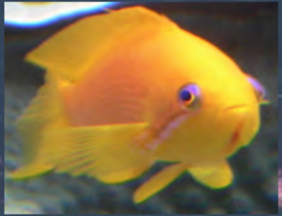
Deep Learning for Higgs Boson Identification and Searches for New Physics at the LHC

Blue Waters Symposium
June 4, 2019

Mark Neubauer
University of Illinois at Urbana-Champaign



The Pursuit of Particle Physics

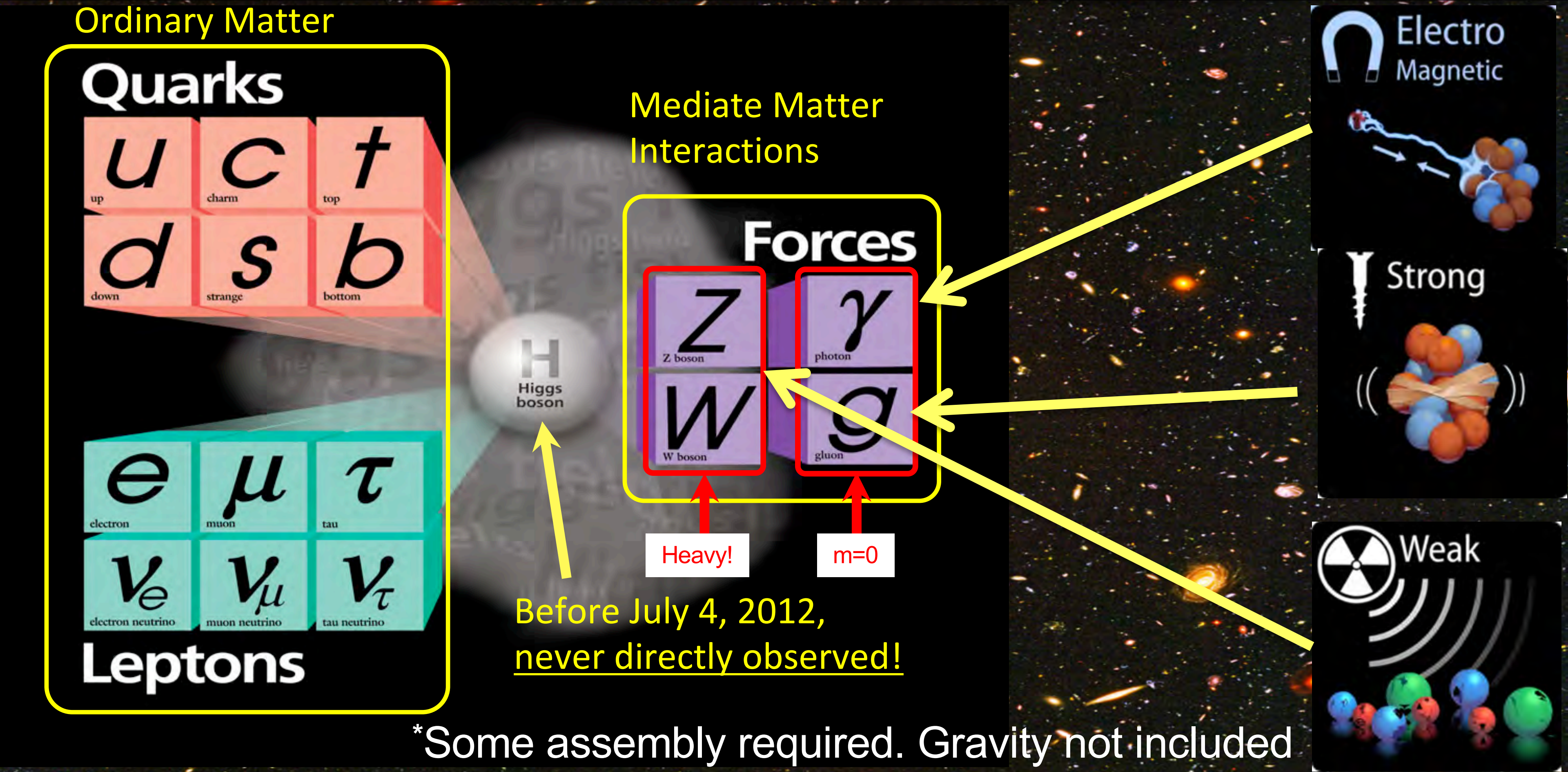
To understand the the Universe at its most fundamental level

Primary questions: What are the

- elementary constituents of matter?
- the nature of space and time?
- forces that dictate their behavior?

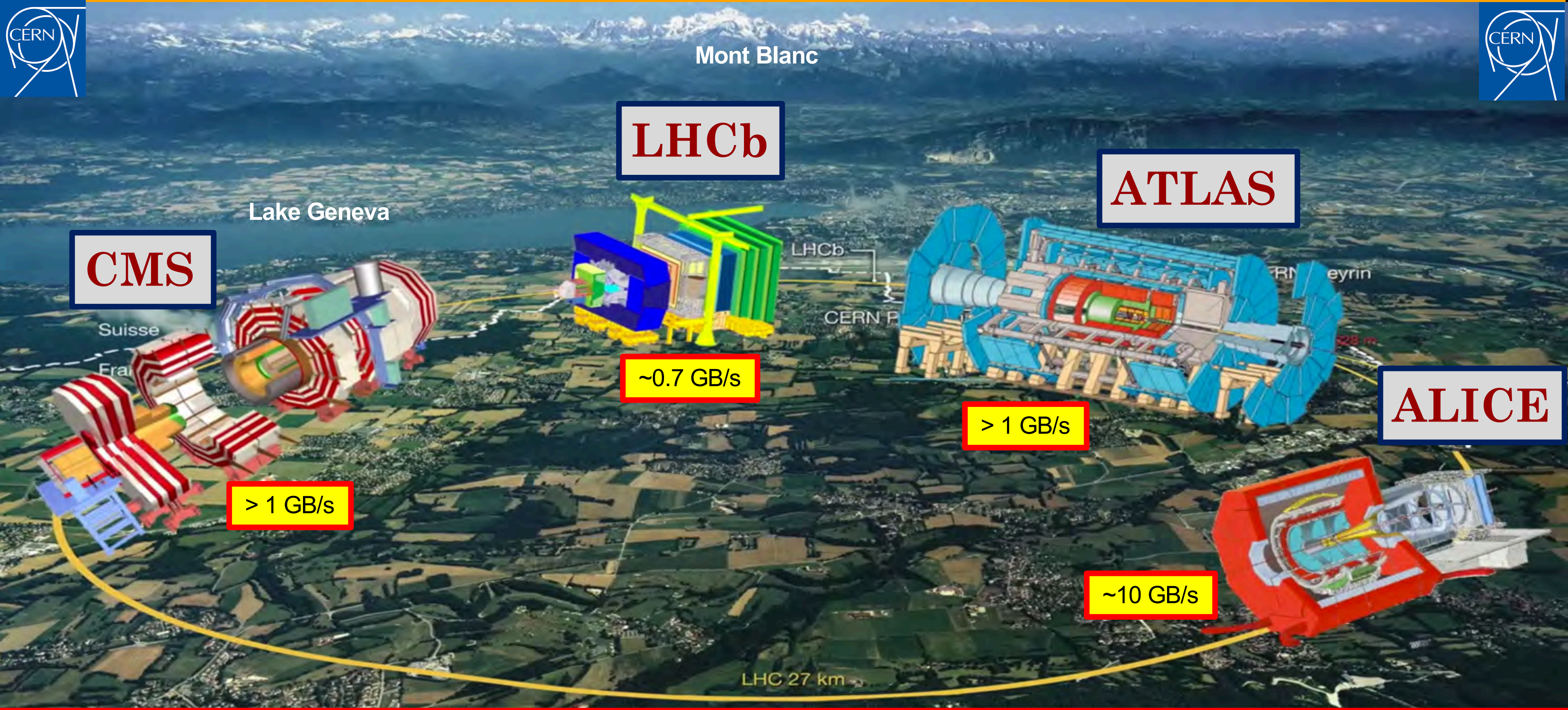
The Standard Model*

(a.k.a. our best theory of Nature)



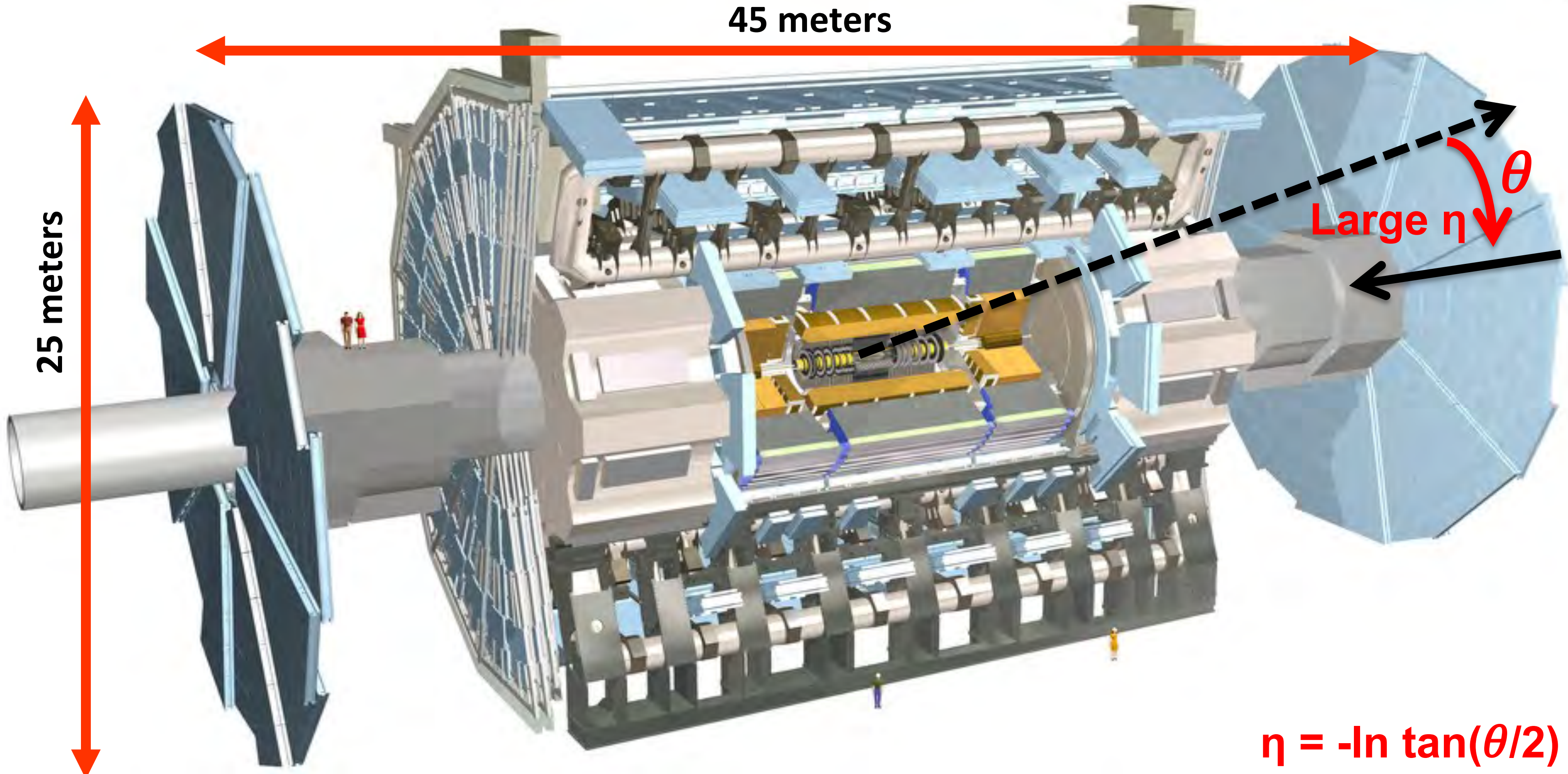
*Some assembly required. Gravity not included

LHC Experiments



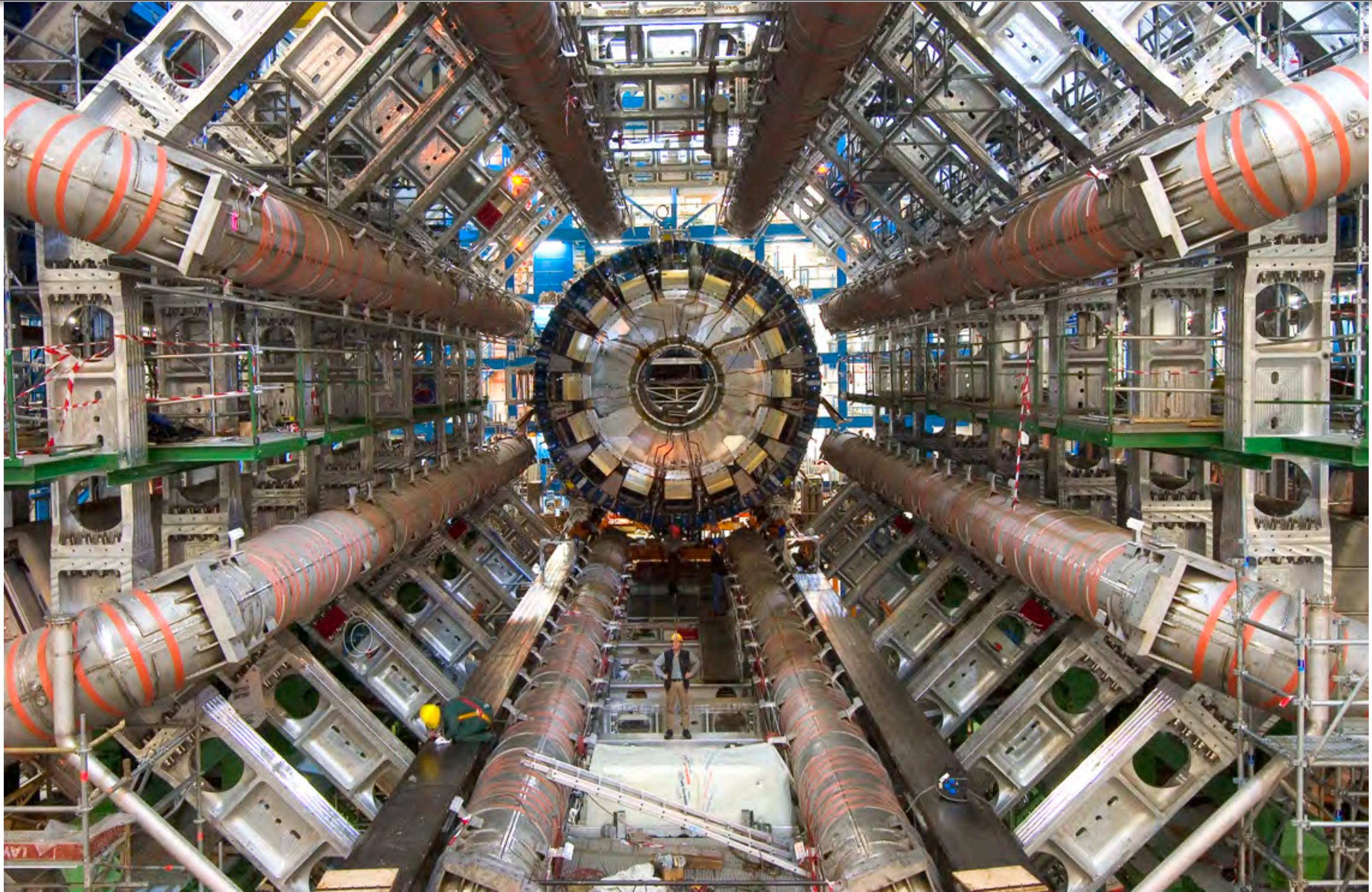
LHC Experiments generate 50 PB/year of science data (during Run 2)

ATLAS Detector

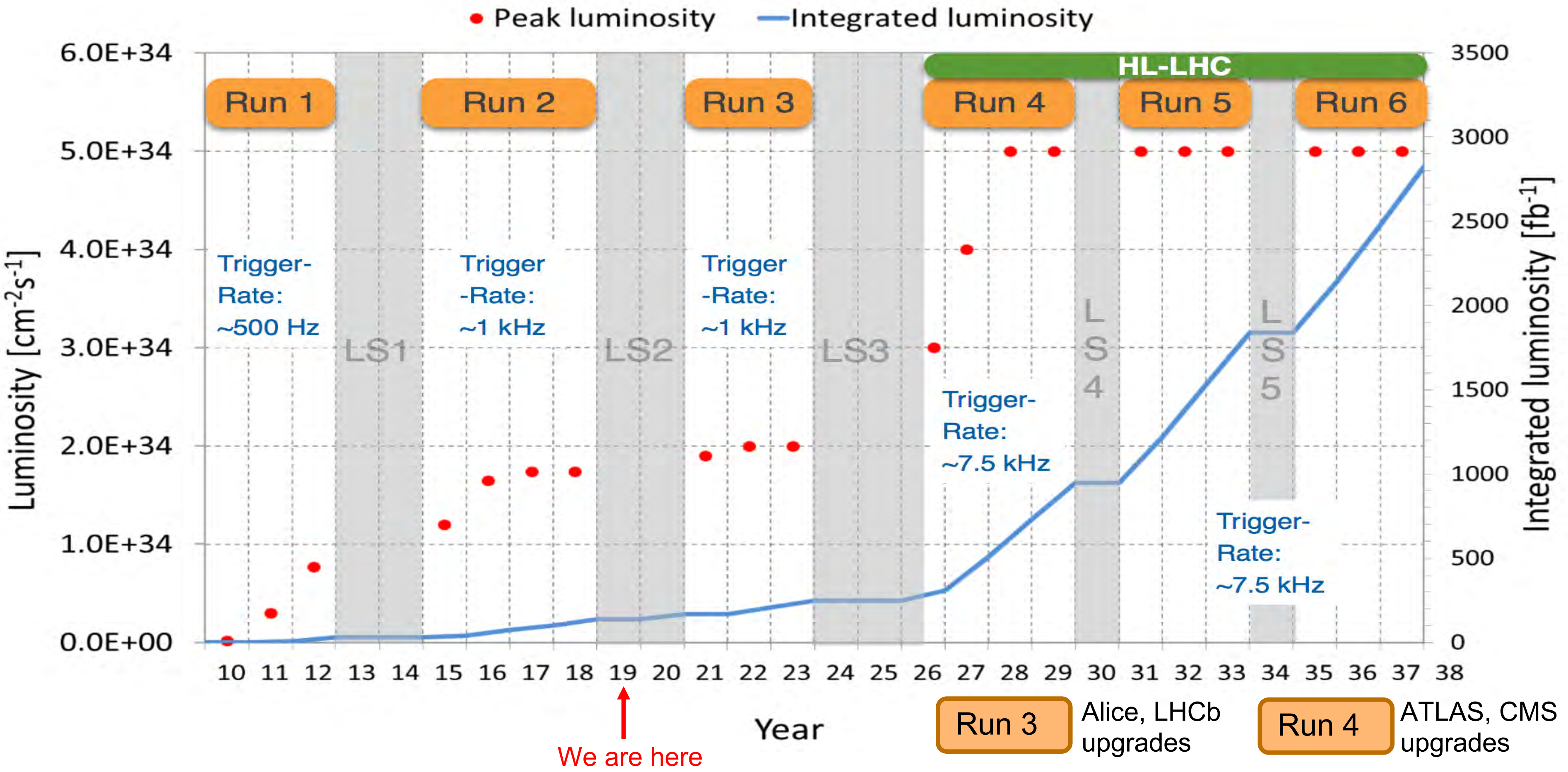


$\eta = -\ln \tan(\theta/2)$

ATLAS Detector



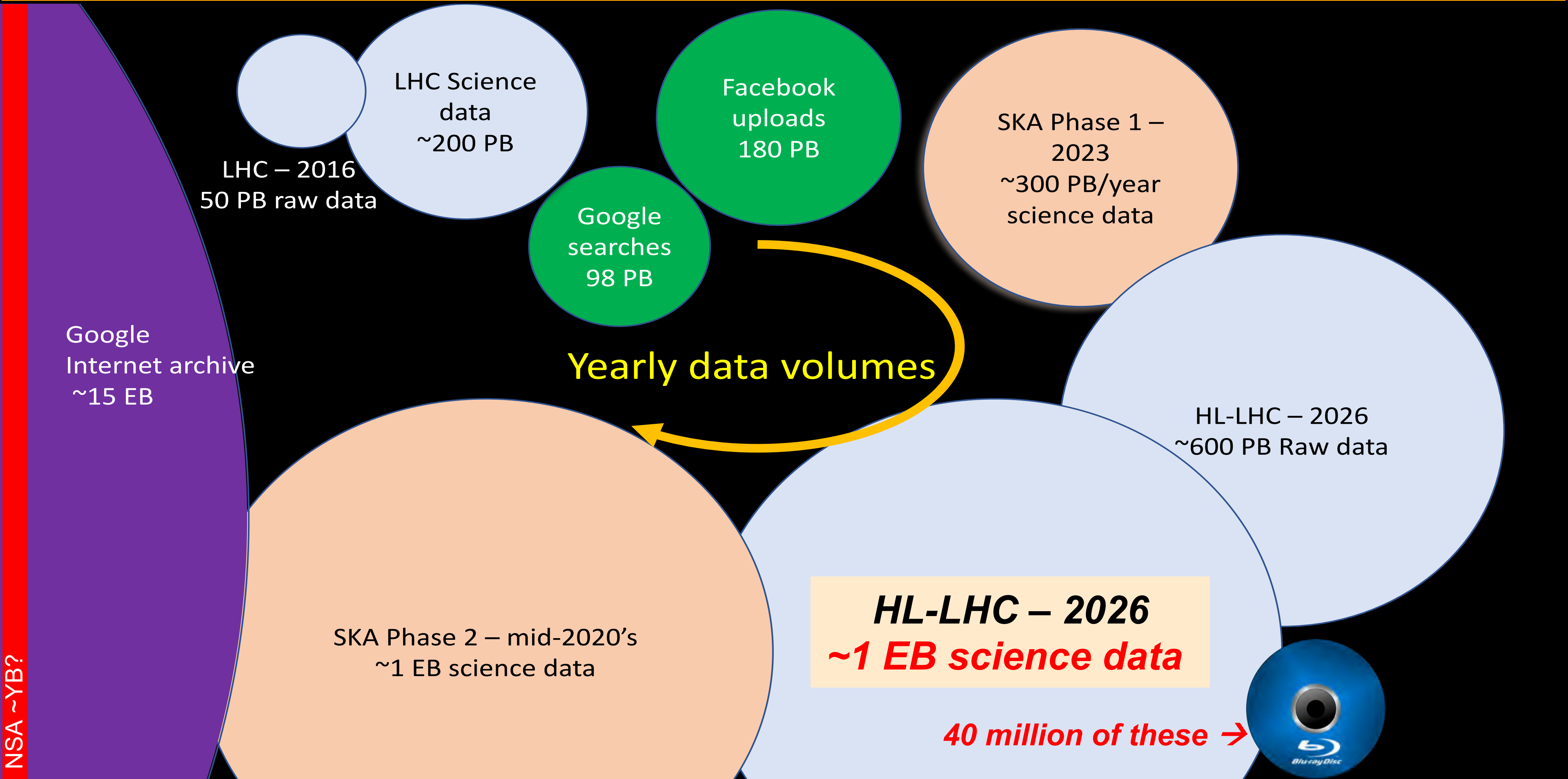
LHC Schedule



Run 3 Alice, LHCb upgrades

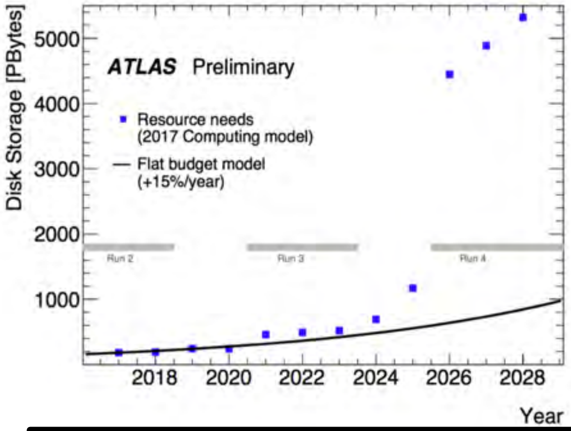
Run 4 ATLAS, CMS upgrades

LHC as Exascale Science



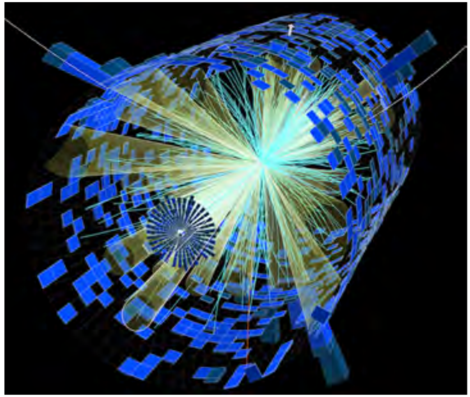
IRIS-HEP

Computational and Data Science Challenges of the High Luminosity Large Hadron Collider (HL-LHC) and other HEP experiments in the 2020s



The HL-LHC will produce exabytes of science data per year, with increased complexity: an average of 200 overlapping proton-proton collisions per event.

During the HL-LHC era, the ATLAS and CMS experiments will record ~10 times as much data from ~100 times as many collisions as were used to discover the Higgs boson (and at twice the energy).

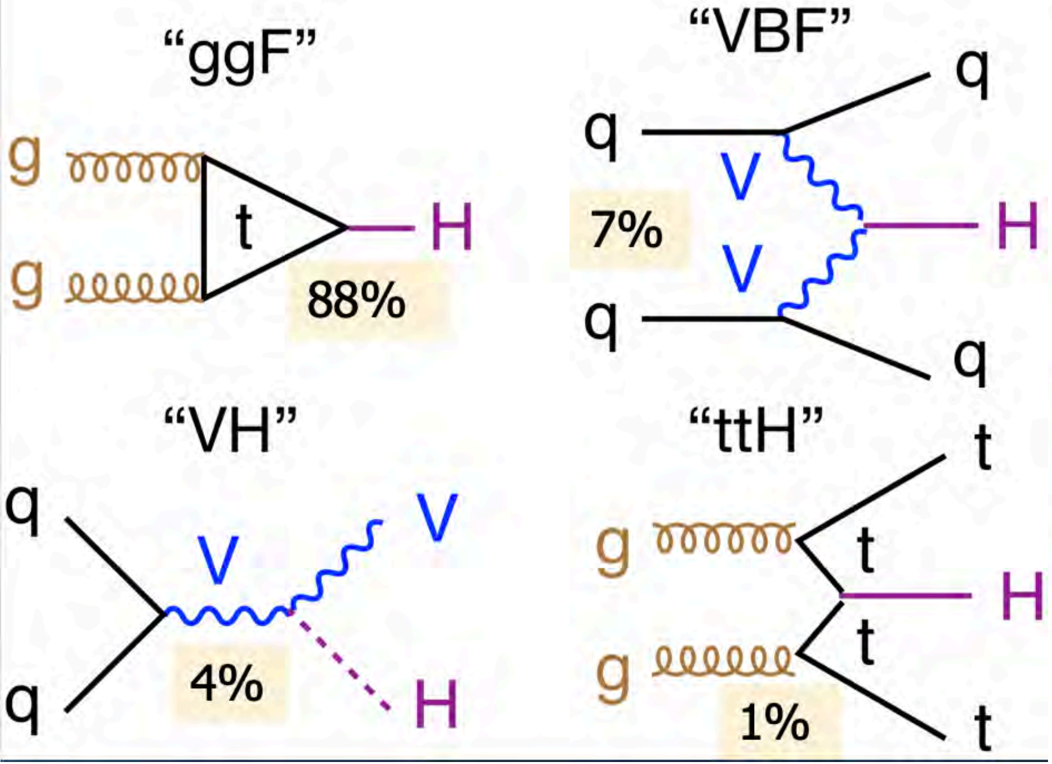


→ ***Institute for Research and Innovation in Software for High-Energy Physics*** **(IRIS-HEP)**

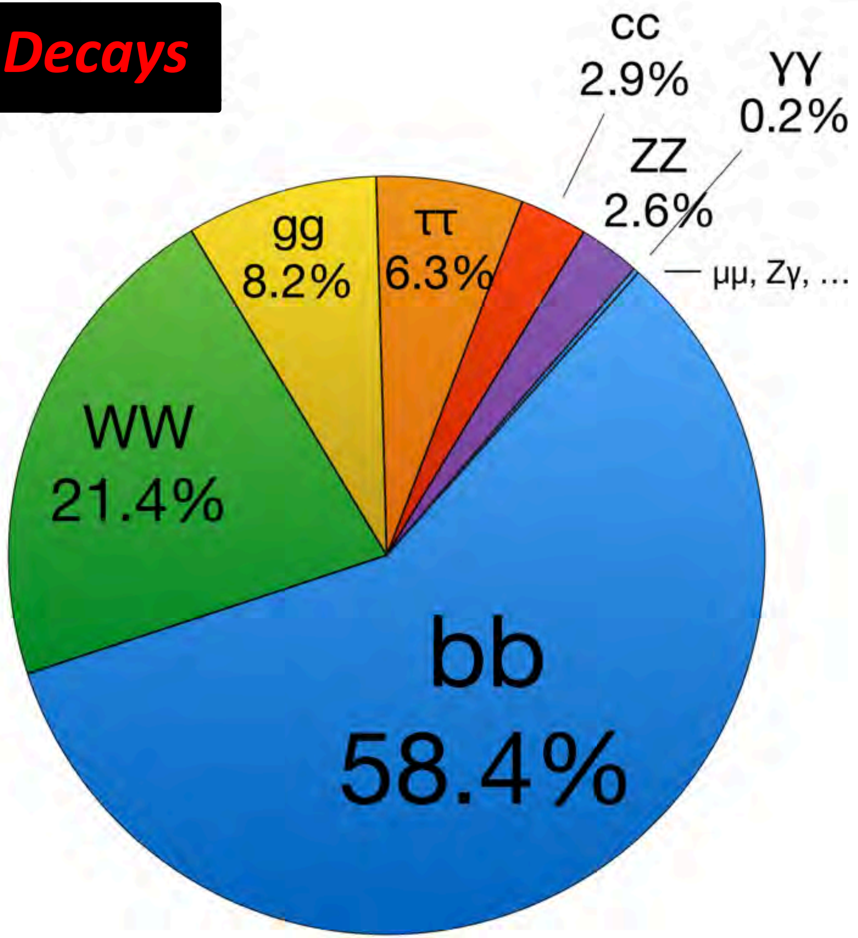
U. Illinois and NCSA are working within IRIS-HEP to develop innovative analysis systems and algorithms; and intelligent, accelerated data delivery methods to support low-latency analysis

Higgs Boson Production & Decay @ LHC

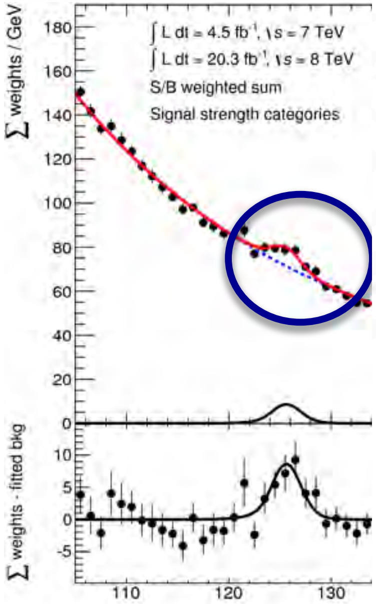
Production

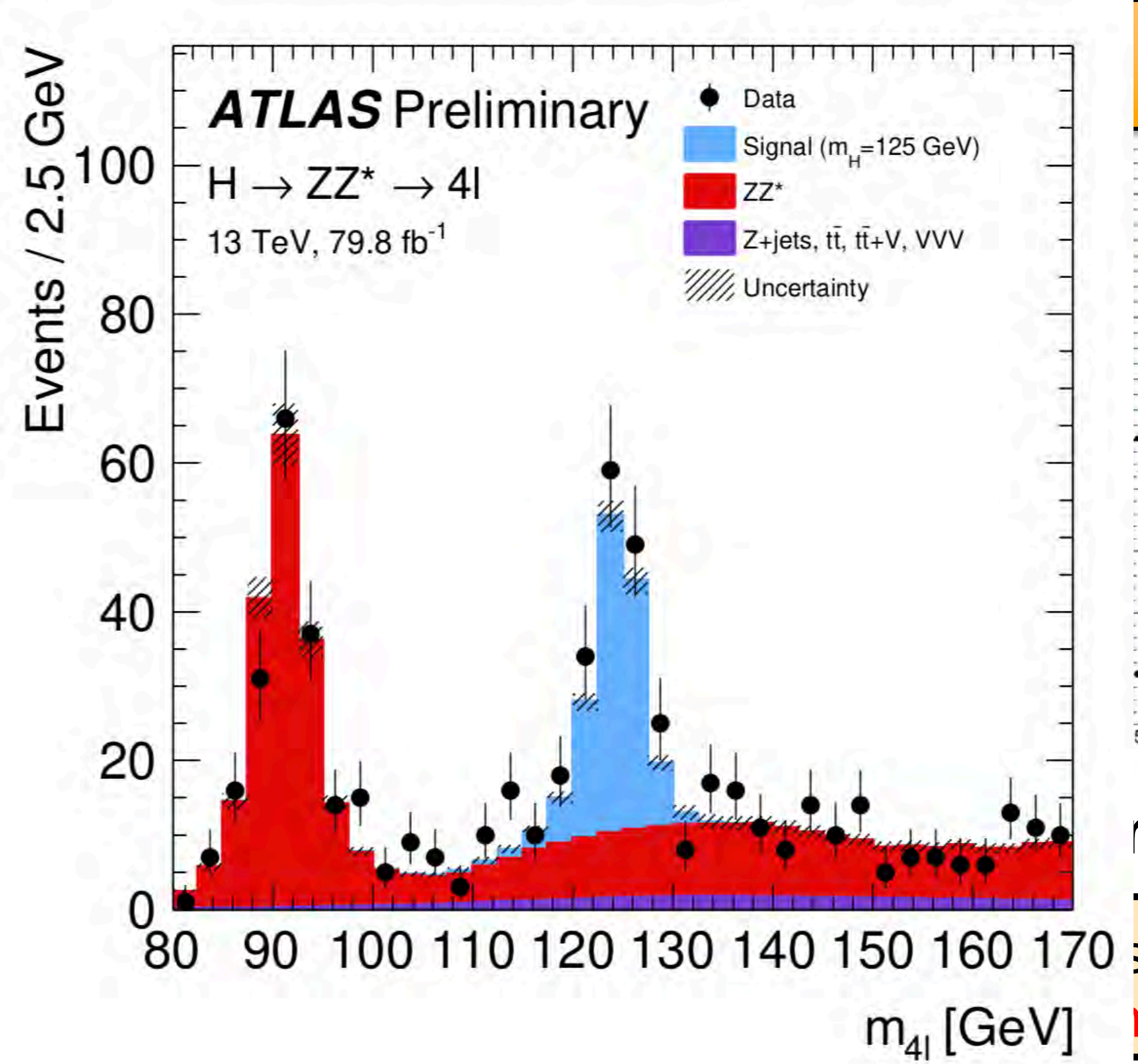


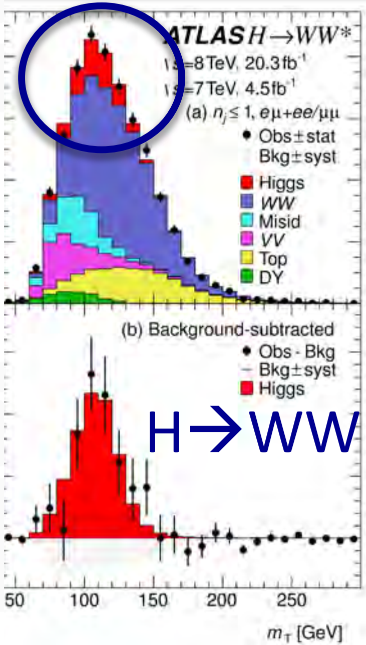
Decays



Hi… 2)





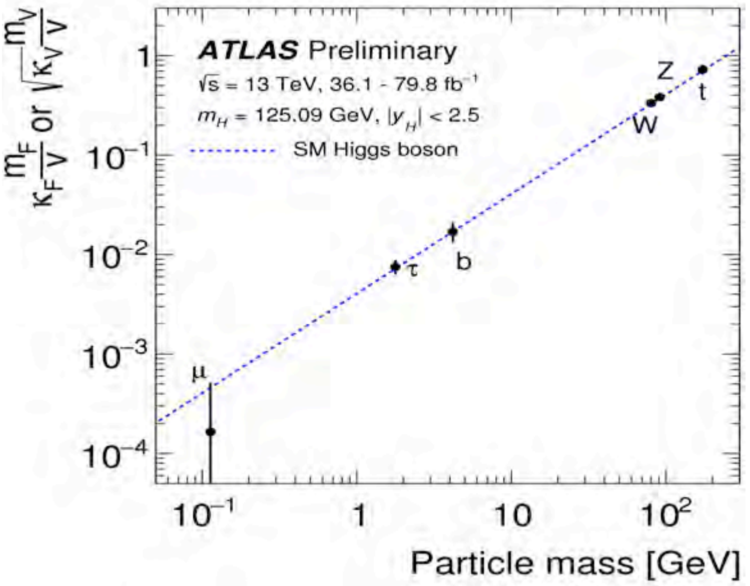


2013 Nobel p… rt

A new … son with mass
125 G… ne-SM physics

Higgs Boson Pair Production

- No new physics (yet) using this tool – The Higgs boson we discovered in 2012 *looks very much like the one in the Standard Model*

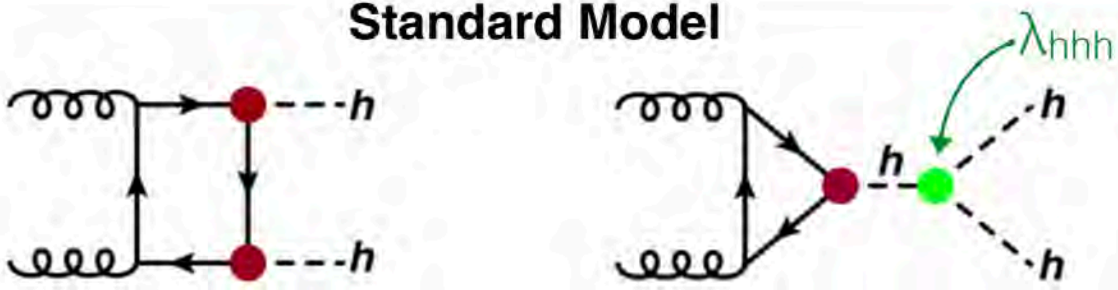


- But… "*Good luck seldom comes in pairs, but bad luck never walks alone*" (Chinese proverb)

- Next LHC frontier: ***hh* production**

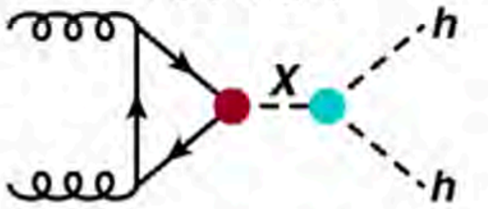
Higgs Boson Pair Production

Standard Model



33.4 fb @ 13 TeV (significant destructive interference)

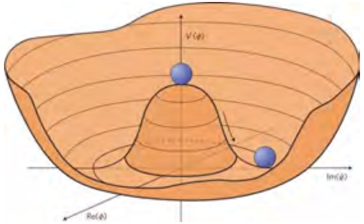
Beyond Standard Model



KK gravitons, heavy higgs, …

tthh vertex, coloured scalars, …

Measuring λ_{hhh} is important since it probes the shape of the Higgs boson potential



Standard Model:

$$\lambda_{hhh} = \frac{m_h^2}{2v^2}$$

Measuring *hh* production is interesting since it measures λ_{hhh}

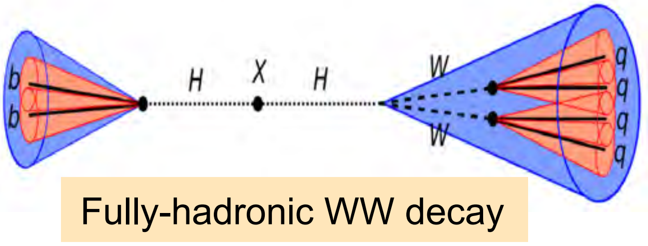
hh production is 1000x smaller than single *h* production (in SM)

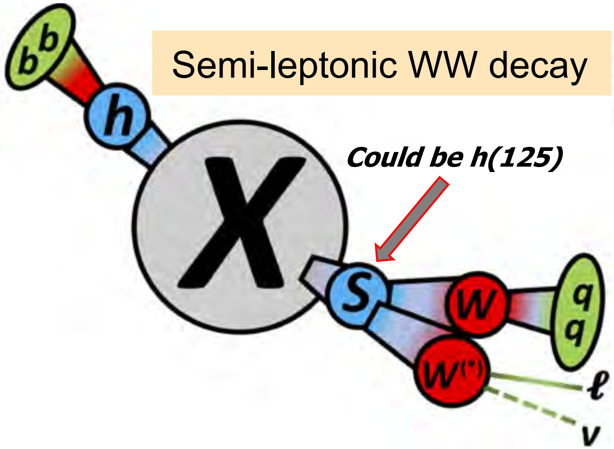
But… the *hh* rate can be enhanced by new physics!

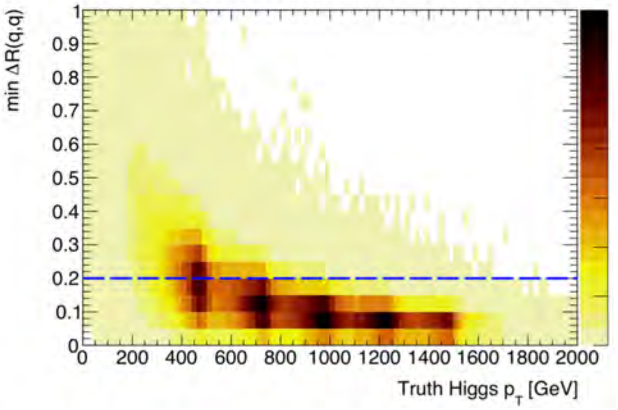
We are searching for *hh* production via the decay of heavy new particles

Resonant *hh* detection is Challenging

For heavy particles decaying to hh, the Higgs bosons are highly boosted and their decay products very close to one another







- We are using Machine Learning to identify boosted Higgs bosons from *X*→*hh* production, focusing on $h \rightarrow WW^{(*)}$ tagging
- We are using **Blue Waters** to develop, test and optimize this ML-based tagger, in collaboration with Indiana & Gottingen U.

Matrix Element Method

Probability density („weight") for event **x** given hypothesis α?

Possible uses:

Sample likelihood → M.L. parameter fit $\prod_{i \in events} P(\mathbf{x}_i \vert \alpha)$	Neyman-Pearson discriminant [4] → Hypothesis testing/search for rare process $P(\mathbf{x} \vert S) / \sum_i r_i P(\mathbf{x} \vert B_i)$

... Can be computed!

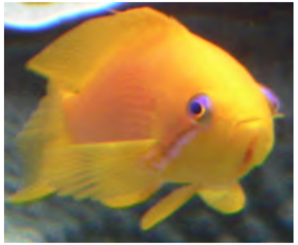
$$P(\mathbf{x}|\alpha) = \frac{1}{A_\alpha \sigma_\alpha} \int d\Phi(y) \frac{dx_1 dx_2}{x_1 x_2 s} f(x_1) f(x_2) |\mathcal{M}_\alpha(y, x_1, x_2)|^2 W(\mathbf{x}|y) \epsilon_\alpha(y)$$

Theoretical hypothesis (Matrix Element) + Parton shower + Detector (transfer functions, efficiencies) + Experimental information (whole event **x**)

We are using **Blue Waters** to develop Deep Neural Networks to approximate this important calculation → a sustainable method

Scalable Cyberinfrastructure for Science

- We use Blue Waters to perform large-scale data processing, simulation & analysis of ATLAS data
  - E.g. 35M events were processed over ~1wk period in 2018
  - See our paper on HPC/HTC integration here here
- We using Blue Waters to develop HPC integration for scalable cyberinfrastructure to increase the discovery reach of data-intensive science using artificial intelligence and likelihood-free inference methods → SCAILFIN & IRIS-HEP



Scalable Cyberinfrastructure for Artificial Intelligence and Likelihood-Free Inference

K.-P.-H. Anampa[1] *J. Bonham*[2] *K. Cranmer*[4] (PI) *B. Galewsky*[3] *M. Hildreth*[1] (PI)
D. S. Katz[2,3] (co-PI) *C. Kankel*[1] *I.-E. Morales*[4] *H. Mueller*[4] (co-PI) ***M. Neubauer***[2,3] (PI)

[1]*University of Notre Dame* [2]*University of Illinois* [3]*National Center for Supercomputing Applications* [4]*New York University*

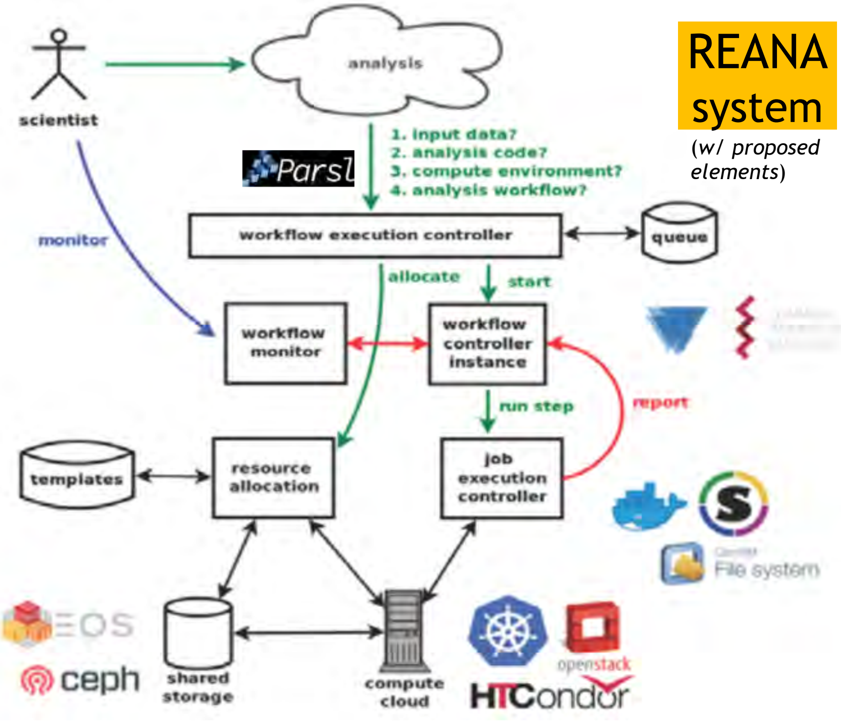
scailfin.github.io

OAC-1841456, 1841471, 1841448

Main Goal

- To deploy artificial intelligence and likelihood-free inference methods and software using scalable cyberinfrastructure (CI) to be integrated into existing CI elements such as the *REANA system*, to ***increase the discovery reach of data-intensive science***

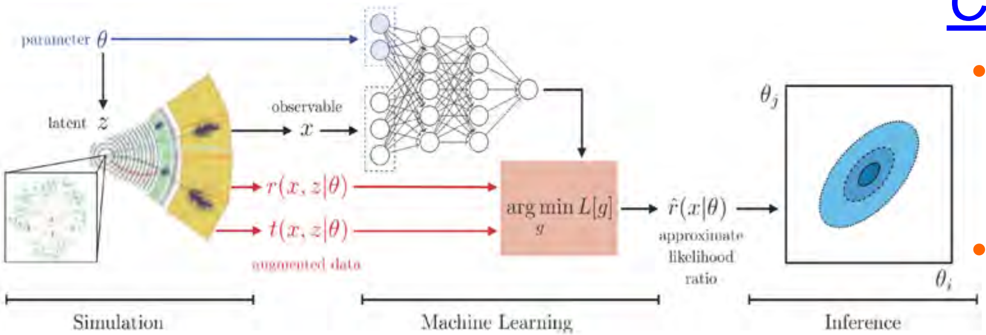
REANA system
(w/ proposed elements)



The SCAILFIN Project

Likelihood-Free Inference

- Methods used to constrain the parameters of a model by finding the values which yield simulated data that closely resembles the observed data

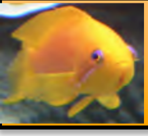


Catalyzing Convergent Research

- Current tools are limited by a lack of scalability for data-intensive problems with computationally-intensive simulators
- Tools will be designed to be scalable and immediately deployable on a diverse set of computing resources, including ***HPCs***
- Integrating common workflow languages to drive an optimization of machine learning elements and to orchestrate large scale workflows **lowers the barrier-to-entry** for researchers from other science domains

Science Drivers

- Analysis of data from the Large Hadron Collider is the **primary science driver**, yet the technology is sufficiently generic to be applicable to other scientific efforts



SCAILFIN Project Activities

REANA Deployment and Application Development

- Established a shared REANA development cluster at NCSA
- REANA implementation of new ML applications (e.g. MadMiner & *t*-quark tagging)
- Ongoing studies of Matrix Element Method approximations using deep neural networks

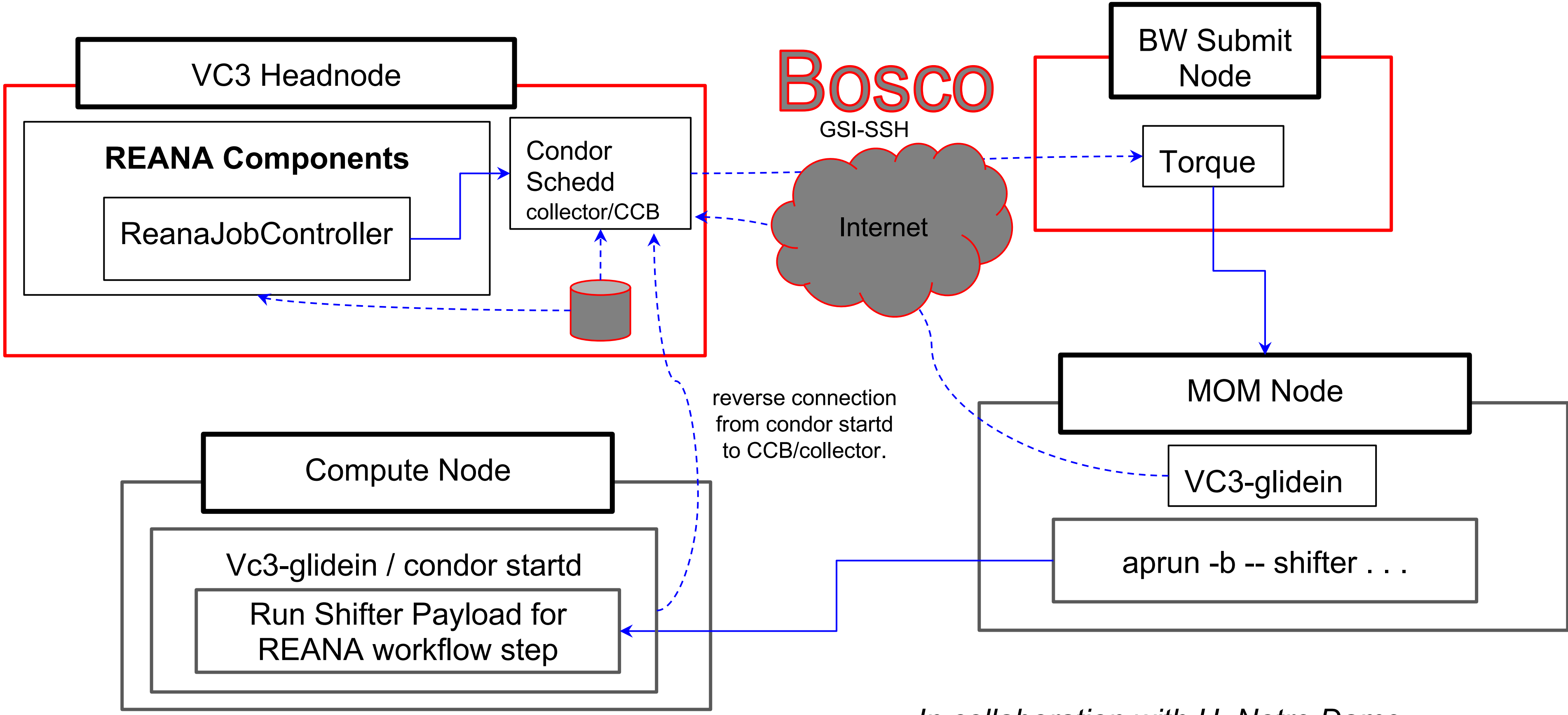
Parsl Integration

- *Parsl*: Annotate python functions to enable them to be run in parallel on laptops, OSG, supercomputers, clouds, or a combination without otherwise changing the original python program and developing capability to export workflow to CWL
- We have ported a REANA example workflow to Parsl

HPC Integration

- Using VC3 infrastructure to configure and set up edge service head node on a cluster at ND
- REANA runs on head node, submits jobs to HPC batch queue using HTCondor
- Jobs are now successfully submitted to worker nodes
  - "Hard problems" and new infrastructure ~finished; "simple issues" like file and executable transfer still to be solved for full chain to work
- Integration and testing on the **Blue Waters** Supercomputer is well underway

SCAILFIN on Blue Waters



In collaboration with U. Notre Dame

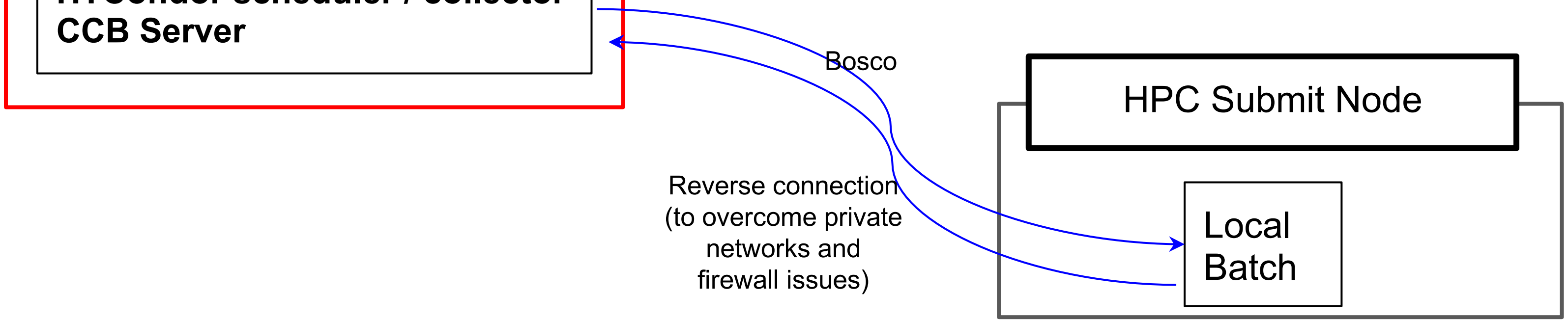
Summary

- We have used the Blue Waters supercomputer to advance frontier science in high-energy particle physics
  - Development and optimization of deep-learning methods for booted Higgs boson identification and ab-initio event-likelihood determination for signal and background hypotheses
  - Development of scalable cyberinfrastructure for ML applications on HPC
- Having a Blue Waters allocation has also helped us establish new collaborators and strengthen existing partnerships
- ***We would like to thank the NSF and the Blue Waters team for delivering and operating such a wonderful resource on the University of Illinois campus!***

SCAILFIN and VC3

We utilize VC3 for remote connections to clusters.

- Virtual Clusters for Community Computation allows users to create a “virtual cluster” with a user defined head-node.



- This head-node will have a local REANA-CLUSTER running with a modified job-controller component specially tuned to launch jobs to the head-node’s HTCondor scheduler.

- VC3 will launch HTCondor glide-ins to the remote HPC facility to accept jobs submitted to the local Scheduler. BOSCO will translate requirements from HTCondor to a variety of common HPC schedulers (PBS/Torque, SLURM, SGE, etc.)