Universiteit Utrecht

[Faculty of Science
Information and Computing Sciences]

# Non-Redundant Subgroup Discovery in Large and Complex Data

Matthijs van Leeuwen[1] & Arno Knobbe[2]

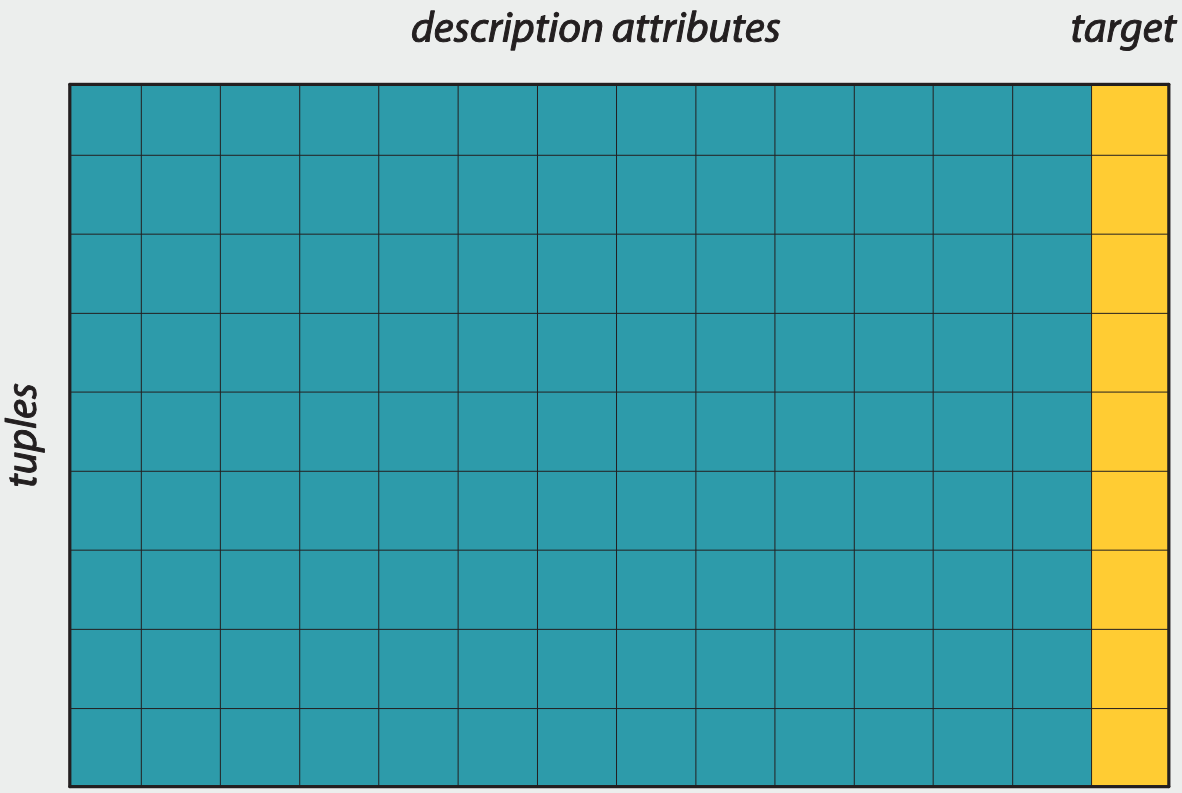
[1] Algorithmic Data Analysis, Universiteit Utrecht
[2] Data Mining Group, Universiteit Leiden

Universiteit Leiden

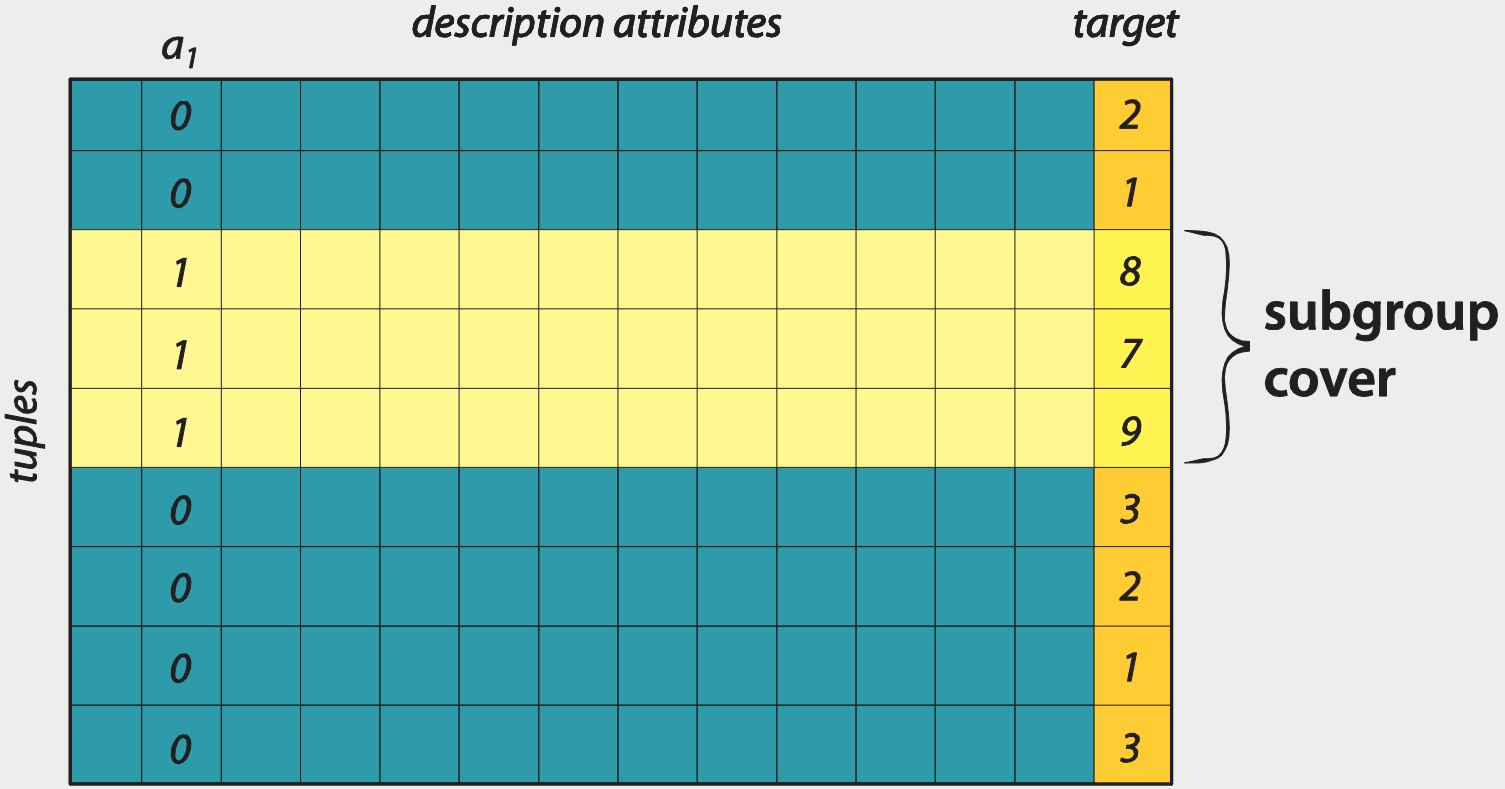
# Subgroup Discovery

- The task
  - Find regions in the data that deviate from the rest, with respect to a given target
- Some details
  - Patterns (or descriptions) define subgroups
  - Quality measure scores subgroups
  - Return top-*k* wrt quality
  - Exhaustive or heuristic search
  - Data can be binary, ordinal, numeric, ...

# Example

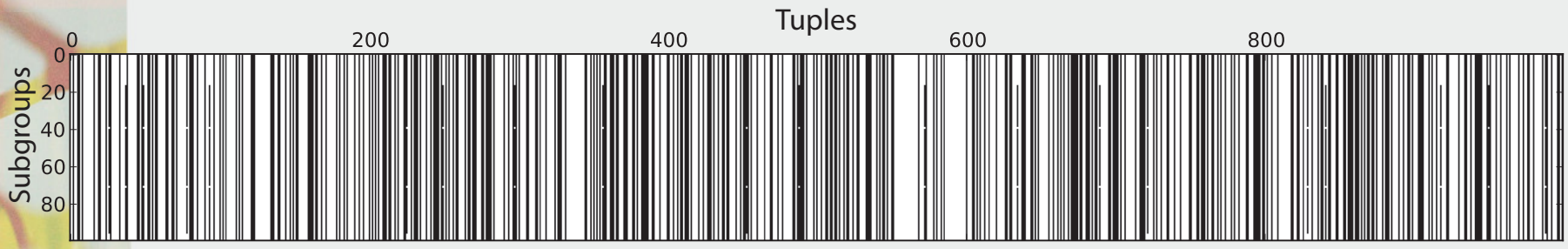


# Example



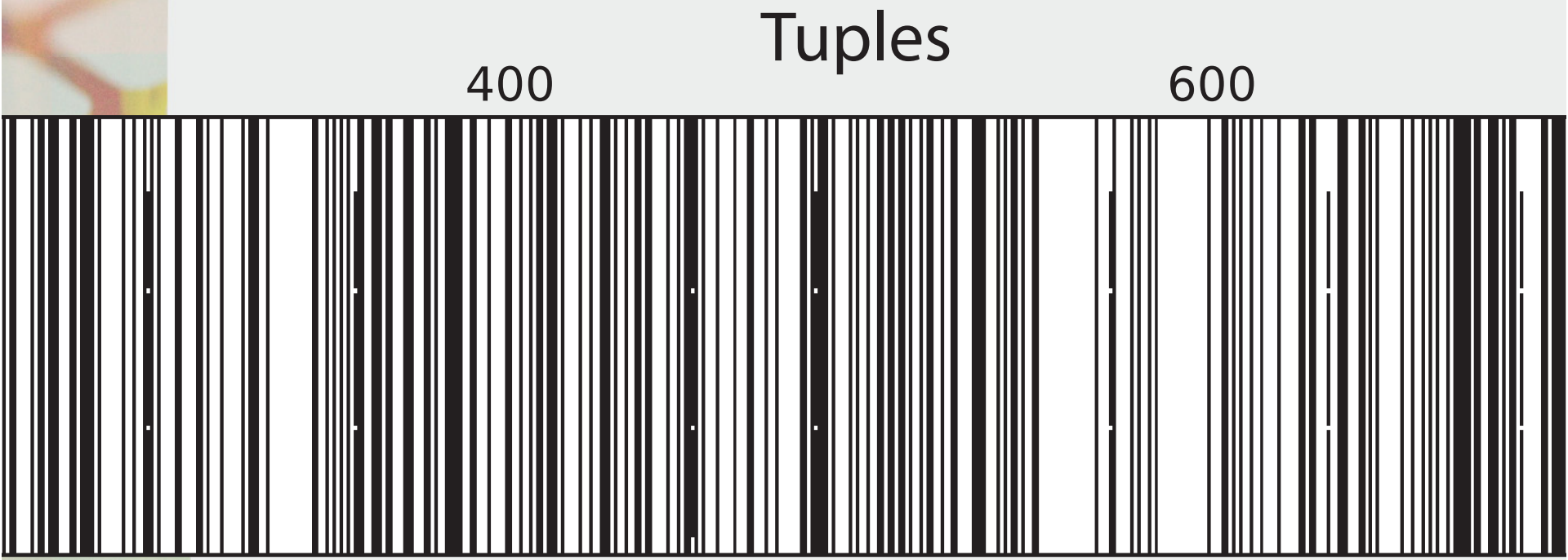

*Description:* $a_1 = 1$
*Quality:* *very good*

# Problem #1




- Subgroup covers for top-100, using exhaustive search
- Dataset: Credit-G
- Quality measure: weighted relative accuracy

# Problem #1



# Problem #1

- Top-4 descriptions
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank && purpose != vacation
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank && purpose != other
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank && personal_status != female_single

# Problem #1

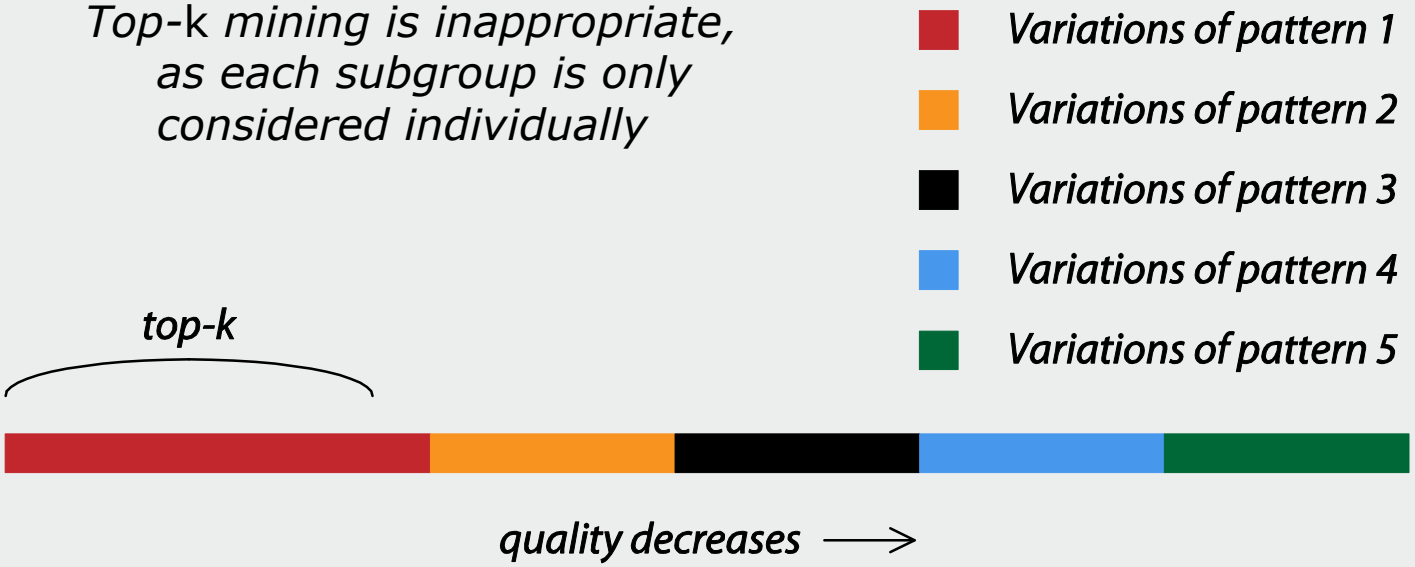
- Top-4 descriptions
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank **&& purpose != vacation**
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank **&& purpose != other**
  - checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank **&& personal_status != female_single**

  - *Identical subgroup covers: 390 tuples, quality = 0.78*

# Problem #1

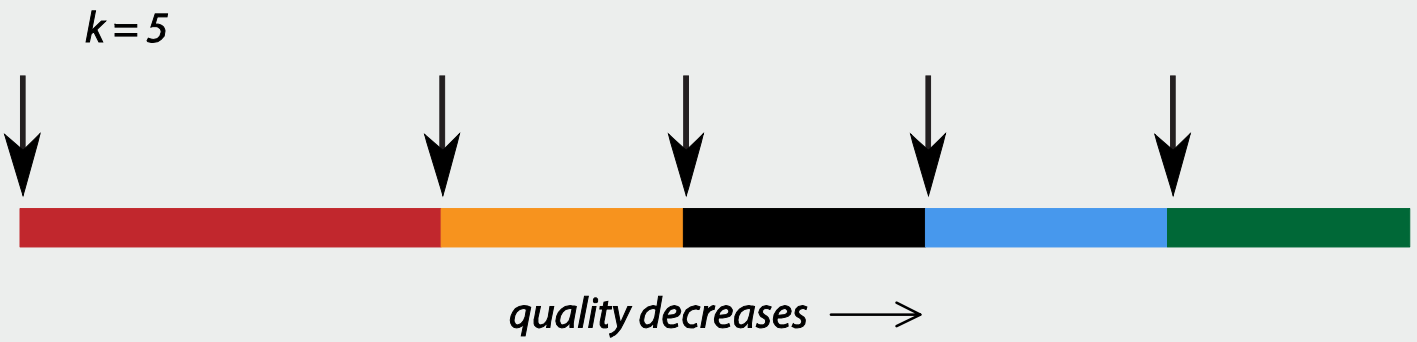
- Redundancy
  - Top-*k* contains many variations of the same theme
  - Other interesting patterns not found

*Top-k mining is inappropriate, as each subgroup is only considered individually*



# Non-Redundant Subgroup Set Mining

- Subgroup sets
  - Considering individual subgroups not good enough
  - Consider *subgroup sets* instead
- Goal
  - Find a *non-redundant* set of *k* high-quality subgroups



# Problem #2

- Complex and large data
  - Search space explosion
    - Many attributes
    - Attributes of high cardinality (numeric attributes!)
    - Correlated attributes
  - Multiple target attributes (= Exceptional Model Mining):
    - Candidate testing time-consuming
    - Often no optimistic estimates available

*Exhaustive search not the way to go*

# Heuristic Search

- Beam search
  - Commonly used for subgroup discovery
  - Fast and effective
  - But ... suffers from the same redundancy problems as exhaustive search
- Goal
  - Modify standard beam search such that it can be used to mine non-redundant subgroup sets

# Degrees of Redundancy

- A subgroup set is *non-redundant* if all its subgroups have substantially different

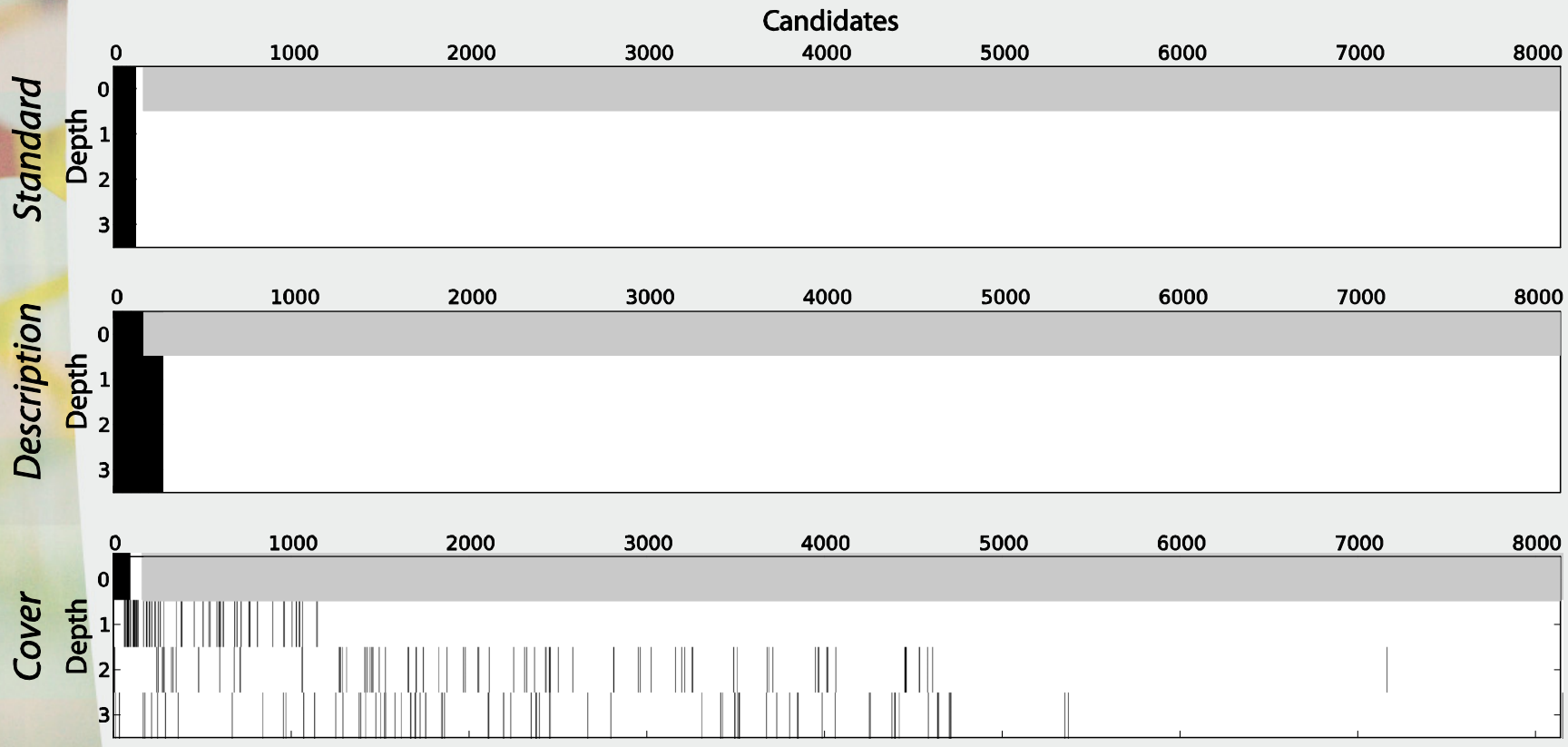
  1. subgroup descriptions / patterns, *or*
  2. subgroup covers, *or*
  3. exceptional models (in case of EMM).

- Each degree is more strict than its predecessor.

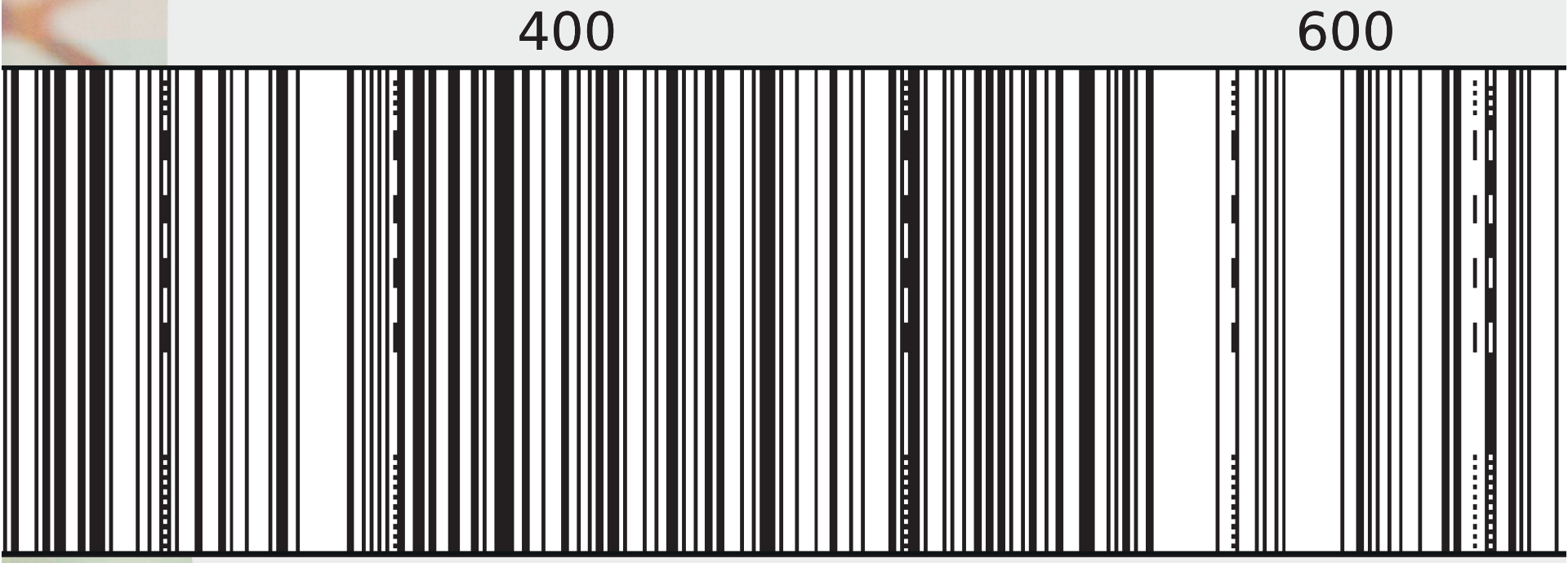
# Non-Redundant Beam Search

- Algorithm
  - Standard beam search
  - Except: do not use top-*k* as beam
  - Instead, select diverse subgroup set
- Beam selection strategies
  - Depends on desired degree of redundancy elimination
  - Heuristic strategy for each of the three degrees

# Beam Selection Strategies in Action



# Top-*k* Beam Selection



# Description-based Beam Selection

# Cover-based Beam Selection

# Subgroup Discovery Results

- Aggregated over 3 datasets
  - *Adult*
  - *Credit-G*
  - *Mushroom*
- Aggregated over 2 quality measures
  - *Weighted relative accuracy*
  - *Weighted KL-based measure*
- Averages per experiment are shown

# Subgroup Discovery Results

| Search | #cands | time (m) | CR |
|---|---|---|---|
| DFS | 403801872 | 1553 | 1.10 |
| Standard | 88641 | 0.3 | 1.23 |
| Description | 88508 | 1.0 | 0.98 |
| Cover | 89116 | 49 | 0.37 |

- CR = Cover Redundancy (lower is more diverse)
- No significant differences in highest obtained qualities
  - Beam search allows larger search spaces, i.e. multiple constraints on a single attribute
  - This regularly leads to better solutions than achievable with DFS

# Exceptional Model Mining Results

- Aggregated over 4 datasets
  - *Adult (different variant)*
  - *Emotions*
  - *Mammals*
  - *Yeast*
- Aggregated over 2 quality measures
  - *Weighted KL-based measure*
  - *Weighted Krimp Gain*
- Averages per experiment are shown

# Exceptional Model Mining Results

| Search | #cands | time (m) | CR |
|---|---|---|---|
| Standard | 244830 | 8 | 1.53 |
| Description | 244659 | 49 | 1.36 |
| Cover | 244830 | 62 | 0.48 |
| Model | 255993 | 143 | 1.07 |

- CR = Cover Redundancy (lower is more diverse)
- No significant differences in highest obtained qualities

# Conclusions

- Non-Redundant Subgroup Set Mining
  - Consider *subgroup sets* instead of individual subgroups
  - Applies to both classical SD and EMM
- Non-Redundant Beam Search
  - Standard beam search with modified beam selection
  - Experiments show that algorithm is fast and effective
  - Optimal subgroups are discovered
- No need for exhaustive search?
  - Proposed methods find diverse yet high-quality subgroup sets
  - Much larger search spaces can be tackled

# Thank you for your attention!

Universiteit Utrecht

Universiteit Leiden

[Faculty of Science
Information and Computing Sciences]