



Finding the real patterns

Geoff Webb

Monash University

When I do data mining

I don't want to find JUNK!

Overview

False discoveries in pattern discovery

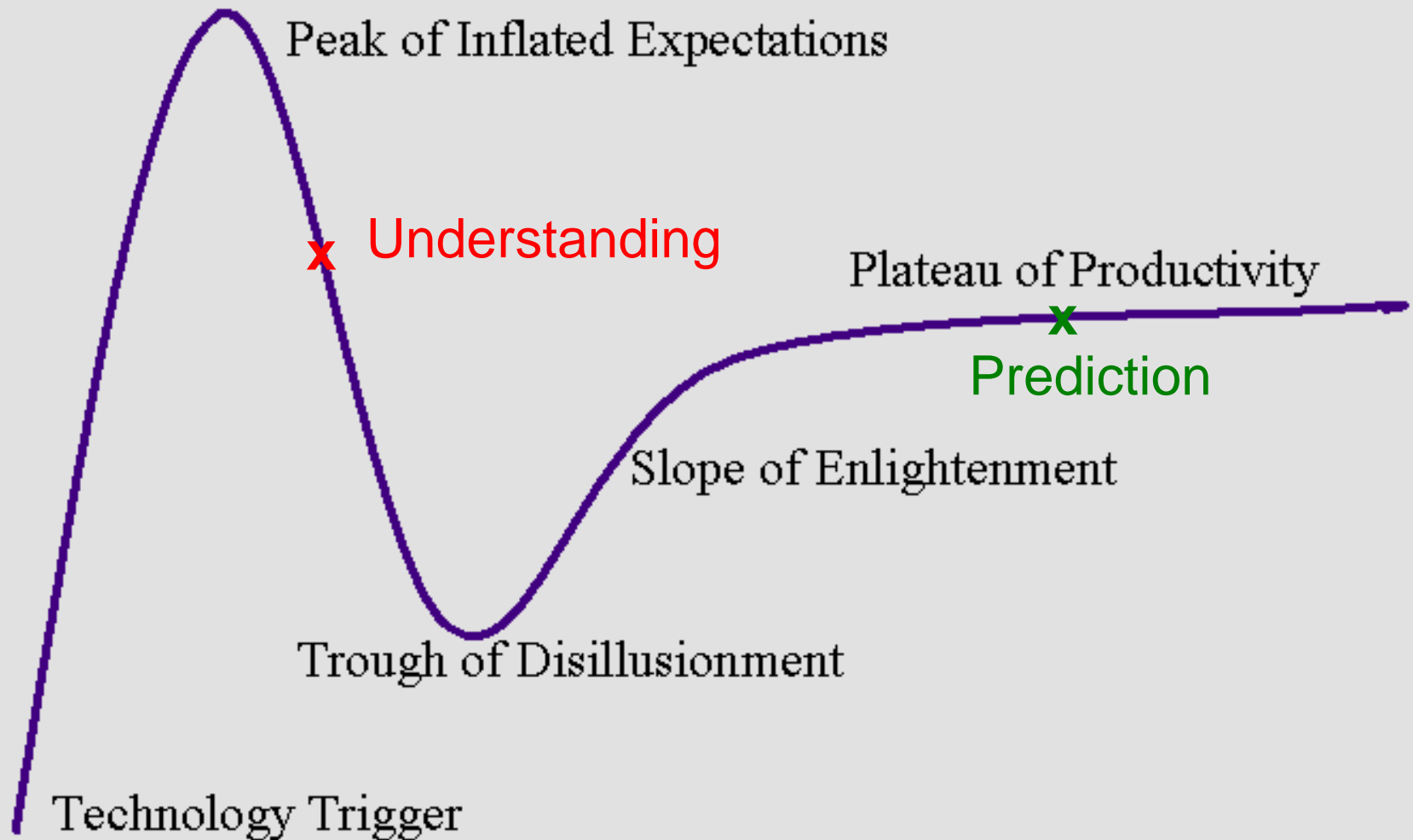
- **What is pattern discovery?**
- **Why it is important?**
- **Limitations of standard techniques**
- **These limitations can be ameliorated**

Prediction vs understanding

Data mining comprises two distinct forms of activity

- **Prediction**
 - decision trees, SVM, neural networks, naïve Bayes, linear regression, ...
 - relatively mature
- **Understanding**
 - clustering, pattern discovery, association rules, ...
 - relatively immature

Gartner Hype Cycle



The perils 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

bruises=f & *gill-attachment=f* & gill-spacing=c & ring-number=o
→ poisonous

[Coverage=0.406 (3296); Support=0.388 (3152); Confidence=0.956]

bruises=f & gill-spacing=c & *veil-color=w* & ring-number=o
→ poisonous

[Coverage=0.406 (3296); Support=0.388 (3152); Confidence=0.956]

Perils of model selection (cont.)

- **Data mining systems often make arbitrary choices**
 - **without warning!**
- **A system may have no basis on which to select models, but an expert often will**
 - **ease / cost of operationalisation**
 - **comprehensibility / compatibility with existing knowledge and beliefs**
 - **social / legal / ethical / political acceptability**

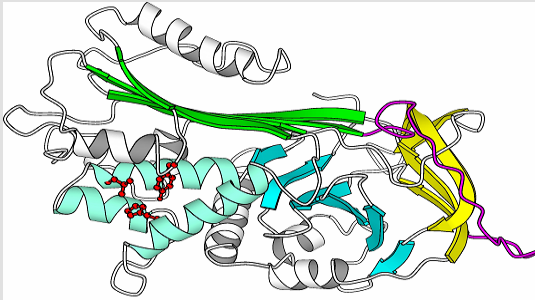
Pattern discovery

- **Pattern discovery seeks all patterns that satisfy user-defined constraints**
- **The user can select from these patterns**
 - **can use criteria that might be infeasible to quantify**



Patterns

- **Rules:**
 - $\langle \text{antecedent} \rangle \rightarrow \langle \text{consequent} \rangle$
- **Itemsets**
 - $\langle \text{condition}_1 \rangle \& \langle \text{condition}_2 \rangle \& \dots$
- **Sequences**
 - $\langle \text{event}_1 \rangle, \langle \text{event}_2 \rangle, \dots$
- **Structures**



Association rule and frequent pattern discovery

- First developed for market-basket analysis



- but very widely applicable
- Requires *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
- **Feast or famine**
 - minimum support is a crude control mechanism
 - often results in too few or too many associations
- **Cannot handle dense data**
- **Cannot prune search space using constraints on relationship between antecedent and consequent**
 - eg confidence
- **Minimum support may not be relevant**
 - cannot be sufficiently low to capture all valid rules
 - cannot be sufficiently high to exclude all spurious rules

Very low support patterns 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);
Confidence estimate=0.601; Lift estimate=192.06]

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

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

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

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

Very high support patterns can be spurious

Data file: covtype.data 581012 cases / 125 values

ST15=0 → ST07=0 [Coverage=1.000 (581009);
Support=1.000 (580904); Confidence=1.000]

ST07=0 → ST15=0 [Coverage=1.000 (580907);
Support=1.000 (580904); Confidence=1.000]

ST15=0 → ST36=0 [Coverage=1.000 (581009);
Support=1.000 (580890); Confidence=1.000]

ST36=0 → ST15=0 [Coverage=1.000 (580893);
Support=1.000 (580890); Confidence=1.000]

ST15=0 → ST08=0 [Coverage=1.000 (581009);
Support=1.000 (580830); Confidence=1.000]

ST08=0 → ST15=0 [Coverage=1.000 (580833);
Support=1.000 (580830); Confidence=1.000]

..... *197,183,686 such rules have highest support*

Soil Type

- **ST01 to ST40 are binary variables encoding the soil type of a region**
- **ST01=1 entails ST02=0, ... ST40=0**
- **Hence true associations are**
 - **STX=1 \rightarrow STY=0**
 - **STX=0 \rightarrow STY=1**
- **But almost 99% of cases are either ST01=1 or ST03=1 so ST02=0, ST04=0 ... ST40=0 all have support above 0.99.**
- **So almost all combinations form rules with high support and confidence!**

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 (aka top-*k*) 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
- $\text{lift}(A \rightarrow C) = \text{confidence}(A \rightarrow C) / F(C)$
 - proportional increase in confidence in context of antecedent
- $\text{leverage}(A \rightarrow C) = \text{support}(A \rightarrow C) - F(A) \times F(C)$
 - difference between observed and expected frequency
 - also known as *interest*

My techniques for k -optimal rule discovery

- **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**

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.

Applications

- I. I. Artamonova, G. Frishman, M. S. Gelfand, & D. Frishman (2005) Mining sequence annotation databanks for association patterns. *Bioinformatics* 21: 49-57. **Bioinformatics**
- Hei-Chia Wang, Yi-Shiun Lee and Tian-Hsiang Huang (2006) Gene Relation Finding Through Mining Microarray Data and Literature. In *Transactions on Computational Systems Biology V*, Springer: Berlin, pp. 83-96. **Web**
- Georgii E, Richter L, Ruckert U, Kramer S (2005) Analyzing microarray data using quantitative association rules. *Bioinformatics* 21: 123-129. **Web**
- Siu, K.K.W., et. al. (2005). Identifying markers of pathology in SAXS data of malignant tissues of the brain. *Nuclear Instruments & Methods in Physics Research A*, 548:140-144. **Robotics**
- Eirinaki, M. Vazirgiannis, I. Varlamis: (2003) SEWeP: using site semantics and a taxonomy to enhance the Web personalization process. *KDD 2003*: 99-108. **Finance**
- Thomas Hellström (2003) Learning Robotic Behaviors with Association Rules. *ASEAS Transactions on Systems*. Editor: Nikos Mastorakis. ISBN 1109-2777. **Finance**
- Jianxin Jiao, Yiyang Zhang: (2005) Product portfolio identification based on association rule mining. *Computer-Aided Design* 37(2): 149-172. **Health**
- Damien McAullay, Graham J. Williams, Jie Chen, Huidong Jin: (2005) A Delivery Framework for Health Data Mining and Analytics. *Australian Computer Science Conference 2005*: 381-390. **Health**
- J. Papaparaskevas, et. al. (2001) The use of Data Mining Techniques in Antibiotic Resistance Surveillance, Tech. Report, Dept. Informatics, Athens University of Economics and Business. **Management**
- Srinivas Vinnakota and Nina S.N. Lam, (2006) Socioeconomic inequality of cancer mortality in the United States: A spatial data mining approach. *International Journal of Health Geographics*. 5: 9. **Management**
- Tsironis L., Bilalis N., Moustakis V., (2001) Using inductive Machine Learning to support Quality Management, in *Proceedings of the 3rd International Conference on Design and Analysis of Manufacturing Systems*, Tinos Island. **Spatio-temporal Data**
- Mennis, J. and Liu, J.W., (2005) Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. *Trans. GIS*, 9(1): 13-18. **Data Management**
- Orna Raz (May 2004) Helping Everyday Users Find Anomalies in Data Feeds, Ph.D. Thesis - Software Engineering, Carnegie-Mellon University. **Data Management**

False discoveries

- **Massive search leads to high risk of false discoveries**
 - eg 100 observations, two independent events each occurring with 0.5 probability,
 - the probability of perfect correlation is 7.8×10^{-31} .
 - if there are 1000 events then there are $2^{1000} = 1.07 \times 10^{301}$ antecedent – consequent pairs.
 - ⇒ 10^{270} perfect correlations
- **What constitutes a false discovery depends on 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 ,
 - $\text{confidence}(A \ \& \ X \rightarrow B) \approx \text{confidence}(A \rightarrow B)$
 - $\text{lift}(A \ \& \ X \rightarrow B) \approx \text{lift}(A \rightarrow B)$
 - Eg *pregnant & Data Mining Researcher* \rightarrow *oedema*
- Special case: redundant rules
 - condition X is entailed by condition A
 - all standard metrics of interest, inc. confidence, lift and leverage, identical for specialisation & generalisation
 - Eg *pregnant & female* \rightarrow *oedema*
 - 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!

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**

Randomization tests

- Randomize data to instantiate the null hypothesis
eg variables are independent
- Run data mining algorithm
- Note most extreme value for some measure
eg support
- Repeat many times
- Run data mining algorithm on original data and accept any patterns more extreme than the α (eg 0.05) of most extreme values from randomized data
- Megiddo & Srikant, 1998; Gionis, Mannila, Mielikäinen & Tsaparas, 2006

Limitations of randomization testing

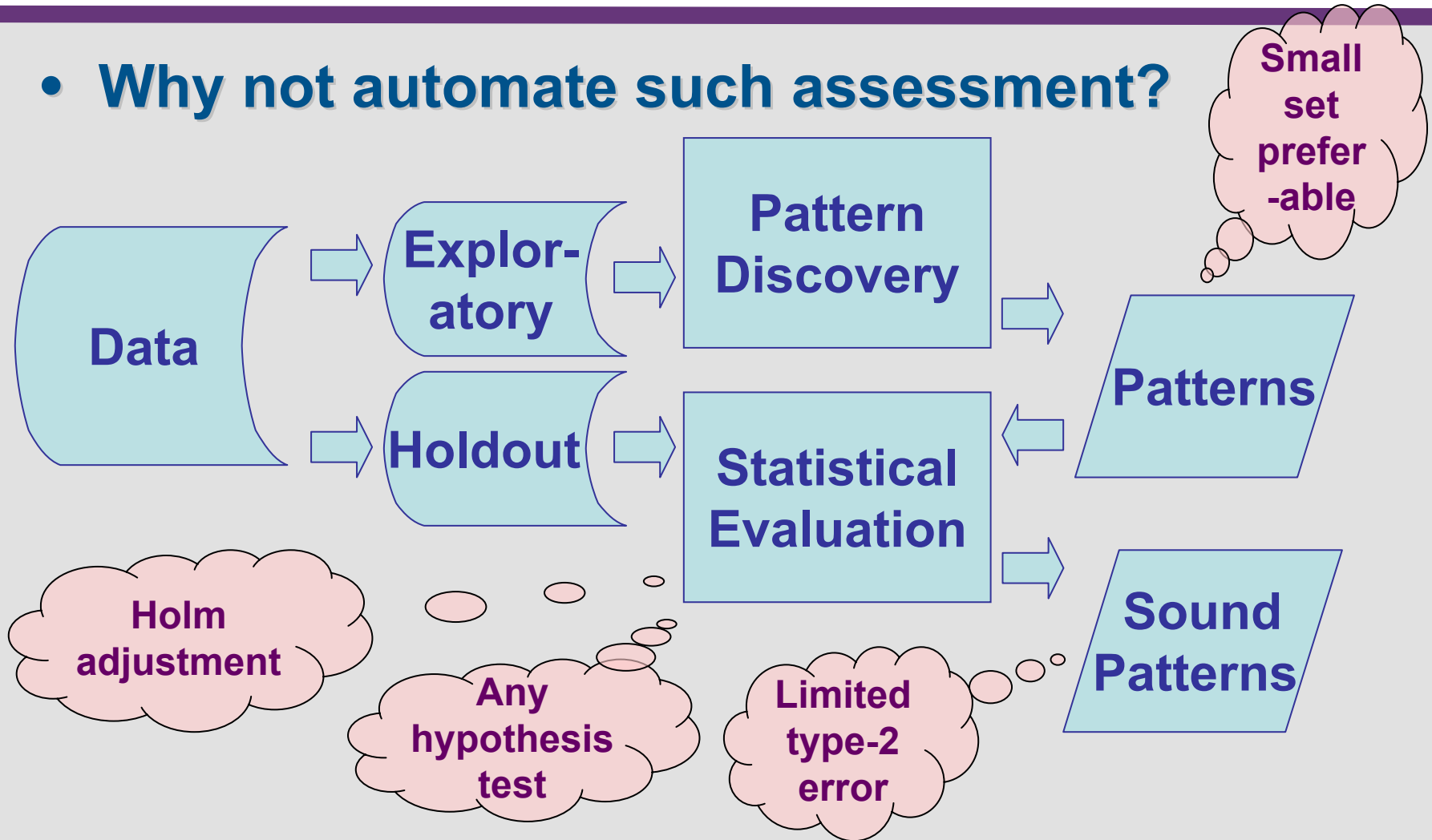
- **Computationally intensive**
- **Must be possible to instantiate the null hypothesis through a single randomization**
- **OK for all variables are independent**
- **Not possible for higher-order interactions**
eg pregnant & female → oedema

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?

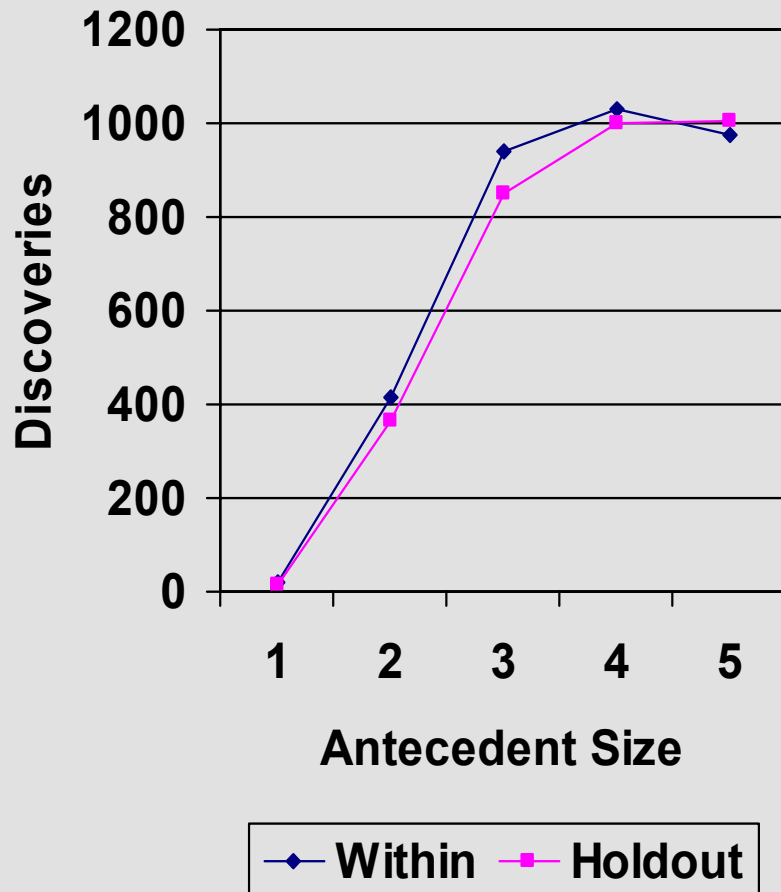


Direct adjustment

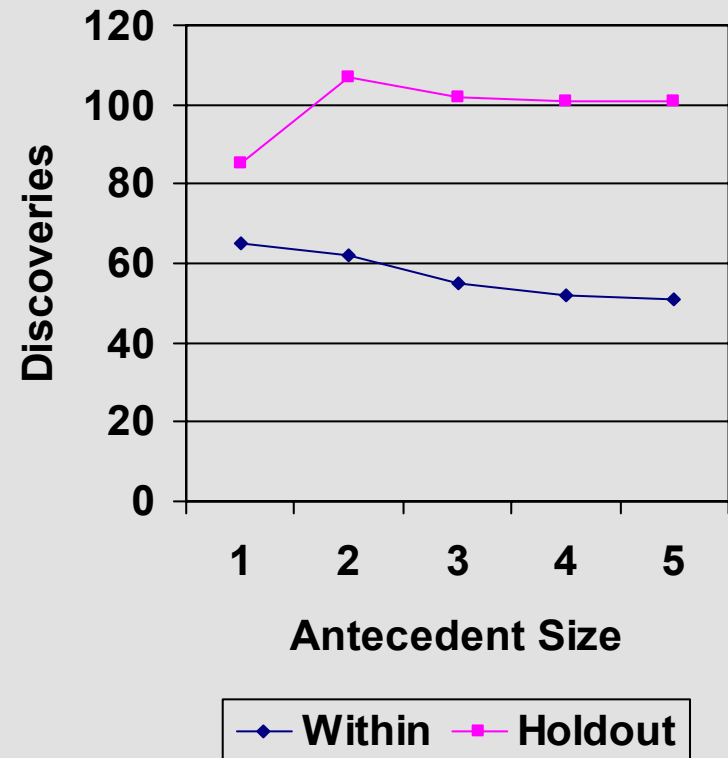
- An alternative is to perform significance tests during search with a Bonferroni adjustment for the size of the search space
 - BMS-WebView-1 1.25×10^{12} 4.00×10^{-14}
 - Covtype 1.55×10^{09} 3.23×10^{-11}
 - KDDCup98 8.76×10^{17} 5.71×10^{-20}
 - Retail 5.05×10^{19} 9.90×10^{-22}
- In most cases less powerful than holdout evaluation, but better support for *k*-optimal pattern discovery

Comparison on real data

Letter Recognition



Retail



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 confidence than all its generalisations
 - Eg, is $\text{confidence}(\text{pregnant \& female} \rightarrow B) >$
 - $\text{confidence}(\text{pregnant} \rightarrow B)$
 - $\text{confidence}(\text{female} \rightarrow B)$
 - $\text{confidence}(\text{true} \rightarrow B)$

Detecting spurious rules (cont)

- **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 Holm adjustment for strict control of type-1 error**

Spurious rules case study: high support & confidence non-redundant rules

Name	Records	Attribute -values	# Rules	# Accepted	%
bms webview	59,601	497	22,135	1,747	8
covtype	581,012	125	10,018	0	0
ipums.la.99	88,443	1,874	9,857	288	3
kddcup98	52,256	19,662	9,863	40	<1
letter-recognition	20,000	74	7,978	952	12
mush	8,124	127	8,957	1,266	14
retail	88,162	16,470	11,656	97	1
shuttle	58,000	34	9,760	876	9
splice-junction	3,177	243	8,937	132	1
ticdata-2000	5,822	689	10,438	30	<1

KDDCUP98: 99.5% of rules rejected

The following 40 rules passed holdout evaluation

...

ETH12<=0 → HC15<=0 [Coverage=0.987 (25786); Support=0.946 (24722); Confidence=0.959; Lift=1.00]

...

The following 9843 rules failed holdout evaluation, adjusted critical value = 5.09E-06

...

NOEXCH=0 & ETH12<=0 → HC15<=0 [Coverage=0.984 (25703); Support=0.943 (24644); Confidence=0.959; Lift=1.00]

...

NOEXCH=0 & ETH12<=0 & MDMAUD_F=X → HC15<=0 [Coverage=0.981 (25629); Support=0.940 (24573); Confidence=0.959; Lift=1.00]

...

NOEXCH=0 & ETH12<=0 & ADATE_2>=9706 & MDMAUD_R=X → HC15<=0 [Coverage=0.981 (25623); Support=0.940 (24567); Confidence=0.959; Lift=1.00]

...

Summary

- **Data mining for understanding is at an early stage in the hype cycle**
- **Pattern discovery**
 - **avoids the ‘perils of model selection’**
 - **empowers user to select the most useful patterns**
- **Minimum support can get overloaded**
- ***K*-optimal pattern discovery**
 - **circumvents need for minimum-support constraint**
 - **user specifies how to identify interesting patterns**
 - **user directly controls the number of patterns discovered**
- **If you mine for patterns without statistical evaluation, expect to find fool’s gold!**