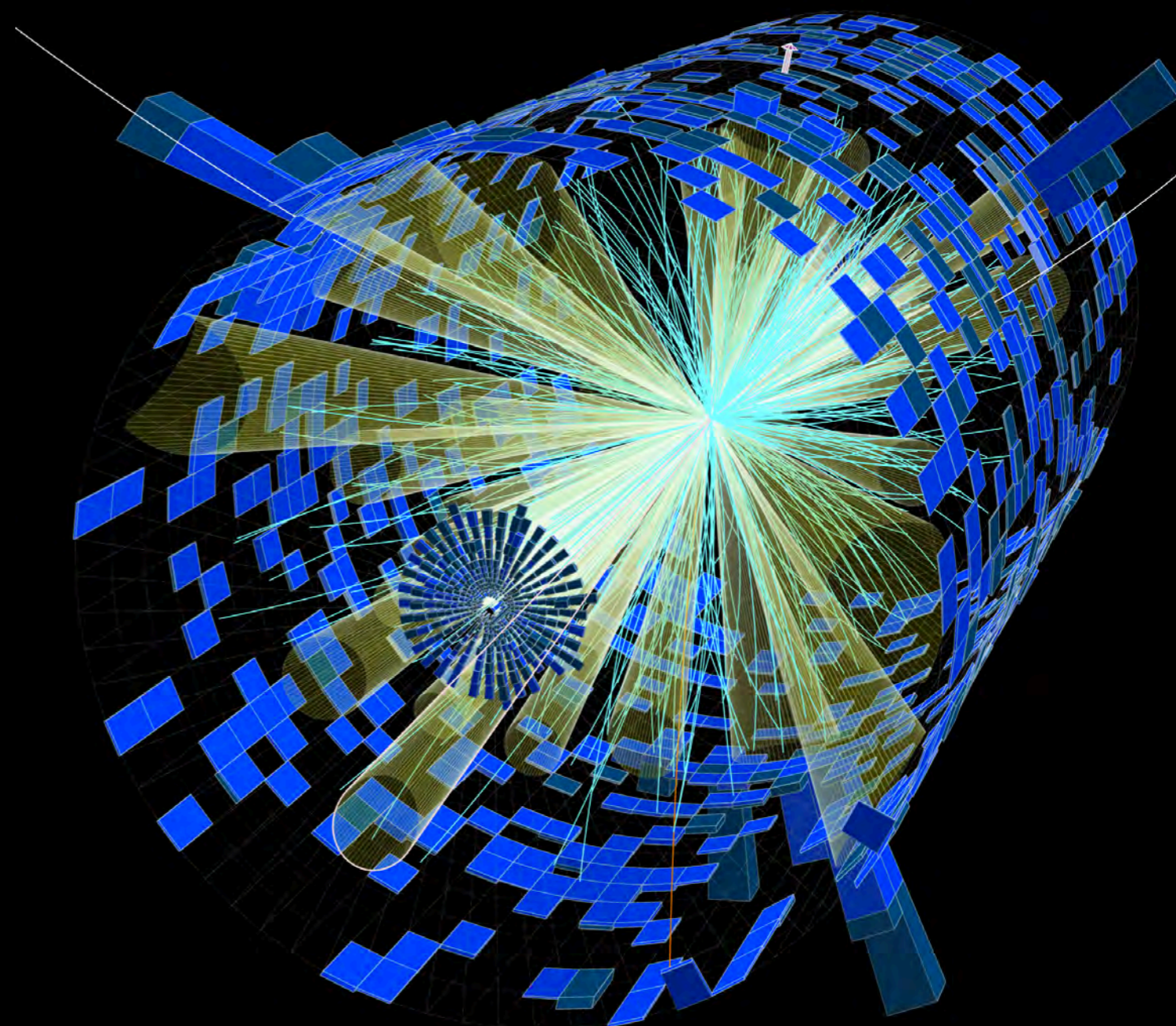




# ACCELERATING PHYSICS

WITH ADVANCED, AUTOMATED WORKFLOWS



**@KyleCranmer**

New York University  
Department of Physics  
Center for Data Science  
CILVR Lab

# THE LAST 10 YEARS

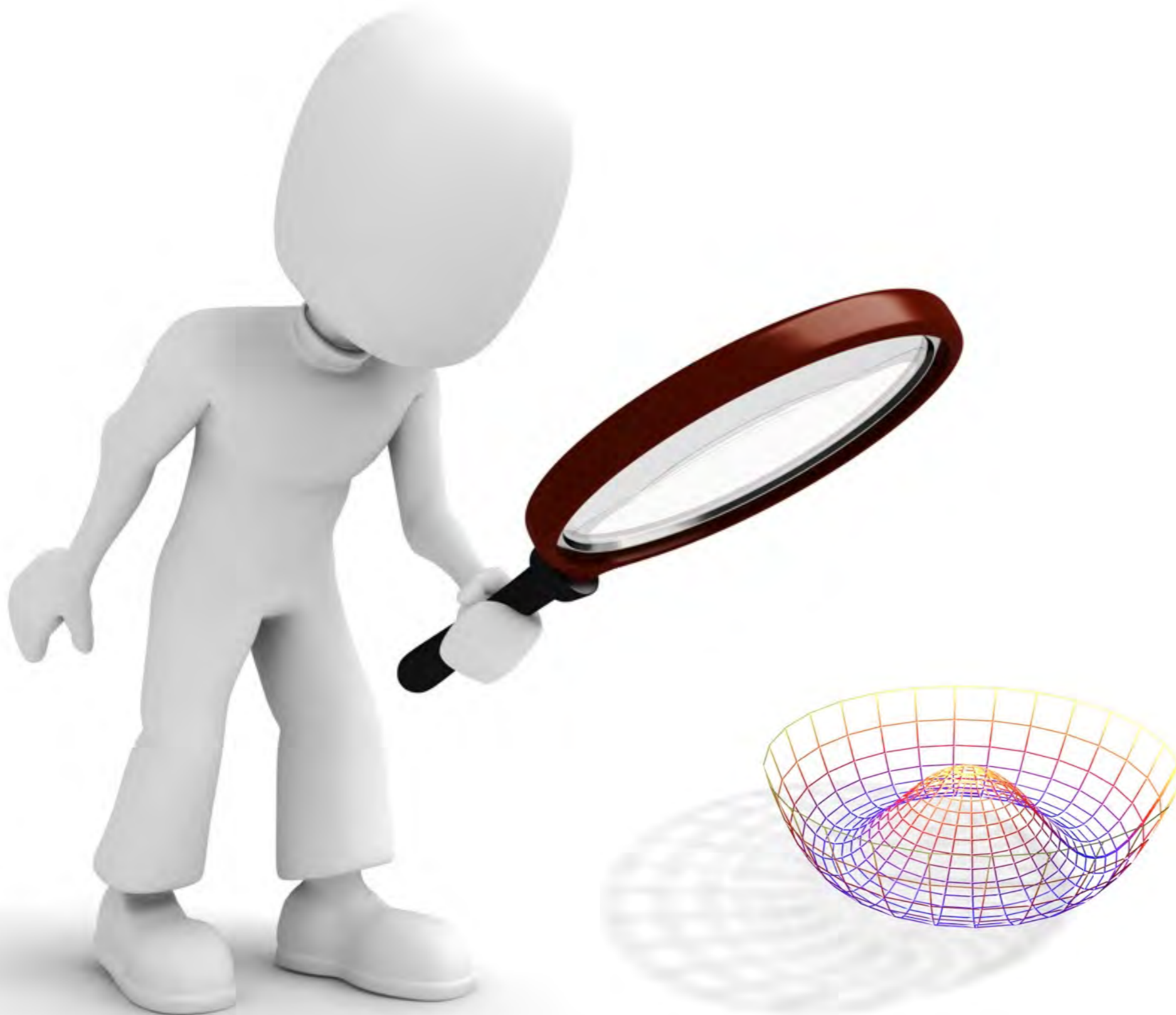
Before pontificating on the next 10 years, let me share experience with two projects over the last 10 years

- Collaborative Statistical Modeling
- Reinterpreting the searches for new physics at the LHC

Collaborations in High-Energy Physics have ~3000 people

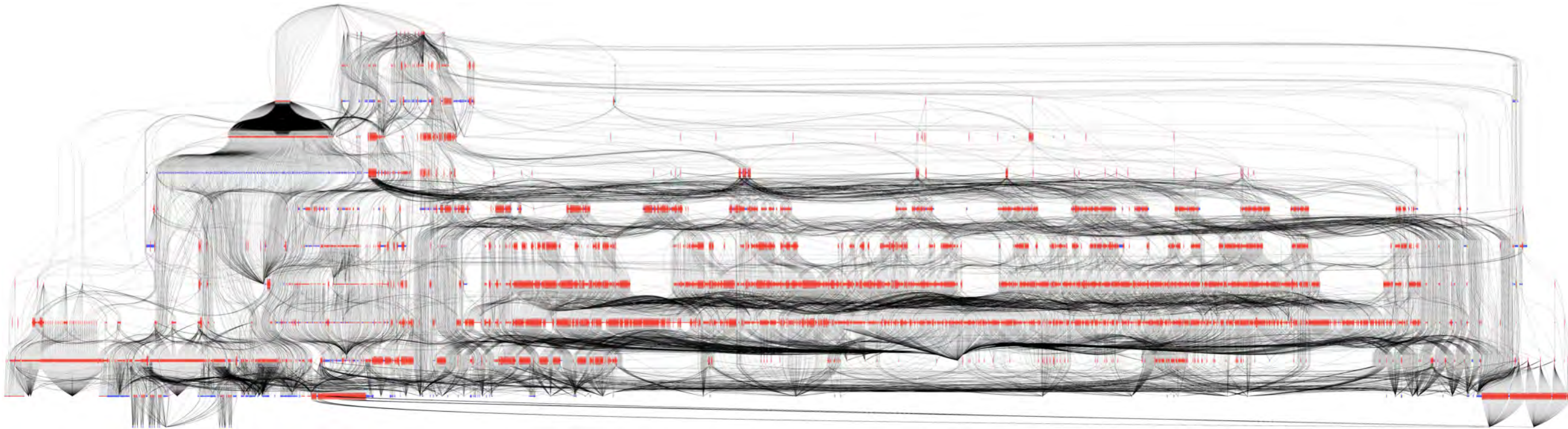
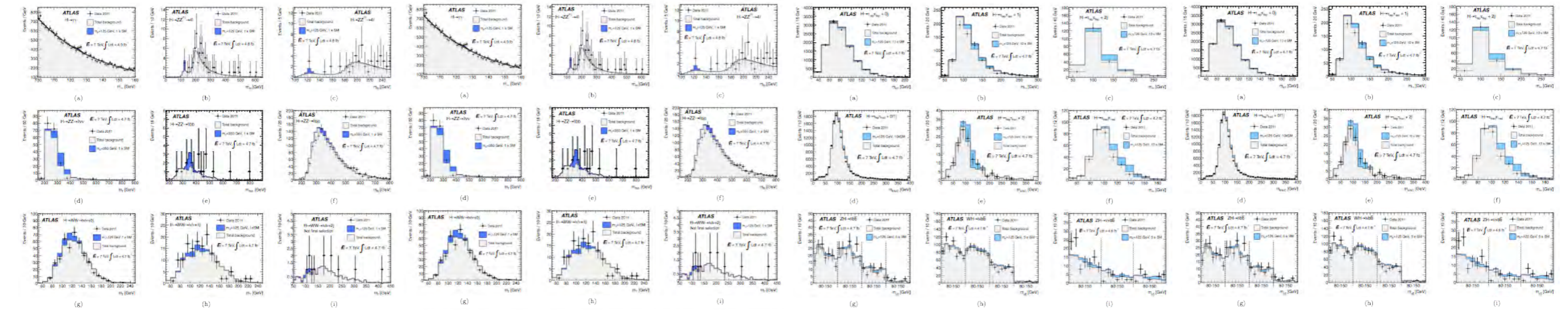
- Microcosm for science with sociological challenges for change

Searching for the Higgs required combining evidence from many sources.  
Not a meta-analysis, but a single combined statistical analysis



# COLLABORATIVE STATISTICAL MODELING

Technical solution sharing low-level likelihoods introduced c. 2010

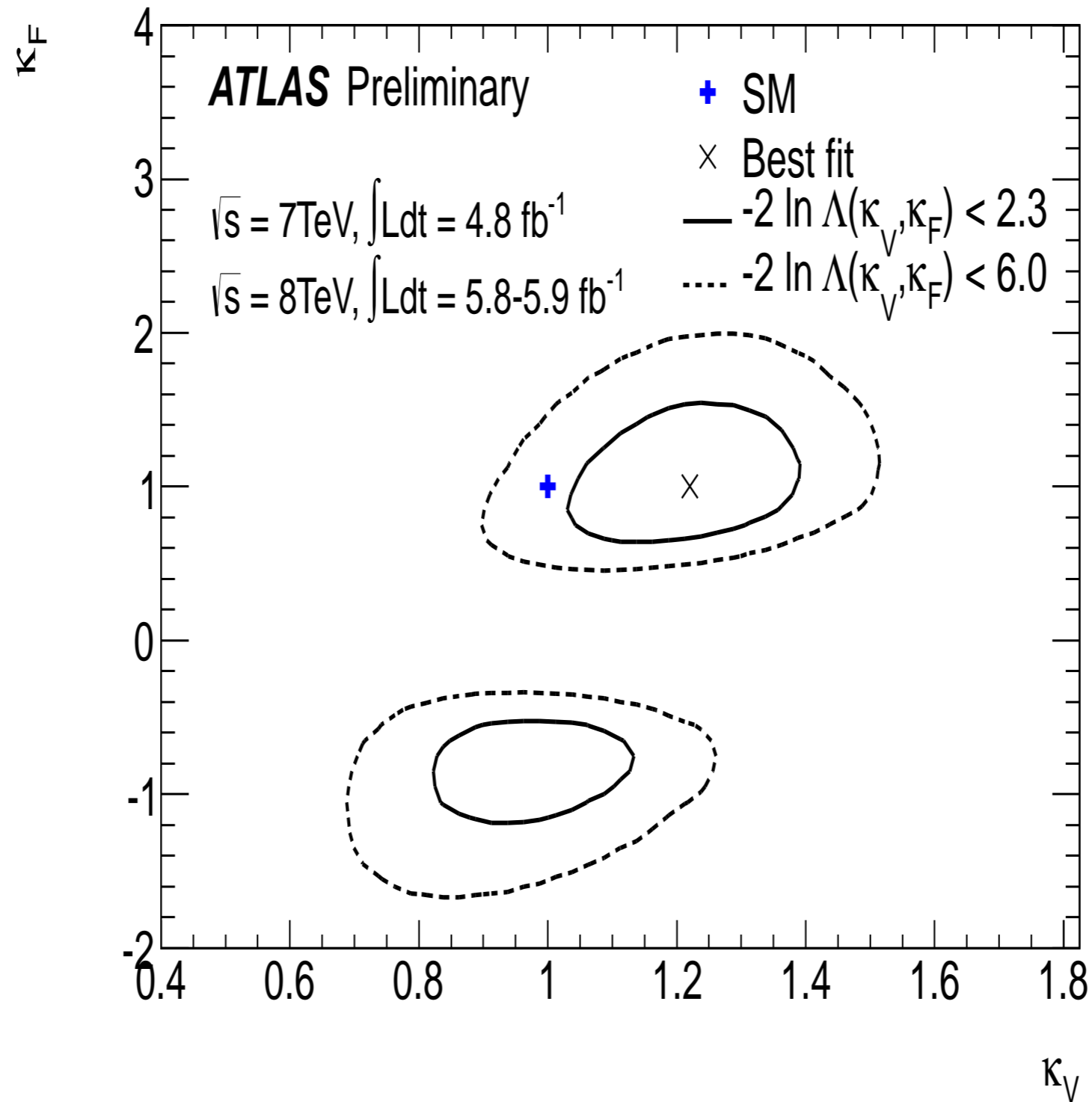


$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathcal{S}} f_p(a_p | \alpha_p)$$



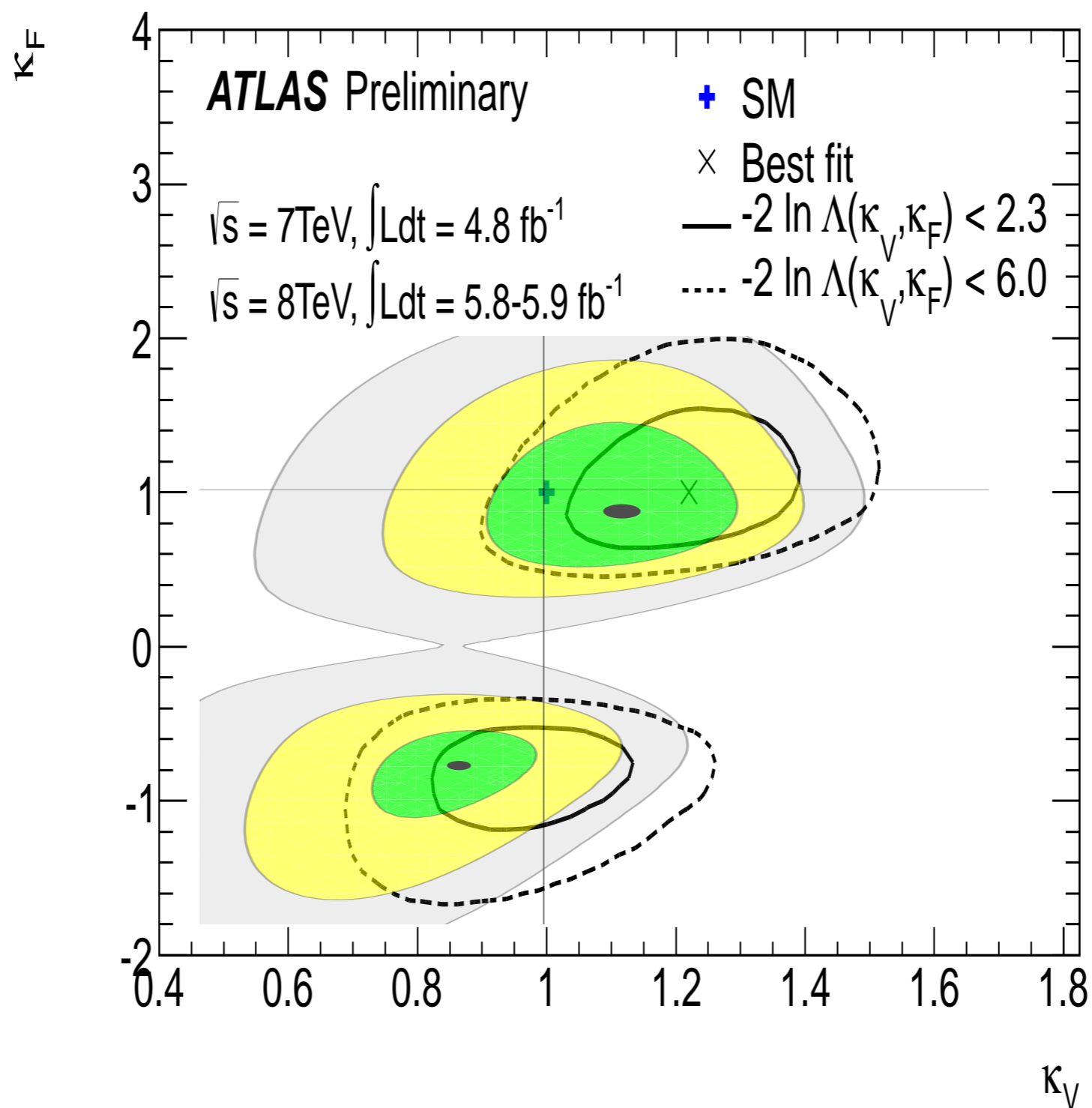
# REPRODUCIBILITY PROBLEM

*Not possible for others to reproduce results from paper.*



# REPRODUCIBILITY PROBLEM

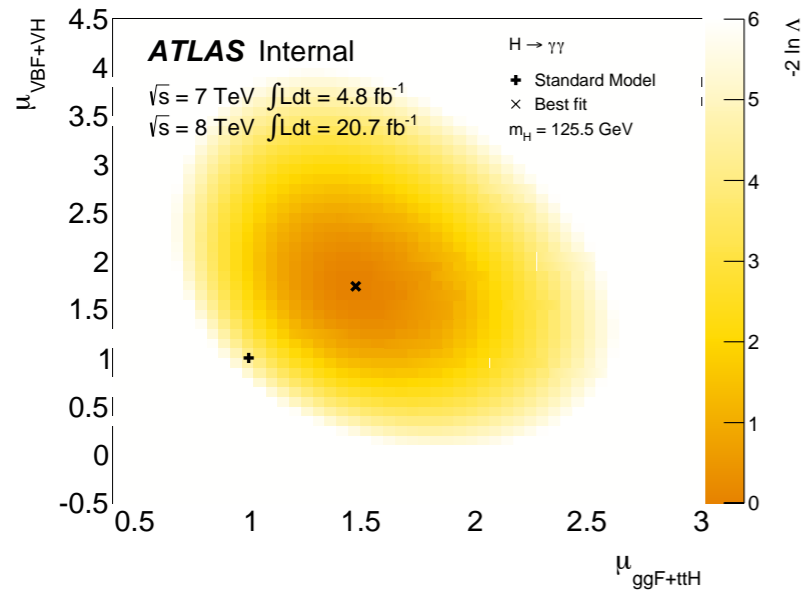
*Not possible for others to reproduce results from paper.*



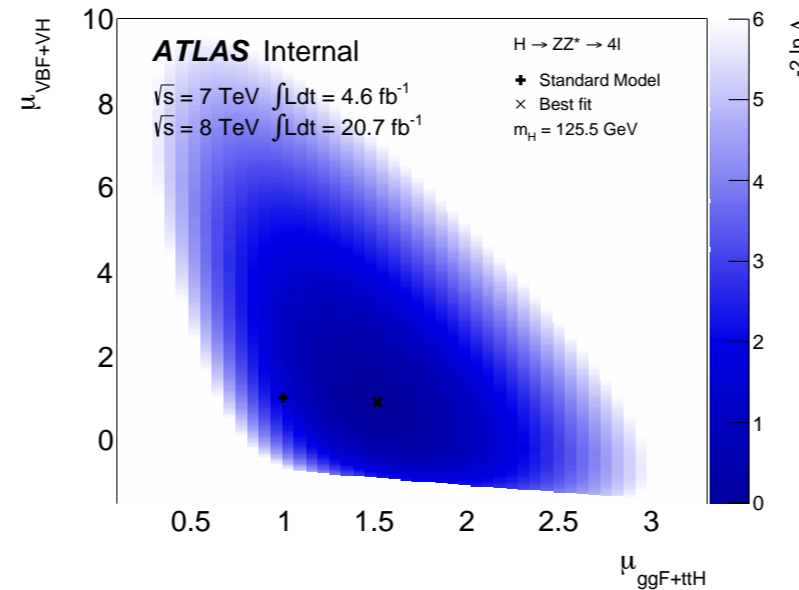
# WHAT INFO TO SHARE AND HOW TO RETRIEVE IT

**First step:** publish likelihood scans for communicating LHC Higgs results.

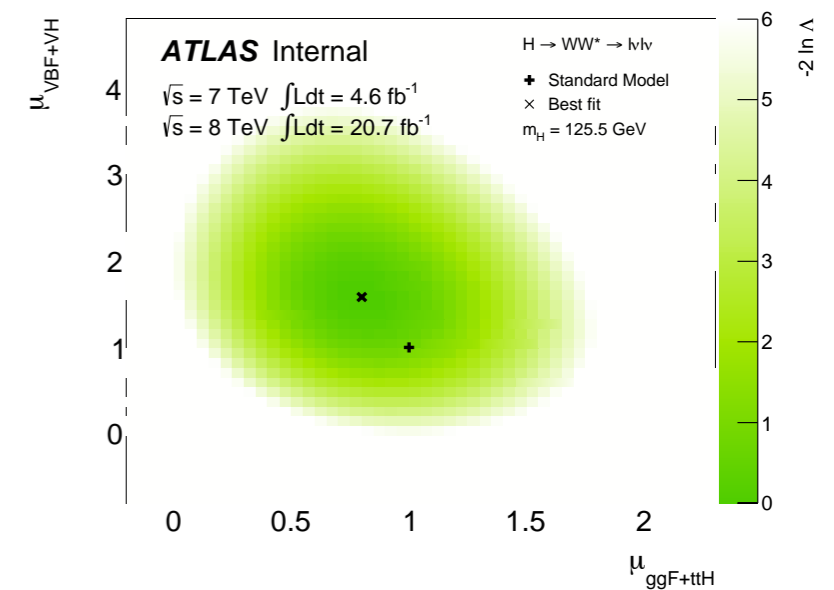
<http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44>



<http://doi.org/10.7484/INSPIREHEP.DATA.RF5P.6M3K>



<http://doi.org/10.7484/INSPIREHEP.DATA.26B4.TY5F>



Welcome to [INSPIRE](http://inspirehep.net), the High Energy Physics information system. Please direct questions, comments or contact [feedback@inspirehep.net](mailto:feedback@inspirehep.net).

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HEP



23

Blogged by 3  
Tweeted by 6

[Click for more details](#)

Information Citations (7) Files

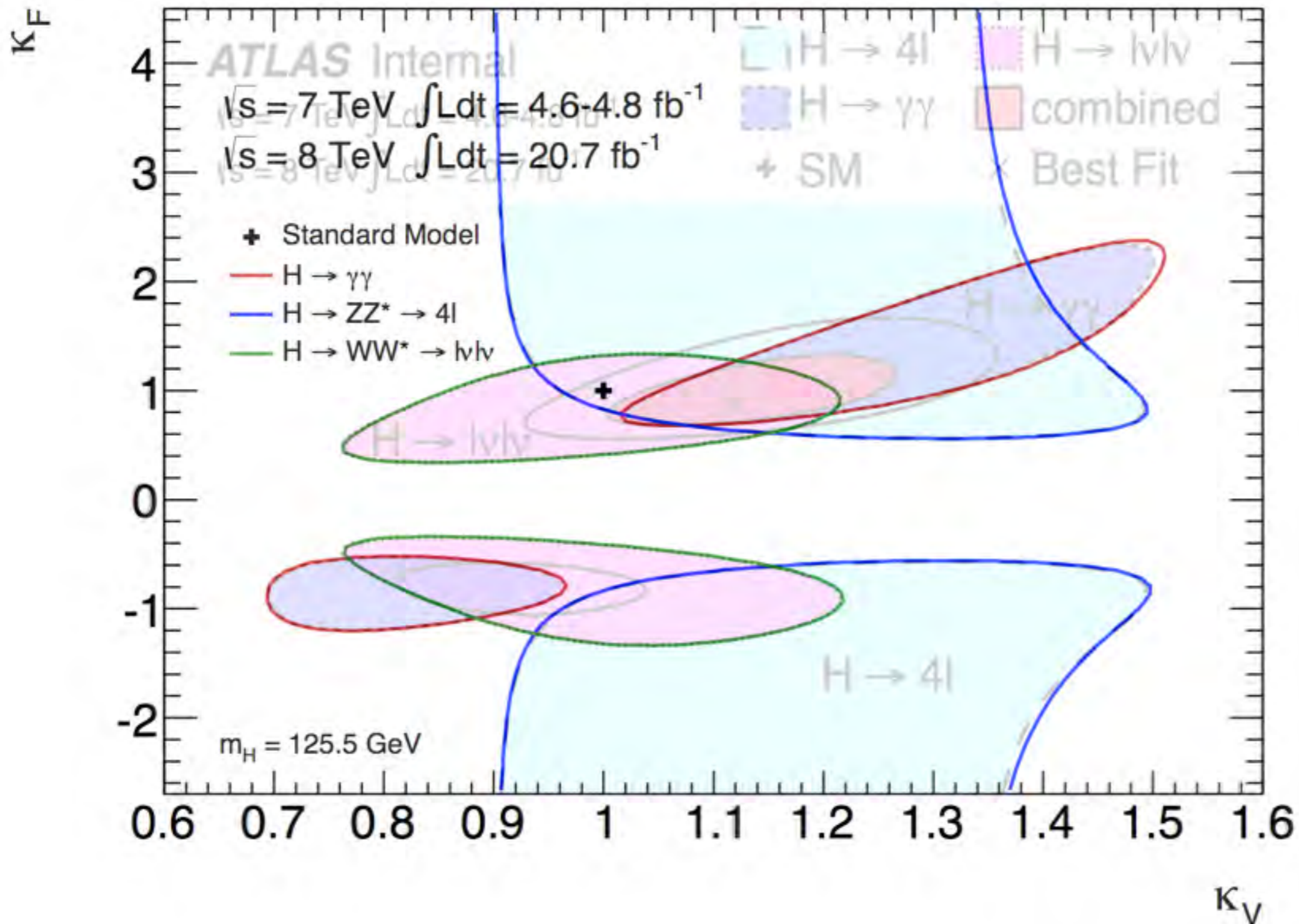
**Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC**

ATLAS Collaboration (Aad, Georges (Freiburg U.) [...]) [Show all 2923 authors](#)

Cite as: ATLAS Collaboration ( 2013 ) HepData, <http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44>

# LIKELIHOODS ON HEPDATA

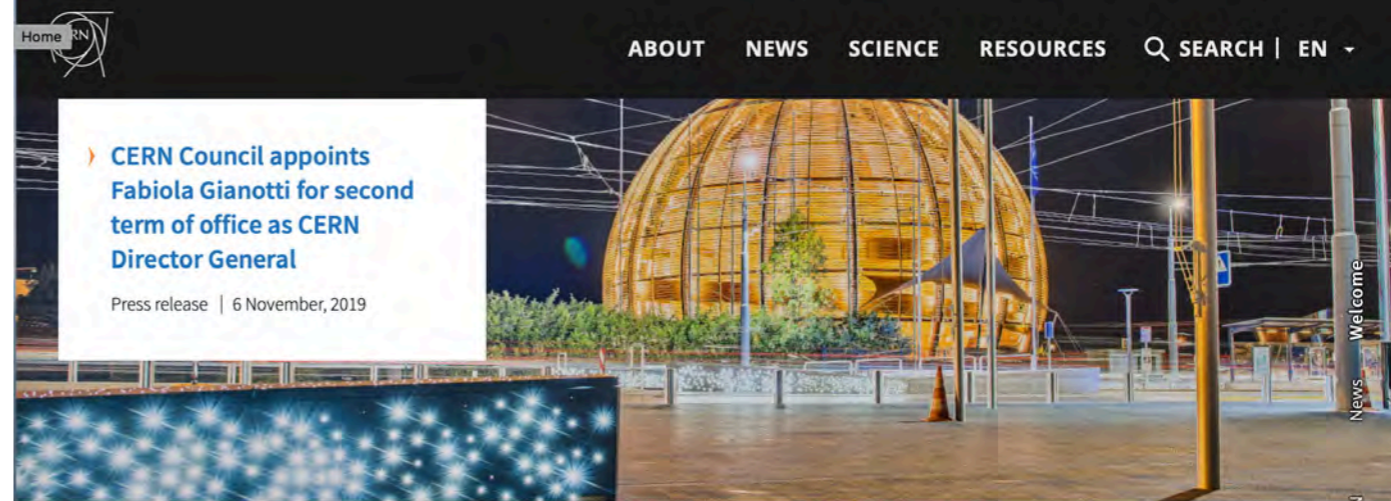
*Reproducing derived results from original paper!*



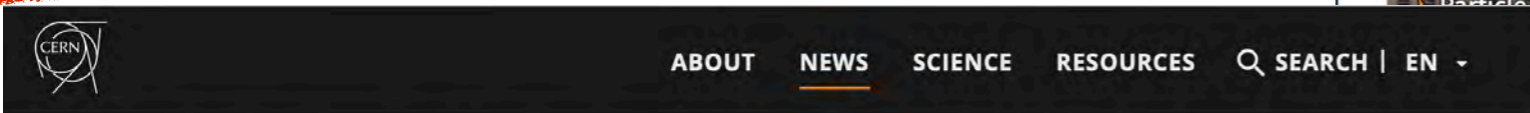
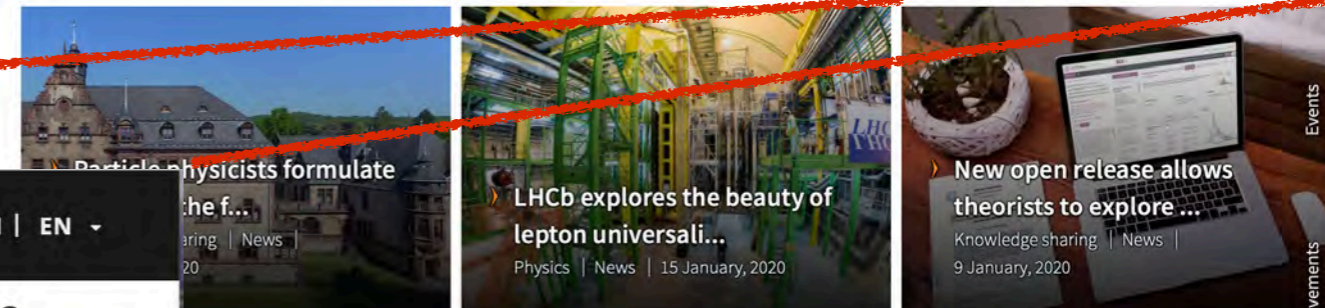
# MAKING IT STANDARD

10 years later: community embraces publishing likelihoods as a standard

- Moved to JSON schema



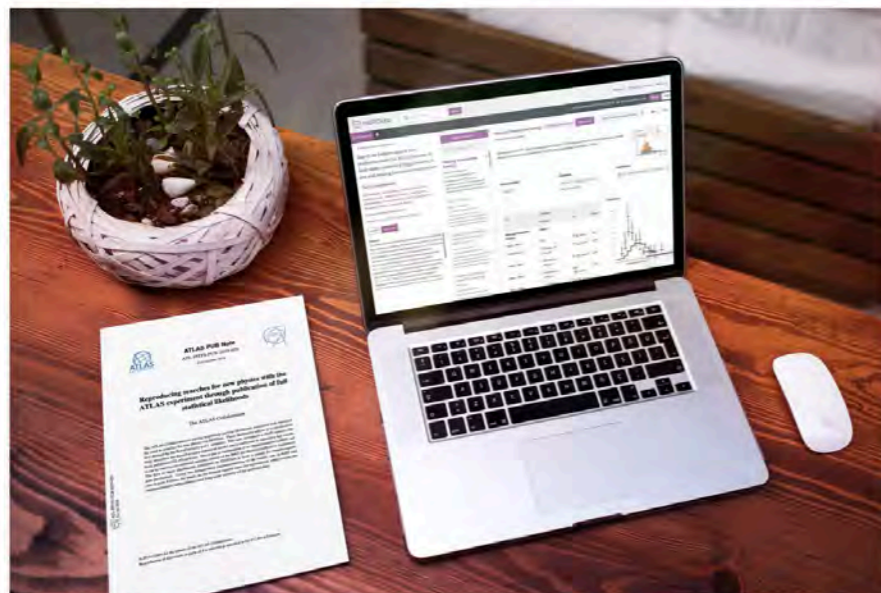
## LATEST NEWS



## New open release allows theorists to explore LHC data in a new way

The ATLAS collaboration releases full analysis likelihoods, a first for an LHC experiment

9 JANUARY, 2020 | By Katarina Anthony

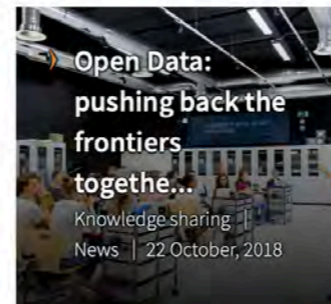


Explore ATLAS open likelihoods on the HEPData platform (Image: CERN)

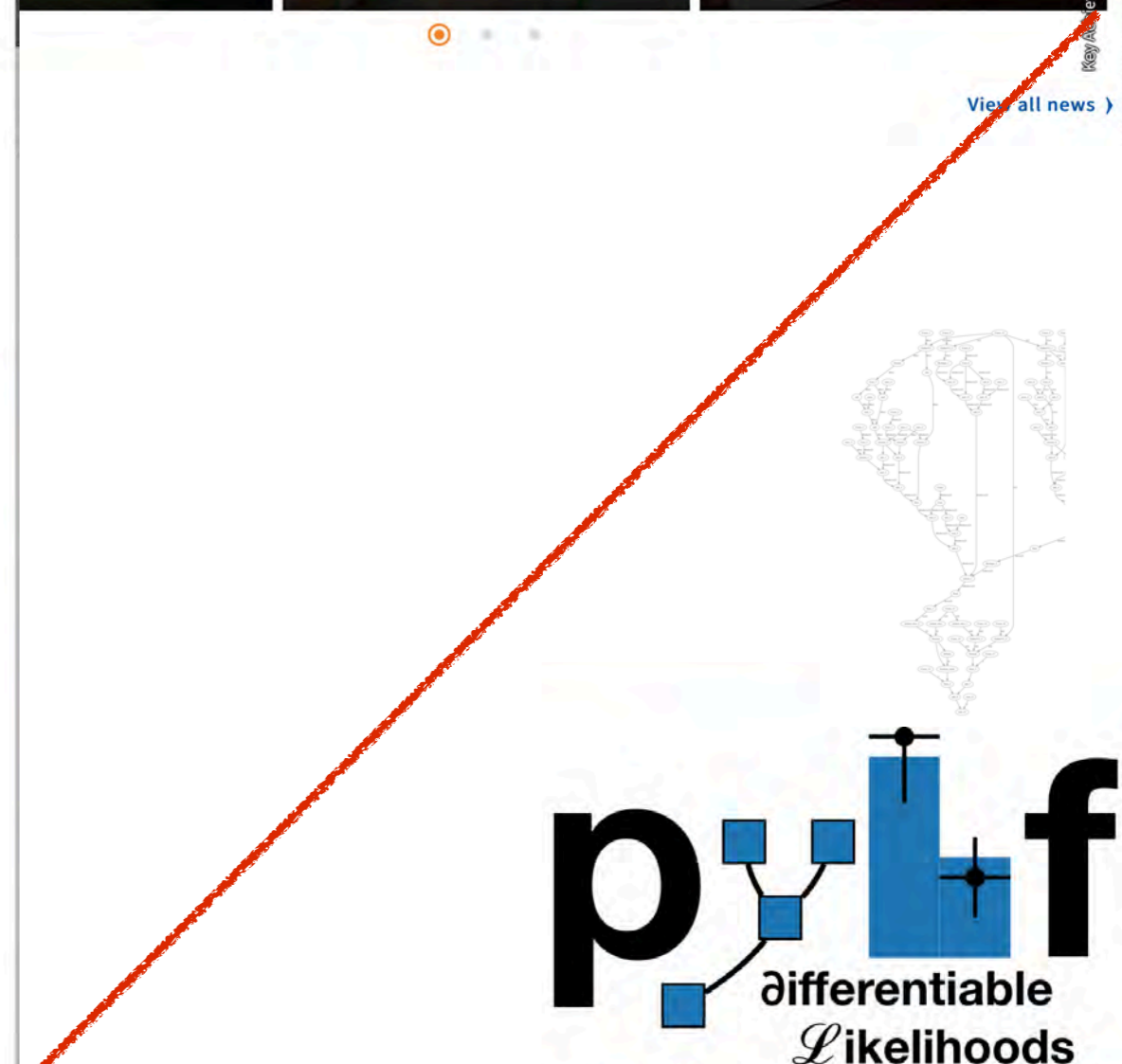
What if you could test a new theory against LHC data? Better yet, what if the expert knowledge needed to do this was captured in a convenient format? This tall order is now on

Display a menu y from the ATLAS collaboration, with the first open release of full analysis likelihoods

### Related Articles



[View all news >](#)





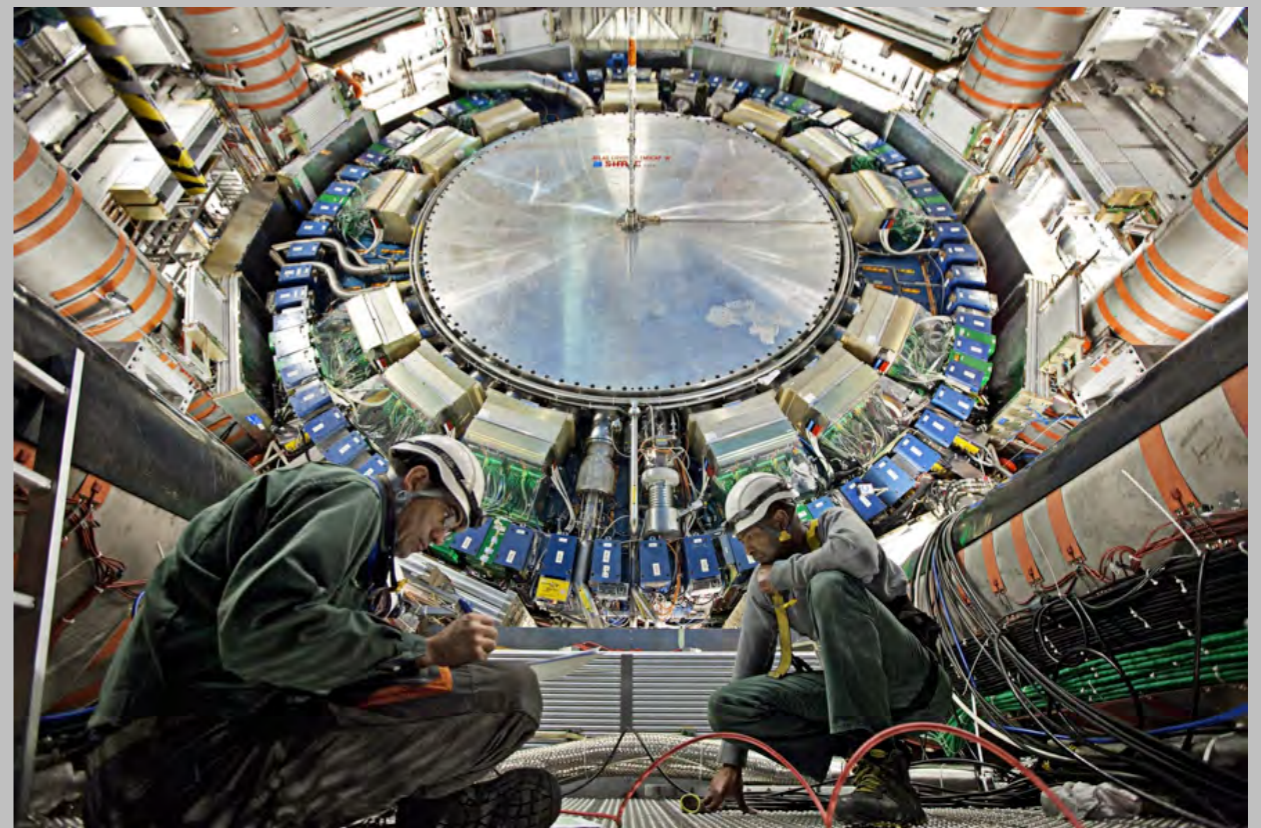
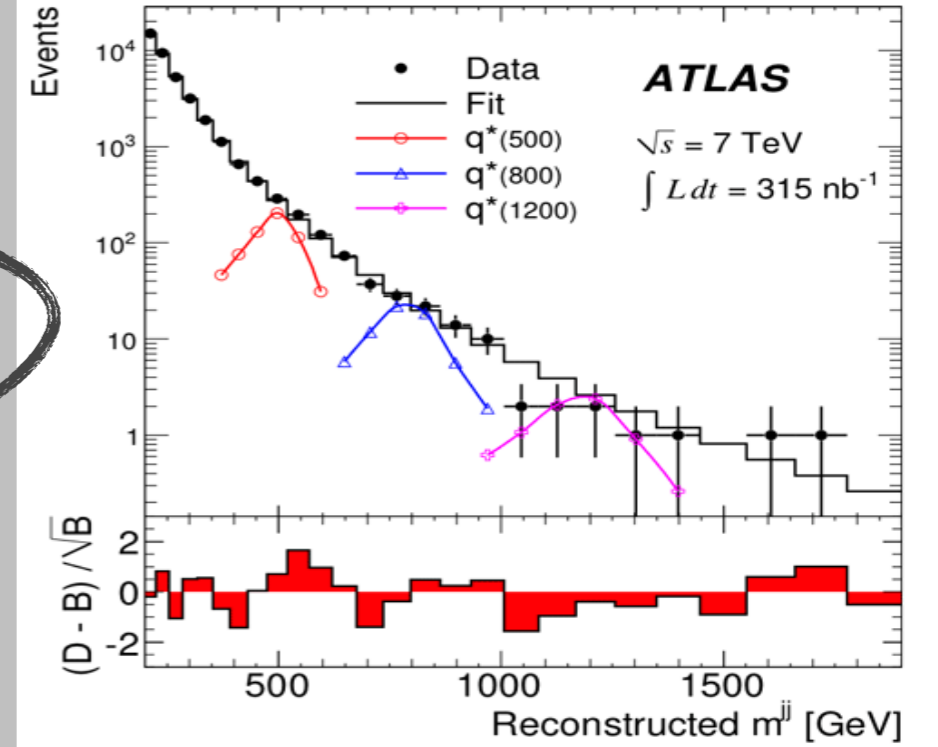
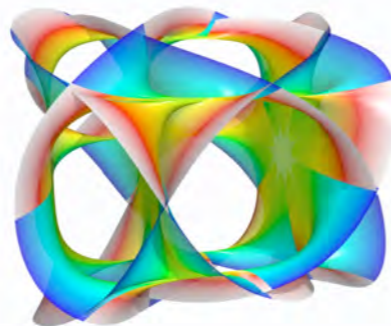
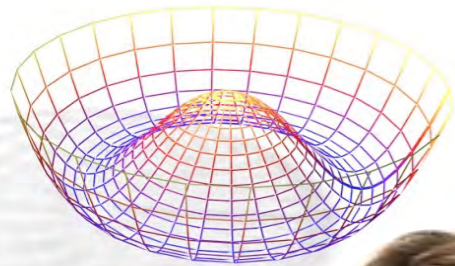
# THEORY

# SERVICE

$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i \partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} |(i \partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi|^2 - V(\phi)}_{\text{W}^\pm, \text{Z}, \gamma, \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 \bar{L} \phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

Q

A



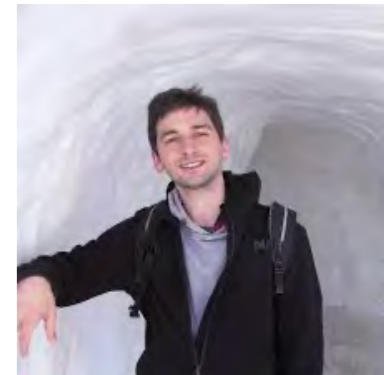
# BUILD IT AND THEY WILL COME

In 2010 we identified a use-case with high scientific value for community

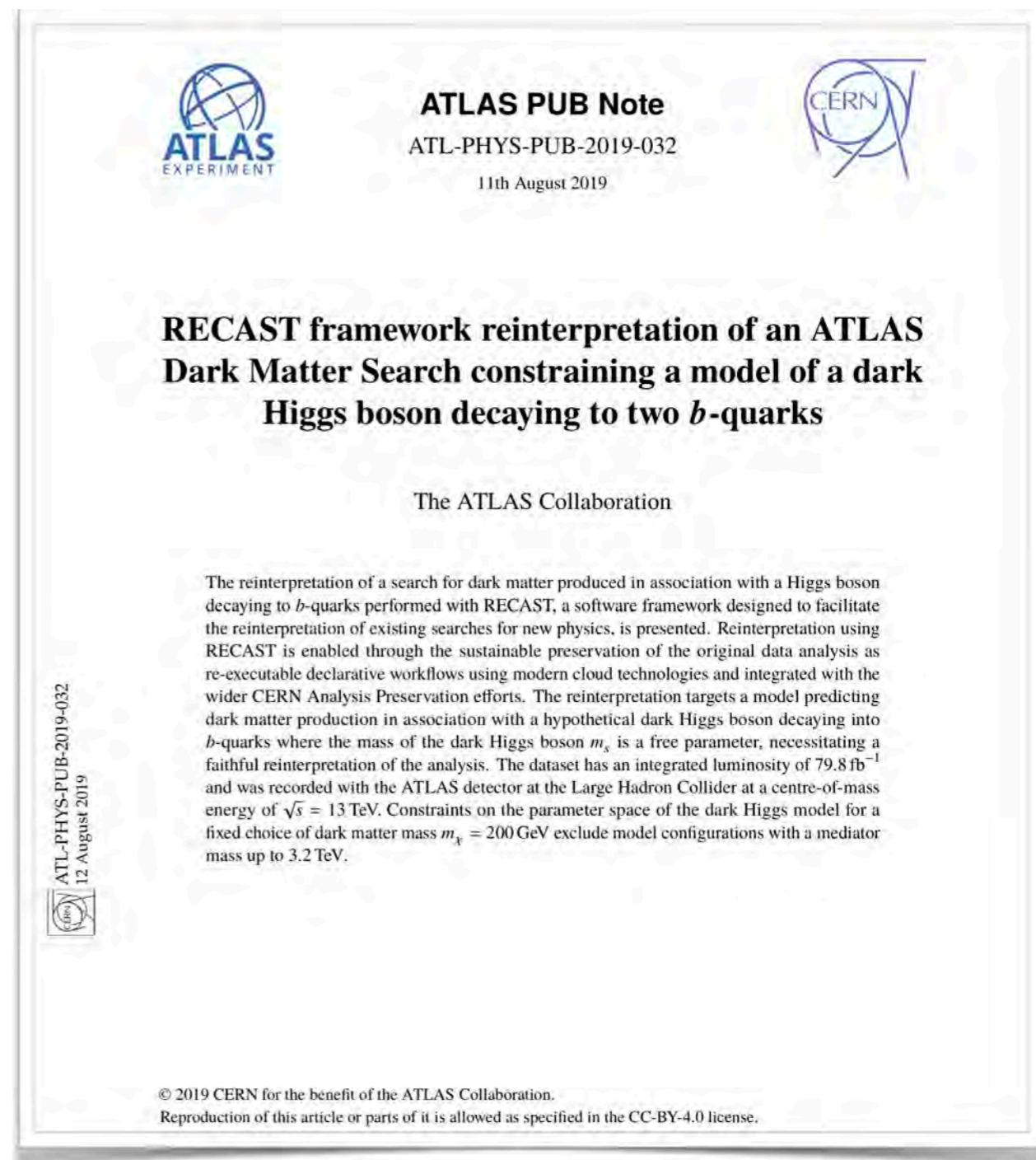
- Conservative narrative compared to “open data”
- Not conservative enough for many. Lots of resistance
- People said it couldn't be done, our workflows are too complicated
- Hard to get effort to work on it.

Got lucky with an amazing student that took a risk and just built it.

- Containers & Cloud technology
- 9 years later ...



Lukas Heinrich



# SHIFTING FROM REPRODUCIBILITY TO REUSE

## Open is not enough

Xiaoli Chen<sup>1,2</sup>, Sünje Dallmeier-Tiessen<sup>1\*</sup>, Robin Dasler<sup>1,11</sup>, Sebastian Feger<sup>1,3</sup>, Pamfilos Fokianos<sup>1</sup>, Jose Benito Gonzalez<sup>1</sup>, Harri Hirvonsalo<sup>1,4,12</sup>, Dinos Kousidis<sup>1</sup>, Artemis Lavasa<sup>1</sup>, Salvatore Mele<sup>1</sup>, Diego Rodriguez Rodriguez<sup>1</sup>, Tibor Šimko<sup>1\*</sup>, Tim Smith<sup>1</sup>, Ana Trisovic<sup>1,5\*</sup>, Anna Trzcinska<sup>1</sup>, Ioannis Tsanaktsidis<sup>1</sup>, Markus Zimmermann<sup>1</sup>, Kyle Cranmer<sup>6</sup>, Lukas Heinrich<sup>6</sup>, Gordon Watts<sup>7</sup>, Michael Hildreth<sup>8</sup>, Lara Lloret Iglesias<sup>9</sup>, Kati Lassila-Perini<sup>4</sup> and Sebastian Neubert<sup>10</sup>

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.



Reproducible research data analysis platform

Flexible

Run many computational workflow engines.



Scalable

Support for remote compute clouds.



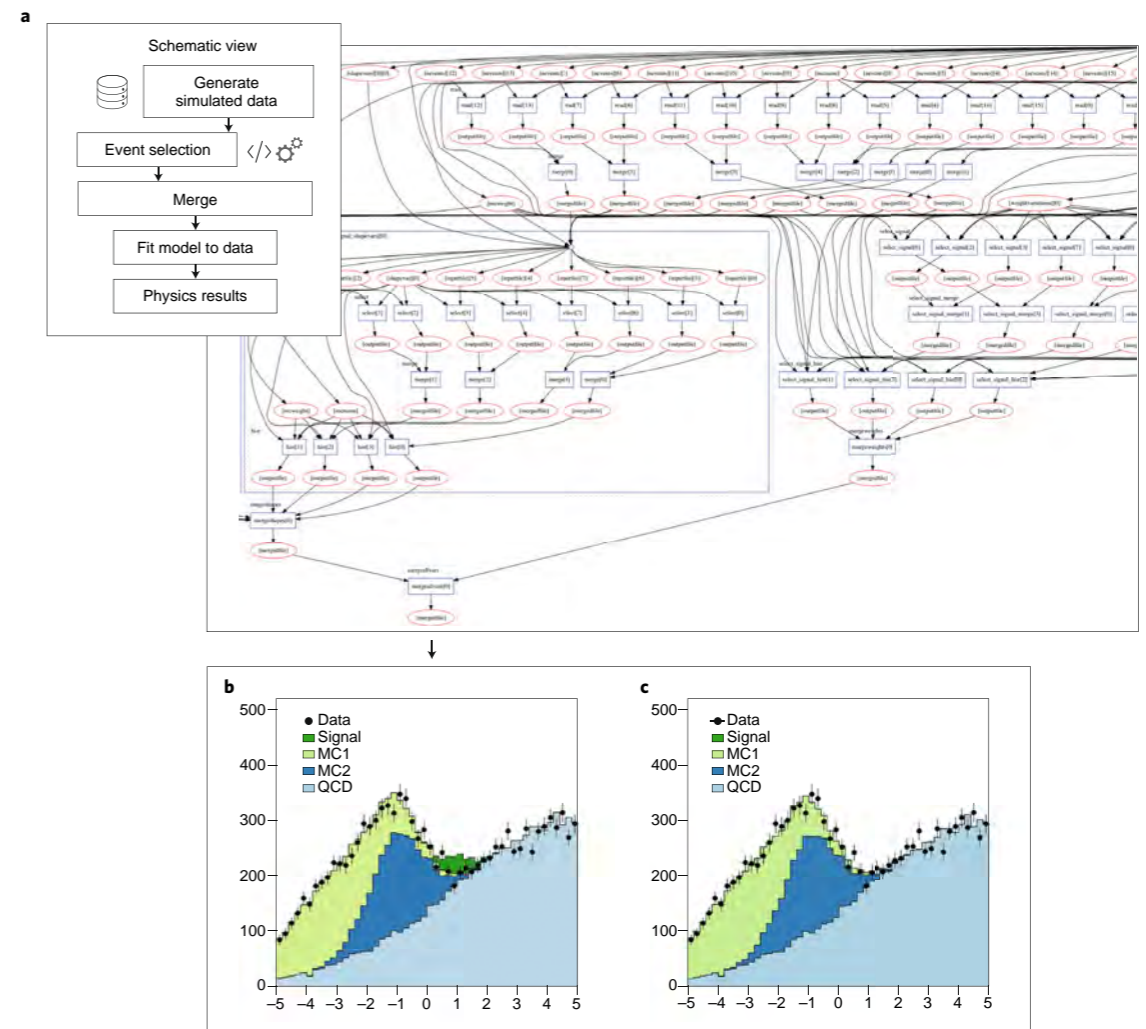
Reusable

Containerise once, reuse elsewhere. Cloud-native.



Free

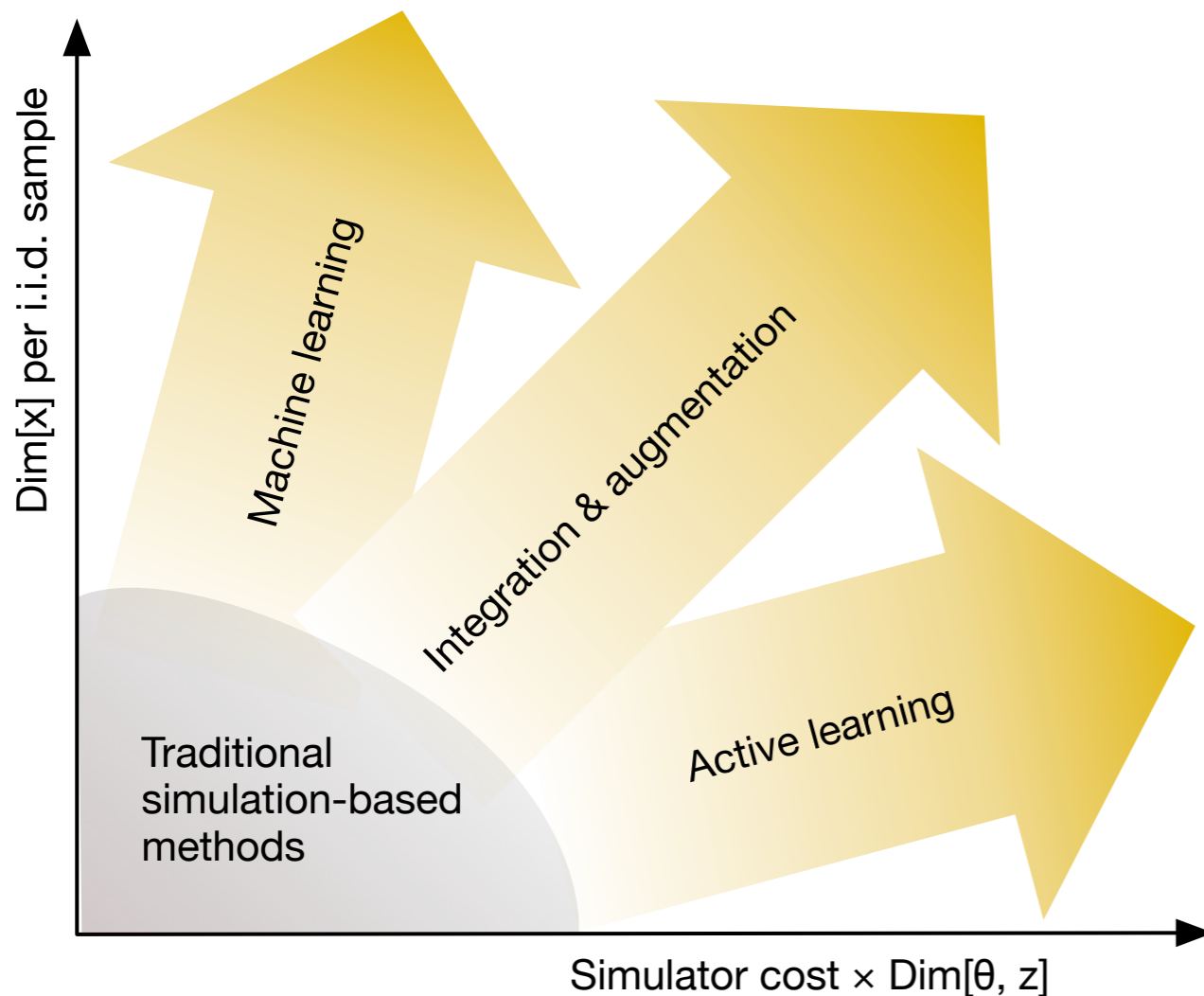
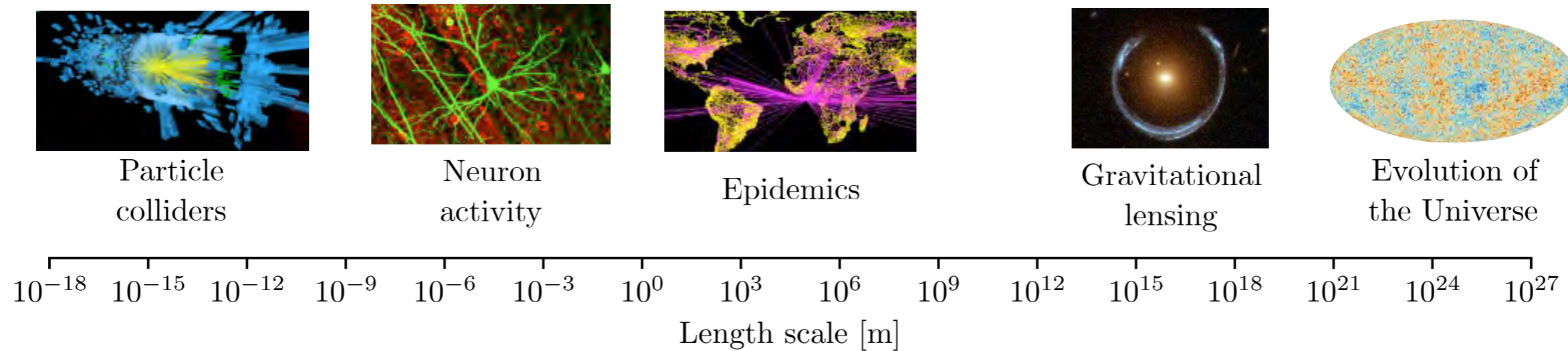
Free Software. MIT licence. Made with ❤️ at CERN.



**2 | Example of a complex computational workflow on REANA mimicking a beyond the standard model (BSM) analysis.** This figure shows an example where the experimental data is compared to the predictions of the standard model with an additional hypothesized signal component. The example permits one to study the complex computational workflows used in typical particle physics analyses. **a–c.** The computational workflow (**a**) may consist of several tens of thousands of computational steps that are massively parallelizable and run in a cascading ‘map-reduce’ style of computations distributed across compute clusters. The workflow definition is modelled using the Yadage workflow specification and produces an upper limit on the signal strength of the BSM process. A typical search for BSM physics consists of simulating a hypothetical signal process (**c**), as well as the background processes predicted by the standard model with properties consistent with the hypothetical signal (marked dark green in **b**)). The background often consists of simulated background estimates (dark blue and light green histograms) and data-driven background estimates (light blue histogram). A statistical model involving both signal (dark green histogram) and background components is built and fit to the observed experimental data (black dots). **b.** Results of the model in its pre-fit configuration at nominal signal strength. We can see the excess of the signal over data, meaning that the initial setting does not describe the data well. The post-fit distribution would scale down the signal in order to fit the data. This REANA example is publicly available at ref. <sup>35</sup>. For icon credits, see Fig. 1.

Looking forward

# SIMULATION-BASED INFERENCE



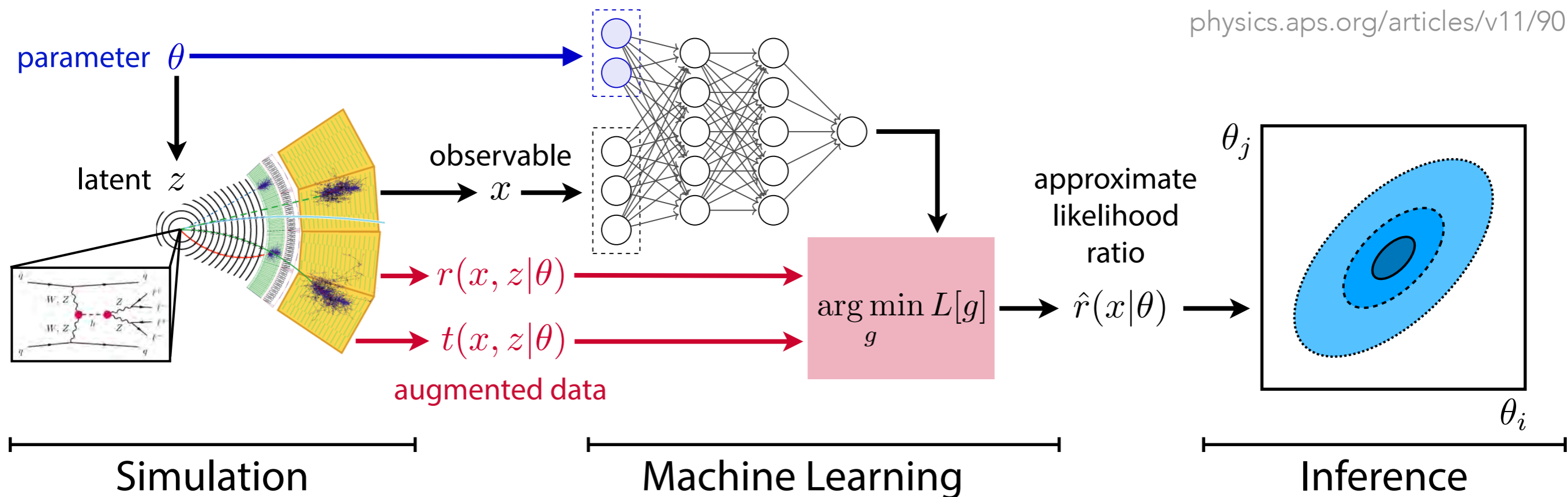
Many areas of science have simulations based on some well-motivated mechanistic model.

However, the aggregate effect of many interactions between these low-level components leads to an intractable inverse problem.

The developments in machine learning and AI have the potential to effectively bridge the microscopic - macroscopic divide & aid in the inverse problem.

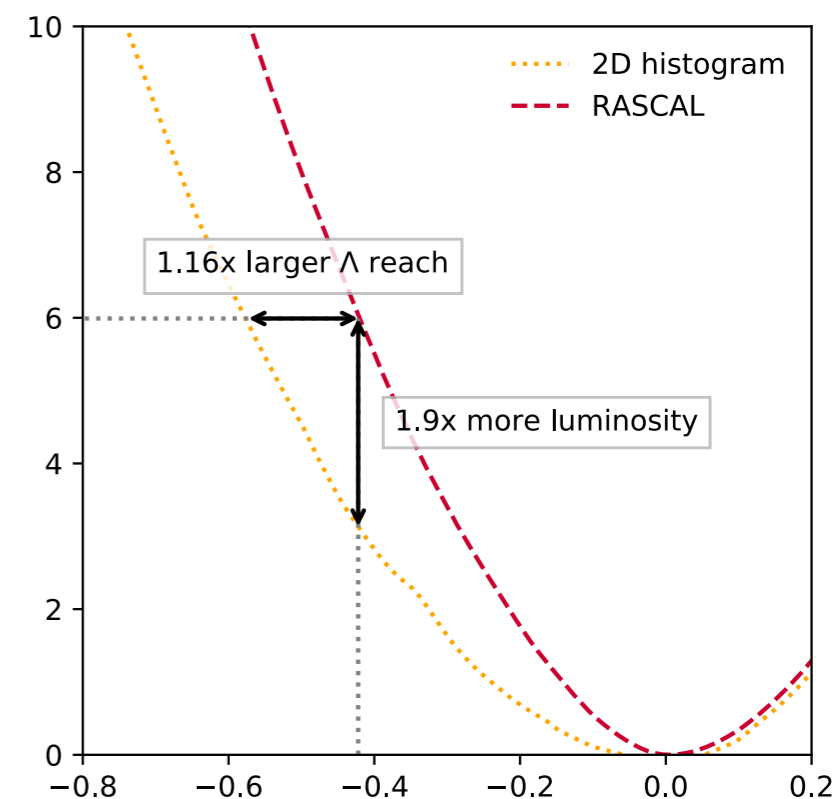
- they can provide effective statistical models that describe emergent macroscopic phenomena that are tied back to the low-level microscopic (reductionist) model

# SIMULATION-BASED INFERENCE



**Accelerating science...**

Similar sensitivity achieved with half the data



# SYNTHESIS

active learning / sequential design / black box optimization



Active Sciencing



reusable workflows

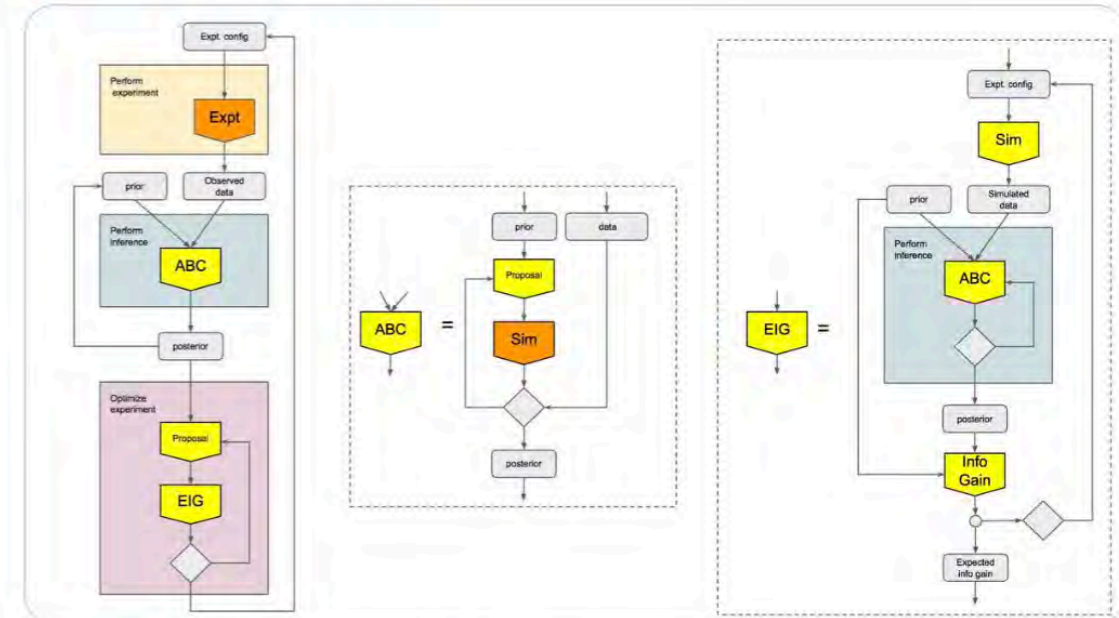


simulation-based /  
likelihood-free  
inference engines



**Kyle Cranmer** @KyleCranmer · Jun 11, 2017

Demo for YComb research  
active learning + workflows + implicit models = #ActiveSciencing  
[@lukasheinrich\\_](#) [@gloupe](#)  
[github.com/cranmer/active...](https://github.com/cranmer/active_science)



4 22 61



**Danilo J. Rezende** @DeepSpiker · Jul 19, 2017

This is great!

1 3



**Kyle Cranmer** @KyleCranmer · Jul 19, 2017

Thanks!!!

1 2



**Danilo J. Rezende**  
@DeepSpiker

Replying to [@KyleCranmer](#) [@lukasheinrich\\_](#) and [@gloupe](#)

You have the full loop of the scientific method in a python notebook :)

3:12 PM · Jul 19, 2017 · [Twitter for iPhone](#)

## QUESTION 1:

Where do you see opportunities in your domain to accelerate scientific discovery through advanced and automated workflows? Could you give an example for advanced and automated workflows in your area?

- Reuse of existing workflows ("pipelines")
  - Often workflow developed for one purpose can be used to address a different use-case (eg. reinterpret the results of a search for some alternate theory)
  - Streamline on-boarding by "forking" and "remixing" existing workflows
- Development of large-scale, multi-purpose experimental facilities. Think of a meta-version of experimental design where same apparatus or facility (eg. LHC experiments) is used for multiple scientific goals
  - Difficult to iterate on a coherent Technical Design Report when baselines are manually updated. Automation allows for quicker iterations with consistent versioning
- Facilitate Simulation-based Inference & Active Learning
  - Requires running computer simulation and down-stream analysis workflows many times

## QUESTION 2:

With regard to using automated workflows and data science tools, where do you wish your field to be going in the next 10 years?

- Differentiable Programming (autodiff) & Probabilistic Programming
- Active learning / Reinforcement Learning / Control
- Simulation-based inference
- Containerization / Cloud-native / Functions as a service
- Direct Integration into literature system

What are major obstacles?

- Nay-sayers in positions of power. Tendency to be too general.

What does it take to get there?

- Persistence.
- Limit scope for specific use case at first. **Build it and they will come!**

## QUESTION 3

What changes in funding, organizational, and incentive structures would you like to see to realize the opportunities in your domain?

- Need support for R&D aligned to a specific use-case
- Ideally embedded in programs that provide base support for “doing the science”
- ~1 or 2 technical FTE for 1-2 years has huge impact



Alfred P. Sloan  
FOUNDATION



DASPOS



The SCALFIN Project  
[scaifn.github.io](https://scaifn.github.io)

