The First Deep Learning Search for Technosignatures with the GBT

(Peter) Xiangyuan Ma | Advisor: Dr. Cherry Ng

LAIDLAW FOUNDATION

BERKELEY SETI RESEARCH CENTER

UNIVERSITY OF TORONTO

遠馬 印驤

What is SETI?

What is SETI?

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)

Q: What is intelligence?

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?
A: We look for *radio transmitters.*

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?
A: We look for *radio transmitters.*
Q: Why radio?

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?
A: We look for *radio transmitters.*
Q: Why radio?
A: It's one of the cheapest/fastest forms of com.

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?
A: We look for *radio transmitters.*
Q: Why radio?
A: It's one of the cheapest/fastest forms of com.
Q: What do we look for?

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?
A: We look for *radio transmitters.*
Q: Why radio?
A: It's one of the cheapest/fastest forms of com.
Q: What do we look for?
A: Radio transmissions beyond Earth are *Narrowband doppler drifting signals (technosignature)*

What is SETI?

A: Stands for the Search for *Extraterrestrial Intelligence* (SETI)
Q: What is intelligence?
A: We look for signs of *technology beyond Earth*
Q: What technology?
A: We look for *radio transmitters.*
Q: Why radio?
A: It's one of the cheapest/fastest forms of com.
Q: What do we look for?
A: Radio transmissions beyond Earth are *Narrowband doppler drifting signals (technosignature)*

If we find Narrowband doppler drifting signal beyond Earth → *Potential Intelligent life beyond Earth*

That’s simple!

That’s simple! Where the Aliens at?

Okay but hold on … there’s a problem

What’s the problem?

The Problem

*If we find Narrowband doppler drifting signal **beyond Earth*** → *Potential Intelligent life beyond Earth*

So What’s the Problem?

1. ***Narrowband doppler drifting signal***
1. ***beyond Earth***

So What's the Problem?

1. ***Narrowband doppler drifting signal***

1. ***beyond Earth***

We end up finding our own RFI! - Makes sense since we are intelligent … but we need to find things beyond Earth!

So What's the Problem?

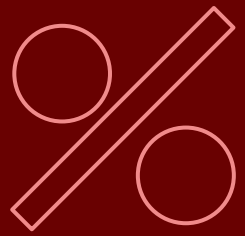
1. ***Narrowband doppler drifting signal***

1. ***beyond Earth***

FALSE POSITIVE problem

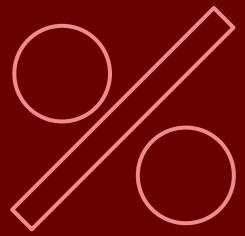
So What’s the Problem?

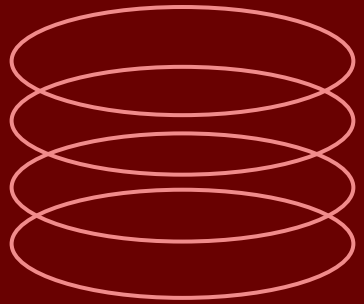
1. RFI False Pos.



So What’s the Problem?

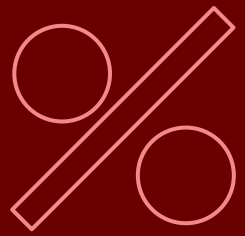
1. RFI False Pos.
2. Big data

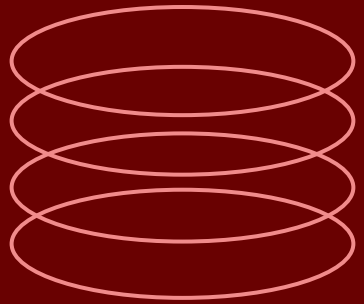


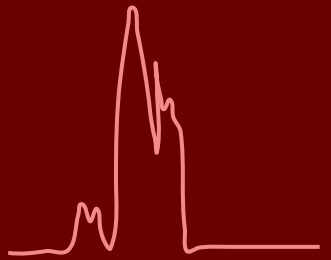


So What's the Problem?

1. RFI False Pos.
2. Big data
3. Signal Bias



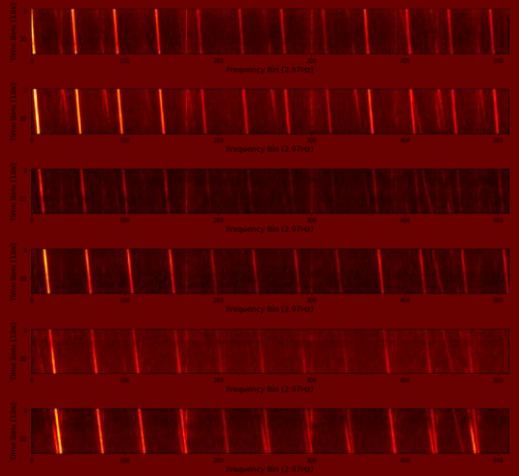




So What's the Problem?

1. RFI False Pos.

We keep finding signals from our selves because of RFI

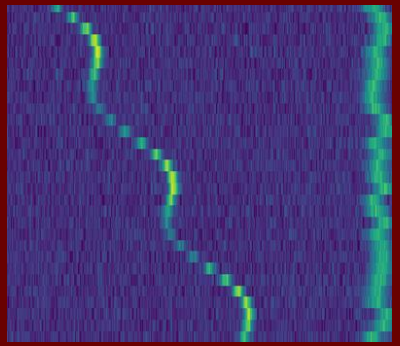


2. Big data

We have PB worth of data to search through

3. Signal Bias

There are actually an infinite number of signal morphologies to look for.



So What's the Problem?

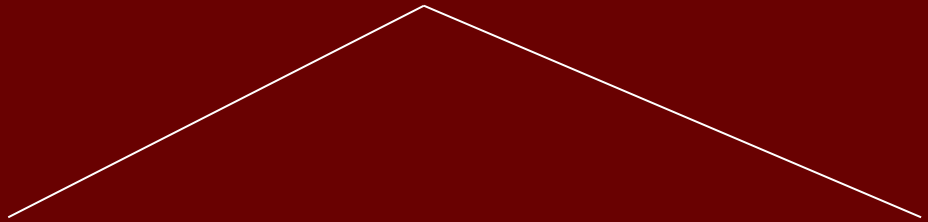
1. RFI False Pos.
2. Big data
3. Signal Bias

Deep Learning???

So What's the Problem?

1. RFI False Pos.
2. Big data
3. Signal Bias

Deep Learning???



Automatic RFI rejection

Reduce Bias in detection

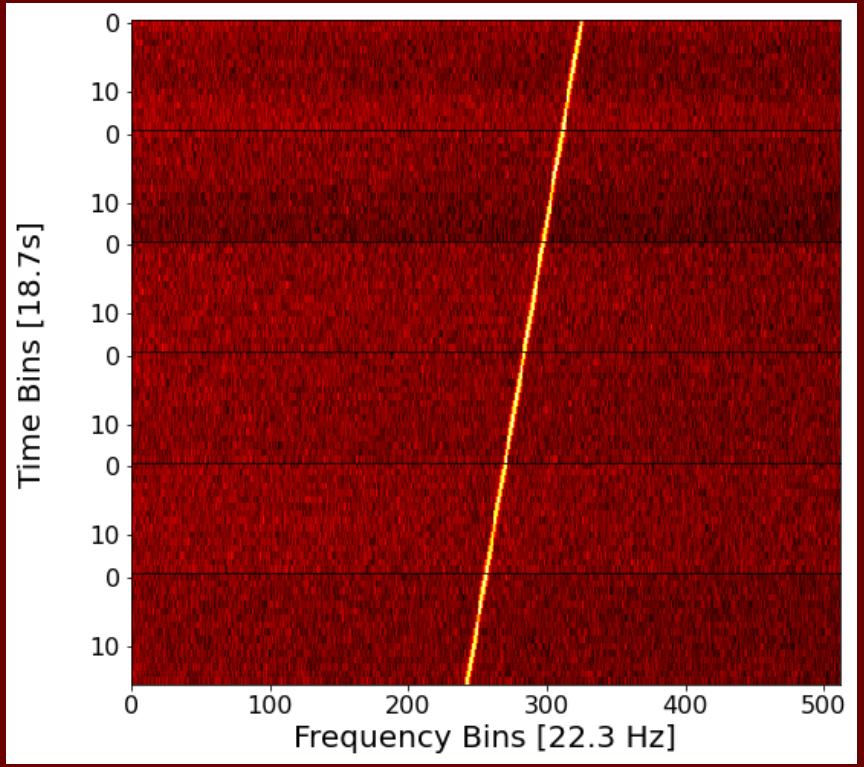
How would a human do it?

The Data

What does the data we work with look like?

The Data

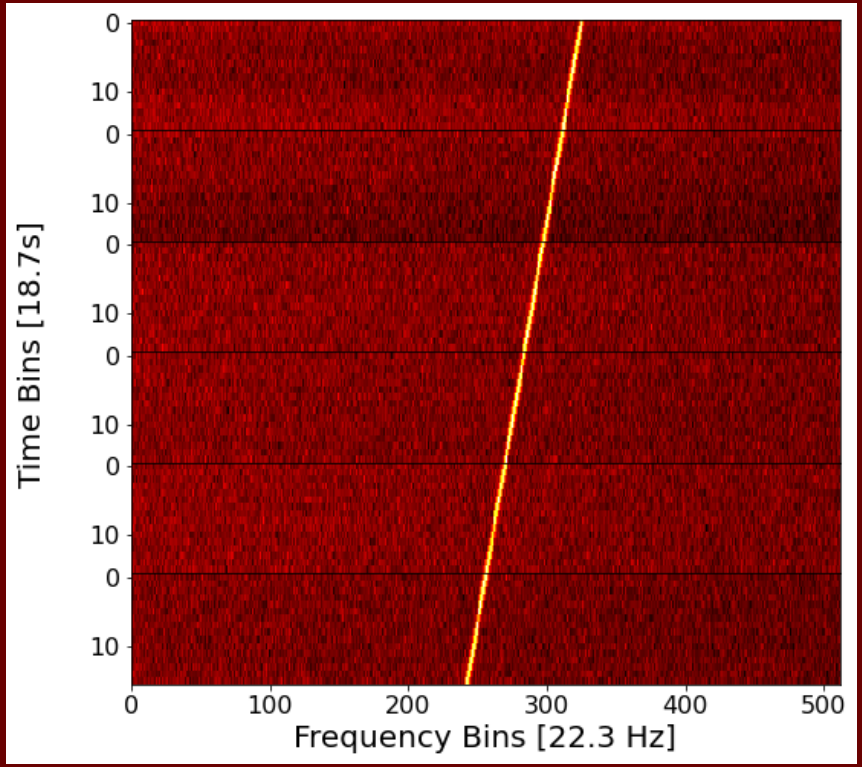
We work with **Dynamic Spectra** (Waterfall plots)



The Data

We work with **Dynamic Spectra** (Waterfall plots)

Data is **[time, polarization, frequency]**



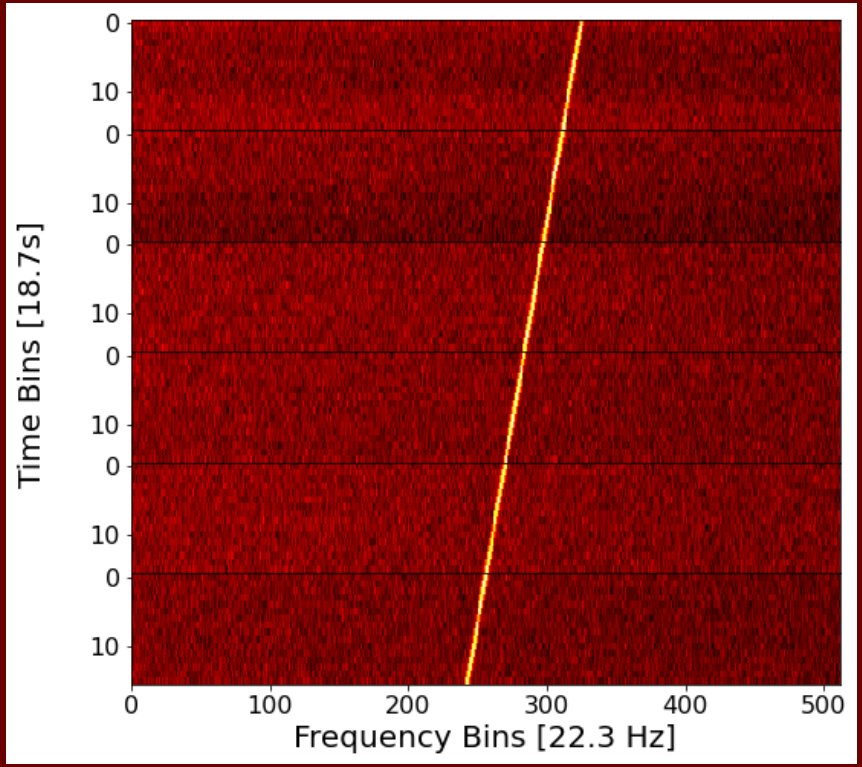
The Data

We work with **Dynamic Spectra** (Waterfall plots)

Data is **[time, polarization, frequency]**

Time resolution: **18s** => 4.8min obs = 16 total time bins

Frequency resolution: **2.84Hz** ~ 3.18 x 10^{8} freq bins
903MHz



The Data

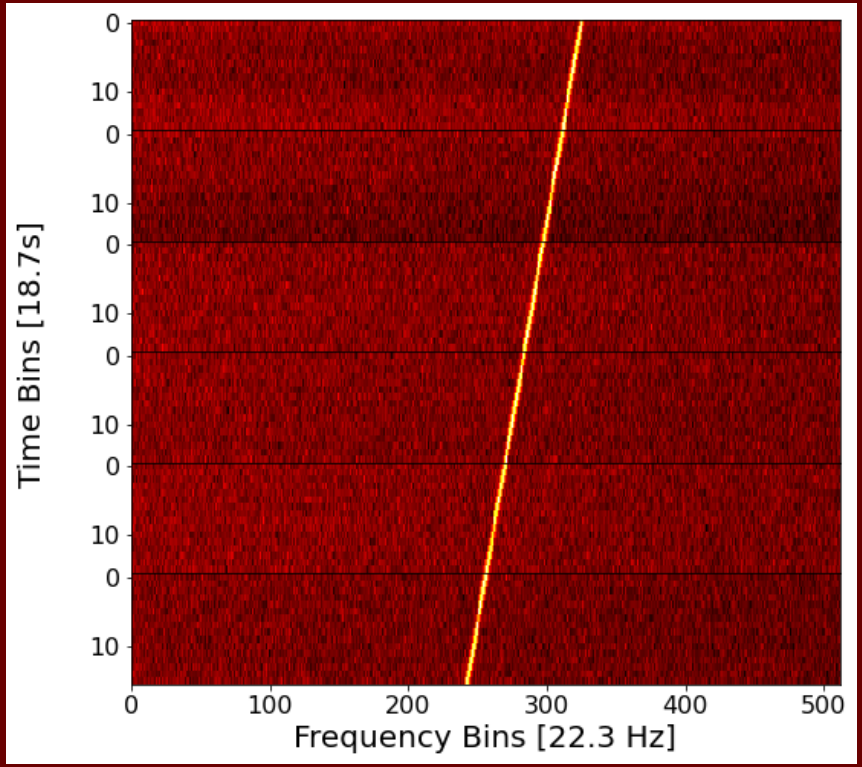
We work with **Dynamic Spectra** (Waterfall plots)

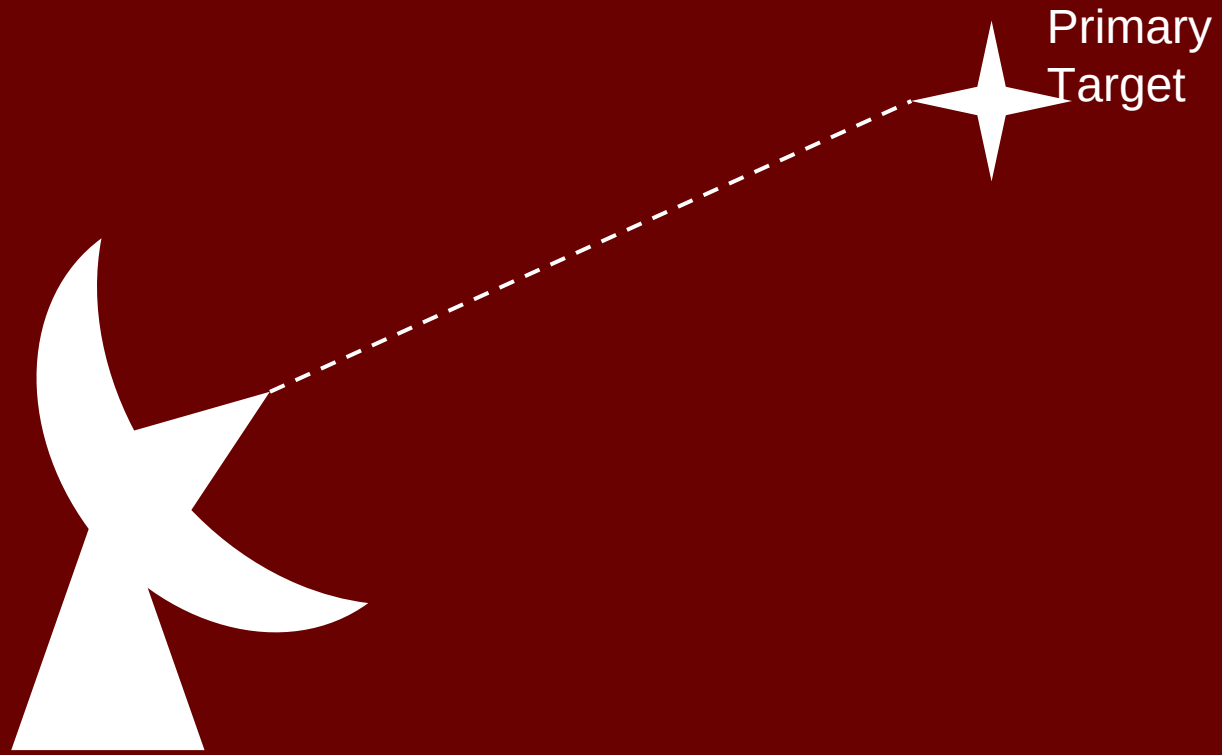
Data is **[time, polarization, frequency]**

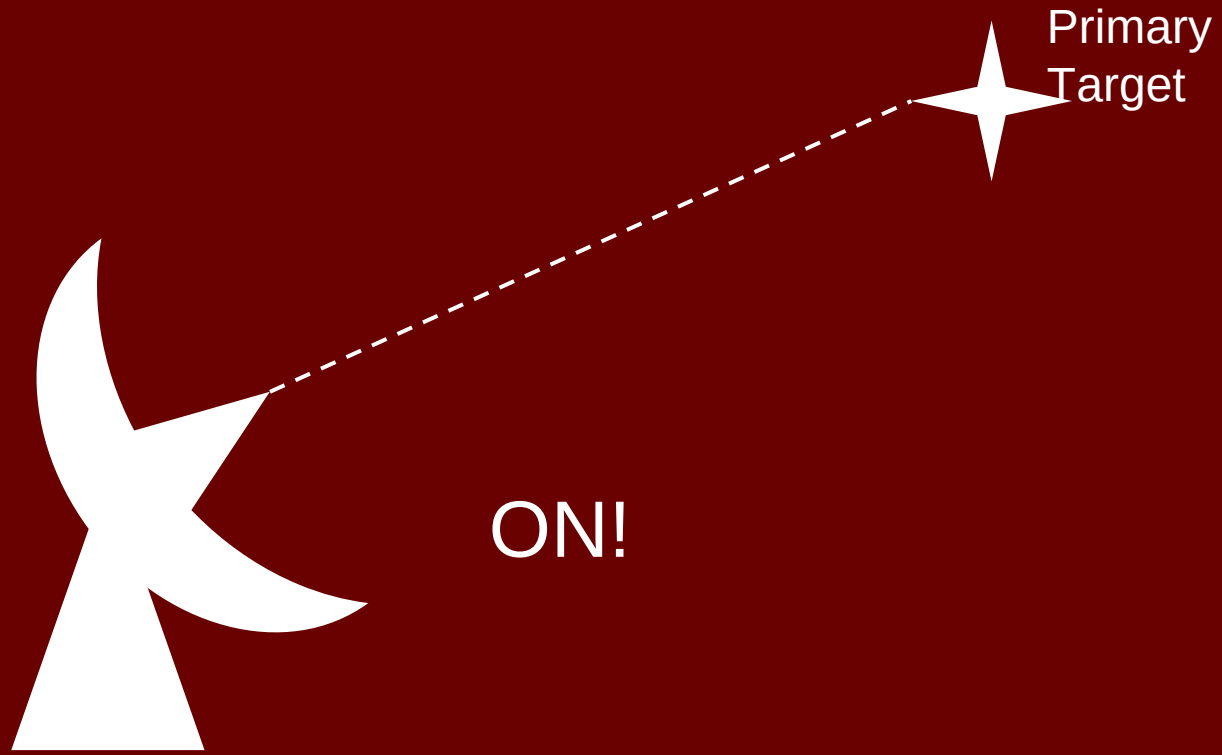
Time resolution: **18s** => 4.8min obs = 16 total time bins

Frequency resolution: **2.84Hz** ~ 3.18 x 10^{8} freq bins 903MHz

Cadence: 6 adjacent observations




Primary
Target


Primary
Target
ON!

Primary
Target

ON!

Generated Data Product

Primary Target

Secondary Target

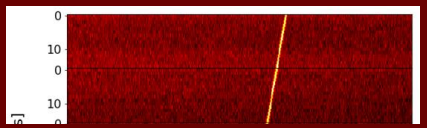
Generated Data Product

ON!

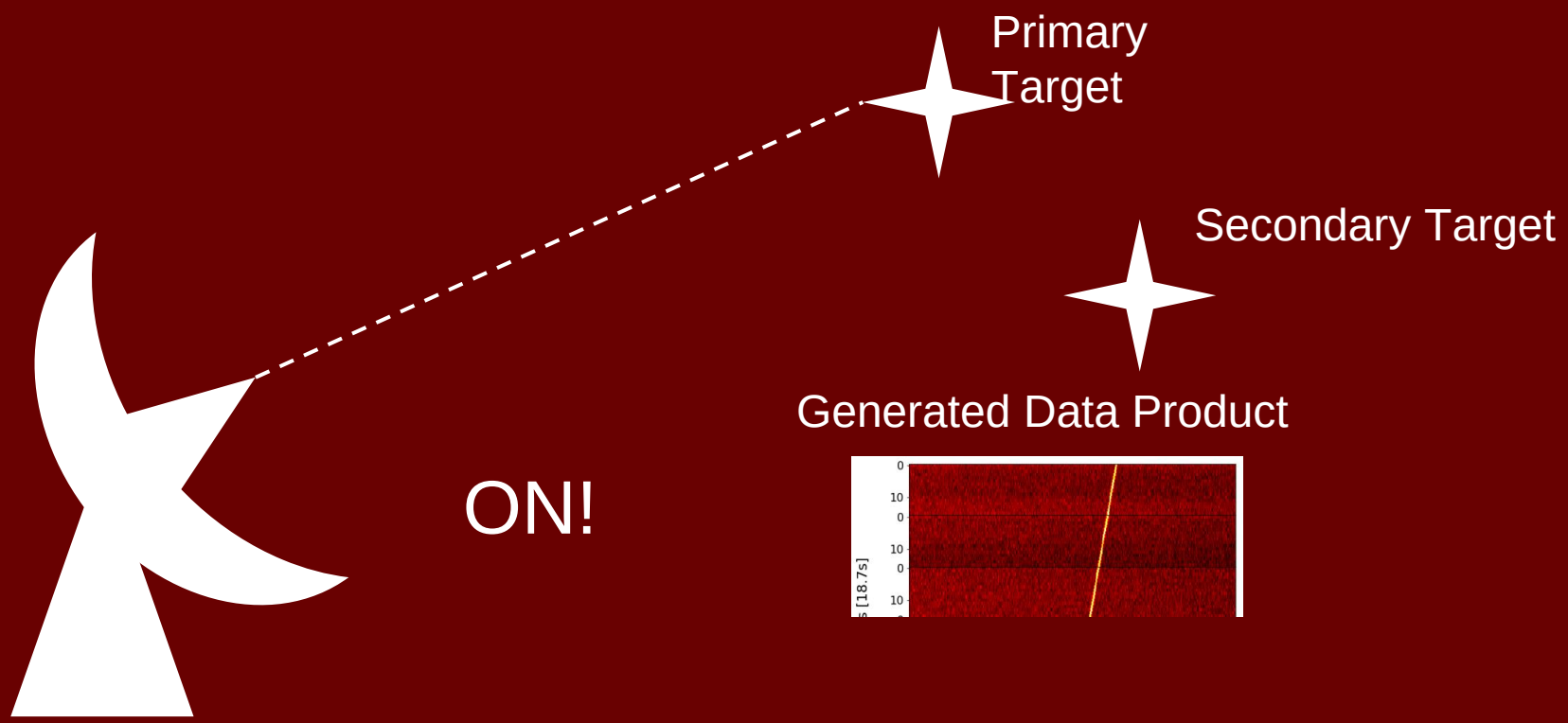
Primary Target

Secondary Target

Generated Data Product



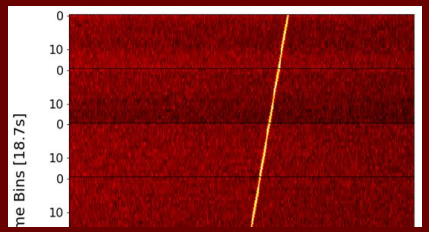
OFF!


Primary
Target
Secondary Target
Generated Data Product
ON!
0
10
0
10
0
10
[18.7s]

Primary Target

Secondary Target

Generated Data Product



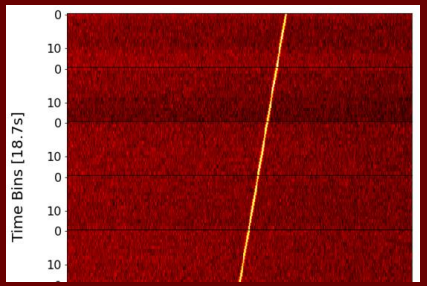
OFF!

Primary Target

Secondary Target

ON!

Generated Data Product

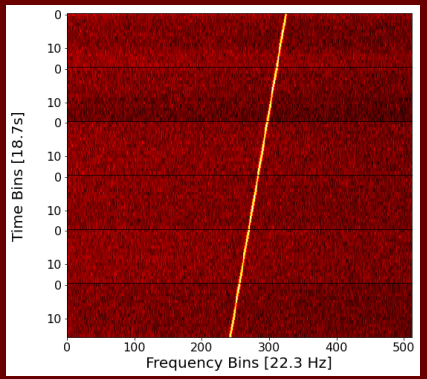


Primary Target

Secondary Target

Generated Data Product

OFF!



What Do We Look For?

ON

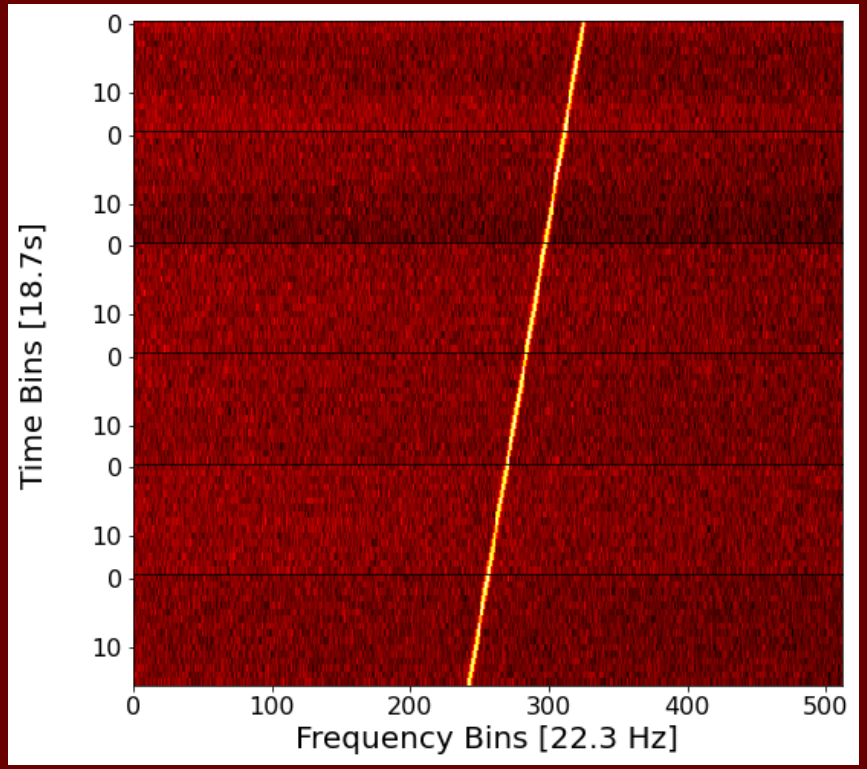
OFF

ON

OFF

ON

OFF



This is RFI

ON

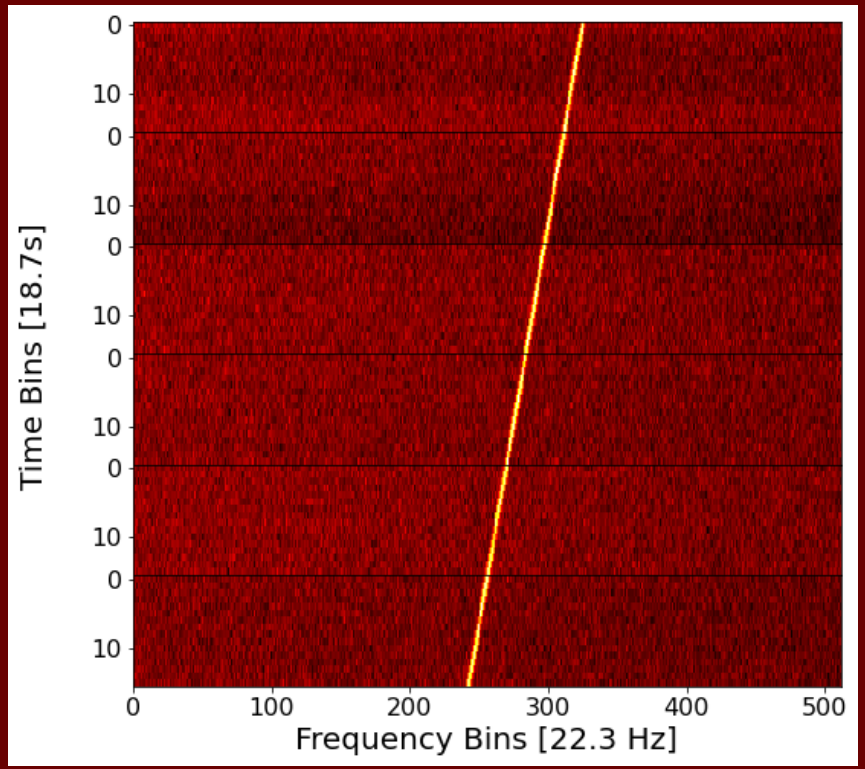
OFF

ON

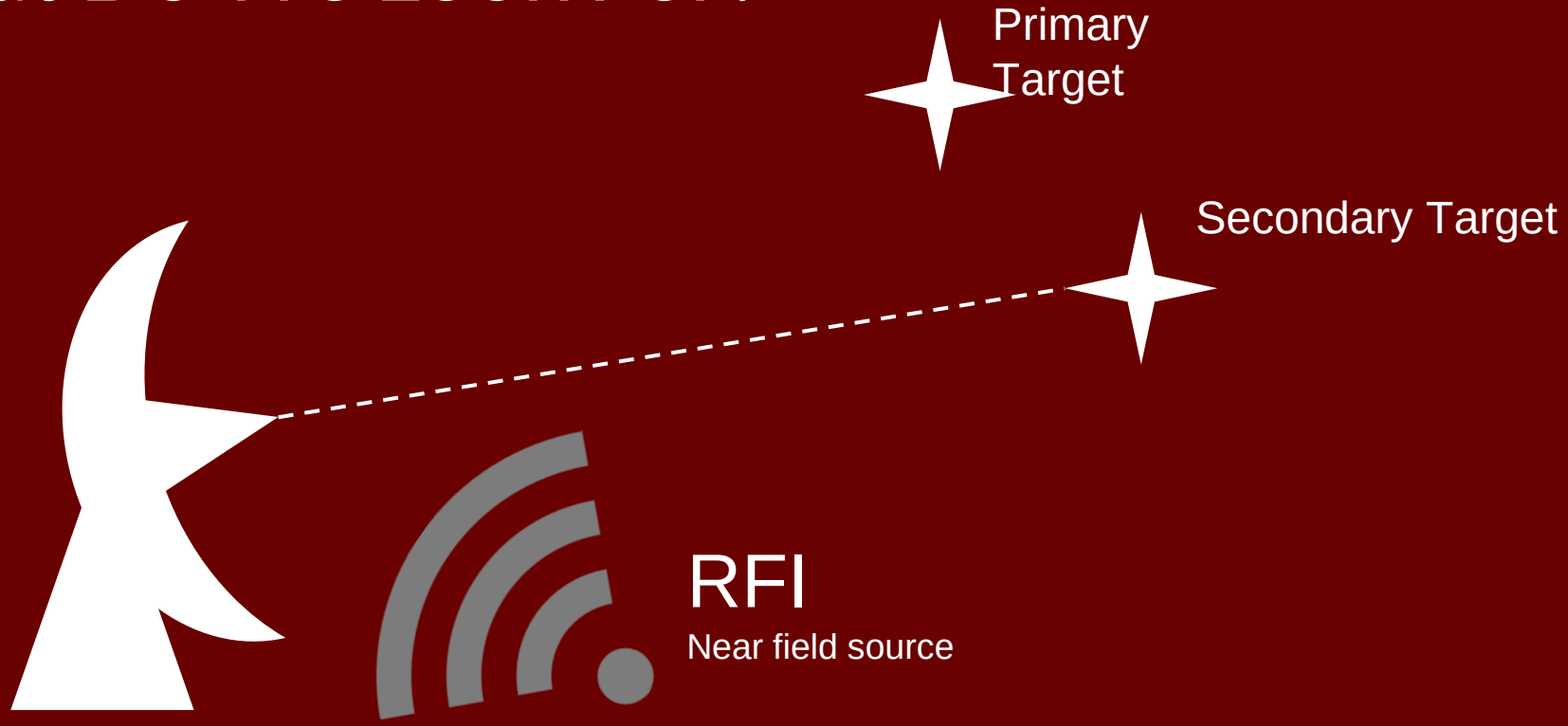
OFF

ON

OFF



What Do We Look For?



More RFI Examples

ON

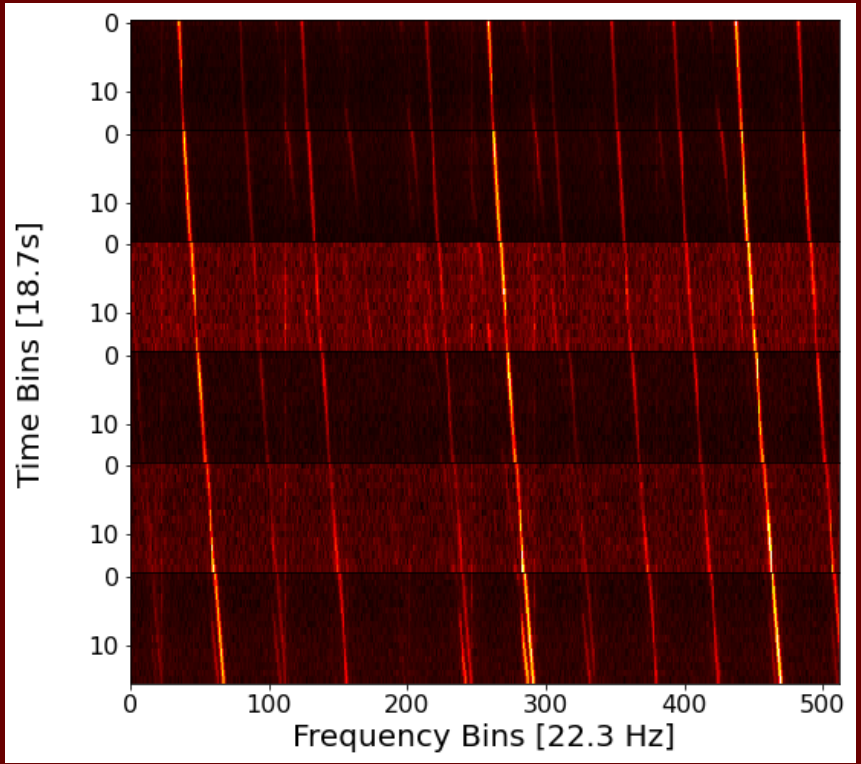
OFF

ON

OFF

ON

OFF



Ideal SETI Candidate

ON

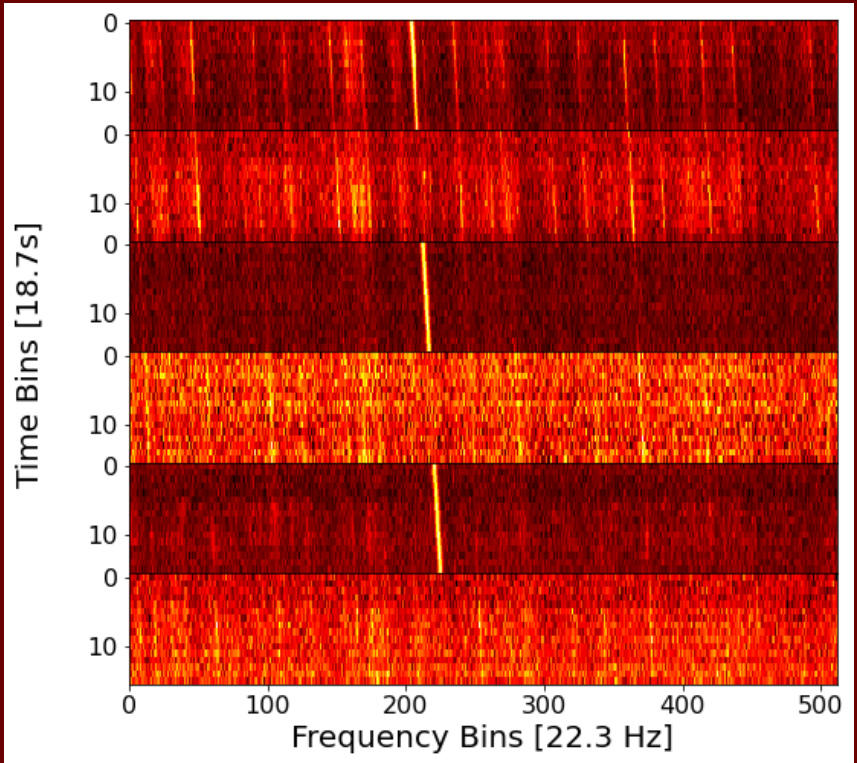
OFF

ON

OFF

ON

OFF



Ideal SETI Candidate

ON

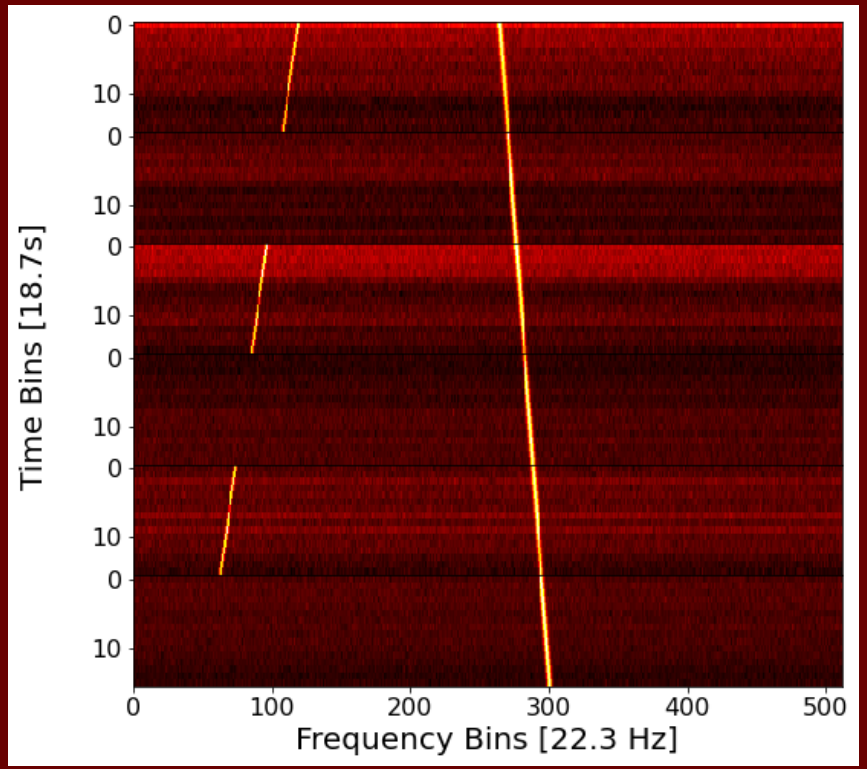
OFF

ON

OFF

ON

OFF

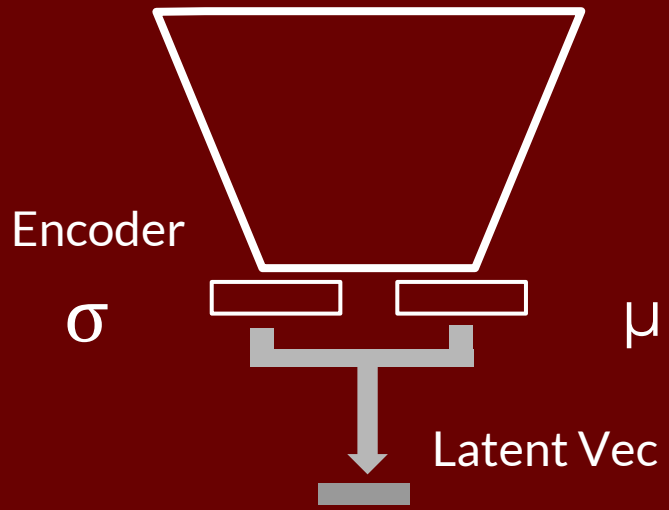


How do we ***teach*** computers to do this?

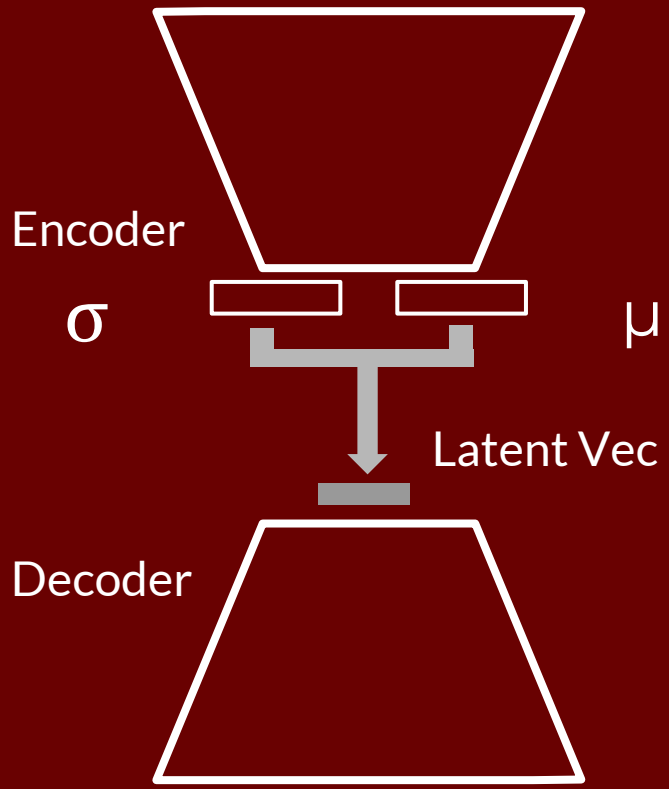
β-Variational Autoencoder + Random Forest Classifier

β-Variational Autoencoder

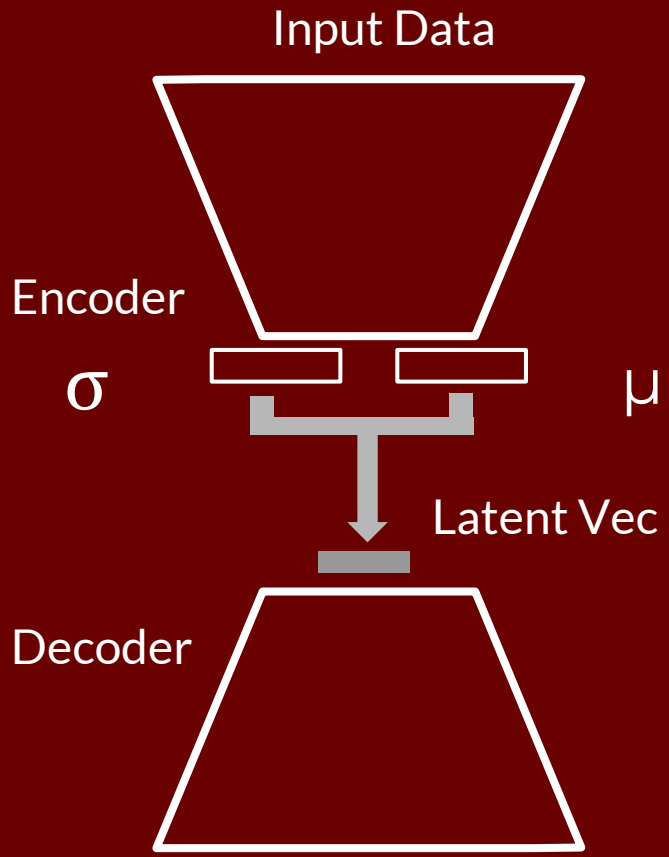
β-Variational Autoencoder



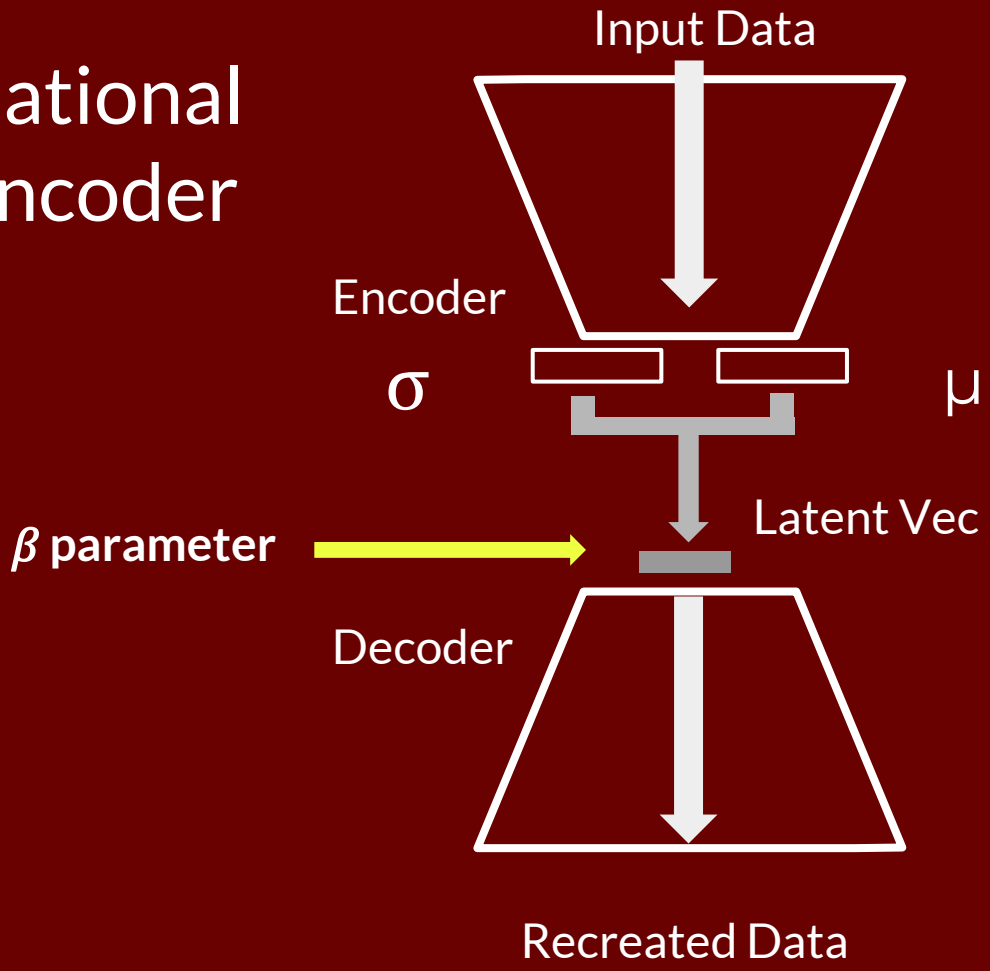
β-Variational Autoencoder



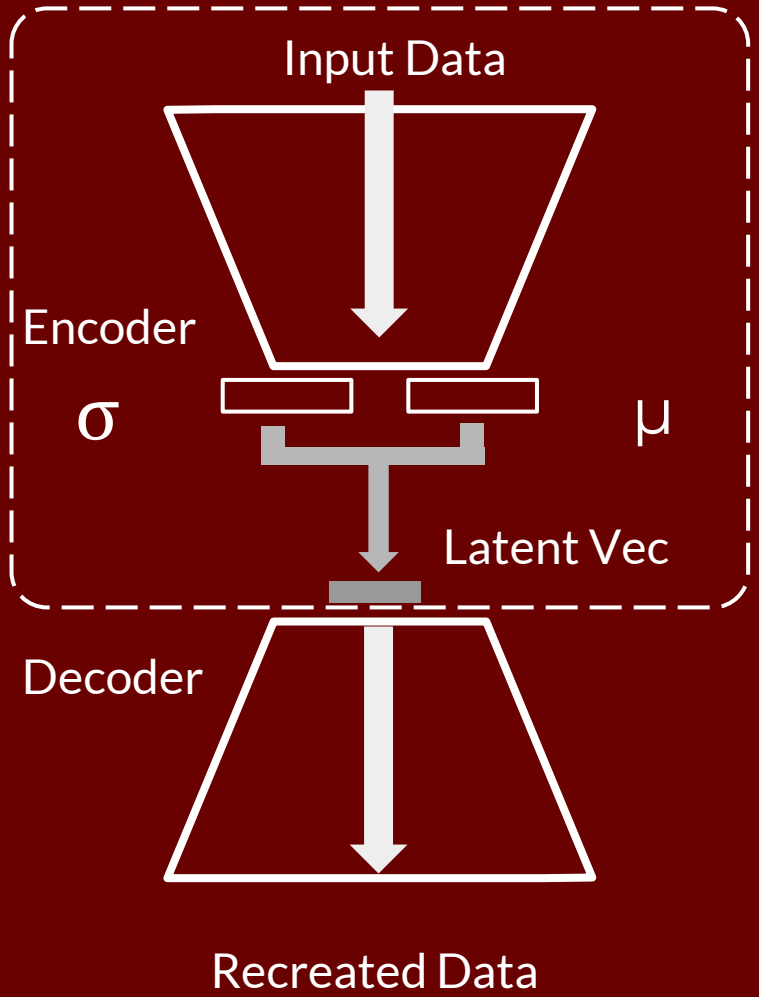
β-Variational Autoencoder

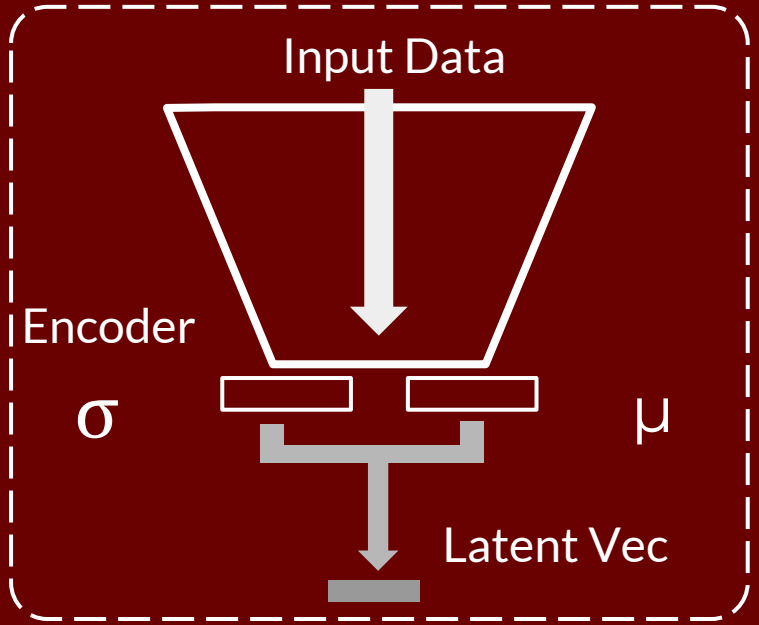


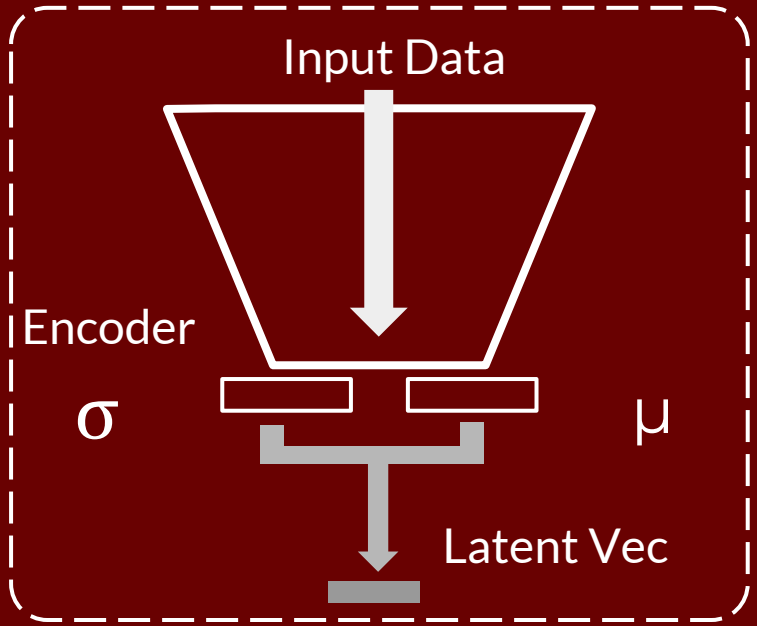
β-Variational Autoencoder



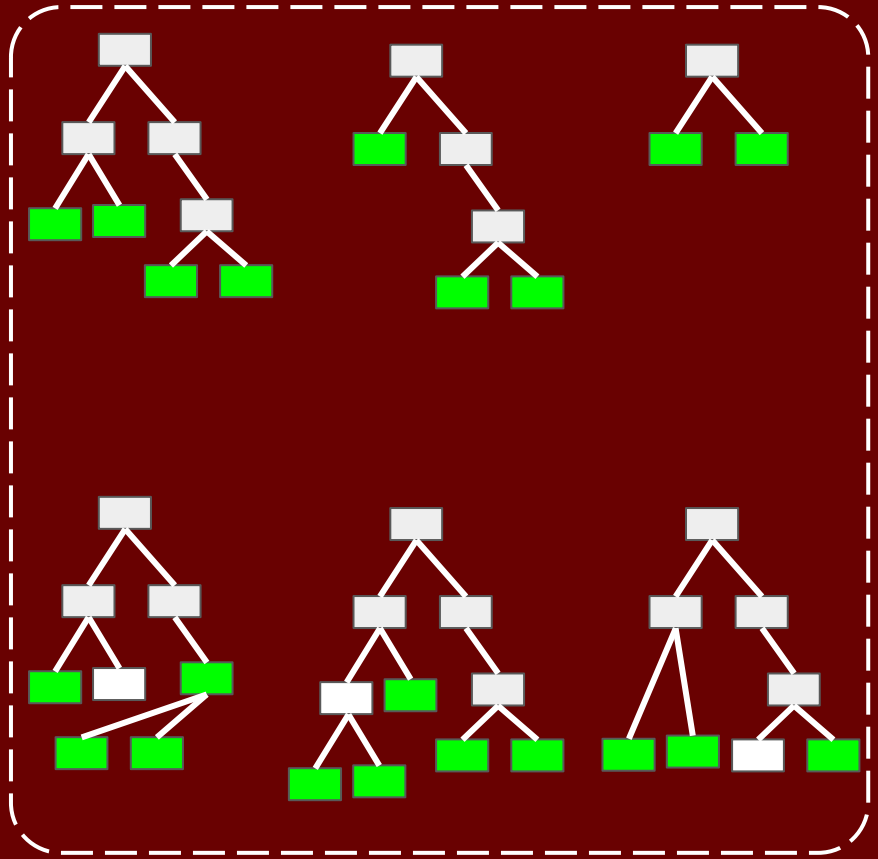
β-Variational Autoencoder

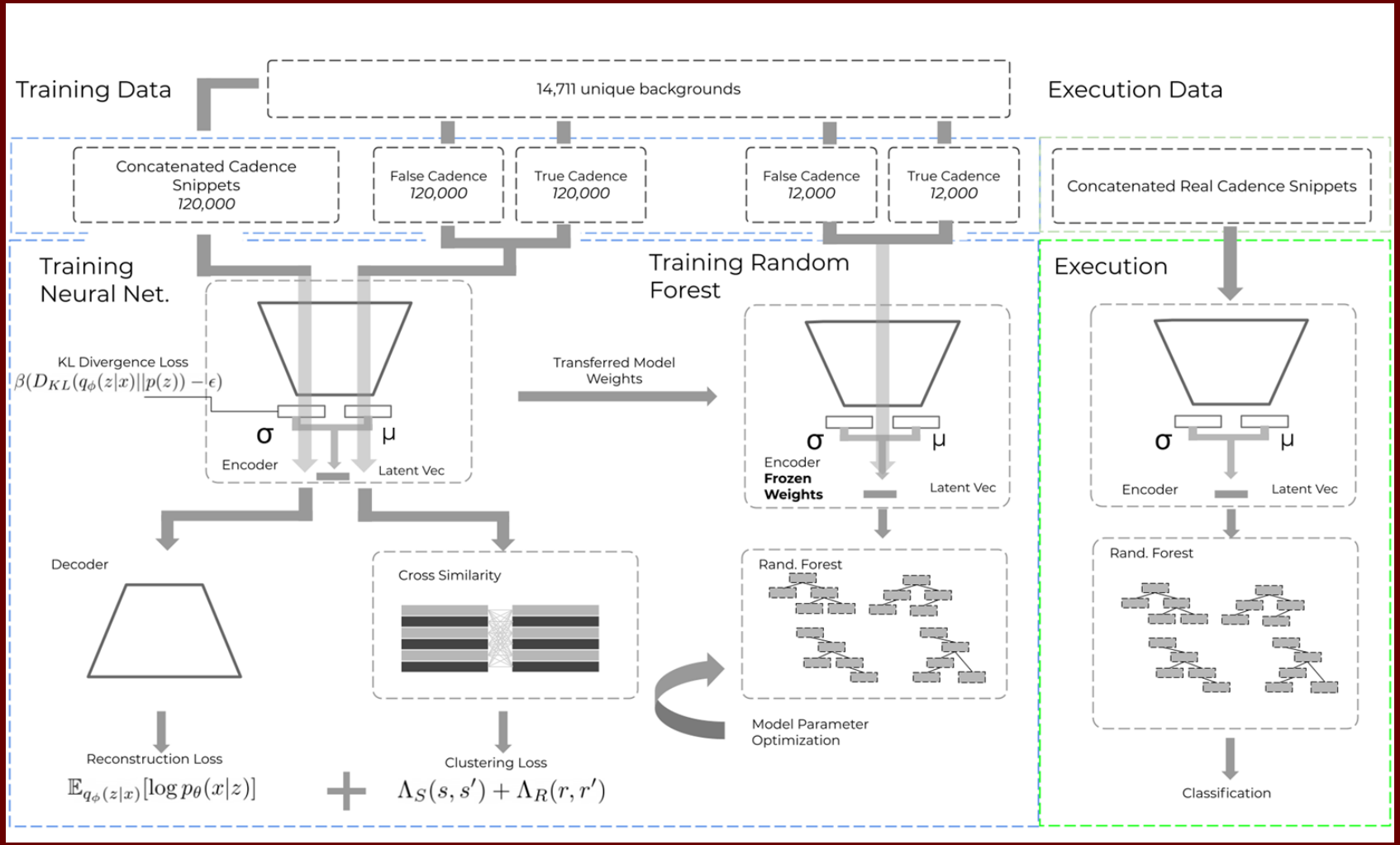



Input Data
Encoder
σ
μ
Latent Vec



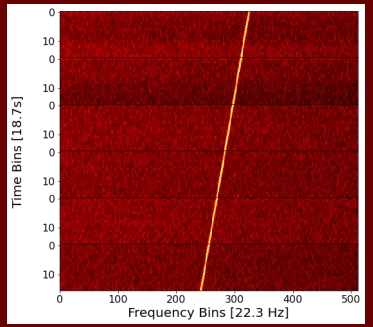
Random Forest

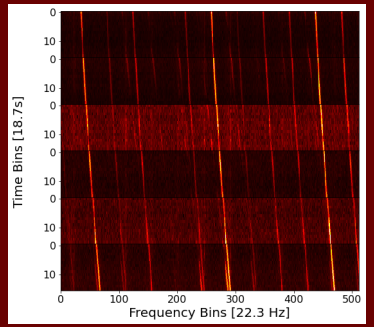


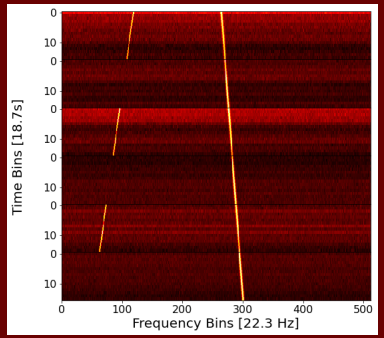

Training Data
14,711 unique backgrounds
Execution Data
Concatenated Cadence Snippets
120,000
False Cadence
120,000
True Cadence
120,000
False Cadence
12,000
True Cadence
12,000
Concatenated Real Cadence Snippets
Training Neural Net.
Training Random Forest
Execution
KL Divergence Loss
$\beta(D_{KL}(q_\phi(z|x)||p(z)) - \epsilon)$
σ
μ
Encoder
Latent Vec
Transferred Model Weights
Encoder
Frozen Weights
Latent Vec
Decoder
Cross Similarity
Rand. Forest
Rand. Forest
Reconstruction Loss
$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$
Clustering Loss
$\Lambda_S(s, s') + \Lambda_R(r, r')$
Model Parameter Optimization
Classification

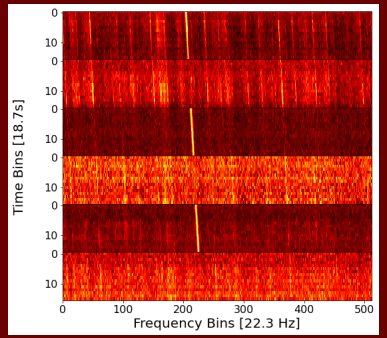
What data was this trained on??

The Training Data



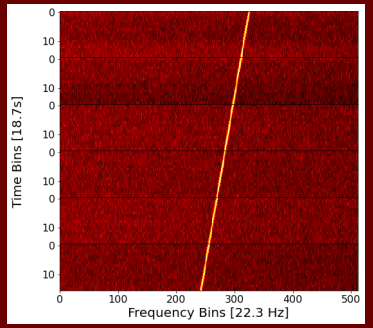


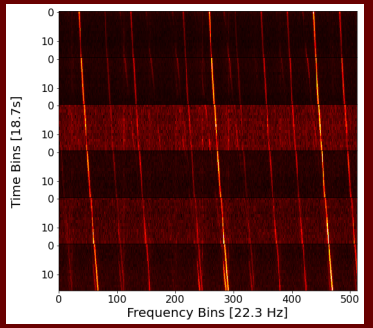


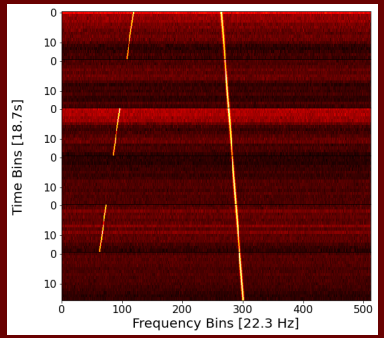


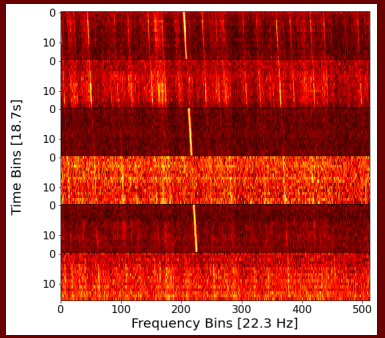
The Training Data

We simulated signals on real **Dynamic Spectra**





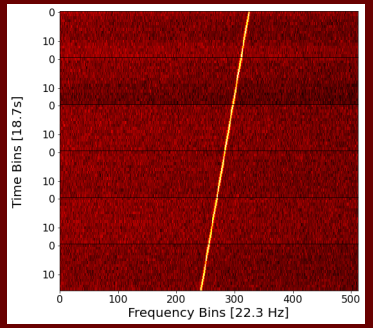


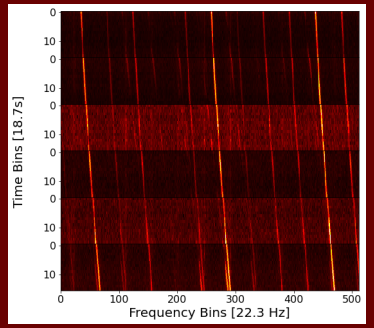


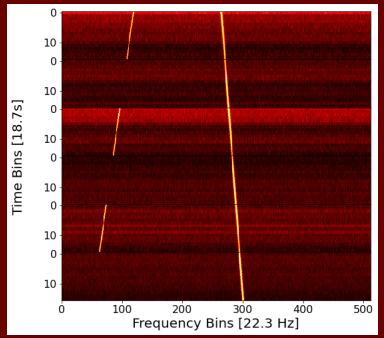
The Training Data

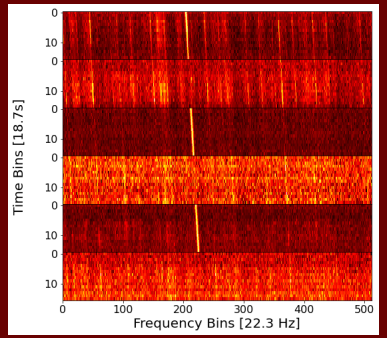
We simulated signals on real **Dynamic Spectra**

We used SETIGEN, which drew signals into observations







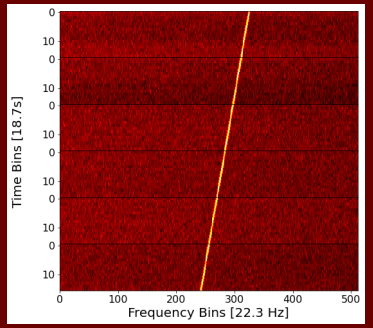


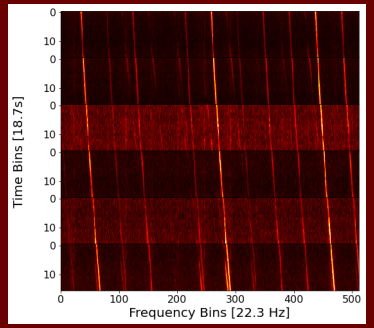
The Training Data

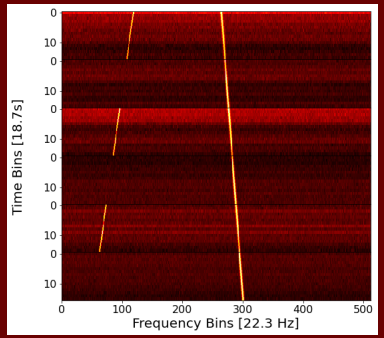
We simulated signals on real **Dynamic Spectra**

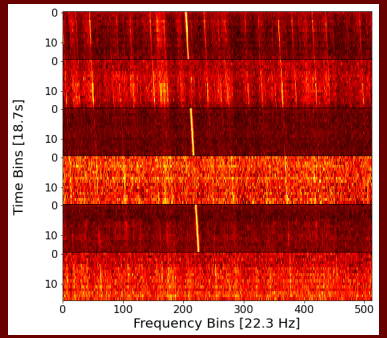
We used SETIGEN, which drew signals into observations

We randomly sampled parameters of SNR, drift, width









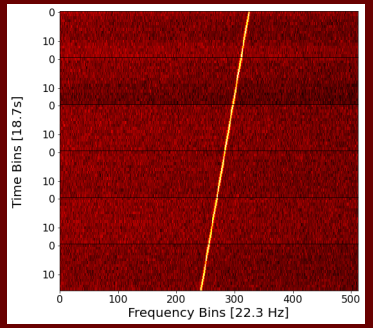
The Training Data

We simulated signals on real **Dynamic Spectra**

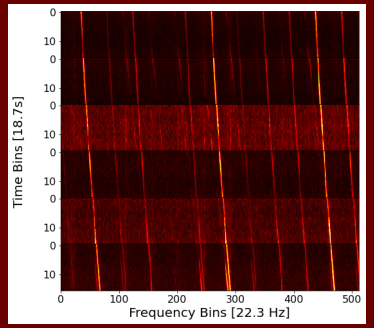
We used SETIGEN, which drew signals into observations

We randomly sampled parameters of SNR, drift, width

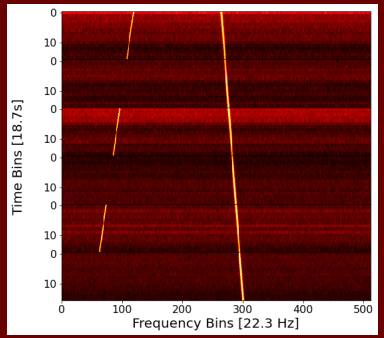
We created 4 categories, empty, false, true, single true.



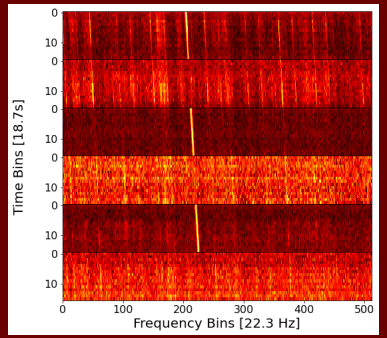
False



False Empty



True



Single True

The Training Data

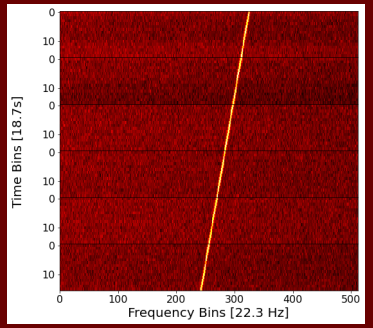
We simulated signals on real **Dynamic Spectra**

We used SETIGEN, which drew signals into observations

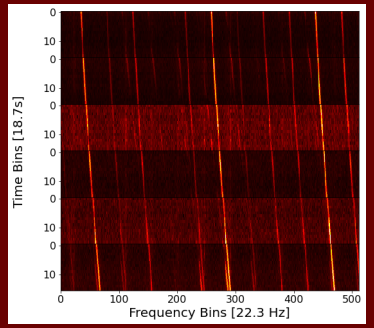
We randomly sampled parameters of SNR, drift, width

We created 4 categories, empty, false, true, single true.

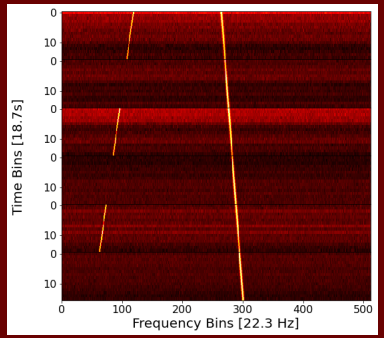
Each sample contains a set of 6 observations each of a Small window of the actual band.



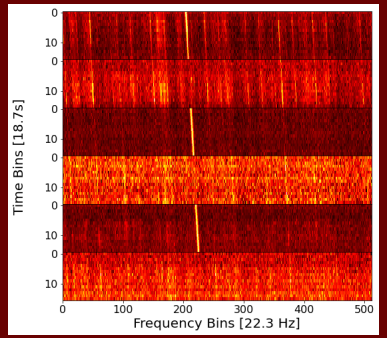
False



False Empty



True

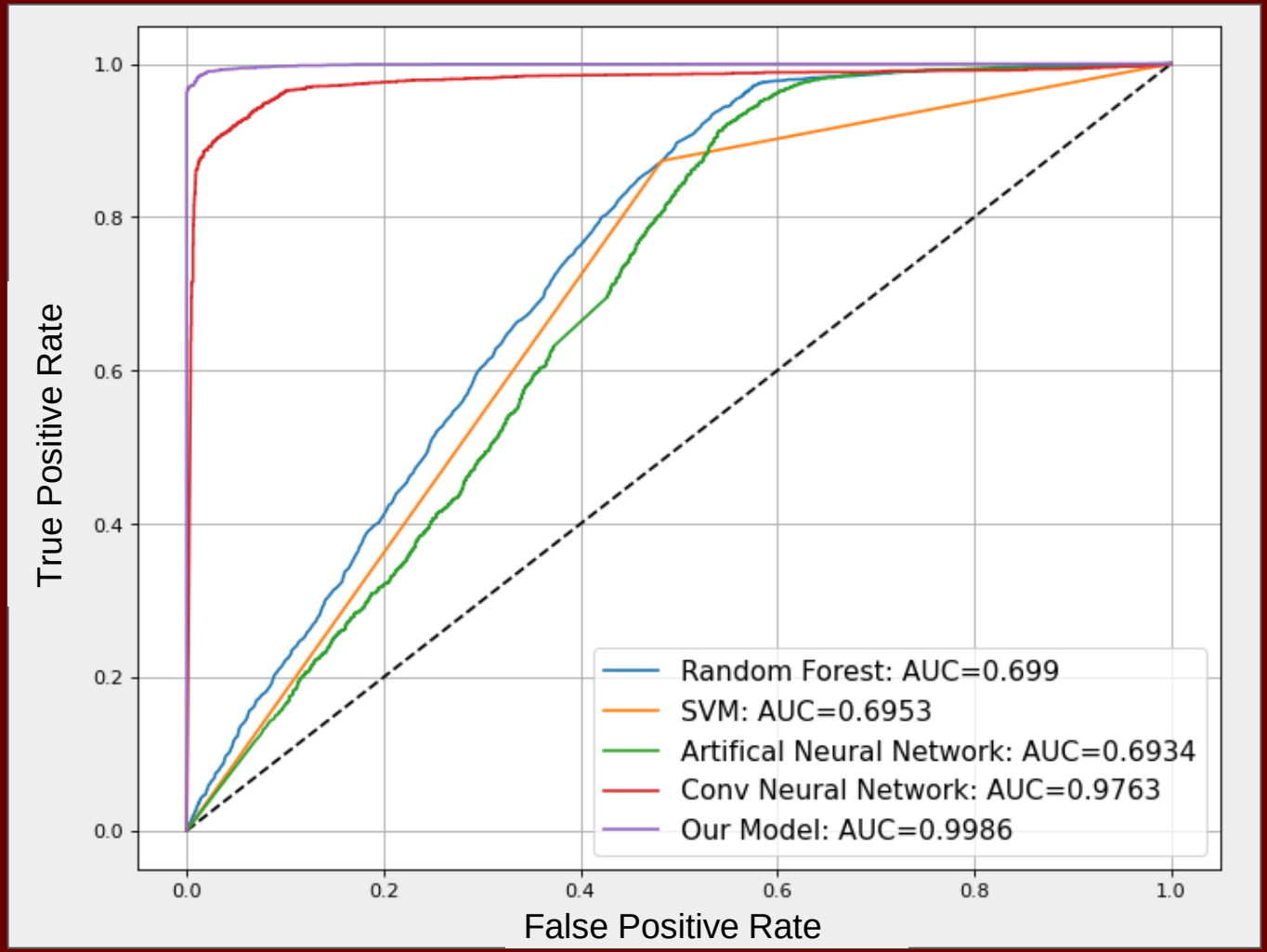


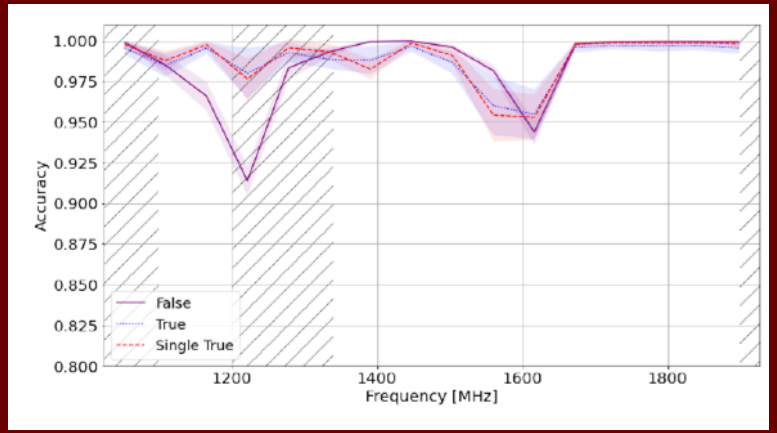
Single True

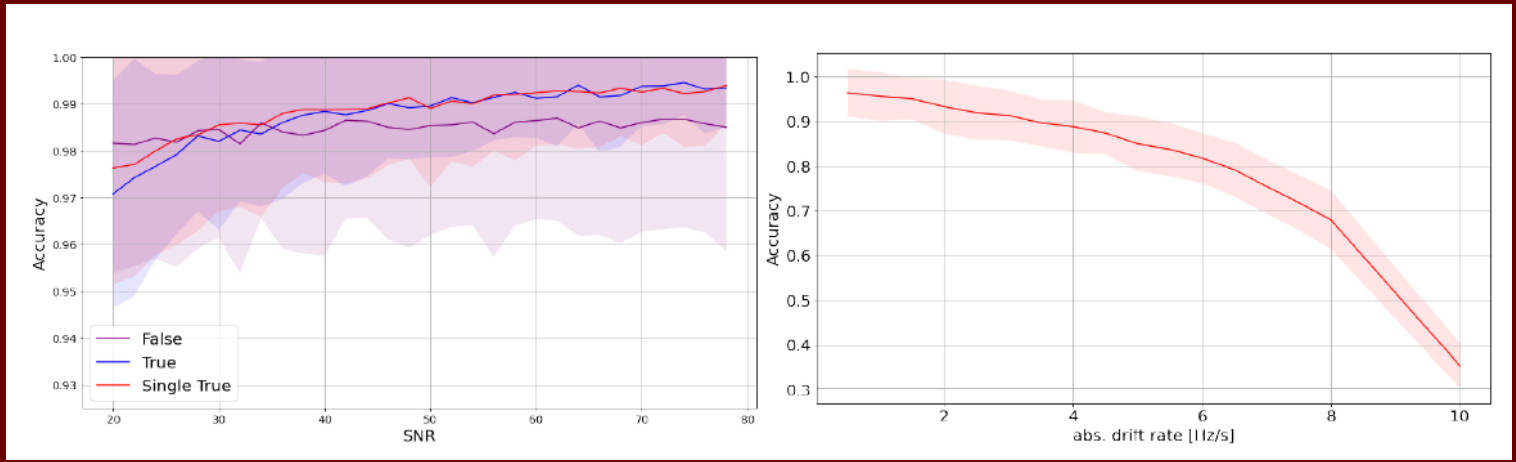
Test set was done similarly.

[just different parameters and backgrounds]

How well does it perform?


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest: AUC=0.699
SVM: AUC=0.6953
Artifical Neural Network: AUC=0.6934
Conv Neural Network: AUC=0.9763
Our Model: AUC=0.9986


1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
Accuracy
False
True
Single True
1200
1400
1600
1800
Frequency [MHz]

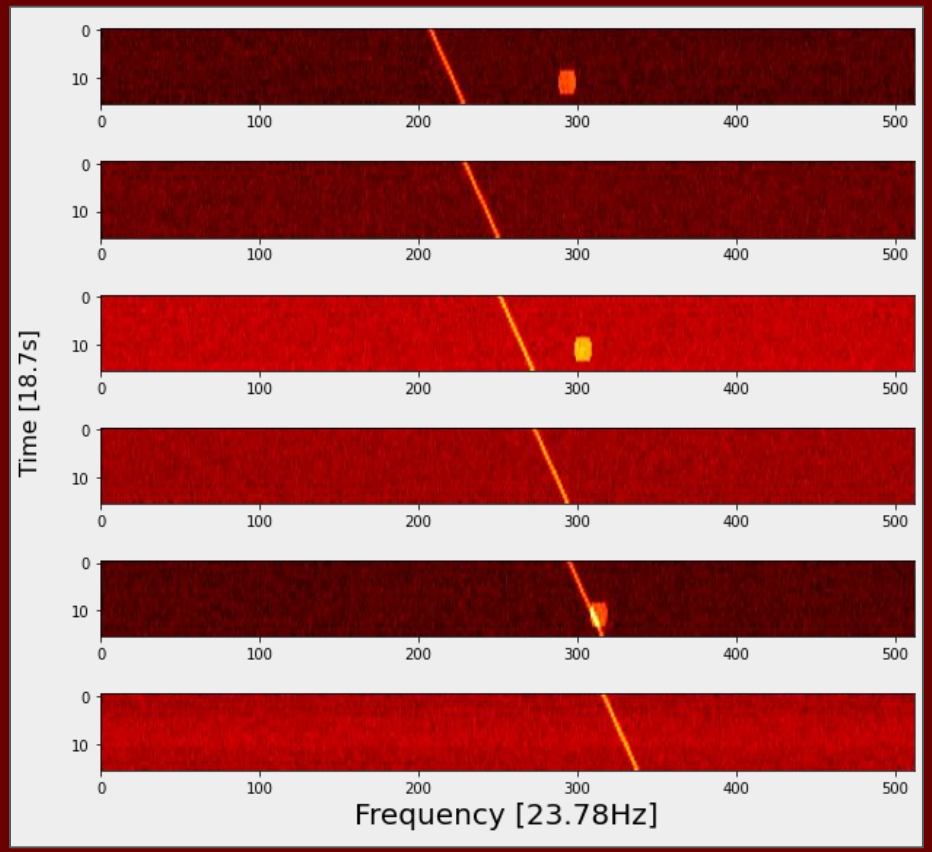

1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
Accuracy
False
True
Single True
20
30
40
50
60
70
80
SNR
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Accuracy
2
4
6
8
10
abs. drift rate [Hz/s]

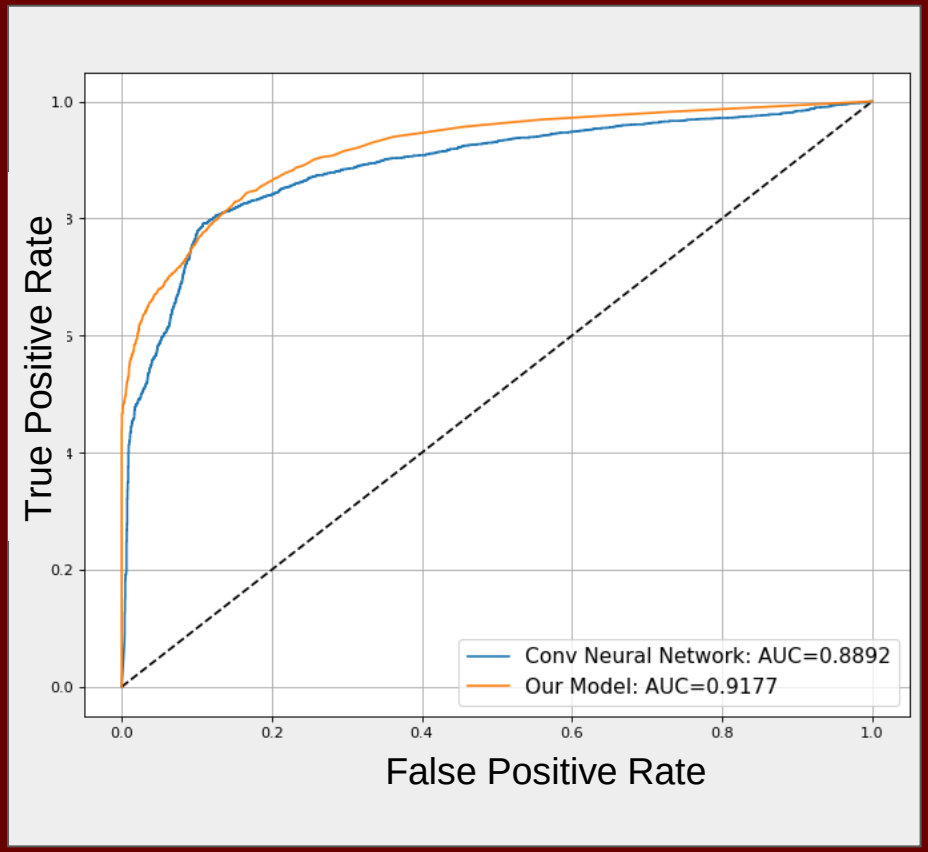
How well does it perform?

Overall Accuracy: 97.5%

Overall AUC 0.998

Can It Detect signals its never seen before?





Conducting The Search

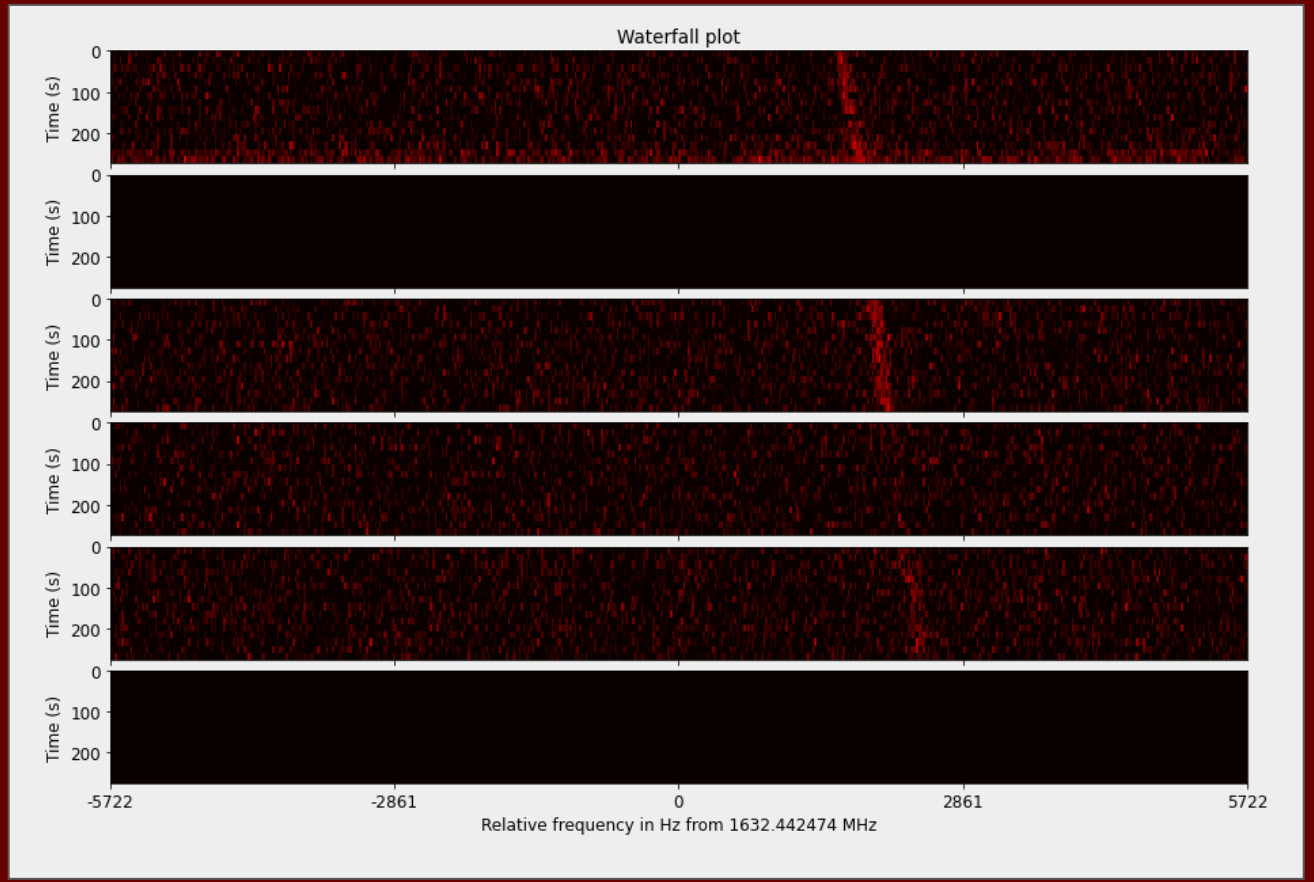
Goal: Conduct the first ML Search on GBT L band data at a max drift rate of 8.6Hz/s with 820 unique stars and 120 TB of data

Runtime: **16-18mins** / 30 Mins Obs [faster than RT]

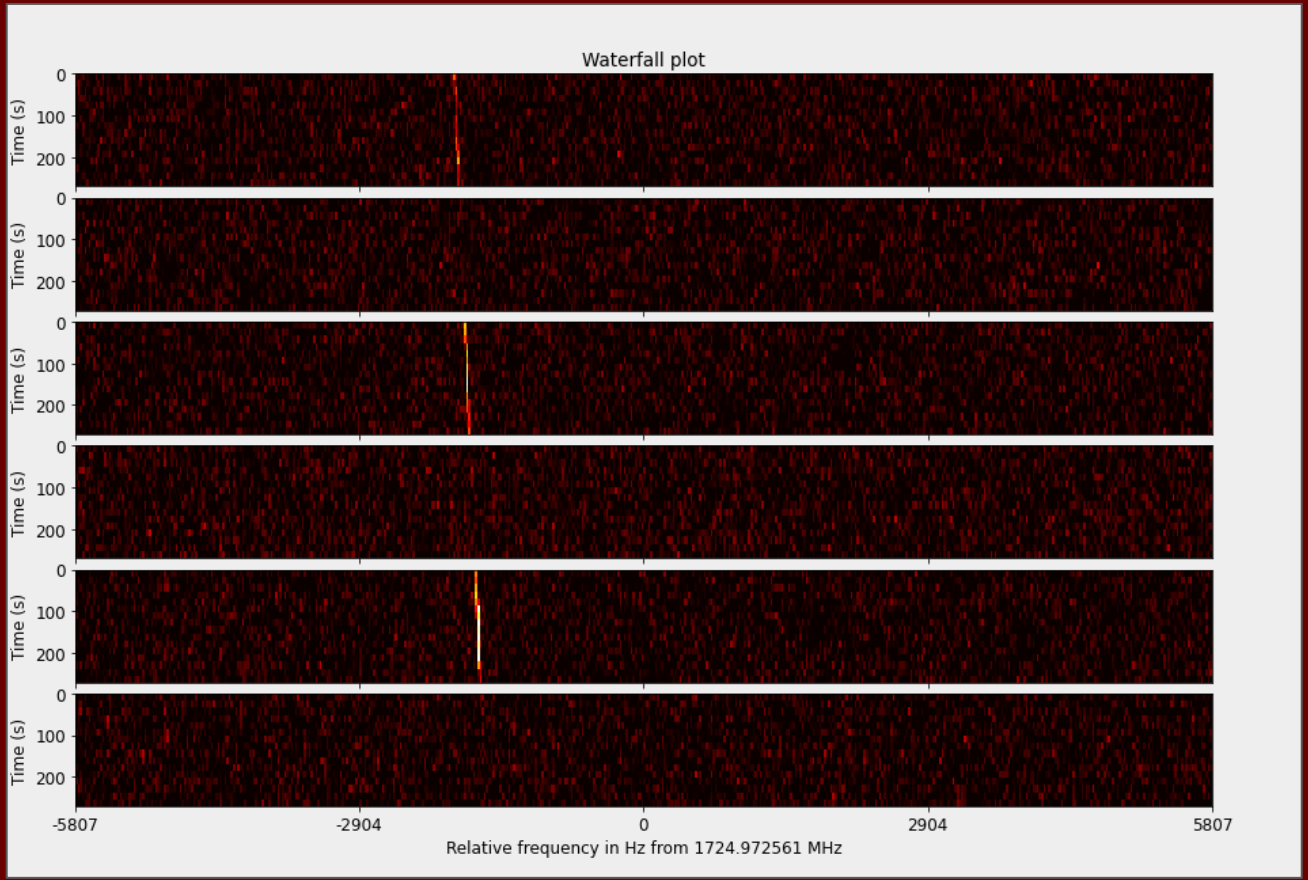
Resources: 120 CPU cores, 16 Nvidia Titan X’s

Approximately 1 week for all GBT L band search

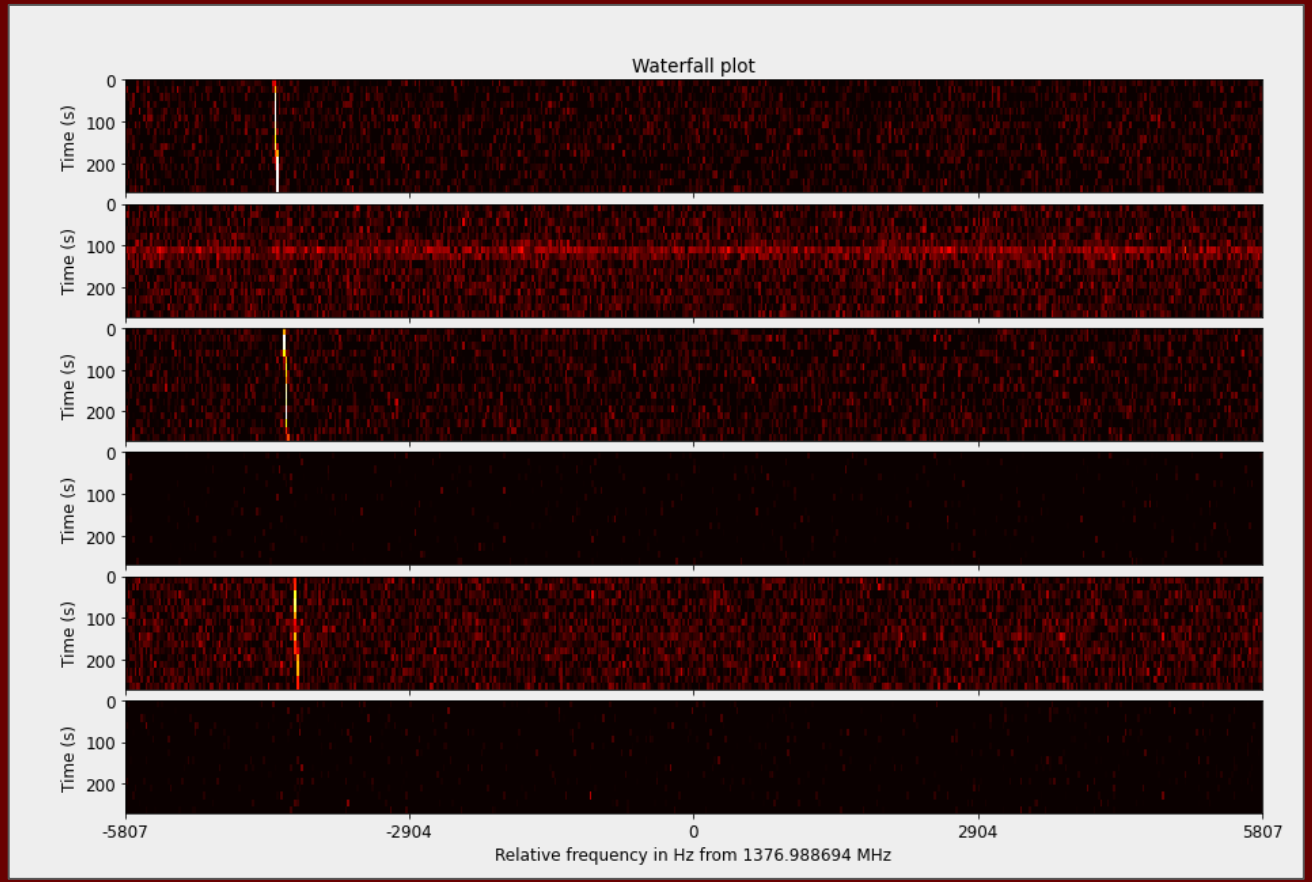
Results - Never found before



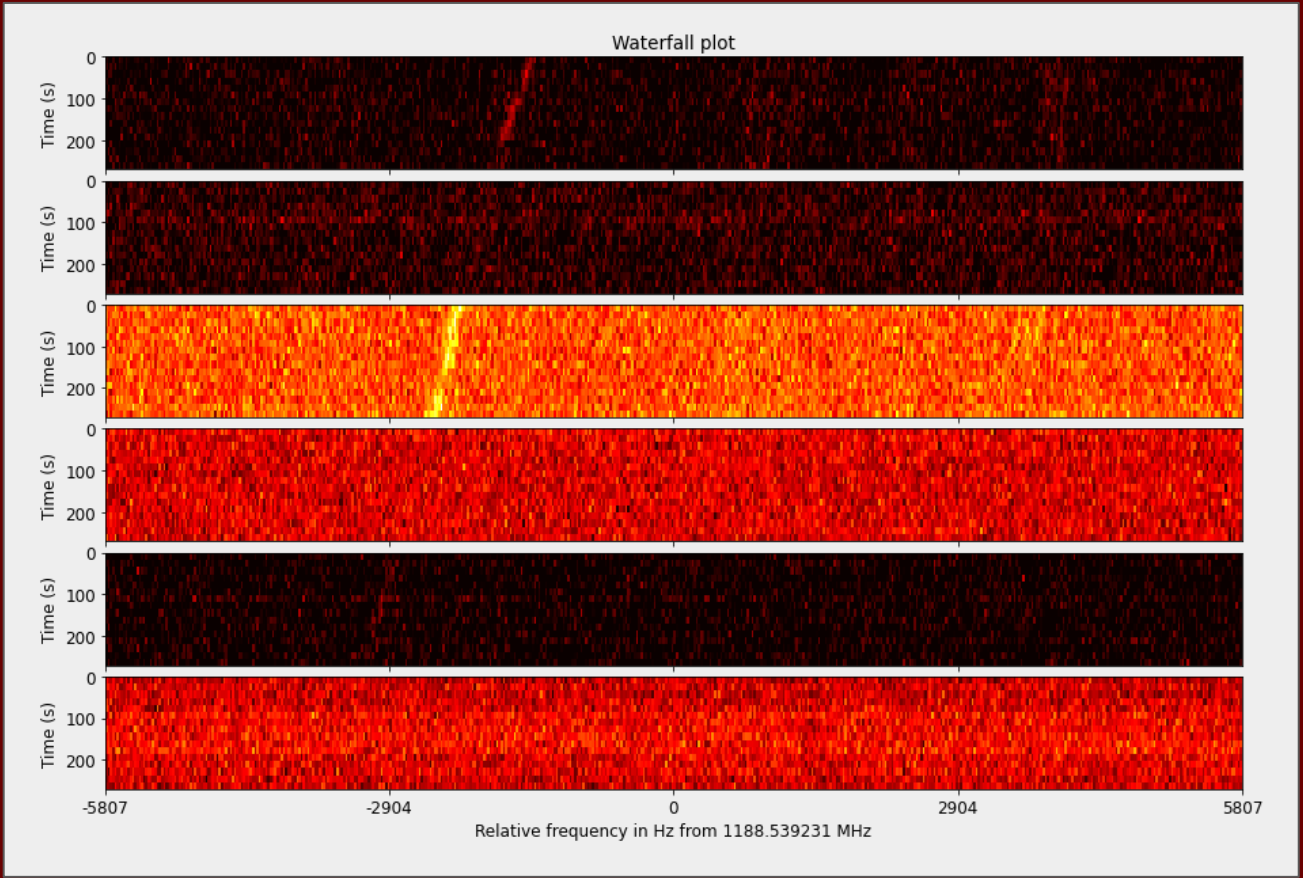
Results - Never found before



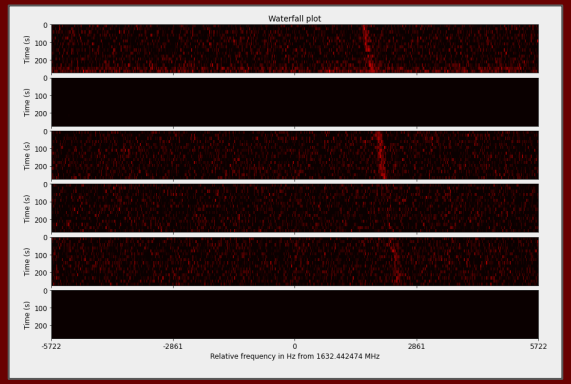
Results - Never found before

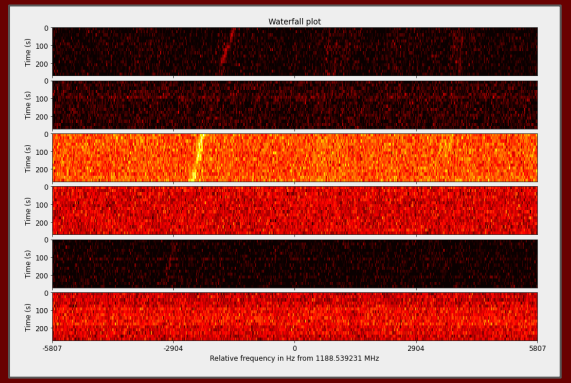


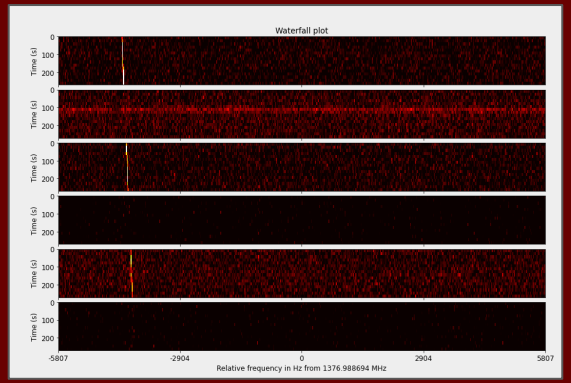
Results - Never found before

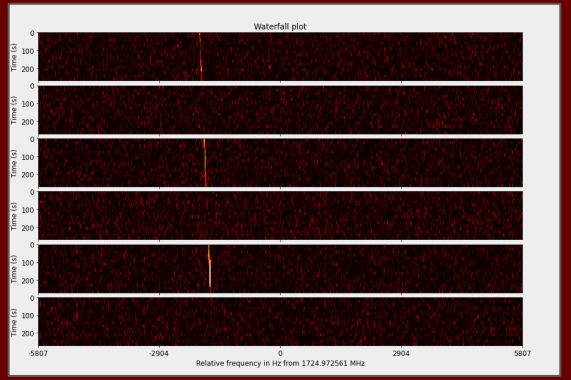


Results - Never found before









Details of our findings are documented in our paper under review @ Nature Astronomy.

Find me at:

Personal Website: Peterma.ca

Email: Peterxy.ma@gmail.com

Twitter: @peterma02

Github: @Petchma

LAIDLAW FOUNDATION

BERKELEY SETI RESEARCH CENTER

UNIVERSITY OF TORONTO

遠馬印驤