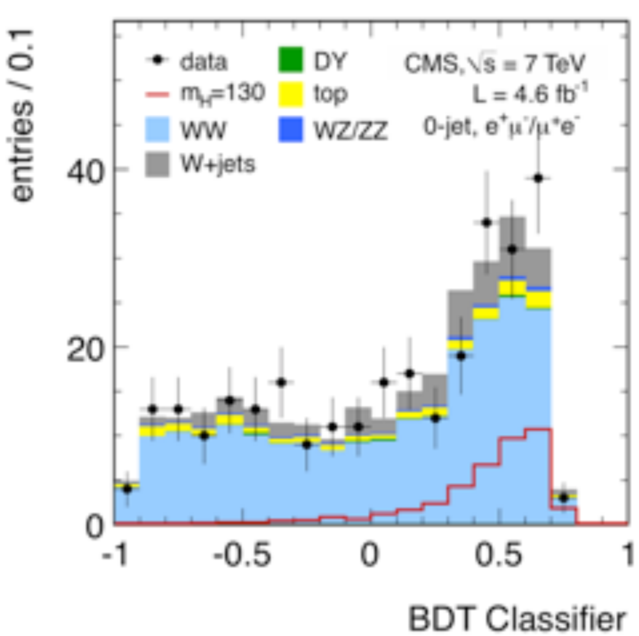
Review of HEP Analysis Strategies

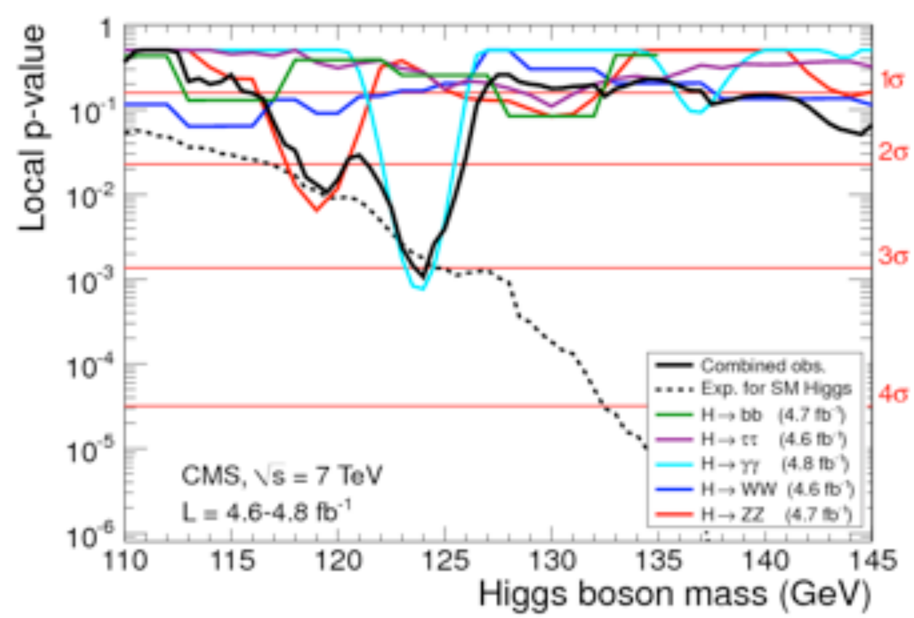
Markus Klute
Massachusetts Institute of Technology
May 22nd, 2012

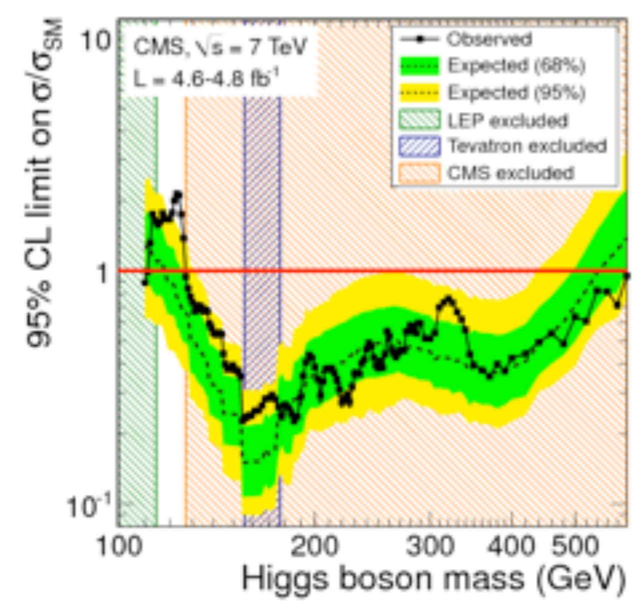
CHEP 2012
Computing in High Energy and Nuclear Physics 2012 • New York • United States

HEP Data Analysis

- Analysis
  - statistical interpretation of an ensemble of events collected in a HEP experiment
  - typically, we have a model for signal and background
  - extract properties from statistical analysis
- Examples
  - OPAL, AMS, LHCb and CMS
  - search for the Higgs boson in CMS



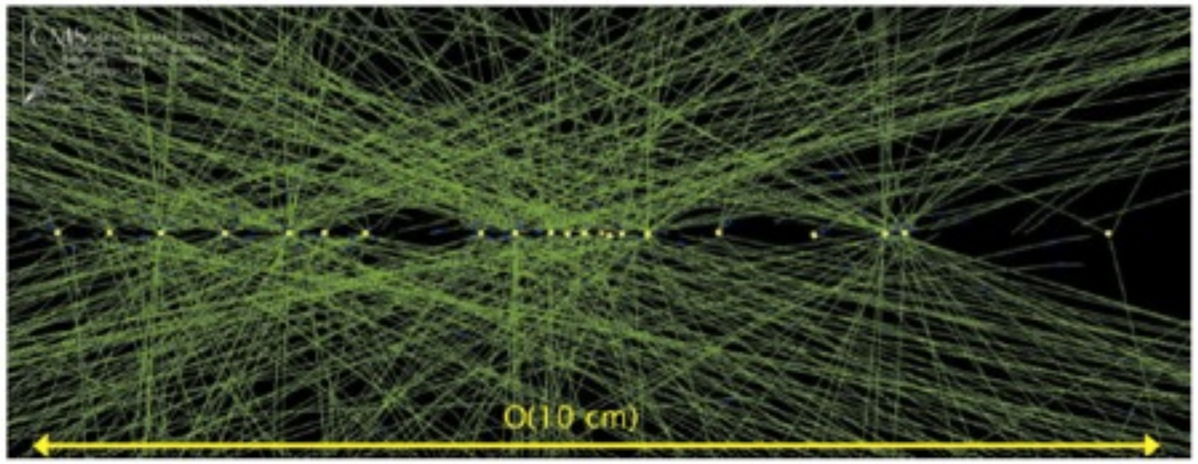




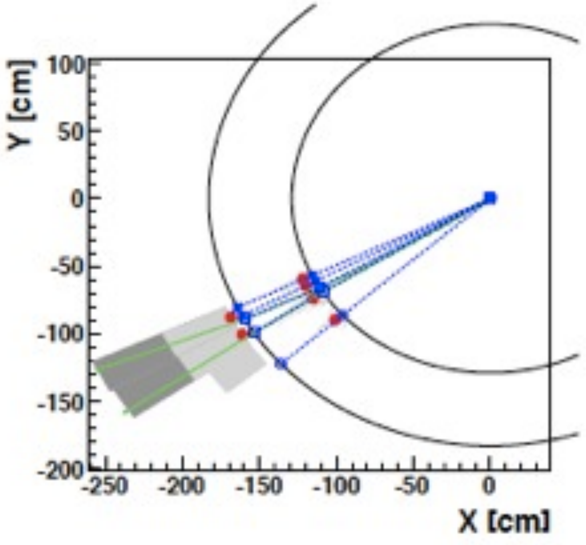
HEP Event

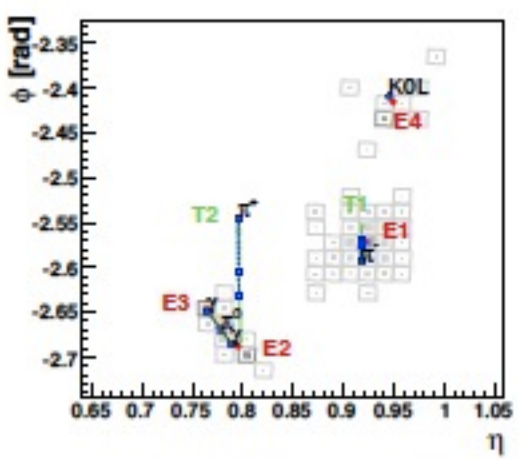
- Event
  - readout cycle of the experiment
    - bunch crossing in accelerator structure
    - interaction of particle with detector
  - event consists of sub-detector measurements
  - particles are identified in a reconstruction step

pile-up

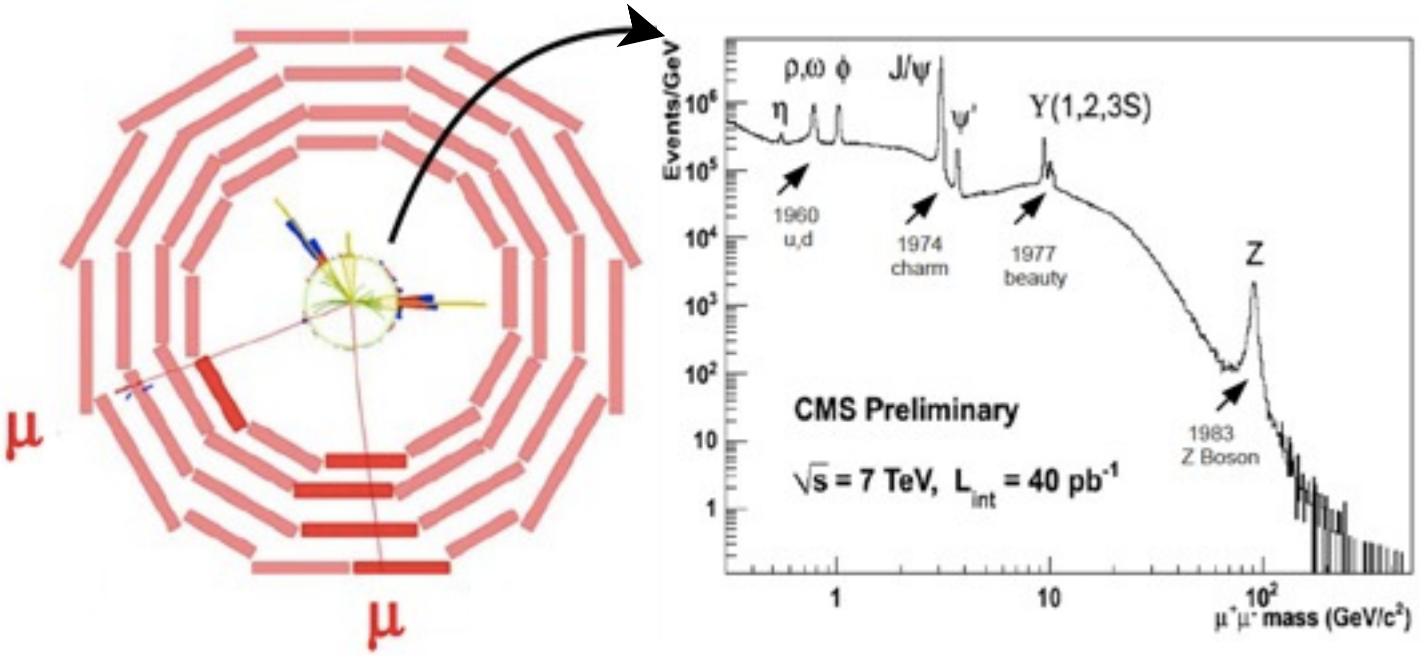


event reconstruction

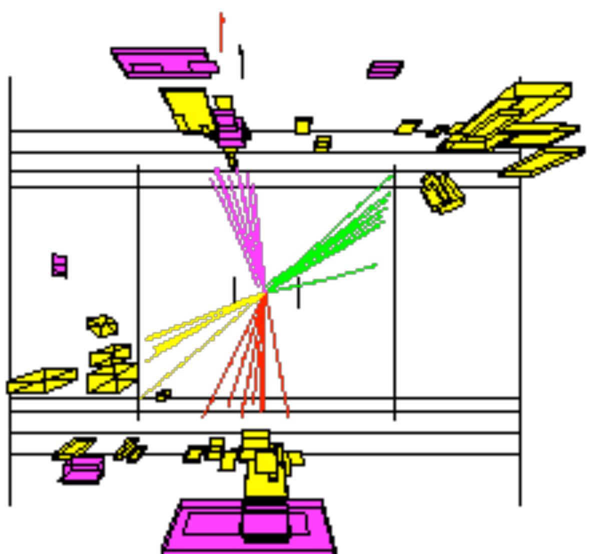
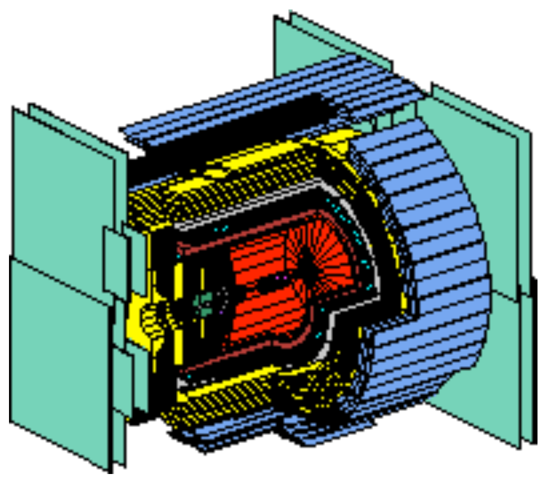




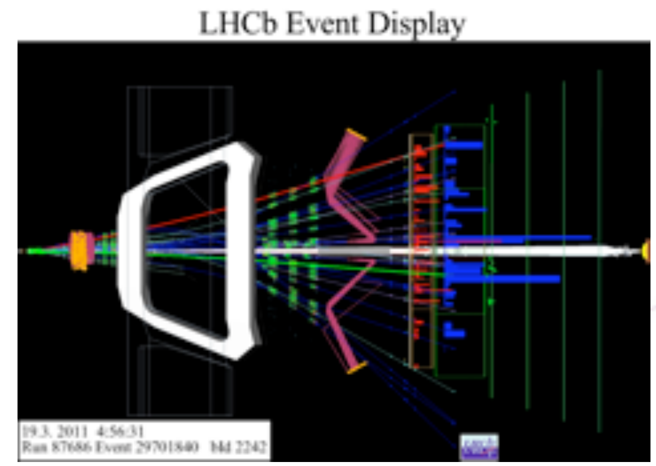
event and event sample

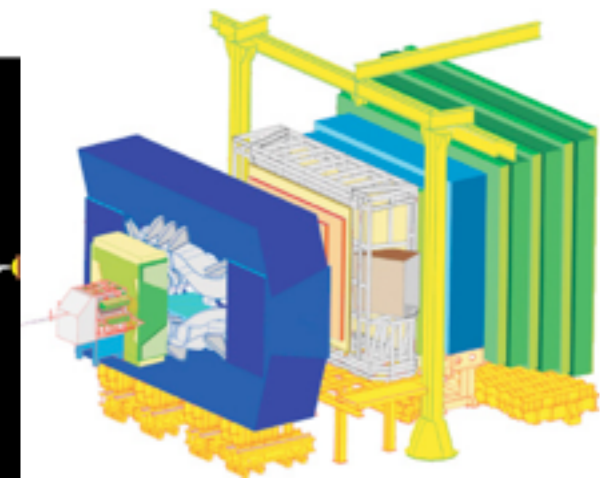


Event Samples and Analysis Infrastructure

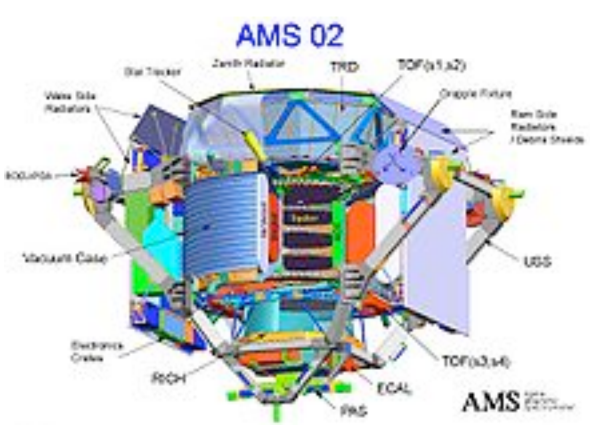



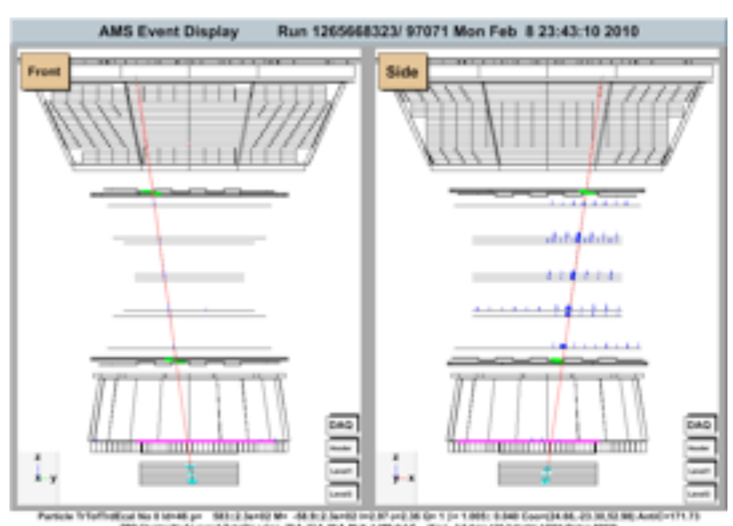
OPAL at LEP



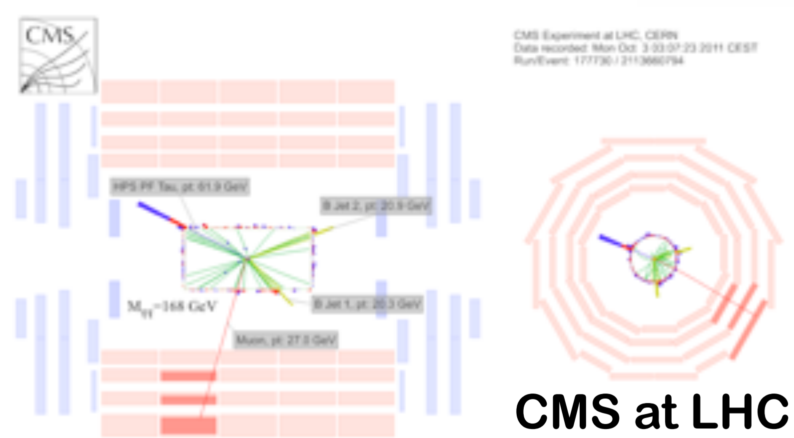


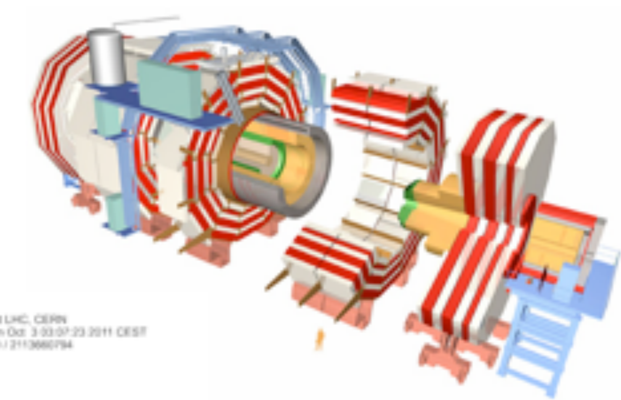
LHCb at LHC





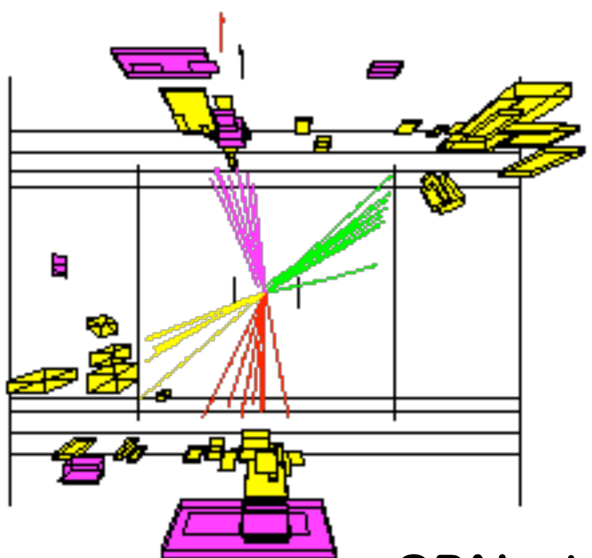
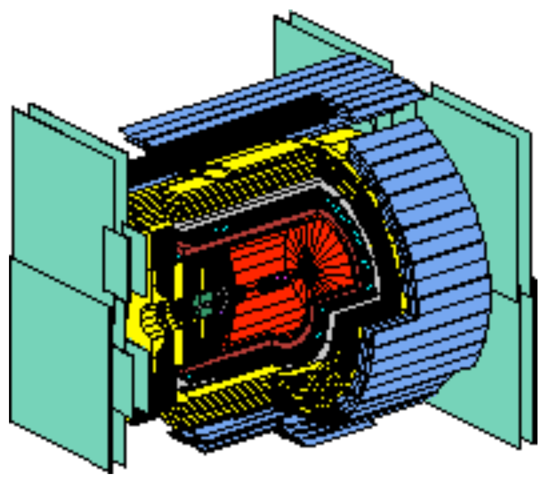
AMS at ISS





CMS at LHC

Example: OPAL at LEP

OPAL at LEP

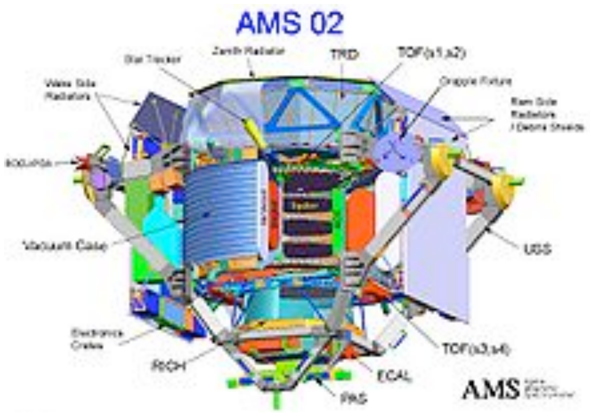
- OPAL (and the other LEP exp.) used analysis frameworks based on PAW
- Analysis software mainly in FORTRAN
- GEANT detector simulation
- HEP community transitioned to ROOT/C++ and other object-oriented programming at the end of the experiment lifetime

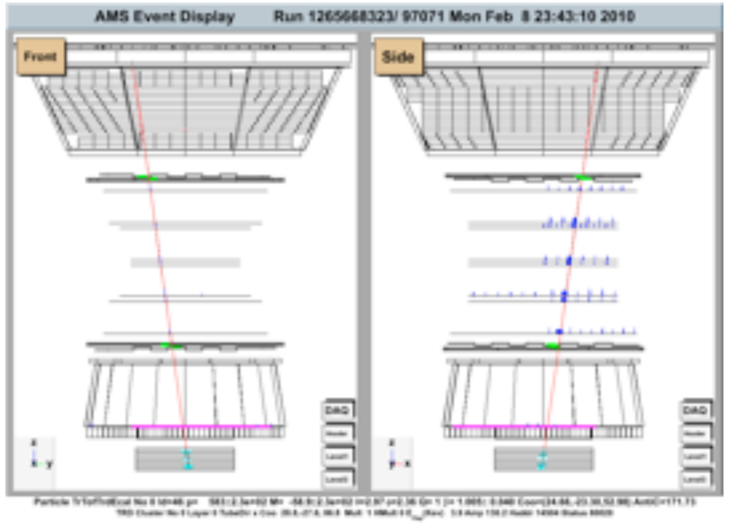
- OPAL in 1991
  - trigger rate 4Hz
  - event size 12-40 kB
- OPAL total data volume
  - reconstructable dataset ~2 TB
  - analysis dataset ~300 GB
- Storing and reconstructing OPAL data in 1991 was a challenge
- Parallel event processing on "computing farm"
- Trivial with today's technology
- Event reconstruction took 30-60s at the time (on a 17SPECMark CPU)

Example: AMS at the ISS

- Alpha Magnetic Spectrometer Experiment on the International Space Station
- Computing challenge is the bandwidth limitation of 10Mb/s
- Detector output of ~7Gb/s is reduced using dedicated algorithms (zero suppression)
- Detector can buffer ~one week of data
- Primary data archive with 2 month buffer using laptop on ISS
- Data is transmitted using satellites to White Sands Ground terminal and via Huntsville to the AMS center at CERN
- AMS produces ~36TB of raw data per year
- Analysis framework is based on ROOT

Astronaut Don Pettit installs hard drive on AMS laptop



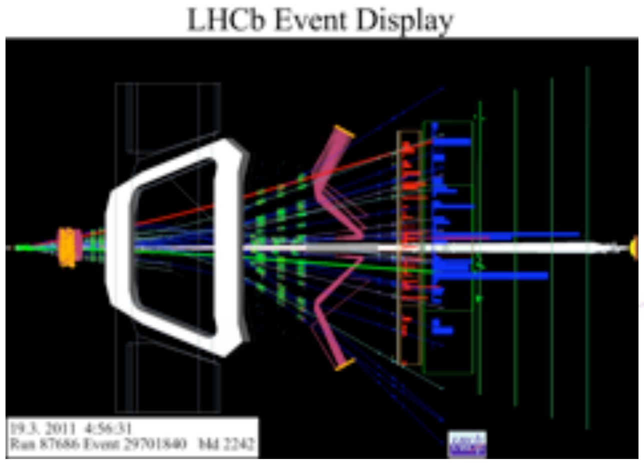


AMS at ISS

special thanks to Paolo Zuccon for the content of this slide

Example: LHCb at the LHC

- LHC used luminosity leveling to control pile-up conditions
- Nominal interaction rate in ~20 MHz
- Reduced to 1MHz by L0 hardware trigger
- High level trigger farm (26k procs) reduced the rate to ~3.5kHz
- RAW event size ~50kB, resulting in ~1PB/y
- USER defined pre-selections reduce event sample to 10%
- Pre-selections are updated 2-3 times per year
- 150kB per reconstructed event. A micro format of 10kB per events used for large selections



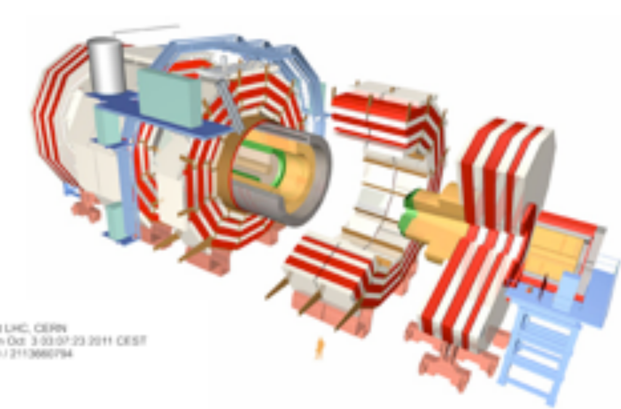
LHCb at LHC

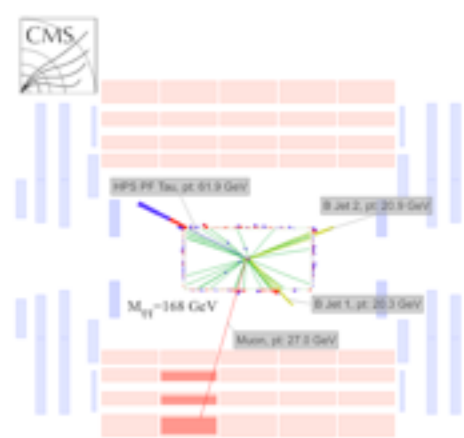
- Software framework Gaudi also used by other experiments
- GEANT4 detector simulation
- Analysis framework based on ROOT

Example: CMS at the LHC

- Reduce events rates using a hardware trigger (100kHz) and high level trigger farm (300Hz)
- Event samples are split in streams of ~50 Hz or less based of trigger signatures (e.g. DoubleMu)
- Event size for data
  - RAW (460kB), RECO (680kB), AOD (230kB)
- Total volume of data and MC stored on tape ~30PB
- 2011 AOD samples ~700TB. Multiple copies needed to support analysis
- USER access data (AOD) via grid submission
- GEANT4 detector simulation
- Analysis framework based on ROOT

Individual grid user for analysis jobs





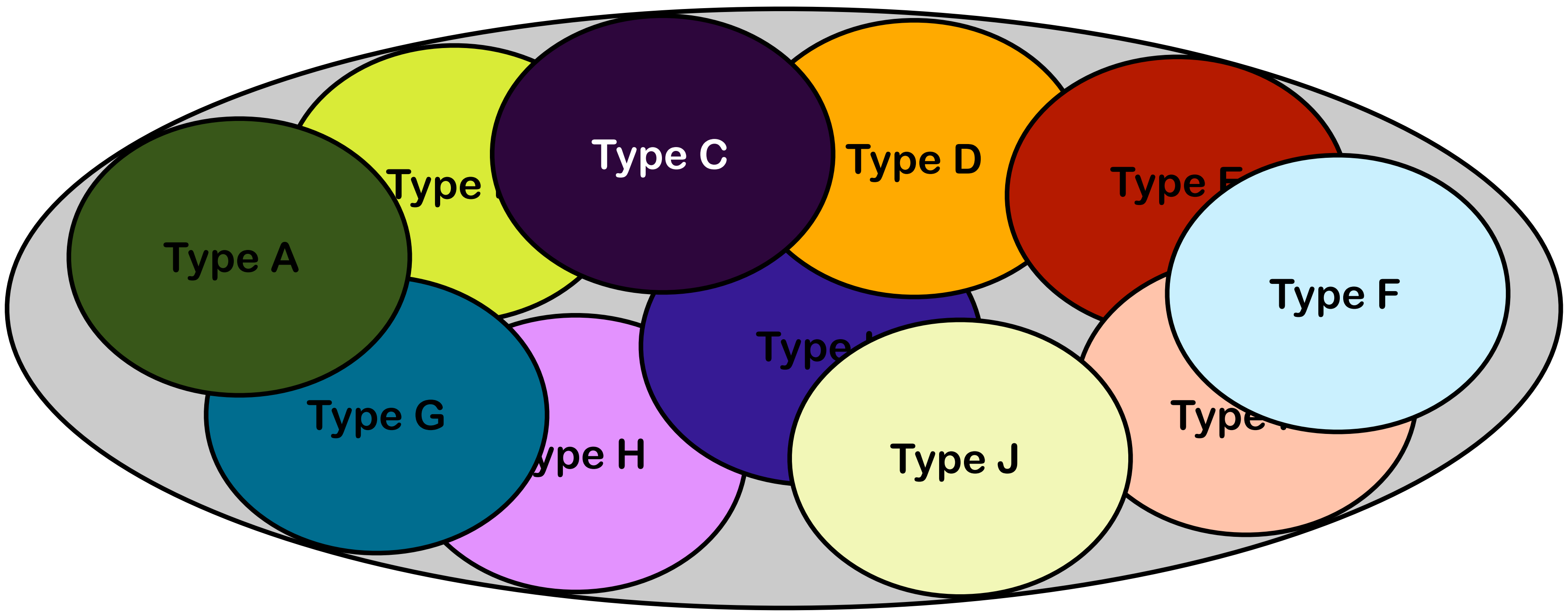
CMS at LHC

Event Samples and Analysis Infrastructure

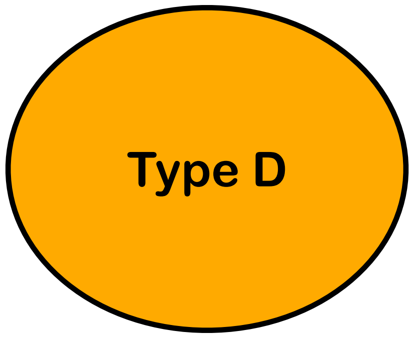
- Prerequisite for an effective analysis is the availability of datasets of manageable size
- Data reduction by data format
  - RAW → (CPU intensive) RECO format → Analysis format → USER defined → more USER defined
- Data reduction by event selection
  - reduce volume for analysis use
  - selections are applied at all data format levels
- Both methods have pros and cons
  - typically trade CPU and flexibility with storage
- Working point for an experiment (collaboration) needs optimization
  - Guideline (stating the obvious)
    - samples needed many times have to be small and always available
    - trade between event count and size / information
    - samples needed rarely can be large (RECO samples)
    - calculate derived quantities to minimize storage requirements, e.g. jets, electrons, missing ET

Event Samples and Analysis Infrastructure

- Data reduction by event selection
- Classify events in stream/skims which use a well defined event selection



- Subset of events contains all events need to perform an analysis
- Volume of events sample of "type D" can be much reduced wrt the total volume



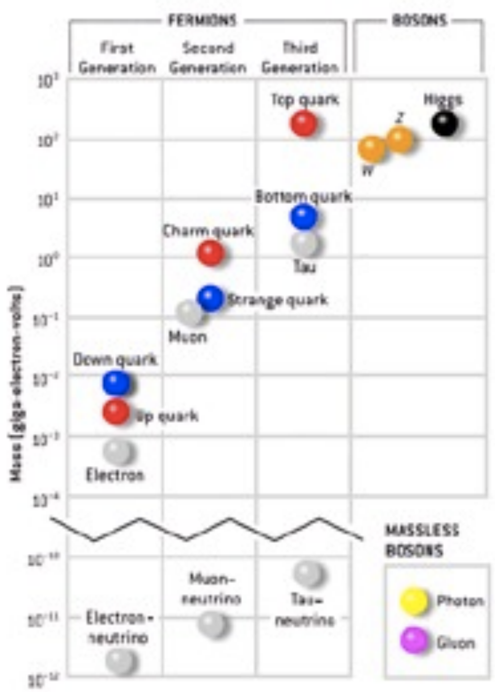
e.g. /DoubleMu

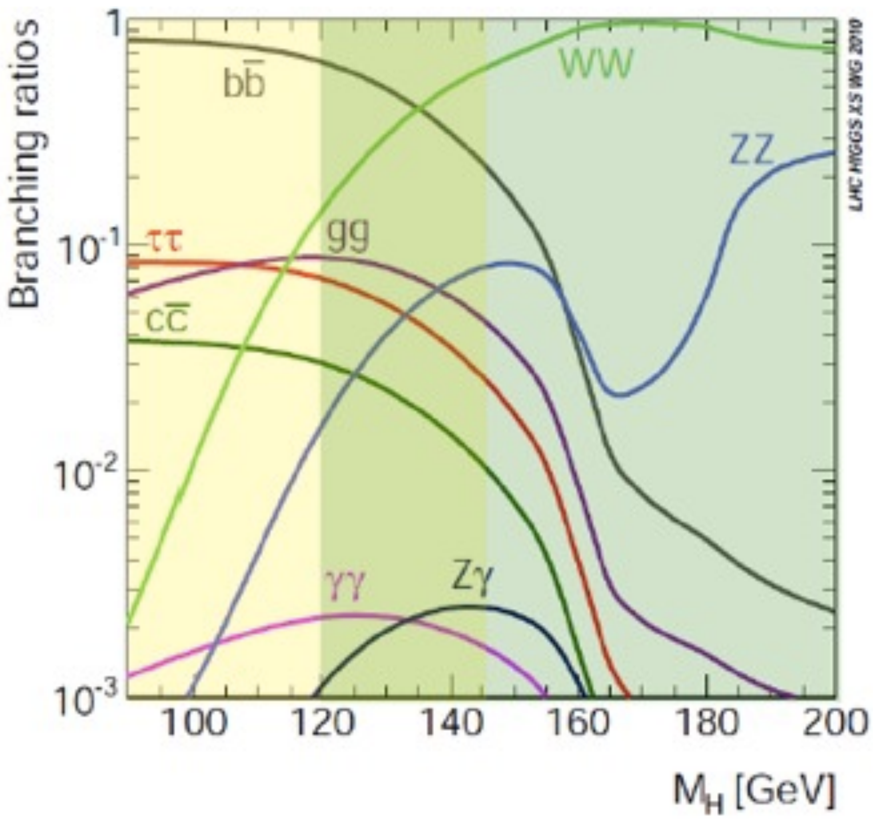
Event Samples and Analysis Infrastructure

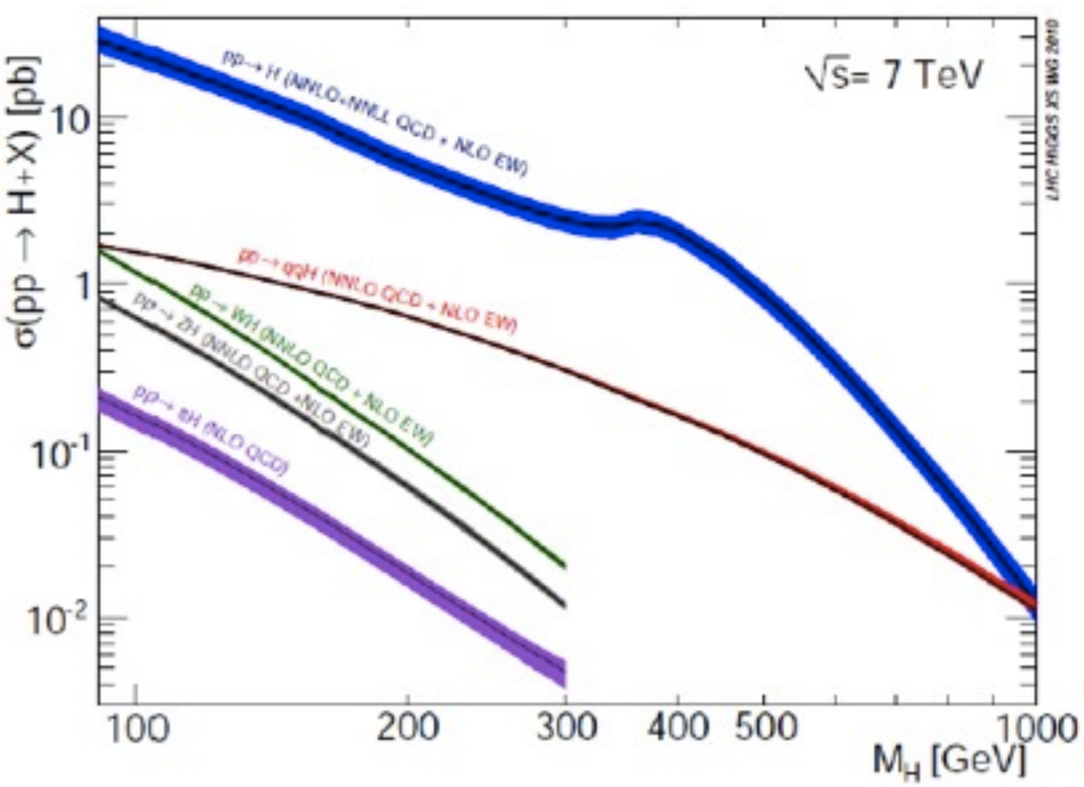
- Experiments adopted computing / analysis strategy according to given challenges
- Main concepts a very similar across experiments with strong links between computing and analysis models
- Large scale experiments profit from improvements in computing technologies (grid)
- Standardized and collaborative tools
    - modeling of detectors with GEANT
    - physics using Monte Carlo generators with standardized interfaces
    - ROOT is main analysis tool and includes classes for statistical analysis

Example Analysis: H → WW in CMS

- Study the origin of electroweak symmetry breaking
  - how do W and Z bosons acquire mass?
  - can we explain fermion masses?
  - Higgs mechanism give answers
  - new particle is proposed in the SM: the Higgs Boson
    - mass is free parameter
    - all other properties are predicted in the SM
- Search for the Higgs Boson at the LHC

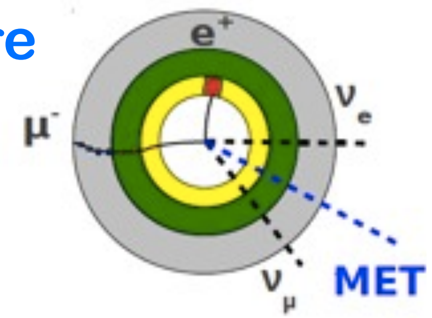


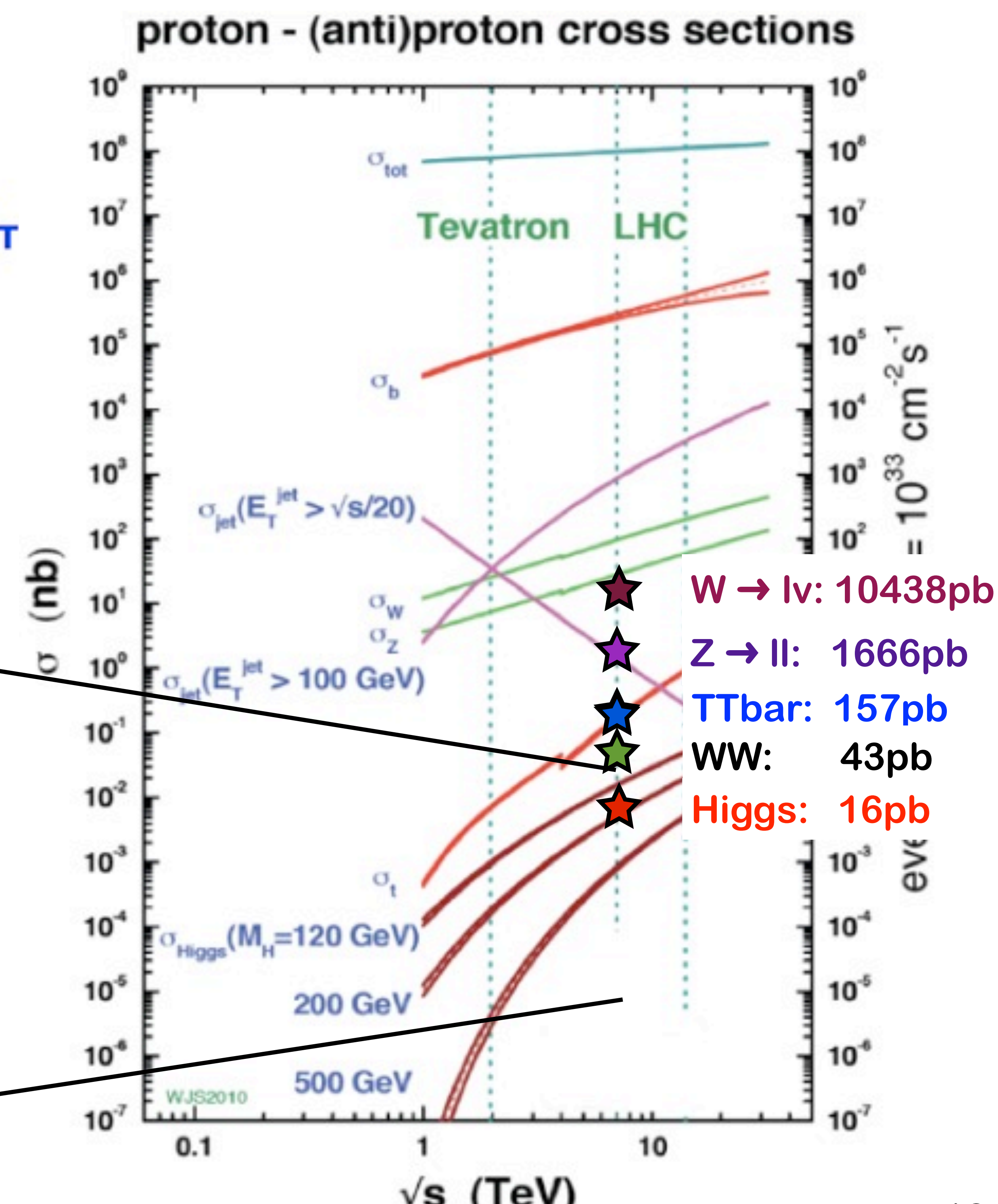


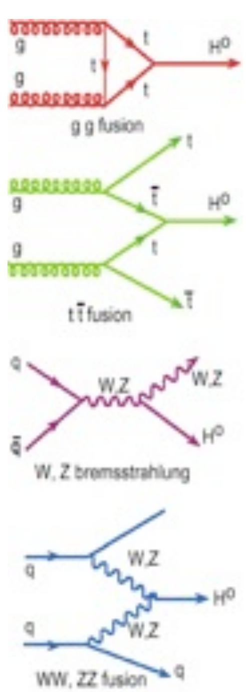


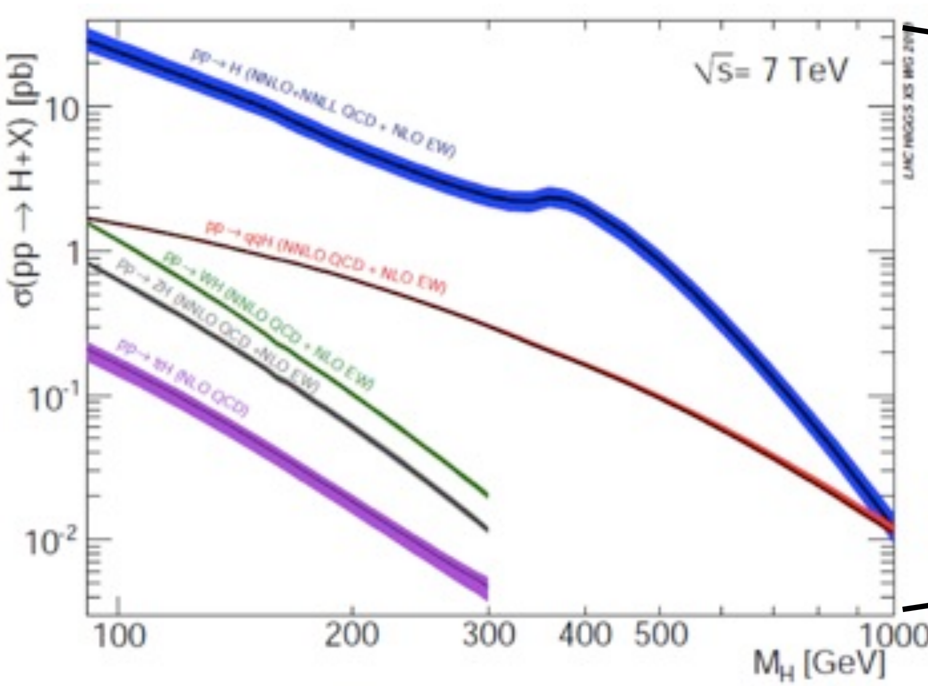
Example Analysis: H → WW in CMS

- Clean experimental signature
  - 2 leptons + 2 neutrinos
  - No mass peak
  - analysis is all about the background estimation
  - backgrounds have large cross sections



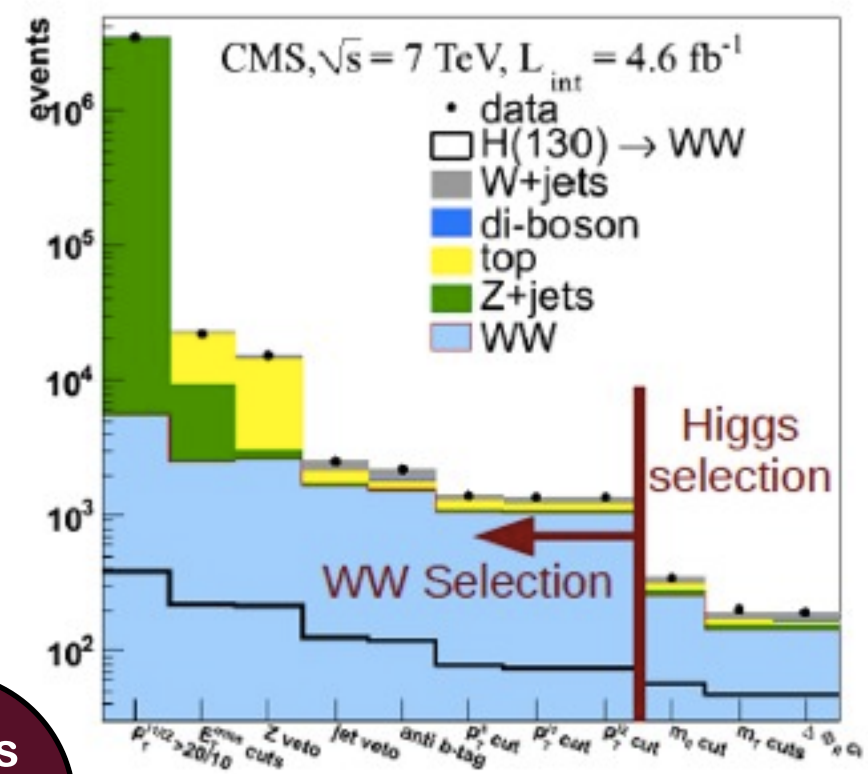


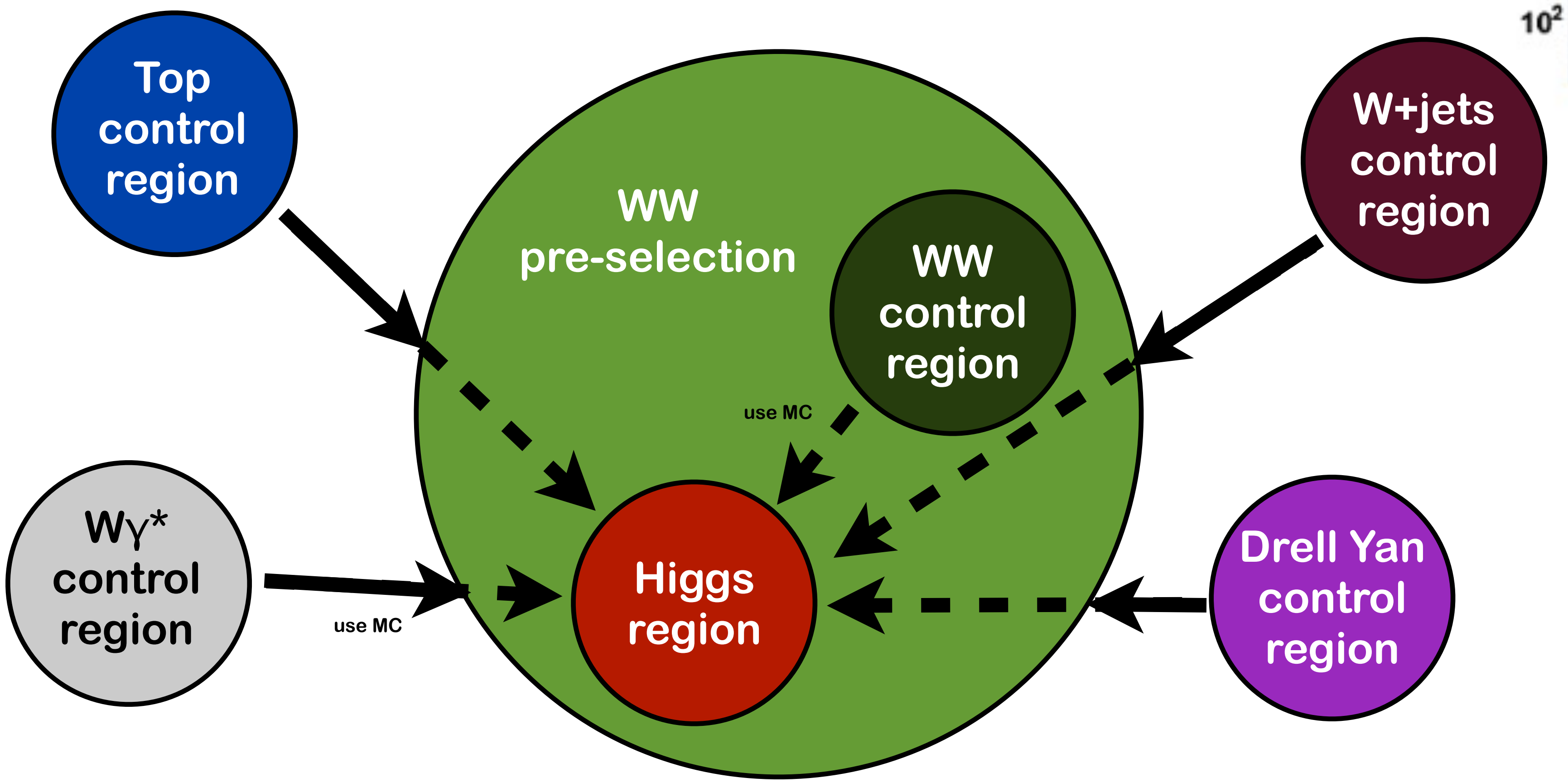




Example Analysis: H → WW in CMS

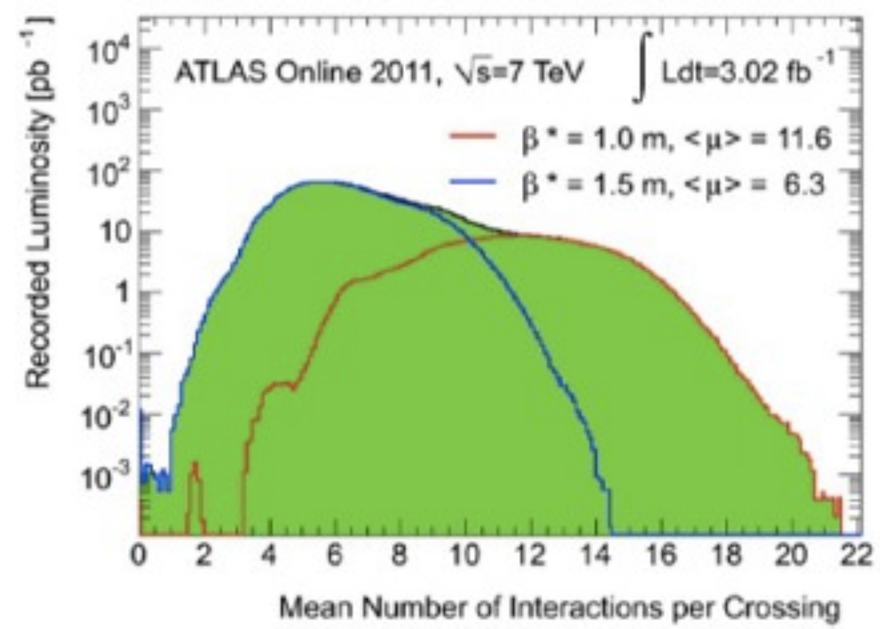
- Analysis strategy
  - WW pre-selection
    - establish WW signature
    - data driven estimates of main backgrounds
  - Higgs selection
    - discriminate Higgs against WW background
    - cut-based selection or multivariate analysis discriminator



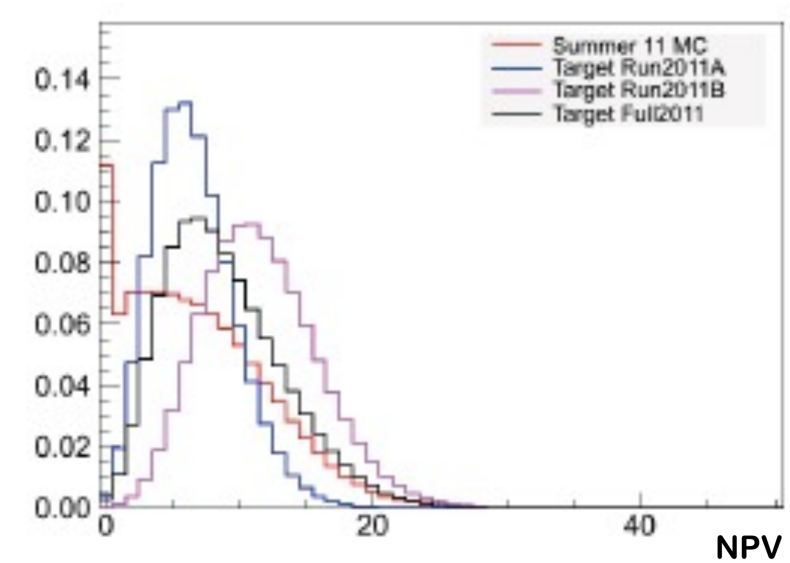


Example Analysis: H → WW in CMS

- Data streams and reduction in analysis flow
  - /DoubleMu 63M events
  - /DoubleElectron 21M events
  - /MuEG 25M events
  - ~3.5M after pre-selection
  - ~100 events after WW selection
- Monto Carlo needs
  - 20M signal events
    - produced for 25 mass points from 110 GeV < mH < 600 GeV
    - POWHEG and PYTHIA
  - 110M background events
    - MADGRAPH, PYTHIA and POWHEG
  - MC generators are interfaced in software framework or used via standard LHE format
- Pile-up (in-time and out-of-time)
  - large effects on measurements of missing transfers energy, isolation of leptons, jet measurements, etc.
  - distribution unknown a priori
  - re-weighting of MC is required to match collision data
    - MC samples have been (digitized and) reconstructed using two different PU scenarios
- Beam spot
- Center-of-mass energy



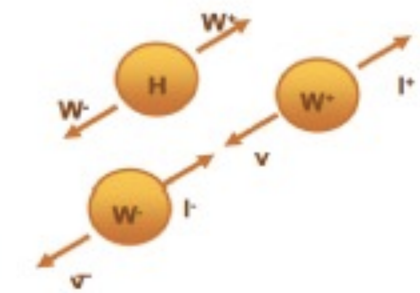
LHC beam parameter modified in Sep. 2011



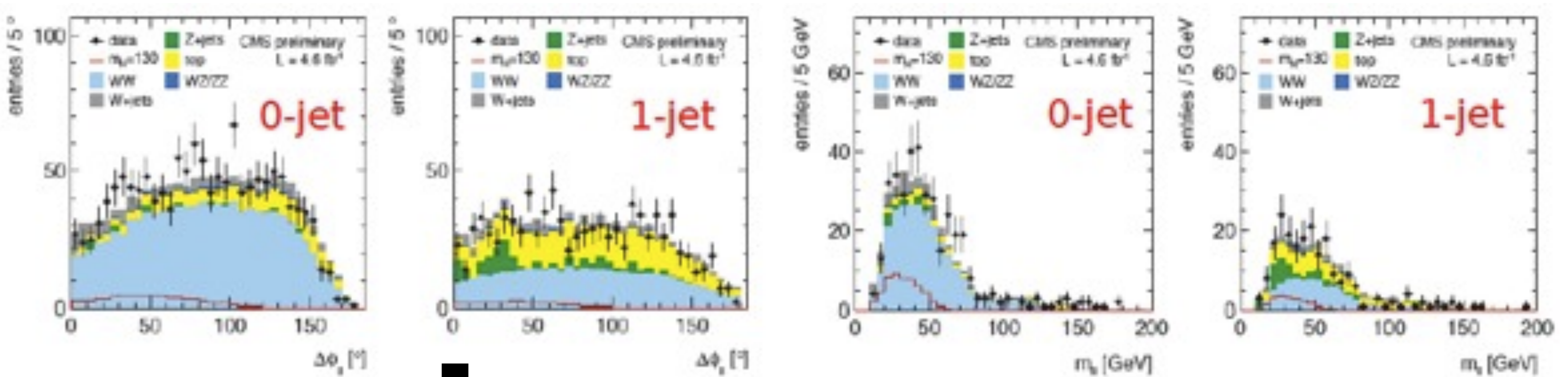
PU re-weighting

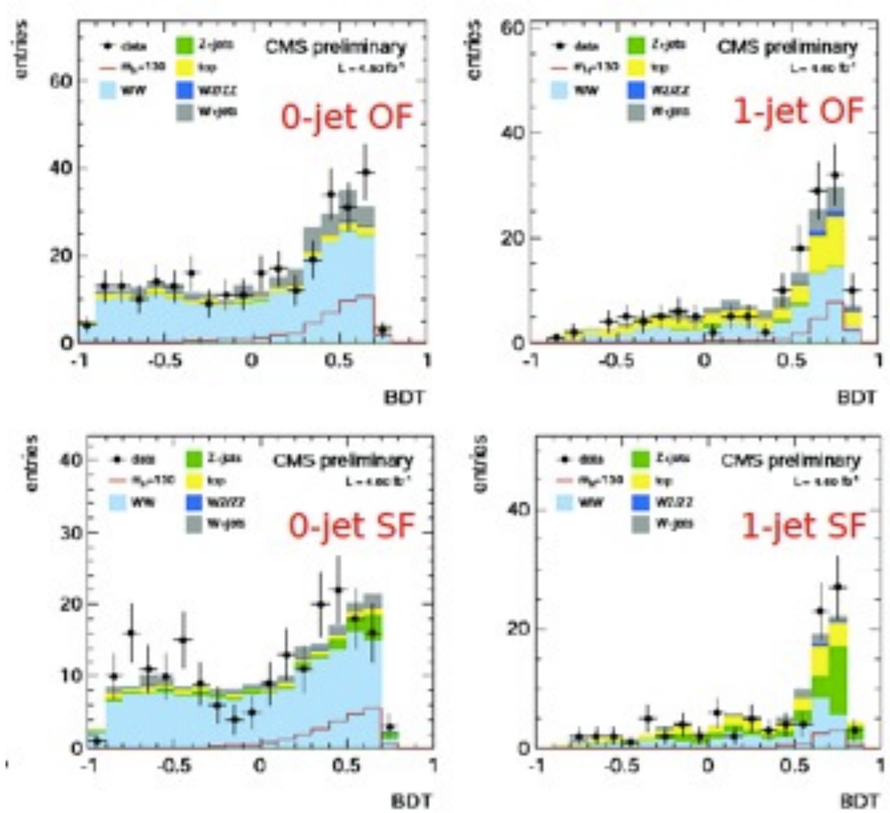
Example Analysis: H → WW in CMS

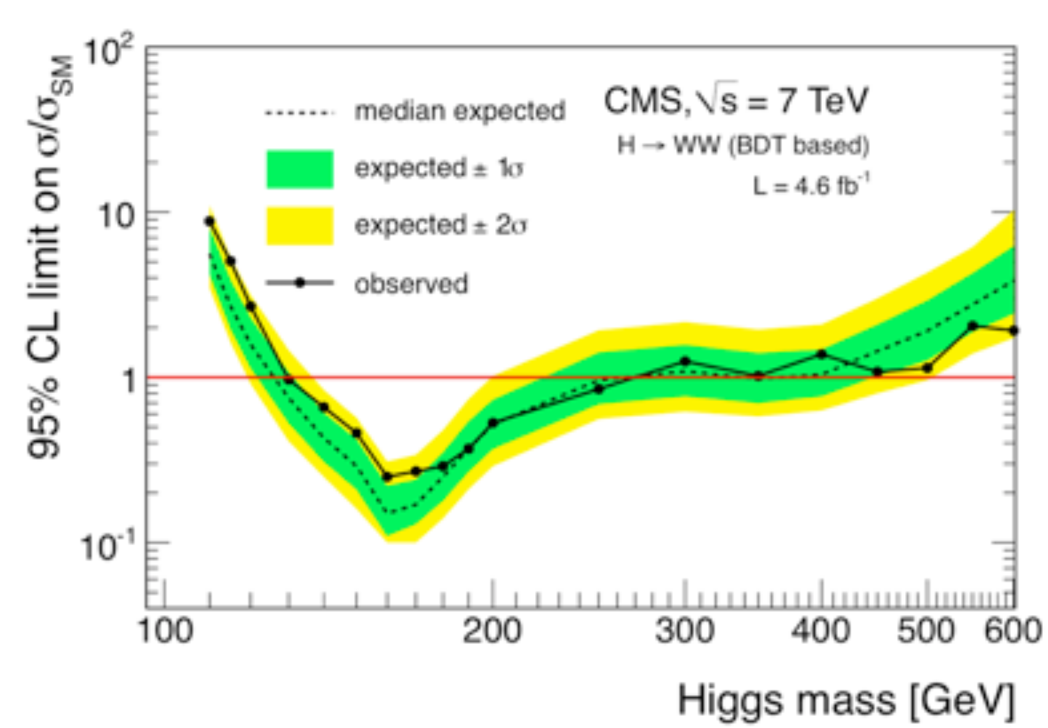
- Boosted decision tree explores event kinematics
  - Multivariate classifier are also used in lepton selection
- Classification by # of jets and lepton flavor explore regions with varying S/B



Higgs is scalar boson







Expected range: 127 < M_H < 270 GeV
Observed range: 129 < M_H < 270 GeV

Statistical Interpretation

- General framework
  - signal strength modifier $\sigma = \mu \cdot \sigma_{\text{SM}}$.
  - nuisance parameter θ_i
  - likelihood $\mathcal{L}(\text{data} \mid \mu \cdot s(\theta) + b(\theta)) = \mathcal{P}(\text{data} \mid \mu \cdot s(\theta) + b(\theta)) \cdot p(\tilde{\theta}|\theta)$
  - construct test statistic
- Quantify an excess

$$q_0 = -2 \ln \frac{\mathcal{L}(\text{data} \mid b(\hat{\theta}_0))}{\mathcal{L}(\text{data} \mid \hat{\mu} \cdot s(\hat{\theta}) + b(\hat{\theta}))}, \qquad \hat{\mu} \geq 0,$$

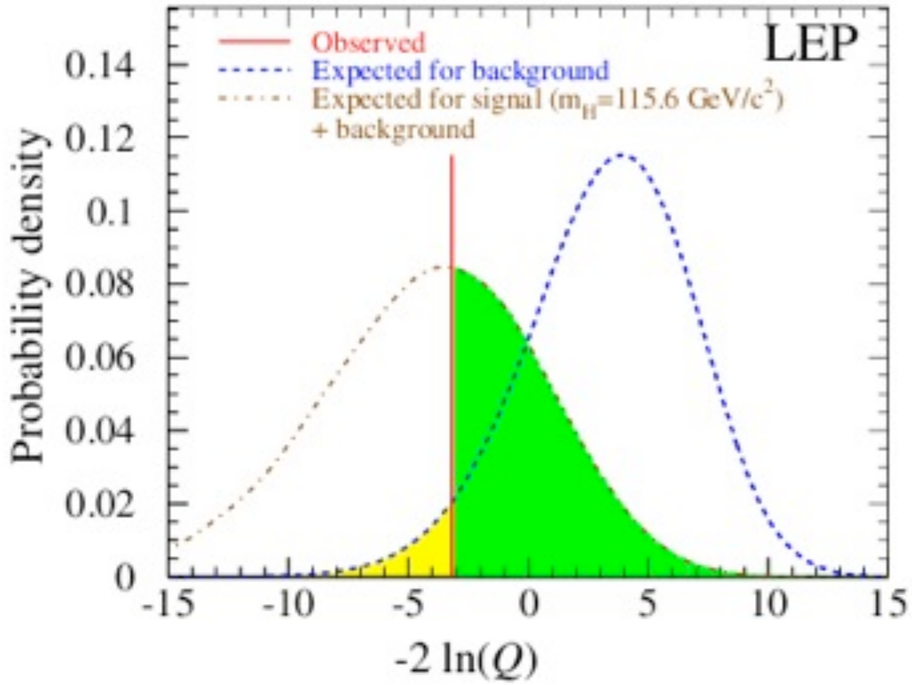
- Quantify the absence of a signal

$$q_\mu = -2 \ln \frac{\mathcal{L}(\text{data} \mid \mu \cdot s(\hat{\theta}_\mu) + b(\hat{\theta}_\mu))}{\mathcal{L}(\text{data} \mid \hat{\mu} \cdot s(\hat{\theta}) + b(\hat{\theta}))}, \qquad 0 \leq \hat{\mu} < \mu,$$

  - modified frequentist CLs (similar to method used at LEP)

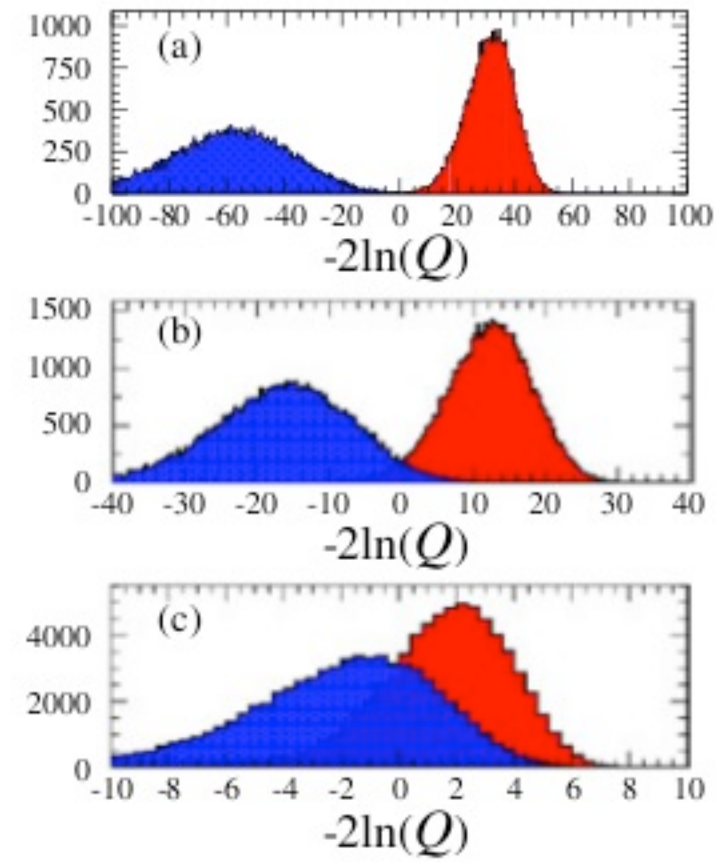
$$\text{CL}_{\text{s+b}} = P\left(q_\mu \geq q_\mu^{obs} \mid \mu \cdot s + b\right),$$
$$\text{CL}_{\text{b}} = P\left(q_\mu \geq q_\mu^{obs} \mid b\right),$$

$$\text{CL}_{\text{s}} = \frac{\text{CL}_{\text{s+b}}}{\text{CL}_{\text{b}}}$$



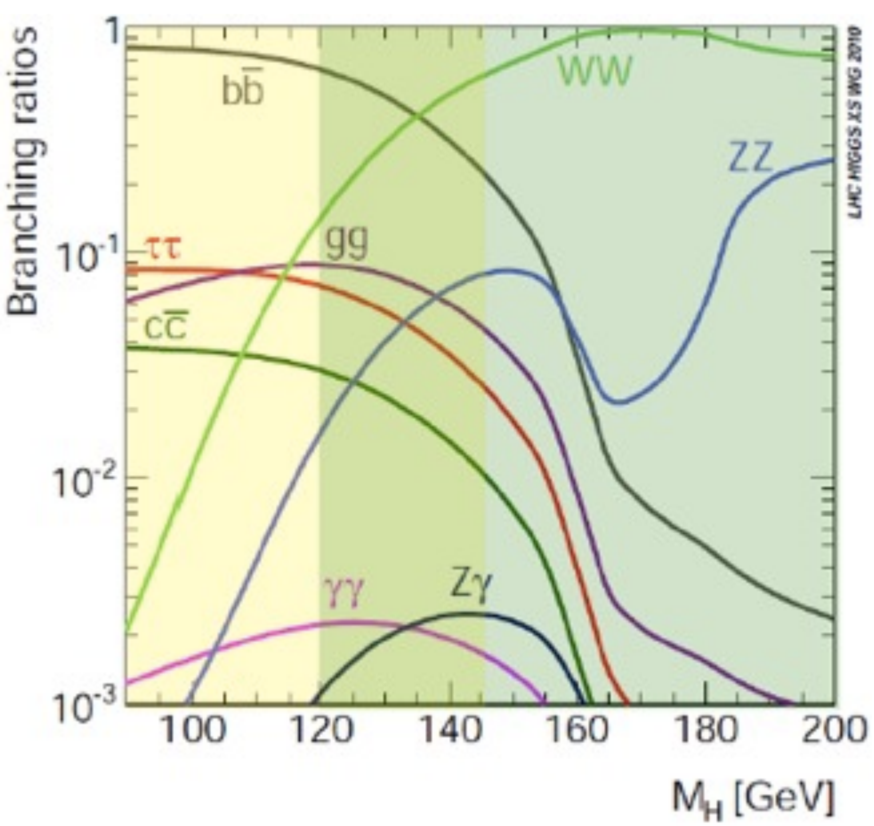
Statistical Interpretation

- If $CL_s \leq \alpha$ for $\mu = 1$, conclude that signal is excluded at $(1-\alpha)$ CL
  - conservative approach given that proper CLs can not be determined w/o knowing the signal beforehand
- CL_{s+b} and CL_b are determined from independent toys
  - O(1000) needed to have a statistical precision of O(1%)
  - toys are numerical integrations of the likelihood functions
- Limit calculation procedure
  - data considered binned or unbinned
  - models can be number (cut & count), shapes or parametrized models that make predictions on presence of data
  - determine q_μ for an ensemble of μ values from a maximum likelihood fit of the model of data and calculate CL_{s+b} and CL_b
  - systematic uncertainties incorporated as nuisance parameters to the fitted likelihood functions L and integrated out
  - 95% CL upper limit
    - determine the value of μ where $CL_s \leq 5\%$



Example: CMS Higgs Combination

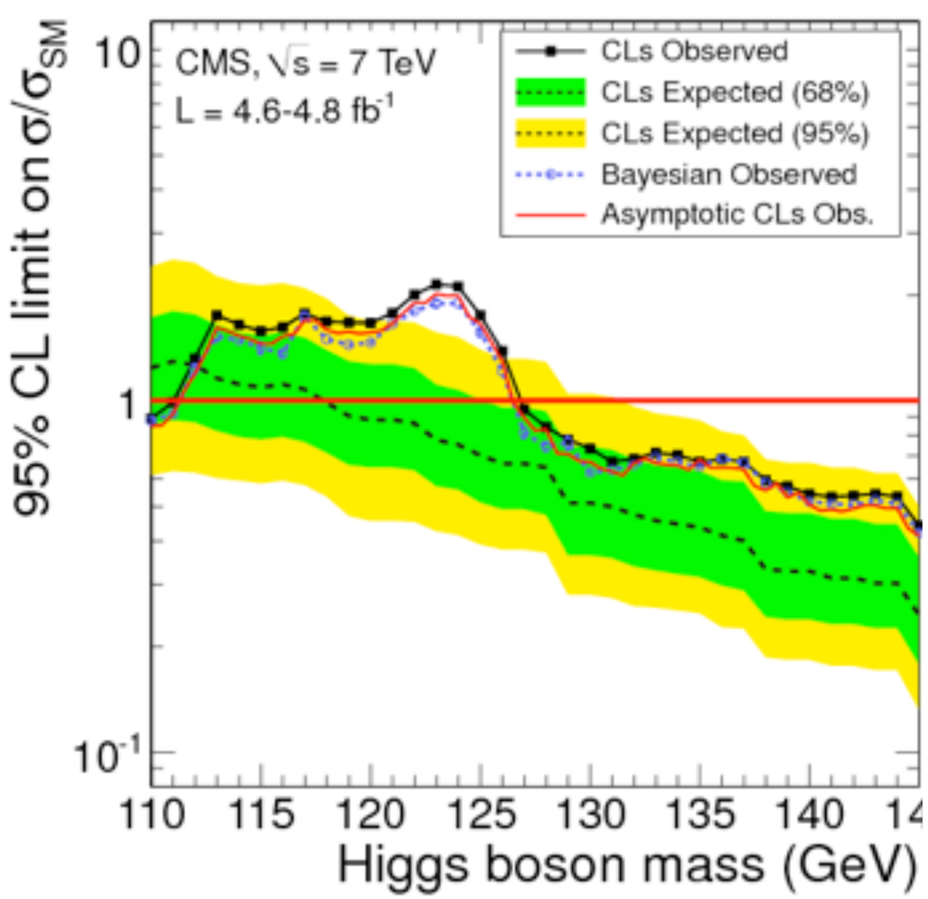
- **Limit calculation in CMS Higgs search**
  - tools based on ROOTs statistic classes, alternatives available and used to cross check results
  - 43 channels (and growing)
  - 156-222 nuisance parameters (depending on tested mass hypothesis)
  - 183 mass hypotheses (from 110 GeV ≤ mH ≤ 600 GeV)

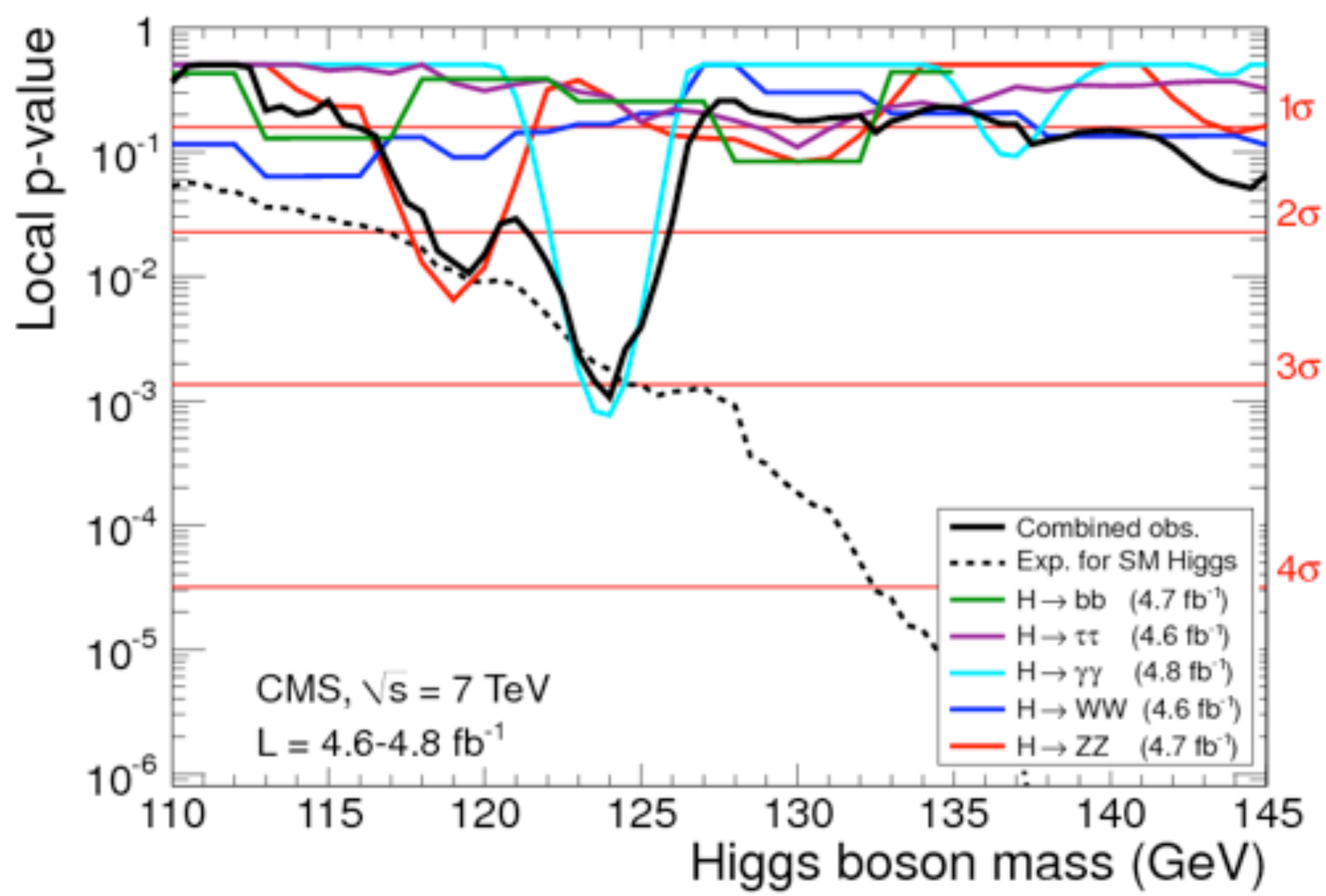


Channel	m_H range (GeV)	Luminosity (fb^{-1})	Sub-channels	m_H resolution	Reference
$H \to \gamma\gamma$	110–150	4.8	5	1–3%	[60]
$H \to \tau\tau$	110–145	4.6	9	20%	[61]
$H \to bb$	110–135	4.7	5	10%	[62]
$H \to WW^* \to 2\ell 2\nu$	110–600	4.6	5	20%	[63]
$H \to ZZ^{(*)} \to 4\ell$	110–600	4.7	3	1–2%	[64]
$H \to ZZ \to 2\ell 2\nu$	250–600	4.6	2	7%	[65]
$H \to ZZ^{(*)} \to 2\ell 2q$	130–164 200–600	4.6	6	3% 3%	[66]
$H \to ZZ \to 2\ell 2\tau$	190–600	4.7	8	10–15%	[67]

Example: CMS Higgs Combination

- CPU requirements
  - toy evaluation for CL_{s+b} and CL_b typically ~1min
  - μ (95% CL) typically determined from O(1000) toys per mass hypothesis and O(10) discrete values of μ
  - results in ~3 CPU years
  - calculation can be parallelized. Submission of O(10000) grid jobs, i.e. results can be obtained within a few days
  - alternative statistical method (asymptotic calculation) can be used to speed up the process





What works, what can be improved

- CMS (and other experiments) are able to perform complex analyses
  - task becomes more complex as we search for rare processes
- Access to large data and MC samples is challenging
  - changing conditions require new productions and reprocessing
  - latencies in production and distribution
  - tails in availability limit the final result
  - data driven techniques limit the dependency on MC
- Computing resources
  - binding of CPU with data location has large overhead and causes inefficiency
- Advanced software frameworks based on common tools
  - GEANT for simulation
  - standard MC generator or interfaces
  - ROOT is main analysis framework
    - includes classes for statistical analysis

Summary

- General characteristics of an analysis in high energy physics
- Event samples in various high energy physics experiments
- Example analysis: $H \to WW$ in CMS
- Statistical interpretation: Combination of Higgs searches in CMS
- What works, what can be improved

- LHC Higgs results will be updated for ICHEP this summer