

Classification in Time Domain Astrophysics

Dr. Rob Lyon
University of Manchester (SKA Group)
robert.lyon@manchester.ac.uk



@scienceguyrob

Collaborators

- **John Brooke (retired)**
- **Sally Cooper (University of Manchester)**
- **DRAGNET team (Hessels et. al.)**
- **Joshua Knowles (University of Birmingham)**
- **Lina Levin-Preston (University of Manchester)**
- **Mitchell Mickaliger (University of Manchester)**
- **Ben Stappers (University of Manchester)**
- **Chia Min Tan (University of Manchester)**
- **SKA Group Time Domain Team (Oxford, INAF)**

Talk Overview

- Candidate classification problem
- Characteristics
- Data
- Constraints
- Possible approaches & current solution

Paper

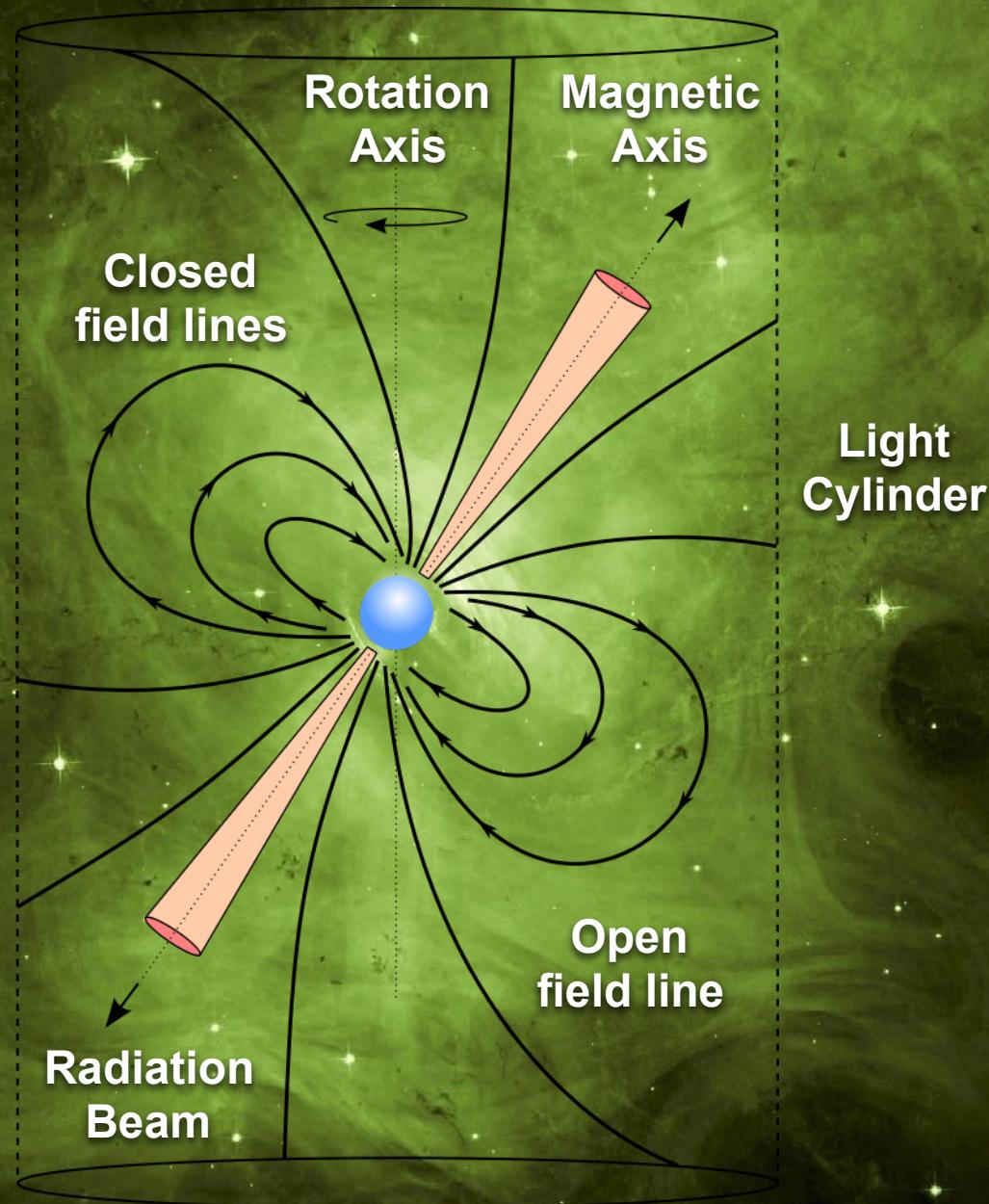
R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles,
MNRAS, 459 (1): 1104-1123, 2016. DOI:10.1093/mnras/stw656.

Candidate Selection

- Goal: automatically categorise signals
- A classification problem
- Heavily imbalanced data
- Encountered on-line and off-line
- Data non-stationary
- Input data not i.i.d
- Data labelling costly
- Real-time constraints
- Conceptually simple, difficult in practice!

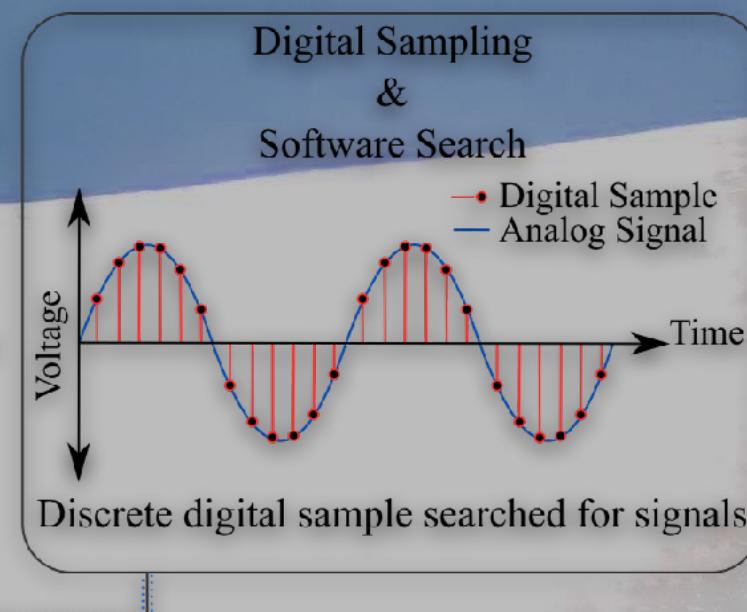
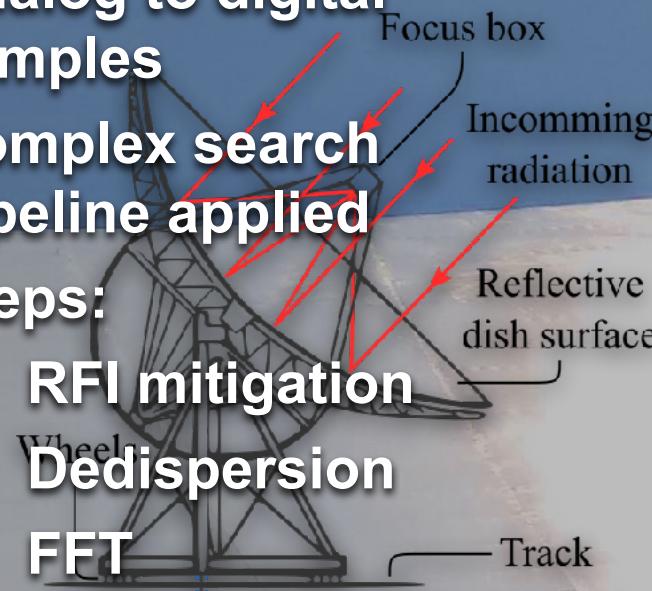
Pulsars

- Stellar remnants
- Very dense
- ~20 km diameter
- Produce radio emission
- Very useful for science



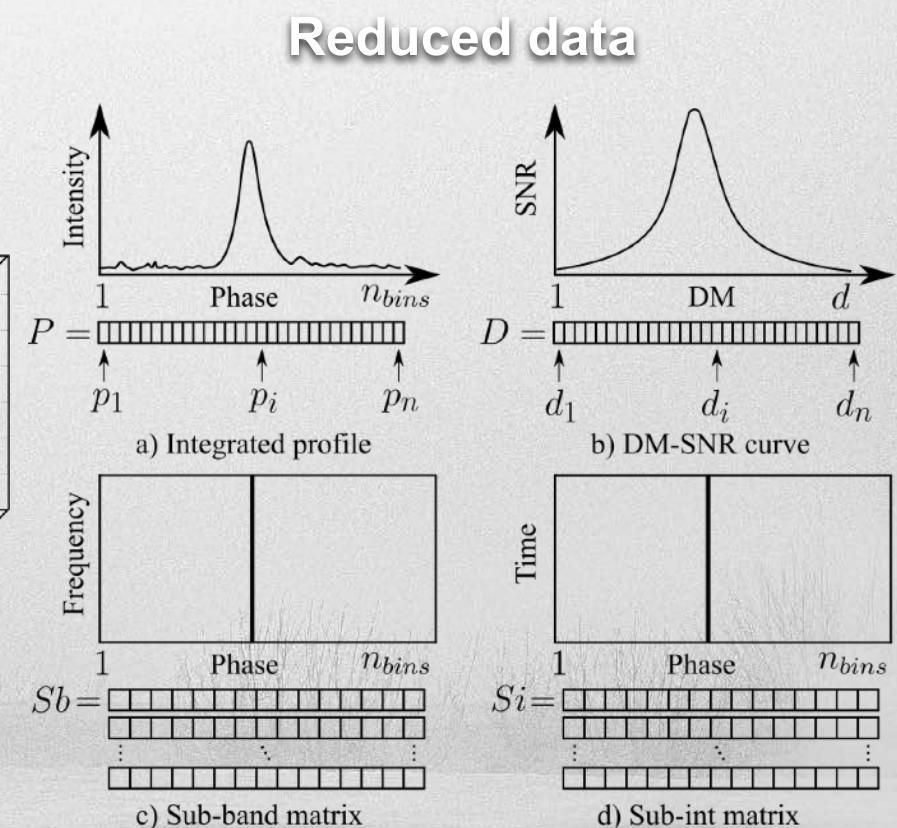
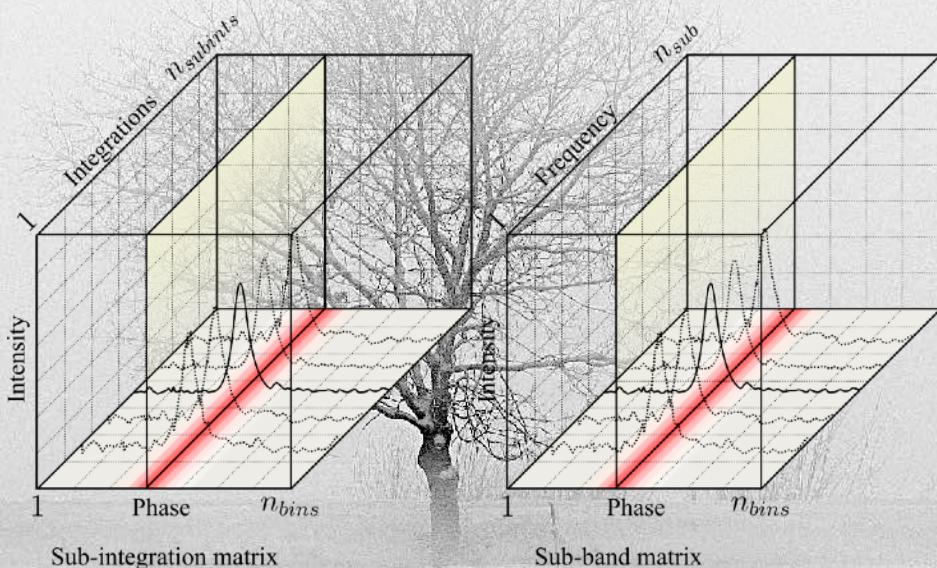
Data Capture

- Analog to digital samples
- Complex search pipeline applied
- Steps:
 - RFI mitigation
 - Dedisperion
 - FFT
 - Harmonic summing
 - Detection
 - Sifting



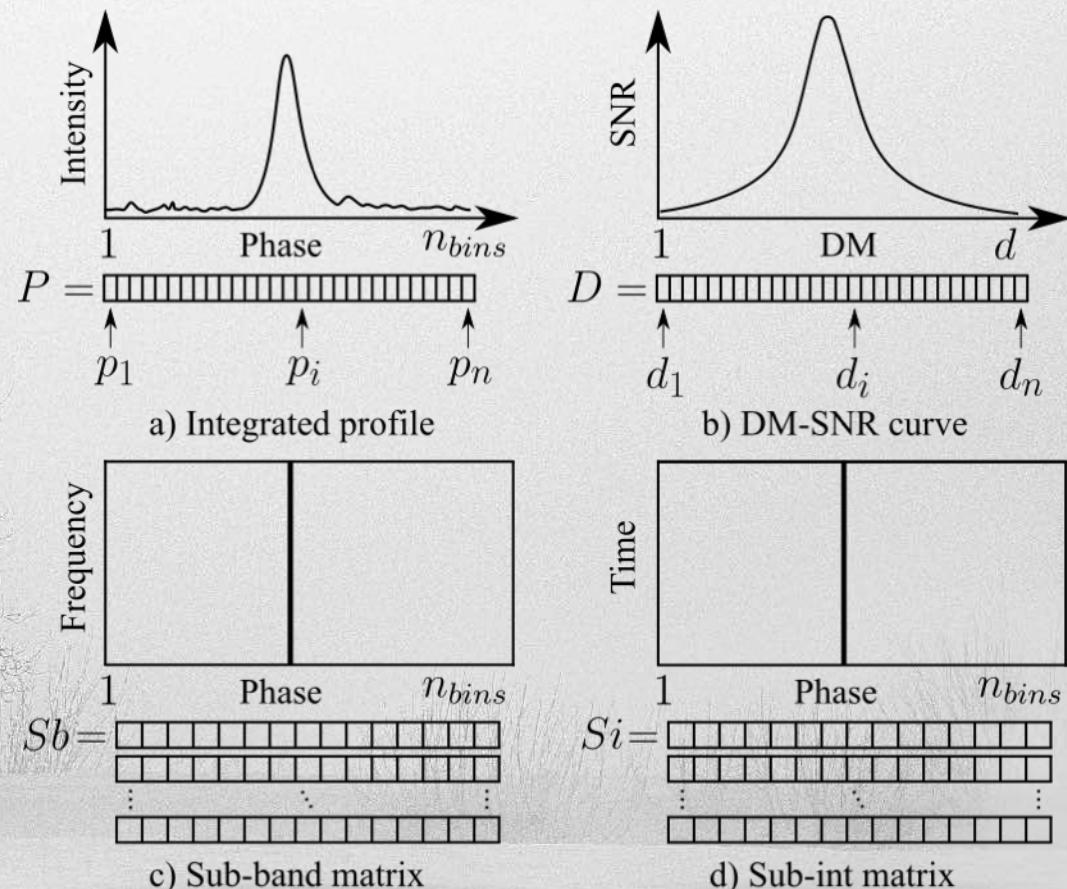
Data Products (1)

Data produced via signal processing



Data Products (2)

- A candidate file
- Describes a signal detection
- Covers time and frequency space
- At least one per detection (duplicates!)
- Classification appears simple - things do get fuzzy!



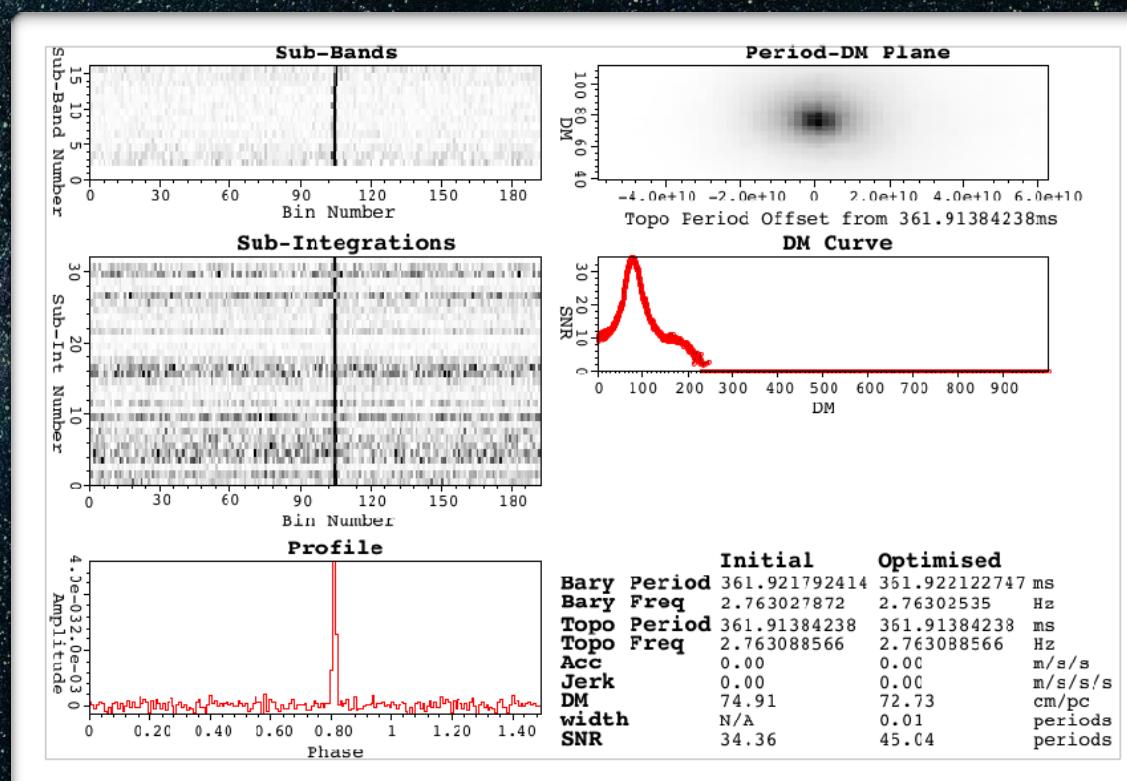
What to look out for

- **Clear pulse with a defined peak**
- **Evidence that the signal persists in time, i.e. an indication of a periodic source**
- **Evidence that the signal persists in frequency - pulsars are broadband emitters**
- **A DM value greater than zero**
- **Evidence of other effects, such as scintillation?**



Candidate Examples (1)

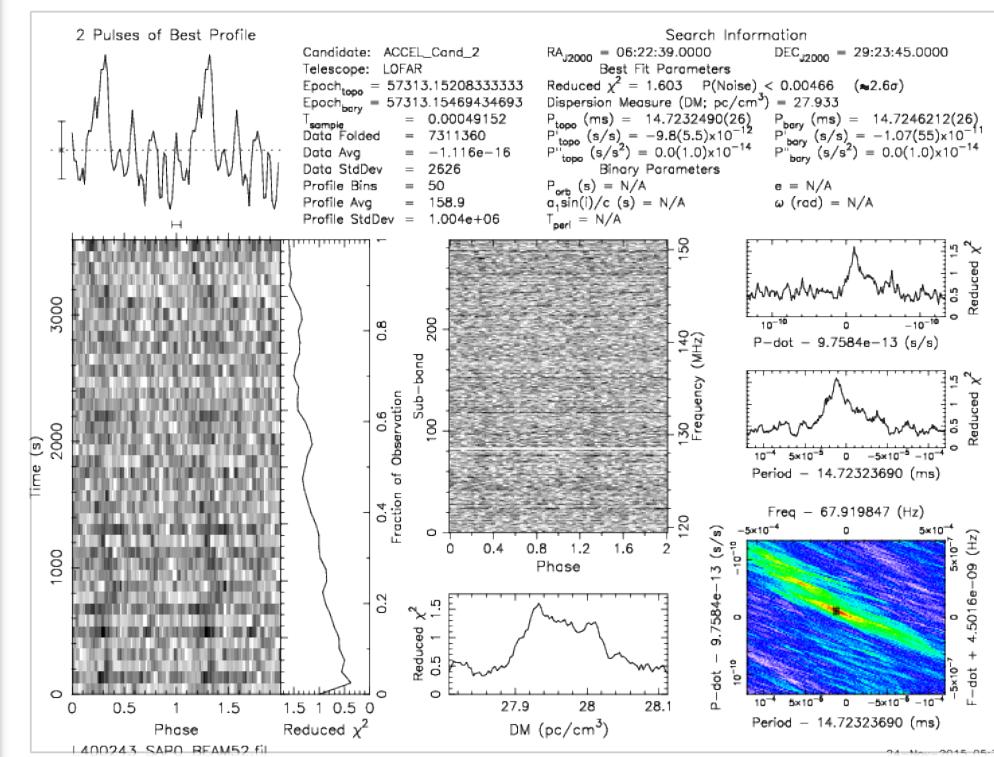
- **Defined peak**
- **Clear DM-SNR peak**
- **Consistent in time**
- **Persistent in frequency**
- **Obvious pulsar!**



Credit: HTRU Collaboration.

Candidate Examples (2)

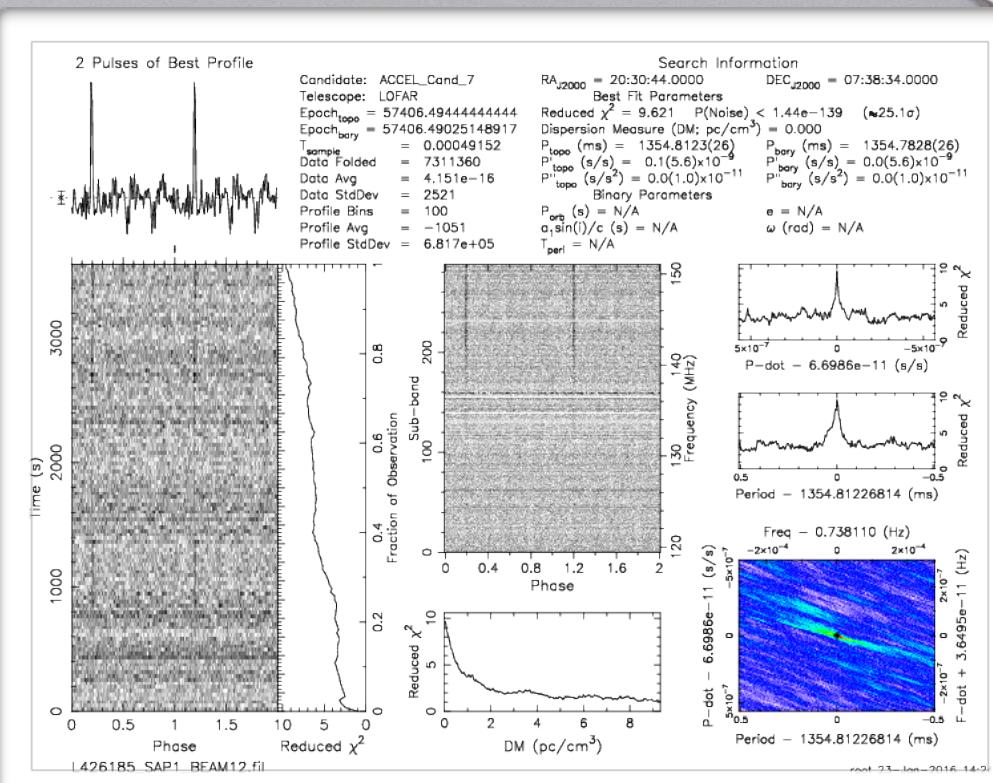
- Clearly defined peak?
- Some persistence in time?
- A DM value > zero
- Labelled as noise



Credit: LOTAAS Collaboration (Chia Min Tan et. al.).

Candidate Examples (3)

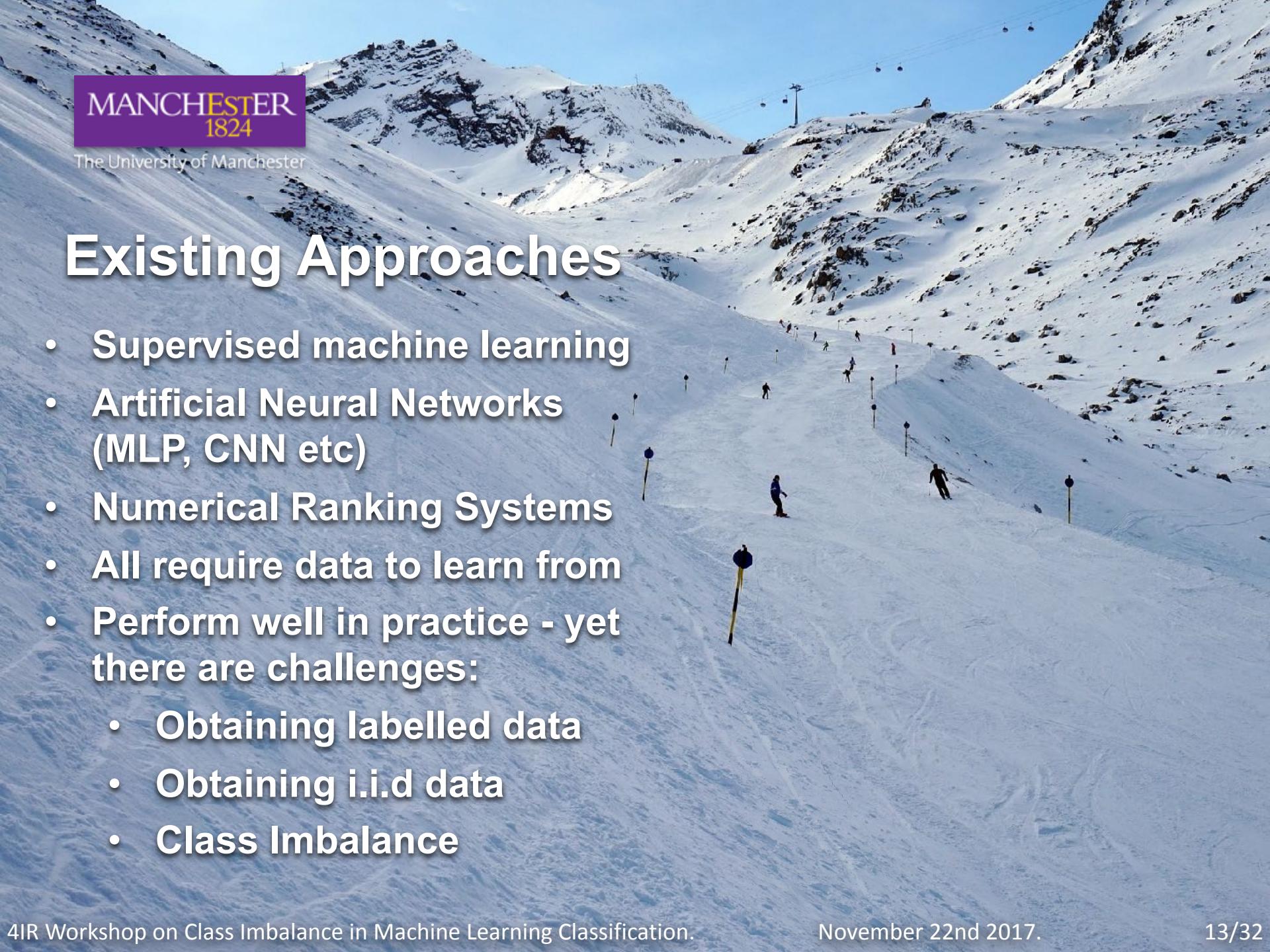
- Clearly defined peak?
- Some persistence in time?
- Some persistence in frequency?
- RFI



Credit: LOTAAS Collaboration (Chia Min Tan et. al.).

Existing Approaches

- Supervised machine learning
- Artificial Neural Networks
(MLP, CNN etc)
- Numerical Ranking Systems
- All require data to learn from
- Perform well in practice - yet there are challenges:
 - Obtaining labelled data
 - Obtaining i.i.d data
 - Class Imbalance



Labelling

- Data difficult to acquire & process.
- Fundamentally hard to assign labels.
- Process is time consuming.
- Humans prone to error given the above.
- Ground truth cannot be verified without physically pointing a telescope at a candidate - expensive!!
- Data sharing not widespread - but data from different telescopes not necessarily compatible...

i.i.d. Assumption

- Feature distributions can change, causing a drop in accuracy due to i.i.d violation.
- Mostly occurs when training data differs from real-word data.
- Also occurs when data is subject to distributional drift.

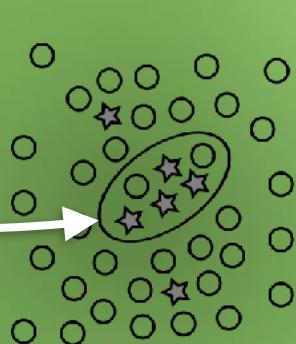


Class Imbalance (1)

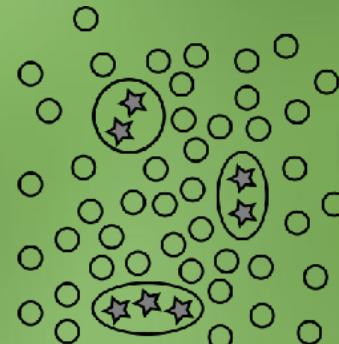
- Class ratios are skewed toward one or more category
- Pulsar search ratio is 1 pulsar to 10,000 non-pulsar examples (not a worse case)
- Imbalance can be intrinsic or extrinsic
- Given that class distributions drift over time, we have a difficult problem
- What are it's characteristics? What does the imbalance actually do?

Class Imbalance (2)

Class overlap

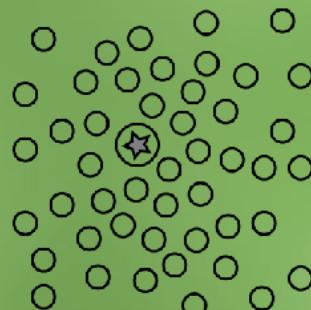


a)



b)

Absolute
rarity



c)

Disjuncts



Potential Solutions?

- Generate artificial examples to train on?
- Random over or under-sampling?
- Ditch binary classification, try anomaly/outlier/novelty detection, one-class learning, semi-supervised learning?
- Must be careful - this is a discovery discipline, and we can't introduce bias.
- What about unknowns?

Recap

- Goal is to classify pulsar candidates
- Multiple classes under consideration
- Classification difficult even for experts
- Manual classification costly & time consuming
- Turned to machine learning for help
- Progress made but things are about to change!

Square Kilometre Array

Two telescopes, one **HUGE** processing challenge.

- Changes the classification problem
- Greatly increases data capture rate
- Requires real-time computation
- Introduces a number of constraints



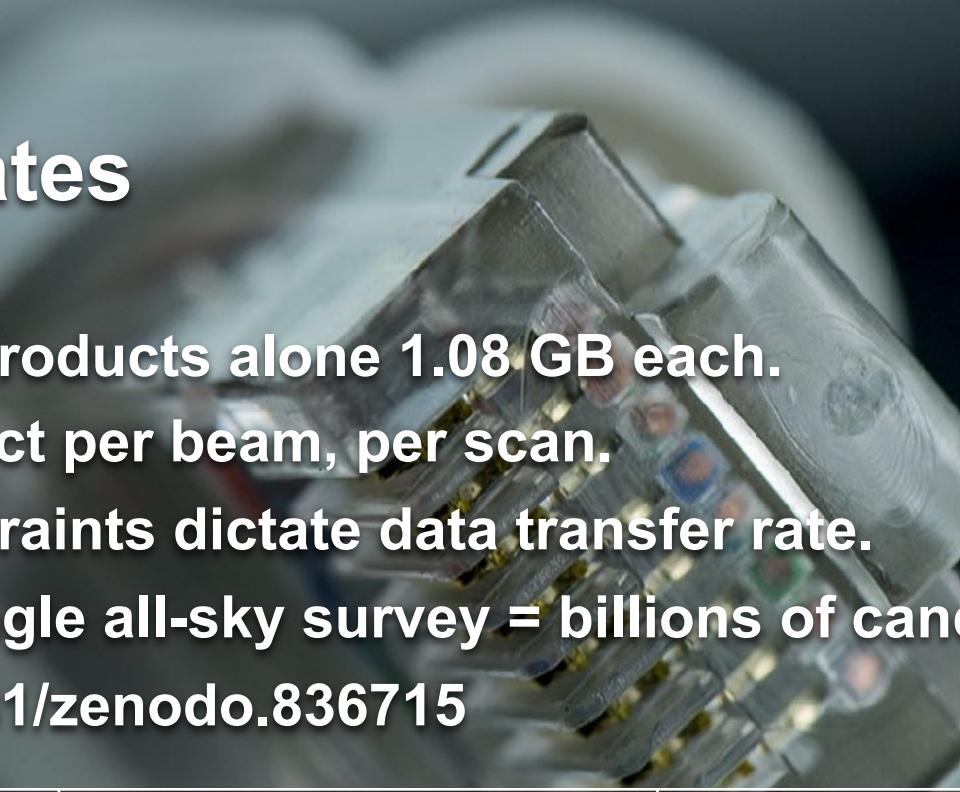
Credit: SKAO

Constraints

- **Computational constraints** - must be resource efficient
- **Real-time constraints** - must process the data quickly
- **Storage constraints** - only a portion of the data can be stored
- **Architectural constraints** - must run our systems on a massive heterogeneous computing resource, the SDP
- **Archiving constraints** - must archive our data with along with metadata, allowing auditing and tracking
- **Operational constraints** - the system must be tuneable by users whilst executing autonomously
- Then there's cost, scalability, maintainability....

SKA Data rates

- Enormous.
- Reduced data products alone 1.08 GB each.
- One data product per beam, per scan.
- Real-time constraints dictate data transfer rate.
- ~50 PB for a single all-sky survey = billions of candidates.
- See DOI: [10.5281/zenodo.836715](https://doi.org/10.5281/zenodo.836715)



	SKA-Low (500 beams)	SKA-Mid (1500 beams)
Data per beam (GB)	~1.08	~1.08
Data per scan (TB)	~0.54	~1.62
Approx. Data rate Gbit/s	~43.22	~130.66

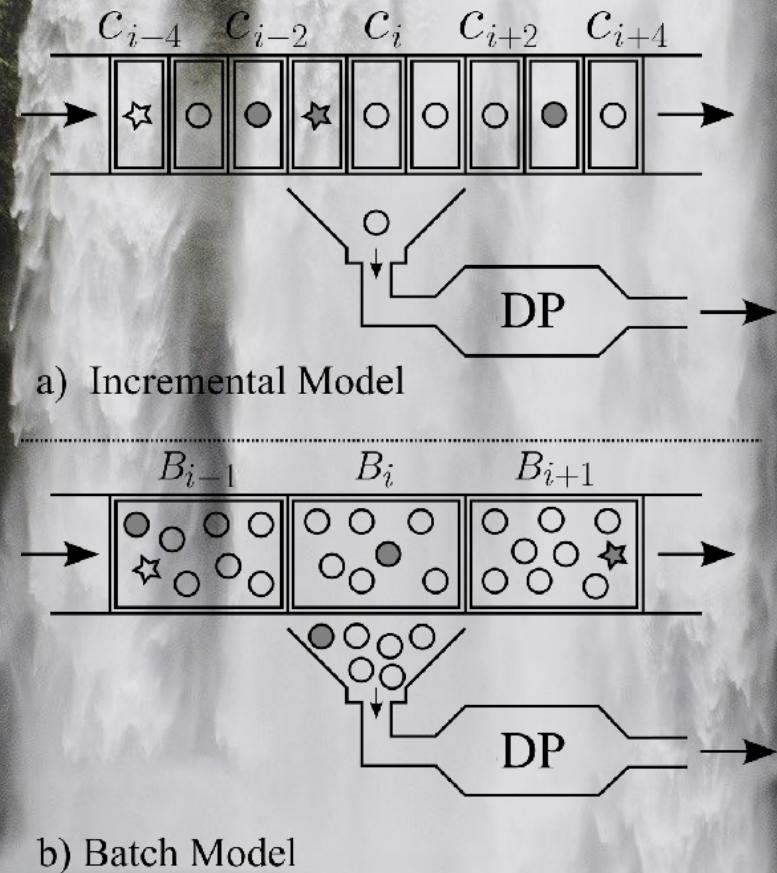
Real-time Constraints

- Data storage restrictions mean we can't store everything.
- For pulsars we must process data as quickly as possible.
- For transients, real-time processing important for rapid followup.
- Must make a decision quickly.
- Also constrained by computational power available.

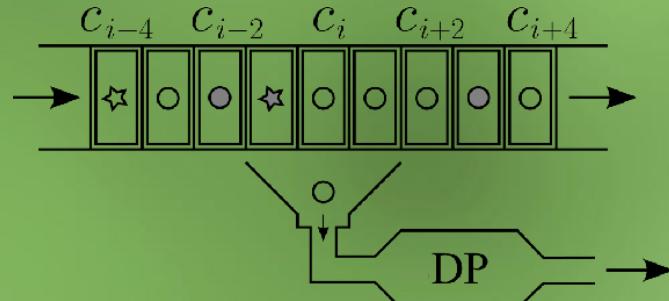


Steaming Data

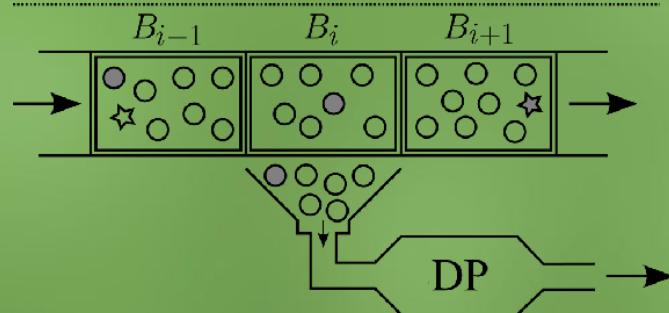
- **High data rates and real-time constraints, require a streaming model.**
- **Makes traditional off-line learning impractical.**



Class Imbalance Worsens

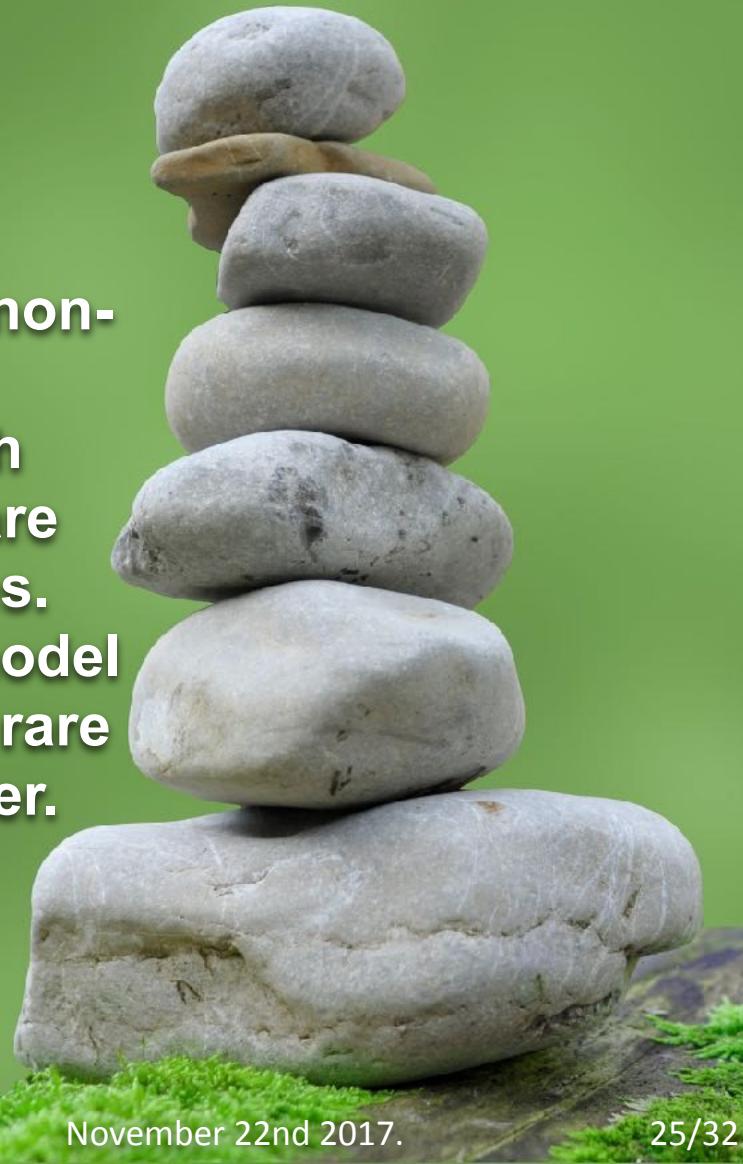


a) Incremental Model



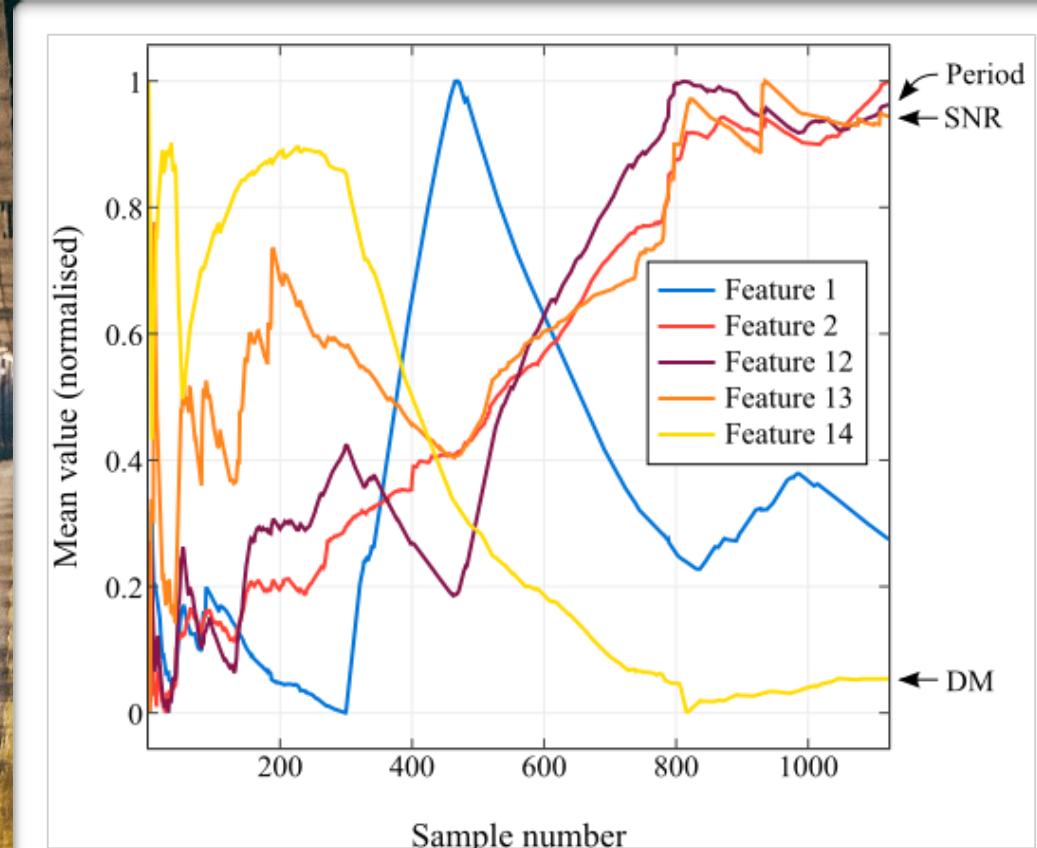
b) Batch Model

- Worsened by non-stationarity.
- Arbitrary batch contains no rare class examples.
- Incremental model might not see rare examples either.

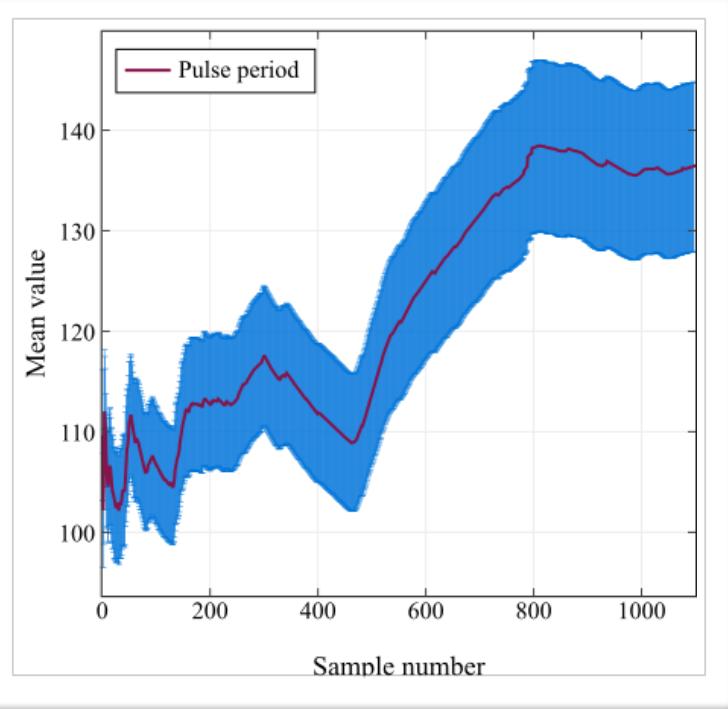


Non-stationarity (1)

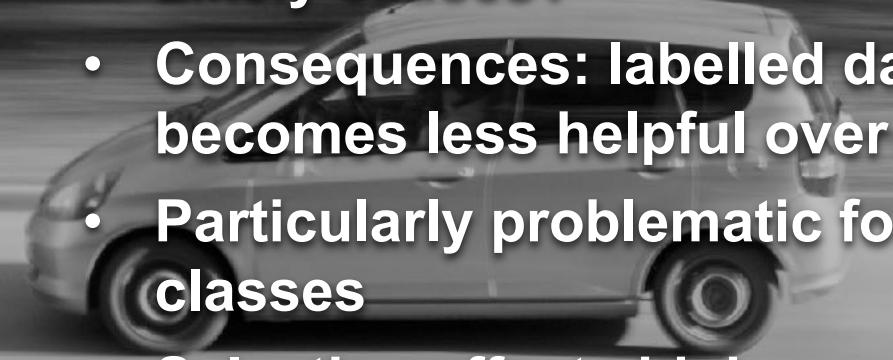
- Data distributions changes over time
- Can affect one or more classes
- Cause violations in i.i.d assumption
- Evidence?



Non-stationarity (2)



- **Likely causes?**
- **Consequences: labelled data becomes less helpful over time**
- **Particularly problematic for RFI classes**
- **Selection effects higher up processing chain?**
- **Requires adaptable learning**

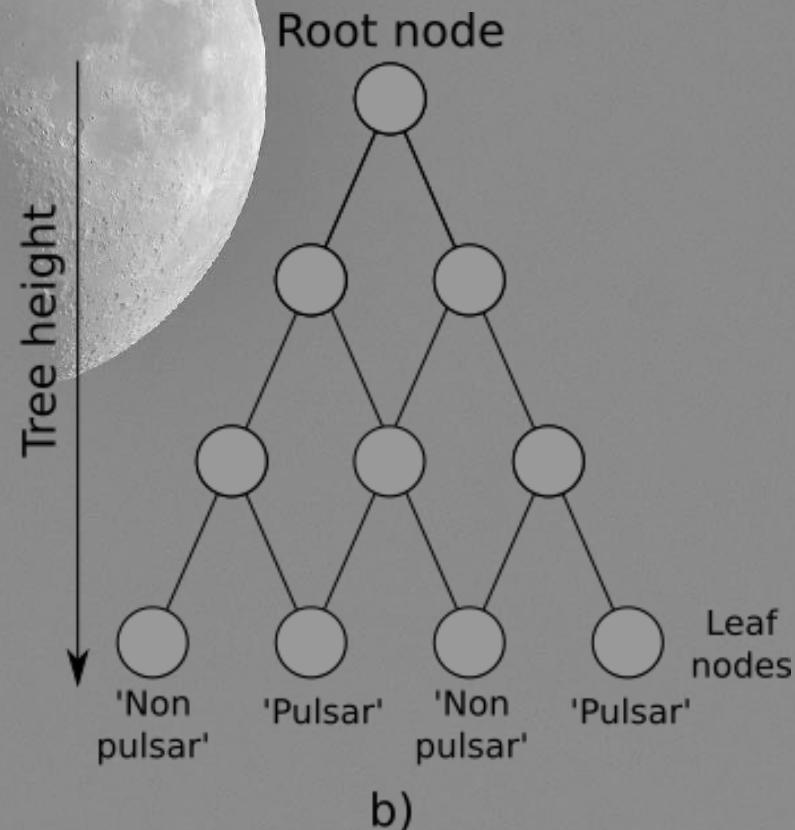
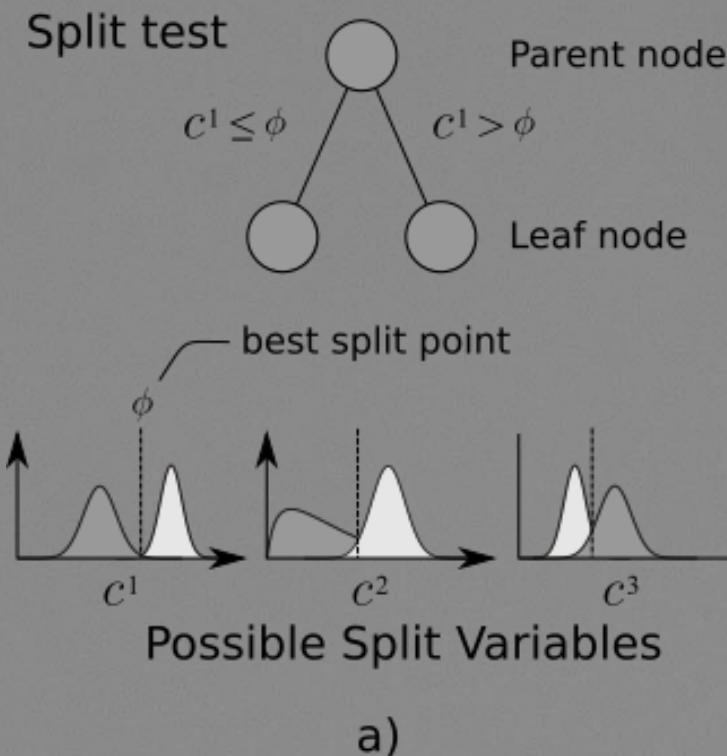


MANCHESTER
1824

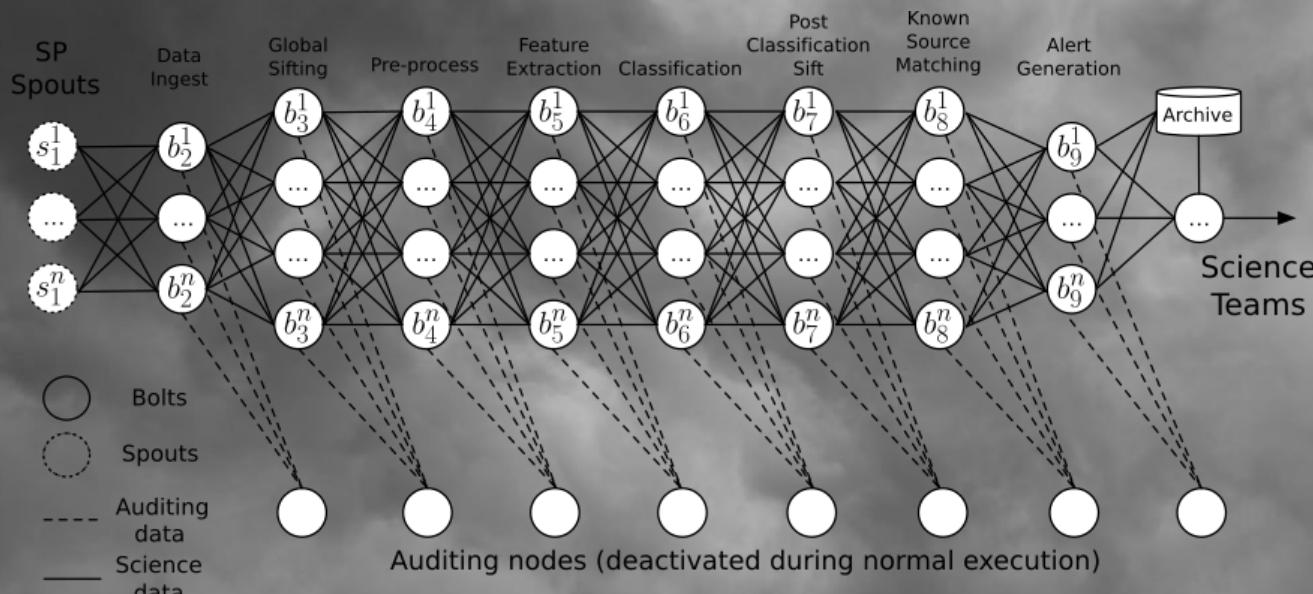
The University of Manchester



Tree Learning



Experimental Prototypes



Credit: Lyon et. al. in prep.

99.9 % Accuracy

81 % Recall

**225 second
processing time**

**Progress but not
enough!**

The Future

- **50 years since Gordon Moore's paper “Cramming more components onto integrated circuits”, Electronics, Volume 38, Number 8, April 19, 1965.**
- **Computing power increasing - GPUs/multi-processors.**
- **This may mean more data.**
- **More sophisticated computing and analytics?**
- **More systems and technologies to learn?**
- **As a community we can meet these challenges.**

Questions & Thank For Listening



@scienceguyrob



robert.lyon@manchester.ac.uk

Acknowledgements

- Work supported by grant EP/I028099/1 from the EPSRC.
- Uncredited images obtained from <https://pixabay.com> (Creative Commons License, no attribution required).
- SKA image credit: the SKA Organisation.

More Information:

Paper DOI: [10.1093/mnras/stw656](https://doi.org/10.1093/mnras/stw656)

Notebook DOI: [10.5281/zenodo.883.844](https://doi.org/10.5281/zenodo.883.844)

Dataset DOI: [10.6084/m9.figshare.3080389.v1](https://doi.org/10.6084/m9.figshare.3080389.v1) or <https://archive.ics.uci.edu/ml/datasets/HTRU2>