# Constraint-based Pattern Sampling

**Arnaud Soulet**

**Université de Tours, LIFAT, Blois**

CAVIAR
Contraintes et Apprentissage

**8ème journée CAVIAR**
**8 avril 2025**

université de TOURS

LIFAT EA 6300

What is it?

# Constraint-based Pattern Sampling


**Arnaud Soulet**

**Université de Tours, LIFAT, Blois**


CAVIAR
Contraintes et Apprentissage

université de TOURS

LIFAT EA 6300

**8ème journée CAVIAR**
**8 avril 2025**

# Frequent pattern mining

*[Agrawal et al., 93;Mannila and Toivonen, 97]*

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

Itemset FGH with
$freq(FGH, \mathcal{D}) = 3$

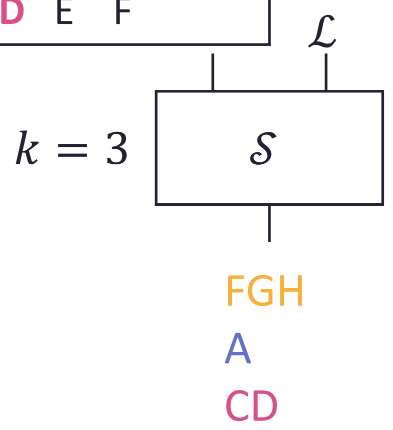
**Discovering relevant local correlations in itemset language $\mathcal{L} = 2^{\{A,B,C,D,E,F,G,H\}}$ considering the dataset $\mathcal{D}$**

- **Drawbacks of pattern mining:**
  1. Long response time
  2. Threshold definition of constraints: $freq(X, \mathcal{D}) \geq \gamma$
  3. Overwhelming number of mined patterns

# Pattern sampling with frequency:
*[Al Hasan et al., 09;Boley et al., 11]*

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C **F G H** |
| t3 | **A** B **C D** E **F G H** |
| t4 | B E **F G H** |
| t5 | **C D** E F |




**Each pattern $X \in \mathcal{L}$ is drawn with a probability proportional to its frequency $\boldsymbol{freq(X, \mathcal{D})}$.**

$FGH$ is 3 times more likely to be drawn than $A$ because its frequency is 3 times greater.

# Pattern sampling with frequency: The two-step random procedure *[Boley et al., 11]*

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

**How do you generate a sample without extracting all the frequent patterns?**

| Itemsets |
|---|
| A |
| B, B |
| … |
| CD, CD |
| … |
| FGH, FGH, FGH |
| … |
| ABCDEFGH |

# Pattern sampling with frequency: The two-step random procedure *[Boley et al., 11]*

$\mathcal{D}$

| Tid | Items | ω |
|---|---|---|
| t1 | D G H | 7 |
| t2 | C F G H | 15 |
| t3 | A B C D E F G H | 255 |
| t4 | B E F G H | 31 |
| t5 | C D E F | 15 |

⓿

| Tid | Itemsets |
|---|---|
| t1 | D, G, H, DG, DH, GH, DGH |
| t2 | C, F, G, H, CF, CG, CH, FG,..., **FGH** |
| t3 | **A**, B, C, D, E, F, G, H, **CD**,..., **FGH** |
| t4 | B, E, F, G, H, BE, BF, BG,..., **FGH** |
| t5 | C, D, E, F, **CD**, CE, CF, DE, DF, EF,... |

**⓿ Calculate ω the number of itemsets per transaction**

# Pattern sampling with frequency: The two-step random procedure *[Boley et al., 11]*

$\mathcal{D}$

| Tid | Items | | | | | | | | ω |
|---|---|---|---|---|---|---|---|---|---|
| t1 | | | | D | | | G | H | 7 |
| t2 | | | C | | | F | G | H | 15 |
| t3 | A | B | C | D | E | F | G | H | 255 |
| t4 | | B | | | E | F | G | H | 31 |
| t5 | | | C | D | E | F | | | 15 |

➊ →

| Tid | Itemsets |
|---|---|
| t1 | D, G, H, DG, DH, GH, DGH |
| t2 | C, F, G, H, CF, CG, CH, FG,…, **FGH** |
| t3 | **A**, B, C, D, E, F, G, H, **CD**,…, **FGH** |
| t4 | B, E, F, G, H, BE, BF, BG,…, **FGH** |
| t5 | C, D, E, F, **CD**, CE, CF, DE, DF, EF,… |

**➊ Draw a transaction $t \in \mathcal{D}$ with a probability proportional to the number of itemsets ω contained in $t$**

# Pattern sampling with frequency: The two-step random procedure *[Boley et al., 11]*

$\mathcal{D}$

| Tid | Items | ω |
|---|---|---|
| t1 | D G H | 7 |
| t2 | C F G H | 15 |
| t3 | A B C D E F G H | 255 |
| t4 | B E F G H | 31 |
| t5 | C D E F | 15 |

➊ →

| Tid | Itemsets |
|---|---|
| t1 | D, G, H, DG, DH, GH, DGH |
| t2 | C, F, G, H, CF, CG, CH, FG,…, FGH |
| t3 | A, B, C, D, E, F, G, H, CD,…, FGH |
| t4 | B, E, F, G, H, BE, BF, BG,…, FGH |
| t5 | C, D, E, F, CD, CE, CF, DE, DF, EF,… |

➋ →

| Itemsets |
|---|
| A |
| B, B |
| … |
| CD, CD |
| … |
| FGH, FGH, FGH |
| … |
| ABCDEFGH |

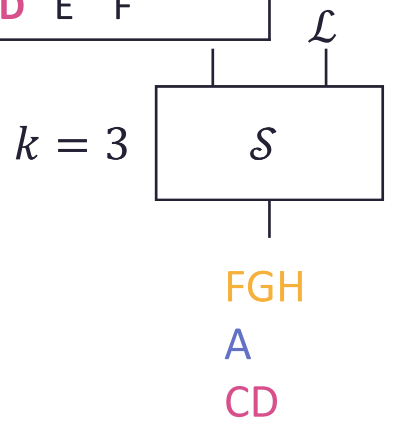
➊ **Draw a transaction $t \in \mathcal{D}$ with a probability proportional to the number of itemsets ω contained in $t$**

➋ **Draw uniformly an itemset from $t$**

# Pattern sampling with frequency: Interests for user-centric pattern mining

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




- **Controlled size of the pattern sample**
- **Very fast extraction with the two-step random procedure**
- **Useful profiling of the dataset for different mining processes:**
  - Feature construction *[Boley et al., 11]*
  - Outlier detection *[Giacometti et al., 16]*
  - Interactive pattern mining *[Dzyuba et al., 16;Hien et al., 23]*

Why????

# Constraint-based Pattern Sampling

**Arnaud Soulet**

**Université de Tours, LIFAT, Blois**

CAVIAR
Contraintes et Apprentissage

université de TOURS

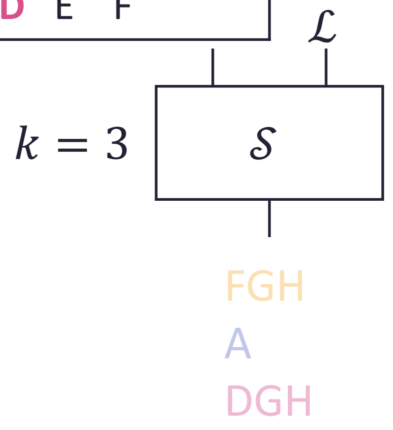
LIFAT EA 6300

**8ème journée CAVIAR**
**8 avril 2025**

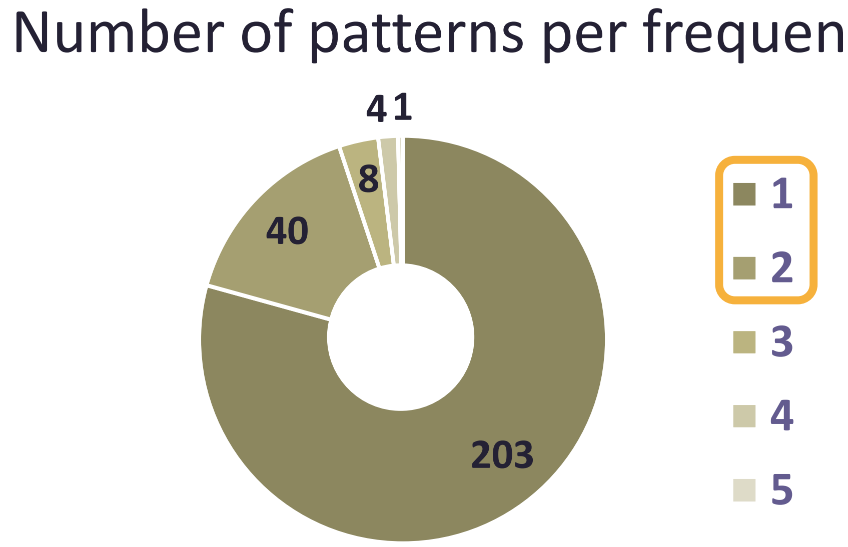
# Pattern sampling with frequency: Curse of the long tail *[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




**Despite the draw bias, the sampling focuses on non-frequent patterns :**




86% of the drawn patterns have a frequency of less than 2!

# Pattern sampling with frequency: Curse of the long tail *[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| t1 | | | | D | | | G | H |
| t2 | | | C | | | F | G | H |
| t3 | A | B | C | D | E | F | G | H |
| t4 | | B | | | E | F | G | H |
| t5 | | | C | D | E | F | | |

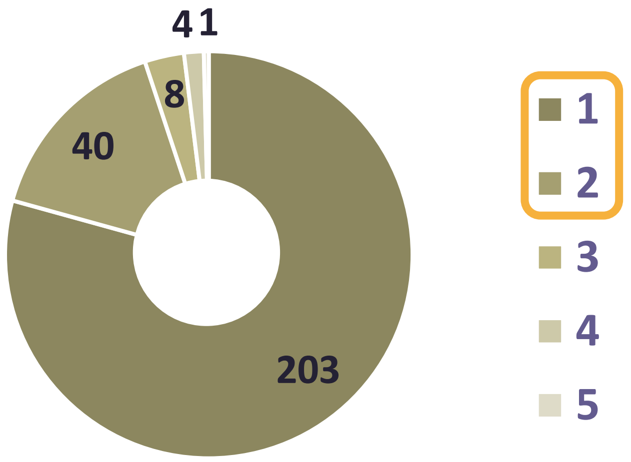
$\mathcal{L}$

$k = 3$ $\mathcal{S}$

FGH
A
DGH

**Despite the draw bias, the sampling focuses on non-frequent patterns :**

Number of patterns per frequency



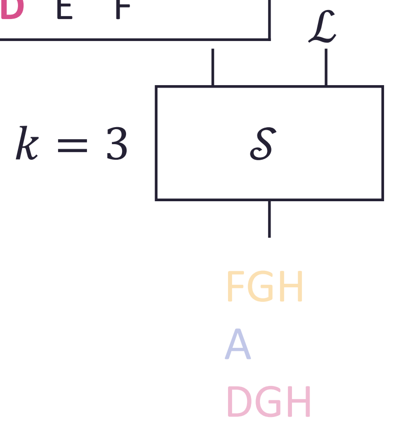

73% of the drawn patterns have a frequency of less than 2!

Bias amplified with $freq(x, \mathcal{D})^2$!!!

# Pattern sampling with frequency: Curse of the long tail *[Diop et al., 2018]*

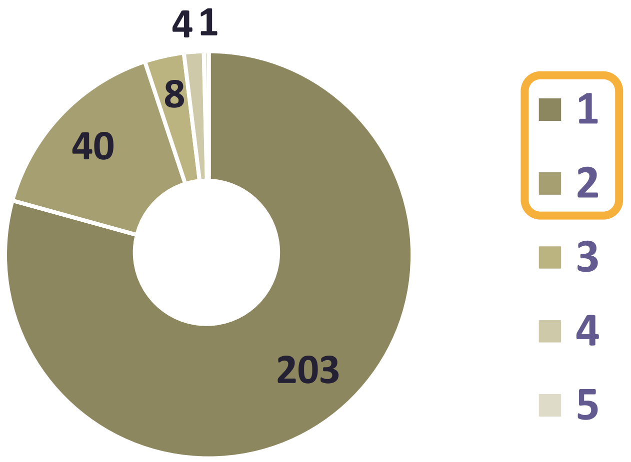
$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




**Despite the draw bias, the sampling focuses on non-frequent patterns :**

Number of patterns per frequency



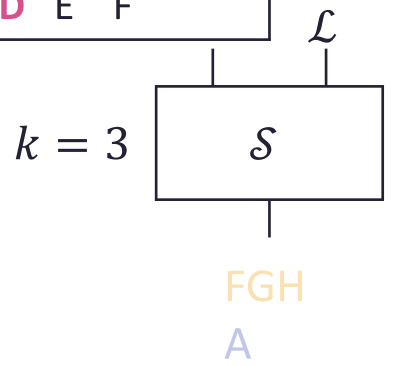

53% of the drawn patterns have a frequency of less than 2!

Bias amplified with $freq(x, \mathcal{D})^3$!!!!!!

# Pattern sampling with frequency: Curse of the long tail *[Diop et al., 2018]*

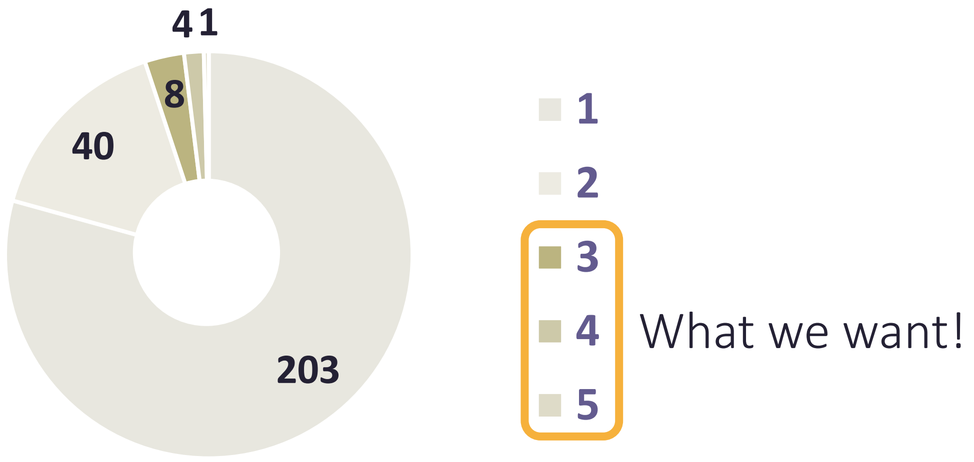
$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




**Despite the draw bias, the sampling focuses on non-frequent patterns :**

Number of patterns per frequency




**Our goal: Pattern sampling with constraint for avoiding non-frequent itemsets**

# Short Pattern Sampling

**Arnaud Soulet**

**Université de Tours, LIFAT, Blois**

CAVIAR
Contraintes et Apprentissage

**8ème journée CAVIAR**
**8 avril 2025**

université de TOURS

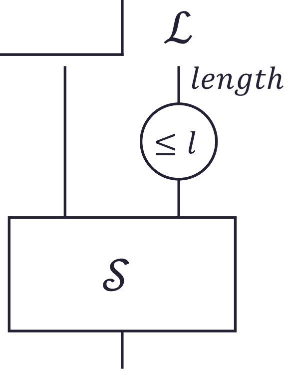
LIFAT EA 6300

# Pattern sampling with length constraint

*[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




- **Key idea:** short patterns are more frequent

➔ Maximum length constraint

# Pattern sampling with length constraint

*[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items | ω |
|---|---|---|
| t1 | D G H | 7 |
| t2 | C F G H | 14 |
| t3 | A B C D E F G H | 92 |
| t4 | B E F G H | 25 |
| t5 | C D E F | 14 |

➓ ⟷

ρ

| Tid | len.=1 | len.=2 | len.=3 |
|---|---|---|---|
| t1 | 3 (D, G, H) | 3 (DG, DH, GH) | 1 (DGH) |
| t2 | 4 | 6 | 4 (FGH) |
| t3 | 8 (A) | 28 (CD) | 56 (FGH) |
| t4 | 5 | 10 | 10 (FGH) |
| t5 | 4 | 6 (CD) | 4 |

$$\boldsymbol{length(X) \leq 3}$$

**⓪ Calculate ω the number of itemsets per transaction and ρ the number of itemsets per transaction and per length**

# Pattern sampling with length constraint

*[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items | ω |
|---|---|---|
| t1 | D G H | 7 |
| t2 | C F G H | 14 |
| t3 | A B C D E F G H | 92 |
| t4 | B E F G H | 25 |
| t5 | C D E F | 14 |

➊

ρ

| Tid | len.=1 | len.=2 | len.=3 |
|---|---|---|---|
| t1 | 3 (D, G, H) | 3 (DG, DH, GH) | 1 (DGH) |
| t2 | 4 | 6 | 4 (FGH) |
| t3 | 8 (A) | 28 (CD) | 56 (FGH) |
| t4 | 5 | 10 | 10 (FGH) |
| t5 | 4 | 6 (CD) | 4 |

➊ **Draw a transaction $t \in \mathcal{D}$ w.r.t ω**

# Pattern sampling with length constraint

*[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items | ω |
|---|---|---|
| t1 | D G H | 7 |
| t2 | C F G H | 14 |
| t3 | A B C D E F G H | 92 |
| t4 | B E F G H | 25 |
| t5 | C D E F | 14 |

❶

ρ

❷

| Tid | len.=1 | len.=2 | len.=3 |
|---|---|---|---|
| t1 | 3 (D, G, H) | 3 (DG, DH, GH) | 1 (DGH) |
| t2 | 4 | 6 | 4 (FGH) |
| t3 | 8 (A) | 28 (CD) | 56 (FGH) |
| t4 | 5 | 10 | 10 (FGH) |
| t5 | 4 | 6 (CD) | 4 |

❶ **Draw a transaction $t \in \mathcal{D}$ w.r.t ω**

❷ **Draw a length $l$ w.r.t ρ**

# Pattern sampling with length constraint

*[Diop et al., 2018]*

$\mathcal{D}$

| Tid | Items | ω |
|---|---|---|
| t1 | D G H | 7 |
| t2 | C F G H | 14 |
| t3 | A B C D E F G H | 92 |
| t4 | B E F G H | 25 |
| t5 | C D E F | 14 |

➊

ρ

➋

| Tid | len.=1 | len.=2 | len.=3 |
|---|---|---|---|
| t1 | 3 (D, G, H) | 3 (DG, DH, GH) | 1 (DGH) |
| t2 | 4 | 6 | 4 (FGH) |
| t3 | 8 (A) | 28 (CD) | 56 (FGH) |
| t4 | 5 | 10 | 10 (FGH) |
| t5 | 4 | 6 (CD) | 4 |

➌

| Itemsets |
|---|
| A |
| B, B |
| … |
| CD, CD |
| … |
| FGH, FGH, FGH |
| … |

➊ **Draw a transaction $t \in \mathcal{D}$ w.r.t ω**

➋ **Draw a length $l$ w.r.t ρ**

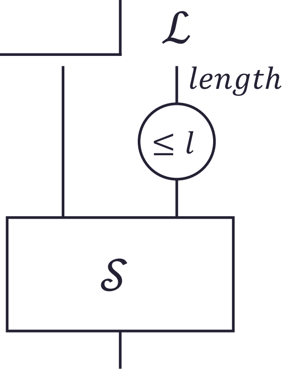
➋ **Draw uniformly an itemset of length $l$ in $t$**

# Pattern sampling with length constraint

*[Diop et al., 2018]*

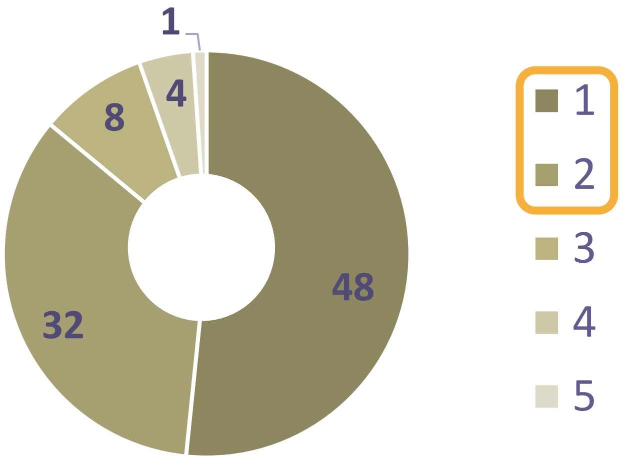
$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




**Despite the length constraint ($length(X) \leq 3$), the sampling focuses on non-frequent patterns :**

Number of patterns per frequency




73% of the drawn patterns have a frequency of less than 2!

# Frequent Pattern Sampling

**Arnaud Soulet**

**Université de Tours, LIFAT, Blois**

CAVIAR
Contraintes et Apprentissage

**8ème journée CAVIAR**
**8 avril 2025**

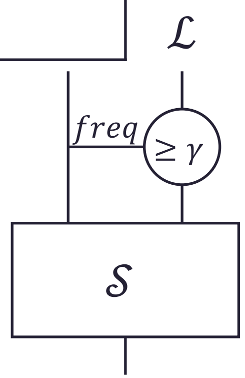
université de TOURS

LIFAT EA 6300

# Challenge of frequent pattern sampling

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




- **Families of pattern sampling methods:**
  - ✓ Stochastic methods *[Al Hasan et al., 08]*
  - ✓ Constraint programming *[Dzyuba et al., 16]*
  - ✗ Multi-step random procedure *[Boley et al., 11]*

  Possible to push constraints but poor efficiency

- **Constraint-based multi-step random procedure**
  - ✓ Syntactic constraints can be pushed into the language $\mathcal{L}$ *[Diop et al., 18]*
  - ✗ Frequency-based constraint

**How to push the frequency constraint into the sampling procedure?**

# The old pots make the best soups

*[Bonchi et al., 03]*

## ExAnte: Anticipated Data Reduction in Constrained Pattern Mining

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2],
Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

*[El-Hajj et Zaiane, 03]*

## COFI-tree Mining: A New Approach to Pattern Growth with Reduced Candidacy Generation

Mohammad El-Hajj
Department of Computing Science
University of Alberta Edmonton, AB, Canada
mohammad@cs.ualberta.ca

Osmar R. Zaïane
Department of Computing Science
University of Alberta Edmonton, AB, Canada
zaiane@cs.ualberta.ca

### Abstract

*Existing association rule mining algorithms suffer from many problems when mining massive transactional datasets. Some of these major problems are: (1) the repetitive I/O disk scans, (2) the huge computation involved during the candidacy generation, and (3) the high memory dependency. This paper presents the implementation of our frequent itemset mining algorithm, COFI, which achieves its efficiency by applying four new ideas. First, it can mine using a compact memory based data structures. Second, for each frequent item assigned, a relatively small independent tree is built summarizing co-occurrences. Third, clever pruning reduces the search space drastically. Finally, a simple and non-recursive mining process reduces the memory requirements as minimum candidacy generation and counting is needed to generate all relevant frequent patterns.*

posed method uses a pruning technique that dramatically saves the memory space. These relatively small trees are constructed based on a memory-based structure called FP-Trees [11]. This data structure is studied in detail in the following sections. In short, we introduced in [8] the COFI-tree stucture and an algorithm to mine it. In [7] we presented a disk based data structure, inverted matrix, that replaces the memory-based FP-tree and scales the interactive frequent pattern mining significantly. Our contributions in this paper are the introduction of a clever pruning technique based on an interesting property drawn from our top-down approach, and some implementation tricks and issues. We included the pruning in the algorithm of building the tree so that the pruning is done on the fly.

#### 1.1 Problem Statement

The problem of mining association rules over market

# The old pots make the best soups

ExAnte: Anticipated Data Reduction
in Constrained Pattern Mining

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2],
Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

- **Using pre-processing on the dataset for decreasing the search space** ***[Bonchi et al., 03]***
- **Recursive reduction**
  - Anti-monotone constraint : $freq(X, \mathcal{D}) \geq 3$
  - Monotone constraint : $length(X) \geq 4$

# The old pots make the best soups

**ExAnte: Anticipated Data Reduction in Constrained Pattern Mining**

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2], Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

- **Recursive reduction**
  - **Anti-monotone constraint :** $\boldsymbol{freq(X,\mathcal{D}) \geq 3}$
  - Monotone constraint : $length(X) \geq 4$

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

# The old pots make the best soups

## ExAnte: Anticipated Data Reduction in Constrained Pattern Mining

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2], Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

- **Recursive reduction**
  - Anti-monotone constraint : $freq(X, \mathcal{D}) \geq 3$
  - **Monotone constraint :** $\boldsymbol{length(X) \geq 4}$

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

# The old pots make the best soups

## ExAnte: Anticipated Data Reduction in Constrained Pattern Mining

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2], Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

- **Recursive reduction**
  - **Anti-monotone constraint :** $\boldsymbol{freq(X, \mathcal{D}) \geq 3}$
  - Monotone constraint : $length(X) \geq 4$

$\mathcal{D}$

| Tid | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| t1 | | | | D | | | G | H |
| t2 | | | C | | | F | G | H |
| t3 | A | B | C | D | E | F | G | H |
| t4 | | B | | | E | F | G | H |
| t5 | | | C | D | E | F | | |

# The old pots make the best soups

ExAnte: Anticipated Data Reduction in Constrained Pattern Mining

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2], Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

- **Recursive reduction**
  - Anti-monotone constraint : $freq(X,\mathcal{D}) \geq 3$
  - **Monotone constraint : $\boldsymbol{length(X) \geq 4}$**

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

# The old pots make the best soups

## ExAnte: Anticipated Data Reduction in Constrained Pattern Mining

Francesco Bonchi[1,2,3], Fosca Giannotti[1,2], Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it

**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

- **Recursive reduction**
  - **Anti-monotone constraint :** $\boldsymbol{freq(X, \mathcal{D}) \geq 3}$
  - Monotone constraint : $length(X) \geq 4$

$\mathcal{D}$

| Tid | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| t1 | | | | D | | | G | H |
| t2 | | | C | | | F | G | H |
| t3 | A | B | C | D | E | F | G | H |
| t4 | | B | | | E | F | G | H |
| t5 | | | C | D | E | F | | |

# The old pots make the best soups

ExAnte: Anticipated Data Reduction in Constrained Pattern Mining


Francesco Bonchi[1,2,3], Fosca Giannotti[1,2], Alessio Mazzanti[1,3], and Dino Pedreschi[1,3]

[1] Pisa KDD Laboratory*
http://www-kdd.cnuce.cnr.it
[2] ISTI - CNR Area della Ricerca di Pisa, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy
Giannotti@cnuce.cnr.it
[3] Department of Computer Science, University of Pisa
Via F. Buonarroti 2, 56127 Pisa, Italy
{bonchi,mazzanti,pedre}@di.unipi.it



**Abstract.** Constraint pushing techniques have been proven to be effective in reducing the search space in the frequent pattern mining task, and thus in improving efficiency. But while pushing anti-monotone constraints in a level-wise computation of frequent itemsets has been recognized to be always profitable, the case is different for monotone constraints. In fact, monotone constraints have been considered harder to push in the computation and less effective in pruning the search space. In this paper, we show that this prejudice is ill founded and introduce ExAnte, a pre-processing data reduction algorithm which reduces dramatically both the search space and the input dataset in constrained frequent pattern mining. Experimental results show a reduction of orders of magnitude, thus enabling a much easier mining task. ExAnte can be used as a pre-processor with any constrained pattern mining algorithm.

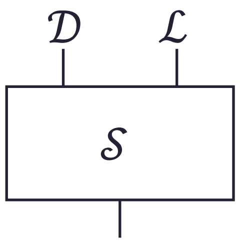

- **Recursive reduction**
  - Anti-monotone constraint : $freq(X,\mathcal{D}) \geq 3$
  - **Monotone constraint : $\boldsymbol{length(X) \geq 4}$**

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

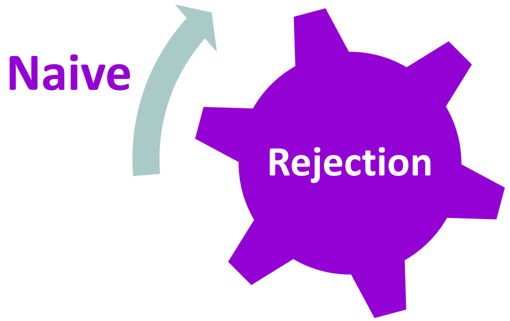
# Key ideas

**Benefiting from existing pattern sampling methods**

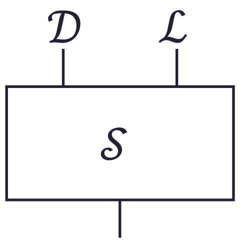



The language $\mathcal{L}$ cannot be changed.

# Key ideas



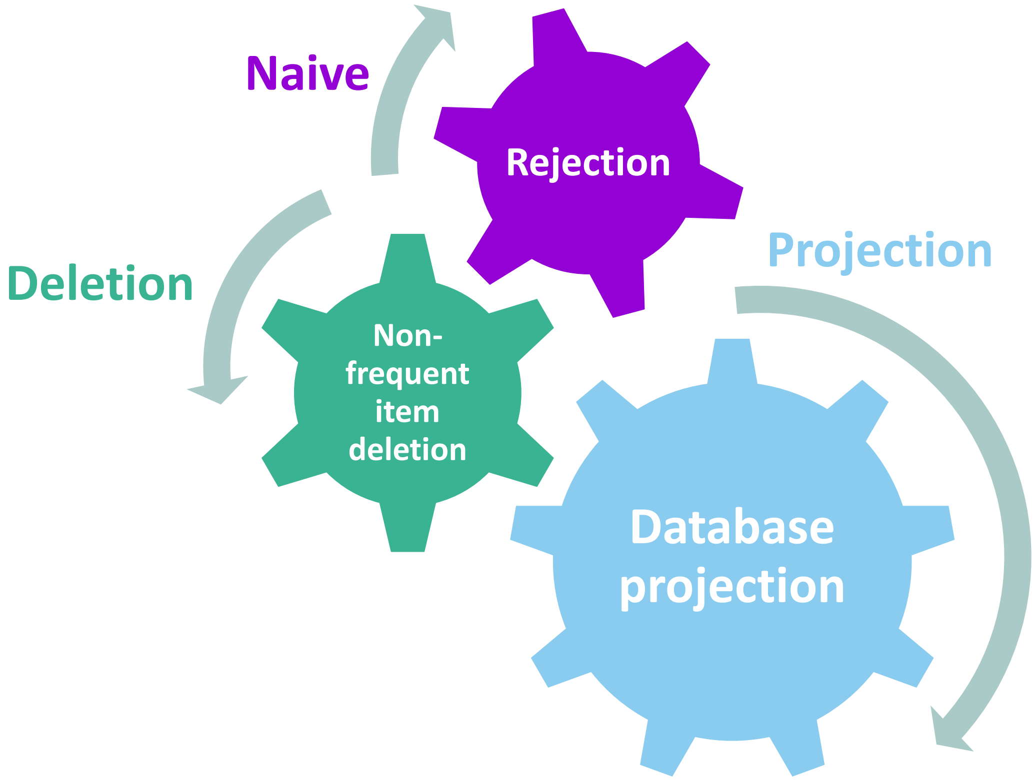

**Benefiting from existing pattern sampling methods**



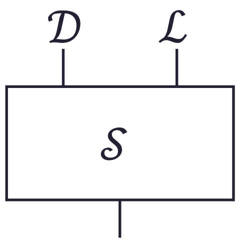

The language $\mathcal{L}$ cannot be changed.

❶ **Filtering the output:** rejection step

# Key ideas




**Benefiting from existing pattern sampling methods**




The language $\mathcal{L}$ cannot be changed.

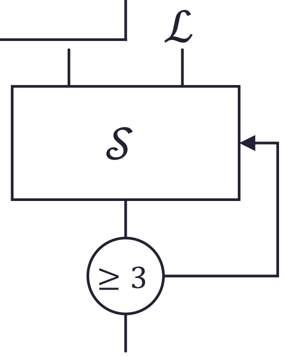
➊ **Filtering the output:** rejection step

➋ **Filtering the dataset $\mathcal{D}$:** deletion + projection

# Naive method: sampling with rejection (1)

$\mathcal{D}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




FGH (3), ~~A~~ (1), ~~CD~~ (2),…

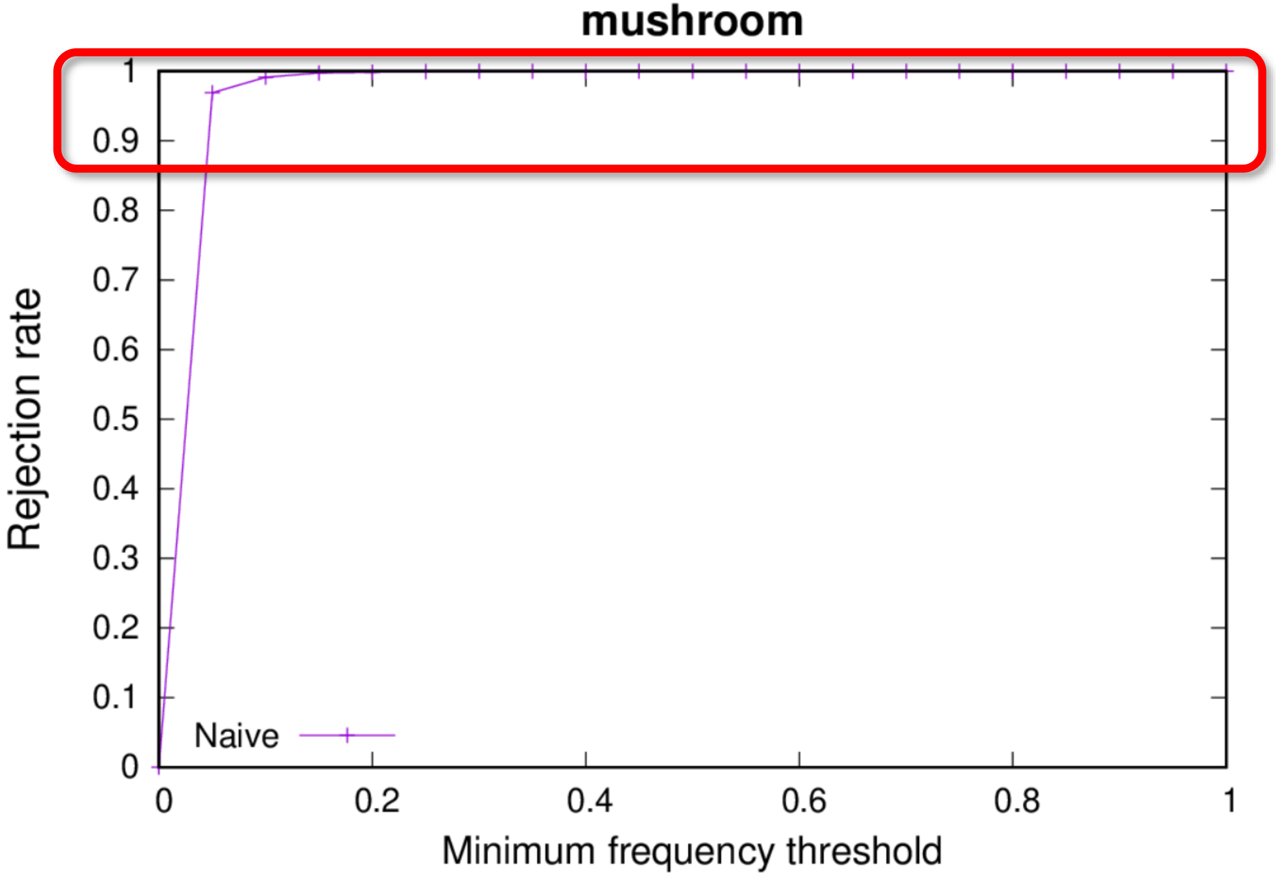
**Principle:**

1. Draw a pattern
2. Compute its frequency
3. Reject it if its frequency is less than $\gamma$

**Rejection rate with $\gamma = 3$: 86%**

# Naive method: sampling with rejection (2)

Lower is better



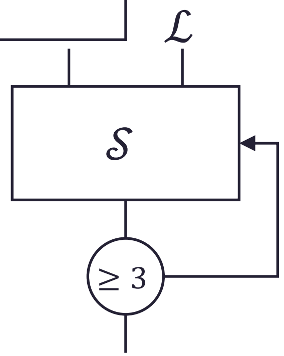

The rejection rate is too high!

**Naive method does not work.**

# Deletion method: discarding non-frequent items (1)

$\mathcal{D}_3$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |




FGH (3), ~~A~~ (1), ~~CD~~ (2),…

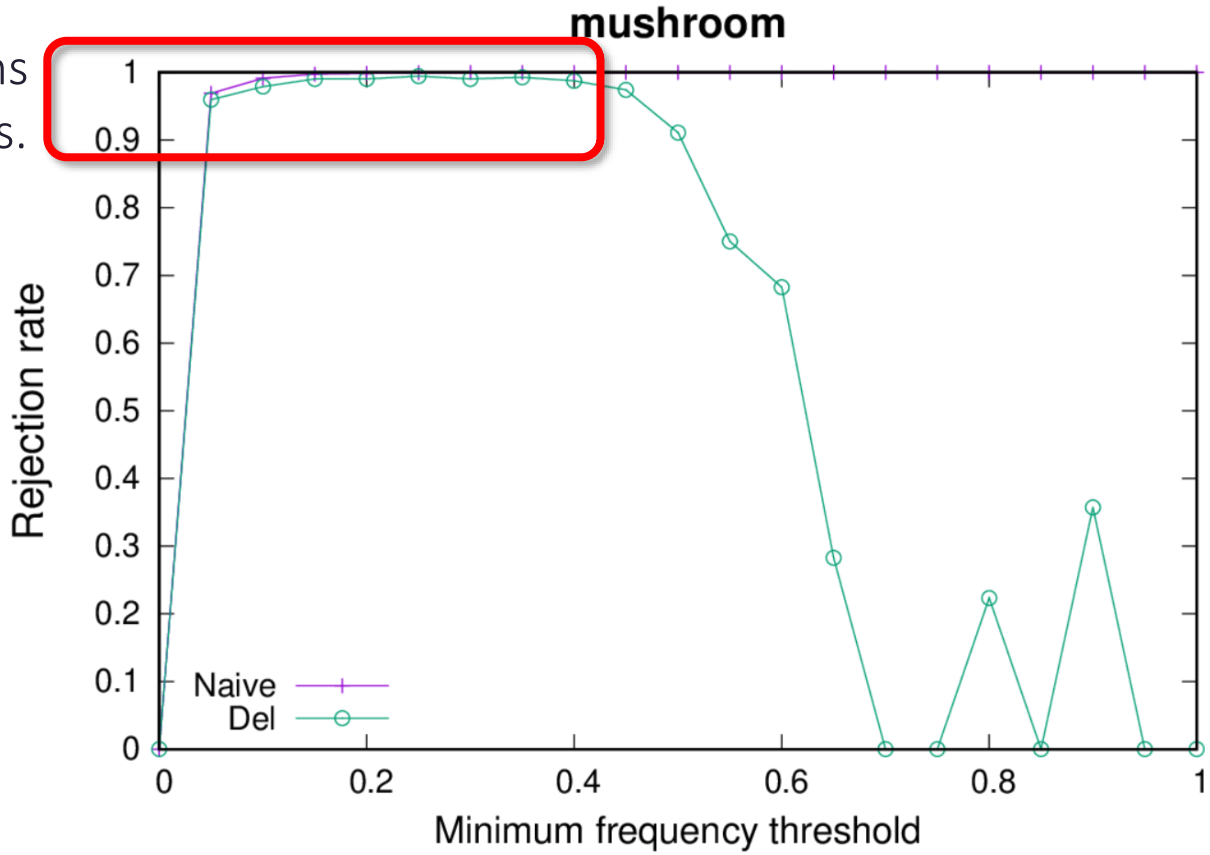
**Principle:** Delete all the non-frequent items (and then, all the itemsets containing at least one non frequent item)

Deletion of A ➔AC, AD, ACD,…

**Rejection rate with $\gamma = 3$: 86% ➔ 52%**

# Deletion method: discarding non-frequent items (2)

The rejection rate remains too high for low thresholds.




**How can this approach be generalized to non-frequent item pairs?**

# Projection method: adding item sampling (1)

$\mathcal{D}^{(C)}$

| Tid | Items | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| t1 | | | | D | | | G | H |
| t2 | | | C | | | F | G | H |
| t3 | A | B | C | D | E | F | G | H |
| t4 | | B | | | E | F | G | H |
| t5 | | | C | D | E | F | | |

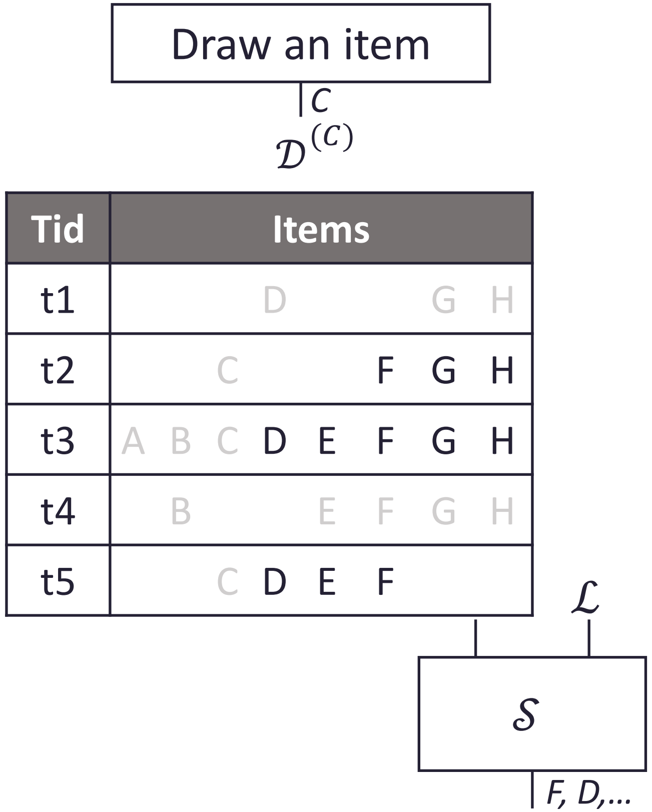
$> C$

Projected database for the item $C$
(considering the lexicographic order)

**Projected database for the item $i$:**

It contains only the transactions where the item $i$ occurs and the items greater than $i$.

# Projection method: adding item sampling (1)

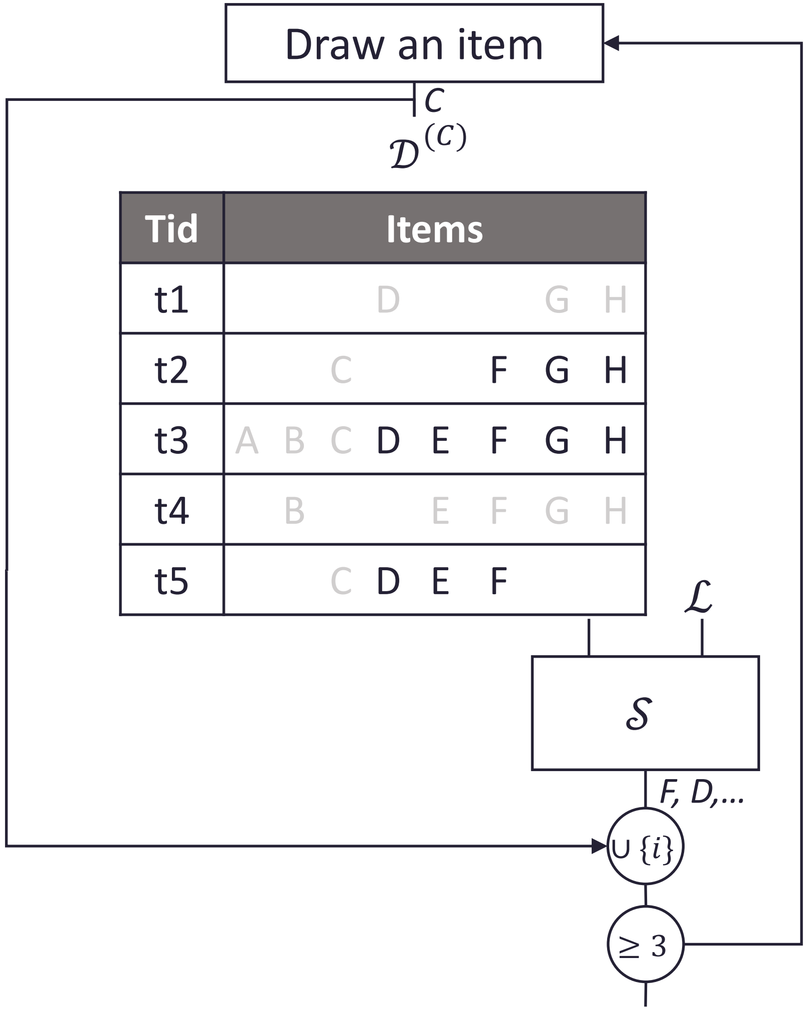



| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

**Principle:**

1. Draw the first item $i$ proportionally to $\omega(i)$
2. Sample an itemset $Y$ in the projected dataset $\mathcal{D}^{(i)}$

# Projection method: adding item sampling (1)




| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

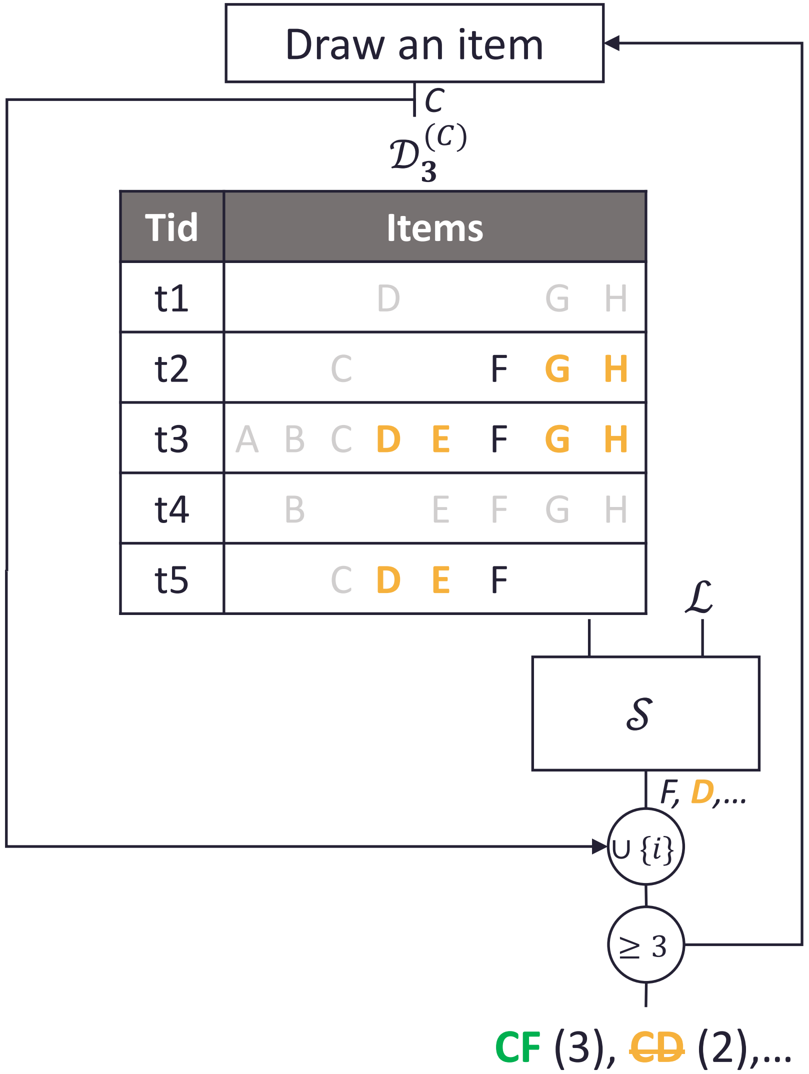
CF (3), ~~CD~~ (2),...

**Principle:**

1. Draw the first item $i$ proportionally to $\omega(i)$
2. Sample an itemset $Y$ in the projected dataset $\mathcal{D}^{(i)}$
3. Return the pattern $\{i\} \cup Y$ (if its fequency $\geq \gamma$)

**Rejection rate with $\gamma = 3$: 86% → 52% → 52%**

# Projection method: adding item sampling (1)




| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F G H |
| t4 | B E F G H |
| t5 | C D E F |

**Principle:**

1. Draw the first item $i$ proportionally to $\omega(i)$
2. Sample an itemset $Y$ in the projected dataset $\mathcal{D}_\gamma^{(i)}$ (with deletion!)
3. Return the pattern $\{i\} \cup Y$ (if its fequency $\geq \gamma$)

**Rejection rate with $\gamma = 3$: 86% → 52% → 0%**

# Projection method: adding item sampling (2)

$\mathcal{D}_3^{(C)}$

| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F **G** **H** |
| t3 | A B C **D** **E** F **G** **H** |
| t4 | B E F G H |
| t5 | C **D** **E** F |

$\omega(C) =$ 6 occurrences

$\mathcal{D}_3^{(D)}$

| Tid | Items |
|---|---|
| t1 | D **G** **H** |
| t2 | C F G H |
| t3 | A B C D **E** **F** **G** **H** |
| t4 | B E F G H |
| t5 | C D **E** **F** |

$\omega(D) =$ 3 occurrences

$\mathcal{D}_3^{(E)}$

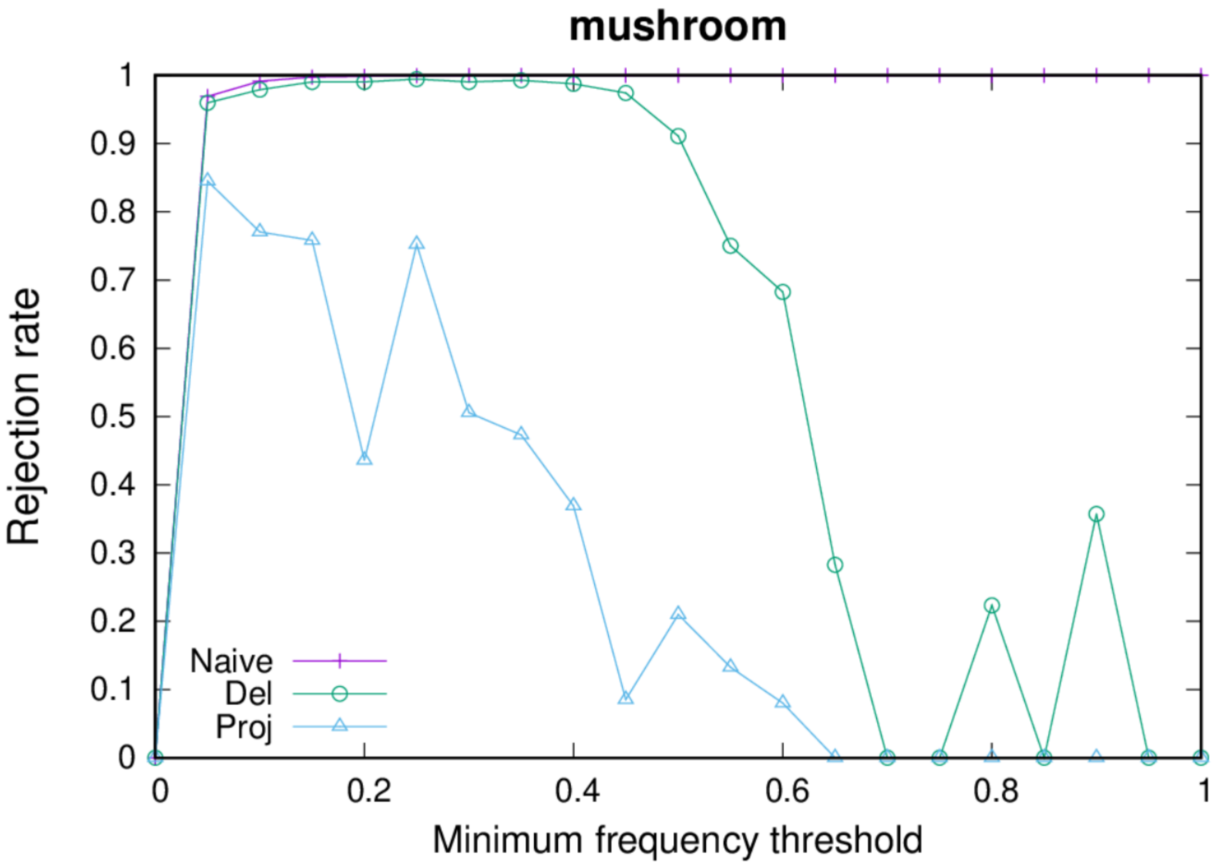
| Tid | Items |
|---|---|
| t1 | D G H |
| t2 | C F G H |
| t3 | A B C D E F **G** **H** |
| t4 | B E F **G** **H** |
| t5 | C D E F |

$\omega(E) =$ 6 occurrences

…

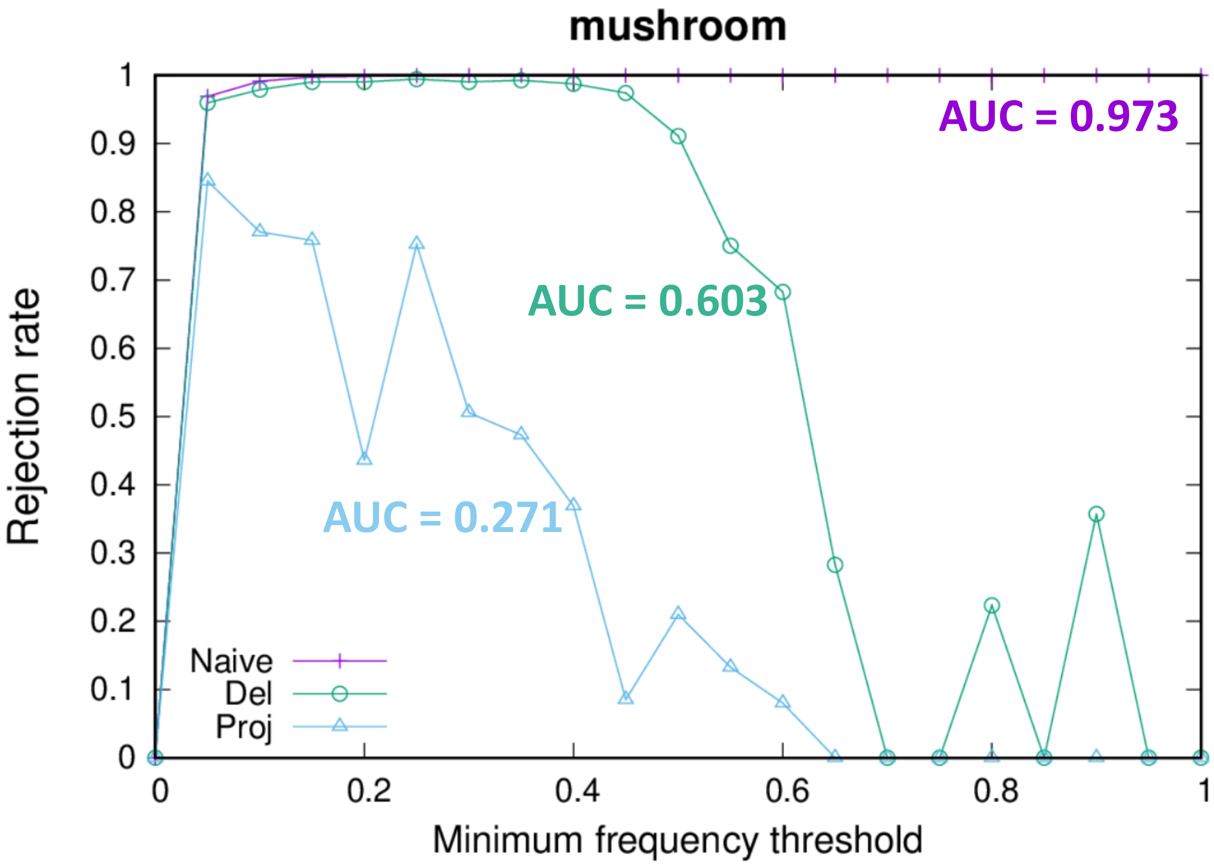
**Example for $\mathcal{D}_3^{(C)}$: $\emptyset$ and F in t2 for leading to C and CF (idem for t3 and t5)**

# Projection method: adding item sampling (3)



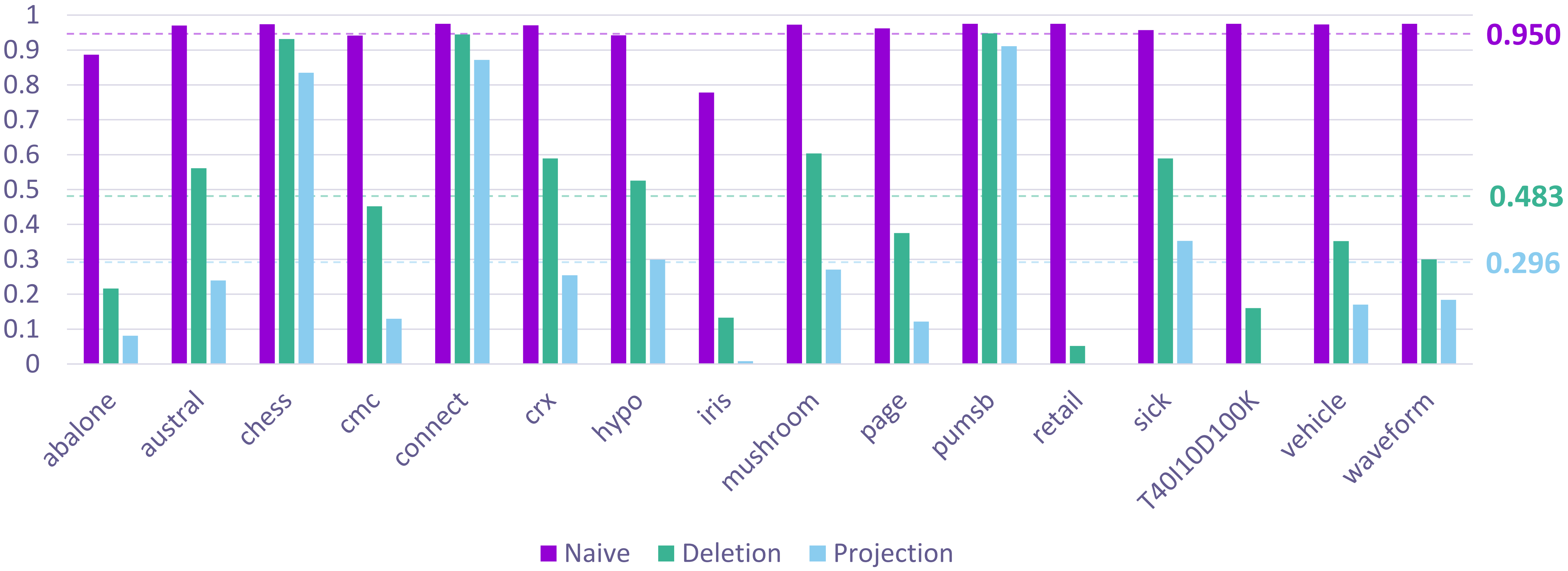

**Projection method works well for any threshold.**

# Projection method: adding item sampling (3)



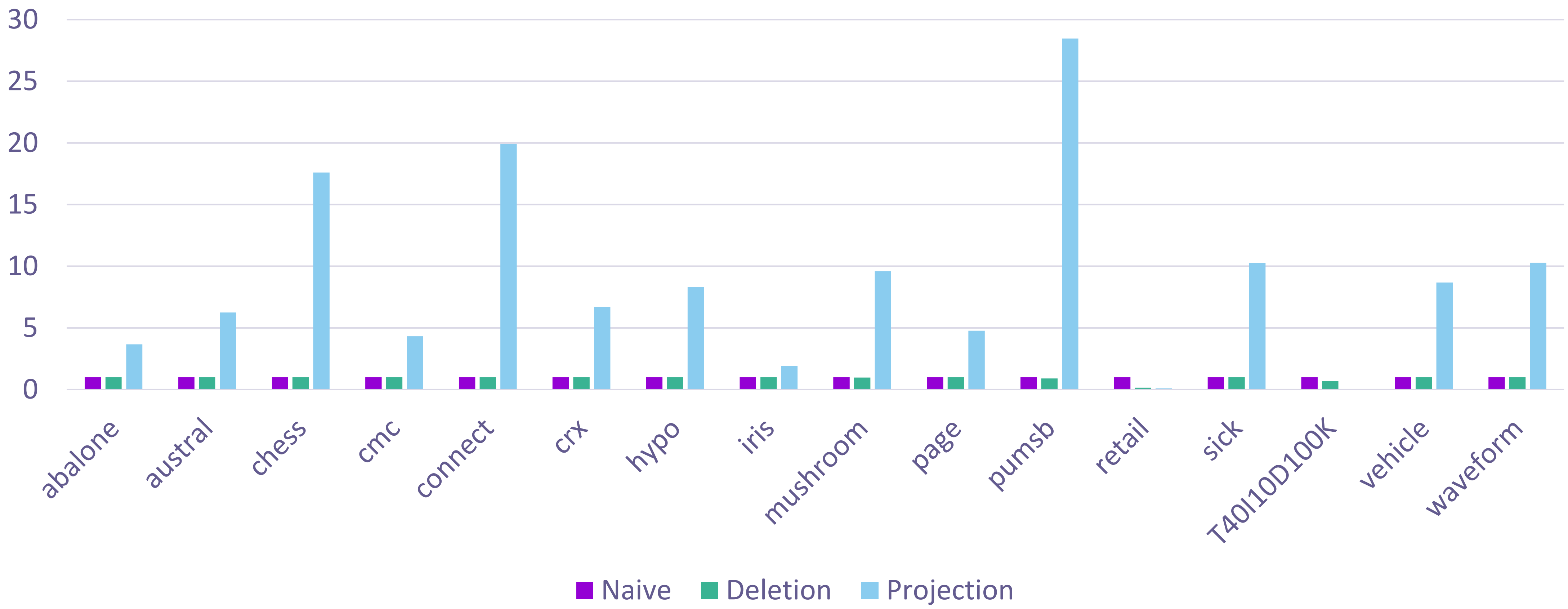

**Projection method works well for any threshold.**

# AUC of rejection rates




**Projection method works well for any threshold and any dataset.**

# Storage of projected datasets (worst case)




**Projection method requires more data storage (x30).**

# Conclusion

- **Lesson 1: Complementary of sampling and constraints**
  - Sampling : Controlled number of patterns + fast
  - Constraint : Elimination of non-relevant patterns

- **Lesson 2: Syntactic constraints can be pushed into sampling**
  - Not so hard to apply
  - Maximum length constraint not so efficient for avoiding the curse of the long tail

- **Lesson 3: Syntactic constraints can be derived from non-syntactic constraints**
  - Generic: Any frequent pattern sampling method can be reused
  - Efficient: Low rejection ratio at the cost of extra storage