The University of Manchester
Jodrell Bank Observatory

MANCHESTER 1824

ANNA SCAIFE
JODRELL BANK CENTRE FOR ASTROPHYSICS
@RADASTRAT

REAL-TIME CLASSIFICATION AT SKA-SCALE FOR TIME DOMAIN ASTROPHYSICS

WITH CONTENT FROM ROB LYON @SCIENCEGUYROB

The University of Manchester
Jodrell Bank Observatory

MANCHESTER 1824

ANNA SCAIFE

JODRELL BANK CENTRE FOR ASTROPHYSICS

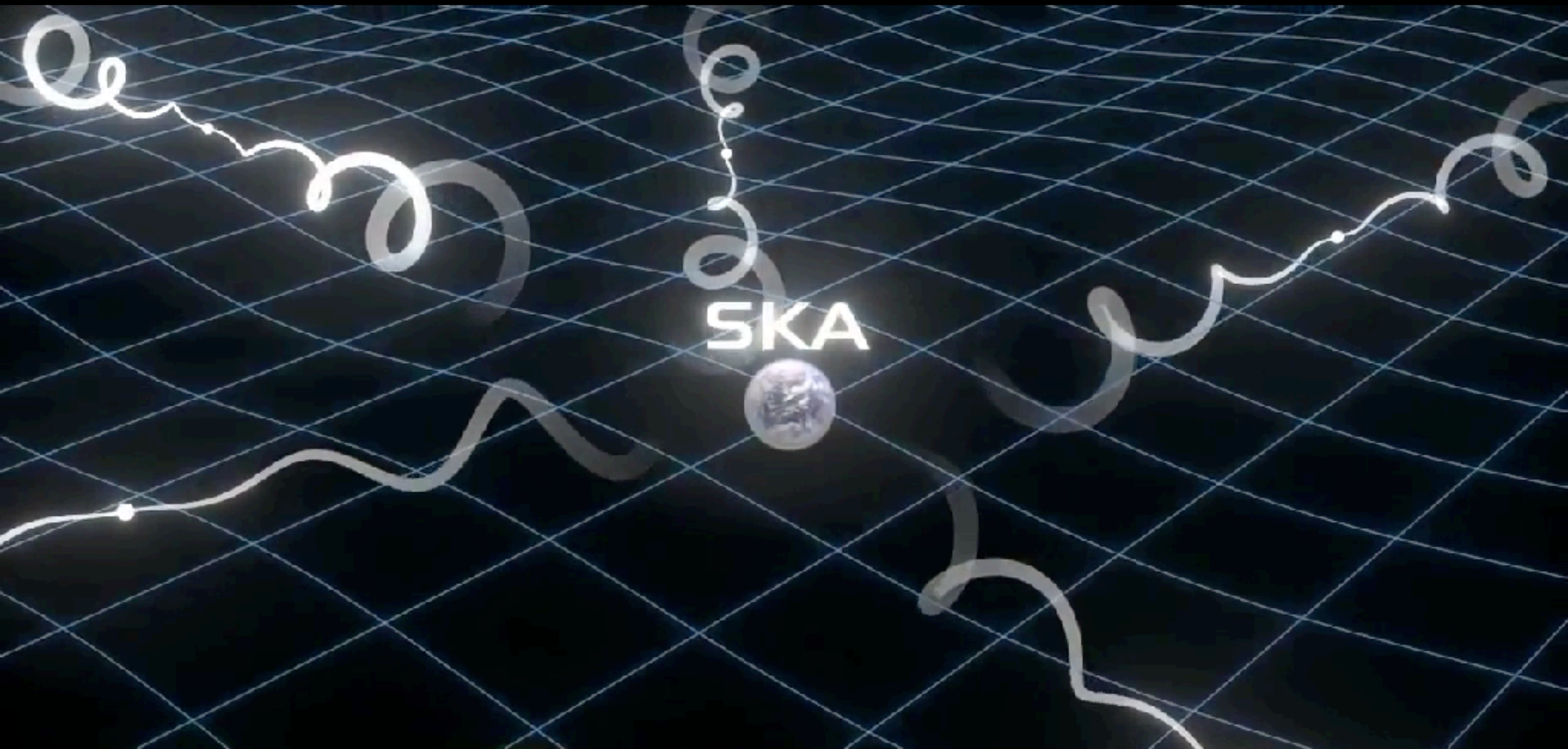
@RADASTRAT

REAL-TIME CLASSIFICATION AT SKA-SCALE FOR TIME DOMAIN ASTROPHYSICS

WITH CONTENT FROM ROB LYON @SCIENCEGUYROB

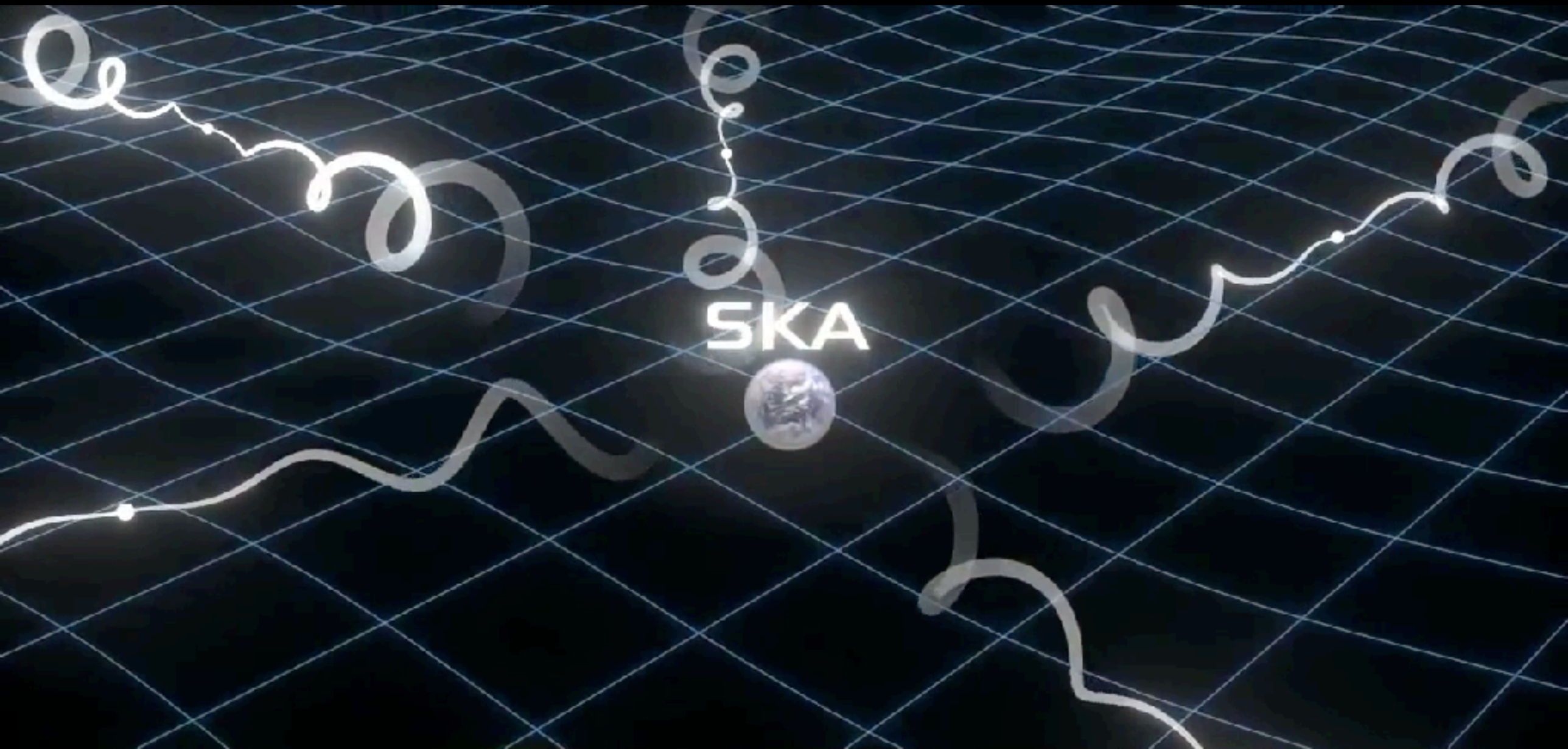
GRAVITATIONAL WAVES

- RIPPLES IN SPACE-TIME
- USE A NETWORK OF PULSARS TO DETECT THEM
- CAUSED BY INTER-ACTING SUPER-MASSIVE BLACK HOLES



GRAVITATIONAL WAVES

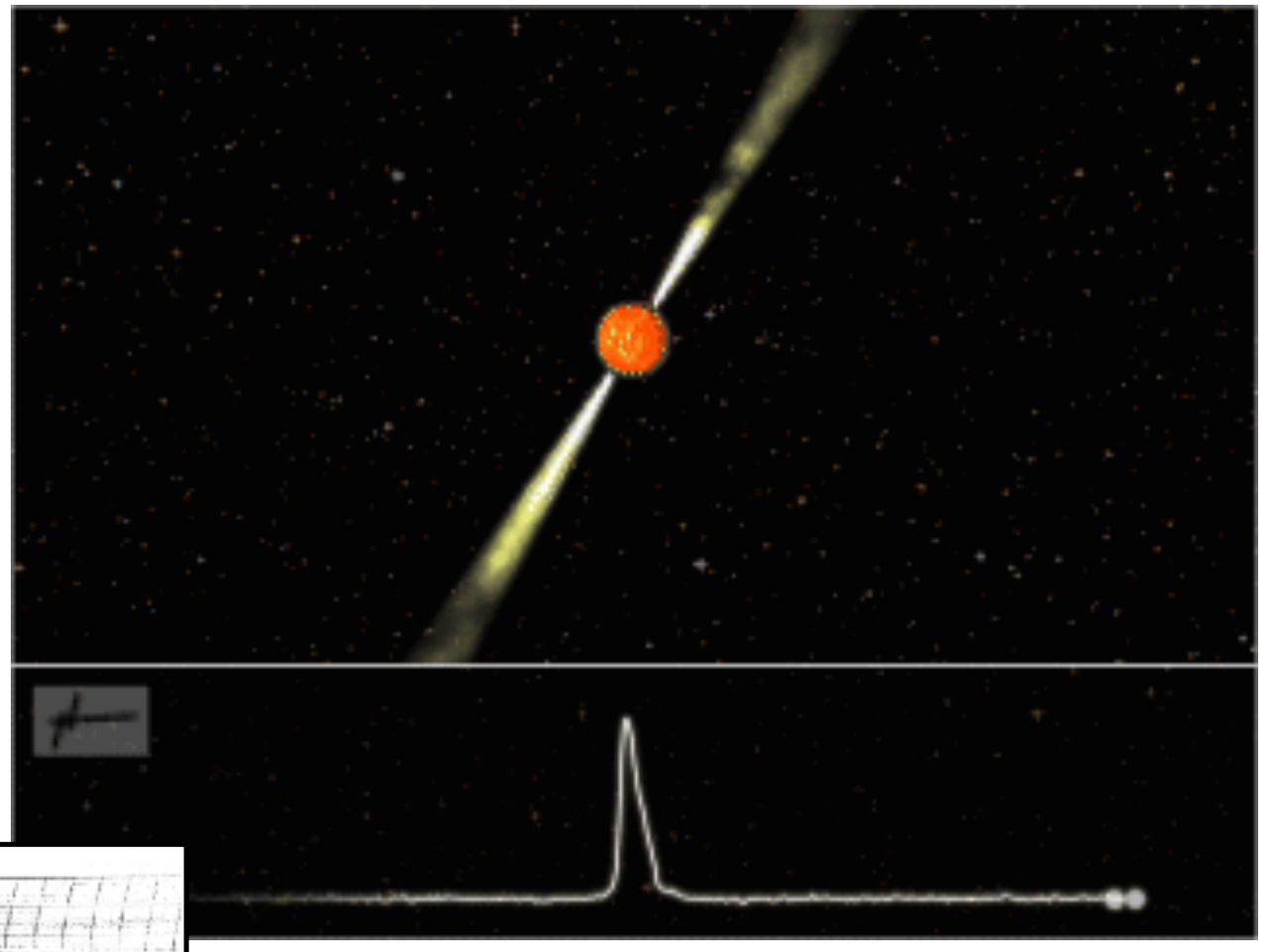
- RIPPLES IN SPACE-TIME
- USE A NETWORK OF PULSARS TO DETECT THEM
- CAUSED BY INTER-ACTING SUPER-MASSIVE BLACK HOLES



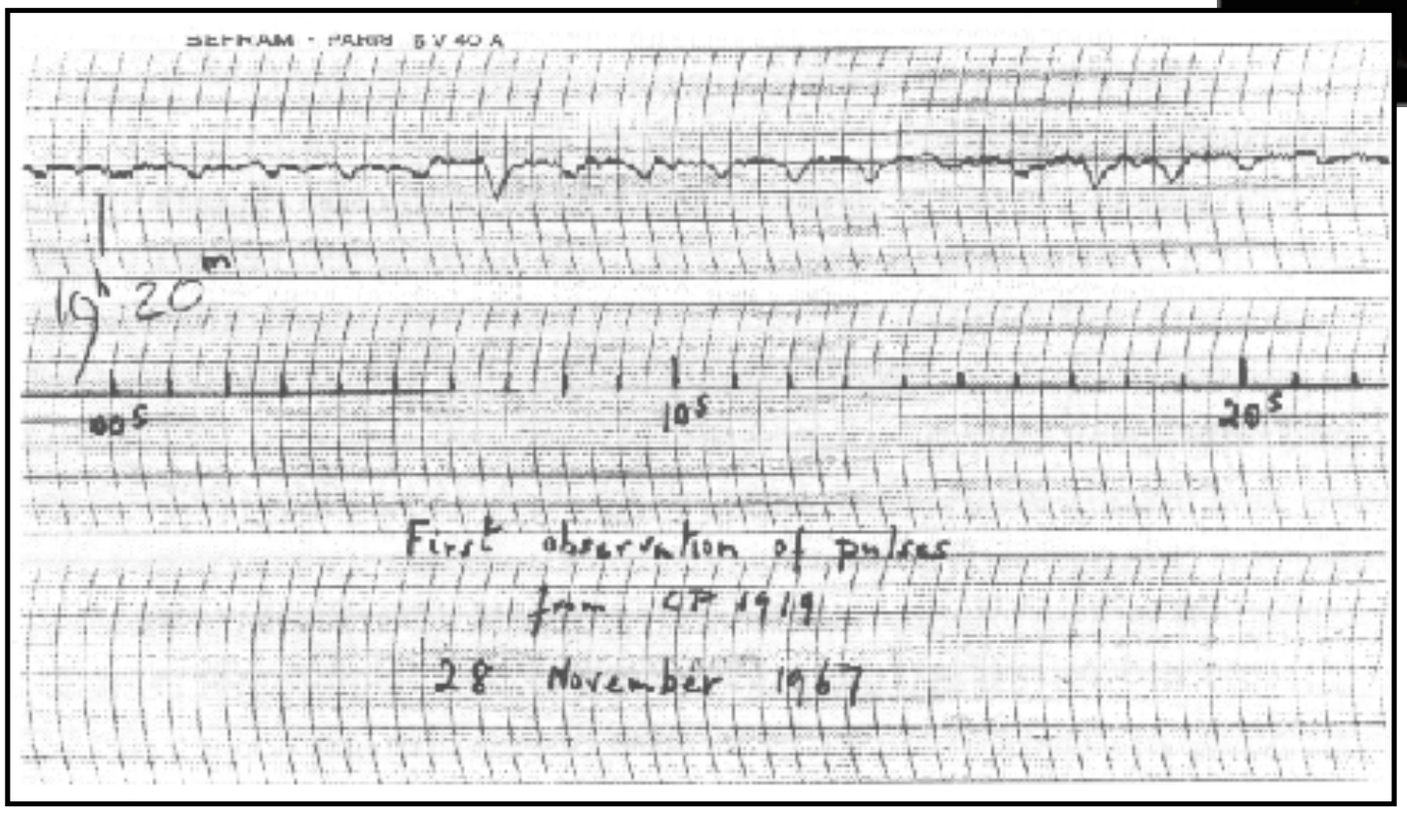
WHAT IS A PULSAR?

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html



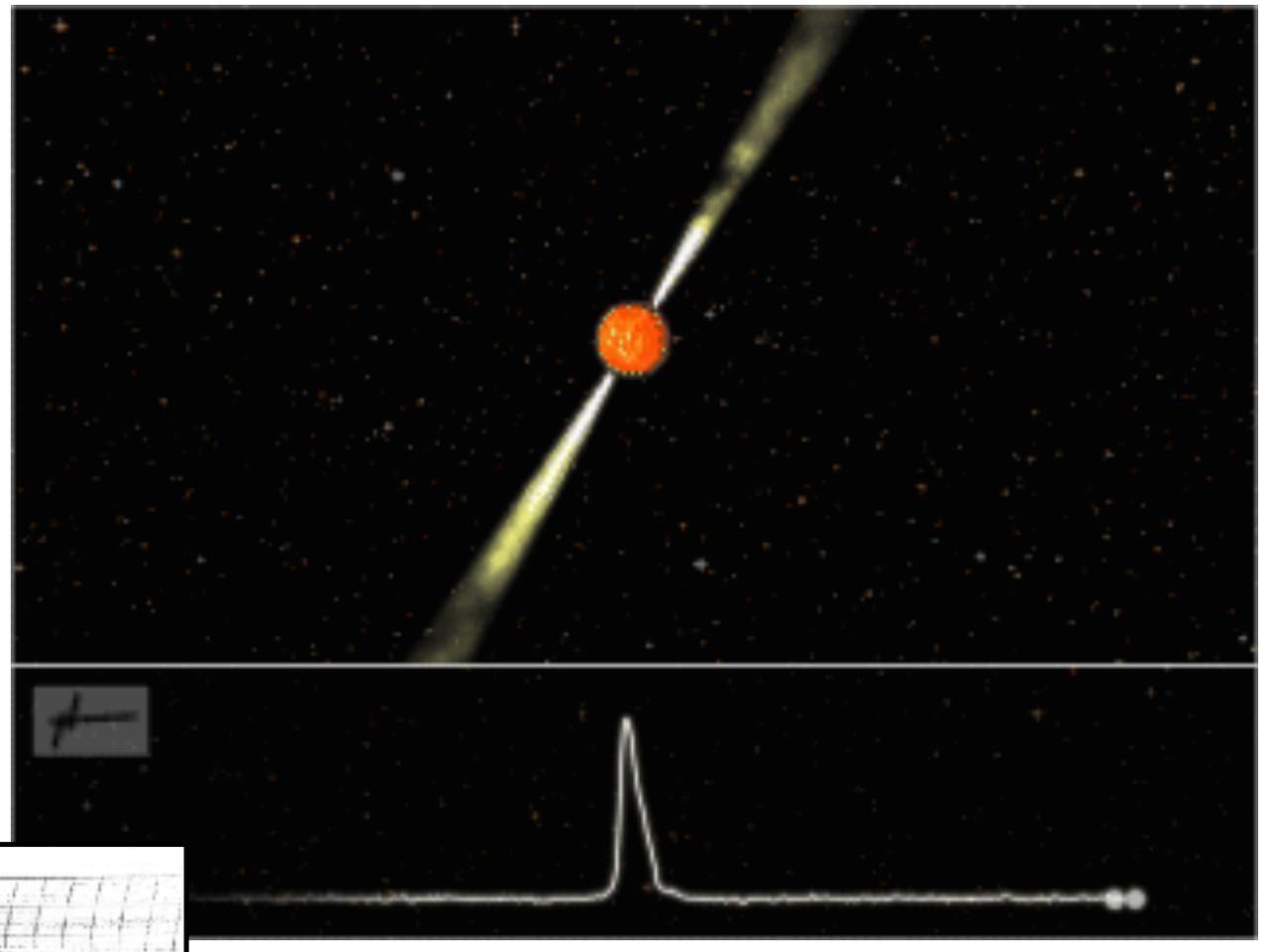
An artist's impression of a pulsar. Image credit: Joeri van Leeuwen, License: CC-BY-AS



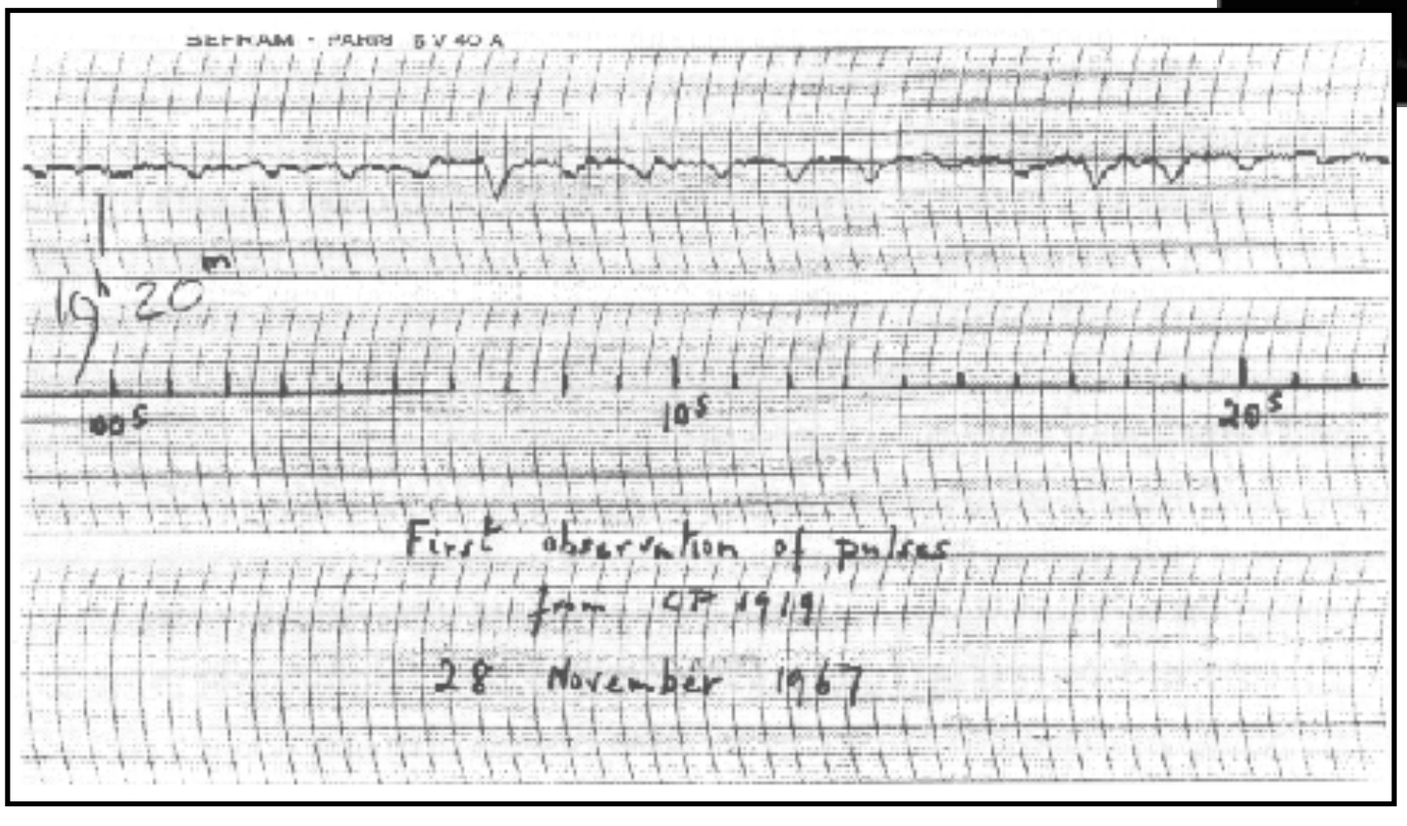
WHAT IS A PULSAR?

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html



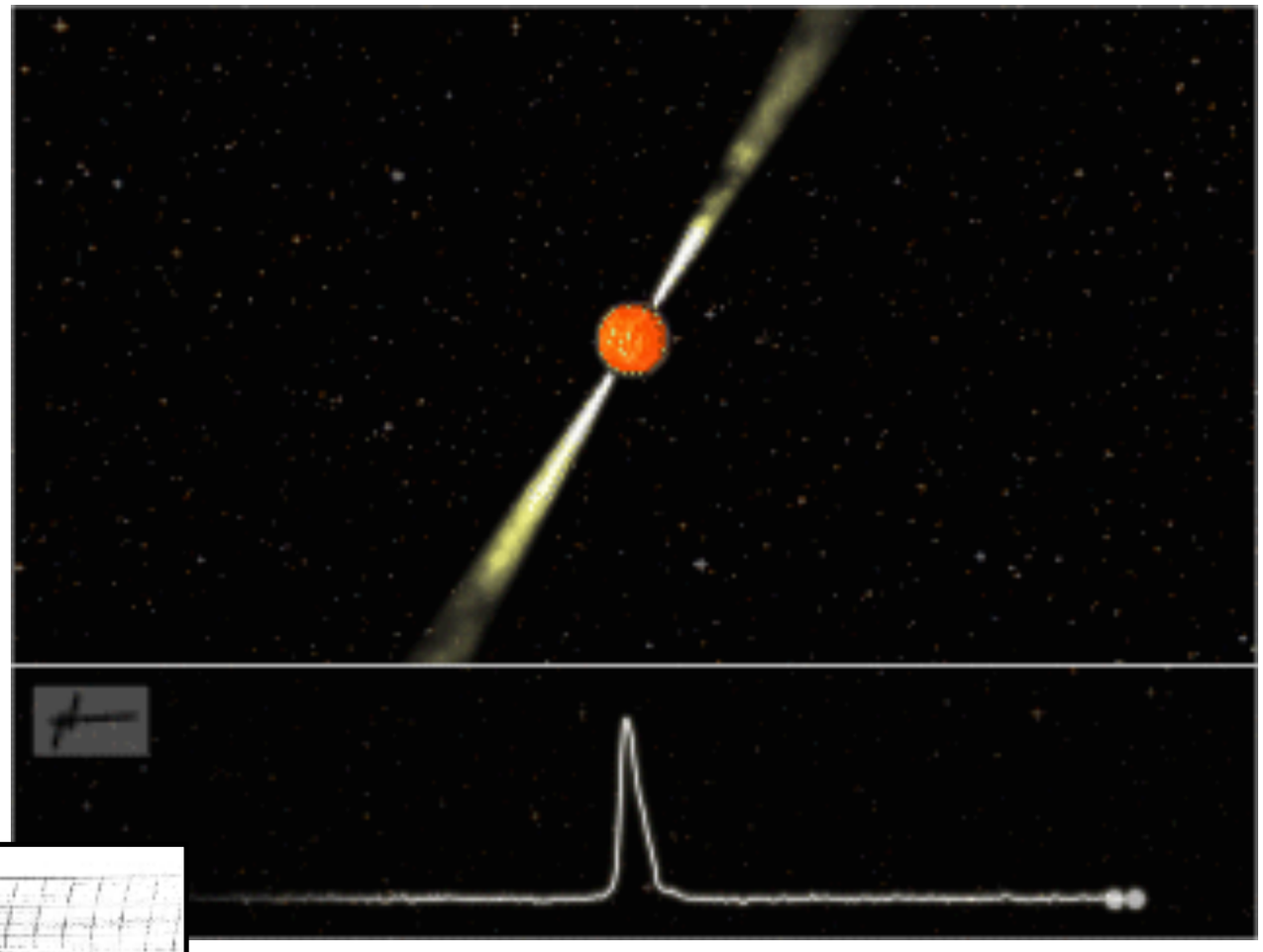
An artist's impression of a pulsar. Image credit: Joeri van Leeuwen, License: CC-BY-AS



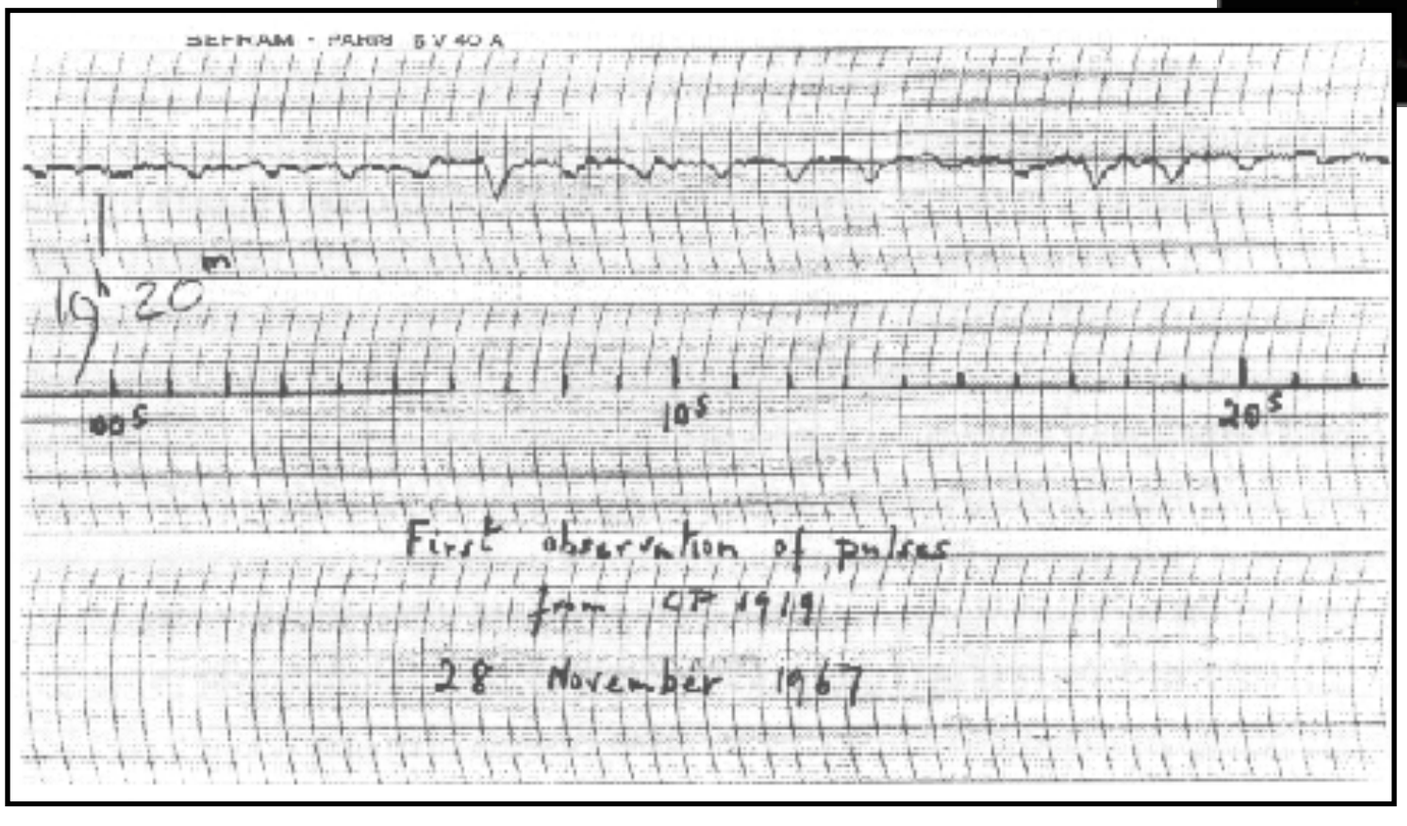
WHAT IS A PULSAR?

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html



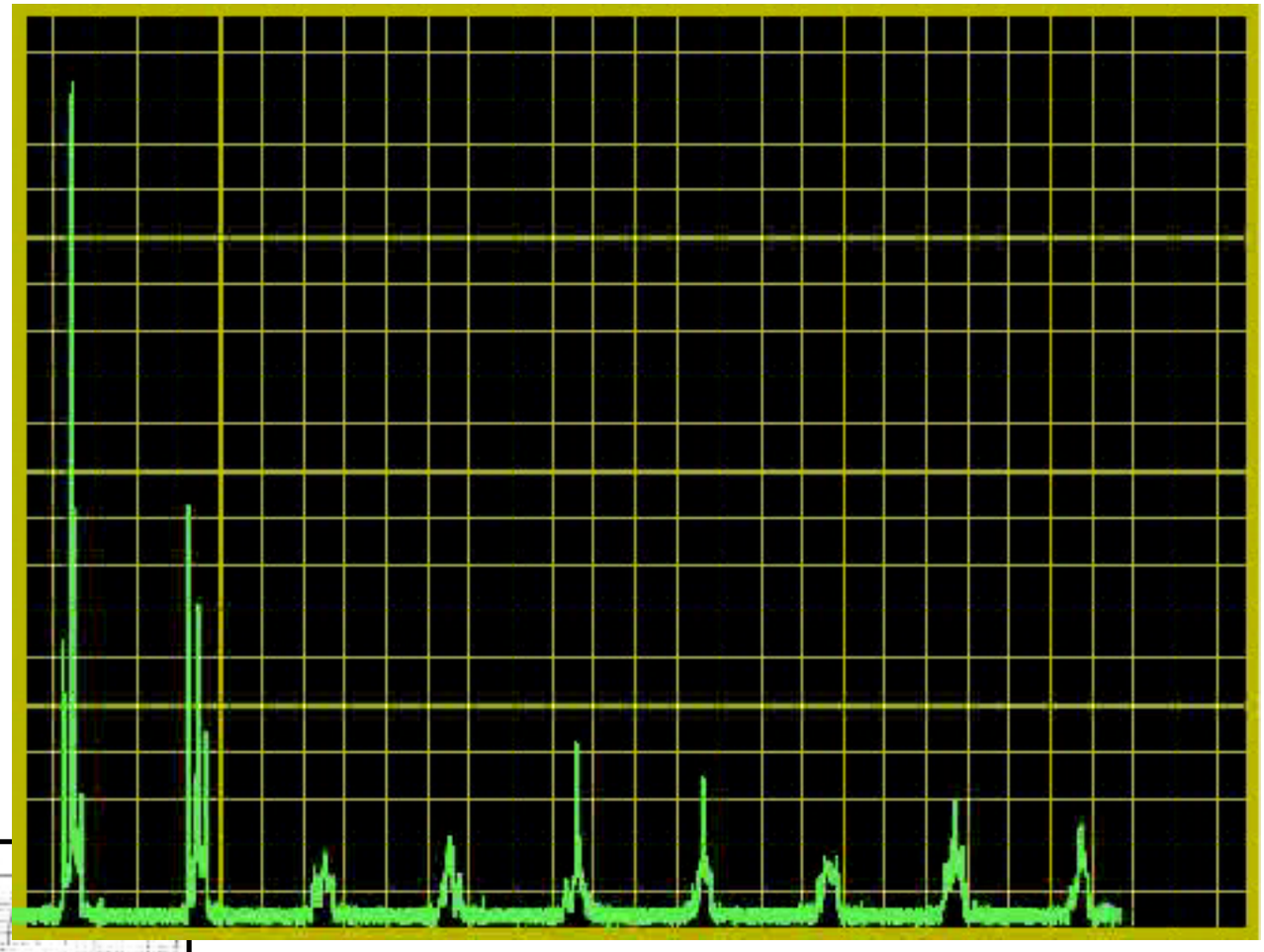
An artist's impression of a pulsar. Image credit: Joeri van Leeuwen, License: CC-BY-AS

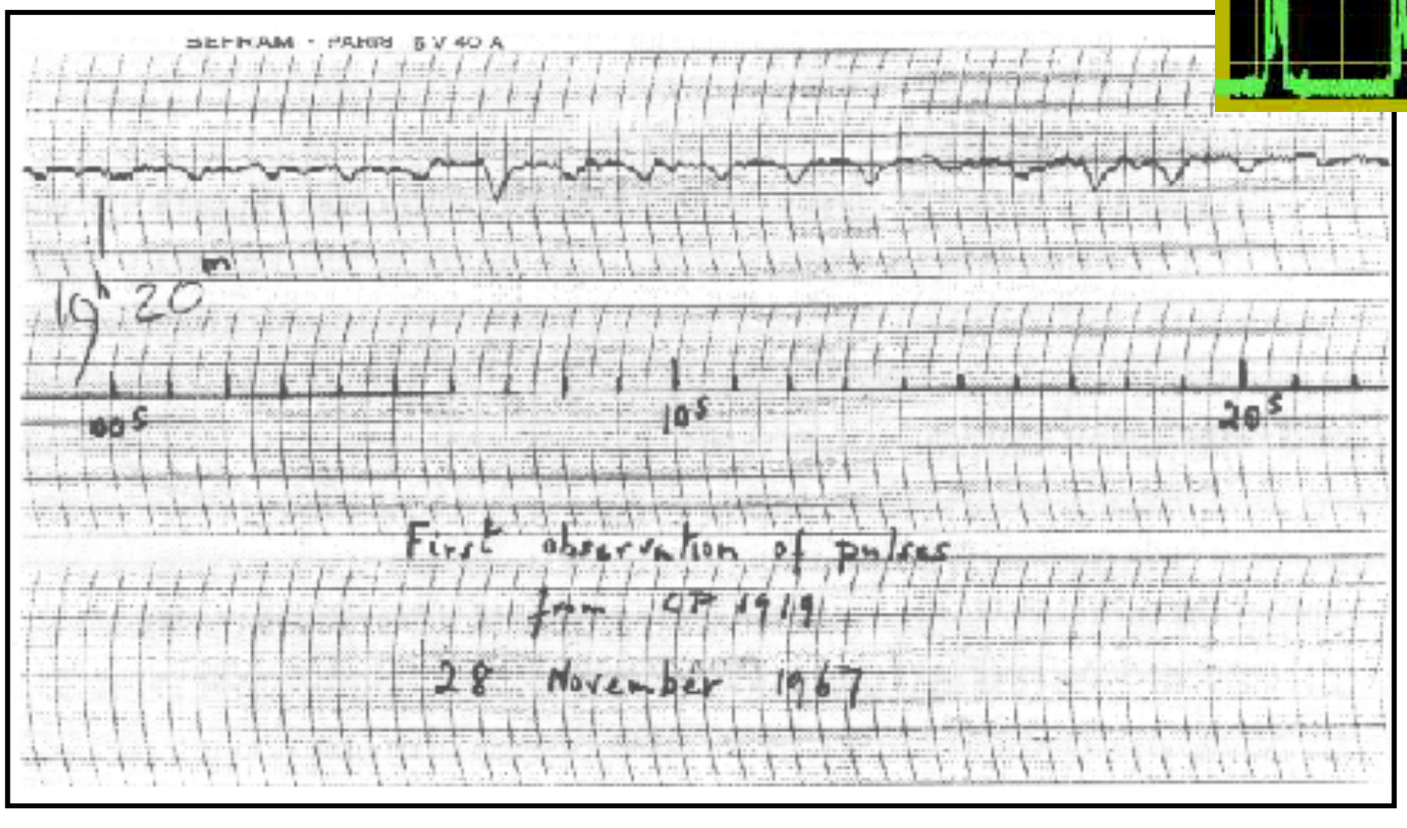


WHAT IS A PULSAR?

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html

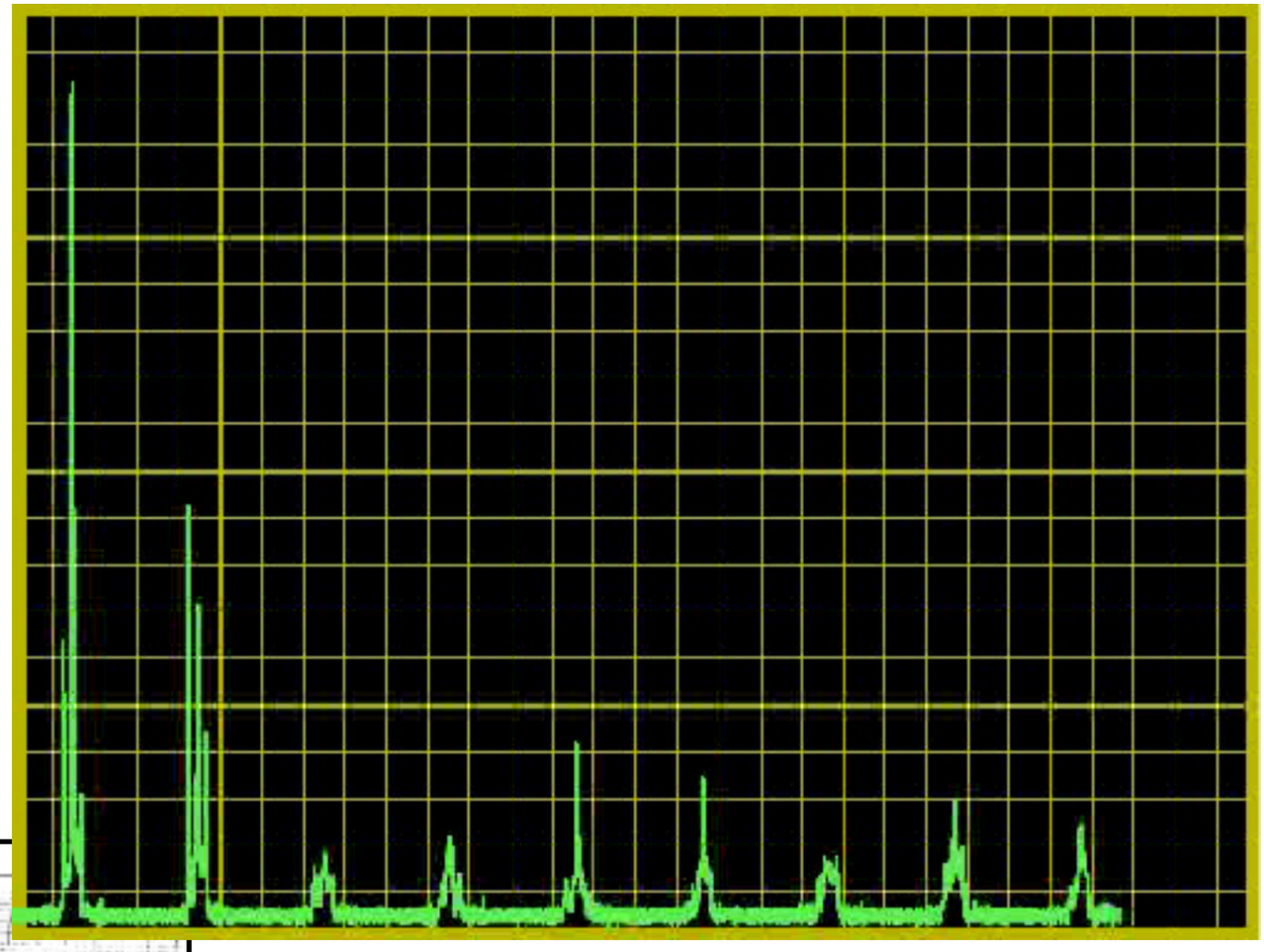


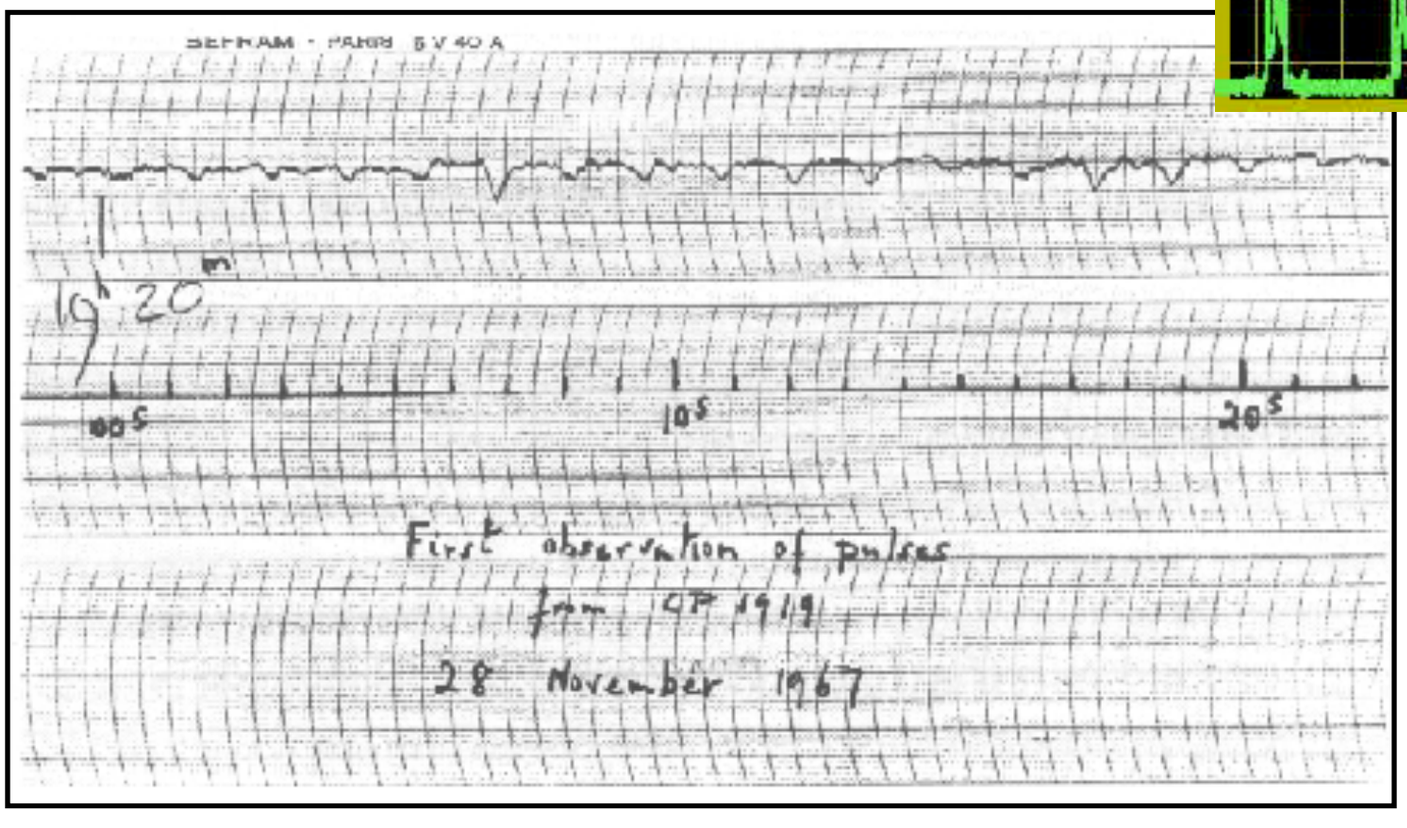


WHAT IS A PULSAR?

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html



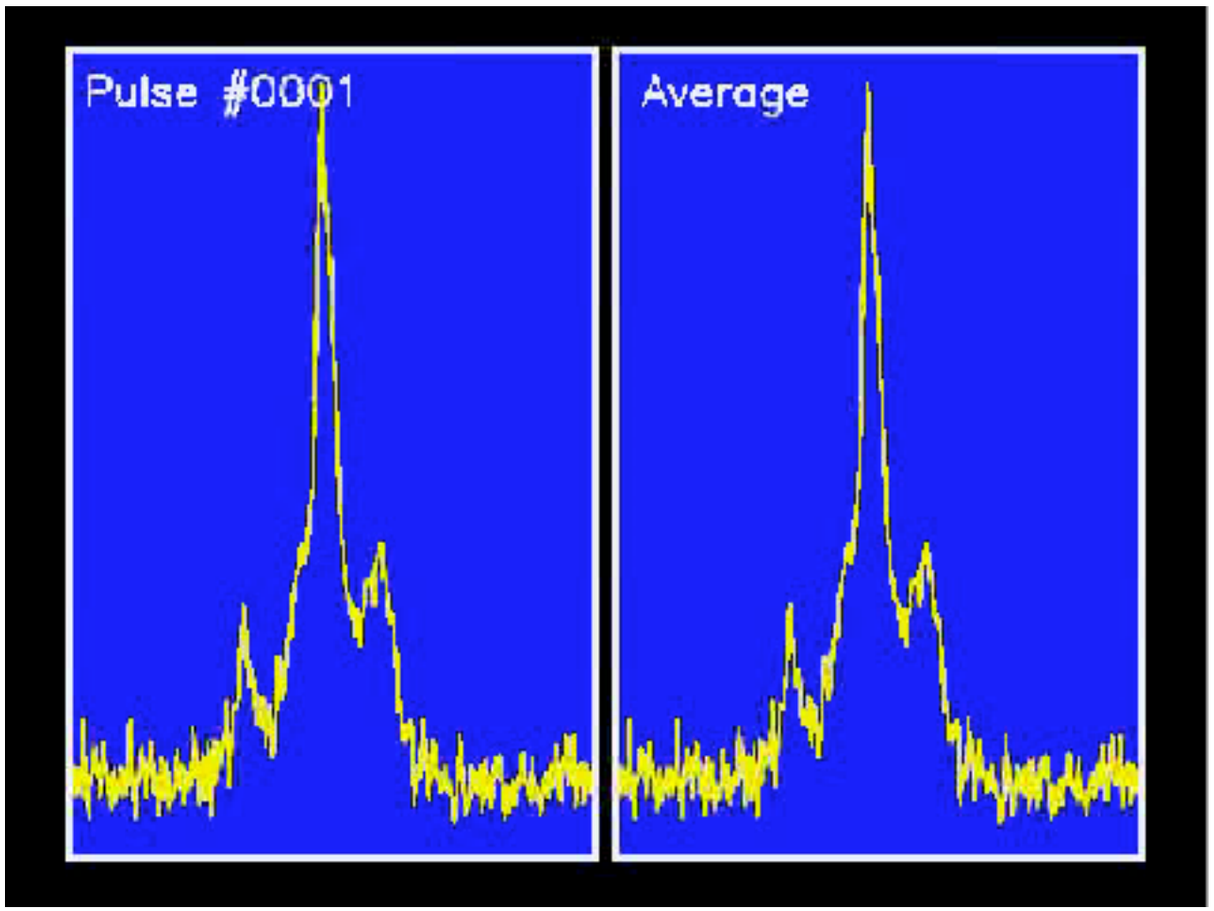


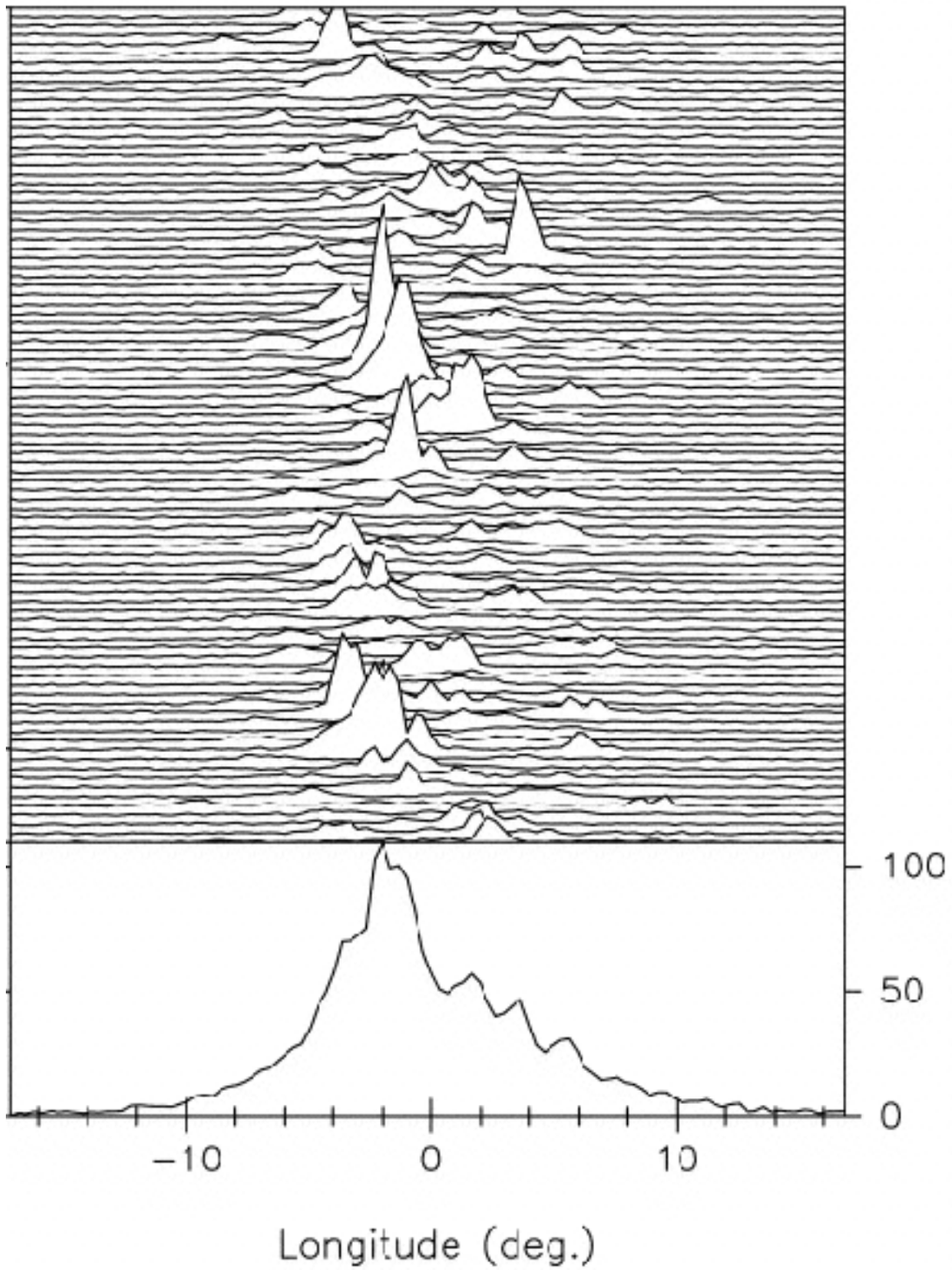
https://www.cv.nrao.edu/~sransom/web/Ch6.html

INTEGRATED PROFILE

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html



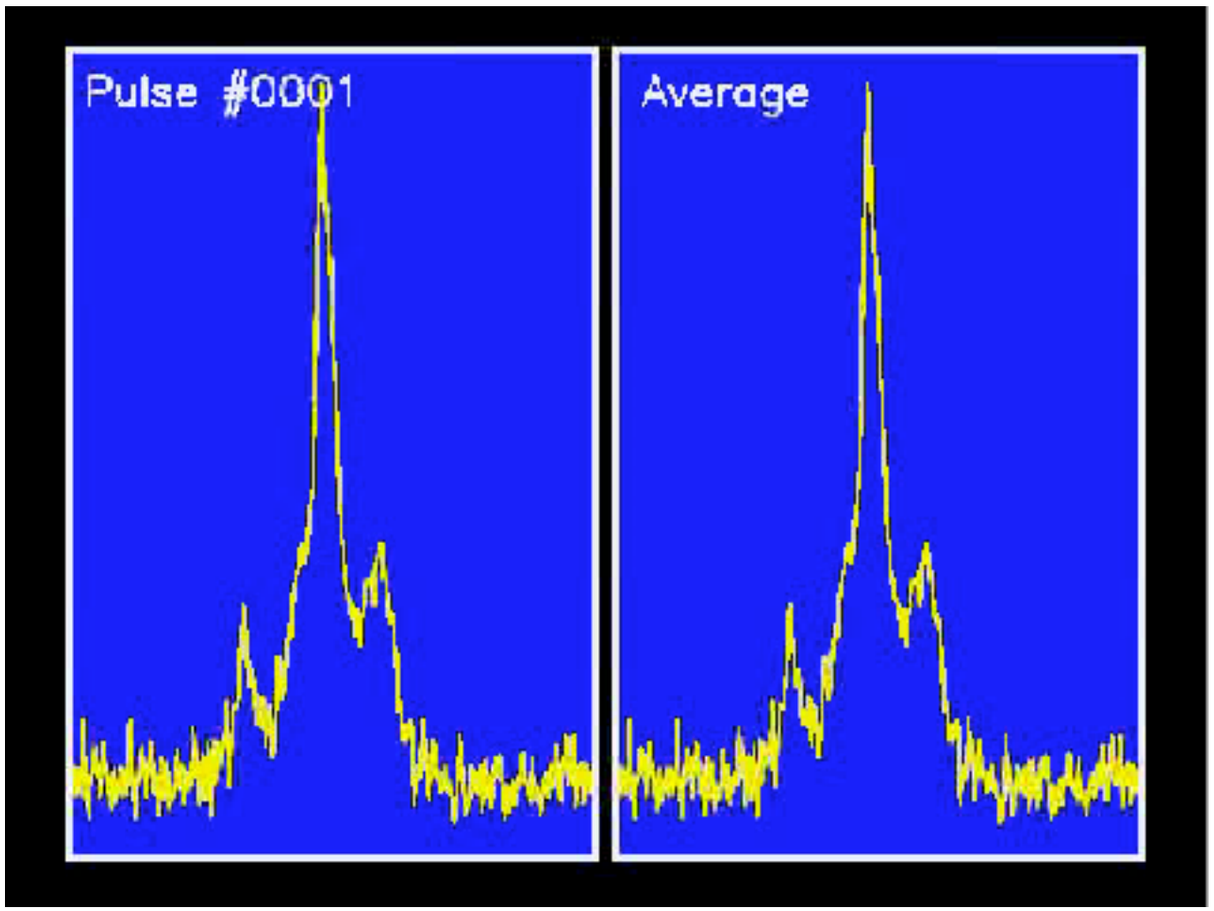


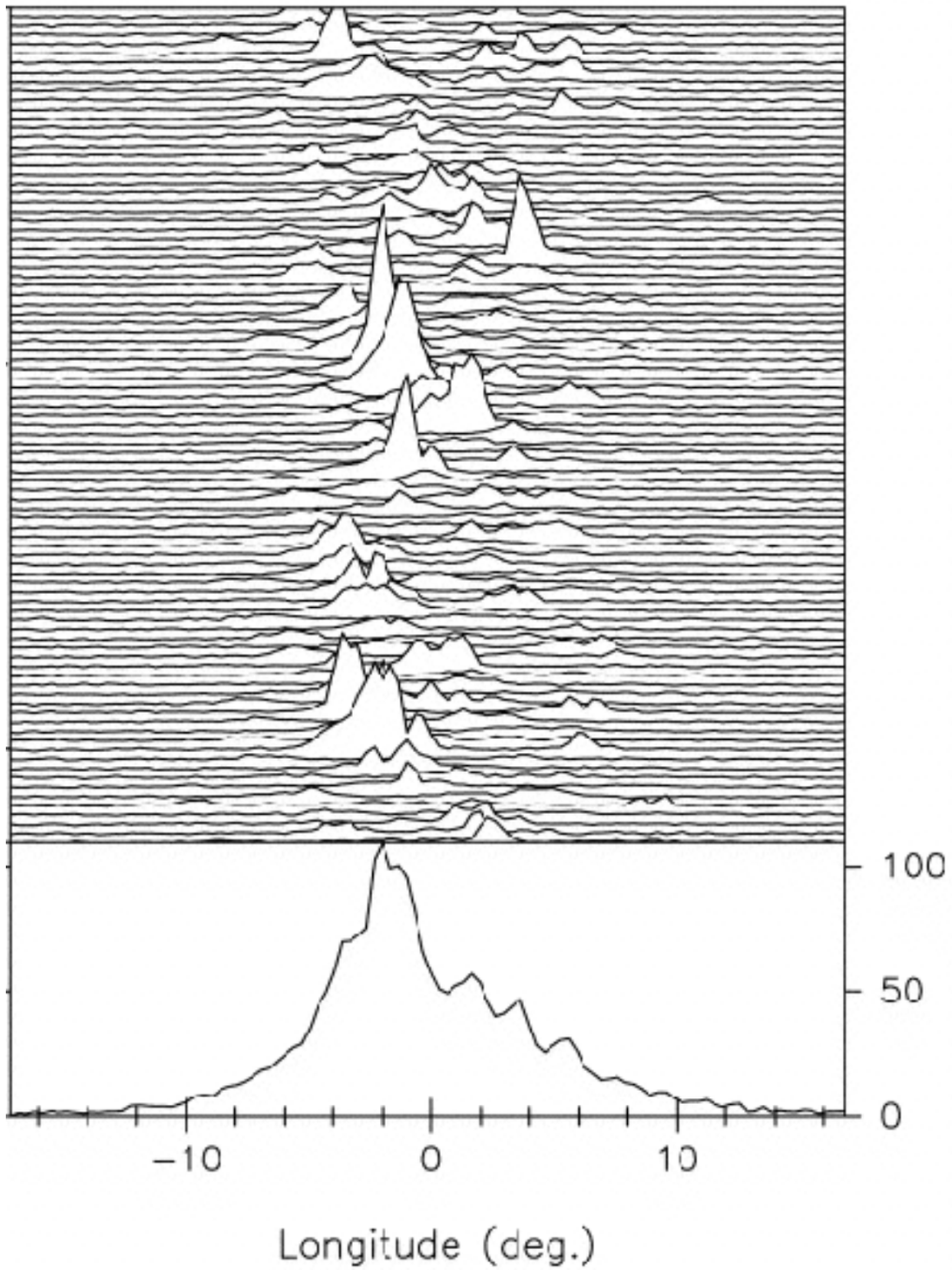
https://www.cv.nrao.edu/~sransom/web/Ch6.html

INTEGRATED PROFILE

PSR B0329+54: This pulsar is a typical, normal pulsar, rotating with a period of 0.714519 seconds, i.e. close to 1.40 rotations/sec.

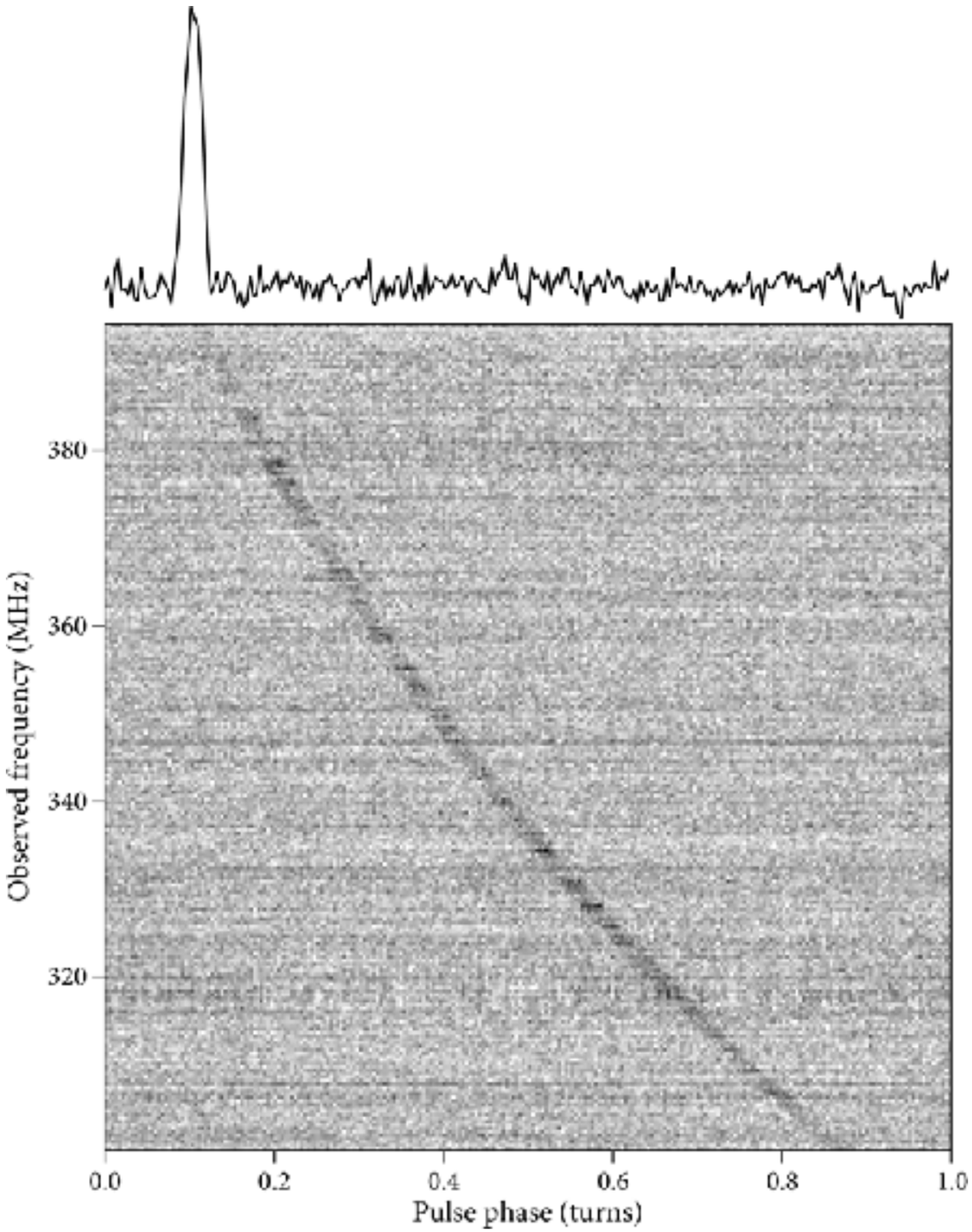
http://www.jb.man.ac.uk/distance/frontiers/pulsars/section1.html





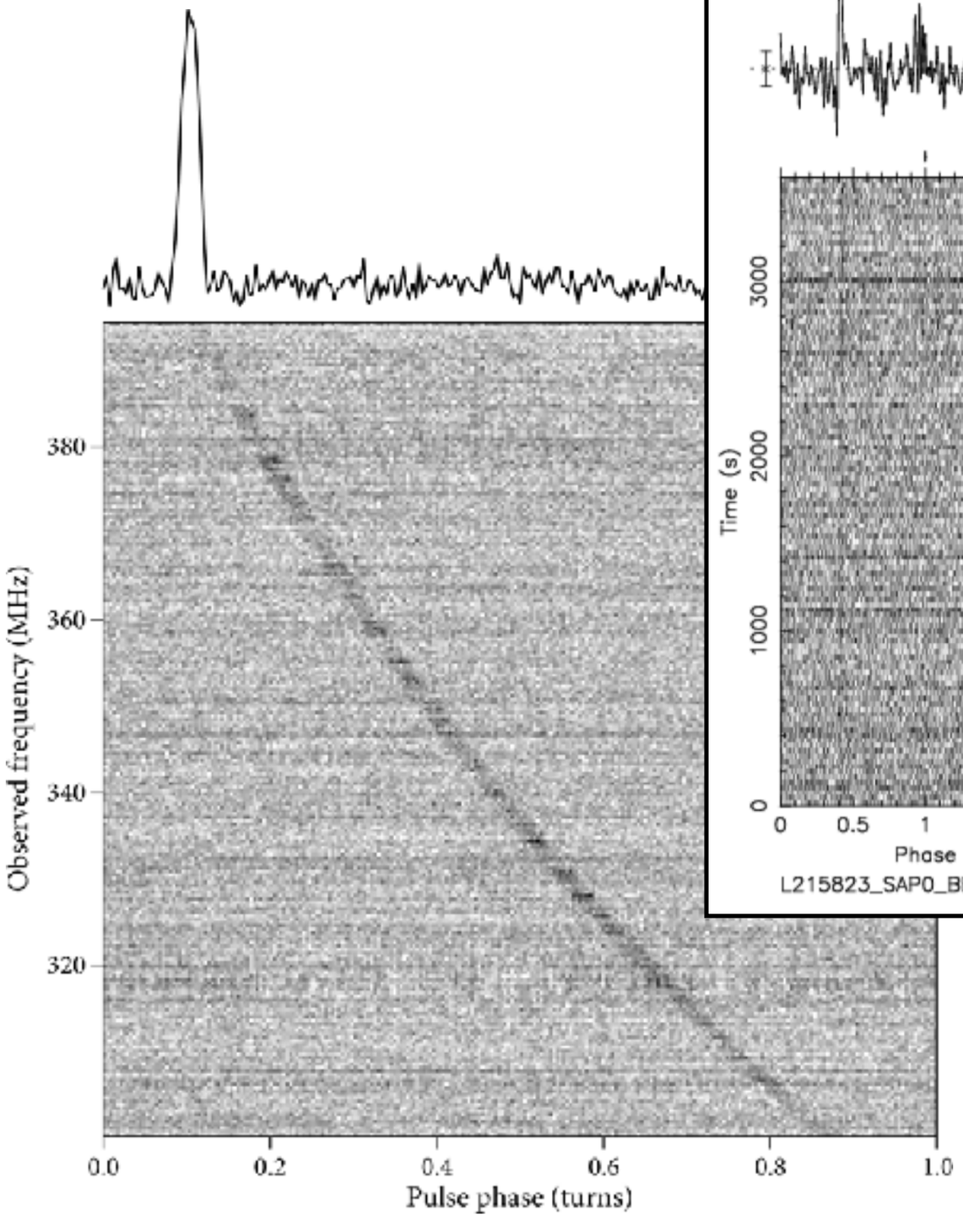
https://www.cv.nrao.edu/~sransom/web/Ch6.html

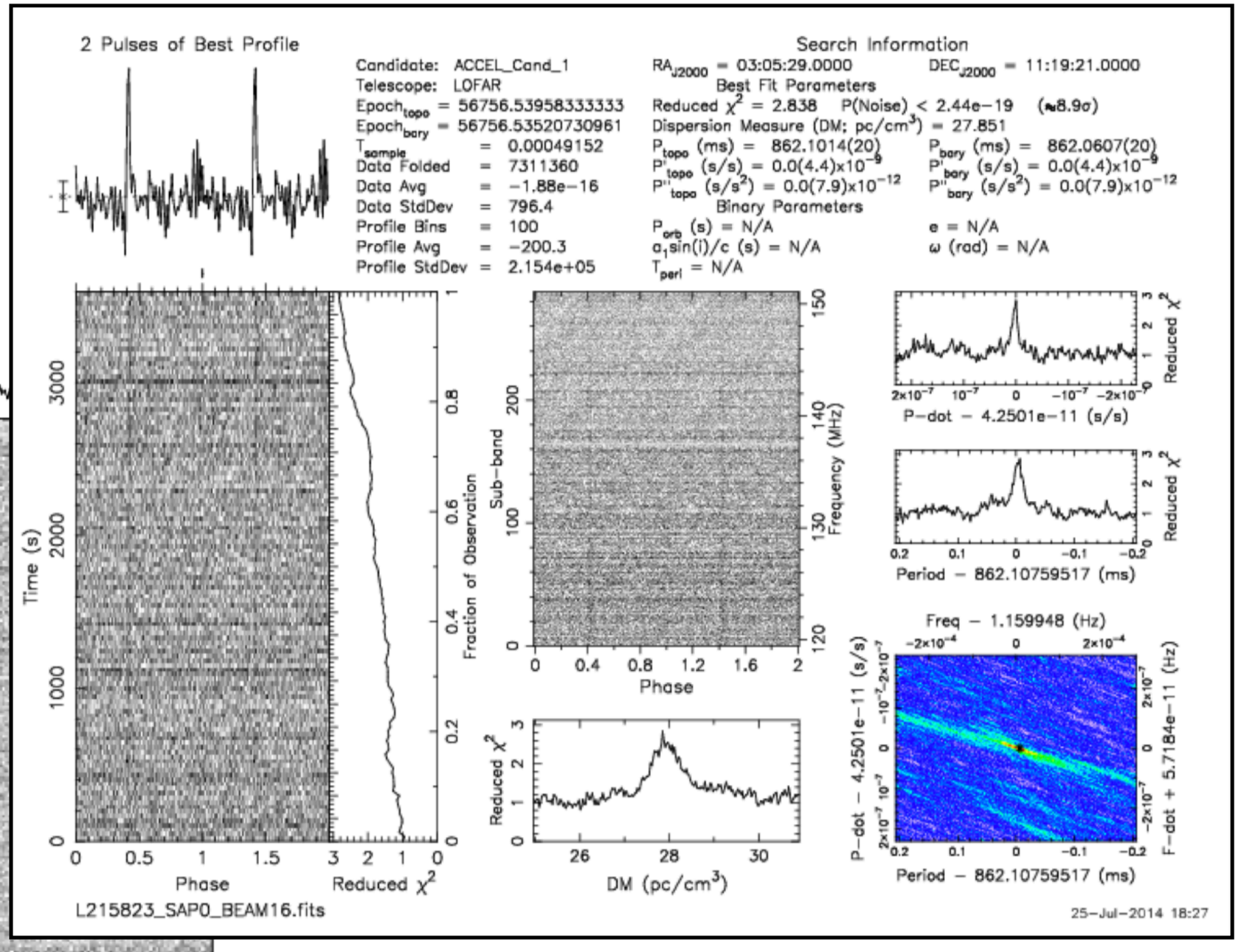
DM-SNR CURVE



https://www.cv.nrao.edu/~sransom/web/Ch6.html

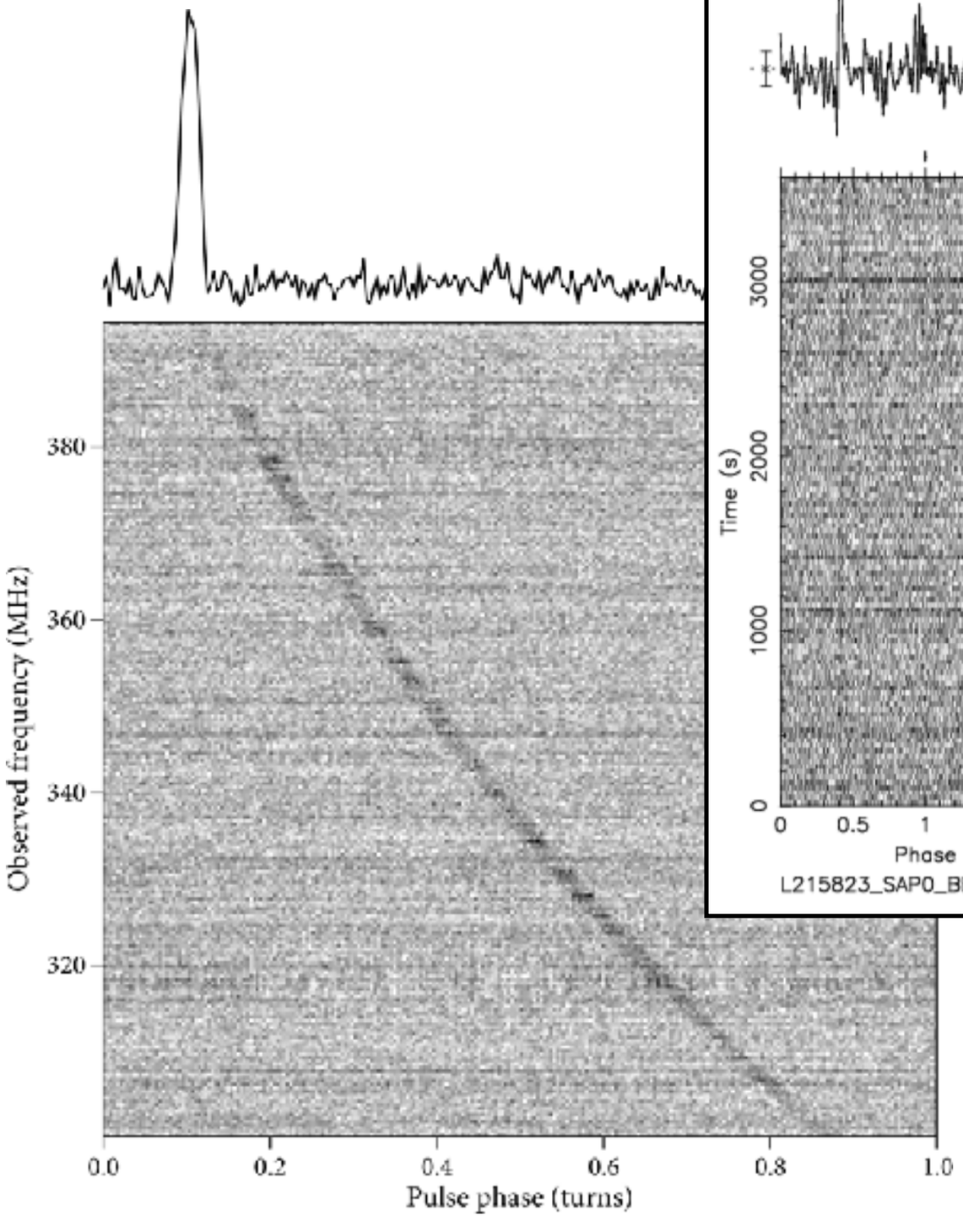
DM-SNR CURVE

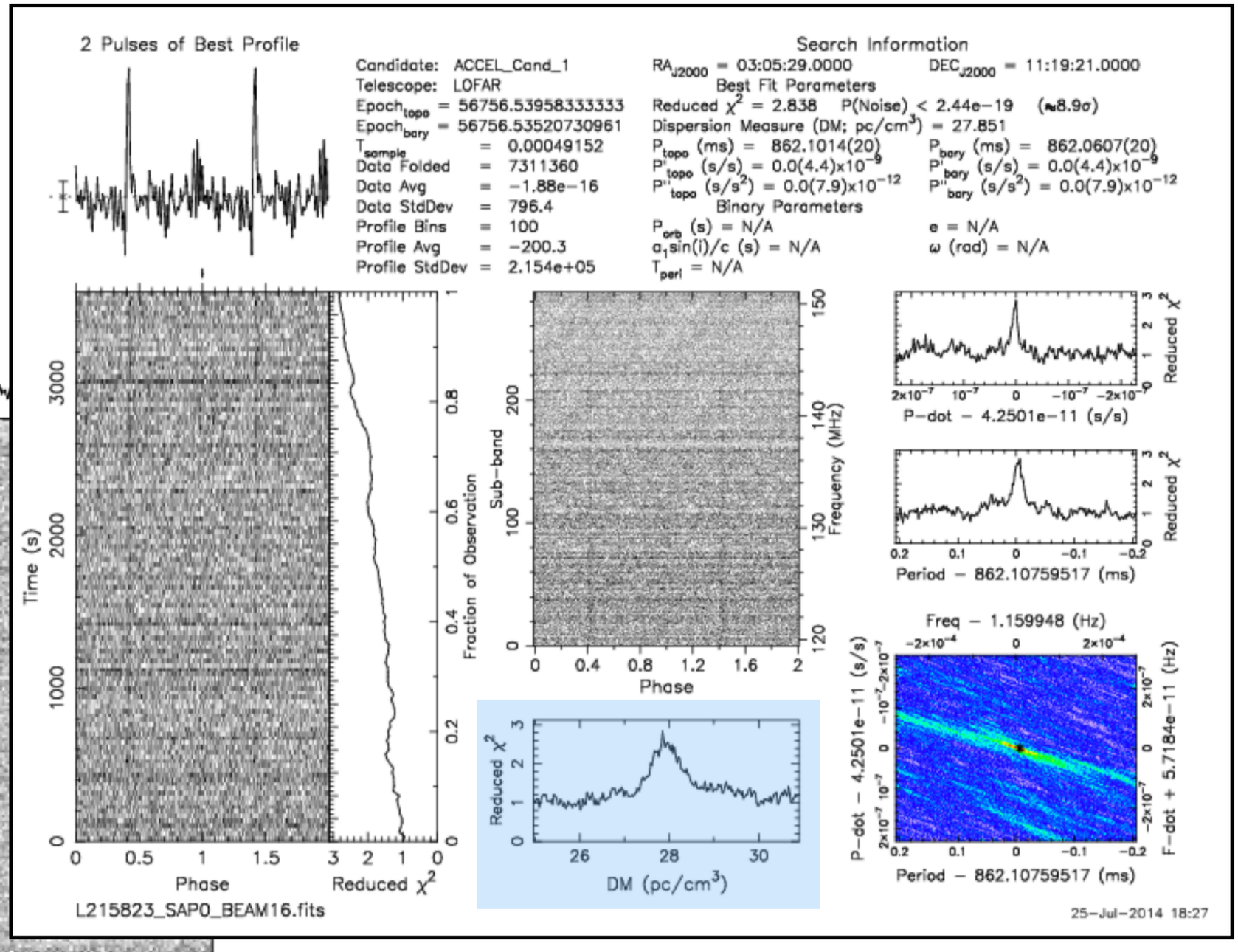




https://www.cv.nrao.edu/~sransom/web/Ch6.html

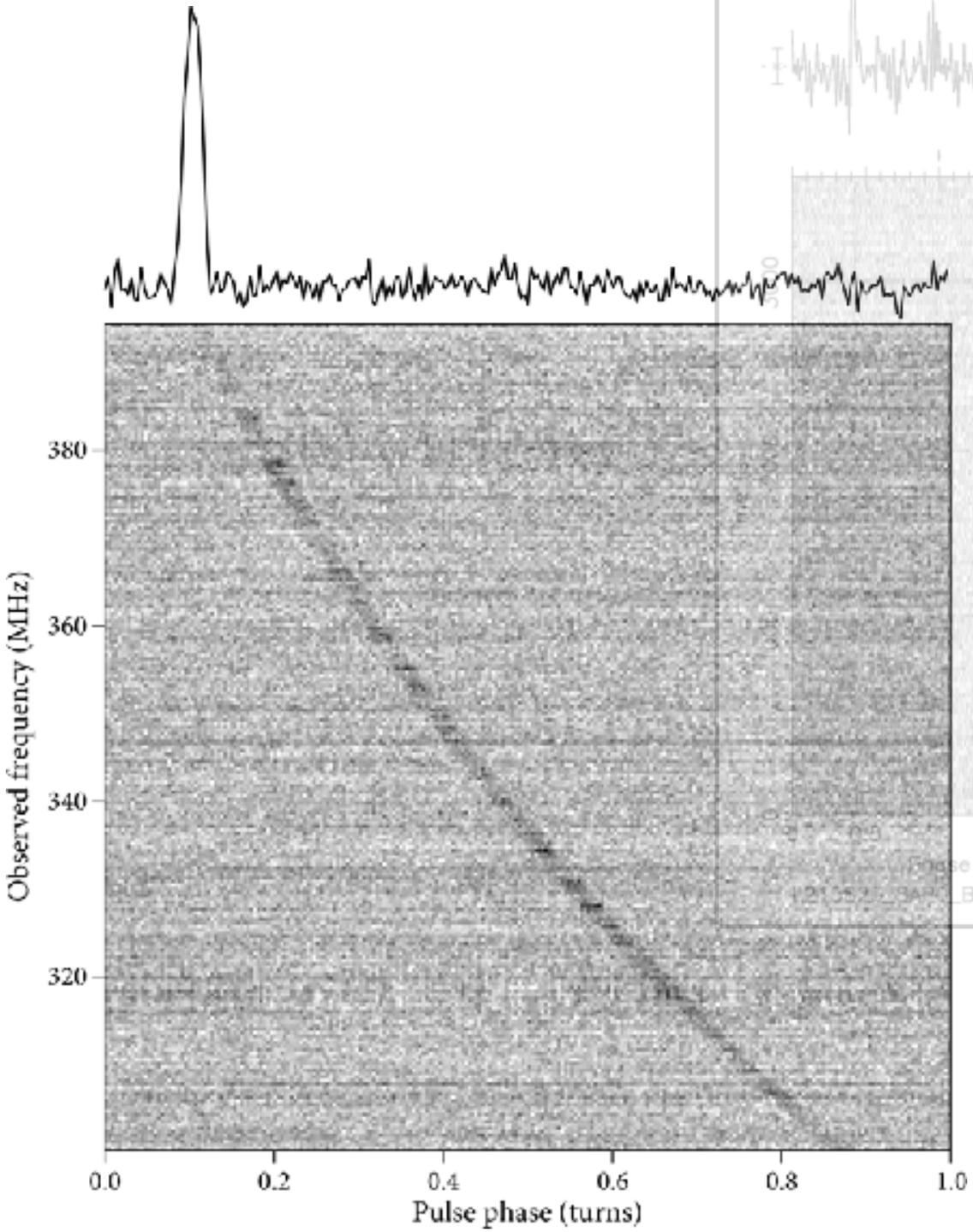
DM-SNR CURVE

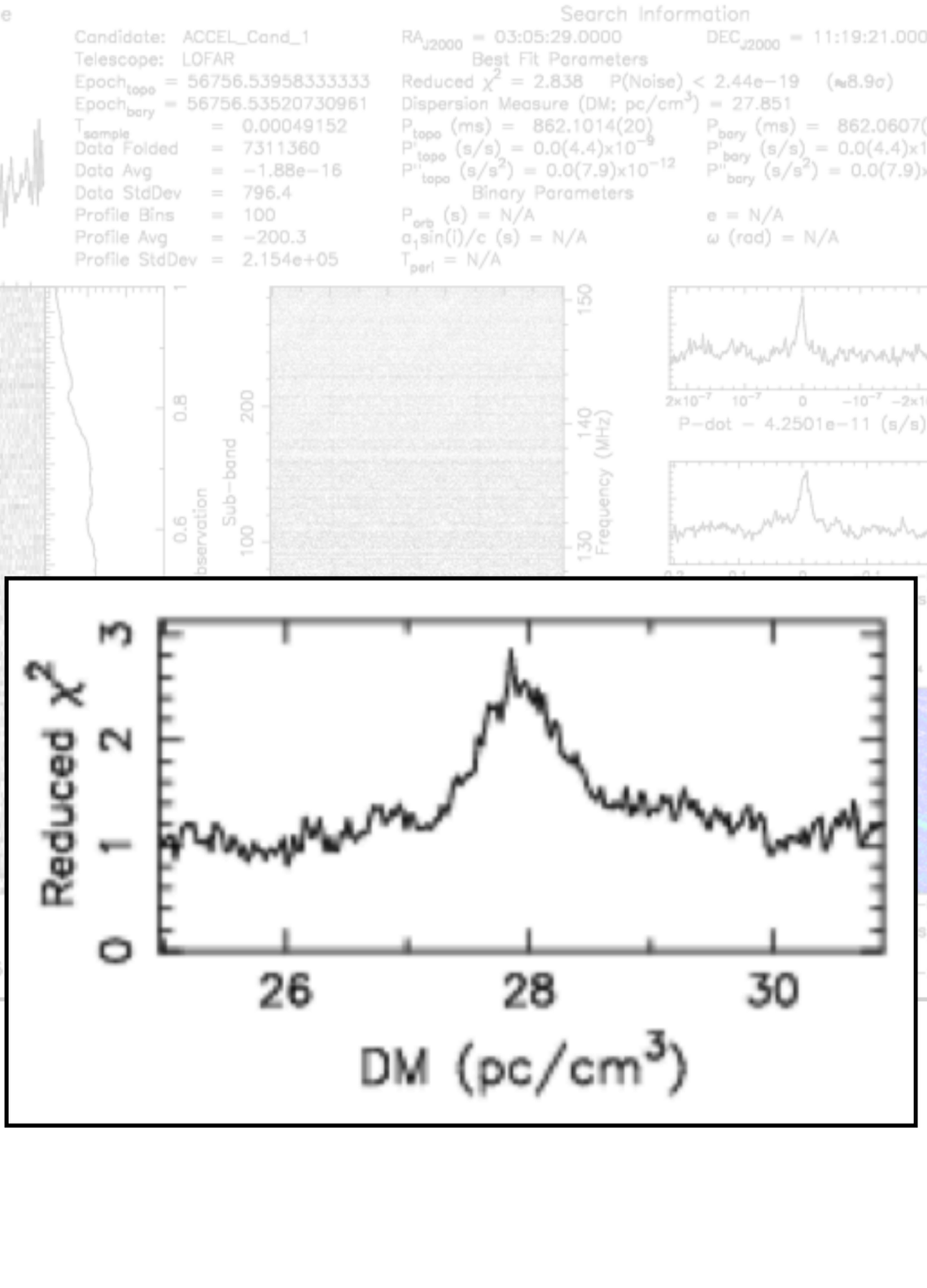




https://www.cv.nrao.edu/~sransom/web/Ch6.html

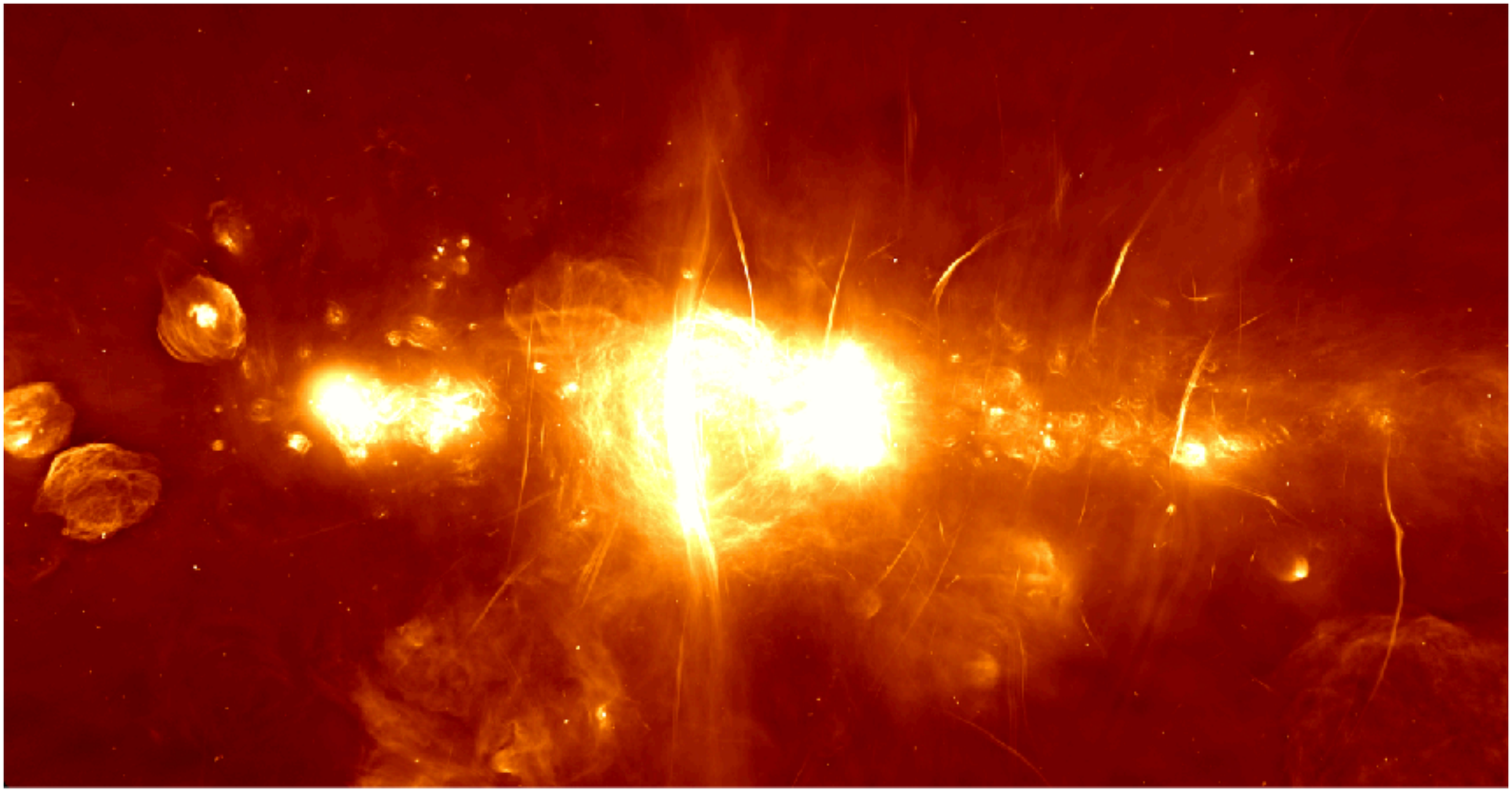
DM-SNR CURVE





http://www.jb.man.ac.uk/distance/frontiers/pulsars/section3.html

WHAT IS A PULSAR?

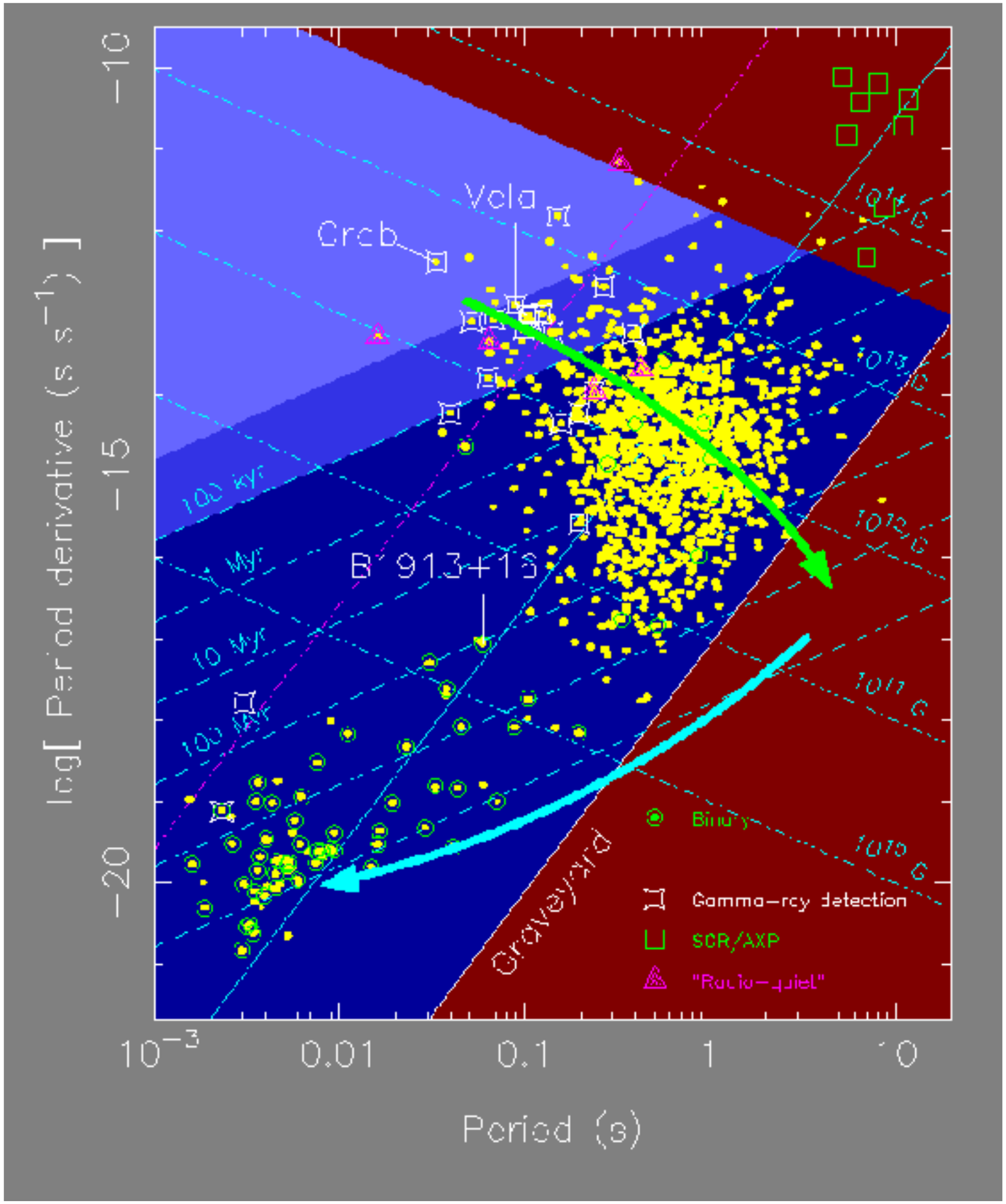


http://www.jb.man.ac.uk/distance/frontiers/pulsars/section3.html

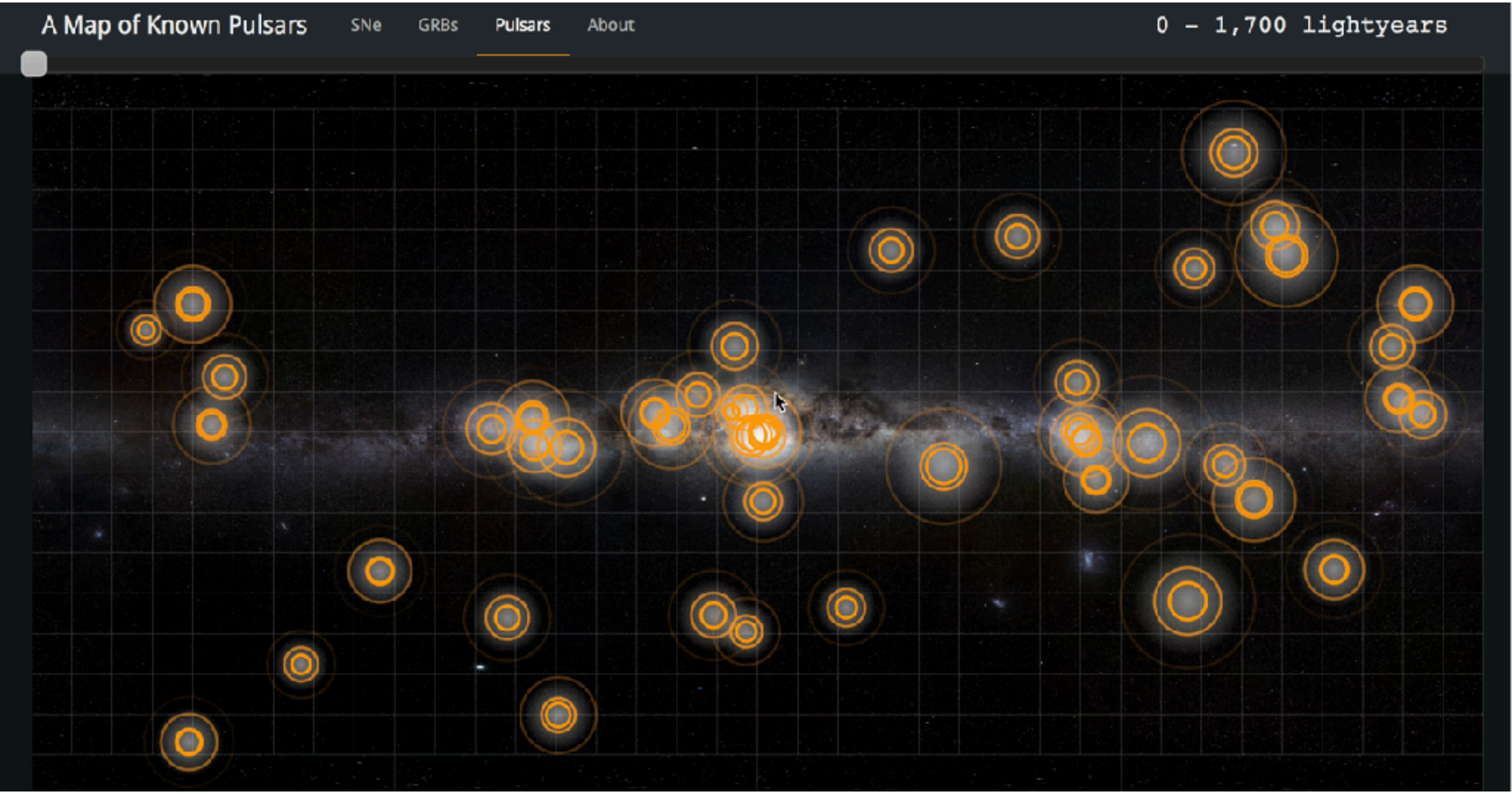
CRAB NEBULA, A.D. 1054

http://www.jb.man.ac.uk/distance/frontiers/pulsars/section3.html

WHAT IS A PULSAR?

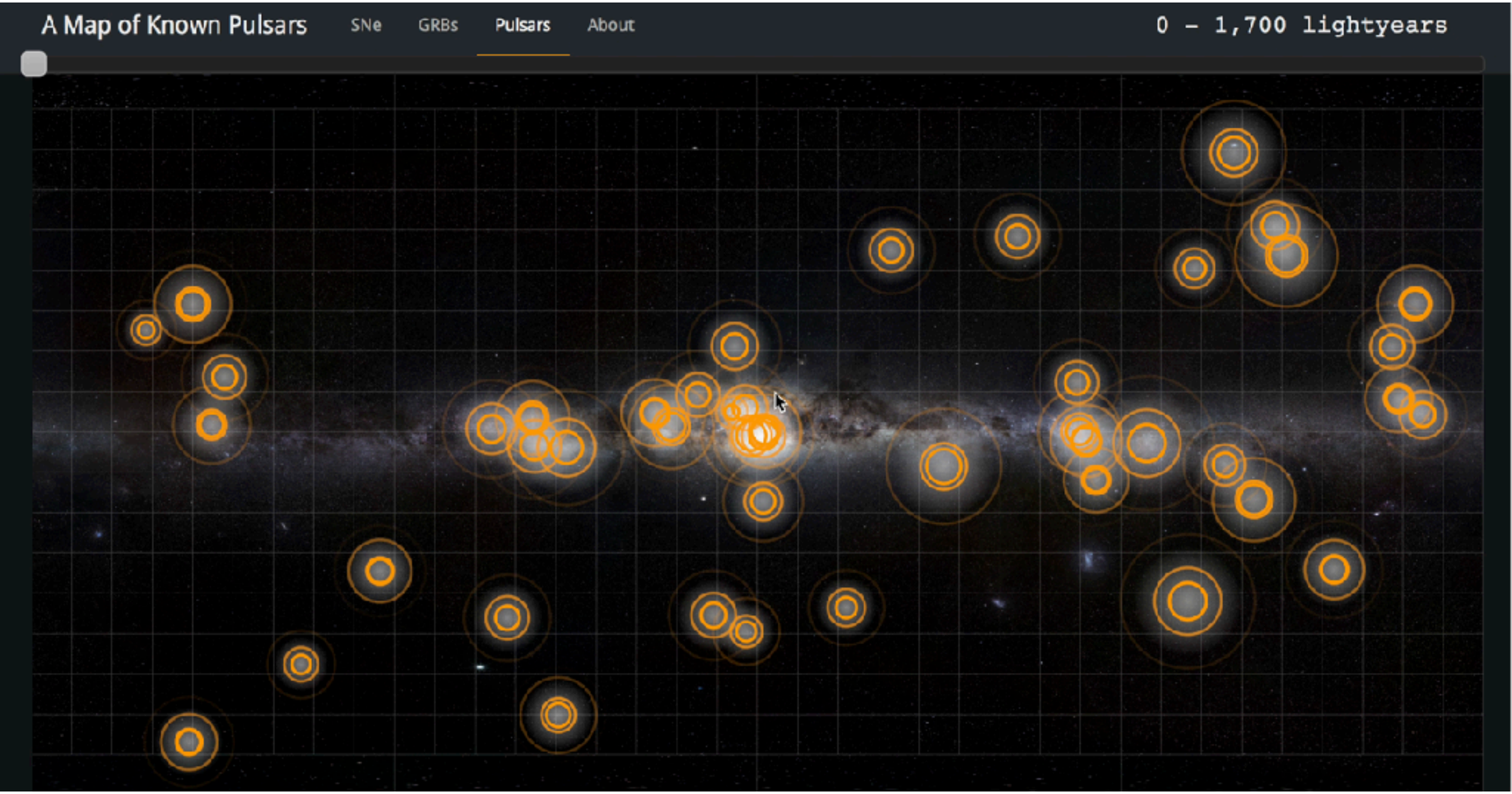


http://www.jb.man.ac.uk/distance/frontiers/pulsars/section3.html



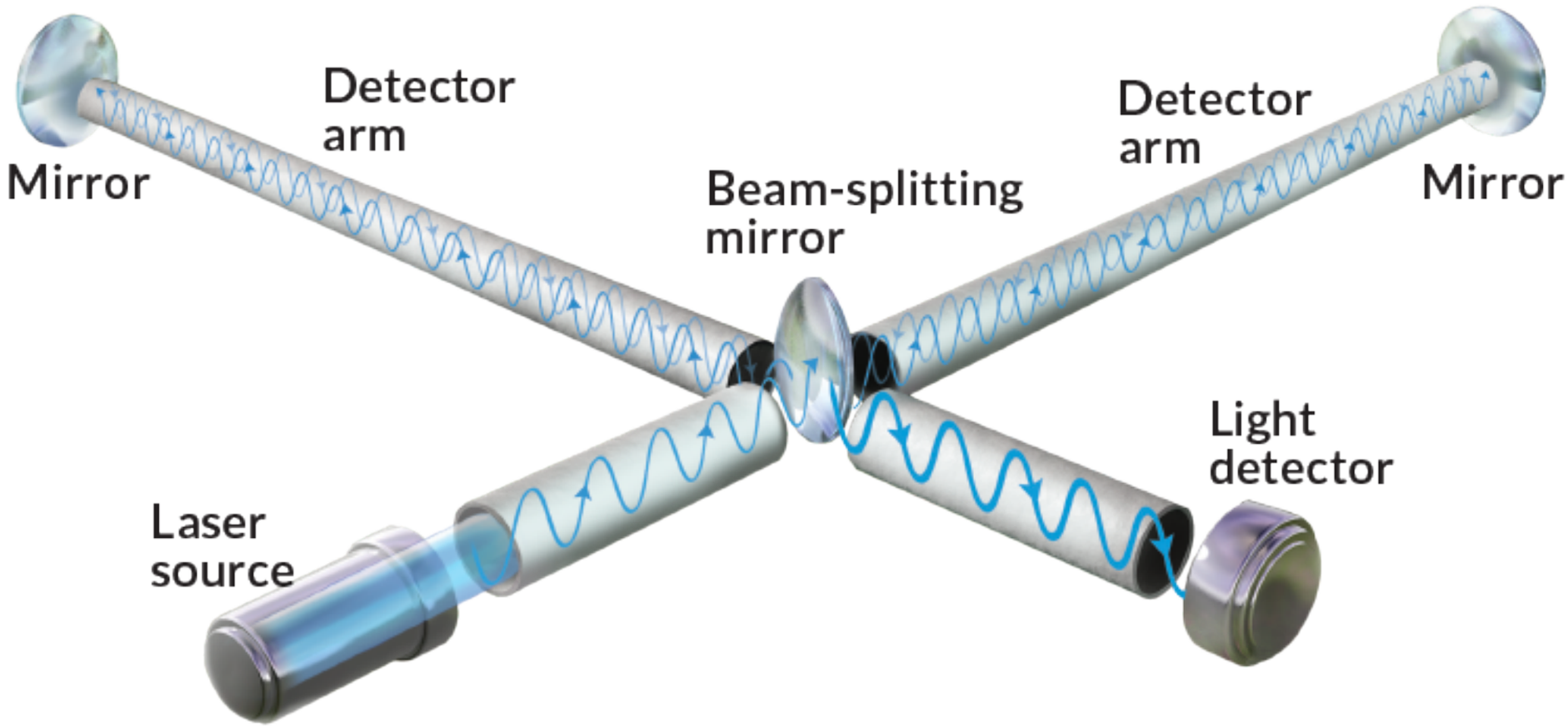
http://www.ishivvers.com/maps/pulsars.html

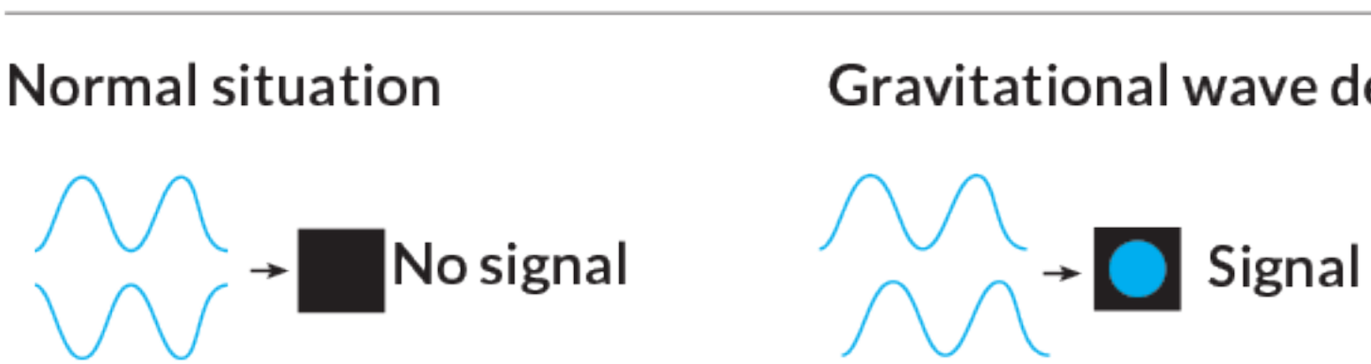
http://www.jb.man.ac.uk/distance/frontiers/pulsars/section3.html



http://www.ishivvers.com/maps/pulsars.html

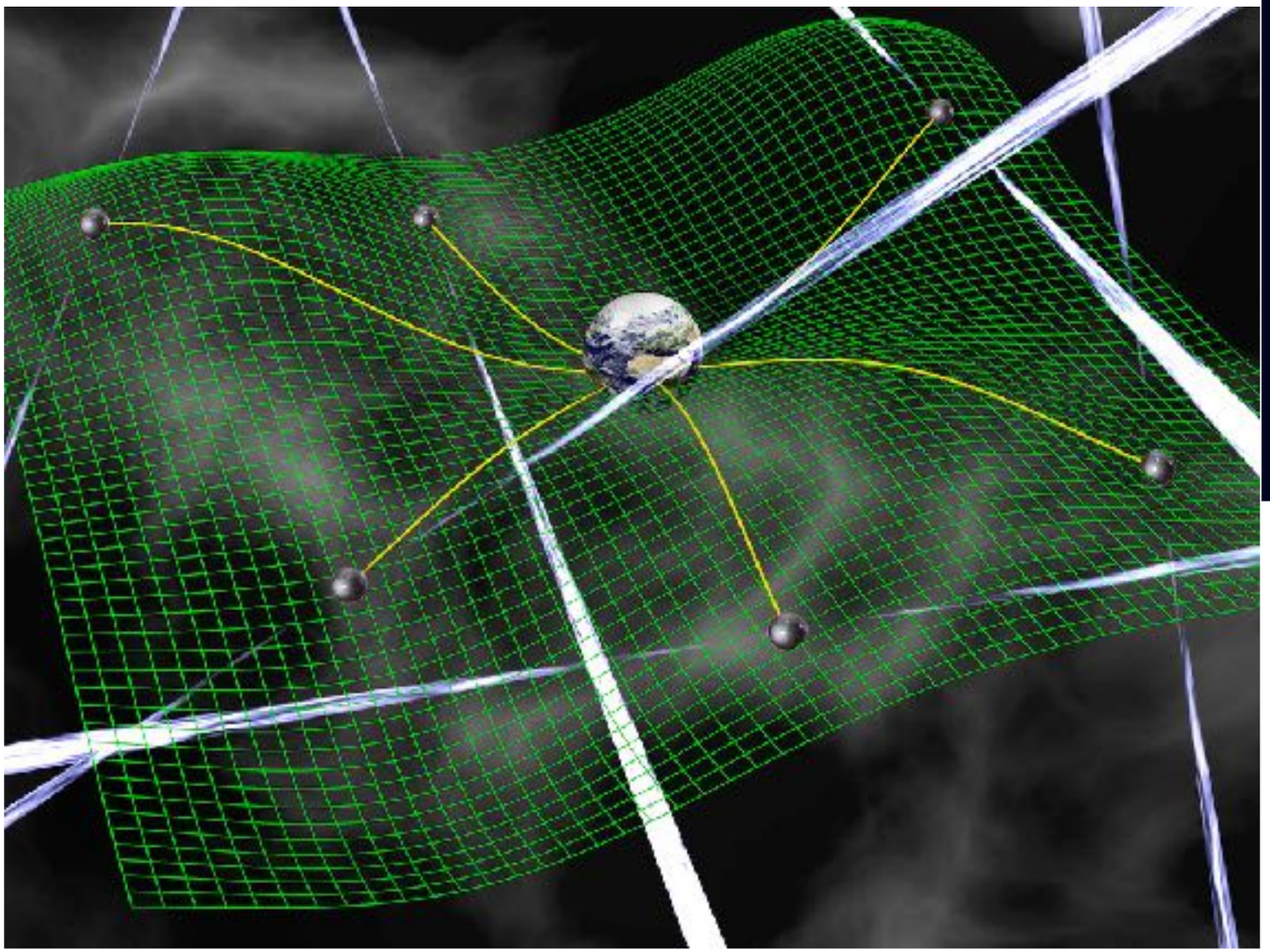
MEASURING GRAVITATIONAL WAVES



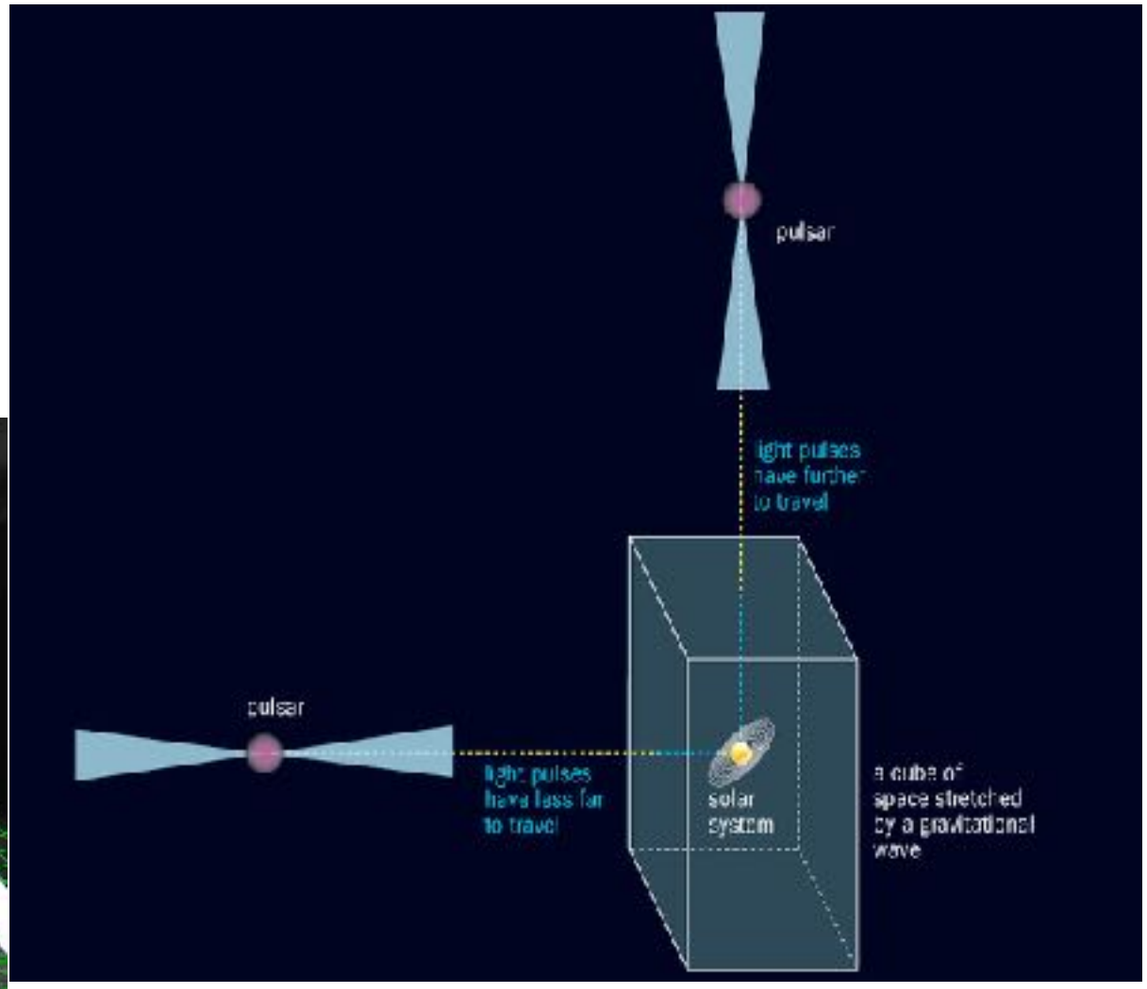


https://www.sciencenews.org/article/gravitational-waves-explained

MEASURING GRAVITATIONAL WAVES



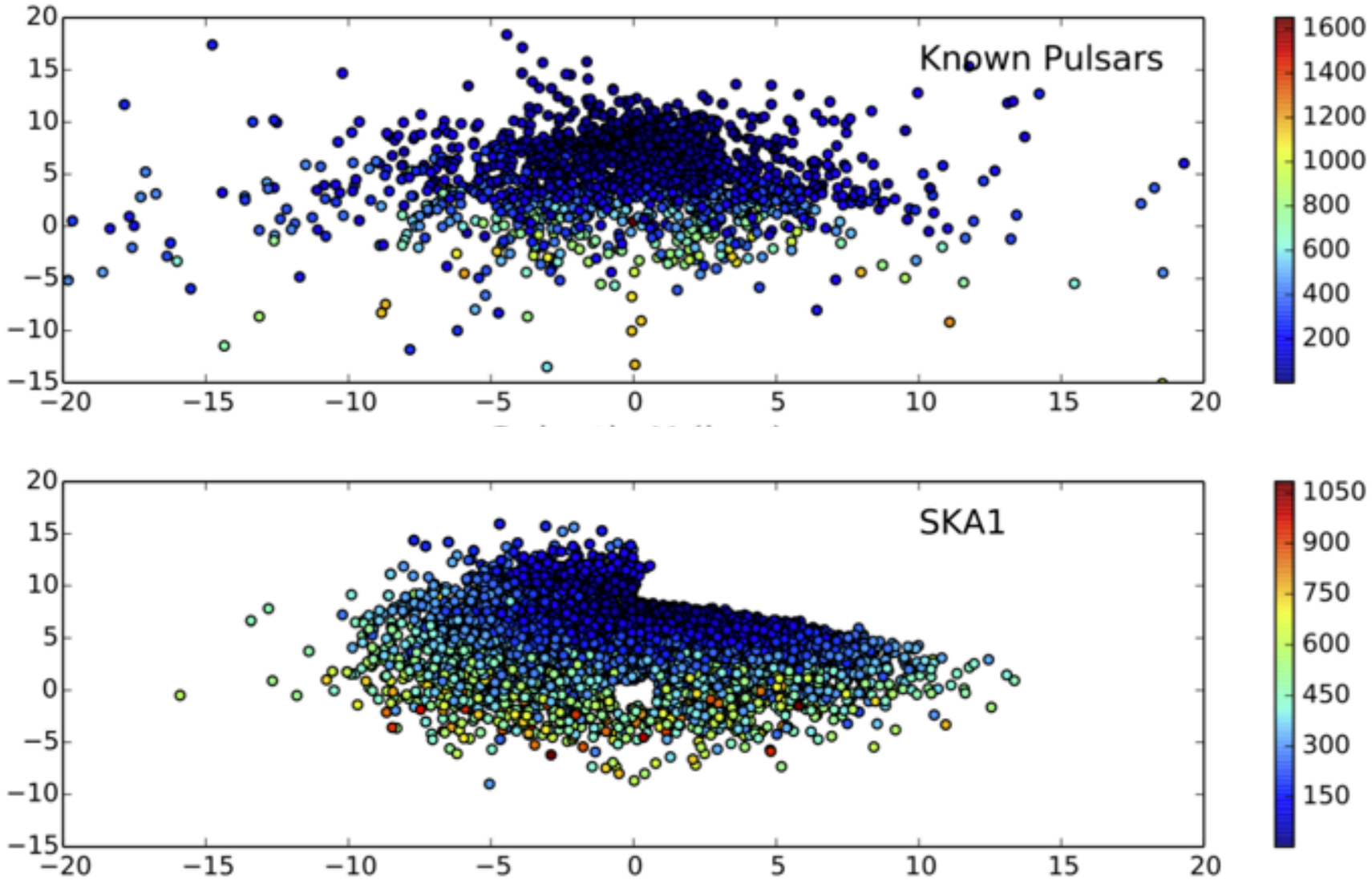
Credit: David Champion



http://live.iop-pp01.agh.sleek.net/2014/09/25/hunting-gravitational-waves-using-pulsars/

Keane et al. 2015: https://hal-insu.archives-ouvertes.fr/insu-01343105

PULSARS WITH THE SKA



SKA1-MID is expected to find **10,000 new pulsars**

https://arxiv.org/pdf/1603.05166.pdf

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.

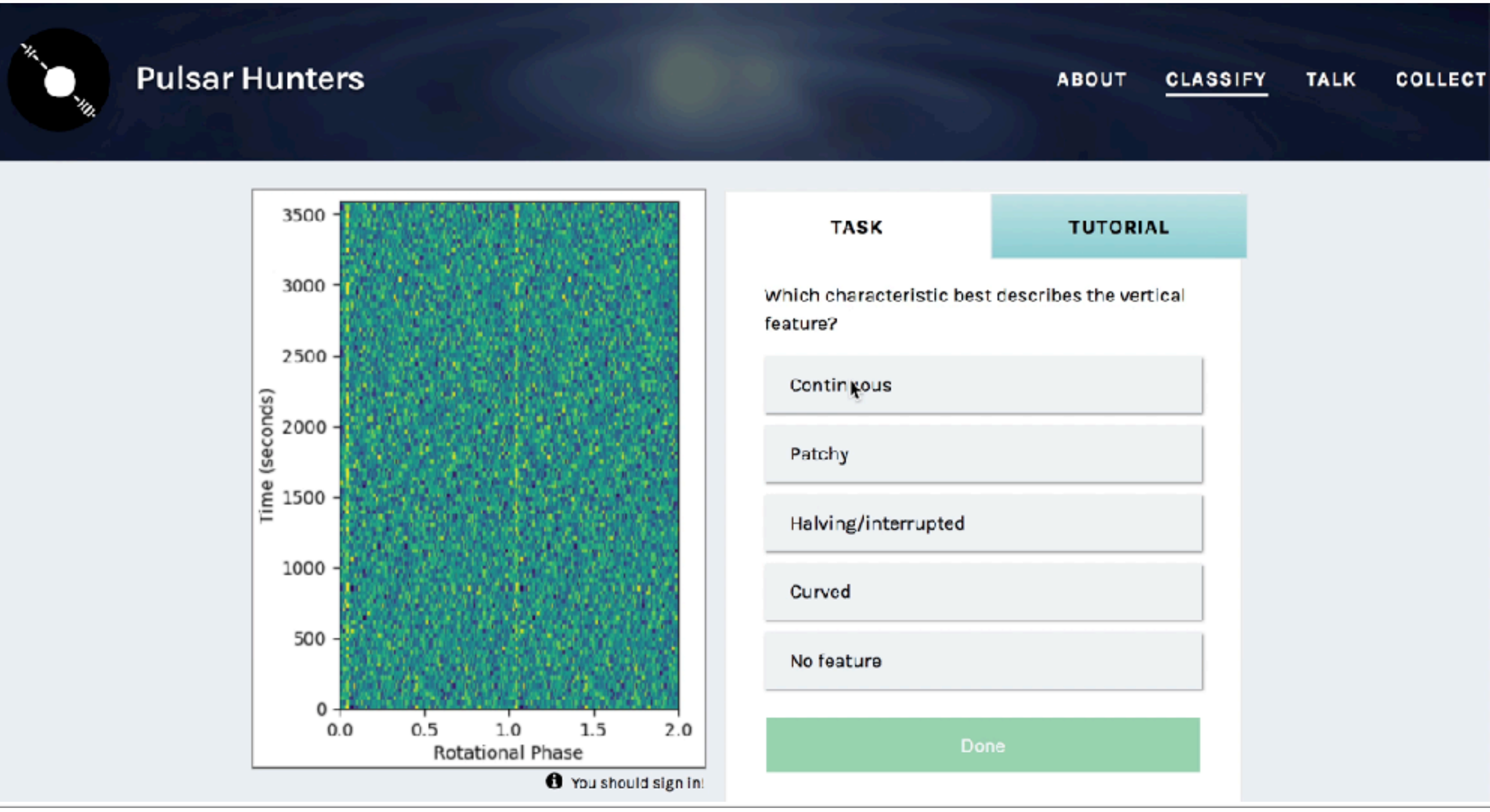
- We need to deal with the **rising candidate numbers**.
- **Data rate is also rising**… we can no longer store all the data and repeatedly re-process it - we need to be right the first time around (or as often as possible).

https://arxiv.org/pdf/1603.05166.pdf

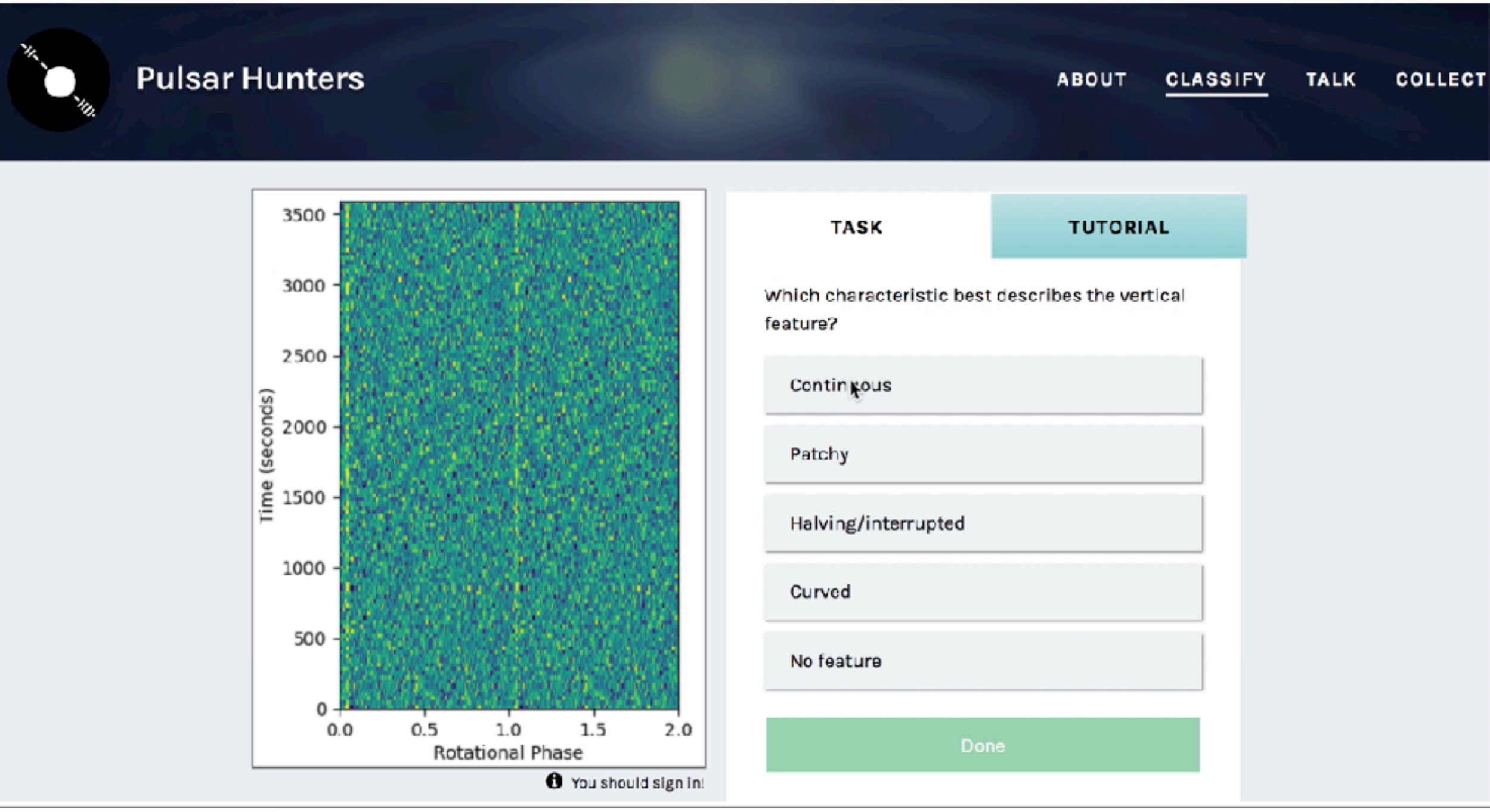
Survey	Year	Candidates	Per Sq. Degree
2nd Molonglo Survey(Manchester et al. 1978)	1977	2,500	~0.1
Phase II survey (Stokes et al. 1986)	1983	5,405	~1
Parkes 20 cm survey (Johnston et al. 1992)	1988	~ 150,000	~188
Parkes Southern Pulsar Survey (Manchester et al. 1996)	1991	40,000	~2
Parkes Multibeam Pulsar 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.

- There are far more non-pulsars than true pulsars… this is a machine learning **class imbalance problem**.
- Typically there are ~10,000 non-pulsars for every true pulsar.



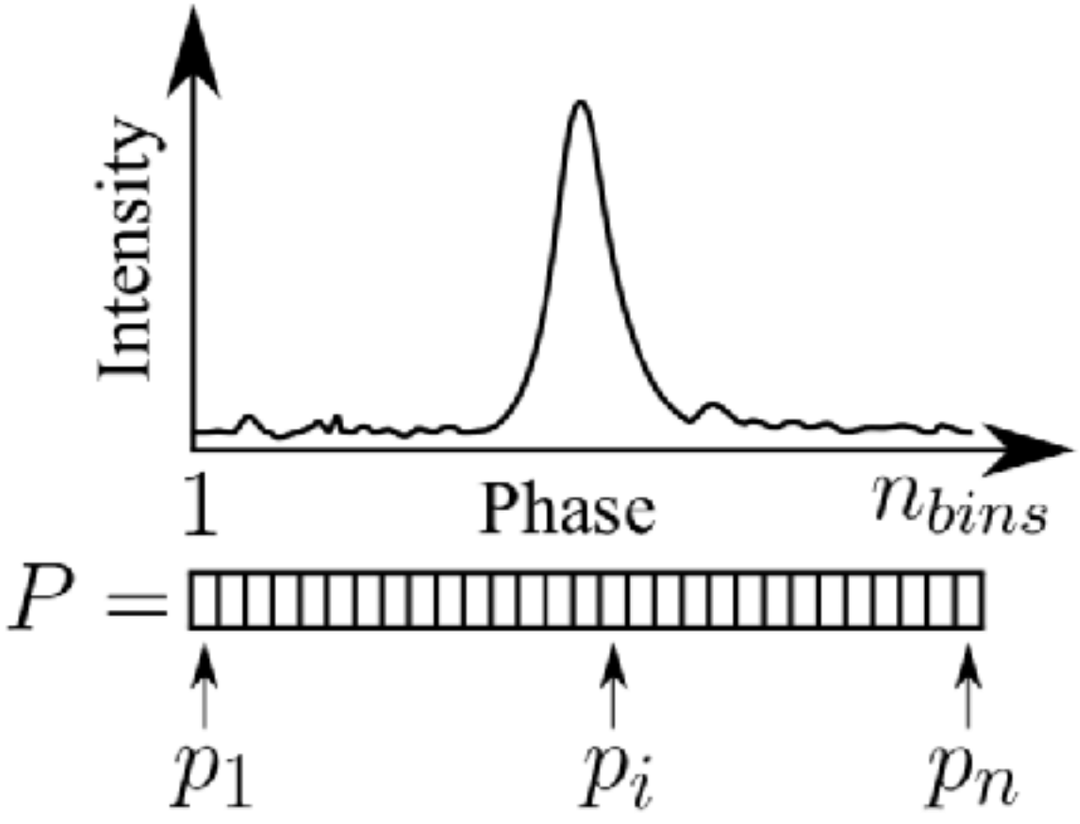
https://www.zooniverse.org/projects/zooniverse/pulsar-hunters

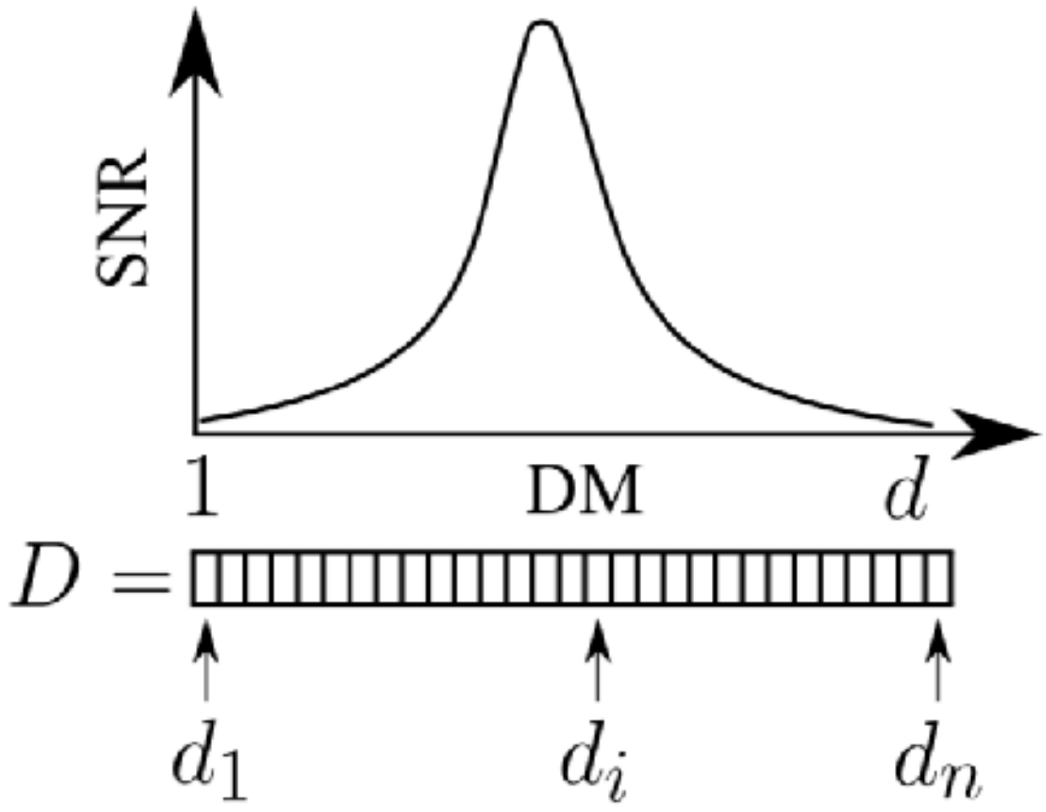


https://www.zooniverse.org/projects/zooniverse/pulsar-hunters

Lyon et al. 2016

For each pulsar candidate we have **8** pieces of data - these are the ***features.***





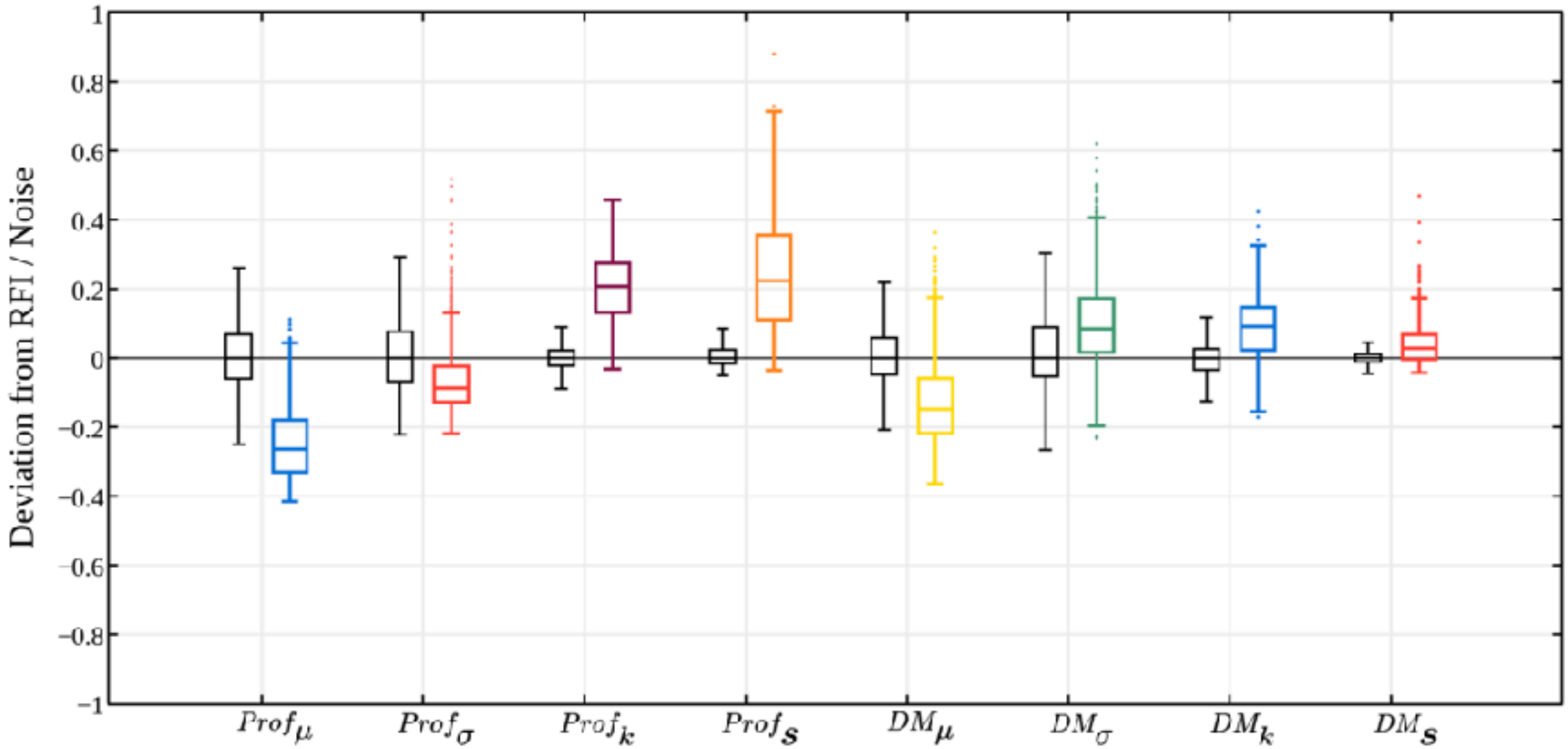
Integrated Profile

- Mean
- Standard deviation
- Kurtosis
- Skewness

DM-SNR Curve

- Mean
- Standard deviation
- Kurtosis
- Skewness

Lyon et al. 2016

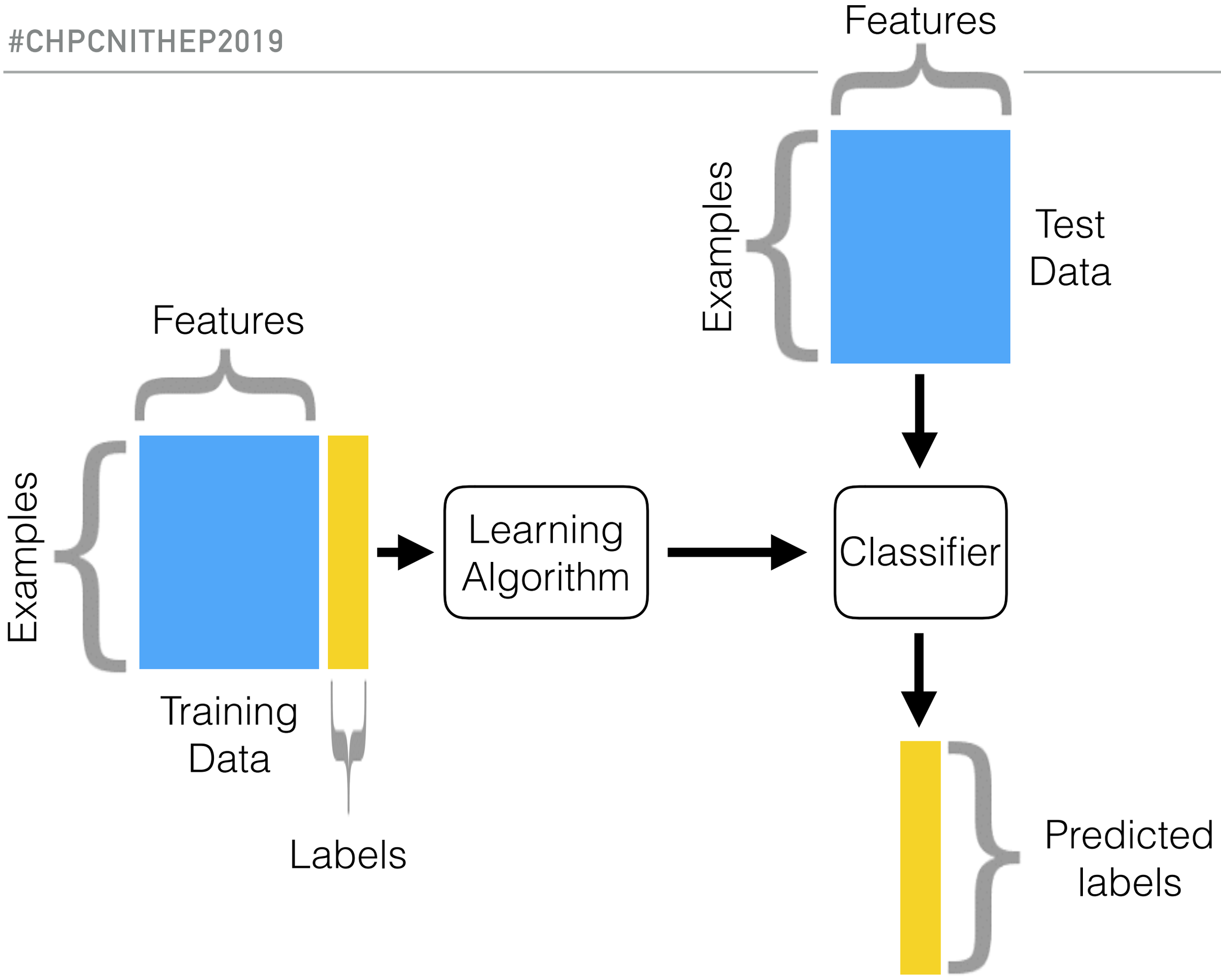


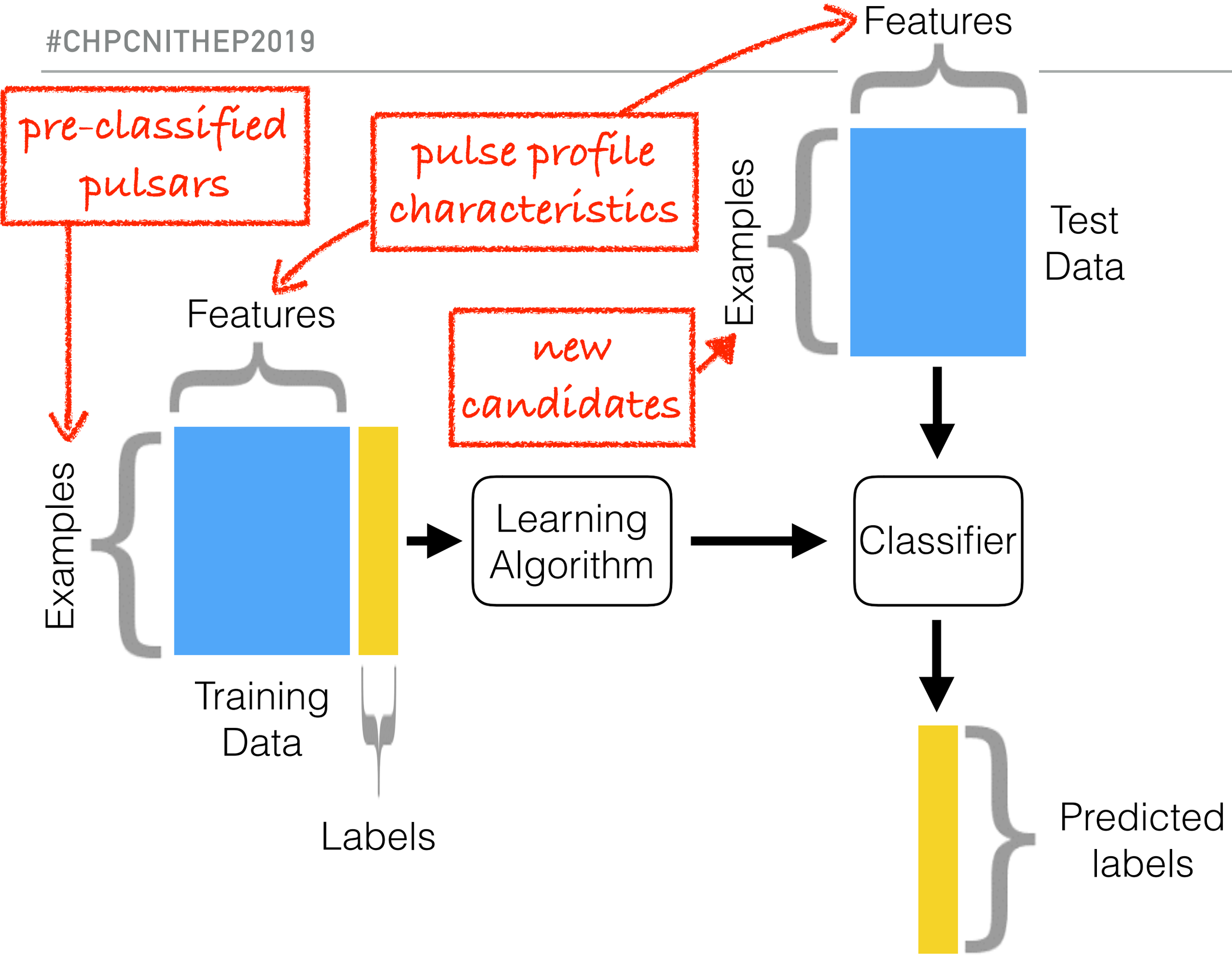
Integrated Profile

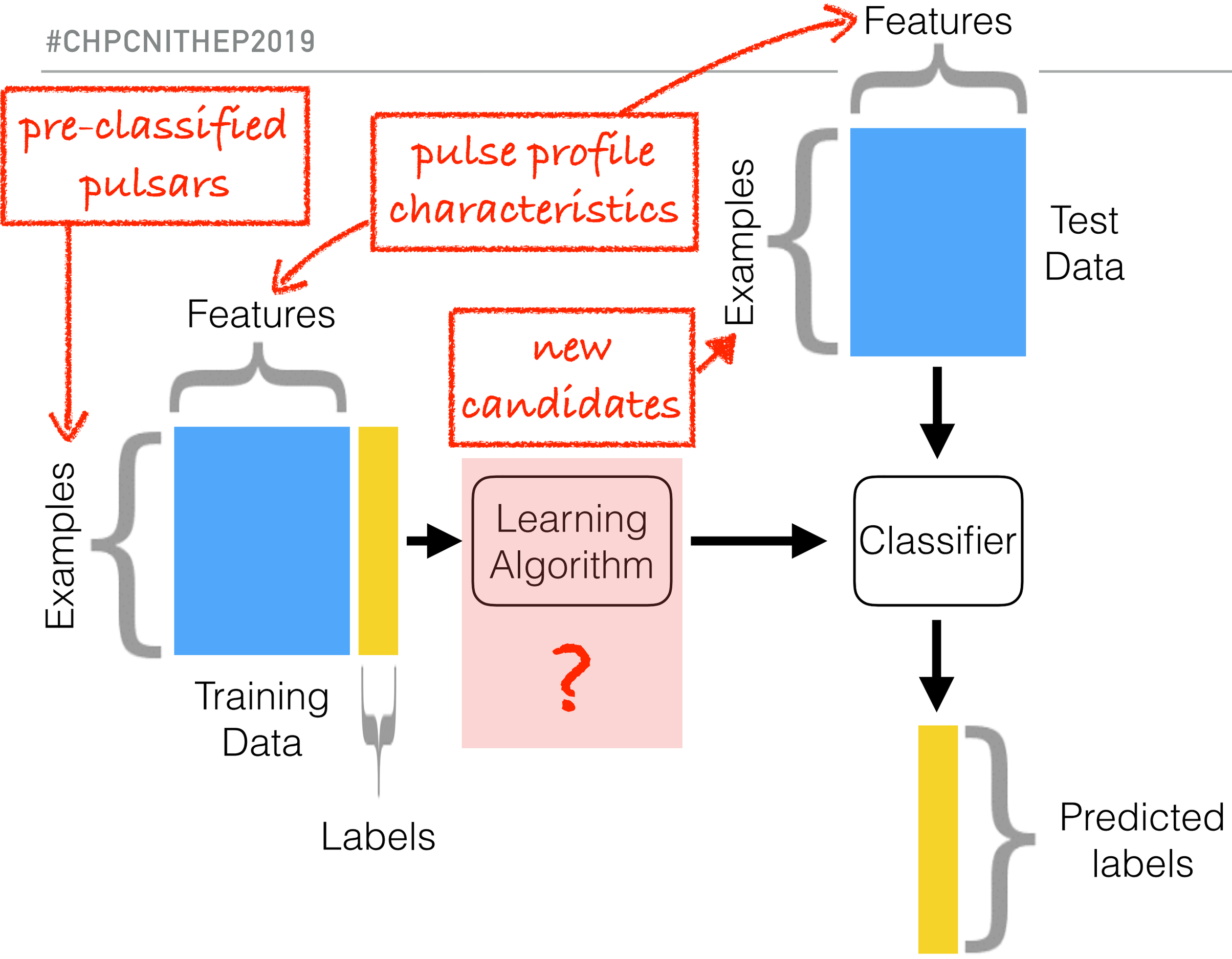
- Mean
- Standard deviation
- Kurtosis
- Skewness

DM-SNR Curve

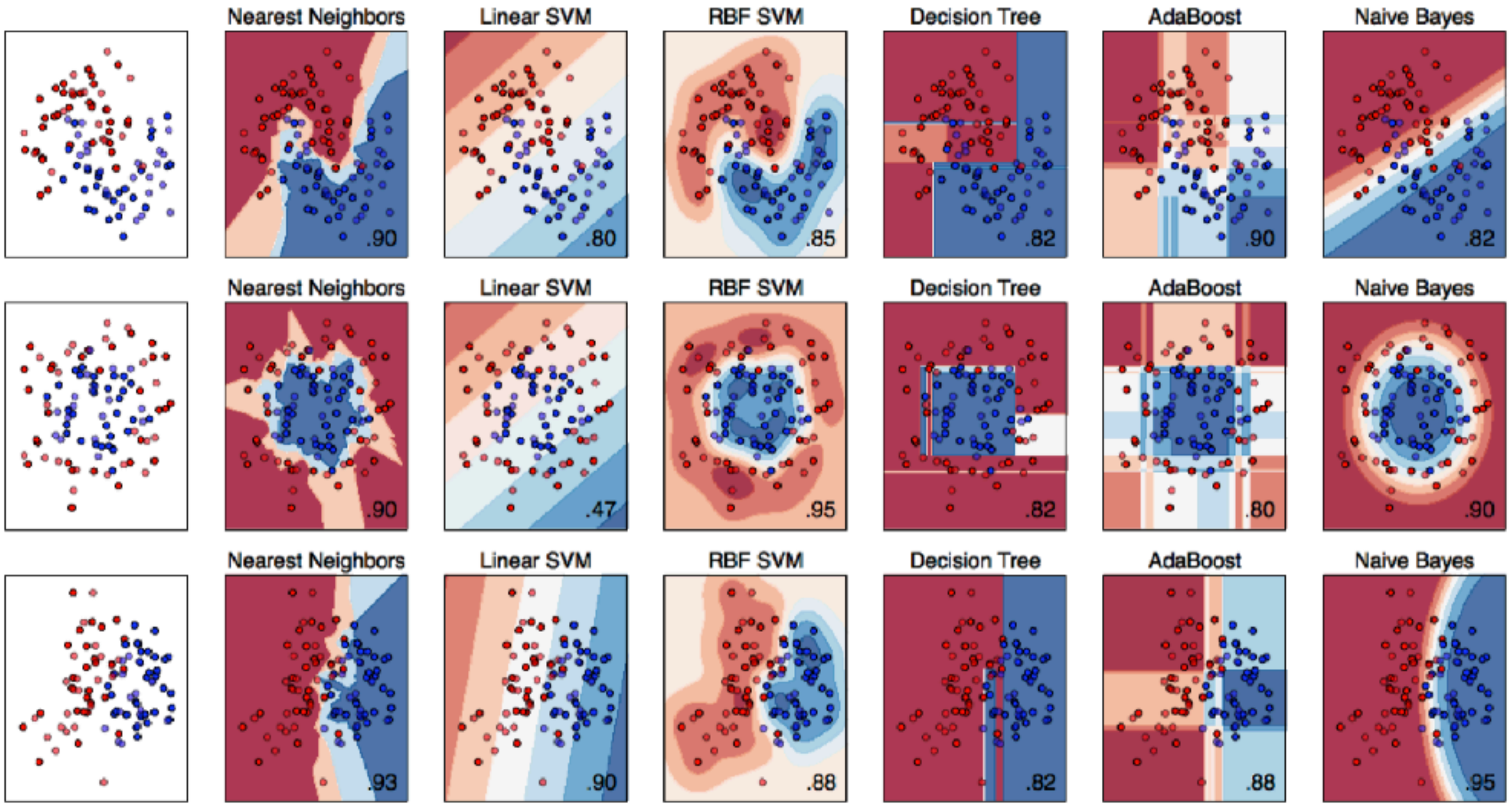
- Mean
- Standard deviation
- Kurtosis
- Skewness


Features
Examples
Training Data
Labels
Learning Algorithm
Classifier
Features
Examples
Test Data
Predicted labels


pre-classified pulsars
pulse profile characteristics
new candidates
Features
Examples
Test Data
Features
Examples
Training Data
Labels
Learning Algorithm
Classifier
Predicted labels

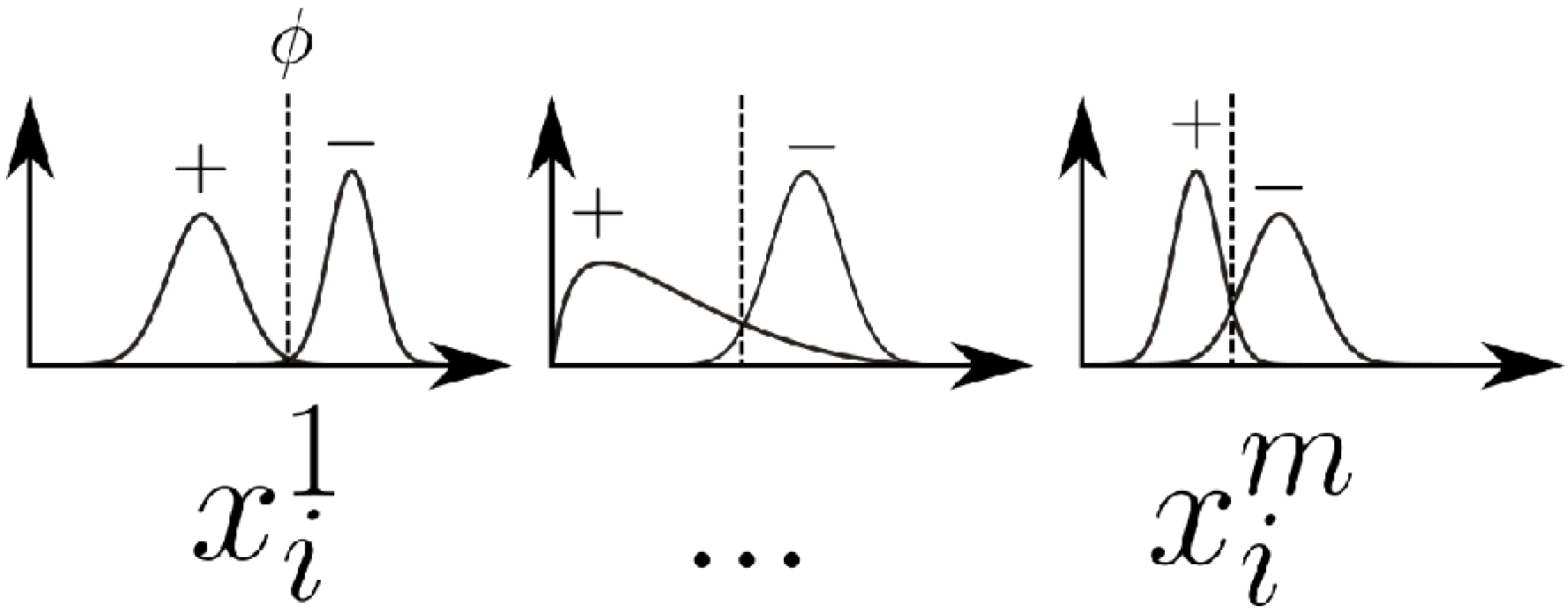

pre-classified pulsars
pulse profile characteristics
new candidates
Features
Examples
Test Data
Features
Examples
Training Data
Labels
Learning Algorithm
?
Classifier
Predicted labels

WHAT KIND OF MACHINE LEARNING?

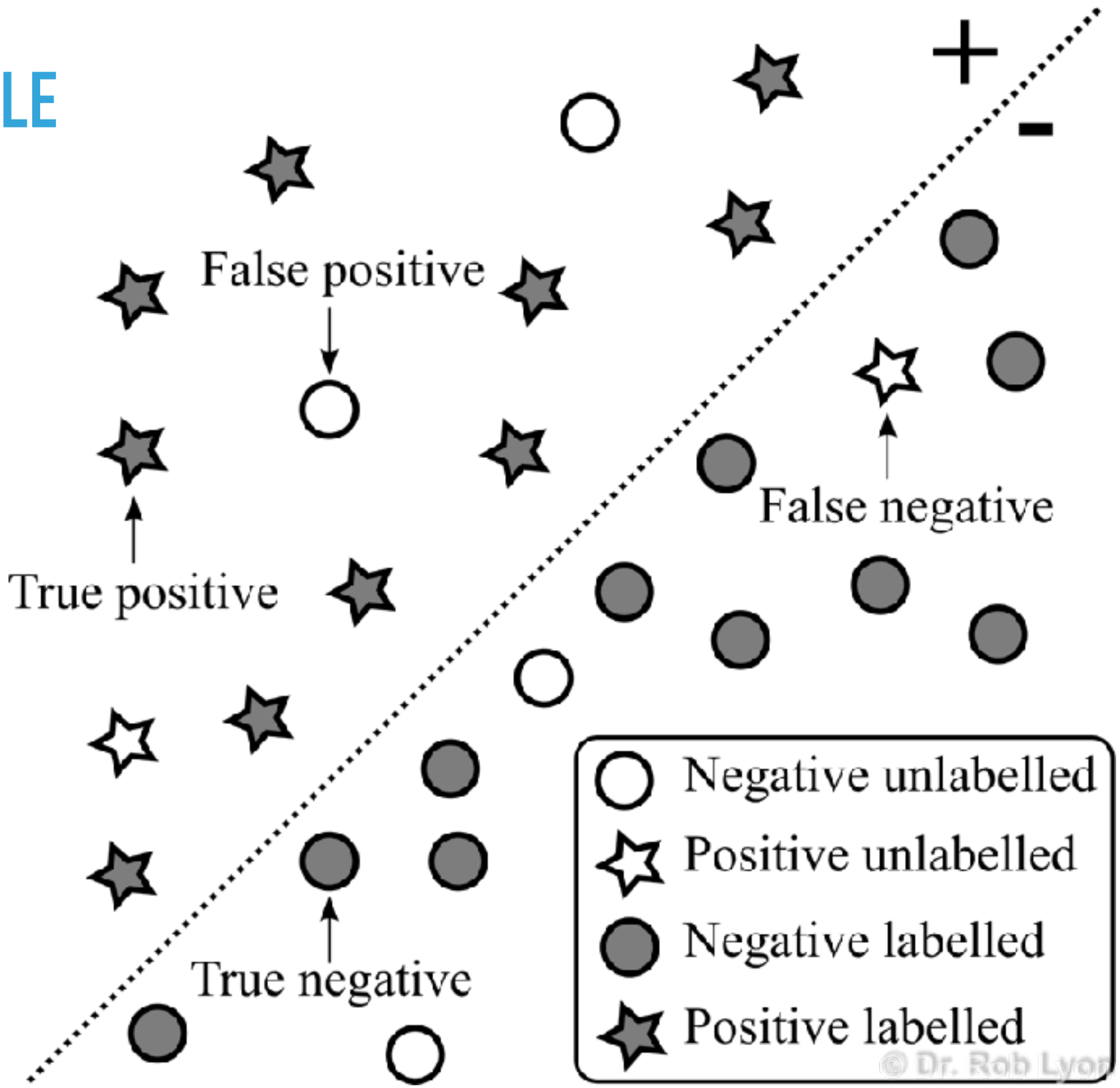


DECISION RULE

if s(x) > ϕ THEN *predict* y = +ve
if s(x) < ϕ THEN *predict* y = -ve



DECISION RULE



	Truth	
Prediction	Positive	Negative
Positive	True Positive (TP)	False Positive (FP) Type I Error
Negative	False Negative (FN) Type II Error	True Negative (TN)

$$Precision = \frac{TP}{TP + FP}$$ % of **+ve** predictions that are truly **+ve**

$$Recall = \frac{TP}{TP + FN}$$ % of true **+ves** that are predicted **+ve**

$$Specificity = \frac{TN}{TN + FP}$$ % of true **-ves** that are predicted **-ve**

EVALUATING CLASSIFIER PERFORMANCE

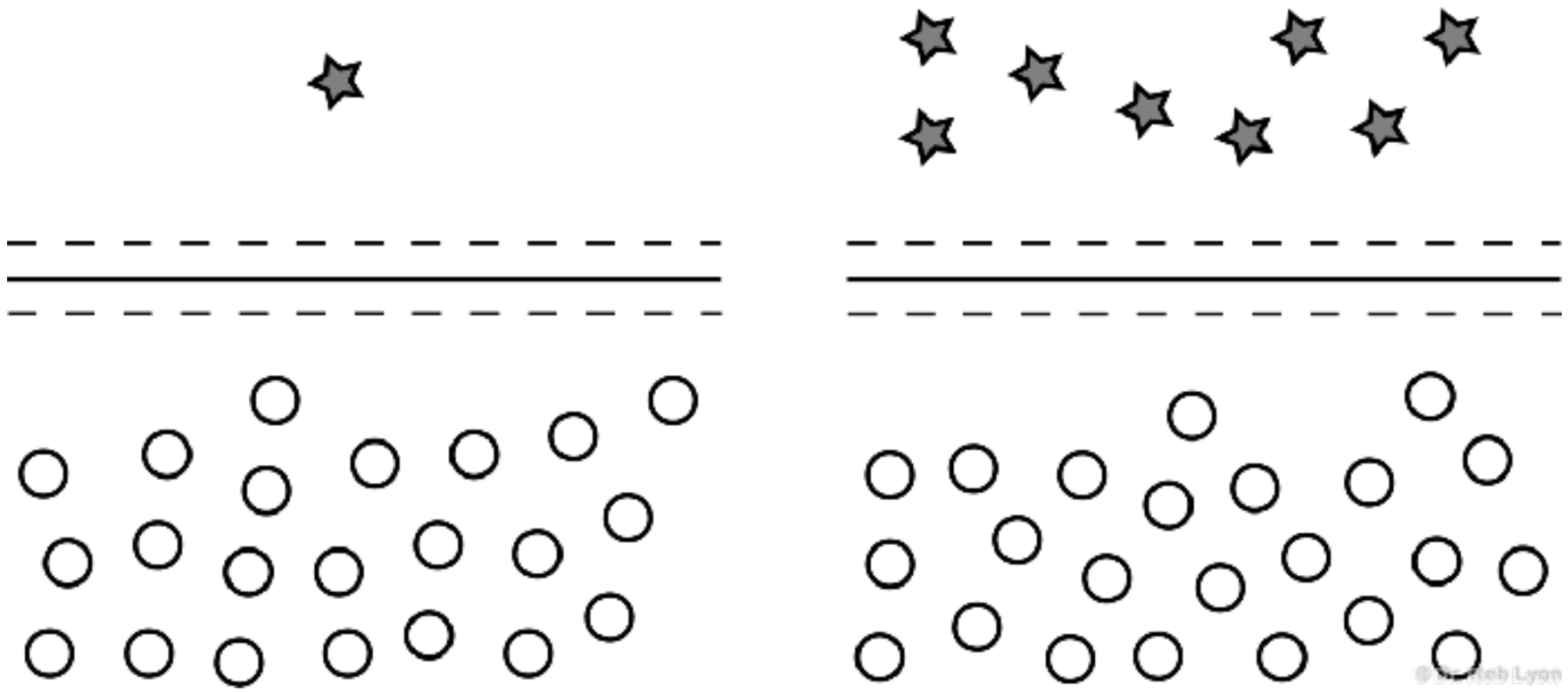
Statistic	Description	Definition
Accuracy	Measure of overall classification accuracy.	$\frac{(TP+TN)}{(TP+FP+FN+TN)}$
False positive rate (FPR)	Fraction of negative instances incorrectly labelled positive.	$\frac{FP}{(FP+TN)}$
G-Mean	Imbalanced data metric describing the ratio between positive and negative accuracy.	$\sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$
Precision	Fraction of retrieved instances that are positive.	$\frac{TP}{(TP+FP)}$
Recall	Fraction of positive instances that are retrieved.	$\frac{TP}{(TP+FN)}$
F-Score	Measure of accuracy that considers both precision and recall.	$2 \times \frac{precision \times recall}{precision+recall}$
Specificity	Fraction of negatives correctly identified as such.	$\frac{TN}{(FP+TN)}$

Lyon et al. 2016

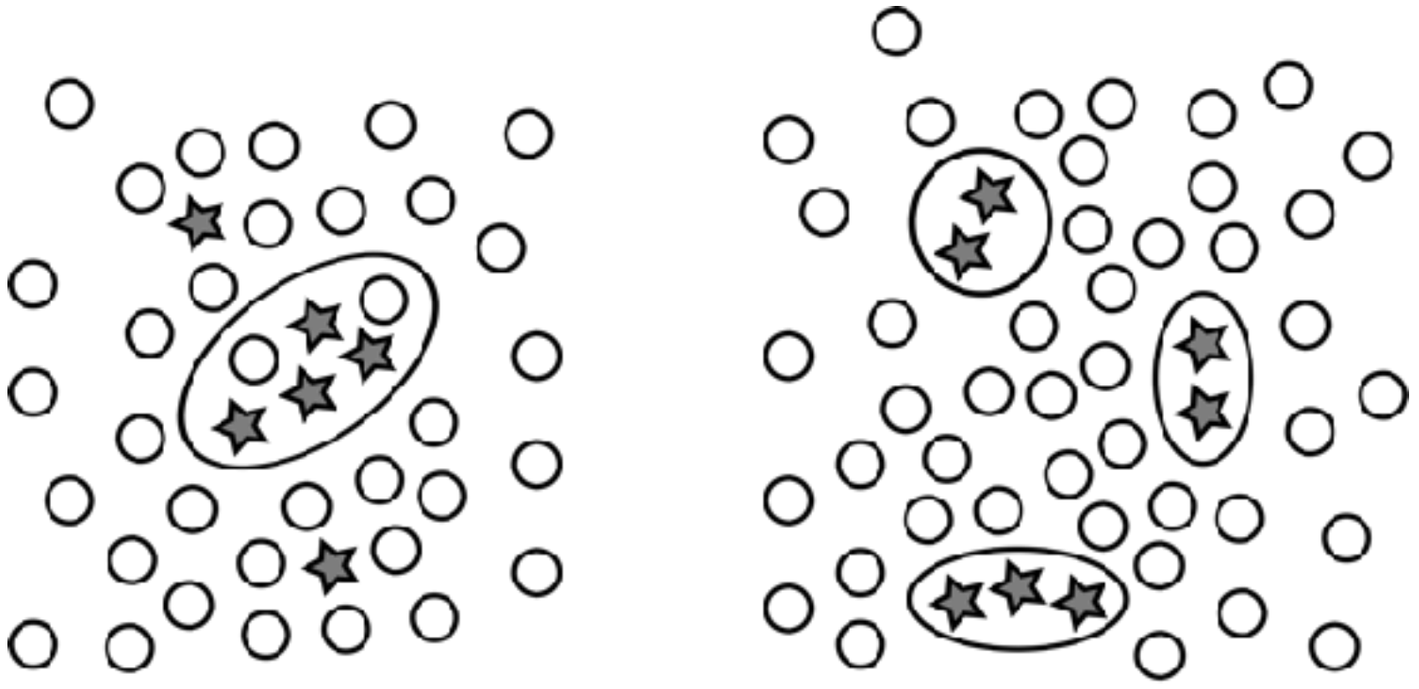
For imbalanced classification problems, we can also treat **rarity** as **importance** and try to minimise the ***expected cost*** of mis-classification:

$$R = (1/p_{-ve}) \times FP + (1/p_{+ve}) \times FN$$

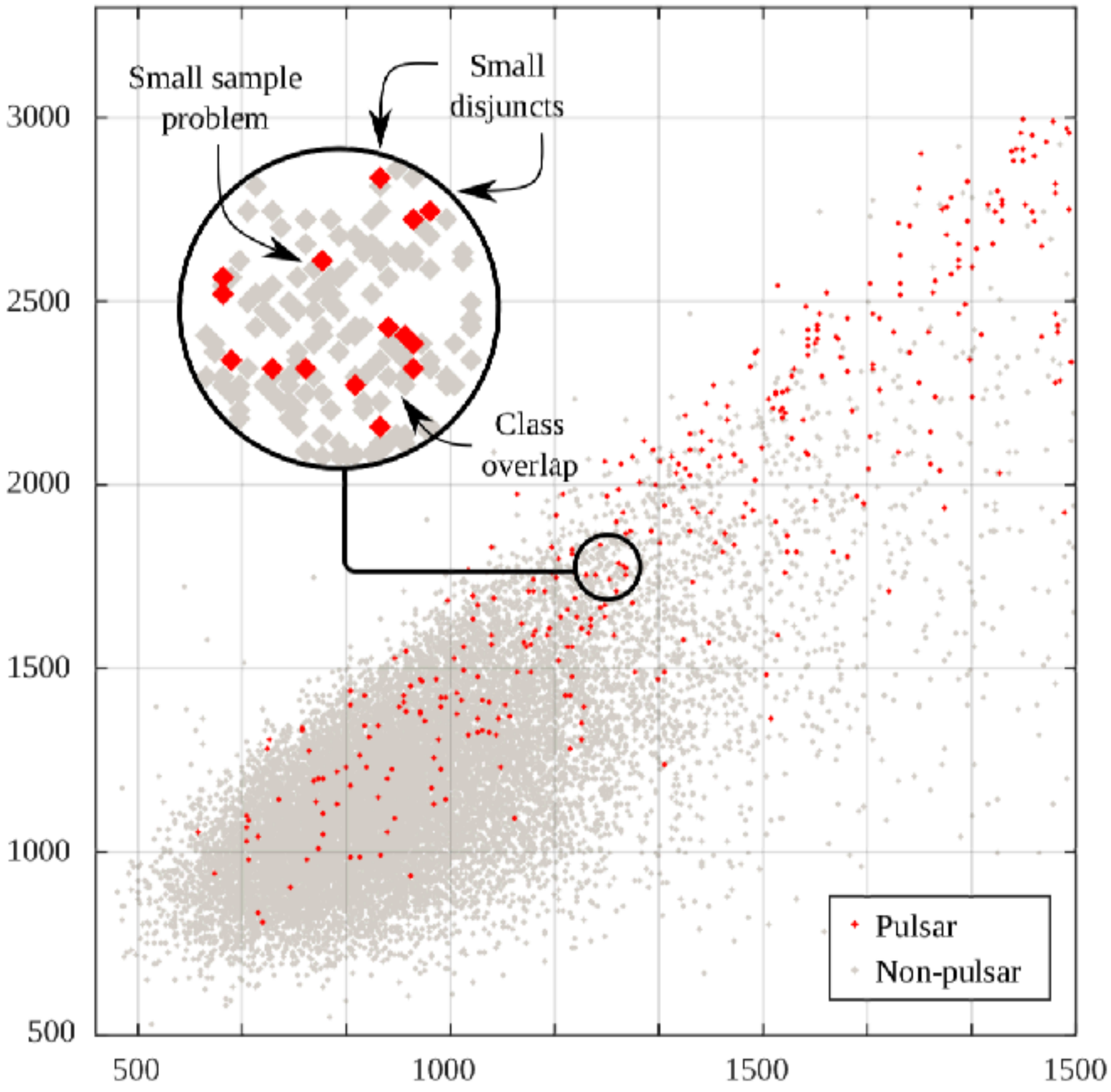
IMBALANCE IS NOT ALWAYS A PROBLEM



. . . BUT IT OFTEN IS.



. . . BUT IT OFTEN IS.



POSSIBLE APPROACHES

- Do nothing.
- Force balance in the dataset.
- Modify algorithm.
- Write a new algorithm.
- Treat it as an anomaly detection problem.
- Get more data.

POSSIBLE APPROACHES

- Oversample minority class
- Undersample majority class
- Introduce synthetic examples

- Do nothing.
- Force balance in the dataset.
- Modify algorithm.
- Write a new algorithm.
- Treat it as an anomaly detection problem.
- Get more data.

POSSIBLE APPROACHES

- Weight input data
- Modify the loss function
- Calibrate probability estimates

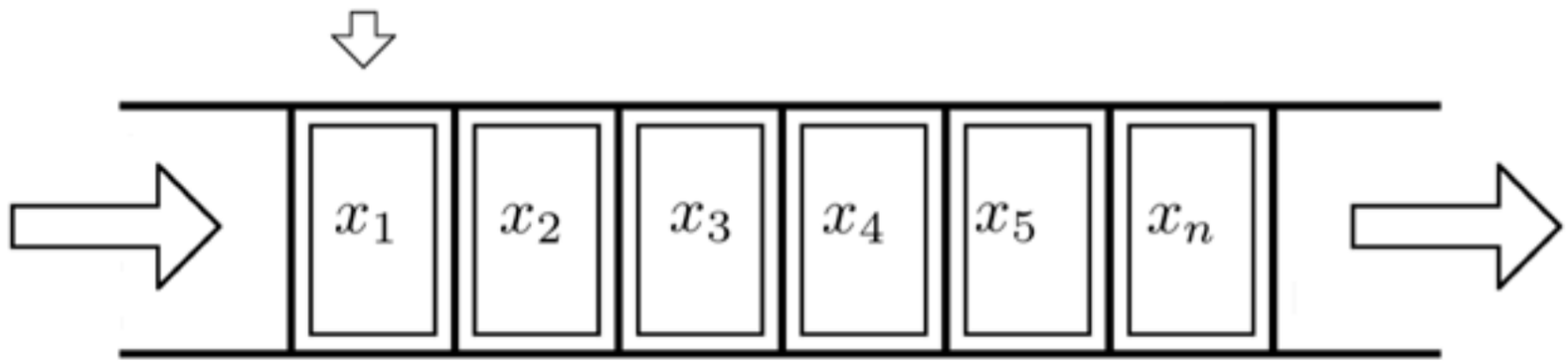
- Do nothing.
- Force balance in the dataset.
- Modify algorithm.
- Write a new algorithm.
- Treat it as an anomaly detection problem.
- Get more data.

POSSIBLE APPROACHES

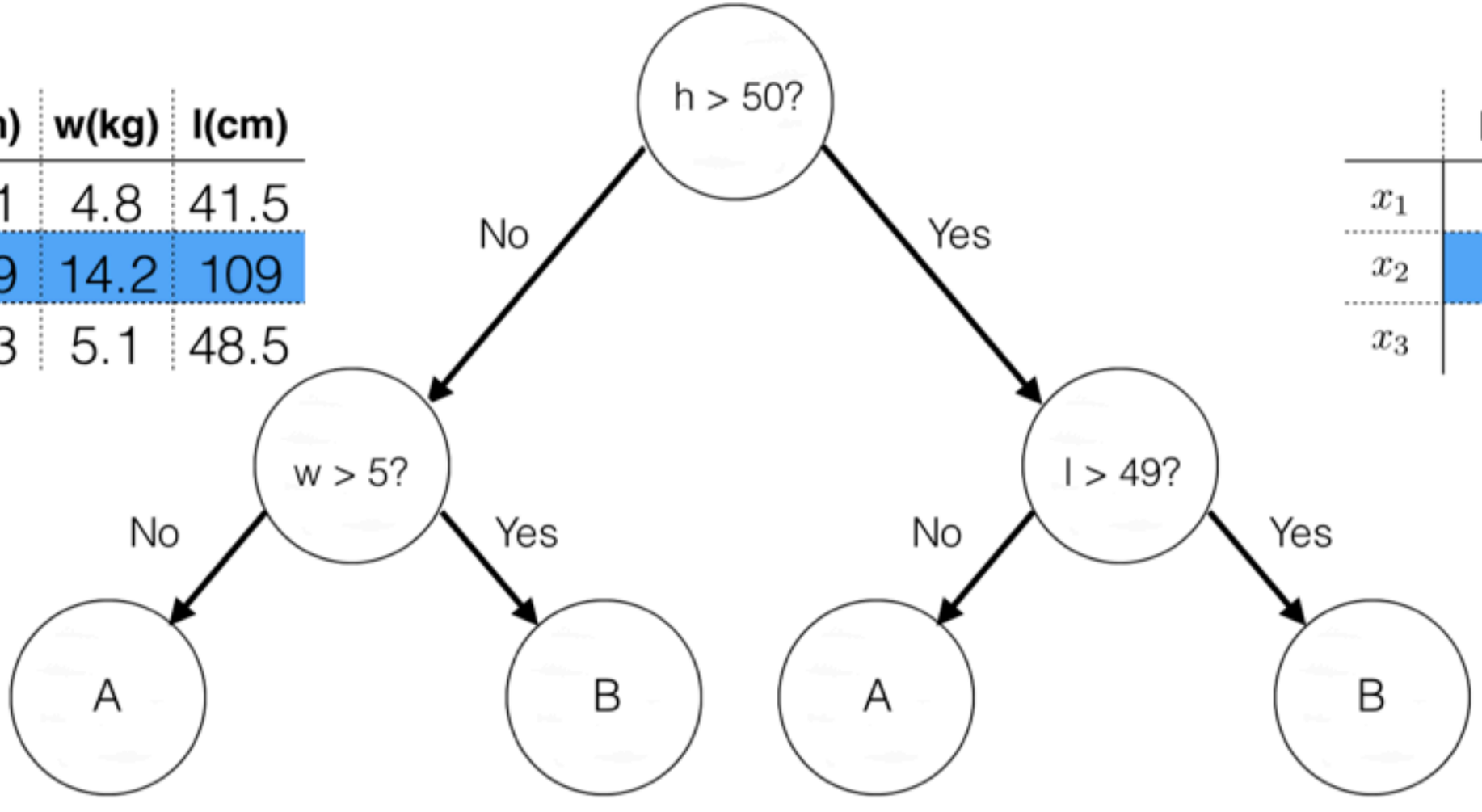
- Do nothing.
- Force balance in the dataset.
- Modify algorithm.
- Write a new algorithm.
- Treat it as an anomaly detection problem.
- Get more data.

- Not always possible
- Probably part of the reason you have an imbalance in the first place...

DECISION TREES

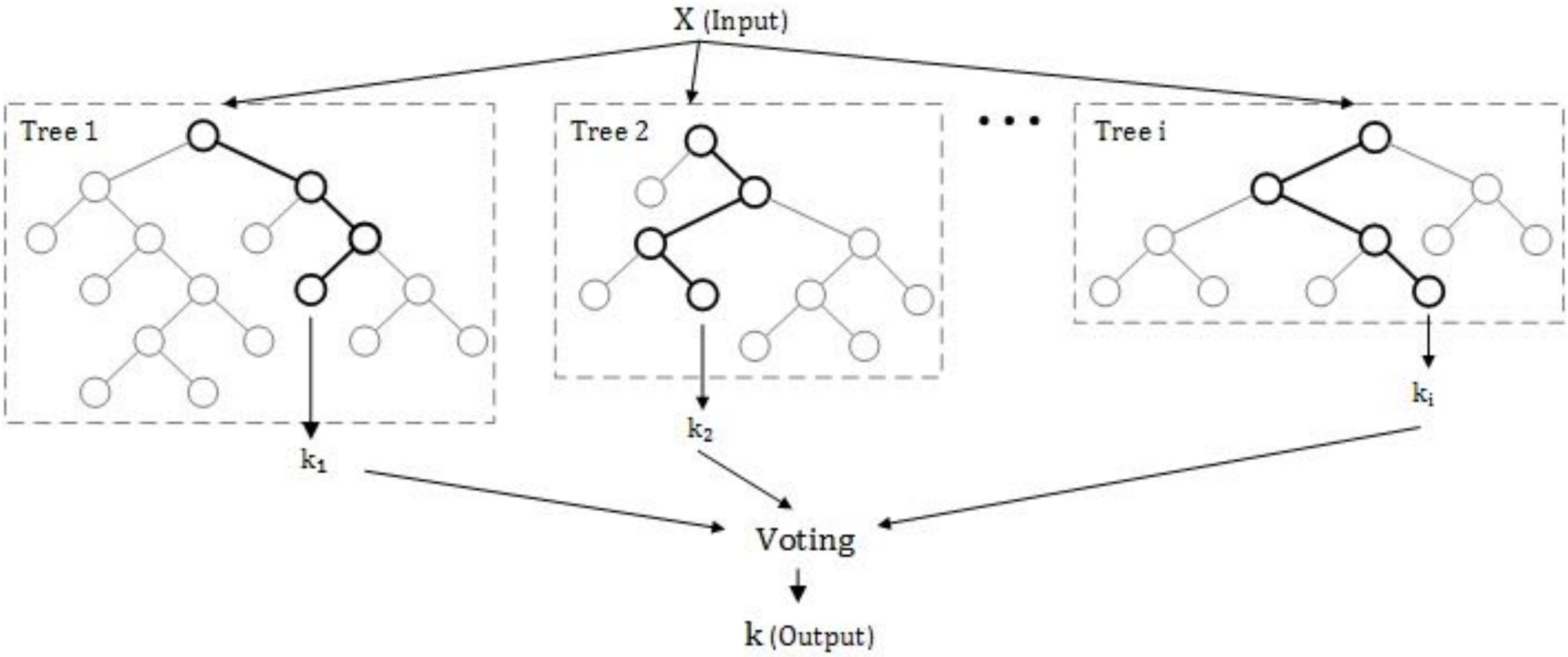


	h(cm)	w(kg)	l(cm)
x_1	32.1	4.8	41.5
x_2	67.9	14.2	109
x_3	35.3	5.1	48.5



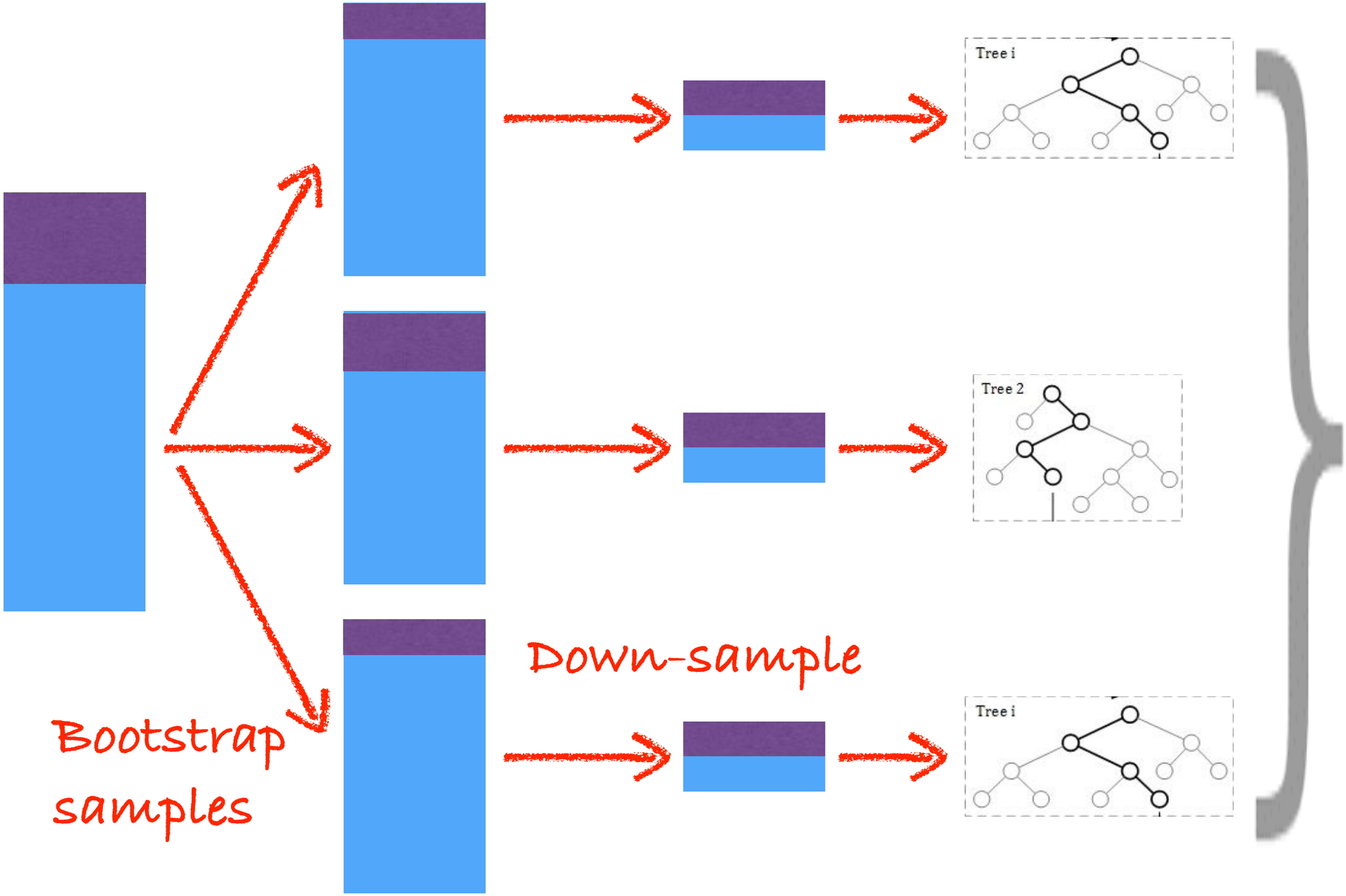
	Predictions
x_1	A
x_2	B
x_3	A

RANDOM FOREST CLASSIFIERS



Each tree uses **a subset of the data** and **a subset of the features**
(this is known as *bagging*)

BAGGED UNDERSAMPLING

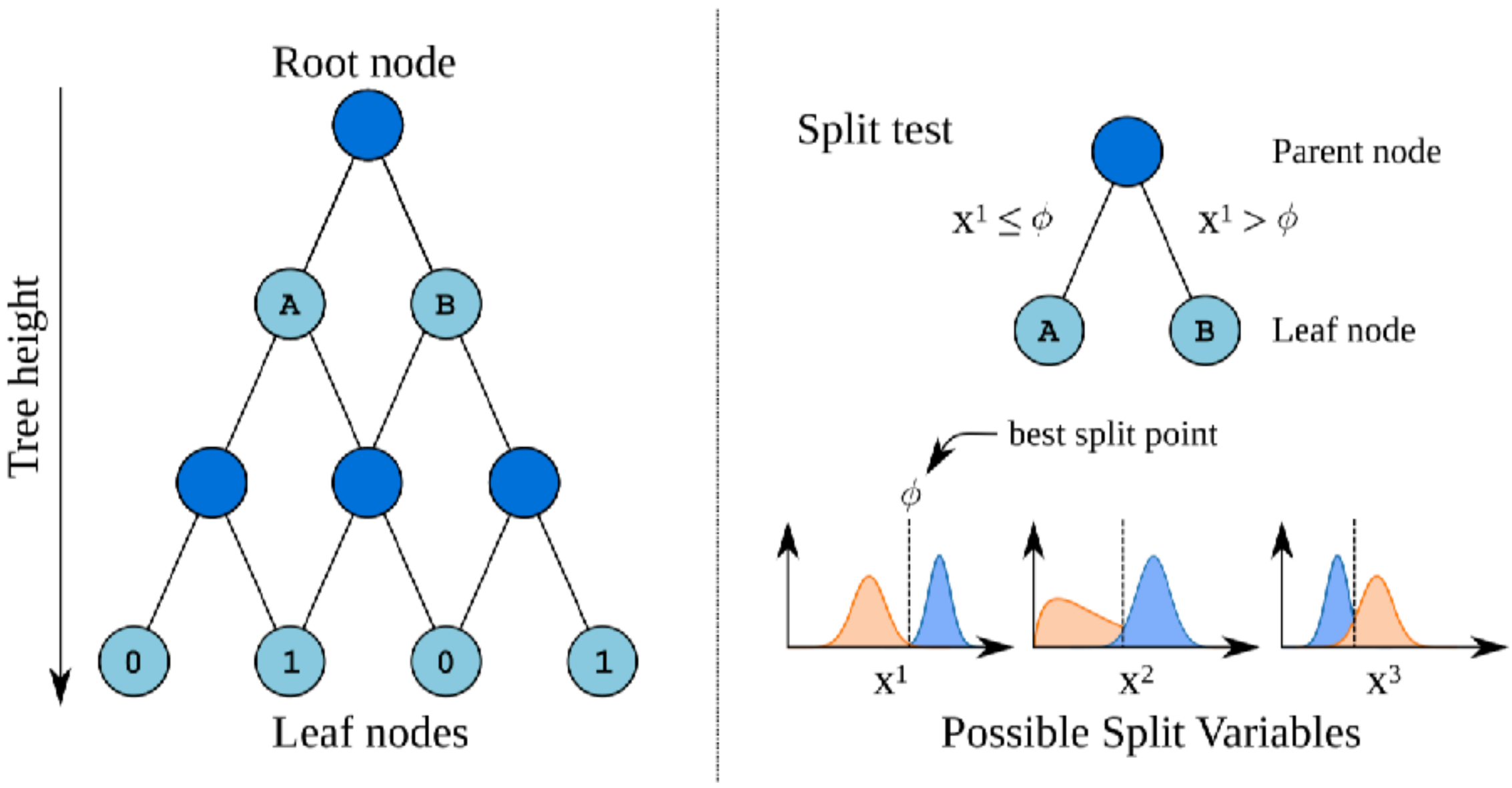


SKA PULSAR CLASSIFICATION

1,000 candidates ***per second***

> 80,000,000 candidates ***per day***

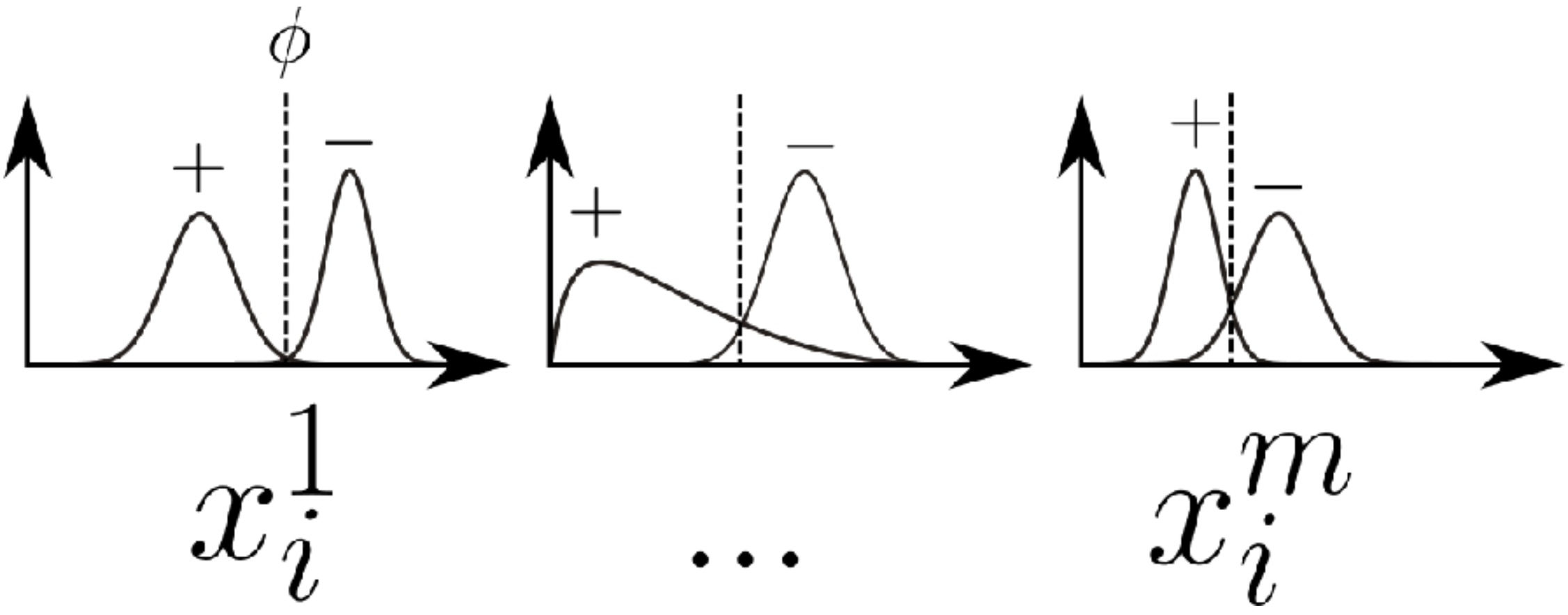
SKA PULSAR CLASSIFICATION: GH-VFDT STREAM CLASSIFIER



Gaussian Hellinger Very Fast Decision Tree

https://arxiv.org/pdf/1405.2278.pdf

SKA PULSAR CLASSIFICATION: GH-VFDT STREAM CLASSIFIER

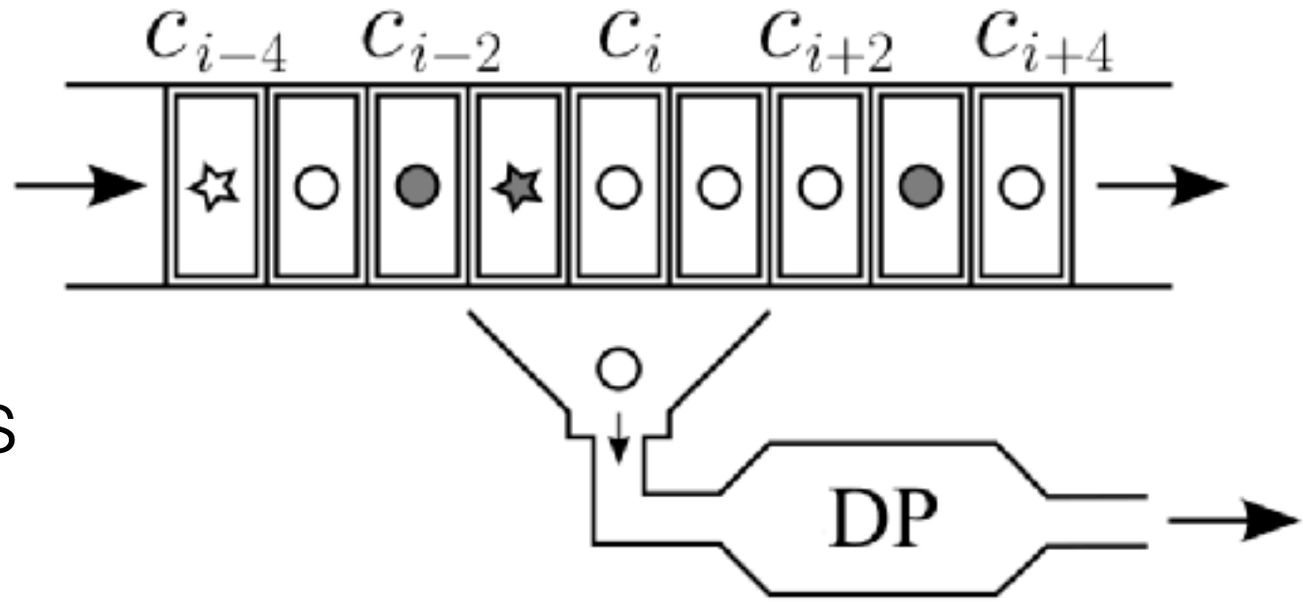


Hellinger Distance = separation between two distributions

https://arxiv.org/pdf/1810.06012.pdf

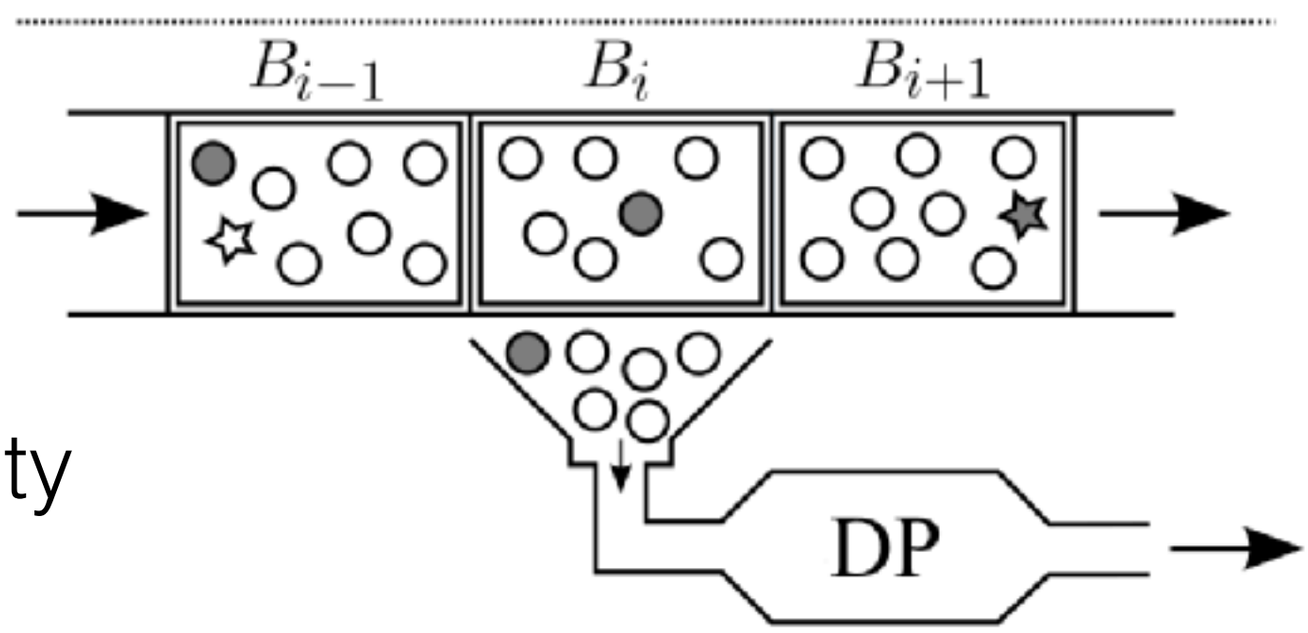
REAL-TIME PROCESSING

good for
isolated outliers



a) Incremental Model

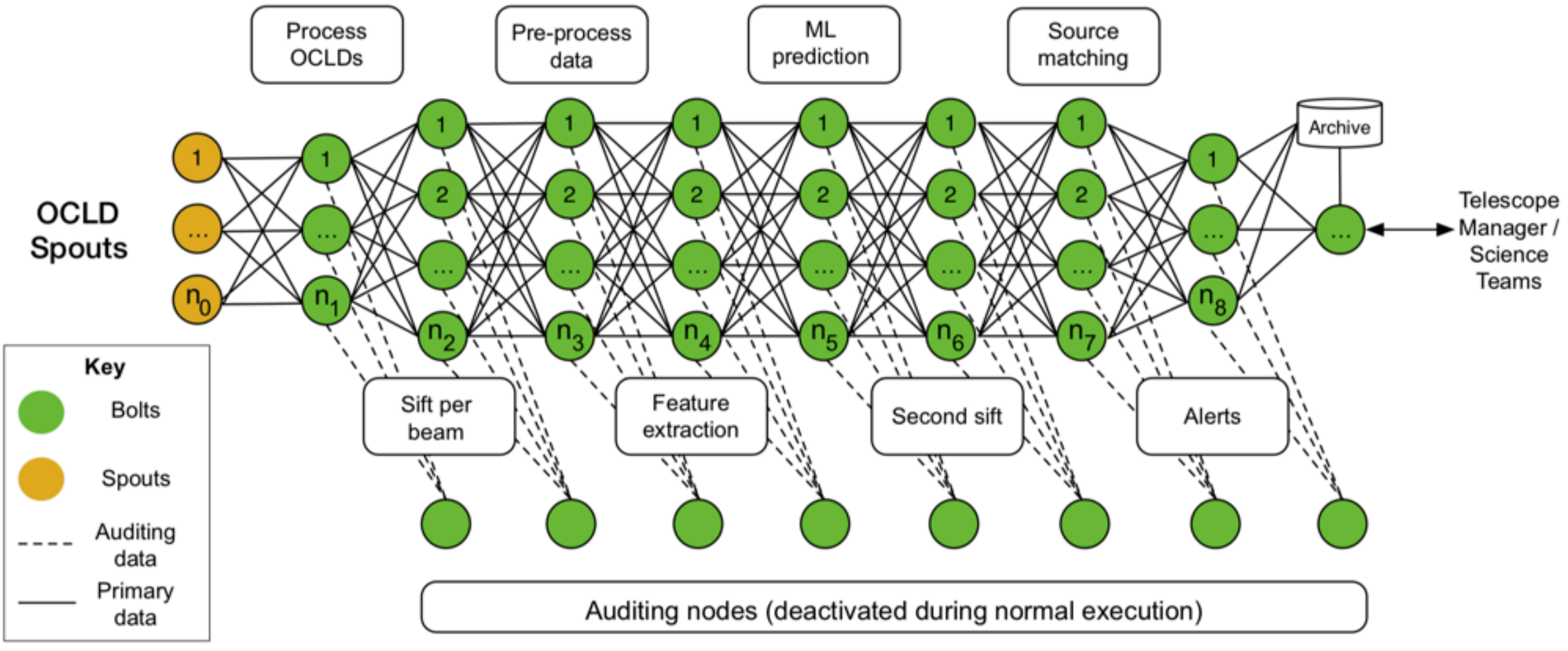
good for
localised similarity



b) Batch Model

https://arxiv.org/pdf/1810.06012.pdf

SKA REAL-TIME PROCESSING



https://github.com/scienceguyrob/GHVFDT

https://arxiv.org/pdf/1810.06012.pdf

TUTORIAL

https://github.com/as595/NITheP