BERKELEY SETI RESEARCH CENTER

S9307 Artificial Intelligence in Search of Extraterrestrial Intelligence

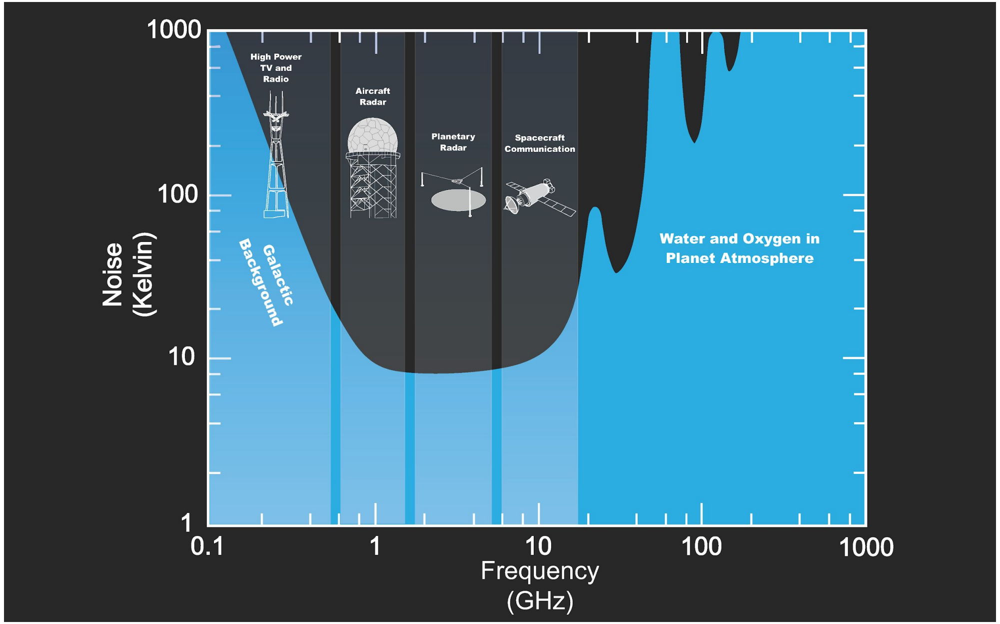
Yunfan Gerry Zhang
PhD Candidate, UC Berkeley

GPU Technology Conference 2019

Artwork by Danielle Futselaar

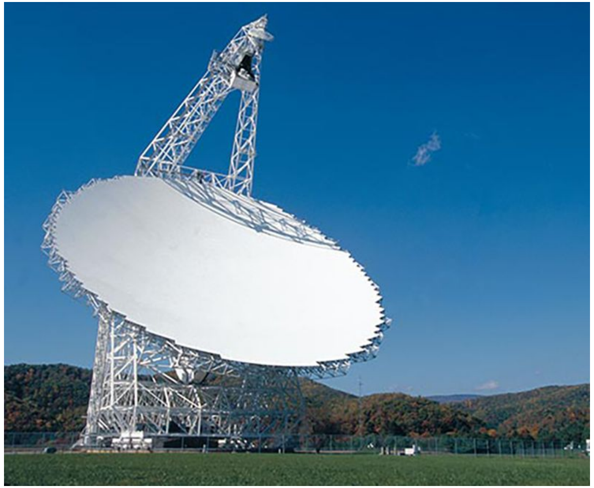
Search for Extraterrestrial Intelligence (SETI)

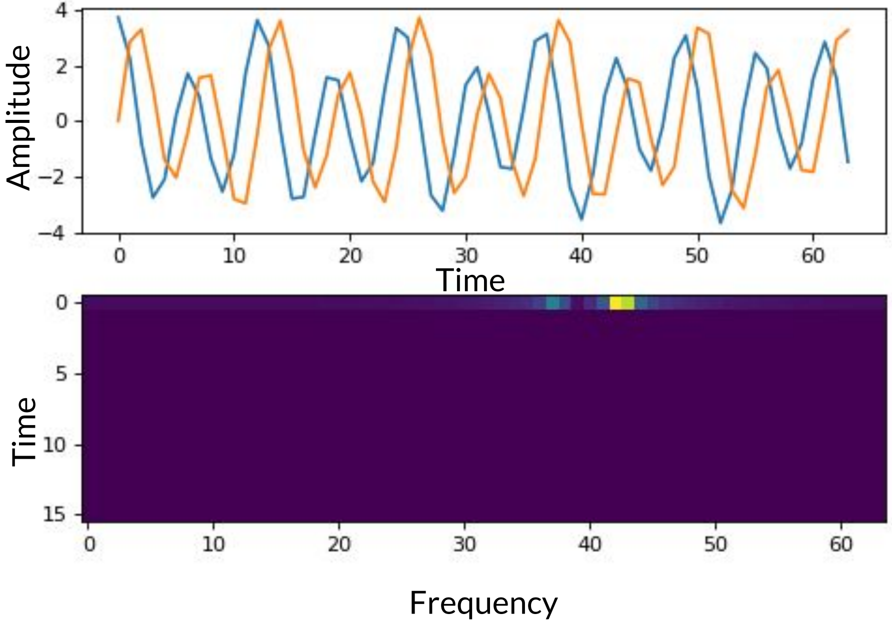
- Technological signals from space.
- Radio band of transparency.
- Main challenges:
  - Unknown signal of interest
  - Unlabeled data
  - Unbalanced data with radio frequency interference (RFI)
- Need algorithm with minimal human supervision



Source: seti.berkeley.edu

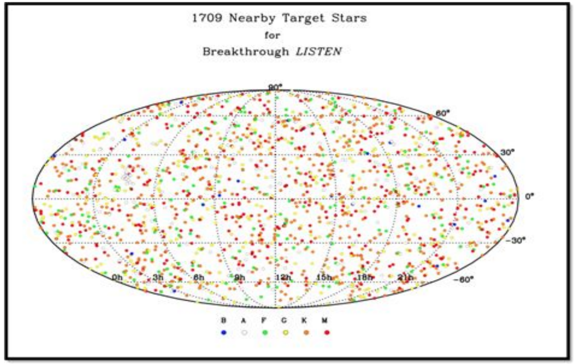
Where does RF data come from?



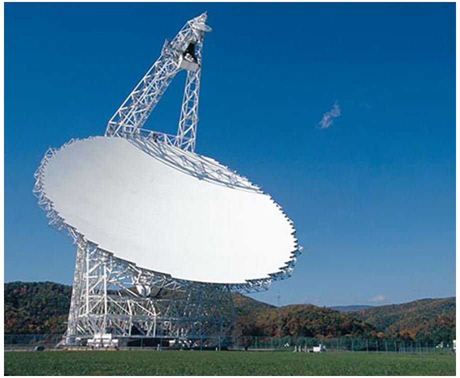


Breakthrough Listen

- Telescopes: Green Bank Telescope, Parkes Telescope, MeerKat Array
- Mission: 1 million stars, 100 galaxies narrowband search.

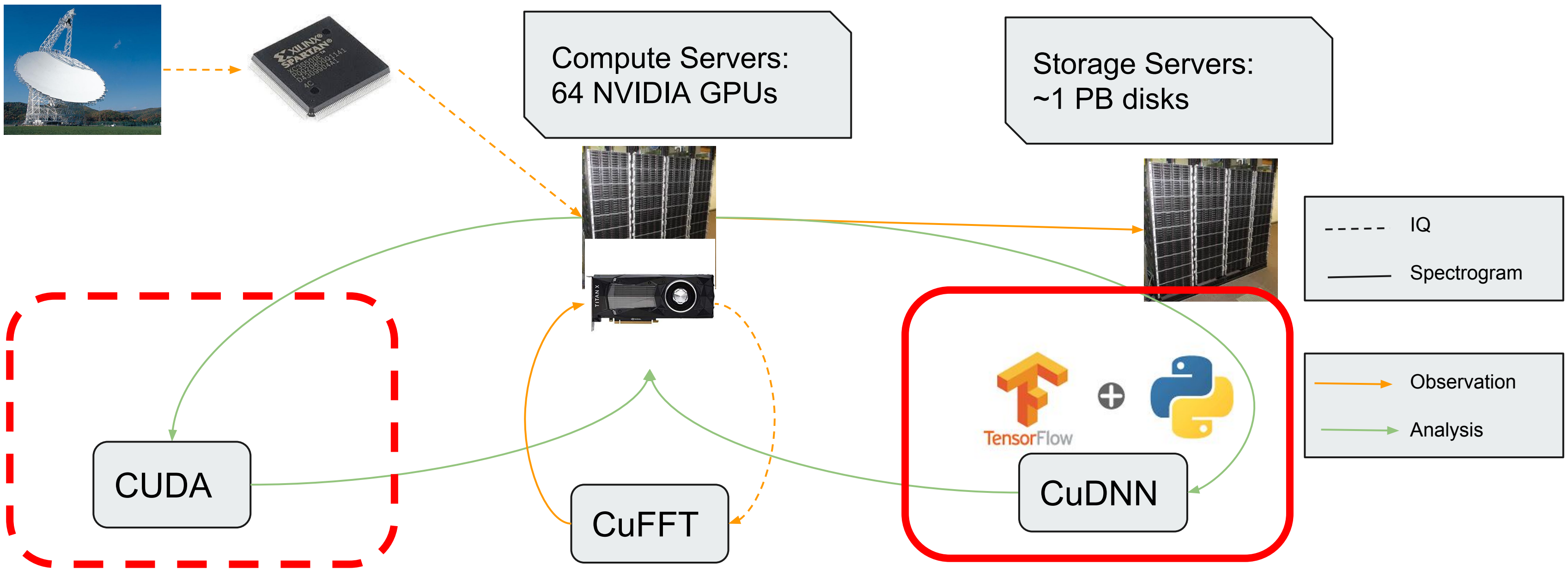


Source: [1]



- Data rate: 1PB/day IQ, 10 GHz bandwidth
- Need massively parallel hardware for data processing

GPU essential from observation to science



Goals of AI and Machine Learning

- Classification
- Regression/Clustering
- Understanding

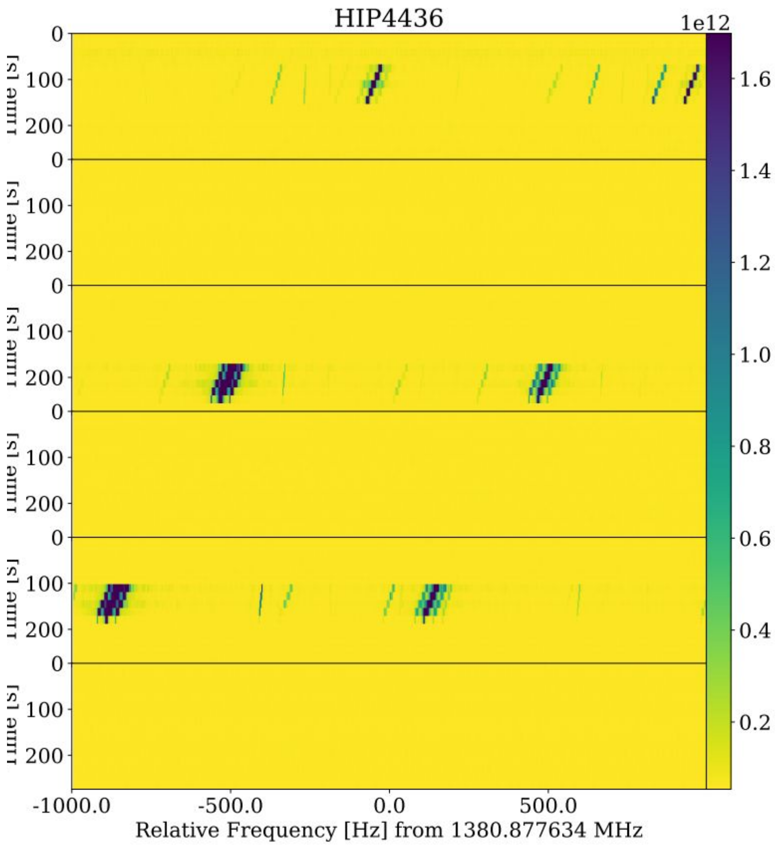
- Detect known signal
- Detect unknown signal
- Characterize the data domain

Outline

1. Core topics
   a. Fast radio bursts
   b. Blind detection
   c. Representation learning
   d. Predictive anomaly
2. Other topics
   a. IQ signal processing and modulation classification
   b. Narrowband algorithm

Preliminaries I: Spatial Filtering

- Simultaneous or sequential observations of multiple areas of the sky.
- Signal in multiple areas:
    - local RFI
- Signal in one area:
    - potential candidate



Source: [2]

Preliminaries II:
How spectrograms differ from camera images?

- Resolutions:
  - (0.3ms, 0.35MHz), (1s, 0.3kHz), (18s, 2.8Hz)
- Data shapes (5 mins, S-band):
  - (1e6, 1e4), (273, 3e5), (16, 3e8)
- Information sparsity
- Large variations in signal support

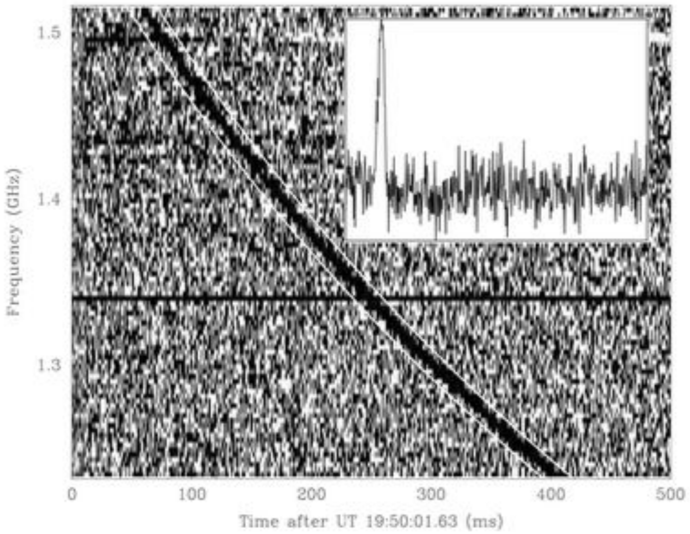
Deep learning architecture considerations

- Known signals:
  - Fixed size sliding window with targeted resolutions
- Unknown signals:
  - Use energy detection to reduce sparsity
- Image pyramid
- Attention mechanisms

I. Finding known signals

Fast Radio Bursts

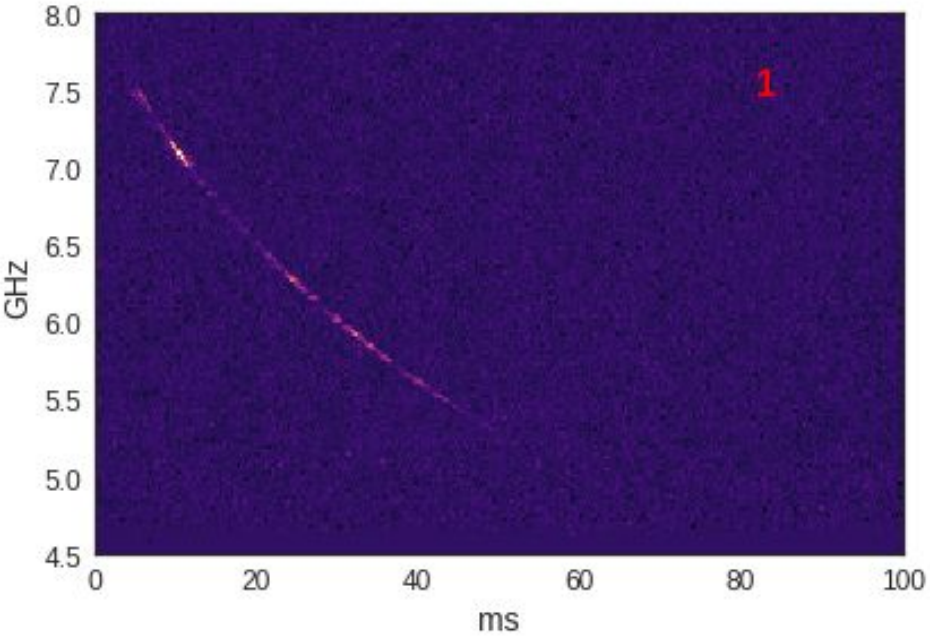
- Millisecond-duration signals of unknown origin.
- Quadratic dispersion with large dispersion measure, suggesting extra-galactic source.
- One has been observed to repeat (FRB121102), leading to localization in a dwarf galaxy 3 billion light years away.



Source: [3]

Deep Learning Detection

- Observation on August 26, 2017
- 21 bursts originally reported
- **72 DL discovered**



Search

NVIDIA DEVELOPER

NEWS CENTER

ARTIFICIAL INTELLIGENCE AUTONOMOUS VEHICLES DESIGN & VISUALIZATION GAME DEVELO
VIRTUAL REALITY

Comments 254 Shares

AI Spots Mysterious Signals Coming from Deep in Space

September 10, 2018

Fast radio bursts are some of the most mysterious high-energy astrophysical phenomena in the entire universe. They are intense blasts of radio emissions that last just milliseconds in duration and are thought to originate from distant galaxies. The exact nature of the objects is uncertain, but they could point to extraterrestrial intelligence.

MWC 2019
Asia
Fundings & Exits

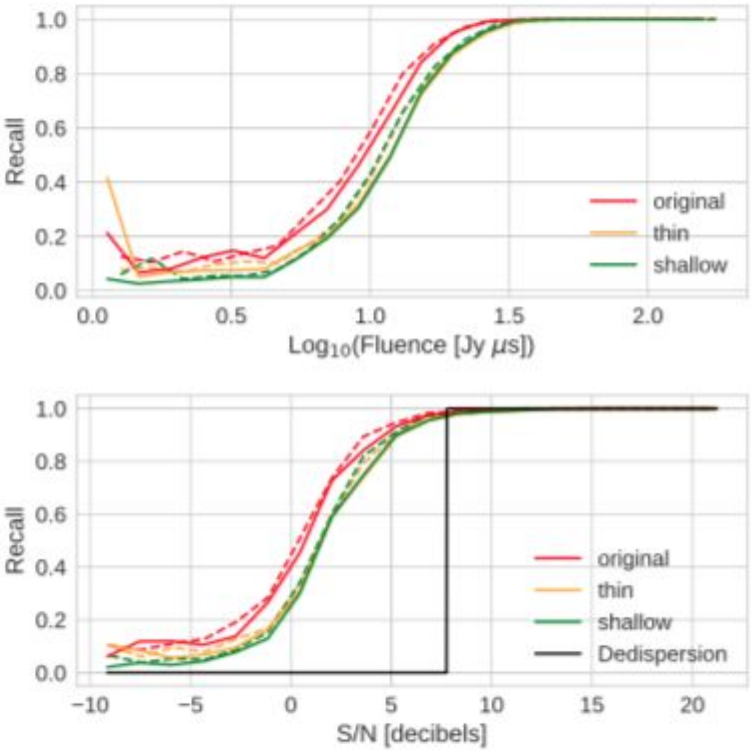
The perennial optimists at the Search for Extraterrestrial

Challenges and Solutions

- Highly imbalanced data and few positive examples
  - Solution: Simulate positive examples and inject on infinite supply of negative examples
  - Model: binary classifier on fixed size input
- Large input size and information sparsity:
  - Chop into fixed size window frame
  - Concatenation with pooling only tower (image pyramid)
  - Initial data rate reduction through large filters and strides
- Reason why deep learning can be effective
  - High modulations and local 2-dimensional detection

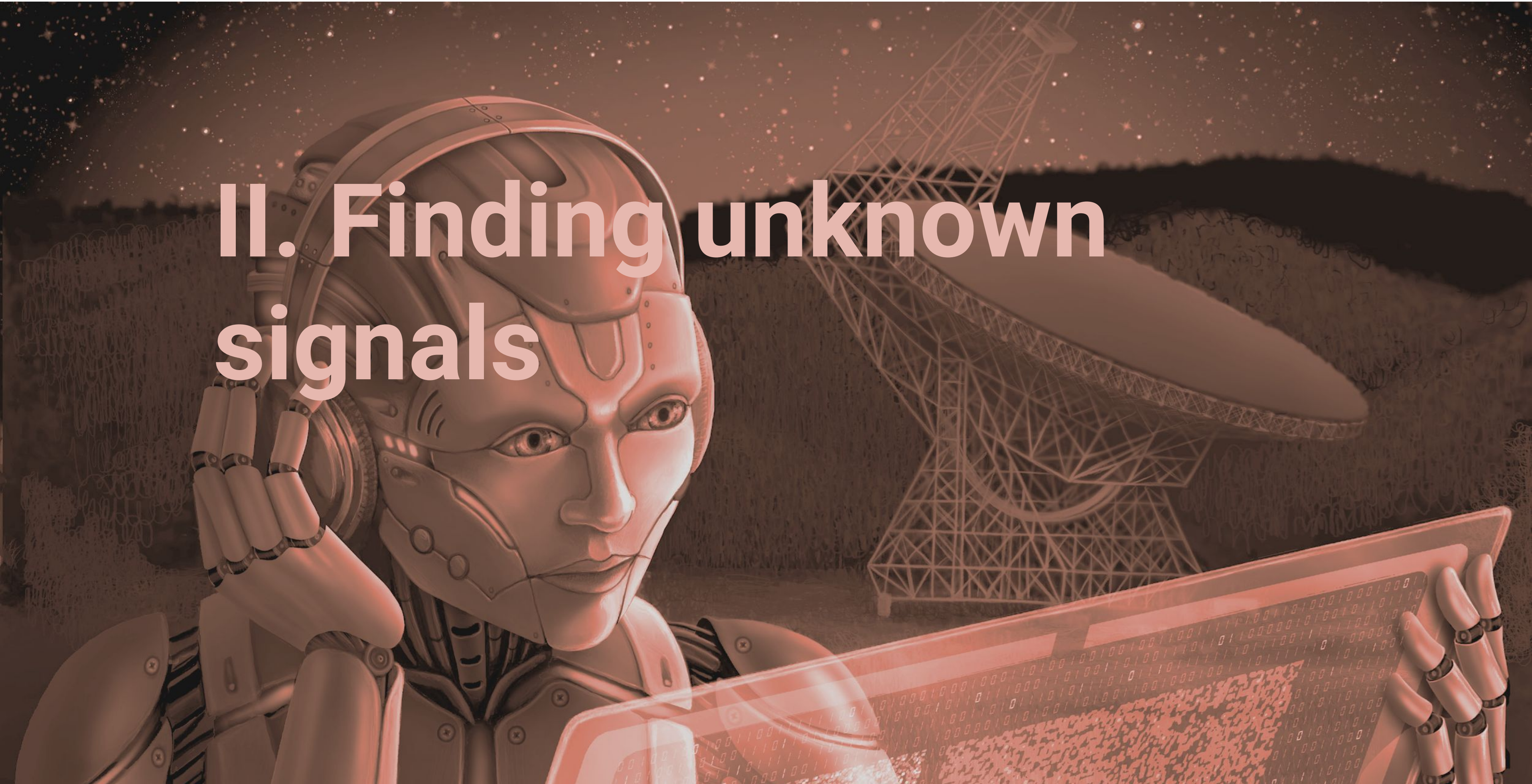
Model and performance

- Residual Network (27 layers).
- Inference speed:
  - 70 times faster than real time on single GTX 1080
  - Depends on frequency and time resolution of input
- Evaluation
  - Ambiguous ground truth
  - 93 believable out of ~300 (chosen threshold)
- Data and code available from:
  - https://seti.berkeley.edu/frb-machine/
- Paper: arXiv 1802.03137

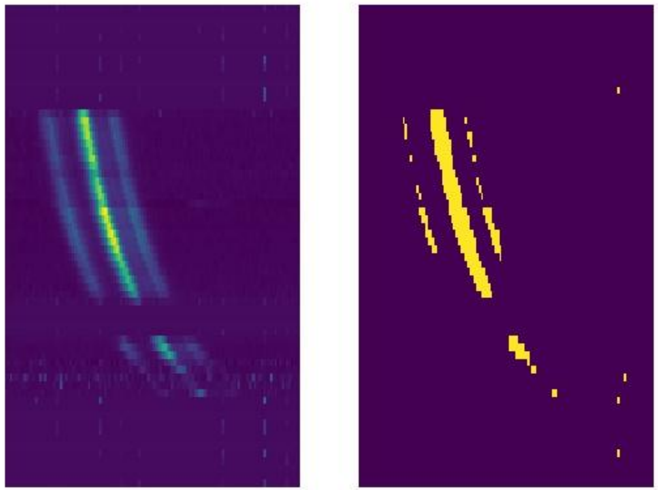


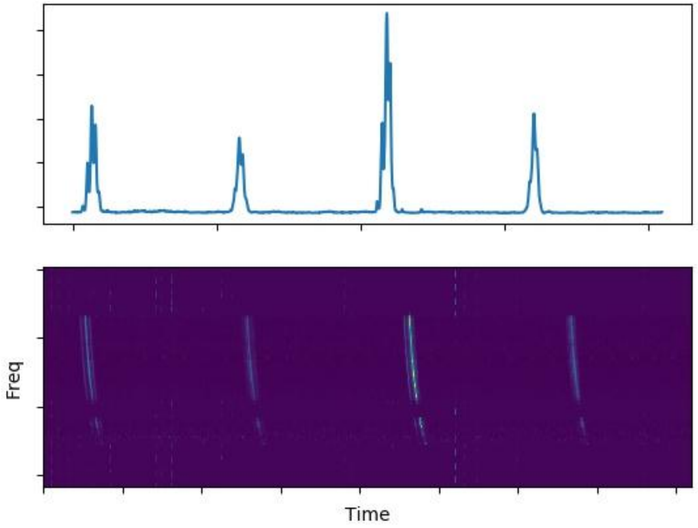
Source: [4]

II. Finding unknown signals



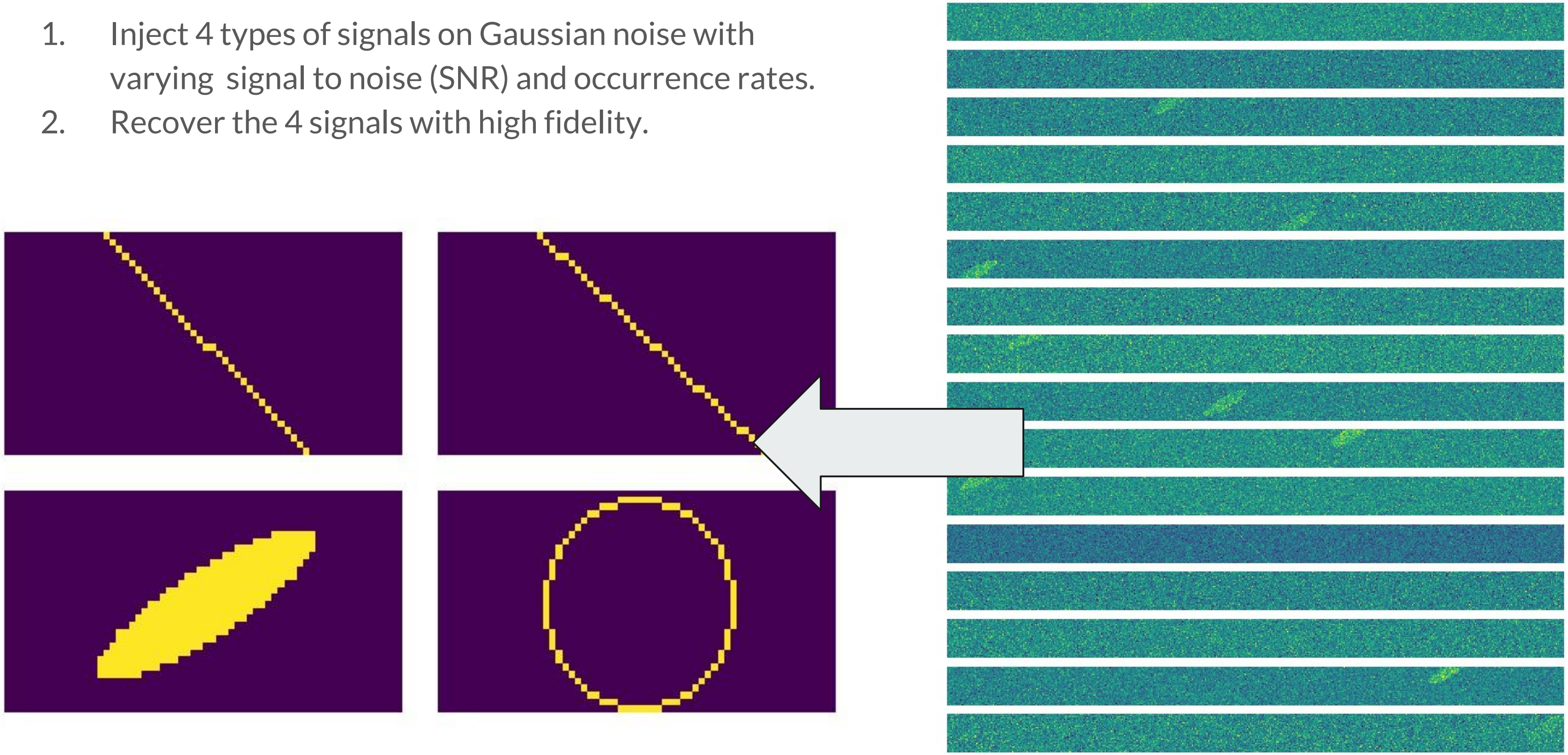
Dedispersion as Convolution





Problem formulation:

1. Inject 4 types of signals on Gaussian noise with varying signal to noise (SNR) and occurrence rates.
2. Recover the 4 signals with high fidelity.

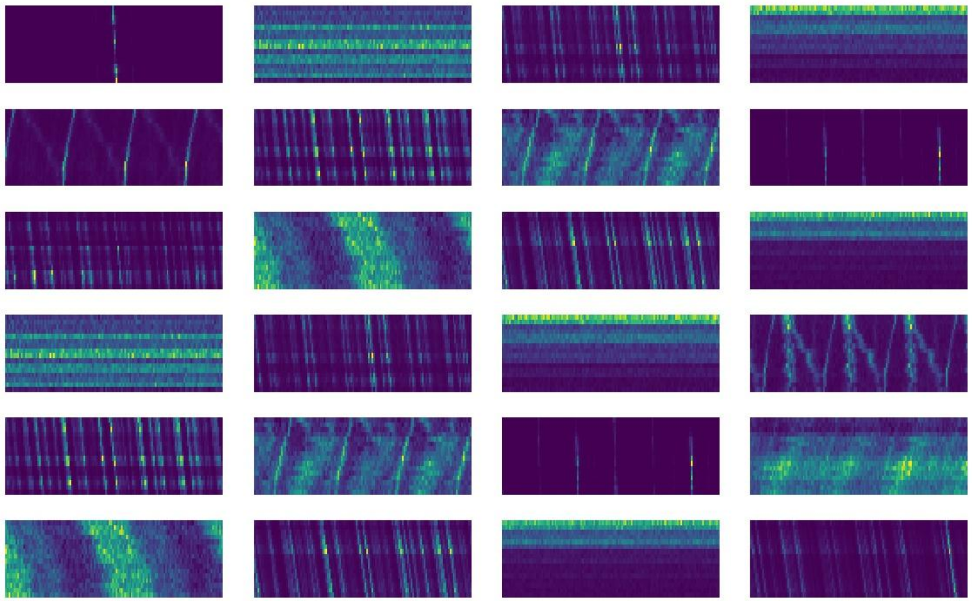


Approach

- Map: Energy detection
- Reduce:
  - Clustering.
  - Dimensionality reduction.

Map: Energy detection

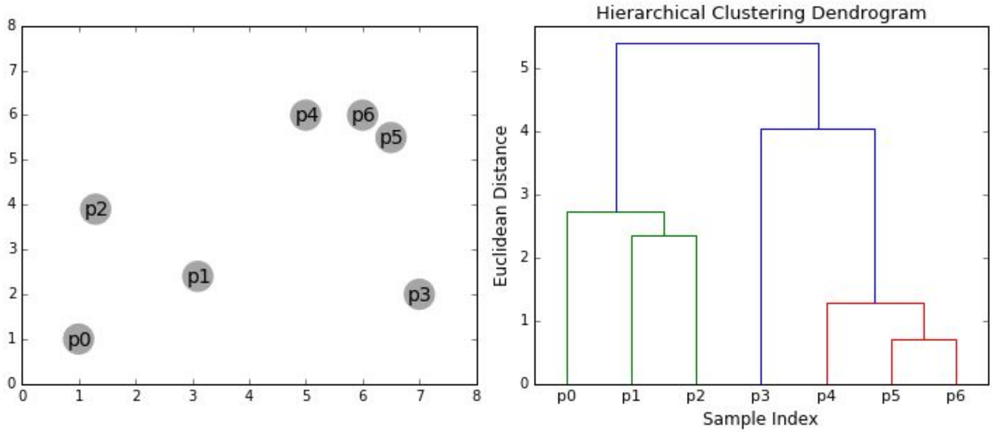
- Energy detection = threshold pixel values
- Finding patterns that do not match noise distribution (pixel-joint).
- Entropy computationally forbidding
    - curse of dimensionality.



Phase 1: Hierarchical clustering and PCA

$$\mathbf{Q} \propto \mathbf{X}^T\mathbf{X} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$$

PCA to reduce dimensionality

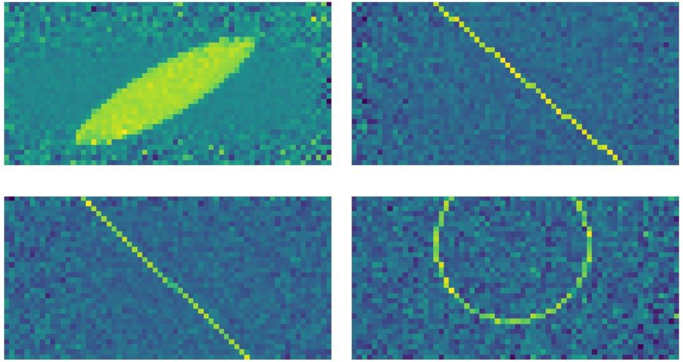


Source: [5]

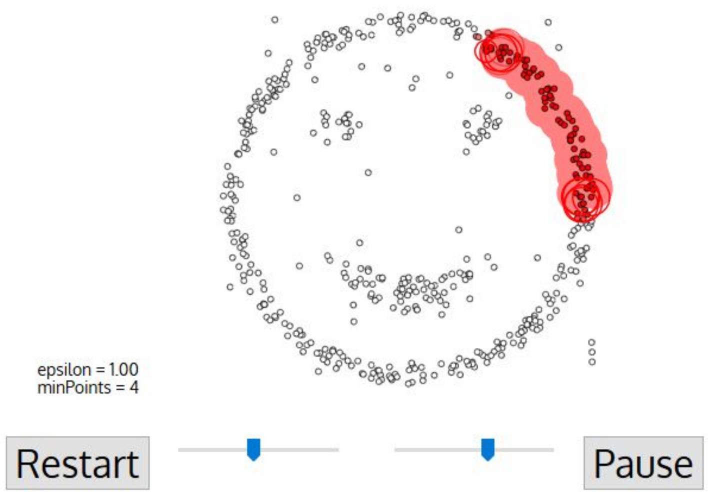
$$D = 1 - \mathtt{max}_{\delta x} \frac{|A(x) * B(x + \delta x)|}{\sqrt{|A * A||B * B|}}$$

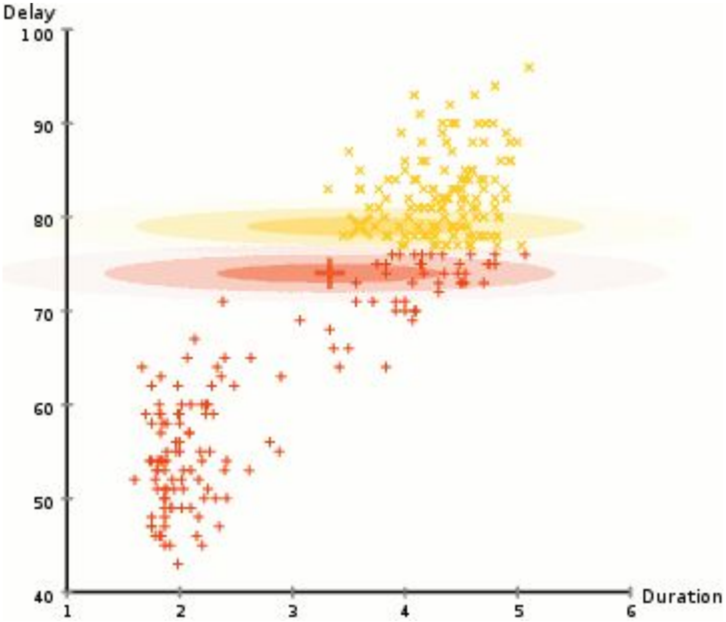
Phase 1

- Initialization
  - Map: High threshold energy detection
  - Reduce: Hierarchical clustering and PCA



Phase 2: GMM and DBSCAN

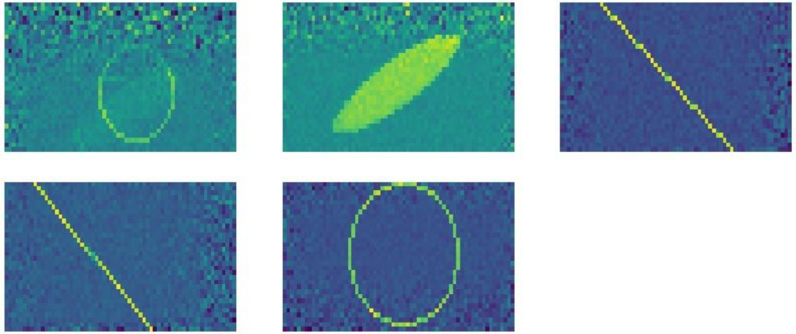


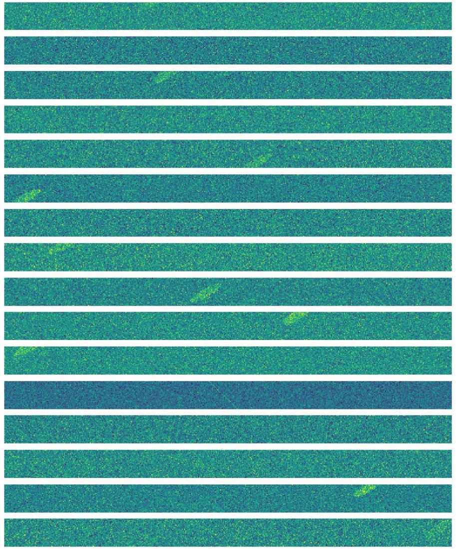


Source: [5]

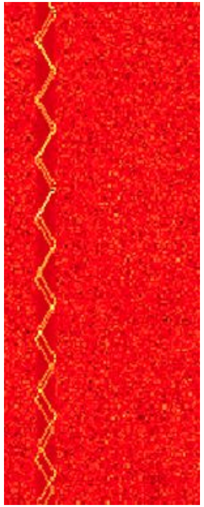
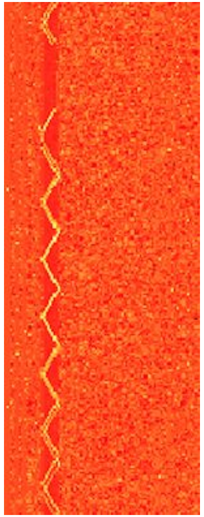
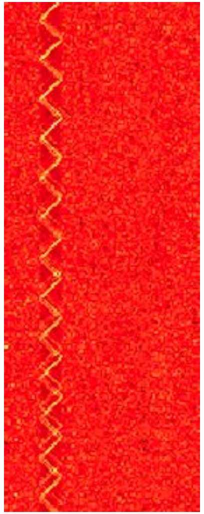
Phase 2

- Continued Learning
  - Map: Energy detection
  - Reduce 1: For existing templates, variance helps identify new examples (GMM)
  - Reduce 2: DBSCAN to locate any new clusters.
- After initial clustering, inject new signal, a circle of lower radius.

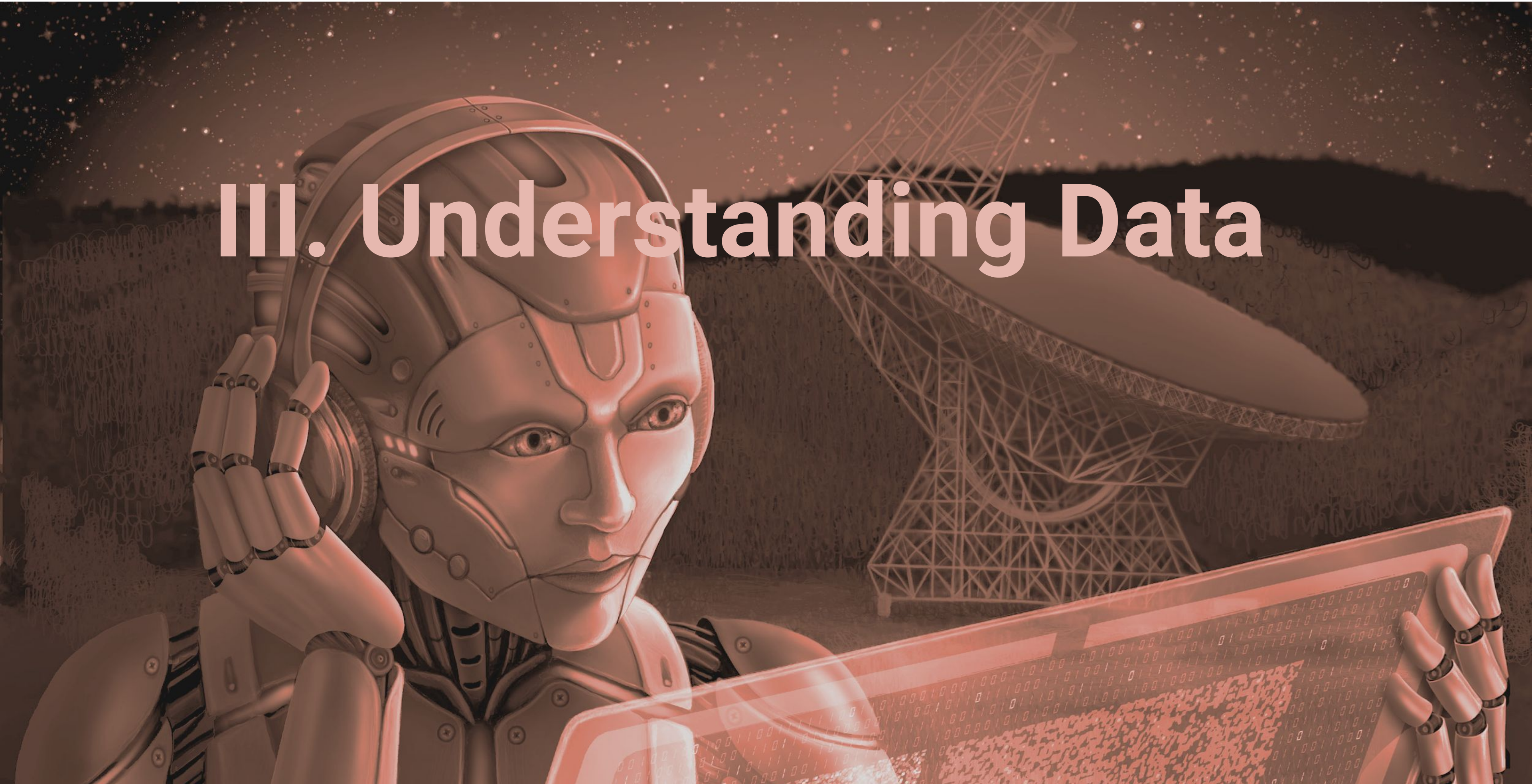




Are these similar?


III. Understanding Data



What does it mean to understand?

- Know the data comes from Fourier transforms of polyphase filterbank of complex voltage captured with receiver that……..

Or...

- Learn data distribution
  - Predict masked samples
  - Retrieve similar samples
  - Point out anomalies
  - Reduce noise on data
  - Generate new data
- Goal: develop core module usable in various scenarios

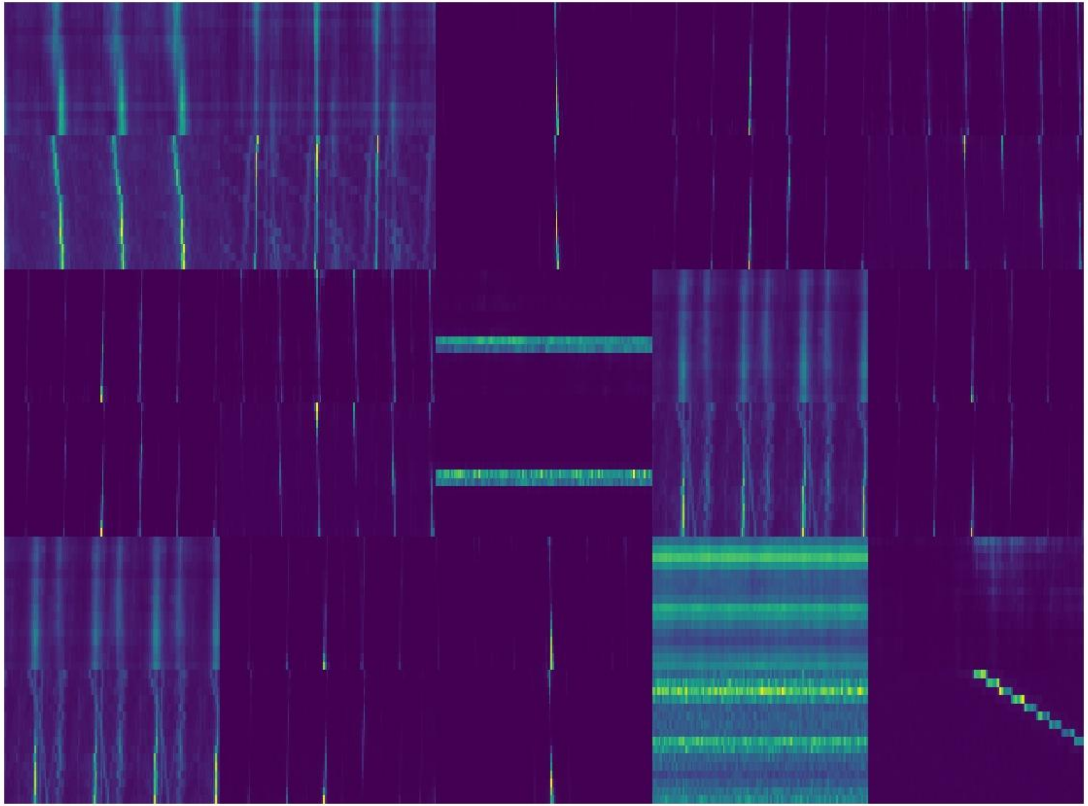
Learning Data Distribution

- Autoregressive model (e.g. PixelCNN)
  - Learns likelihood of data sample P(x)=p(x1)p(x2|x1)P(x3|x1,x2)...
- Latent Variable models
  - Compress data into compact representation.
  - Auto-encoder and its many variants.
  - Auxiliary tasks: rotation prediction, jigsaw puzzle solving, adversarial discrimination etc.
  - Latent variable + clustering objective

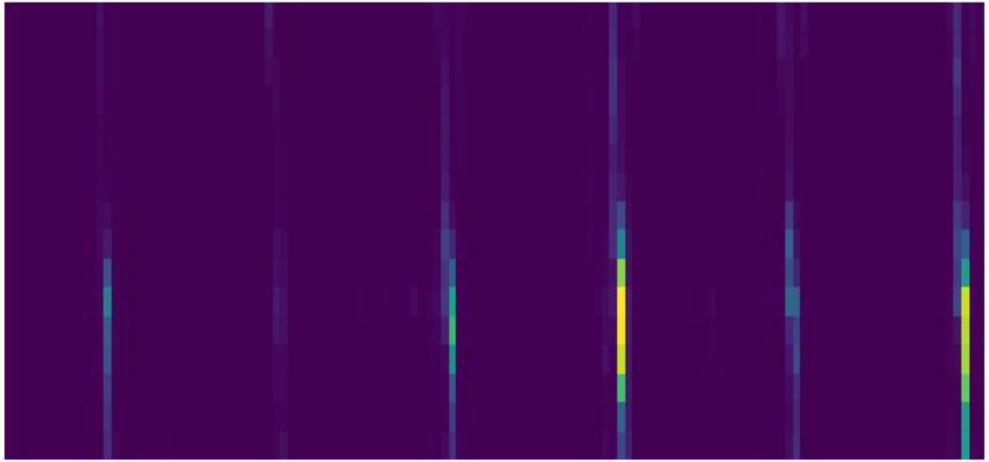
Reconstruction

Convolutional encoder, fully connected decoder.

2048 input → 64 hidden vector length.



Latent Space Interpolation

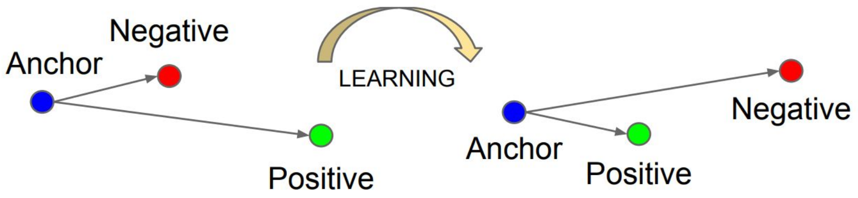


Convolutional encoder, fully connected decoder.

GMM clustering (10 clusters).

How to improve the representation?

- Clustering objectives
  - Potential risk of mis-clustering
- Translation invariant auto-encoder
  - Partial view of signals
- Semi-supervised learning
  - Human labels
  - **Coarse channel (noisy labels)**
  - **Permutation of multi-frame observations**
  - **Robustness to perturbations (translation, scale etc.)**
- More expressive architecture



Source: [6]

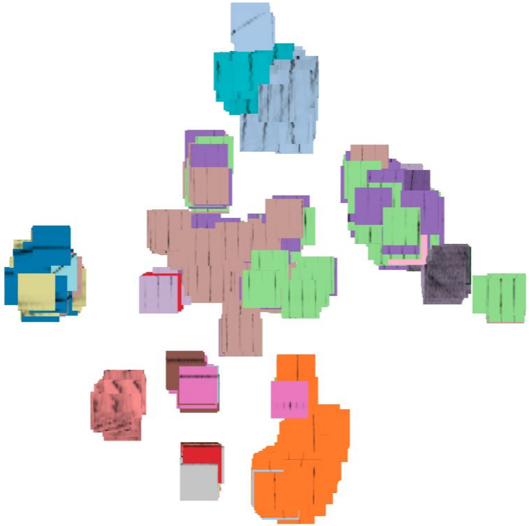
With triplet-loss and coarse channel

Loss function

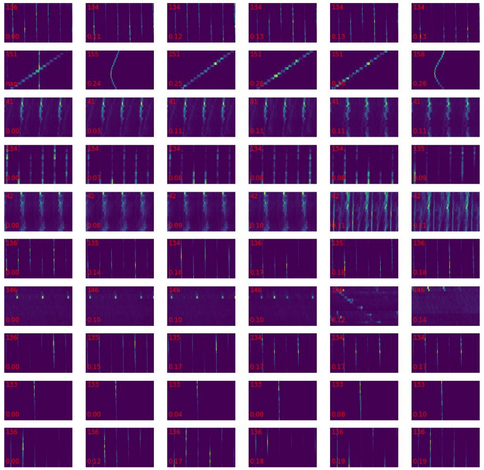
$$\mathscr{L} = \alpha\mathscr{L}_{reconstruct} + \beta\mathscr{L}_{triplet}$$

- Noisy data:
  - low α
- Noisy label:
  - low β

Tensorboard demo



Evaluation



Top 5 accuracy

Evaluate top 5 candidates with 500 queries in test set of 10000

Model \ Experiment	0 added noise	-10 dB (no retraining)	-10 dB training
Coarse channel	79.0%		
FC (β=0)	95.6%	86%	
FC (α=3β)	98.8%	86%	97.7%
Conv (α=3β)	99.8%	78%	98.9%

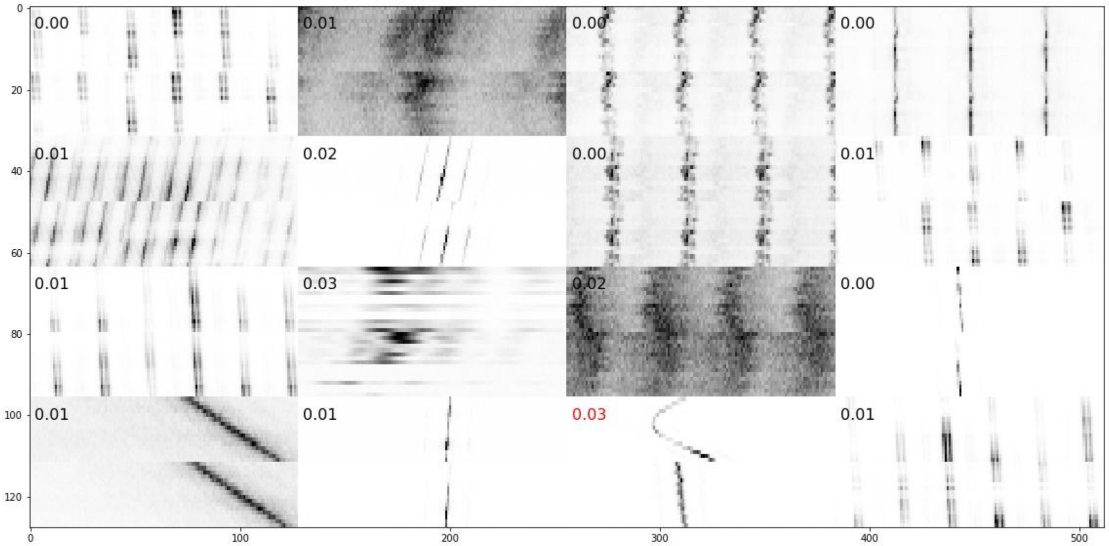
Data Query

Database searching and anomaly detection { z: (img, meta)}

Dot product distance (|z|=1):

- d = 1 - z · z_

Webapp: http://35.192.106.72/



High level applications

- SETI search pipeline: beam comparison
- Outlier detection
- RFI environment characterization

ML/astronomy paradigm separation!

Stay tuned for publication, blog post, data and code release!

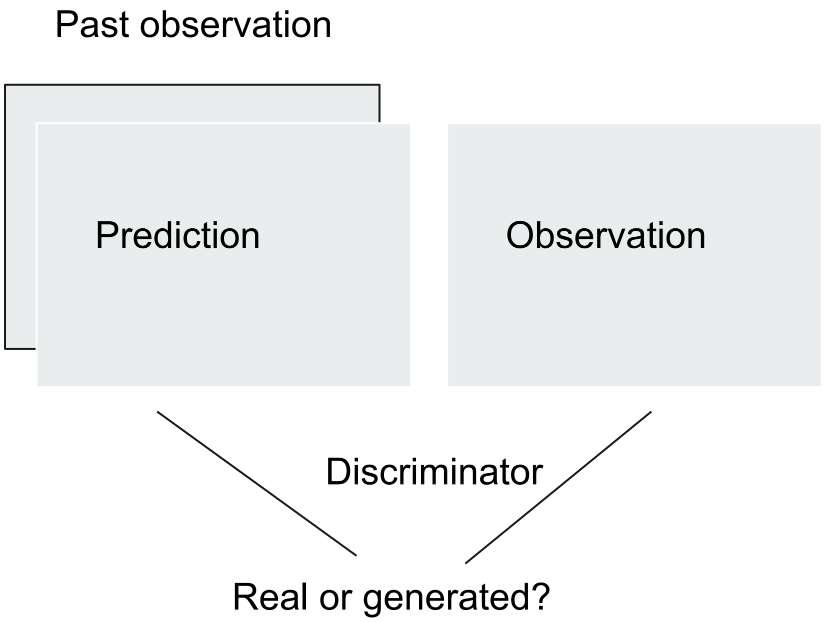
III -b. Sequential data

Predictives Anomaly Detection on Spectrograms

- Detect anomalies by predicting future observations
- RFI filtering in same framework.
- Time series prediction: RNN and LSTM
- Spatial/frequency dimension: convolution
- Challenge: noise is not predictable
- Solution: introduce discriminator

$$L_g = \log(1\text{-}D(G_{\text{future}}))$$

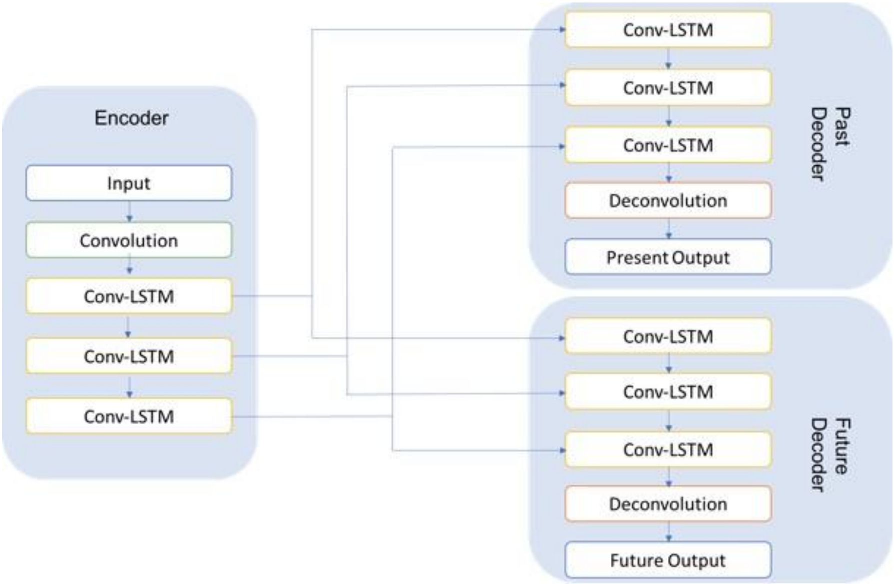
$$L_d = \log(D(G_{\text{future}})) + \log(1\text{-}D(x_{\text{future}})).$$



Architecture

- Convolutional LSTM baseline
- Dual decoder
  - Better representation
  - Learn data distribution
- Multiple frames at a time
- Generative Adversarial Loss
  - Regulated training to counter instability

$$L_G = \alpha(L_{\ell 2\text{-future}} + L_{\ell 2\text{-past}}) + \beta L_{\ell 2\text{-feature}} + L_g,$$



Source: [7]

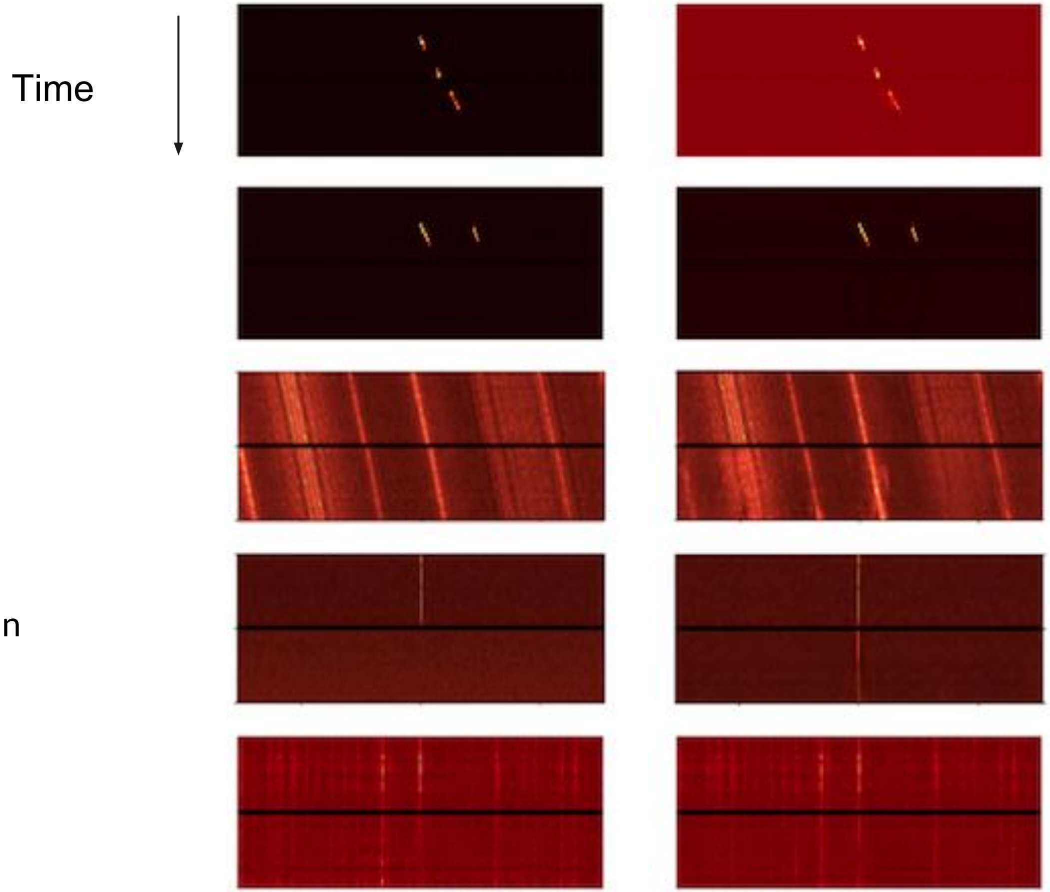
Prediction Results

Dataset:

20000 instances of 256 X 16 candidate spectrograms.

Advantages:

- High fidelity prediction
- Understands discontinuity of signals
- Agnostic to signal type
- Self-supervised learning needs no human labels



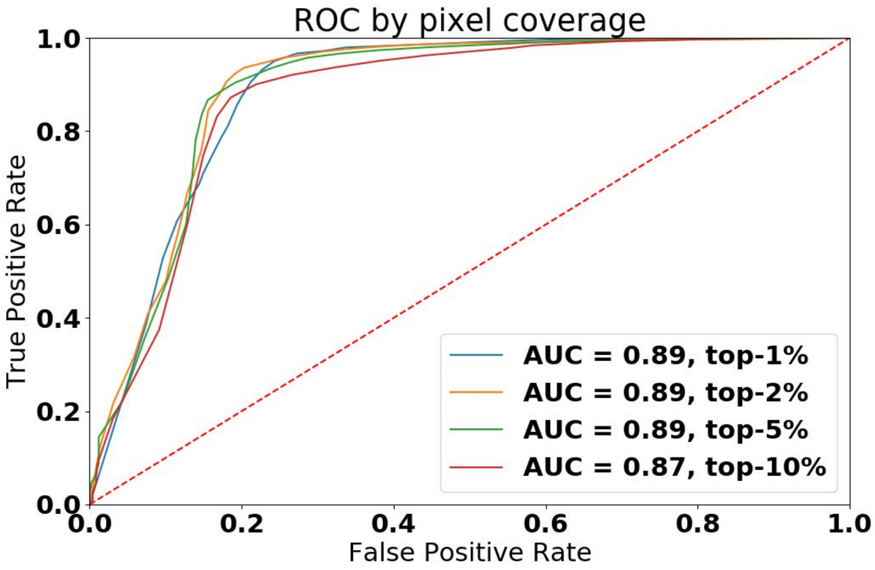
Source: [7]

Anomaly Detection Evaluation

Pair correspondence with top pixel coverage:

$$\tau \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\|m_1 \& m_2\|}{\|m_1 | m_2\|},$$

False positives due to selection criterion, not prediction model.



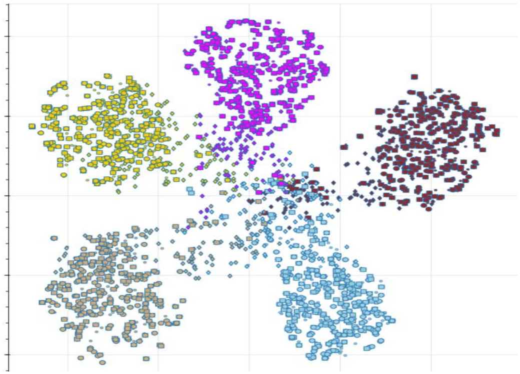
Source: [7]

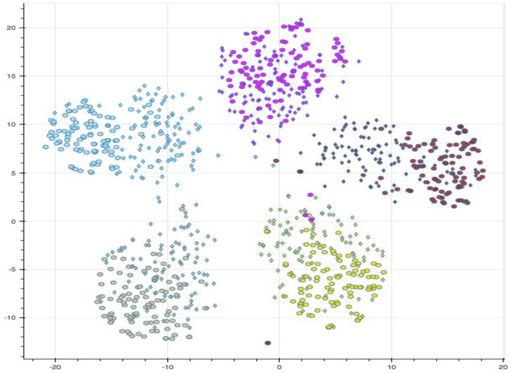
IV Other topics

Other related projects

Time series (IQ) data:

- Signal modulation classification
- GNUradio visualization and inference
- Adversarial domain adaptation





GPU algorithms of signal search

- e.g. Massively parallel narrowband search

```
__global__ void sweep(float *g_idata, float *g_odata,
const int *delay_table, const int nfreqs, const int ntimes, const int
ndelays) {
    int tx = threadIdx.x;  int ty = threadIdx.y;
    int bx = blockIdx.x;   int by = blockIdx.y;
    int bdx = blockDim.x;  int bdy = blockDim.y;
    int i = bdx * bx + tx; int j = bdy * by + ty;
    int p = INDEX(j,i,nfreqs); //j is delays, i is freqs

    int delay;

    __syncthreads();
    // each core computes one output pixel
    for ( int t=0; t<ntimes; t++) {
        delay = delay_table[INDEX(t,j,ndelays)];
        if (delay+i >= 0 && delay+i < nfreqs){
            g_odata[p] += g_idata[t*nfreqs + i + delay];
        }
    }
}
```

Conclusion

- Radio SETI has challenges such as large data volume, and uncertain signal of interest.
- NVIDIA GPUs are indispensable for data reduction, parallel search algorithms, and deep learning based analysis.
- Large input, varying signal support and information sparsity motivates algorithm designs.
- Supervised classification works for detecting known signals (e.g. FRB).
- Clustering useful for characterizing unlabeled dataset.
- Deep representation learning core to wide range of SETI tasks.
- Predictive spatial filtering effective for sequential data.

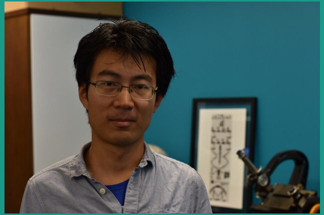
BERKELEY SETI
RESEARCH CENTER

David MacMahon
Chief Engineer

Dr. Vishal Gajjar
Postdoctoral Researcher:
Pulsar Astronomy

Dr. Andrew Siemion
Director

Howard Isaacson
Research Associate



Yunfan Gerry Zhang
Graduate Student:
Machine Learning and
Data Science

Dr. Steve Croft
Outreach Specialist

Emilio Enriquez
Graduate Student:
SETI astronomy

Matt Lebofsky
System Administrator and
Information Scientist

Thank you!

Contact:

yf.g.zhang@gmail.com
yunfanz@berkeley.edu

Image Sources:

[1]:H. Isaacson et. al. “The Breakthrough Listen Search for Intelligent Life: Target Selection of Nearby Stars and Galaxies”, ASP 2017.
[2]: J. E. Enriquez, et. al. “The Breakthrough Listen Search for Intelligent Life: 1.1-1.9 GHz Observations of 692 Nearby Stars,” ApJ 2017
[3] Lorimer D. et. al. “A bright millisecond radio burst of extragalactic origin” 2017
[4] Zhang Y.G. et. al. Fast Radio Burst 121102 Pulse Detection and Periodicity: A Machine Learning Approach, ApJ 2018
[5]:https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
[6] Schroff, F. et. al. FaceNet: A Unified Embedding for Face Recognition and Clustering, 2015
[7]: Zhang Y.G. et. al. “Self-supervised Anomaly Detection for Narrowband SETI”, IEEE GlobalSIP, 2018.