Towards a Spanish LISA Global Fit Pipeline

Ivan Martin Vilchez

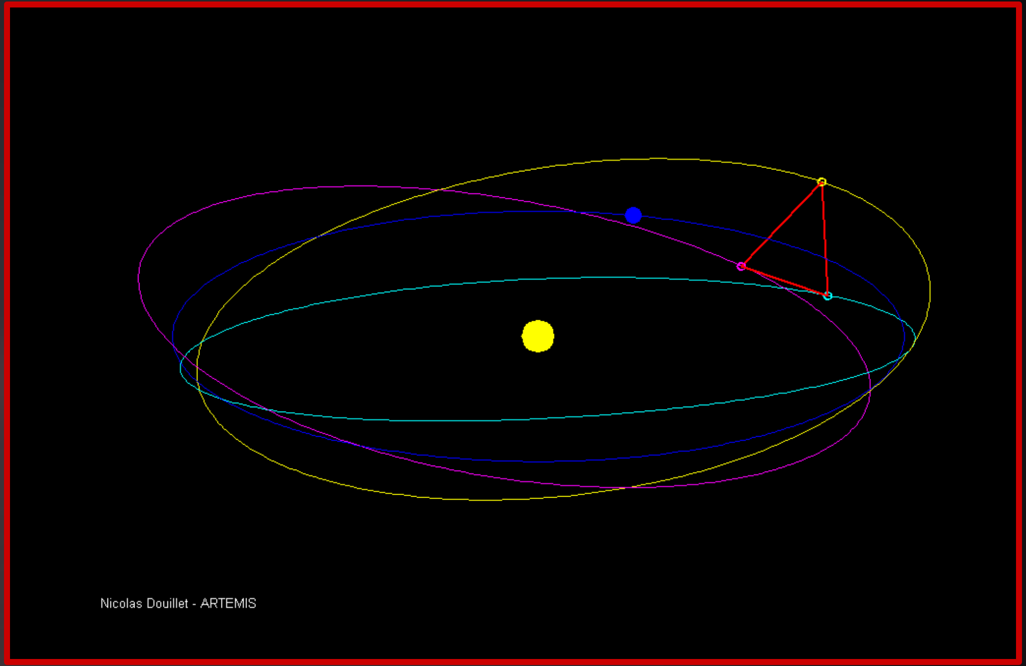
2nd BIG&C Meeting

IFAE, January 22nd 2024

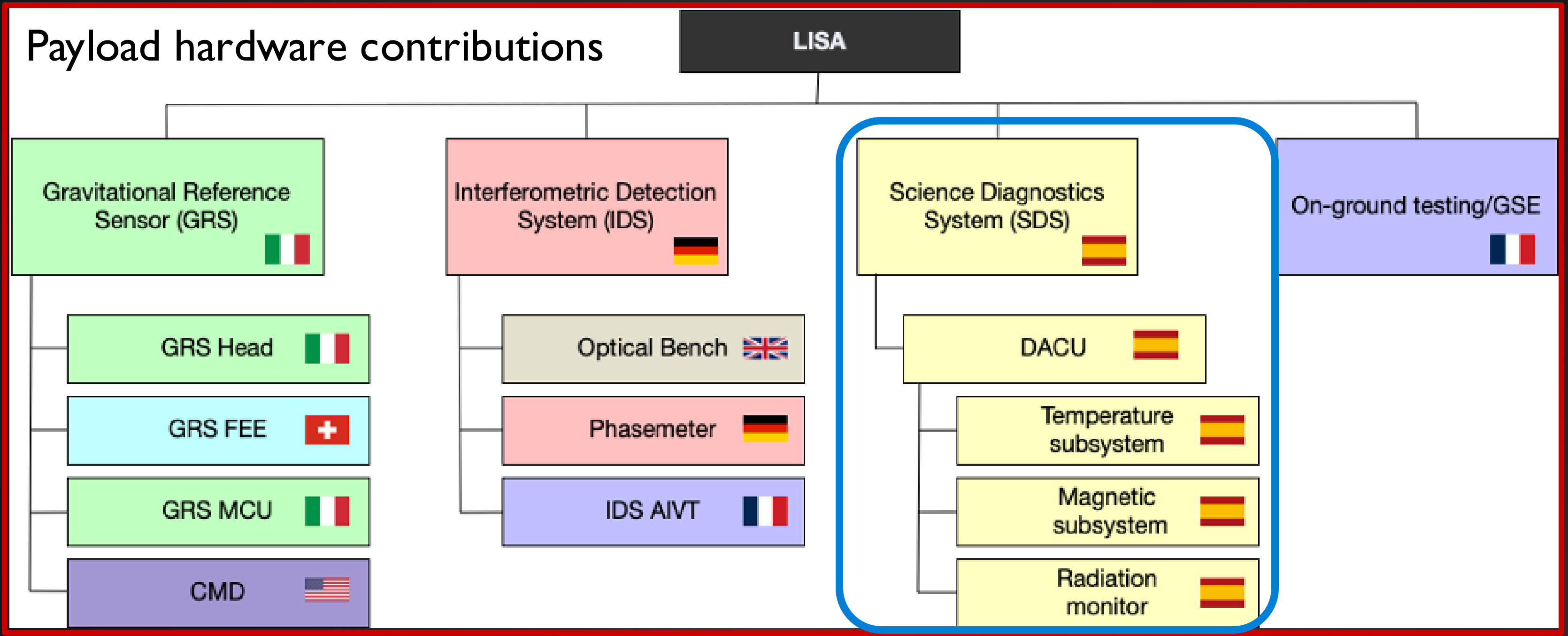
Institute of Space Sciences
EXCELENCIA MARÍA DE MAEZTU
CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
IEEC Institut d'Estudis Espacials de Catalunya
UAB Universitat Autònoma de Barcelona

The LISA mission

- Space-based GW observatory
- 3 spacecrafts in an interferometer configuration
- Arm length: 2.5M km
- Heliocentric orbit behind Earth
- ESA-led, with NASA as junior partner
- Estimated launch: ~2035
- Estimated adoption: in 3 days!



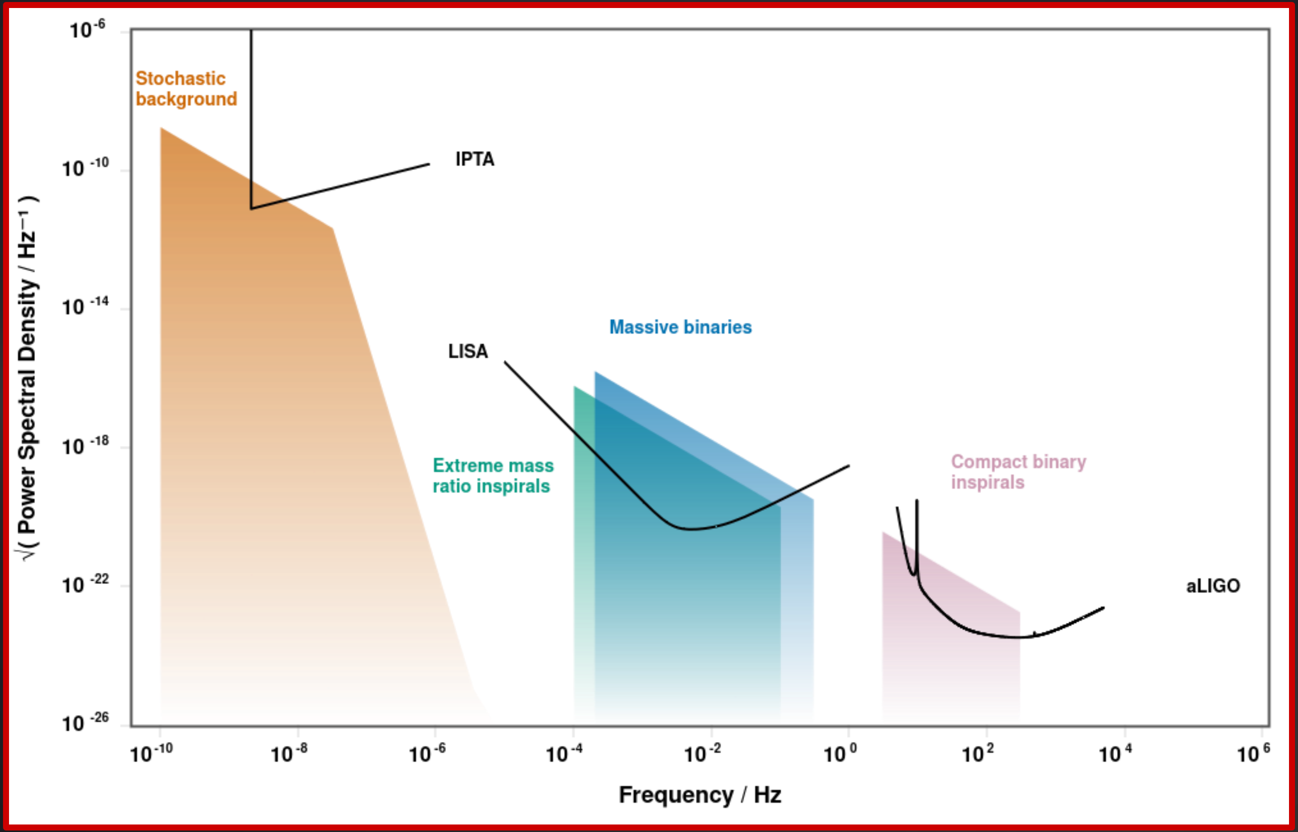
Spanish contribution to LISA



(source: ESA)

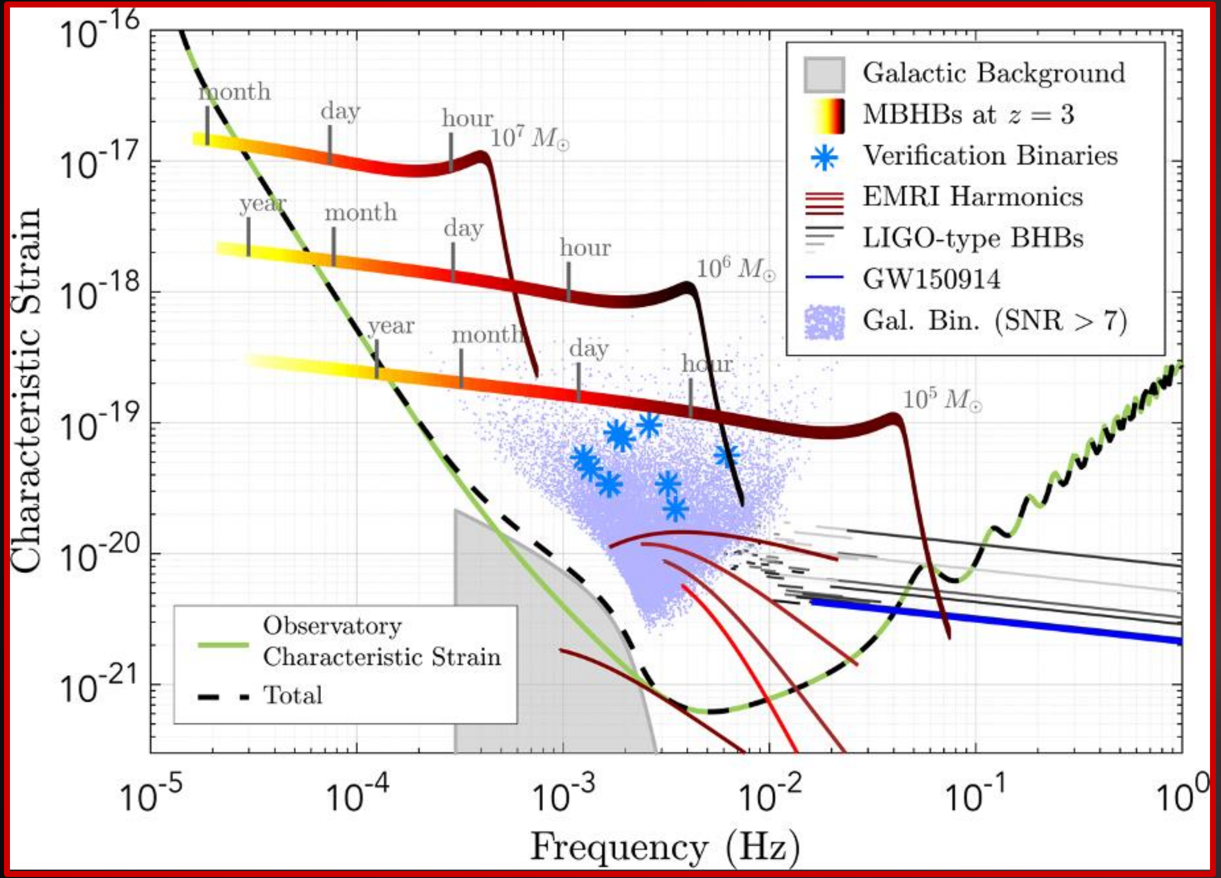
- In the ground segment, Spain is committing to:
  - Establishing a Data Processing Centre (DPC)
  - Developing and running a Global Fit pipeline

The LISA frequency band

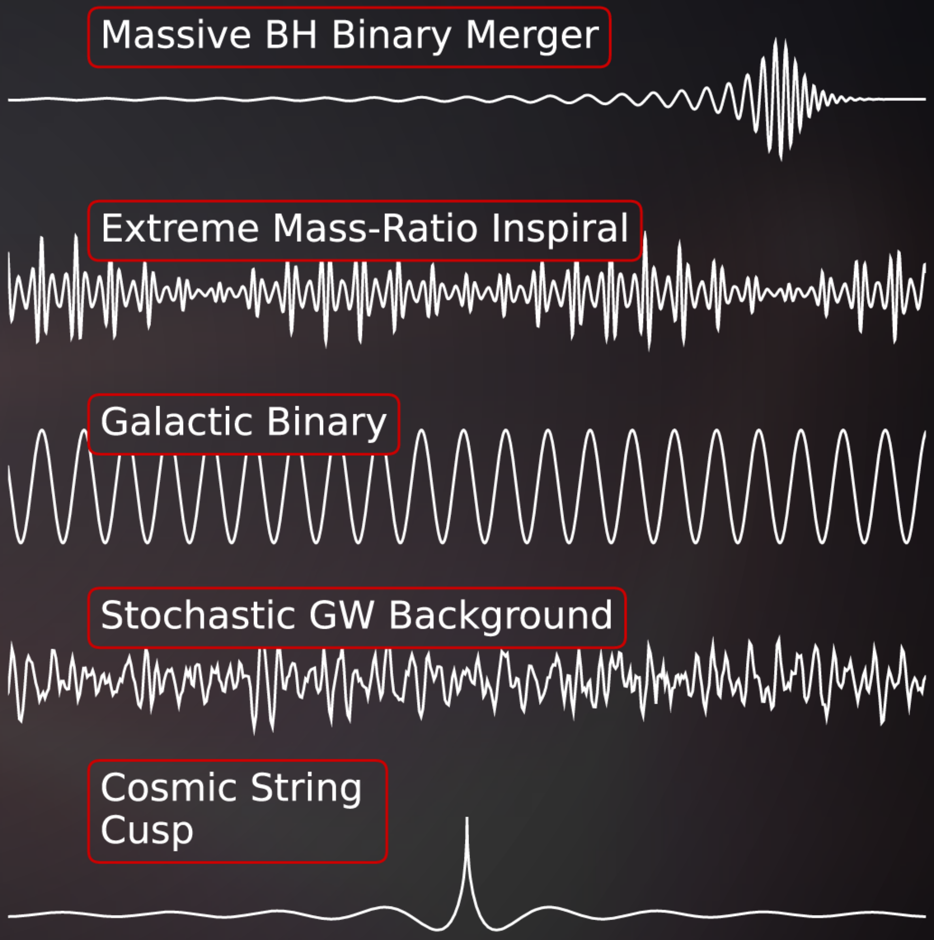


(source: C. Moore et al., gwplotter.com)

GW sources in LISA



(source: K. Danzmann et al., LISA mission proposal, arXiv:1702.00786)



LISA Science Objectives

- SO1: Study the formation and evolution of compact binary stars and the structure of the Milky Way Galaxy
- SO2: Trace the origins, growth and merger histories of massive Black Holes
- SO3: Probe the properties and immediate environments of Black Holes in the local Universe using EMRIs and IMRIs
- SO4: Understand the astrophysics of stellar-mass Black Holes
- SO5: Explore the fundamental nature of gravity and Black Holes
- SO6: Probe the rate of expansion of the Universe with standard sirens
- SO7: Understand stochastic GW backgrounds and their implications for the early Universe and TeV-scale particle physics
- SO8: Search for GW bursts and unforeseen sources

LISA Ground Segment

L0 data

Raw phasemeter data

L1 data

Time-Delay Interferometry

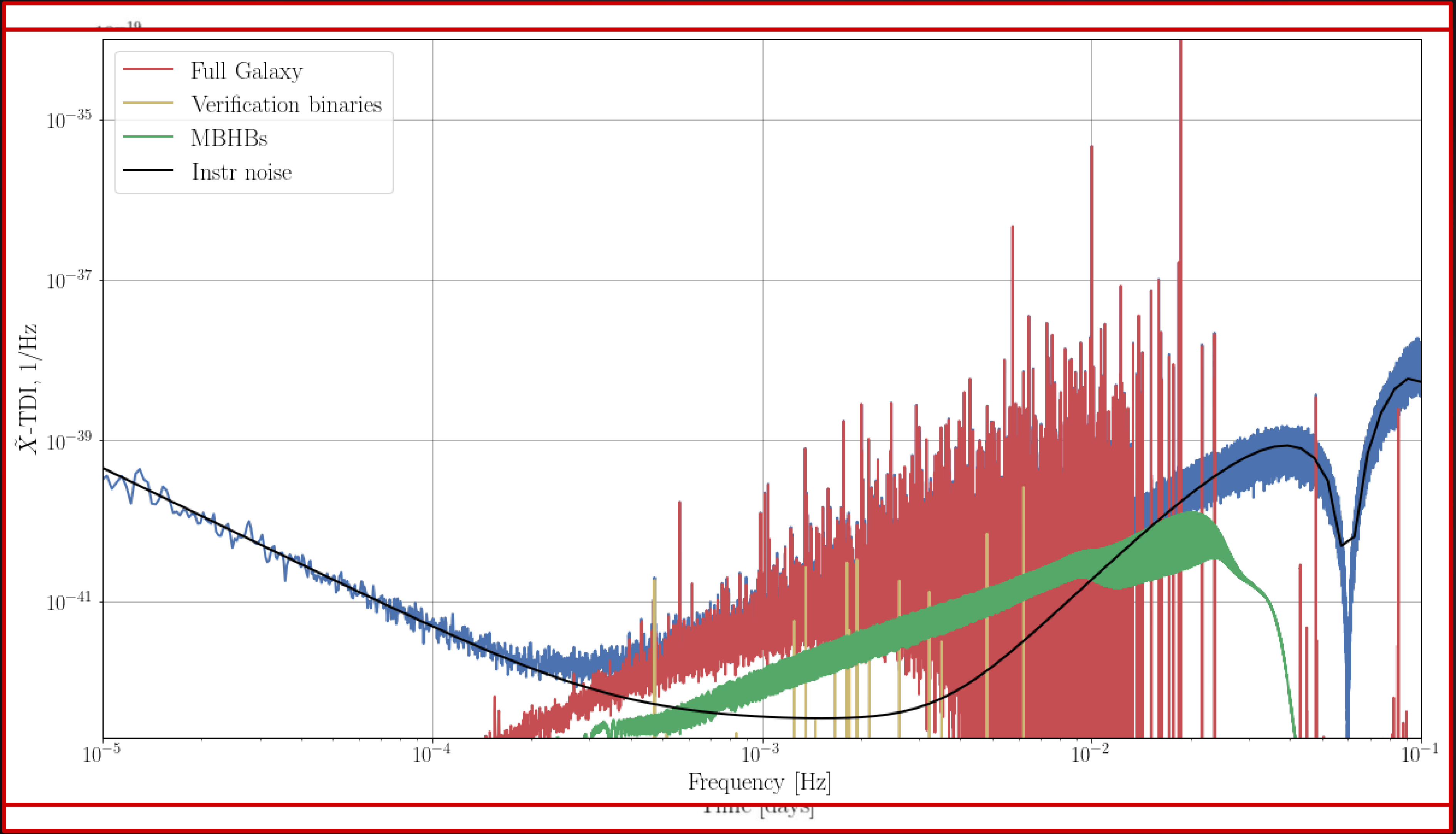
L2 data

Identified source waveforms

L3 data

Source catalogs for astrophysicists

The LISA Data Challenges



Fundamentals of GW data analysis

- Our signal is $\mathbf{d}(t;\boldsymbol{\theta}) = \mathbf{h}(t;\boldsymbol{\theta}) + \mathbf{n}(t)$

  deterministic
  depends only on $\boldsymbol{\theta}$

  realization of a
  random process

- If we assume $\mathbf{n}(t) = \mathbf{d}(t) - \mathbf{h}(t;\boldsymbol{\theta})$ is a Gaussian, stationary random process:

$$p(\mathbf{n} = \mathbf{n}_0) \propto \exp\left[-\frac{1}{2}\,(\mathbf{n}_0|\mathbf{n}_0)\right] = \exp\left[-\frac{1}{2}\,(\mathbf{d}-\mathbf{h}|\mathbf{d}-\mathbf{h})\right]$$

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

$$\mathrm{SNR}[\mathbf{h}] \simeq \sqrt{(\mathbf{h}|\mathbf{h})}$$

$$\rho(\mathbf{h};\mathbf{d}) = \frac{(\mathbf{h}|\mathbf{d})}{\sqrt{(\mathbf{h}|\mathbf{h})}}$$

- Multiple data channels? The inner product then becomes:

$$(\mathbf{a}|\mathbf{b}) = 2\sum_{i,j}\int_0^\infty \mathrm{d}f\; \tilde{\mathbf{a}}_i^*(f)[\mathbf{S}_n(f)^{-1}]^{ij}\tilde{\mathbf{b}}_j(f)\,.$$

(it simplifies to a sum over streams if the channels are independent, e.g. TDI AET)

Bayesian inference

Likelihood $\log p(\,\mathbf{d}\mid\boldsymbol{\theta}\,) = \sum_{n=\{A,E,T\}} 2\,(\,\mathbf{h}(\boldsymbol{\theta})\mid\mathbf{d}\,)_n - (\,\mathbf{h}(\boldsymbol{\theta})\mid\mathbf{h}(\boldsymbol{\theta})\,)_n$

integrals!

forward simulations!

$$p(\,\boldsymbol{\theta}\mid\mathbf{d}\,) = \frac{p(\,\mathbf{d}\mid\boldsymbol{\theta}\,)\,p(\boldsymbol{\theta})}{p(\mathbf{d})}$$

Prior

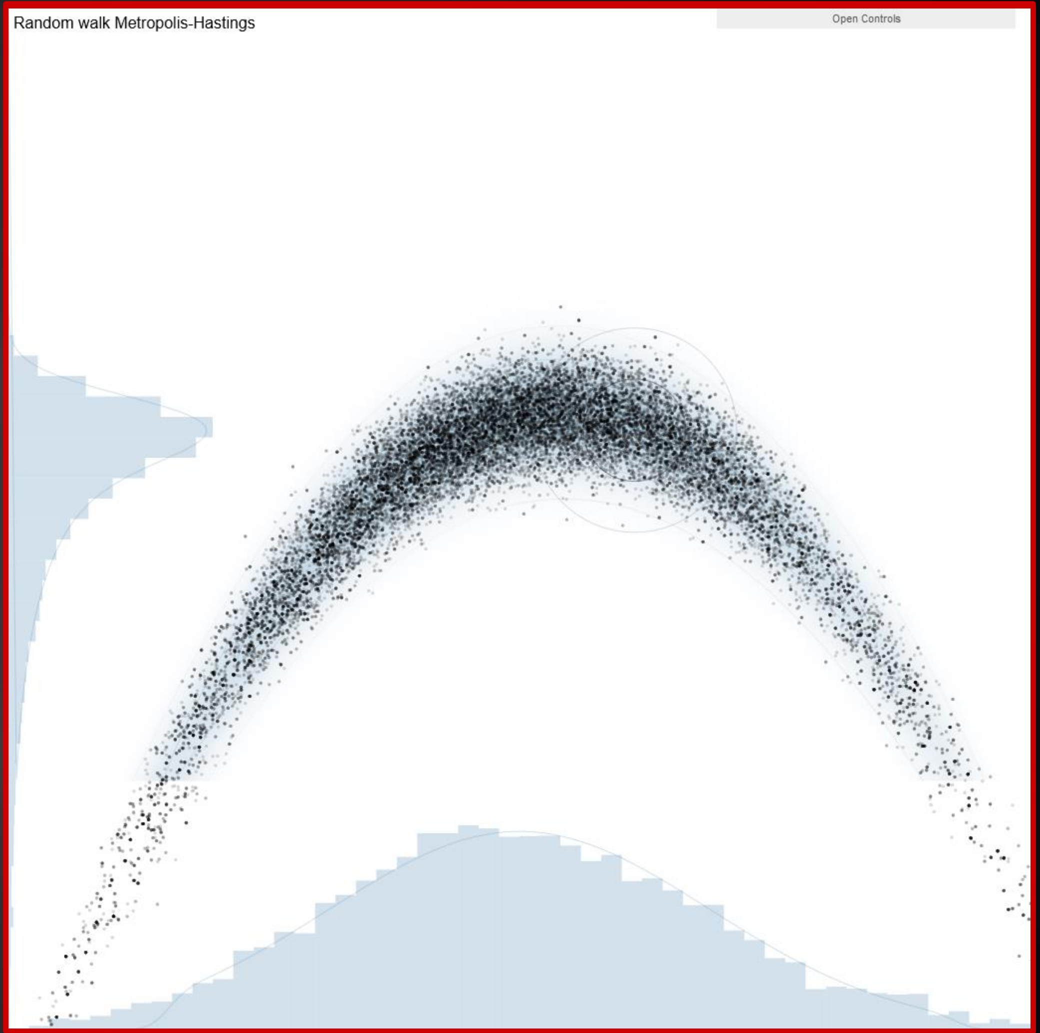
Posterior

Evidence

$$p(\mathbf{d}) = \int p(\,\mathbf{d}\mid\boldsymbol{\theta}\,)\,d\boldsymbol{\theta}$$

Markov Chain Monte Carlo (MCMC)

- Random number generation that can follow an arbitrary distribution *eventually*
- Propose points, then accept/reject with some criterion
- Large zoo of algorithms, some more suitable than others depending on situation
- Lots of evaluations of the likelihood ⇒ potentially slow



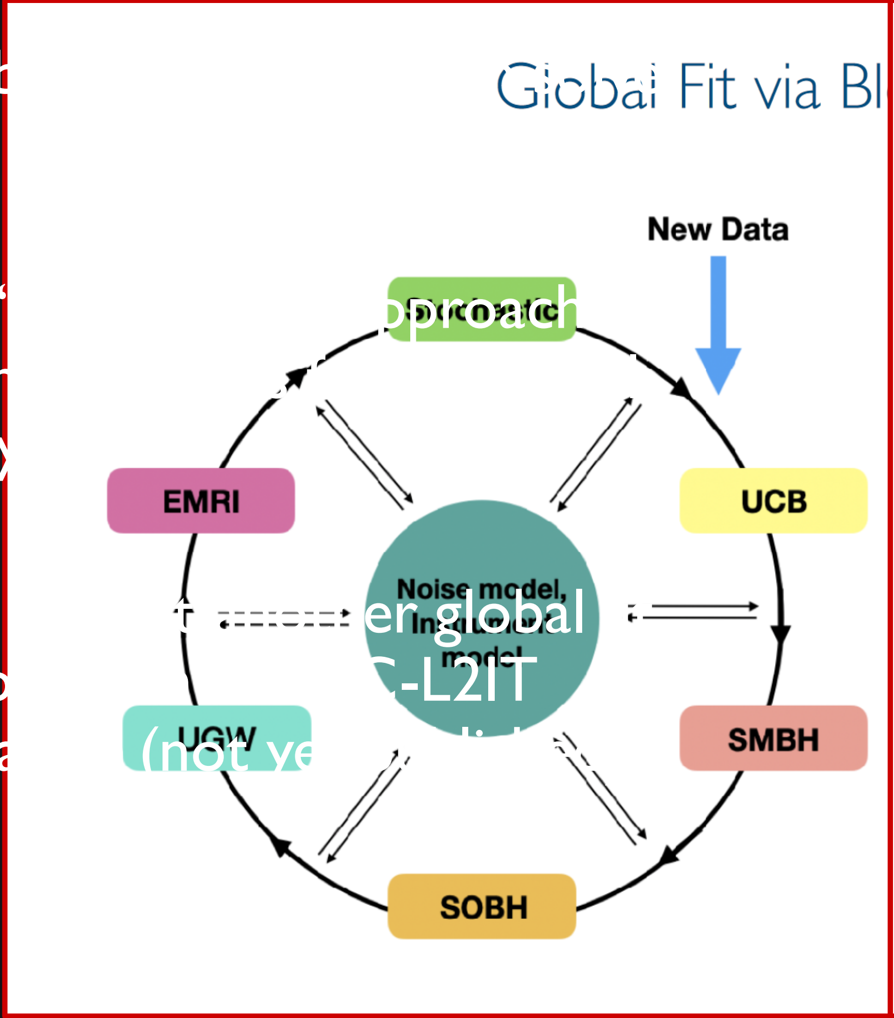
(source: Chi Feng, *The Markov Chain Monte Carlo Interactive Gallery*)

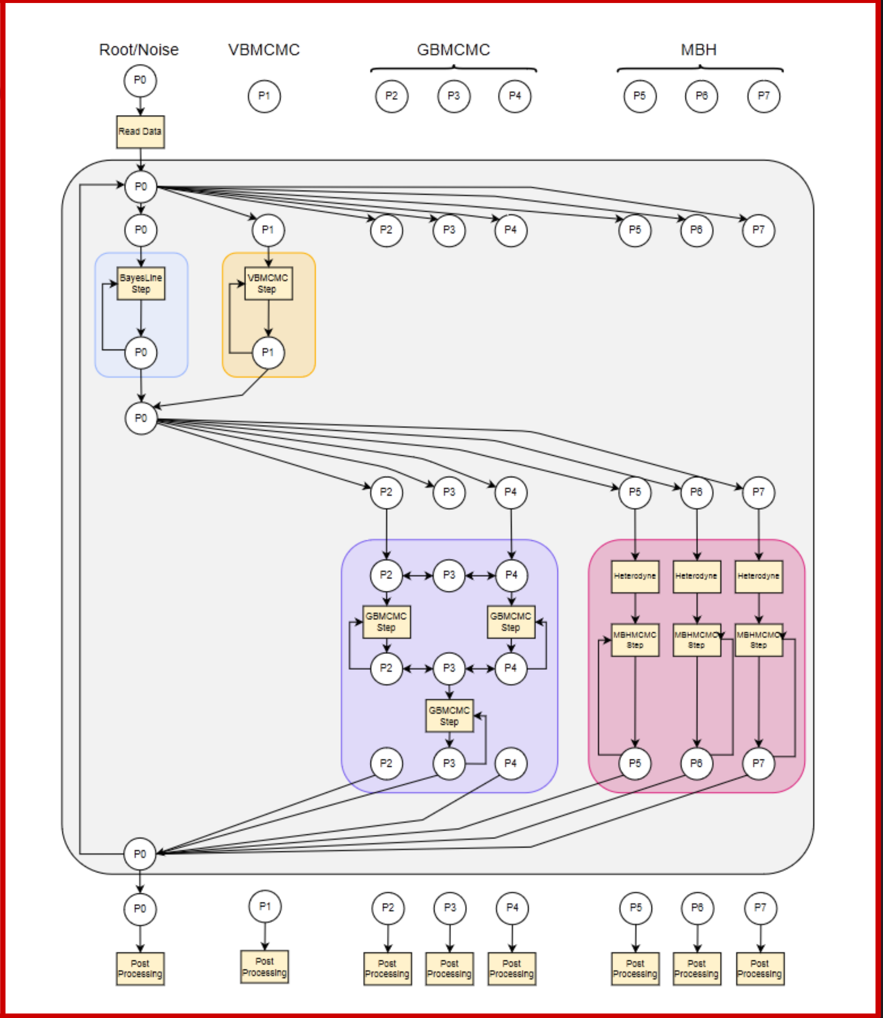
The need for a Global Fit

- In reality, our data will contain:
  - Non-stationary instrumental noise (+glitches, gaps)
  - $\mathcal{O}(10)$ MBHBs/year
  - $\mathcal{O}(10^7)$ GBs (!!!)
    - Of these, $\mathcal{O}(10^4)$ will be individually resolvable
    - There are 37 known GBs: verification binaries
  - $\mathcal{O}(1-1000)$ EMRIs
  - $\mathcal{O}(10^4)$ SBBHs
  - SGWBs
  - maybe some surprises...?
- Since one source may interfere with another, this all needs to be extracted and estimated, at once!

First Global Fit prototype

- First glob
  LDC2
- Kind of “ pproach
  combinin
  source ty
- There is t er global
  prototyp C-L2IT
  collabora (not ye





(source: T. Littenberg & N.Cornish, 2023)

(slide ~~stolen~~ graciously borrowed from: N. Cornish & T. Littenberg, “The LISA Global Fit @ Montana-MSFC”, Towards LISA catalogs workshop)

Computational requirements: the DDPC

- Based on LDC1 results, some studies have been made on the computational cost of 1 year of LISA analysis.
  - Assuming current techniques and usage of CPUs. Using GPUs may accelerate this.
  - Storage requirement is very high for intermediate data (*"scratch volume"*) but reasonable for the end products.
  - We need to account for potential bugs/recalibration and having alternative Global Fits as backup/reassurance ⇒ multiply all these numbers by a factor of ~6.
- We need to build a DDPC to analyse all this data.
- It is imperative that we find ways to increase the efficiency of data analysis! Either through "tricks" in existing algorithms or by using entirely new ones.

Requirements for analysis of 1 year of LISA data for individual sources

Source type	CPU-hours	Scratch volume	Informative volume
Galaxy	(180-250)K	(260-2000)GB	120GB
MBHBs	(1.2-300)K	(5-50)TB	(0.6-6)GB
EMRIs	(4-6)M	(16-24)TB	(12-20)GB
SBBHs	14M	100TB	200MB
Noise	(0.5-3)K	(50-260)GB	(1-5)GB

Per year of data	CPU-hours	Scratch volume	Informative volume
#1 With SBBH	30M	500TB	160GB
#1 without SBBH	17M	225TB	160GB
#2 with SBBH	(14.5-52.5)M	500TB	160GB
#2 without	(7.5-15)M	225TB	160GB
Low-latency	550K	52TB	6GB

Table 2: Summary of the resources required for detection and characterization of all individually resolvable sources in 1 year of LISA data.

(source: S. Babak, "Computational cost and storage for L1, L2, L3 production", technical report ref. LISA-TN-001, 2021)

The Spanish LISA Data Analysis collaboration

- Currently beginning to take shape
- The plan: develop a Global Fit pipeline that uses more "experimental" ML techniques (+ all the tools/infrastructure needed to support it)
- Currently, we have collaborators from:

Institute of Space Sciences
EXCELENCIA MARÍA DE MAEZTU
CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
IEEC
Institut d'Estudis Espacials de Catalunya
UNIVERSITAT DE BARCELONA
IFIC
INSTITUT DE FÍSICA CORPUSCULAR

- Interested? Get in touch!

Thank you!