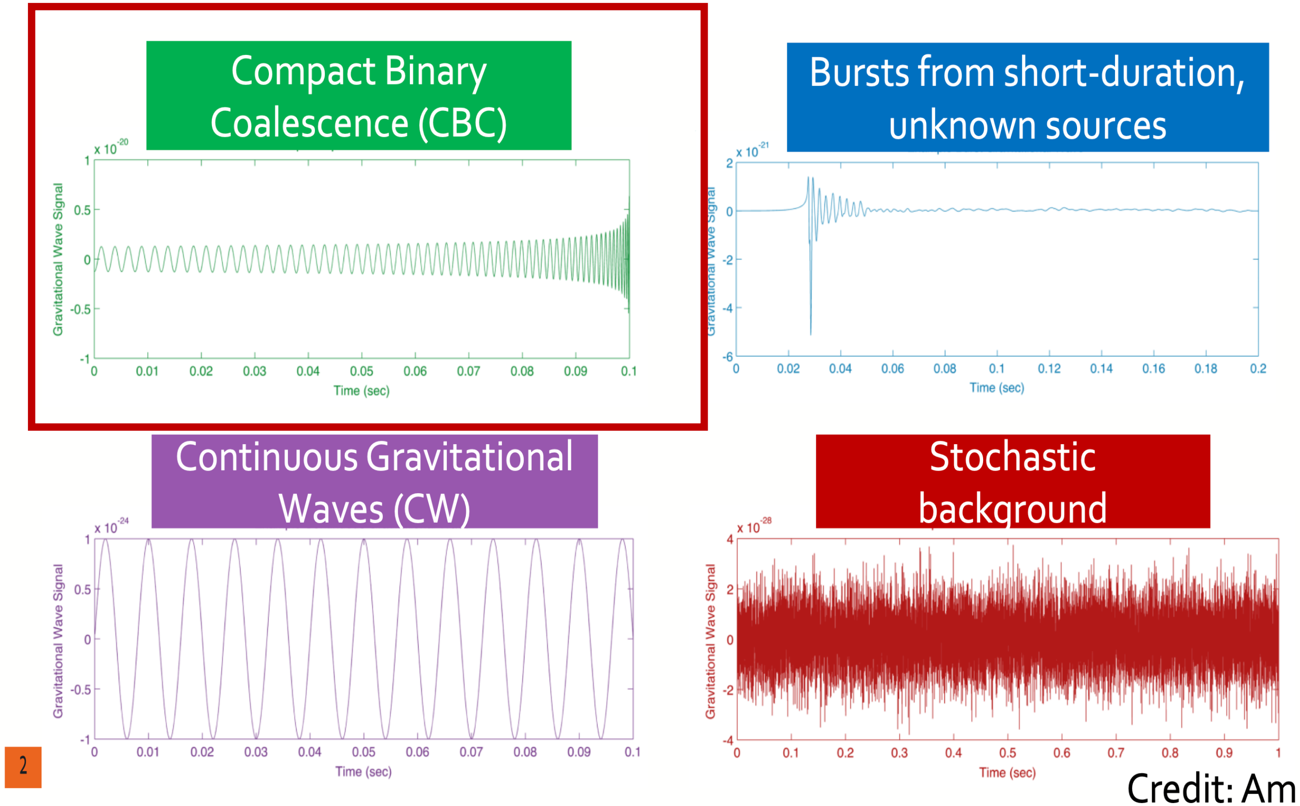
A brief Introduction to Data Analysis procedures in LVK

Ester Ruiz Morales

UPM & IFT-Madrid Virgo Group

- Data analysis is a huge topic!
- I present a very limited scope, reduced to:
  - CBC data
  - Mainly in off-line data processing for the GWTC.



Credit: Amber Stuver

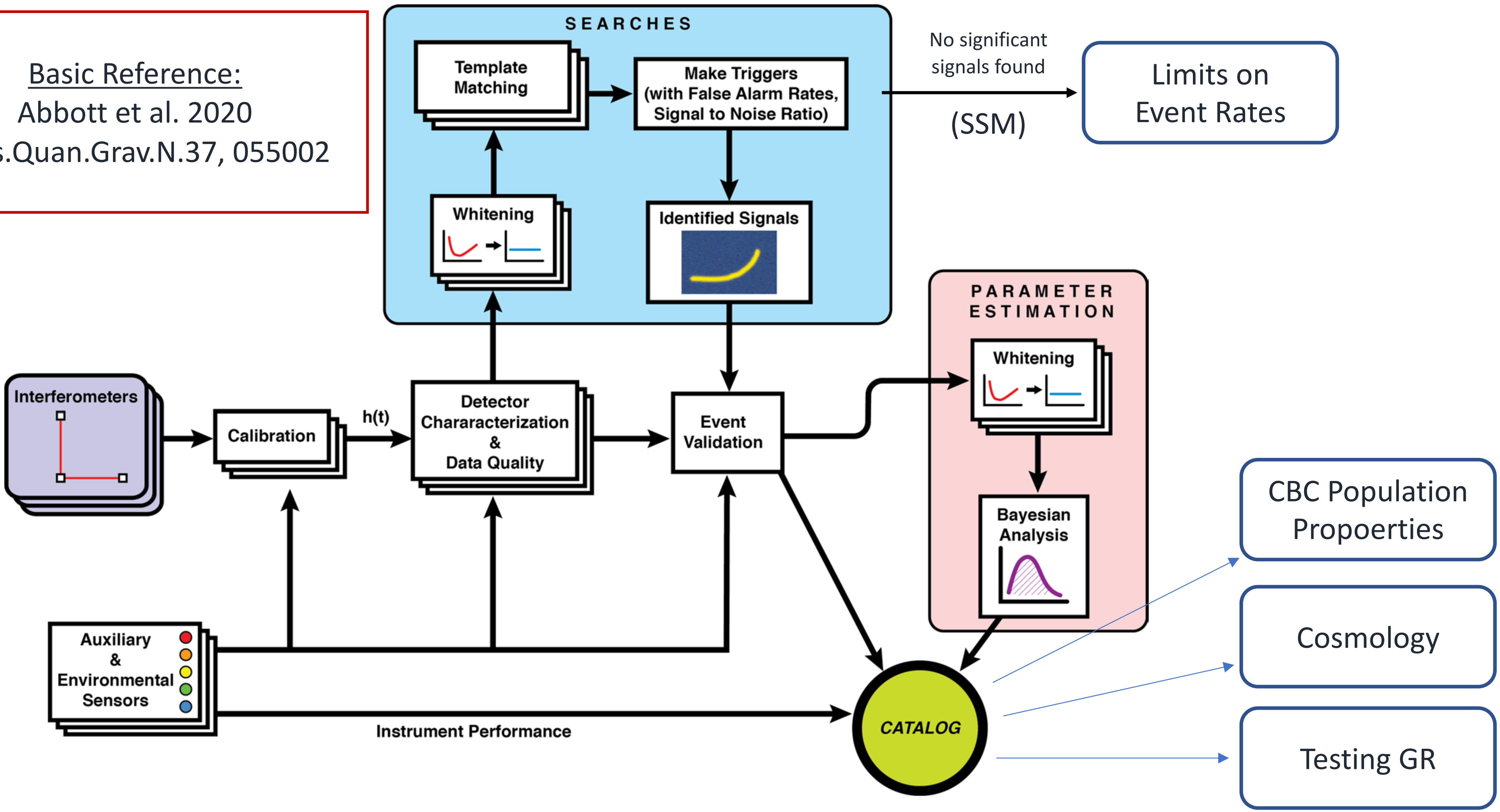
- For Low Latency information, see for instance Cardiff 22 LVK meeting slides:

https://dcc.ligo.org/DocDB/0184/G2201664/003/LowLatency_Plenary.pdf

- Describe the basic methods/procedures, main properties & limitations.
- Many of them are automatized, but I won’t describe that either.

Main steps in data processing

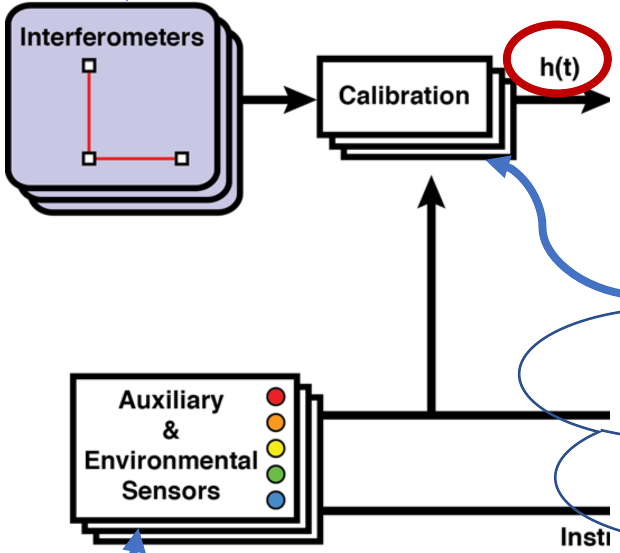
Basic Reference:
Abbott et al. 2020
Class.Quan.Grav.N.37, 055002



Data processing – Step 1: What is really measured?

<u>raw data:</u>
time-varying intensity of the laser light measured at the interferometer output

<u>data used for analysis:</u>
gravitational-wave strain amplitude d(t)

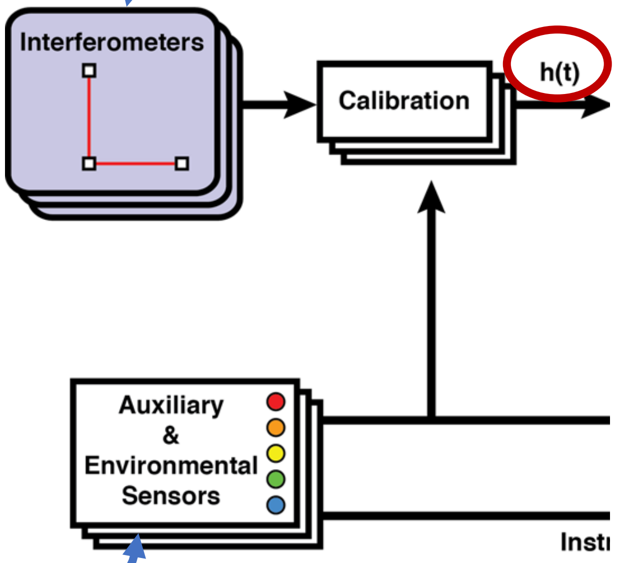


complex process ($\Delta L \sim 1\times10^{-19}$ m)
converts the changes of arm length difference into
the <u>time series of the strain</u> of a passing gravitational wave h(t),
sampled at ~ 16kHz,
Calibration valid in (10Hz – 5kHz) frequency range

Record the time series of the detector & environment state

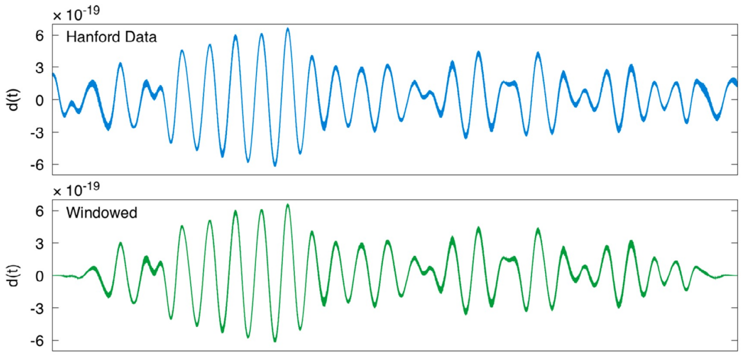
Data processing – Step 1: What is really measured?

<u>raw data:</u>
time-varying intensity of the laser light measured at the interferometer output



Record the time series of the detector & environment state

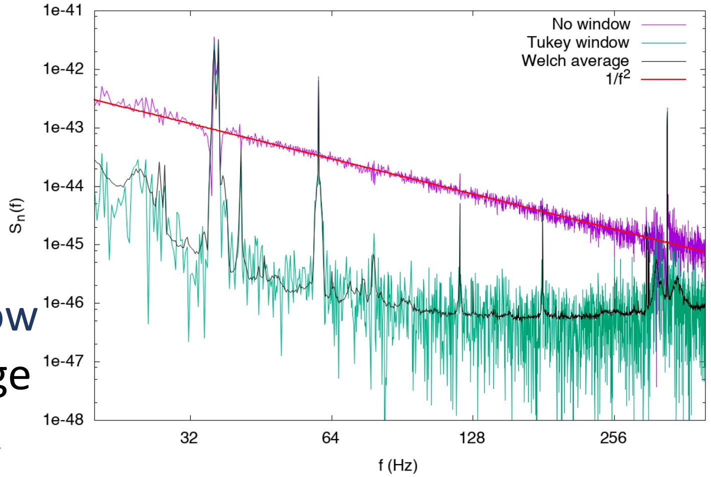
<u>data used for analysis:</u>
gravitational-wave strain amplitude d(t)



PSD of the noise is not known, must be estimated from data

Most of the strain amplitude is simply NOISE
ONLY Aprox. stationary
with a Gaussian component + glitches

FFT
Window
Average



PSD: Power Spectral Density

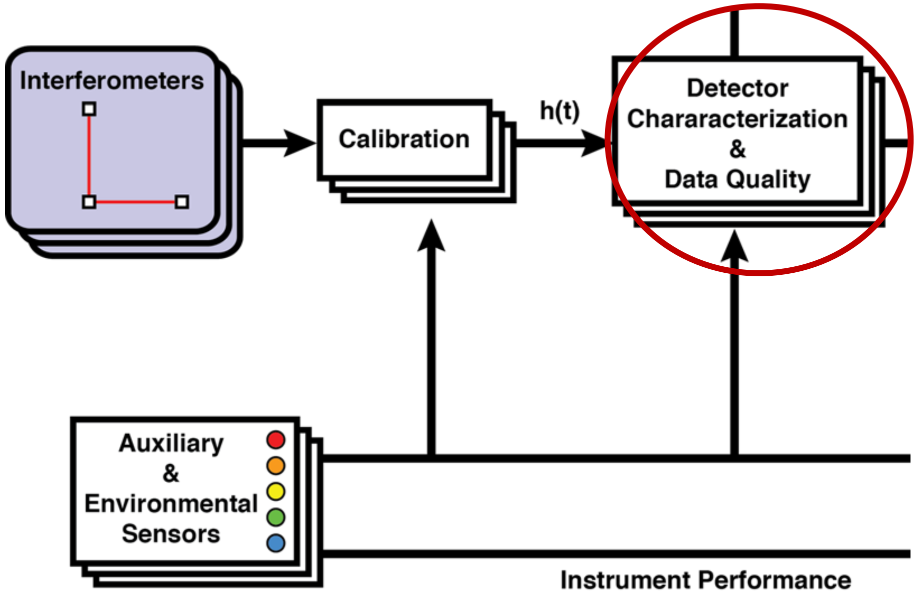
<u>ESSENTIAL TO CHARACTERIZE NOISE WELL</u>
PSD needed for the matched filtered search and PE

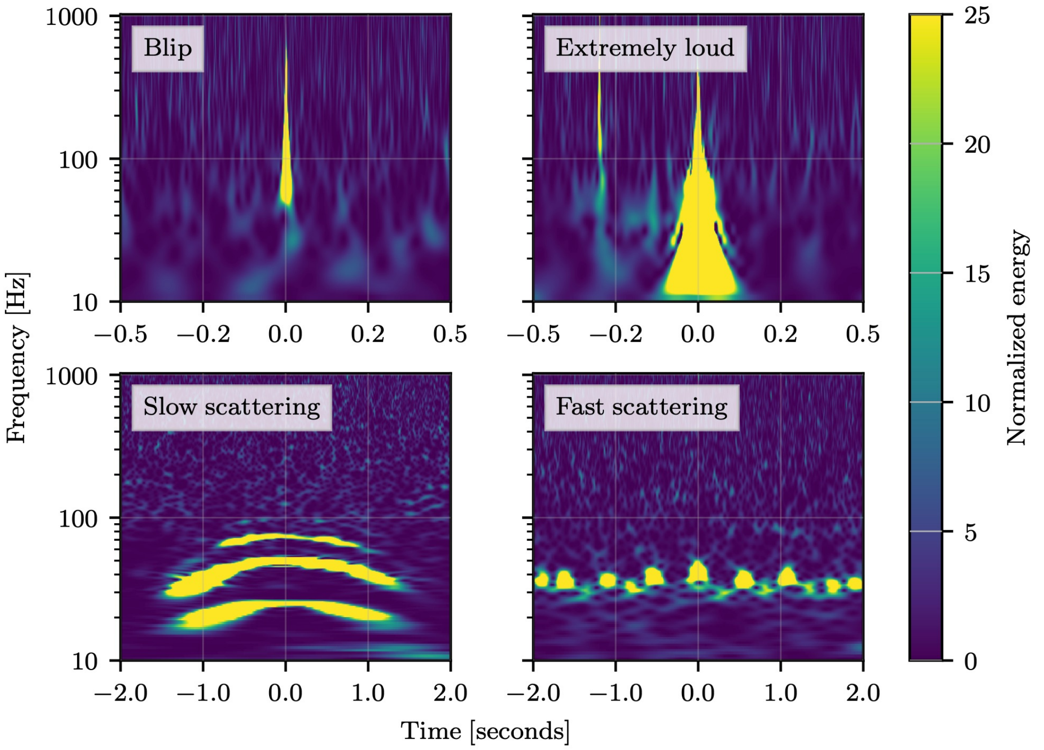
Data processing – Step 2: Check data quality

DetChar & DQ functions:

1.- Noise characterization and mitigation

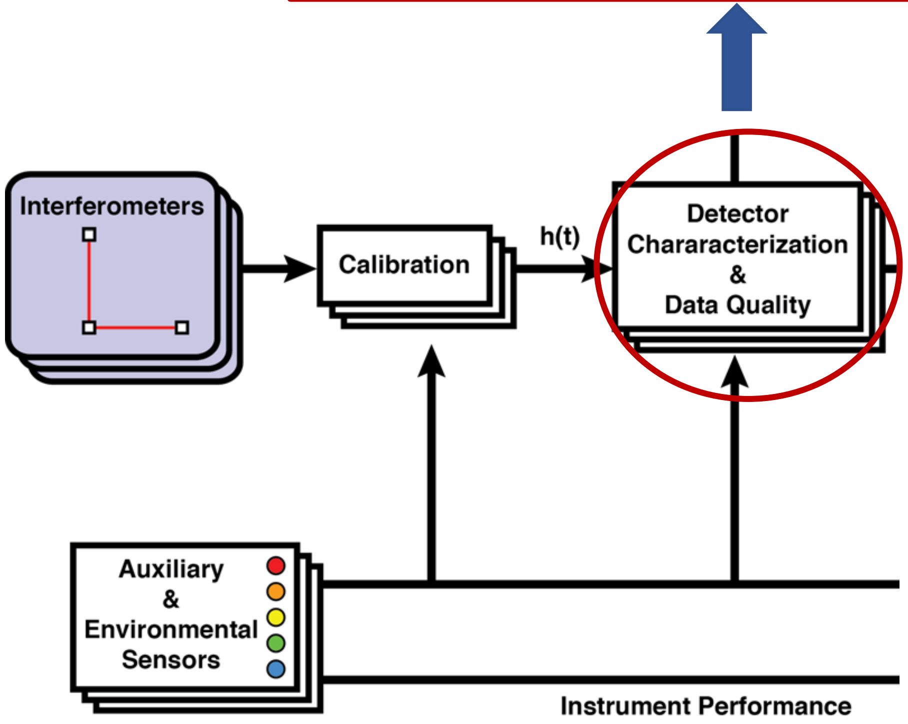
- Find sources of noise and work in on-site mitigation
- Characterization of transient noise in the detectors





Data processing – Step 2: Check data quality

<u>Output: Cat2 + gatedData</u>
ready for searches &
public access



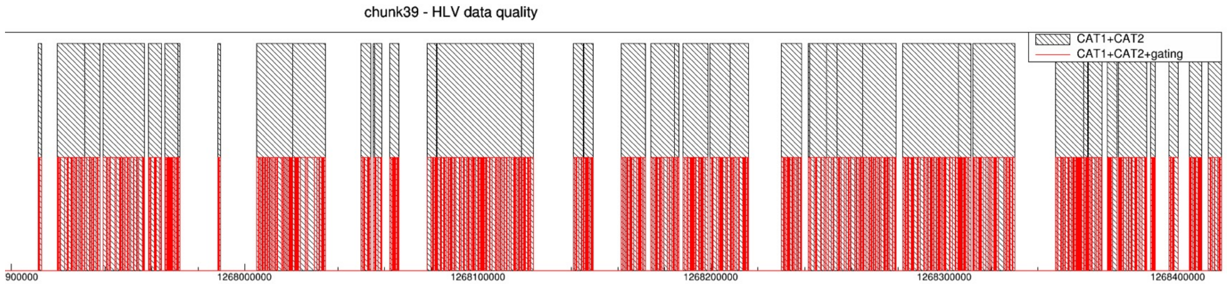
DetChar functions:

1.- Noise characterization and mitigation

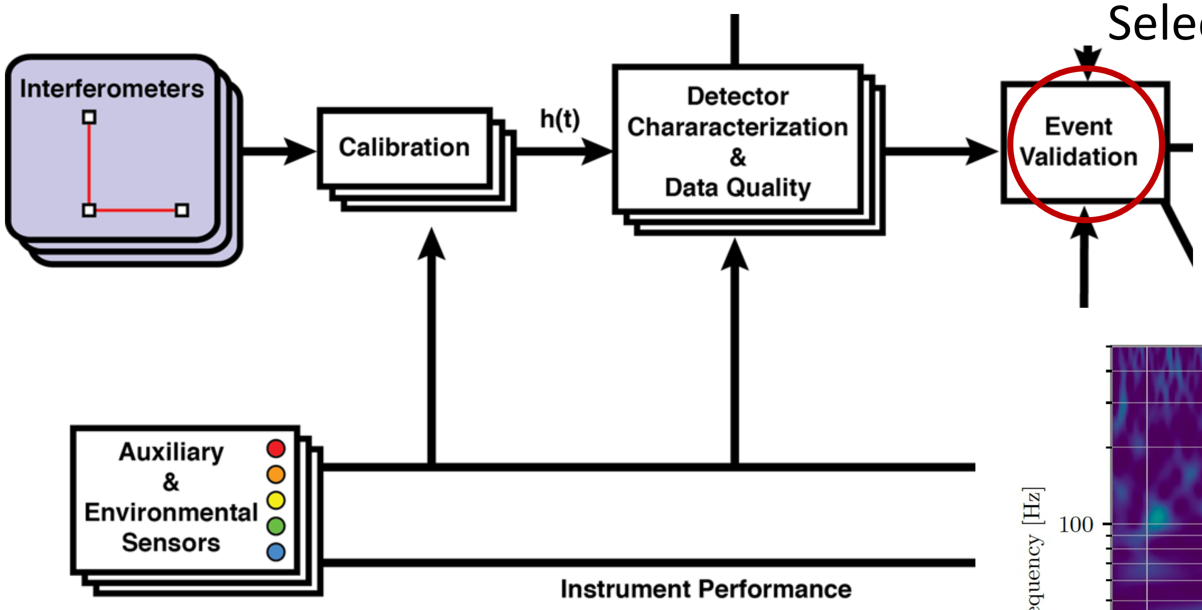
- Find sources of noise and work in on-site mitigation
- Characterization of transient noise in the detectors

2.- Data Quality vetoes on strain time series

Select t when detectors are working without problems



Data processing – Step 2: Check data quality



DetChar functions:

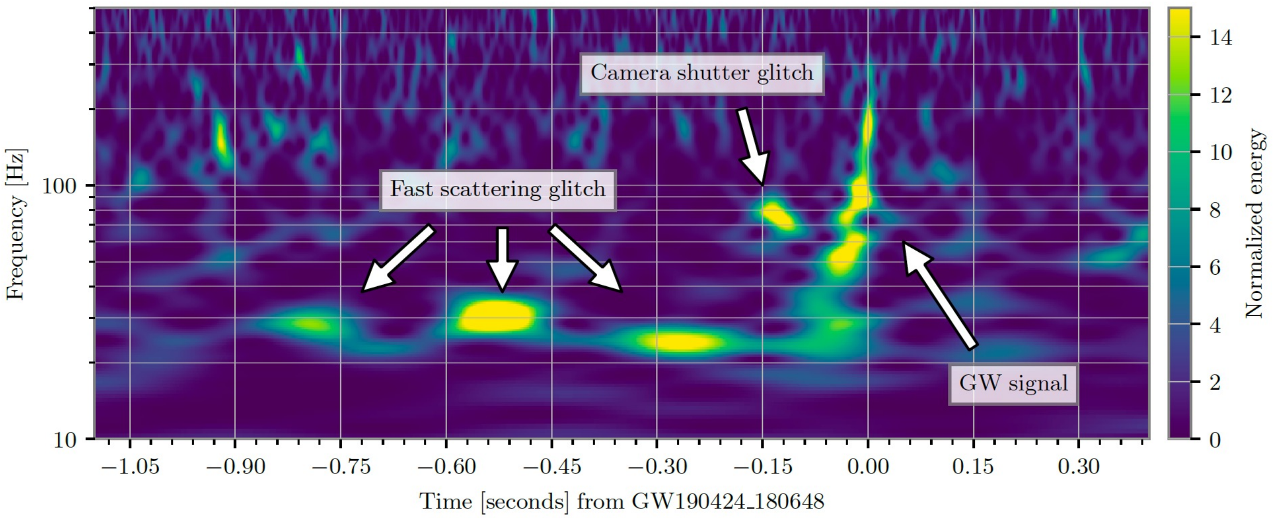
1.- Noise characterization and mitigation

- Find sources of noise and work in on-site mitigation
- Characterization of transient noise in the detectors

2.- Data Quality vetoes on strain time series

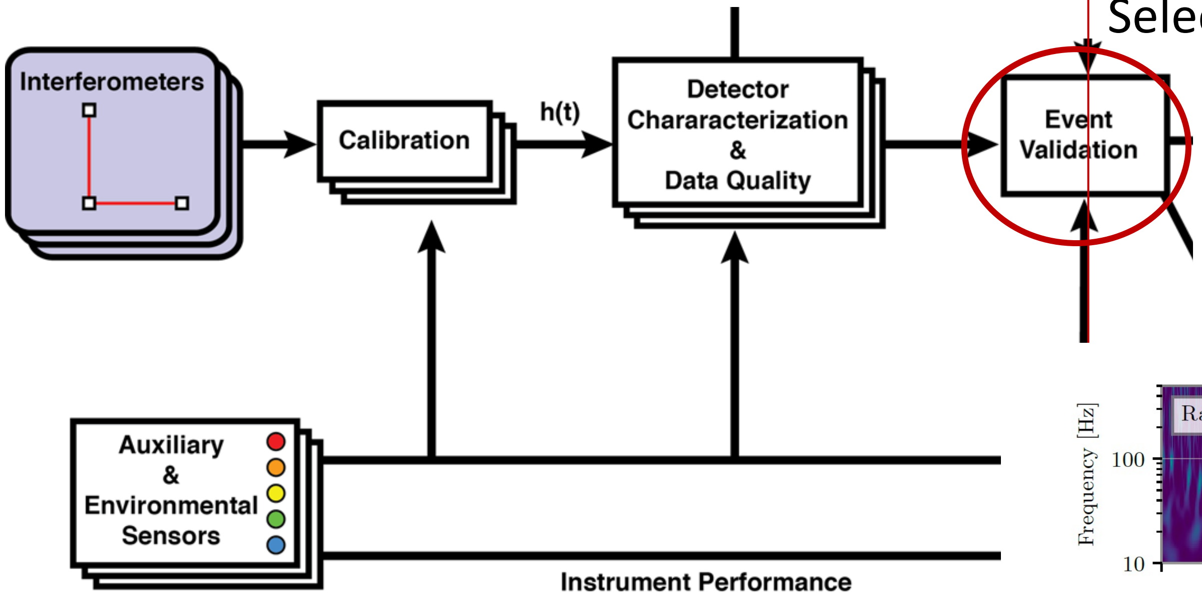
Select t when detectors are working without problems

3.- Event validation for candidates found by the search



Detector Characterization and Mitigation of Noise in Ground-Based Gravitational-Wave Interferometers
Published in: *Galaxies* 10 (2022) 1, 12

Data processing – Step 2: Check data quality



DetChar functions:

1.- Noise characterization and mitigation

- Find sources of noise and work in on-site mitigation
- Characterization of transient noise in the detectors

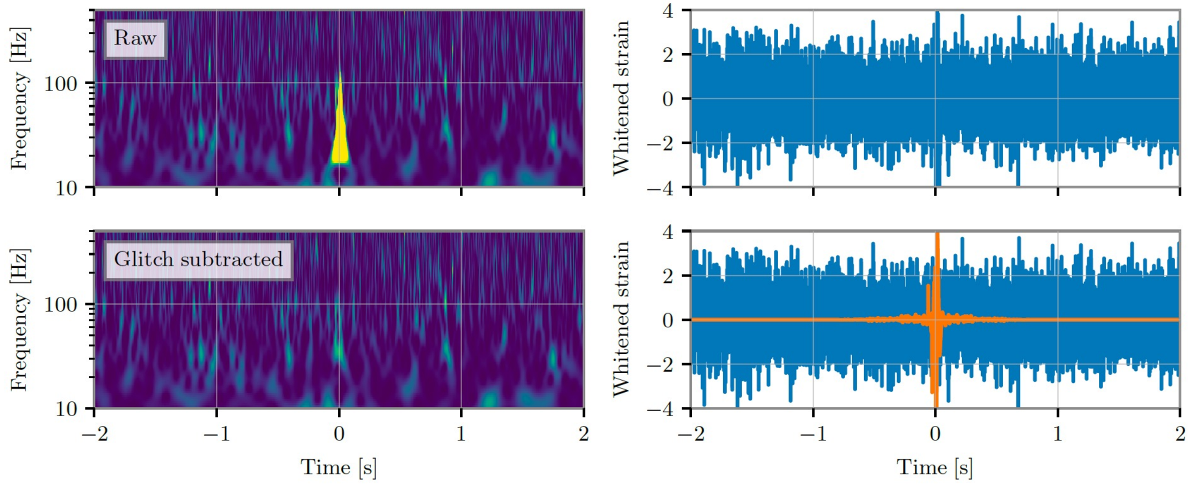
2.- Data Quality vetoes on strain time series

Select t when detectors are working without problems

3.- Event validation for candidates found by the search

4.- Noise substraction: deglitching, gating..

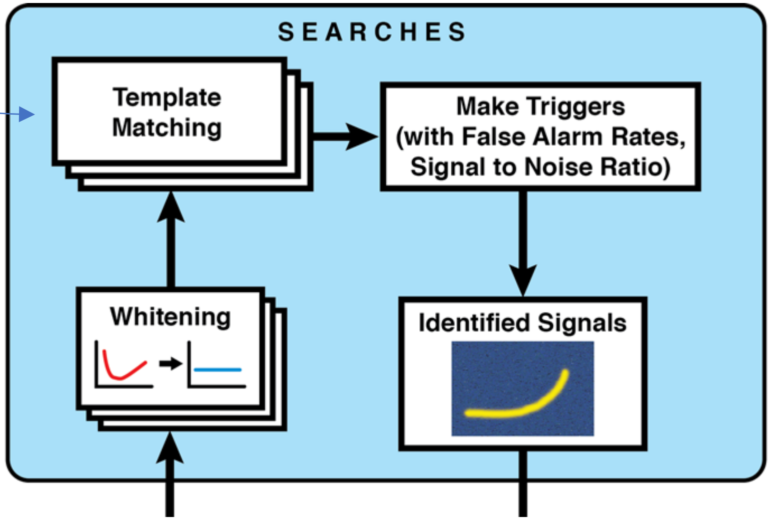
20% of events, "Delicate" processes…



Noise sustraction software: BayesWave
New glitch + signal modelling methods:
https://arxiv.org/abs/2205.13580

Data processing – Step 3: Matched filtered searches

Signal identification: Match filtering of d(t)



GraceDB

Candidate identified!

1.- A common bank of templates for filtering is designed for each category: BBH,NSBH & BNS.

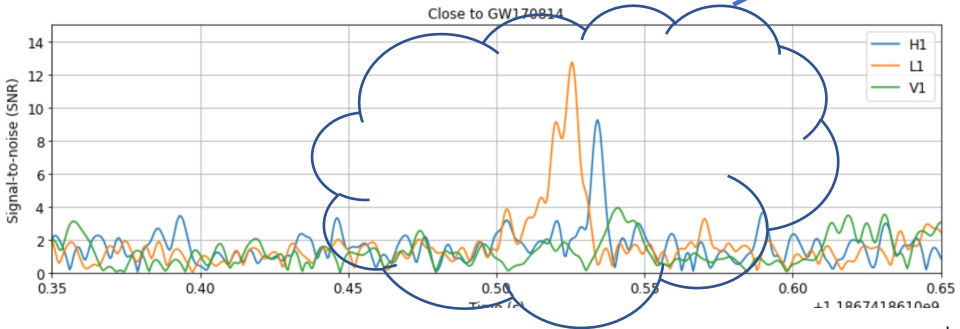
2.- For each template h(t) w. param μ

$$\mathbf{h}(\boldsymbol{\theta}) = A\mathbf{p}(t,\boldsymbol{\mu})\cos\phi + A\mathbf{q}(t,\boldsymbol{\mu})\sin\phi$$

calculate the SNR time series

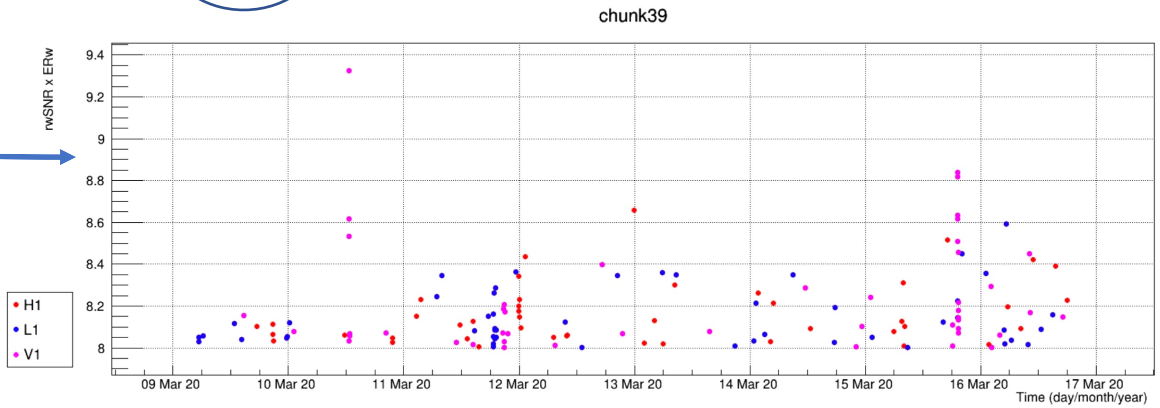
$$\rho(t,\boldsymbol{\mu}) \equiv \sqrt{(\mathbf{d}\mid\mathbf{p}(t,\boldsymbol{\mu}))^2 + (\mathbf{d}\mid\mathbf{q}(t,\boldsymbol{\mu}))^2}$$

$$(\mathbf{a}|\mathbf{b}) = 2\int_0^\infty \frac{\tilde{a}(f)\tilde{b}^*(f) + \tilde{a}^*(f)\tilde{b}(f)}{S_n(f)}\,\mathrm{d}f\,.$$



SNR quantifies the likelihood that the observed data contains a GW signal.

3.- As result of filtering, we get a collection of triggers:



4.- Cluster triggers in time in each detector, choose the most representative, look for coincident triggers in more than one detector to select signal candidates.

Data processing – Step 3': Event Ranking

SEARCHES

We get a HUGE amount of triggers, most of them caused by NOISE !!

Need to solve 2 problems:

1.- How do we rank the triggers to select the most "signal like" ones?

- SNR is optimal ranking statistics in Gaussian Noise: the higher the SNR the higher the likelihood that data contains a signal

- But non-gaussian glitches produce HIGH SNR!

- A reweighted-SNR (using chi^2 methods) is used as RE

- Each pileline uses his own RE.

2.- How do we assign an statistical significance to each candidate event in terms of its ranking statistic?

Data processing – Step 3: Event Ranking

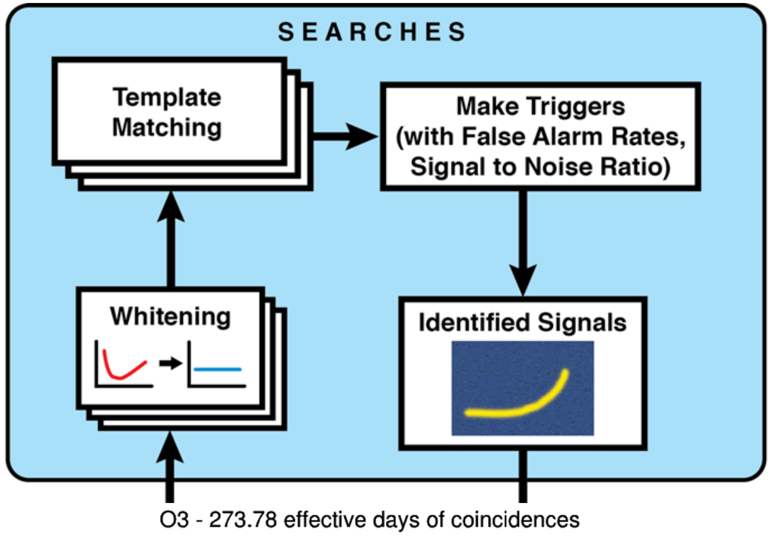
1.- Assign a FAR value to the event = the rate of background triggers with ranking statistics value equal to or greater than the RE of the event.

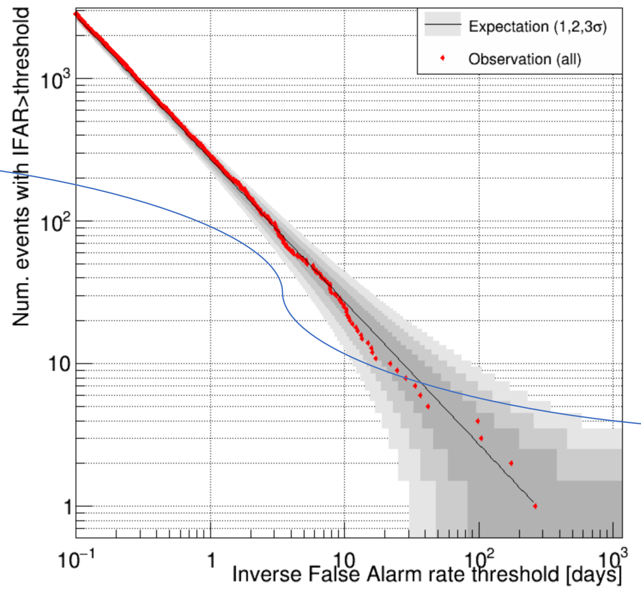
2.- The background distribution of the ranking statistic is estimated in a data driven way, by running the search over time-shifted detector data, so that coincidences become not physical.

3.- Plot the cumulative FAR distribution for background and data, outstanding events would clearly appear, as in GWCT1 plots

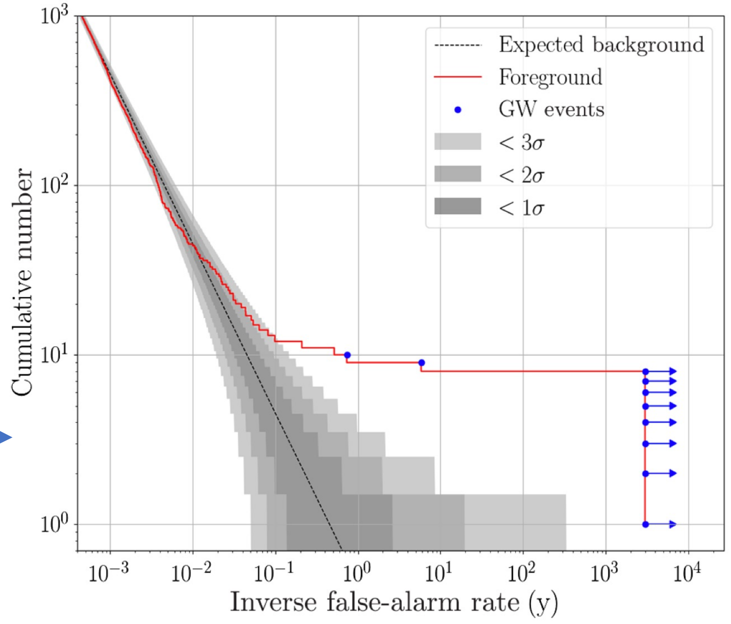
4.- The FARxTobs = estimate of the probability of there being at least one noise trigger with a FAR this low or lower in the observed time.

Typically: FAR< 2/year for O3a/b runs





MBTA in SSMO3b



PyCBC in GWTC1

Data processing – Step 3: Event Ranking

Problem:

Since FAR depends exponentially on the RE values, FAR values can differ by orders of magnitude among pipelines!

Name	Inst.	cWB FAR (yr^{-1})	cWB SNR	cWB p_{astro}	GstLAL FAR (yr^{-1})	GstLAL SNR	GstLAL p_{astro}	MBTA FAR (yr^{-1})	MBTA SNR	MBTA p_{astro}	PyCBC-broad FAR (yr^{-1})	PyCBC-broad SNR	PyCBC-broad p_{astro}	PyCBC-BBH FAR (yr^{-1})	PyCBC-BBH SNR	PyCBC-BBH p_{astro}
GW191103_012549	HL	–	–	–	–	–	–	27	9.0	0.13	4.8	9.3	0.77	0.46	9.3	0.94
GW191105_143521	HLV	–	–	–	24	10.0	0.07	0.14	10.7	> 0.99	0.012	9.8	> 0.99	0.036	9.8	> 0.99
GW191109_010717	HL	< 0.0011	15.6	> 0.99	0.0010	15.8	> 0.99	1.8×10^{-4}	15.2	> 0.99	0.096	13.2	> 0.99	0.047	14.4	> 0.99
GW191113_071753	HLV	–	–	–	–	–	–	26	9.2	0.68	1.1×10^{4}	8.3	< 0.01	1.2×10^{3}	8.5	< 0.01
GW191126_115259	HL	–	–	–	80	8.7	0.02	59	8.5	0.30	22	8.5	0.39	3.2	8.5	0.70
GW191127_050227	HLV	–	–	–	0.25	10.3	0.49	1.2	9.8	0.73	20	9.5	0.47	4.1	8.7	0.74
GW191129_134029	HL	–	–	–	$< 1.0 \times 10^{-5}$	13.3	> 0.99	0.013	12.7	> 0.99	$< 2.6 \times 10^{-5}$	12.9	> 0.99	$< 2.4 \times 10^{-5}$	12.9	> 0.99
GW191204_110529	HL	–	–	–	21	9.0	0.07	1.3×10^{4}	8.1	< 0.01	980	8.9	< 0.01	3.3	8.9	0.74
GW191204_171526	HL	$< 8.7 \times 10^{-4}$	17.1	> 0.99	$< 1.0 \times 10^{-5}$	15.6	> 0.99	$< 1.0 \times 10^{-5}$	17.1	> 0.99	$< 1.4 \times 10^{-5}$	16.9	> 0.99	$< 1.2 \times 10^{-5}$	16.9	> 0.99
GW191215_223052	HLV	0.12	9.8	0.95	$< 1.0 \times 10^{-5}$	10.9	> 0.99	0.22	10.8	> 0.99	0.0016	10.3	> 0.99	0.28	10.2	> 0.99
GW191216_213338	HV	–	–	–	$< 1.0 \times 10^{-5}$	18.6	> 0.99	9.3×10^{-4}	17.9	> 0.99	0.0019	18.3	> 0.99	7.6×10^{-4}	18.3	> 0.99
GW191219_163120	HLV	–	–	–	–	–	–	–	–	–	4.0	8.9	0.82	–	–	–

From GWTC3

Data processing – Step 3: Event Ranking

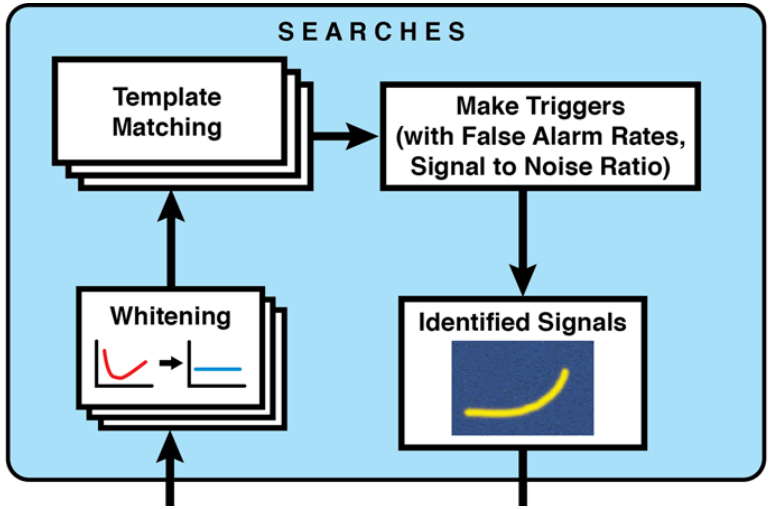
1.- Assign a FAR value to the event = the rate of background triggers with ranking statistics value equal to or greater than the RE of the event.

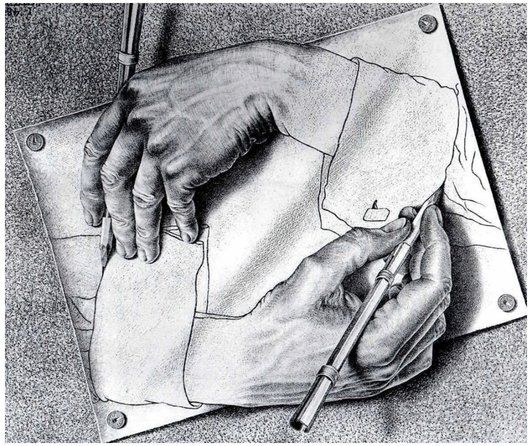
2.- The background distribution of the ranking statistic is estimated in a data driven way, by running the search over time-shifted detector data, so that coincidences become not physical.

3.- Plot the cumulative FAR distribution for background and data, outstanding events would clearly appear, as in GWCT1 plots

4.- The FARxTobs = estimate of the probability of there being at least one noise trigger with a FAR this low or lower in the observed time.

Typically: FAR< 2/year for O3a/b runs





¿Does Pastro have this risk?

- Present situation:

FAR eliminated as statistical criteria from the catalogs after GWTC1.

For an evaluation of the FAR from Gaussian Noise in GD Detectors, see our recent paper: arXiv:2209.05475

- Since GWTC2.1 Pastro introduced to quantify the "probability of astrophysical origin".

Characteristics:
- Event goes to catalog if Pastro >0.5.
- Each pipeline computes its own.
- Based on prior knoledge of the population properties and rates!

Main steps in data processing

See Jose Francisco's talk for the rest of topics!

Thank You!