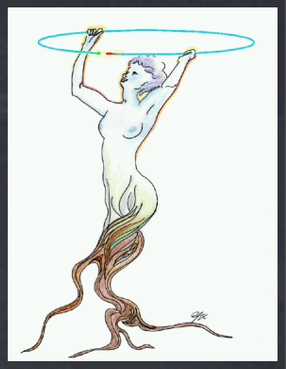


Overview of New ROOT Statistical Software

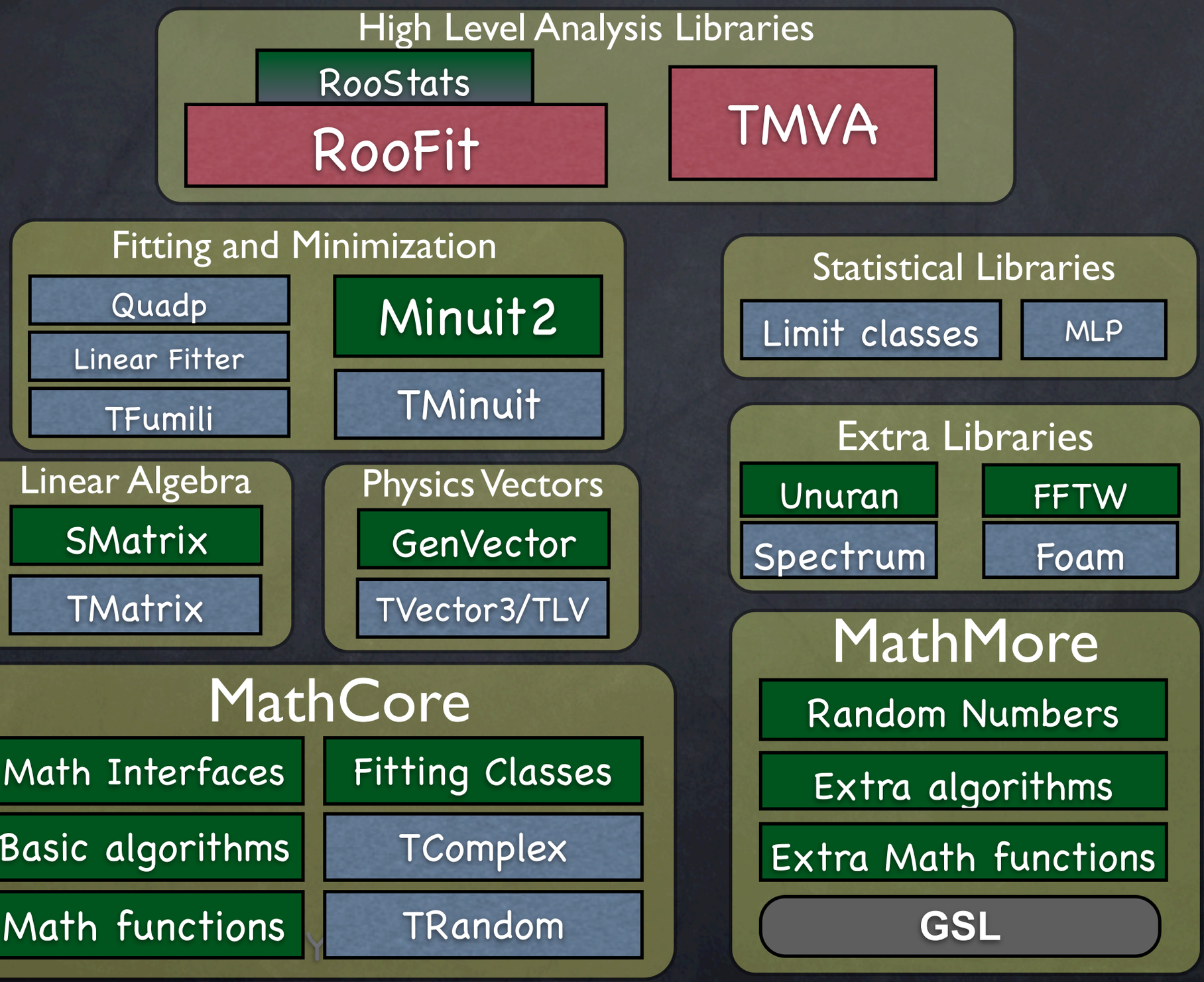
Lorenzo Moneta (CERN- PH/SFT)

ROOT Math: D. Gonzalez Maline, Lorenzo Moneta
RooFit: D. Kirkby, W. Verkerke
RooStats: K. Cranmer, L.M., G. Schott, W. Verkerke
TMVA: A. Höcker, P.Speckmayer, J.Stelzer, H.Voss
and many other contributors

Outline

- Recent developments in core ROOT statistical software
  - new developments in MathCore library
- Recent developments in RooFit
- RooStats
- TMVA

ROOT Math/Stat Libraries



New in MathCore

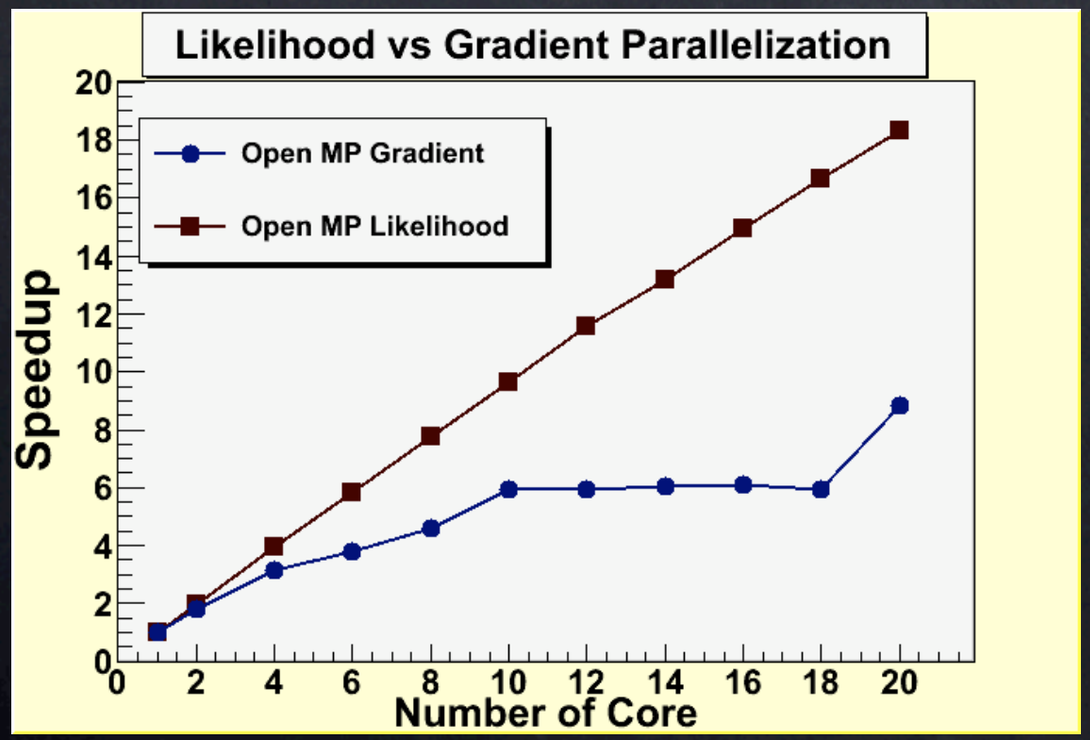
- Improved functions in TMath
  - Implemented mathematical and statistical functions using new set of implementations
  - better numerical precision and in some cases better CPU efficiency
- Released new numerical algorithm classes
  - classes implementing integration, derivation, root finders
  - provide also complementary implementation in MathMore
  - mechanism to use the alternative implementation via the plug-in manager
  - interfaces to function evaluation
- New ROOT fitting classes

Fitting Improvements

- Re-engineered fitting and minimization classes
  - maintained full backward compatibility
- Separate Fitter and Minimization interfaces
  - multiple implementations of Minimizer interface
    - Minuit, Minuit2, Fumili, GSL minimizers
    - possible to use (and mix) the various minimization engines
- Decouple fitting from data source
  - BinData and UnBinData classes
- Fit result object which can be stored and retrieved
  - keep full result information (parameters, errors, covariance matrix)
- Parallel fitting and minimization
- New GUI for fitting (see presentation of D. Gonzalez Maline)

Fitting Parallelization

- Parallelization in Minimization in the gradient calculation (independent of user code)
- Log-likelihood parallelization (splitting the sum)
  - more efficient but more demanding on thread safety of provided code



- OpenMP (multi-threads)
- MPI for multi-process in a cluster

(see poster from A. Lazzaro)

Example: unbinned fit with 20 parameters

RooFit

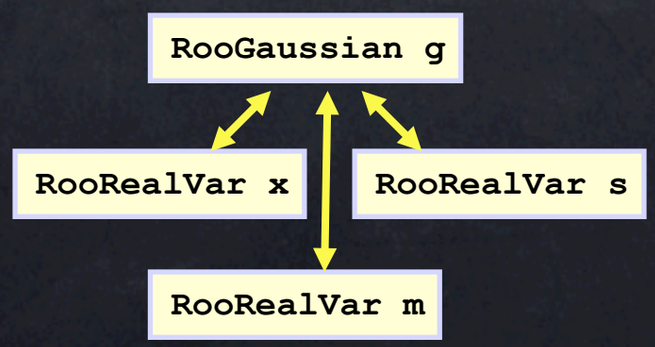
- Toolkit for data modeling (by W. Verkerke and D. Kirkby)
  - model distribution of observable x in terms of parameter p and q
    - probability density function (pdf): $\mathcal{P}(x;p,q)$
- Functionality for building the pdf's
  - complex model building from standard components
- All models provide the functionality for
  - maximum likelihood fitting
  - toy MC generator
  - visualization

RooFit

Mathematical concepts are represented as C++ objects

Mathematical concept		RooFit class
variable	x	`RooRealVar`
function	$f(x)$	`RooAbsReal`
PDF	$f(x)$	`RooAbsPdf`
space point	$\vec{x}$	`RooArgSet`
integral	$\int_{x_{\min}}^{x_{\max}} f(x)dx$	`RooRealIntegral`
list of space points		`RooAbsData`

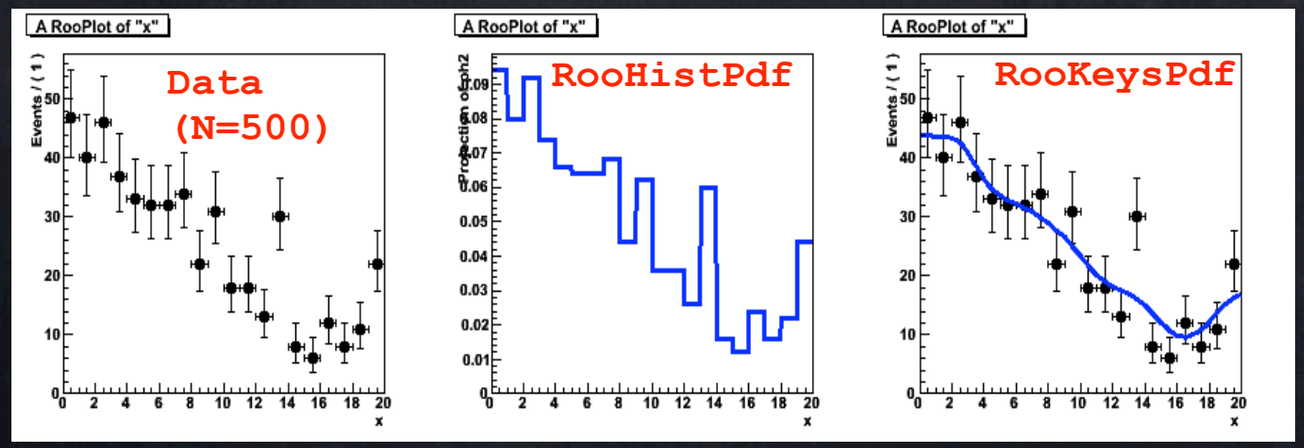
$G(x,m,s)$




```
RooRealVar x("x","x",2,-10,10)
RooRealVar s("s","s",3) ;
RooRealVar m("m","m",0) ;
RooGaussian g("g","g",x,m,s)
```

Model Building

- Complex model building using standard components
- pdf composition using addition, product and convolution
  - standard set of pdf classes
  - user provided pdf (formula or C++ classes)
  - non-parametric pdf (obtained from data)
    - histogram or kernel density estimators (available also for multi-dimensions)



Workspace

- Maintain a complete description of all the model
  - with possibility to save entire model in a ROOT file
- Workspace class in RooFit (`RooWorkSpace`)
  - probability density functions
  - (multiple) data sets
  - model configuration
- All information will be available for further analysis and combination of results
  - full likelihood for bayesian analysis
  - probability density functions for frequentists analysis
- Common format for combining and sharing physics results

Example Combination

ATLAS-CMS combination from persisted workspaces

Read ATLAS workspace

```
TFile* f = new TFile("atlas.root") ;
RooWorkspace *atlas = f->Get("atlas") ;
```

Read CMS workspace

```
TFile* f = new TFile("cms.root") ;
RooWorkspace *cms = f->Get("cms") ;
```

Construct combined LH

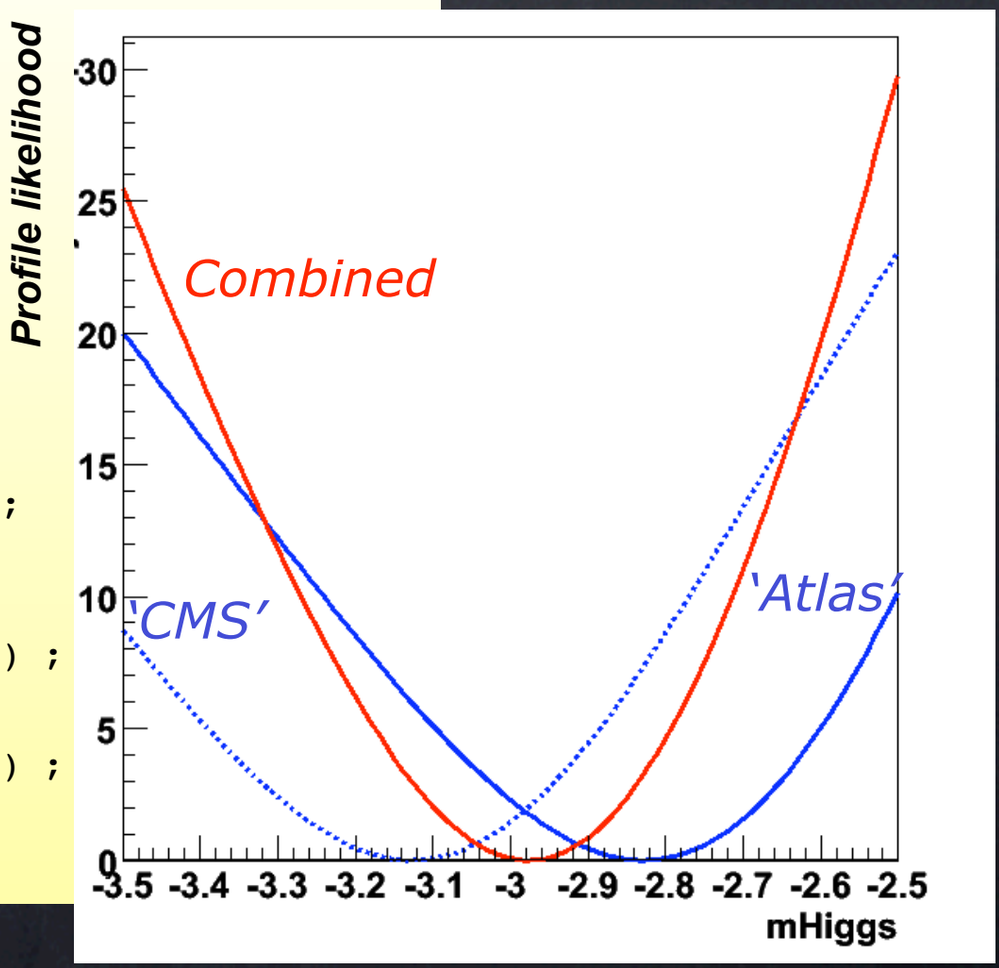
```
RooAddition nllCombi("nllCombi","nll CMS&ATLAS",
            RooArgSet(*cms->function(“nll”),
                      *atlas->function(“nll”))) ;
```

Construct profile LH in mHiggs

```
RooProfileLL pllCombi("pllCombi","pll",
                       nllCombi,*atlas->var("mHiggs")) ;
```

Plot Atlas,CMS, combined profile LH

```
RooPlot* mframe = atlas->var("mHiggs")->frame(-3.5,-2.5) ;
atlas->function(“nll”)->plotOn(mframe)) ;
cms->function(“nll”)->plotOn(mframe),LineStyle(kDashed)) ;
pllCombi.plotOn(mframe,LineColor(kRed)) ;
mframe->Draw() ;
```



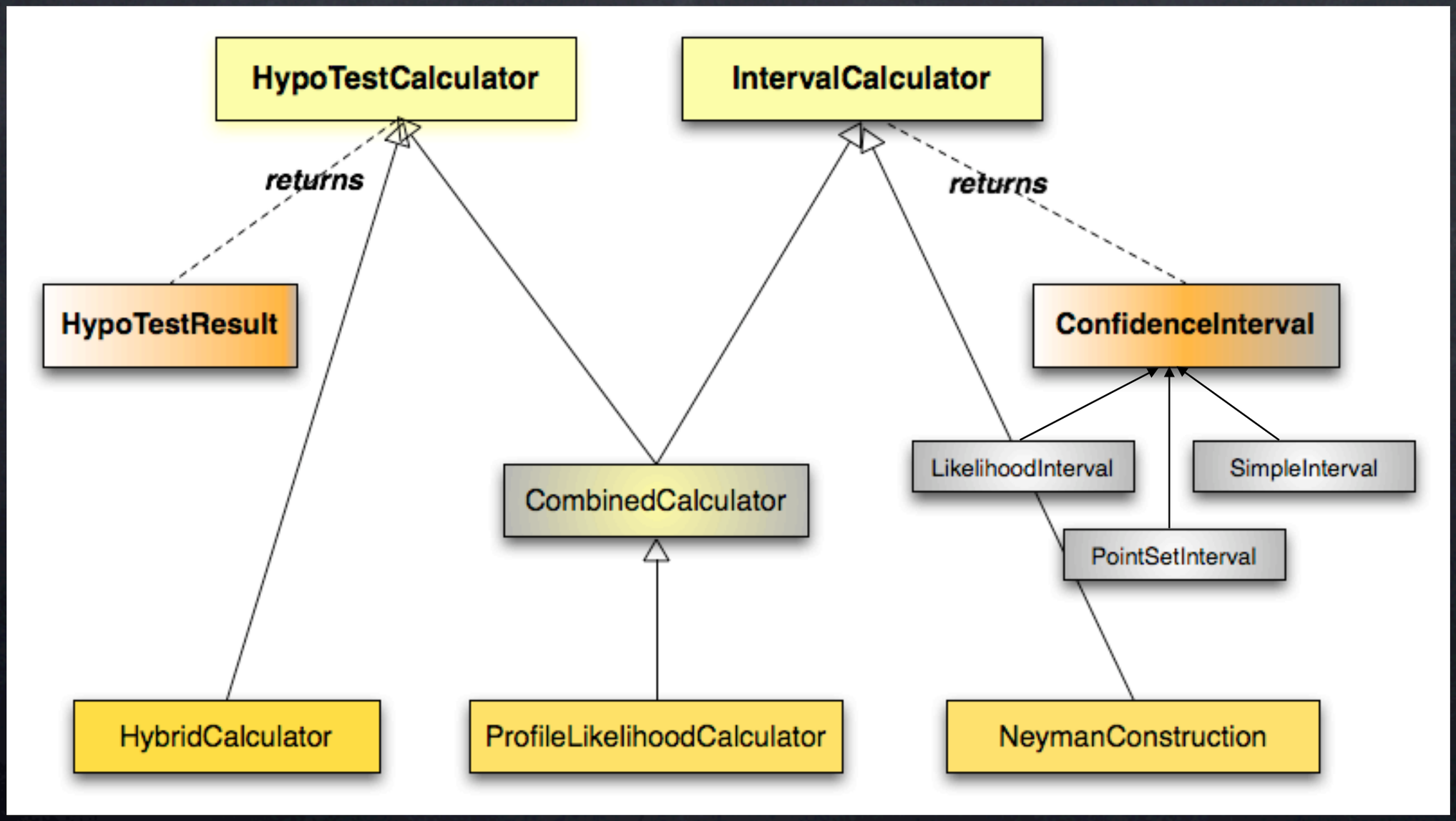
Profile Likelihood result for ATLAS and CMS and combination

RooStats

- New project to create advanced statistical tools needed by LHC experiments
  - built on top of RooFit
    - provides convenient description of models probability density function or likelihood functions
- Common framework for statistical calculations
- Joint contribution between ATLAS, CMS ROOT and RooFit from K.Cranmer (ATLAS), G. Schott (CMS), W. Verkerke (RooFit), Lorenzo Moneta (ROOT) and other contributors (D. Piparo, M. Pelliccioni from CMS)
  - developments monitored by ATLAS and CMS statistical committees

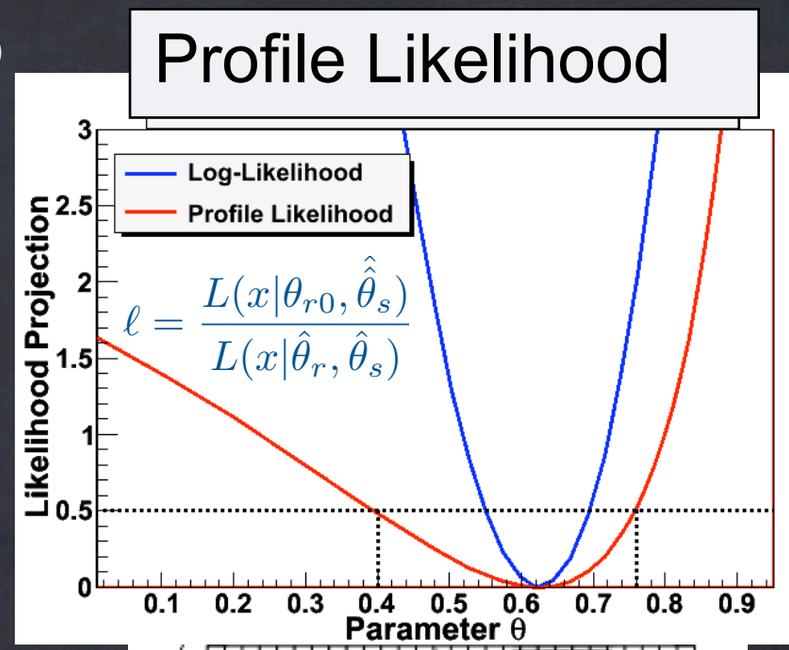
RooStats Interfaces

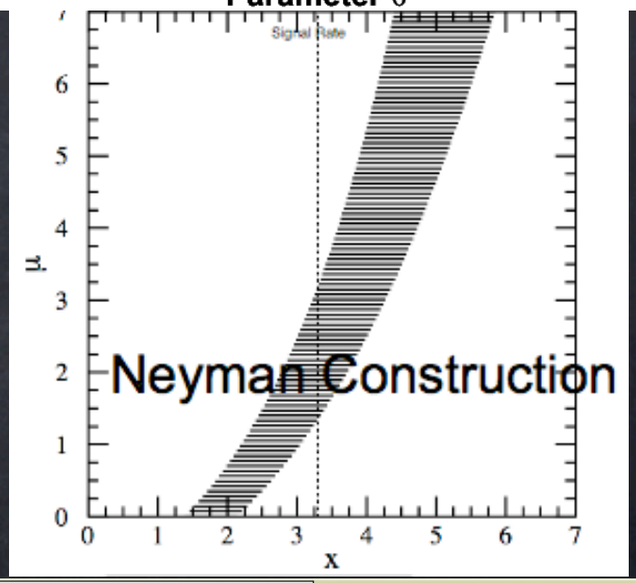
- Common interfaces to statistical tools
  - clear mapping to statistical concepts

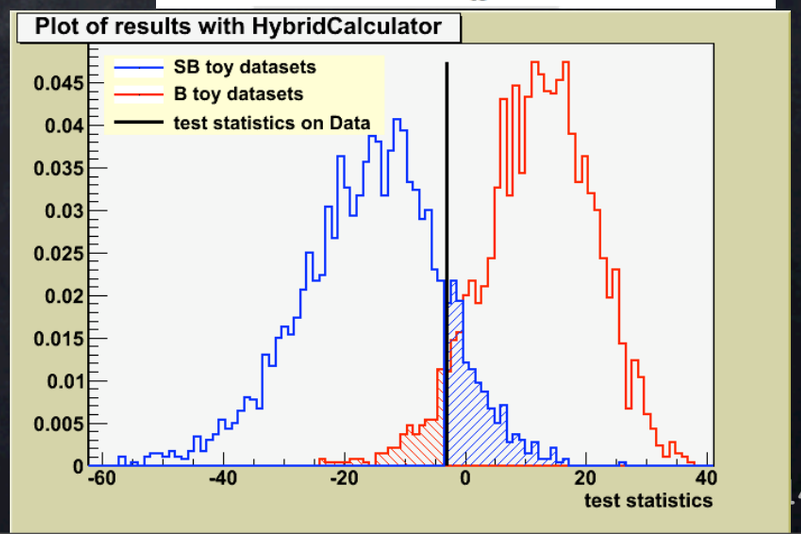


Statistical Methods

- Interval Calculator
  - Profile Likelihood method
    - uses asymptotic distribution
  - Neyman construction
    - full frequentist method
    - various ordering rule possible
- Hypothesis Test Calculator
  - Profile Likelihood
  - Hybrid calculator
    - frequentist method but with bayesian treatment of systematic errors (used at LEP and Tevatron)







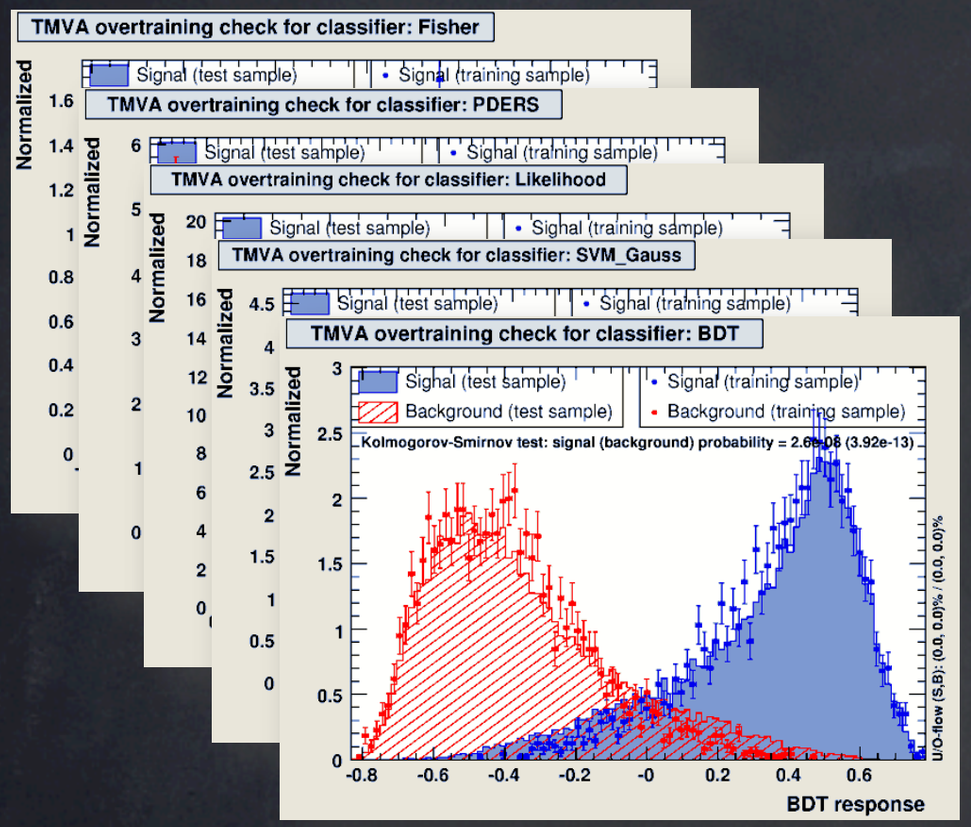
RooStats Status and Plans

- First versions of ProfileLikelihood calculator and HybridCalculator in ROOT 5.22 (December 2008) combining ATLAS and CMS code
- Neyman construction expected for next release (Feldman-Cousins method)
  - set of new classes needed for Neyman (sampling of a test statistics)
- Forseen to integrate Bayesian tools
  - interfacing to BAT package (see next talk of D. Kollar)
- Designing a new user API for model building
  - starting from RooStatsCMS (see poster from D. Piparo)
- Focusing also on parallelization
  - parallel MC sampling (toy generation) and scans

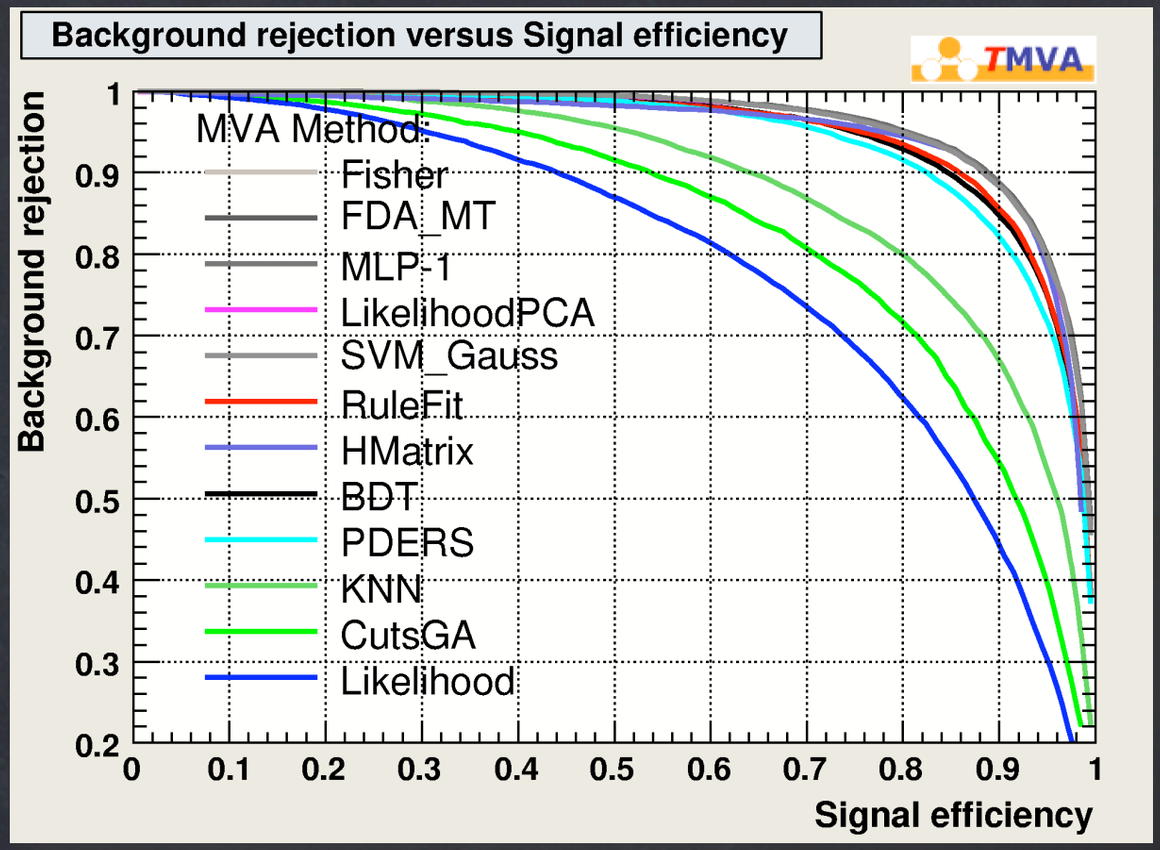
Multi-Variate Analysis

- ROOT provides some classes for MVA analysis
  - Neural Network class (`TMultiLayerPerceptron`)
- TMVA: toolkit for multi-variate analysis distributed within ROOT
  - from *A. Höcker, P. Speckmayer, J. Stelzer, H. Voss and many other contributors*
  - provides currently a large variety of methods for signal/background classification
  - simple and convenient usage
    - pre-processing tools
    - common interfaces to methods

TMVA Classifiers



Classifier output variable



ROC Curve

Scan over classifier output variable creates set of (ε_{sig},ε_{bkgd}) points

- See the TMVA User Guide for classifier training
- New Web page for classifier tuning parameters

TMVA Methods

- Linear Classifiers
  - Cut based
  - Projective likelihood estimator
  - Linear Fisher Discriminant
- Non-Linear Classifiers
  - Neural Network
  - PDE: Range Search, kNN, Foam, multi-dim Likelihood
  - Function Discriminant Analysis
- Modern Classifiers
  - Boosted Decision Trees
  - Support Vector Machine
  - Learning via rule ensembles

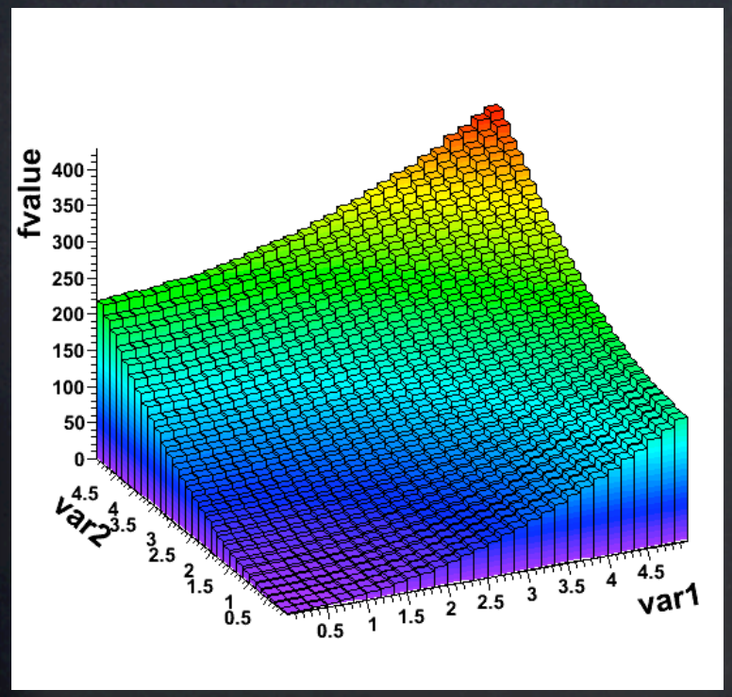
See TMVA User Guide for a detailed description

New in TMVA

- Version 4 is being released with various nice new features:
  - Support for regression in addition to classification
  - generic boost or bag of any classifier
  - new method PDE-FOAM
- Planned for following releases
  - multi-class classification
  - automatic tuning using cross validation
- see the TMVA poster by P. Speckmayer

Regression in TMVA

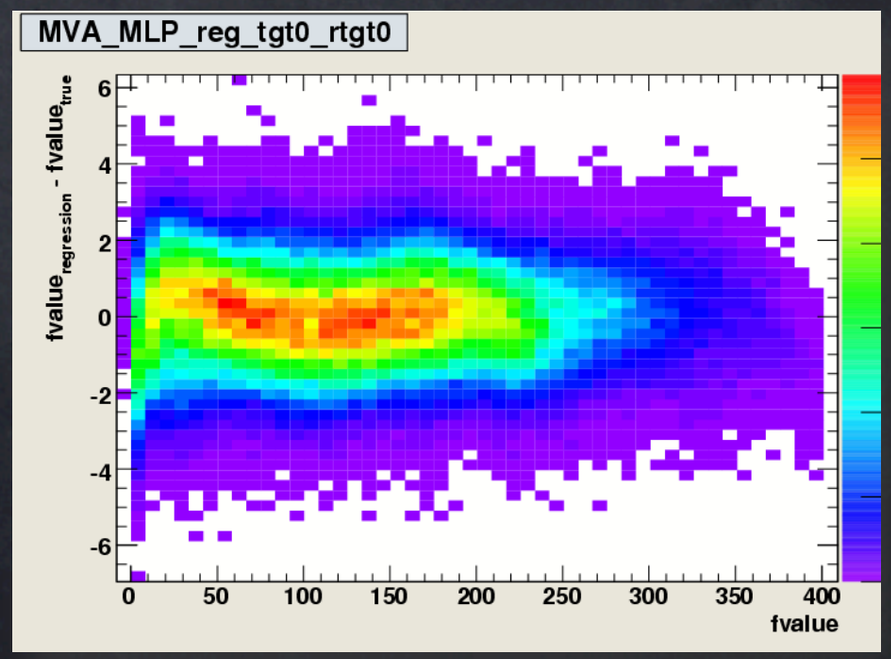
- Use MV methods to predict value of one (or more) dependent variable(s)
  - predict a functional dependence $f(x_1,..,x_N)$
  - different training needed (need a regression target instead of signal & background)



Target as function of 2 variables

Example using MLP

predict $f(x_1,x_2)$



Estimated value - true value as function of true value

Conclusions

- Large collection of statistical tools available in ROOT
- Considerable efforts in developing common tools for physics analysis
  - complex fitting (RooFit)
  - multivariate analysis (TMVA)
  - new statistical framework (RooStat)
- Working on improving the overall quality and better usability
- Important to ensure the correctness of tools we are using
  - large effort on improving validation and test suites
  - good documentation for reference and maintainability

References

- ROOT : new Web site with additional documentation)
  - Reference Guide
  - USER Guides (including TMVA and RooFit)
- RooFit: http://roofit.sourceforge.net/
  - large collections of tutorials available in `$ROOTSYS/tutorials/roofit`
- RooStats: https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome
- TMVA: http://tmva.sourceforge.net/
- ROOT Talk Forum (for support, requests and discussions)
  - a thread for only Math and Statistical topics
- ROOT Savannah for reporting bugs