DMV – Subgroup Discovery

Alexandre Termier

General idea of Subgroup Discovery

- Again, goal = finding interesting patterns in the data

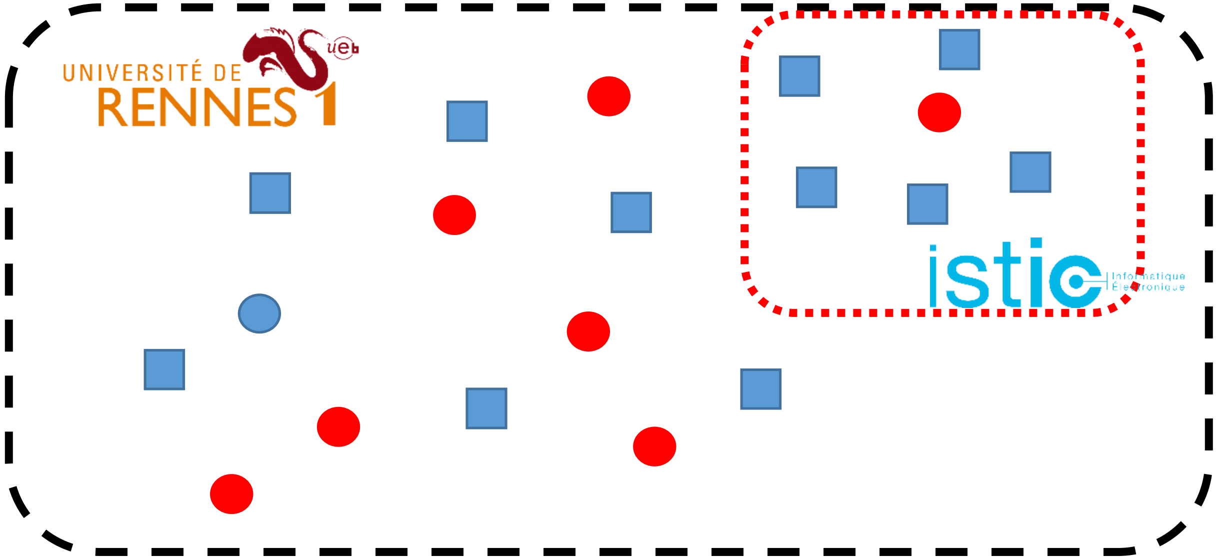
- Flexible definition of interestingness measure / quality function
    - Exploits settings similar to classification or regression (target attribute(s))

- Attributes can be nominal, ordinal or numerical

Subgroup Discovery Task

Find description of subsets in the data that differ significantly of the total population with respect to a target concept

Ex: *80% of students in Computer Science are males, compared to 50% in all students of the University*



Pattern language

- Pattern: rule of the form

  conjuction of selectors => property on target concept (quality function)

- Selector:
  - Nominal attribute: *attribute = value*
  - Ordinal/numeric attribute: *attribute* $\in$ *[minValue, maxValue]*

- Ex:
  - Study = ‘Computer Science’ => Higher % of ‘Male’ than student population
  - Sex = ‘Female’ **and** Age $\in$ [40,60] => Higher ‘Number of activities’ than total population

Quality functions – Binary/Nominal target

- Quality function:

  $$q\text{: } \textit{Pattern space} \rightarrow \mathbb{R}$$

- Most quality functions for binary targets rely on the confusion matrix

- Many possible measures

  Good reference:
  Francisco Herrera, Cristóbal J. Carmona, Pedro González, María José del Jesús:
  An overview on subgroup discovery: foundations and applications. Knowl. Inf. Syst. 29(3): 495-525 (2011)

Target (columns) / Subgroup description (rows)

	True	False
True	a	b
False	c	d

n = a+b (extent of subgroup)
N = a+b+c+d (all population)

Ex.: Weighted Relative Accuracy

		Target	
		True	**False**
Subgroup description	**True**	a	b
	False	c	d

n = a+b (extent of subgroup)
N = a+b+c+d (all population)

- Weighted Relative Accuracy (WRAcc):
  - Measure of « unusualness »
  - Intuitively: Coverage of subgroup * accuracy gain

- Formula:

$$WRAcc(P) = \frac{n}{N} \cdot (t_P - t_{all}) = \frac{n}{N} \cdot (\frac{a}{n} - \frac{a+c}{N})$$

Example with binary target

Income	Sex	Age	Education level	Married	Has Children
High	M	>50	High	Y	Y
High	M	>50	Medium	Y	Y
High	F	40-50	Medium	Y	Y
High	M	40-50	Low	N	Y
Medium	M	30-40	Medium	Y	Y
Medium	M	>50	High	Y	N
Low	M	<30	High	Y	N
Medium	F	<30	Medium	Y	N
Low	F	40-50	Low	Y	N
Low	M	40-50	Medium	N	N
Medium	F	>50	Medium	N	N
Low	F	<30	Low	N	N
Low	F	30-40	Medium	N	N
Low	F	40-50	Low	N	N
Low	M	<30	Low	N	N
Medium	F	30-40	Medium	N	N

Target: Income='High'

Ex. from slides of M. Atzmueller

Example with binary target

Income	Sex	Age	Education level	Married	Has Children
High	M	>50	High	Y	Y
High	M	>50	Medium	Y	Y
High	F	40-50	Medium	Y	Y
High	M	40-50	Low	N	Y
Medium	M	30-40	Medium	Y	Y
Medium	M	>50	High	Y	N
Low	**M**	**<30**	**High**	**Y**	**N**
Medium	F	<30	Medium	Y	N
Low	F	40-50	Low	Y	N
Low	M	40-50	Medium	N	N
Medium	F	>50	Medium	N	N
Low	F	<30	Low	N	N
Low	F	30-40	Medium	N	N
Low	F	40-50	Low	N	N
Low	**M**	**<30**	**Low**	**N**	**N**
Medium	F	30-40	Medium	N	N

Target: Income='High'

Subgroup: Sex='M' **and** Age<30
n=2, a=0
WRAcc = 2/16*(0/2-4/16) = -0.03125

Ex. from slides of M. Atzmueller

Example with binary target

Income	Sex	Age	Education level	Married	Has Children
High	**M**	**>50**	**High**	**Y**	**Y**
High	**M**	**>50**	**Medium**	**Y**	**Y**
High	**F**	**40-50**	**Medium**	**Y**	**Y**
High	M	40-50	Low	N	Y
Medium	**M**	**30-40**	**Medium**	**Y**	**Y**
Medium	**M**	**>50**	**High**	**Y**	**N**
Low	**M**	**<30**	**High**	**Y**	**N**
Medium	**F**	**<30**	**Medium**	**Y**	**N**
Low	**F**	**40-50**	**Low**	**Y**	**N**
Low	M	40-50	Medium	N	N
Medium	F	>50	Medium	N	N
Low	F	<30	Low	N	N
Low	F	30-40	Medium	N	N
Low	F	40-50	Low	N	N
Low	M	<30	Low	N	N
Medium	F	30-40	Medium	N	N

Target: Income='High'

Subgroup: Sex='M' **and** Age<30
n=2, a=0
WRAcc = 2/16*(0/2-4/16) = -0.03125

Subgroup: Married='Y'
n=8, a=3
WRAcc = 8/16*(3/8-4/16) = 0.0625

Ex. from slides of M. Atzmueller

Example with binary target

Income	Sex	Age	Education level	Married	Has Children
High	M	>50	High	Y	Y
High	M	>50	Medium	Y	Y
High	F	40-50	Medium	Y	Y
High	M	40-50	Low	N	Y
Medium	M	30-40	Medium	Y	Y
Medium	M	>50	High	Y	N
Low	M	<30	High	Y	N
Medium	F	<30	Medium	Y	N
Low	F	40-50	Low	Y	N
Low	M	40-50	Medium	N	N
Medium	F	>50	Medium	N	N
Low	F	<30	Low	N	N
Low	F	30-40	Medium	N	N
Low	F	40-50	Low	N	N
Low	M	<30	Low	N	N
Medium	F	30-40	Medium	N	N

Target: Income='High'

Subgroup: Sex='M' **and** Age<30
n=2, a=0
WRAcc = 2/16*(0/2-4/16) = -0.03125

Subgroup: Married='Y'
n=8, a=3
WRAcc = 8/16*(3/8-4/16) = 0.0625

Subgroup: HasChildren='Y'
n=5, a=4
WRAcc = 5/16*(4/5-4/16) = **0.17**

Ex. from slides of M. Atzmueller

Quality functions – numeric target

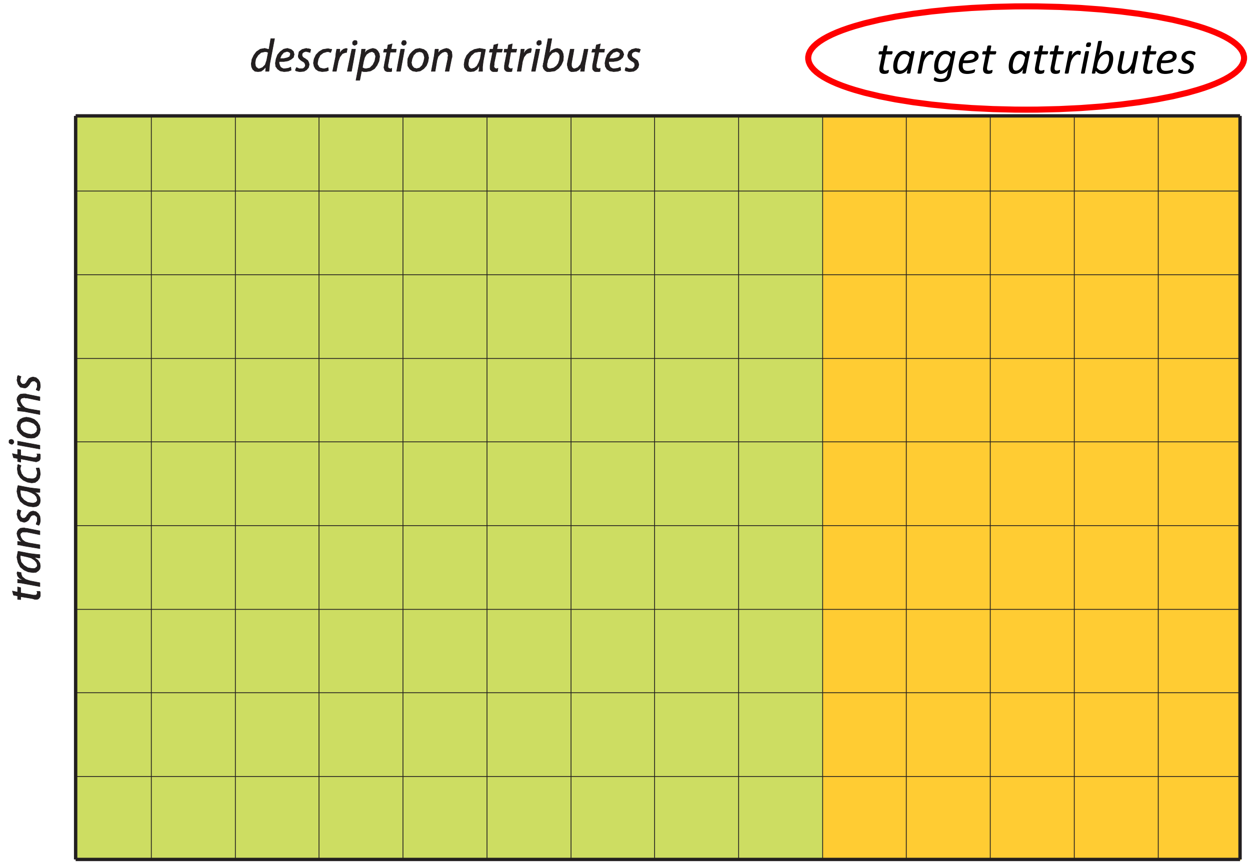
- Replace percentages of presence by mean of target value
- Formula:

$$q_M^a(Patt) = \left(\frac{n}{N}\right)^a \cdot (m_{patt} - m_{all}) \qquad a \in [0;1]$$

m_{patt} = mean of target value on rows covered by subgroup
m_{all} = mean of target value in the whole dataset

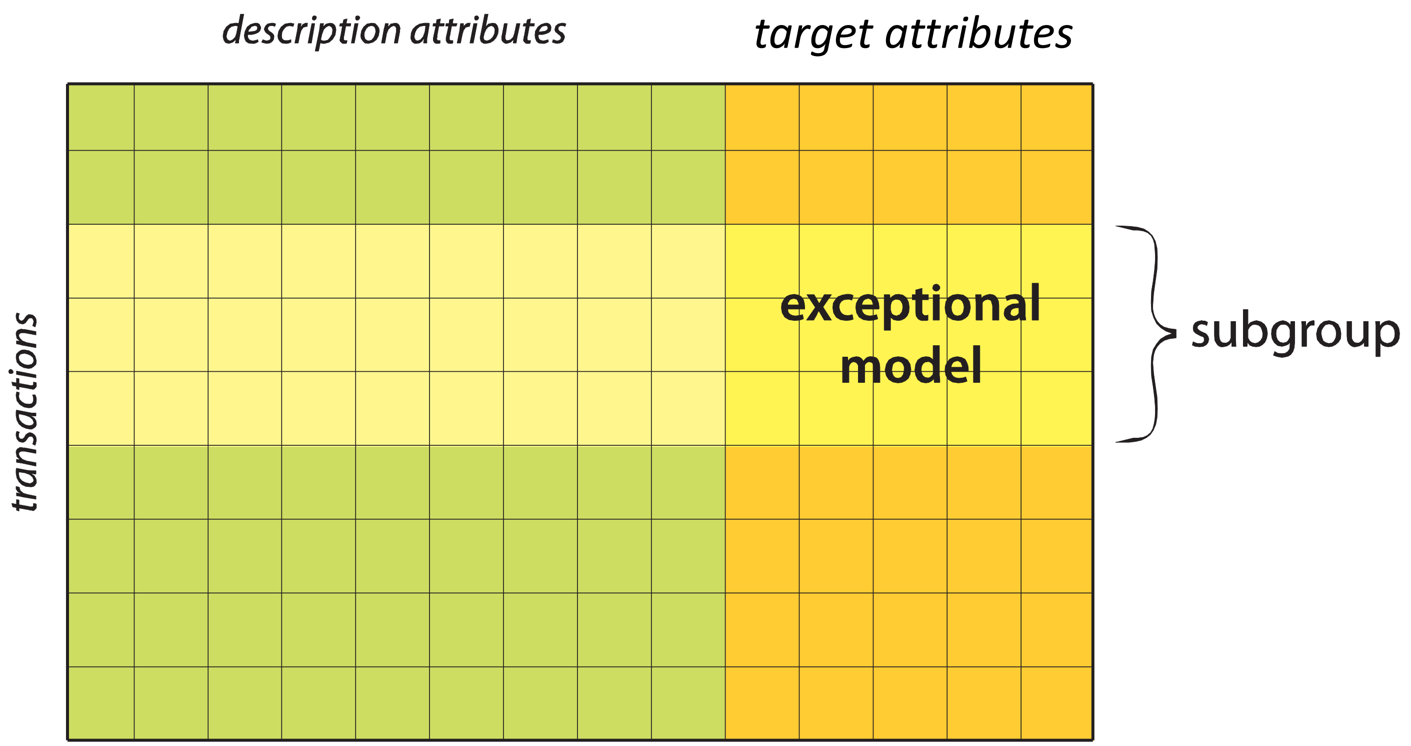
- a = 0 -> *mean gain* (do not consider size of subgroup)
- a = 0.5 -> *mean test*
- a = 1 -> *impact*

What about multiple targets?



Slide courtesy of M. van Leeuwen

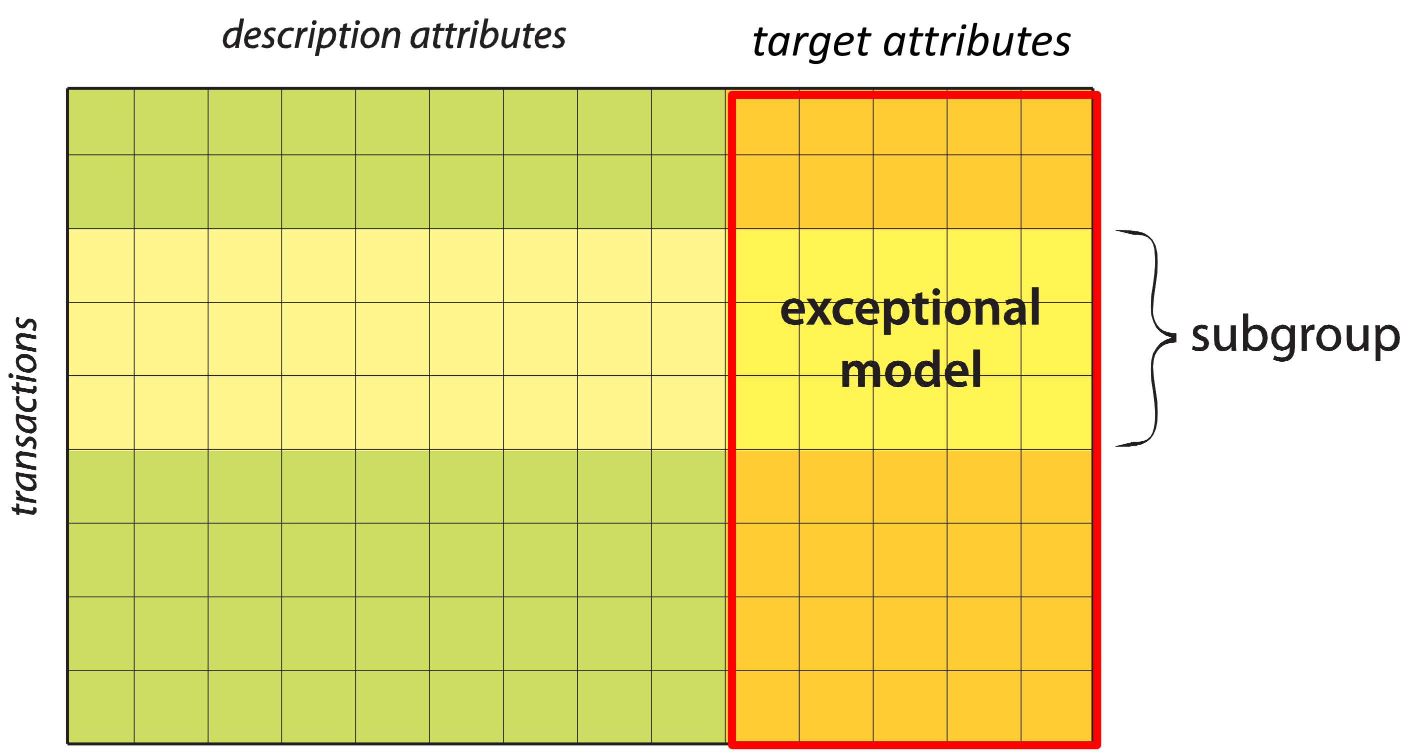
Exceptional Model Mining



Subgroup model (on target attributes) substantially different from model on complement or all data

Slide courtesy of M. van Leeuwen

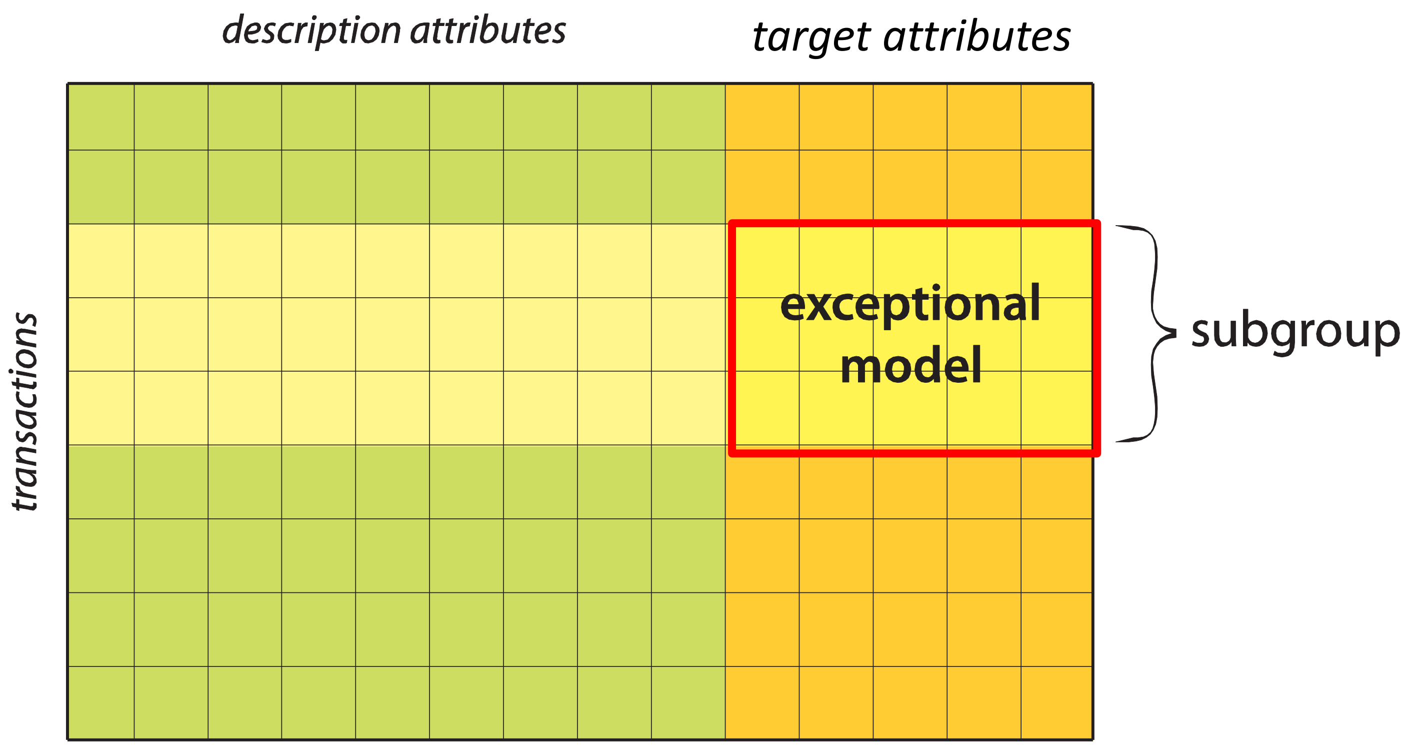
Exceptional Model Mining



Subgroup model (on target attributes) substantially different from model on complement or all data

Slide courtesy of M. van Leeuwen

Exceptional Model Mining



Subgroup model (on target attributes) substantially different from model on complement or all data

Slide courtesy of M. van Leeuwen

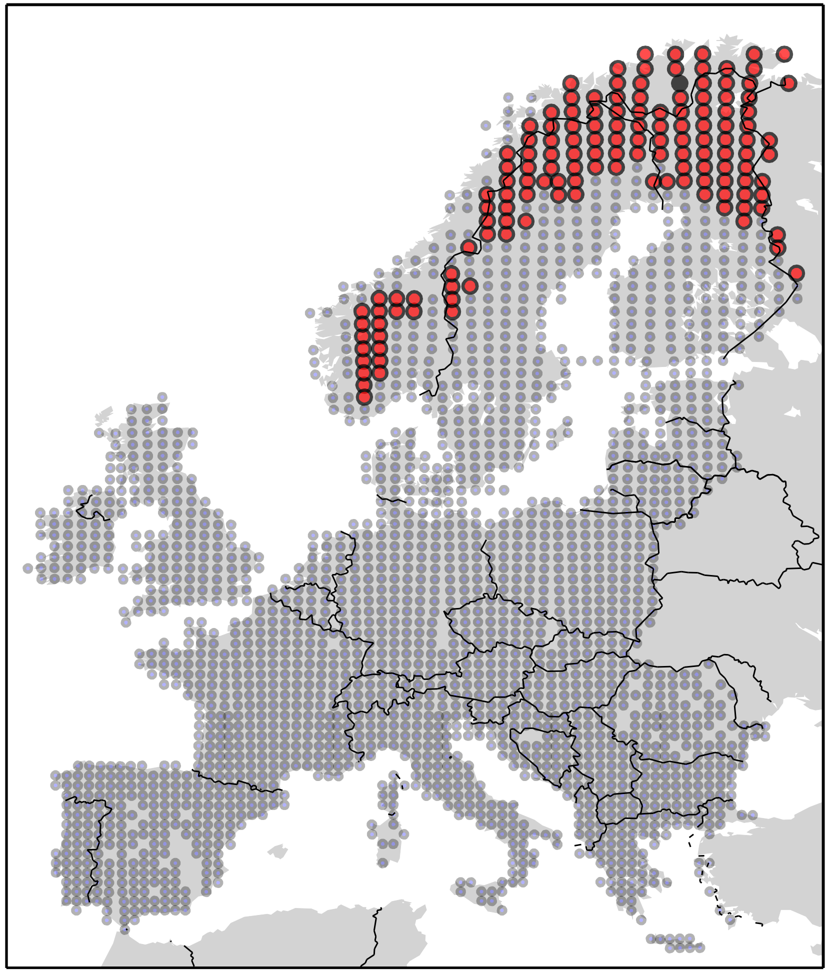
Exceptional model mining (EMM)

- Important parameter: a *class* for a model of interest over target attributes
  - Ex: linear model, gaussian process, conjuction of values…
- From this class:
  - A model is inferred from all the data
  - A model is inferred from the subgroup
  - The subgroup model is **exceptional** if it significantly differs from the model for all the data

EMM ex. 1

Model class: conjunction of literals

Target attributes: presence/absence of animals



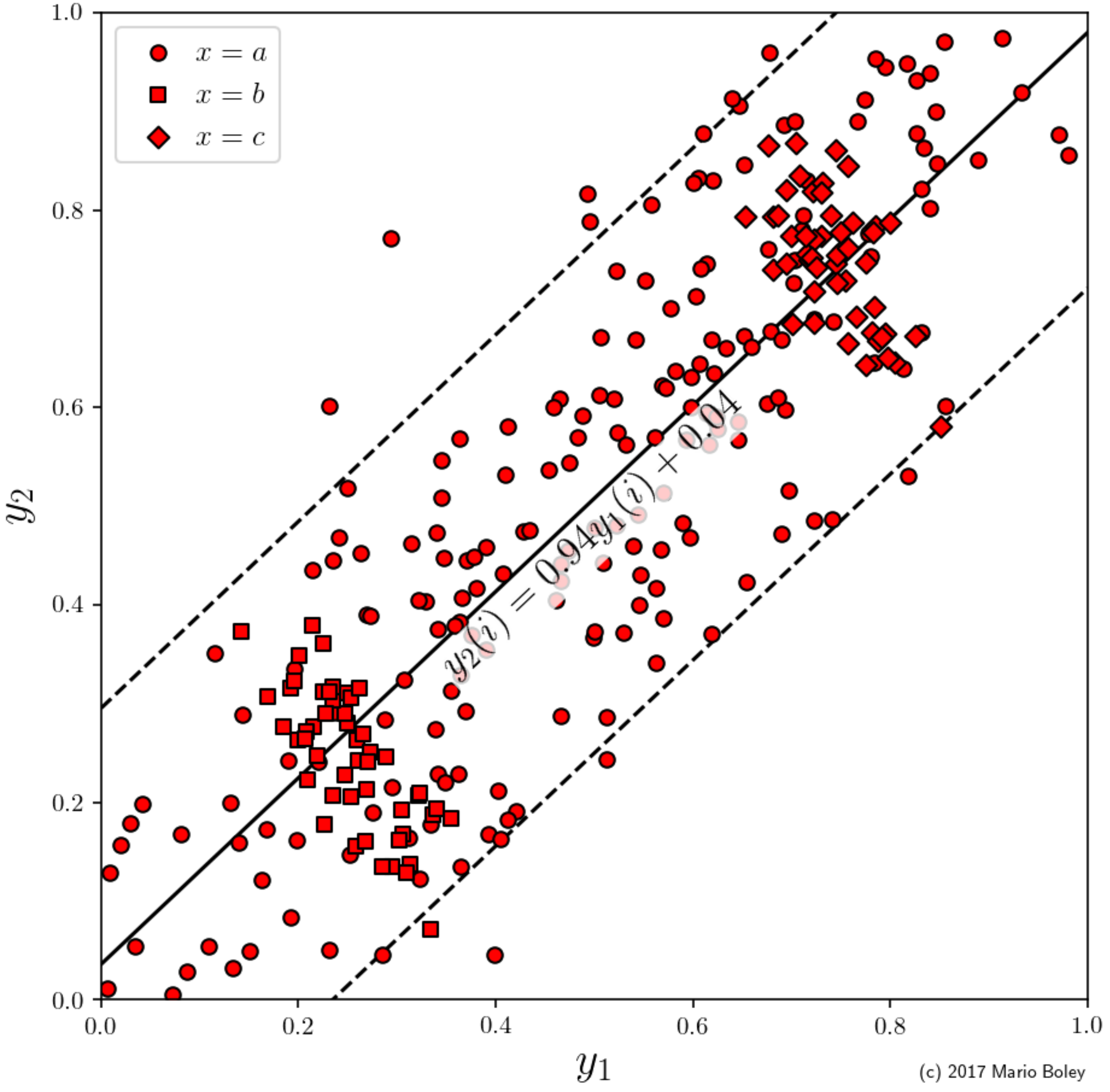
- IF (*subgroup description*)
  - (max temperature in September <= 11.1°C AND
  - max temperature in April <= 3.47°C AND
  - max temperature in November >= -2.56°C)
- THEN (*exceptional model*)
  - Arctic fox AND
  - Skunk bear AND
  - Norway lemming AND
  - Elk all occur relatively often (compared to all of Europe)

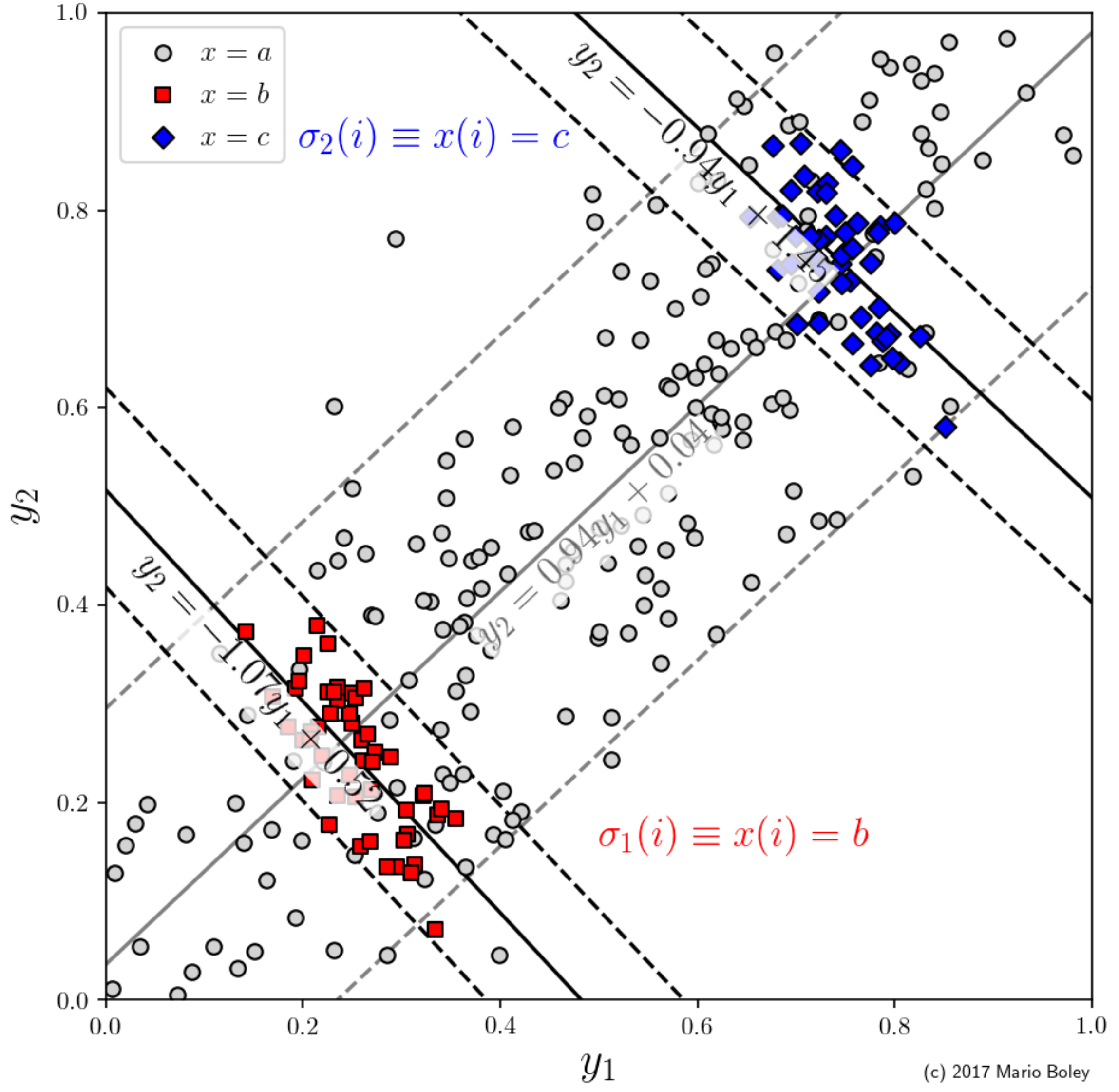
Slide adapted from M. van Leeuwen

EMM ex. 2

Model class: linear relation

Target attributes: y_1, y_2

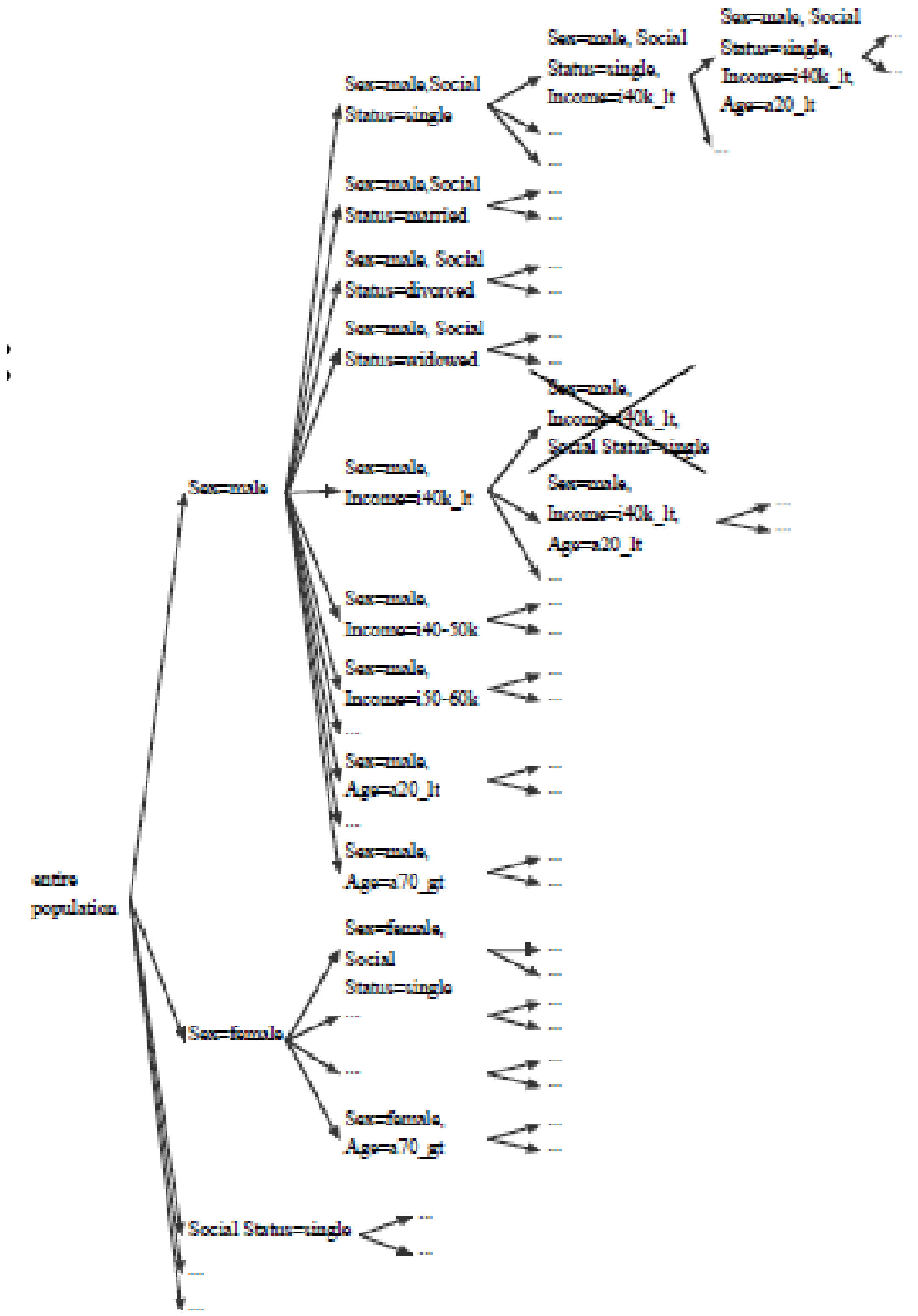




Searching subgroups

- Exhaustive search
  - DFS
  - Efficient data structures (~FP-tree)
  - Pruning

  - Some algorithms
    - SD-Map [Atzmueller & Puppe 2006]
    - SD-Map* [Atzmueller & Lemmerich 2009]
    - Merge-SD [Grosskreutz & Rueping 2009]



Slide adapted from M. Atzmueller

Pruning in exhaustive search

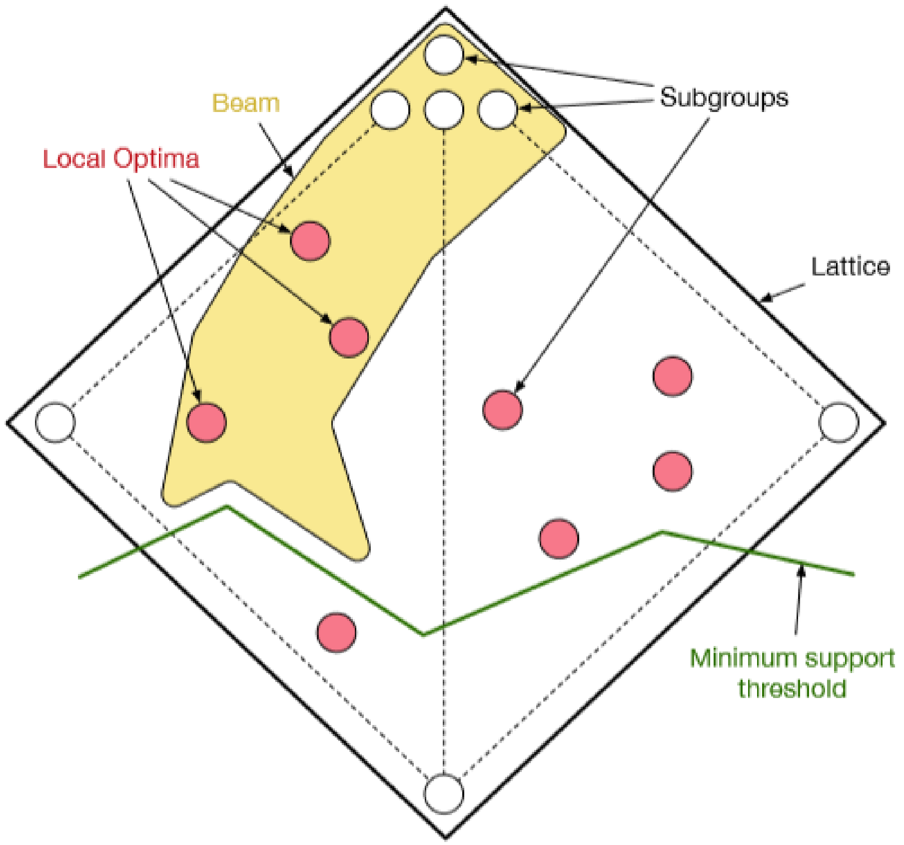
Mostly based on top-k pruning

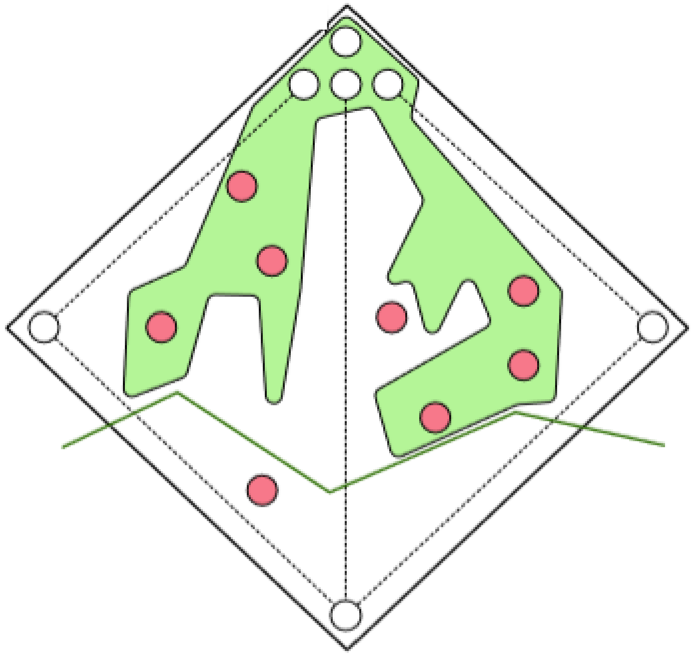
- Idea: Output the top-k subgroups in decreasing order of quality measure

- Optimistic estimates
  - Function *oe* s.t. $P' \supset P \Rightarrow oe(P) > quality(P')$
    - no refinement of *P* can exceed the quality *oe(P)*
  - For many quality function *oe* can be found
  - In top-k setting, if in a branch *oe(P)* < quality of worst patt. of top-k -> branch can be abandonned

Heuristic search

Mostly beam search
[van Leeuwen et al, 2012]





Recent work use Monte-Carlo Tree Search (MCTS)
[Bosc et al., 2018]

Figures courtesy of G. Bosc

The perils of top-*k* mining

1. Expressive but redundant pattern languages
   - Many descriptions for the same cover
   - Results in uniform top-*k*'s
2. Problems irrespective of search
   - Exhaustive: find all redundant patterns
   - Beam search: repeated top-*k* selection

Neither of these problems is specific to SD!

Slide courtesy of M. van Leeuwen

Top-4 descriptions

1. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank
2. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank && purpose != vacation
3. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank && purpose != other
4. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank && personal_status != female_single

Slide courtesy of M. van Leeuwen

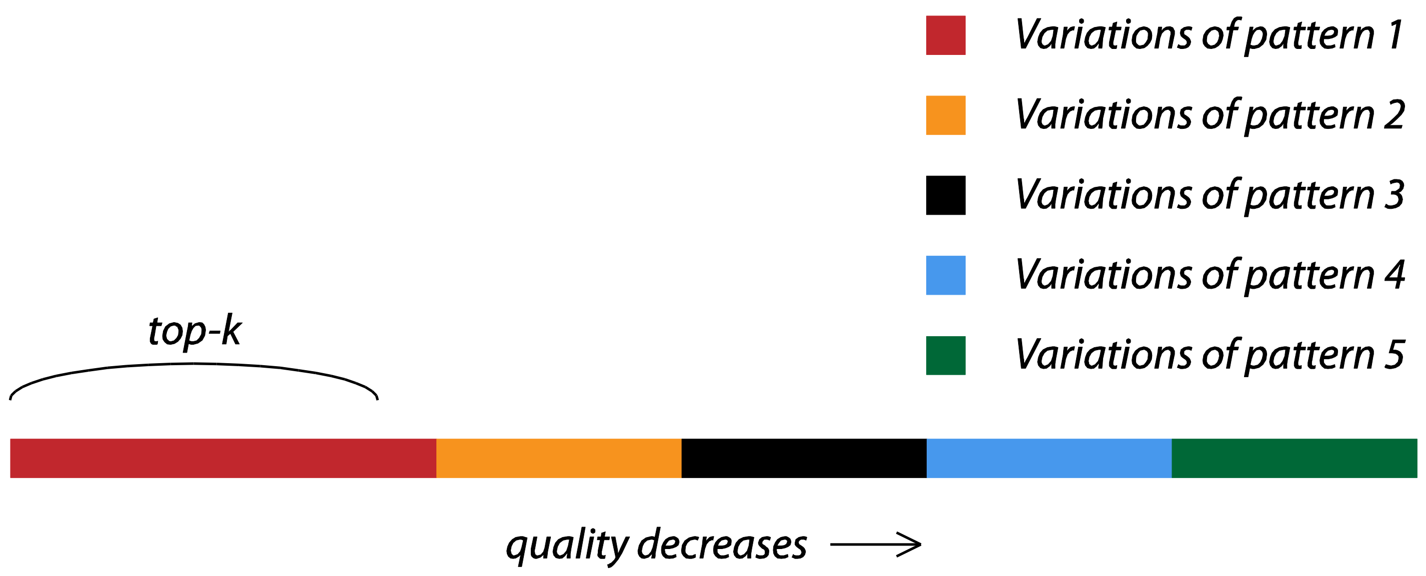
Top-4 descriptions

1. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank
2. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank **&& purpose != vacation**
3. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank **&& purpose != other**
4. checking_status != <0 && checking_status != 0<=X<200 && other_parties != co_applicant && other_payment_plans != bank **&& personal_status != female_single**

Slide courtesy of M. van Leeuwen

Redundancy in top-*k*

- Many variations of the same theme
- Other interesting patterns not found

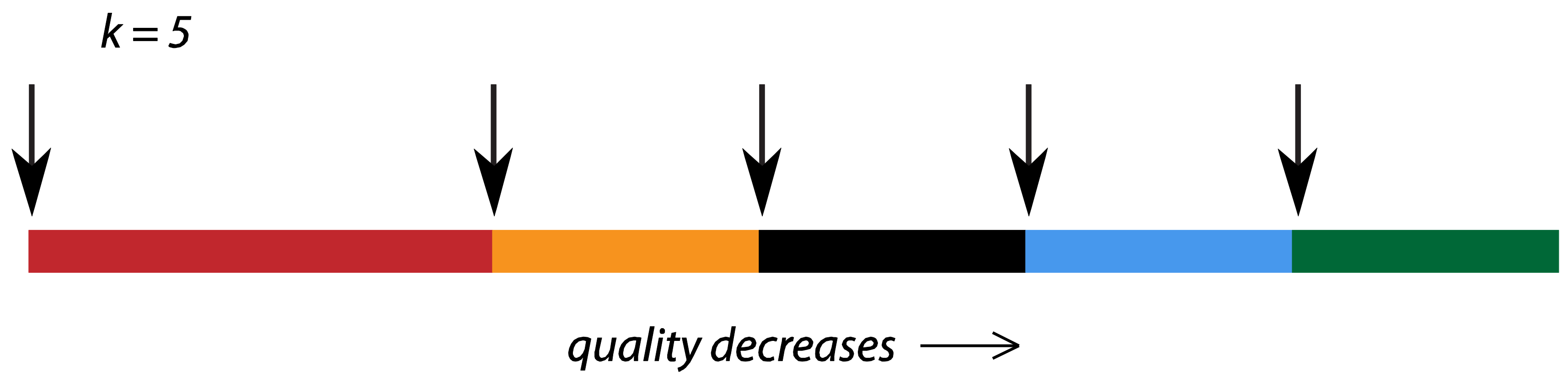


Slide courtesy of M. van Leeuwen

Subgroup Set Selection

Inspired by pattern set mining techniques

1. Given a set of candidate subgroups
2. Find a diverse set of k high-quality subgroups



Slide courtesy of M. van Leeuwen

Diverse Subgroup Set Discovery

Integrate pattern selection into search

Diverse beam search

- Use subgroup set selection for choosing beam
- I.e. diverse top-*k* instead of strict top-*k*

Slide courtesy of M. van Leeuwen

Degrees of Diversity

A subgroup set is diverse if all its subgroups have substantially different

1. subgroup descriptions,
2. subgroup covers,
3. exceptional models.

Each degree is more strict than its predecessor.

Slide courtesy of M. van Leeuwen

Software

- Vikamine: http://www.vikamine.org/ (Atzmueller)
  - Subgroup Discovery and Analytics
  - Comes with a GUI

- DSSD : http://www.patternsthatmatter.org/software.php (van Leeuwen)
  - Diverse Subgroup Set Discovery

- Read-KD library : http://www.realkd.org/realkd-library/
  - EMM + DSSD