HEP(e)xML

Introduction to HEP-ex

Johannes Wagner

Berkeley
UNIVERSITY OF CALIFORNIA

HEPxML overview

- Workshop series consisting of 5 separate workshops
  - Each will be focused on a different ML algorithm and one of its applications to particle physics
  - Focus is on deep learning and its applications rather than physics explicitly
  - Many examples will be from collider physics, but there's more out there!
- Knowledge of basic python and coding concepts **REQUIRED**
  - No previous knowledge of particle physics or machine learning is necessary
- These are intended to be practical and hands-on
  - Do the exercises and ask questions to get the most out of them!

Our team

- Luc Le Pottier, 2nd year PhD under Prof. Wang
- Radha Mastandrea, 2nd year PhD under Dr. Nachman
- Ryan Roberts, 4th year PhD under Prof. Wang
- Johannes Wagner, 3rd year PhD under Prof. Gray
- Robin Xiong, 4th year PhD under Prof. Gray

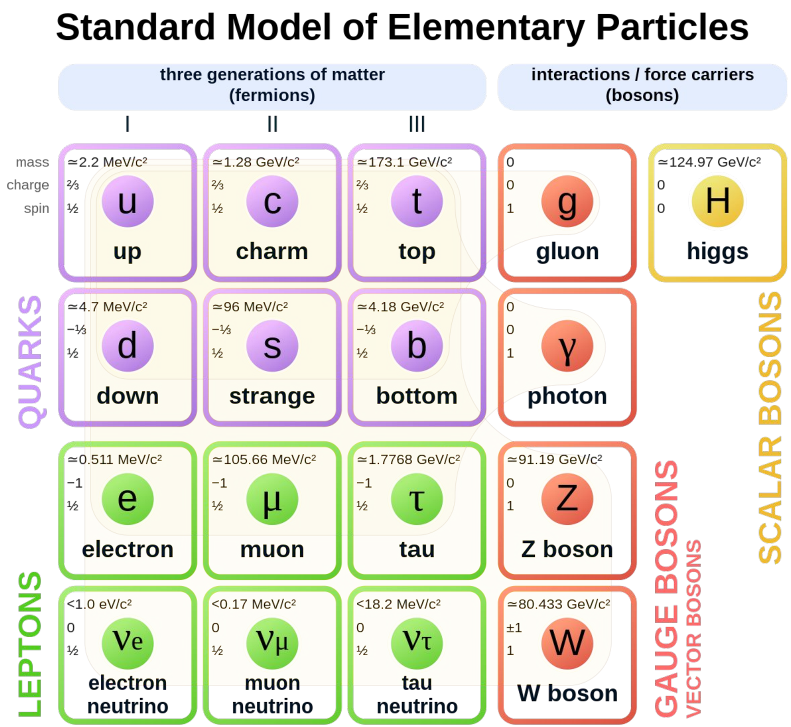
If it disagrees with experiment it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is – if it disagrees with experiment it is wrong. That is all there is to it.

Richard Feynman

What is HEP-ex?

- An acronym for **EX**perimental **H**igh **E**nergy **P**hysics (i.e. particle physics)
- A way for us to test predictions made by the Standard Model of Particle Physics (SM) and other theories
    - SM = our current state-of-the art theory describing weak, strong and EM particle interactions (everything except gravity)

Source: [Wikipedia]



What are we looking for?

- Particle discovery: detection of new particles
  - Examination of new theories (e.g. particulate dark matter, supersymmetry, etc.)
  - Usually requires looking for interaction signatures specific to whatever particle we're looking for (indirect detection)
- Precision measurements of physical quantities
  - Testing of current theories by measuring relevant parameters
  - If measurement deviates from theory prediction, this could mean new physics
  - Quantities of interest are mostly related to interaction strengths and intrinsic particle characteristics

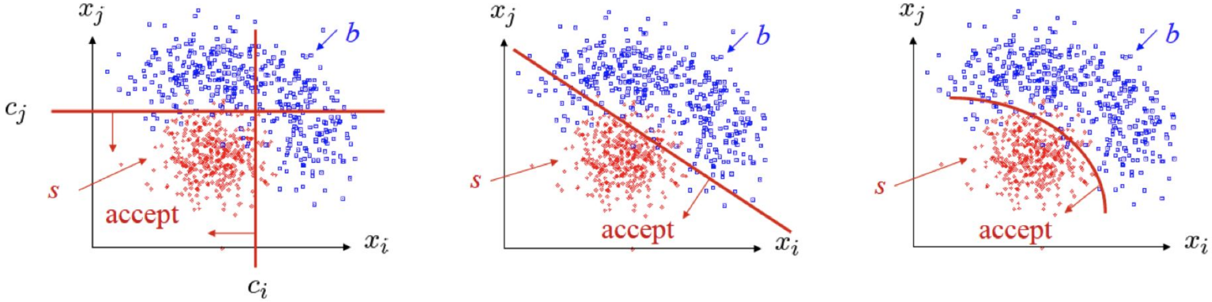
Signal vs. background

- It’s impossible to build a detector that ONLY measures the kind of events you’re interested in
- The main challenge of particle physics is distinguishing events we’re interested in (**signal**) from those we’re not interested in (**background**)
  - What this looks like depends entirely on the relevant physics and experimental setup
- Some background signatures will be easy to see, some will be hard to tell apart from signal events

Berkeley
UNIVERSITY OF CALIFORNIA

How do we deal with backgrounds?

- We need to have a good understanding of our backgrounds!
    - We simulate events based on standard model predictions to achieve this
- Impossible to say whether individual events are signal or background, but we can make statements about statistical aggregates

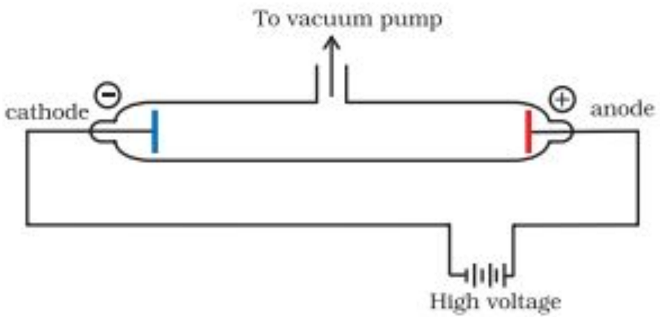


Source: Proceedings of Science

The simplicity of the past

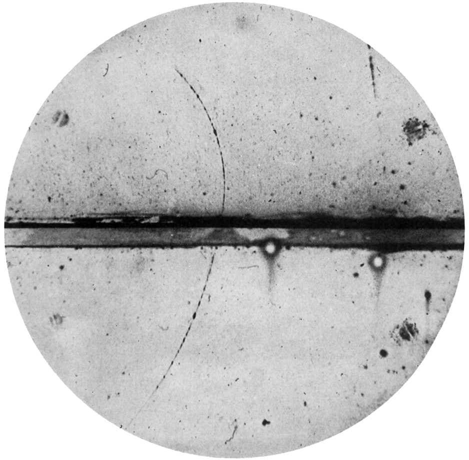
- Small scale experiments to analyze stable building blocks of matter
- Could be set up by a handful of people and data analyzed by hand

Source: [The Fact Factor]



Cathode ray (Electron discovery)

Source: [Wikipedia]

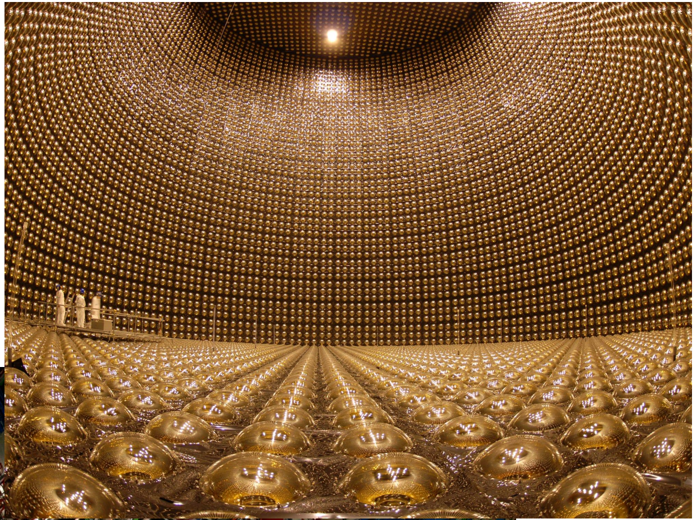


Cloud chamber (Positron discovery)

Berkeley
UNIVERSITY OF CALIFORNIA

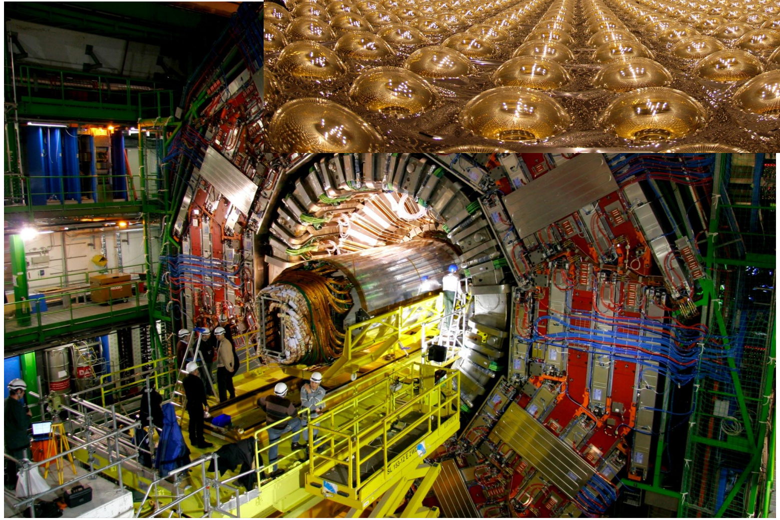
The reality of the present

- Search for unstable particles and exotic processes
- Requires complicated detectors and enormous amounts of data
    - Doing anything in small groups or by hand completely impossible



Super Kamiokande

Source: Business Insider



CMS

Source: CMS

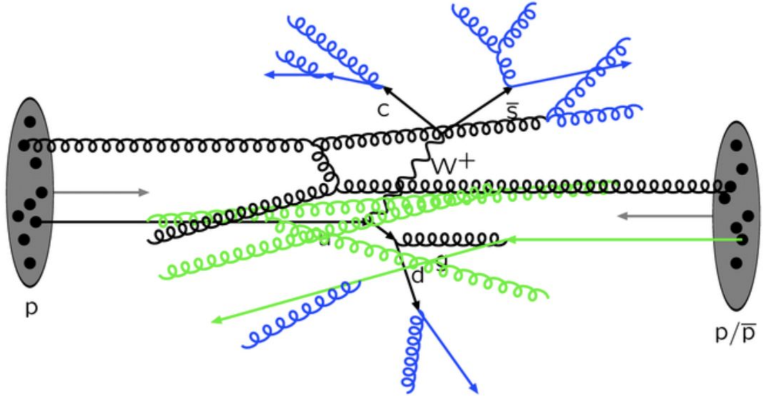
Berkeley UNIVERSITY OF CALIFORNIA

The 3 frontiers of HEP-ex

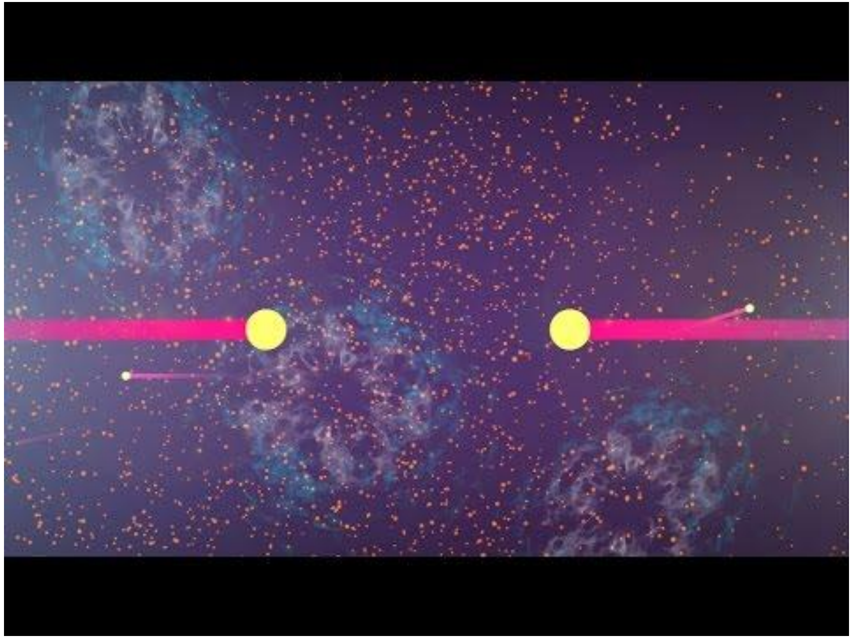
- Energy frontier (collider physics)
  - Create experiments that give us access to higher energy interactions
  - Allows us to look for new particles that might require high energies to be created
- Intensity frontier (neutrino physics)
  - Create experiments that are sensitive to very rare interactions
  - Allows us to study exotic particles such as neutrinos in more detail
- Cosmic frontier (astroparticle physics)
  - Create experiments that allow us to measure particles originating from the cosmos
  - Allows us to study the universe through particle detection

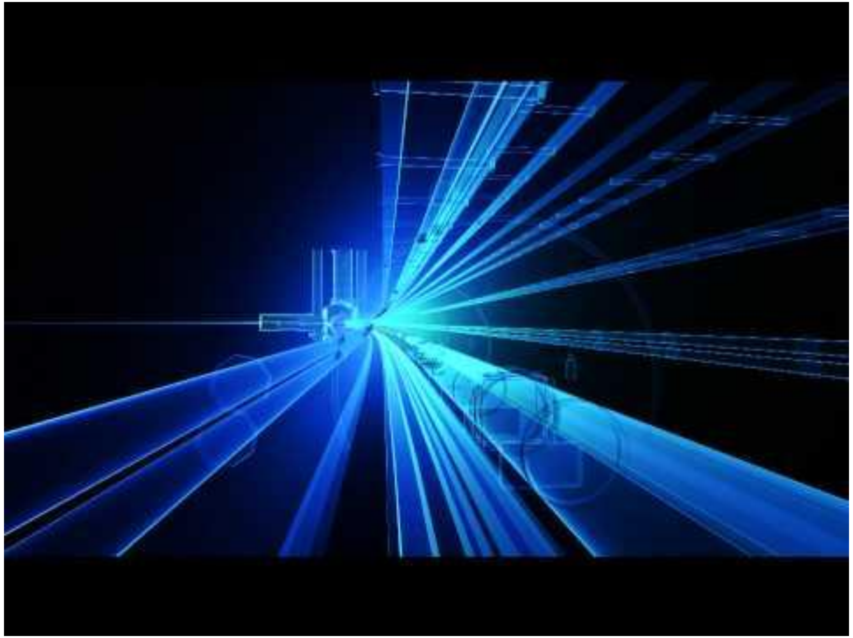
Collider physics

Source: [Florian Bechtel]



- High-energy collisions create complicated decay structures that allow us to probe a wide range of subatomic physics
- There are **many** different possible products that arise from a 2 proton collision at high enough energies
  - Quantum mechanical effects can generate particles via mass-energy equivalence (as long as no conservation laws are broken)





Neutrino detection

- Neutrinos are light, neutral particles that interact only via the weak force
  - They come in 3 flavors: electron, muon and tau
  - Neutrinos oscillate between flavors as they travel
  - Need to be detected indirectly via interactions with detector material
- Two types of neutrino experiments: reactor and cosmic
  - Relevant for both intensity and cosmic frontier
  - Detectors need to be isolated from surroundings to reduce backgrounds
- Many open questions: Mass? Oscillation parameters? Additional neutrino flavors?

The HEP-ex pipeline

1. Build a detector and measure some particle interactions
2. Reconstruct relevant information from measured events
3. Perform a statistical analysis on reconstructed data to extract desired physical quantity

Raw data
- Detector hits
- Images

⇨

Reconstructed objects
- Particle trajectories
- Energies, momenta
- Higher-level objects

⇨

Particle parameters
- Cross sections
- Branching ratios
- Masses

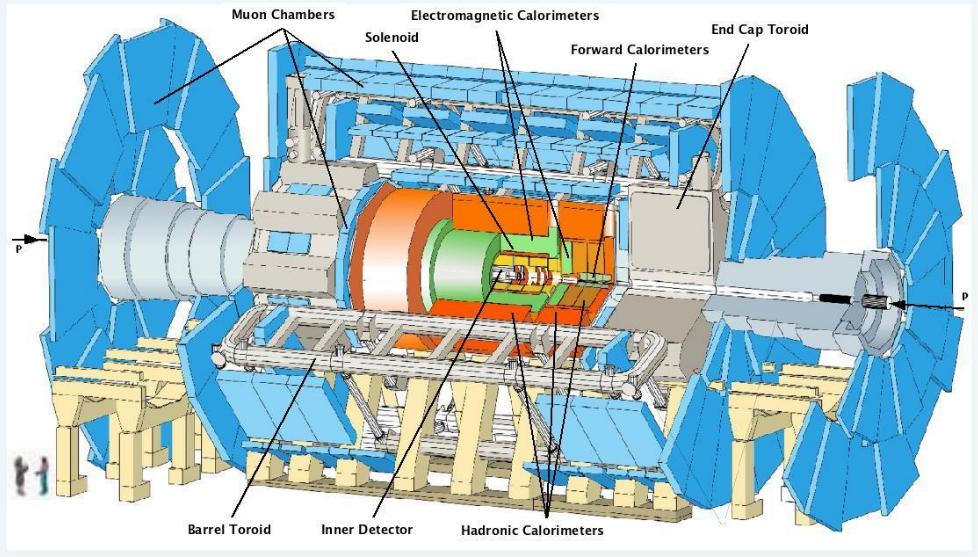
⇨

Interaction parameters
- Couplings
- Mixing parameters

Berkeley
UNIVERSITY OF CALIFORNIA

Particle detectors

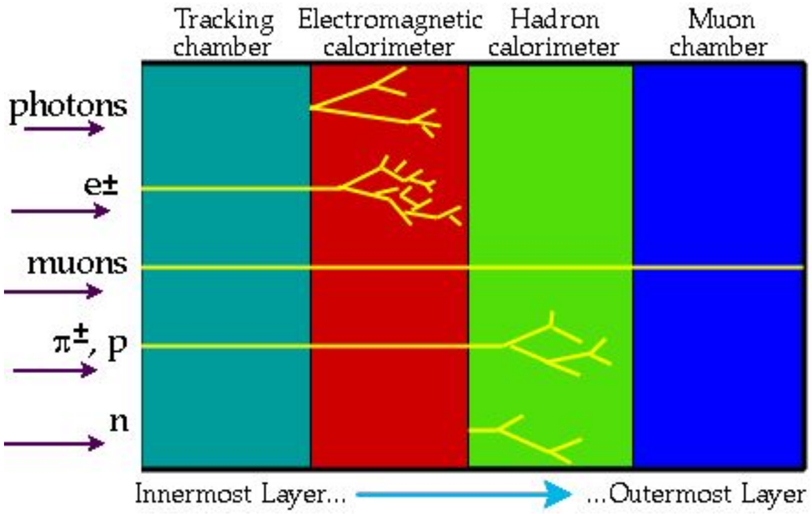
- Particle detector hardware is **very** complicated and diverse
- Modern detectors contain submodules (each with many segments) that measure particle energy deposits and trajectories



Source: [UCI]

Trackers and calorimeters

- Tracking chambers are useful for measuring trajectories
  - Charged particles interact with material and thus leave tracks
  - Particles don't lose much energy
- Calorimeters are useful for measuring energy deposits
  - These slow particles down, causing them to deposit most of their energy
- Magnetic fields allow for the measurement of particle momentum
  - Possible via curvature of tracks

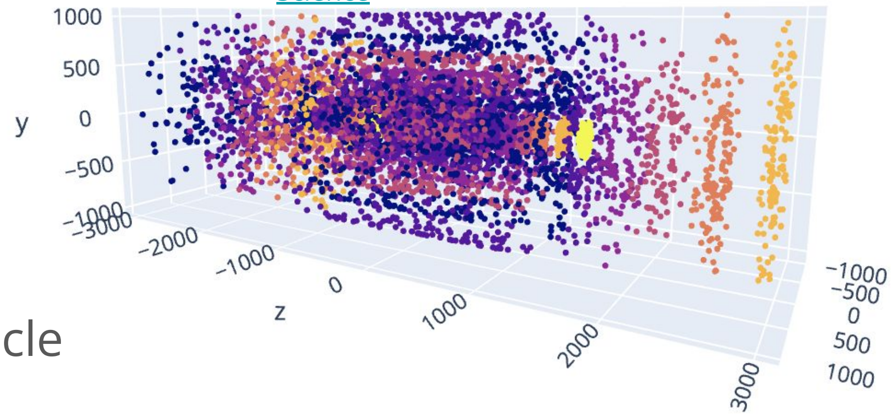


Source: UCI

Berkeley UNIVERSITY OF CALIFORNIA

Track reconstruction

Source: [Towards Data Science]

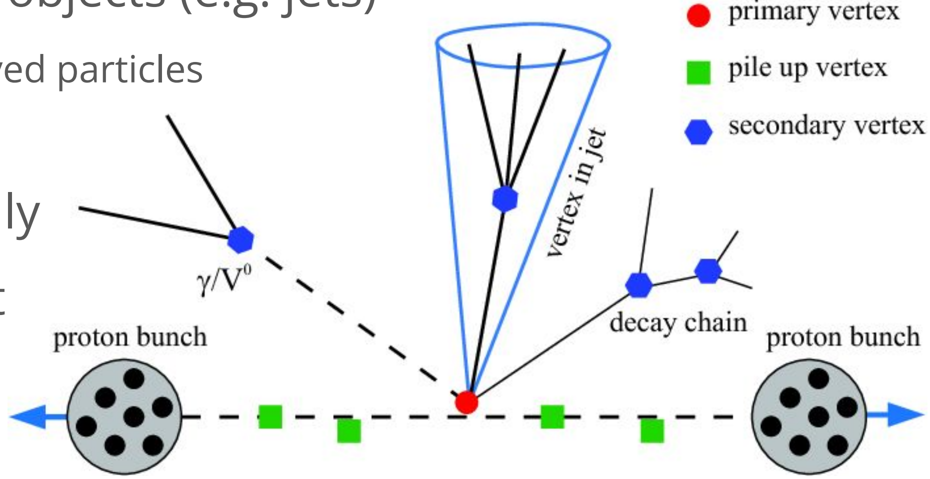


- Raw data from measurements are generally very messy
  - Consists of energy deposit information in each detector segment (“hits”)
- Need to reconstruct individual particle energies/trajectories first
  - Sometimes we can draw conclusions on what type of particle corresponds to a given track based on its visibility in various detector layers

Berkeley
UNIVERSITY OF CALIFORNIA

Examining event structure

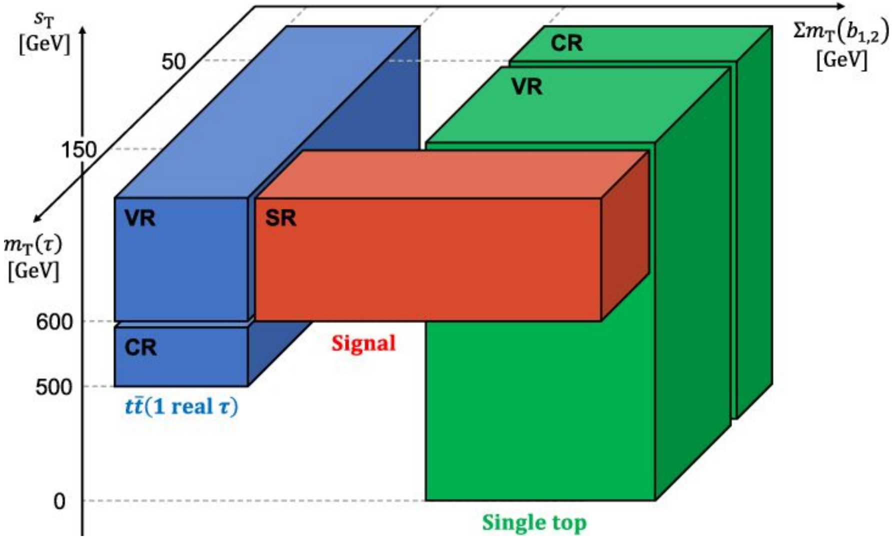
- Once tracks are reconstructed, they can be classified and grouped into decay vertices or other higher-level objects (e.g. jets)
  - Often times, we're interested in short-lived particles that decay rapidly
- From this, we can calculate physically relevant quantities within our event
  - These quantities can be used to separate our signal from our background (e.g. invariant mass of decay vertex)



Source: IEEE

Event selection

- Need to figure out which part of the parameter space is expected to contain most of our signal
    - This will be our signal region (SR)
- Next we need to define regions that are as pure as possible in each of our remaining backgrounds, which are called control regions (CR)



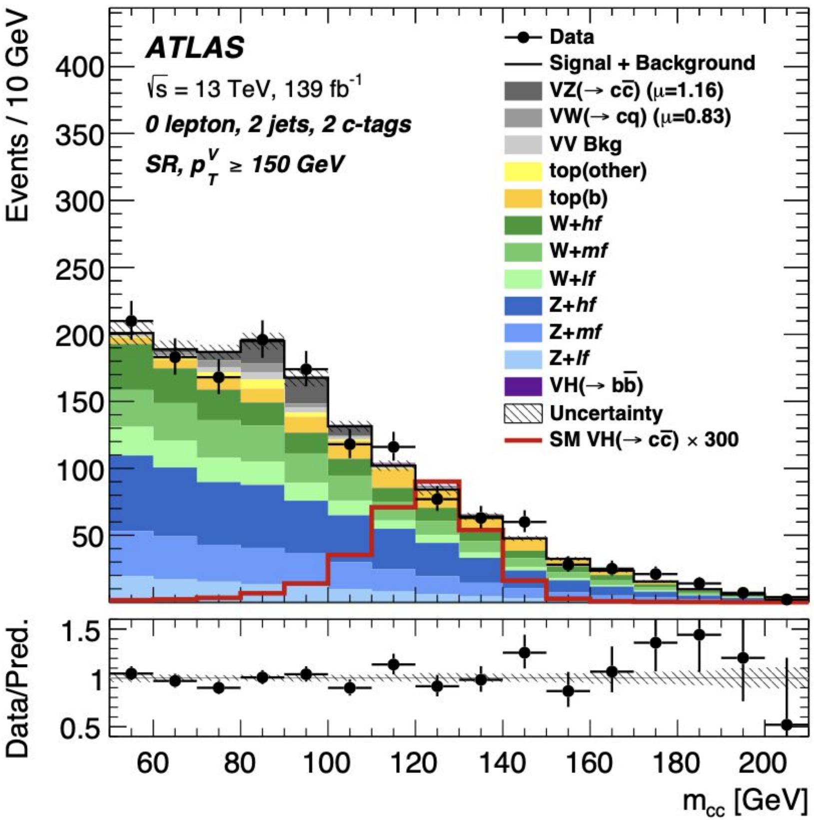
Source: [Phys Rev D]

Berkeley UNIVERSITY OF CALIFORNIA

Example: SR in H->cc

Want to look for a Higgs decaying into 2 charm quarks (i.e. charm jets)

Common selections	
Central jets	≥ 2
Signal jet p_{T}	≥ 1 signal jet with $p_{\mathrm{T}} > 45$ GeV
c-jets	One or two c-tagged signal jets
b-jets	No b-tagged non-signal jets
Jets	2, 3 (0- and 1-lepton); 2, ≥ 3 (2-lepton)
p_{T}^{V} regions	75–150 GeV (2-lepton) > 150 GeV
$\Delta R(\mathbf{jet1},\ \mathbf{jet2})$	$75 < p_{\mathrm{T}}^{V} < 150$ GeV: $\Delta R \leq 2.3$ $150 < p_{\mathrm{T}}^{V} < 250$ GeV: $\Delta R \leq 1.6$ $p_{\mathrm{T}}^{V} > 250$ GeV: $\Delta R \leq 1.2$



Source: [Eur. Phys. J. C]

Signal regions only!

Berkeley UNIVERSITY OF CALIFORNIA

Importance of control regions

- Why can't we just use our knowledge of backgrounds in SR from MC to isolate the signal?
    - Because our MC simulations are very imperfect!
- Control regions allow us to match MC to observational data for each of our backgrounds
    - From this we can infer what our backgrounds look like in data in the SR
- We can then use this knowledge to "filter out" our signal events within SR
    - In practice this is done by a "maximum likelihood fit"

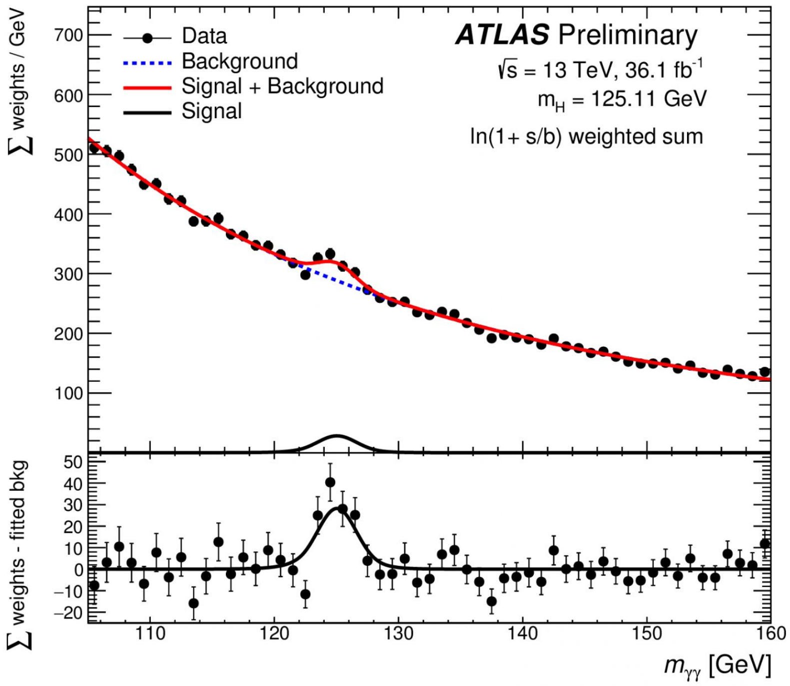
Berkeley
UNIVERSITY OF CALIFORNIA

Likelihood fits

- Need to define a distribution function that describes our signal and background in all regions
  - Informed by MC simulations (signal and background modeling)
  - This function will have a variety of parameters that can be tuned by the fit, one of which is our POI (parameter of interest) while the others are called nuisance parameters (NP)
- The fit will perform a maximization of our likelihood function that chooses the set of parameters that gives the best agreement with our data
  - It will match our model, which is based on MC to the data as best it can
  - From this we get a best-fit value of our POI (signal strength) as well as our NPs

Example: Higgs discovery

- Select H->$\gamma\gamma$ candidate events
  - Reconstruct mass of $\gamma\gamma$ decay vertex
- Split candidate events into SRs and CRs based on expected backgrounds
- Small bump in data shows statistically significant deviation from background only hypothesis



Source: ATLAS

Where does machine learning come in?

- Modern particle physics experiments use ML extensively throughout their event reconstruction pipeline
- ML algorithms also often used to separate different kinds of event signatures in an analysis
  - Neural network output scores can be used as variables in a likelihood fit
- Increasingly elaborate experimental setups require more and more efficient and performant algorithms
  - We **cannot** get by without utilizing ML and AI techniques anymore

Berkeley
UNIVERSITY OF CALIFORNIA

Machine learning in practice

- Machine learning algorithms are almost always implemented in python
    - Julia is a new language that is slowly becoming more relevant for this (more friendly to heterogeneous computing environments including GPU’s, TPU’s, etc.)
- Two main packages currently used in research/industry:
    - Tensorflow + Keras (affiliated with Google)
    - pyTorch (affiliated with Meta)
- Core concepts transfer easily between both
    - We will use pyTorch throughout this workshop as the syntax is a bit nicer

Hands-on exercise #1

Intro to numpy and pyTorch

Exercise agenda

- Basic knowledge of python is assumed
  - If you don’t know python or are not comfortable with it find a partner who can help you
- Exercise #1: numpy (Credit: Stanford University)
  - Introduction to numpy used in Stanford’s “Intro to Deep Learning” class
- Exercise #2: pyTorch (Credit: pyTorch Team)
  - We will only do the “Tensors” and “Datasets & Data Loaders” chapters today
- Work in teams and let us know if you have any questions!
- Please take a moment to fill out the feedback form when you’re done