The Chosen Few:
On Identifying Valuable Patterns

CMILE 2007, Warsaw, Poland
21.09.2007

Björn Bringmann, **Albrecht Zimmermann**
Katholieke Universiteit Leuven, Belgium

Cut the Crap

CMILE 2007, Warsaw, Poland
21.09.2007

Björn Bringmann, **Albrecht Zimmermann**
Katholieke Universiteit Leuven, Belgium

Cut the Crap

ICDM 2007, Omaha, USA
30.10.2007

Björn Bringmann, **Albrecht Zimmermann**
Katholieke Universiteit Leuven, Belgium

Idea of Data Mining

Idea of Data Mining

We torture the data until it confesses

Idea of Data Mining

We torture the data until it confesses

Problem with Torture

Idea of Data Mining

We torture the data until it confesses

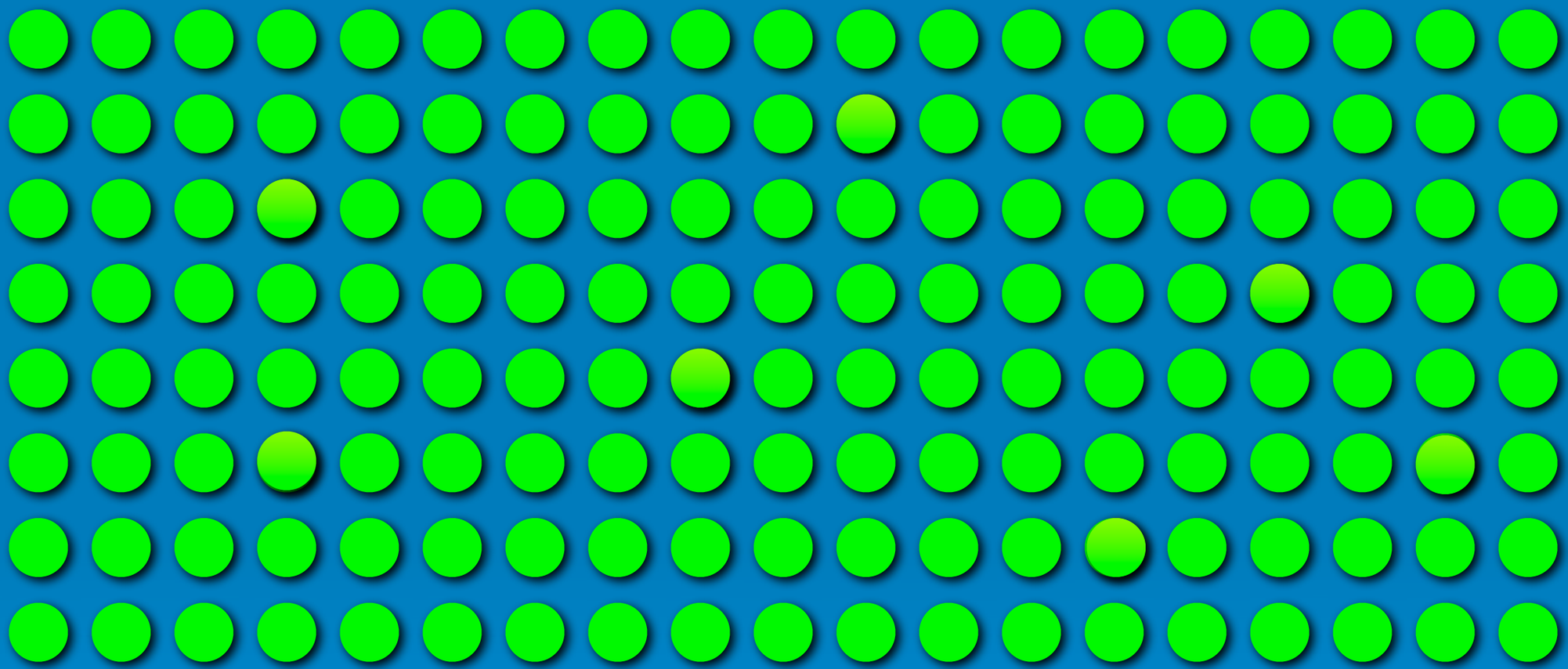
Problem with Torture

It's easy to get answers,
however, it's hard to get the *truth*!

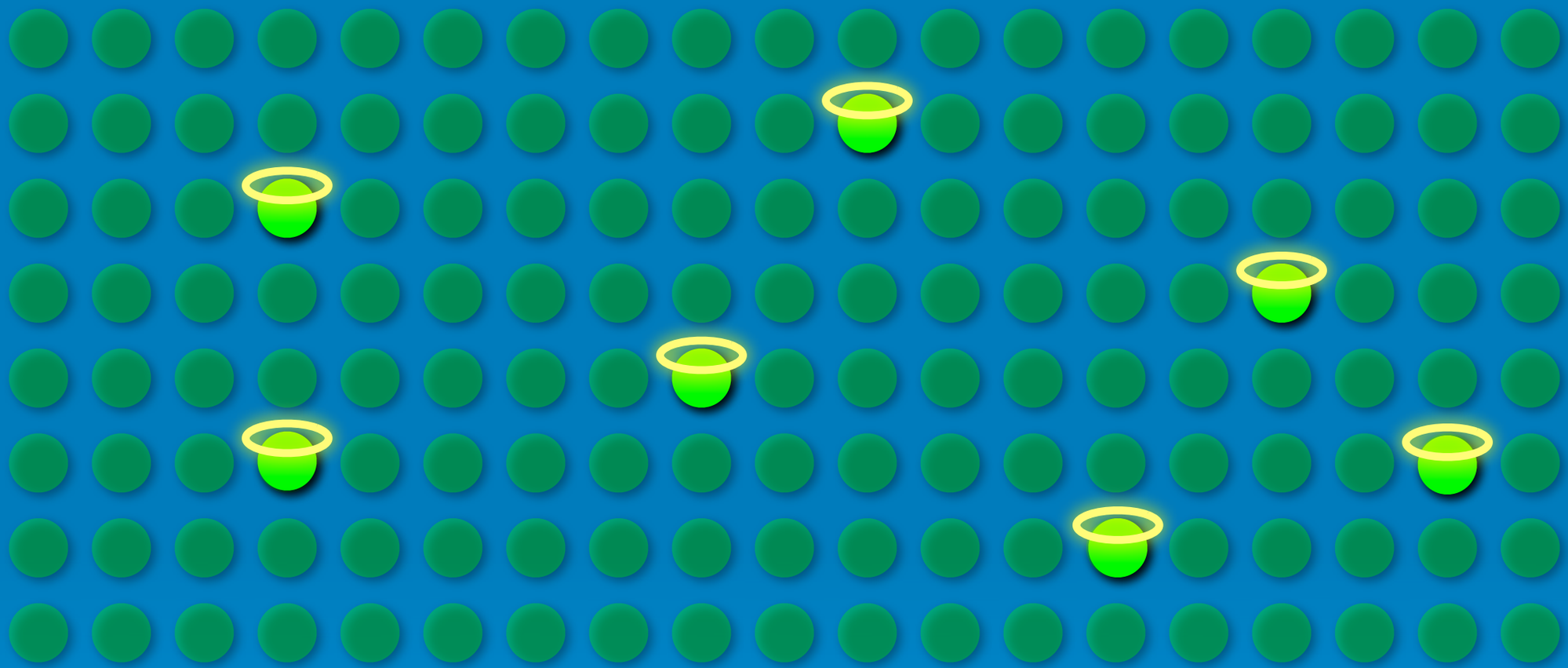
Frequent Pattern Mining

Frequent Pattern Mining

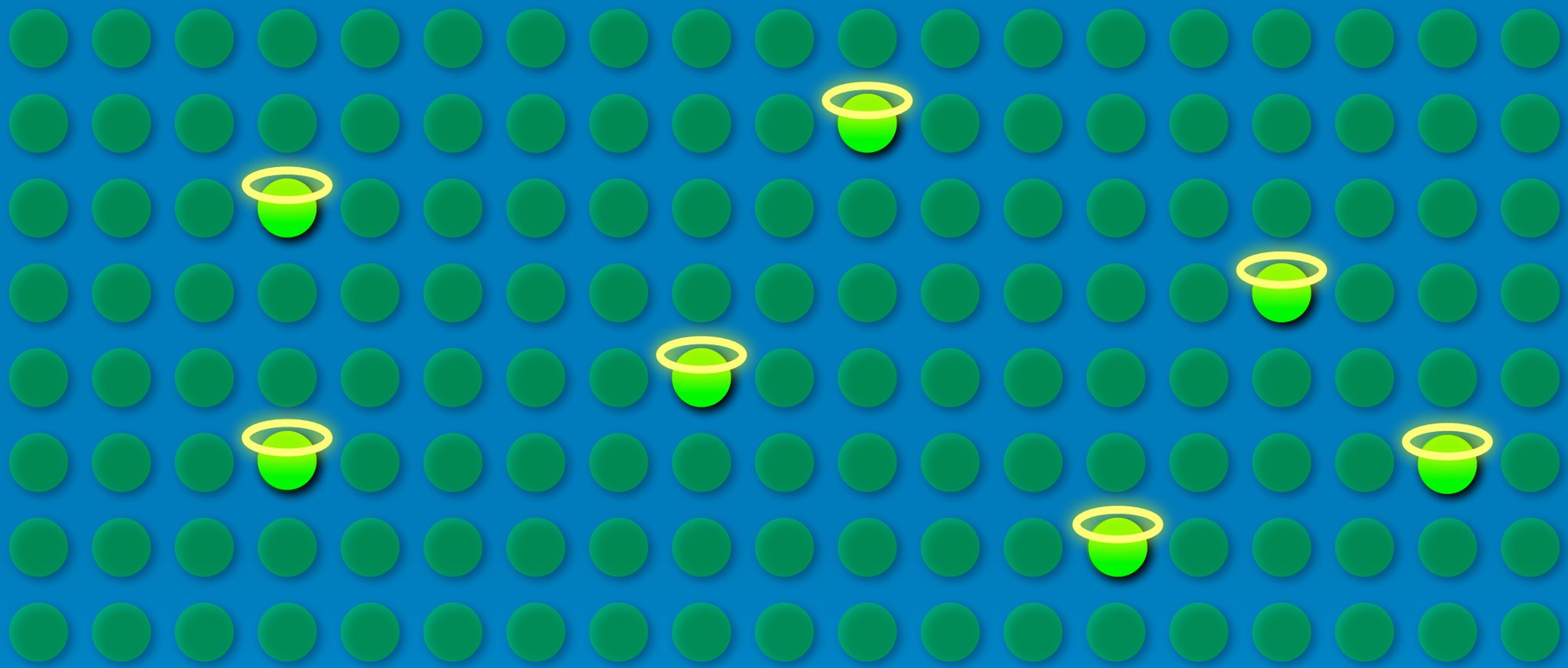
Frequent Pattern Mining



Frequent Pattern Mining



Frequent Pattern Mining



How to identify the valuable patterns ?

The Trigger

Don't be afraid of simpler patterns - [ECML2006]

- Mine top k ***correlated*** pattern (X^2 wrt. class)
- Use patterns as binary features
- Compare different structural features

The Trigger

Don't be afraid of simpler patterns - [ECML2006]

- Mine top k ***correlated*** pattern (X^2 wrt. class)
- Use patterns as binary features
- Compare different structural features

Coverage of the patterns found is ***very redundant***

Semantic Constraints for the Subset

- small enough for human inspection
- no redundancy
- full pattern set information

Semantic Constraints for the Subset

- small enough for human inspection
- no redundancy
- full pattern set information

Only information available is the database

Semantic Constraints for the Subset

- small enough for human inspection
- no redundancy
- full pattern set information

Only information available is the database

information of set ***S***
~
partition ***S*** induces on the data

Semantic Constraints for the Subset

- small enough for human inspection
- no redundancy
- full pattern set information

Only information available is the database

Semantic Constraints for the Subset

- small enough for human inspection
- no redundancy
- full pattern set information

Only information available is the database

Partition

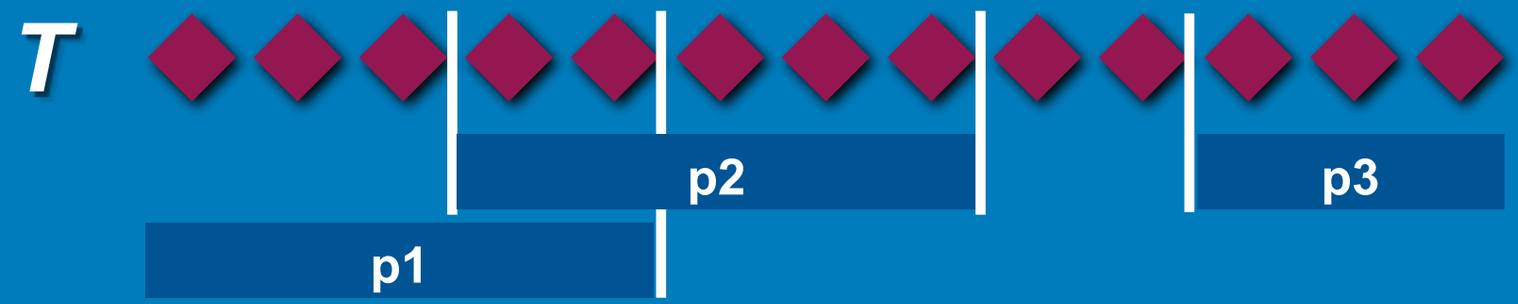
Given a *set* of *examples* ***T*** and *pattern* set ***S***

Partition

Given a *set* of *examples* ***T*** and *pattern* set ***S***

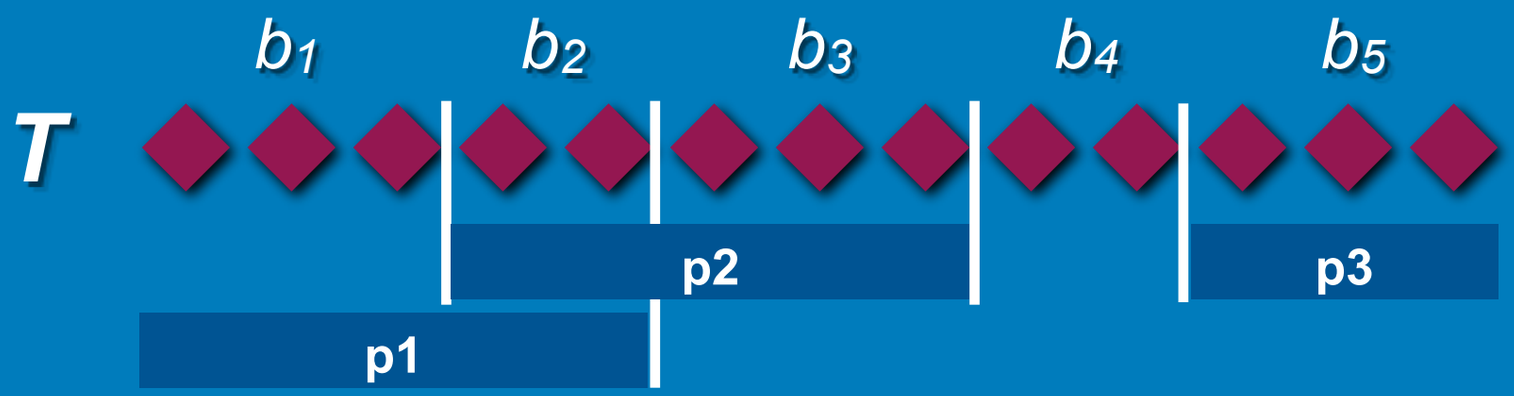
Partition

Given a *set* of *examples* ***T*** and *pattern* set ***S***



Partition

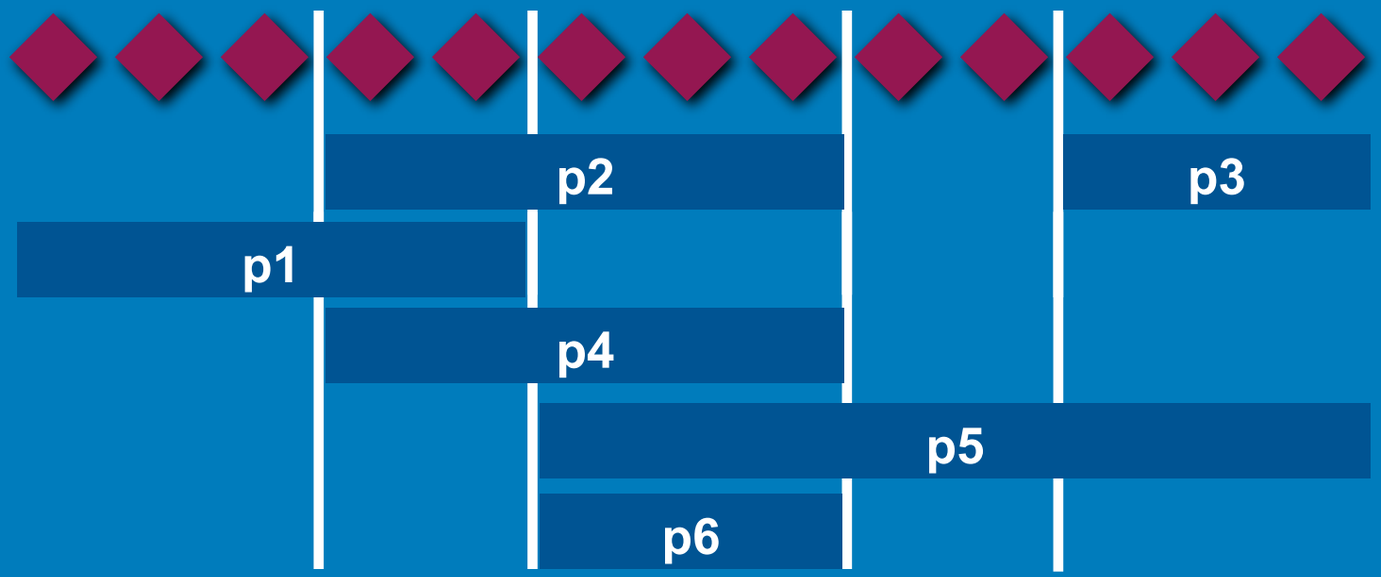
Given a *set* of *examples* ***T*** and *pattern* set ***S***



a partition consists of a set of blocks

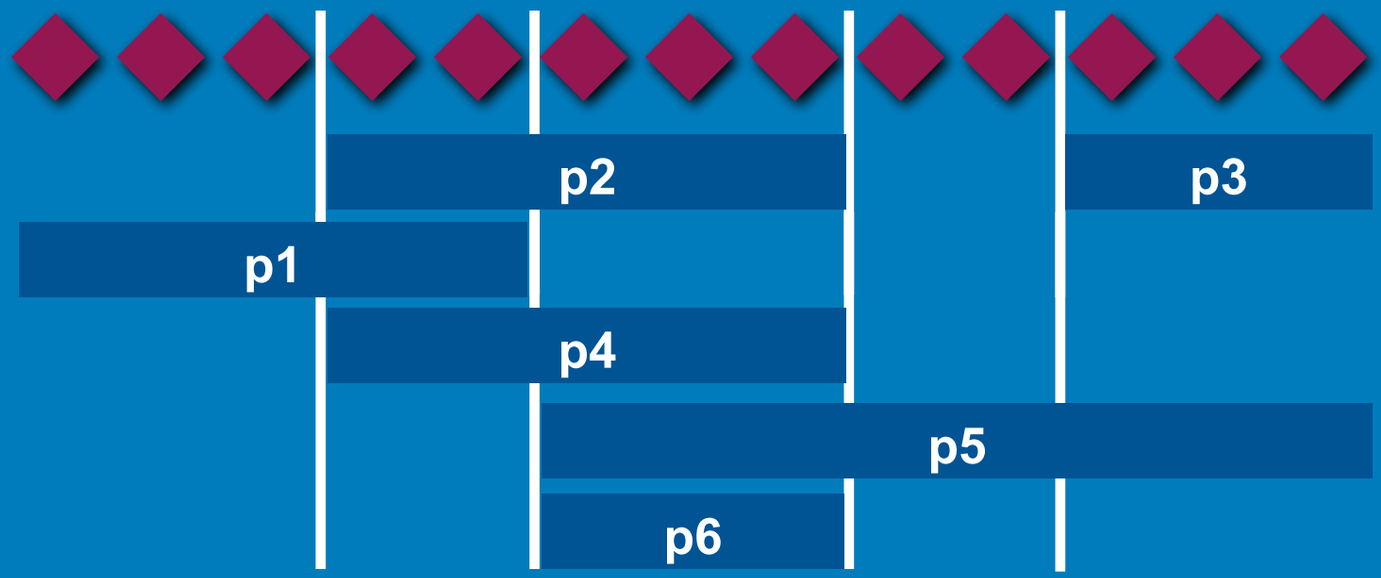
Redundancy

Given a *set* of *examples* ***T*** and *pattern* set ***S***



Redundancy

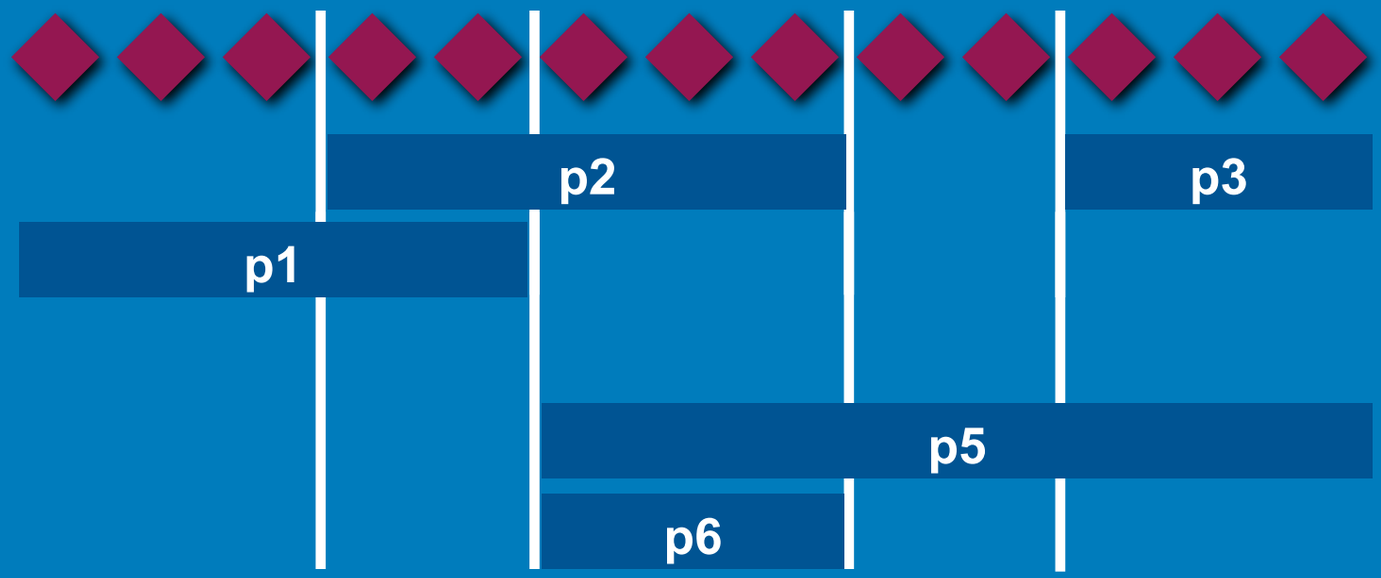
Given a *set* of *examples* ***T*** and *pattern* set ***S***



- $p_4 \approx p_2$

Redundancy

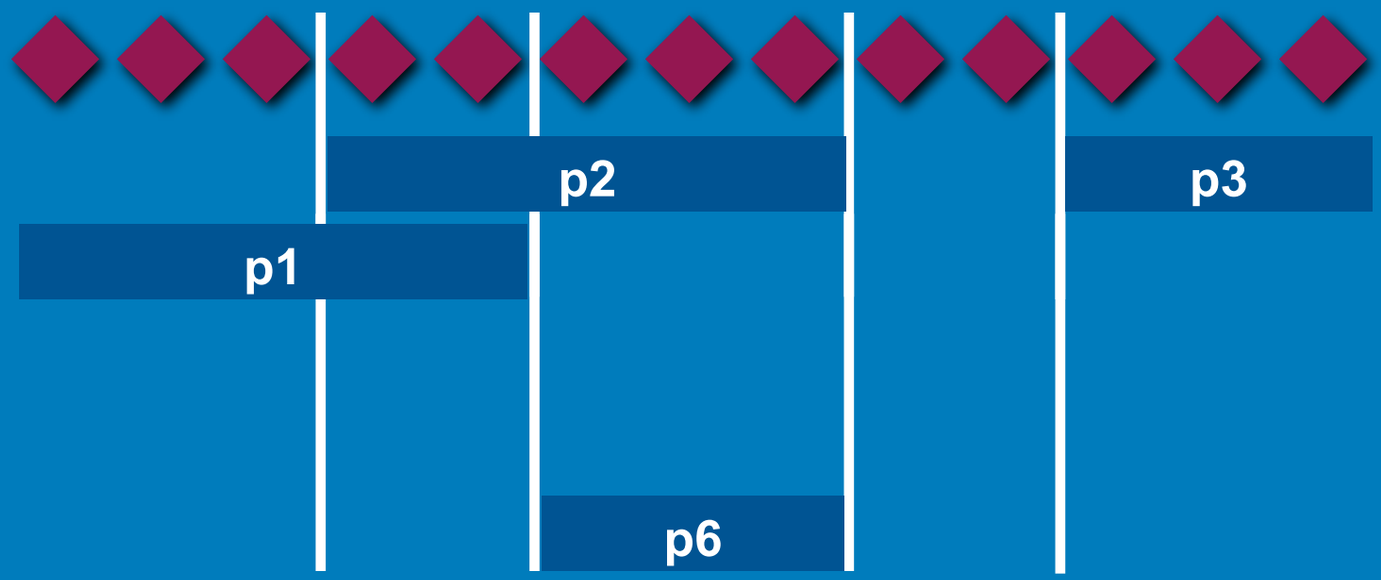
Given a *set* of *examples* ***T*** and *pattern* set ***S***



- $p_4 \approx p_2$
- $p_5 \approx \neg p_1$

Redundancy

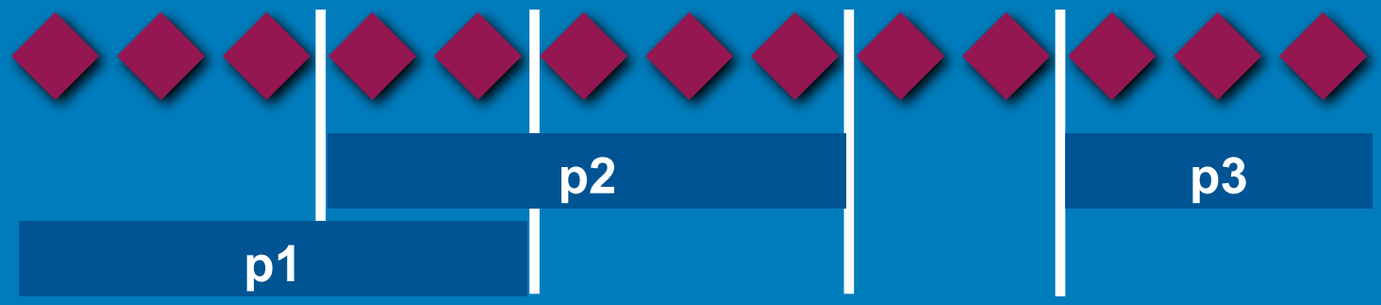
Given a *set* of *examples* ***T*** and *pattern* set ***S***



- $p_4 \approx p_2$
- $p_5 \approx \neg p_1$
- $p_6 \approx p_2 \wedge \neg p_1$

Redundancy

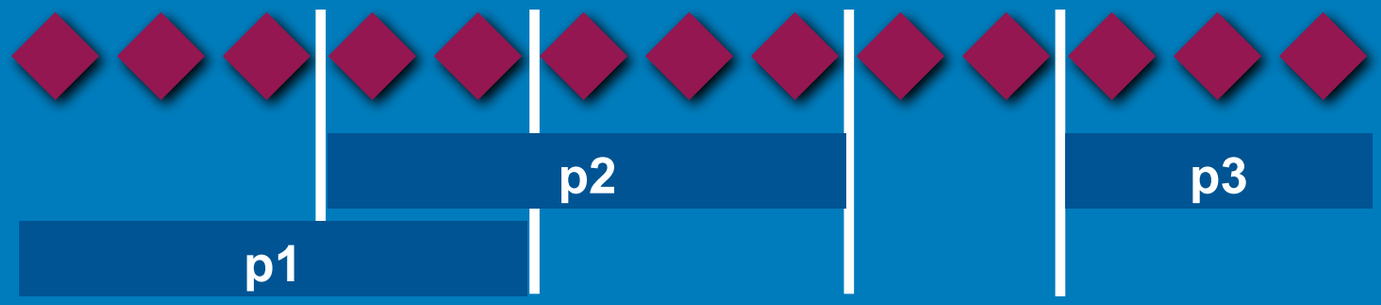
Given a *set* of *examples* ***T*** and *pattern* set ***S***



- $p_4 \approx p_2$
- $p_5 \approx \neg p_1$
- $p_6 \approx p_2 \wedge \neg p_1$

Redundancy

Given a *set* of *examples* ***T*** and *pattern* set ***S***



$$partitionsize \leq 2^{|S|}$$

- $p_4 \approx p_2$
- $p_5 \approx \neg p_1$
- $p_6 \approx p_2 \wedge \neg p_1$

A Measure For Redundancy

formalising the constraints

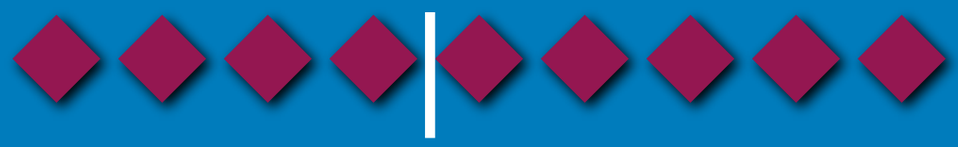
$$\Phi\,(T, S^*, p) \rightarrow [0,1]$$

- 3 measures Φ based on
  - ***q****uotient*: splits induced by ***p***
  - ***i****nference*: predictability of ***p*** from **S***
  - ***c****lustering*: rand index **S*** vs **S*** ∪ **p**

Quotient: Φ_Q

Fraction of split blocks

Umformulierung möglich so dass Upper Bound berechnet werden kann

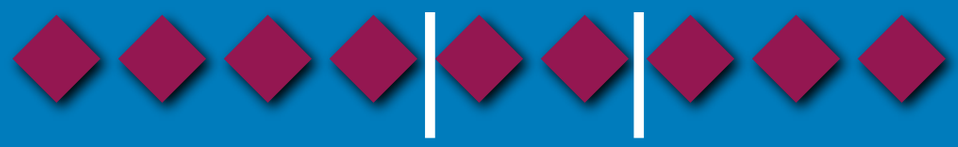


blocks	*fraction*
2	

Quotient: Φ_Q

Fraction of split blocks

Umformulierung möglich so dass Upper Bound berechnet werden kann

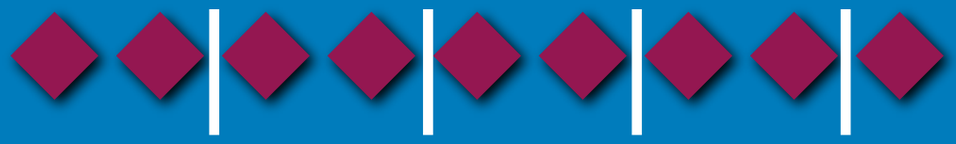


blocks	*fraction*
2	
3	0.66

Quotient: Φ_Q

Fraction of split blocks

Umformulierung möglich so dass Upper Bound berechnet werden kann

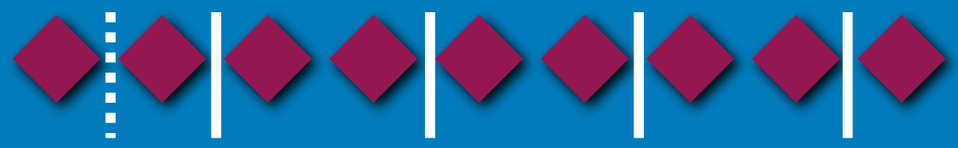


blocks	*fraction*
2	
3	0.66
5	0.60

Quotient: Φ_Q

Fraction of split blocks

Umformulierung möglich so dass Upper Bound berechnet werden kann

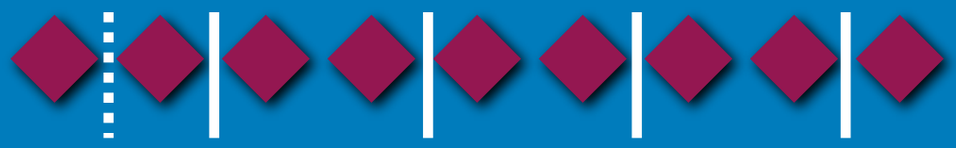


blocks	*fraction*
2	
3	0.66
5	0.60
6	0.83

Quotient: Φ_Q

Fraction of split blocks

Umformulierung möglich so dass Upper Bound berechnet werden kann



blocks	*fraction*
2	
3	0.66
5	0.60
6	0.83

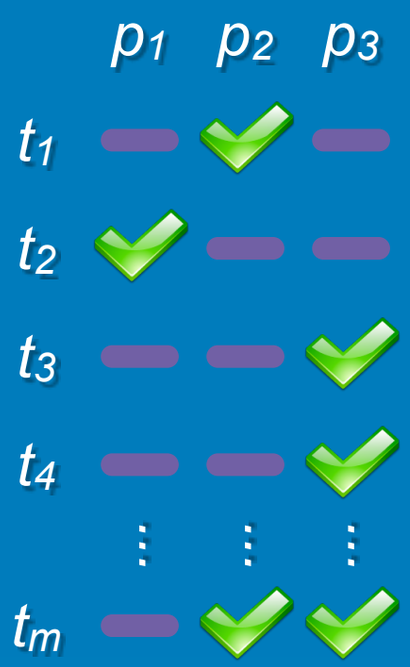
Φ_Q = 1- S* blocks / S^* ∪ p_i blocks

- easy evaluation
- rejects low granularity improvements

Inference: Φ_I

Predictability of new pattern

Learner hier eigentlich überflüssig: Blöcke entweder negativ oder positiv labeln und "zählen"

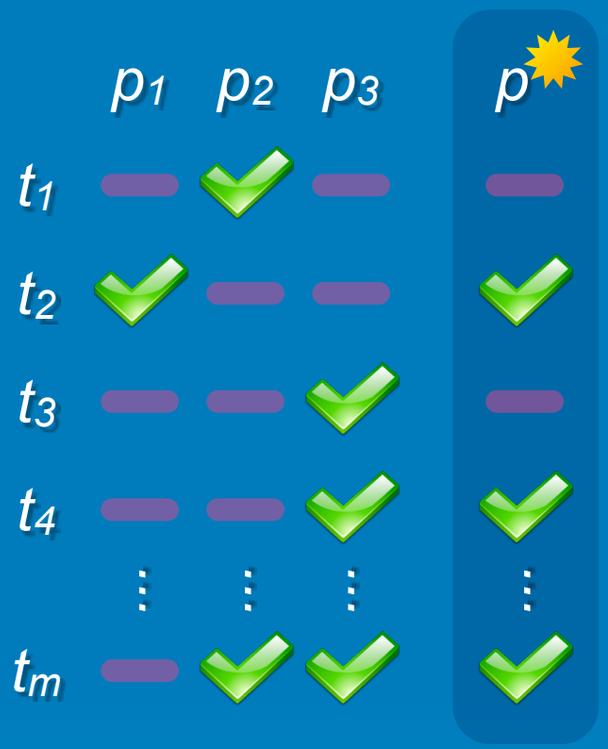


*) JRip - WEKA, unpruned

Inference: Φ_I

Predictability of new pattern

Learner hier eigentlich überflüssig: Blöcke entweder negativ oder positiv labeln und "zählen"



Hypotheses

$p^* = p_1 \vee (p_2 \wedge p_3)$

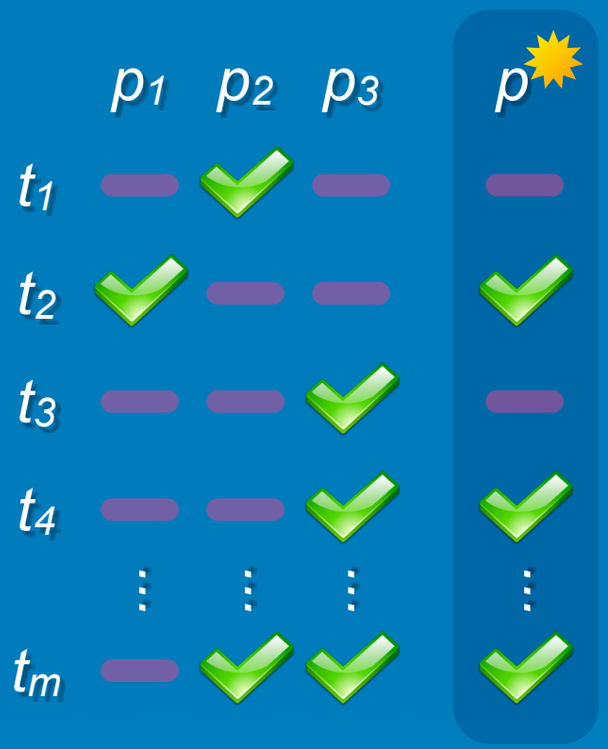
$p^* = p_1 \vee p_3$

*) JRip - WEKA, unpruned

Inference: Φ_I

Predictability of new pattern

Learner hier eigentlich überflüssig: Blöcke entweder negativ oder positiv labeln und "zählen"



Hypotheses

$p^* = p_1 \vee (p_2 \wedge p_3)$

$p^* = p_1 \vee p_3$

Φ_I = error of hypotheses

- Selects balanced splits
- Reduces pattern dependencies

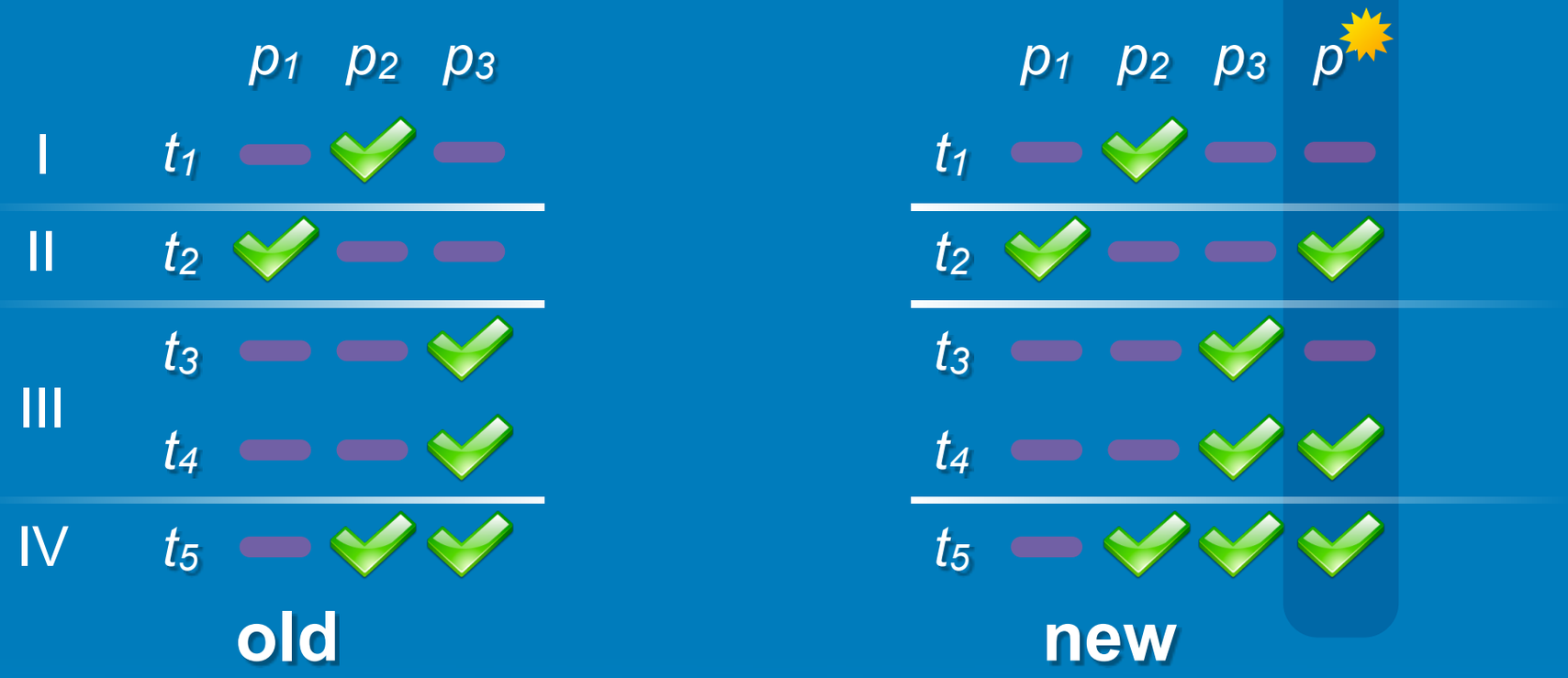
*) JRip - WEKA, unpruned

Clustering: Φ_C

Change of clustering

Clustering: Φ_C

Change of clustering

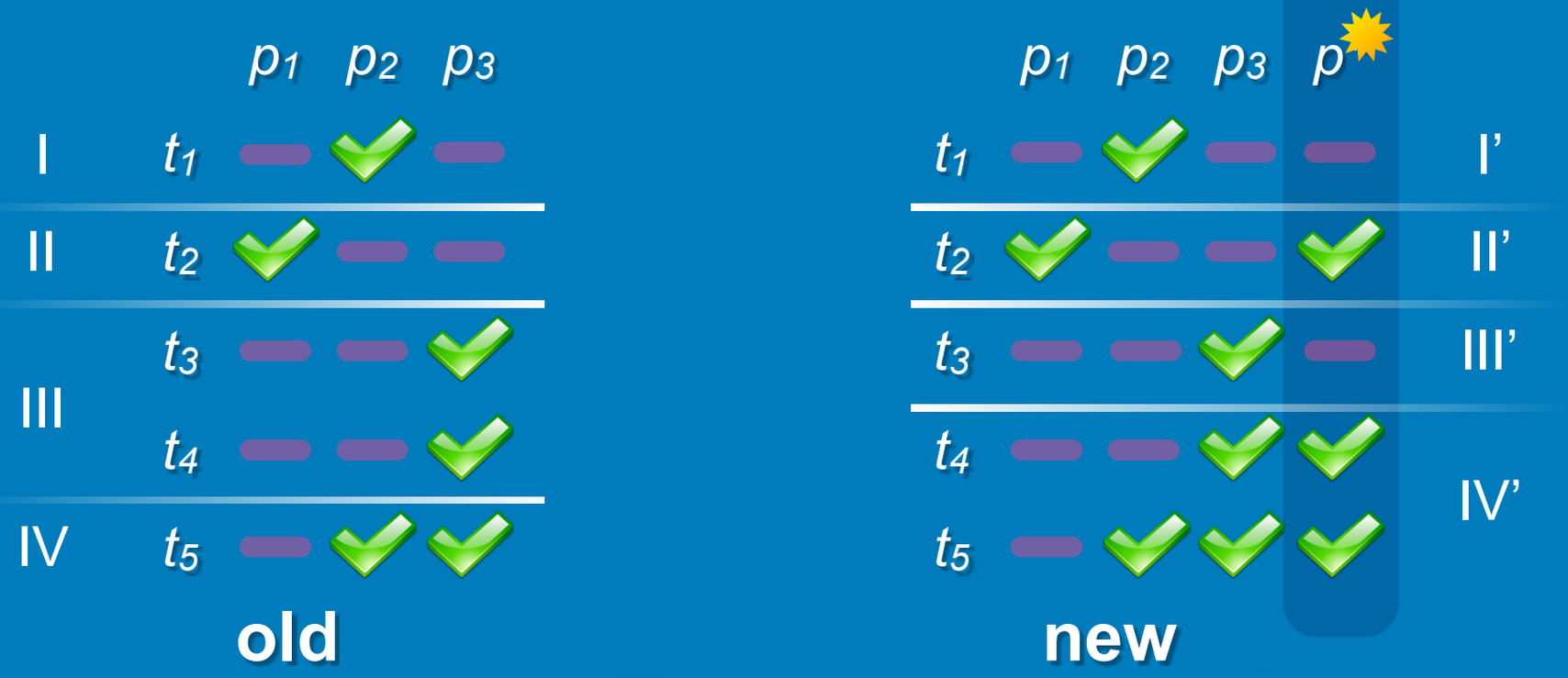


Clustering: Φ_C

Change of clustering

Clustering: Φ_C

Change of clustering



Clustering: Φ_C

Change of clustering

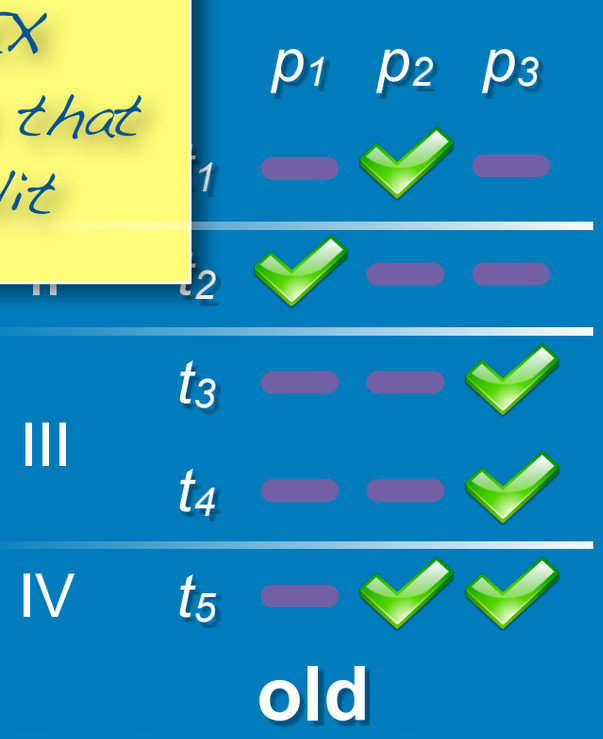
Φ_C = 1 - **rand** (old *vs.* new clustering)

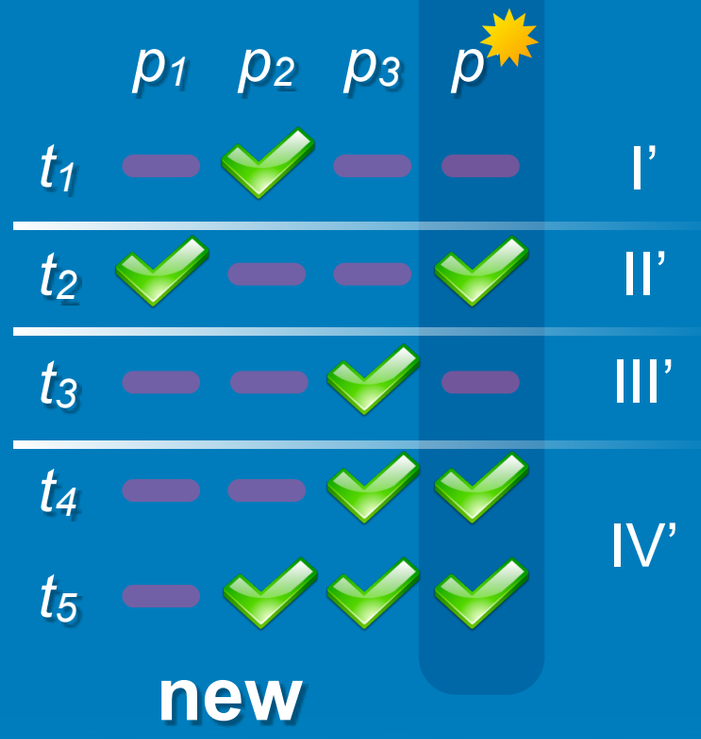
- balanced splits or
- numerous splits

Clustering: Φ_C

Change of clustering

RAND INDEX
number of pairs that don't join/split





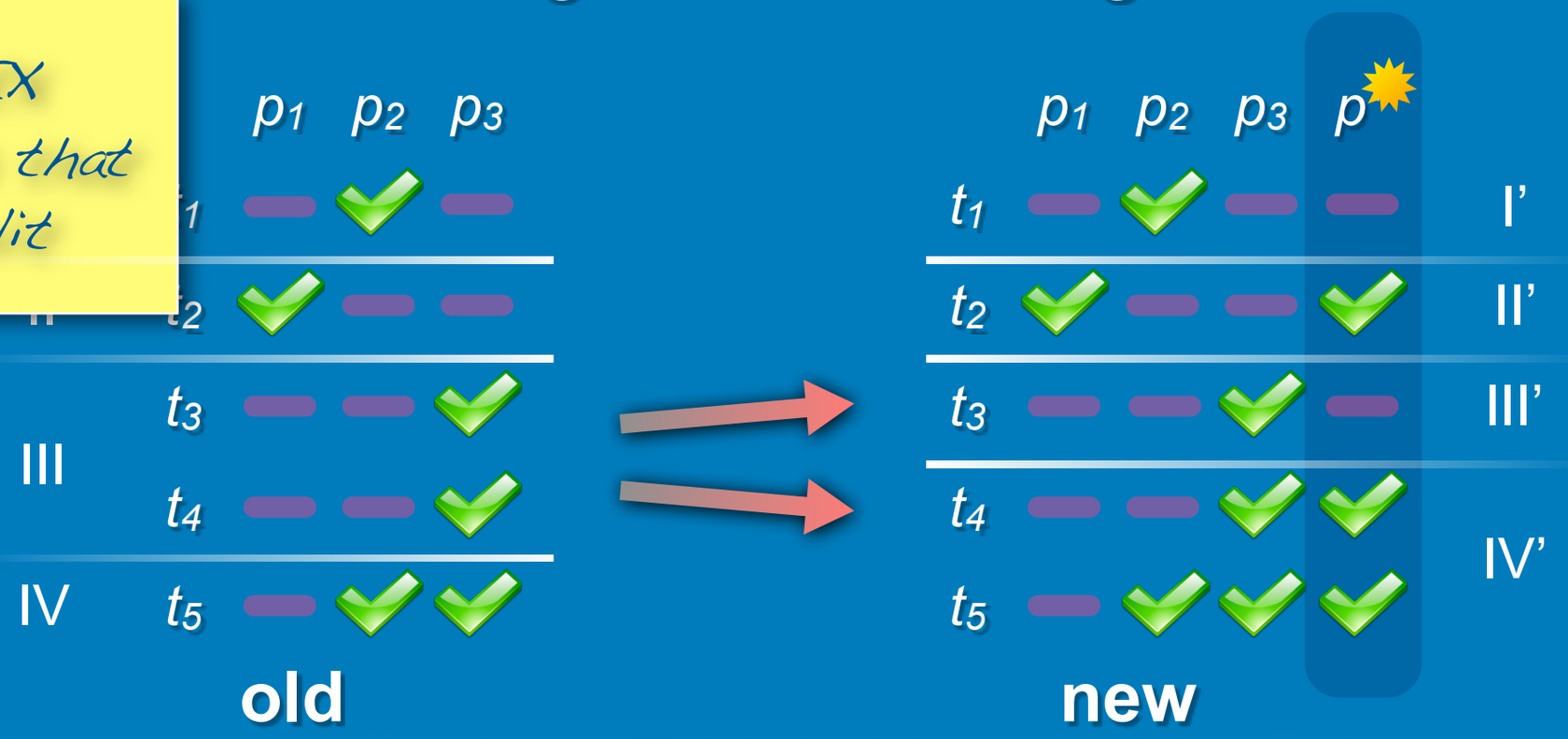
Φ_C = 1 - **rand** (old *vs.* new clustering)

- balanced splits or
- numerous splits

Clustering: Φ_C

Change of clustering

RAND INDEX
number of pairs that don't join/split



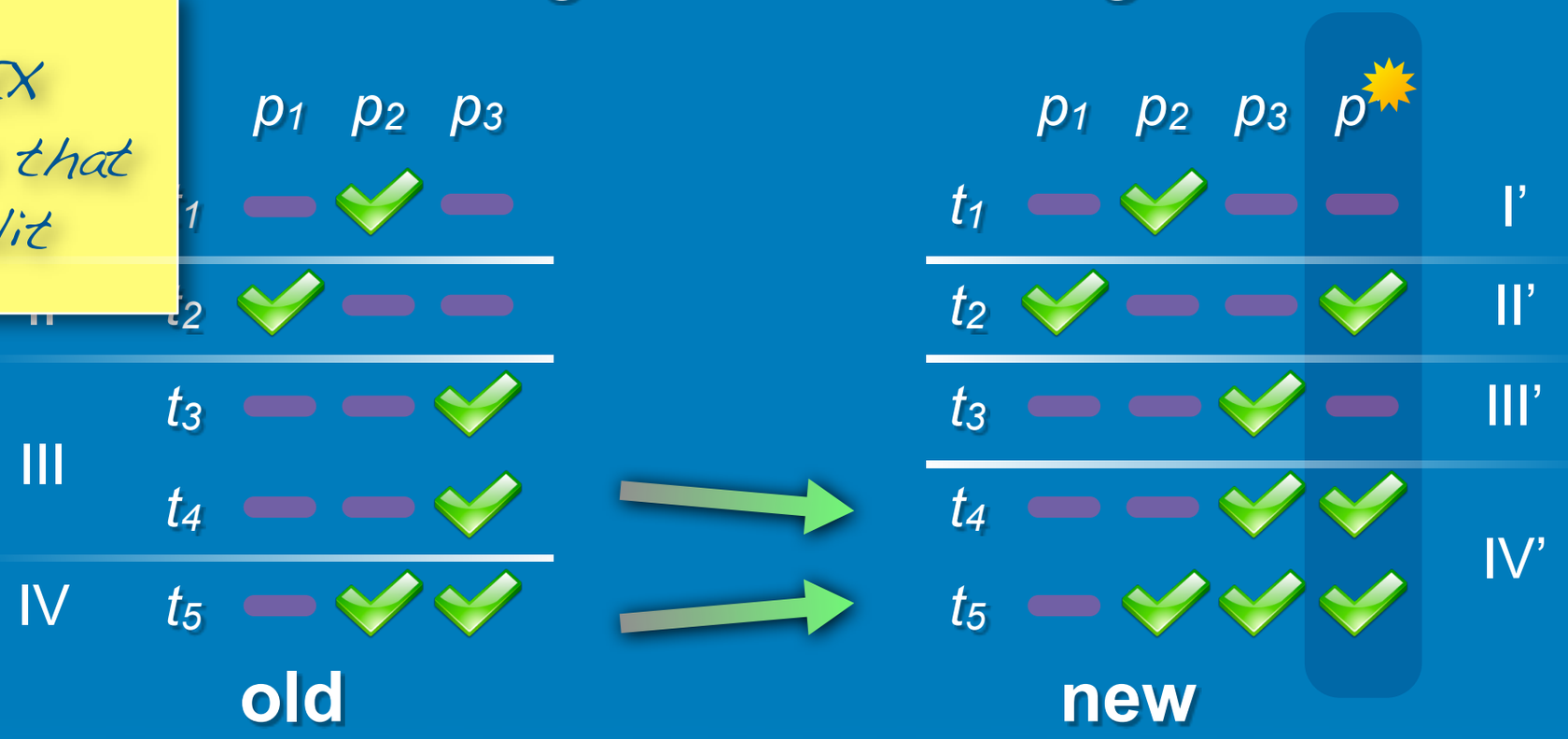
Φ_C = 1 - **rand** (old *vs.* new clustering)

- balanced splits or
- numerous splits

Clustering: Φ_C

Change of clustering

RAND INDEX
number of pairs that don't join/split



Φ_C = 1 - **rand** (old *vs.* new clustering)

- balanced splits or
- numerous splits

The Approach

Space of possible subsets ***S**** of ***S*** is huge

The Approach

Space of possible subsets S^* of S is huge

Basic Greedy Approach

for $i = 1$ to $|S|$ do
if $\Phi(T, S^*, p_i) > t$
$S^* = S^* \cup p_i$

The Approach

Space of possible subsets S^* of S is huge

Basic Greedy Approach

for $i = 1$ to $|S|$ do
if $\Phi(T, S^*, p_i) > t$
$S^* = S^* \cup p_i$

Processing order of patterns relevant !

The Approach

Space of possible subsets S^* of S is huge

Basic Greedy Approach

for $i = 1$ to $|S|$ do
if $\Phi(T, S^*, p_i) > t$
$S^* = S^* \cup p_i$

Processing order of patterns relevant !

All Φ reject fully redundant patterns

Experimental Setup

[classification, categorical attr., big enough]

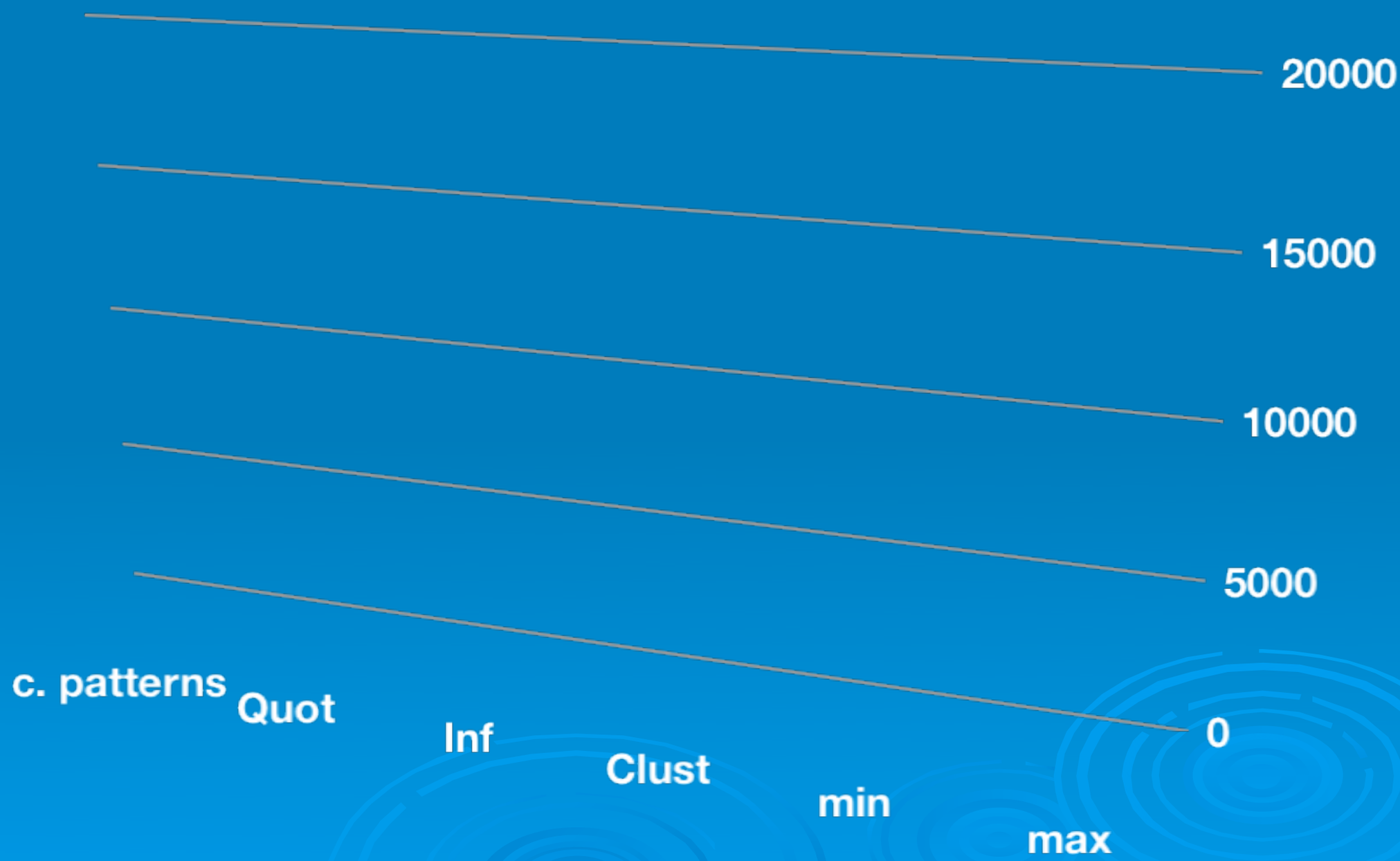
Experimental Setup

- Six UCI Datasets, each two support thresholds
  [classification, categorical attr., big enough]
- Four orders on closed itemsets
  - support ascending
  - support descending
  - length ascending
  - length descending
- Three Measures
  - Quotient
  - Inference
  - Clustering
- Eight acceptance thresholds

1152 Experiments

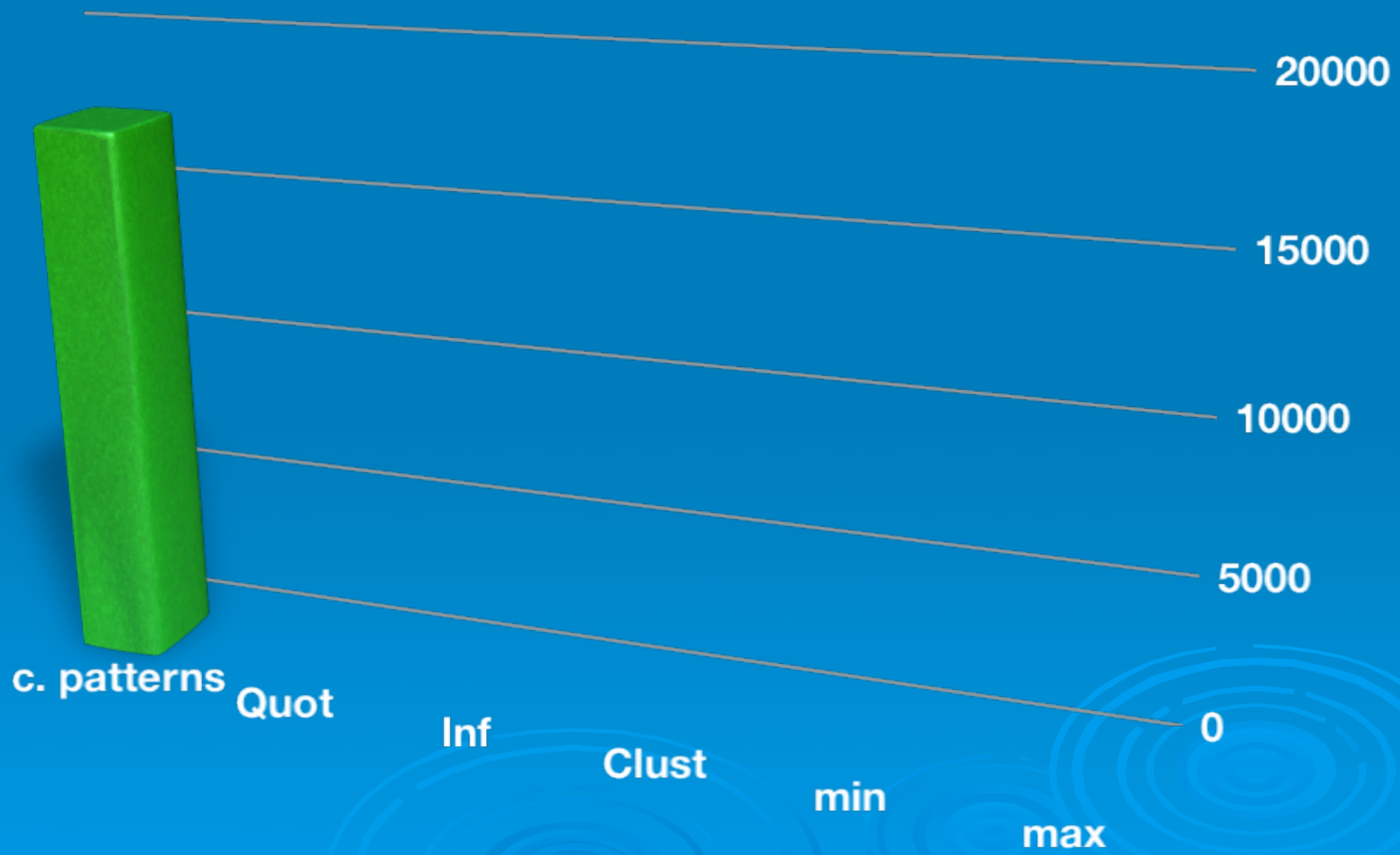
Primary Tumor

339 Instances, 10% support



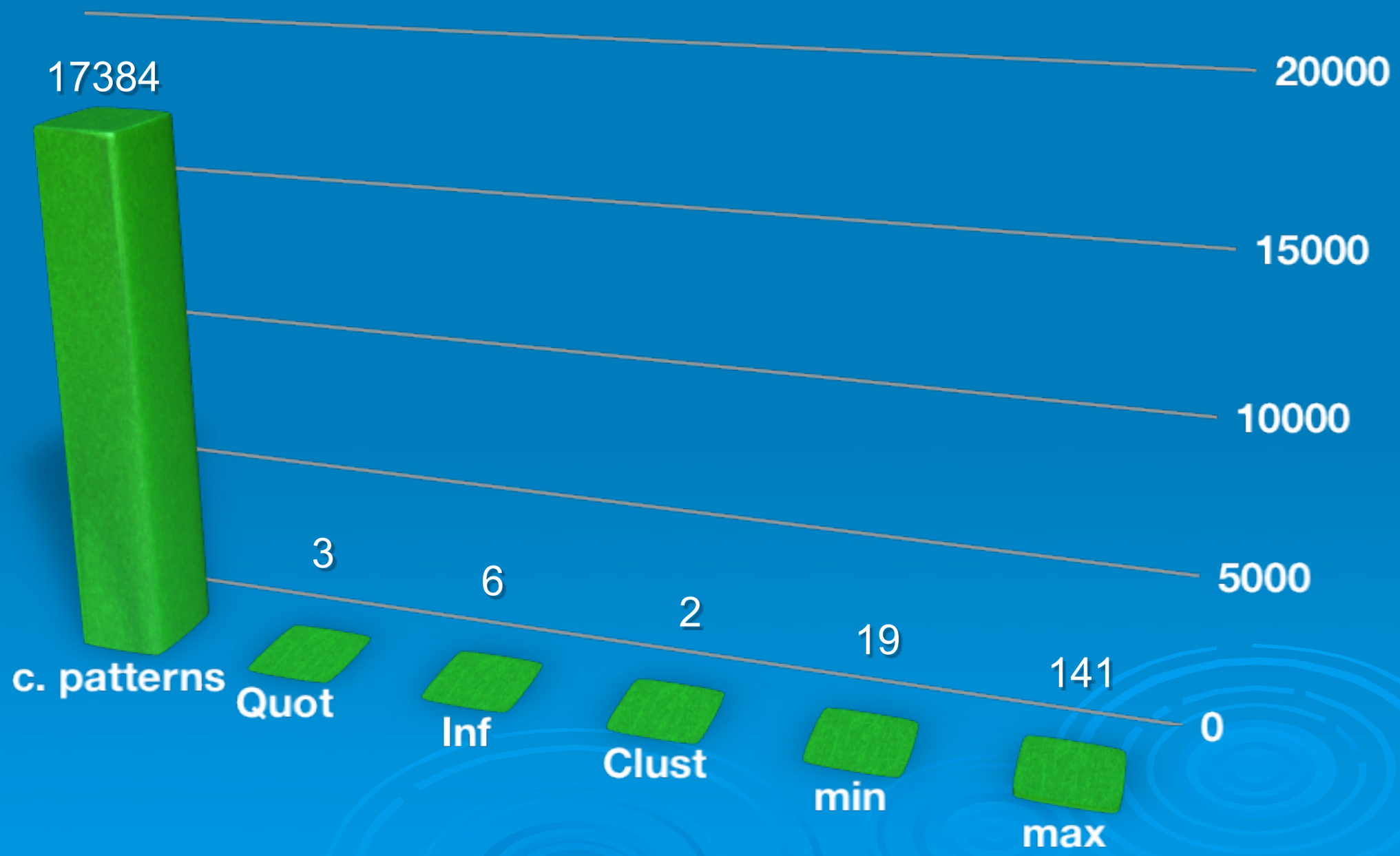
Primary Tumor

339 Instances, 10% support



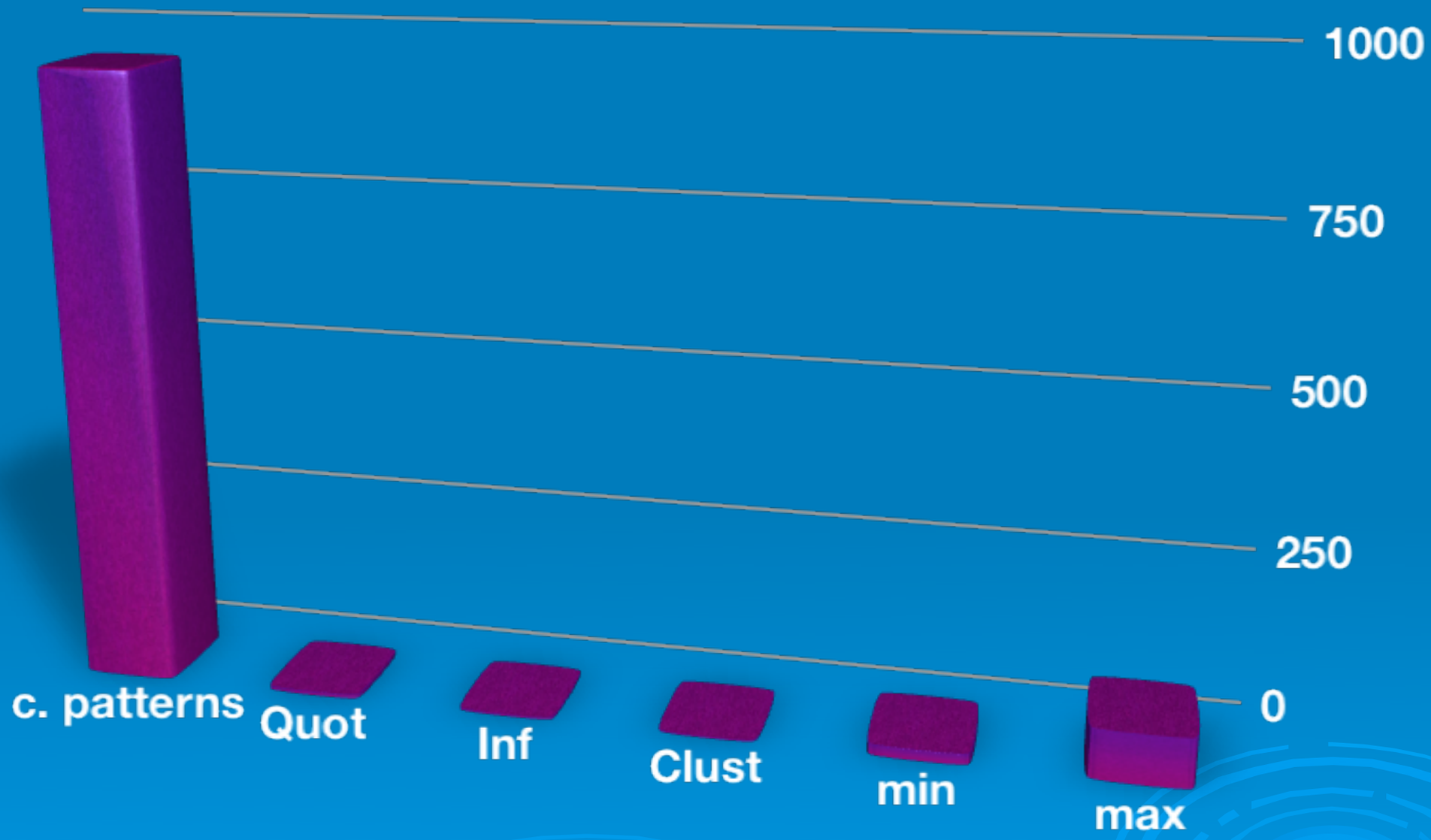
Primary Tumor

339 Instances, 10% support



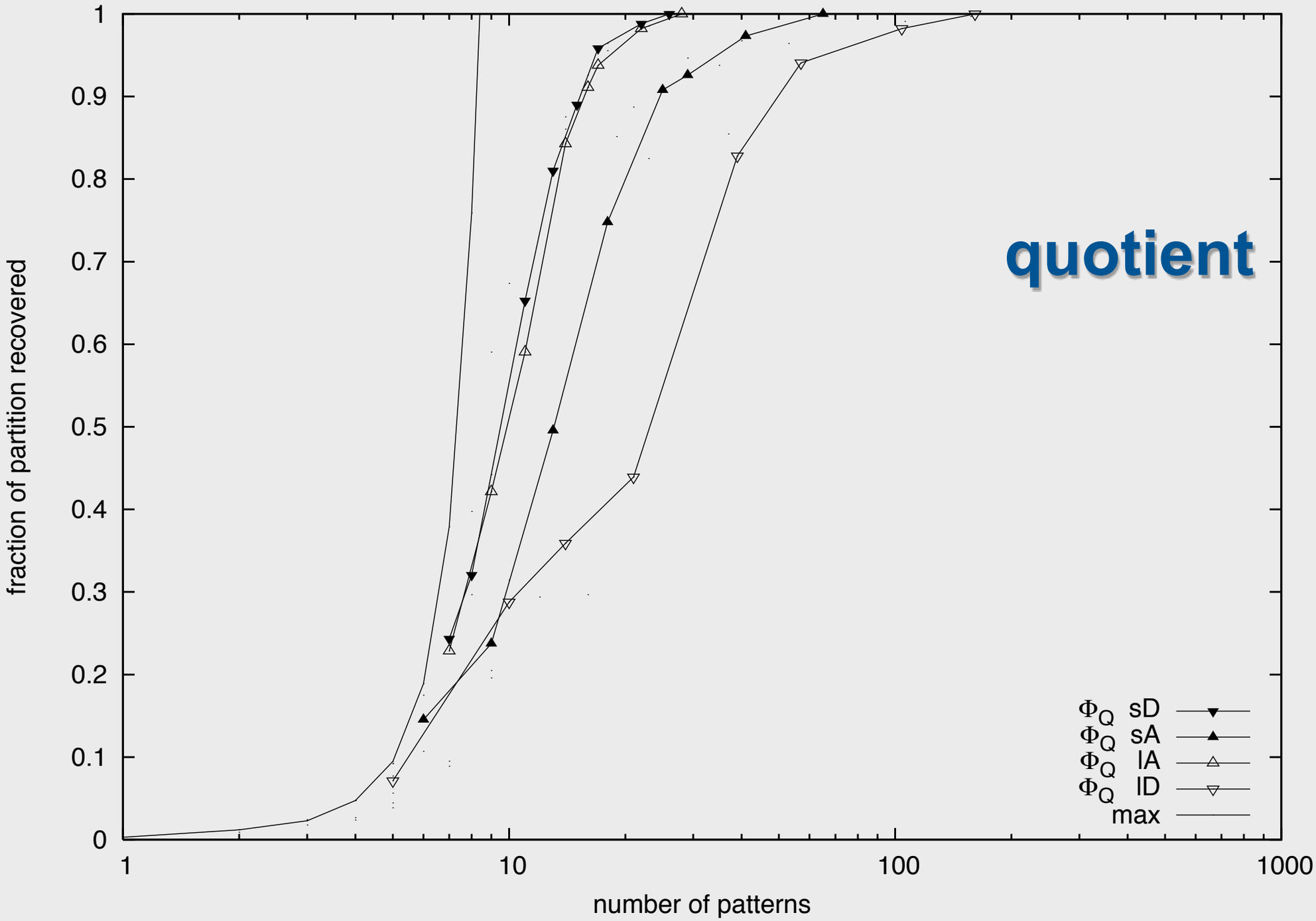
Tic Tac Toe

958 Instances, 5% support



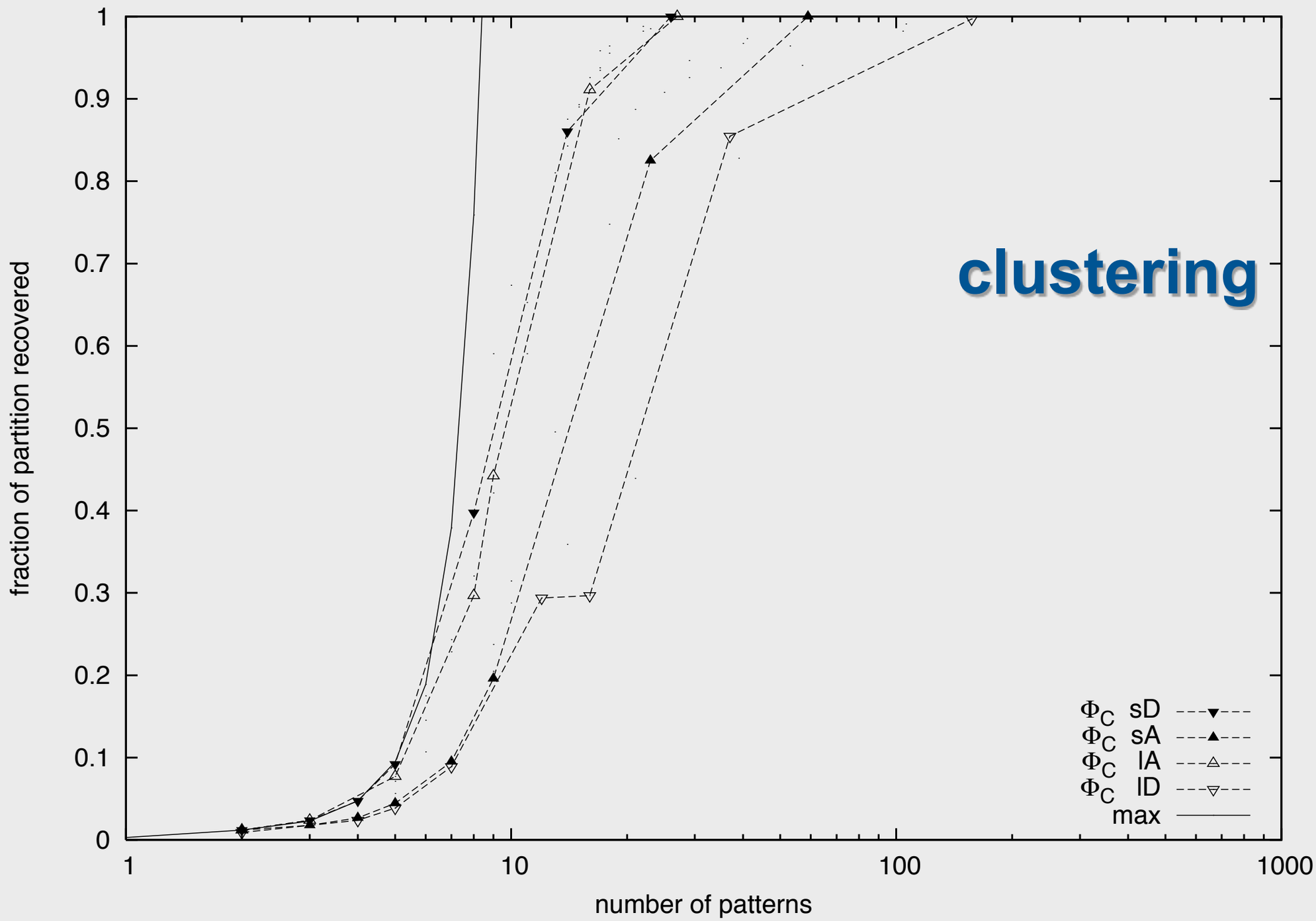
Partition Recovery

voting, support 25%



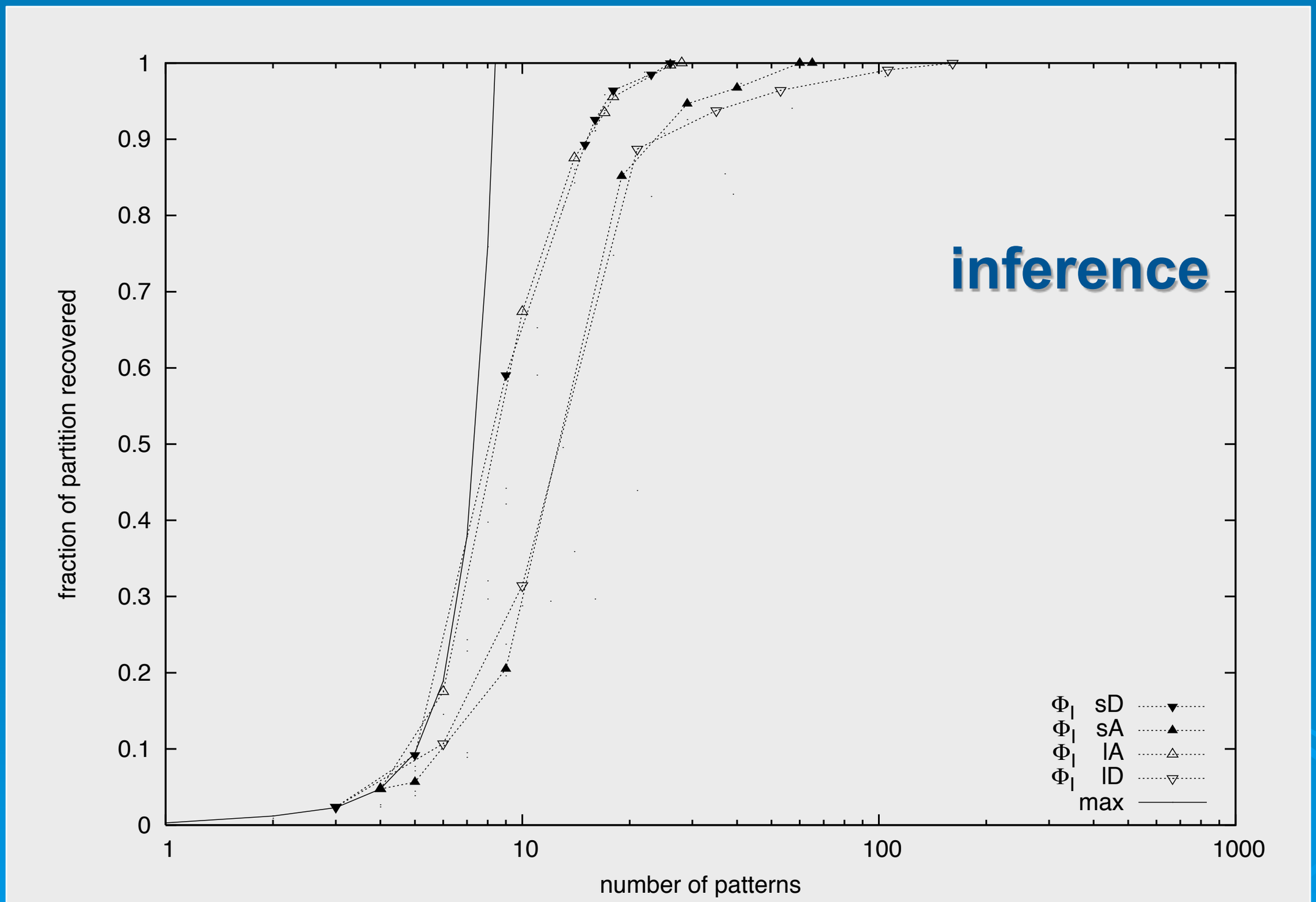
Partition Recovery

voting, support 25%



Partition Recovery

voting, support 25%



Support ↓ vs. Length ↑

- Short itemsets not always high support

Support ↓ vs. Length ↑

- Short itemsets not always high support

- Low acceptance threshold
  - Similar induced partitions (rand > 95%)

- High acceptance threshold
  - Depending on measure

Ascending ↑ vs. Descending ↓

Assumption

different information of low support patterns

Ascending ↑ vs. Descending ↓

Assumption
different information of low support patterns

- Inference
- High support can infer low support

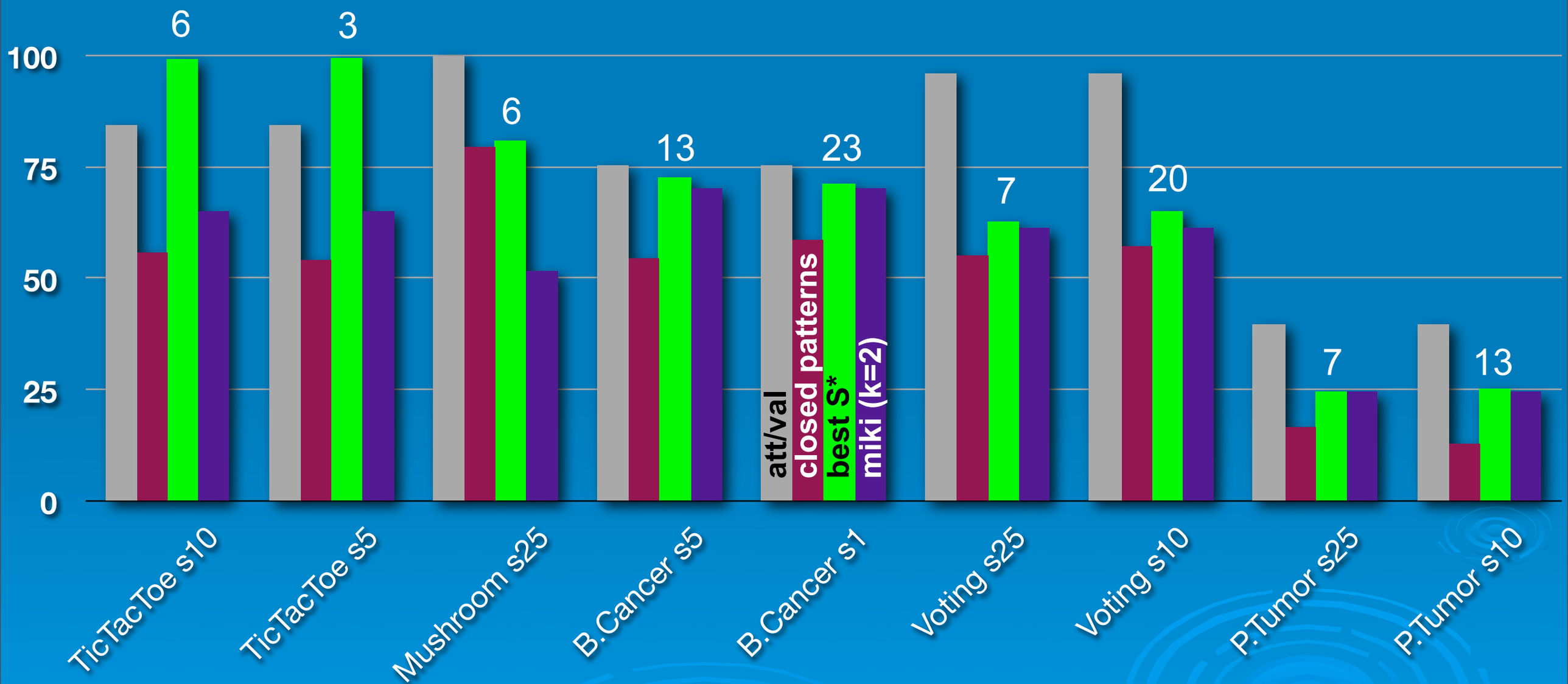
Ascending ↑ vs. Descending ↓

Assumption
different information of low support patterns

- Inference
  - High support can infer low support
- Partition similarity
  - Highly similar partitions (rand)
  - Except clustering with coarse partitions

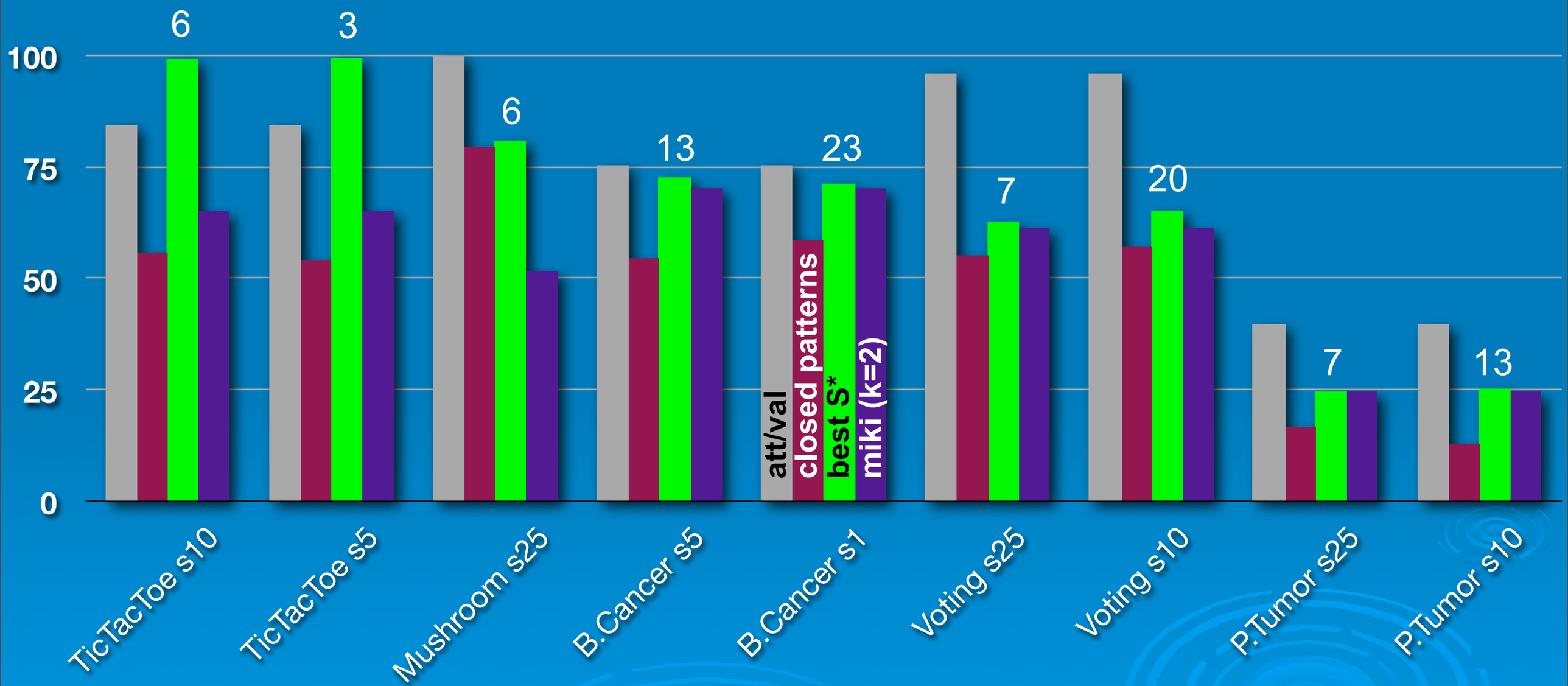
Prediction Quality

C4.5 pruned, 10 fold



Prediction Quality

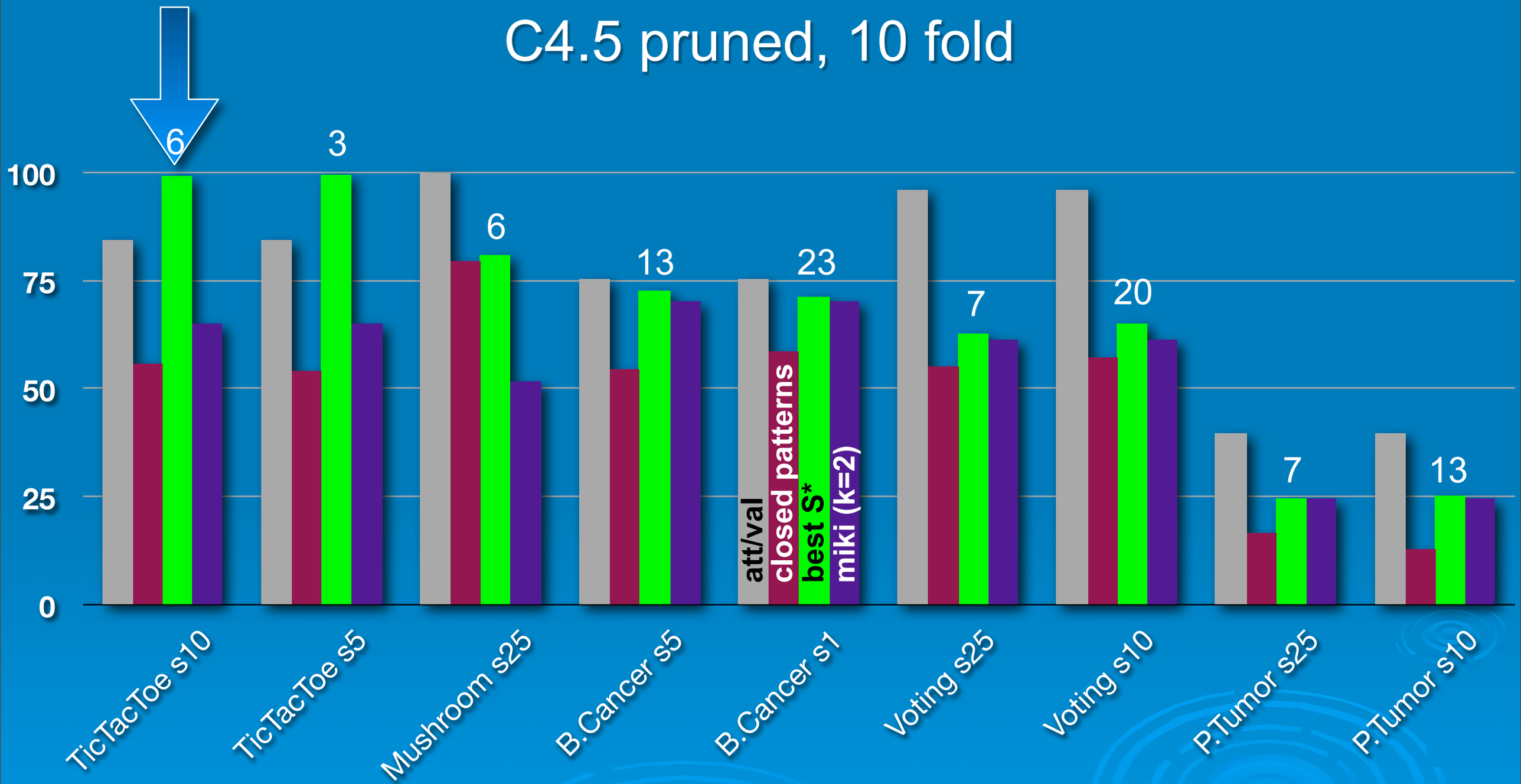
C4.5 pruned, 10 fold



overfitting when using all closed patterns

Prediction Quality

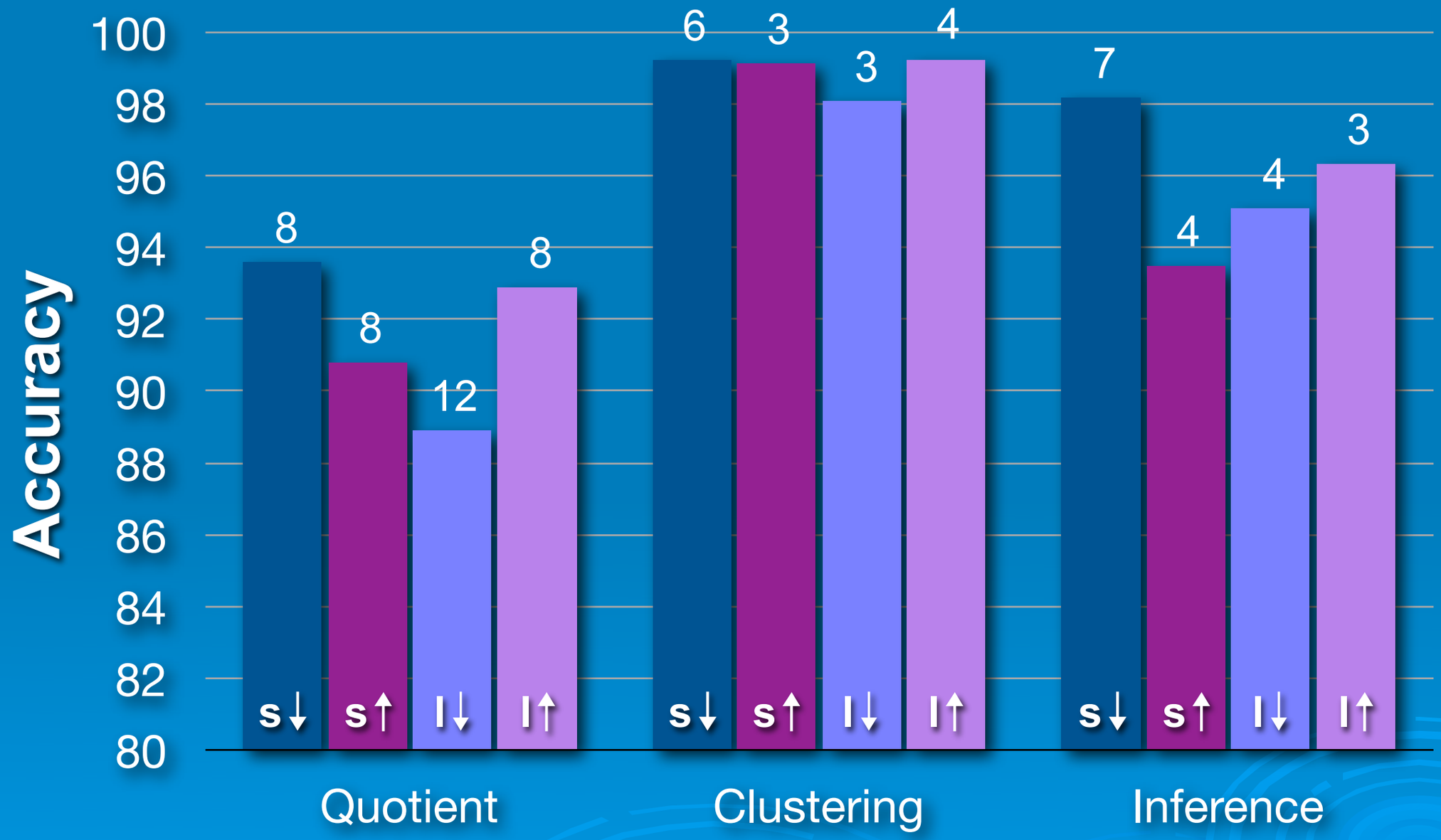
C4.5 pruned, 10 fold



overfitting when using all closed patterns

Prediction Quality

Tic Tac Toe, support 10%



Conclusions

- Introduced flexible mechanism
- and it works!

- Dramatic reduction in set size
  - Human-understandable
- Small sets = better accuracy
  - To a certain degree

Conclusions

- Introduced flexible mechanism
- and it works!

- Dramatic reduction in set size
  - Human-understandable
- Small sets = better accuracy
  - To a certain degree

The Chosen Few: On Identifying Valuable Patterns

ICDM 2007, Omaha, USA

Thank you very much

The Chosen Few: On Identifying Valuable Patterns
ICDM 2007, Omaha, USA

Thank you very much

The Chosen Few: On Identifying Valuable Patterns

ICDM 2007, Omaha, USA