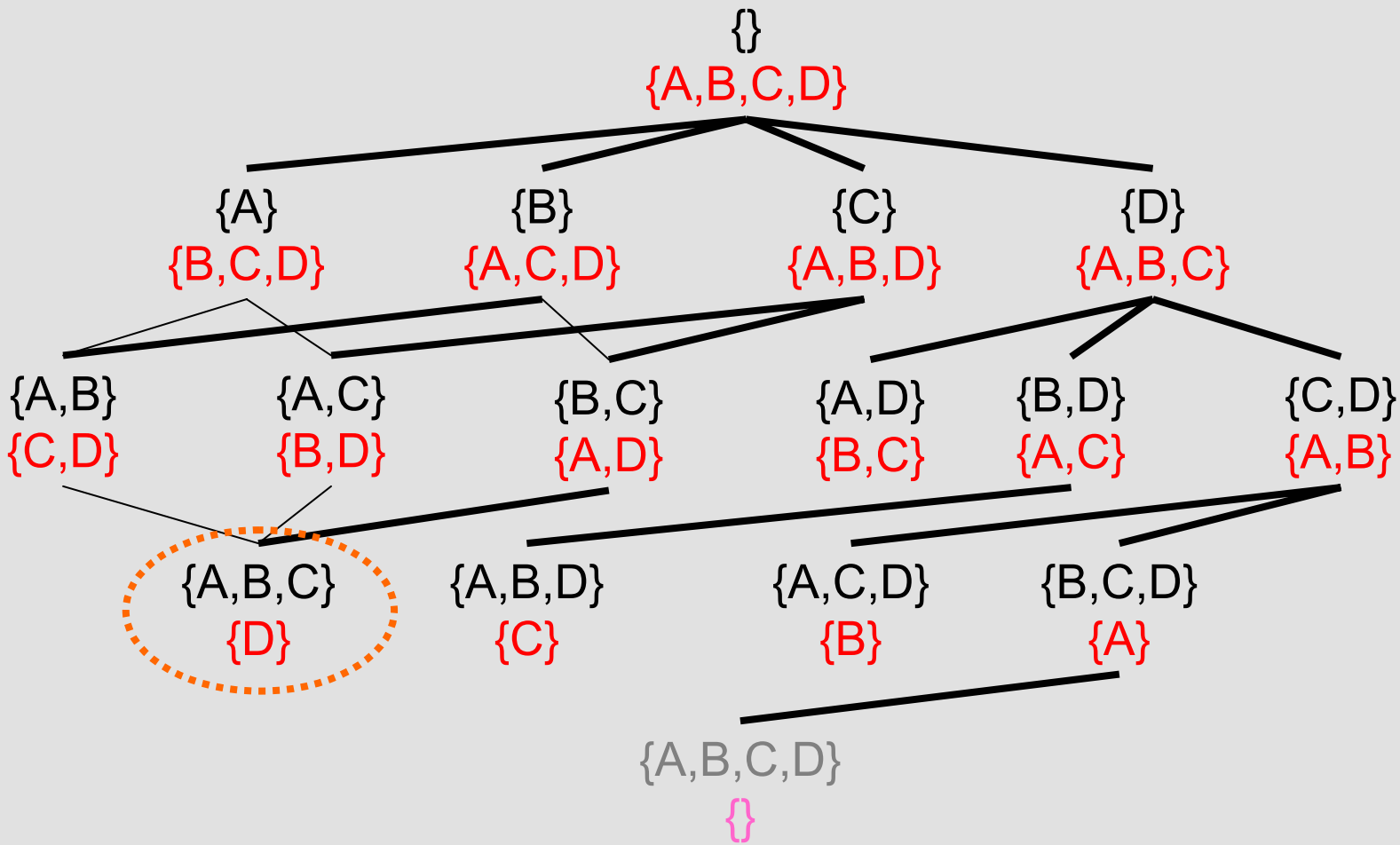
False discoveries

- Massive search leads to high risk of false discoveries
  - eg 100 observations, two independent events each occurring with 50% probability,
    - the probability of perfect correlation is 7.8E-31.
  - if there are 1000 events then there are 2^{1000} = 1.07E+301 antecedent – consequent pairs.
- What constitutes a false discovery depends upon the analytic objective
- Usually should include rules where
  - antecedent and consequent are independent
  - antecedent and consequent are independent given a generalisation of the antecedent

Spurious rules

- If condition *X* is unrelated to conditions *A* and *B*,
  - $strength(A \ \& \ X \rightarrow B) \approx strength(A \rightarrow B)$
  - $lift(A \ \& \ X \rightarrow B) \approx lift(A \rightarrow B)$
  - Eg *pregnant* & *Californian* $\rightarrow B$
- Special case: redundant rules
  - condition *X* is entailed by condition A
  - all standard metrics of interest, inc. strength, lift and leverage, identical for specialisation & generalisation
  - Eg *pregnant* & *female* $\rightarrow B$
  - redundant rules subset of improvement ≤ 0 rules
- One core rule can result in many *spurious rules*
- If problem ignored, majority of rules can be spurious!

Need to test up the generalization lattice



Testing independence

- Cannot perform simple test of independence because of multiple comparisons problem
  - used previously (eg Webb, Butler & Newlands, 2003) as a statistically unsound filter
- Cannot perform simple adjustment such as Bonferroni or Benjamini-Hochberg because rule spaces are so large, eg 2^{1000} (> 1.0E+301)
  - would result in unacceptable type-2 error
    - eg α = 5.0E-303
  - previous approaches (eg Bay & Pazzani) have adjusted only for the number of rules 'considered'
    - not adequate
- Can only use randomization techniques for simple tests

Discovery as hypothesis generation

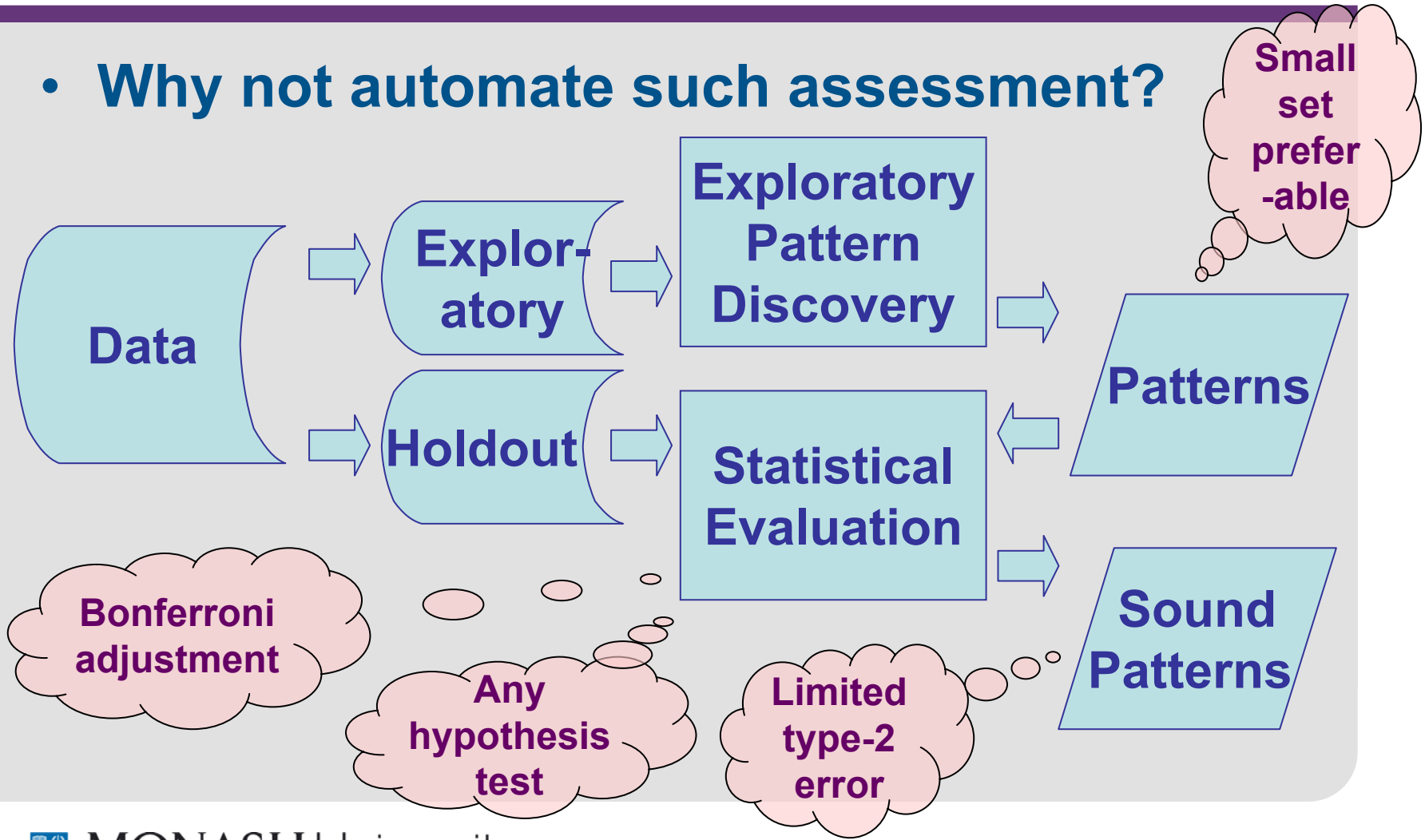
- Important to trade-off the risks of both type-1 and type-2 errors
- Perhaps best viewed as hypothesis generation, recognising that 'discovered' patterns require independent assessment

Hypothesis testing: proposal

- Why not automate such assessment?
- Partition data into exploratory and holdout sets
- Perform exploratory pattern discovery on exploratory set
- Select small set of patterns of potential interest
- Apply hypothesis tests on holdout data using correction such as Bonferroni or Benjamini-Hochberg for the number of patterns so tested
- Can perform any hypothesis test
- Risk of type-2 error constrained by small adjustment

Hypothesis testing: proposal

- Why not automate such assessment?



Detecting spurious rules

- Assuming interest only in positive associations
  - P(C | A) > P(C)
- For any rule $A \rightarrow C$, want to assess whether it has higher strength than all its generalisations
  - Eg, is strength(*pregnant* & *female* $\rightarrow$ *B*) >
    - strength(*pregnant* $\rightarrow$ *B*)
    - strength(*female* $\rightarrow$ *B*)
    - strength(*true* $\rightarrow$ *B*)

Detecting spurious rules (cont)

- Could use log linear analysis
  - but, may have low expected frequencies
  - want a one-tailed analysis
  - do not need to identify most parsimonious model, only whether a more parsimonious model exists
- Perform one-tailed Fisher exact tests with respect to each generalisation
  - Reject if *any* test does not exceed critical value
    - no need to adjust for multiple comparisons with respect to the multiple tests for a single rule
- Use Bonferroni adjustment for strict control of type-1 error

Case study: Ten widely used data sets

Name	Description	Records	Preds
BMS webview	products viewed at a commercial website	59,601	497
covtype	forest cover data	581,012	125
ipums.la.99	Los Angeles census data	88,443	1,874
kddcup98	charity donors	52,256	19,662
letter-recog'n	digital image recognition	20,000	74
mush	identification of poisonous mushrooms	8,124	127
retail	retail market basket data	88,162	16,470
shuttle	records of space shuttle flight data	58,000	34
splice-junction	DNA sequence records	3,177	243
ticdata-2000	insurance risk assessment	5,822	689

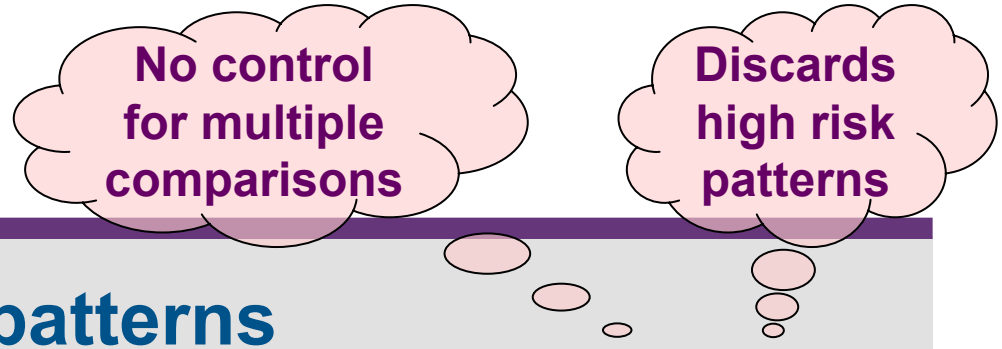
Spurious rules case study

Name	Records	Preds	Rules rejected
bms webview	59,601	497	170
covtype	581,012	125	998
ipums.la.99	88,443	1,874	973
kddcup98	52,256	19,662	995
letter-recognition	20,000	74	541
mush	8,124	127	891
retail	88,162	16,470	590
shuttle	58,000	34	666
splice-junction	3,177	243	748
ticdata-2000	5,822	689	996

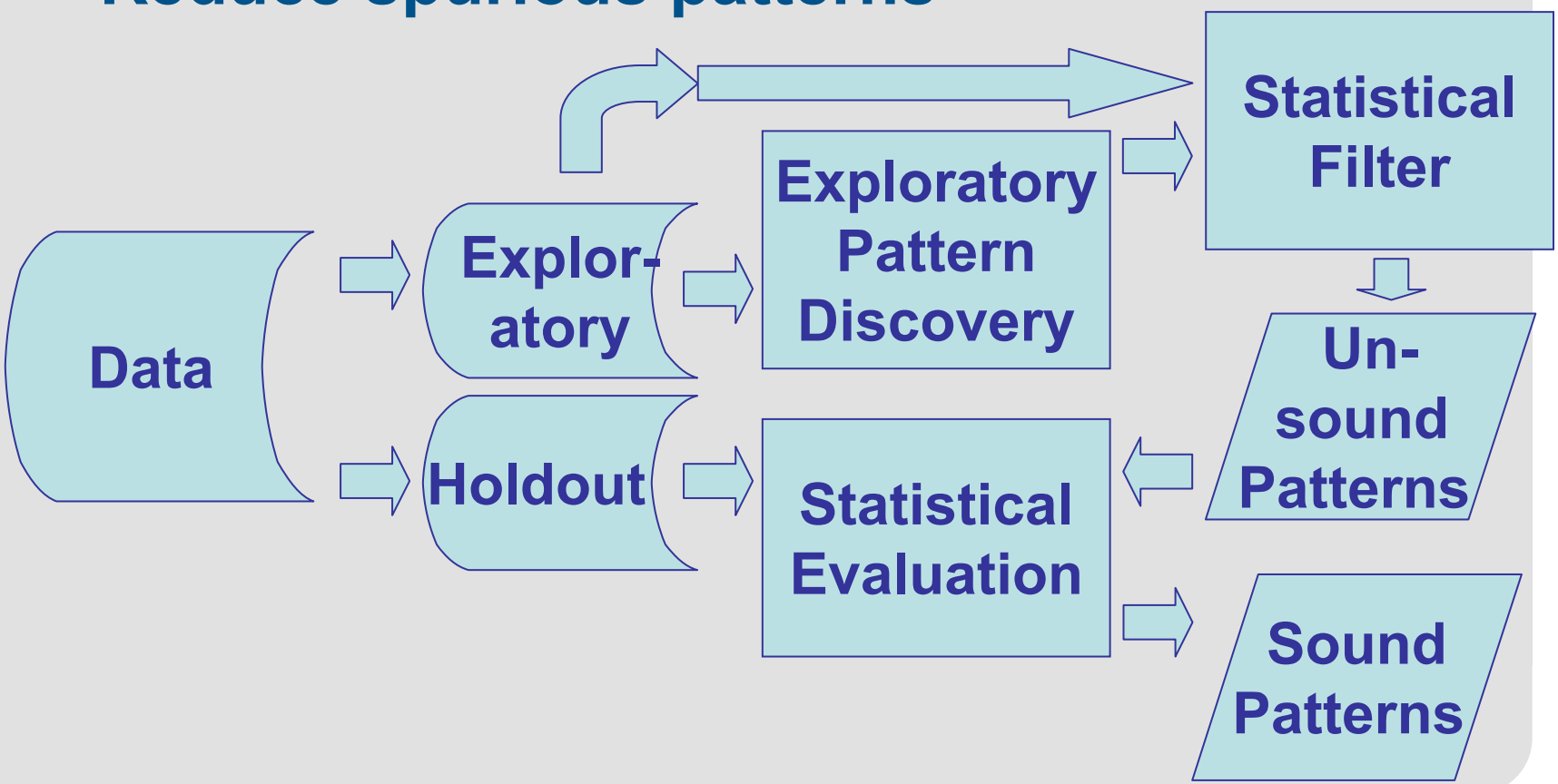
Filtering

- ↑ No. rules subjected to holdout evaluation
  - ↓ adjusted critical value
  - ↑ type-2 error
- Hence, want to select small number of rules
- Hence, want to exclude rules that are unlikely to pass holdout evaluation
- Solutions: perform statistical test during exploratory pattern discovery
  - unsound, but…
  - identifies high risk patterns that can be discarded

Filtering



- Reduce spurious patterns



Case Study: Failed Rules

dataset	filter			
	none	redundant	improvement	significant
bms webview	170	170	155	117
covtype	998	815	143	132
ipums.la.99	973	959	481	388
kddcup98	995	992	939	576
letter-recognition	541	524	421	291
mush	891	469	128	115
retail	590	590	519	408
shuttle	666	595	312	178
splice-junction	748	727	699	651
ticdata-2000	996	996	988	862

Summary

- Avoid the evils of model selection
- Minimum support gets overloaded
- K-optimal exploratory pattern discovery provides an efficient and effective alternative
- Statistically sound exploratory data analysis is desirable and achievable