

Tijl De Bie

Akis Kontonasios

Eirini Spyropoulou

A framework for mining interesting pattern sets



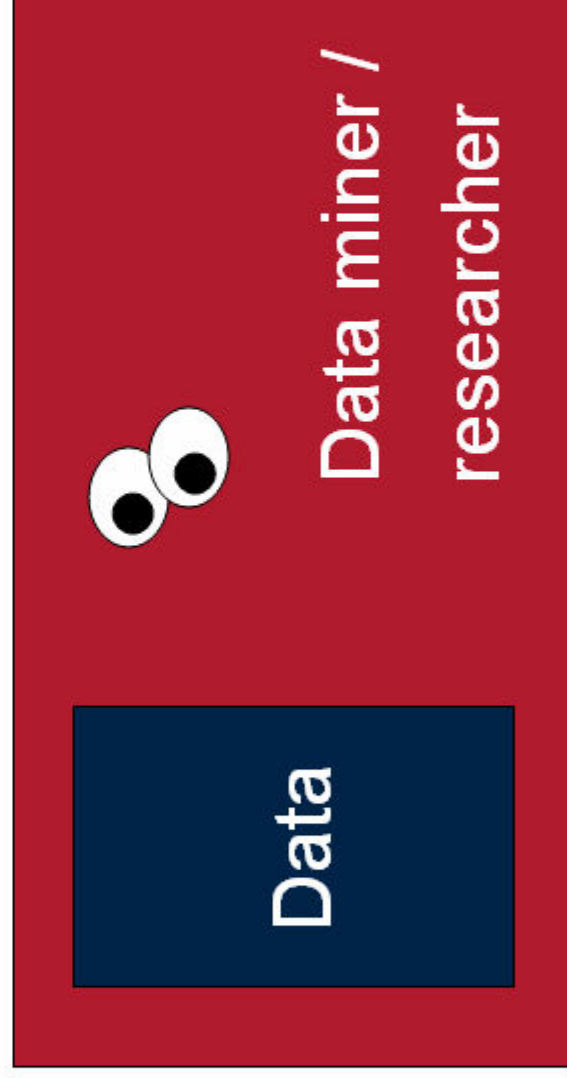
A zoo of interestingness measures

- Survey (Geng and Hamilton, 2006):
38 different probabilistically inspired interestingness measures for association rules / itemsets
- Often not relevant
- Often redundant

- Support
- Confidence
- Coverage
- Recall
- Specificity
- Accuracy
- Lift
- Leverage
- ...
- Support x size (tiles)
- MDL-based (KRIMP)
- Comparison to subsets (Tatti)

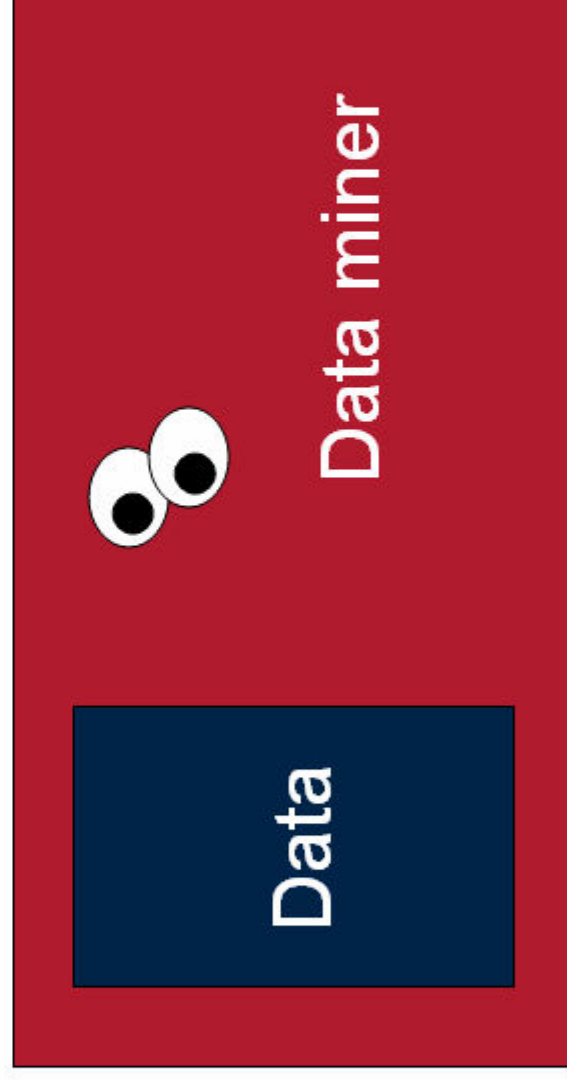
A zoo of interestingness measures

- Exploratory data mining = intentionally ill-defined?
 - Each interestingness measure has benefits
 - Choice can be task dependent
- All the above: ‘**objective**’ measures



Interesting to whom?

‘Objective interestingness measure’ = a mirage

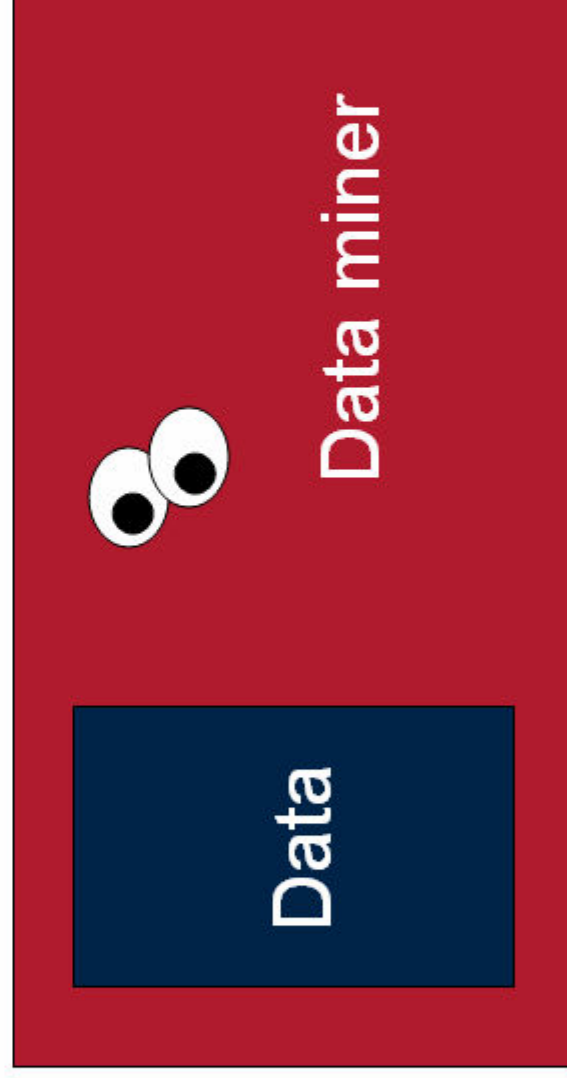


Interesting to whom?

‘Objective interestingness measure’ = a mirage

Subjective interestingness: interesting to a subject (the data miner)

(Tuzhilin and Silberschatz, 1996)

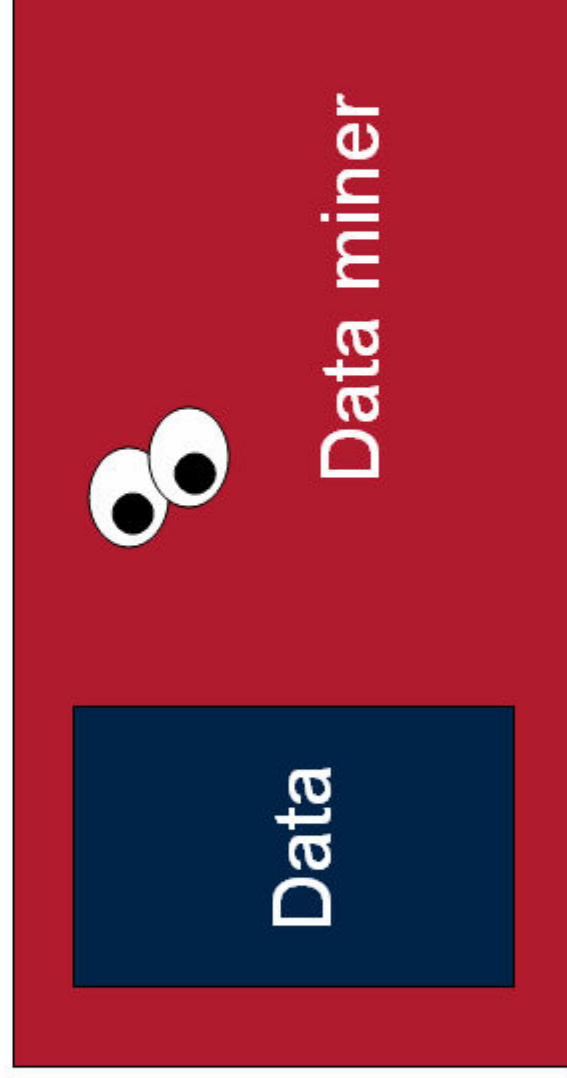


Interesting to whom?

‘Objective interestingness measure’ = **a mirage**

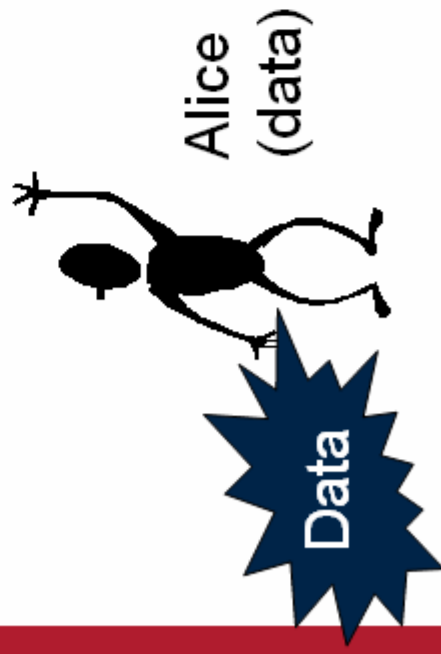
Subjective interestingness: **interesting to a subject** (the data miner)

(Tuzhilin and Silberschatz, 1996)

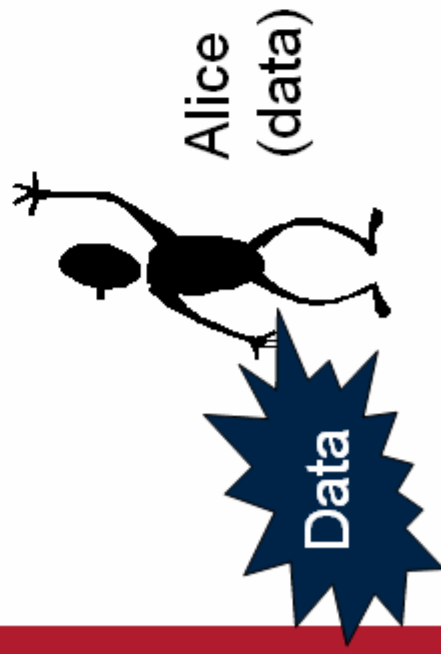


Data mining
researcher

Data mining as 2-way communication

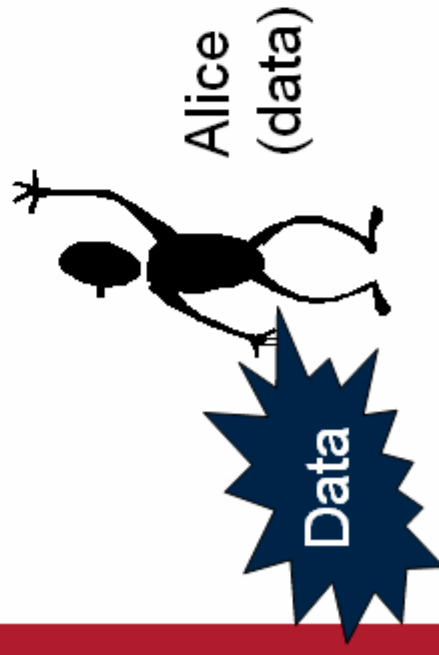


Data mining as 2-way communication



Data mining as 2-way communication

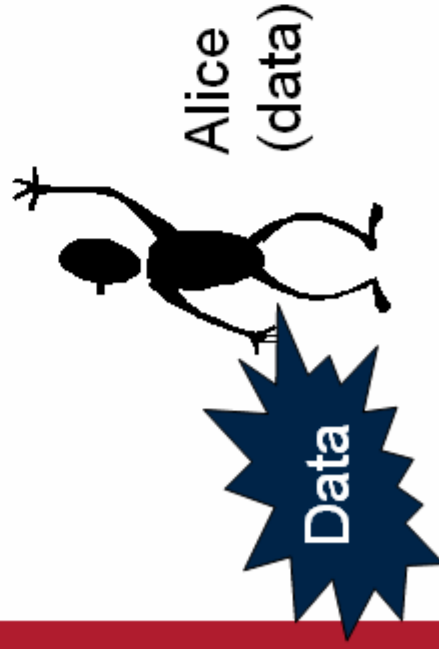
Tell me what you know about the data.



Data mining as 2-way communication

Tell me what you know about the data.

I have **expectations** about certain aspects of the data.

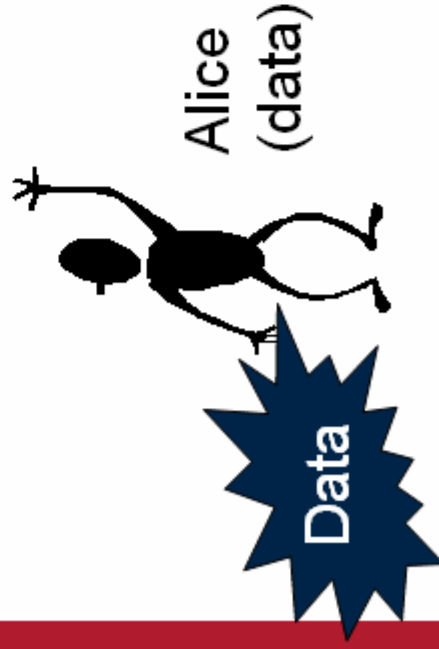


Data mining as 2-way communication

Tell me what you know about the data.

I have **expectations** about certain aspects of the data.

Great. I will compress the database using an optimal Shannon code with respect to the **a background distribution** reflecting these expectations.



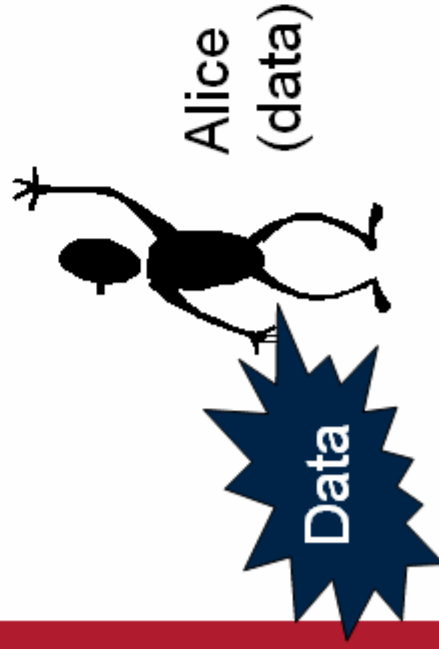
Data mining as 2-way communication

Tell me what you know about the data.

Great. I will compress the database using an optimal Shannon code with respect to the **a background distribution** reflecting these expectations.

I **have expectations** about certain aspects of the data.

Wait, I suspect certain types of pattern in the data,
 → First send me any **interesting patterns** of that kind present in the data.

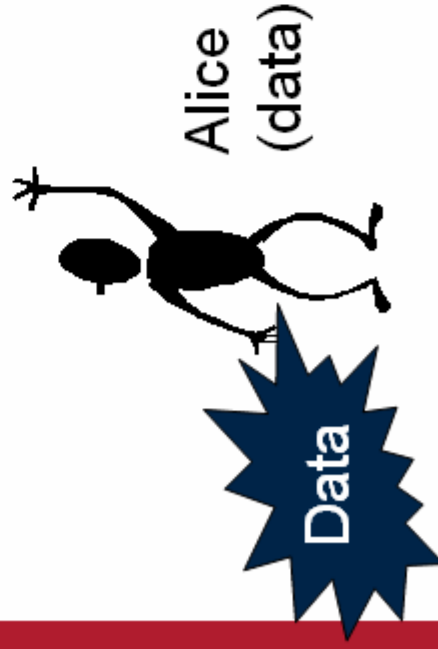


Data mining as 2-way communication

Tell me what you know about the data.

Great. I will compress the database using an optimal Shannon code with respect to the **a background distribution** reflecting these expectations.

Deal, this allows me to further **compress** the overall description!



I **have expectations** about certain aspects of the data.

Wait, I suspect certain types of pattern in the data,
 → First send me any **interesting patterns** of that kind present in the data.



Data mining as 2-way communication

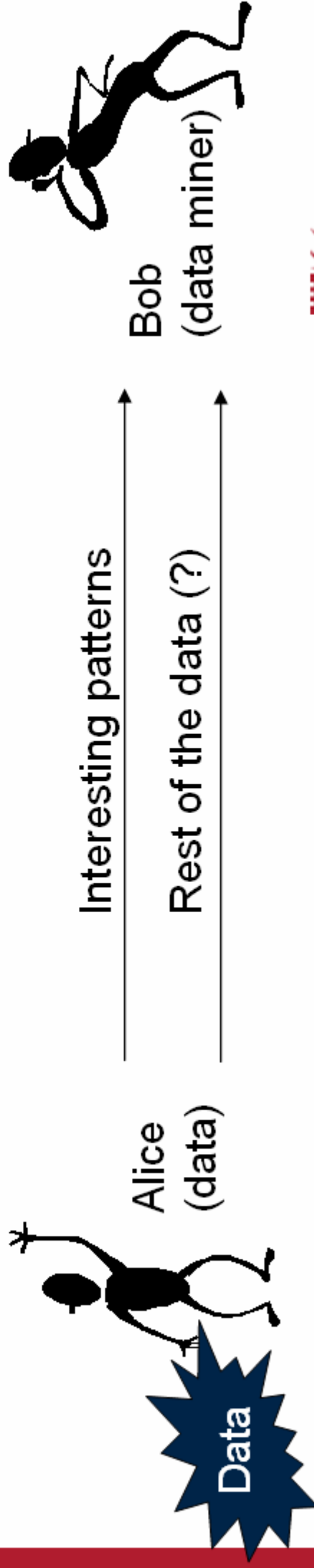
Tell me what you know about the data.

Great. I will compress the database using an optimal Shannon code with respect to the **a background distribution** reflecting these expectations.

Deal, this allows me to further **compress** the overall description!

I **have expectations** about certain aspects of the data.

Wait, I suspect certain types of pattern in the data,
 → First send me any **interesting patterns** of that kind present in the data.



Components of a framework

Formalize
prior expectations

Probabilistic background model

Prior expectations as constraints

Components of a framework

Formalize
prior expectations

Probabilistic background model

Prior expectations as constraints

Choose
pattern syntax

Chosen by data miner

Task dependent

Components of a framework

Formalize
prior expectations

Probabilistic background model

Prior expectations as constraints



Self-information



Description length (complexity)

Contrast patterns
with prior expectations

Choose
pattern syntax

Chosen by data miner

Task dependent



Case study

- Binary database
- Prior expectations on row and column sums
- Tiles as patterns

Case study – Background model

Formalize
prior expectations

Probabilistic background model

Prior expectations as constraints



Self-information



Description length (complexity)



Task dependent

Chosen by data miner

Choose
pattern syntax

Contrast patterns
with prior expectations

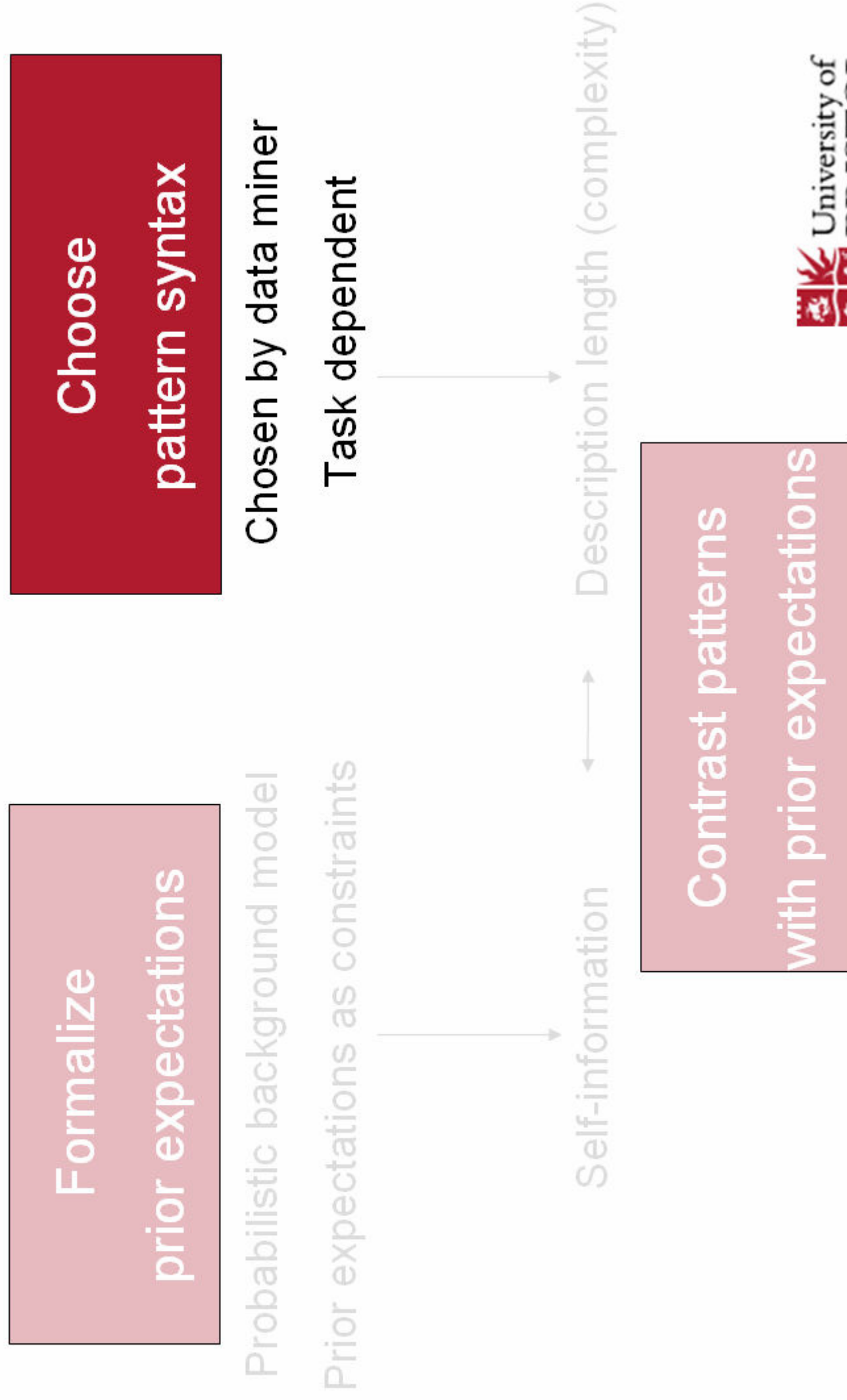
Case study – Background model

- Binary database \mathbf{D}
- Expectations about row and column sums
→ constraints on the background model
- Pick **Maximum Entropy** (MaxEnt) distribution
- Product of independent Bernoulli distributions:

$$P(\mathbf{D}) = \prod_{i,j} P_{ij}(\mathbf{D}(i, j))$$

$$P_{ij}(\mathbf{D}(i, j)) = \begin{cases} P_{ij} & \text{if } \mathbf{D}(i, j) = 1 \\ 1 - P_{ij} & \text{if } \mathbf{D}(i, j) = 0 \end{cases}$$

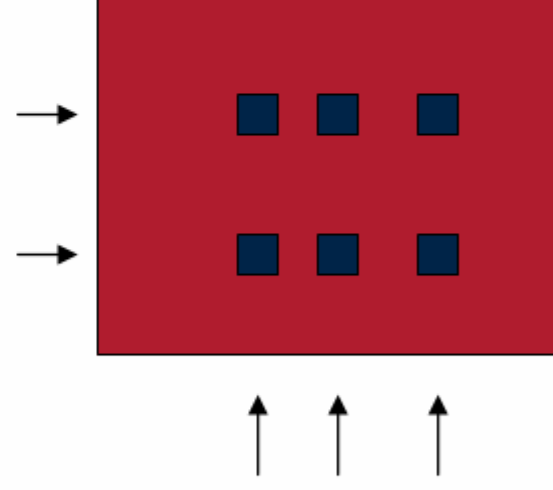
Case study – Tiles as patterns



Case study – Tiles as patterns

- Pattern syntax
- E.g. tiles (Geerts, Goethals, Mielikäinen, 2004), i.e. pairs of:
 - row set
 - column set

such that there are 1's in the intersection



Case study – Which tiles are interesting?

Formalize
prior expectations

Probabilistic background model

Prior expectations as constraints



Self-information



Description length (complexity)



Contrast patterns
with prior expectations

Choose
pattern syntax

Chosen by data miner

Task dependent

Case study – Which tiles are interesting?

- Self-information (SI) of a tile pattern
 - The negative log-probability under the background model
 - Smaller probability
= more informative w.r.t. prior expectations
- Description length (DL) of a tile pattern
 - Perceived complexity of a tile
 - E.g. an affine function of the tile's circumference
- $SI / DL = \text{'information density'}$

Case study – Interesting sets of tiles

- Data miner has limited processing capacity
 - Bounds the total description length of tile patterns
- Find the pattern set that:
 - Has **maximal total self-information**
 - Subject to an **upper bound on the total description length**
- Maximally compresses the data
- Weighted budgeted set coverage problem
 - Hard to solve exactly, but greedy is a good approximation

Case study – Results

Rows = KDD abstracts

Columns = stemmed words

KDD

Mining interesting pattern sets (current paper)	$ I $	Tiling databases as described in [7]	$ I $
machin support svm vector	25	data paper	389
art state	39	algorithm propos	246
labeled learn supervised unlabeled	10	data mine	312
associ mine rule	36	base method	202
express gene	25	result show	196
frequent itemset	28	problem	373
graph larg network social	15	data set	279
column row	13	approach	330
algorithm faster magnitud order	12	model	301
algorithm data paper propos real synthetic	27	present	296
answer question	18	larg	286
nearest neighbor	13	applic	271
classifi featur machin support text vector	9	perform	266
precis recal	14	real	255
decis tree	33	inform	240

[7] Geerts, Goethals, Mielikäinen, 2004

Final thoughts and conclusions

- Current focus
 - Other constraints of the form $\sum_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}) = d$
 - Other data types (networks, relational databases)
 - Other pattern types (syntax)
 - Other contrast measures
- **Open problem:** the empirical results trap...
 - Risk of missing the target, we should not test:
 - number of patterns,
 - speed,
 - classification accuracy...
 unless they are actually the goal
 - We tend to use text (but...)