Multidimensional classification analysis of kleine Welle triggers in LIGO S5 run

Soma Mukherjee for the LSC
University of Texas at Brownsville

GWDAW12, Boston, December 13-16, 2007

LIGO DCC # G070814-00-Z

A quick walk through …

•This study involves implementation of classification methods (non-parametric/hierarchical and parametric, s-means) to see presence of structures in higher dimensional parameter space. Often features embedded in higher dimensions are not elucidated in simple 1 or 2 dimensional study.

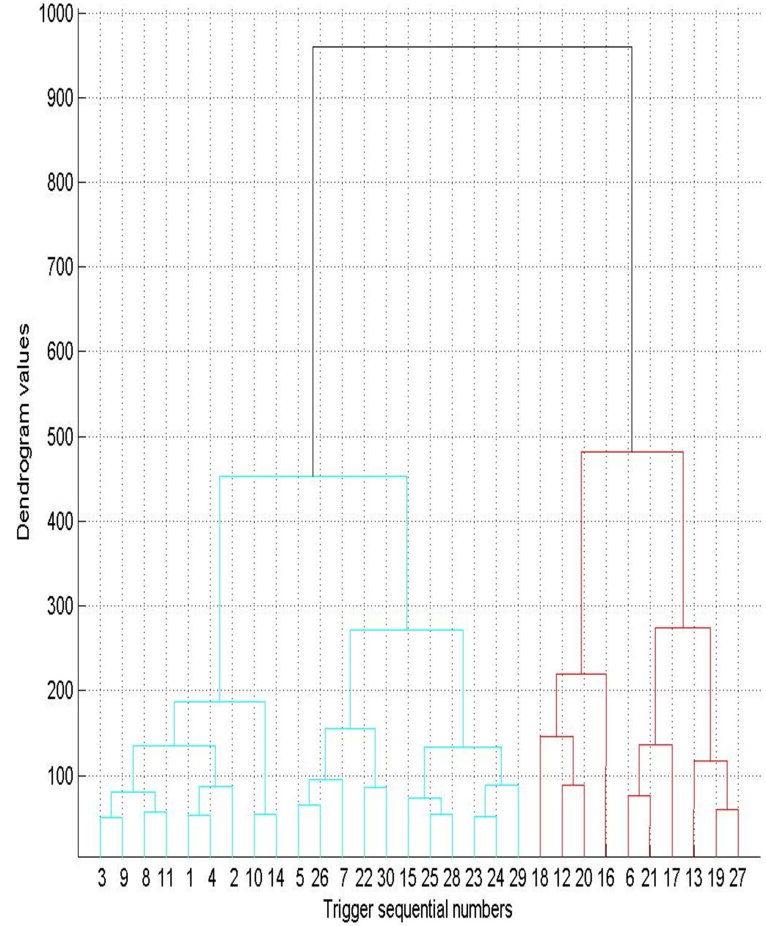
•References for LIGO classification analysis: S. Mukherjee : Past LSC, F2F and GWDAW talks, Burst and Glitch group telecons, published paper in CQG).

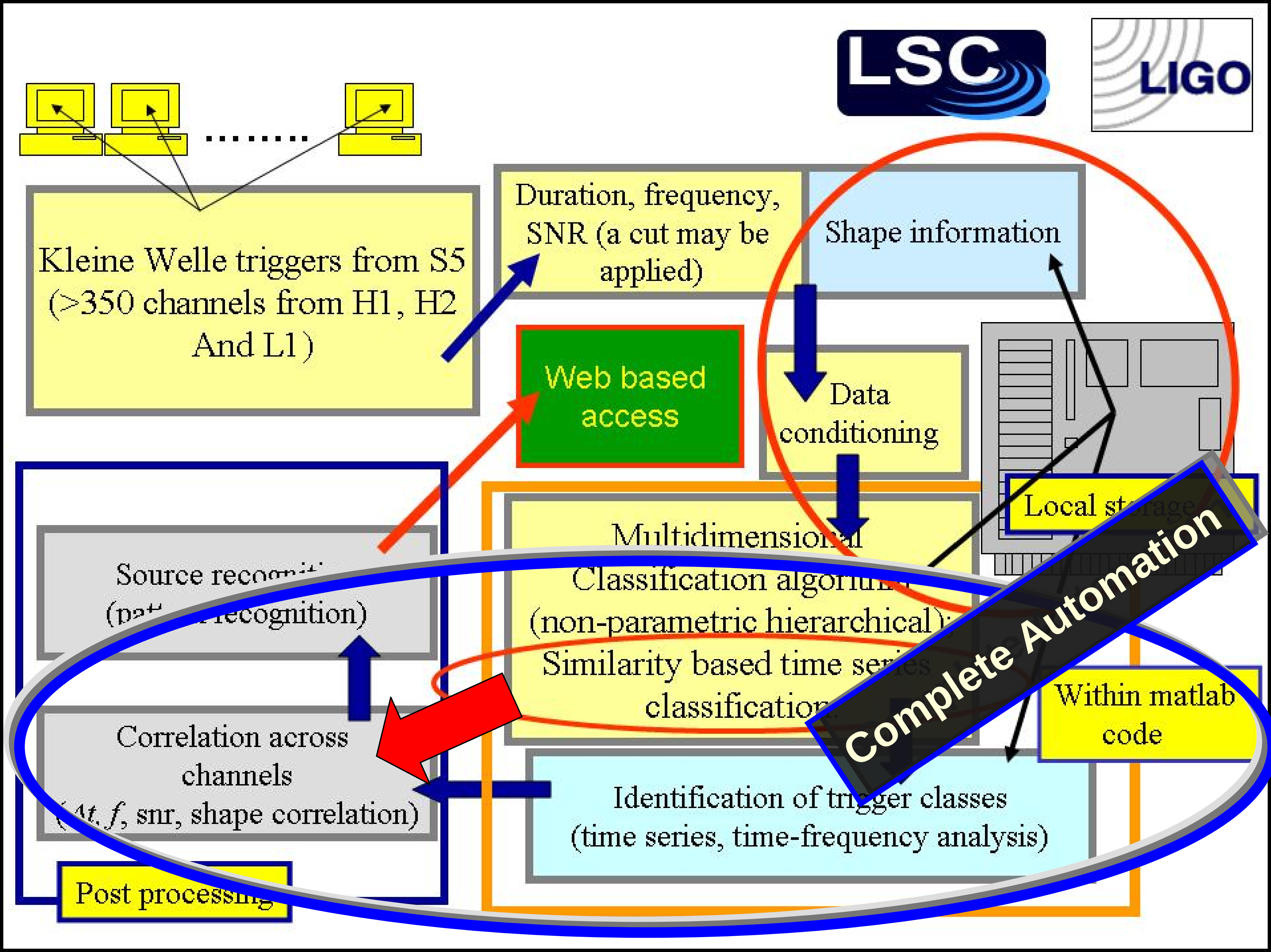
•kleine Welle *(Blackburn, L. et. al. 2005 LIGO-050158-00-Z)* is an algorithm that picks up burst triggers from the gravitational wave, auxiliary and environmental channels in LIGO. It generates several gigabytes of trigger database containing information about the physical properties of the burst triggers. The purpose of this analysis is to mine the trigger database to see if the triggers can be categorized in different groups that share common properties. This will lead to effective dimensionality reduction of the problem since the number expected groups will be a countable small number and each group, to some extent, uniform in character. The physical motivation here is that this could become a powerful veto mechanism.

Algorithm : Hierarchical classification

The algorithm is based on computation of distances between data points in the multi-dimensional space. A variance minimization criterion is used to group the objects into statistically distinct classes. The metric may be chosen in several different ways. The group formation stage is guided by *complete-linkage* criterion, i.e. largest distance between objects in two groups.

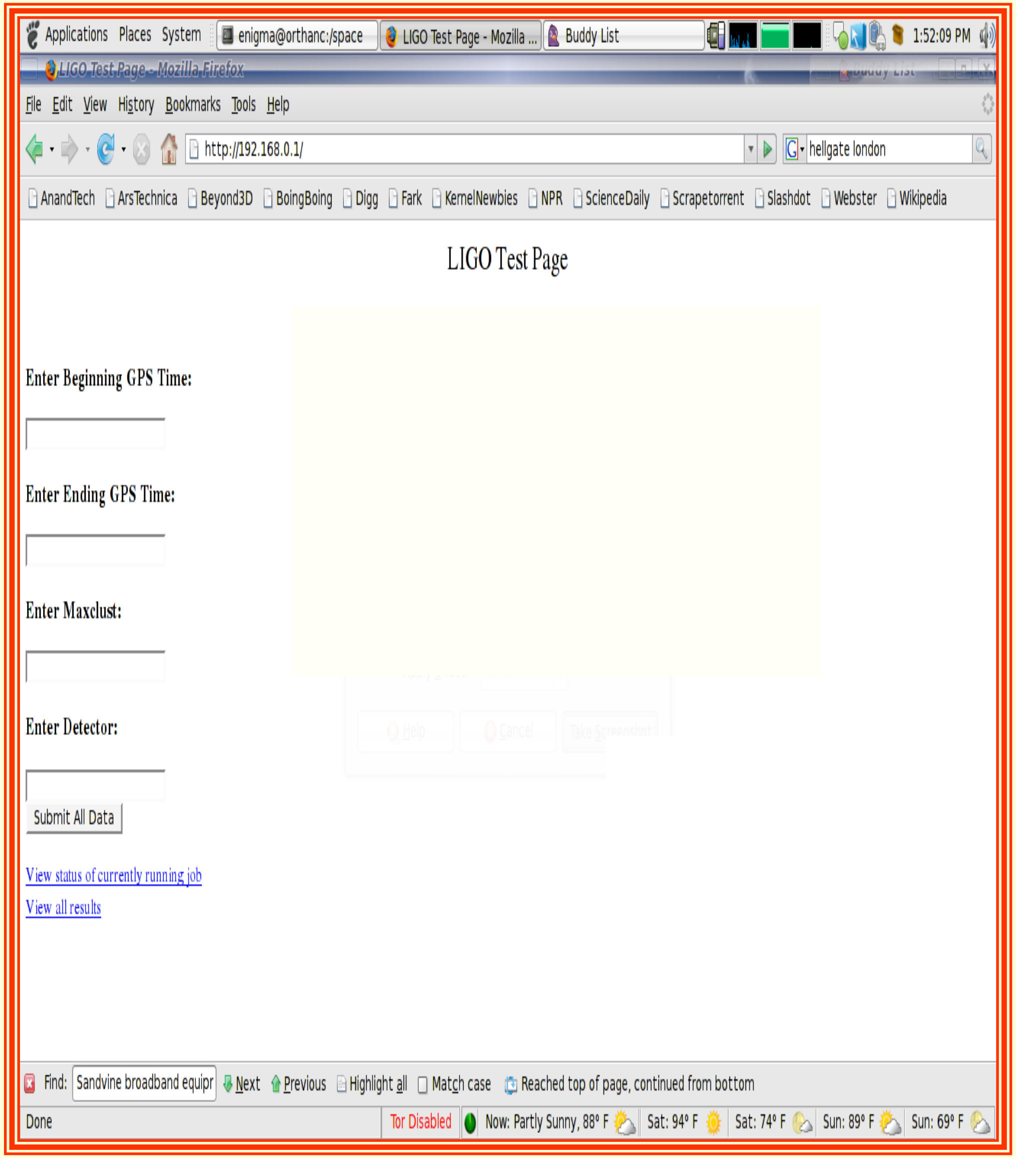
Example




LSC
LIGO
Kleine Welle triggers from S5 (>350 channels from H1, H2 And L1)
Duration, frequency, SNR (a cut may be applied)
Shape information
Web based access
Data conditioning
Local storage
Multidimensional Classification algorithm (non-parametric hierarchical); Similarity based time series classification
Within matlab code
Identification of trigger classes (time series, time-frequency analysis)
Source recognition (pattern recognition)
Correlation across channels (Δt, f, snr, shape correlation)
Post processing
Complete Automation

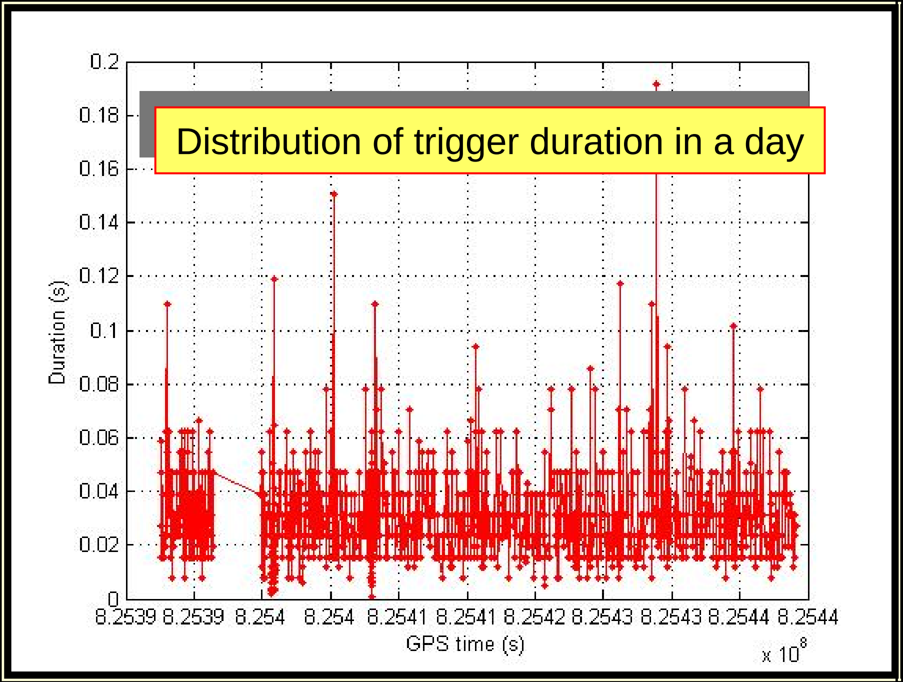
Pipeline Automation

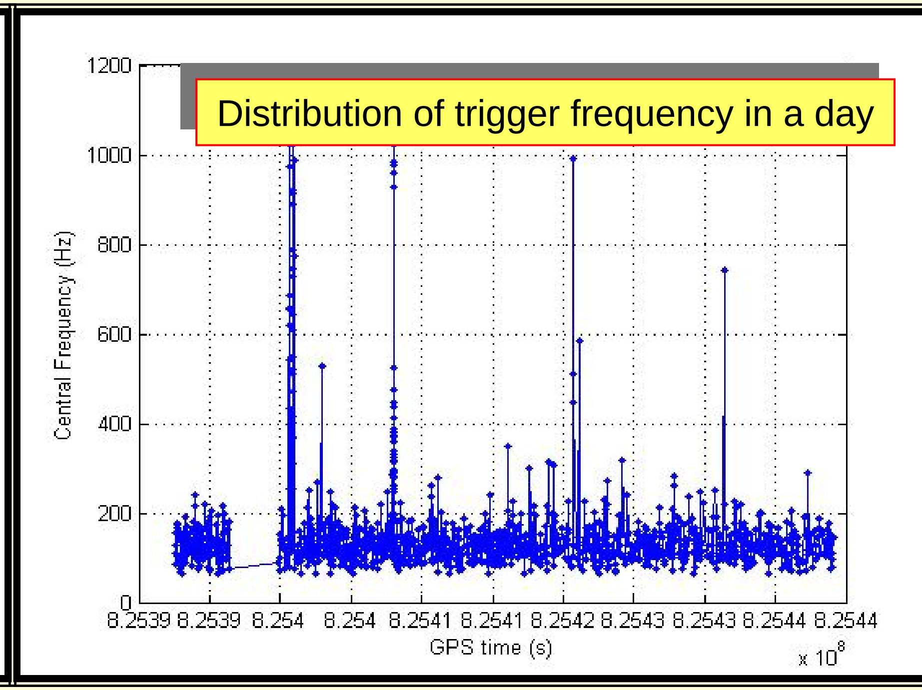
- Interactive interface : input GPS times and detector.
- Optimization : 1 year KW database analyzed in ~2.5 days with ~20 nodes in the local cluster Funes.
- Achieved complete automation in analysis, archiving and web-based dissemination of results.

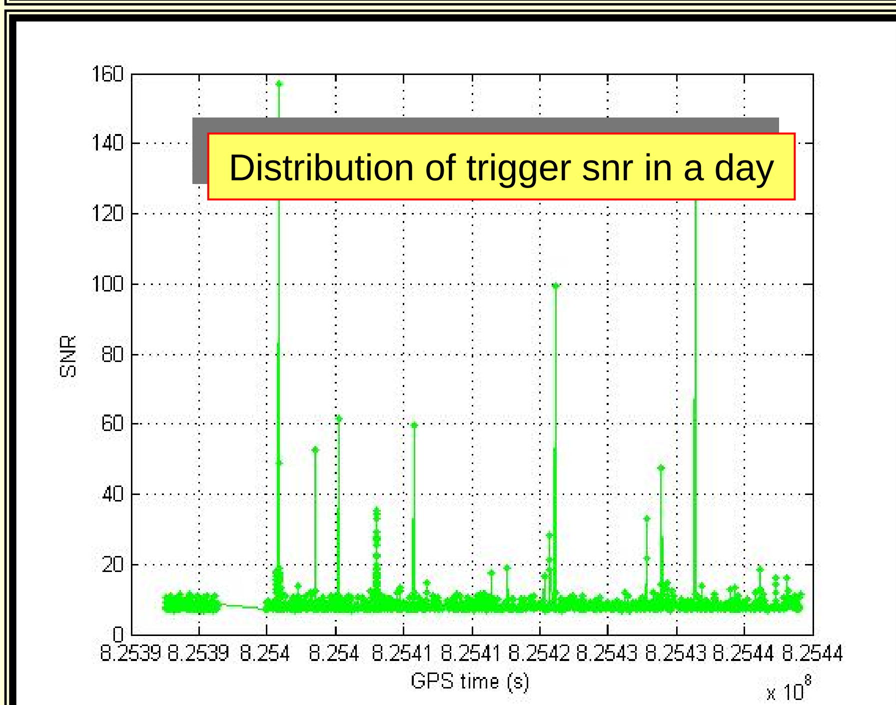


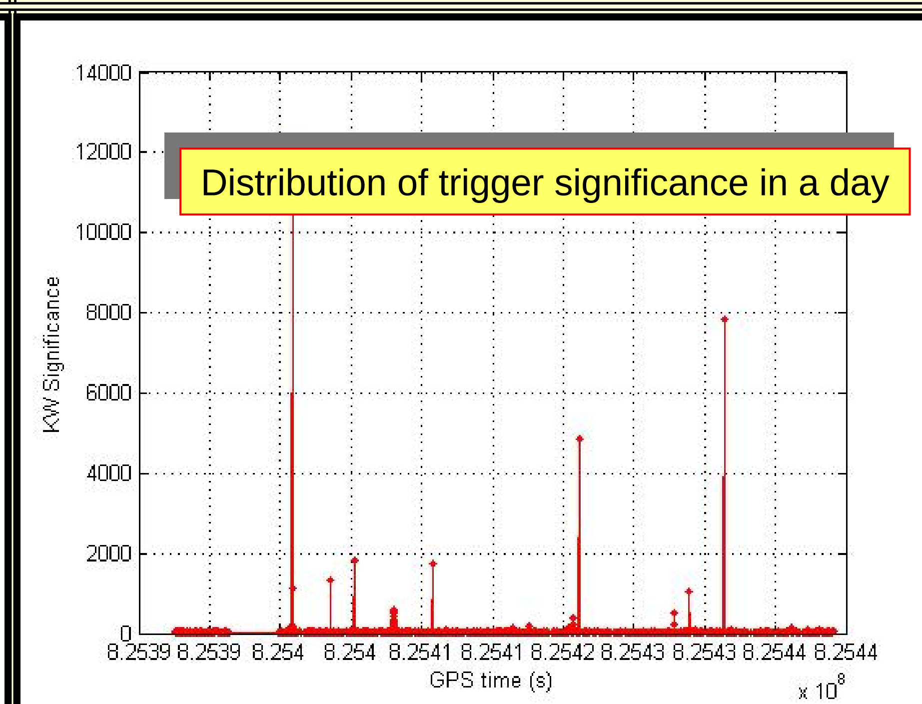
Post-analysis & Information Extraction I

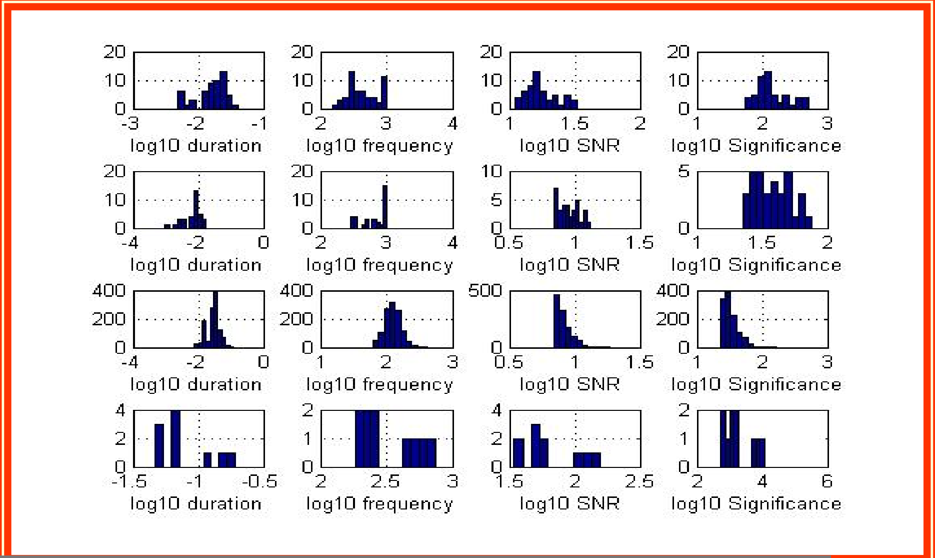
- Trigger Visualization : Extended Analysis
  - A look at the time series and time-frequency plots for each trigger in the different classes found (snr = 6 and above at the moment).
    - Data conditioning
      - Whitening and cleaning
      - Filtering over 256 Hz around central frequency
      - Decimation
    - Plots of conditioned time series and spectrograms
- Across the channel correlation and correspondence : Targeted Analysis
  - Select a window around the time of interest
  - Construct a database of all channels in that window
  - Look at classes across channels
  - Run *S-means** to see how the waveforms correspond
  - Infer and catalog correlated channels.


Distribution of trigger duration in a day
Duration (s)
0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16 0.18 0.2
8.2539 8.2539 8.254 8.254 8.2541 8.2541 8.2542 8.2543 8.2543 8.2544 8.2544
GPS time (s)
x 10^8

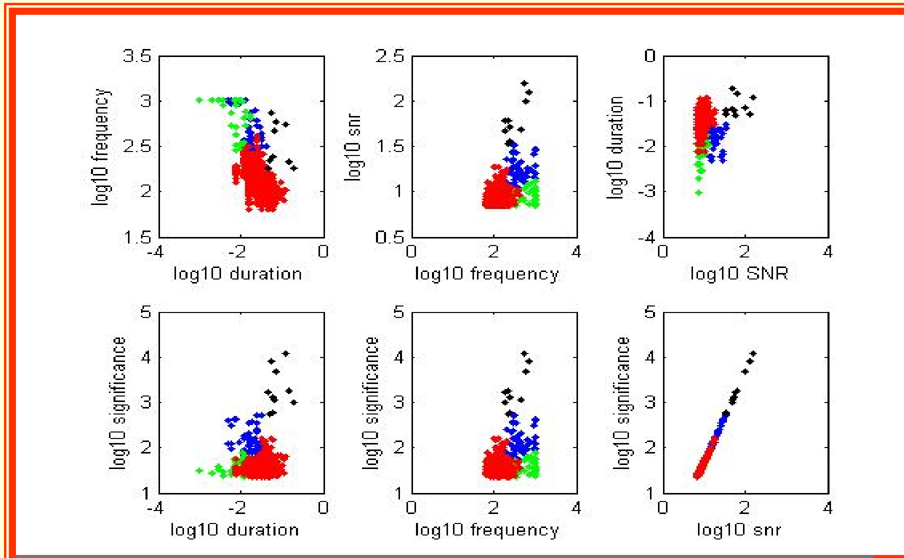

Distribution of trigger frequency in a day
Central Frequency (Hz)
0 200 400 600 800 1000 1200
8.2539 8.2539 8.254 8.254 8.2541 8.2541 8.2542 8.2543 8.2543 8.2544 8.2544
GPS time (s)
x 10^8


Distribution of trigger snr in a day
SNR
0 20 40 60 80 100 120 140 160
8.2539 8.2539 8.254 8.254 8.2541 8.2541 8.2542 8.2543 8.2543 8.2544 8.2544
GPS time (s)
x 10^8

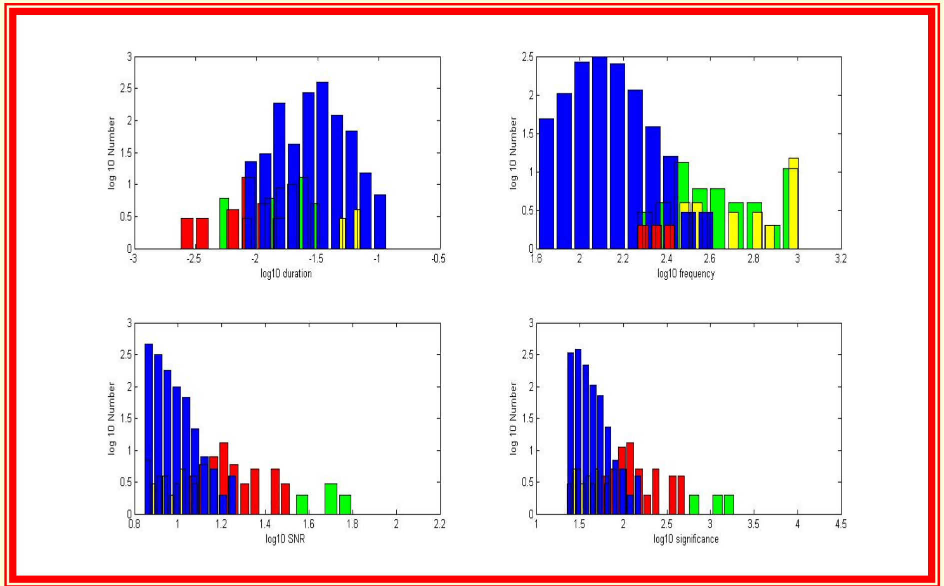

Distribution of trigger significance in a day
KW Significance
0 2000 4000 6000 8000 10000 12000 14000
8.2539 8.2539 8.254 8.254 8.2541 8.2541 8.2542 8.2543 8.2543 8.2544 8.2544
GPS time (s)
x 10^8



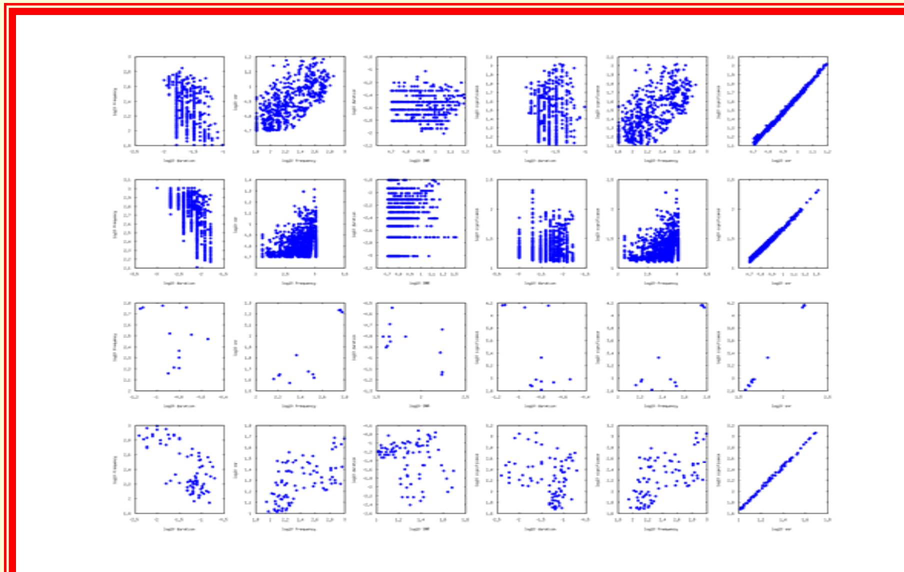
Distribution of individual class properties in a day in H1:LSC-DARM_ERR



2D relative scatter of individual class members in a day in H1:LSC-DARM_ERR



Relative distribution of individual class members in a Day in H1:LSC-DARM_ERR



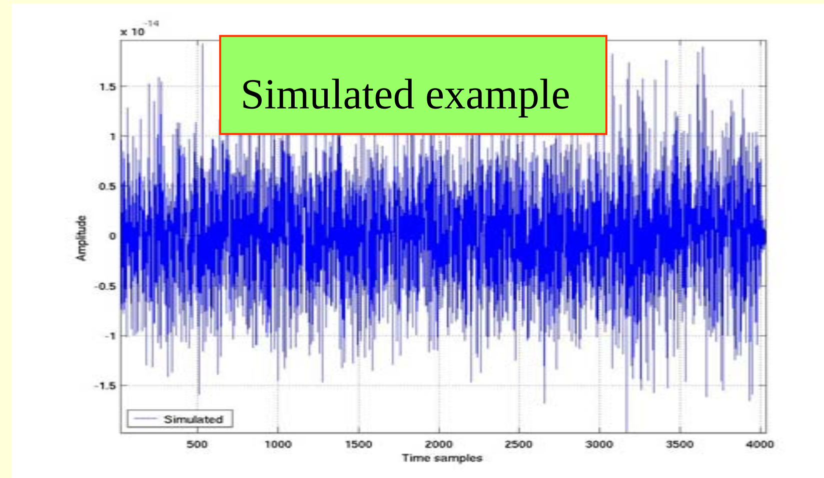
2D scatter of individual class properties in a day In H1:LSC-DARM_ERR

Post-analysis & Information Extraction II

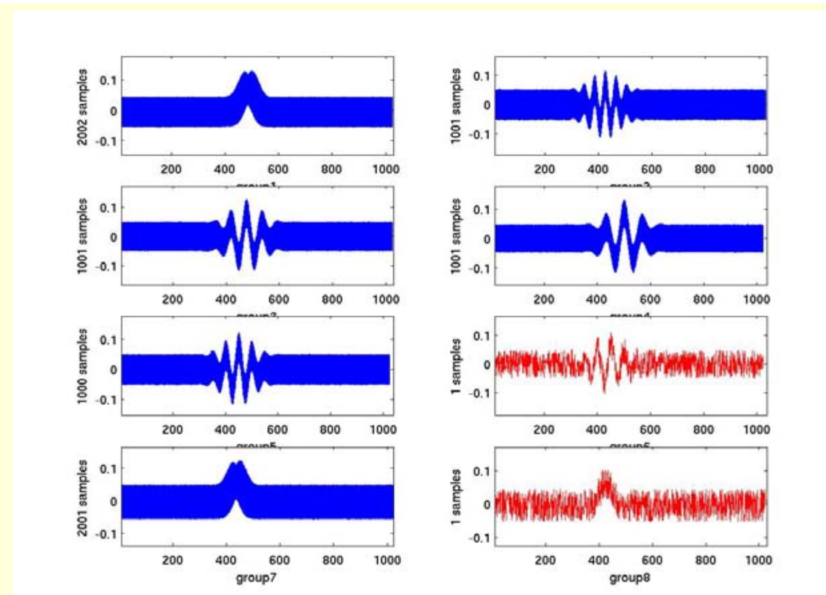
- Across the channel correlation and correspondence : Targeted Analysis
  - Select a window around the time of interest
  - Construct a database of all channels in that window
  - Look at classes across channels
  - Run *S-means** to see how the waveforms classify in the DARM_ERR
  - Infer and catalog correlated channels

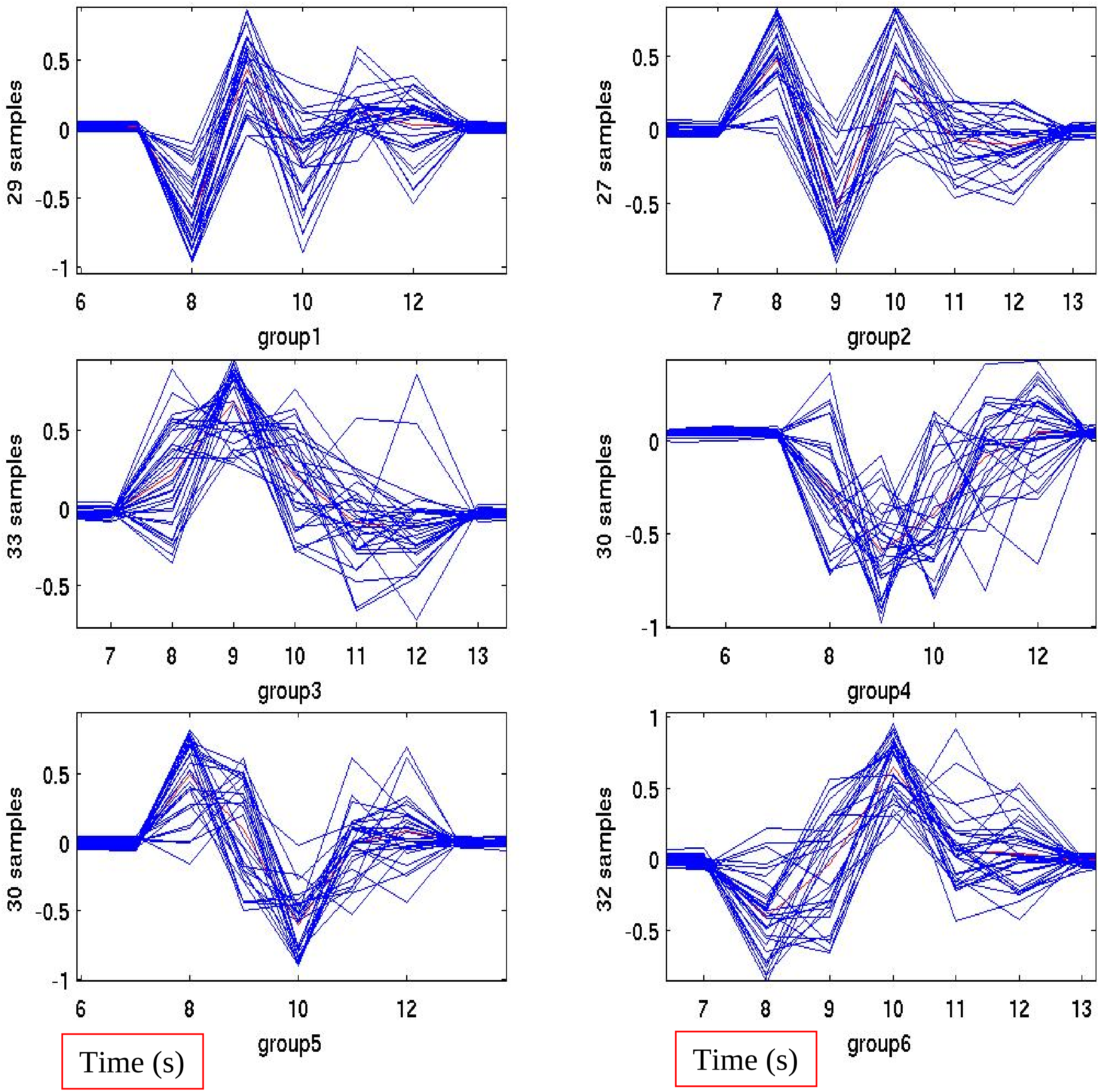
*Lei, Tang, Mukherjee, Mohanty, Iglesias, 2007, ECML/PKDD.

Algorithm : Similarity driven time series classification



This algorithm is based on *s-means* which classifies data by assignment of *k* centroids chosen *a priori* and then partitioning data based on association of data points to the nearest centroid. In the final step, the algorithm minimizes an objective function, which in this case is a squared error function. The method is an iterative one where the centroids are re-calculated until stability is reached.

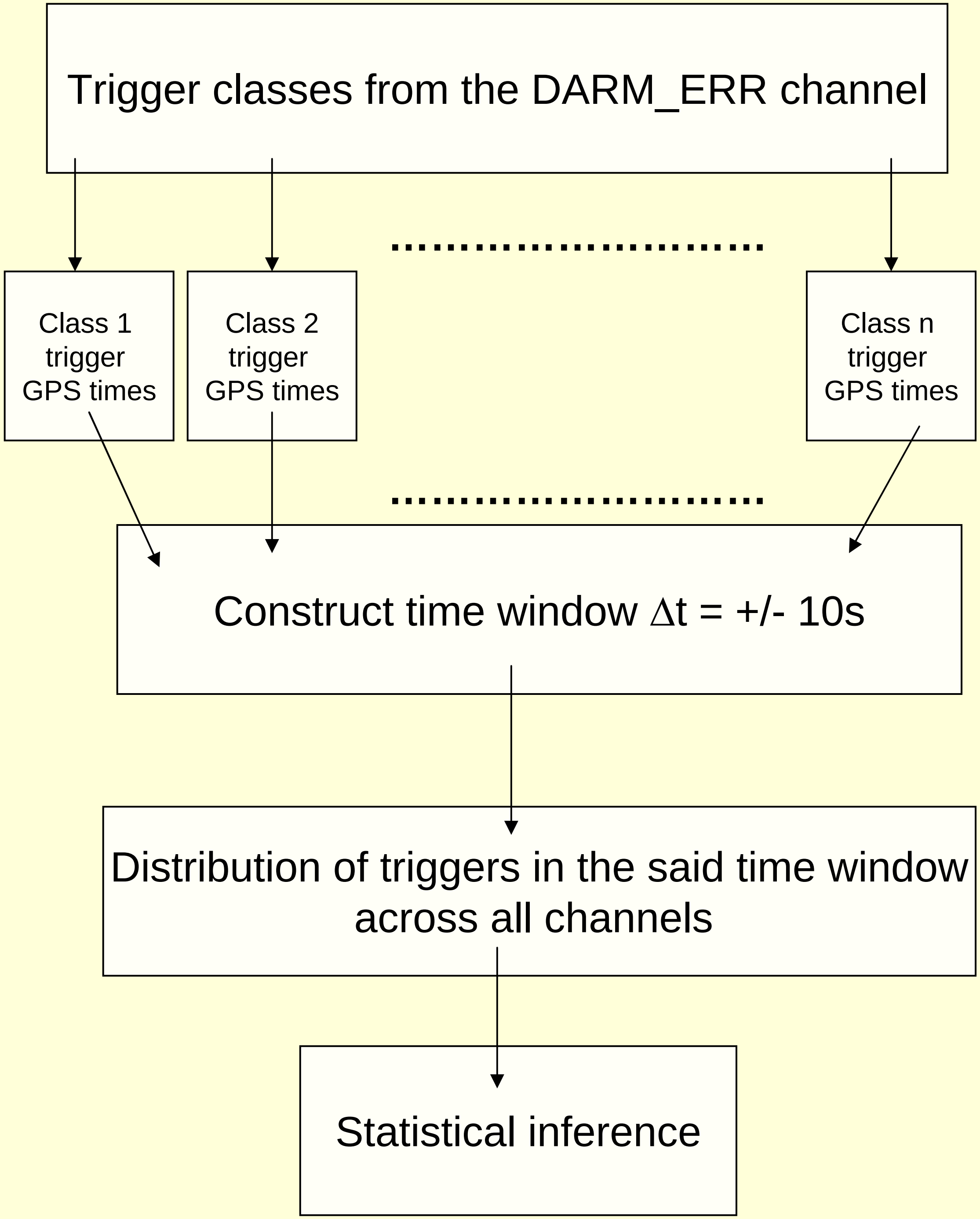




L1:LSC-DARM_ERR

**GPS : 862588800-
862675200 s.**

**6 distinct groups found by Similarity driven classification algorithm. These classes are based on the shape of the triggers obtained by retaining 128 Hz around the central frequency and band passing.
SNR cut= 30**

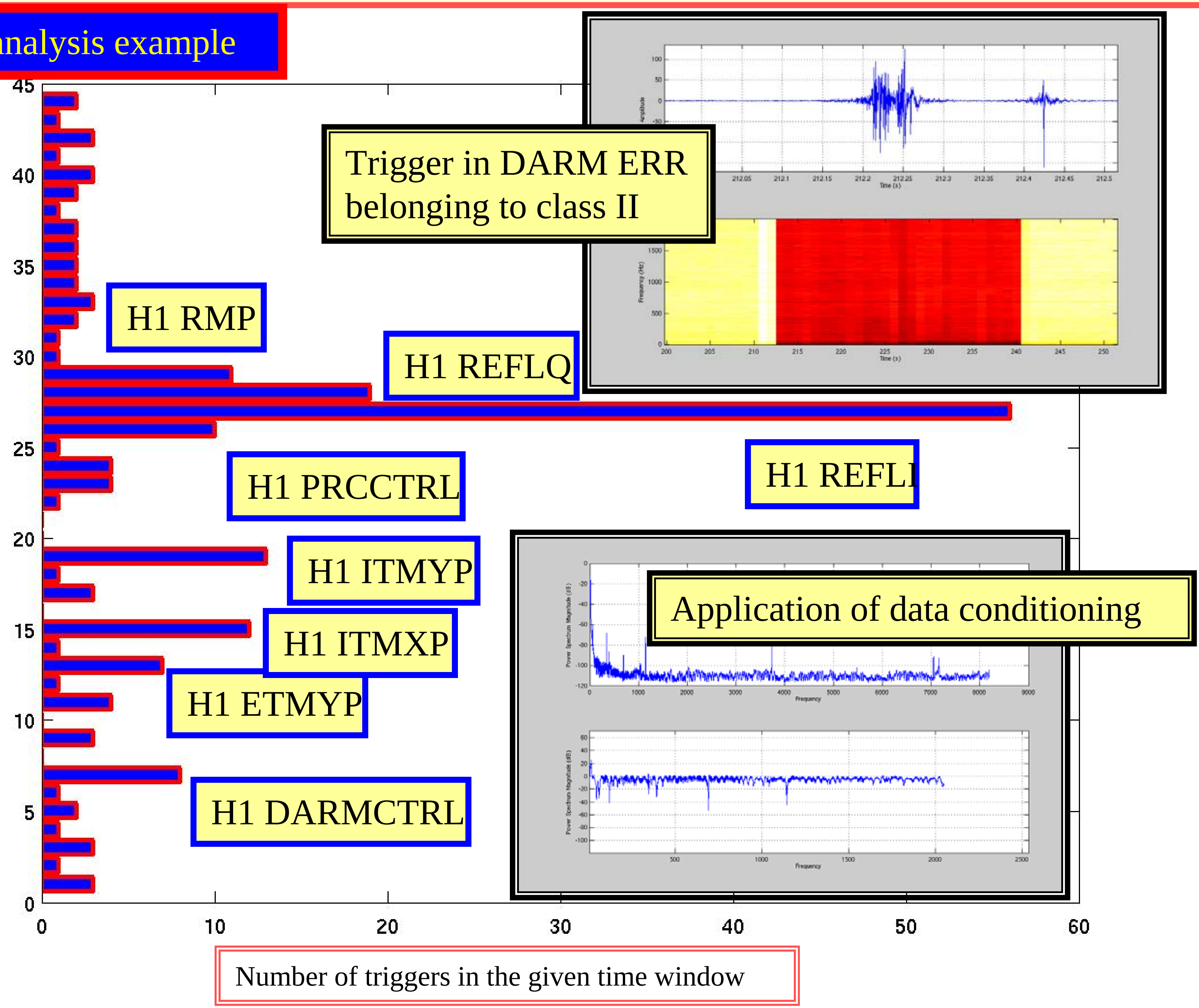


Conclusions:

Trigger *A* belonging to Class *m* in DARM_ERR channel in a certain interferometer shows statistically significant large values in the trigger distribution of the auxiliary channels in the said time window.
Accumulated results over the S5 run lead to identification of possible vetoes.

	A	B	C	D	E	F	G
1	Channel Name	GPS Time Window = 20s; GPS Central = 830661844; GPS+/-10=830661834-830661854					
2	s5_h1_asac_830620800_830707200.trg (3)	830661844.31	830661844.44	830661844.37	577	1.68E+009	30314.7
3		830661844.45	830661844.52	830661844.46	392	3.60E+007	657.47
4		830661844.47	830661844.47	830661844.47	1024	1.53E+006	30.83
5	s5_h1_asdc_830620800_830707200.trg (1)	830661844.25	830661844.75	830661844.39	69	1.25E+008	3936.39
6	s5_h1_asi_830620800_830707200.trg (3)	830661844.33	830661844.46	830661844.36	574	5.88E+002	30657.43
7		830661844.47	830661844.47	830661844.47	1024	5.11E-001	26.16
8		830661844.47	830661844.5	830661844.49	306	1.72E+000	128.55
9	s5_h1_asq_830620800_830707200.trg (1)	830661844.31	830661844.52	830661844.35	578	1.00E+003	30229.79
10	S5_h1_bsp_830620800_830707200.trg (2)	830661836.21	830661836.22	830661836.22	256	3.44E-013	17.62
11		830661844.25	830661844.5	830661844.4	15	5.85E-012	325.81
12	S5_h1_bsy_830620800_830707200.trg (1)	830661844.25	830661844.53	830661844.43	24	3.37E-012	811.73
13	s5_h1_darmctrl_830620800_830707200.trg (8)	830661844.31	830661844.48	830661844.35	555	5.91E+004	25024.88
14		830661844.42	830661844.42	830661844.42	1024	4.18E+001	17.76
15		830661844.44	830661844.45	830661844.44	1024	6.99E+001	29.73
16		830661844.54	830661844.55	830661844.54	709	1.06E+004	4338.32
17		830661844.55	830661844.58	830661844.57	479	1.45E+003	638.73
18		830661844.59	830661844.59	830661844.59	1024	4.91E+001	20.9
19		830661844.59	830661844.6	830661844.59	512	6.36E+001	22.91
20		830661844.6	830661844.61	830661844.6	256	6.79E+001	37.15
21	s5_h1_darmctrlexcdaq_830620800_830707200.trg (0)						
22	s5_h1_darmerrhifreq_830620800_830707200.trg (3)	830661844.32	830661844.35	830661844.33	5046	2.51E-005	144897.04
23		830661844.35	830661844.37	830661844.36	4950	1.59E-005	91974.64
24		830661844.53	830661844.54	830661844.53	2946	6.66E-007	3964.21
25	s5_h1_etmxexcdaq_830620800_830707200.trg (0)						
26	s5_h1_etmxp_830620800_830707200.trg (4)	830661844.31	830661844.42	830661844.33	304	5.87E+002	13547.18
27		830661844.45	830661844.47	830661844.46	128	6.88E-001	21.53
28		830661844.47	830661844.59	830661844.53	32	7.82E-001	38.93
29		830661845.92	830661845.93	830661845.93	428	1.52E+000	35.34
30	s5_h1_etmxy_830620800_830707200.trg (1)	830661844.25	830661844.56	830661844.34	287	1.63E+003	14754.95
31	s5_h1_etmyp_830620800_830707200.trg (7)	830661844.31	830661844.5	830661844.34	303	1.68E+002	11571.78
32		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Targeted analysis example



Future, Timescales, Logistics

- Specific Applications
  - Combination of Hierarchical classification and S-means
    - Example of direct immediate application – Waveburst reconstructed waveforms.
  - CVclust (in algorithm development stage)
- Other databases
  - Tracksearch
  - Virgo, GEO, …

Estimated timelines

- December 10 2007 – All S5 analyzed and results made available.
- December 2007 LSC-Virgo/GWDAW – Results from the
  - Trigger Visualization
  - Across the channel correlation

March F2F/LSC-Virgo – Estimated completion of the first two bulleted activities, continuation of more (e.g. other databases, other detectors) …

This work is supported by the National Science Foundation grant PHY-0555842 (2006).