Data Analysts Without Borders: A Case Study

Warren G. Anderson
University of Wisconsin - Milwaukee

UNIVERSITY of WISCONSIN MILWAUKEE

Image courtesy of NASA Goddard NR group

Outline

- Foundation
  - Four Data Analysis (DA) groups
- Formulation
  - GWs, IFOs, Data, Likelihood, Power, Power $^{2.0}$
- Application
  - Searches for modeled signals, backgrounds, bursts
- Migration
  - Real life, burst or glitch, STAMP
- Conclusion

Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category		

Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	

Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration

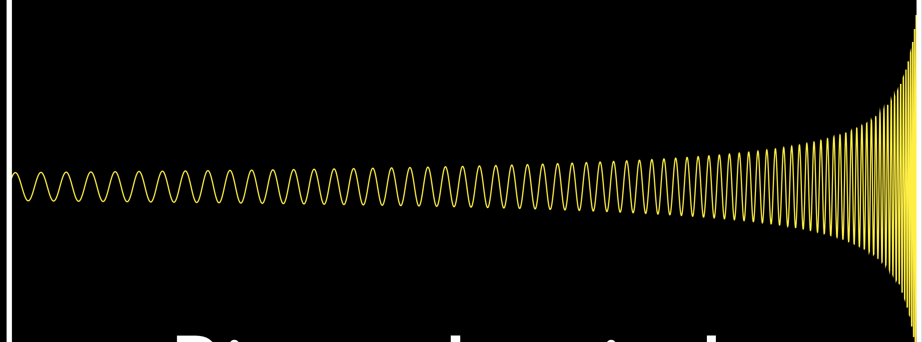
Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration
Theoretical Waveform		

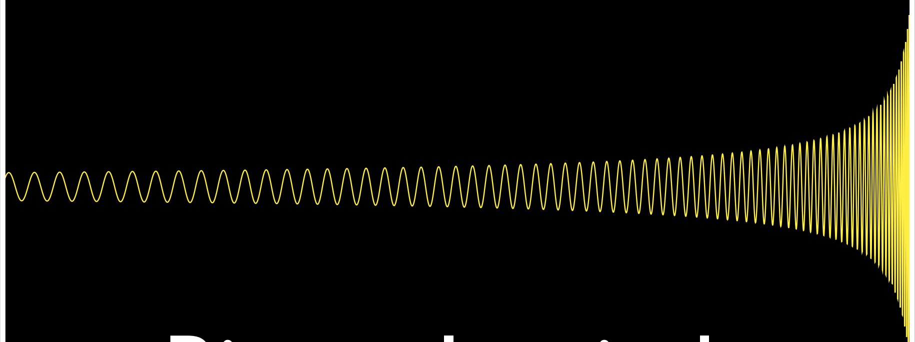
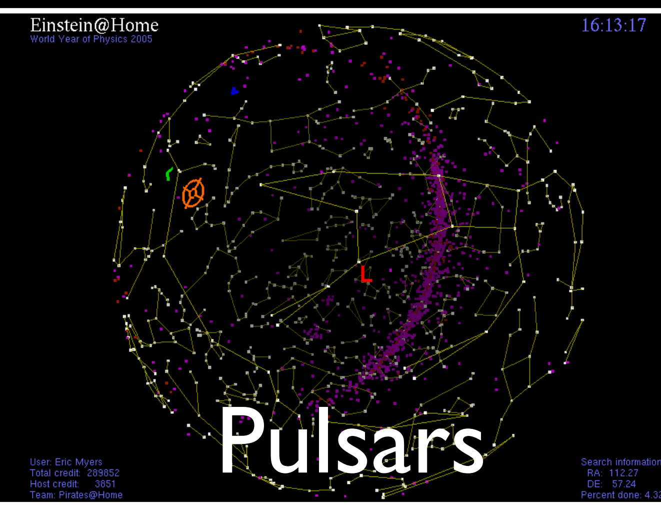
Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration
Theoretical Waveform	 Binary Inspirals	

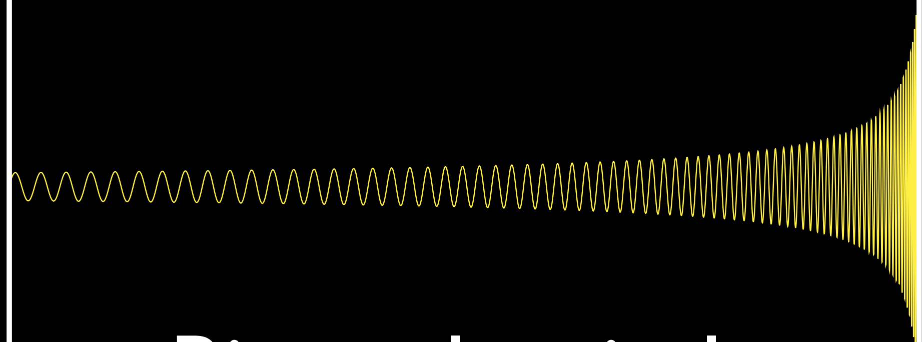
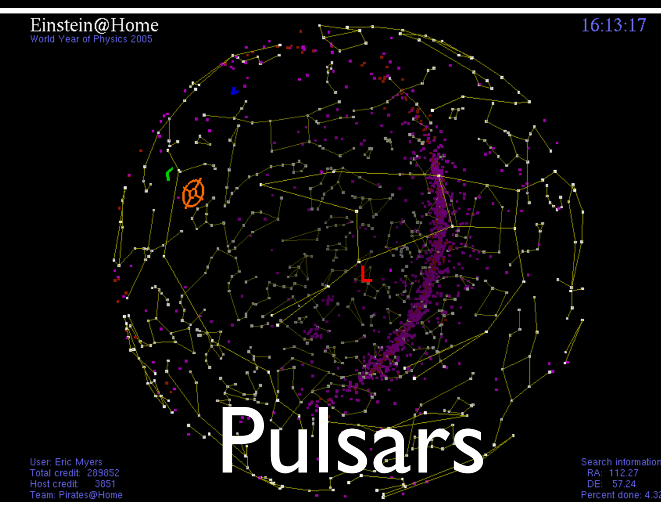
Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration
Theoretical Waveform	 Binary Inspirals	 Pulsars

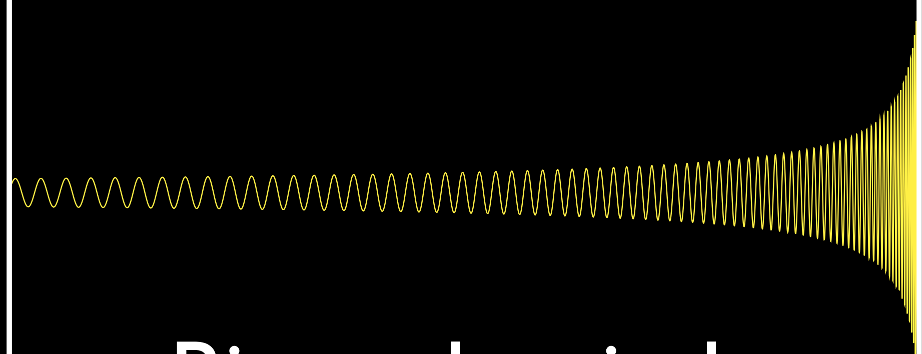
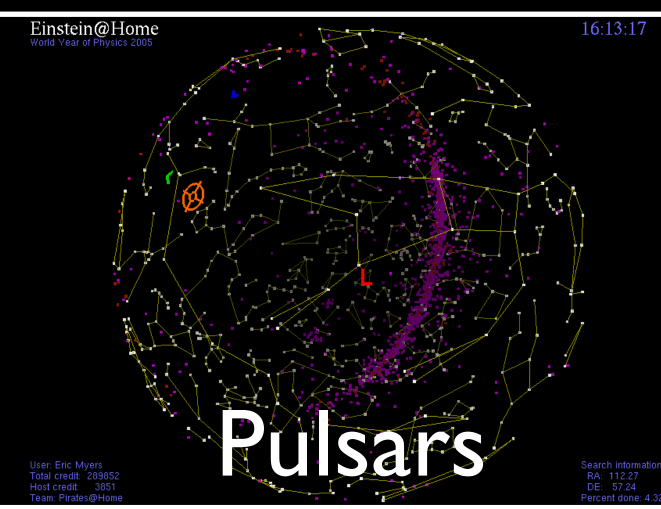
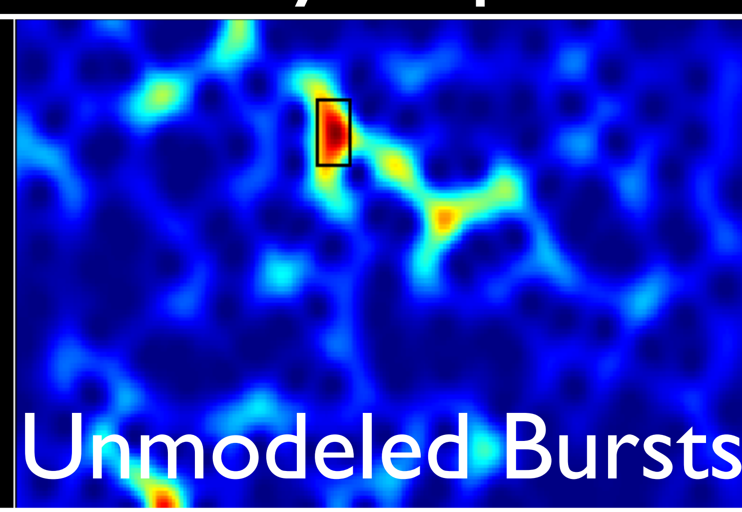
Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration
Theoretical Waveform	 Binary Inspirals	 Pulsars
No Theoretical Waveform		

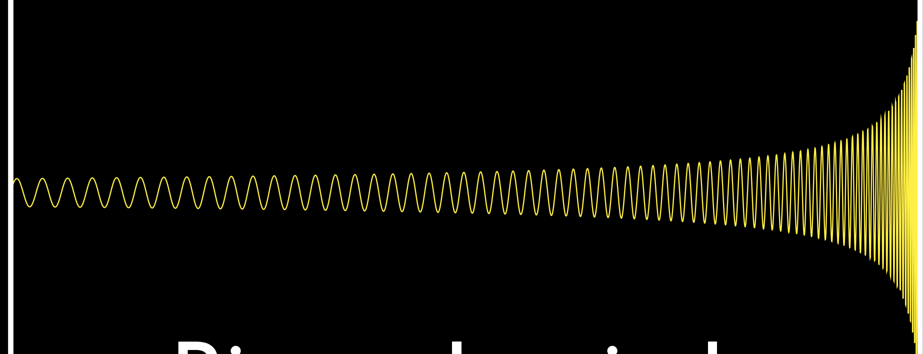
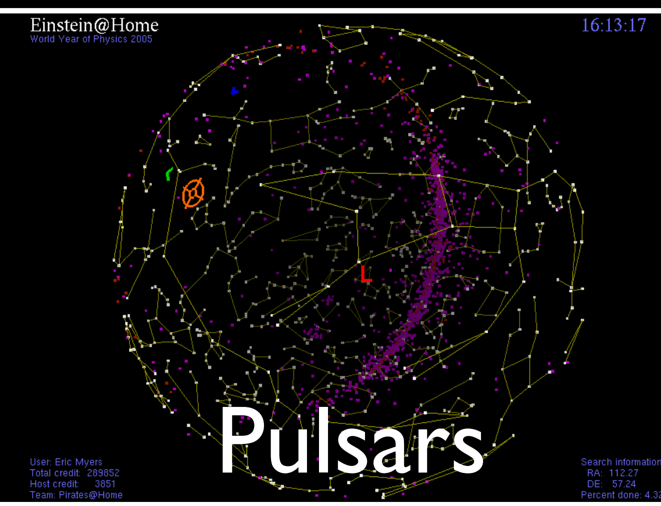
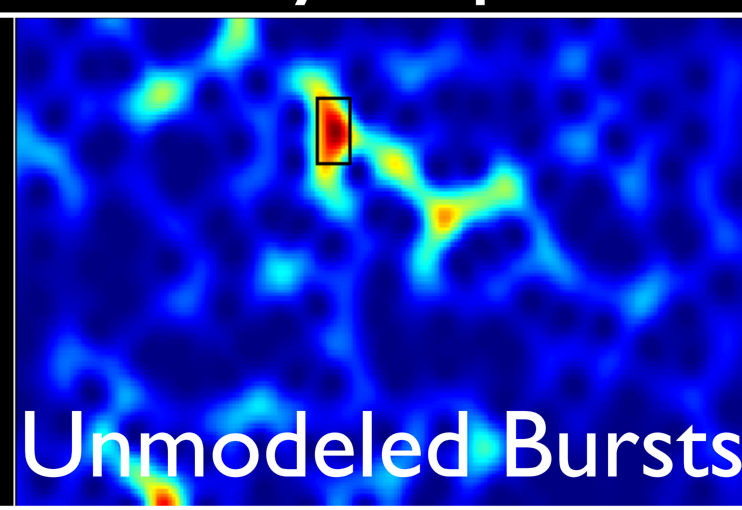
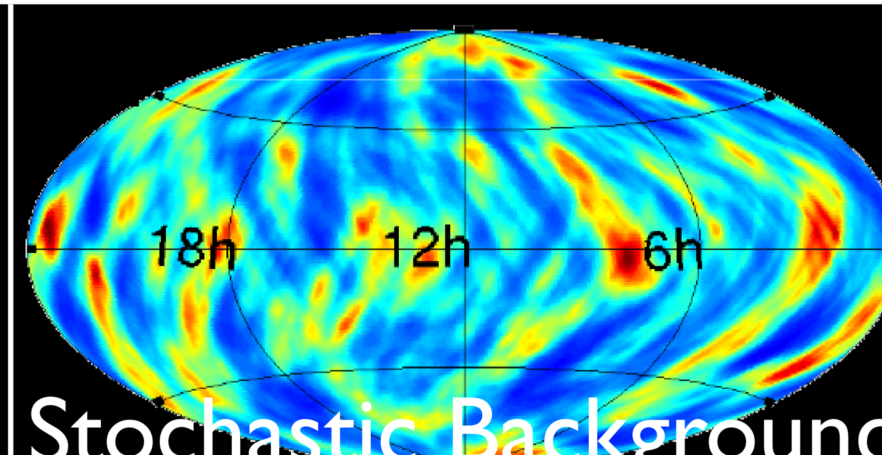
Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration
Theoretical Waveform	 Binary Inspirals	 Pulsars
No Theoretical Waveform	 Unmodeled Bursts	

Foundation

- Ground-based gravitational wave DA is traditionally divided into four categories:

Category	Short Duration	Long Duration
Theoretical Waveform	 Binary Inspirals	 Pulsars
No Theoretical Waveform	 Unmodeled Bursts	 Stochastic Background

Foundation

- Data analysts are usually divided into four groups based on the categories of signal they are looking for.
- Questions:

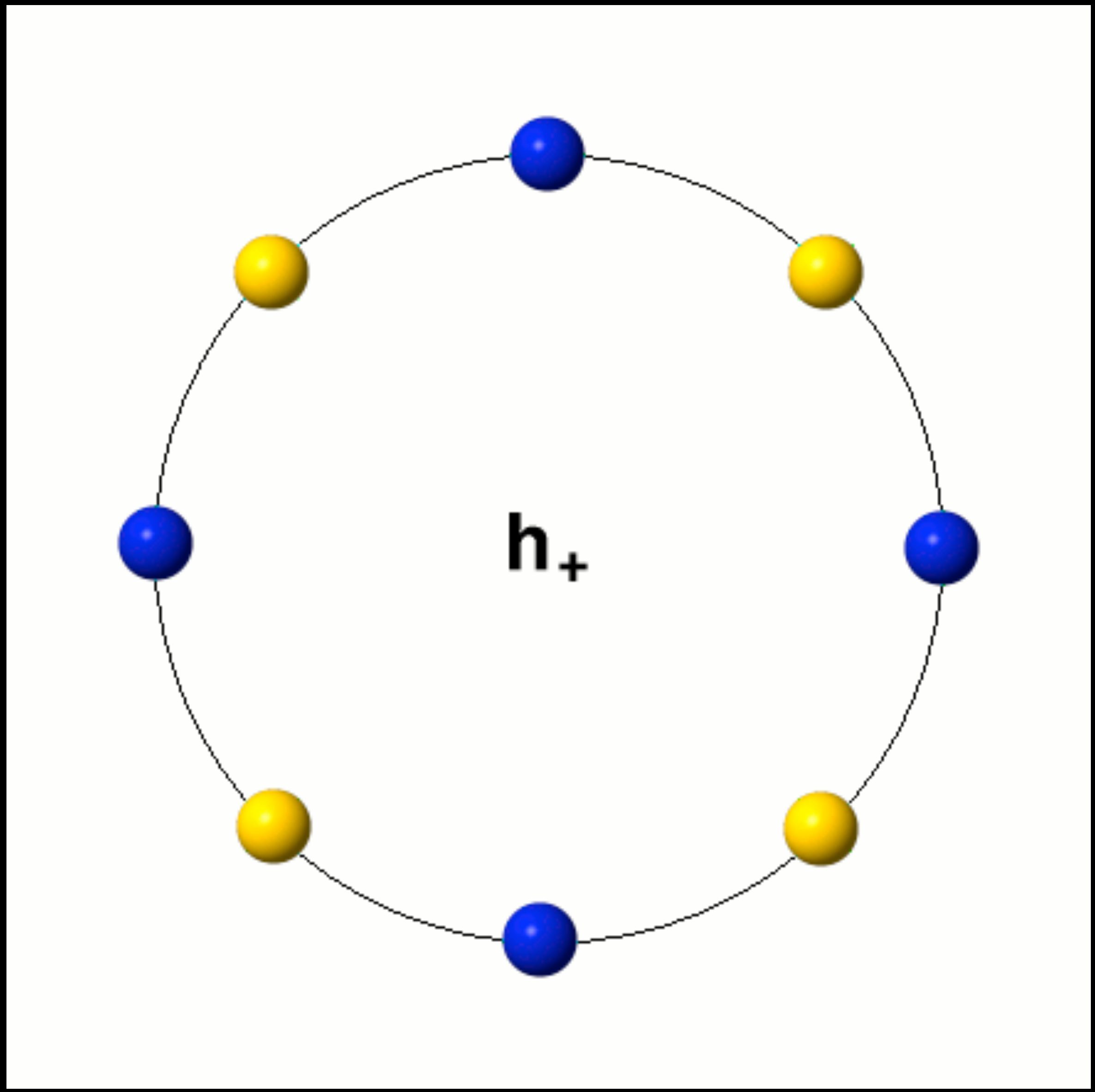
  "Are these categories fundamental or just convenient?"

  "Do all signals fit neatly into just one category?"

  "Is there one search method that is optimal for all signals in each category?"

Formulation: GWs

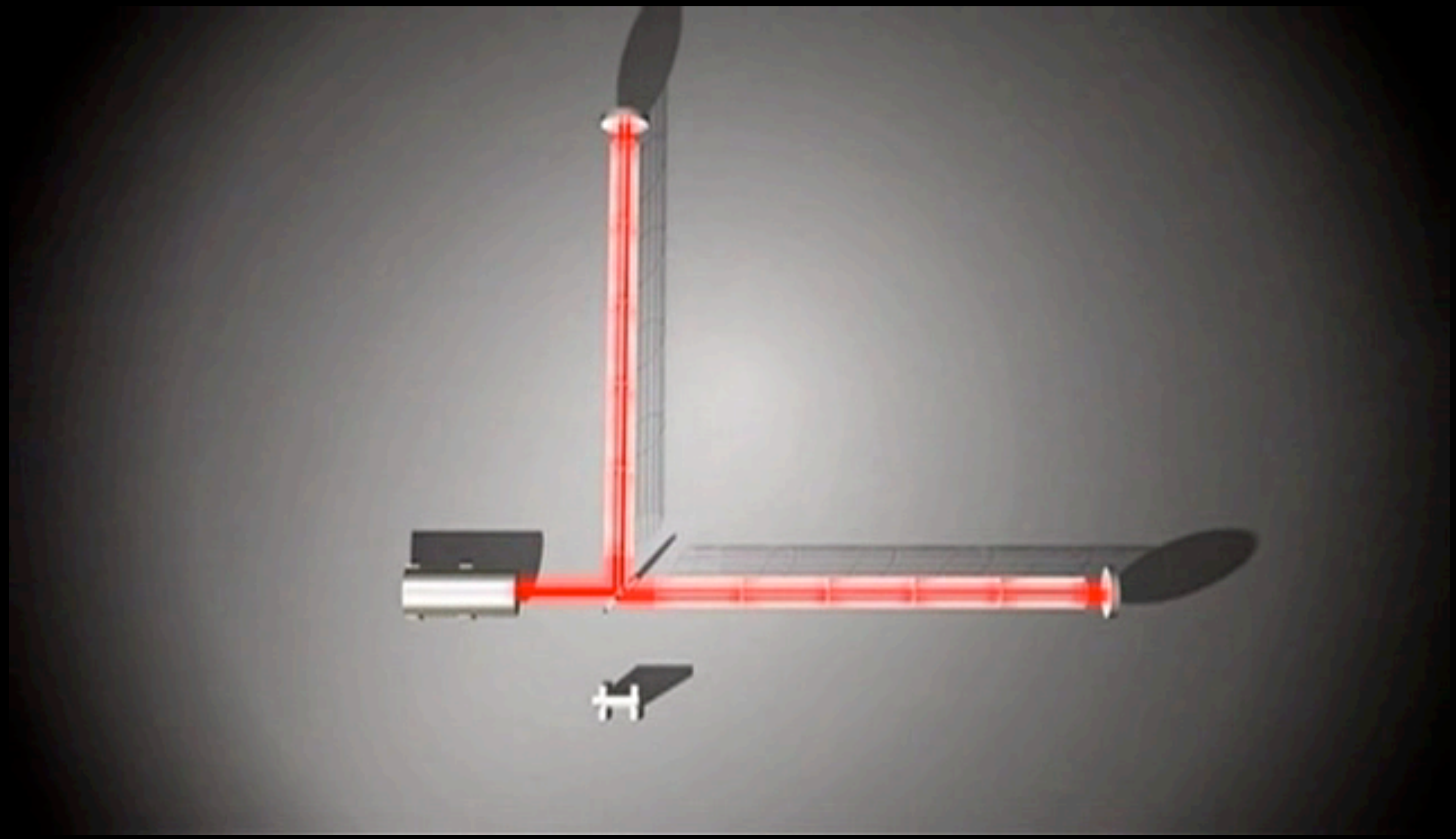
Formulation: GWs



Formulation: IFOs

From "Einstein's Messengers"
National Science Foundation

Formulation: IFOs



From "Einstein's Messengers"
National Science Foundation

Formulation: Data

- From our instruments we get strain data $h(t)$

$$h(t) = n(t) + s(t)$$

Noise *Signal*

Formulation: Data

- From our instruments we get strain data $h(t)$

$$h_i = n_i + s_i$$

Noise *Signal*

- Signal (usually) is deterministic (possibly $s_i \approx 0$)
- Noise is stochastic - eg Gaussian (ideally)

$$p(n) \propto \exp\left(-\frac{n \cdot n}{2\,\sigma}\right)$$

Formulation: Likelihood

- Neyman-Pearson: Optimal statistic is ***Likelihood***

$$\Lambda[h] = \int \frac{p(h|n+s)}{p(h|n)} \mathcal{D}[s]$$

- The measure $\mathcal{D}[s]$ projects the integrand from the space of all possible signals down onto the subspace of signals we are searching for, denoted by Σ .

Formulation: Likelihood

- Neyman-Pearson: Optimal statistic is ***Likelihood***

$$\Lambda[h] = \int \frac{p(h|n+s)}{p(h|n)} \boxed{\mathcal{D}[s]}$$

Signal

- The measure $\mathcal{D}[s]$ projects the integrand from the space of all possible signals down onto the subspace of signals we are searching for, denoted by Σ .

Formulation: Likelihood

- Neyman-Pearson: Optimal statistic is ***Likelihood***

$$\Lambda[h] = \int \frac{p(h|n+s)}{p(h|n)} \boxed{\mathcal{D}[s]}$$

Signal

- Probabilities of h derive from distribution of n, eg

$$p(h|n+s) \propto e^{-\frac{(h-s)\cdot(h-s)}{2\,\sigma}}$$

Formulation: Likelihood

- Neyman-Pearson: Optimal statistic is ***Likelihood***

$$\Lambda[h] = \int \boxed{\frac{p(h|n+s)}{p(h|n)}} \boxed{\mathcal{D}[s]}$$

Noise *Signal*

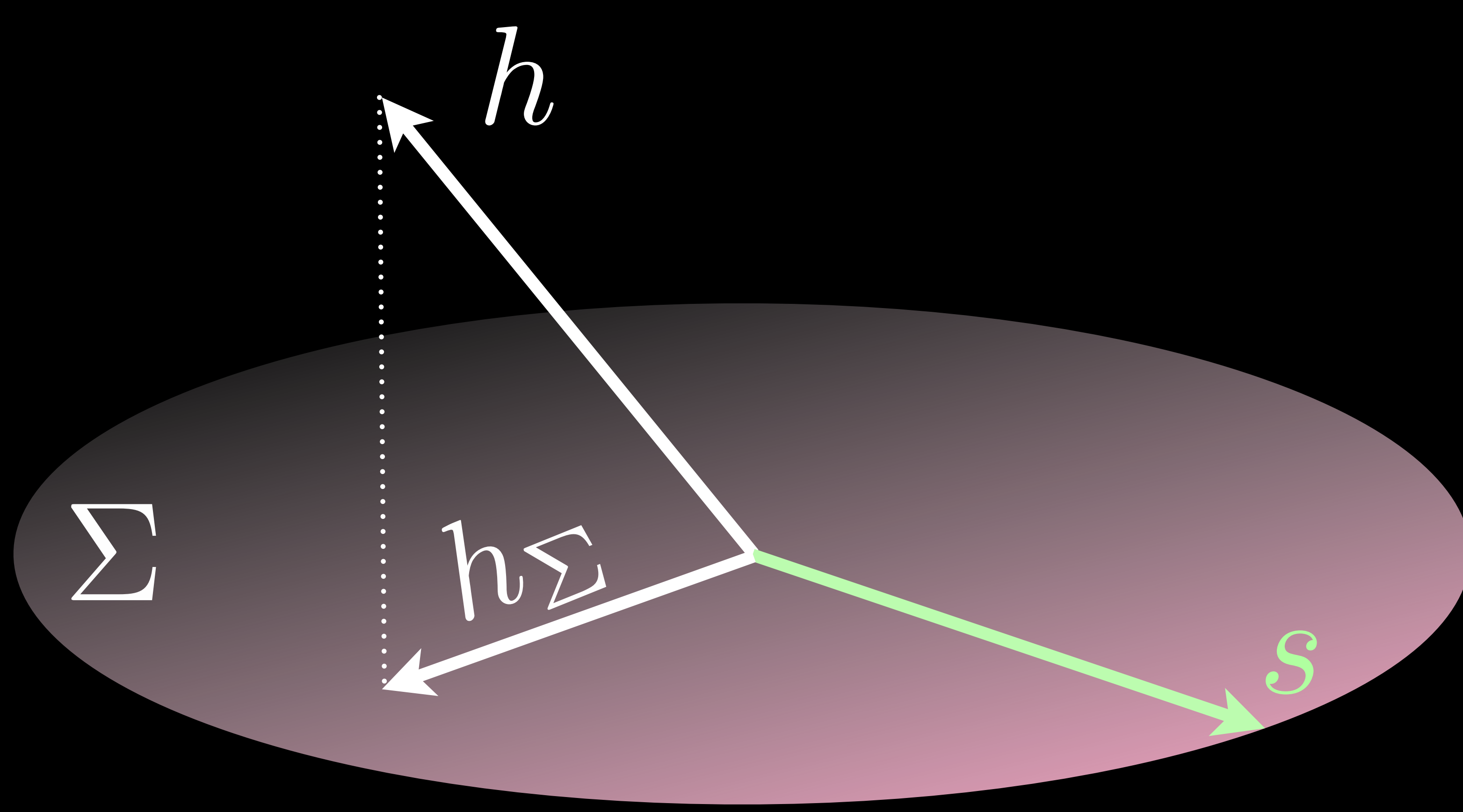
- Probabilities of h derive from distribution of n, eg

$$p(h|n+s) \propto e^{-\frac{(h-s)\cdot(h-s)}{2\,\sigma}}$$

Formulation: Power

- Putting this together for Gaussian noise:

$$\Lambda \propto \int_{\Sigma} \exp(h_{\Sigma} \cdot s - s \cdot s/2)\, ds$$



Formulation: Power

- Putting this together for Gaussian noise:

$$\Lambda \propto \int_{\Sigma} \exp(h_{\Sigma} \cdot \boldsymbol{s} - \boldsymbol{s} \cdot \boldsymbol{s}/2)\, d\boldsymbol{s}$$

- Note that Λ is monotonic in length of projection, $|h_{\Sigma}|$, so $|h_{\Sigma}|$ is equivalent statistic.
- ***Power*** of projected data, $|h_{\Sigma}|^2$, is also eqivalent optimal statistic and is most commonly used.

Formulation: Power$^{2.0}$

- For more than one detector, simply substitute:

$$h = \underbrace{\boxed{A^k(\Omega)}}_{\text{Detector Antenna Pattern}} \underbrace{\boxed{h_k(t)}}_{\text{Detector Data}}$$

Formulation: Power$^{2.0}$

- For more than one detector, simply substitute:

$$h = A^k(\Omega)\, h_k(t)$$

- *Eg*, for two detectors with Gaussian noise, optimal statistic is

$$\begin{aligned} |h_\Sigma|^2 &= |\ [A_1 h_1 + A_2 h_2]_\Sigma\ |^2 \\ &= A_1^2 |h_1^2|_\Sigma + A_2^2 |h_2^2|_\Sigma \\ &\quad + 2A_1 A_2 [h_1 \cdot h_2]_\Sigma \end{aligned}$$

Formulation: Power$^{2.0}$

- For more than one detector, simply substitute:

$$h = A^k(\Omega)\, h_k(t)$$

- *Eg*, for two detectors with Gaussian noise, optimal statistic is

$$|h_\Sigma|^2 = |\ [A_1 h_1 + A_2 h_2]_\Sigma\ |^2$$

$$= \boxed{A_1^2|h_1^2|_\Sigma + A_2^2|h_2^2|_\Sigma \quad \text{auto power}}$$

$$+\, 2A_1 A_2[h_1 \cdot h_2]_\Sigma$$

Formulation: Power$^{2.0}$

- For more than one detector, simply substitute:

$$h = A^k(\Omega)\, h_k(t)$$

- *Eg*, for two detectors with Gaussian noise, optimal statistic is

$$|h_\Sigma|^2 = \left|\ [A_1 h_1 + A_2 h_2]_\Sigma\ \right|^2$$

$$= \boxed{A_1^2 |h_1^2|_\Sigma + A_2^2 |h_2^2|_\Sigma \quad \text{auto power}}$$

$$+ \boxed{2 A_1 A_2 [h_1 \cdot h_2]_\Sigma \quad \text{cross power}}$$

Application: we know s.

- For the cases where we know the signal $\hat{s}$ we are looking for, the measure becomes a Dirac delta:

$$\mathcal{D}[s] = \delta(s - \hat{s})\, ds$$

- Integrating against this measure, $h_\Sigma = h \cdot \hat{s}/|\hat{s}|^2$.
- This is nothing but the *signal-to-noise* ratio of the matched filter search.
- If you are searching for one of a finite set of signals, search for each individually - template bank.
- This is how we search for binaries and pulsars.

Application: s is random.

- For stochastic backgrounds, by central limit theorem, resulting signal is random and Gaussian.
- Sum of Gaussians, $h = n + s$, is Gaussian. $h_\Sigma = h$.
- If $\sigma_n^2 \gg \sigma_s^2$ then $h_\Sigma \cdot h_\Sigma \approx n \cdot n$ regardless of whether signal exists, so can't use auto-power.
- However, for two instruments with data h_1 and h_2 of length N, noises n_1 and n_2 are uncorrelated, so $n_1 \cdot n_2 \sim \sqrt{N}$, but s_1 and s_2 are correlated, so $s_1 \cdot s_2 \sim N$.
- So, in the limit of large enough N, $h_1 \cdot h_2 \sim s_1 \cdot s_2$.

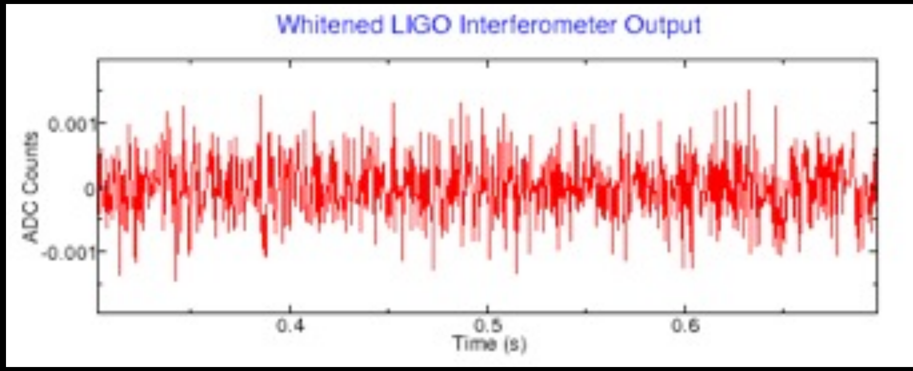
Application: s not known.

- For a determinate signal that is not completely known we can again apply likelihood.
- In this case, $\mathcal{D}[s]$ encodes what we know about the signal.
- Eg, to search for signals that last Δt seconds, use data segments h_Σ of that duration. If additionally signal frequencies are known to be in band Δf, use data segments h_Σ restricted to that band.
- So, for a single detector, an optimal statistic is auto-power for data segments restricted to $\Delta t\ \Delta f$.

Application: s not known.

- How do we practically restrict to data segments of dimension $\Delta t\ \Delta f$?

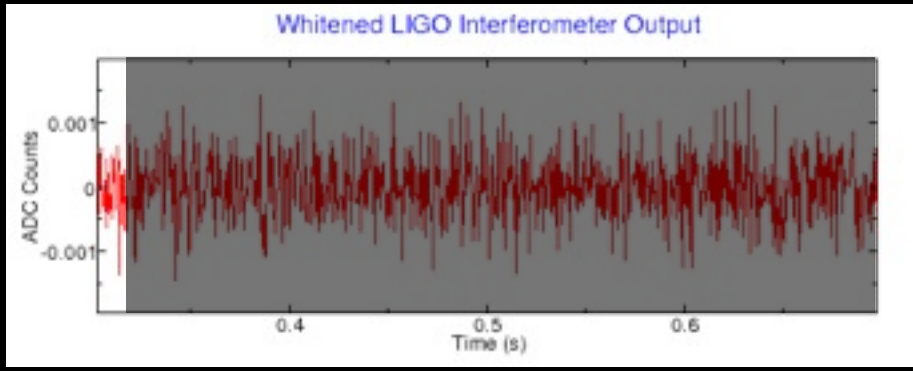
  - take a slice of detector data and Fourier transform it.



Application: s not known.

- How do we practically restrict to data segments of dimension $\Delta t\ \Delta f$?

  - take a slice of detector data and Fourier transform it.

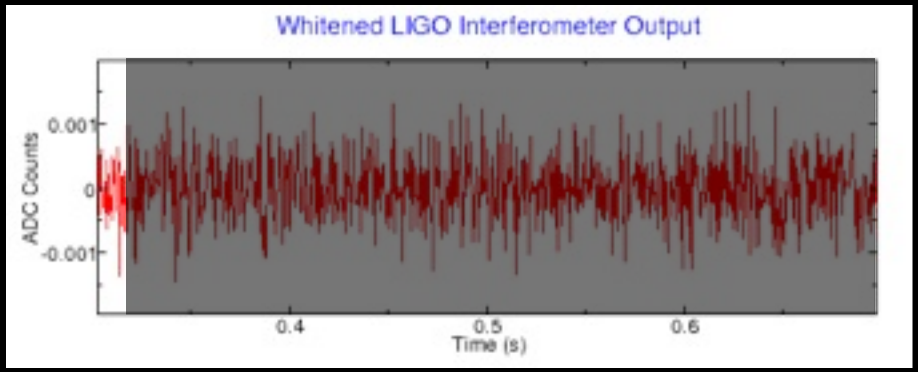


Application: s not known.

- How do we practically restrict to data segments of dimension $\Delta t \; \Delta f$?

  - take a slice of detector data and Fourier transform it.

  - plot the Fourier coefficient magnitudes on a vertical line.

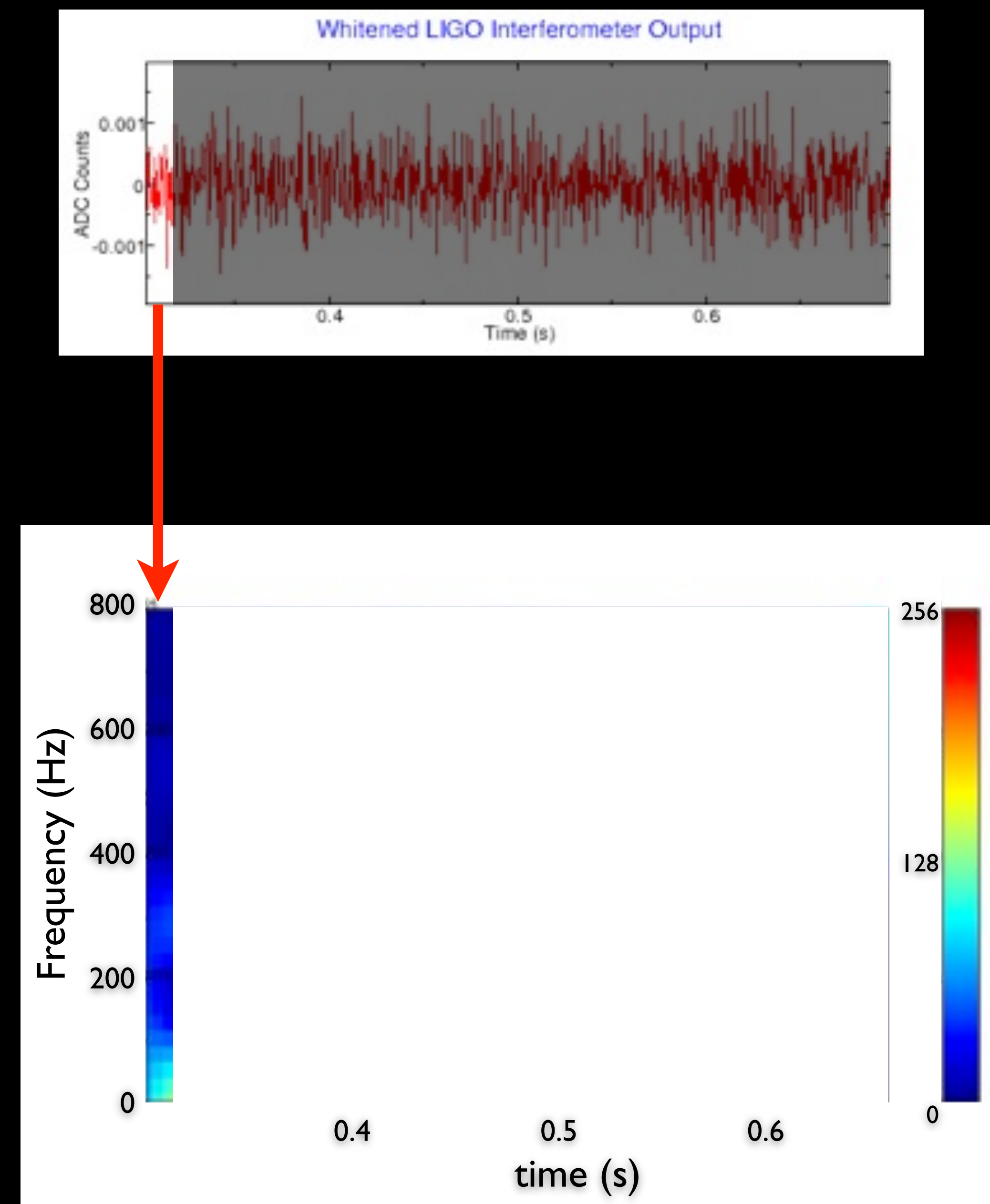


Application: s not known.

- How do we practically restrict to data segments of dimension $\Delta t \; \Delta f$?

  - take a slice of detector data and Fourier transform it.

  - plot the Fourier coefficient magnitudes on a vertical line.



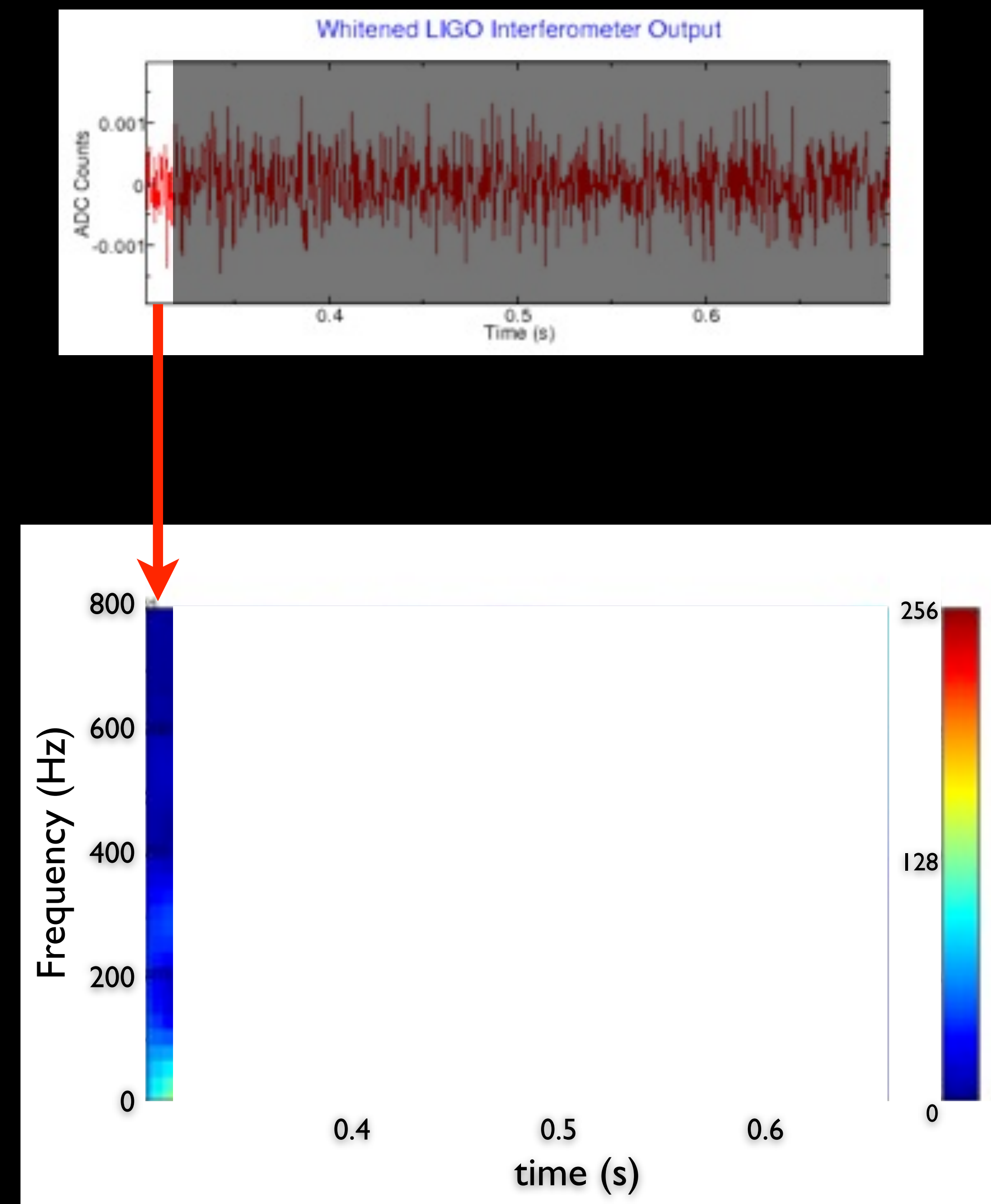
Application: s not known.

- How do we practically restrict to data segments of dimension $\Delta t \; \Delta f$?

  - take a slice of detector data and Fourier transform it.

  - plot the Fourier coefficient magnitudes on a vertical line.

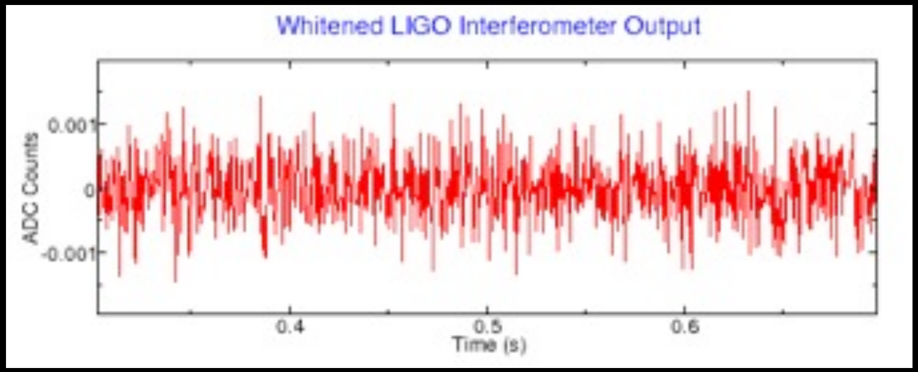
  - repeat for subsequent slices of data.



Application: s not known.

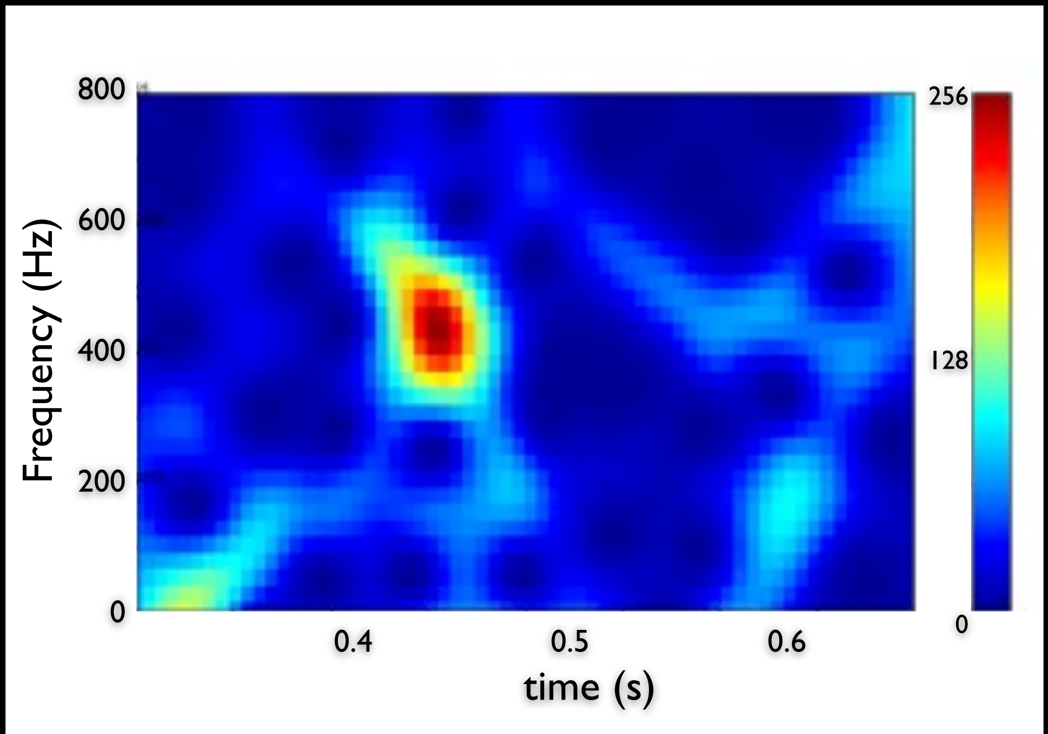
- How do we practically restrict to data segments of dimension $\Delta t\ \Delta f$?

  - take a slice of detector data and Fourier transform it.

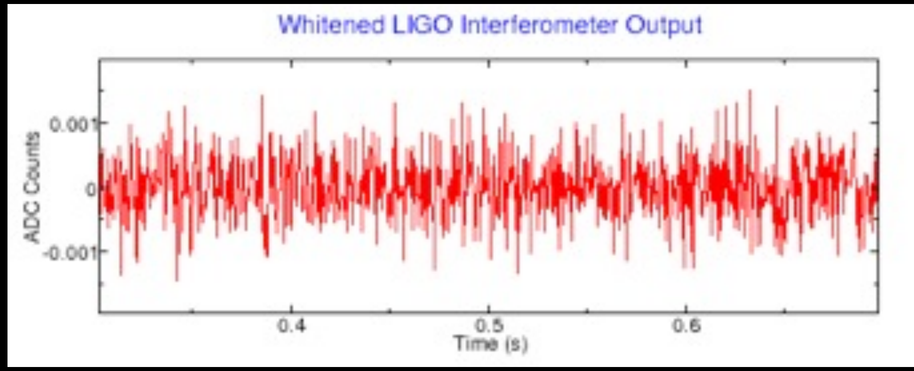
  - plot the Fourier coefficient magnitudes on a vertical line.

  - repeat for subsequent slices of data.

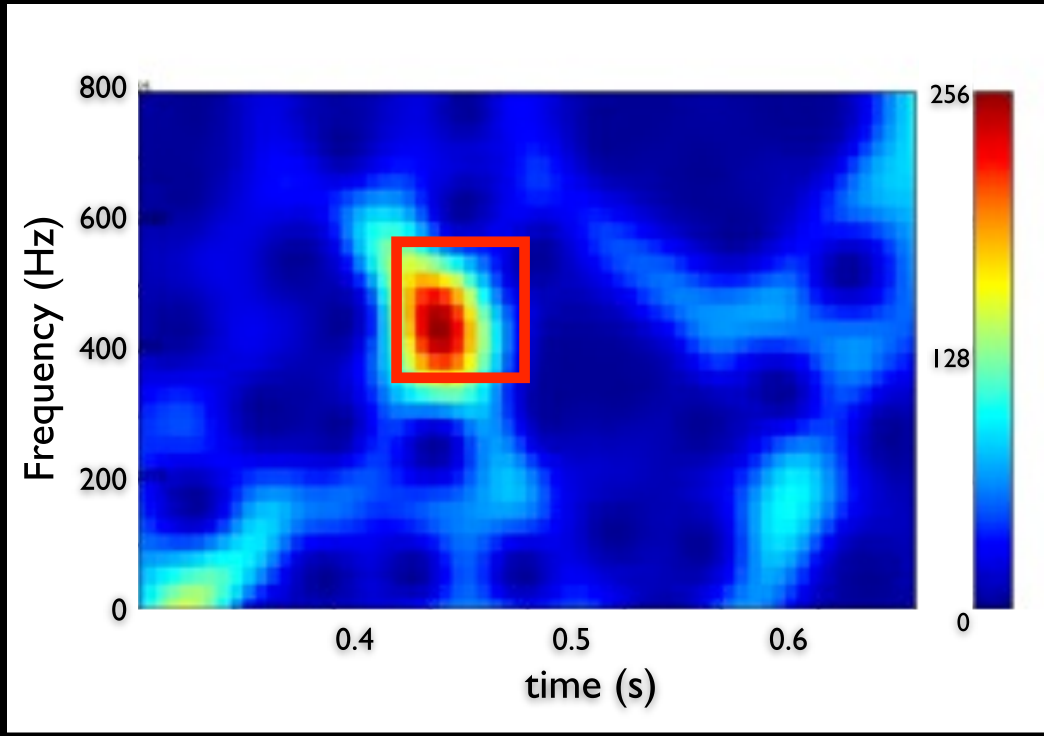
Application: s not known.

- How do we practically restrict to data segments of dimension $\Delta t \, \Delta f$?



- take a slice of detector data and Fourier transform it.

- plot the Fourier coefficient magnitudes on a vertical line.



- repeat for subsequent slices of data.

- then we can search for boxes with statistical significance.

Application: real life

- In an ideal world, these four data analysis groups and these methods would be the end of the story, but ...
- In reality:
  - signals are never exactly known
  - no theoretical model for noise statistics
  - false alarm probabilities can't be calculated
  - narrowband noise "lines" can mask signals
  - "glitches" (burst of noise) mimic real signals
  - noise not really uncorrelated between detectors
  - ...

Migration: burst? glitch?

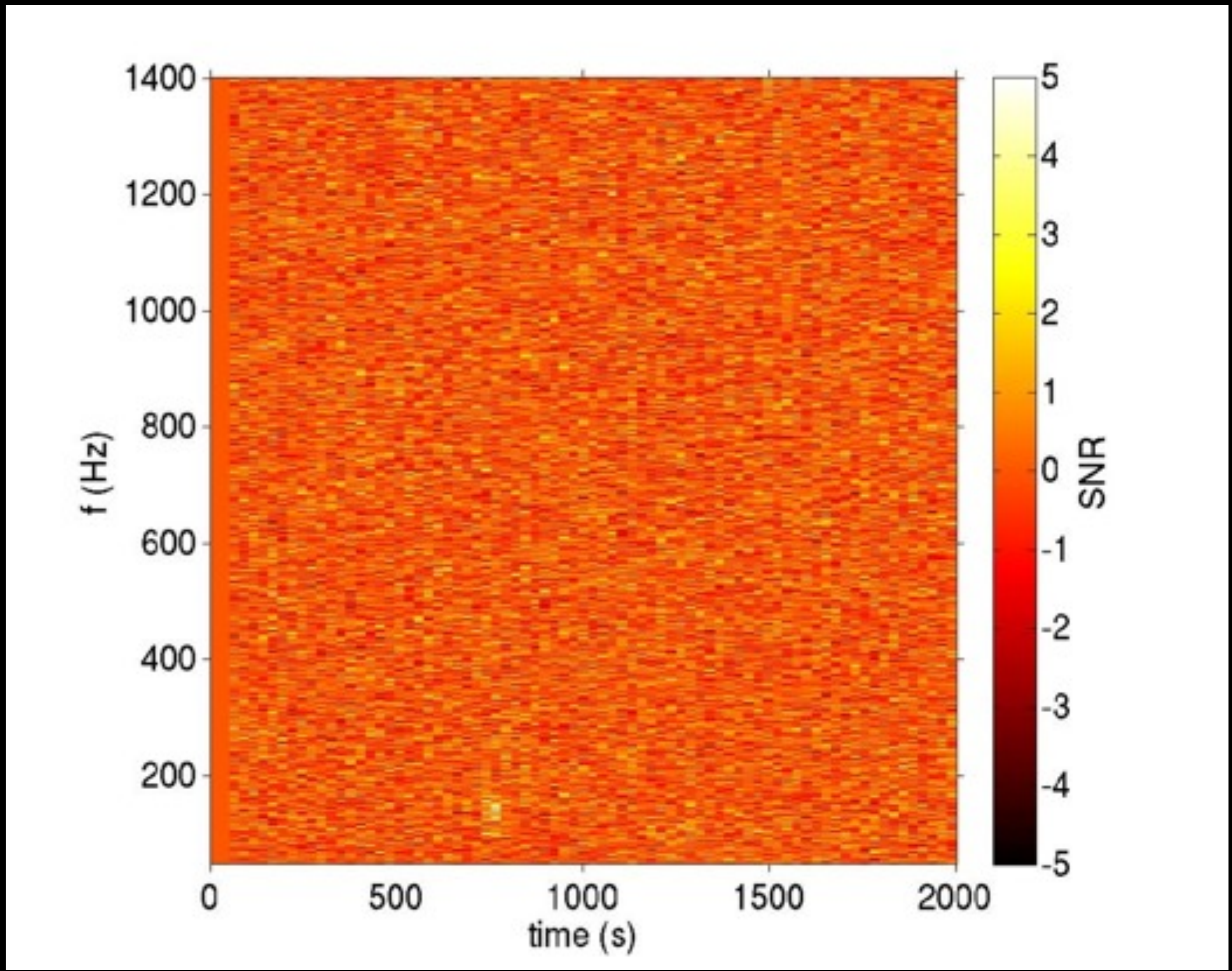
- Consider the problem looking for bursts in real noise.
- Power no longer optimal because loud glitches also cause large auto-power.
- Question: Who knows how to search for signals when you can't tell the signal from a detector's noise?
- Answer: Analysts who look for stochastic backgrounds!

Migration: STAMP

- Idea: Use cross-power to look for unmodeled burst signals in glitchy data.
- Led by: LIGO stochastic analysis group
- Called: STAMP - Stochastic Transient Analysis Multi-detector Pipeline.
- Uses: analysis code and expertise from
  - stochastic analyses
  - unmodeled burst analyses
  - pulsar analyses

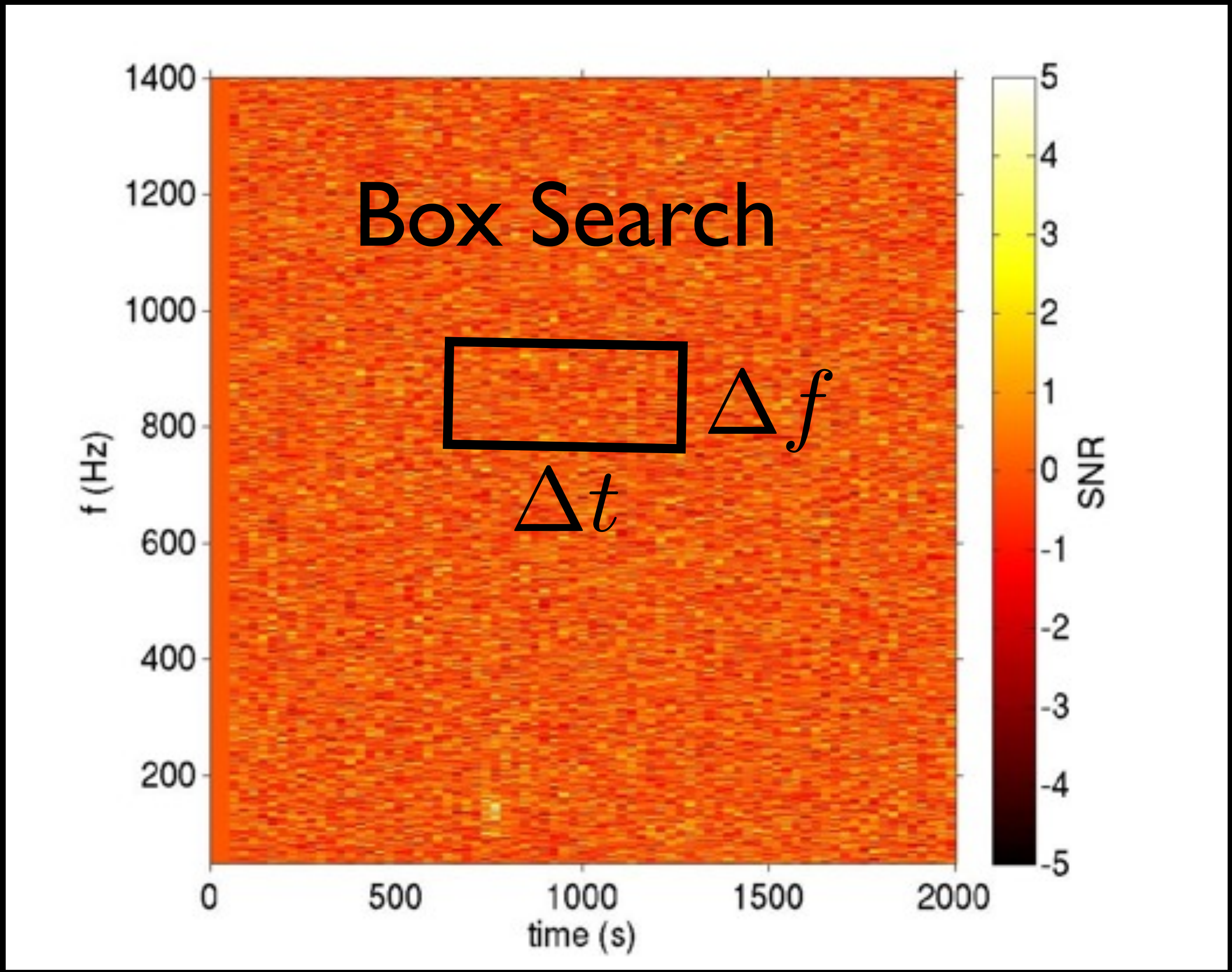
Migration: how it works

- Uses a TF representation of cross-power to project onto signal space.



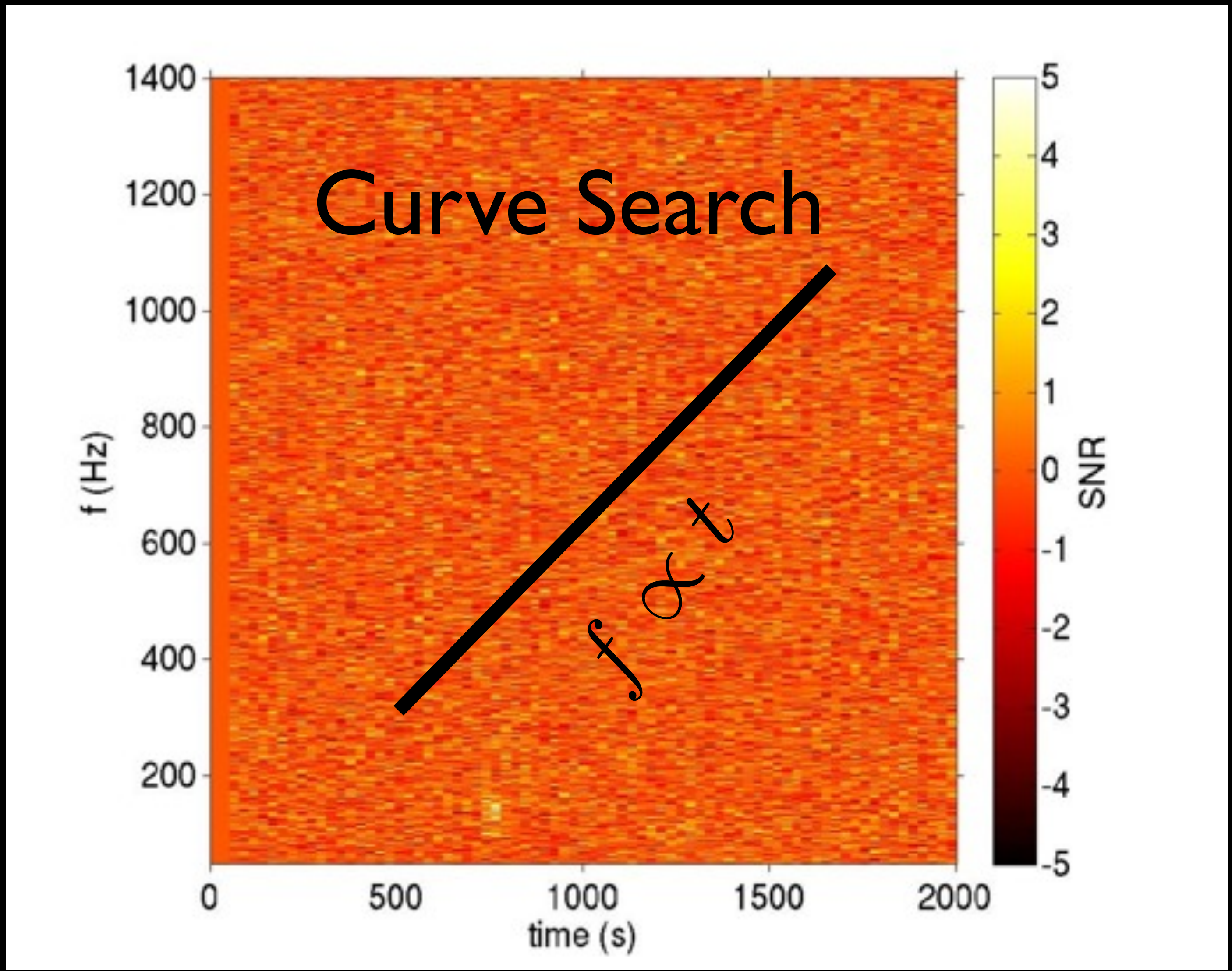
Migration: how it works

- Uses a TF representation of cross-power to project onto signal space.



Migration: how it works

- Uses a TF representation of cross-power to project onto signal space.



Migration: activity

- So far, the STAMP group have produced:
  - a methods paper (*PRD 83, 083004*)
  - a detector noise paper (*CQG 28, 235008*)
  - a long GRB upper limits paper (*PRD 88, 122004*)
- Group is currently working on:
  - an all-sky search
  - an neutron star r-mode search
- Contributors include:

  current - Marie Anne Bizouard, Samuel Franco, Patrice Hello, Nelson Christensen, Eric Thrane, Shivaraj Kandhasamy, Tanner Prestegard, Patrick Meyers, Jialun Luo, Michael Coughlin, Bernard Whiting, Antonis Mytidis.

  past - Christian Ott, Steven Dorsher, Stefanos Giampanis, Vuk Mandic, Peter Raffai, WGA

Conclusion

- All "four types" of LIGO-Virgo data analysis have a lot in common:
  - use the same data
  - based on likelihood
  - the only difference is signal space we project on.
- Expertise and tools are portable across many analyses.
- STAMP is proving to be an example of fruitful interactions between analysts from different camps.