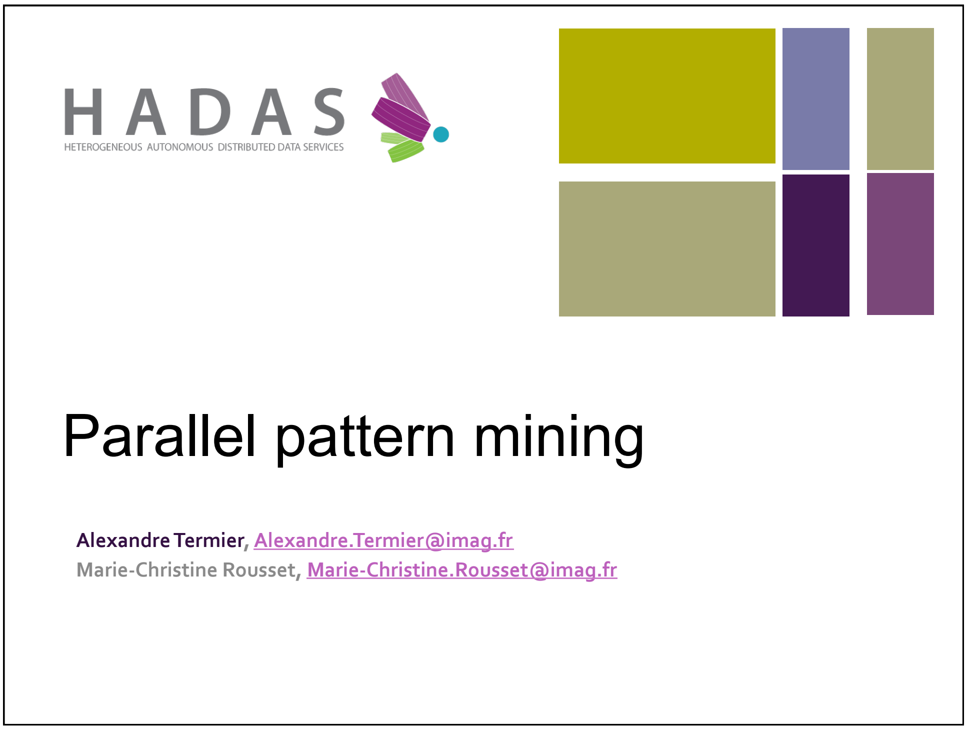
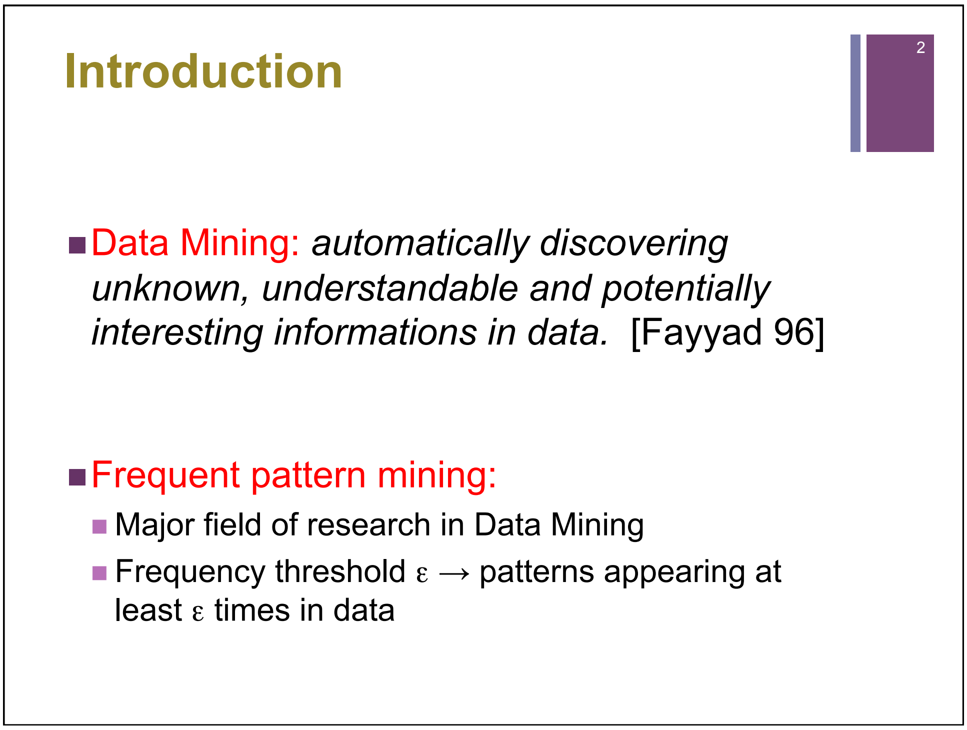


HADAS
HETEROGENEOUS AUTONOMOUS DISTRIBUTED DATA SERVICES

Parallel pattern mining

Alexandre Termier, Alexandre.Termier@imag.fr
Marie-Christine Rousset, Marie-Christine.Rousset@imag.fr

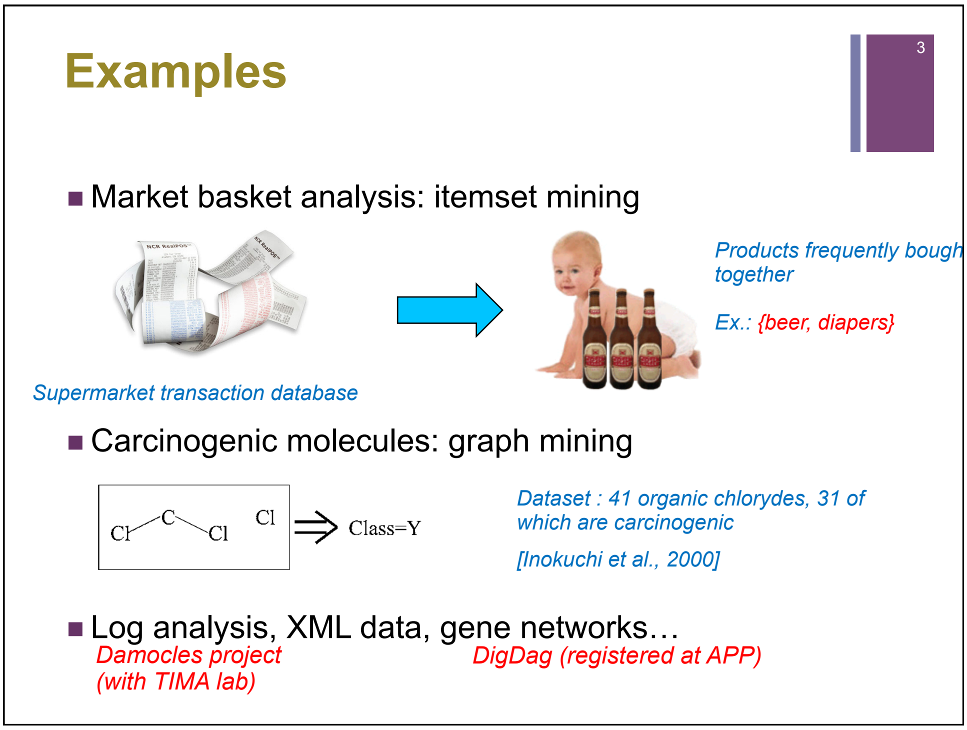


Introduction

2

- Data Mining: *automatically discovering unknown, understandable and potentially interesting informations in data.* [Fayyad 96]

- Frequent pattern mining:
  - Major field of research in Data Mining
  - Frequency threshold ε → patterns appearing at least ε times in data



Examples

3

- Market basket analysis: itemset mining

Cl
C
Cl
Cl
Class=Y

Supermarket transaction database

Products frequently bough together

Ex.: {beer, diapers}

- Carcinogenic molecules: graph mining

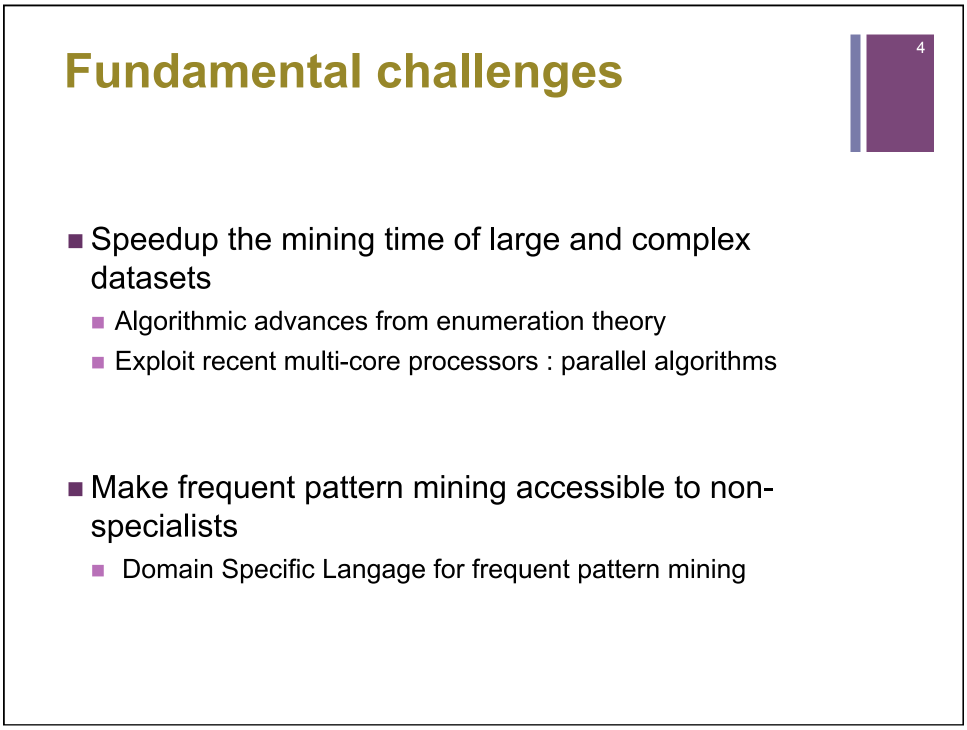
Dataset : 41 organic chlorydes, 31 of which are carcinogenic

[Inokuchi et al., 2000]

- Log analysis, XML data, gene networks…

Damocles project (with TIMA lab)

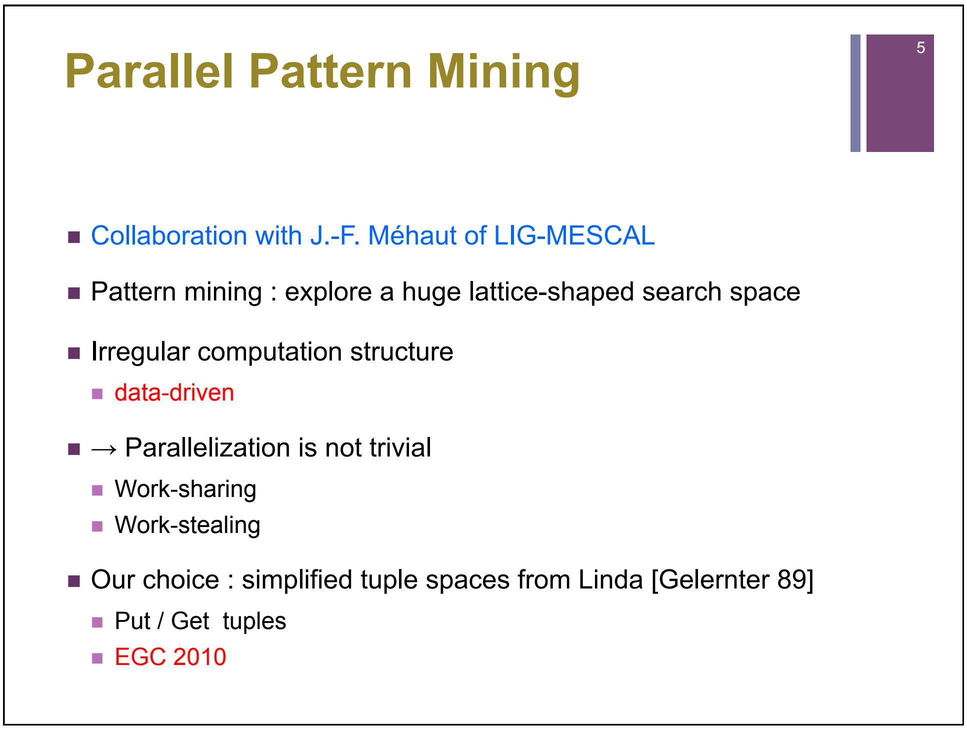
DigDag (registered at APP)



Fundamental challenges

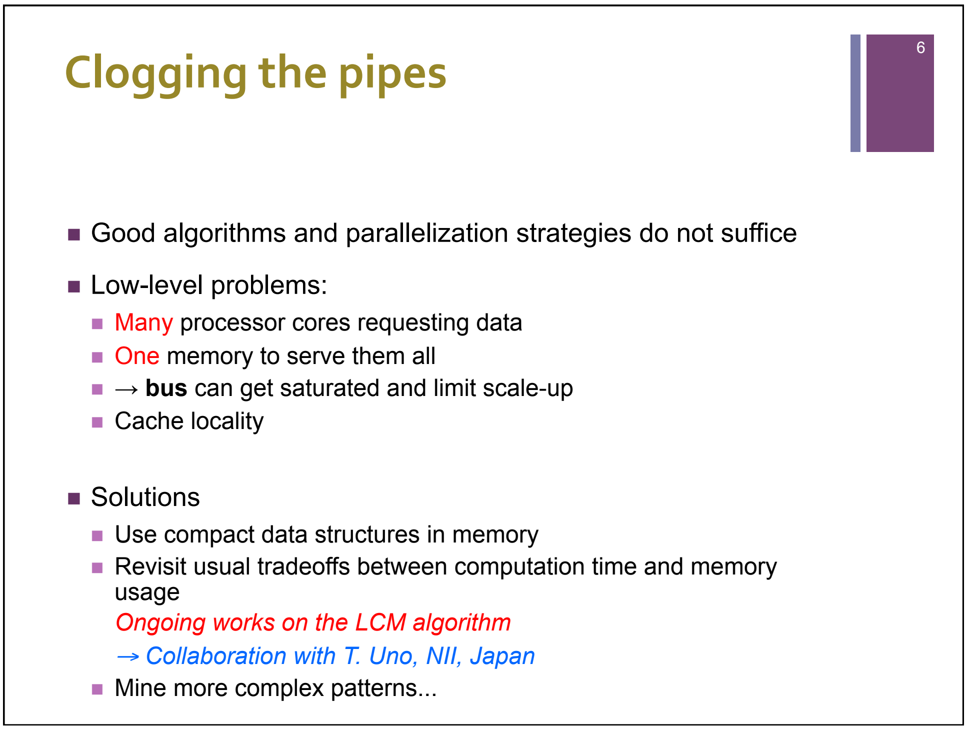
4

- Speedup the mining time of large and complex datasets
  - Algorithmic advances from enumeration theory
  - Exploit recent multi-core processors : parallel algorithms
- Make frequent pattern mining accessible to non-specialists
  - Domain Specific Langage for frequent pattern mining



Parallel Pattern Mining

- Collaboration with J.-F. Méhaut of LIG-MESCAL
- Pattern mining : explore a huge lattice-shaped search space
- Irregular computation structure
  - data-driven
- → Parallelization is not trivial
  - Work-sharing
  - Work-stealing
- Our choice : simplified tuple spaces from Linda [Gelernter 89]
  - Put / Get tuples
  - EGC 2010

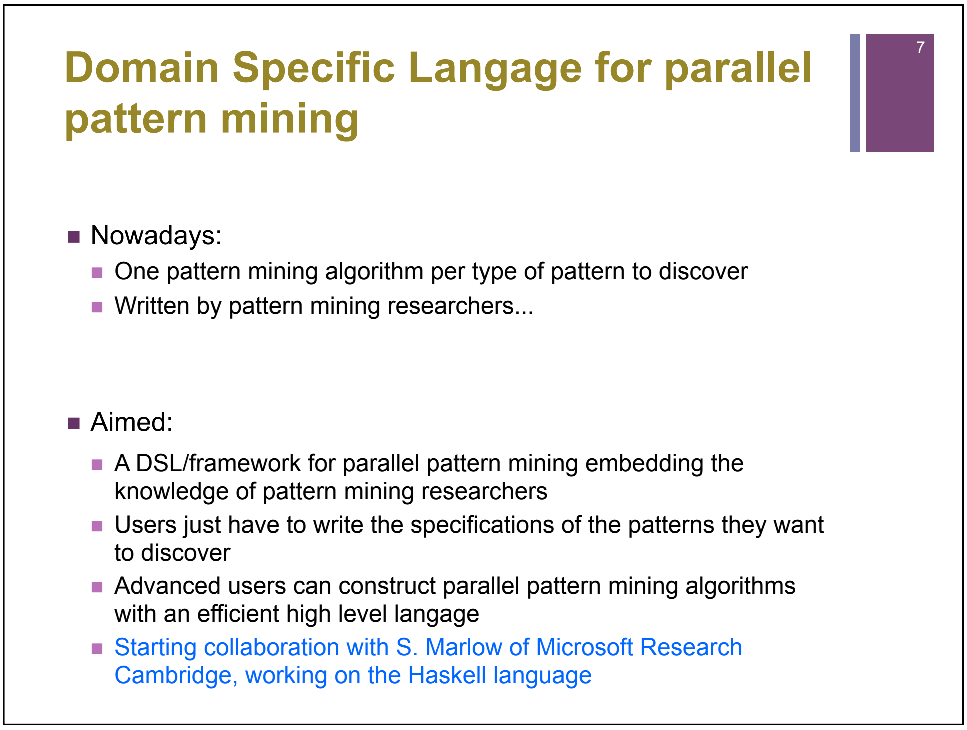


Clogging the pipes

- Good algorithms and parallelization strategies do not suffice
- Low-level problems:
  - Many processor cores requesting data
  - One memory to serve them all
  - → **bus** can get saturated and limit scale-up
  - Cache locality
- Solutions
  - Use compact data structures in memory
  - Revisit usual tradeoffs between computation time and memory usage

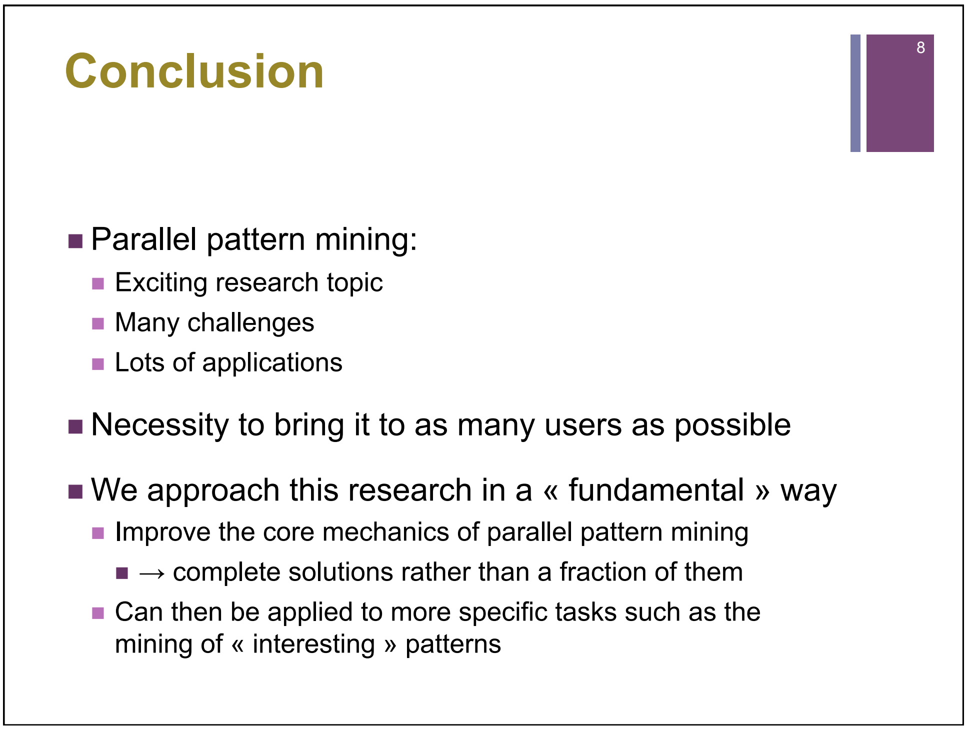
    Ongoing works on the LCM algorithm

    → Collaboration with T. Uno, NII, Japan
  - Mine more complex patterns...



Domain Specific Langage for parallel pattern mining

- Nowadays:
  - One pattern mining algorithm per type of pattern to discover
  - Written by pattern mining researchers...
- Aimed:
  - A DSL/framework for parallel pattern mining embedding the knowledge of pattern mining researchers
  - Users just have to write the specifications of the patterns they want to discover
  - Advanced users can construct parallel pattern mining algorithms with an efficient high level langage
  - Starting collaboration with S. Marlow of Microsoft Research Cambridge, working on the Haskell language



Conclusion

- Parallel pattern mining:
  - Exciting research topic
  - Many challenges
  - Lots of applications
- Necessity to bring it to as many users as possible
- We approach this research in a « fundamental » way
  - Improve the core mechanics of parallel pattern mining
    - → complete solutions rather than a fraction of them
  - Can then be applied to more specific tasks such as the mining of « interesting » patterns