{ Mining, Sets, of, Patterns }

A tutorial at ECMLPKDD2010

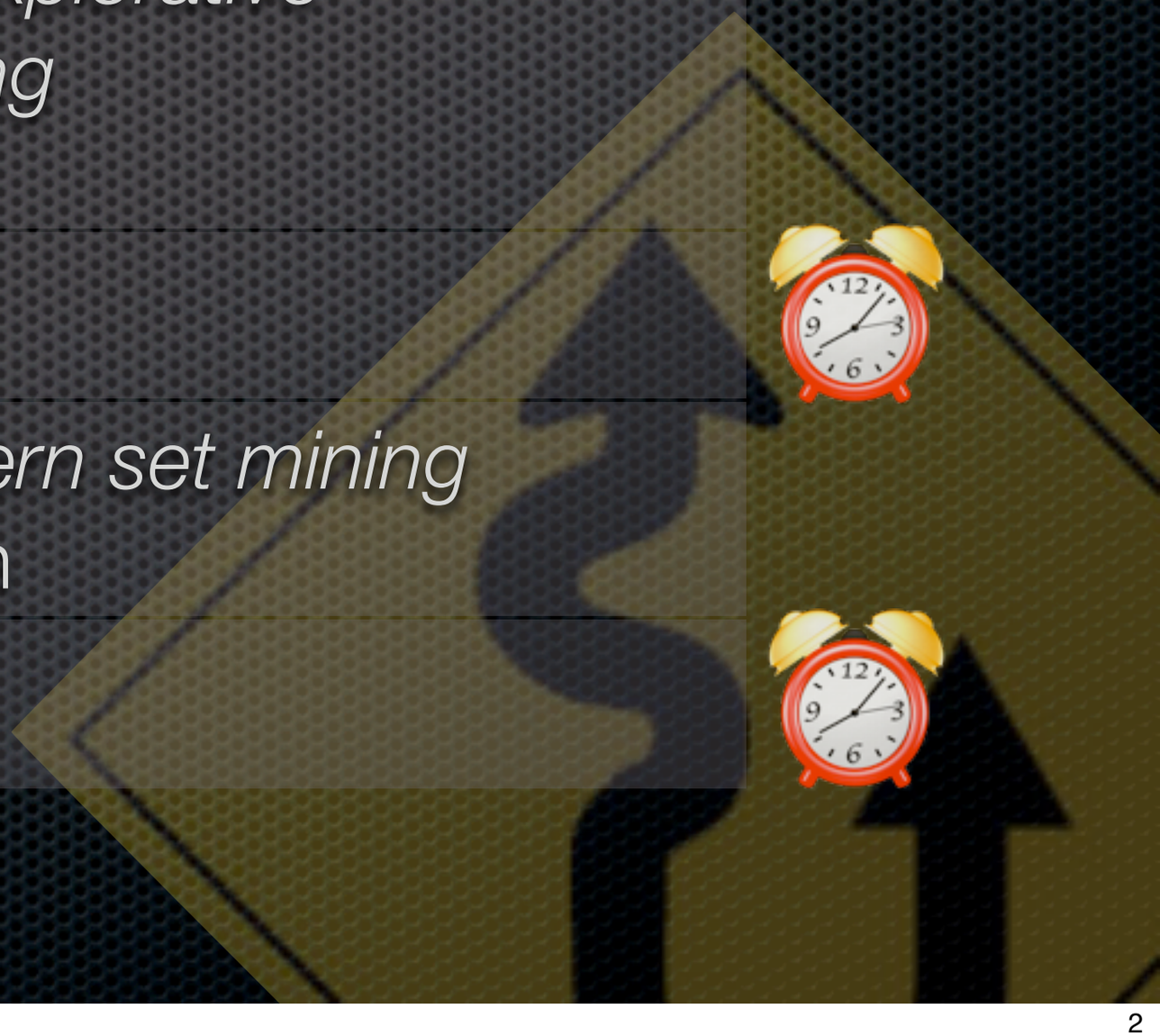
September 20, 2010, Barcelona, Spain

by

B. Bringmann, S. Nijssen, N. Tatti, J. Vreeken, A. Zimmermann

Overview Tutorial

00:00	*Introduction* Siegfried Nijssen
00:45	*Unsupervised, explorative pattern set mining* Jilles Vreeken
01:30	Break
02:00	*Supervised pattern set mining* Björn Bringmann
02:45	End



Practical information

- Even though we did our best to achieve otherwise:

WARNING

This TUTORIAL is neither complete nor unbiased

REFERENCES are not necessarily authoritative or complete

- More information (including references): http://www.cs.kuleuven.be/conference/msop/

Part I

Introduction

Overview part I

Patterns

Pattern sets

- Definitions
- Motivations
- Dimensions
- Algorithms

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

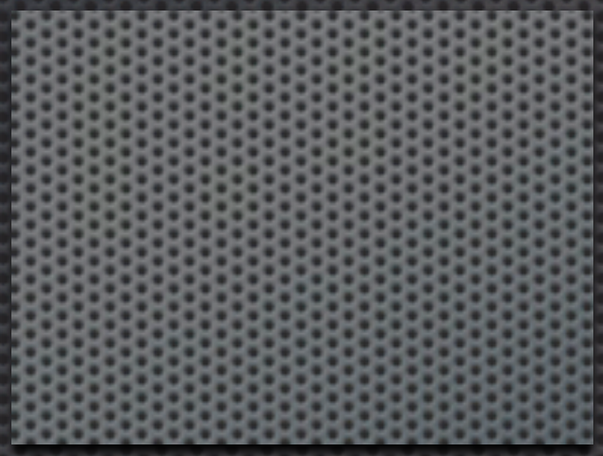
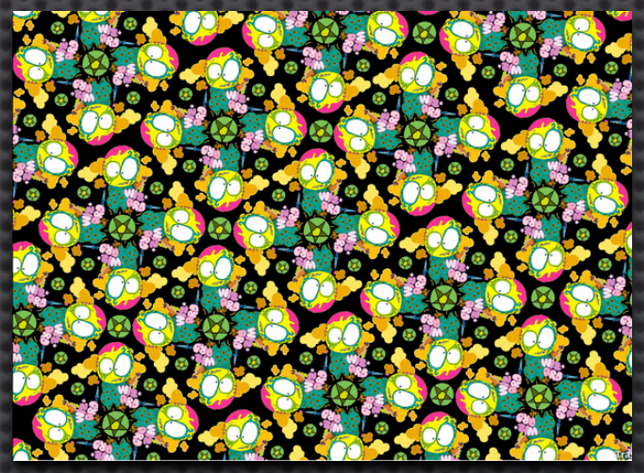
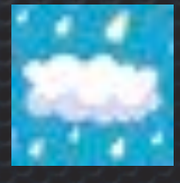
Overview part I

Patterns

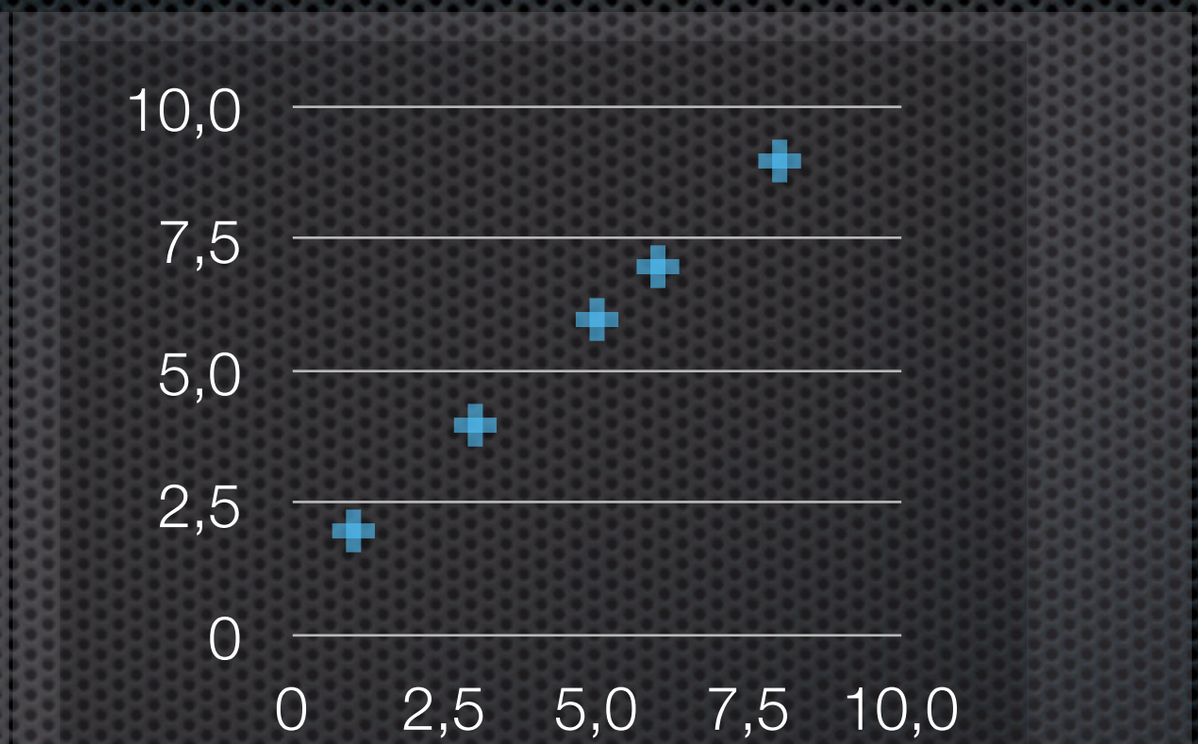
- Definitions
- Motivations
- Dimensions
- Algorithms

What is a pattern?

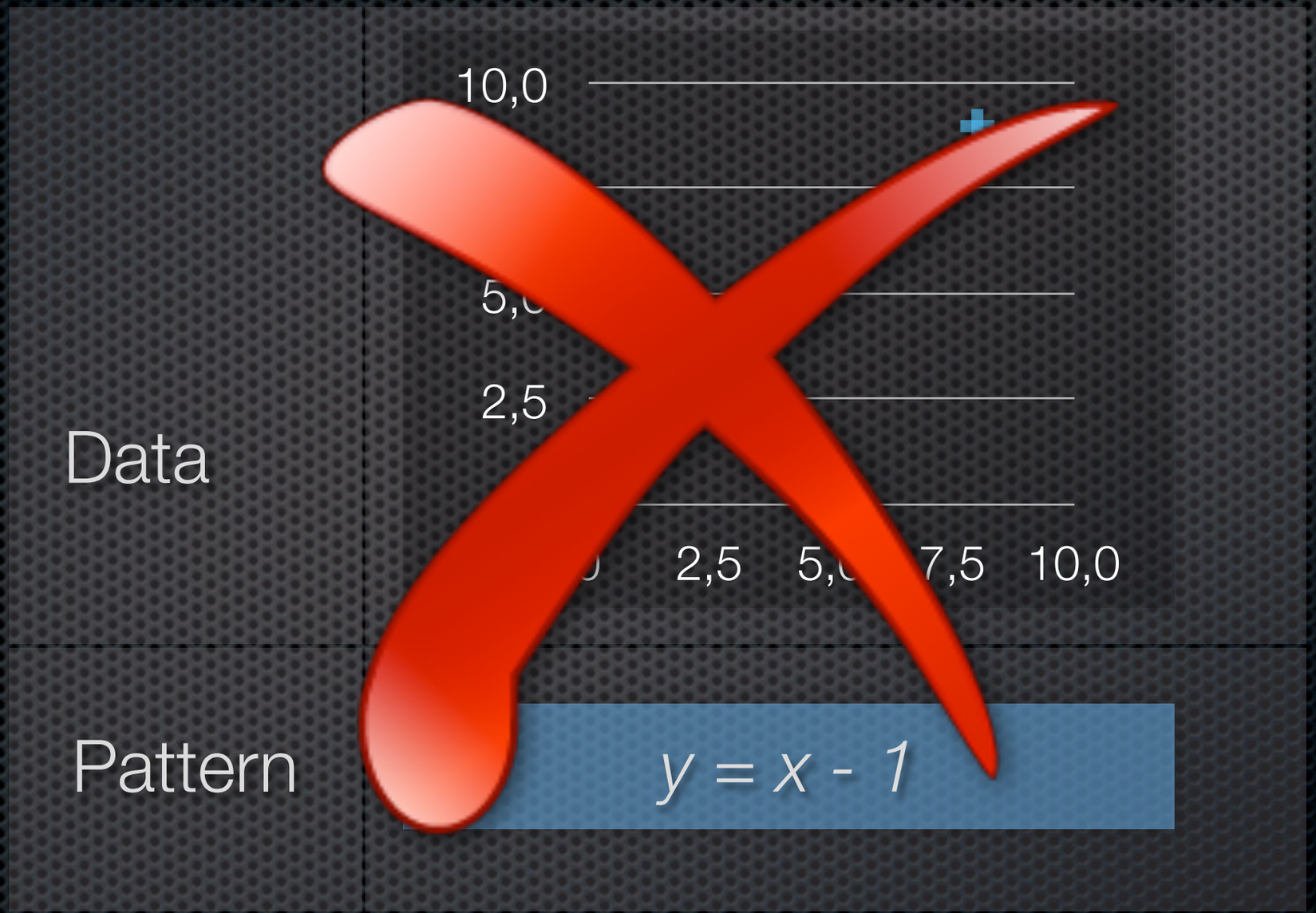
Recurring structure

Data			
Pattern			

What is a pattern?

Data	
Pattern	$y = x - 1$

What is a pattern?



What is a pattern?

In this tutorial we are looking for

- Recurring structures ...
- ... in enumerable, discrete domains

Hence we do not consider a regression model to be a pattern...

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

What is a pattern?

Example 1: Frequent Itemset in Market Basket Data

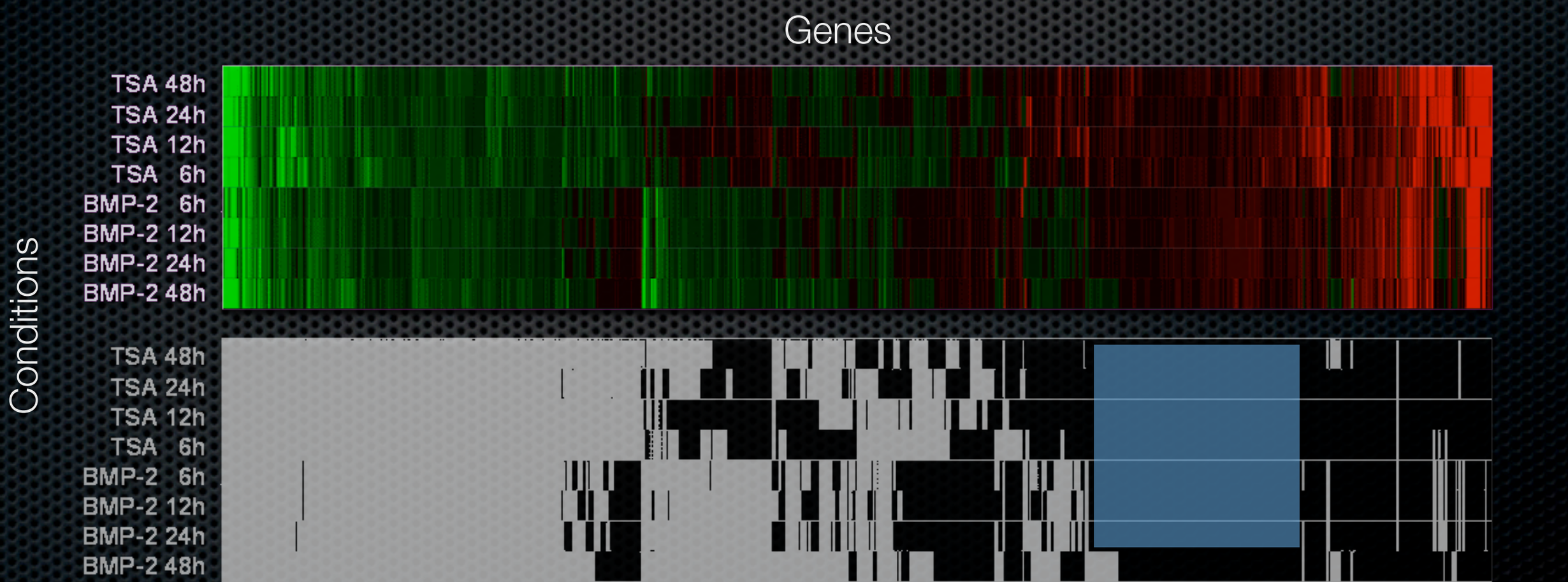
What is a pattern?

Example 1: Frequent Itemset in Market Basket Data

support( )=3

What is a pattern?

Example 2: Co-cluster in Gene Expression Data



Lyssiotis et al.

What is a pattern?

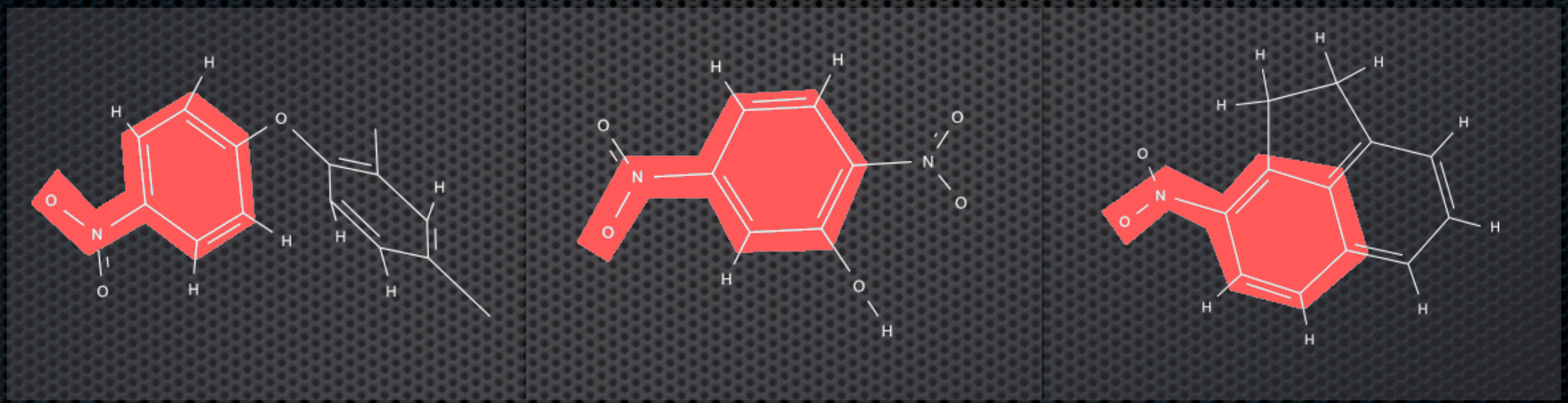
Example 3: Conjunctive Formula in UCI Data

```
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
```

Petal length >= 2.0
and Petal width <= 0.5

What is a pattern?

Example 4: Frequent Subgraph in Molecules



What is a pattern?

Recurring structure in enumerable, discrete domain

Enumerable, discrete domains:
itemsets, graphs, sequences, trees, ...

Recurrence as determined by constraints:
support constraint, size constraint, area constraint, ...

The problem: too many patterns

Too many patterns...

Solution 1: constraint-based mining

Solution 2: pattern set mining

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

Solution 1:
pattern constraints

Constraint on **each** pattern individually based on

- background knowledge
- condensed representations
- class labels

Constraints: background knowledge

- Support constraints
- Syntactical constraints

- Statistical constraints
  - difference with expectation
  - taxonomies

= diapers

Constraints: condensed representations

If we pass a pattern through the data, we obtain another pattern

Constraints: condensed representations

- Closed patterns

  Pasquier et al.

- Free/generator patterns

  Pasquier et al.

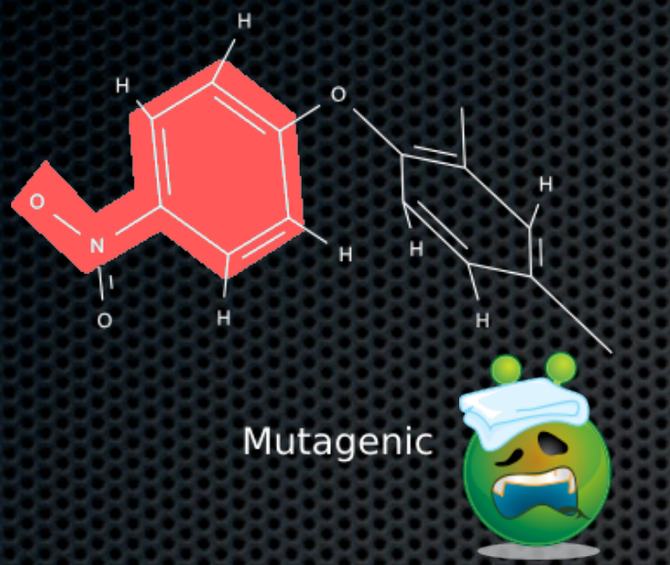
- Maximal frequent patterns

  Bayardo

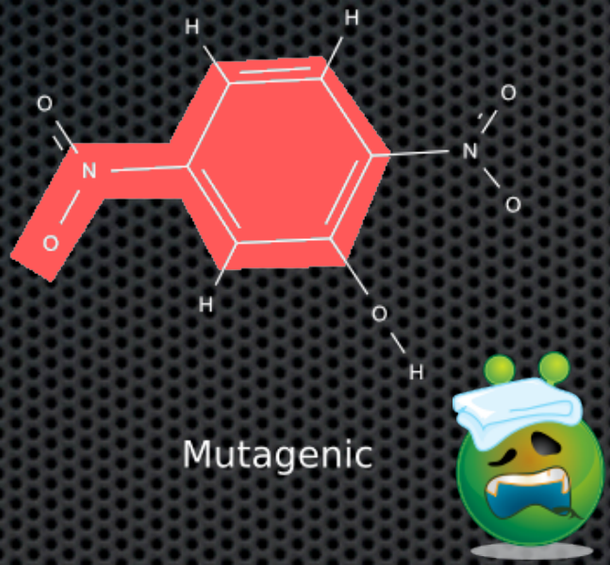
- Non-derivable patterns

  Calders et al.

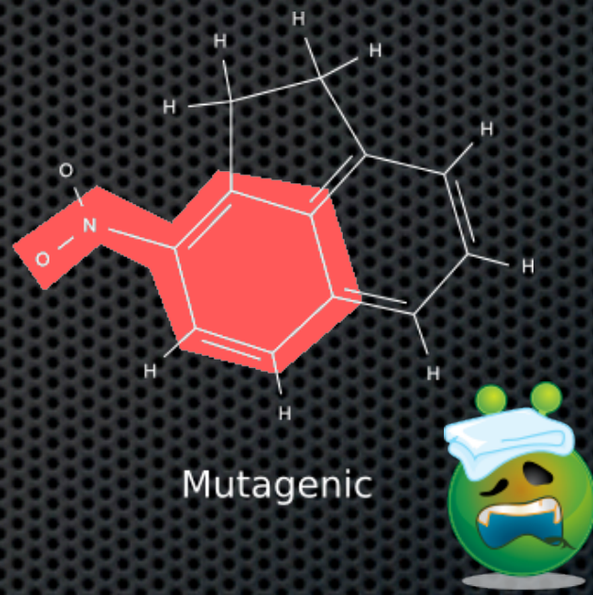
Constraints: class labels



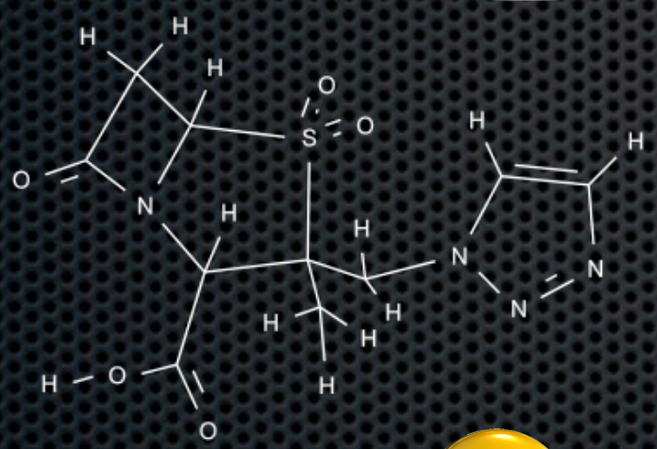
Mutagenic



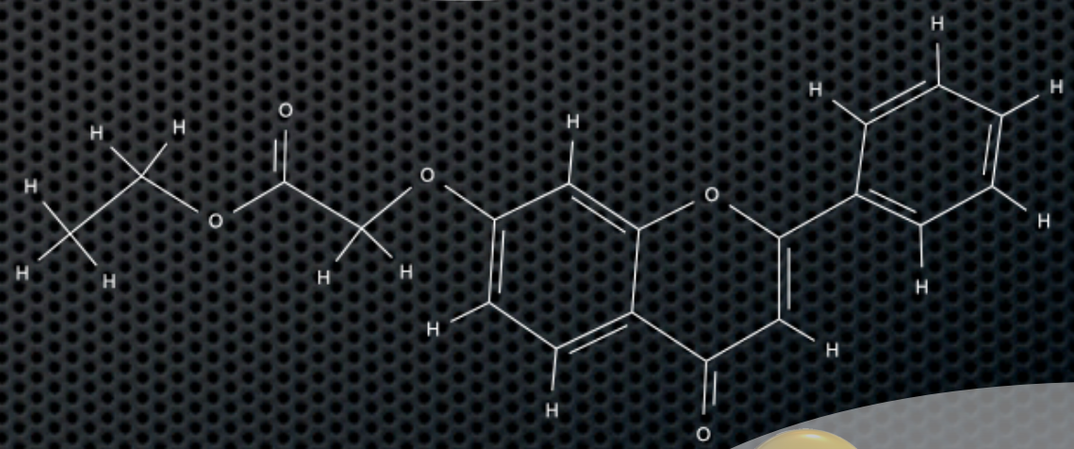
Mutagenic



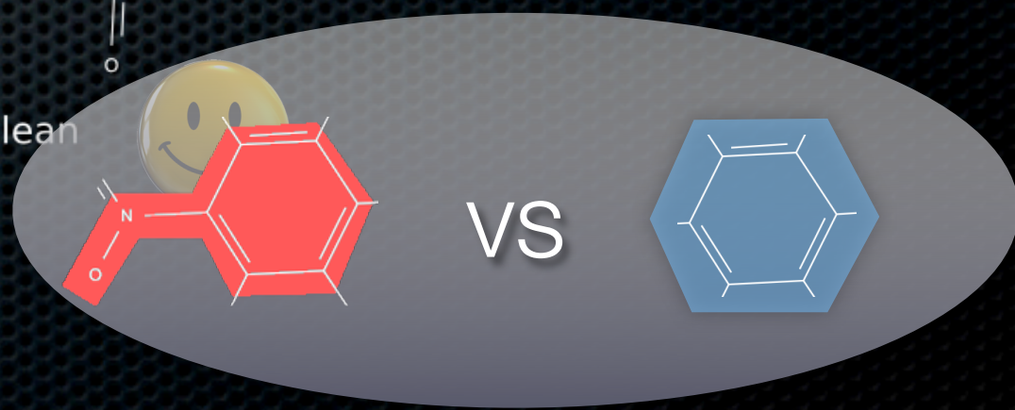
Mutagenic



Clean

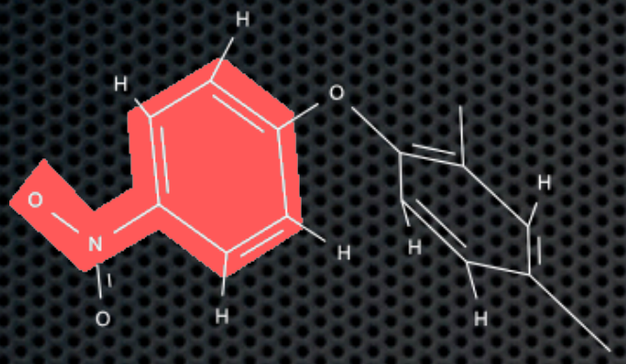


Clean

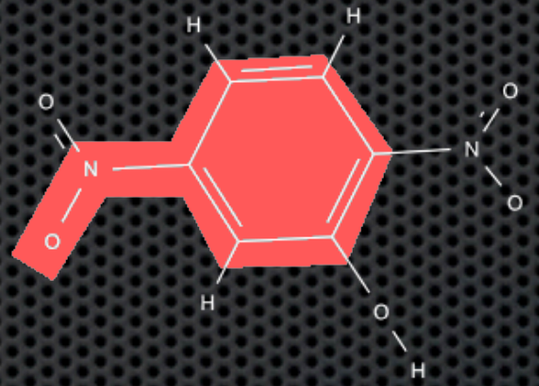


Constraints: class labels

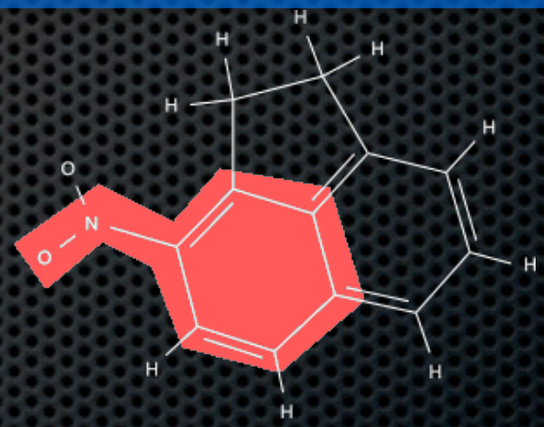
GIVEN database *D*, target *c*, threshold *t*, class of patterns
FIND **all** patterns *p* with *f(p,D,c)>t*



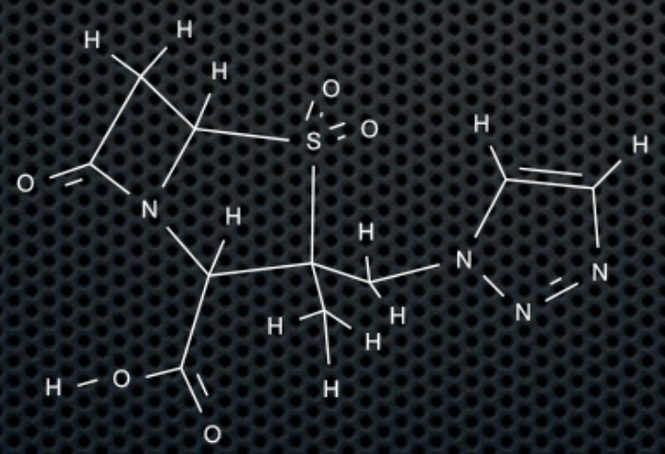
Mutagenic



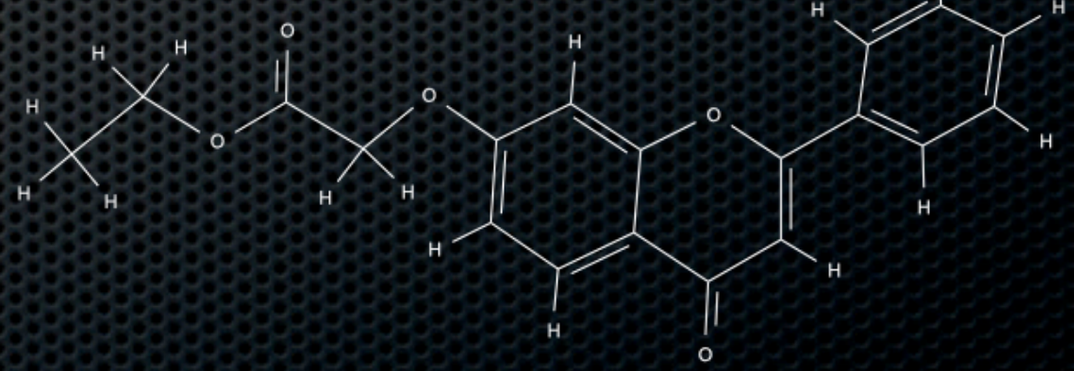
Mutagenic



Mutagenic



Clean



Clean

Constraints: class labels

Constraints: class labels

Many different names for this setting — Novak, Webb and Lavrac

Pattern name	Typical measure	
Emerging pattern	Growth rate	Dong et al.
Contrast set	Difference in rel support	Bay et al.
Correlated pattern	Chi2	Morishita et al.
Subgroup	Weighted relative accuracy	Kloesgen et al.
Discriminative pattern	Information gain	Cheng et al.
Class association rule	Confidence	Liu et al.

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms

How to find patterns?

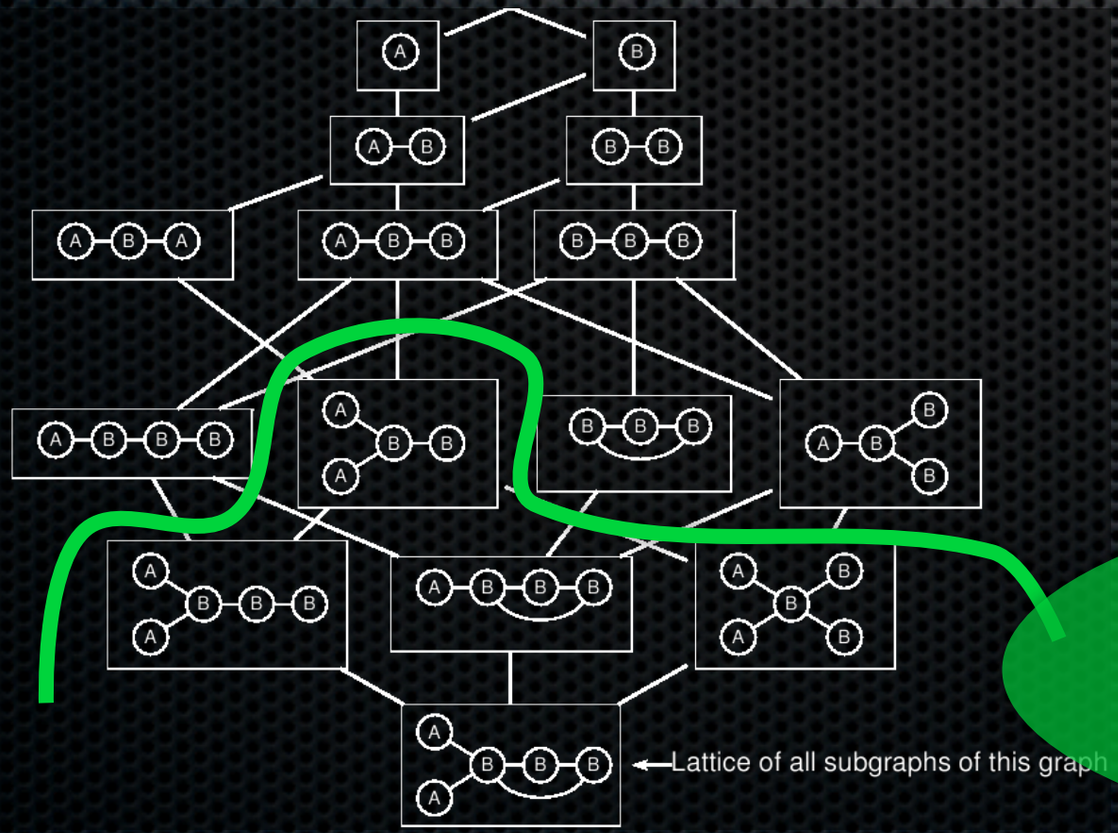
In principle two ways:

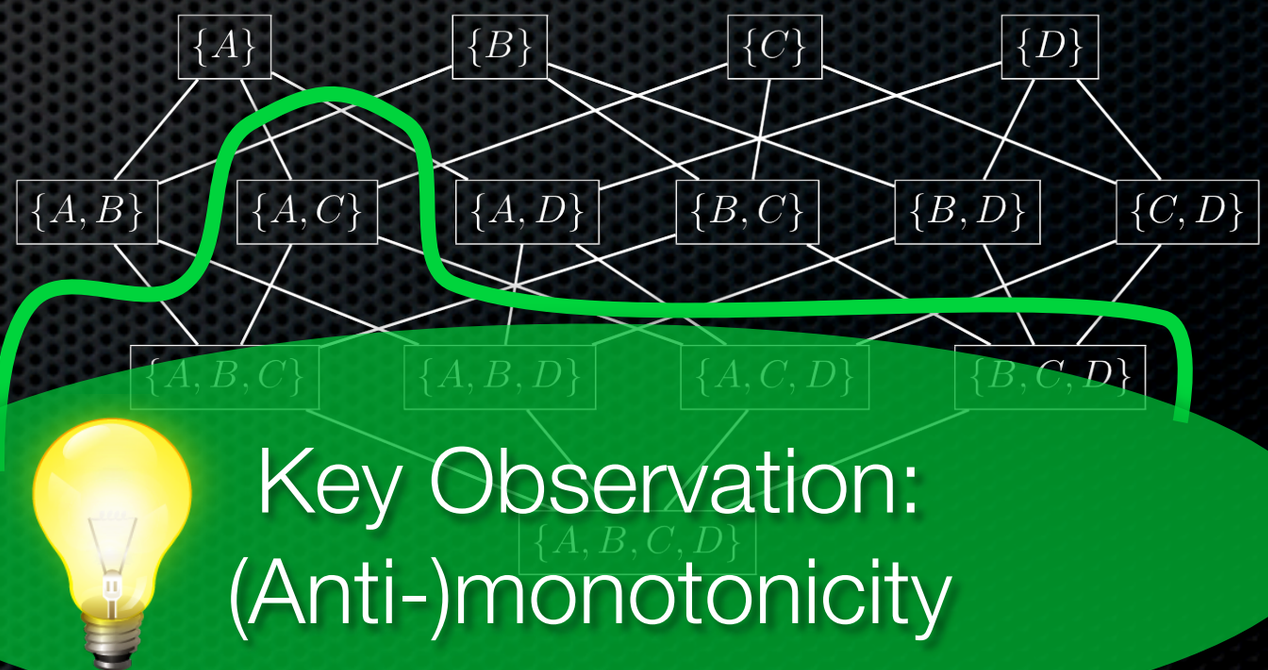
- Greedy / heuristic
  - ➕ Fast
  - ❌ Overlooks solutions
- Complete search
  - ➕ Finds everything
  - ❌ Slower

How to find patterns?

Complete search under constraints often feasible

GIVEN database *D*, constraint **φ** on *D*, class of patterns *C*
FIND all patterns *p* in class *C* satisfying **φ**





Key Observation:
(Anti-)monotonicity

Lots of solutions...
what's their problem?

Overview part I

Patterns

A

Motivations

- Definition
- Dimensions
- Algorithms

Overview part I

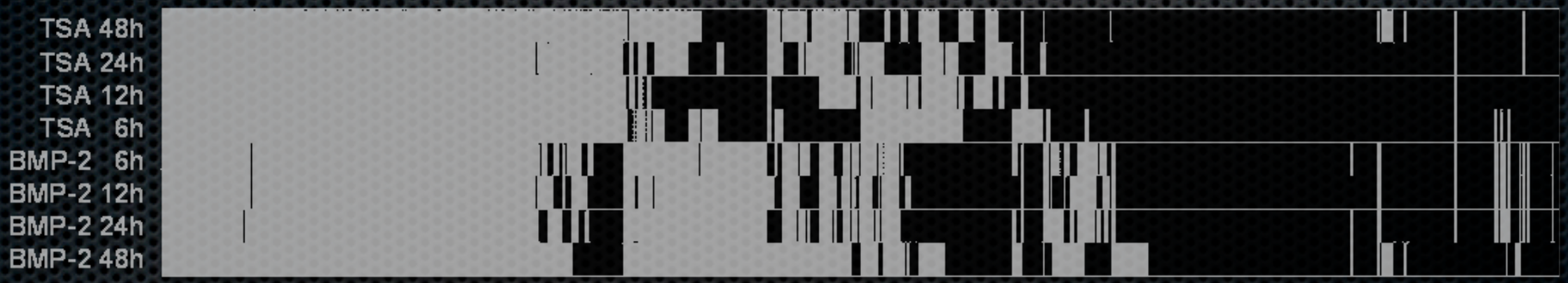
Pattern sets

Motivations

- Definition
- Dimensions
- Algorithms

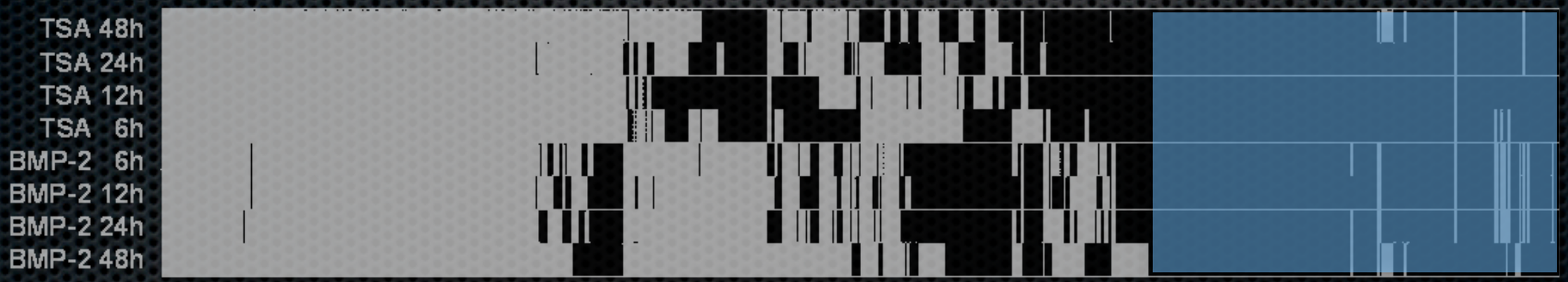
The problem - complex pattern relationships

Unsupervised descriptive task



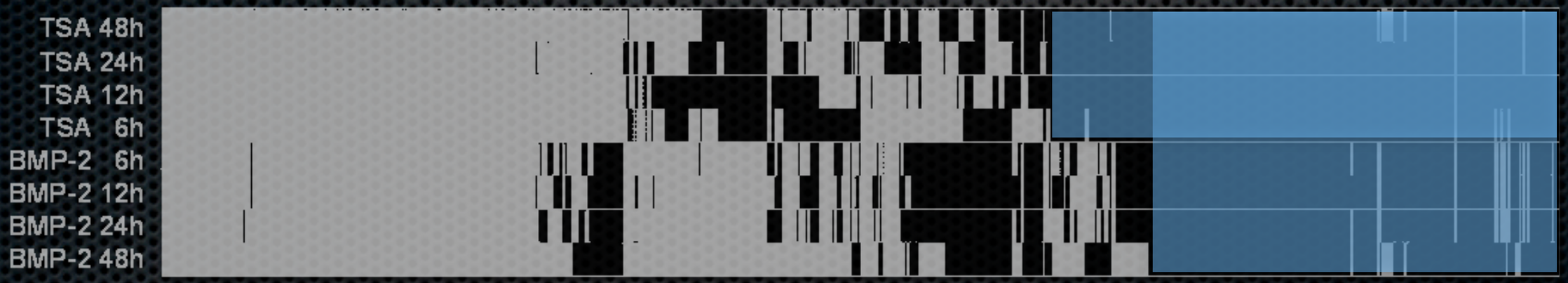
The problem - complex pattern relationships

Unsupervised descriptive task



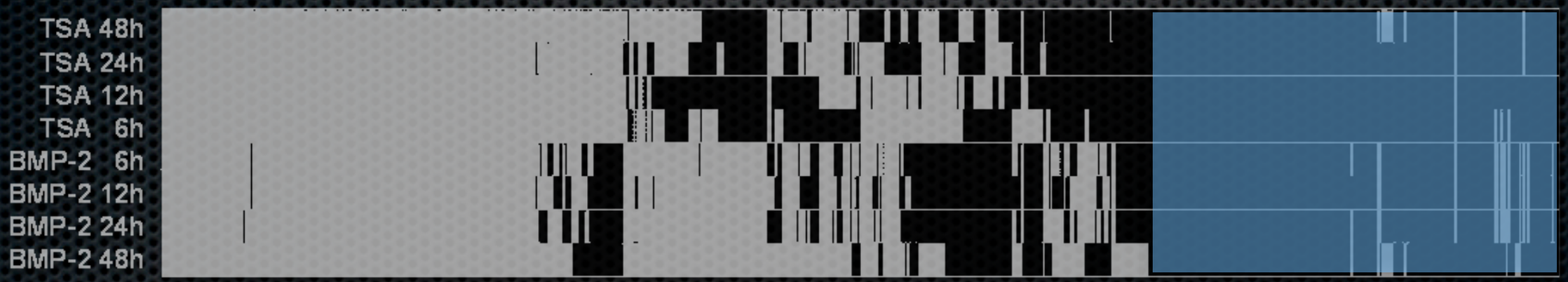
The problem - complex pattern relationships

Unsupervised descriptive task



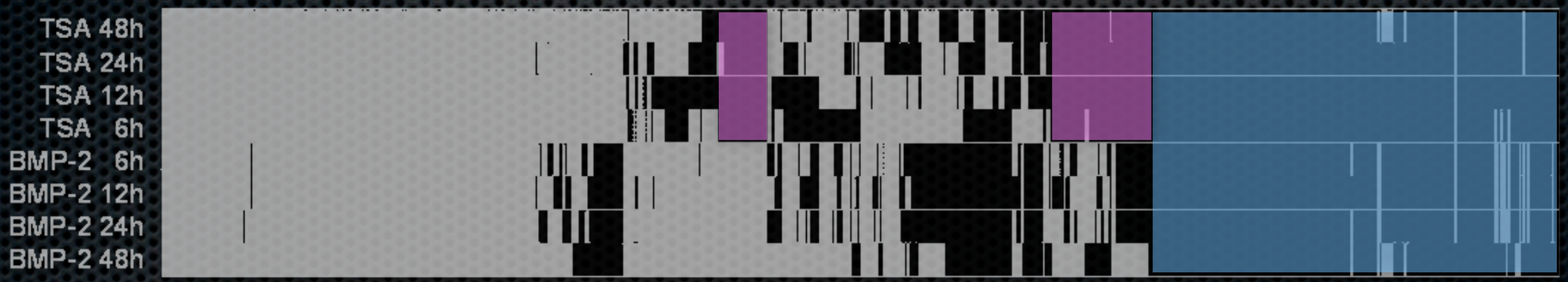
The problem - complex pattern relationships

Unsupervised descriptive task



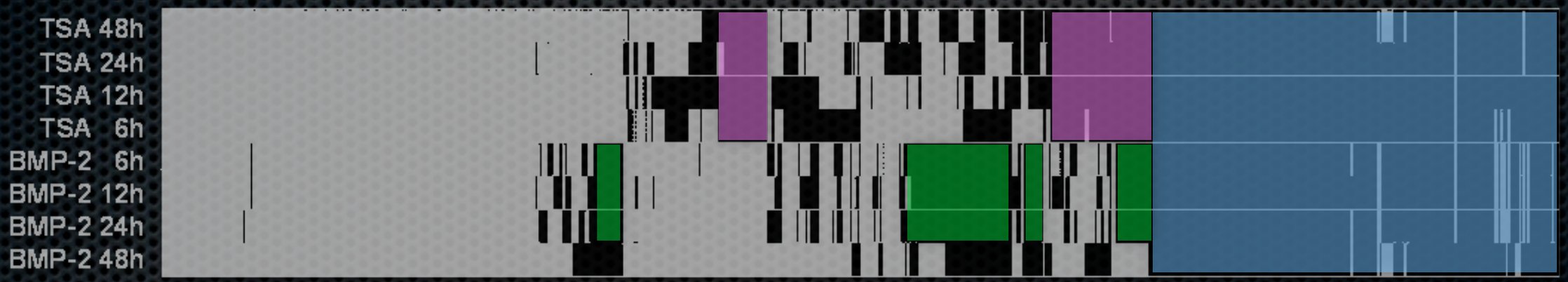
The problem - complex pattern relationships

Unsupervised descriptive task



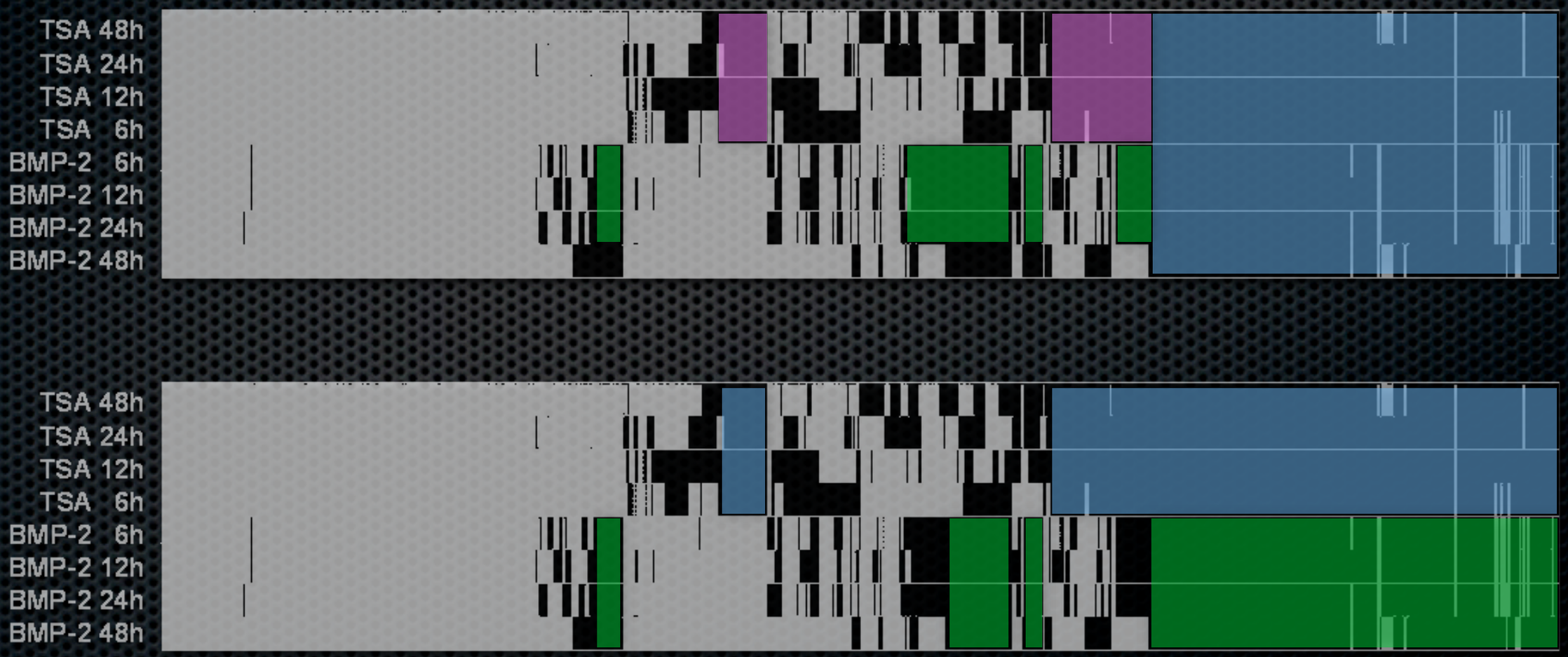
The problem - complex pattern relationships

Unsupervised descriptive task



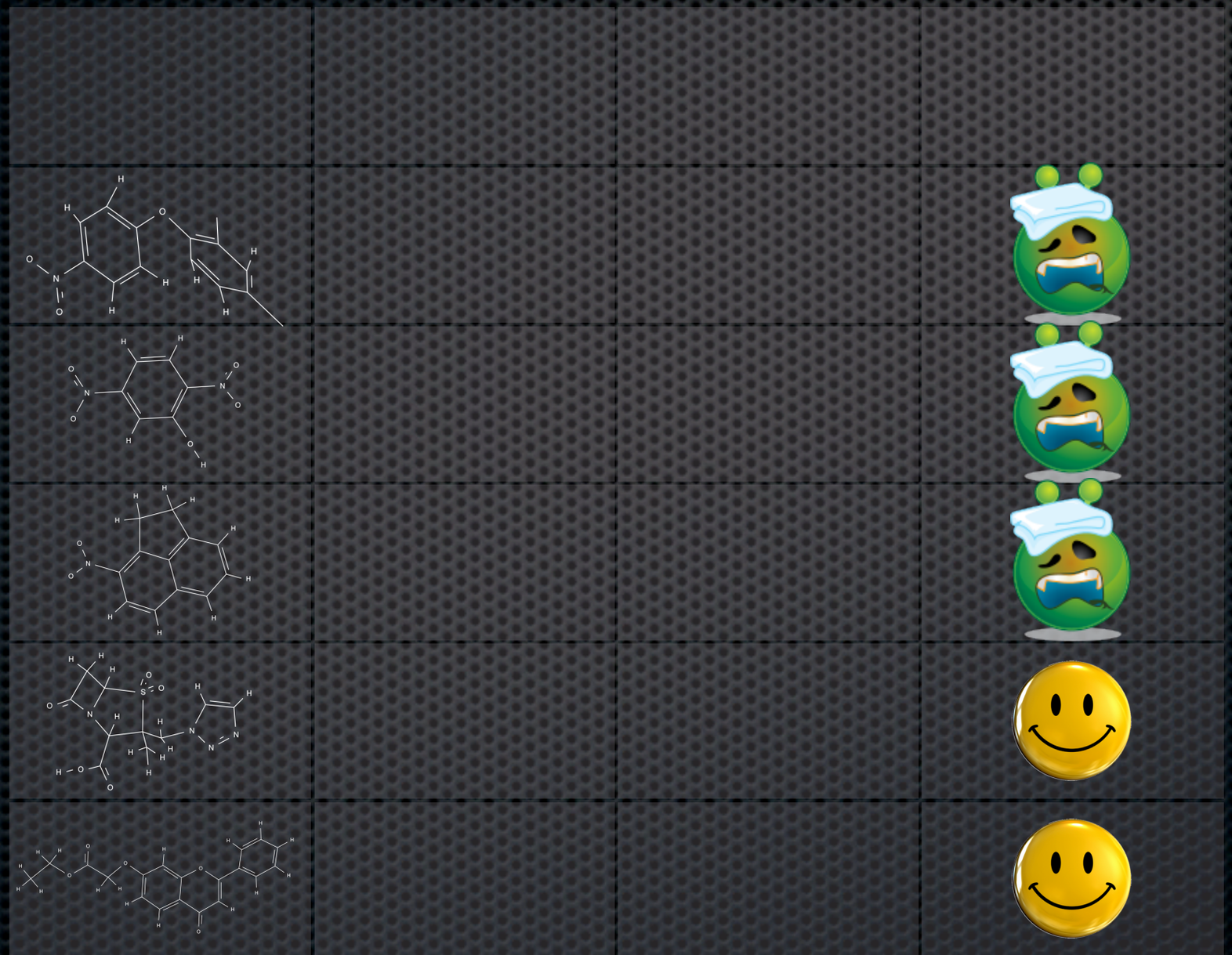
The problem - complex pattern relationships

Unsupervised descriptive task



The problem - complex pattern relationships

Supervised predictive task



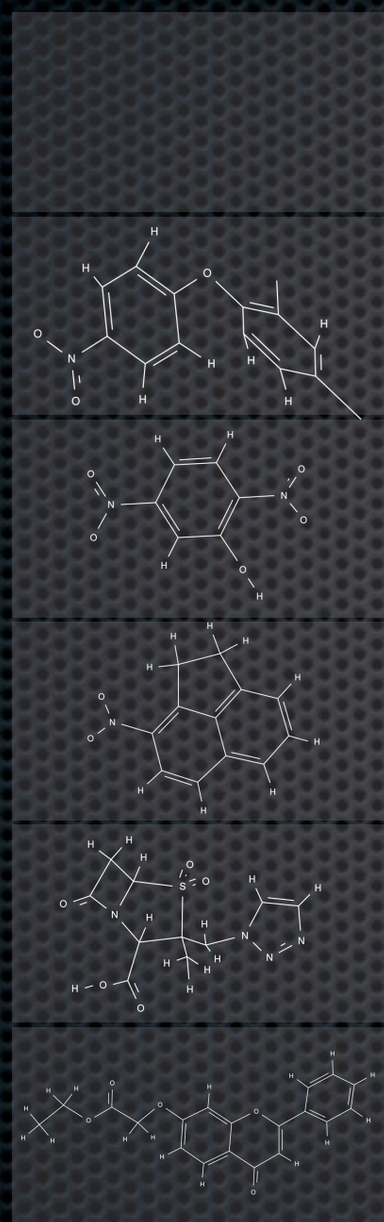
The problem - complex pattern relationships

Supervised predictive task

All patterns mined

P1

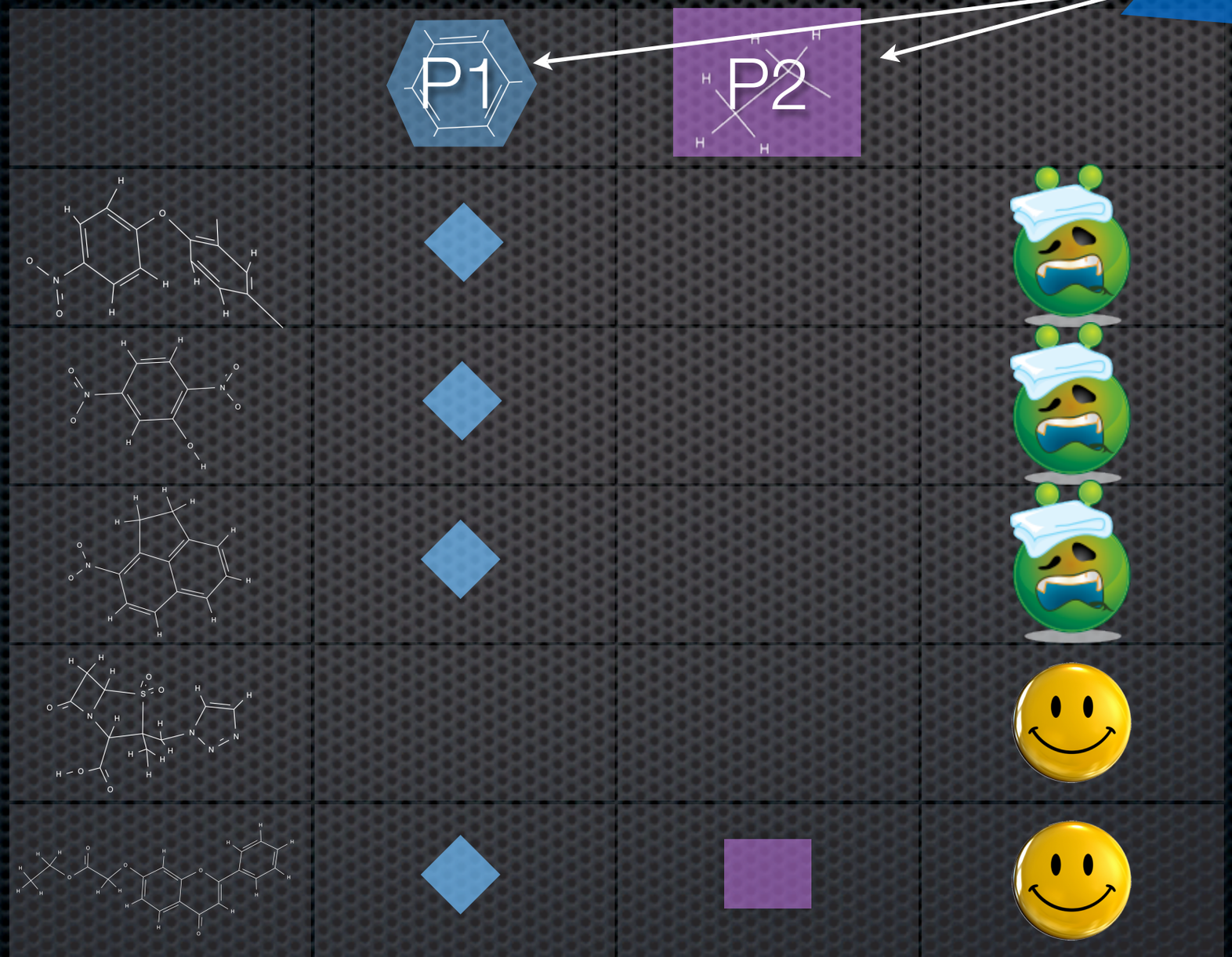
P2



The problem - complex pattern relationships

Supervised predictive task

All patterns mined



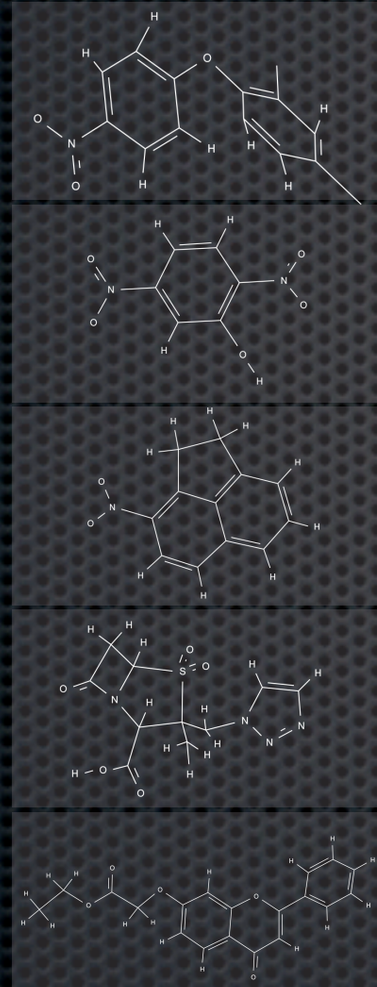
The problem - complex pattern relationships

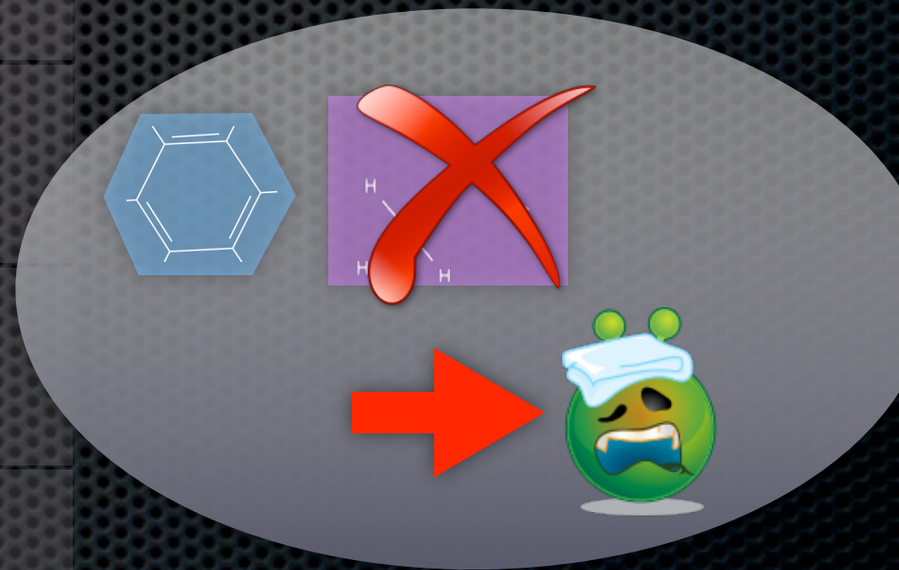
Supervised predictive task

All patterns mined

P1

P2





Overview part I

Pattern sets

Motivations

- Definitions
- Dimensions
- Algorithms

Overview part I

Pattern sets

- Motivations
- Definitions
- Dimensions
- Algorithms

Pattern set mining

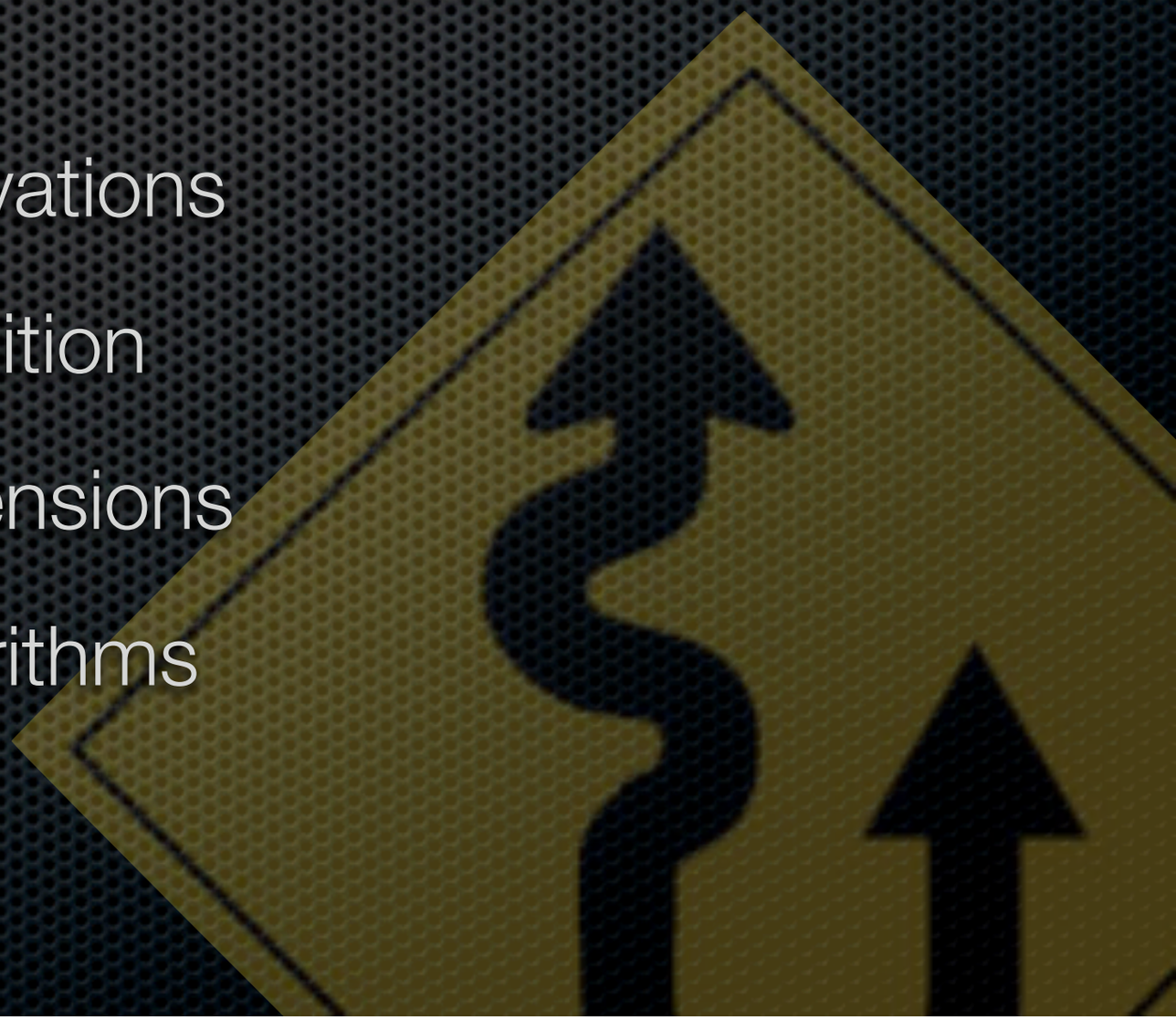
GIVEN a data mining task

FIND an interrelated set of patterns useful for this task

Overview part I

Pattern sets

- Motivations
- Definition
- Dimensions
- Algorithms



Overview part I

Pattern sets

- Motivations
- Definition
- Dimensions
- Algorithms

Patterns vs Pattern sets

	unsupervised	supervised
pattern mining	**no** target **no** relationships	**relevant** to target **no** relationships
pattern set mining	**no** target **relationships**	**relevant** to target **relationships**

part II **part III**

Task dimensions

	unsupervised	supervised
descriptive	**Association Analysis** Tiling (Co-)Clustering Probabilistic models	Subgroup discovery Exceptional model mining
predictive	Predictive clustering	**Classification** Regression

part II **part III**

Task dimensions

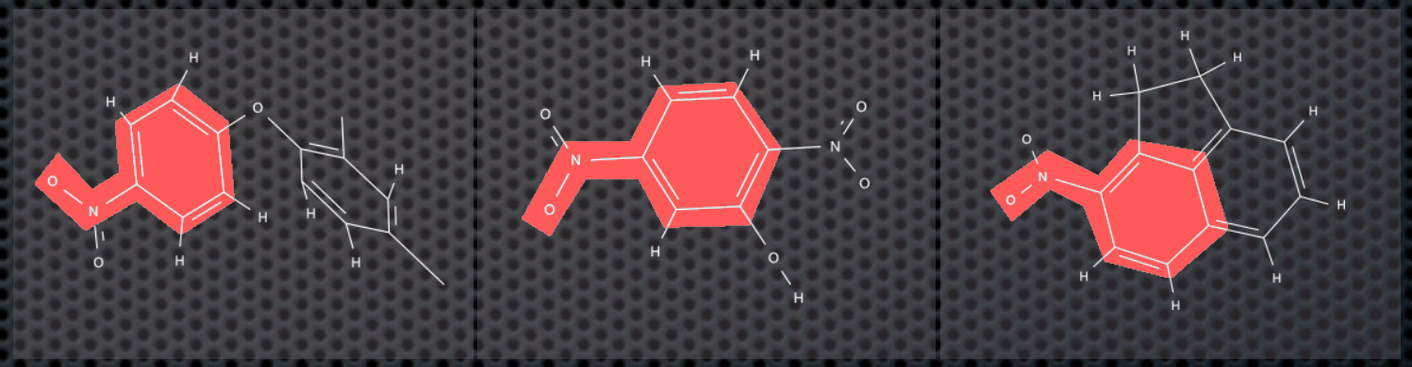
- **Supervised vs unsupervised**
- **Predictive vs descriptive**

Task dimensions

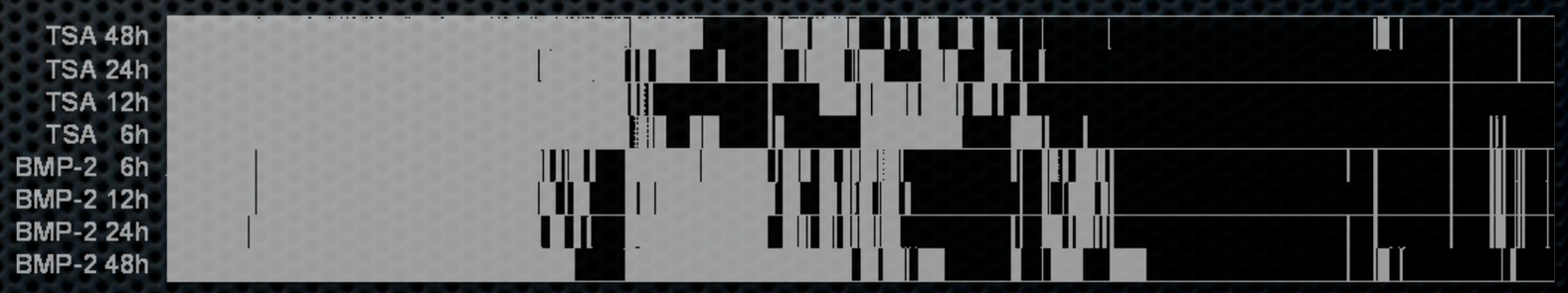
- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data

Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data



VS



Task dimensions

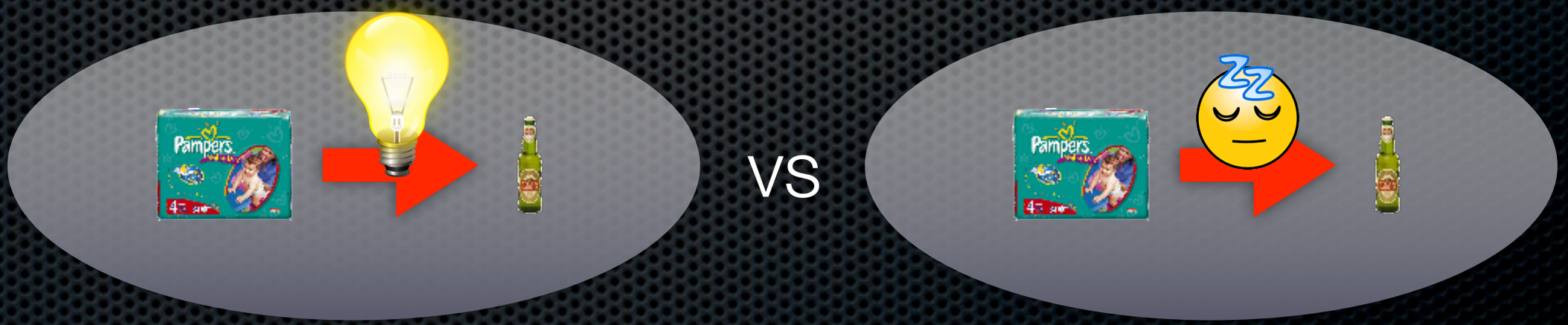
- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data

Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained

Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained



Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained

Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained
- Interpretable model vs Black box

Overview part I

Pattern sets

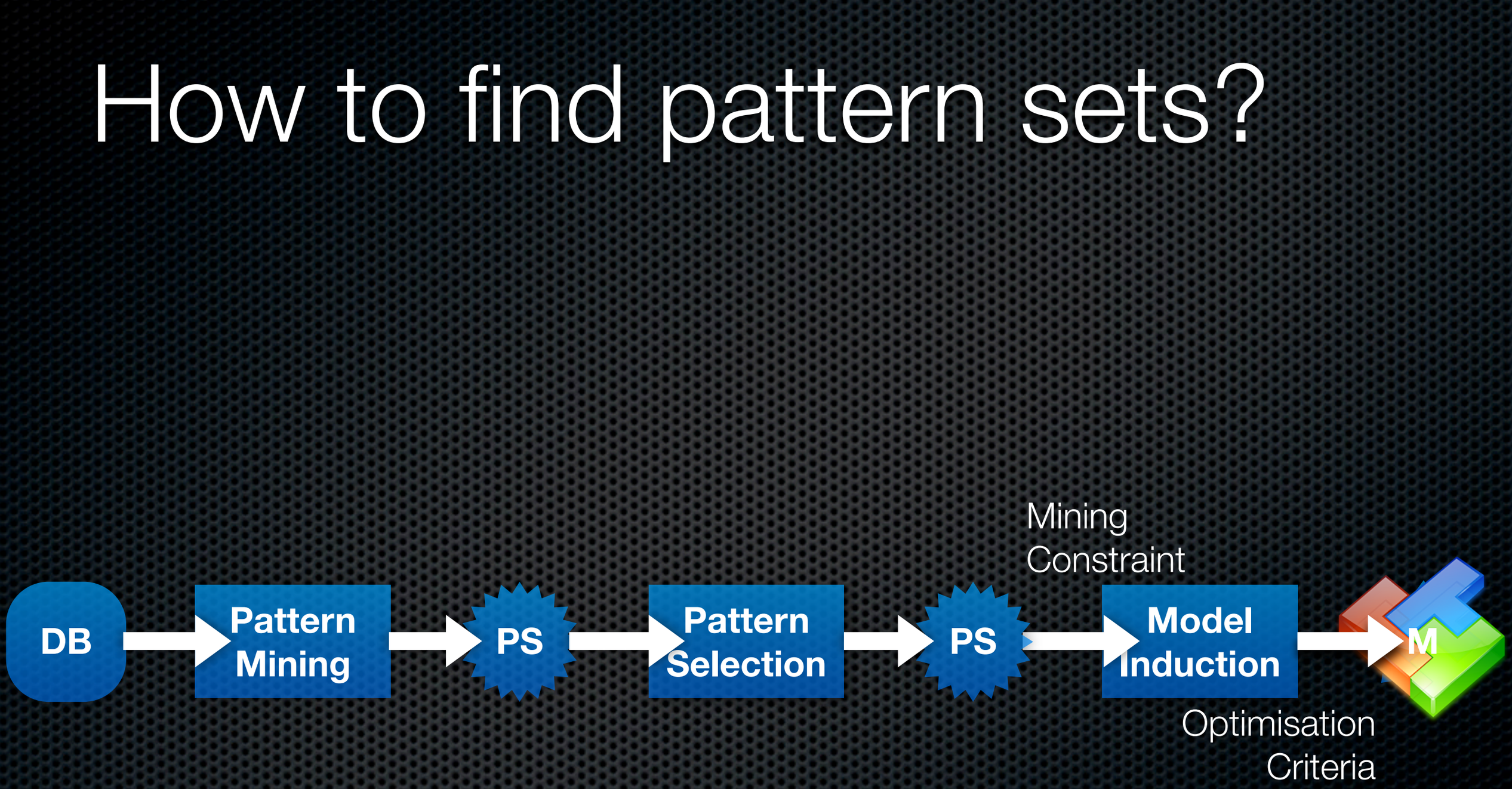
- Motivations
- Definitions
- Dimensions
- Algorithms

Overview part I

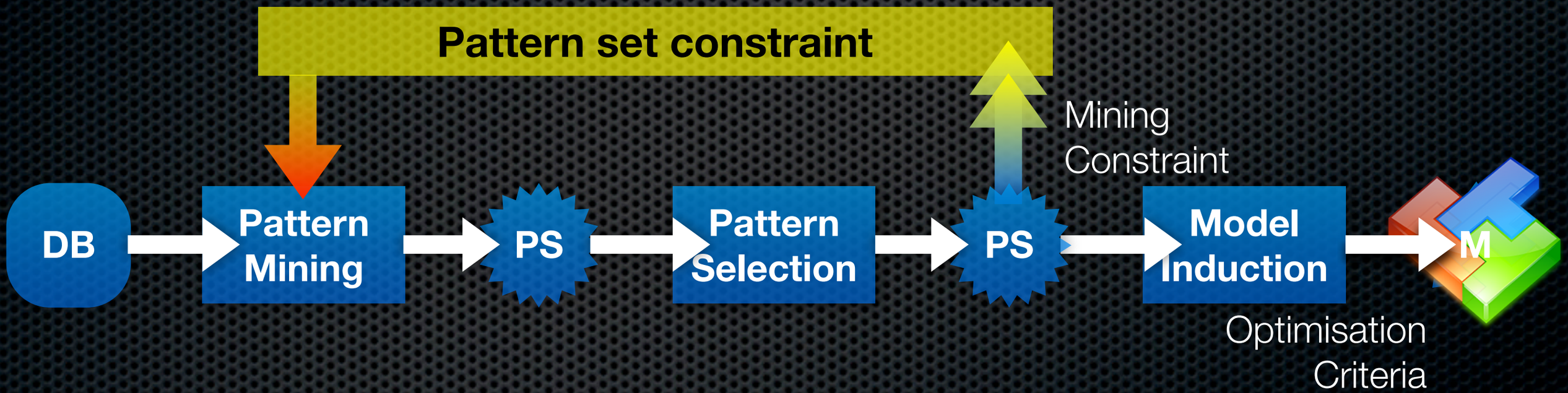
Pattern sets

- Motivations
- Definitions
- Dimensions
- Algorithms

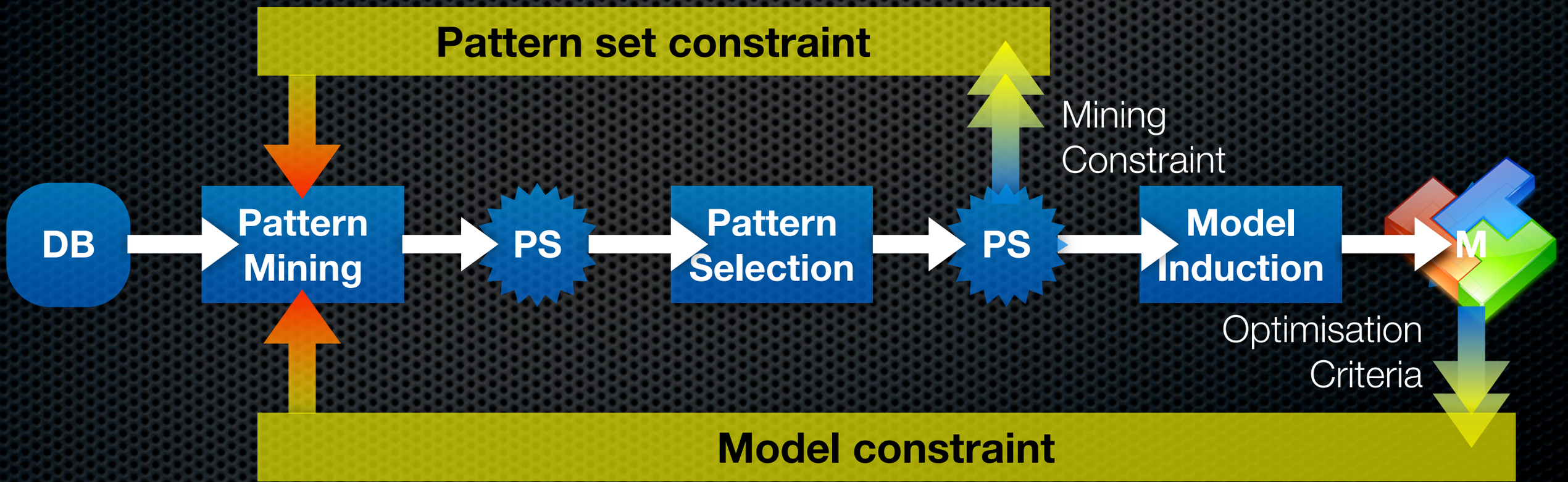
How to find pattern sets?



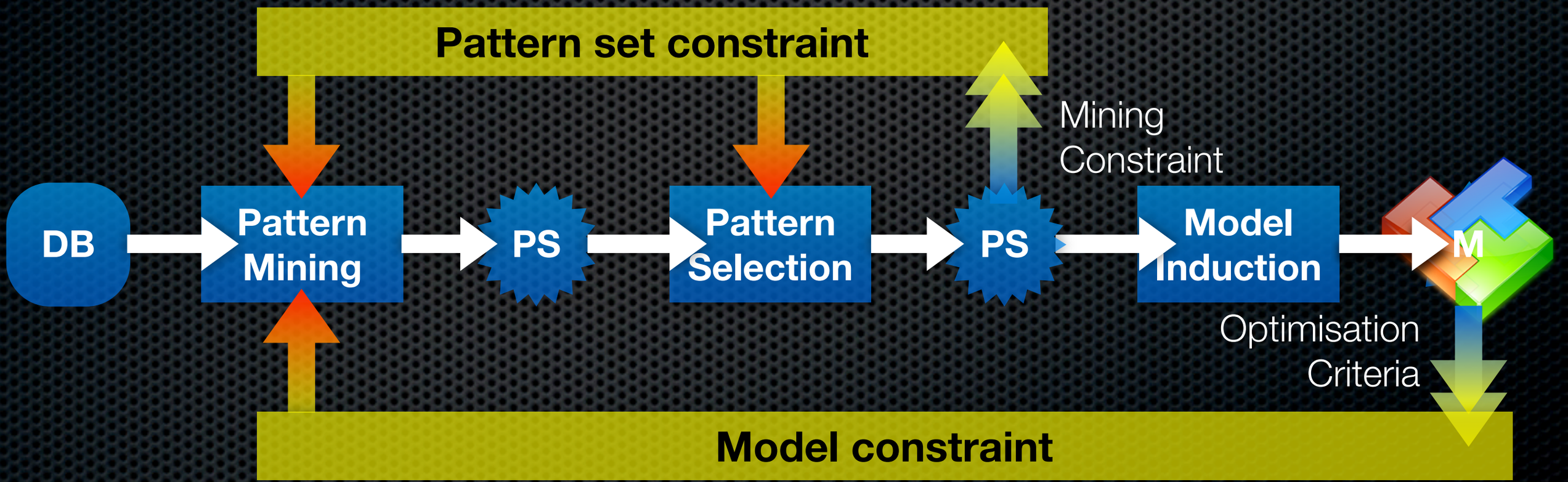
How to find pattern sets?



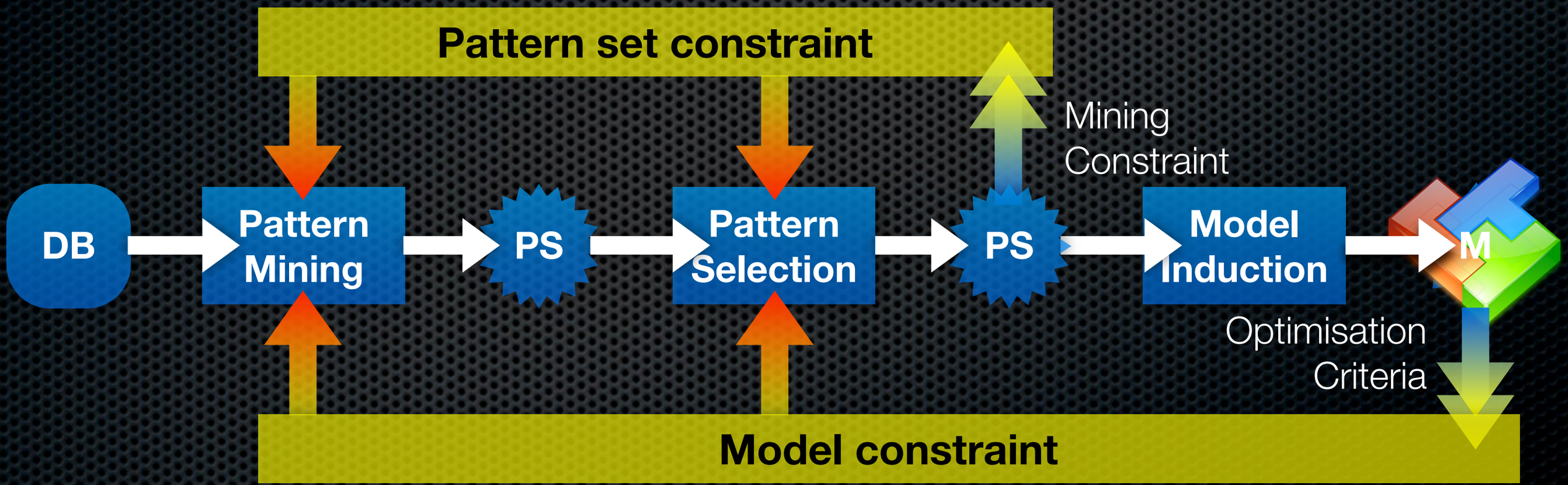
How to find pattern sets?



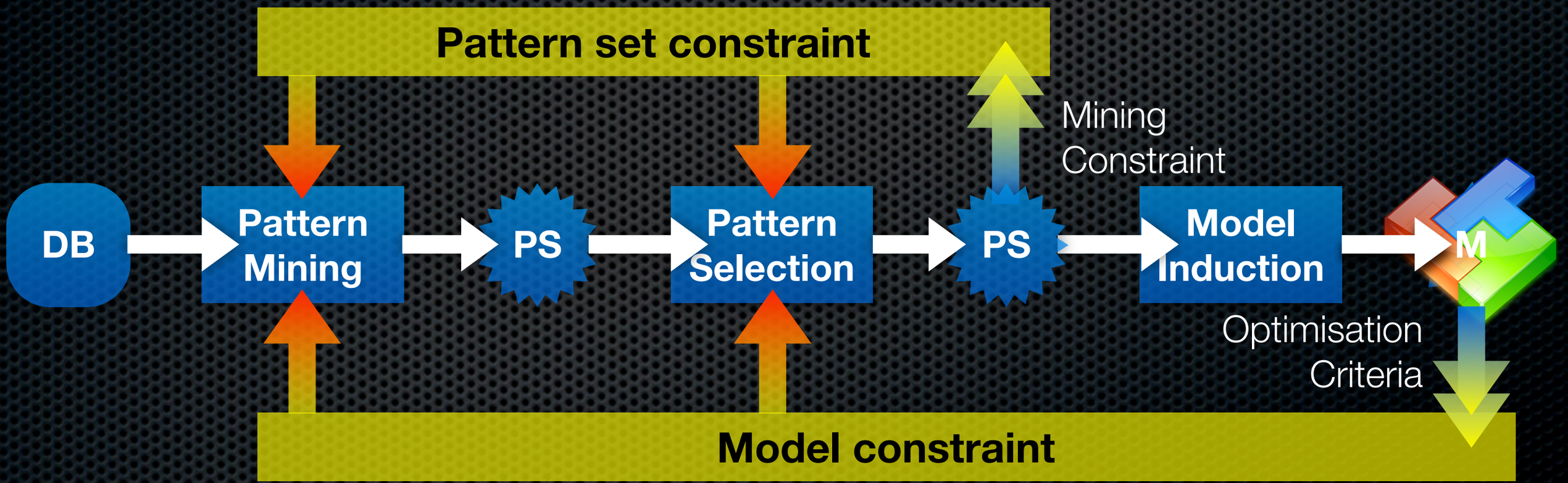
How to find pattern sets?



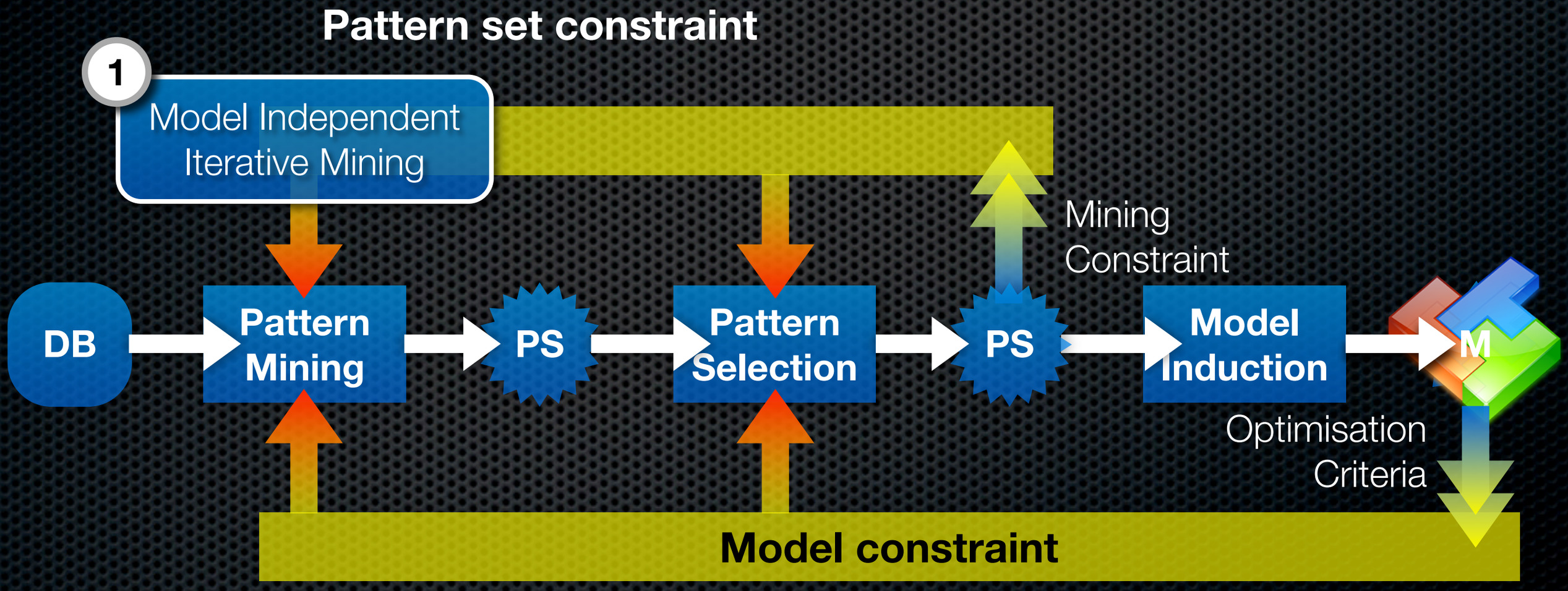
How to find pattern sets?



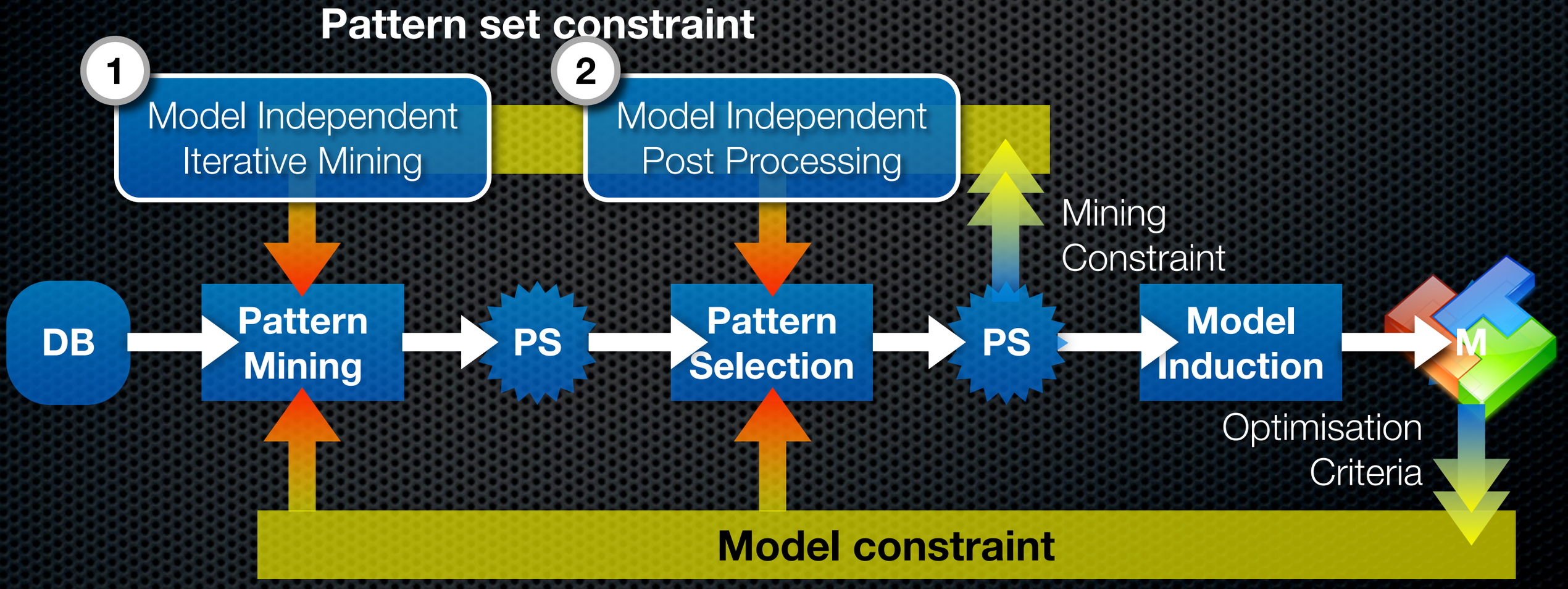
How to find pattern sets?



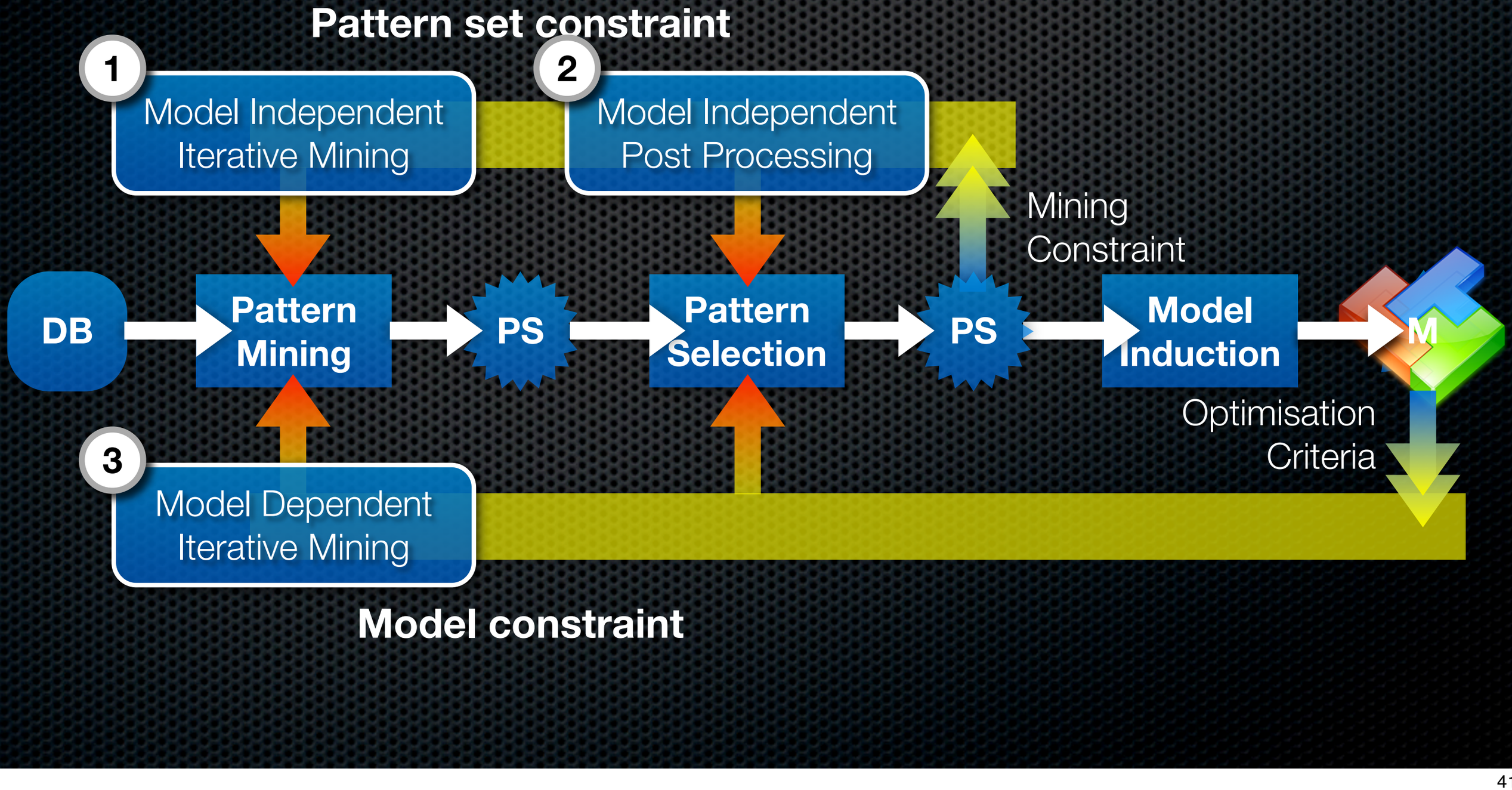
How to find pattern sets?



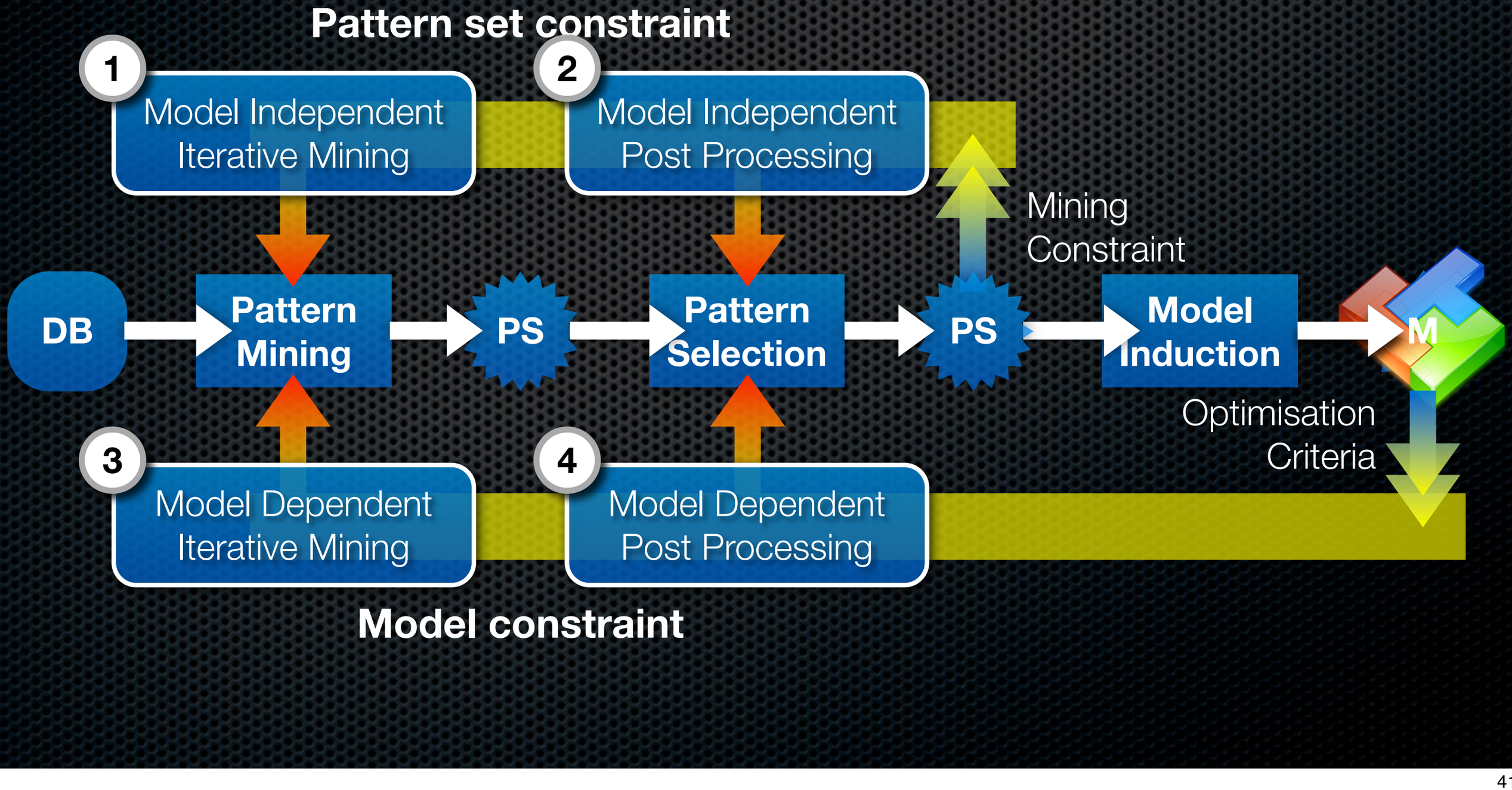
How to find pattern sets?



How to find pattern sets?



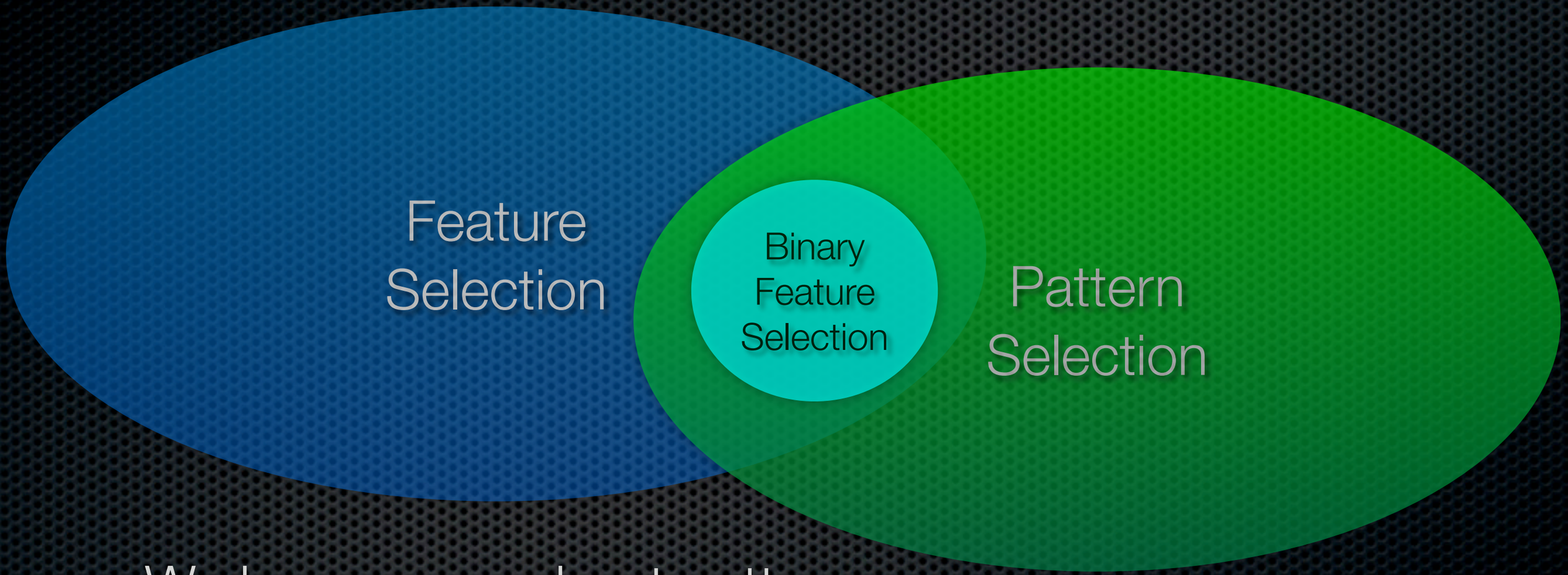
How to find pattern sets?



Pattern set = Feature set

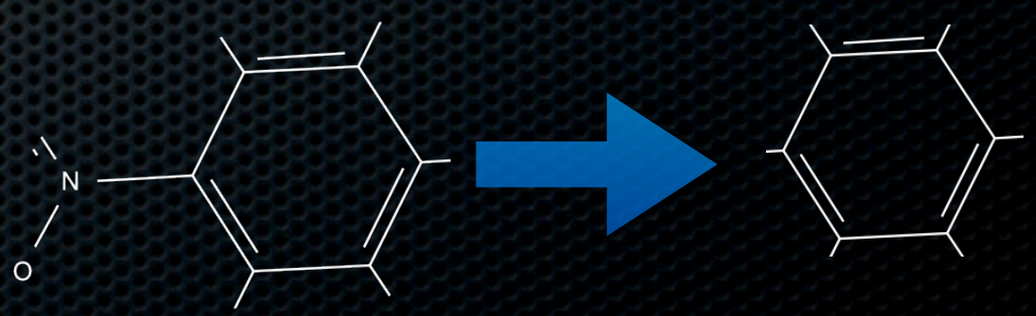
Pattern set = Feature set

Feature vs pattern selection



We know more about patterns

- constraints used
- generality relationships



Overview

Unsupervised
Pattern set mining
Part II

Supervised
Pattern set mining
Part III

How to score pattern sets

How to find pattern sets

End of Part I