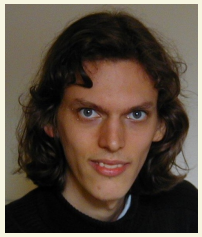
Integrating Constraint Programming And Itemset Mining



Siegfried Nijssen and *Tias Guns*

DTAI, K.U.Leuven, Belgium

ECML/PKDD 2010, Barcelona

Constraint-based Itemset Mining

Analysing purchases (e.g. items = books),
to find interesting patterns (sets of items)

Interestingness: <u>constraints</u>

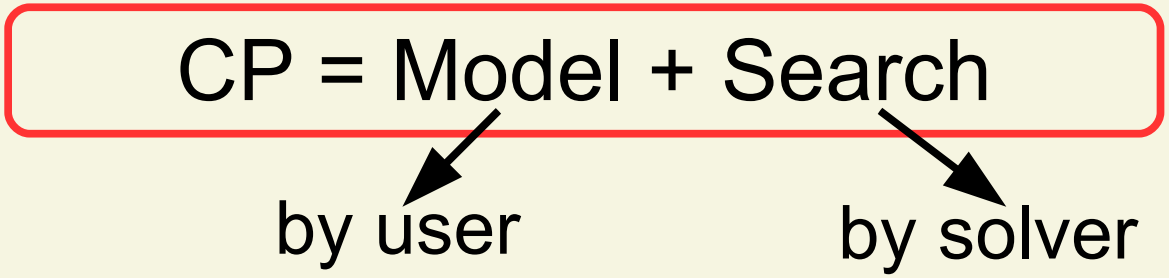
- Frequent sets, closed sets, correlated sets, ...
- Maximum size, minimum price, average cost, ...

Many constraints, many algorithms

Constraint Programming

General methodology for solving constraint satisfaction problems.

- Wide range of applications
- Each constraint is independent
- Freely combine constraints



Constraint Programming (CP) for IM

Model

- Variables: booleans

: $[i_1 \dots i_n]$  : $[t_1 \dots t_m]$

- Constraints: Many + combinations

Search

- <u>Out-of-the-box</u> Gecode CP solver (2008)

<u>Integrating</u> solver/miner principles:
practical AND theoretical benefits

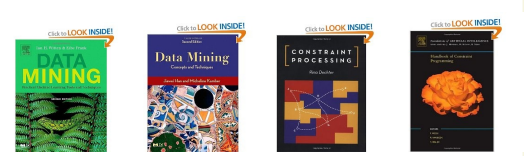
Overview

1. Motivation
2. **Principles of IM & CP**
3. Integrating IM & CP
4. Theoretical benefits
5. Practical benefits
6. Conclusions

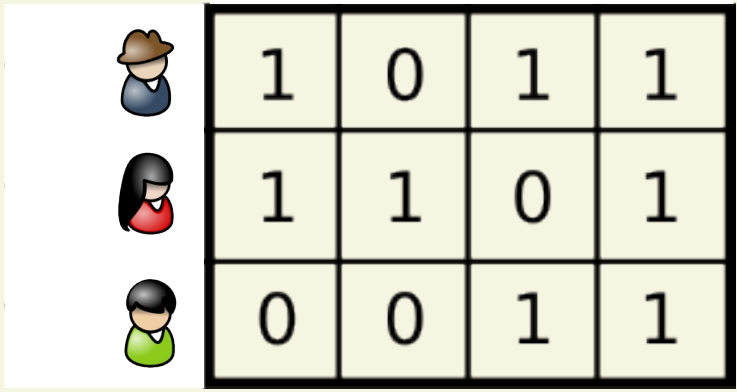
Principles of IM and CP

Will compare them on following implementations:

- **Eclat:** Simple and effective itemset miner
- **FIMCP:** our 2008 system, using the out-of-the-box Gecode CP solver



Itemset Mining principles



1	0	1	1
1	1	0	1
0	0	1	1

- Search strategy *Level-wise, BFS, DFS*
- Representation of data

1	0	1	1
1	1	0	1
0	0	1	1

1	0	1	1
1	1	0	1
0	0	1	1

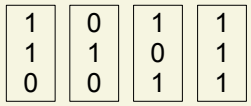
1 0 1 1
1 1 0 1
0 0 1 1

- Representation of sets

1 0 1 1
{1, 3, 4}

0 1 0 0
{2}

Comparison IM principles

	Eclat Miner	Gecode CP Solver
Search Strategy	DFS	DFS (binary)
Repres. of data	Shared, vertical 	In constraints (up to 4 copies)
Repres. of sets	Sparse or Dense	Sparse {1, 3, 4}

Constraint Programming principles

- Types of variables *Bool, Int, Set, ...*
- Supported constraints *Clause, Sum, AllDifferent, ...*
- Constraint activation

  On change of domain, on change of upper/lower bound, on change of specific value, ...

Comparing CP principles

	Eclat Miner	Gecode CP Solver
Types of vars.	Boolean vector (set)	Bool, Int, Set, ...
Constraints	Few, hard to combine	Many, easy to add/combine
Constraint activation	Fixed order (in algorithm)	On domain change

Overview

1. Motivation
2. Principles of IM & CP
3. **Integrating IM & CP**
4. Theoretical benefits
5. Practical benefits
6. Conclusions

Integrating IM & CP

We created a new CP solver called DMCP,

- using principles of both IM and CP
- implementing constraints for itemset mining

Integration 1/3

	Eclat Miner	Gecode CP Solver	Our **DMCP** CP Solver
Search strategy	DFS	DFS (binary)	DFS (binary)
Repres. of data	Shared, vertical	In constraints (up to 4 copies)	Shared matrix (default: vertical)

- Data shared (read-only) by constraints
- Horizontal, positive and negative *views* available

Integration 2/3

	Eclat Miner	Gecode CP Solver	Our **DMCP** CP Solver
Repres. of sets	Sparse or Dense	Sparse	Sparse or Dense
Types of vars.	Boolean vector (set)	Bool, Int, Set, ...	Boolean vector (set)

- Represented by lower and upper bound:

0, 0/1, 0/1, 1

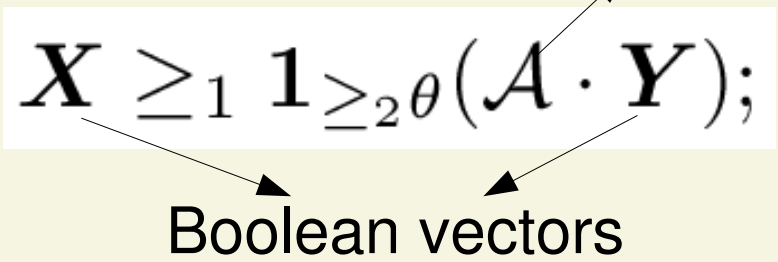
Min: {0, 0, 0, 1}

Max: {0, 1, 1, 1}

Integration 3/3

	Eclat Miner	Gecode CP Solver	Our **DMCP** CP Solver
Constraints	Few, hard to combine	Many, easy to add/combine	Some, easy add/combine
Constraint activ.	Strict order (in algorithm)	On domain change	Change of lower/upper bound

- General matrix constraint:



Overview

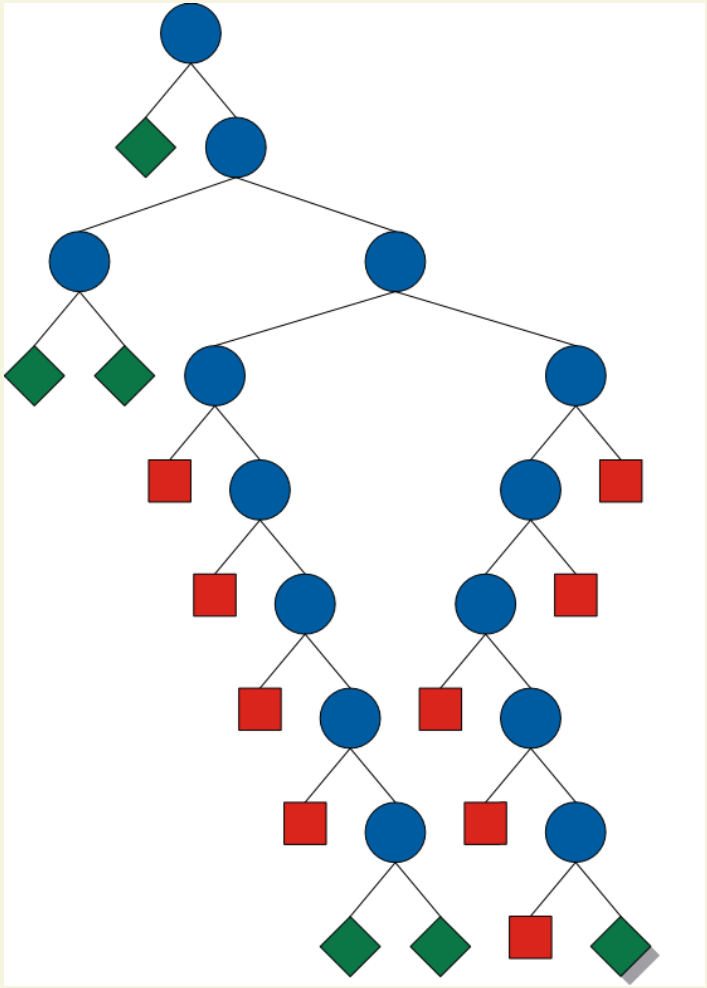
1. Motivation
2. Principles of IM & CP
3. Integrating IM & CP
4. **Theoretical benefits**
5. Practical benefits
6. Conclusions

Theoretical benefits

Complexity analysis:

Polynomial constraints
+ non-failing siblings
= polynomial delay

Search tree:



Theoretical benefits

Can prove polynomial delay of

- Frequent itemset mining
- Closed itemset mining
- A related graph mining problem*
- more ?

* M. Boley et al. Efficient closed pattern mining in strongly accessible set systems. PKKD 2007

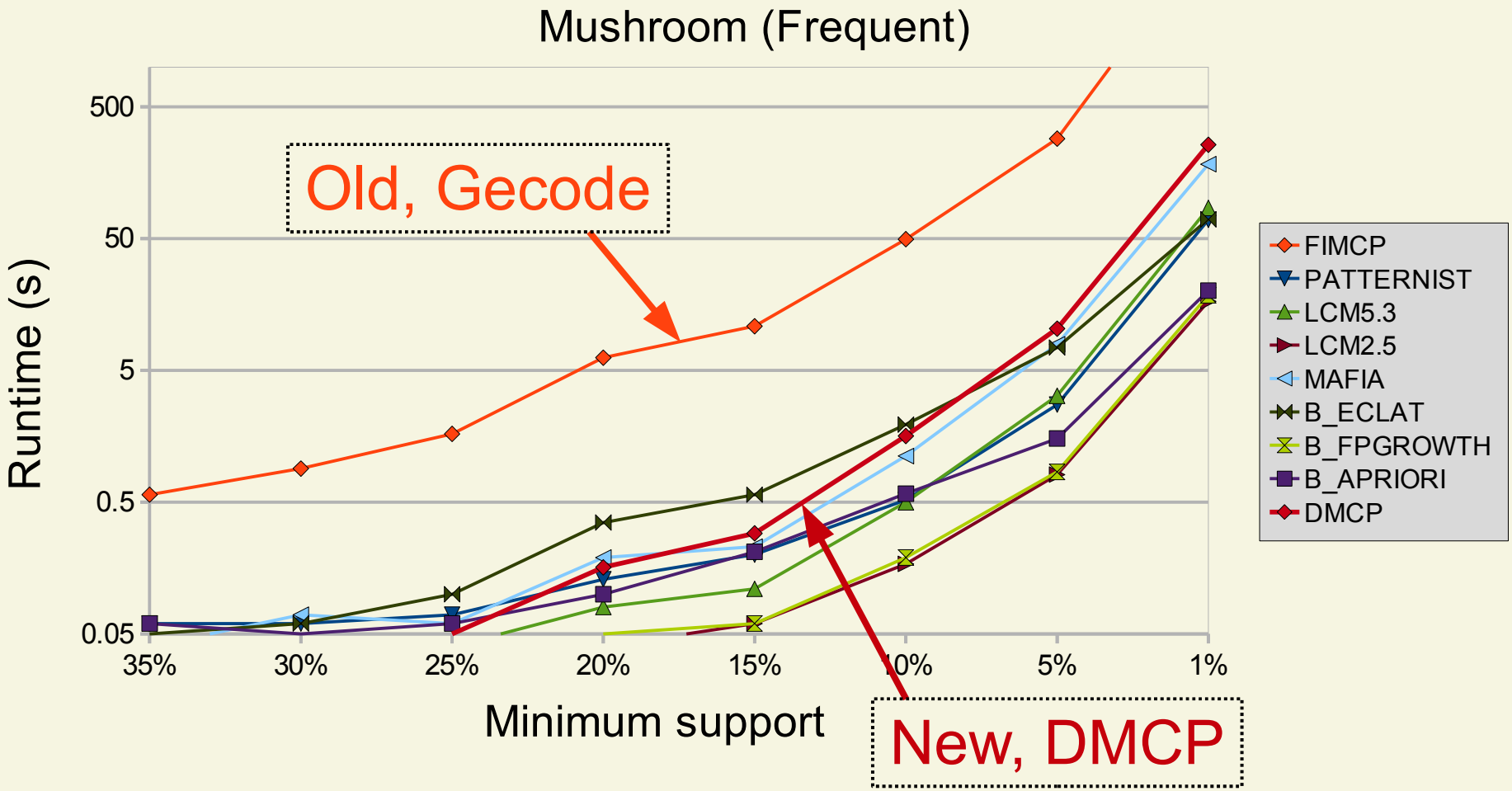
Overview

1. Motivation
2. Principles of IM & CP
3. Integrating IM & CP
4. Theoretical benefits
5. **Practical benefits**
6. Conclusions

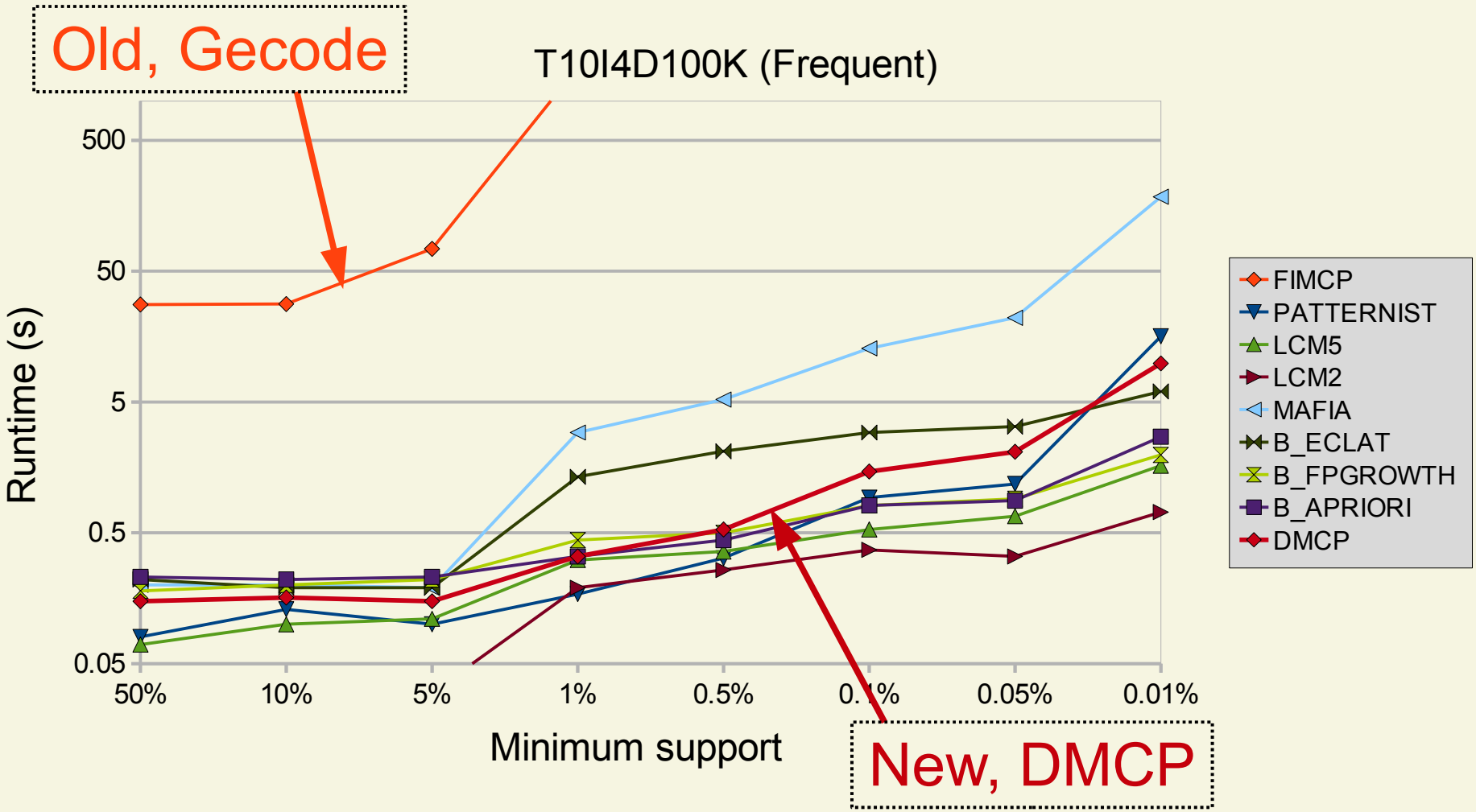
Mining systems

- DMCP: our new integrated CP/IM system.
- FIMCP: our Gecode based system
- PATTERNIST: constraint-based itemset miner
- LCM: 'winner' of the FIMI competition
- ECLAT, FPGrowth and APRIORI,
  as implemented by C. Borgelt

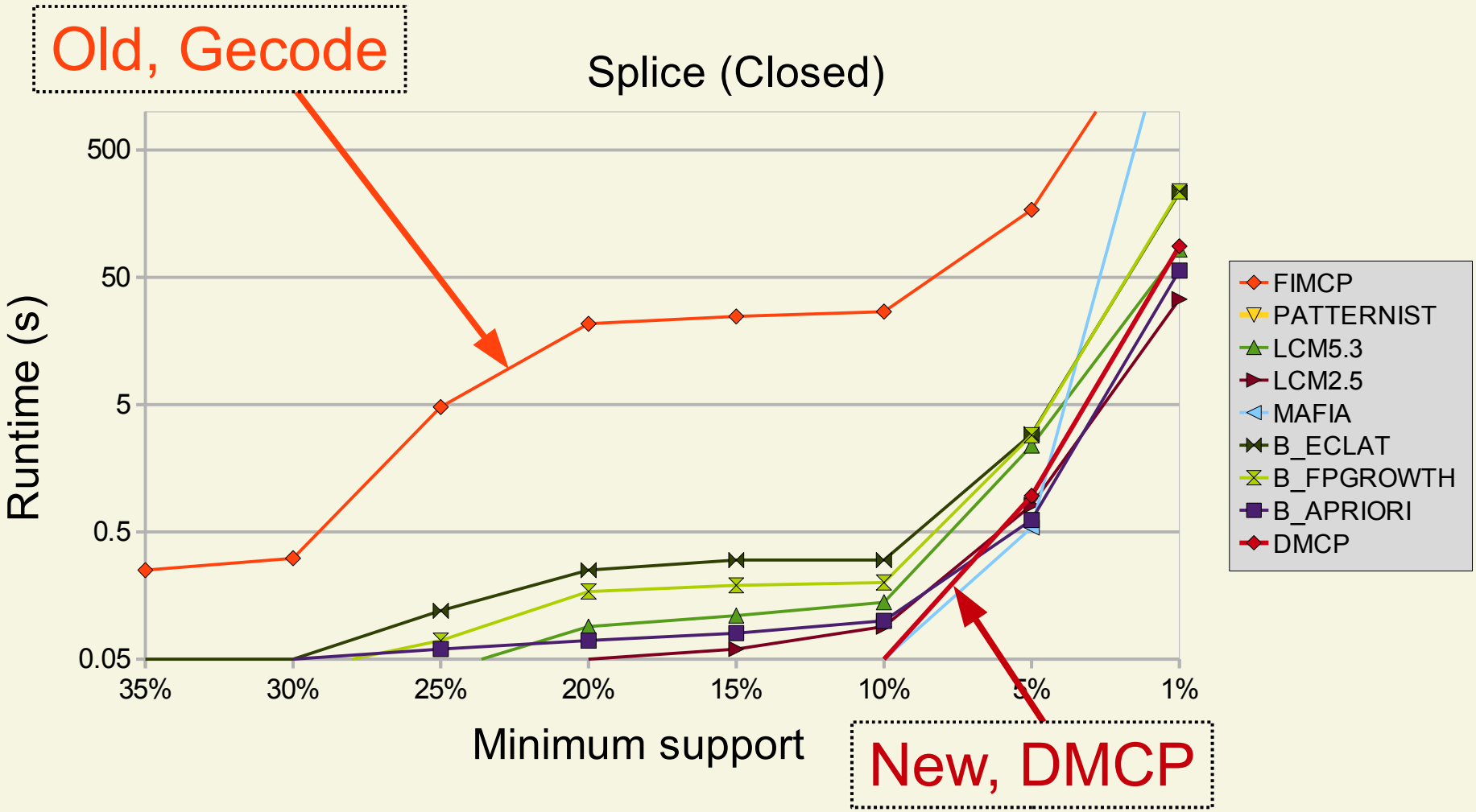
Frequent Itemset Mining



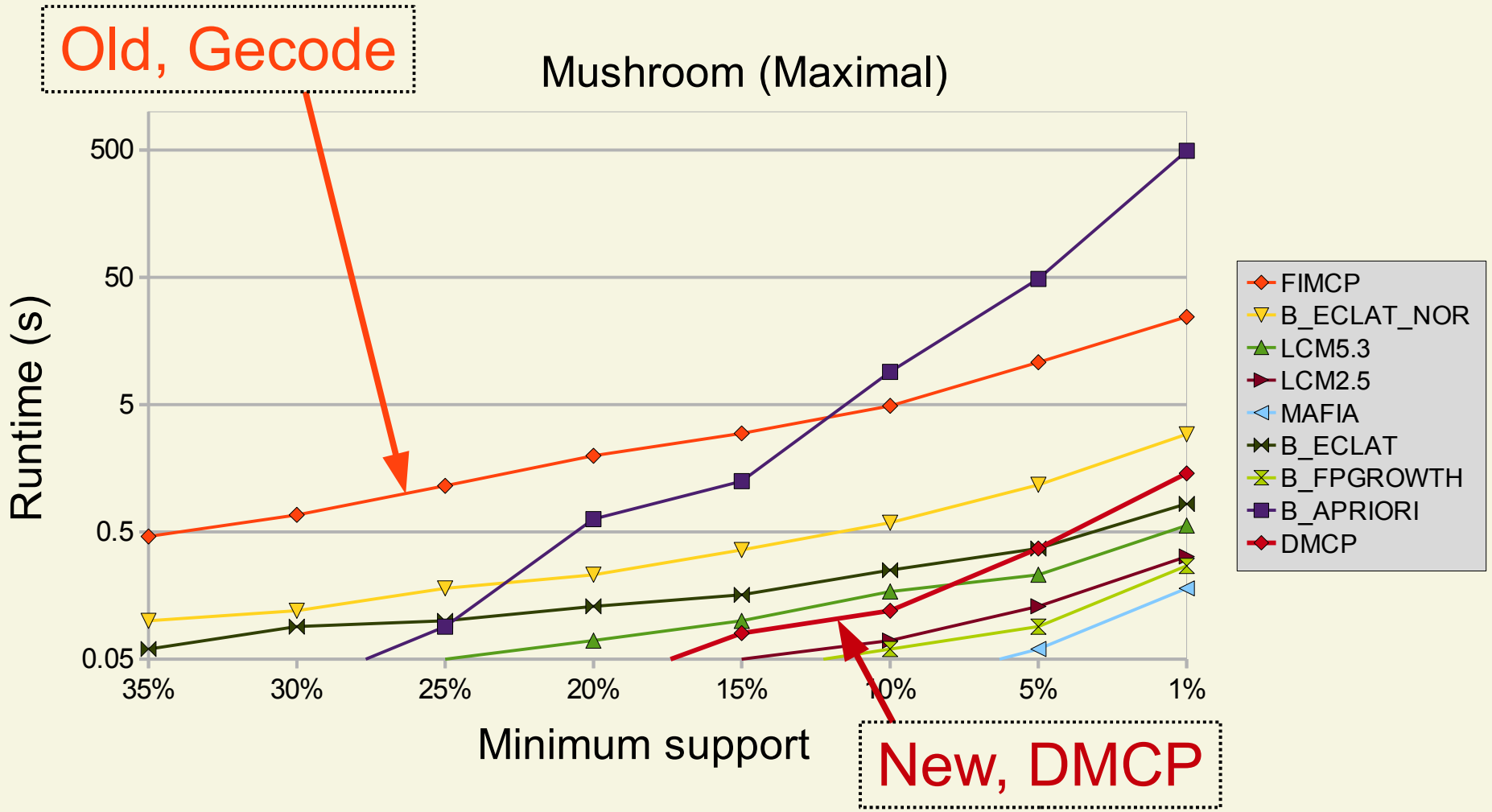
Frequent Itemset Mining, scaling



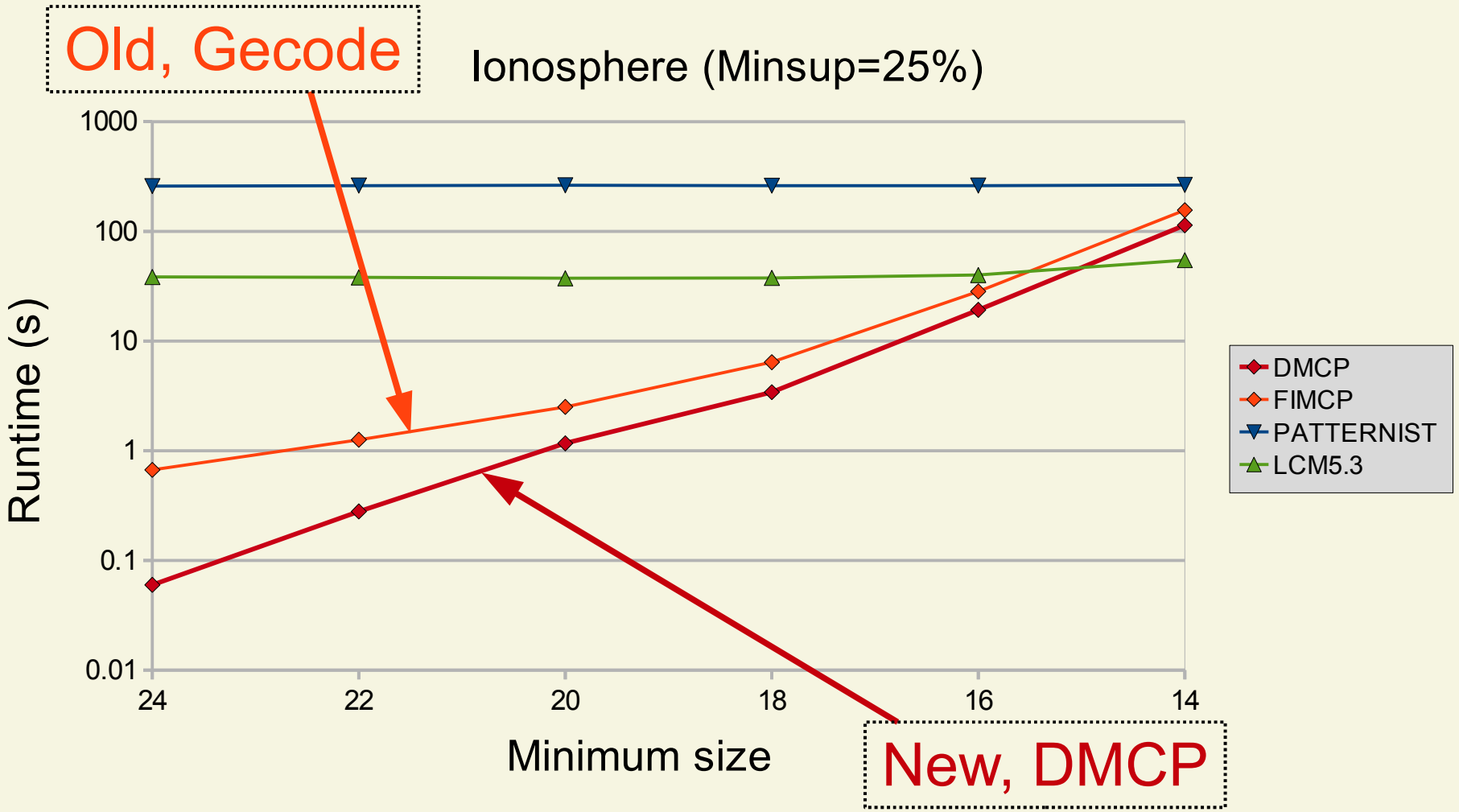
Closed Itemset Mining



Maximal Itemset Mining



Minimum size (monotone)



Overview

1. Motivation
2. Principles of IM & CP
3. Integrating IM & CP
4. Theoretical benefits
5. Practical benefits
6. **Conclusions**

Conclusion

Advantages of CP modelling:

Model

- Easily add new constraints
- Freely combine constraints

Search

Advantage of IM/CP solver integration:

- Theoretical: polynomial delay analysis
- Practical: remove efficiency/scalability gap

Open questions

- Integrate IM principles in existing CP solver ?
- More efficient solving of typical CP problems ?
- Other mining strategies (e.g. FPgrowth) in CP ?
- Adding CP principles to other pattern domains (e.g. sequences, trees, graphs) ?

Thank you for listening

Questions?

http://dtai.cs.kuleuven.be/CP4IM