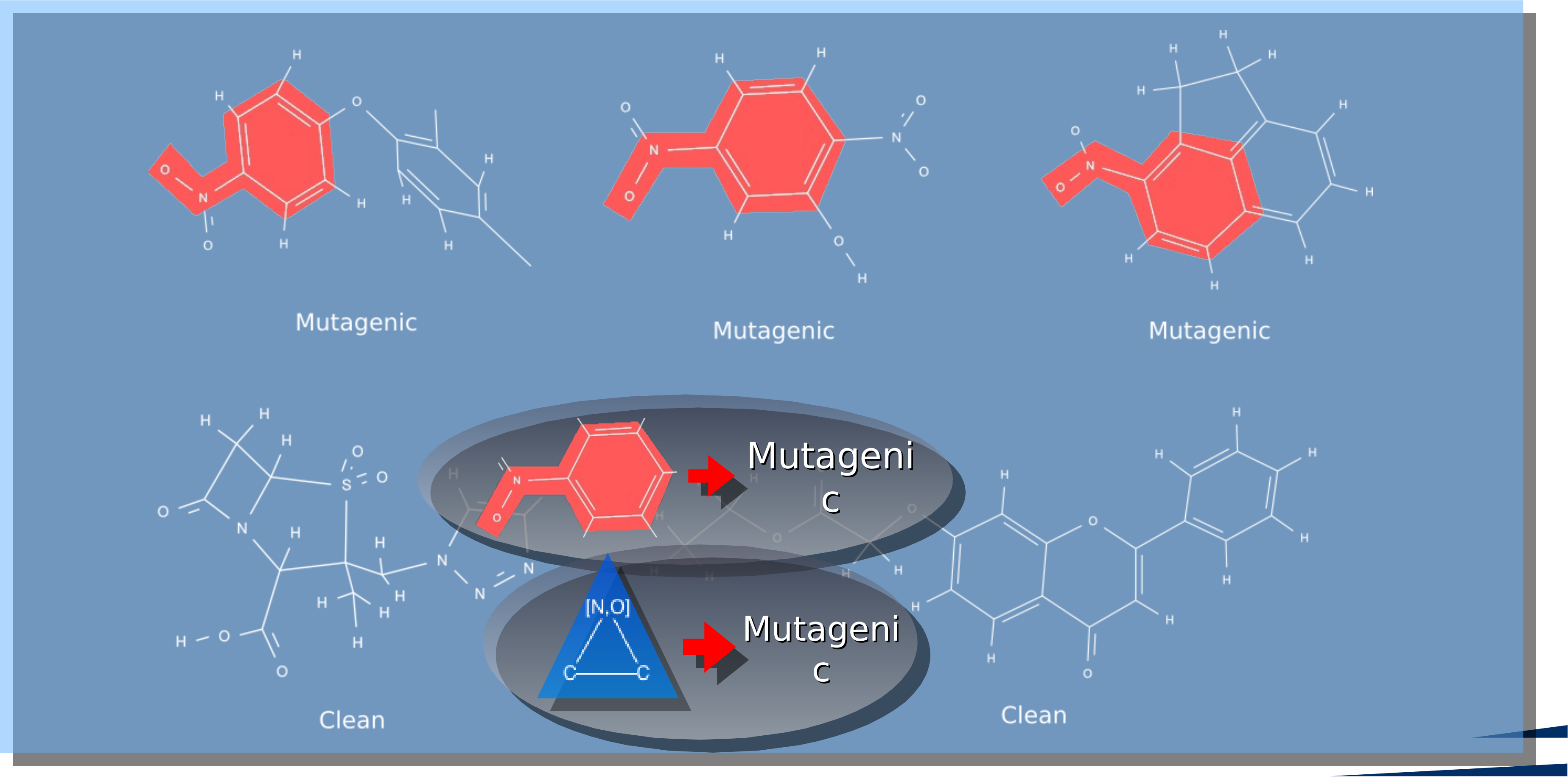
Siegfried Nijssen

- Master in computer science (Leiden, 2000)
- PhD in computer science (Leiden, 2006)
- Post doc in Leuven (KU Leuven)
- Docent (Leiden)

- Machine learning
- Data mining
- Artificial intelligence

Graph Mining



Graph Mining

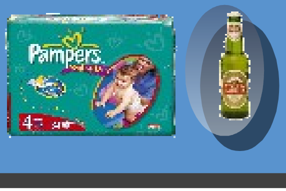
- Can we efficiently modify a graph mining system such that it supports *multiple* node labels?

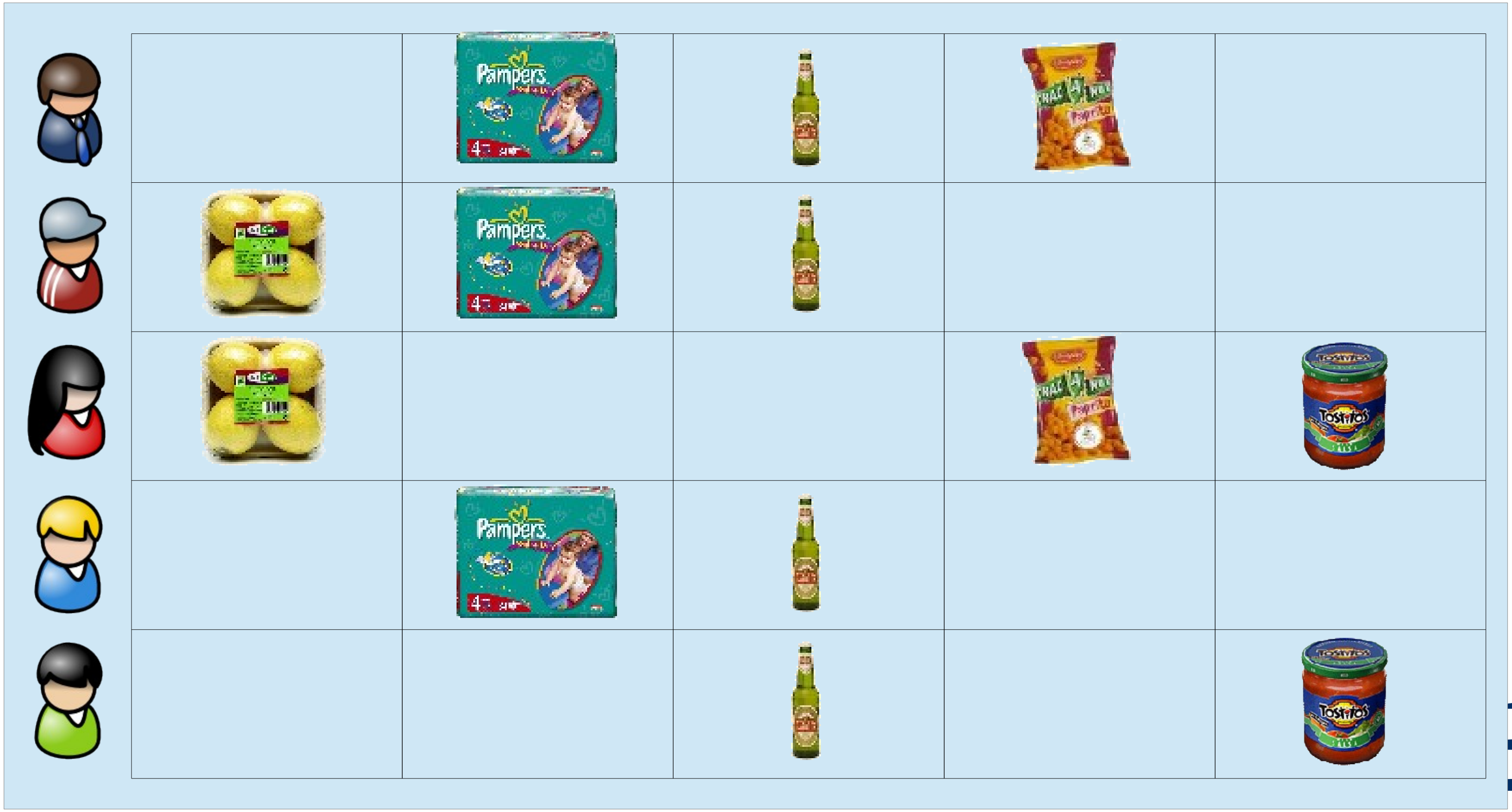
  O, hydrogen donor
  O, hydrogen acceptor
  N, hydrogen donor
  N, hydrogen acceptor
  …

- Requirements:
  - An interest in efficient programming in C++
  - An interest in graph theory

Declarative Data Mining

- Market basket data

support()=3



Declarative Data Mining

- Apriori
- FP-Growth
- Eclat
- SD-Apriori
- DDPMine
- Gaston
- gSpan
- FFSM
- TreeMiner
- LCM
- MaxMiner
- DualMiner
- Molfea
- CorrMine
- EclatV
- Mafia
- kDCI
- ARMOR
- AIM
- COFI-tree
- DCI closed
- WinePI
- MinePI
- … … …

Declarative Data Mining

- “An SQL for data mining” using “constraint programming”

```
int: NrI;
int: NrT;
int: Freq;

array [1 . . NrT] of set of 1 .. NrI : TDB;

var set of 1..NrI: Items ;

constraint card ( cover ( Items , TDB ) ) >= Freq ;

solve satisfy;
```

Declarative Data Mining

- Can an effective declarative data mining system be built in Python, based on “Numberjack” and “sckit-learn”?

- Requirements:
  - An interest in programming in Python
  - An interest in algorithms
  - An interest in artificial intelligence
  - An interest in declarative programming

Mining a Conference

- European Conference on Machine Learning and Principles of Knowledge Discovery in Databases (ECMLPKDD)
- 450 conference submissions, with 1350 reviews
- 150 journal submissions, with 450 reviews
- Different types of data:
  - Text: reviews, abstracts
  - Attribute-value data: topical categories, nationalities, accepted or not
  - Network data: co-authorship graphs, citation graphs

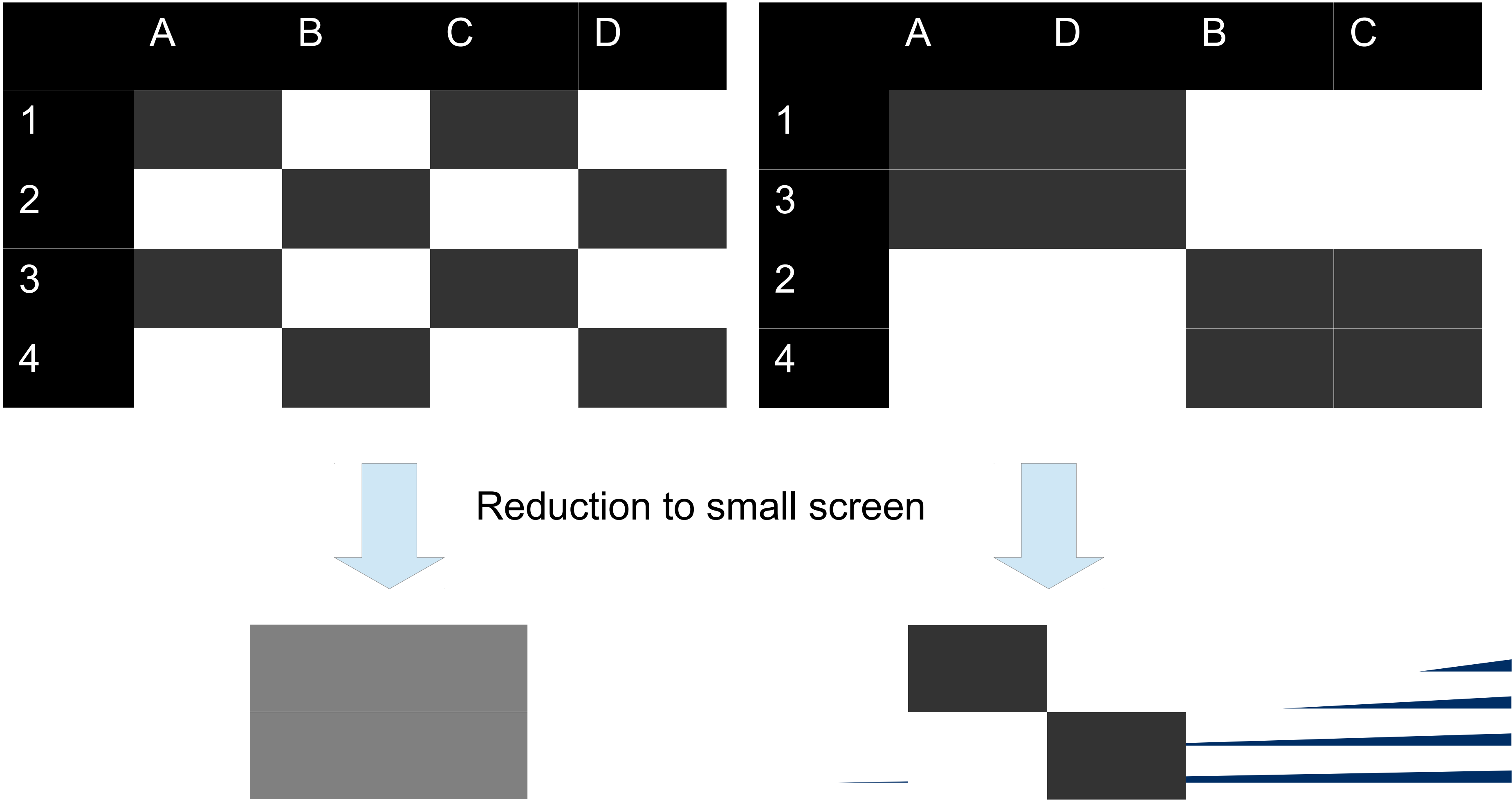
Mining a Conference

- **Goal:** to answer questions on this data
  - Can we predict whether a paper is accepted?
  - Can we predict the length of a review?
  - Can we predict the verdict of a review based on its text?
  - Are there large differences between subfields of machine learning and data mining?
  - Can we predict whether a paper should receive a summary reject?
  - Can we predict how long it will take to review a paper?
- … while also using network data

Mining a Conference

- **Required:**
  - Interest in data mining, machine learning and a little bit of statistics
  - Interest to use programs such as Weka
  - Interest to implement in SQL, Python
- **Desirable:**
  - Interest in scraping web pages
  - Interest in network mining
  - Interest in R

Patterns in Data Visualization



	A	B	C	D
1	■		■	
2		■		■
3	■		■	
4		■		■

	A	D	B	C
1	■	■		
3	■	■		
2			■	■
4			■	■

Reduction to small screen

Patterns in Data Visualization

- What do the data visualizations look like for different types of patterns?

- Requirements:
  - An interest in making visualizations in Python, C++, …
  - An interest in running existing data mining programs in C++
  - An interest in algorithms