CERN openlab Overview

SKA meeting

May 17, 2017

Maria Girone

CERN openlab CTO

Introduction

- CERN openlab has been created to support the computing and data management goals set by the LHC programme
  - 15 years of innovative projects between CERN groups, experiments and leading IT companies
- CERN openlab is working to solve some of the key technical challenges facing the LHC
  - Mutual benefit for industry and research communities
  - Collaborating with other research communities **beyond HEP**
- Ever-increasing interest in CERN openlab
  - well established mechanism at CERN of partnership between industry and research communities
  - Provides an umbrella to facilitate industry interactions: IP, NDA, collaboration agreements

CERN openlab

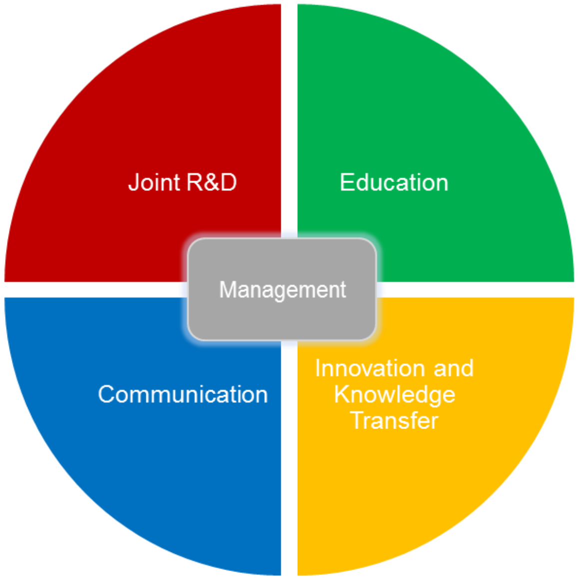
Members

Partners	intel, ORACLE, SIEMENS, HUAWEI
Contributors	rackspace the open cloud company, SEAGATE, IDT, BROCADE
Associates	Yandex, ComTrade
Research	EMBL-EBI, GSI, Newcastle University, Kazan Federa Univers, Innopolis University, King's College London, INFN Istituto Nazionale di Fisica Nucleare, Fcimpulse

Objectives

A science – industry partnership to drive R&D and innovation with over 15 years of success

- **Evaluate** state-of-the-art technologies in a challenging environment and improve them
  Test in a research environment today technologies will be used in many business sectors tomorrow
- **Train** next generation of engineers/employees.
  Promote education and cultural exchanges
- **Disseminate** results and outreach to new audiences
- **Collaborate** and exchanges ideas to create knowledge and innovation



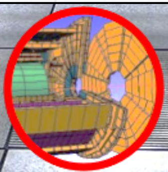
CERN openlab V Research Areas

Data acquisition and filtering
Collecting data

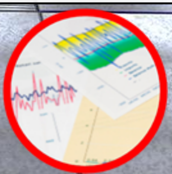
Networks and connectivity
Connecting resources

Data storage architectures
Storing and serving data

Compute management and provisioning (cloud)
Managing resources for processing



Computing platforms, data analysis, simulation
Improving processing and code efficiency



Data analytics/Machine Learning
Extracting information

Medical applications

Overview of our Current Projects

> **Data acquisition**

> **Networks and connectivity**

> **Data storage architectures**

> **Compute provisioning and management**

> **Computing platforms**

> **Data analytics**

intel — High-Throughput Computing Collaboration (HTCC)
IDT — RapidIO for data acquisition

BROCADE — Flow Optimizer Software

ORACLE — Database Technology and Monitoring
SEAGATE — Alternative Storage Architecture
COMTRADE — EOS Productisation

rackspace (the open cloud company) — Containers
ORACLE — Database Cloud
ORACLE — Java EE
HUAWEI — OpenStack at Scale

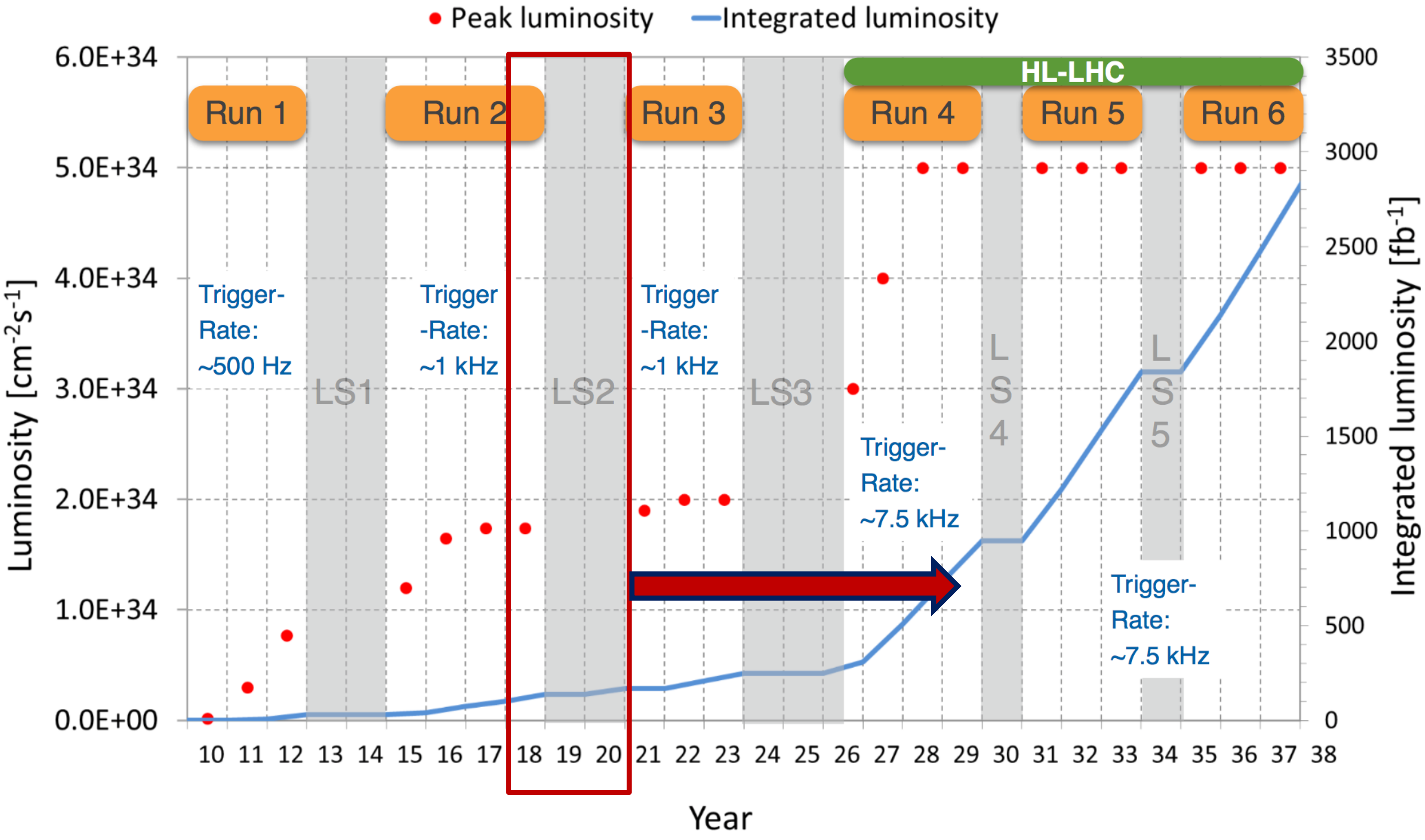
intel — Code Modernization (Geant V, BioDynamo)
ARM porting, optimization and benchmarking

intel — Big Data Analytics
ORACLE — Analytics-as-a-service
SIEMENS — Industrial Control and Monitoring
Yandex — Data popularity and anomaly detection

Currently 27 FTEs, 16 ongoing projects

~40 openlab summer students

LHC Schedule



Run 3 Alice, LHCb upgrades

Run 4 ATLAS, CMS upgrades

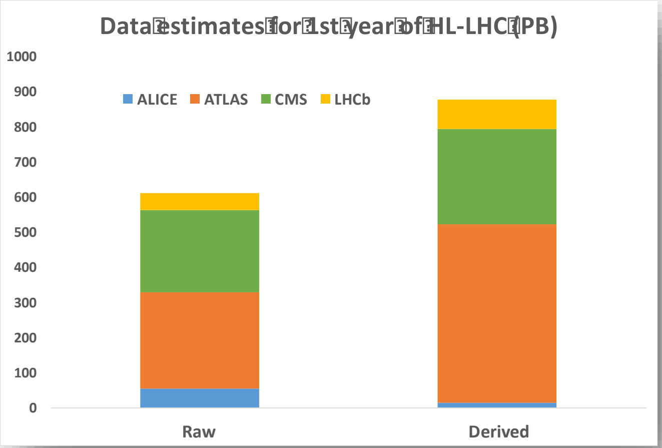
LHC Run3 and Run4 Scale and Challenges

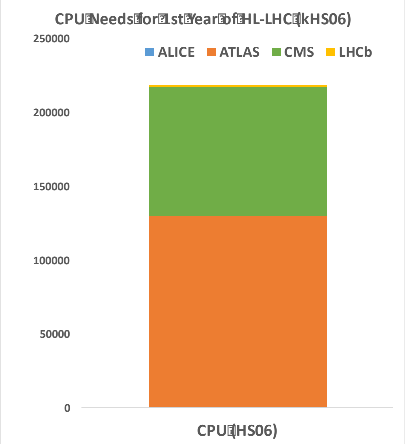
2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 …… 2030?

First run | LS1 | Second run | LS2 | Third run | LS3 | HL-LHC | FCC?

- Raw data volume for LHC increases exponentially and with it processing and analysis load
- Technology at ~20%/year will bring x6-10 in 10-11 years
- Estimates of **resource needs at HL-LHC x10** above what is realistic to expect from technology with reasonably **constant or decreasing funding**

Technology revolutions are needed



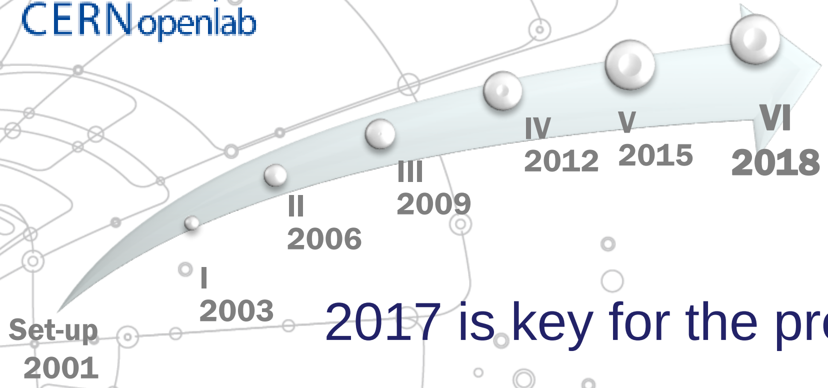


Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016

Courtesy of I. Bird

Preparing for the Next Phase



2017 is key for the preparation of the next phase (2018-2020)

- We had a series of brainstorming workshops in March and April 2017
  - Engaging the scientific community to present requirements and needs
  - Identify with industry and research partners potential solutions and common use cases and challenges

Phase VI Research Areas (2018-2020)

- **Data Analytics and Machine Learning**
  - Physics (simulation, reconstruction, particle classification, trigger, data quality, anomaly detection, ..)
  - Engineering (Control systems, infrastructure optimization, …)
- **Computing Platforms and Software**
  - Architectures
  - Software modernization/acceleration
- **Data Center Technologies and Infrastructures**
  - Networks
  - Cloud Computing
  - Storage and databases
  - Data Center Architectures (disaggregation)

Data Center Technologies and Infrastructures

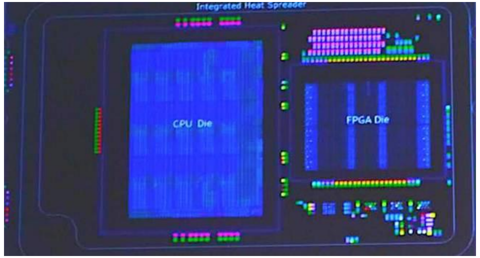
- **More efficient collection, processing and storage of data**
  - Coupling of processing and storage (e.g. use of technologies like NVRAM)
  - Coupling of processing and network (e.g. use of high speed/low latency interconnects)
  - Storage Improvements (e.g. affordable, reliable and long lasting storage); evolution of the tape archive model

- **CERN is considering building a new large scale data center at the Prevessin site**
  - Ambitious goals on efficiency, capacity and density to handle many of the functions currently provided at the experiment premises
  - Looking at processing and storage models for on-site systems and dynamic resource provisioning expansion into remote facilities

Compute Platforms and Software

LHC experiments are looking **alternative architectures**

- Adoption of GPU, co-processors, or FPGA technologies, ..



- Experiments will need to process more events that are much more complex
  - Existing technologies are growing in number of cores and instruction sizes
    - Need to more efficiently parallelize the code, but also execute more instructions in parallel per cycle
  - Huge effort needed in **code modernization** and code evolution

- Blurring the online and offline boundaries
  - Allows more events to be collected if processing happens only once
  - Streamlines event collection and archiving

Data Analytics and Machine Learning

The community is opening to modern techniques in the field and turns to industry who is leading the way

- Interest from (young) researchers to learn new skills
- Wide interest for applications in physics, engineering and other sciences

Proposed Proof of Concepts (POCs)

- Anomaly Detection and Preemptive Maintenance
- Resource Infrastructure Optimization
- Streamlining analysis access
- Event Categorization and Triggering
- Physics Object Identification (particle reconstruction)

Where ML can help in a LHC Experiment

Online Data Taking (real time)

We need to select which events to keep

We need to make sure the detector is working properly

We need to access the quality of the data we take

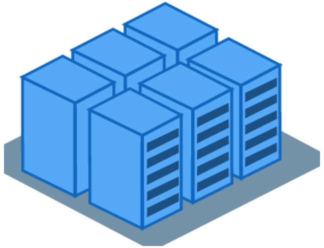
Offline Event Processing (centralised)

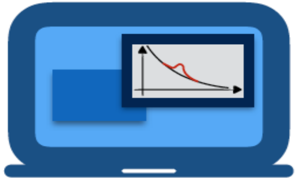
Charged particles as helixes (connecting dots)

Clusters of energies in calorimeters

Muons = inner track + outer tracks

Close-by particles ➡ jets





Data Analysis (by users)

Analyses targeting specific physics processes

Event selection to enhance Sig/Bkg ratio

Statistical analysis to highlight presence of a signal

Courtesy of M. Pierini

Where ML can help in a LHC experiment

Online Data Taking (real time)

Fast trigger algorithms for topology classification based on image recognition

Fast reconstruction algorithms (clustering, tracking, classification based on trained networks)

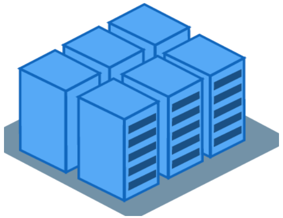
Data quality monitoring and detector operations (correlating data from many sources)

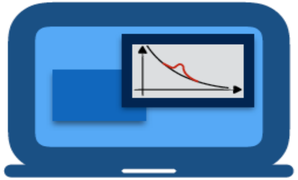
Offline Event Processing (centralised)

Event indexing based on topology classification (learned algorithms trained on simulation)

Fast simulation (optimized for agreement between simulation and data)

Fast reconstruction algorithms (image recognition and deep networks)





Data Analysis (by users)

Data reduction with improved data analytics

Jet Tagging based on learned algorithms

Searches for new physics

Quality monitoring

Courtesy of M. Pierini

Opportunities for Collaboration

- The LHC experiments need help to adopt new technologies and new techniques to meet the goals for HL-LHC
  - New data centers will be built
  - New architectures will be tested and deployed
  - New techniques will be advanced in machine learning and advanced data analytics
- This is one of the largest challenges in **Scientific Big Data**
  - Unique for consistency and reproducibility
  - It is a good testbed for developing and demonstrating tools at scale

DEEP-EST

- DEEP – EST EC funded project at Jülich

(Dynamical Exascale Entry Platform - Extreme Scale Technologies) I

- A HPC system prototype designed for high performance Data analytics workflows (HPDA)

- CERN and ASTRON are participating for application testing
  - CERN for CMS Physics Data Analysis Reduction
  - ASTRON (SKA Data Analysis) J. Romein

- The project will start in July 2017
  - Possibility to collaborate on this?

Conclusions

- There is tremendous interest from industry to work together
  - Both industry and our community have a lot to gain
  - Industry via CERN openlab can offer opportunities for education and training in areas where expertise is needed (parallelization, ML, DA, ..)
- The goal of the next years of R&D is to design deployable solutions for the HL-LHC
  - There are synergies with the challenges highlighted by SKA
- CERN openlab provides the ground for collaboration of researchers (not only HEP)
  - Very welcome by industry as well

Backup slides

Where CERN openlab is helping in ML

- We have a number of ongoing projects with Intel, Oracle, Siemens and Yandex
  - Anomaly detection for industrial controls , data certification and data quality monitoring, physics data reduction, ML in (fast) simulation – Geant V, …

- We have established active discussions and work with other companies, including IBM, Microsoft, E4, NVIDIA, Google, Atos, …
  - Targeting to next phase (2018-2020)

Why the Interest from Industry?

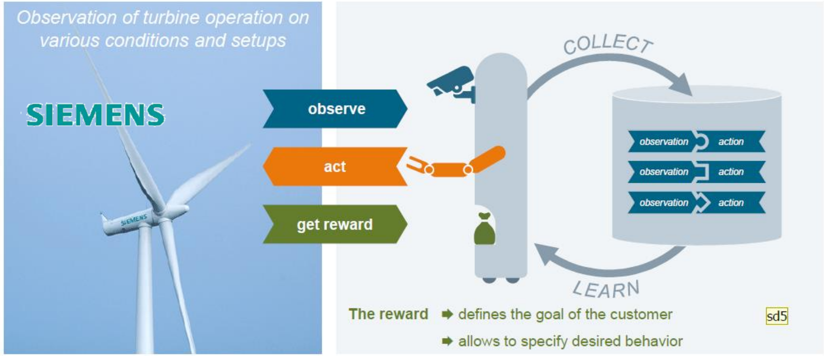
- Machine Learning is a massive emerging market
  - Used in social media, e-commerce, finance, etc.
- Lots of development activities in industry
  - to have the best techniques and support libraries
  - to ensure that the next generation of hardware platforms are designed for ML

- We can offer challenging environments and smart people
  - but we need to find common challenges

Monitoring, Automation , Anomaly Detection

- Network security and fraud detection
- Industrial monitoring and predictive failures
  - Looking at optimizing performance of complex systems
    - Minimize costs and improve resource utilization

Reinforcement learning

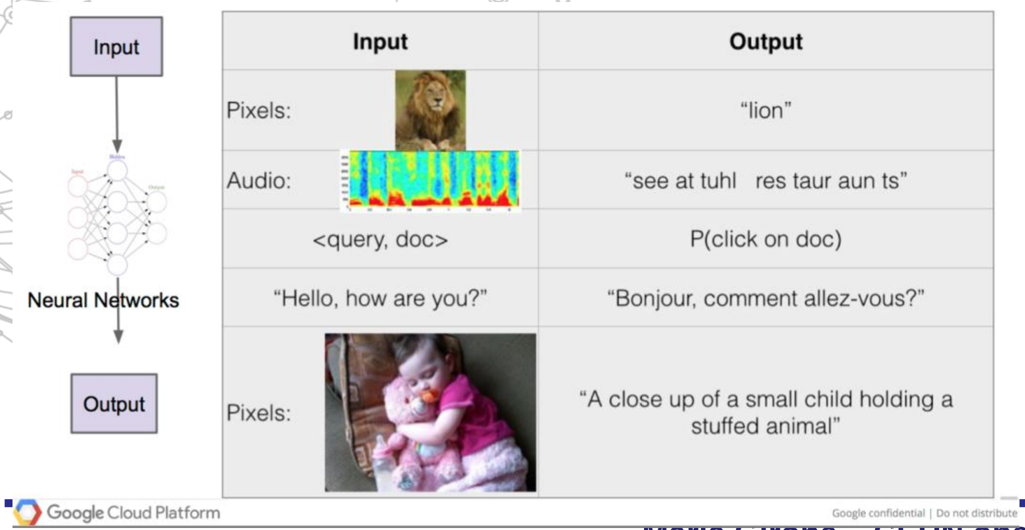


- Detector Health
  - Complex system monitoring to minimize downtime and reduce operations costs
- Resource Utilization
  - Scheduling
  - Data placement
  - I/O optimization
- Quality Monitoring/Data certification
  - Automated assessment of quality of the data produced, online and offline
- LHC magnets, industrial controls, …

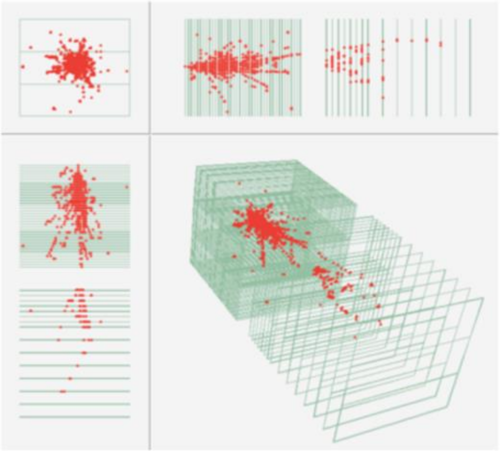
Visualization

- Image and object recognition
  - Sort image catalogs
  - Object recognition
  - Computer vision

- Object reconstruction and identification
  - Physics objects identification using visualization techniques

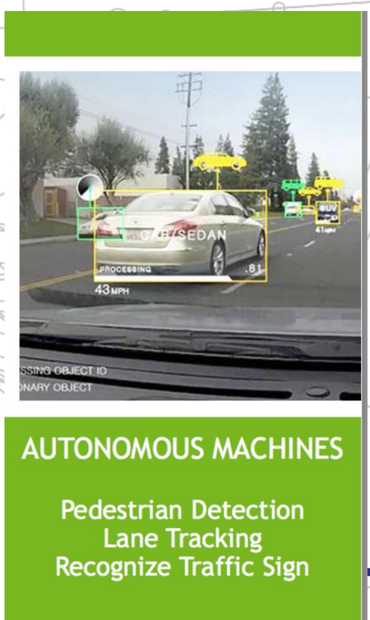


Input	Output
Pixels:	"lion"
Audio:	"see at tuhl res taur aun ts"
<query, doc>	P(click on doc)
"Hello, how are you?"	"Bonjour, comment allez-vous?"
Pixels:	"A close up of a small child holding a stuffed animal"



Decision Making

- During workshops we spoke about whether the speed we need to make some decisions is challenging

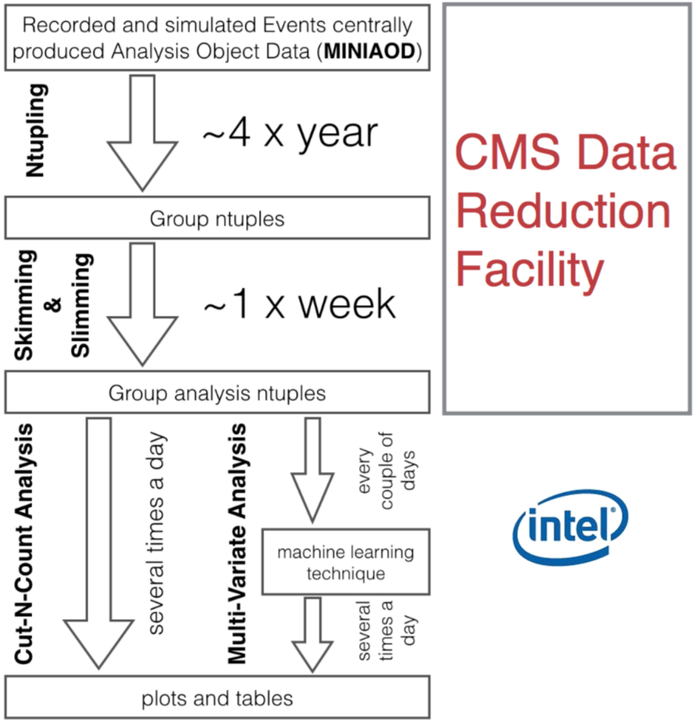


- Self Driving Cars
  - Decisions needed in a few milliseconds
- Triggering
  - L1 trigger processing time~10µs
  - HLT processing time ~30ms

Big Data

- The size and complexity of industrial datasets have grown
  - HEP has PB and industry has EB

- New toolkits and systems collectively called “Big Data” technologies have emerged and we are investigating within CERN openlab how to transform and reduce time and resources needed to physics



Hardware

- Current industry investments are in hardware developments aimed to improve efficiency on machine learning and data analytics
  - Many-core co-processors, FPGA, GPUs, ultra-fast storage close to the CPU
- Industry has recognized the need to work together on a common R&D programme to maximize the capabilities of new hardware technologies
  - **heterogeneous platform test environments to our community are key**
    - Important to CERN IT and CERN openlab