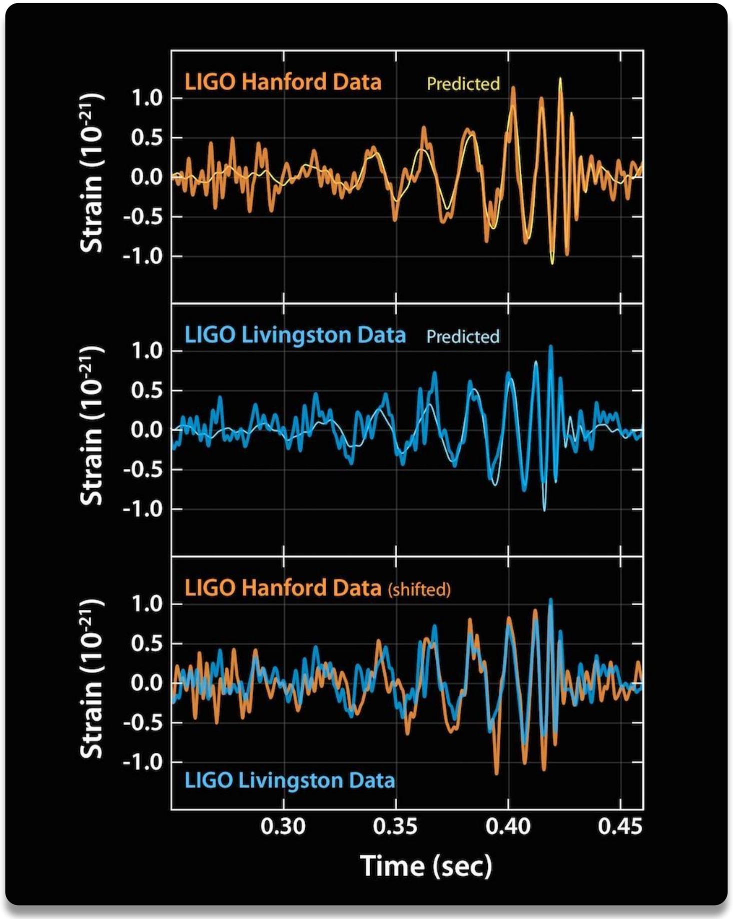
Finding the origin of noise transients in LIGO data with machine learning

REBECCA WHITE

Introduction – Gravitational Waves (GWs)



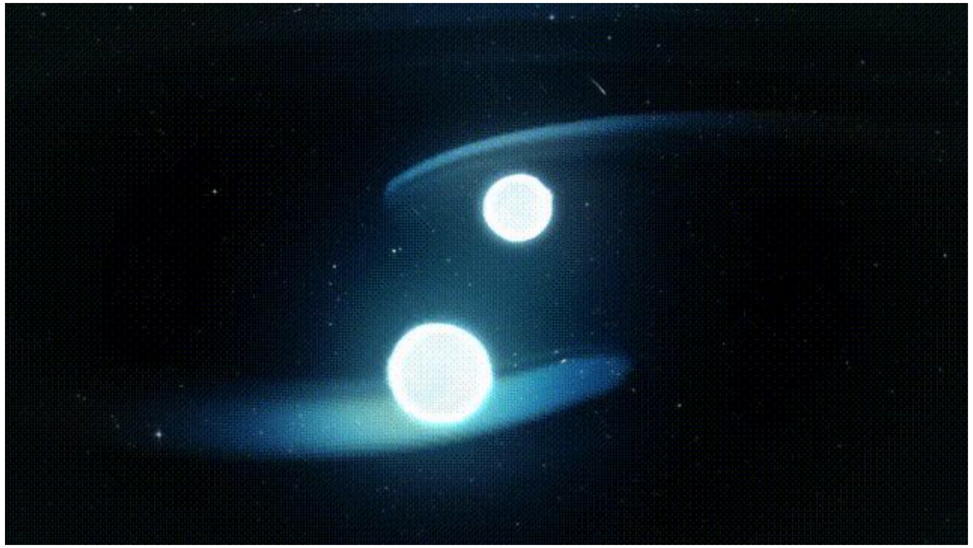
- Predicted by Einstein in 1916
- Most caused by high-energy events like collisions between black holes/neutron stars
- Can also be caused by single-spinning neutron stars or be from the formation of the universe
- First detected by LIGO on September 14th 2015

Introduction – LIGO

- Laser Interferometer Gravitational Wave Observatory
- Two detectors in Hanford, Washington and Livingston, Louisiana
- Virgo – detector in Pisa, Italy
- The detectors have detected 14 confirmed GW events, most of them being Binary Black Hole (BBH) collisions through 3 observing runs
- Capable of measuring a change in distance on the order of 10^{-19} meters

Introduction – Why Study GWs?

- EM radiation and other ways to study the universe can be obstructed through astrophysical objects and phenomena
- Analyze events that can’t be “seen” though other waves
- There are waves from the beginning of our universe that haven’t been altered while traveling through space



Introduction – Noise Problem

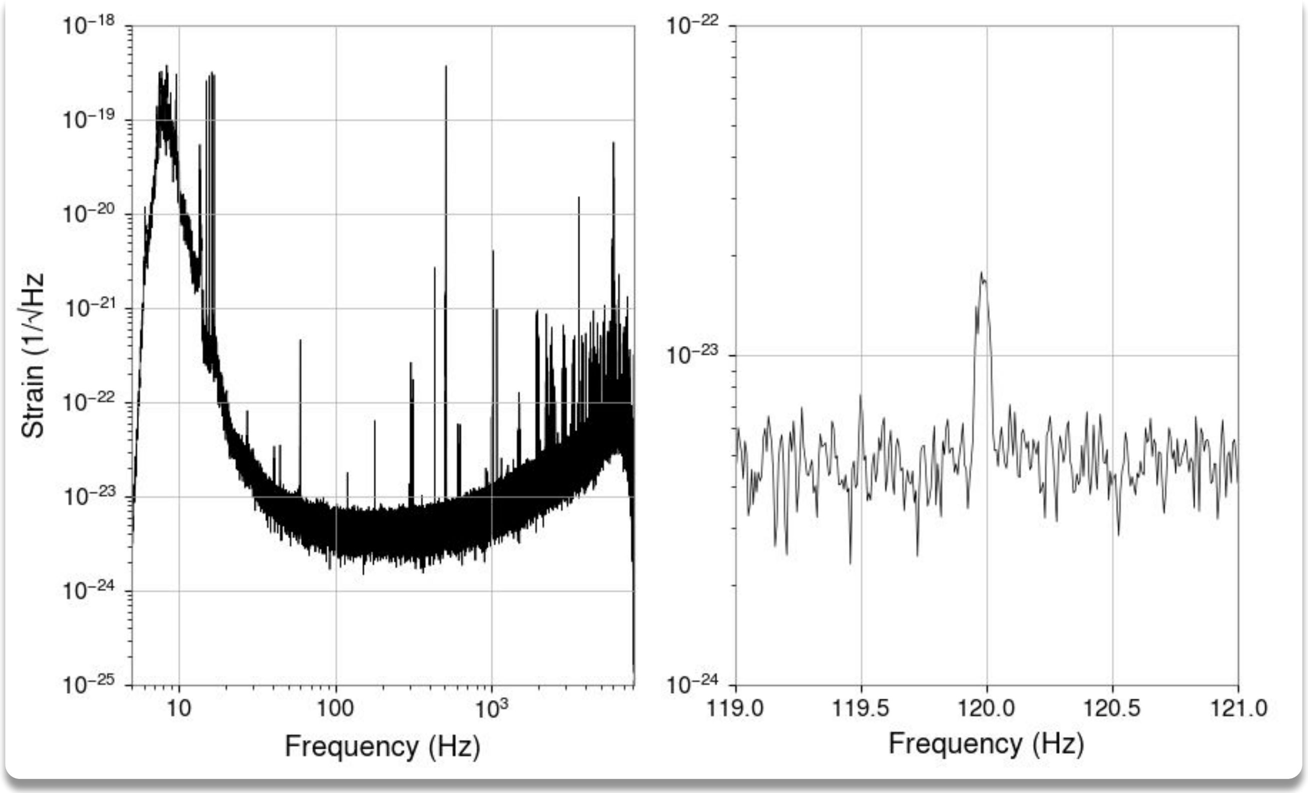
- The detectors are extremely sensitive and are impacted by non-astrophysical disturbances
- Instrumental and environmental noise increases the false alarm rate
- Noise has made it impossible to find long-transient and continuous waves (CWs)
- Removing noise artifacts will greatly improve the statistical significance of GW events and help us possibly find CWs
- Before removal, noise needs to be better understood and classified

Proposed Solution

- Use machine learning methods to identify noisy times in the detector output as well as identify the sources of said noises so the detector can be adjusted to remove the noise
- "We introduce two ML algorithms that provide simple, yet robust methods to mine the data of auxiliary channels and infer the origin of noise transients in the main detector output."
- Ultimately, make fast/effective methods that data analysts can use with little tuning

Inputs & Raw Data

- List of times when certain noises occur with already known origins
- Raw data is time series data in HDF5 and Frame files
- Can be graphed over the frequency domain to make it easier to notice noise/glitches



Methods – Random Forest (RF)

- Construct multiple decision trees and then use them to determine a predicted result by taking the result with the most "votes"
- Pros:
  - Considered highly accurate because of the number of decision trees
  - Cancels out biases
  - Can determine the most contributing features
- Cons:
  - Can be slow in generating predictions
  - Can be difficult to interpret

Methods – Genetic Programming (GP)

- Compares each executed hypothesis with a qualified label to make a fitness value and the ones with higher fitness values are most likely to be chosen for the next generation so each generation is more likely to be accurate and solve the problem than the previous
- Pros:
  - Doesn’t require derivative information
  - Easily parallelizable
  - Provides more than one solution
  - Useful when the search is large
- Cons:
  - No guarantees in optimality or quality of the solutions
  - The “fitness value” is calculated repeatedly which could be computationally expensive

Data Sets Used

- Two sets of glitches with known origin from the first and second observing runs:
  - X-arm end station (EX) magnetometer glitches (2049 data points)
  - Air compressor coupling (42 data points)
    - While the number is low, the physical properties have been well characterized and are well understood
- Both are extreme cases and therefore help when testing the effectiveness of the algorithms

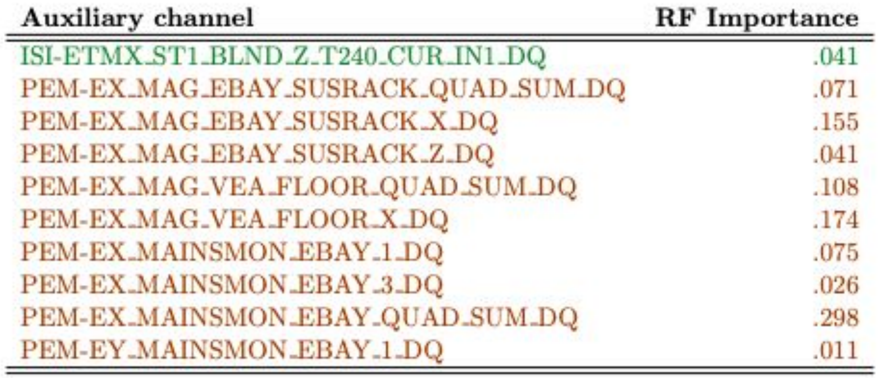
Results – General

- Ran RF algorithm on the training sets
- Varied the number of estimators and the iteration threshold – no significant change in results
- RF results were validated with the GP that was run with multiple different parameter configurations

Different colors denote different detector subsystems and auxiliary channels: Seagreen = Armlength Stabilization (ALS), orchid = Alignment Sensing and Control (ASC), goldenrod = Photon Calibrator (CAL) royal blue = Hydraulic External Pre-Isolator (HPI), olive green = Internal Seismic Isolation (ISI), violet = Length Sensing and Control (LSC), sienna = Physical and Environmental Monitor (PEM), turquoise = Suspension (SUS), magenta = Thermal Compensation (TCS).

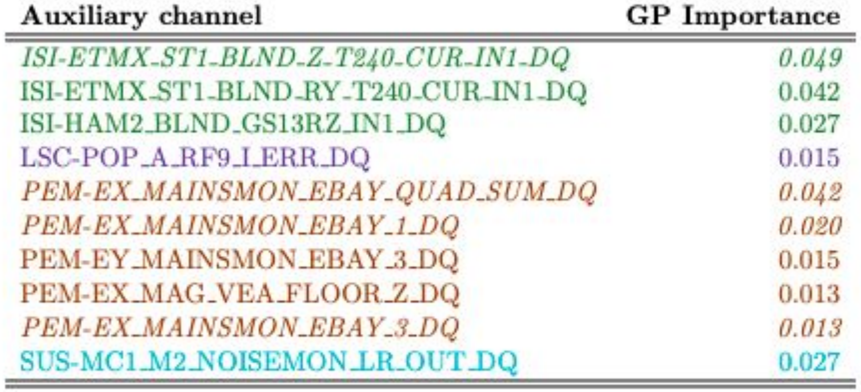
Results – Magnetometer

- "The RF algorithm correctly identifies the voltage monitor of the EX electronics bay as the origin of the noise transients."
- It also pointed to the origin of the noise being electromagnetic – this is consistent with what was found to be the cause
- The GP results validated the RF results for the origin of the noise



Auxiliary channel	RF Importance
ISI-ETMX_ST1_BLND_Z_T240_CUR_IN1_DQ	.041
PEM-EX_MAG_EBAY_SUSRACK_QUAD_SUM_DQ	.071
PEM-EX_MAG_EBAY_SUSRACK_X_DQ	.155
PEM-EX_MAG_EBAY_SUSRACK_Z_DQ	.041
PEM-EX_MAG_VEA_FLOOR_QUAD_SUM_DQ	.108
PEM-EX_MAG_VEA_FLOOR_X_DQ	.174
PEM-EX_MAINSMON_EBAY_1_DQ	.075
PEM-EX_MAINSMON_EBAY_3_DQ	.026
PEM-EX_MAINSMON_EBAY_QUAD_SUM_DQ	.298
PEM-EY_MAINSMON_EBAY_1_DQ	.011

Table 1: Auxiliary channels with nonzero RF importance for the magnetometer set. Different colors denote instrumental and environmental auxiliary channels corresponding to different detector subsystems: Olive green = Internal Seismic Isolation (ISI), sienna = Physical and Environmental Monitor (PEM).



Auxiliary channel	GP Importance
ISI-ETMX_ST1_BLND_Z_T240_CUR_IN1_DQ	*0.049*
ISI-ETMX_ST1_BLND_RY_T240_CUR_IN1_DQ	0.042
ISI-HAM2_BLND_GS13RZ_IN1_DQ	0.027
LSC-POP_A_RF9_I_ERR_DQ	0.015
PEM-EX_MAINSMON_EBAY_QUAD_SUM_DQ	*0.042*
PEM-EX_MAINSMON_EBAY_1_DQ	*0.020*
PEM-EY_MAINSMON_EBAY_3_DQ	0.015
PEM-EX_MAG_VEA_FLOOR_Z_DQ	0.013
PEM-EX_MAINSMON_EBAY_3_DQ	*0.013*
SUS-MC1_M2_NOISEMON_LR_OUT_DQ	0.027

Table 3: Auxiliary channels with GP importance larger than 0.012 for the magnetometer set. As in Fig. 4, different colors denote instrumental and environmental auxiliary channels corresponding to different detector subsystems. Channels in italic denote those selected also by the RF algorithm (see Table 1).

Results – Histograms

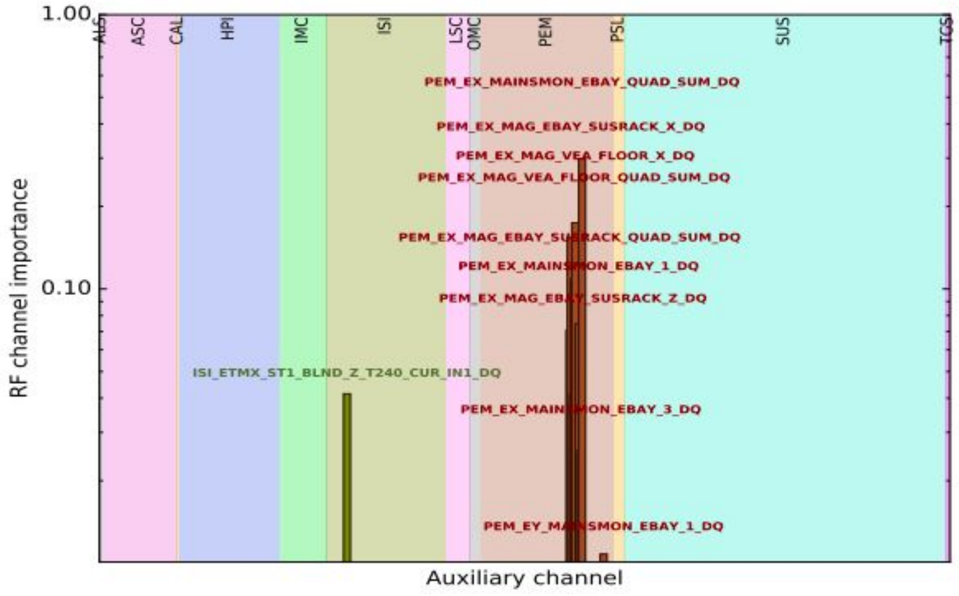


Figure 4: Histogram of RF channel importance for the magnetometer set from the data in Table 1. Different colors denote different detector subsystems and auxiliary channels: Seagreen = Arm-length Stabilization (ALS), orchid = Alignment Sensing and Control (ASC), goldenrod = Photon Calibrator (CAL) royal blue = Hydraulic External Pre-Isolator (HPI), lime green = Input Mode Cleaner (IMC), olive green = Internal Seismic Isolation (ISI), violet = Length Sensing and Control (LSC), gray = Output Mode Cleaner (OMC), sienna = Physical and Environmental Monitor (PEM), orange = Pre-Stabilized Laser (PSL), turquoise = Suspension (SUS), magenta = Thermal Compensation (TCS). The plot clearly shows how the glitches arise from an environmental electromagnetic disturbance in the EX station.

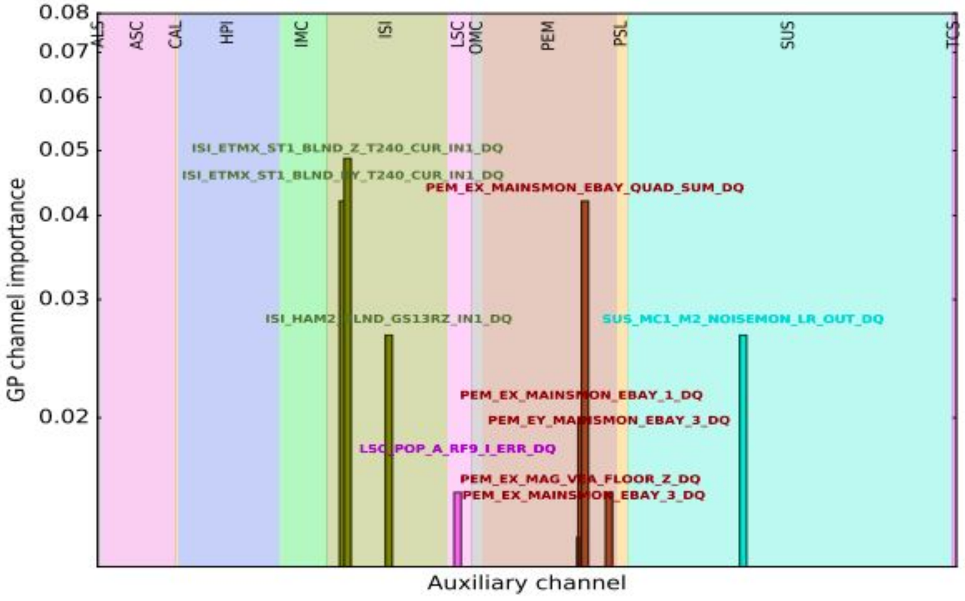
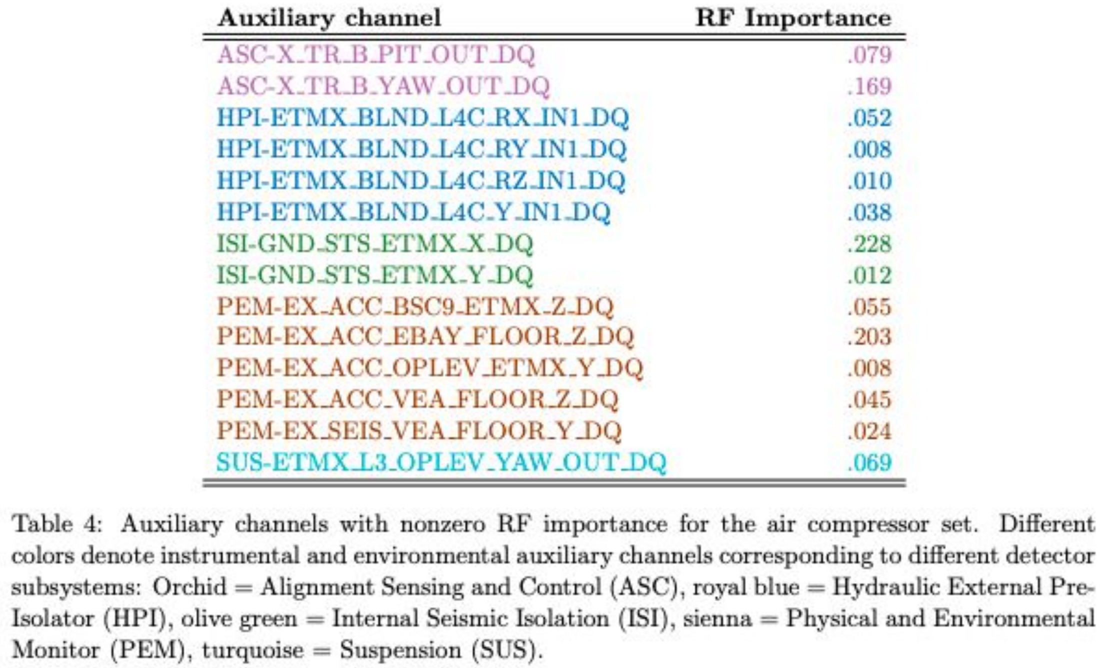


Figure 6: Channel importance for the magnetometer set from Karoo GP. Only auxiliary channels with GP importance > 0.012 are shown.

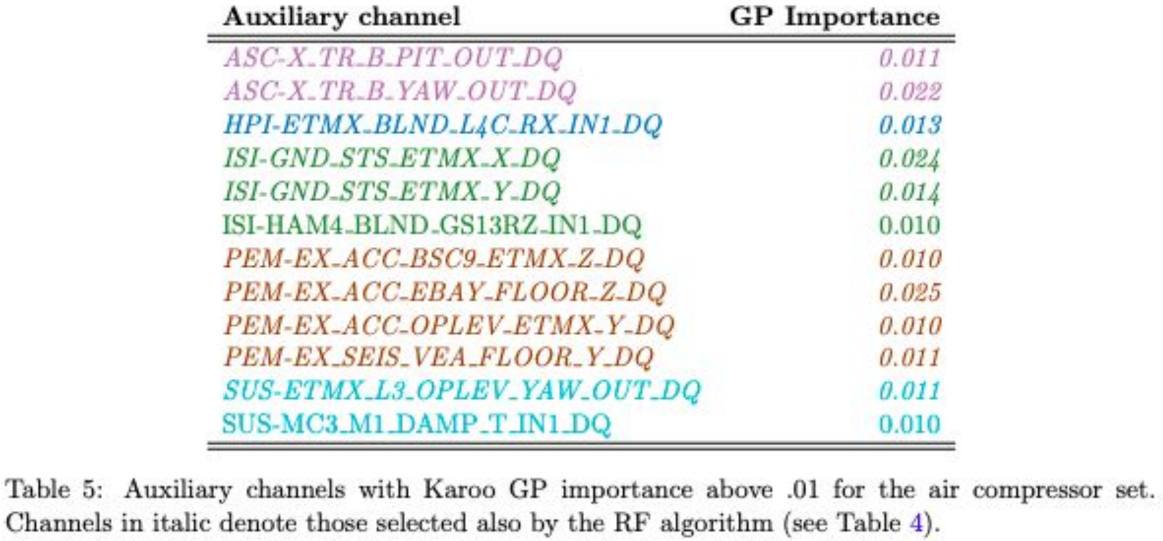
Results – Air Compressor

- After data preparation, the dataset was reduced to 16 noise transients
- “If the origin of mechanical couplings can be inferred with just a limited number of recorded glitches as soon as they appear in the detector, the method may prove very useful for commissioning purposes.”
- The RF identified the glitches as having seismic origin – which is correct
- The GP results validated the RF results for the origin of the noise



Auxiliary channel	RF Importance
ASC-X_TR_B_PIT_OUT_DQ	.079
ASC-X_TR_B_YAW_OUT_DQ	.169
HPI-ETMX_BLND_L4C_RX_IN1_DQ	.052
HPI-ETMX_BLND_L4C_RY_IN1_DQ	.008
HPI-ETMX_BLND_L4C_RZ_IN1_DQ	.010
HPI-ETMX_BLND_L4C_Y_IN1_DQ	.038
ISI-GND_STS_ETMX_X_DQ	.228
ISI-GND_STS_ETMX_Y_DQ	.012
PEM-EX_ACC_BSC9_ETMX_Z_DQ	.055
PEM-EX_ACC_EBAY_FLOOR_Z_DQ	.203
PEM-EX_ACC_OPLEV_ETMX_Y_DQ	.008
PEM-EX_ACC_VEA_FLOOR_Z_DQ	.045
PEM-EX_SEIS_VEA_FLOOR_Y_DQ	.024
SUS-ETMX_L3_OPLEV_YAW_OUT_DQ	.069

Table 4: Auxiliary channels with nonzero RF importance for the air compressor set. Different colors denote instrumental and environmental auxiliary channels corresponding to different detector subsystems: Orchid = Alignment Sensing and Control (ASC), royal blue = Hydraulic External Pre-Isolator (HPI), olive green = Internal Seismic Isolation (ISI), sienna = Physical and Environmental Monitor (PEM), turquoise = Suspension (SUS).



Auxiliary channel	GP Importance
ASC-X_TR_B_PIT_OUT_DQ	*0.011*
ASC-X_TR_B_YAW_OUT_DQ	*0.022*
HPI-ETMX_BLND_L4C_RX_IN1_DQ	*0.013*
ISI-GND_STS_ETMX_X_DQ	*0.024*
ISI-GND_STS_ETMX_Y_DQ	*0.014*
ISI-HAM4_BLND_GS13RZ_IN1_DQ	0.010
PEM-EX_ACC_BSC9_ETMX_Z_DQ	*0.010*
PEM-EX_ACC_EBAY_FLOOR_Z_DQ	*0.025*
PEM-EX_ACC_OPLEV_ETMX_Y_DQ	*0.010*
PEM-EX_SEIS_VEA_FLOOR_Y_DQ	*0.011*
SUS-ETMX_L3_OPLEV_YAW_OUT_DQ	*0.011*
SUS-MC3_M1_DAMP_T_IN1_DQ	0.010

Table 5: Auxiliary channels with Karoo GP importance above .01 for the air compressor set. Channels in italic denote those selected also by the RF algorithm (see Table 4).

Results – Histograms

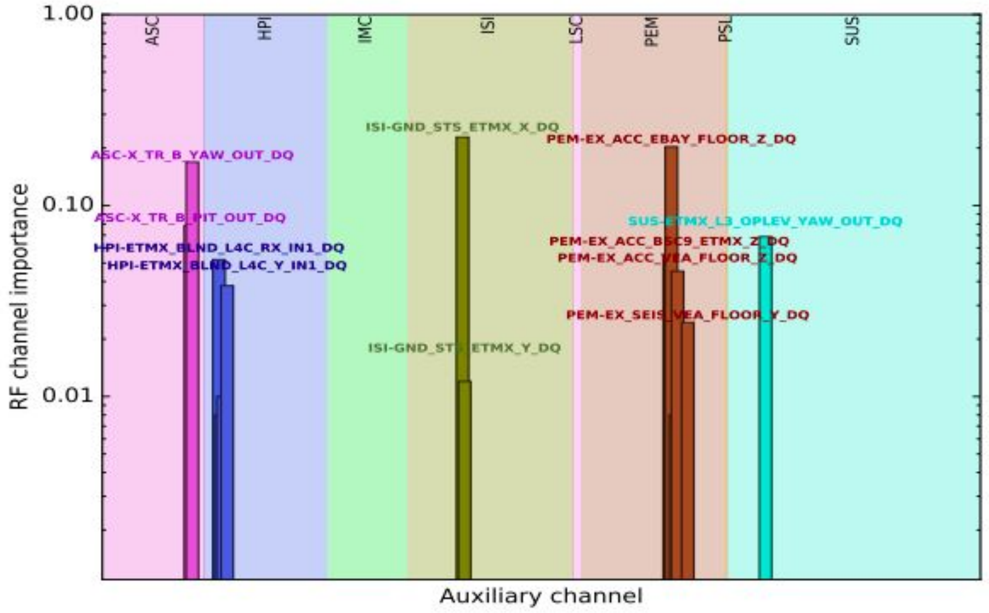


Figure 7: Histogram of RF channel importance for the air compressor set from the data in Table 4. Different colors denote different detector subsystems and auxiliary channels: Orchid = Alignment Sensing and Control (ASC), royal blue = Hydraulic External Pre-Isolator (HPI), lime green = Input Mode Cleaner (IMC), olive green = Internal Seismic Isolation (ISI), violet = Length Sensing and Control (LSC), sienna = Physical and Environmental Monitor (PEM), orange = Pre-Stabilized Laser (PSL), turquoise = Suspension (SUS).

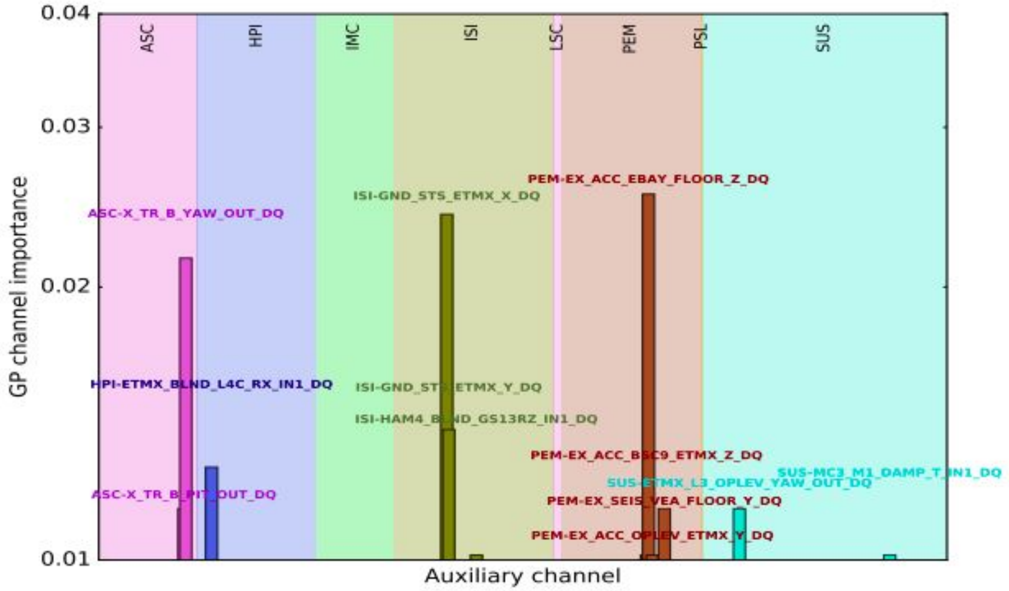


Figure 8: Channel importance for the air compressor set from Karoo GP. Only auxiliary channels with importance > 0.01 are shown.

Conclusion

- ML methods can be used to infer the causes of glitches
- This could help solve non-astrophysical noise problems sooner
- This could also help with noise cancellation after data has been collected
- The results might be improved if other ML concepts are implemented

Accessing Data and Methods

- LIGO Strain data: https://www.gw-openscience.org/data/
- RF algorithm: https://scikit-learn.org/stable/
- GP algorithm: https://kstaats.github.io/karoo_gp/