Machine Learning & Science Data Processing

Rob Lyon

robert.lyon@manchester.ac.uk

SKA Group
University of Manchester

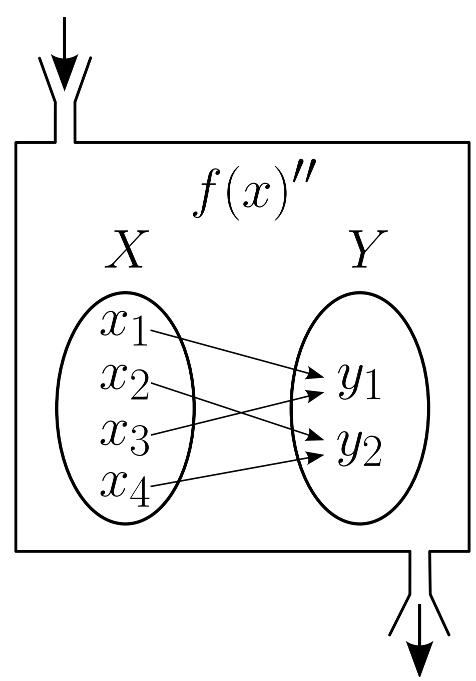
Machine Learning (1)

- Collective term for branch of A.I.
- Uses statistical tools to make decisions over data ‘intelligently’.
- Appearance of intelligence is an illusion backed up by functions.
- So how does it work?

Machine Learning (1)

- Collective term for branch of A.I.
- Uses statistical tools to make decisions over data ‘intelligently’.
- Appearance of intelligence is an illusion backed up by functions.
- So how does it work?

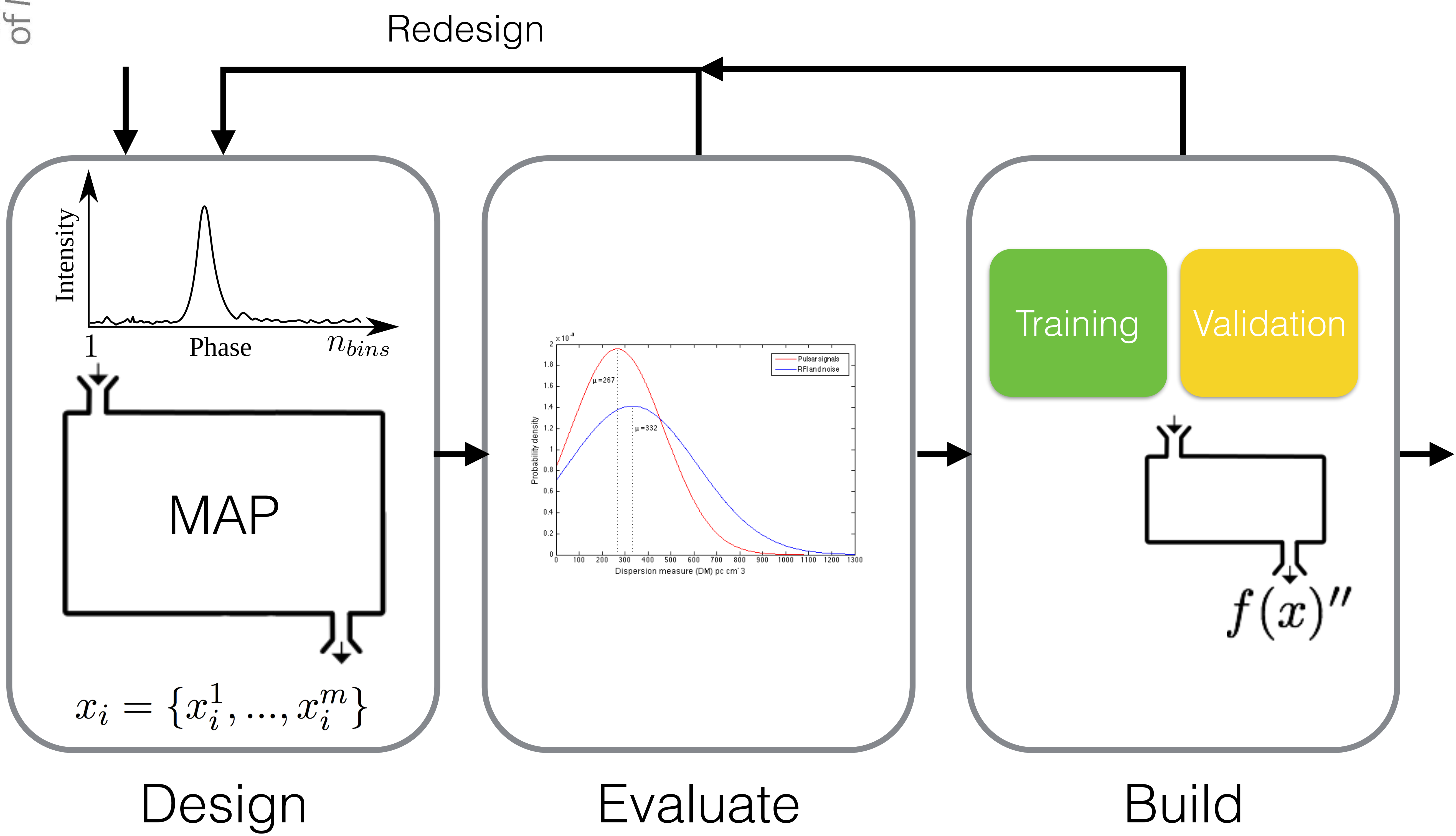
$$f(x) \longrightarrow f(x)' \longrightarrow f(x)'' \longrightarrow$$



Observe   Collect Data   Build Model

The University of Manchester

Machine Learning (2)



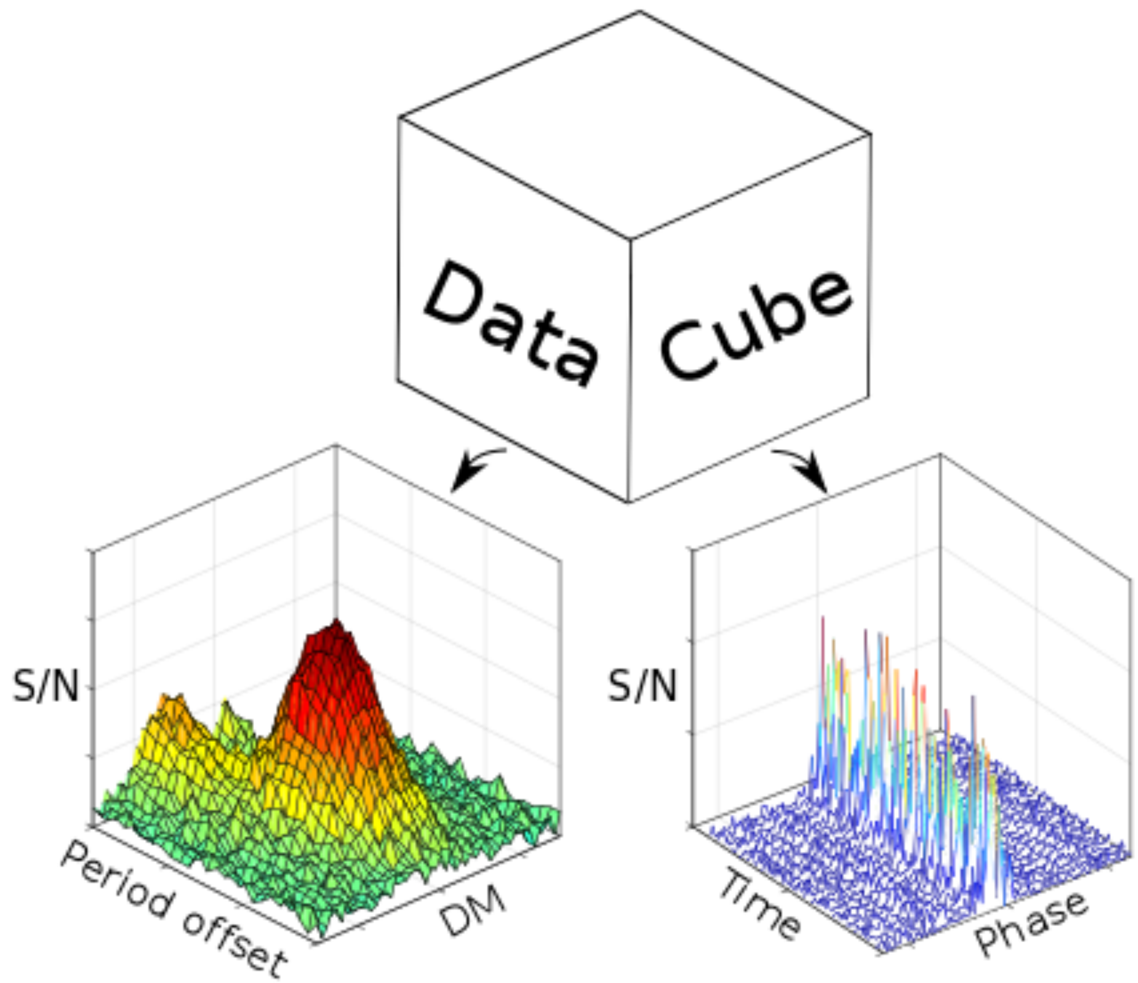
Machine Learning & SDP

Machine Learning & SDP

- SDP converts / filters CSP data in to products useful for science.

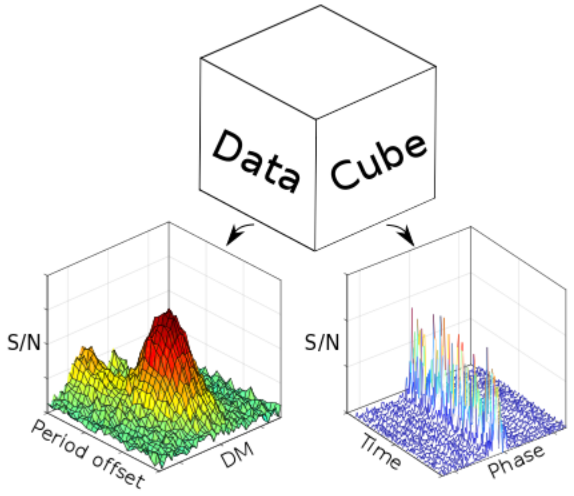
Machine Learning & SDP

- SDP converts / filters CSP data in to products useful for science.



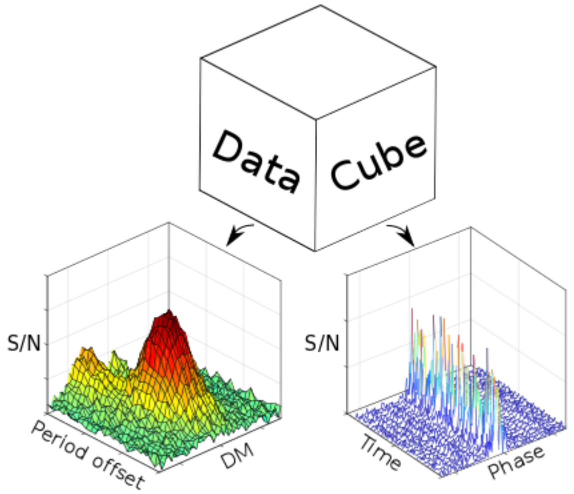
Machine Learning & SDP

- SDP converts / filters CSP data in to products useful for science.



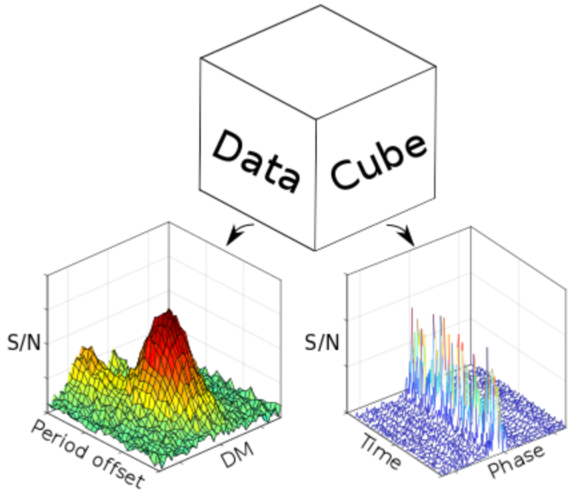
Machine Learning & SDP

- SDP converts / filters CSP data in to products useful for science.
- Includes pulsar timing, single pulse search (transients signals, FRBs) and periodicity search (pulsars).



Machine Learning & SDP

- SDP converts / filters CSP data in to products useful for science.
- Includes pulsar timing, single pulse search (transients signals, FRBs) and periodicity search (pulsars).
- For single pulse and periodicity search, CSP data products describe potential observations of astrophysical phenomena - new discoveries?

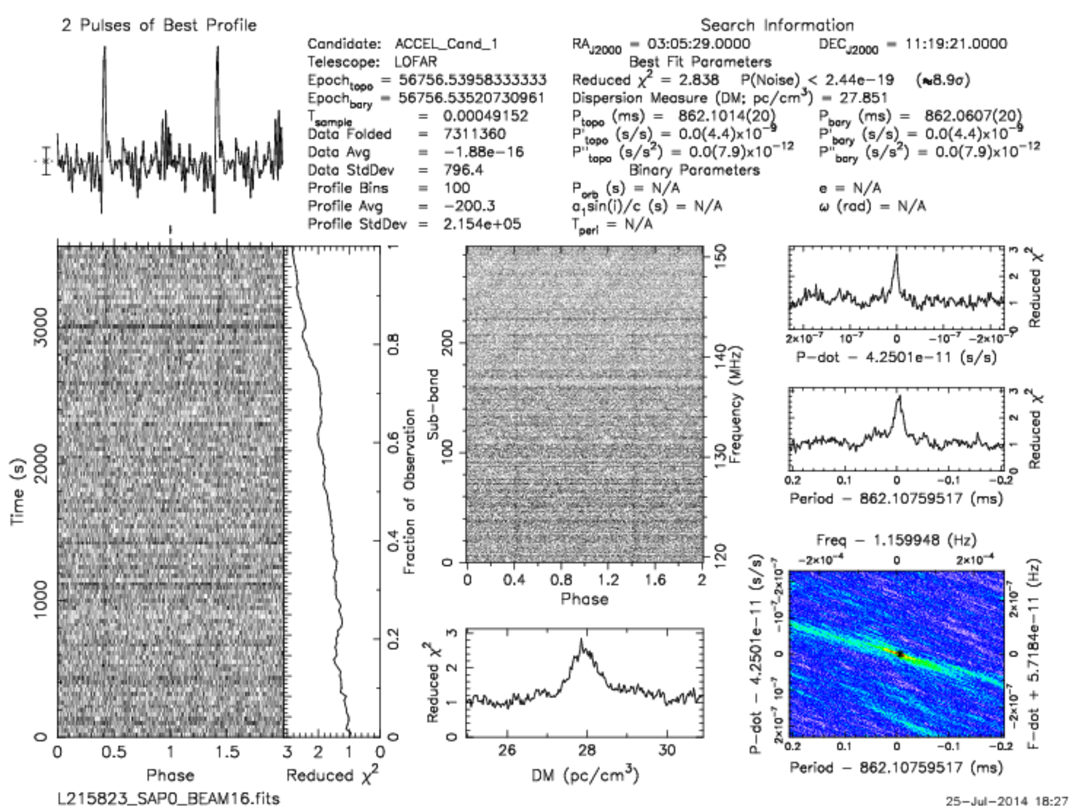


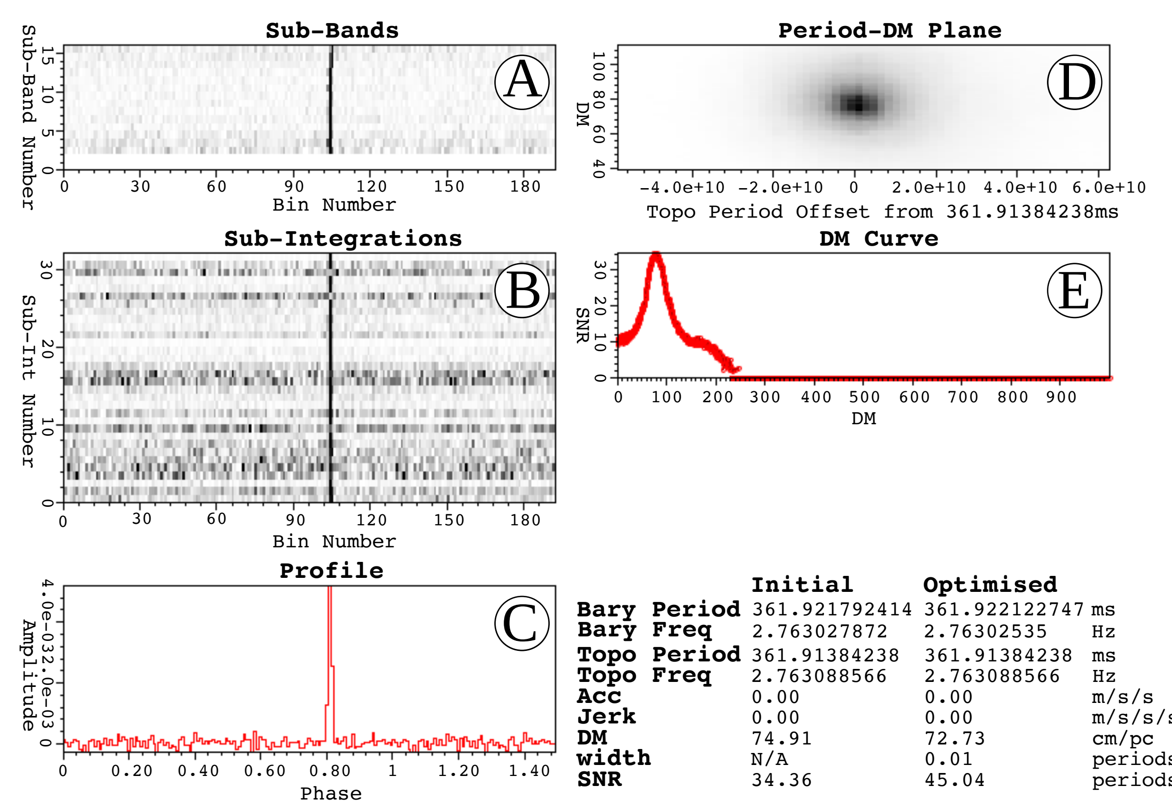
MANCHESTER 1824

The University of Manchester

Existing Approaches

- Applied to candidate selection for single pulse and periodicity search.
- Supervised machine learning algorithms.
- Learn from fixed-size training sets of examples.
- Variety of algorithms used, with varying computational requirements.





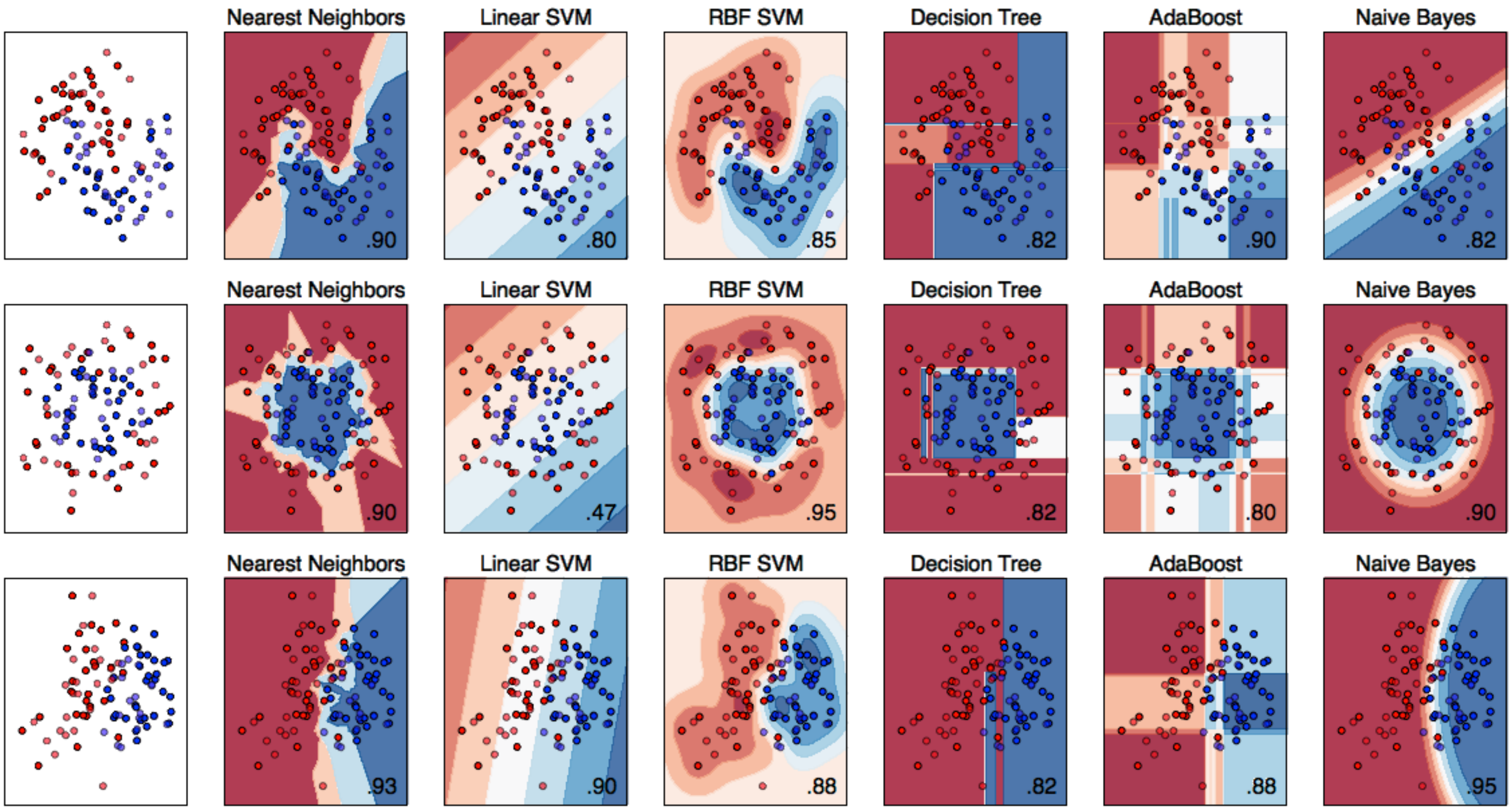
Existing Approaches

- Applied to candidate selection for single pulse and periodicity search.
- Supervised machine learning algorithms.
- Learn from fixed-size training sets of examples.
- Variety of algorithms used, with varying computational requirements.

MANCHESTER
1824

The University of Manchester

Which method?



Issues With ML at Scale

MANCHESTER 1824

The University of Manchester

Issues with ML at SKA Scale

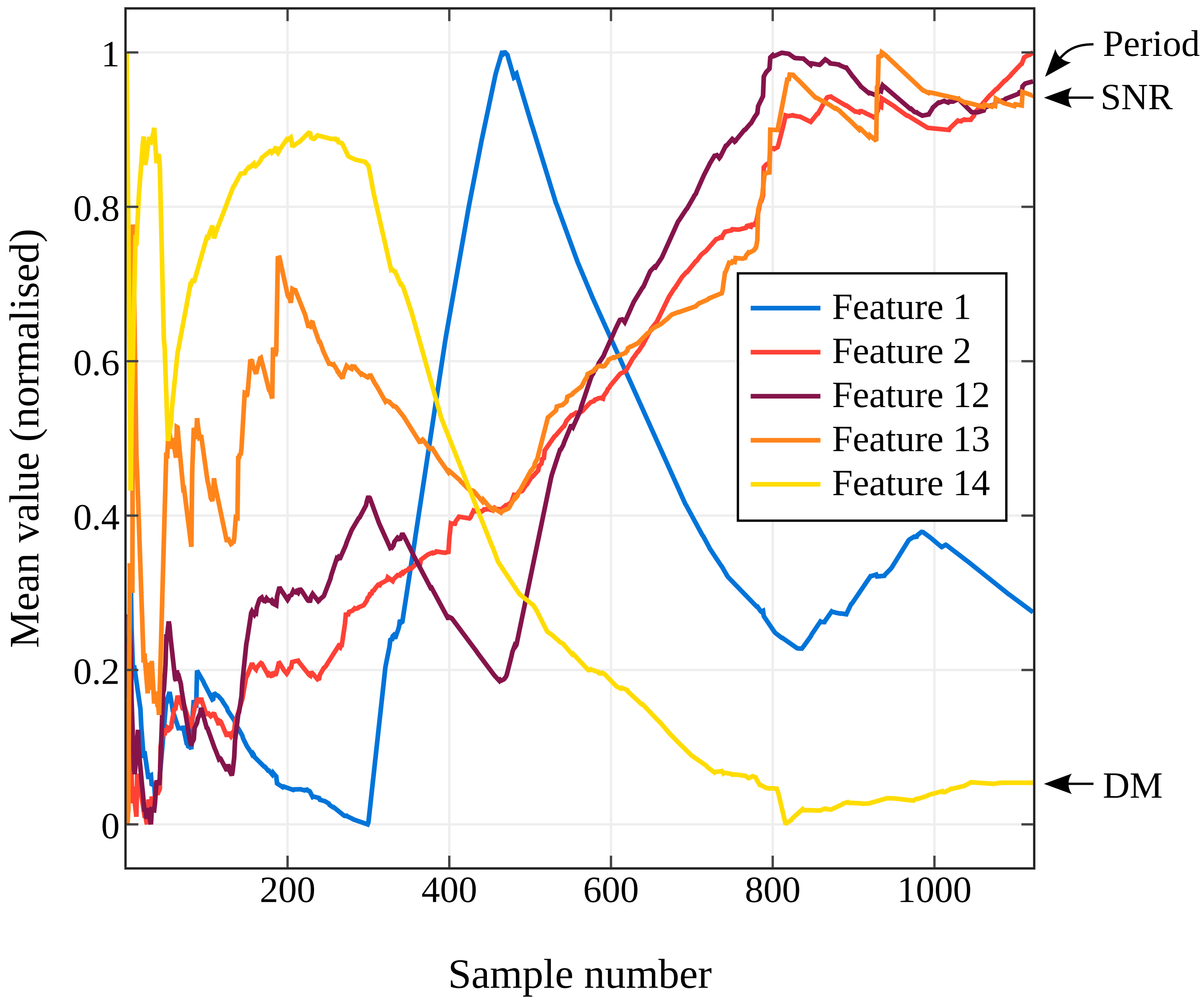
- ML typically very accurate if training data is good.
- Problems:

  1. Not optimised to minimise resource use.
  2. Non-adaptive, and retraining with more examples can be expensive (depending on the algorithm).

- Other issues: training data hard to obtain, classifier decisions often hard to audit.

Practical Issues with ML at SKA Scale (1)

- Adapting to distributional change advantageous.
- Rapidly adapting to new training examples important for discovery.

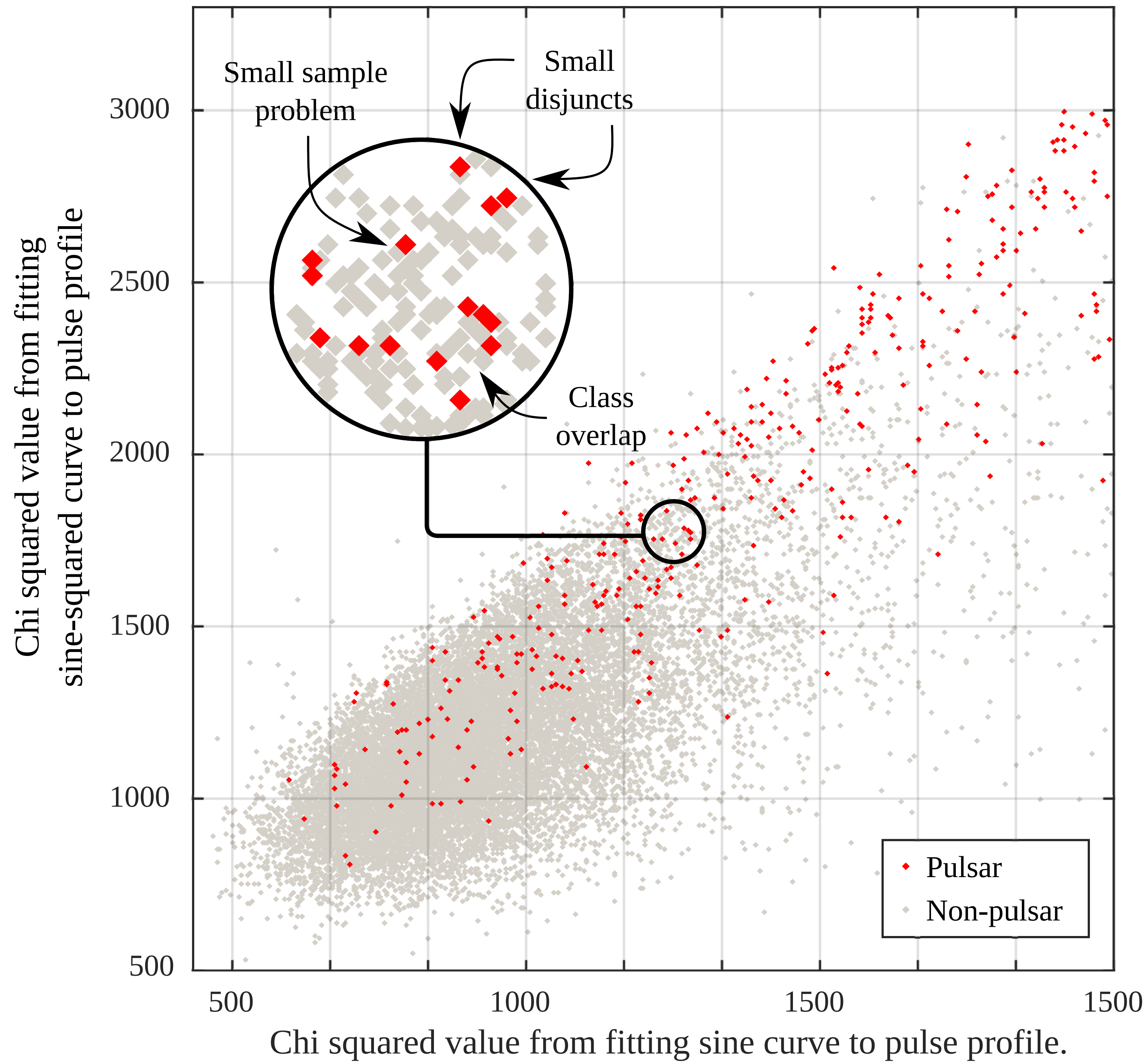


MANCHESTER 1824

The University of Manchester

Structural Issues with ML at SKA Scale (2)

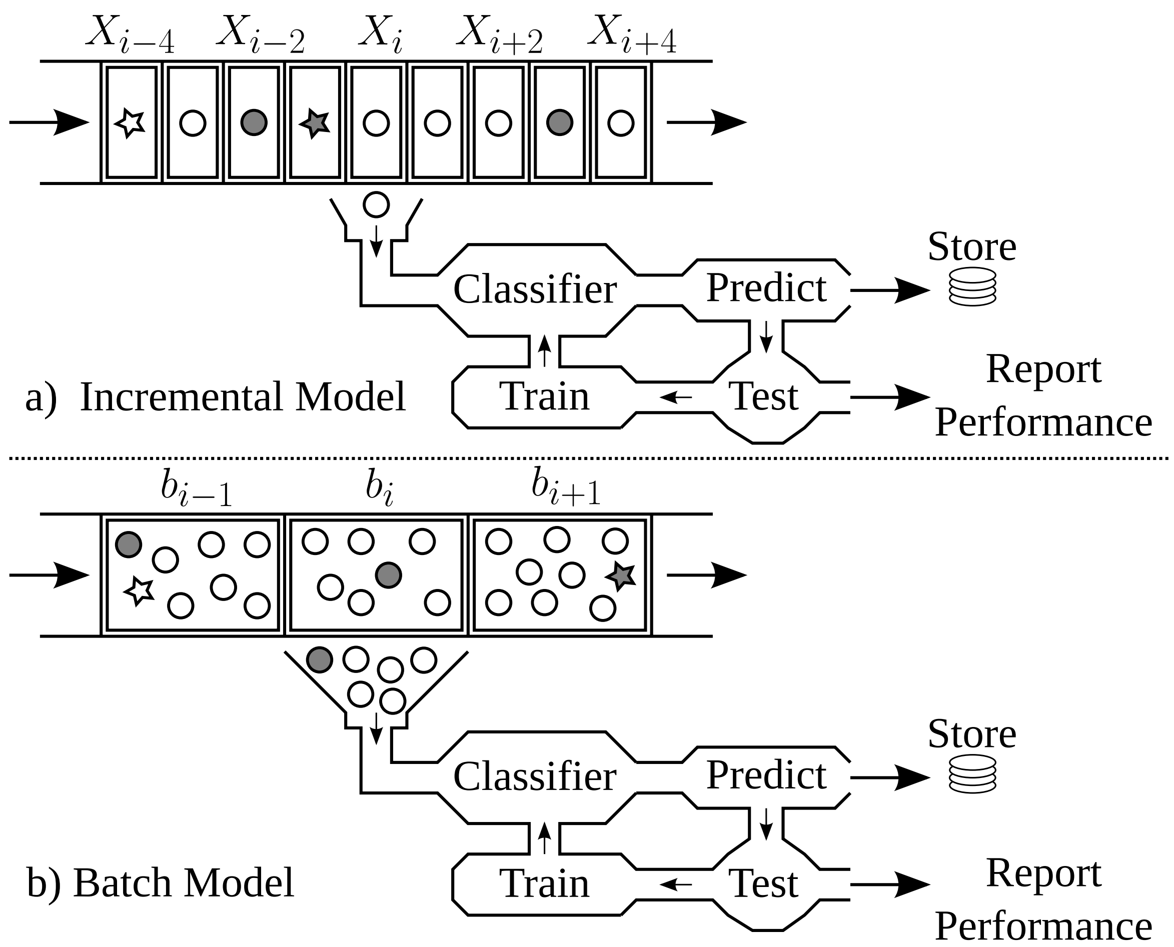
- Performance issues due to imbalance.
- How to acquire training examples?
- How to incorporate expert feedback?
- How to audit classifications?



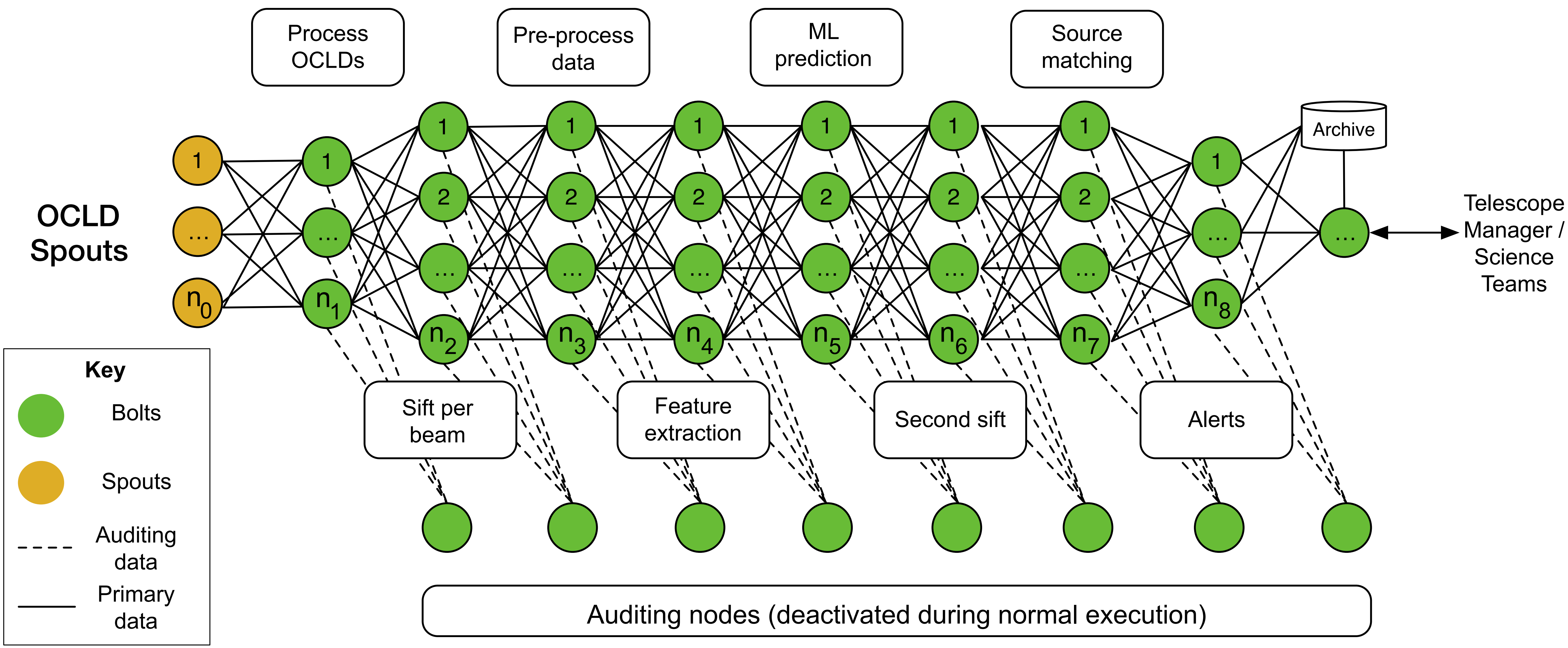
Exploring solutions

Possible SDP Approach

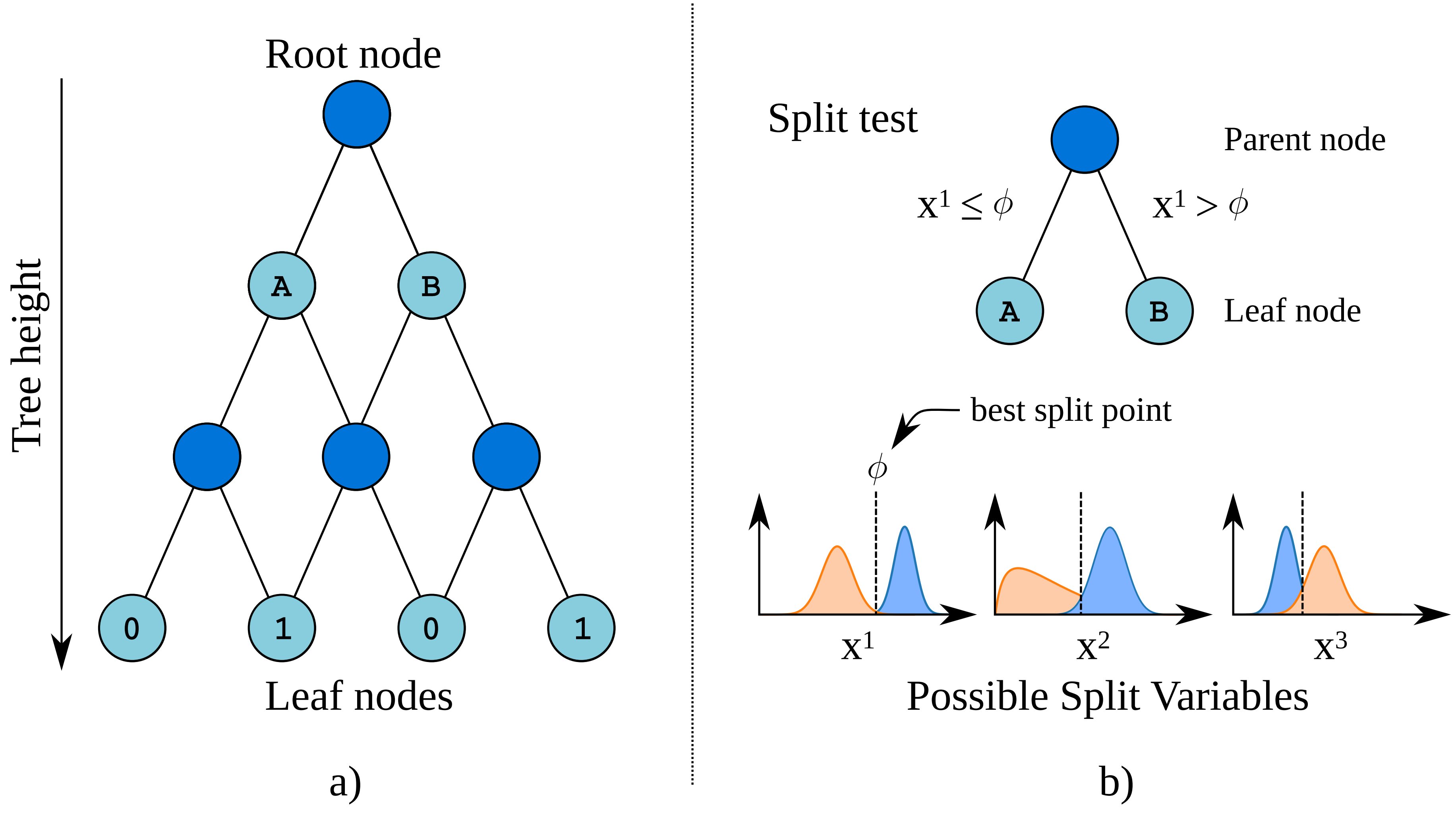
- Data stream learning methods.
- Very low resource requirements.
- Able to adapt to concept drift.
- Able to learn from new training examples observed over time.



Incremental Stream Prototype



Stream Classifier: GH-VFDT



MANCHESTER 1824

The University of Manchester

Algorithm Performance

Dataset	Algorithm	G-Mean	F-Score	Recall	Precision	Specificity	FPR	Accuracy
HTRU 1	C4.5	0.962*	0.839*	0.961	0.748	0.962	0.038	0.962
	MLP	**0.976**	0.891	**0.976**	0.820	0.975	0.025*	0.975
	NB	0.925	0.837*	0.877	0.801	0.975	0.025*	0.965
	SVM	0.967	0.922	0.947	0.898	0.988	0.012	0.984
	GH-VFDT	0.961*	**0.941**	0.928	**0.955**	**0.995**	**0.005**	**0.988**
HTRU 2	C4.5	0.926	0.740	**0.904**	0.635*	0.949*	0.051*	0.946*
	MLP	**0.931**	0.752	0.913	0.650*	0.950*	0.050*	0.947*
	NB	0.902	0.692	0.863	0.579	0.943	0.057	0.937
	SVM	0.919	0.789	0.871	0.723	0.969	0.031	0.961
	GH-VFDT	0.907	**0.862**	0.829	**0.899**	**0.992**	**0.008**	**0.978**
LOTAAS 1	C4.5	0.969	0.623	0.948	0.494	0.991	0.009	0.990
	MLP	**0.988**	0.846*	**0.979**	0.753	0.998	0.002	0.997*
	NB	0.977	0.782	0.959	0.673	0.996	0.004	0.996
	SVM	0.949	**0.932**	0.901	**0.966**	**0.999***	**0.001***	**0.999**
	GH-VFDT	0.888	0.830*	0.789	0.875	**0.999***	**0.001***	0.998*

See "*Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach*", Lyon et al, accepted for publication in MNRAS, 2016.

Other results in: "*Hellinger Distance Trees for Imbalanced Streams*", Lyon et al., ICPR, 2014.

Prototype Performance

- Local tests on a single machine (Quad Core i7)
- 1,000 candidates per second
- Approx. 2 seconds for a candidate to move through the system.
- Relatively easy to configure & program.
- Possible problems:
    - you may want to send more than a tuple.
    - you may want data to go backwards.

Open Questions

- How to acquire unlimited supply of accurately labelled data?
- To what extent does pulsar data drift?
- Will a multi-class approach improve classification performance?
- What will the final compute environment look like?
- How do we keep track of training data and validate our approaches?
- Are our features good enough?

MANCHESTER
1824

The University
of Manchester

Summary

- Structural issues with ML at scale
- Practical issues with ML at scale
- Success depends on understanding the data
- Prototype pipelines under development

Thanks for listening!

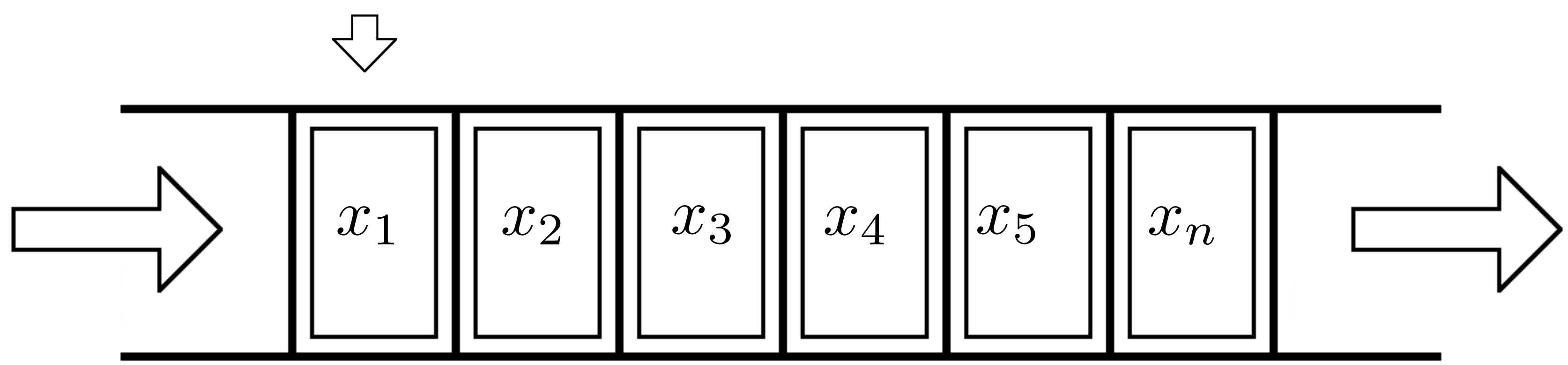
robert.lyon@manchester.ac.uk

Candidate Numbers

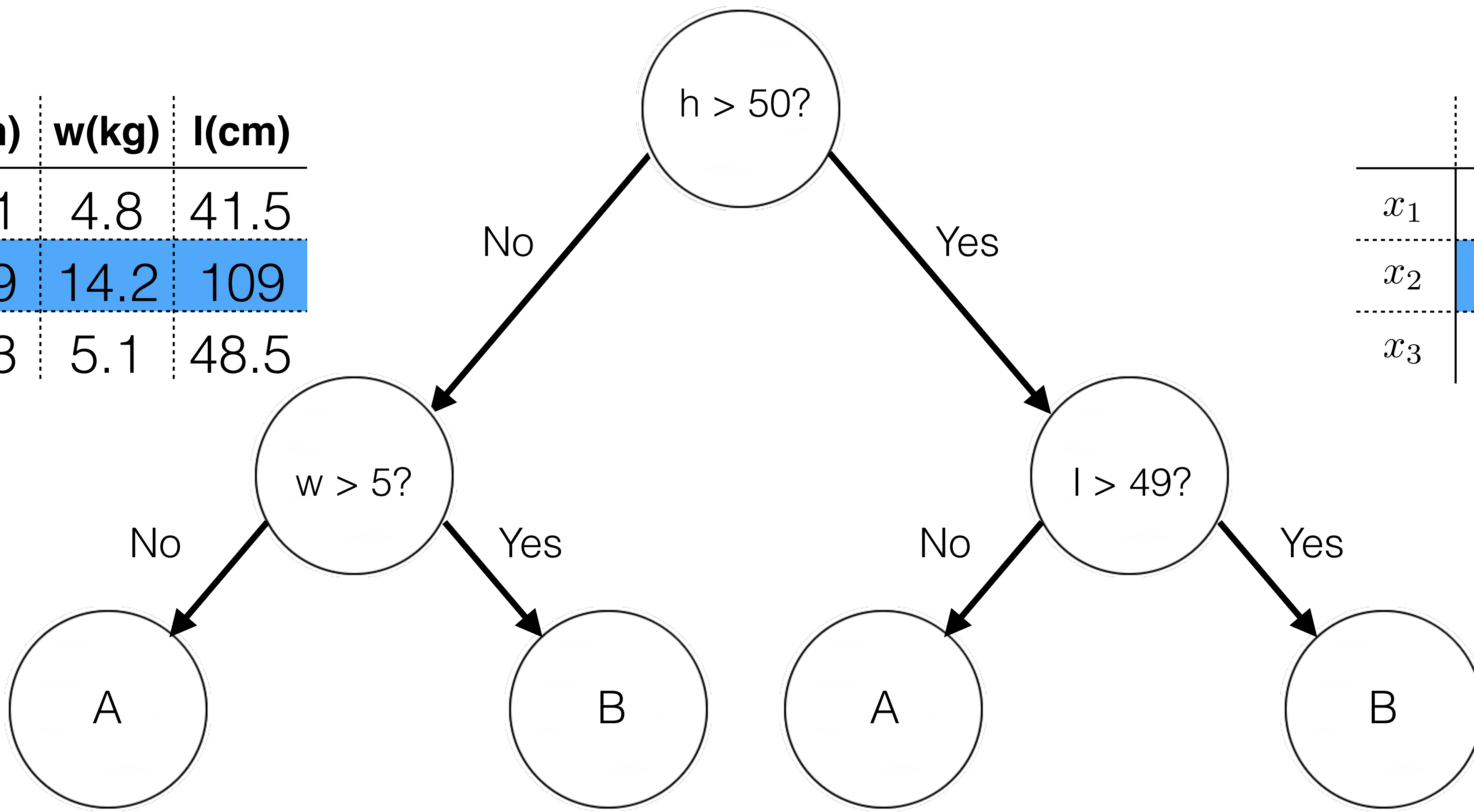
Survey	Year	Candidates	Per Sq. Degree
2nd Molonglo Survey(Manchester et al. 1978)	1977	2, 500	~0.1
Phase II survey (Stokes et al. 1986)	1983	5, 405	~1
Parkes 20 cm survey (Johnston et al. 1992)	1988	~ 150, 000	~188
Parkes Southern Pulsar Survey (Manchester et al. 1996)	1991	40, 000	~2
Parkes Multibeam Pulsar Survey (Manchester et al. 2001)	1997	8, 000, 000	~5,161
Swinburne Int. Lat. Survey (Edwards et al. 2001)	1998	> 200, 000	~168*
Arecibo P-Alfa all configurations (Cordes et al. 2006; Lazarus 2012; P-Alfa Consortium 2015)	2004	> 5, 000, 000	~16,361*
6.5 GHz Multibeam Survey (Bates et al. 2011a; Bates 2011)	2006	3, 500, 000	~77,778 †
GBNCC survey (Stovall et al. 2014)	2009	> 1, 200, 000	~89*
Southern HTRU (Keith et al. 2010)	2010	55, 434, 300	~1,705
Northern HTRU (Barr et al. 2013; Ng 2012)	2010	> 80, 000, 000	~2,890*
LOTAAS (Cooper, private communication, 2015)	2013	39, 000, 000	~2,000

Table 1. Reported folded candidate numbers. Note * indicates a lower bound on the number of candidates per square degree, calculated from incomplete candidate numbers. † indicates very long integration times, with further details supplied in Tables 2 & 3.

Tree Classification Example



	h(cm)	w(kg)	l(cm)
x_1	32.1	4.8	41.5
x_2	67.9	14.2	109
x_3	35.3	5.1	48.5



	Predictions
x_1	A
x_2	B
x_3	A

The University of Manchester

Apache Storm Cluster