Progress towards an MHD documentation database

Dave Pretty

9th CWGM, ANU, 28 Jan 2012

MHD Documentation database (MDDB): Overview

Use data mining to generate a "bottom-up" database to complement "top-down" databases (confinement and profile databases).

- Top-down: smaller number of carefully chosen shots.
- Bottom-up: Use machine-learning / data mining to find patterns from all shots.

MHD documentation -> provide description / metadata for different MHD activity identified.

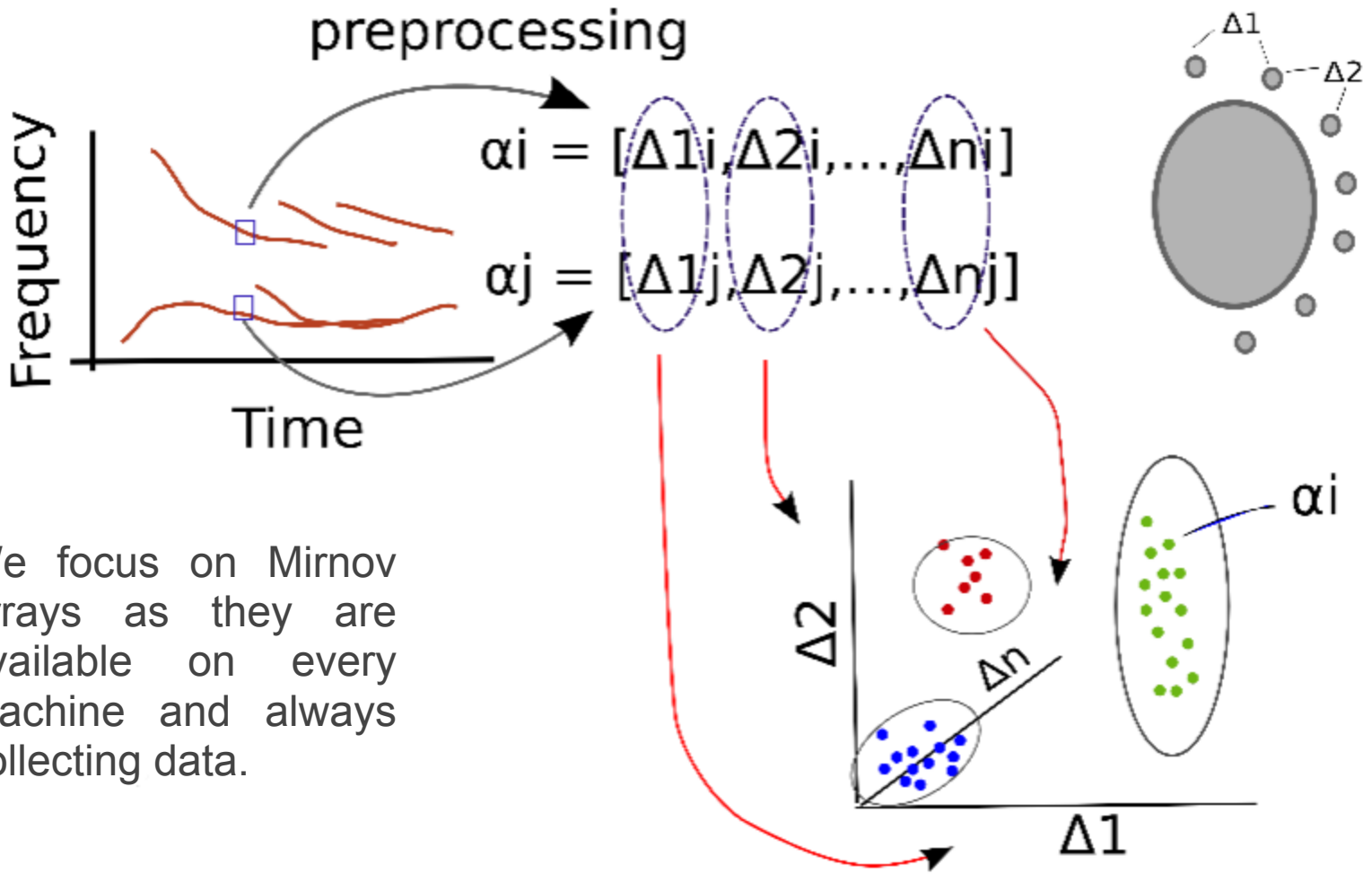
MDDB: Story so far

- Have shown that we can identify physics of MHD modes using data mining techniques: H1, HJ, TJ-II (Pretty et al, ISHW 2009)

- However, different datasets were analysed with different algorithms (improving our techniques over time).

- We have now started the process of applying the same process to all datasets, and included data from LHD and W7AS.

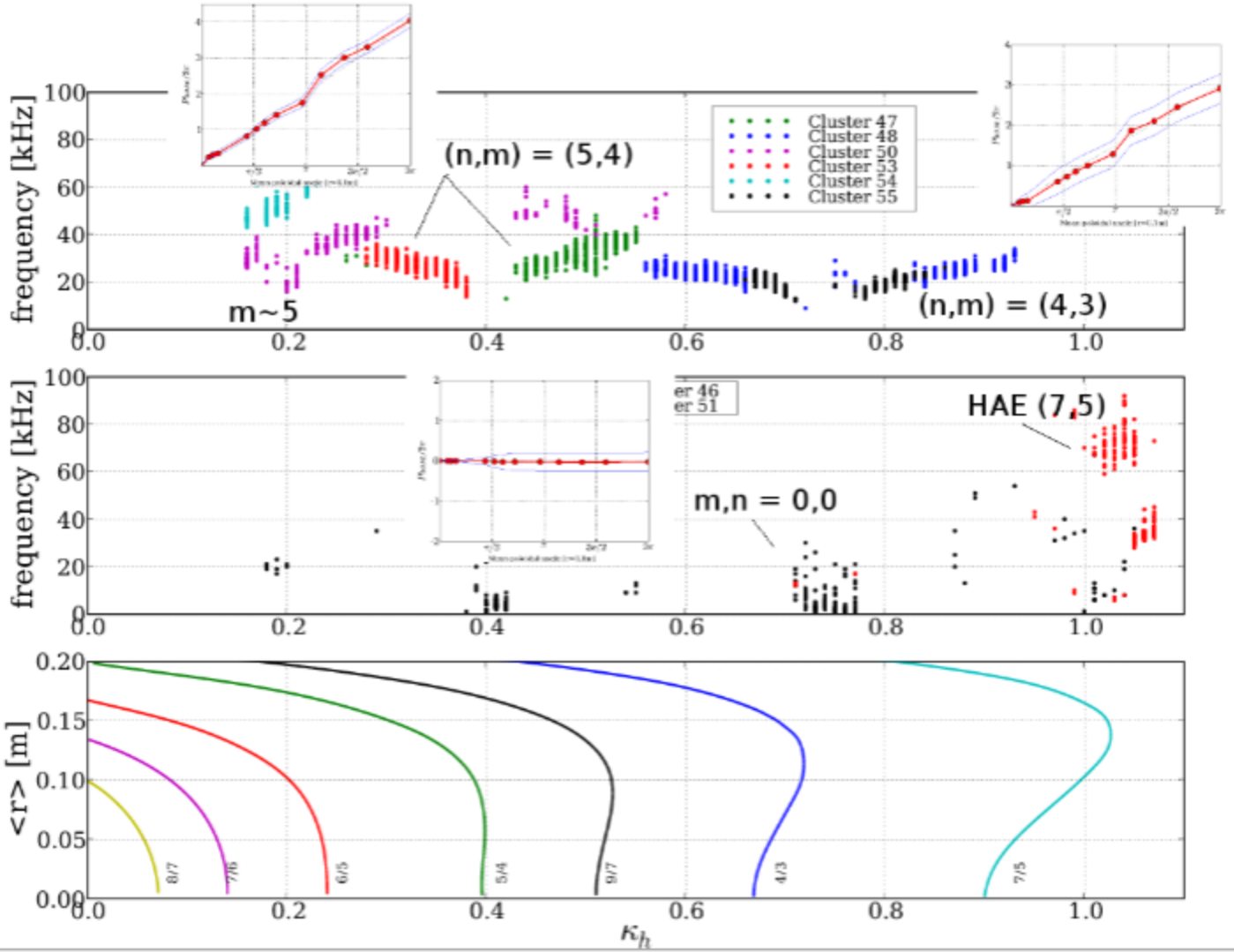
Overview of process



We focus on Mirnov arrays as they are available on every machine and always collecting data.

Example of results: H1 rotational transform scan

~100 shots, ~100 different rotational transform profiles



Target dataset for MDDB 1.0

	H-1	Heliotron J	LHD	TJ-II	W7AS
Shots	58000 - 64050	26430 - 46324	84000 - 94000	17200 - 28500	51000 - 54359
Active mirnov coils	11 from pol. array	13 from pol. array	12 from hel. array	14 from pol. array	15 from pol. array
Sample rate	1 MHz	1 MHz	1 MHz	312.5 kHz	333 kHz

Already done:

- All shots above (range of **~50,000 shots**) have been pre-processed on local servers. (Except W7AS, where pre-processing was done at ANU)
- Reduced data (processed fluctuation data) totalling **~80 Gb (compressed)** has been transferred to ANU computers
- Total number of fluctuation data points > 8x107

However...

These datasets are too big for the clustering algorithms we have been using!

Minimal LHD dataset filesize for clustering is 11 Gb.

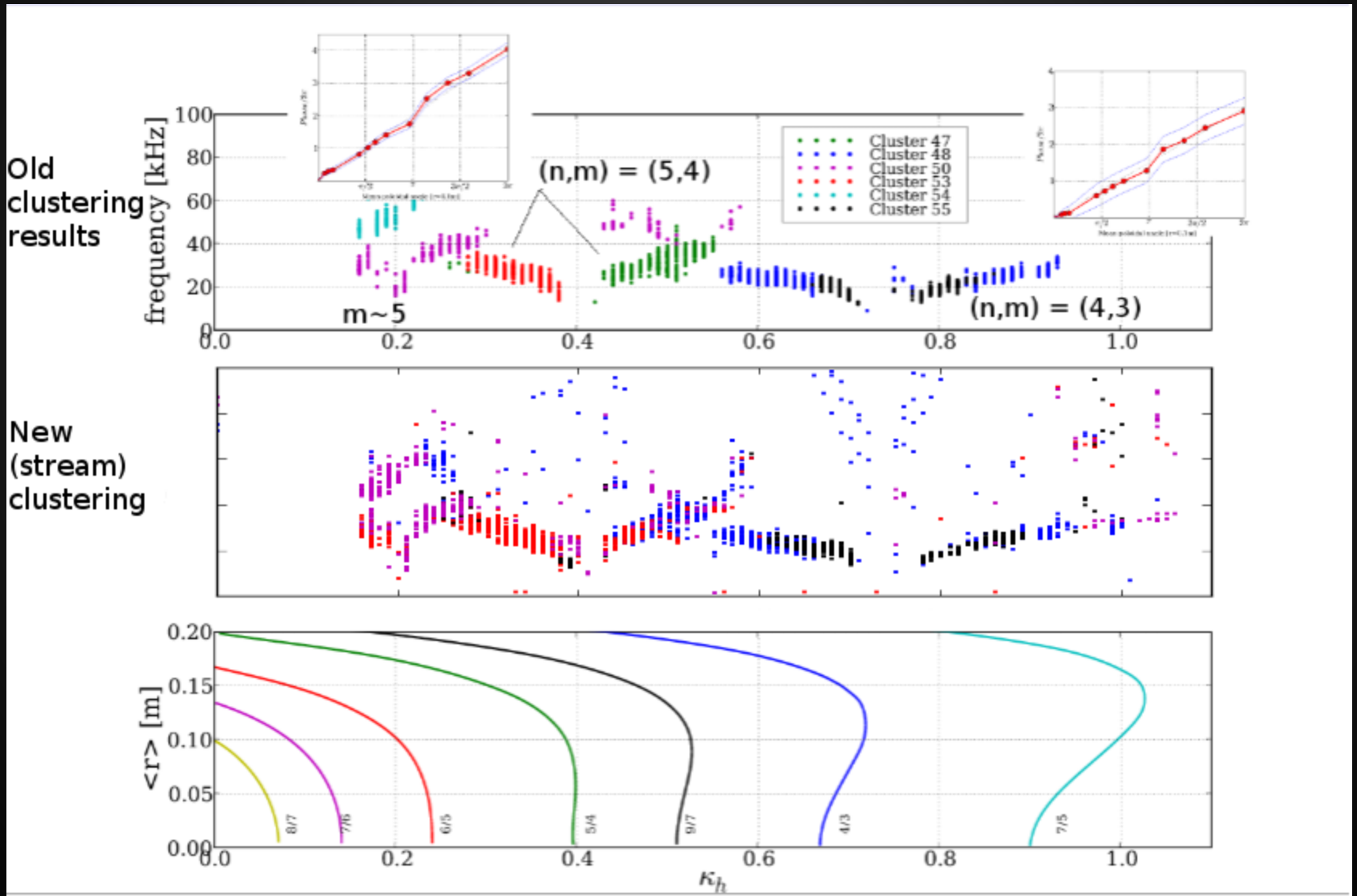
We require a method of clustering where the dataset does not have to be stored in computer memory

Solution is known as 'data stream clustering', where the data file is read and processed in sequence and not stored in memory.

Cutting-edge computer science, we are now using a stream clustering algorithm published only last month [Michael Shindler, Alex Wong, and Adam Meyerson. Fast and Accurate k-means for Large Datasets. In NIPS, 2011]

Comparison of old and new (stream) clustering results

Using standard H1 iota-scan dataset, shots 58043-58153



Proposed MDDB database structure

Mode Table

Mode ID	Channel A	Channel B	Delta phase (cluster parameters, e.g.mean, var)

Documentation Table

Mode ID	Documentation

Instance Table

Instance ID	Shot	Time (0.5 ms interval)	Frequency	Signal strength

Mapping Table

Instance ID	Mode ID	Likelihood

Supplementary Table

Instance ID	ne	B	iota	heating	etc...	etc...

Next steps

- More testing of stream clustering, until we are confident that it gives expected results.
- Do clustering - possibly using Australian supercomputer (NCI National Facility).
- Populate mode, instance and mapping tables
- Work with collaborators to populate documentation and supplementary tables.

Thanks to collaborators at Kyoto University, Ciemat, NIFS and IPP for their continued support