Machine Learning

Sergei Gleyzer

PH-SFT Group Meeting

Nov. 2, 2015

Outline

- What is Machine Learning
- Machine Learning in HEP

Machine Learning

What is Machine Learning?

- Study of algorithms that
  improve their performance **P**
  for a given task **T**
  with more experience **E**

Sample tasks: identifying faces, Higgs bosons

In Computer Science

Machine learning already preferred approach to

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

ML is growing

- Improved algorithms
- Increased data capture
- Software too complex to write by hand

A Little History

1950s: First methods invented

1960-80s: Slow growth, focus on knowledge

1990s: growth of computing power, new learning methods, data-centric focus

2000s-10s: wide use of machine learning in all spheres of research and industry

2010s: improvement of learning, parallelism, deep learning

Data, data…

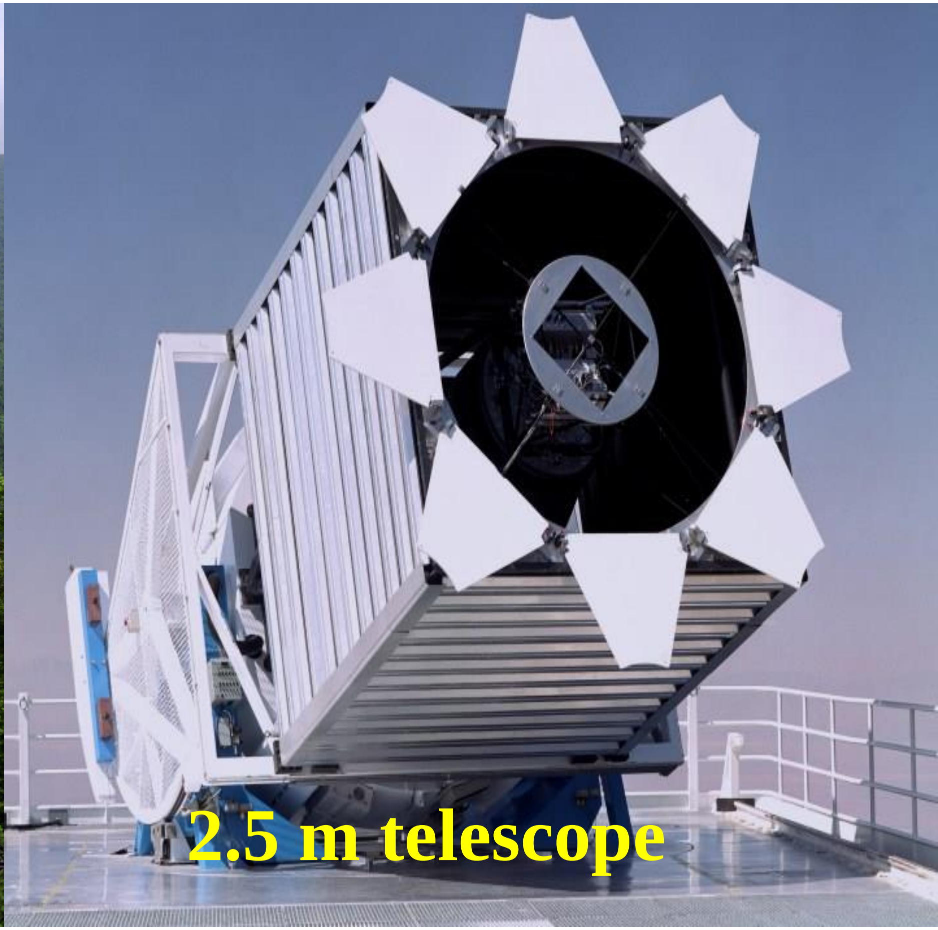
"As of 2012, about 2.5 Exabytes (EB) of data are created each day, and that number is doubling every 40 months"

Harvard Business Review, Oct. 2012

1 Exabyte = 1 000 000 000 Gb

Big Science → Big data? Yes

Sloan Digital Sky Survey

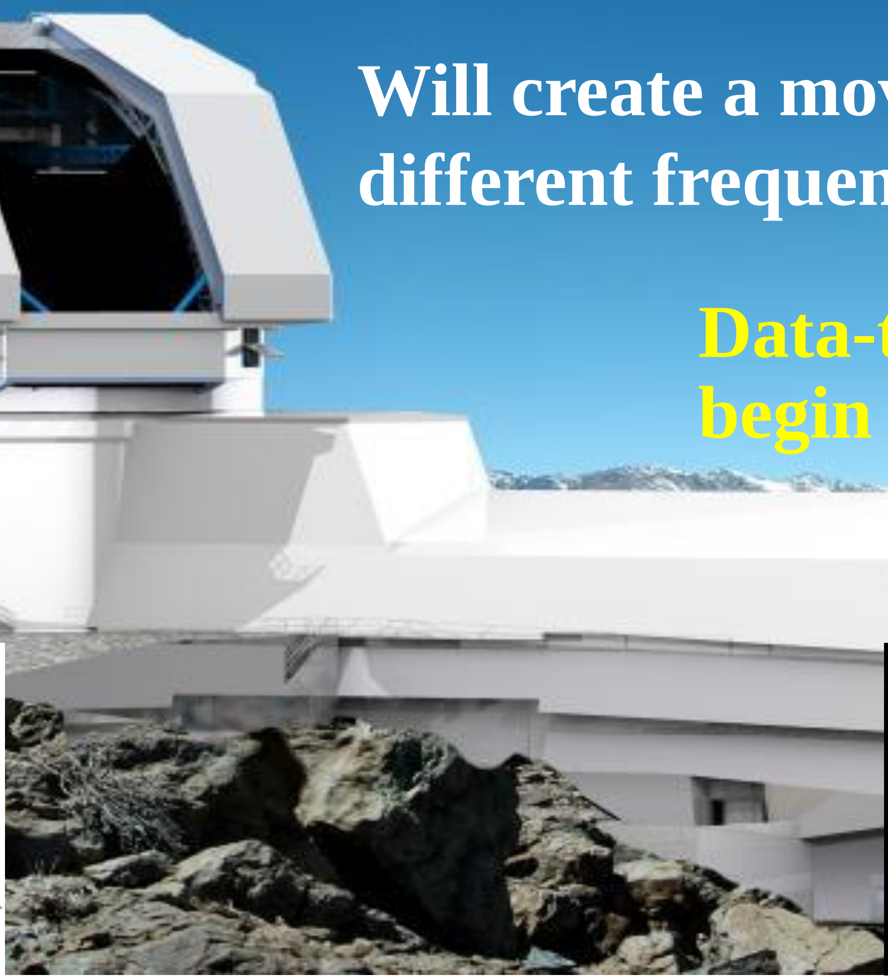


Collected more data in the first two weeks

than was collected in the history of astronomy

Large Synoptic Survey Telescope

3200 Megapixel camera



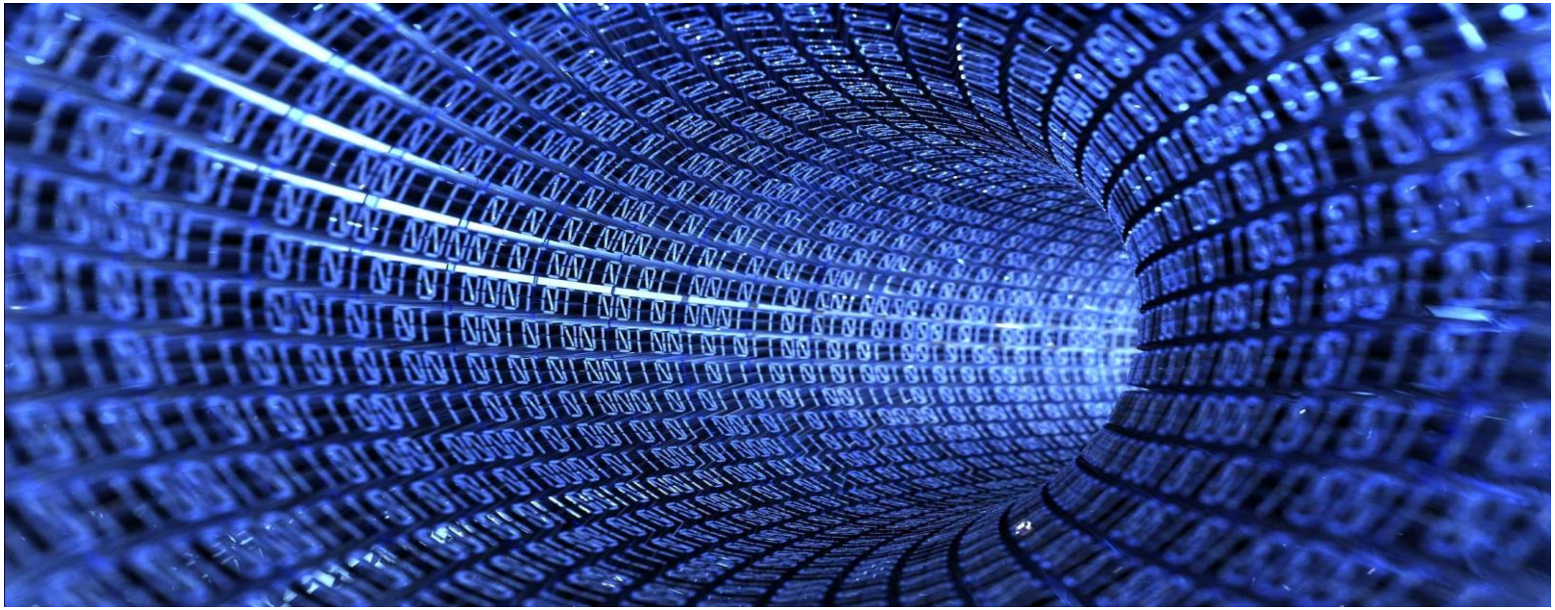
Will create a movie of the sky in different frequencies for ~10 years

Data-taking expected to begin in 2022

Large Synoptic Survey Telescope

Big Data

Project	Expected Data	Period
SDSS	100 Tb	2000 - 2015
LSST	100 000 Tb	2022 - 2032
LHC	15 000 000 Tb	2010 - 2035



ML in HEP

- **Classification**
  - **Particle ID:** b-tagging, primary vertex tagging, is this a photon or a jet?
  - **New Physics Searches**
    - Used by many physics analyses to search for new physics
      - Is this a possible Higgs or SUSY event or background?
- **Function Estimation, Regression:**
  - **Calorimetry**
    - Particle energy deposited in non-compensating calorimeter better measured by function of individual energy deposits, cluster shapes obtained with ML methods

Classification

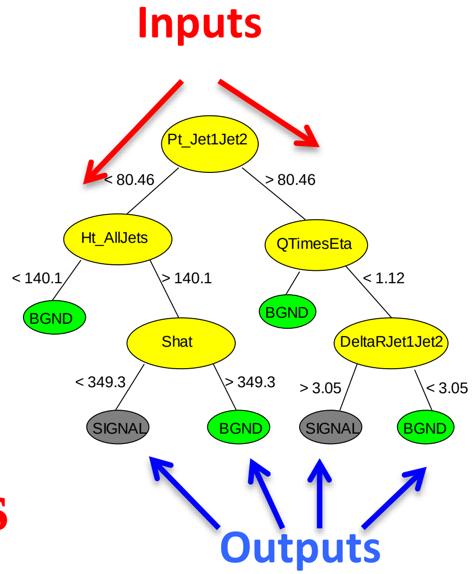
Distinguish f(x), g(x) using Training set of observations

{**inputs** , **outputs**}

Pass **observations** to a learning algorithm

neural network, decision tree

that produces **outputs** in response to **inputs**

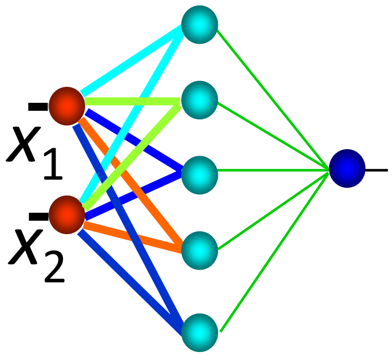


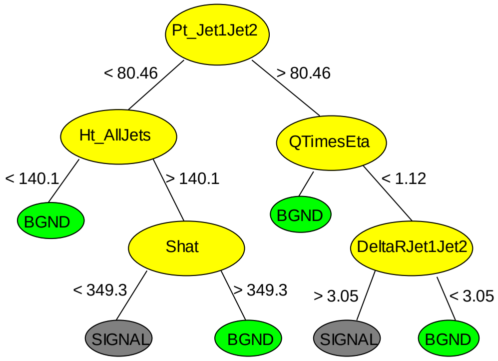
Use another set of observations to evaluate

Popular Methods

Incomplete list of learning algorithms:

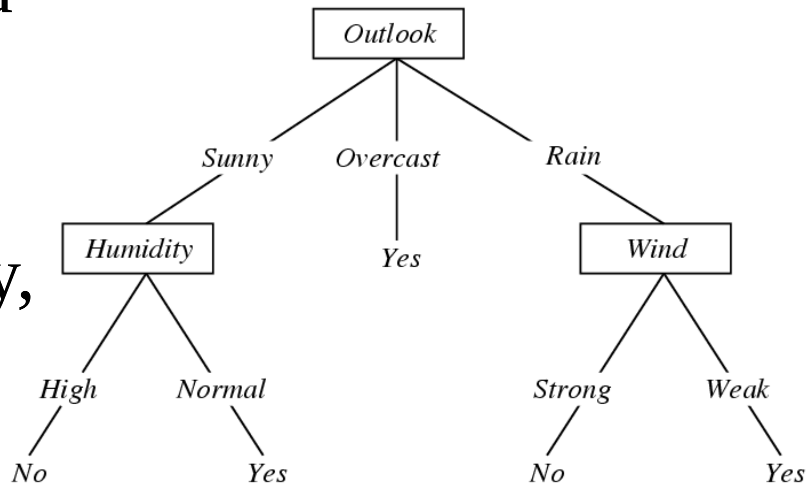
- Fisher Discriminant
- Quadratic Discriminant
- Support Vector Machines
- Decision Trees
- Neural Networks
- Bayesian Neural Networks
- Genetic Algorithms
- Random Forest





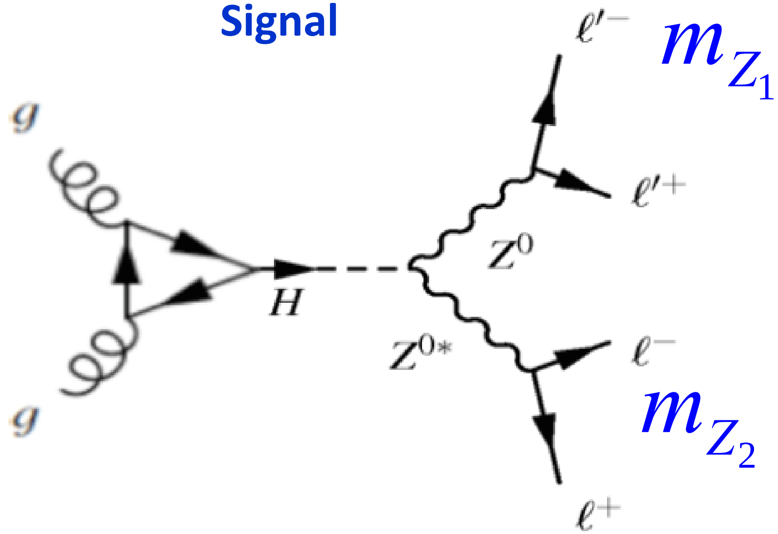
Decision Trees

- Think of **decision trees** as **multidimensional histograms**
  - Bins are recursively constructed
  - Each associated to the value (or class) of f(x) to be approximated
  - Tennis-Playing Decision Tree: f(outlook, humidity, wind, T)



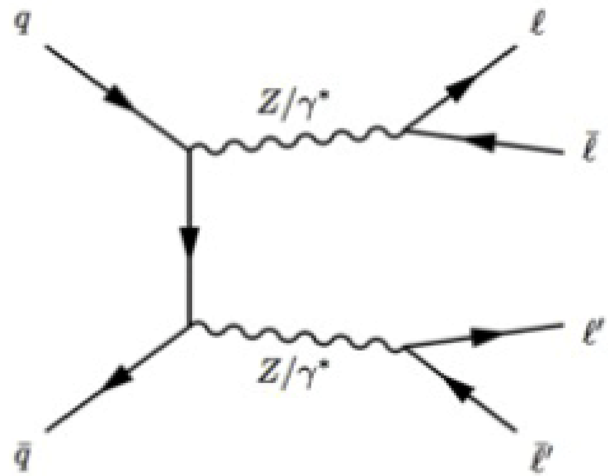
$H \rightarrow ZZ \rightarrow$ 4 leptons

Signal



$$pp \rightarrow H \rightarrow ZZ \rightarrow \ell^+ \ell^- \ell'^+ \ell'^-$$

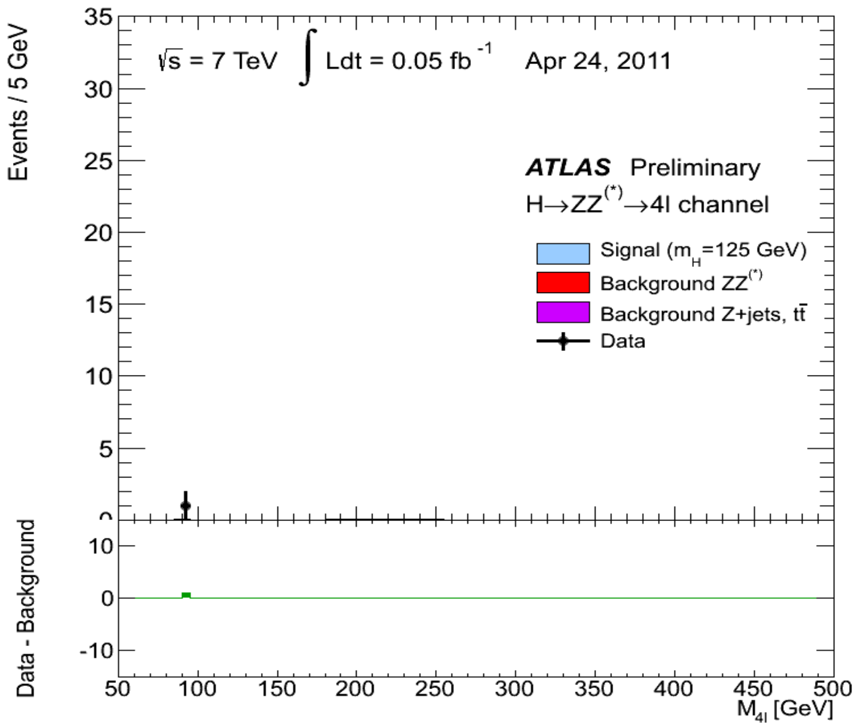
Background



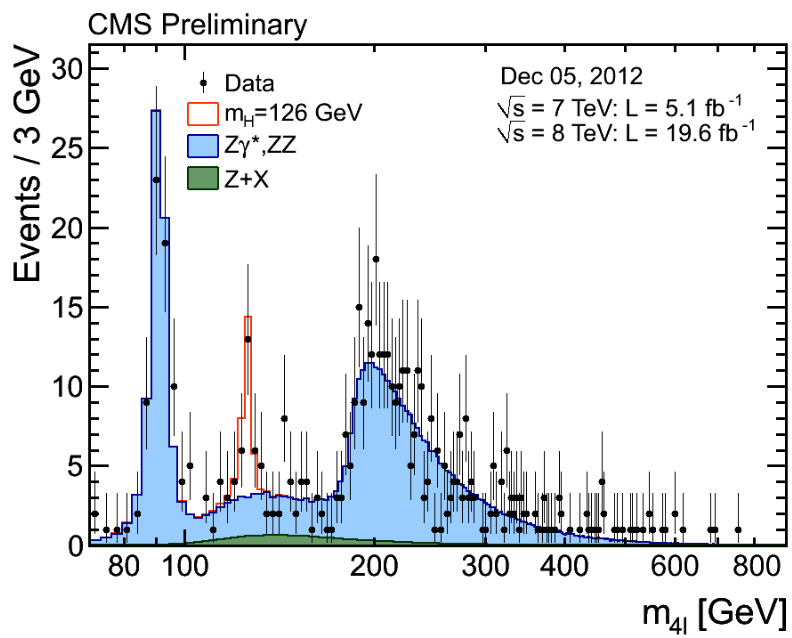
$$pp \rightarrow ZZ \rightarrow \ell^+ \ell^- \ell'^+ \ell'^-$$

Illustrative example

H → ZZ → 4 leptons

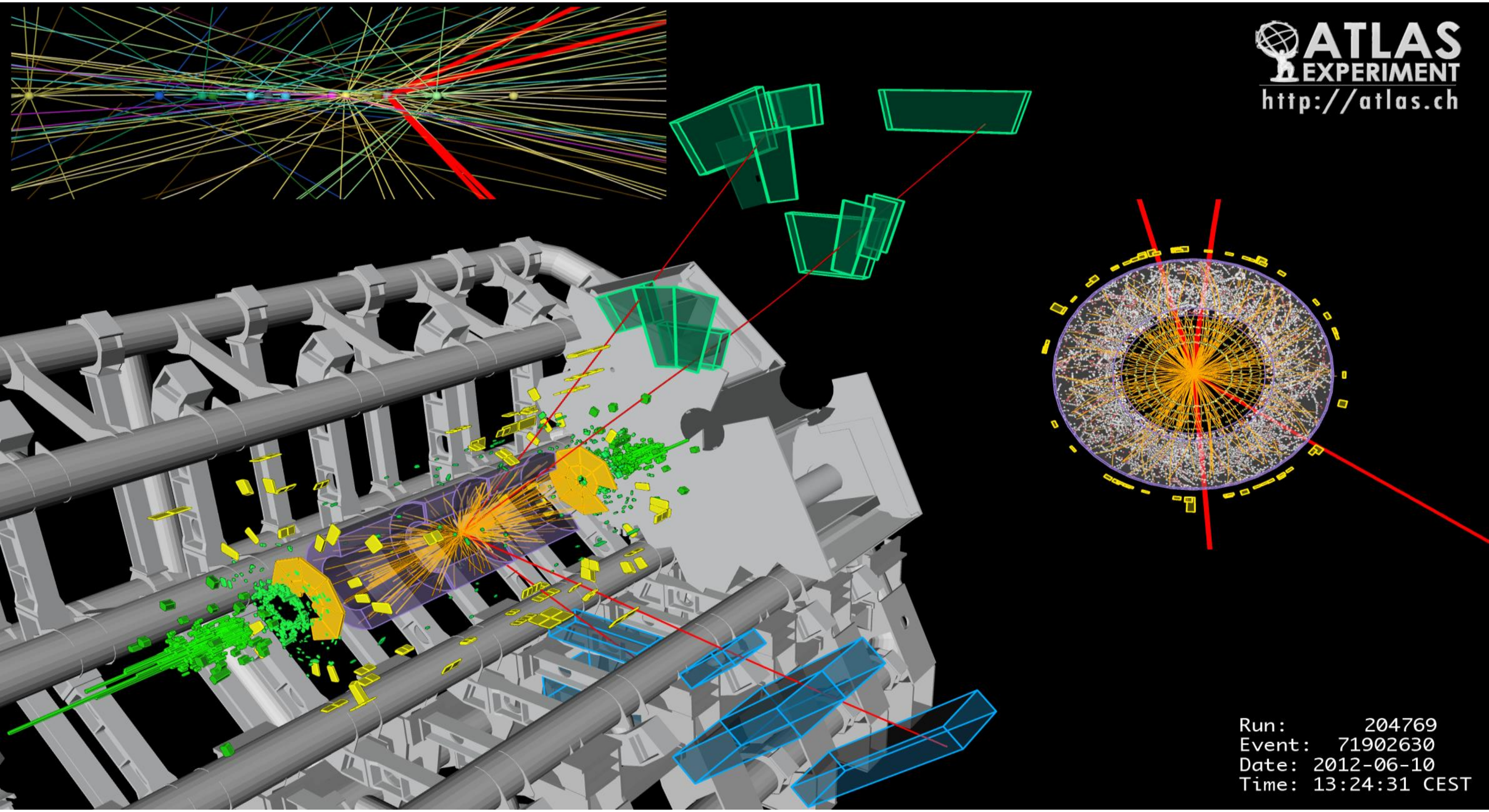


ATLAS

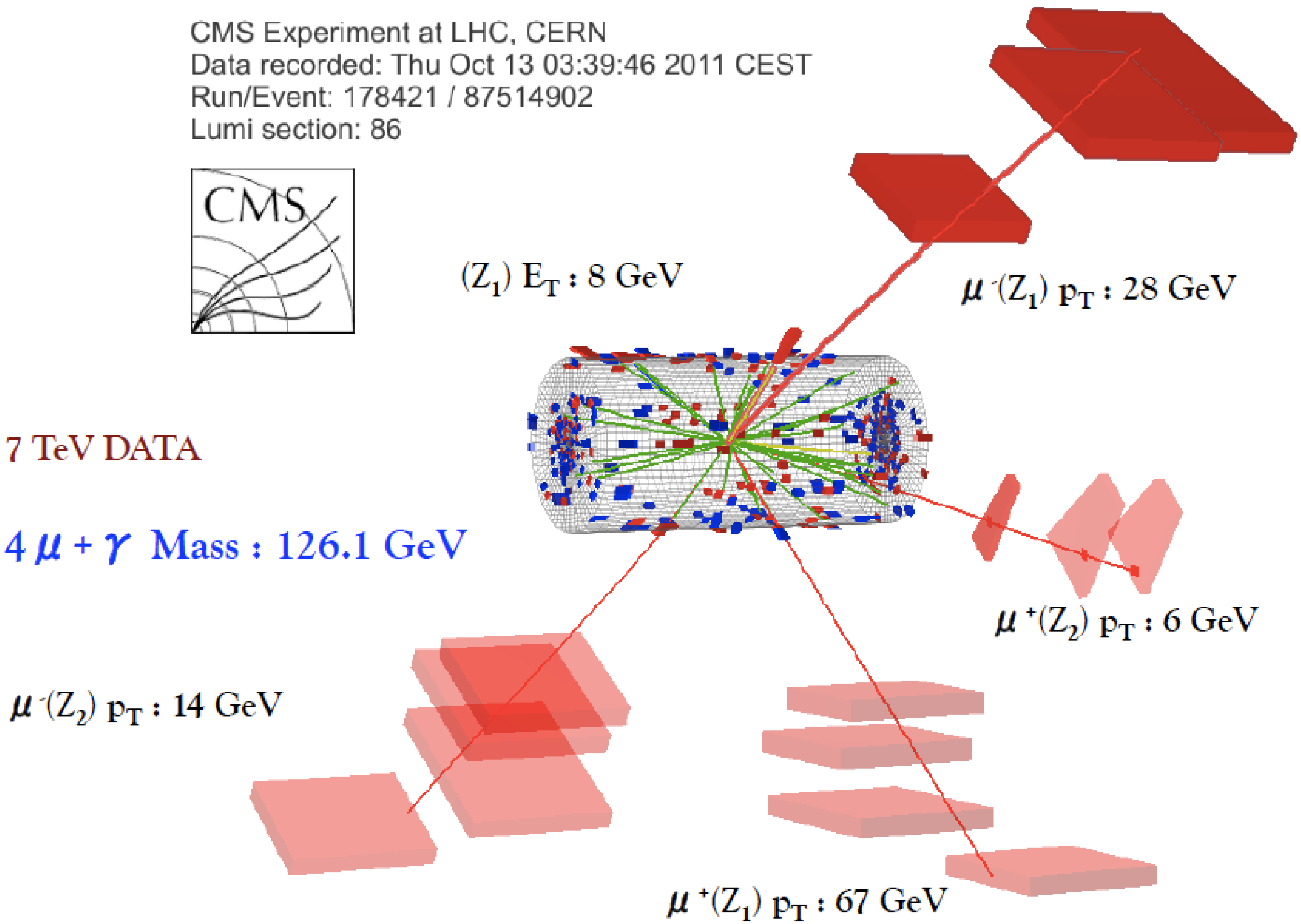


CMS

4-lepton Event ATLAS

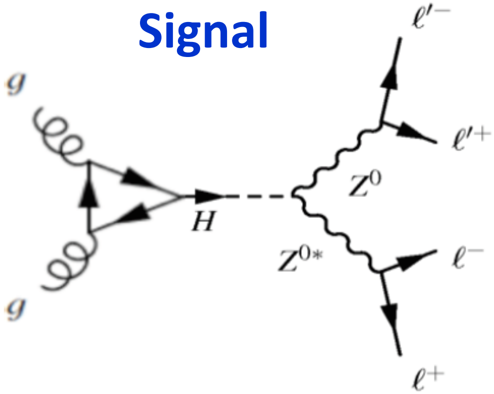


4-lepton Event CMS



H → ZZ → 4 leptons

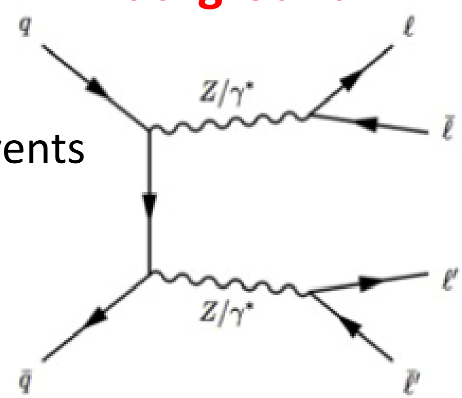
Signal



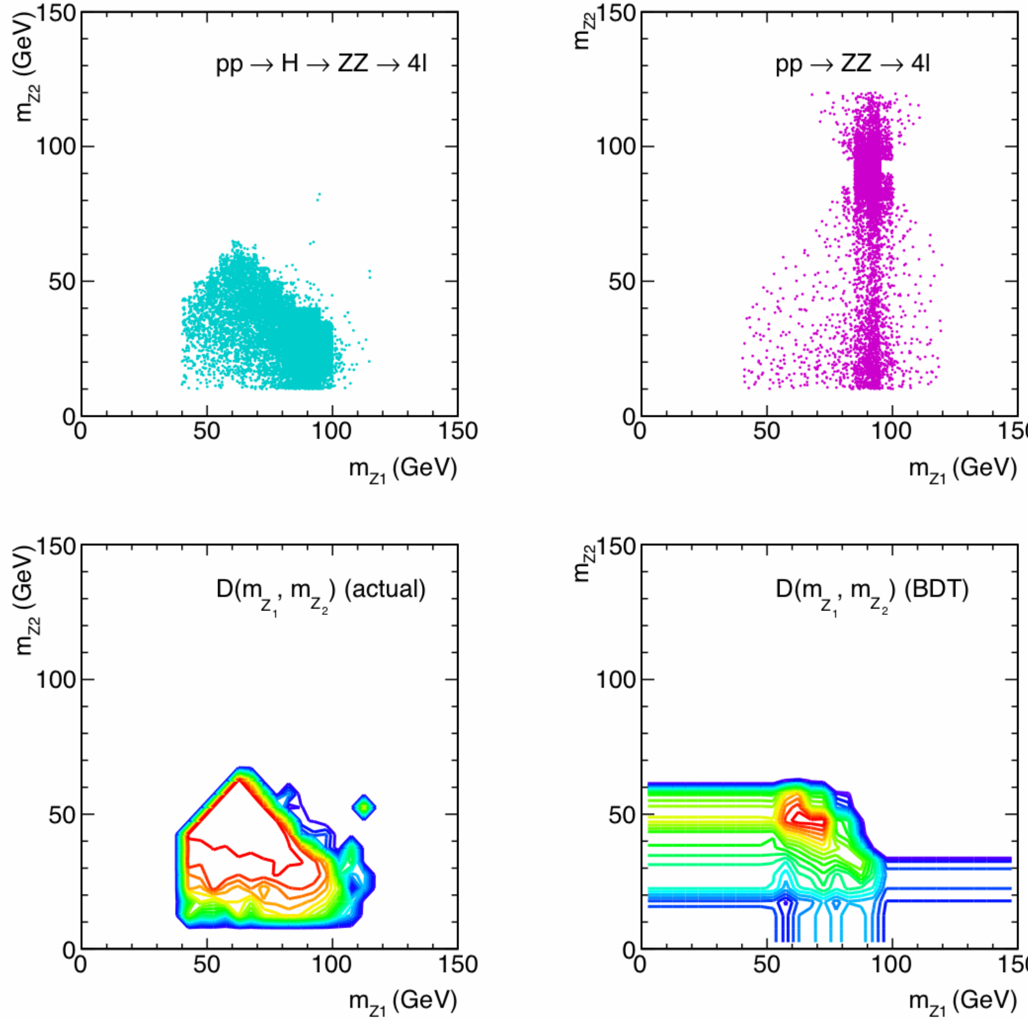
$$pp \rightarrow H \rightarrow ZZ \rightarrow \ell^{+}\ell^{-}\ell'^{+}\ell'^{-}$$

Background

200 trees
min 100 events
per leaf

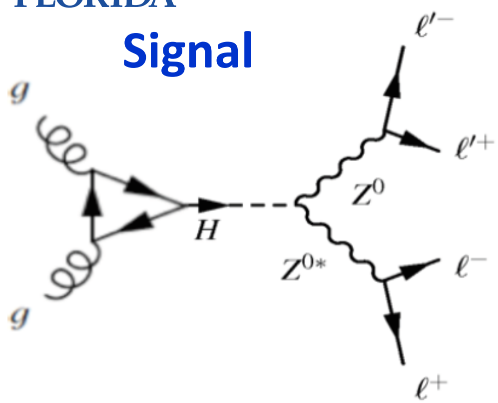


$$pp \rightarrow ZZ \rightarrow \ell^{+}\ell^{-}\ell'^{+}\ell'^{-}$$



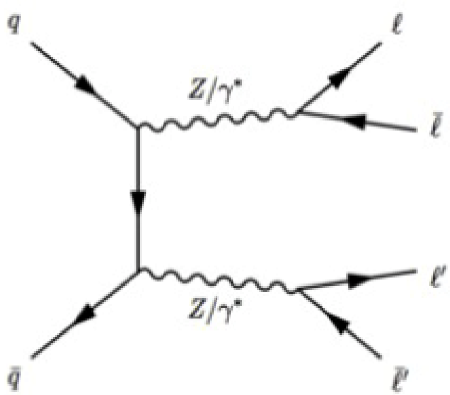
H → ZZ → 4 leptons

Signal

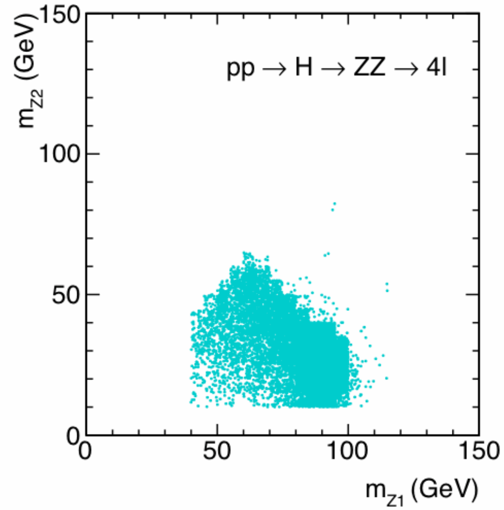


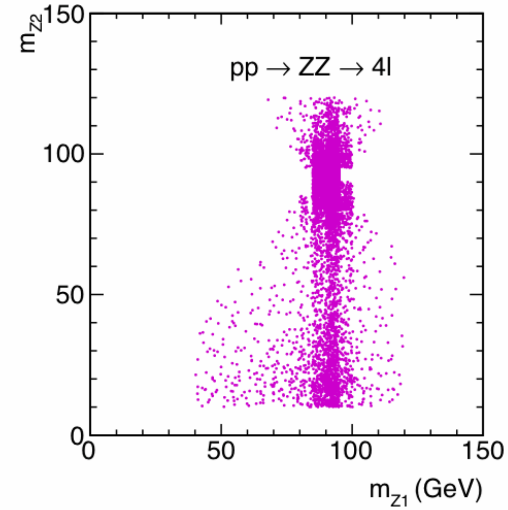
$$pp \rightarrow H \rightarrow ZZ \rightarrow \ell^+ \ell^- \ell'^+ \ell'^-$$

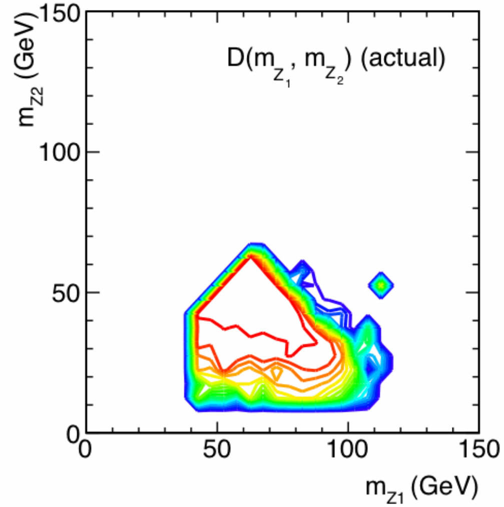
Background

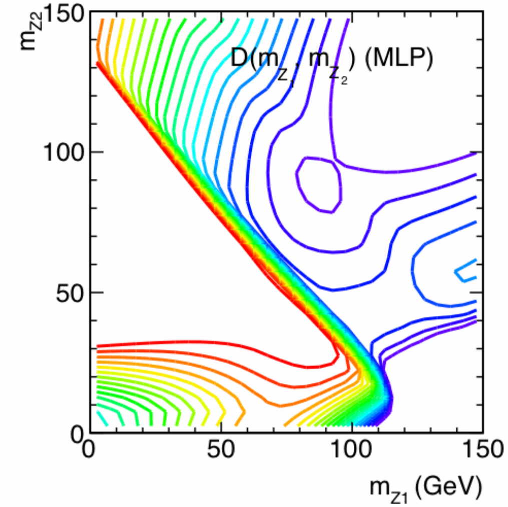


$$pp \rightarrow ZZ \rightarrow \ell^+ \ell^- \ell'^+ \ell'^-$$

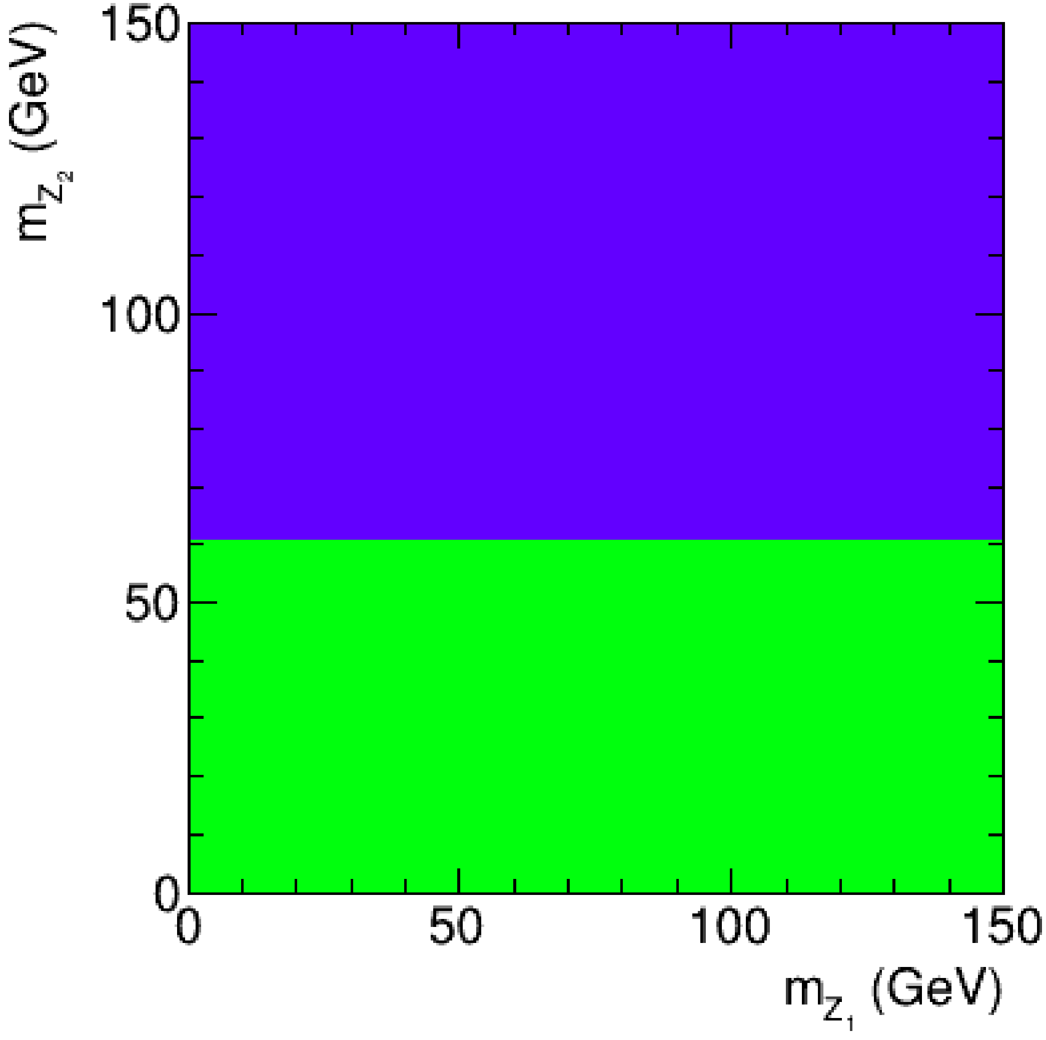




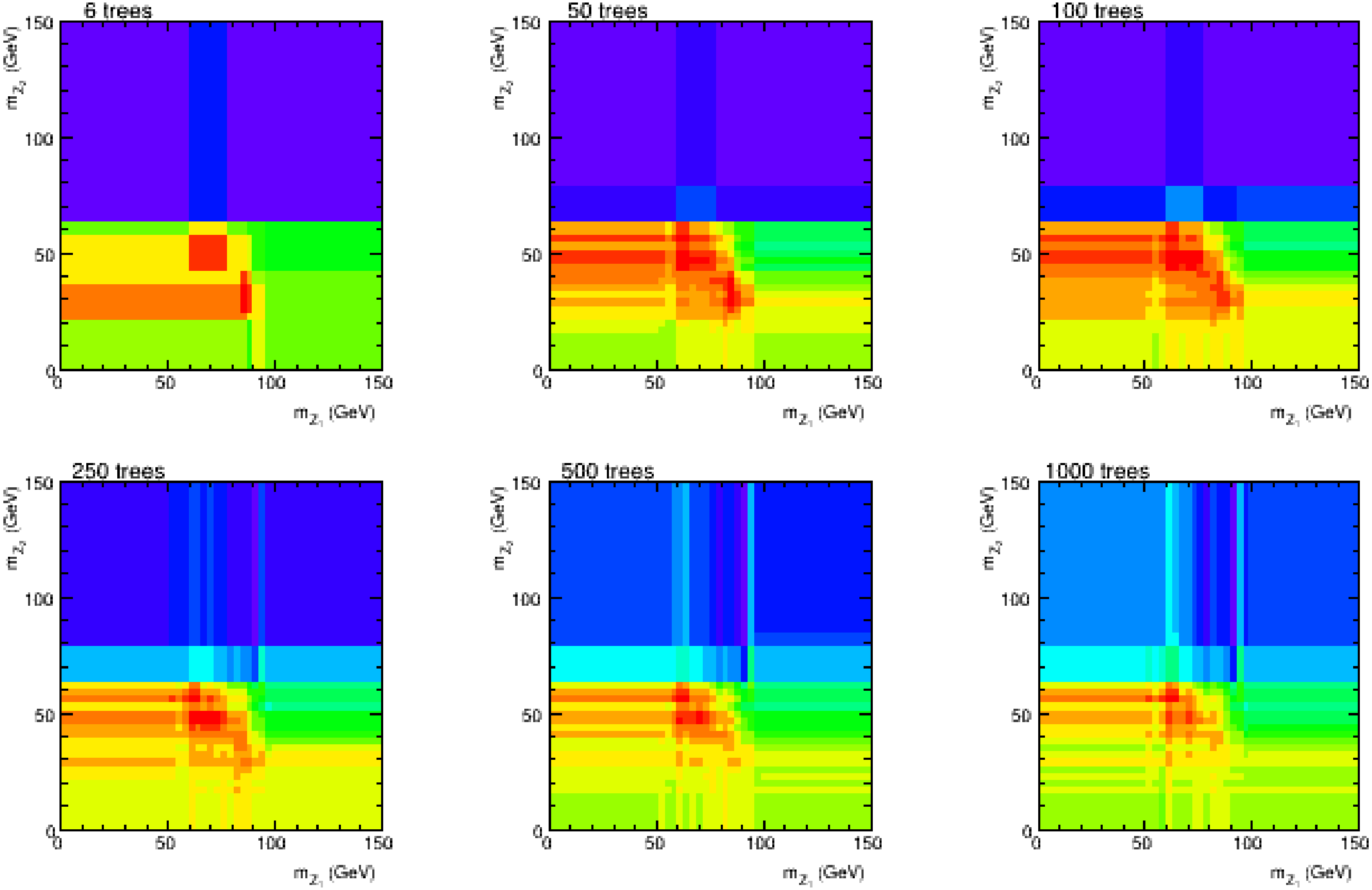




First 100 Decision Trees



Averaging over a Forest



Function Estimation

- From classification to regression:
  - change the evaluation criteria used in the learning algorithm
    - from **maximum separation gain** to **minimal variance**
  - **Inputs:** Training examples {<$x^{(i)}$, $y^{(i)}$>} of unknown function f. $x^{(i)}$,$y^{(i)}$: electromagnetic shower information, other calorimetric variables
  - **Target Output:** hypothesis h that best approximates target function f (calorimeter energy)

Modern ML

- **Large ensembles** of classifiers
- **Deep** vs. shallow **learning**
  - Neural nets with many more hidden layers
- **Bayesian** approaches
- **Combination** of semi/un-supervised learning with supervised learning

Summary

- **Machine Learning** use is exploding everywhere including **HEP**
- See next talks for:
  - **Bringing ML innovation to HEP** with Challenges
  - **Latest ML features in ROOT+TMVA**
  - **LHC ML Working Group (IML)** and **Data Science at the LHC Workshop**

Resources

Literature

G. James, et al. "Introduction to Statistical Learning" Springer 2013

C.M. Bishop "Pattern Recognition and Machine Learning" Springer 2006

J. R. Quinlan "C4.5: Programs for Machine Learning" Morgan Kaufmann 1992

Talks and Tutorials

DESY Statistics School 2014 talk + tutorials

https://indico.desy.de/getFile.py/access?contribId=15&resId=1&materialId=slides&confId=9288

TMVA @ Root Users Workshop 2013

http://indico.cern.ch/event/217511/contribution/37/material/slides/0.pdf

DS'15 Workshop http://cern.ch/DataScienceLHC2015