Physics at the Terascale: Computing Challenges and Solutions at the Large Hadron Collider

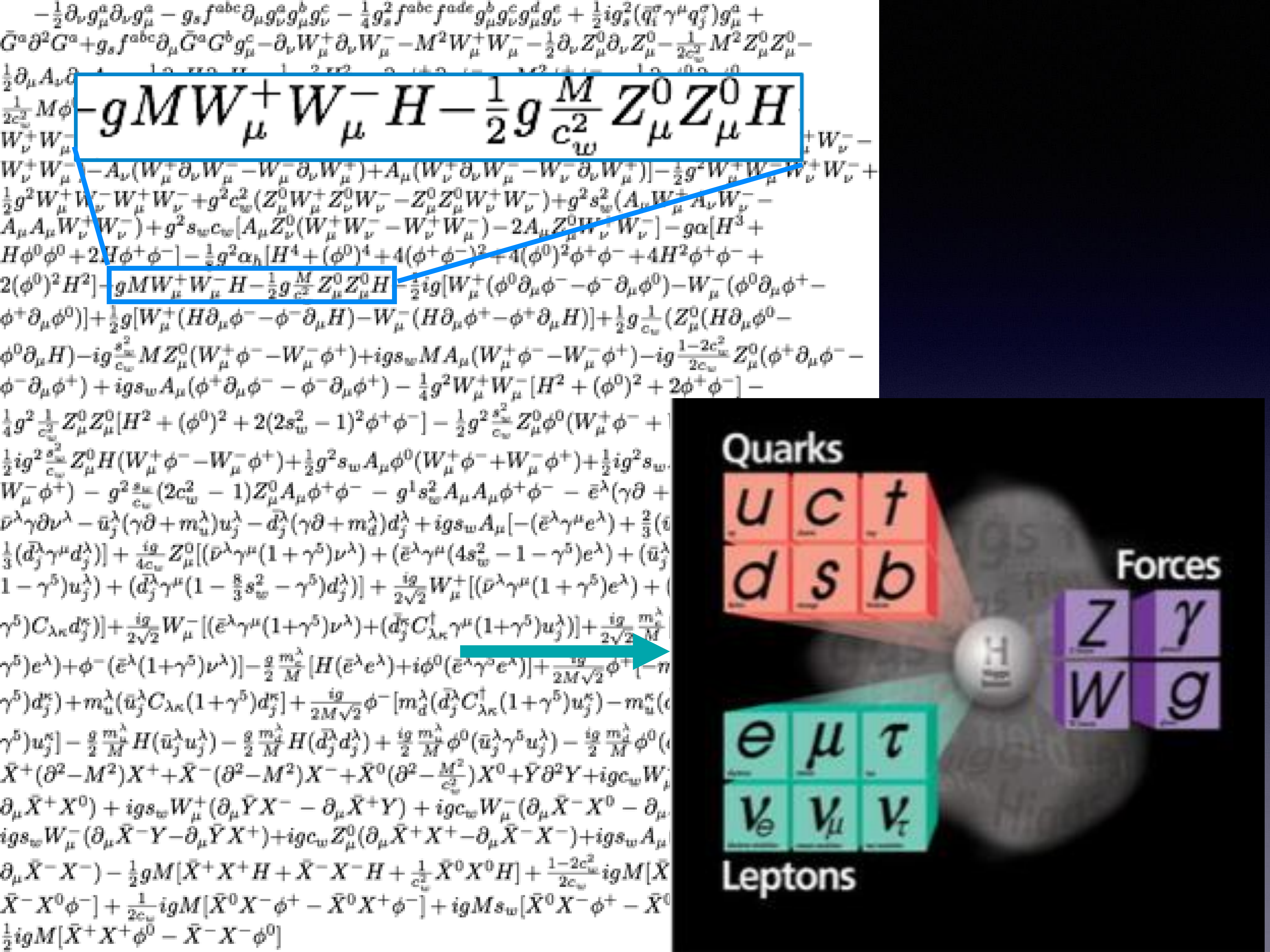
Andrew Melo
Vanderbilt University

About Me

- Nashville Native
  - Went to MLK for High School
- Sewanee C'07 - Computer Science and Physics
- Vanderbilt C’16 - PhD
- Currently a Post-Doc w/Vanderbilt

What’s the Point?

<<enter physics here>>

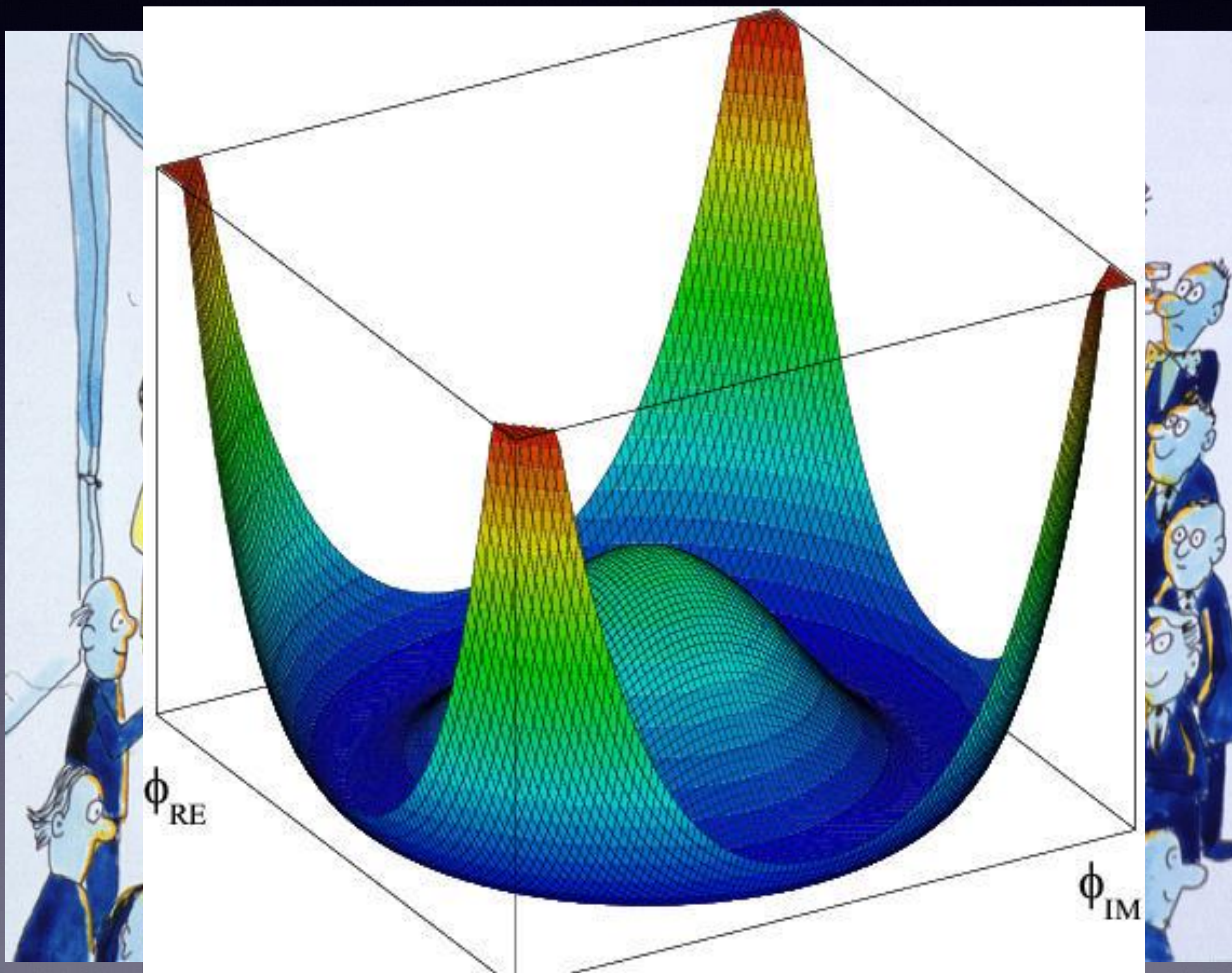


$-gMW_{\mu}^{+}W_{\mu}^{-}H-\frac{1}{2}g\frac{M}{c_{w}^{2}}Z_{\mu}^{0}Z_{\mu}^{0}H$
Quarks
u c t
d s b
Forces
Z γ
W g
H
e μ τ
ν_e ν_μ ν_τ
Leptons

The Higgs Boson

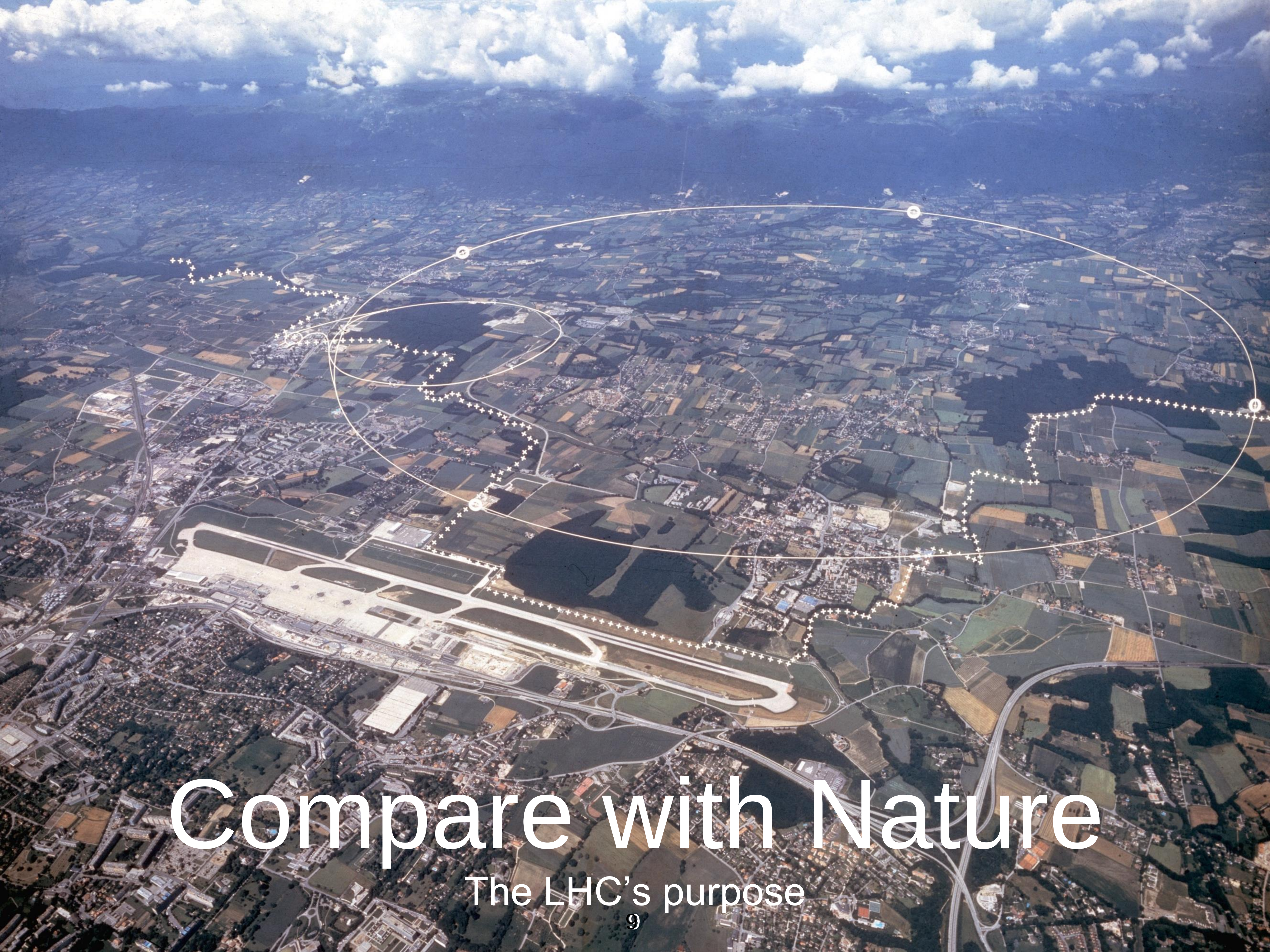
NOT “The God Particle”

What Is the Higgs Field?



Predict effects

- To compare with data, we need to simulate observables the detector would record
- QCD and QED don’t have closed solutions
- To get enough statistics, we need O(1-100M) events
- Each event takes ~<u>1min</u> to simulate
  - It gets worse…(later)
- There are ~100 different backgrounds that need to be simulated
  - Plus upgrades for different detector alignments…



Compare with Nature

The LHC’s purpose

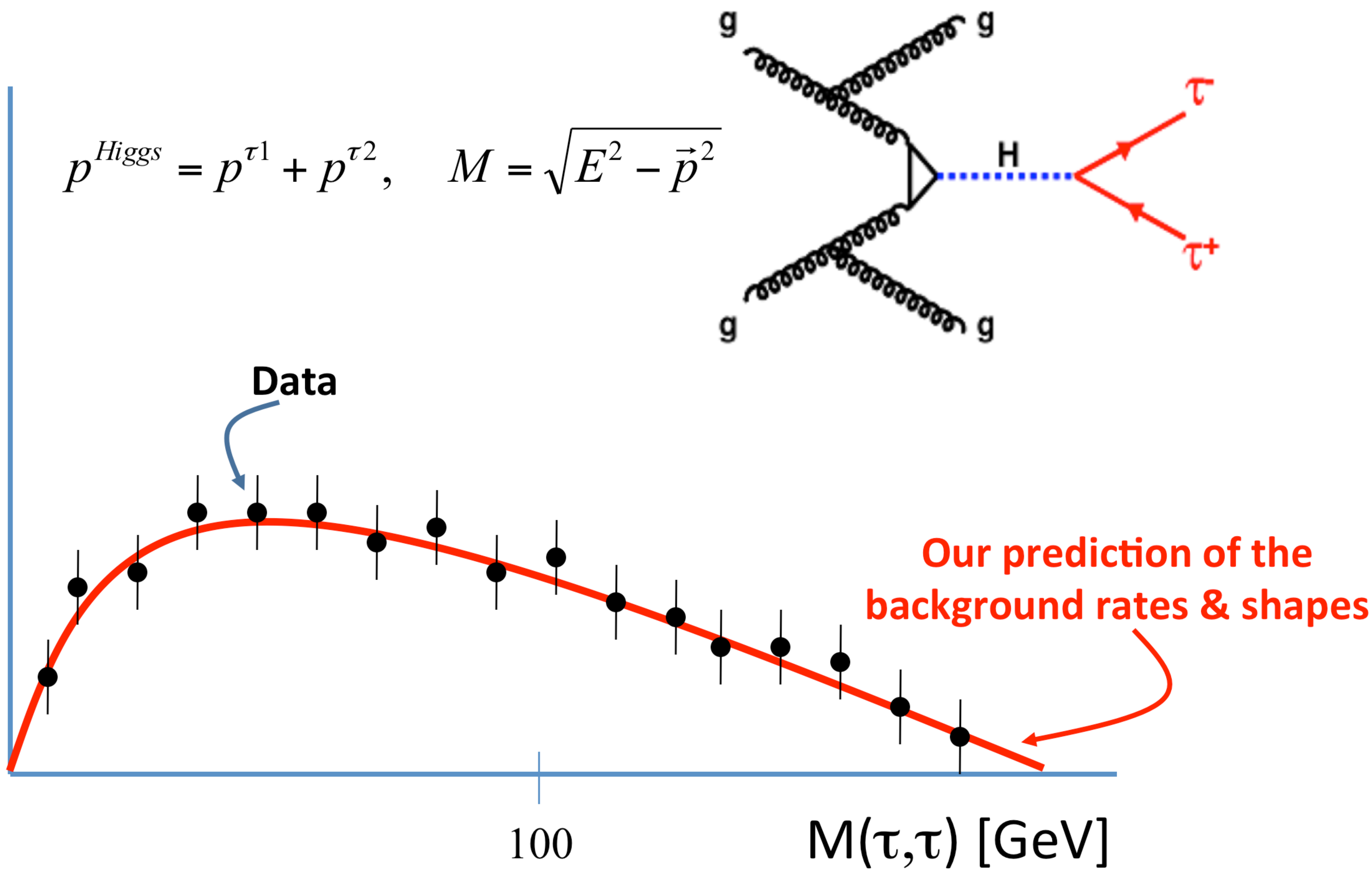
How Is New Physics Found?

- Simulate the signal and all relevant backgrounds
- Record data from a detector
- Foreach event in (signal, background, data):
  - If event passes analysis-specific selections:
    - Save event somewhere to the side
- If sum(signal+backgrounds) matches data:
  - Something is probably wrong, do a ton of cross-checks
  - If it still matches:
    - PressConference()

Hunting for "Bumps"

Suppose there was NO Higgs …
What would our distributions look like?

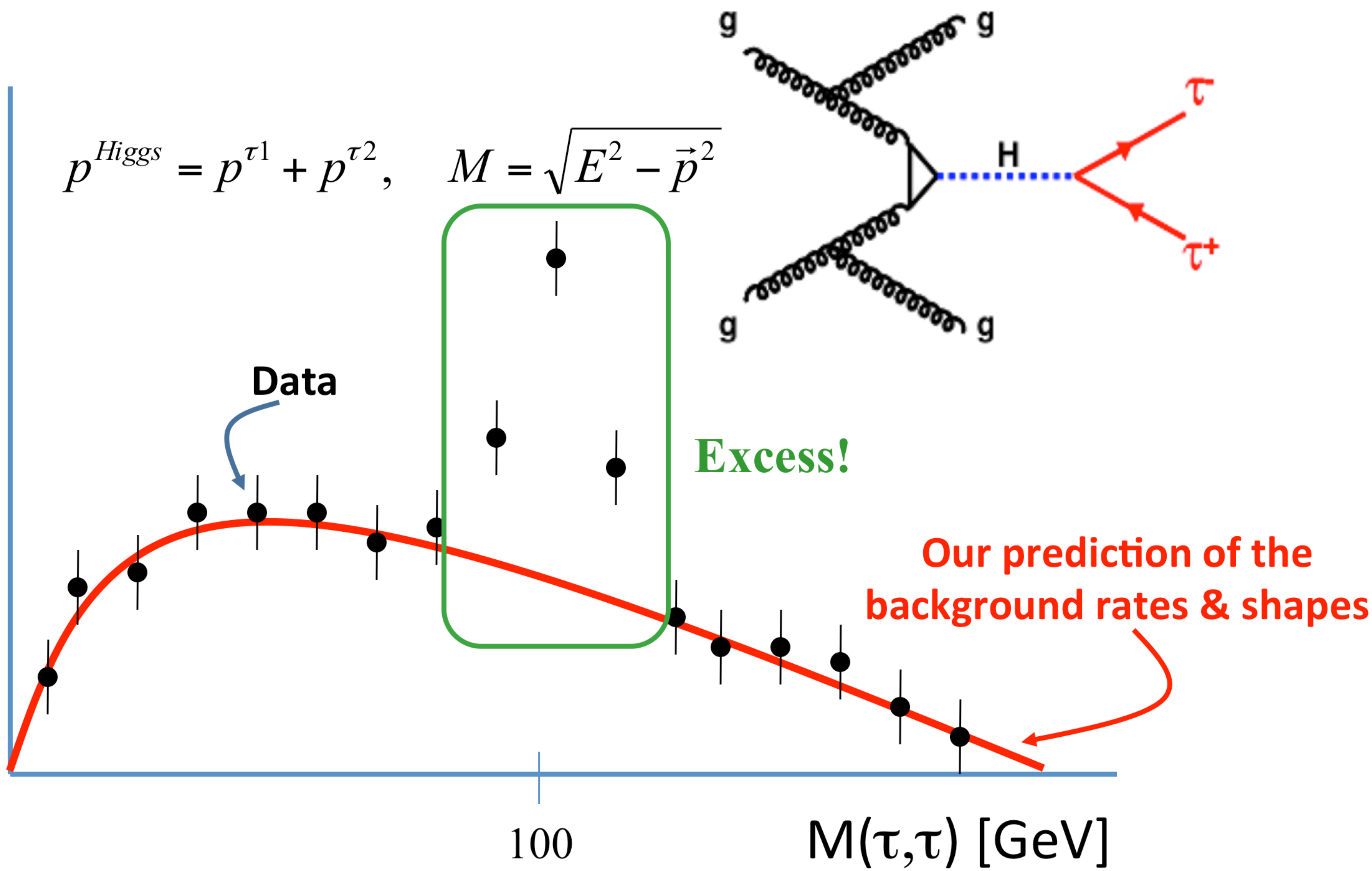
$$p^{Higgs} = p^{\tau 1} + p^{\tau 2}, \quad M = \sqrt{E^2 - \vec{p}^{\,2}}$$



Thanks: Alfredo Gurrola

Hunting for “Bumps”

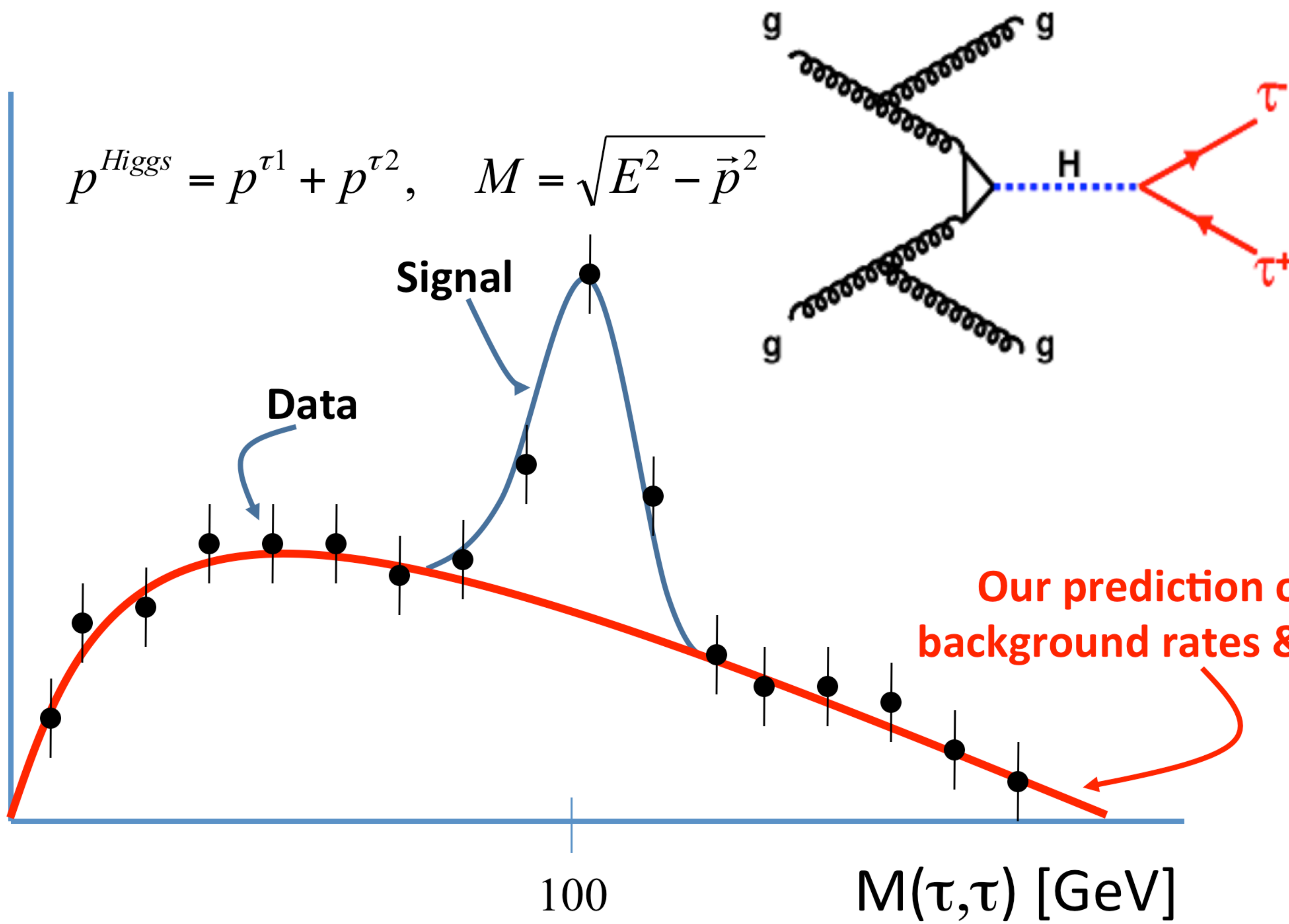
Suppose there WAS a Higgs …
What would our distributions look like?



Thanks: Alfredo Gurrola

Hunting for "Bumps"

Suppose there WAS a Higgs …
What would our distributions look like?



Thanks: Alfredo Gurrola

The Problem

How Many H Are Around Us?

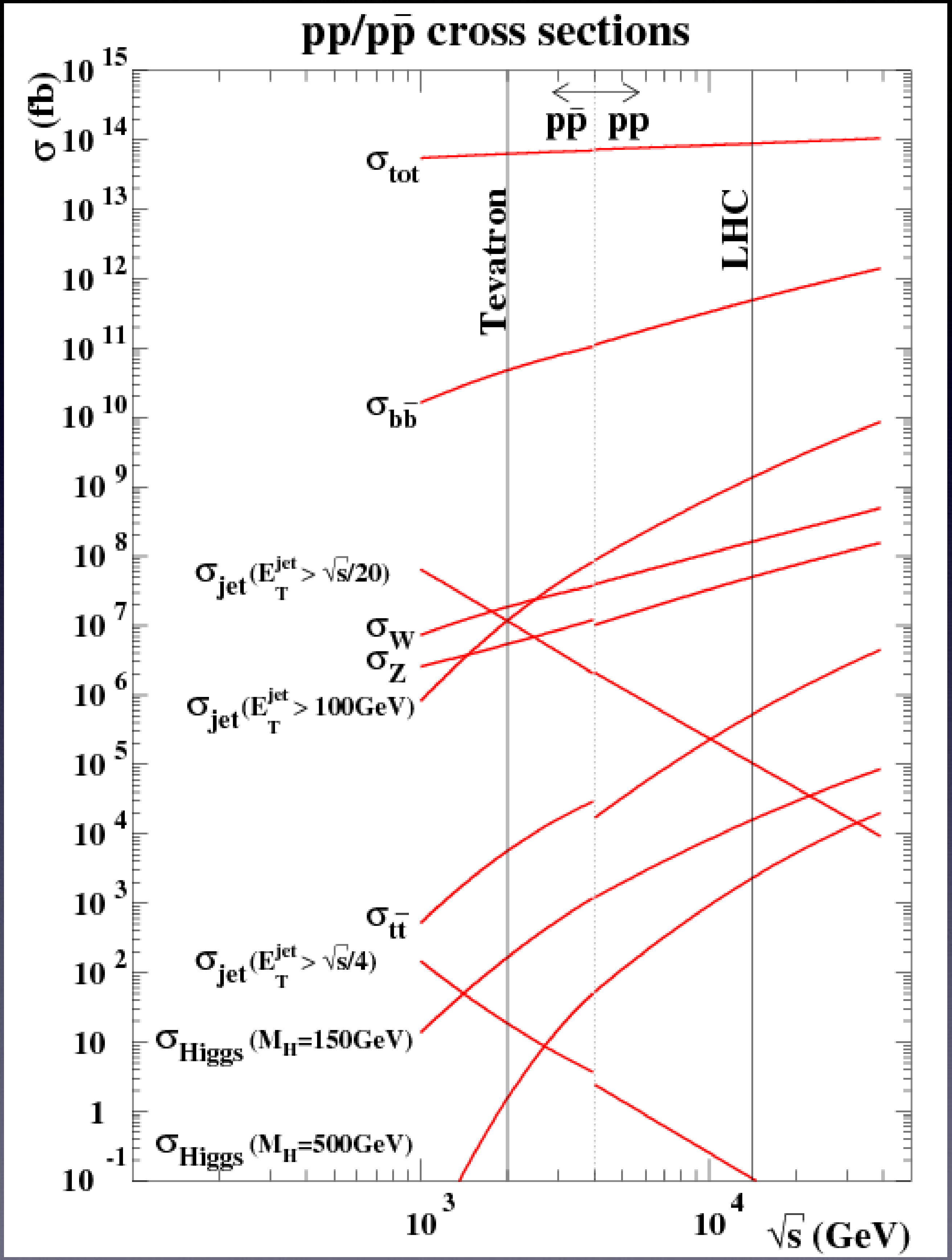
0

The Higgs weighs 125 GeV/c^2 - Need 125GeV of energy to produce one

Equivalent to 10.7 million degrees K

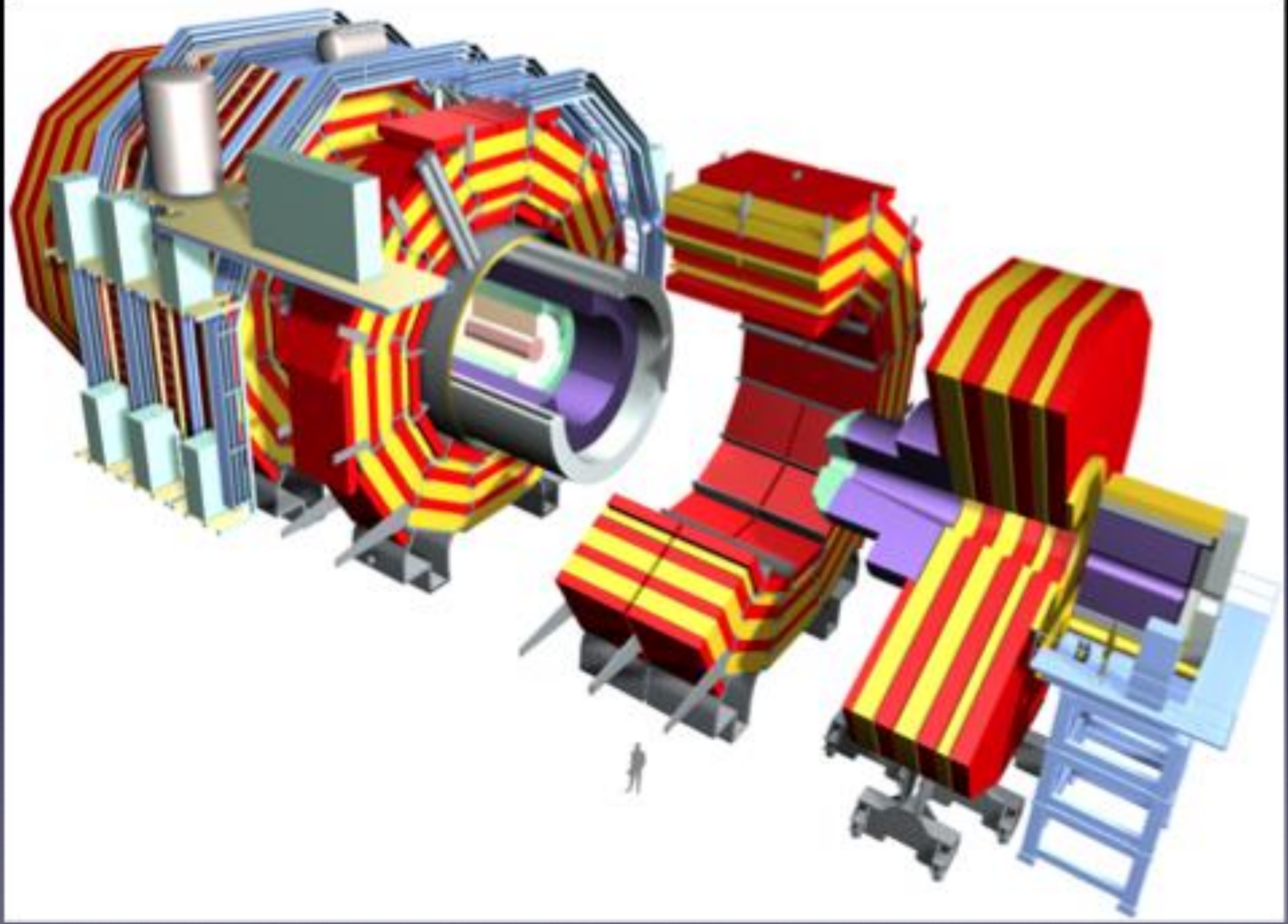
“We’re gonna need a bigger machine”

- Higher energies "unlock" new processes
- At increasing energies, heavier particles are increasingly preferred



The LHC Provides Higher Energies

- One proton-proton collision = 10^{-2}(Barn^{-1})
- Higgs production cross section @ 8TeV = 10^{-12}Barn
- Chance of producing a Higgs in a given collision: **10^{-14}**
- We need a LOT of collision events to produce even one higgs!



How Much Data Is Produced?

CMS Trigger System

Bunch Crossing Rate - 40MHz

Level1 Trigger - 100KHz

High Level Trigger (HLT) - 1KHz

To: Data AQuisition (DAQ) and Stable Storage

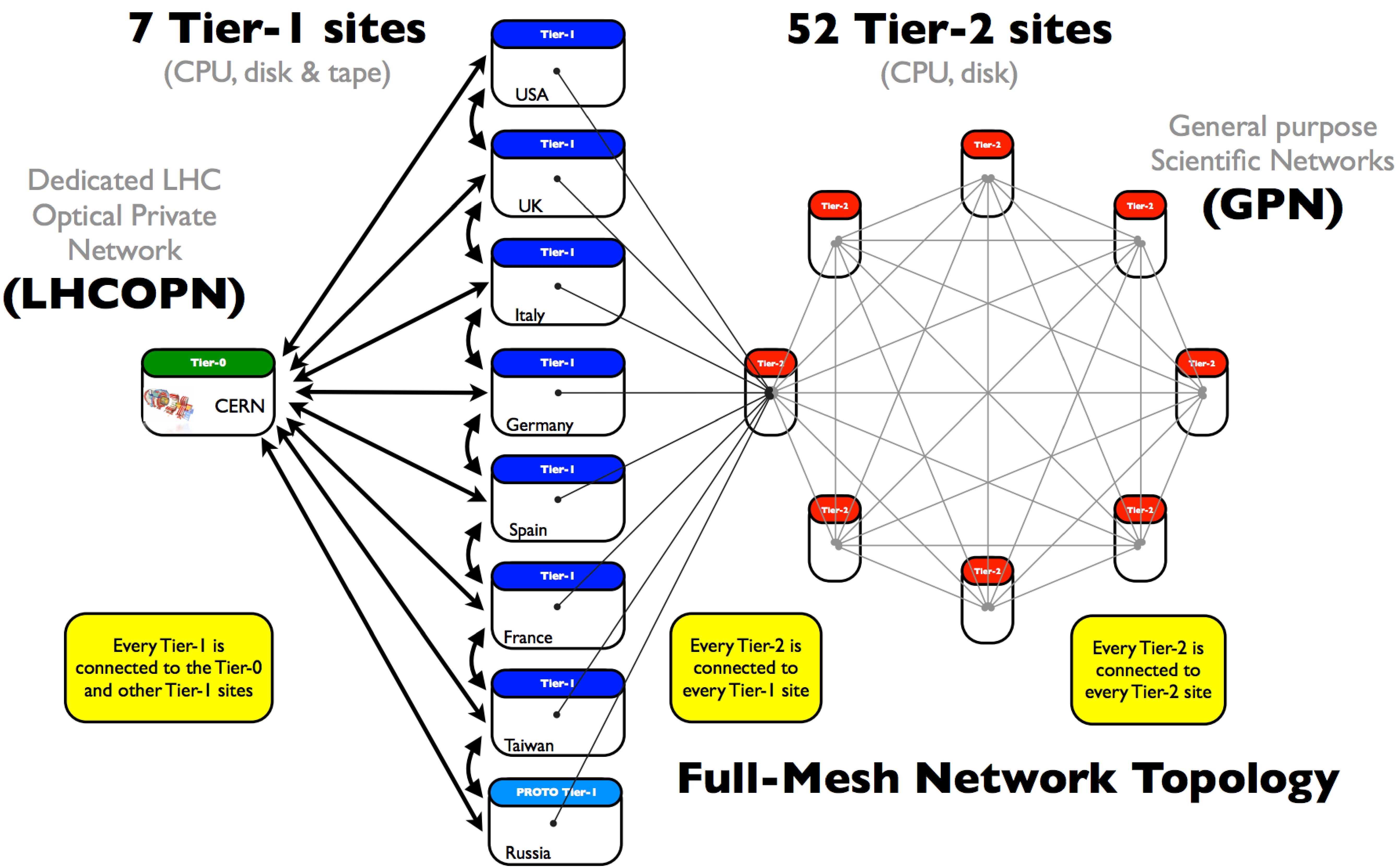
CMS Offline System

- Once the raw data is streamed to a buffer, it needs to be reshuffled to a permanent storage format and injected to be made available for subsequent steps to analyze
- The "offline" system happens asynchronously
  - The "online" system must be active while the detector is running
- I work here!

The Grid

- Was doing "The Cloud" before clouds were hip
- Allows the experiment to transfer data between and run jobs at ~80 sites in ~30 countries
- Federation of authentication and authorization (authn/authz) across administrative domains
- A consequence of the reality of funding

CMS Computing Infrastructure



Main data flows

Archiving

Tape based: grouping by physics content of related files on separate sets of tapes (tape families) have to be defined before files start arriving at a Tier-1 site

T0 → T1

- **RAW data**
  - RAW data is recorded at Tier-0
    - Stored on tape at CERN as "cold" backup copy
  - RAW data is distributed across the Tier-1s via the LHCOPN network links
    - Archived on tape and current year kept on disk for quick access

T2 → T1

- **Simulation output**
  - Output of Monte Carlo generation and simulation is produced mainly on Tier-2s (CPU intensive workflows)
  - Archived on tape distributed across the Tier-1s via the LHCOPN and GPN network links

T1 → T1

- **Analysis Object Data (AOD) production**
  - Output of reconstructions of RAW data and simulation is produced on Tier-1s (strong I/O capabilities required)
  - Archived on tape at Tier-1 site that has archival copy of input and ran reconstruction workflow

Serving

T1 → T2

- **Analysis Object Data (AOD) access**
  - Access to reconstructed data and simulated events on Tier-2 sites
    - Samples are distributed to the Tier-2s via the GPN network links
    - Tier-2 disk is logically separated into managed portions for common samples and unmanaged areas for user files (ntuples)physics samples

Main workflows

Production workflows

T0

- **Data recording**
  - Collisions are recorded by the detector, selected by the trigger and sent from the detector pit to the CERN computing center in binary format
  - 10% of the selected collisions are express repacked into the CMS ROOT format and reconstructed
    - Latency of express reconstruction is 1 hour to allow for prompt alignment and calibration workflows and data quality monitoring
  - 100% of the selected collisions are repacked into the CMS ROOT format and promptly reconstructed
    - Prompt reconstruction starts 48 hours after events have been recorded to incorporate updated calibrations

T1/2

- **Simulation**
  - Collisions are generated using theory software packages and the detector response is simulated using GEANT4
    - CPU intensive workflow
    - Needs little or no input

T1

- **Data and simulated event reconstruction**
  - Collisions are reconstructed with different software versions and/or calibrations
    - RAW data and simulated events archived on Tier-1s need to be pre-staged to disk for efficient access
    - Reconstruction of simulated events needs to access additional datasets as input to simulate additional interactions in the detector (PileUp)

Analysis

T2

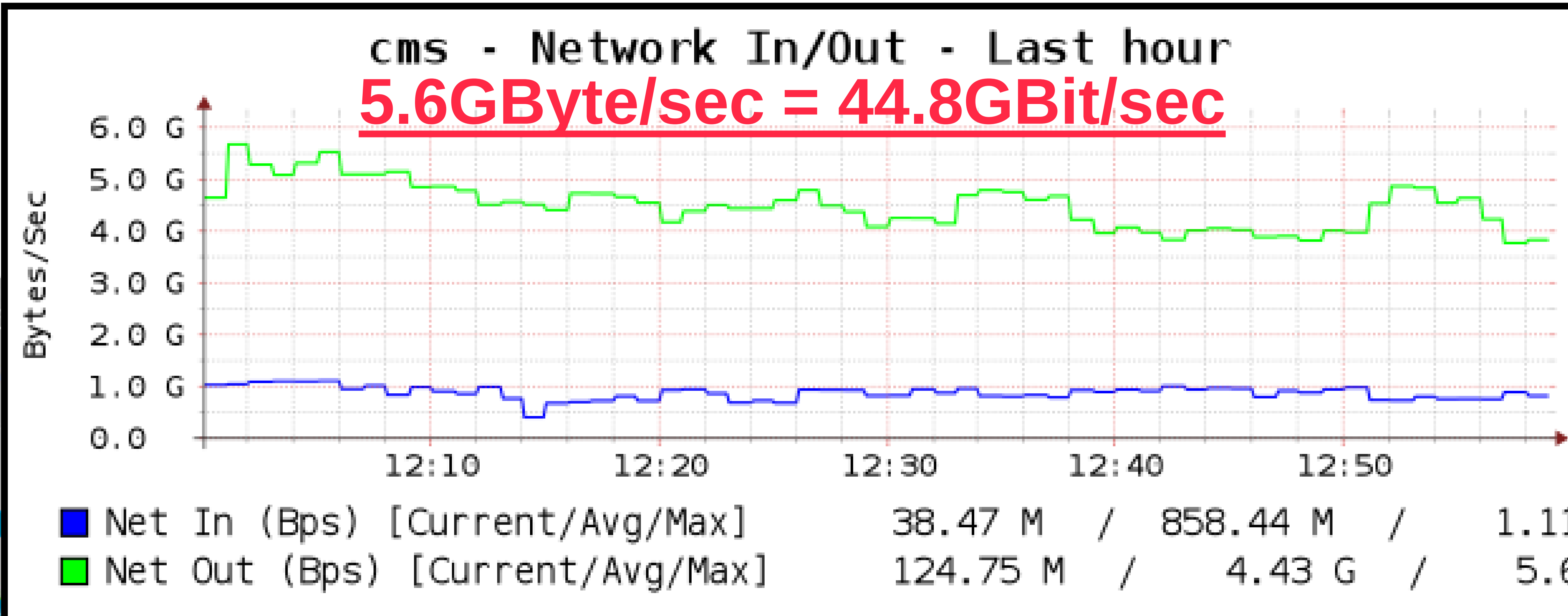
- **Data and simulated event analysis**
  - Collisions are accessed by physicists using officially released and self-written code
    - Input is distributed beforehand through transfer system and analysis jobs are sent to data location
    - Output is stored on Tier-2s outside the transfer system and catalogues but can be elevated
  - Multi-user access instead of single user production mode

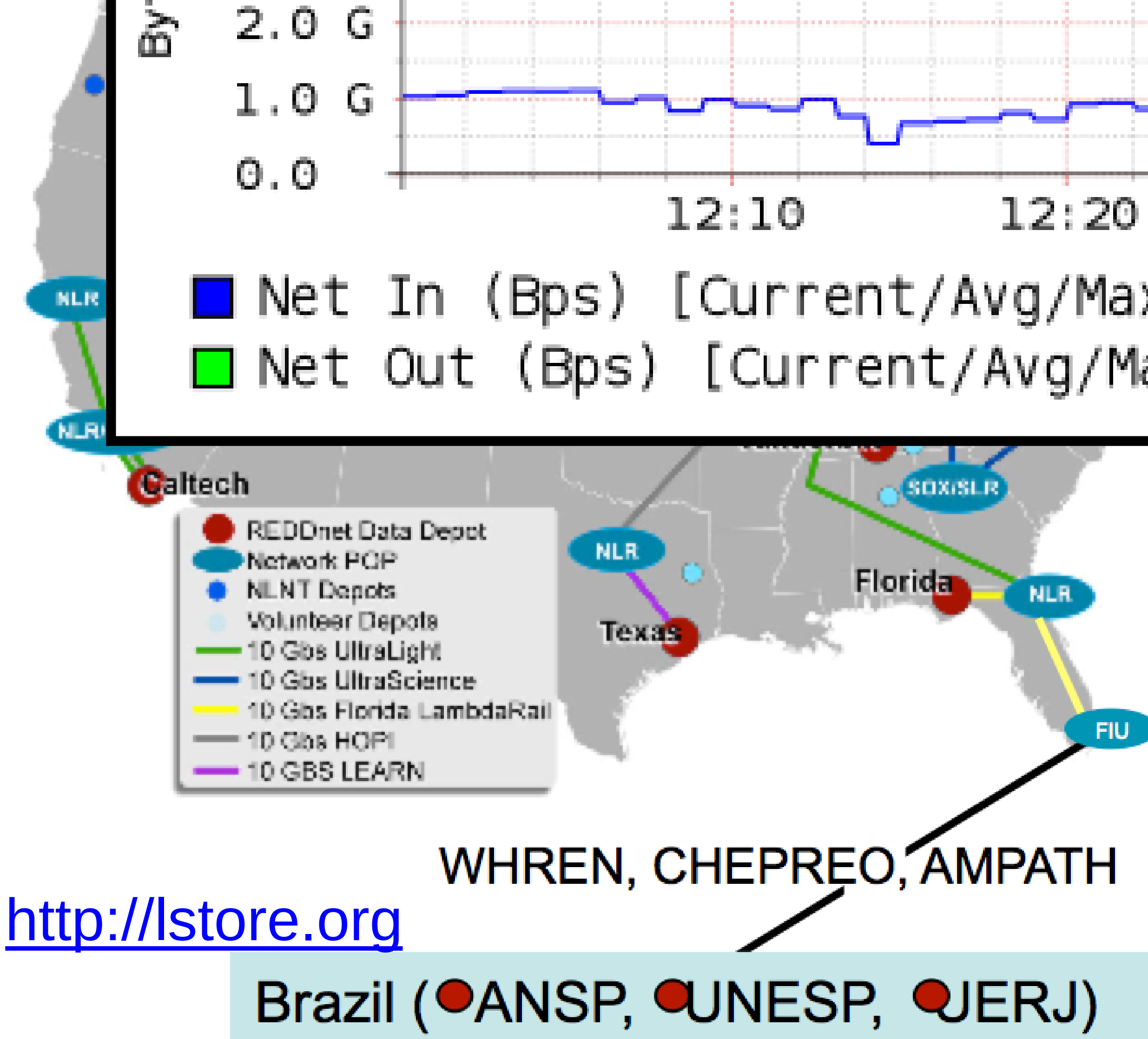
Storing PBytes of Data

- Rule #1 - Has to be cheap
- Rule #2 - Has to be fast
- Rule #3 - Has to be reliable

http://lstore.org

LStore: Vanderbilt's Solution



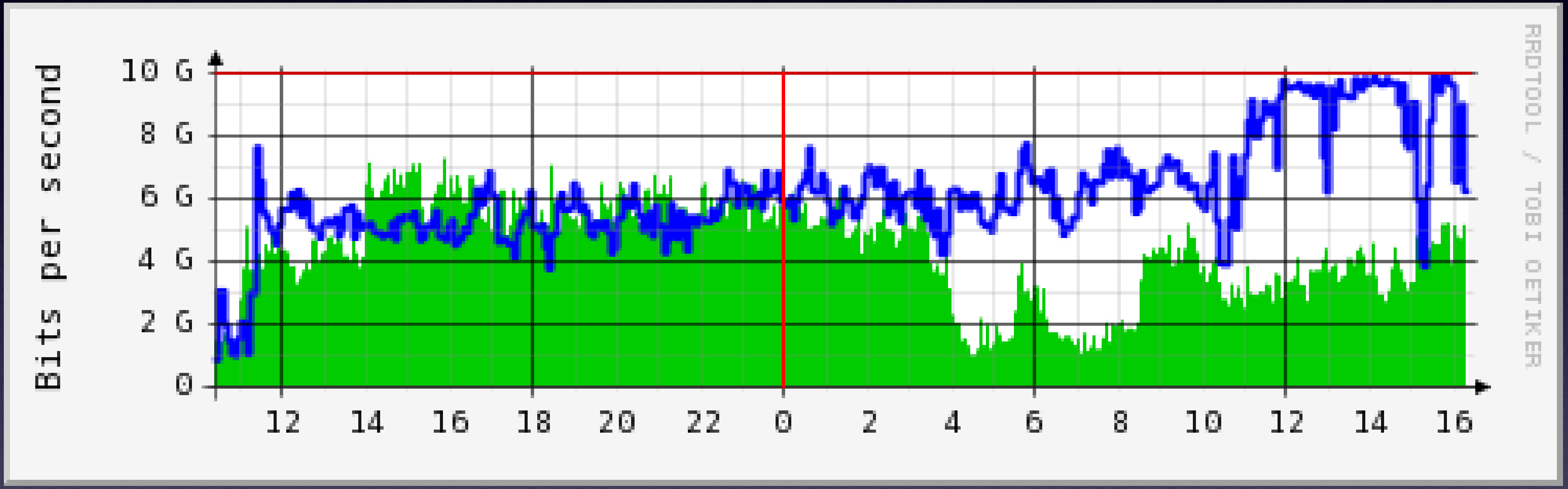


– Satellite Imagery (AmericaView)
– High Energy Physics: LHC
– Terascale Supernova Initiative
– Structural Biology
– *Vanderbilt Campus Depot*

http://lstore.org

Transferring over the WAN

- Managed to fully saturate a 10GBit/s link!



- Vanderbilt ITS was VERY unhappy

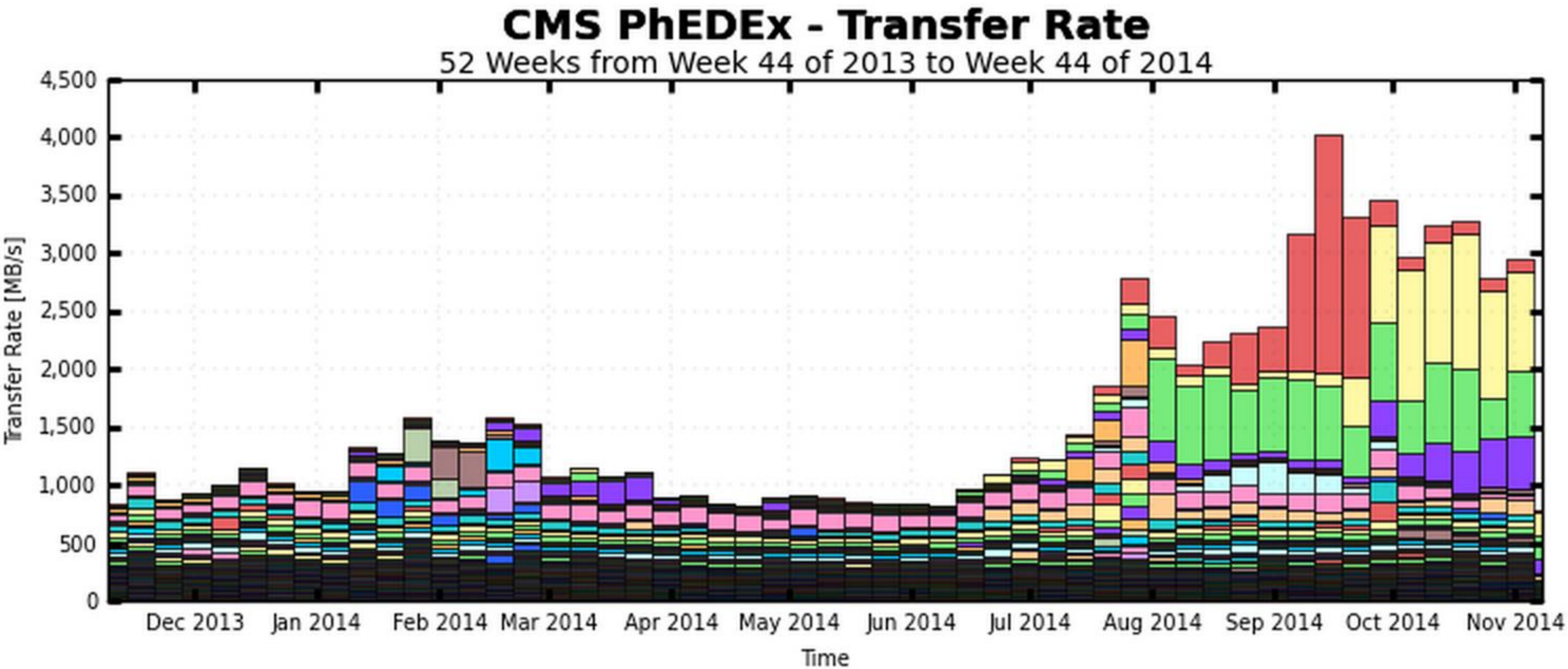
> ==========
> A traffic policer will be configured on the research 6509 to protect enterprise Internet access. It will effectively limit traffic from the research networks to <9Gbps. The policer will not be applied to the interfaces until an outage is actually occurring, so there is not an expected service impact from this change.
>
> The impetus for this change was the recent Google outage caused by link saturation.

http://github.com/PerilousApricot/gridftp-lfs

Global Bulk Transfers

- PhEDEx is responsible for delivering data globally
    - Clever acronym, say it aloud
- Global agents make queues for files pending transfer to sites
- Local agents at each site handle moving the files they need

~200PB/Year



CMS PhEDEx - Transfer Rate
52 Weeks from Week 44 of 2013 to Week 44 of 2014
Transfer Rate [MB/s]
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Dec 2013
Jan 2014
Feb 2014
Mar 2014
Apr 2014
May 2014
Jun 2014
Jul 2014
Aug 2014
Sep 2014
Oct 2014
Nov 2014
Time

Workflow Management

- A typical analysis may use O(PB) of data
- What tools are needed to leverage our computing resources to enable these workflows?
- Remember: resources are at sites with varying levels of support and performance
  - Assume the worst and expect to retry
- Choose to optimize throughput over latency
  - We have more jobs than CPUs
  - Everything will have to wait anyway

Optimize for Users Needs

- Production system
  - Requests are well defined and long-term: “re-reconstruct all of the data from 2012”
  - Should be extensively automated: nTasks >>> nHumans
- Analysis system
  - Requests are short-term and ill-defined: users are REAL good at breaking things
  - Can fall back to the user more often

WMAgent and CRAB

- Each is built off the same framework (I spent 3 years here)
- WMAgent - for production
  - Complete lifecycle for data from detector to scientist-usable form
  - A central request manager generates WorkQueueElements for distributed and isolated WMAgents to consume
    - Long queues = resiliency to failure!
- CRAB - for analysis (I worked on the current rewrite)
  - Lets a user say “Let me find all of the events with two electrons”

http://github.com/dmwm/WMCore

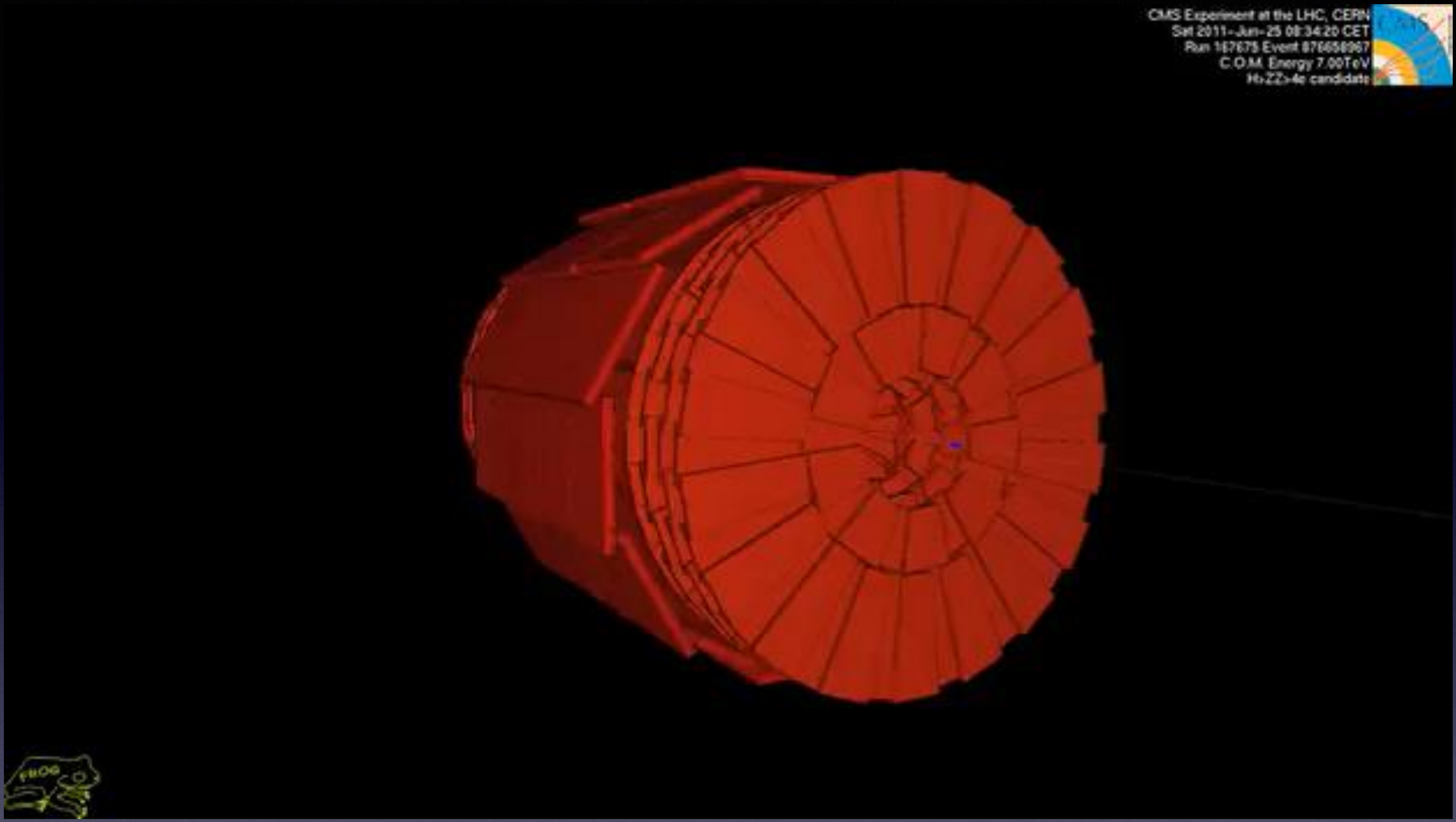
CMS SoftWare (CMSSW)

- Software framework and executables which handle reading/writing/analyzing all CMS data
- ~1.5M lines of C++
  - Previous team member on C++ standardization committee
  - Very active in GCC development
- C++ modules linked by simple python configuration language
- Makes the easy stuff easy and the hard stuff possible

http://github.com/cms-sw/cmssw

Future Plans

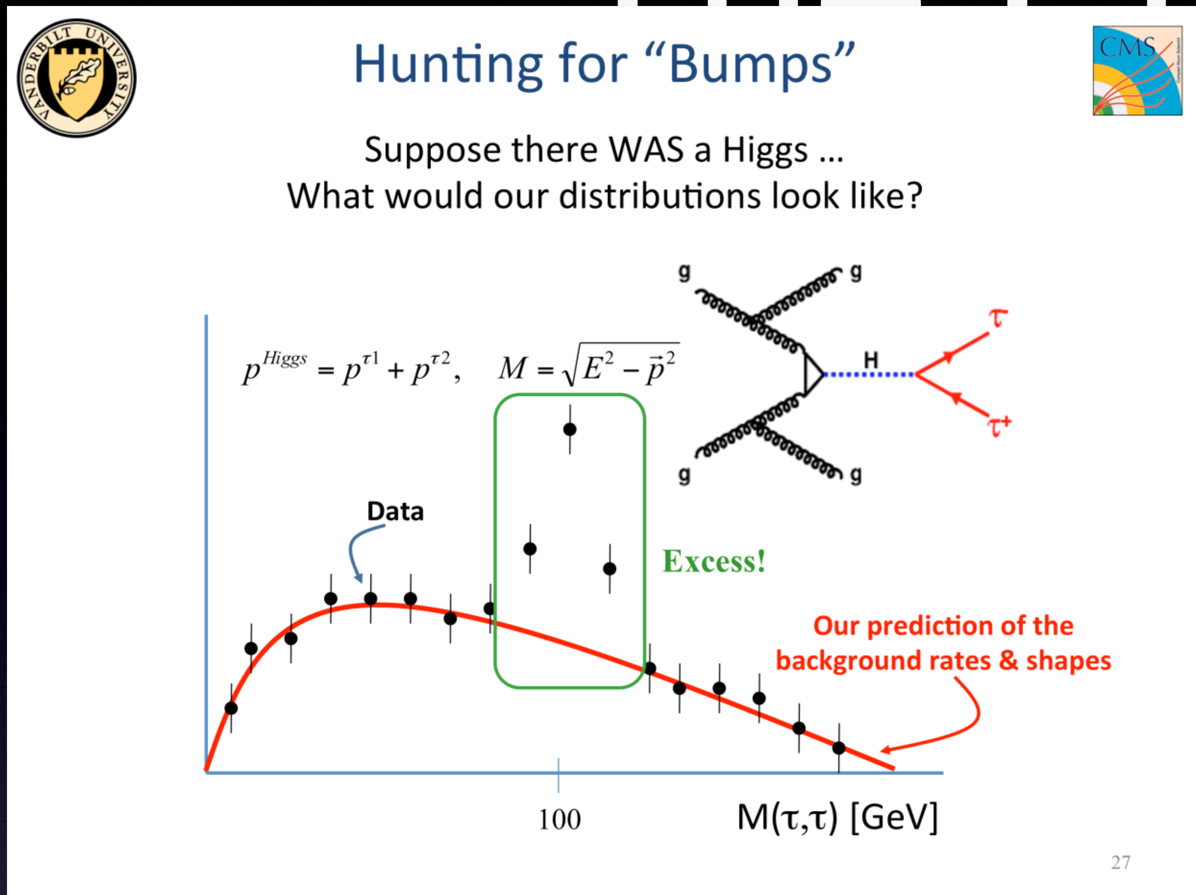
- LHCs upgrade completed last year, which means not only a higher luminosity, but much more complex events
- GPU-ization of algorithms
  - Tracking/Simulation
- Better networking
  - Most sites moving to 100GBt WAN links
- Simplified on-disk formats
  - Faster to read
  - Smaller on disk
- Multicore processing

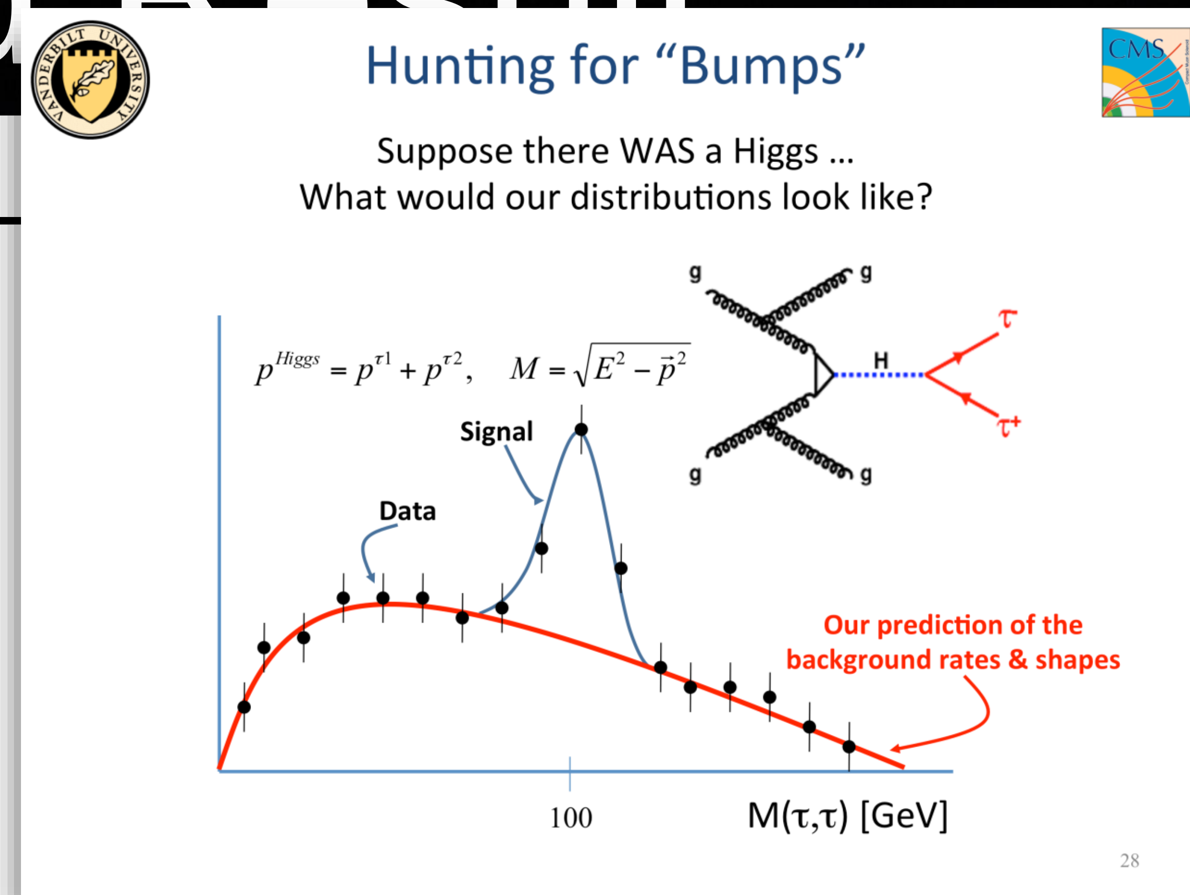


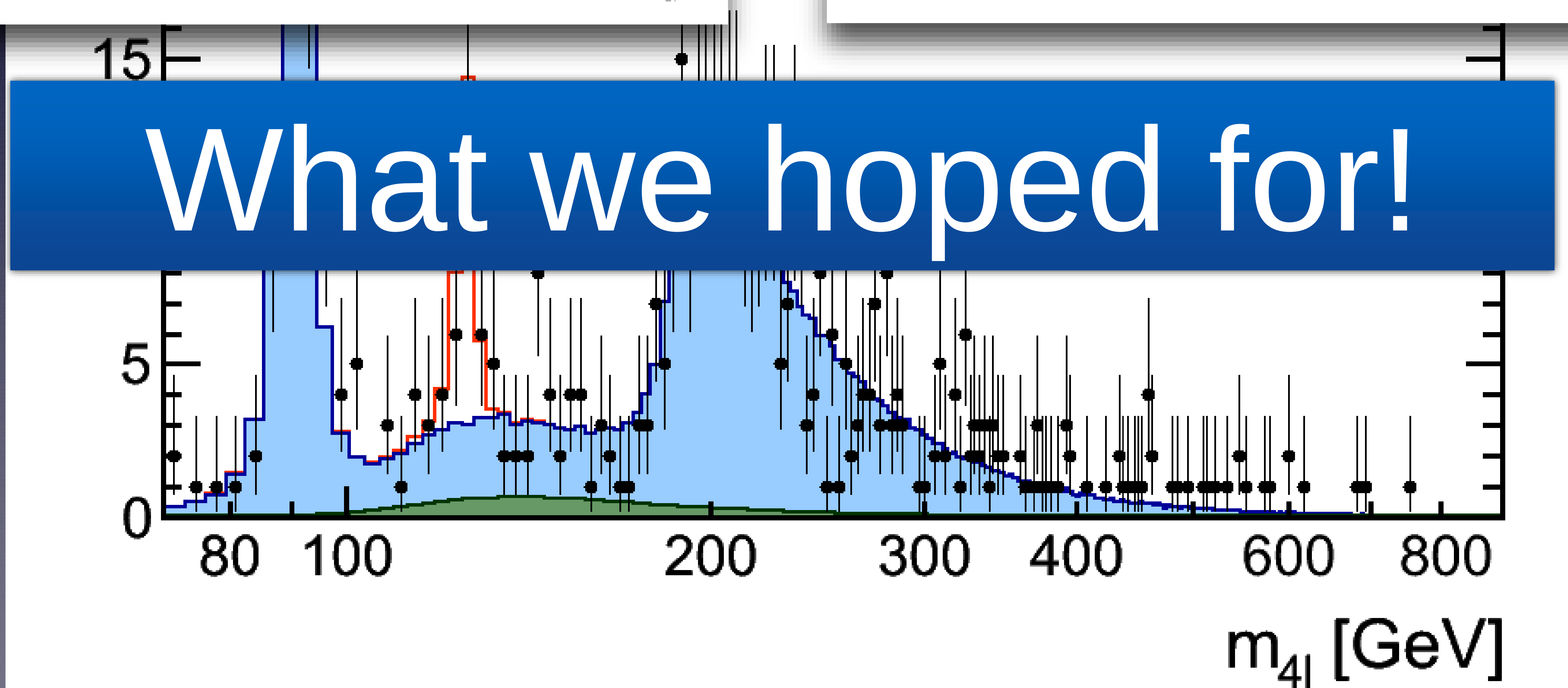
The End Result

One candidate Higgs -> 4mu event

The End Result

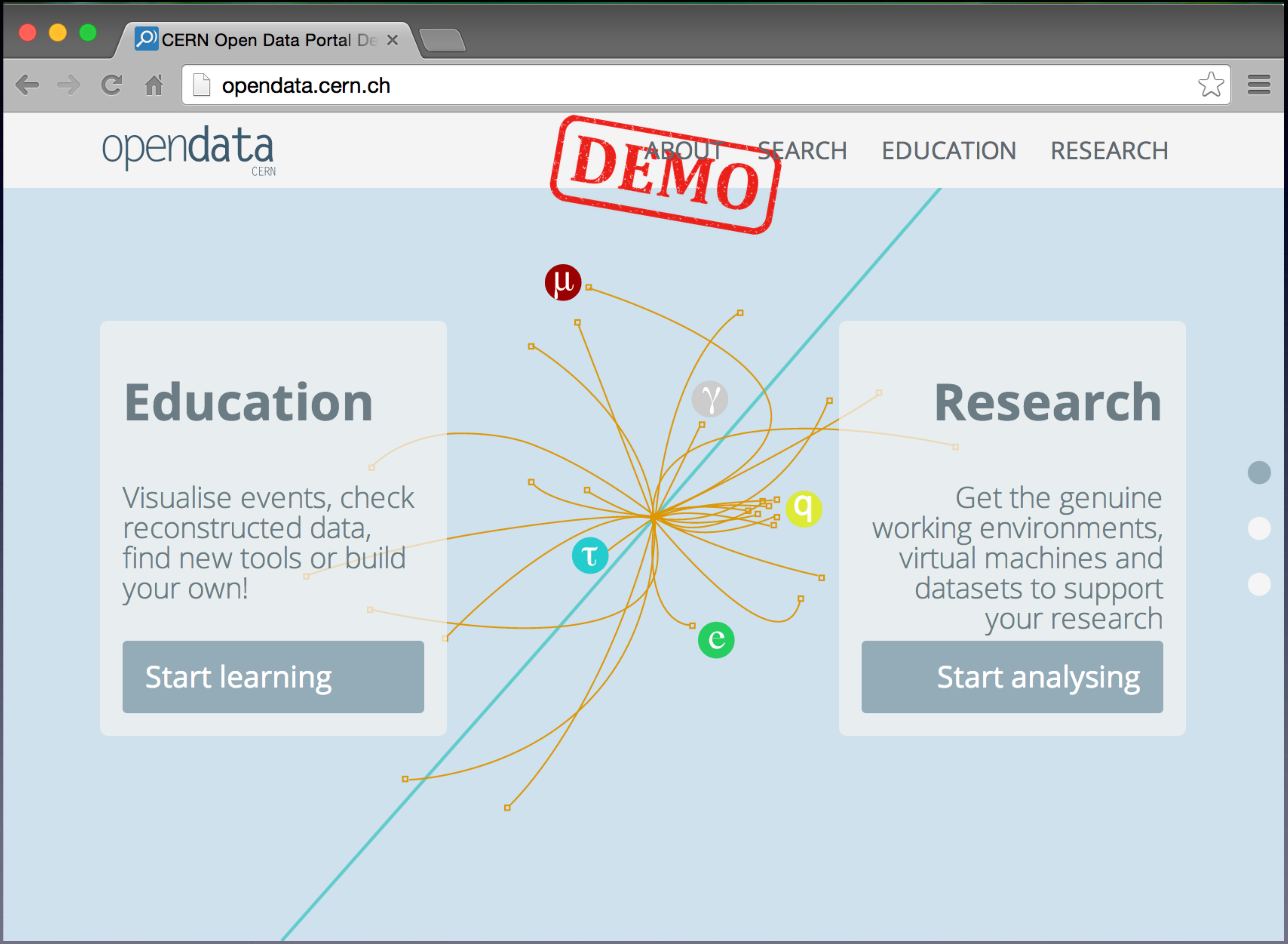






What we hoped for!

See more!


CERN Open Data Portal De
opendata.cern.ch
opendata
CERN
DEMO
ABOUT
SEARCH
EDUCATION
RESEARCH
Education
Visualise events, check reconstructed data, find new tools or build your own!
Start learning
Research
Get the genuine working environments, virtual machines and datasets to support your research
Start analysing