

# Association Discovery

**Geoff Webb**

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

# Outline

---

- Association Discovery
- Interestingness
- Geoff's four principles for identifying associations that are unlikely to be interesting
  - in the absence of prior beliefs
- Geoff's approach to identifying associations that are likely to be interesting

# But wait!

---

- Why are we talking about association discovery?
- Subgroup/contrast/emerging pattern discovery
  - find description  $x$  such that  $P(x | y) > P(x)$
- Rule discovery
  - find description  $x$  such that  $P(y | x) > P(y)$
- $P(y | x) > P(y) \equiv P(x | y) > P(x)$ 
  - Webb, G. I., S. Butler, and D. Newlands (2003). On Detecting Differences Between Groups. In *KDD-2003*, pp. 256-265.
- Association discovery finds rules for all  $y$

# Claiming my place in history!

---

- OPUS is a very general and efficient algorithm for rule discovery
  - Webb, G. I. (1995). **OPUS: An Efficient Admissible Algorithm For Unordered Search.** *Journal of Artificial Intelligence Research* 3. AAAI Press, pp. 431-465.
- Hence OPUS is a very general and efficient algorithm for association, subgroup, contrast & emerging pattern discovery
  - The JAIR paper describes variants for finding all sets that satisfy specified constraints and for finding  $k$  sets that optimise an objective function under specified constraints

# Association discovery

---

- Association discovery is one of the signature data mining techniques
  - but fielded applications are few
  - why?
- In contrast: *Magnum Opus*
  - steady sales
  - scientific applications

## Scientific Applications of Magnum Opus

1. Il. Artamonova, G. Frishman, MS. Gelfand, & D. Frishman (2005) [Mining sequence annotation databanks for association patterns](#). *Bioinformatics*, 21: 49-57.
2. A. Bartholomeusz, S. Locarnini, L. Yuen, A. Ayres & M. Littlejohn (2007) [Multidrug Resistance and Cross-Resistance Pathways in HBV as a Consequence of Treatment Failure](#). *Journal of Hepatology* 46: S192.
3. C ássia Blondet Baruque, L úcia Blondet Baruque & Rubens Nascimento Melo (2006) [Using Data Mining for the Refresh of Learning Objects Digital Libraries](#). In *Proceedings of the International Conference on Engineering Education, ICEE-2006*.
4. C.B. Baruque, M.A. Amaral, A. Barcellos, J.C. da Silva Freitas and C.J. Longo (2007) [Analysing users' access logs in Moodle to improve e learning](#). In *Proceedings of the 2007 ACM Euro American conference on Telematics and information systems*. pp. 1-4.
5. D. Bhosale (2006) [AlcoZone: An Adaptive Hypermedia Based Personalized Alcohol Education](#). Masters Thesis. Virginia Polytechnic Institute and State University.
6. Jie Chen, Hongxing He, Huidong Jin, Damien McAullay, Graham Williams, & Chris Kelman, (2006) [Identifying Risk Groups Associated with Colorectal Cancer](#), *Lecture Notes in Computer Science*, Volume 3755, Pages 260 - 272.
7. R. Damaševičius (2009) [Analysis of Academic Results for Informatics Course Improvement Using Association Rule Mining](#). In G.A. Papadopoulos et al. (eds.), *Information Systems Development*, Springer, Berlin, pp 357-363.
8. George Dimitoglou & Shmuel Rotenstreich (2007) [A System for Association Rule Discovery in Emergency Response Data](#). In Sobh, Tarek (Ed.) *Innovations & Advanced Techniques in Computer & Information Sciences & Engineering*. Springer. Pages 193-199.
9. Daniel Druckman, Richard Harris, & Johannes F ürnkrantz (2006) [Modeling International Negotiation: Statistical and Machine Learning Approaches](#). In Robert Trappl (Ed.) *Modeling International Negotiation Statistical and Machine Learning Approaches*. Springer Netherlands, pp. 227-250.
10. J. P. Early & C. E. Brodley (2005) [Behavioral features for network anomaly detection](#), In M.A. Maloof (Ed.) *Machine Learning and Data Mining for Computer Security: Methods and Applications*, pp. 107-124, Springer.
11. Magdalini Eirinaki, Michalis Vazirgiannis, Iraklis Varlamis: (2003) [SEWeP: using site semantics and a taxonomy to enhance the Web personalization process](#). *KDD 2003*: 99-108.
12. E. Georgii, L. Richter, U. Ruckert, & S. Kramer (2005) [Analyzing microarray data using quantitative association rules](#). *Bioinformatics* 21: 123-129.
13. Thomas Hellström (2003) [Learning Robotic Behaviors with Association Rules](#). *WSEAS Transactions on Systems*. Editor: Nikos Mastorakis. ISBN 1109-2777.
14. J. Jiao, S. Pokharel, L. Zhang, & Y. Zhang (2005) [Coordination of product and process variety in mass customization with data mining approach](#). *Proc 10th Annual Int. Conference on Industrial Engineering – Theory, Applications and Practice*. Pages 342-348.
15. J. Jiao & Y. Zhang: (2005) [Product portfolio identification based on association rule mining](#). *Computer-Aided Design* 37(2): 149-172
16. J. Jiao, Y. Zhang, & M. Helander (2006) [A Kansei mining system for affective design](#). *Expert Systems with Applications*, 30(4): 658-673.

17. J. Jiao, Y. Zhang, & M. Helander (2006) [Analytical Customer Requirement Analysis Based on Data Mining](#). In Voges, K. & Pope, N. (Eds.), *Business Applications and Computational Intelligence*, ISBN: 1-59140-702-8, Chapter XII, pp. 227-247.
18. J. Jiao, L. Zhang, Y. Zhang, & S. Pokharel (2007) [Association rule mining for product and process variety mapping](#). *International Journal of Computer Integrated Manufacturing*, 21(1): 111-124.
19. J. Jiao, Q.L. Zu, & M. Helander (2007) [Analytical modeling and evaluation of customer citarasa in vehicle design](#). *IEEE Int. Conf. Indust. Eng. and Eng. Management*. pp.1277-1281.
20. Roger J. Jiao, Qianli Xu, Jun Du, Yiyang Zhang, Martin Helander, Halimahtun M. Khalid, Petri Helo & Cheng Ni (2007) [Analytical Affective Design With Ambient Intelligence For Mass Customization And Personalization](#). *International Journal of Flexible Manufacturing Systems*, 19(4): 570-595
21. D. McAullay, G.J. Williams, J. Chen, & H.Jin: (2005) [A Delivery Framework for Health Data Mining and Analytics](#). *Australian Computer Science Conference 2005*: 381-390.
22. J. Mennis (2006) [Socioeconomic-Vegetation Relationships in Urban, Residential Land: The Case of Denver, Colorado](#), *Photogrammetric Engineering & Remote Sensing*, 72(8):933.
23. J. Mennis & J.W. Liu, (2005) [Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change](#). *Transactions in GIS*, 9(1): 13-18.
24. S. K. Moon, S. R. T. Kumara, & T. W. Simpson, (2006) [Data Mining and Fuzzy Clustering to Support Product Family Design](#). *Proc. DETC06, 2006 ASME Design Engineering Technical Conferences*, Philadelphia, PA, Paper No. DETC2006/DAC-99287.
25. S.K. Moon, T.W. Simpson, & S.R.T. Kumara (2010) [A methodology for knowledge discovery to support product family design](#). *Annals of Operations Research*, 174(1): 201-218.
26. H.K. Nehemiah & A.Kannan (2006) [A Diagnostic Decision Support System For Adverse Drug Reaction Using Temporal Reasoning](#). *The International Journal of Artificial Intelligence and Machine Learning*. 6(2): 79-86.
27. J. Papaparaskevas, Y. Batistakis, M. Halkidi, C. Amanatidis, M. Kanellopoulou, M. Vazirgiannis, E. Papafragas, & A. Vatopoulos (2001) [The use of Data Mining Techniques in Antibiotic Resistance Surveillance](#), Tech Rep, Dept Informatics, Athens University Economics & Business.
28. Q.L. Nguyen, I. Pilgermann, A. Gill, A. Guhr, T. Zhang, K. von Eckardstein, T. Picht, J. Veelken, R. L. Martuza, A. von Deimling, & A. Kurtz (2010) [Identification of diagnostic serum protein profiles of glioblastoma patients](#). *Journal of Neuro-Oncology*.
29. Orna Raz (2004) [Helping Everyday Users Find Anomalies in Data Feeds](#), Ph.D. Thesis - Software Engineering, Carnegie-Mellon University.
30. K.K.W. Siu, S.M. Butler, T. Beveridge, J.E. Gillam, C.J. Hall, A.H. Kaye, R.A. Lewis, K. Mannan, G. McLoughlin, S. Pearson, A.R. Round, E. Schultke, G.I. Webb, & S.J. Wilkinson (2005). [Identifying markers of pathology in SAXS data of malignant tissues of the brain](#). *Nuclear Instruments and Methods in Physics Research A*, 548:140-146. [[Pre-publication PDF](#)]
31. D. Strand (2005) [Controlling a Robot using Association Rules with a Temporal Component](#). Masters Thesis. Department of Computer Science, Umea University.
32. L. Tsiaronis, N. Bilalis, & V. Moustakis, (2005) [Using machine learning to support quality management](#). *The TQM Magazine*, 17(3): 237 - 248.
33. S Vinnakota & N. S.N. Lam, (2006) [Socioeconomic inequality of cancer mortality in the United States: A spatial data mining approach](#). *Int. Journal Health Geographics*. 5: 9.
34. Hei-Chia Wang, Yi-Shiun Lee, & Tian-Hsiang Huang (2006) [Gene Relation Finding Through Mining Microarray Data and Literature](#). In C. Priami et al. (Eds.): *Transactions on Computational Systems Biology V*, LNBI 4070, pp. 83 – 96.

# Association Discovery

---

- Find things that are associated with one another
- How is this different from statistical association analysis?
  - It's not – we just do it badly?
  - Market basket data?
  - High volume data?
  - High dimensional data?
  - Associations between values rather than variables!
  - Focus on utility!

# Why does Association Discovery matter?

---

- Exploratory data analysis
- Local models
  - Global models may trade-off some local optimality for the sake of global optimality

# Why does Association Discovery matter?

---

- Avoids the evils of model selection

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

→ poisonous

[Coverage=3296; Support=3152; Confidence=0.956]

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

→ poisonous

[Coverage=3296; Support=3152; Confidence=0.956]

- Empowers the user to select useful local models

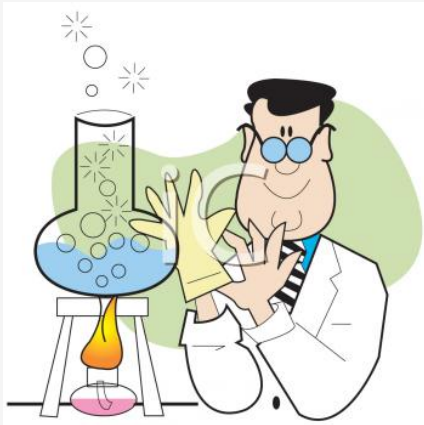
# Association Discovery Research

---

- Much research into efficient techniques
- Little consideration of what techniques should do

# Interestingness

---



# Geoff's 1<sup>st</sup> principle

---

- A conjunction of terms is unlikely to be interesting if its frequency can be predicted by assuming independence between a partition thereof
- pregnant, oedema, random
  - pregnant, oedema
  - random
- male, poor-eyesight, prostate cancer, glasses
  - male, prostate cancer
  - poor-eyesight, glasses

# Geoff's 2<sup>nd</sup> principle

---

- A conjunction of terms is unlikely to be interesting if its frequency can be predicted by the frequency of its specialisations
- flammable, oxygen, heat, fire
- flammable, oxygen, fire
- flammable, heat, fire
- oxygen, heat, fire
- flammable, fire
- oxygen, fire
- heat, fire

# Geoff's 3<sup>rd</sup> principle

---

- **Appropriate statistical testing should be applied to assess the first two principles**

# False Discoveries

---

- We typically want to discover associations that hold in the process that generated the data
- The massive search involved in association discovery results in a massive risk of false discoveries
  - associations that appear to hold in the sample but do not hold in the generating process

# Massive Search

---

- Retail
  - 3182 items
  - $2^{3182}$  possible rules
  - $> 10^{16}$  rules with antecedent  $\leq 4$  items
- Probability is high of very improbable co-occurrences

# Statistical tests

---

- Cannot just apply statistical test to each rule
  - Critical value of 0.05 applied to  $10^{16}$  rules wrt random data should produce  $> 10^{14}$  false discoveries!
- Solutions
  - Randomization tests
  - Bonferroni correction
  - Holdout evaluation

# Randomization tests

---

- Repeat many times:
  - shuffle data
  - apply association discovery to shuffled data
- Measure the 5<sup>th</sup> percentile (or value corresponding to desired critical value) of some statistic in resulting discoveries and accept discoveries that exceed that value

# Advantages of randomization tests

---

- Automatically take account of the relationships between rules that may complicate other approaches to testing.
- Easy to implement.

# Randomization Tests: Concerns

---

- If only consider one statistic wrt all associations,
  - low power
  - may miss many true associations
    - will only find associations with extreme values on the one metric

# Randomization Tests: Concerns

---

- Stochastic
  - Results may vary
- Independence between all items not the only null-hypothesis!
  - pregnant & random-val → oedema

# KDDCUP98: 10,000 rules, max sup, lift $\geq 2.0$

---

Found 10,000 rules / 28 rules passed holdout evaluation,  $\alpha = 5.0005E-006$

**91 <=HU3<=96 -> 4<=HU4<=9**  
[Coverage=11466; Support=11465; Lift=2.28]

**91 <=HU3<=96 & ADATE\_2>=9706 & RFA\_2R=L -> 4<=HU4<=9**  
[Coverage=11464; Support=11463; Lift=2.28]  
Holdout coverage = 11573, holdout support = 11571

**MDMAUD=XXXX & 91 <=HU3<=96 & MDMAUD\_R=X & MDMAUD\_A=X**  
**-> 4<=HU4<=9**  
[Coverage=11444; Support=11443; Lift=2.28]  
Holdout coverage = 11535, holdout support = 11533

**91 <=HU3<=96 & MAXADATE>=9702 -> 4<=HU4<=9**  
[Coverage=11414; Support=11413; Lift=2.28]  
Holdout coverage = 11532, holdout support = 11530

# Within search Bonferroni correction

---

- Apply statistical test while doing search
- Apply Bonferroni correction
  - divide critical value by size of search space
  - Eg retail
    - $> 10^{16}$  rules with antecedent  $\leq 4$  items
    - $\alpha \approx 5 \times 10^{-18}$
- Can *layer* the critical values
  - antecedent 1: critical value =  $9.2 \times 10^{-011}$
  - antecedent 2: critical value =  $5.5 \times 10^{-015}$
  - antecedent 3: critical value =  $1.0 \times 10^{-018}$
  - antecedent 4: critical value =  $2.4 \times 10^{-022}$

# Advantages of within search testing

---

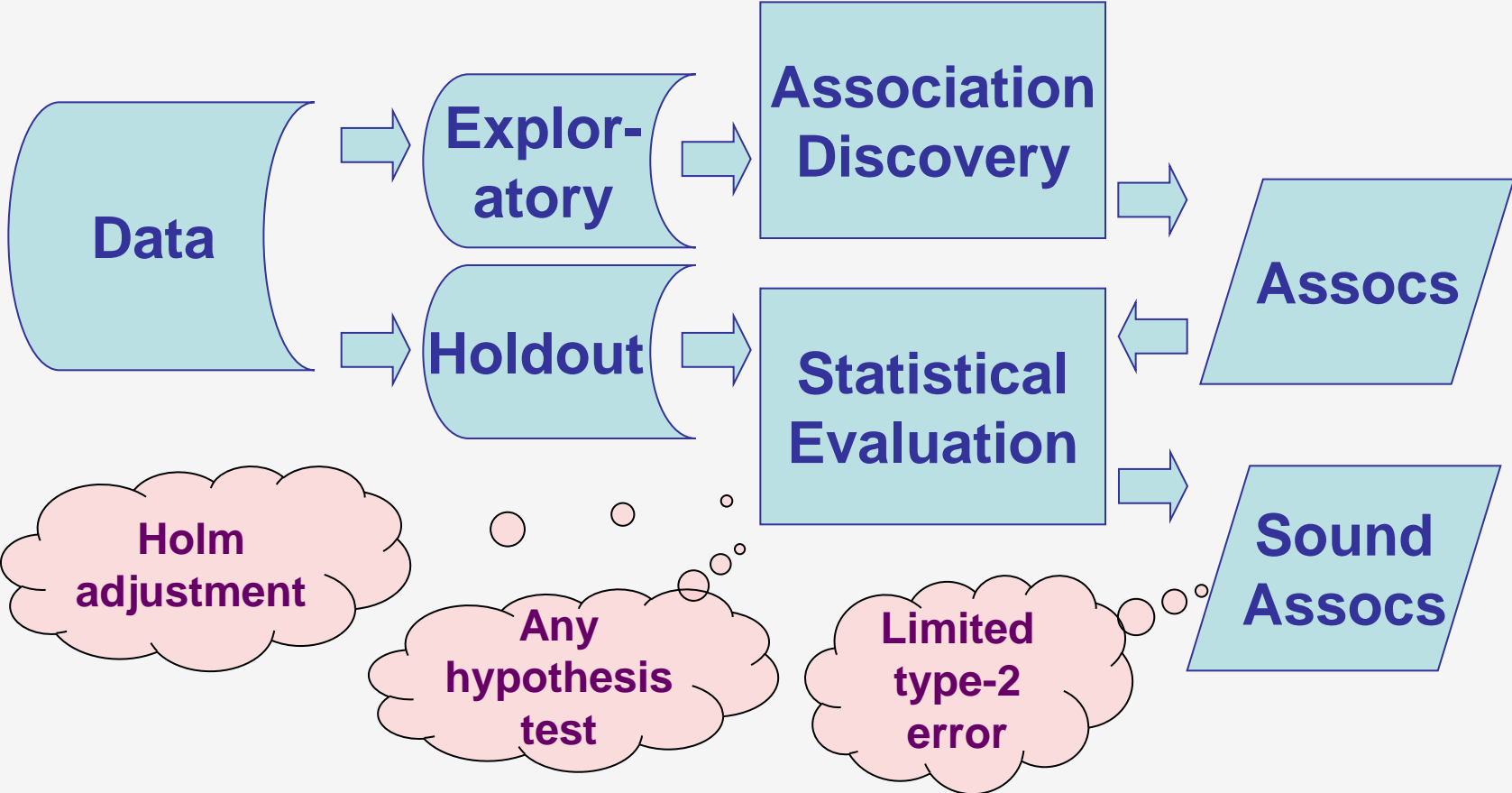
- Can be used to prune search space
- Supports top-k techniques
- Can apply any statistical test

# Disadvantages of within search testing

---

- Slightly lower power than holdout
- Computational issues if computationally expensive tests used
- Requires that the size of the search space can be determined

# Holdout evaluation



# Strengths of holdout evaluation

---

- Can apply any statistical test
- High power
  - Low miss rate
- Does not require that the size of the search space can be determined

# Disadvantages of holdout evaluation

---

- Does not support top-k techniques
- Different random divisions into exploratory and holdout data will result in different outcomes

# Geoff's 4<sup>th</sup> principle (conjecture?)

---

- If a conjunction of terms is unlikely to be interesting then any rule composed from those terms is unlikely to be interesting
- pregnant & random → oedema
- pregnant & oedema → random
- random → pregnant & oedema
- pregnant → random & oedema

# Are rules a good representation?

---

- Association discovery usually finds rules
- Why organise associated items into an antecedent and a consequent?
- An association between two items will be represented by two rules
  - three items – nine rules
  - four items – twenty-eight rules
  - ...
- It may not be apparent that all the resulting rules represent a single multi-item association

# Rules

bruises?=t -> ring-type=p  
[Coverage=3376; Support=3184; Lift=1.93; p<4.94E-322]

ring-type=p -> bruises?=t  
[Coverage=3968; Support=3184; Lift=1.93; p<4.94E-322]

stalk-surface-above-ring=s & ring-type=p -> bruises?=t  
[Coverage=3664; Support=3040; Lift=2.00; p=6.32E-041]

stalk-surface-below-ring=s & ring-type=p -> bruises?=t  
[Coverage=3472; Support=2848; Lift=1.97; p=9.66E-013]

stalk-surface-above-ring=s & stalk-surface-below-ring=s &  
ring-type=p -> bruises?=t  
[Coverage=3328; Support=2776; Lift=2.01; p=0.0166]

stalk-surface-above-ring=s & stalk-surface-below-ring=s ->  
ring-type=p  
[Coverage=4156; Support=3328; Lift=1.64; p=5.89E-178]

stalk-surface-above-ring=s & stalk-surface-below-ring=s ->  
bruises?=t  
[Coverage=4156; Support=2968; Lift=1.72; p=1.47E-156]

stalk-surface-above-ring=s -> ring-type=p  
[Coverage=5176; Support=3664; Lift=1.45; p<4.94E-322]

ring-type=p -> stalk-surface-above-ring=s  
[Coverage=3968; Support=3664; Lift=1.45; p<4.94E-322]

stalk-surface-below-ring=s & ring-type=p -> stalk-surface-  
above-ring=s  
[Coverage=3472; Support=3328; Lift=1.50; p=3.05E-072]

stalk-surface-above-ring=s & ring-type=p -> stalk-surface-

below-ring=s  
[Coverage=3664; Support=3328; Lift=1.49; p=3.05E-072]

bruises?=t -> stalk-surface-above-ring=s  
[Coverage=3376; Support=3232; Lift=1.50; p<4.94E-322]

stalk-surface-above-ring=s -> bruises?=t  
[Coverage=5176; Support=3232; Lift=1.50; p<4.94E-322]

stalk-surface-below-ring=s -> ring-type=p  
[Coverage=4936; Support=3472; Lift=1.44; p<4.94E-322]

ring-type=p -> stalk-surface-below-ring=s  
[Coverage=3968; Support=3472; Lift=1.44; p<4.94E-322]

bruises?=t & stalk-surface-below-ring=s -> stalk-surface-  
above-ring=s  
[Coverage=3040; Support=2968; Lift=1.53; p=1.56E-036]

stalk-surface-below-ring=s -> stalk-surface-above-ring=s  
[Coverage=4936; Support=4156; Lift=1.32; p<4.94E-322]

stalk-surface-above-ring=s -> stalk-surface-below-ring=s  
[Coverage=5176; Support=4156; Lift=1.32; p<4.94E-322]

bruises?=t & stalk-surface-above-ring=s -> stalk-surface-  
below-ring=s  
[Coverage=3232; Support=2968; Lift=1.51; p=1.56E-036]

bruises?=t -> stalk-surface-below-ring=s  
[Coverage=3376; Support=3040; Lift=1.48; p<4.94E-322]

stalk-surface-below-ring=s -> bruises?=t  
[Coverage=4936; Support=3040; Lift=1.48; p<4.94E-322]

# Itemsets

---

**bruises?=t & stalk-surface-above-ring=s &  
stalk-surface-below-ring=s & ring-type=p**

**[Coverage=2776; Leverage=928.9;  $p < 4.94E-322$ ]**

# But how to measure interest?

---

- Most measures for selecting rules measure the degree to which the antecedent and consequent deviate from independence
  - Lift, leverage, etc
- However, we do not want to measure interest for an itemset as deviation from independence between all items
  - Would allow any irrelevant item to be added to any interesting itemset
  - pregnant & oedema & random-val

# Itemset interest

---

- Measure smallest deviation from independence between any two subsets of the itemset
  - lowest value of measure of interest for any rule that can be formed from all items allowing multiple items in both antecedent and consequent
  - eg leverage =  $\min[\text{sup}(a \& b) - \text{sup}(a)\text{sup}(b)]$  such that  $a \subset I, b \subset I, a \cup b = I$  and  $a \cap b = \{\}$

# Supervised descriptive rule discovery

---

- Rules are useful where
  - interested in associations with a specific item
    - eg cancer
  - interested in contrasts/subgroups/....

# What's wrong frequent pattern discovery?

---

- Discontinuity in objective 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 low enough to capture all valid rules
  - cannot be high enough to exclude all spurious rules

# Very infrequent patterns can be significant

---

Data file: Brijs retail.itl, 88162 cases / 16470 items

237 -> 1

[Coverage=3032; Support=28; Lift=3.06; p=1.99E-007]

237 & 4685 -> 1

[Coverage=19; Support=9; Lift=157.00; p=5.03E-012]

1159 -> 1

[Coverage=197; Support=9; Lift=15.14; p=1.13E-008]

4685 -> 1

[Coverage=270; Support=9; Lift=11.05; p=1.68E-007]

168 -> 1

[Coverage=293; Support=9; Lift=10.18; p=3.33E-007]

4382 -> 1

[Coverage=72; Support=8; Lift=36.83; p=6.26E-011]

168 & 4685 -> 1

[Coverage=9; Support=7; Lift=257.78; p=6.66E-011]

# Very high support patterns can be spurious

---

Data file: covtype.data 581012 cases / 125 values

ST15=0 → ST07=0

[Coverage=581009; Support=580904; Confidence=1.000]

ST07=0 → ST15=0

[Coverage=580907; Support=580904; Confidence=1.000]

ST15=0 → ST36=0

[Coverage=581009; Support=580890; Confidence=1.000]

ST36=0 → ST15=0

[Coverage=580893; Support=580890; Confidence=1.000]

ST15=0 → ST08=0

[Coverage=581009; Support=580830; Confidence=1.000]

ST08=0 → ST15=0

[Coverage=580833; Support=580830; Confidence=1.000]

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

# What's the alternative to minimum support?

---

- Top- $k$  techniques
  - Allow user to specify how many associations should be discovered
  - Allow the user to specify the objective function
    - Interestingness metric
  - Use objective function together with  $k$  to make computation feasible
  - Find top- $k$  associations within constraints implied by Geoff's four principles

# Summary

---

- Much more attention has been paid to finding associations efficiently than to which ones to find
- While we cannot be certain what will be interesting, the following probably won't
  - frequency explained by independence between a partition
  - frequency explained by specialisations
- Statistical testing is essential
- If an itemset is not interesting then neither is a rule composed from all its elements
- Itemsets often provide a much more succinct summary of association than rules

# References

---

- Webb, G. I. (1995). OPUS: An Efficient Admissible Algorithm For Unordered Search. Journal of Artificial Intelligence Research 3. AAAI Press, pp. 431-465. [[Link to paper via JAIR website](#)]
  - Webb, G. I. (2000). Efficient Search for Association Rules. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000) Boston, MA. New York: The Association for Computing Machinery, pp. 99-107. [[Pre-publication PDF](#)][[Link to paper via ACM Portal](#)]
  - Webb, G. I., S. Butler, and D. Newlands (2003). On Detecting Differences Between Groups. In Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003) pp. 256-265. [[PDF](#)][[Paper via ACM Portal](#)]
  - Webb, G. I. and S. Zhang (2005). k-Optimal-Rule-Discovery. Data Mining and Knowledge Discovery 10(1):39-79. [[Prepublication PDF](#)][[Link to paper via Springerlink](#)]
  - Webb, G.I. (2007). Discovering Significant Patterns. Machine Learning 68(1): 1-33. [[Pre-publication PDF](#)][[Link to paper via Springerlink](#)]
  - Webb, G.I. (2008). Layered Critical Values: A Powerful Direct-Adjustment Approach to Discovering Significant Patterns. Machine Learning 71(2-3): 307-323 [Technical Note] [[Pre-Publication PDF](#)][[Link to paper via Springerlink](#)]
  - Webb, G.I. (2010). Self-Sufficient Itemsets: An Approach to Screening Potentially Interesting Associations Between Items. Transactions on Knowledge Discovery from Data 4: 3:1-3:20. [[Pre-Publication PDF](#)][[Link to paper via ACM Digital Library](#)]
  - Webb, G.I. (2010) Magnum Opus version 4.6.3. Computer Software. <http://www.giwebb.com>
-