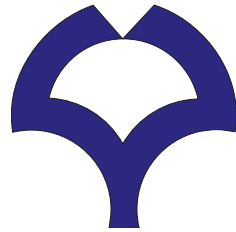


September 17, 2015

FIT 2015



統計的パターンマイニング

大阪大学 産業科学研究所

さきがけ研究者

杉山 磨人 Mahito Sugiyama

Outline

- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

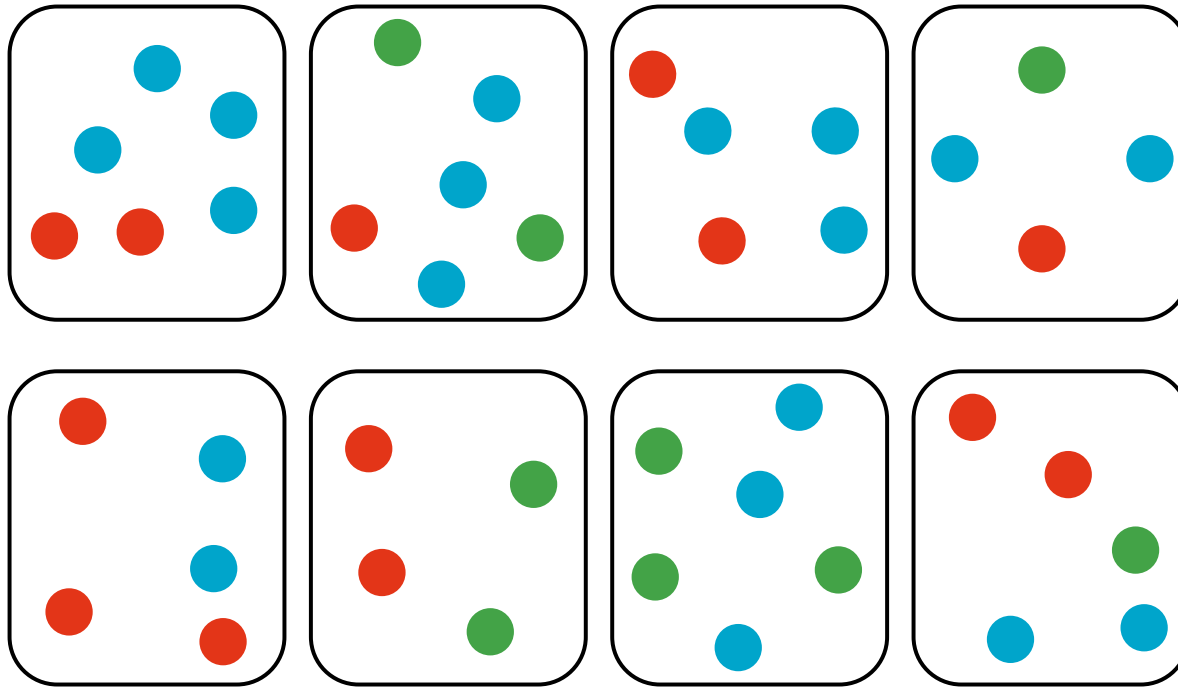
Outline

- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

Itemset Mining

- Find interesting **combinatorial patterns** from massive data

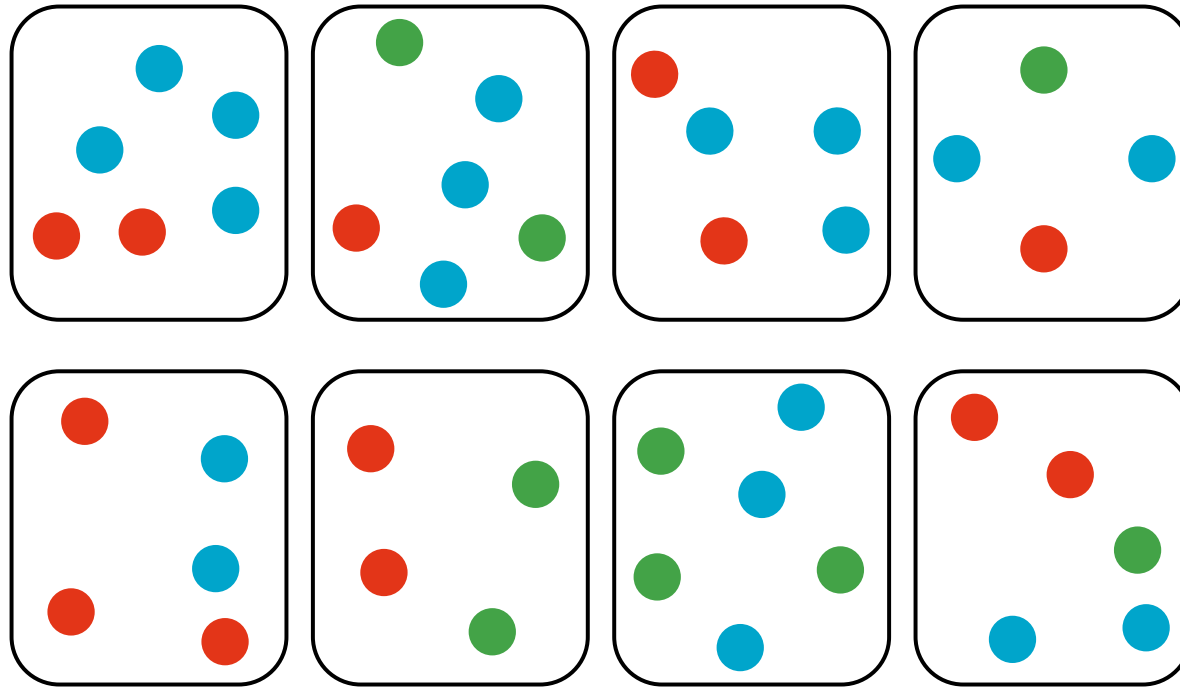
Database



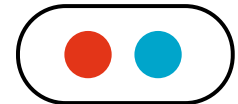
Itemset Mining

- Find interesting **combinatorial patterns** from massive data

Database



Itemset
(combination
of colors)



Support: 6

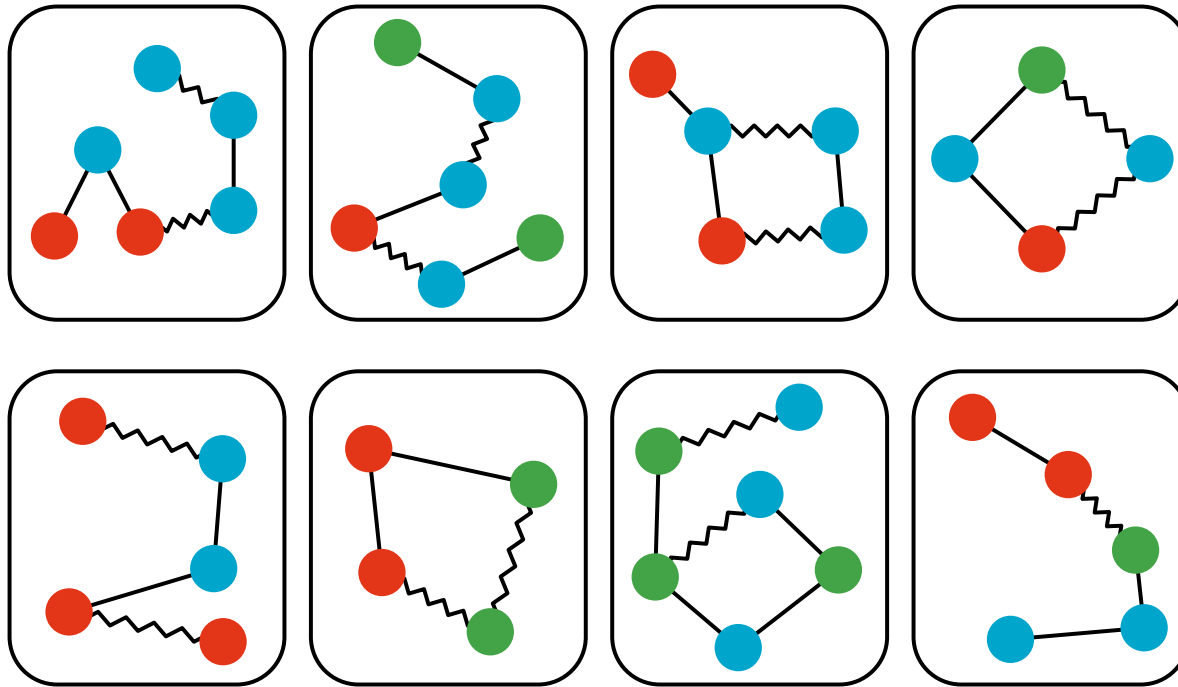


Support: 3

Subgraph Mining

- Find interesting **combinatorial patterns** from massive data

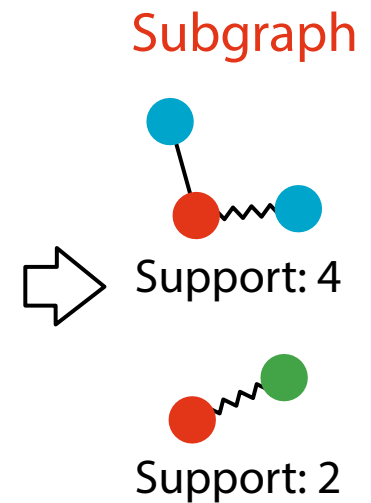
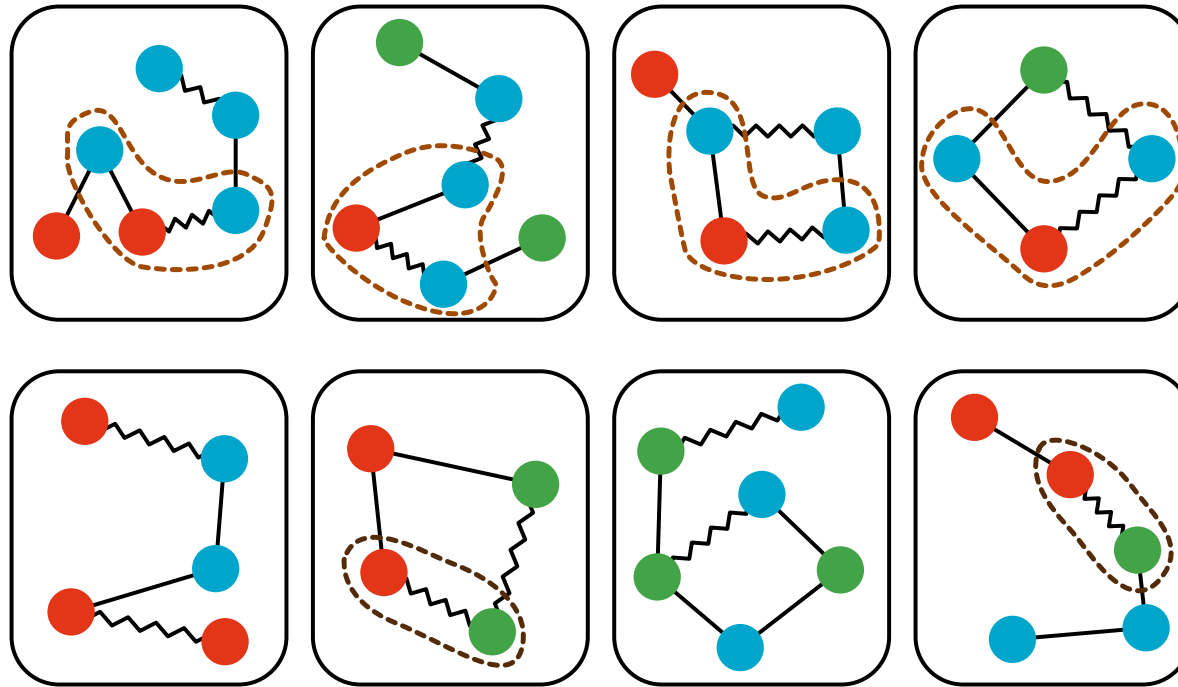
Database



Subgraph Mining

- Find interesting **combinatorial patterns** from massive data

Database



Apriori on Itemset Lattice

Transaction
database

ID 1: ● ●

ID 2: ● ● ●

ID 3: ● ●

ID 4: ● ● ●

ID 5: ● ●

ID 6: ● ●

ID 7: ● ●

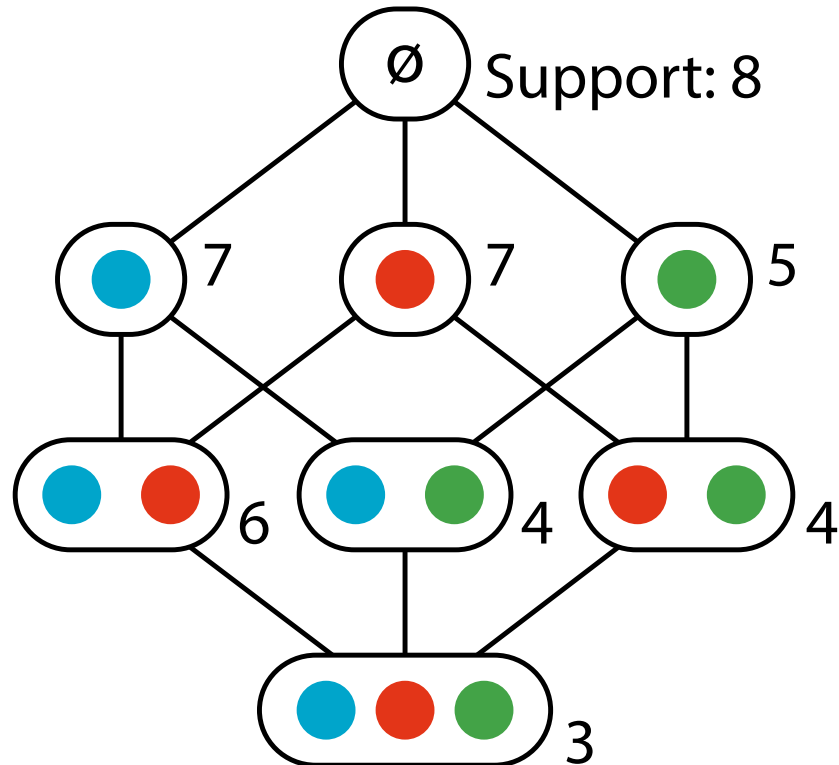
ID 8: ● ● ●

Apriori on Itemset Lattice

Transaction
database

ID 1: ● ●
ID 2: ● ● ●
ID 3: ● ●
ID 4: ● ● ●
ID 5: ● ●
ID 6: ● ●
ID 7: ● ●
ID 8: ● ● ●

Itemset lattice



Apriori principle:

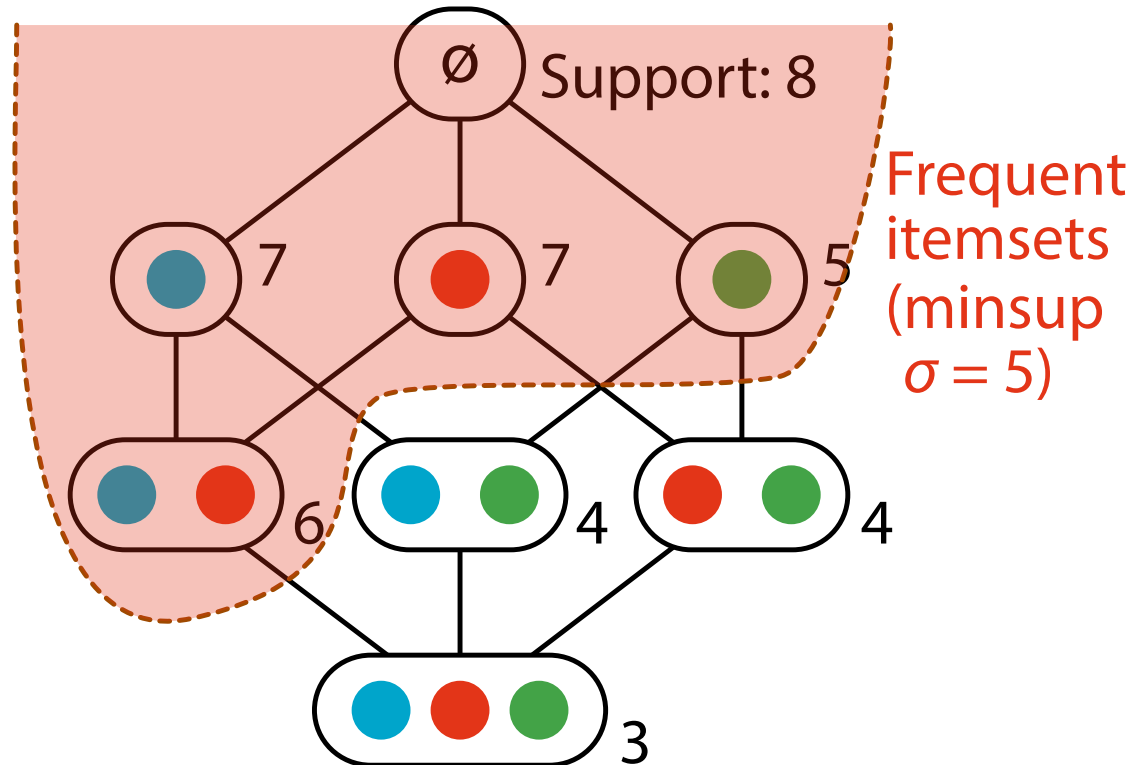
Support is monotonically decreasing

Apriori on Itemset Lattice

Transaction database

ID 1: ● ●
ID 2: ● ● ●
ID 3: ● ●
ID 4: ● ● ●
ID 5: ● ●
ID 6: ● ●
ID 7: ● ●
ID 8: ● ● ●

Itemset lattice



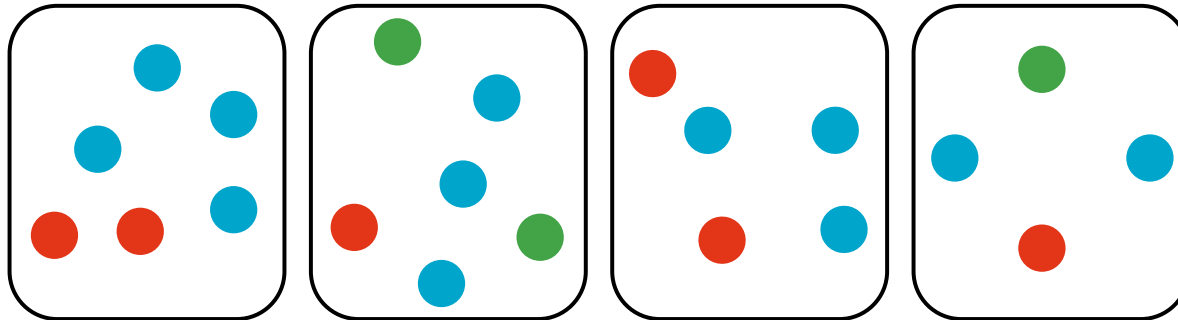
Apriori principle:
Support is monotonically decreasing

Discriminative Itemset Mining

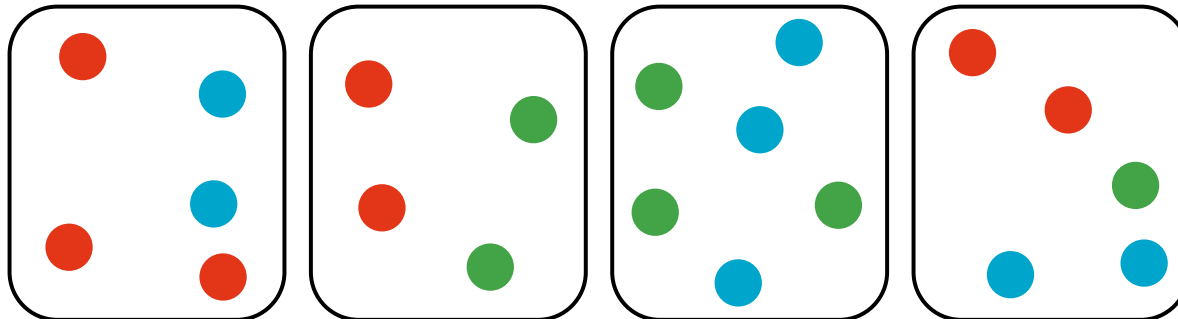
- Find **discriminative patterns** from **supervised** data
 - e.g. **Genome-wide association studies** in Bioinformatics

Database

Positive

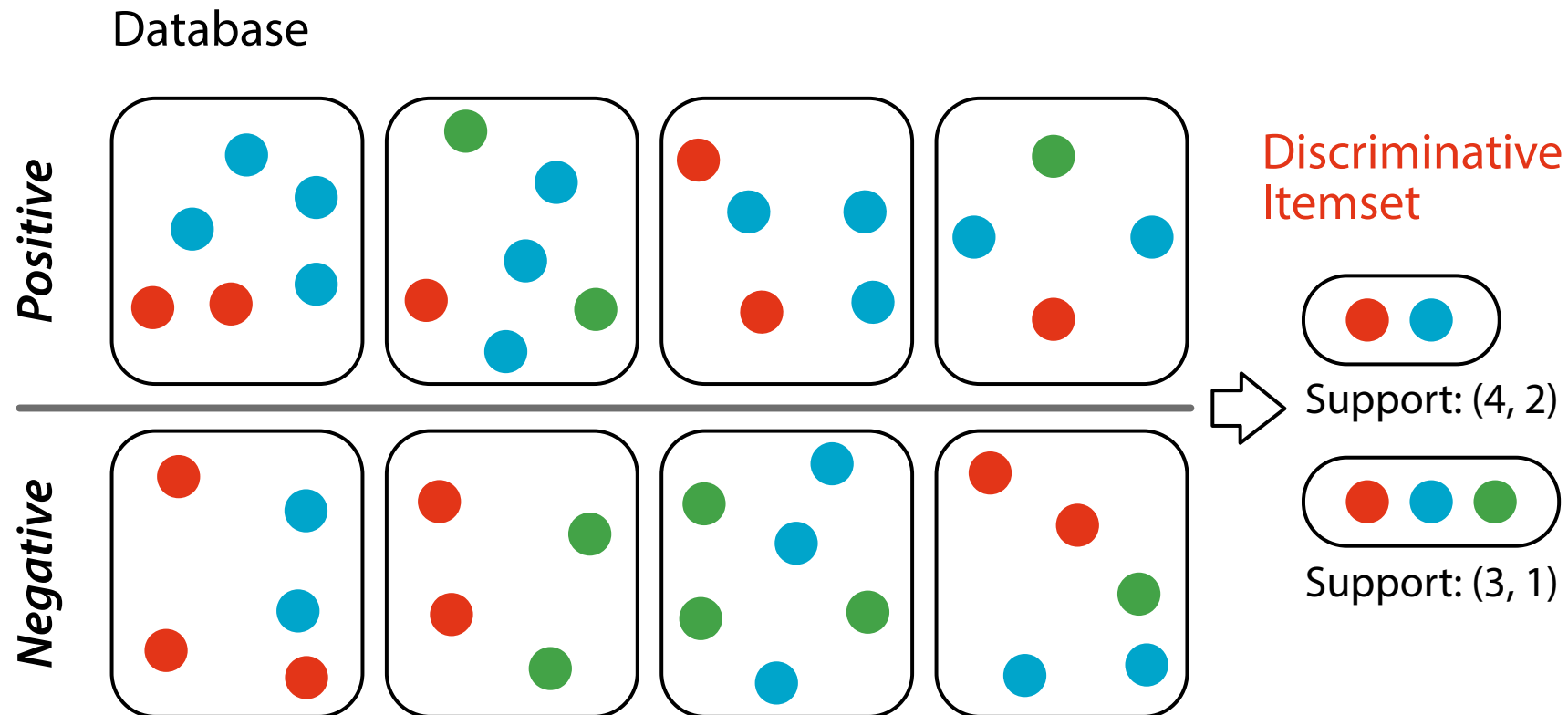


Negative



Discriminative Itemset Mining

- Find **discriminative patterns** from **supervised** data
 - e.g. **Genome-wide association studies** in Bioinformatics

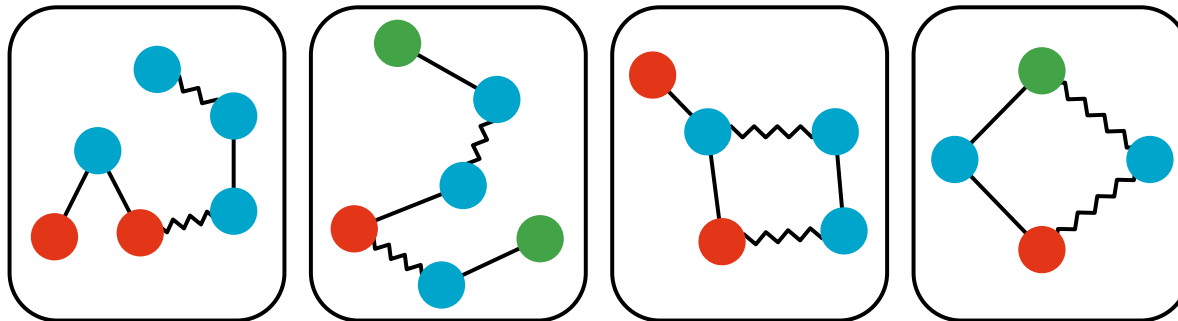


Discriminative Subgraph Mining

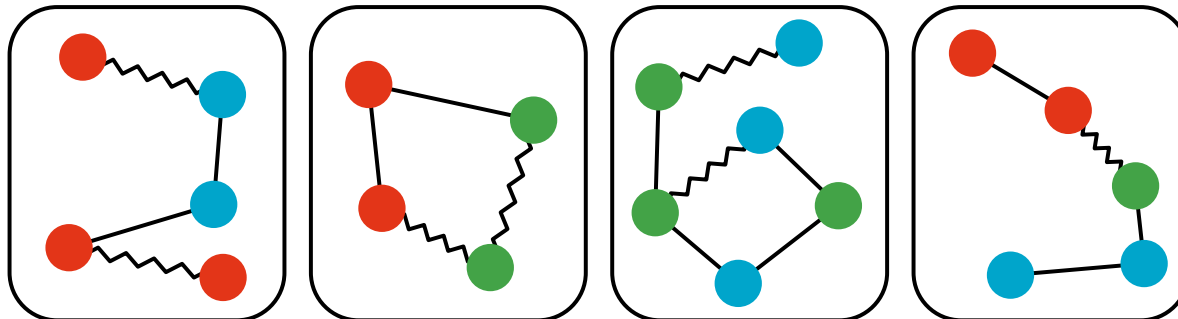
- Find **discriminative patterns** from **supervised data**
 - e.g. **Drug discovery**

Database

Positive



Negative

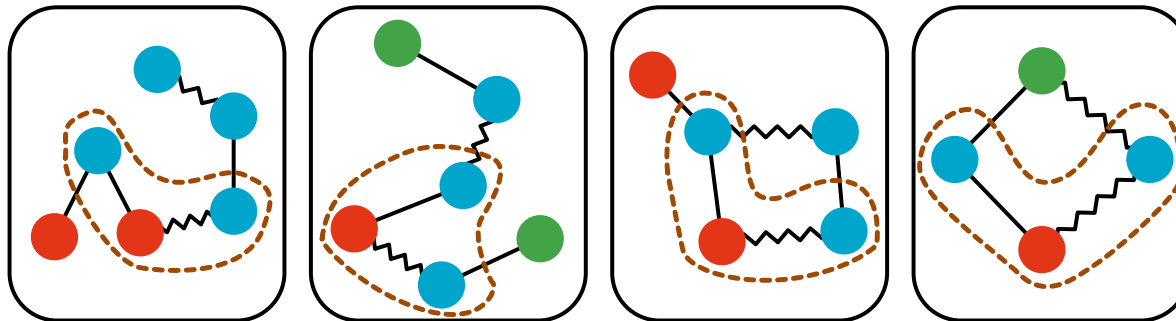


Discriminative Subgraph Mining

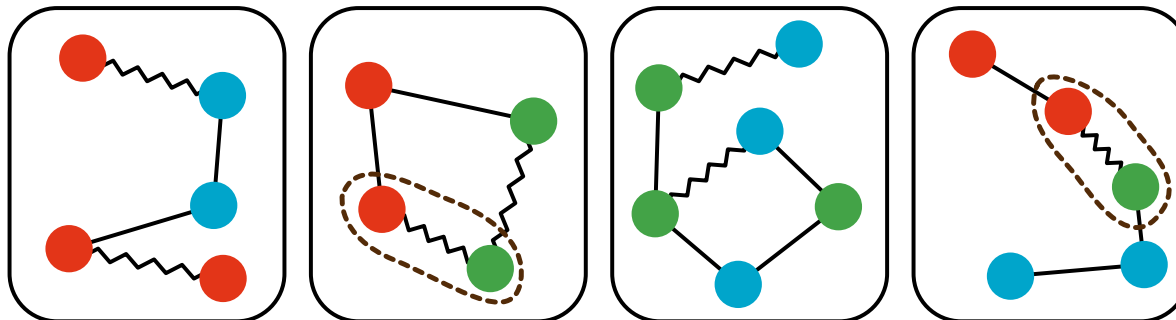
- Find **discriminative patterns** from **supervised data**
 - e.g. **Drug discovery**

Database

Positive



Negative



Discriminative
Subgraph



Support: (4, 0)

Support: (0, 2)

Agenda

- In discriminative pattern mining:
 1. How to measure the **discriminability** of patterns?
 2. How to enumerate all discriminative patterns?

Agenda

- In discriminative pattern mining:
 1. How to measure the **discriminability** of patterns?
 2. How to enumerate all discriminative patterns?
- *Answer to 1:*
 - Compute the **p-value** via *statistical hypothesis testing*
 - **Discriminative** pattern \iff **(Statistically) Significant** pattern
- *Answer to 2:*
 - Integrate evaluation of discriminability and enumeration of patterns

Outline

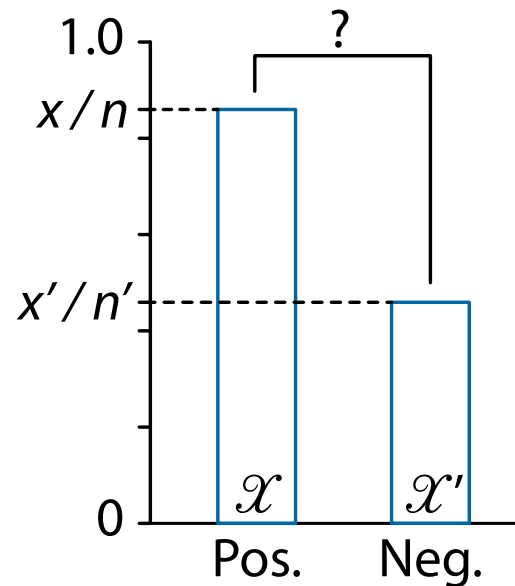
- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

Computing p -value of Pattern

- Given positive and negative sets of transactions $\mathcal{X}, \mathcal{X}'$
 - $|\mathcal{X}| = n, |\mathcal{X}'| = n' (n \leq n')$
- The p -value of each pattern H is determined by the Fisher's exact test
 - $x = |\{X \in \mathcal{X} \mid H \subseteq X\}|$

	Occ.	Non-occ.	Total
\mathcal{X} (Pos.)	x	$n - x$	n
\mathcal{X}' (Neg.)	x'	$n' - x'$	n'
Total	$x + x'$ $= \sigma$	$(n - x) + (n' - x')$	$n + n'$

Support



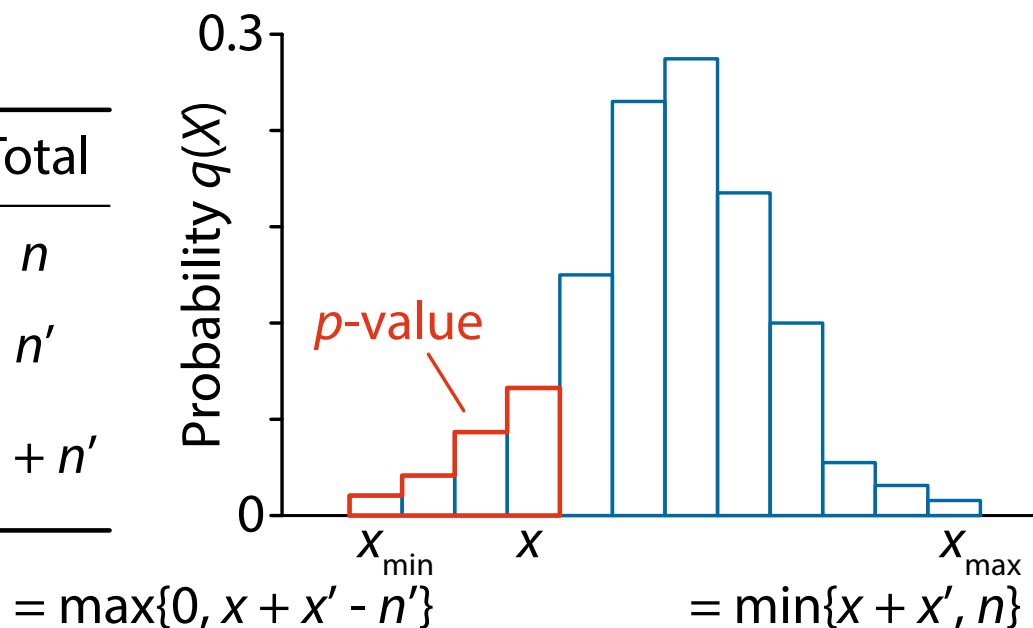
Fisher's Exact Test

- The probability $q(x)$ of obtaining x and x' is given by the hypergeometric distribution:

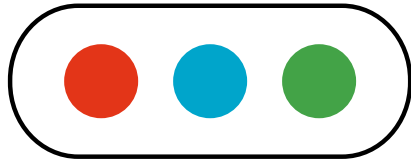
$$q(x) = \frac{\binom{n}{x} \binom{n'}{x'}}{\binom{n+n'}{x+x'}}$$

	Occ.	Non-occ.	Total
\mathcal{X} (Pos.)	x	$n - x$	n
\mathcal{X}' (Neg.)	x'	$n' - x'$	n'
Total	$x + x'$ $= \sigma$	$(n - x) + (n' - x')$	$n + n'$

Support



Hypothesis Test for Each Pattern



Alternative hypothesis
is true

Null hypothesis
is true

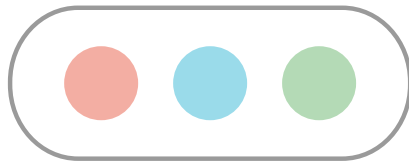
Declared significant ($p\text{-value} < \alpha$)	True Positive	False Positive (Type I Error)
Declared non-significant	False Negative (Type II Error)	True Negative

Null hypothesis:

The occurrence of the pattern is
independent from classes

Alternative hypothesis: The occurrence of the pattern is
associated with classes

Hypothesis Test for Each Pattern



Alternative hypothesis
is true

Null hypothesis
is true

Declared
significant
($p\text{-value} < \alpha$)

True Positive

False Positive
(Type I Error)

Declared
non-signif

False Negative

True Negative

Binary desicion (**significant or not**) via
the user specified threshold α

Null hypothesis:

The occurence of the pattern is
independent from classes

Alternative hypothesis: The occurence of the pattern is
associated with classes

Outline

- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

Multiple Testing

Pattern: 

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

Fisher's exact test: $p\text{-value} = 0.429$

Multiple Testing

Pattern:



Pat

	Occ.	Non-occ.	Total
--	------	----------	-------

Positive	3	1	4
----------	---	---	---

Pos

Negative	1	3	4
----------	---	---	---

Neg

Total	4	4	8
-------	---	---	---

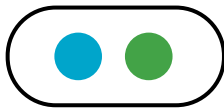
Tot

Fisher's exact test: $p\text{-value} = 0.486$

Fisher's exact test: $p\text{-value} = 0.429$

Multiple Testing

Pattern:



Pat

Pat

Pos

Pos

Neg

Tot

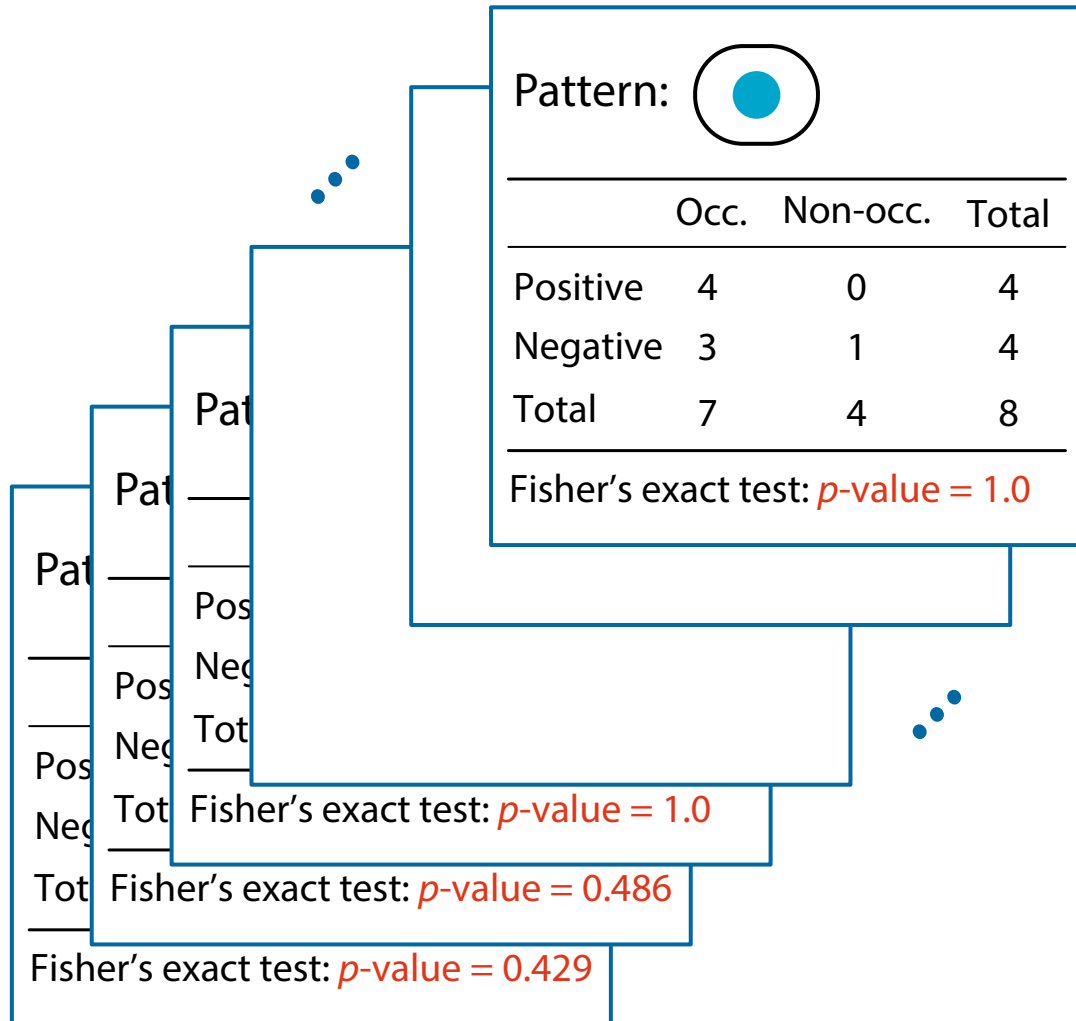
	Occ.	Non-occ.	Total
Positive	2	2	4
Negative	2	2	4
Total	4	4	8

Fisher's exact test: $p\text{-value} = 1.0$

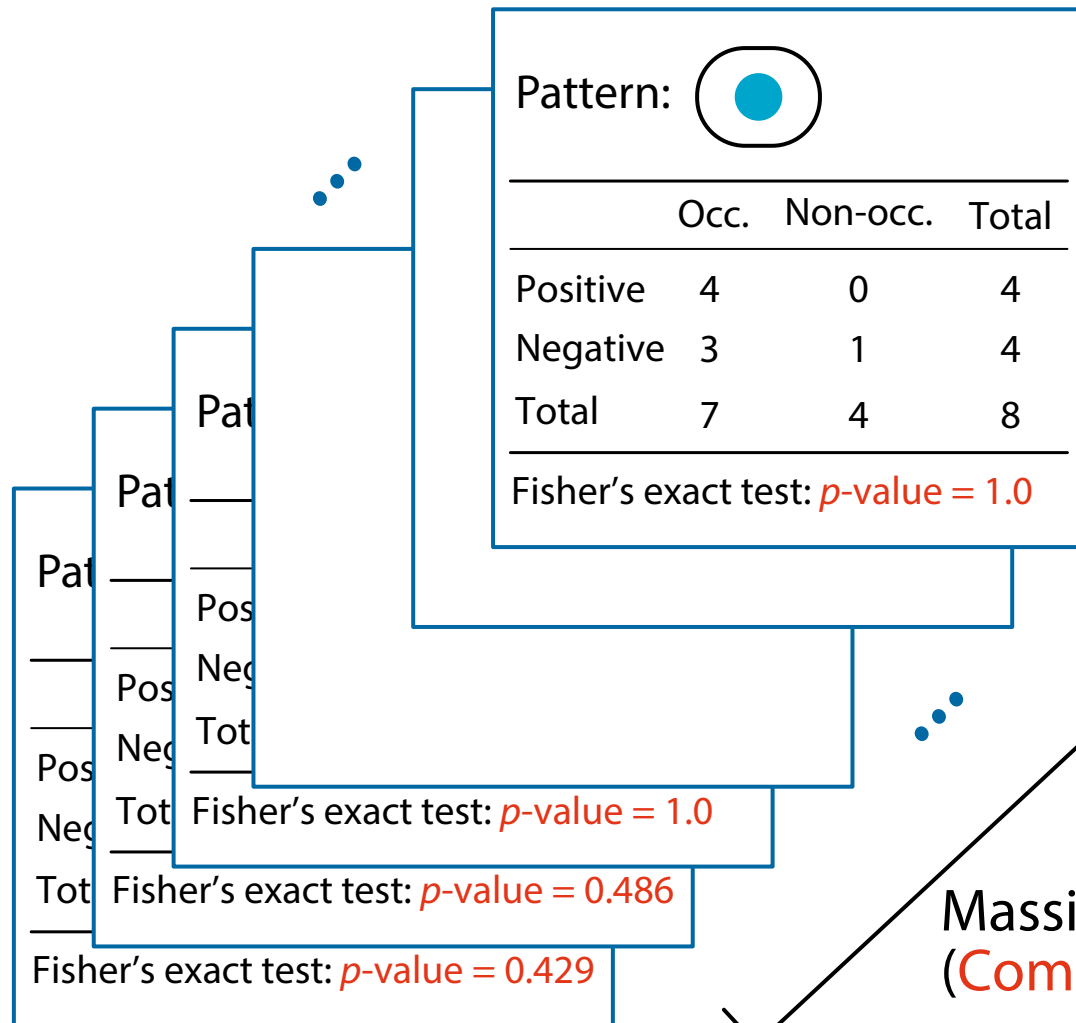
Fisher's exact test: $p\text{-value} = 0.486$

Fisher's exact test: $p\text{-value} = 0.429$

Multiple Testing



Multiple Testing



Task: Enumerate **all** significant patterns

Massive number of patterns
(**Combinatorial explosion!**)

Multiple Testing Correction

- In each test, [probability of having a false positive] $\leq \alpha$
- If we repeat m tests, αm patterns can be false positives
 - Too many if m is large!
- 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] $\leq \alpha$
- If we repeat m tests, *am patterns can be false positives*
 - Too many if m is large!
- 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*
 - One of the standard criteria in multiple testing
 - $\text{FWER} = 1 - (1 - \alpha)^m$ if patterns are independent

Controlling the FWER

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

Controlling the FWER

- $\text{FWER} = \Pr(\text{FP} > 0)$
 - FP: Number of false positives
- To achieve $\text{FWER} = \alpha$, change the significance level for each pattern from α to δ
 - δ : corrected significance level
 - $\delta \leq \alpha$
- Objective is to optimize (maximize) δ :
$$\delta^* = \underset{\delta}{\operatorname{argmax}} \text{FWER}(\delta) \quad \text{s.t. } \text{FWER}(\delta) \leq \alpha$$
 - $\text{FWER}(\delta)$: FWER at corrected significance level δ
 - Cannot be evaluated in closed form (simple but not easy!)
 - Bonferroni correction is popular: $\delta_{\text{Bon}}^* = \alpha/m$

Outline

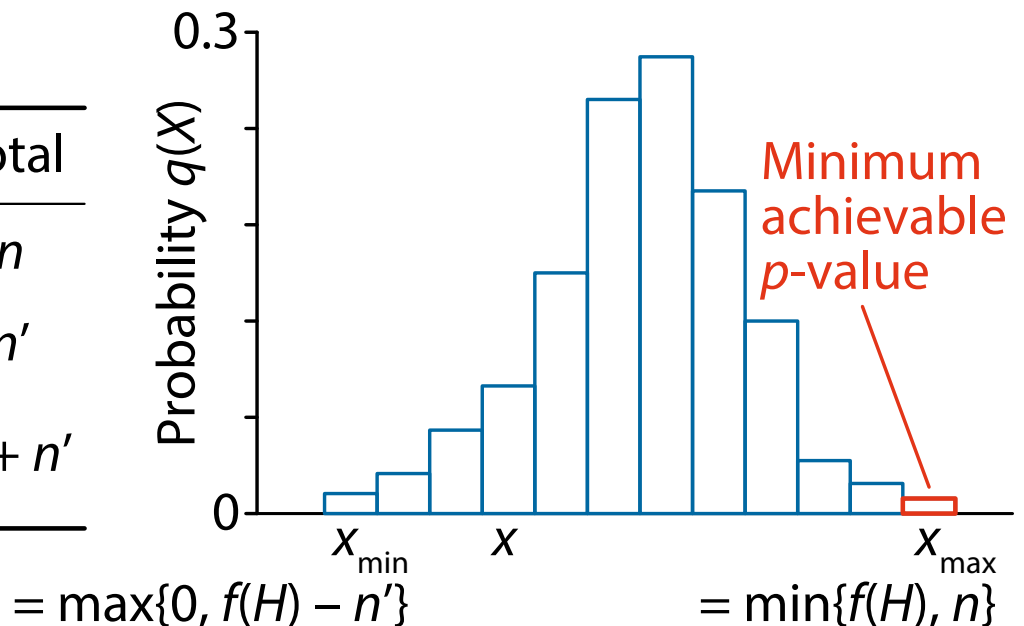
- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

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

- Consider the **minimum achievable p -value** $\Psi(\sigma)$ of a pattern H for its **support** $\sigma = |\{X \in \mathcal{X} \cup \mathcal{X}' \mid H \subseteq X\}|$
 - $\Psi(\sigma) = \min\{p(x) \mid x_{\min} \leq x \leq x_{\max}\}$
 - $x_{\min} = \max\{0, \sigma - n'\}$, $x_{\max} = \min\{\sigma, n\}$

	Occ.	Non-occ.	Total
\mathcal{X} (Pos.)	x	$n - x$	n
\mathcal{X}' (Neg.)	x'	$n' - x'$	n'
Total	$x + x'$ $= \sigma$	$(n - x) + (n' - x')$	$n + n'$

Support



Computing $\Psi(\sigma)$

- Consider the **minimum achievable p -value** $\Psi(\sigma)$ of a pattern H for its **support** $\sigma = |\{X \in \mathcal{X} \cup \mathcal{X}' \mid H \subseteq X\}|$

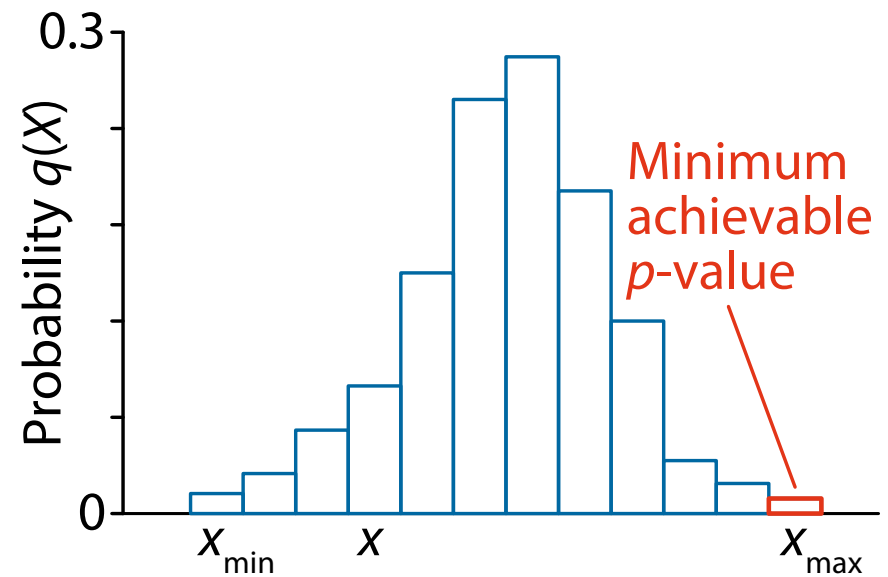
$$\Psi(\sigma) = \binom{n}{\sigma} / \binom{n + n'}{\sigma}$$

	Occ.	Non-occ.	Total
\mathcal{X} (Pos.)	σ	$n - \sigma$	n
\mathcal{X}' (Neg.)	0	n'	n'
Total	σ	$(n - \sigma) + n'$	$n + n'$

Most biased case ($\sigma < n$)

$$= \max\{0, f(H) - n'\}$$

$$= \min\{f(H), n\}$$



Testability

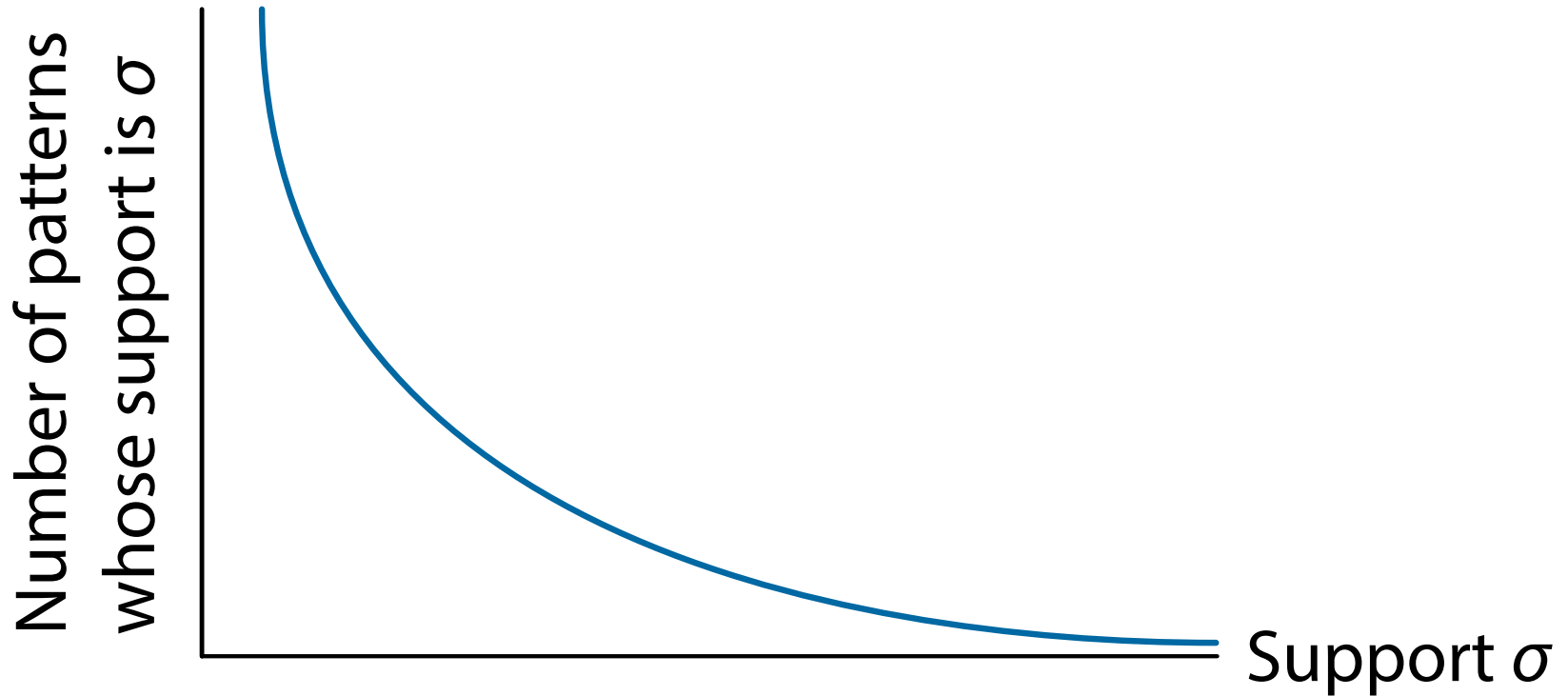
- Consider the **minimum achievable p -value** $\Psi(\sigma)$ of a pattern H for its **support** $\sigma = |\{ X \in \mathcal{X} \cup \mathcal{X}' \mid H \subseteq X \}|$

$$\Psi(\sigma) = \binom{n}{\sigma} / \binom{n + n'}{\sigma}$$

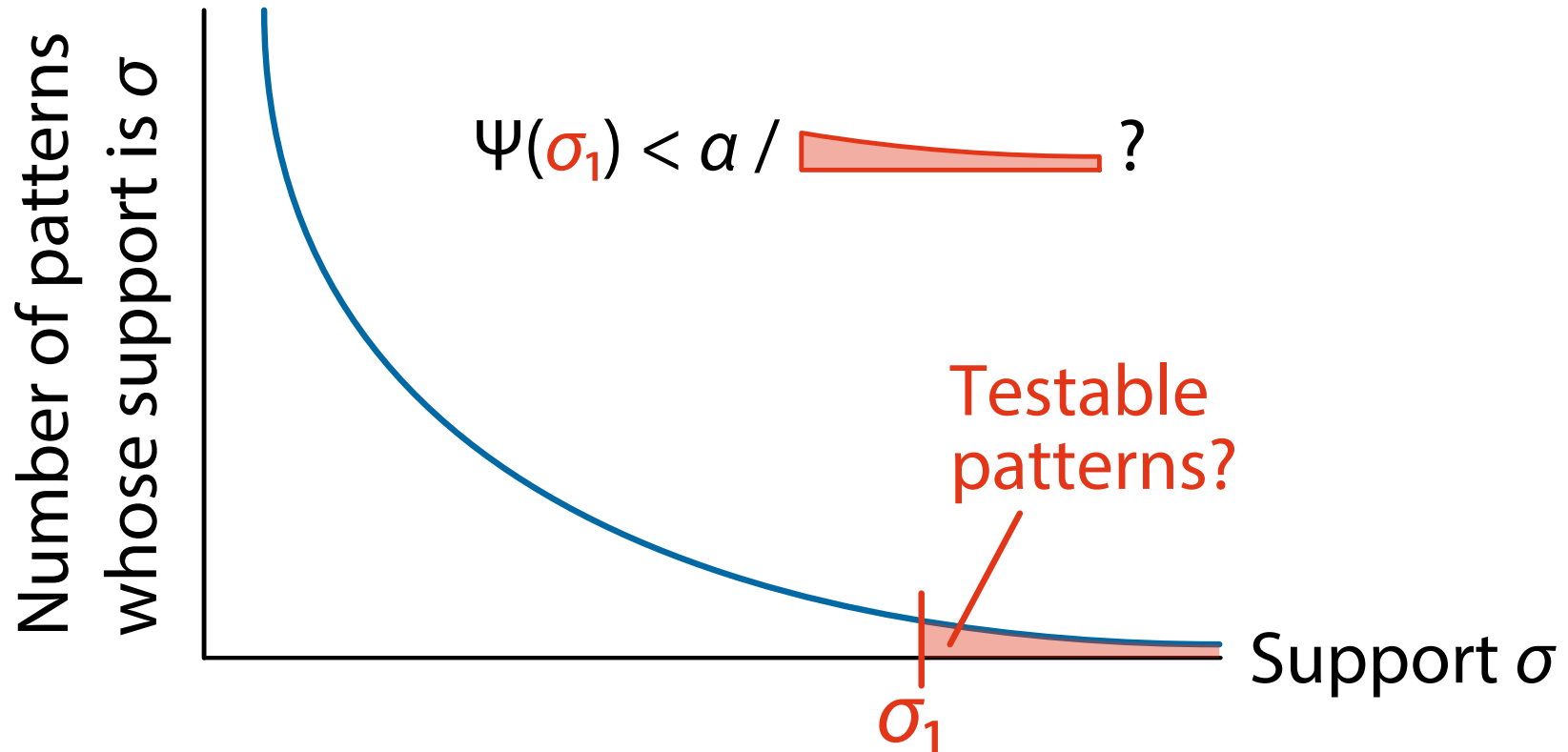
- Tarone (1990) pointed out (and Terada et al. (2013) revisited):

*For a pattern H , if $\Psi(\sigma)$ is larger than the significance threshold, this is **untestable** and we can ignore it*

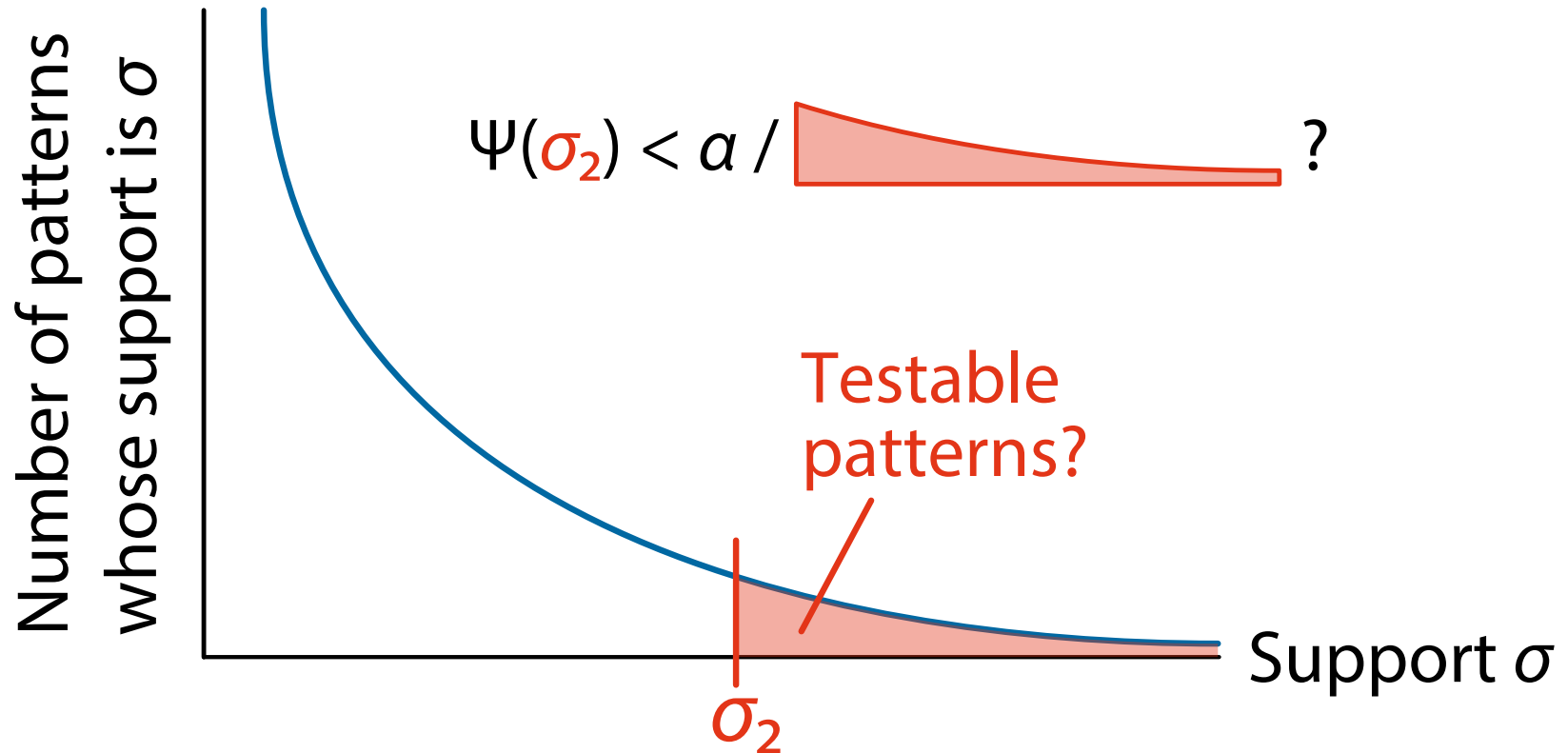
Finding Testable Patterns



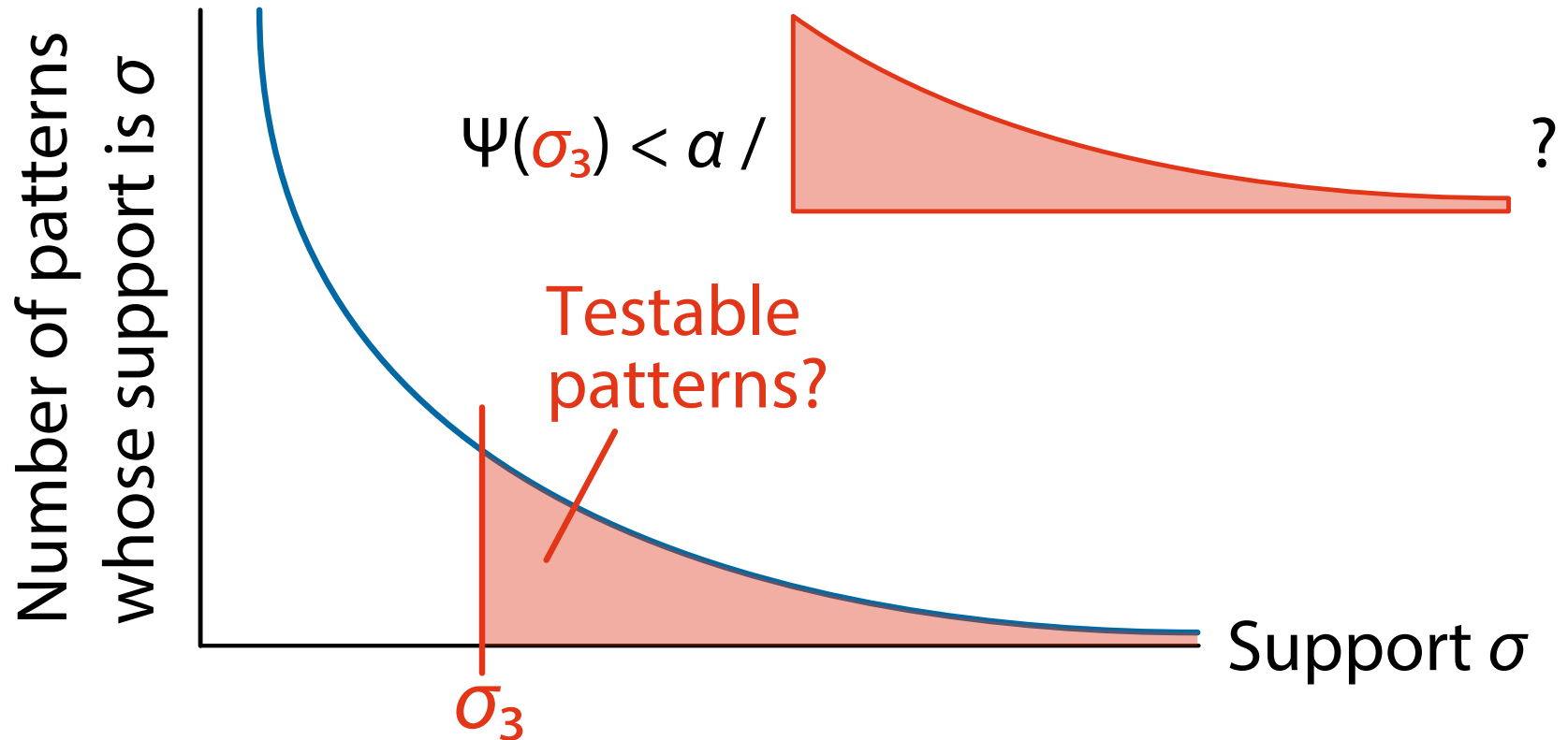
Finding Testable Patterns



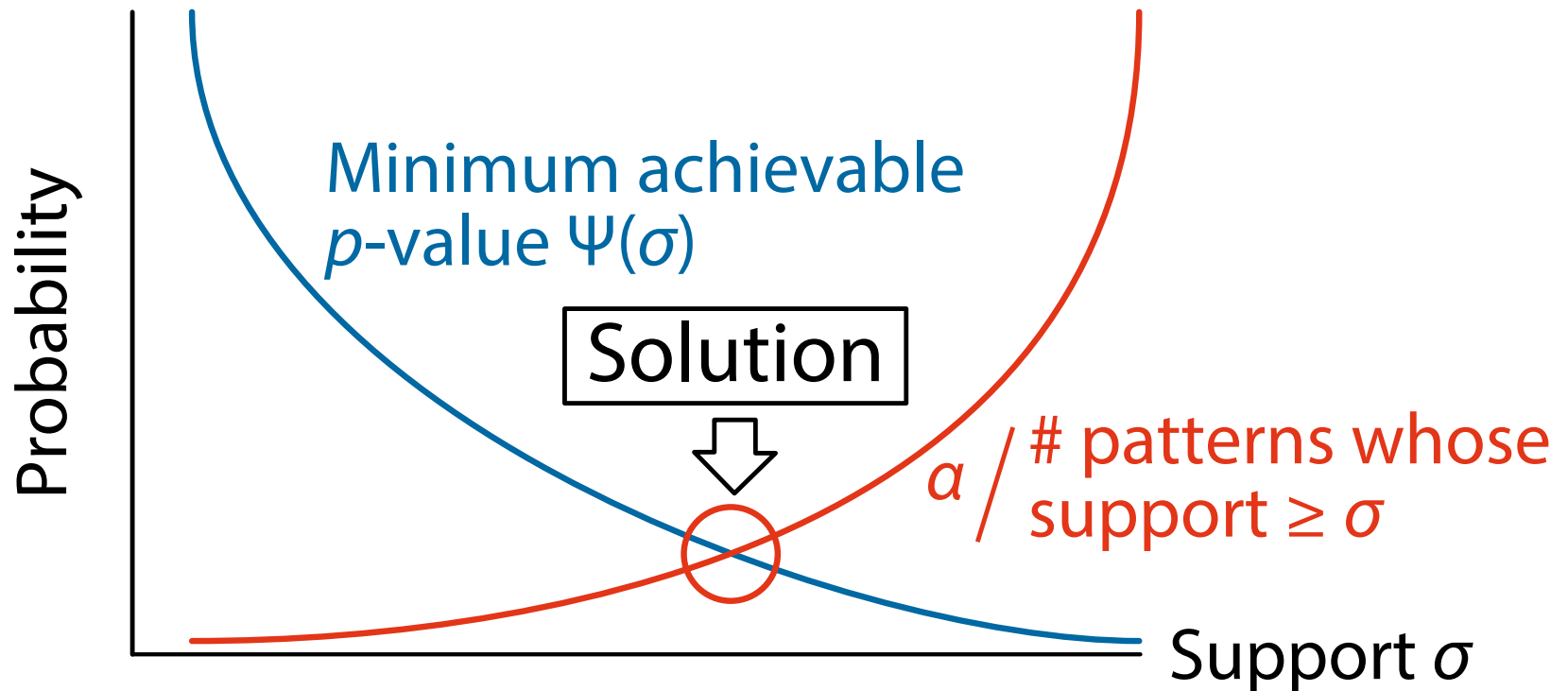
Finding Testable Patterns



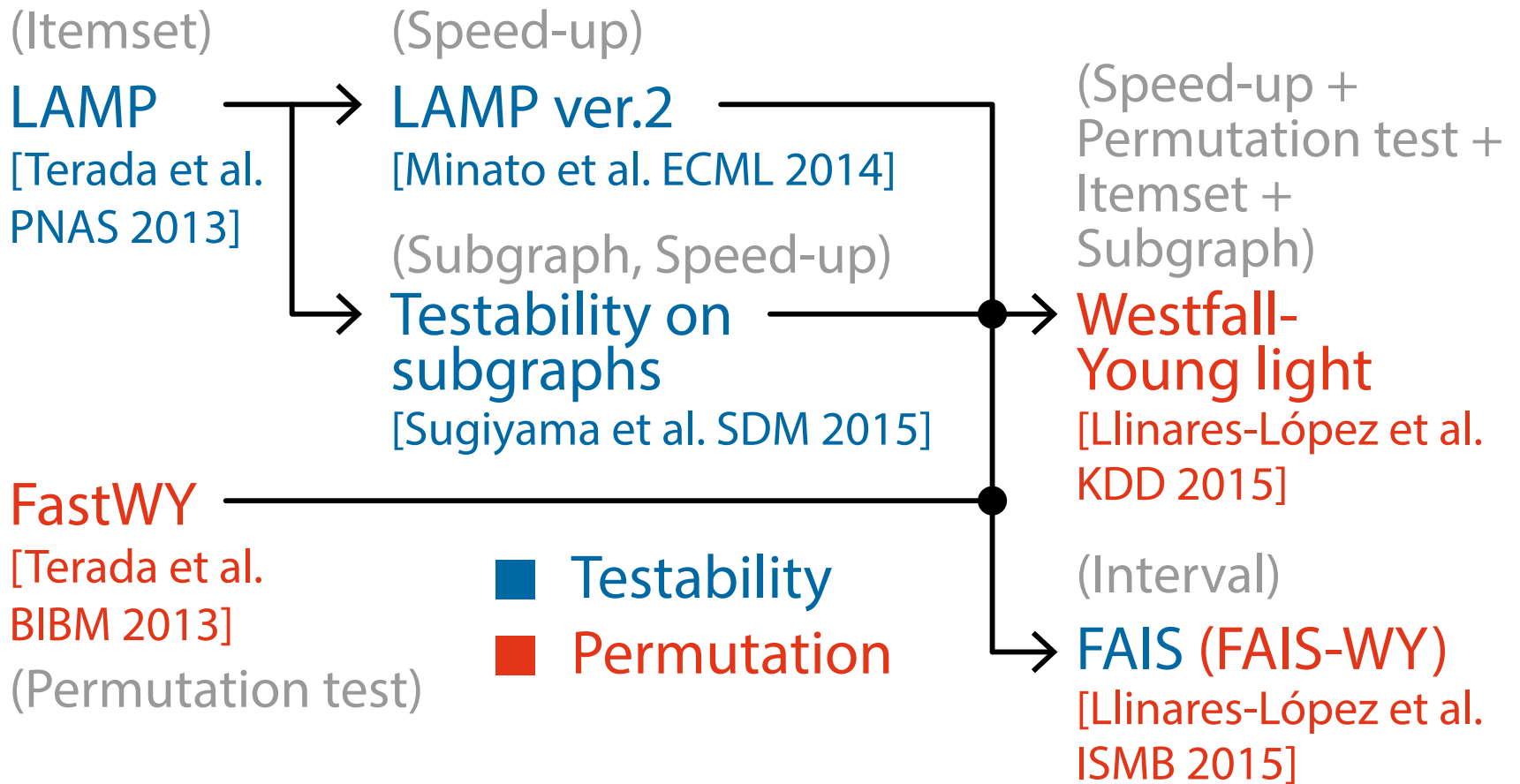
Finding Testable Patterns



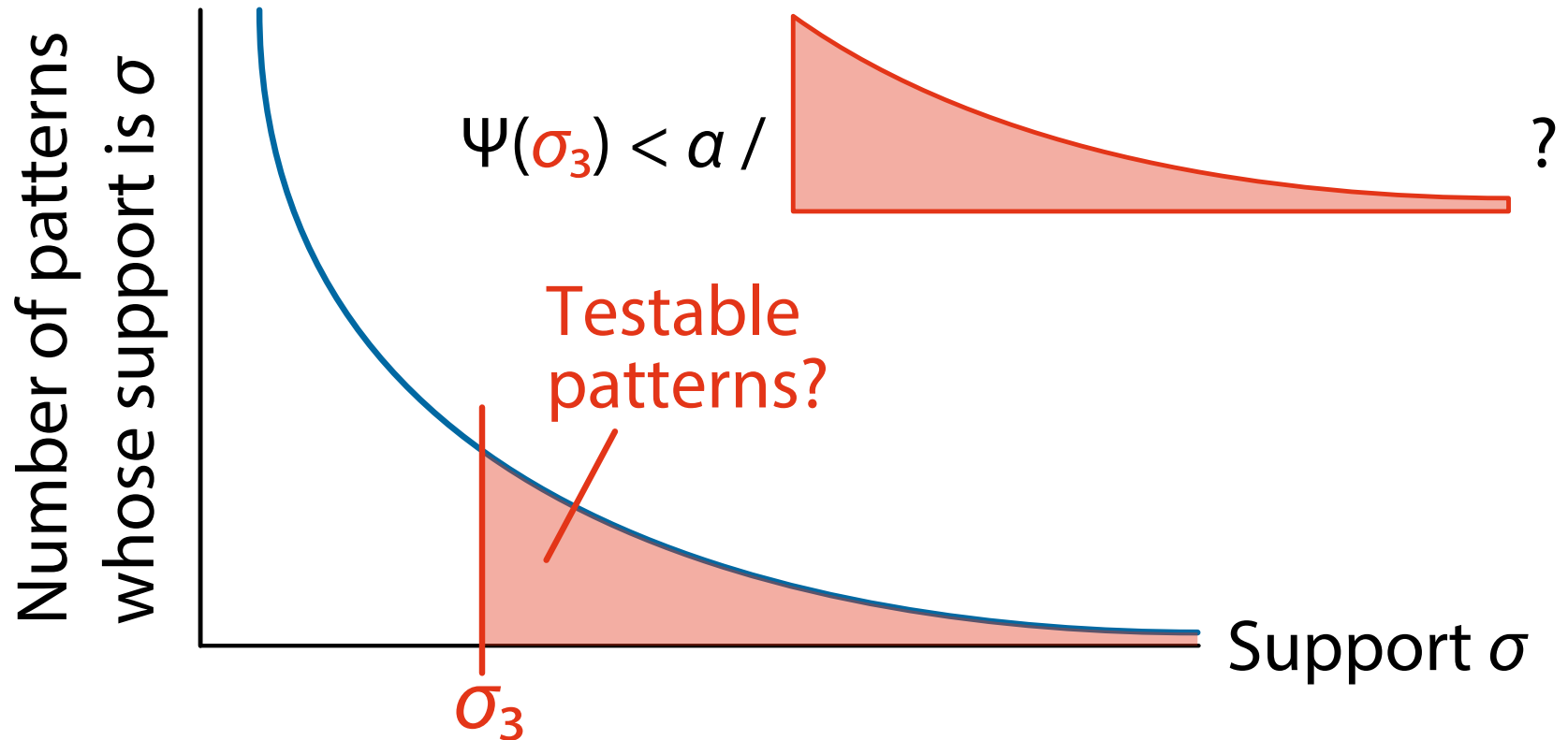
Finding Testable Patterns



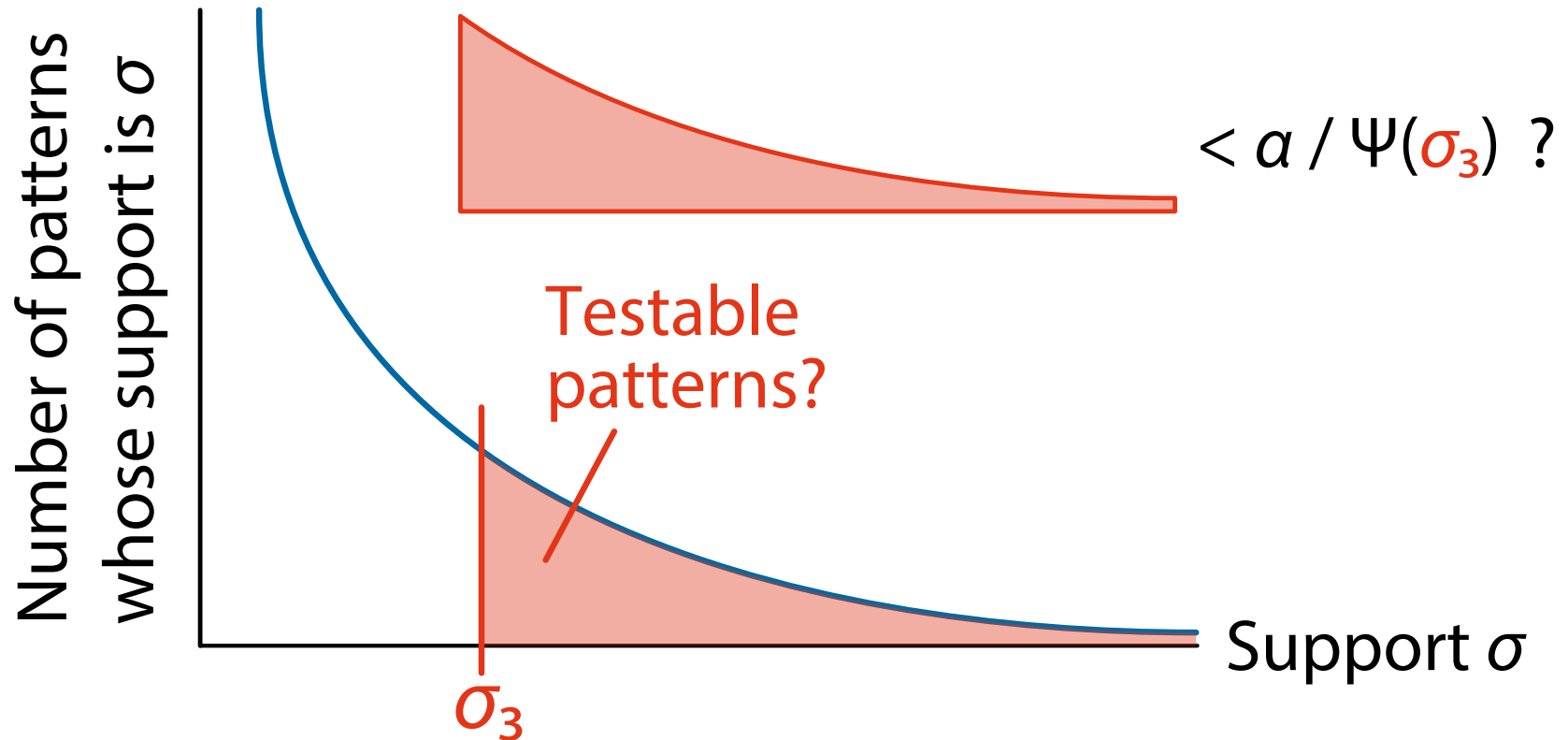
State-of-the-art



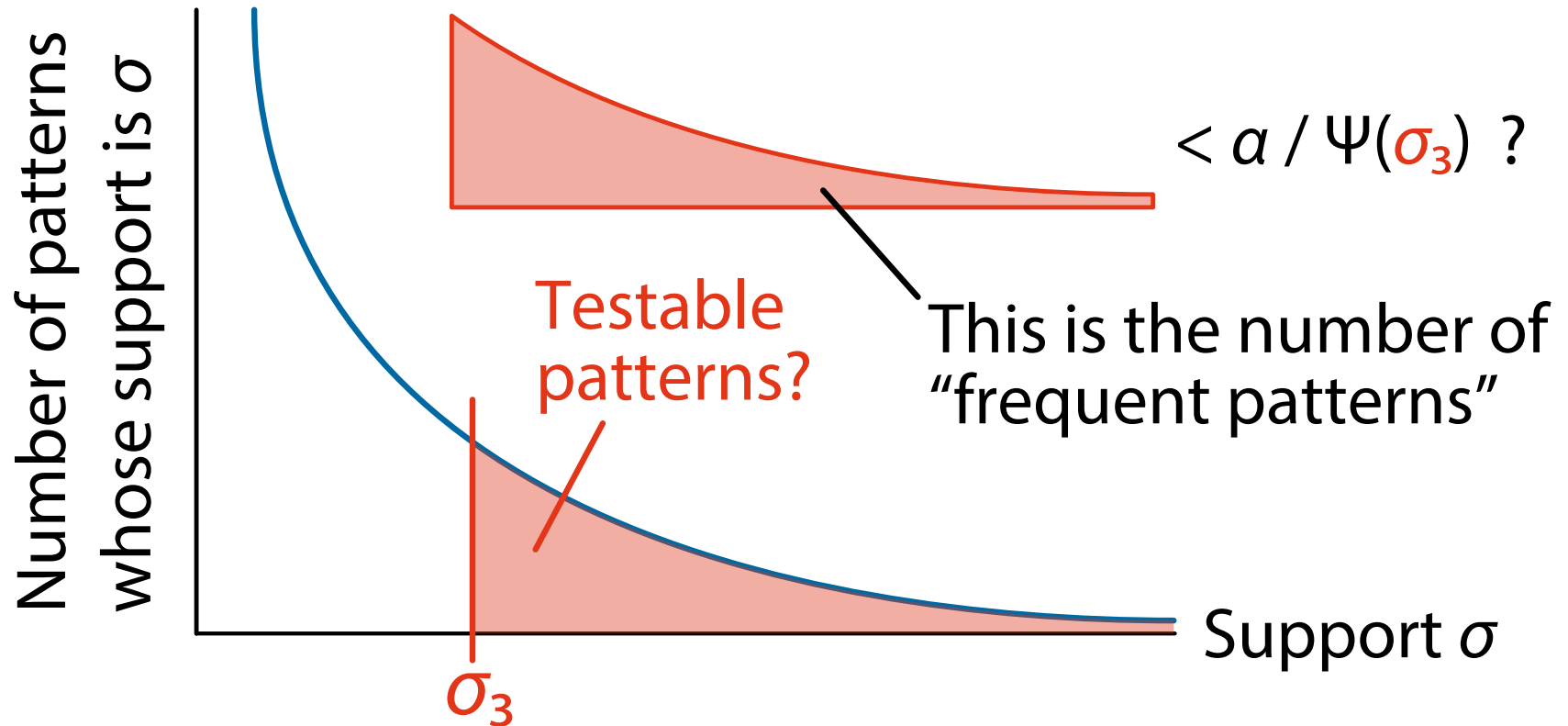
How to Find Testable Patterns?



How to Find Testable Patterns?

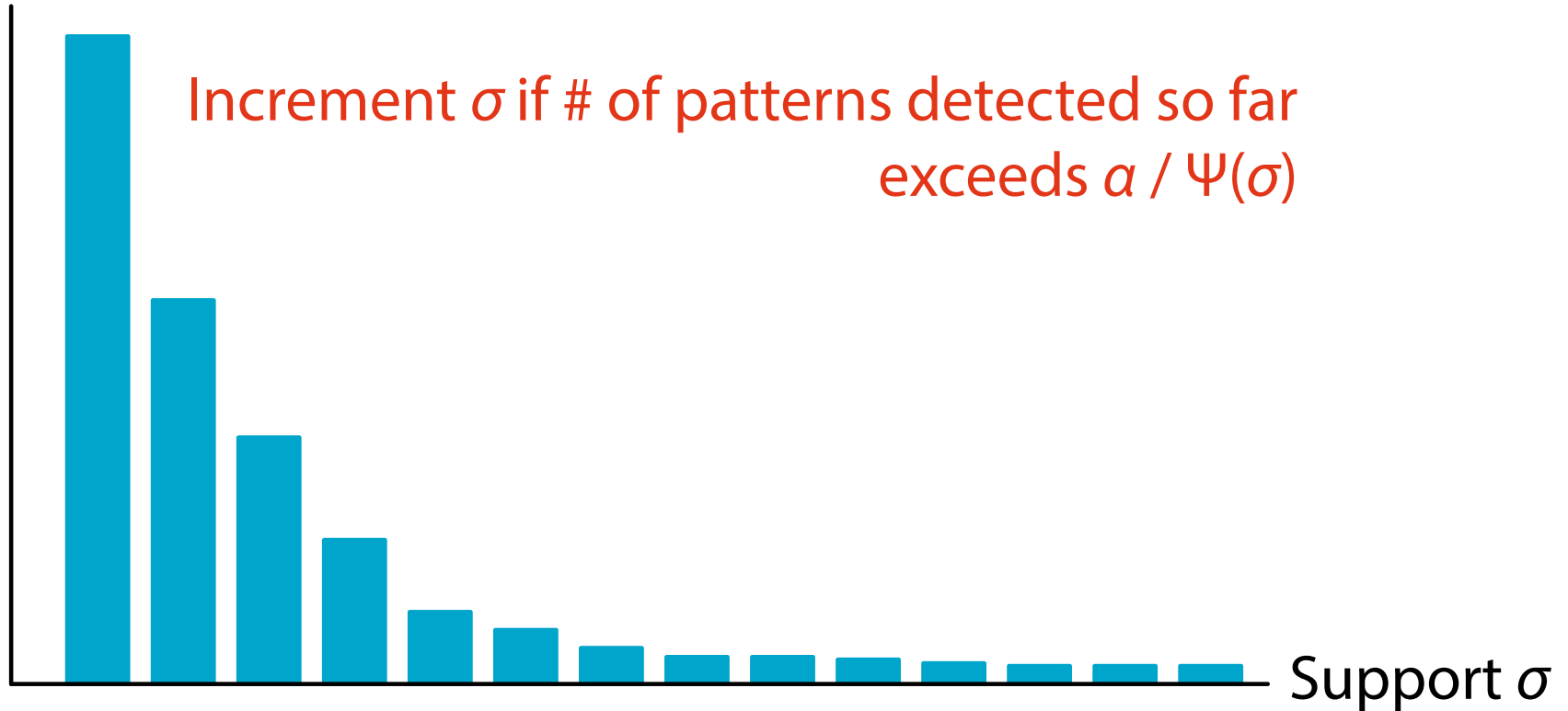


How to Find Testable Patterns?



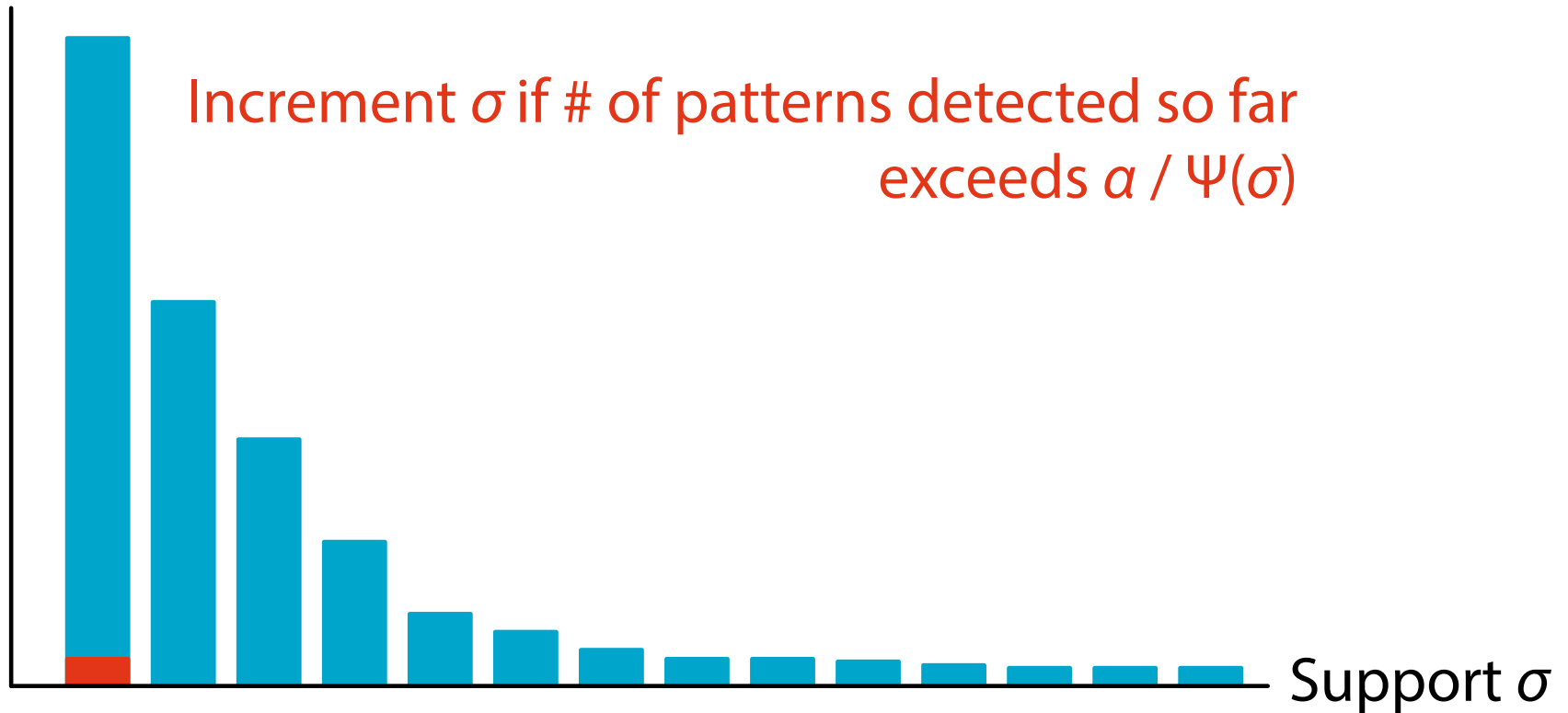
LAMP ver.2

of patterns



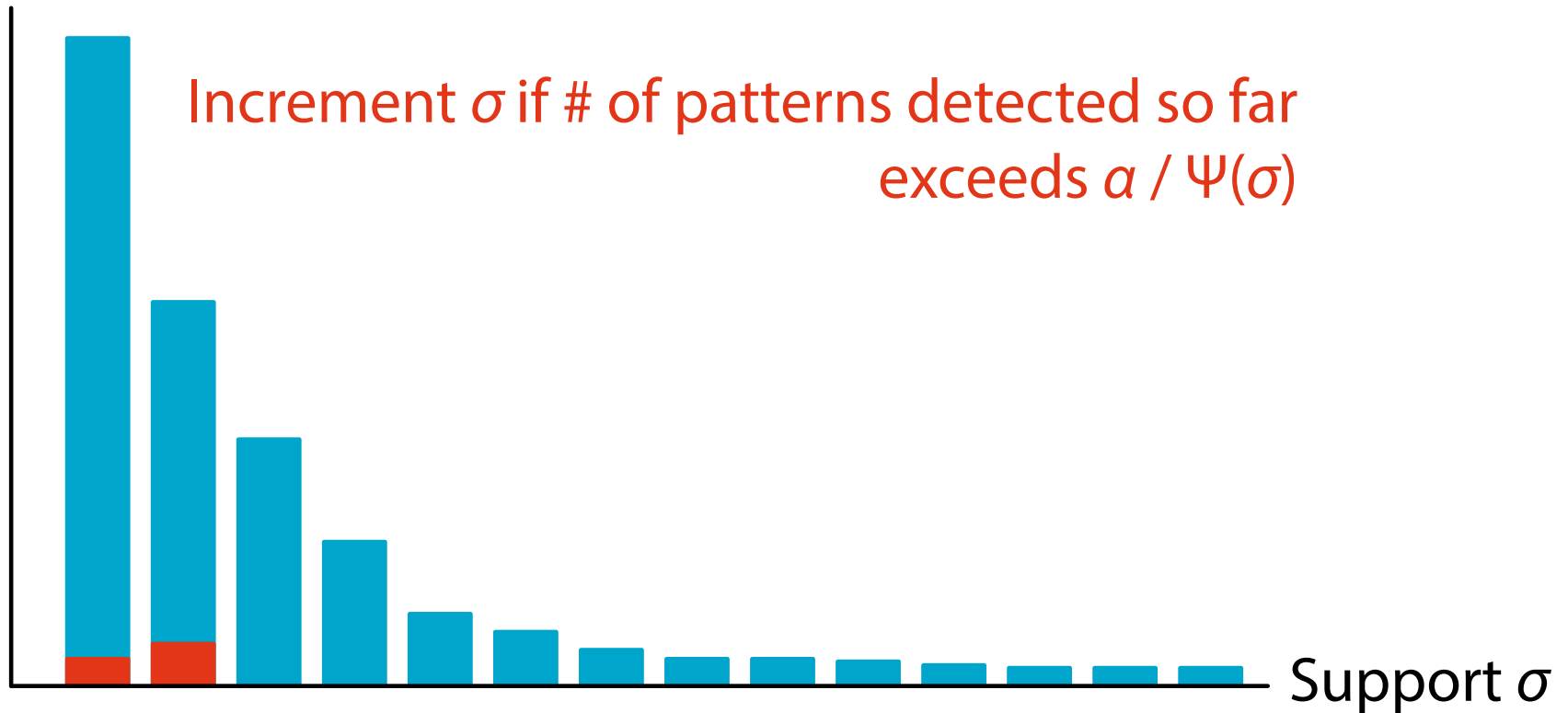
LAMP ver.2

of patterns



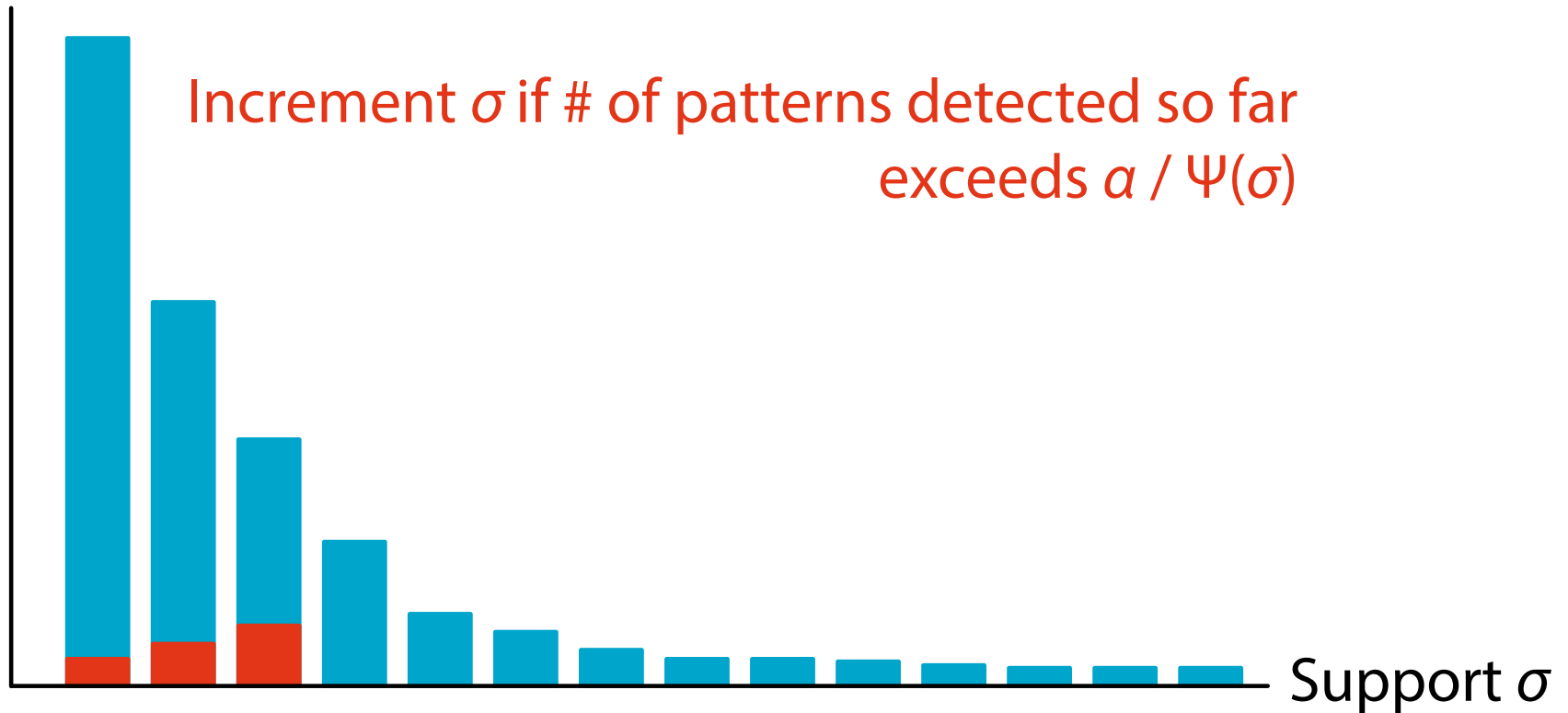
LAMP ver.2

of patterns



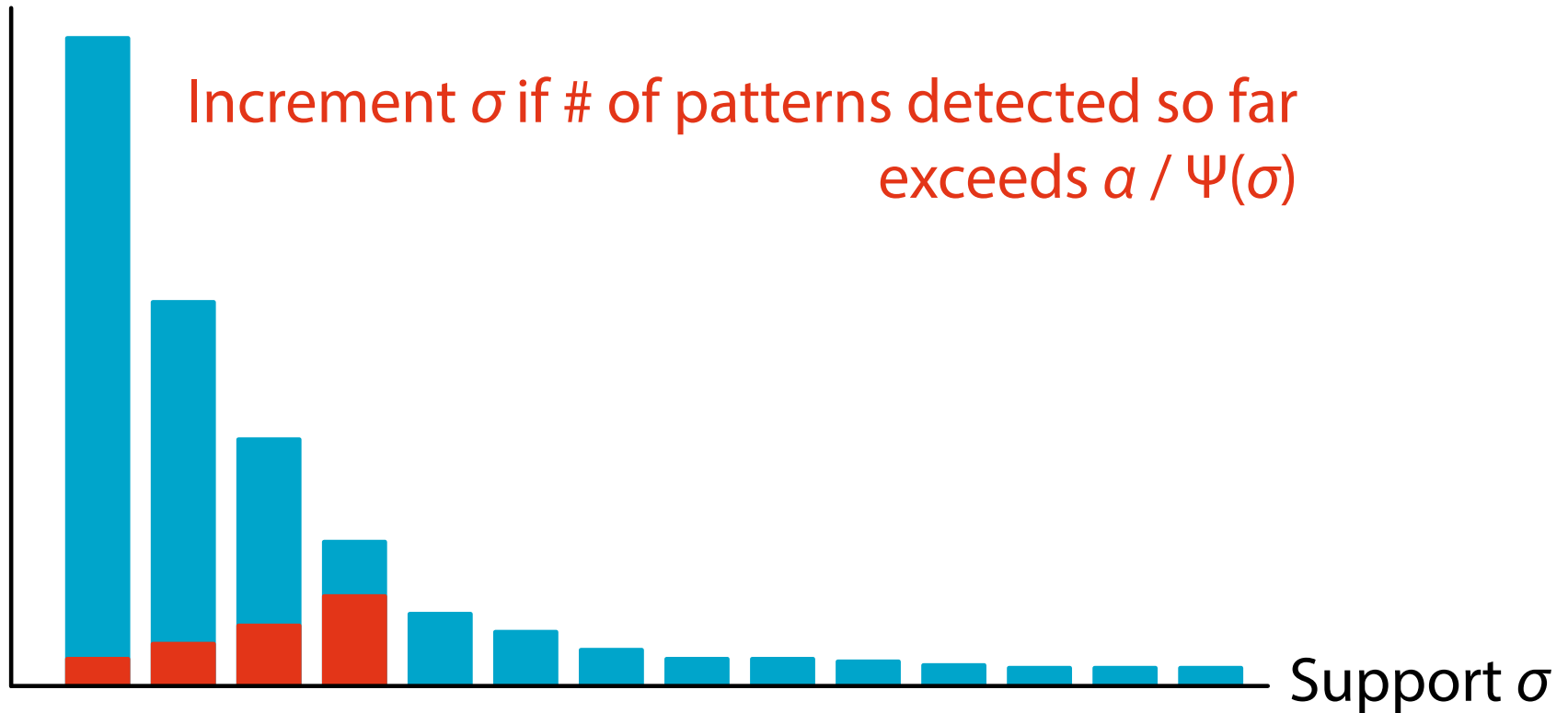
LAMP ver.2

of patterns



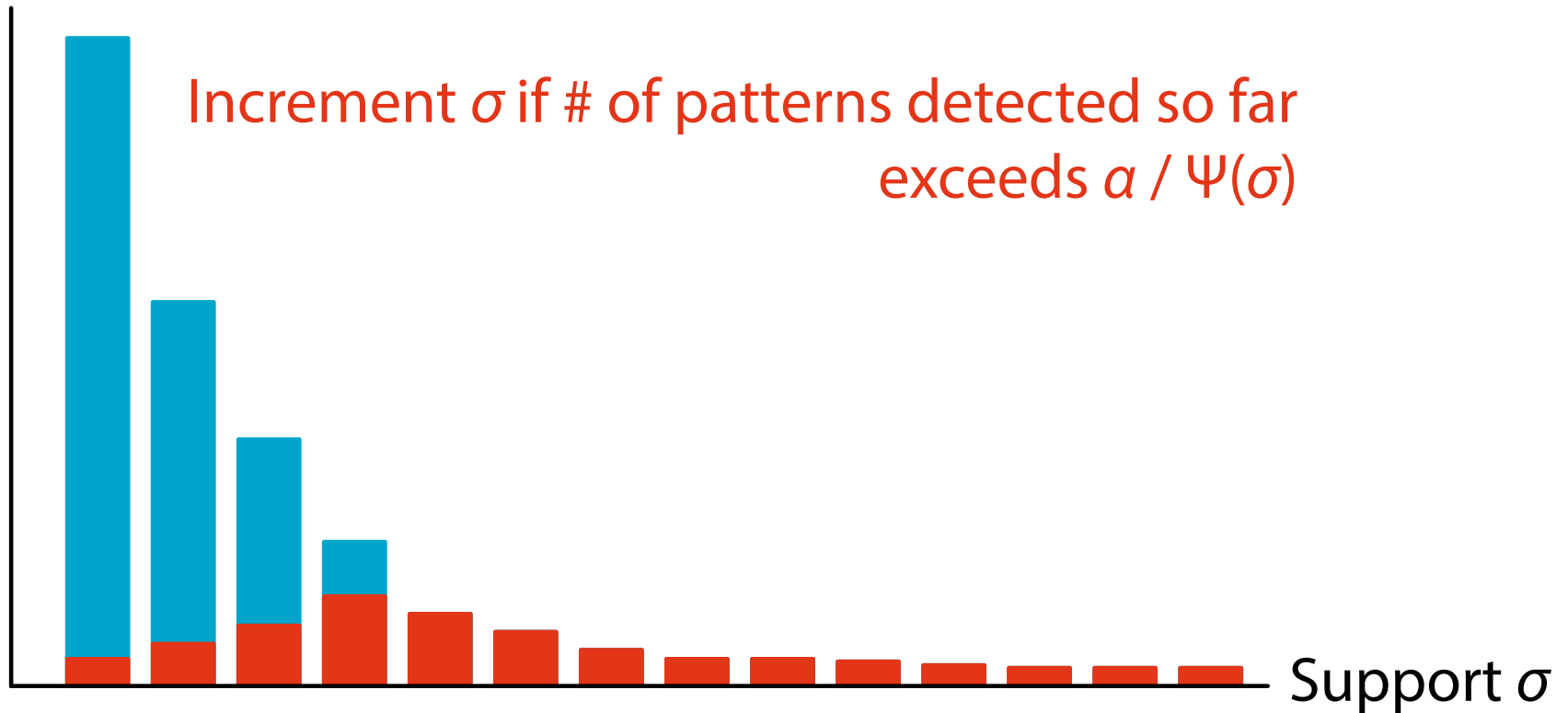
LAMP ver.2

of patterns



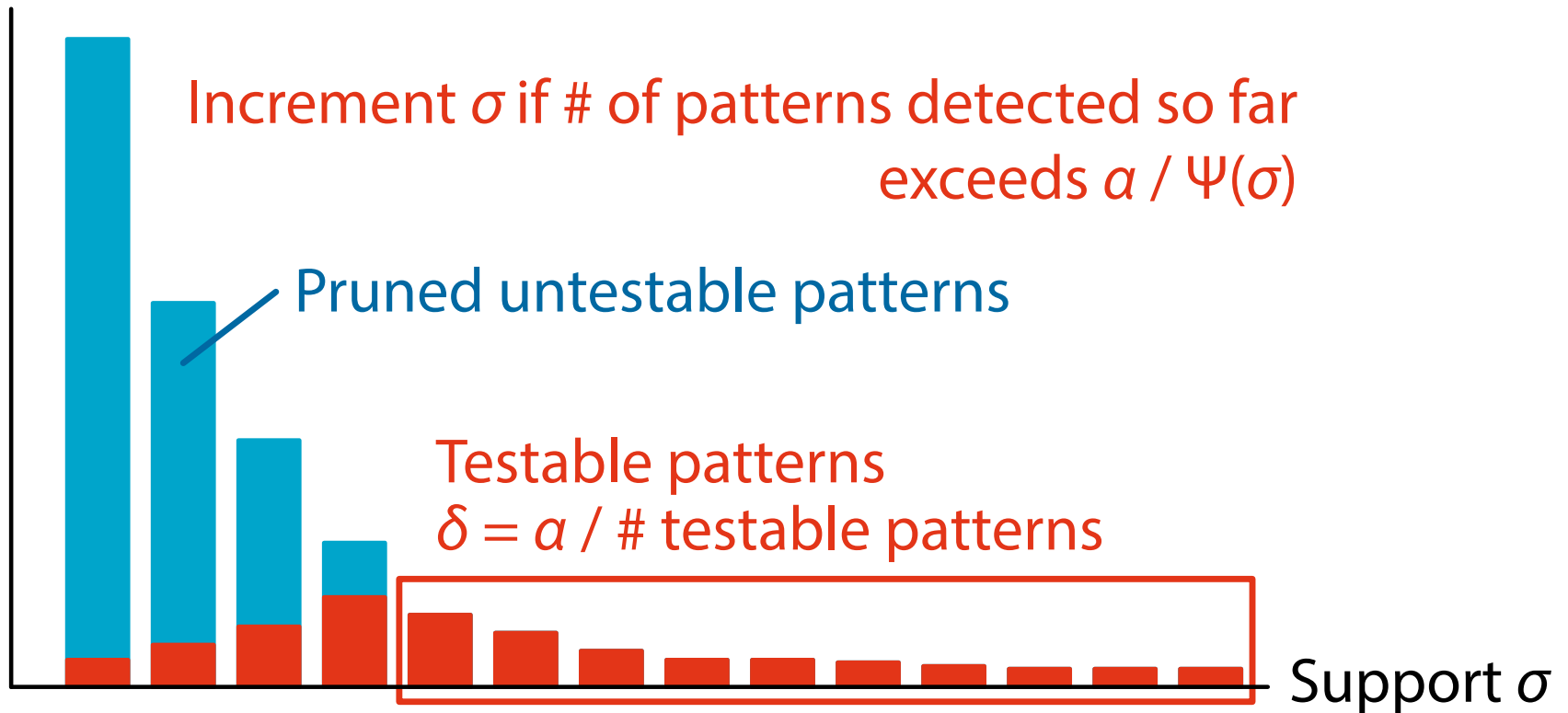
LAMP ver.2

of patterns

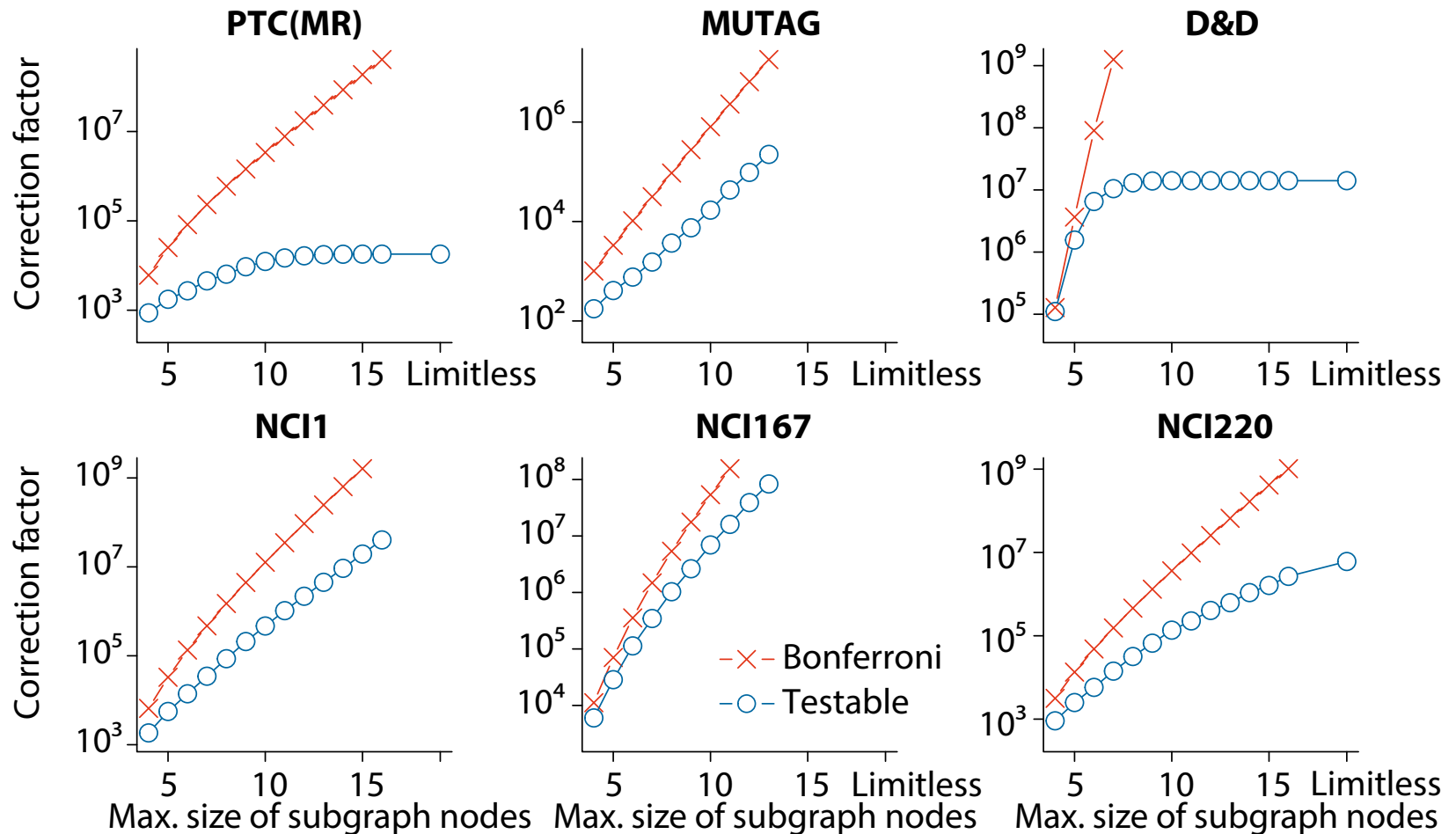


LAMP ver.2

of patterns



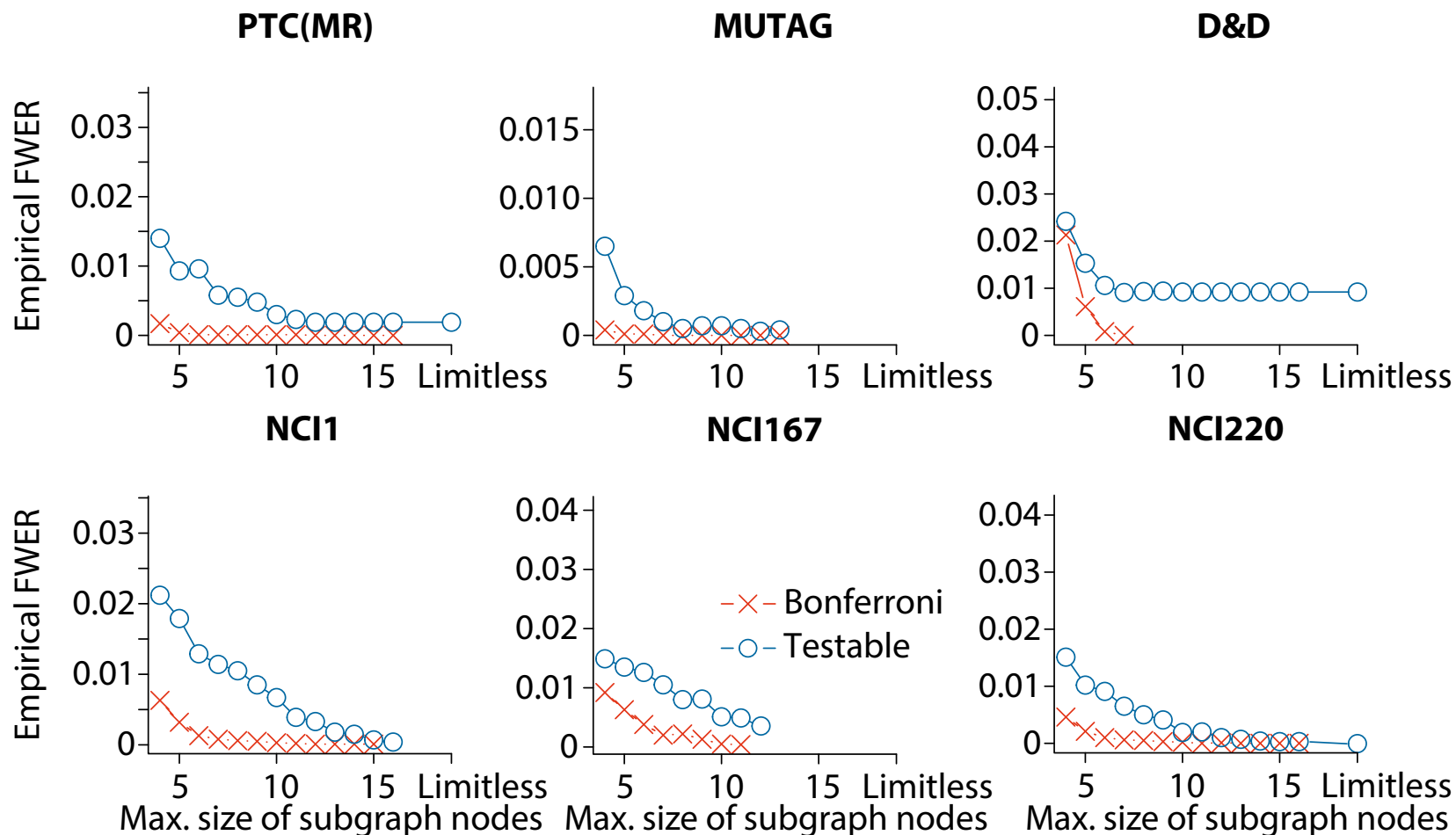
Testable Patterns in Subgraph Mining



Outline

- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

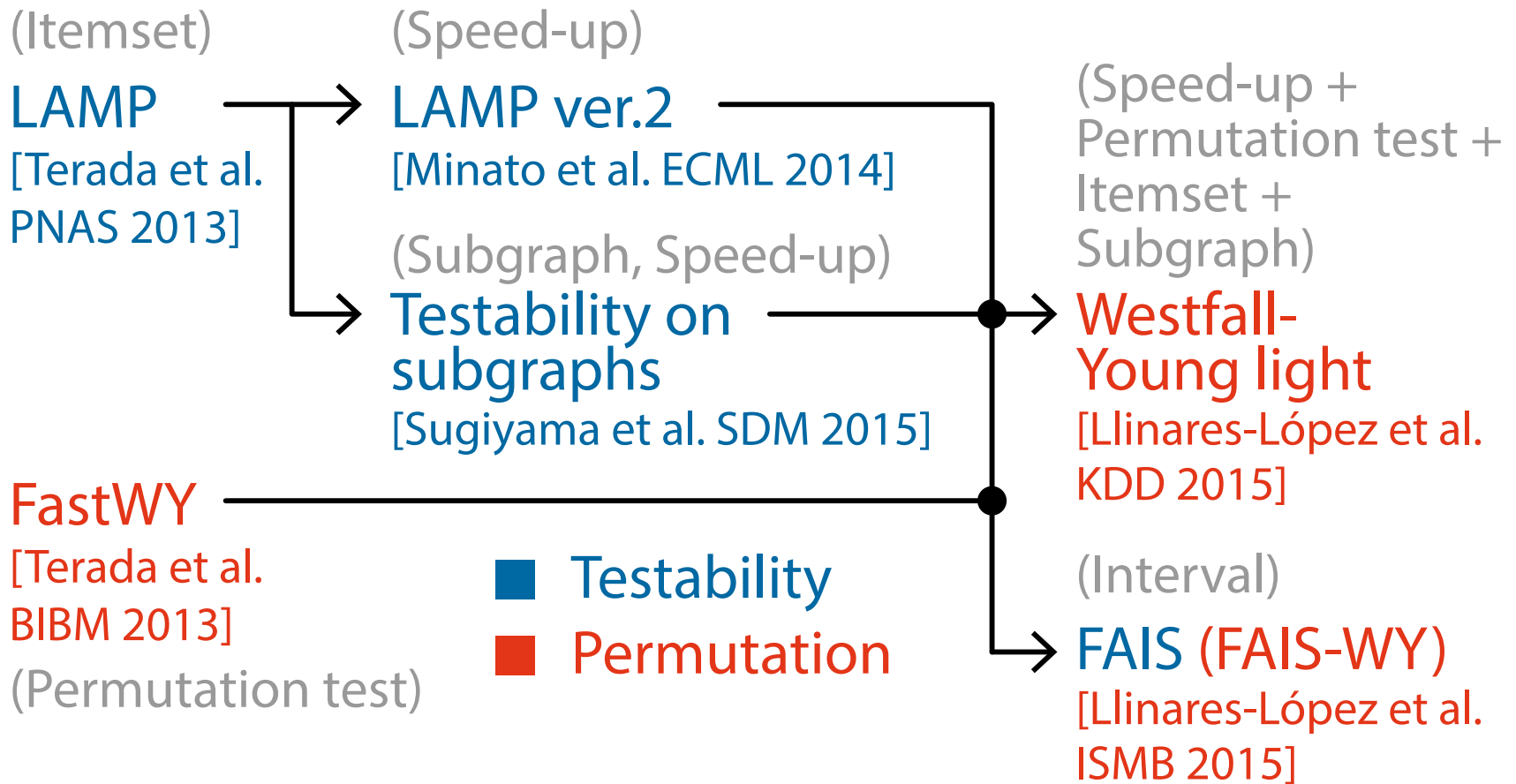
FWER Is Still Too Low!



Take Dependencies into Account

- **Problem:** Dependencies between patterns are not considered
- **Solution:** Permutation test
 - Repeat random permutation of class labels ($10^3 \sim 10^4$ times)
 - Get the null distribution of p -values
 - The optimal correction factor can be obtained

State-of-the-art



Westfall-Young Permutation

1. Randomly permute class labels
2. Compute p -values for all patterns using the permuted class labels
3. Find the minimum p -value p_{\min} among them
 - $FP > 0 \iff p_{\min} < \delta$
 - FP: Number of false positives
4. Repeat steps 1 to 3 h times and obtain $p_{\min}^1, p_{\min}^2, \dots, p_{\min}^h$
 - $FWER(\delta) \approx |\{i : p_{\min}^i \leq \delta\}| / h$
5. δ^* is the α -quantile of $p_{\min}^1, p_{\min}^2, \dots, p_{\min}^h$

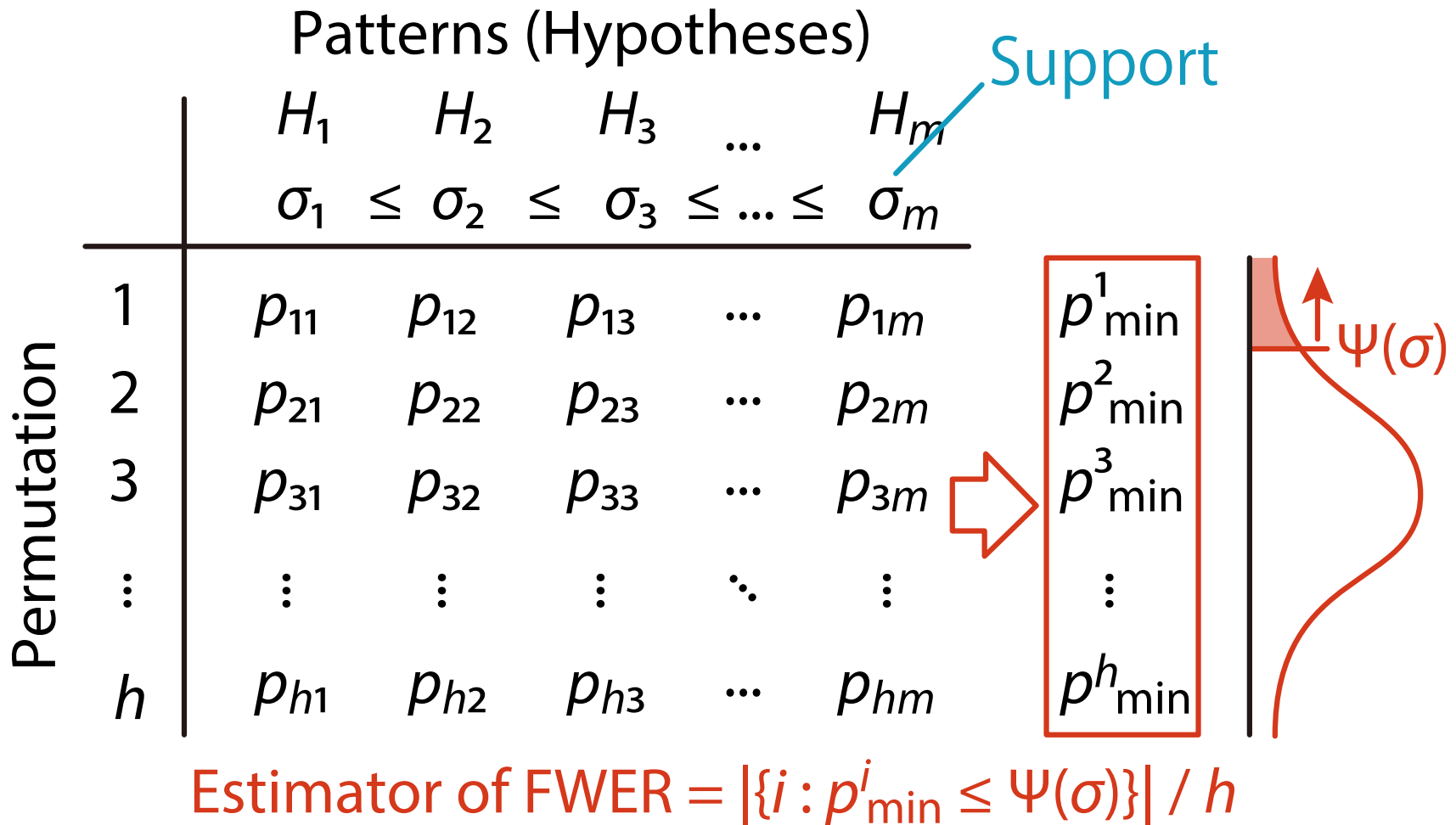
Westfall-Young Permutation

		Patterns (Hypotheses)						
		H_1	H_2	H_3	...	H_m		
Permutation	1	p_{11}	p_{12}	p_{13}	...	p_{1m}	<div>Sort and find α-quantile</div> <div> p_{\min}^1 p_{\min}^2 p_{\min}^3 \vdots p_{\min}^h </div> <div> </div>	
	2	p_{21}	p_{22}	p_{23}	...	p_{2m}		
	3	p_{31}	p_{32}	p_{33}	...	p_{3m}		
	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots		
	h	p_{h1}	p_{h2}	p_{h3}	...	p_{hm}		

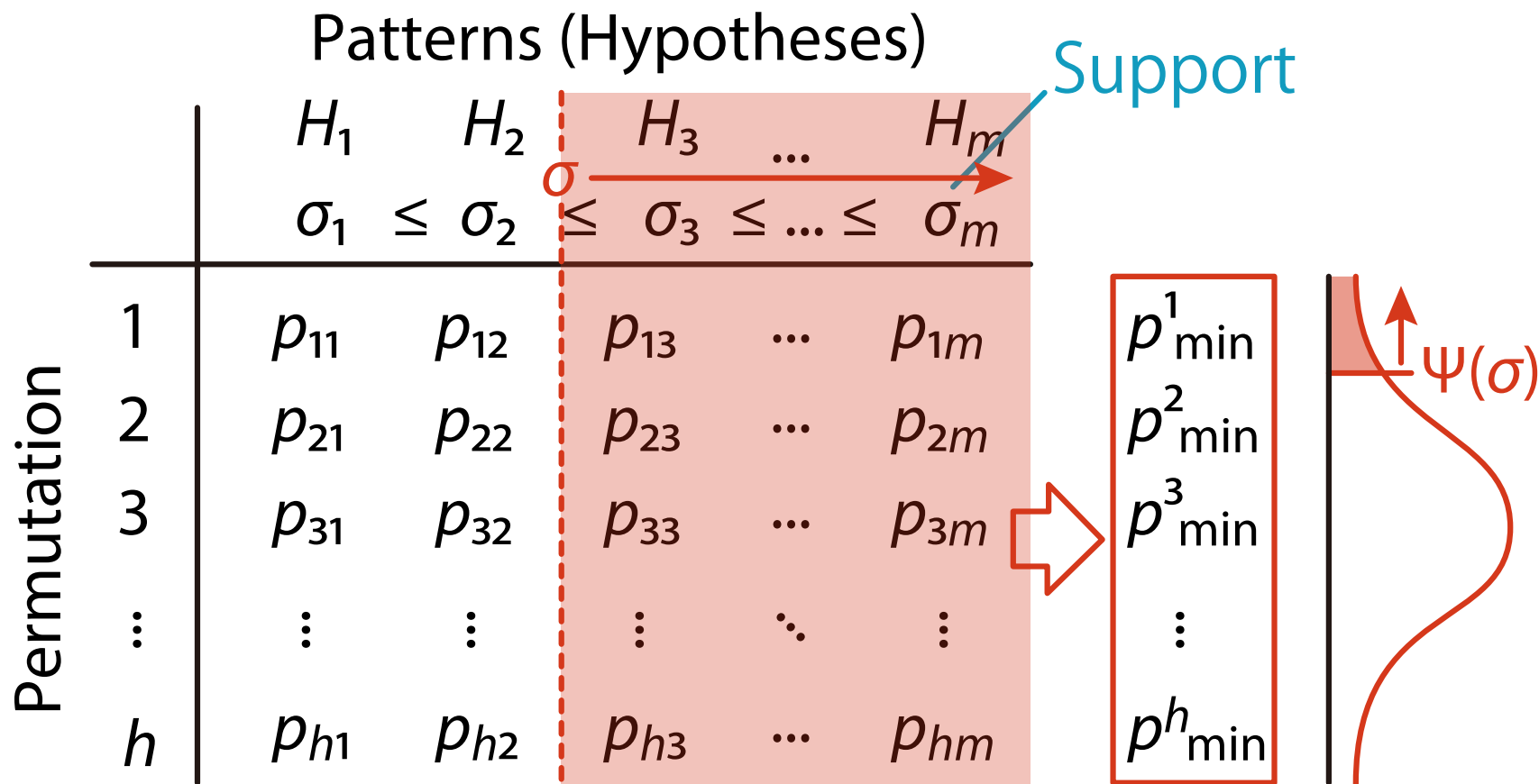
Using Support for Estimating FWER

		Patterns (Hypotheses)					Support Sort and find α -quantile
		H_1	H_2	H_3	...	H_m	
		σ_1	$\leq \sigma_2$	$\leq \sigma_3$	$\leq \dots \leq$	σ_m	
Permutation	1	p_{11}	p_{12}	p_{13}	...	p_{1m}	p_{\min}^1
	2	p_{21}	p_{22}	p_{23}	...	p_{2m}	p_{\min}^2
	3	p_{31}	p_{32}	p_{33}	...	p_{3m}	p_{\min}^3
	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	h	p_{h1}	p_{h2}	p_{h3}	...	p_{hm}	p_{\min}^h

Estimating FWER



Estimating FWER

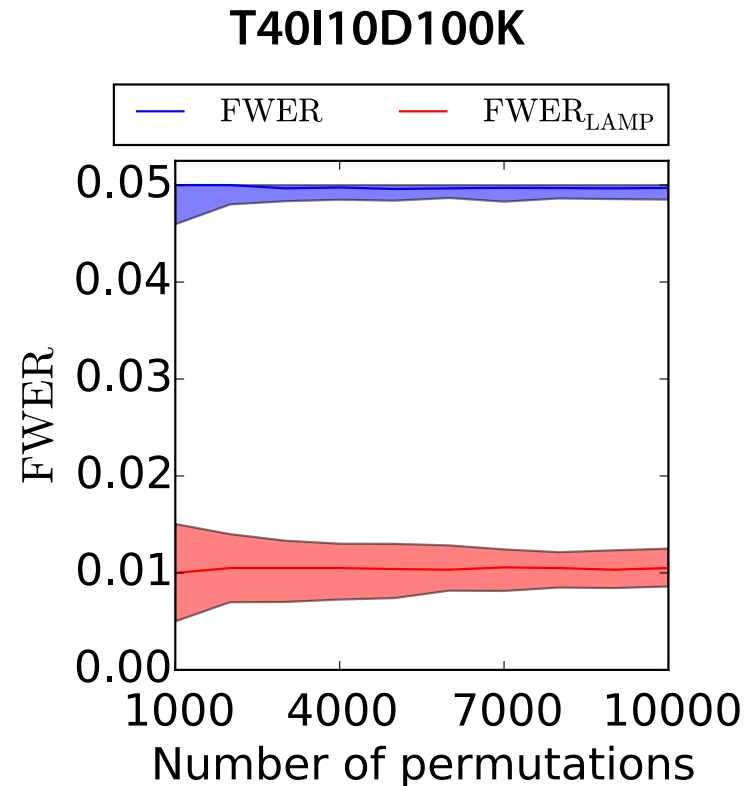
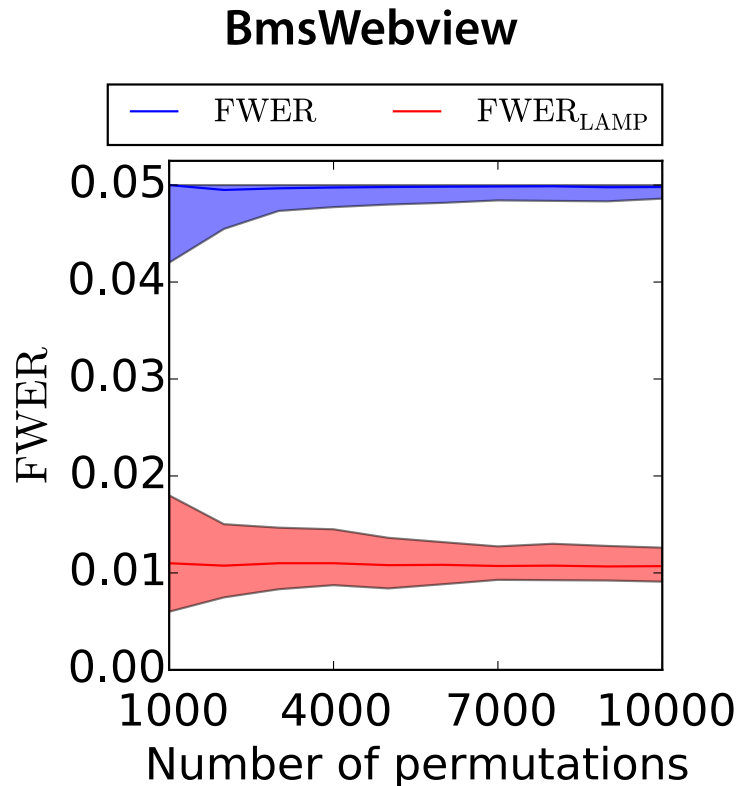


$$\text{Estimator of FWER} = |\{i : p_{\min}^i \leq \Psi(\sigma)\}| / h$$

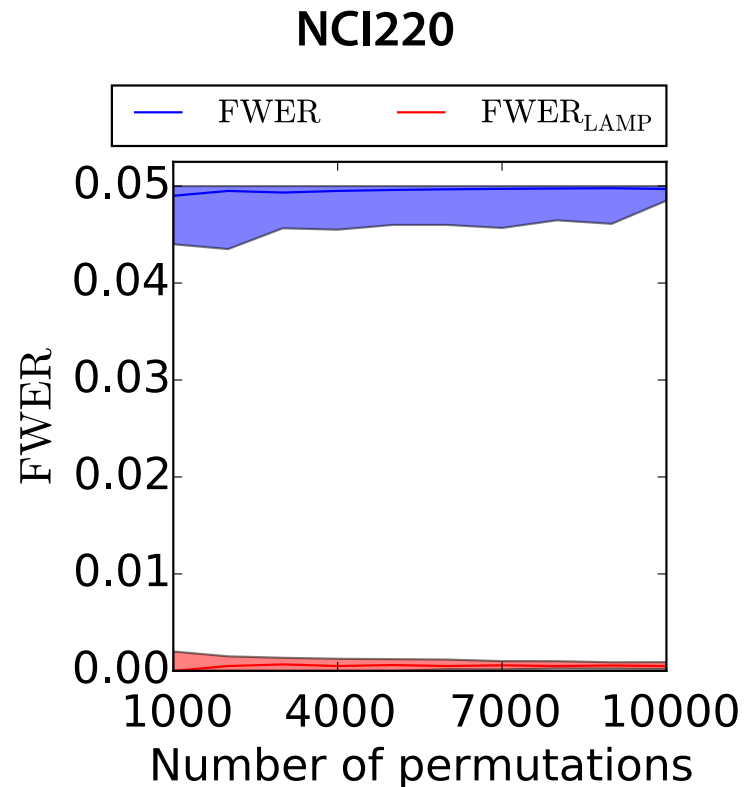
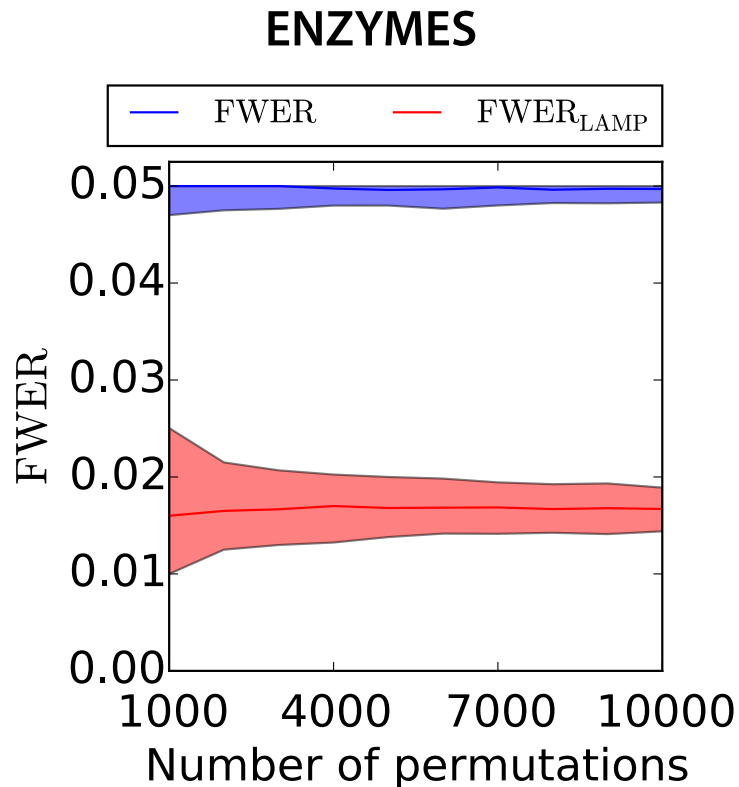
“Westfall-Young light”

- Precompute h permuted labels; $\sigma \leftarrow 1$; $p_{\min}^i \leftarrow 1$
- **Westfall-Young light** does the following whenever a miner (like LCM) finds a new frequent pattern H :
 - **for** $i \leftarrow 1$ **to** h **do**:
 - $p^i \leftarrow$ the p -value of H for i th permutation
 - $p_{\min}^i \leftarrow \min\{p_{\min}^i, p^i\}$
 - $\text{FWER} \leftarrow |\{i : p_{\min}^i \leq \Psi(\sigma)\}| / h$
// $\Psi(\sigma)$ is the min. achievable p -value at σ
 - **while** $\text{FWER} > \alpha$ **do**:
 - $\sigma \leftarrow \sigma + 1$ // σ is the **minimum support**
 - $\text{FWER} \leftarrow |\{i : p_{\min}^i \leq \Psi(\sigma)\}| / h$
 - Go children of H

FWER in Itemset Mining



FWER in Subgraph Mining



Outline

- (Discriminative) Pattern mining
- Statistical hypothesis testing of patterns
- Multiple hypothesis testing in pattern mining
- Testable patterns
- Permutation testing in pattern mining
- Conclusion

Conclusion

- The area of **significant pattern mining** is emerging
 - Find **statistically significant combinatorial patterns** while controlling false positive rate
- Pattern mining, a classical yet central topic in data mining, can be enriched by introducing statistical assessment
 - Can be applied in scientific fields such as biology

Appendix

Papers about Testability

- Tarone, R.E.:
A modified Bonferroni method for discrete data
Biometrics (1990)
- [LAMP] Terada, A., Okada-Hatakeyama, M., Tsuda, K., Sese, J.:
Statistical significance of combinatorial regulations,
Proc. Natl. Acad. Sci. USA (2013).
- [LAMP ver.2] Minato, S., Uno, T., Tsuda, K., Terada, A., Sese, J.:
**Fast Statistical Assessment for Combinatorial Hypotheses
Based on Frequent Itemset Mining**
ECML PKDD 2014
- Sugiyama, M., Llinares, F., Kasenburg, N., Borgwardt, K.:
Significant Subgraph Mining with Multiple Testing Correction,
SIAM SDM 2015

Papers about Permutation Testing

- Westfall, P. H. and Young, S. S.
Resampling-based multiple testing: Examples and methods for p -value adjustment
John Wiley & Sons (1993)
- [FastWY] Terada, A. and Tsuda, K. and Sese, J.:
Fast Westfall-Young permutation procedure for combinatorial regulation discovery,
IEEE BIBM 2013
- [Westfall-Young light] Llinares, F., Sugiyama, M., Papaxanthos, L., Borgwardt, K.:
Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing, ACM SIGKDD 2015

Frequent Itemset Miners

- **[Apriori]** Agrawal, R. and Imieliński, T. and Swami, A.:
Mining association rules between sets of items in large databases,
ACM SIGMOD 1993
- **[FP-Growth]** Han, J. and Pei, J. and Yin, Y.:
Mining frequent patterns without candidate generation,
ACM SIGMOD 2000
- **[LCM]** Uno, T. and Asai, T. and Uchida, Y. and Arimura, H.:
An efficient algorithm for enumerating closed patterns in transaction databases,
DS 2004
(won FIMI'04 competition)

Frequent Subgraph Miners

- **[AGM]** Inokuchi, A. and Washio, T. and Motoda, H.:
An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, PKDD 2000
- **[gSpan]** Yan, X. and Han, J.:
gSpan: Graph-based substructure pattern mining, ICDM 2002
- **[GASTON]** Nijssen, S. and Kok, J. N.:
A Quickstart in Frequent Structure Mining Can Make a Difference, KDD 2004
- **(comparison)** Wörlein, M. and Meinl, T. and Fischer, I. and Philippsen, M.
A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, and Gaston, PKDD 2005
 - We used GASTON as it is the fastest