# Outsourcing astrophysics data analysis to the real experts

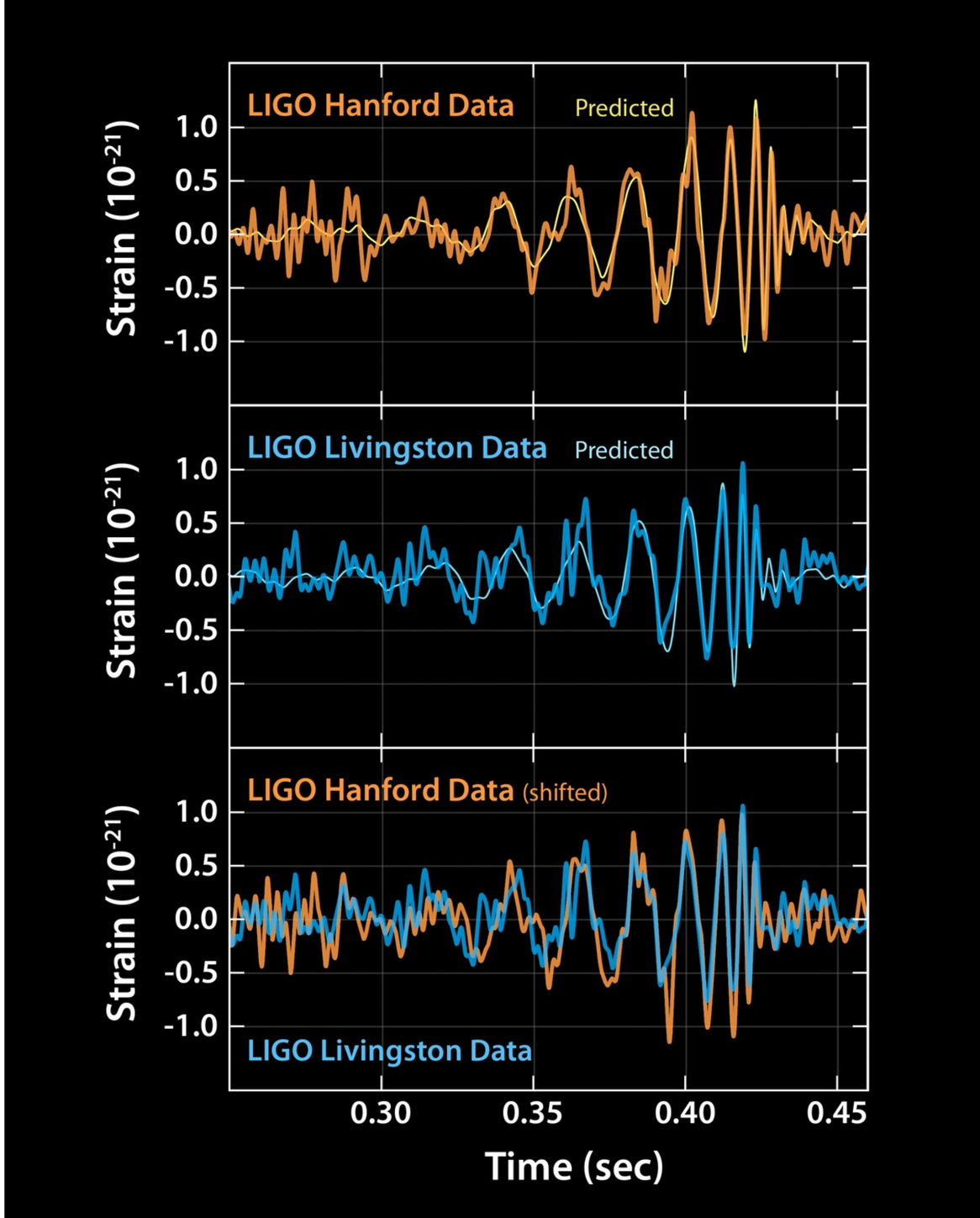
Joe Bayley - University of Glasgow
Chris Messenger - University of Glasgow
Michael Williams - University of Glasgow

G2Net meeting - 29th September 2022

LIGO-Virgo Collaboration, arXiv:1602.03837 (2016)

# Outline

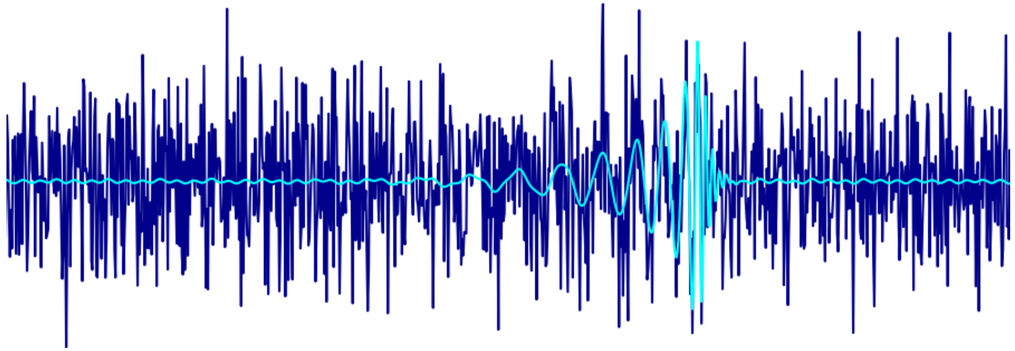
- Motivation
- Kaggle and our challenge
  - Kaggle, the challenge itself, the process, and the results
- Summary

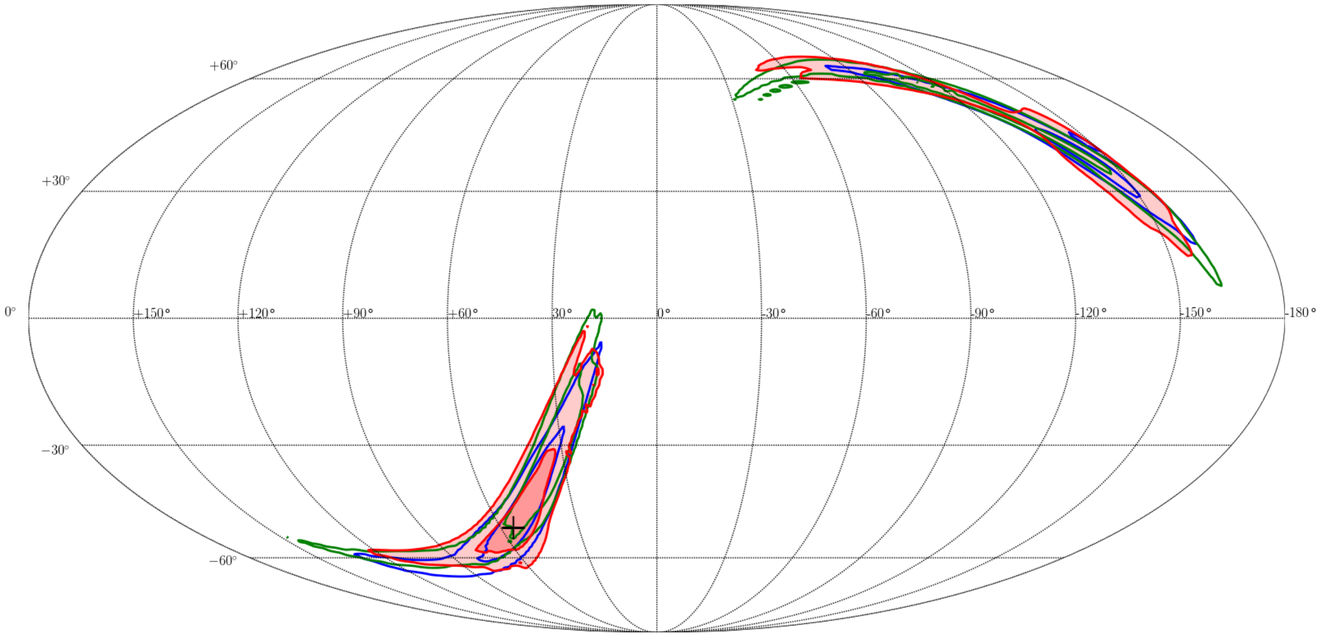


Gabbard *et al*, Nature (2021)

# Motivation

- The number of papers on ML applications to GW data has grown rapidly in recent years (see Cuoco, *et al* Mach. Learn.: Sci. Technol (2020) for a review)
- Most problems and ML tools have been attempted but there is still lots of room for improvement
- Some of the most recent and exciting work has been on rapid parameter estimation
- CBC searches were one of the early classification problems looked at
- However, there is still no serious CBC ML search pipeline
- Plus, we're mostly self-trained in ML so not "experts"





# Kaggle - what is it?

kaggle

Create

Home

Competitions

Datasets

Code

Discussions

Courses

More

Search

Sign In

Register

## Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the documentation or learn about Community Competitions.

Host a Competition

Search competitions

Filters

All competitions | Featured | Research | Getting Started | Playground | Analytics | Community

### Get Started

See all

**New to Kaggle?**

These competitions are perfect for newcomers.

**Titanic - Machine Learning from Disaster**

Start here! Predict survival on the Titanic ...

Getting Started

14143 Teams

**Knowledge** Ongoing

**House Prices - Advanced Regression Techniques**

Predict sales prices and practice feature ...

Getting Started

4247 Teams

**Knowledge** Ongoing

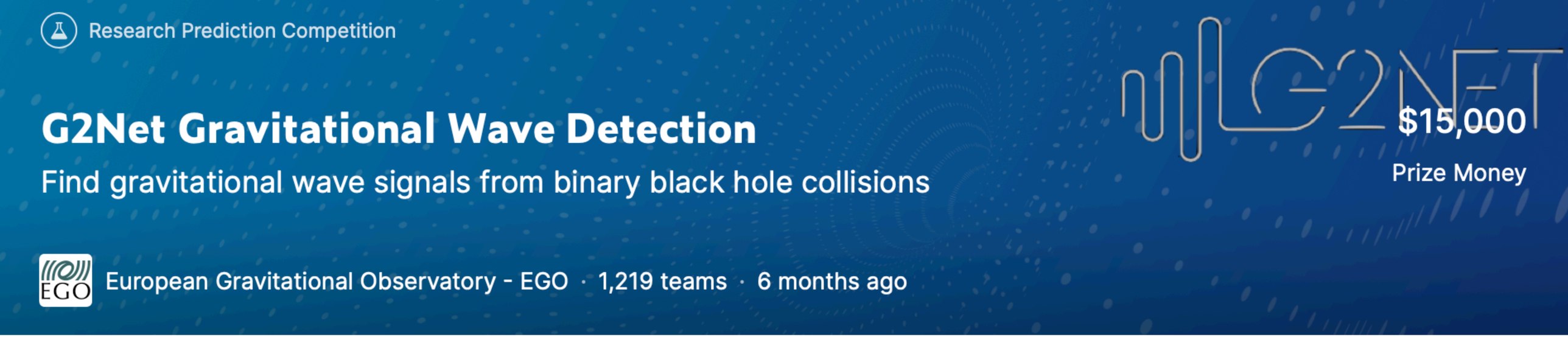
**Digit Recognizer**

Learn computer vision fundamentals with ...

Getting Started

2065 Teams

**Knowledge** Ongoing




# The challenge

https://www.g2net.eu

What did we ask people to do?

# Data description (for the competitors)

In this competition you are provided with a training set of time series data containing simulated gravitational wave measurements from a network of 3 gravitational wave interferometers (LIGO Hanford, LIGO Livingston, and Virgo). Each time series contains either detector noise or detector noise plus a simulated gravitational wave signal. The task is to identify when a signal is present in the data (`target=1`).

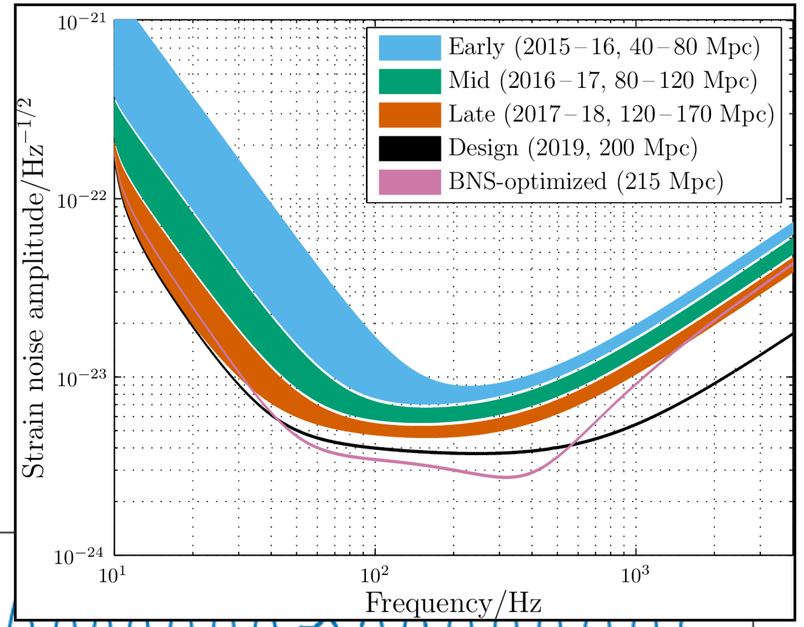
The parameters that determine the exact form of a binary black hole waveform are the masses, sky location, distance, black hole spins, binary orientation angle, gravitational wave polarisation, time of arrival, and phase at coalescence (merger). These parameters (15 in total) have been randomised according to astrophysically motivated prior distributions and used to generate the simulated signals present in the data, but are not provided as part of the competition data.

Each data sample (`npy` file) contains 3 time series (1 for each detector) and each spans 2 sec and is sampled at 2,048 Hz.

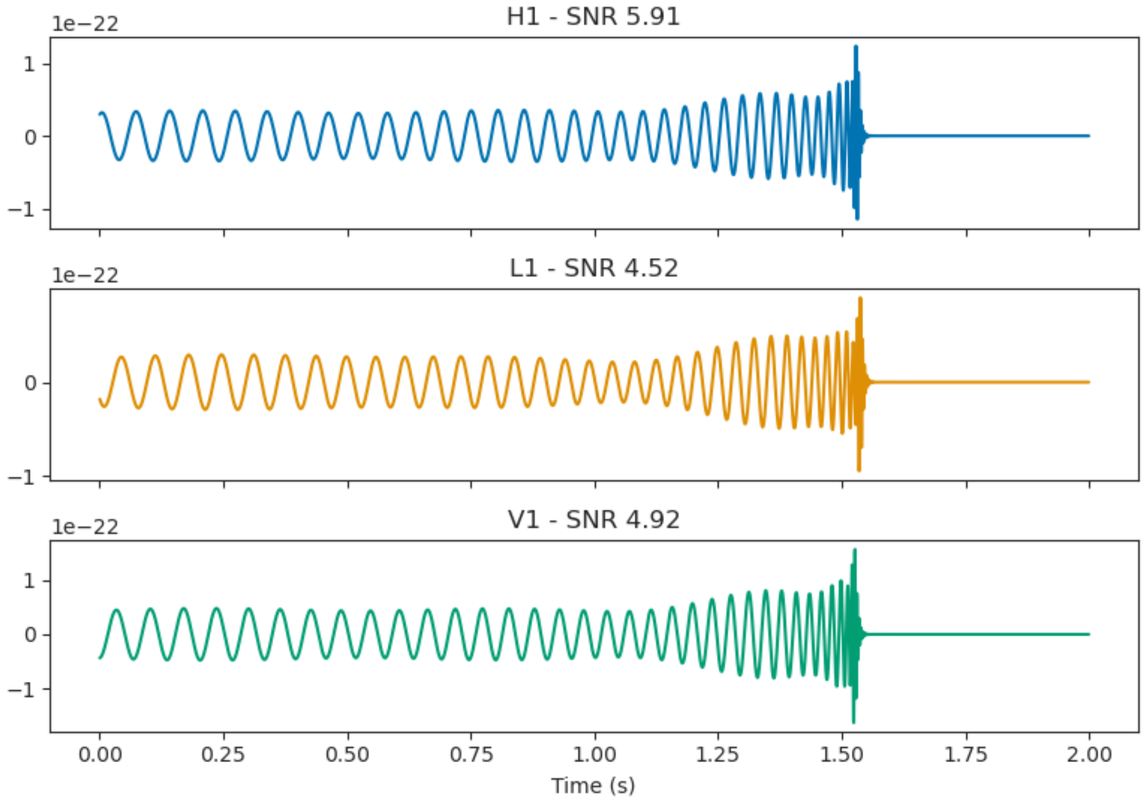
The integrated signal-to noise ratio (SNR) is classically the most informative measure of how detectable a signal is and a typical level of detectability is when this integrated SNR exceeds ~8. This shouldn't confused with the instantaneous SNR - the factor by which the signal rises above the noise - and in nearly all cases the (unlike the first gravitational wave detection GW150914) these signals are not visible by eye in the time series.

LIGO-Virgo Collaboration, Living Rev Relativity (2018)

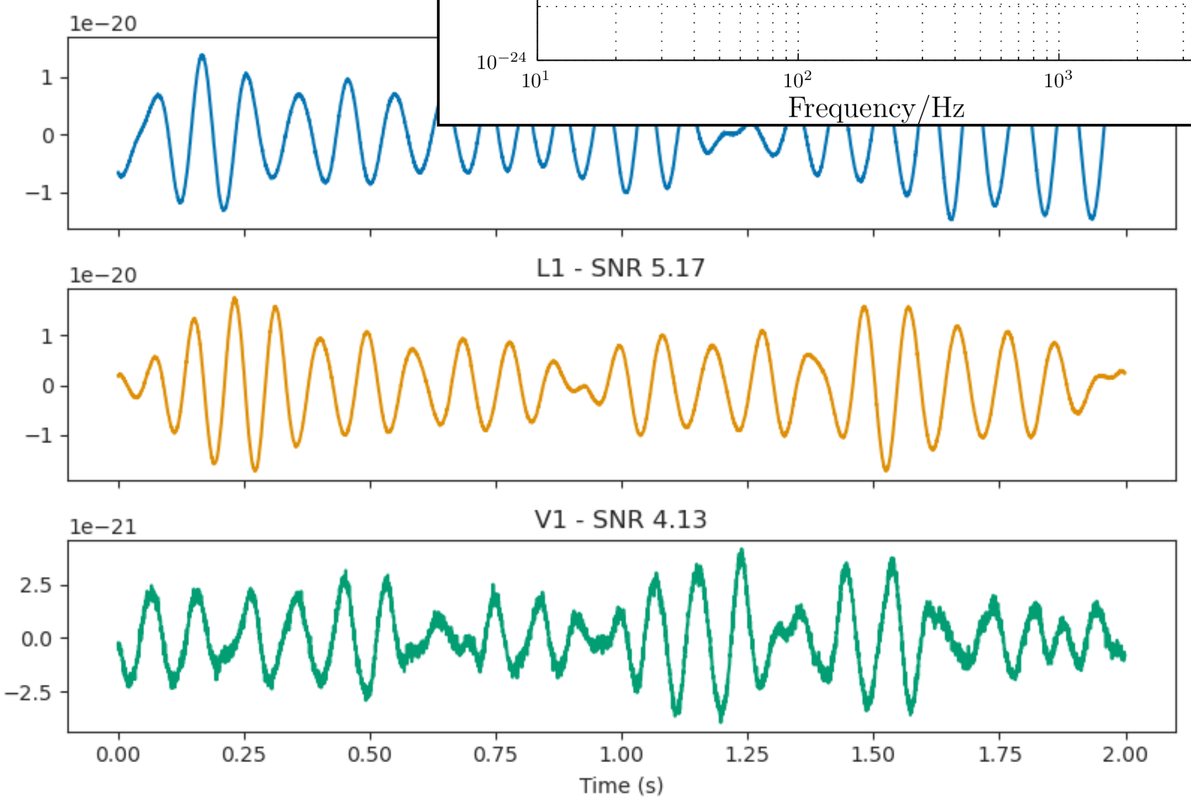
# The data




## signal

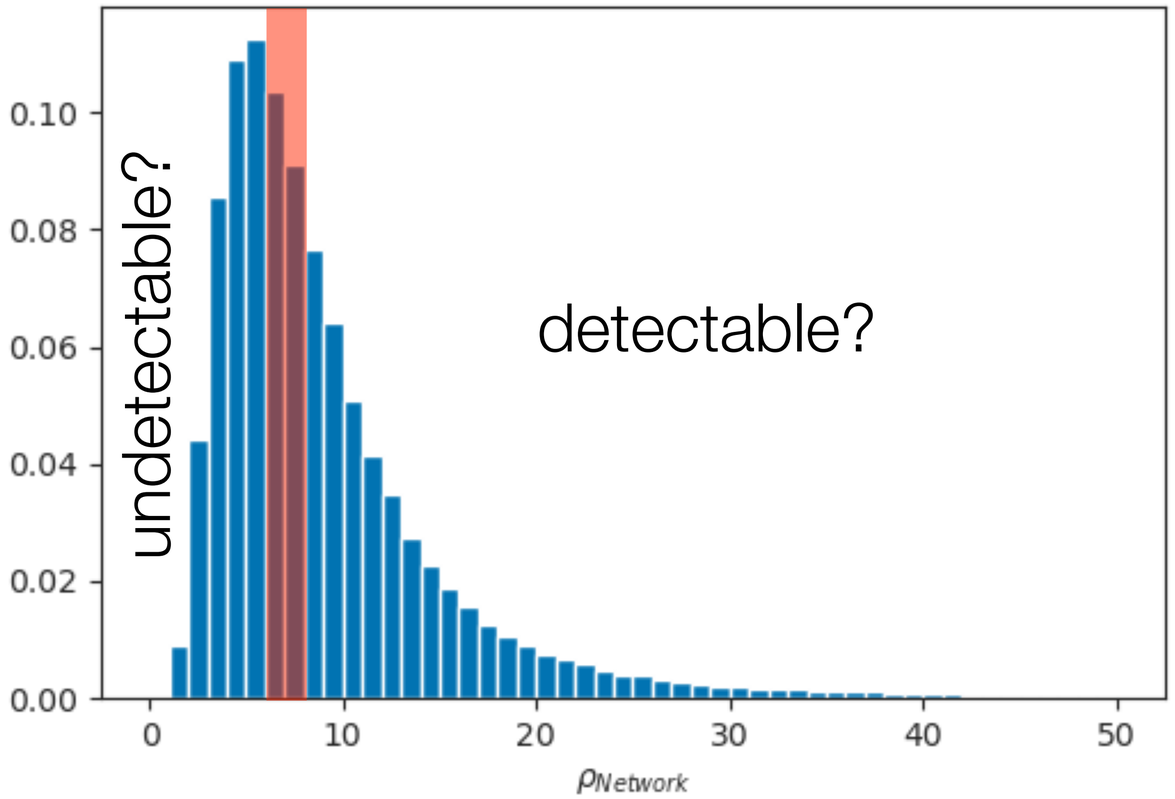



## signal+noise



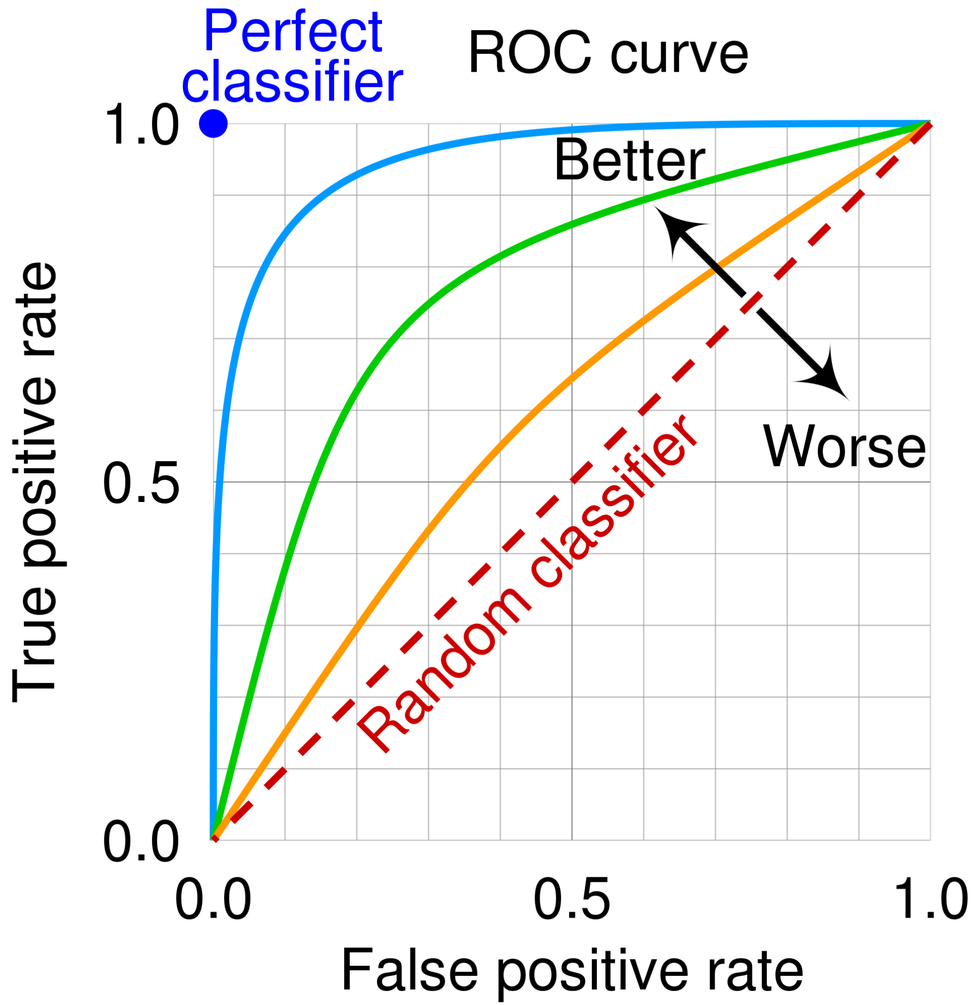
# Data description (hidden from competitors)

- Detector noise was generated using the expected O4 Advanced interferometer Power Spectral Densities.
- Signal parameters were sampled from standard astrophysical distributions.
- However, the SNR distribution was tuned by limiting the redshift to z=0.5.
- This gave us the ability to set the difficulty of the challenge.
- Reverse engineering of the problem is a concern - Kagglers are sneaky and smart, so great care had to be taken



# Data practicalities

- Training data consisted of 500K items (~55GB)
- Testing data consisted of 220K items (~25GB)
- Defining a metric - how do we decide who wins?
  - Standard practice in GW astronomy is to define a sensible False Positive Rate and try to maximise the True Positive Rate
  - Closest metric available within Kaggle was the Area Under the Curve (AUC)
  - Leaves the potential for analyses with best sensitivity at very low FPR to lose.



# Who's involved

Maggie Demkin
Customer Success

Walter Reade
Data Scientist

Chris Zerafa
Elena Cuoco

Michael Williams
Phd Student

https://www.g2net.eu

### Novice

You've joined the community.

- [x] Register!

### Contributor

You've completed your profile, engaged with the community, and fully explored Kaggle's platform.

- [x] Run 1 notebook or script
- [x] Make 1 competition submission
- [x] Make 1 comment
- [x] Give 1 upvote

### Expert

You've completed a significant body of work on Kaggle in one or more categories of expertise. Once you've reached the expert tier for a category, you will be entered into the site wide Kaggle Ranking for that category.

| Competitions | Datasets | Notebooks | Discussions |
|---|---|---|---|
| ☐ 2 bronze medals | ☐ 3 bronze medals | ☐ 5 bronze medals | ☐ 50 bronze medals |

### Master

You've demonstrated excellence in one or more categories of expertise on Kaggle to reach this prestigious tier. Masters in the Competitions category are eligible for exclusive Master-Only competitions.

| Competitions | Datasets | Notebooks | Discussions |
|---|---|---|---|
| ☐ 1 gold medal | ☐ 1 gold medal | ☐ 10 silver medals | ☐ 50 silver medals |
| ☐ 2 silver medals | ☐ 4 silver medals | | ☐ 200 medals in total |

### Grandmaster

You've consistently demonstrated outstanding performance in one or more categories of expertise on Kaggle to reach this pinnacle tier. You're the best of the best.

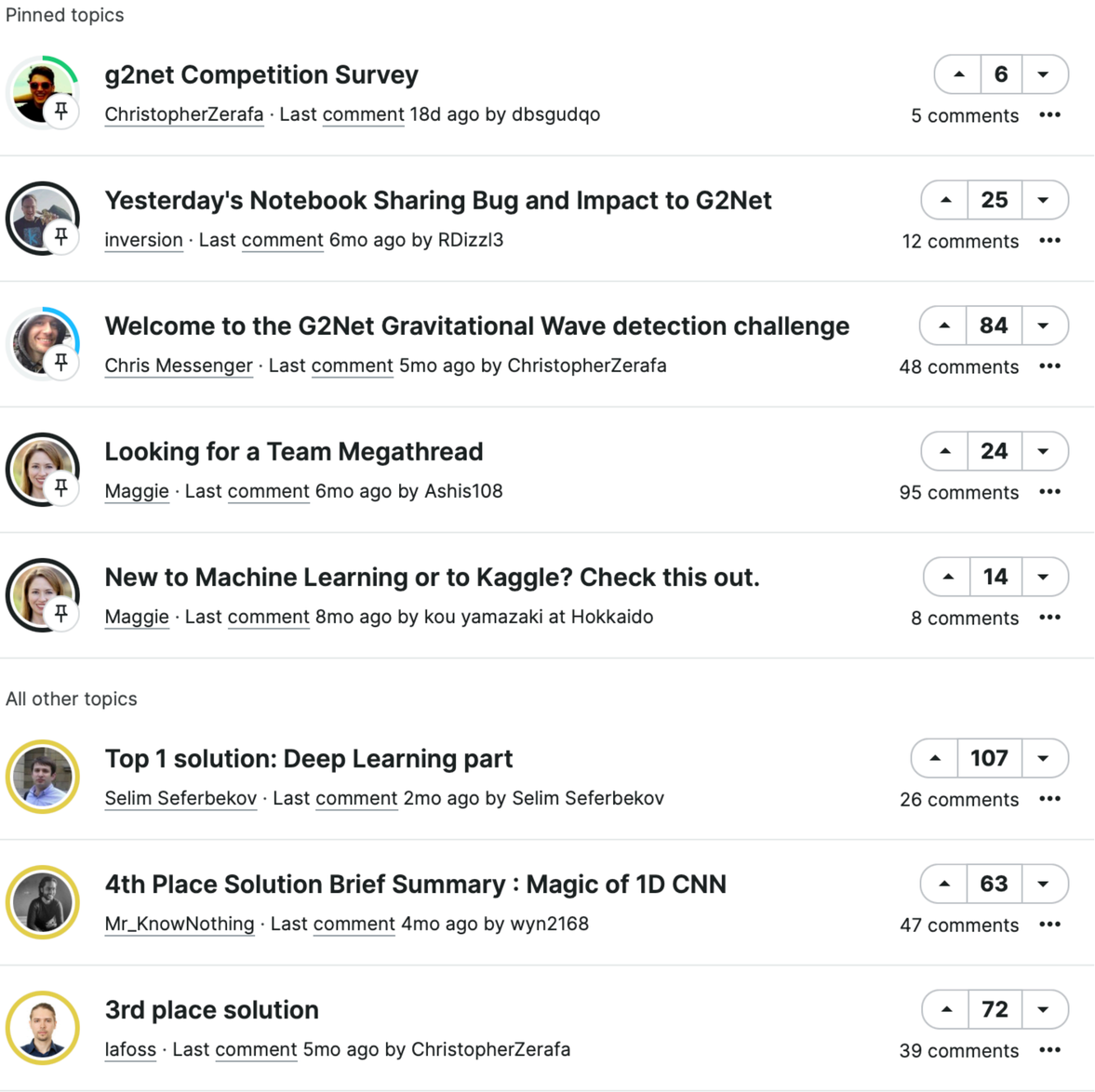
| Competitions | Datasets | Notebooks | Discussions |
|---|---|---|---|
| ☐ 5 gold medals | ☐ 5 gold medals | ☐ 15 gold medals | ☐ 50 gold medals |
| ☐ Solo gold medal | ☐ 5 silver medals | | ☐ 500 medals in total |

# The timeline

- **July 2020**: First contact with Kaggle
  - Meetings held every ~month
- **October 2020**: Initial trial dataset generated and sent to Kaggle
- **January 2021**: Started drafting the documentation
- **Early 2021**: Found out about Google prize money
- **April 2021**: Decided on the competition metric
- **28th April 2021**: Contract signed
- **30th June 2021**: The competition launched
  - The competition was live for 3 months
- **30th September 2021**: Competition ended

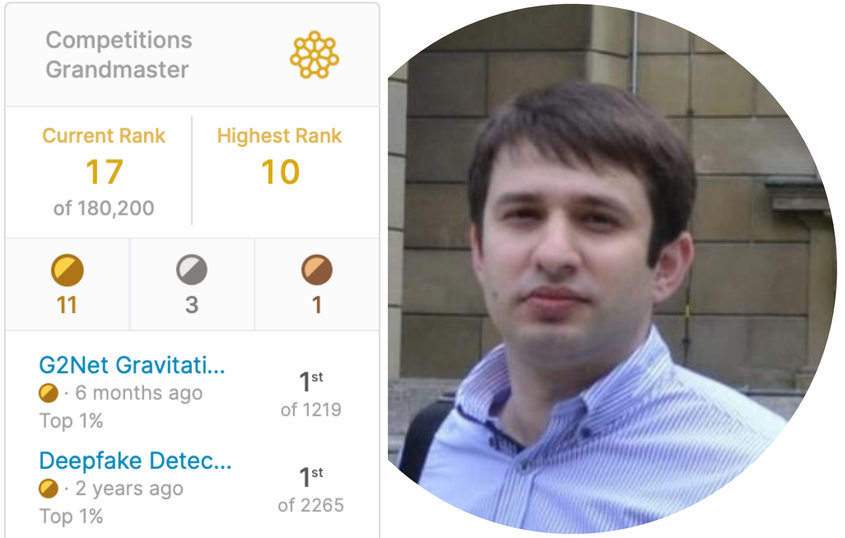
# During the challenge

- Interest in the challenge grew quickly - ended with ~1200 teams
- This was exemplified by the number of messages on the challenge forum
- Fortunately, most messages were between competitors (it's a quite friendly and helpful environment)
- Occasionally, Michael or myself would get asked something and we would answer carefully
- This was relatively low effort
- **After the challenge** - we are still working on the complete meta-analysis and hope to publish soon



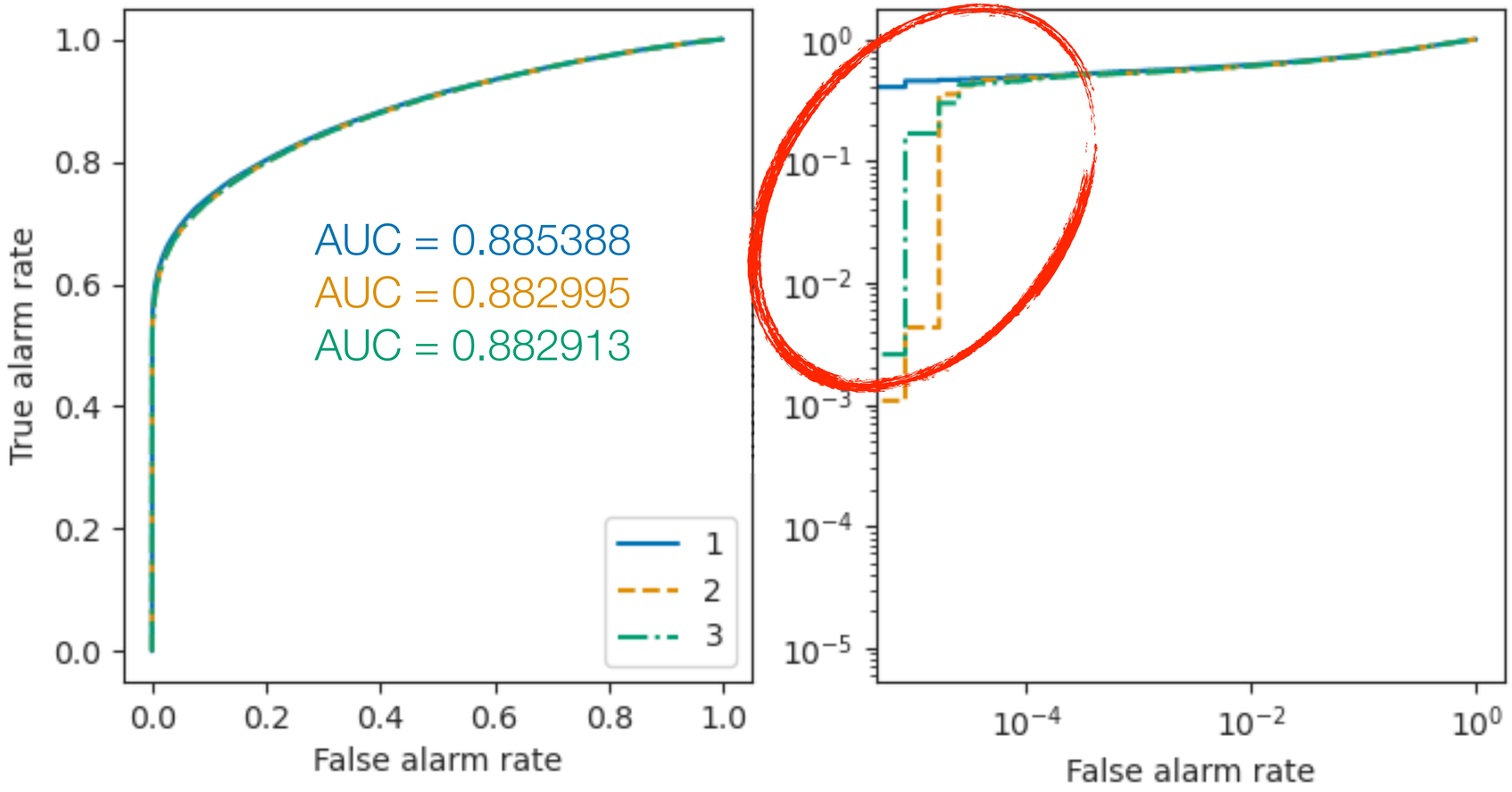
# 1st place (AUC=0.885388)

- Selim Seferbekov is a Computer Vision Engineer at Mapbox, Minsk, Belarus - joined Kaggle 6 years ago
  - Enormous experience in competitive Machine Learning, very basic understanding in Digital Signal Processing (DSP).
  - Attracted by the unusual competition topic
- Denis Kanonic is an Engineering Manager at Goldmine, studying for PhD in Computer Science
  - Profound experience with DSP in radio communications, significant reverse-engineering experience, some degree of familiarity with Gravitational Waves detection.
  - Attracted by the DSP-related competition.
- Both spent 2-4h daily for 1.5 month.
- Won $6000 for 1st place




100th place had AUC = 0.877268

# Results - ROC curves



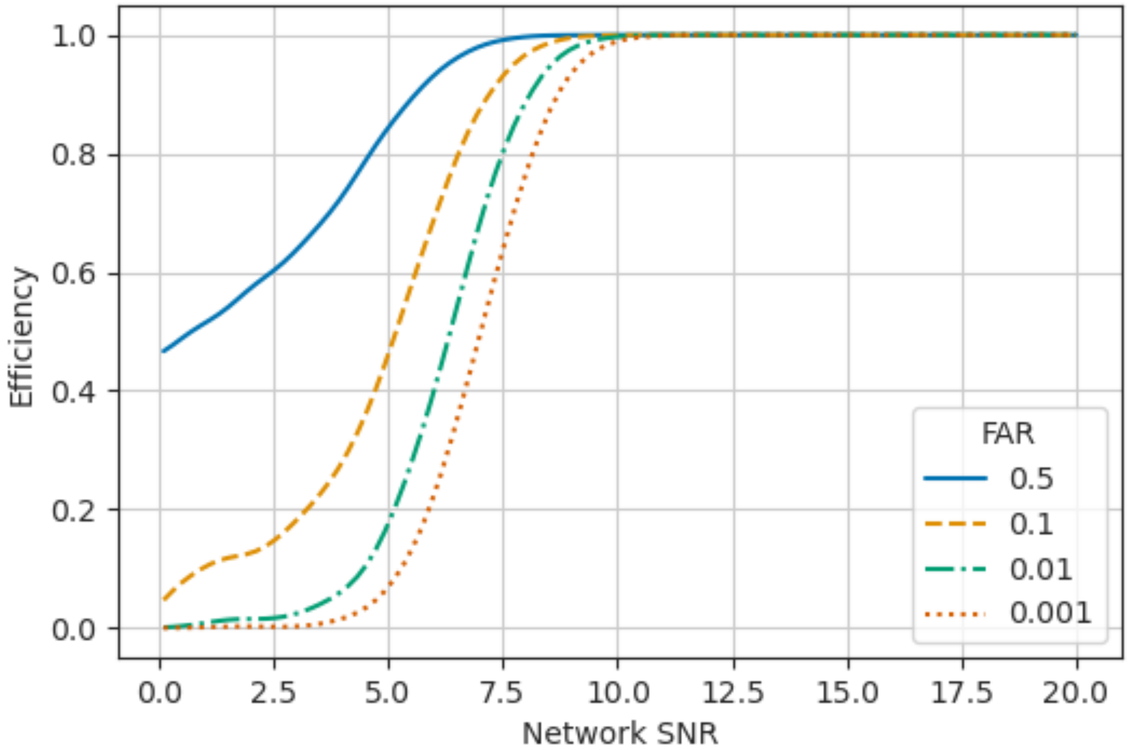
# Summary

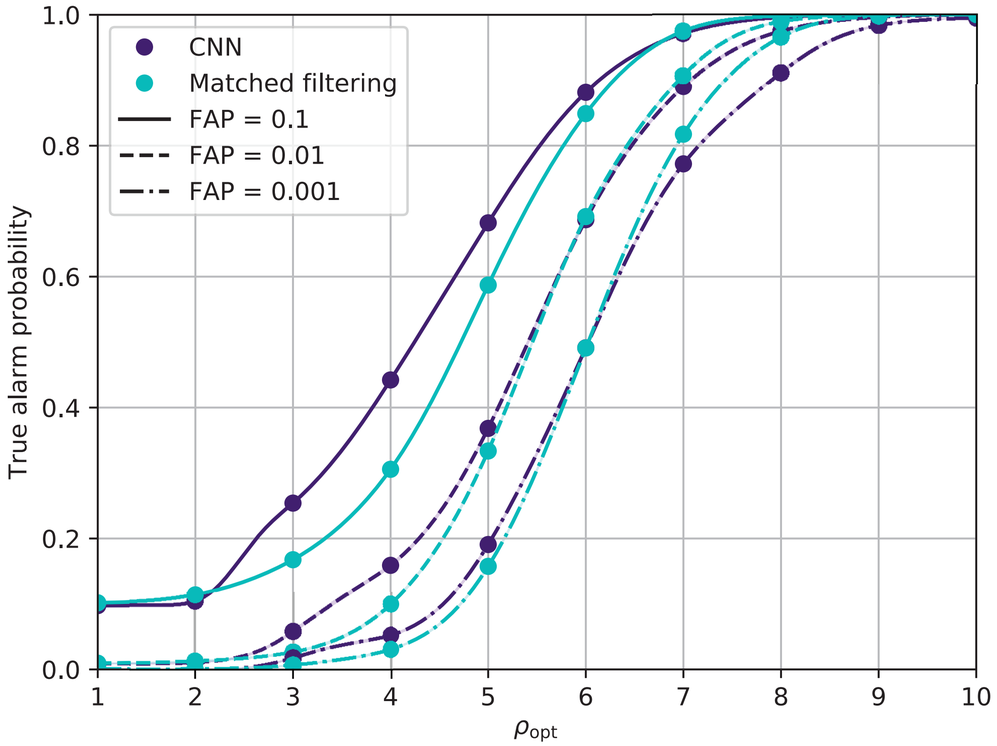
- There is a vast wealth of untapped knowledge and skill that we can harness to enhance our scientific impact
- This is a 2-way situation where the Kaggle community genuinely enjoys learning about the astrophysics behind our problems
- **We are very hopeful for our next challenge - the results could be very exciting (see the talk from Rodrigo)**

# Extra Slides

# Results - Efficiency

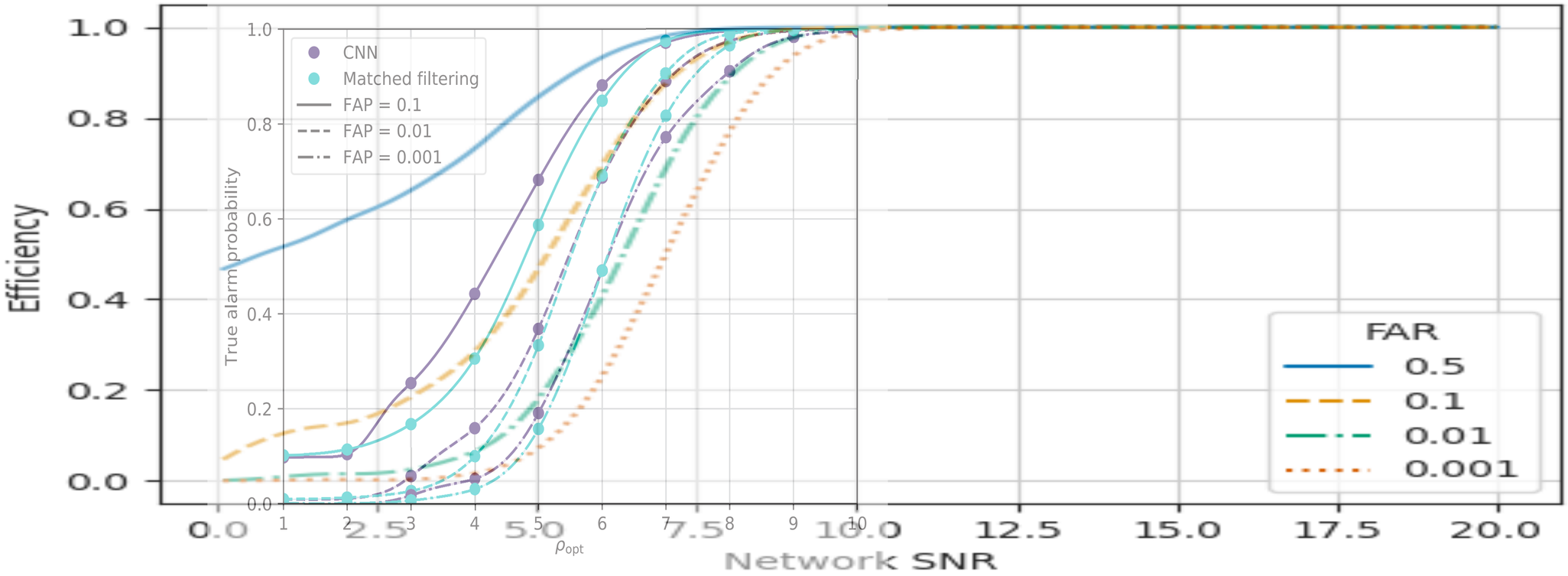
Gabbard *et al*, PRL (2018)



# Results - Efficiency

Gabbard *et al*, PRL (2018)

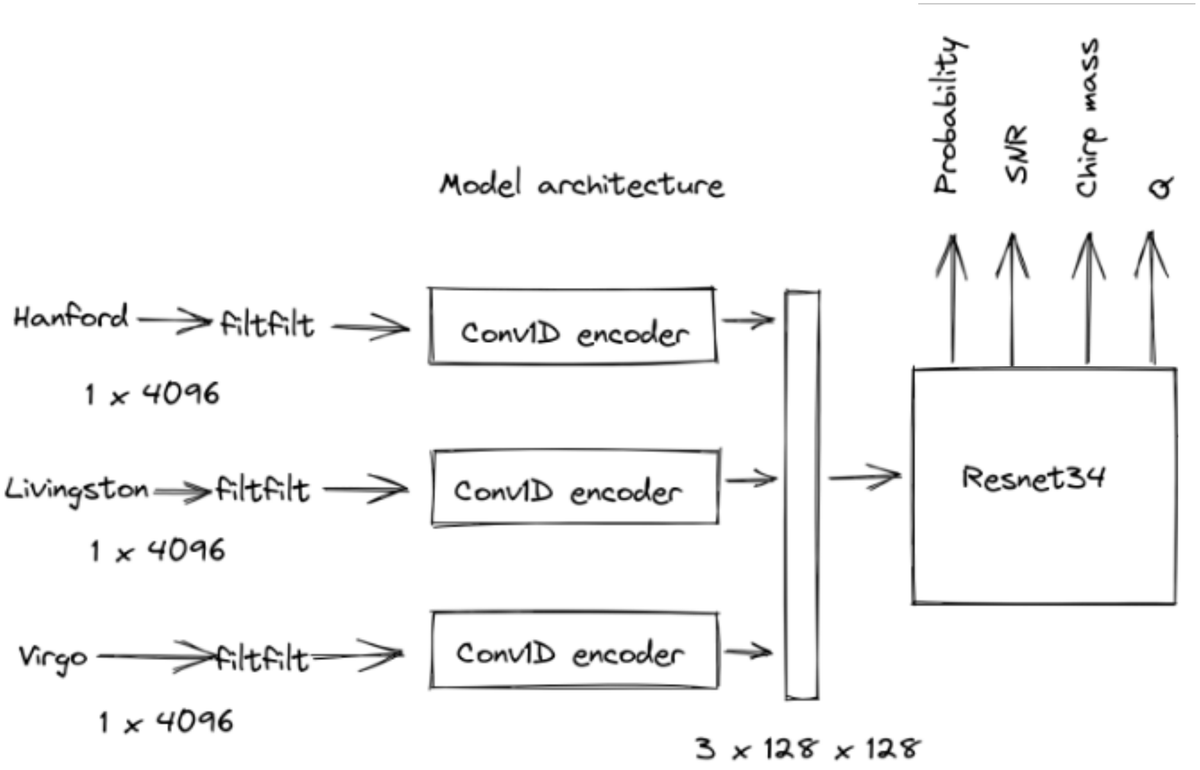


# 1st place interesting findings

- The most important trick used was custom Conv1D model with multiple large kernels
- The element that made them stand out from others was that they identified that 1D CNNs performed better than 2D CNNS and that they generated synthetic dataset for pre-training which helped to avoid severe overfitting
- They found that around 30% of positive sample cannot be identified by any model due to theoretical limit (so called SNR wall)
- When training on hard negative/hard positive samples, the model generalizes to predict signals/noise with reversed probability.
- Model execution time
  - pre-training takes ~2 days (fine tuning takes ~2 hours)
  - 20 mins to analyse all 220K testing data (

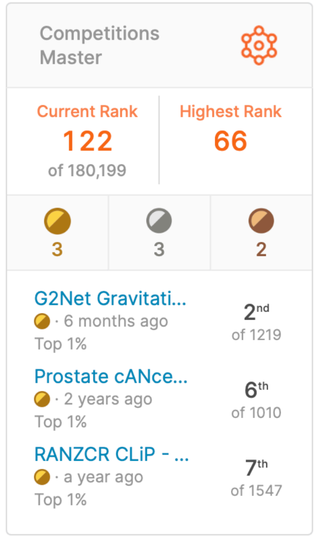
# 1st place model

- Able to generate unlimited amounts of training data (avoid overfitting) - **reverse engineered the training data using our own GW tools!**
- Pre-train on “home-made” data which allowed access to hidden parameters
- Needed to use learnable frontend to transform 1D data into more suitable time-frequency or time-feature 2D domain
- Required a separate frontend for each channel to eliminate the need of data whitening
- Needed to use lean encoder to limit overfitting
- pre-training takes ~2 days (fine tuning takes ~2 hours)
- 20 mins to analyse all 220K testing data (300 times faster than real time)



# 2nd place (0.88299)

- Hiroshi Yoshihara Machine learning engineer at Aillis Inc in Tokyo, Japan
- M.Sc. in health economics / epidemiology and Doctoral candidate in public health Professional background:
- Participated in many computer vision competitions.
- In total, spent more than 200 hours on it.
- Won $5000 for 2nd place
- Model summary: Front end generation of 2D features to a standard backend (MANY models)



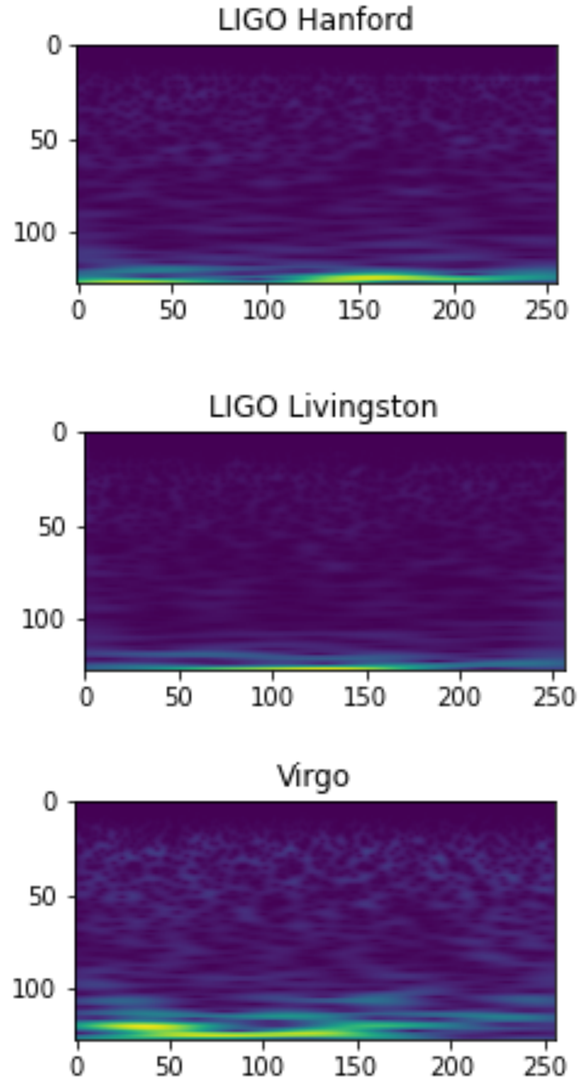


| # | Frontend | Backend | Input size | CV | Public LB | Private LB | Comment |
|---|---|---|---|---|---|---|---|
| 1 | CWT | efficientnet-b2 | 256 x 512 | 0.8779 | 0.8797 | 0.8782 | |
| 2 | CWT | efficientnet-b2 | 128 x 1024 | 0.87841 | 0.8801 | 0.8787 | |
| 3 | CWT | densenet201 | 256 x 512 | 0.87762 | 0.8796 | 0.8782 | |
| 4 | CWT | xcit-tiny-p16 | 384 x 768 | 0.87794 | 0.8800 | 0.8782 | |
| 5 | CWT | efficientnet-b7 | 128 x 1024 | 0.87957 | 0.8811 | 0.8800 | |
| 6 | CWT | efficientnet-b4 | 256 x 1024 | 0.87942 | 0.8812 | 0.8797 | |
| 7 | CWT | efficientnet-b2 | 128 x 1024 | 0.87875 | 0.8802 | 0.8789 | difference CWT params |
| 8 | WaveNet | efficientnet-b2 | 128 x 1024 | 0.87846 | 0.8809 | 0.8794 | |
| 9 | WaveNet | efficientnet-b6 | 128 x 1024 | **0.87982** | 0.8823 | 0.8807 | |
| 10 | WaveNet | densenet201 | 128 x 1024 | 0.87831 | 0.8818 | 0.8804 | |

| # | Frontend | Backend | Input size | CV | Public LB | Private LB | Comment |
|---|---|---|---|---|---|---|---|
| 11 | CNN | efficientnet-b6 | 128 x 1024 | **0.87982** | 0.8823 | 0.8808 | |
| 12 | WaveNet | effnetv2-m | 128 x 1024 | 0.87861 | **0.8831** | **0.8815** | |
| 13 | CNN | effnetv2-m | 128 x 1024 | 0.87847 | 0.8817 | 0.8799 | |
| 14 | WaveNet | effnetv2-l | 128 x 1024 | 0.87901 | 0.8829 | 0.8811 | |
| 15 | WaveNet | efficientnet-b6 | 128 x 1024 | 0.8797 | 0.8817 | 0.8805 | Channel-wise |
| 16 | WaveNet | efficientnet-b3 | 256 x 1024 | 0.87948 | 0.8820 | 0.8803 | |
| 17 | WaveNet | resnet200d | 128 x 1024 | 0.87791 | 0.881 | 0.8797 | |
| 18 | ResNet1d-18 | | - | 0.87663 | 0.8804 | 0.8785 | |
| 19 | WaveNet | | - | 0.87698 | 0.8796 | 0.8784 | |
| 20 | DenseNet1d-121 | | - | 0.86826 | 0.8723 | 0.8703 | |

# 2nd place model (actually lots of models)

- Neural network architecture played the most important role for improving the performance.
- In addition to a conventional spectrogram frontend + 2d-CNN, several trainable 1d-CNN based frontends (wavegram), and Complex Morlet wavelet transform, Wavenet, and a multi-scale CNN was used as a frontend before a 2d-CNN backend were attempted.
- For models with frontend--backend architectures, the depth and size of backend was correlated with model performance.
- After strong denoising by applying bandpass filter, 1d-CNN with no 2d-CNN backends also performed well. One-dimensional versions of ResNet and DenseNet were used.
- Simplified WaveNet without gated activation was also used because it converged much faster than it with gated attention.
- Improvement was found by using pseudo labels for the test dataset
- In order to increase diversity of prediction, 20 models were used and combined using Ridge regression.

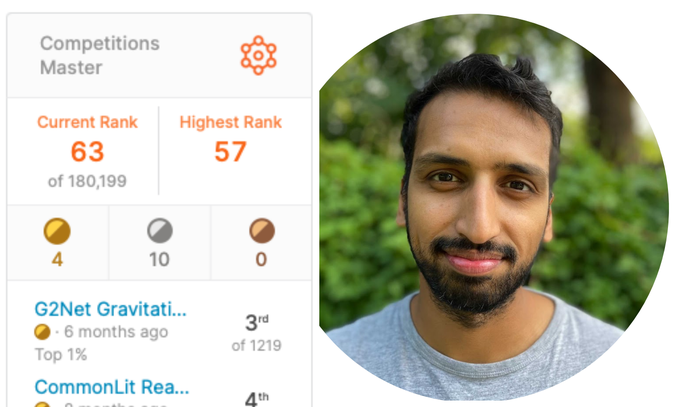


# 2nd place interesting findings

- Things that worked
  - Bandpass filter, adding gaussian noise, flipping waveform, test time augmentation, complex morlet wavelet transform, 1d-CNN as frontend (feature extractor), 1d-CNN, deeper backbone, semi-supervised learning (pseudo label), label smoothing during SSL.
- Things that did not work
  - Signal whitening, many types of augmentation (swapping channel, discrete wavelet transform denoising, shifting time, masking frequency bands, and etc.), melspectrogram (and many other SFTF based spectrograms), complex convolution, focal loss [6], online hard example mining loss [7], SSL using mean teacher framework.
- Setup and execution time
  - I used a local workstation with a NVIDIA GeForce RTX 3090 (24GB) and multiple cloud instances with a NVIDIA Tesla A100 (40GB). Surprisingly, training / inference time on both types of machine was almost the same.
  - In general, it took 2000 to 6000 seconds/epoch to train a frontend-backend model depending on its backbone architecture (no SSL). For a 1d-CNN model, it took only 1000 to 2000 seconds/epoch to train, which is much faster.
  - Inference time was around 500 to 1500 seconds for a frontend-backend model, and 200 to 500 seconds for a 1d-CNN model (no SSL / no TTA).

# 3rd place (0.88299)

- Anjum Sayed: Masters in Physics, Data Scientist, 12 years in the energy industry as a petrophysicist, Experience in the previous BirdClef competition was helpful in G2Net
- Maxim Shugaev: PhD in Applied Physics, Research Scientist in Computer Vision at Intelligent Automation, Inc., Kaggle Grandmaster (20+ competitions)
- Isamu Yamashita: Masters in Computer Science, Data Scientist in Canon. Inc., many years in the printer camera industry as a software engineer
- Ruiqi (Richard) Xing: PhD in Theoretical physics, Quant Researcher in Financial Industry, Prior machine learning and CNN knowledge helped in G2Net
- Ziyue (Vincent) Wang: Masters in Financial Engineering, Algorithmic Quant Researcher at BNPP
- Won $4000 for 3rd place
- Model summary: Whitening to Q-Transform to 2D pre-trained CNN backend



Competitions Master

| Current Rank | Highest Rank |
|---|---|
| 63 of 180,199 | 57 |

| 4 | 10 | 0 |
|---|---|---|

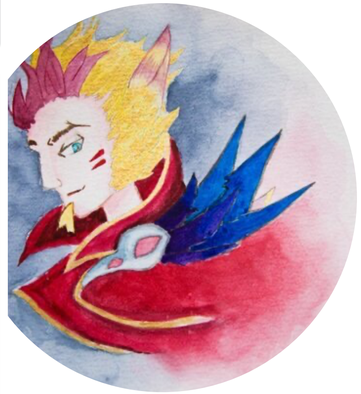
| | |
|---|---|
| G2Net Gravitati... · 6 months ago Top 1% | 3rd of 1219 |
| CommonLit Rea... · 8 months ago Top 1% | 4th of 3633 |
| RSNA Intracrani... · 2 years ago Top 1% | 5th of 1345 |

Competitions Grandmaster

| Current Rank | Highest Rank |
|---|---|
| 150 of 180,199 | 94 |

| 5 | 12 | 4 |
|---|---|---|

| | |
|---|---|
| G2Net Gravitati... · 6 months ago Top 1% | 3rd of 1219 |
| HuBMAP - Hack... · a year ago Top 1% | 4th of 1200 |
| OSIC Pulmonary... · a year ago Top 1% | 9th of 2097 |

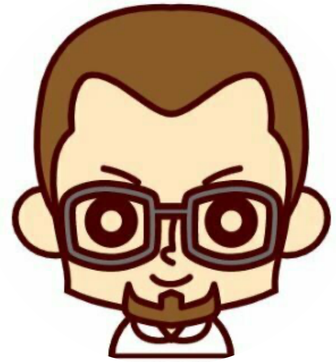


Competitions Contributor

Unranked

| 1 | 0 | 0 |
|---|---|---|

| | |
|---|---|
| G2Net Gravitati... · 6 months ago Top 1% | 3rd of 1219 |
| Optiver Realize... 3 months ago Top 69% | 2,656th of 3852 |
| DonorsChoose.... 4 years ago Top 26% | 149th of 580 |



Competitions Master

| Current Rank | Highest Rank |
|---|---|
| 81 of 180,199 | 39 |

| 6 | 6 | 3 |
|---|---|---|

| | |
|---|---|
| ASHRAE - Great... · 2 years ago Top 1% | 1st of 3614 |
| G2Net Gravitati... · 6 months ago Top 1% | 3rd of 1219 |
| Indoor Location ... · a year ago Top 1% | 7th of 1170 |

Competitions Master

| Current Rank | Highest Rank |
|---|---|
| 355 of 180,199 | 334 |

| 2 | 1 | 2 |
|---|---|---|

| | |
|---|---|
| G2Net Gravitati... · 6 months ago Top 1% | 3rd of 1219 |
| TensorFlow - H... · 2 months ago Top 1% | 11th of 2025 |
| Jane Street Mar... · 7 months ago Top 2% | 84th of 4245 |