Improved Machine Learning Algorithm using Information Theoretic Feature Selection for Classifying Noise Artifact of Gravitational-Wave Data

John J. Oh (NIMS)

on behalf of

Edwin J. Son, Sang Hoon Oh (NIMS), Kyungmin Kim, Hyunkyu Lee (Hanyang Univ.), Young-Min Kim (Pusan National Univ.), Lindy Blackburn (Caltech), Ruslan Vaulin (MIT), Kazuhiro Hayama (Osaka Univ.)

2014. 6. 21 (Sat)

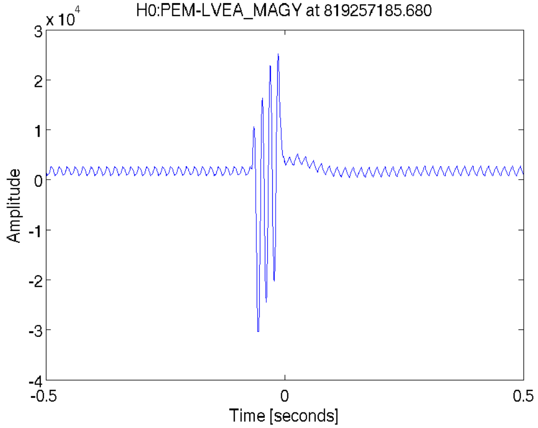
The 6th Korea-Japan Workshop on KAGRA

NAOJ, Japan

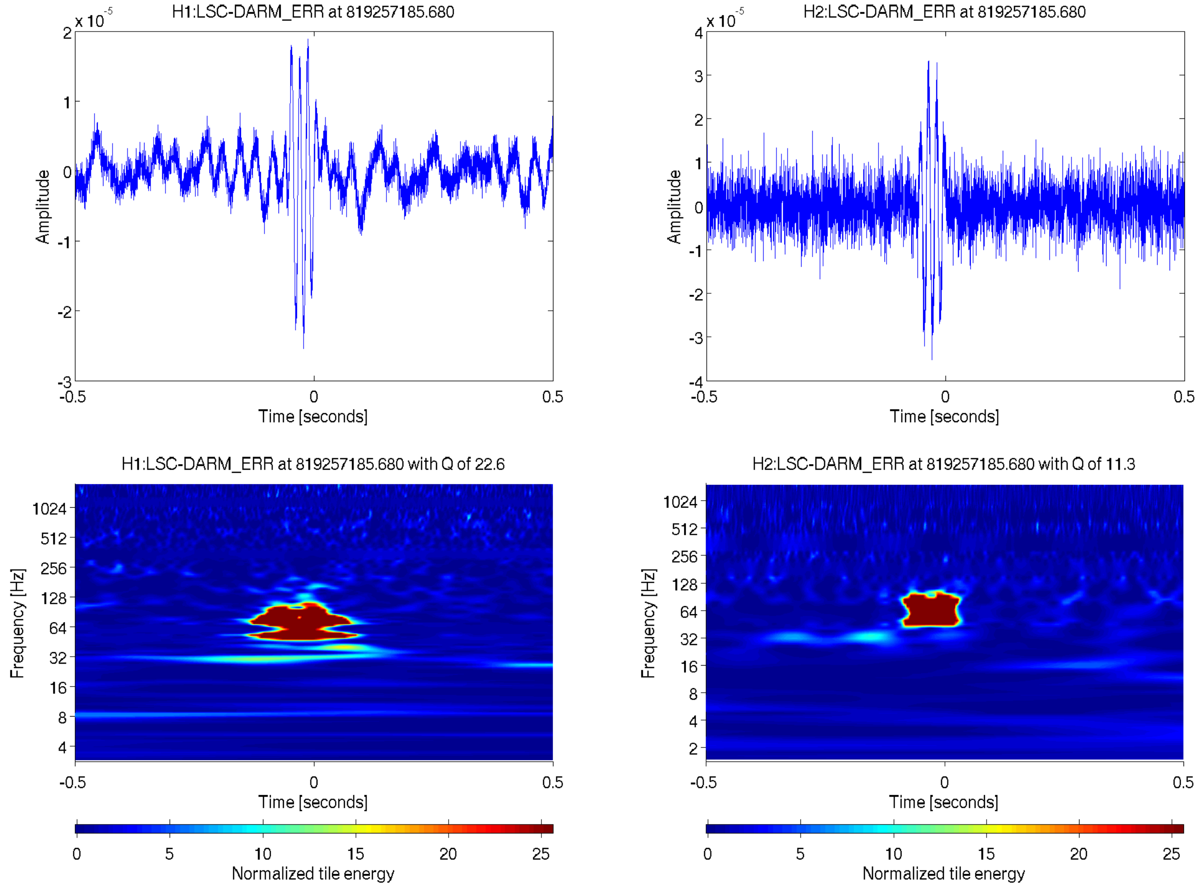
Contents

Introduction

Power line glitch (noise transient)



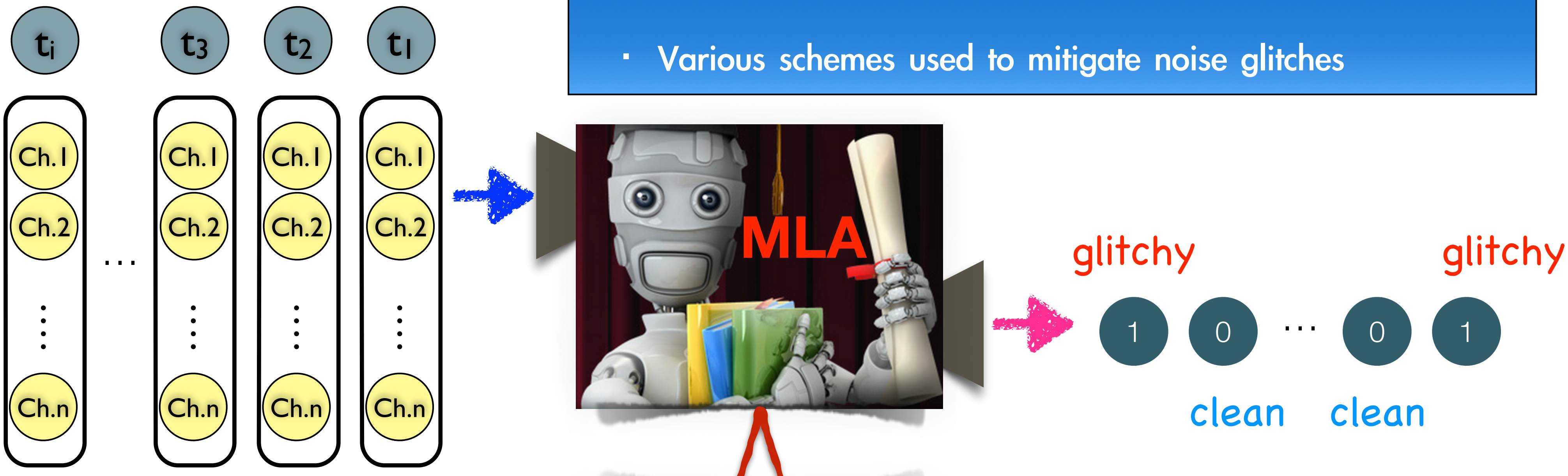
Magnetometer channel

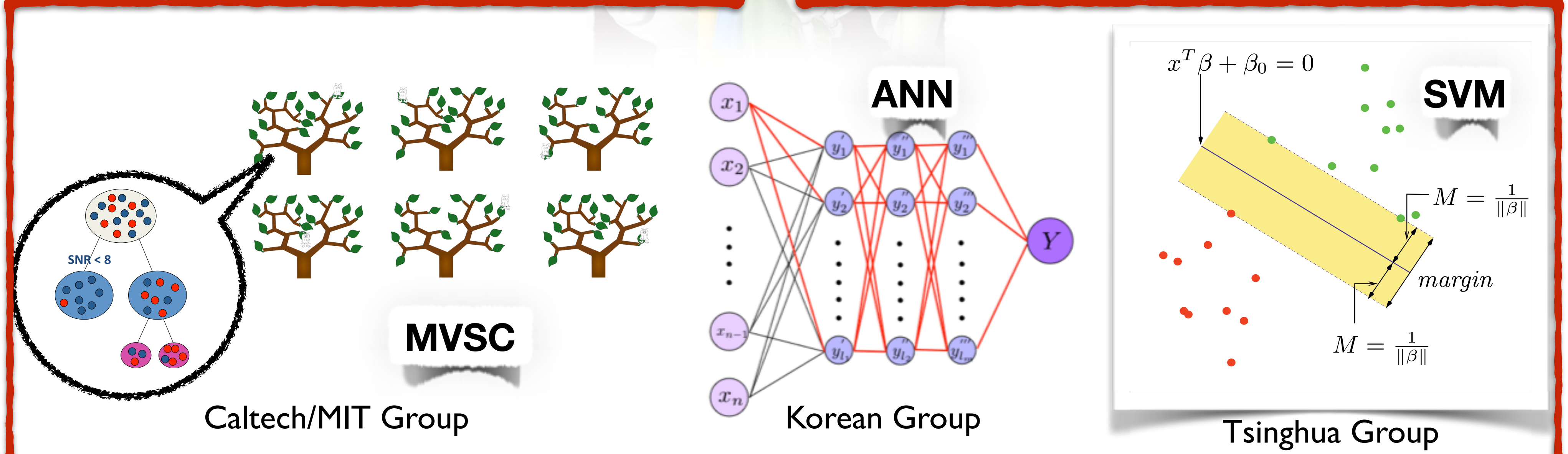


- initial LIGO ~1,000 aux channels
- advanced LIGO ~10,000 aux channles

Introduction

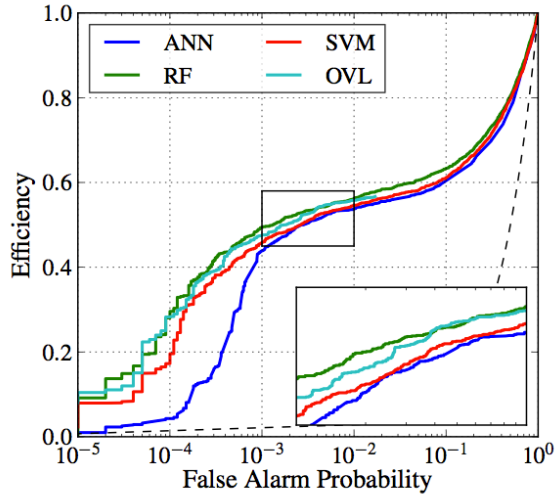
- Reducing noise artifacts in Auxiliary channels:
  - maintain the "data quality" for DA
  - monitor the instrumental abnormality
- Various schemes used to mitigate noise glitches

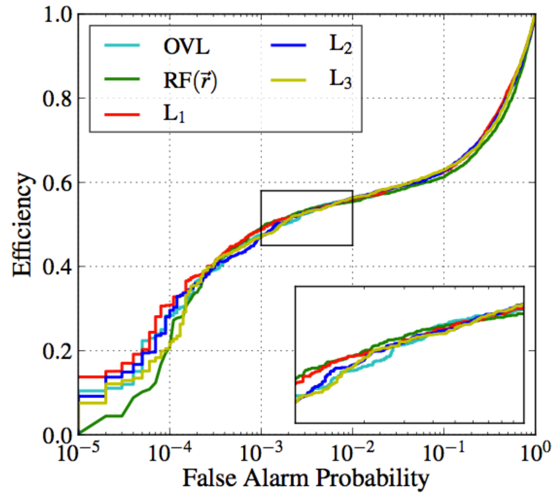




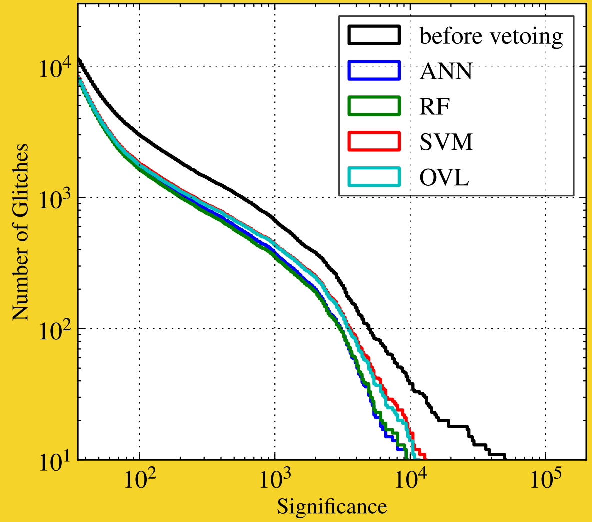
Introduction

Biswas, Blackburn, Hodge, Oh, Oh, Son, Vaulin, Kim, Kim, Lee, et.al., Phys. Rev. D 88, 062003, 2013

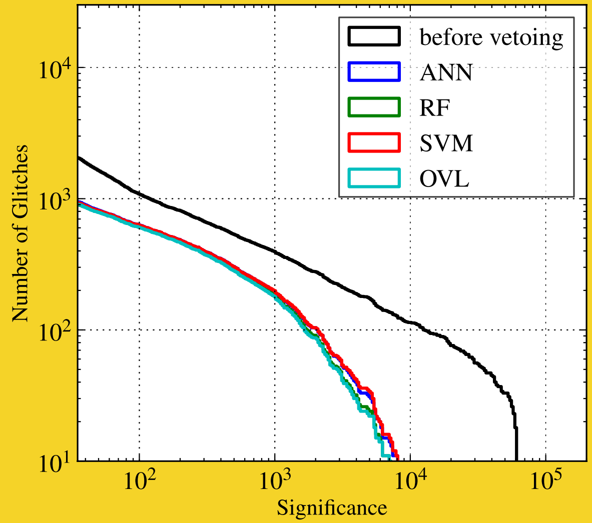




S4: 30% reduction
S6: 55% reduction

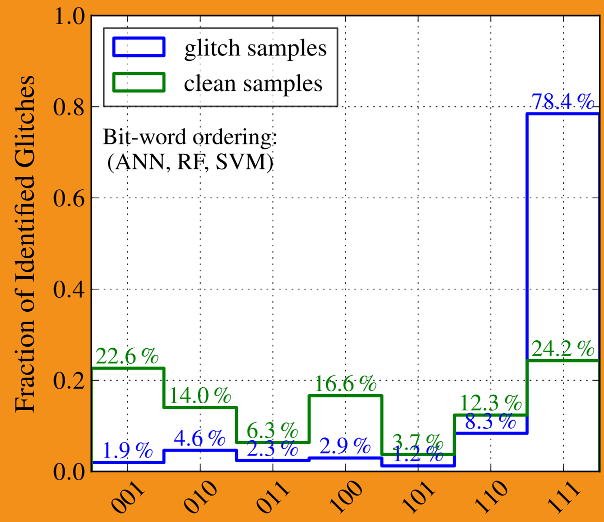


(a) S4 glitches

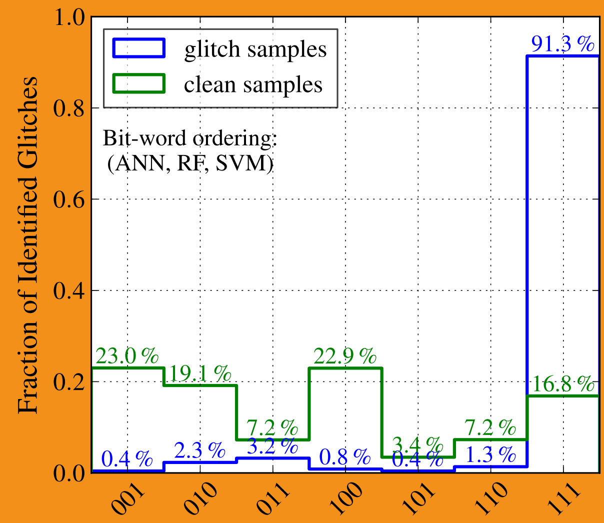


(b) S6 glitches

Redundancy b.t. MLAs



(a) S4 bit-word histogram for MLAs



(b) S6 bit-word histogram for MLAs

Introduction

Input Feature Selection and Algorithms for ANNs

Dimensional Reduction

- **Rotation**
  - **Linear Method**
    - Principal Component Analysis (PCA)
    - Partial Least-Squares (PLS) (World, 1966)
  - **Non-linear Method**
    - Independent Component Alaysis (ICA)
    - Non-linear PCA (NLPCA)
- **Clustering**
  - Learning Vector Quantization (LVQ)
  - Self-Organizing Map (SOM) (Bowden et al, 2002)

Feature Selection

- **Model-based**
  - **Wrapper**
    - Nested
      - Forward Selection (Constructive ANNs)
      - Backward Elimination
      - Nested Subset (e.g. Increasing Delay Order)
    - Global Search
      - Exhaustive Search
      - Heuristic Search (e.g. GA-ANN)
    - Ranking
      - Single-Variable Ranking (SVR)
      - GRNN Input Determination Algorithm (GRIDA)
  - **Embedded**
    - Optimization
      - Direct Optimization (e.g. Lasso)
      - Evolutionary ANNs
    - Weight-based
      - Stepwise Regression
      - Pruning (e.g. OBD (Le Cun et al., 1990))
- **Filter (Model-Free)**
  - **Correlation (linear)**
    - Rank (Maximum) Pearson Correlation
    - Ranked (Maximum) Spearman Correlation
    - Forward Partial Correlation Selection
    - Time-series analysis (Box & Jenkins, 1976)
  - **Information Theoretic (nonlinear)**
    - Entropy
      - Entropy (minimum) Ranking
      - Minimum Entropy
    - Mutual Information (MI)
      - Rank (maximum) MI
      - MI Feature Selection (MIFS) (Battiti, 1994)
      - MI w/ ICA (ICAIVS) (Back & Trappenberg, 2001)
      - Partial Mutual Information (PMI) (Sharma, 2000)
      - Joint MI (JMI) (Bonnlander & Weigend, 1994)

Robert May, Graeme Dandy and Holger Maier, "Review of Input Variable Selection Methods for Artificial Neural Networks", Artificial Neural Networks - Methodological Advances and Biomedical Applications, Edited by Kenji Suzuki, ISBN 978-953-307-243-2, 362 pages, Publisher: InTech, Chapters published.

Introduction

- **Project Goal:**
  - Reducing number of input features by selecting mostly contributed features : computational speed-up
  - Removing redundant and/or harmful features to the classification performance by feature selection

- **Methods:**
  - Normalized Mutual Information Feature Selection (Nonlinear)

- **Data:**
  - ALL_S6_959126400_hveto_channels_signif_dt (101,819 samples/ 35 channels / 2 attributes)

Information Theoretic Method (MiGANN)

- **Mutual Information Coefficient: (Information Theory)**
  - mutual information of two discrete random variables:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

  where p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y.
  - Intuitively, it measures the information that X and Y share: how much knowing one of these variables reduces uncertainty about the other.
  - If both are independent variables, I(X;Y) = 0, no mutual information to share.

- **Algorithm: NMIFS**

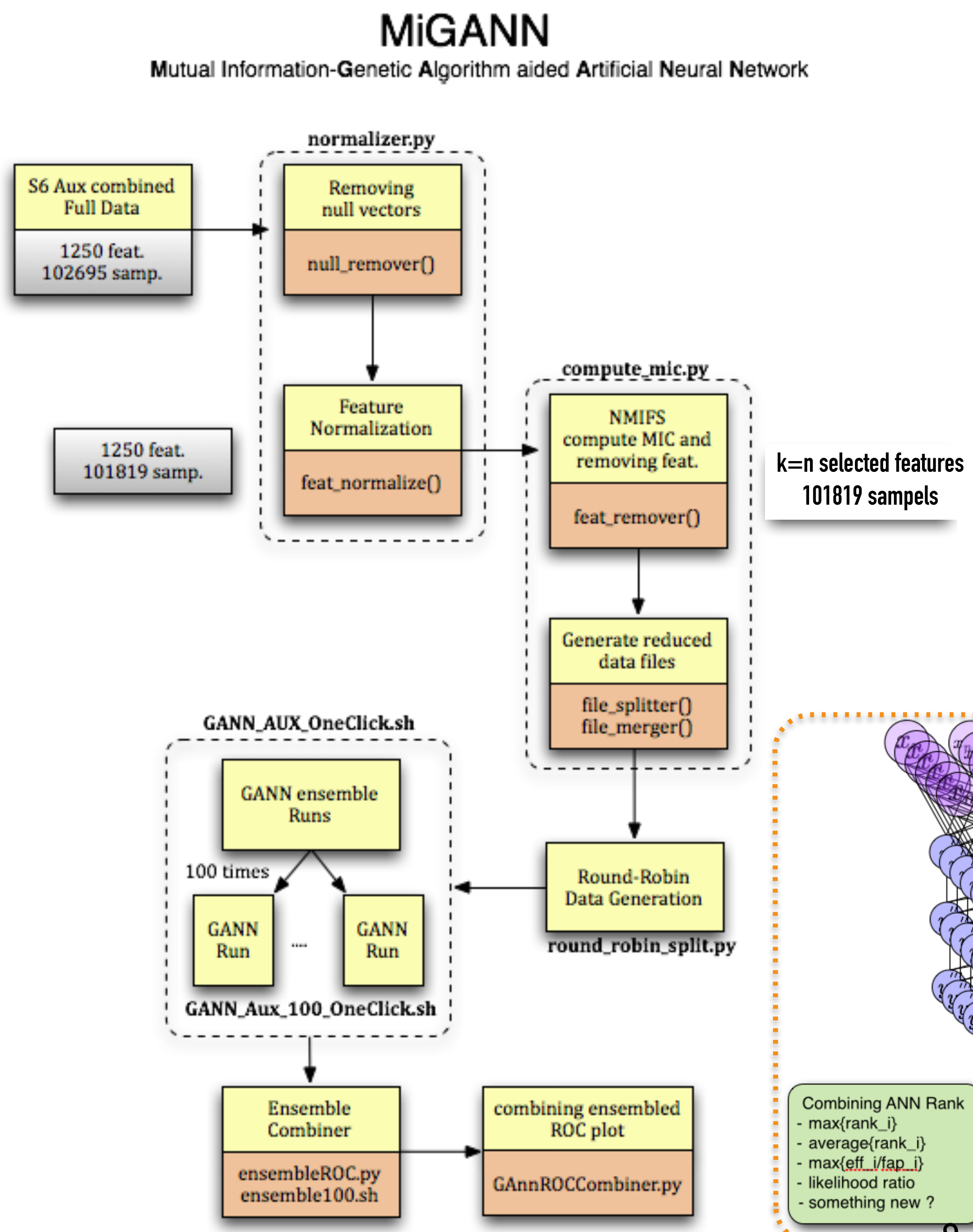
Ref.) Pablo A. Estevez, Michel Tesmer, Claudio A. Perez, and Jacek M. Zurada, "Normalized Mutual Information Feature Selection", IEEE Transactions on Neural Networks, Vol. 20, No2. 189 (2009)

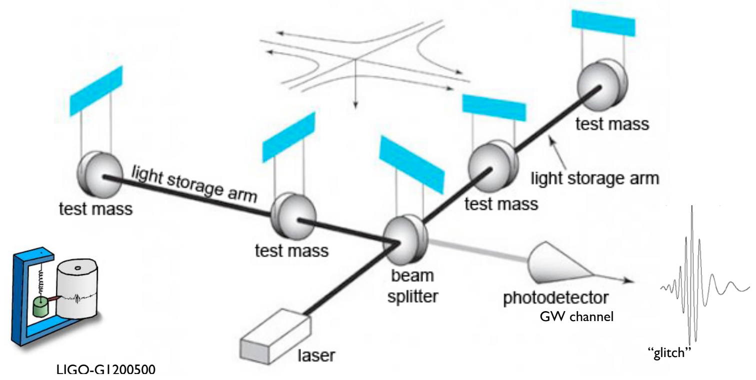
1. Initialization: Set $F = \{f_i/i = 1, \cdots, n\}$, (initially N-features) and $S = \{\}$
2. Compute MI w.r.t Classes: $I(f_i; C)$ for each $f_i \in F$.
3. Select the first feature: FIND $\hat{f}_i = max_{i=1,\cdots,N}\{I(f_i;C)\}$ and set $F \leftarrow F \backslash \{\hat{f}_i\}$ and set $S \leftarrow \{\hat{f}_i\}$.
4. Greedy selection: REPEAT until $|S| = k$.
   - Compute the MI between features: $I(f_i; f_s)$ for all pairs of (f_i, f_s), with $f_i \in F$ and $f_s \in S$
   - Select the next features: Select features $f_i \in F$ that maximize:

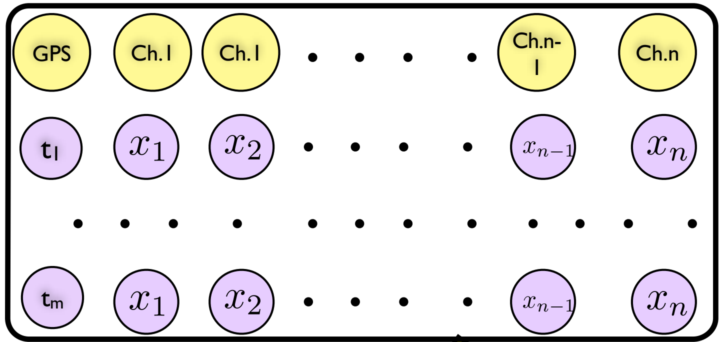
$$G \equiv I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} I_n(f_i; f_s). \quad \text{Set } F \leftarrow F \backslash \{f_i\} \text{ and set } S \leftarrow \{f_i\}.$$

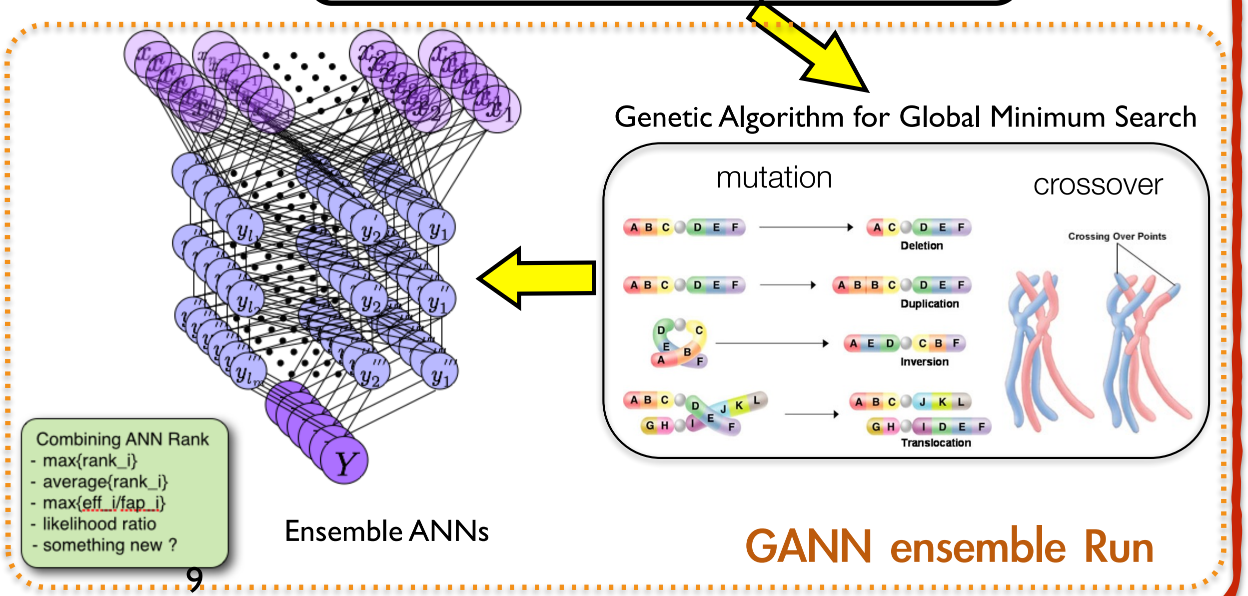
5. Output: the set S containing the selected features:

Information Theoretic Method (MiGANN)





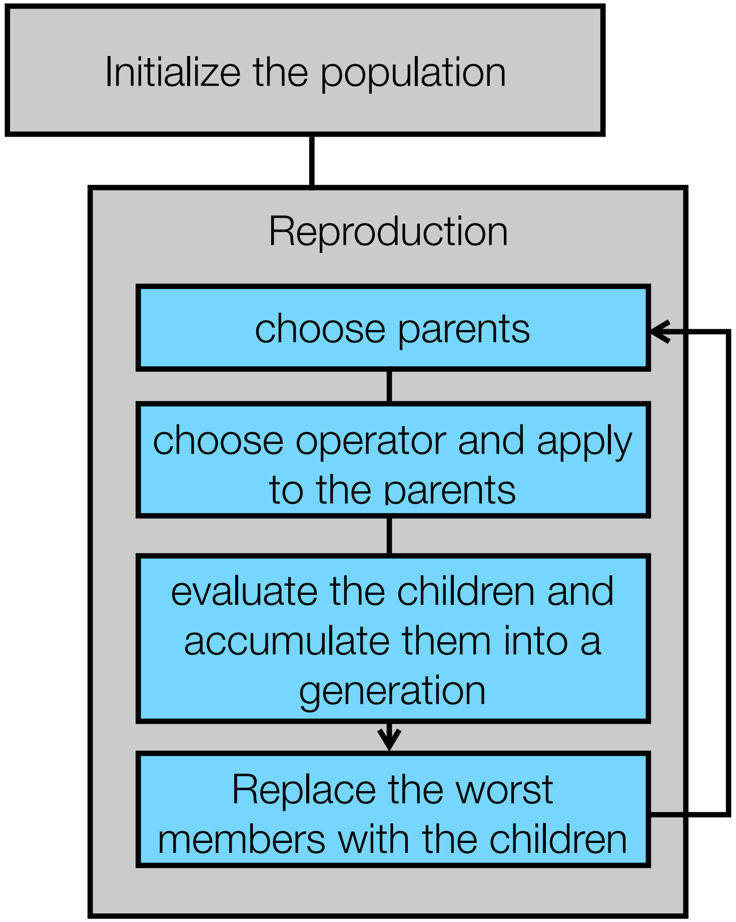


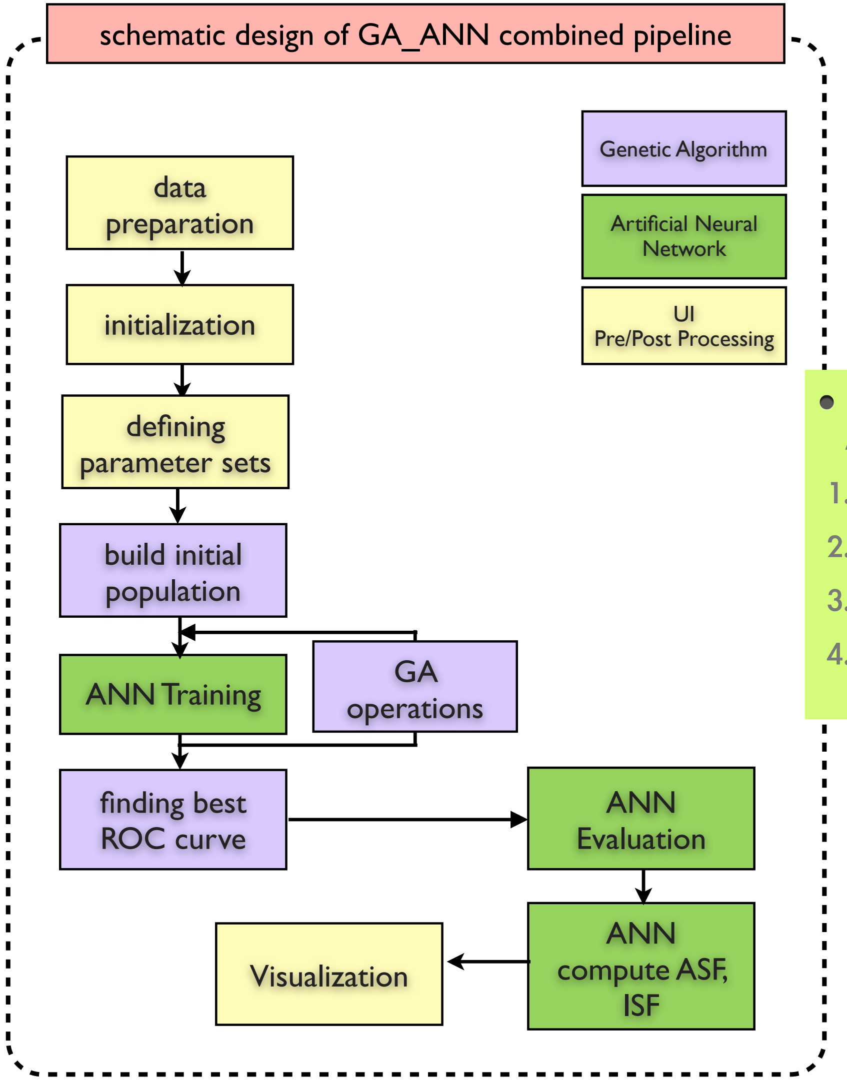


Information Theoretic Method (MiGANN)

- Optimization / Machine learning (loosely) based on biological evolution; ***natural selection of genes***

Search global optimum





- GA's incorporation with ANN
  1. Feature selection
  2. Topology selection
  3. Weight selection
  4. To learn neural network learning algorithm

Application to S6 Aux.Chan.Data

ANN

L1_OMC-QPD1_P_OUT_DAQ_32_2048=667289.817642
L1_OMC-QPD2_Y_OUT_DAQ_32_2048=660339.342513
L1_OMC-QPD2_P_OUT_DAQ_32_2048=648880.835468
L1_OMC-PZT_LSC_OUT_DAQ_8_1024=611644.083158
L1_OMC-QPD3_P_OUT_DAQ_8_1024=560136.453594
L1_OMC-QPD1_SUM_OUT_DAQ_32_2048=464637.187728
L1_OMC-QPD2_SUM_OUT_DAQ_32_2048=340147.925119
L1_ISI-OMC_CONT_RZ_IN1_DAQ_8_1024=321934.370699
L1_LSC-REFL_Q_32_2048=281685.246617
L1_OMC-QPD4_P_OUT_DAQ_8_1024=247231.541518
L1_OMC-QPD4_Y_OUT_DAQ_8_1024=238743.180353
L1_OMC-PZT_VMON_AC_OUT_DAQ_32_2048=213446.834119
L1_OMC-QPD3_Y_OUT_DAQ_8_1024=210633.289612
L1_ISI-OMC_GEOPF_H1_IN1_DAQ_8_1024=201769.172767
L1_ASC-WFS4_IP_8_256=186508.980033
L1_ASC-WFS3_IP_8_256=174079.131805
L1_ASC-RM_P_8_256=147316.587591
L1_LSC-POB_I_1024_4096=141248.58001
L1_LSC-PRC_CTRL_32_2048=140997.938081
L1_LSC-POB_I_32_2048=132375.465581
L0_PEM-HAM1_ACCZ_8_1024=130893.99712
L1_ASC-ETMX_P_8_256=127461.52814
L1_ASC-ETMY_P_8_256=114186.921959
L1_ISI-OMC_GEOPF_H2_IN1_DAQ_8_1024=108914.685549
L1_ASC-ITMY_P_8_256=108810.943927
L0_PEM-EY_SEISY_8_128=108150.738939
L1_ISI-OMC_GEOPF_V2_IN1_DAQ_8_1024=96507.9823112
L1_ASC-ITMX_P_8_256=91580.4862366
L1_ASC-WFS2_QP_8_256=84060.8204434
L1_ASC-WFS2_IP_8_256=75559.9162881
L1_ASC-WFS1_QP_8_256=73135.1890178
L1_OMC-DUOTONE_OUT_DAQ_1024_4096=69635.8853255
L1_SUS-ETMY_SENSOR_SIDE_8_256=69559.2061476
L1_ASC-QPDY_Y_8_128=63663.2104693
L1_SEI-ETMX_Y_8_128=61582.3383791

Only 19 ANN Channels matched with OVL

OVL

L1_LSC-POB_Q_1024_4096
L1_OMC-PZT_LSC_OUT_DAQ_8_1024
L1_ISI-OMC_GEOPF_H2_IN1_DAQ_8_1024
L0_PEM-LVEA_SEISZ_8_128
L1_ISI-OMC_GEOPF_H1_IN1_DAQ_8_1024
L1_ASC-ITMY_P_8_256
L0_PEM-LVEA_BAYMIC_8_1024
L1_ASC-BS_P_8_256
L1_LSC-POB_Q_32_2048
L1_ASC-ITMX_P_8_256
L1_ASC-WFS1_QY_8_256
L1_SUS-ETMX_SENSOR_SIDE_8_256
L1_SUS-ETMY_SENSOR_SIDE_8_256
L0_PEM-EX_SEISX_8_128
L1_ASC-ITMY_Y_8_256
L1_OMC-QPD1_P_OUT_DAQ_32_2048
L1_ASC-WFS2_IP_8_256
L1_ASC-ETMX_Y_8_256
L1_OMC-PZT_VMON_AC_OUT_DAQ_32_2048
L1_OMC-QPD1_SUM_OUT_DAQ_32_2048
L1_ASC-WFS2_QY_8_256
L1_SUS-RM_SUSPIT_IN_8_32
L1_OMC-QPD2_Y_OUT_DAQ_32_2048
L1_OMC-QPD2_SUM_OUT_DAQ_32_2048
L1_OMC-QPD2_P_OUT_DAQ_32_2048
L1_ASC-ITMX_Y_8_256
L1_ASC-WFS2_IY_8_256
L1_OMC-QPD3_Y_OUT_DAQ_8_1024
L1_LSC-PRC_CTRL_32_2048
L1_OMC-QPD3_P_OUT_DAQ_8_1024
L0_PEM-HAM1_ACCZ_8_1024
L1_ASC-WFS1_QP_8_256
L0_PEM-EY_MAGX_1_1024
L1_OMC-QPD4_P_OUT_DAQ_8_1024
L0_PEM-LVEA_MAGY_1_1024

Top OVL 35 channels

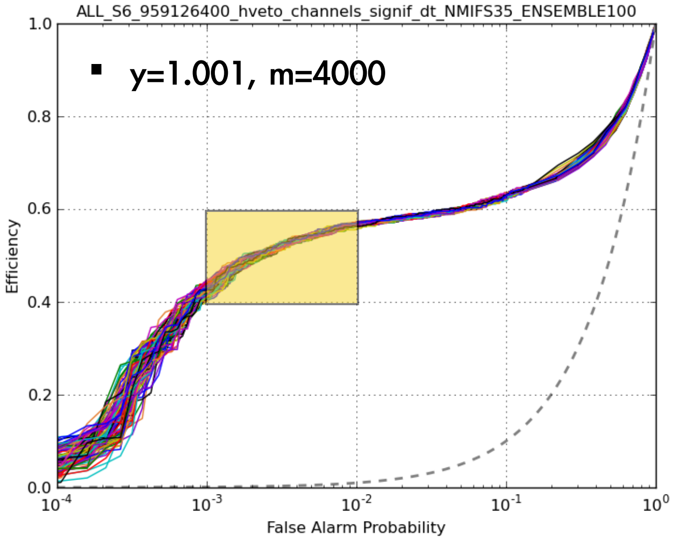
Mutual Information

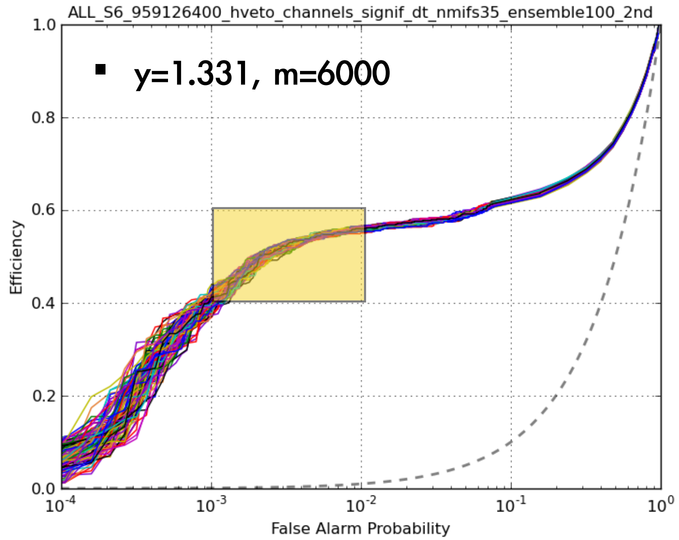
Highly Correlated Channels: Unorm
0:L1_SEI-LVEA_STS2_X_8_256_signif
630:L1_OMC-QPD3_P_OUT_DAQ_8_1024_signif
635:L1_OMC-QPD4_Y_OUT_DAQ_8_1024_signif
640:L1_OMC-QPD4_P_OUT_DAQ_8_1024_signif
765:L1_OMC-QPD1_P_OUT_DAQ_32_2048_signif
766:L1_OMC-QPD1_P_OUT_DAQ_32_2048_dt
767:L1_OMC-QPD1_P_OUT_DAQ_32_2048_dur
768:L1_OMC-QPD1_P_OUT_DAQ_32_2048_freq
769:L1_OMC-QPD1_P_OUT_DAQ_32_2048_npts
770:L1_OMC-QPD2_P_OUT_DAQ_32_2048_signif
772:L1_OMC-QPD2_P_OUT_DAQ_32_2048_dur
773:L1_OMC-QPD2_P_OUT_DAQ_32_2048_freq
774:L1_OMC-QPD2_P_OUT_DAQ_32_2048_npts
775:L1_OMC-QPD2_Y_OUT_DAQ_32_2048_signif
780:L1_OMC-QPD1_SUM_OUT_DAQ_32_2048_signif
781:L1_OMC-QPD1_SUM_OUT_DAQ_32_2048_dt
782:L1_OMC-QPD1_SUM_OUT_DAQ_32_2048_dur
783:L1_OMC-QPD1_SUM_OUT_DAQ_32_2048_freq
784:L1_OMC-QPD1_SUM_OUT_DAQ_32_2048_npts
785:L1_OMC-QPD2_SUM_OUT_DAQ_32_2048_signif
786:L1_OMC-QPD2_SUM_OUT_DAQ_32_2048_dt
787:L1_OMC-QPD2_SUM_OUT_DAQ_32_2048_dur
788:L1_OMC-QPD2_SUM_OUT_DAQ_32_2048_freq
789:L1_OMC-QPD2_SUM_OUT_DAQ_32_2048_npts
810:L1_LSC-REFL_Q_32_2048_signif
830:L1_LSC-PRC_CTRL_32_2048_signif
940:L1_ASC-WFS3_IP_8_256_signif
950:L1_ASC-WFS4_IP_8_256_signif
975:L1_ASC-ETMX_P_8_256_signif
985:L1_ASC-ETMY_P_8_256_signif
995:L1_ASC-ITMX_P_8_256_signif
997:L1_ASC-ITMX_P_8_256_dur
1005:L1_ASC-ITMY_P_8_256_signif
1007:L1_ASC-ITMY_P_8_256_dur
1015:L1_ASC-RM_P_8_256_signif

27 matched with OVL!
6 matched with ANN!

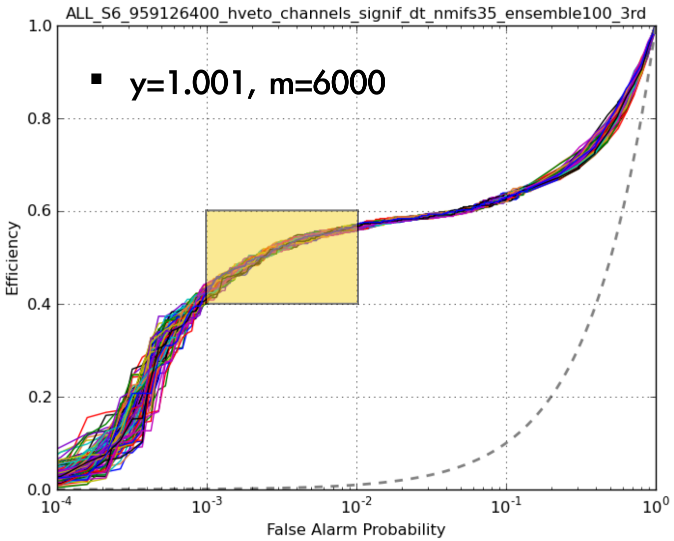
Application to S6 Aux.Chan.Data

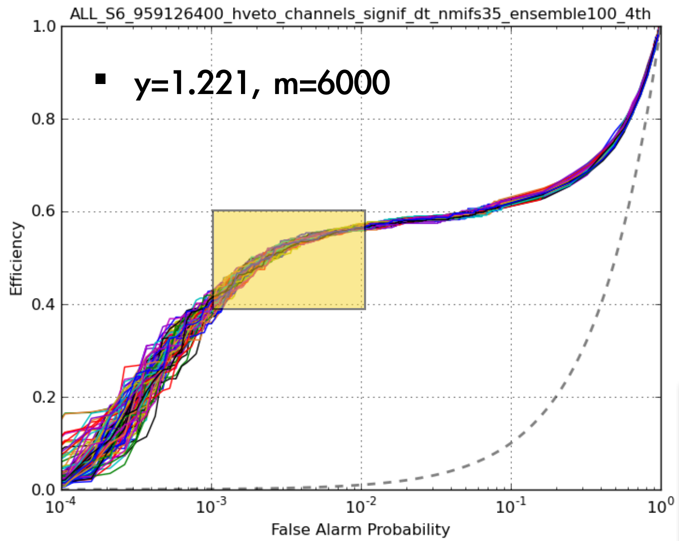
- 70 features reduced to 35 features via NMIFS algorithm
- 100 ensemble runs and plot a combined ROC with averaged rank of 20 random sampled ranks
- independent 100 runs - make 100 combined ROC plots





- n35n35n1
- Running time ~ 41 hours
- 40~43% @ 0.1% FAP
- 58~59% @1.0% FAP



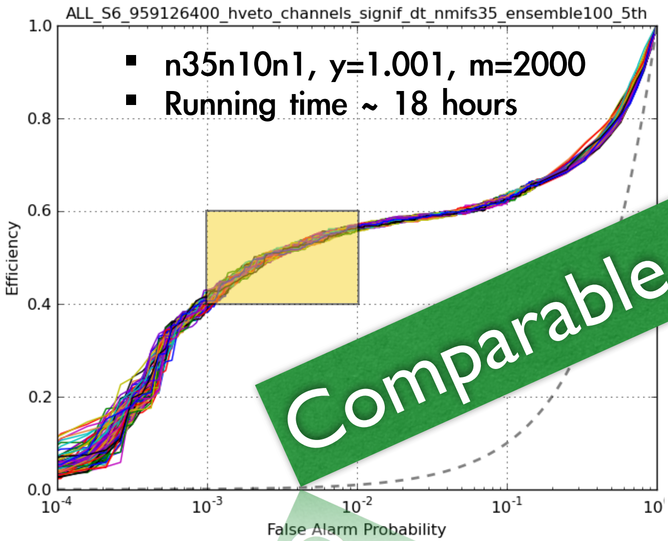


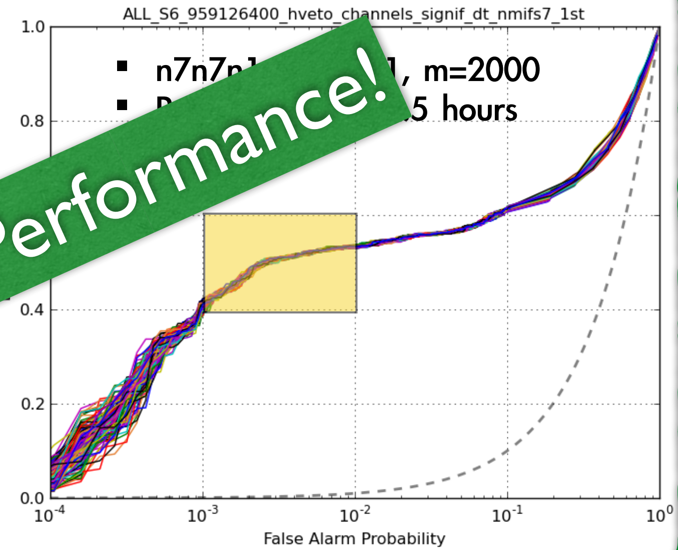
- n35n15n1
- Running time ~ 20 hours
- 40~43% @ 0.1% FAP
- 58~59% @1.0% FAP

Computing Resource: ANNE Cluster @ NIMS
- 6 cores Single node cluster
- 24 GB Memory

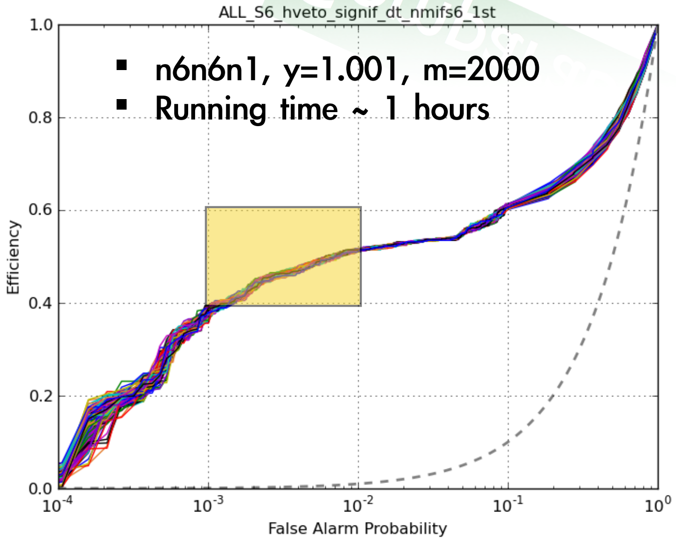
Application to S6 Aux.Chan.Data

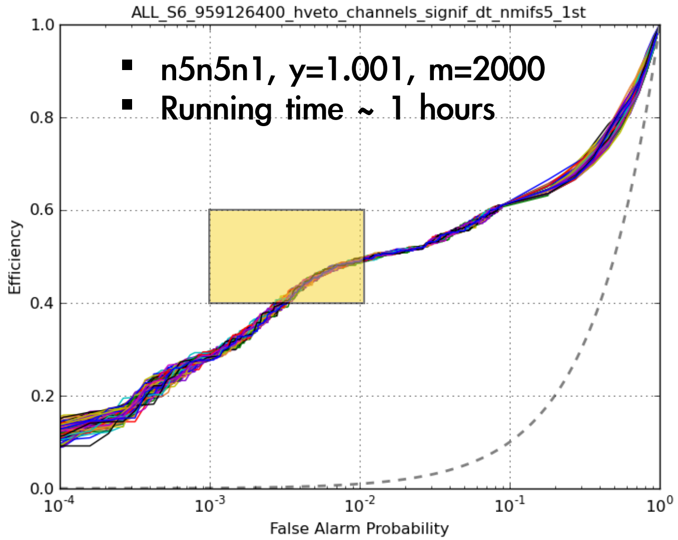
- 70 features reduced to more simpler features via NMIFS algorithm
- 100 ensemble runs and plot a combined ROC with averaged rank of 20 random sampled ranks
- independent 100 runs - make 100 combined ROC plots





- 70 → 35 / 7 both show comparable performance at our decision FAP ~ 0.1% (shaded yellow region)
- However, computationally 7-reduced feature case has huge advantages ~ 10 times faster.
- This selection rule is very crucial for selecting features among 1250 full features in AuxChannel data.



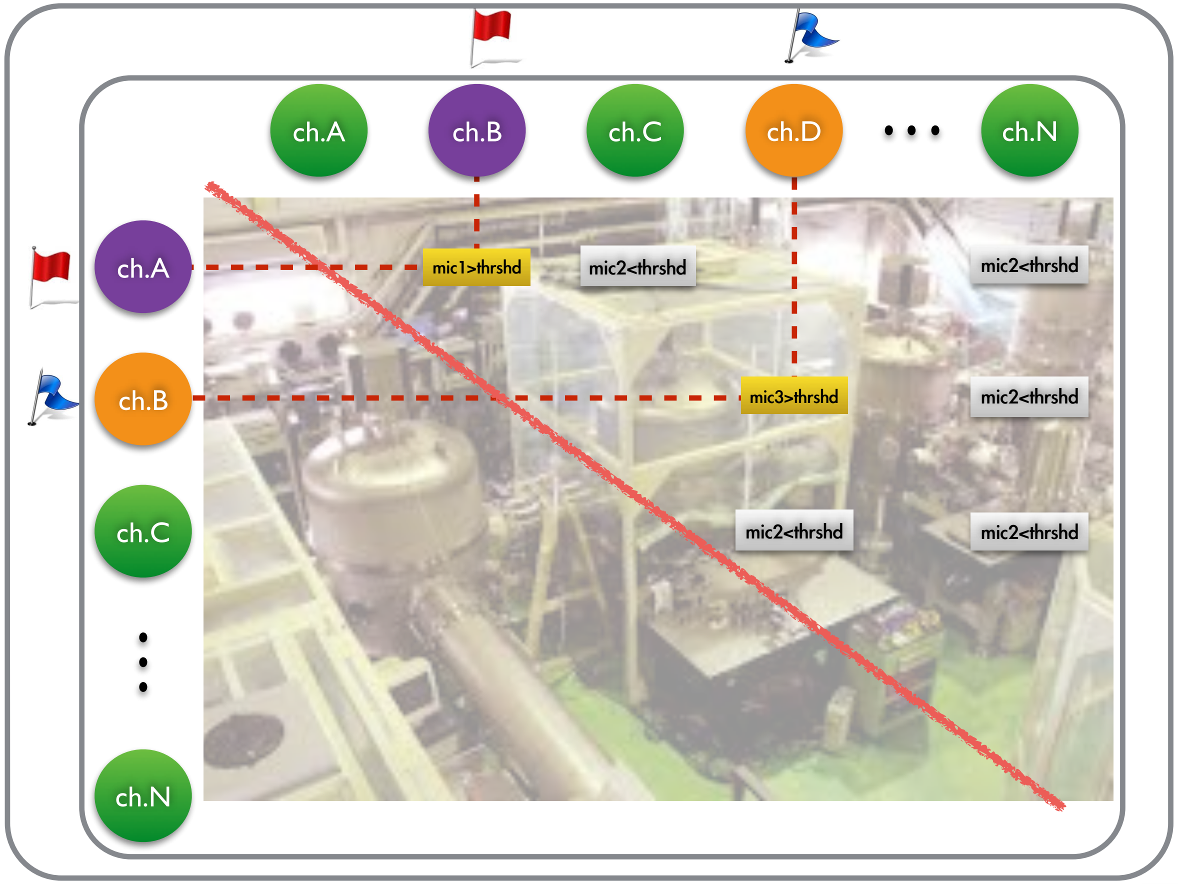


Q1) How can we find the optimal number of reduced features?
Q2) What if 1250 full feature input?
- 10% or 7 or so? varying in cases?
- More studies needed

Application to Multichannel Correlation in KAGRA

"Mutual Information Coefficient" Analysis

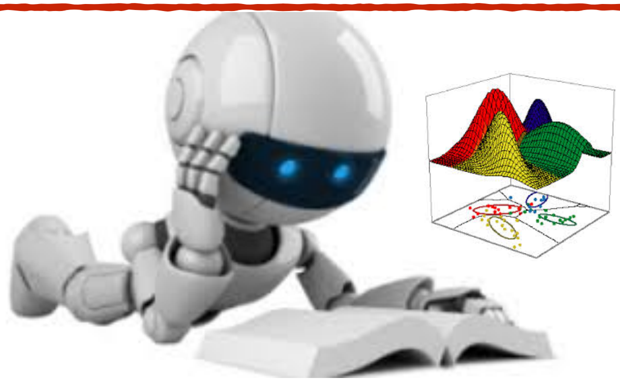
measuring correlation between two different channels

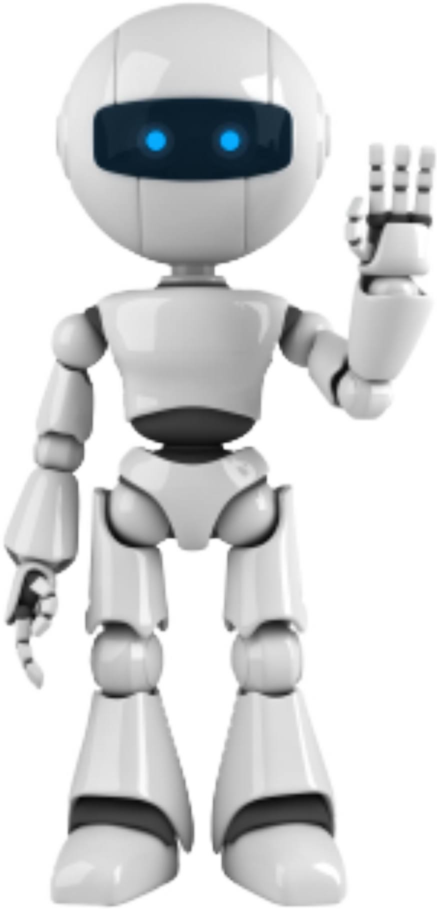


- Monitoring the correlations between channels
- Fixing instrumental via channel monitoring
- Finding more glitches that are harmful for data quality
- More studies needed in this direction

Future Plan

- Find a systematic way of minimum number of features, giving comparable performance and maximal computing speed-up.
    - Full data analysis with 1,250 features around 10^4 samples for one week data
    - Selected Channel Analysis: Comparing OVL and ANNs
- KAGRA nonlinear correlation analysis btw aux. channels. using "Mutual Information Coefficient (MIC)"






THANK YOU