GdR ondes gravitationnelles
Rencontre du groupe de travail “méthodes d’analyse des données”
2024-10-16

Analysis of ground-based detector data with a focus on matched filtering

Tito Dal Canton

Modeling the data of a ground-based GW detector

Continuous time series of spacetime strain measurement, sampled at ~10 kHz, contaminated with noise:

$$s(t) = n_{\mathrm{easy}}(t) + n_{\mathrm{hard}}(t) + \sum_i h_i(t; \vec{\lambda}_i)$$

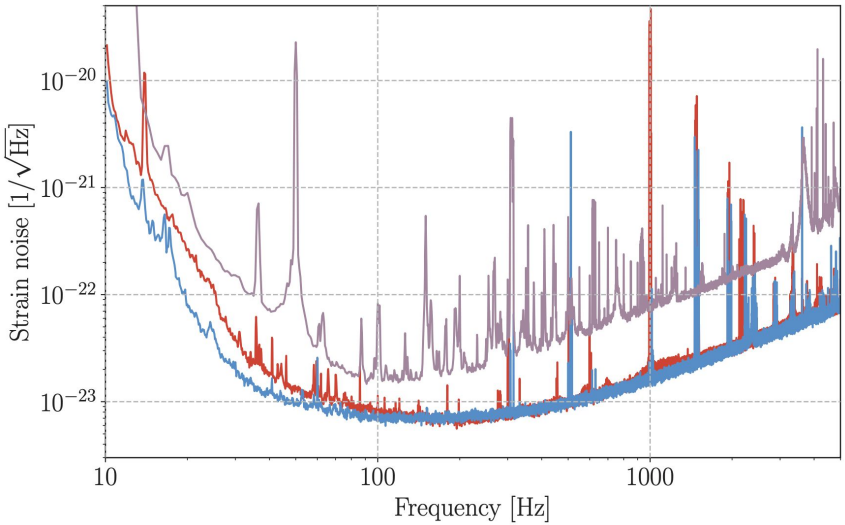
Detector noise that is easy to predict or model. Usually fundamental, and determines the sensitivity of the detector.

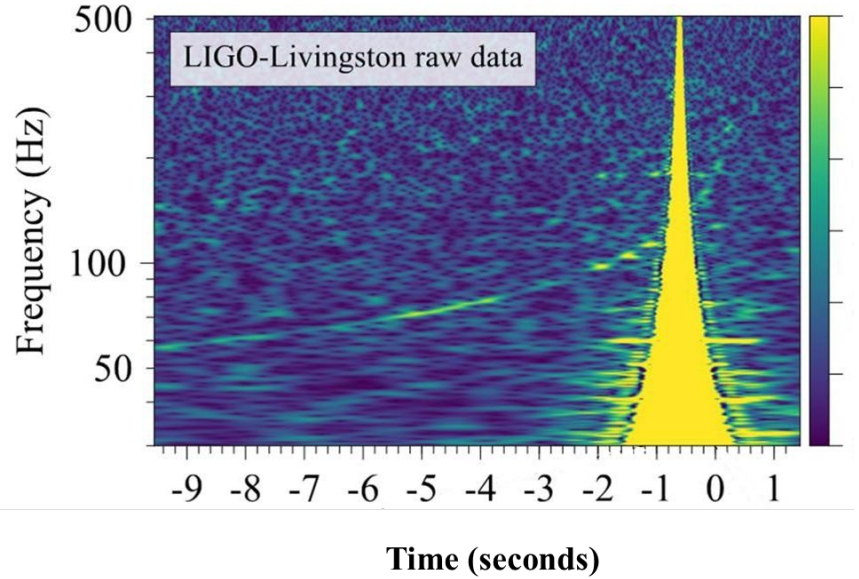
Detector/environment noise that is hard or impossible to predict or model. Usually technical, and determines the quality of the data.

Superposition of all gravitational-wave signals, each with its own vector of parameters describing the source and possibly larger structures, or even the whole Universe.

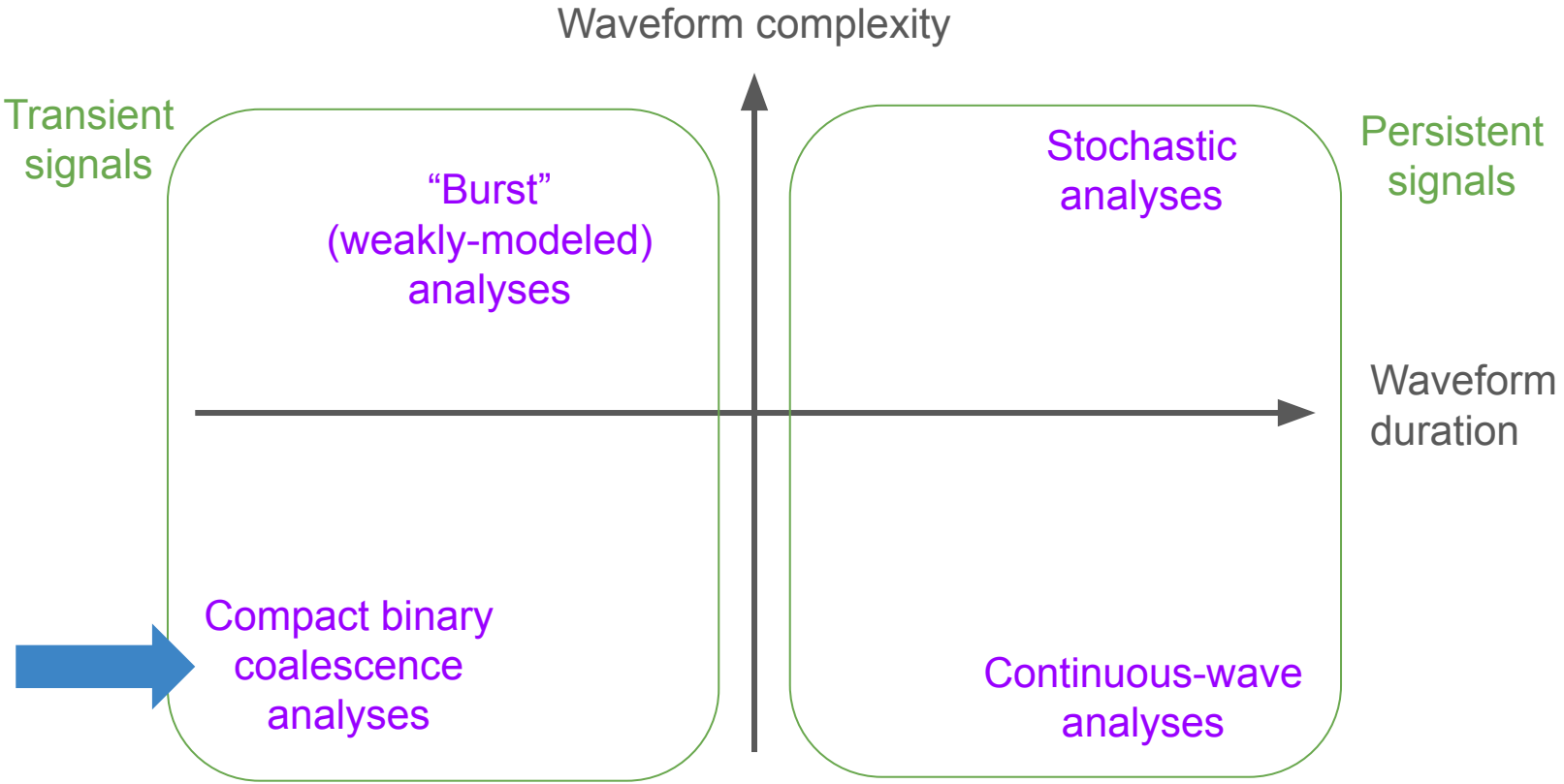
$$h = (x^i x^j - y^i y^j) h_{ij} = D^{ij} h_{ij} = F_+ h_+ + F_\times h_\times$$

Modeling the data of a ground-based GW detector





Taxonomy of ground-based data analysis methods



Common assumptions for ground-based transient searches

1. Separable signals:
   the duration of a signal is much shorter than the time between signals (on average).
   → Separate the analysis problem into
   detection (search) and parameter estimation

2. Fixed detector geometry:
   the rotation of the Earth is negligible during the duration of a signal.
   → Detector response is just a constant scaling factor in the waveform model.

3. (In the case of CBC searches) simplified waveform model:
   neglect precession, higher-order modes, eccentricity & tidal effects.
   Signal can be expressed as $h(t;\vec{\theta}) = A(t;\vec{\theta})\cos\Psi(t;\vec{\theta})$ or $\tilde{h}(f;\vec{\theta}) = B(t;\vec{\theta})\exp\left[i\Phi(t;\vec{\theta})\right]$

Maximum likelihood formalism

For the first identification of an unknown GW transient, we use a maximum likelihood approach:

$$\frac{P(M_A|d)}{P(M_B|d)} = \frac{\mathcal{Z}(d|M_A)}{\mathcal{Z}(d|M_B)} \frac{\pi(M_A)}{\pi(M_B)} \longrightarrow \mathcal{R} = \frac{\max_{\theta\in\Theta} \mathcal{L}}{\max_{\theta\in\Theta_0} \mathcal{L}}$$

Maximum likelihood ratio

A: noise + signal hypothesis
B: noise only (null) hypothesis

$$p(\theta|d,M) = \frac{\mathcal{L}(d|\theta,M)\ \pi(\theta|M)}{\mathcal{Z}(d|M)} \longrightarrow \begin{cases} \hat{\theta} = \arg\max \mathcal{L} \\ \boldsymbol{\Sigma}_\theta = \left. \dfrac{\partial^2 \mathcal{L}}{\partial\theta_i \partial\theta_j} \right|_{\hat{\theta}} \end{cases}$$

Maximum likelihood parameter estimate

Covariance matrix

Solving the ML problem gives only an approximate answer to the original problem.
In many cases not a very good one.

The signal + Gaussian noise likelihood function

Data model $\tilde{\boldsymbol{s}} = \tilde{\boldsymbol{n}} + \tilde{\boldsymbol{h}}$ with stationary Gaussian noise $P(\tilde{\boldsymbol{n}}) \approx \dfrac{\exp\left(-\frac{1}{2}\sum_i \tilde{n}_i^2/\sigma_i^2\right)}{\sqrt{(2\pi)^N \prod_i \sigma_i^2}}$

$$\mathcal{L}(\boldsymbol{s}|\boldsymbol{h}) \propto \exp\left(-\frac{1}{2}\sum_i (\tilde{s}_i - \tilde{h}_i)(\tilde{s}_i - \tilde{h}_i)^* \sigma_i^{-2}\right)$$ Whittle likelihood

Assume known noise PSD → The only free parameters are signal parameters

ML ratio becomes $\mathcal{R} = \dfrac{\max_h \mathcal{L}(\tilde{s}|h)}{\mathcal{L}(\tilde{s}|h=0)}$ simpler in log: $\ln \mathcal{R} = \max_h \left(\langle s|h\rangle - \frac{1}{2}\langle h|h\rangle\right)$

with the noise-weighted inner product between discrete-time signals *a* and *b* $\langle a|b\rangle = \sum_i \dfrac{\tilde{a}_i \tilde{b}_i^*}{\sigma_i^2}$

(Note that I am not being very careful with constant factors in these expressions)

Matched filtering and signal-to-noise ratio

Want to maximize $\ln \mathcal{R} = \max_h \left(\langle s|h \rangle - \frac{1}{2} \langle h|h \rangle \right)$. Re-express the signal as $h = a h_{\rm norm}$

Then maximizing over *a* has a closed-form solution:

$$\ln \mathcal{R} = \max_{h_{\rm norm}} \left(\frac{1}{2} \frac{\langle s|h_{\rm norm} \rangle^2}{\langle h_{\rm norm}|h_{\rm norm} \rangle} \right) = \max_{h_{\rm norm}} \left(\frac{1}{2} \rho^2 \right)$$

with the (amplitude-maximized) signal-to-noise ratio (SNR)

$$\rho = \frac{\langle s|h_{\rm norm} \rangle}{\langle h_{\rm norm}|h_{\rm norm} \rangle^{1/2}} = \frac{\langle s|h \rangle}{\sqrt{\langle h|h \rangle}}$$

Template waveform

Matched filter

Maximizing over an overall phase shift is also possible if we use instead two templates that differ by a 90 deg phase rotation:

$$\rho = \frac{\sqrt{\langle s|h_{\rm I} \rangle^2 + \langle s|h_{\rm Q} \rangle^2}}{\sqrt{\langle h|h \rangle}}$$

← This is what people usually mean with "SNR" in LVK analyses

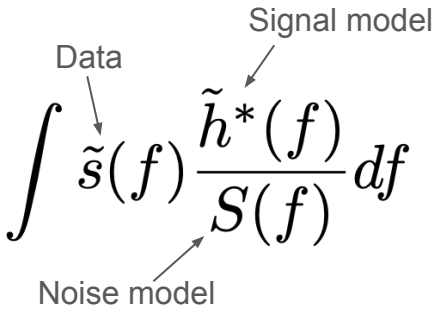
Matched filtering and signal-to-noise ratio

Signal-processing interpretation(s)

- Linear filter optimized to estimate the amplitude of a signal with known shape (hence *matched* filter)
- Correlation between the whitened data and the whitened template

Geometric interpretation

- N-dimensional inner product with a unit vector



$$\int \tilde{s}(f) \frac{\tilde{h}^*(f)}{S(f)} df$$

$$\int \frac{\tilde{s}(f)}{S^{1/2}(f)} \frac{\tilde{h}^*(f)}{S^{1/2}(f)} df$$

$$\vec{s} \cdot \hat{h}$$

Interpreting the signal-to-noise ratio

Remember that ultimately we are computing the LLR between “Gaussian noise + signal h” and “Gaussian noise only”.

How does the SNR behave if there is no signal?

Each <s|h> term is a unit-norm linear filter applied to Gaussian whitened data
→ Normal random variate

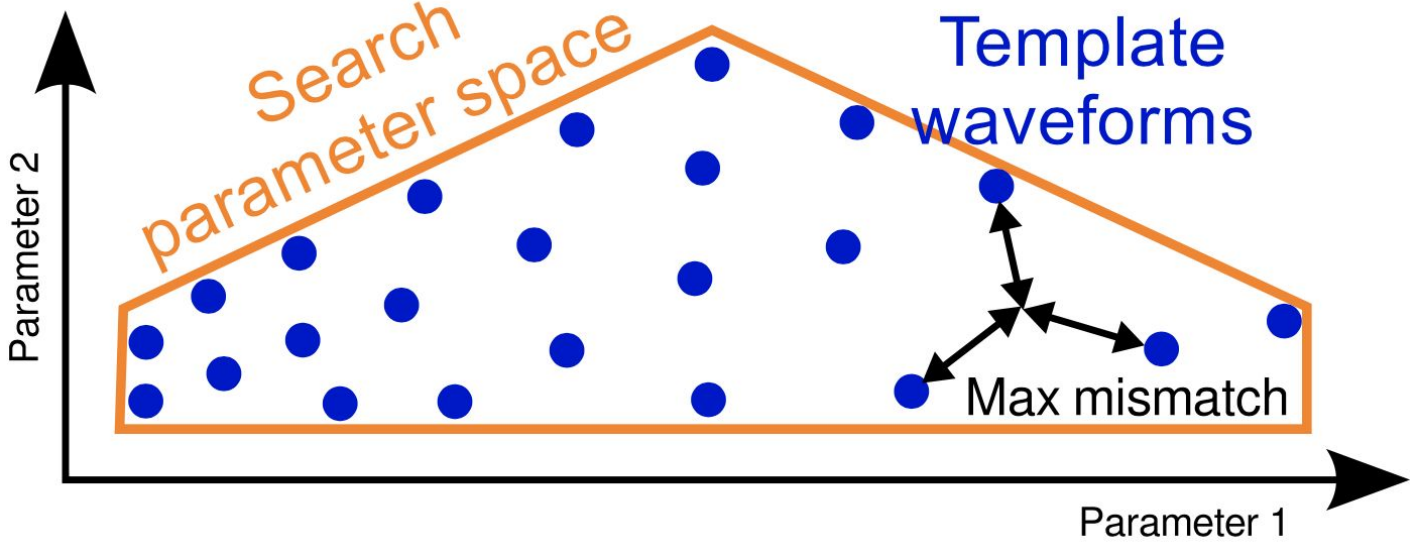
→ ϱ^2 distributed as a central χ^2 random variate → ϱ is on average ~1 far from the signals

Maximize the SNR ↔ Maximize the LLR for Gaussian noise

→ As SNR grows above some $\mathrm{SNR}_{\mathrm{min}}$ (typically 4-8) the null hypothesis for Gaussian noise becomes less and less likely.

Assuming that the signals are well separated, finding the local maxima of the SNR over the remaining parameters will point out the (strongest) signals.
This generally requires a numerical search.

*Mis*matched filtering: the template bank



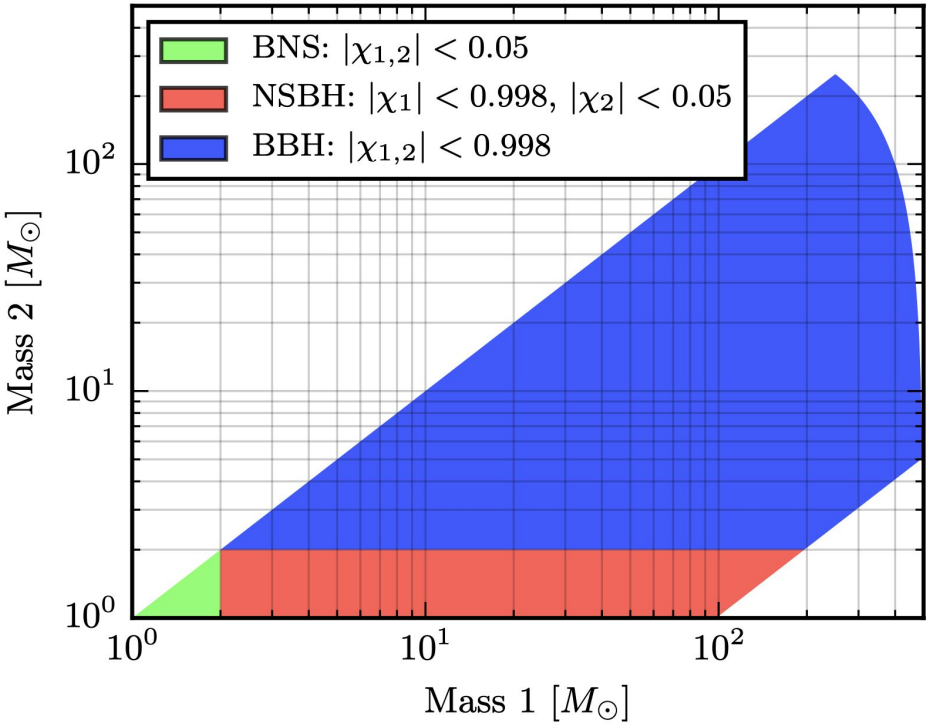
Discrete bank
Approximations
…

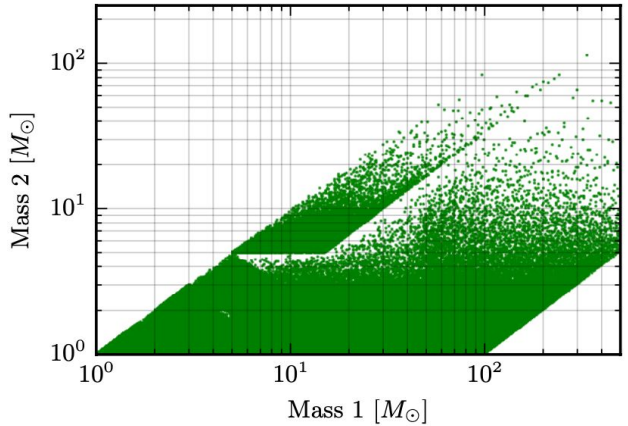
→ Loss of SNR $\varphi := \frac{\rho_{\text{observed}}}{\rho_{\text{optimal}}} < 1$ Fitting factor

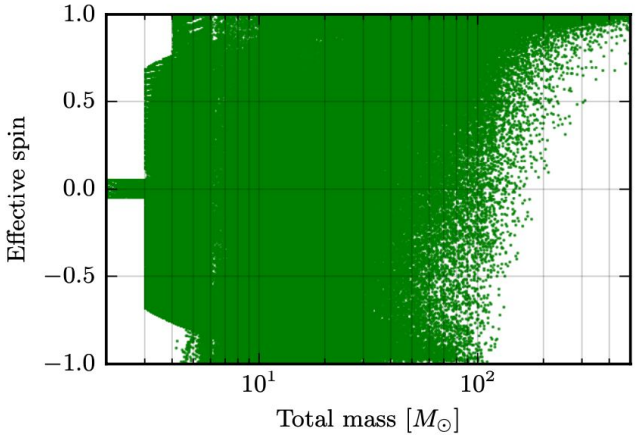
→ Loss of CBC detections $\frac{n}{n_{\text{optimal}}} \propto \varphi^3$

→ Bank must achieve $\varphi \approx 1$

*Mis*matched filtering: the template bank

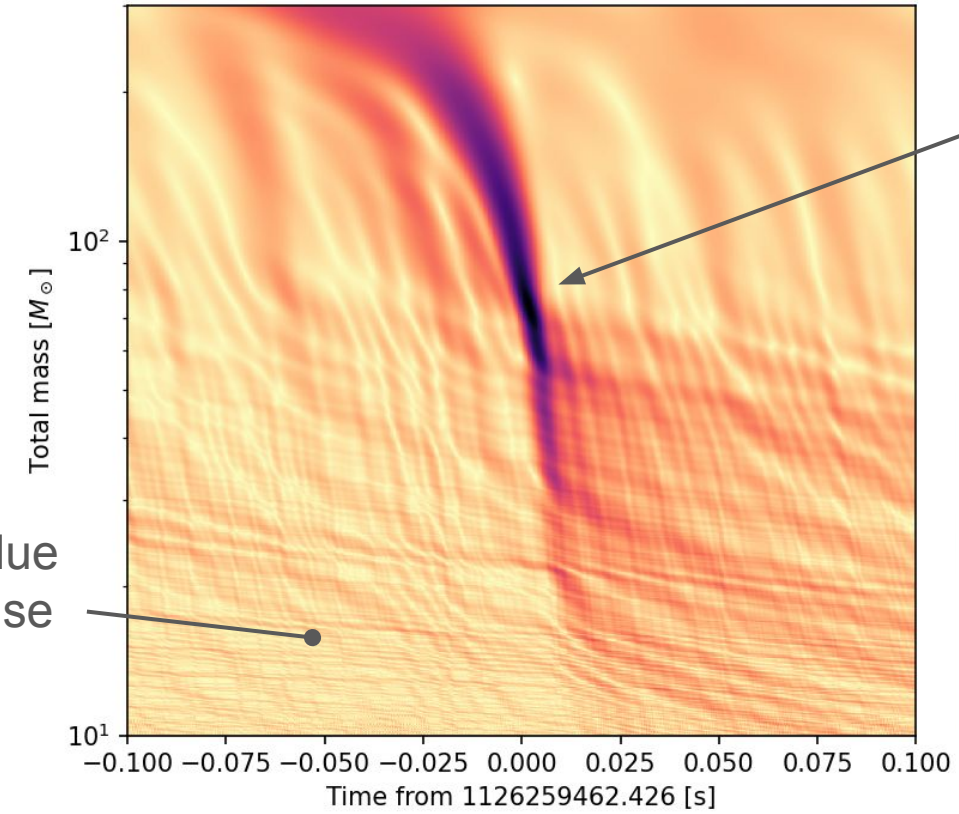






10^5-10^6 templates for CBC searches with LIGO/Virgo/KAGRA

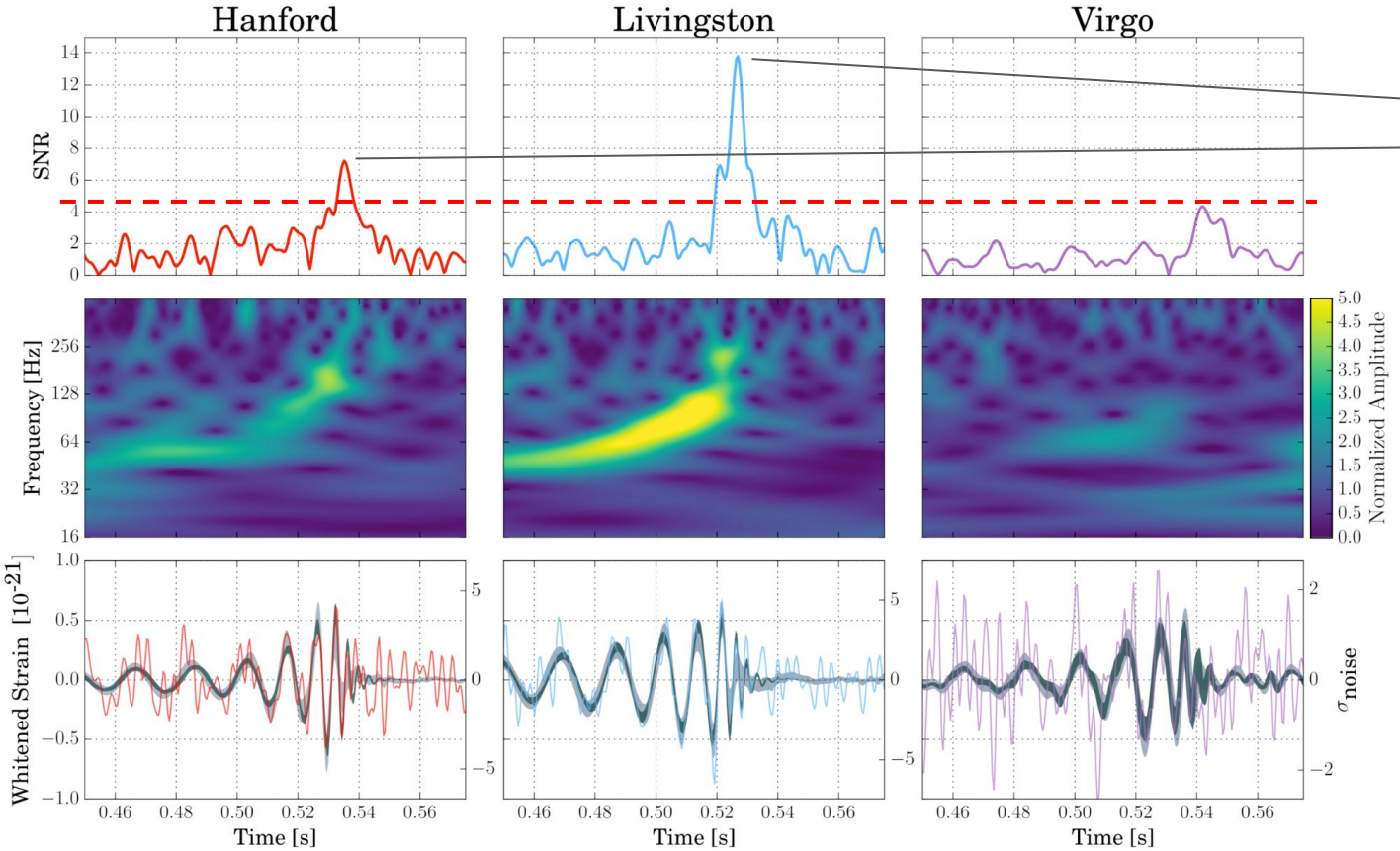
Projection of the SNR profile over two search parameters



Local maximum due to a signal (SNR ~ 10)

Local maxima due to Gaussian noise (SNR ~ 1)

From the SNR to candidate events



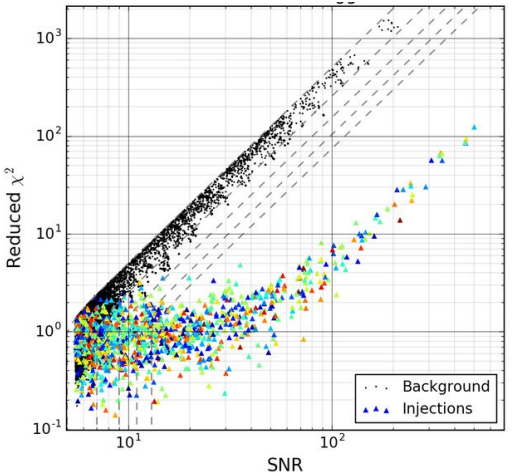
Dealing with “hard” noise post-facto

Data — Template waveform

$$\rho \propto \int \frac{\tilde{s}(f)\tilde{h}^{*}(f)}{S_{\mathrm{n}}(f)}\,\mathrm{d}f$$

Noise spectral density

SNR “proportional to the data”
→ Local fluctuations of the noise will reflect in the SNR
→ Large SNR no longer implies a trigger is astrophysical



Solution: signal-based discrimination statistics

- Time-frequency χ^2: check distribution of SNR over frequency
- Autocorrelation χ^2: check shape of SNR peak over merger time
- Bank χ^2: check shape of SNR peak over template parameters

Common statistical property:

- Distributed like a central χ^2 under Gaussian noise
  or Gaussian noise + matched signal
- Distributed like a noncentral χ^2 under Gaussian noise + mismatched signal

Combining data from different detectors

Incoherent methods

Solve the ML problem separately for each detector.

Identify triggers separately in each detector.

Time coincidence between detectors with a coincidence window accounting for the light travel time between detectors.

Rank each coincident candidate with an incoherent SNR-like quantity

$$\rho^2_{\mathrm{net}} = \sum_d \rho^2_d$$

Fully coherent methods

Solve the full ML problem simultaneously for all detectors with a common signal.

Requires exploring a larger search space (e.g. sky location).

Beneficial for many detectors (more constrained likelihood).

Semicoherent methods

Identify triggers separately in each detector.

Express the network likelihood in terms of the single-detector triggers.

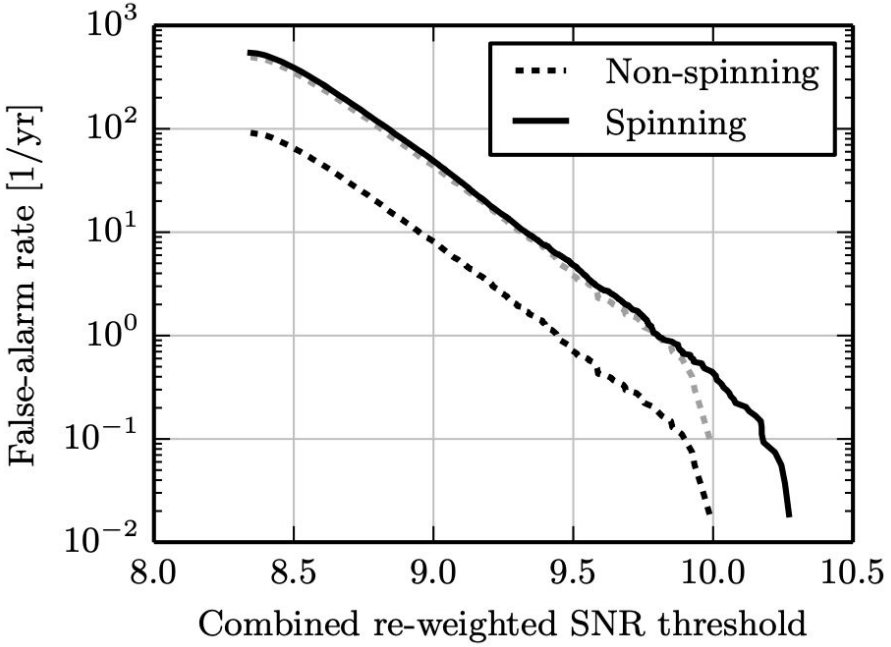
Statistical significance of candidate events

How often does instrumental noise produce a candidate event ranked higher than what I got?
→ False-alarm rate (FAR)

Generate a “null” distribution of ranking statistic from a large sample of unphysical events:

- By time-sliding data from different GW detectors
- By extrapolating the bulk of the ranking statistic.

Obtain a map to “look up” the FAR associated with a given ranking statistic.



E.g. FAR $\lesssim$ 1/100 yr
→ Candidate is unlikely to come from noise

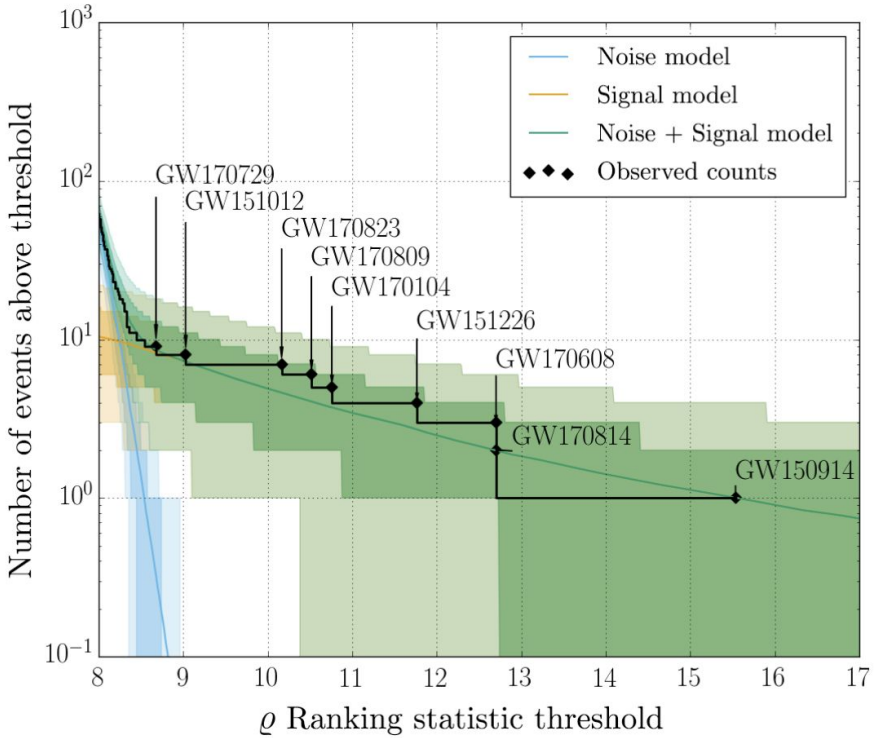
Statistical significance of candidate events

How probable is a candidate event to be of astrophysical origin?
→ P(astro) or p_astro

Construct a model for the rate density of signal $f(\boldsymbol{\lambda})$ and background $b(\boldsymbol{\lambda})$ candidates over the space of candidate parameters $\boldsymbol{\lambda}$

$$0 \leq P(\mathrm{astro}) = \frac{f(\boldsymbol{\lambda})}{f(\boldsymbol{\lambda}) + b(\boldsymbol{\lambda})} \leq 1$$

0: candidate is certainly of terrestrial origin
0.5: ambiguous origin
1: candidate is certainly astrophysical



Modern implementations of matched-filter searches for CBCs

"Pipelines" developed by different teams

GstLAL

Time-domain matched filter using a singular value decomposition of the templates.

MBTA

Frequency-domain matched filter using a two-band decomposition of the templates.

PyCBC

Direct frequency-domain matched filter.

SPIIR

Time-domain fully coherent matched filter

Online (low latency)

Results available ~10 s after data acquisition.

Used to produce rapid alerts for electromagnetic followup observations.

Offline (archival)

Results available hours to weeks after data acquisition.

Used for "more careful" analyses, to compile ultimate event catalogs like GWTC.

See https://emfollow.docs.ligo.org/userguide/ for more info.

Links to papers

General introduction to ground-based data analysis

A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals
https://iopscience.iop.org/article/10.1088/1361-6382/ab685e

Technical implementation of discrete-time FD matched filtering

FINDCHIRP: An Algorithm for detection of gravitational waves from inspiraling compact binaries
https://arxiv.org/abs/gr-qc/0509116

Description of a complete CBC search pipeline based on matched filtering

The PyCBC search for gravitational waves from compact binary coalescence
https://arxiv.org/abs/1508.02357

Latest LVK catalog

GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run
https://doi.org/10.1103/PhysRevX.13.041039

Links to tutorials

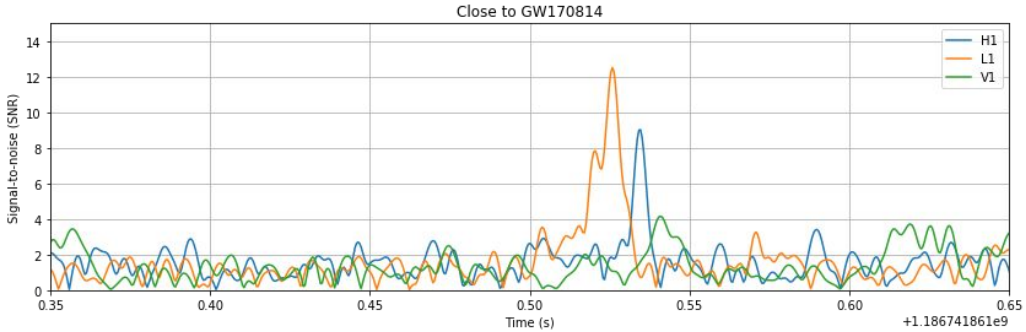
https://github.com/gwastro/PyCBC-Tutorials

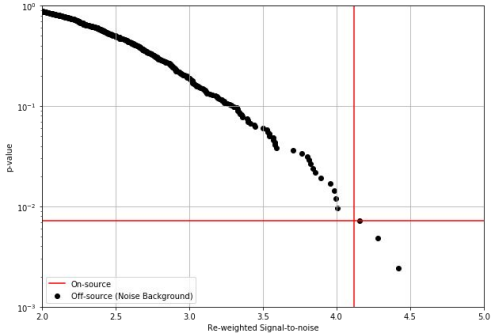
Tutorial 1: Accessing Gravitational-wave data

Tutorial 2: Data visualization and basic signal processing

Tutorial 3: Matched filtering to identify signals

Tutorial 4: Signal consistency and basic significance testing





Thank you!