Discovering pulsars using machine learning

V. Balakrishnan, P. Padmanabh, E.D. Barr, D.J. Champion,
H.R. Klöckner, M. Kramer and SAP Data Hub Team

Max Planck Institute for Radio Astronomy, Germany

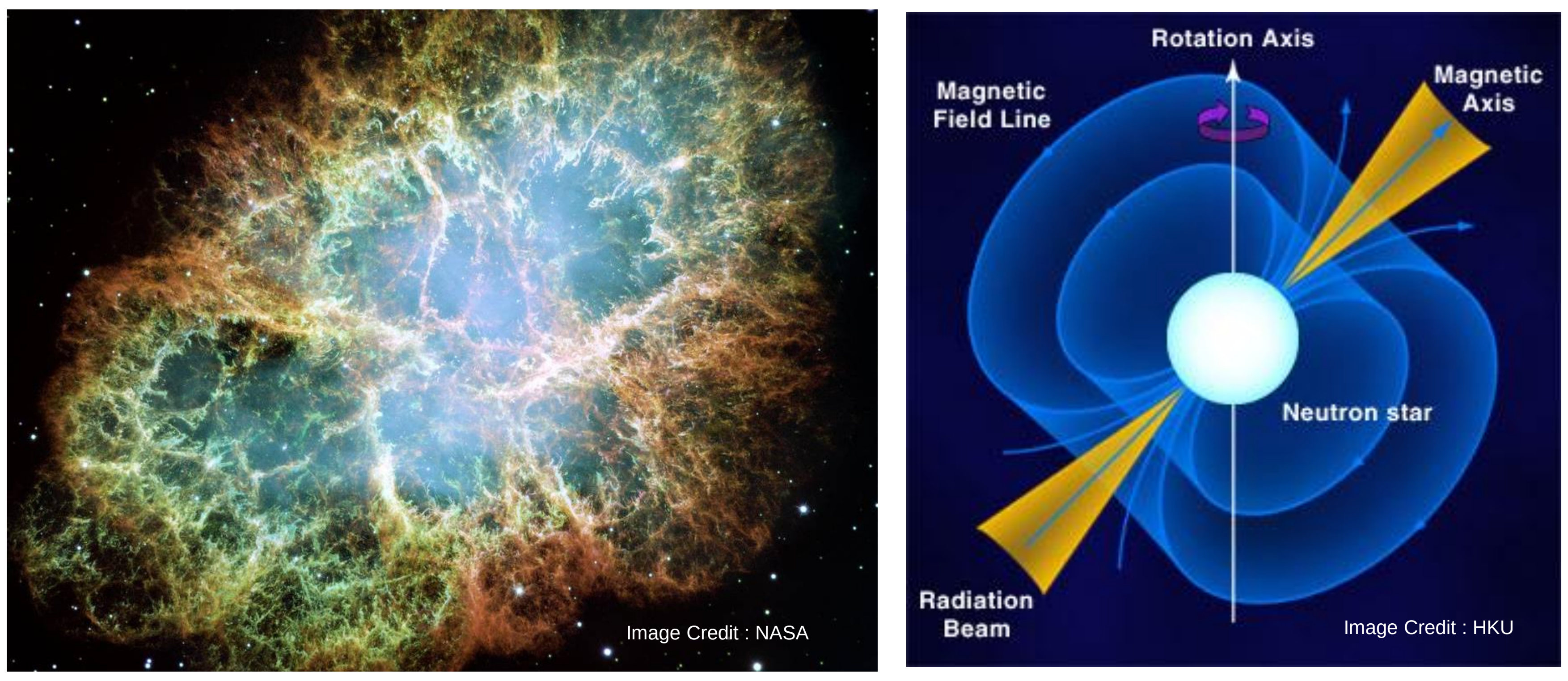

Fig. 1 : a) Crab Nebula – A supernova remnant with a pulsar at its centre b) Schematic of a pulsar

Pulsar science

Pulsars are highly magnetized rapidly **rotating neutron stars** born in the cores of supernova explosions (see Fig. 1 a). Neutron stars are some of the densest objects known in the observable universe. For example, they typically have masses similar to the Sun packed in a sphere with the diameter of Paris.

Pulsars are like cosmic lighthouses that emit beams of electromagnetic radiation seen as pulses on Earth (see Fig. 1 b). They rotate up to 700 times a second. Pulsars provide a unique opportunity to test many aspects of fundamental physics in extreme environments that cannot be replicated in laboratories on Earth, e.g. **Einstein's theory of general relativity**.

In order to explore these interesting questions, we need to search for pulsars using radio telescope surveys that sample the sky with **high time resolution**.

Why machine learning?

Searching for new pulsars is a **Big Data problem**. The techniques used to detect the radio emission of pulsars, are designed to find periodic signals in noisy data. Such signals are stored in a reduced data cube referred to as a **pulsar 'candidate'**. Currently, we inspect different projections of this data cube manually to decide if the signal is a real detection or a false positive (see Fig. 4).

The upcoming radio telescope, **MeerKAT** (online in 2018, see Fig. 2) is expected to generate **3.8 billion candidates (~5 PB)** in total. Thus, manually inspecting candidates is impractical.

Pulsar astronomers have turned to algorithms based on supervised learning to differentiate pulsars from man-made radio interference (e.g. 50 Hz voltage mains, satellite signals, airport radar collectively called **Radio Frequency Interference**) .

Fig. 2 : MeerKAT radio telescope array in South Africa

Comparing machine learning models using SAP Data Hub

In order to compare different machine learning (ML) models in real time from high throughput data streams, we are currently using **SAP Data Hub** which is a scalable container orchestration system (see Fig. 3) built using **Kubernetes** and **Docker**. **Apache Kafka** is used to stream candidates from the data storage cluster to the pipeline built on the **SAP HANA Vora** analytics system. These candidates are then ranked based on different ML models and stored in a database along with their corresponding metadata. The aim of this system is to provide a flexible platform for astronomers to inject new algorithms and test their reliability across a broad range of datasets in a short span of time.

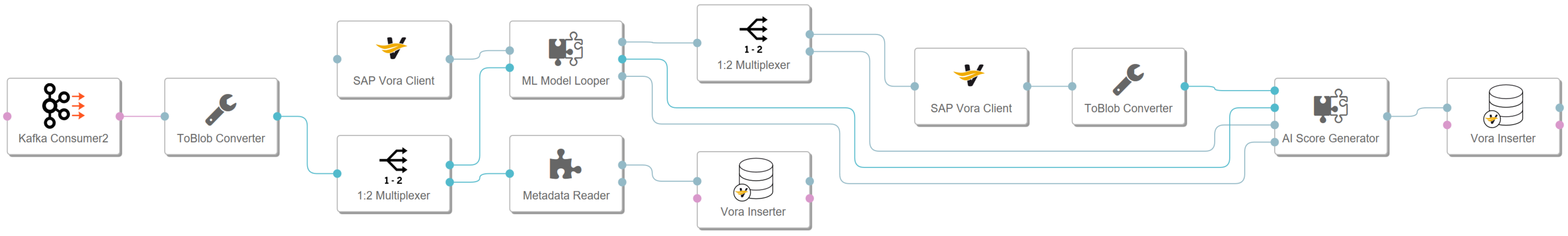

Fig. 3: Schematic of the pipeline for comparing different machine learning algorithms to discover new pulsars. Each block represents an operator with a specific functionality.

What is a pulsar candidate?

A pulsar candidate is a three dimensional array consisting of elapsed time since the start of a telescope observation, observing frequency range and the rotational phase of the pulsar.
Astronomers inspect various projections of this array by averaging over each axes in order to determine if a pulsar signal is in the data.

Steps to verify a pulsar detection:
1) Check if the signal is persistent through time (e.g. Fig. 4a).
2) Check if the signal seen across a broad range of frequencies (e.g. Fig. 4b)
3) Average over time and frequency to inspect for a pulse-like profile.

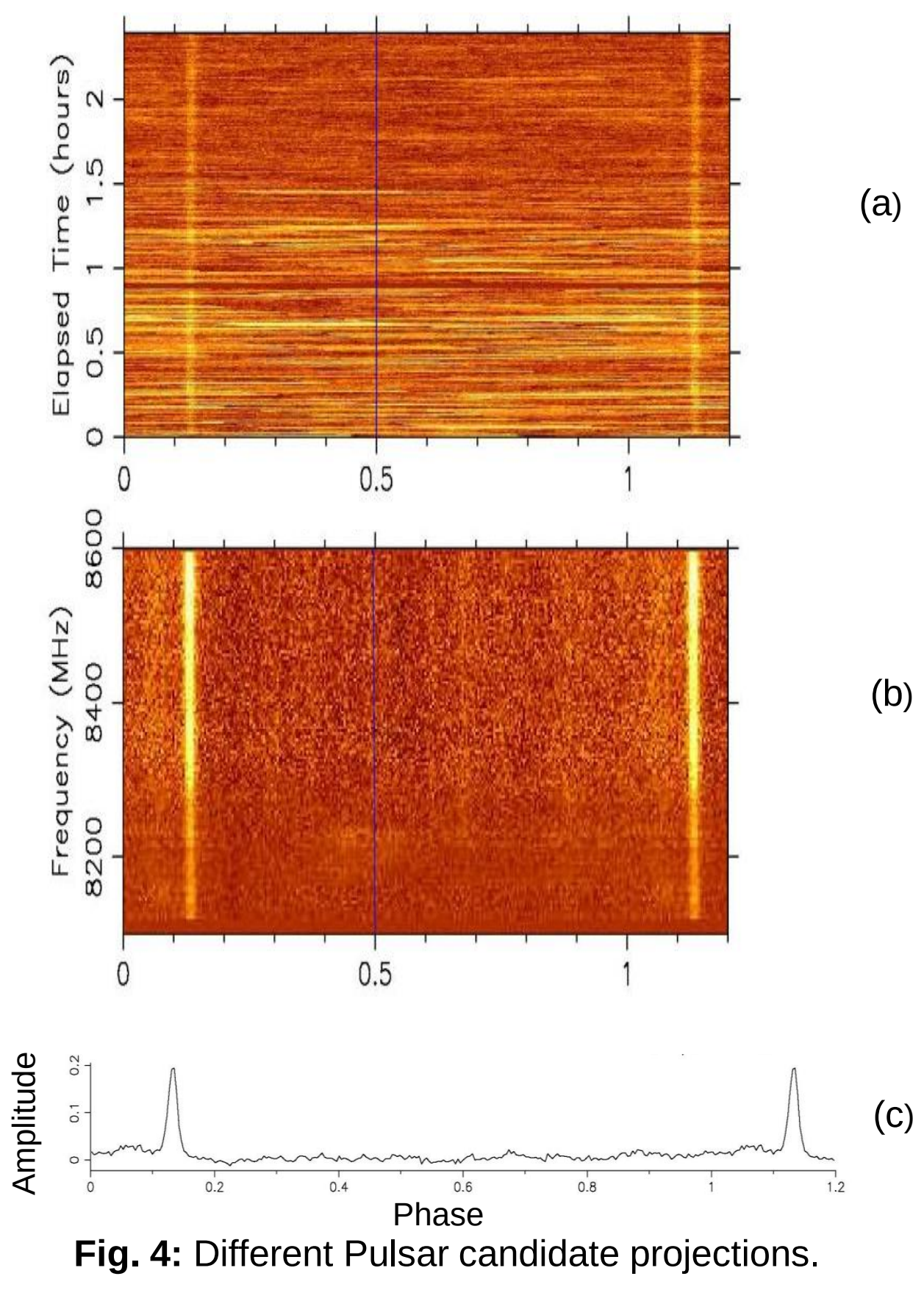

Fig. 4: Different Pulsar candidate projections.

Work in progress

- Setup different pulsar classification systems on SAP Data Hub (see Fig. 3) to compare their relative performances.
- Reprocess archival candidate data (175 million candidates) through this new framework in order to check for missed pulsar detections.
- In general, the radio frequency interference environment around telescopes is dynamic. Thus, using a pre trained ML model for this scenario is not ideal. We are currently exploring ML models that can make decisions adaptively.
- Develop a new robust classifier based on **semi supervised learning** that can retrain our models **offline** as more unlabelled candidates are produced from MeerKAT.
- This framework when developed will act as an important quality control and fault detection mechanism for radio telescope data. Unexpected signals could be a new scientific discovery!

Pulsar classification systems

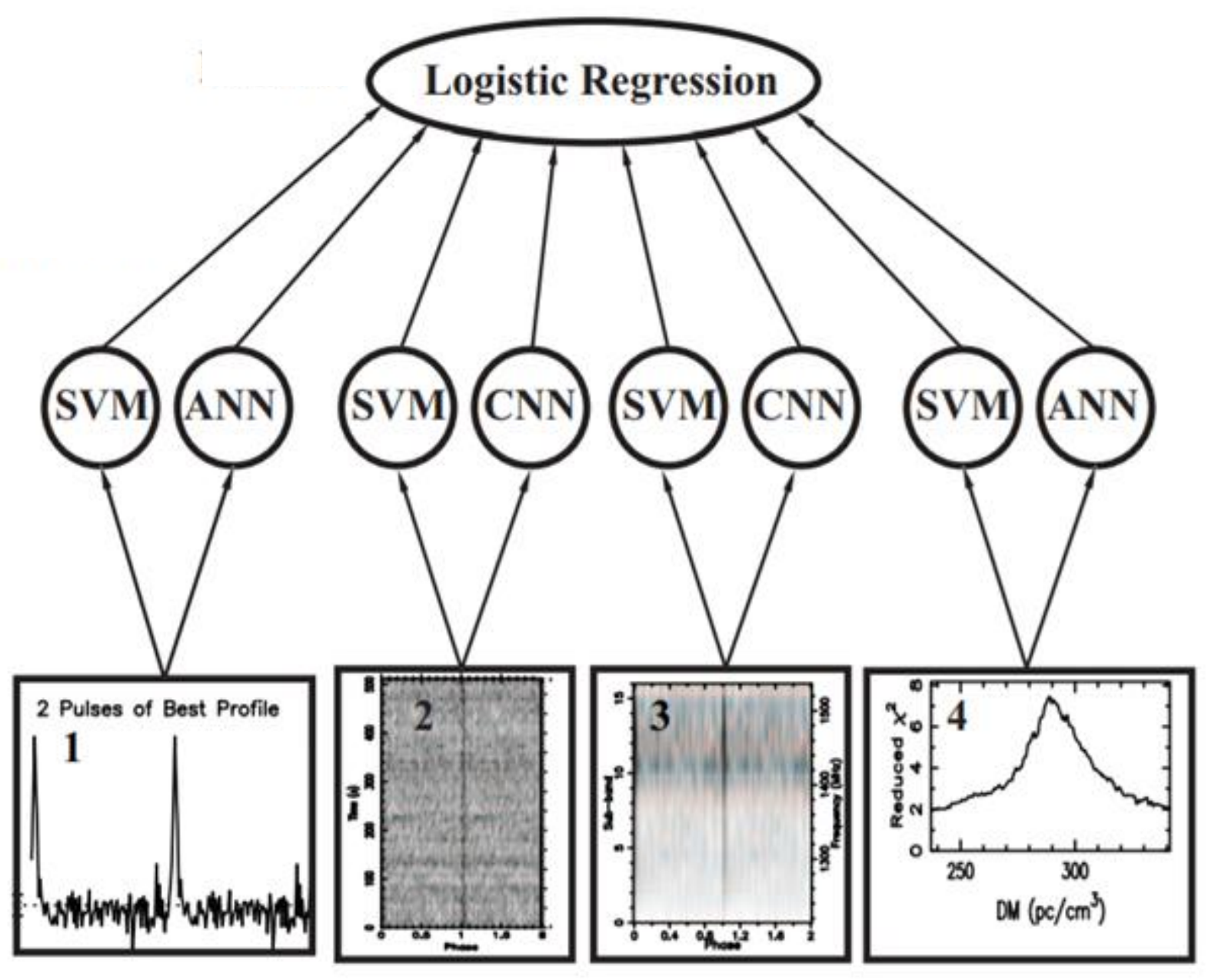

Fig. 5: Example of a two level ML based classifier system used to find pulsars.

In order to filter pulsar signals from a large number of candidates, astronomers either use a ranking system based on pre-assigned weight for each feature in the candidate cube (e.g. Lee et al. 2013) or an ML based approach where models are trained with a pre-labelled dataset (e.g. Zhu et al. 2014, Morello et al. 2014). Recent approaches also include a tree based ML classifier designed specifically for online processing (e.g. Lyon et al. 2016).

Each system has been successful in discovering new pulsars from telescopes around the world. However, more work can be done to understand the strengths and weaknesses of each approach.

References

- Zhu W. W. et al. 2014, ApJ, 781, 2
- Keith M. J. et al. 2010, MNRAS, 409, 619-627
- Lyon R.J. et al. 2016, MNRAS, 459, 1104
- Morello V. et al. 2014,MNRAS, 443, 1651
- Lynch R. S. et al. 2013, IAU Symposium, 291
- Lee K. J. et al. 2013, MNRAS, 433, 688