Modelling Real Valued Data by Maximum Entropy

incorporating expectations on arbitrary sets of cells

Akis Kontonasios, **Jilles Vreeken** & Tijl De Bie

University of Bristol
FWO Vlaanderen
Universiteit Antwerpen

Identifying Interesting Patterns in Real-Valued Data

through iterative Maximum Entropy modelling

Akis Kontonasios, **Jilles Vreeken** & Tijl De Bie

University of BRISTOL
FWO VLAANDEREN
Universiteit Antwerpen

Maximum Entropy Models for Iteratively Identifying Subjectively Interesting Structure in Real Valued Data

Akis Kontonasios, **Jilles Vreeken** & Tijl De Bie

University of BRISTOL

FWO VLAANDEREN

Universiteit Antwerpen

Question at hand

Given a data mining result, how **interesting** is it *with regard to what we already know*?

What is interesting?

something that
increases our knowledge
about the data

What is good?

something that
reduces our uncertainty
about the data

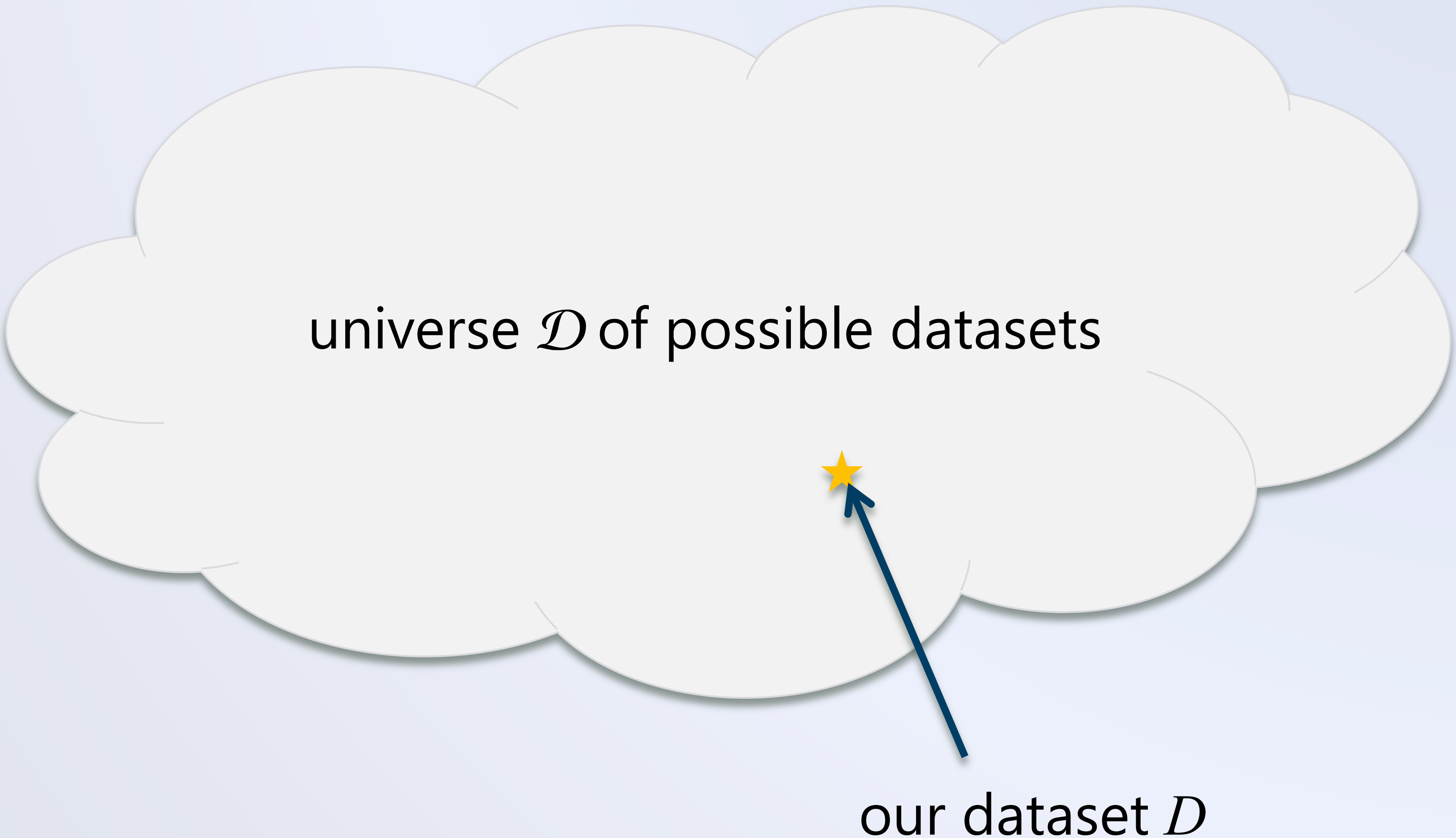
(ie. increases the likelihood of the data)

What is really good?

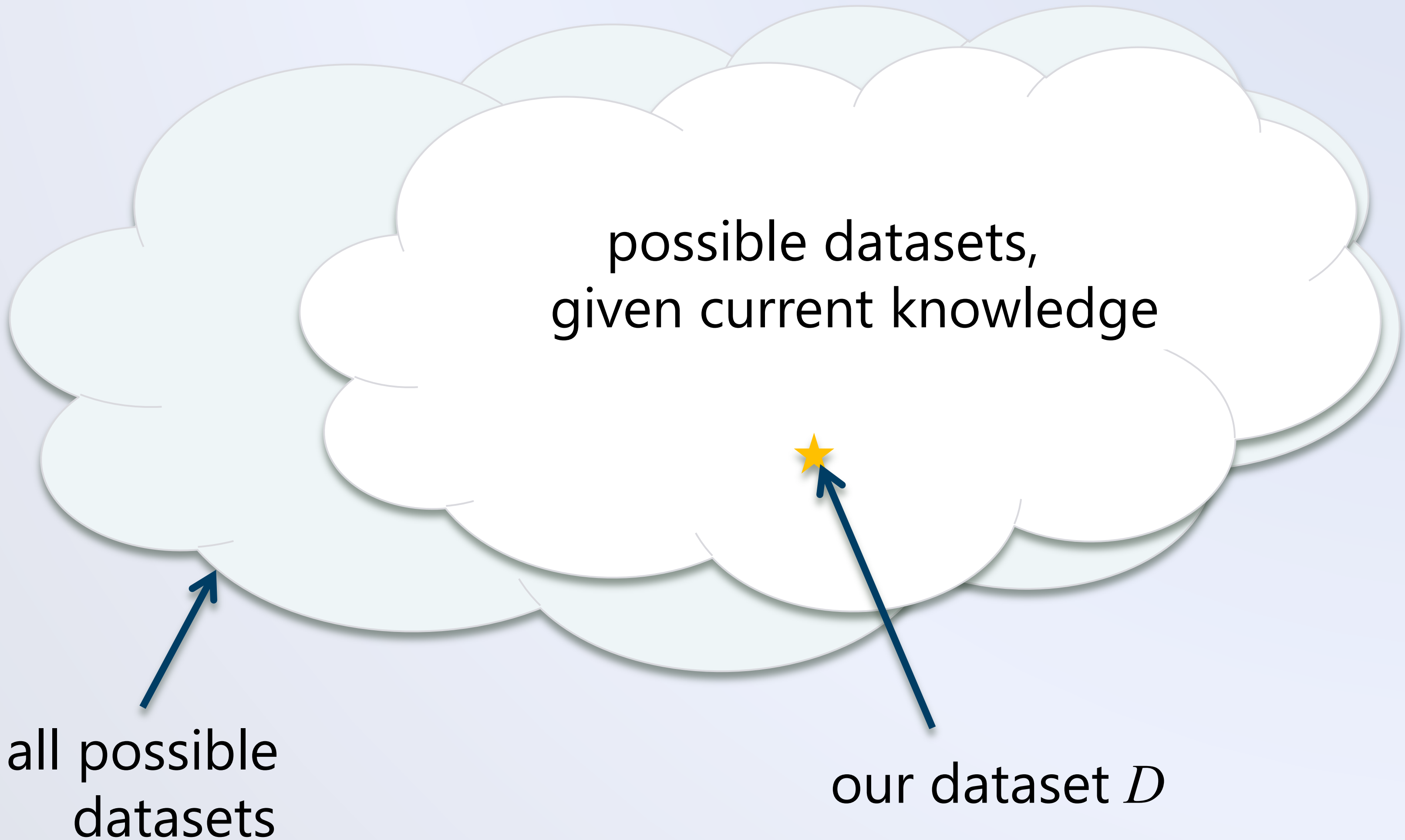
something that,
in **simple terms**,
strongly reduces our uncertainty
about the data

(maximise likelihood, but avoid overfitting)

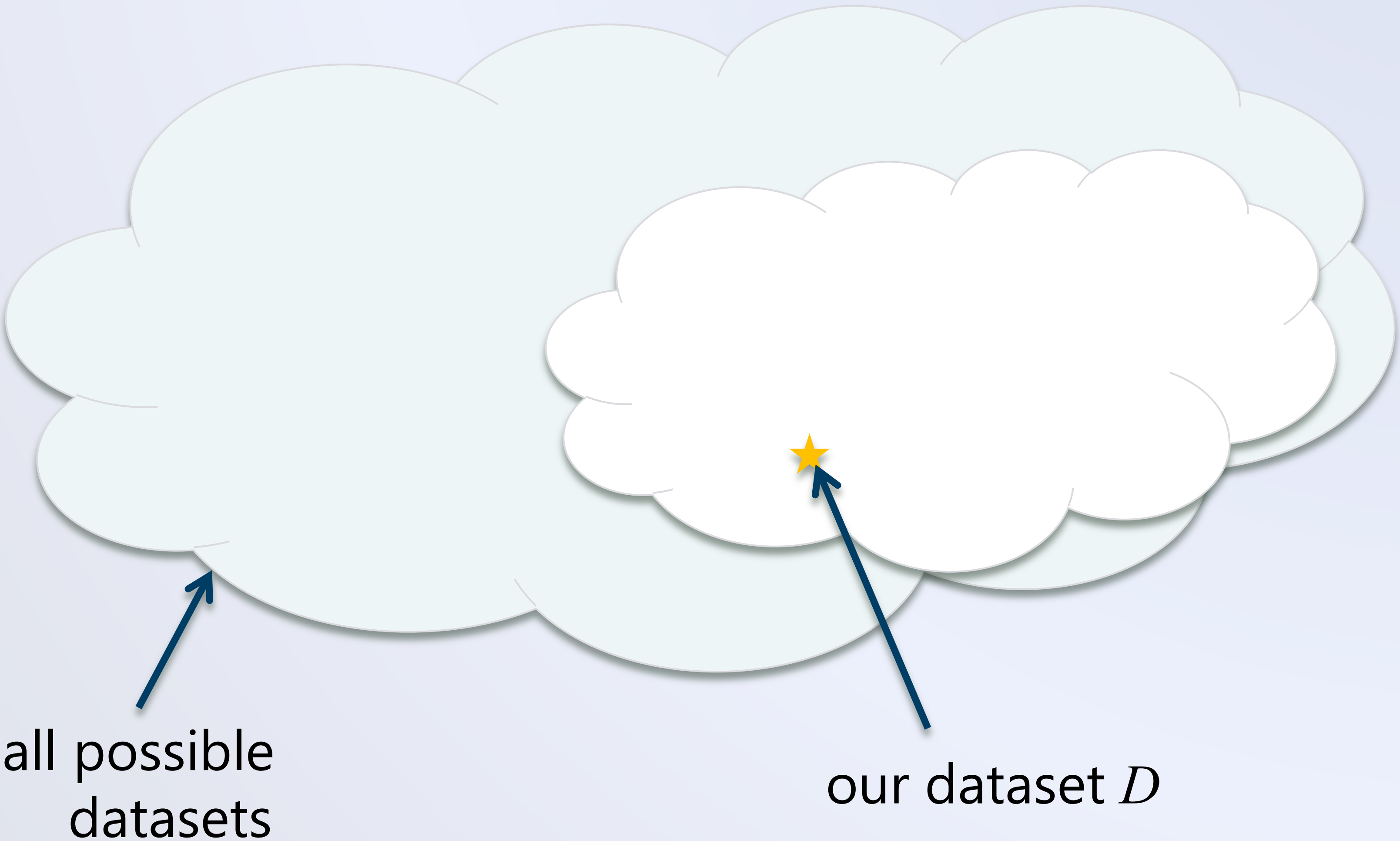
Let's make this visual



Given what we know

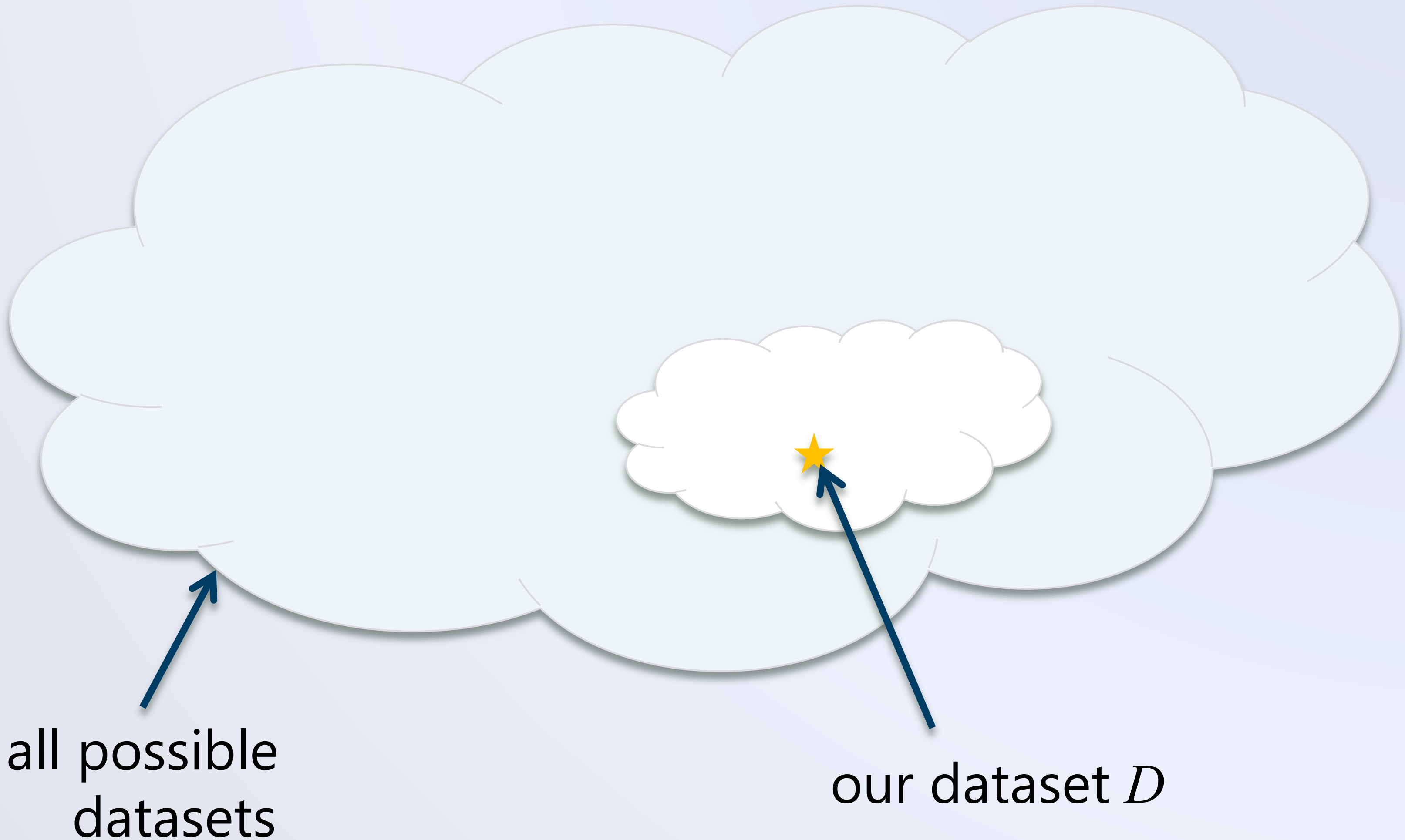


More knowledge…



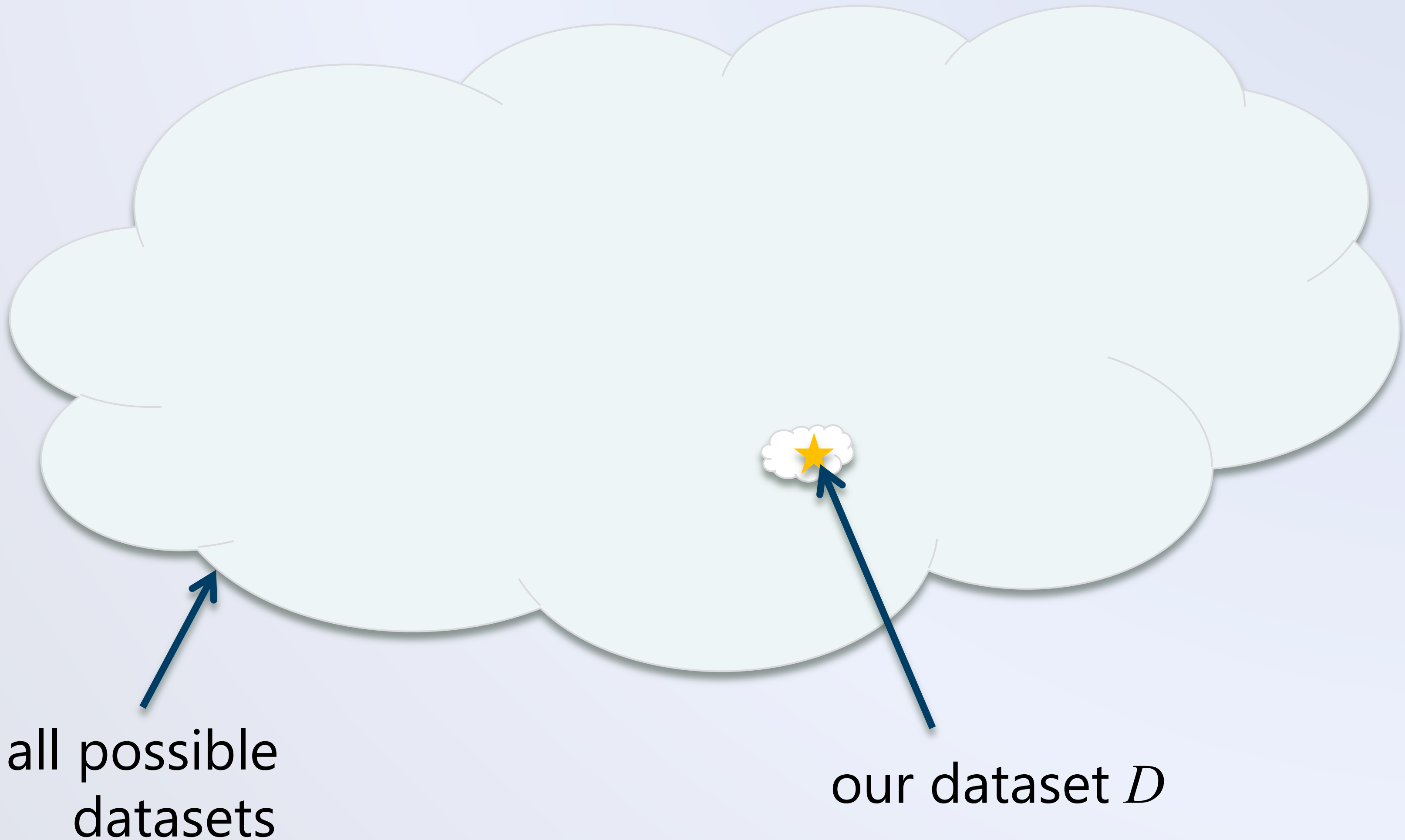
dimensions, margins, patterns P_1 and P_2

Fewer possibilities…



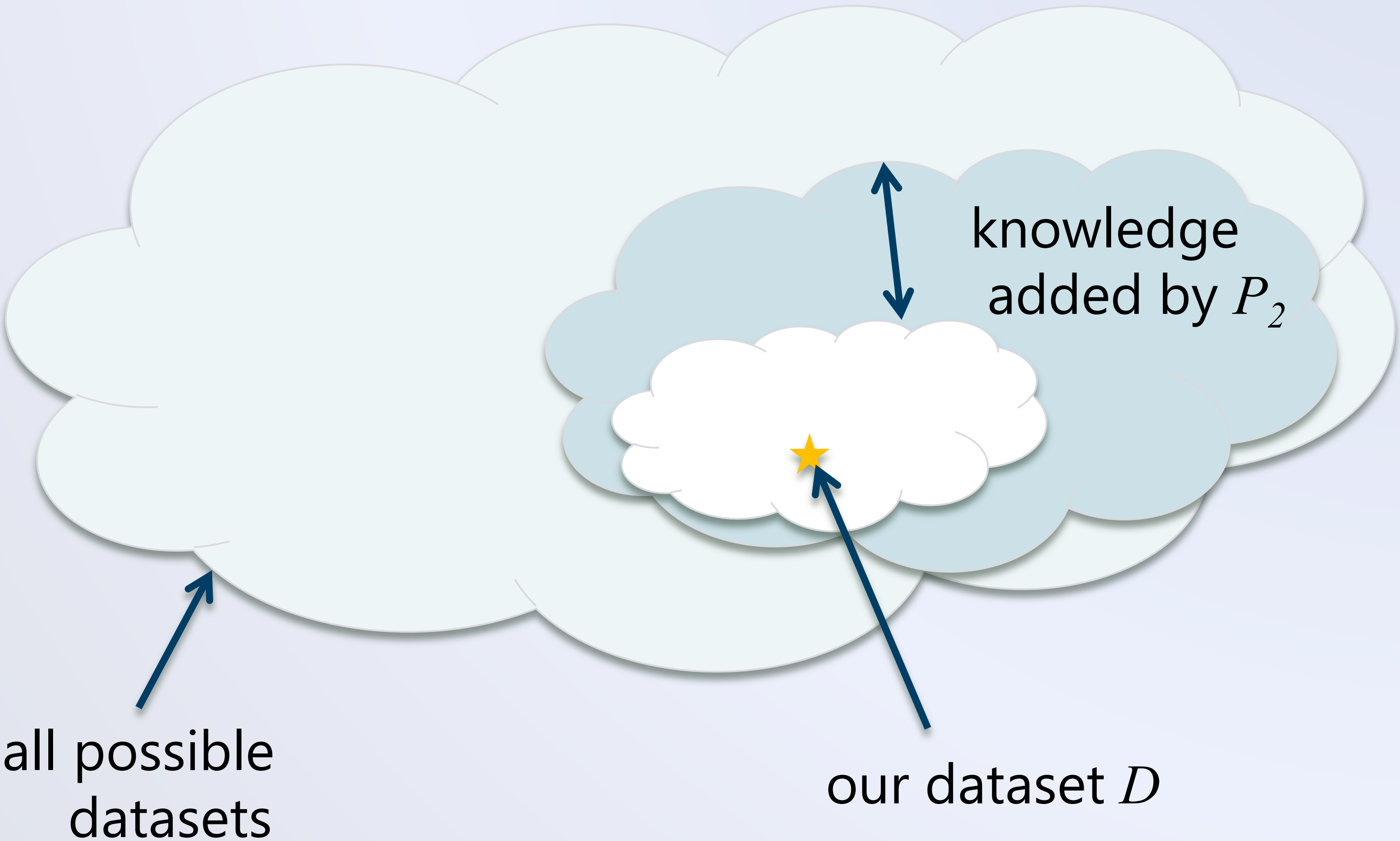
dimensions, margins,
the key patterns

Less uncertainty.



Maximising certainty

dimensions, margins, patterns P_1 and P_2



(iterative data mining, Hanhijärvi et al. 2009)

How can we define

‘uncertainty’ and ‘simplicity’?

interpretability and **informativeness**
are intrinsically **subjective**

Measuring Uncertainty

We need access to the **likelihood**
of data D given background knowledge B

$$P(D \mid B)$$

such that we can calculate the **gain** for X

$$P(D \mid B \cup X) - P(D \mid B)$$

...which distribution should we use?

Maximum Entropy principle

'the best distribution p^* satisfies the background knowledge, but makes **no further** assumptions'

very useful for data mining:

unbiased measurement of **subjective interestingness**

(Jaynes 1957; De Bie 2009)

MaxEnt Theory

To use MaxEnt, we need **theory** for modelling data given background knowledge

Binary Data

- margins (De Bie, '09)
- tiles (Tatti & Vreeken, '12)

Real-valued Data

- margins (Kontonasios et al. '11)
- arbitrary sets of cells (**now**)

MaxEnt Theory

To use MaxEnt, we need **theory** for modelling data given background knowledge

Binary Data

- margins (De Bie, '09)
- tiles (Tatti & Vreeken, '12)

Real-valued Data

- margins (Kontonasios et al. '11)
- arbitrary sets of cells **(now)**

allow for **iterative** mining

MaxEnt for Real-Valued Data

Our model can incorporate

means, variance, and higher order moments,
as well as **histogram** information

over **arbitrary** sets of cells

MaxEnt for Real-Valued Data

,9	,8	,7	,4	,5	,5	,5
,7	,8	,9	,3	,5	,3	,5
,8	,8	,8	,6	,3	,4	,2
,7	,9	,7	,7	,3	,2	,5
,2	,8	,7	,8	,4	,4	,1
,3	,6	,9	,8	,3	,8	,3
,2	,1	,3	,4	,5	,3	,2

MaxEnt for Real-Valued Data

,9	,8	,7	,4	,5	,5	,5
,7	,8	,9	,3	,5	,3	,5
,8	,8	,8	,6	,3	,4	,2
,7	,9	,7	,7	,3	,2	,5
,2	,8	,7	,8	,4	,4	,1
,3	,6	,9	,8	,3	,8	,3
,2	,1	,3	,4	,5	,3	,2

Pattern 1

- {1-3}x{1-4}
- mean 0.8

Pattern 2

- {2,3} x {3-5}
- mean 0.8

Pattern 3

- {5-7} x {3-5}
- mean 0.3

MaxEnt for Real-Valued Data

,9	,8	,7	,4	,5	,5	,5		,6
,7	,8	,9	,3	,5	,3	,5		,6
,8	,8	,8	,6	,3	,4	,2		,6
,7	,9	,7	,7	,3	,2	,5		,6
,2	,8	,7	,8	,4	,4	,1		,5
,3	,6	,9	,8	,3	,8	,3		,6
,2	,1	,3	,4	,5	,3	,2		,3
,5	,7	,7	,6	,4	,4	,3		,5

Pattern 1

- {1-3}x{1-4}
- mean 0.8

Pattern 2

- {2,3} x {3-5}
- mean 0.8

Pattern 3

- {5-7} x {3-5}
- mean 0.3

MaxEnt for Real-Valued Data

,5	,5	,5	,5	,5	,5	,5		
,5	,5	,5	,5	,5	,5	,5		
,5	,5	,5	,5	,5	,5	,5		
,5	,5	,5	,5	,5	,5	,5		
,5	,5	,5	,5	,5	,5	,5		
,5	,5	,5	,5	,5	,5	,5		
,5	,5	,5	,5	,5	,5	,5		
								,5

Pattern 1

- {1-3}x{1-4}
- mean 0.8

Pattern 2

- {2,3} x {3-5}
- mean 0.8

Pattern 3

- {5-7} x {3-5}
- mean 0.3

MaxEnt for Real-Valued Data

,6	,7	,7	,7	,5	,6	,4		,6
,7	,6	,6	,6	,4	,4	,6		,6
,6	,7	,7	,6	,5	,5	,3		,6
,6	,6	,7	,6	,5	,4	,5		,6
,5	,7	,6	,6	,5	,4	,3		,5
,5	,7	,7	,6	,5	,6	,3		,6
,3	,6	,6	,3	,2	,2	,2		,3
,5	,7	,7	,6	,4	,4	,4		,5

Pattern 1

- {1-3}x{1-4}
- mean 0.8

Pattern 2

- {2,3} x {3-5}
- mean 0.8

Pattern 3

- {5-7} x {3-5}
- mean 0.3

(Kontonasios et al., 2011)

MaxEnt for Real-Valued Data

,8	,8	,8	,6	,4	,4	,4		,6
,8	,8	,8	,6	,4	,4	,4		,6
,8	,8	,8	,6	,4	,4	,4		,6
,8	,8	,8	,6	,4	,4	,4		,6
,2	,6	,6	,6	,4	,5	,4		,5
,3	,6	,6	,6	,6	,6	,6		,6
,1	,3	,3	,3	,4	,4	,3		,3
,5	,7	,7	,6	,4	,4	,4		,5

Pattern 1

- **{1-3}x{1-4}**
- **mean 0.8**

Pattern 2

- {2,3} x {3-5}
- mean 0.8

Pattern 3

- {5-7} x {3-5}
- mean 0.3

(this paper)

Simplicity?

Likelihood alone is insufficient

does not take size, or complexity into account

as practical example of our model:

Information Ratio

for tiles in real valued data

Information Ratio

$$\frac{Information\ \ Content}{Description\ \ Length}$$

$$InfContent(p) = L(D \mid \mathcal{B}) - L(D \mid \mathcal{B} + p)$$

$$DescLength(p) = L(rows(p)) + L(cols(p)) + L(stat(p))$$

Results

	It 1	It 2	It 3	It 4	It 5	Final
1.	**A2**	**B3**	**A3**	**B2**	**C3**	**A2**
2.	A4	B4	B2	C3	C4	**B3**
3.	A3	B2	C3	C4	C2	**A3**
4.	B3	A3	C4	C2	D2	**B2**
5.	B4	C3	C2	B4	D4	**C3**
6.	B2	C4	B4	D2	D3	**C2**
7.	C3	C2	D2	D4	D1	**D2**
8.	C4	D2	D4	D3	A5	**D3**
9.	C2	D4	D3	D1	*21*	**A5**
10.	D2	D3	B1	A5	B5	**B5**

Synthetic Data

- random Gaussian
- 4 'complexes' (ABCD) of 5 overlapping tiles
- (x2 + x3 big with low overlap)

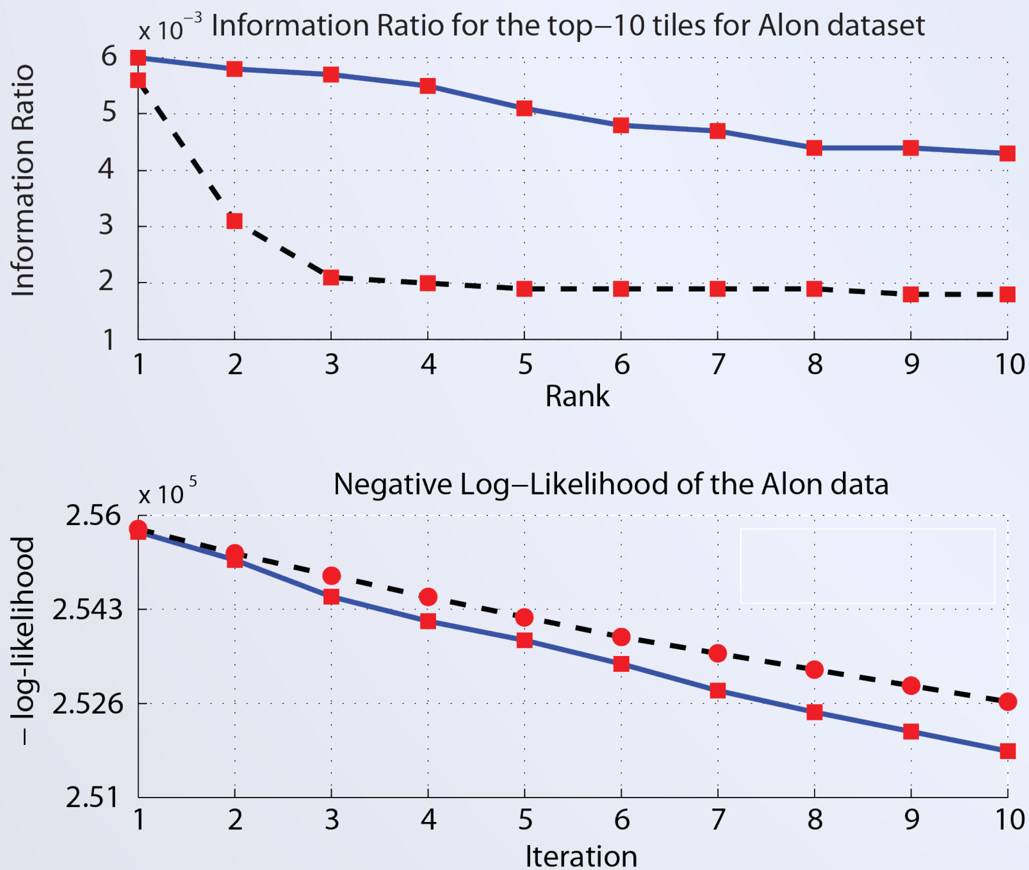
Patterns

- real + random tiles

Task

- Rank on InfRatio, add best to model, iterate

Results



Real Data

- gene expression

Patterns

- Bi-clusters from external study

Legend:

solid line	histograms
dashed line	means/var

Conclusions

Maximum Entropy modelling

- allows for **subjective** interestingness measurement

For **real-valued** data, we can now

- model **expectations** over **arbitrary sets of cells**
- measure the InfRatio for tiles
- pre-requisites for **iterative data mining**

Future work includes

- richer data types and statistics
- develop e.g. subspace cluster selection algorithms

Thank you!

Maximum Entropy modelling

- allows for **subjective** interestingness measurement

For **real-valued** data, we can now

- model **expectations** over **arbitrary sets of cells**
- measure the InfRatio for tiles
- pre-requisites for **iterative data mining**

Future work includes

- richer data types and statistics
- develop e.g. subspace cluster selection algorithms