November 17, 2017

Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

Supervised Pattern Mining

Data Mining 04 (データマイニング)

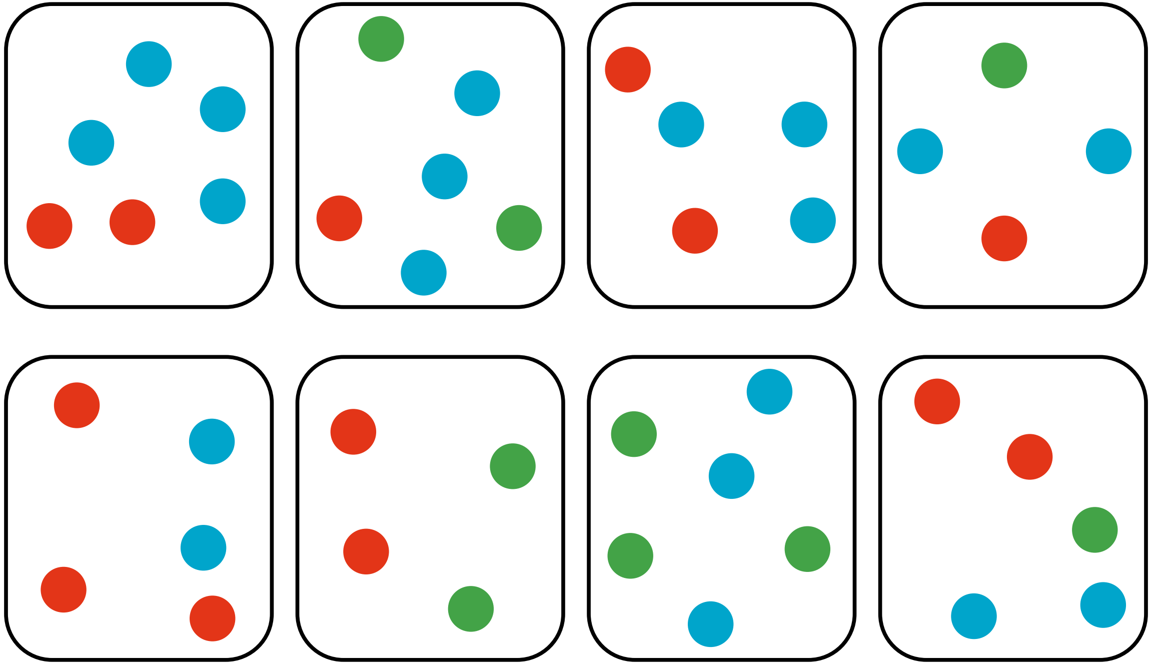
Mahito Sugiyama (杉山麿人)

Today's Outline

- Pattern mining with class labels (supervision)
  - Various measures
- Significant pattern mining
  - Statistical tests
  - Testable patterns
  - Controlling the FWER (Family-Wise Error Rate) by Tarone's testability trick

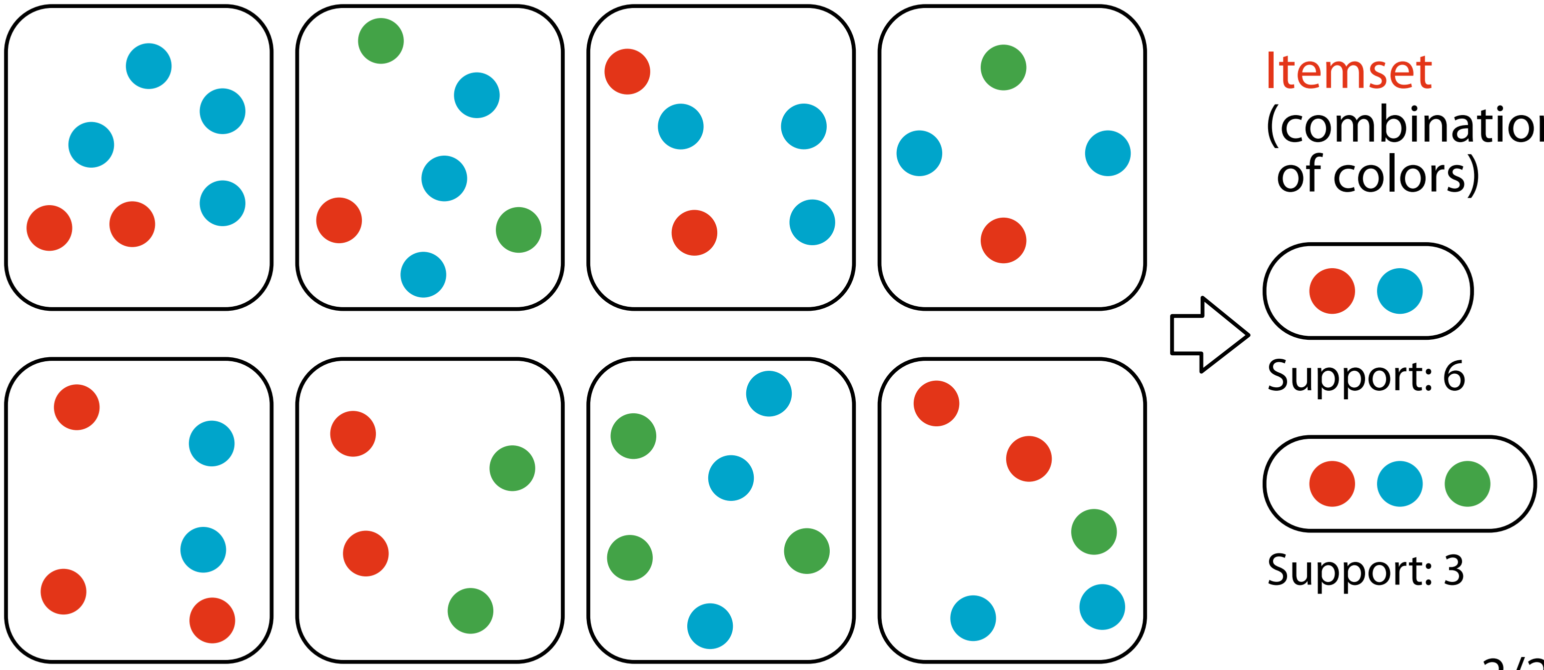
Itemset Mining

- Find interesting combinatorial patterns from massive data



Itemset Mining

- Find interesting combinatorial patterns from massive data

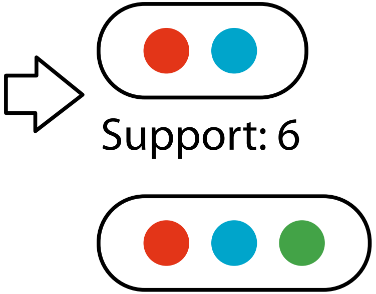


Itemset Mining (Binary Representation)

- Find interesting combinatorial patterns from massive data

ID1	1	1	0
ID2	1	1	1
ID3	1	1	0
ID4	1	1	1
ID5	1	1	0
ID6	0	1	1
ID7	1	0	1
ID8	1	1	1

Itemset
(combination of colors)



Support: 6

Support: 3

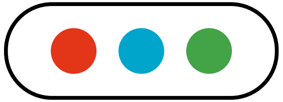
Itemset Mining (Binary Representation)

- Find interesting combinatorial patterns from massive data

ID1	1	1	0
ID2	1	1	1
ID3	1	1	0
ID4	1	1	1
ID5	1	1	0
ID6	0	1	1
ID7	1	0	1
ID8	1	1	1

Itemset
(combination of colors)

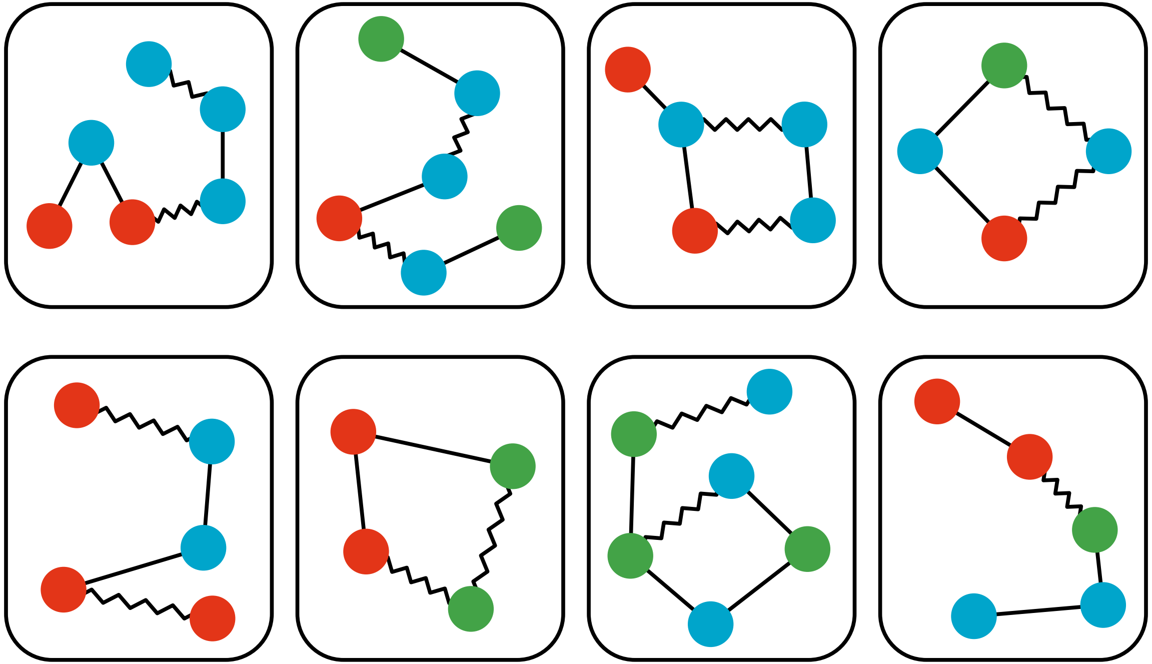
Support: 6



Support: 3

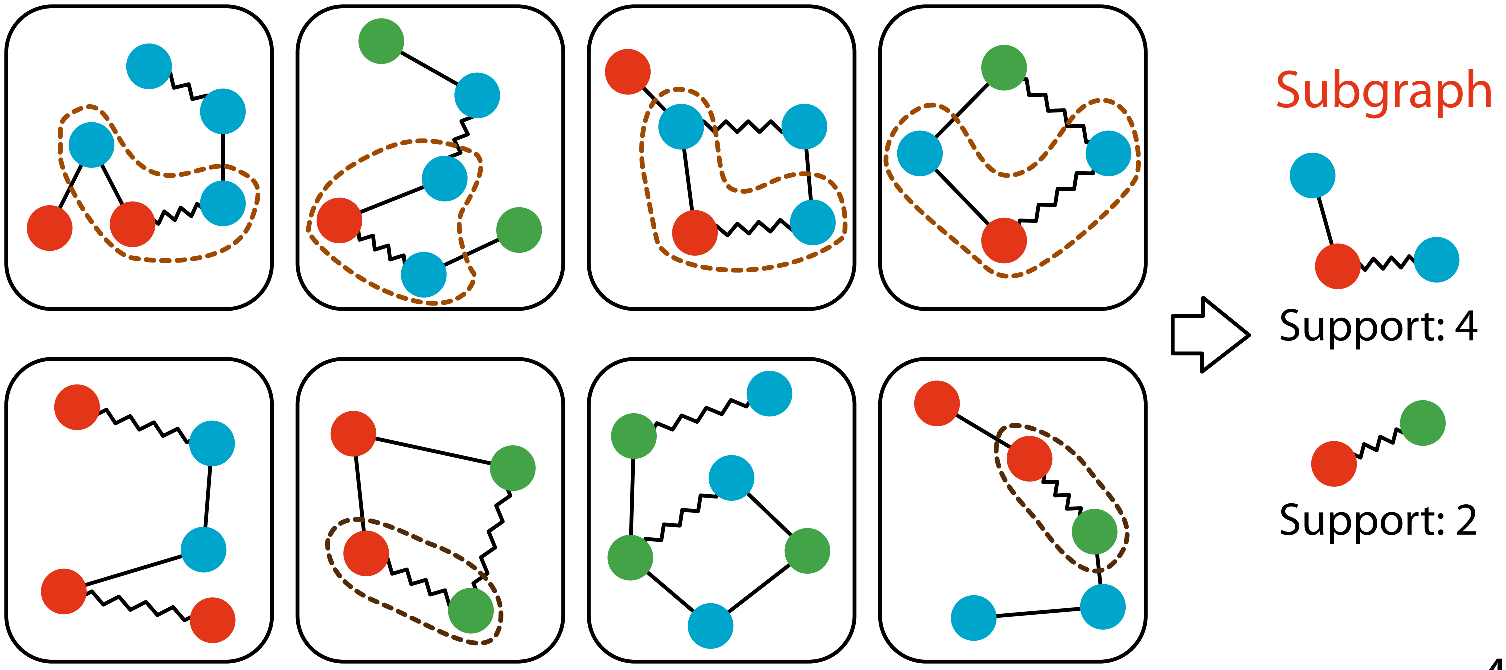
Subgraph Mining

- Find interesting combinatorial patterns from massive data



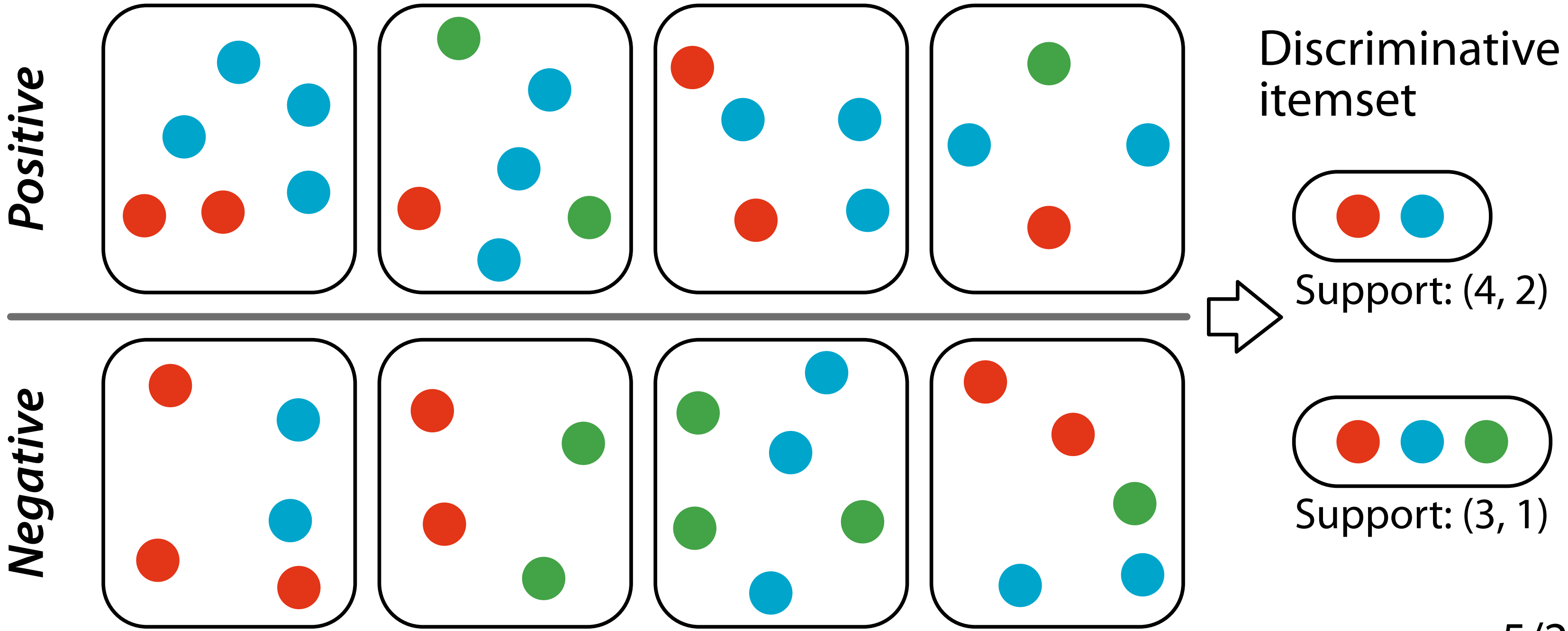
Subgraph Mining

- Find interesting combinatorial patterns from massive data



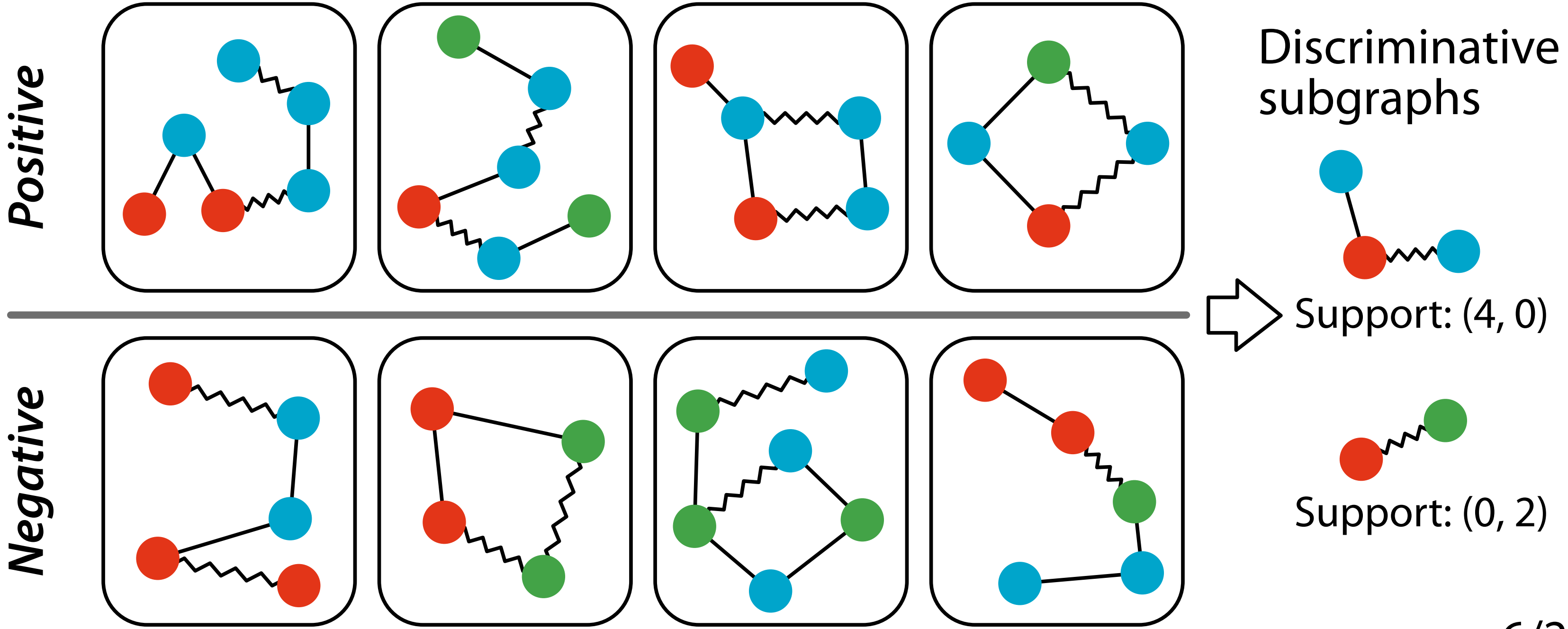
Supervised Itemset Mining

- Find discriminative patterns from supervised data



Supervised Subgraph Mining

- Find discriminative patterns from supervised data



Contingency Table

	Occurrence	Non-occurrence	Total
Positive	$\mathrm{supp}_C(x)$	$\lvert C\rvert - \mathrm{supp}_C(x)$	$\lvert C\rvert$
Negative	$\mathrm{supp}_{\bar{C}}(x)$	$\lvert\bar{C}\rvert - \mathrm{supp}_{\bar{C}}(x)$	$\lvert\bar{C}\rvert$
Total	$\mathrm{supp}(x)$ $= \mathrm{supp}_C(x) + \mathrm{supp}_C(x)$	$\lvert D\rvert - \mathrm{supp}(x)$	$\lvert D\rvert$

Contingency Table

	Occurrence	Non-occurrence	Total
Positive	n_{11}	n_{12}	c_1
Negative	n_{21}	n_{22}	c_2
Total	s	s'	d

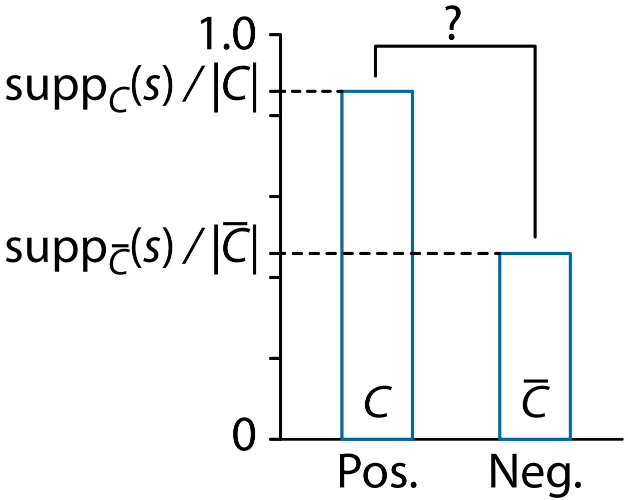
Various Measures

- Confidence: n_{11}/d
- Growth rate (relative risk): n_{11}/n_{21}
- Support difference (risk difference): $n_{11} - n_{21}$
- Mutual information:
$$\frac{n_{11}}{d}\log\frac{n_{11}/d}{c_1 s/d^2} + \frac{n_{12}}{d}\log\frac{n_{12}/d}{c_1 s'/d^2} + \frac{n_{21}}{d}\log\frac{n_{21}/d}{c_2 s/d^2} + \frac{n_{22}}{d}\log\frac{n_{22}/d}{c_2 s'/d^2}$$
- Subgroup discovery measure (weighted relative accuracy):
$(c_1/d)((n_{11}/c_1) - (c_1/d))$

Computing *p*-value of Pattern

- Given positive and negative sample sets $C, \bar{C}$ such that $D = C \cup \bar{C}$
- The *p*-value of each pattern s is assessed by the Fisher's exact test

	Occ.	Non-occ.	Total
C (Pos.)	$\mathrm{supp}_C(s)$	$\lvert C\rvert - \mathrm{supp}_C(s)$	$\lvert C\rvert$
$\bar{C}$ (Neg.)	$\mathrm{supp}_{\bar{C}}(s)$	$\lvert\bar{C}\rvert - \mathrm{supp}_{\bar{C}}(s)$	$\lvert\bar{C}\rvert$
D (Total)	$\mathrm{supp}(s)$	$\lvert D\rvert - \mathrm{supp}(s)$	$\lvert D\rvert$

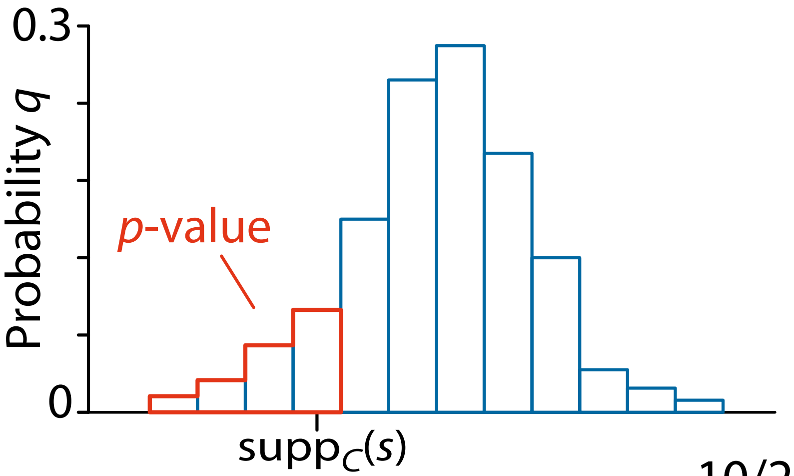


Fisher's Exact Test

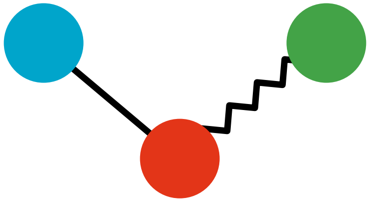
- Probability $q(\mathrm{supp}_C(s))$ is given by hypergeometric distribution:

$$q(\mathrm{supp}_C(s)) = \binom{|C|}{\mathrm{supp}_C(s)}\binom{|\bar{C}|}{\mathrm{supp}_{\bar{C}}(s)}\bigg/\binom{|D|}{\mathrm{supp}(s)}$$

	Occ.	Non-occ.	Total
C (Pos.)	$\mathrm{supp}_C(s)$	$\lvert C\rvert - \mathrm{supp}_C(s)$	$\lvert C\rvert$
$\bar{C}$ (Neg.)	$\mathrm{supp}_{\bar{C}}(s)$	$\lvert \bar{C}\rvert - \mathrm{supp}_{\bar{C}}(s)$	$\lvert \bar{C}\rvert$
D (Total)	$\mathrm{supp}(s)$	$\lvert D\rvert - \mathrm{supp}(s)$	$\lvert D\rvert$



Hypothesis Test for Each Pattern



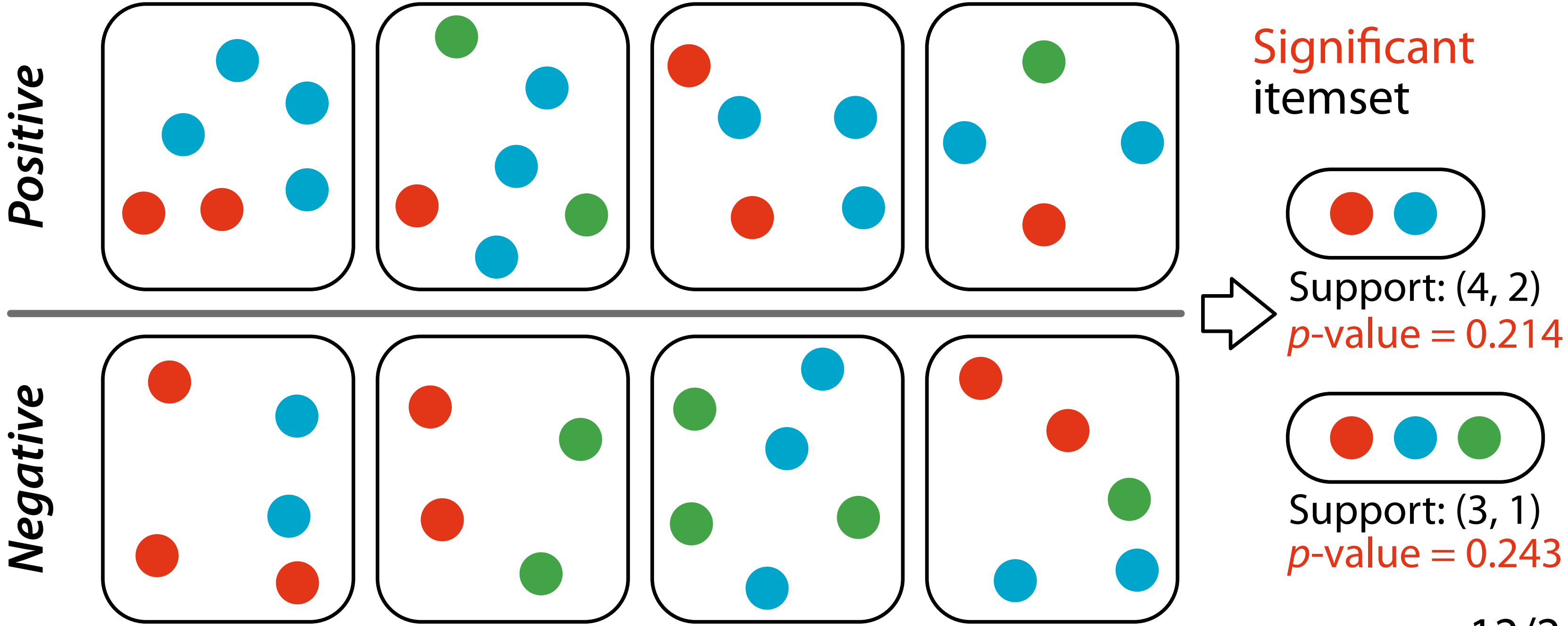
	Alternative hypothesis is true	Null hypothesis is true
Declared significant (p-value $< \alpha$)	True Positive	False Positive (Type I Error)
Declared non-significant	False Negative (Type II Error)	True Negative

Null: Occurence of pattern is independent from classes

Alternative: Occurence of pattern is associated with classes

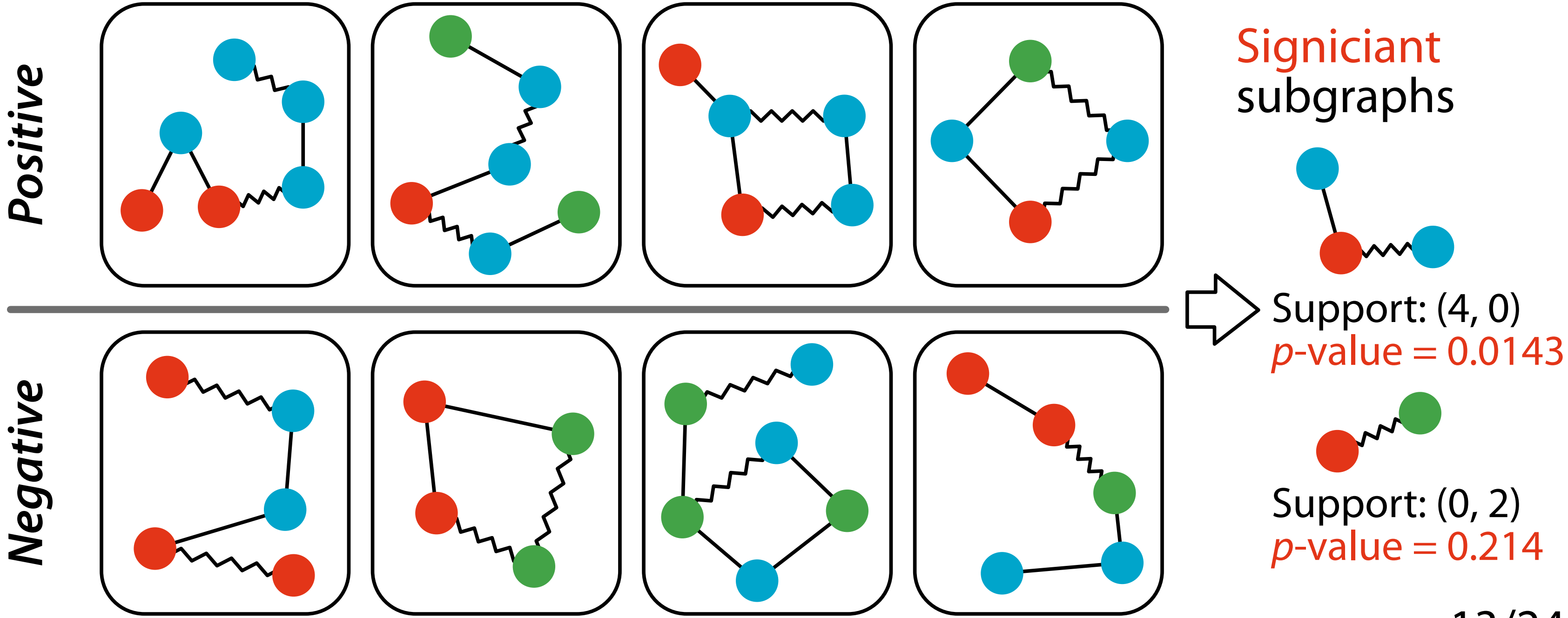
Significant Pattern (Itemset) Mining

- Find discriminative patterns from supervised data



Significant Subgraph Mining

- Find discriminative patterns from supervised data



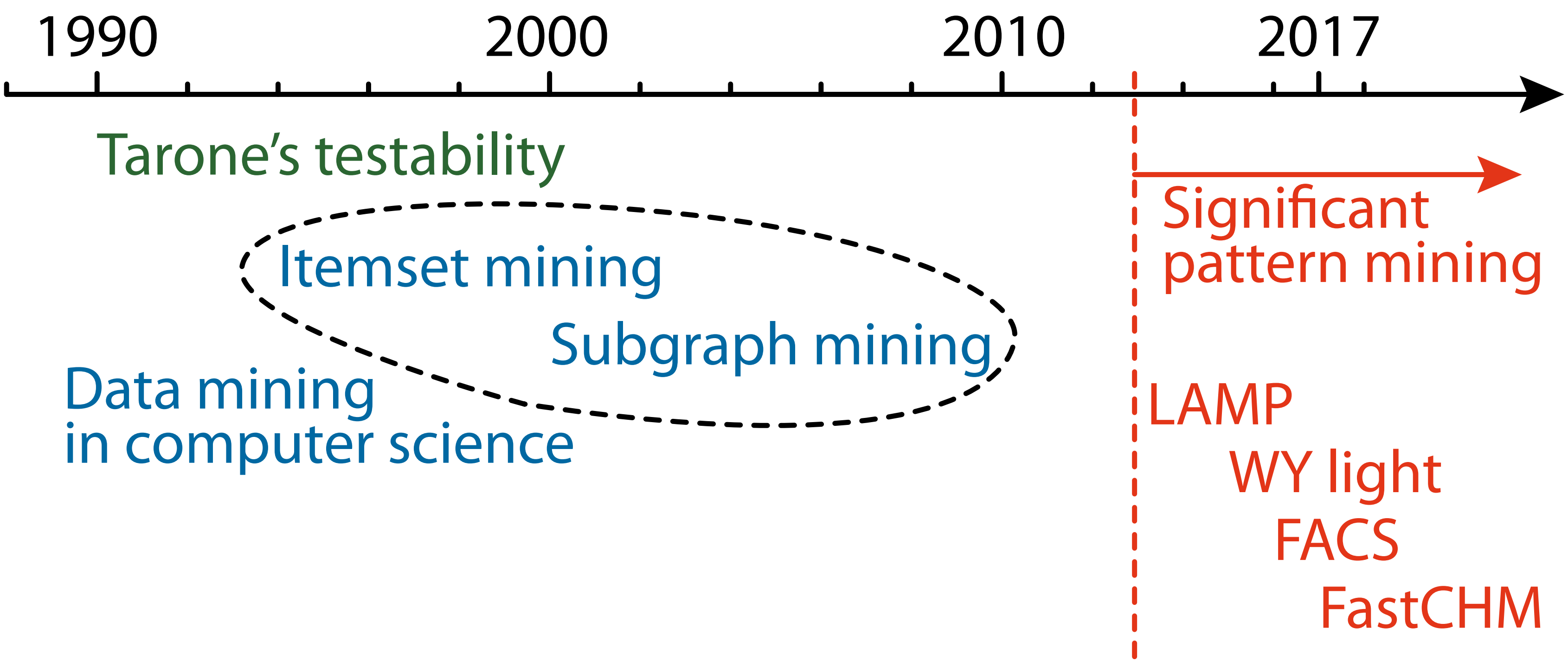
Challenges and Solutions of SPM

1. (Computational) How to check all patterns with avoiding combinatorial explosion?

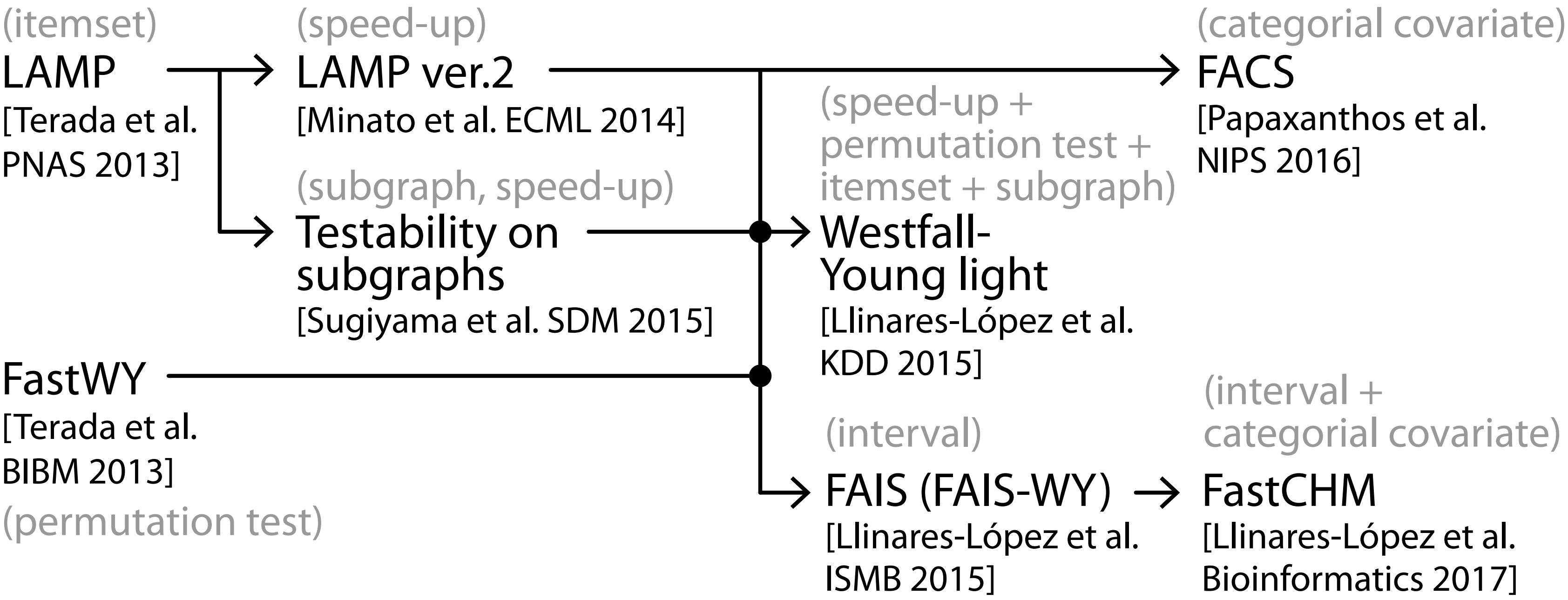
2. (Statistical) How to measure the statistical association (i.e. *p*-value) with correcting for multiple testing with avoiding combinatorial explosion?

- **Answer: Tarone's trick + Apriori principle**
  - The Tarone's trick to define patterns that are irrelevant
  - The Apriori principle to efficiently prune such patterns using the partial order structure of patterns

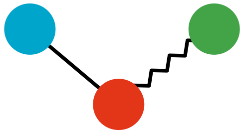
Timeline


1990
2000
2010
2017
Tarone's testability
Itemset mining
Subgraph mining
Data mining
in computer science
Significant
pattern mining
LAMP
WY light
FACS
FastCHM

Summary of SPM Methods



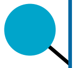
Multiple Testing



	Occ.	Non-occ.	Total
Positive	4	0	4
Negative	2	2	4
Total	6	2	8

Fisher's exact test: p-value $= 0.429$

Multiple Testing



	Occ.	Non-occ.	Total
Positive	3	1	4
Negative	1	3	4
Total	4	4	8

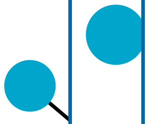
Fisher’s exact test: p-value = 0.486

Fisher’s exact test: p-value = 0.429

Multiple Testing

	Occ.	Non-occ.	Total
Positive	4	0	4
Negative	3	1	4
Total	7	4	8

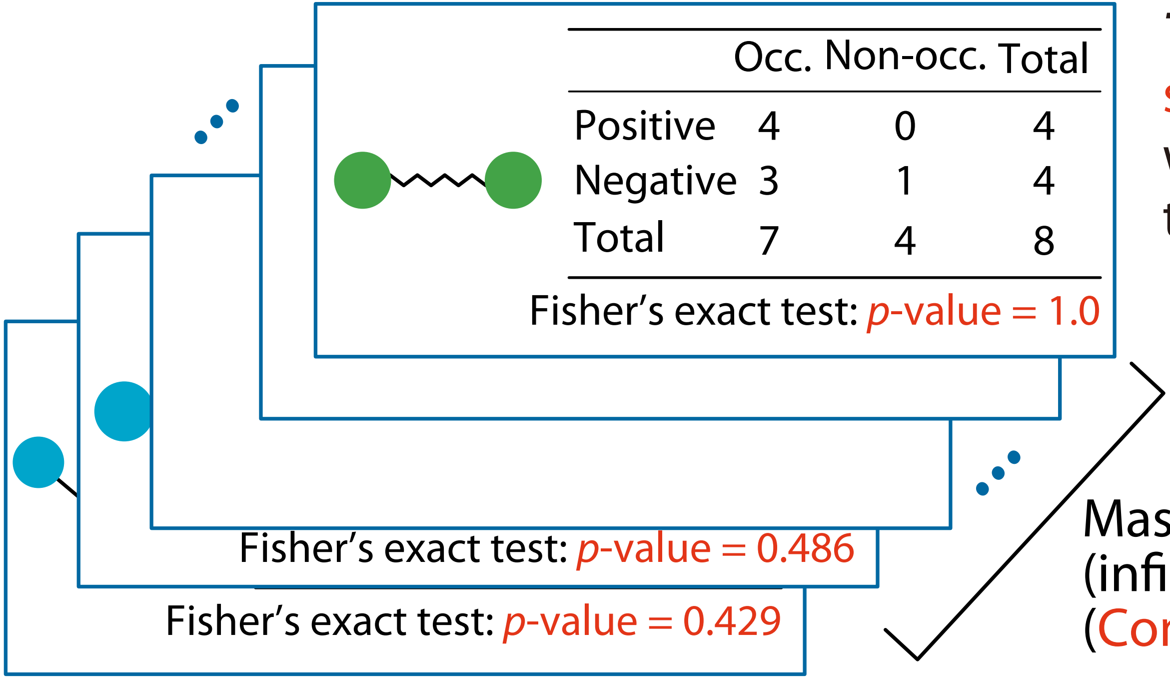
Fisher's exact test: *p*-value = 1.0



Fisher's exact test: *p*-value = 0.486

Fisher's exact test: *p*-value = 0.429

Multiple Testing



	Occ.	Non-occ.	Total
Positive	4	0	4
Negative	3	1	4
Total	7	4	8

Fisher's exact test: p-value = 1.0

Fisher's exact test: p-value = 0.486

Fisher's exact test: p-value = 0.429

Task: Enumerate all significant patterns while controlling the FWER

Massive number of (infinitely many) patterns (Combinatorial explosion!)

Multiple Testing Correction

- In each test, [probability of having a false positive] ≤ α
- If we repeat m tests, αm patterns can be false positives
  - Too many if m is large! For example in itemset mining:
    - For 100000 items, #patterns = 2^{100000}
    - Set significance level $\alpha = 0.01$
    - Number of false positives: $0.01 \cdot 2^{100000} = 10^{30101}$

Multiple Testing Correction

- In each test, [probability of having a false positive] ≤ α
- If we repeat m tests, αm patterns can be false positives
  - Too many if m is large! For example in itemset mining:
    - For 100000 items, #patterns = 2^{100000}
    - Set significance level $\alpha = 0.01$
    - Number of false positives: $0.01 \cdot 2^{100000} = 10^{30101}$
- FWER (family-wise error rate): ***Probability of having more than one false positives among all patterns***
  - FWER = $1 - (1 - \alpha)^m$ if patterns are independent

Controlling the FWER

- FWER = Pr(FP > 0)
  - FP: Number of false positives
- To achieve FWER = α, change the significance level for each pattern from α to δ ($\delta \leq \alpha$), the corrected significance level

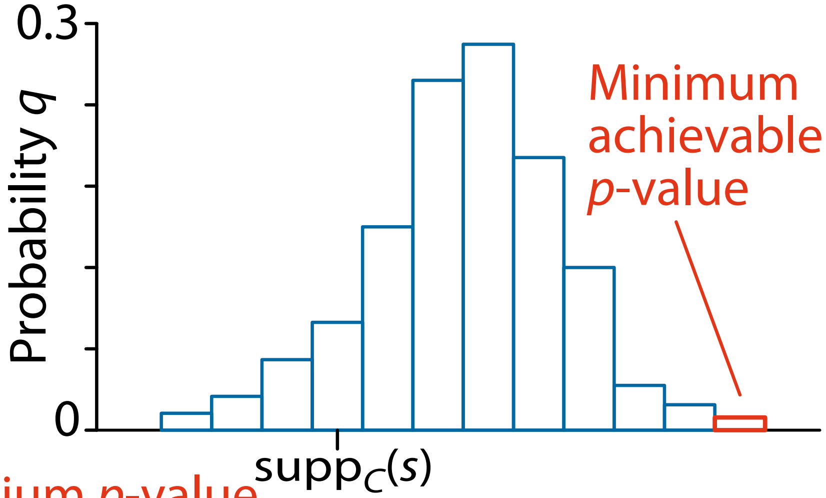
Controlling the FWER

- FWER = Pr(FP > 0)
  - FP: Number of false positives
- To achieve FWER = α, change the significance level for each pattern from α to δ ($\delta \le \alpha$), the corrected significance level
- **Objective:** Maximize FWER(δ) subject to FWER(δ) $\le \alpha$
  - FWER(δ): FWER at corrected significance level δ
    - Cannot be evaluated in closed form (simple but not easy!)
  - Bonferroni correction is popular: $\delta^*_{\mathrm{Bon}} = \alpha / m$

Minimum Achievable p-value $\Psi(\sigma)$

- Consider the minimum achievable p-value $\Psi(s)$ of a pattern s for its support $\mathrm{supp}(s)$

	Occ.	Non-occ.	Total
C (Pos.)	$\mathrm{supp}(s)$	$\lvert C\rvert - \mathrm{supp}_C(s)$	$\lvert C\rvert$
$\bar{C}$ (Neg.)	0	$\lvert\bar{C}\rvert - \mathrm{supp}_{\bar{C}}(s)$	$\lvert\bar{C}\rvert$
D (Total)	$\mathrm{supp}(s)$	$\lvert D\rvert - \mathrm{supp}(s)$	$\lvert D\rvert$



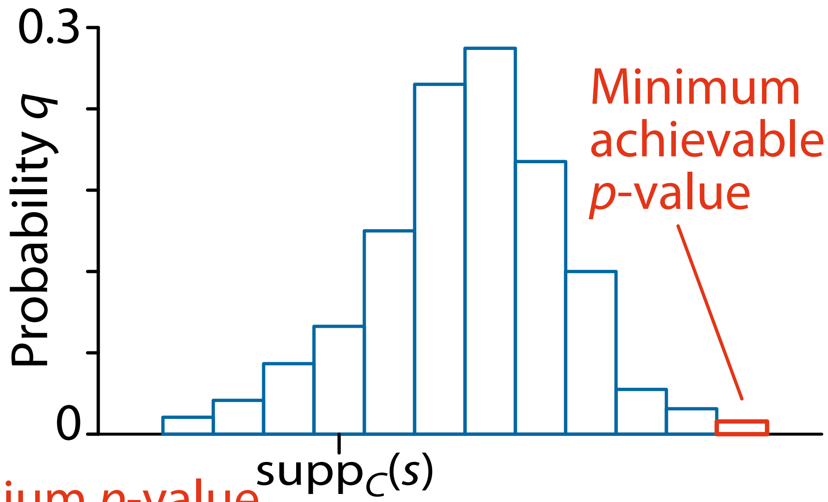
Most biased case that achieves the minium p-value

Computing $\Psi(\sigma)$

Minimum achievable p-value $\Psi(s) = \binom{|C|}{\mathrm{supp}(s)} \Big/ \binom{|D|}{\mathrm{supp}(s)}$

	Occ.	Non-occ.	Total
C (Pos.)	$\mathrm{supp}(s)$	$\lvert C\rvert - \mathrm{supp}_C(s)$	$\lvert C\rvert$
$\bar{C}$ (Neg.)	0	$\lvert \bar{C}\rvert - \mathrm{supp}_{\bar{C}}(s)$	$\lvert \bar{C}\rvert$
D (Total)	$\mathrm{supp}(s)$	$\lvert D\rvert - \mathrm{supp}(s)$	$\lvert D\rvert$

Most biased case that achieves the minium p-value



Tarone's Testability Trick

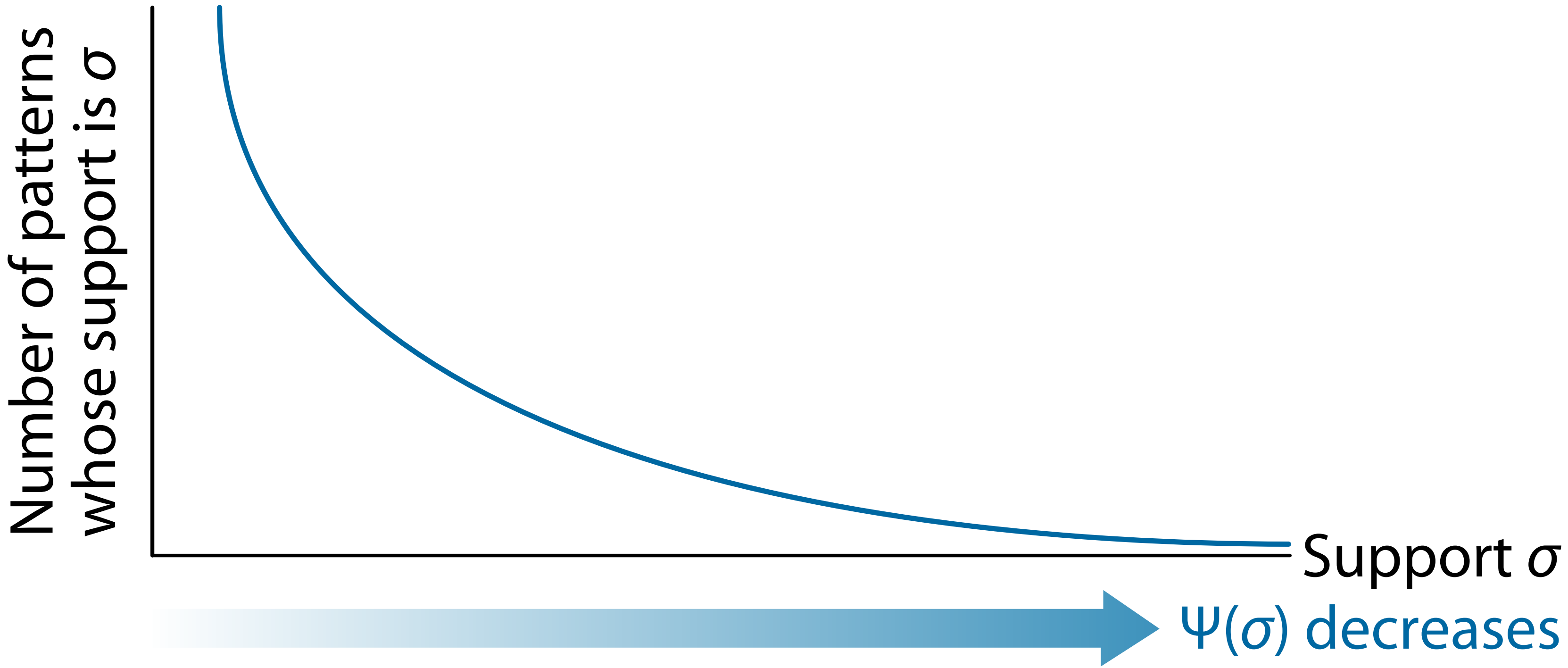
Minimum achievable p-value $\Psi(s) = \binom{|C|}{\mathrm{supp}(s)} \Big/ \binom{|D|}{\mathrm{supp}(s)}$

- Tarone (1990) pointed out (and Terada et al. (2013) revisited):

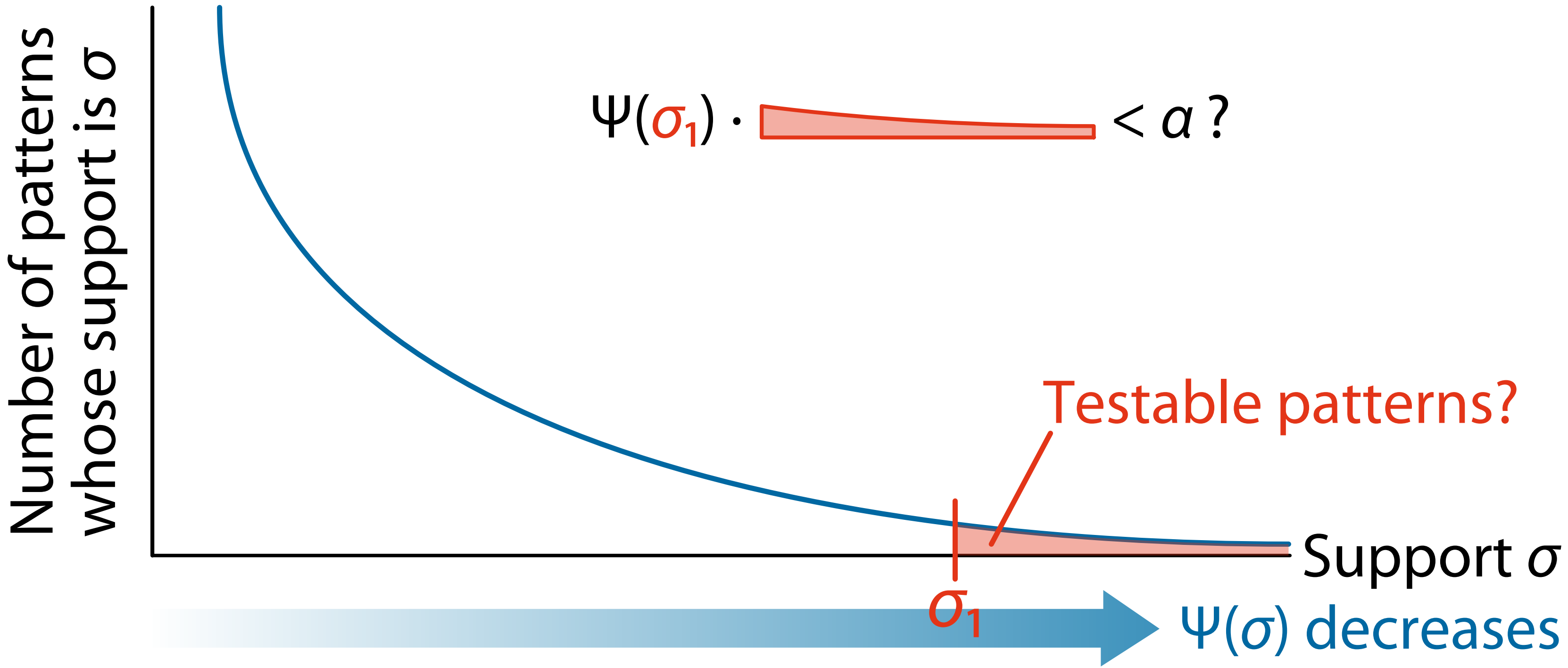
  For a pattern s with its support $\mathrm{supp}(s)$*, if the minimum achievable p-value* $\Psi(s)$ *is larger than the significance threshold, this is untestable and we can ignore it*

  - Significance threshold = α / [# testable subgraphs]
  - Untestable subgraphs can never be significant

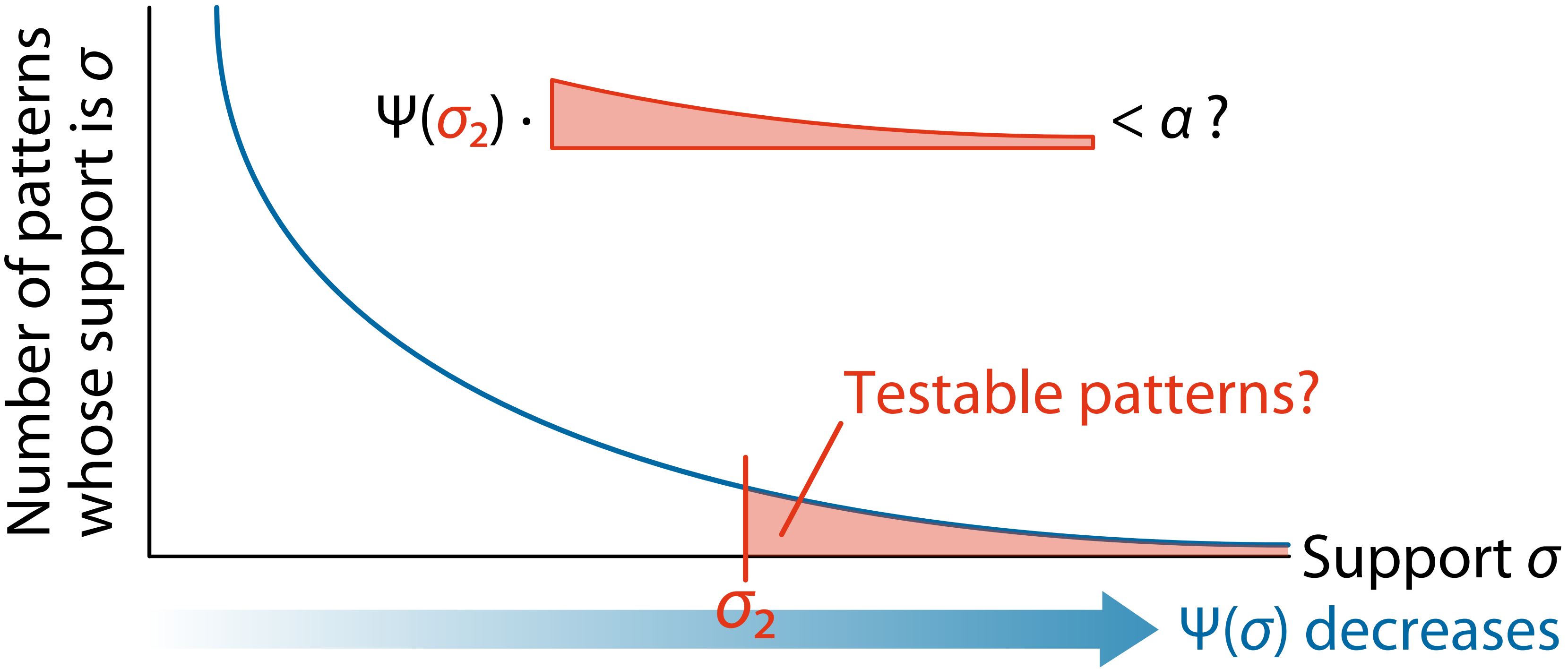
Finding Testable Patterns



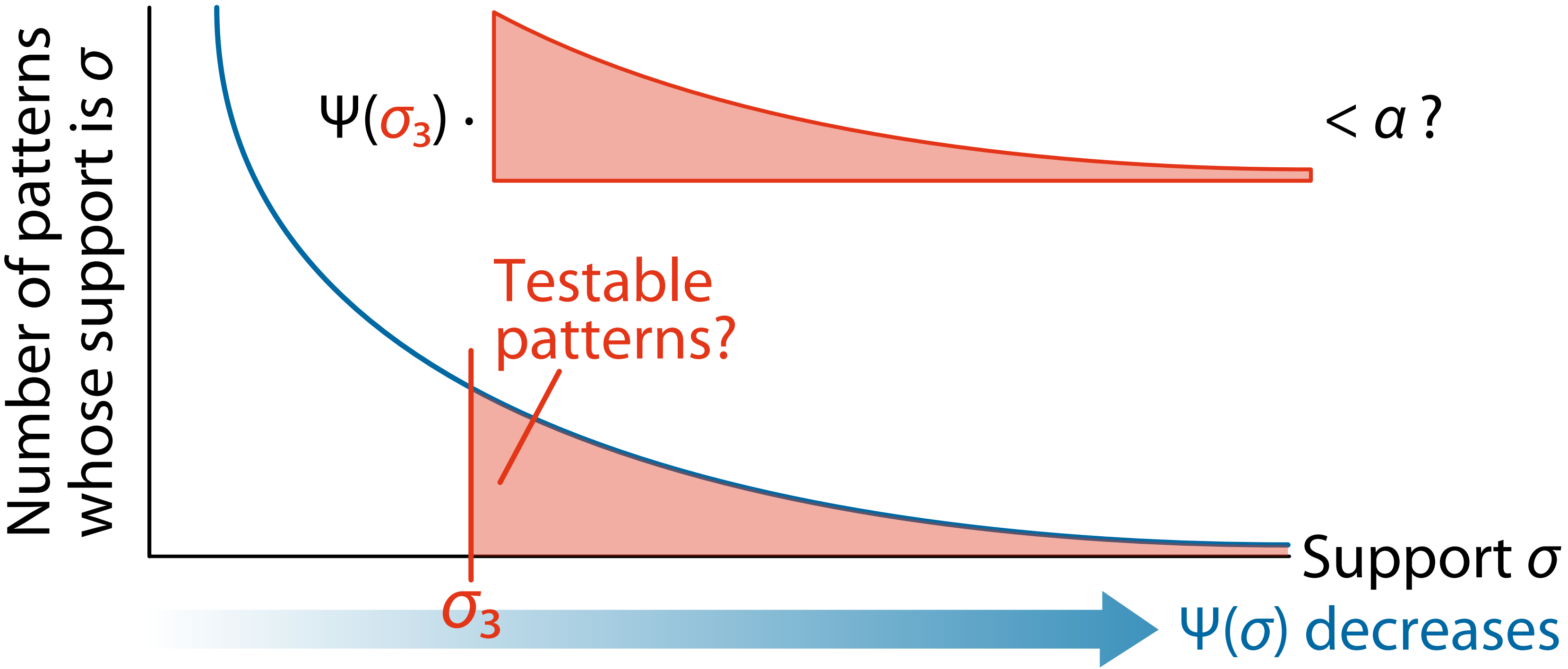
Finding Testable Patterns



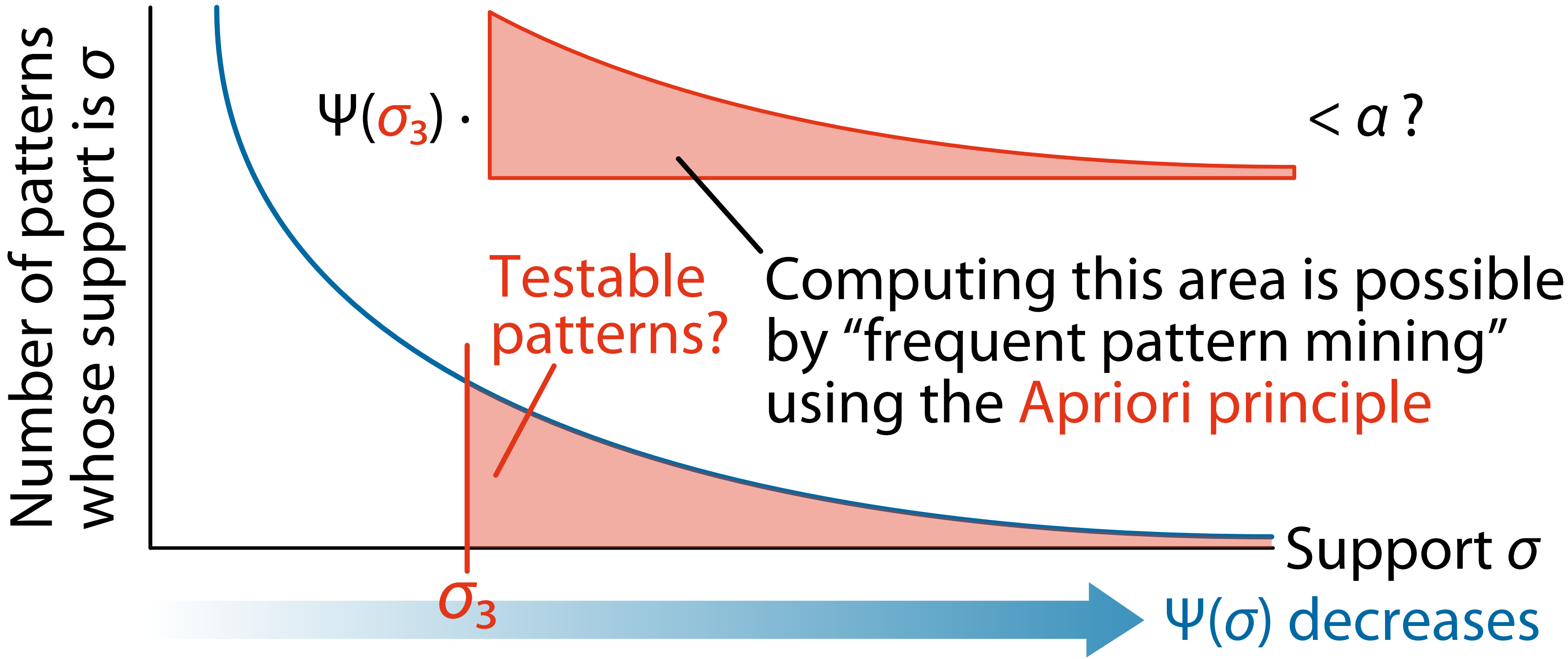
Finding Testable Patterns



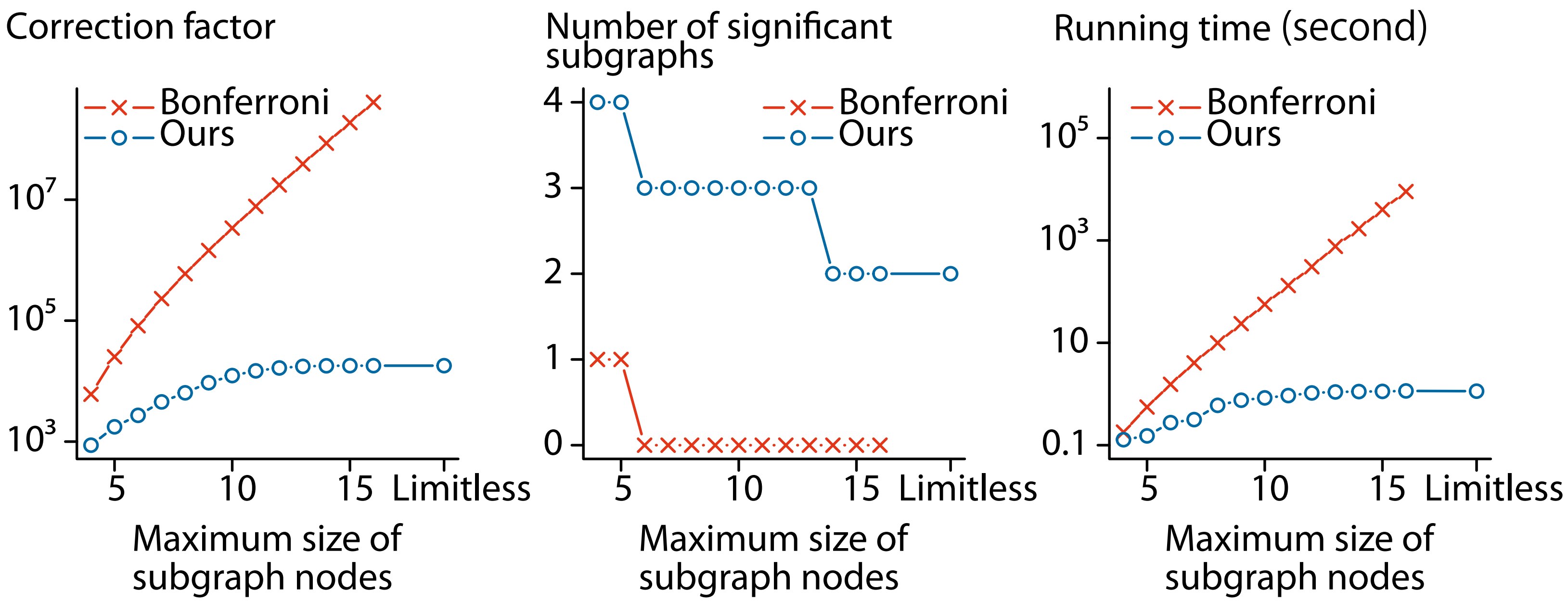
Finding Testable Patterns



Finding Testable Patterns



Power of Testability



The PTC (Predictive Toxicology Challenge) dataset with 601 chemical compounds

Conclusion

- Significant pattern mining is introduced
  - Find statistically significant subgraphs while controlling the FWER
  - pattern mining (data mining) + multiple testing correction (statistics)
- Open problems: How to treat continuous data?
  - Continuous features → mostly solved
    M. Sugiyama and K. Borgwardt:
    Finding Significant Combinations of Continuous Features,
    arXiv:1702.08694
  - Continuous response values → not solved yet