

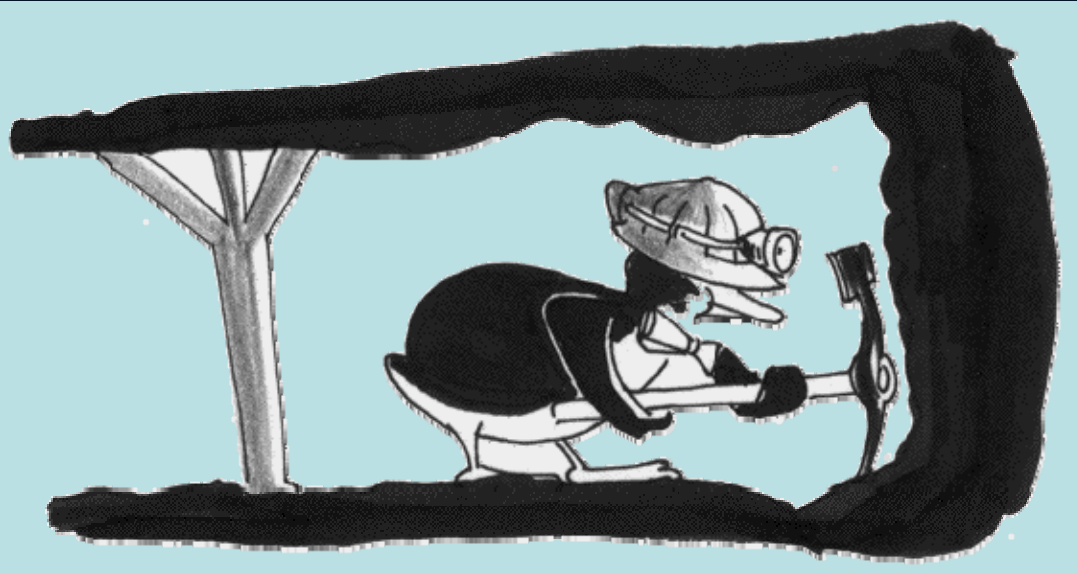


THE AUSTRALIAN NATIONAL UNIVERSITY

Data Mining in Physics

Frank Detering, Boyd Blackwell, Markus Hegland, David Pretty
Research School of Physical Sciences and Engineering &
Mathematical Sciences Institute

Data Mining: Prospecting in Physics



Data Mining:

"the science of extracting useful information from large data sets or databases" (D. Hand, H. Mannila, P. Smyth (2001))

Physics at PRL

H1 is currently at pulse# 64581

20-50MB of data/pulse

Large amount of data + complex experiment

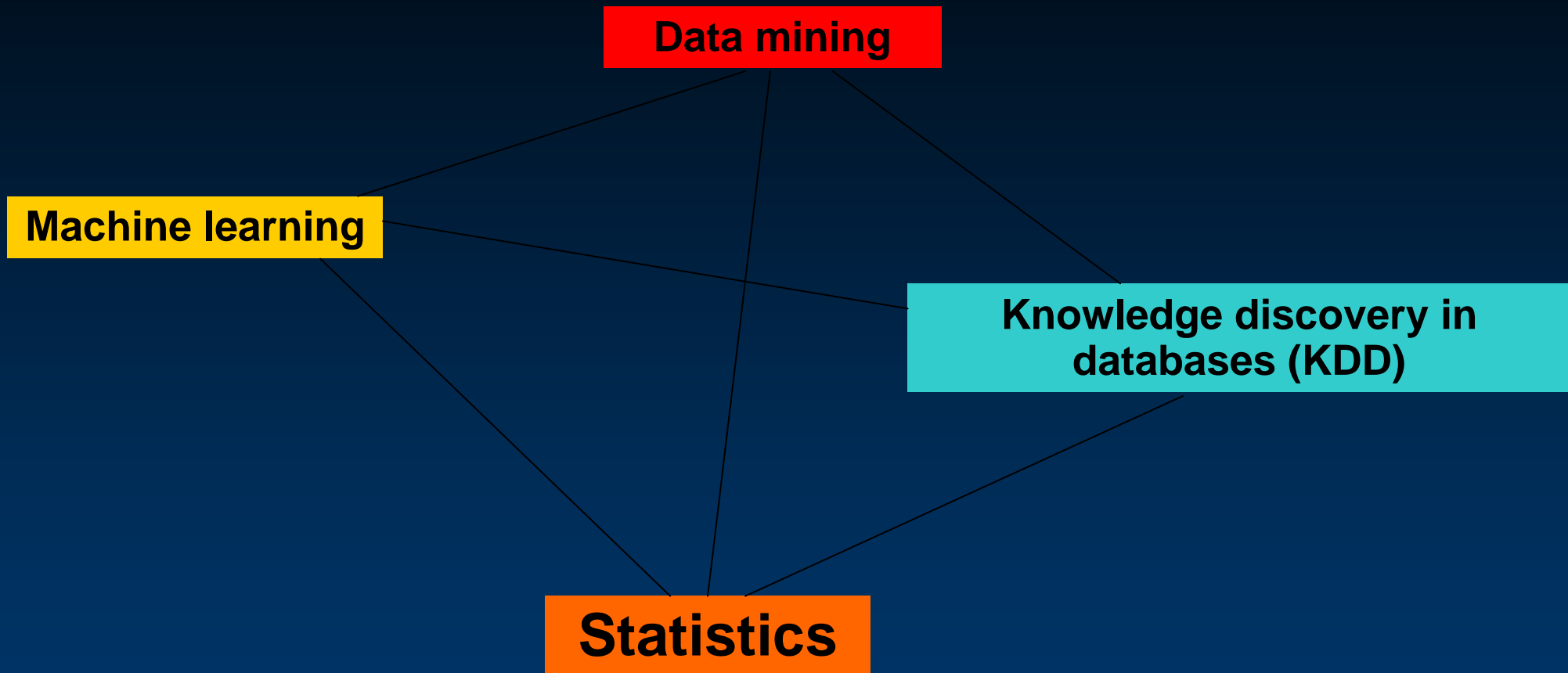
-> **ideal candidate for data mining**



Guideline

- What is “data mining” and what is so special about it?
- A toolbox of statistical techniques and their uses outside the hard sciences
(and where physicists can learn from Walmart)
- The three steps of data mining
- H1 National Facility: the grand data mining challenge
 - Organizing the data
 - Magnetic fluctuation measurements
 - Applications of data mining techniques

Data Mining: Just a statistical technique?



Statistics and statistical techniques form the base of them all!
There exist many different data mining techniques.

Examples of Data Mining in Government and Business

Tax Office

Business task: income tax declarations

DM task: “unusual” declarations that deserve human review

→ Clustering, Regression, Classification

Bank

Business task: Credit risk analysis

DM task: find the important factors

→ Classification, Clustering, ...

E-Business (e.g. Google ads)

Business task: Placing your ads

DM task: Estimate customer interests

→ Association rules, Clustering

Supermarkets

Business task: Arrangement and promotions

DM task: find the frequent item combinations

→ Association rules, ...

One DM technique: Market Basket (or *what are association rules?*)

1. Are there frequent patterns in the composition of the market basket?
2. Are there “causal” relationships between the purchase of the items?

Statistically speaking



Nappies already in the caddy,
customer male, on a Friday

60%
probability



Nappies **AND** beer

The 3-steps of Data Mining

Preparation and Reduction (*from Data repository to warehouse*)

“Mining” for information and nuggets of knowledge

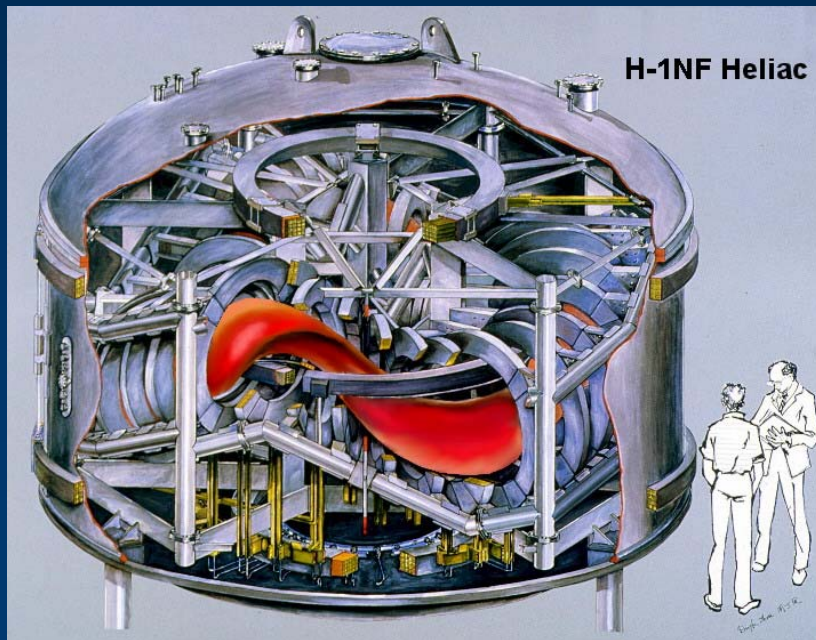
(*applying a DM algorithm*)

Many iterations
to identify the *real* nuggets

Visualization and
Analysis of the results

(*The human intervention*)

- The Heliac is a flexible “magnetic bottle”
- Investigating the step beyond the new ITER experiment in France
- World-leading advanced measurement systems



Argon plasma in H-1

Data Mining questions to H1:

- Provide a useful data structure (similar to a Data Warehouse)
- Was the shot *good*? (“machine learning”)
- Questions when looking at wave phenomena:
 - How to reduce the data? (the pre-processing)
 - How do frequency, spatial mode and rotational transform relate? (“clustering”)
 - How do modes interact in time and frequency? (“association rule”)

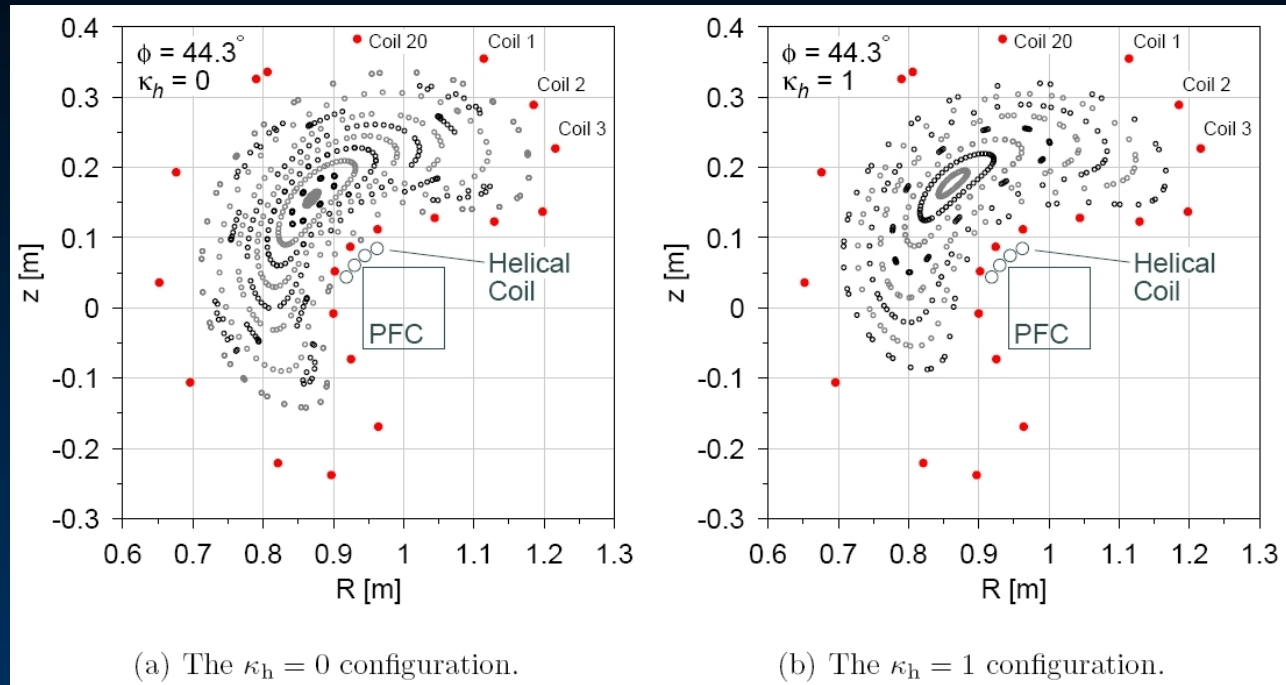
H1 data organization

- Raw data stored in remotely accessible data **repository**
- Main experimental parameters stored in a summary database (updated automatically, accessible via the web)
- **Here is an extract** from the summary database:

shot	N_e18	gas	M_{eff}	k_h	rf_power	MBytes
58142	0.91	H/He	4	0.89	56.7	13.2
58141	0.92	H/He	4	0.87	56.2	13.2
58140	0.81	H/He	4	0.85	56.6	13.2

- A simple but important step, especially for:
 - Book-keeping and knowledge transfer on multi-user facilities
- Next step: combination with processed raw data

Raw data = Magnetic probe array



Multi-channel time series

→ mode and spectrogram (k, f, t)

Mode analysis not simple

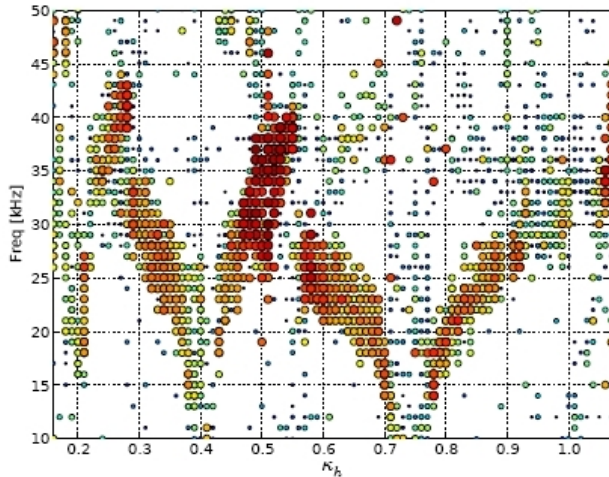
→ robust method for mode separation needed

Methods of choice:

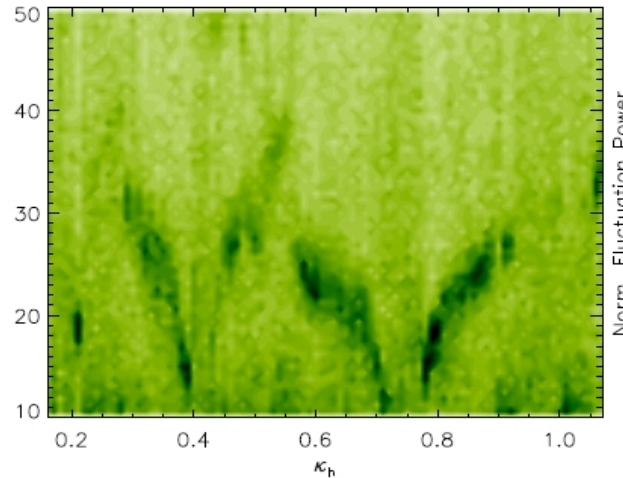
Singular value decomposition (SVD),
Principal component analysis (PCA)

Both find maximally separated, orthogonal signal components.

1. Clustering in frequency and spatial mode (k, \hat{n})



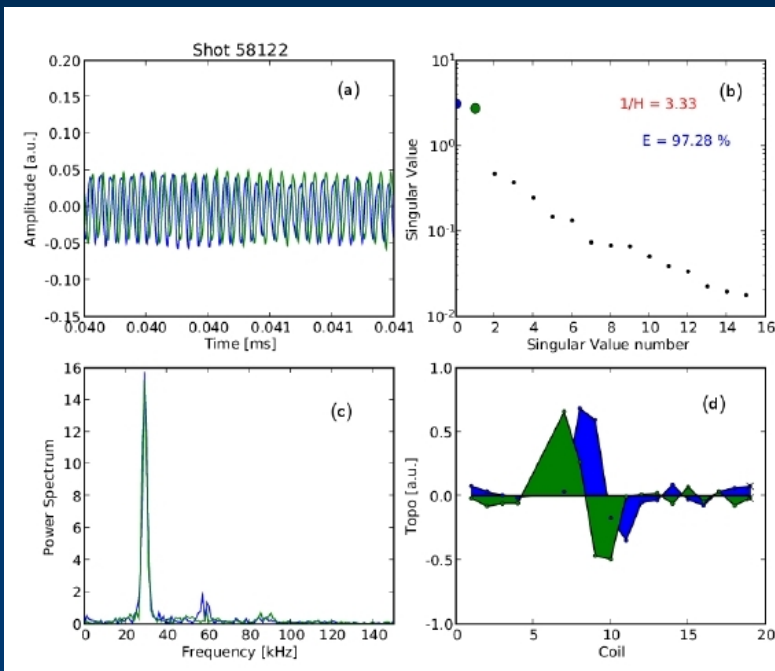
(a) Magnetic fluctuations



(b) Electron density fluctuations

Parameter scan reveals
“whale tails”

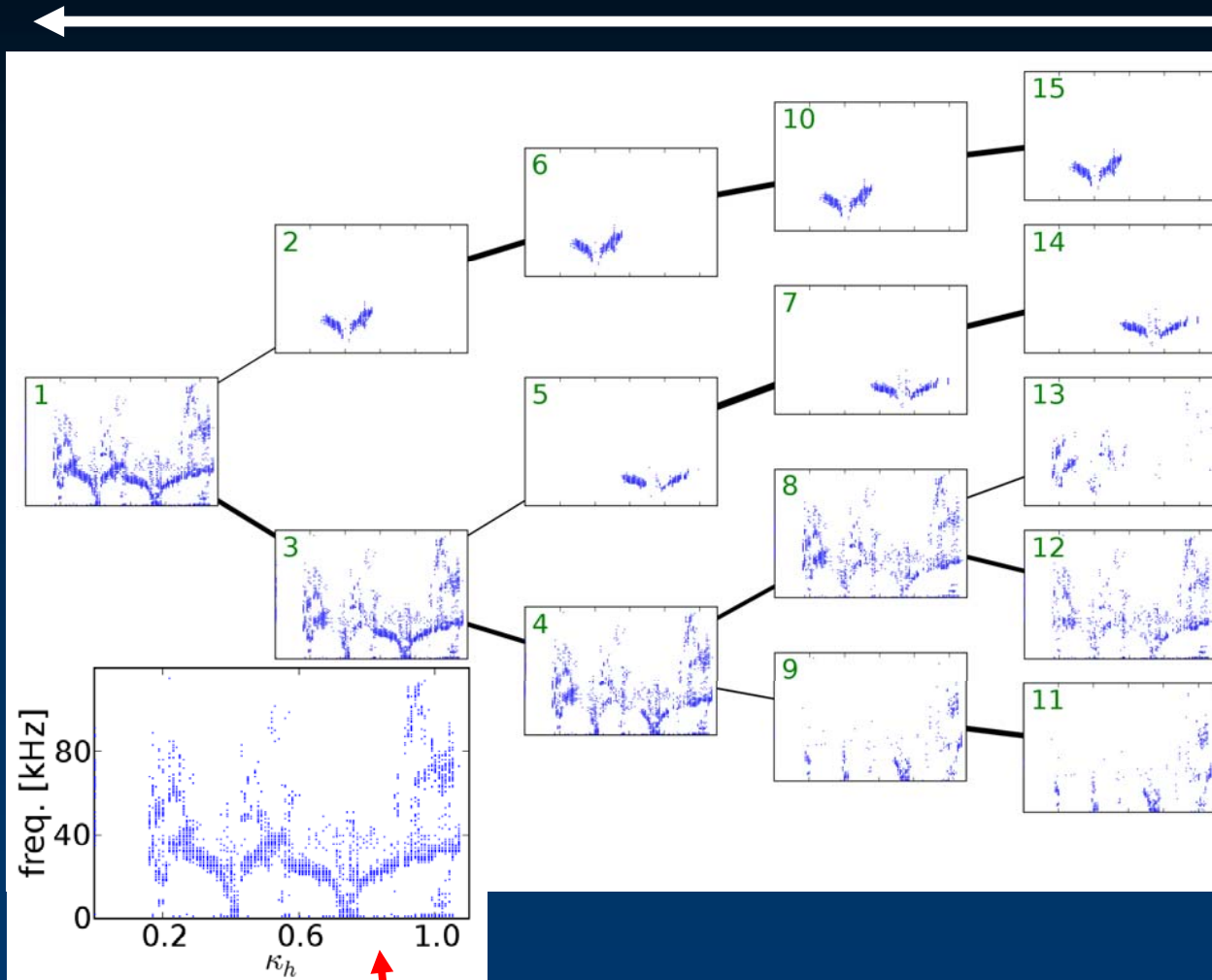
Is there an underlying
structure?



Ansatz (on short time segments):

1. SVD identifies dominating mode (possibly rotating)
2. Reconstruct phase difference at this mode at peak frequency
3. Perform clustering in the space of phase differences

Classification by hierarchical clustering

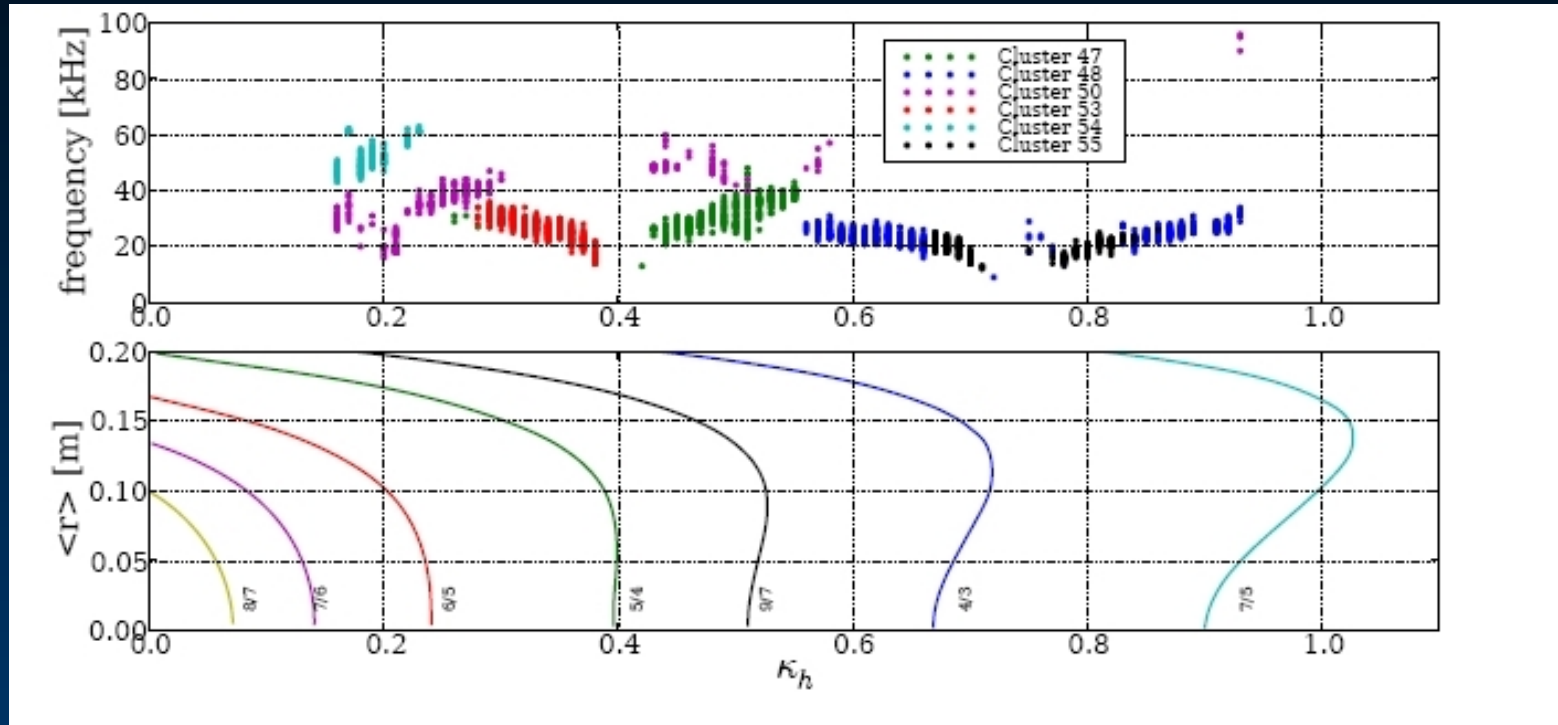


Procedure outline:

- Decrease the number of clusters
- Visualize
- Identify *strong* branches as significant clusters

All the peak frequencies

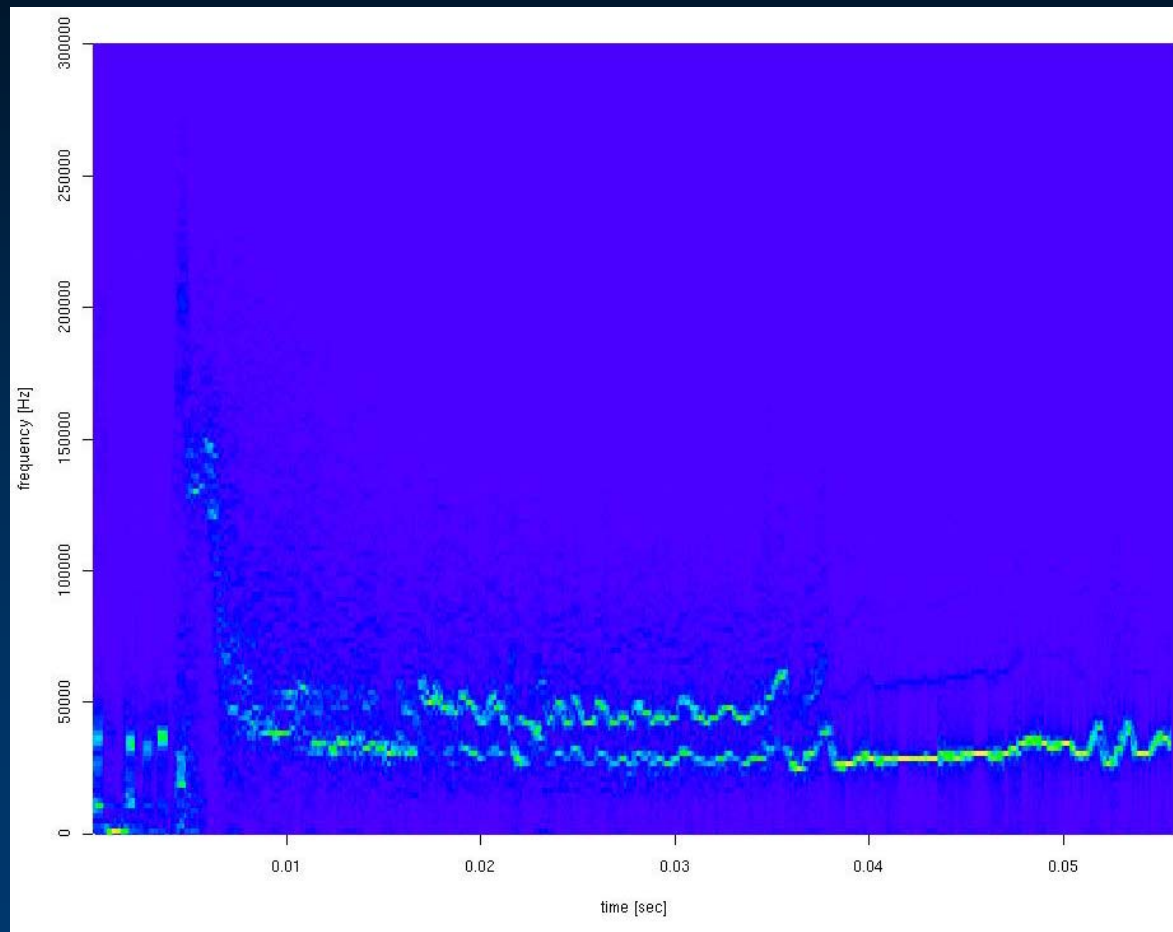
Clustering results of a 100 shot parameter scan



“Whale tails” fall into small number of clusters \rightarrow same spatial mode
 Procedure now employed on different machines (TJ-II, Heliotron J)

2. Using more information: time evolution

Typical *pre*-preprocessed measurement (spectrogram of strongest PCA)




How do events at different time, frequency and other parameters relate?

→ What are the *association rules* ?

Reminder: Association Rules



60%
probability

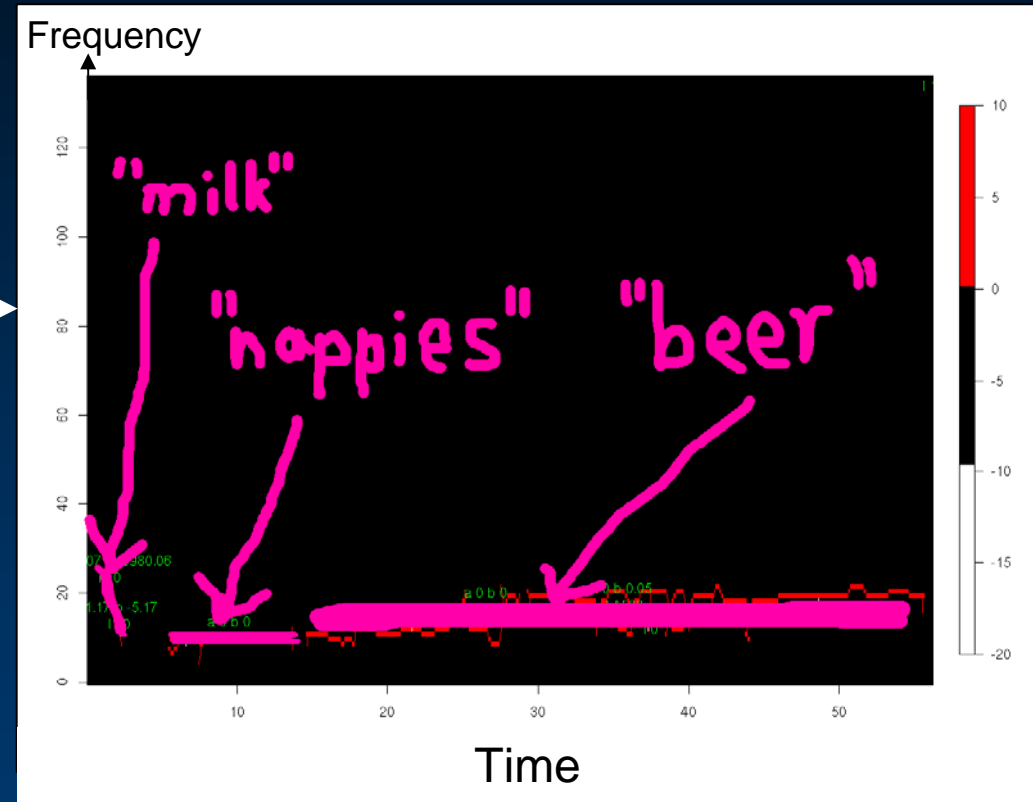
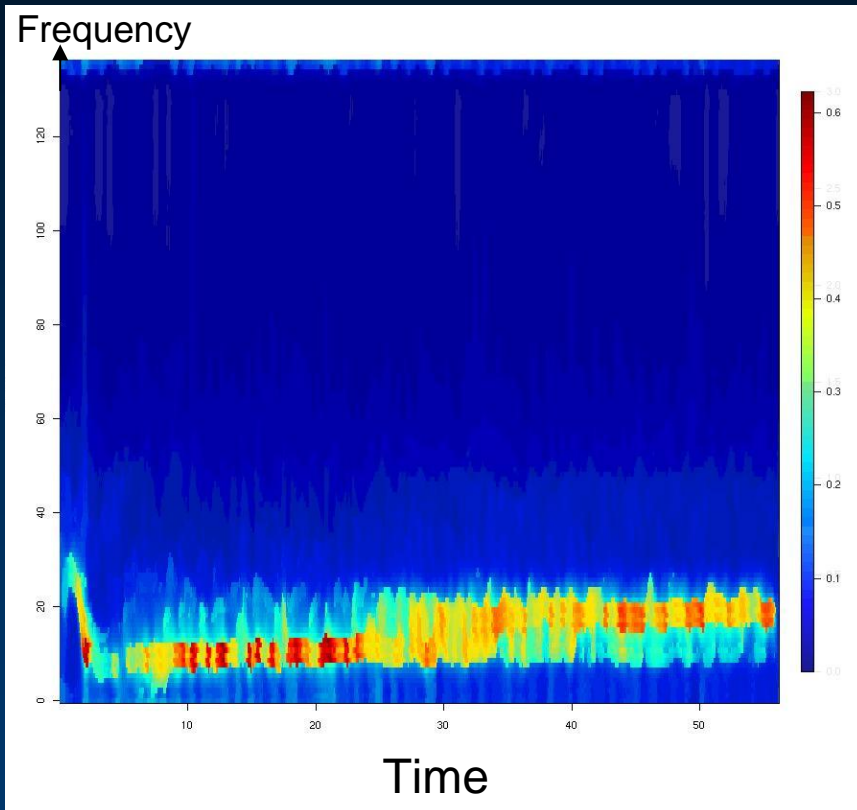


Nappies already in the caddy,
customer male, on a Friday

Nappies AND beer

→ Reduce spectrogram to **items (features)**

Image processing – data reduction



Smoothing, histogram techniques, image segmentation, curve fitting

Next step: “Feature database”

Shot + PCA	Time [ms]	Frequency [kHz]	Slope [kHz/ms]	Ne18	Rel. Energy [%]	
5804301	1.80	68.36	-30.61	0.08	2.97	
5804301	30.00	70.31	0.72	1.11	67.50	
5804301	17.28	74.22	-4.65	0.95	0.42	...
5804301	20.76	80.08	-0.02	1.04	0.11	
5804301	23.04	80.08	6.10	1.07	0.22	
5804301	35.16	46.88	0.00	1.19	0.15	
5804301	39.84	46.88	0.00	1.20	0.14	

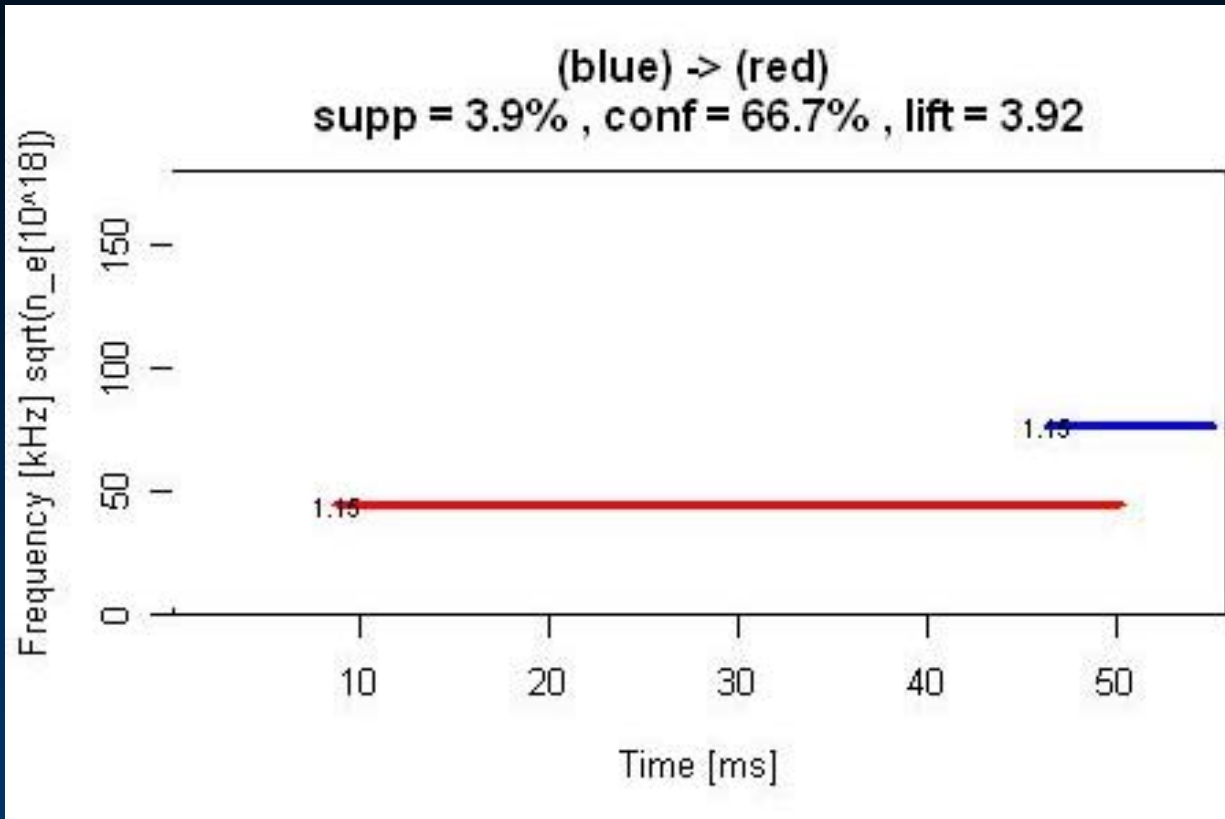
Raw data -> Feature database (5000x smaller)

Standard format and aggregate information → flexible use

Search tool (e.g. when did a chirp occur?)

direct link to DM tools for: clustering, association rules, etc.

First association rule!



How to read an association rule:

**“yy% of the experiments:
 If (blue) then (red) with
 xx% probability”**

yy% = support

xx% = confidence

Some observations:

large parameter space
 large number of rules

→→

→ small transaction lists despite large database

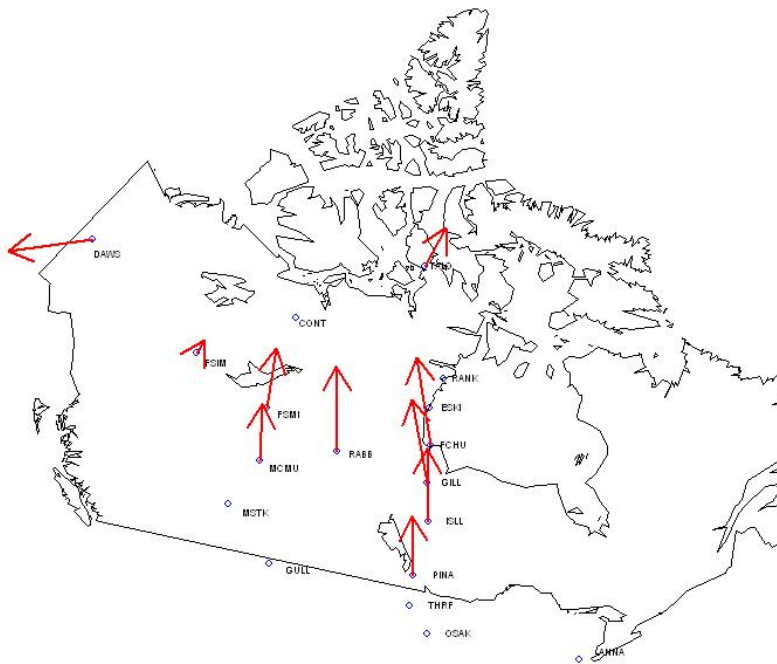
→ measure of *interestingness* needed

more iterations (in the *three DM steps*) needed
 the *real* work has begun

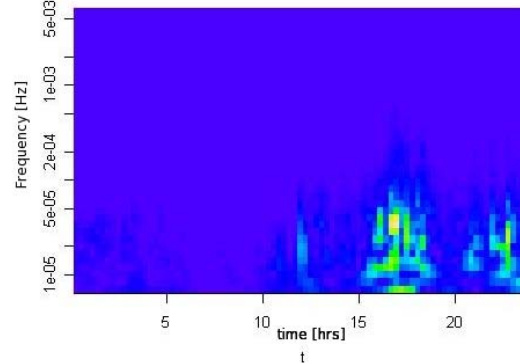
Another immediate candidate

CARISMA (formerly CANOPUS) magnetometer array

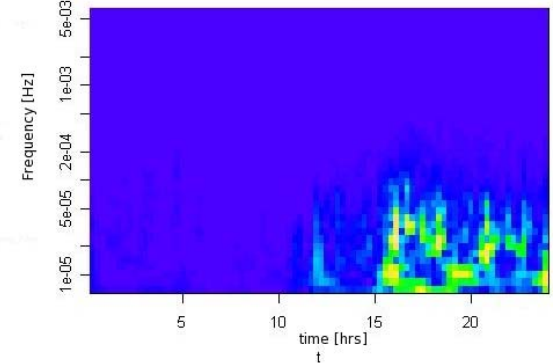
real Main PCA, day = 20070927 Hour: 119



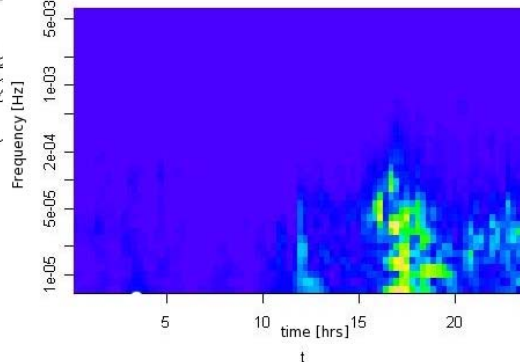
Continuous wavelet transform, day = 20070927 chronos#1



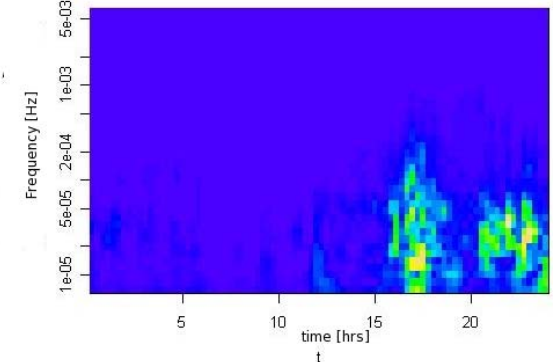
Continuous wavelet transform, day = 20070927 chronos#2



Continuous wavelet transform, day = 20070927 chronos#3



Continuous wavelet transform, day = 20070927 chronos#4



Different scales but similar multi-channel time-series data

PCAs = main flow patterns

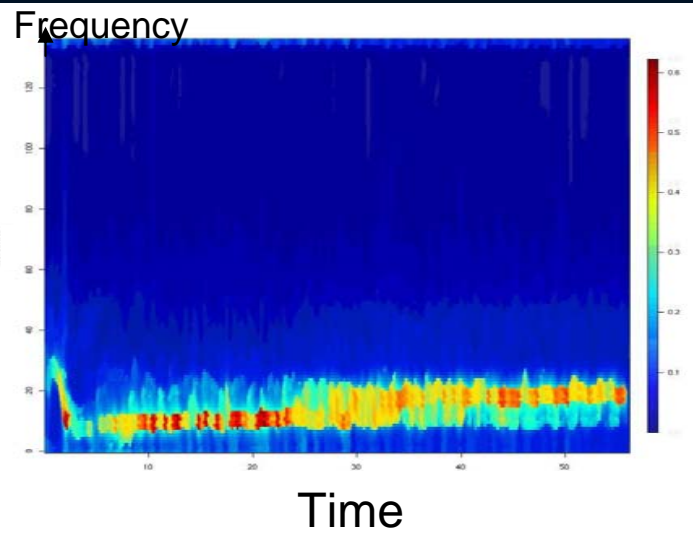
Wavelet transform = bursts, ULFs, etc.

Conclusions

Data mining does not replace the scientist
→ *be wary but not afraid*

- A proper electronic logbook increases the effective *time of useful life* of experimental data
- Established robust data reduction for multi-channel time series
- Combination with the summary logbook
(à la Data Repository → Data Warehouse)
Makes data amenable to wide range to data mining tools
- After reduction $O(1000-10000)$ results can be easily shared with colleagues
- Association rules enable to **complement** the single experiment forensics (i.e. **plasma experiments**)

Mining for information and visualization



Feature database = Transaction list

Feature tree for mining patterns

