



# Openlab Machine Learning and Data Analytics Workshop Summary

HSF Meeting

LAL Orsay, May 3<sup>rd</sup> 2016

Maria Girone

CERN openlab CTO



CERNopenlab

# About the Event

## CERN openlab Machine Learning and Data Analytics workshop

29 April 2016  
CERN  
Europe/Zurich timezone

  
**Search**

Material at <https://indico.cern.ch/event/514434/timetable/#20160429.detailed>

Technical notes: M. Martin Marquez, A. Romero Marin

	<b>CMS: ML and DA Challenges</b> <i>Dr. Jean-Roch Vlimant</i> 31-3-004 - IT Amphitheatre, CERN 09:10 - 09:55	
10:00	<b>LHCb: ML and DA Challenges</b> <i>Andrey Ustyuzhanin</i> 31-3-004 - IT Amphitheatre, CERN 09:55 - 10:25	
	<b>Networking Coffee</b> 31-3-004 - IT Amphitheatre, CERN 10:30 - 11:00	
11:00	<b>ATLAS: ML and DA Challenges</b> <i>David Rousseau</i> 31-3-004 - IT Amphitheatre, CERN 11:00 - 11:45	
	<b>ALICE: ML and DA Challenges</b> <i>Michele Floris</i> 31-3-004 - IT Amphitheatre, CERN 11:45 - 12:15	
12:00	<b>Challenges for Industrial Control Systems</b> <i>Manuel Gonzalez Berges</i> 31-3-004 - IT Amphitheatre, CERN 12:15 - 12:35	
	<b>Machine Learning and Data Analytics at Intel</b> <i>Marie-Christine Sawley</i> 31-3-004 - IT Amphitheatre, CERN 13:30 - 13:50	
14:00	<b>Machine Learning and Data Analytics at Cloudera</b> <i>Tom White</i> 31-3-004 - IT Amphitheatre, CERN 13:50 - 14:10	
	<b>Machine Learning and Data Analytics at Siemens</b> <i>Volker Tresp</i> 31-3-004 - IT Amphitheatre, CERN 14:10 - 14:30	
	<b>Machine Learning and Data Analytics at IBM</b> <i>Costas Bekas</i> 31-3-004 - IT Amphitheatre, CERN 14:30 - 14:50	
	<b>Machine Learning and Data Analytics at Google</b> <i>Alex Osterloh</i> 31-3-004 - IT Amphitheatre, CERN 15:20 - 15:40	
	<b>Machine Learning and Data Analytics at Microsoft</b> <i>Alexandre Gattiker</i> 31-3-004 - IT Amphitheatre, CERN 15:40 - 16:00	
16:00	<b>Machine Learning and Data Analytics at Cisco</b> <i>Enzo Fenoglio</i> 31-3-004 - IT Amphitheatre, CERN 16:00 - 16:20	
	<b>Machine Learning and Data Analytics at Yandex</b> <i>Andrey Ustyuzhanin</i> 31-3-004 - IT Amphitheatre, CERN 16:20 - 16:40	

# Experiment Plans

- All LHC experiments presented in the morning
  - There was significant overlap in the experiment plans and needs in the areas of machine learning and data analytics
    - › Event Categorization and Triggering
    - › Physics Object Identification
    - › Anomaly Detection
    - › Resource Optimization
    - › Stream lining analysis access

# Event Categorization and Triggering

- The challenges of the HL-LHC (ATLAS,CMS), and of Run3 (ALICE and LHCb), were highlighted

- Data rates and event complexity both go up dramatically

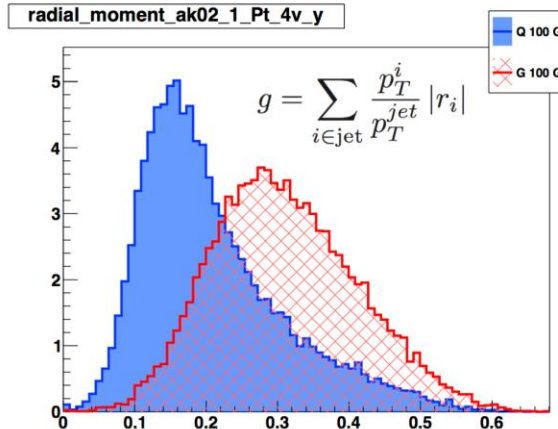
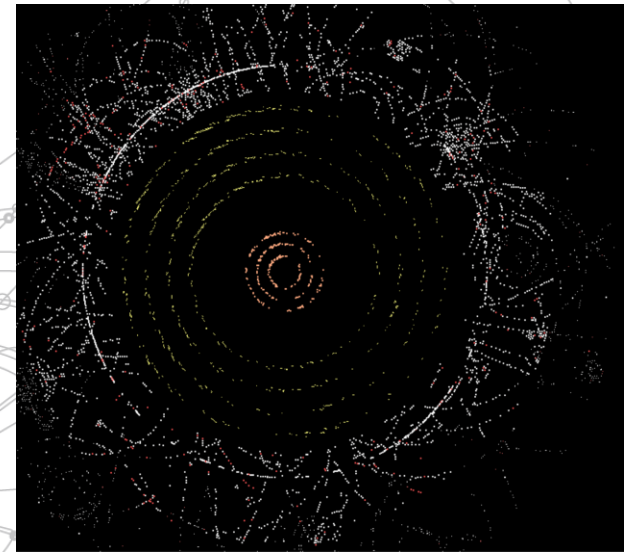
- CMS is investigating using machine learning techniques to perform real-time event categorization

- This would be a natural extension of their data scouting work

- LHCb already performs a similar selection as part of their successful “Turbo Stream”

# Physics Object Identification

- Physics Object identification was raised by ALICE, ATLAS, CMS and LHCb
  - It is a potentially interesting machine learning problem with the potential to improve analysis
    - › Considered for finding the tracks to be given to the track fitter
    - › Discriminating types of jets



# Resource Optimization and Anomaly Detection

■ Several experiments proposed machine learning applications for computing resource optimization

- How to best lay out data across of complex and distributed environment?
- How to schedule processing?
- How to find grid intrusions?

■ This area had the most obvious overlap with ongoing industry projects

- Could be done as a common project across the 4 experiments

# Data Analysis

- The increase in data volume expected by the experiments in Run3 and Run4 changes the operations mode and opens up possibilities for analysis
  - In ALICE and LHCb events will leave the detector essentially reconstructed with final calibrations
    - › Analysis can start immediately (maybe even online)
  - ATLAS and CMS will both have much higher triggers and a desire to streamline analysis
- Interest to look at industry tools for improved data analysis
  - SPARK, Hadoop, etc.
    - › CERN openlab is helping to set up a project with Intel. Interest by other companies
- Community building around analysis challenges, e.g.
  - <https://www.kaggle.com/c/flavours-of-physics>

- In the end there were 9 presentations from industry
  - Intel, Cloudera, Siemens, IBM, Google, Microsoft, Cisco, Yandex, and Nvidia
    - › Some CERN openlab partners but a few new faces
  - Good balance of groups who have historically made hardware and groups that have historically made solutions
    - › There is a lot of overlap these days



# Resource Optimization

- There was an interesting presentation from Siemens on their work in machine learning resource optimization for industrial applications
  - Controls and monitoring for Steel Mills and wind turbines
  - Online machine learning to better use resources and optimize the systems
    - › Many of the same issues we face with resource optimization, but for companies with more money at stake and well established methods



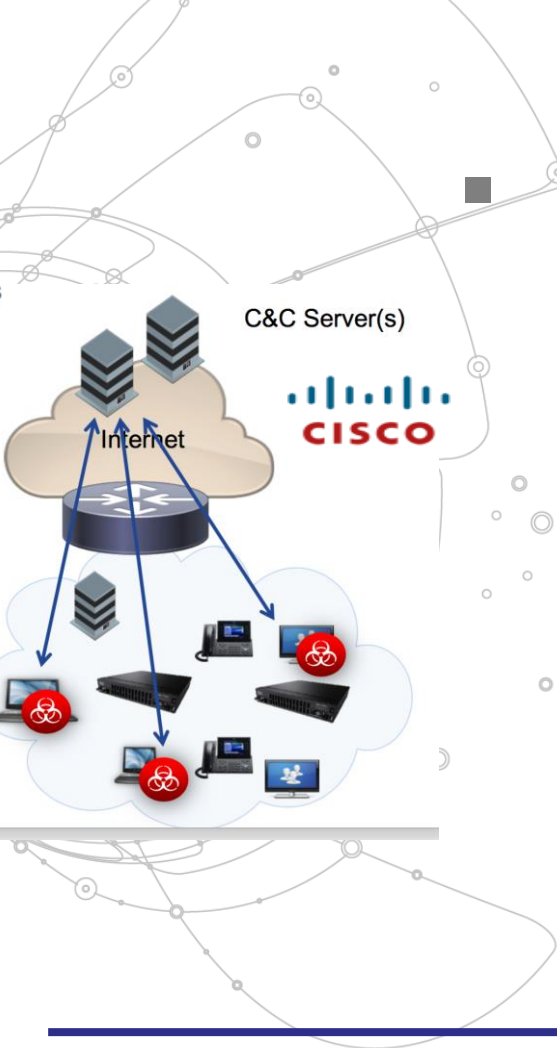
Observation of turbine operation on various conditions and setups



# Anomaly Detection

■ Several companies are looking at anomaly detection for security and networking

- Systems learn what regular operations look like and then events outside the norm are flagged
  - › Used for network operations and grid security

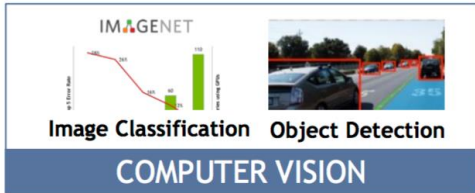


Google

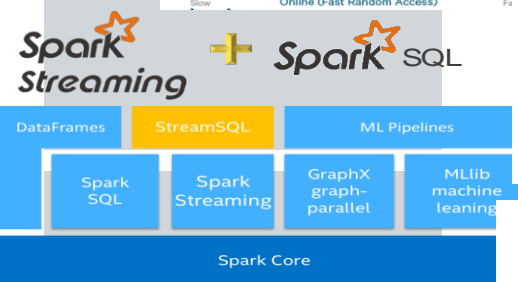
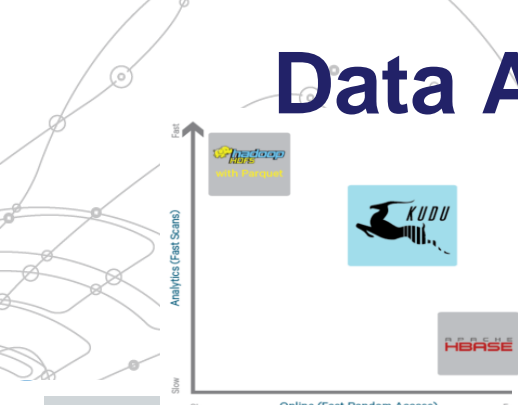
# Physics Object Identification

Industry is not doing a lot of physics

- What they are doing is a lot of computer visualization
  - › Used in self driving cars and image sorting (Google)
  - › Object detection (Nvidia)
  - › Image search (Yandex)
- We need to learn how to phrase our physics identification problems as computer visualization problems
  - › Once a machine can identify an event by looking at it, a lot of possibilities open up

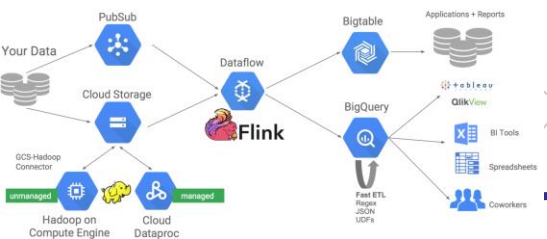


# Data Access and Analysis Optimization



You may view the project here:  
<https://github.com/Intel-bigdata/spark-streaming>

Enterprise Big Data Architecture on Google



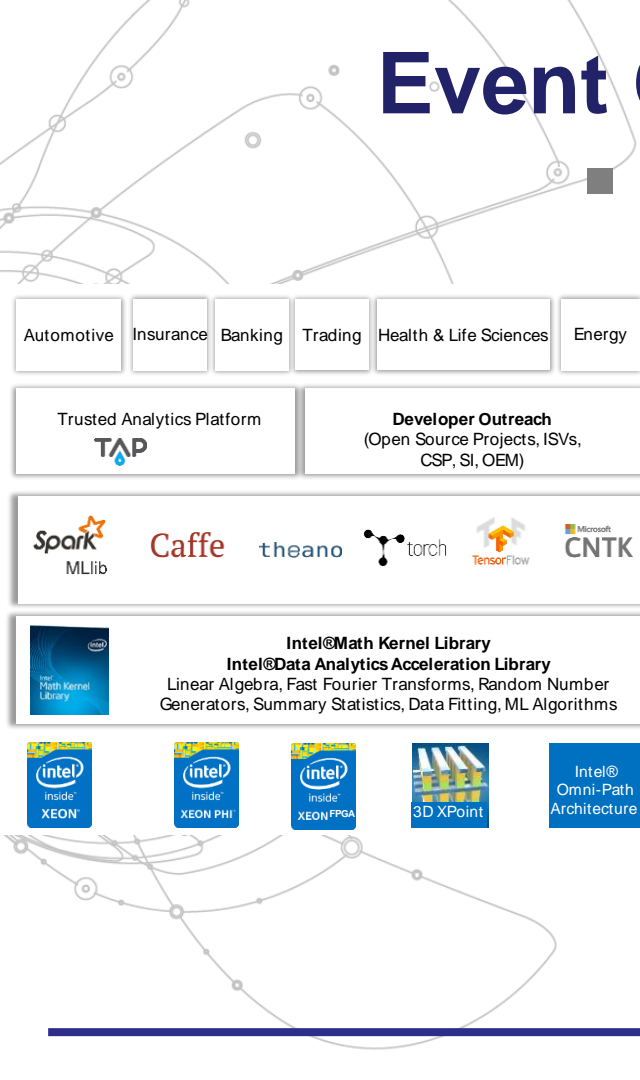
[panda.cisco.com](https://panda.cisco.com)












- This is one of the active areas of development
  - Companies like Cloudera specialize in it (Hadoop, Spark, and additions)
  - Intel, IBM and Microsoft both have research in this area
  - Cisco has developed Platform for Data Analytics (PANDA)
  - Google has done generations of products
- ## Most of the tools are open source
- Many are built on similar underlying components

# Event Categorization and Triggering

Industry is not working on anything as specific as high energy physics real time event classification

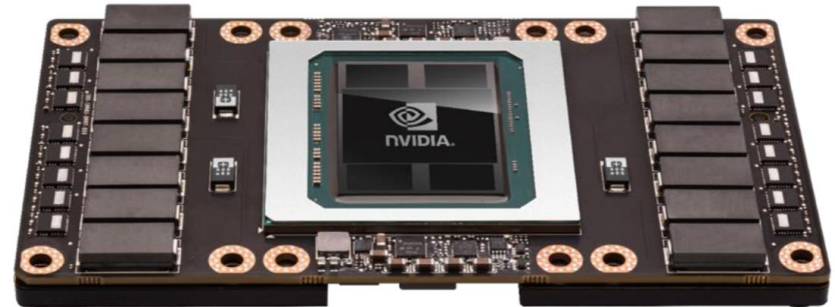
- However, there is a lot of progress in the field of deep learning and specifically unsupervised learning that is useful in this problem
- Industry developed frameworks to help facilitate deployment
  - › Google Tensor Flow, Nvidia DIGITS, Yandex ML-HEP are just three examples



Automotive	Insurance	Banking	Trading	Health & Life Sciences	Energy
Trusted Analytics Platform <b>TAP</b>			Developer Outreach (Open Source Projects, ISVs, CSP, SI, OEM)		
 Spark MLlib	 Caffe	 theano	 torch	 TensorFlow	 CNTK
Intel®Math Kernel Library Intel®Data Analytics Acceleration Library Linear Algebra, Fast Fourier Transforms, Random Number Generators, Summary Statistics, Data Fitting, ML Algorithms					
					

# Hardware Improvements

- In addition to the software and solution work, there are hardware improvements
  - New hardware from both Intel and Nvidia is applying a lot more computing capacity to these resource intensive applications





# Observations

Machine Learning and Data Analytics are big research areas for industry  
(excellent talks from all 9 industry participants)

- Big investments beyond what would be considered the core business
  - Intel has platforms for machine learning and big investments in data access
  - Cisco has a data analytics framework
  - Siemens have established themselves in machine learning
  - Nvidia is committed to scientific computing
- There is an interest in working with us
  - While our dataset size is large, it is not as diverse or complicated as many industry applications
  - We have an interesting challenge and a pool of smart people to help find solutions that are often applicable to other problems
  - We are a good place to train future data scientists

# Collaborating with Industry: CERN openlab

A unique science – industry partnership to drive R&D and innovation with over a decade of success

- Evaluate state-of-the-art technologies in a challenging environment and improve them

- Test in a research environment today what will be used in many business sectors tomorrow

- Train next generation of engineers/employees

- Disseminate results and outreach to new audiences





# The CERN openlab forum

- Working with industry is useful because they bring a lot of complementary capabilities and experience
  - However, they do not have the same history of collaboration that the collaborations have. It is a much more competitive environment
  - CERN openlab has a history and a structure with NDAs to help facilitate the interactions

■ The CERN openlab discussion with the experiments and industry on machine learning and data analytics went very well

- Significant overlap in the experiment use-cases in this area
- High level of activity in complementary areas from the industry community

■ It was a good first step and we should follow-up potentially exploring more limited areas in more depth

- Challenges in Machine Learning CERN openlab white paper