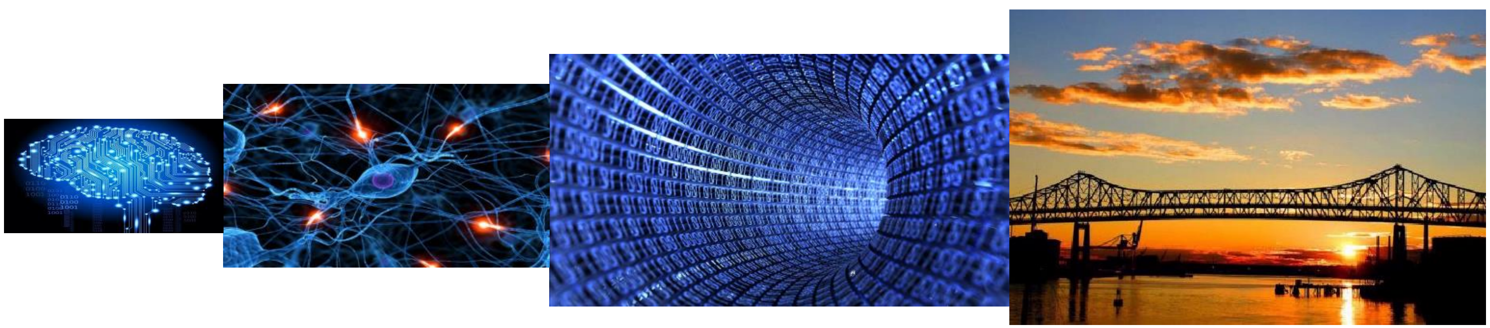
HEP-CS: Machine Learning

Sergei V. Gleyzer

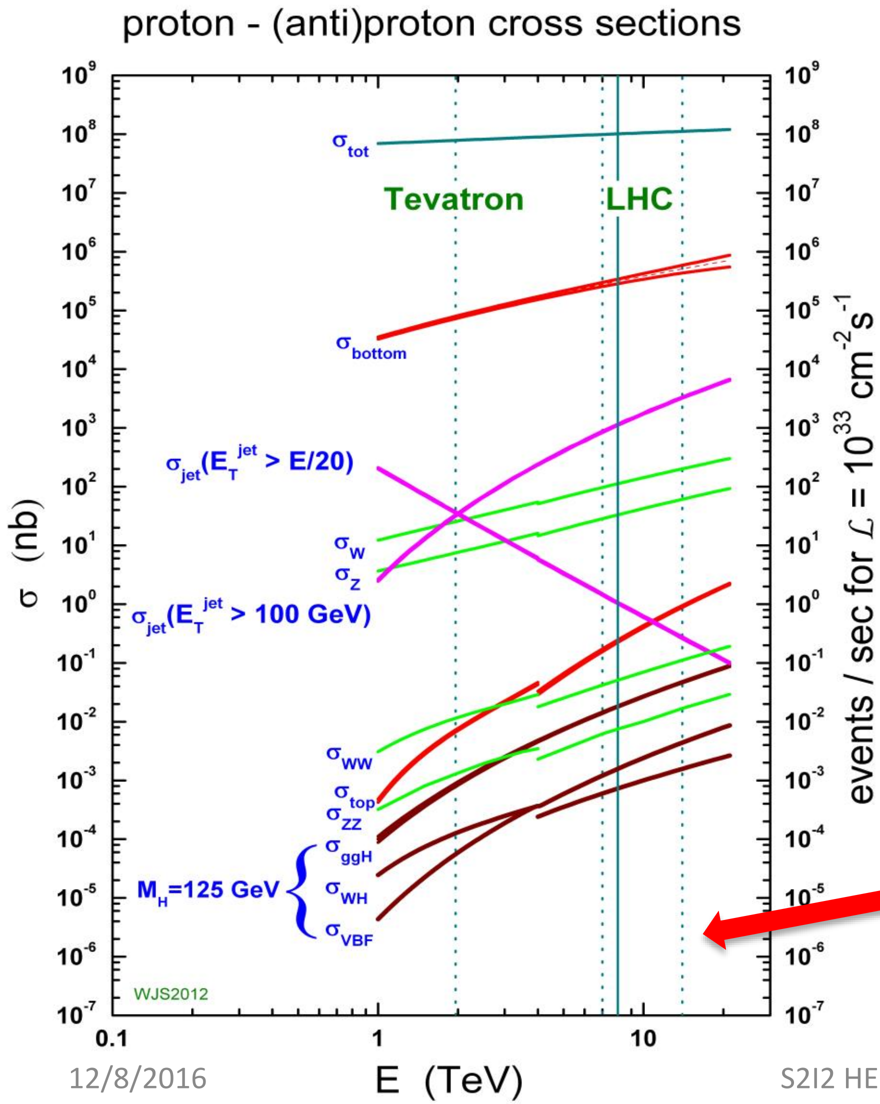
University of Florida



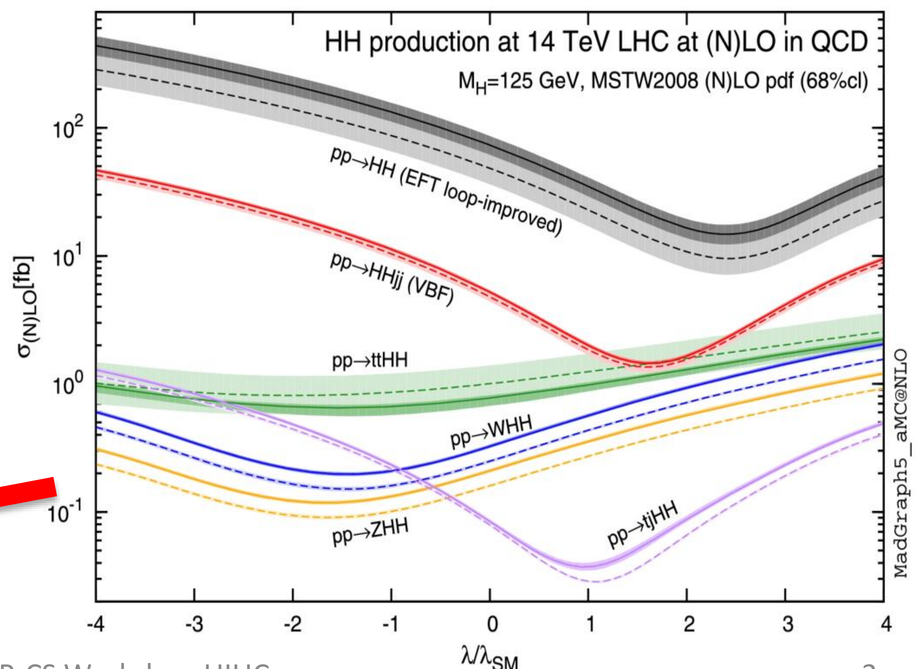
CS-HEP S2I2 Workshop

Dec. 8, 2016

Upcoming Challenges



Orders of magnitude between signals and backgrounds

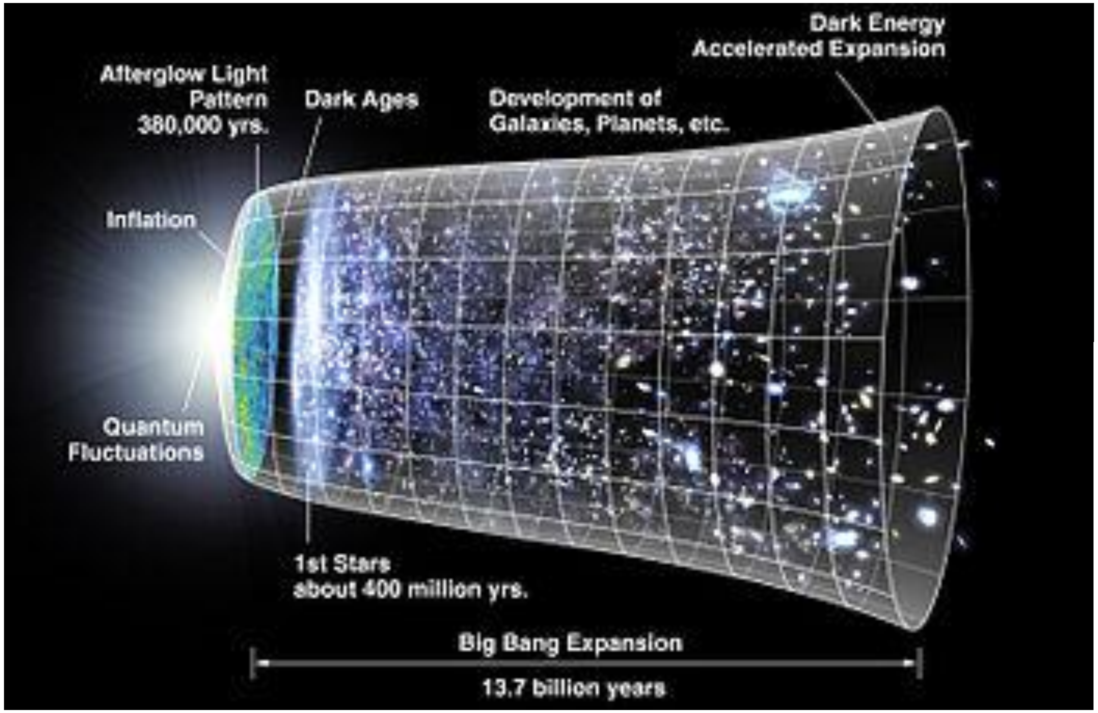


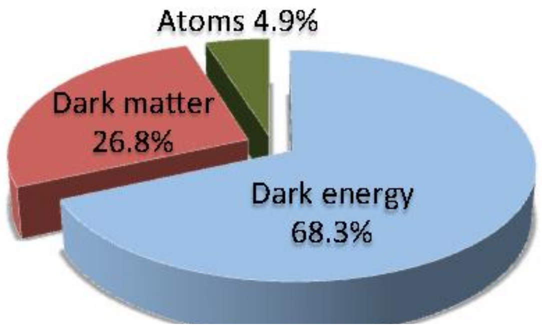
Upcoming Challenges

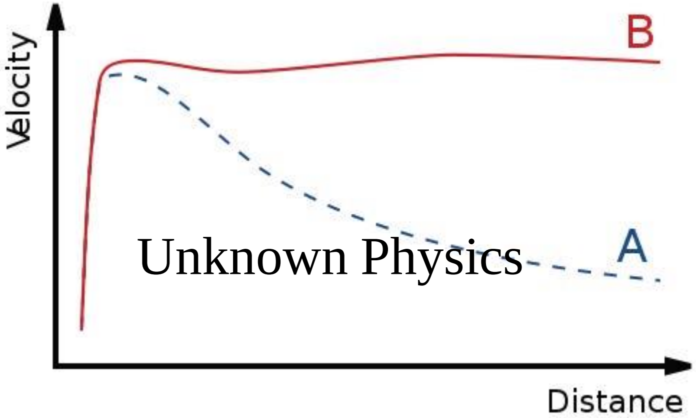
Today

Near future

Upcoming Challenges







Data size:

– LHC **15,000,000 Tb** 2010 - 2035

– Resources not up as fast as data volumes

Machine Learning in HEP

Machine Learning

General Approach:

- Given **training** data $T_D = \{y, \mathbf{x}\} = (y,x)_1$ $(y,x)_N$, **function space** $\{f\}$ and a **constraint** on these functions, teach a machine to learn the **mapping** $y = f(x)$

In Computer Science

Machine learning already preferred approach to

- Speech recognition, Natural language processing
- Computer vision, Robot control
- Medical outcomes analysis

Machine Learning field is growing fast

- Improved algorithms
- Increased data capture

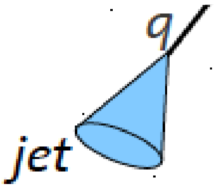
Machine Learning in HEP

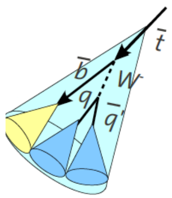
Machine Learning is already at the core of what we do today

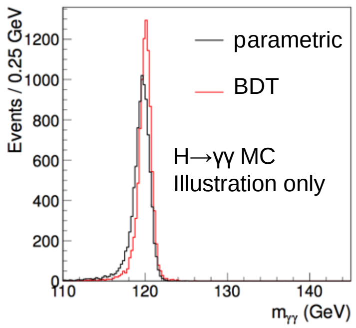
- **Automated** way to achieve better signal to background separation
- **Improved** detector performance and related measurements
  - Flavor tagging of jets
  - Particle Energies with Regression Methods
- Majority of HEP analyses already rely on some type of Machine Learning

HEP Applications

- **I. Classification**
  - **Particle Identification**
    - a photon or a jet?
  - **Advanced Pattern Recognition**
    - Hits clustering, jet substructure
  - **Searches for new Physics**
    - Event Classification: is this a Higgs or not?
  - **Data Quality Monitoring**
    - Outliers

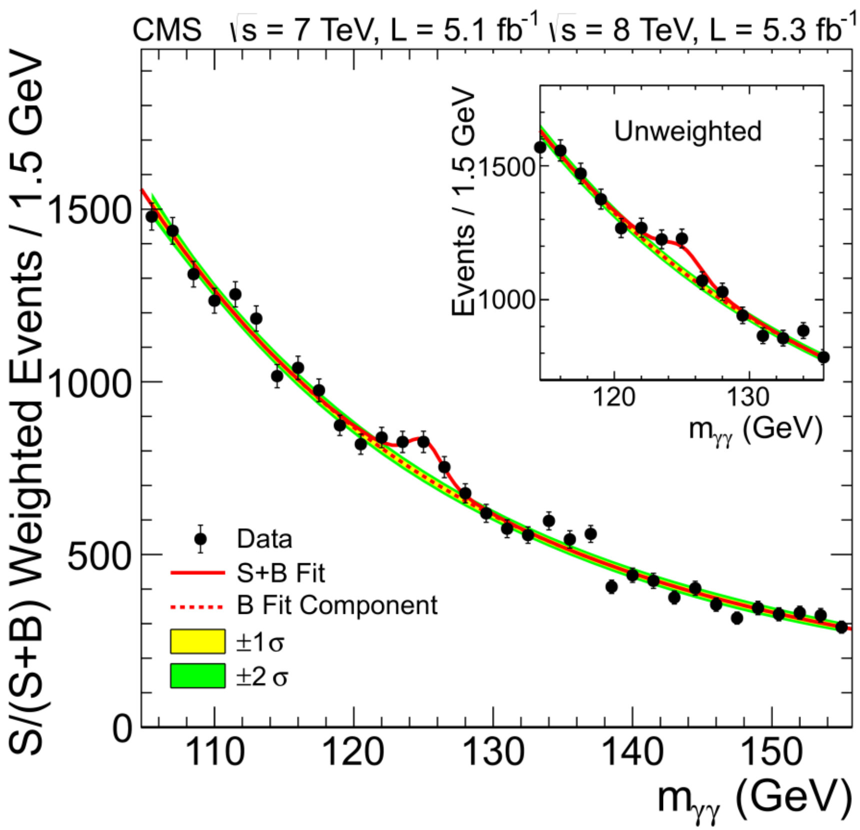






- **II. Function Estimation, Regression**
  - **Calorimetry**
    - Particle energy deposited in calorimeter better measured by function of individual energy deposits obtained with ML methods
  - **Energy/Momentum** regressions: photons, electrons, b-jets

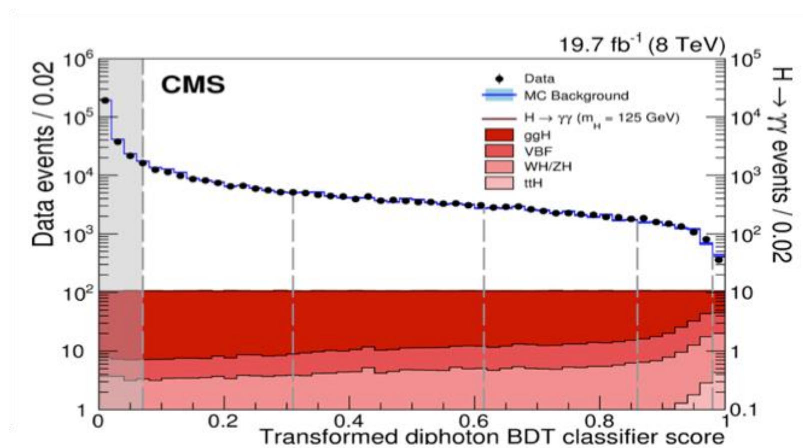
Higgs Discovery



Improvement in analysis from all four areas

Machine Learning used in Higgs Discovery

- Event selection
- Identification of particles
- Identification of interactions
- Energy regression



Machine Learning

- **Naturally collaborative and cross-cutting**
  - Statistics
  - Theoretical Computer Science
  - Mathematics
  - Physics
- Excellent place for CS-HEP collaboration

Towards Future

- **Require powerful ML algorithms**
- **Smart use of resources**
  - GPUs, clusters, spark, HPC
  - Efficiency of application
    - Latency, memory management
- **Further ML applications in data analysis and detectors**

HEP-CS Collaboration

Collaboration Areas

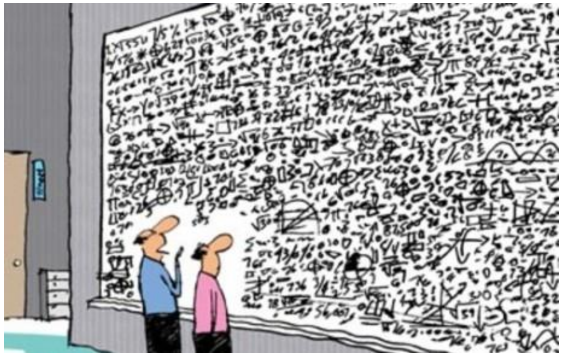
- **Software**
- **Algorithms**
- **Acceleration (hardware)**
- **Applications**

Software

- **ROOT framework has been around HEP for 20+ years**
  - I/O, histograms, statistics, data analysis
  - Core developers + eco-system
- **TMVA machine learning toolkit ~10 years old (integrated in ROOT ~3 years ago)**
  - Modernized over the past year
  - Easy to use, basic and advanced ML methods
  - Used by about 50% of HEP analysts
    - Others rely on external tools

Algorithms

- **Most natural area of collaboration**
  - Taking interesting directions from theoretical CS and applying to HEP problems
  - Looking at HEP problems and suggesting (1, 2…n) solutions
  - Work together on the R&D



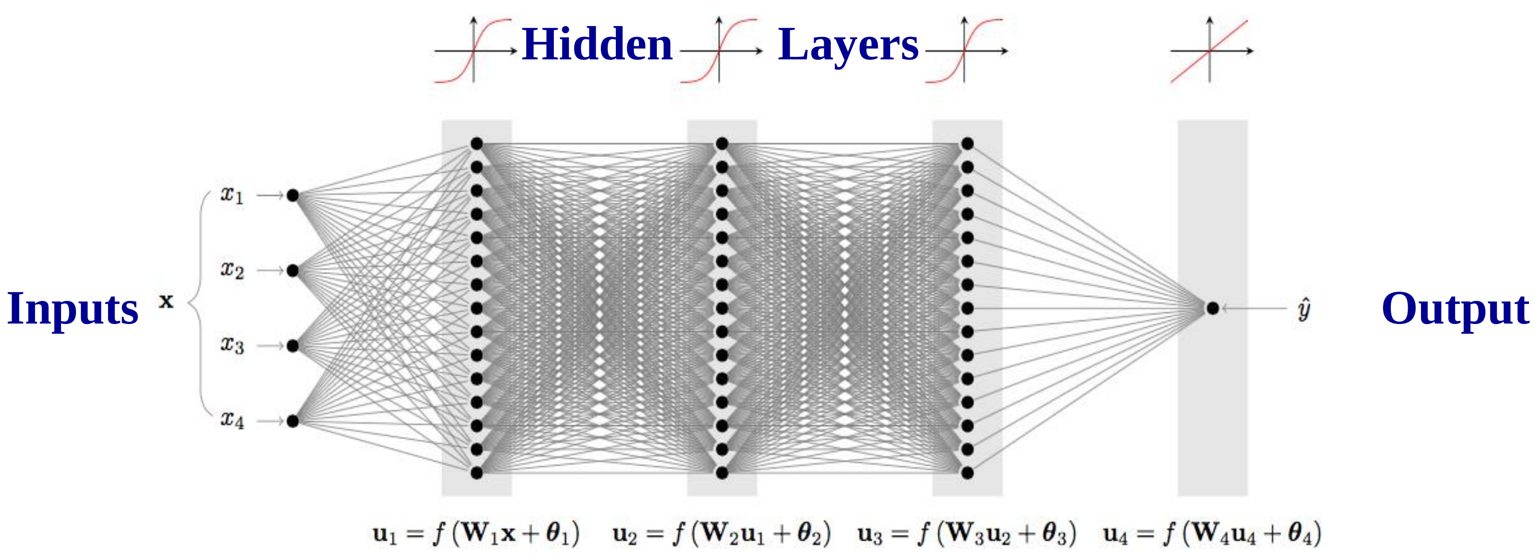
Algorithms in Use

ML algorithms in HEP:

- Fisher, Quadratic
- Naïve Bayes (Likelihood)
- Kernel Density Estimation
- Random grid search
- Boosted decision trees
- Rule ensembles
- Random forests
- Deep learning (neural networks)
  - feed-forward, recurrent, convolutional, LTSM, Bayesian
- Support vector machines
- Genetic algorithms

Deep Learning

Powerful Machine Learning method based on Deep Neural Networks (DNN) that achieves significant performance improvement in classification tasks



Deep Learning

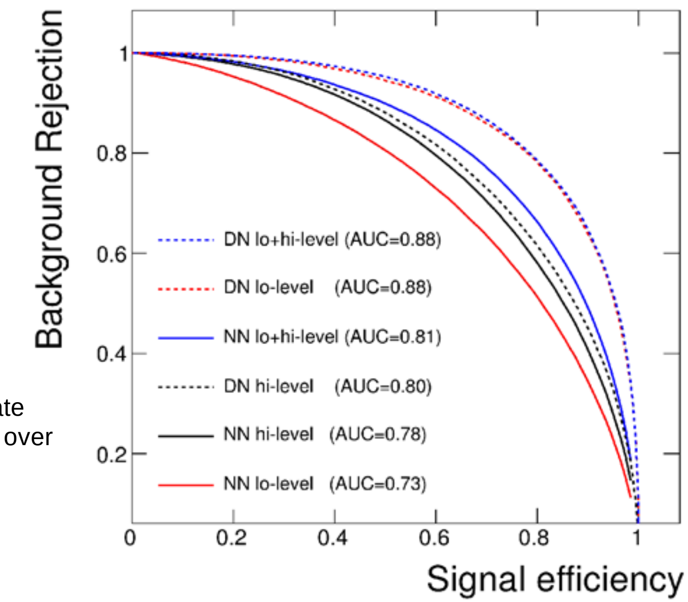
Higgs Boson Example:

P. Baldi, et. al. 2014

Tuning deep neural network architectures.

Hyper parameters	Choices
Depth	2,3,4,5,6 layers
Hidden units per layer	100,200,300,500
Learning rate	0.01, 0.05
Weight decay	0, 0.00001
Pre-training	none, autoencoder
	multi-task autoencoder
Input features	low-level, high-level
	complete set

Best:
- 5 hidden layers
- 300 neurons per layer
- Tanh hidden units, sigmoid output
- No pre-training
- Stochastic gradient descent
- Mini batches of 100
- Exponentially-decreasing learning rate
- Momentum increasing from .5 to .99 over 200 epochs
- Weight decay = 0.00001



8% improvement

In major HEP experiments mechanisms exist for making computer scientists authors on HEP papers

Deep Learning

Since 2014 ~10 deep learning HEP papers:

- Jet images and deep learning: link
- Jet substructure and deep learning: link
- Parton shower uncertainties and jet substructure: link
- Deep learning for ttHiggs link
- Nova link
- Daya Bay link
- Next: link
- Microboone: link

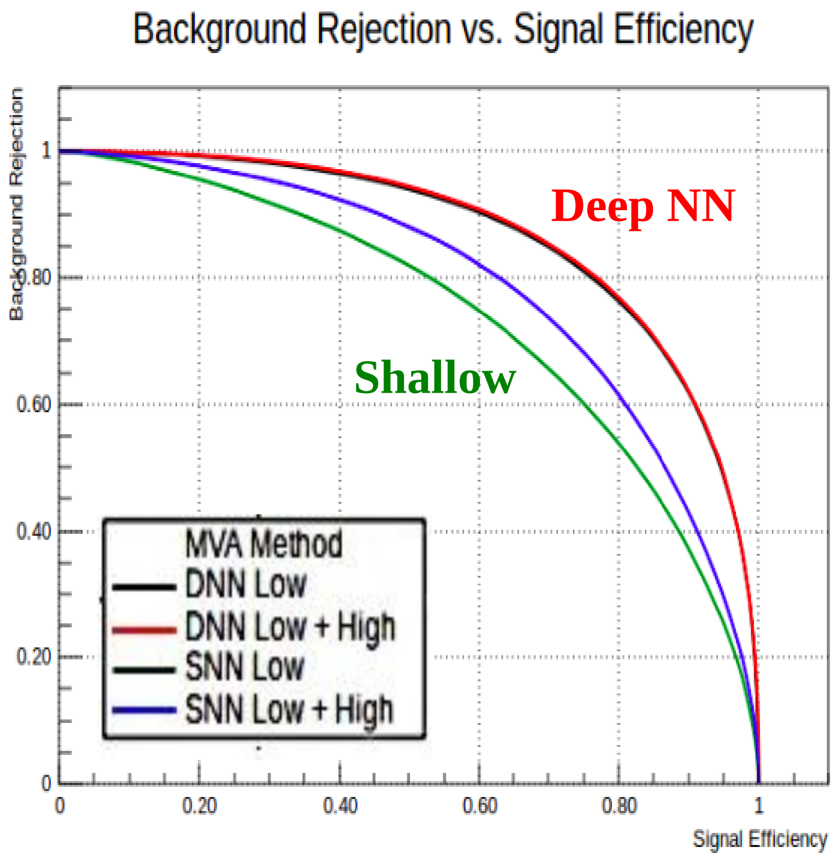
Google Summer of Code

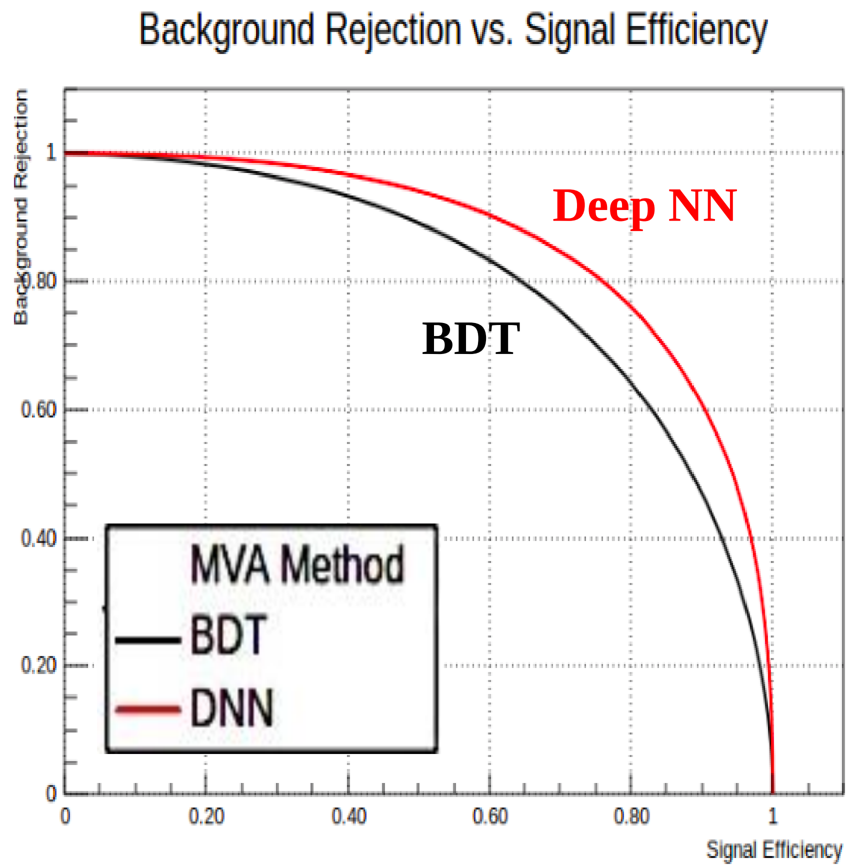
- **CERN Software for Experiments group participates since 2011**
  - This year 12 students (mostly CS ph.d. students)
  - Lots of useful cs and software engineering work
  - Excellent impact on our eco-system
  - Will expand to an umbrella association
    - Via HEP Software Foundation
    - More mentors, projects, students
    - Greater HEP involvement and impact
    - Train future developers (CS students interested in HEP)

Project Areas

- **Simulation**
  - **Geant4 and GeantV**
  - **Sixtrack (particle tracking)**
  - **Blond (beam dynamics)**
- **Data analysis tools**
  - **Interactive ROOT Graphics**
  - **Machine Learning**
    - **first time this year, significant student interest**
- **Other utilities and tools**

Deep Learning



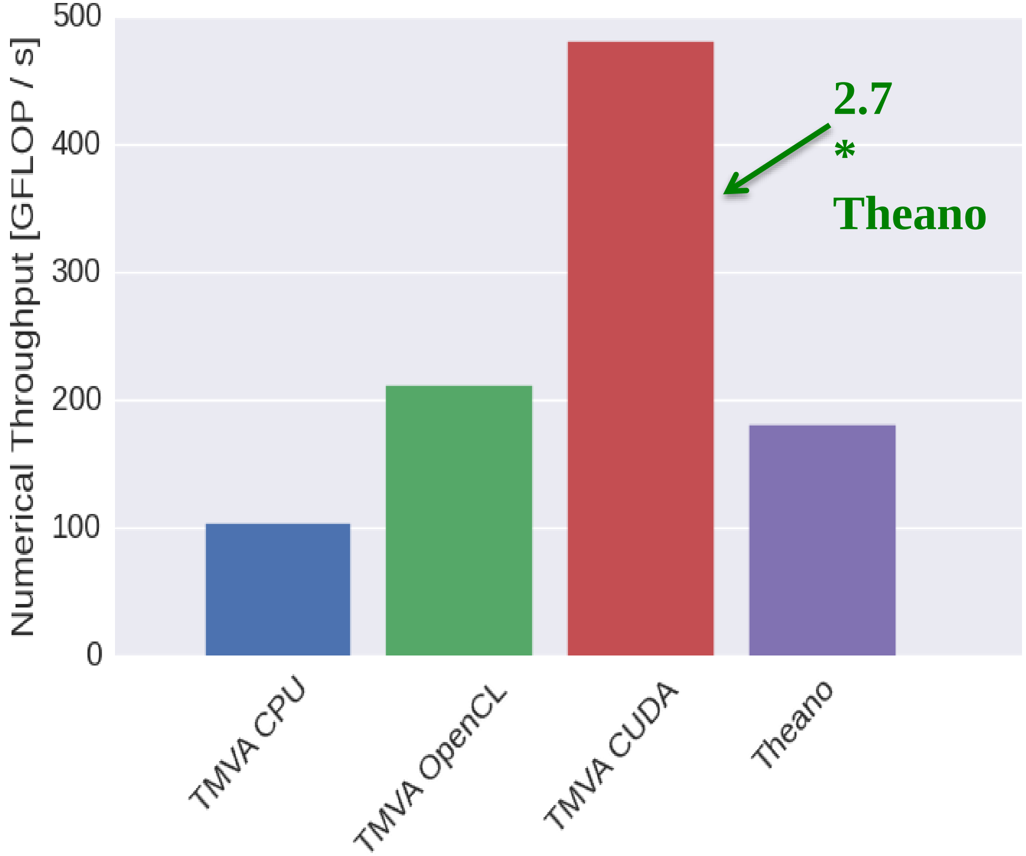


GSoC example

Significant improvements in performance

Deep Learning

Throughput Comparison



Single precision

Excellent throughput compared to Theano on same GPU

Beyond Classification

Function Estimation

Given enough data, estimate a function?

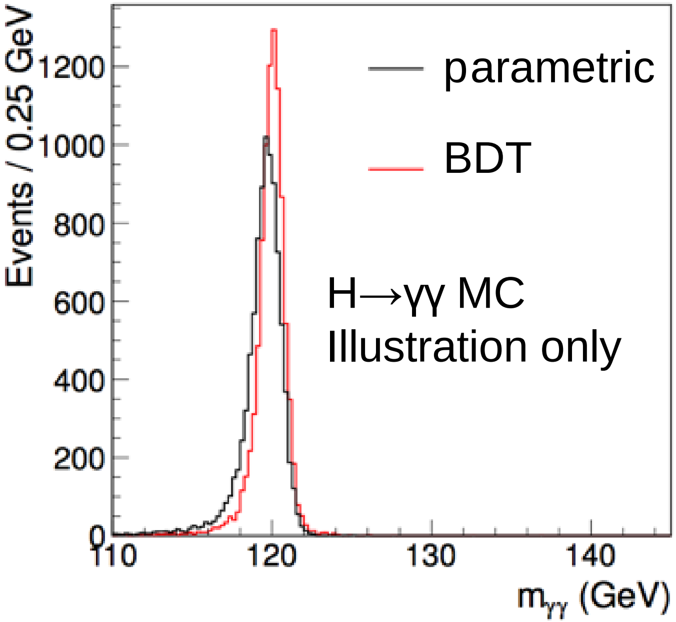
Problem posed by Gauss (1805): Estimate trajectory of comet from observations

Solution: Minimize difference between measurements and predictions by varying model parameters

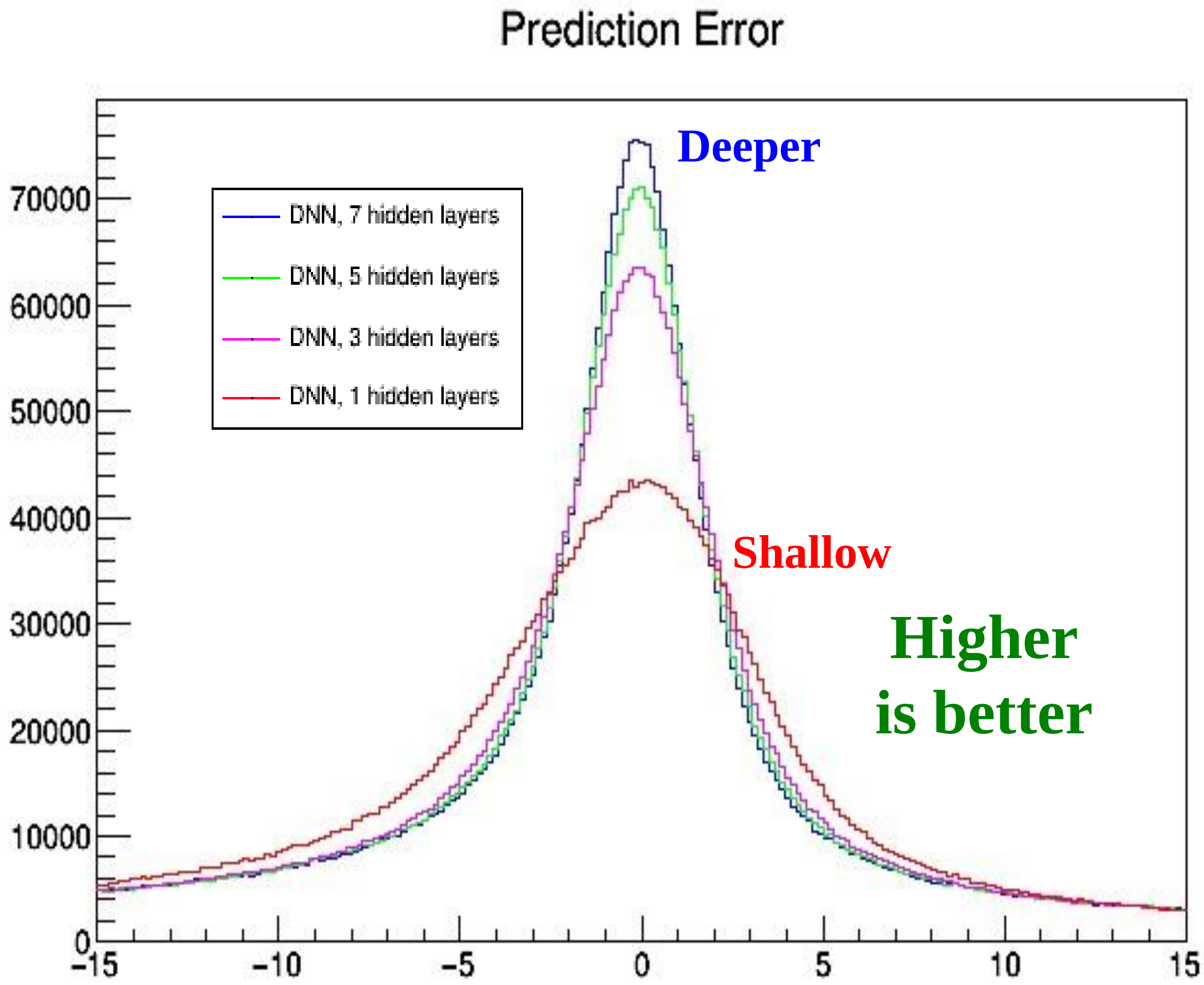
Regression in HEP

Machine learning regression

- Improve detector resolution (~10-30%)
- Example: estimate particle energy
- First implementations based on shallow methods
  - Neural networks, BDTs
- Applications:
  - Electrons and photons
  - B-jets, muons



Deep Learning Regression



Multi-objective Models

- Take into account **dependencies** between output attributes (their correlations)
- **improved performance results** compared to single-objective models, especially in ensembles
- usually smaller and easier to interpret
- applicable to detector simulations

Unsupervised Learning

Most of Machine-Learning in HEP has been focused on supervised learning

- **Labeled data, answers are known**

ML research shows better results when combining supervised and unsupervised learning

Building Bridges

Interesting Areas

- **Tracking**
  - Hundreds of particle trajectories
    - Algorithms smarter than "Kalman filters"
    - Low-level data
- **Calorimetry**
  - Posed as an image problem
- **Trigger**
  - Currently throwing away 99.9% of all events
    - Better use of this data, smarter decisions

Interesting Areas

- **Identification of interesting (different) physics**
  - Unsupervised learning
- **Faster detector simulation**
  - We spend a lot of computing power doing this
  - Replace with ML-based systems
- **Better vertexing**
  - Still using methods from 20-30 years ago

TMVA Interfaces

Interfaces to External ML Tools

- **RMVA interface to R**
- **PyMVA interface to scikit-learn**
- **PyKeras interface to Keras**
  - **High-level interface to Theano, TensorFlow deep-learning libraries**

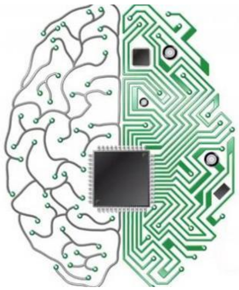
Interactive ML

Added support for interactive ML with Jupyter integration

- **Interactive training**
- **Model tuning**
- **Visualizations**

Acceleration Hardware

- **Optimized hardware for machine learning training and application**
  - **Acceleration**
    - GPUs
    - Combined FPGA-CPU systems
  - **Neuromorphic computing**
  - **Other…**



Computing Resources

Better ML training strategy and regularization

- **significant progress in overcoming overtraining**
- **more data → better outcomes**
- **availability and optimal use of resources for training become key**
  - **Use of GPUs, clusters, spark, HPC etc.**
  - **Flexible programming model**

CWP

HEP Software Foundation

- Community White Paper
  - link to CWP
  - Machine Learning
    - Identification of challenges
    - Roadmap to address them
  - Important to think of these issues now
    - Impact on how we design our software

Inter-experimental LHC Machine Learning Working Group [iml.cern.ch]

- **Exchange between HEP and ML communities**
- **Sharing of ML expertise and experience among LHC experiments**
- **ML Forum and Education (Tutorials)**
- **ML software development and maintenance**
  - **Connection to other efforts: AMVA4NewPhysics, Diana-HEP, DS@LHC, HSF**

Summary

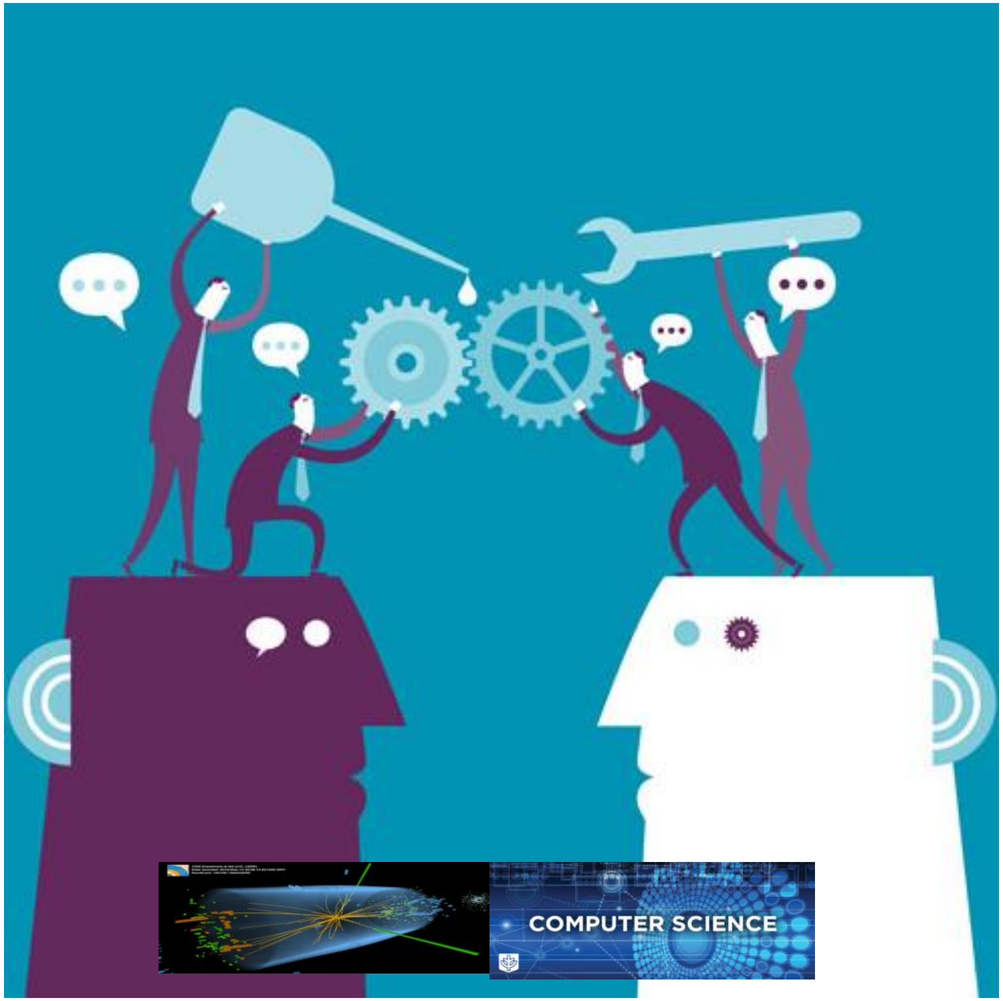
HL-LHC physics and computing challenges will require significant progress:

- **Higher backgrounds and pileup, data volume, unknown new physics**
  - Machine learning offers a promising direction
  - An opportunity to examine new areas of ML applications to HEP

Summary

- **Classification has been the primary focus for ML in HEP**
  - **Significant progress with Deep Learning**
    - 10-20% improvement in classification
    - Progress beyond fully interconnected architectures
- **Other areas becoming increasingly important**
  - **Machine-Learning Regression**
    - 10-25% improvements in detector resolution
    - Good promise with Deep Learning
  - **Unsupervised learning**

HEP-CS Collaboration Model




Thank You