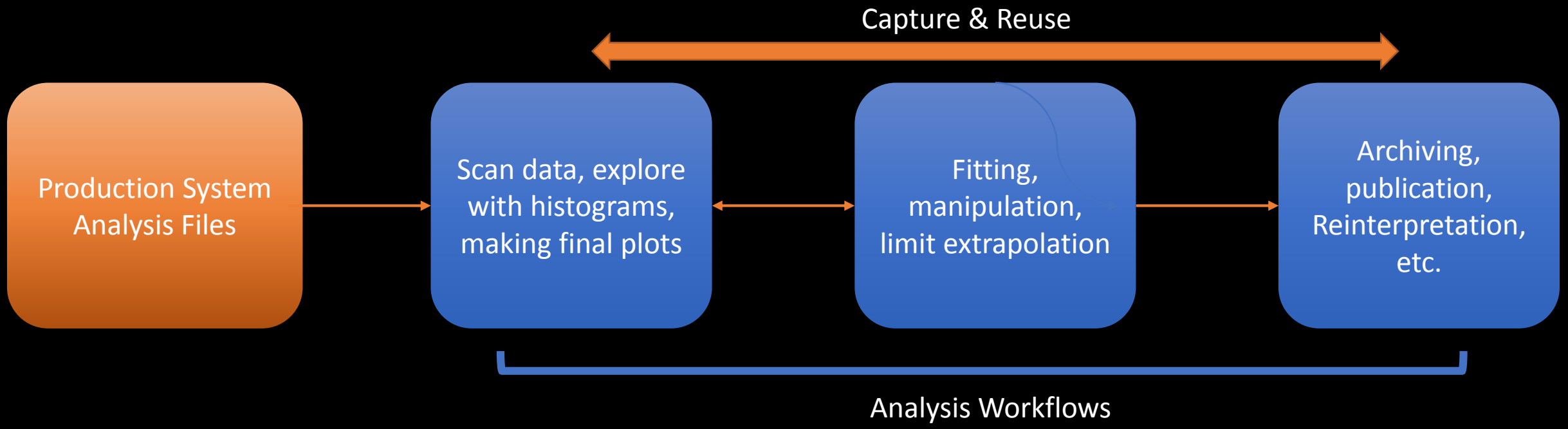

CHEP
2019
NSF
Analysis Workflows
G. Watts (UW)
iris
hep
Institute for Research & Innovation
in Software for High Energy Physics

Analysis Workflows

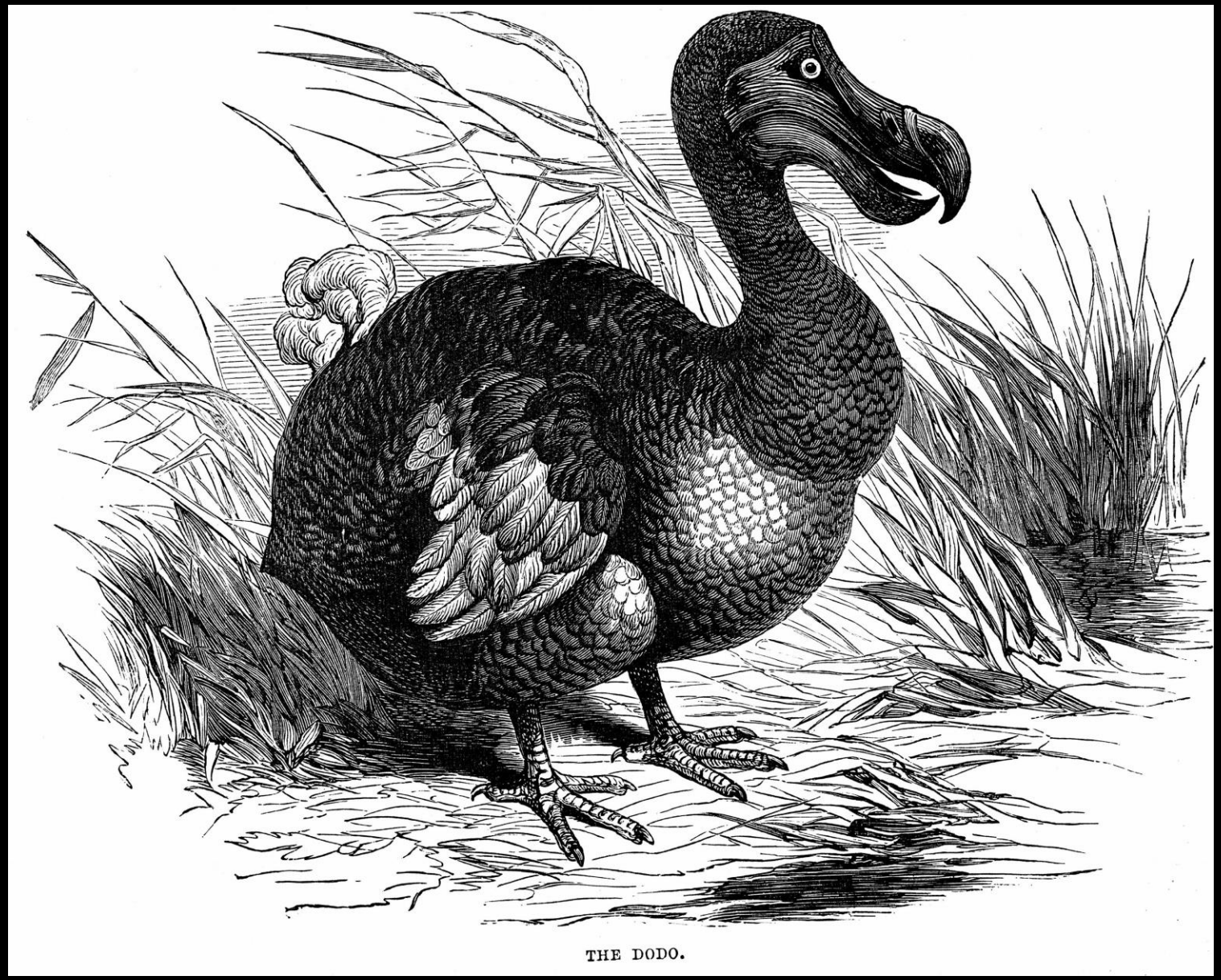


Challenges

I have my opinion.

Don't confuse me with facts!

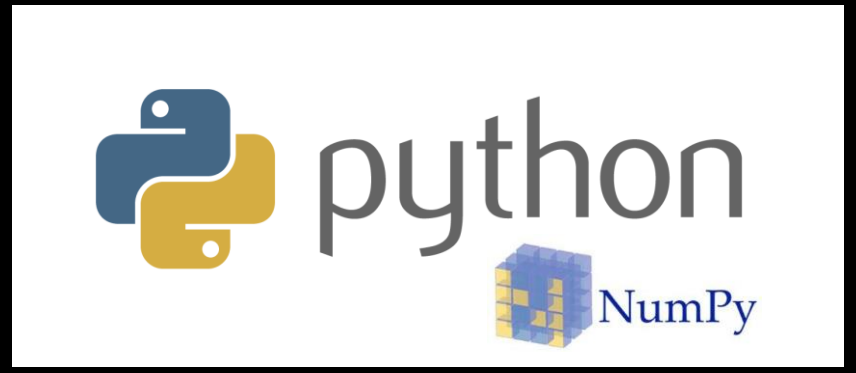
Just leading questions today…

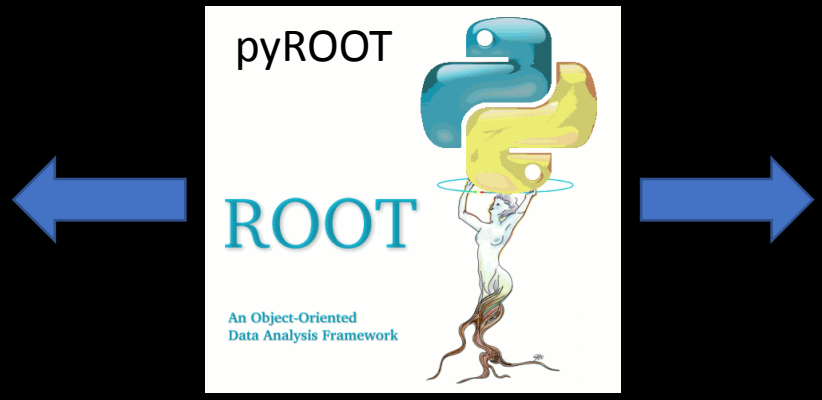


Challenges

Two Ecosystems

Two.five Ecosystems





Capture & Reuse

Production System Analysis Files

Scan data, explore with histograms, making final plots

Fitting, manipulation, limit extrapolation

Archiving, publication, Reinterpretation, etc.

As an example…

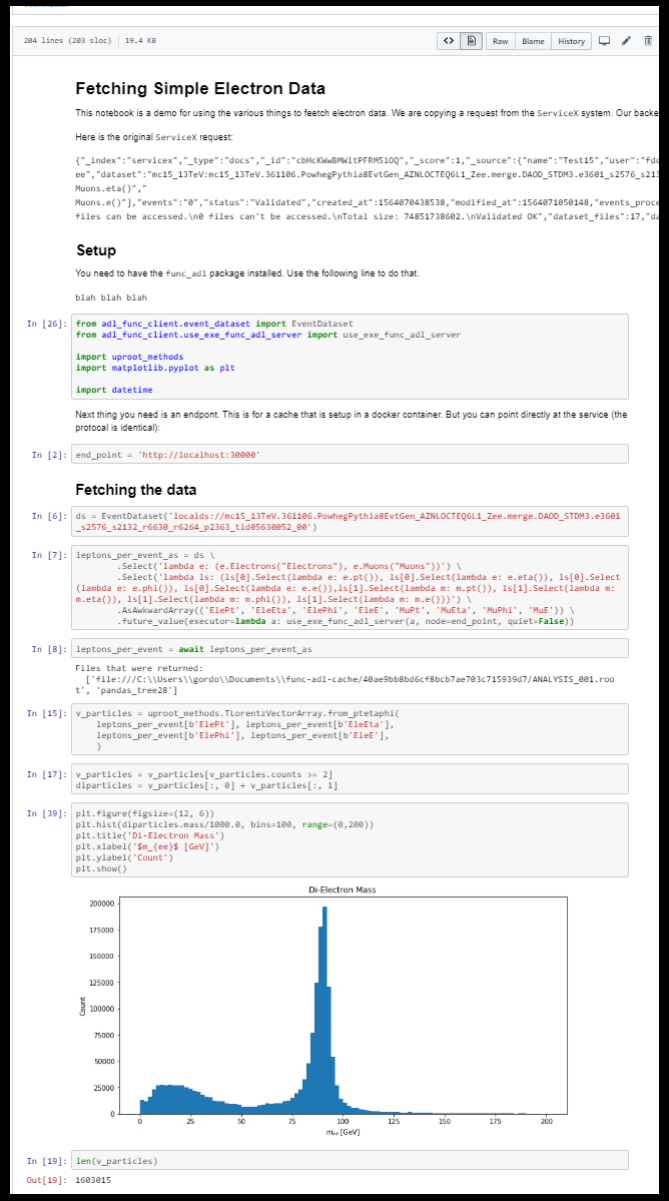
Awkward Array

Interactive Analysis

TTree::Draw

How much should we support this in the field? Integration?

What would this do to our IoC Frameworks?



204 lines (203 sloc) | 19.4 KB

Raw Blame History

Fetching Simple Electron Data

This notebook is a demo for using the various things to feetch electron data. We are copying a request from the ServiceX system. Our backe

Here is the original ServiceX request:

{"_index":"servicex","_type":"docs","_id":"cbHcKWwBMW1tPFRM5lOQ","_score":1,"_source":{"name":"Test15","user":"fd
ee","dataset":"mc15_13TeV:mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.merge.DAOD_STDM3.e3601_s2576_s21
Muons.eta()","
Muons.e()"],"events":"0","status":"Validated","created_at":1564070438538,"modified_at":1564071050148,"events_proc
files can be accessed.\n0 files can't be accessed.\nTotal size: 74851738602.\nValidated OK","dataset_files":17,"d

Setup

You need to have the func_adl package installed. Use the following line to do that.

blah blah blah

In [26]:
```
from adl_func_client.event_dataset import EventDataset
from adl_func_client.use_exe_func_adl_server import use_exe_func_adl_server

import uproot_methods
import matplotlib.pyplot as plt

import datetime
```

Next thing you need is an endpont. This is for a cache that is setup in a docker container. But you can point directly at the service (the protocal is identical):

In [2]:
```
end_point = 'http://localhost:30000'
```

Fetching the data

In [6]:
```
ds = EventDataset('localds://mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.merge.DAOD_STDM3.e3601
_s2576_s2132_r6630_r6264_p2363_tid05630052_00')
```

In [7]:
```
leptons_per_event_as = ds \
        .Select('lambda e: (e.Electrons("Electrons"), e.Muons("Muons"))') \
        .Select('lambda ls: (ls[0].Select(lambda e: e.pt()), ls[0].Select(lambda e: e.eta()), ls[0].Select
(lambda e: e.phi()), ls[0].Select(lambda e: e.e()),ls[1].Select(lambda m: m.pt()), ls[1].Select(lambda m:
m.eta()), ls[1].Select(lambda m: m.phi()), ls[1].Select(lambda m: m.e()))') \
        .AsAwkwardArray(('ElePt', 'EleEta', 'ElePhi', 'EleE', 'MuPt', 'MuEta', 'MuPhi', 'MuE')) \
        .future_value(executor=lambda a: use_exe_func_adl_server(a, node=end_point, quiet=False))
```

In [8]:
```
leptons_per_event = await leptons_per_event_as
```

Files that were returned:
['file:///C:\\Users\\gordo\\Documents\\func-adl-cache/40ae9bb8bd6cf8bcb7ae703c715939d7/ANALYSIS_001.roo
t', 'pandas_tree28']

In [15]:
```
v_particles = uproot_methods.TLorentzVectorArray.from_ptetaphi(
    leptons_per_event[b'ElePt'], leptons_per_event[b'EleEta'],
    leptons_per_event[b'ElePhi'], leptons_per_event[b'EleE'],
    )
```

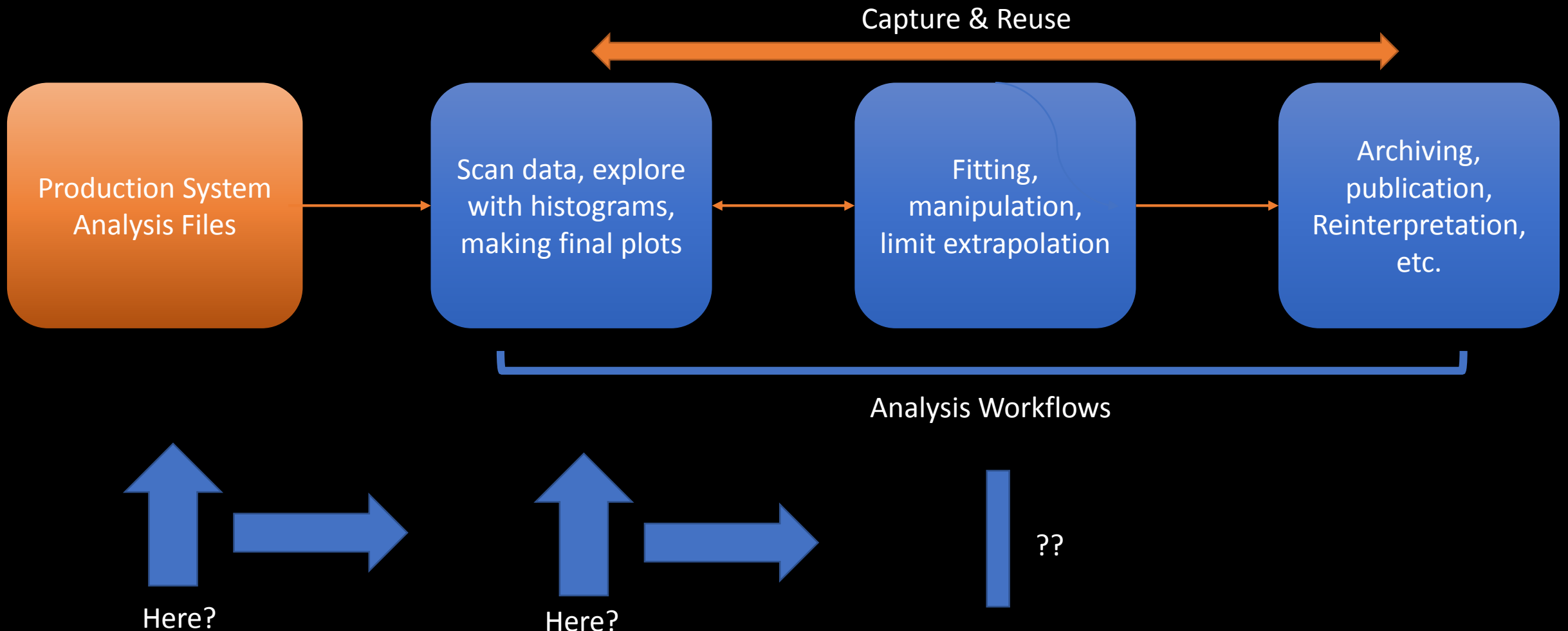
In [17]:
```
v_particles = v_particles[v_particles.counts >= 2]
diparticles = v_particles[:, 0] + v_particles[:, 1]
```

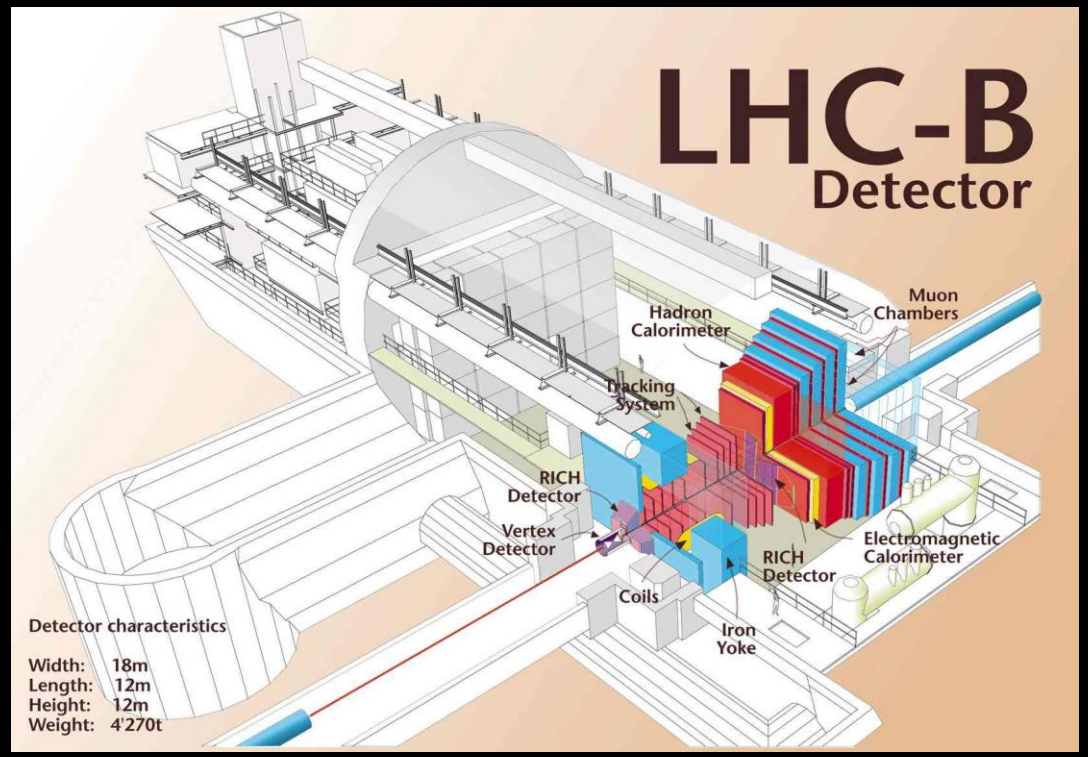
In [39]:
```
plt.figure(figsize=(12, 6))
plt.hist(diparticles.mass/1000.0, bins=100, range=(0,200))
plt.title('Di-Electron Mass')
plt.xlabel('$m_{ee}$ [GeV]')
plt.ylabel('Count')
plt.show()
```


In [19]:
```
len(v_particles)
```

Out[19]: 1603015

What does Interactive Analysis Encompass?





Here?

High Level Trigger

Here?

The New Student Challenge

New undergrad. Give them a dataset name, and say "plot $Z \rightarrow e^{+}e^{-}$"

Note I did not say anything about credentials

Interactive Analysis

Simple things should happen in “simple time”*

*Internet Time (seconds, less than a minute)

How long should it take to do a full analysis?
(skipping experiment’s approval process)

Final Analysis

EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP-2012-218

Accepted by: Physics Letters B

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

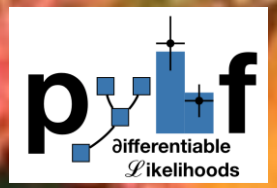
Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s}$ = 7 TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s}$ = 8 TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

arXiv:1207.7214v2 [hep-ex] 31 Aug 2012

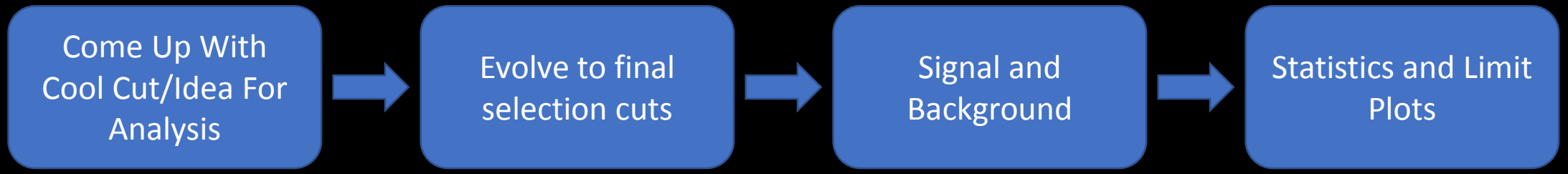
nature

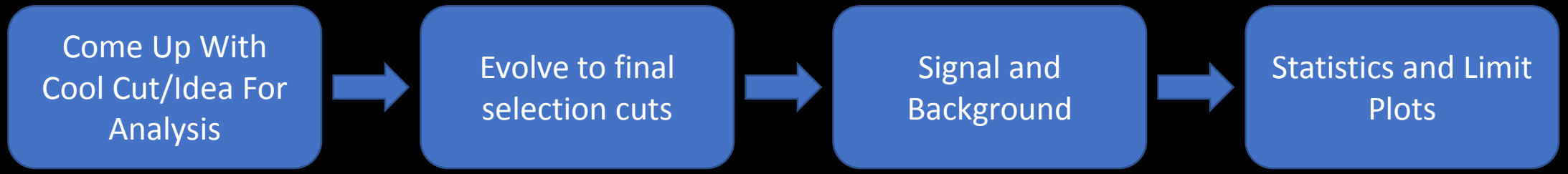
If interactive analysis is fueled by Jupyter notebooks…



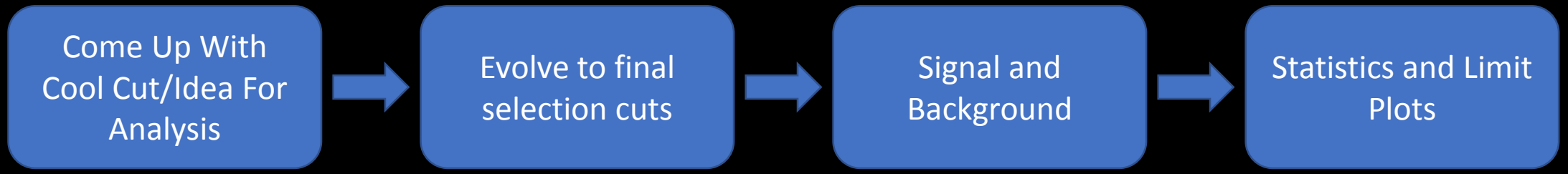
Maybe there is more than one challenge here…


Final Analysis
Interactive Analysis

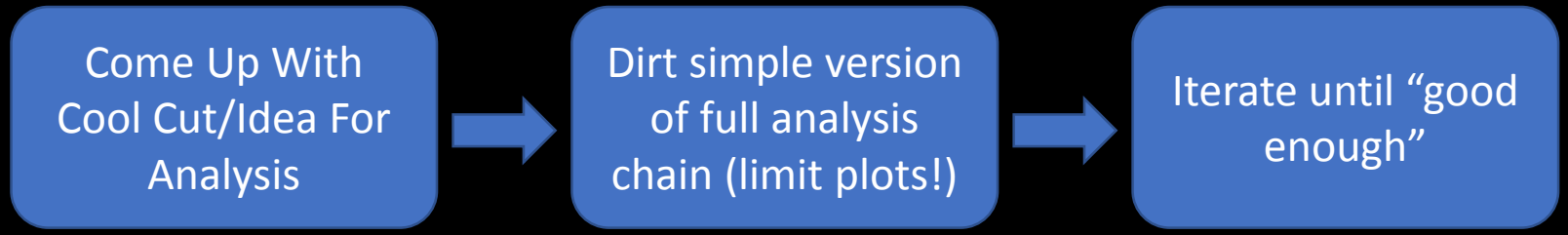

Come Up With Cool Cut/Idea For Analysis
Evolve to final selection cuts
Signal and Background
Statistics and Limit Plots



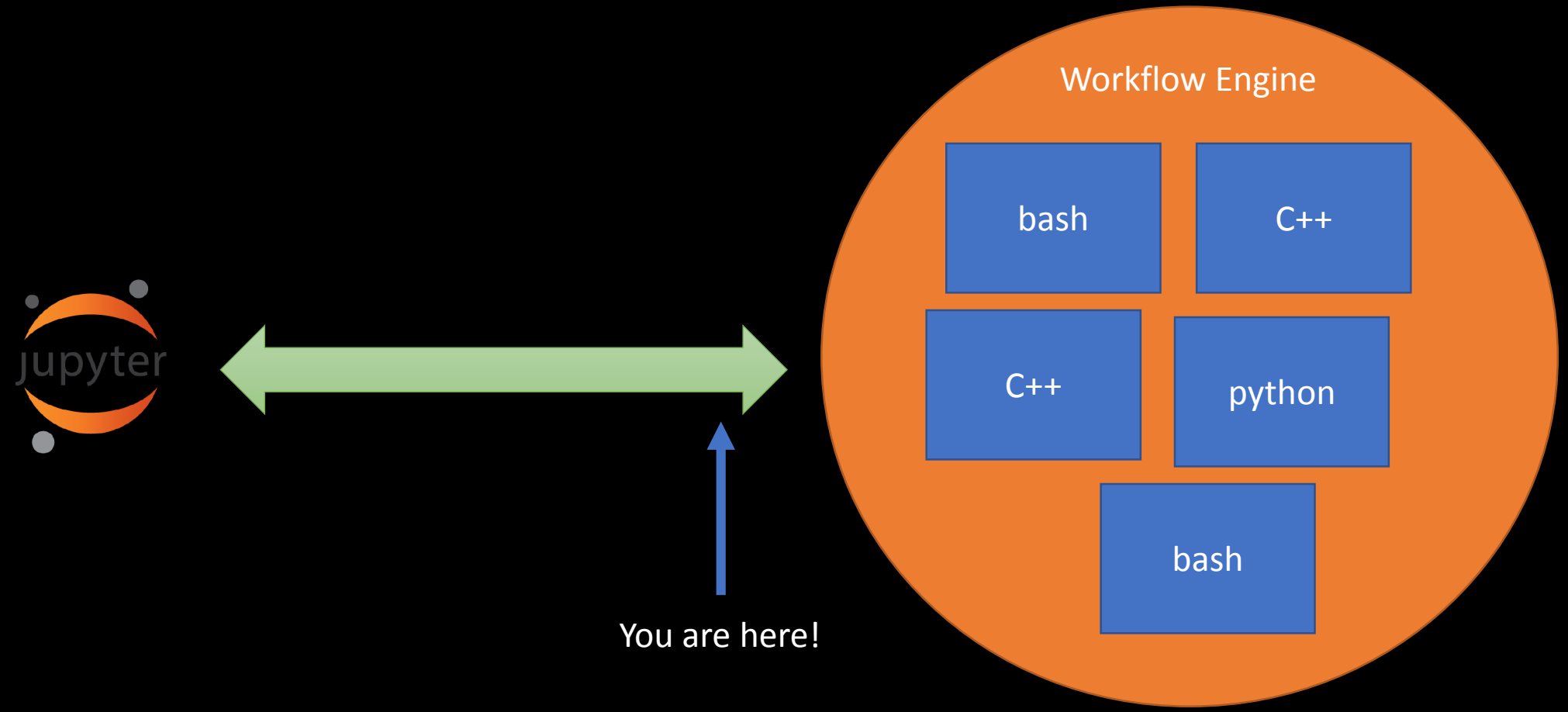
How would Ben do an analysis?
Why in CI, of course!



Don’t forget to be blind!

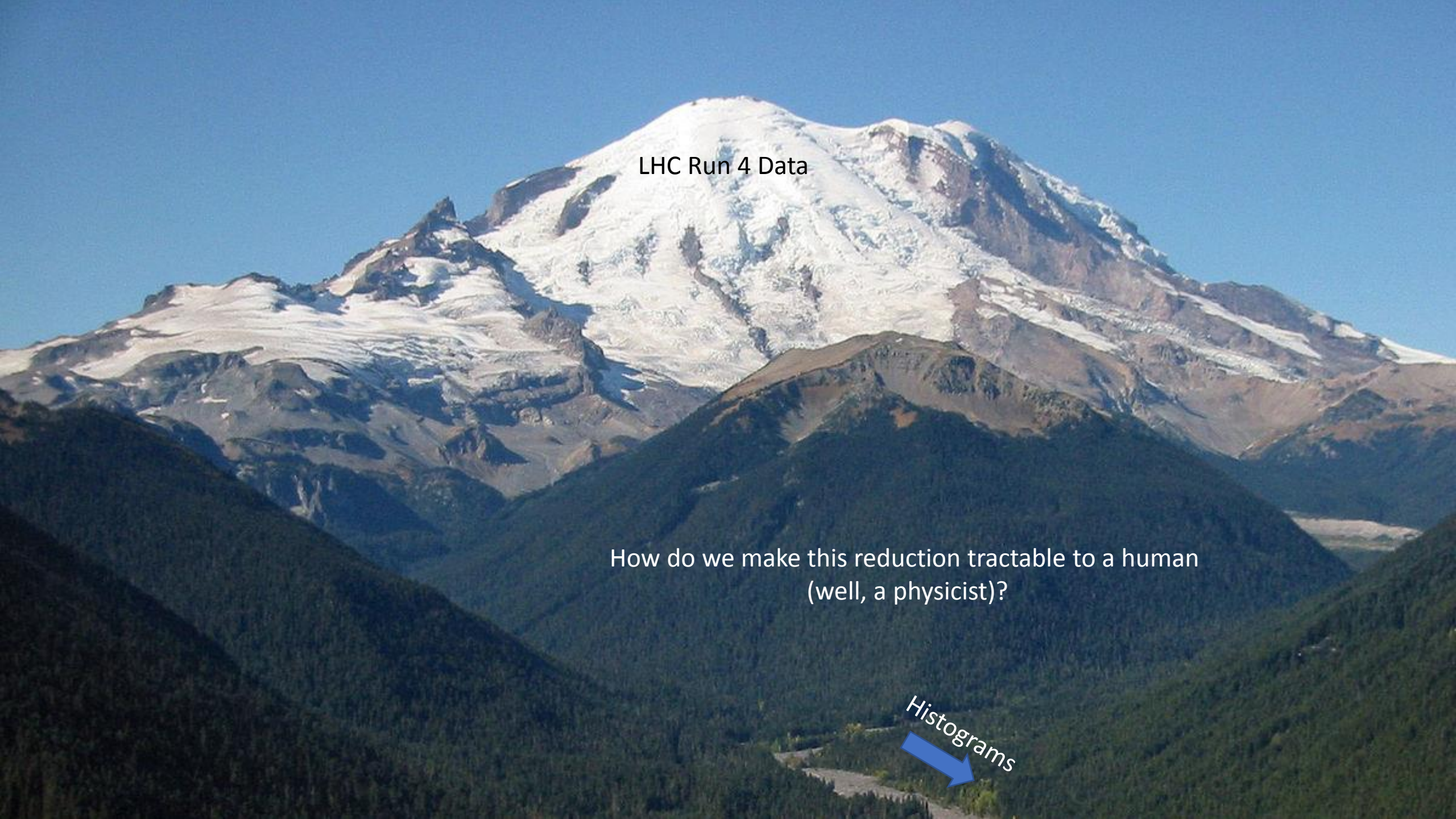


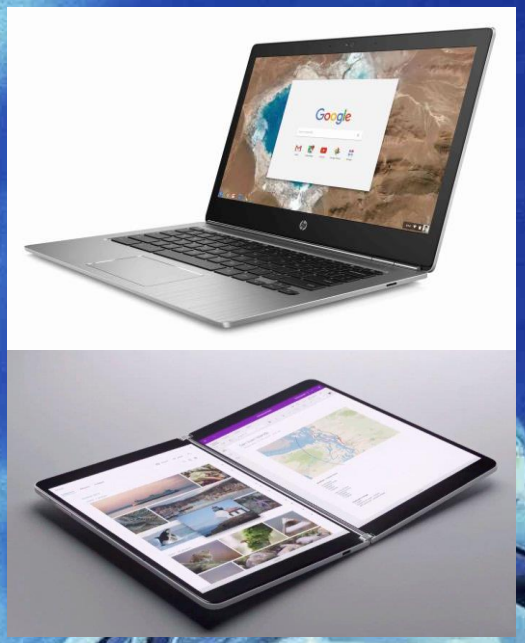
Could this be a goal?
Would this work?



But where should we be?


LHC Run 4 Data
Histograms


LHC Run 4 Data
How do we make this reduction tractable to a human (well, a physicist)?
Histograms

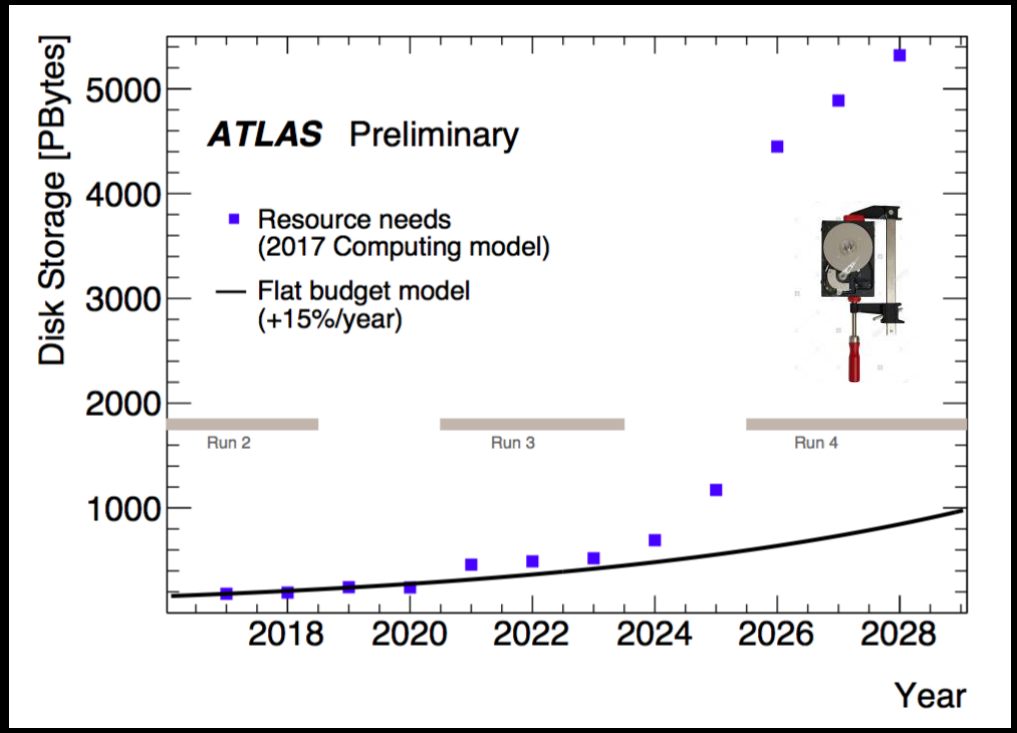

Google

Work Anywhere

Change machines…

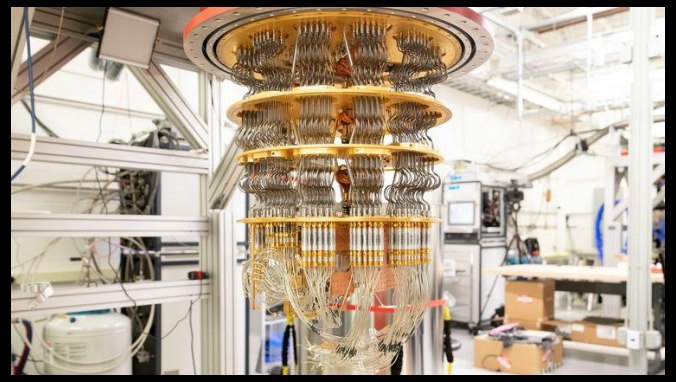
Local Data?

Machine Learning?
(Distributed Learning As A Service)?



Is there something we can do that will change behavior enough to have an effect on this plot?

Conclusions



- What challenges have I missed?
- Part of the reason I knew of these challenges is I've seen people trying to address them
  - There are possible solutions around for all of these
- Do we have a coherent picture?