Iterative Data Mining

Jilles Vreeken

26 June 2014 (TADA)

UNIVERSITÄT DES SAARLANDES

CLUSTER OF EXCELLENCE

max planck institut informatik

Service Announcement #1

Evaluation Forms

1. Hand forms out (me)
2. Fill forms out (you)
3. Collect forms (you)
4. Put forms in envelop (you)
5. Bring envelop back to Evelyn (one ‘volunteer’ and me)

Service Announcement #2

The Exam

type: oral
when: September 11th
time: individual
where: E1.3 room 0.16
what: all material discussed in the lectures, plus one assignment (your choice) per topic

The Re-Exam

type: oral
when: October 1st
time: individual
where: E1.3 room 001

Service Announcement #3

Master thesis projects

- in principle: yes!
- in practice: depending background, motivation, interests, and grades --- plus, on whether I have time
- interested? mail me and/or Pauli

Student Research Assistant (HiWi) positions

- in principle: maybe!
- in practice: depends on background, grades, and in particular your **motivation** and **interests**
- interested? mail me and/or Pauli, include CV and grades

Service Announcement #4

Introduction

- Is DM science?
- DM in action

Tensors

- Introduction to tensors
- Tensors in DM
- Special topics in tensors

Information Theory

- MDL + patterns
- Entropy + correlation
- MaxEnt + iterative DM

Mixed Grill

- Influence Propagation
- Redescription Mining
- ***<special request>***

Service Announcement #4

<special request>?

Let us know (asap, mail) what topic **you** would like to see discussed

sors

nsors

Info

tion

ing

- MaxEnt + iterative DM

- ***<special request>***

Service Announcement #5

Introduction

Tensors

Information Theory

Mixed Grill

Wrap-up + <*ask-us-anything*>

Servi

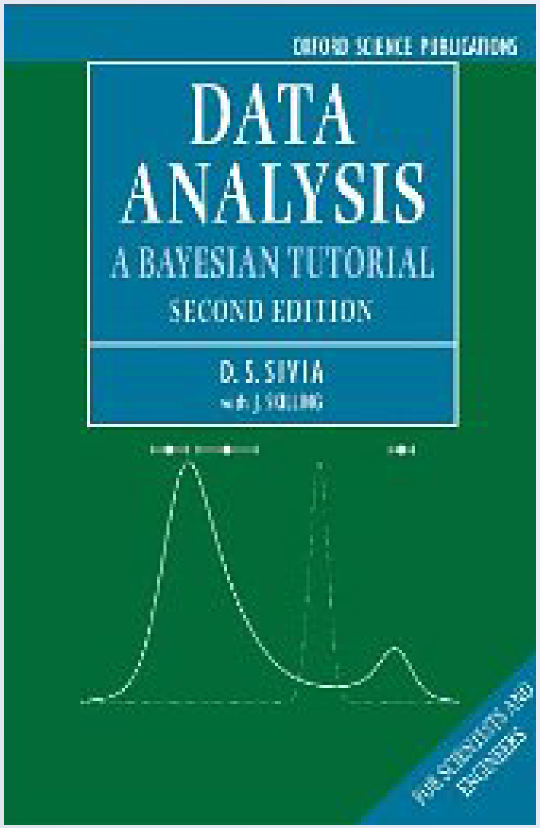
<ask-us-anything>?

Yes!

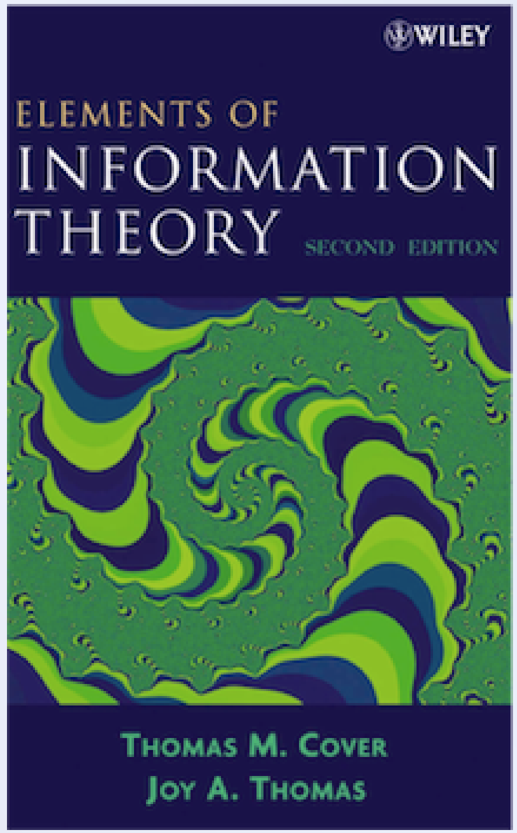
Prepare questions on **anything*** you've always wanted to ask Pauli and/or me. We'll answer **on the spot**

* preferably related to TADA, data mining, machine learning, science, the world, etc.

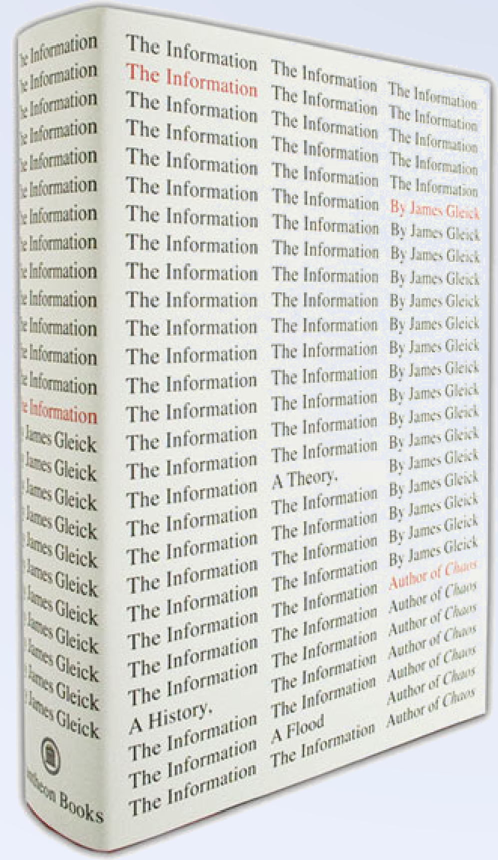
Good Reads



Data Analysis: a Bayesian Tutorial
D.S. Sivia & J. Skilling
(very good, but skip the MaxEnt stuff)



Elements of Information Theory
Thomas Cover & Joy Thomas
(very good textbook)



The Information
James Gleick
(great light reading)

Iterative Data Mining

Jilles Vreeken

26 June 2014 (TADA)

UNIVERSITÄT DES SAARLANDES

CLUSTER OF EXCELLENCE

max planck institut informatik

Question of the day

How can we find things
that are interesting with regard to
what we already know?

How can we measure
subjective interestingness?

What is interesting?

something that
increases our knowledge
about the data

What is a good result?

something that
reduces our uncertainty
about the data

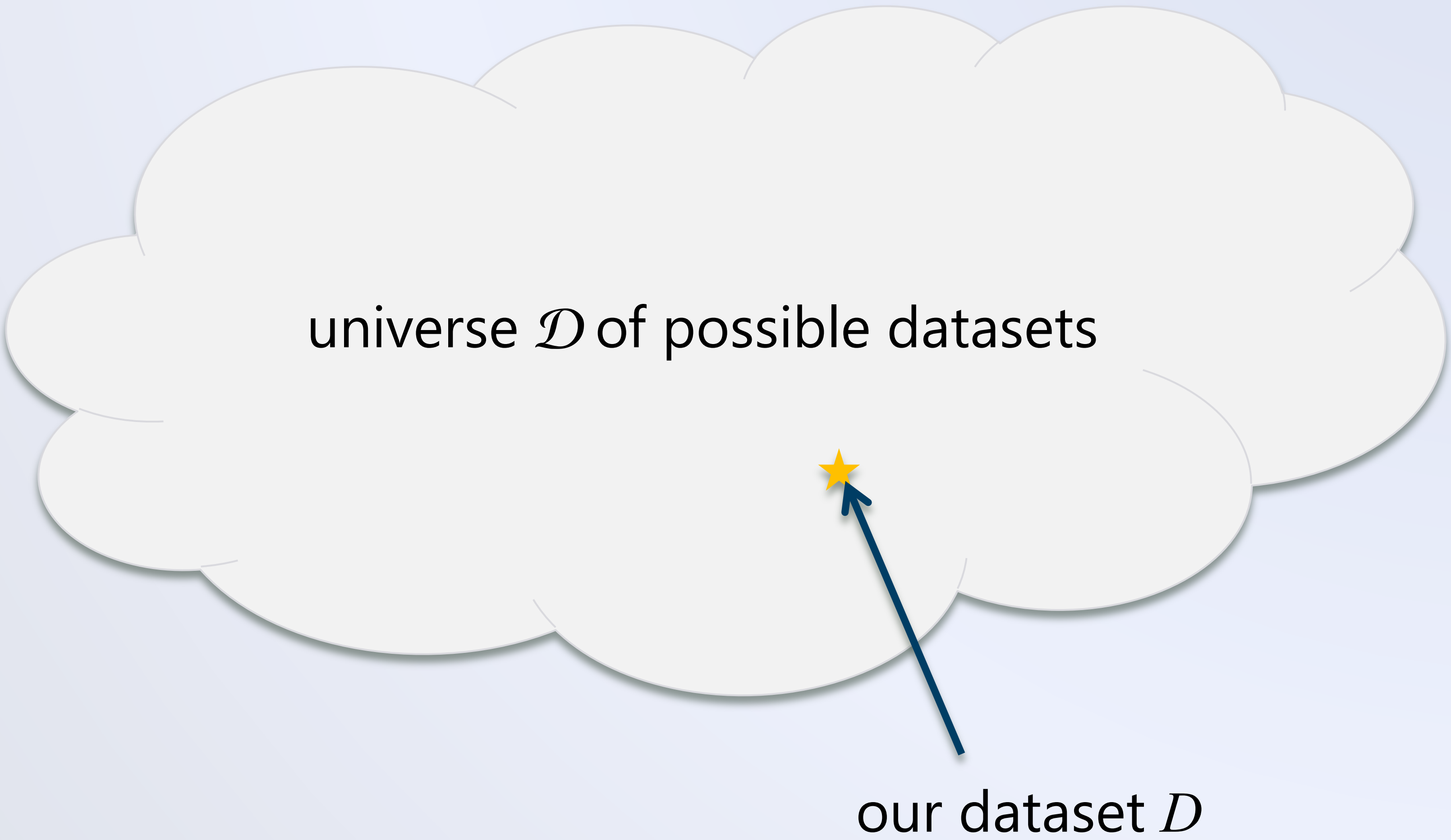
(ie. increases the likelihood of the data)

What is really good?

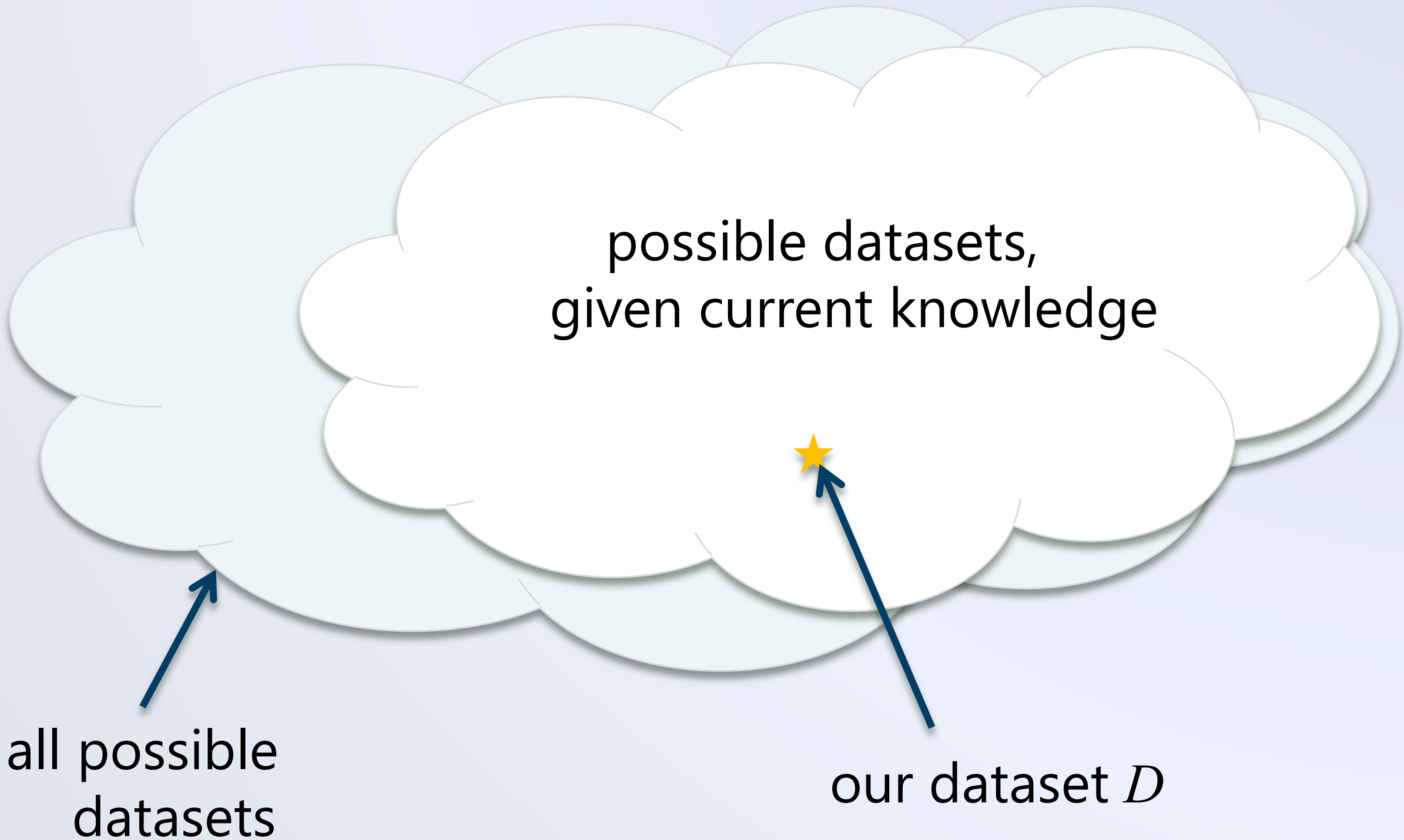
something that,
in **simple terms**,
strongly reduces our uncertainty
about the data

(maximise likelihood, but avoid overfitting)

Let's make this visual

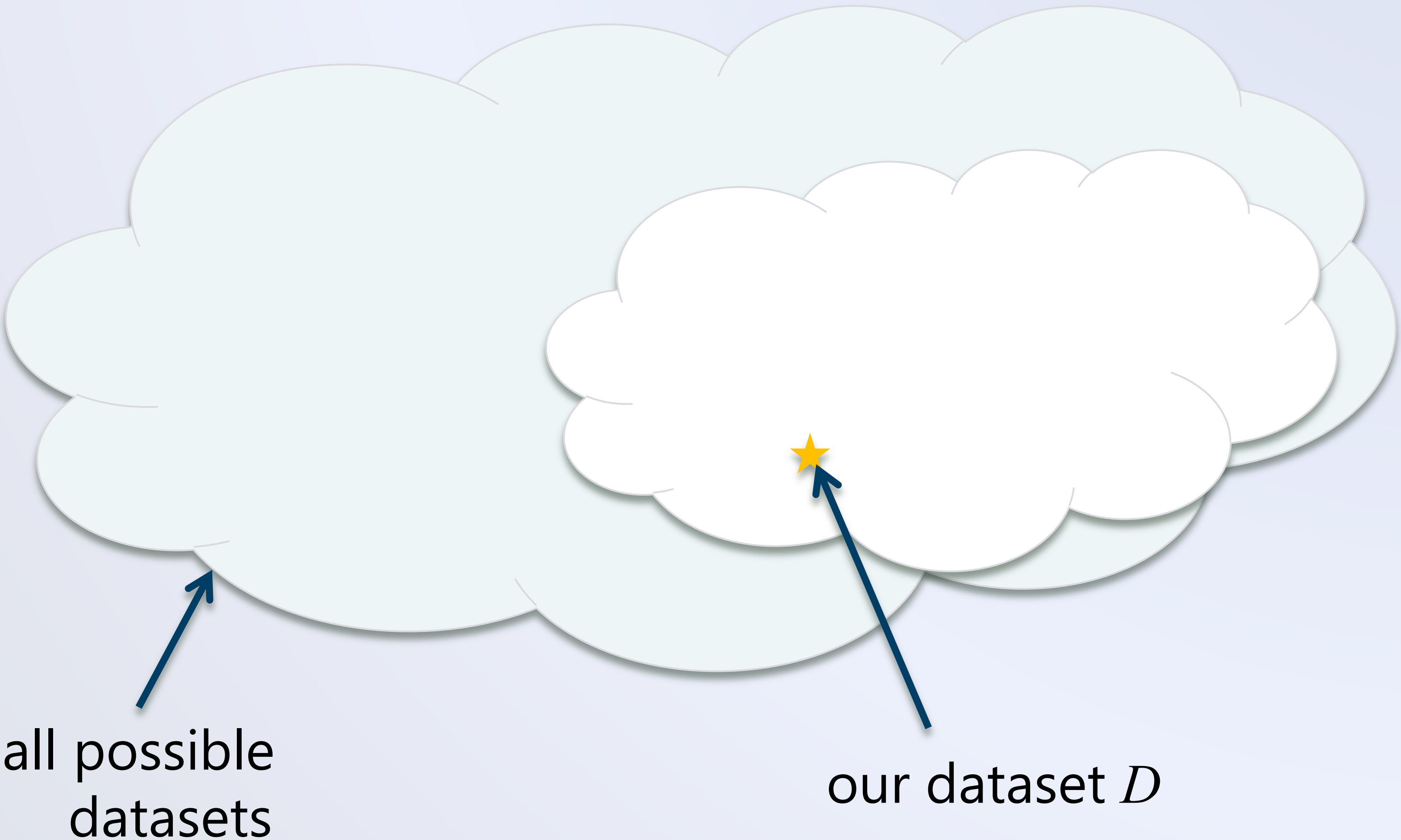


Given what we know



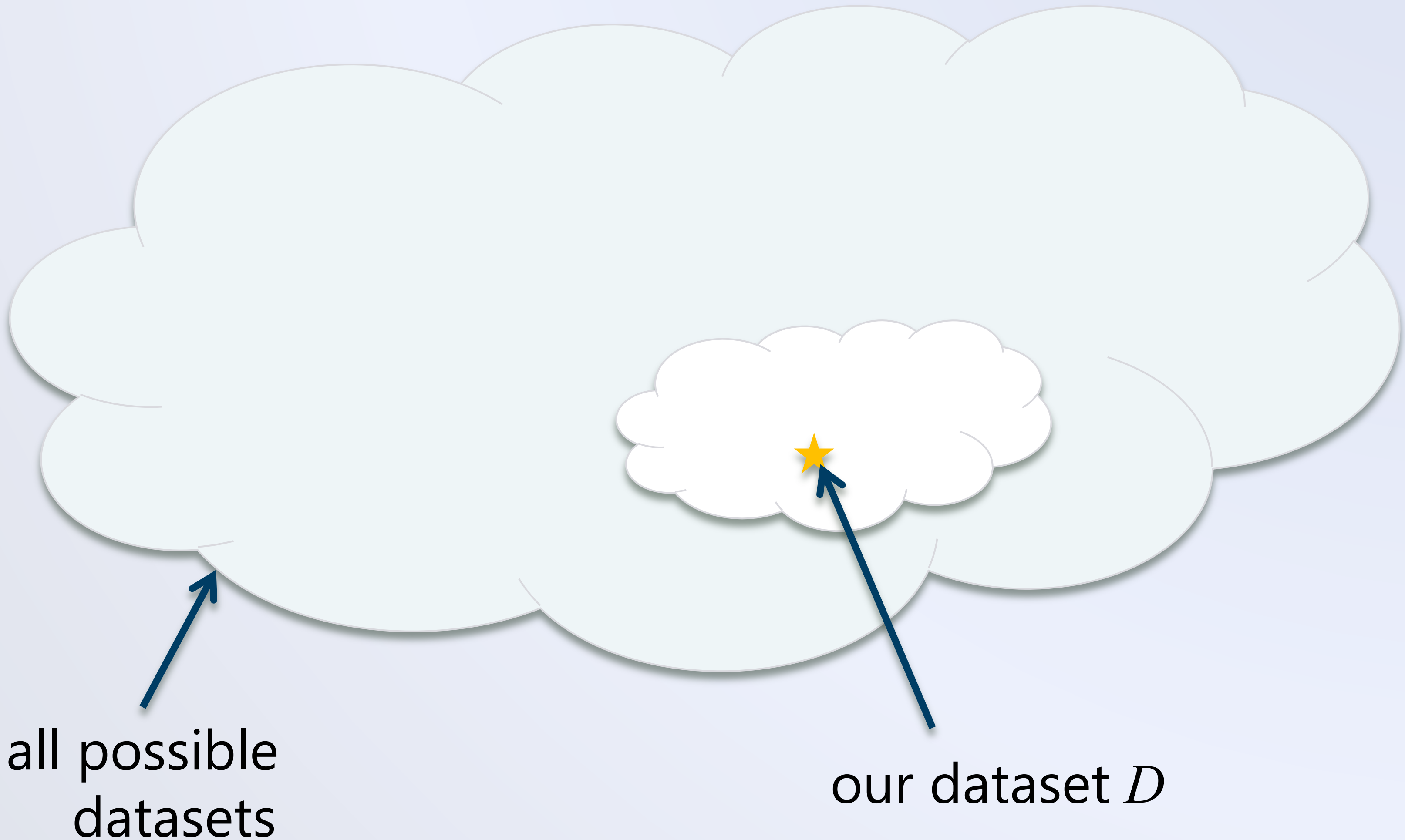
dimensions, margins, pattern P_1

More knowledge...



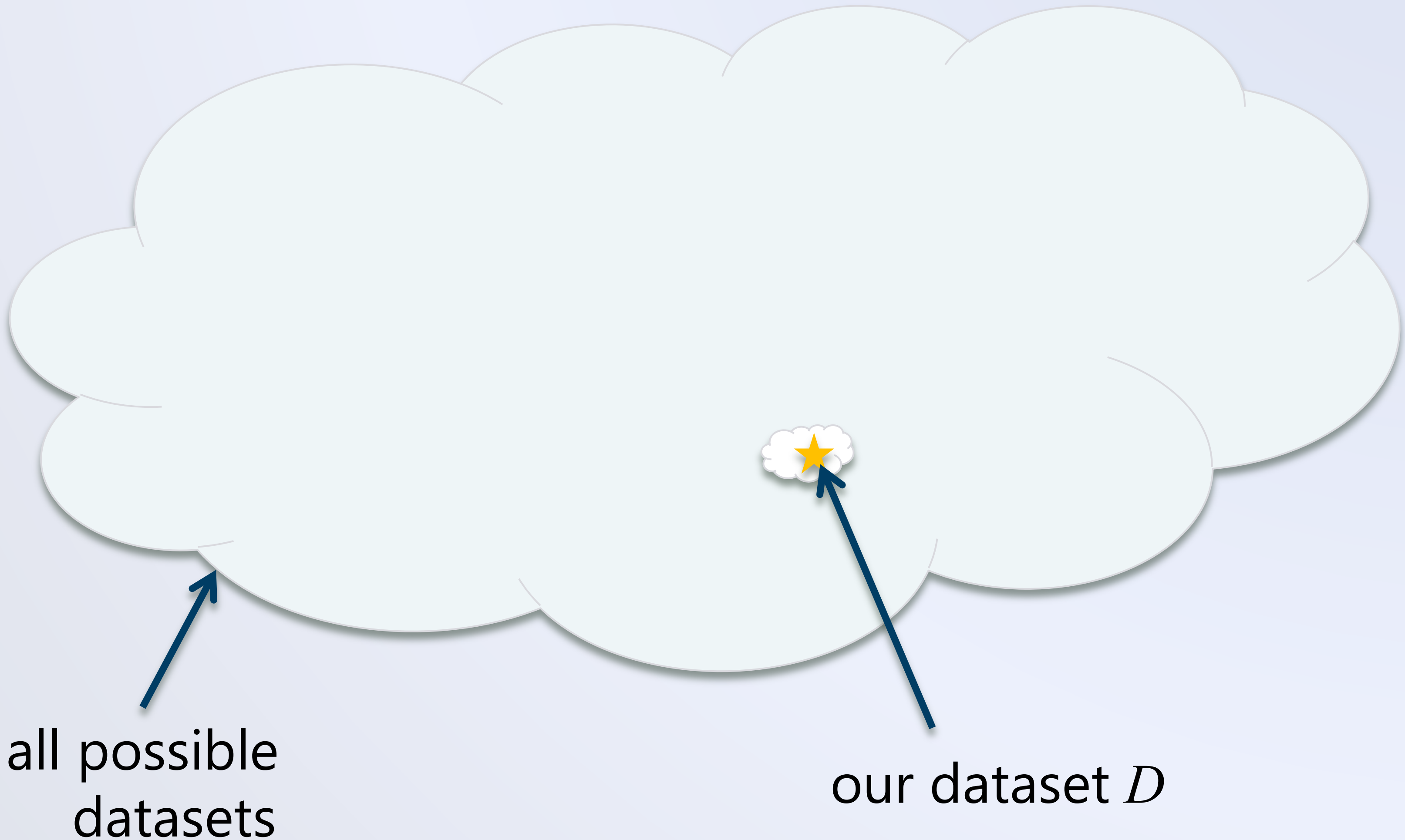
dimensions, margins,
patterns P_1 and P_2

Fewer possibilities…



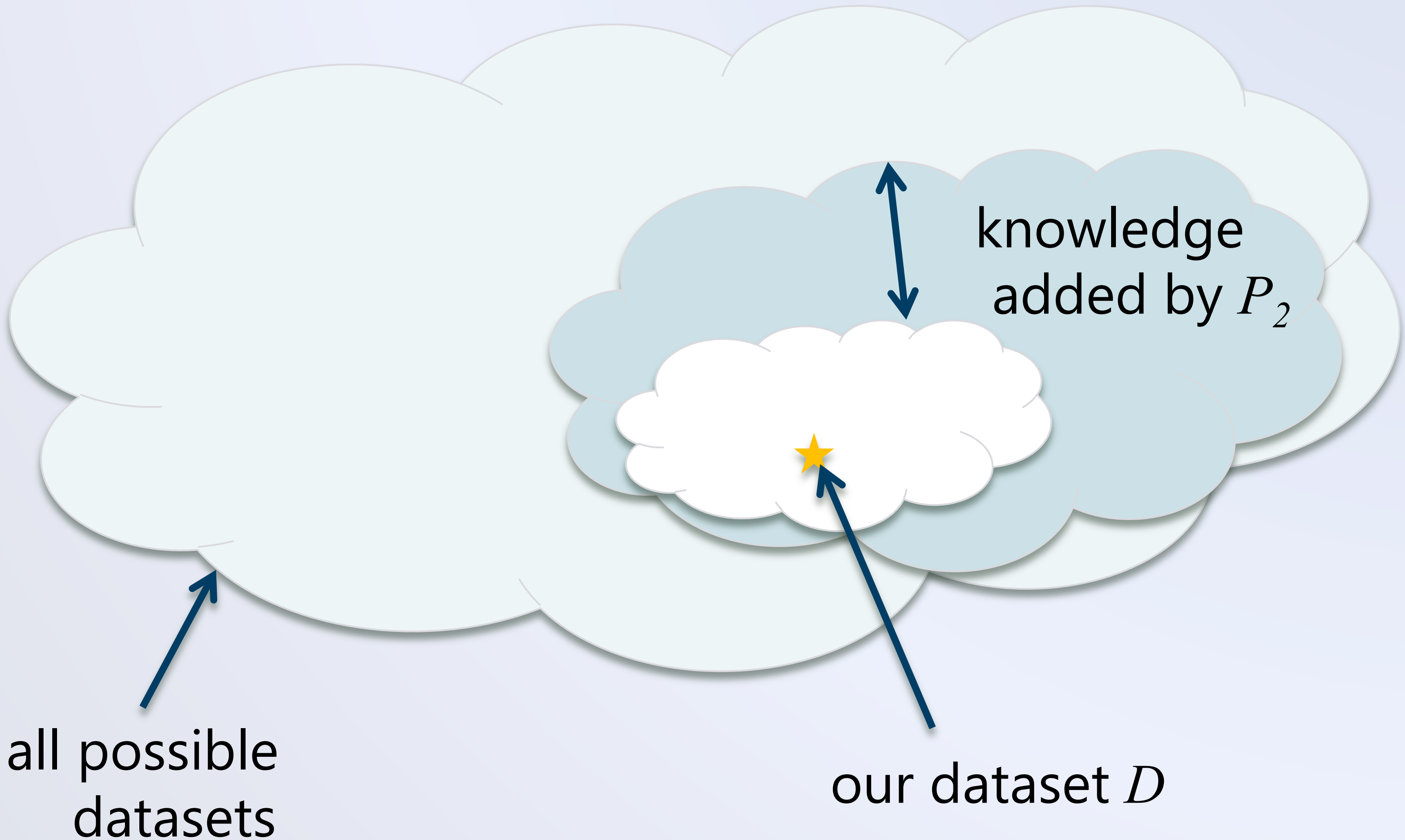
dimensions, margins,
the key structure

Less uncertainty.



Maximising certainty

dimensions, margins, patterns P_1 and P_2



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?

Measuring Surprise

We need access to the **likelihood**
of result X given background knowledge B

$$P(X \mid B)$$

such that we can mine the data for
X that have a low likelihood,
that are ***surprising***

…which distribution should we use?

Measuring Surprise

We need access to the **likelihood**
of result X given background knowledge B

This is called the
p-value
of result X

such that we can mine the data for
X that have a low likelihood,
that are ***surprising***

…which distribution should we use?

Measuring Surprise

We need access to the **likelihood**
of re… given background knowledge B

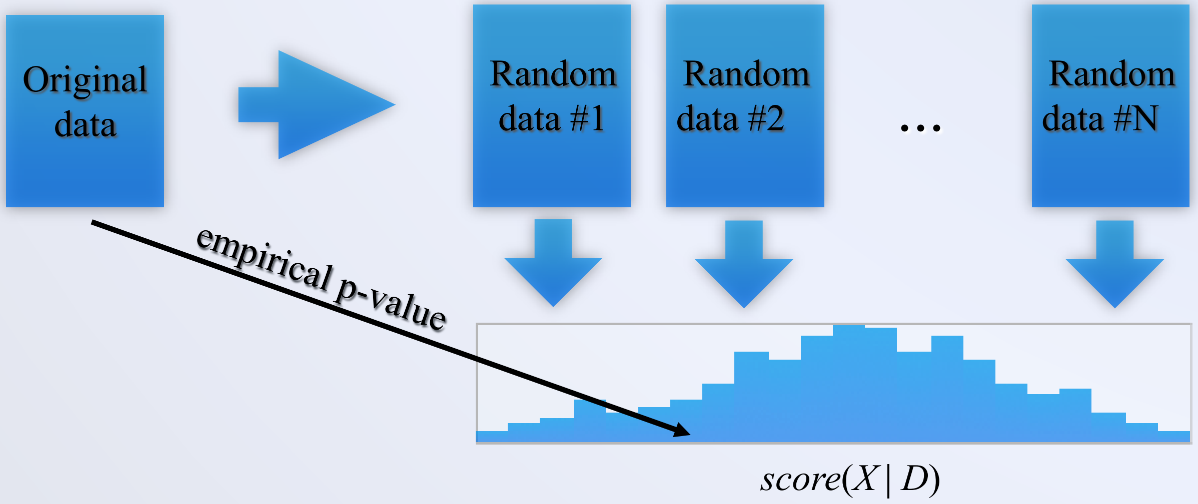
Standard practice:
if the ***p-value*** of result $X < 0.5$
its **significant**

such that we ca…
X that have a low like…
that are ***surprising***

…which distribution should we use?

Approach 1: Randomization

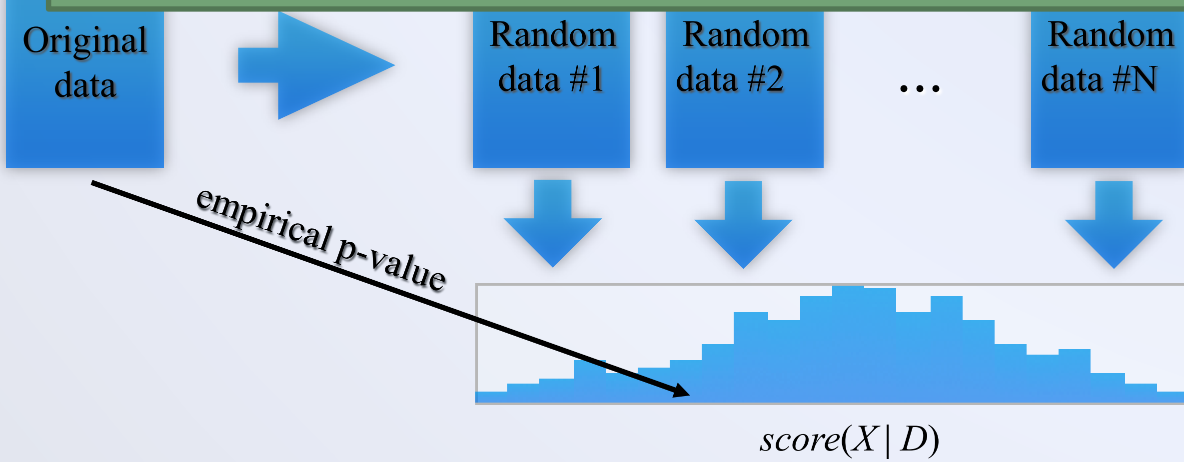
1. Mine original data
2. Mine random data
3. Determine probability



Approach 1: Randomization

1. Mine original data
2. Mine random data

The fraction of ***better*** 'randoms' is the
empirical p-value
of result X

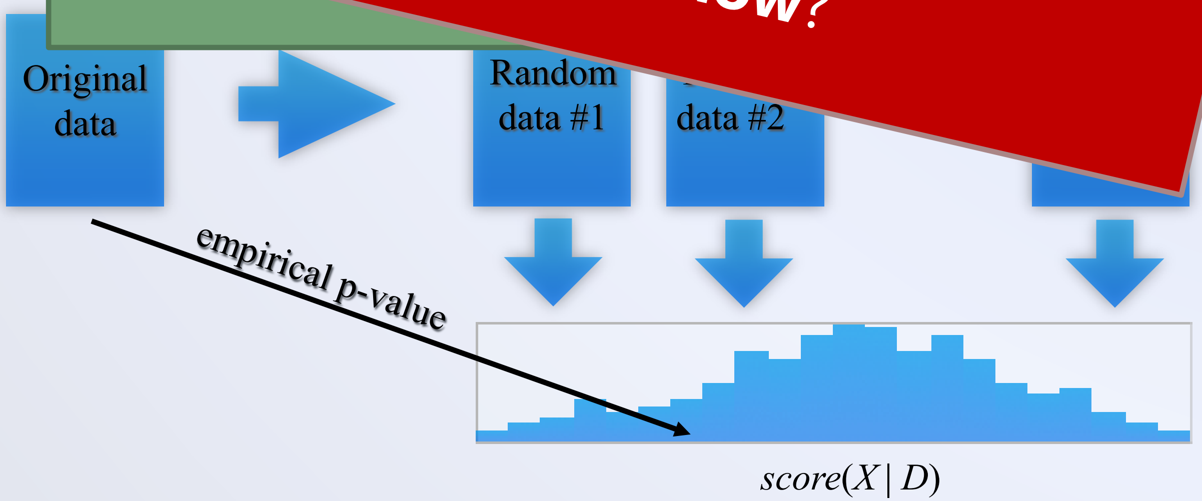


Approach 1: Randomization

1. Mine original data
2. Mine random data

...doms' is the

Random data?
Random **how**?



Random Data

So, we need data that

- maintains our background knowledge, and
- is otherwise completely random

How can we get our hands on that?

Swap Randomization

Let there be data

1	1	1	0	1	1	1
0	1	1	0	1	0	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
0	1	1	1	0	0	0
0	1	1	1	0	1	0
0	0	0	0	1	0	0

(swap randomization, Gionis et al. 2005)

Swap Randomization

Say we only know overall density.

How to sample random data?

1	1	1	0	1	1	1	
0	1	1	0	1	0	1	
1	1	1	1	0	0	0	
1	1	1	1	0	0	1	
0	1	1	1	0	0	0	
0	1	1	1	0	1	0	
0	0	0	0	1	0	0	
							27

(swap randomization, Gionis et al. 2005)

Swap Randomization

Didactically, let us instead consider a Monte-Carlo Markov Chain

Very simple scheme

1. select two cells at random,
2. swap values,
3. repeat until *convergence*.

1	1	1	0	1	1	1	
0	1	1	0	1	0	1	
1	1	1	1	0	0	0	
1	1	1	1	0	0	1	
0	1	1	1	0	0	0	
0	1	1	1	0	1	0	
0	0	0	0	1	0	0	
							27

(swap randomization, Gionis et al. 2005)

Swap Randomization

Margins are easy understandable for binary data, how can we sample data with same margins?

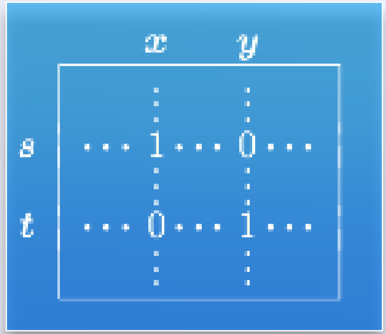
1	1	1	0	1	1	1		6
0	1	1	0	1	0	1		4
1	1	1	1	0	0	0		4
1	1	1	1	0	0	1		5
0	1	1	1	0	0	0		3
0	1	1	1	0	1	0		4
0	0	0	0	1	0	0		1
3	6	6	4	3	2	3		27

(swap randomization, Gionis et al. 2005)

Swap Randomization

By MCMC!

1. randomly find submatrix



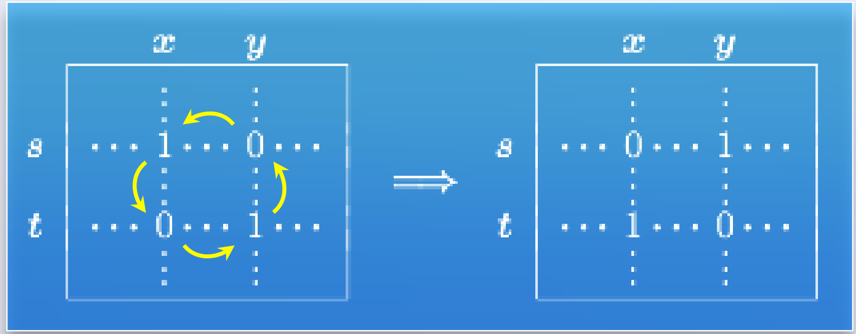
1	1	1	0	1	1	1	6
0	1	1	0	1	0	1	4
1	1	1	1	0	0	0	4
1	1	1	1	0	0	1	5
0	1	1	1	0	0	0	3
0	1	1	1	0	1	0	4
0	0	0	0	1	0	0	1
3	6	6	4	3	2	3	27

(swap randomization, Gionis et al. 2005)

Swap Randomization

By MCMC!

1. randomly find submatrix



2. swap values

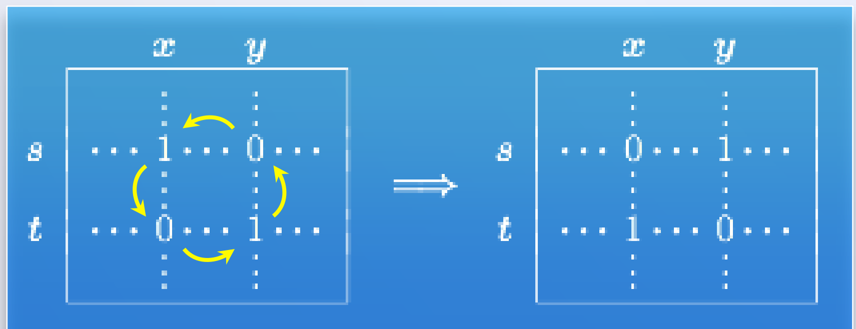
1	1	1	0	1	1	1	6
0	1	1	0	1	0	1	4
1	1	1	1	0	0	0	4
1	1	1	1	0	0	1	5
0	1	1	1	0	0	0	3
0	1	1	1	0	1	0	4
0	0	0	0	1	0	0	1
3	6	6	4	3	2	3	27

(swap randomization, Gionis et al. 2005)

Swap Randomization

By MCMC!

1. randomly find submatrix



2. swap values
3. repeat until *convergence*

1	1	1	0	1	1	1	6
0	1	1	1	0	0	1	4
1	1	1	0	1	0	0	4
1	1	1	1	0	0	1	5
0	1	1	1	0	0	0	3
0	1	1	1	0	1	0	4
0	0	0	0	1	0	0	1
3	6	6	4	3	2	3	27

(swap randomization, Gionis et al. 2005)

Static Models

Many ways to test **static** null hypothesis

assuming distribution,
swap-randomization, MaxEnt

What can we use this for?

ranking based on **static significance**

mining the **top-k** most significant patterns,

but **not** suited for iterative mining

Dynamic Models

For **iterative** data mining,
we need models that can maintain
the **type** of information (eg. patterns)
that we **mine**

Randomization is powerful

- variations exists for many data types (Ojala '09, Henelius et al '13)
- can be pushed beyond margins (see Hanhijärvi et al 2009)
- but… has key disadvantages

Approach 2: **Maximum Entropy**

'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)

Constraints and Distributions

Let B be our set of constraints

$$B = \{f_1, \ldots, f_n\}$$

Let C be the set of admissible distributions

$$C = \{\, p \in \mathbf{P} \mid p(f_i) = \tilde{p}(f_i) \text{ for } f_i \in B\}$$

We need the
most **uniformly** distributed $p \in \mathbf{P}$

Uniformity and Entropy

Uniformity $\leftrightarrow$ Entropy

$$H(p) = -\sum_{x \in \mathbf{X}} p(X = x) \log p(X = x)$$

tells us the **entropy** of a (discrete) distribution p

Maximum Entropy

We want access to the distribution p^* with **maximum entropy**

$$p_B^* = \operatorname{argmax}_{p \in C} H(p)$$

better known as the **maximum entropy model**

It can be shown that p^* is well defined there *always** exist a unique p^* with maximum entropy for any constrained set C

* that's not completely true, some esoteric exceptions exist

Some examples

Mean and

■ interval?	uniform
■ variance?	Gaussian
■ positive?	exponential
■ discrete?	geometric
■ …	

But… what about distributions
for like data, patterns, and stuff?

MaxEnt Theory

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

Patterns

- itemset frequencies (Tatti '06, Mampaey et al. '11)

Binary Data

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

Real-valued Data

- margins (Kontonasios et al. '11)
- sets of cells (Kontonasios et al. '13)

Exponential Form

Let p be a probability density satisfying the constraints

$$\int_S p(x) f_i(x) dx = \alpha_i \quad \text{for } 1 \leq i \leq m \quad .$$

Then we can write the MaxEnt distribution as

$$p^*(x) = p_\lambda(x) = \propto \begin{cases} \exp\left(\lambda_0 + \sum_{f_i \in B} \lambda_i \cdot f_i(x)\right) & D \notin \mathcal{Z} \\ 0 & D \in \mathcal{Z} \end{cases} \quad ,$$

where we choose the lambdas to satisfy the constraints

(Csizar 1975)

Inferring the Model

The problem is convex – yay!

This means we can use
any convex optimization strategy.

Standard approaches include
iterative scaling,
gradient descent,
conjugate gradient descent,
Newton's method,
etc.

Inferring the Model

Optimization requires calculating *p*

for datasets and *tiles*
this is **easy**

for itemsets and frequencies, however,
this is **PP-hard**

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

Current state of the art can incorporate

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

over **arbitrary** sets of cells

(Kontonasios et al. 2013)

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

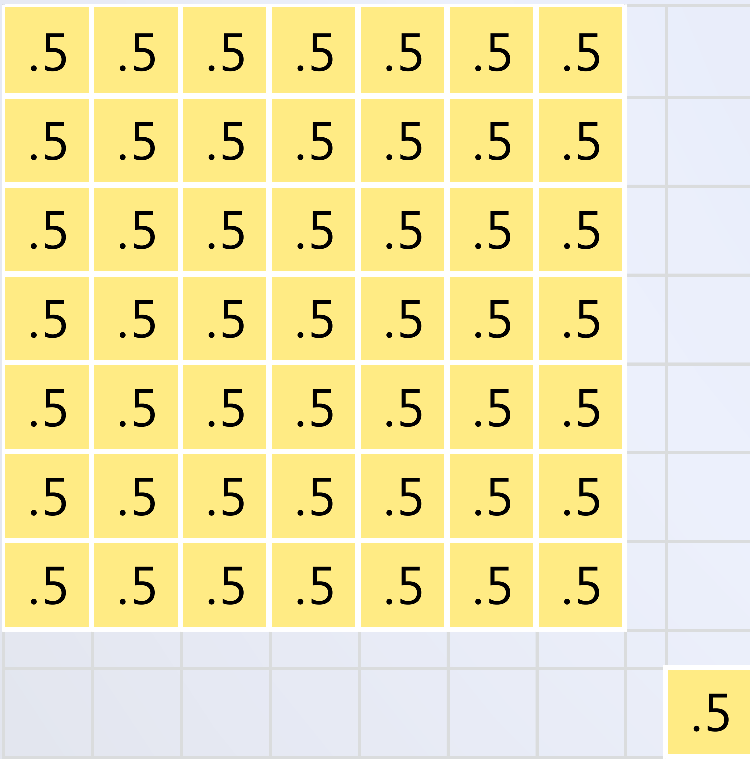
Pattern 2

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

Pattern 3

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

MaxEnt for Real-Valued Data



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

(Kontonasios et al. 2013)

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)

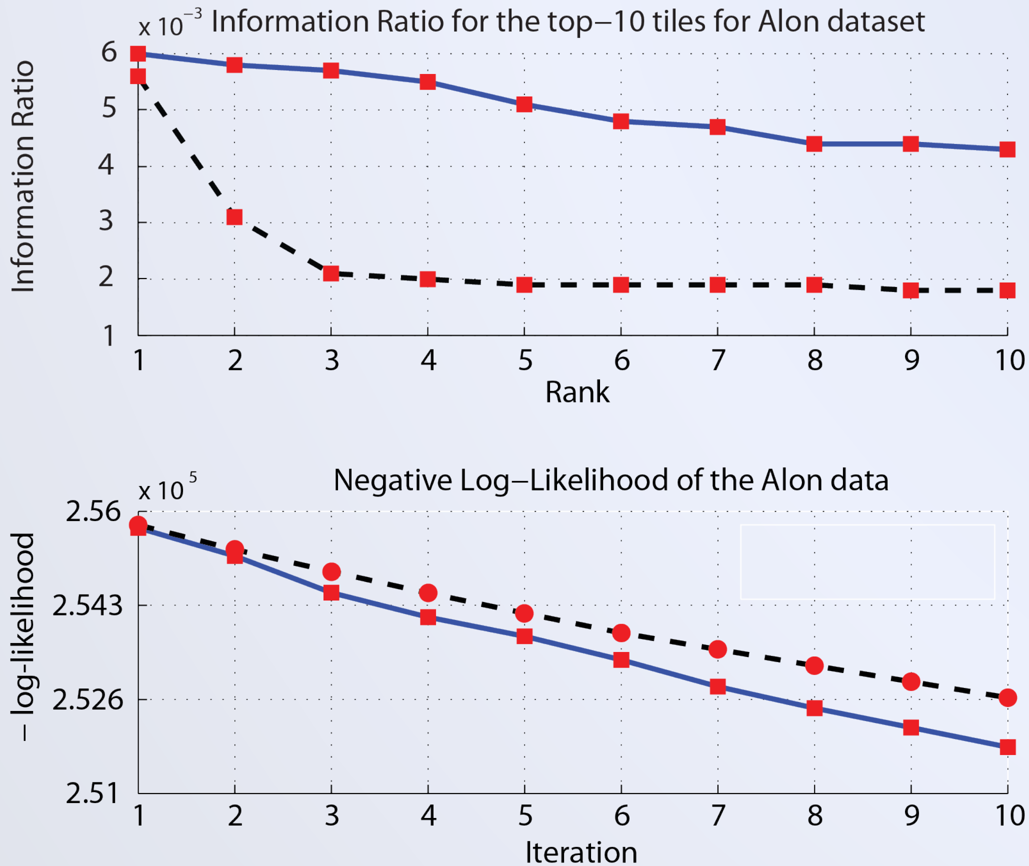
Patterns

- 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

Significance testing is important
- choosing a good model (and test) is difficult

Randomization
- simple yet powerful – difficult to extend – **empirical** p-values

Maximum Entropy modelling
- complex yet powerful –inferring can be NP-hard – **exact** p-values
- can be defined for **anything** …if you can derive the model…

Iterative Data Mining
- mine most informative thingy, update model, repeat.

Thank you!

Significance testing is important

- choosing a good model (and test) is difficult

Randomization

- simple yet powerful – difficult to extend – **empirical** p-values

Maximum Entropy modelling

- complex yet powerful –inferring can be NP-hard – **exact** p-values
- can be defined for **anything** …if you can derive the model…

Iterative Data Mining

- mine most informative thingy, update model, repeat.