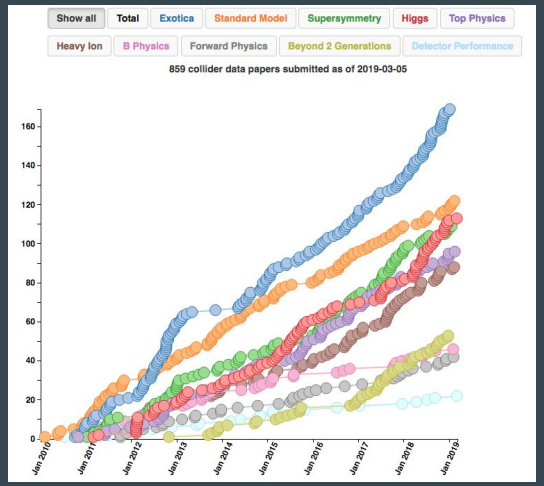
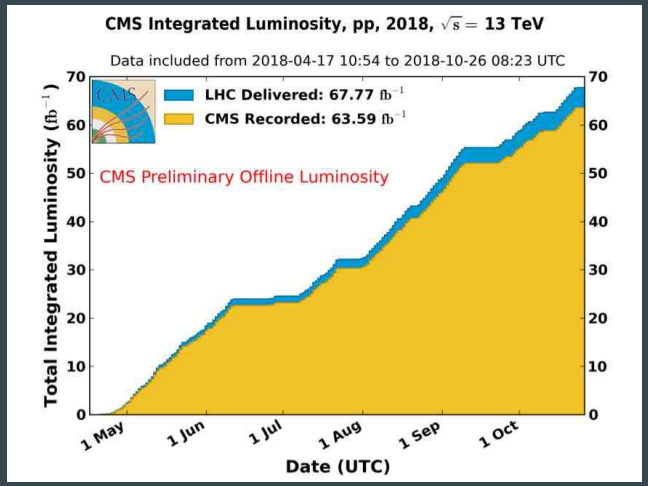
Analysis challenges towards HL-LHC

David Lange
November 2, 2019

LHC and it's experiments have already achieved a lot

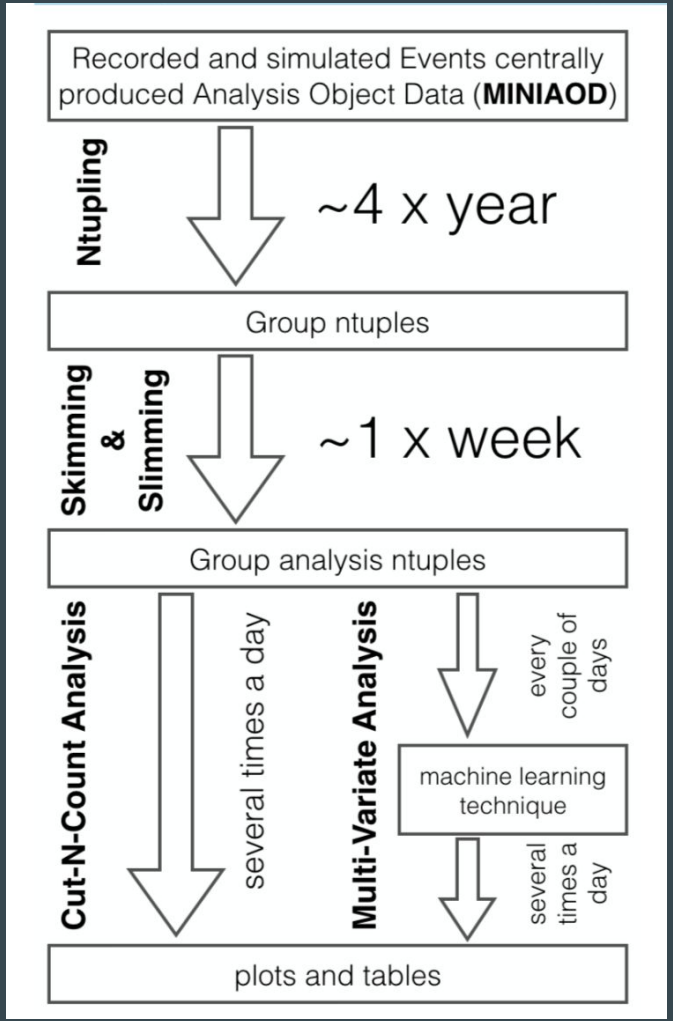
- LHC completed Run 2 at the end of 2018
- Both the two blind and offline proved to be able to maximize the use of Run 2 and could accommodate the unexpected
- Analysis operations is now in full swing...



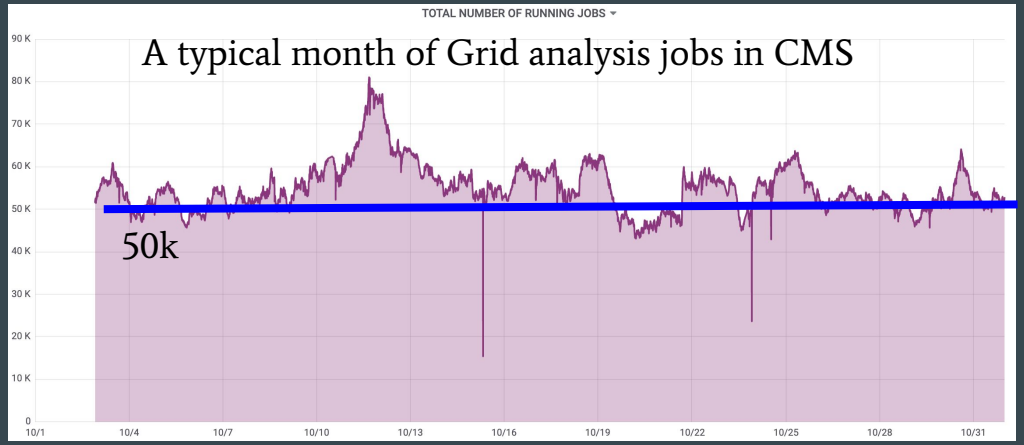


How do we do analysis today?

Centrally produced data is reduced in a multistep process to arrive a format suitable for performing interactive analysis



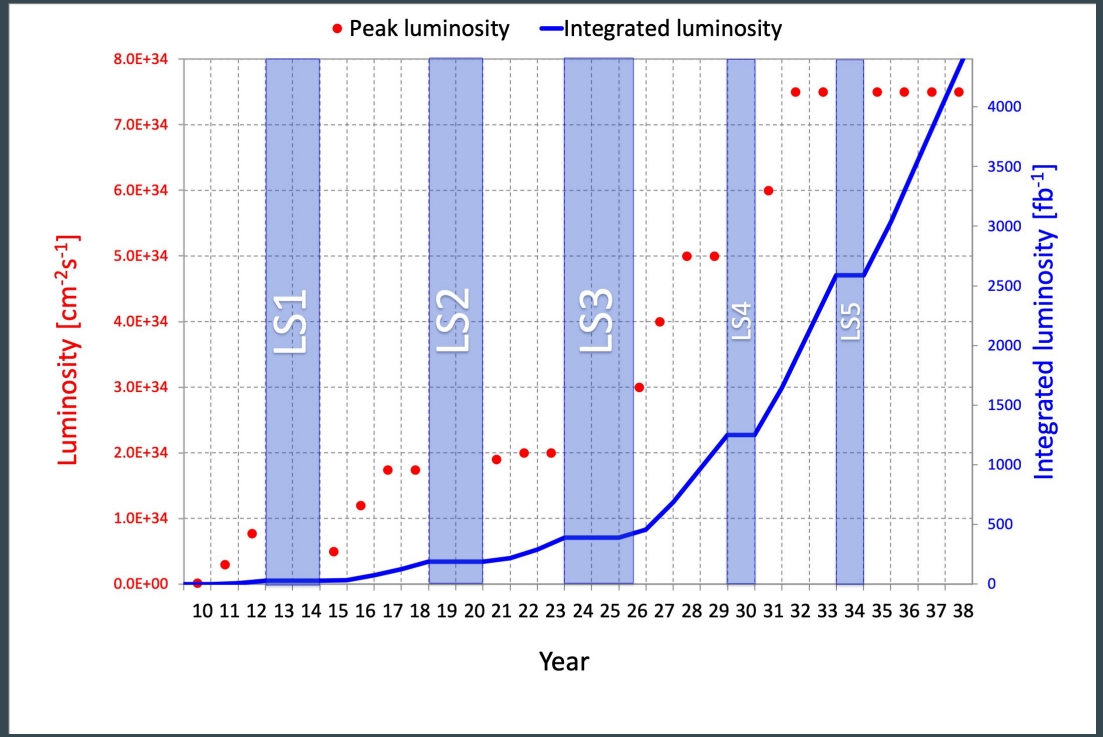
How do we do analysis today?



CMS devotes 50k cores on any given day for distributed analysis processing.

Over time, more and more has been pushed into centrally produced data formats (great!). However, substantial processing is still needed given the Run 2 event rates and having 3 years to analyze together.

HL-LHC brings an entirely new scale for analysis

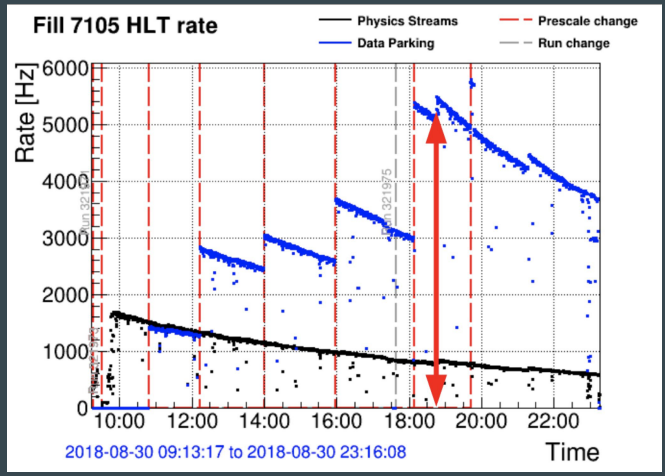


What data rates are expected?

Naive extrapolation go from L=1.9e34 (@PU 50-55) up to L=7.5e34 (@PU 200) explains the expected need of HLT output at ~ 7.5 kHz, mostly coming from single object triggers

This sets the scale of event rates to be analyzed

- Unless we want to reduce / descope a part of the physics program
- Unless we can use less inclusive trigger approaches - to be studied, but “failed” for RunII



We don’t.. we want to maximize the physics program

What does this mean for analysts?

To spin over a year of data raking...

50 TB of data for each kB/event of data tier size

600 CPU days for each millisecond/event of processing needed

[Today seconds is a more appropriate unit than ms]

Eventually analysts will want to process 10 years worth of data together...

Does this scale towards HL-LHC?

Like many other components it's hard to imagine the current approach to analysis scaling up by factors of 10 in event rates and event sizes

Fortunately there are substantial opportunities to modernize the HEP approach to data reduction and analysis

Strengths

HEP analysis methods are robust and mature. Our tools have enabled a huge range of run 2 analysis

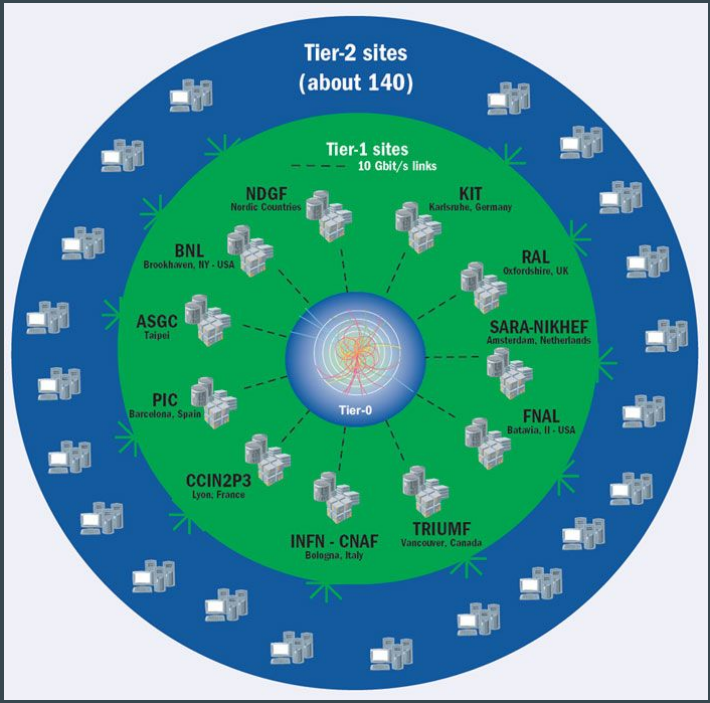
I won't talk about these.

Instead let's discuss opportunities we have to scale towards HL-LHC scales

Distributed, many user, and rapid?

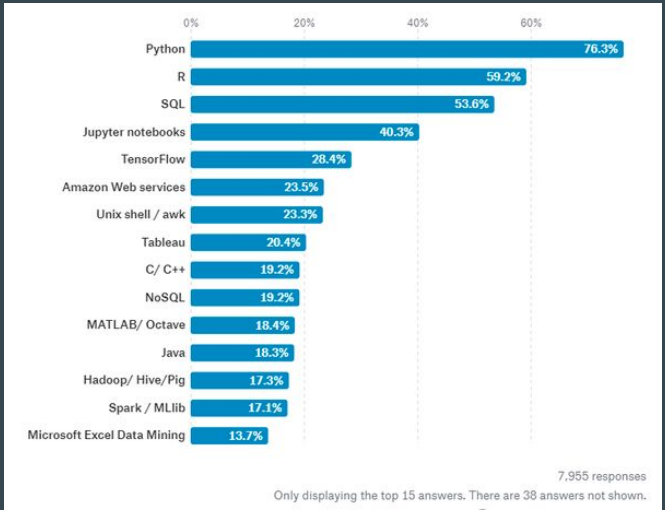
How do systems avoid duplication?

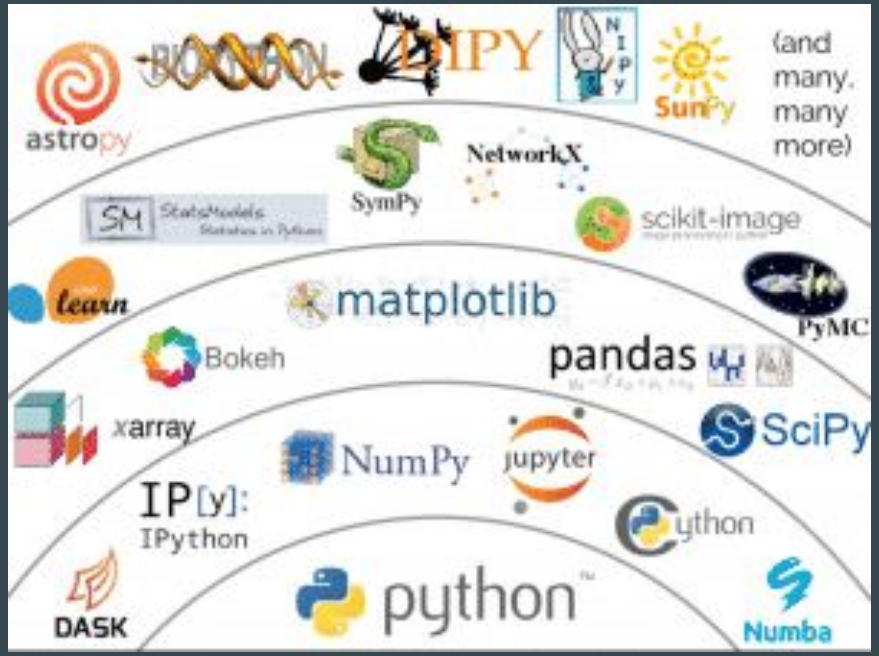
Fulfill R&D promises while scaling up from 1 to many users (who are likely to ask for resources in correlated ways)



Opportunities with data science community

We have all seen the rise of open source tools in the wider data science community. Fortunately HEP adopter Python early on

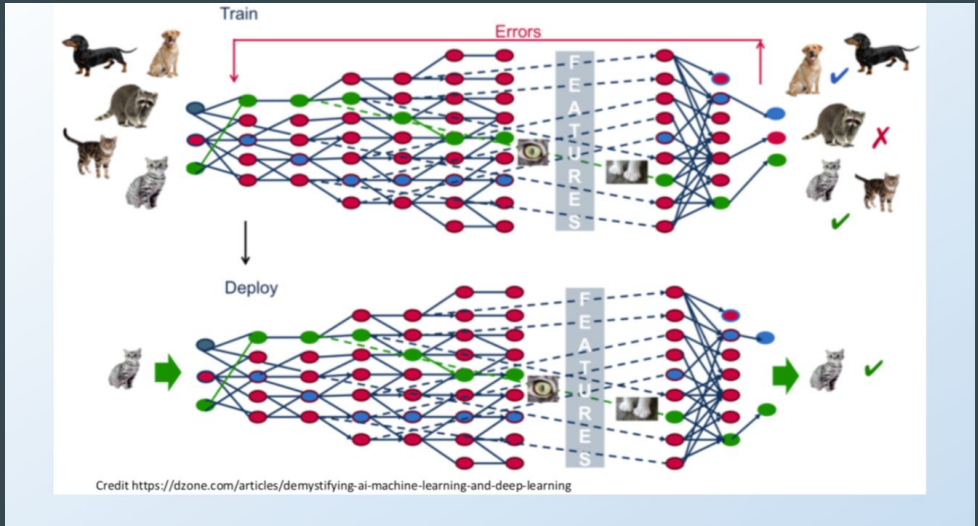




Opportunities with AI

Will training neural networks become the primary activity of HEP researchers?

Will adoption increase the analysis processing needs rather than reduce?



Opportunities with simulation

A bigger reliance on “fast” simulation may be required to fit physics scope into resource budgets for HL-LHC

Hopefully the loss of modeling accuracy is small. If not, how can tools help analysts cope with a Monte Carlo vs data agreement that is not as good as today?

Deep Learning for fast sim

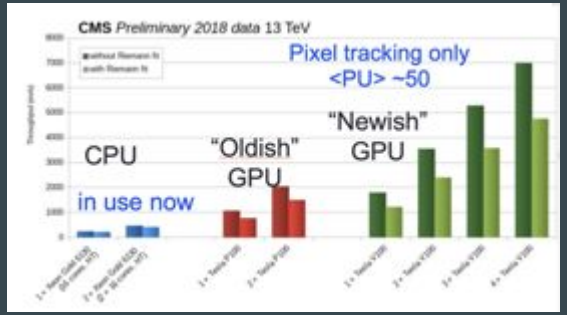
S. Vallecosra (ACAT17)

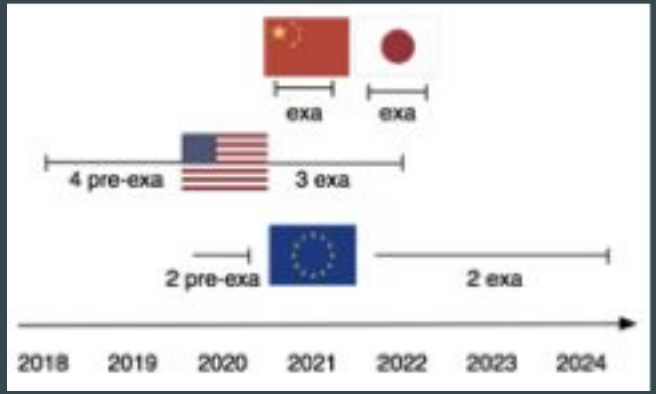
- Generic approach
- Can encapsulate expensive computations
- DNN inference step is generally faster than algorithmic approach
- Already parallelized and optimized for GPUs/HPCs.
- Industry building highly optimized software, hardware, and cloud services.

Opportunities with heterogeneous hardware and HPC

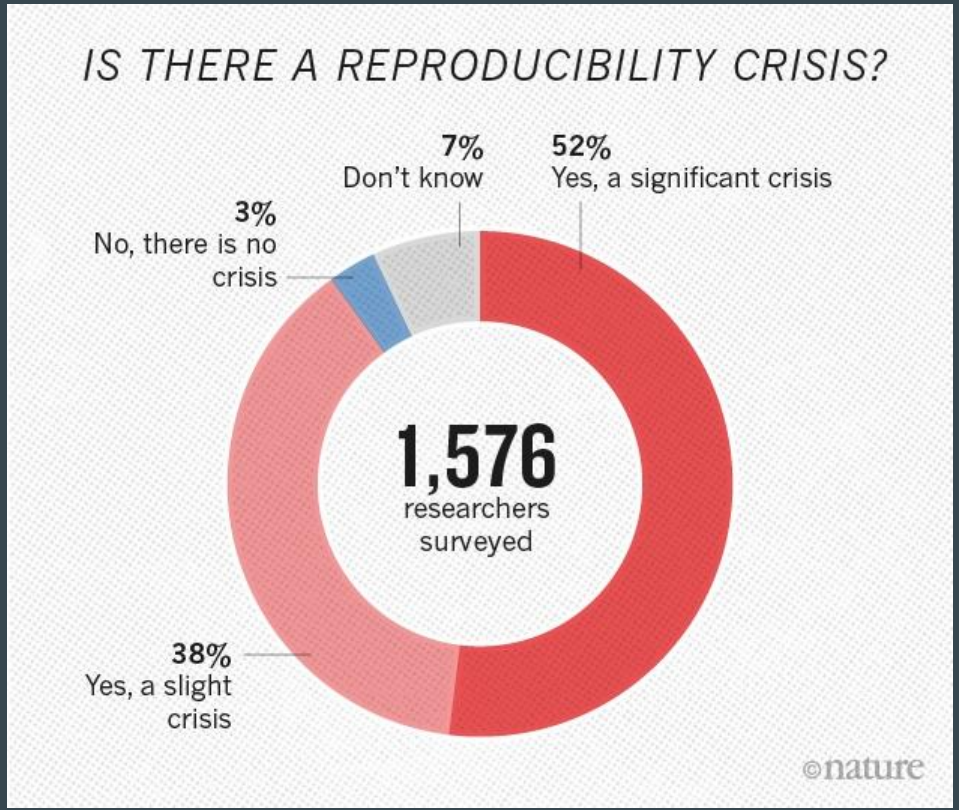
How will analysis codes follow the lead of reco/trigger?

Of course most machine learning toolkits give this capability for free





Opportunities with reproducibility



Finally - Facilitating the transition to widespread use

HEP researchers will need help to transition their tools to take advantage of today's R&D results that should help them scale up their work to HL-LHC datasets

We should be sure to include that adoption in our planning