Managing Many Systematic Uncertainties Simultaneously

Agostino Di Iorio,
Alberto Orso Maria Iorio,
Luca Lista

University of Naples Federico II
& INFN Napoli

Outline

- Short introduction about treatment of systematic uncertainties
- Application in simultaneous template fitting
- Technical implementation issues
- Ideas for possible improvements

Systematics and nuisance parameters

- The dependence of a probabilistic model on sources of systematic uncertainty is modeled via nuisance parameters
- Those parameters may be known from external measurements with some uncertainty
- Data samples can constrain nuisance parameters and reduce the original uncertainties

- Different approaches in Bayesian or frequentist applications, but the resulting effect is similar

Nuisance pars. in Bayesian approach

- Notation: μ = parameter(s) of interest,
  θ = nuisance parameter(s)
  x = data sample

μ is usually the 'signal strength' (i.e.: $\sigma/\sigma_{\text{th}}$) in case of a search for a new (or specific SM) signal

- Posterior probability P of all unknown parameters:

$$P(\mu, \theta | x) = \frac{L(x; \mu, \theta)\pi(\mu, \theta)}{\int L(x; \mu', \theta')\pi(\mu', \theta')\mathrm{d}\mu'\mathrm{d}\theta'}$$

- $P(\mu|x)$ obtained as marginal PDF of μ by integration over nuisance parameters θ:

$$P(\mu|x) = \int P(\mu, \theta | x)\mathrm{d}\theta = \frac{\int L(x; \mu, \theta)\pi(\mu, \theta)\,\mathrm{d}\theta}{\int L(x; \mu', \theta')\pi(\mu', \theta')\mathrm{d}\mu'\mathrm{d}\theta'}$$

Profile likelihood (frequentist)

- Test statistic based on a likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \quad \begin{array}{l} \longleftarrow \text{ Fix } \mu\text{, fit } \theta \\[1.5em] \longleftarrow \text{ Fit both } \mu \text{ and } \theta \end{array}$$

- Different ‘flavors’ of test statistics exist
  - E.g.: deal with unphysical $\mu < 0$, etc. …
- The distribution of $q_\mu = -2 \ln \lambda(\mu)$ is used to determine the signal parameter μ and/or set upper limits to new signal
- The distribution of the test statistic for $\mu=0$ may be asymptotically approximated to a χ^2 with one degree of freedom (for one parameter of interest = μ)
  - Wilks’ theorem and other properties

Simultaneous fits

- A complementary dataset, or control sample, y, is used to constrain nuisance parameters θ
  - Calibration data, background estimates from independent data samples, …
- Statistical problem formulated in terms of both the main data sample (x) and the control sample (y) assumed statistically independent

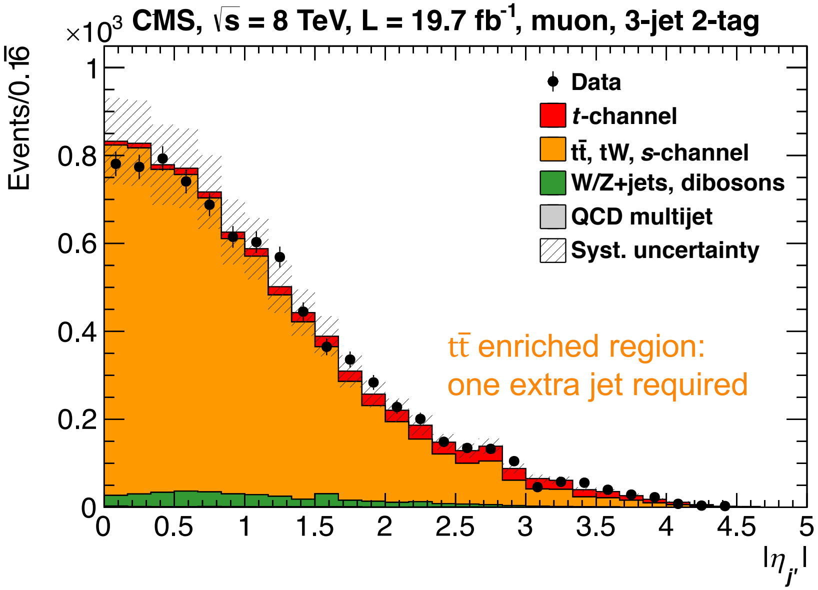
$$L(x, y; \mu, \theta) = L_x(x; \mu, \theta) L_y(y; \mu, \theta)$$

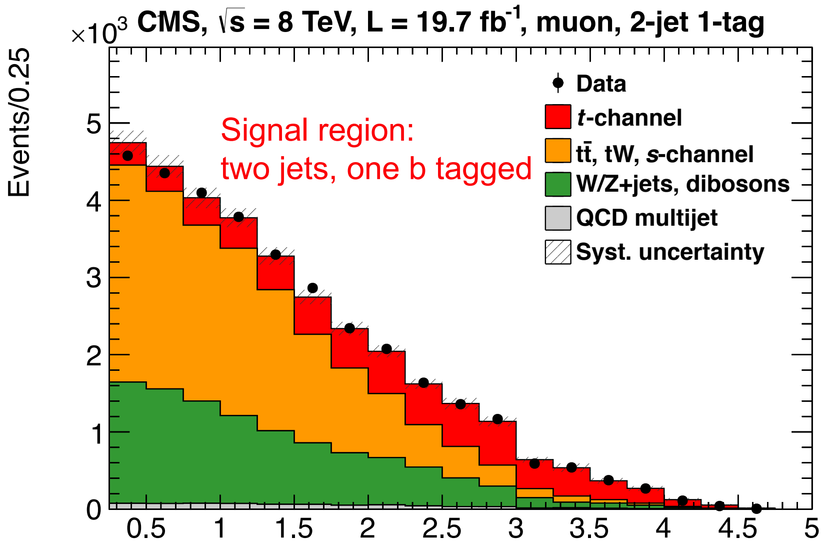
  - L_y does not depend on μ only if there is no signal contamination in the control sample
- Control samples data are not always available
  - Calibrations from test beam, data stored in different formats or analyzed with different software framework, …
- Simplest case; simplified PDF given a ‘nominal’ value θ^{nom}
  - Gaussian, log-normal, Gamma, …

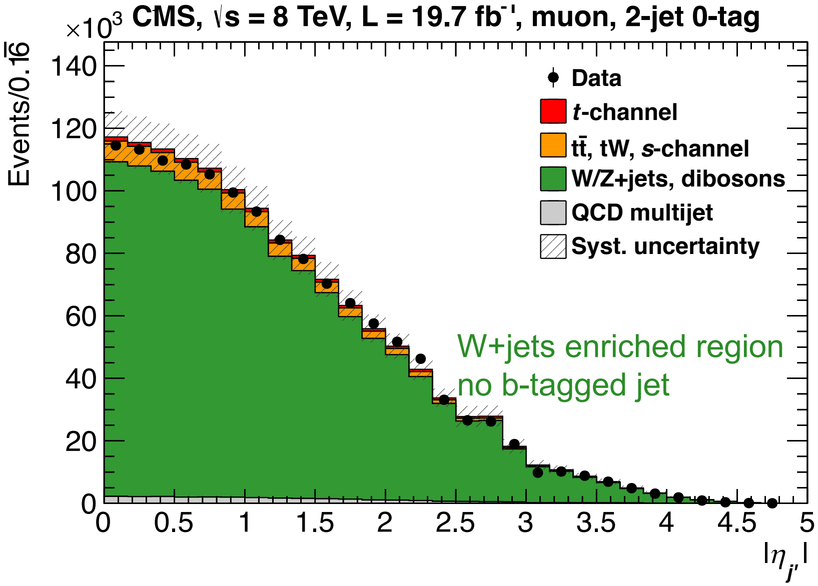
$$L(x, \theta^{\mathrm{nom}}; \mu, \theta) = L_x(x; \mu, \theta) L_{\theta^{\mathrm{nom}}}(\theta^{\mathrm{nom}}; \theta)$$

Fitting control regions

- Control regions and signal region can be fit simultaneously
- Effectively, background yields measured from background-enriched regions are extrapolated to signal regions
  - Scale factors predicted from simulation
- Categories:
  - 2 jets, 1 b tag (signal enriched)
  - 3 jets, 2 b tags ($t\bar{t}$ enriched)
  - 2 jets, 0 b tags (W+jets enriched)





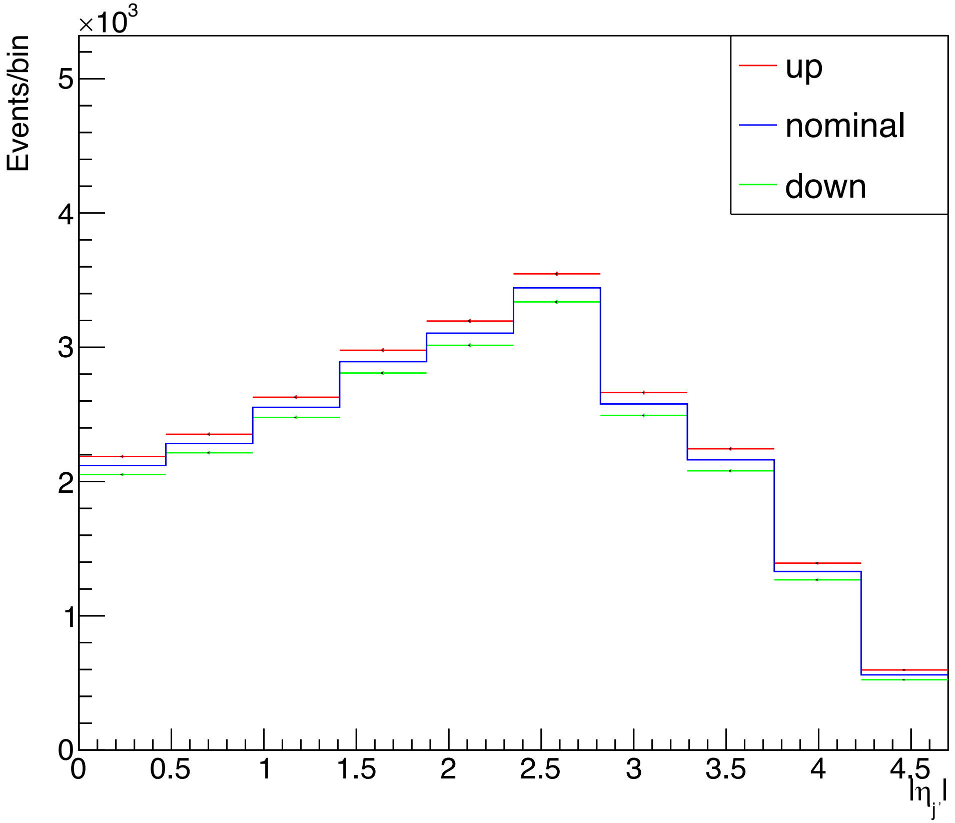


|η_j'|

Measurement of single-top production at LHC, JHEP06(2014)090

Systematics with templates

- Simulation provides samples with a nuisance parameter modified by ± one sigma
  - "up" / "down" variations
- Intermediate values (or outside $\pm 1\sigma$) are determined with interpolation (extrapolation)
  - Linear, parabolic (inter/extra)polation



RooStats

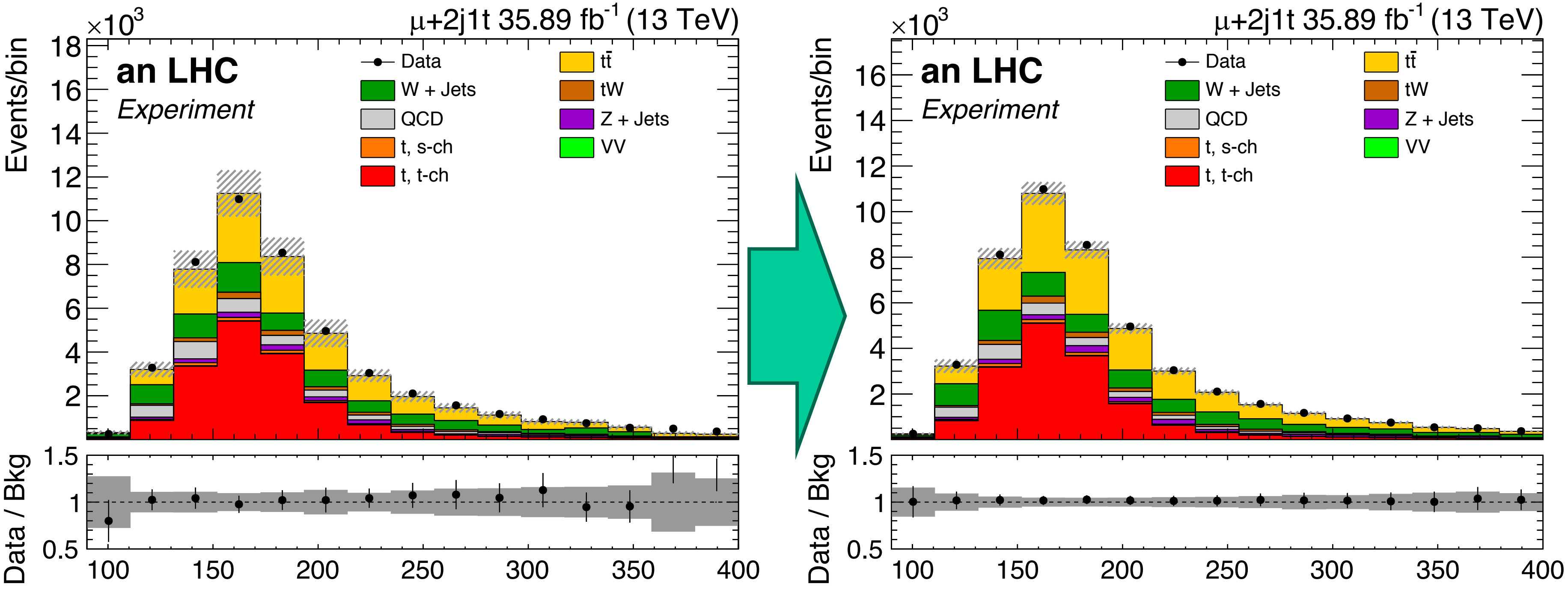
- Most of the methods adopted in High Energy Physics are implemented in the RooStats C++ framework
- Convenient modeling of PDF via RooFit package
  - PDFs from templates determined from ROOT histograms (`RooHistPdf` class)
  - PDF models and data with parameter definition stored in a convenient file format (`RooWorkspace`)
- Asymptotic approximations available, allow to save CPU time avoiding intensive toy Monte Carlo generation
  - G. Cowan et al., Eur.Phys.J.C71:1554,2011

Sources of uncertainties

- Systematic uncertainties may affect the rate (i.e.: cross section) or shape (i.e.: distribution) of a process or both
  - Luminosity
  - Pile up modeling in simulation
  - Jet Energy Scale
  - b-tagging efficiency, mis-id, flavor dependence
  - Mu, e selection, reconstruction and trigger efficiencies
  - Theory modeling:
    - Individual cross section predictions
    - Shape and normalization due to renorm./factor. Scales
    - PDF models
    - Parton shower modeling
    - Generator choice
    - . . .
  - Monte Carlo simulation
    - Limited sample size
  - . . .

Results of fit (1)

- Measurement of parameter of interest
- Nuisance parameters determined from data

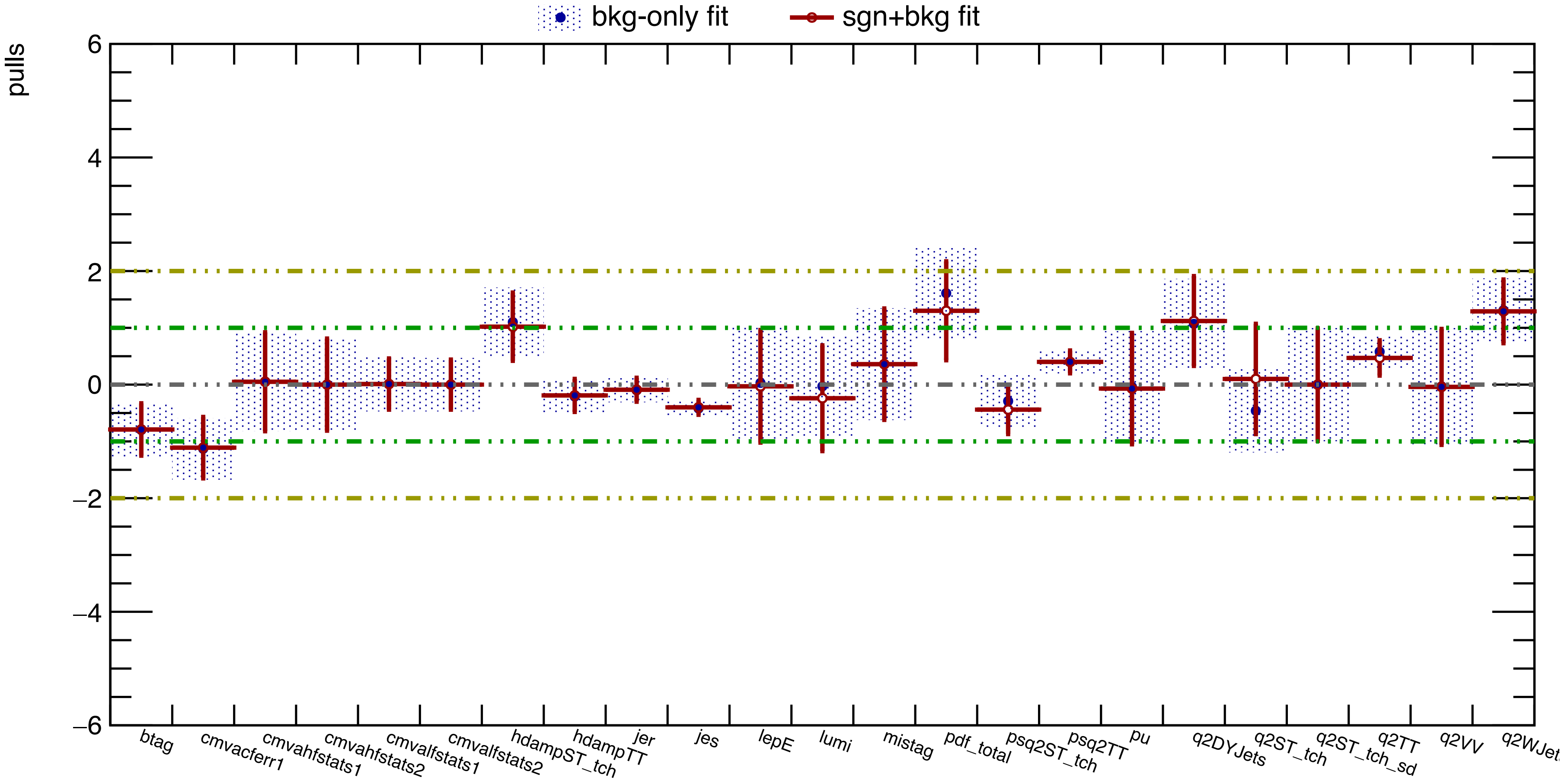


Before fit

After fit

Results of fit (2)

- Constraint of systematic uncertainties



The CMS Higgs combine tool

- Many analyses in CMS use a command-line, datacard-driven, python-powered tool originally developed for the combination of multiple Higgs production/decay channels

- Documentation open to public access:

  https://cms-hcomb.gitbooks.io/combine/content/

Data-cards example

```
# Simple counting experiment, one signal and a few background processes
# Simplified version of H->WW analysis from gitHub documentation
imax 1 number of channels
jmax 3 number of backgrounds
kmax 5 number of nuisance parameters

# just one region (bin = bin1), 0 events observed
bin bin1
observation 0
bin               bin1     bin1     bin1     bin1
process           ggH      qqWW     ggWW     others
process           0        1        2        3
rate              1.47     0.63     0.06     0.22

#systematic uncertainties
lumi     lnN      1.11     -        1.11     -
xs_ggH   lnN      1.16     -        -        -
WW_norm  gmN 4    -        0.16     -        -
xs_ggWW  lnN      -        -        1.50     -
bg_others lnN     -        -        -        1.30
```

Spectra shape naming conventions

- Data and simulation spectra (shapes) are stored as histograms with proper naming convention
  - E.g.: `singleTopTch_muon_2j1t_jesUp` and many more combinations
- Book-keeping may become an issue
  - Histograms may be arranged in different files with overloaded names, or in the same files with different names or in the same file but different ROOT sub-directories
  - Separators, usually underscores, are used in histogram titles to match tags with various meanings
- Higgs combine tool provides a flexible definition via wildcards

```
shapes <process> <channel> <file> <histo-name> <histo-name-for-syst>
```

- E.g. (`$XYZ` is replaced with actual value) :

```
shapes * * htt_mt.input_8TeV.root $CHANNEL/$PROCESS
      $CHANNEL/$PROCESS_$SYSTEMATIC
shapes ggH * htt_mt.input_8TeV.root $CHANNEL/$PROCESS$MASS
      $CHANNEL/$PROCESS$MASS_$SYSTEMATIC
shapes qqH * htt_mt.input_8TeV.root $CHANNEL/$PROCESS$MASS
      $CHANNEL/$PROCESS$MASS_$SYSTEMATIC
shapes VH * htt_mt.input_8TeV.root $CHANNEL/$PROCESS$MASS
      $CHANNEL/$PROCESS$MASS_$SYSTEMATIC
```

MC statistical uncertainty

- Limited simulation statistics in each bin is also a source of uncertainty

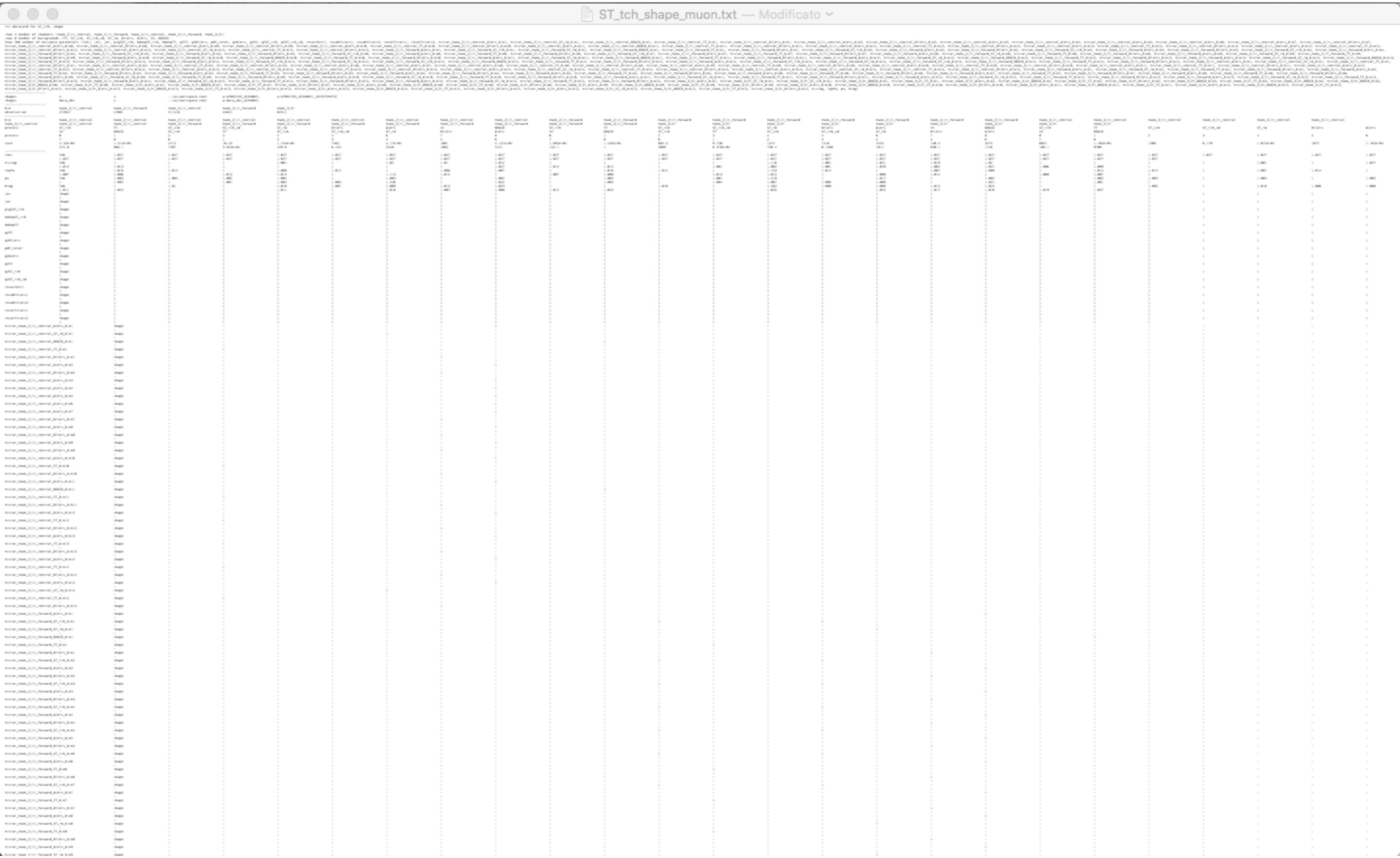
One parameter per bin

=

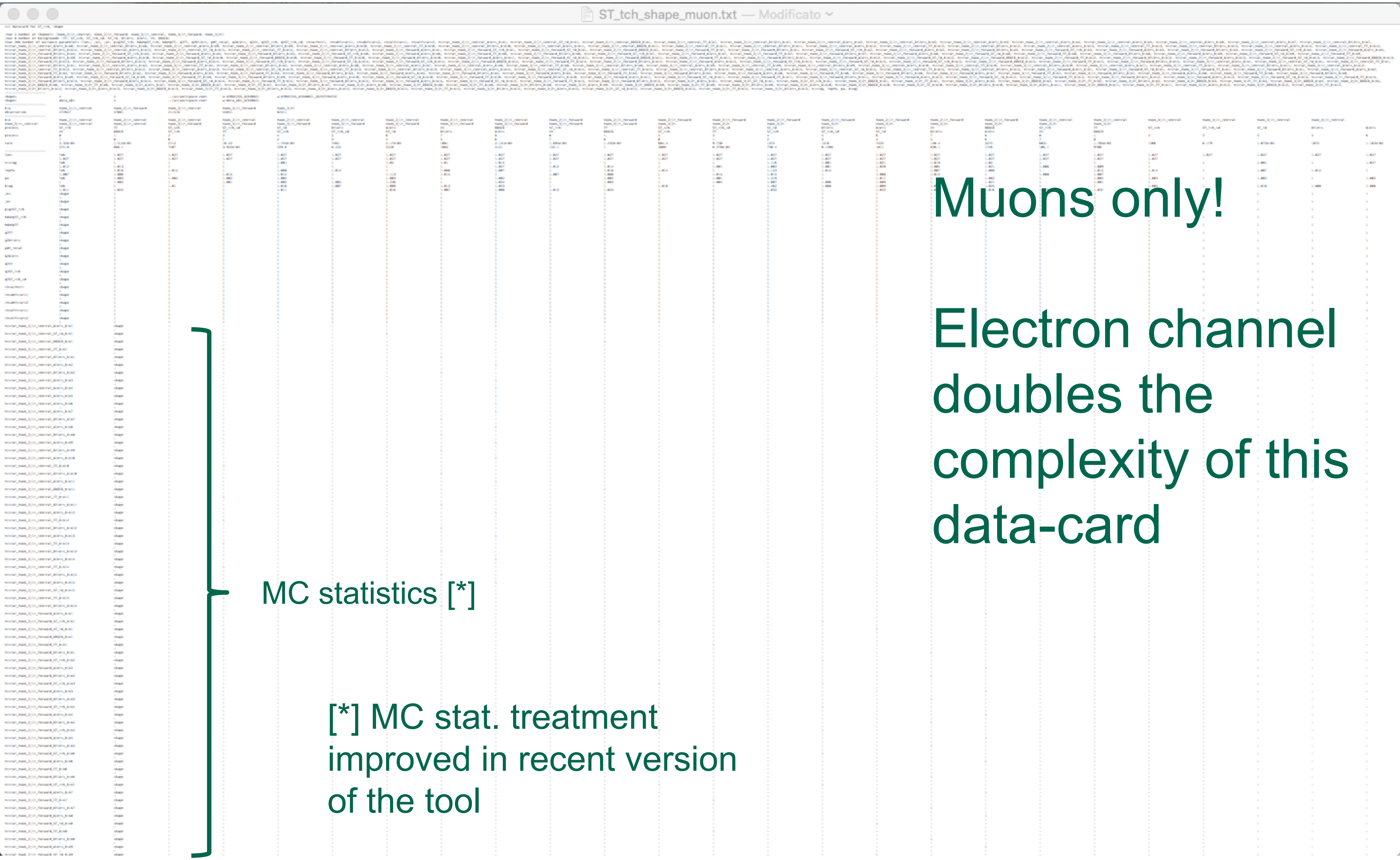
Many parameters!

- The previously-presented treatment requires two spectra (up/down) for each bin (!!!), each varied up and down by its statistical uncertainty
  - **Redundant**: uncertainty is already stored in ROOT histograms!
- Uncertainties in bins with large number of entries may be neglected, simplifying the problem
  - Typical of exponentially falling spectra

Realistic data-cards



Realistic data-cards



Muons only!

Electron channel doubles the complexity of this data-card

MC statistics [*]

[*] MC stat. treatment improved in recent version of the tool

Applying constraints

- Background in signal region constrained from control region
- Scale by bin-dependent factor α_i
  - $h_i^{(\mathrm{sig})} = h_i^{(\mathrm{bkg})}\alpha_i$
  - α_i determined from Monte Carlo samples
- Histogram content in each bin depends on the value of nuisance parameters
  - Scaled histogram represented by a customized **`RooAbsPdf`** object
- RooFit helper class: **`RooFormulaVar`**
- From online tutorial:

```
RooFormulaVar wFunc("w","event weight","(x*x+10)",x);
```

- Parameter name are ‘encoded’ into strings, which may require convoluted code to define strings in complex cases
  - **Bugs only spotted at run time!**

Code example

```
const Config& cfg = Config::get();
 string name = "bkg_CR_" + sampleName_ + "_"+ controlRegionName_ + "_bin" + stringBin;
 string descr = "Bkg. CR " + sampleName_ + " yield in SR " + controlRegionName_ +
   ", bin " + stringBin;
 if(cfg.hasSystematics()) {
   string formula = "@0*(";
   RooArgList args;
   args.add(*signalRegionBins_[bin]);
   unsigned int count = 1;
   for(unsigned int syst = 0; syst < cfg.numSystematics(); ++syst) {
     ostringstream ssc1; ssc1 << count; formula += "@" + ssc1.str();
     ostringstream ssc2; ssc2 << ++count; formula += "*@" + ssc2.str();
     args.add(container_.systematicParameter(syst));
     args.add(*slopes_[syst][bin]);
     if(syst < cfg.numSystematics() - 1) formula += " + ";
   }
   formula += ")";
   controlRegionBins_.push_back(
     make_shared<RooFormulaVar>(name.c_str(), descr.c_str(), formula.c_str(), args));
 } else {
   RooArgList args(*signalRegionBins_[bin], *ratioCRSR_[bin]);
   controlRegionBins_.push_back(
     make_shared<RooFormulaVar>(name.c_str(), descr.c_str(), "@0*@1", args));
 }
```

Automatic data-cards generation

- Large data-cards can be automatically generated with ad-hoc software
  - One extra layer on top of Higgs combine tool, which is already a layer on top of RooStats
- Uncertainties assigned to blocks/groups of samples in one shot
- Possible improved management of statistical uncertainties
  - E.g.: only consider least populated bins

Possible simpler organization

- Spectra in data and simulation can be categorized using the following ‘classes’:

- Data / Simulation process
  - Single top, $t\bar{t}$, W+jets, QCD, etc.
- Signal/control regions (sometimes called category in analysis notes)
  - Signal region: 2j1b; control regions: 2j0b, 2j2b, 3j, etc.
- Channel
  - Semileptonc decays to electrons, muons; full hadronic decays
- Distribution
  - Specific spectrum for a given process, region and channel

- Uncertainties and nuisance parameters may pertain to a specific class

Parameter organization

- Parameters may be common to groups of distributions
  - Common to all spectra:
    - Luminosity, jet-energy scale, b-tag, …
  - Common to a process:
    - Theory uncertainties (renorm./factor. scale, affect both shape and rate)
  - Common to a decay channel:
    - Muon, electron efficiencies (reconstruction, isolation, trigger)
  - Possibly even common to a (control/signal) region
    - Not used in the considered case
  - Specific to a single spectrum:
    - Statistical uncertainty from simulation in each bin

```
HistoDirectory /afs/cern.ch/work/u/user/fits

Categories muon_2j1t_central muon_2j1t_forward muon_3j1t_central muon_3j1t_forward muon_3j2t
CategoryFiles muon muon muon muon muon #Same file for all categories, in this case

#Variables whose spectra is saved in the workspace
VariableNames h_2j1t_topMass_mtw_G_50_AND_etajprime_L_2p5 h_2j1t_topMass_mtw_G_50_AND_etajprime_G_2p5
h_3j1t_topMass_mtw_G_50_AND_etajprime_L_2p5 h_3j1t_topMass_mtw_G_50_AND_etajprime_G_2p5 h_3j2t_topMassLeading
#Variable name used by RooFit
RooRealVar topMass
#MC samples (signal, background) for each process
SignalSample ST_tch     single top t-channel
BackgroundSample TT     ttbar
BackgroundSample ST_sch             single top s-channel
BackgroundSample ST_tch_sd          single top t-channel_sd
BackgroundSample ST_tW              single top tW
BackgroundSample DYJets             Drell-Yan
BackgroundSample WJets  W + jets
BackgroundSample VV     diboson
BackgroundSample DDQCD  QCD
#Rate parameters to be fit from data-driven processes (QCD in this case)
#RateParam <par name> <region to fit> <process> <region to fit> <process> . . .
RateParam QCD_muon_2j1t muon_2j1t_forward DDQCD muon_2j1t_central DDQCD
#lumi is a global rate uncertainty
LumiUncertainty 0.027
#normalization only, applied to all processes
NormSystematics mistag lepMu pu btag
#other systematics. Specific to one process if named <syst>_<process>, otherwise common to all
Systematics jes jer psq2ST_tch hdampST_tch hdampTT q2TT q2DYJets pdf_total q2WJets q2VV q2ST_tch q2ST_tch_sd
cmvacferr1 cmvahfstats1 cmvahfstats2 cmvalfstats1 cmvalfstats2
#still problematic the insertion of stat. uncertainties. All bins enumerated. Omitted here for simplicity
```

This **meta-data-card** generates:

- the complex data-cards shown before
- the RooFit workspace from histogram files

Possible approaches for a general solution

- Goal:
  - More easily management of the most commonly used cases
    - We already have a large number of use-cases in place
- Possible solutions:
  - Extension of the CMS Higgs combine interface
    - Code publicly available in gitHub, but integrated in CMS software release system
    - Promote it as common HEP tool?
  - Extension of ROOT/RooStats
    - Usable by the entire HEP community
    - C++ or python interfaces (or both)?
  - Should data-cards be entirely replaced by a python scripts?
- Any more thoughts?

Pseudo code, just brainstorming…

```
Processes singleTop, ttbar, Wjets, QCD
Channels electron, muon, hadronic, hadBoosted
Regions (electron, muon).(2j1b, 2j0b, 2j2b),
        hadronic.5j1b, hadBoosted.2j1FatJet1b

HistoNames $Process_$Region_$Channel_topMass

NuisanceParameters
  lumi, btagScale, jetEScale, jetEResol
  (singleTop, ttbar).(renScale, mcScale)
  (Wjets, QCD, ttbar, singleTop).mcScale
  electron.elEffScale
  muon.muEffScale
  Wjets.stat[bins: 10-20]
  QCD.stat[bins: *]
```

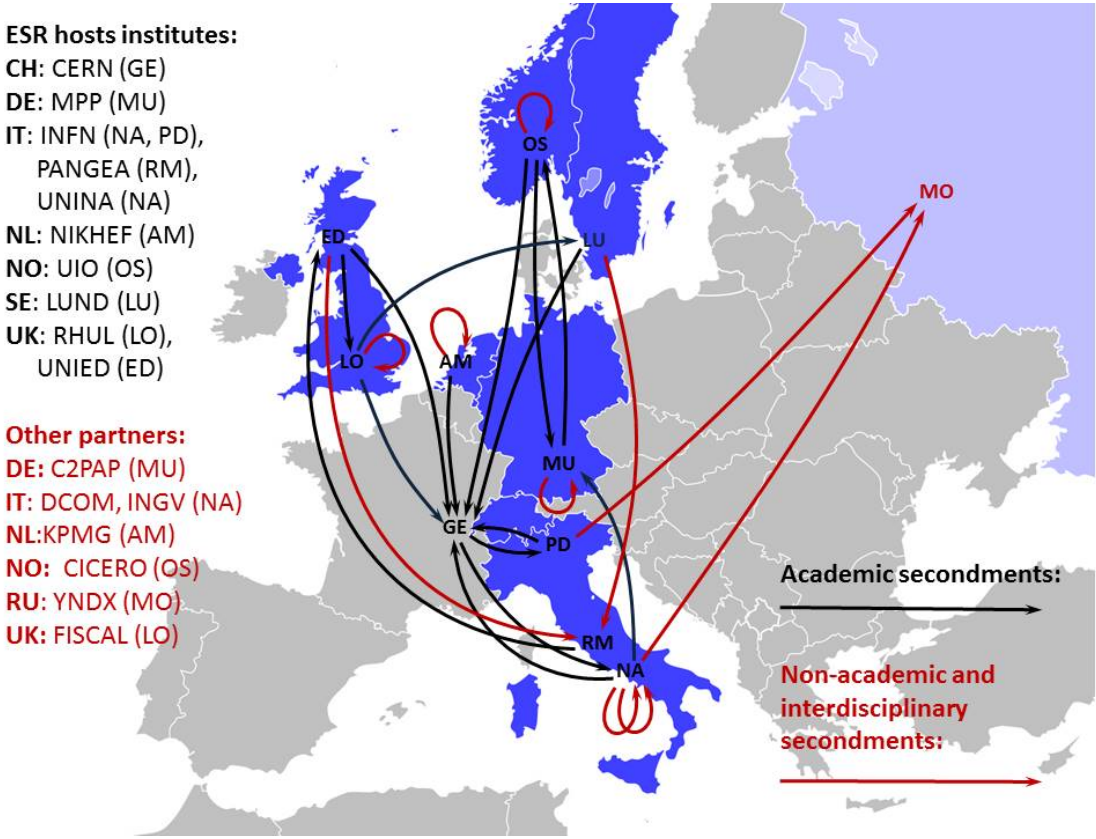
CMS combine tool recently implemented syst. grouping and improved MC stat treatment

Python script may be an effective replacement to data cards

Insights

- International Training Network of Statistics for High Energy Physics and Society

- INSIGHTS is a 4-year Marie Sklodowska-Curie Innovative Training Networks project for the career development of 12 Early Stage Researchers (ESRs) at 10 partner institutions across Europe.
- INSIGHTS is focused on developing and applying latest advances in statistics, and in particular machine learning, to particle physics
- CERN is part of the network with deep interconnection with the ROOT development team



ESR hosts institutes:
CH: CERN (GE)
DE: MPP (MU)
IT: INFN (NA, PD),
PANGEA (RM),
UNINA (NA)
NL: NIKHEF (AM)
NO: UIO (OS)
SE: LUND (LU)
UK: RHUL (LO),
UNIED (ED)

Other partners:
DE: C2PAP (MU)
IT: DCOM, INGV (NA)
NL:KPMG (AM)
NO: CICERO (OS)
RU: YNDX (MO)
UK: FISCAL (LO)

Academic secondments:

Non-academic and interdisciplinary secondments:

https://www.insights-itn.eu/

Future developments

- Insights' Early-Stage Researchers have been selected
- Will shortly start working on different statistical tools and applications
- One of the projects proposes development for the presented problem
- Inputs and suggestions are welcome!
- We are in the early stage for these developments!

Conclusions

- Most of data analyses at LHC, both precision measurements and search for physics beyond the SM, require simultaneous statistical analysis of many data samples to constrain systematic uncertainties
- Managing the achieved complexity requires a substantial amount of coding and challenges the structure of the present software interfaces
- Ad-hoc solutions and mini-framework are implemented in experiment and for specific analyses
- A common implementation in the framework of RooFit/RooStats/ROOT tools is desirable in order to simplify the management of many applications