Mining the strongest emerging patterns characterizing patients affected by diseases due to atherosclerosis

Bruno Crémilleux and Arnaud Soulet and François Rioult

GREYC - CNRS UMR 6072

Université de Caen - France

Context

STULONG data (atherosclerosis).

twenty-year long longitudinal study of the risk factors of the atherosclerosis (1417 men, former Czechoslovakia)

Main goal of the STULONG study:

identify atherosclerosis risk factors and follow the development of these risk factors and their impacts

Our contribution in this discovery challenge:

search of strong emerging patterns characterizing patients affected by diseases due to atherosclerosis

Motivations

- evaluate the potential impact of the search of emerging patterns in domains like medicine
- discover useful emerging patterns in atherosclerosis data

what is it difficult?

- a huge number of potential emerging patterns

  ➥ requires powerful Knowledge Discovery in Databases (KDD) methods.

- selection of “interesting” emerging patterns

  ➥ **strong emerging patterns.**

Emerging patterns: running example

emerging pattern (EP): pattern whose frequency increases significantly from one data set to another

Patient	Items					
P_1	A	B	C	D		$\mathcal{D}_1$
P_2	A	B	C	D		
P_3	A	B	C			
P_4	A			D	E	
P_5	A	B	C			$\mathcal{D}_2$
P_6		B	C	D	E	
P_7		B	C		E	
P_8		B			E	

2 data sets: $\mathcal{D}_1$ and $\mathcal{D}_2$

5 **items**: A, B, C, D, E (e.g., level of reached education, level of total cholesterol)

pattern: set of items

$\mathcal{F}$: **frequency**

$\mathcal{F}(ABC, \mathcal{D}_1) = 3$

$\mathcal{F}(ABC, \mathcal{D}_2) = 1$

Emerging patterns: definition

X: pattern *growth rate* of X from $\mathcal{D}_2$ to $\mathcal{D}_1$:

$$GR_1(X) = \frac{|\mathcal{D}_2| \times \mathcal{F}(X, \mathcal{D}_1)}{|\mathcal{D}_1| \times \mathcal{F}(X, \mathcal{D}_2)}$$

$\rho > 1$: X is a ρ-**EP** (or EP) from $\mathcal{D}_2$ to $\mathcal{D}_1$ if $GR_1(X) \geq \rho$

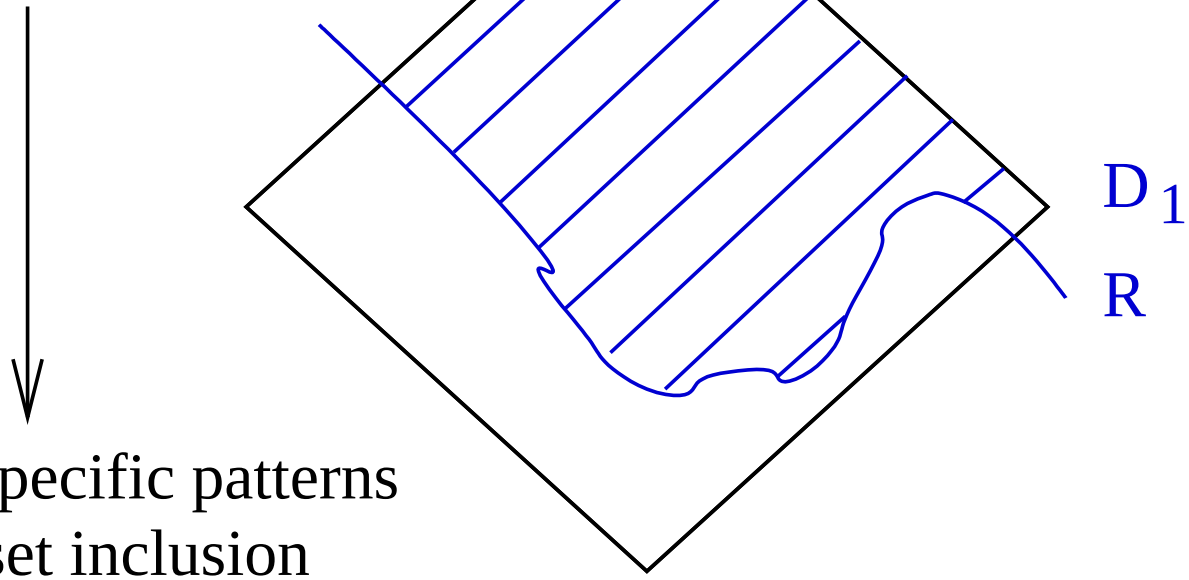
Running example: $\rho = 3$ from $\mathcal{D}_2$ to $\mathcal{D}_1$

- ABC is an EP ($GR_1(ABC) = 3$)
- BCD is not an EP ($GR_1(BCD) = 2$)
- $ABCD$ is a jumping emerging pattern ($GR_1(ABCD) = \infty$)

It is easy to generalize the definition to k data sets

Common approaches: use of borders

most general patterns
w.r.t. set inclusion



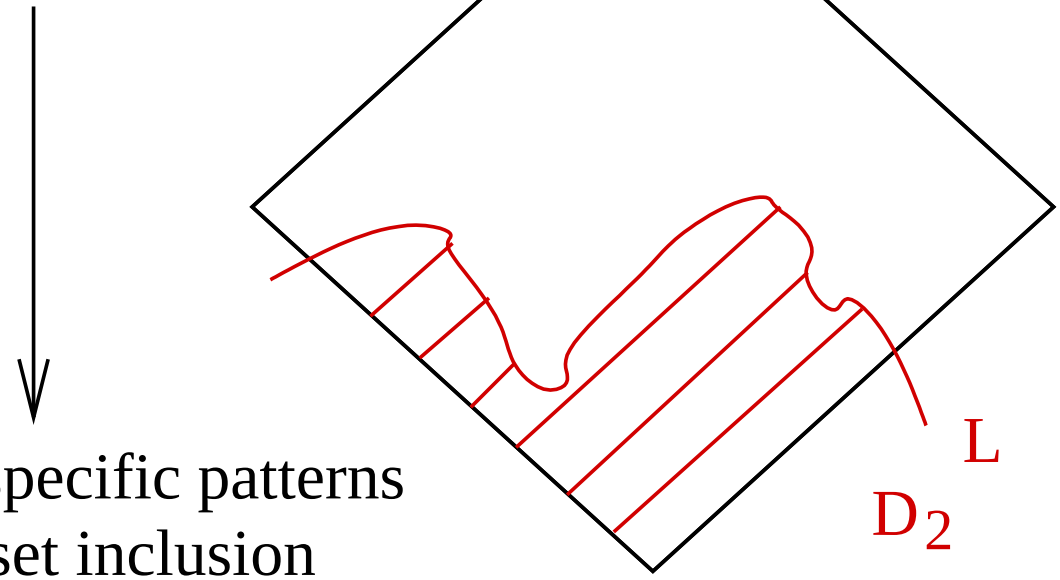
most specific patterns
w.r.t. set inclusion

γ: frequency threshold

$\mathcal{R}$: right-hand bound: maximal frequent patterns in $\mathcal{D}_1$ ($\mathcal{F}(X) \geq \gamma$)

Common approaches: use of borders

most general patterns
w.r.t. set inclusion



most specific patterns
w.r.t. set inclusion

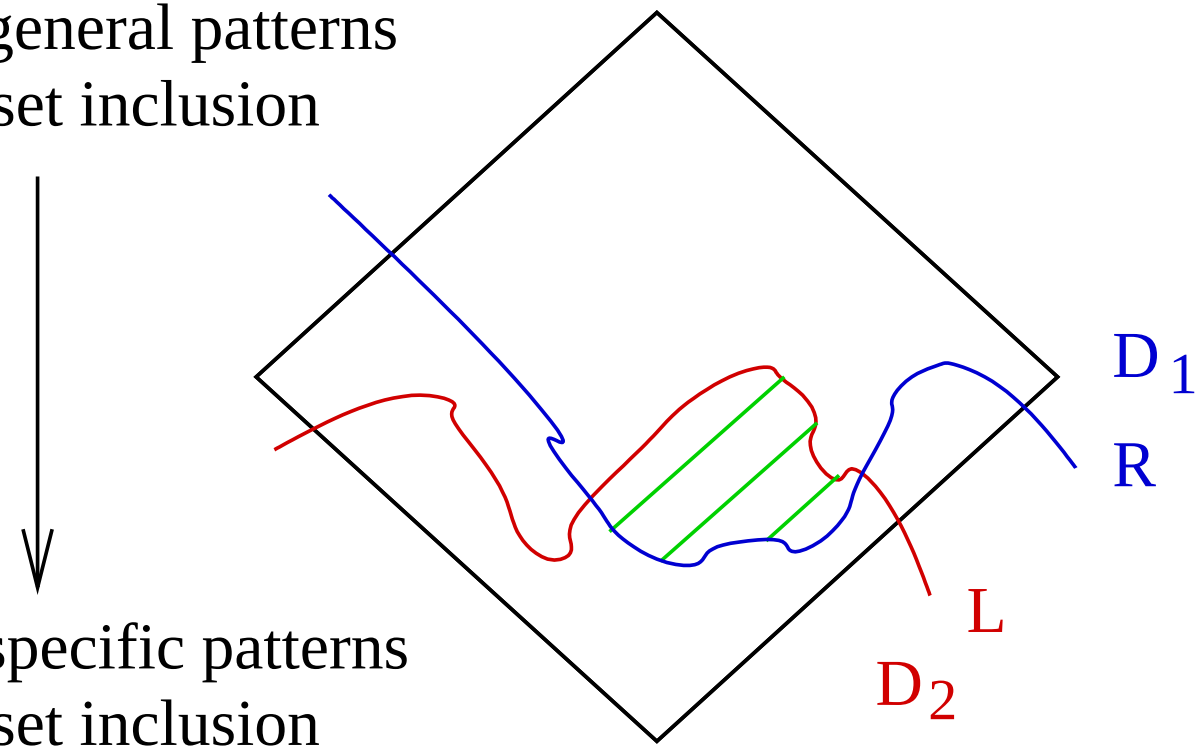
γ: frequency threshold

$\mathcal{R}$: right-hand bound: maximal frequent patterns in $\mathcal{D}_1$ ($\mathcal{F}(X) \geq \gamma$)

$\mathcal{L}$: left-hand bound: minimal infrequent patterns in $\mathcal{D}_2$ ($\mathcal{F}(X) \leq \gamma/\rho$)

Common approaches: use of borders

most general patterns
w.r.t. set inclusion



most specific patterns
w.r.t. set inclusion

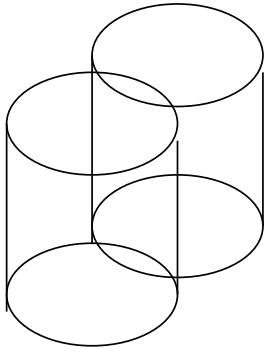
γ: frequency threshold

$\mathcal{R}$: right-hand bound: maximal frequent patterns in $\mathcal{D}_1$ ($\mathcal{F}(X) \geq \gamma$)

$\mathcal{L}$: left-hand bound: minimal infrequent patterns in $\mathcal{D}_2$ ($\mathcal{F}(X) \leq \gamma/\rho$)

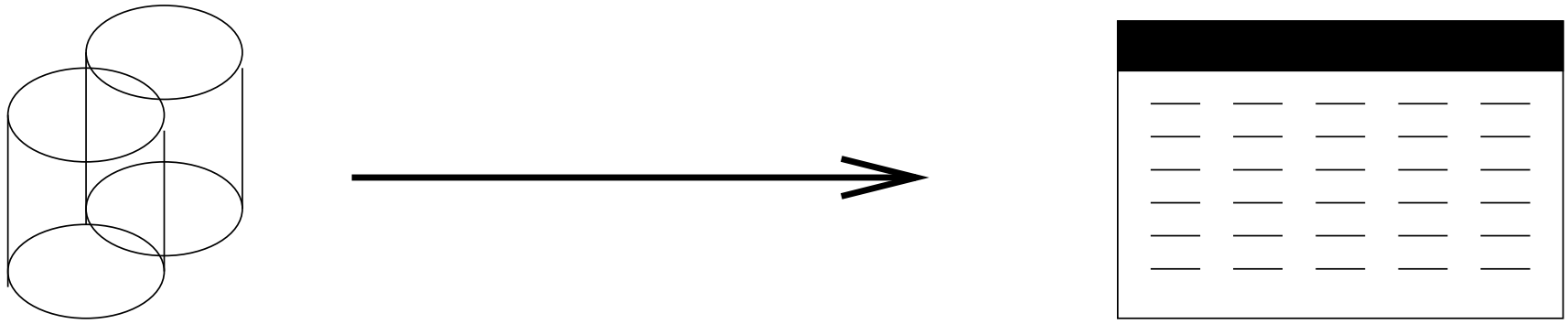
EPs are in $[\mathcal{L}, \mathcal{R}]$

Our proposition: use of condensed representations



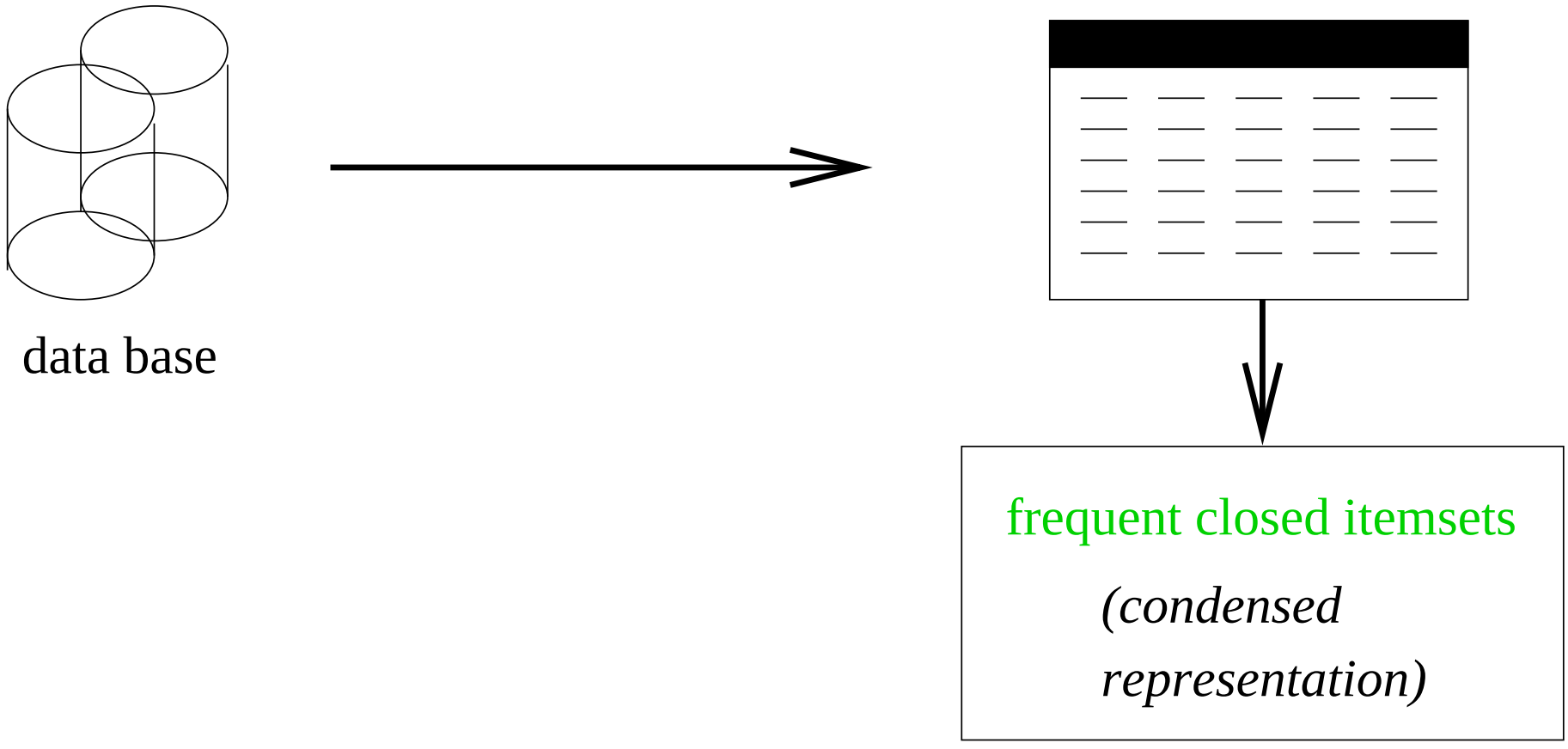
data base

Our proposition: use of condensed representations

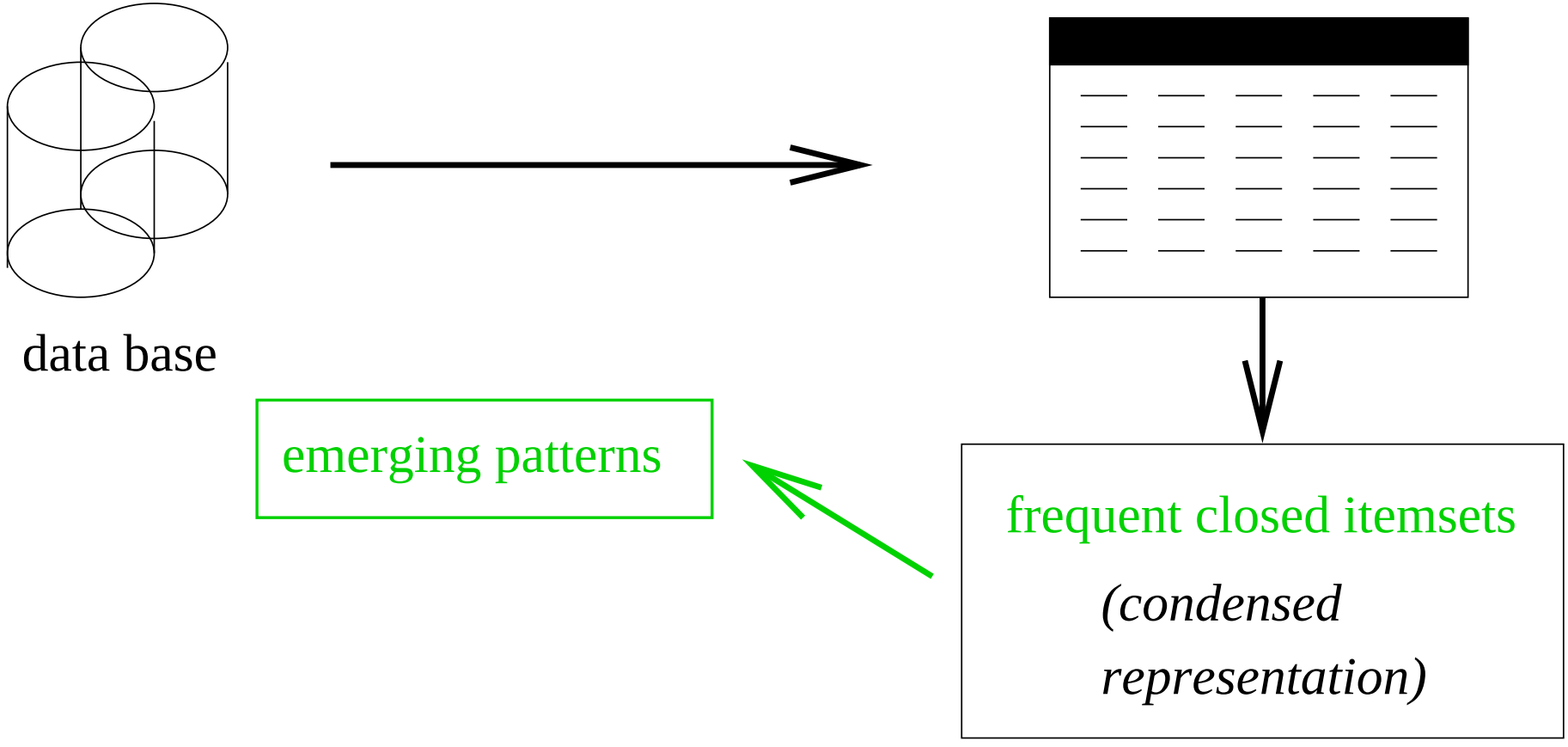


data base

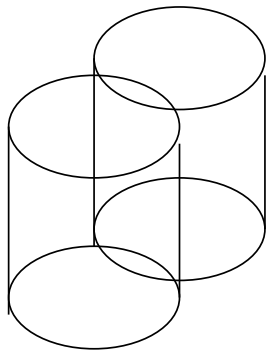
Our proposition: use of condensed representations



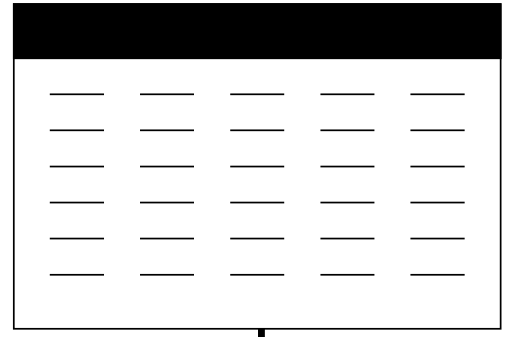
Our proposition: use of condensed representations



Our proposition: use of condensed representations



data base



frequent closed itemsets

(condensed representation)

emerging patterns

characterization of classes

clustering

classification

missing values

rules

...

Condensed representations based on closed patterns

closed patterns in a database $\mathcal{D}$:

➠ the maximal set of items shared by a set of examples.

Running example:

- ABC is a closed pattern on $\mathcal{D}_1$
- BC is not a closed pattern on $\mathcal{D}_1$ (A is always present when BC is present)

$h(X, \mathcal{D})$: **closure** of X in $\mathcal{D}$. It is a *closed pattern* with the same frequency of X: $\mathcal{F}(X, \mathcal{D}) = \mathcal{F}(h(X, \mathcal{D}), \mathcal{D})$

Running example: $h(BC, \mathcal{D}) = ABC$

There are efficient algorithms to extract all frequent closed patterns

Strong emerging patterns

$\mathcal{D} = \bigcup_i \mathcal{D}_i$ (k classes)

strong emerging pattern (SEP) from $\mathcal{D} \backslash \mathcal{D}_i$ to $\mathcal{D}_i$

➠ EP based on a closed frequent pattern in $\mathcal{D}_i$

SEPs have two main advantages:

- their growth rate is directly obtained from the condensed representation of closed patterns in $\mathcal{D}$
- SEPs have the highest growth rates:

$$GR_i(X) \leq GR_i(h(X, \mathcal{D}_i))$$

Running example:

- BC is not a closed pattern in $\mathcal{D}_1$: $GR_1(BC) = 1$
- $h(BC, \mathcal{D}_1) = ABC$: $GR_1(ABC) = 3$

Aim of experiments

from the features available when patients are entered in the study (i.e. features of the table `Entry`)

- `Experiment 1`: distinguish the patients *who died* from atherosclerosis from the others
- `Experiment 2`: distinguish patients *who came down* with the observed cardiovascular diseases from those who stayed healthy

long-term information needed:

➥ focus on patients from `normal studied group`, `intervened risk group` and `control risk group` (899 patients).

Experiment 1: data preparation

assumption: all patients dying from atherosclerosis are recorded in the table `Death`

Table `Death`: 165 patients died due to atherosclerosis:

➠ values `myocardial infarction`, `coronary heart disease`, `stroke` and `general atherosclerosis` of the attribute `PRICUMAR`

join these patients with patients followed during all the study:

➥ 124 patients remain

Experiment 2: data preparation

assumption: all patients suffering from atherosclerosis are recorded in the table `Control`

remark: these patients may suffer from another disease

From the web pages:

patient affected by a cardiovascular disease:

➠ illness based on attributes `HODN1`, `HODN2`, `HODN3`, `HODN11`, `HODN12`, `HODN13`, `HODN14`, `HODN21`, `HODN23`

join these patients with patients followed during all the study:

➥ 281 patients remain

Data preparation (cont'd)

keep all attributes of the table `Entry` except the deletion of:

- the attribute KONSUP: may introduce a bias
- the attributes relating to risk factors: already taken into account by other attributes
- attributes relating to the personal anamnesis: very low frequencies of values
- the attributes ROKNAR (year of birth) and ROKVSTUP (year of entry into the study)

  ➥ replaced by the age of the patient when he was entered in the study.

Data preparation (cont'd)

- binary segmentation of the attributes `CHLST` (cholesterol) and `TRIGL` (triglycerides) according to the thresholds given in the web pages
- segmentation in three values of each other continuous attribute

➥ finally: 119 items.

Characteristics of the files and results

- `athero`: patients who died (`Experiment 1`) or ill (`Experiment 2`)
- `healthy`: other patients.

15: absolute frequency SEPs with at most 8 items

	Experiment 1		Experiment 2	
	athero	healthy	athero	healthy
No. of patients	124	624	281	618
γ (%)	12.1%	2.4%	5.3%	2.4%
$GR \in [1..2[$	32606	2,278,346	510,901	2,845,756
$GR \in [2..5[$	6254	1,229,359	69609	605,312
$GR \in [5..\infty[$	47	94,921	1038	61168
JEP	132	387,203	2690	16916

Results:
Experiment 1 (atherosclerosis)

items of SEPs	GR	$\mathcal{F}$ (%)
the way to work takes around 1 hour ; smoker of 21 and more cigarettes per day ; smoking during 21 and more years ; do not drink liquors	6.71	11.3
weight $\leq$ 74 kg ; blood pressure II diastolic > 92 mm Hg ; normal urine	6.29	11.3
height $\leq$ 1.72 m ; blood pressure II diastolic > 92 mm Hg	3.91	16.9
blood pressure II diastolic > 92 mm Hg	1.72	32.3
age of entry in the study $\in$ [43,47] ; moderate activity after his job ; level of total cholesterol $\geq$ 200 mg/dL	∞	18.5

Results: Experiment 1 (healthy)

items of SEPs	GR	$\mathcal{F}$ (%)
partly independent worker ; blood pressure II systolic $\leq$ 118 mm Hg	11.7	9.46
reached education: university ; level of total cholesterol < 200 mg/dL	8.35	6.7
age of entry in the study $\in$ [44,47] ; level of total cholesterol < 200 mg/dL	8.15	6.6
age of entry in the study $\leq$ 43 years	2.21	30.3
reached education: university ; blood pressure II diastolic $\leq$ 78 mm Hg	∞	8.7
age of entry in the study $\leq$ 43 years ; mainly standing at work	∞	5.0
non-smoker ; blood pressure I systolic $\leq$ 120 mm Hg	∞	4.5

Results:
Experiment 2 (atherosclerosis)

items of SEPs	GR	$\mathcal{F}$ (%)
1 or 2 cups of coffee per day ; height < 1.72 m ; blood pressure I diastolic $\in$ [75,92] ; skinfold above musculus triceps > 11	7.48	6.0
more than 6 sugar lumps per day ; skinfold above musculus triceps > 11	2.30	8.2
height $\leq$ 1.72 m ; blood pressure I systolic > 135	2.00	14.6
drinking of alcohol: occasionally ; drinking of vine ; up to half a litre of vine per day ; level of triglycerides > 150 mg/dL	∞	14.2
reached education: secondary school ; drinking of vine ; up to half a litre of vine per day ; blood pressure II diastolic $\in$ [78,92]	∞	12.5

Results: Experiment 2 (healthy)

items of SEPs	GR	$\mathcal{F}$ (%)
single ; do not drink coffee	8.64	3.1
lower limbs pain is non-ischaemic ; blood pressure I diastolic $\leq$ 75 mm Hg	8.64	3.1
mainly walks at work ; drink daily more than 1 litre of beer	5.12	7.3
partly independent worker ; blood pressure I systolic $\leq$ 120 mm Hg ; blood pressure I diastolic $\leq$ 75 mm Hg	5	7.1
drinking of 10° beer ; daily consumption of 2 at 6 sugar lumps ; blood pressure I diastolic $\leq$ 75 mm Hg ; normal urine	∞	7.3
blood pressure II systolic $\in$ [118,138] ; blood pressure II diastolic > 92 mm Hg	∞	3.8

Results (cont'd)

- `Experiment 1`: a lot of SEPs to `atherosclerosis` have the item "smoking during 21 and more years", the blood pressure seems to have an important role.
- `Experiment 2`: all JEPs to `atherosclerosis` have at least 4 items.

It is likely that most associations highlighting by SEPs are expected (and already known) by physicians:

➥ but SEPs quantify such associations (e.g., how much increases the risk of atherosclerosis with respect to precise features?)

Conclusion

- *strong emerging patterns* (best growth rate, efficient method to extract them from condensed representation)
- search of SEPs characterizing patients with respect to atherosclerosis: it allows to quantify associations highlighted by SEPs
- *further work:*
  - how to select the most interesting SEPs (or EPs) from a data set?
  - build condensed representations of EPs including the patterns and their growth rates.