Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

Characteristic Relational Patterns

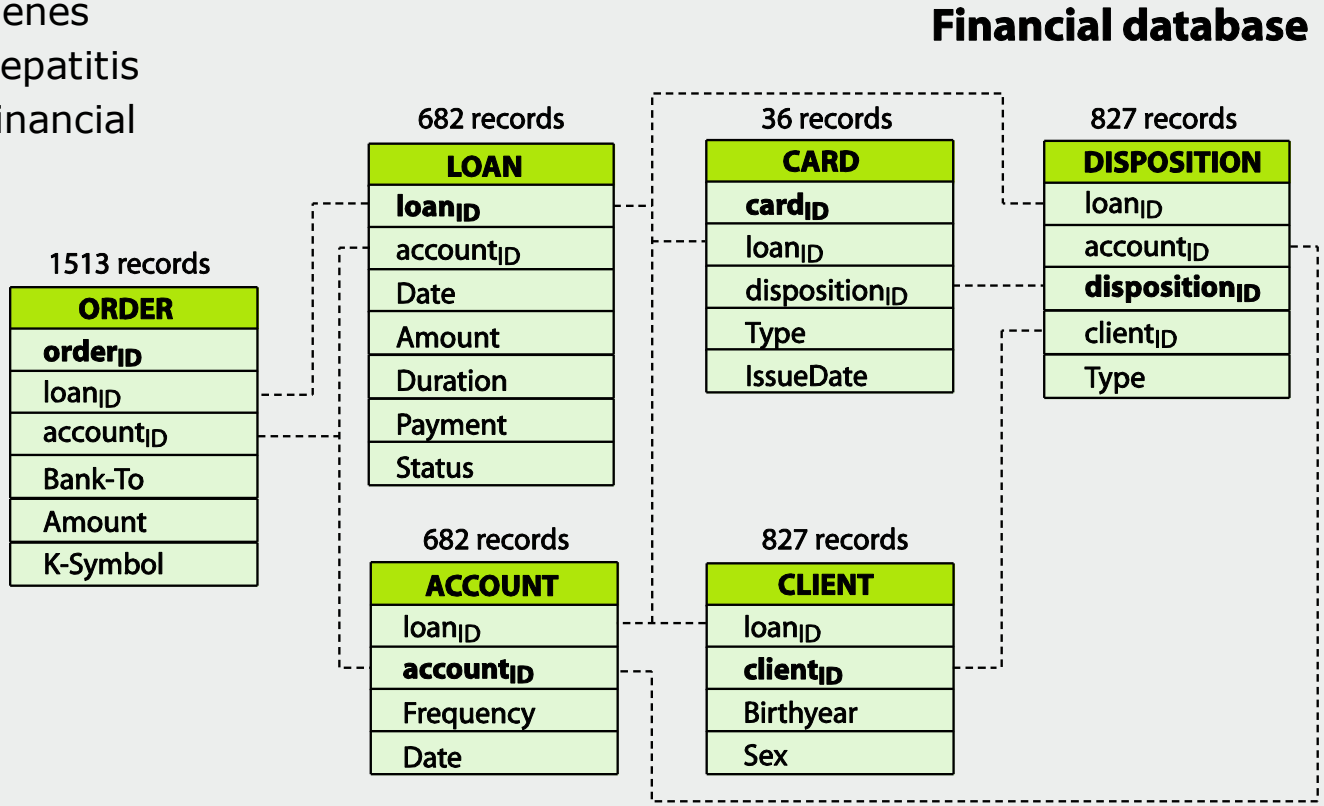
Arne Koopman & Arno Siebes
Algorithmic Data Analysis, Universiteit Utrecht

Characterising the Database

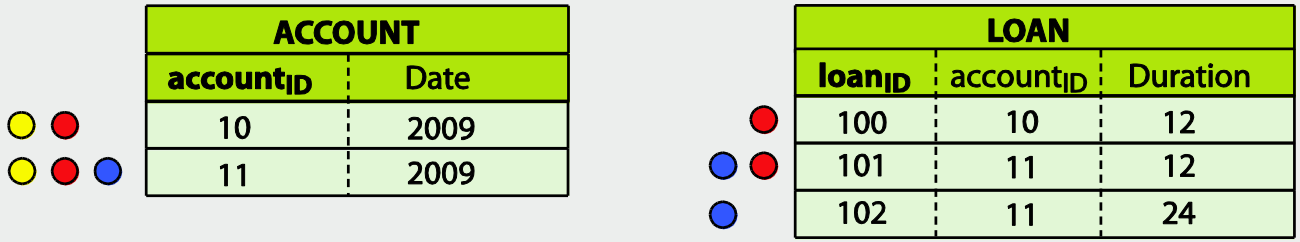
- Relational database models
  - Local models: frequent pattern mining
  - Global models: probabilistic relational model
- Characterising the database
  - Combine patterns to form a global model
- Experiments

Relational Databases

- KDD cup relational databases
  - Genes
  - Hepatitis
  - Financial



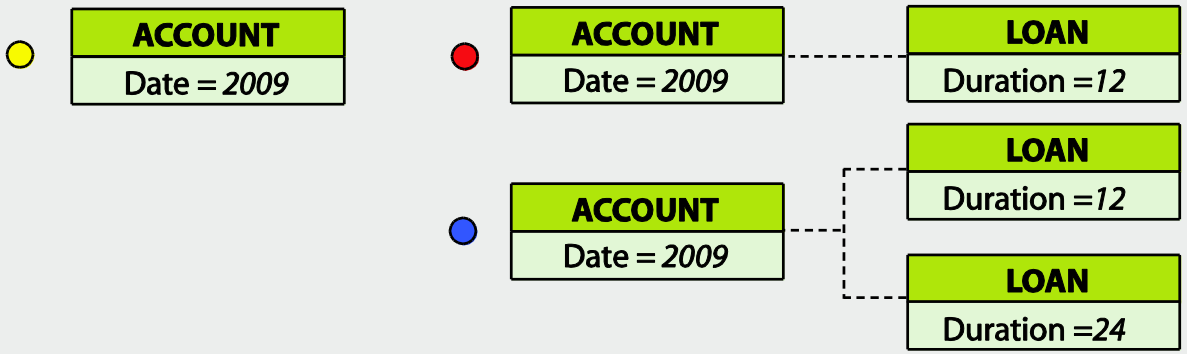
Local Models



ACCOUNT	
$account_{ID}$	Date
10	2009
11	2009

LOAN		
$loan_{ID}$	$account_{ID}$	Duration
100	10	12
101	11	12
102	11	24

Patterns



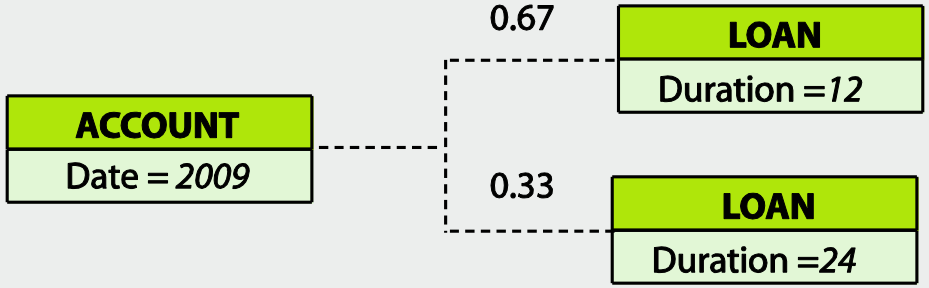
- Frequent pattern mining: too many patterns!

Global Model

ACCOUNT	
$account_{ID}$	Date
10	2009
11	2009

LOAN		
$loan_{ID}$	$account_{ID}$	Duration
100	10	12
101	11	12
102	11	24

Probabilistic Relational Model



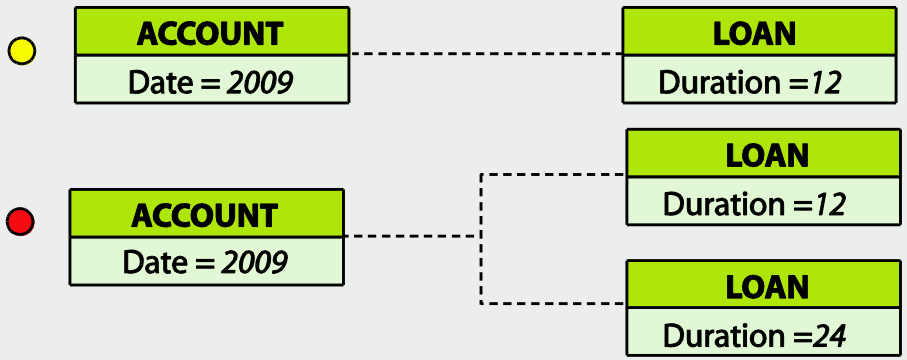
- Probabilistic Relational Models: local co-occurrence information lost

Combined Model

ACCOUNT	
account$_{ID}$	Date
10	2009
11	2009

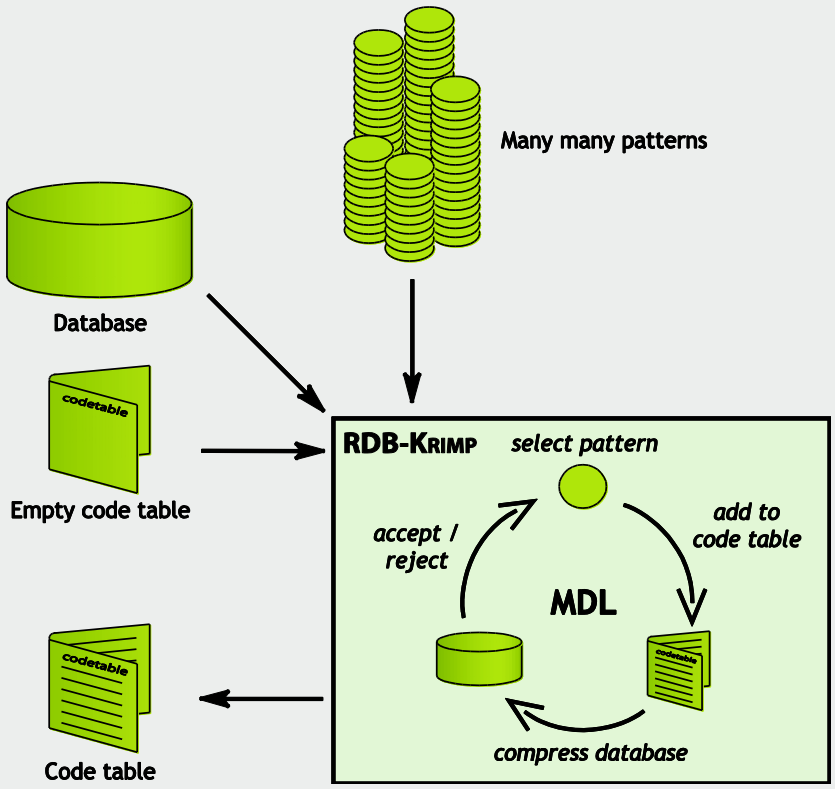
LOAN		
loan$_{ID}$	account$_{ID}$	Duration
100	10	12
101	11	12
102	11	24

Code Table



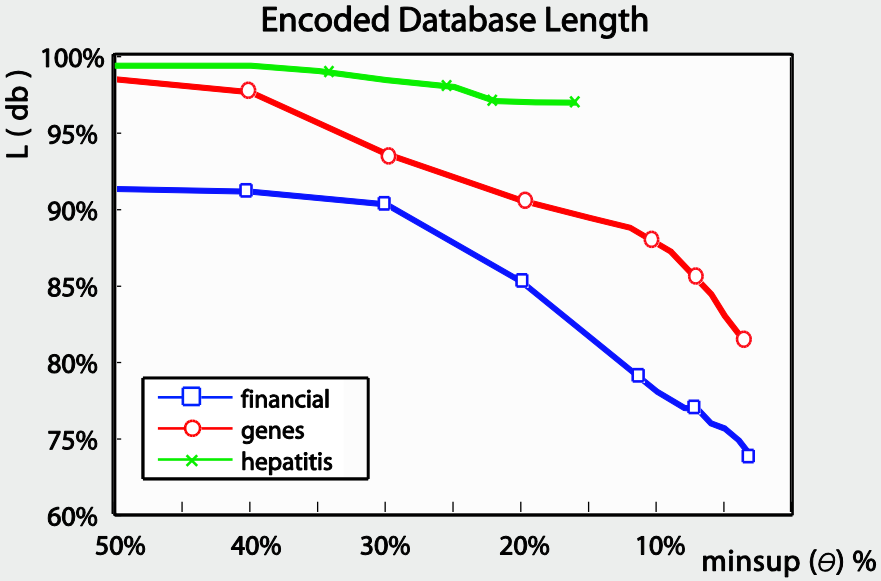
- Relational Code Table: compact and lossless description of the complete database

RDB-KRIMP



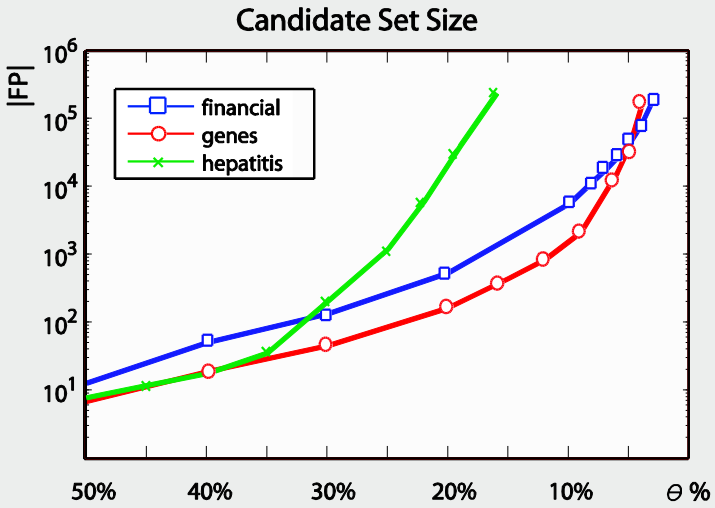
- RDB-KRIMP selects patterns that describe the database well
- Candidates are frequent relational patterns
- Describing patterns are placed in a code table
  - shortness code length proportional to usage

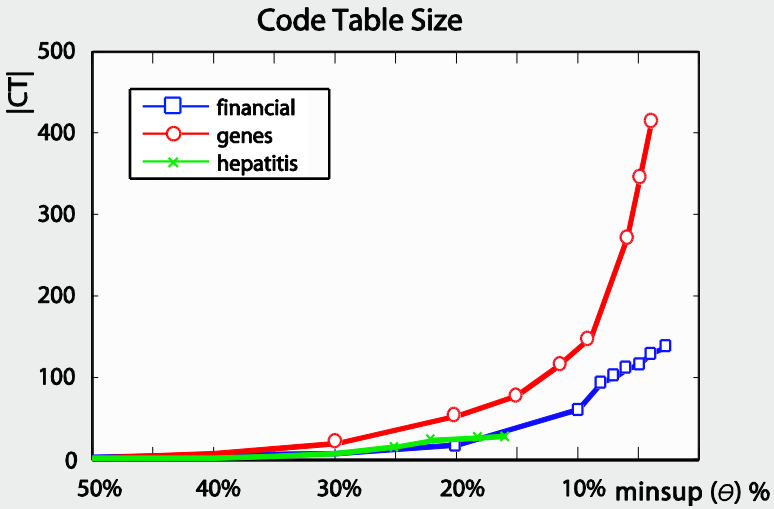
Compression - 1



- Code tables encode the database
  - Increasingly better encoded sizes for lower minimum supports

Compression - 2





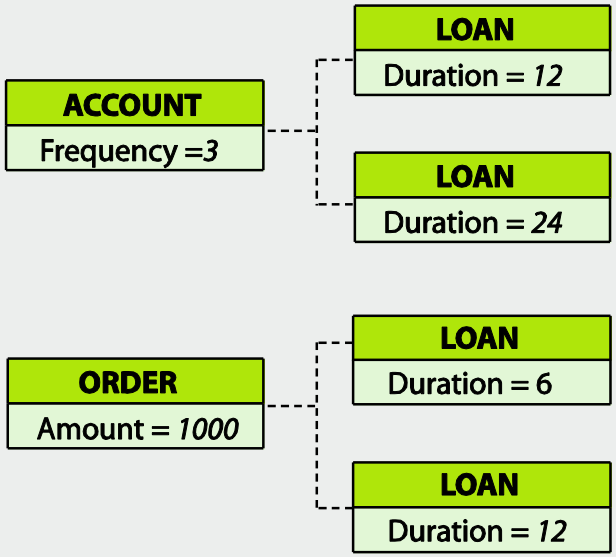
- Code tables encode the database
  - Increasingly better encoded sizes for lower minimum supports
  - Candidate set grow exponentially
  - Code tables stay compact

Pattern Languages

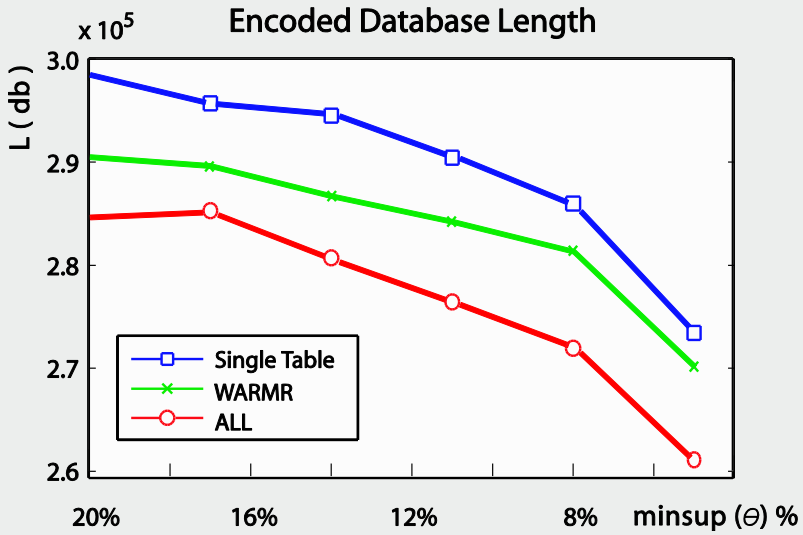
local table

WARMR

ALL: FARMER w/o target



Pattern Complexity



- More complex patterns lead to better descriptions.
- Thus, they encode MDL-relevant structure

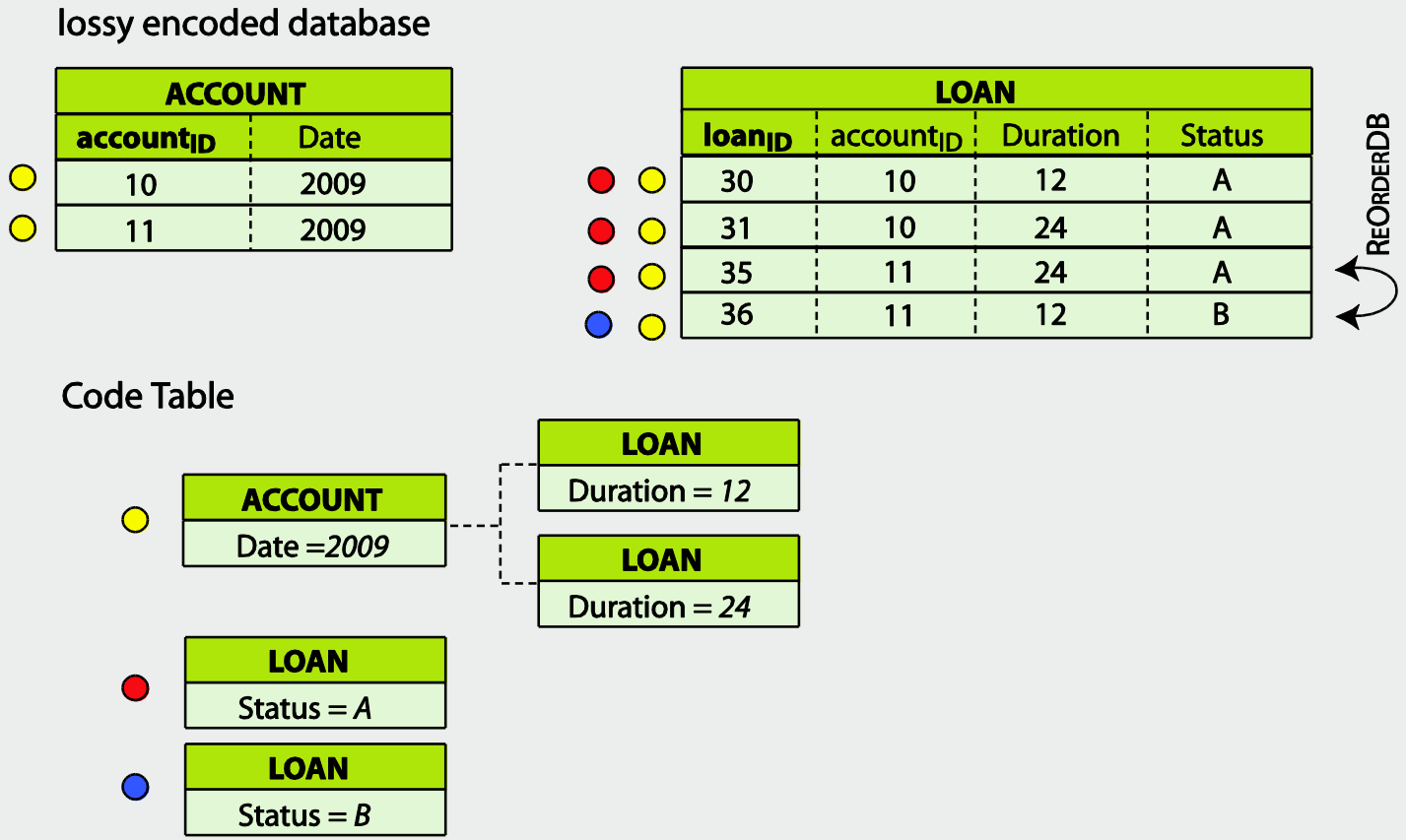
Database	single table		WARMR		all	
	L	CT	L	CT	L	CT
Financial	91%	29	76%	130	76%	117
Genes	87%	72	86%	191	83%	342
Hepatitis	99%	5	98%	13	97%	26

Conclusions

- Code tables describe the database while preserving local information
- Code tables stay compact
  - Stay compact for low minimum support values
  - Reductions up to 4 orders of magnitude
- Richer patterns lead to better models
  - Smaller encodings
  - Better descriptions without target tables

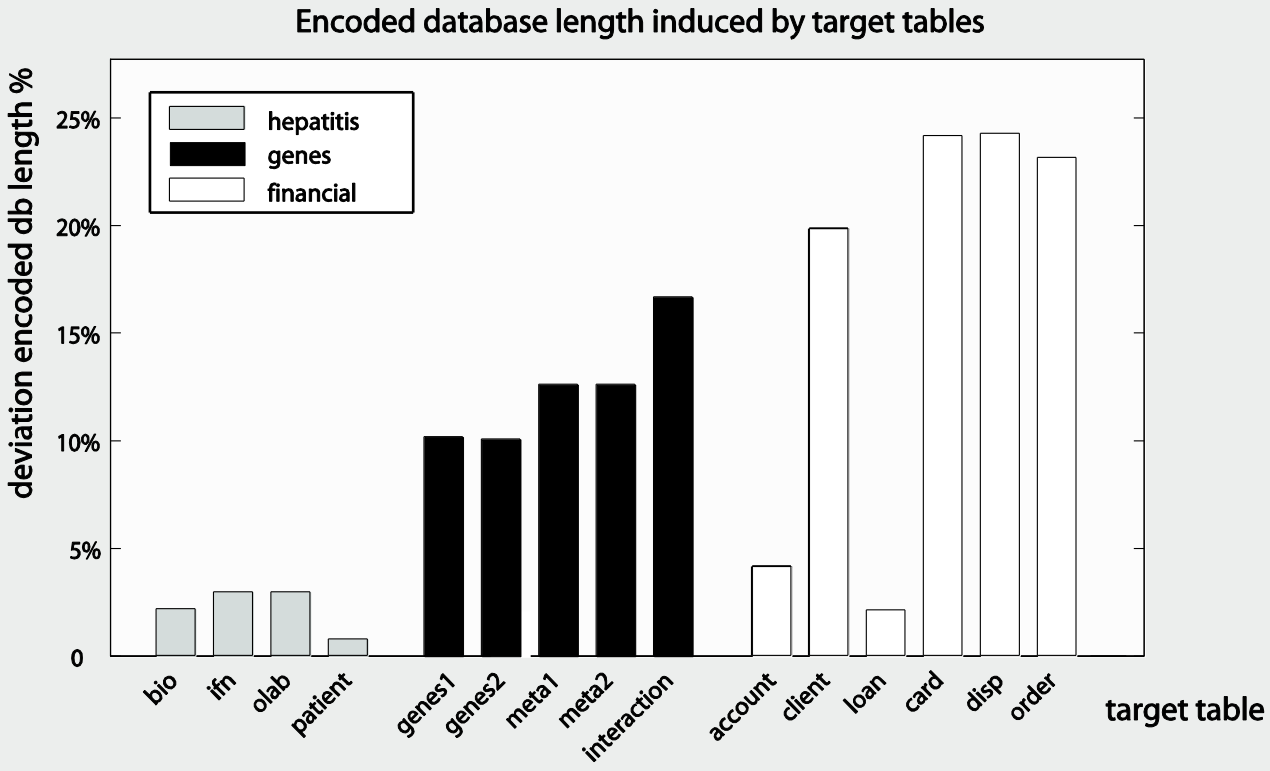
Questions?

Database Encoding



- Reordering allows for a lossless encoding

Target Tables

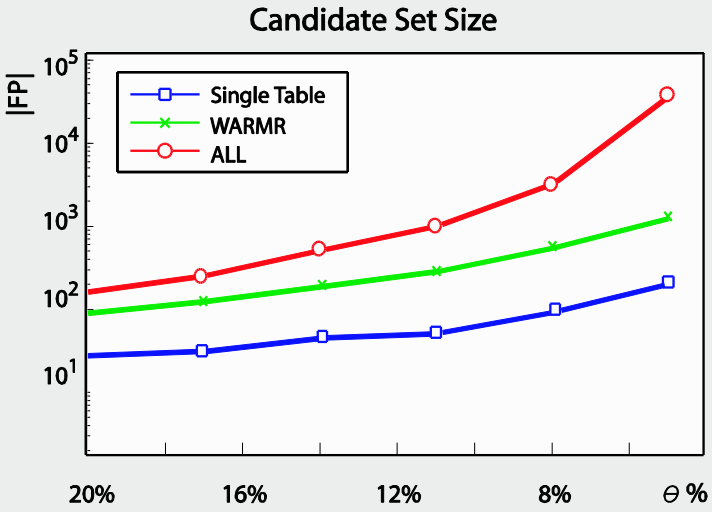


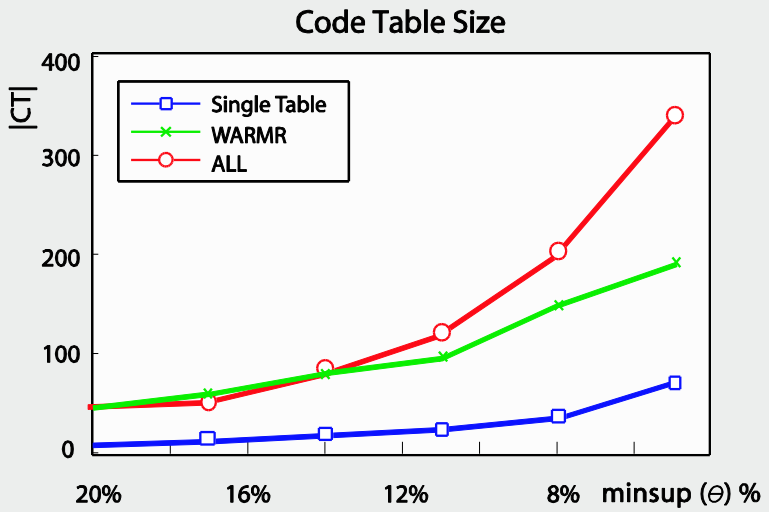
- We obtain better encodings without a target table

Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

Pattern Complexity





- In all cases reductions are obtained
- Additional rich patterns lead better encodings