Challenges in GW data analysis and computing

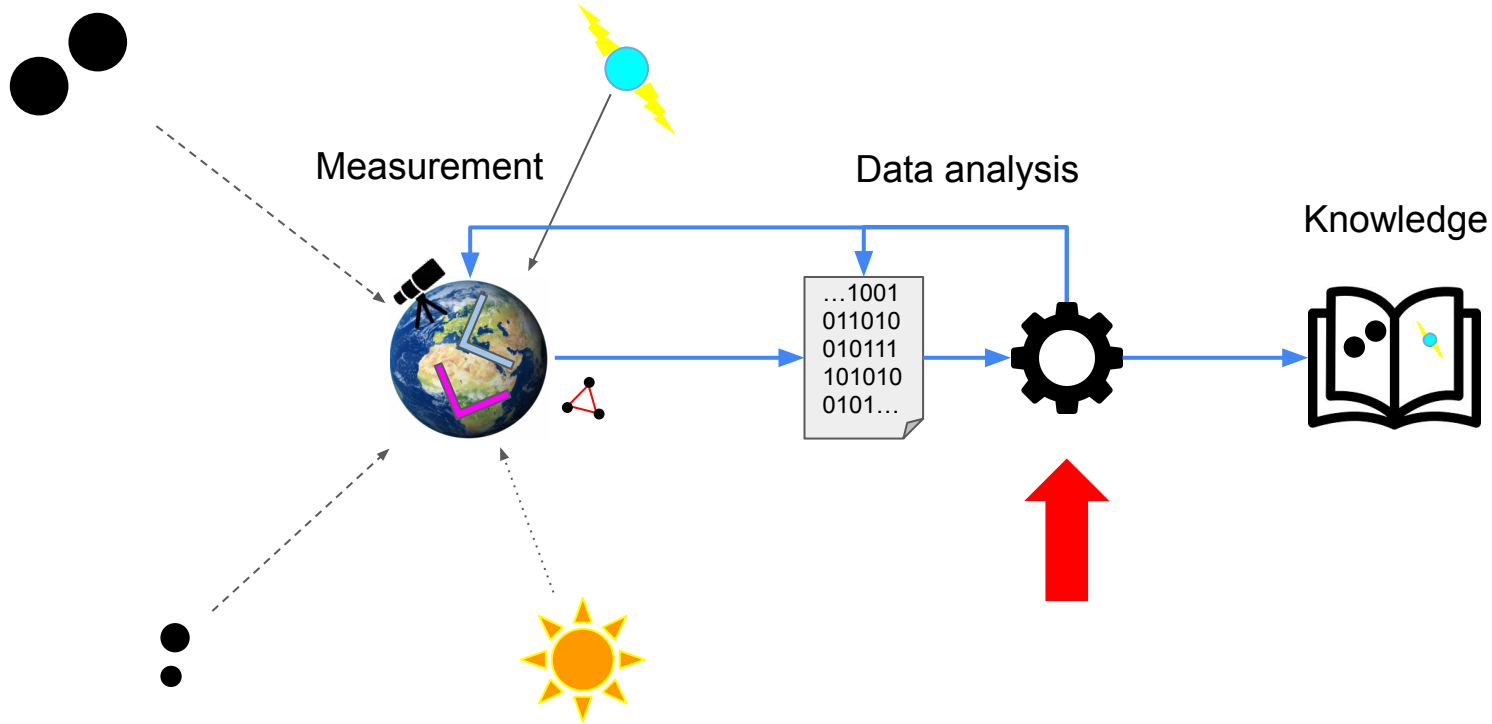
Tito Dal Canton

ISAPP summer school on gravitational waves
June 2021

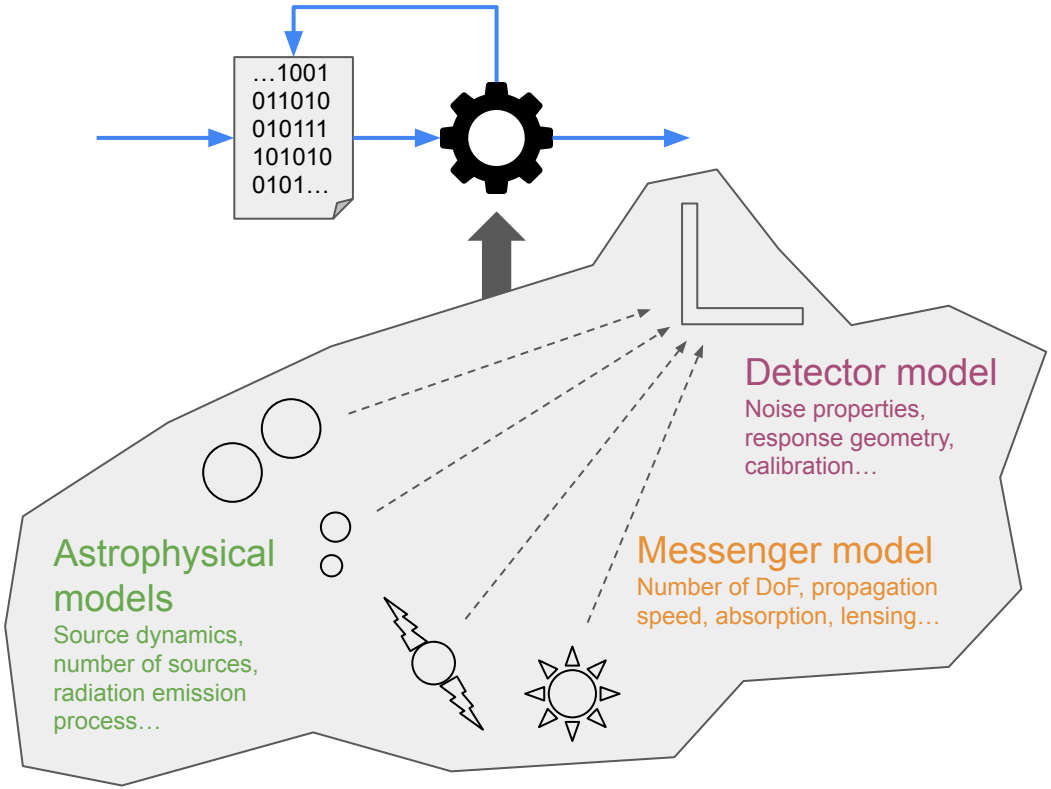
Overview

- Overview of GW data analysis challenges
- General approaches to these challenges
- Challenges from current analyses
  - LIGO & Virgo
- Expected challenges
  - LISA
  - ET & CE

The basic problem



Models



Model considerations

- Amount of physical detail
- Number of parameters
- Prior knowledge
- Computational cost

"All models are wrong; some are useful"

"[…] as simple as possible, but not simpler"

Data analysis

What should happen inside  ?

1. Predict what the data would look like given a certain model
2. Calculate how close that prediction is to the data
3. Select which model, or which choice(s) of parameters, best represents reality

Possible ways:

- Likelihood approach
  - Compute P(data|model,parameters) ← data - model(parameters)
  - Maximize or sample/integrate over parameters
- Machine learning approach
  - Train a classifier by showing it many examples of simulated data
  - Apply classifier to observational data

In both cases, we need many evaluations of the model

Likelihood

Assuming additive Gaussian noise, P(data|model,parameters) is something like

$$\log p\left(\boldsymbol{y}|\boldsymbol{\theta}\right) = -\frac{1}{2}\left[\log\left|\boldsymbol{\Sigma}\right| + \left(\boldsymbol{y}-\boldsymbol{h}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}-\boldsymbol{h}\right)\right]$$

Covariance of multivariate Gaussian realization

Data vector

Model vector

General challenges:

- Assumption of Gaussian noise
- Models predicting the same data (degeneracies)
- Large size of $\boldsymbol{h}$ and $\boldsymbol{\Sigma}$

Complexity of compact binary merger models

Inspiral vs inspiral-merger-ringdown

No spins, aligned spins, misaligned spins

Fundamental quadrupole mode vs higher-order modes

BH-only vs NS matter

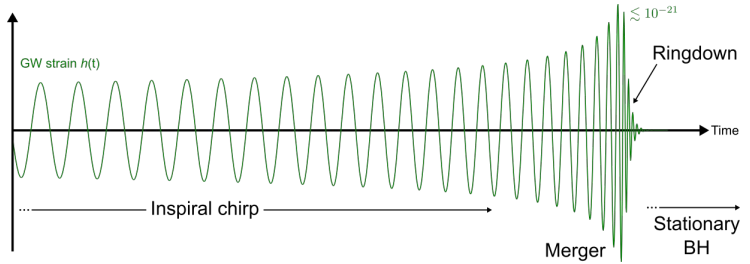
Quasicircular orbit vs eccentric orbit

Environmental effects (SMBH, AGN disk, DM halo)

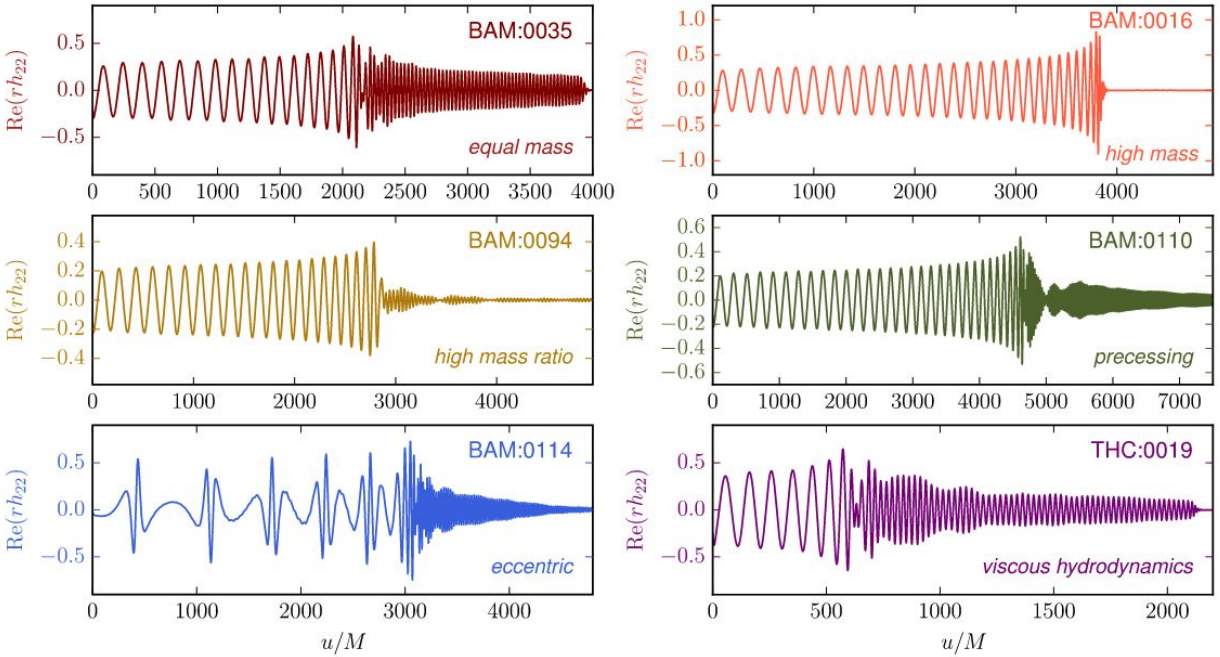
Deviations from general relativity

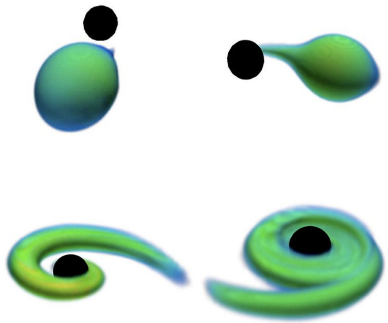
Numerical relativity

(Semi)analytic models (pN, EOB)



Complexity of compact binary merger models



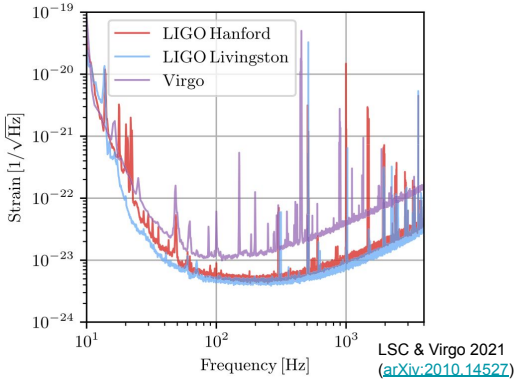


F. Foucart

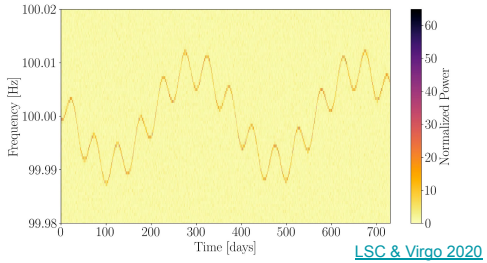
Dietrich et al 2018 (arXiv:1806.01625) & http://www.computational-relativity.org

Complexity of detector models

Do the detectors move during the duration of the signal?



LSC & Virgo 2021 (arXiv:2010.14527)



LSC & Virgo 2020

Binary neutron star inspiral

Duration (20 Hz → 1 kHz): ~160 s

→ Earth rotates by less than 1 deg

→ Antenna patterns approximately constant

$$h(t) \approx F_+(\theta,\phi,\psi)h_+(t)+F_\times(\theta,\phi,\psi)h_\times(t)$$
$$\tilde{h}(f) \approx F_+(\theta,\phi,\psi)\tilde{h}_+(f)+F_\times(\theta,\phi,\psi)\tilde{h}_\times(f)$$

Continuous-wave GW from a rotating neutron star

Duration: entire observation period!

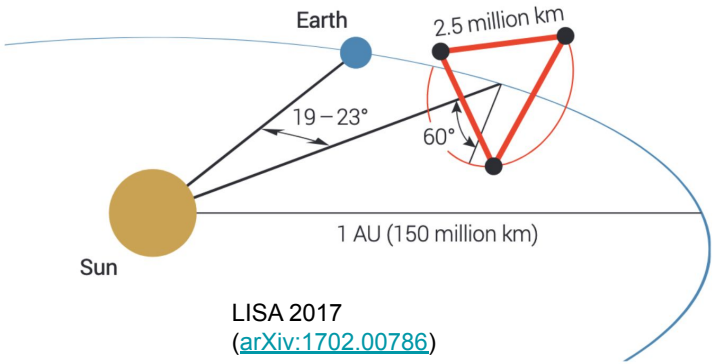
→ Detectors constantly rotating, Earth orbiting the Sun

→ Time-dependent modulation of signal

$$\tilde{h}(f) \neq F_+(\theta,\phi,\psi)\tilde{h}_+(f)+F_\times(\theta,\phi,\psi)\tilde{h}_\times(f)$$

Complexity of detector models

How does the gravitational wavelength compare to the size of the detectors?



LISA 2017
(arXiv:1702.00786)

$f_{GW} \lesssim 1$ kHz, $\lambda \gtrsim 300$ km, $L_{Virgo} = 3$ km
→ Long-wavelength approximation
→ $h(t) \approx F_+(\ldots,t)h_+(t)+F_\times(\ldots,t)h_\times(t)$

$f_{GW} \sim 0.1$ Hz, $\lambda \sim 3\times10^9$ m, $L_{LISA} \sim 2.5\times10^9$ m
→ Wavelength comparable to detector size
→ $h(t) \approx F_+(\ldots,t,f)h_+(t)+F_\times(\ldots,t,f)h_\times(t)$

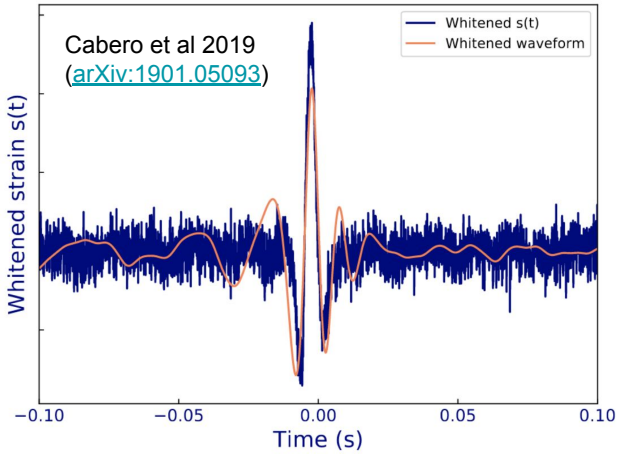
Complexity of detector models

Do the statistical properties of the detector noise change over the duration of the signal?

Do the detectors produce glitches that look like astrophysical signals?

How much do you trust the calibration of the detectors?

Are data gaps frequent enough to intersect the astrophysical signals?



A very basic challenge

- 1 month of LIGO Hanford, LIGO Livingston and Virgo data with Gaussian and stationary noise (PSD known) sampled at 1024 Hz → ~10^9 samples
- Hypothesis 0:
  No astrophysical signals; data is just detector noise
  → No parameters
- Hypothesis 1:
  Data is detector noise + 1 quasicircular BBH merger with zero spins
  → 8 unknown parameters
- Hypothesis 2:
  Data is detector noise + 2 quasicircular BBH mergers with zero spins
  → 16 unknown parameters
- …

A very basic challenge

Which of those models are supported by the data?

→ **Bayesian model selection**: compute the evidence for each model,

$$Z = \int P(\text{data}|\text{model}, \vec{\lambda}) P(\vec{\lambda}) \mathrm{d}\vec{\lambda}$$

Likelihood

Prior

Parameter vector

…then compare the evidences weighted by prior probabilities of the models.

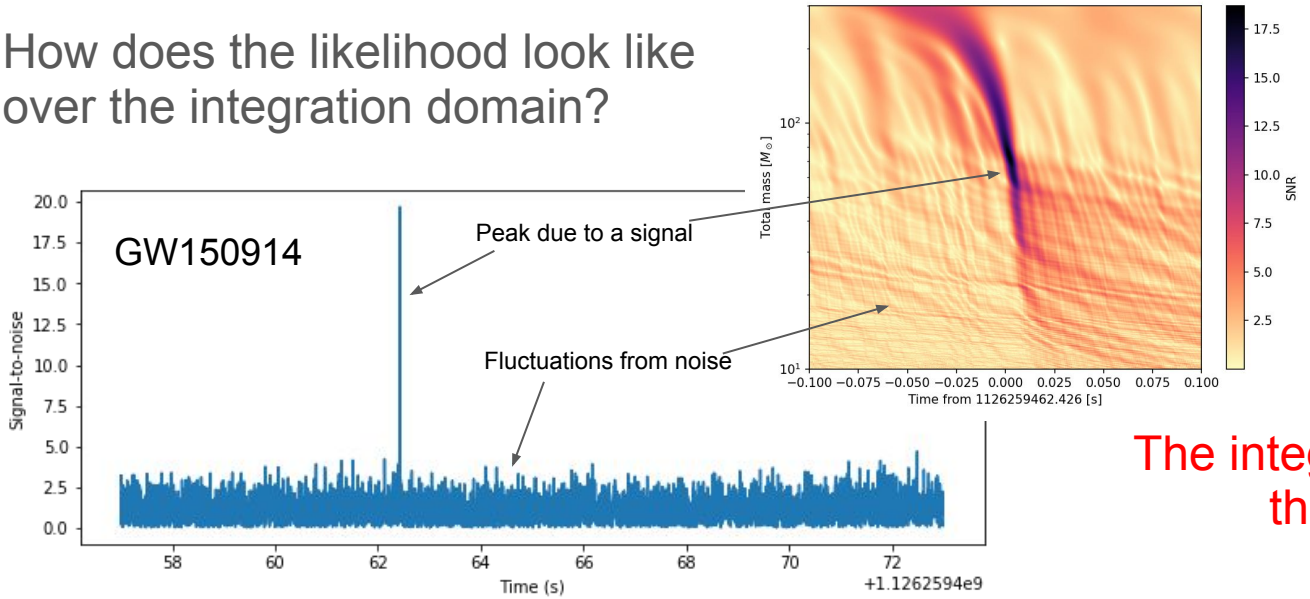
A very basic challenge

$$Z = \int P(\text{data}|\text{model}, \vec{\lambda}) P(\vec{\lambda}) \mathrm{d}\vec{\lambda}$$

Hypothesis 1 → Integral over 8 dimensions

Hypothesis 10 → Integral over 80 dimensions

How does the likelihood look like over the integration domain?



The integration must resolve the peak of the signal!

A very basic challenge

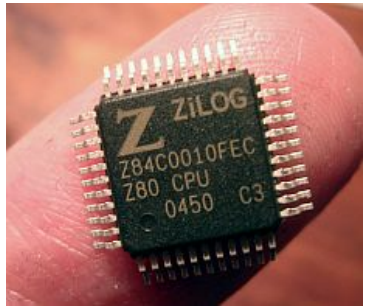
Suppose that:

- We can resolve the peak by scanning just 10 points per dimension.
- We must complete the analysis in 1 month using a single-core computer.

Then

- For hypothesis 1, we need to calculate the likelihood at 10^8 points
  → Must calculate one likelihood value in ~10 ms ✔
- But for hypothesis 80, we have 10^{80} points!
  → One likelihood value in ~10^{-74} s ✘

Compare with the clock period of a 3 GHz CPU core.
Even a cluster of 10000 CPU cores will be useless!



A cute CPU, but not good for our problem

…and remember that we have 10^9 data samples, with ideal noise with known PSD.

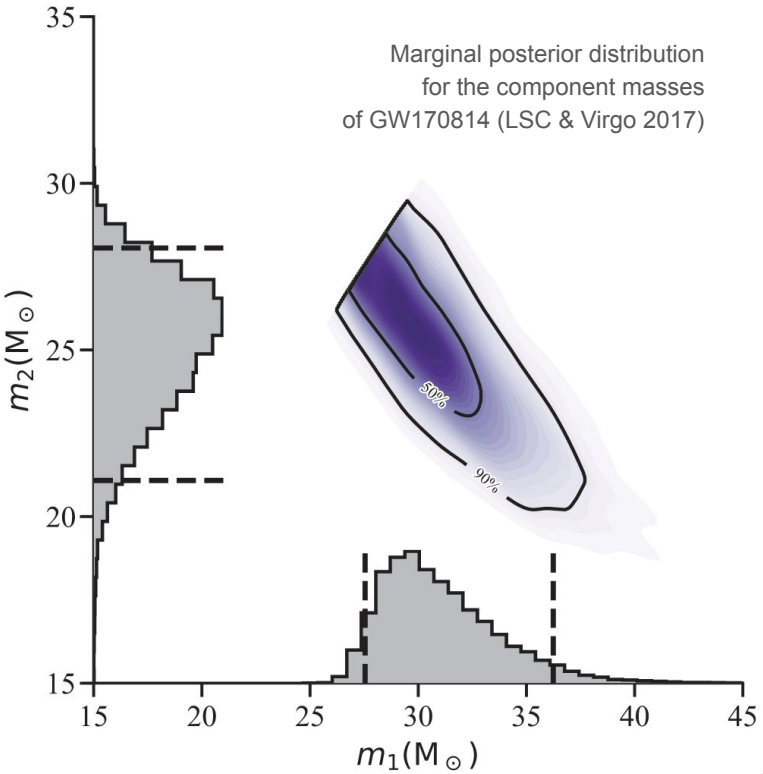
A very basic challenge

Suppose we magically manage to carry out the computation, and find support for hypothesis 1. What can we conclude about the masses of that BBH merger?

→ **Bayesian inference**: compute the posterior distribution for the BBH model over its 8-dimensional parameter space, then marginalize over all parameters except the two masses.

$$P(m_1, m_2|\mathrm{data}, \mathrm{H}_1) \propto \int P(\mathrm{data}|\mathrm{H}_1, \vec{\lambda}')P(\vec{\lambda}')\mathrm{d}\vec{\lambda}'$$

And we have another integral over 6 dimensions!



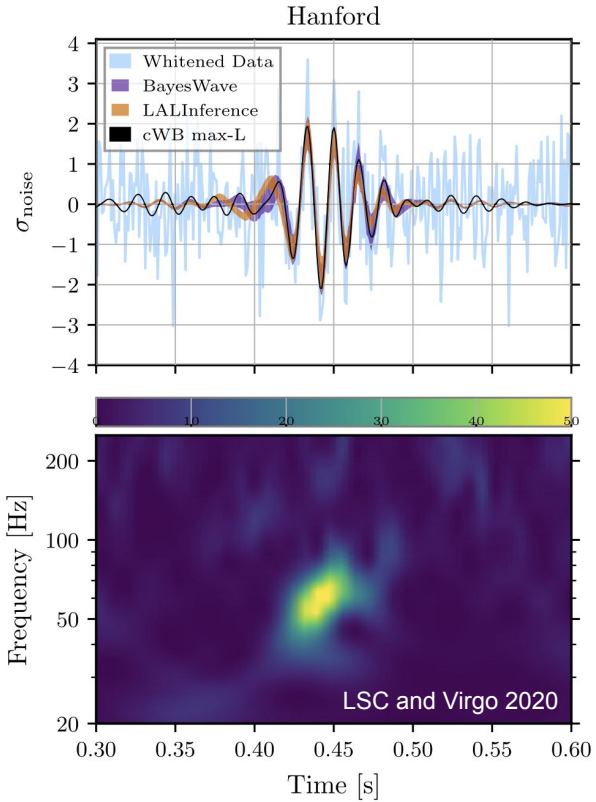
So what can we do?

Break a giant problem into many small problems

Solving the problem as presented is clearly impossible… But do we have to?

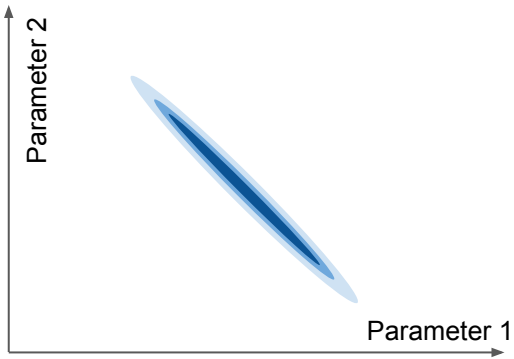
- Do the *N* possible astrophysical signals “interfere” with each other?
  Hint: are their peaks in the likelihood well separated?
- Divide months of data into ~1 hour long segments
- Assume signals are rare: zero or one per segment
- Separate the problem of searching from the problem of parameter estimation
  a. Use simpler models to roughly identify the most challenging parameters
  b. Reduce the data to what is minimally necessary to study a possible signal
  c. Use an expensive, more accurate model to analyze the reduced data

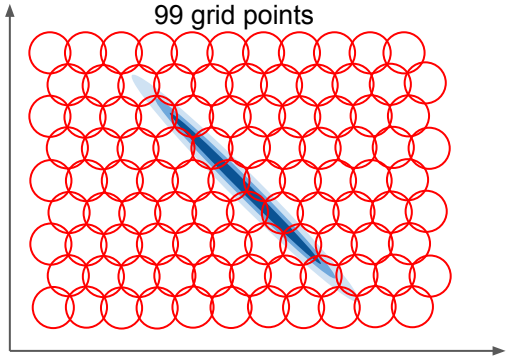
Simplify the models as much as possible

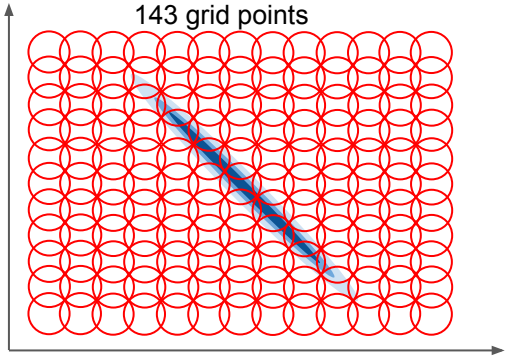


- Reduce the amount of physics at the source → fewer parameters / smaller prior volume / fewer computations
  - Example: neutron stars have small spins and merge at ≳1 kHz
- Phenomenological models
  - Example: search using wavelets instead of accurate compact binary merger waveforms
- Idealize the detectors
  - Example: stationary Gaussian noise, evenly-sampled data
    → $\mathbf{\Sigma}$ becomes diagonal in the Fourier domain
    → $\mathbf{\Sigma}^{-1}$ becomes a trivial operation

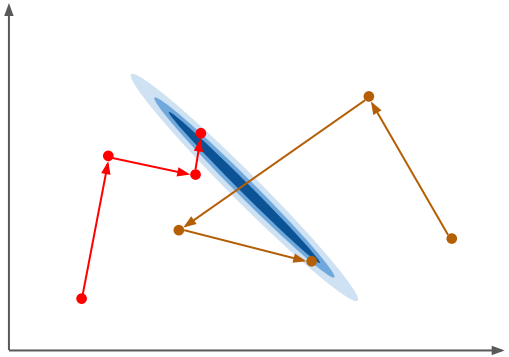
Use clever methods to explore *N*-dimensional likelihoods







- Analytic maximization or integration
- Efficient placement of grid points
- Markov-chain Monte Carlo, parallel tempering, nested sampling, particle-swarm optimization, differential evolution, genetic algorithms, transdimensional MCMC…



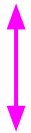
Carefully evaluate the number of operations required

- Example: Fourier transform
  - Naive Discrete Fourier Transform ⟶ Complexity ~N^2
  - Fast Fourier Transform ⟶ Complexity ~$N \ln N$ (Gauss-Cooley-Tukey algorithm)
- Take 1 hour segment of data sampled at 2048 Hz ⟶ 7×10^6 samples
- DFT: ~10^{13} operations
- FFT: ~10^8 operations
- What looks like a one-day problem is really a one-second problem

Make sure the code is implemented efficiently

Avoid unnecessary operations

```
for data_segment in data:
    for params in template_bank:
        template = generate_waveform(params)
        snr = matched_filter(data_segment, template)
        find_peaks(snr)
```



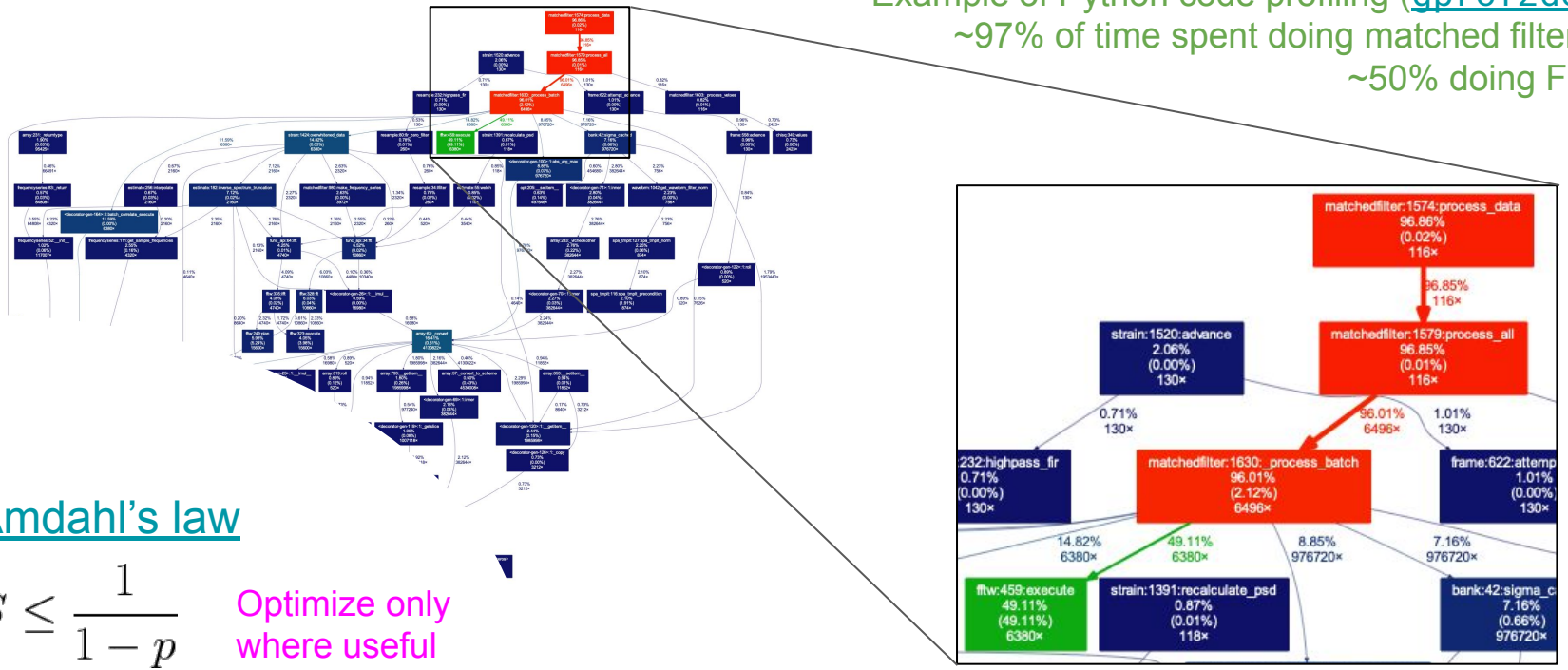
Which version requires more calculations?

```
for params in template_bank:
    template = generate_waveform(params)
    for data_segment in data:
        snr = matched_filter(data_segment, template)
        find_peaks(snr)
```

Make sure the code is implemented efficiently

Know where your code is spending time

Example of Python code profiling (gprof2dot):
~97% of time spent doing matched filtering
~50% doing FFTs



Amdahl's law

$$S \leq \frac{1}{1-p}$$

Optimize only where useful

Use the appropriate computational tools

- General-purpose libraries that are already optimized
  - Numpy/Scipy, FFTW, GNU Scientific Library
- Computer clusters + job schedulers & workflow management tools
  - Open Science Grid
  - HTCondor, Pegasus, MPI
  - Talk to computing experts in LIGO, Virgo and KAGRA
- Single-instruction-multiple-data
- Many-core CPUs, GPUs
- Dedicated (“exotic”) hardware
  - FPGAs
  - ASICs

Sacrifice some science to simplify the problem

Do you expect to find signals in a particular region of the parameter space?

Would you learn more things by exploring a particular region?

Are regions already excluded by previous observations?

Are some regions much more expensive to explore than others?

Examples from present analyses

Stellar-mass compact binary mergers in LIGO/Virgo

Assumptions:

- Rare signals
  Non-overlapping
- No precession
  Spins aligned with the orbital axis
- No eccentricity
- No matter effects
  Neutron stars treated as black holes!
- Ignore higher-order modes
- Short signals
  Antenna pattern functions and time delays are time-independent
- …and others

Complete model for signal observed by a detector:

$$h(t) \approx A(t;\vec{\lambda})\cos[\Psi(t;\vec{\lambda})]$$
$$\tilde{h}(f) \approx B(f;\vec{\lambda})\cos[\Phi(f;\vec{\lambda})]$$

With parameter vector λ comprising only:

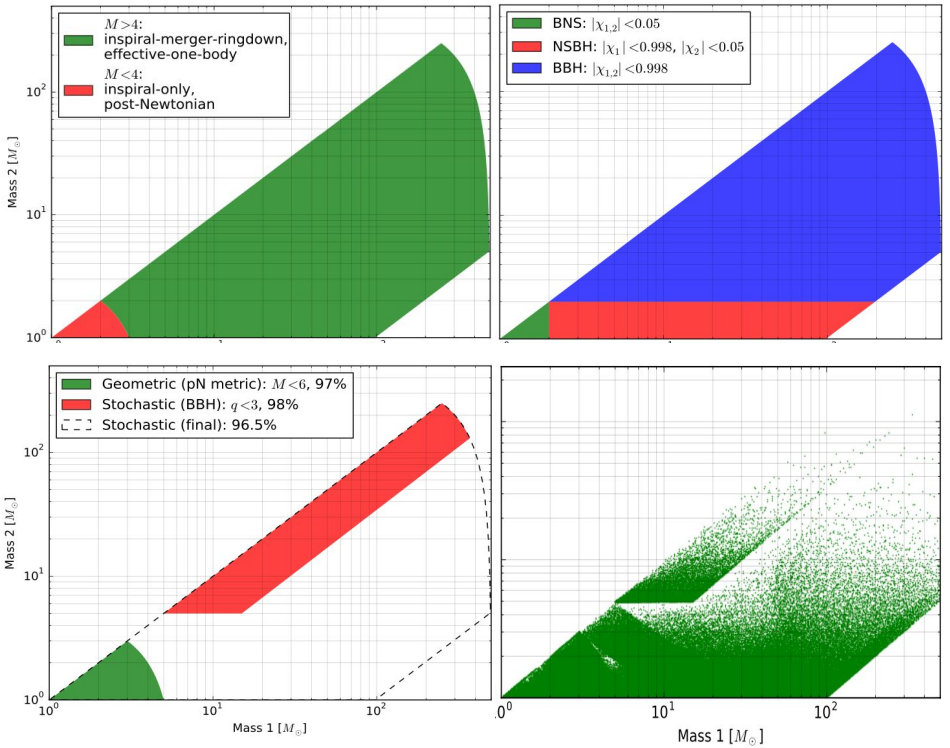
- Amplitude
  (incl. distance, orbital orientation, sky location)
- Merger time
- Merger phase
- Two masses
- Two spin components

Such a waveform can be evaluated in 1-10 ms.

Stellar-mass compact binary mergers in LIGO/Virgo

Exploration of parameter space

- Amplitude → Analytic
- Merger time → Fast Fourier Transform
- Merger phase
  → Two templates out of phase by $\pi/2$ (also via FFT)
- Masses and spins
  → Bruteforce grid search over ~4×10^5 points (*template bank*)



Dal Canton & Harry 2017 (arXiv:1705.01845)
Keppel 2013 (arXiv:1303.2005)
Roy et al 2017 (arXiv:1702.06771)

Stellar-mass compact binary mergers in LIGO/Virgo

6 months of data, 2048 samples per second, 3 detectors

Divide data into 1000 s segments → 5×10^4 segments

4×10^5 templates

→ We need to do 2×10^{10} independent FFTs (plus cheaper operations)

Want to finish in 10 days?
→ Need a machine that can do 2×10^4 FFTs per second
→ Divide the workload over 10^4 CPU cores

In principle, this is definitely feasible 👍 In practice, remember those words about efficient code and appropriate computing tools ⚠️

Stellar-mass compact binary mergers in LIGO/Virgo

Candidates from search phase → Parameter estimation using nearby data

1. Assume masses and spins are known
   → Approximate spatial localization in seconds
   (Singer & Price 2016, arXiv:1508.03634)
2. Relax many assumptions; use waveform models with precession, higher-order modes, matter effects…
   → Full parameter estimation in hours to many days (for a single event!)
   Cost typically dominated by the waveform model

⚠️ More sensitive detectors → higher detection rate → higher cost of PE ⚠️

Stellar-mass compact binary mergers in LIGO/Virgo

Stationary Gaussian noise → Whittle likelihood

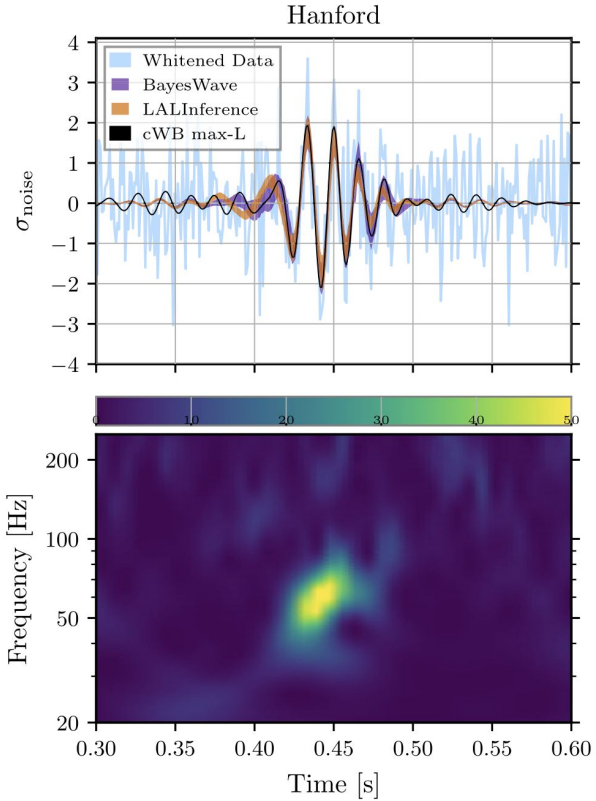
$$p(\vec{d}|\vec{\theta}, H) \propto \exp\left[-2\sum_{i}^{N} \frac{|h_i(\vec{\theta}) - \vec{d}_i|^2}{\tau S_n(f_i)}\right]$$

$$N = f_{\rm Nyq}/\delta f \approx f_{\rm Nyq} T_{\rm chirp} \approx 10^4$$

- Reduced order quadratures
  Smith et al 2016 (arXiv:1604.08253)
- Multiband interpolation
  Vinciguerra et al 2017 (arXiv:1703.02062)
- Relative binning / Heterodyning
  Zackay et al 2018 (arXiv:1806.08792)

Method	Waveform Evaluations
Relative binning [this work]	63
Reduced order quadrature [7]	1740
Multi-band interpolation [6] [a]	$\sim 10^4$
Full FFT grid	$\sim 10^7$

GW190521: an example of model degeneracy



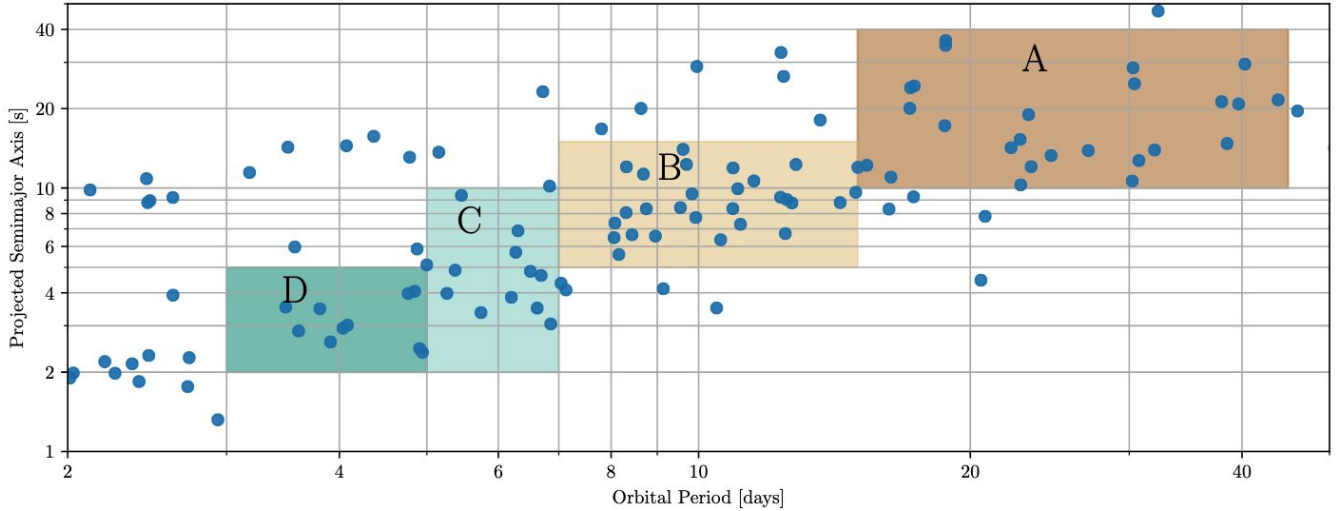
After two years, what happened is still under debate.

- A merger of BHs with precessing spins
  - LIGO & Virgo 2020 (arXiv:2009.01075)
- A merger of BHs in an AGN disk
  - Graham et al 2020 (arXiv:2006.14122)
- A merger of BHs in an eccentric orbit
  - Romero-Shaw et al 2020 (arXiv:2009.04771)
  - Gayathri et al 2020 (arXiv:2009.05461)
- A head-on collision of boson stars
  - Calderon-Bustillo et al 2021 (arXiv:2009.05376)

…essentially the same signal in LIGO & Virgo data.

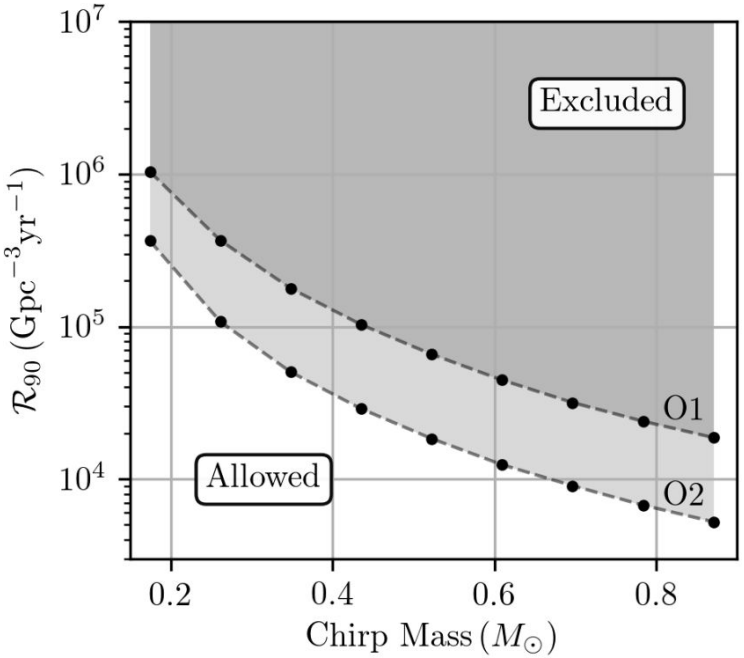
Continuous GWs from unknown pulsars in LIGO's O3a run

- LIGO & Virgo Collaborations, 2021 (arXiv:2012.12128)
- Only use LIGO data, not Virgo
- Impossible to explore the whole space → Focus on a reduced portion



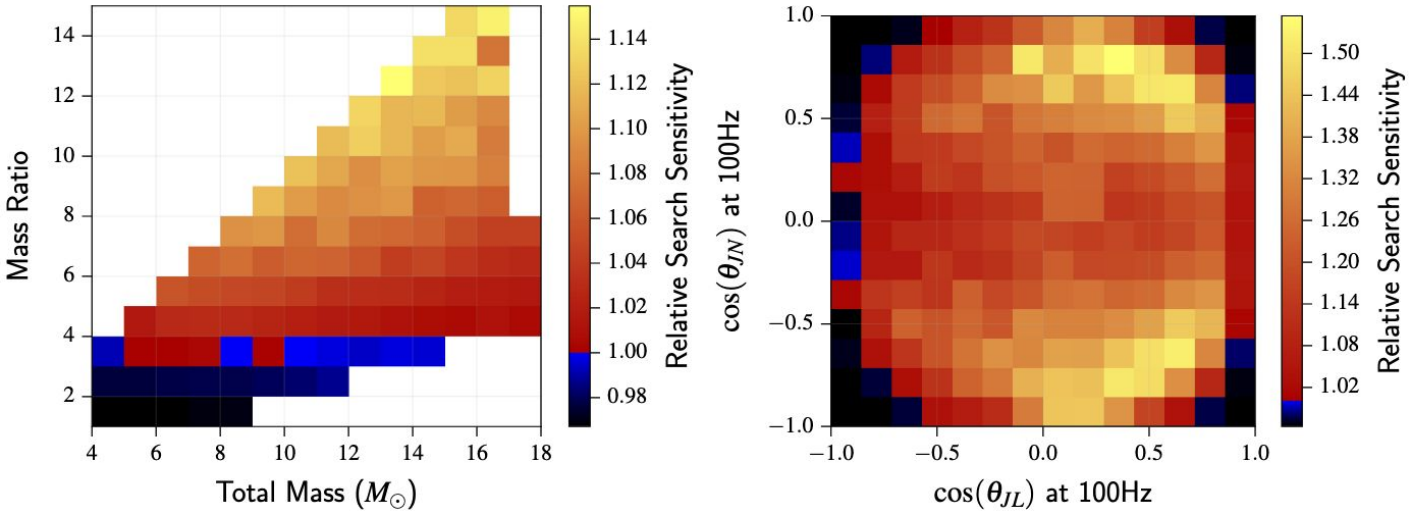
Mergers of sub-solar-mass compact objects in LIGO/Virgo

- LIGO & Virgo Collaborations, 2019 (arXiv:1904.08976)
- ~10^6 templates (~2x BNS-BBH-NSBH searches)
- Start templates at 45 Hz instead of the typical ~20 Hz
- 8% SNR loss → ~22% reduction in sensitive volume compared to higher-mass searches
- Limited range of chirp mass explored



Compact binary mergers with precession in LIGO/Virgo

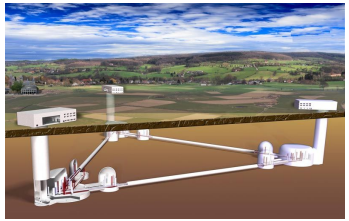
- Harry et al 2016 (arXiv:1603.02444), Indik et al 2017 (arXiv:1612.05173)
- ~10× more templates than aligned-spin banks
- Not yet systematically applied to real data

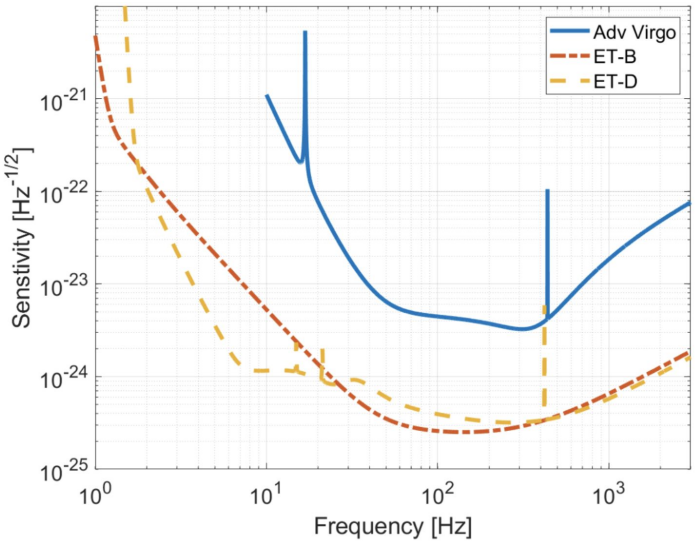


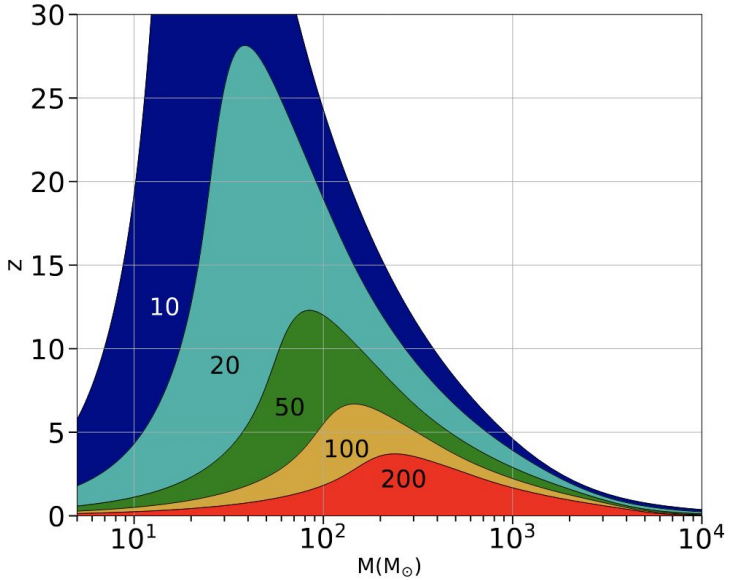
Examples from the future

You (and your students) will have to deal with these!

3G detectors: long, loud, frequent inspirals

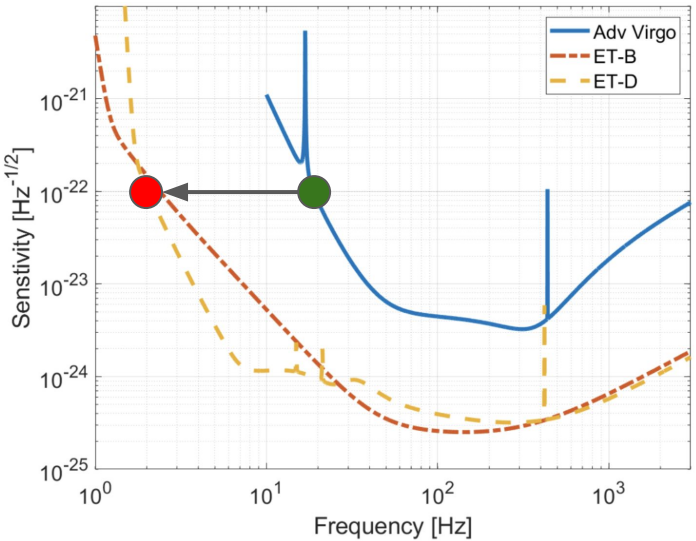






Maggiore et al 2020 (arXiv:1912.02622)

3G detectors: long, loud, frequent inspirals



Maggiore et al 2020 (arXiv:1912.02622)

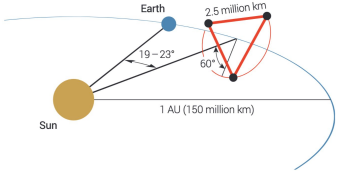
```
$ lalapps_chirplen --m1 1.4 --m2 1.4 --flow 20
0: Reached requested termination frequency
fStart according to Tev = 1.999983e+01 Hz
fStop  according to Tev = 1.513428e+03 Hz
length according to Tev = 1.607861e+02 seconds
Ncycle according to Tev = 5144.283297

$ lalapps_chirplen --m1 1.4 --m2 1.4 --flow 2
0: Reached requested termination frequency
fStart according to Tev = 2.000024e+00 Hz
fStop  according to Tev = 1.499925e+03 Hz
length according to Tev = 7.370359e+04 seconds
Ncycle according to Tev = 236079.041636
```

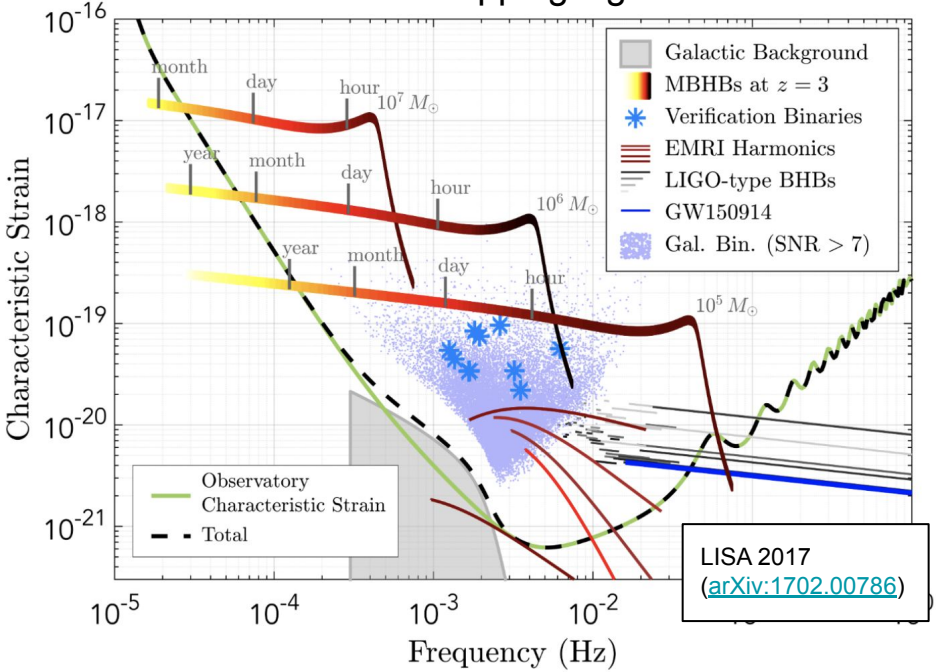
3G detectors: long, loud, frequent inspirals

- Binary neutron star mergers in band for hours
  - Detector motion no longer negligible
  - Current bruteforce matched filtering likely prohibitive
  - Regimbau et al 2012 (arXiv:1201.3563)
- One merger every 1-10 minutes
  - Overlap between signals
  - Current search methods may no longer apply
  - Current approach to parameter estimation will be impossible
- Events with SNR ~ 100
  - Statistical uncertainties may become smaller than model systematics
  - Source of “noise” for weaker signals?
- ≳3 orders of magnitude larger cost (CPU and RAM)
- Bagnasco et al 2020

LISA: a signal-dominated observatory

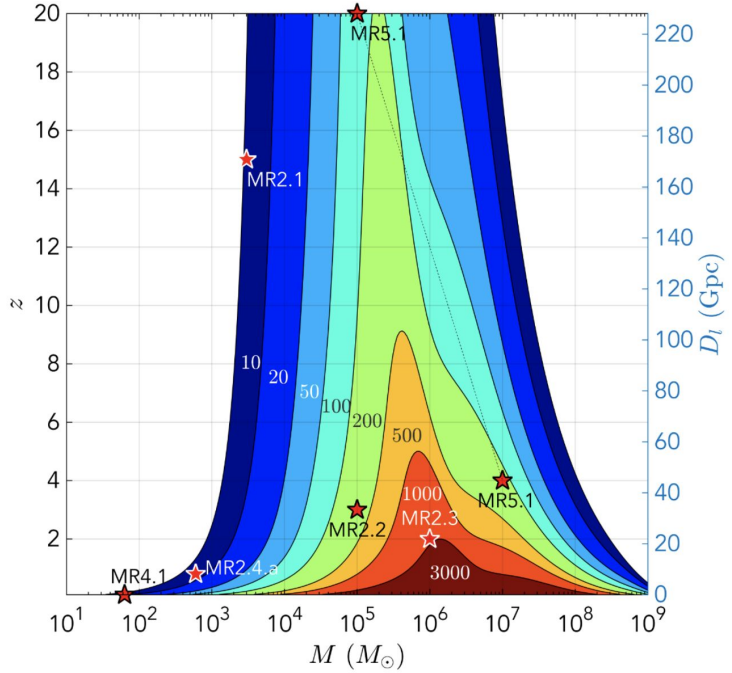


$\gtrsim 10^4$ overlapping signals

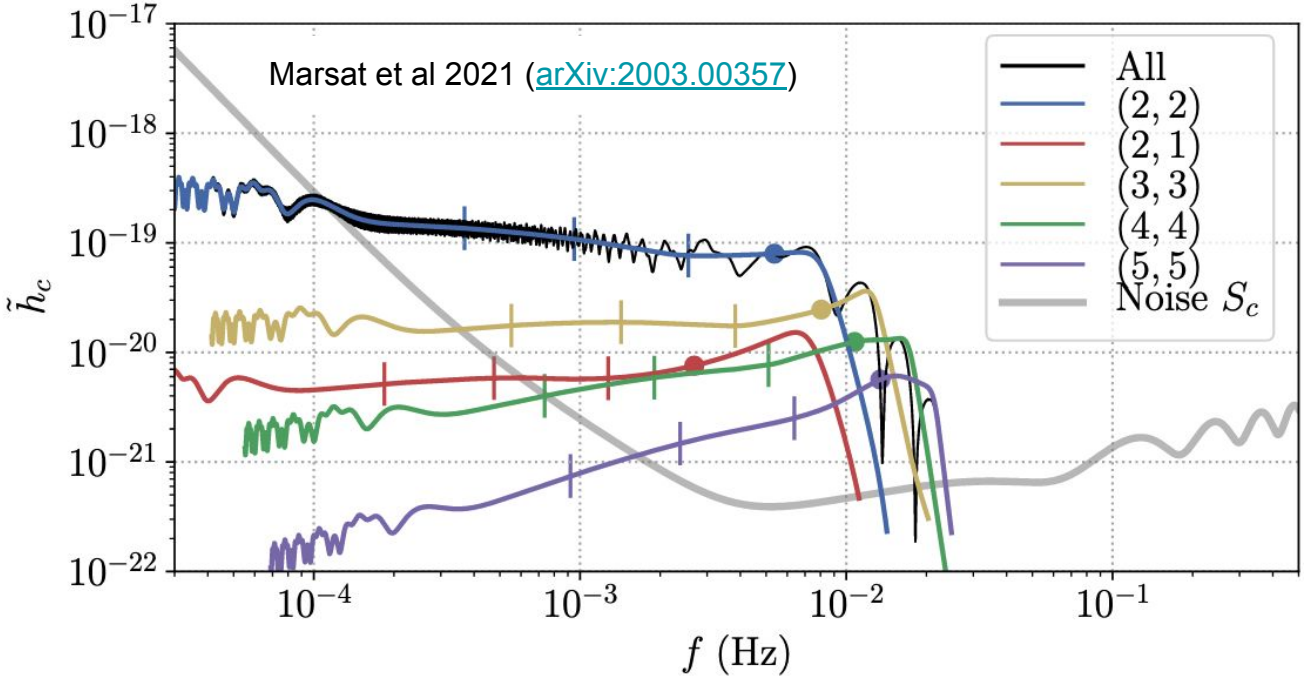


LISA 2017 (arXiv:1702.00786)

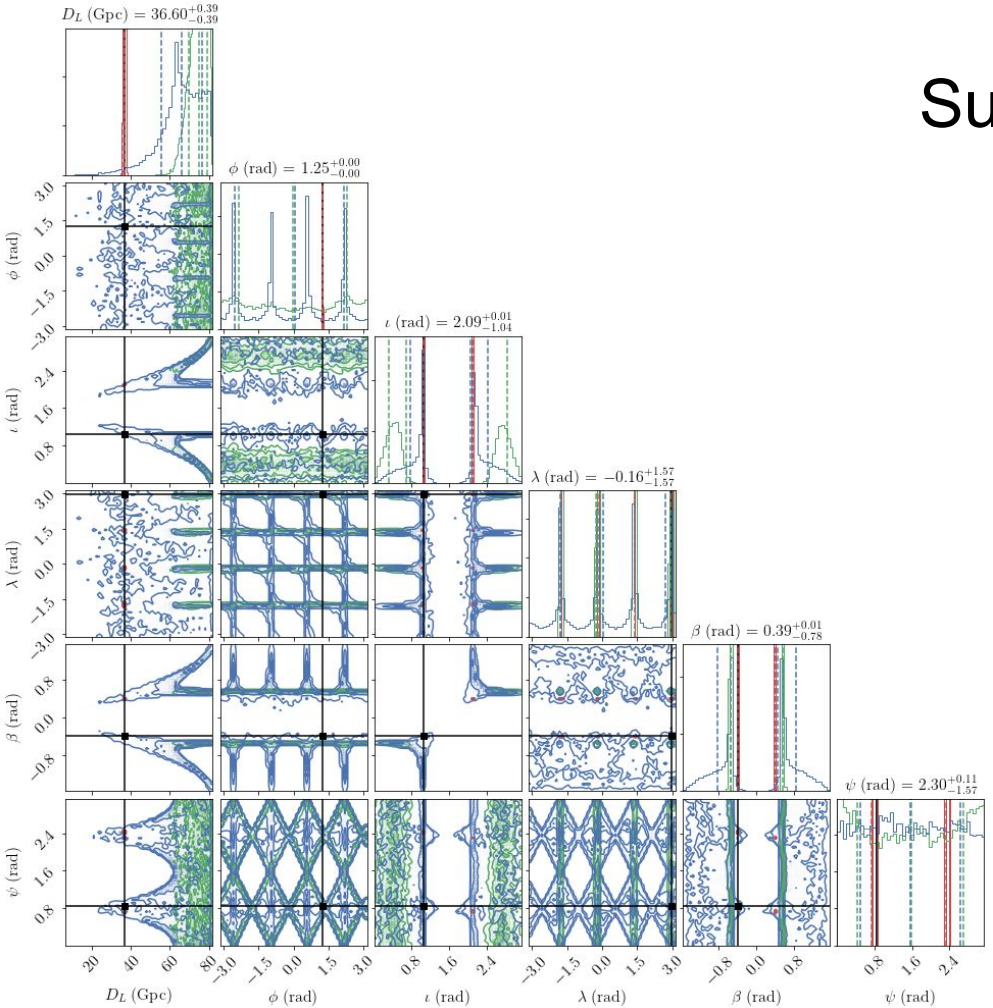
…including very loud ones



Supermassive BBHs in LISA



Supermassive BBHs in LISA



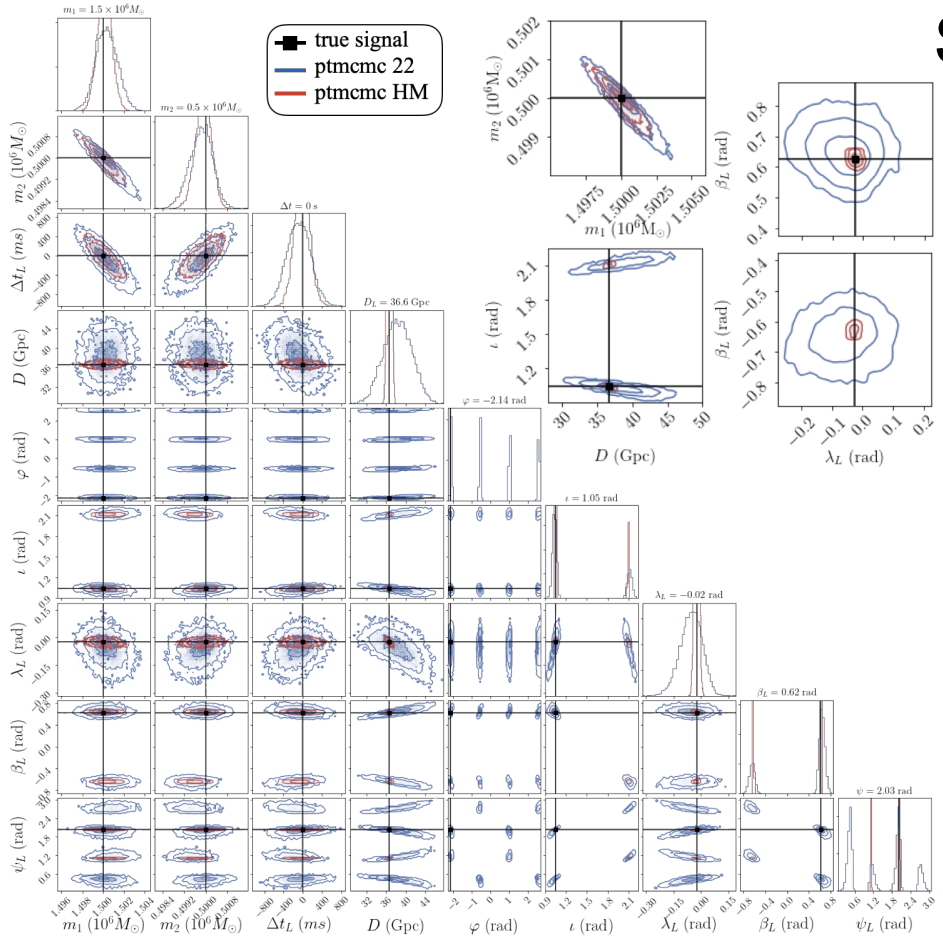
Parameter estimation neglecting LISA motion and higher-order modes

↓

Extreme degeneracies make the likelihood very complicated

Marsat et al 2021 (arXiv:2003.00357)

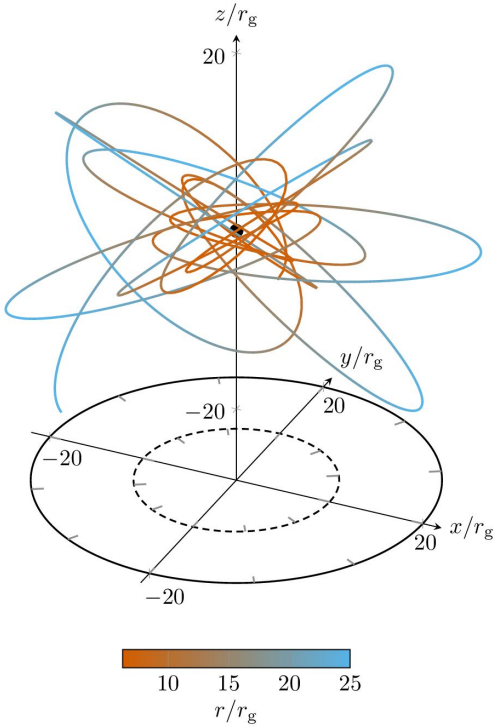
Supermassive BBHs in LISA



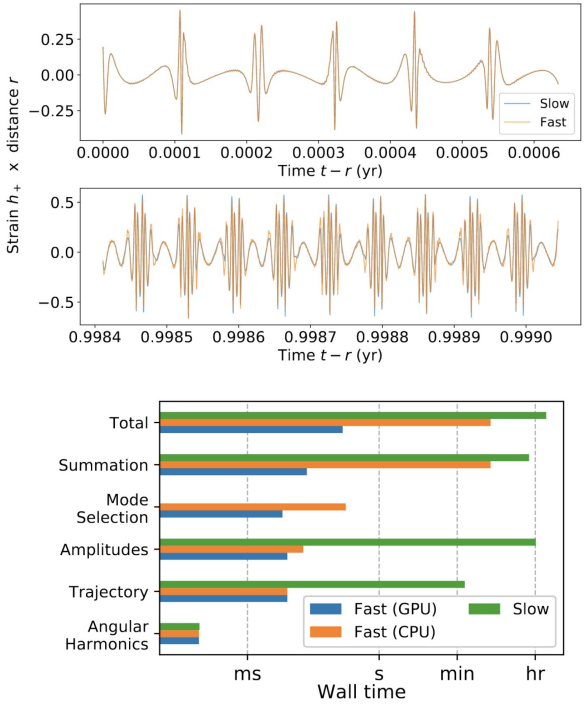
Parameter estimation
including LISA motion,
with dominant mode only
or with higher-order modes

Marsat et al 2021
(arXiv:2003.00357)

Extreme-mass-ratio inspirals in LISA



Berry et al 2019 (arXiv:1903.03686)

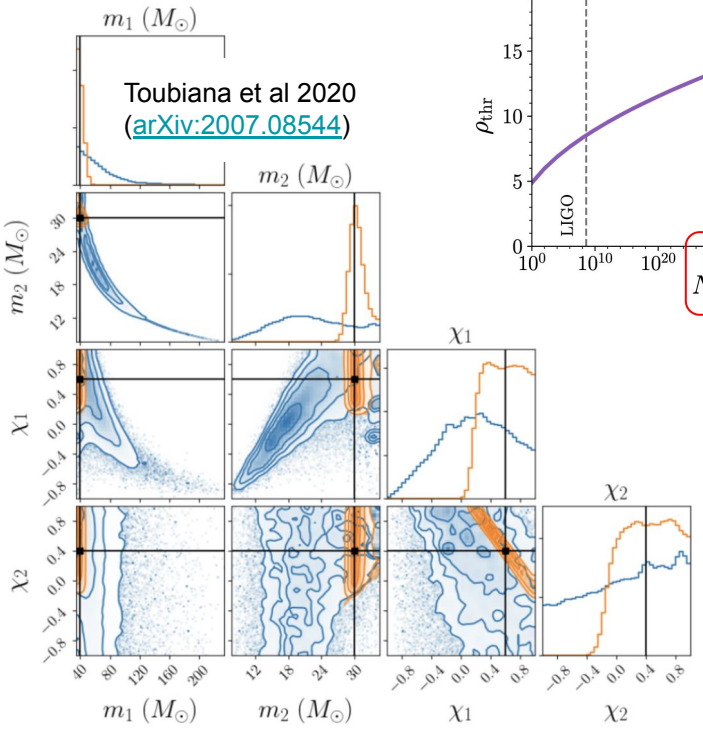


Chua et al 2021 (arXiv:2008.06071)

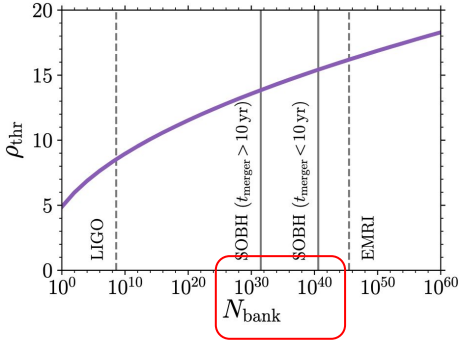
- Extremely rich and complicated orbits
- Long-lived signals
- Need a waveform model accurate over ~10^4-10^5 orbits, and computationally cheap
- Transient resonances complicate the use of adiabatic approximations
- These signals also overlap with all the other ones…

Stellar-mass compact binary mergers in LISA

- Years-long signals
  → LISA moves during them
- Chirping, not narrow-band
- High-frequency → Wavelength comparable to detector size
- Precession and eccentricity
  → 17 params for each signal
- Relatively weak (SNR ~10)
- Environmental effects?
  → More parameters

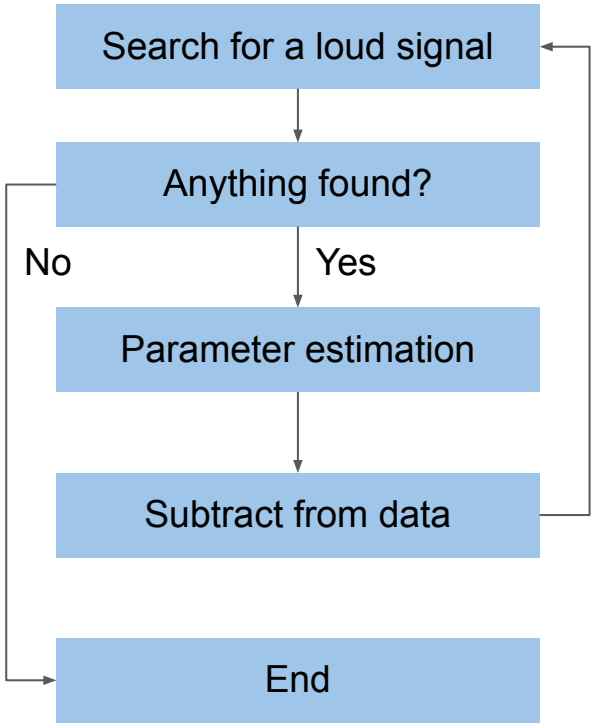


Toubiana et al 2020
(arXiv:2007.08544)



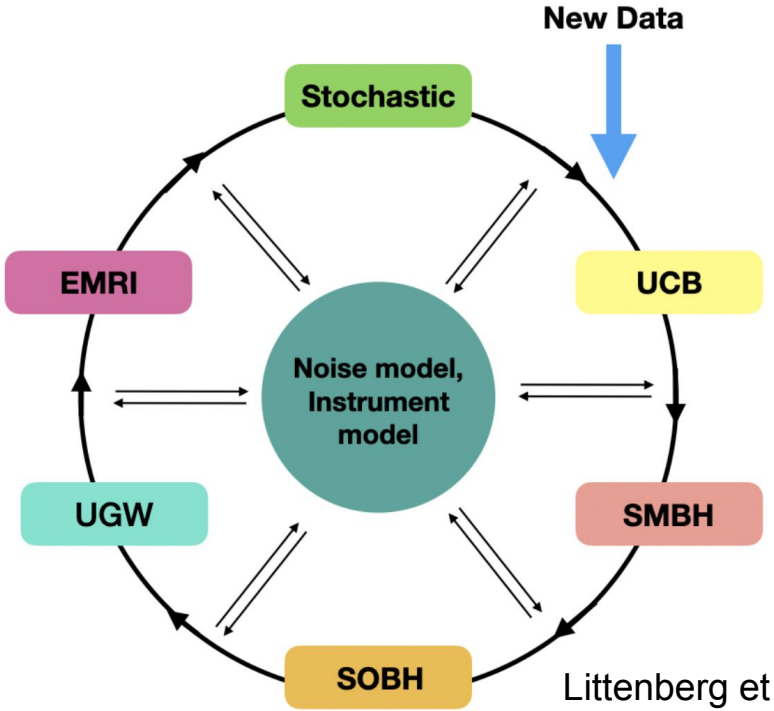
Moore et al 2019
(arXiv:1905.11998)

LISA data analysis: iterative search and subtraction



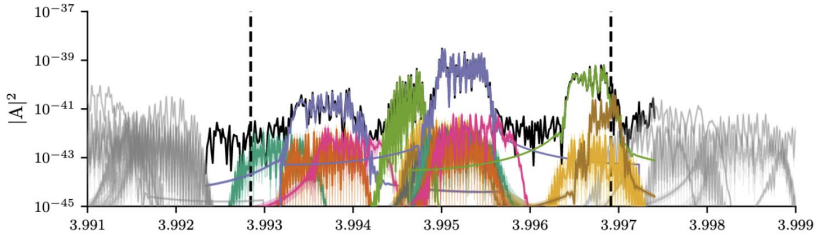
- Sequential, not parallelizable
- Need models that subtract the entire signal reliably

LISA data analysis: simultaneous global fit



Littenberg et al 2020
(arXiv:2004.08464)

- Transdimensional MCMC: handle models with varying number of parameters
- Separation of sources where possible
- Approximations and optimizations to speed up the likelihood
- Demonstrated for one class of LISA sources; more work needed for the full solution



Preparing for the challenges

Simulate the expected data!

- CBC rates known fairly well
- Keep up with modern signal models; include as much physics as possible
- LISA data challenge: https://lisa-ldc.lal.in2p3.fr/
- Need something equivalent (up to date) for 3G detectors
- Multi-observatory data challenges?

There are more challenges to talk about…

- Unknown/unexpected signals
- Other detectors
  - Pulsar timing arrays
  - Lunar GW detectors
  - NEMO
- Multimessenger astronomy
  - Organizing joint observations
  - Reducing the latency of reporting results
- Organizing results for ~10^6 events

…and potential solutions as well

- Machine learning