BROWN

Data Analysis: Algorithms and Tools

Thomas Speer
Brown University

Introduction

XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Are we ready for LHC era experiments?

Introduction

XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Are we ready for LHC era experiments?

- The LHC presents an enormous challenge
  - Huge amount of data
  - Small signals with large backgrounds

Introduction

XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Are we ready for LHC era experiments?

- The LHC presents an enormous challenge
  - Huge amount of data
  - Small signals with large backgrounds
- Challenge for all 3 tracks !
  - Track 1: Computing Technology for Physics Research
  - Track 2: Data Analysis - Algorithms and Tools
  - Track 3: Methodology of Computations in Theoretical Physics

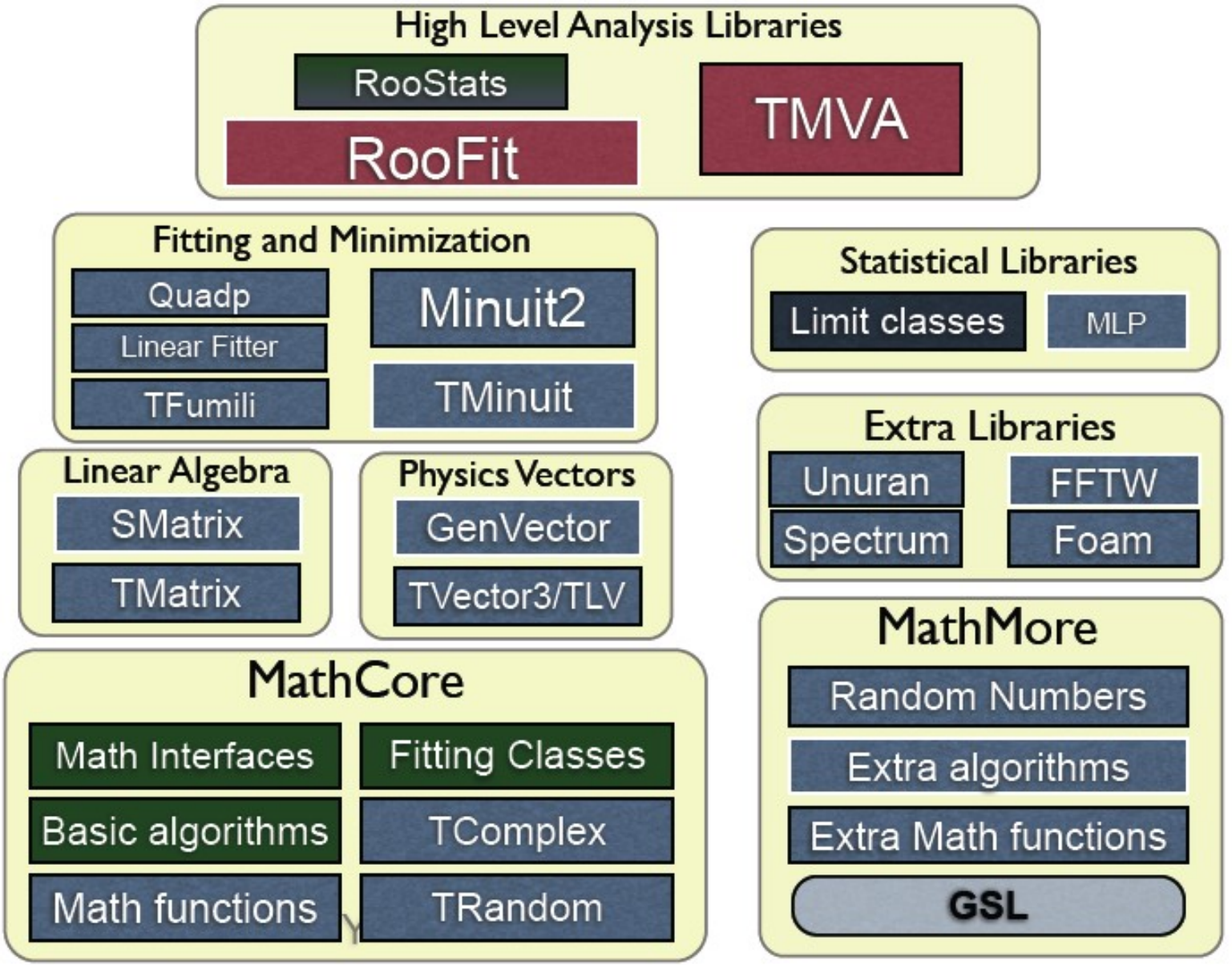
Introduction

XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Are we ready for LHC era experiments?

- The LHC presents an enormous challenge
  - Huge amount of data
  - Small signals with large backgrounds
- Challenge for all 3 tracks !
  - Track 1: Computing Technology for Physics Research
  - Track 2: Data Analysis - Algorithms and Tools
  - Track 3: Methodology of Computations in Theoretical Physics
- Challenge spawns many activities, research and innovation in a vast field

Structure of ROOT Math & Statistical Libraries:

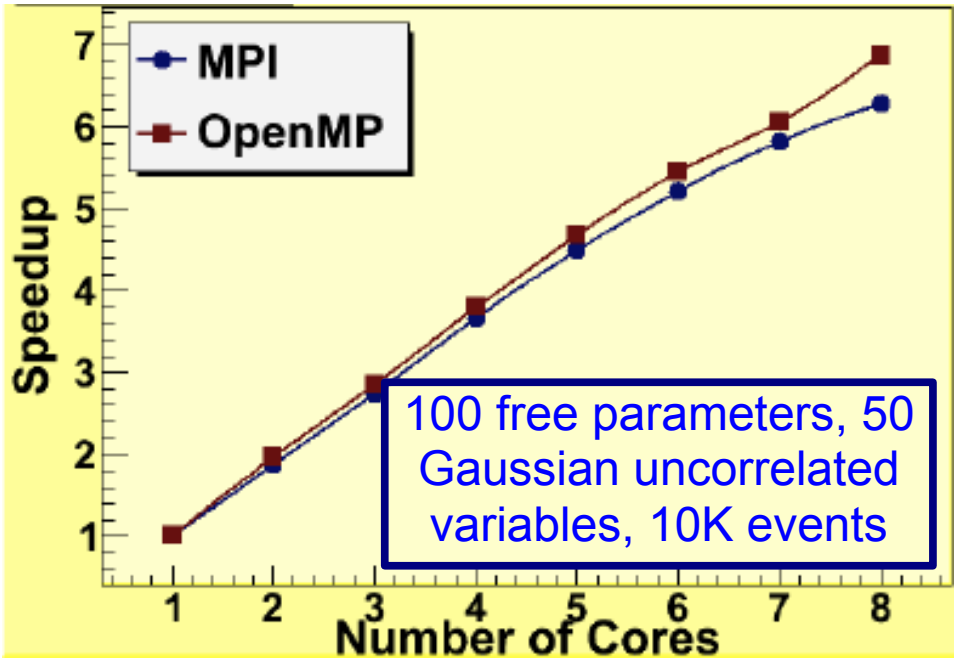


L. Moneta: Improvements of the ROOT Fitting and Minimization Classes

- Large collection of math and statistical tools available in ROOT
  - Large number of improvements, too many to list alll here!
    - Improvement of basic functionality of MathCore (e.g. numerical algorithms)
    - Improved fitting and minimization classes
    - RooStats: New statistical framework for discovery, confidence levels and result combination
  - Improved overall quality and better usability (e.g. GUI FitPanel)
  - Improved modularity and removal of duplications
- Considerable efforts from external contributors in developing tools for physics analysis
  - RooFit: Complex fitting
  - TMVA: Multivariate analysis
  - RooStats
- Large effort on improving validation and test suites
- Good documentation for reference and maintainability

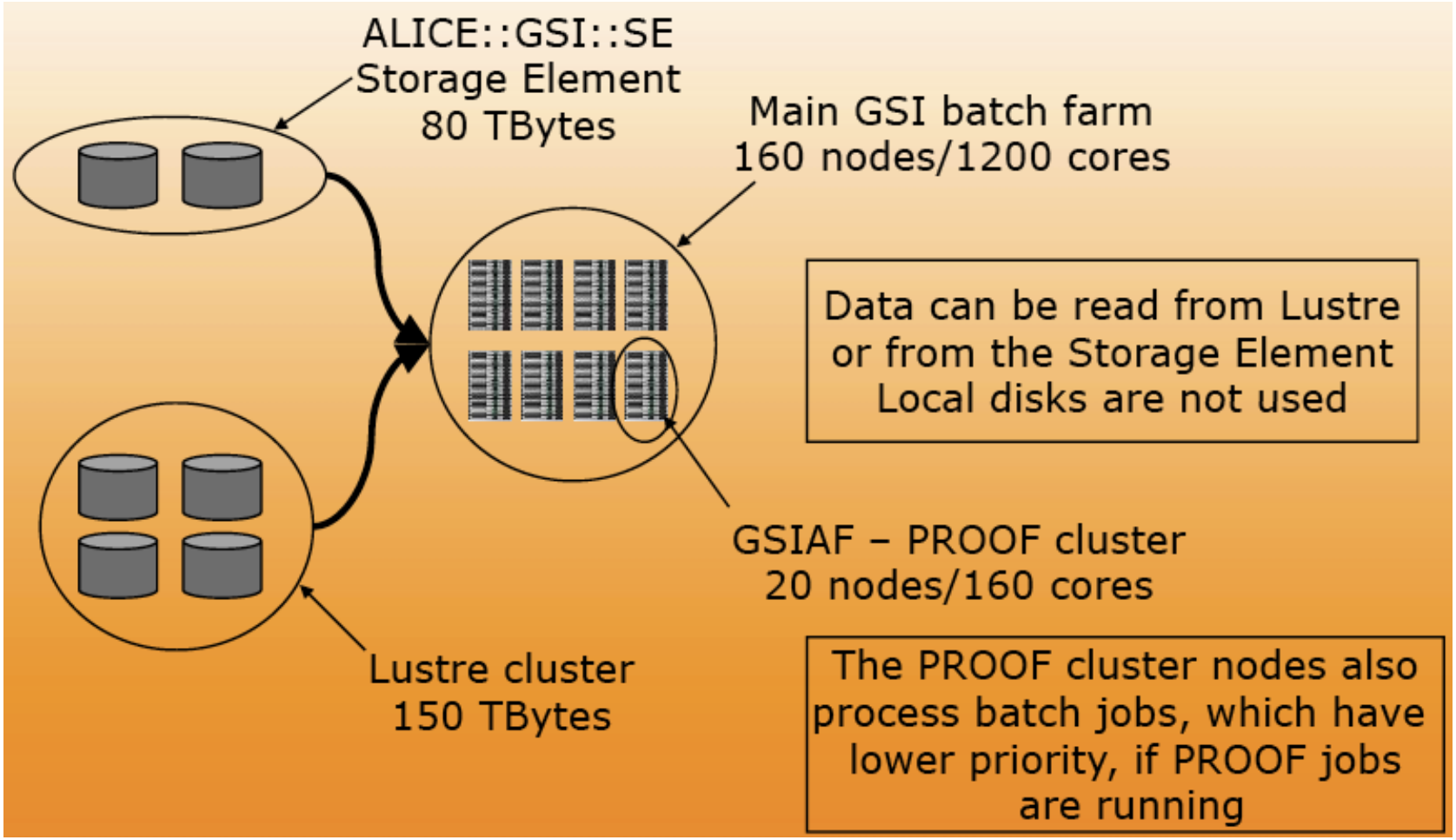
A. Lazzro: MINUIT Package Parallelization

- Large datasets are expected at the LHC
- Many data analysis techniques are very CPU-intensive
  - Optimization problems, some with many free parameters
- Parallelize algorithms to benefit from new multi/many core architectures or clusters
- Parallelization of Maximum Likelihood Fits in RooFit & Minuit2
  - RooFit implements the possibility to split the likelihood calculation over different threads
    - Likelihood calculation is done on sub-samples and the results are collected and summed
  - Split the derivative calculation over several OpenMP/MPI processes
  - Minuit2: parallelization of the minimization and log-likelihood calculation



A. Kreshuk: Interactive Data Analysis with PROOF at GSI

- Interactive data analysis for ALICE using PROOF at tier-2
- Possibility to switch dynamically resources between gird-jobs, jobs from the local batch system and the GSI Analysis Facility (GSIAF)
- Creating PROOF clusters on demand

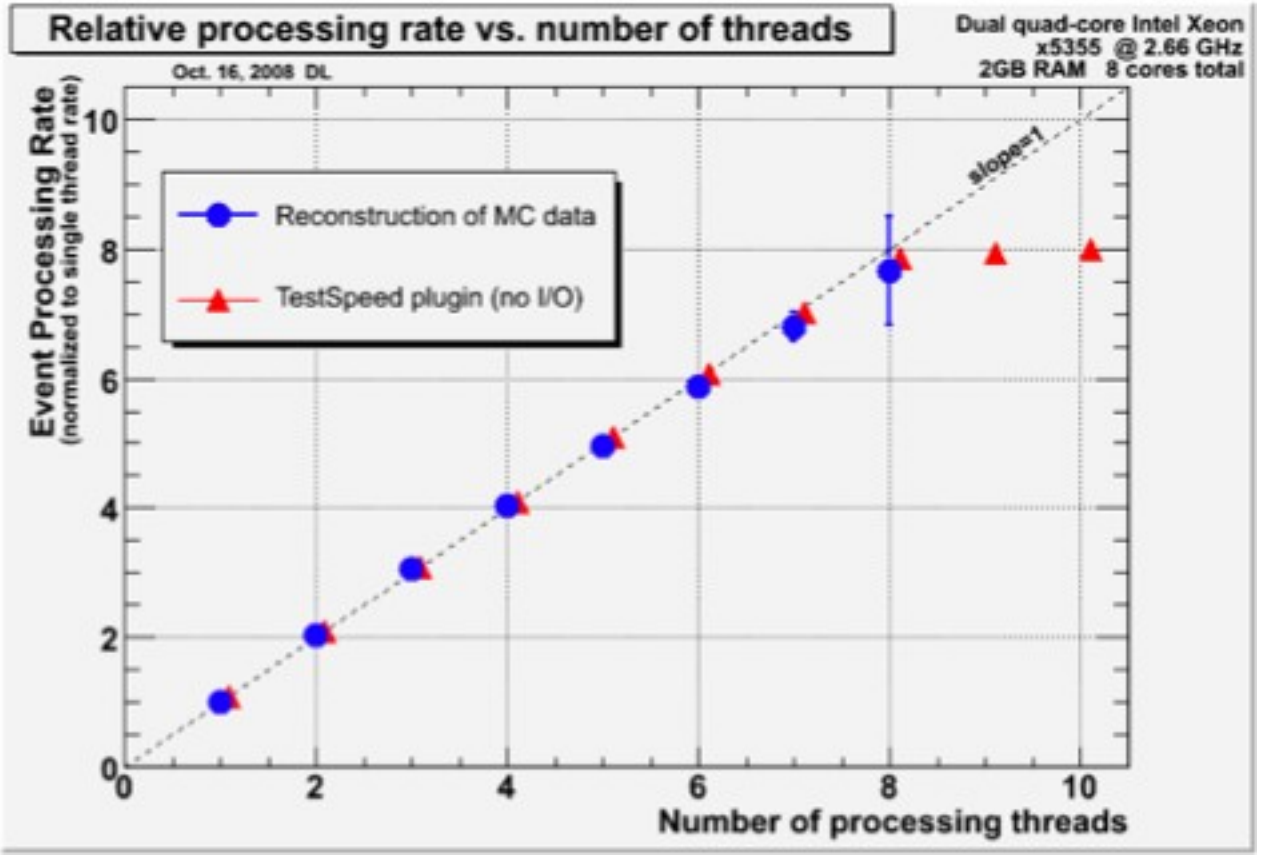


D.Lawrence: Multi-threaded Event Processing with JANA

GlueX experiment on the 6 GeV electron accelerator at jefferson Lab

The Event Processing Framework JANA includes the following features:

- C++ , object-oriented, STL
- Plug-ins
  - Reconstruction Algorithms
  - Event (Data) sources
  - Event Processors
- Multi-threaded
- Data on demand



Multi-variate analysis methods

- Excellent plenary presentation by Harrison Prosper (→ link)
- MVA methods gaining in acceptance in the last years
  - Several high-profile results published
  - Get much attention at the LHC experiments since 2 / 3 years
  - Experiments want to get the most out of their data ! (and be the first...)

Multi-variate analysis methods

- Excellent plenary presentation by Harrison Prosper (→ link)
- MVA methods gaining in acceptance in the last years
  - Several high-profile results published
  - Get much attention at the LHC experiments since 2 / 3 years
  - Experiments want to get the most out of their data ! (and be the first...)
- Several packages implementing MVAs available:
  - Standardized implementations
  - Common platform & interface for all MVA classifiers
  - Consistent training, testing and evaluation of the MVAs
  - TMVA, SPR, WEKA, R

Multi-variate analysis methods

- Excellent plenary presentation by Harrison Prosper (→ link)
- MVA methods gaining in acceptance in the last years
  - Several high-profile results published
  - Get much attention at the LHC experiments since 2 / 3 years
  - Experiments want to get the most out of their data ! (and be the first...)
- Several packages implementing MVAs available:
  - Standardized implementations
  - Common platform & interface for all MVA classifiers
  - Consistent training, testing and evaluation of the MVAs
  - TMVA, SPR, WEKA, R
- This year, less emphasis in presentations on new classifiers:
  - RIPPER
  - PDE-FOAM
- Presentations on examples of usage and comparison with. other methods

J.Stelzer: TMVA

- Implementation of many MVA methods:
  - Integrated in ROOT:
  - Open-source: http://tmva.sf.net
  - Common platform / interface for all MVA classifiers
  - Common data pre-processing capabilities
  - Provide common data input and analysis framework (ROOT scripts)
  - Train and test all classifiers on same data sample and evaluate consistently
  - Classifier application w/ and w/o ROOT, through macros, C++ executables or python
- TMVA 4 to come with new features
  - Current stable TMVA version 3.9.5 for ROOT 5.22 (middle of December)

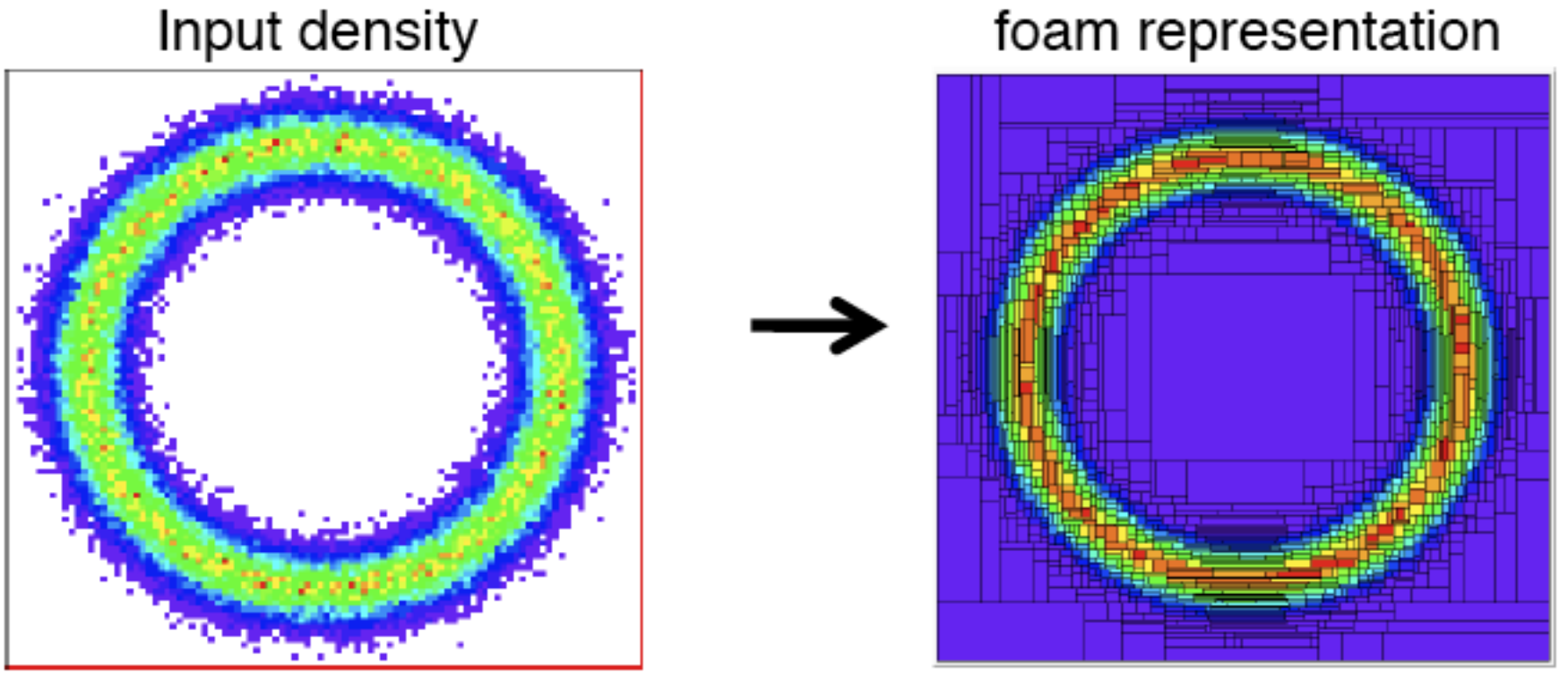
J.Stelzer: TMVA

New developments foreseen:

- Data Regression
- Categorization: multi-class classification
- Automated classifier tuning: using cross-validation method
- Generic boost or bag of any classifiers
- Composite classifiers (parallel training in different phase space regions)
- Input data handling
  - Arbitrary combination of dataset transformations possible
  - Changed TMVA framework in handling datasets and classifiers
  - Implemented regression training for most classifiers
  - User interface extended.
- New Method: PDE Foam
- TMVA to be made multi-threaded and run on multi-core architectures

D. Dannheim: PDE-FOAM

- PDE-FOAN: probability-density estimation method based on self-adapting binning method to divide d-dimensional phase space in finite number of hyper rectangles (cells)
- Iterative cell-split algorithm, starting from base cell containing all events
  - Density of distribution is sampled by counting events in a box of fixed size which is moved randomly across the cell
  - Split cell along plane with highest gain in variance reduction

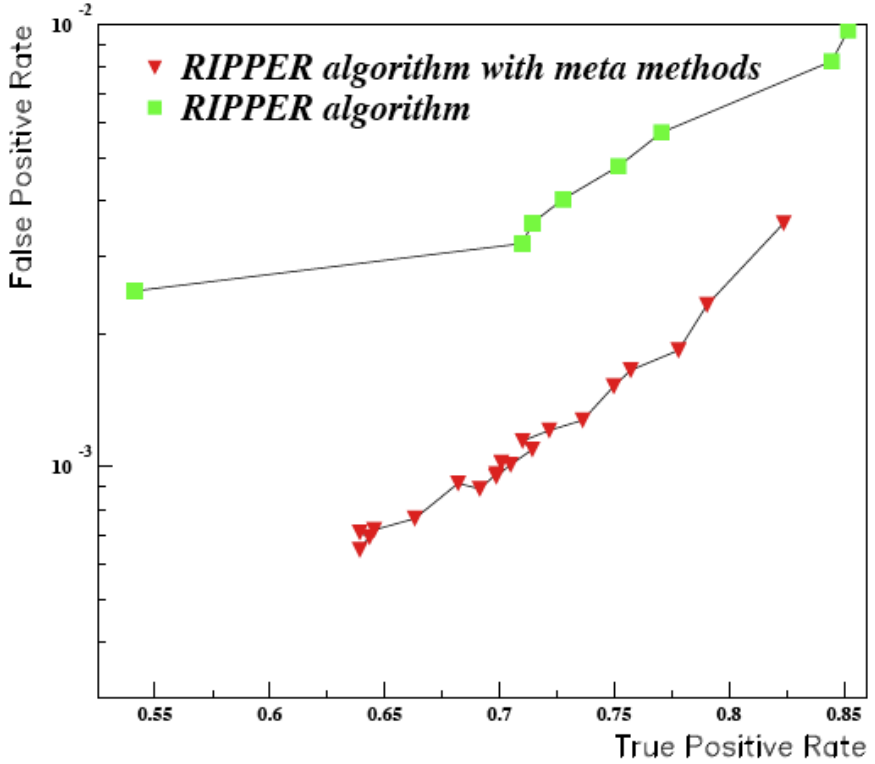


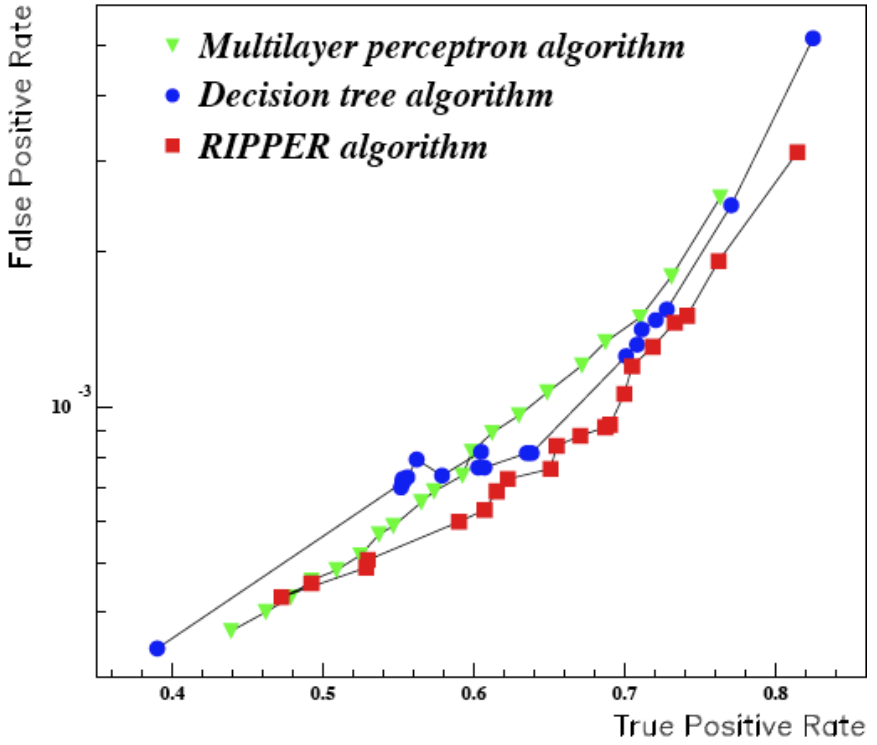
D. Dannheim: PDE-FOAM

- Foam of cells:
  - Few large cells in phase-space regions with constant likelihood density
  - Many small cells in regions with high gradients of likelihood density
- Preserve only binned averaged density information after training phase
  - Fast, memory efficient classification, independent of training sample size
  - Reduces sensitivity to statistical fluctuations for small training samples
- Implemented in TMVA
- Compared performance for various toy models and parameter settings
  - Performance exceeds PDE-RS for small training samples
  - Reduced classification time, independent of size of training sample
  - Reduced memory consumption, independent of size of training sample
- Adapted PDE-Foam for reconstruction of event variables

M. Britsch: RIPPER

- RIPPER:
  - Direct rule based classifier
  - Instance weighting: automated sampling/weighting of instances according to cost
  - Bagging
- Example: LHCb MC $\Lambda \rightarrow p^+ + \pi^-$





Examples of application at the LHC

A few examples of application were shown:

- Tau-tagging:
  - mostly 1-prong or 3-prong decays
  - Well-collimated calorimeter cluster
  - Small number of associated charged tracks
  - Displaced secondary vertex
- B-tagging:
  - Lifetime: τ ~1.5 ps ($c\tau$ ~450μm) $\rightarrow$ p = 20 GeV/c, decay length ~ 1.8 mm
    - secondary (tertiary) decay vertex; displaced tracks with large IP
  - High mass (~5.2 GeV) and decay multiplicity (~5 charged tracks)
  - Decay kinematics (e.g. rapidities)
  - Hard b-quark fragmentation function:
    - decay products with larger p_T relative to b hadron flight direction
  - Semi-leptonic decays (per lepton flavour: BR ~ 11%, ~20% incl. cascade)
- Large number of variable (8-10), non with sufficient separation power

M. Wolter: Tau identification using MVAm in ATLAS

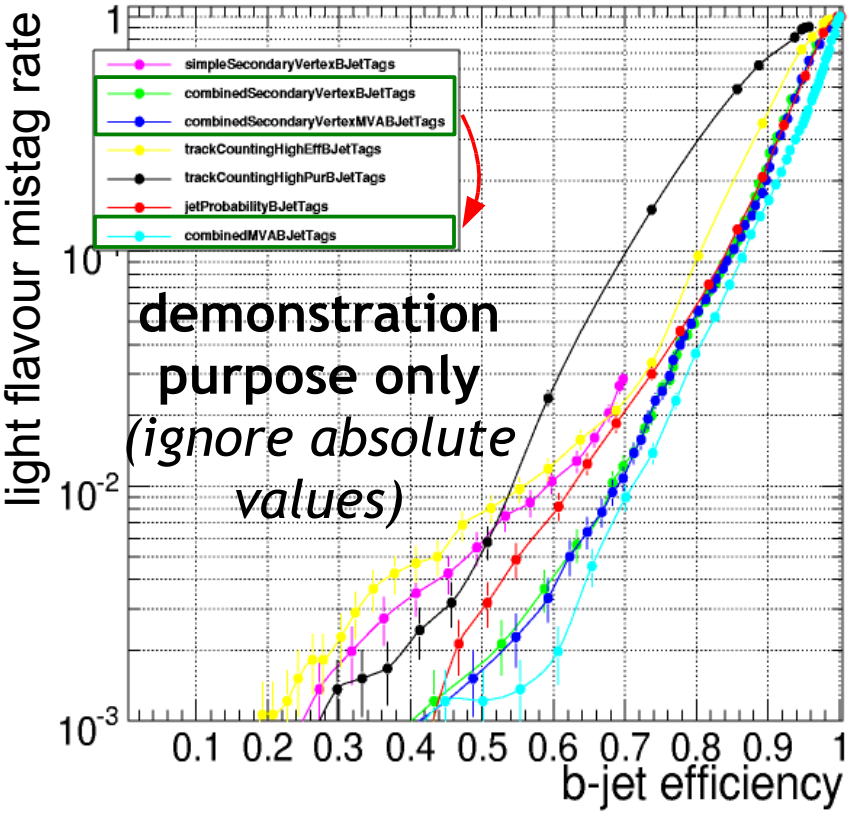
Several methods tried:

- Cut analysis:
  - baseline method, fast, robust, transparent
- Projected Likelihood:
  - popular and well performing tool
- PDE-RS – Probability Density Estimator with Range Searches:
  - robust and efficient, but large samples of reference candidates needed
- Neural Network – Stuttgart Neural Network Simulator (SNNS), feed-forward NN with two hidden layers:
  - Very fast, low memory consumption when trained network is converted to the C code
- Boosted Decision Tree:
  - fast and simple training, insensitive to outliers, good performance.
- For classification problems no single “best” method exists!
- What matters is also simplicity and speed of learning and fast and robust classification!

C.Saout: *b*-Tagging Algorithms in the CMS

Several different algorithms implemented, with different characteristics:

- Track Counting: sort tracks by descending Signed IP Significances (3D)
  - robust, simple and fast → suitable for HLT
- Jet probability: total probability that all tracks originate from PV
- Soft-lepton tagger, using electrons or muons
  - less reliance on IP and on tracker, fast, suitable for HLT
- Simple Secondary Vertex:
  - significance of flight distance
  - Robust wrt to misalignment
- Combined Secondary Vertex:
  - Using all variables in likelihood or NN
  - Highest performance, but more complex

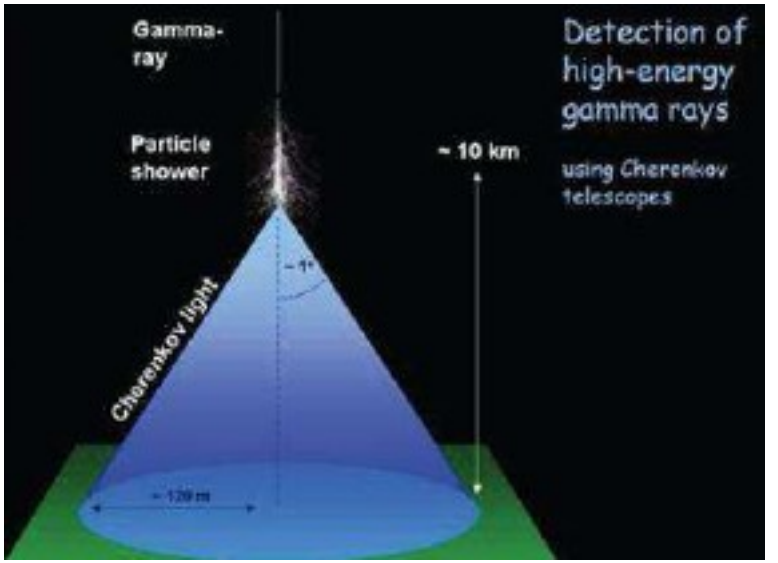


Examples of application at the LHC

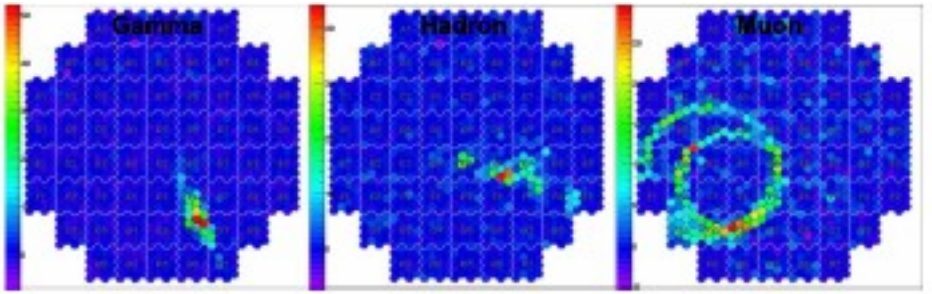
- Many algorithms developed
  - Different characteristics, for different uses (e.g. HLT)
- Multivariate methods do show improvements
  - Further gain from MVA expected to be small
  - Complex algorithms
- But this is all Monte Carlo!
- Concentrate on preparation for real data:
  - Robust,simple algorithms needed first (mis-alignement, etc)
  - Validation of input data
  - Comparison of Monte Carlo with data
  - Measurement of efficiency and mistag rates from data (see poster by V.Bazterra “Strategies for btagging calibration using data at CMS”)
  - Calibration of MVA-based algorithms
  - Systematics...

S. Khatchadourian: ANN Level 2 Trigger for HESS

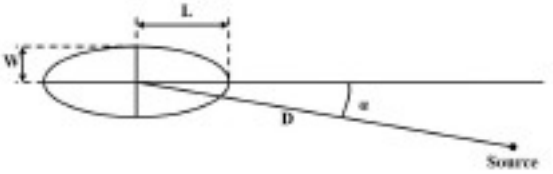
- HESS/HESS 2: Detection of High-energy gamma rays
- Detection of Cherenkov light:
  - As a high energy cosmic ray particle hits the atmosphere it creates an extensive air shower by interaction with the atmosphere.
  - 4 (HESS 2: 5) Cherenkov Telescopes located in Namibia



S. Khatchadourian: ANN Level 2 Trigger for HESS



- Large amount of data
  - Trigger system to minimize the data flo
- Two level trigger system:
  - L1 : eliminate the NSB (thresholds, etc)
  - L2 : classify particles (γ, μ, Proton): pattern recognition
- Two approaches for L2 :



  - Classical method: Filter based on Hillas parameters
  - Neural System: Composed of 2 stages
    - Preprocessor: Describe image with small number of parameters (Zernike momentes)
    - Neural classifier: Multi-layer Perceptron
  - NN: Issue with timing: to be implemented in FPGA

	Gamma	Muon	Proton
Hillas filter	60%	56%	37%
NN with Hillas	71%	84%	82%
NN with Zernike	95%	58%	41%

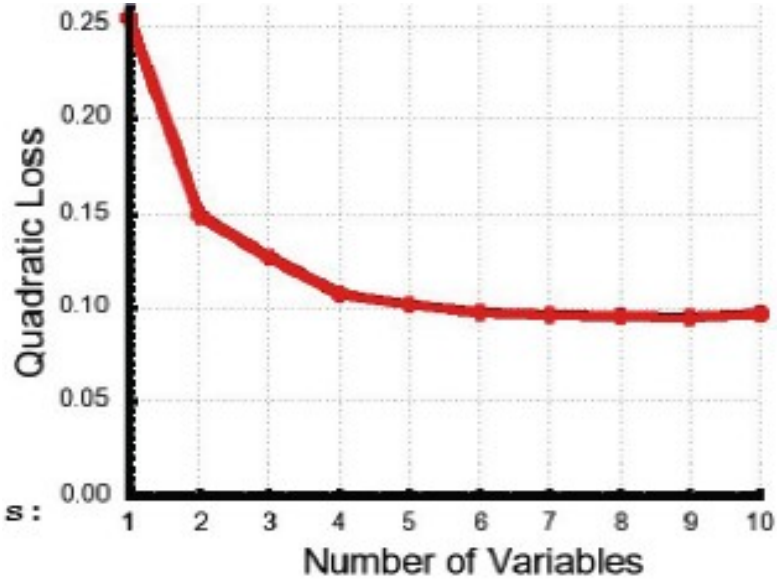
Gamma recognized	Hadron recognized
76%	80%
85%	70%

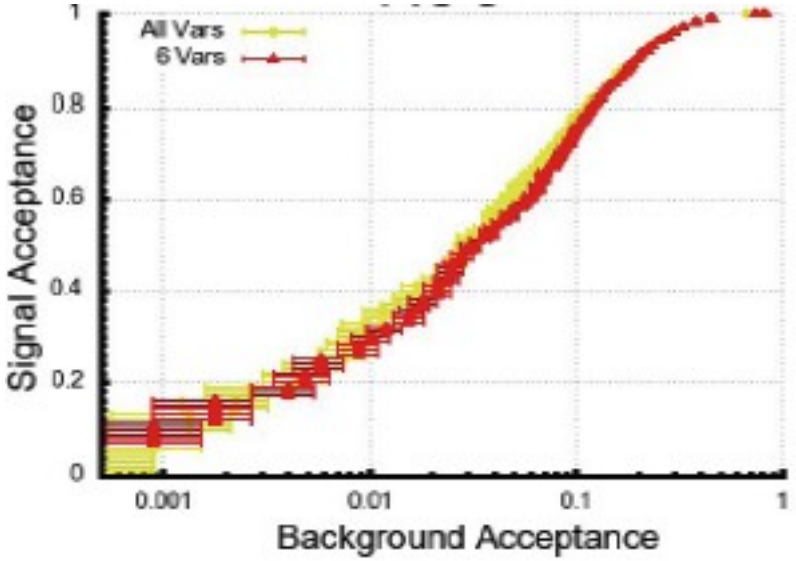
G. Palombo: Feature Selection Algorithms for Supervised Learning

- In same applications, large number of variables, huge number of events
  - Usually, not all the variables are useful for a classification problem
  - Selection of the most powerful discriminating features needed
- Feature Selection: reduce the feature-set to the smallest subset that gives the same or better quality of separation between signal and background as the full set does
  - Independent of classifier
  - Reduces data analysis time.
  - Easier analysis and interpretation of the results
  - For some algorithms: Can improve performance
- Several FS algorithms presented
  - *Add n Remove r*: Start with a null set. Add *s* variables at a time. For each added variable, train the classifier and compute FOM. Choose the variables that improve FOM most. Then remove *r* variables that improve FOM least. Continue as long as it is possible to improve FOM.
- Comparison on several datasets shown

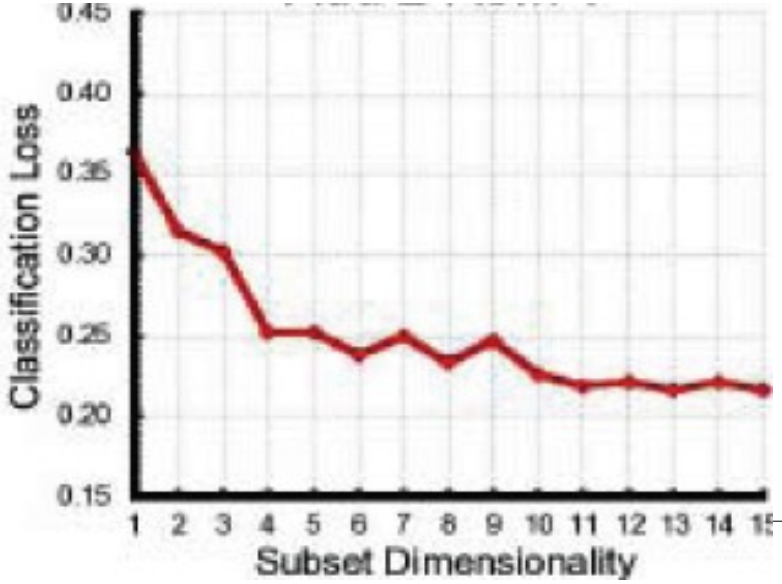
G. Palombo: Feature Selection Algorithms

- Magic Gamma-ray Telescope data: 10 variables – 4 irrelevant





- Arrhythmia dataset: Multi-class classification problem–12 classes
  - 261 variables, 250 irrelevant!

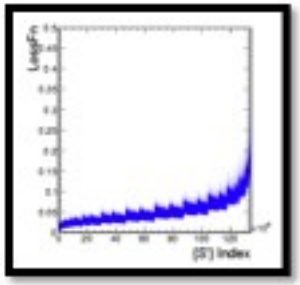


Features set	accuracy (%)	p-value
Best 11 Add2Rem1	80.95	
Best 11 Add1Rem0	76.19	0.49 (-)
Best 3 Add2Rem1	71.90	0.04 (-)
Best 4 Add2Rem1	75.24	0.22 (-)
All 261 SPR	75.95	0.10 (-)

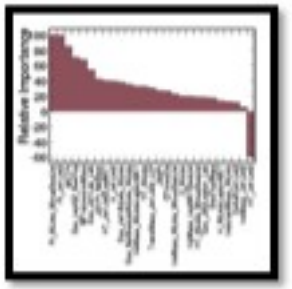
S. Gleyzer: PARADIGM

- PARADIGM: Decision Making Framework for HEP
  - Tool that helps make decisions based on critical information
  - Classifier independent, provides a way to compare and select an optimal choice for a particular analysis
- Variable Selection and Reduction
- Classifier Relevance, Performance & Improvement
- Analysis Optimization

Global Loss Function



Relative Variable Importance

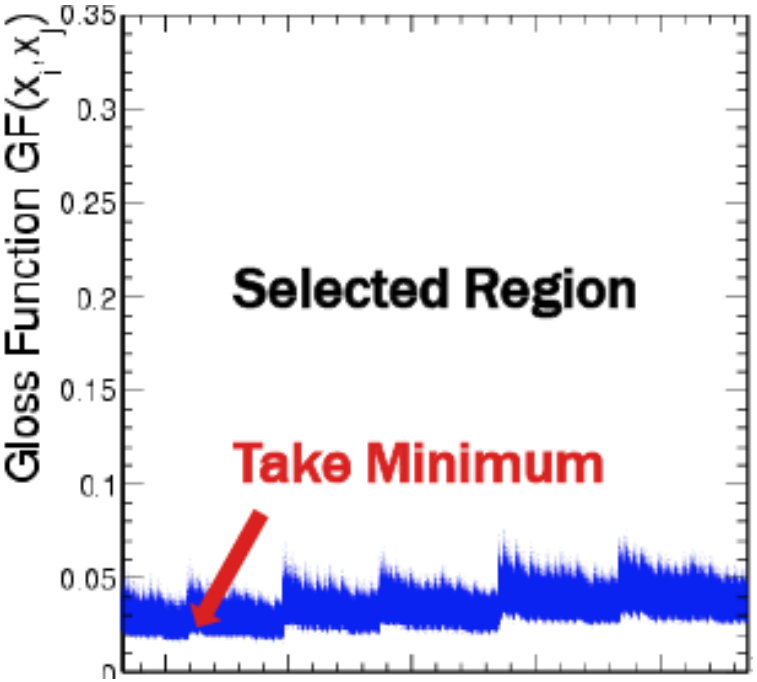


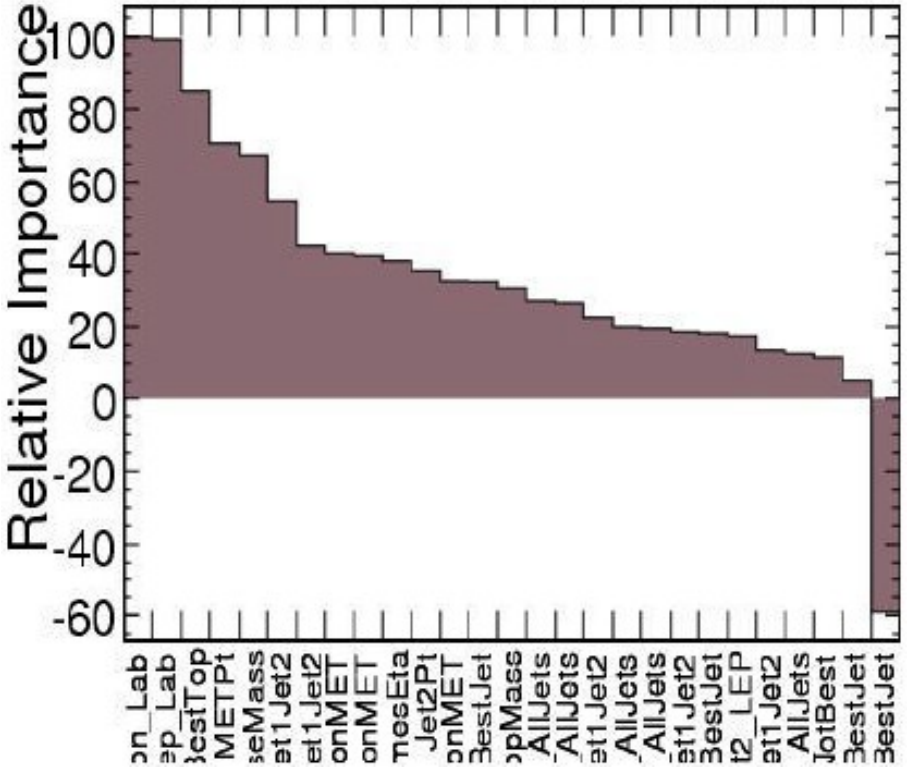
Reduce Variables
Choose Classifier

Improve Classifier

S. Gleyzer: PARADIGM

- Global Loss Function
  - Variable Reduction, Classifier Selection
  - Global measure of loss: global optimization of predictive performance
  - Select variable subsets for global removal
  - Select optimal classifier for given problem: minimum area under the curve





- Relative Variable Importance
  - Variable Selection, Classifier Improvement
  - Measures whether variables are desirable or valuable relative to other variables for a particular problem

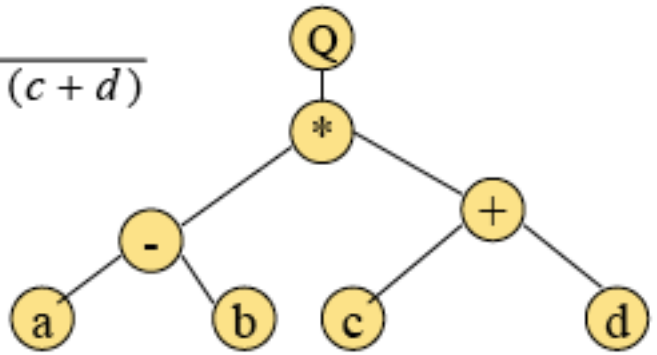
L. Teodorescu: Enhanced Gene Expression Programming

- Evolutionary computation simulates the natural evolution on a computer
  - Generate a population of individuals with increasing fitness to environment
- GEP: Works with two entities, chromosomes and expression trees

Candidate solution represented by an expression tree (ET)

ET encoded in a chromosome: read ET left - right and top - down

$\sqrt{(a-b)\cdot(c+d)}$



Q*-+abcd

Q means sqrt

Chromosome – has one or more genes of equal length

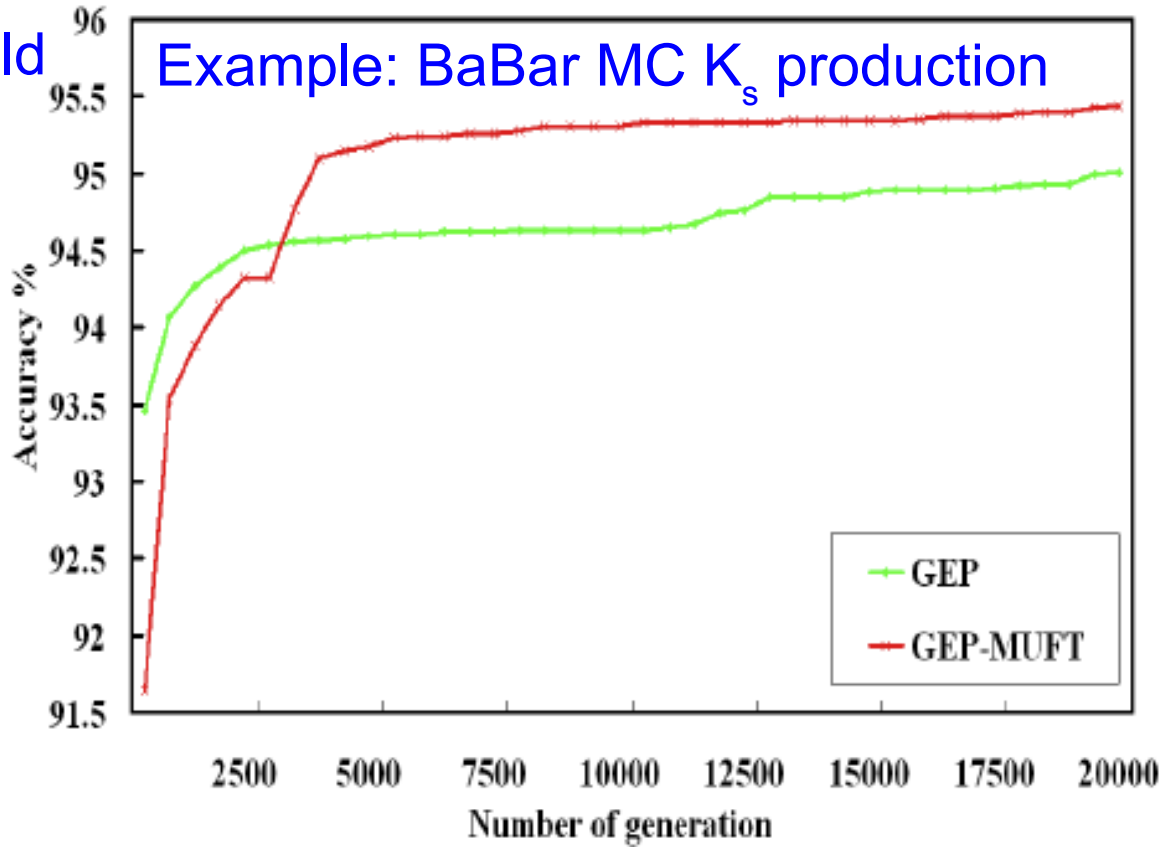
Gene – head: contains both functions and terminals (length h)
- tail: contains only terminals (length t)

- Reproduction: Genetic operators applied on chromosomes
  - Recombination: exchange parts of two chromosomes
  - Mutation: change the value of a node
  - Transposition: move part of chromosome to another location

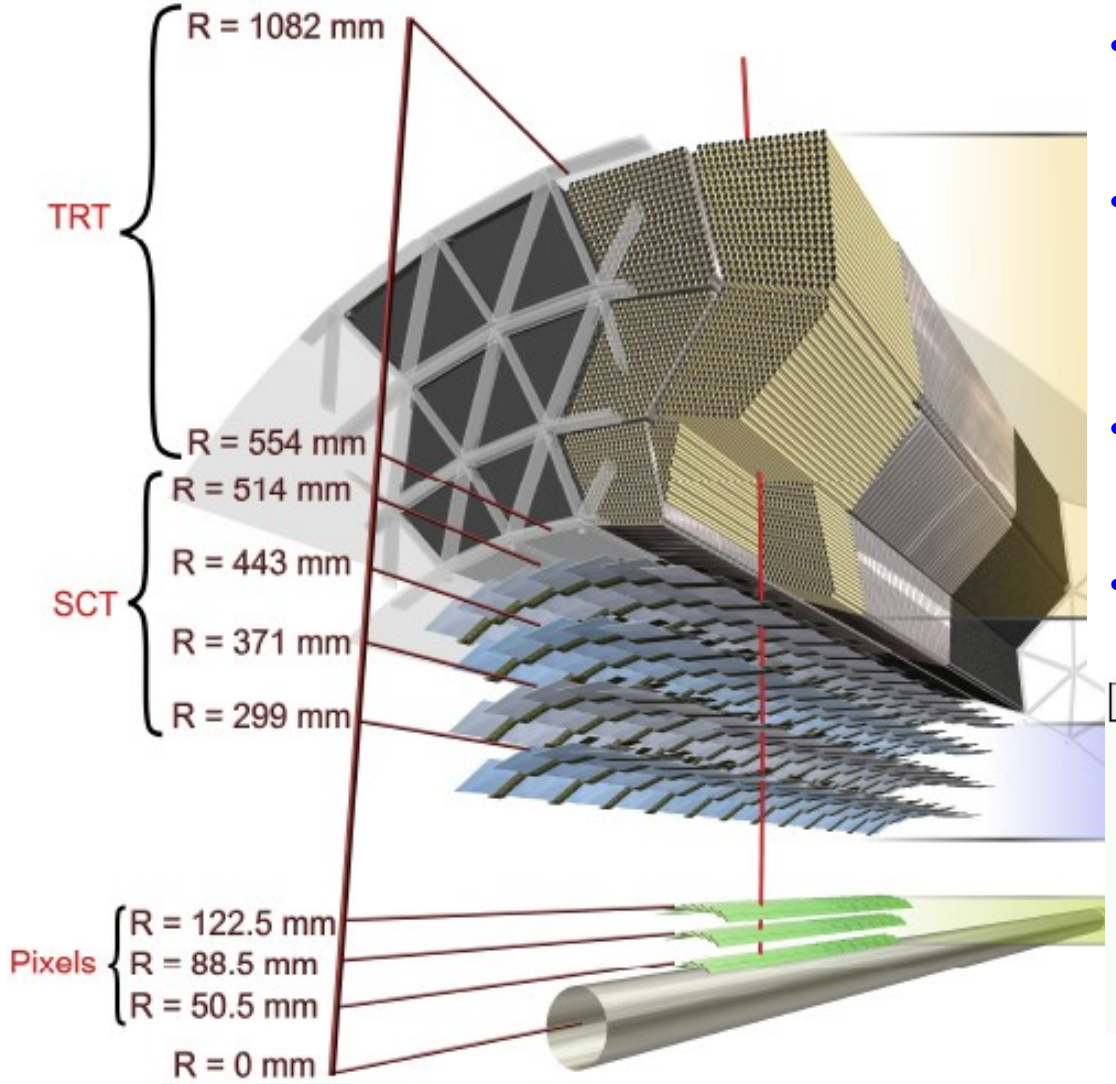
L. Teodorescu: Enhanced Gene Expression Programming

Several new developments since ACAT07:

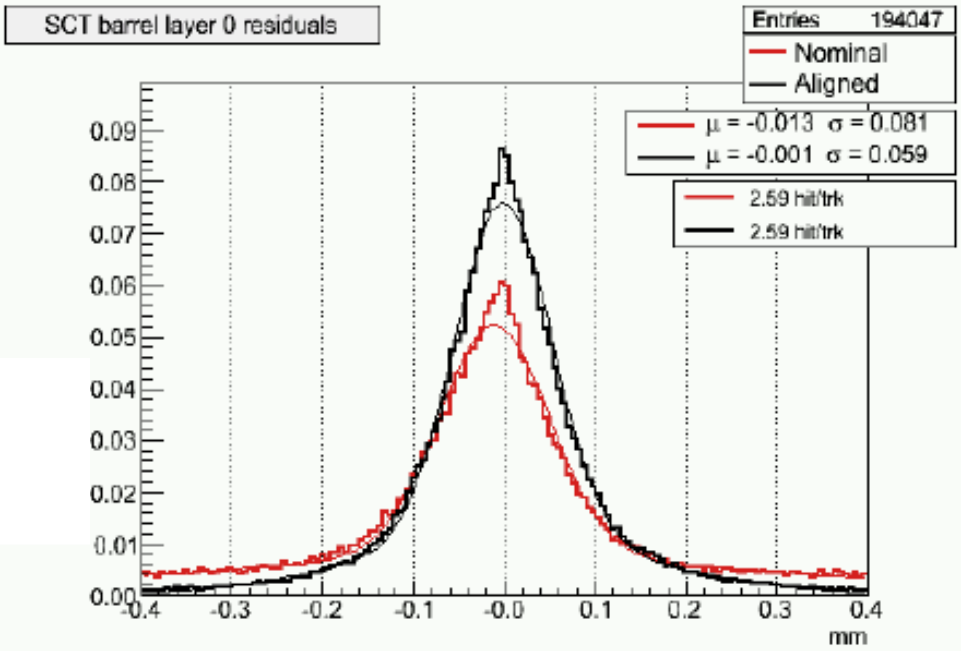
- Different ordering of symbols in chromosome:
  - Keeps the proximity of the genetic material during the translation process → expected lower destructive effect of the genetic operators
- Controlled evolution through fitness threshold
  - Eliminate the weak individuals from the evolution process
- Dynamic classification threshold
  - Threshold value adapted to each individual
- Improvements:
  - earlier convergence
  - slightly higher accuracy



J.Alison: Alignment of the ATLAS Inner Detector



- Trackers of LHC experiments are very complex detectors
- Very sensitive to mis-alignment
  - Directly affects physics measurements
- Comprehensive description of alignment algorithm
- First results with cosmic rays



Conclusion

- Large number of presentations: 26 Presentations
  - And another few which have been presented in track 1 !
  - Very high quality
  - Many posters as well !
- Large number of topics, on many diverse subjects
  - Not all were related to the LHC
  - Sorry for my LHC-bias !
- I was not be able to mention each presentation
  - Subjective choice
  - Apologies for not mentioning yours!