Projects proposal at the CMS Experiment

CERN Openlab Machine Learning and Data Analytics Workshop

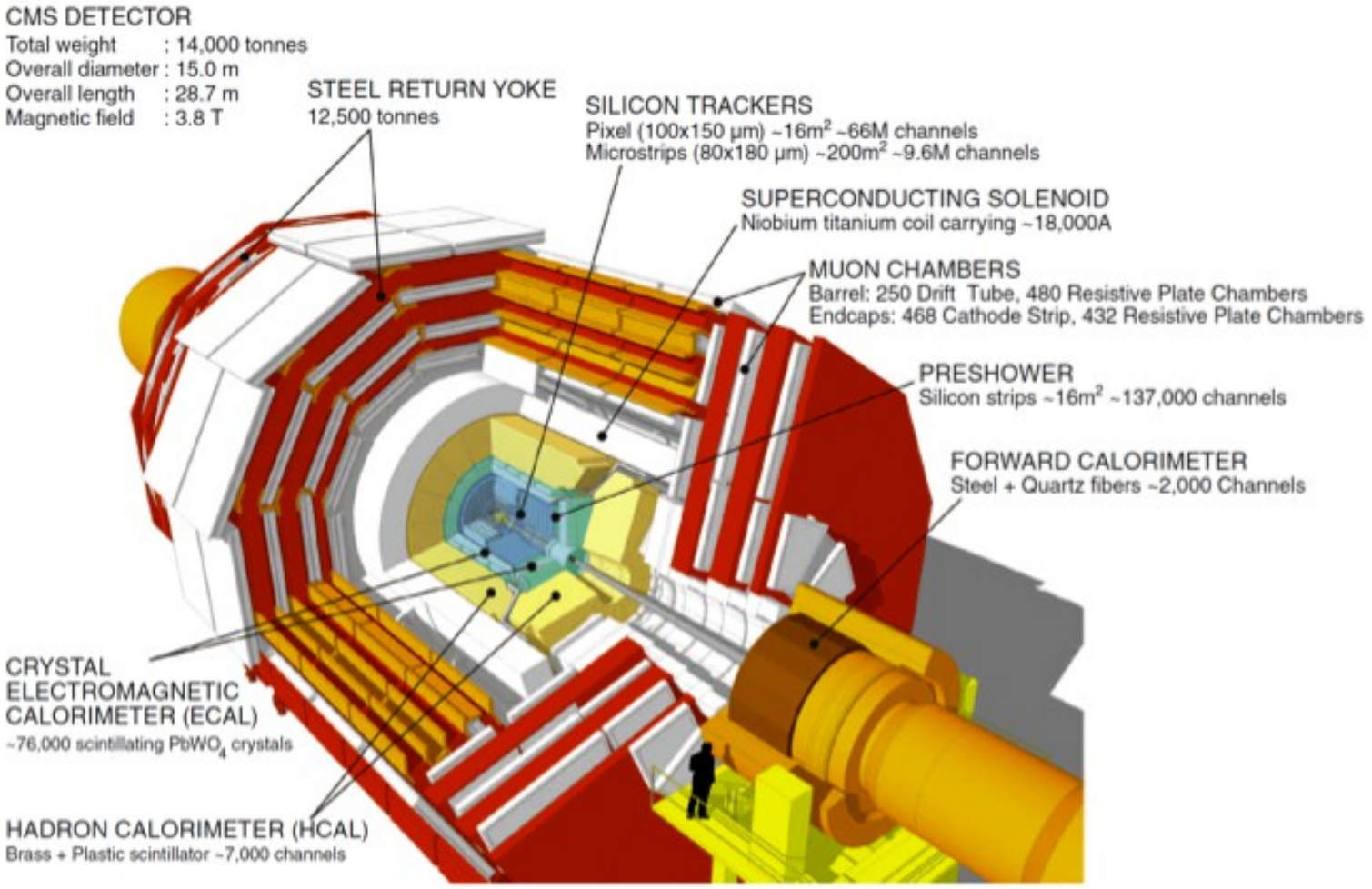
29 April 2016

An overview of possible topics for collaboration with Openlab's member companies, focused on big data, analytics and machine learning...

Goals

- Solve HEP problems with **potentially existing** solutions from the private sectors.
- **Shape the new technologies** to the unique needs of high energy physics.
- **Acquire knowledge** on tools used in industry. **Building collaboration** with industry. **Prepare** students for transition outside of academia

Compact Muon Solenoid (CMS)

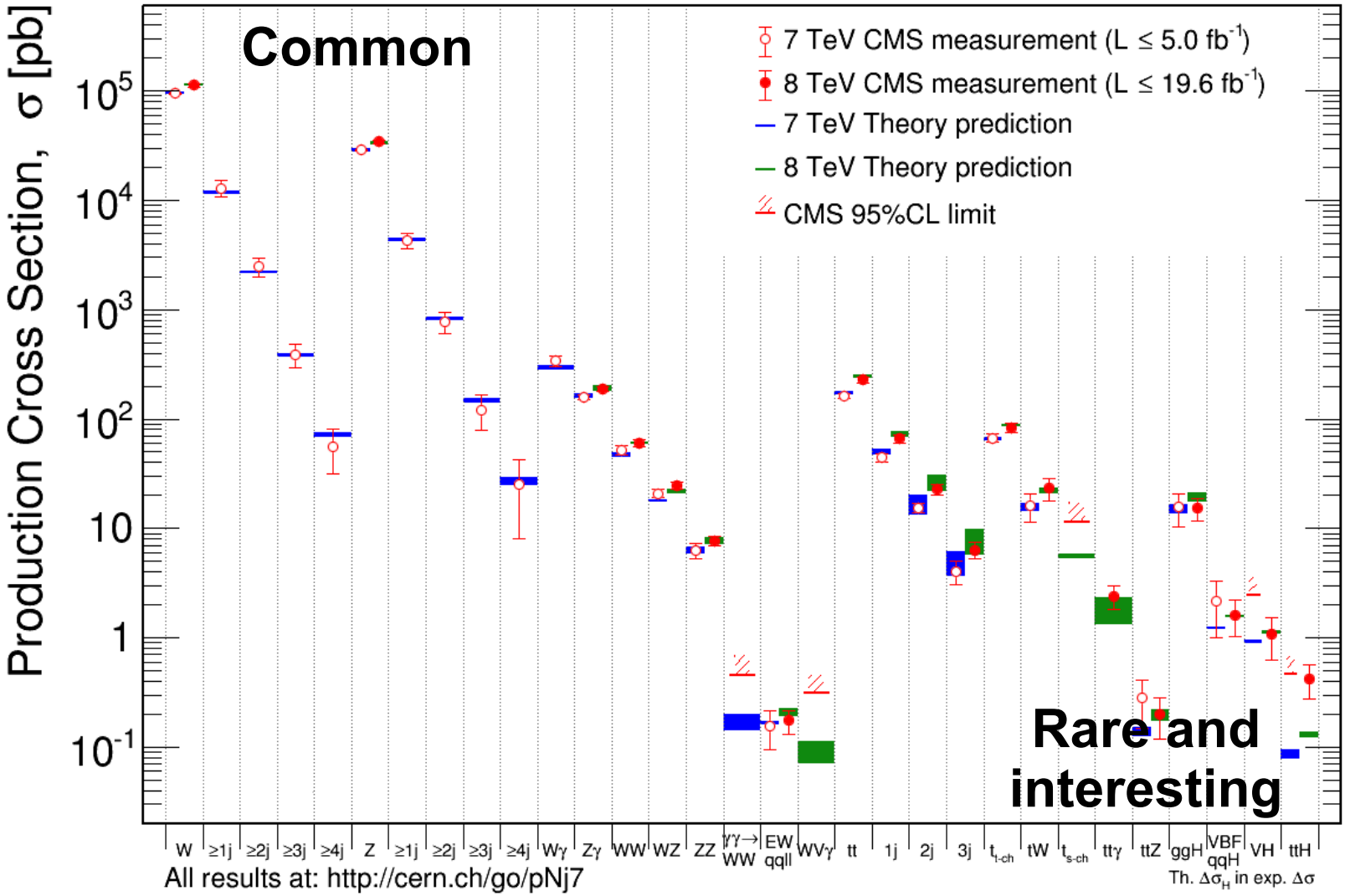


Highly heterogeneous system

Raw data is 100M channels sampled every 25 ns : 1Pb/s

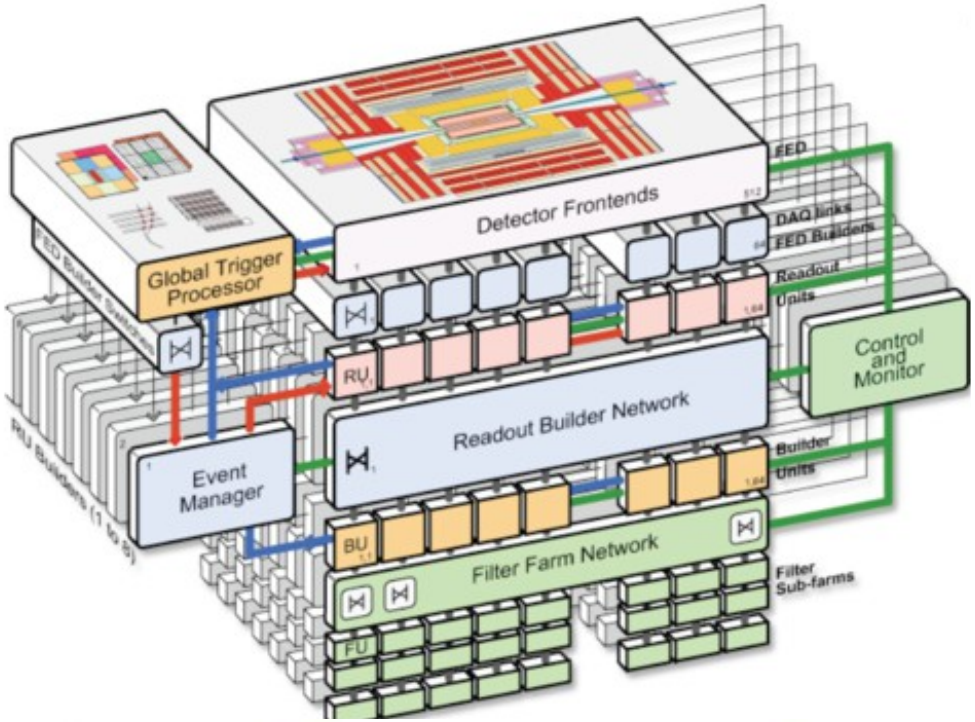
50EB per day in readout and online processing.

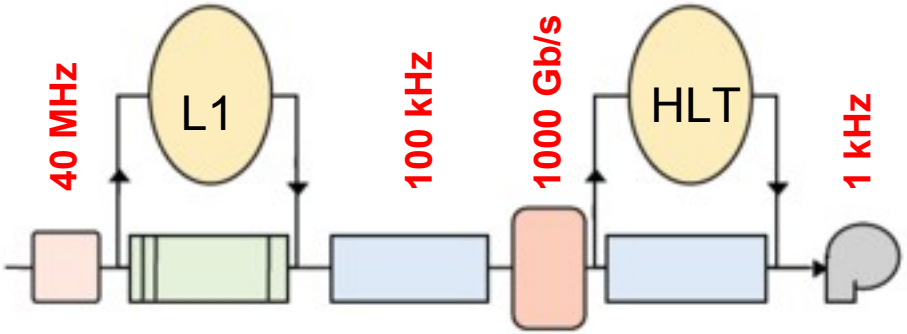
Scale of the Problem

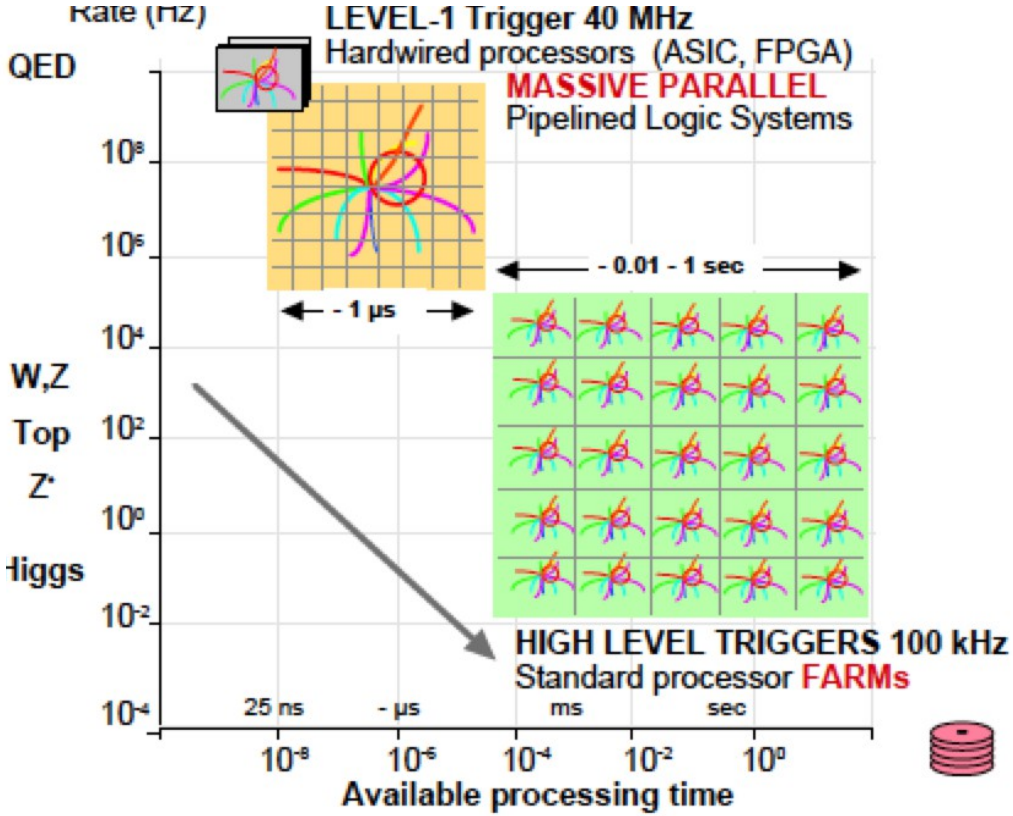


Many orders of magnitude rejection in order to select interesting events

Trigger Decision

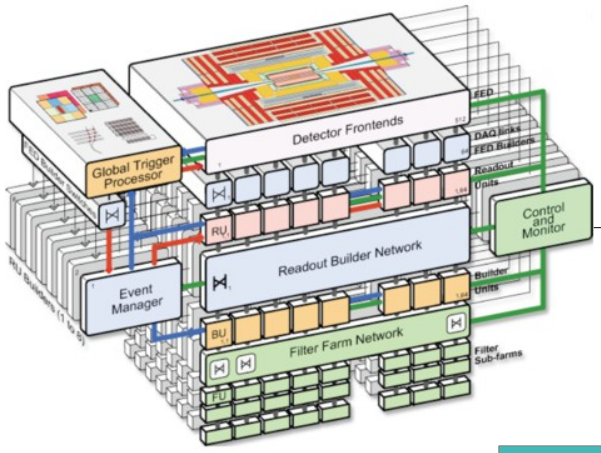






- Massively parallel electronic infrastructure makes a rough selection
- Refined decision in a software defined trigger

Data and Simulation Production



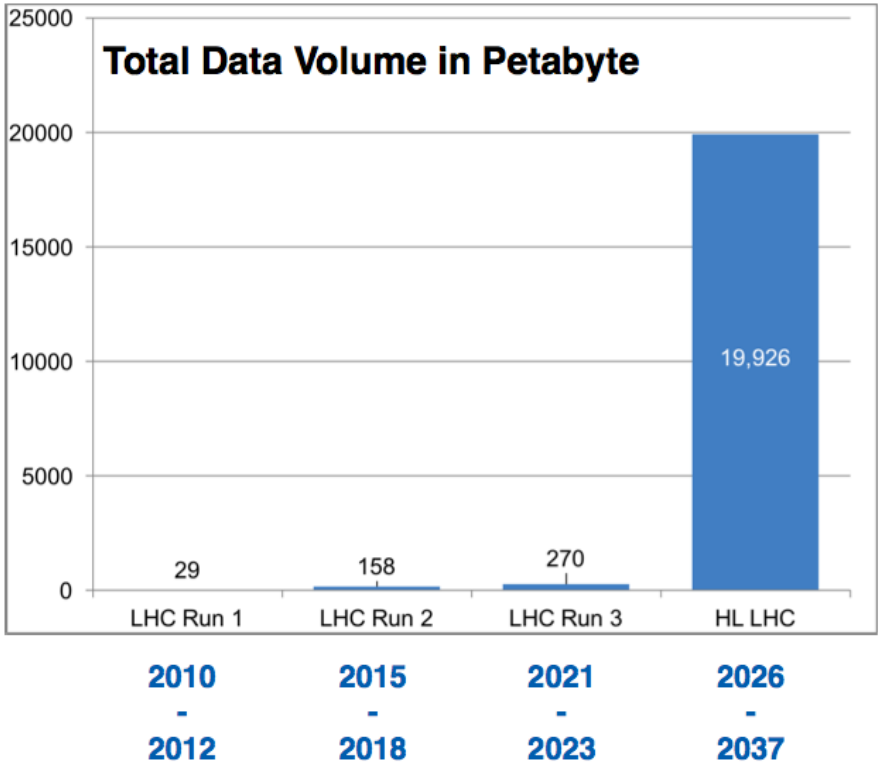
- Hundreds of computer centers (100-10k cores per site) in a full mesh structure
- Increased use as a cloud resources (any job anywhere)
- Increasing use of additional cloud and HPC resource
- Real time data processing at Tier0
- Data and Simulation production at Tier1 and Tier2
- High bandwidth networks

Upcoming Challenges for HEP data

- Event filtering in HL-LHC (circa 2025)
  - Hardware output rate 500-750 kHz (7x)
  - Output rate 5-7.5 kHz (7x)
  - Throughput 25-40 GB/s (20x)
  - Online computing power 5-11MHS06 (50x)
  - ✗ Large cost in construction and operation

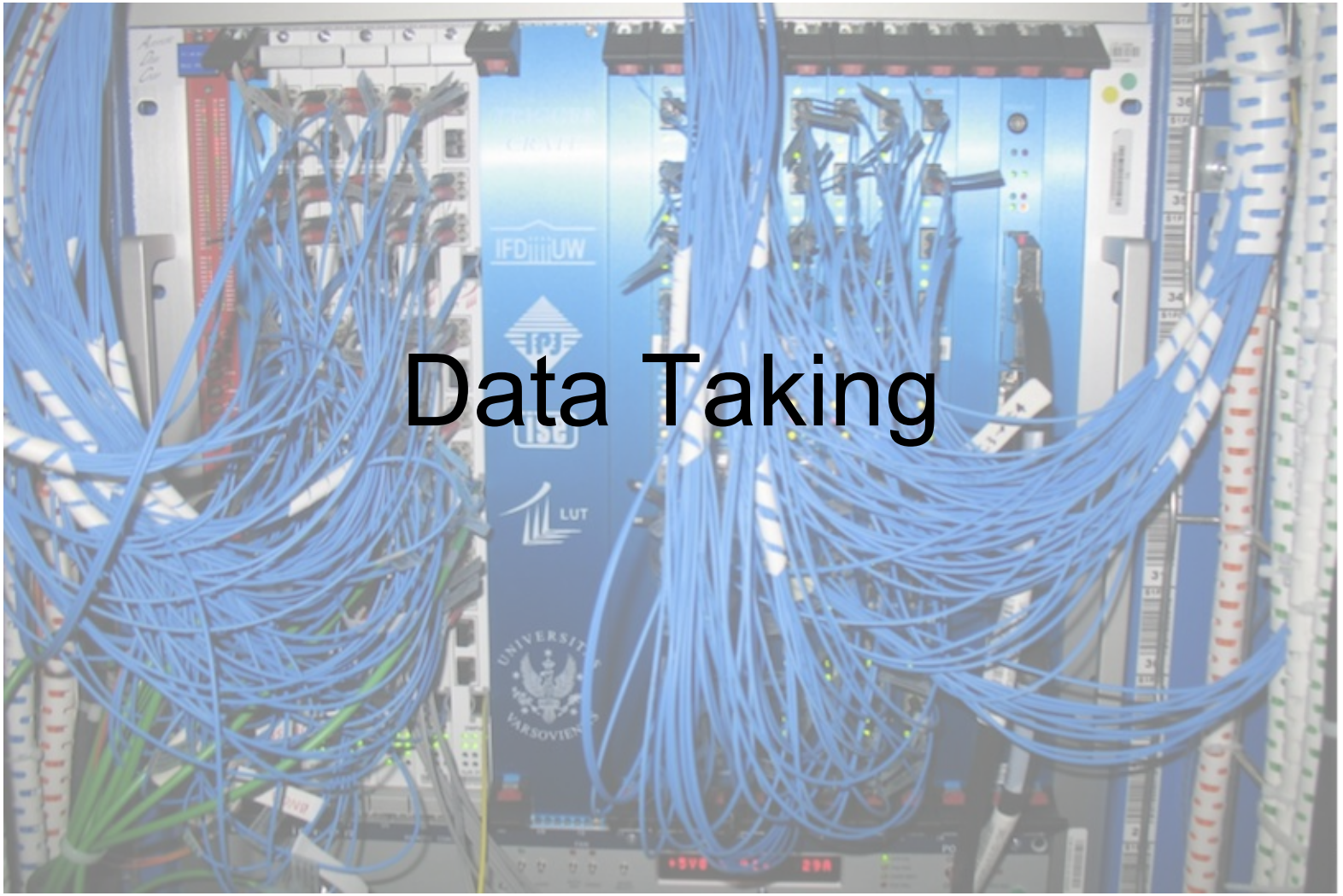
- 50x in data volume
- Raw data processing
  - 20-45x larger time per event
  - ✗ Resource needs growth beyond prediction of growth in budget

- ➔ Online and Offline/Grid processing
  - ➔ Large volume of data in readout and filtering
  - New algorithm not necessarily more accuracy required but definitely running faster
  - Any 1% gain is a lot of budget
  - Huge amount of data for analysis



Projects Outline

- Data taking
  - Real time event categorization
  - Data monitoring & certification robot
- Data Reconstruction
  - Calorimeter reconstruction
  - Boosted object jet tagging
- Data Processing
  - Computing Resource Optimization
  - Predicting data popularity
  - Intelligent networking
- Data Analysis
  - CMS assistance service
  - Big data reduction and analysis
  - Model independent search


Data Taking

Real Time Event Categorization

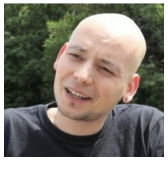
Contact : Anderson, Fisk, Pierini

- ✔ Hardware event filtering (~100kHz) are designed on crude calculation due to limited pipeline depth
- ✔ Software event filtering input is limited by bandwidth from the detector
- ✔ Software event filtering throughput (couple kHz) is limited by storage planning and realtime offline data processing
- ✘ Event **selection is approximate** due to computation budget
- ✘ Events rejected are **lost forever**

- Going **beyond the traditional approach** and study events in real time
- Cover physics **phase space otherwise uncovered**
- Extract lightweight analysis information from otherwise rejected events
- Indexing data with big data tools
  - Demonstration with elasticsearch
  - Change of the analysis with indexing
  - Explore and histogram data with flexible queries

- ➔ Looking forward to industry partners
  - Deploy **big data solution servers**
  - **Accelerate event indexing** at unprecedented rate
- ◆ One summer student this year. Other participation most welcome





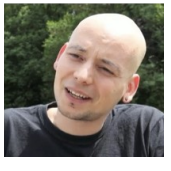
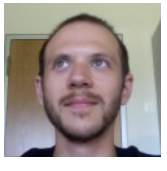
Data Monitoring/Certification Robot

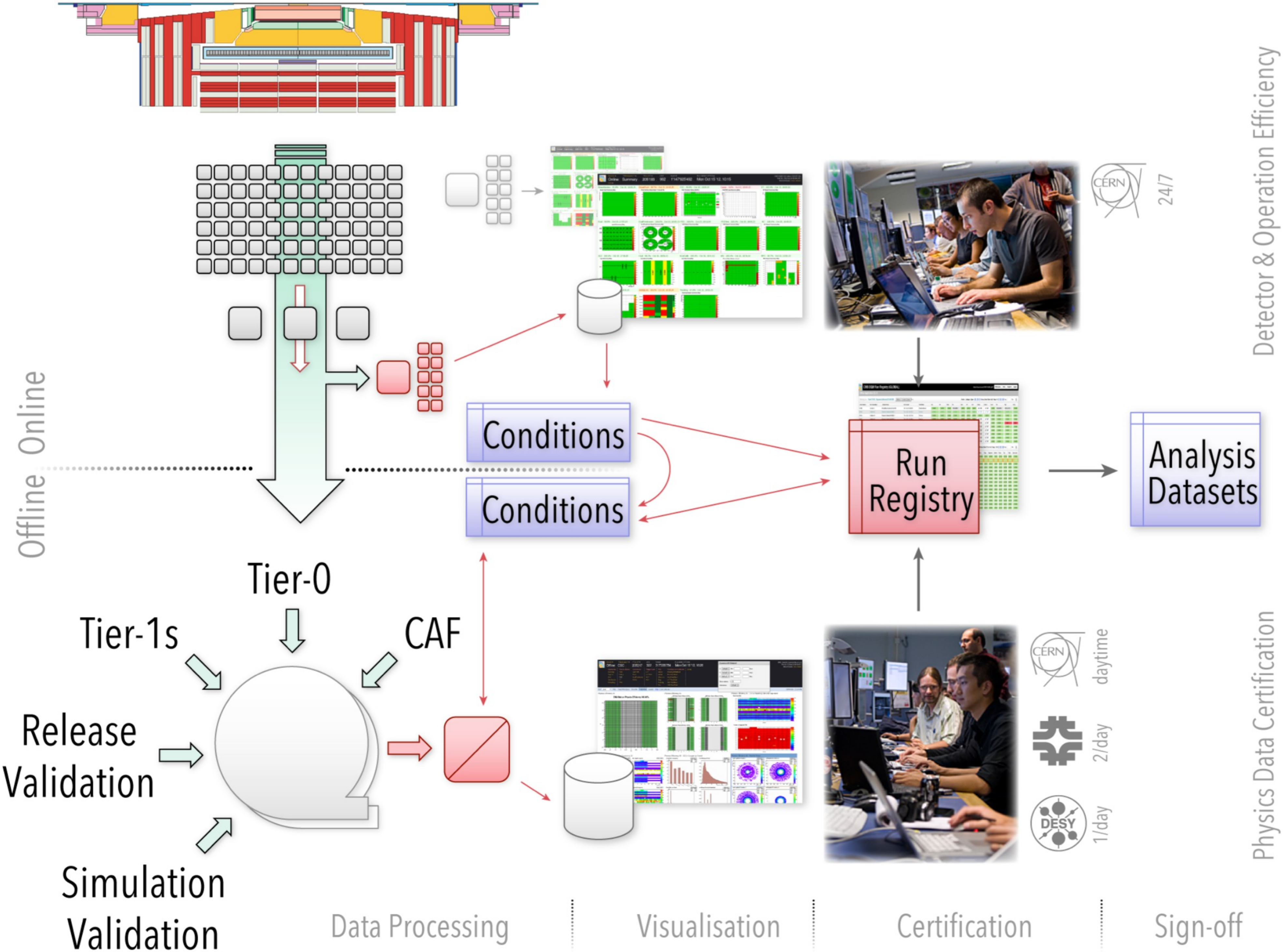
Contact : Pierini, De Guio, Vlimant

- Intensity of beam delivered by the LHC is decreasing with time
- Data quality of about ten sub-detectors is monitor by ~30s timeframe with tens of thousand histograms, trend plots, layout, summaries, ...
- Catches major issues
- Labor intensive task

- Minor defects are often discovered in the aftermath during final analysis
- Not enough time to humanly review all the plots to assess data quality
- Quality control wide spread in industry
- Approaches to reduce manpower overhead
  - Review data integrity all thousands of indicators using big data mining technique
  - Train an algorithm on already certified/rejected data

- Looking forward to industry partner to
  - Develop quality control applications to unique and complex LHC data
  - Deploy infrastructure to meet the challenge
- On-going project with Yandex. Other partners most welcome.



Detector & Operation Efficiency
24/7
Online
Offline
Conditions
Conditions
Run Registry
Analysis Datasets
Tier-0
Tier-1s
CAF
Release Validation
Simulation Validation
daytime
2/day
1/day
Physics Data Certification
Data Processing
Visualisation
Certification
Sign-off

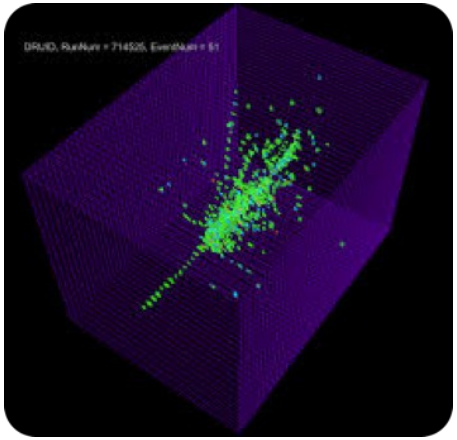
Data Reconstruction

Calorimeter Pattern Recognition

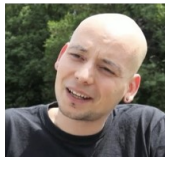
Contact : Pierini, Vlimant

- ✔ Particles emerging from collisions are brought to a stop in calorimeters
- ✔ **Shower of particles** are created in the detector
- ✔ Depth, intensity and topology of the shower are characteristic of particle type and kinematic
- ✔ Timing of the energy deposition can be measured and used for disambiguation of overlapping path
- ✔ Accuracy of measurements is correlated with the **granularity of the detector**
- ✔ Next generation of calorimeter will be way more granular than contemporary ones
  - ✘ Conventional **algorithm cannot cope** with the increase granularity

- New algorithm do not need to be better, it **needs to be faster**
- **Pattern recognition science** has boomed in industry over the last decade
- Particle identification with deep learning pattern recognition shows promising results
  - ➢ Need to bring it to the next level
  - ✘ Deep learning requires **computation acceleration**



➔ Looking forward to an industry partner to
- ➢ Leverage **modern pattern recognition** technique
- ➢ Help **applying deep learning** to a unique dataset
- ➢ Help setting up a “get-started” **cluster for physicists**
- ➢ Provide expertise in building a **deep-training facility**



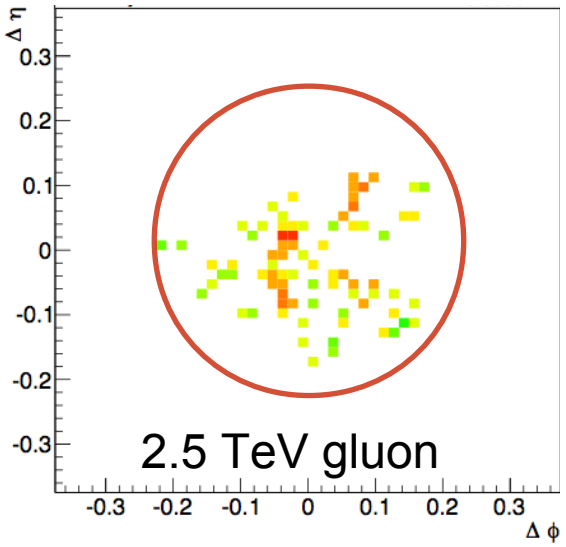
Boosted Objects Imaging

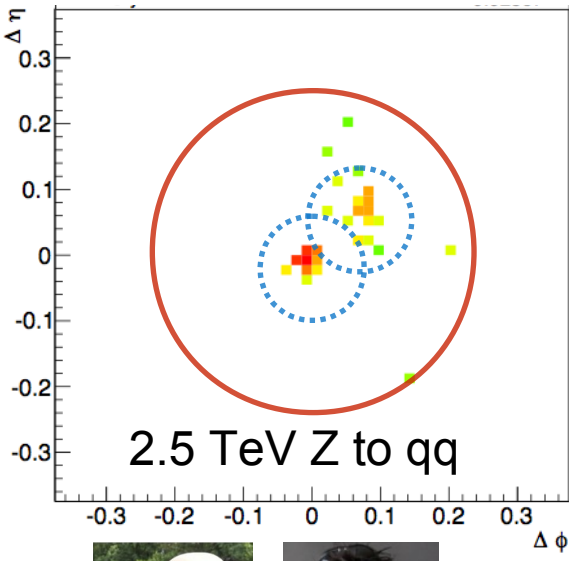
Contact : Pierini, Vlimant

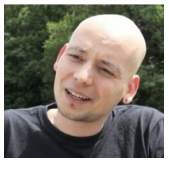
- ✔ Decays of high momentum particles are boosted along the initial direction of the particle
- ✔ Identifying these objects is an essential port of the LHC physics program
- ✘ Technique exist to disambiguate, falling short on dense “jet” cases with many overlapping particles
- ✘ Identification at the level of event filtering at high rate is impossible due to algorithm computation

- New algorithm do not need to be better, it **needs to be faster**
- Particle identification with deep learning pattern recognition shows promising results
  - ➢ Need to bring it to the next level
  - ✘ Deep learning requires computation acceleration

➔ Looking forward to an industry partner to
- ➢ Leverage **modern pattern recognition** technique
- ➢ Help **applying deep learning** to a unique dataset
- ➢ Help setting up a “get-started” **cluster for physicists**
- ➢ Provide expertise in building a **deep-training facility**








Data Processing

Computing Resource Optimization

Contact : Bonacorsi, Lange, Vlimant

- LHC Grid is composed of multiple computing center of various size, reliability, dedication, ...
- HEP data processing is mostly data intensive, in some cases not even suitable for remote access computing
- Distributed computing system have inherent failure rate
- Subtle balance between reading the data from remote and transferring data with respect to network usage and computing efficiency

- Name of the game is optimizing data movement, workload and network links to achieve the **best throughput**. Present in industry to some extend
- Exploring on how to **control this complex system**
- **Mining monitoring big data** using ML technique

→ Looking forward to industry partner to
  - Solve the **scheduling problem** and **reduce processing latency**
  - Pioneer in **controlling a worldwide computing system**



Predicting Data Popularity

Contact : Bonacorsi, Boccali, Kuznetsov

- ✔ Thousands of users need to access thousands of dataset across the LHC grid
- ✔ Data location and replication factor matters for fast turn-over
- ✔ Disk space is costly and requires a tight management. Cannot afford several replicas of all datasets

- We are exploring possibilities to predict popularity of datasets
  - **Extracting trends** from dataset usage
  - **Predicting dataset relevance** prior to usage
- Quite **common in industry** (amazon, ...)
- Seed to dynamic data placement system
  - Reduce transfer latencies
  - Speed-up analysis turn-over
- Initial studies using classification on meta-data indicate good accuracy (http://arxiv.org/abs/1602.07226)
  - ✗ Cost of training is prohibitive
  - ✗ Training stability far in the future is compromise

→ Looking forward to industry partner that can elevate
  - Trend extraction and prediction algorithms
  - Dedicated **large scale training platform** (Spark ML, Azure, ...)



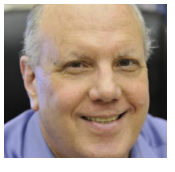
Intelligent Networking

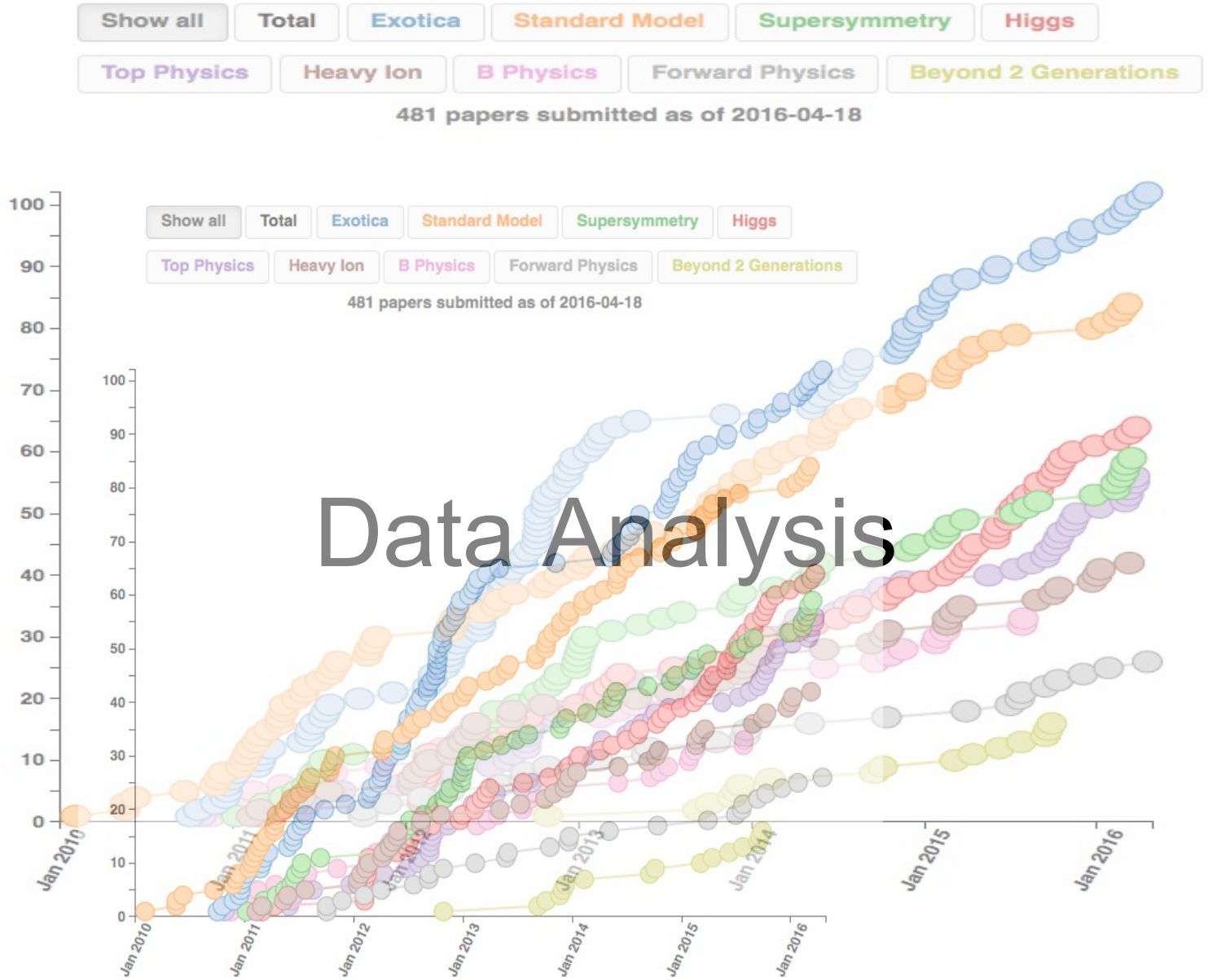
Contact : Newman, Vlimant

- Entering the exa-scale era with the HL-LHC in 2025
- Worldwide networks have finite bandwidth
- Dynamic circuit allows to prioritize and reserve traffic
- Emerging software defined network (SDN) community
- Trends of data movement and network utilization require a change in operation

- Dynamically shape the network topology to the needs
- Consider non-network boundary conditions and requirements
  - Computing-storage-network elements optimization

- Looking forward to industry partner to
  - Bring network optimization solutions to the scientific network
  - Help instrumenting sites with state of the art network elements
  - Participate in exa-scale networking demonstration
  -



Data Analysis

Exploiting Scientific Knowledge

Contact : Elmer, Kuznetsov, Vlimant

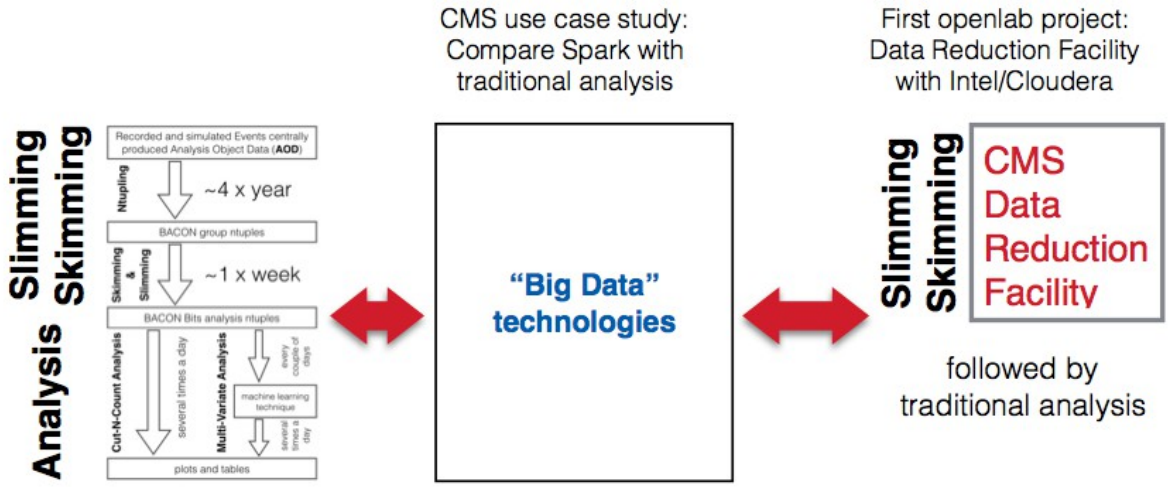
- ✔ HEP experiment build one-of-a-kind instrument, used and maintained over decades
- • These are used by long term staff and very large **number of transients** (e.g students) which work for a few years and move on
- ✗ Relevant information is often **unstructured and heterogeneous**: notes, twikis, forums, e-log, papers, theses, databases ...
  - ◆ Heterogeneous relevance of data (not everyone's answer is relevant, ...)
  - ◆ Heterogeneous information content (text, table, diagrams, histograms, ...)
  - ◆ Heterogeneous source of information (twiki, forum, data services, ...)
- ✔ Useful for continued operation, training and significant potential for use in data and knowledge preservation

- ➔ Looking forward to industry partner
  - ◆ to help **extract the knowledge** from highly heterogeneous substrate
  - ◆ To build and **maintain knowledge bases**, taxonomies, ...
  - ◆ Exploring how it can be used to **support science** in the long term



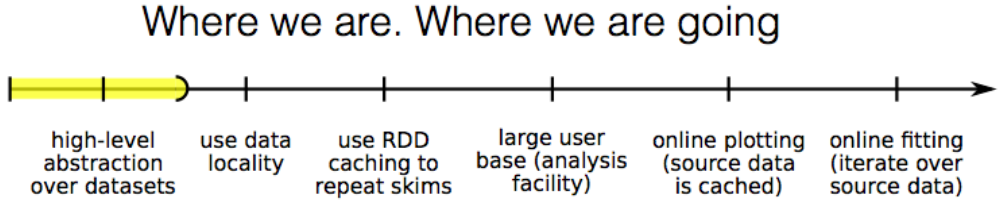
Big Data Reduction & Analysis

Contact : Gutsche, Fisk, Pivarski



- ✓ Produce Petabytes of analysis dataset efficiently
- ✓ Analyze Petabytes of data with as little latency as possible

- System needs to support **thousands of concurrent analysis**
- CMS analysis with Big Data technology
  - **Demonstrator with Spark** in progress



- → Looking forward to industry partners for
  - Model for **easy-to-deploy solution**
  - Balancing data locality and computing load
  - Combining C++(ROOT) and Java
- First collaboration with Intel/Cloudera. Other partners are most welcome

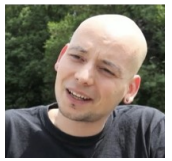
Model Independent Search

Contact : Pierini, Vlimant

- After the Higgs discovery the LHC entered exploratory phase without a concrete golden model to search for
- Plethora of signals to search for : “something” in a hay stack
- Inclusive analysis for family of signal processes does not have full coverage
- Analysis has to be tuned for sensitivity to the specifics of the signal
  - Time consuming, labor intensive

- Pilot project with unsupervised learning (SOM, NADE, ...) showed promising results
  - Need to take it to the next level

- Looking for industry partner to
  - Extract **categories of unforeseen event** using
  - Develop algorithm for **detecting rare patterns**



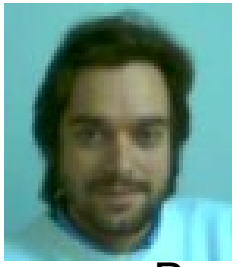

Summary

Unique high energy physics challenges.

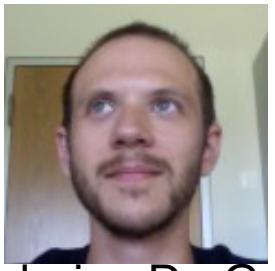
Project proposals accompanied by encouraging supplementary work.

Looking forward to working on these with Openlab partners.

Dustin Anderson
Caltech
dustin.james.anderson@cern.ch



Tommaso Boccali
Univ. Pisa/INFN
Tommaso.Boccali@cern.ch

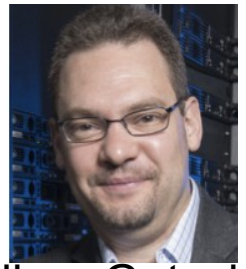


Federico De Guio
CERN
federico.de.guio@cern.ch

Daniele Bonacorsi
Univ. Bologna/INFN
Daniele.Bonacorsi@bo.infn.it

Peter Elmer
Princeton
Peter.Elmer@cern.ch

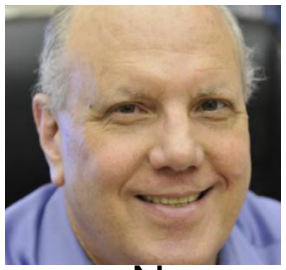
Marco Meoni
Univ. Pisa/INFN
marco.meoni@cern.ch



Oliver Gutsche
Fermilab
gutsche@fnal.gov

Ian Fisk
Fermilab
ian.fisk@cern.ch

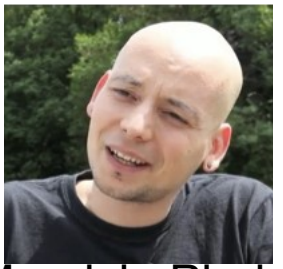
Valentin Kuznetsov
Cornell University
vkuznet@gmail.com



Harvey Newman
Caltech
newman@hep.caltech.edu

Jean-Roch Vlimant
Caltech
vlimant@cern.ch

David Lange
Princeton
David.Lange@cern.ch



Maurizio Pierini
CERN
Maurizio.Pierini@cern.ch

Jim Pivarski
Princeton
jpivarski@gmail.com