Using Random Forest To Rank-Order Potential Gravitational-Wave Events

Kari Hodge[1]

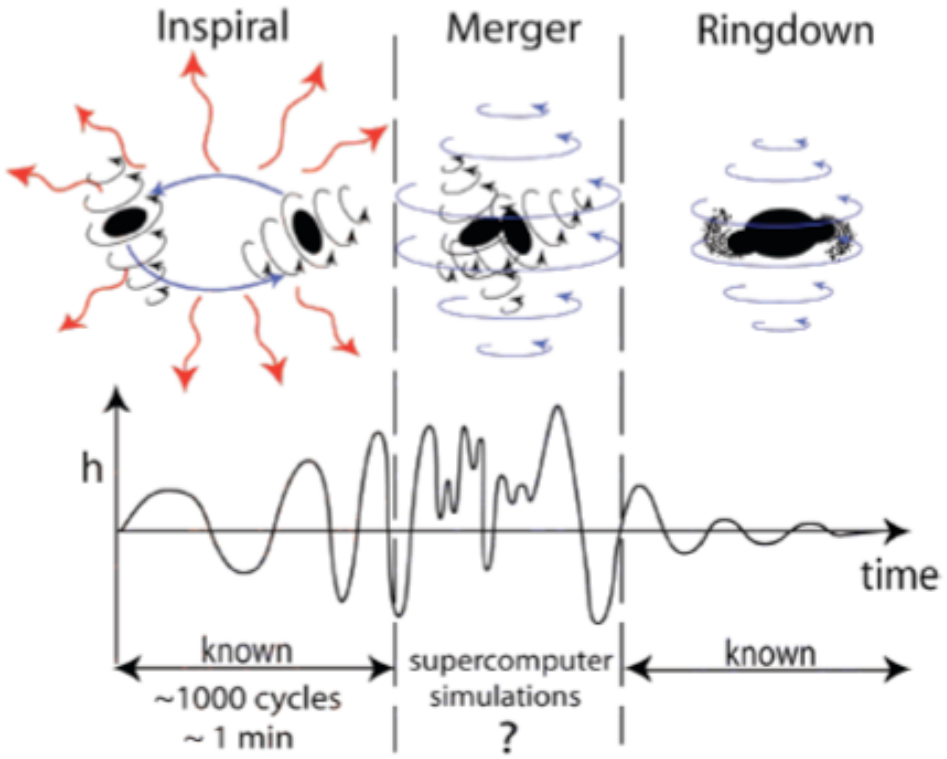
[1]California Institute of Technology, LIGO

Multivariate Analysis Workshop
11 February 2008

LIGO-G080030-00-Z

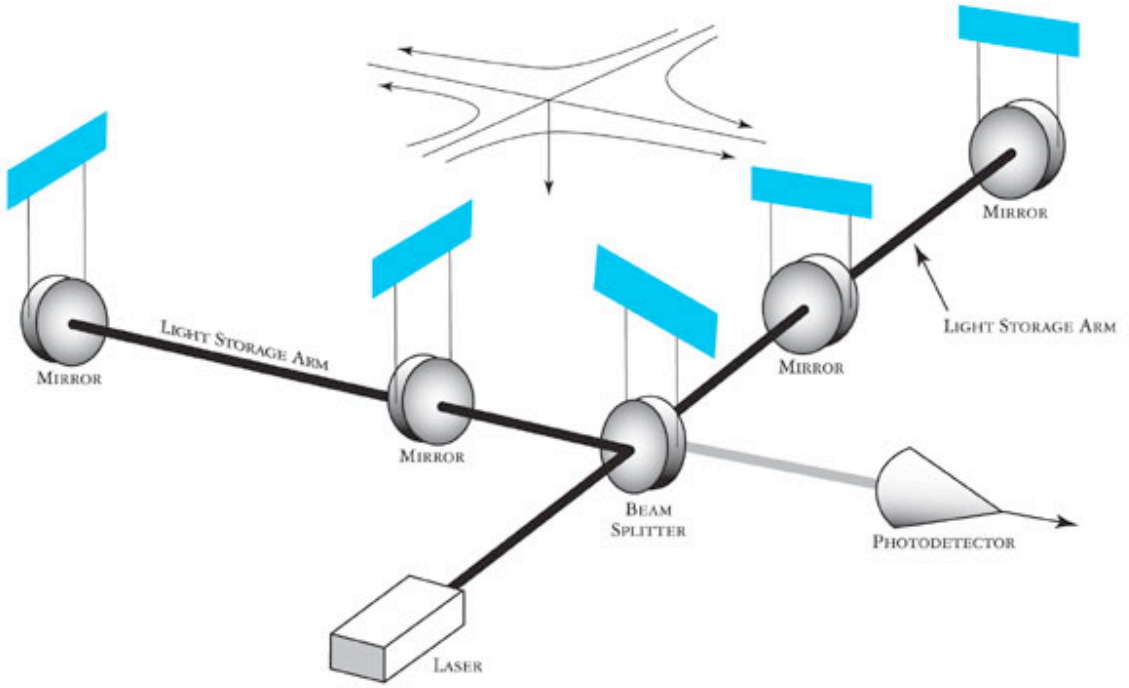
Compact Binary Coalescences

- Compact Binaries:
  - » Two neutron stars
  - » Two black holes
  - » A neutron star and a black hole
- The gravitational waveform emitted by the system during the inspiral phase of the coalescence has been modeled with General Relativity
  - » Second order Post-Newtonian templates



Detectors

- Large Michelson interferometers
  - » H1: 4 km
  - » H2: 2 km
  - » L1: 4 km

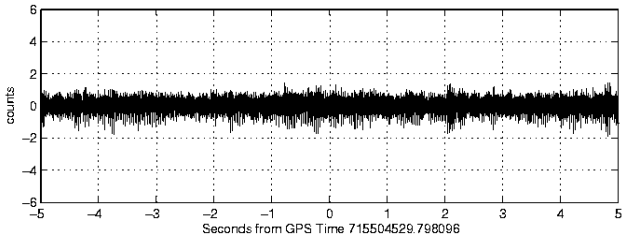


http://www.ligo.caltech.edu/LIGO_web/PR/scripts/facts.html

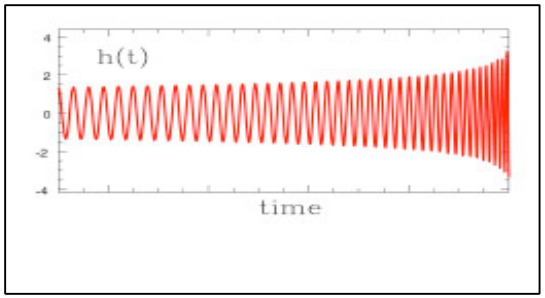
Inspiral Pipeline

- Data from the detectors is broken into segments of time, which are compared to a bank of waveform templates
  - » When a segment from a detector triggers a template, parameters that describe the event (mass, SNR, etc.) are produced
- Gravitational-wave candidate event: triggers from more than one detector are similar in time and mass
  - » The pipeline produces the values of the parameters for each detector it was seen in
- Follow up on the candidates
  - » None of the candidates has yet proven to be a gravitational-wave!

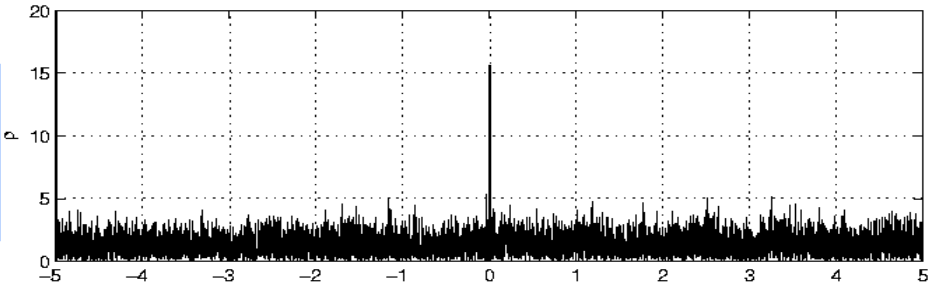
Matched Filtering



Filter to suppress high/low freq



SNR



Coalescence Time

Duncan Brown LIGO-G060580-00-Z

Signal Based Vetoes

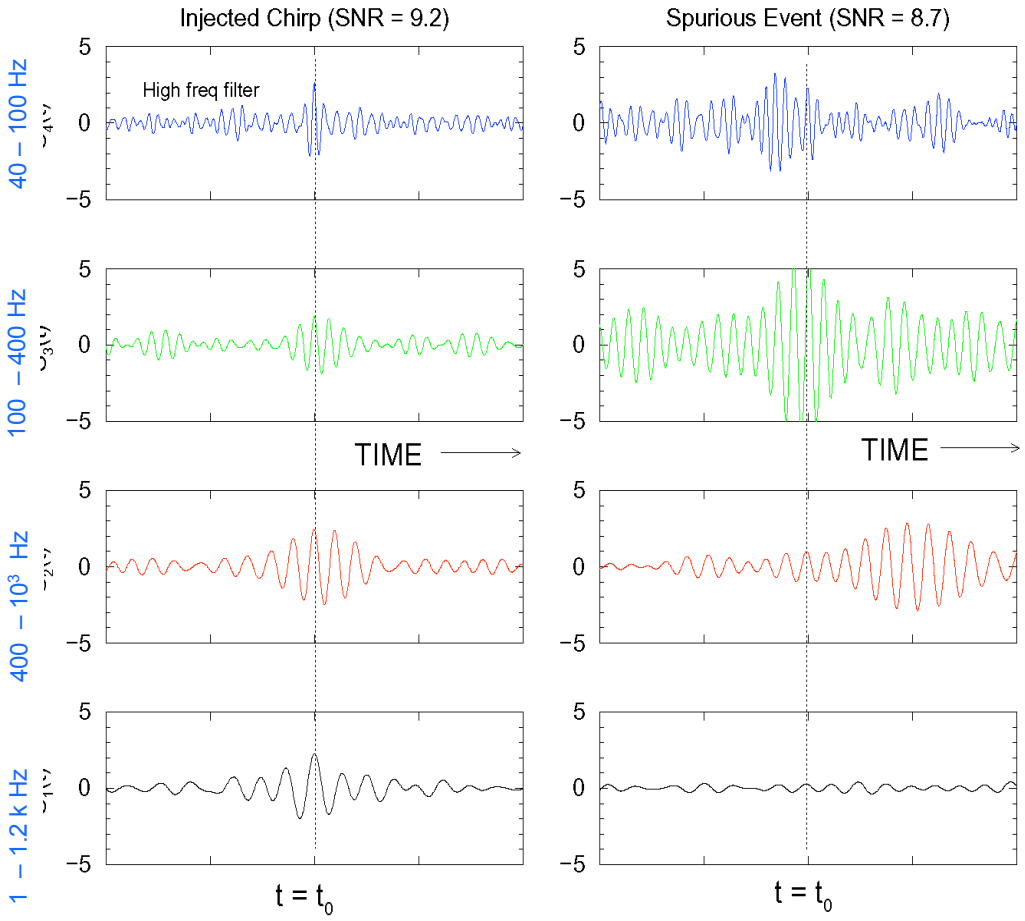
- Any large glitch in the data can caused the matched filter to have a large SNR output
- Signal based vetoes check that the matched filter output is consistent with a signal

$$\chi^2(t) = p\sum_{i=1}^{p} |\rho_i(t) - \rho(t)/p|^2$$

- Require that:

$$\frac{\chi^2}{p + \delta^2\rho^2} < \text{threshold}$$

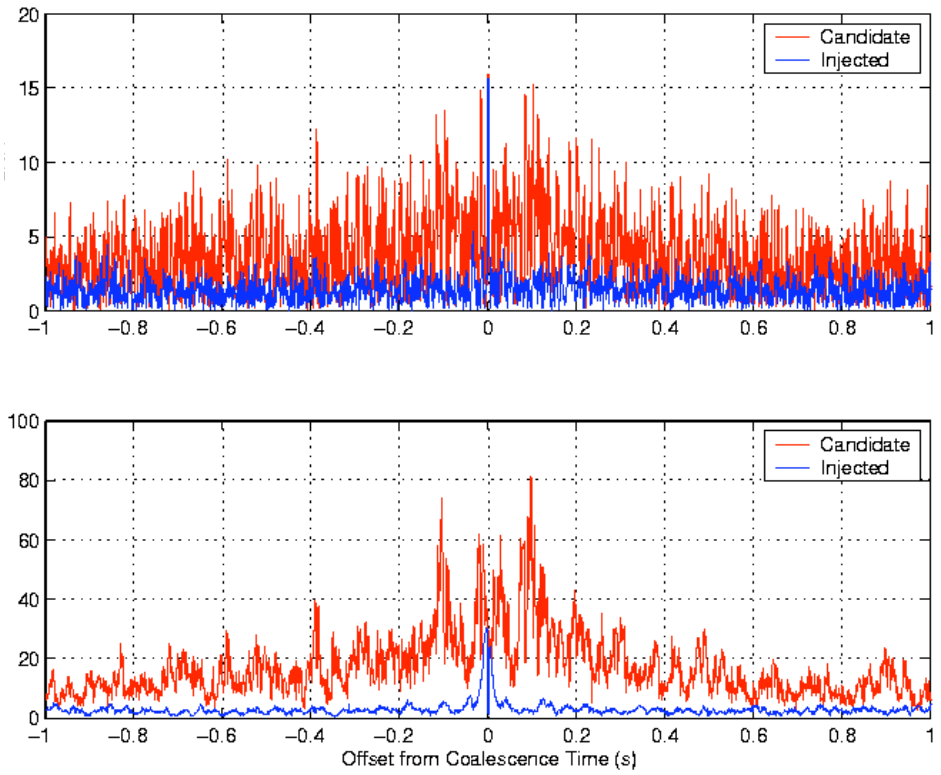
- r^2 veto duration is a measure of the time that chi^2 is above threshold



Duncan Brown LIGO-G060580-00-Z

Glitches in the Data

- Glitches can still be a problem, even with signal based vetoes (particularly in higher mass searches)
- A lot of work in the LSC is devoted to finding, identifying and eliminating glitches
- Loud glitches reduce our range (and hence rate) by hiding signals
  - » Reduces the volume of the sky we can see by (reduction in rate)^3
- Even if a template has excellent overlap with signals, if it picks up lots of glitches we have a problem



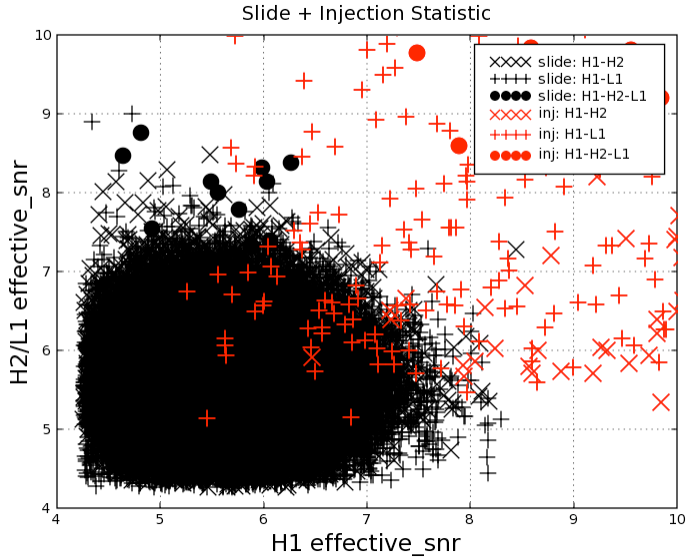
Duncan Brown LIGO-G060580-00-Z

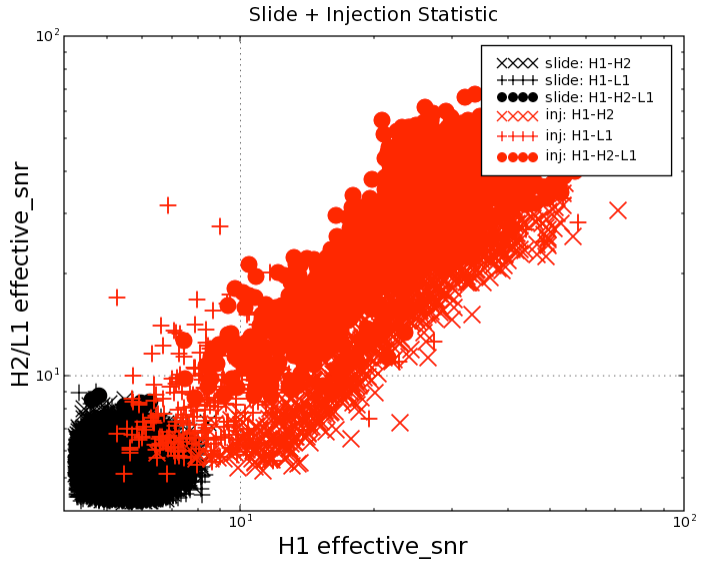
Follow-up Candidate Events

- Follow-ups are necessary because of the glitches that ring up triggers and pass the signal based vetoes
- Currently, the candidates are ranked according to the sum of squares of effective SNR for each detector

$$\rho^2_{eff} = \frac{\rho^2}{\sqrt{\left(\frac{\chi^2}{2p-2}\right)\left(1+\frac{\rho^2}{250}\right)}}$$

- The top-ranking candidates on this list are subjected to rigorous examination





Analysis of Random Forest

- Random forest technology can take into account the correlations between the many parameters that describe a candidate event and create a more robust rank-ordering statistic

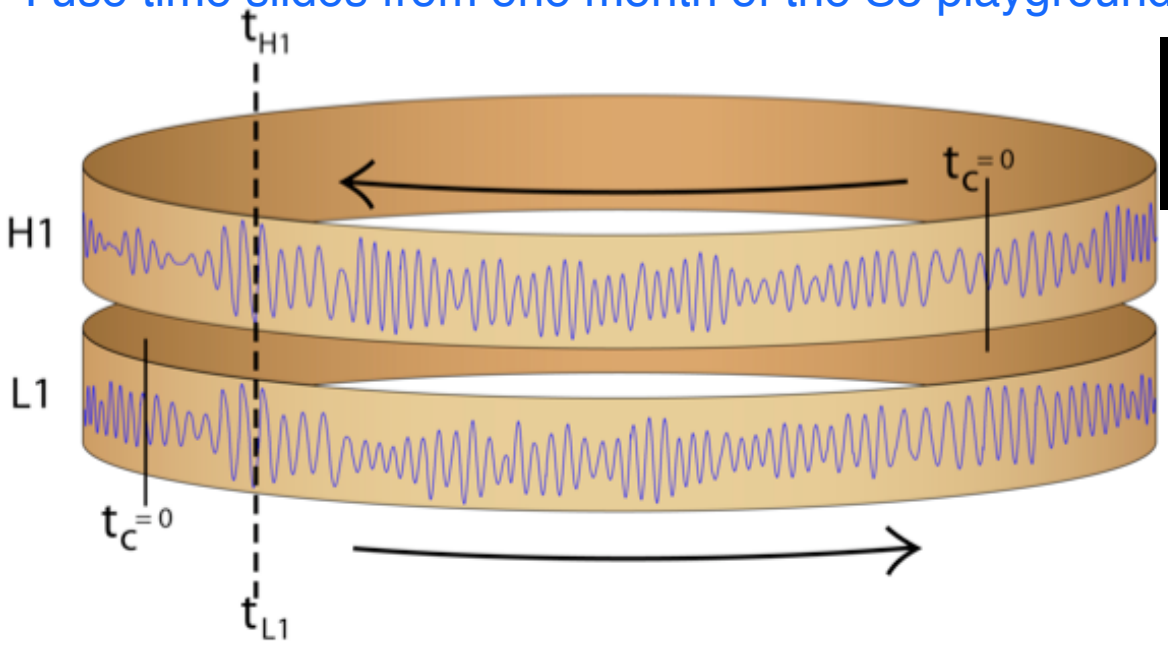
- I use simulated gravitational waves "injected" into the data as the signal for my analysis
  - » From the LSC's 1st year S5 Low-Mass Compact Binary Coalescence Analysis*
  - » I am only using H1-L1 coincidences for the moment since they are harder to classify as signal than triply coincident events

I have 9,569 injections

see: Drew Keppel's GWDAW talk LIGO-G070820-00-Z

Background

- Noise in the detectors:
  - » Seismic motion, thermal disturbances, quantum fluctuations (shot noise)
- Time slides estimate the background:
  - » The data streams from two detectors are slid integer multiples of 5 seconds from each other and run through the Inspiral Pipeline
  - » These accidental coincidences can't be gravitational waves
  - » I use time slides from one month of the S5 playground



I have a total of 267,689 time slides

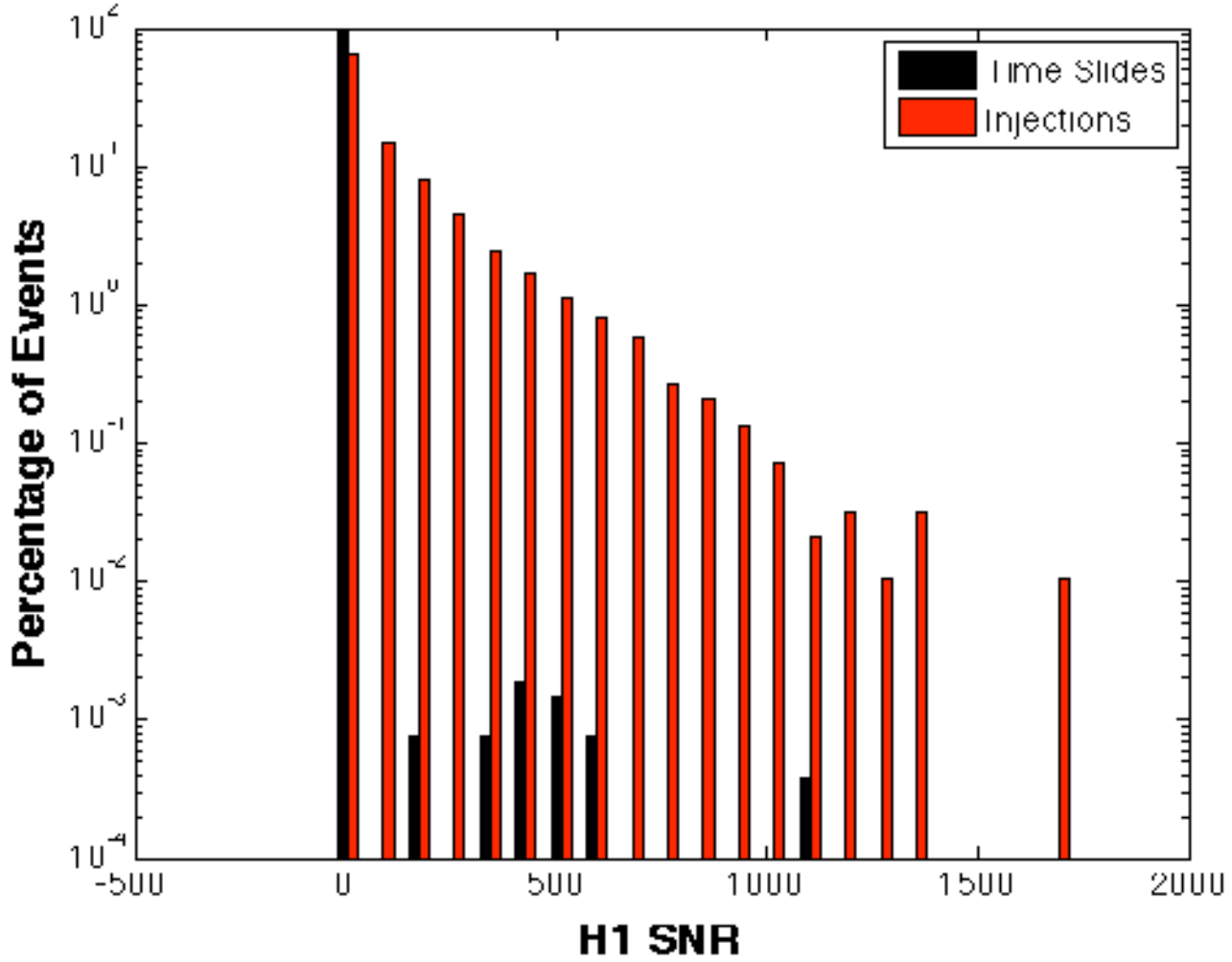
Input Parameters

- Single detector parameters
  - SNR
  - chi^2
  - r^2 veto duration
  - SNR$_{eff}^2$

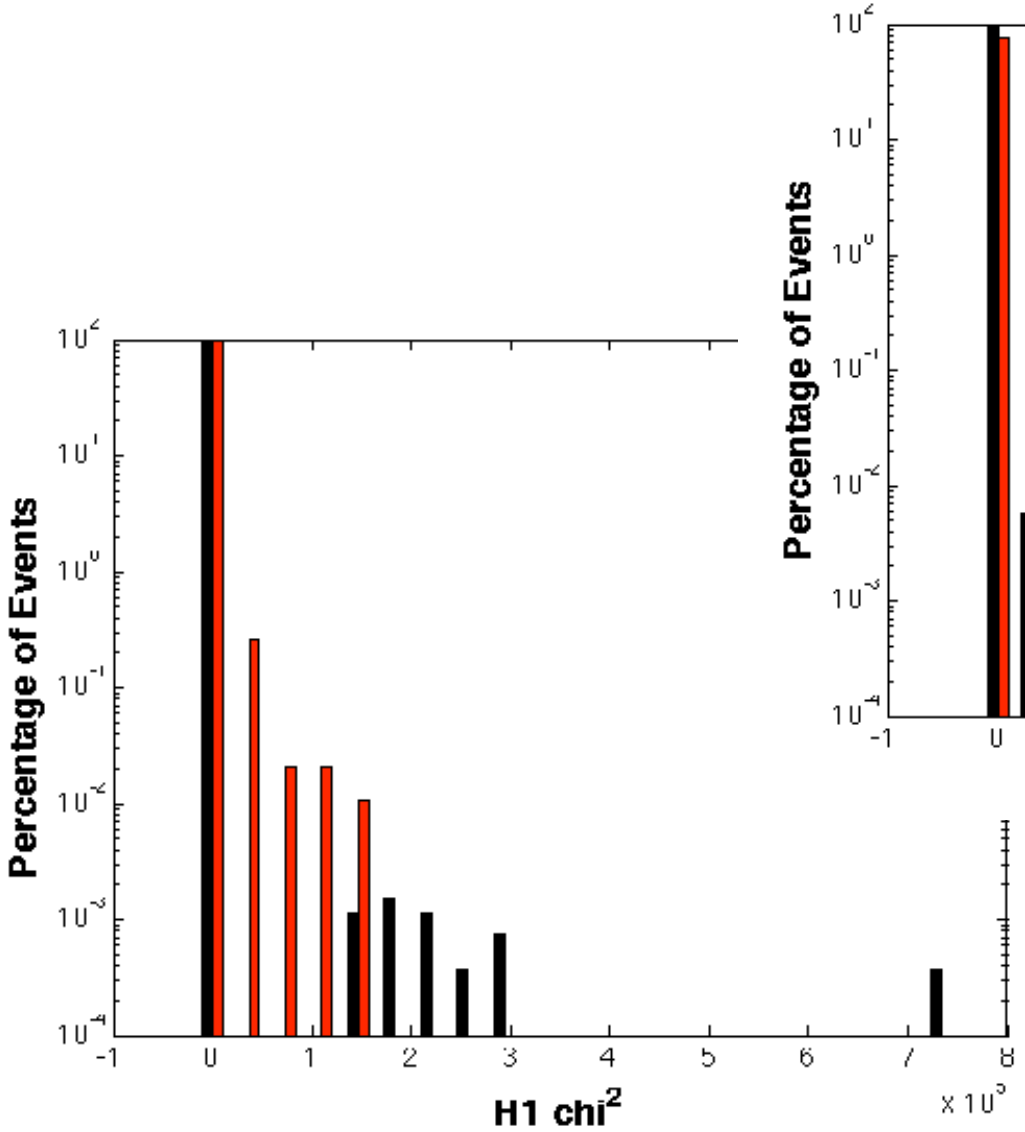
$$\rho_{eff}^2 = \frac{\rho^2}{\sqrt{\left(\frac{\chi^2}{2p-2}\right)\left(1+\frac{\rho^2}{250}\right)}}$$

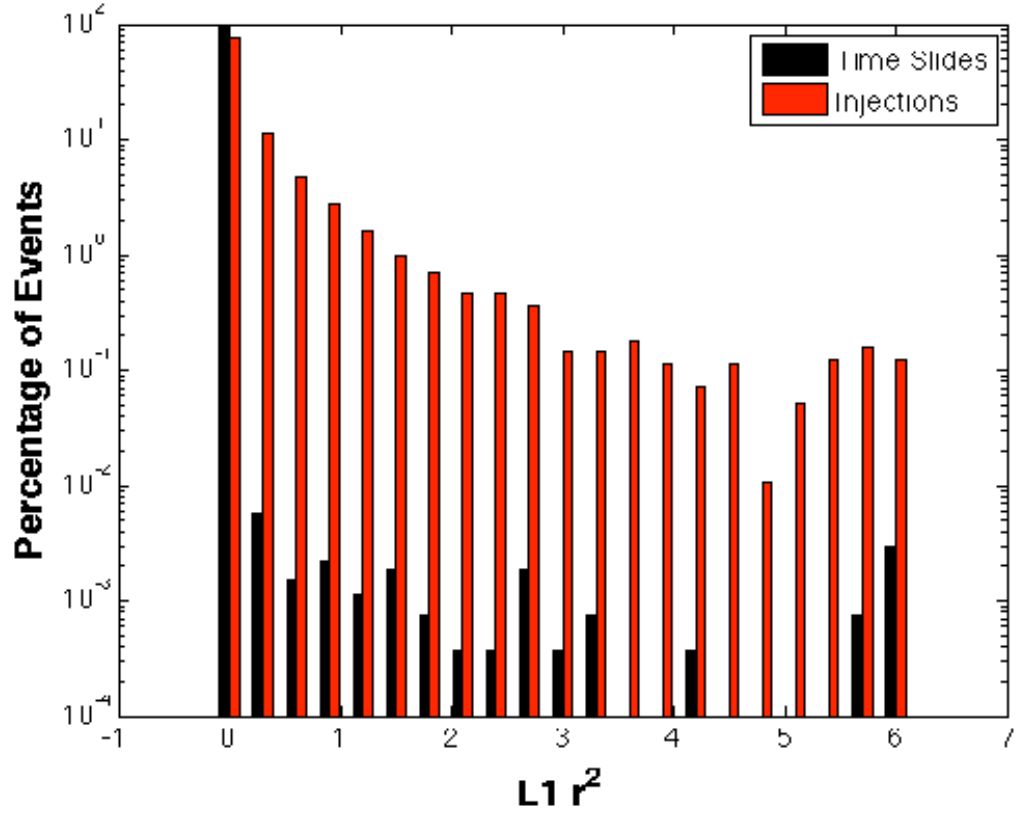
- Coincidence parameters
  - Difference in coalescence time
  - Relative difference in chirp mass

Single Detector Parameters

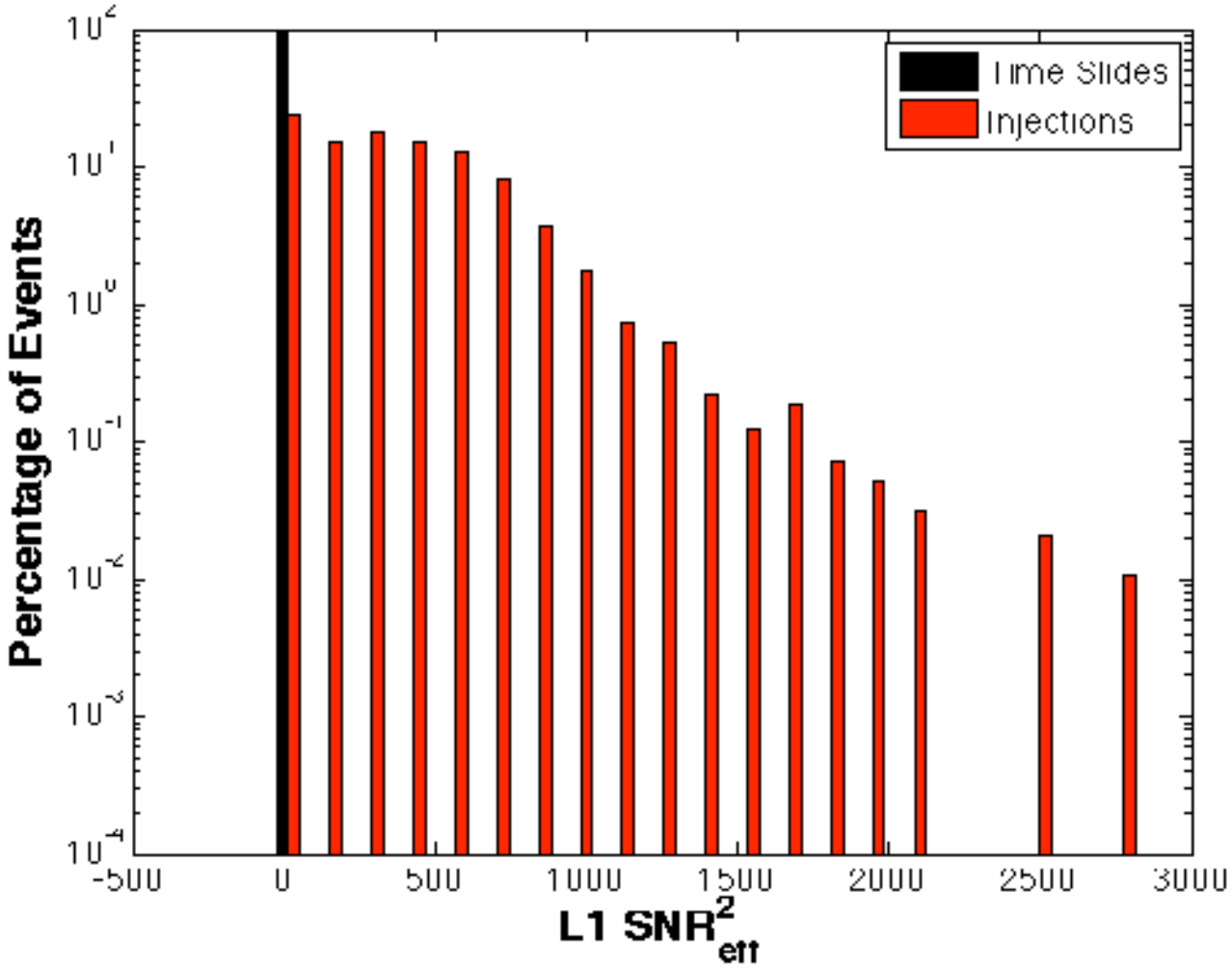


Single Detector Parameters

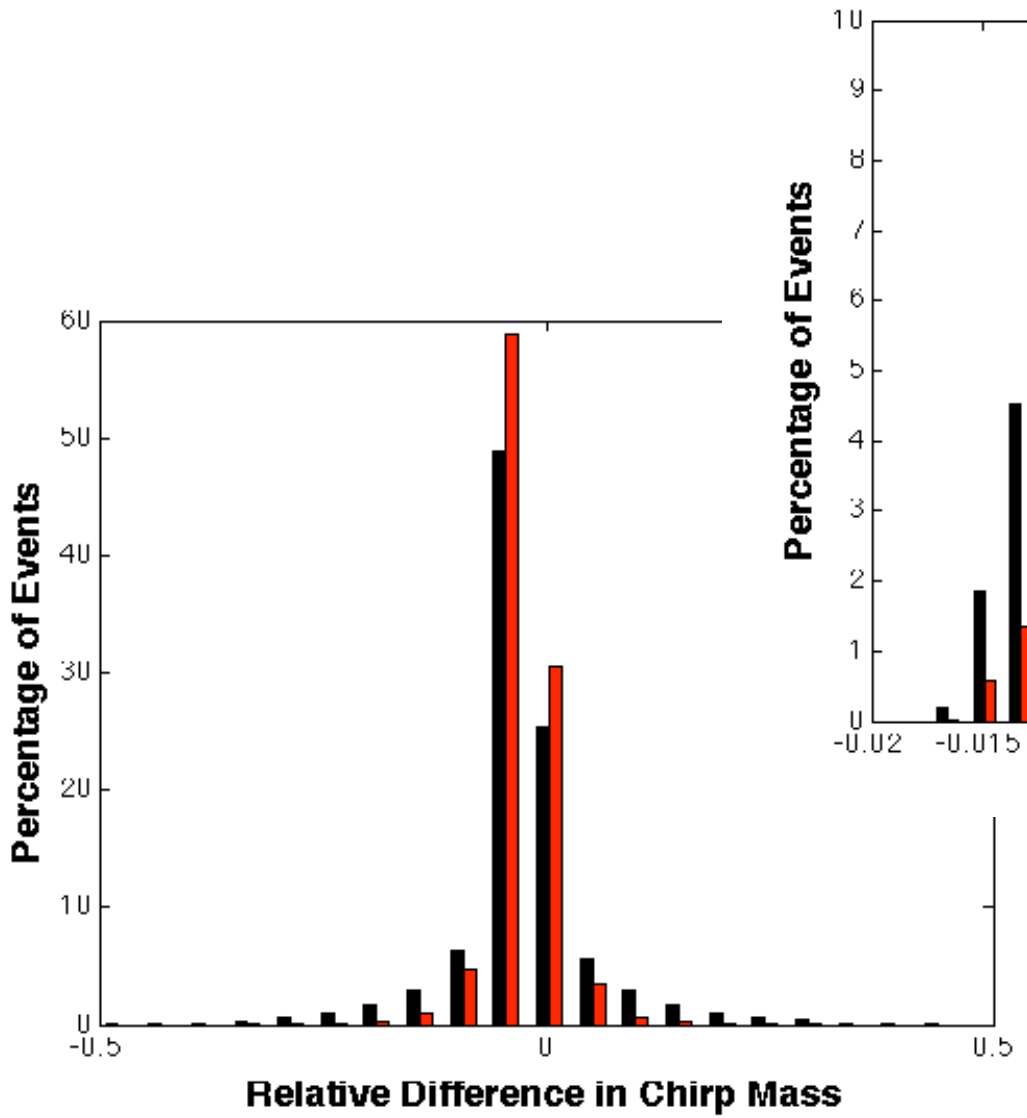


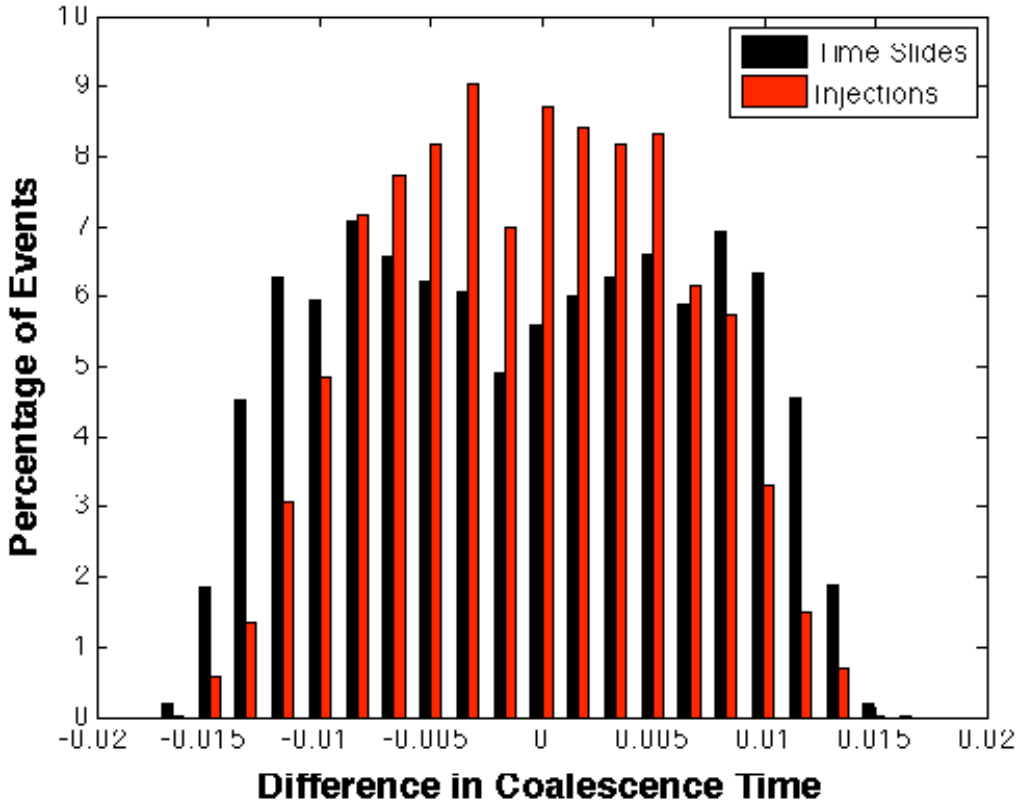


Single Detector Parameters

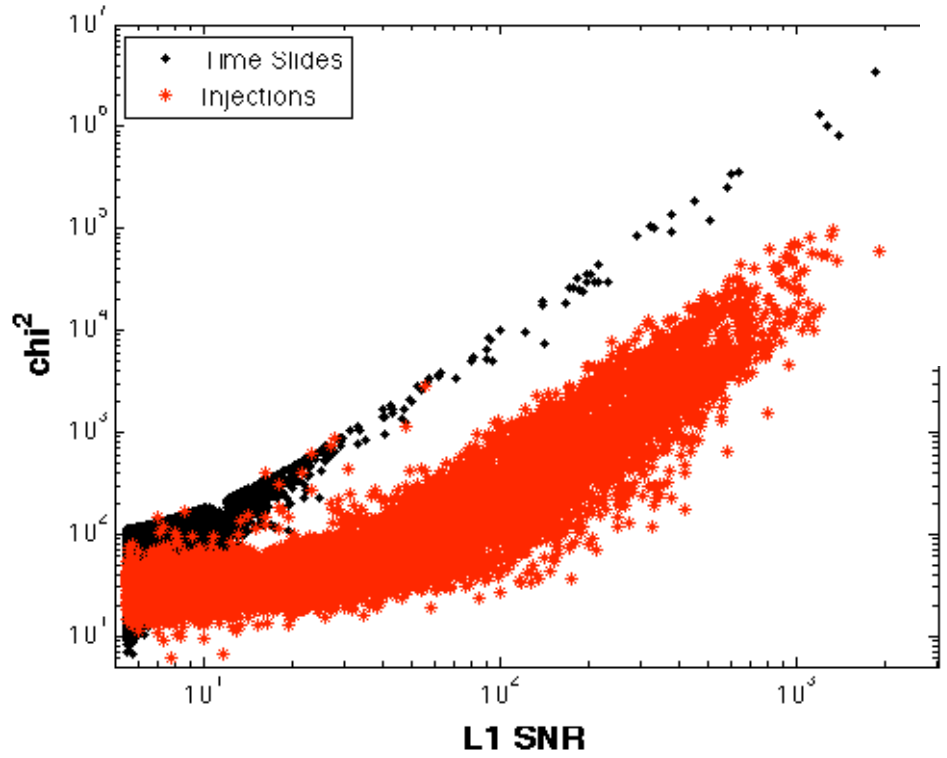


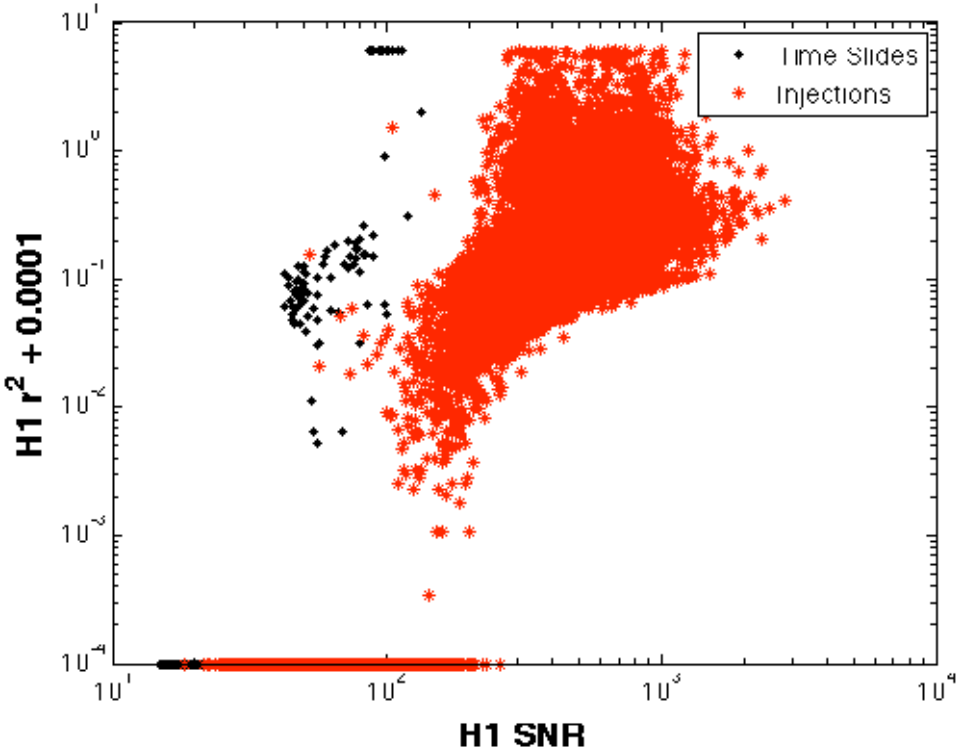
Coincidence Parameters



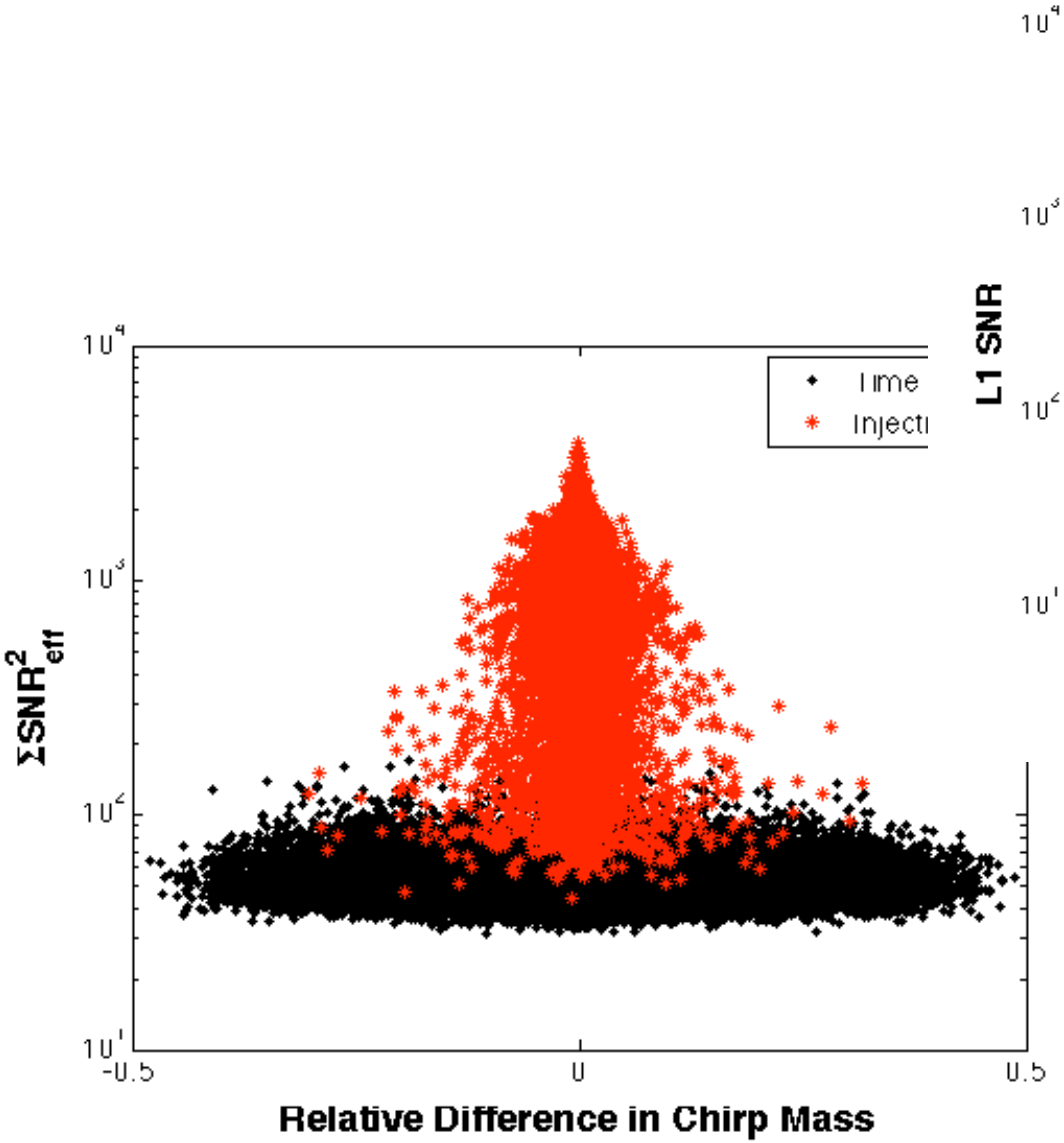


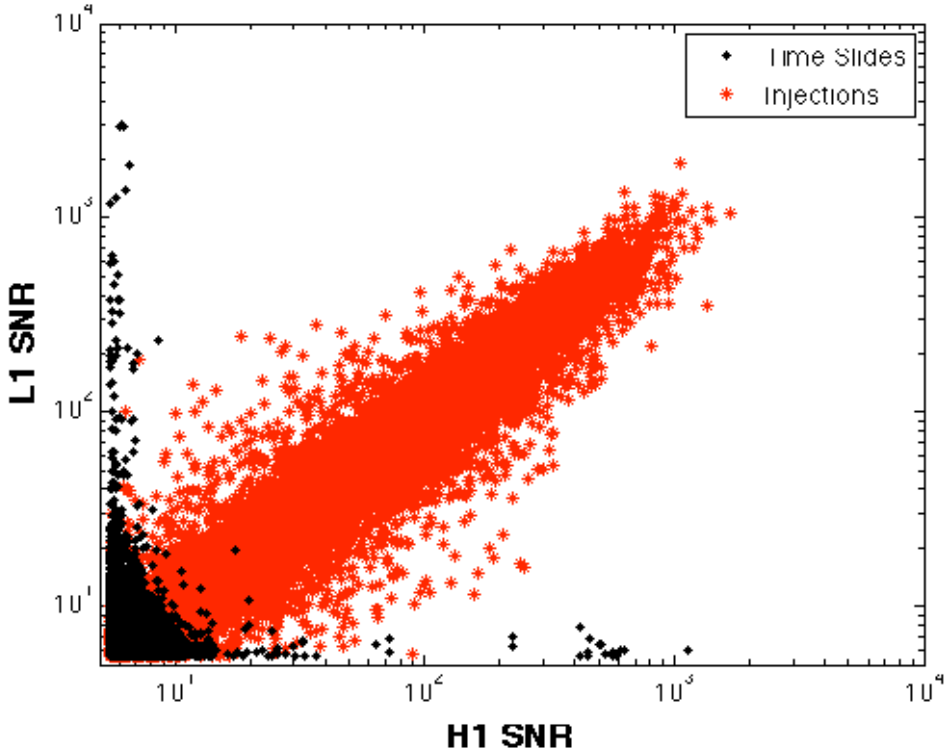
Two Dimensions





Two Dimensions





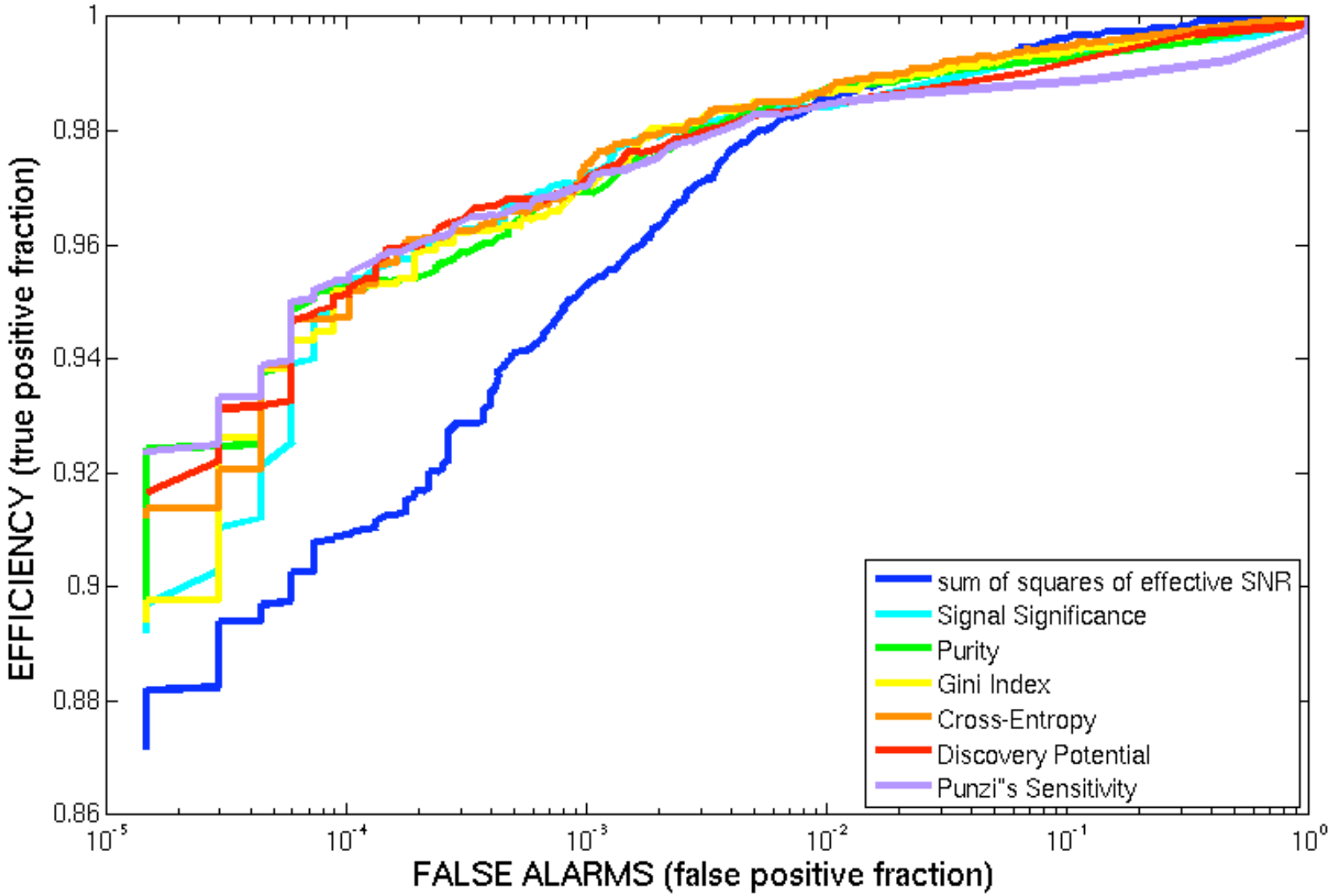
SprBaggerDecisionTree

- I use SprBaggerDecisionTree to train a random forest of bagged trees
  - » This algorithm creates many decision trees, each a bootstrap replica of the training sample. If you start with N training events, then each tree will also have N events, but these events are chosen with replacement
- The random forest technology will sample up to 4 out of 10 of the variables for each split on the tree
- I build 100 trees
  - » Specify each has a minimum of 5 events per leaf
- ~300,000 time slides and ~10,000 injections
  - » 1/2 for training
  - » 1/4 for validation
  - » 1/4 for testing

Criteria for Optimization

- The goal of each tree is to optimize a certain criterion
- SprBaggerDecisionTree gives the option of 9 different criteria
  - » The results for the best of these, as compared to the sum of squares of effective SNR, are summarized on the next plot

Criterion Performance

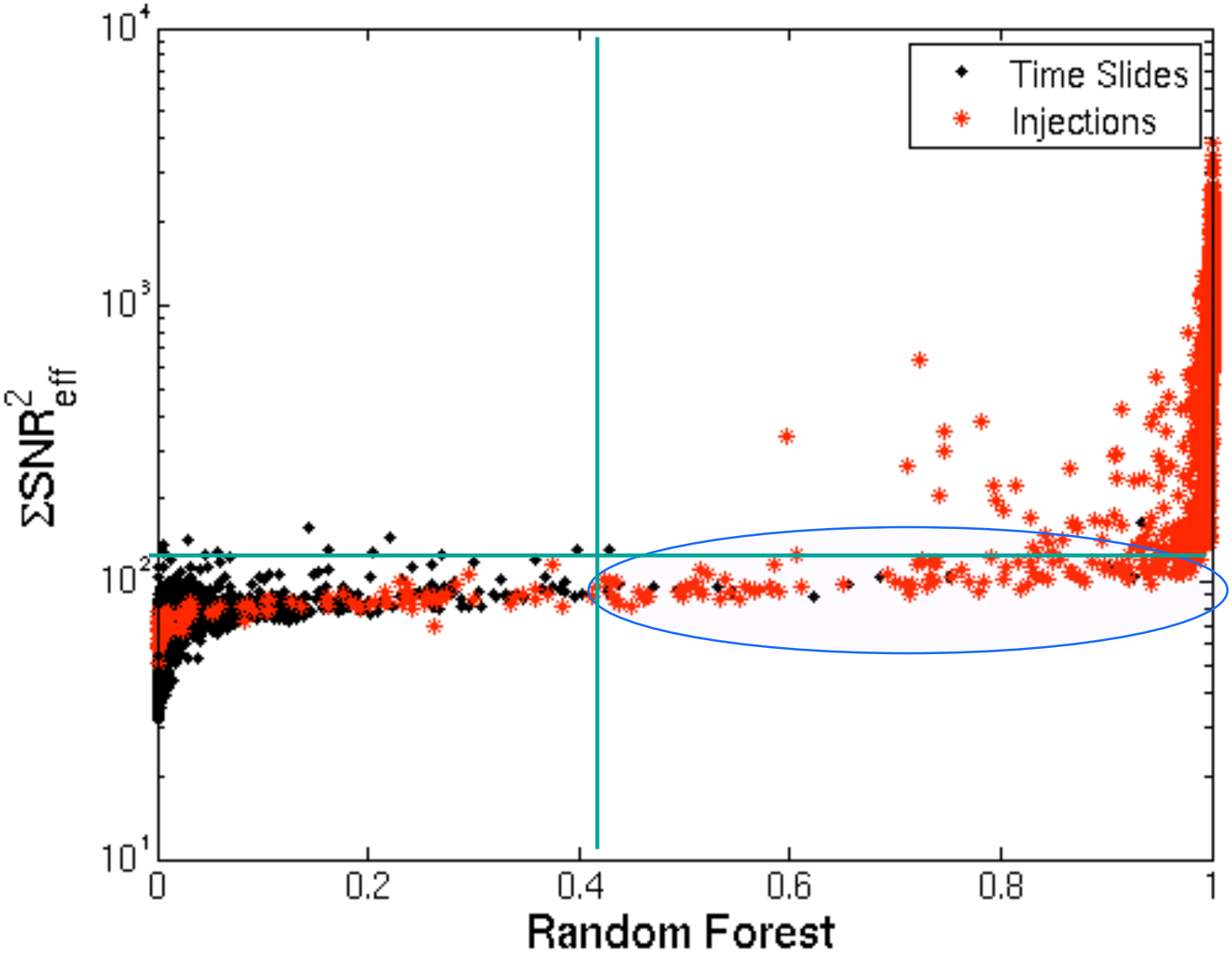


Cross-Entropy

- If you want to live in the region where your false alarm fraction is between 1/1000 and 1/100, then Cross-Entropy gives the best results

Variable	Splits	Delta FOM
dt	1680	193.15895
H1 SNR	2287	255.79869
L1 SNR	2185	254.86952
H1 chi^2	2159	206.18350
L1 chi^2	2499	289.57461
H1 r^2	41	5.58936
L1 r^2	51	8.59637
$(dM)_{rell}$	2360	216.16415
H1 SNR_{eff}^2	2625	264.50807
L1 SNR_{eff}^2	2594	270.51128

Improvement in Region of Weak Signal



Conclusion

- The random forest separates injected signals from accidental coincidences more effectively than the current ranking statistic
- More optimization of the leaf size, number of sampled parameters, etc. will lead to improved results