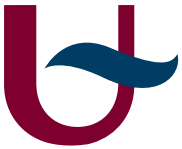
DRule: A Density-based Approach to Discovering Quantitative Association Rules

Thomas Van Brussel Bart Goethals

Universiteit Antwerpen



Overview

Association Rules

- Association rules used to discover relations between attributes

$$\{\text{Beef}, \text{Onion}\} \Rightarrow \{\text{Bun}\}$$

TID	Beef	Onion	Bun	Eggs
1	✓	✓	✓	
2	✓	✓		✓
3	✓	✓	✓	✓
4	✓			✓

- Support $= {}^{2}/_{4}$ and Confidence $= {}^{2}/_{3}$
- Well-researched

Quantitative Association Rules

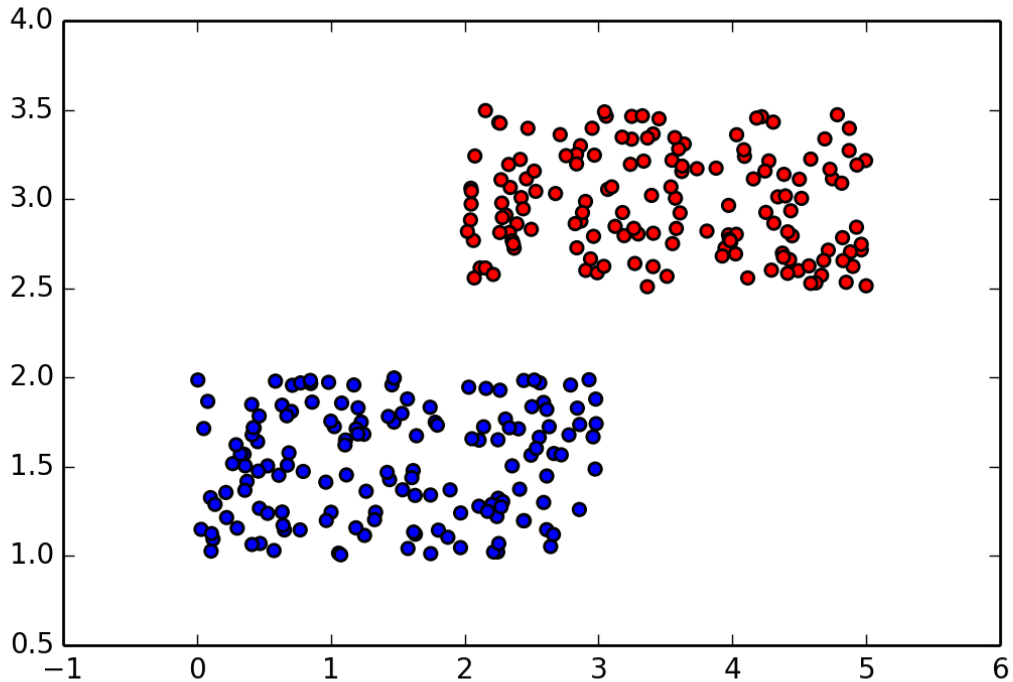
- Numerical data

$$\text{Latitude}[49.72, 50.05] \ \wedge \ \text{Longitude}[77.76, 79.1] \Rightarrow \text{Focal depth}[0.0, 9.0].$$

Quantitative Association Rules

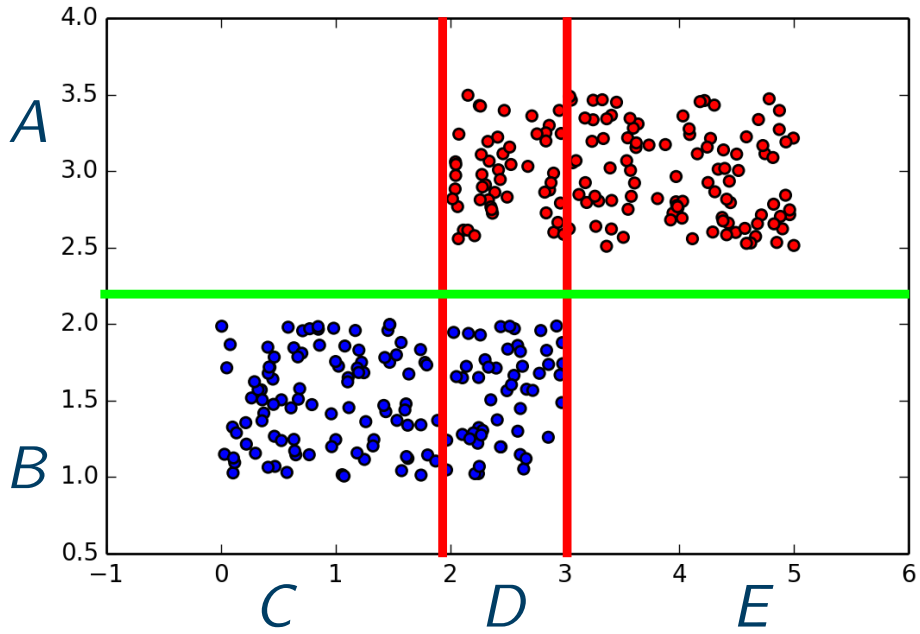
- Numerical data

$$\text{Latitude}[49.72, 50.05] \ \wedge \ \text{Longitude}[77.76, 79.1] \Rightarrow \text{Focal depth}[0.0, 9.0].$$

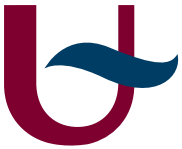


Other Approaches

- Discretization



- Overlapping regions cause issues
- Can cause loss of information



Other Approaches

Mostly optimization strategies

- Image segmentation algorithms
  - Limited in the number of dimensions
- Genetic algorithms
  - Follow a rule template

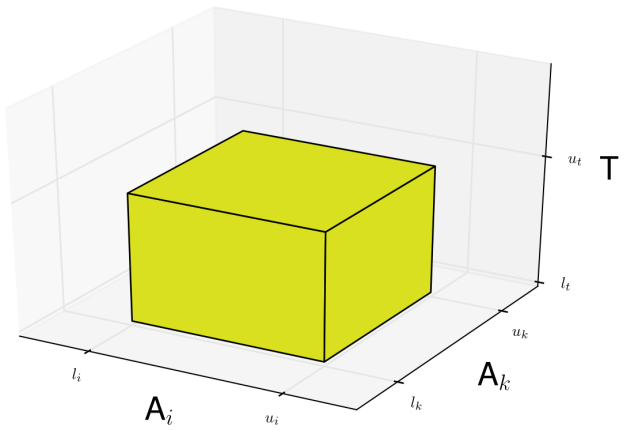
Not capable of finding all rules in the data. Restrictions might be very severe.

Intuition

- Rules of the form

$$A_i[l_i, u_i] \wedge \ldots \wedge A_k[l_k, u_k] \Rightarrow T[l_t, u_t]$$

- Support, confidence not enough
  - Use all points to maximize
- Apply different measure: **density**

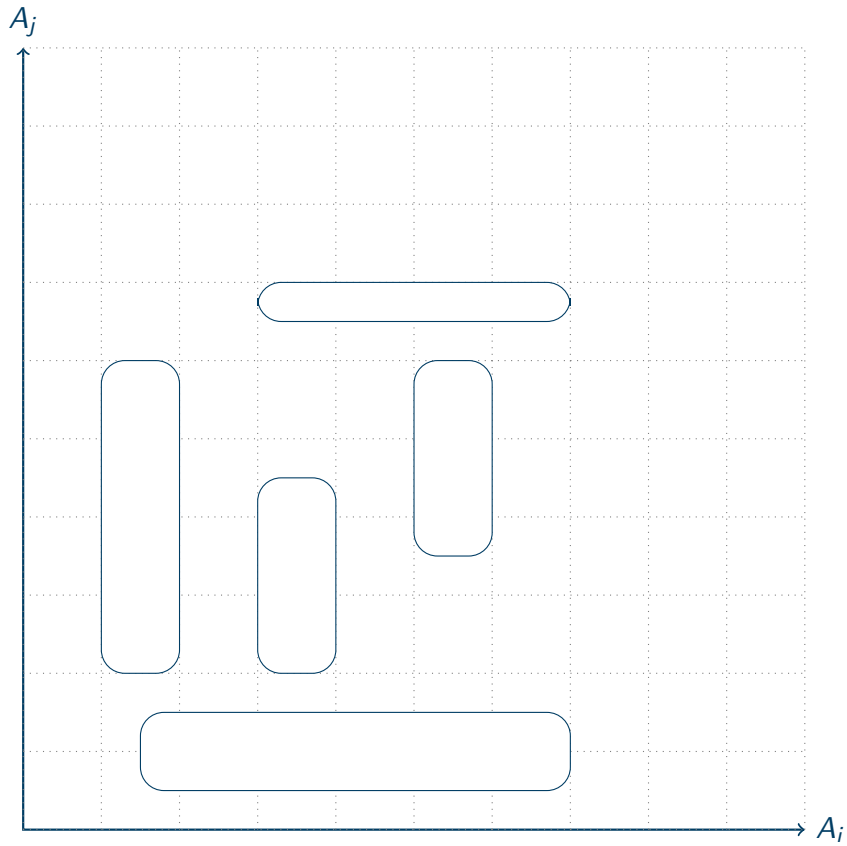


$$|N(x) = \{y \mid d(x, y) \leq \epsilon\}| \geq \texttt{min\_pts}$$

- Only require density for dimensions seperately
  - Density in all dimensions can miss good rules

Algorithm – Simple Example

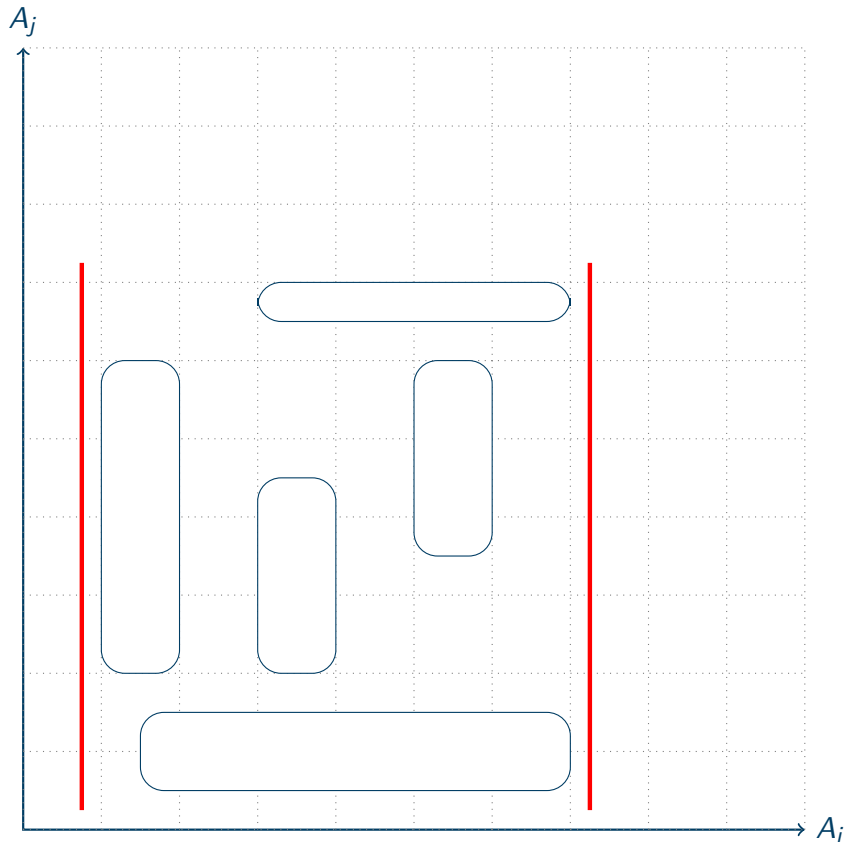
Every rectangle is a dense region



Algorithm – Simple Example

Every rectangle is a dense region

1. Find all regions in dimension A_i

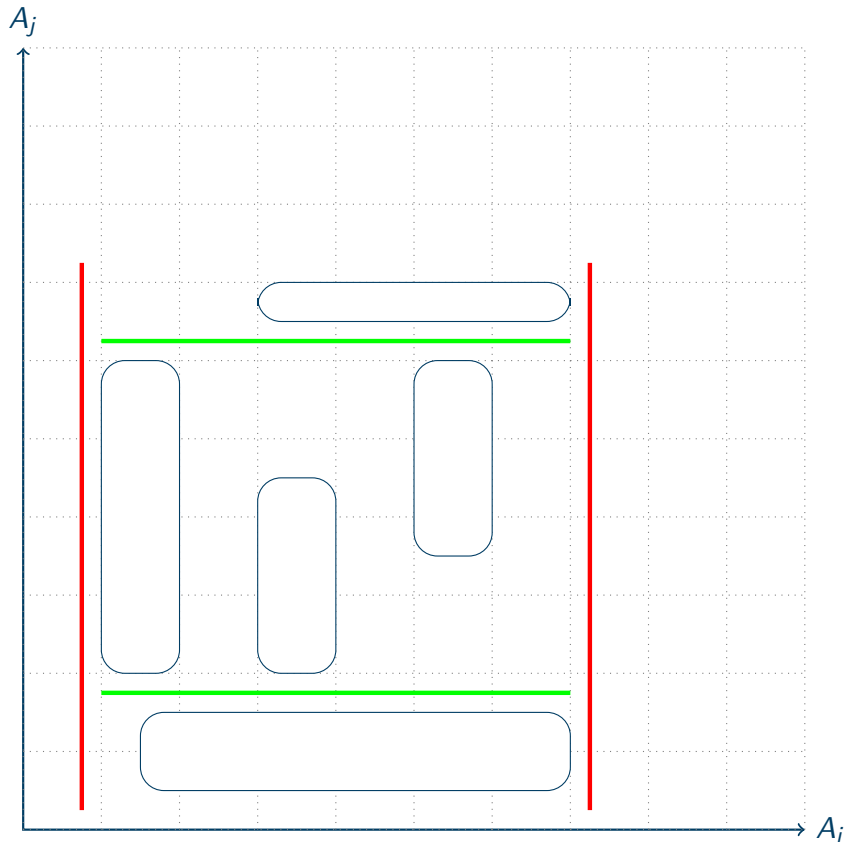


Algorithm – Simple Example

Every rectangle is a dense region

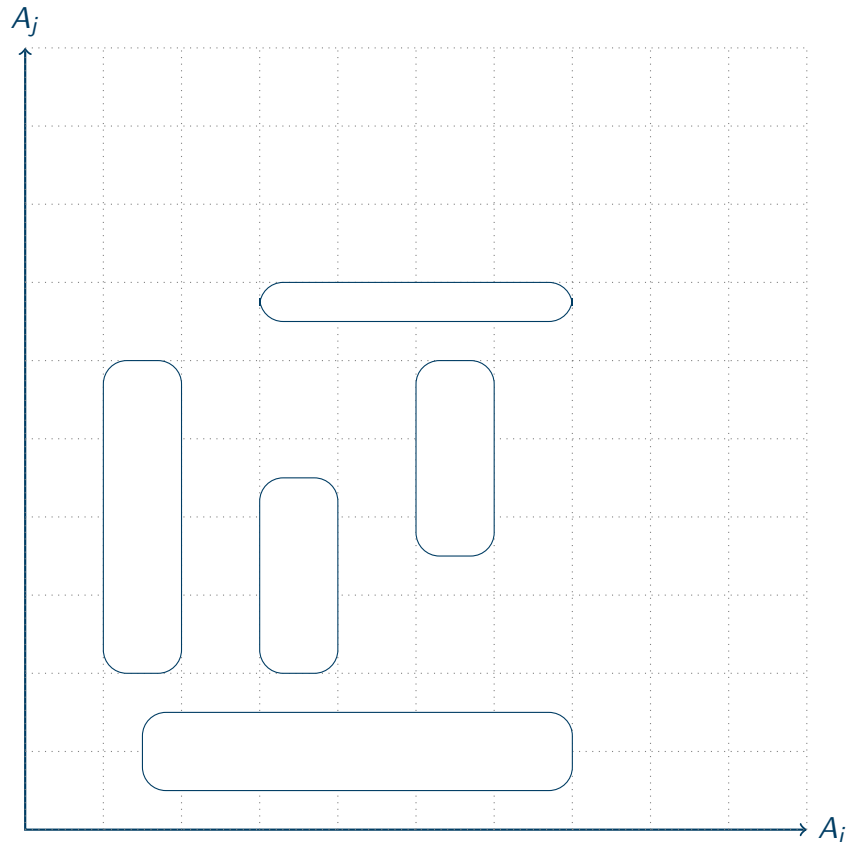
1. Find all regions in dimension A_i
2. Find all regions in dimension A_j

Found three regions for rule generation
(What if we want them all?)



Algorithm – Simple Example

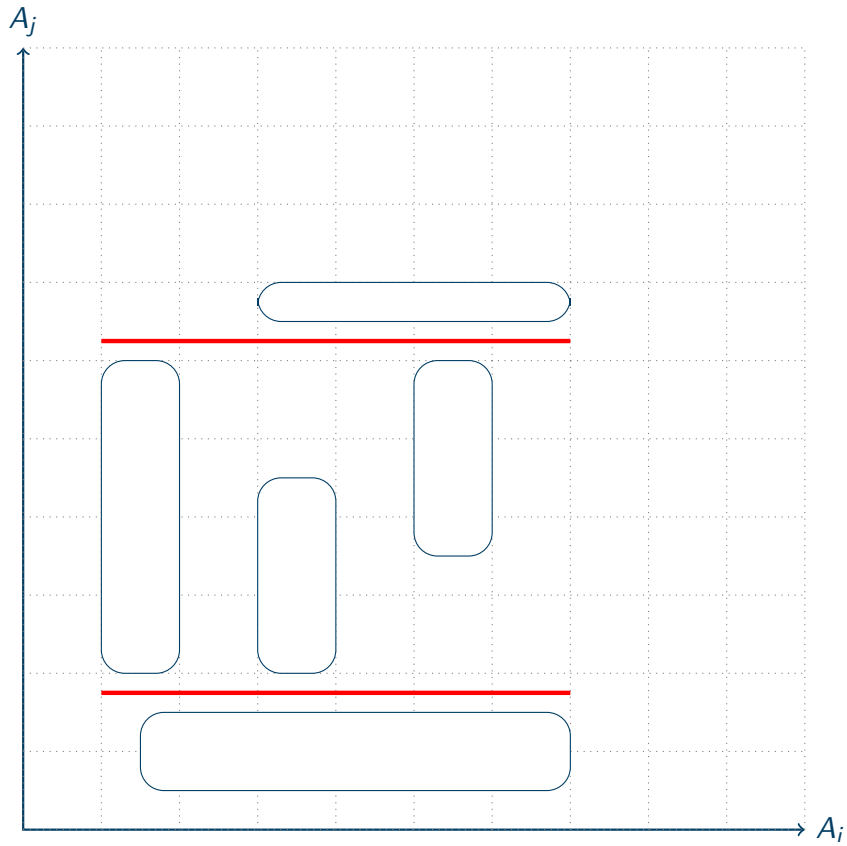
Repeat starting from A_j



Algorithm – Simple Example

Repeat starting from A_j

1. Find all dense regions in dimension A_j

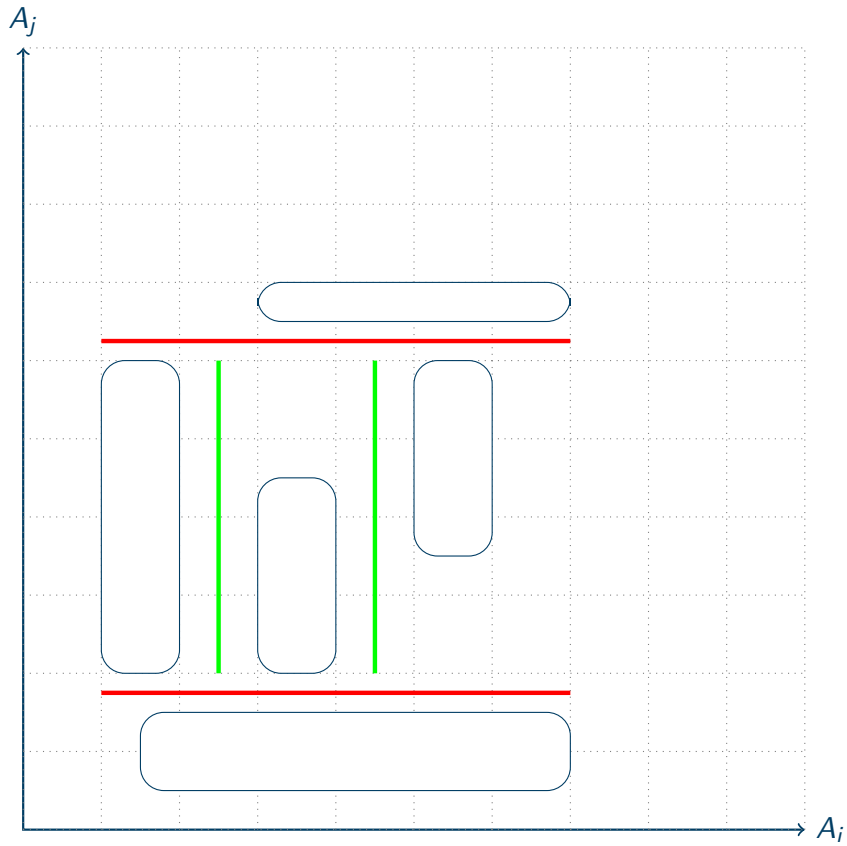


Algorithm – Simple Example

Repeat starting from A_j

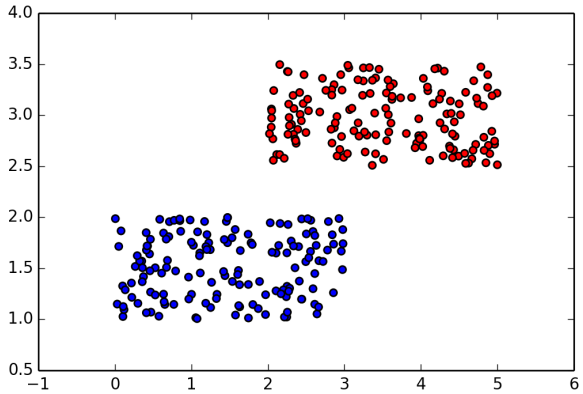
1. Find all dense regions in dimension A_j
2. Scan resulting regions for dense regions in dimension A_i

Finds all smaller regions



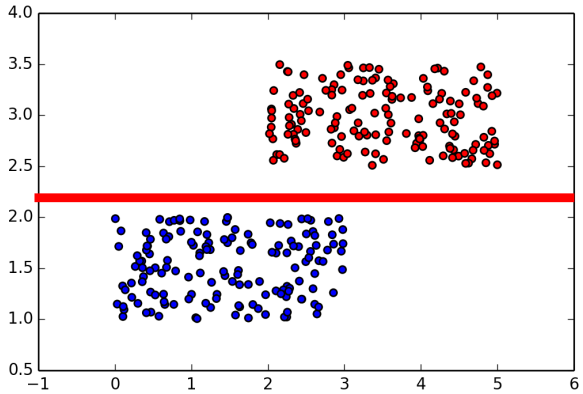
Algorithm – Overlap

- Overlap in horizontal dimension
- Regions easily separated by algorithm
- Difficult for discretization



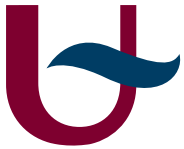
Algorithm – Overlap

- Overlap in horizontal dimension
- Regions easily separated by algorithm
- Difficult for discretization



Example rules:

- $Y[2.5, 3.5] \Rightarrow X[2, 5]$ (red points)
- $X[0, 3] \Rightarrow Y[1, 2]$ (blue points)

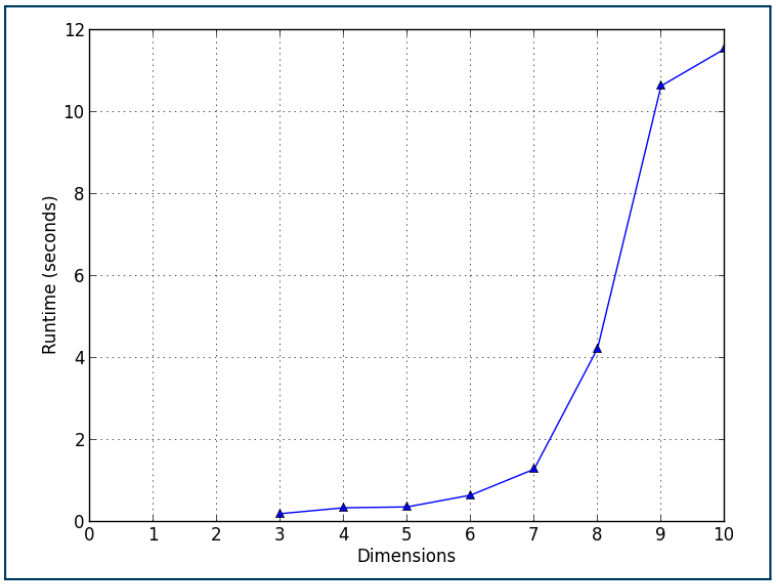


Conclusion

Novel method:

- exhaustively mine quantitative association rules
- while not having to preprocess the data (discretization) or
- specify what rules we want in advance and
- allowing for overlap between rules.

Future work



- Not enough pruning (yet)
  - Many patterns
  - Techniques from classical association rule mining might prove effective
- Efficiency is still a concern
  - Search through dimensions can be exponential