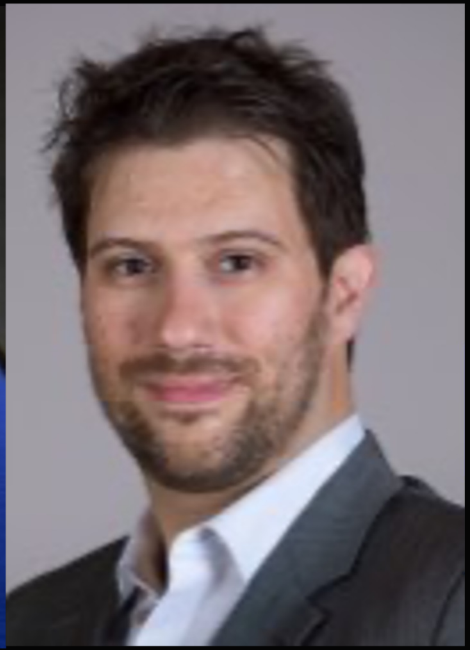
Evolving Research Software towards Next-Generation High-Energy Physics Experiments

David Lange

Princeton University

Introductions

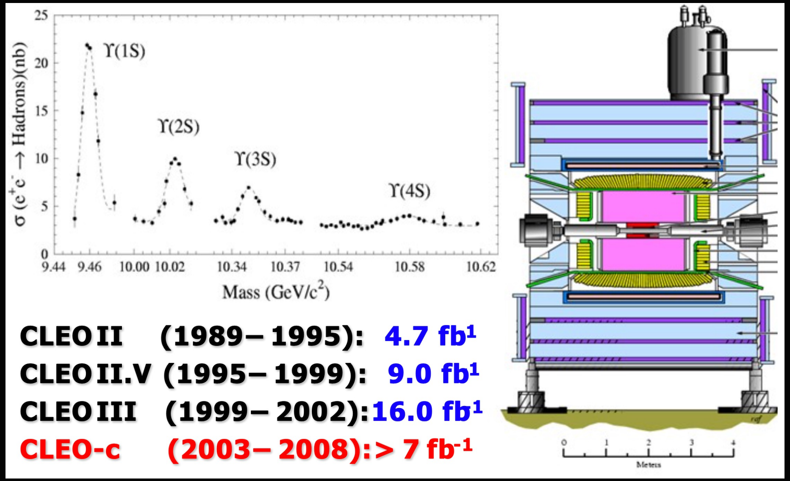


Princeton University: Peter Elmer, David Lange (PI)
University of Massachusetts, Amherst: Rafael Coelho Lopes de Sa, Verena Martinez Outschoorn

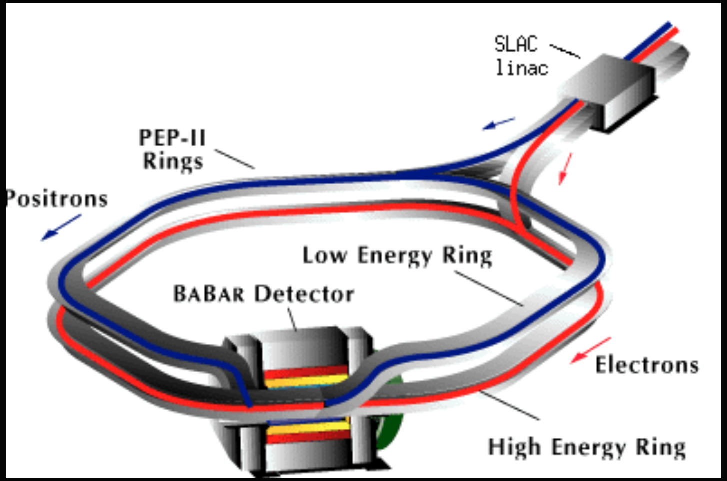
My career in high-energy physics

CLEO-II(.5) @ Cornell



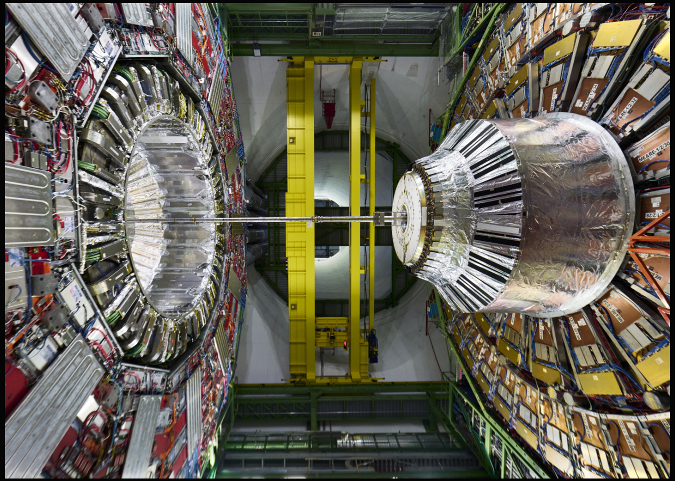
- Silicon detector calibration + operations
- Form-factor analysis in semileptonic decays
- Event generators

BABAR @ SLAC



- CP violation analysis
- RPC detector operations + software
- Event generators (“EvtGen”)
- Event reconstruction software

CMS @ CERN



- Event reconstruction software
- Simulation techniques
- Computing resource projections

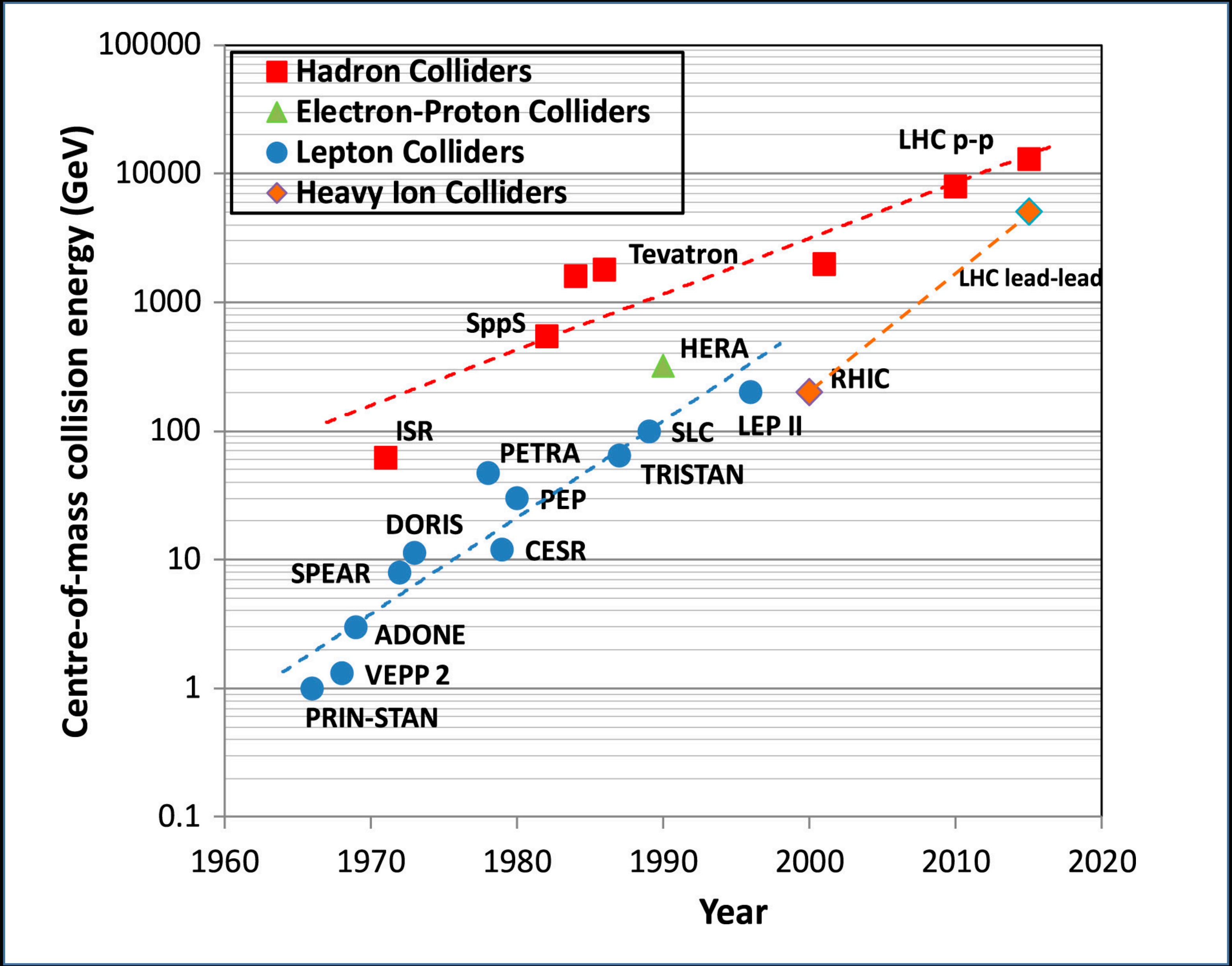
My Goal for Today

- To tell you about experiences working in research software endeavors and how software has however also become a critical element to design and maximize the physics discovery potential of large data intensive science projects.

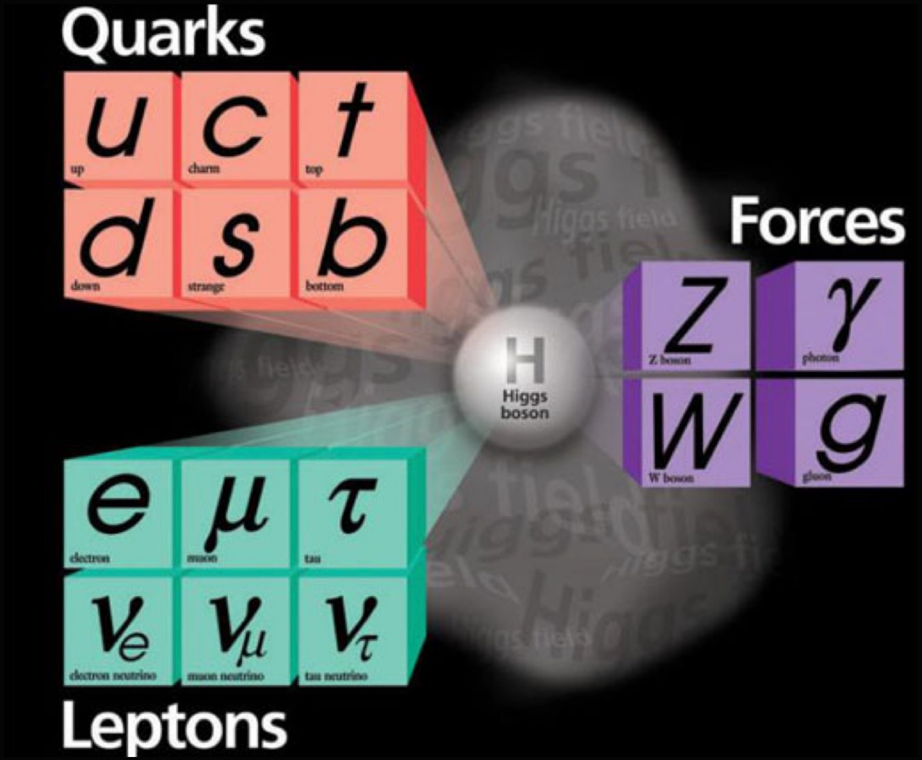
- To convince you of the importance of these research software collaborations to scientific discovery, and to discuss new initiatives for fostering new collaborations.

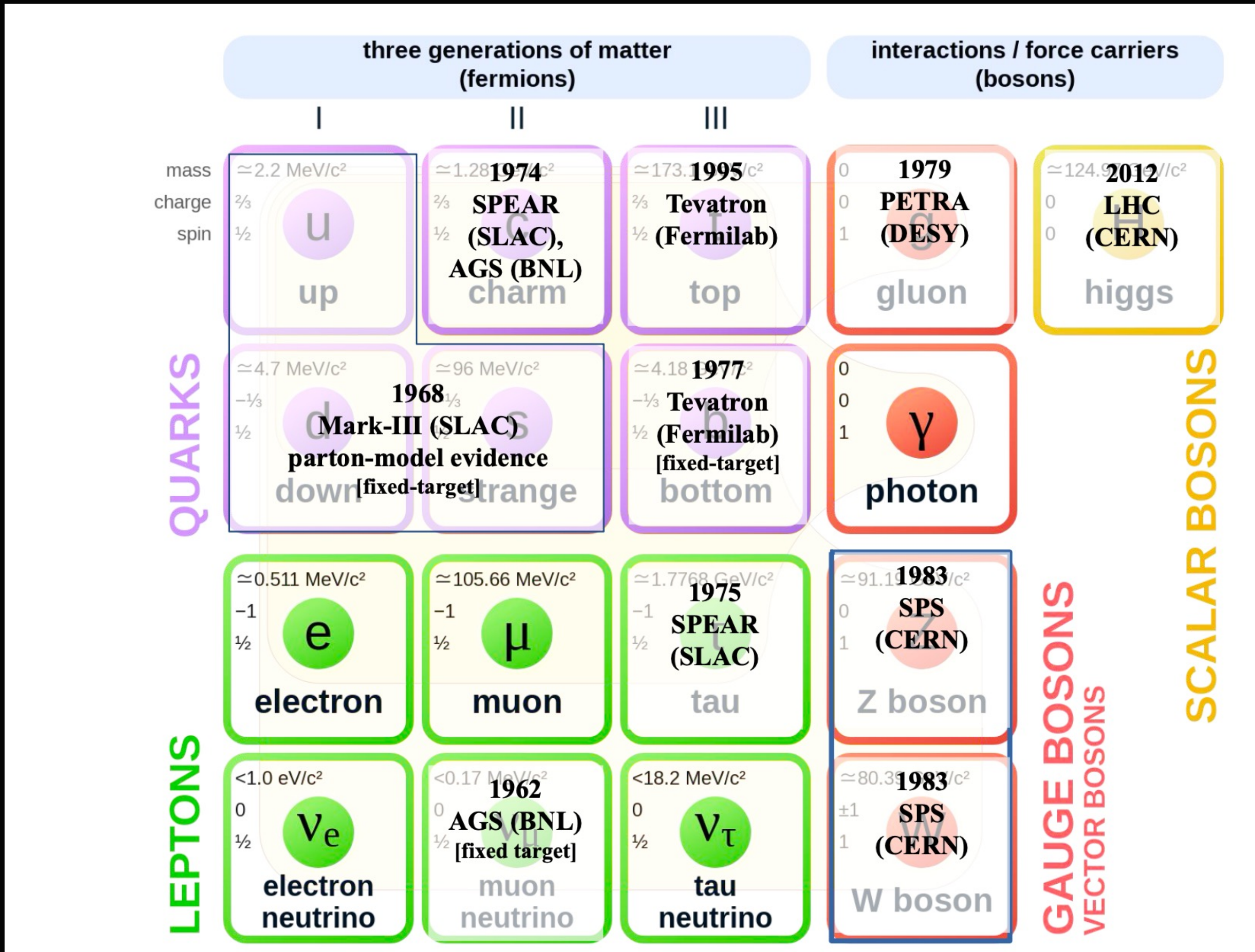
- Nota bene: Examples are often biased towards energy-frontier experiments, and specifically CMS, as that is what I am most familiar with. However it is typically easy to find similar examples elsewhere

High Energy Physics is a facilities driven science



These Facilities allow scientists to discover and test the Standard Model of Particle Physics

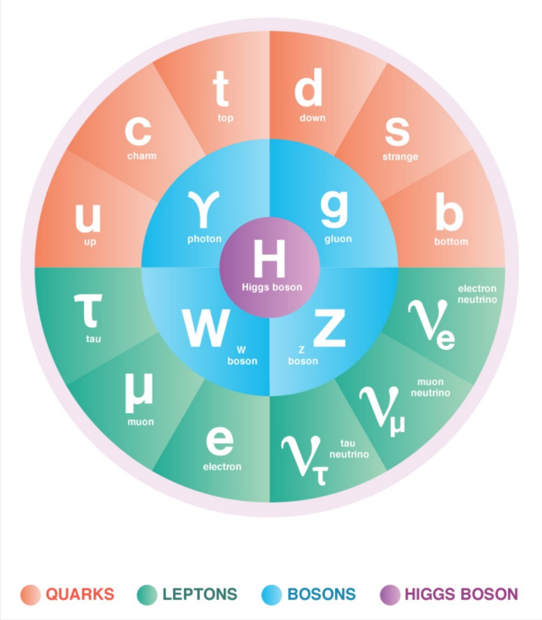




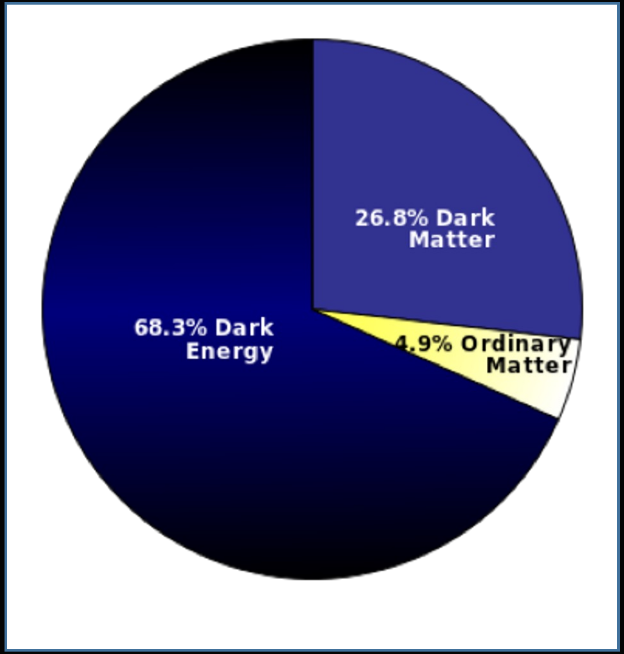
Adapter from source: Wikimedia

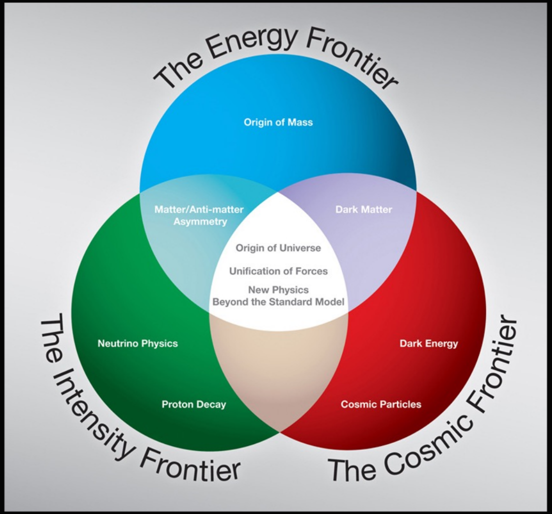
- Tested over more than **30 orders of magnitude**!
  - Photon mass < 10^{-18} eV
  - LHC probes > 10^{12} eV

Science questions continue to drive these facilities



While the Standard Model of Particle Physics describes, often with incredible precision, the vast majority of experimental observations, it is known to be incomplete. It does not (for example) include gravity, and it does not explain neutrino masses, the matter-antimatter asymmetry or dark matter/energy.

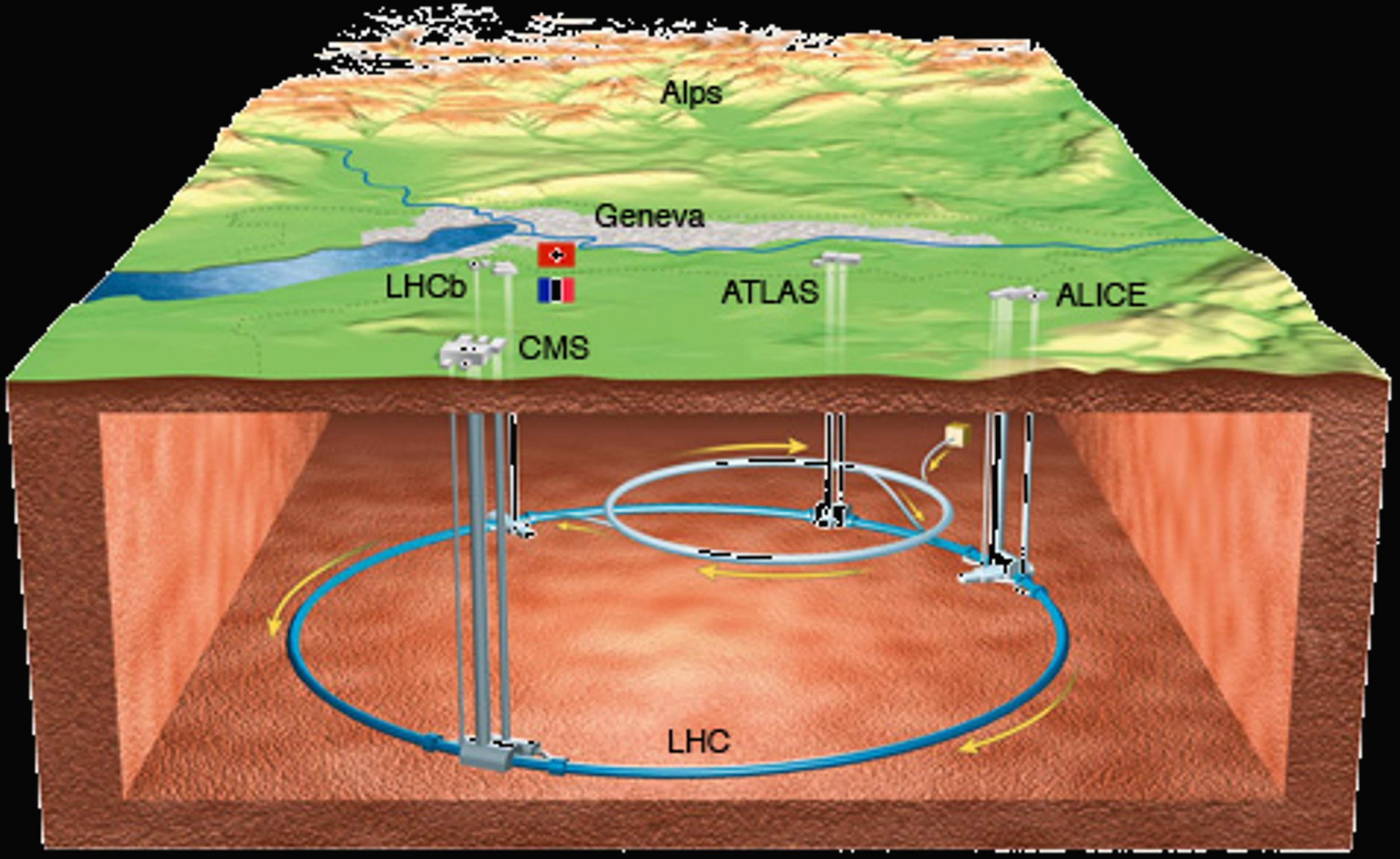


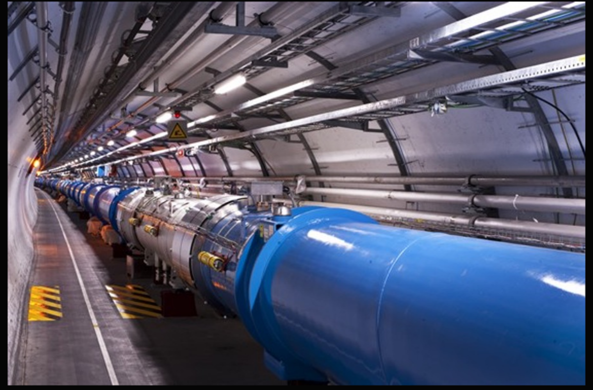


From "Building for Discovery - Strategic Plan for U.S. Particle Physics in the Global Context" - Report of the Particle Physics Project Prioritization Panel (P5):

1) Use the Higgs boson as a new tool for discovery
2) Pursue the physics associated with neutrino mass
3) Identify the new physics of dark matter
4) Understand cosmic acceleration: dark energy and inflation
5) Explore the unknown: new particles, interactions, and physical principles

Large Hadron Collider



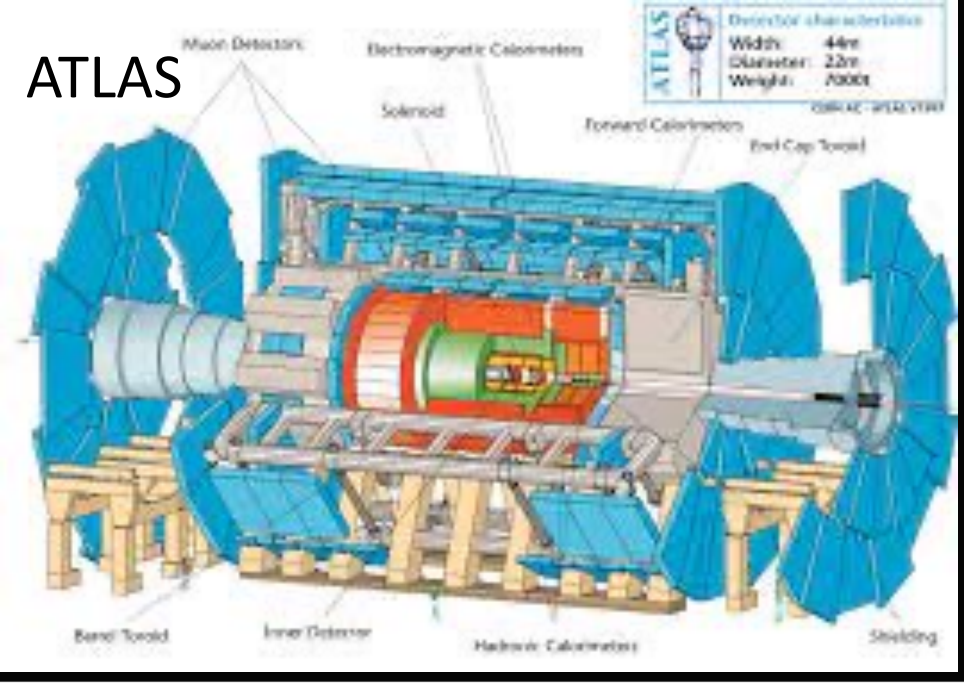


27km circumference tunnel, with larger caverns at points around the ring for detectors

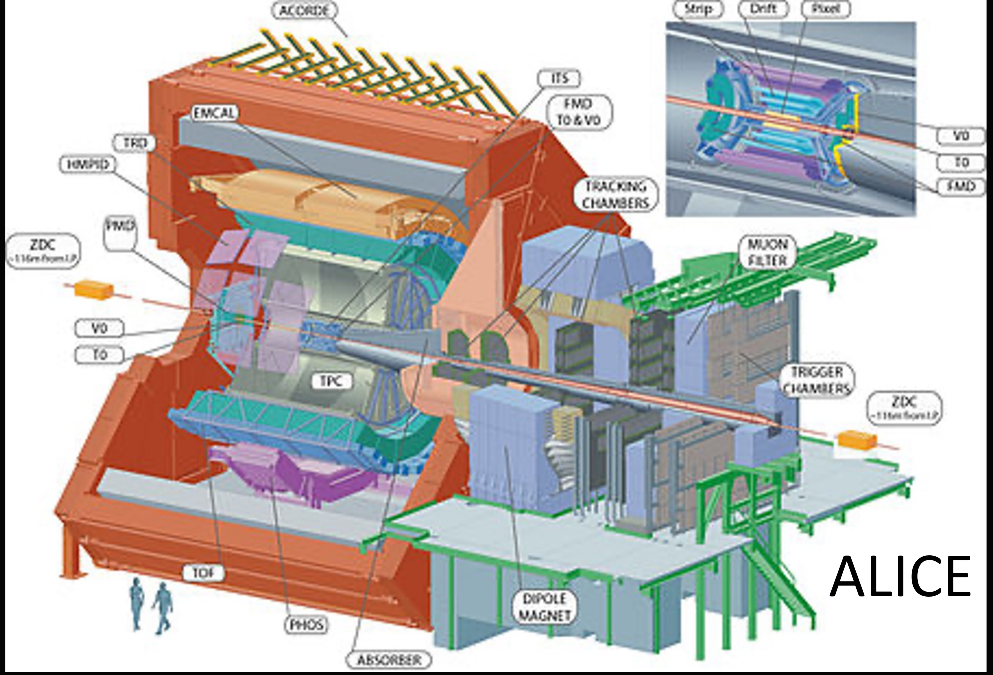
100-150m underground

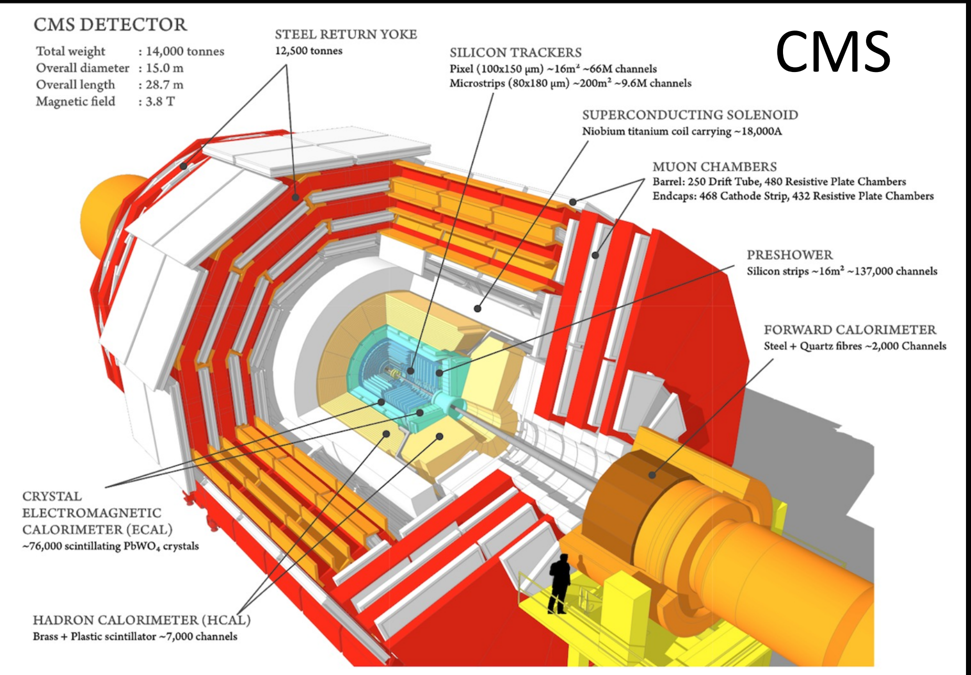
Built in the 1980s for a previous collider (LEP) and expanded with new caverns and access in the early 2000s

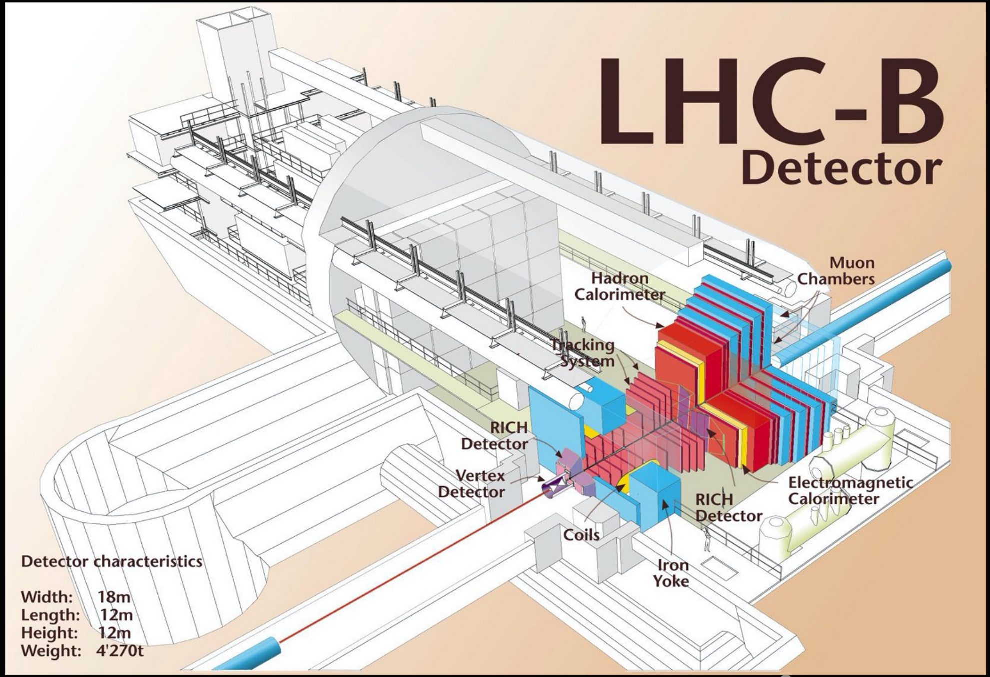
Highest energy collider (design 14 TeV, operating at 13.6TeV) currently available

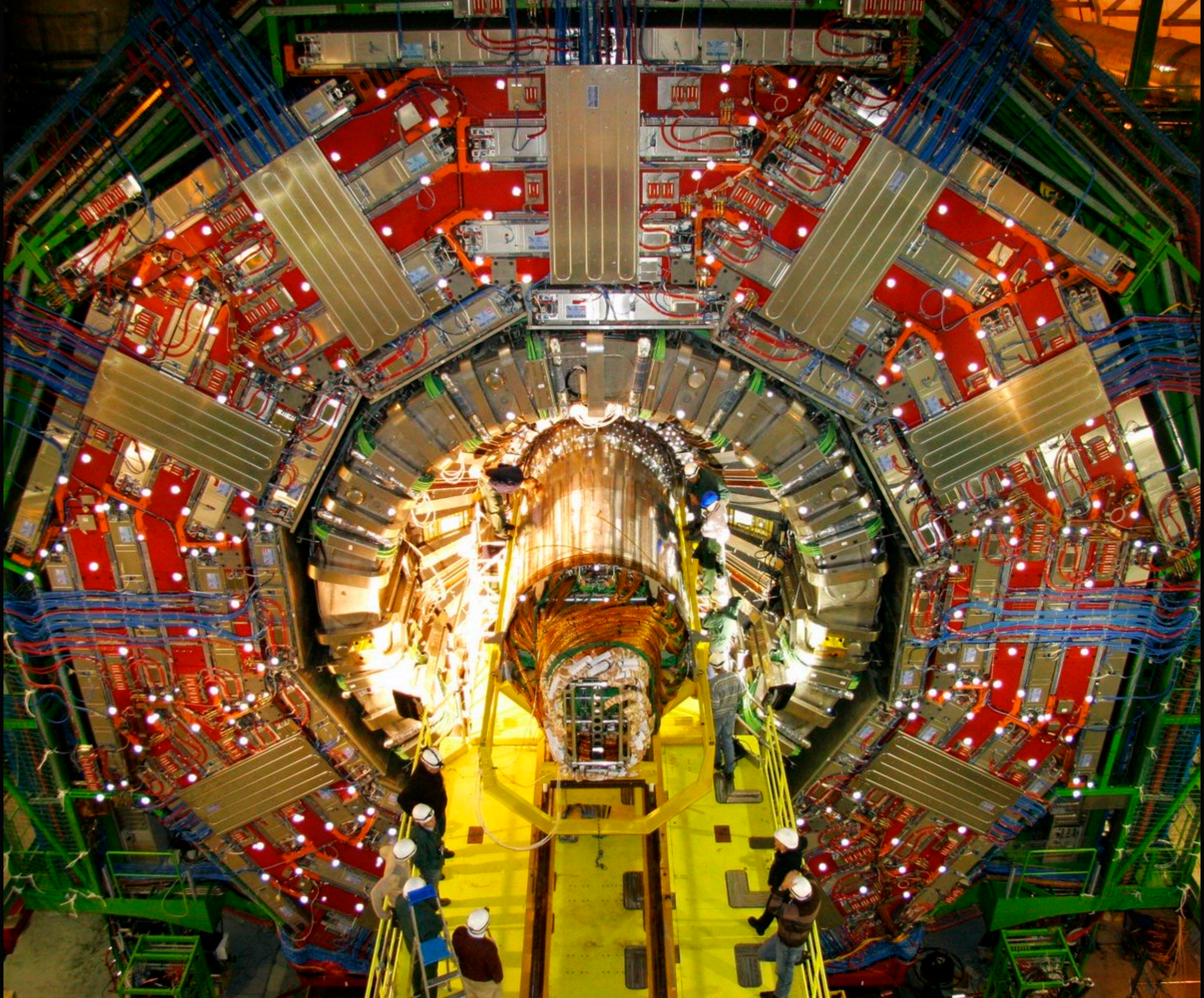


The 4 LHC Experiments



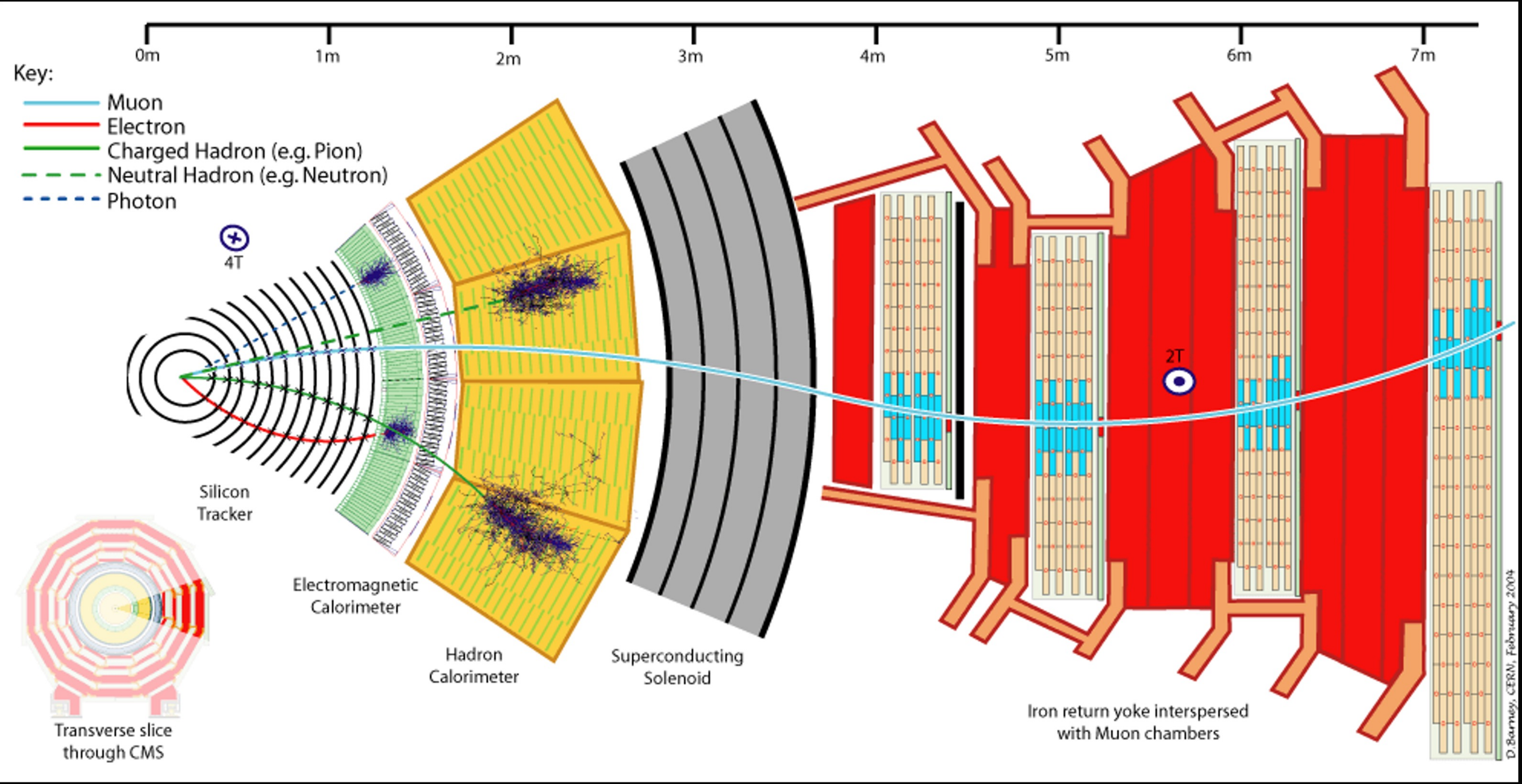






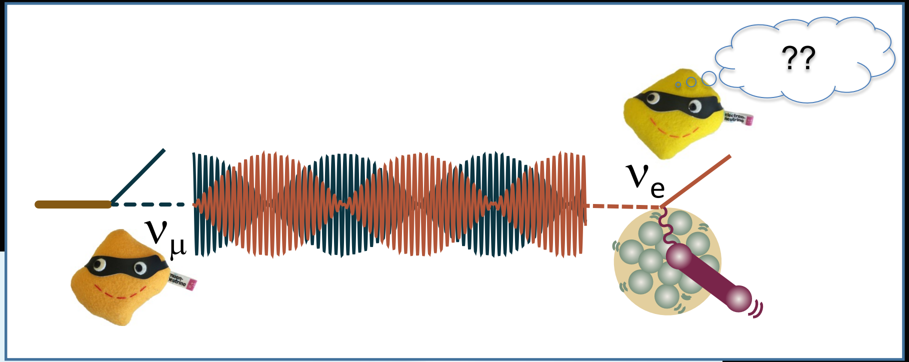

CMS has ~4300 Scientists, Engineers and technicians (includes
800 PhD students) from 41 Countries and 179 institutes

The “onion” of the CMS detector



The DUNE experiment is designed to understand properties of neutrinos

Propagation can change a muon type neutrino into an electron type neutrino

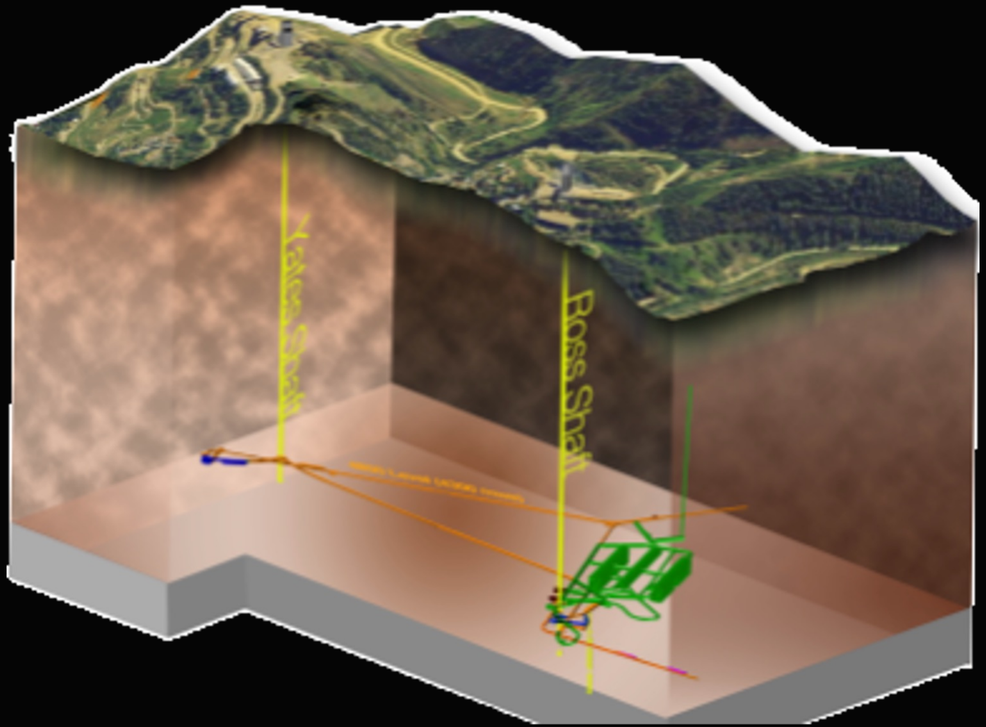


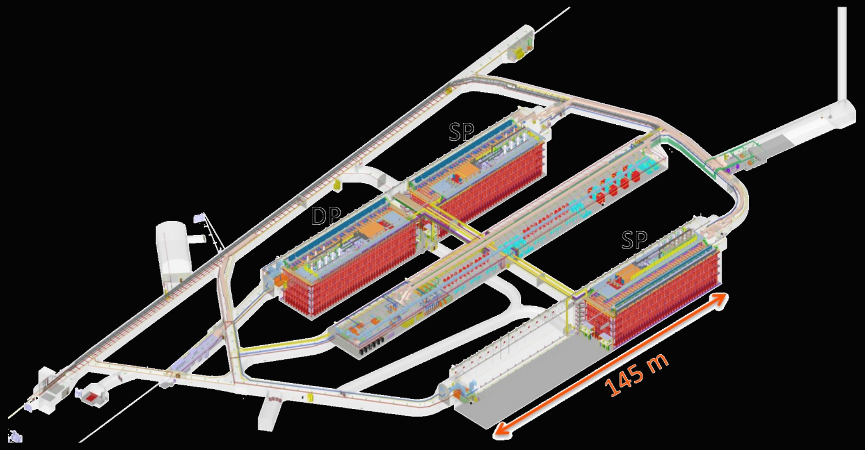
Put a huge LAr detector "DUNE" in the Homestake Gold Mine

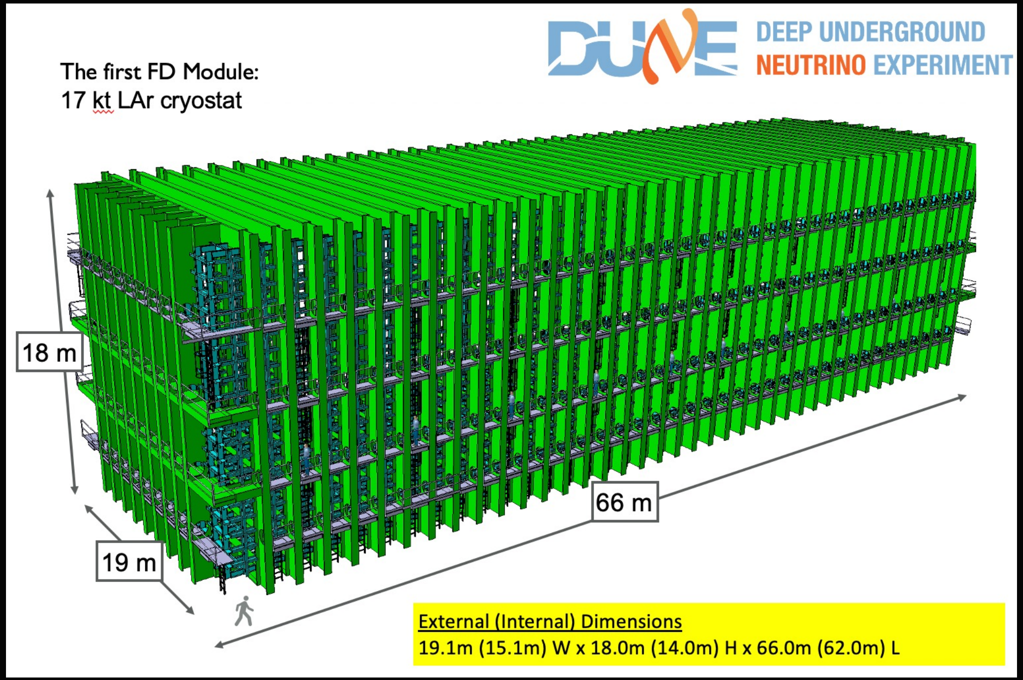
Make a very powerful neutrino beam

Run for 10 yrs

Far Site detector facility under construction

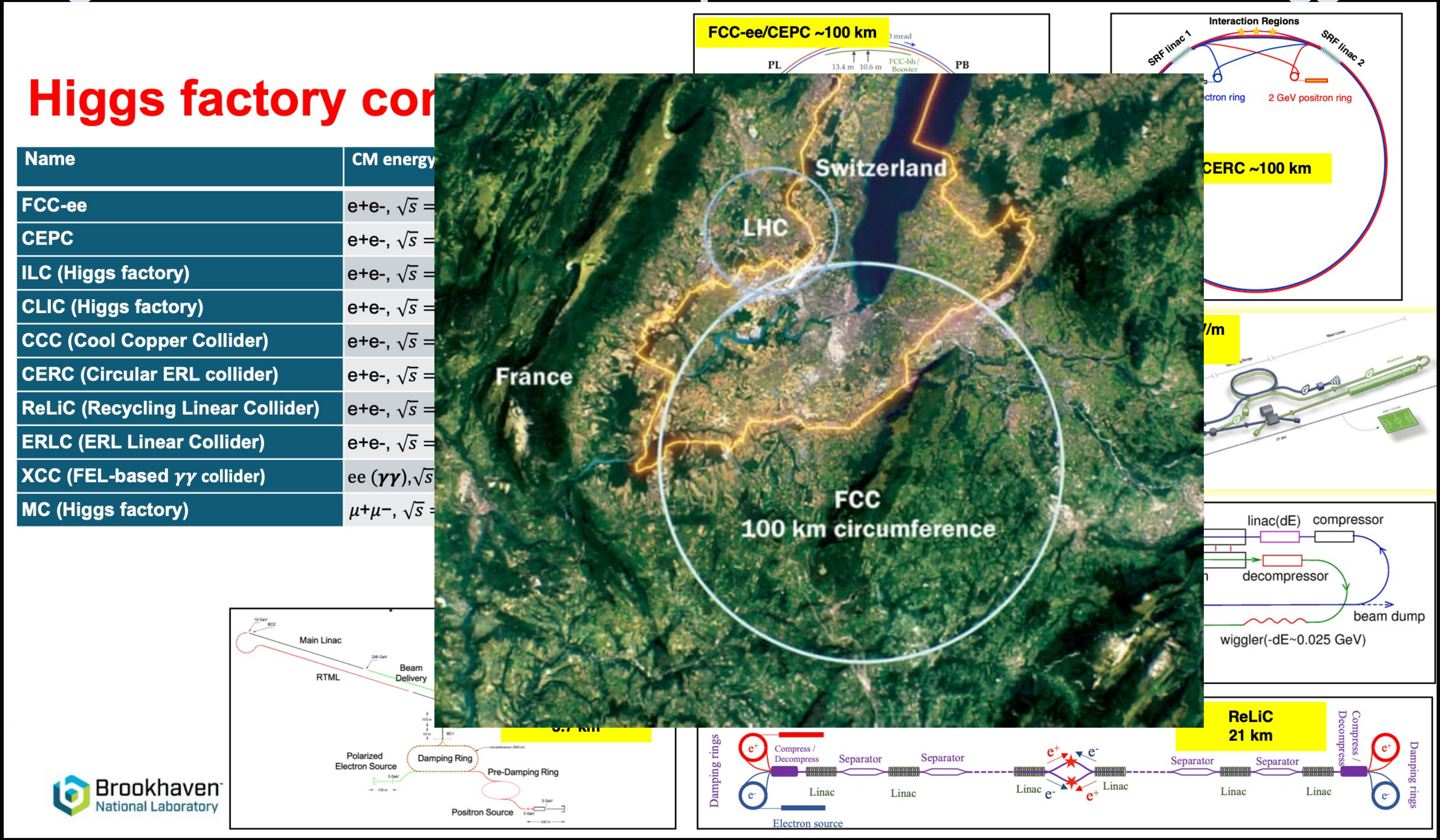




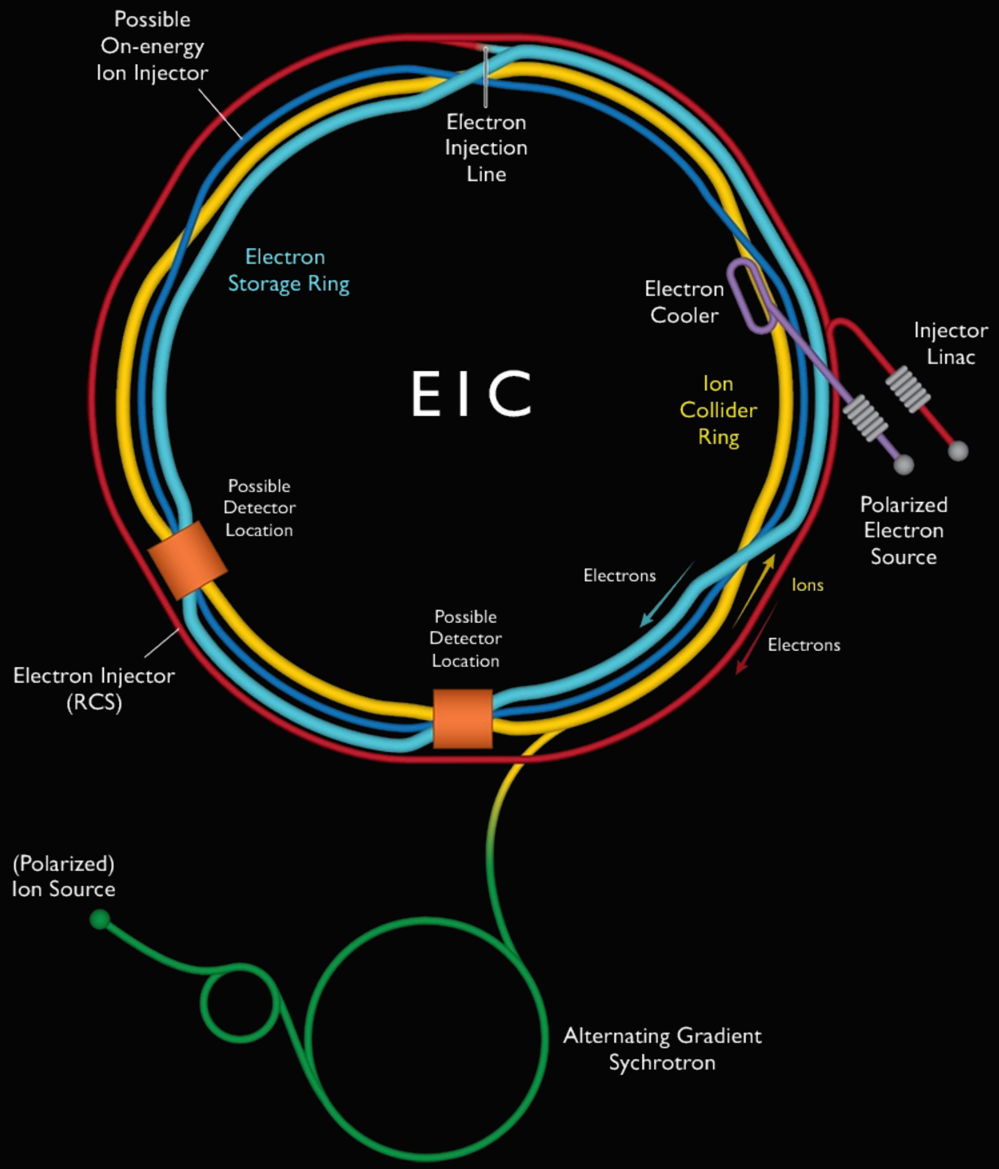


- Major underground excavation removing ~800,000 tons of rock
- Two large caverns housing **four** cryostats and a central utility space
- $4\times17{,}000$ **tons of LAr to fill the cryostats:** ***the target for neutrino interactions***

Next-generation colliders allow precision science with Higgs



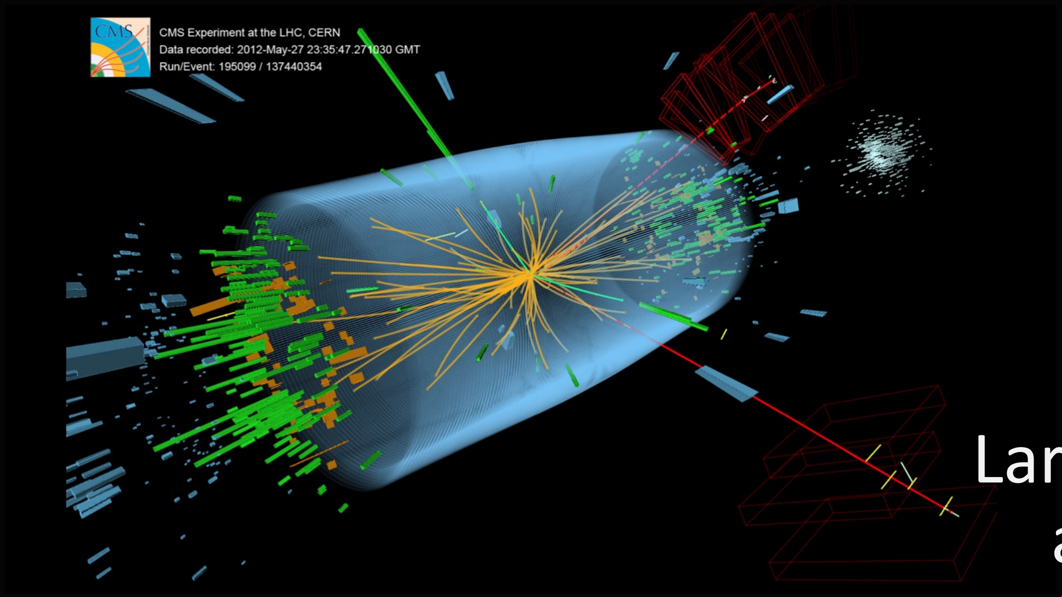
Electron-Ion Collider



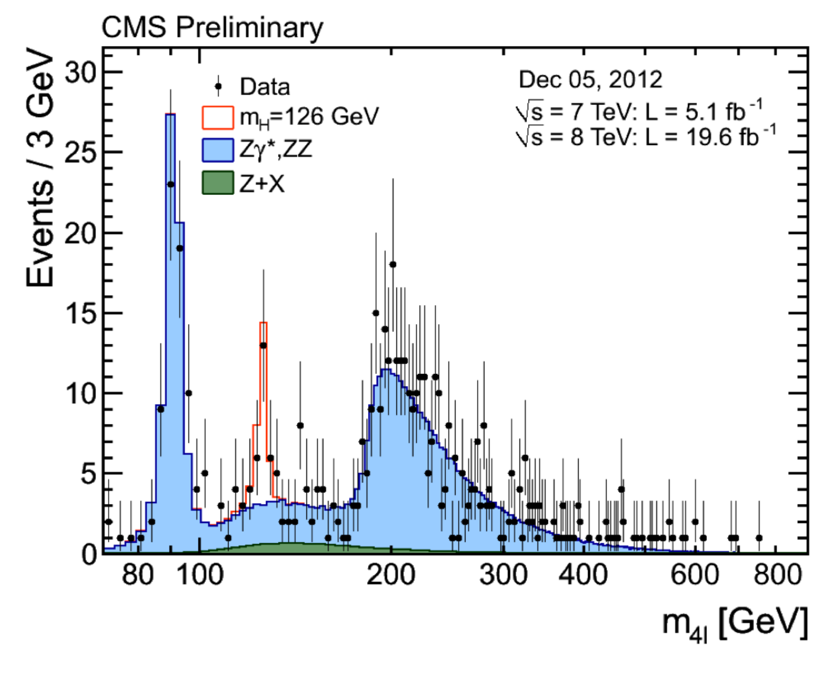
The EIC (to be built at Brookhaven National Lab) will collide electrons and ions to probe quarks and gluons with sub-femtometer resolution to address fundamental questions

- Precision 3D imaging of protons and nuclei
- The proton spin puzzle
- How does the glue bind us together?

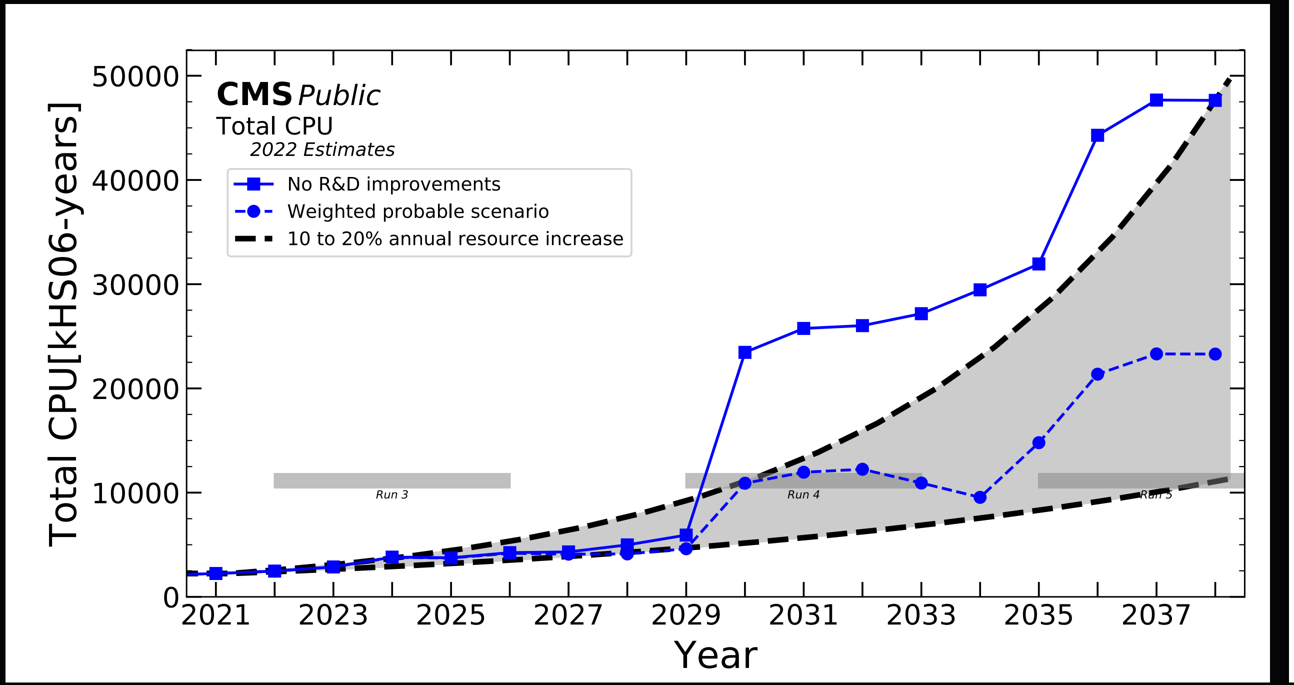
Just as with facilities, HEP scientists rely on large computing infrastructures to do their science

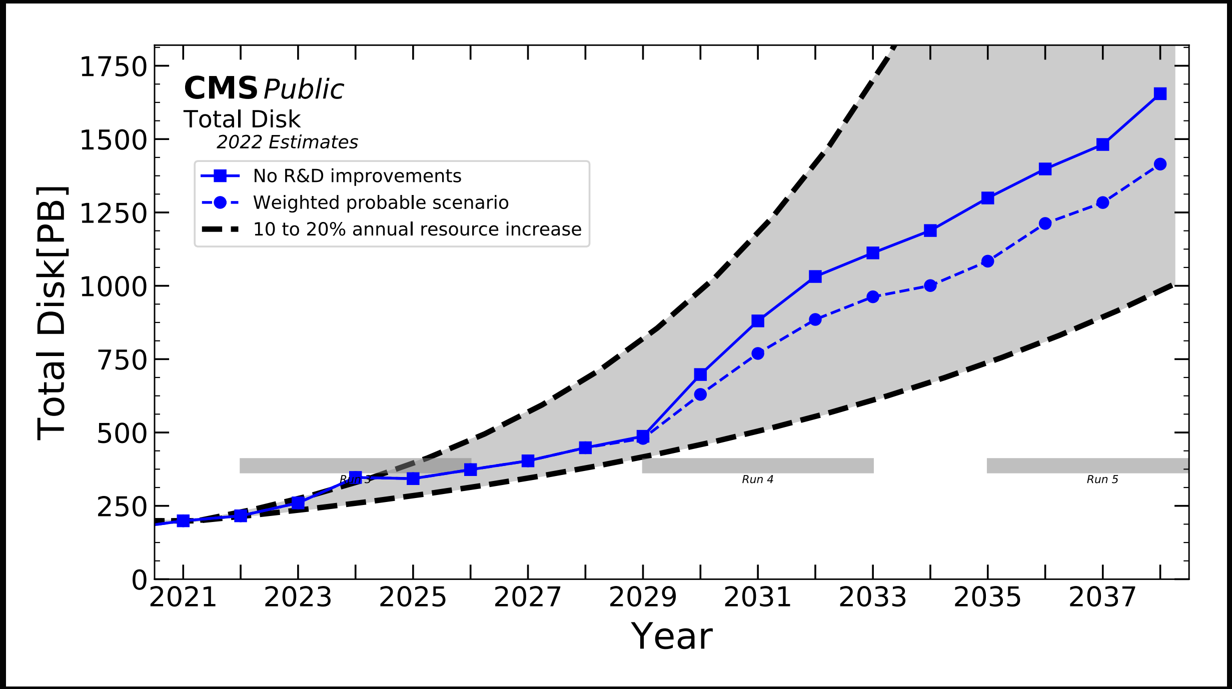


Large scale software and computing infrastructures



Just as with facilities, HEP scientists rely on large computing infrastructures to do their science





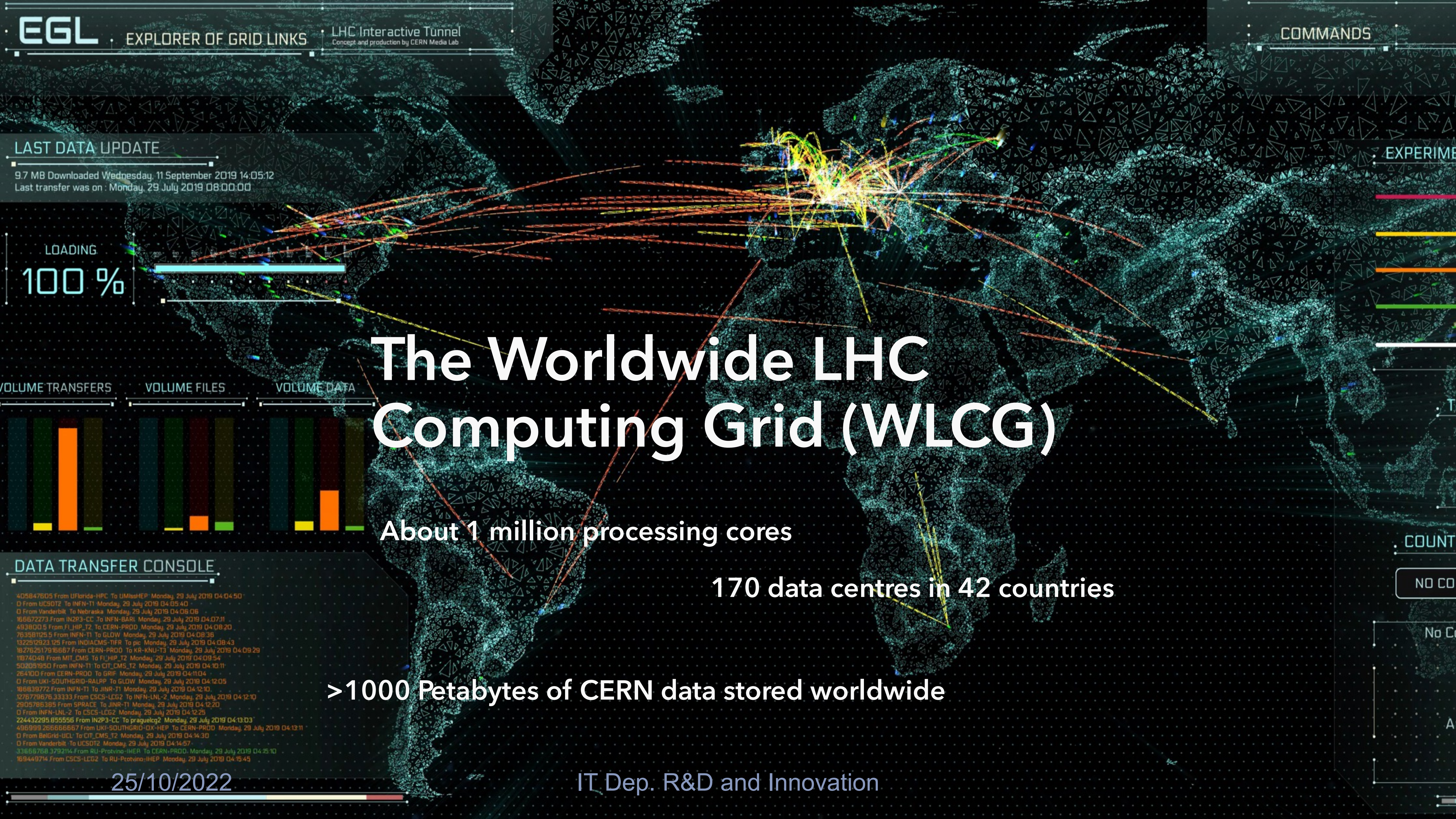
- Today's experiments require O(100k) compute cores and O(100) PB fast-access storage to do data processing and to make data available to analysts

The Worldwide LHC Computing Grid (WLCG)

About 1 million processing cores

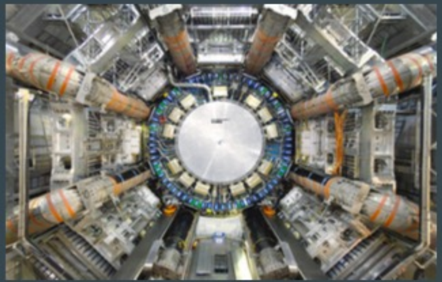
170 data centres in 42 countries

>1000 Petabytes of CERN data stored worldwide



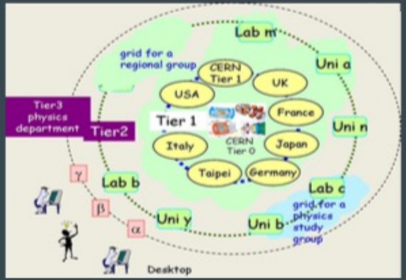
Data flow from collisions to analysis results

Detector

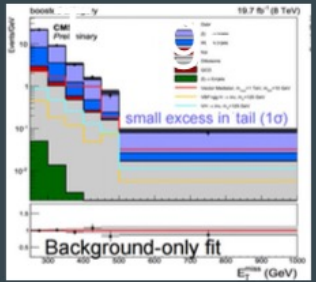


Trigger

Reconstruction



Analysis

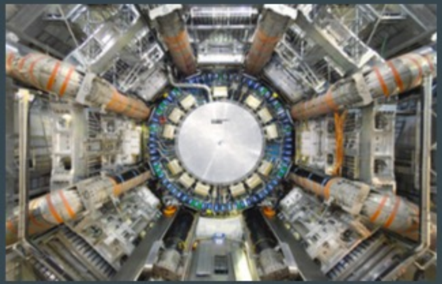


Paper

- 40 MHz collision rate. Potential interesting interactions in every crossing
- ~0.3 PB/second of data for HL-LHC
- Channel- or hit-level signals saved event by event in highly compressed formats

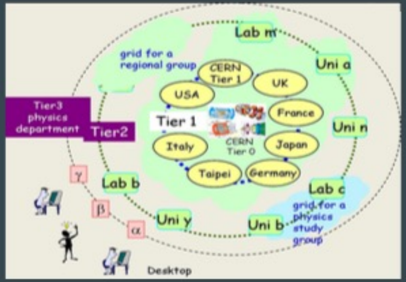
Data flow from collisions to analysis results

Detector

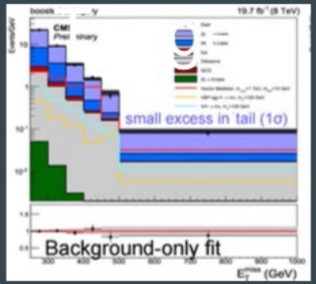


Trigger

Reconstruction



Analysis

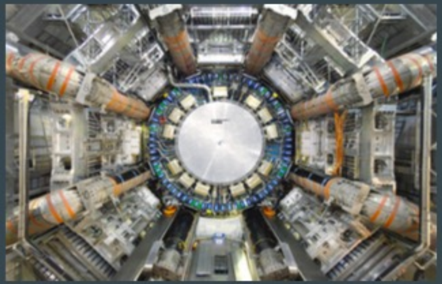


Paper

- 40 MHz reduced first to O(1) MHz input rate (~TB/second) via custom hardware, then to 5-10 kHz via commodity hardware
- In CMS, the second stage trigger shares software codebase with reconstruction application

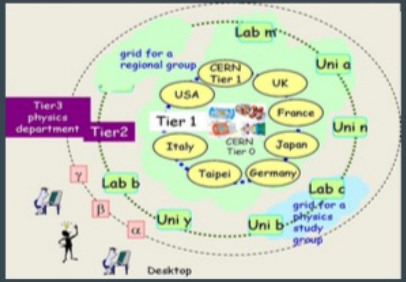
Data flow from collisions to analysis results

Detector

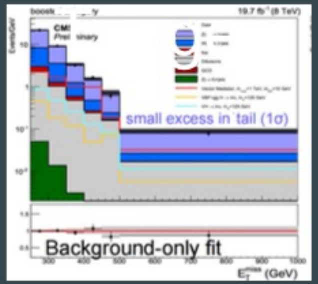


Trigger

Reconstruction



Analysis

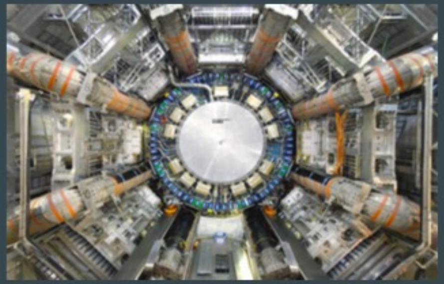


Paper

- 5-10 kHz input rate; each event requires minutes of processing time
- Matching Monte Carlo simulation created using distributed compute system
- Several output formats retained – from 1 kB to 10 MB

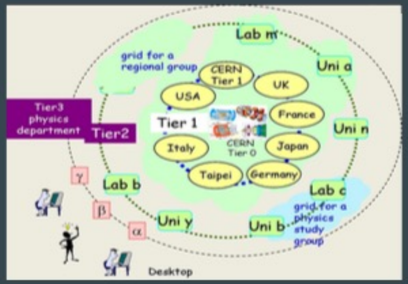
Data flow from collisions to analysis results

Detector

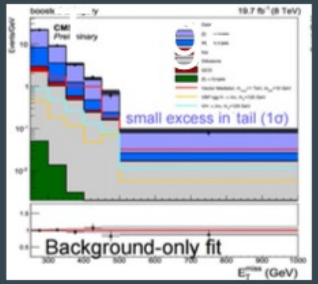


Trigger

Reconstruction



Analysis

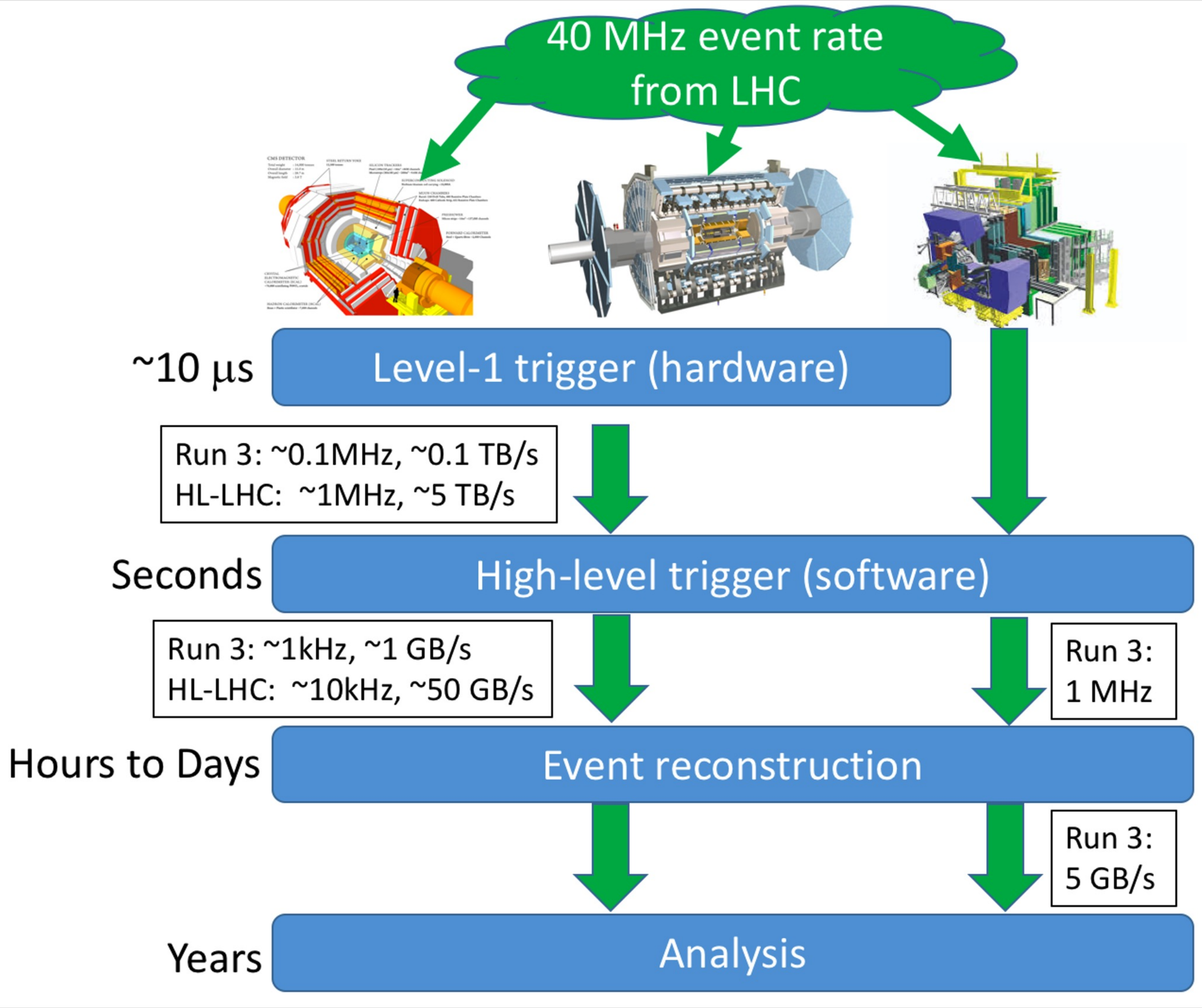


Paper

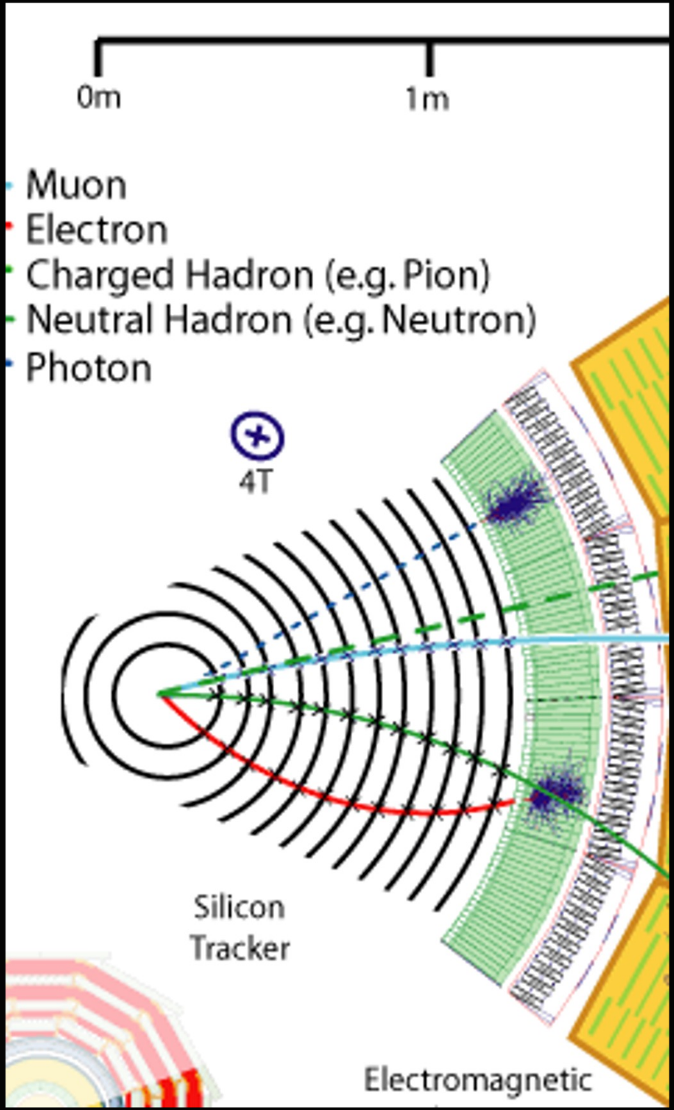
- Analysis groups perform process of reconstruction output down to ntuples using distributed computing system.
- Local compute used to reduce ntuples to plots and eventually complete publication

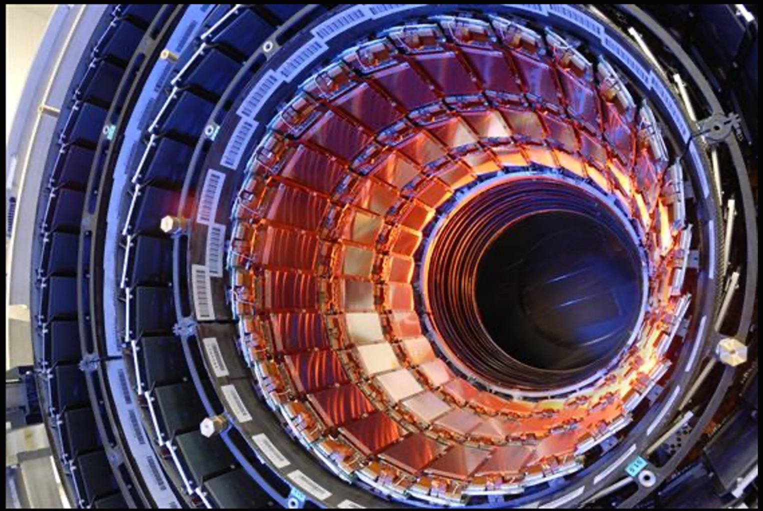
Event trigger

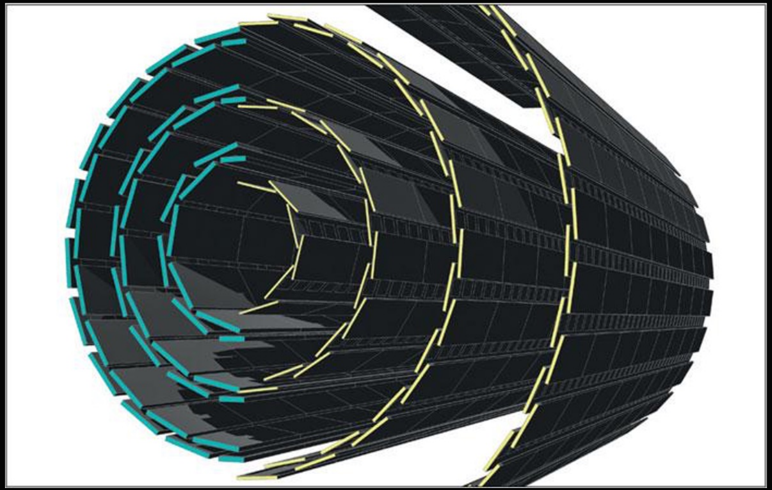
- The LHC collides proton bunches at 40MHz, i.e. a bunch crossing every 25ns
- Common codebases are often used for data reduction (high level trigger) and detailed offline processing (reconstruction) of detector data and simulation
- Each consists of numerous algorithms, primarily developed by HEP research community researchers with varying technical skill sets

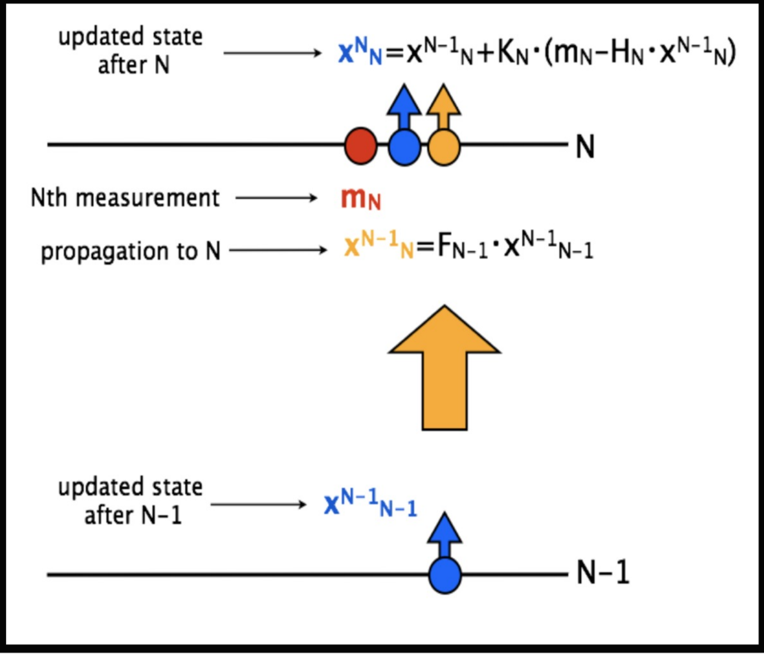


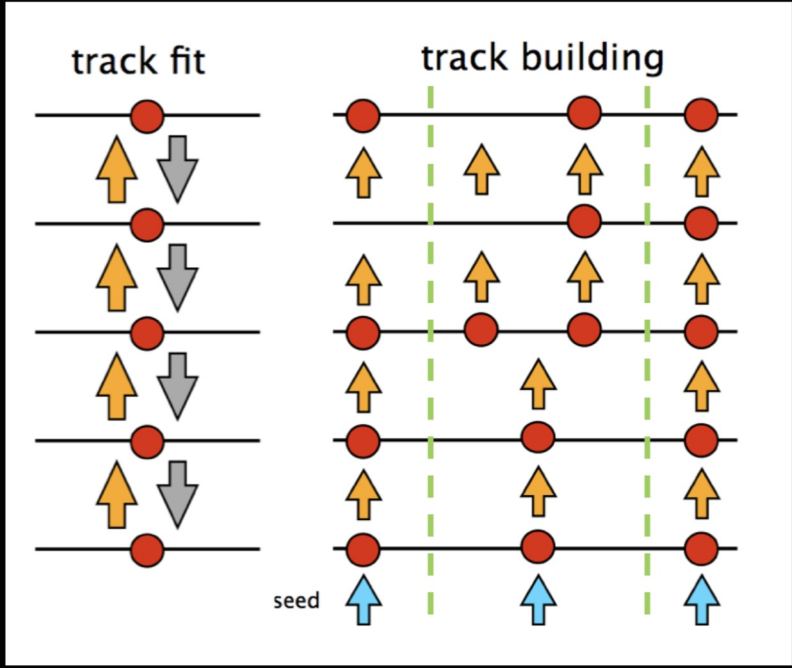
Reconstruction applications process RAW data into "physics objects" for analysis



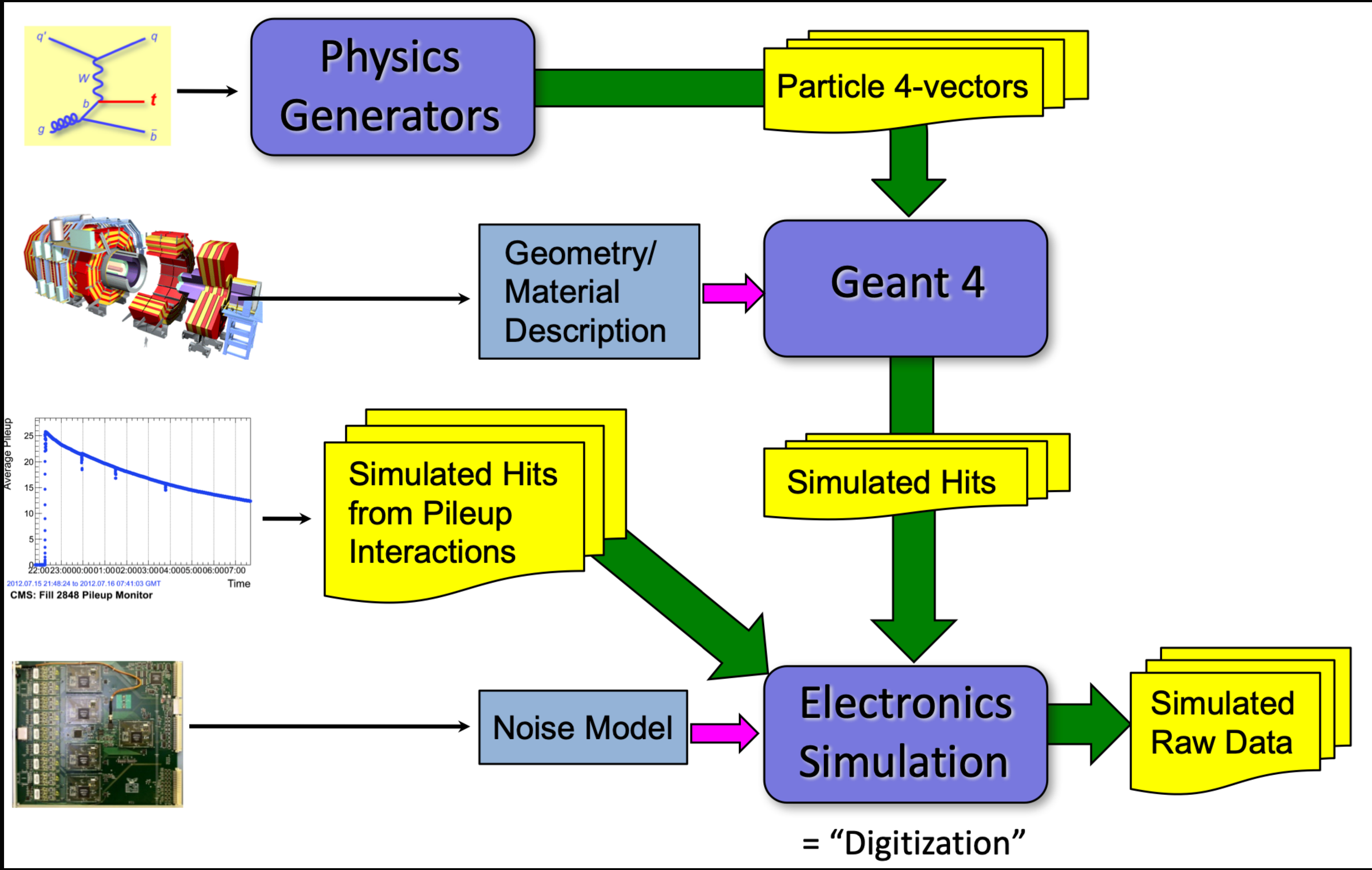






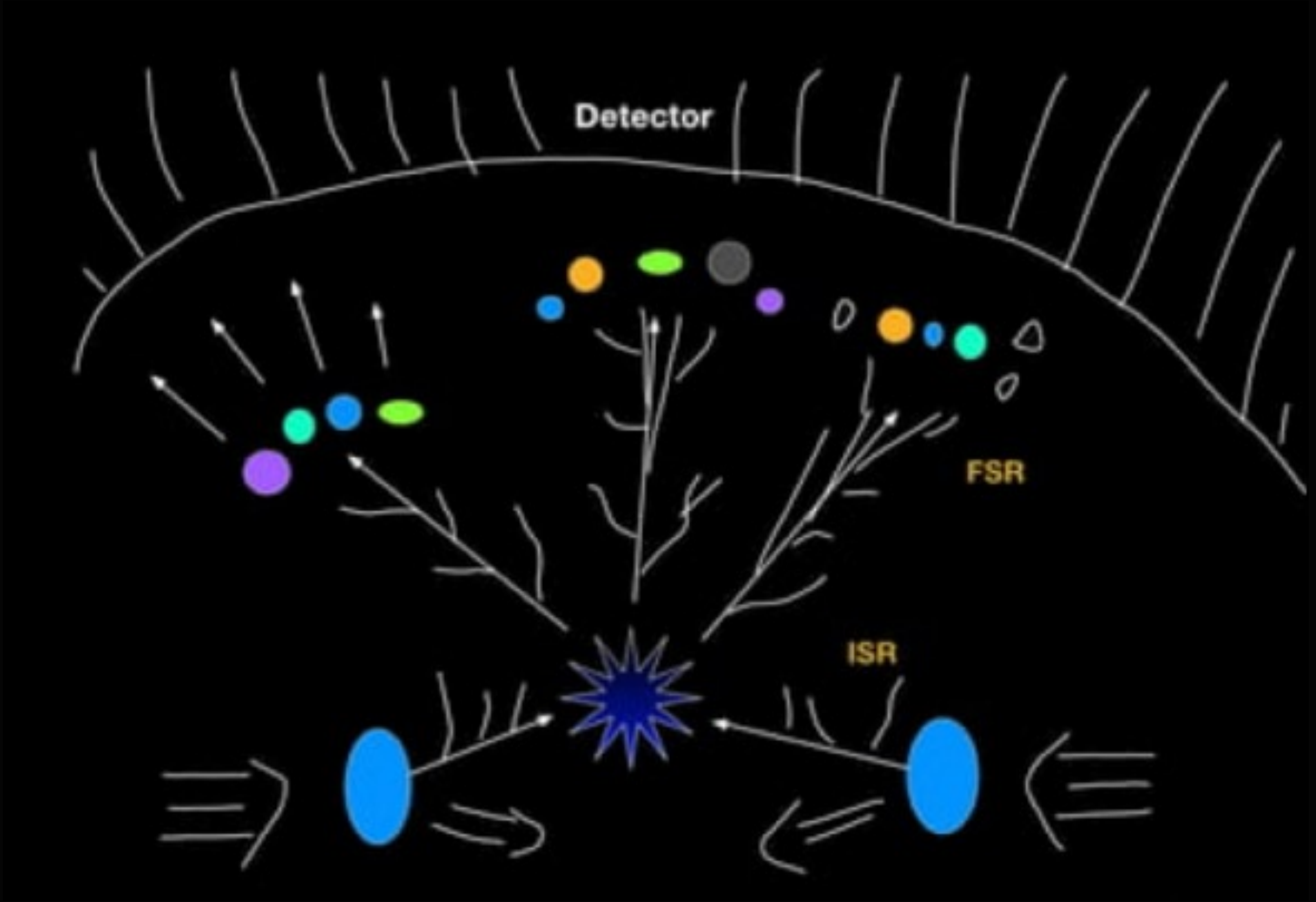


Monte Carlo simulation is an essential tool in HEP



Event Generator view of a collision

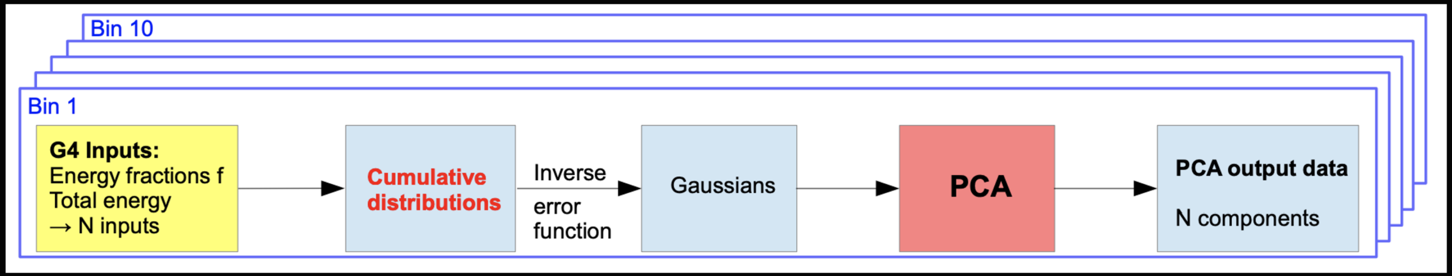
- Event generators simulate high energy collisions of elementary particles (generating momenta and other properties)
- 1 “event” corresponds to 1 collision-scattering
- Matrix element generator
  - Simulates the hard scattering process
- Parton shower generator
  - Soft/collinear radiations

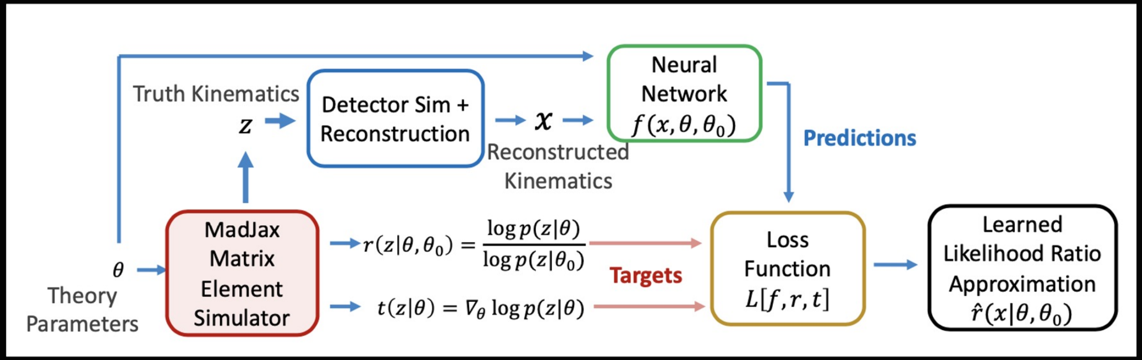


Approaches to detector simulation are increasingly capable

High precision detector modeling required for increasingly complex questions being asked by current and planned experiments.

Fast simulation or fast generation approaches based on modern machine learning techniques

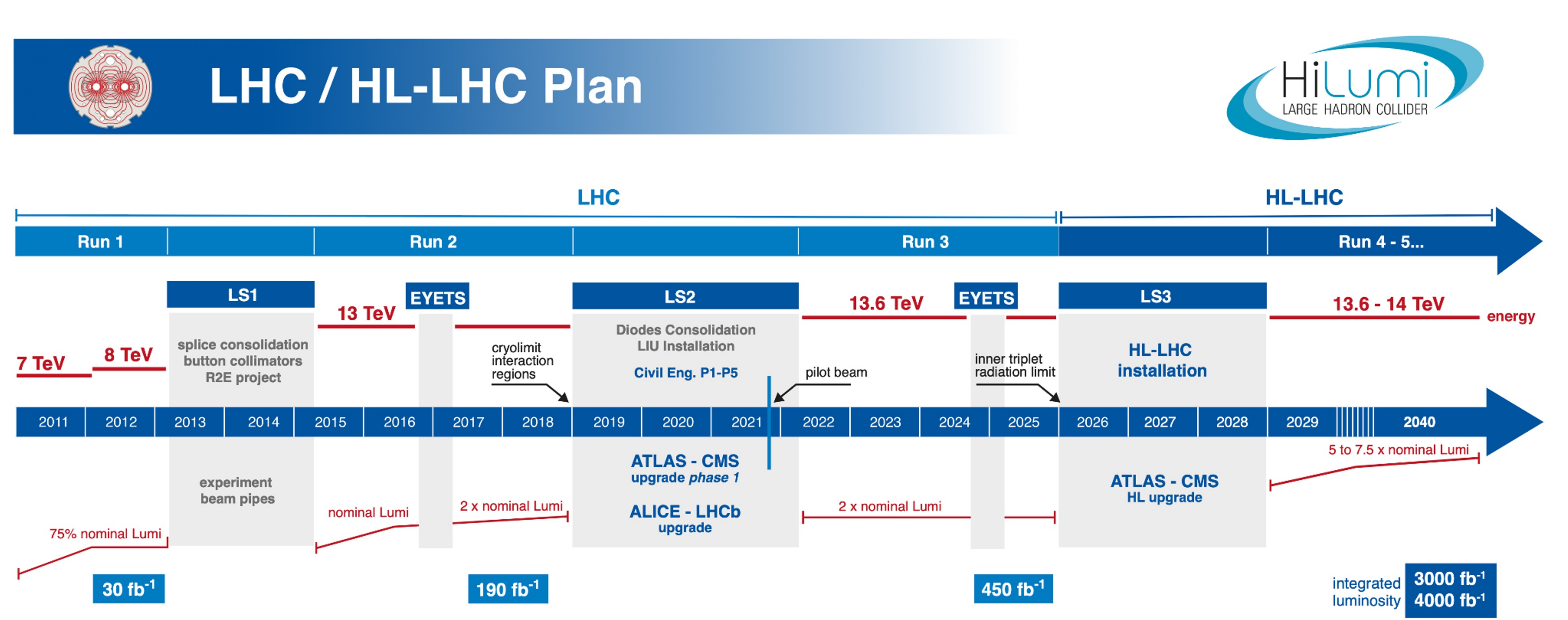




Novel simulation approaches (i.e., simulation based Inference, differentiable programming-based, etc)

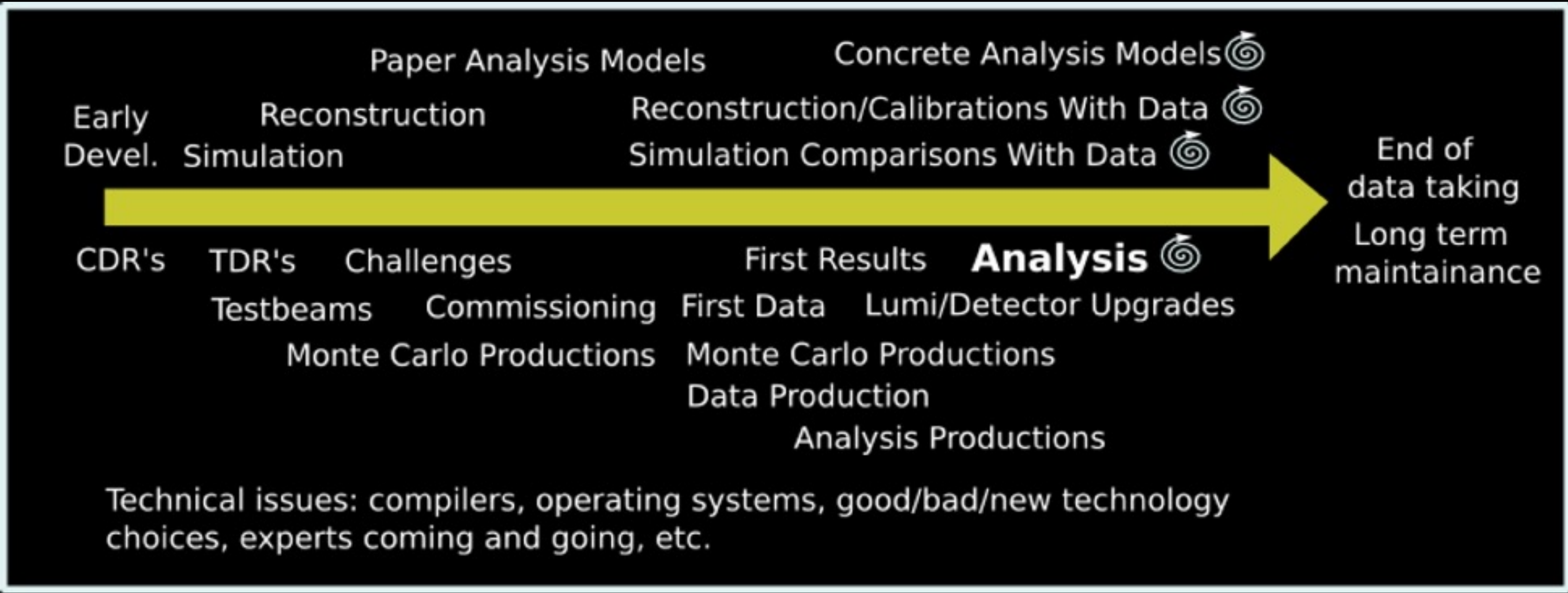
What are the big challenges?

Experimental timescales span decades

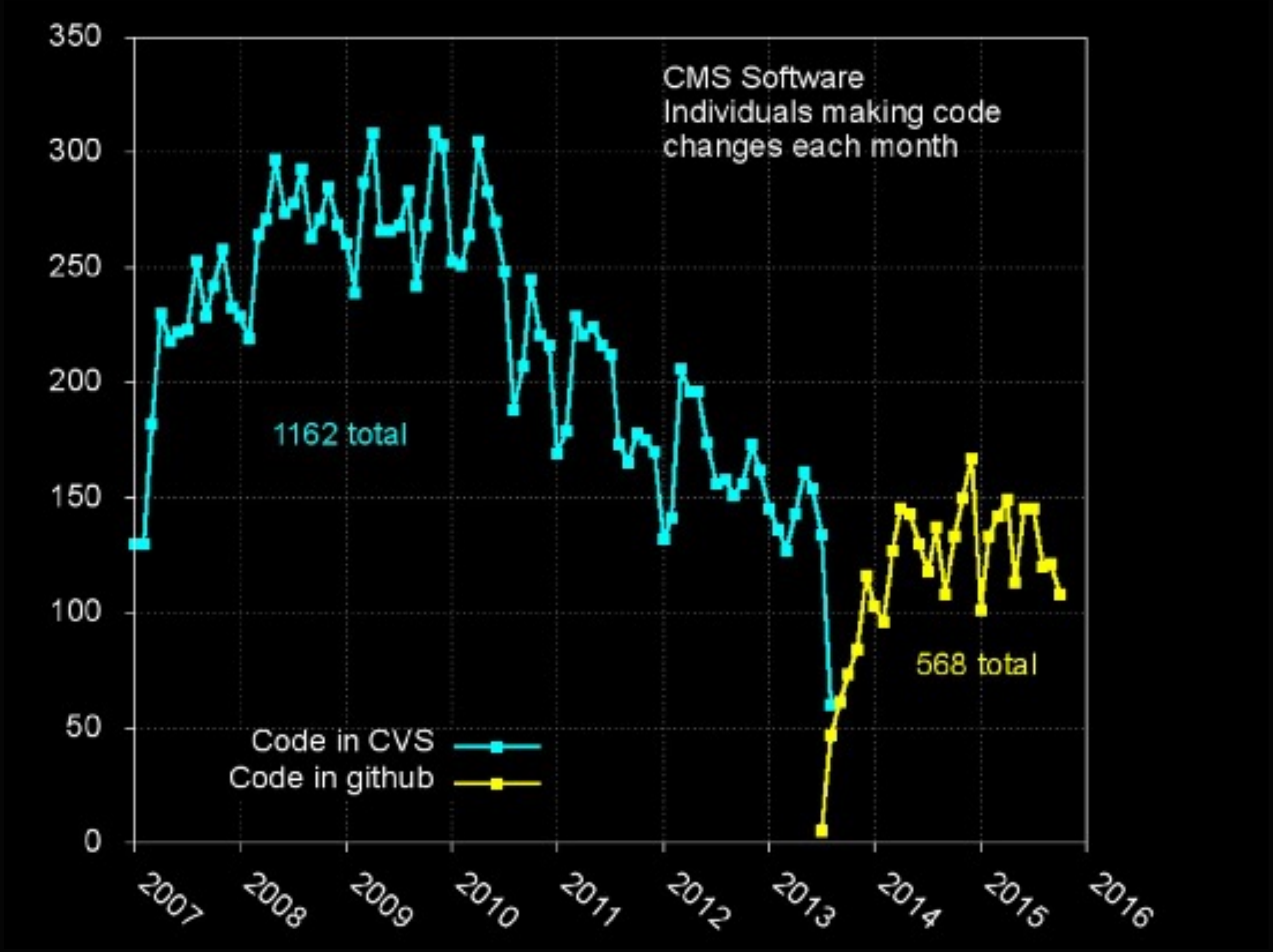


Experiment designs start far before data taking. CMS was formed in 1992 (30 years ago!), expects to run through 2040 and do data analysis for years after that

HEP software lifecycle

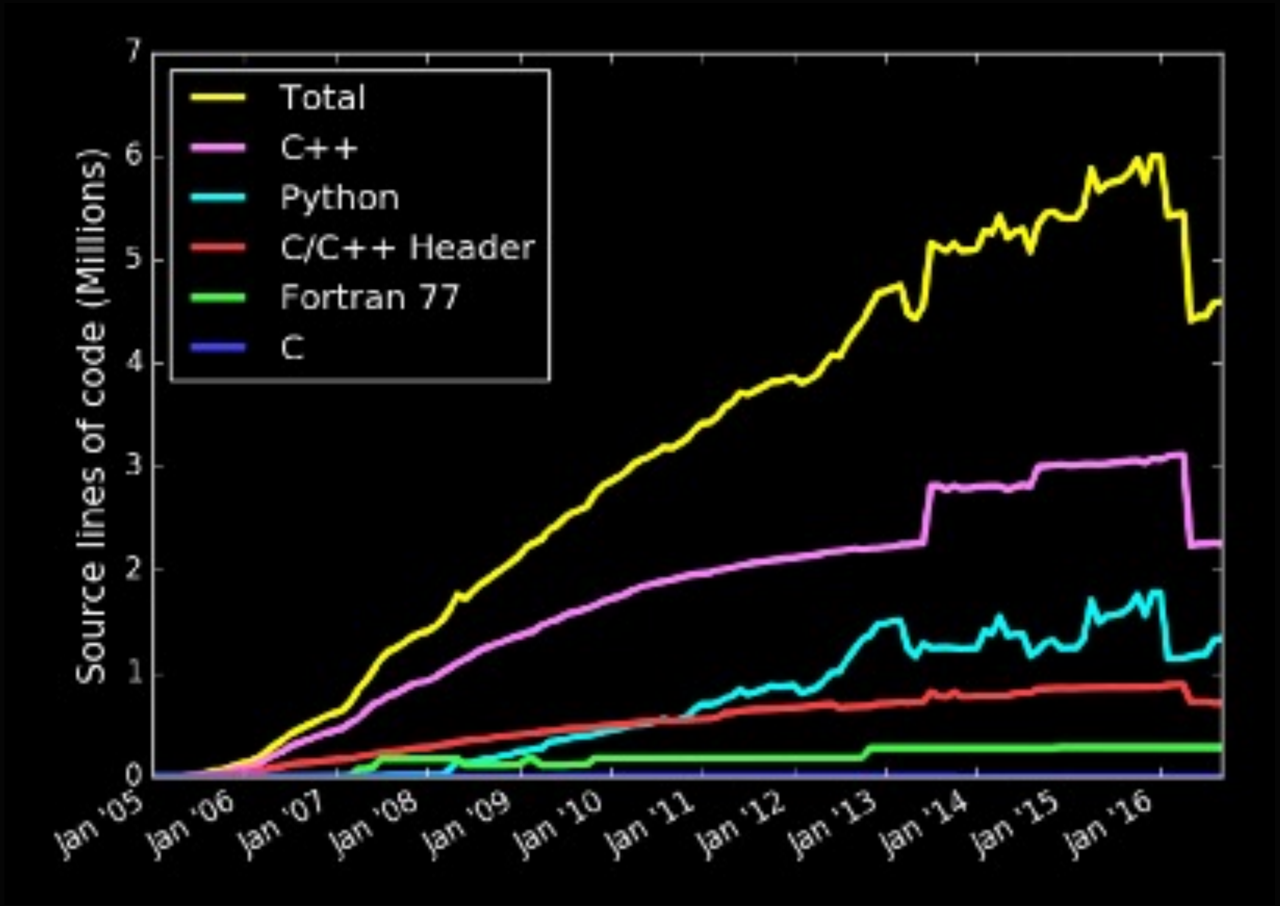


The scale of developer community within experiments



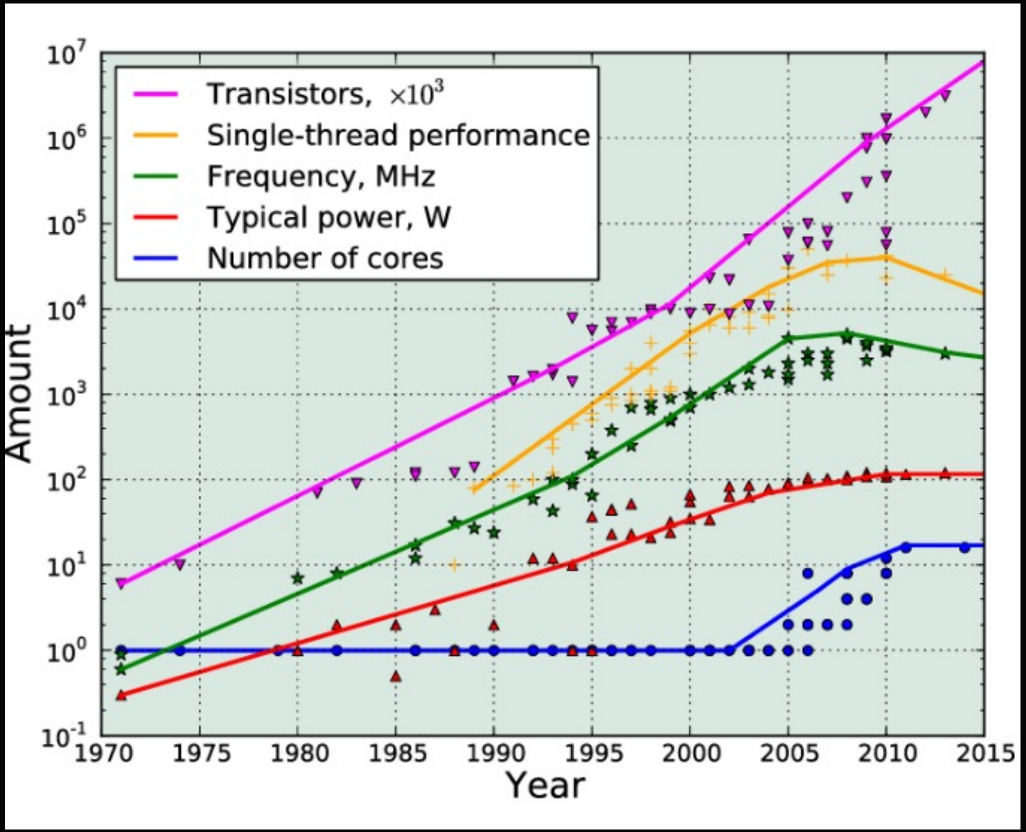
Many developers, typically a handful of true experts

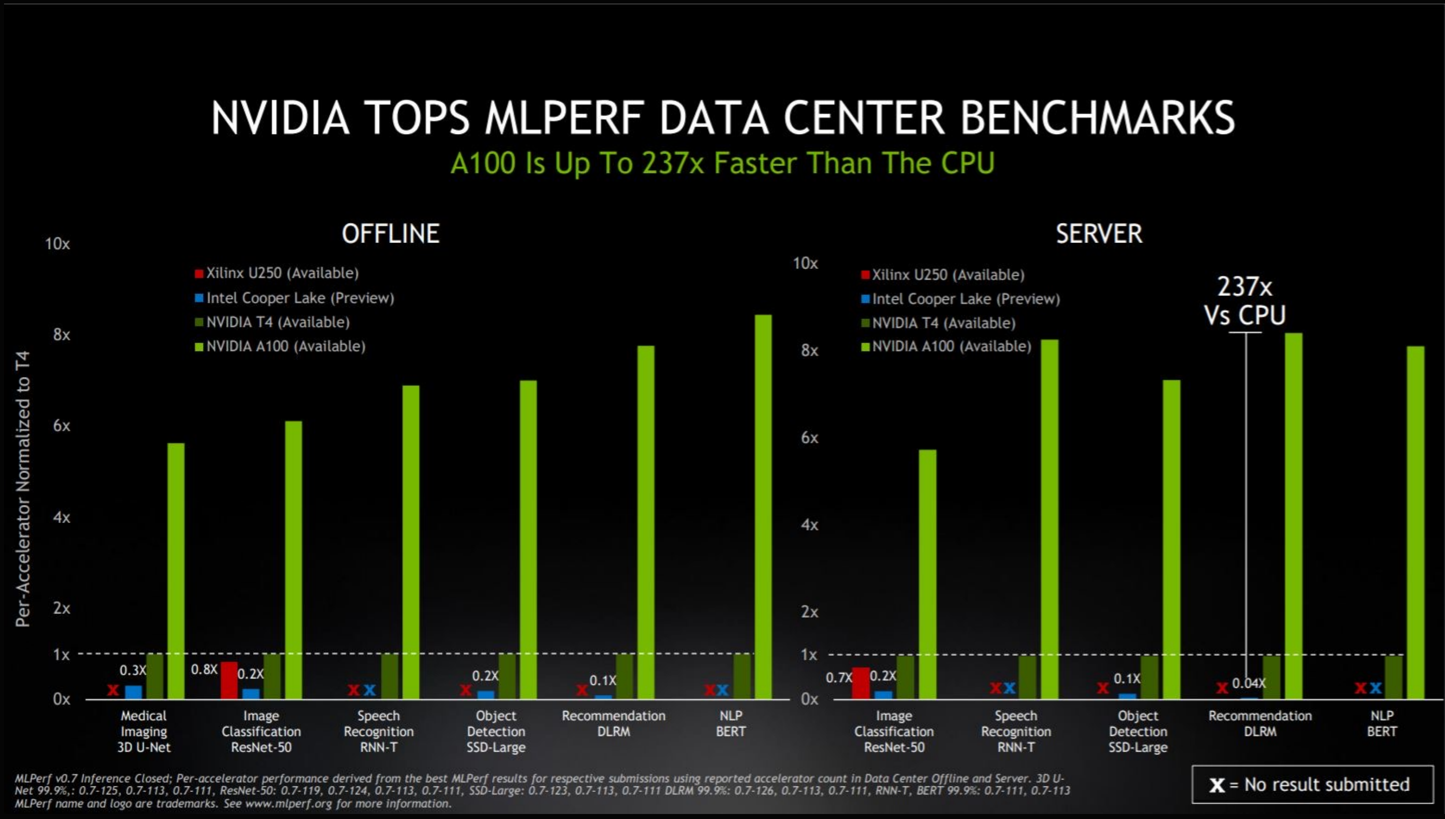
The Scale of the needed code base



Millions of lines of code for CMS - And this excludes most analysis code, event generators, Geant4, etc….

Commodity resources evolve faster during experiment lifecycles

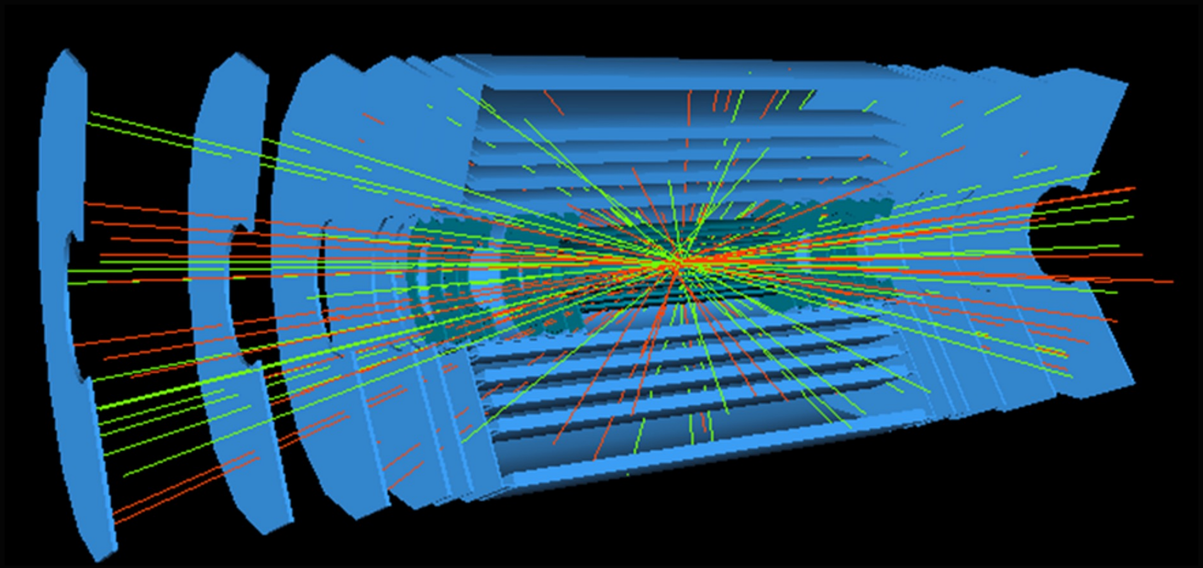


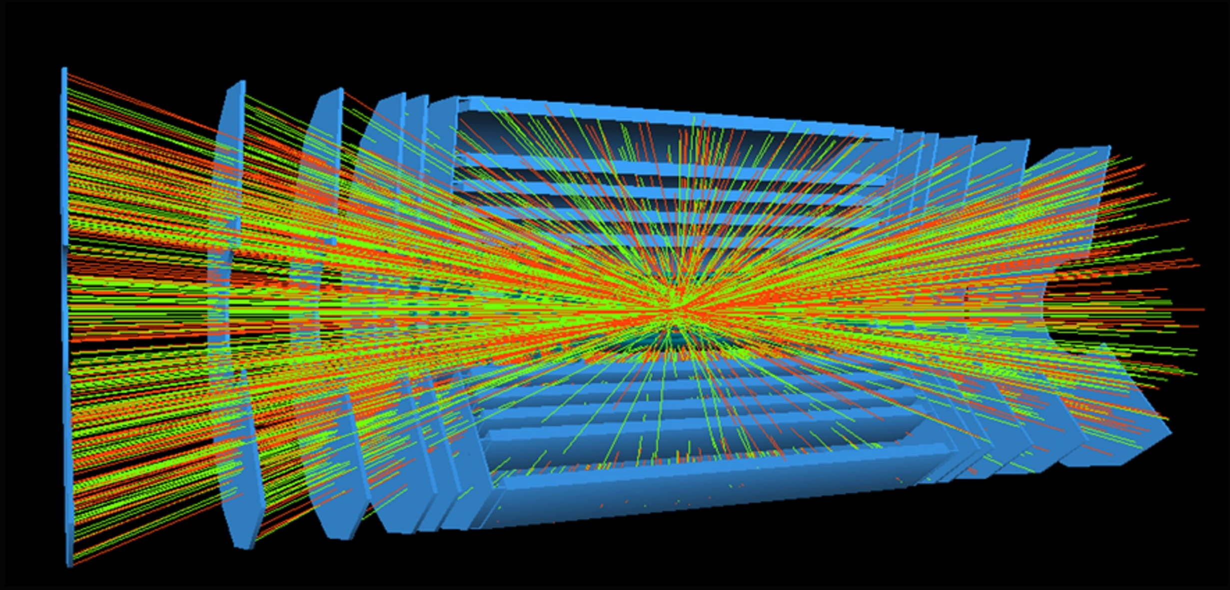


The end of Dennard Scaling: Parallelism has become the ways to faster performance in compute

Parallelism on a chip: GPU performance currently triples every 18 months

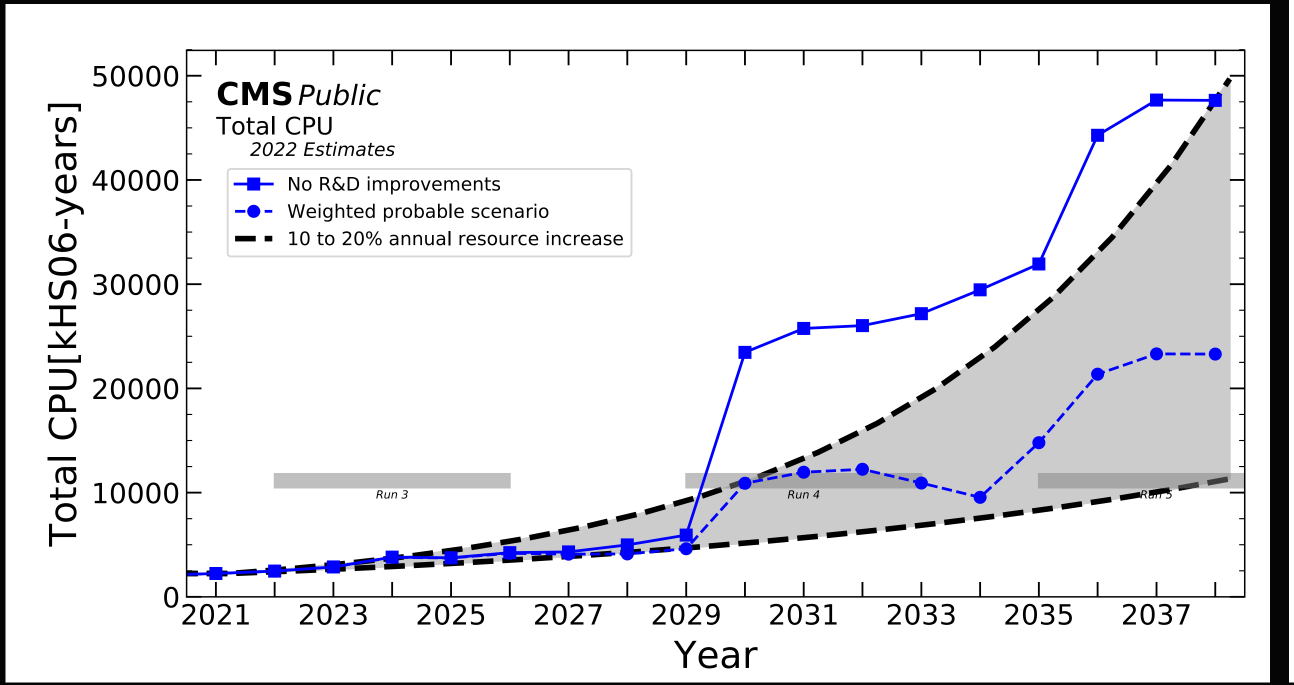
Challenge of next-generation, higher-luminosity or higher-intensity experiments

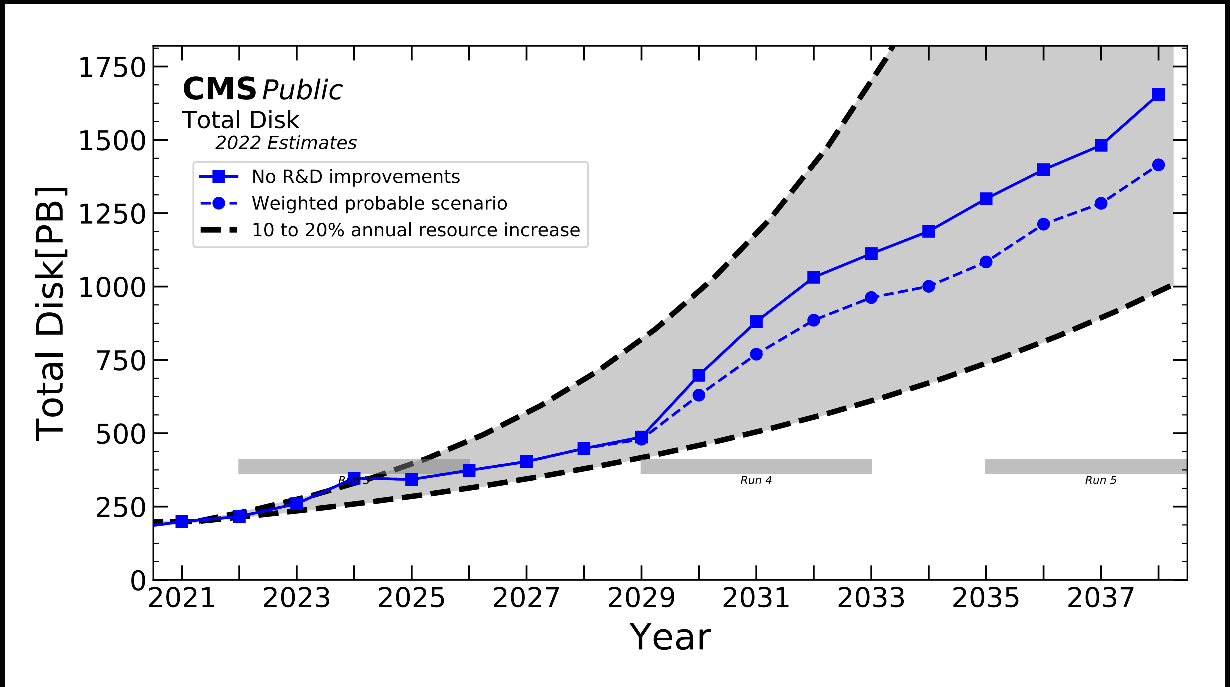




- HL-LHC expects to deliver 200 simultaneous interactions per bunch crossing. 5x more than today.
- More capable detectors are being built to facilitate finding ”needles” in this much bigger “haystack”
- Similarly, the analyst community must develop new approaches to prepare for much higher event rates, higher event complexity, and more detailed detector information.

Future experiments pose even larger computing challenges

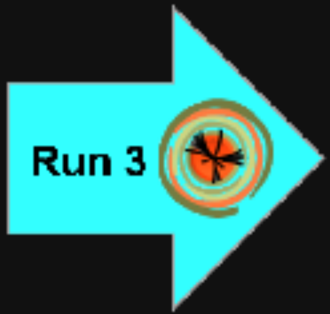




- A naive extrapolation from today's computing model and techniques, even after assuming Moore's Law increases in capabilities, is insufficient to meet the expected resource needs for HL-LHC
  - Technology evolution for processors and storage is an additional challenge
  - New ideas and methods are needed, and software is the key ingredient

Human time is critical: Optimizing analysis is about more than just about pure resources

	LHC (Run 1&2)	HL-LHC (Run 4+)
Analysis Dataset size	10 TB	1,000 TB
Target Scan Turnaround time	Weeks	Hours
Analysis team size (physicists)	5-10	< 5
Primary analysis resource	Laptop	Analysis Facility



LHC analysis:
- Search & Precision Physics
- Simple ML techniques (BDT)
- Reproducibility in its infancy

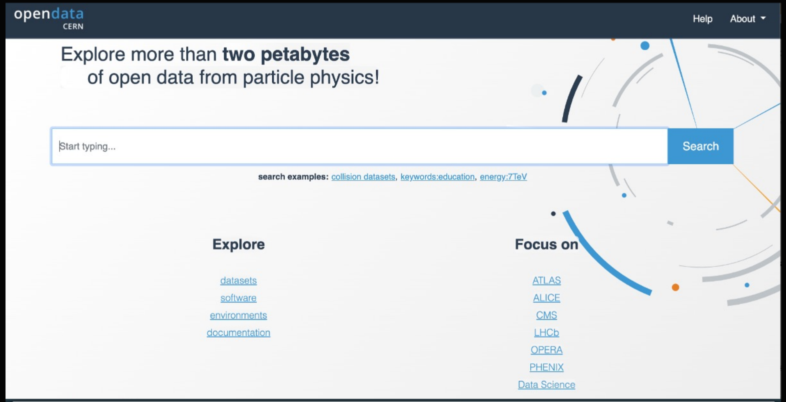
HL-LHC analysis:
- Very High Precision Physics
- Modern ML (Deep Learning)
- Reproducible and Open Data

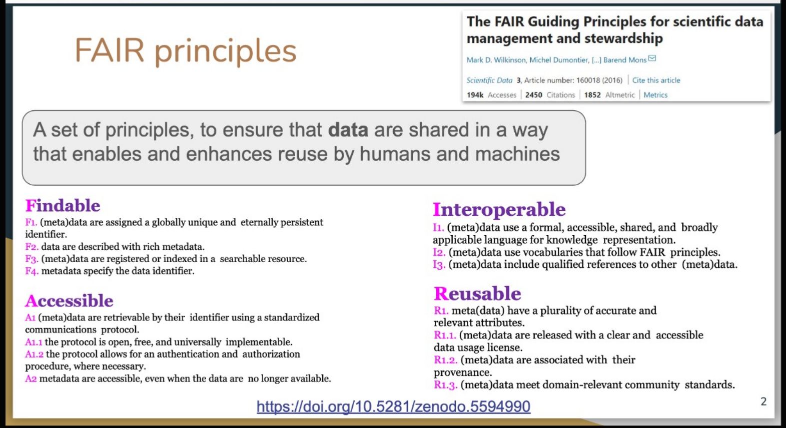
Open data and open software are ever more important

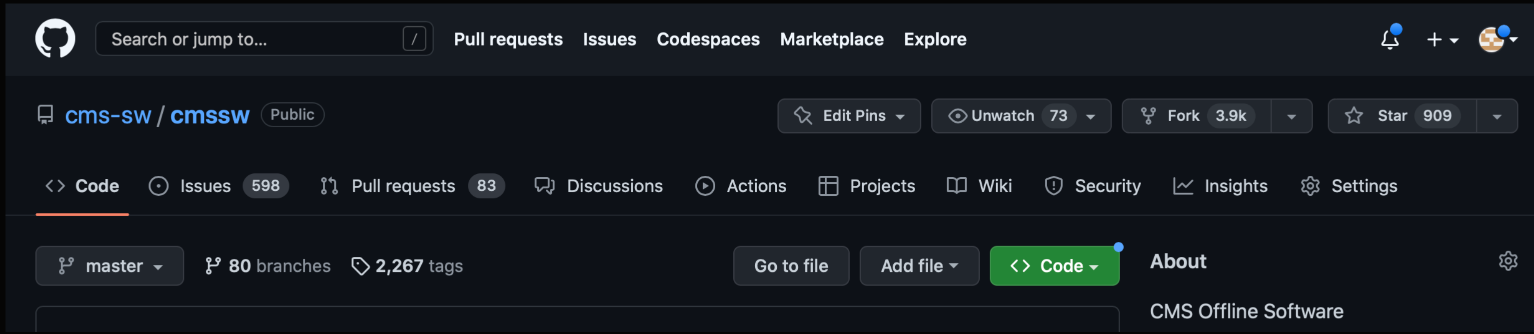
- Reproducibility
- Reusability
- Reinterpretability

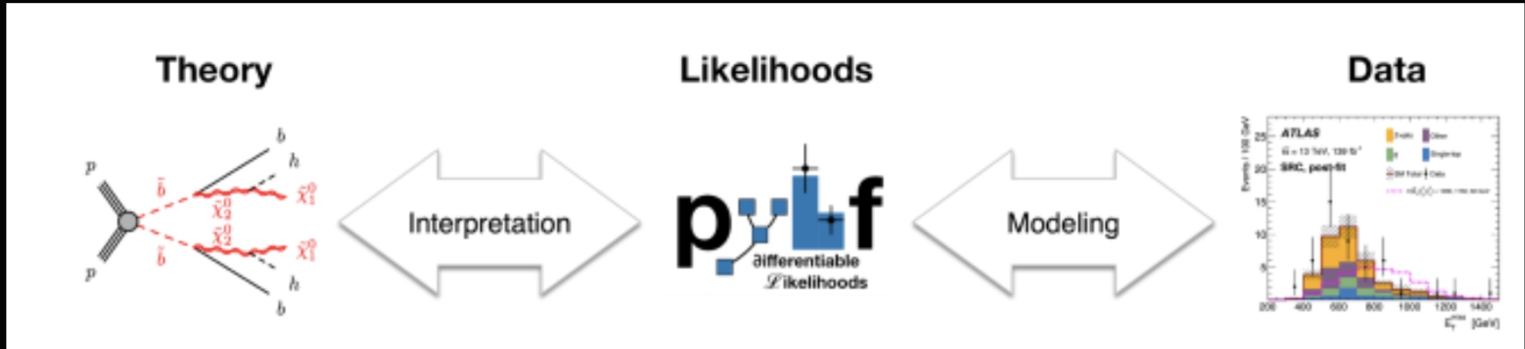
Intentional approach to production, analysis and publication processes fosters research across numerous scientific communities based on HEP our results

- Our community is understanding how best to do this, and what tools are needed to make this "easy" for HEP researchers



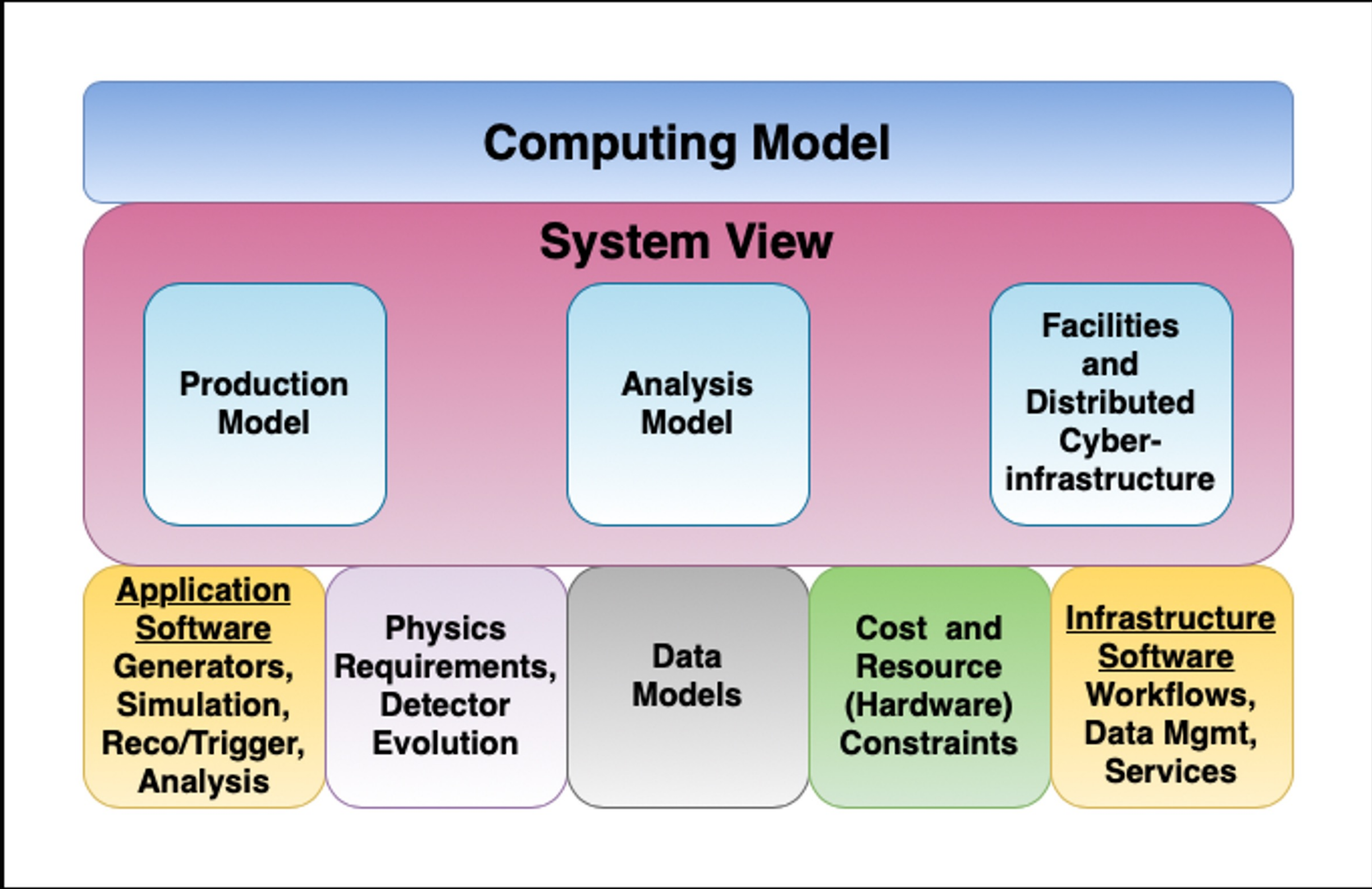






Likelihoods are an essential link between theory and ATLAS data (Image: K. Cranmer/ATLAS)

Taking a system-level view is essential for success



We aim to deliver more than software. Big (Team) Science projects need "Computing Models".

Software enables the science. At the same time, software now often transcends facilities. Is it time to consider software in the context of being infrastructure

Organizing the HEP community to address these challenges

The HEP Software Foundation facilitates cooperation and common efforts in High Energy Physics software and computing internationally.

- The HSF (http://hepsoftwarefoundation.org) was created in early 2015 as a means for organizing our community to address the software challenges of future projects such as the HL-HLC. The HSF has the following objectives:
  - Catalyze new common projects
  - Promote commonality and collaboration in new developments to make the most of limited resources
  - Provide a framework for attracting effort and support to S&C projects
  - Provide a structure to set priorities and goals for work in common projects

Community White Paper

January 2017
UCSD

June 2017
Annecy

ORIGINAL ARTICLE

A Roadmap for HEP Software and Computing R&D for the 2020s

[Computing and Software for Big Science](#) **volume 3**, Article 7 (2019)

"The result: a Programme of Work for the field as a whole, a multifaceted approach to addressing growing computing needs on the basis of existing or emerging hardware."

Eckhard Elsen (CERN Director of Research and Computing), editorial published with CWP/Roadmap

Many workshops, involving a diverse group

- International participants
- Computing Management from the Experiments and Labs
- Individuals interested in the problems
- Members of other compute intensive scientific endeavors
- Members of Industry
- http://s2i2-hep.org/
- https://hepsoftwarefoundation.org/

Individual Papers on the arXiv:

Careers & Training, Conditions Data, DOMA, Data Analysis & Interpretation, Data and Software Preservation, Detector Simulation, Event/Data Processing Frameworks, Facilities and Distributed Computing, Machine Learning, Physics Generators, Security, Software Development, Deployment, Validation, Software Trigger and Event Reconstruction, Visualization

Community White Paper & the Strategic Plan → IRIS-HEP

arXiv 1712.06982 arXiv 1712.06592

The CWP has led to successful collaborative projects (examples..)

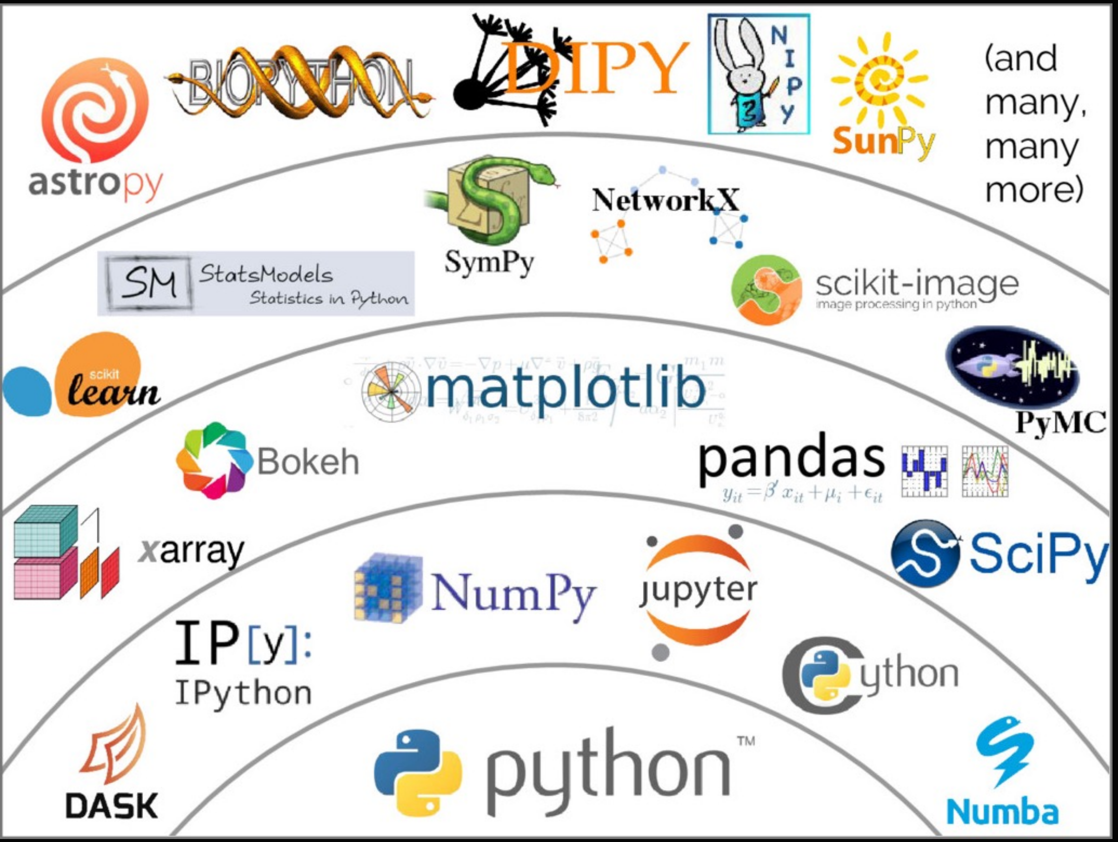
Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)

SoftWare and InFrastruture Technology for High Energy Physics

European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

Leveraging data science for HEP analysis

Scientific Python / PyData vision/ecosystem



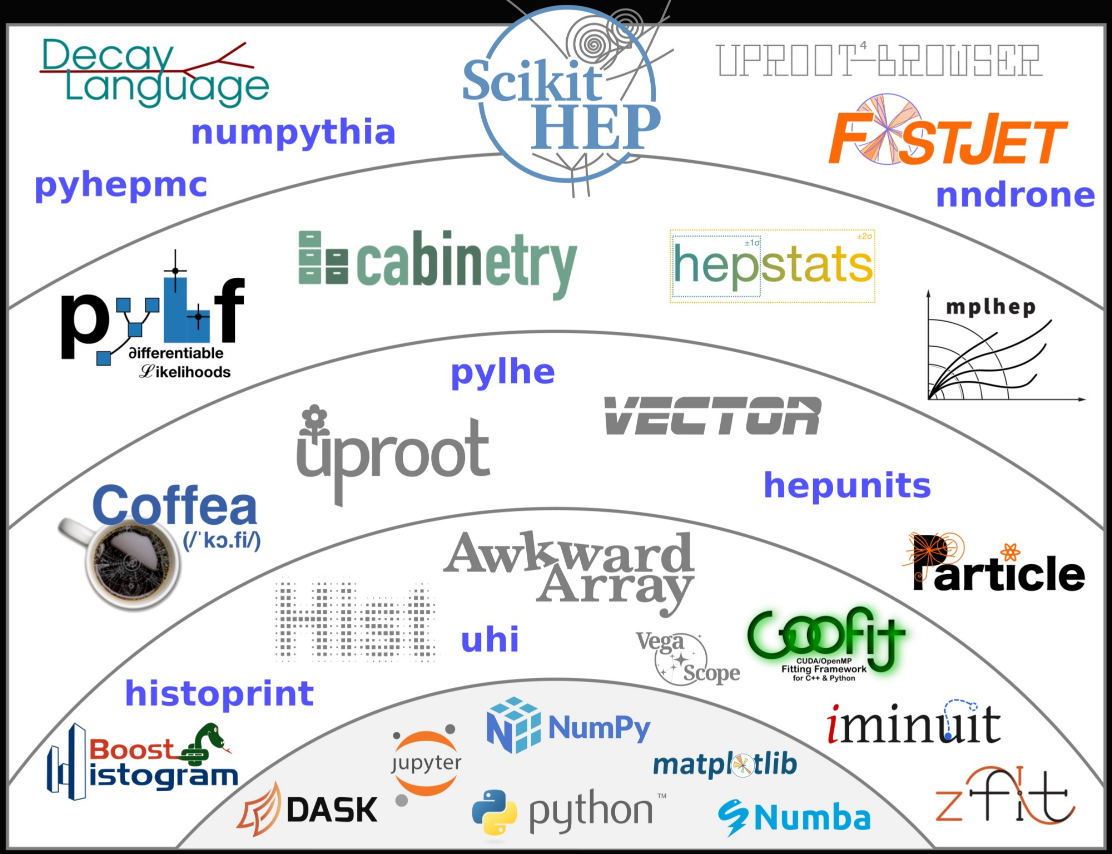
Application Specific

Domain Specific

Technique specific

Foundational

Developing HEP data analysis ecosystem



Awkward Array – numpy for HEP data

```
array = ak.Array([
    [{"x": 1.1, "y": [1]}, {"x": 2.2, "y": [1, 2]}, {"x": 3.3, "y": [1, 2, 3]}],
    [],
    [{"x": 4.4, "y": [1, 2, 3, 4]}, {"x": 5.5, "y": [1, 2, 3, 4, 5]}]
])
```

```
output = []
for sublist in python_objects:
    tmp1 = []
    for record in sublist:
        tmp2 = []
        for number in record["y"][1:]:
            tmp2.append(np.square(number))
        tmp1.append(tmp2)
    output.append(tmp1)
```

2.3 minutes to run (22 GB footprint)

```
output = np.square(array["y", ..., 1:])
```

```
[
    [[], [4], [4, 9]],
    [],
    [[4, 9, 16], [4, 9, 16, 25]]
]
```

4.6 seconds to run (2.1 GB footprint)

(single-threaded on a 2.2 GHz processor with a dataset 10 million times larger than the one shown)

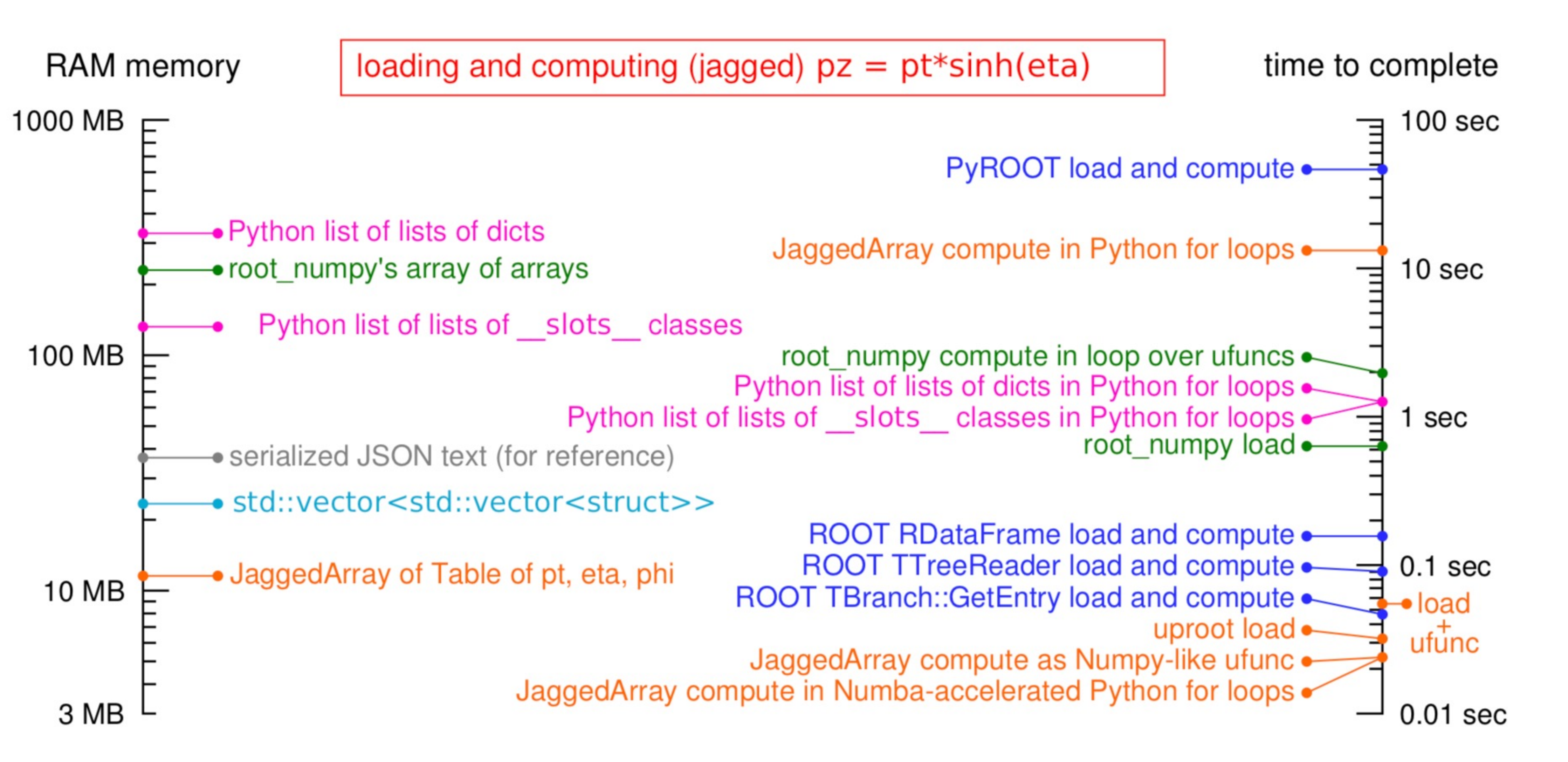
General tool for manipulating JSON-like structures in a NumPy-like way.

Motivated by problems in HEP, but general enough for any irregular data.

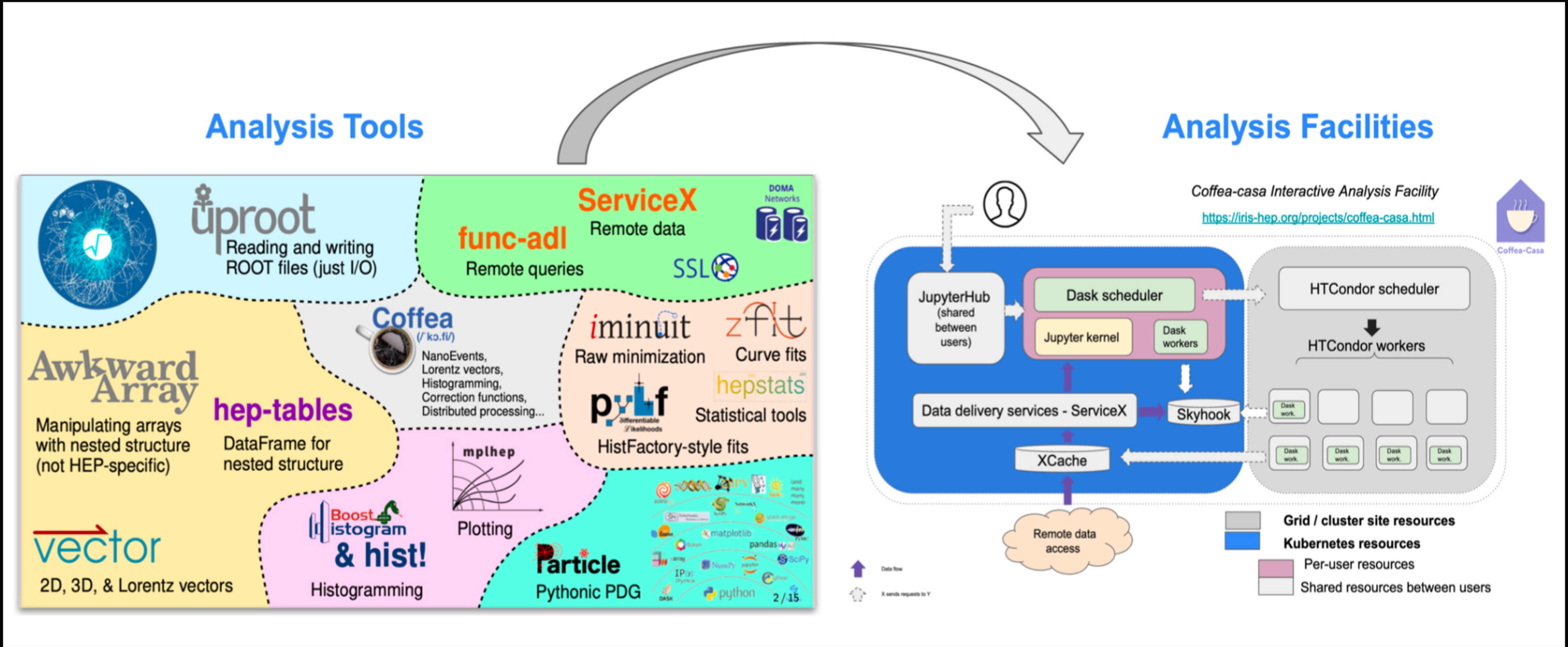
Featured on the PythonBytes podcast.

- General tool for manipulating JSON-like structures in a NumPy-like way
- Motivated by problems in HEP which commonly include irregular, “jagged” data

Exciting results are possible: Orders of magnitude speed ups

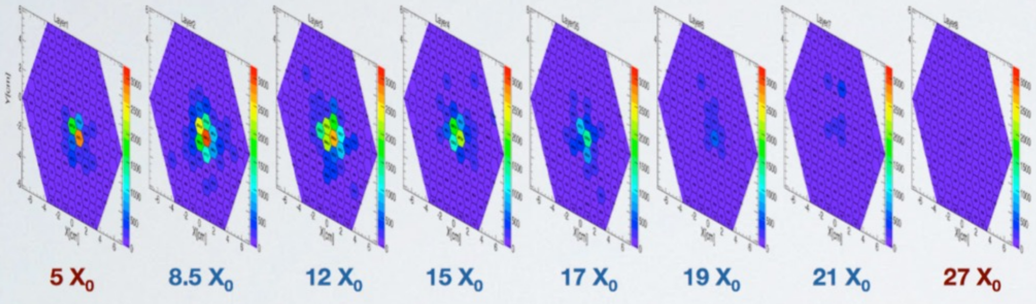


What comes next? Analysis tool chains and facilities rather than just tools

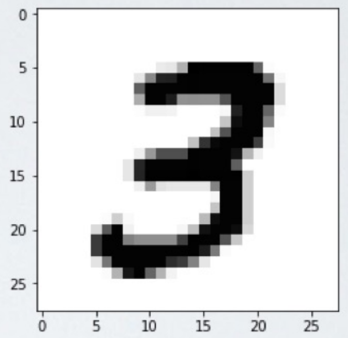


Adoption of AI techniques to solve HEP challenges

HGCAL Data



Sparsified MNIST Data



100 hits with highest energy

$$[\varphi_1, \eta_1, E_1]$$
$$[\varphi_2, \eta_2, E_2]$$
$$\vdots$$
$$[\varphi_{100}, \eta_{100}, E_{100}]$$

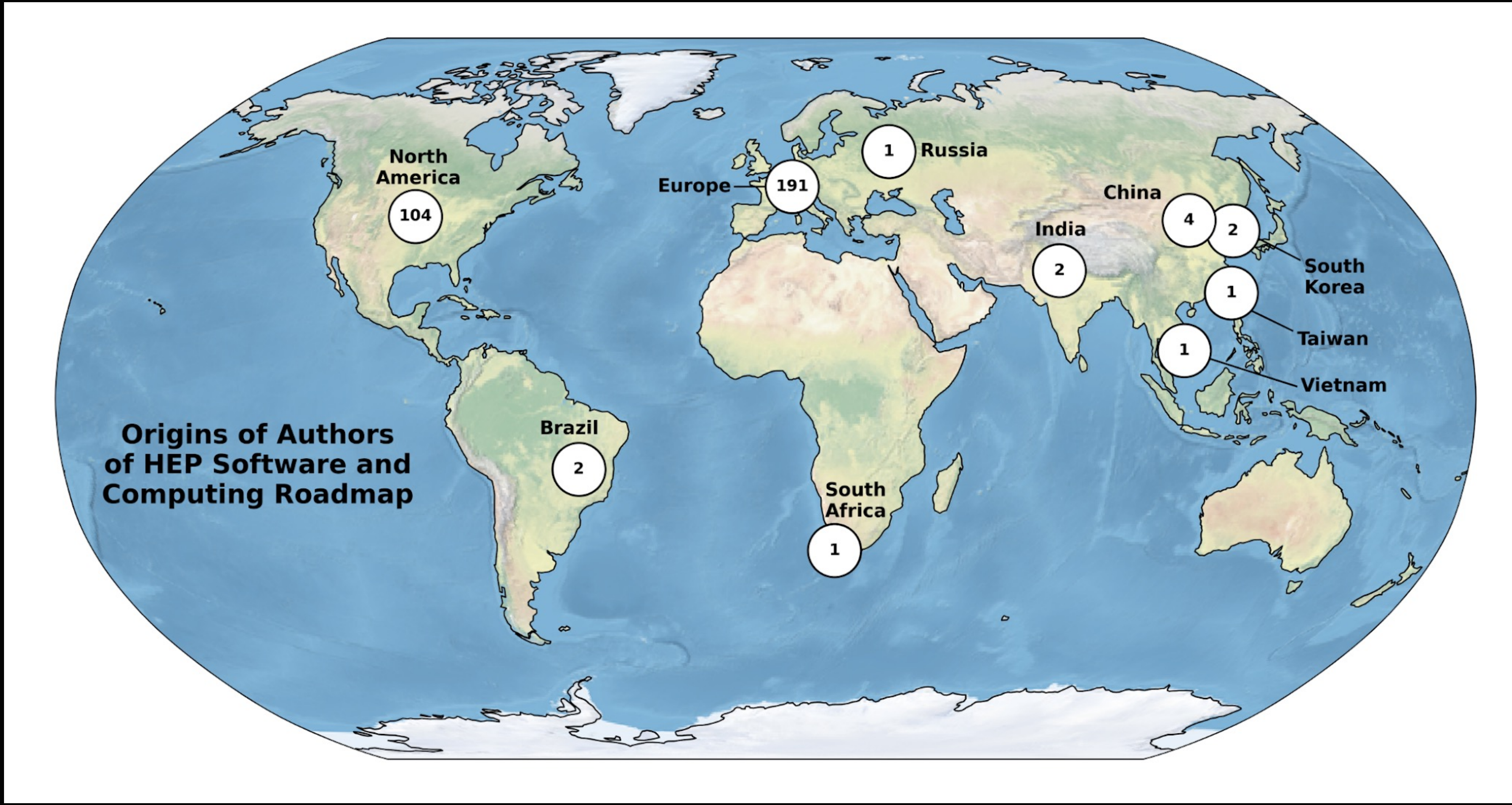
$\longleftrightarrow$

$$[x_1, y_1, I_1]$$
$$[x_2, y_2, I_2]$$
$$\vdots$$
$$[x_{100}, y_{100}, I_{100}]$$

100 pixels with highest intensity

So then why are we here…

Observation: Nearly all authors of the HSF Community Roadmap where from institutions in Europe and the US



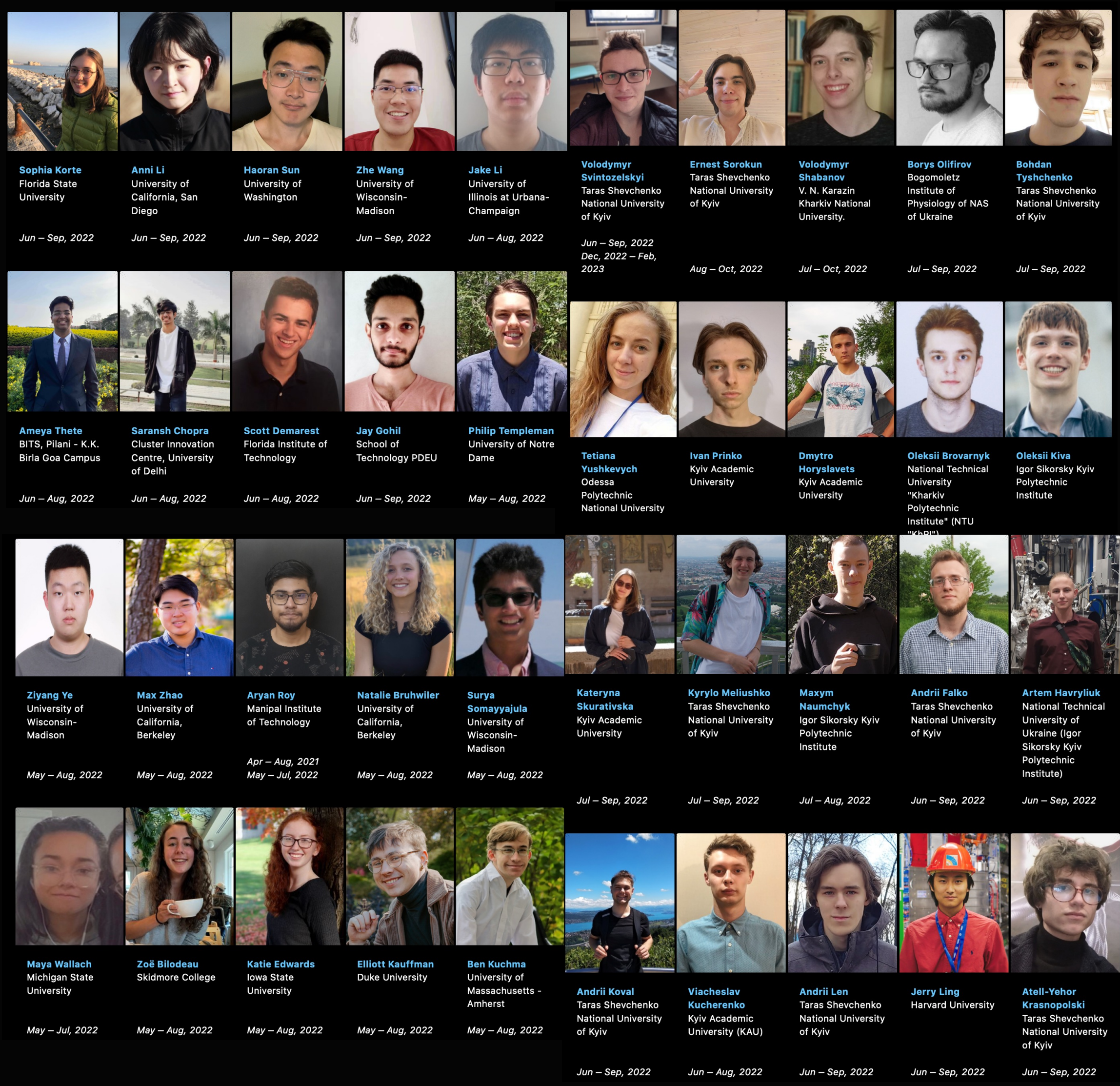
HSF-India project

- HSF-India is a 5 year project funded by the US National Science Foundation that aims to build international research software collaborations between US, European, and India based researchers to reach the science goals of experimental particle, nuclear and astroparticle research.
- Given the growing complexity of our scientific data and collaborations, these collaborations are increasingly important to raise the collective productivity of our research community.
- Intended as a long-term investment in international team science.
- What activities can we fund?
  - Fellowships
  - Researcher exchanges
  - Training events

IRIS-HEP Fellows Program

IRIS-HEPs fellow program.
https://iris-hep.org/fellows.html

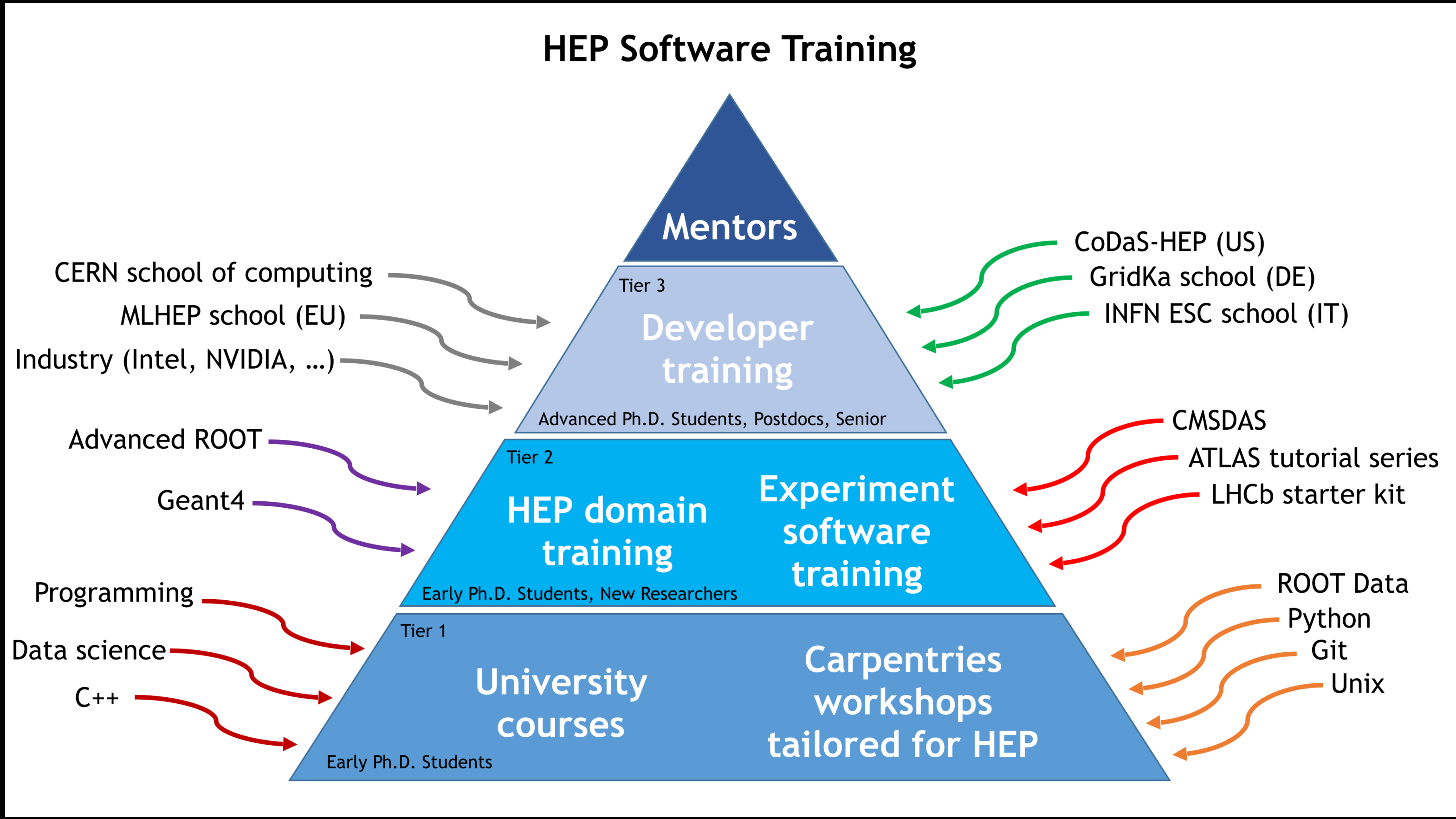
- **Key Insight:** we need to provide incentivized and explicit paths forward for enthusiastic students from the more advanced training schools in HEP (ESC/Bertinoro, CoDaS-HEP, MLHEP, etc.) or for people who become engaged with our software projects in other ways.
- **Project focused:** bring students into contact with “mentors” to work on a specific, pre-defined project, allowing them to grow their software skills and experience working in large projects.



Three broad research themes as a basis for building collaborations

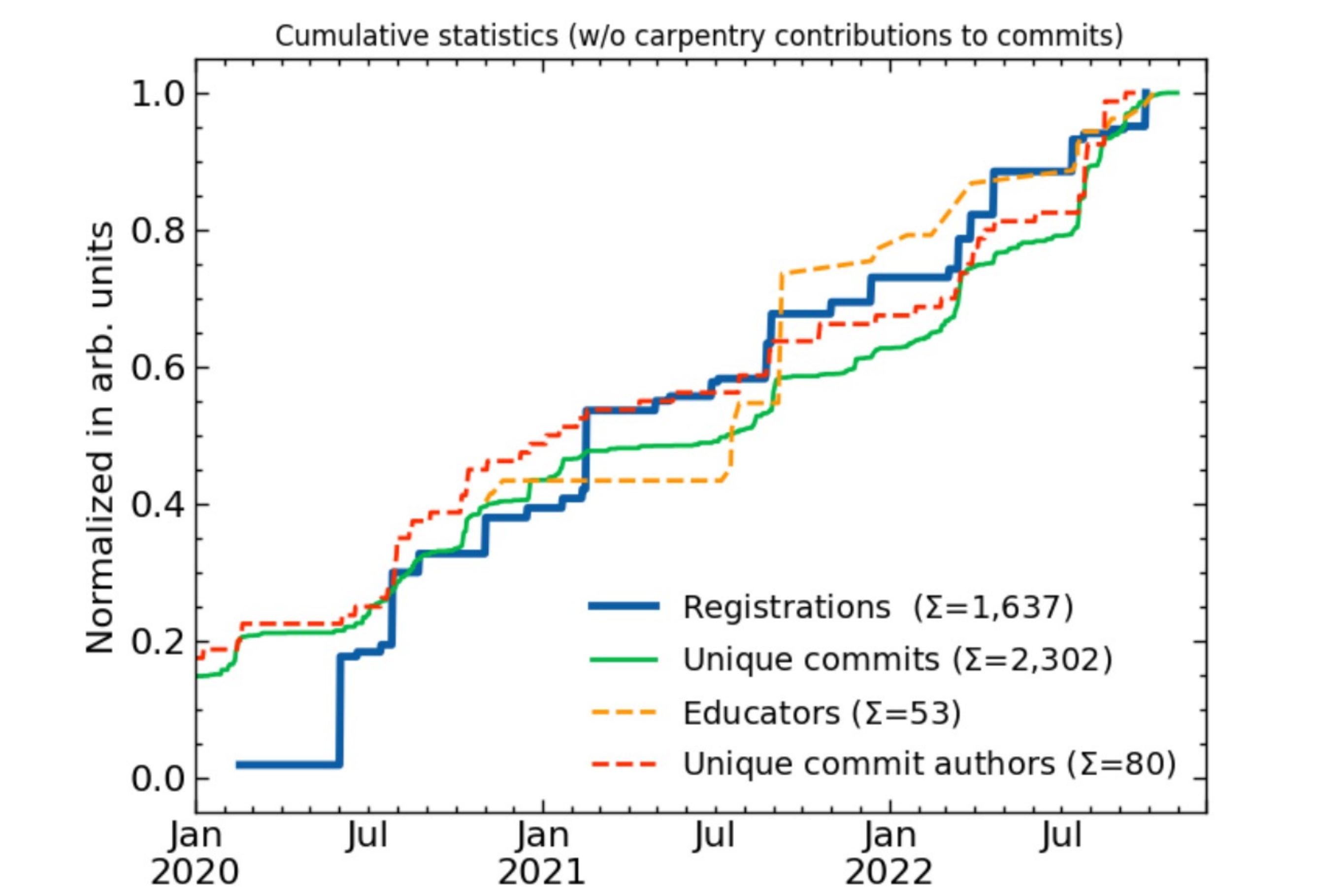
1. Analysis Systems
   - R&D for tools or techniques for analysts
   - Building integrated analysis facilities
2. Simulation tools
   - Event generators, detector simulation, fast simulation techniques, etc
   - R&D on/with existing toolkits (especially cross-experimental)
   - R&D on new techniques
3. Open Science
   - R&D on tools, techniques and systems that enable and encourage reuse, reproducibility, etc of scientific results

Bootstrap collaboration through software training



- A vision for training in HEP: researchers progress (vertically) from basic skills training, through user training in existing software to training in skills needed to develop new research software.

The HSF has accumulated considerable training material and runs regular training courses



The number of commits in our repository and the number of registered learners closely follows the number of our educators. You, too, can make a difference!

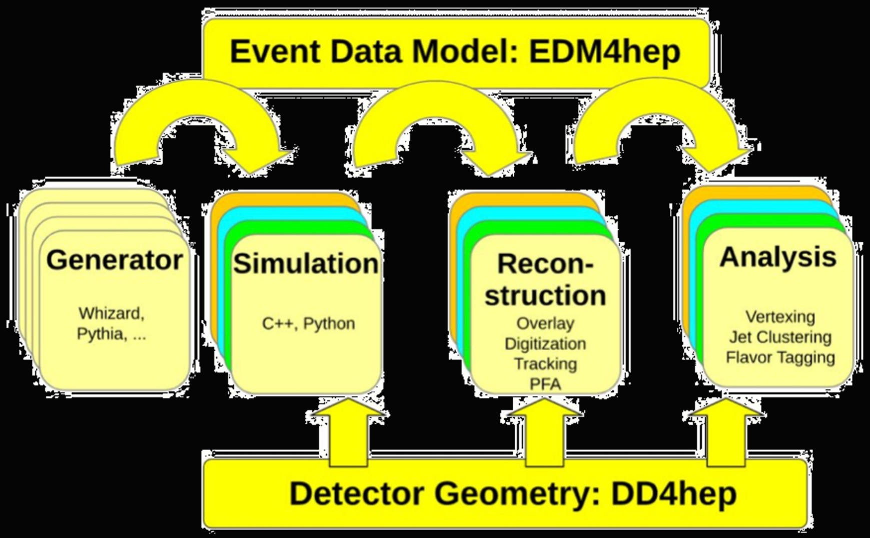
Wrapup

- Many aspects in research software in HEP are both exciting and challenging. Our field is evolving to work cross-experiment and towards and open and more sustainable approach to software infrastructure
- “HSF-India” : We are here to discuss with interested research groups about how our program could be structured and used to support your research goals
  - General attributes of fellowships and research exchanges
  - Specific research ideas/projects with common interest
  - Establish communication channels
  - Identify opportunities for training events
- Thank you for coming!

The analyst community is very interested in this approach (As measured by the IRIS-HEP Slack group)

Examples how experiments are not re-inventing the wheel

Acts is an experiment-independent toolkit for (charged) particle track reconstruction in (high energy) physics experiments implemented in modern C++.



The Key4hep project aims to provide a turnkey software solution for the full experiment life-cycle, based on established HEP community tools. It initially targets future collider communities (CEPC, CLIC, EIC, FCC, and ILC)