Strategy for signal classification to improve data quality for Advanced Detectors gravitational-wave searches

E. Cuoco[1], A.Torres-Forné[2],J.A. Font[2,7], J.Powell[3], R.Lynch[4], D.Trifiró[5 6] , M.Cavaglià[6], Ik S. Heng[2]

[2]Universitat de València, [1]EGO and PISA INFN, [3]SUPA and IGR Glasgow, [4]MIT, [5]Università di Pisa,[6]University of Mississippi, [7]Observatori Astronòmic, Universitat de València

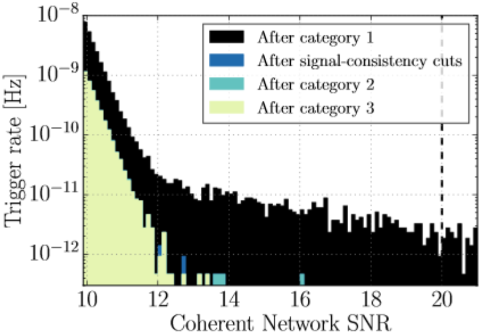
SciNeGHE 2016, October 19

Data Quality for Advanced detectors

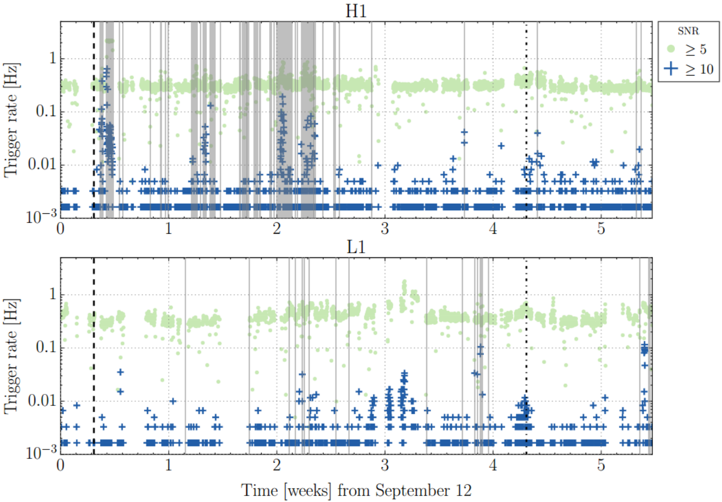
- Transient noise (glitches) can occur within the targeted frequency range and can mimic Gravitational Waves events
- We can cure the problem either in experimental apparatus or by using signal processing strategy
- More than 200000 auxiliary channels are recorded to monitor instrument behaviour and environmental conditions
- In the case of clear correlation within glitches in gravitational wave channel and auxiliary ones, data are discarded from the analysis (vetoed)

Cleaning our triggers distribution

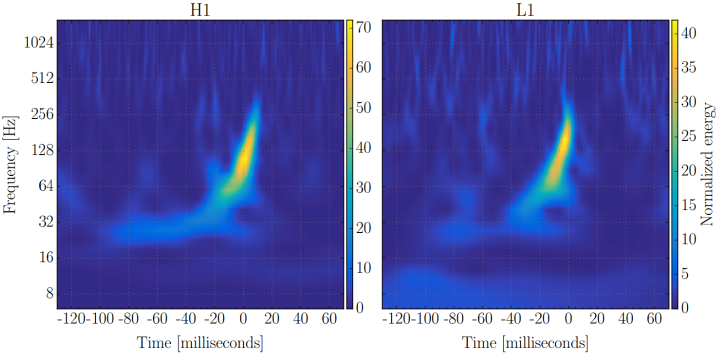
Here an example of the impact of data-quality vetoes and signal consistency requirements on the background trigger distribution from the cWB search for gravitational-wave bursts by coherent network SNR.



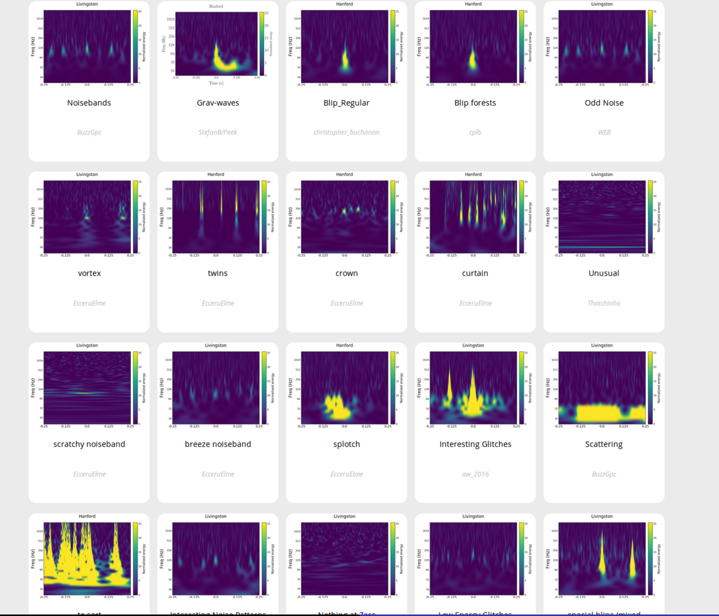
Typical glitchgram for detectors



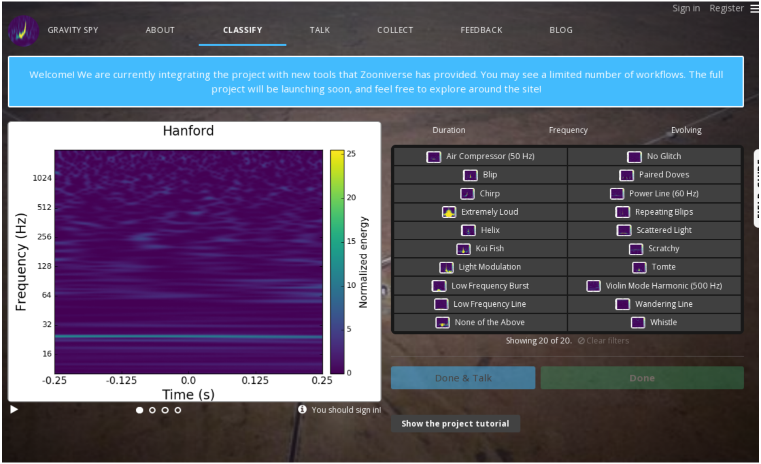
Our 'typical' gravitational waves



Our glitch zoo



GravitySpy project



https://www.zooniverse.org/projects/zooniverse/gravity-spy

Why Glitch Classification?

- As prompt characterization of noise will be critical for improving sensitivity, a fast method for glitch classification was needed.
- We aim to develop methods for automatic classification of glitches.
- We present three methods developed for automatic glitch classification.
- We started using simulated data sets to better understand the performance of the different glitch classifying codes.
- We tested our pipelines on real data (LIGO ER7 data).

Principal Components.

- All three methods use at some stage Principal Components (PCs).
- PCs are a set of orthogonal basis vectors, which are ordered so that the first PC represents the most common feature of a set of waveforms.
- Therefore, a few PCs can be used to represent all the common features of the waveforms.
- The signal model consists of a linear combination of PCs .

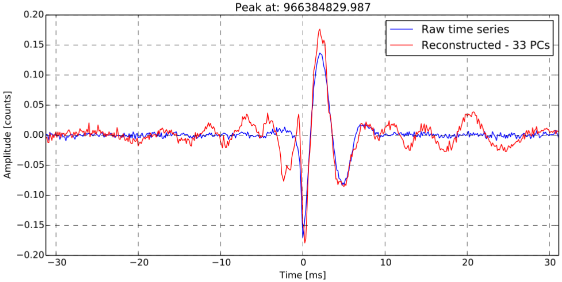


Figure: A glitch reconstructed by PCAT using 33 PCs.

- Results can be strongly effected by the number of Principal Components.
- We use the variance method to choose the ideal number of Principal Components.

PC-LIB

- PC-LIB is an adaptation of the parameter estimation and model selection tool LALInference.
- A set of Principal Components for a type of glitch is made using the high pass filtered time series of fifty glitches for that type.
- A linear combination of the PCs, multiplied by the PC coefficients, is then used as the new signal model in LIB for each different population of noise transient. The different signal models for each glitch population can then be used for Bayesian model selection, which can determine the type of each new noise transient that is detected in the data.

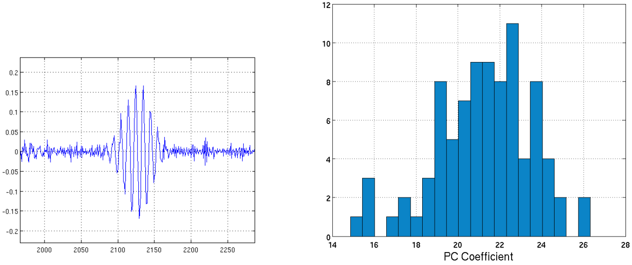
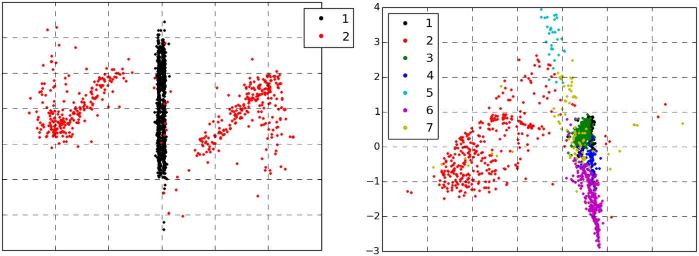


Figure: A glitch reconstructed by PCAT using 33 PCs.

PCAT

- Principal Component Analysis for Transients (PCAT) is a python-based classifier based on Principal Component Analysis
- The time series of whitened glitches are stored in a matrix on which PCA is performed.
- The results of the PCA can be visualized with scatter plots of the principal component coefficients
- PCAT uses the PC coefficients to classify the glitches by applying a Gaussian Mixture Model (GMM) classifier to the coefficients.



WDF-ML: introduction

- WDF-ML consists of an event detection algorithm, Wavelet Detection Filter (WDF), followed by a Machine Learning (ML) classification procedure.
- WDF is part of the Noise Analysis Package (NAP), a C++ library embedded in python, developed by the Virgo Collaboration
- A whitening procedure is applied to the data and is based on a Linear Predictor Filter.
- The parameters are estimated through a parametric Auto Regressive (AR) model fit to the noise PSD.

WDF-trigger

The thresholding function is applied to the wavelet transform of the noisy signal, then the output is inverted and the wavelet transformed. After the wavelet thresholding, we selected the highest coefficients of the wavelet transform which are supposed to contain only the signal and not the noise.

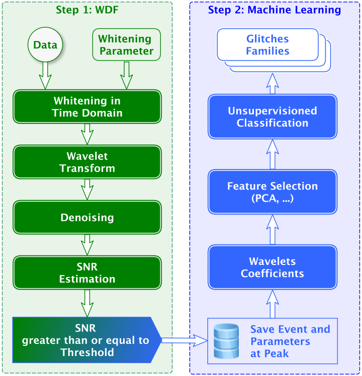
$$E_s = \sqrt{\sum_{k,j} w_{k,j}^2} \tag{1}$$

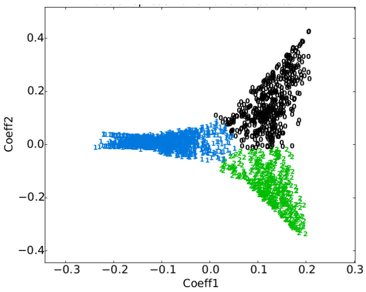
being $w_{k,j}$the wavelet coefficients above the threshold.
In this way E_s represent the signal energy content, so we can build our receiver detector which represents the signal to noise ratio, as

$$SNR = \frac{E_s}{\hat{\sigma}} \tag{2}$$

WDF-ML: Machine Learning

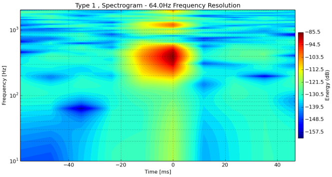
- Completely unsupervised algorithms. No target function
- Wavelets coefficients and Meta data (SNR, Freq,Duration) represents our "features"
- Features selection uses PCA transform an Spectral embedding on 2 dimensions
- The Gaussian Mixture Model (GMM) machine learning classifier is then applied to the outputs of WDF for classification.

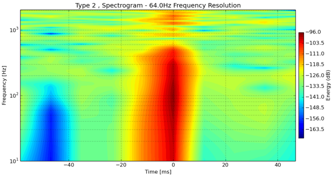




MDC: Data set 1

- To test and compare methods we create a simulated data set in aLIGO Gaussian noise.
- Data set 1 is an ideal data set where all of the glitch types are well separated in frequency and SNR.
- The data set contains 1000 sine Gaussian waveforms and 1000 Gaussian waveforms in simulated Gaussian noise.
- The sine Gaussian waveforms have a frequency $= 400$Hz and an SNR between 5 and 30.
- The Gaussian waveforms are centred at f $= 0$Hz and have an SNR between 20 and 250.





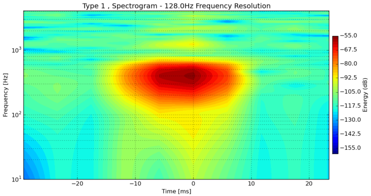
Data Set 1 Results

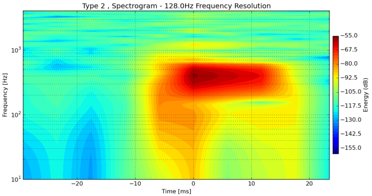
- Table shows the % of detected transients that were classified in each type.
- A few low frequency SG, and low SNR G were in the incorrect classes.
- Overall classification efficiency very good!

	SG	G
PCAT Type 1	99%	0%
PCAT Type 2	1%	100%
LIB Type 1	99.9%	5%
LIB Type 2	0.1%	95%
WDF Type 0	99.5%	2.4%
WDF Type 1	0.3%	46.1%
WDF Type 2	0.2%	51.5%

MDC: Data set 2

- We use a second data set to see if we can classify glitches by waveform morphology only.
- We use 1000 sine Gaussian waveforms and 1000 Ring-down waveforms.
- All waveforms have identical frequency 400Hz and a identical duration 2ms.
- The SNR of the simulated glitches is between 10 and 500.





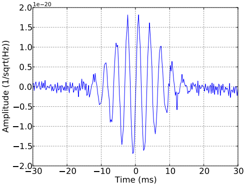
Data Set 2 Results

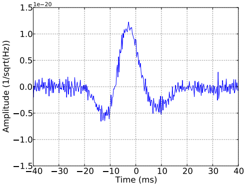
- Table shows the % of detected transients that were classified in each type.
- The few transients in the incorrect class are those with the lowest SNR.
- 5PCs PCAT, 7PCs LIB and 10 PCs WDF-ML.
- All methods can classify by waveform morphology alone.

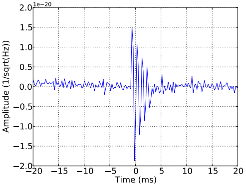
	SG	RD
PCAT Type 1	1.1%	97.4%
PCAT Type 2	98.9%	2.5%
LIB Type 1	97.8%	4.8%
LIB Type 2	2.2%	95.2%
WDF-ML Type 0	8.7%	100%
WDF-ML Type 1	48.0%	0%
WDF-ML Type 2	43.3%	0%

MDC: Data Set 3

- The third data set is to see what happens if different types have a very wide range of parameters.
- The simulated glitches are Gaussian, sine Gaussian and Ring-down waveforms at five second intervals.
- The frequencies are distributed linearly between 40-1500 Hz.
- Majority of the glitches have an SNR between 1 and 300.







Data Set 3 Results

- PCAT 20PCs, LIB 5PCs, WDF-ML 10PCs.
- All methods have the Gaussians in there own class.
- Cannot distinguish between the sine Gaussian and Ring-down waveforms when the parameter range is so large.

	SG	G	RD
PCAT Type 1	15.5%	0%	13.6%
PCAT Type 2	36.8%	0%	41.4%
PCAT Type 3	14.2%	0%	13.0%
PCAT Type 4	9.1%	0%	13.0%
PCAT Type 5	0.8%	0%	0.3%
PCAT Type 6	21.8%	0%	17.2%
PCAT Type 7	1.8%	100%	1.5%
LIB Type 1	39.5%	4.9%	23.8%
LIB Type 2	17.3%	88.3%	23.2%
LIB Type 3	43.3%	6.8%	53.0%
WDF-ML Type 0	89.5%	9.6%	86.9%
WDF-ML Type 1	5.9%	49.7%	7.0%
WDF-ML Type 2	4.6%	40.7%	6.1%

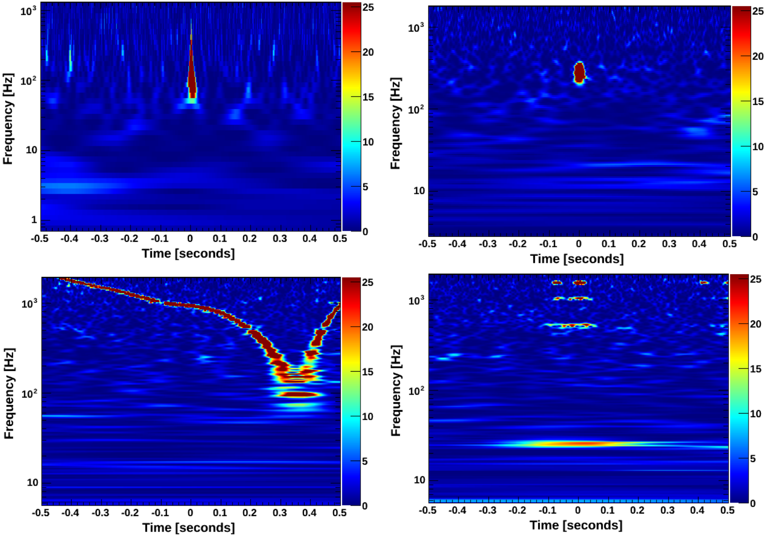
Classification methods for noise transients in advanced gravitational-wave detectors

Class. Quant. Grav., 32 (21), pp. 215012, 2015.

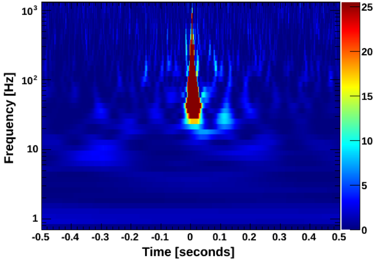
Real Data: ER7 LIGO data

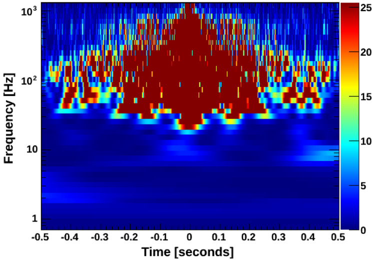
- Data from the 7th aLIGO engineering run (ER7), which began on the 3rd of June 2015 and finished on the 14th of June 2015. The average binary neutron star inspiral range for both Hanford and Livingston detectors in data analysis mode during ER7 was $50 - 60$ Mpc.
- The total length of Livingston data analysed is ~ 87 hours.
- The total length of Hanford data analysed is ~ 141 hours.

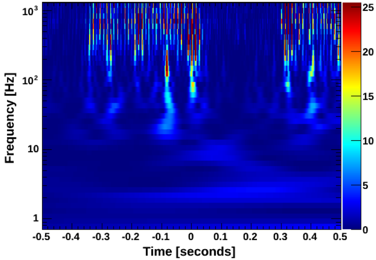
Real Data: ER7 L1

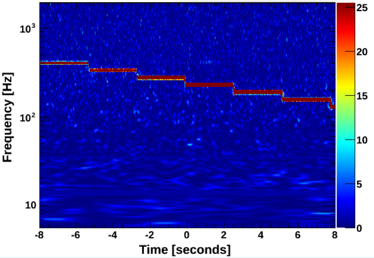


Real Data: ER7 H1









Conclusion

- Jade Powell labeled all the glitches and classified them by eye. This classification is used as reference.
- In the ER7 data from aLIGO Livingston PCAT missed 90 transients and classified 95% of the remaining transients correctly.
- PC-LIB missed 33 transients and classified 98% of the remaining transients correctly.
- WDF-ML classified all transients and 97% of them were correct.
- In aLIGO Hanford PCAT missed 120 transients and classified 99% of the remaining transients correctly.
- PC-LIB missed 6 transients and classified 95% of the remaining transients correctly.
- WDF-ML classified all transients and 92% of them were correct.
- We conclude that our methods have a high efficiency in real non-stationary and non-Gaussian detector noise.

Submitted:
Classification methods for noise transients in advanced gravitational-wave detectors II: performance tests on Advanced LIGO data. (by the authors)

What's next?

- Three different methods have been developed for the fast classification of noise transients.
- Transients are split in to types by waveform morphology first, and then can be split up in to further types by frequency and SNR.
- Results are similar for all methods.
- We plan to use Dictionary Based Algorithm.
- We plan to use Images Deep Learning Classification
- Next we plan on looking at how these codes perform when using data from multiple auxiliary channels.
- We are ready to apply WDF-ML to O1 and O2 run data.