What matters:
Size does, Smarts don’t

LEGO 2008, Antwerp, Belgium
15.09.2008

Albrecht Zimmermann , Björn Bringmann
Katholieke Univeriteit Leuven, Leuven, Belgium

Setting

- Transactional Database
- Binary Class
- Predictive model desired

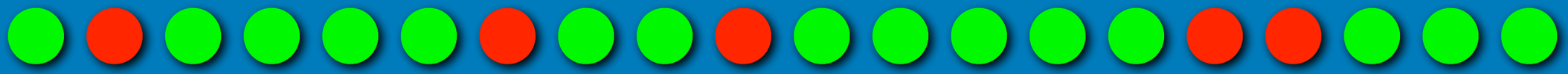
Setting

- Transactional Database
- Binary Class
- Predictive model desired

Mine patterns ***correlating*** with target concept instead of using *frequent* patterns

Pattern Mining

Given: transactional database with binary target class

D = 

Search frequent patterns that we want to use to distinguish between the (two) given classes

Pattern Mining

Given: transactional database with binary target class

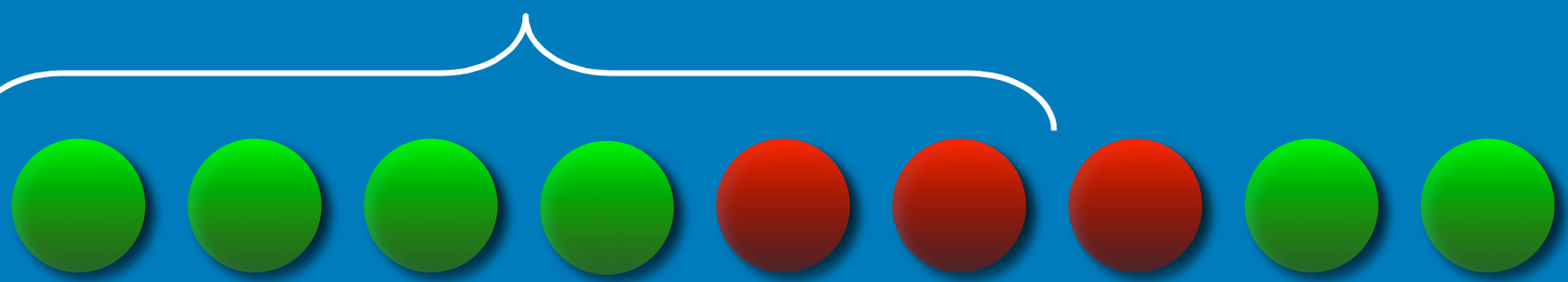
D =

Se uent patterns that we wa distinguish between iven classes

Correlated Pattern Mining

Compare class distribution on all instances against distribution on the covered instances

Covered by a **frequent** pattern



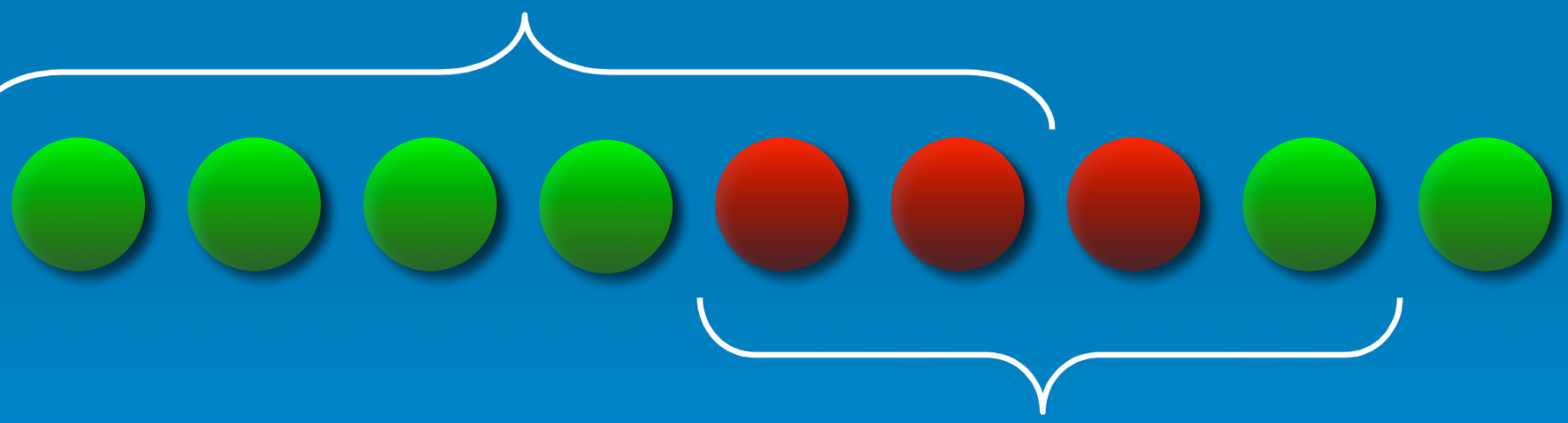
Each pattern quantified by score

Top-*k* mining efficiently extracts *k* best scoring patterns

Correlated Pattern Mining

Compare class distribution on all instances against distribution on the covered instances

Covered by a **frequent** pattern



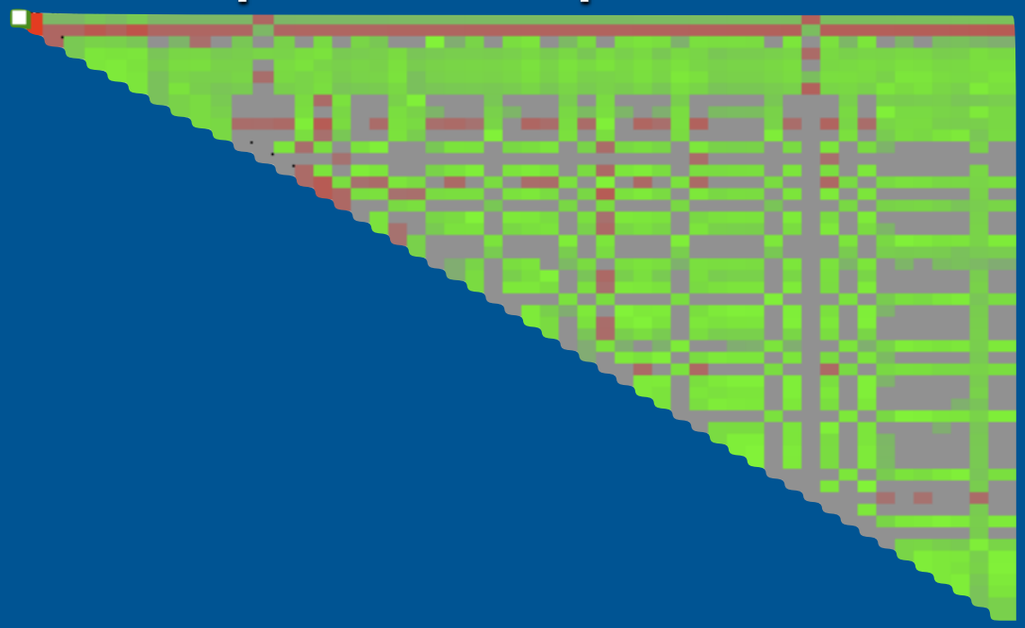
Covered by a **correlating** pattern

Each pattern quantified by score
Top-*k* mining efficiently extracts *k* best scoring patterns

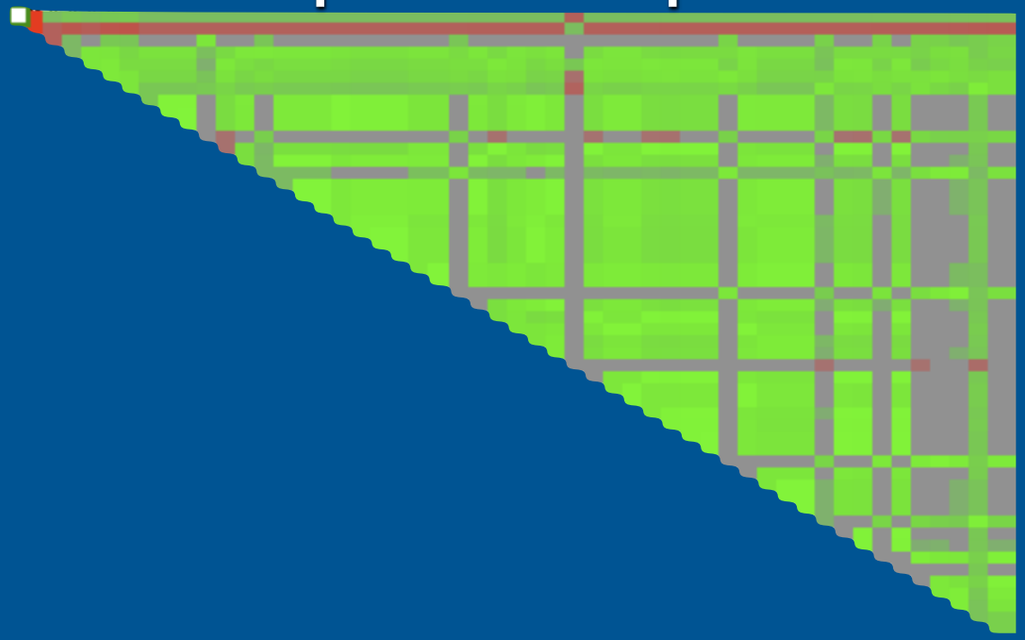
Correlated Pattern Mining

Correlation Matrix

Top 50 Sequences



Top 50 Graphs

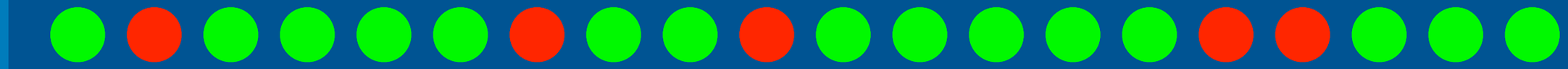


redundancy is still a problem

Each pattern quantified by score
Top-*k* mining efficiently extracts *k* best scoring patterns

Whole loaf vs. sliced bread

Pattern Mining considers whole database at once



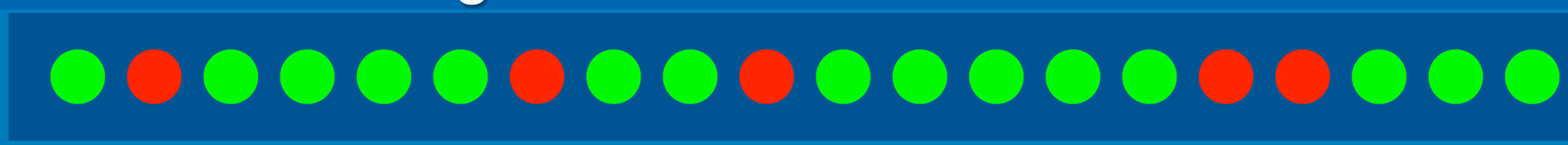
Th_3 = { a, b, c }

{ a, b, d } { a, c, d } { c, d, f } { c, d, f } { b, e, f }

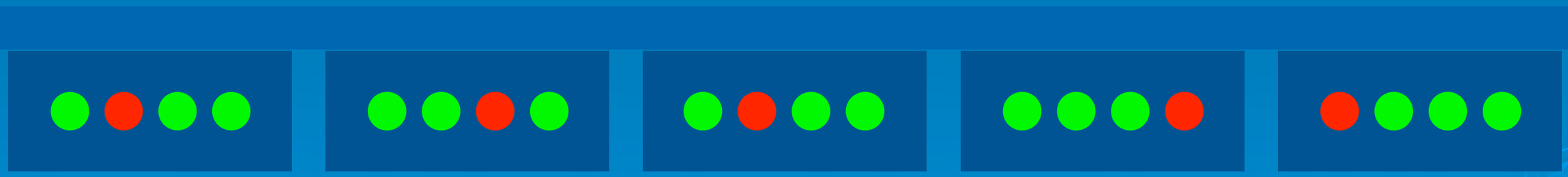
Machine Learning Techniques use parts of data.
Resampling, Validations Sets, Bagging, …

Whole loaf vs. sliced bread

Pattern Mining considers whole database at once



Th_3 = { a, b, c }

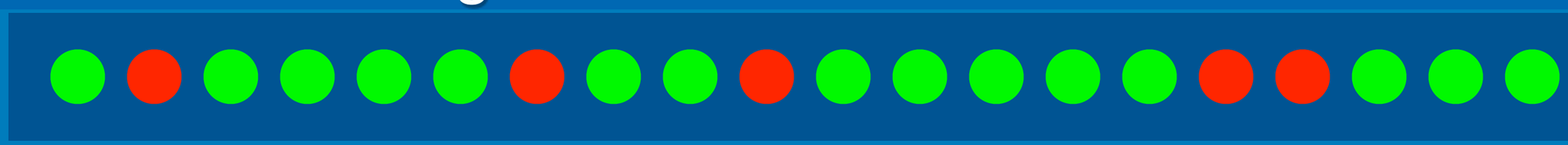


{ a, b, d } { a, c, d } { c, d, f } { c, d, f } { b, e, f }

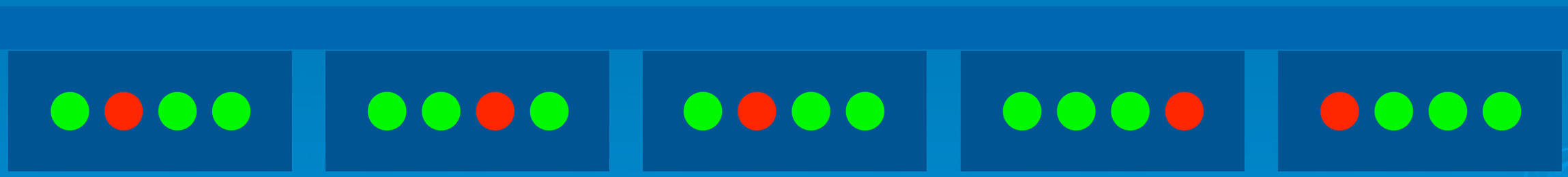
Machine Learning Techniques use parts of data.
Resampling, Validations Sets, Bagging, …

Whole loaf vs. sliced bread

Pattern Mining considers whole database at once



Th_3 = { a, b, c }

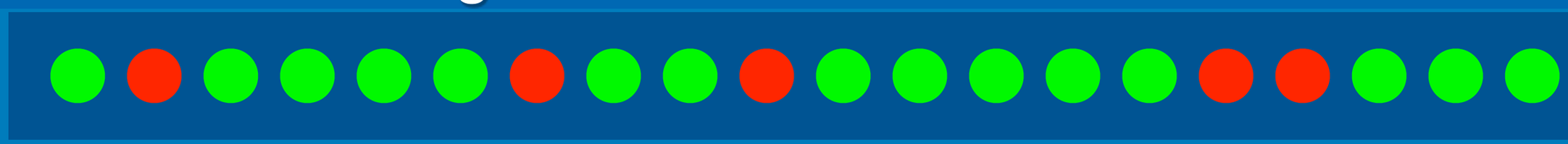


{ a, b, d } { a, c, d } { c, d, f } { c, d, f } { b, e, f }

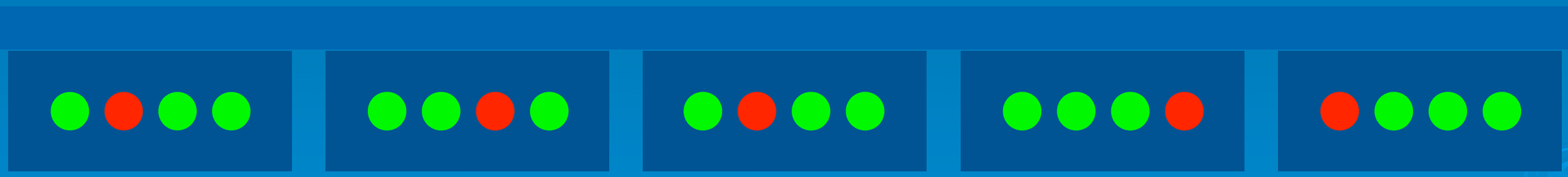
Machine Learning Techniques use parts of data.
Resampling, Validations Sets, Bagging, …

Whole loaf vs. sliced bread

Pattern Mining considers whole database at once



Th_3 = { a, b, c }



{ a, b, d } { a, c, d } { c, d, f } { c, d, f } { b, e, f }

Machine Learning Techniques use parts of data.
Resampling, Validations Sets, Bagging, …

Divide et Impera

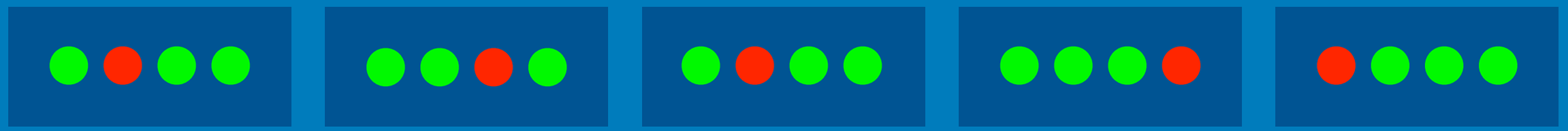
Splitting the database

f equal sized, stratified, non-overlapping folds

f equal sized, stratified, overlapping folds

United we Stand

How to merge *f* pattern sets



d	d	d	d	*4*
	c	c	c	*3*
		f	f	f	*3*
a	a				*2*
b				b	*2*
				e	*1*

United we Stand

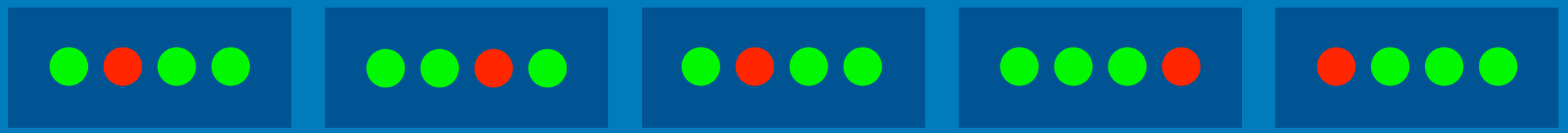
How to merge *f* pattern sets

Sort based on **COUNT**:

d, *c*, *f*, a, b, e

United we Stand

How to merge ***f*** pattern sets

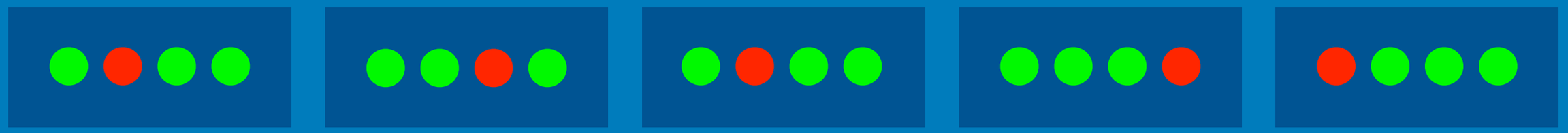


b	a	d	f	e	***3***
a	d	c	d	f	***2***
d	c	f	c	b	***1***

Sort based on average **RANK**:
d(1.6), f(1.2), a(1), b(0.8), c(0.8), e(0.6)

United we Stand

How to merge *f* pattern sets



Sort based on correlation **SCORE** (on whole dataset)

b, *e*, *d*, *c*, a, f

Combinatorics

Split

non-overlap or **overlap** based on **f** folds

Mine

top-**k** correlated pattern from each part

Merge

count, rank, or **score** - take top **n**

f = 3, 5, 7
k = 10, 25, 50, 75, 100
n = **k**

South America
Drug production

Amsterdam
Drug consumption

South America
Drug production


Amsterdam
Drug consumption
Leuven
Drug screening
South America
Drug production

Overview

- Correlated Pattern Mining
- Pattern Languages
- What about...
  - runtime w/o stepwise approach ? **Q1**
  - expressiveness vs. runtime ? **Q2**
  - predictive accuracy vs. runtime ? **Q3**
- Conclusions

Data: A. Karwath, J. Kazius

IQ-Project EU grant IST-FET FP6-516169

Thank you very much

Data: A. Karwath, J. Kazius

IQ-Project EU grant IST-FET FP6-516169

Thank you very much

Data: A. Karwath, J. Kazius

IQ-Project EU grant IST-FET FP6-516169