MINING FLIPPING CORRELATIONS FROM LARGE DATASETS WITH TAXONOMIES

MARINA BARSKY, SANGKYUM KIM,
TIM WENINGER, JIAWEI HAN

VLDB 2012

DEPT OF COMPUTER SCIENCE
UNIV OF ILLINOIS AT URBANA-CHAMPAIGN

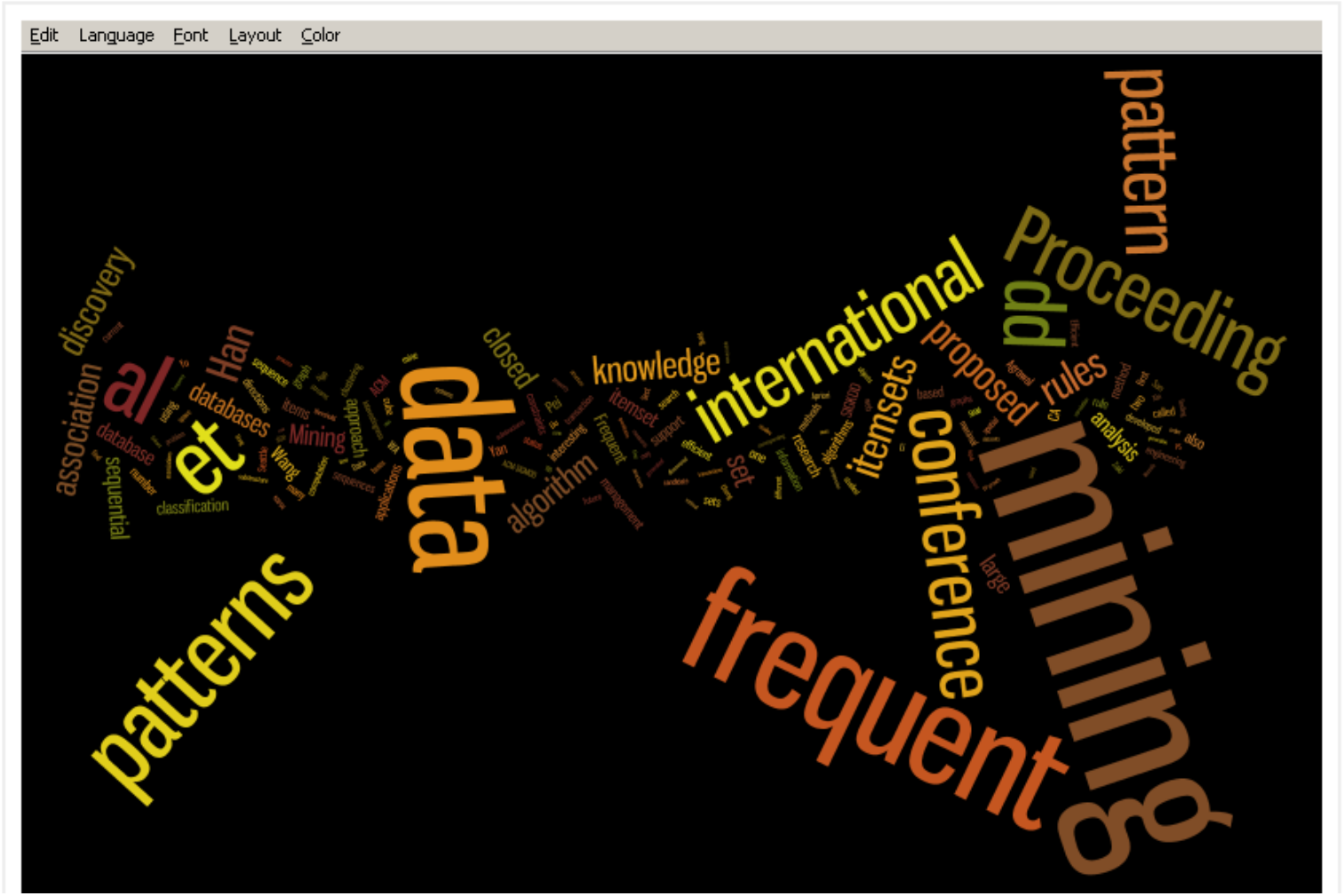
Outline

- Challenge: strong correlations with low support
- Flipping correlation patterns
- Algorithm for mining flipping correlations
- Performance
- Real flipping patterns
- Conclusion and future work

Correlations and frequent itemsets

- Once all frequent itemsets are enumerated, we can find correlation between items in these frequent itemsets
- Computation of frequent itemsets is feasible only for high support thresholds
- Top-frequent itemsets often represent obvious relationships between items

Example: frequent items in papers on frequent pattern mining



Challenge of finding itemsets with low support

In large datasets we can find the top most frequent itemsets

- When we lower the support threshold, the number of frequent itemsets becomes big
- How big? Very big: that we cannot keep in memory all different 2-item combinations, to update their counters

How can we discover non-trivial correlations in large datasets?

- Instead of computing top-frequent, compute top-correlated patterns directly, without enumerating all frequent itemsets

- This presents computational challenges
- Some progress in this direction is in our previous paper

Sangkyum Kim et al., ECML/PKDD (2) 2011: 177-192

Negative correlations

- What if we are also interested in items that rarely appear in the same transaction?

- The *negative correlations* can be useful:
  - To identify competing items: absence of Blu ray and DVD player in the same transaction
  - To discover underrepresented topic combinations: in DBLP –{mobile networks, data cube}
- The set of all itemsets where items are negatively correlated is exponentially large and "the solution remains elusive"

P.-N. Tan et al., 2005.

Challenge: all positive and negative correlations in itemsets with low-to-medium support

- Computing all frequent itemsets with very low support is computationally prohibitive
- Most of the correlation measures for large datasets possess neither monotonicity nor anti-monotonicity properties, and as such cannot be straightforwardly used for pruning purposes.

Outline

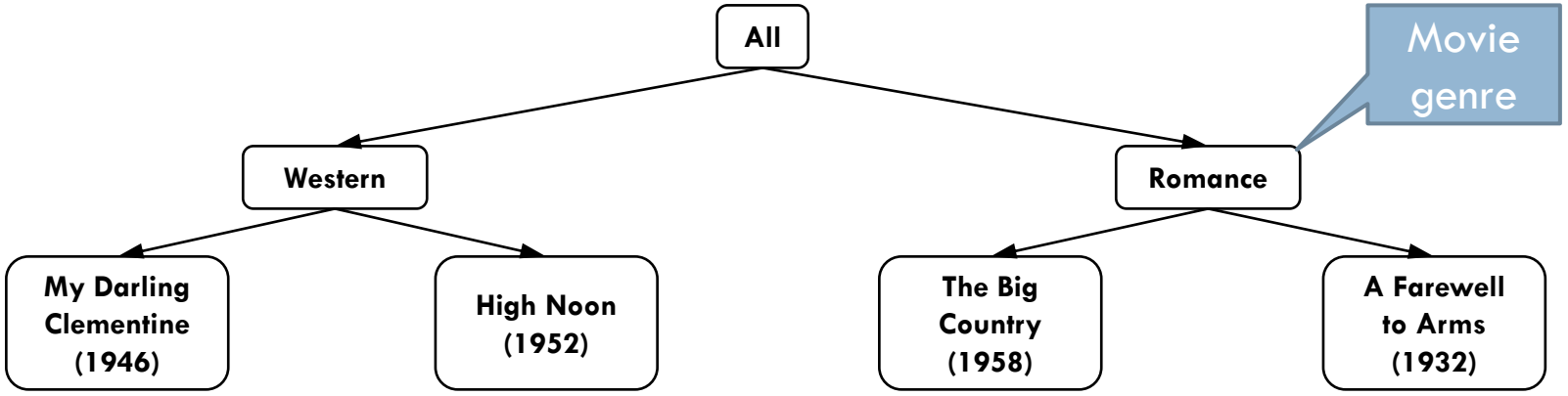
- ✓ Challenge: strong correlations with low support
- Flipping correlation patterns
- Algorithm for mining flipping correlations
- Performance
- Real flipping patterns
- Conclusion and future work

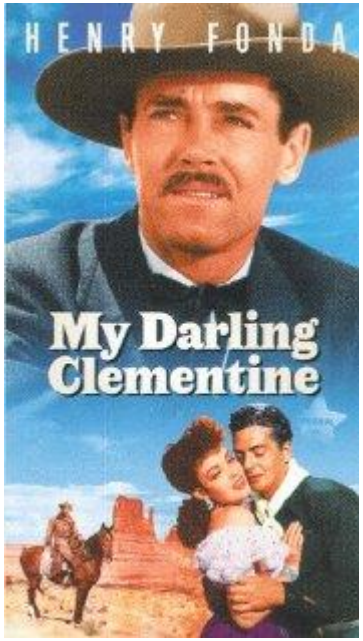
Feasible task with the use of taxonomy

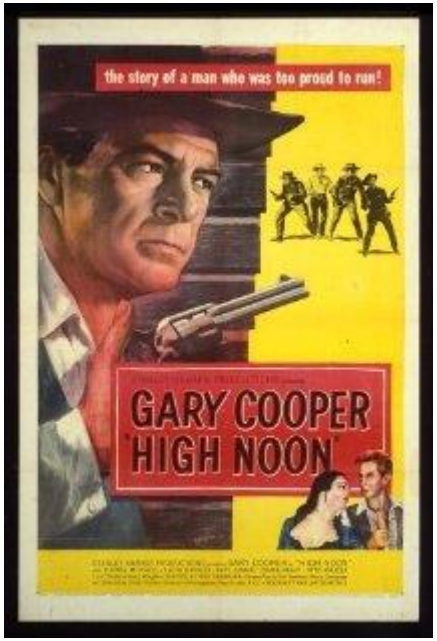
- We cannot compute all positive and negative correlations with low support
- We can find the most surprising positive and negative correlations, which change across the levels of abstraction

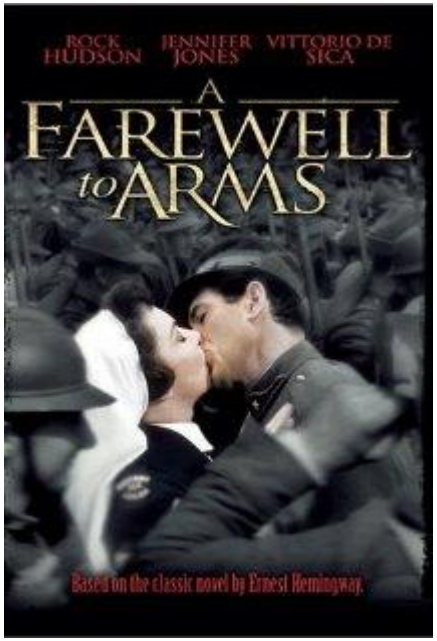
- Items at different levels of abstraction can be modeled as a taxonomy tree

Example of taxonomy: movies

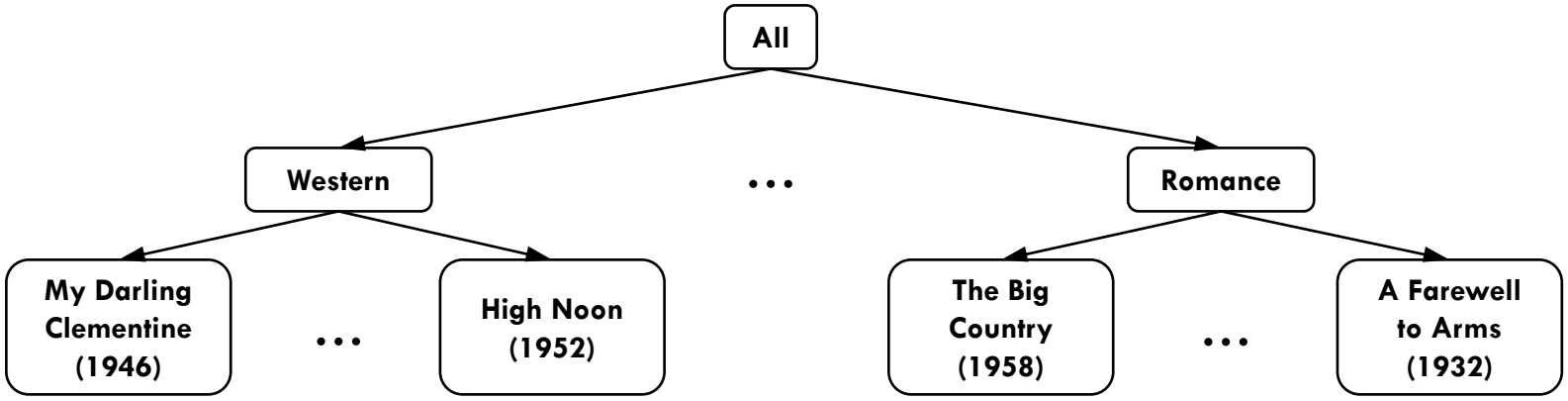


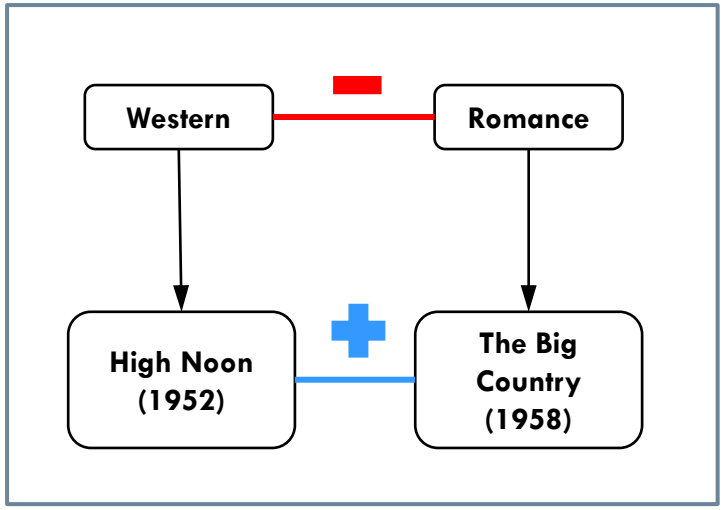






Example: flipping correlations in Movielens dataset





Flipping Correlation Example

People who like **westerns** do not like **romance** movies (**negative correlation**)

Despite this general rule, people who like **High Noon** (western) also like **The big Country** (romance) (**positive correlation**)

Flipping correlations are surprising

- If two groups of items are negatively correlated, but some sub-groups are positively correlated. What is so special about them?
- The positive correlation between two groups of items suggest that the items in both groups behave similarly. But some sub-groups are negatively correlated. Why?

- We leave these questions to domain experts, and our contribution is an efficient computation of all flipping correlations

Outline

- ✔ Challenge: strong correlations with low support
- ✔ Flipping correlation patterns
- Algorithm for mining flipping correlations
- Performance
- Real flipping patterns
- Conclusion and future work

Selecting correlation measure

Two groups of correlation measures

- Null-invariant
- Expectation-based

Null-(transaction) invariance is crucial for large datasets

Measure	Definition	Range	Null-Invariant
$\chi^2(a,b)$	$\sum_{i,j=0,1} \frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0,\infty]$	No
$Lift(a,b)$	$\frac{P(ab)}{P(a)P(b)}$	$[0,\infty]$	No
$AllConf(a,b)$	$\frac{sup(ab)}{max\{sup(a),sup(b)\}}$	$[0,1]$	Yes
$Coherence(a,b)$	$\frac{sup(ab)}{sup(a)+sup(b)-sup(ab)}$	$[0,1]$	Yes
$Cosine(a,b)$	$\frac{sup(ab)}{\sqrt{sup(a)sup(b)}}$	$[0,1]$	Yes
$Kulc(a,b)$	$\frac{sup(ab)}{2}(\frac{1}{sup(a)}+\frac{1}{sup(b)})$	$[0,1]$	Yes
$MaxConf$(a,b)	$max\{\frac{sup(ab)}{sup(a)}, \frac{sup(ab)}{sup(b)}\}$	$[0,1]$	Yes

T. Wu et al., 2010.

Challenge with null-invariant measures

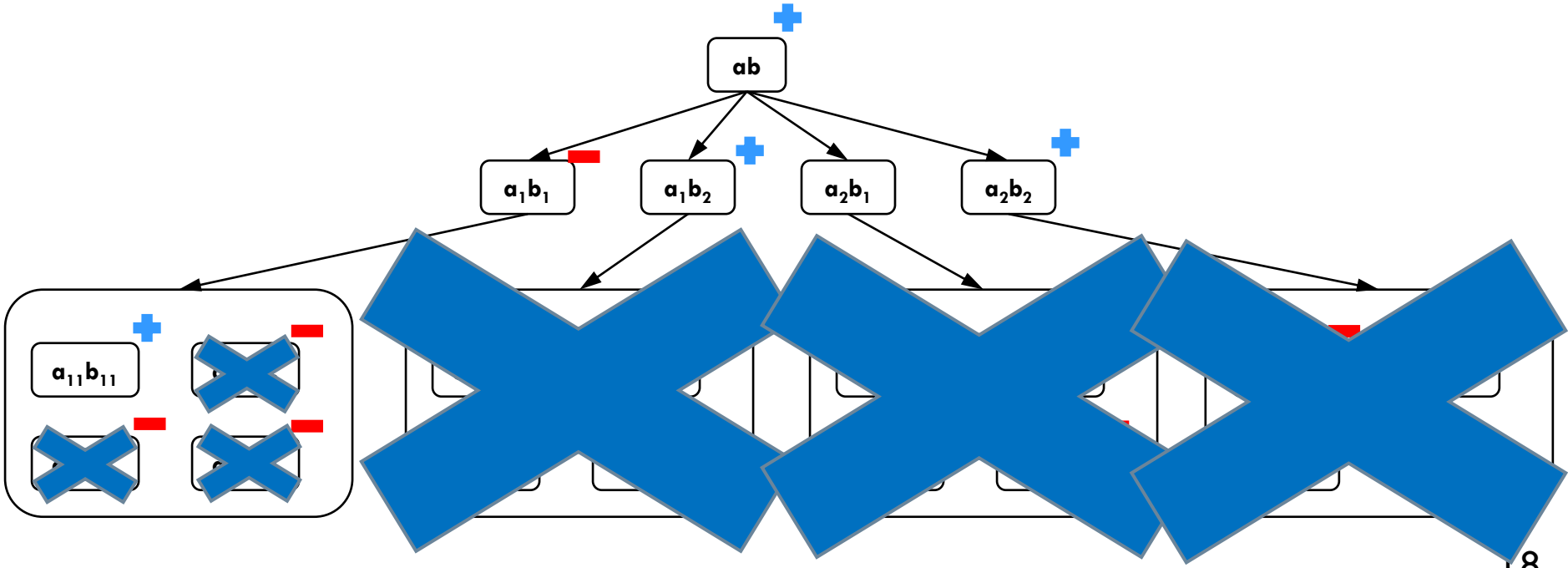
- Some (Cosine, Kulczynsky) are not anti-monotone
- We cannot extract flipping correlations by post-processing all positive and negative correlations, since we cannot compute all positive and negative correlations (see slide 8)
- Solution: incorporate flipping constraints into a mining process

Flipper algorithm: based on three main pruning techniques

1. Pruning non-flipping itemsets
2. Termination of the entire pattern growth
3. Pruning single items and their supersets

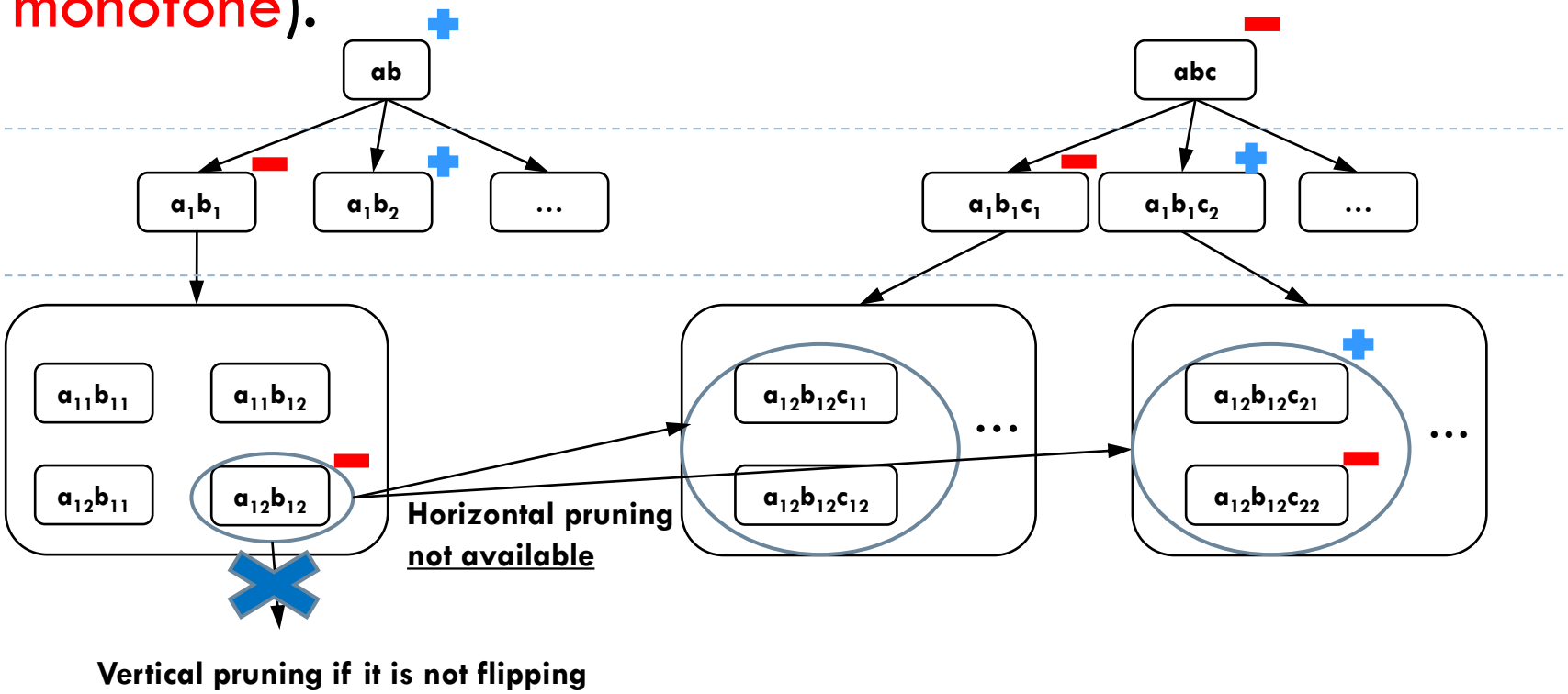
1. Pruning non-flipping patterns (I)

- If both parent itemset (ab) and child itemset (a_1b_2) have the same correlation sign, then they break a flipping sequence and the children of a_1b_2 cannot be a part of flipping pattern – do not test them



1. Pruning non-flipping patterns (II)

- However, a superset of child itemset ($a_{12}b_{12}$) can still be a part of a flipping pattern, since we cannot predict the correlation value of its superset (not anti-monotone).



2. Termination of the Entire Pattern Growth

- We prove that for any null-invariant correlation measure, correlation of the superset cannot be larger than the *max* of correlations of its subsets

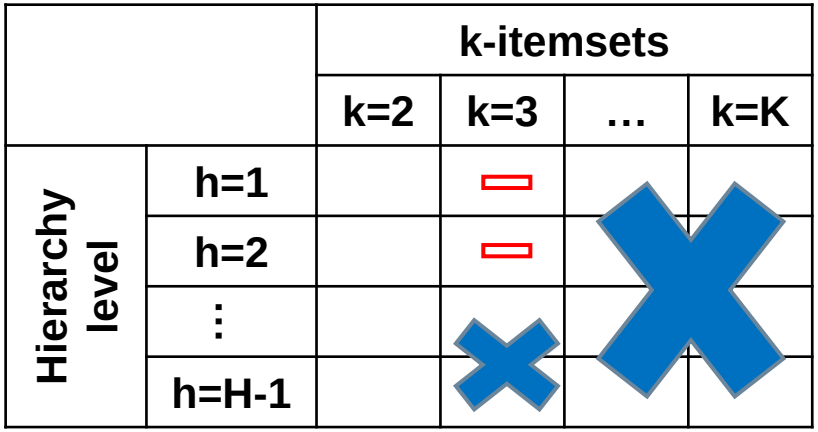
$$Corr(a_1, \cdots, a_{n+1}) \leq \max(Corr(a_1, \cdots, a_n), \cdots, Corr(a_2, \cdots, a_{n+1}))$$

2. Termination of the Entire Pattern Growth

$$Corr(a_1, \cdots, a_{n+1}) \leq \max(Corr(a_1, \cdots, a_n), \cdots, Corr(a_2, \cdots, a_{n+1}))$$

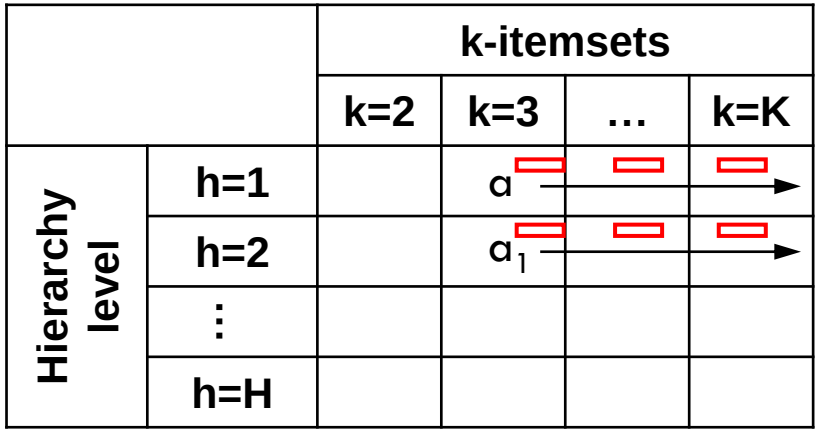
If we adding items to itemsets, and we found that all itemsets in two consecutive cells are non-positive, then there are no more flipping patterns because supersets cannot be positively correlated

We can stop our search right there



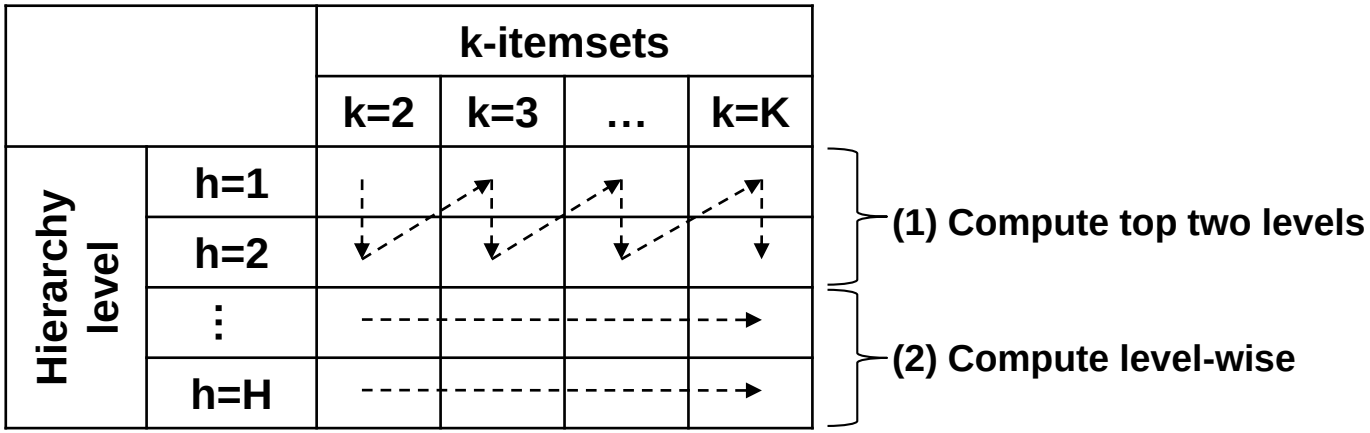
3. Pruning single items and their supersets

- If all itemsets containing item a_1 are non-positive, and all itemsets containing its generalization item a are non-positive, then item a_1 and all its supersets can be removed from further consideration



Order of computation

- To utilize these pruning principles, we need to always compare results for two vertically consecutive cells



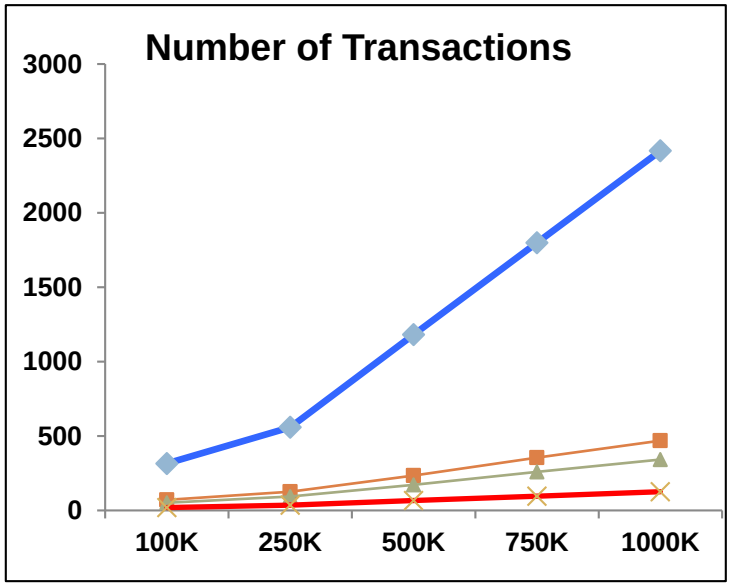
These are the main ideas of the *Flipper* algorithm

Outline

- Challenge: strong correlations with low support
- Flipping correlation pattern
- Algorithm for mining flipping correlations
- Performance
- Real flipping patterns
- Conclusion and future work

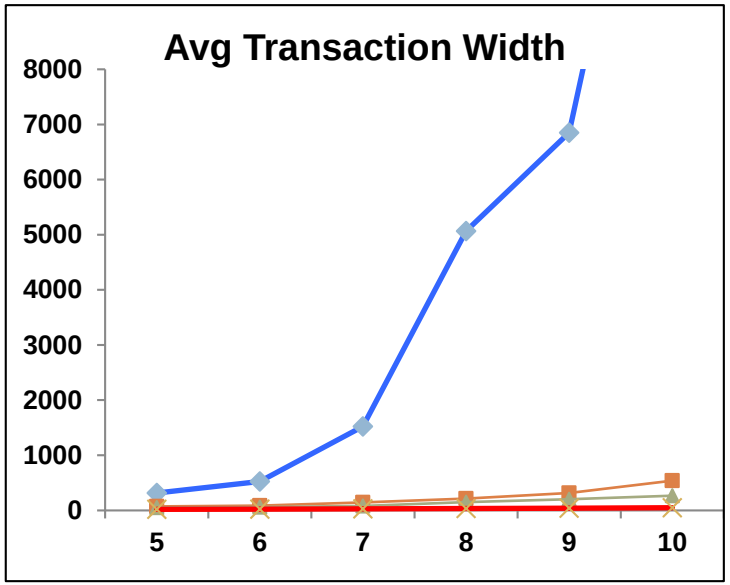
Performance: Synthetic datasets

- Running Time (sec)



Basic support-based

Full Flipper



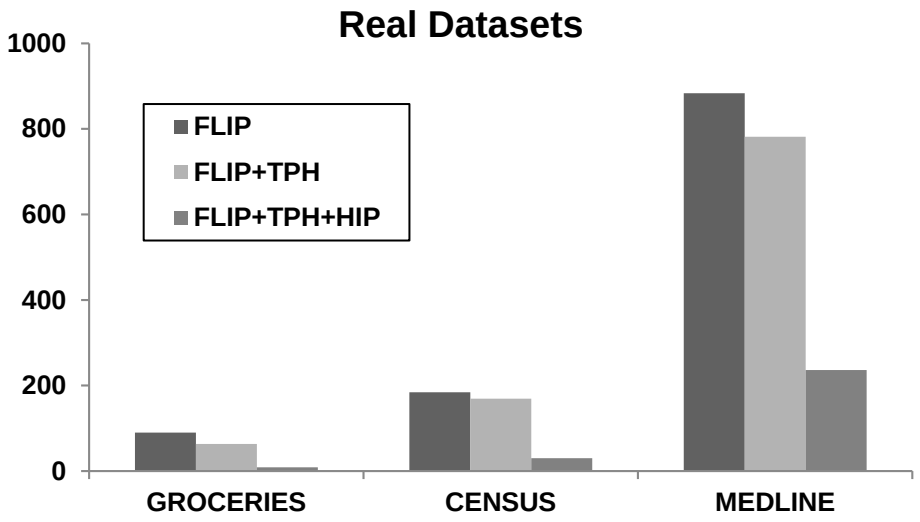
Flipper scales gracefully with the increase of the number of transactions and the average number of items per transaction

Performance: real datasets

- Data Sets

	# Trans	# Pos	# Neg	# Flips
GROCERIES	10K	4.8K	80K	174
CENSUS	32K	140K	73K	232
MEDLINE	6.4M	4.2K	1.6M	430

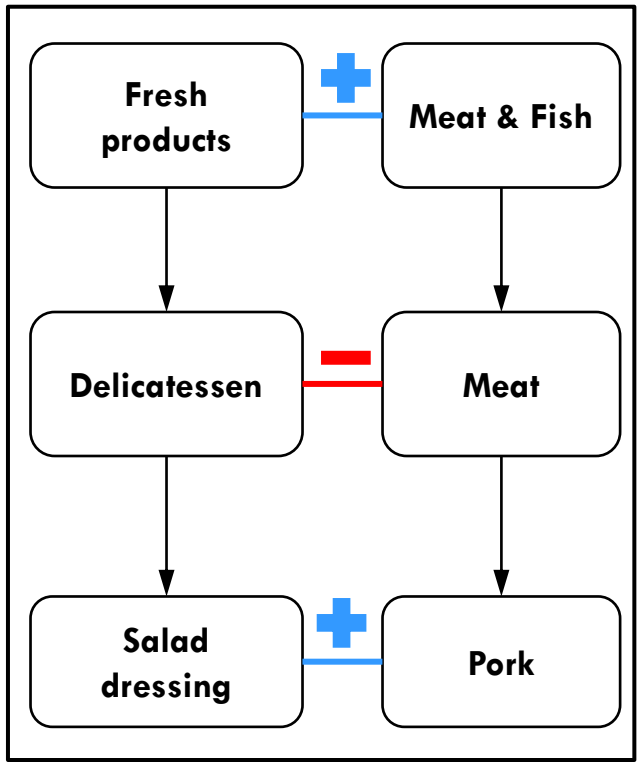
- Running Time (sec)
  - Basic is not included (ran more than 10 hours for the smallest dataset GROCERIES).



Outline

- Challenge: strong correlations with low support
- Flipping correlation pattern
- Algorithm for mining flipping correlations
- Performance
  - Real flipping patterns
  - Conclusion and future work

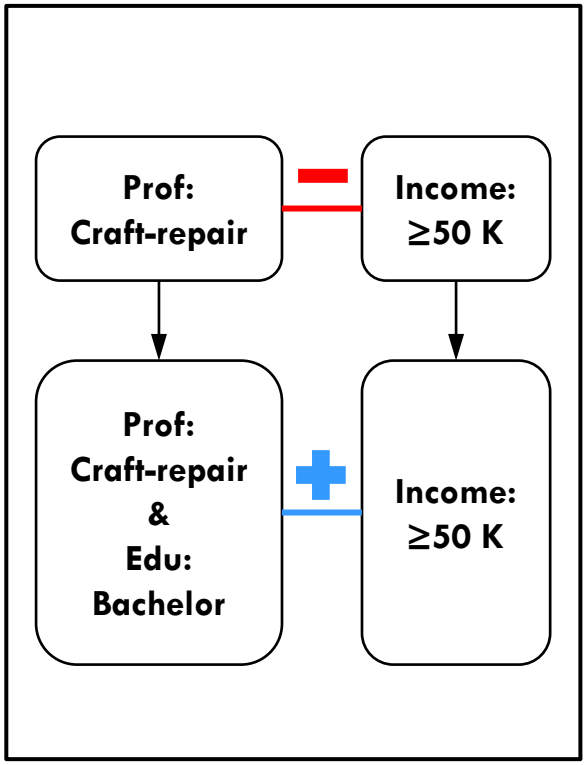
Flipping patterns: discover incorrectly classified items



Re-design store layouts

- pork and salad dressing are positively correlated, while in general meat and delicatessen are negatively correlated.
- This might suggest removing the salad dressing from delicatessen, and moving it closer to the meat department.

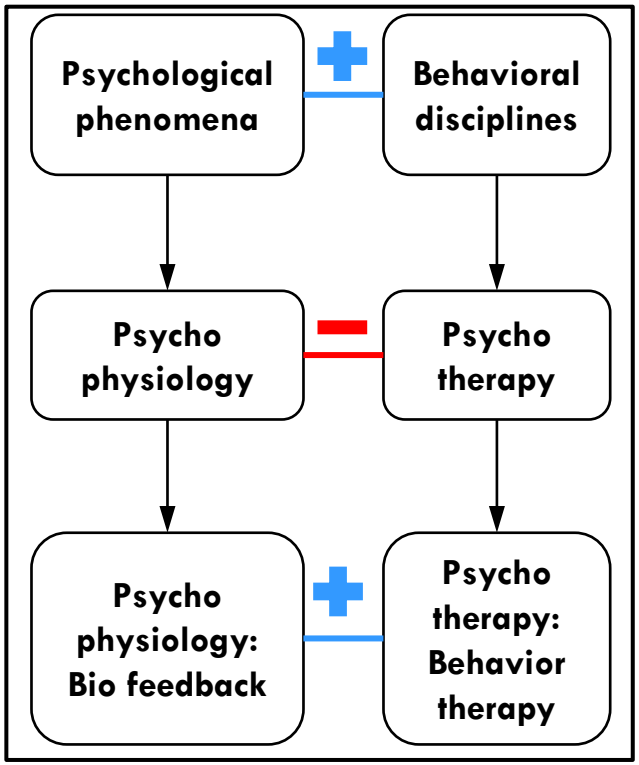
Flipping patterns: contrasting sub-populations



Discover sub-populations with a distinct behaviour

- People working in Craft-repair and having Bachelor degree are positively correlated with high income, unlike all people working in Craft-repair
- Education matters

Flipping patterns: under-represented item combinations

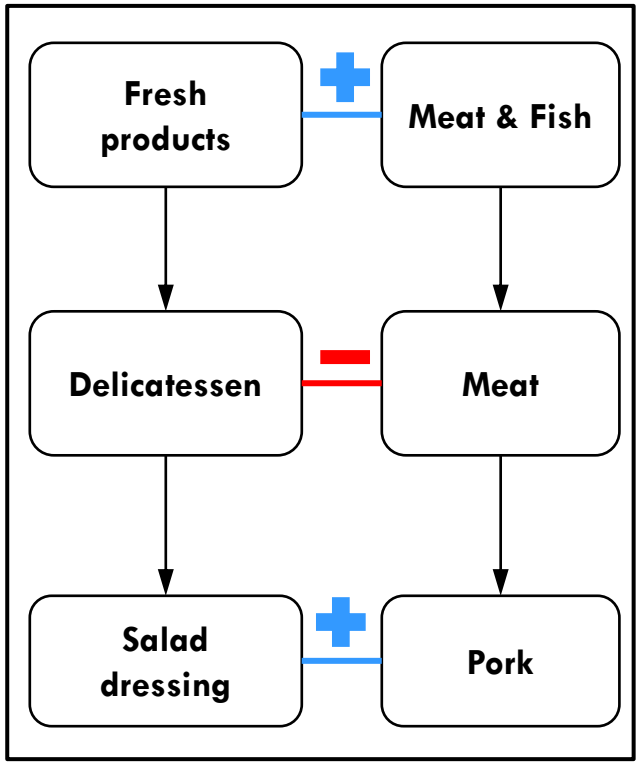


MEDLINE

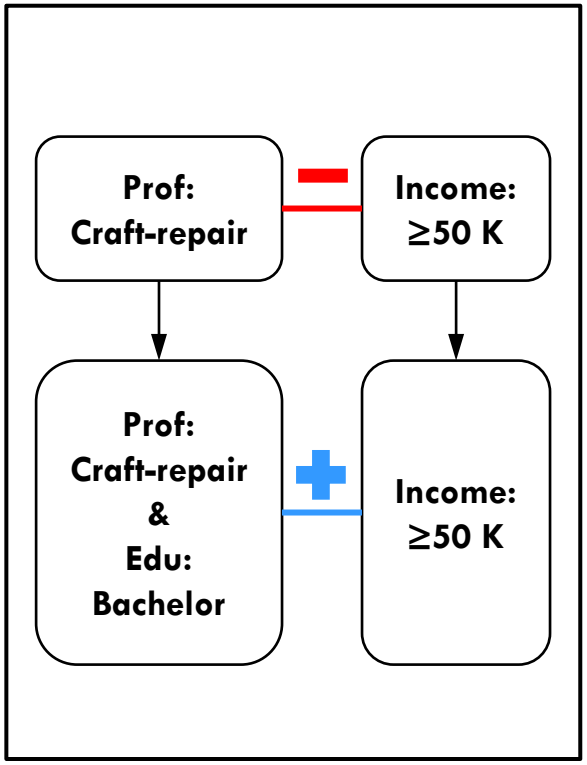
Suggest under-represented research topic combinations

- This pattern suggests the collaboration between two unrelated areas of psychophysiology and psychotherapy.
- However, if one decides to study the combination of such sub-topics as biofeedback and behavior therapy, he finds out that these two are in fact often studied together.

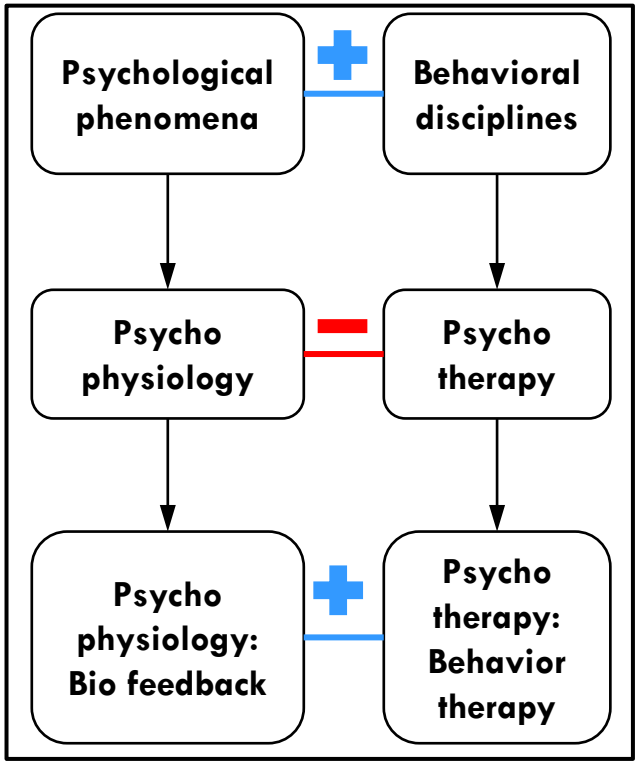
Flipping patterns in real datasets



GROCERIES



CENSUS



MEDLINE

Outline

- Challenge: strong correlations with low support
- Flipping correlation pattern
- Algorithm for mining flipping correlations
- Performance
- Real flipping patterns
- Conclusion and future work

Summary

- Introduced the notion of a *flipping correlation pattern*.
- Developed the *Flipper* algorithm for mining these patterns.
- Algorithm is based on flipping constraints and mathematical properties shared among all null-invariant correlation measures
- Demonstrated the high efficiency of Flipper in experiments with low support thresholds
- Have shown that interesting new patterns can be extracted using the flipping pattern concept.

Future work

- More advanced data structures for improving performance of Flipper
- Top-K “most flipping” patterns
- Computing a set of all discriminative correlations specific for a given subgroup

Thank you for listening

Please email your questions and suggestions to:
mgbarsky@gmail.com