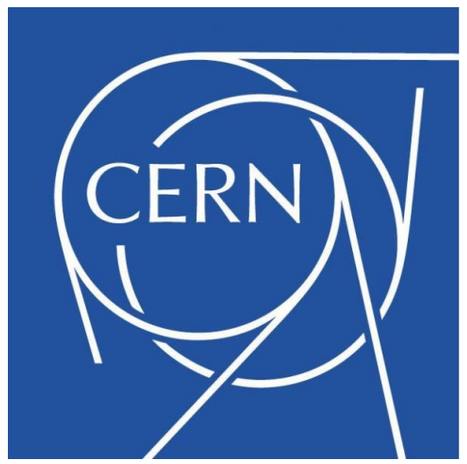
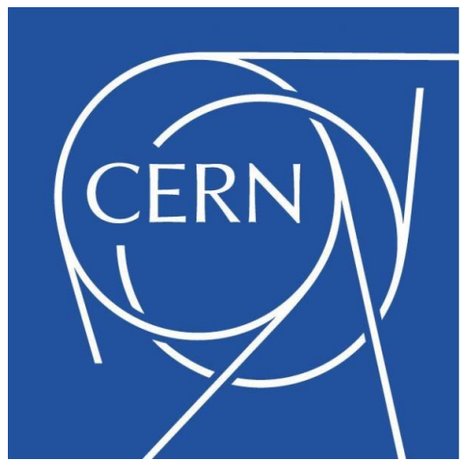
Tools for Data Analysis in HEP

Fons Rademakers, CERN
HEPTech 2015, Budapest, 31-3-2015



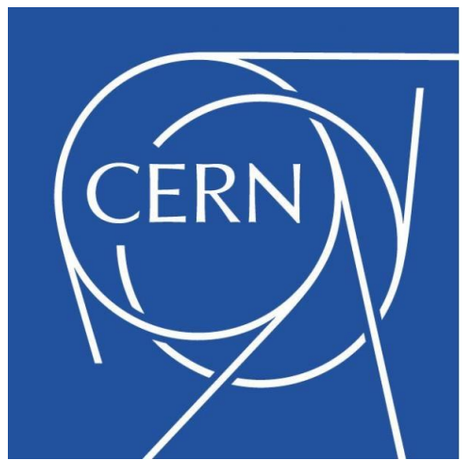
Efficient Data Analysis and Visualisation

Fons Rademakers, CERN
HEPTech 2015, Budapest, 31-3-2015



ROOT: Scientific Data Analysis

Fons Rademakers, CERN
HEPTech 2015, Budapest, 31-3-2015



Outline

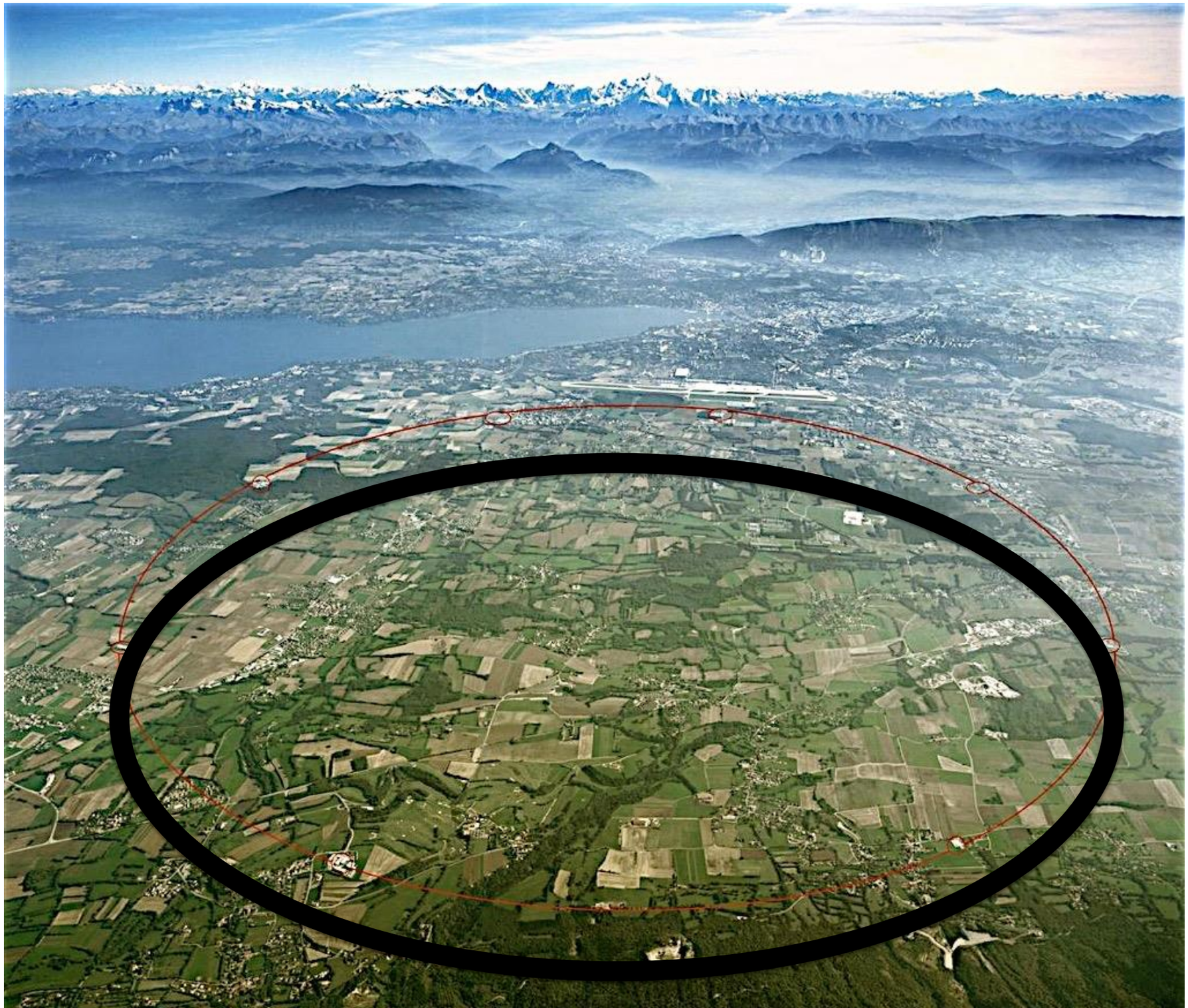
- The Context: CERN, its data, its scientists
- Data Visualisation
- Data Storage
- Statistics
- Extensions
- Conclusion

CERN

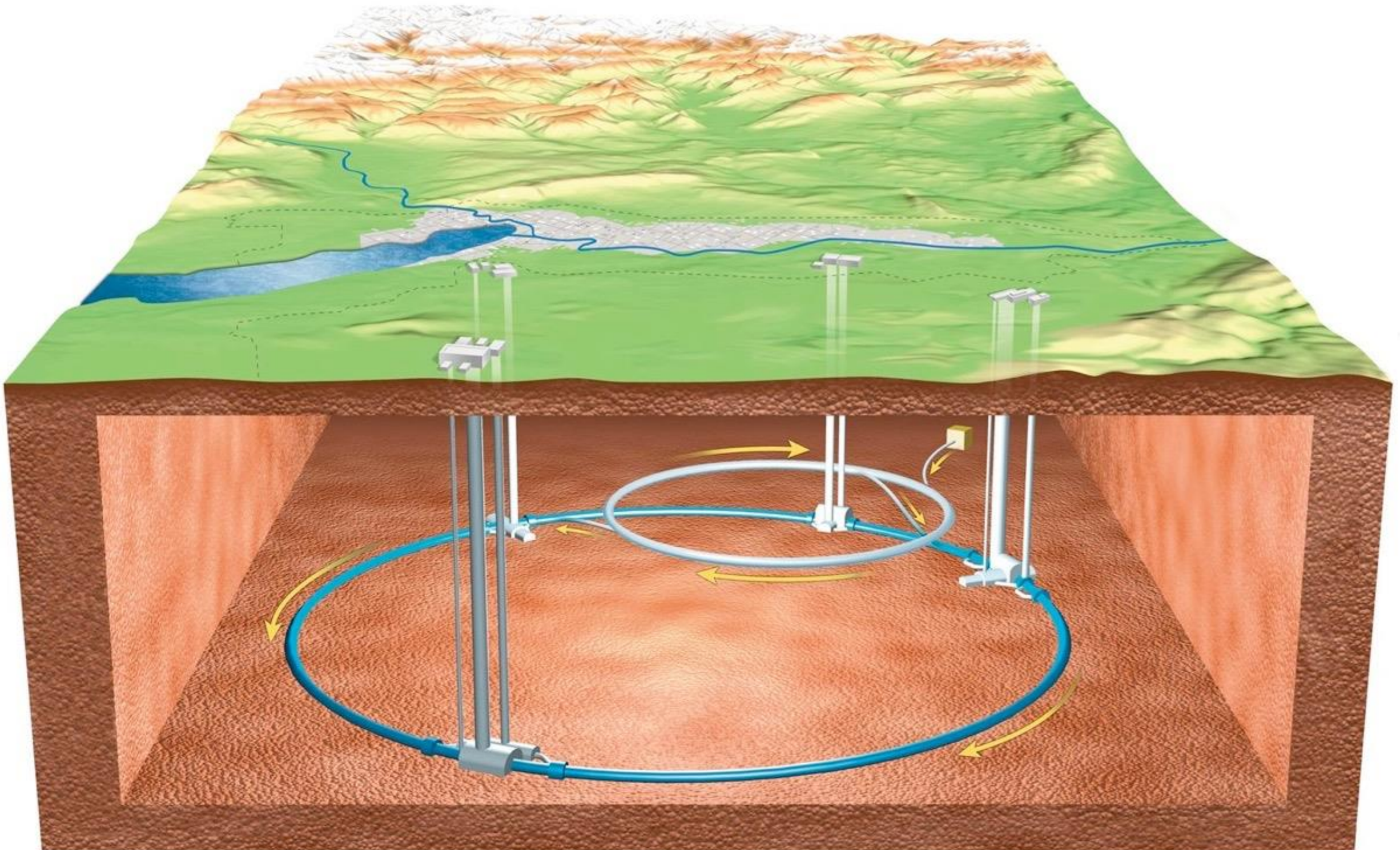
Fact Sheet

- European Organisation for Particle Research, Geneva / Switzerland
- Since 1954: from Heisenberg to Kobayashi to Higgs
- Fundamental research (WWW: inventions happen)
- What is mass? What is the universe made of? Probing smallest scale
  - Higgs, Super Symmetry,...
- Used by 10,000 physicists, 608 universities, 113 nations

Large Hadron Collider

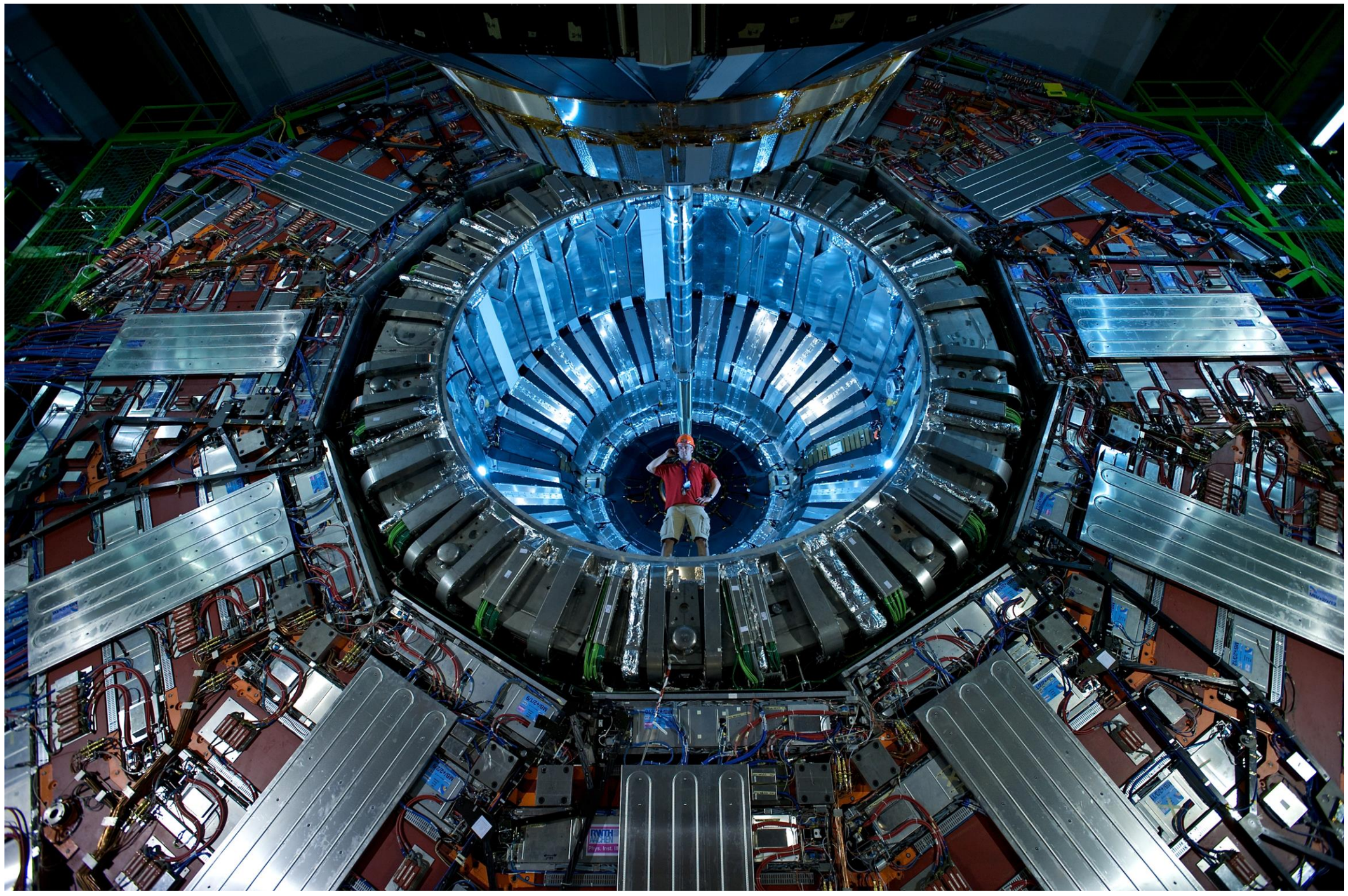


Large Hadron Collider



Large Hadron Collider

A Large Hadron Collider Detector



Large Hadron Collider

- World’s “biggest” particle accelerator - a giant microscope
- Ring 27km long, 100m below Switzerland and France
- Four large experiments
- Expected to run until approx 2030
- 24/7 operations
- Can’t be ordered on the web: do-it-yourself attitude

Physics Results

- LHC is a “discovery machine”
- Higgs!
- Super Symmetry is unlikely
- Standard Model versus Gravity? Crossing our fingers for the new data starting this year!

Computing At CERN

- 15 Petabytes LHC data / year, growing to 400 PB/y by 2023
- 50 million lines of C++ code:
  read detector, find physics objects, filter interesting data, analyze physics. Also: simulate all of that!
  - Fast: data size
  - Correct: scientific results
  - Stable: Higgs is rare
- Distributed resources (Grid/Cloud), >100,000 cores continuously busy

ROOT

ROOT Business Card

- http://root.cern.ch
- C++, with Python interface
- For large volumes of (scientific) data of arbitrary layout defined by C++ types
- Extremely efficient storage: “data base” of samples
- Data analysis (also concurrent: PROOF), statistics
- Visualisation

ROOT Context

+

+

=

Fact Sheet

- Started in 1995, multi-platform for lifetime of >30 years: Linux, MacOS X, Windows*, Solaris*, AIX*, iOS*, HPUX*; i686, x86_64, x32, ARM, PowerPC*

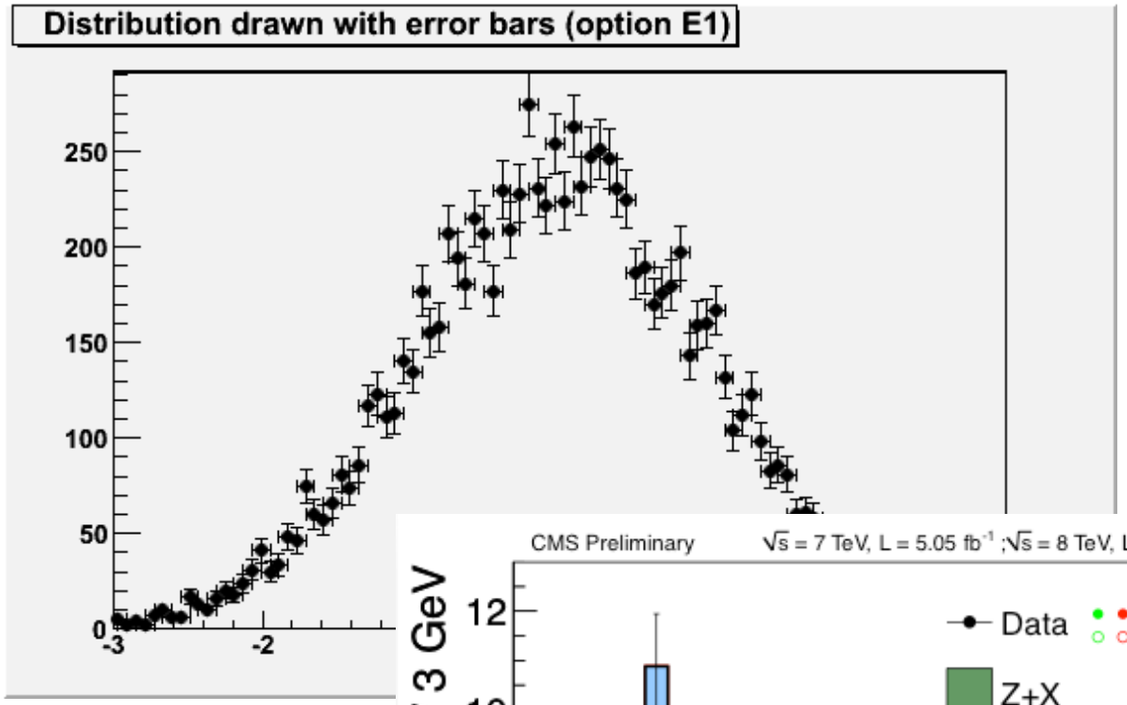
  * not all versions :-(

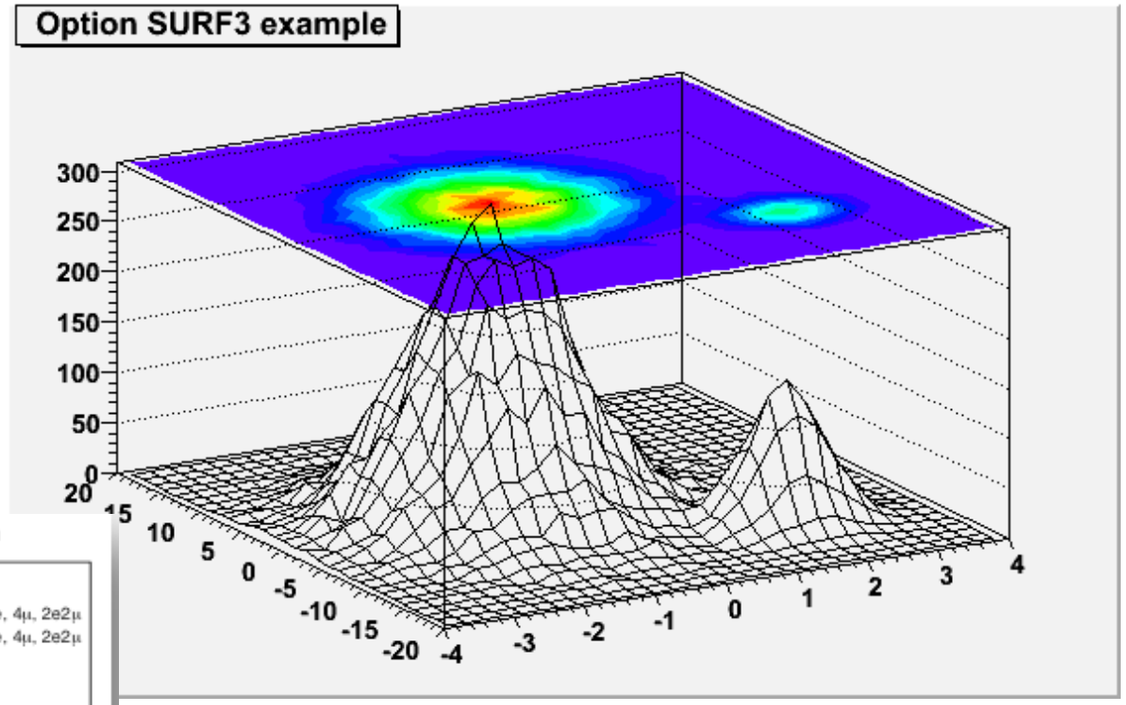
- About 3 million lines of code, 10 developers, open source (LGPL), very active community

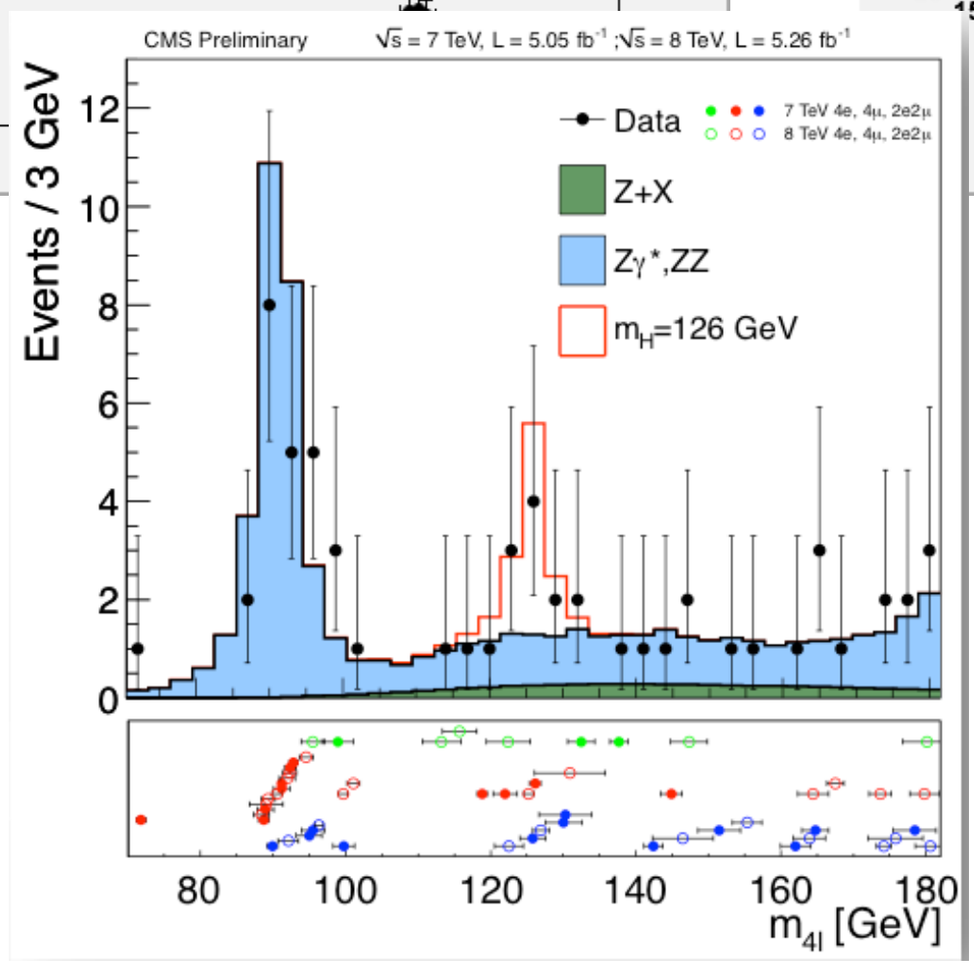
- Used by all High Energy and Nuclear Physics experiments, plus finance, aerospace, astronomy, bio-informatics…

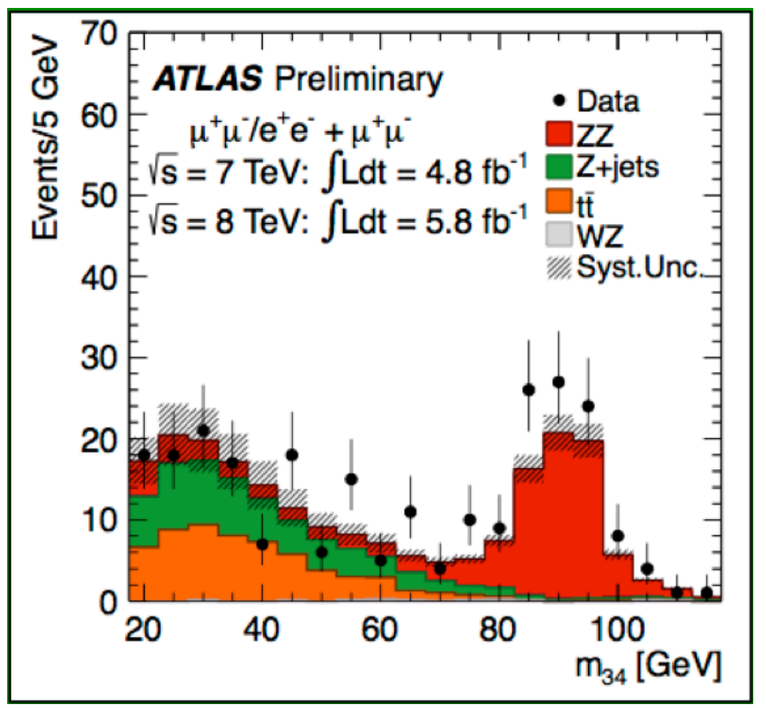
- Two releases per year of production quality: petabytes of data, tens of thousands of users rely on us!

Visualization: Histograms

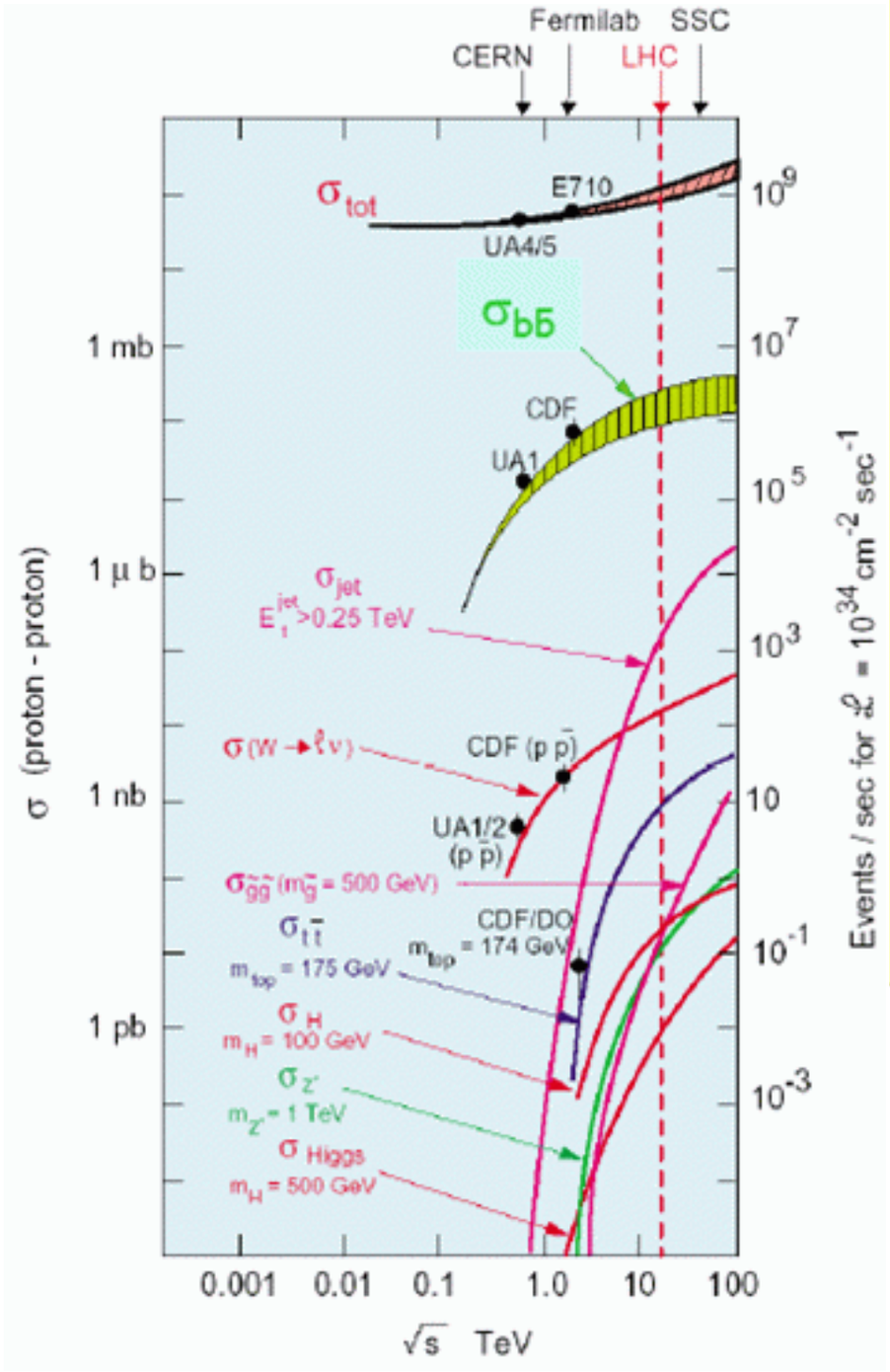


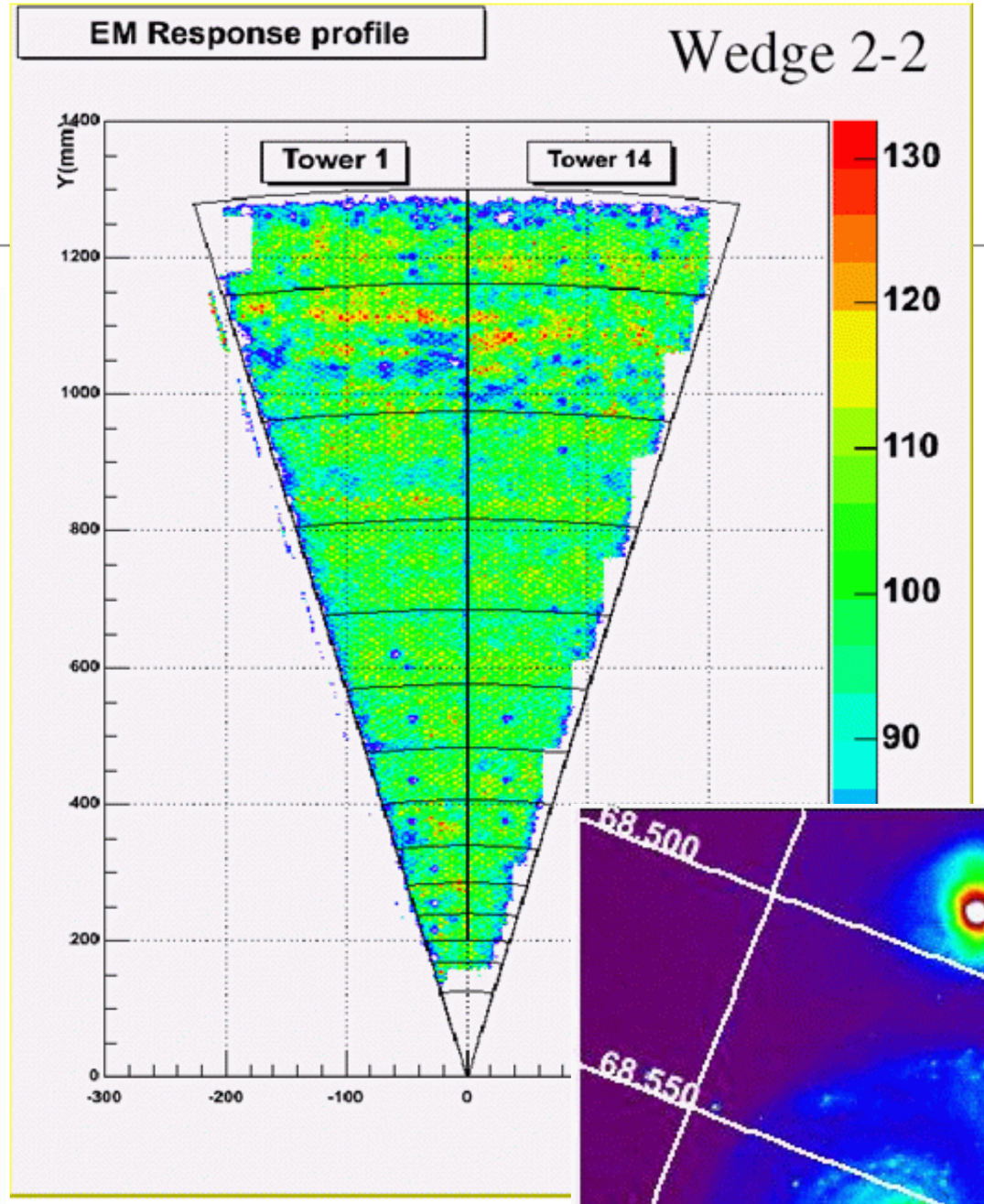


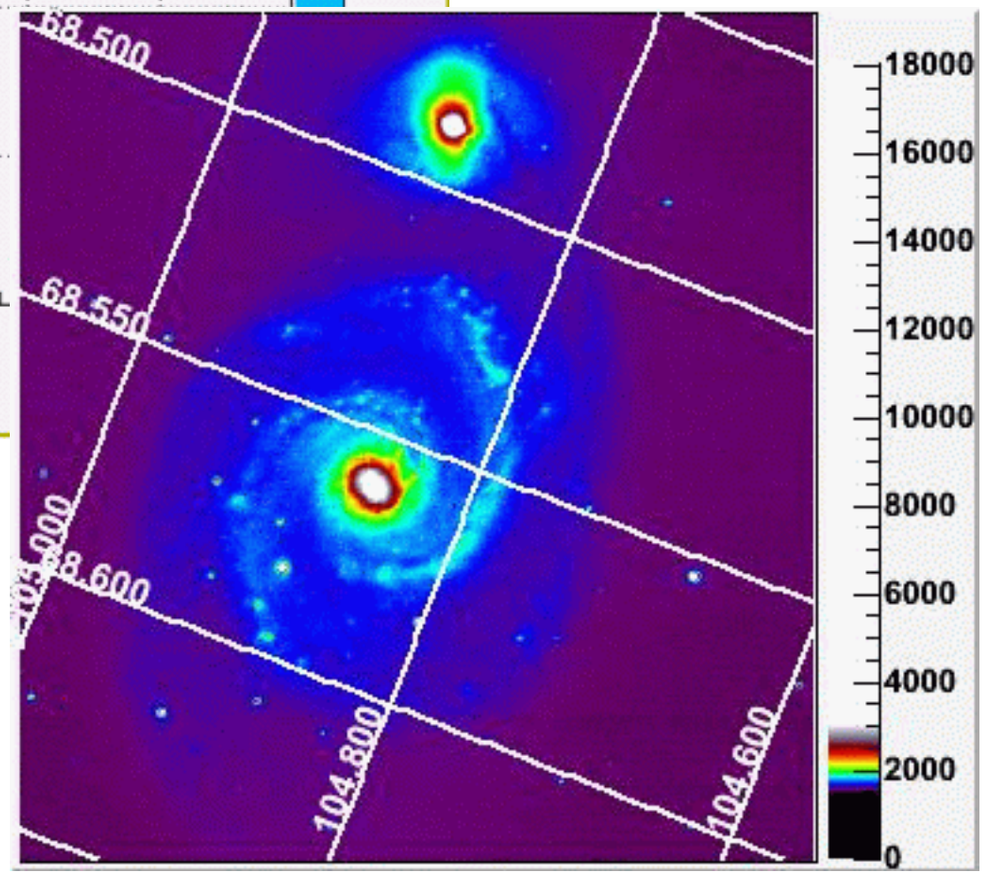




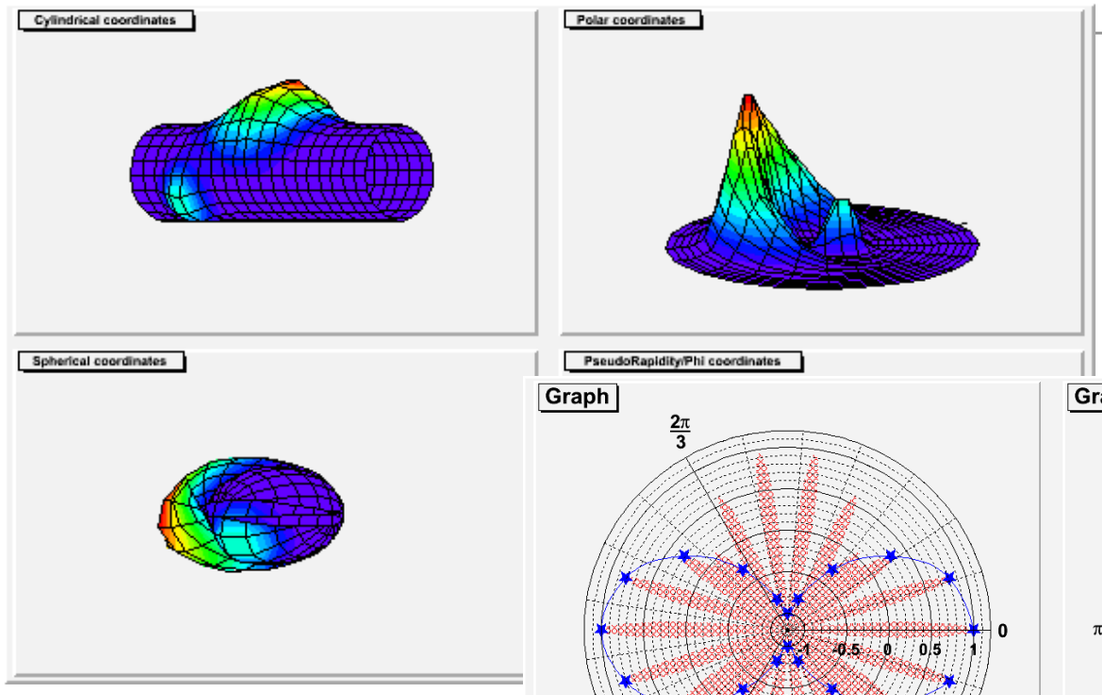
Visualization: 2D

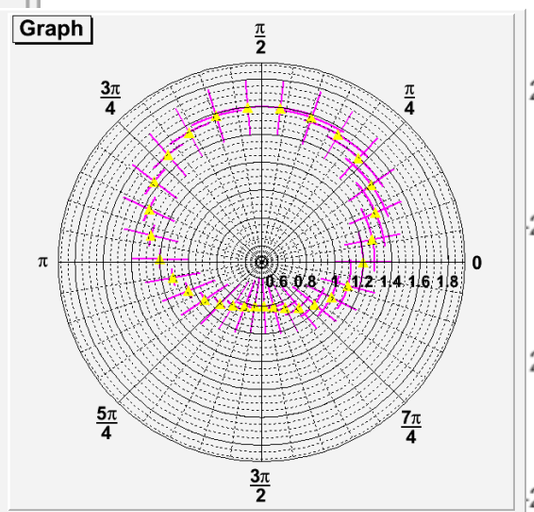


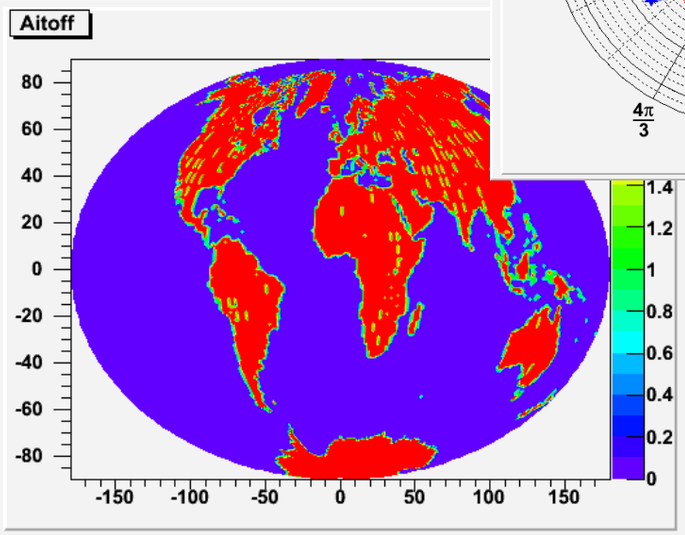


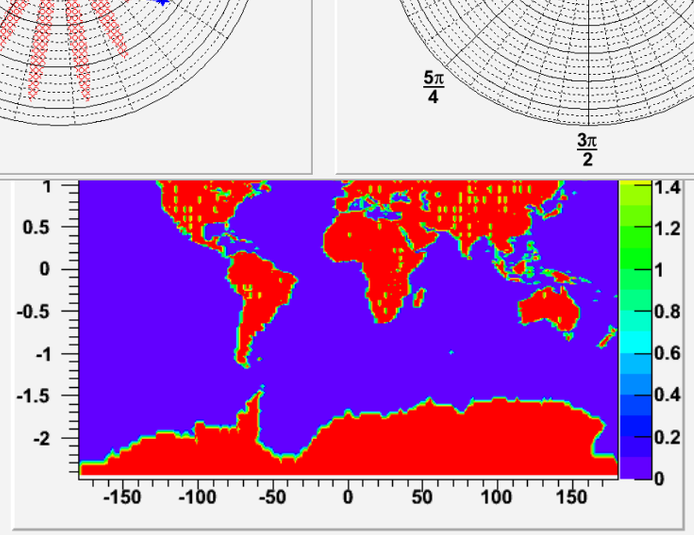


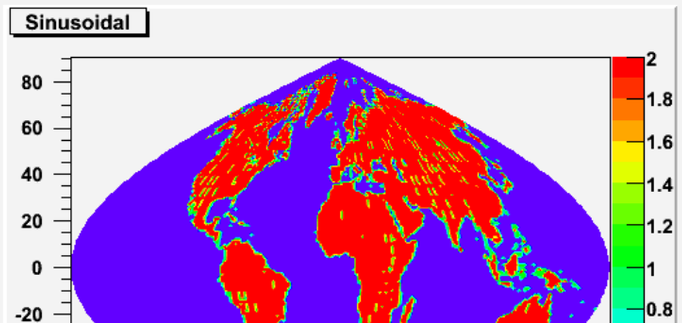
Visualization: 2D Specialties

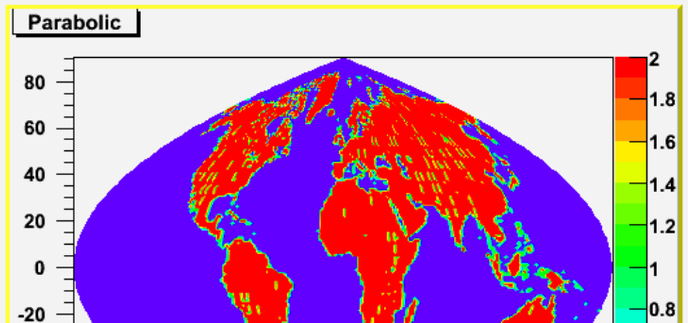


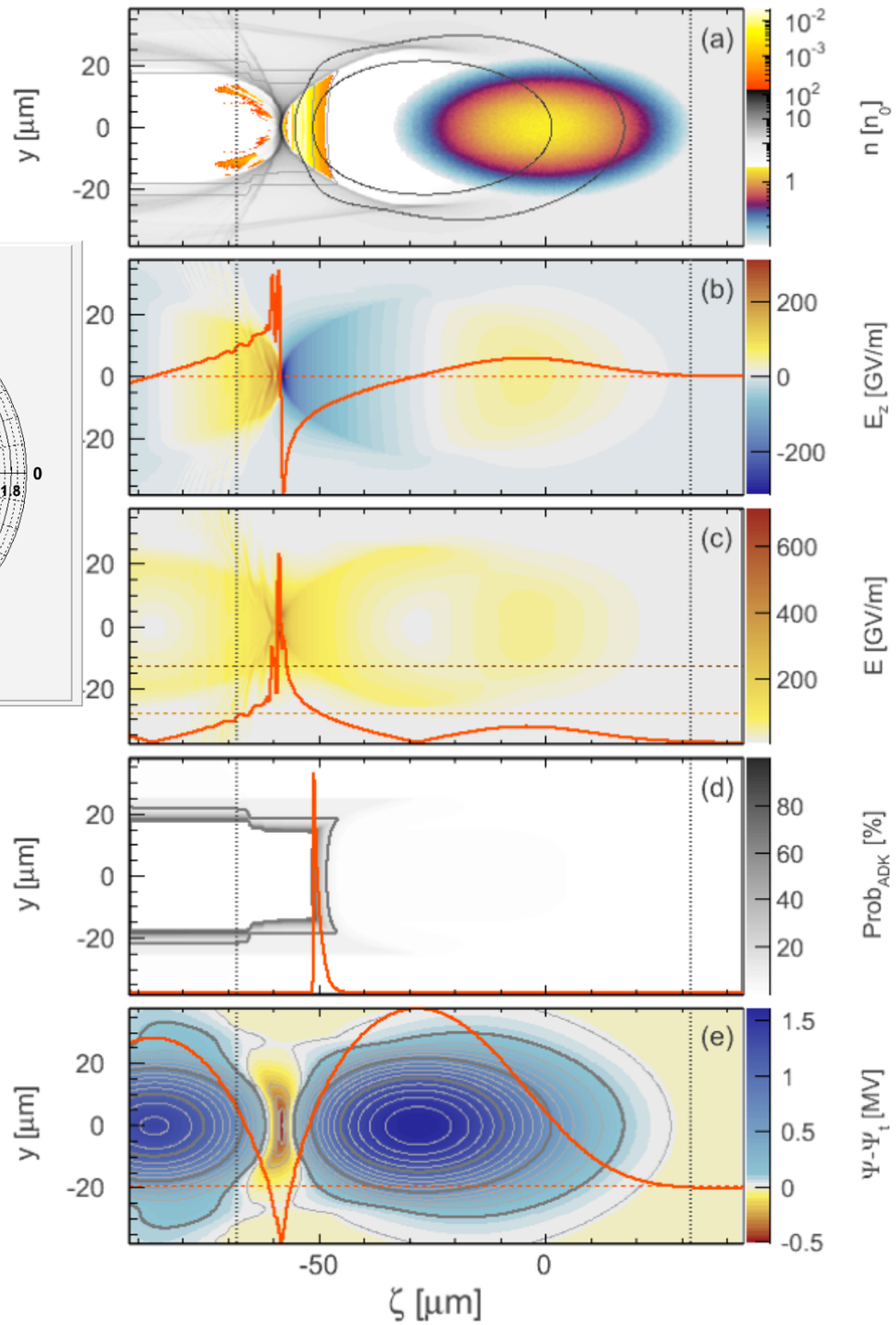




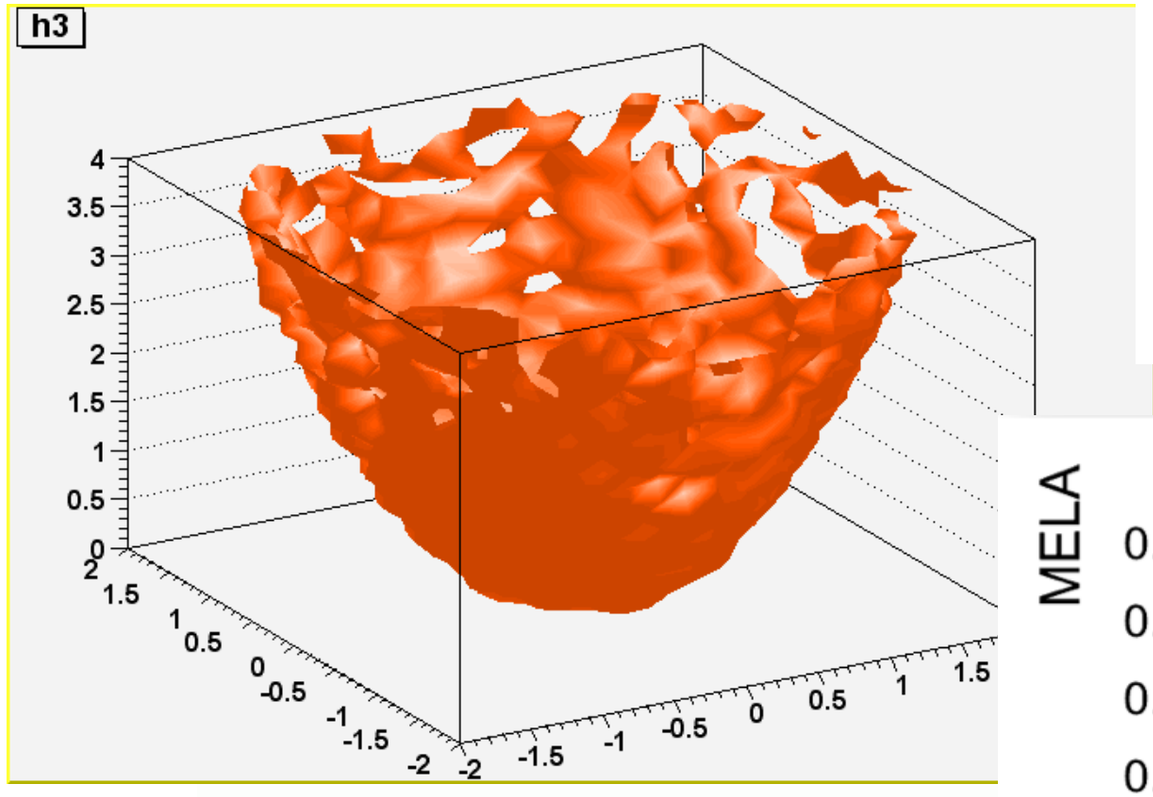


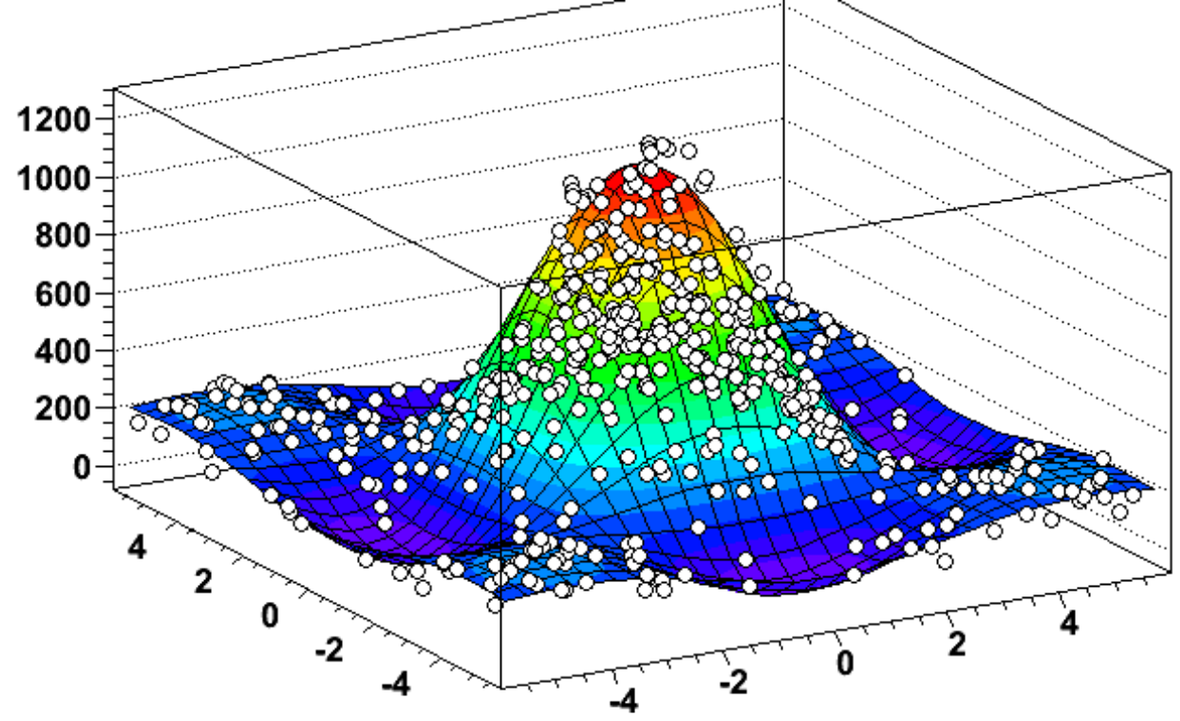


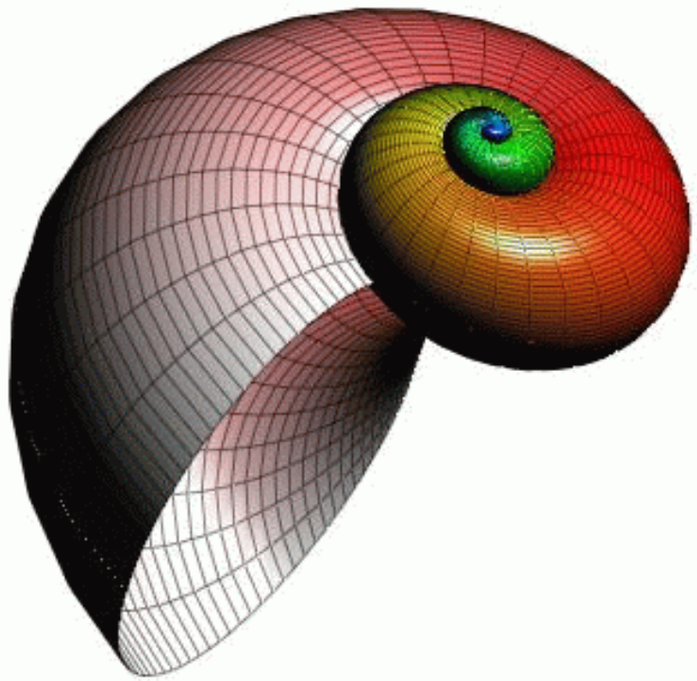


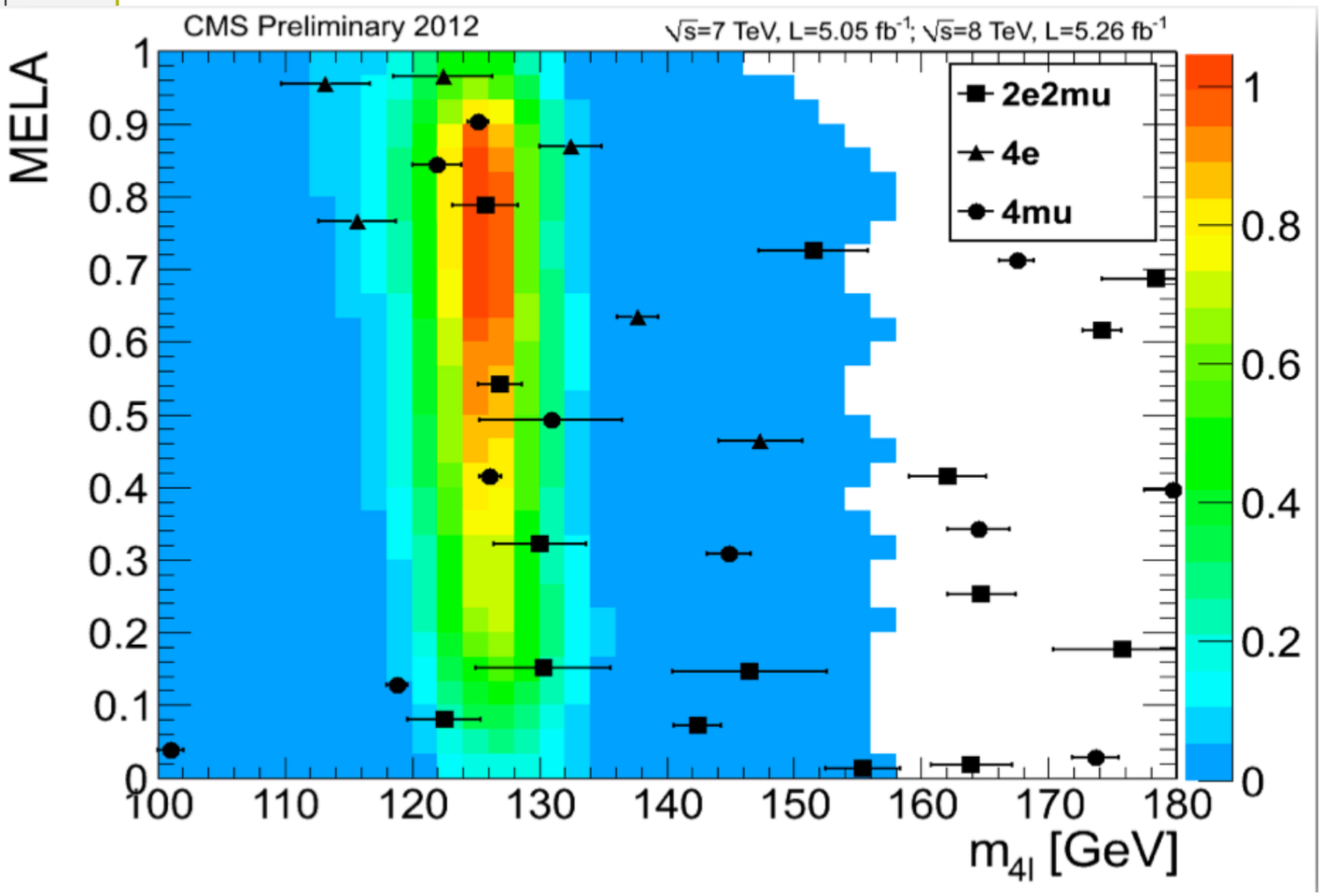


Visualization: 3D

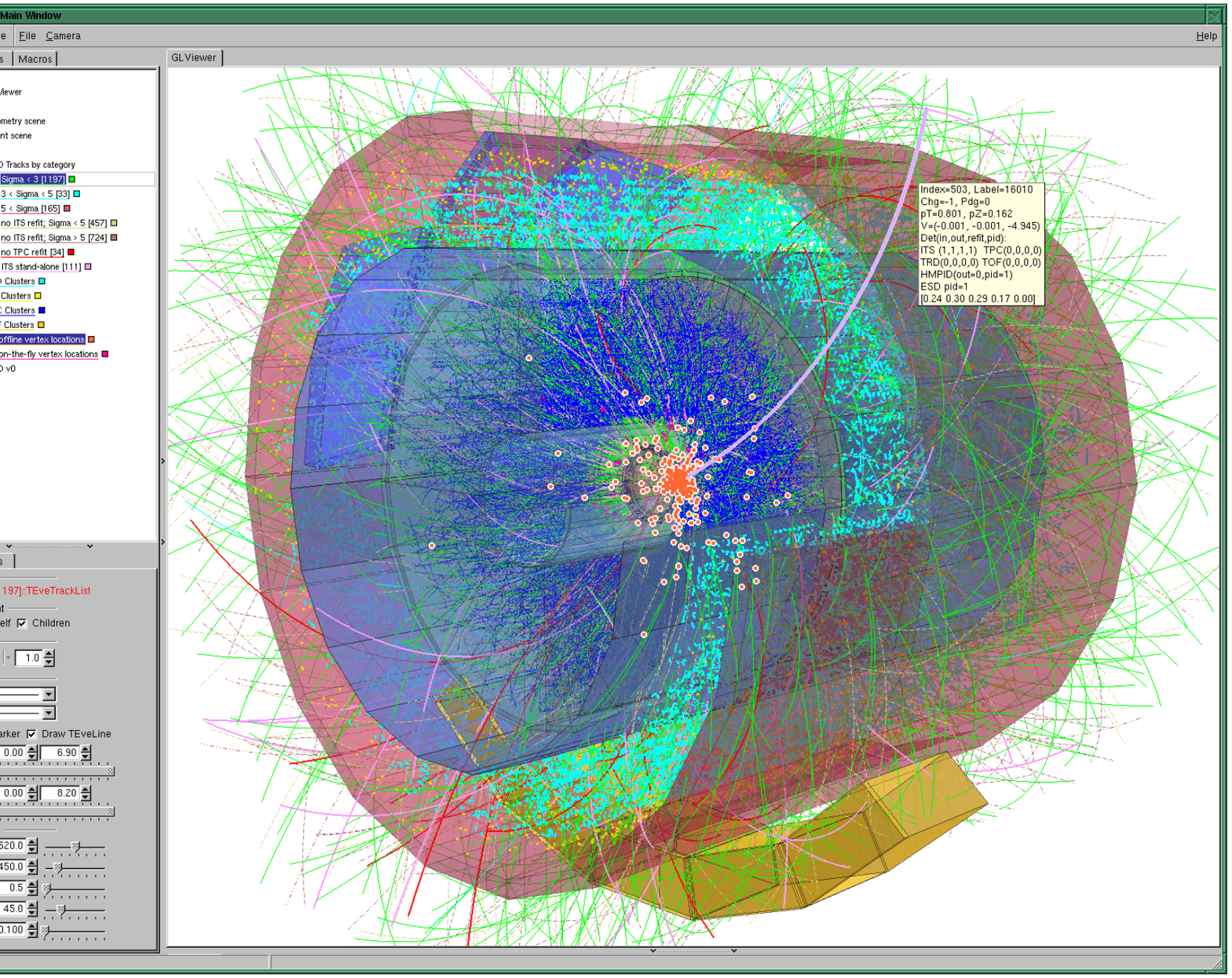


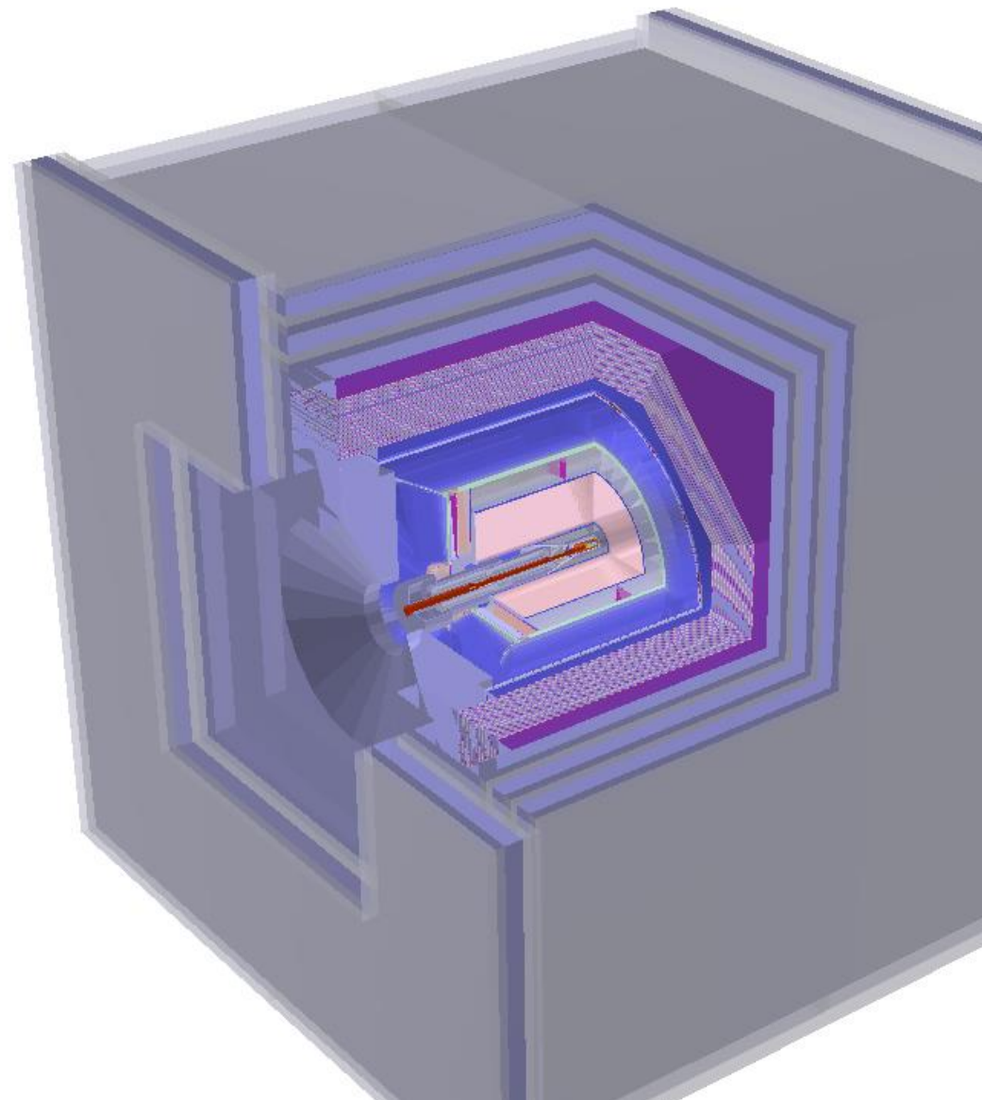






Visualization: CAD-style





Visualisation, GUI

- Many scientific visualisation techniques with advanced palettes
  - + working on interface to ParaView
- Graphics with X11, Cocoa, GL
- ROOT comes with dedicated GUI library, but can be used with Qt, MFC, WPF, Gnome, etc
- Javascript-based interactive visualization, too! https://root.cern.ch/js/3.2/index.htm

Data Storage / “Database”

- Experiments define data structures as C++ types
  - Custom classes of vectors of custom classes of…, i.e. high-dimensional data
- ROOT offers compressed binary storage and database connectors: Postgres, MySQL, Oracle, etc.
- Extremely efficient collection (TTree): column-wise or sequential storage, minimal reads, prefetching (WAN)
- Automatic versioning and schema evolution

“All we do is count.”

– Fons Rademakers

“But we are pretty good at that.”

– Fons Rademakers

Statistics

- High Energy Physics repeats collisions billions of time to measure probability of physics processes

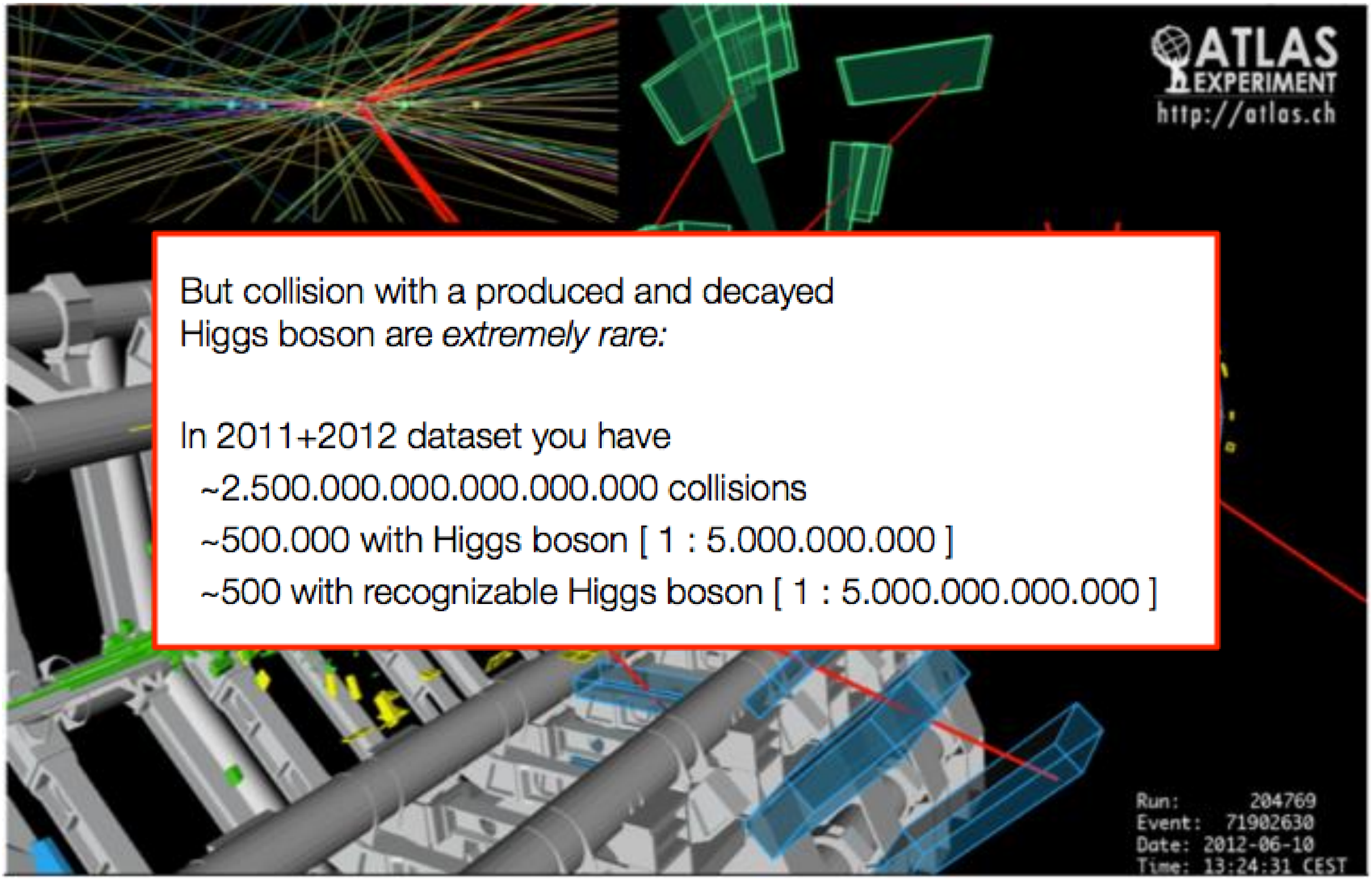
- Physicists handle
  - Conditional probabilities, frequentists versus Bayesian, p-values, t-tests, simulation uncertainty convoluted with measurement uncertainty, correlated multi-parameter minimisations

- We even come up with our own issues, e.g. “look-elsewhere effect”

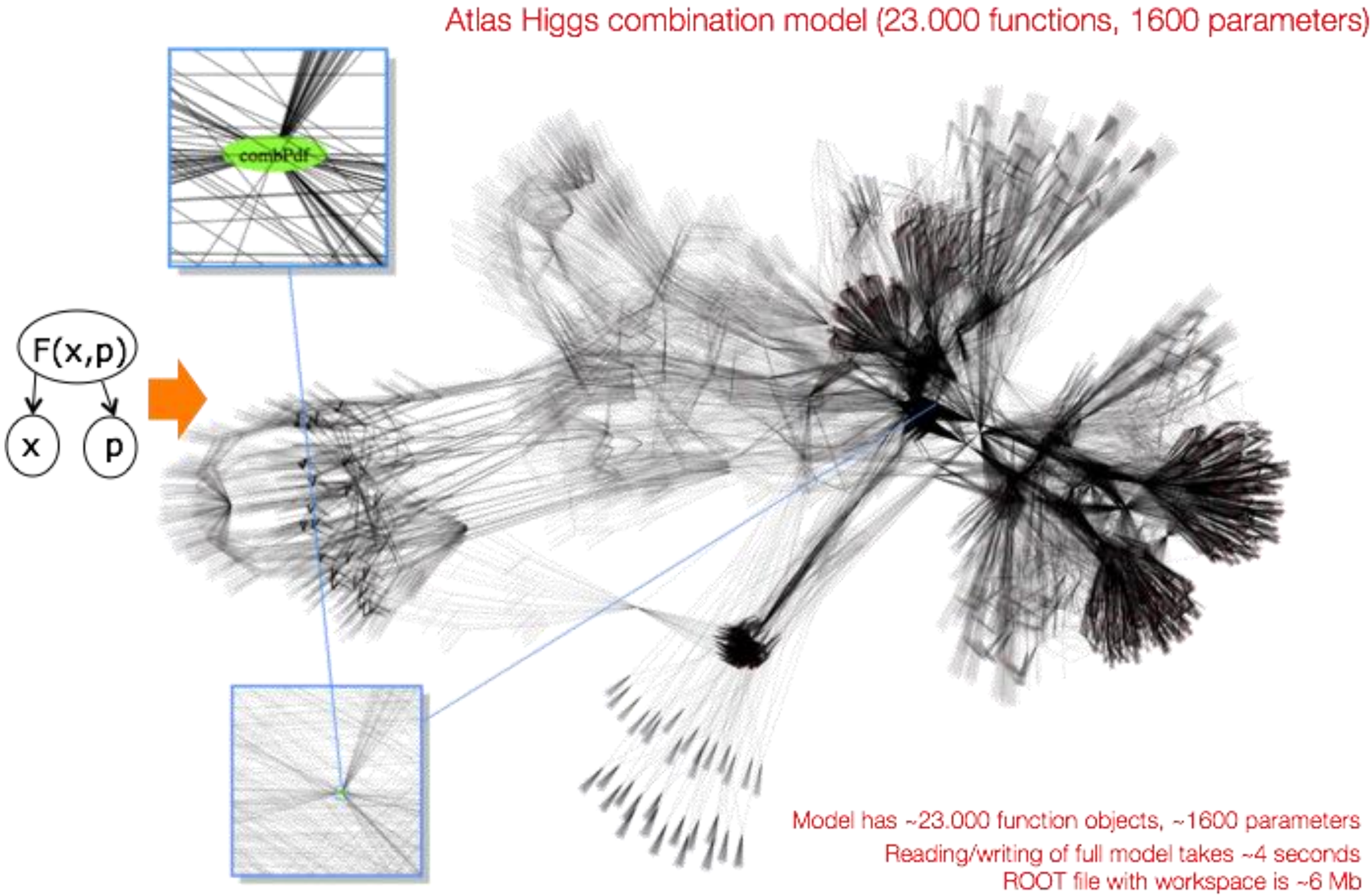
Minimisation

- Central part of multi-parameter statistics (and fitting etc) is minimisation
- High Energy Physics does minimisation since many decades
- Traditional minimisers: Minuit, simplex, GNU scientific library (e.g. simulated annealing)
- Multivariate algorithms (simple Fisher to genetic, kD-trees, boosted decision trees)
- Contributions, e.g. CMA-ES minimiser

Finding the Higgs



RootFit - The Full ATLAS Higgs Combination Model



Contributions to ROOT?

- ROOT is both monolithic and modular, similar to Linux kernel:
  - Minimal prerequisites
  - Large amount of optional dependencies
- Bindings to databases, math libraries, graphics packages
- Domain-specific extensions, e.g. astronomy (FITSIO), CAD, filesystems / storage

Using ROOT

- GUI
- As set of libraries in your application
- Interactive:
  - ROOT has C++ interpreter
  - Python prompt, uses C++ interpreter behind the scenes to be completely dynamic

Conclusion

All Our Data is Big Data

- Our data is Real Big Data
- Real Big Data is an HPC problem
- Have to use as efficient as possible all available computing resources
- No country for Java and Python
- Give ROOT a try
  - It’s powerful, maybe a bit frightening at first :-)