From Raw data to Physics Results

at the LHC

August 2022

PSI Particle Physics Summer School

Anna Sfyrla, UniGe

For any question don't hesitate to get in touch:

anna.sfyrla@unige.ch

What will this lecture be about?

Introduction

- Definitions and basic concepts

Input to the physics

- The data: trigger, data preparation
- The theory: Monte carlo simulations
- Reconstruction, or how to translate detector signals to particles

Physics analyses

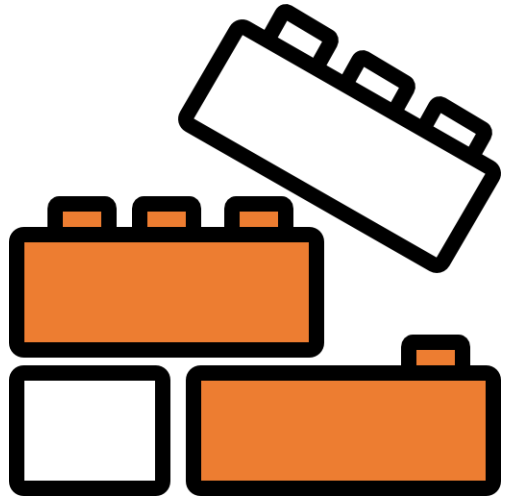
- Through example, step-by-step
- Discussion of analysis methods

Machine Learning in HEP

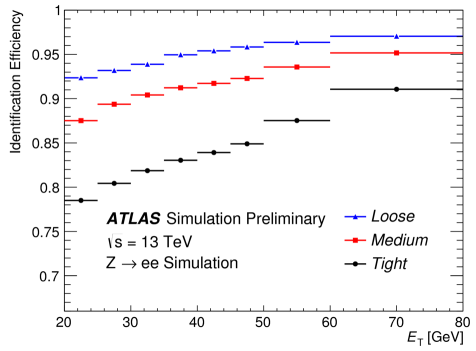
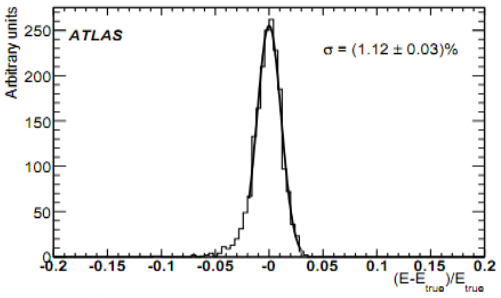
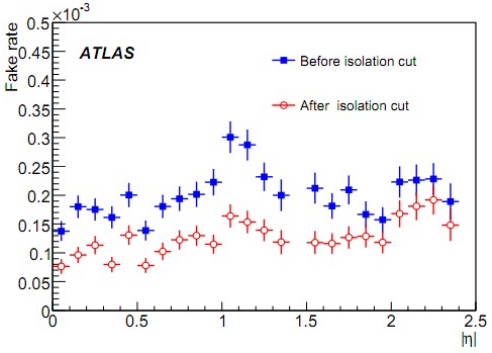
- Just a teaser!

PART 3

Reconstruction



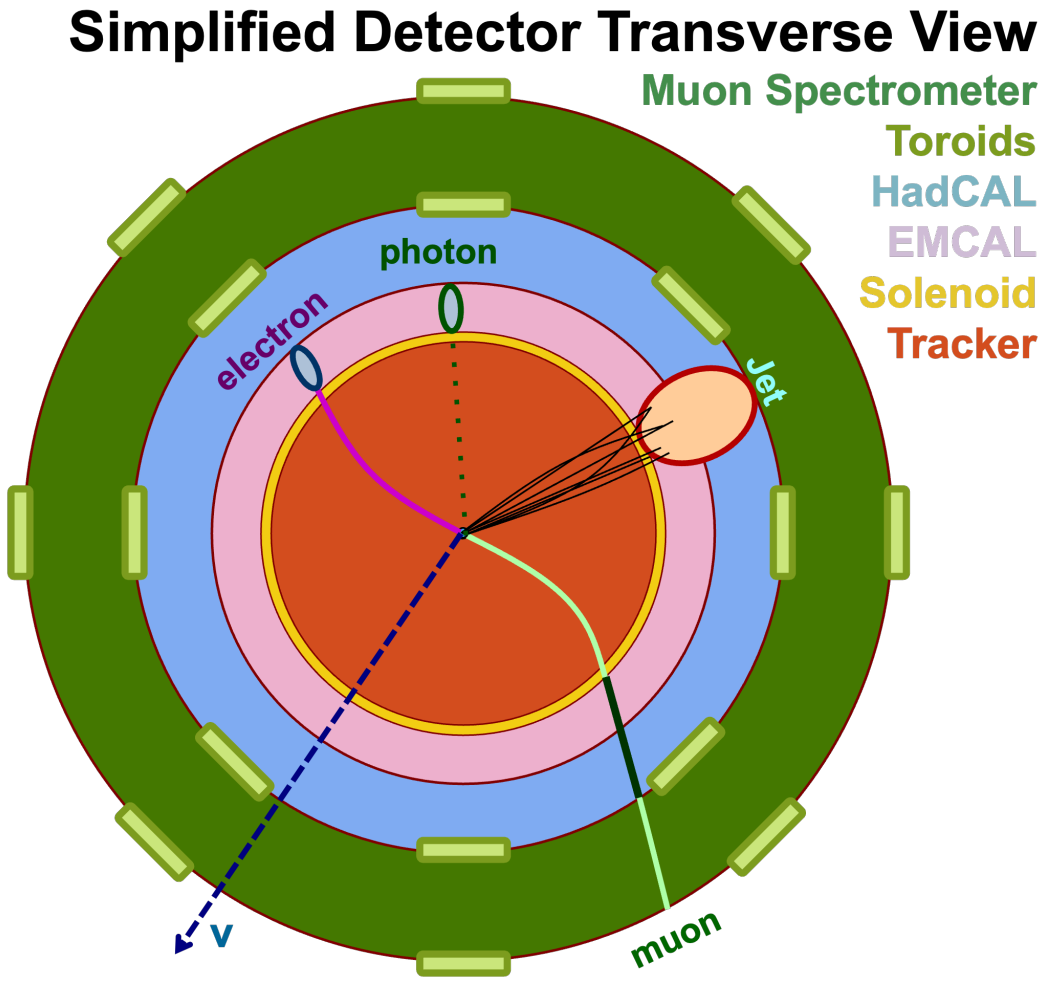
Reconstruction – Figures of Merit

	Definition	Example		Needs be:
Efficiency	how often do we reconstruct the object we are interested in	**electron identification efficiency** = (number of reconstructed electrons) / (number of true electrons) in bins of transverse momentum		**High**
Resolution	how accurately do we reconstruct the quantity	**energy resolution** = (measured energy – true energy) / (true energy)		**Good** *(a small number)*
Fake rate	how often we reconstruct a different object as the object we are interested in	**a jet faking an electron, fake rate** = (Number of jets reconstructed as an electron) / (Number of jets) in bins of pseudorapidity		**Low**

What do we reconstruct?

Tracks and **clusters**

- Combining those:
  - "objects", i.e. "particles"

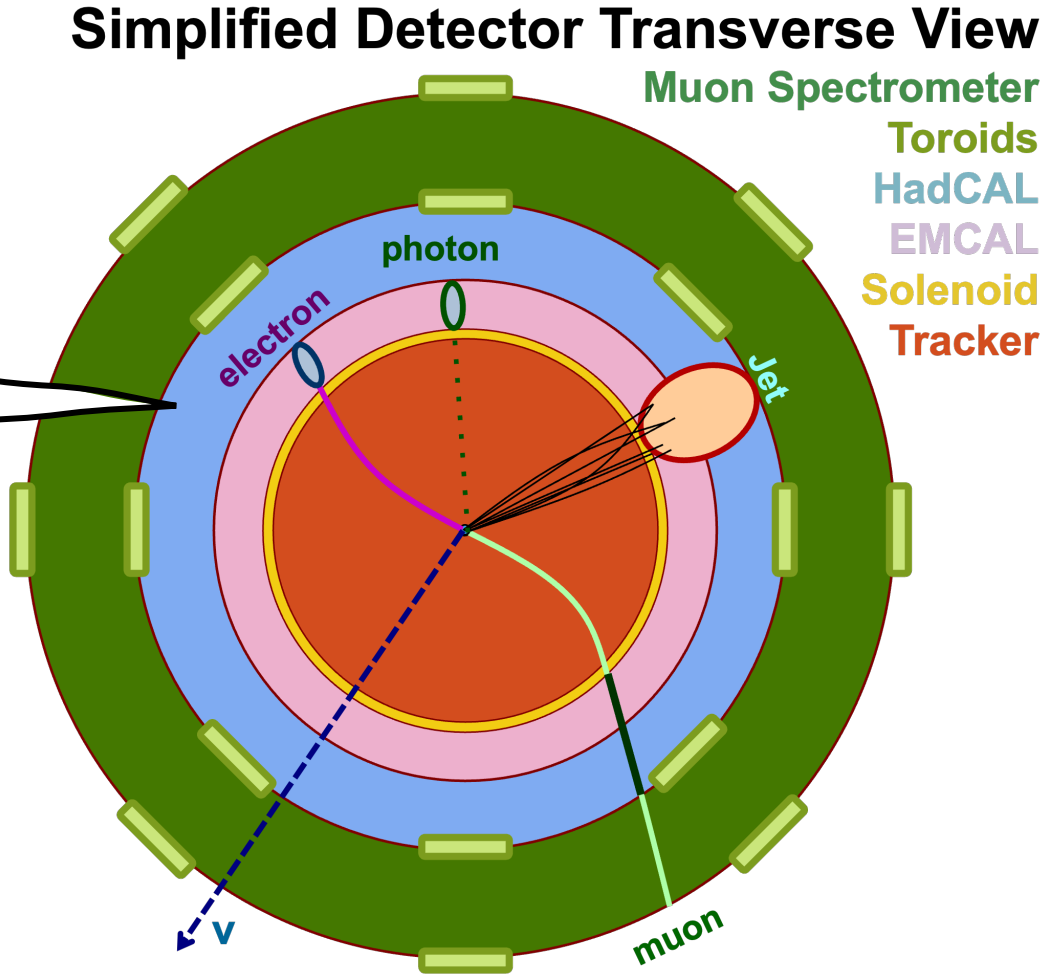


What do we reconstruct?

Tracks and clusters

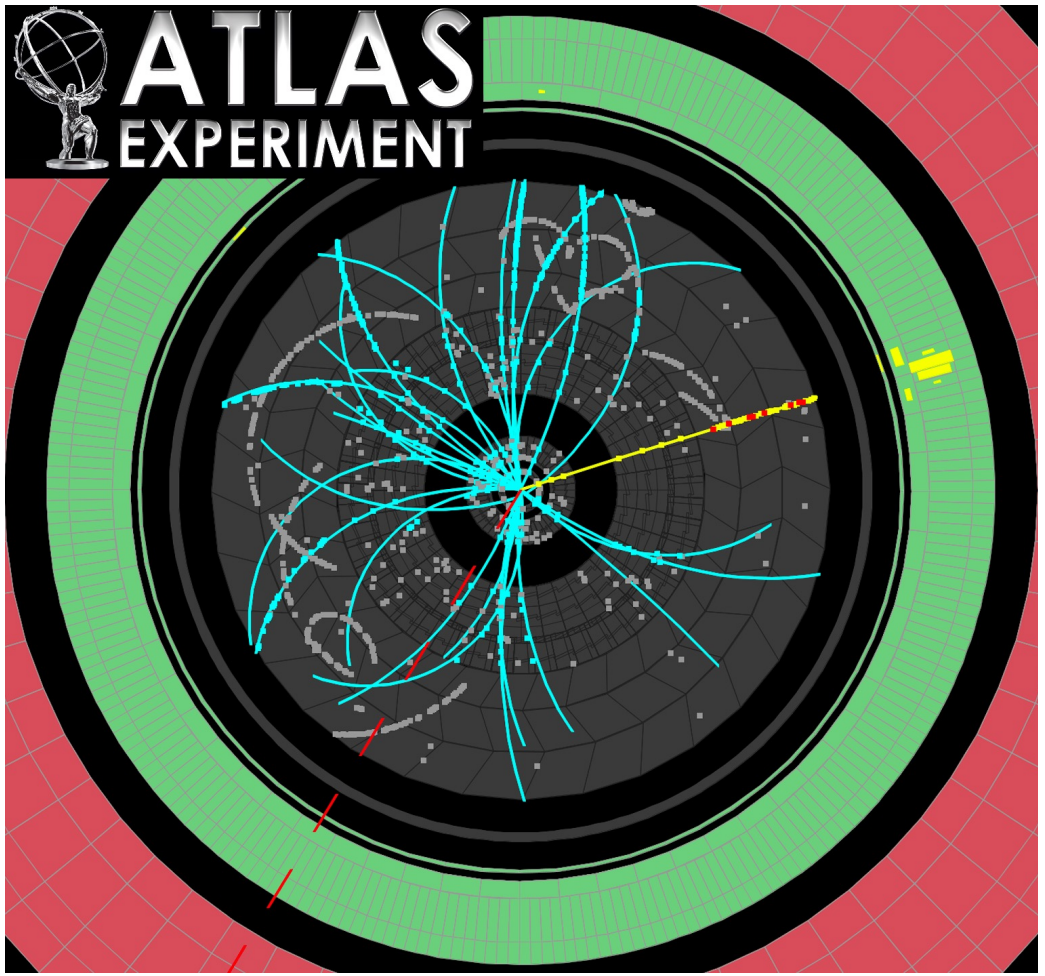
Combining those:

- "objects", i.e. "particles"

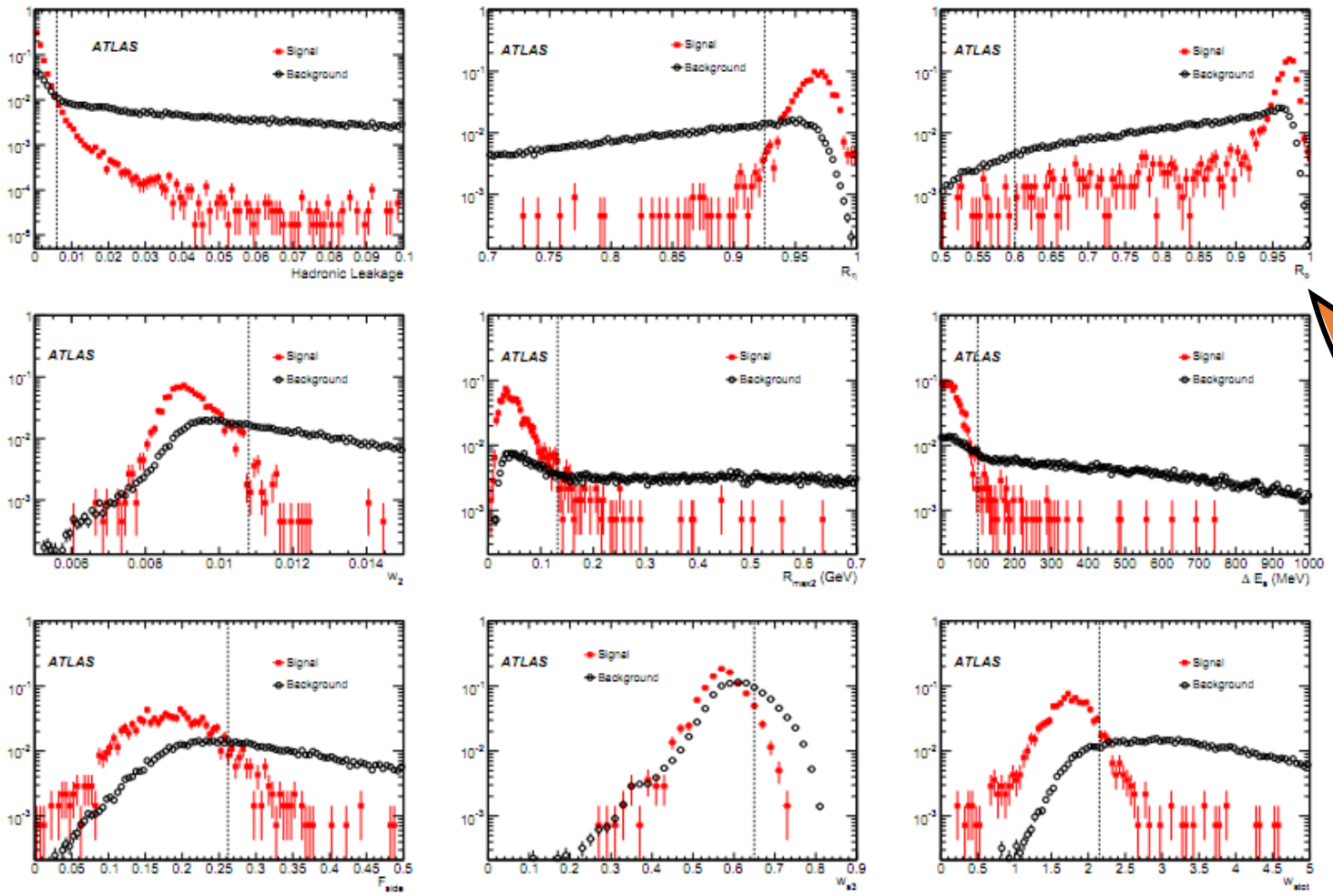


Electron reconstruction

- **Electron momentum measurement** can come from tracking or calorimeter information (or a combination of both)
  - Often have a final calibration to give the best electron energy
- Working points define categories
  - E.g. loose, medium, tight
  - **Trade-off**: Efficiency vs Fakes
- Often want "**isolated electrons**"
  - Require little calorimeter energy or tracks in the region around the electron



Electrons – Identification Algos

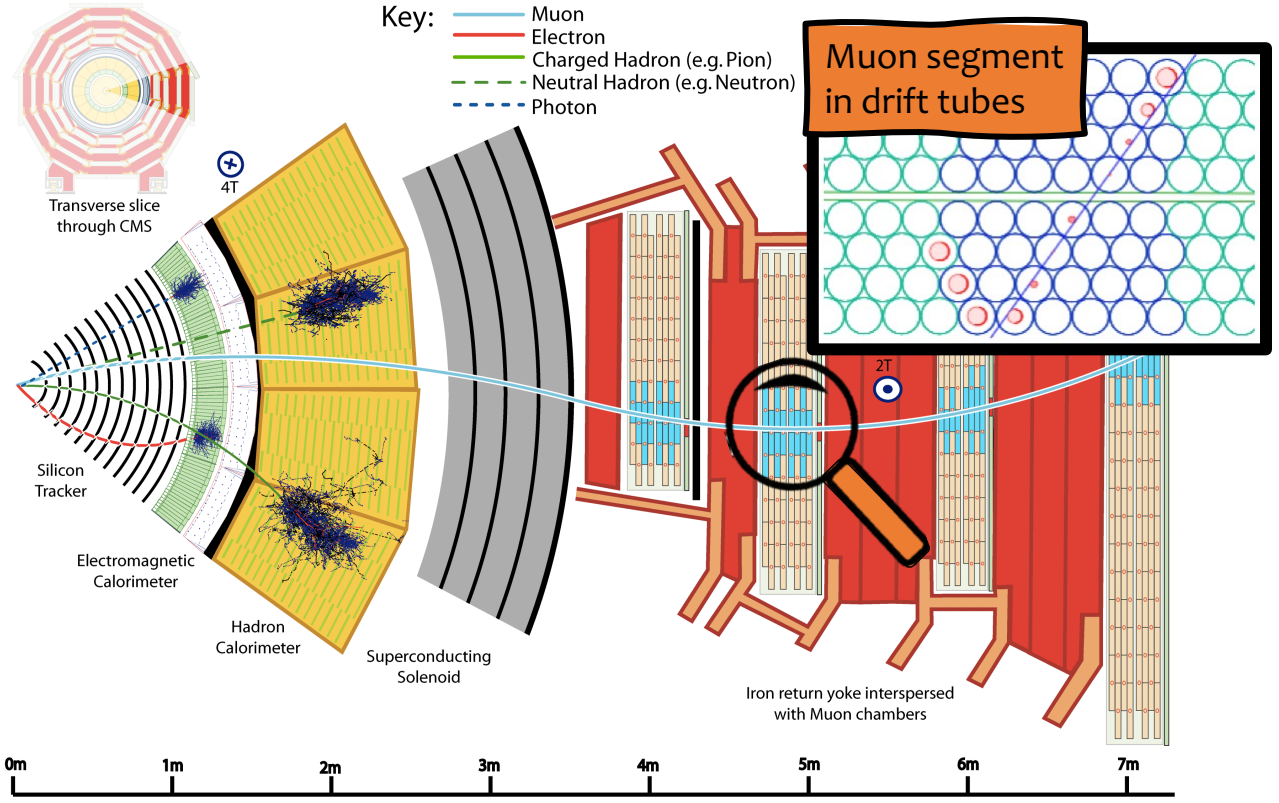


Example of different calorimeter shower shape variables used to distinguish electron showers from jets in ATLAS

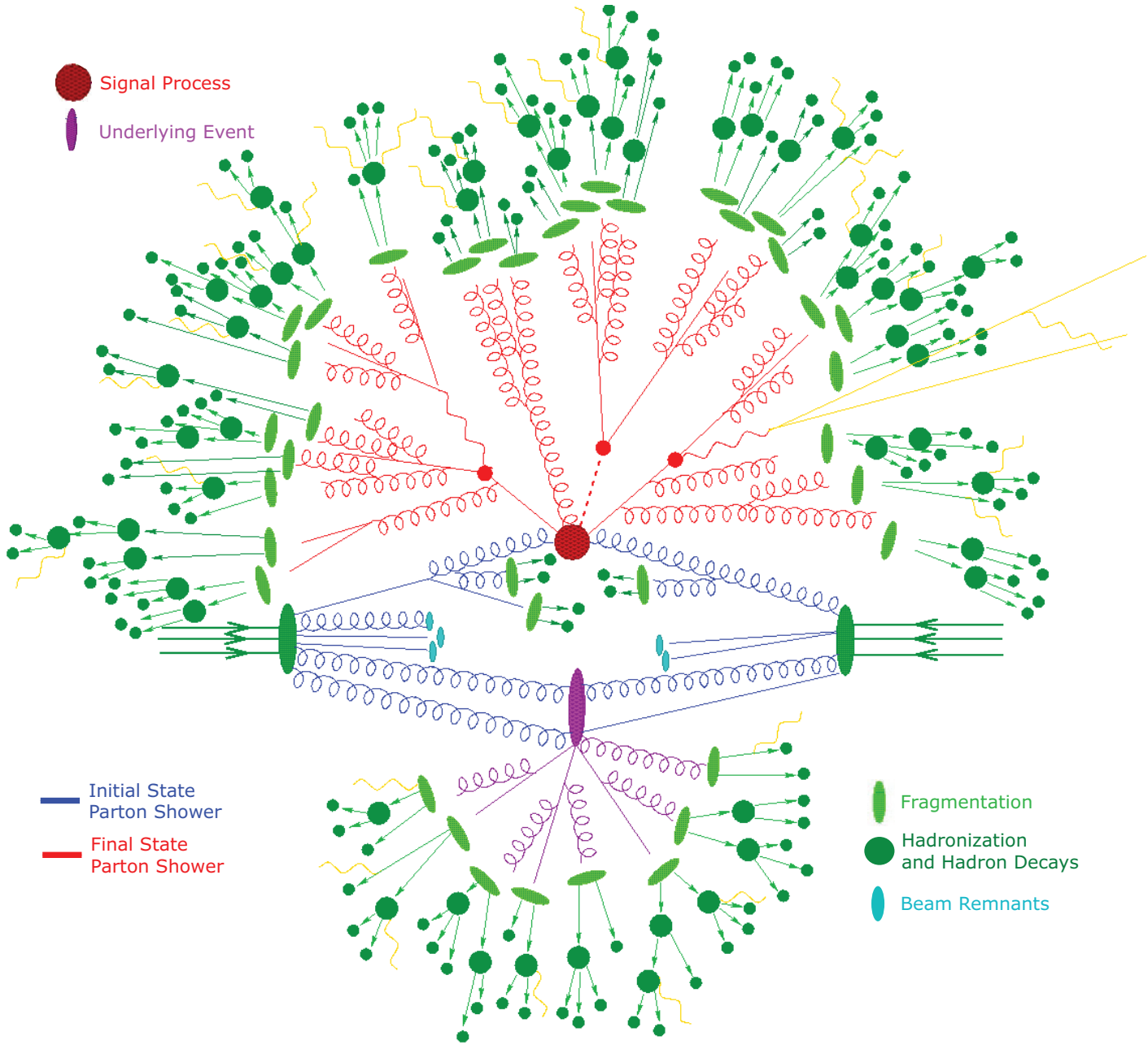
Information can be exploited using multi-variate techniques such as **likelihood discriminants** or **boosted decision trees** or **other machine learning methods**.

Muons

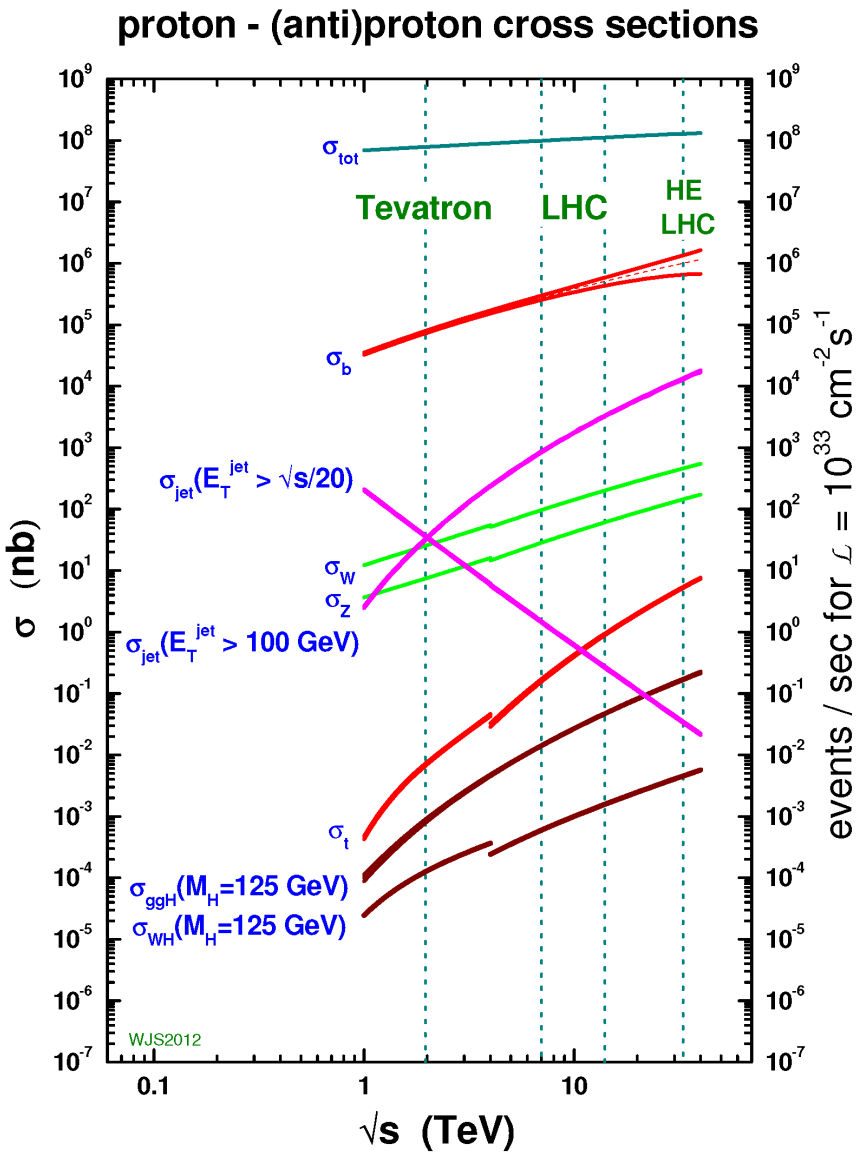
- Combine the **muon segments found in the muon detector with tracks from the tracking detector**
- Momentum of muon determined from bending due to magnetic field in tracker and in muon system
  - Combine measurements to get best resolution
  - Need an accurate map of magnetic field in the reconstruction software
  - Alignment of the muon detectors also very important to get best momentum resolution



JETS


Signal Process
Underlying Event
Initial State
Parton Shower
Final State
Parton Shower
Fragmentation
Hadronization
and Hadron Decays
Beam Remnants

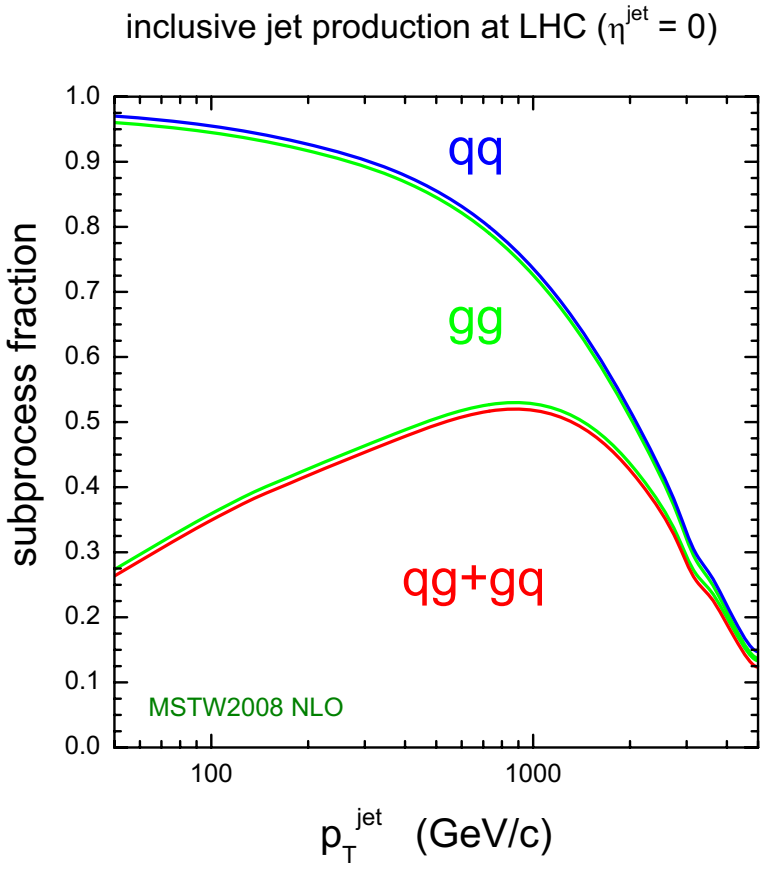
Jet production processes



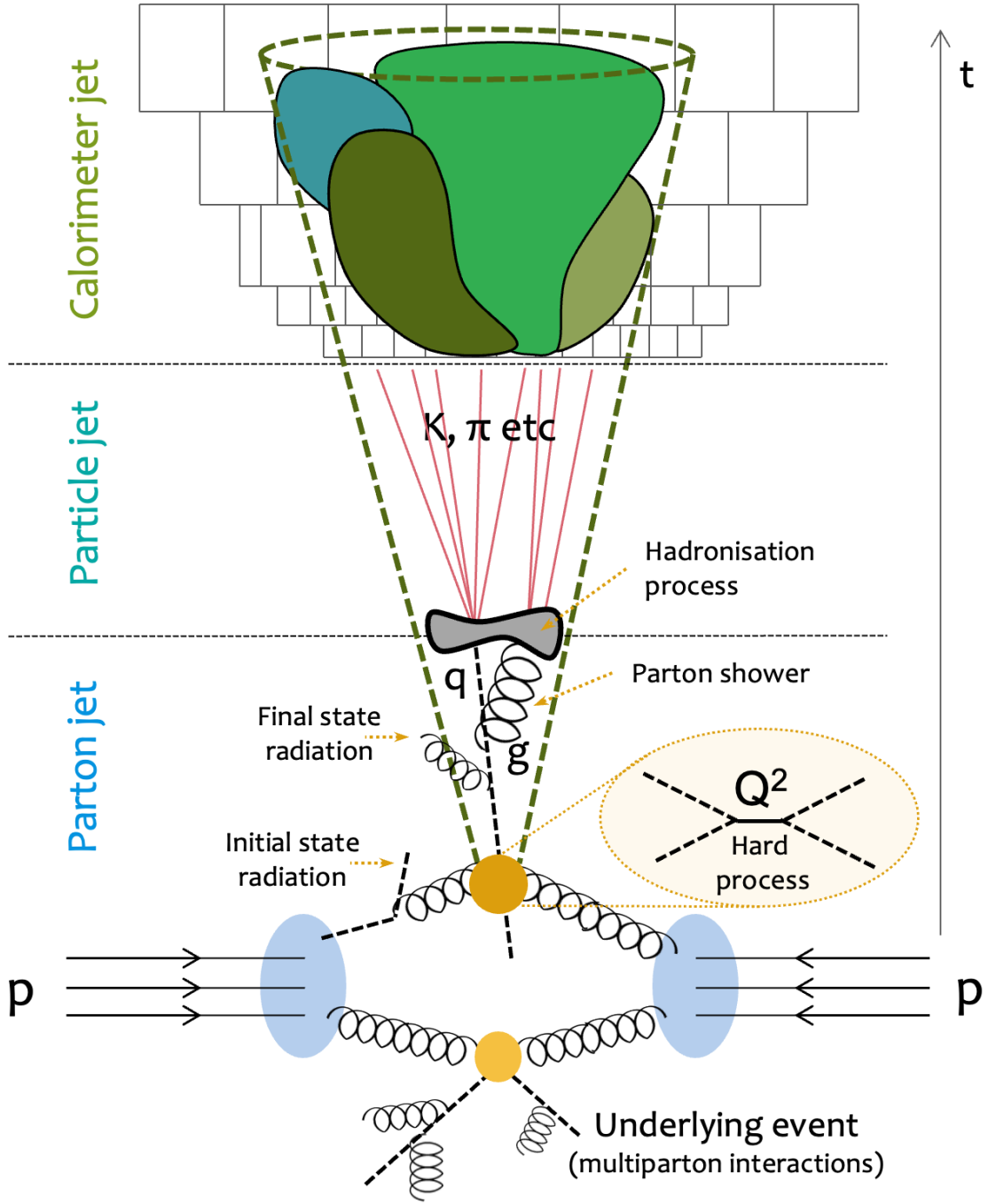
Jets are produced:

- by fragmentation of gluons and (light) quarks in QCD scattering
- by decays of heavy Standard Model particles, e.g. W & Z
- in association with particle production in Vector Boson Fusion, e.g. Higgs
- in decays of beyond the Standard Model particles, e.g. in SUSY

JETS

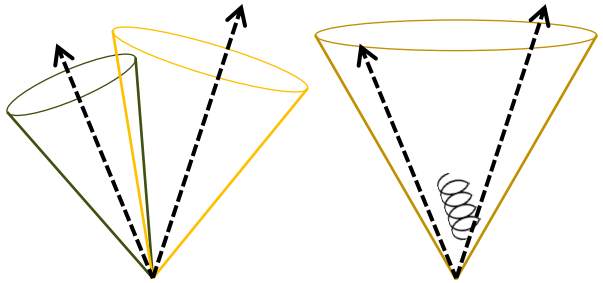


At low energy, jets are more likely produced by gluon fusion.

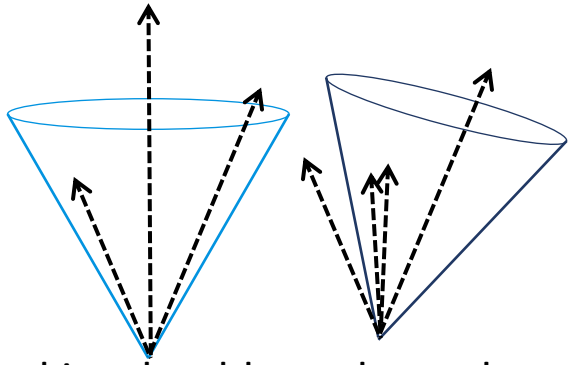


Jet Algorithms

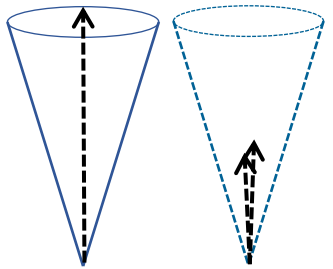
- **Theory requirements:** infrared and collinear safe



Soft gluon radiation should not merge jets



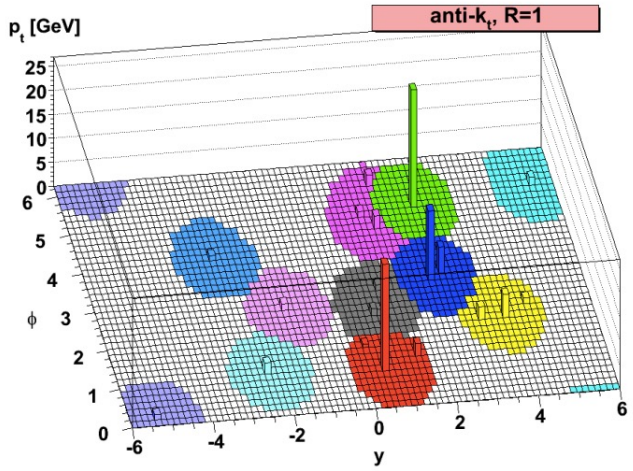
Final jet should not depend on the ordering of the seeds…



…and on signal split in two possibly below threshold

- **Experimental requirements:** Independent to detector technology and data taking conditions, easily implementable

- **Jet algorithm commonly used at the LHC: 'anti-k_t'.** A 'recursive recombination' algorithm. Starts from (topo-)clusters. Hard stuff clusters with nearest neighbor. Various cone sizes (standard R=0.4/0.5, "fat" R=1.0).



arXiv: 0802.1189

Jet Calibration

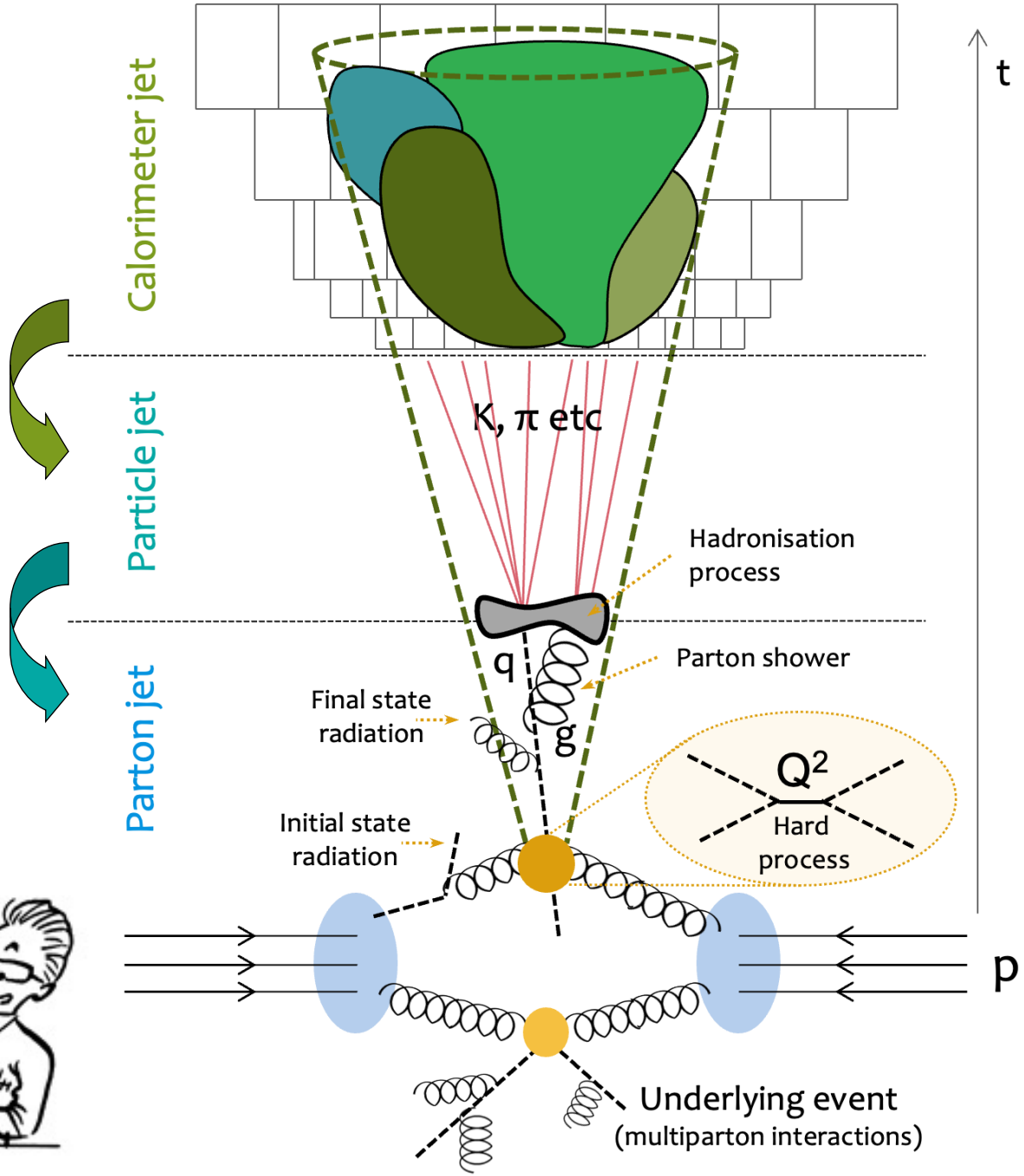
- Correct the energy and position measurement and the resolution.
- Account for:

<u>Instrumental effects</u>

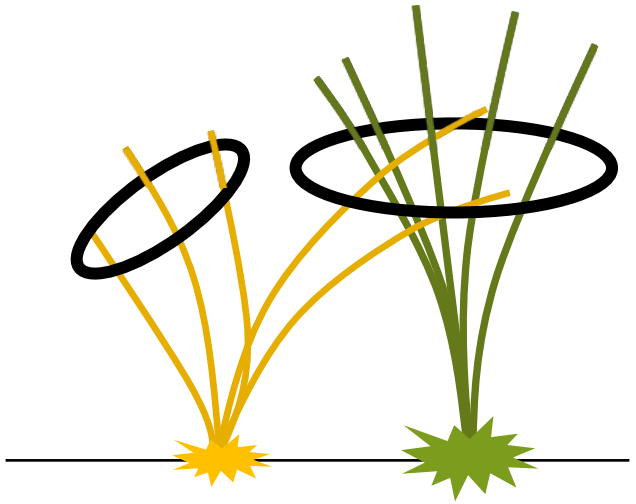
Detector inefficiencies
'Pile-up'
Electronic noise
Clustering, noise suppression
Dead material losses
Detector response
Algorithm efficiency

<u>Physics effects</u>

Algorithm efficiency
'Pile-up'
'Underlying event'



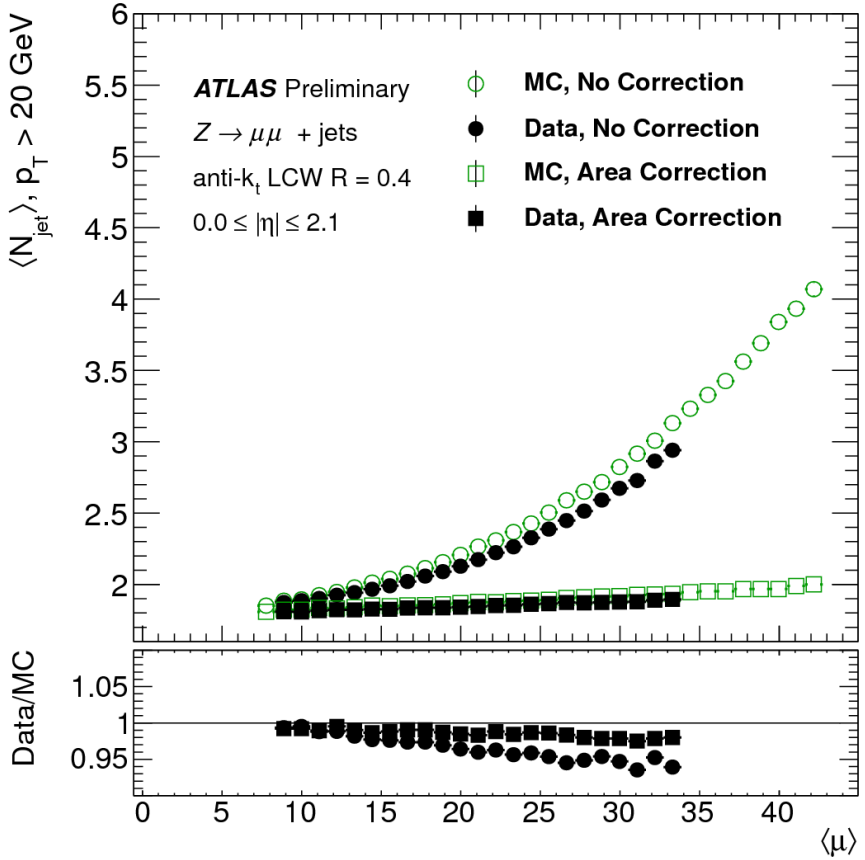
Jets and Pile-Up



Multiple interactions from pile-up

‘Jet-areas’ corrections

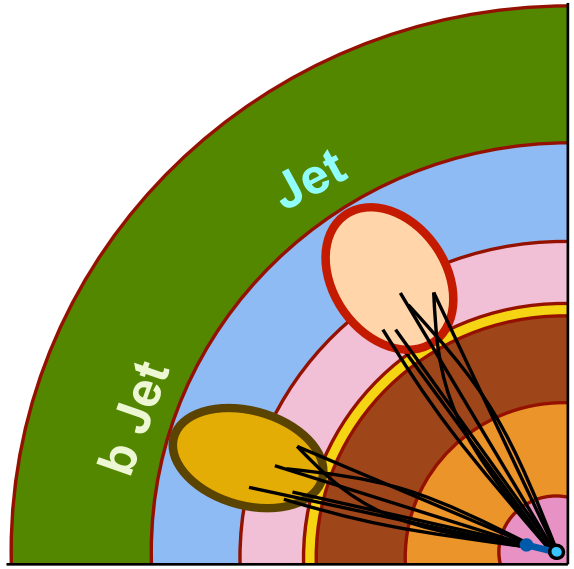
Inspired by arXiv:0707.1378

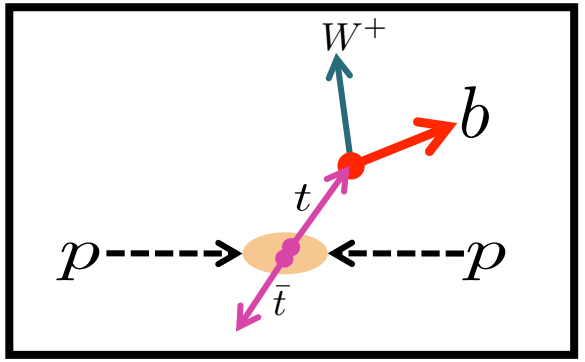


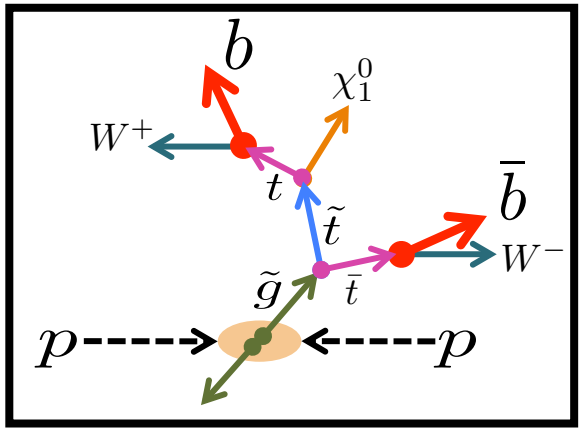
B-Jets

- b-hadrons have a lifetime of ~ 10^{-12} s.
- They travel a small distance (fraction of mm) before decaying.
- A "**displaced vertex**" creates a distinct jet, so b-jets can be tagged (**b-tagged**).
- b-tagging uses sophisticated algorithms, mostly **multi-variate (machine learning)**.

- b-jets create distinct final states, important for both **Standard Model measurements** and **searches for New Physics**.





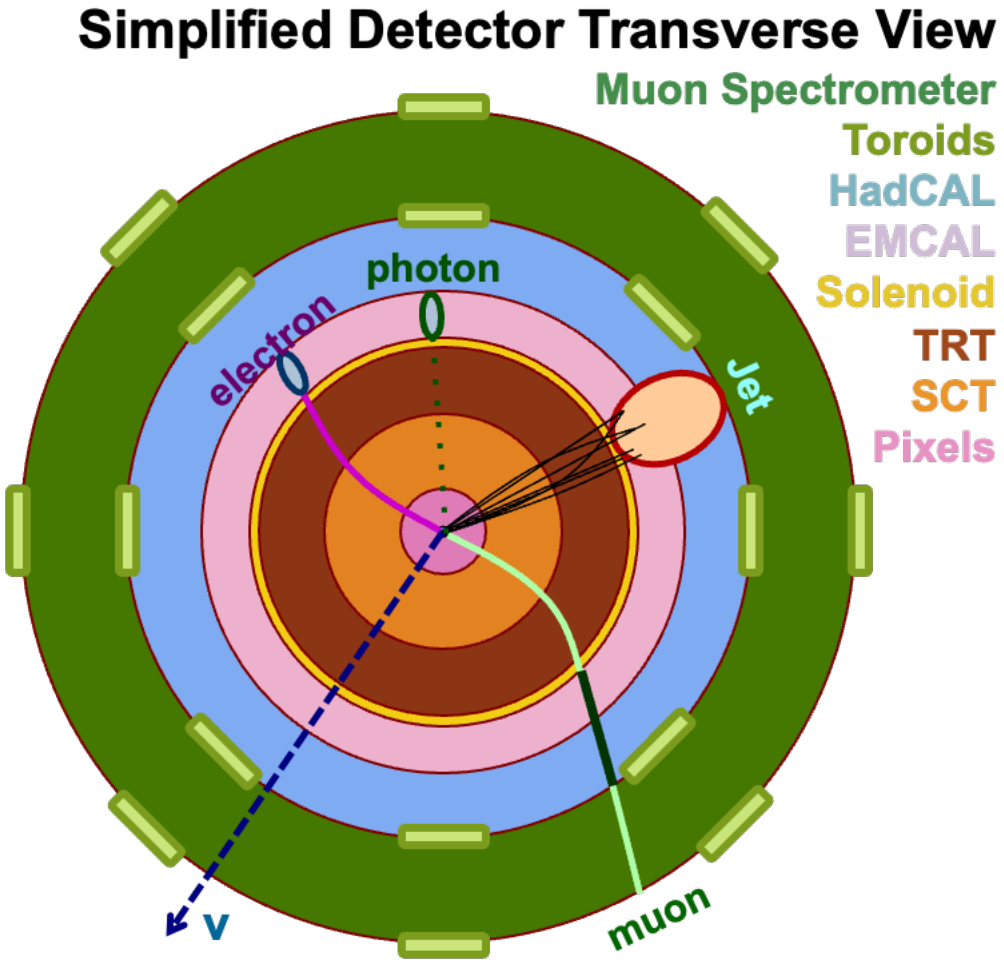


M.E.T.
THE
EXTRA-TERRESTRIAL
THE 20th ANNIVERSARY

Missing Transverse Momentum – ME_T

	I	II	III	
Quarks	2.4 MeV u	1.3 GeV c	170 GeV t	0 γ
	4.8 MeV d	104 MeV s	4.2 GeV b	0 g
Leptons	<2 eV ν_e	<2 eV ν_μ	<2 eV ν_τ	91 GeV Z
	0.5 MeV e	16 MeV μ	1.8 GeV τ	80 GeV W
				126 GeV H

Bosons



In the transverse plane:

$$\Sigma_i \vec{p}_{T,i} = 0$$

So for what we can't directly measure (e.g. neutrinos)

$$E_{\mathrm{T}}^{\mathrm{miss}} = -\Sigma_i \vec{p}_{T,i}$$

Missing Transverse Momentum – ME_T

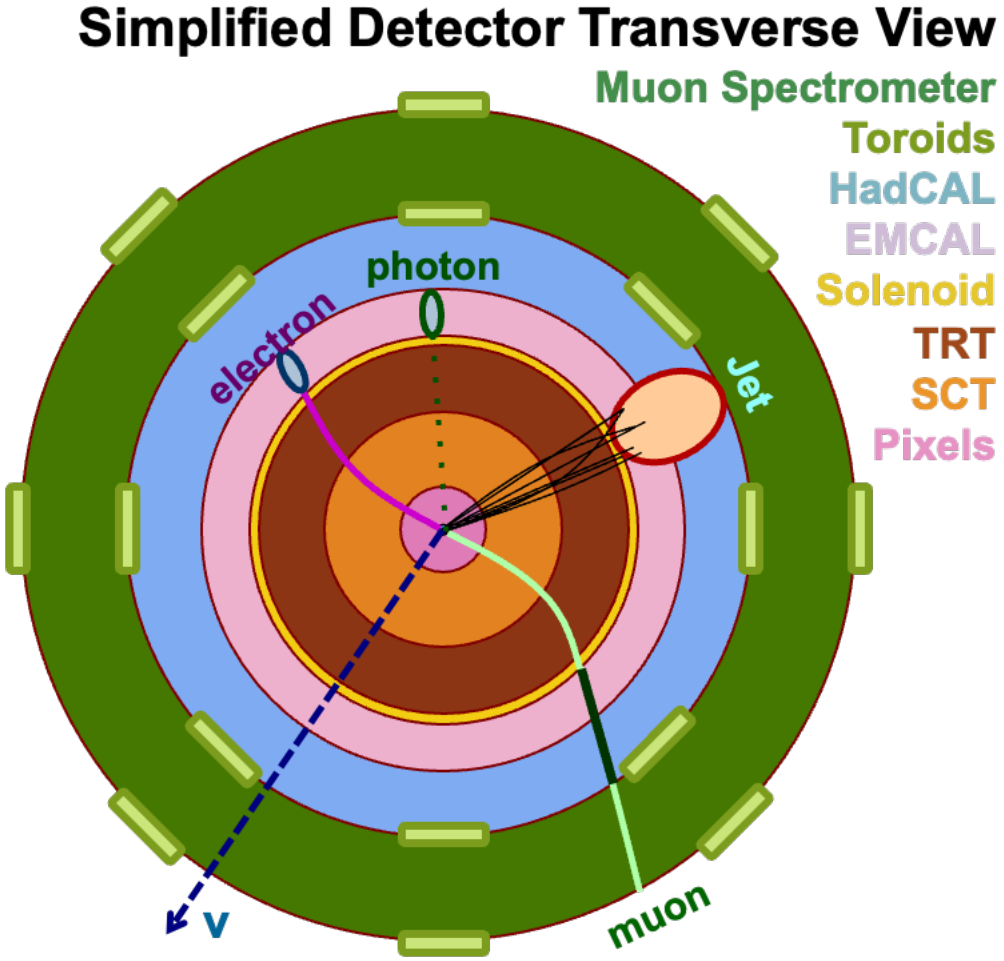
In the transverse plane:

$$\Sigma_i \vec{p}_{T,i} = 0$$

So for what we can't directly measure (e.g. neutrinos)

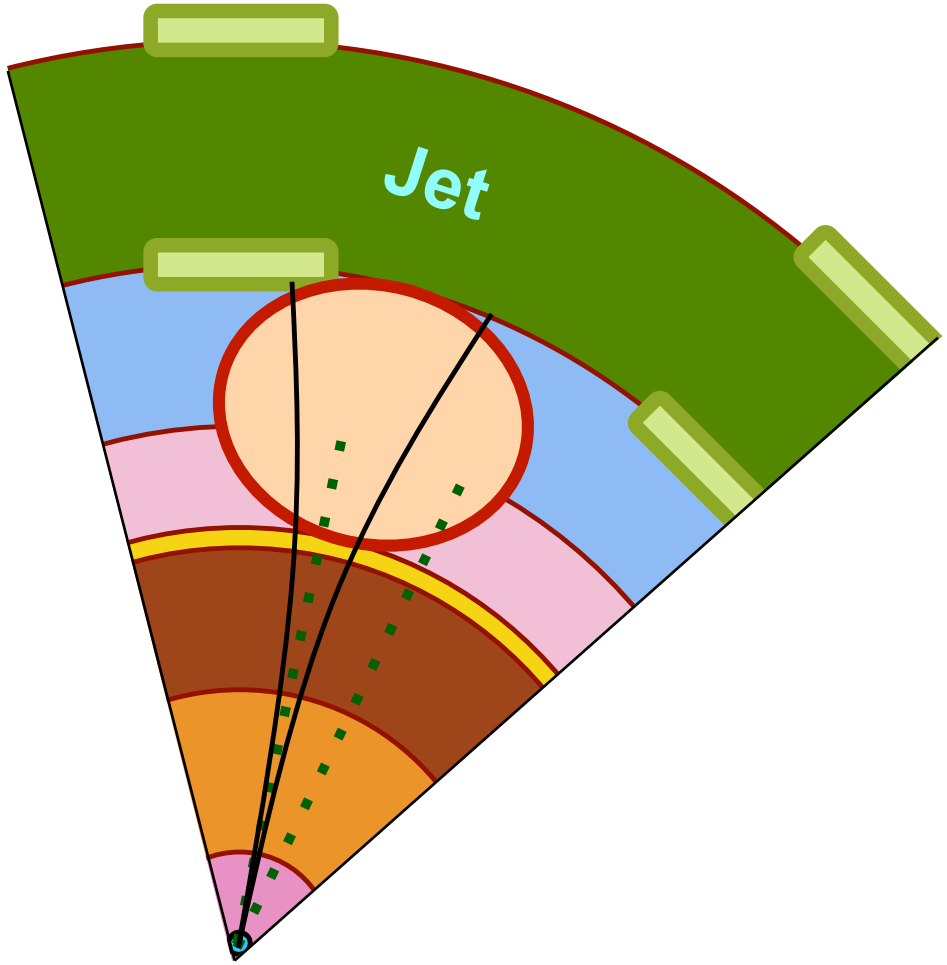
$$E_{\mathrm{T}}^{\mathrm{miss}} = -\Sigma_i \vec{p}_{T,i}$$

Or dark matter candidates!

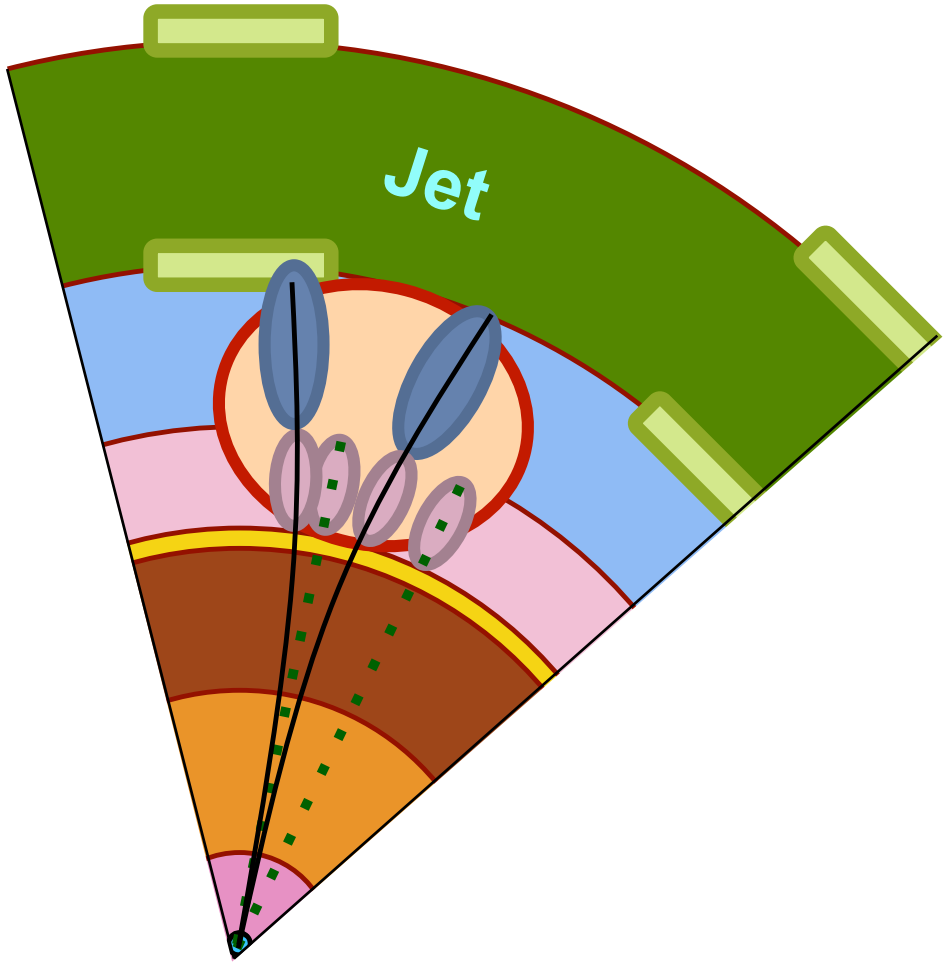


Particle Flow
for Hadronic Reconstruction

Particle flow

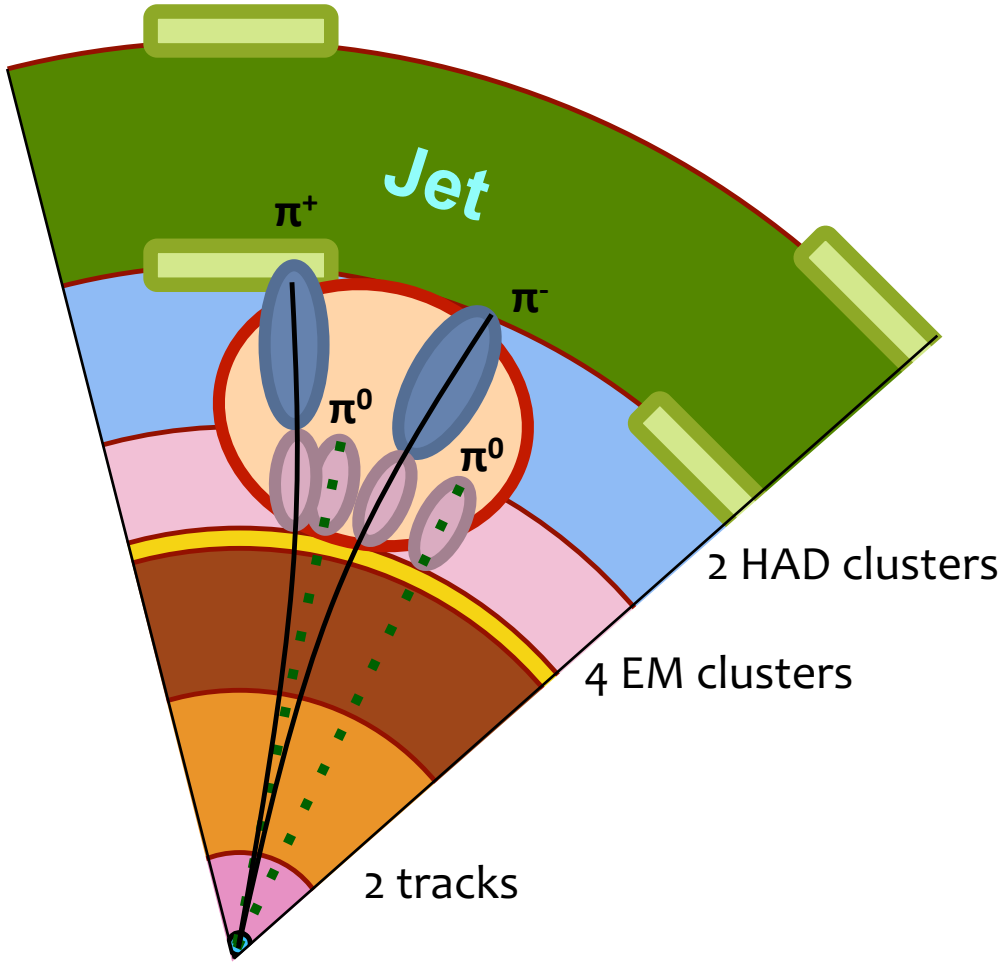


Particle flow

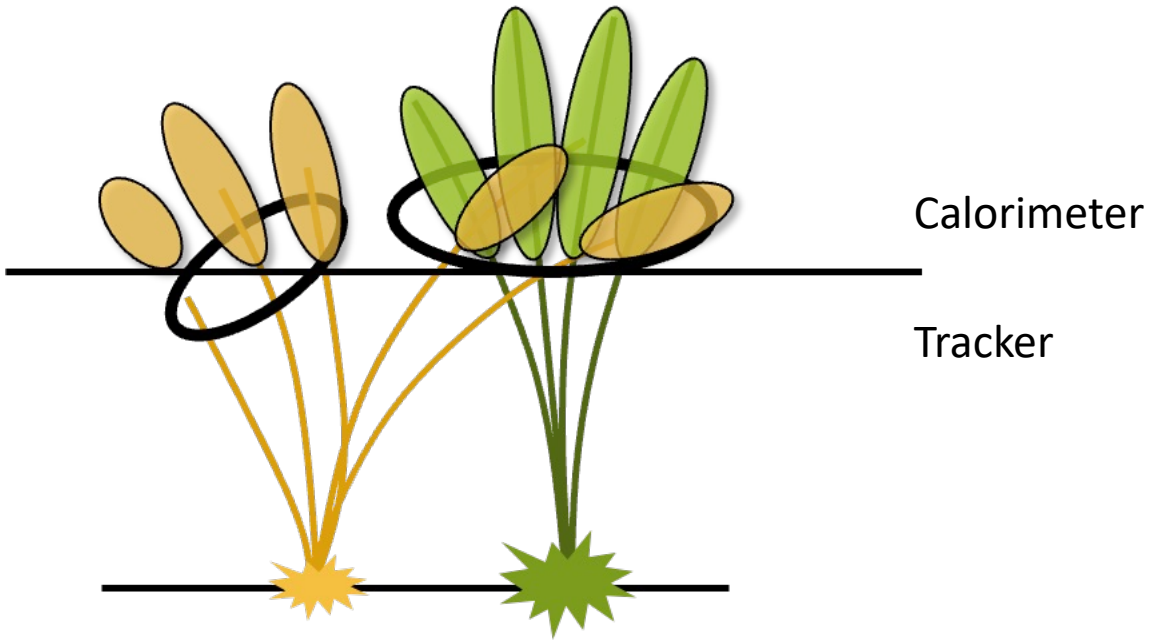


Particle flow

- "Flow of particles" through the detector.
- Reconstruct and identify all particles, photons, electrons, pions, …
- Use best combination of all sub-detectors for measuring the properties of the particles.
- First used at LEP (ALEPH) and then at the LHC (CMS).

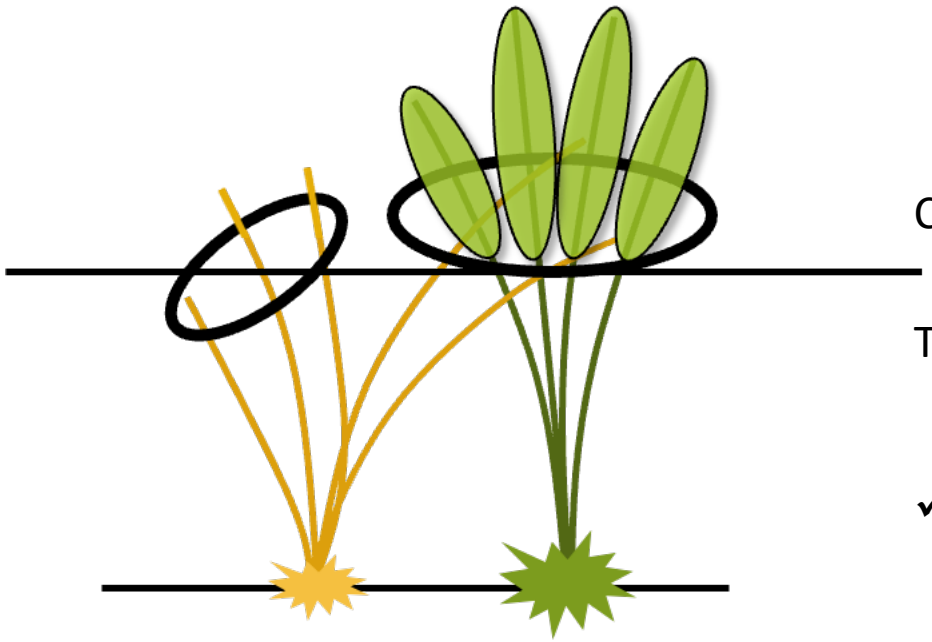


Jets in Pile-up



Multiple interactions from pile-up

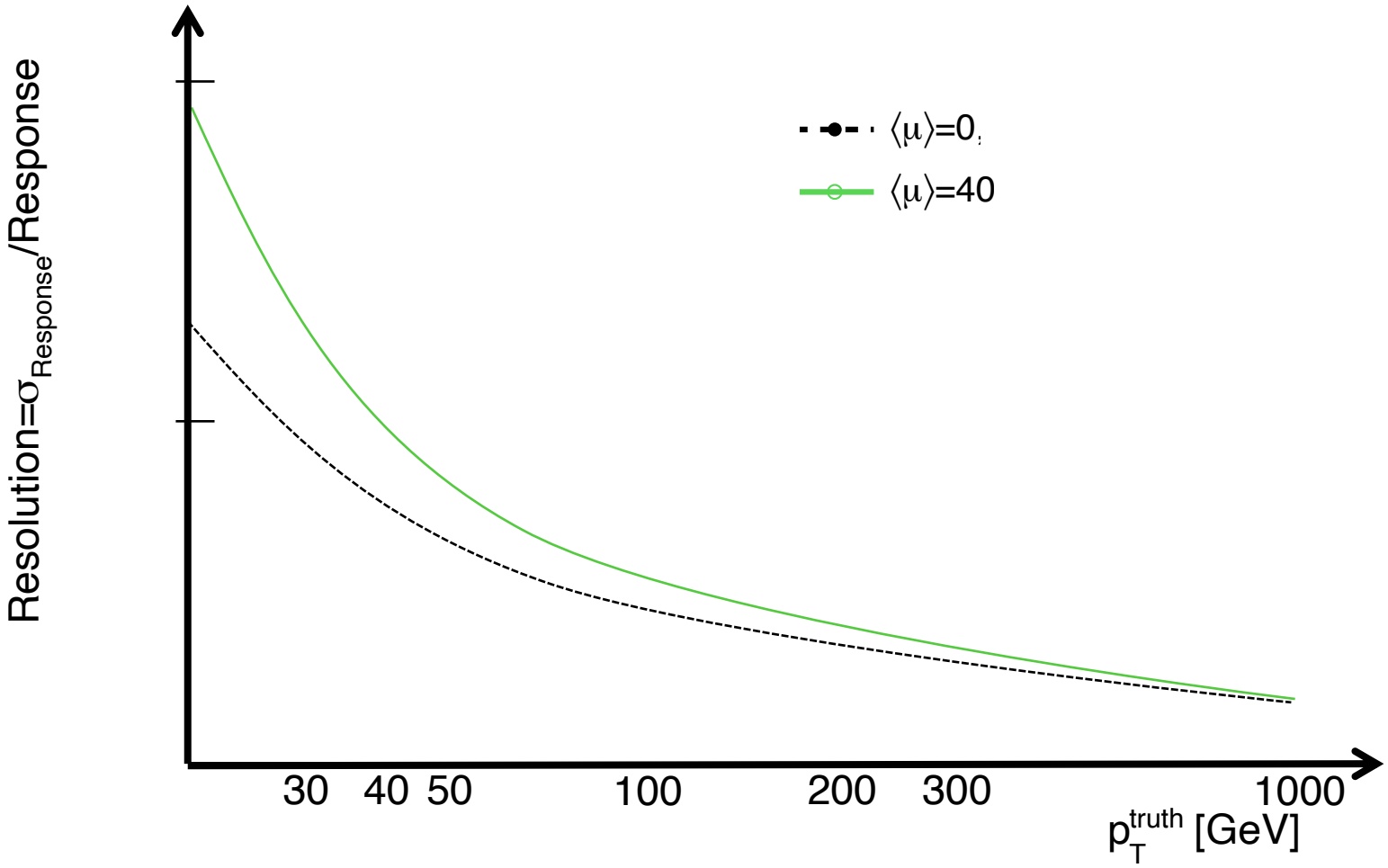
Jets in Pile-up



Multiple interactions from pile-up

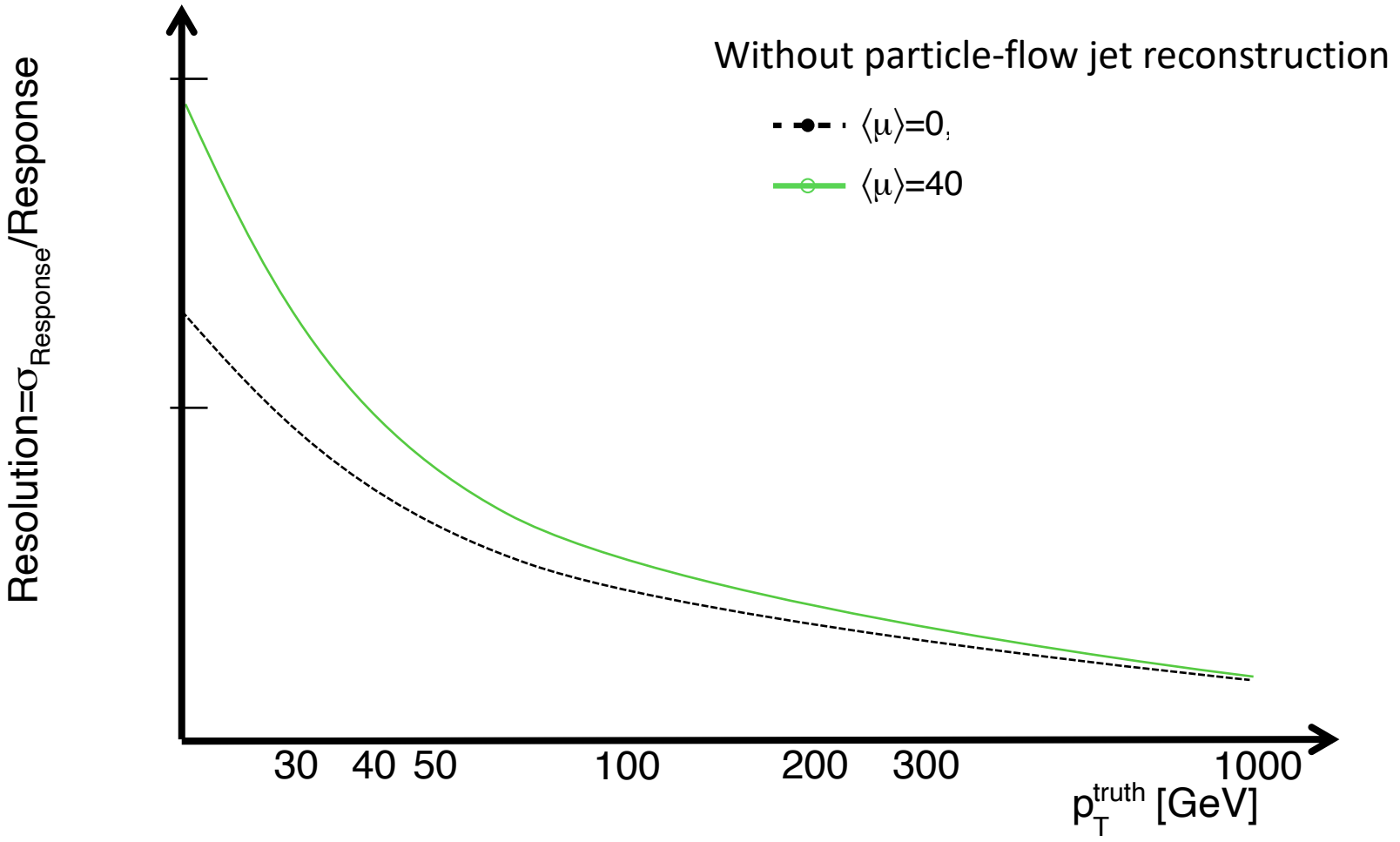
- ✓ Requirement that particles originate from the primary vertex.

Momentum resolution



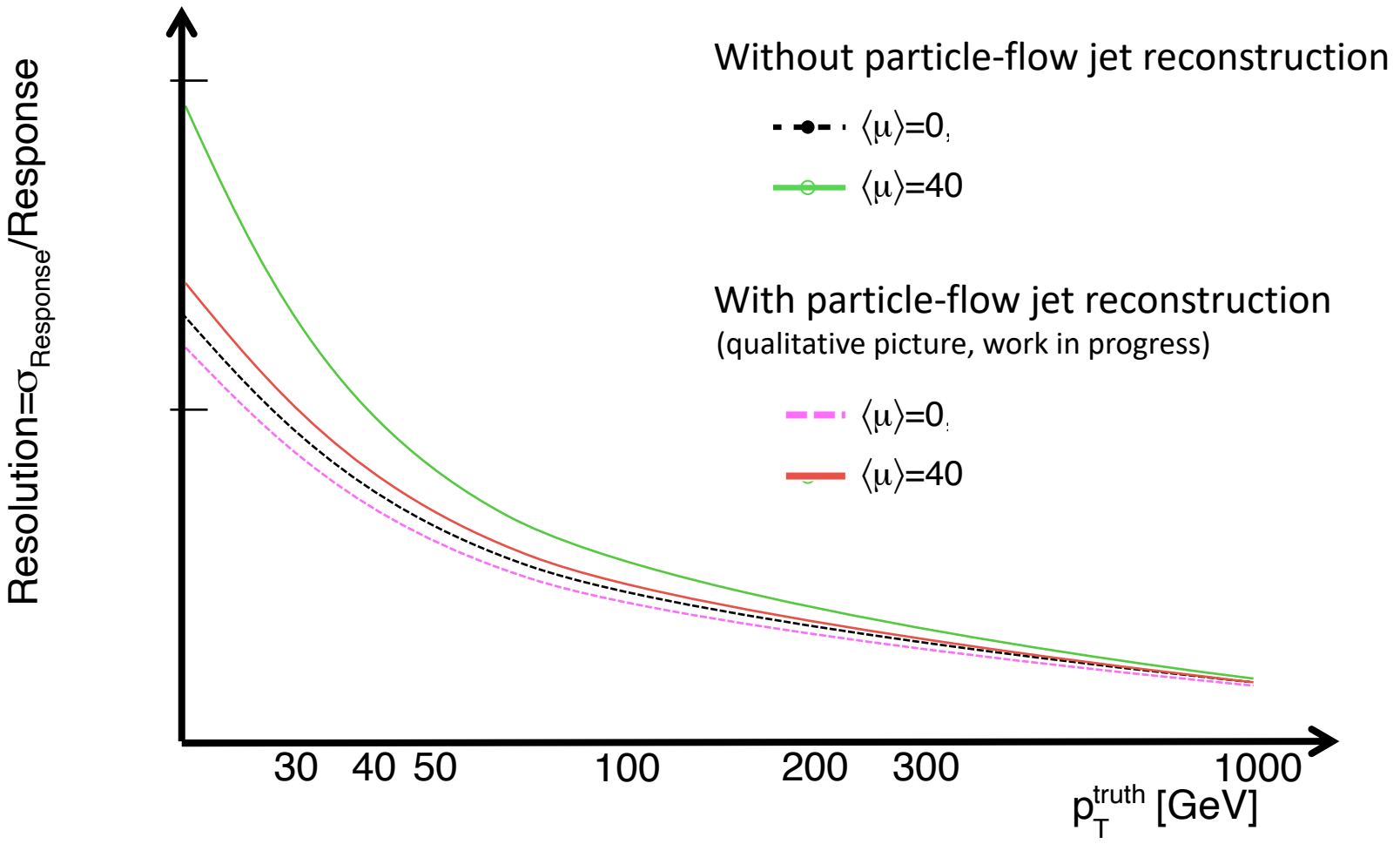
Resolution: the quality with which we measure the jet momentum.

Momentum resolution



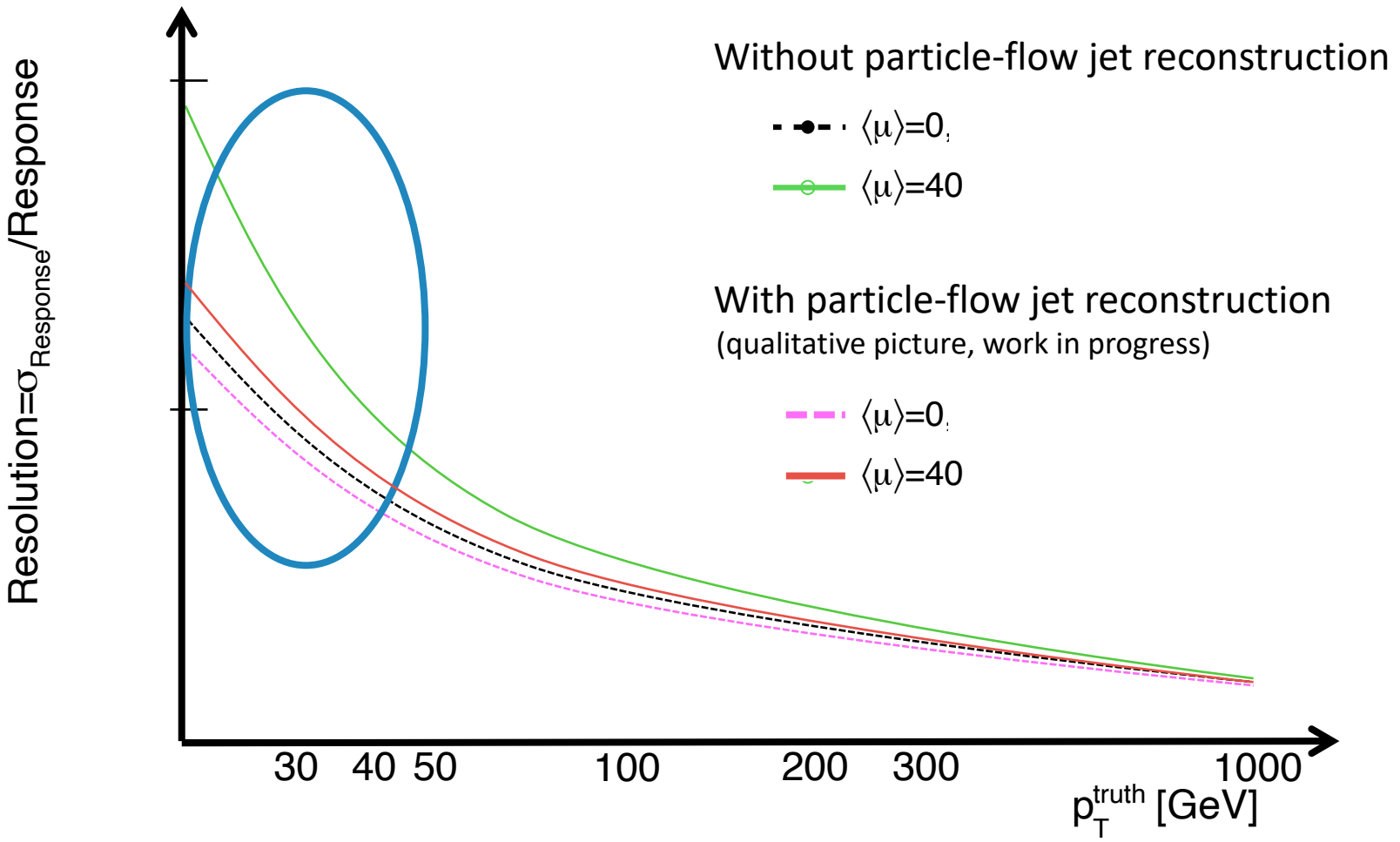
Resolution: the quality with which we measure the jet momentum.

Momentum resolution



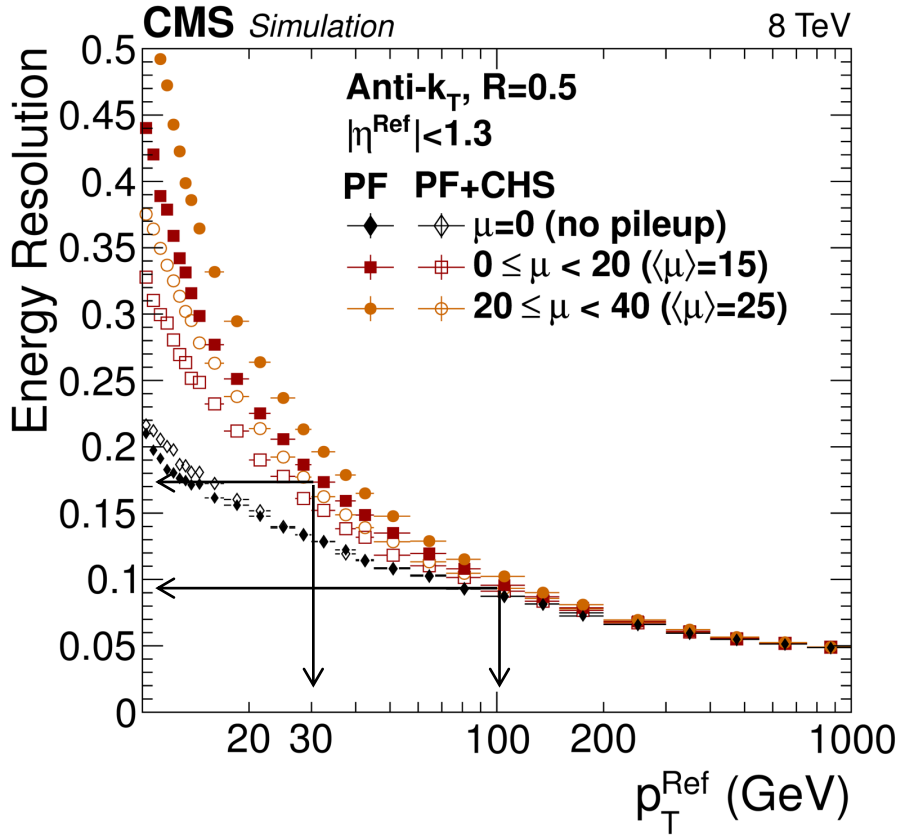
Resolution: the quality with which we measure the jet momentum.

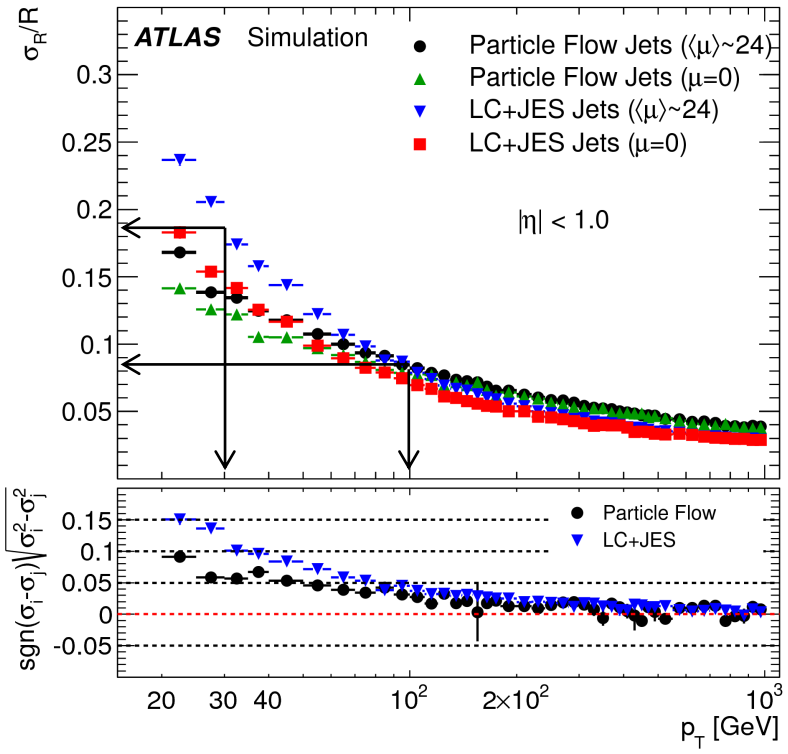
Momentum resolution



Significant improvement for low-pT jets. Similar for MET.

A COMPARISON

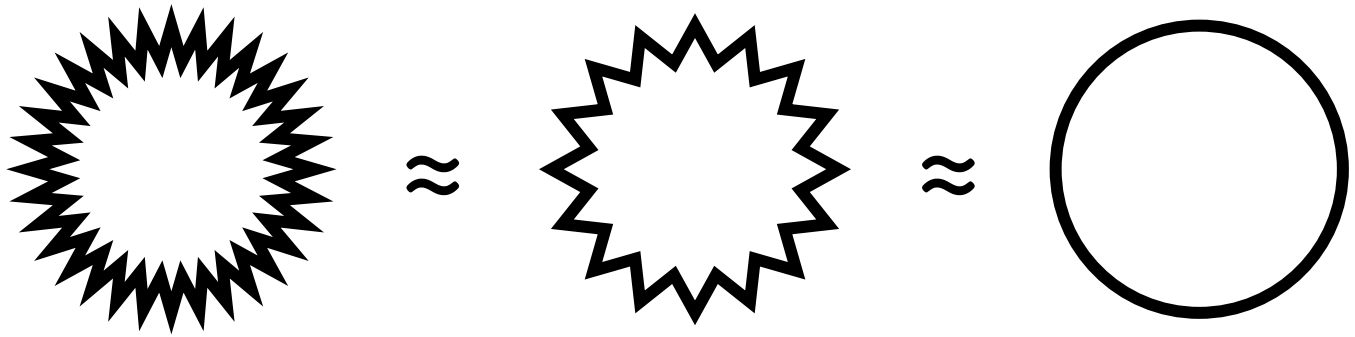




- PF jets (CMS) and calo jets (ATLAS) have similar performance.
- **Particle reconstruction always needs to be optimized depending on the detector technologies and experimental requirements.**

Online reconstruction

Objective: Trigger ("online") reconstruction same as "offline".
Problem: Time. Trigger decision needs to be taken fast.
Solution: Simplification.
Challenge: Clever simplification = good performance.

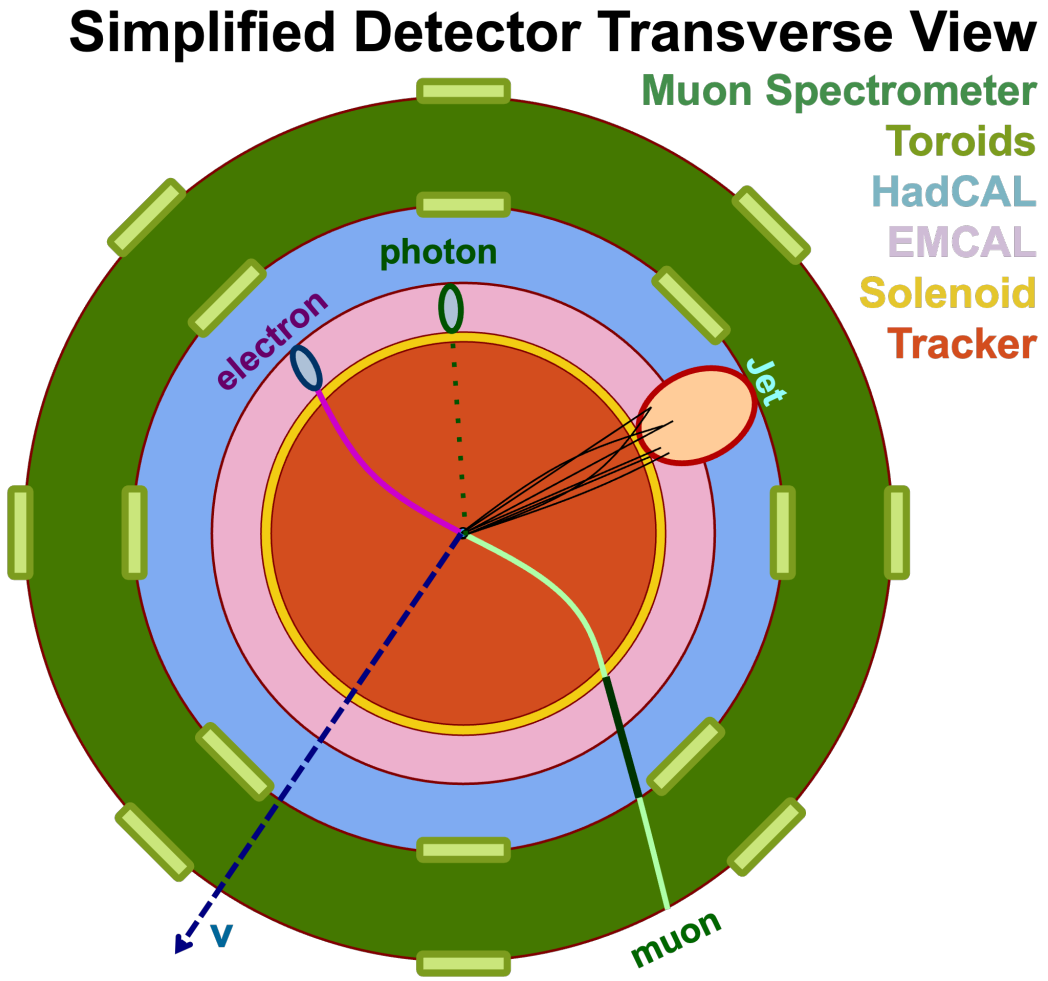


E.g. track reconstruction in regions of interest and simplified MET calculation.

Reconstructing particles

	I	II	III	
Quarks	2.4 MeV u	1.3 GeV c	170 GeV t	0 γ
	4.8 MeV d	104 MeV s	4.2 GeV b	0 g
Leptons	<2 eV ν_e	<2 eV ν_μ	<2 eV ν_τ	91 GeV Z
	0.5 MeV e	16 MeV μ	1.8 GeV τ	80 GeV W
				126 GeV H

Bosons

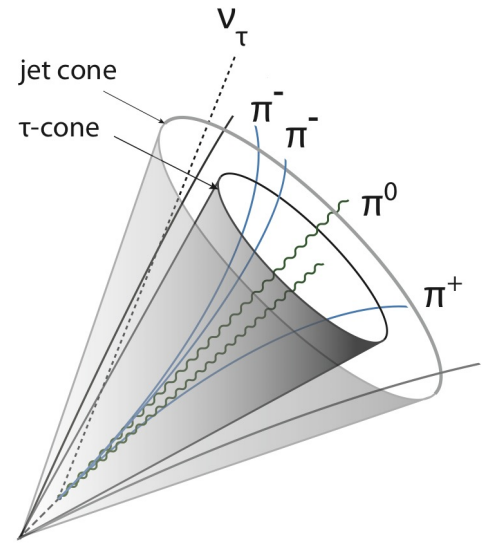


TAUS

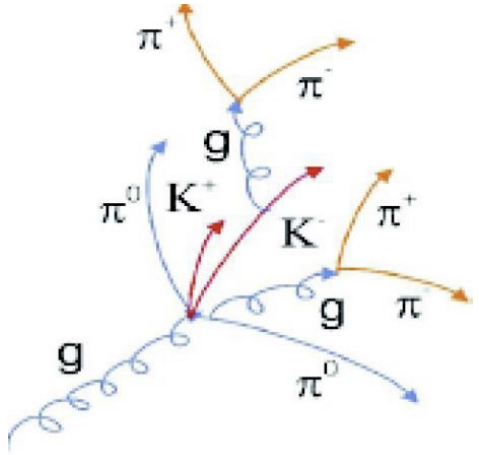
Tau Decay Mode			B.R.
Leptonic		$\tau^{\pm} \rightarrow e^{\pm} + \nu + \nu$	17.8%
		$\tau^{\pm} \rightarrow \mu^{\pm} + \nu + \nu$	17.4%
Hadronic	1-prong	$\tau^{\pm} \rightarrow \pi^{\pm} + \nu$	11%
		$\tau^{\pm} \rightarrow \pi^{\pm} + \nu + n\pi^0$	35%
	3-prong	$\tau^{\pm} \rightarrow 3\pi^{\pm} + \nu$	9%
		$\tau^{\pm} \rightarrow 3\pi^{\pm} + \nu + n\pi^0$	5%
Other			~5%

- Hadronic tau reconstruction extremely challenging
- Using **multi-variate (machine learning)** techniques based on track multiplicity and shower shapes

A tau jet (signal)…

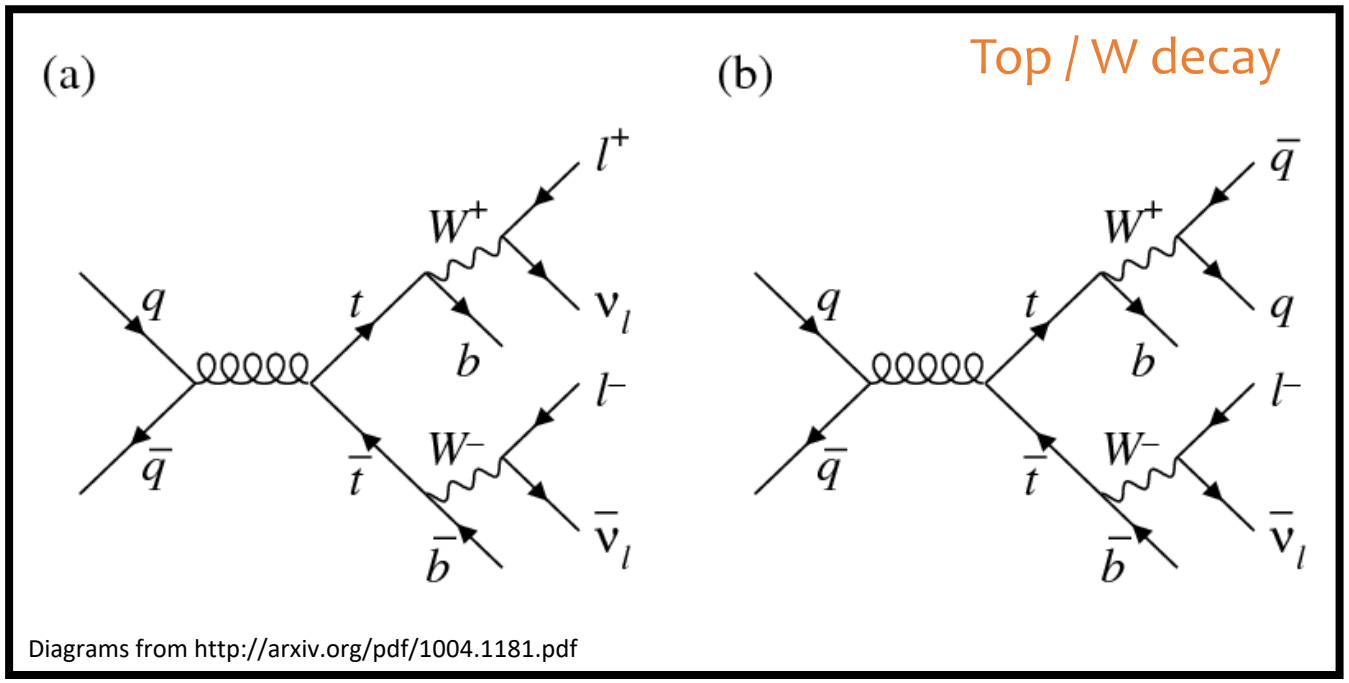


…vs. a QCD jet (background)



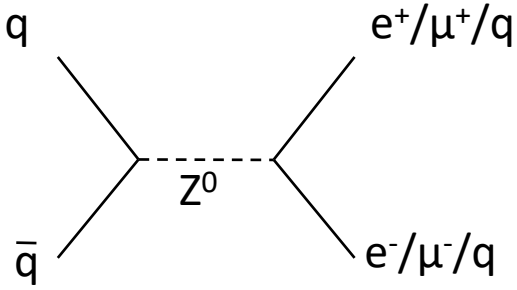
Top, W, Z

Top / W decay

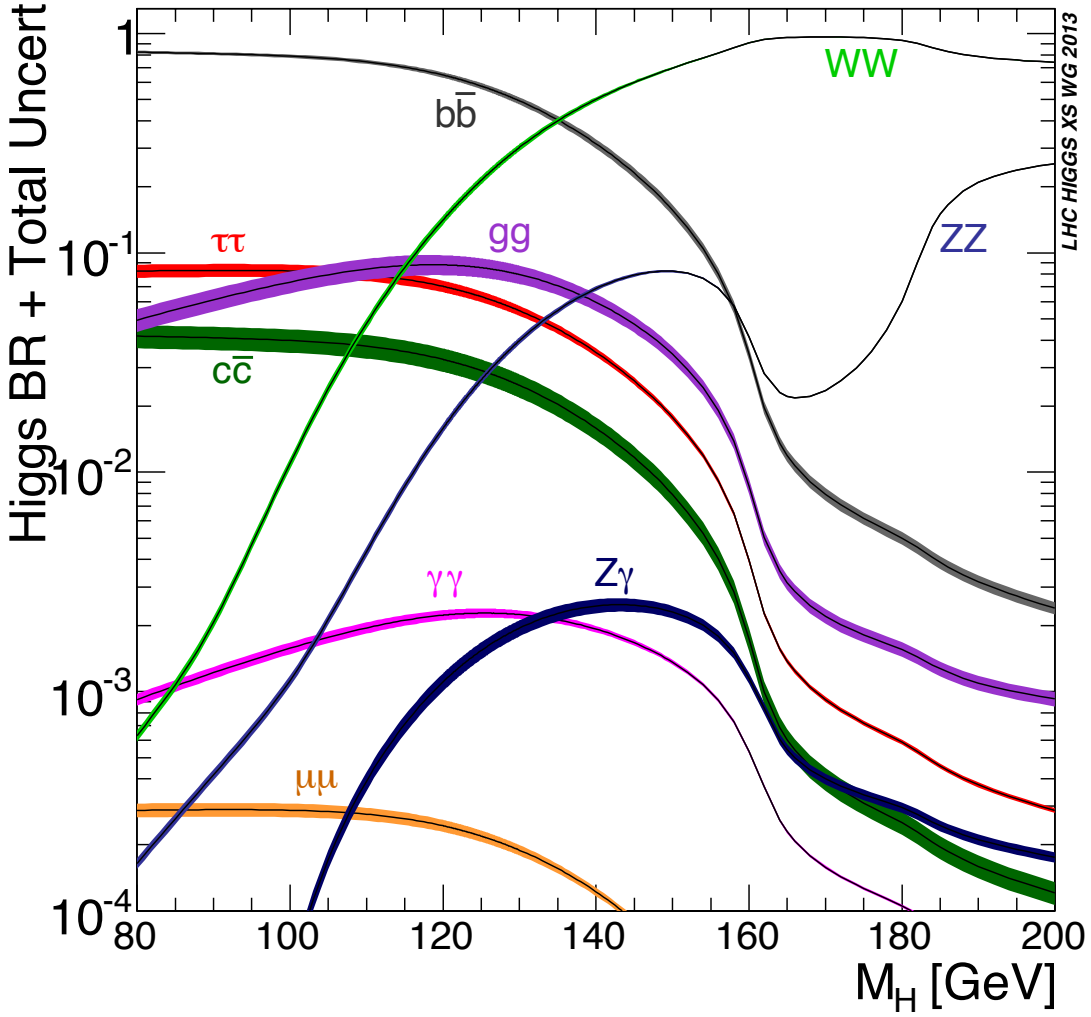


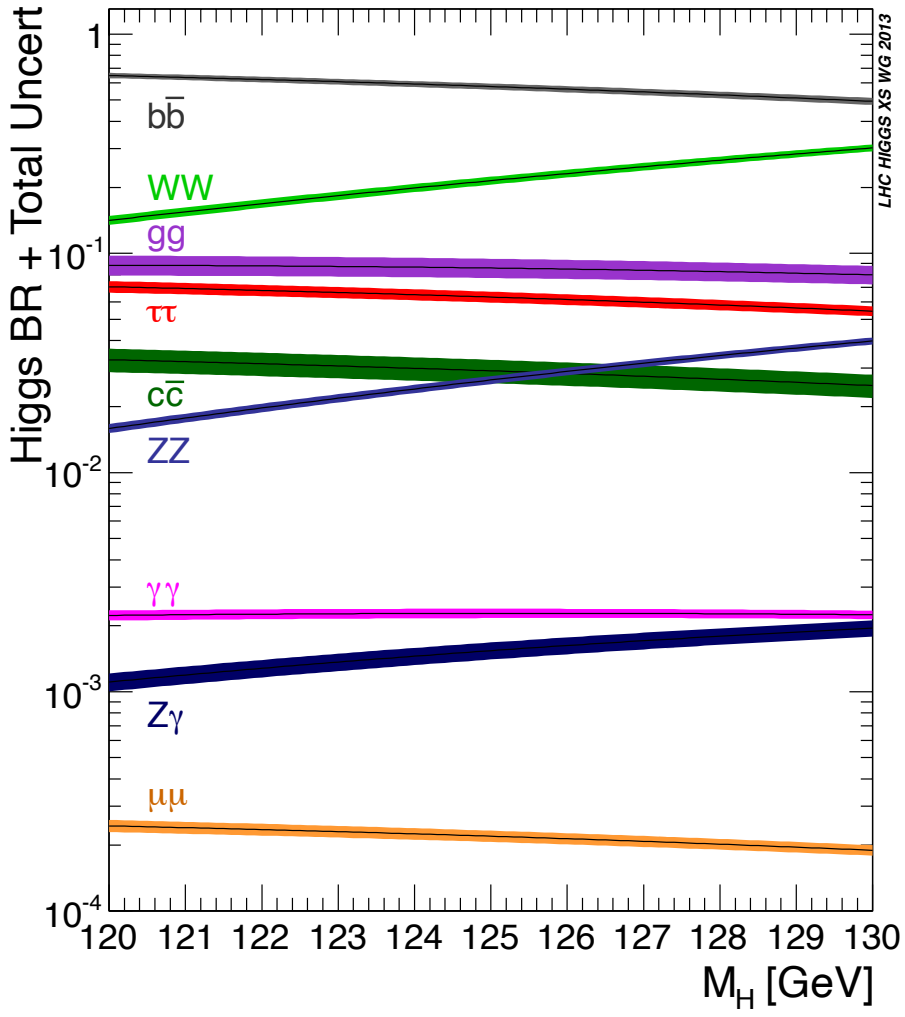
Diagrams from http://arxiv.org/pdf/1004.1181.pdf

Z decay



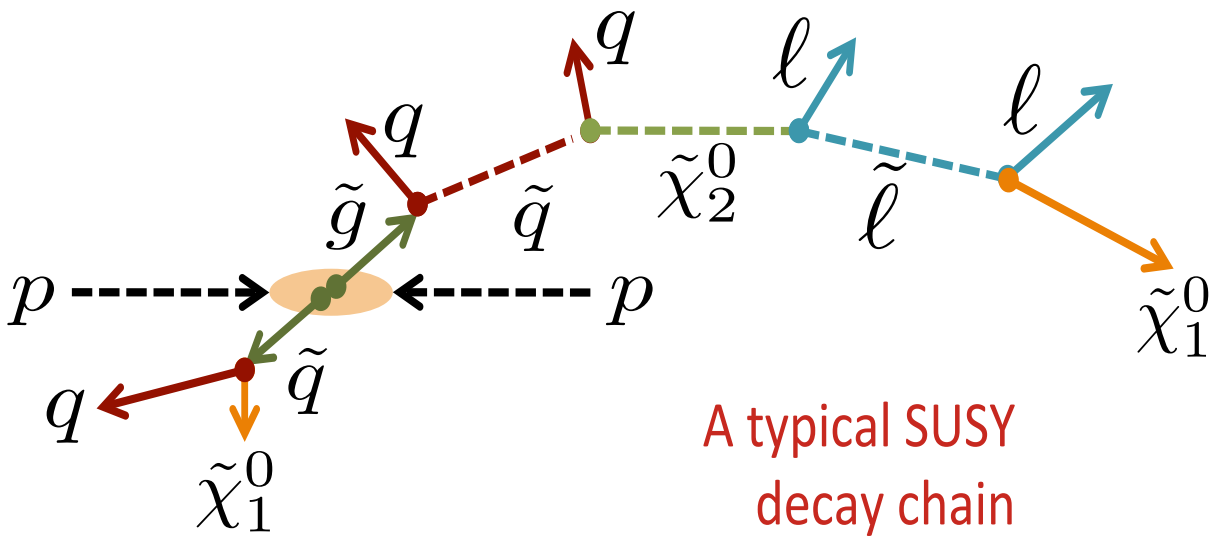
AND THE HIGGS !





How about new particles?

- These decay to Standard Model particles or create ME_T
- E.g.

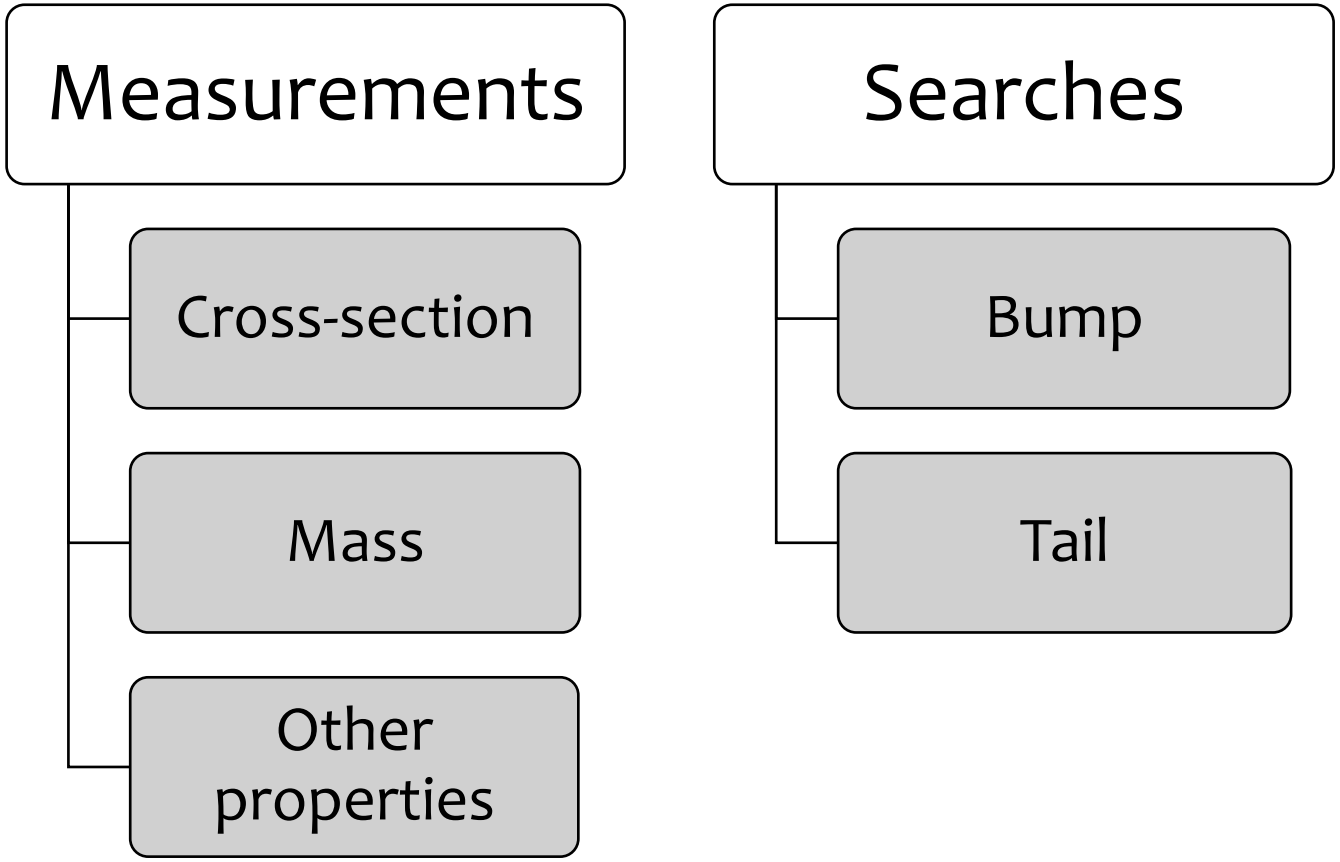


A typical SUSY decay chain

LHC Physics

An analysis Step-by-Step

Physics analyses



Physics analyses

"Systematic" uncertainties are introduced by inaccuracies in the methods used to perform the measurement.

Measurements

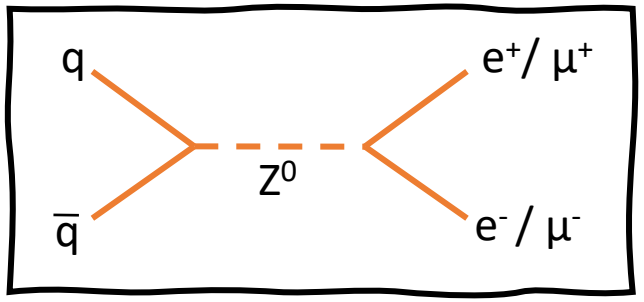
- Allow important tests of the consistency of the theory.
- Typically limited by systematic uncertainties.

Searches

- … For new particles.
- If no signal, set limits on some model.
- If signal, a potential discovery!
- More data typically improve a search.

Simple example:
Measuring the Z^0 cross-section at LHC

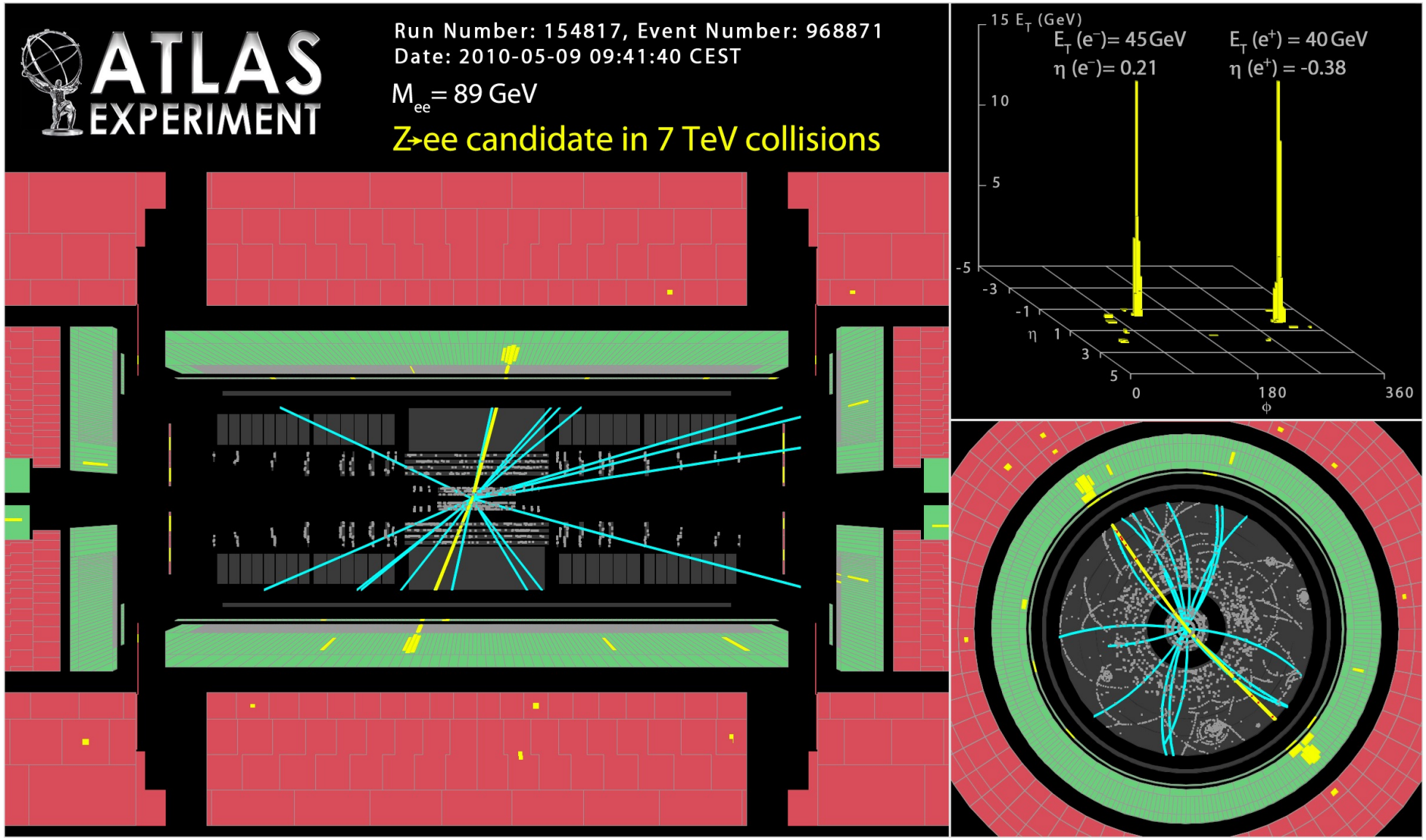
- **Z^0 boson decays to lepton or quark pairs**
  - We can reconstruct it in the e^+e^- or $\mu^+\mu^-$ decay modes



- Discovery and study of the Z^0 boson was a critical part of understanding the electroweak force.

- And now, at the LHC?
  - **Important test of theory**: does the measurement agree with the theoretical prediction at LHC collision energy?
  - **A standard candle** for studying reconstruction and deriving calibrations.
  - Can be used for luminosity determination!

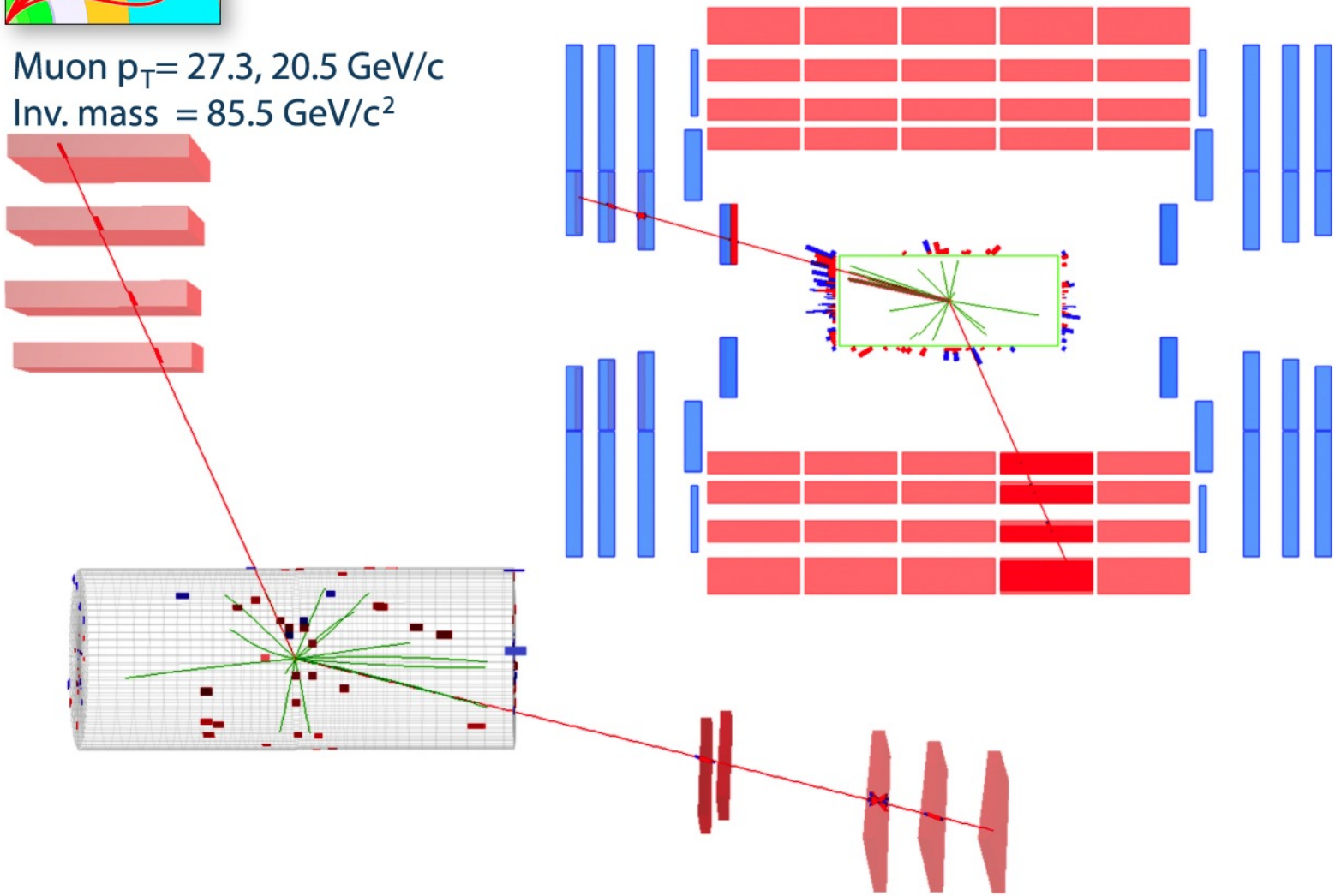
Measuring the Z^0 cross-section at LHC



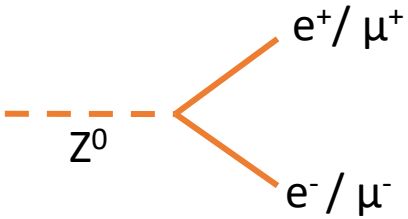
Measuring the Z^0 cross-section at LHC

CMS Experiment at LHC, CERN
Run 136087 Event 39967482
Lumi section: 314
Mon May 24 2010, 15:31:58 CEST

Muon p_T= 27.3, 20.5 GeV/c
Inv. mass = 85.5 GeV/c^2



Reconstructing Z^0's

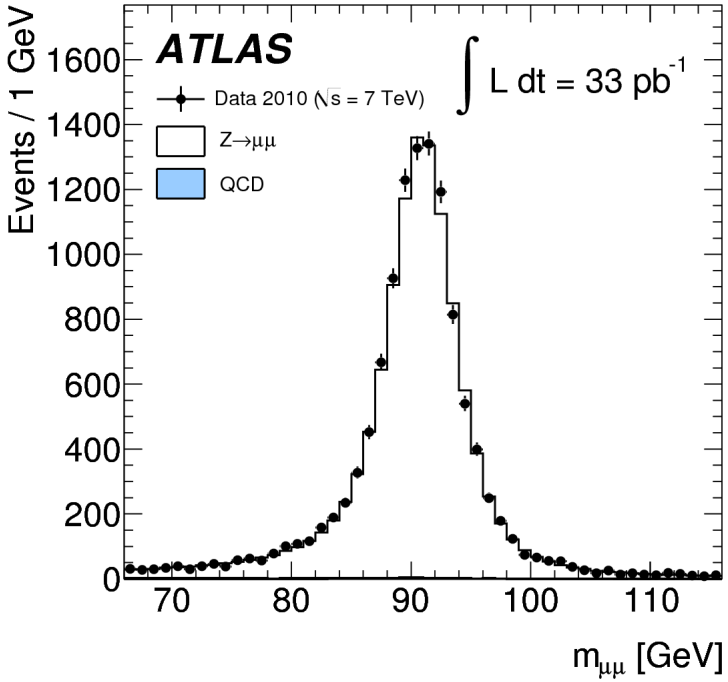


Step-1: Identify the observable of interest

- Identify Z decays using the invariant mass of the 2 leptons
  $M^2 = (L_1+L_2)^2$ where $L_i = (E_i, \underline{\boldsymbol{p}}_i)$ = 4-vector for lepton i
- Under assumption that lepton is massless compared to mass of Z^0
  => $M^2 = 2\,E_1\,E_2\,(1-\cos\theta_{12})$ where θ_{12}= angle between the leptons

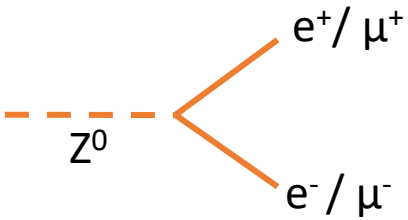
Step-2: Select Z^0 events with 'analysis cuts':

- Events with 2 high momentum electrons or muons
- Require the electrons or muons are of opposite charge
- With di-lepton mass close to the Z^0 mass
  (e.g. $70<m_{l+l-}<110$ GeV)



Very little background in Z^0 mass region!

Reconstructing Z^0's



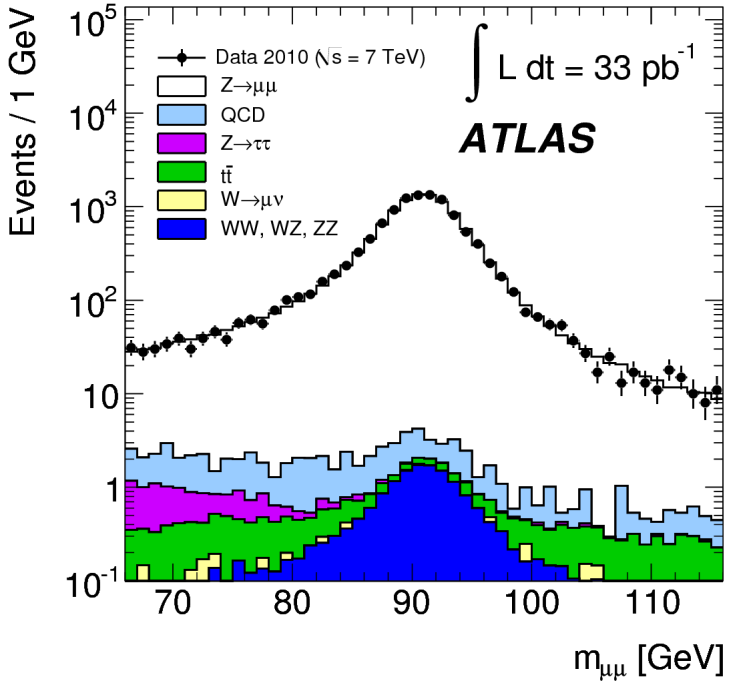
Step-1: Identify the observable of interest

- Identify Z decays using the invariant mass of the 2 leptons
  $M^2 = (L_1+L_2)^2$ where $L_i = (E_i, \underline{\mathbf{p}}_i)$ = 4-vector for lepton i
- Under assumption that lepton is massless compared to mass of Z^0
  => $M^2 = 2\,E_1\,E_2\,(1-\cos\theta_{12})$ where θ_{12}= angle between the leptons

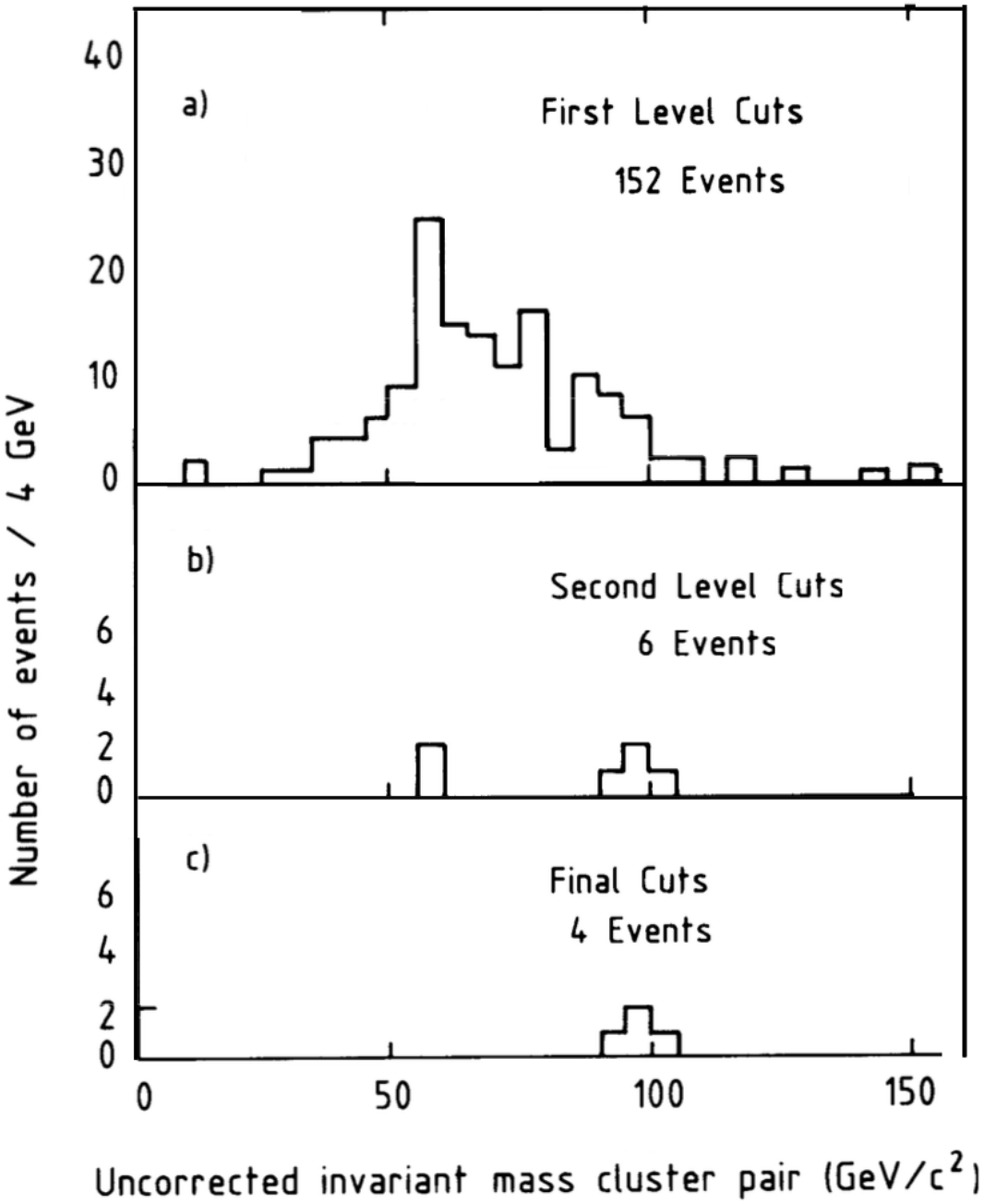
Step-2: Select Z^0 events with 'analysis cuts':

- Events with 2 high momentum electrons or muons
- Require the electrons or muons are of opposite charge
- With di-lepton mass close to the Z^0 mass
  (e.g. $70<m_{l+l-}<110$ GeV)

Very little background in Z^0 mass region!



A step back in time



Z->ee in UA1

Two EM clusters with E_T>25GeV.

As above plus a track with p_T>7GeV pointing to the cluster.
Hadronic and track isolation requirements applied.

A second cluster has also an isolated track.

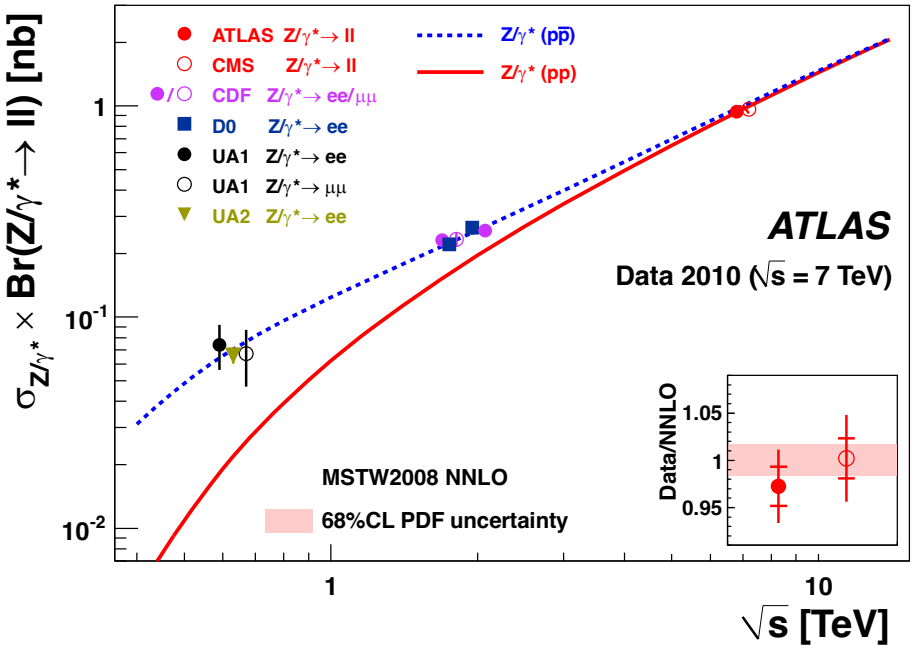
http://www.nobelprize.org/nobel_prizes/physics/laureates/1984/rubbia-lecture.pdf

Measuring Z^0 cross-section

Theoretically

Cross-section calculated for:

- Specific production mechanism (pp, pp, e^+e^-)
- Centre-of-Mass of the collisions (7, 8, 13 TeV at LHC)

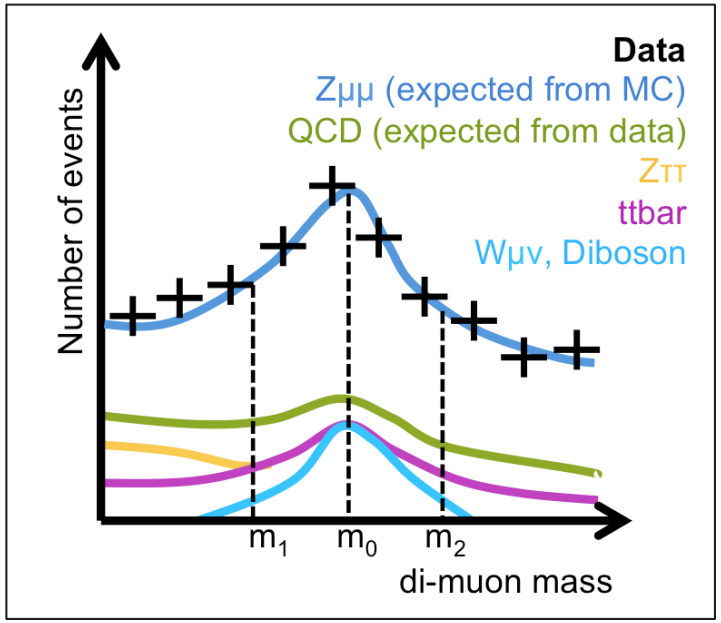


Experimentally

$$\sigma \cdot \mathrm{BR} = \frac{\mathrm{Number\ of\ events}}{\alpha \cdot \epsilon \cdot \mathrm{L}}$$

N of events: N of events on data – N of expected background events
α – acceptance: fraction of events passing selection requirements
ε – efficiency: reconstruction efficiency of relevant objects
L – luminosity

All numbers carry **uncertainties** – both **"statistical"** and **"systematic"**!

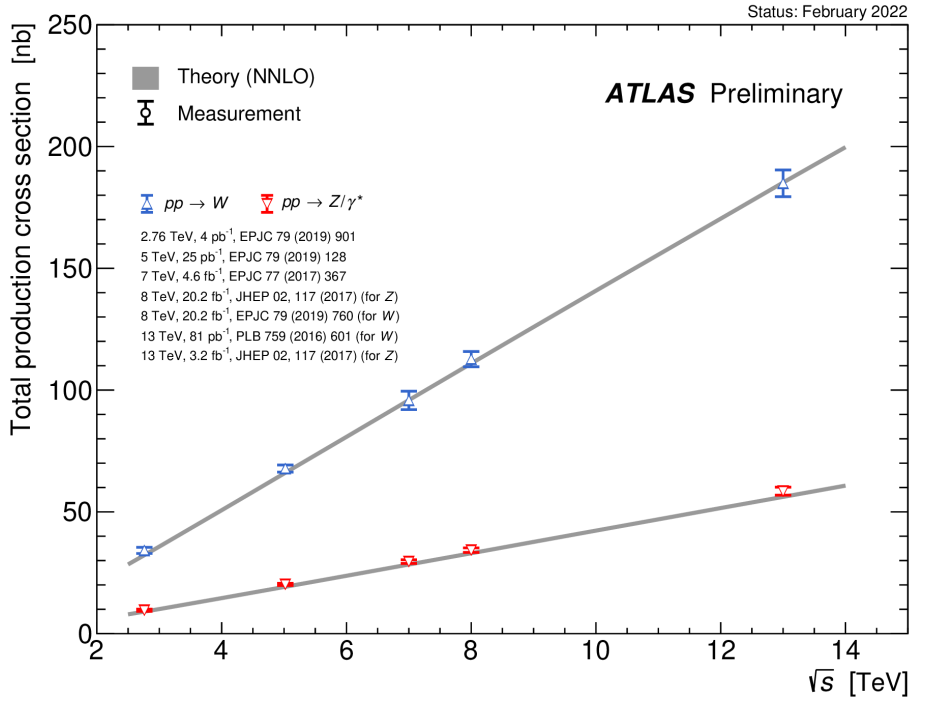


Measuring Z^0 cross-section

Theoretically

Cross-section calculated for:

- Specific production mechanism (pp, pp, e^+e^-)
- Centre-of-Mass of the collisions (7, 8, 13 TeV at LHC)

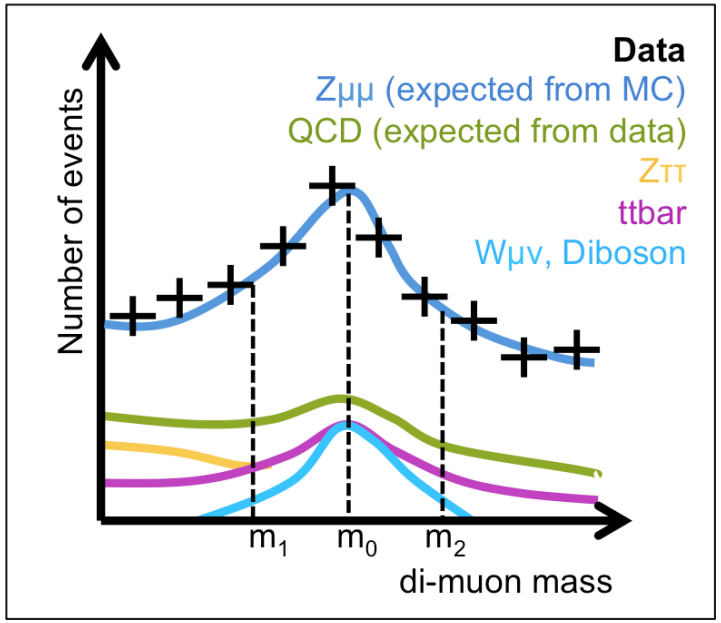


Experimentally

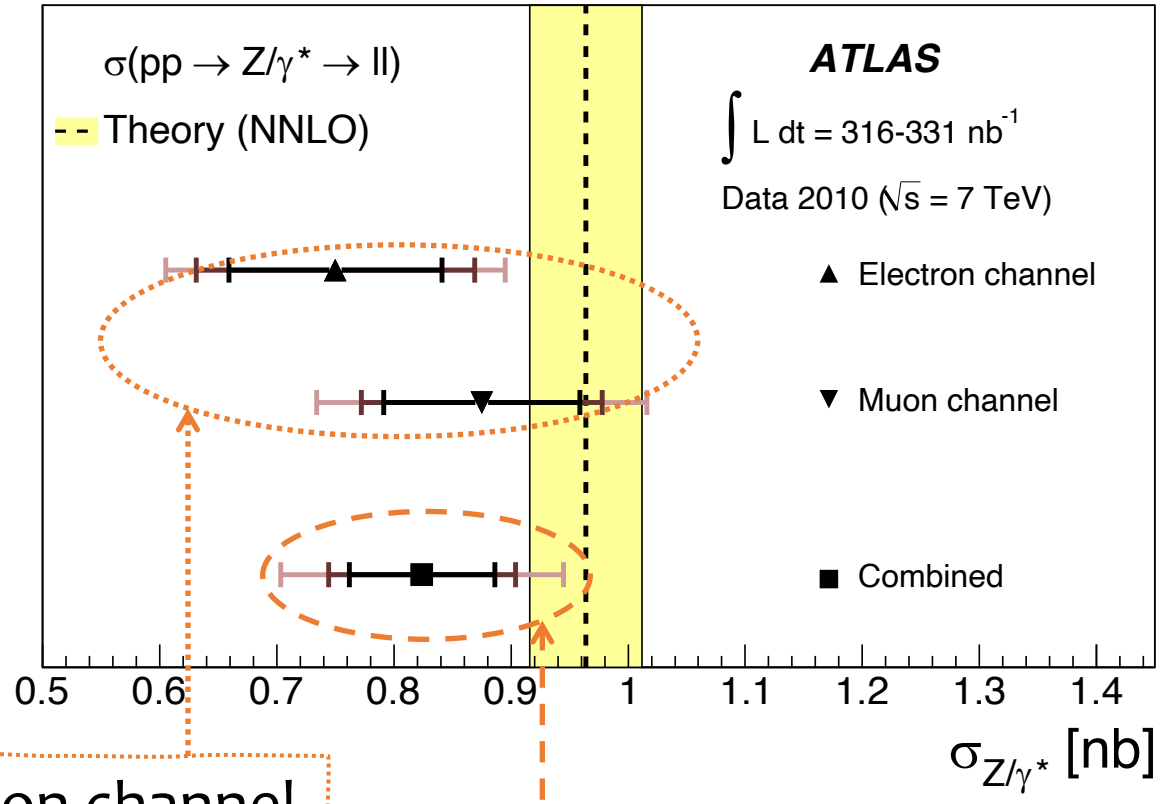
$$\sigma \cdot \mathrm{BR} = \frac{\mathrm{Number\ of\ events}}{\alpha \cdot \epsilon \cdot \mathrm{L}}$$

N of events: N of events on data – N of expected background events
α – acceptance: fraction of events passing selection requirements
ε – efficiency: reconstruction efficiency of relevant objects
L – luminosity

All numbers carry **uncertainties** – both **"statistical"** and **"systematic"**!



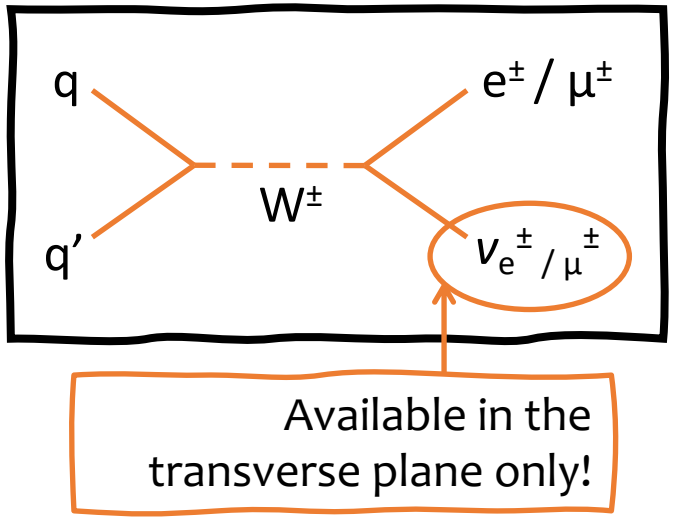
Measuring Z^0 cross-section



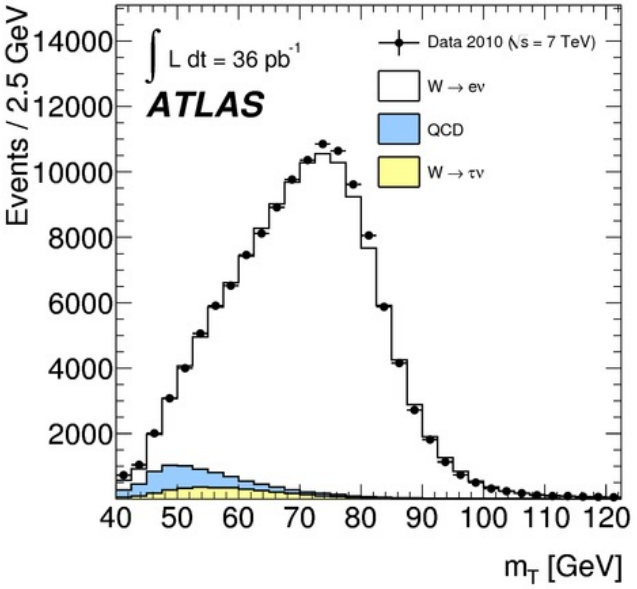
Electron and Muon channel agree within uncertainties

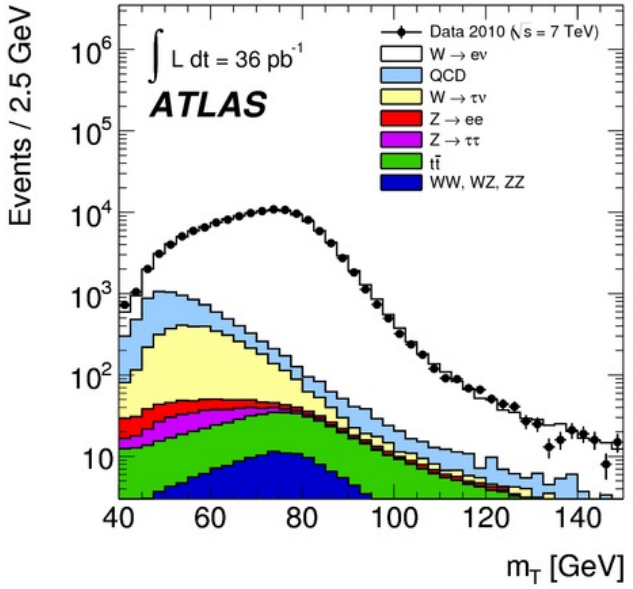
Measurement consistent with prediction within uncertainties

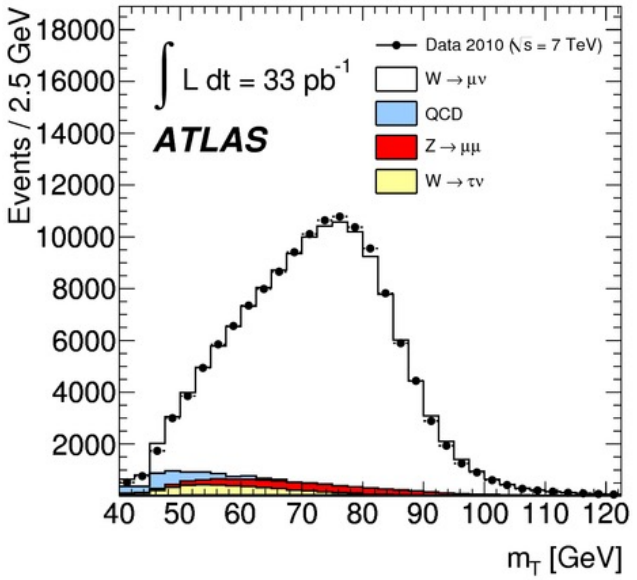
Measuring W cross-section

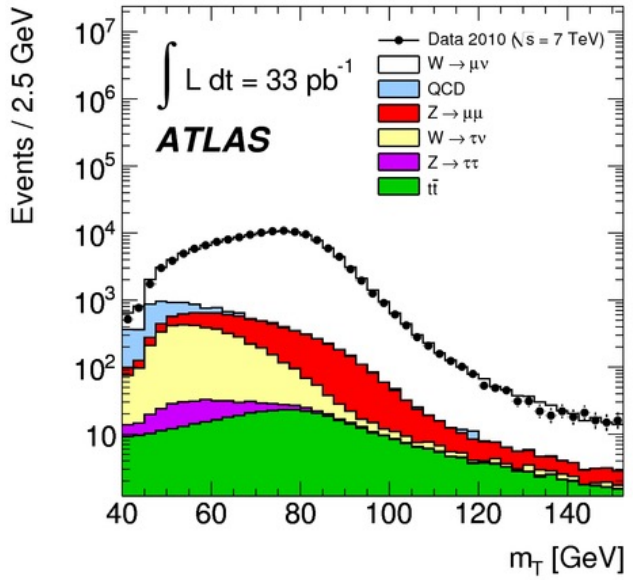


$$M_T^2 = 2\,E_{T1}\,E_{T2}\,(1-\cos\theta_{12})$$

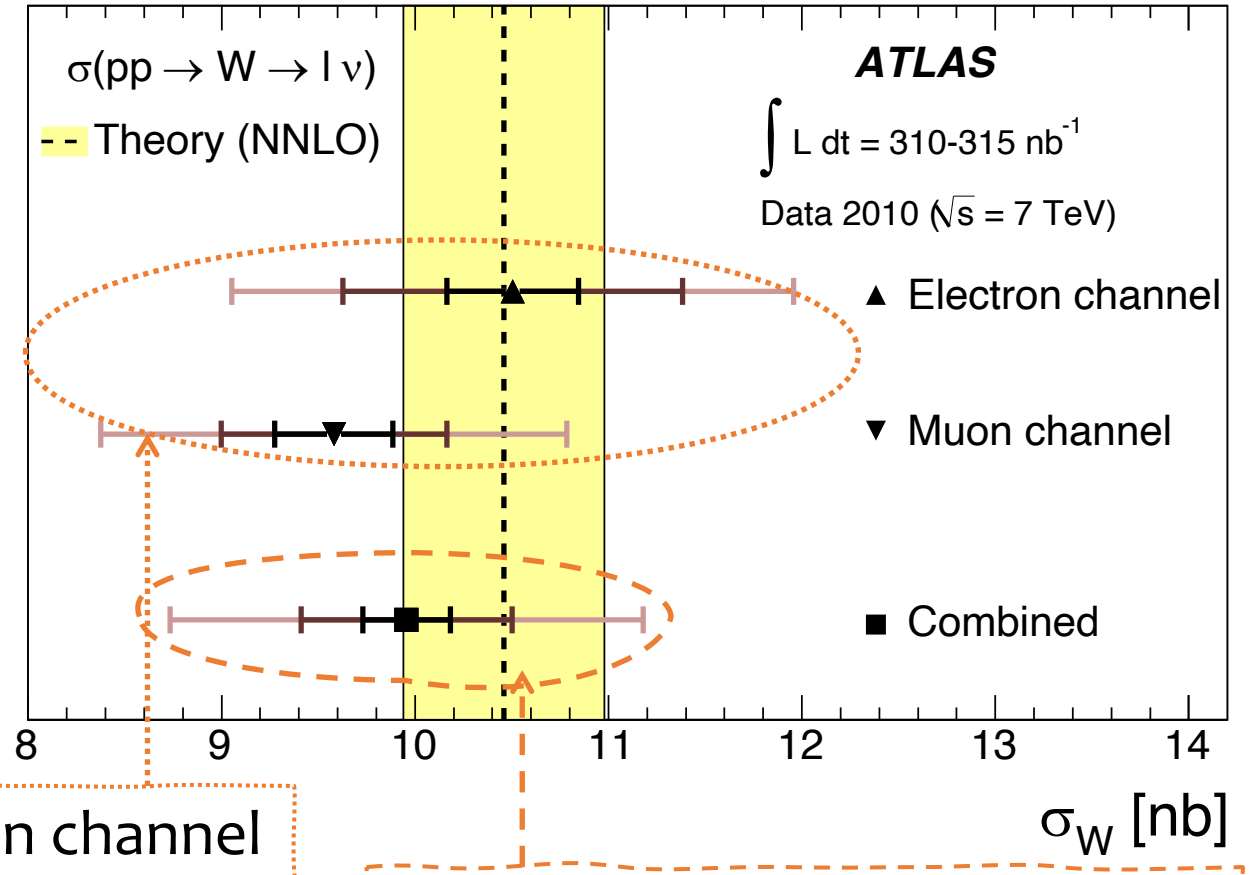








Measuring W cross-section

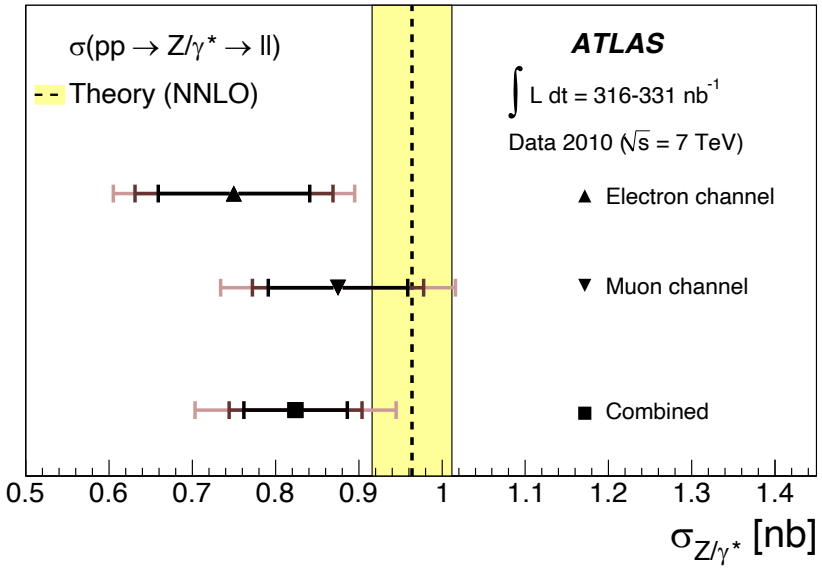


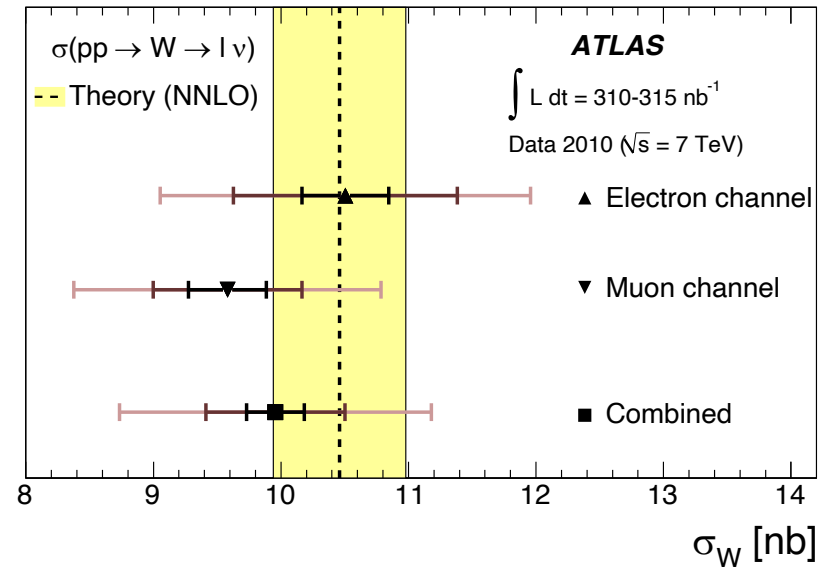
Electron and Muon channel agree within uncertainties

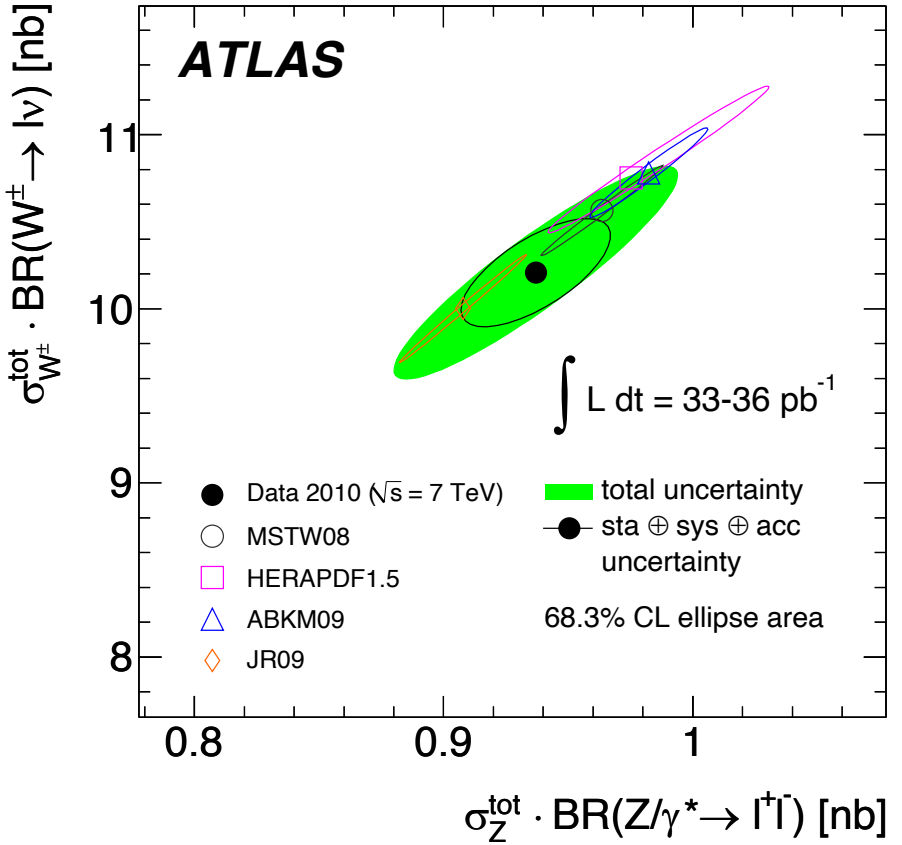
Measurement consistent with prediction within uncertainties

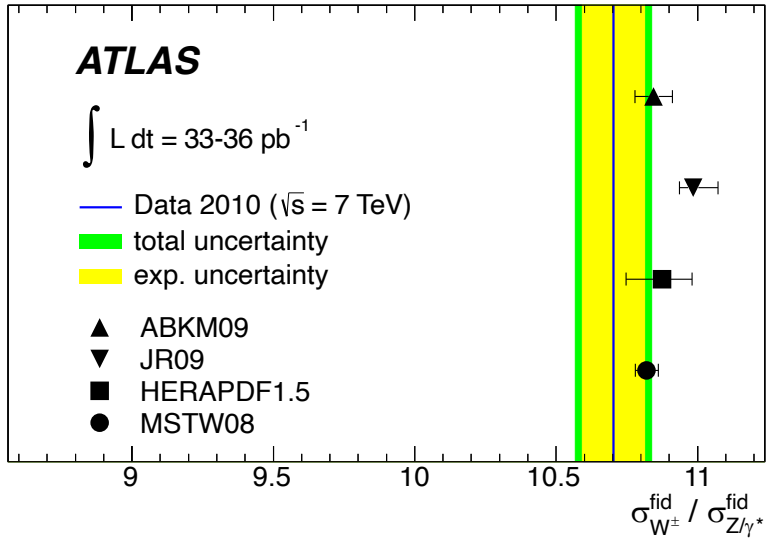
Measuring cross-sections

Ratios



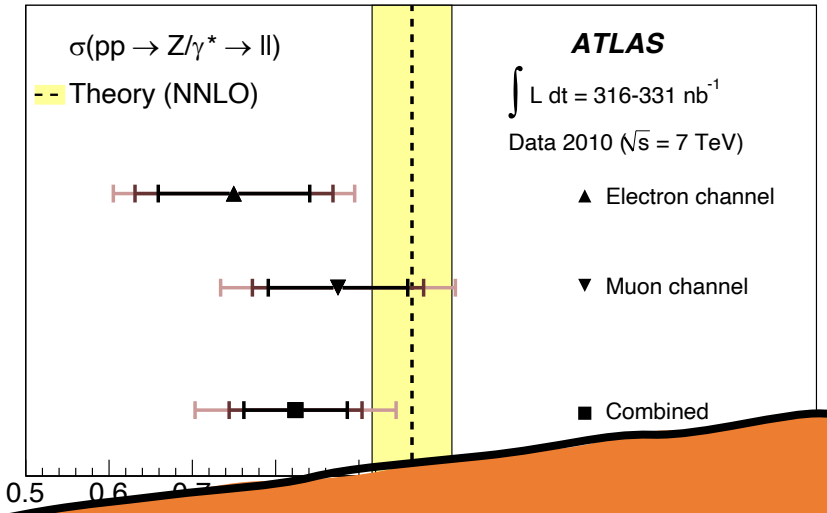


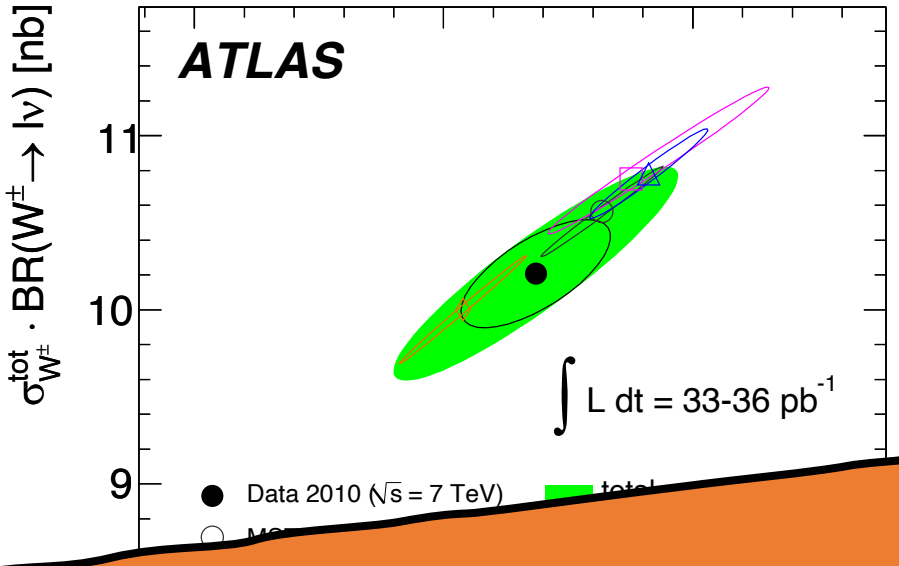




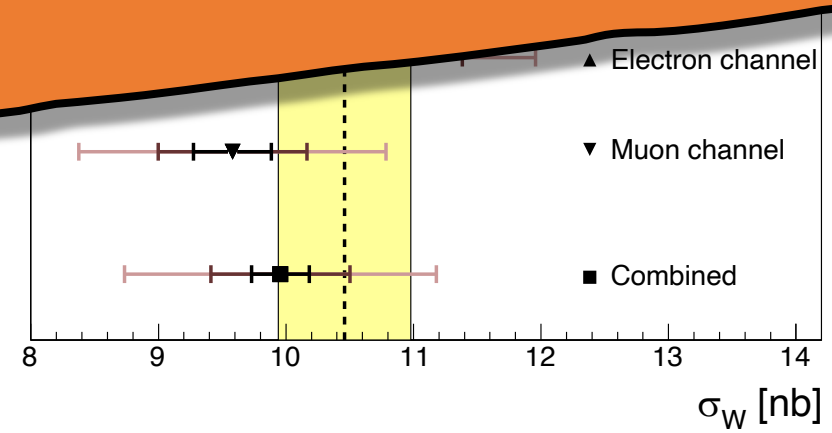
Measuring cross-sections

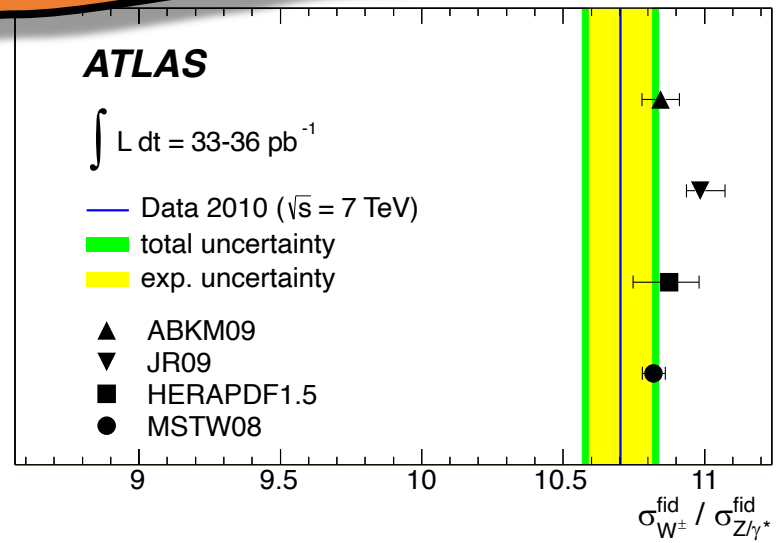
Ratios



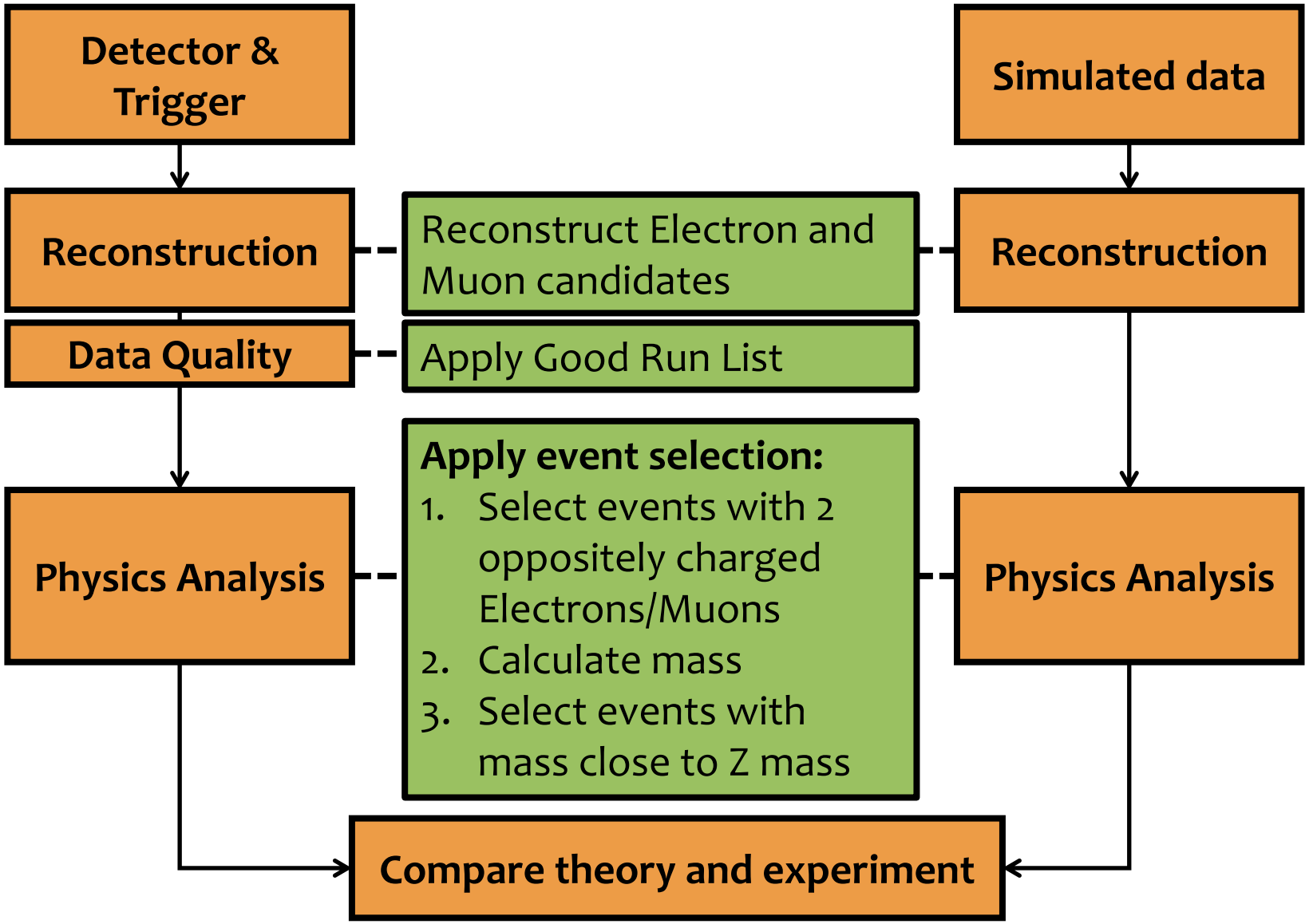


General aim:
find ways to reduce the uncertainties!

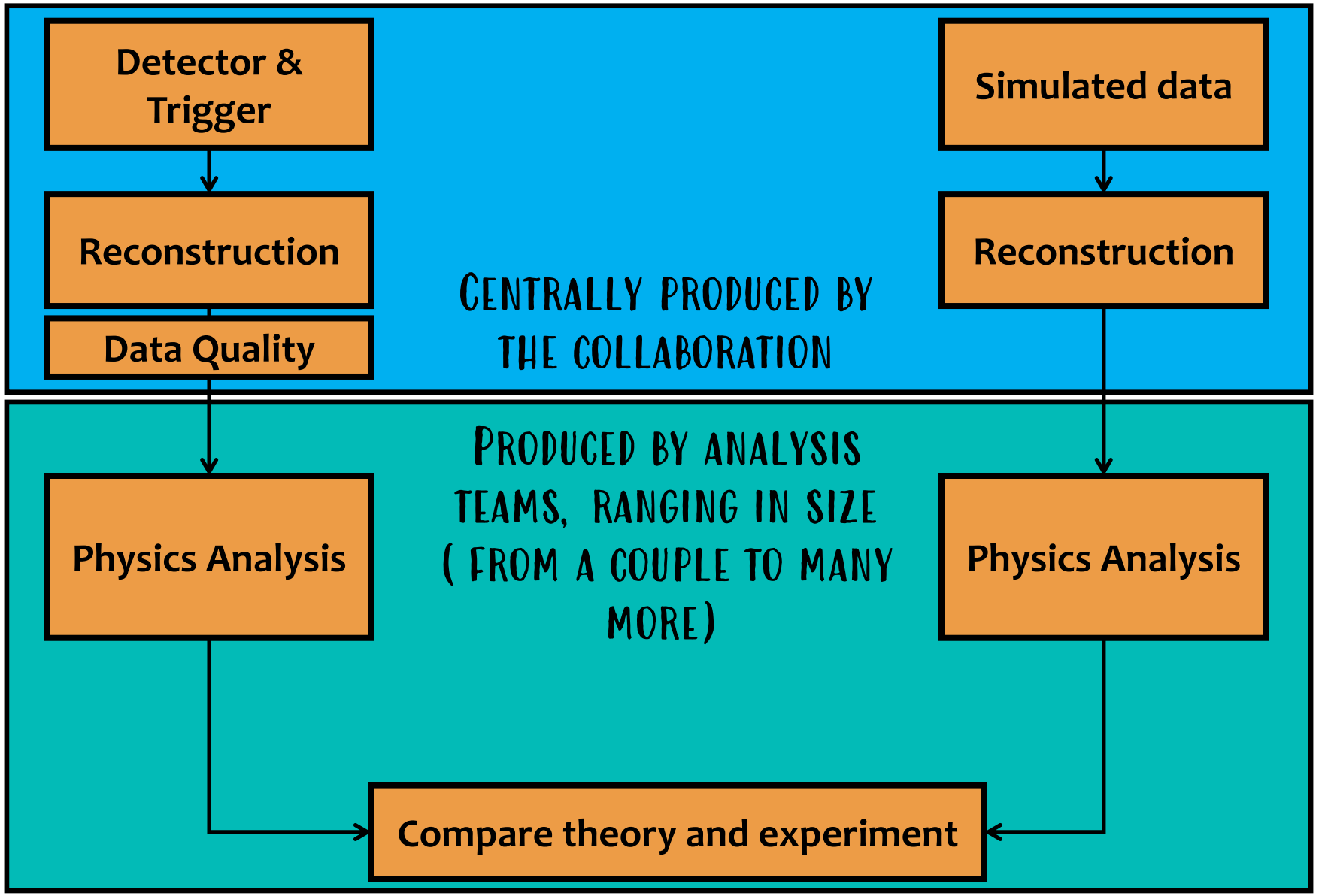


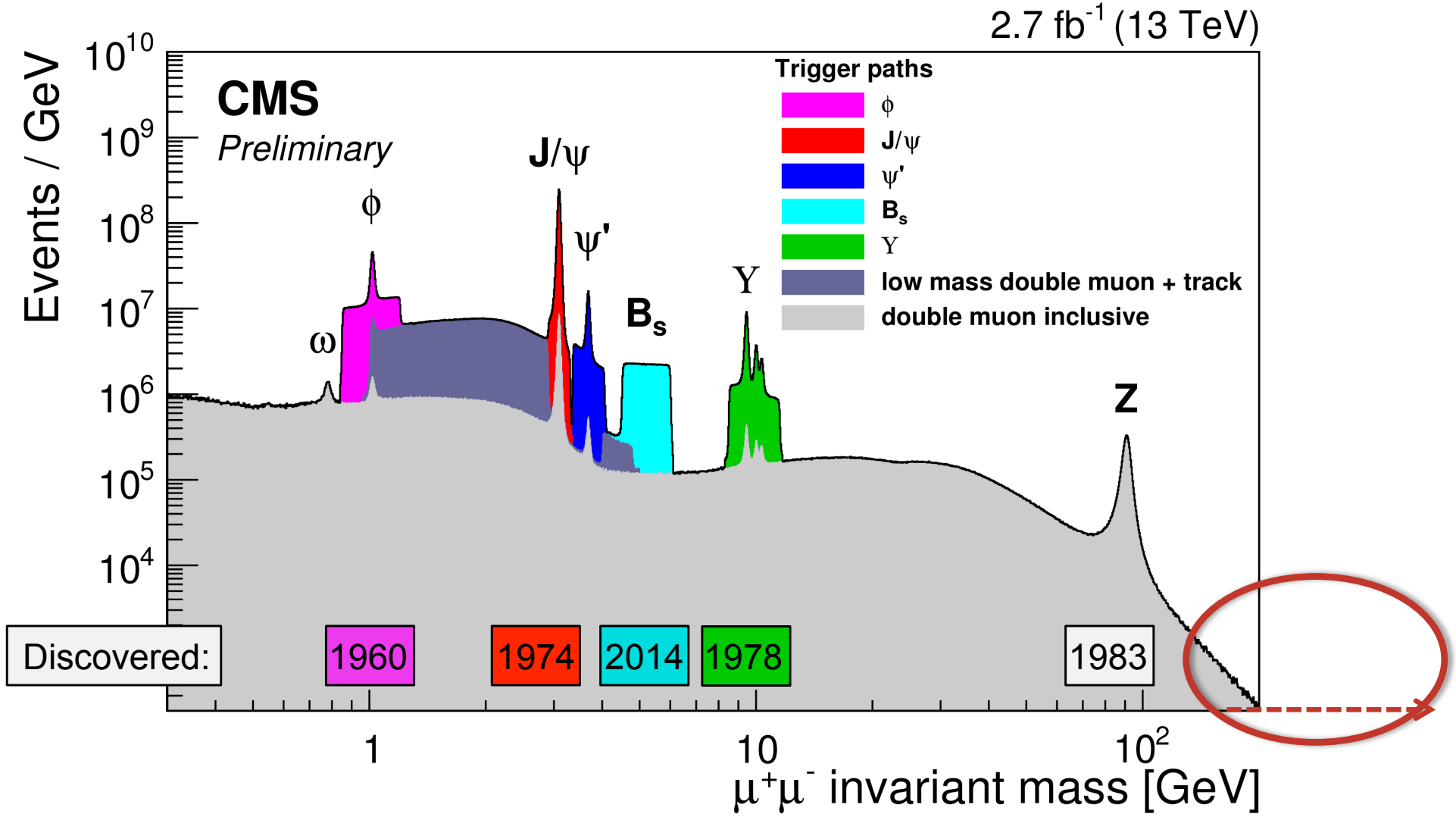


Analysis flow – e.g. cross-section measurement



Analysis flow – e.g. cross-section measurement

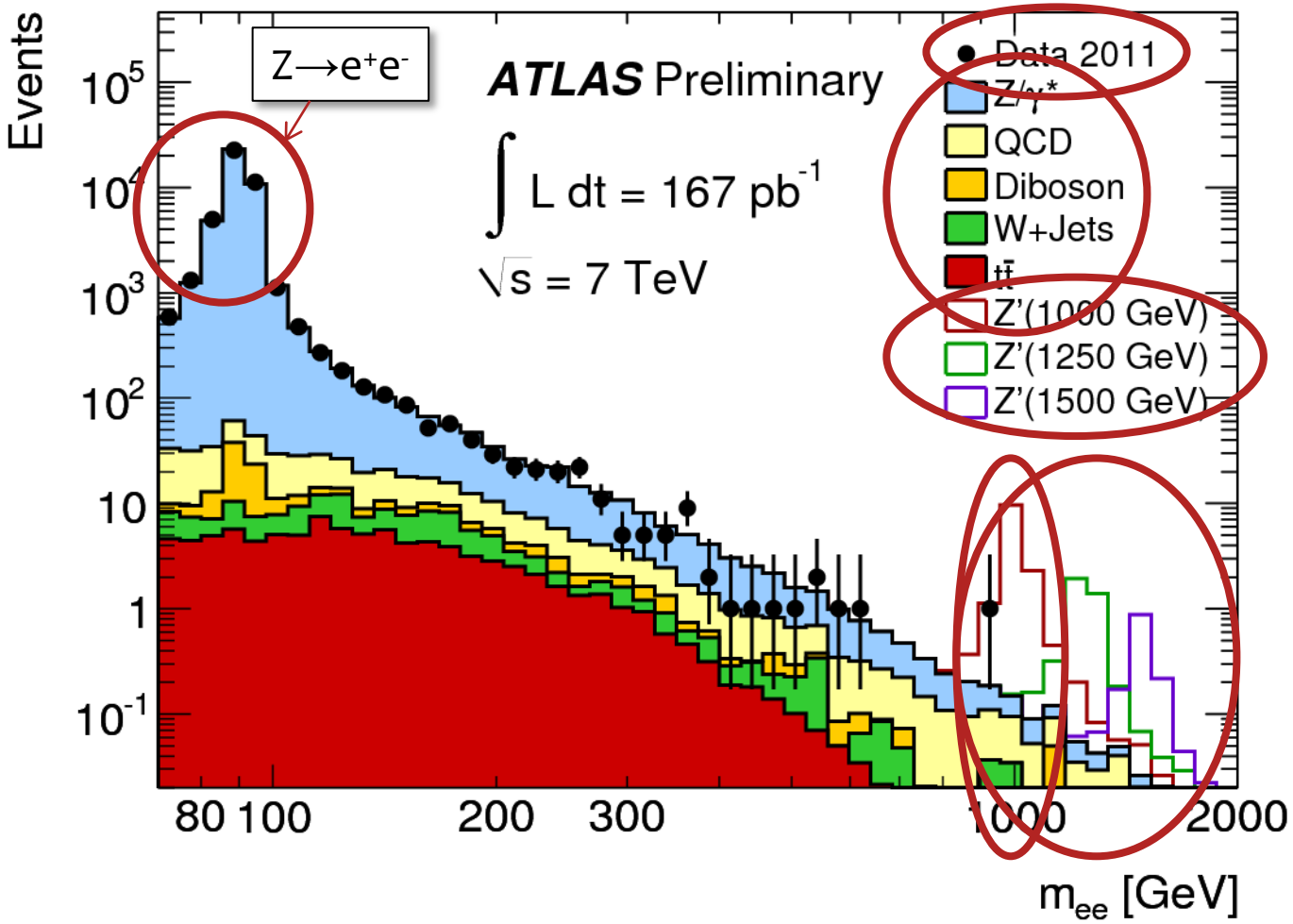




What's out there?

Simple Search example:
Search for a heavy Z'

- Like Z->ee but at higher mass



Select 2 electron candidates and plot their invariant mass for:

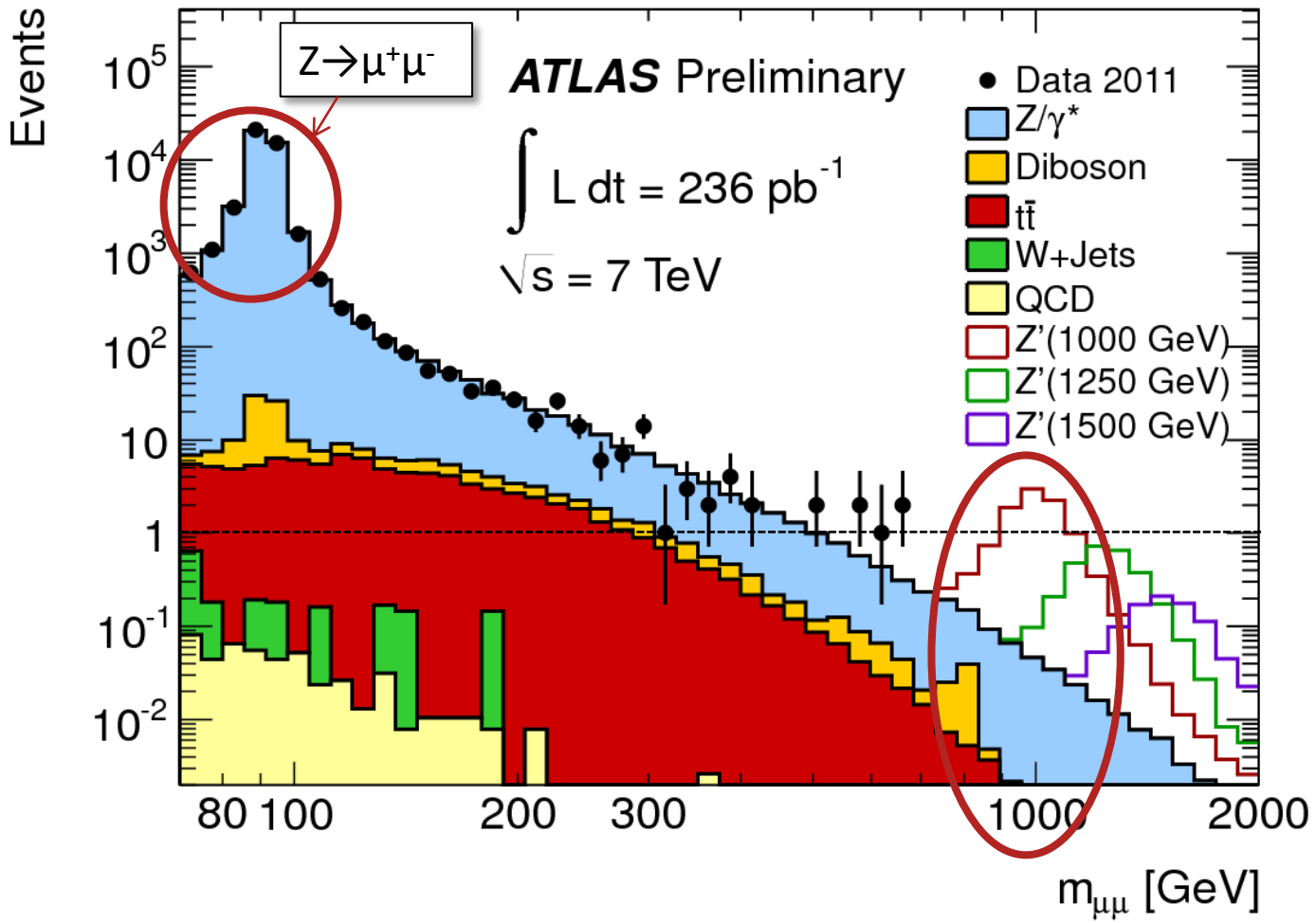
1. **Data**
2. **Simulated background events**
3. **Simulated signal with different masses**

Data inconsistent with a 1TeV Z'

Cross-section decreases with mass (higher the mass of the Z', the more data needed to discover it)

Simple Search example:
Search for a heavy Z'

- And similar for muons



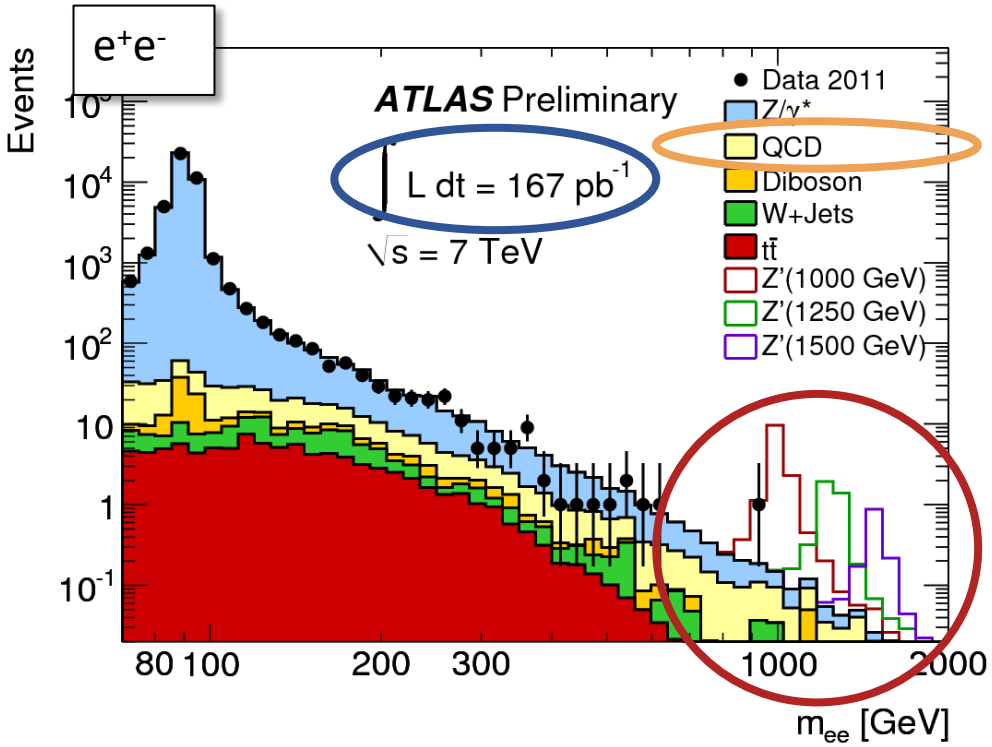
Select 2 electron candidates and plot their invariant mass for:

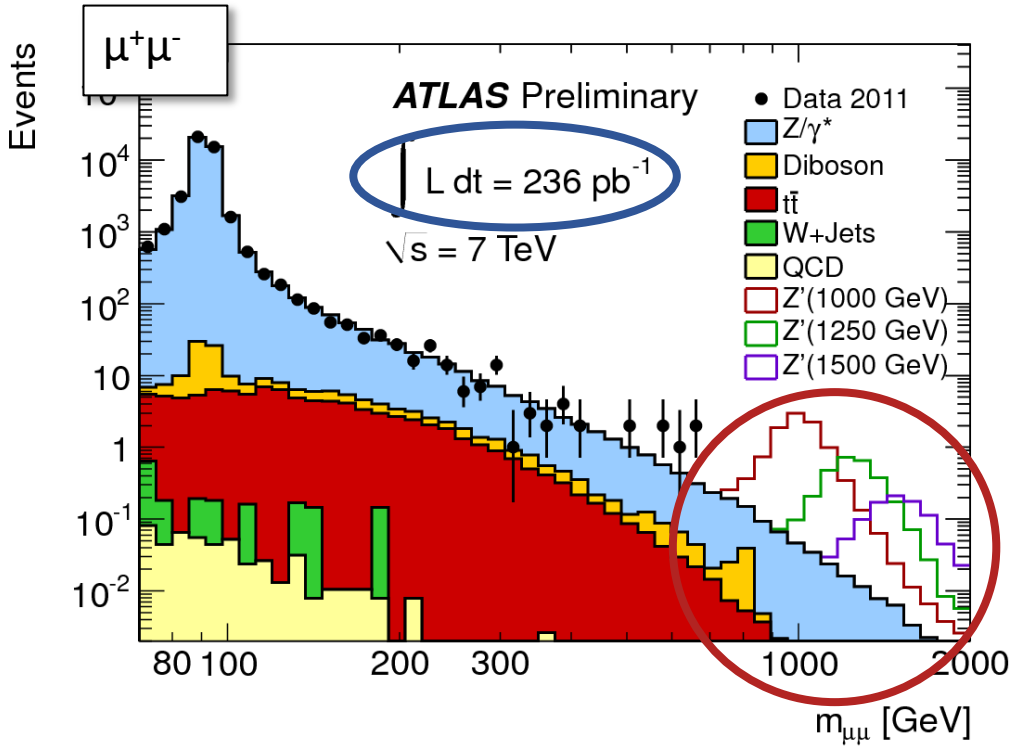
1. **Data**
2. **Simulated background events**
3. **Simulated signal with different masses**

Data inconsistent with a 1TeV Z'

Cross-section decreases with mass (higher the mass of the Z', the more data needed to discover it)

A SMALL COMPARISON

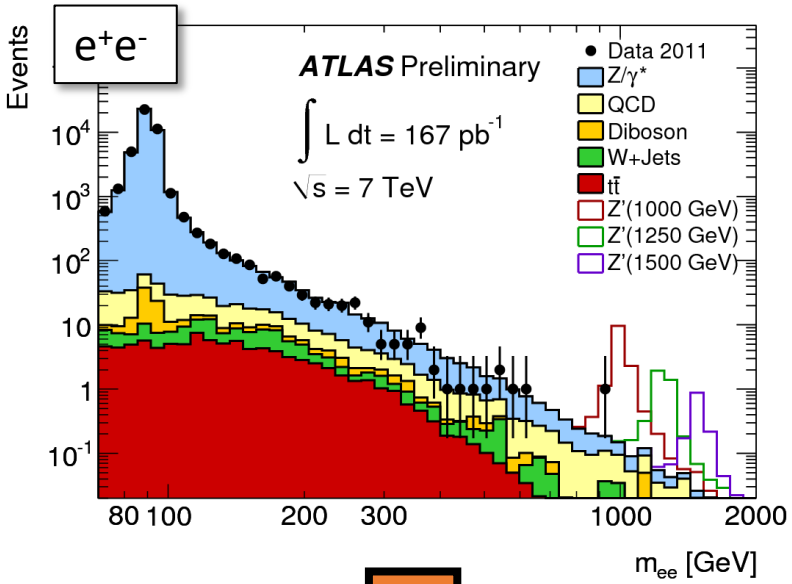


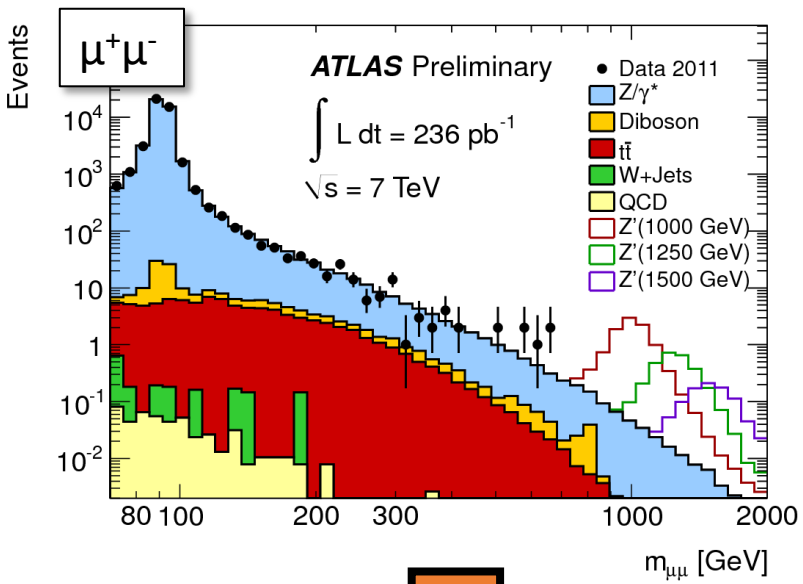


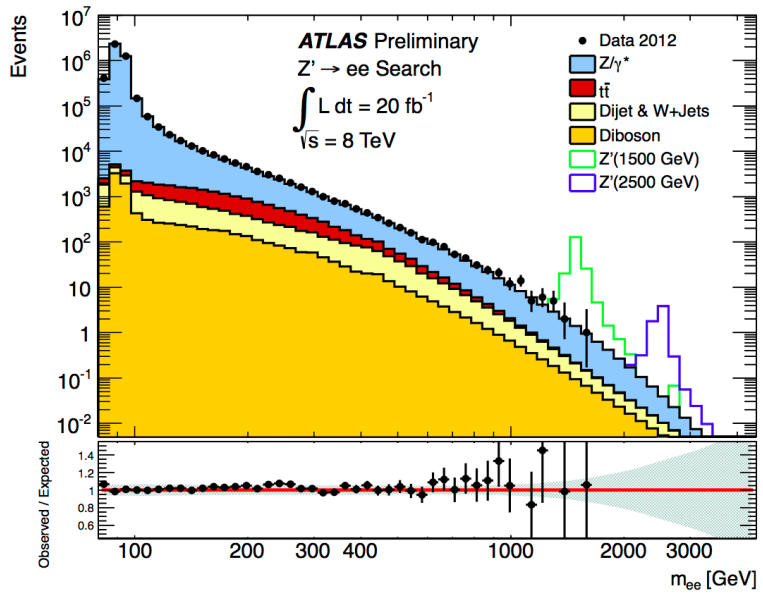
Differences in:

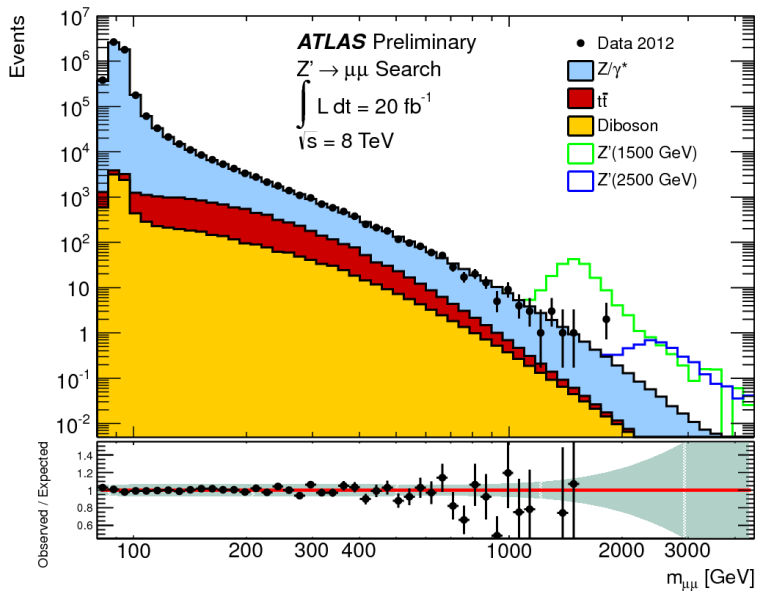
- **Resolution**
- **Background composition**
- **Dataset**

Evolution...

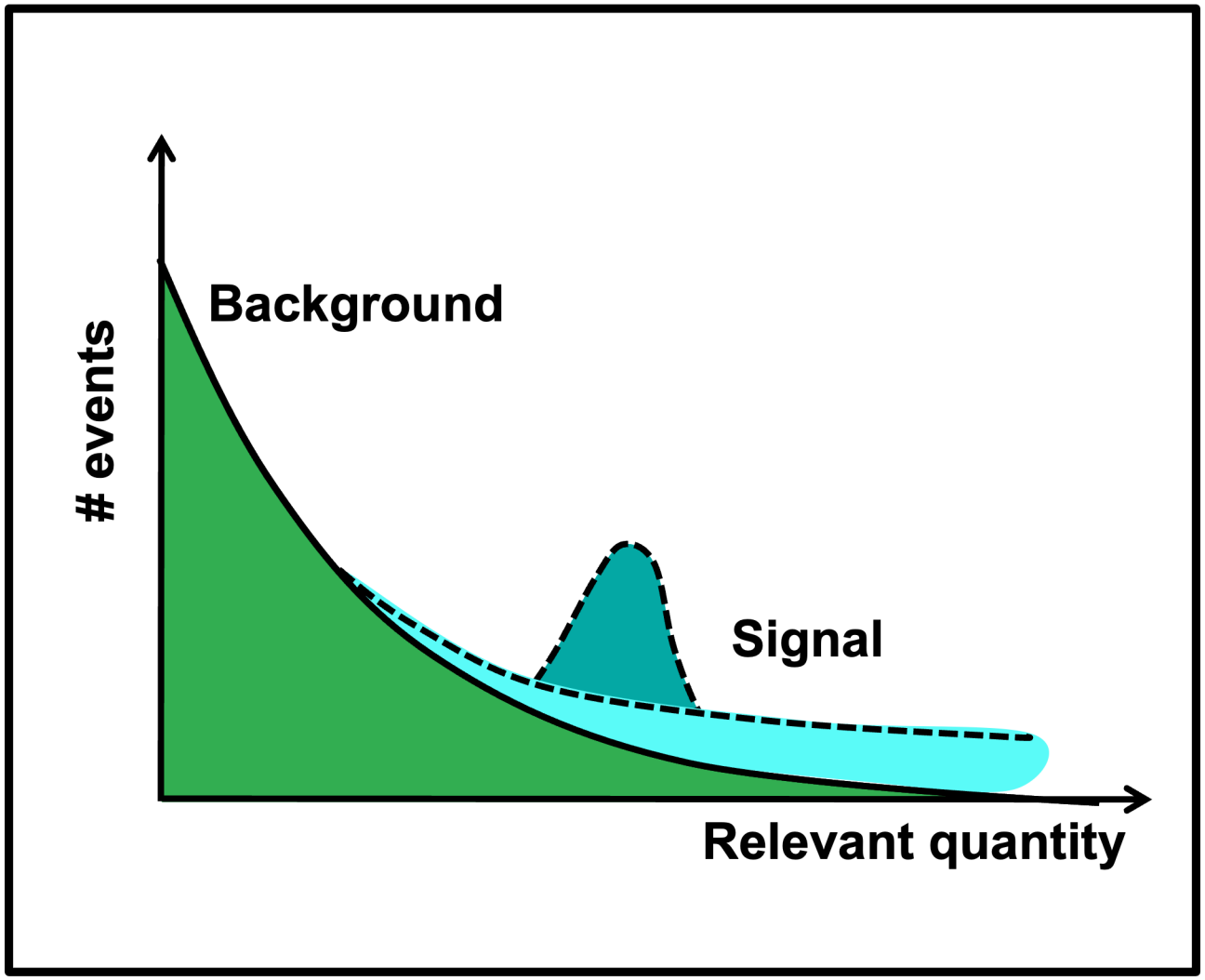

e+e-
ATLAS Preliminary
∫ L dt = 167 pb-1
√s = 7 TeV
Data 2011
Z/γ*
QCD
Diboson
W+Jets
tt̄
Z'(1000 GeV)
Z'(1250 GeV)
Z'(1500 GeV)
Events
mee [GeV]


μ+μ-
ATLAS Preliminary
∫ L dt = 236 pb-1
√s = 7 TeV
Data 2011
Z/γ*
Diboson
tt̄
W+Jets
QCD
Z'(1000 GeV)
Z'(1250 GeV)
Z'(1500 GeV)
Events
mμμ [GeV]

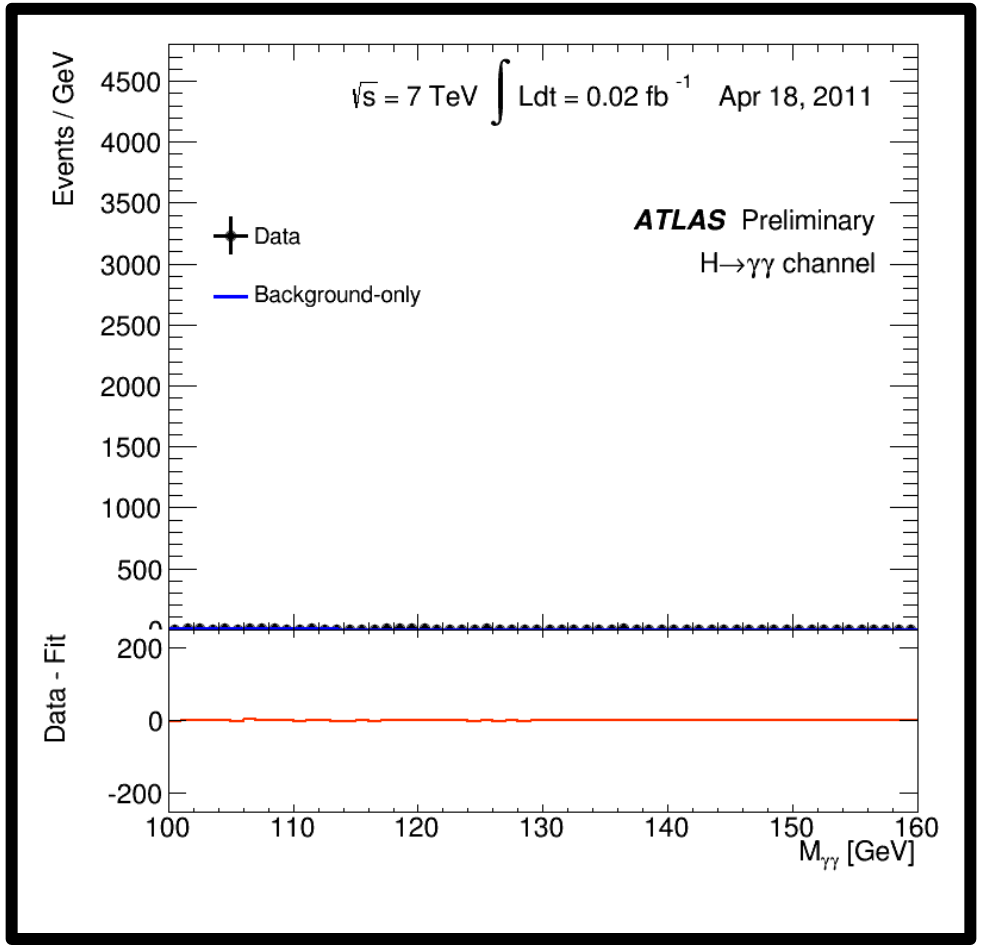

ATLAS Preliminary
Z' → ee Search
∫ L dt = 20 fb-1
√s = 8 TeV
Data 2012
Z/γ*
tt̄
Dijet & W+Jets
Diboson
Z'(1500 GeV)
Z'(2500 GeV)
Events
Observed / Expected
mee [GeV]

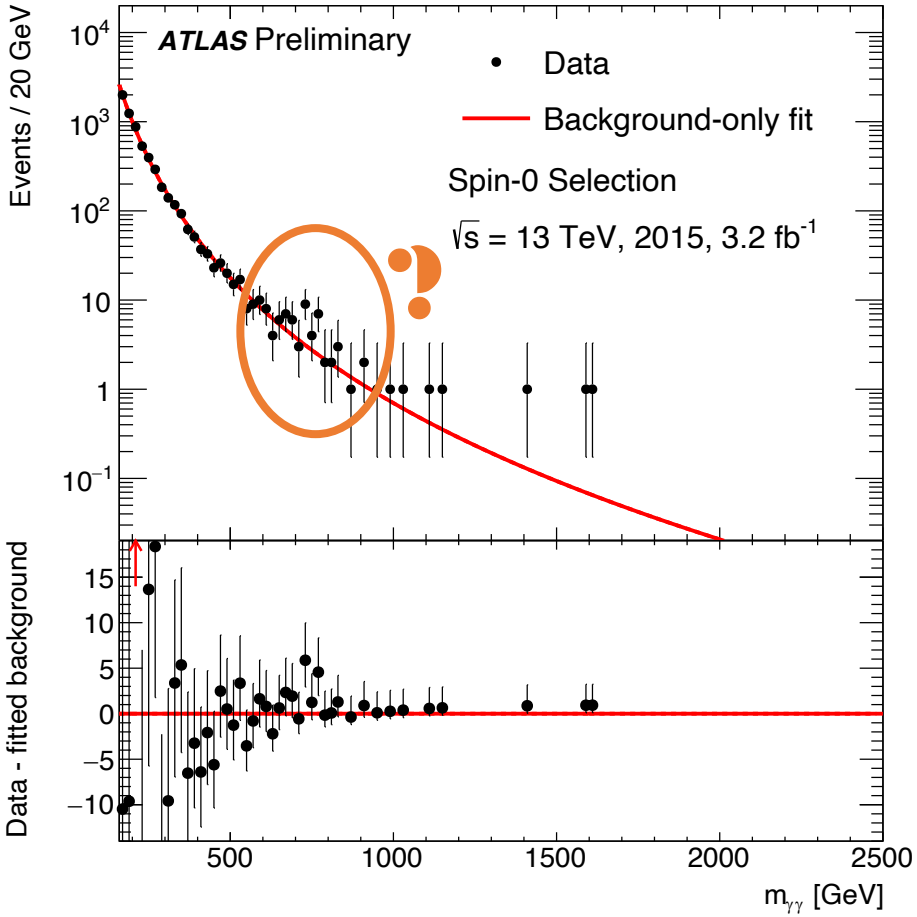

ATLAS Preliminary
Z' → μμ Search
∫ L dt = 20 fb-1
√s = 8 TeV
Data 2012
Z/γ*
tt̄
Diboson
Z'(1500 GeV)
Z'(2500 GeV)
Events
Observed / Expected
mμμ [GeV]

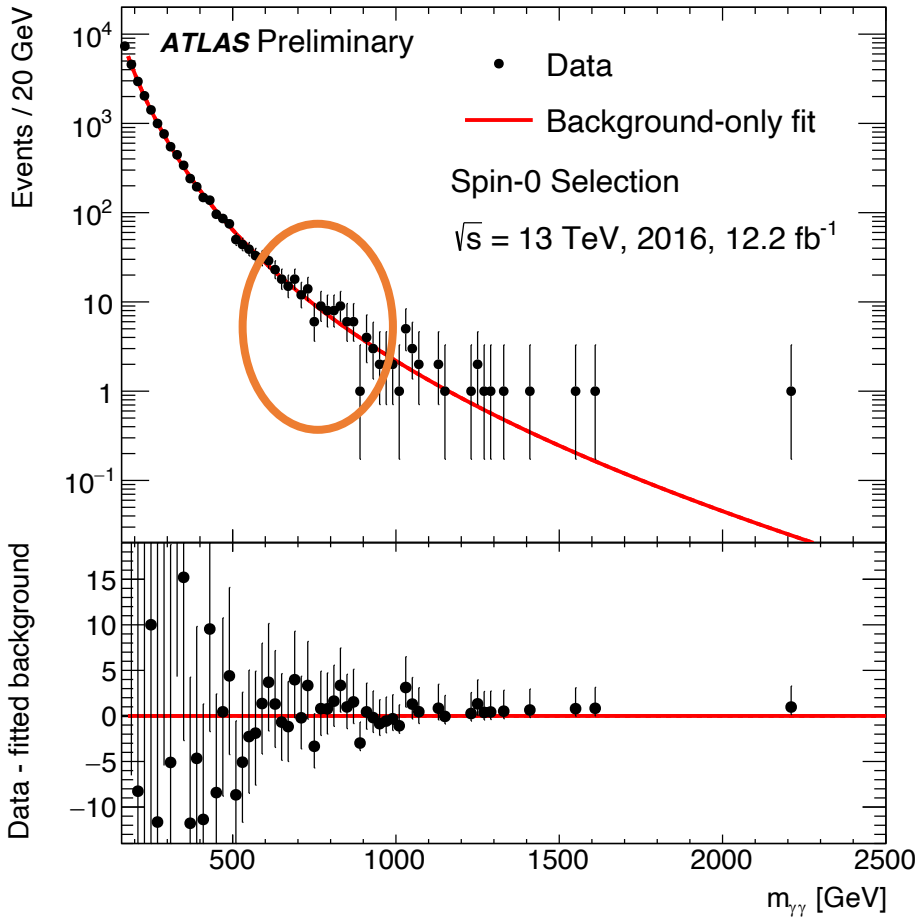
SEARCHES



A WELL-KNOWN BUMP SEARCH

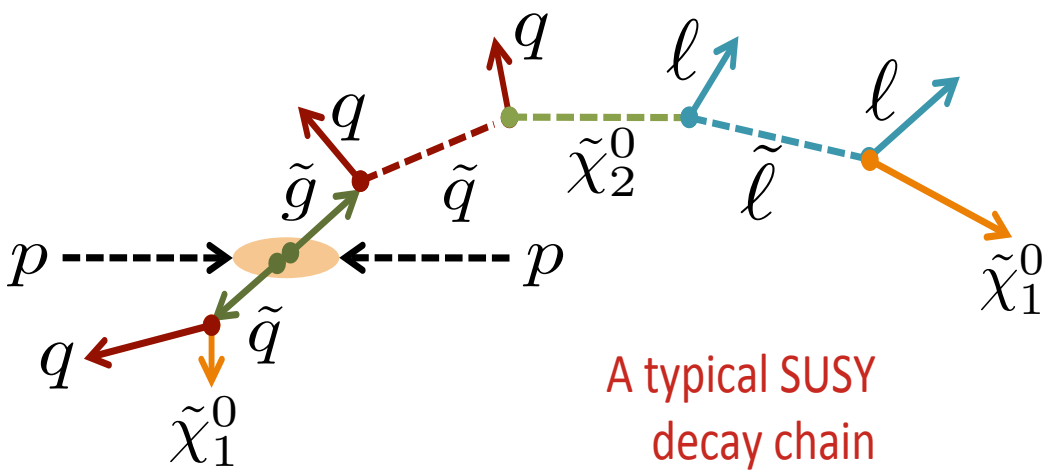



ATLAS Preliminary
Data
Background-only fit
Spin-0 Selection
√s = 13 TeV, 2015, 3.2 fb-1
Events / 20 GeV
Data - fitted background
mγγ [GeV]


ATLAS Preliminary
Data
Background-only fit
Spin-0 Selection
√s = 13 TeV, 2016, 12.2 fb-1
Events / 20 GeV
Data - fitted background
mγγ [GeV]

MARIO
096950
x23
WORLD
1-4
TIME
593
THANK YOU MARIO!
BUT OUR PRINCESS IS IN
ANOTHER CASTLE!

Another SEARCH example:
Search for SUSY in events with large jet multiplicities



A typical SUSY decay chain

Disclaimer:

This is only an example!

There are numerous such searches!

Each of them differs in

- event selections,
- background determinations,
- methodology

Searching for new physics is fun!

From Raw Data to Physics

Instead of summary:

Components of an analysis

Components of a physics analysis

- Data-set and Monte Carlo samples
- Trigger
- Object definitions and event selections
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

Components of a physics analysis

- Data-set and Monte Carlo samples
- Trigger
- Object definitions a
- Background determ
- Systematic uncertai
- Statistical methods
- Results
- [Interpretations]

The data and simulation samples used in the analysis. Data for the measurement / search, simulation to compare data to predictions.

Data-set specifics:
- Data quality ⇒ Good run list.
- Luminosity.

Monte Carlo sample specifics:
- Generator, tunes.
- Statistics.

Components of a physics analysis

- Data-set and Monte Carlo samples
- Trigger
- Object defi
- Background de
- Systematic uncerta
- Statistical methods
- Results
- [Interpretations]

The trigger used to collect the data with.

Trigger specifics:

- Prescales; typically unprescaled triggers are used, prescaled triggers for QCD / high stat measurements.
- Trigger (in)efficiencies.

Components of a physics an

- Data-set and Monte Carlo sa
- Trigger
- Object definitions and event se
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

The exact definition of objects (electrons, muon, jets, …) and how these are combined in selecting events to be analyzed.

Object definition specifics:
- "Flavor" of the identification (loose, medium, tight).
- Calibrations.

Event selection specifics:
- Event cleaning (e.g. from noise and cosmics).
- Momentum, geom. acceptance and multiplicity of objects.
- Higher level cuts, such as invariant mass.
- "**Signal regions**".

Components of a physics analysis

- Data-set and Monte Carlo samples
- Trigger
- Object definitions and event s
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

Events that are imitating the signal we are searching for or measuring.

Background determination specifics:

- Can/must be **data-driven** or **simulation-based**.
- "**Validation regions**" and "**control regions**" required. These can use different triggers wrt signal regions.

Components of a physics analysis

- Data-set and Monte Carlo sa
- Trigger
- Object definitions and event
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

- Any ‘intermediate’ measurement we have performed carries uncertainties (statistical and systematic).
- **“Systematic” uncertainties are introduced by inaccuracies in the methods used to perform the measurement.**
- Efficiencies, acceptance, number of events, luminosity, cross sections used in Monte Carlo scaling…
- Some of them are “centrally” assessed by the performance groups of an experiment. Some of them are analysis-specific.

Components of a physics analysis

- Data-set and Monte Carlo samples
- Trigger
- Object definitions and event selections
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

Dealing with large data-sets, we use statistical methods to make sense of the numbers we measure.

Typical method:
- Do a fit to extract signal from background.

Methodologies can vary a lot, but nowdays they are pretty unified within and across experiments.

Neural nets and other machine learning methods are broadly used, primarily to improve signal over background discrimination!

Components of a physics analysis

- Data-set and Monte Carlo samples
- Trigger
- Object definitions and event selections
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

Produce the results in tables and plots. These include details of what is found in the signal region.

Components of a physics analysis

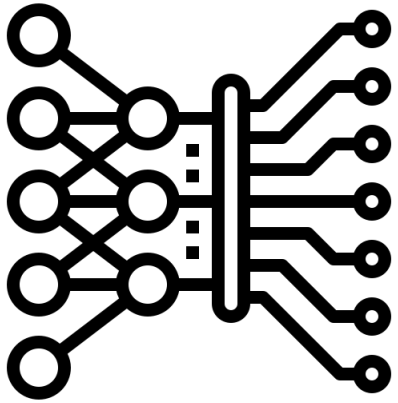
- Data-set and Monte Carlo samples
- Trigger
- Object definitions and event selections
- Background determination
- Systematic uncertainties
- Statistical methods
- Results
- [Interpretations]

Put the results into context:
interpret them in theoretical models.

Machine Learning

How it is used in HEP

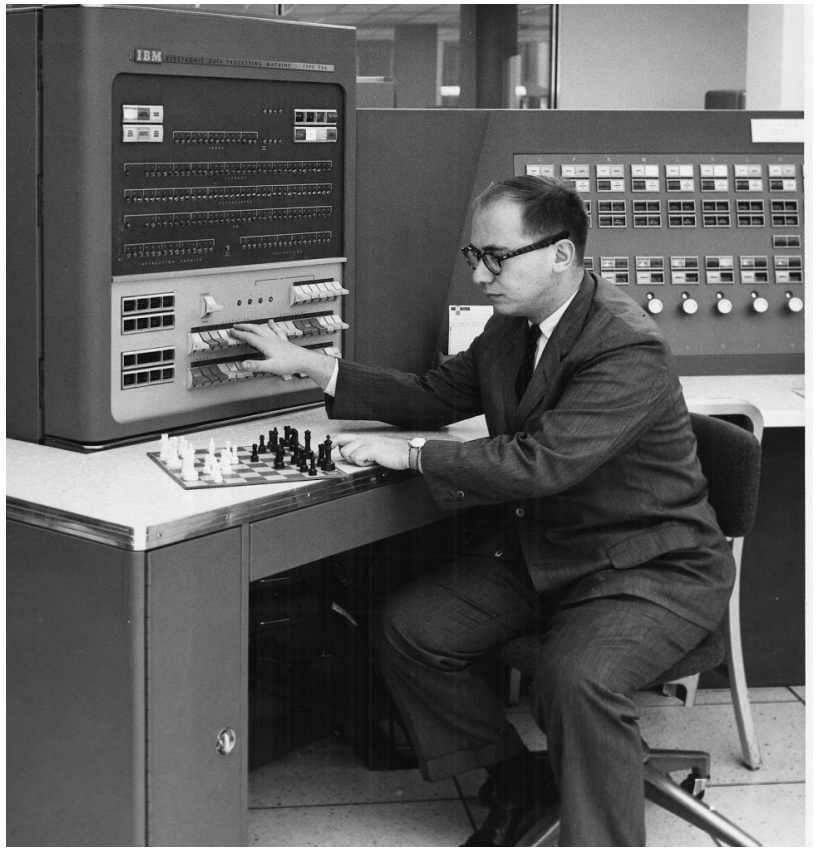
Just a teaser!
The topic is broad and can easily be a long lecture itself!



What is Machine Learning?

"Giving computers the ability to learn without explicitly programming them"

(Arthur Samuel, 1959)

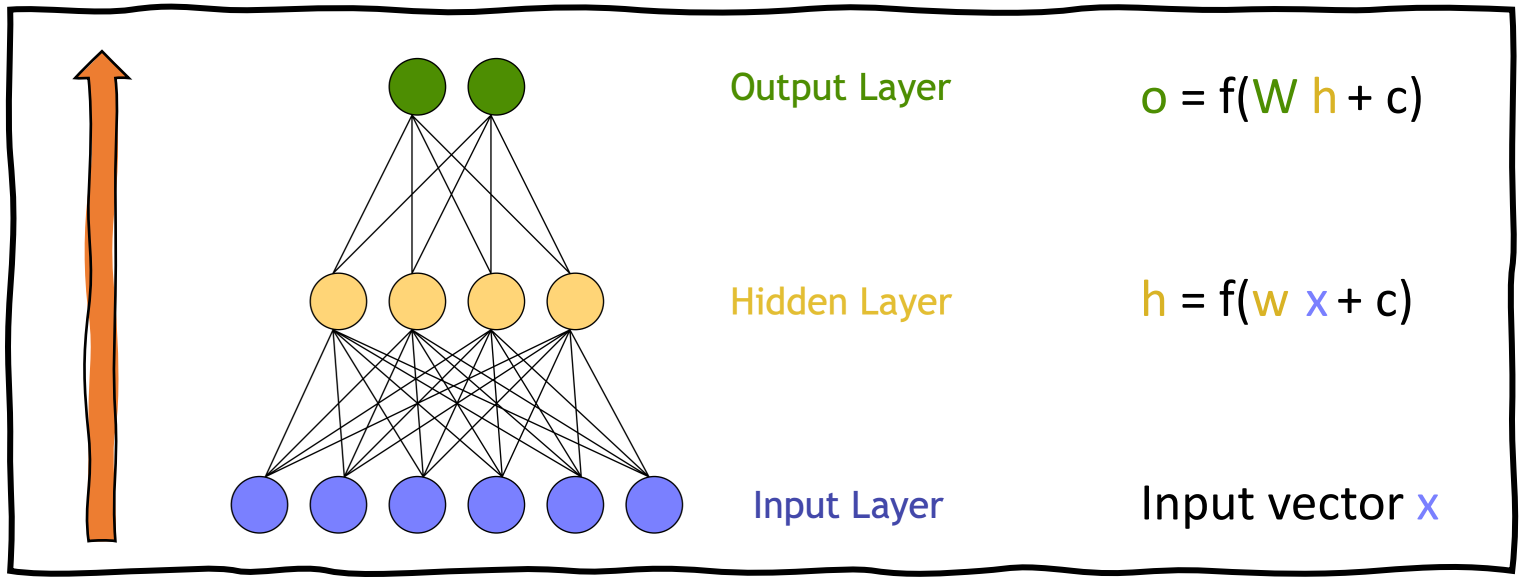


1957 : Alex Bernsteil demonstrates use of IBM's 704 computer AI playing chess against it

1959 : Arthur Samuel uses the term "Machine Learning" for the first time when talking about his checkers program

What is Machine Learning?

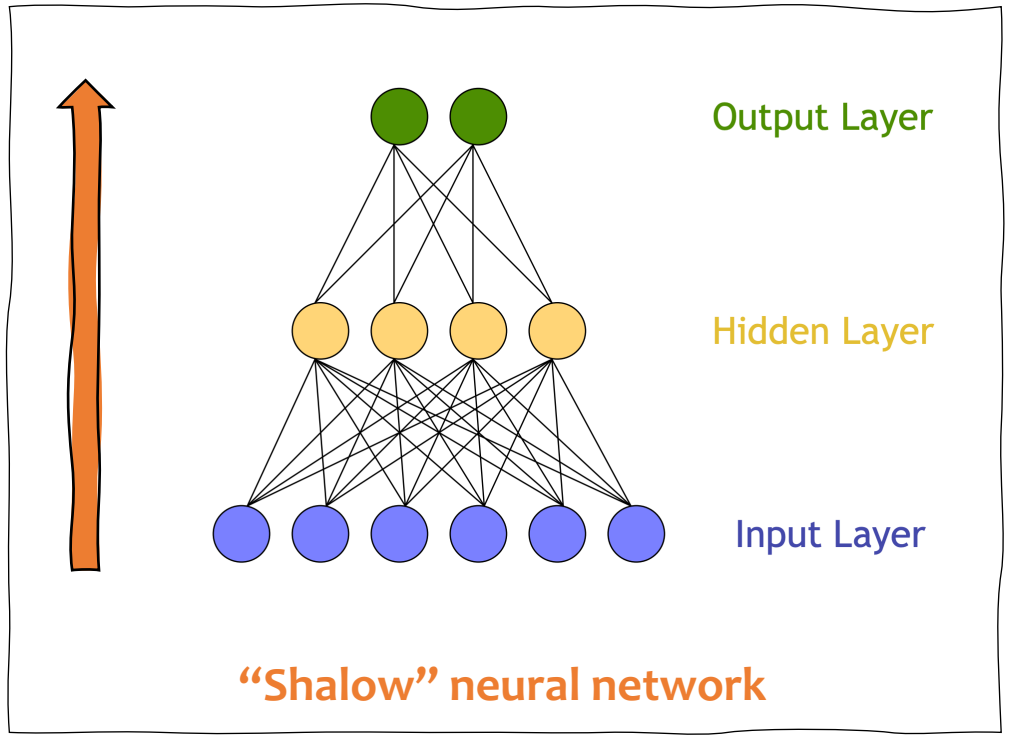
Let's use a Neural Network (NN) as an example

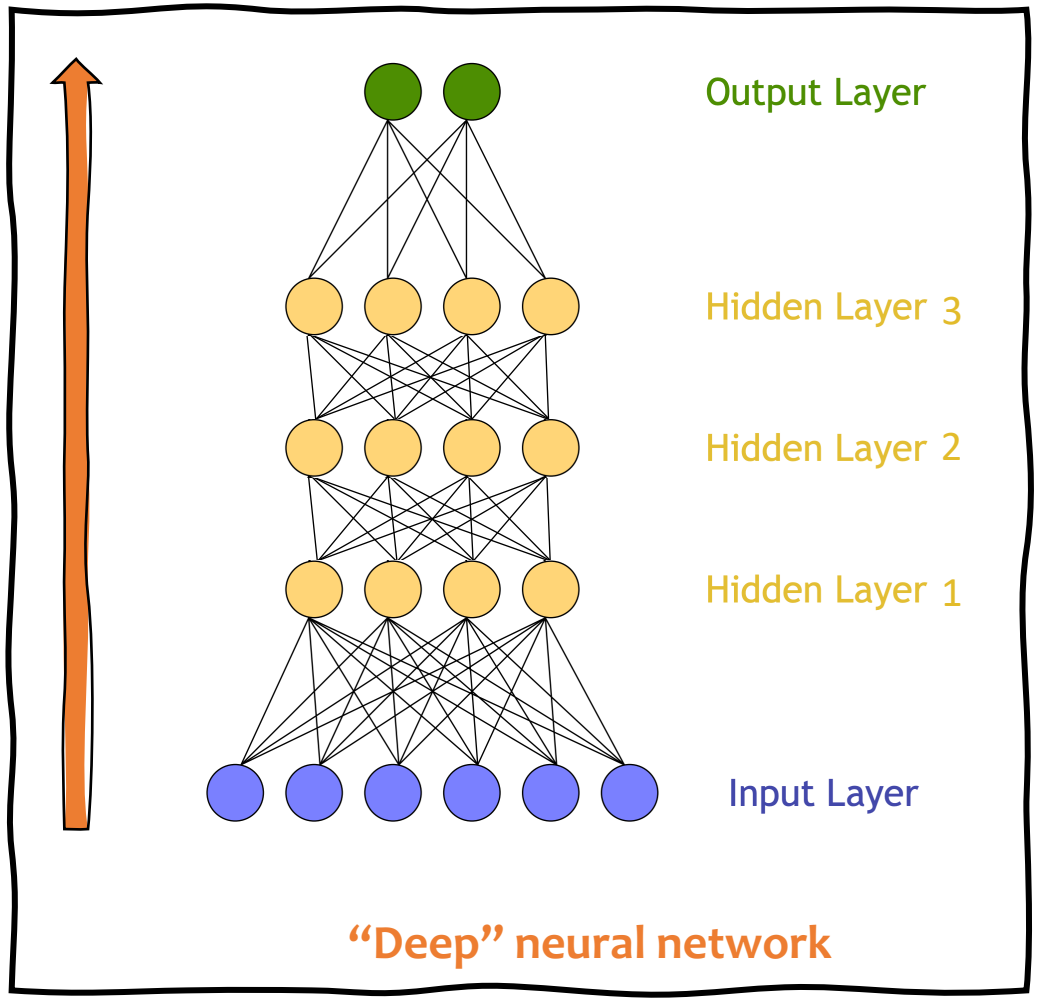


In simple words:

- **NN**: single-valued function (of weights and other parameters) of input values
- **Training**: optimization of that function on test sample (e.g. MC)
  - The function "learns" the model
  - Minimisation of an "error function" (e.g. χ^2) to find optimized weights and parameters
- **Inference**: Use that optimized function on real data

Is it really that simple?



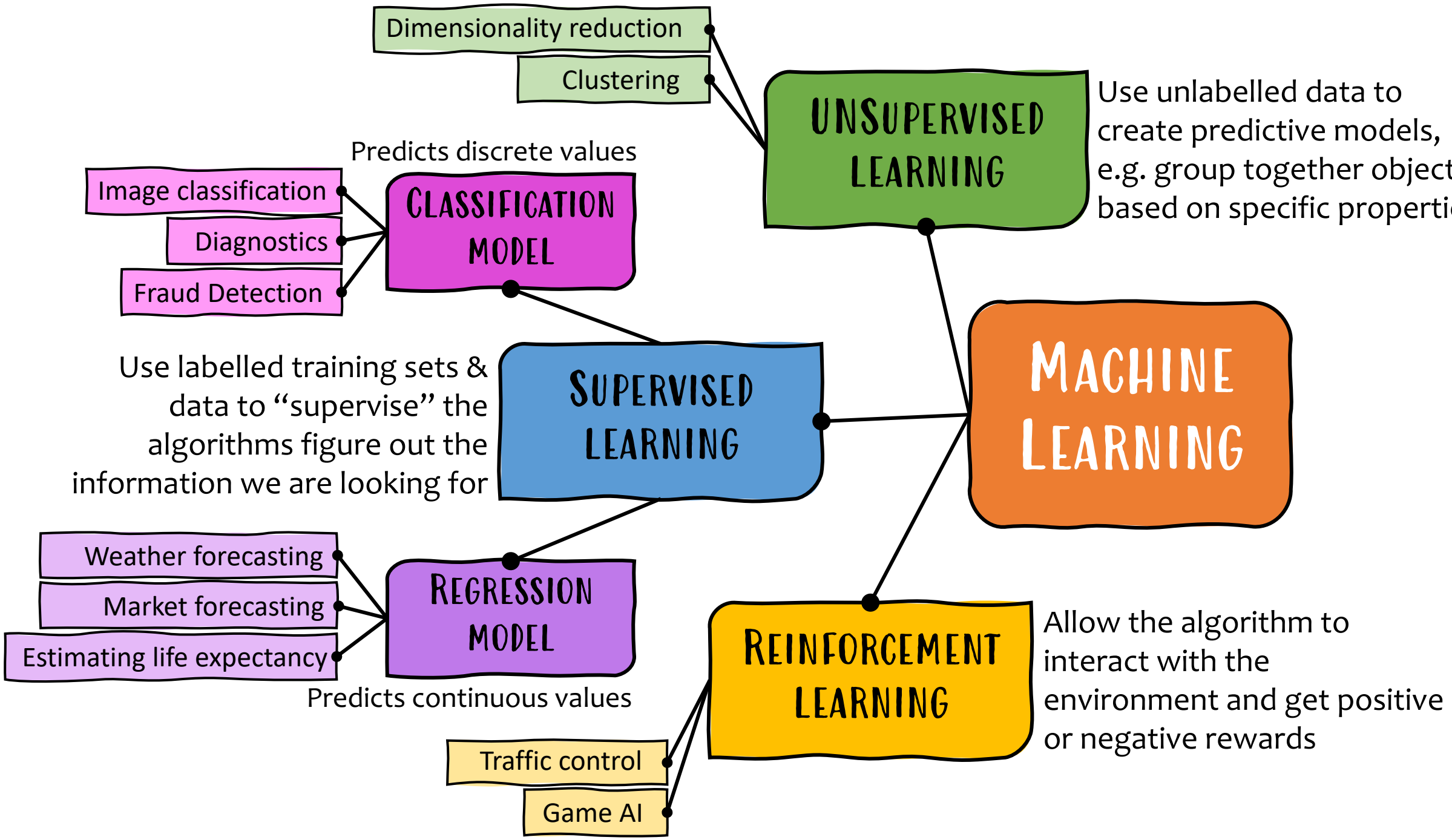


More hidden layers (deep networks) allow:

- Factorised learning of the structure of data
- Progressivelly learning more complex data sets

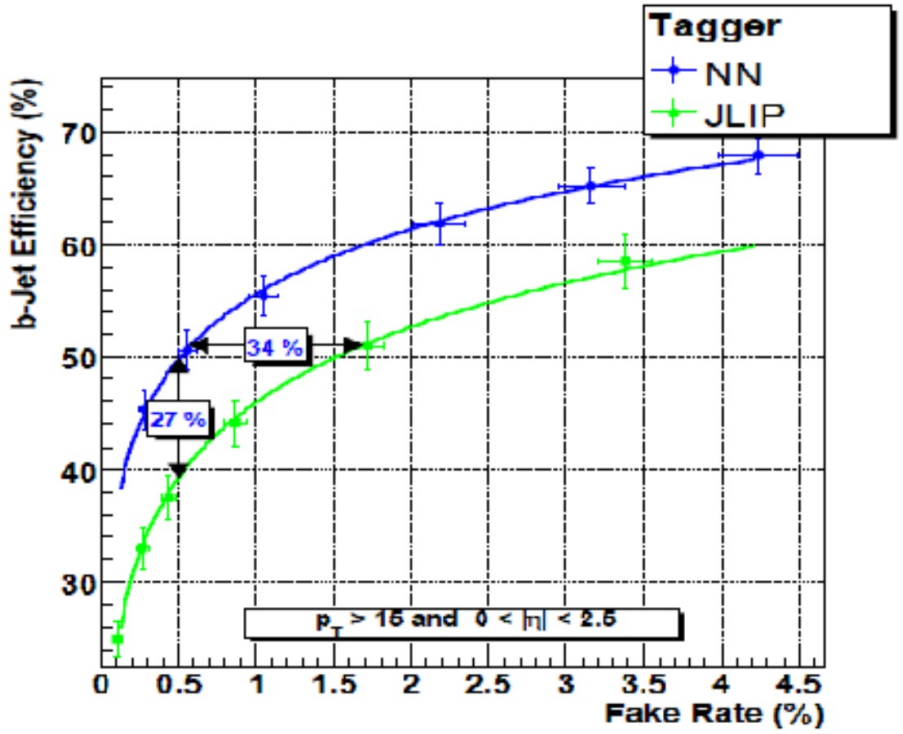
But…

- Very difficult to train
- Decades of research led to great advancements!


Machine Learning
Supervised Learning
Use labelled training sets & data to "supervise" the algorithms figure out the information we are looking for
Classification Model
Predicts discrete values
Image classification
Diagnostics
Fraud Detection
Regression Model
Predicts continuous values
Weather forecasting
Market forecasting
Estimating life expectancy
Unsupervised Learning
Use unlabelled data to create predictive models, e.g. group together objects based on specific properties
Dimensionality reduction
Clustering
Reinforcement Learning
Allow the algorithm to interact with the environment and get positive or negative rewards
Traffic control
Game AI

Machine Learning in HEP

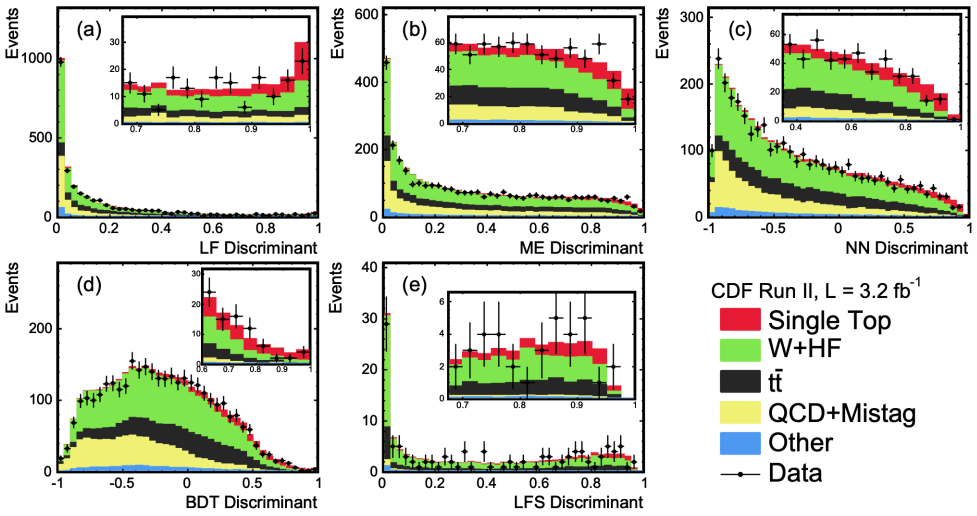
- Used since very long (the 90's, if not before)
- In the past: neural networks and other MVAs (e.g. BDTs).
- Nowdays: More complicated and "deep" networks. CNNs, RNNs, GANs, …
- Used in reconstruction
  - Classic example: b-tagging, already at the Tevatron (since early 2000)
- Used in analyses, e.g. for s/b optimisations
  - Historic example: discovery of single top at the Tevatron (2009)
- Higgs discovery made use of multiple MVAs

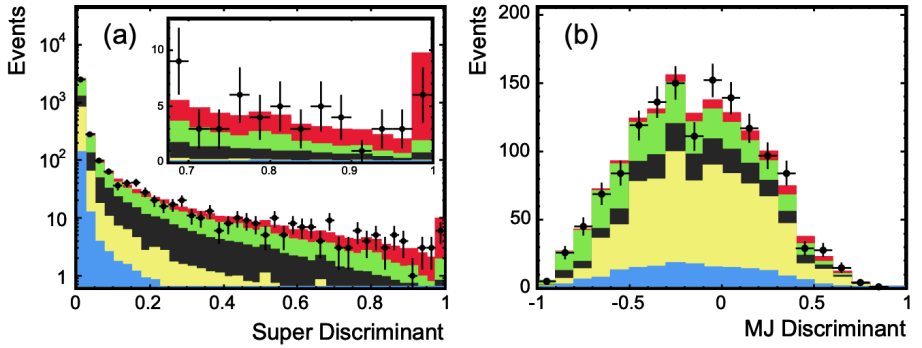


Efficiency and purity for the Do MVA tagger, compared to the "conventional" one

Machine Learning in HEP

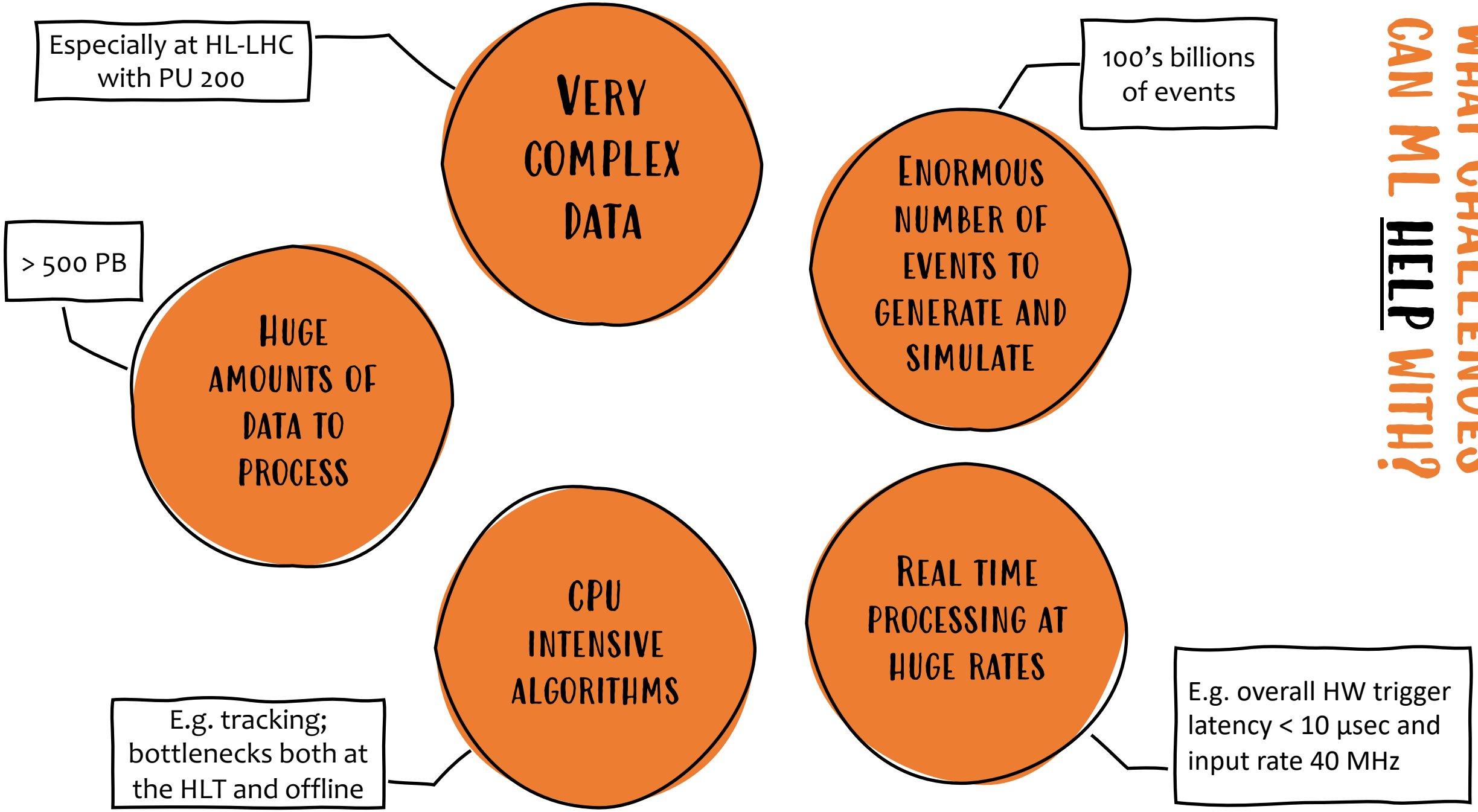
- Used since very long (the 90's, if not before)
- In the past: neural networks and other MVAs (e.g. BDTs).
- Nowdays: More complicated and "deep" networks. CNNs, RNNs, GANs, …
- Used in reconstruction
  - Classic example: b-tagging, already at the Tevatron (since early 2000)
- Used in analyses, e.g. for s/b optimisations
  - Historic example: discovery of single top at the Tevatron (2009)
- Higgs discovery made use of multiple MVAs





Discriminants at the CDF single-top discovery (arXiv:0903.0885)

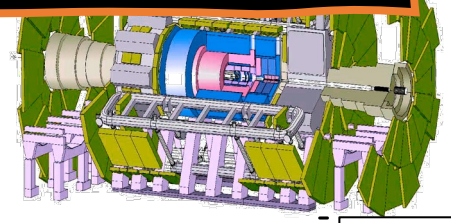
WHAT CHALLENGES CAN ML HELP WITH?



Machine learning in HEP nowdays

1 For improved performance (e.g. b-tagging) and CPU needs (e.g. tracking).

7 Could be used at FE electronics in the future ?



Every 25 ns

Detector

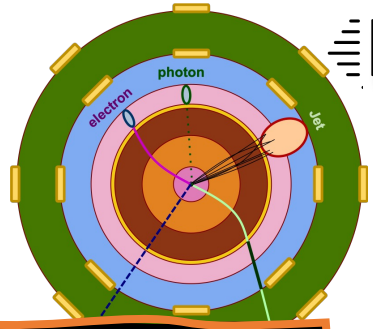
2.5 µs

~0.3 s

Trigger

RAW DATA

Reconstruction



Day(s) - Month(s)

Calibration

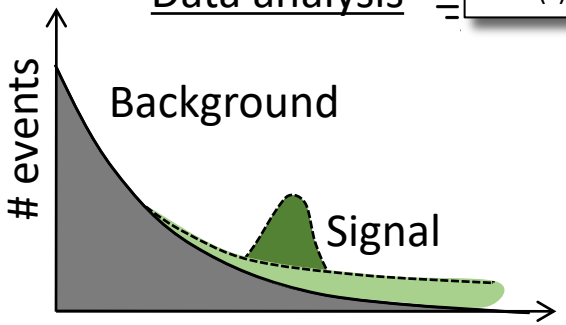
2 For improved performance (e.g. for jets)

5 To enable selections (e.g. tau) and reduce resource needs (e.g. anti-kT jet finding) (esp. in the future)

6 Most importantly to reduce CPU needs (e.g. tracking) (esp. in the future)

3 Rich (challenging) possibilities : S/B, taggers, new methods such as anomaly detection, …

Data analysis



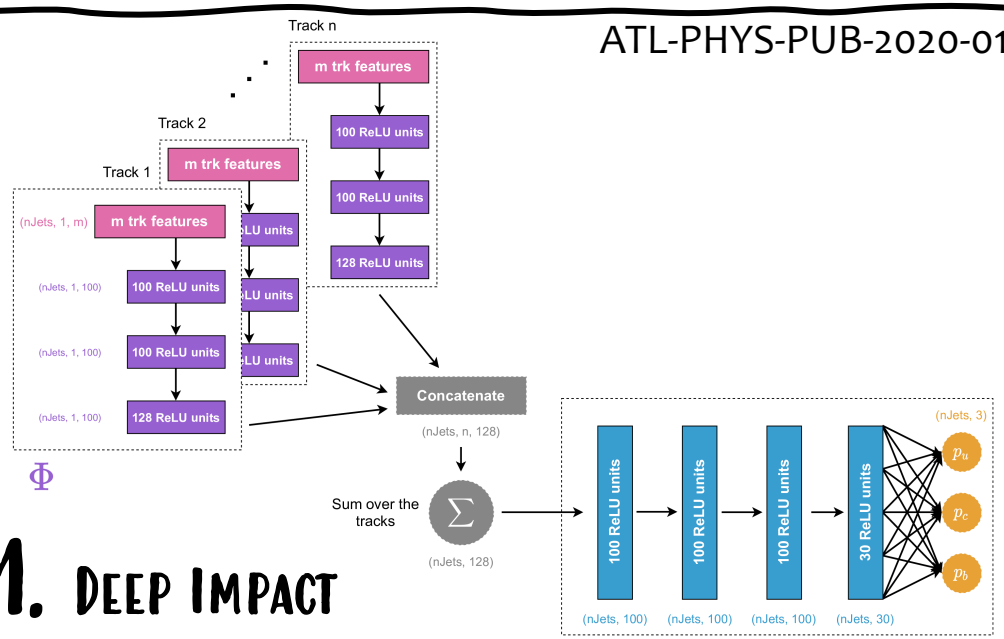
Theory / Simulations

Month(s) - Year(s)

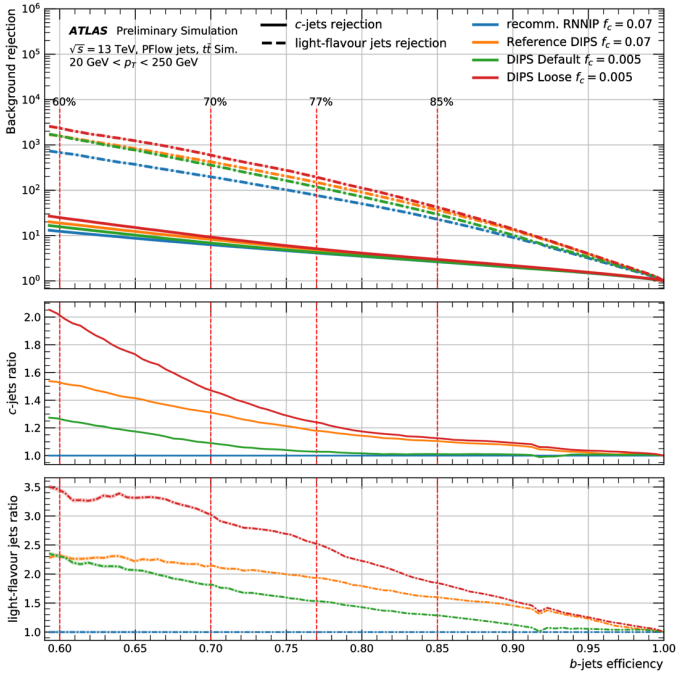
4 Significant recent RnD both for generation and simulation

Two examples from Flavour Tagging

ATL-PHYS-PUB-2020-014



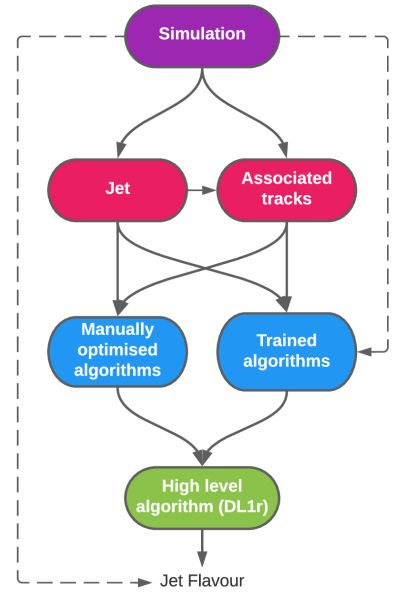
1. Deep Impact Parameter Sets (DIPS)

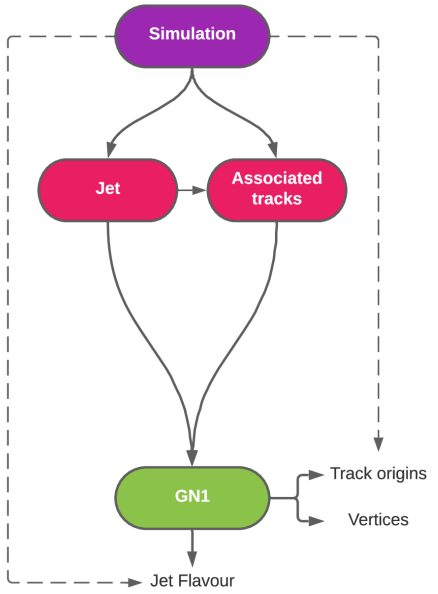


Besides performance, this NN improved by x5 the training time

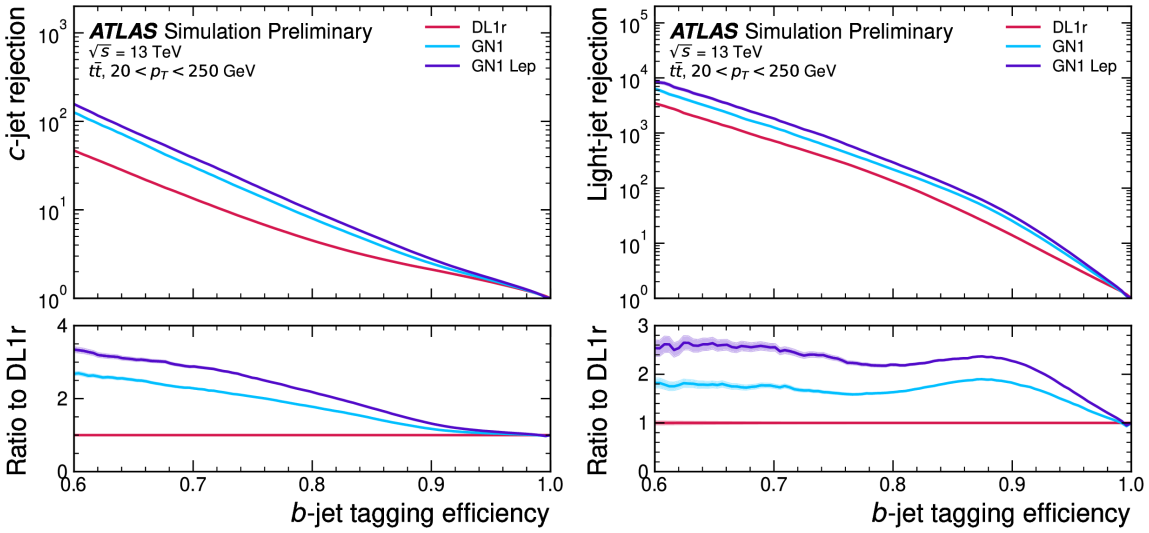
ATL-PHYS-PUB-2022-027

2. Graph Neural Network





Flexible algorithm that requires no inputs from other tagging algorithms



Machine learning in HEP nowdays

To efficiently use ML, we need to understand

- the problem we want to solve
- How ML can help address a specific problem
- What ML algorithm / architecture to use

And then ML can be an extremely useful **tool**!

https://xkcd.com/1838/

Exciting Times coming up in HEP

Good luck in your research!

Please get in touch for question, comments,
or simply feedback on this lecture

anna.sfyrla@unige.ch