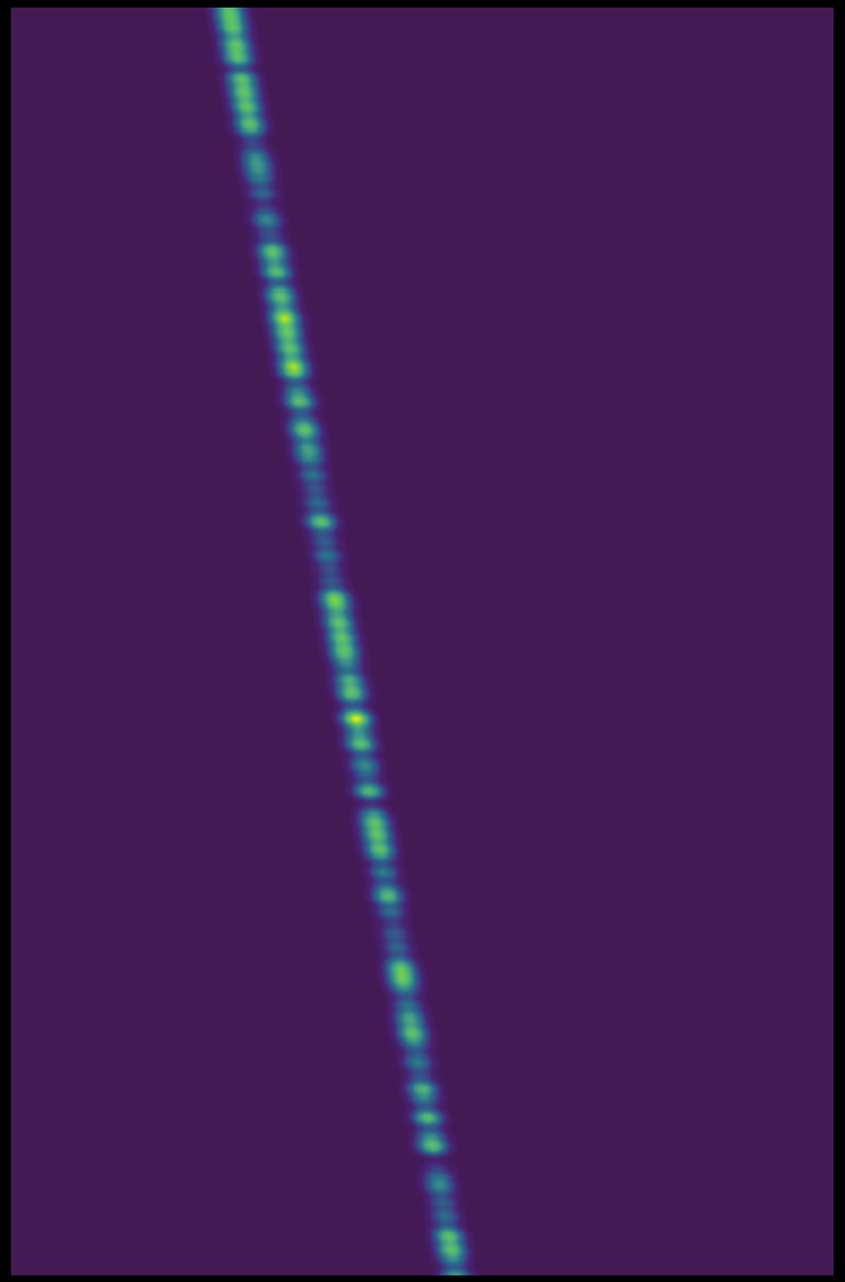


BREAKTHROUGH LISTEN

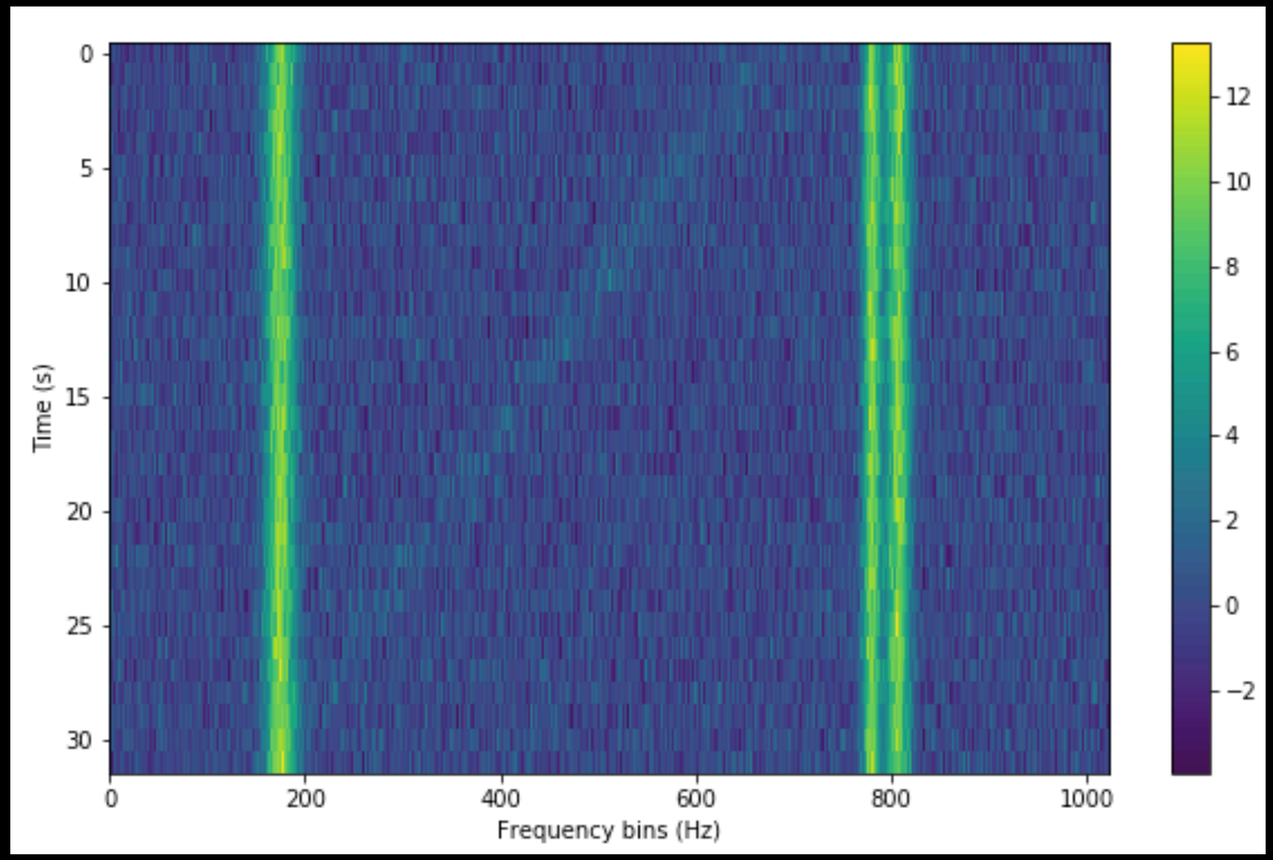
Searching for technosignatures with machine learning

Bryan Brzycki
University of California Berkeley
December 1, 2020

BERKELEY SETI RESEARCH CENTER

How can we build on existing searches?

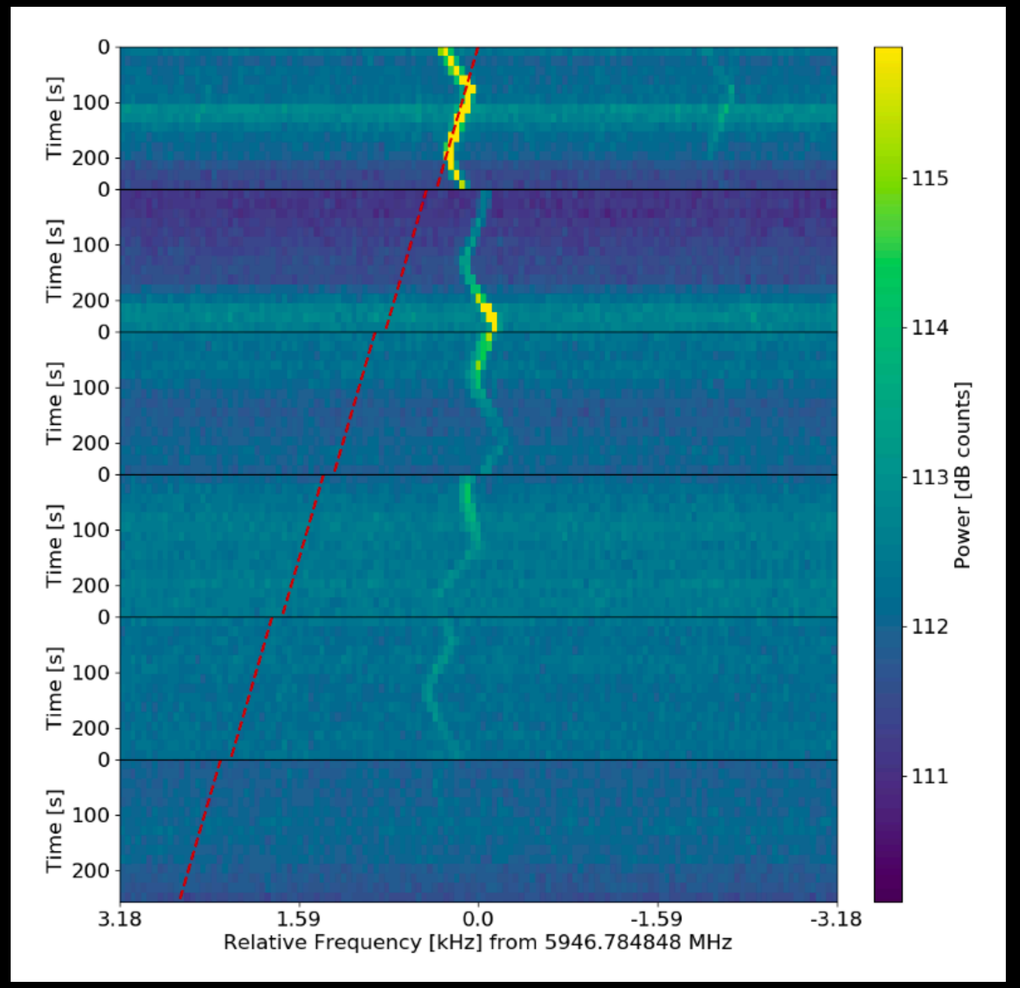
- TurboSETI
  - Struggles in RFI-rich bands
  - Can take a while as you increase range of drift rates to search



Synthetic example data with a dim signal

How can we build on existing searches?

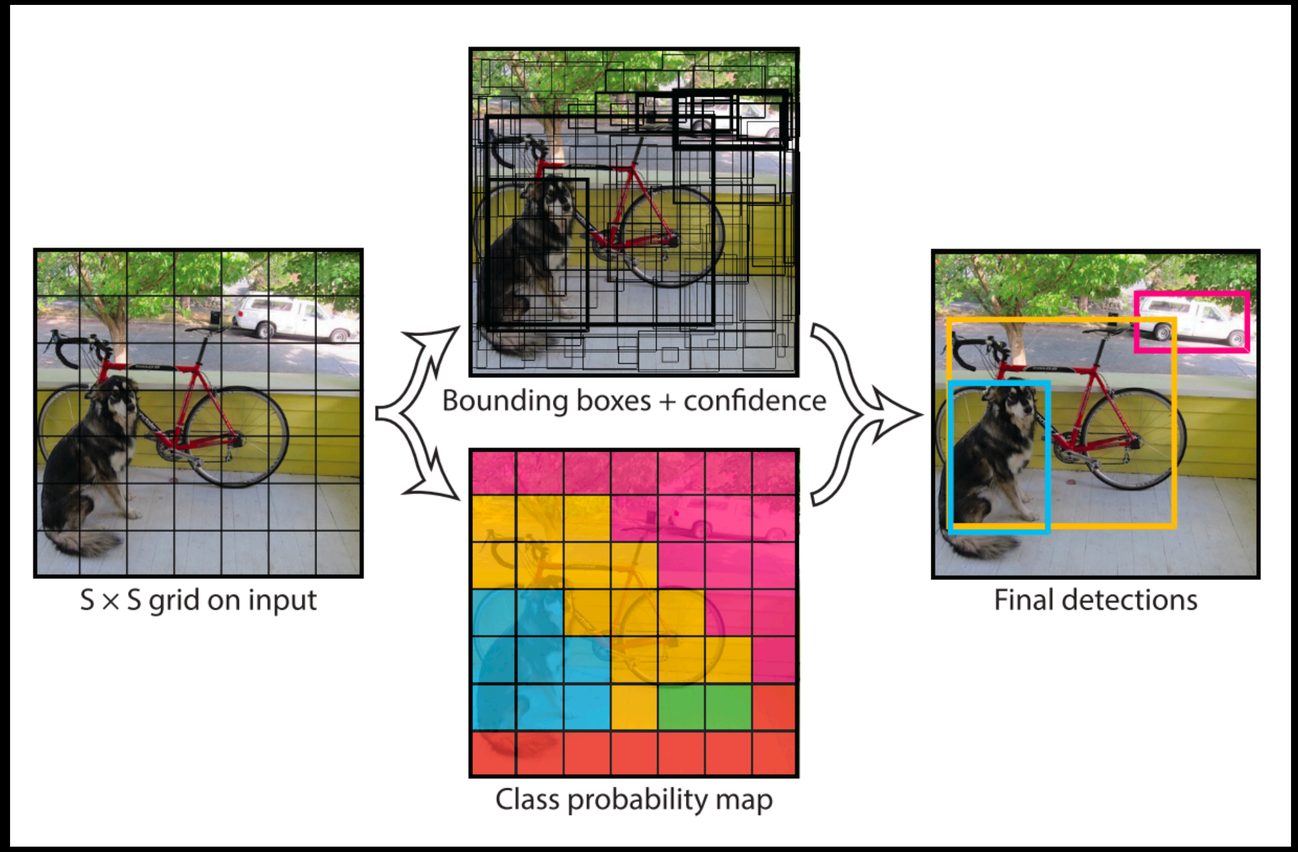
- Current pipelines are built to search for approximately linear signals – there are weird ones out there!
- How can we handle when they arise in on-off cadences?
- Is there a way to search for these directly?



Example on-off cadence (Sheikh et al. 2020)

Let's try machine learning!

- Major advances in machine learning with respect to image analysis
- Computer vision techniques are good at classifying images based on morphological features, and finding high-level objects
- So-called convolution neural networks (CNNs) are the name of the game



Redmon et al. 2016

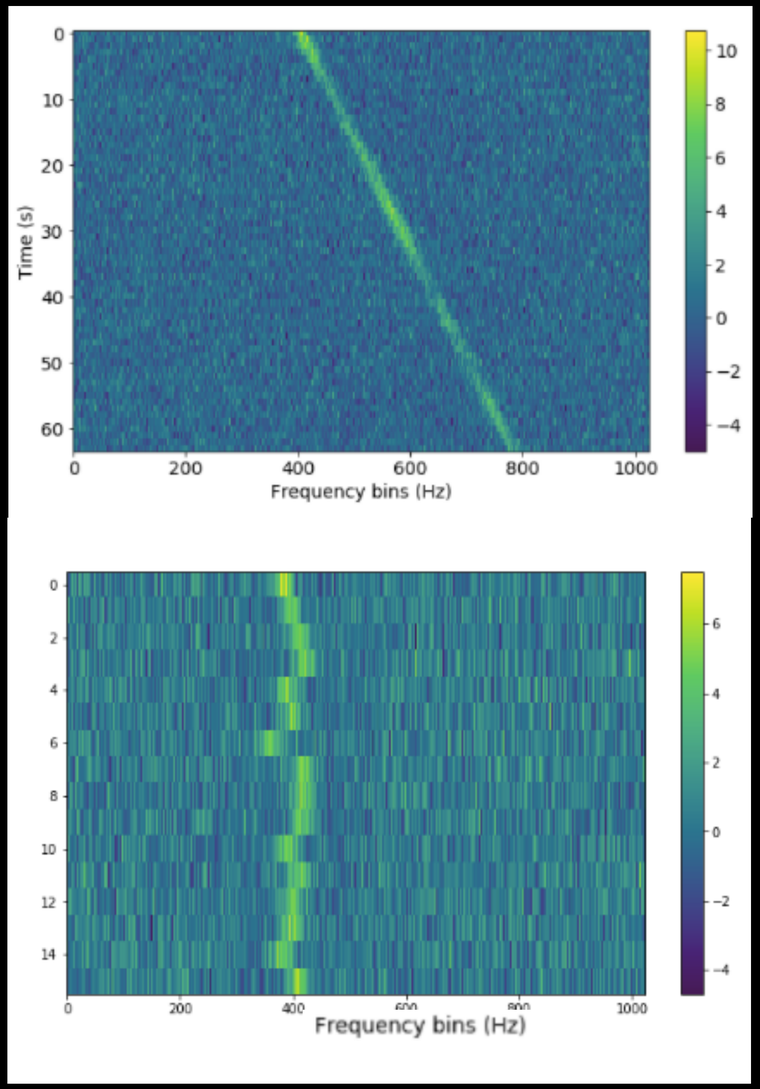
Approaching ML for narrow-band signals

- To train neural networks, we need to have labeled datasets of some sort
- Especially for SETI use cases, it's hard to put together such datasets only from observations

Approaching ML for narrow-band signals

- One solution is to turn to simulations!
- Created **setigen**, a Python module for making synthetic narrow-band signals, which can be directly inserted into observational data
- We've used setigen for ML experiments as well as injection recovery for signal search pipelines (like TurboSETI)!

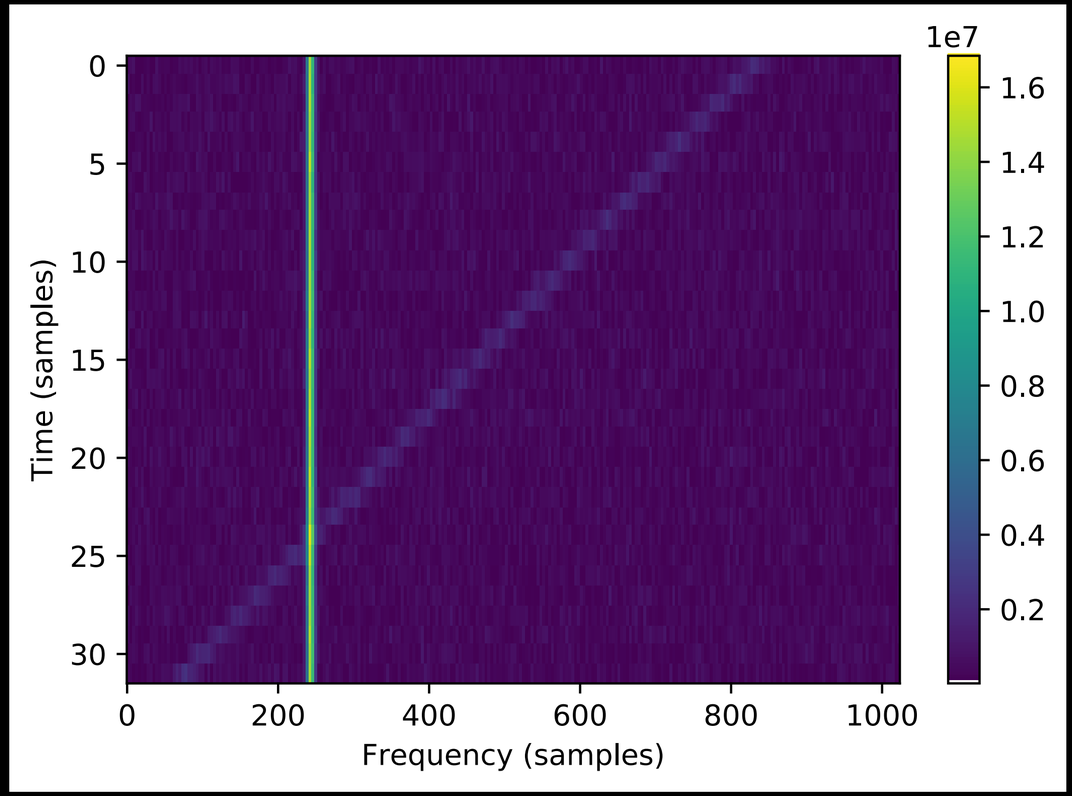
github.com/bbrzycki/setigen



Top: synthetic scintillating signal.
Bottom: synthetic RFI signal.

Signal localization with ML (Brzycki et al. 2020)

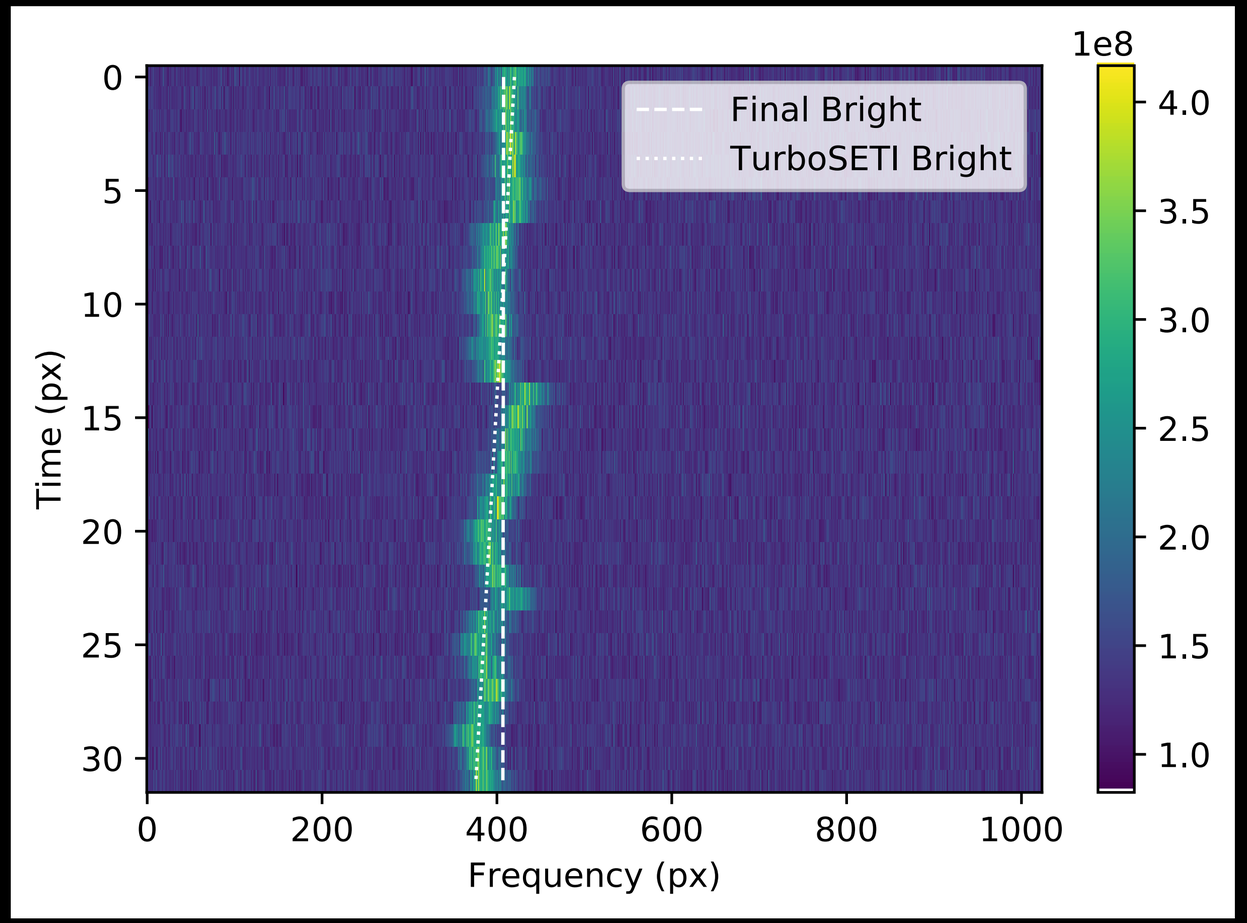
- TurboSETI struggles when multiple signals are within a certain frequency range
- We can generate lots of relevant labeled synthetic data, and train a CNN to find these signals
- Two main datasets:
  - One signal with random drift rate
  - Two signals, one with random drift rate, and the other with 0 drift rate (meant to simulate "bright" RFI)



Example of a frame with 2 synthetic signals, at 25 and 15 dB.

Overall takeways

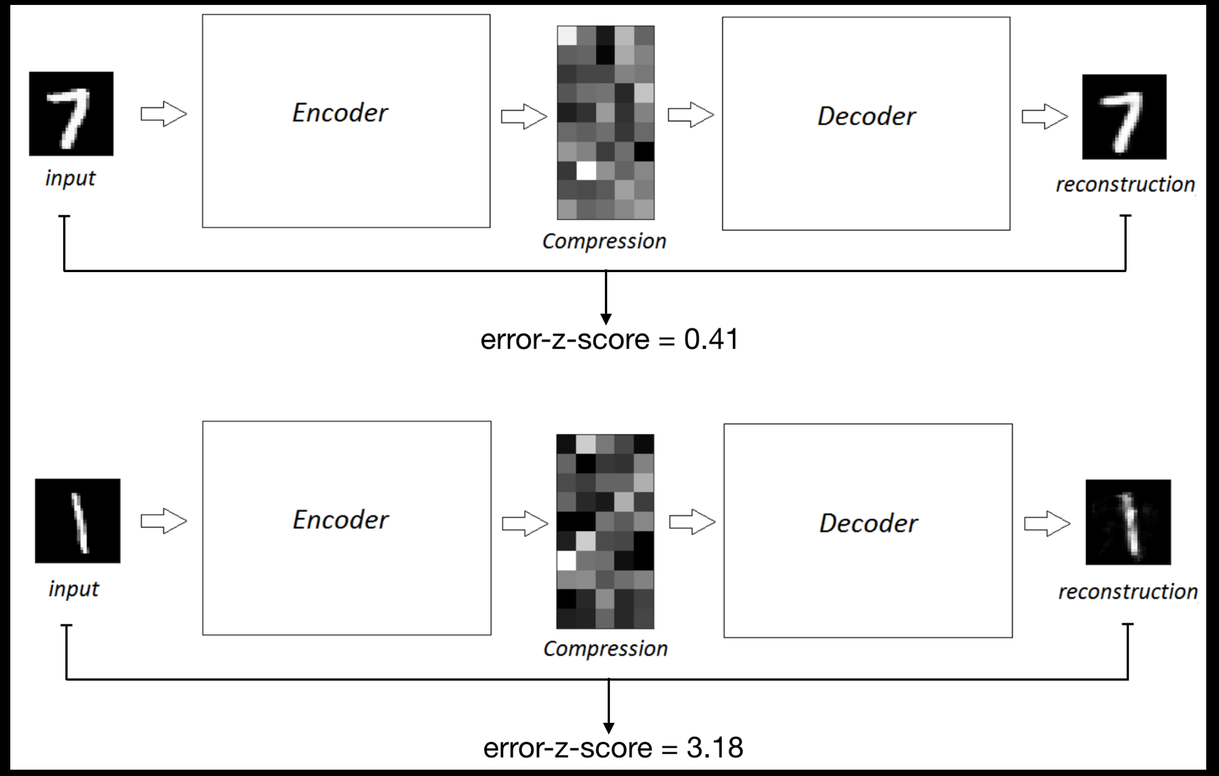
- ML predictions were generally less accurate than TurboSETI, but they were much faster (20-40x)
- Predictions were worse for the two signal dataset, but reasonable for SNR>10 (median errors in the 10s of pixels, out of 1024)
- Ran predictions on complex RFI signals in real observational data, and found that our ML models still obtained reasonable localizations



Observational data frame with real RFI signal, with ML prediction dashed and TurboSETI localization dotted.

Anomaly detection

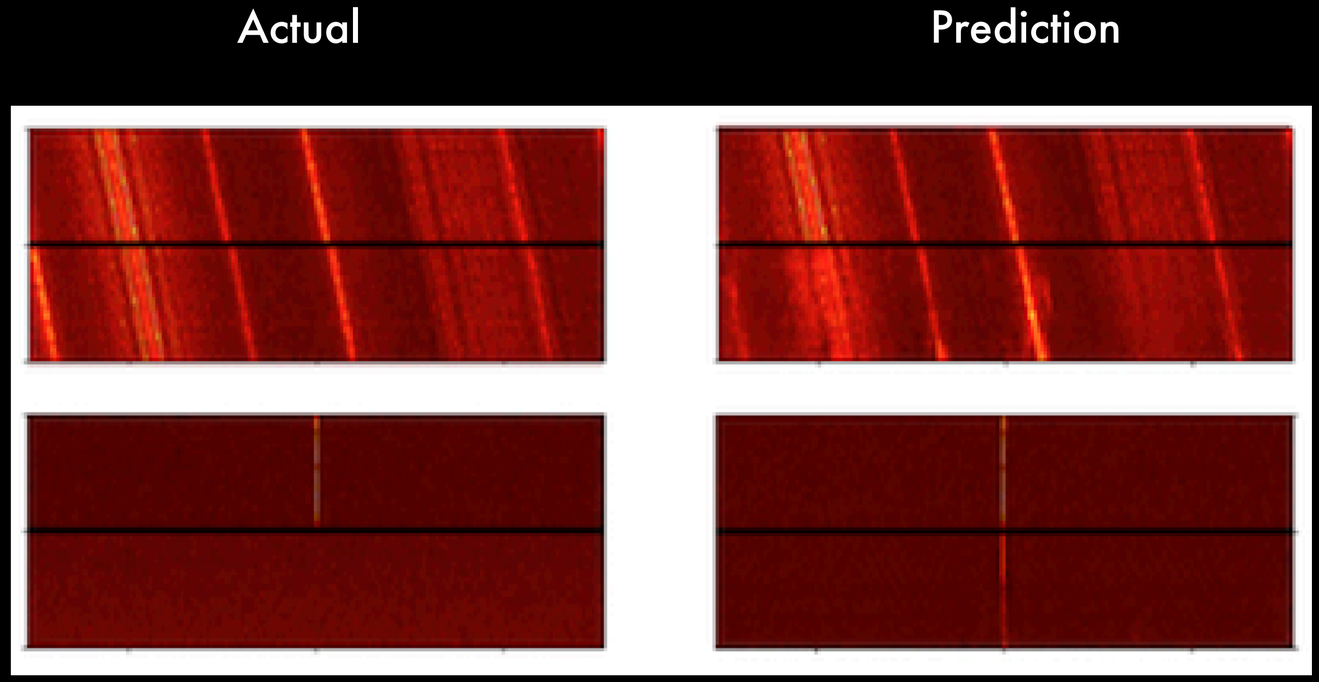
- Find snippets of data different enough from the rest
- Use CNN-based architecture called an *autoencoder*, based on compression and reconstructing input images
- If the model struggles to reconstruct an input, it's anomalous!



Autoencoder schematic (Google Ai-Hub)

Prediction-based anomaly detection (Zhang et al. 2019)

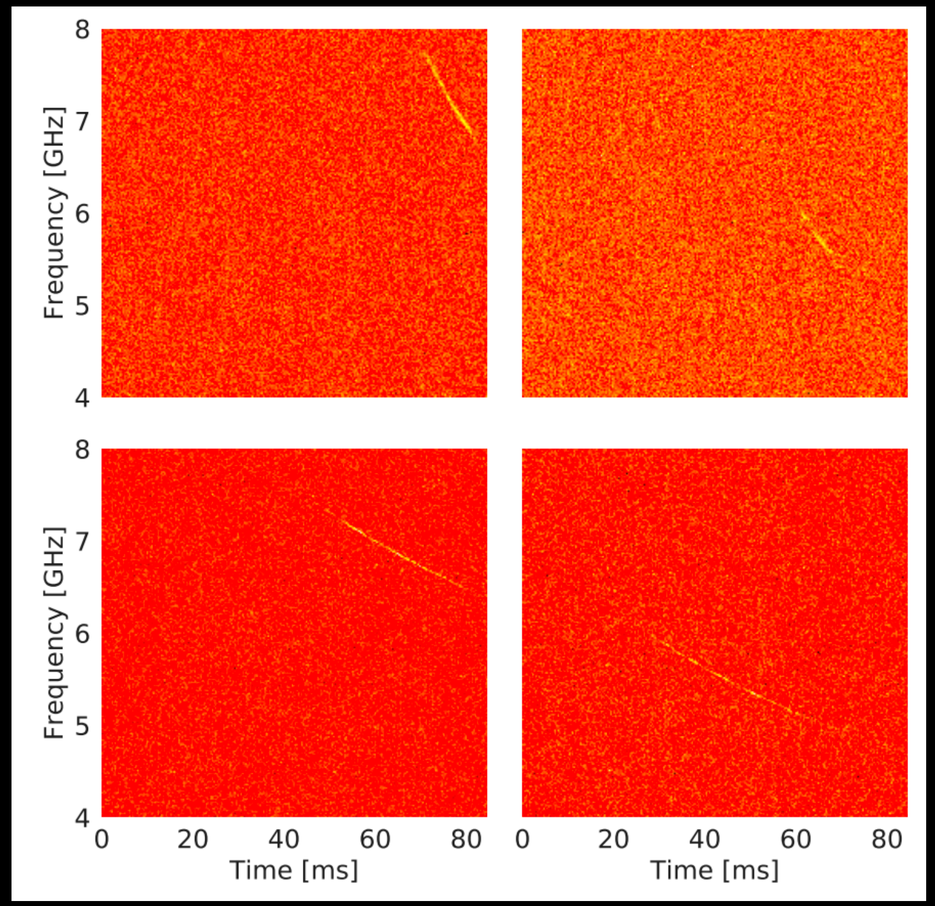
- Generative adversarial network (GAN), similar to an autoencoder
- Train to predict next time steps
- Compare predictions to reality to identify deviations



Prediction examples (Zhang et al. 2019)

Radio signal searches via supervised learning

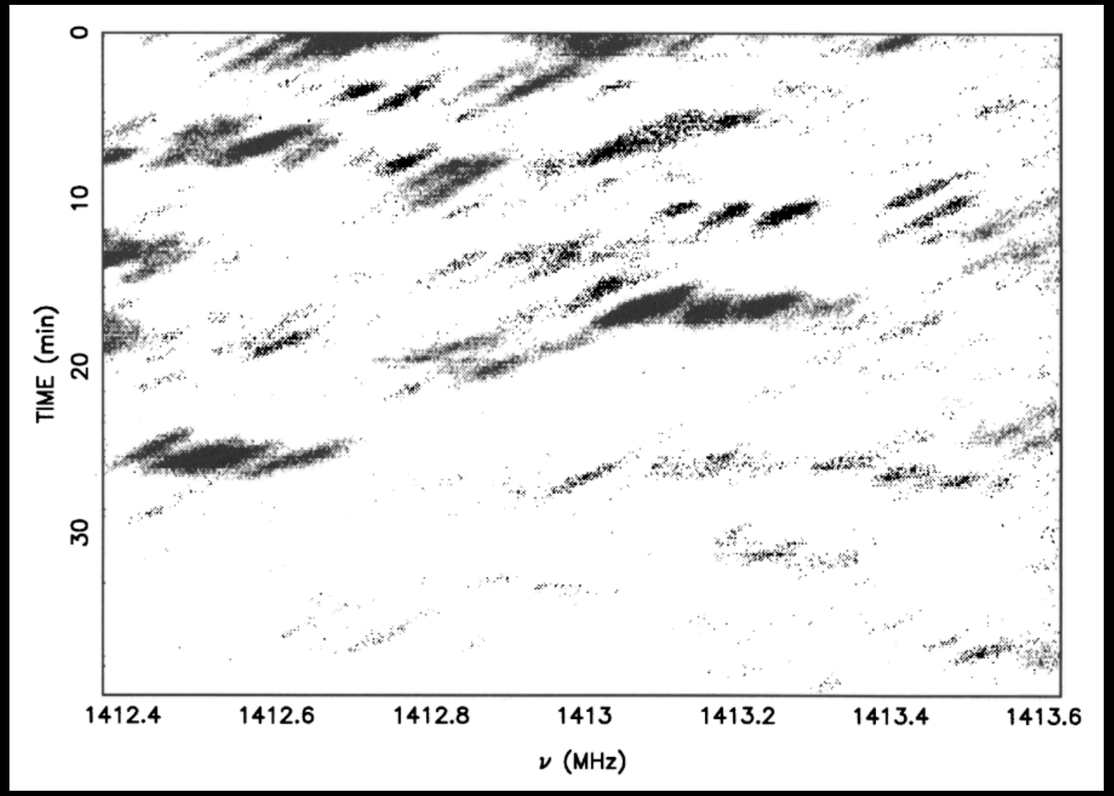
- FRB detection (Zhang et al. 2018)
- Simulate FRB pulses and inject in sample observations to create a labeled dataset
- Detected 72 new pulses from FRB 121102 using the ML pipeline



Synthetic dispersed pulses (Zhang et al. 2018)

SUPERVISED LEARNING - NEW DETECTION PROCEDURE?

- Ongoing research: use astrophysical effects like ISM scintillation and pulse broadening to develop a ML-based search strategy
- Especially relevant for galactic center surveys

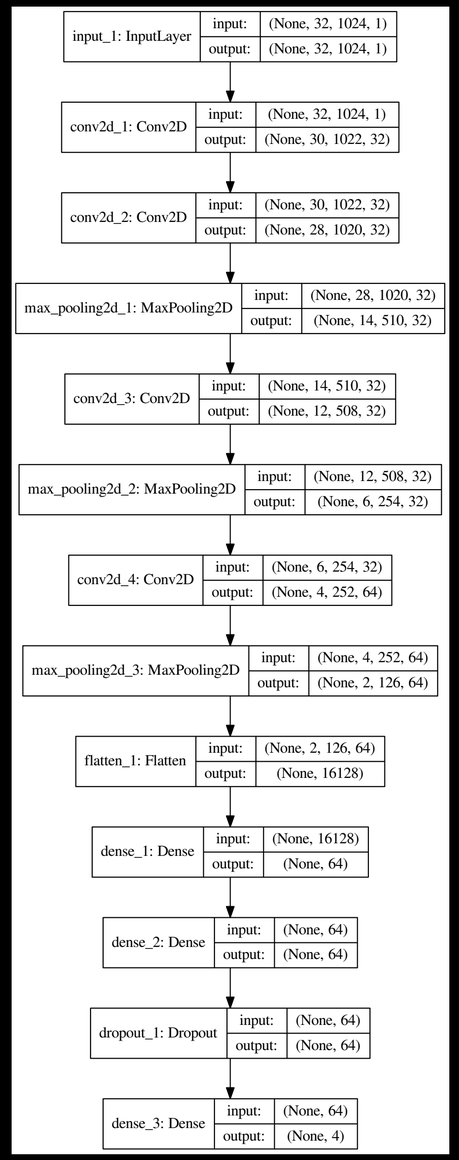


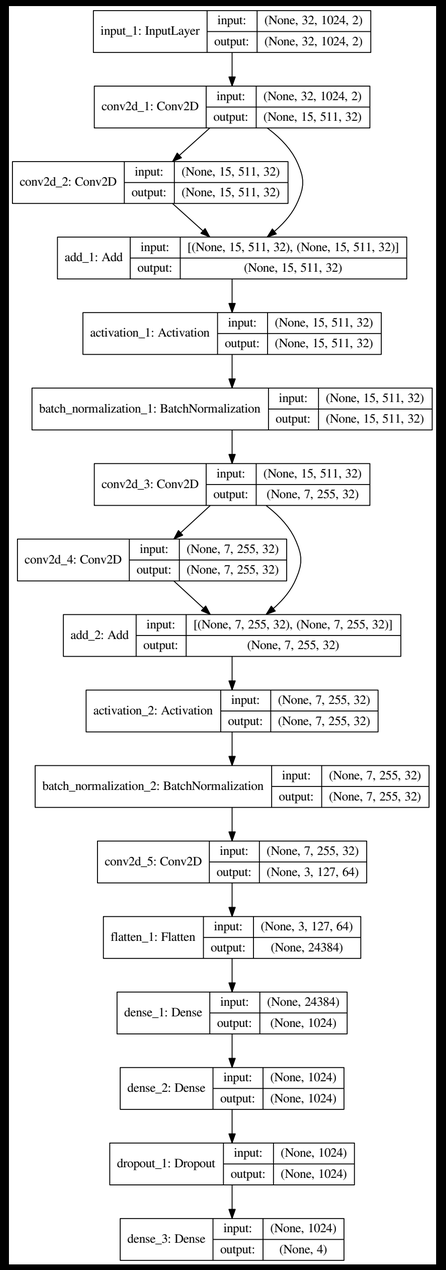
Dynamic spectra of pulsar PSR 1933+16 (Cordes & Lazio 1991)

Thank you!

Model architectures

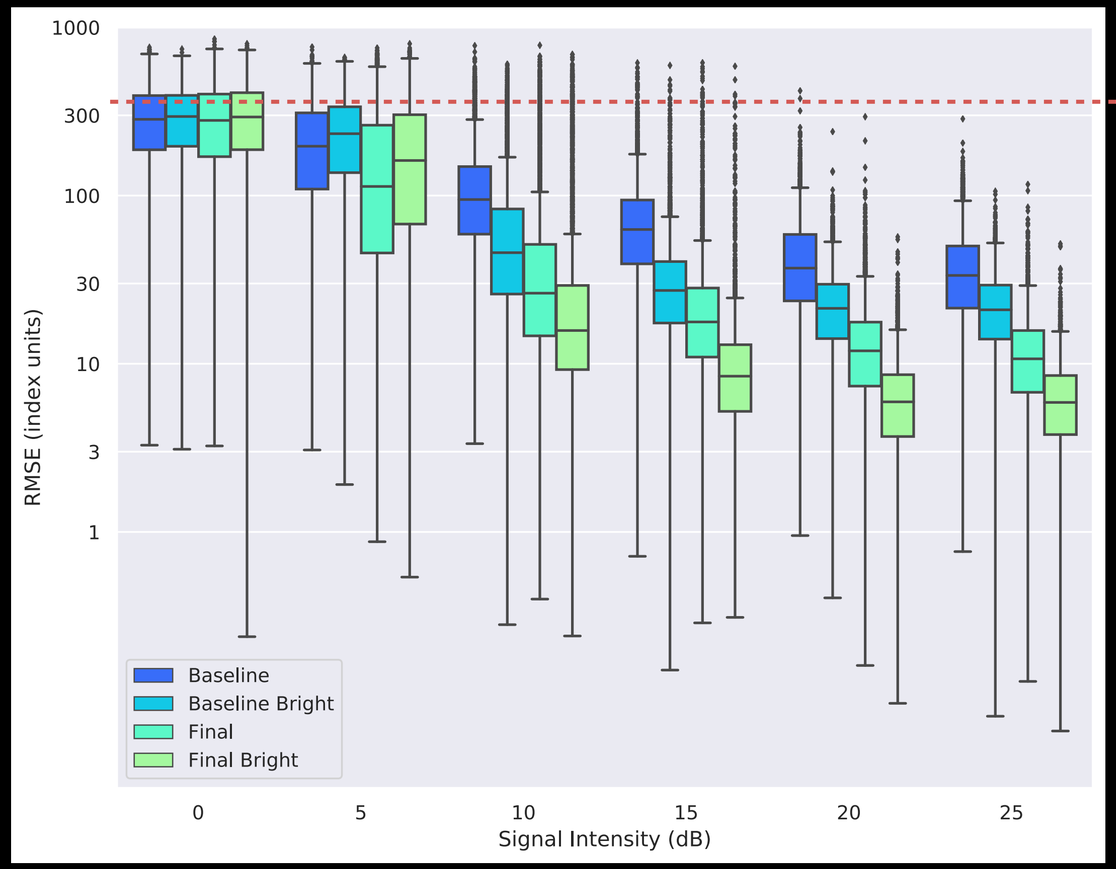
- Used convolutional neural networks, especially suited for image input data
- Created a “baseline” and a “final” model, to compare performance:
  - Baseline model uses convolutional layers, max pooling, and fully connected layers
  - Final model includes residual connections, stride 2 convolutions instead of max pooling, and batch normalization
- In addition to training these models over all input training data, we did alternate training over only 10 - 25 dB signal frames, labeling these as “bright” models



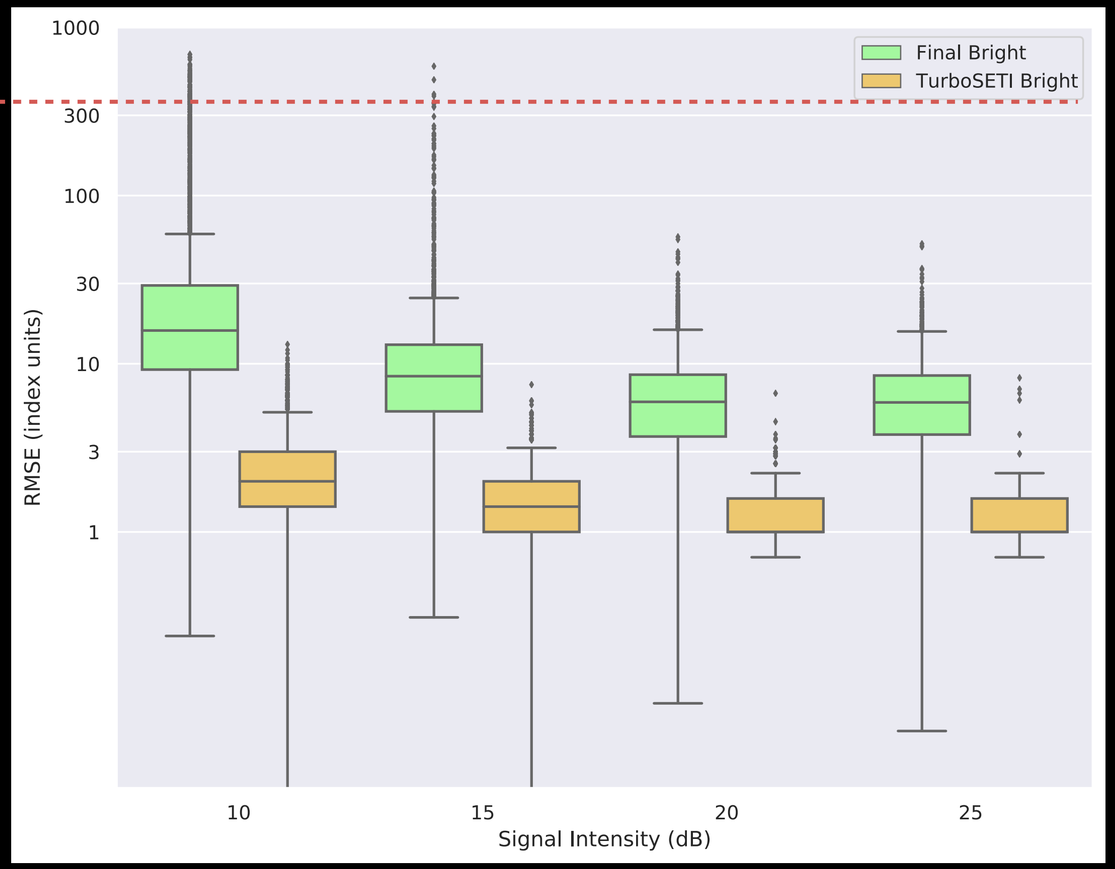


$$\text{RMSE (index units)} = 1024 \times \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2}$$

One signal results on test data

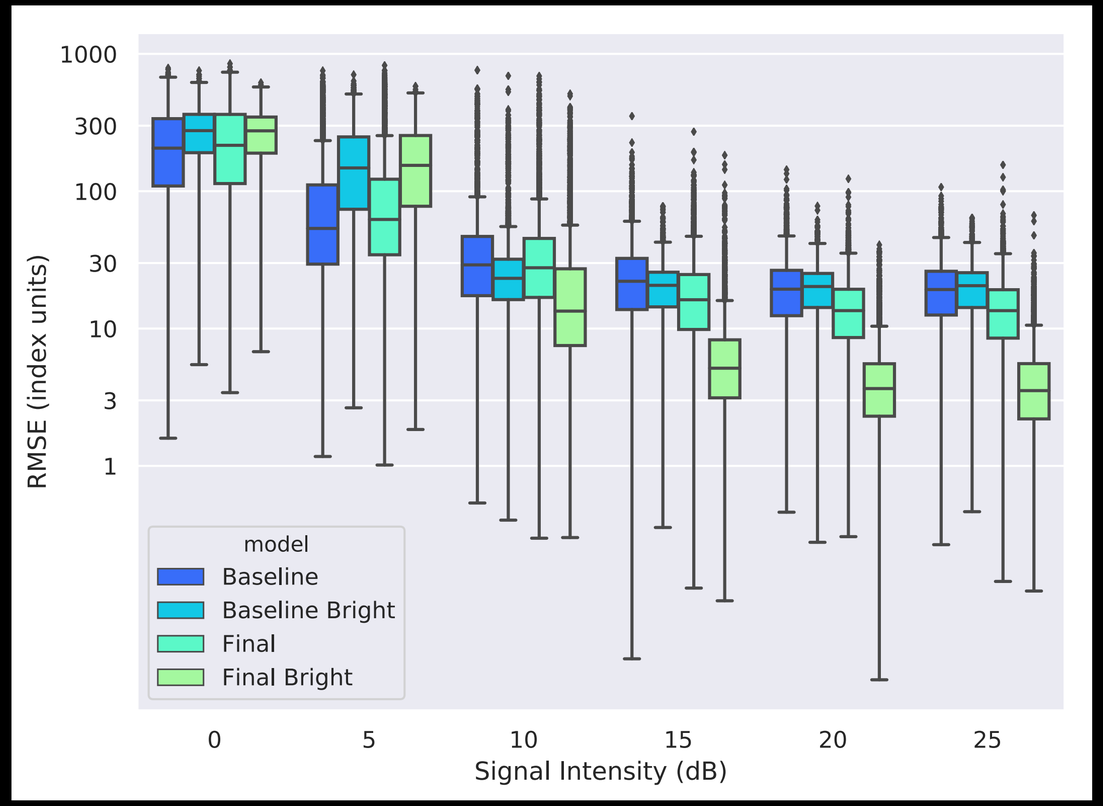


Root mean squared error across different signal intensities, in pixels, for various neural network architectures in the 1 signal case.

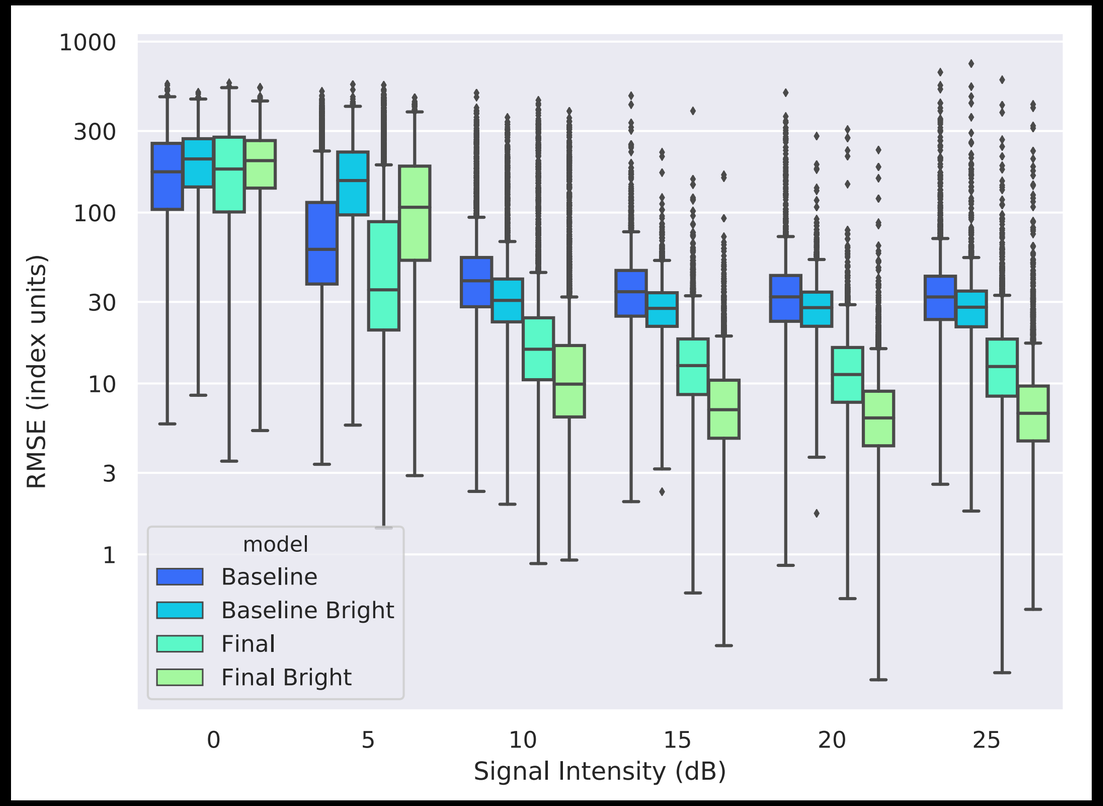


Root mean squared error across different signal intensities, in pixels, compared to TurboSETI performance. Only calculated for SNR > 10.

Signal Localization in Spectrograms



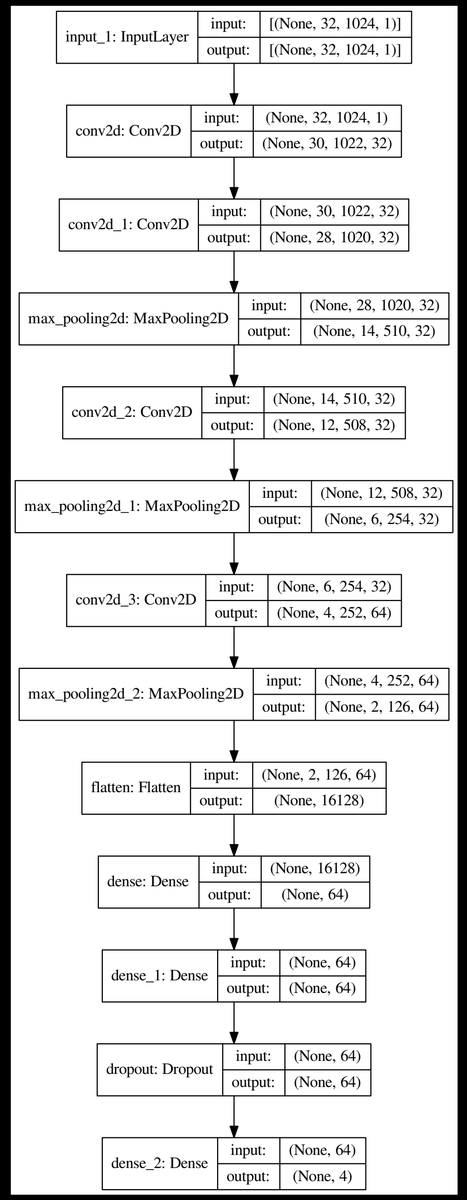
Mean squared error across different signal intensities, in pixels, for 1 signal case (Brzycki et al. 2020)



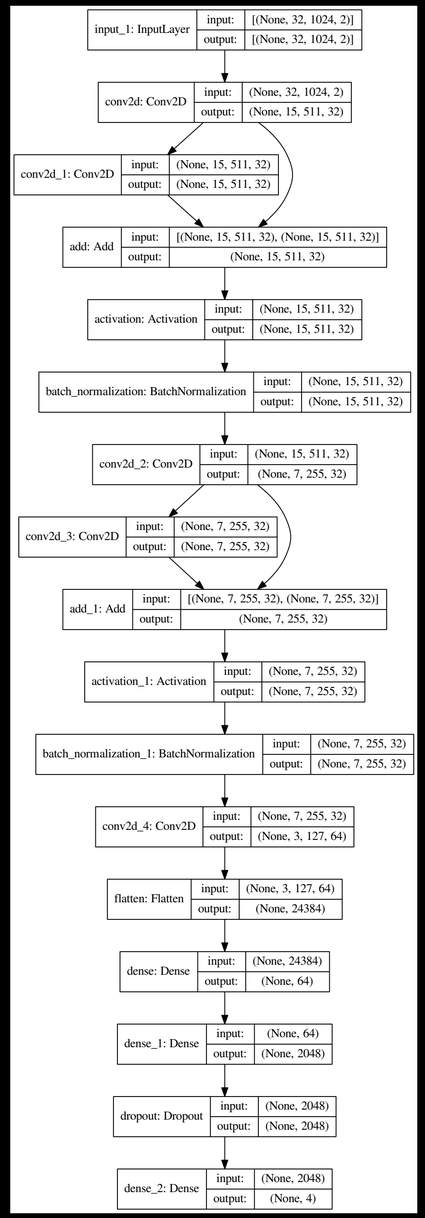
Mean squared error across different signal intensities, in pixels, for 2 signal case (Brzycki et al. 2020)

Localization Models

Baseline

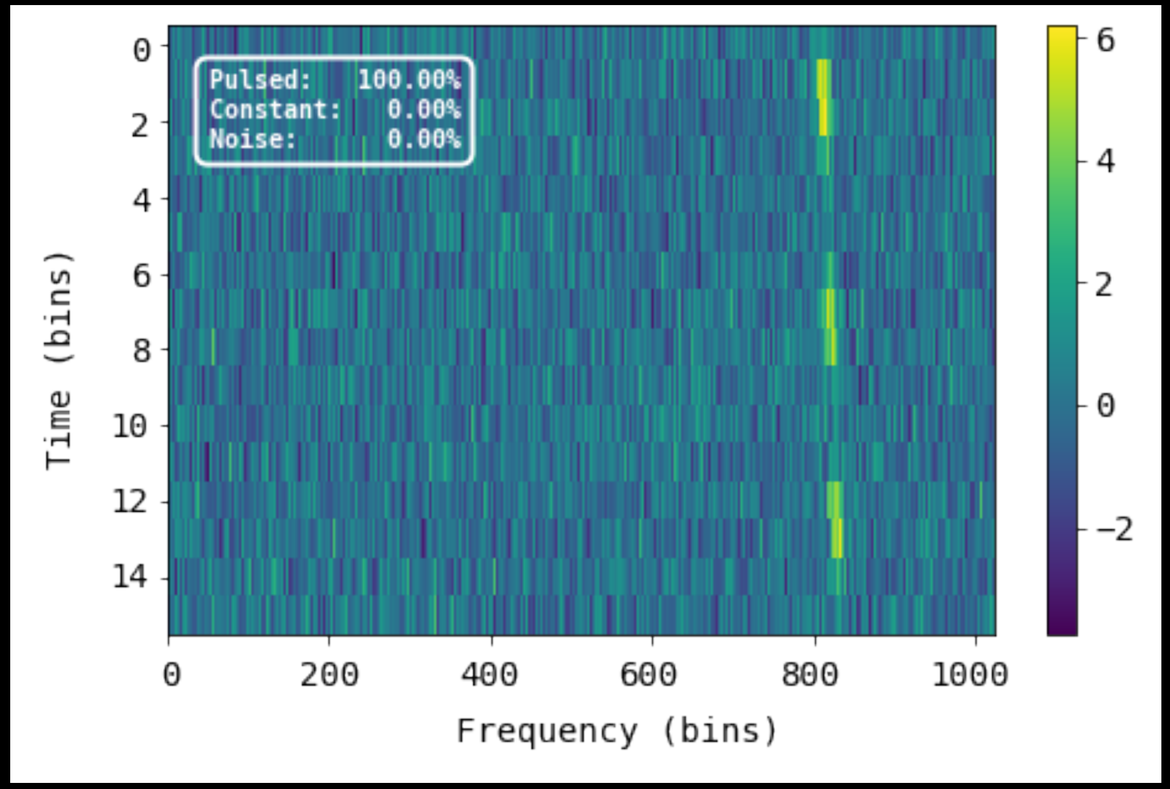


Final



Let's try machine learning!

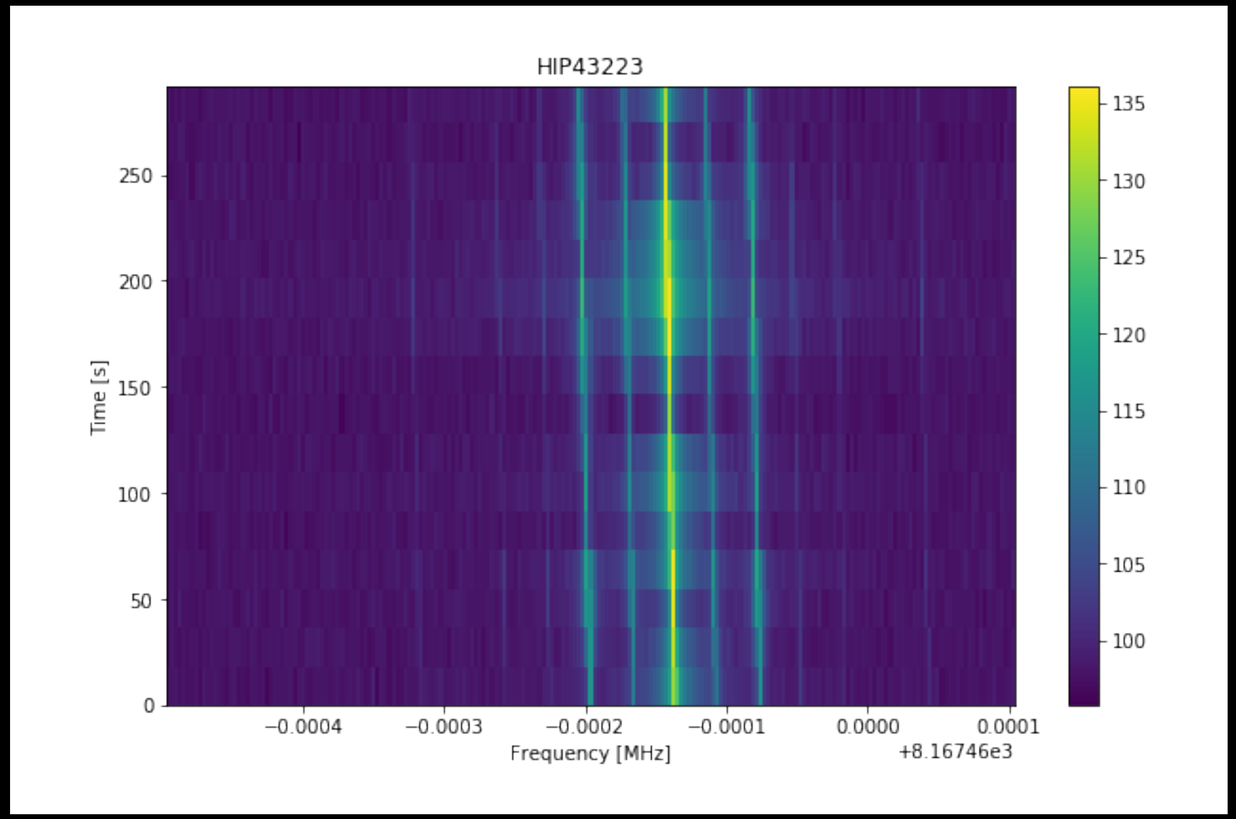
- Major advances in machine learning with respect to image analysis
- Computer vision techniques are good at classifying images based on morphological features, and finding high-level objects
- Potentially allow us to localize multiple signals in one shot, reducing computational costs



Simple example of ML classification (between noise, constant intensity, or pulsed) with a synthetic signal

Searching for ETI: radio technosignatures

- We can visualize BL radio data as waterfall plots (spectrograms), of intensity as a function of frequency and time
- Narrow-band signals generally appear as straight-line paths over time, can be sloped from Doppler acceleration
- Most of what we see is interference (RFI), but perhaps some of these signals are technosignatures!



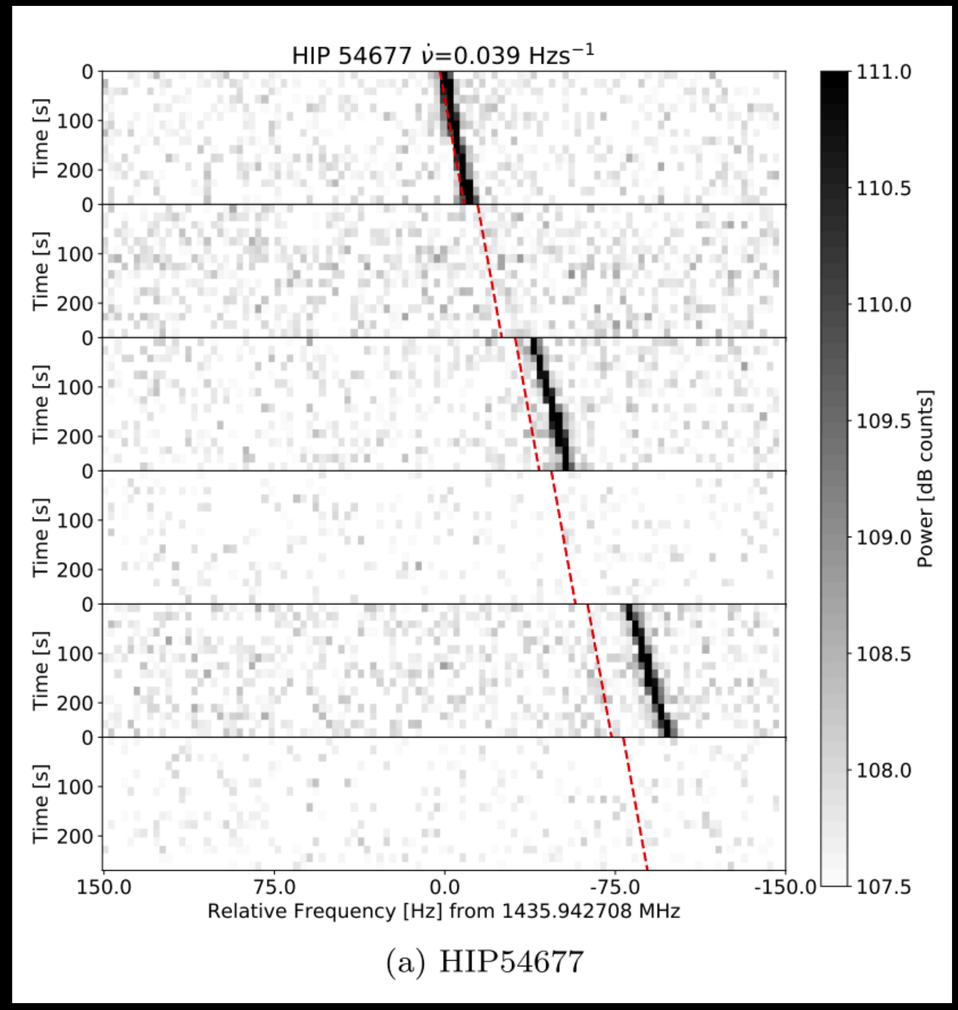
Real narrow-band signal

TurboSETI: standard dedoppler algorithm

- The standard signal search method at BL is TurboSETI, a tree deDoppler algorithm
- Searches frequencies and drift rates (slopes) by integrating over time and finding combinations that maximize SNR
- For each statistically significant signal, ultimately yields a position in starting frequency and Doppler drift rate
- Efficient; eliminates redundant calculations using trees

How to distinguish from human RFI?

- ABACAD / on-off observing cadences
- If a signal can be localized in the sky over the observing period, it's unlikely to be anthropogenic!



Example ABACAD cadence (Price et al. 2020)

Acknowledgements

- My advisor, Andrew Siemion
- The BSRC team
- Breakthrough Listen

Thank you!

- Cordes, J. M. & Lazio, T. J. 1991, ApJ
- Cordes, J. M. & Lazio, T. J., Sagan, C. 1997, ApJ
- Cordes, J. M. & Lazio, T. J. W. 2002, arXiv, astro-ph
- Zhang et al. 2018, ApJ, submitted