Outline

- **Science at the LHC**
  - Triggering on rare signals
  - Data processing and simulation
  - Data movement and computation
  - Search strategy

- **Data Science at the LHC a solution**
  - Advanced tracking algorithms
  - Object identification
  - Faster simulation
  - Low energy computation

- **HL-LHC The computational challenge**
  - The ever increasing event complexity

- **Machine learning in complex modeling**

- **Beyond the computation needs**
  - Model-independent searches

- **Summary**
HEP Complex Data

CMS Experiment at LHC, CERN
Data recorded: Thu Apr 5 05:47:32 2012 CEST
Run/Event: 190401 / 12540576
Lumi section: 75
Orbit/Crossing: 194958546 / 1347

2010  \(\langle PU \rangle = 2\)

2011  \(\langle PU \rangle = 7\)

2012  \(\langle PU \rangle = 21\)

Event taken at random (filled) bunch crossings
Finding Rare Signal

Many orders of magnitude rejection in order to select interesting events
Massively parallel electronic infrastructure makes a prime selection
Refined decision in a software defined trigger
Little processing time for selection: ML for faster algorithm
Data Flow

- **Hundreds of computer centers** (100-10k cores per site)
- 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 between disk storage
- Ground for **control with reinforcement learning**
Simulation of Collisions

- Most analysis have data driven background estimations
- Cross checks and analysis tailoring nevertheless require a large amount of simulated event for the main backgrounds

- **Simulating events is a costly process**
  - Physics process is computed from matrix elements and amplitude calculations
  - Material effect are successfully simulated using GEANT4
  - In-house emulation of digitization

- **Billions of CPUh spend in simulating Monte-Carlo for analysis**
  - Generative models with machine learning

[Diagram: Complex showering process for one Xtal out of 50000 for CMS]
Event Reconstruction

*From individual measurements in sub-detectors to kinematics and properties of particles created in collisions*
Charged Particles

- Particle trajectory bended in a solenoid magnetic field
- Curvature is a proxy to momentum
- Particle ionize silicon pixel and strip throughout several concentric layers
- **Thousands of sparse hits**
- Lots of hit pollution from low momentum, secondary particles

**Seeding**

**Kalman Filter**

- **Explosion in hit combinatorics** in both seeding and stepping pattern recognition
- **Highly time consuming task** in extracting physics content from LHC data
Cost of Tracking

- Charged particle track reconstruction is one of the most CPU consuming task in event reconstruction

- Optimizations (to fit in computational budgets) saturated

- Large fraction of CPU required in the HLT. Cannot perform tracking inclusively

- Extrapolation into HL-LHC era far surpasses growth in computing budget

- Need for faster algorithms

- Approximation allowed in the trigger
  - Apply machine learning to the challenge
Calorimeter Reconstruction

- Energy deposit per Xtal computed from timed samples
- Xtal energies collected in cluster and super-cluster
- Photon, electron, jets energy calibrated

- Multiple step reconstruction process
- **Becomes challenging with higher granularity**
  - Identification and regression: all-in-one with machine learning
Simulation of Collisions

- Most analysis have data driven background estimations
- Cross checks and analysis tailoring nevertheless require a large amount of simulated event for the main backgrounds

- **Simulating events is a costly process**
  - Physics process is computed from matrix elements and amplitude calculations
  - Material effect are successfully simulated using GEANT4
  - In-house emulation of digitization

- **Billions of CPUh spend in simulating Monte-Carlo for analysis**
  - Generative models with machine learning

PbW0₄ CMS, X₀=0.89 cm

Complex showering process for one Xtal out of 50000 for CMS
Search For New Physics

Higgs discovery : we knew what it would look like

New physics searches (Susy, ...) : we don't know what to expect.

→ Unsupervised machine learning
Challenges Ahead
HL-LHC Challenge

- 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
    - Address with machine learning algorithm at the HLT
- 50x in data volume
- Offline data processing
  - 20-45x larger time per event
  - Resource needs growth beyond prediction of growth in budget
  - **Machine learning algorithm for faster processing**
Machine Learning
Machine Learning in a Nutshell

“The science of getting computers to act without being explicitly programmed” - Andrew Ng (Stanford/Coursera)

• part of standard computer science curriculum since the 90s
• inferring knowledge from data
• generalizing to unseen data
• usually no parametric model assumptions
• emphasizing the computational challenges

Balazs Kegl, CERN 2014
Scene Labelling

Farabet et al. ICML 2012, PAMI 2013

- Group and classify what each pixel belongs to
- Real-time video processing with deep
  - Attribute each Xtal to a cluster
  - Attribute tracker hit to a charged particle
Image Captionning

Karpathy, Fei-Fei, CVPR 2015

- Create a description of images
  - Generate a decay process description from collision representation
Deep Learning Application
Advanced Tracking Algorithms

- Investigation may involve
  - Application of scene labeling to seed formation
  - Application of object detection to track assembling
  - More reference algorithms

- Medium/High risk, very high reward problem
  - Exploratory phase on the model definition
Jet Tagging

- Hadronic activity results in bundle of collimated particles
- The more energetic, the more collimated: W-jet
- With even higher energy, even mother particles are collimated top-jet, Higgs-jet
- Available discriminators are performing well. Not taking advantage of the full substructure of the jets
- Image processing methods are natural candidates to perform the classification

Small dataset, 11 categories
60% accuracy on gluon versus any quark. pre-preliminary
Tagging Boosted Object

Top Tagger arXiv: 1501.05968 Almeida, Backovic, Cliche, Lee, Perelstein

Neural net

W tagger arXiv: 1511.05190, Oliveira, Kagan, Mackey, Nachman, Schwartzman

Train

W to QCD discrimination
3D Calorimeter Imaging

100GeV Photon ≠ 100GeV Pi0

LCD Calorimeter configuration
http://lcd.web.cern.ch
5x5 mm Pixel calorimeter
28 layer deep
Photon and pion particle gun

2D x 25 channels
Convolutional NN + dense layers
Limited dataset
95% efficiency
5% fake
Generative Models

- Computer science can generate images, text, sound
- Performing image arithmetic
- Simulation of collision events is very computation intensive
- Faster simulation with such generative models
- Address computing bottleneck
- Enable science program
Deep Learning Hardware
Training and Inference

- GPUs are the workhorse for parallel computing
- Enable training large models, with large dataset
- Deep learning facility clusters

M40
- 7 TFLOps
- 5 k$
- 250W

Emergence of small GPU
- Not dedicated to training
- Strike the balance between Tflops/$ for inference
- Deployment on the grid

M4
- 2.2 TFLOps
- ? k$
- 50W
Neuromorphic Hardware

- Brain inspired **low power silicon** hardware
- Spiking neurons for general computation
- Demonstrated to perform well on **pattern recognition** problems
- **Unsupervised learning** capabilities on some models

- On-going collaboration with iniLab & INI Zurich
  - Aiming at application to calorimeter pattern recognition in level 1-2 of the trigger
  - Potential application as accelerator card

http://www.nature.com/articles/srep14730
Cognitive Computing

- IBM TrueNorth
  - Spiking neuron technology
  - Low power consumption
- New programmatic paradigm
  - Evolving library of modules (corelet)
  - Compiling into minimalistic connectivity spiking neuron
- Application to pattern recognition in HEP
Data analytics beyond the computation needs
Literature Wizard

- IBM Watson Discovery Advisor
- Demonstrated capabilities to make sense of large volume of research papers and provide insights
Self Organizing Analysis

- Train a 4D self organizing map (SOM) on synthetic data composed of 3 backgrounds and one signal
  - Injection performed at varying signal/background ratio
- Interpretation using only backgrounds allow to single out the events from signal
  - Significance of deviation estimated as a function of signal injection

$10^{-4}$ Injected Signal

Significance: $\text{signal/}\sqrt{\text{background}}$
Recurrent Neural Nets

- Recurrent neural nets are adapted to text processing for which the input is not of fixed size.
- **Long Short Term Memory (LSTM)** cells revolutionized the field by considering long range correlations.
- Datasets in HEP has collections of variable size.
- There is **natural ordering of the data in HEP** → Challenge for computer science.
Deep Learning And Symmetries of Nature
Symmetry in Neural Nets

- Introduction of convolutional layers was a **ground-breaking** advancement
- Research on embedding more fundamental symmetries into neural nets
- Symmetries operate on the data or **internal representation** of data
- Next is to implement symmetries observed in HEP to build HEP-specific NN

T.S. Cohen, M. Welling ICML2016
Dark Knowledge

• Absence of interpretability is a sociological show-stopper
• Active field of research in data science
• Example of interpretability
  ➢ Decision tree thanks to box-cut by design
• Convolutional filters in CNN thanks to incorporation of symmetry of the data in the model
Summary

Deep learning has leaped forward over the last decade

Machine learning is a potential solution to several HL-LHC computational challenge

Deep learning can further enable scientific return, by tackling complexity

Engage Data Science Experts on HEP Challenges

http://cern.ch/DataScienceLHC2015
https://indico.cern.ch/event/514434
https://indico.hep.caltech.edu/indico/event/102
Backup Slides
the Compact Muon Solenoid

CMS is a **highly heterogeneous system**
Raw data is 100M channels sampled every 25 ns : 1Pb/s
50EB per day in readout and online processing.
Parametrized Learning

ArXiv:1601.07913 Baldi, Cranmer, Faucett, Sadowski, Whiteson

- Analysis in mass scans are usual optimized for discretized mass points.
- Training a classifier at each mass point
- Training a classifier with the mass as an input
- Such method shows very good extrapolation between point
- Excellent generalization capabilities
- Ideal for continuous hypothesis testing
Pileup Reduction

- Multiple proton-proton interaction per bunch crossing result in multiple overlaid events
- Non uniform energy deposition that needs to be subtracted
- Several methods are utilized to perform this subtraction
  - Apply de-noising with recurrent neural nets within a cone around particle
Predicting Data Popularity

- The large amount of LHC data and simulation is spread across many sites
- Monitoring services are aggregating several features on usage (number of CPU spend using it, number of access, ...)
- Goal is to facilitate access to popular data by replicating popular dataset
- Define data popularity by access counts above 10
- Weekly rolling mechanism of training classifiers and making popularity forecast
- Seasonal effects observed mostly from conferences

\[
\text{TPR} = \frac{TP}{TP+FN} \\
\text{TNR} = \frac{TN}{TN+FP}
\]
Re-weighting with BDT

- Great success of the standard model in predicting observations
- Generation of events in certain phase space are however hard and unconstrained
- Detector/electronic simulation is good but not perfect
- Data/Simulation agreement is de-facto very good but not perfect, and precise estimations require weighting of event to compensate for these imperfections
- Likelihood or binned reweighting is common practice
- Too many bins in high dimensionality leads to fluctuations

- Train BDT on original versus desired
- Use leaves to produce rescaling weights
- Less leaves than bins: less fluctuations
Decision Tree

- Regularly used, still room for development and tailoring of methods: LHCb experience

Train classifier with requirement on flat efficiency arXiv:1410.4140

Fast classification using sparse decision DAG arXiv:1206.6387
LHCb Topological Trigger

arXiv:1210.6861

- Full reconstruction of the event up-front
- Bonzai decision tree contains the tree growth with feature discretization
- Train on 2-, 3-, 4- body decay topologies
- Stable through detector mis-alignment
- Fast execution time used in the time-critical part of the trigger
Pixel Cluster Reconstruction

A neural network clustering algorithm for the ATLAS silicon pixel detector arXiv: 1406.7690

Example of cluster split

Average number of shared measurements in the innermost pixel layer, before (CCA) and after splitting (NN)
Outlier Detection With NADE

• Train a NADE (arXiv:1306.0186) model on mixture of the known backgrounds
• Use a synthetic dataset with small injected signal (10 events OUT OF)
• Log density singles out the injected signal