# HEP in the Cloud Computing and Open Science Era

Lukas Heinrich (CERN), DESY Colloquium

# Big Picture Goals

**Our job: extract as much information from experimental data**

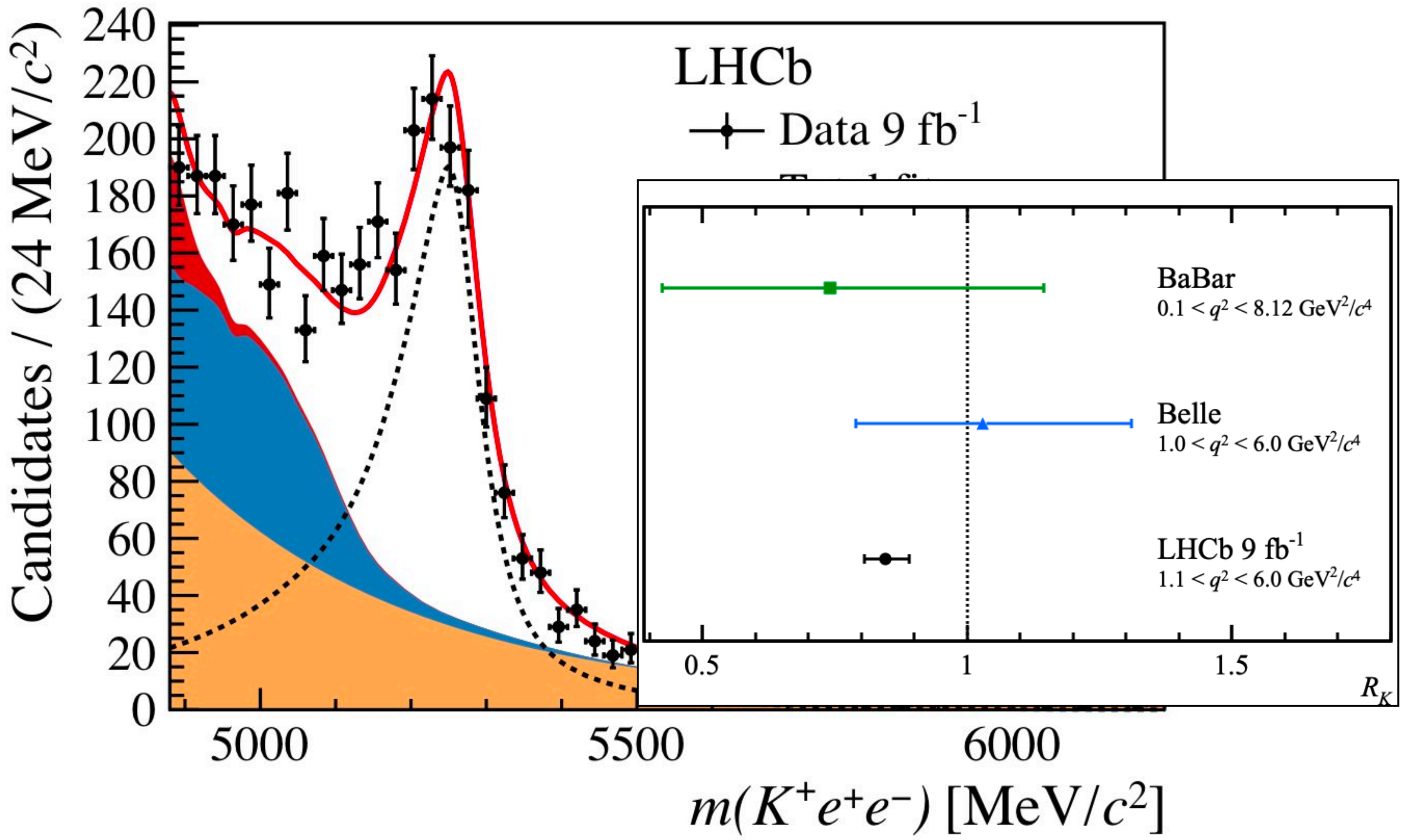
$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})}{p(\text{data})} p(\text{theory})$$

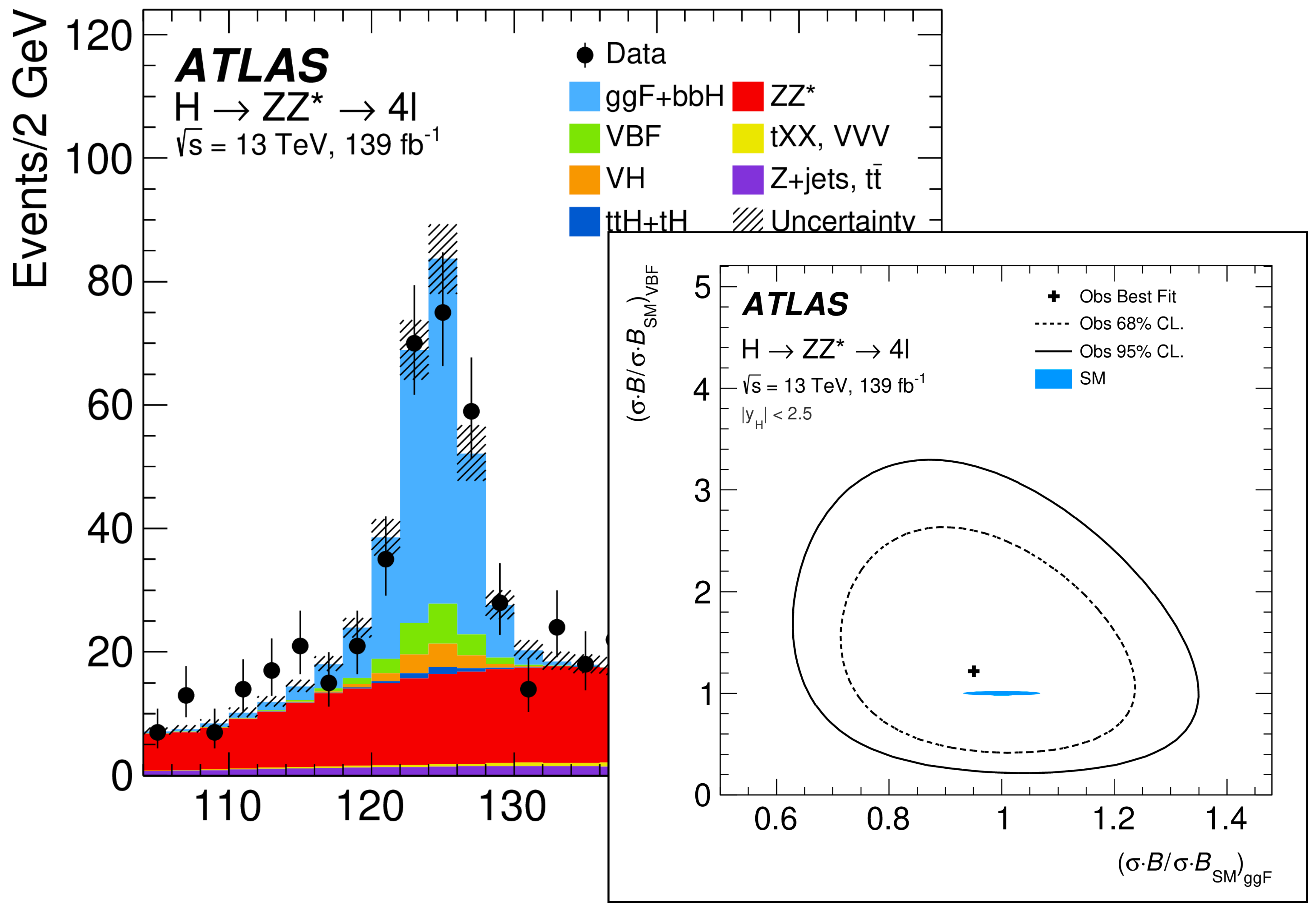
↑ what we want

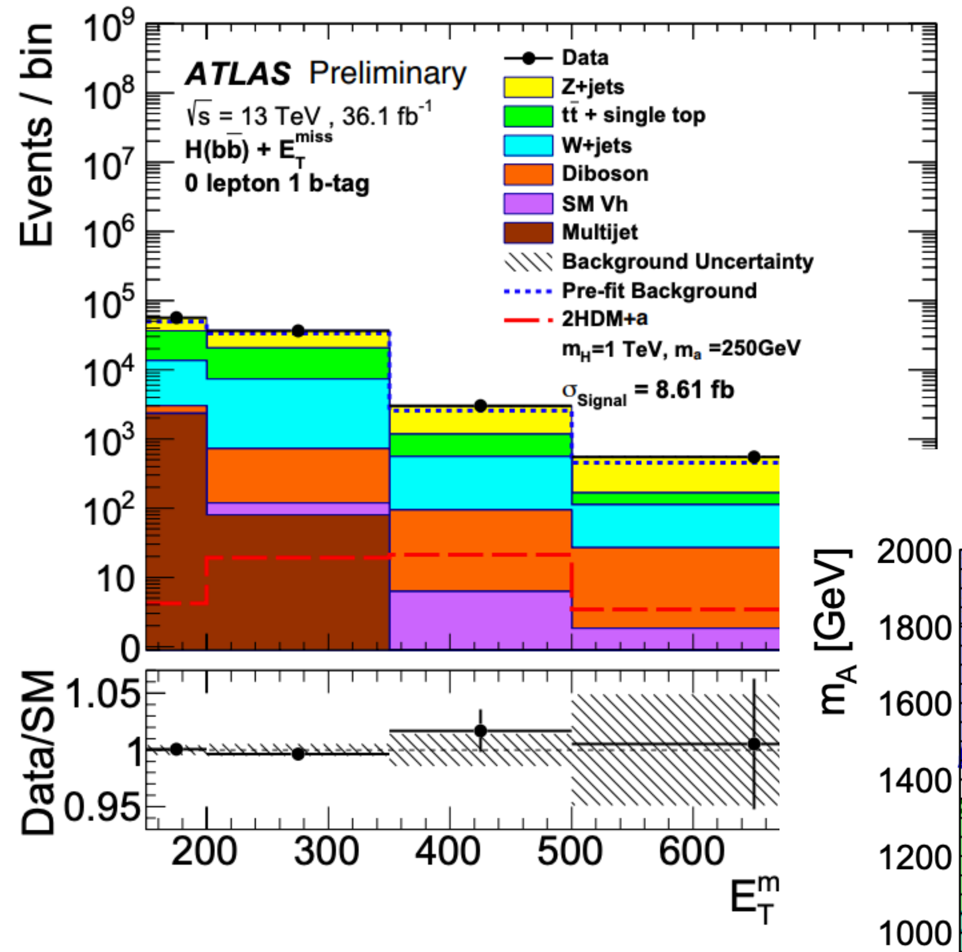
↑ hep-ex

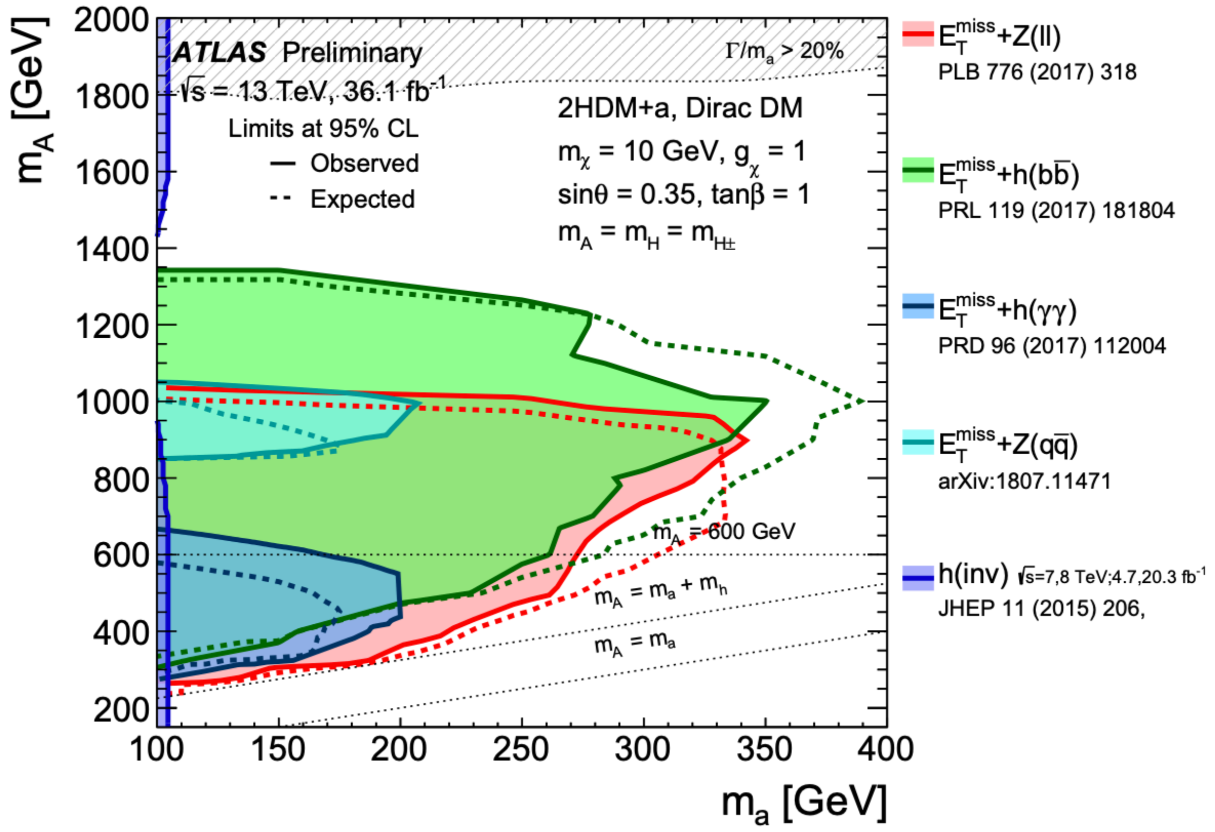
↑ hep-ph/hep-th

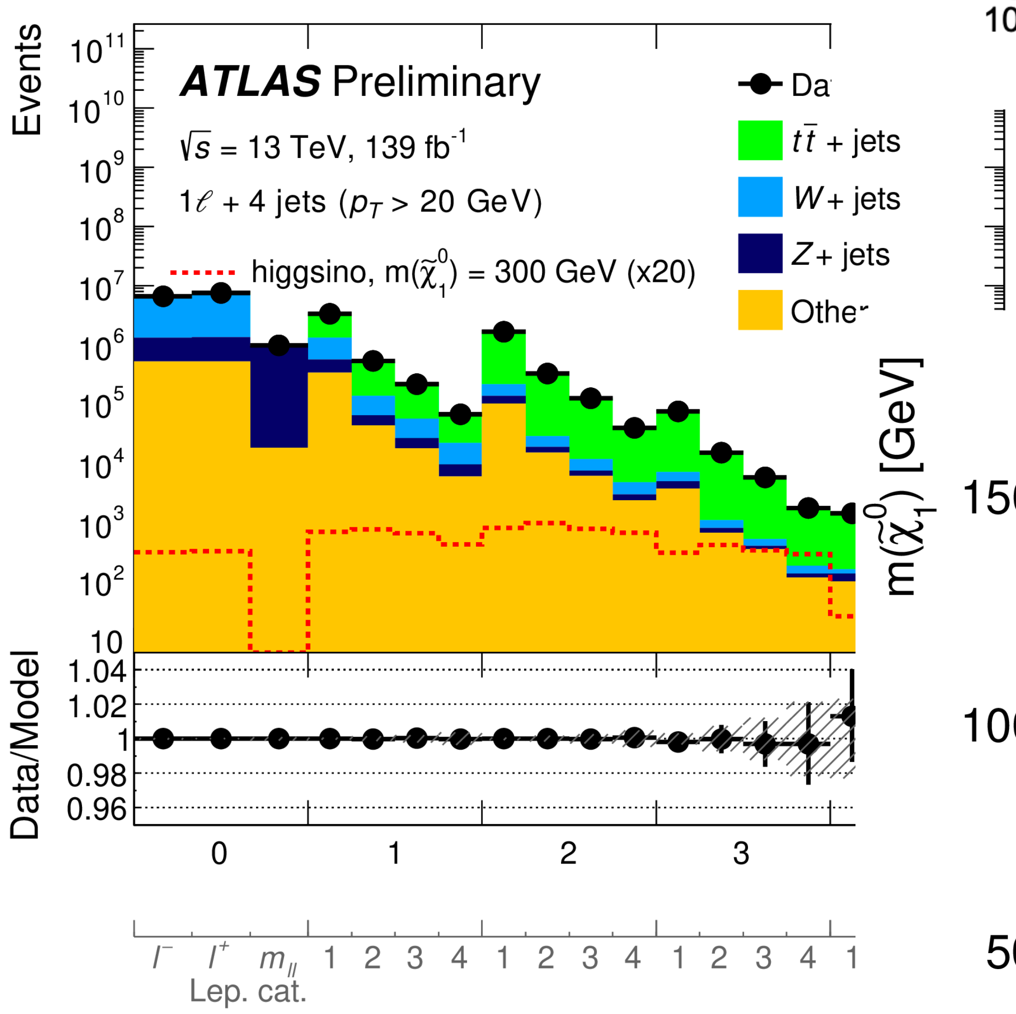
# Big Picture Goals

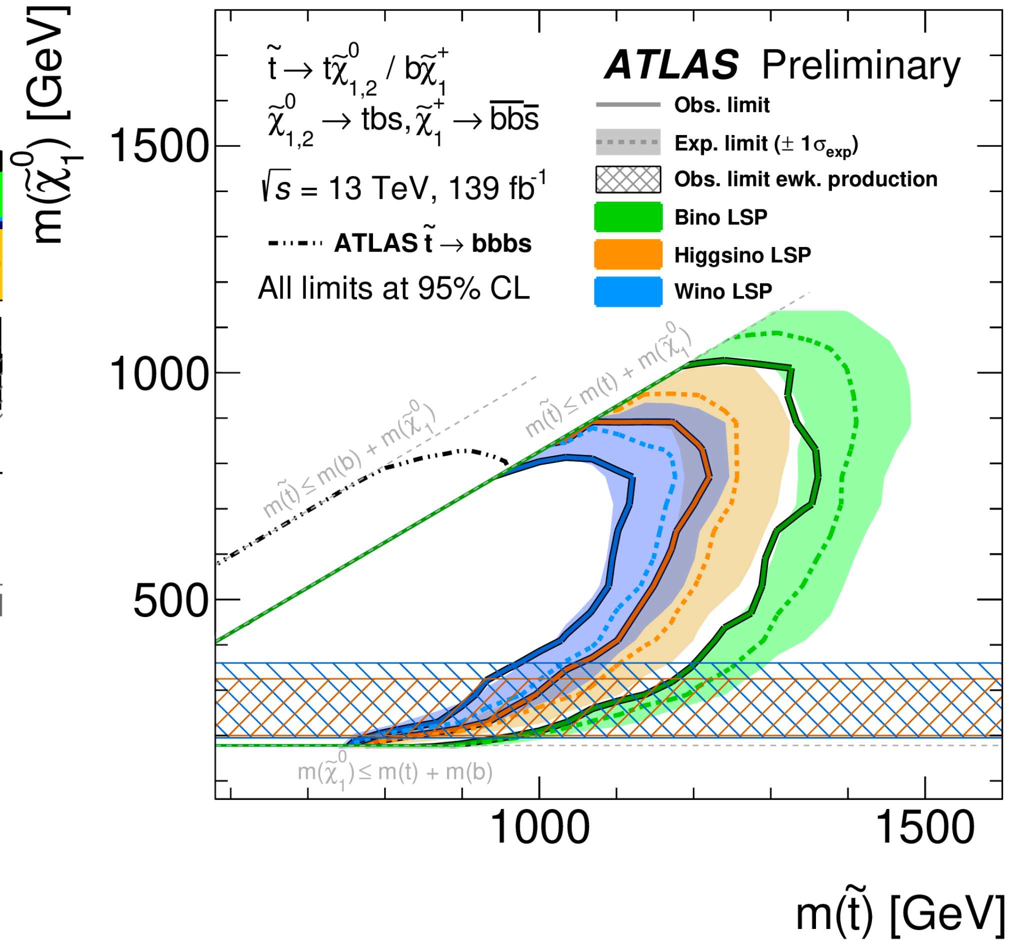


# Big Picture Goals

**Our job: extract as much information from experimental data**

$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})}{p(\text{data})} p(\text{theory})$$

what we want — hep-ex — hep-ph/hep-th

- **What's the best way to do this in our Big Science setting?**
- **How can we collaborate best across theory-experiment divide?**

## Three Challenges of Big Science

1. The data is large:

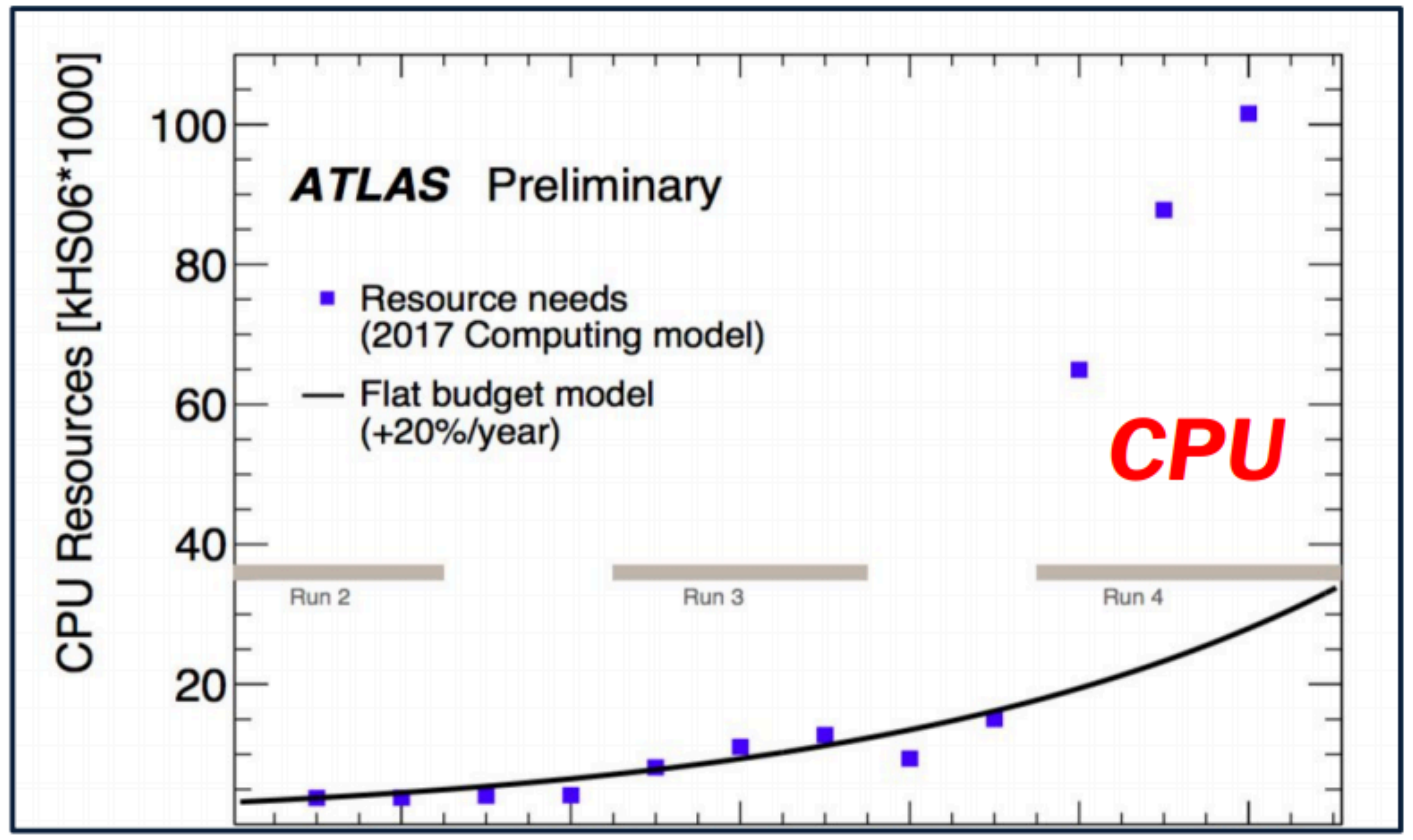
At a project level:

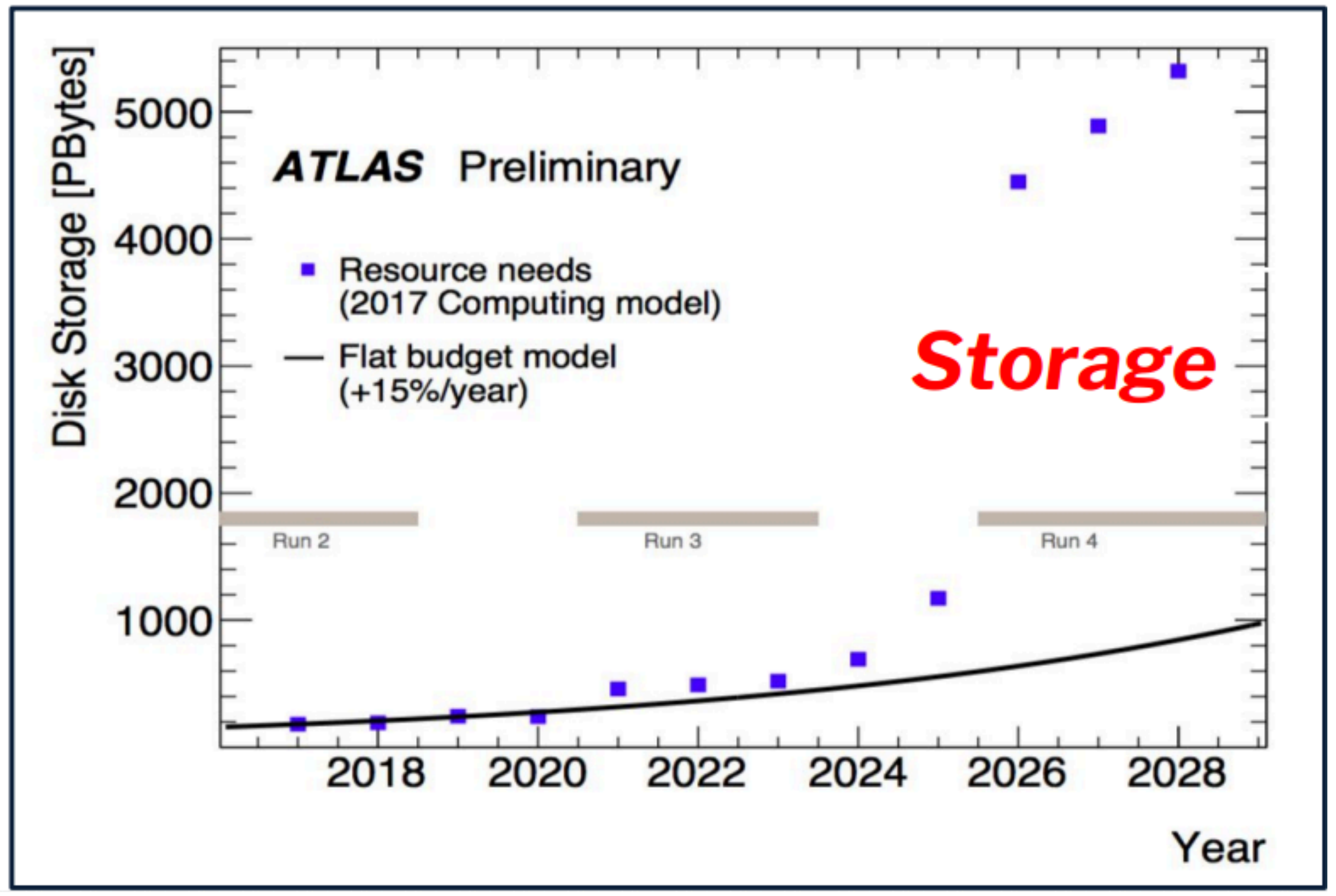
LHC: 1 EB today, O(30) EB HL-LHC (?)
LSST: O(1) EB
SKA: O(0.5) EB/year

At an analysis level:

routinely O(100) TB that analysis teams process as a small group - will be O(PB)

- how do we maintain explorative research?




[LHC Projections]

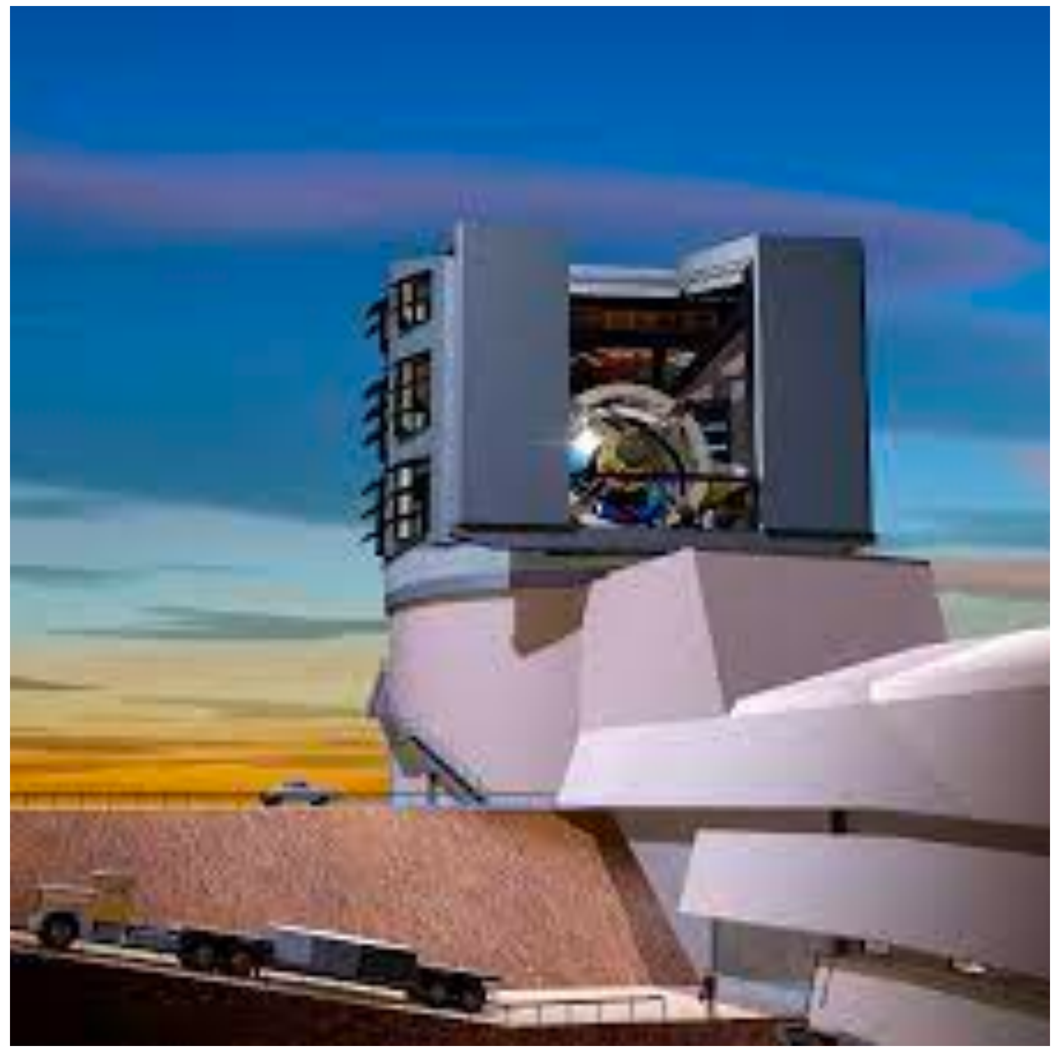
# Three Challenges of Big Science

## 2. The data is unique:

**There is only one LHC / LSST / SKA / ...**

- **What data products are released to the public?**
- **How do we ensure rigorous analysis if public?**

Square Kilometer Array



Robin Observatory / LSST

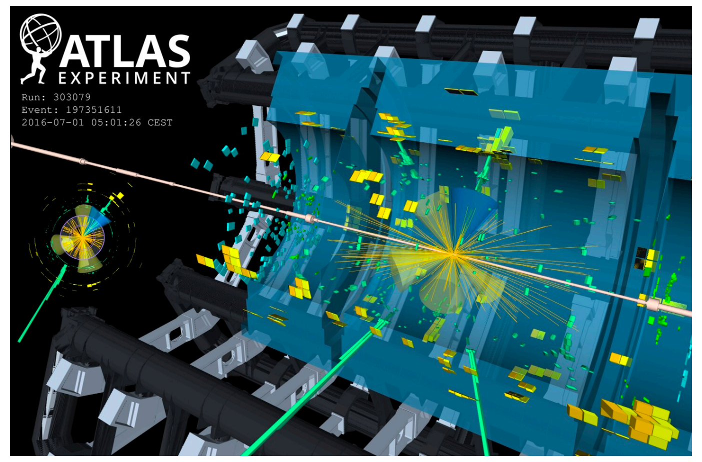
LHC

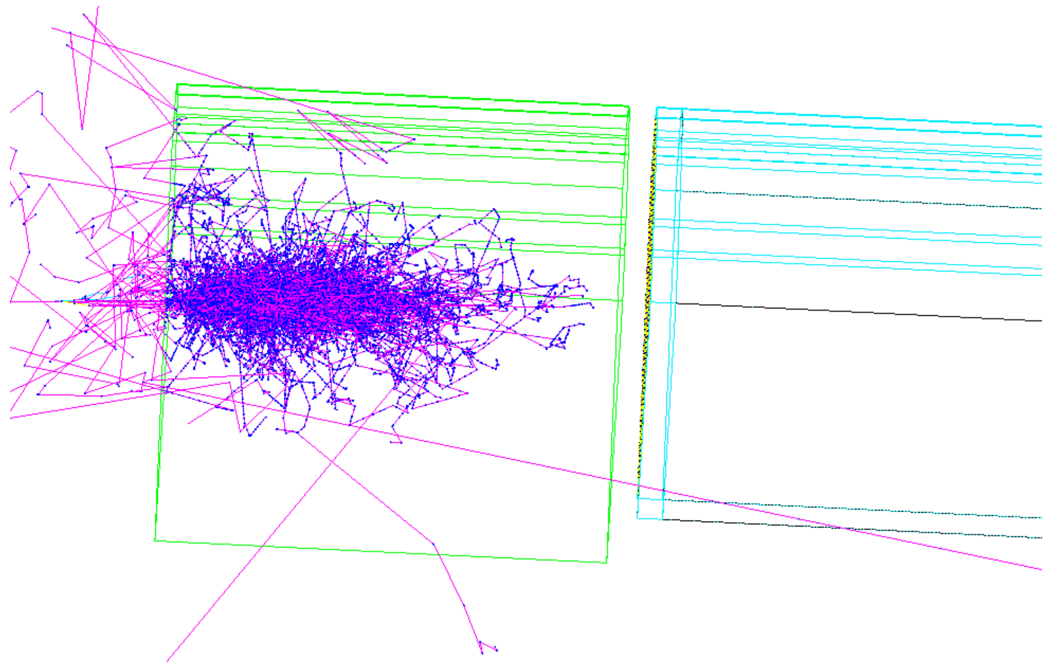
# Three Challenges of Big Science

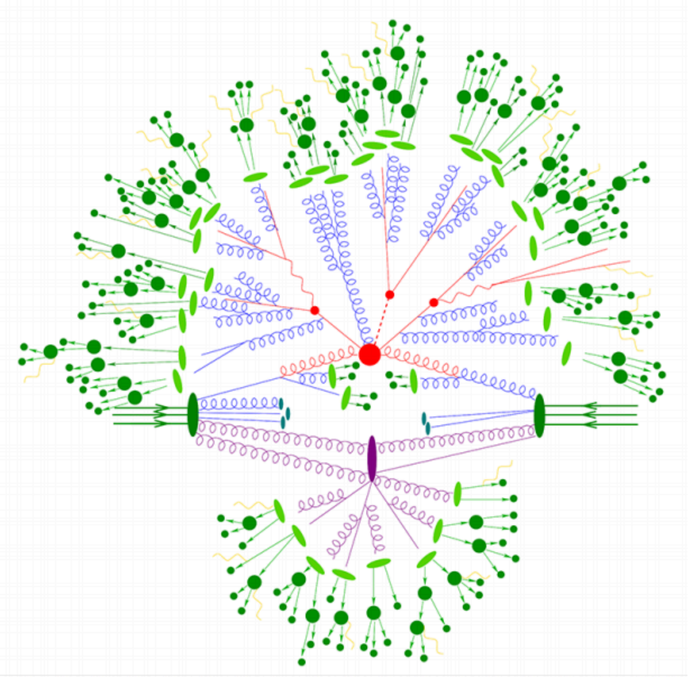
## 3. The data is complex:

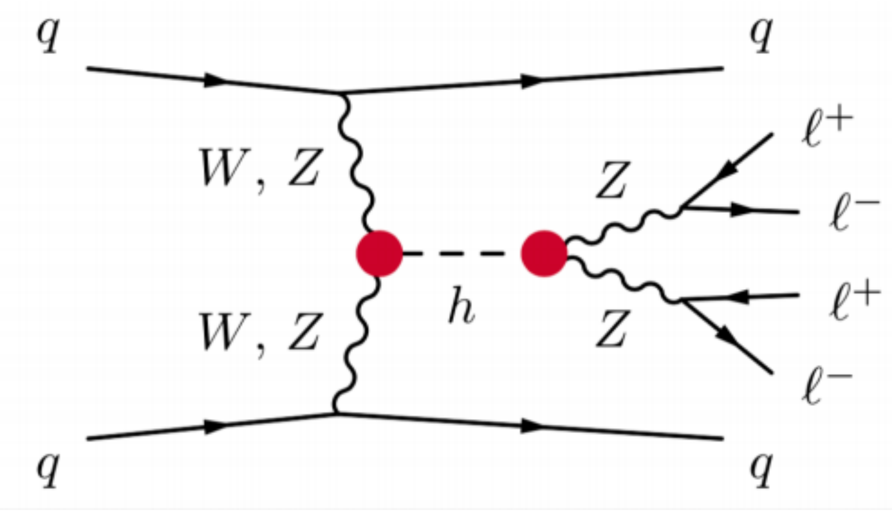
## In HEP:

- complex generative history across many energy scales
- heterogeneous detectors (not a single 100Mpx sensor / image)







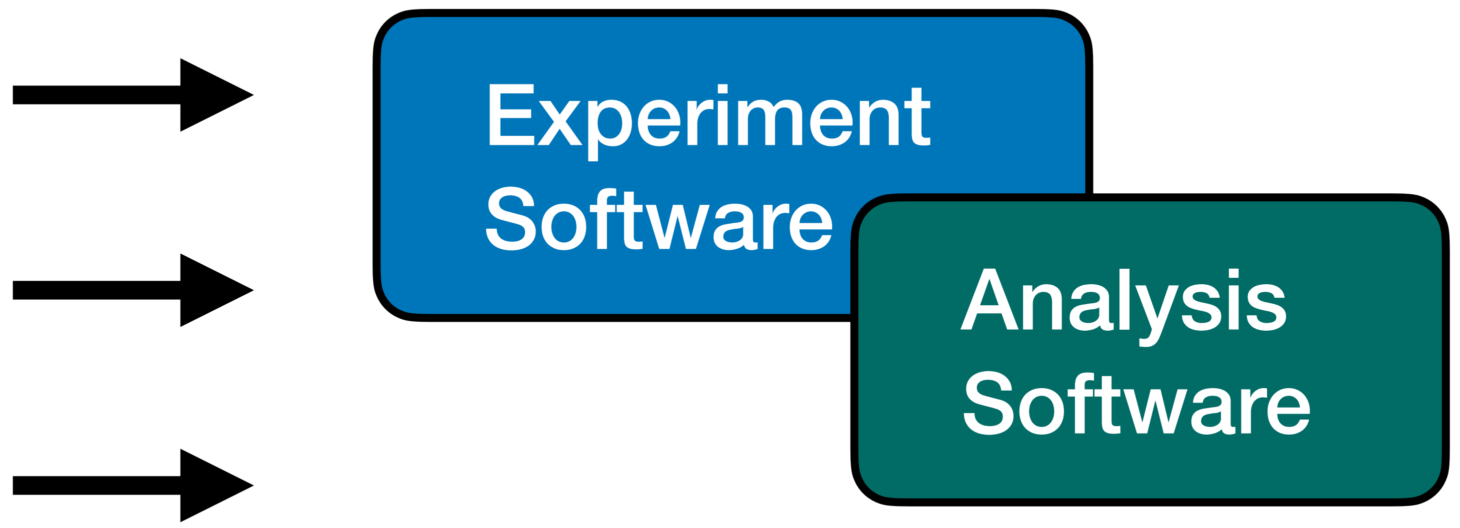

$$p(x|\theta, \theta_{\rm nuis}) = \iiint dz_d dz_h dz_p \, p(x|z_d, \theta_{\rm sensor}) \; p(z_d|z_h, \theta_{\rm geom}) \; p(z_h|z_p, \theta_{\rm pQCD}) \; p(z_p|\theta)$$

high-dimensional integrals: needs MC

# Three Challenges of Big Science

## 3. The data is complex:

- analysis heavily simulation-driven
- heavy reconstruction from raw → physics data
- exceeds data volume & resources of actual data
- extremely software-reliant (O(M) LoC)




**Analyzing any part of the data is very expensive human resources-wise:**

- can we prioritize human time over computing? → interactive analysis
- if we invest so much? → how do we exploit these analyses maximally?

# Trends

**Technical:**

1. **Machine Learning**
   - powerful & creeps into every aspect of scientific workflow
   - raises questions of how analysis details are disseminated (th ↔ ex)

2. **Cloud Computing & Heterogeneous ~~Future~~ Present**
   - anybody can get access to vast amounts of compute
   - new requirements on hardware / software (GPU, CPU, TPU, Dataflow,...)

**can we use these to create new ways to push our science forward?**

# Trends

**Political:**

1. Open & FAIR Data
   - publicly funded research must be accessible

2. Reproducibility
   - Uniqueness of data & analysis → special responsibility

**But how does this work at Big Science scale?**

- usually not focus of those discussions.
- are there useful ways to interpret these buzzwords?

# (Open) Data Analysis at LHC Scale

# Open Data

## Significant Development in HEP

All LHC Experiments agree to release data publicly.

including software to analyze rigorously
- (e.g. systematic uncertainties)

Intent:
- foster more interdisciplinary collaboration (e.g. ML, Theory, cross-experiment R&D, … )

## CERN announces new open data policy in support of open science

**A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative**

11 DECEMBER, 2020

Data storage solutions at the CERN data centre (Image: CERN)

Geneva, 11 December 2020. The four main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have unanimously endorsed a new open data policy for scientific experiments at the Large Hadron Collider (LHC), which was presented to the CERN Council today. The policy commits to publicly releasing so-called level 3 scientific data, the type required to make scientific studies, collected by the LHC experiments. Data will start to be released approximately five years after collection, and the aim is for the full dataset to be publicly available by the close

# Open Data: a growing Ecosystem

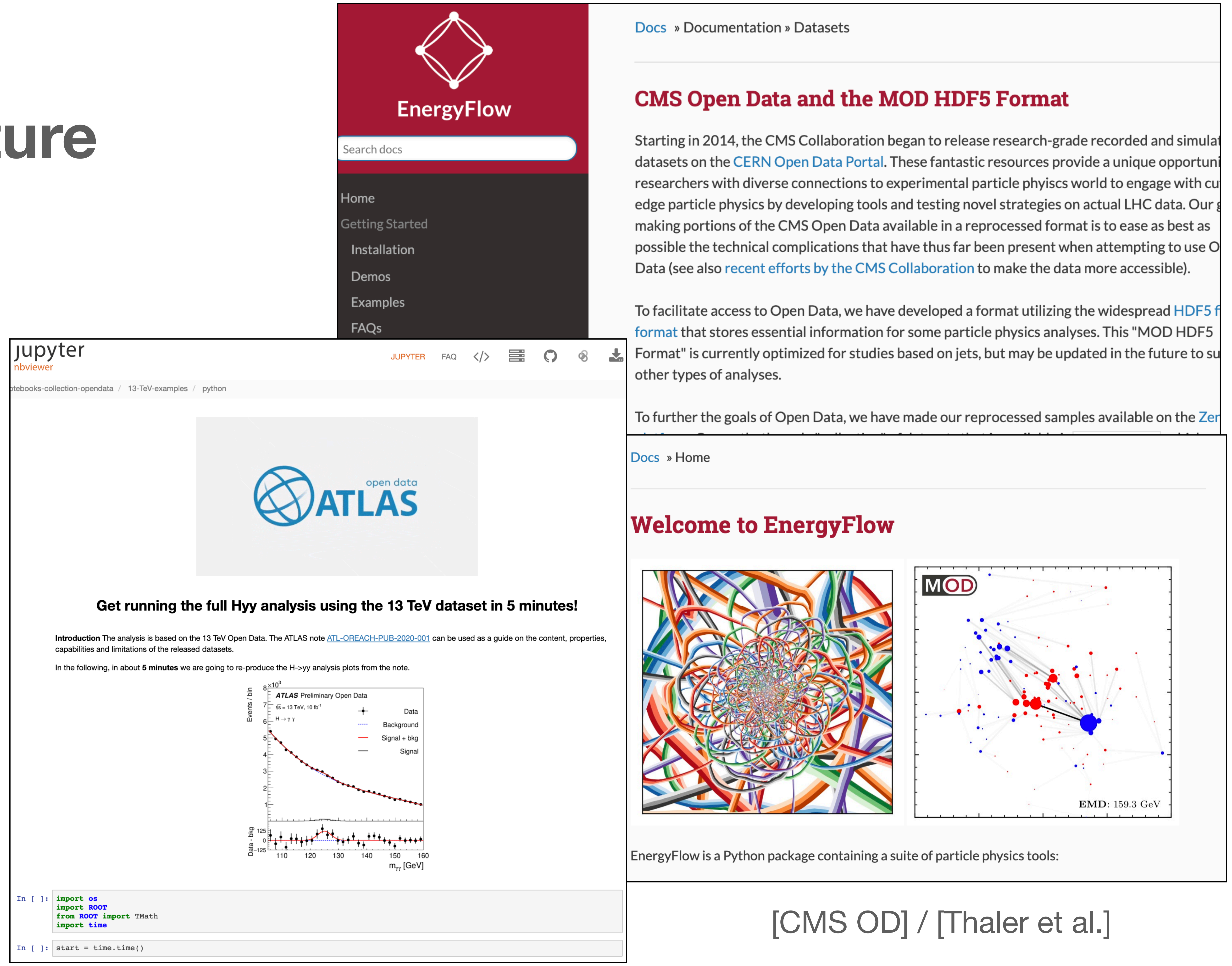
## New paradigm for HEP, we're learning.

## CERN:
- develop supporting cyberinfrastructure (Open Data Portal, Zenodo)

## Experiments:
- how to release data to outsiders

## Theorists:
- develop data analysis expertise
- own ecosystem  of tools

## Can you scale to realistic analysis?



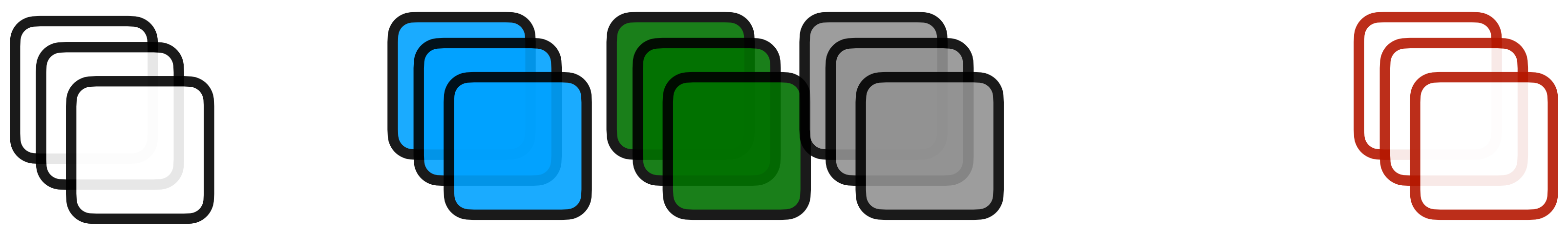

[ATLAS OD]

[CMS OD] / [Thaler et al.]

# Benchmark Example: CMS Open Data Higgs → 4l Analysis

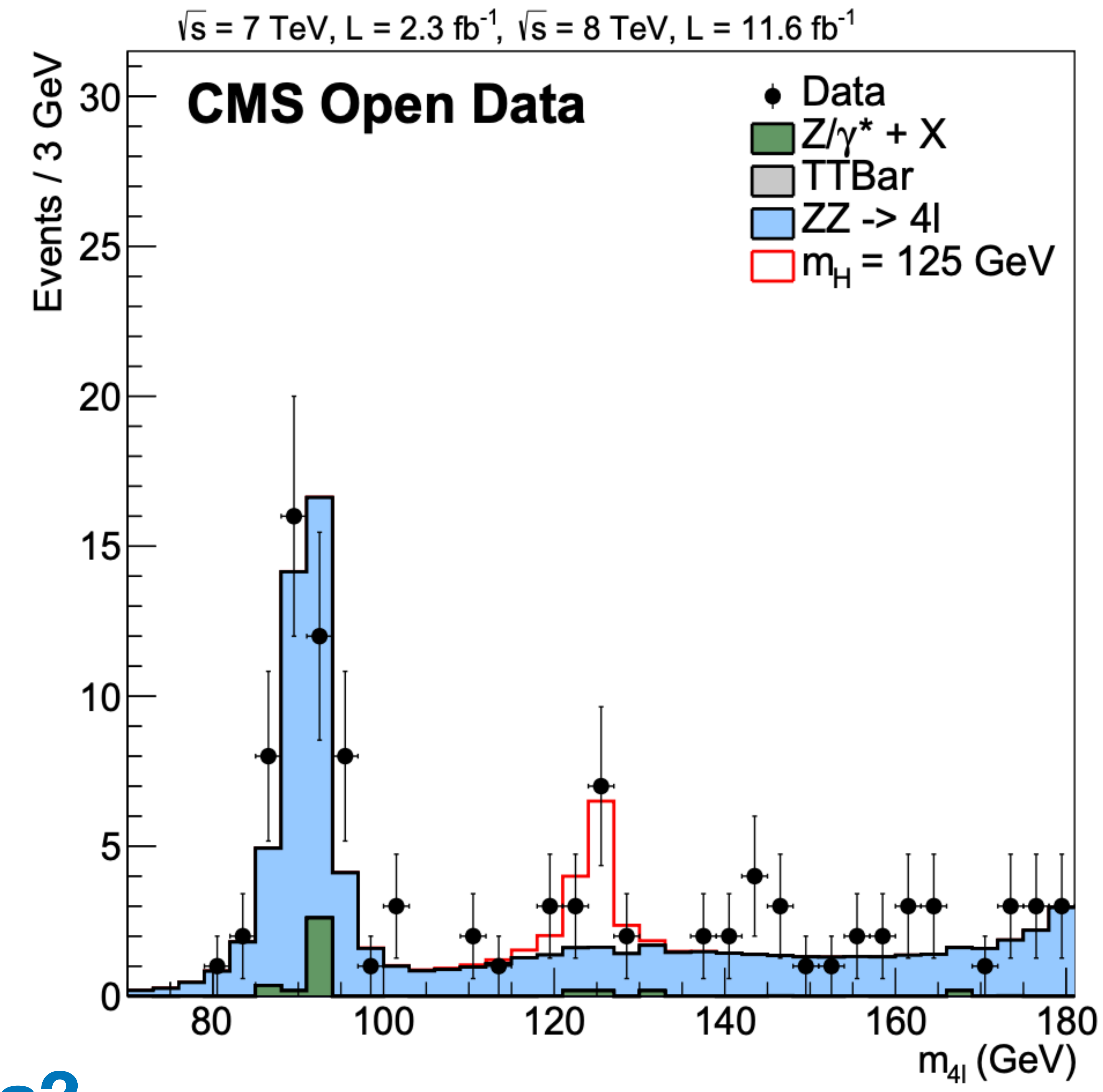
## Anatomy:

- 70TB / 25k files
- Data, Background + Signal Simulation



- C++ based data analysis (CMSSW) (event selection + feature comp)

## The physics is there: but requires large-scale compute




[A. Geiser, N. Jomhari]

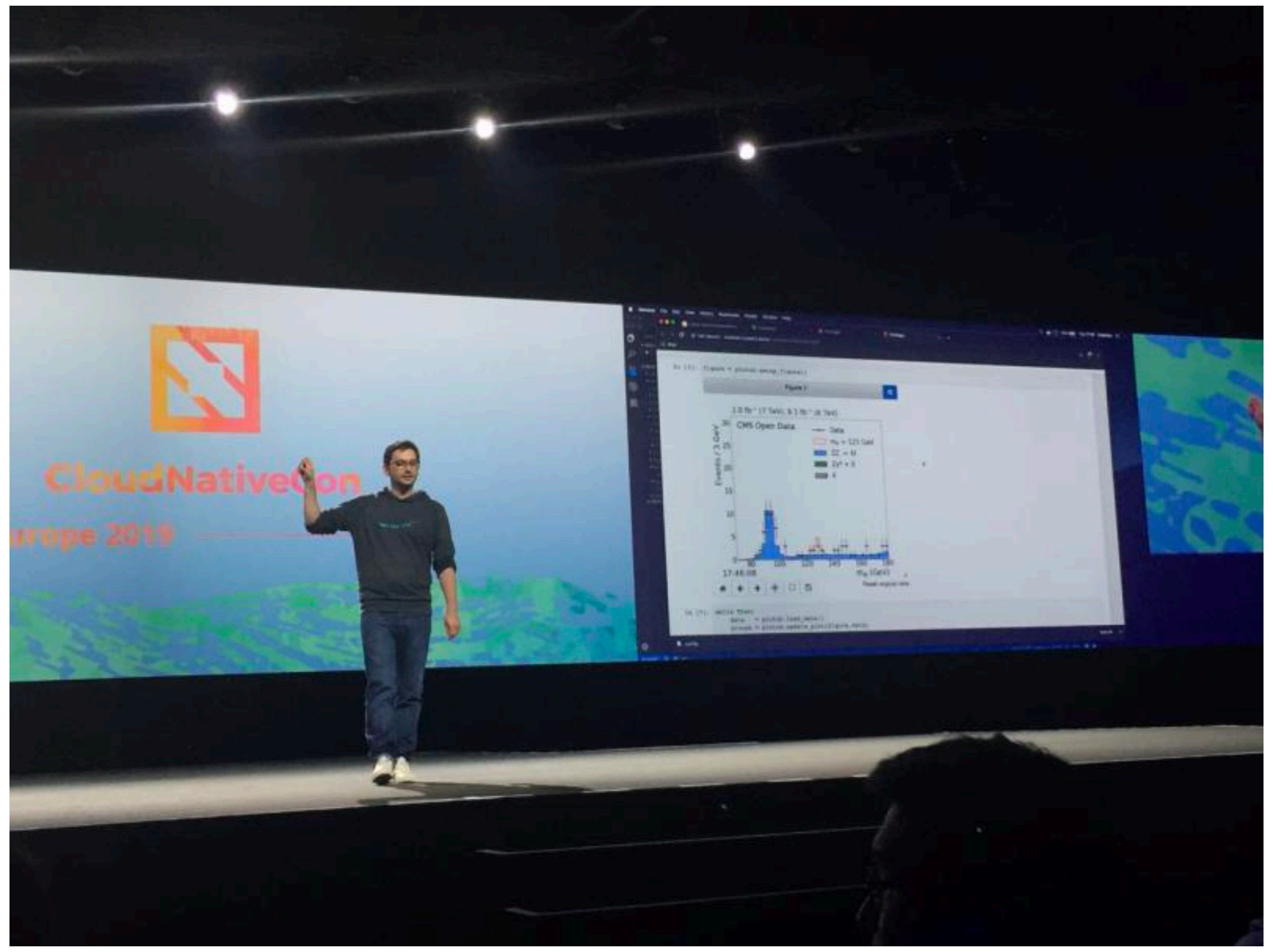
**Question: How quickly can we analyze the data?**

# Benchmark Example: CMS Open Data Higgs → 4l Analysis

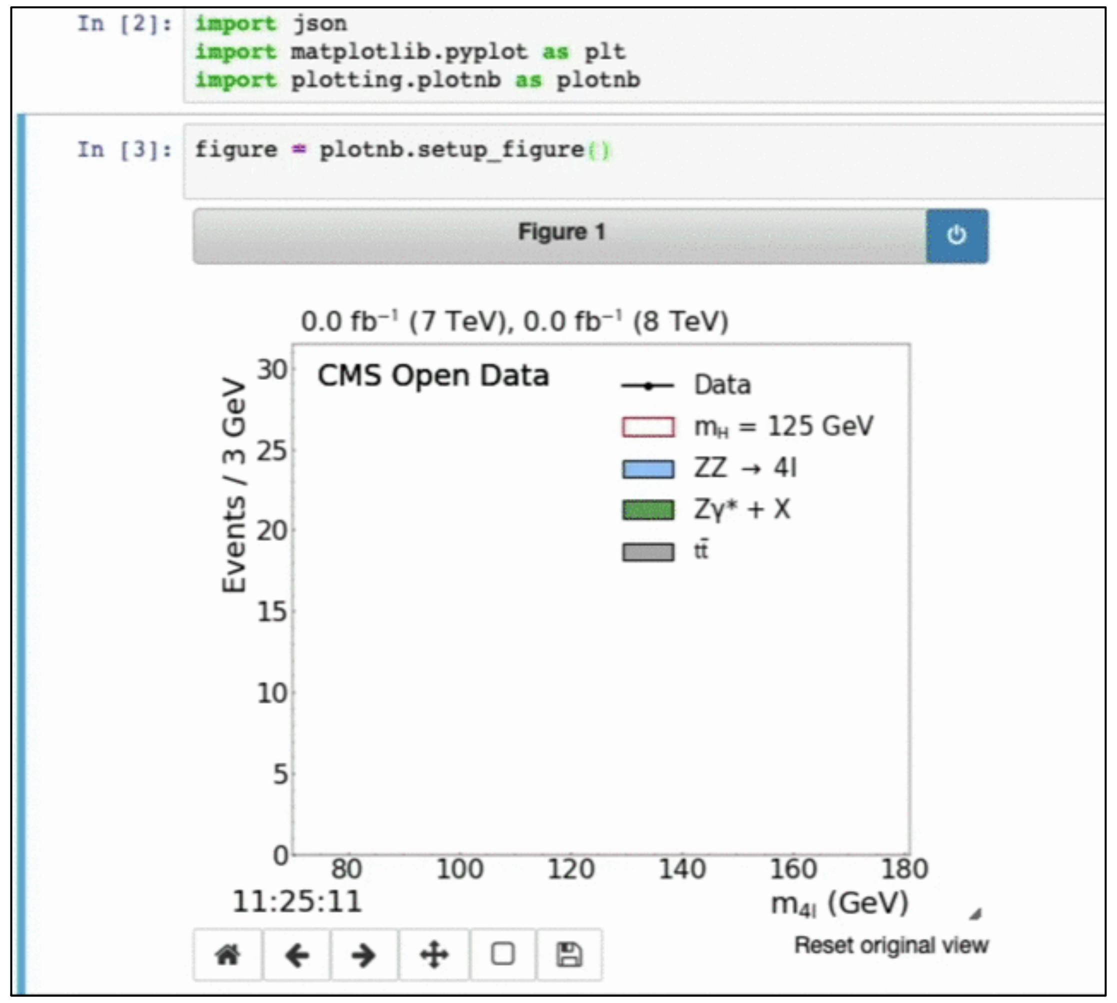
**Answer: 5 minutes with cloud computing techniques → >1Tbps throughput**
**Fast enough to do 70TB analysis in real-time.**

- **Open Data as a tool for R&D to develop interactive analysis systems @ HL-LHC**

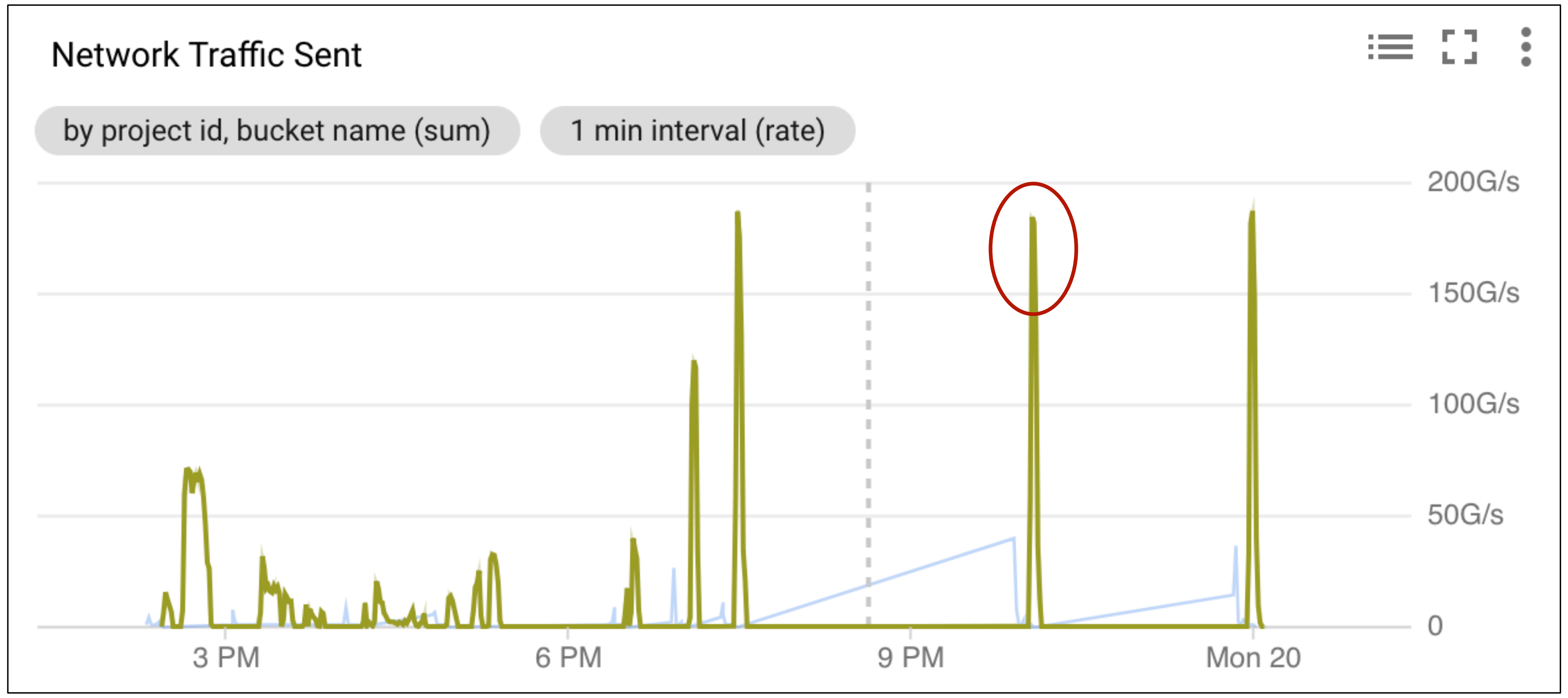
Google Cloud CERN openlab




[KubeCon 2019]



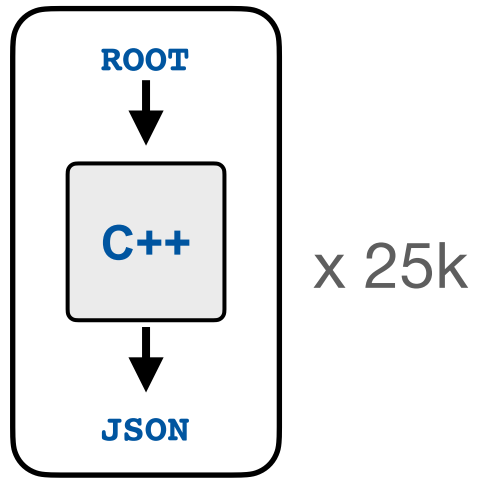

(slightly sped-up version)




[R. Rocha (CERN IT), LH (ATLAS), C. Lange (CMS) ]

# Setup

25k files → 25k cores across 10 clusters



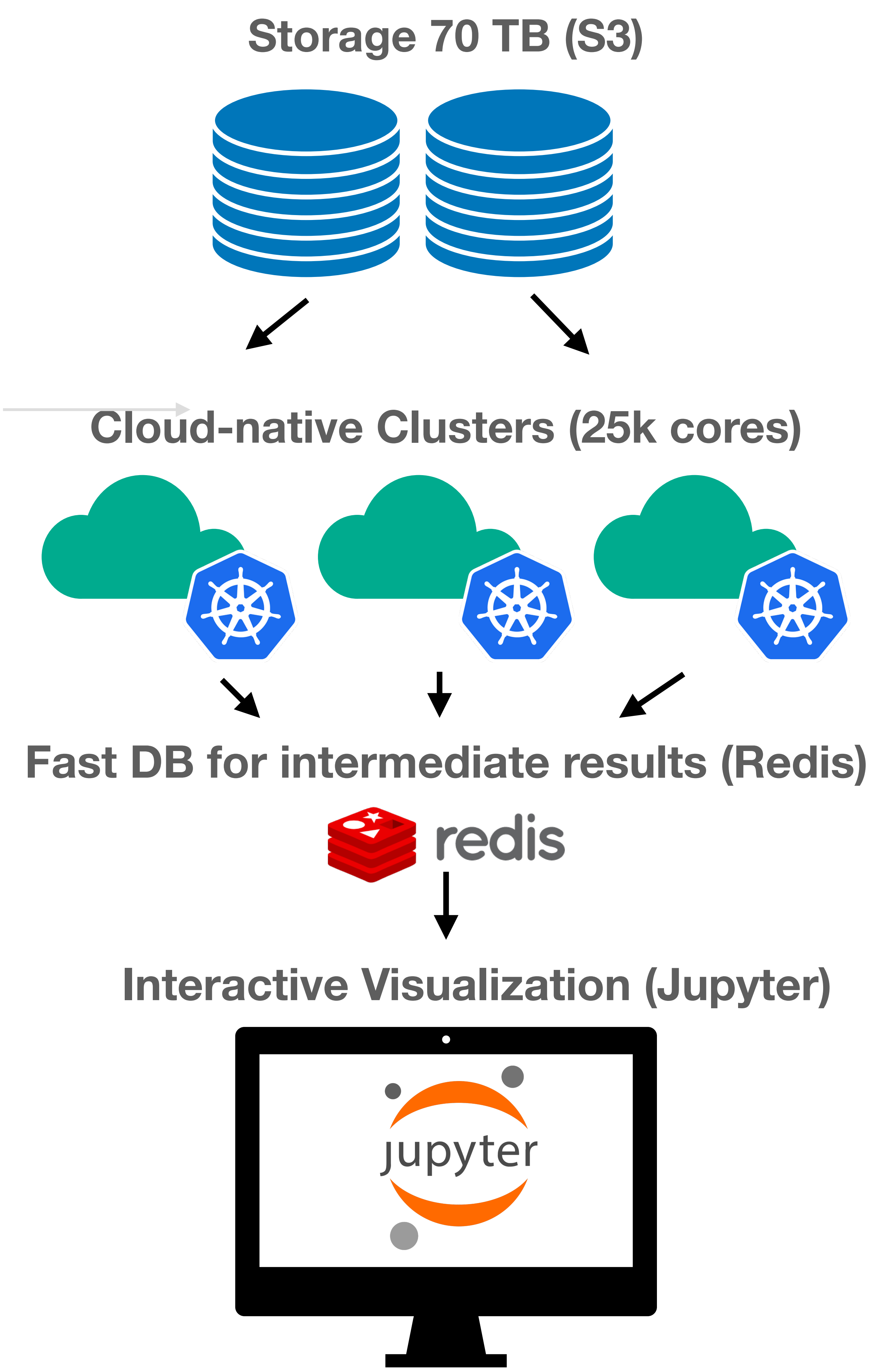

Uses Kubernetes to schedule work

- cluster system originated in Google
- can handle service, batch, interactive work
- containers: easy to deploy HEP s/w

Cloud characteristic: only CPUh matters

- 25k CPU @ 5 minutes = 8 CPU @ 10d
  (not true for physicist-hours, try to minimize time to do physics)

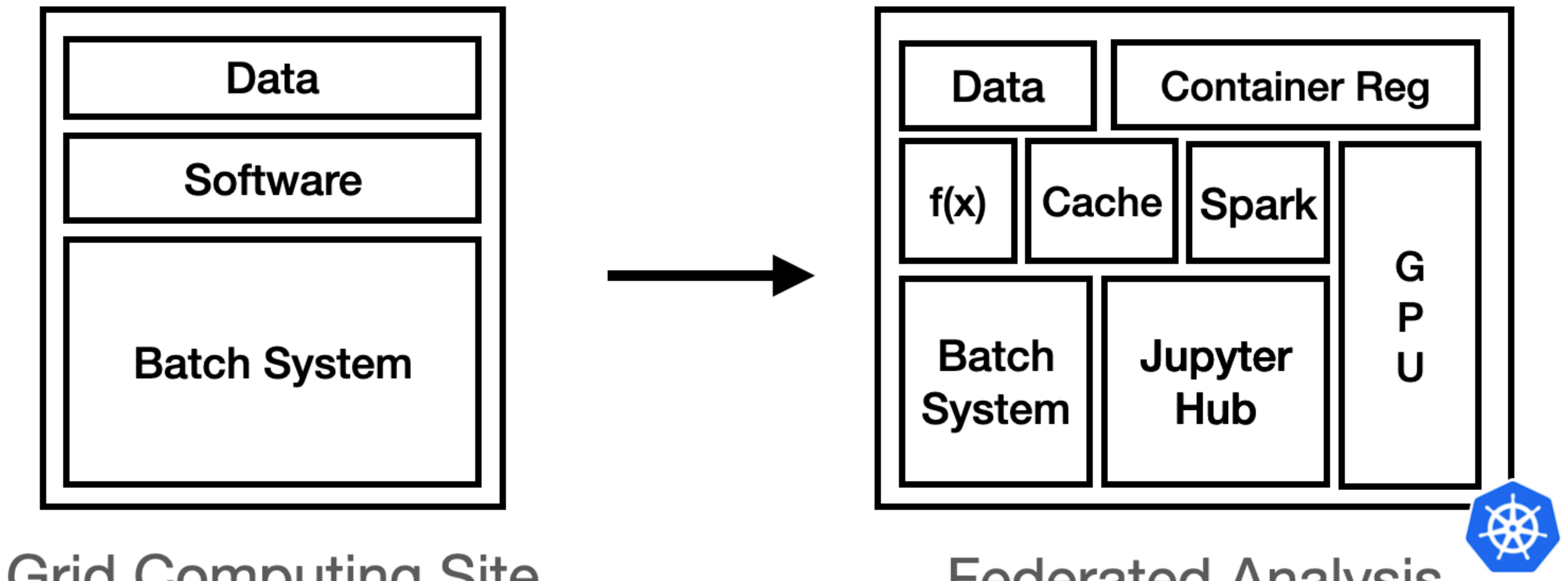
- individual user can get this on-demand
  (as long as you can pay)



# Bringing the technology to HEP

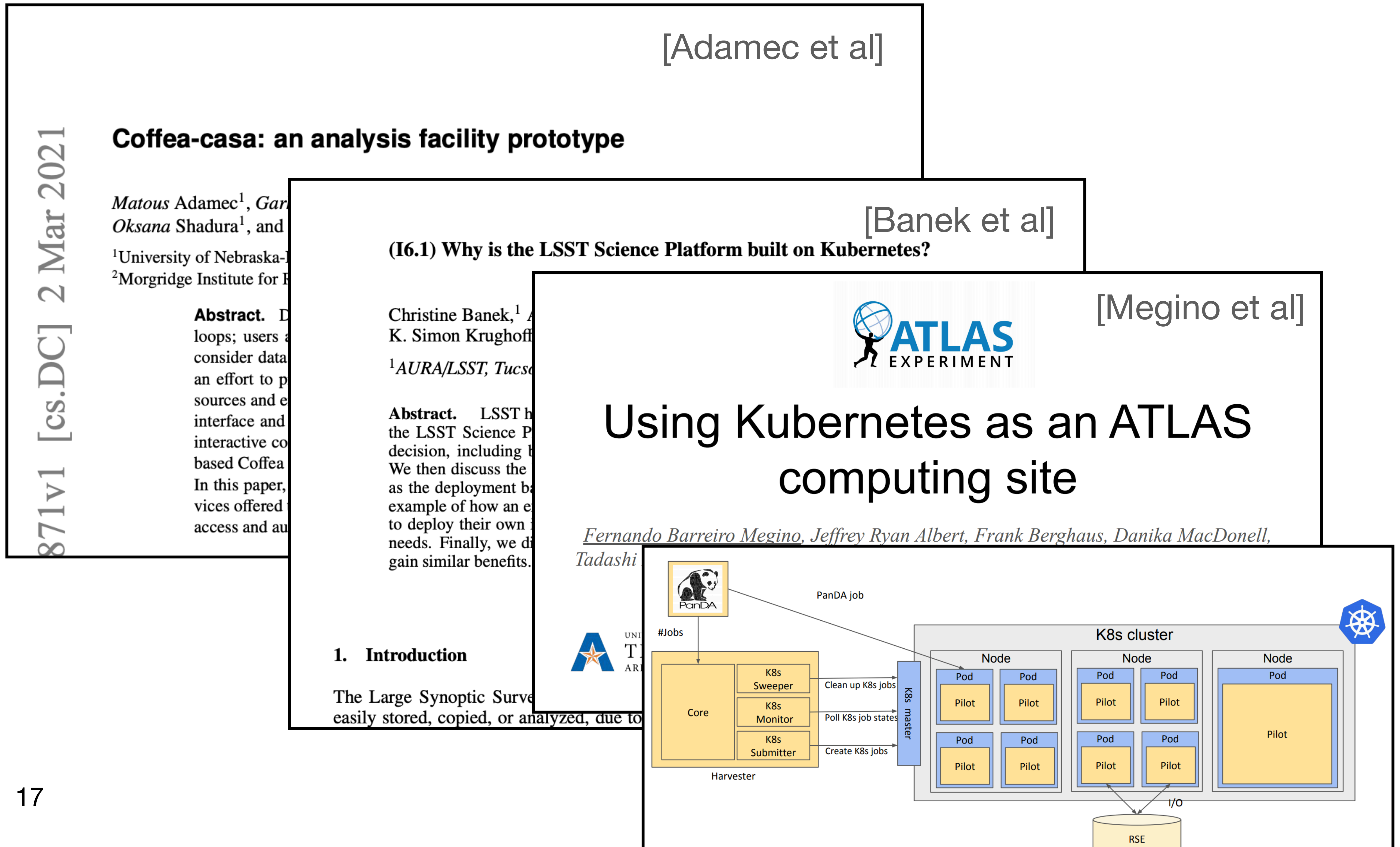
- **push to go beyond standard batch/grid picture**
  **→ Analysis Facilities / Science Platforms (cf: PUNCH4NFDI / IRIS-HEP)**

- **crucial for maintaining physics-driven/explorative data analysis at next phase of physics experiments: technology choices matter.**

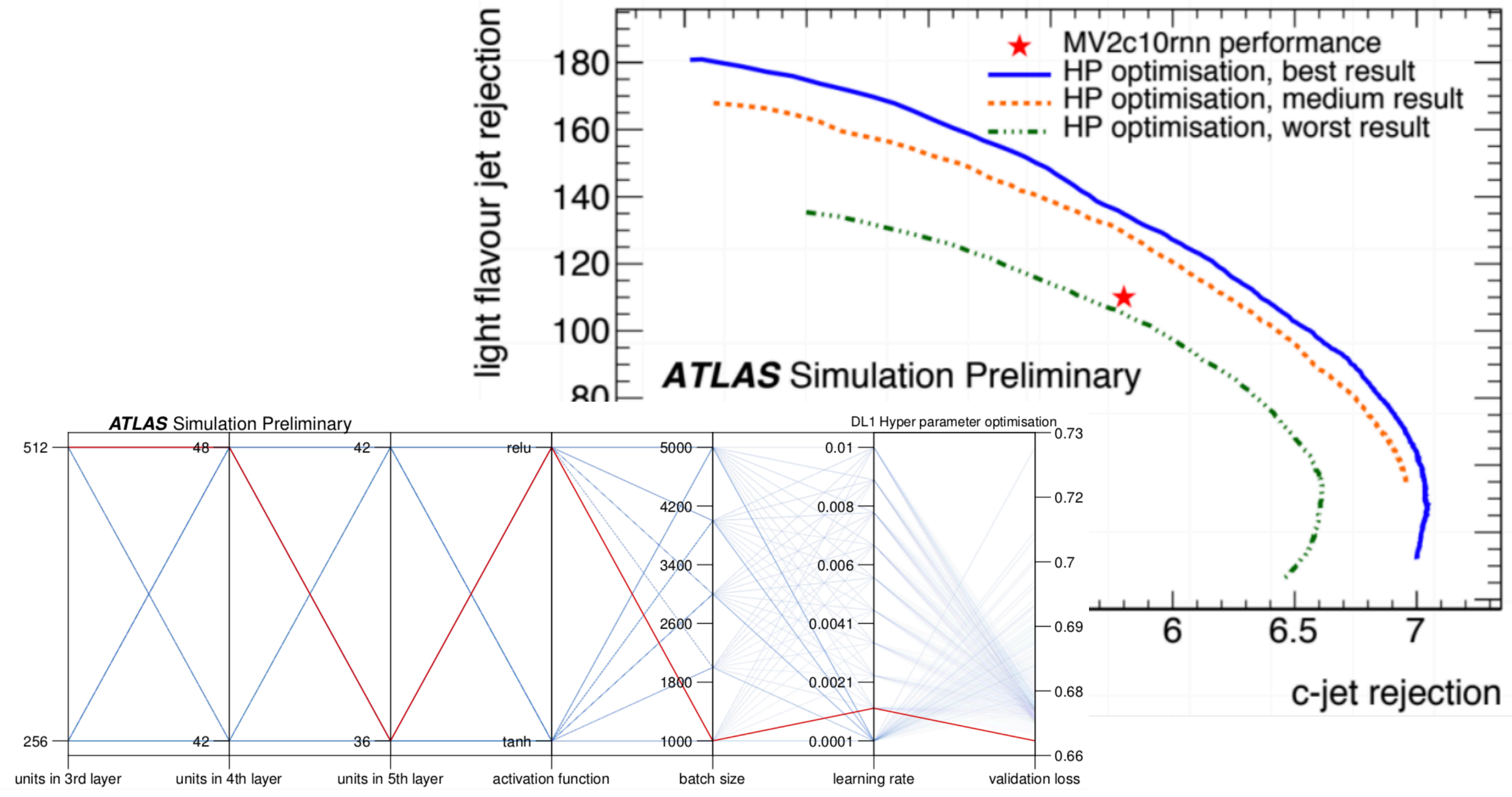



Grid Computing Site

Federated Analysis Facility



# Bringing the technology to HEP

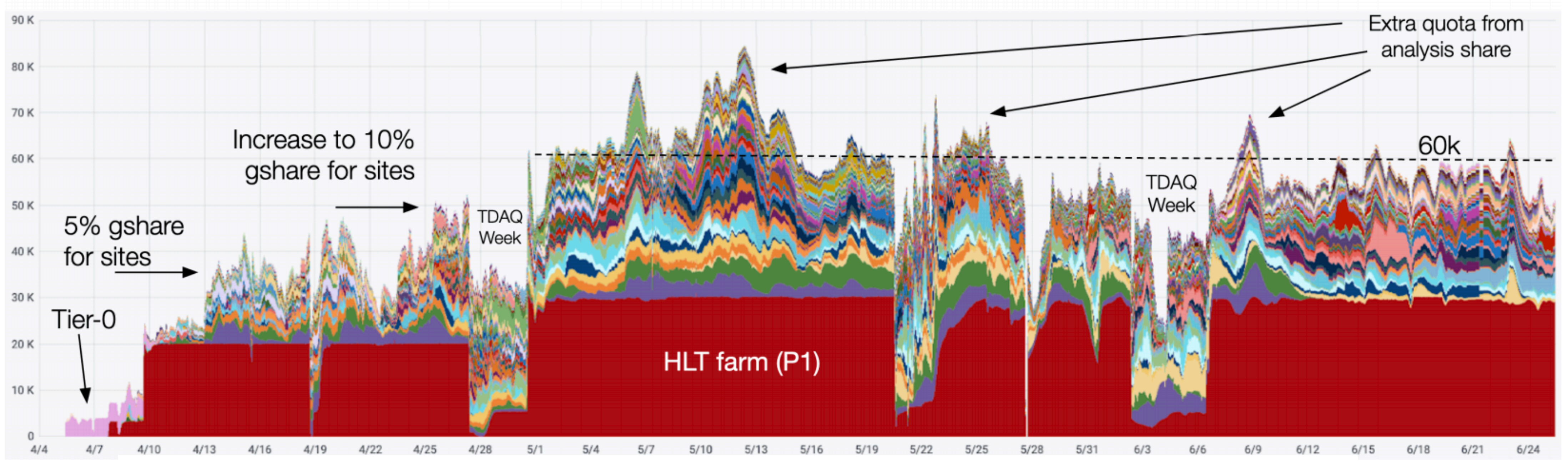


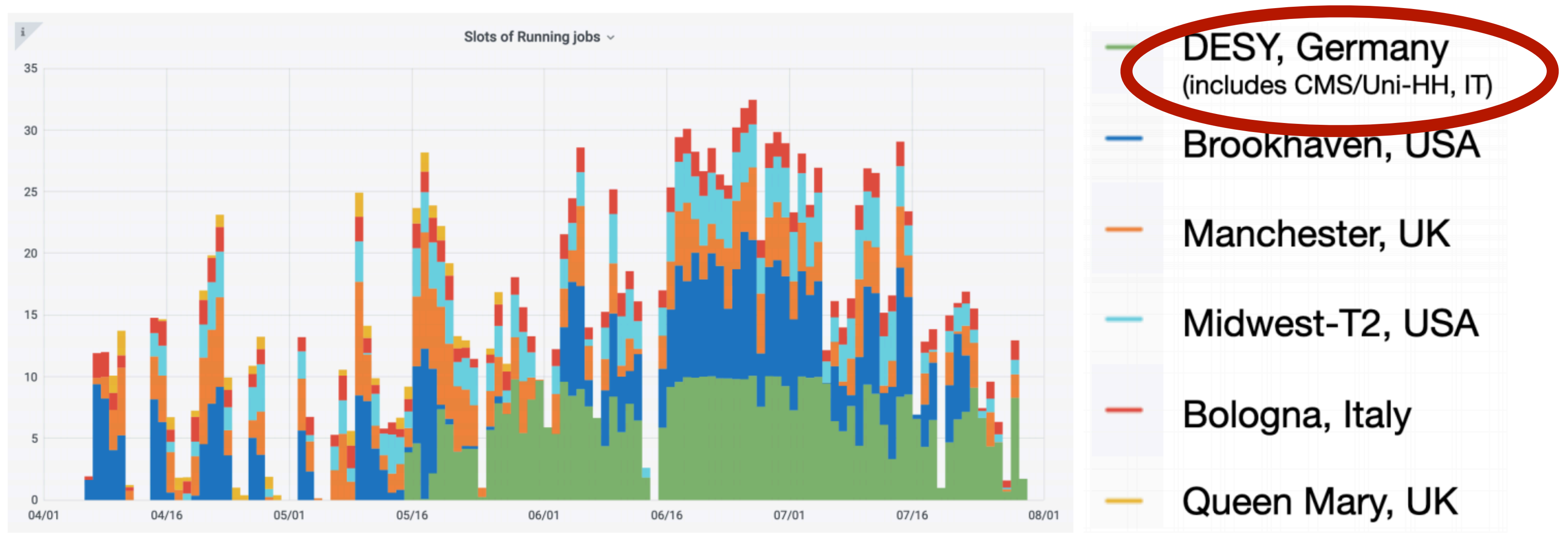

## In production:

- Grid-scale ML on GPUs (ATLAS)
- WLCG for volunteer computing (folding@home, 50k cores)

## Key: Containers allowing user-defined s/w environment

A Forti, M Guth, D Guest, C Pollard, LH




**Folding@Home Teams Overall Rankings**

Teams are color coded based on 24 hour average production.
Color codes are as follows: 0 / 1+ / 100k+ / 250k+ / 500k+ / 1M+ / 2.5M+ / 5M+ / 10M+

| Rank Overall | Team Name | Users Active | Users Total | Change 24hr | Change 7days | Points 24hr Avg | Points Update | Points Today | Points Week | P… |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Default (Team 0) | 1,200 | 2,000 | | | 14,868,491,388 | 1,629,724,245 | 4,966,192,700 | 33,501,103,756 | 732,3… |
| 2 | LinusTechTips_Team | 26,910 | 105,550 | | | 5,072,271,918 | 595,799,121 | 1,704,651,305 | 11,430,427,220 | 378,6… |
| 3 | Curecoin | 3,887 | 23,837 | | | 1,209,843,224 | 154,586,434 | 445,858,938 | 2,809,558,226 | 1,164,5… |
| 4 | PC Master Race - PCMR | 7,746 | 48,847 | | +2 | 1,056,141,646 | 114,336,343 | 329,271,240 | 2,302,740,356 | 83,5… |
| 5 | CERN & LHC Computing | 145 | 214 | +1 | +7 | 794,993,726 | 103,003,621 | 301,591,243 | 1,862,010,985 | 20,4… |
| 6 | NVIDIA Corp | 46 | 74 | | +6 | 943,008,120 | 83,172,296 | 236,982,978 | 1,861,543,403 | 26,3… |
| 7 | folding@evga | 1,003 | 19,610 | | | 644,096,447 | 78,377,970 | 230,860,571 | 1,477,464,856 | 379,0… |
| 8 | VMware | 769 | 1,696 | | +2 | 450,549,575 | 56,778,849 | 172,933,812 | 1,081,783,018 | 18,9… |

A Forti, D. South, I. Glushkov, R. Walker, M. Lassnig, LH

# Is dumping PB of data open enough?

# Beyond Open Data

**Open Data: with great freedom comes great responsibility**

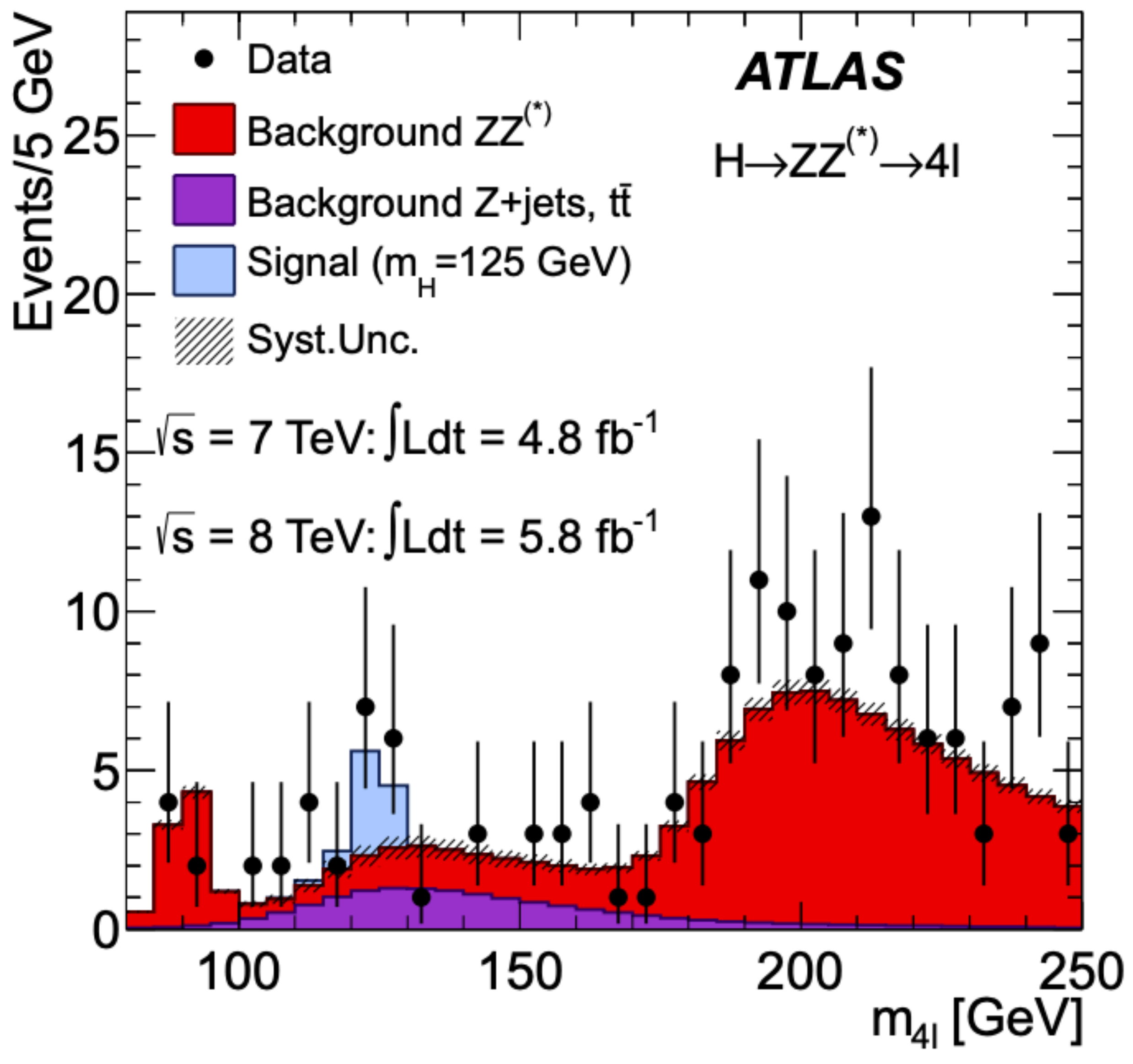
- **allows looking at uncovered corners of the dataset (comp. feasible)**
- **but: cost of developing an analysis *in full rigour* is enormous**

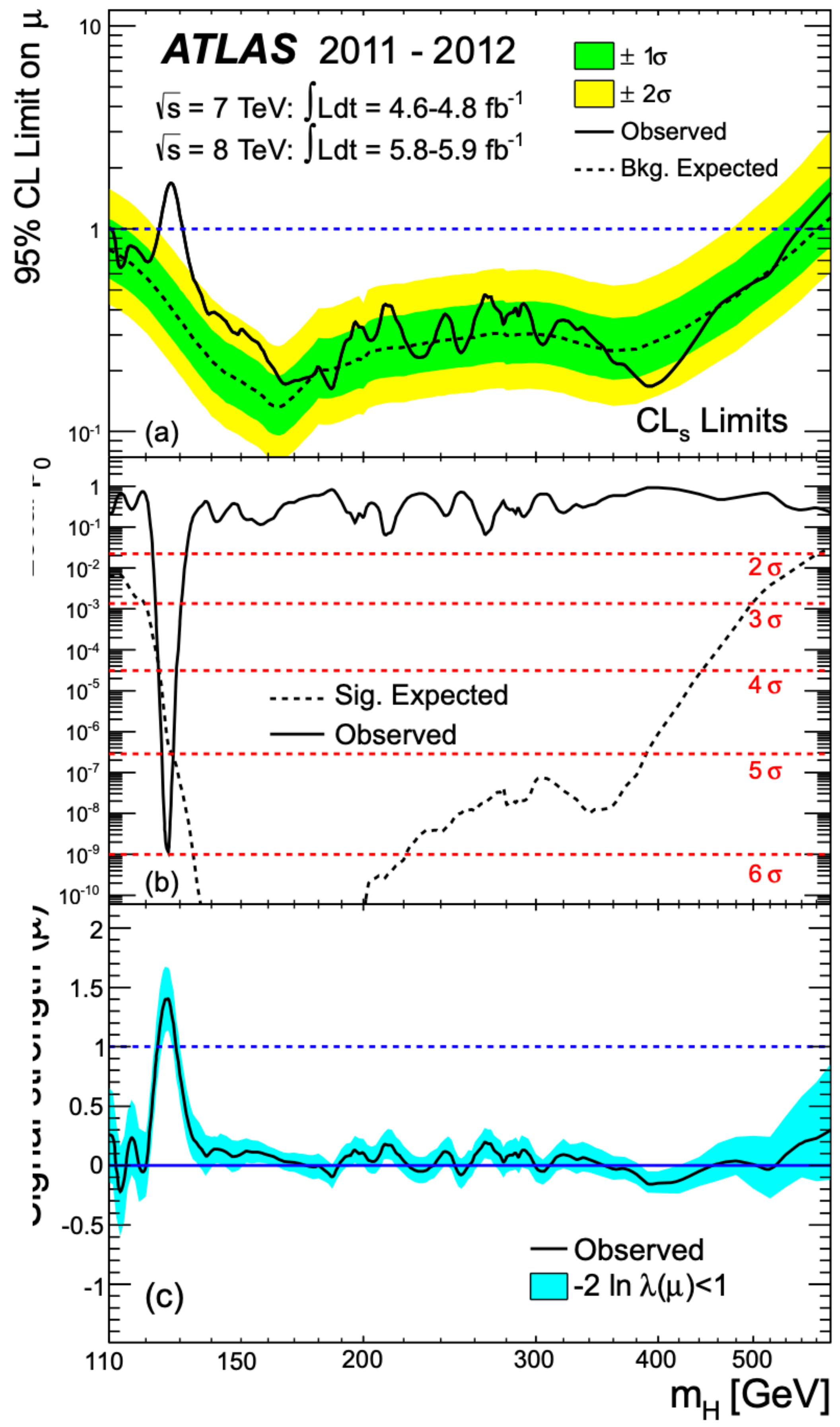
**Are there other ways data can be "open"?**
**→ instead of the source data, release the result data**

**What should we release?**

- **many types of results**

  **Event Counts, post-fit parameters**
  **p-value scans, ....**



# Remember:

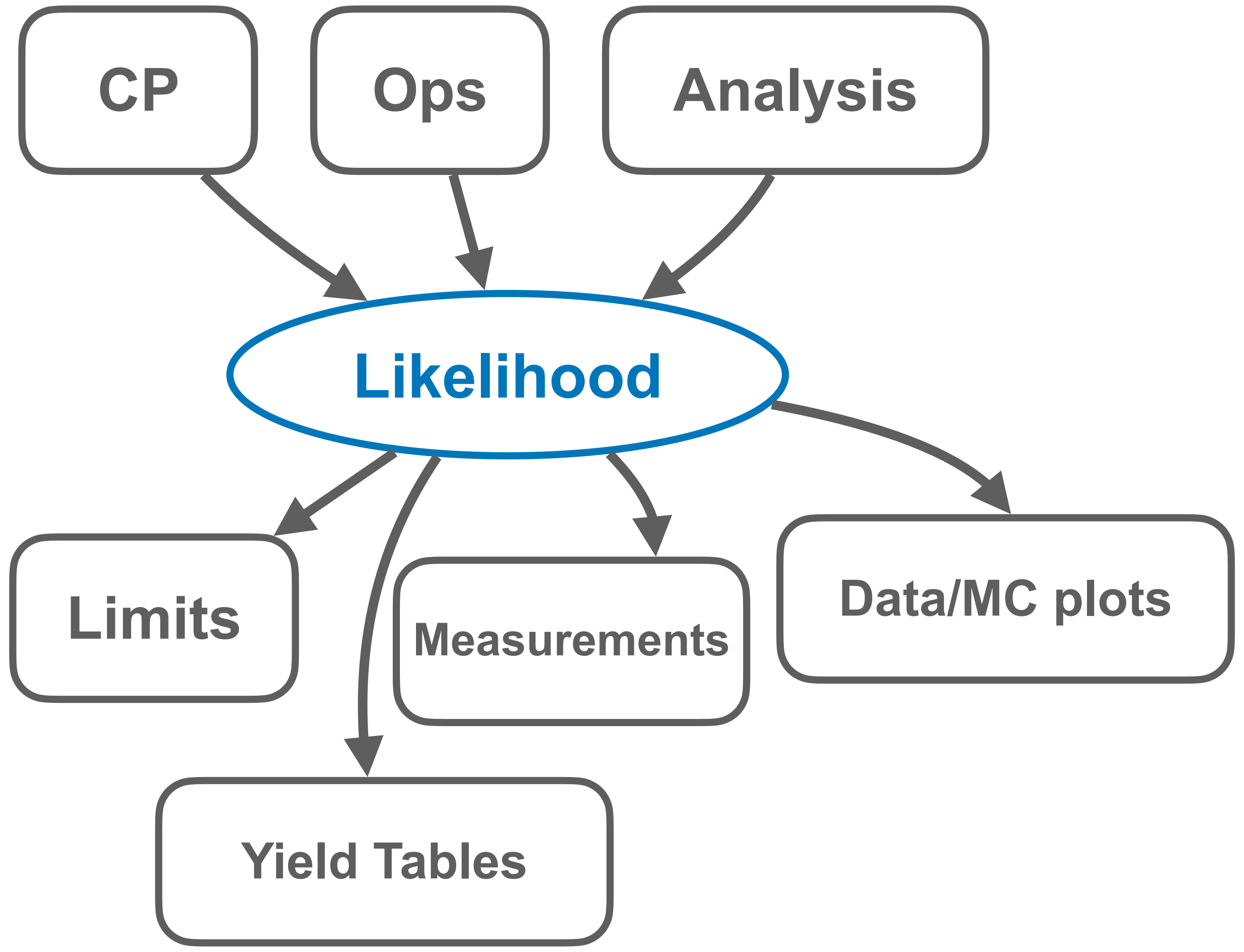
$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})}{p(\text{data})} p(\text{theory})$$

# Likelihood Principle:

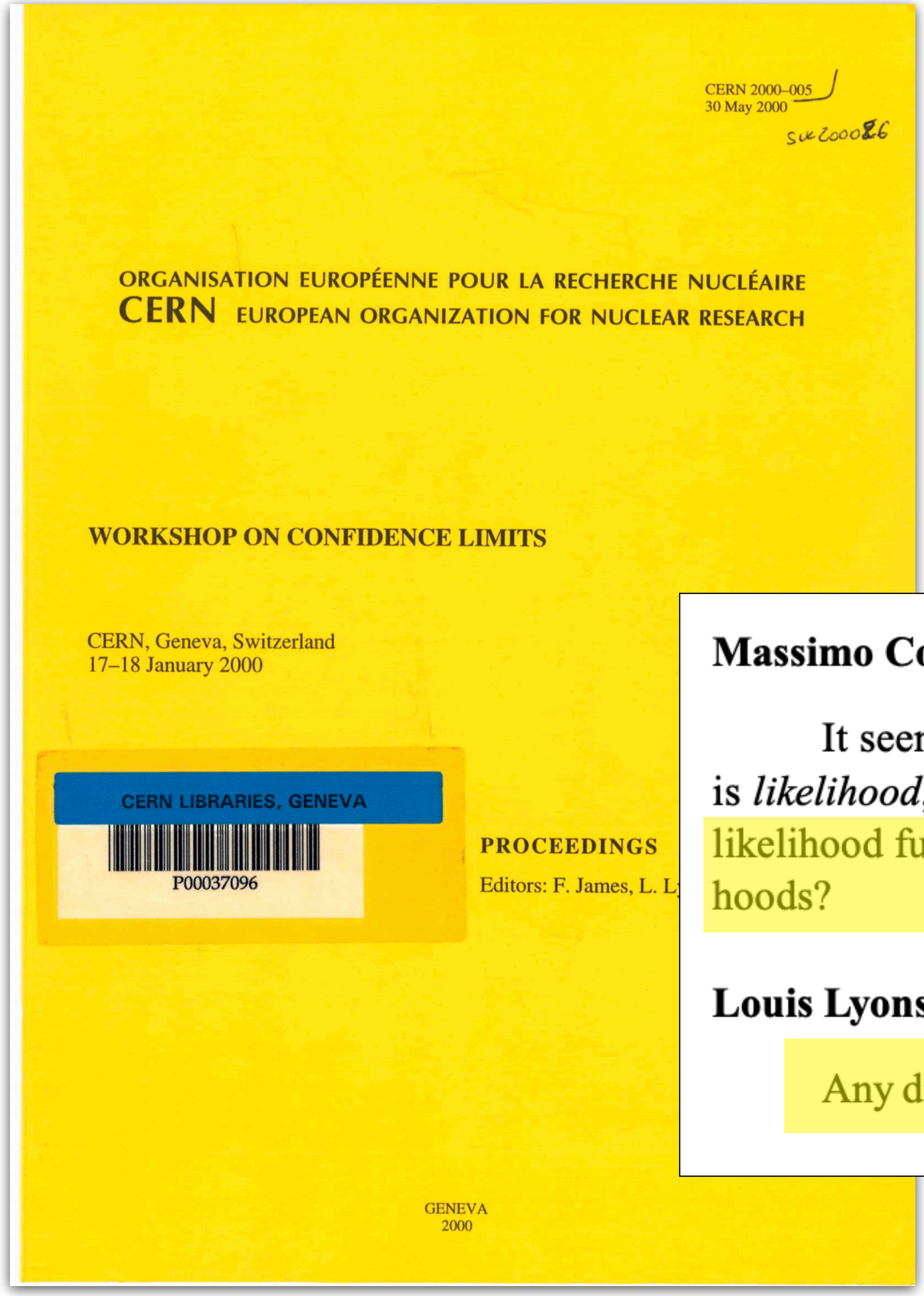
**Full experimental information captured in the likelihood $p(x\,|\,\theta)$!**

- **all analysis details are reflected in it**
- **most other types of results are derived from it.**

**Best, almost lossless summary of the measurements performed by a data analysis**



# 1st PHYSTAT workshop:


CERN 2000–005
30 May 2000

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17–18 January 2000



PROCEEDINGS
Editors: F. James, L. L

GENEVA
2000

- preserve the likelihood!
- universal concept applicable to a broad set of experiments (also beyond HEP)

**Massimo Corradi**

It seems to me that there is a general consensus that what is really meaningful for an experiment is *likelihood*, and almost everybody would agree on the prescription that experiments should give their likelihood function for these kinds of results. Does everybody agree on this statement, to publish likelihoods?

**Louis Lyons**

Any disagreement ? Carried unanimously. That's actually quite an achievement for this Workshop.

# Likelihoods

## How do we archive likelihoods?

## Not straightforward:

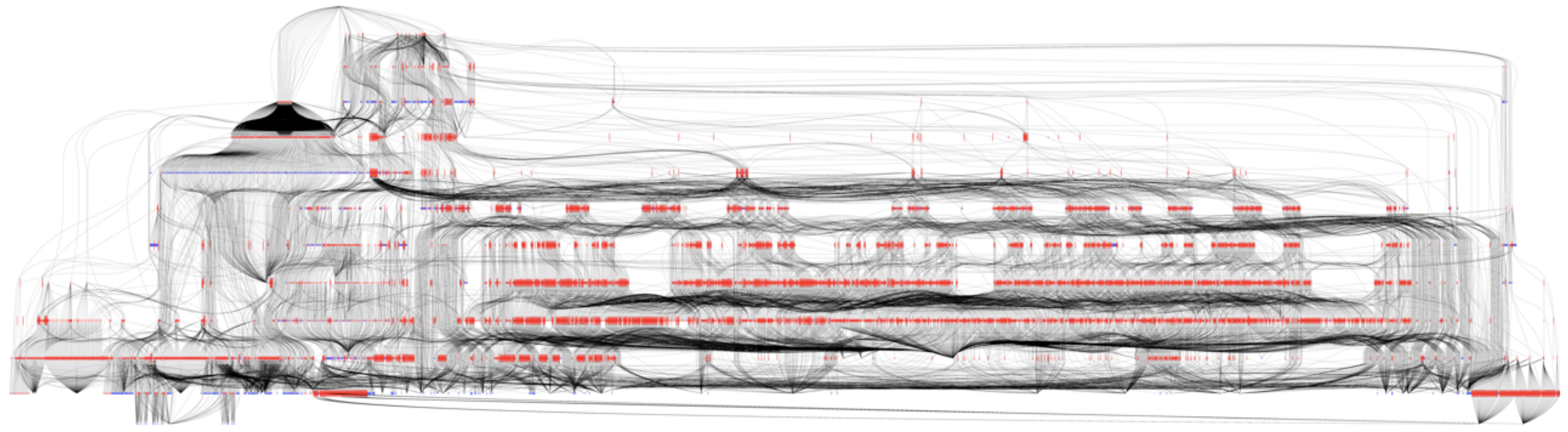
$p(x \,|\, \theta)$ **could by anything! (Open World)**

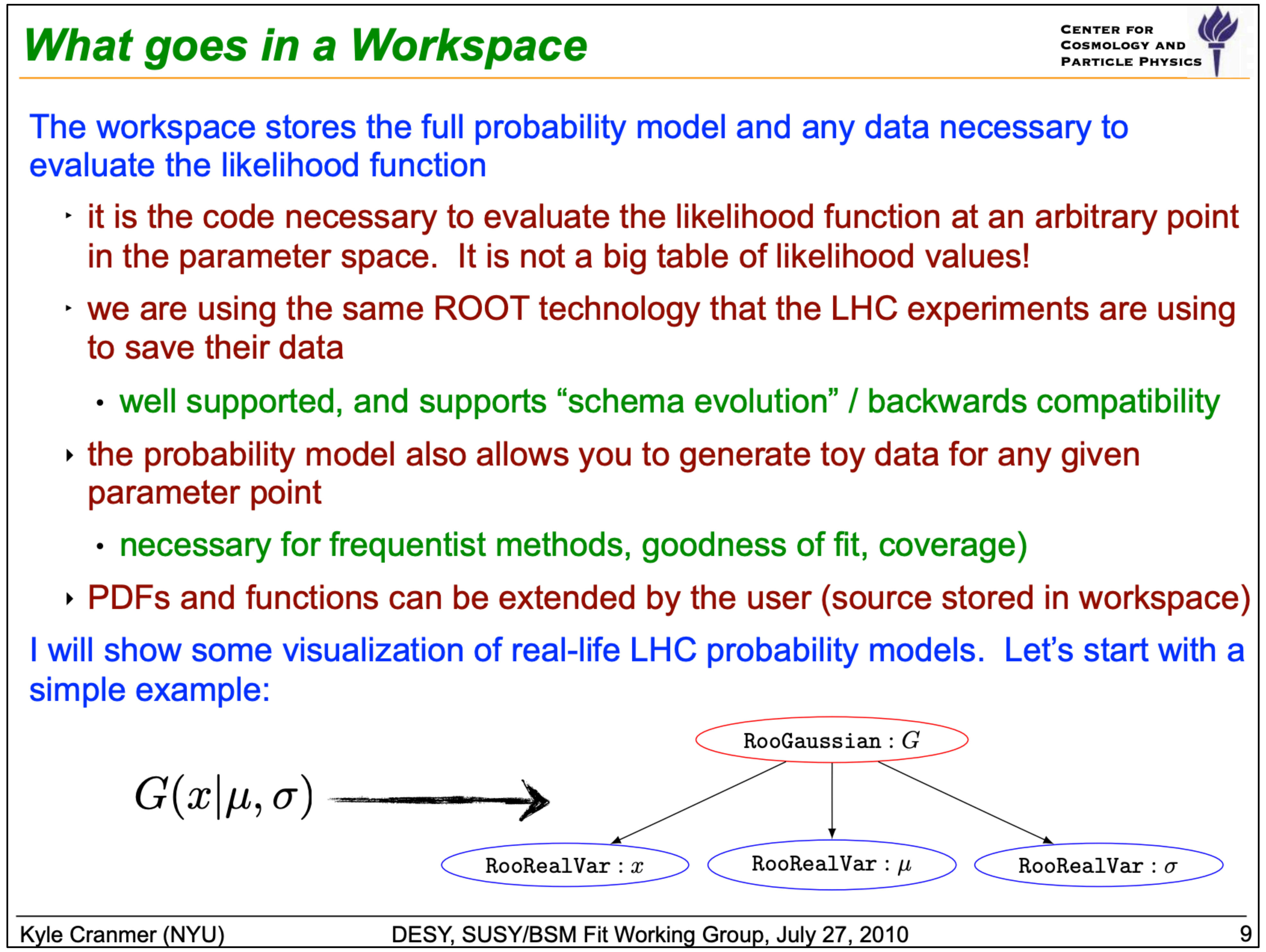
## Obvious Candidate: RooFit workspaces

- Key development in early LHC stats
- designed as "sharable data product"

## But:

- opaque, binary format
- tied to specific data analysis tool






[K. Cranmer @ DESY 2010]

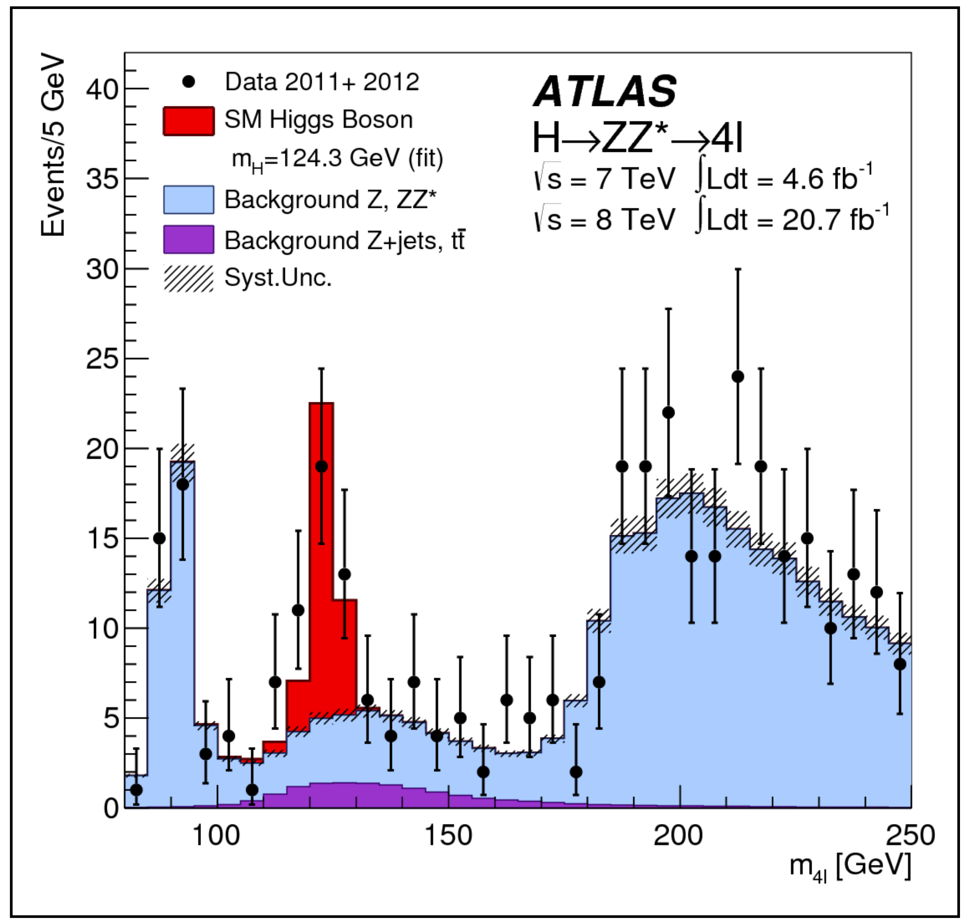
# Likelihoods

## $p(x \mid \theta)$ could by anything... but in reality they're not!

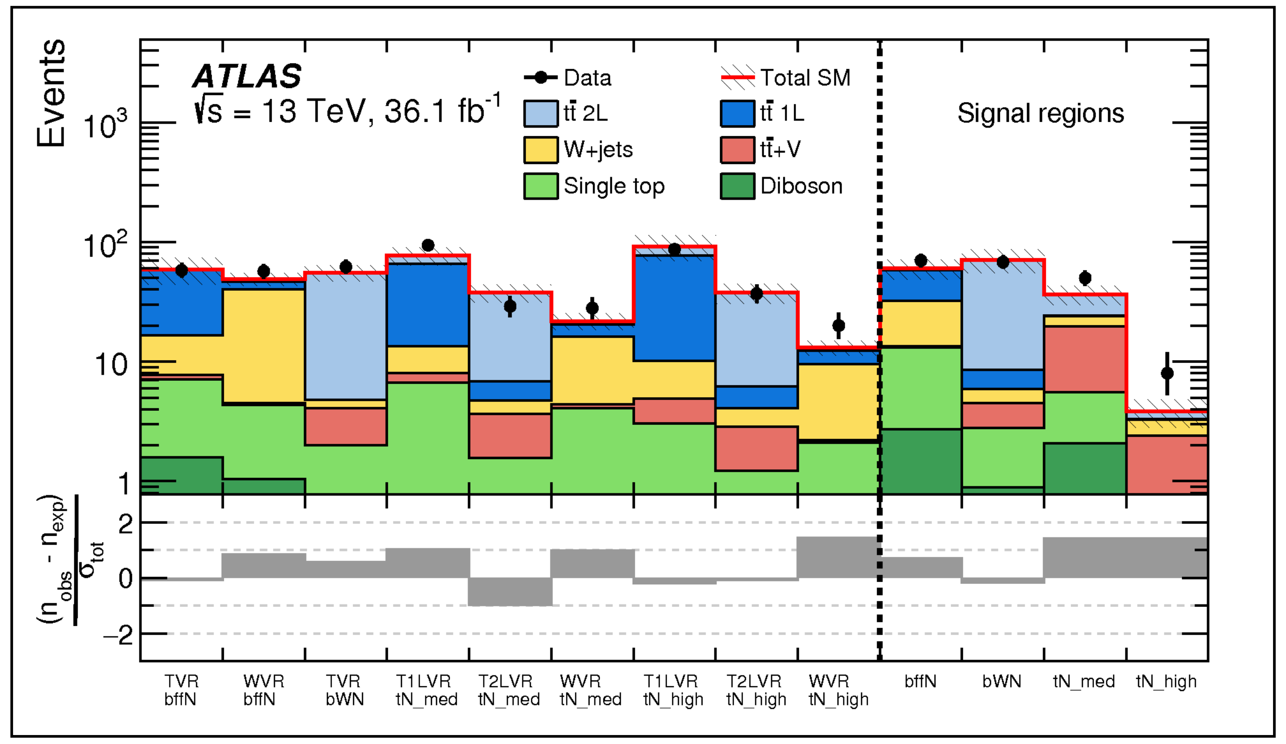
**For LHC a very large fraction of analyses use standardized binned likelihood functions built from data counts + simulation-derived templates**

**ATLAS: HistFactory** →
CMS: HiggsCombine

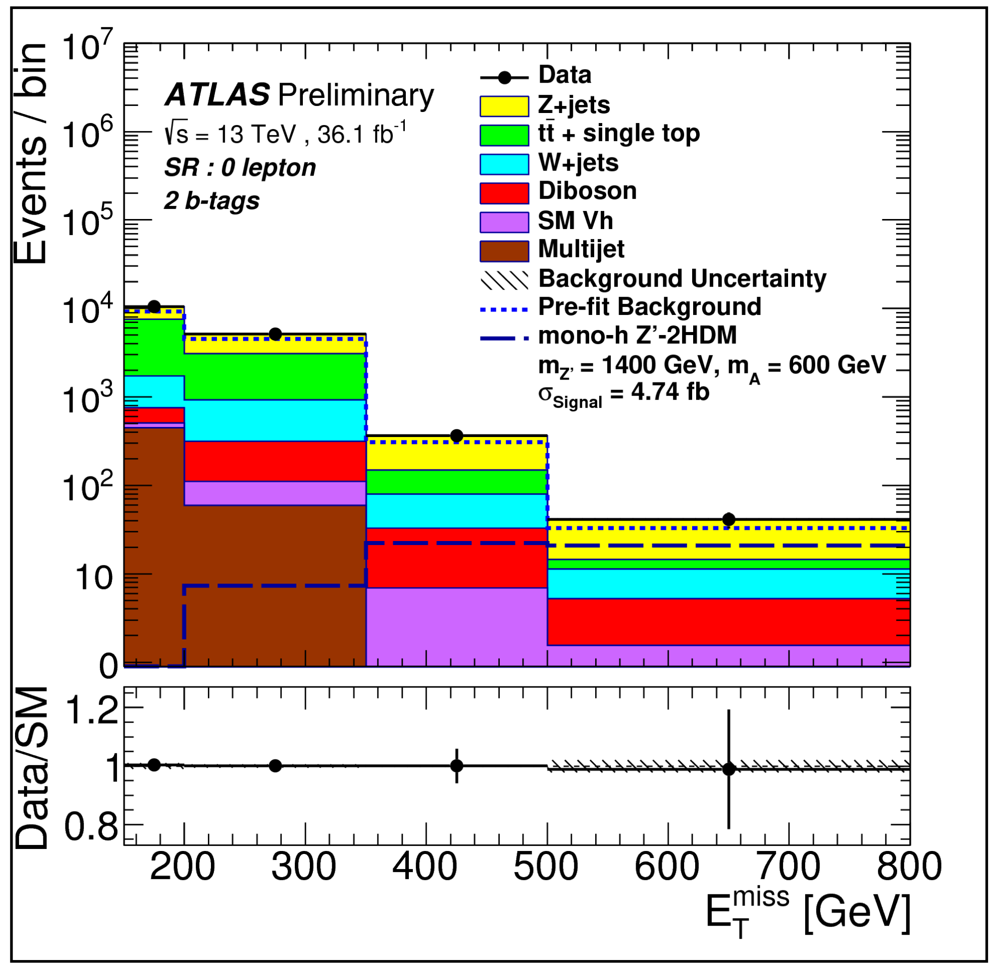
$$f(\boldsymbol{n}, \boldsymbol{a} \mid \boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb} \mid \nu_{cb}(\boldsymbol{\eta}, \boldsymbol{\chi})\right)}_{\text{Simultaneous measurement of multiple channels}} \underbrace{\prod_{\chi \in \mathcal{X}} c_\chi(a_\chi \mid \chi)}_{\text{constraint terms for "auxiliary measurements"}},$$




SM




SUSY




Exotics

# HistFactory

## Fixed Template (Closed World) simplifies likelihood problem:

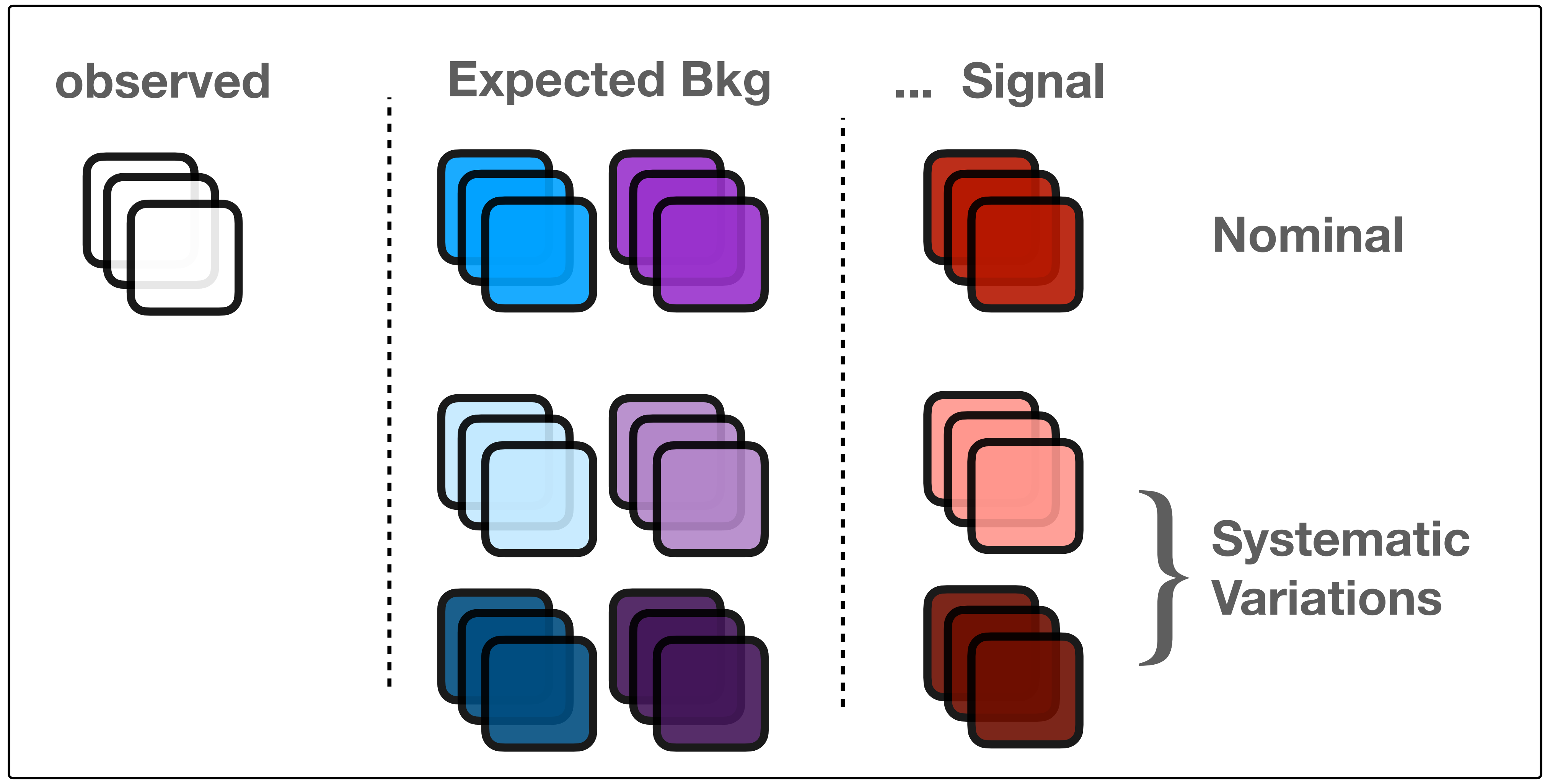
- archive template ingredients (pure tabular data)
- template math defined independent of implementation - not tied to s/w

Fixed Likelihood template

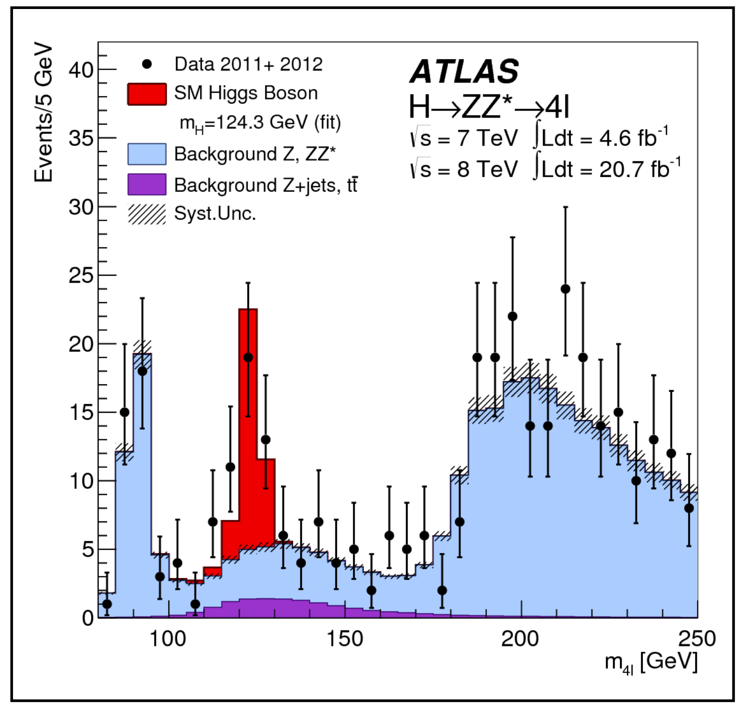
$$f(\boldsymbol{n}, \boldsymbol{a} \mid \boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb} \mid \nu_{cb}(\boldsymbol{\eta}, \boldsymbol{\chi})\right)}_{\text{Simultaneous measurement of multiple channels}} \underbrace{\prod_{\chi \in \boldsymbol{\chi}} c_\chi(a_\chi \mid \chi)}_{\text{constraint terms for "auxiliary measurements"}},$$

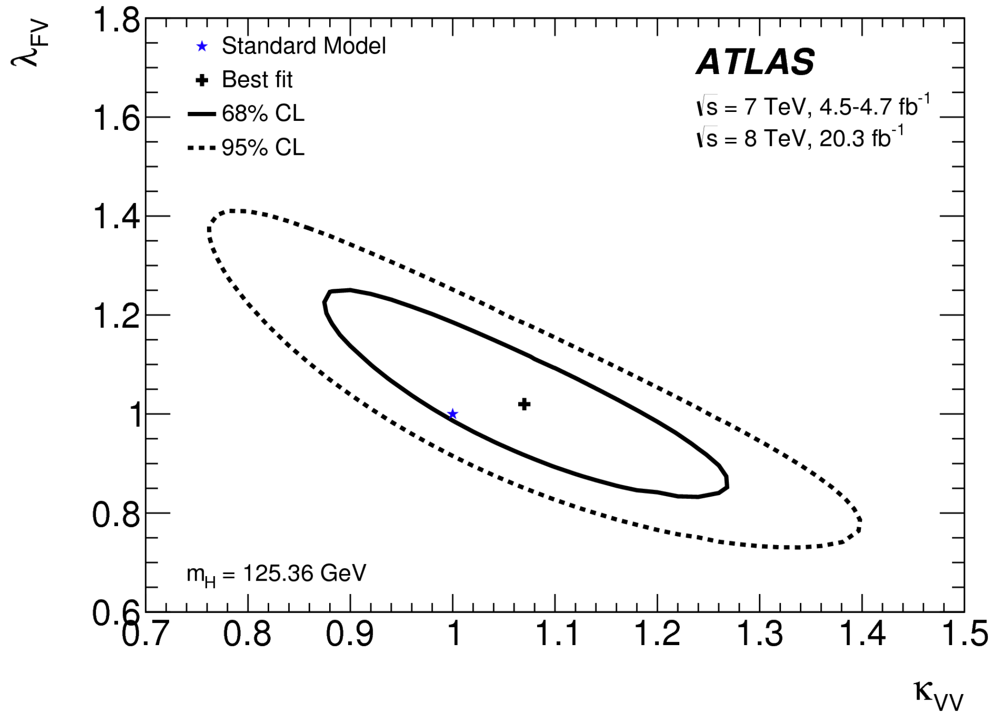
+

Likelihood ingredients

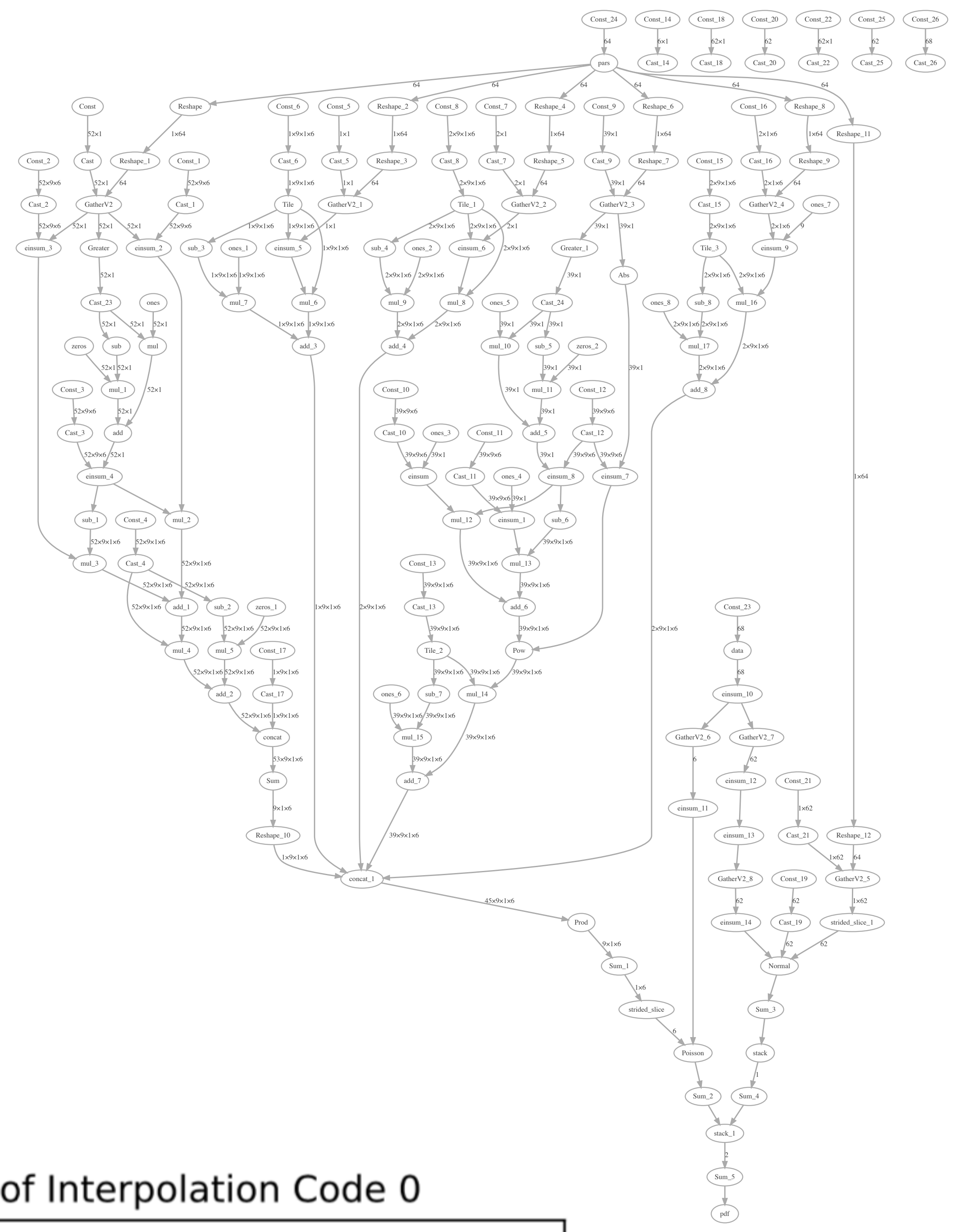



=



# pyhf - fast, pythonic HistFactory

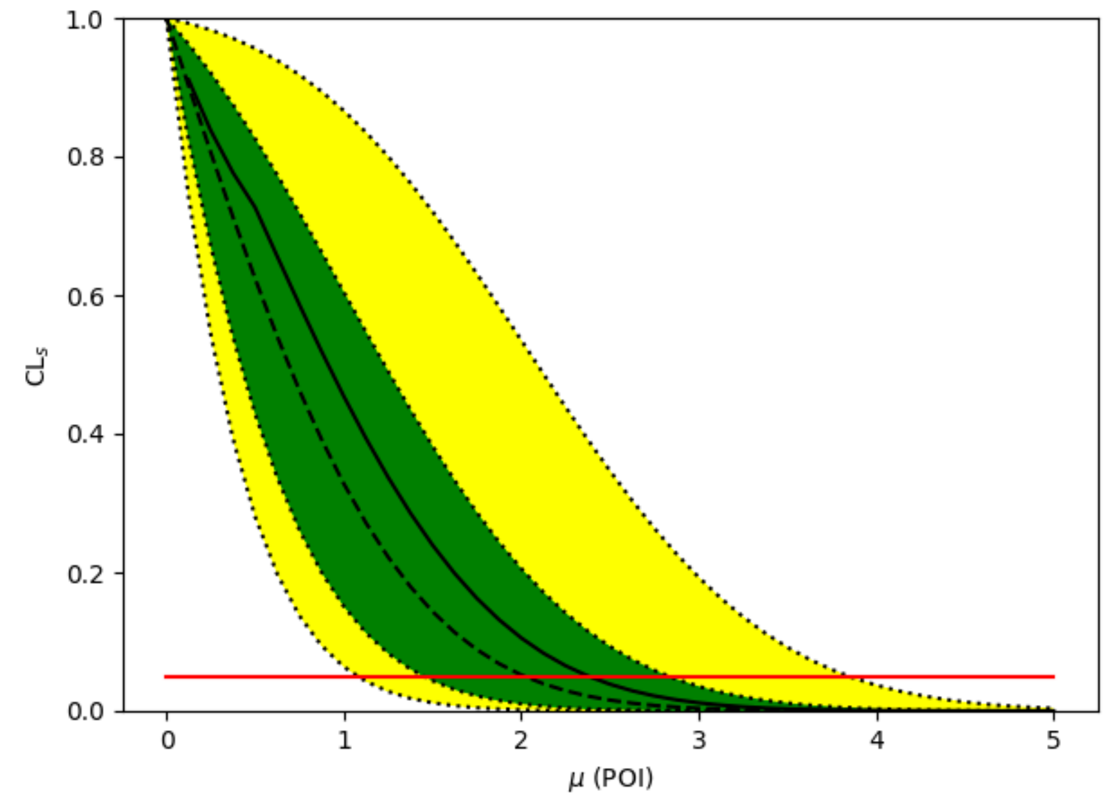


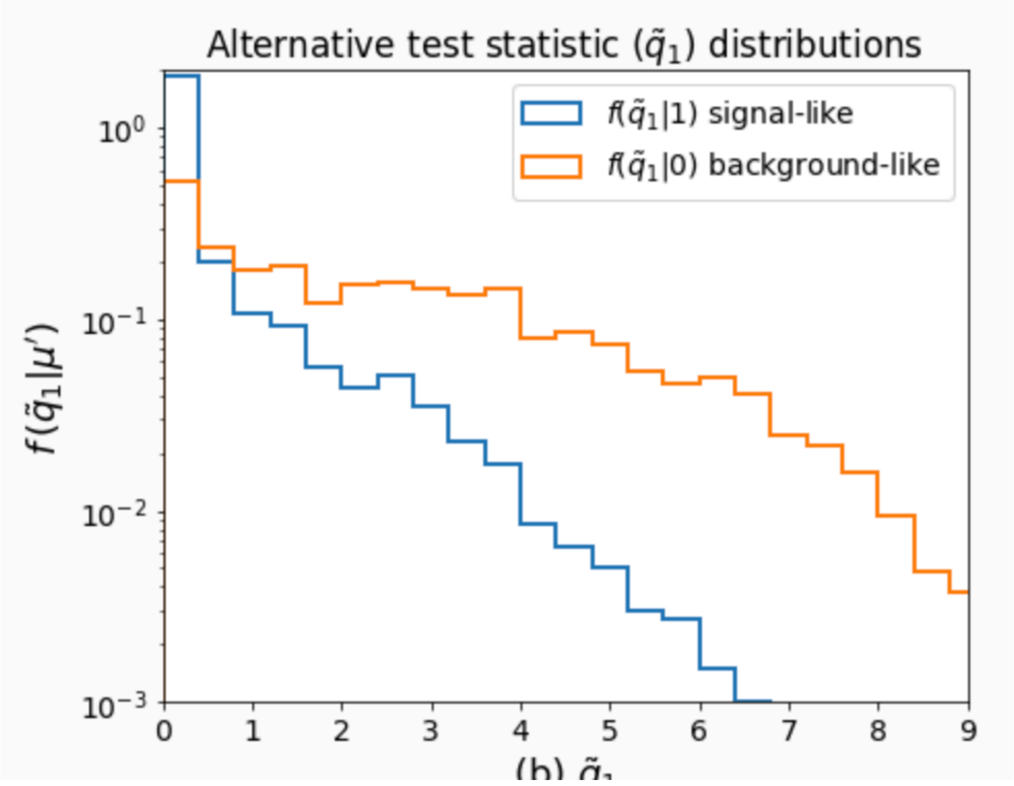
**Statistics tool with likelihoods as vectorized computation based on ML libraries**

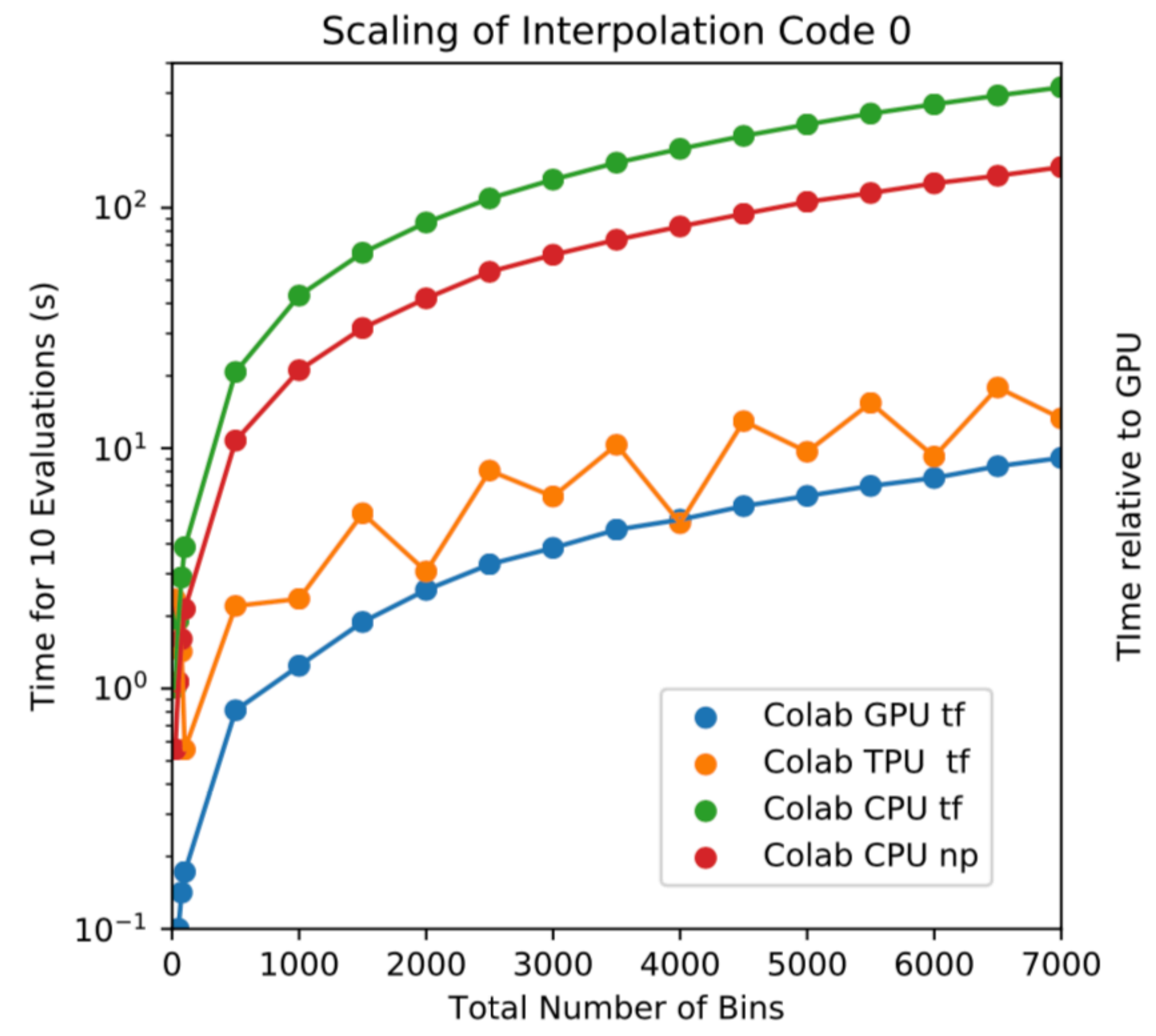
- **run on heterogenous hardware (TPU, GPU)**
- **automatic differentiation for exact** $\nabla_\theta p(x|\theta)$

**Full suite of HEP inference algorithms**

- **Profile Likelihood, Asymptotics & Toys**




[LH, Feickert, Stark]

# pyhf - fast, pythonic HistFactory

## Nice:

- adoption across theory-experiment divide
- developed at LHC, but used in larger range of settings

**Charged Lepton Flavor Violation at the EIC**

Vincenzo Cirigliano, Kaori Fuyuto, Christopher Lee, Emanuele Mereghetti, and Bin Yan

*Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.*

*E-mail:* c… emereghe…

acceptance, obtained from the yields of the two validation regions given in Ref. [63]. We then used signal and background events in a likelihood analysis using pyhf [66], obtaining[3]

$$\mathrm{BR}(t \to q e \tau) \leq 2.2 \cdot 10^{-4}. \quad (6.10)$$

arXiv: 2102.06176

**EIC**

**Sensitivity of Future Hadron Colliders to Leptoquark Pair Production in the Di-Muon Di-Jets Channel**

B. C. Allanach[1], Tyler Corbett[2], Maeve Madigan[a,1]

[1]DAMTP, Unive… [2]The Niels Bohr…

fixed $\mu$.

Suppose the measured data $n_i$ shows no fluctuations above the SM background $b_i$. To set exclusion limits on

The upper limit at 95% CL on $\mu$ is then given by the value of $\mu$ at which $\mathrm{CL}_s = 0.05$. We compute the $\mathrm{CL}_s$ values using pyhf [64], a Python implementation of

arXiv: 1911.04455

**FCC**

ZU-TH 02/20

**Lepton Flavor Violation and Dilepton Tails at the LHC**

Andrei Angelescu,[1,*] Darius A. Faroughy,[2,†] and Olcyr Sumensari[3,4,‡]

$Z' \to \mu^\pm \tau_h^\mp$, respectively. Given that the search focuses on the decay of a heavy resonance, the resulting leptonic pair is expected to fly away back-to back in the azimuthal plane. Hence, in order to reduce the leading backgrounds, the cut

mass bins into a likelihood function based on Poissonian distributions. The 95% confidence level (CL) upper limits were extracted using the $\mathrm{CL}_s$ method [40] with the pyhf package [41]. For High Luminosity (HL) projections, we repeated the

arXiv: 2002.05684

**LHC BSM**

**How to discover QCD Instantons at the LHC**[1]

Simone Amoroso[a] Deepak Kar[b] Ma…

[a]*DESY, Hamburg, Germany*

respective *tight* signal region selection is applied. The systematic uncertainty on the background is estimated as described in the previous sections, and ranges from about 20% at low invariant masses to about 50% at high invariant masses. The limits are then performed as counting experiment with the pyhf package [51]. The results are shown in Fig. 21 for

arXiv: 2012.09120

**LHC QCD**

**Hunting wino and higgsino dark matter at the muon collider with disappearing tracks**

Rodolfo Capdevilla,[a,b] Fede…

diction in $\mathrm{SR}_{1t}^{\gamma}$ has been reduced to 10%. The discovery significance is evaluated from the expected discovery $p$-value, while limits are set at 95% CL using the CLs method [91] with the pyhf software package [92, 93]. Additional lines show the sensitivity of the conservative scenario inflating the background estimates by an order of magnitude. The sensitivity is

arXiv:2102.11292

**μ-Collider**

# pyhf byproduct: JSON serialization

## Plain-text format for HEP Workspaces

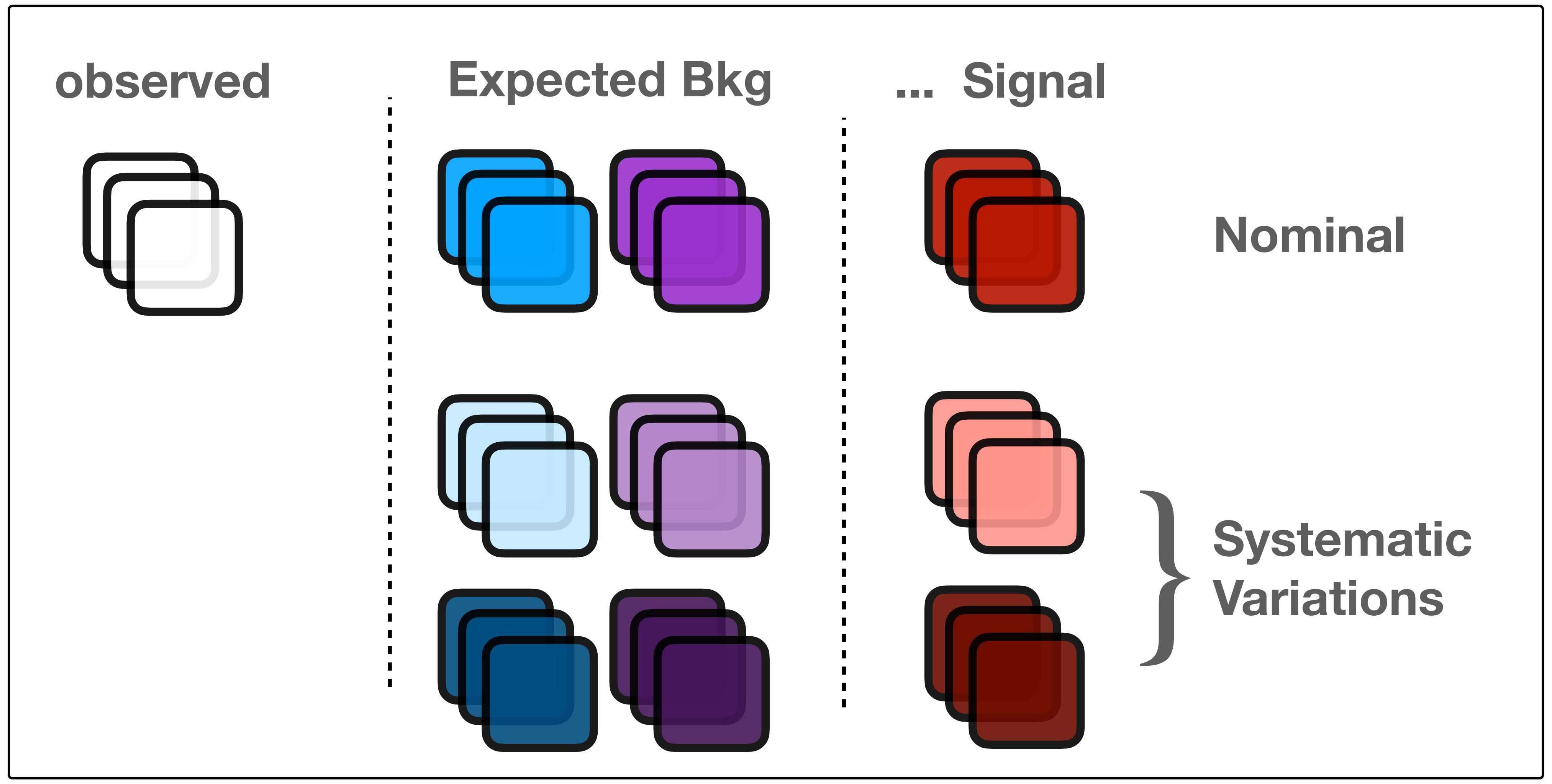
(for those using HistFactory template)

## Ubiquitous format: JSON

- readable in any language
- independent of s/w implementation

## Ideal for long-term archival of likelihoods

→ as promised in 2000





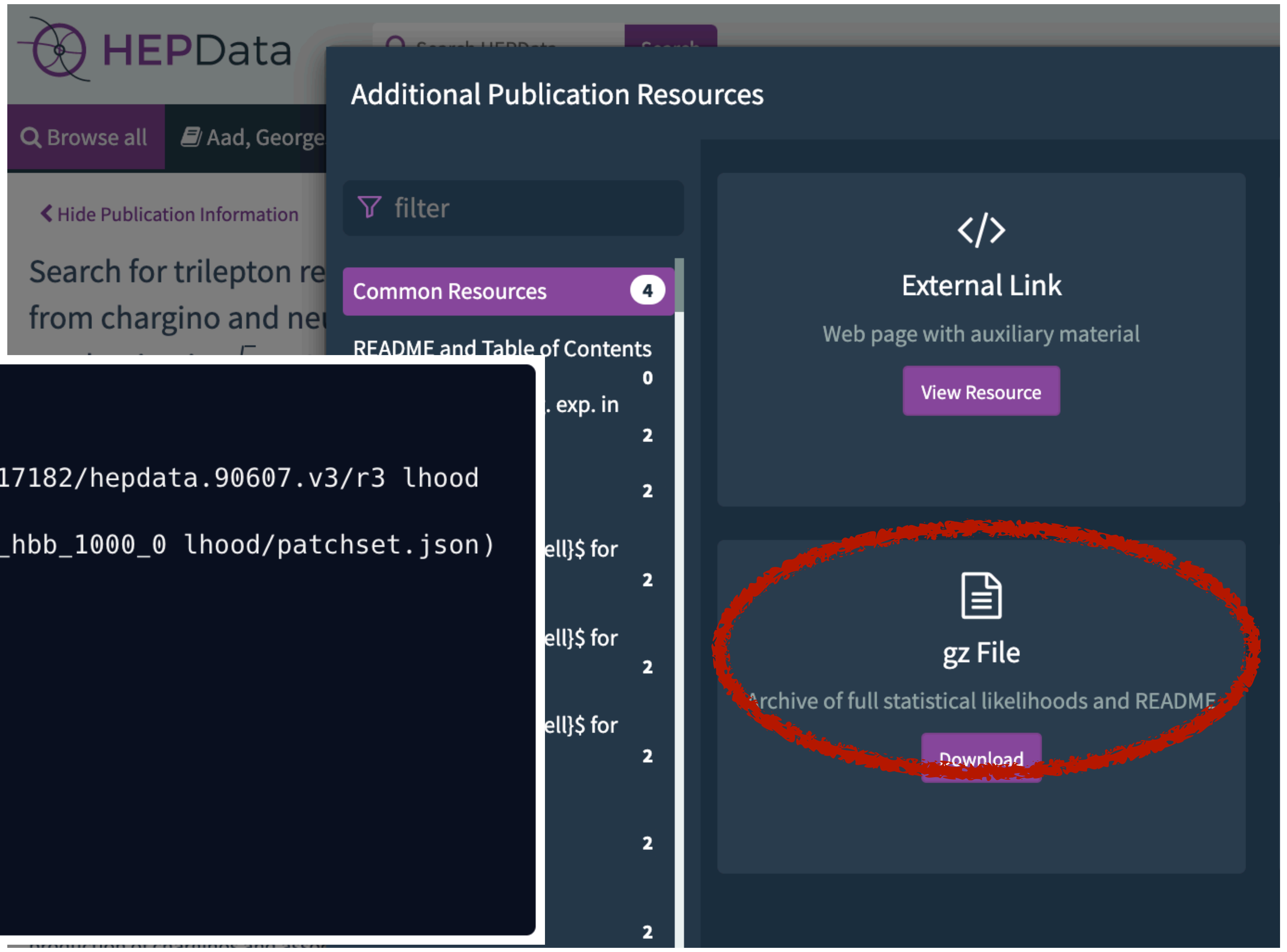
```
{
    "channels": [
        { "name": "singlechannel",
          "samples": [
            { "name": "signal",
              "data": [7.0, 2.0],
              "modifiers": [ { "name": "mu", "type": "normfactor", "data": null} ]
            },
            { "name": "background",
              "data": [50.0, 60.0],
              "modifiers": [ {"name": "uncorr_bkguncrt", "type": "shapesys", "data": [5.0,12.0]} ]
            }
          ]
        }
    ],
    "data": {
        "singlechannel": [50, 60]
    },
    "measurements": [
        { "name": "Measurement", "config": {"poi": "mu", "parameters": []} }
    ]
}
```


[LH, Feickert, Stark]

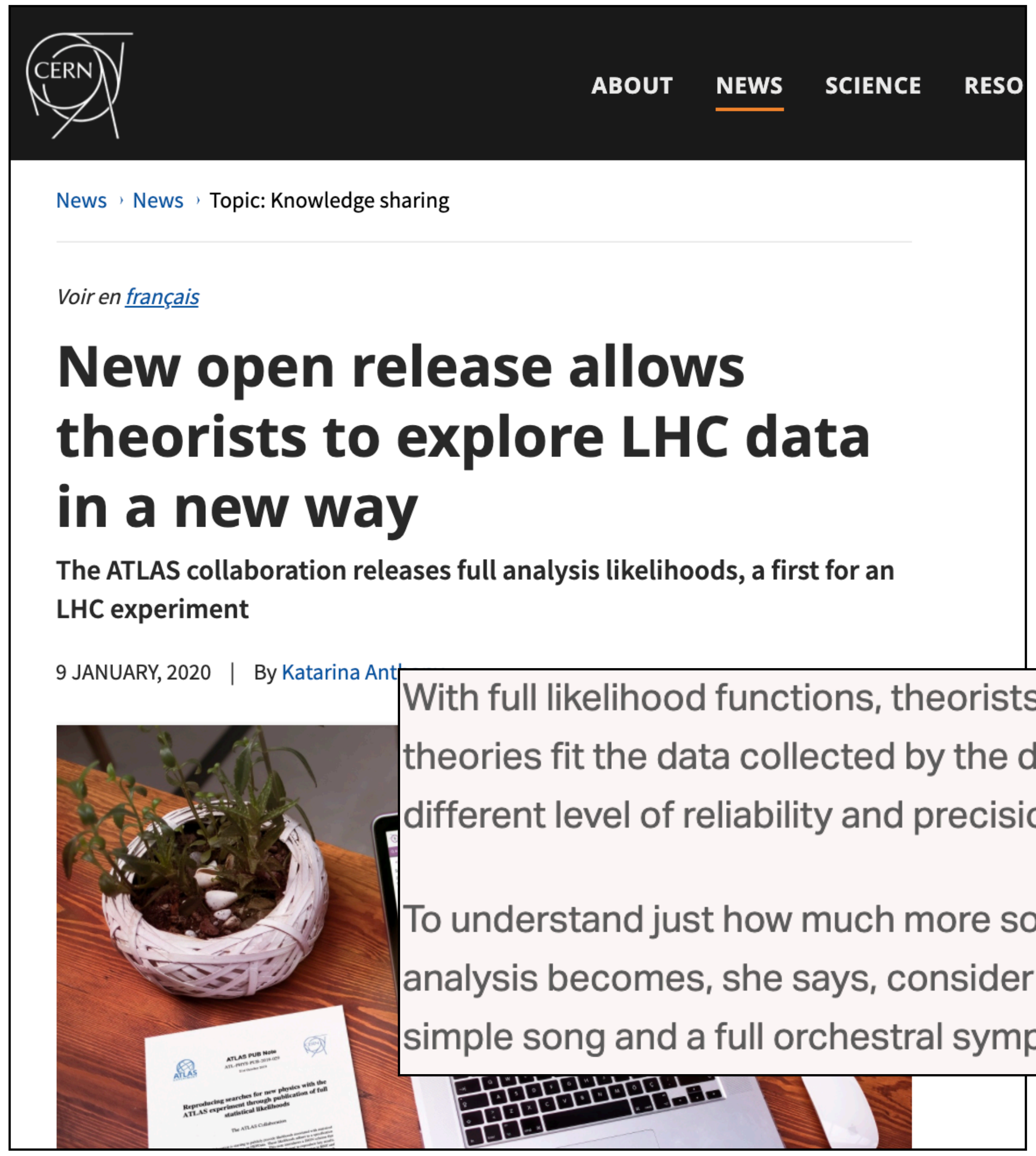
# Likelihood Preservation in ATLAS

**ATLAS first experiment to publish full likelihood**

- **stored on public archive as citable data product: HepData DOI**
- **continuation of long-term effort (cf. simplified l'hoods in CMS)** [N. Wardle et al]
- **best information we have even within experiment**
  **full set of systematics → global fits / combinations / ...**




```
$> pyhf contrib download https://doi.org/10.17182/hepdata.90607.v3/r3 lhood
$> pyhf cls lhood/BkgOnly.json \
   -p <(pyhf patchset extract --name C1N2_Wh_hbb_1000_0 lhood/patchset.json)
{
    "CLs_exp": [
        0.049881095057455506,
        0.12644782067016755,
        0.29258025746896477,
        0.5694144743318884,
        0.8475964832107026
    ],
    "CLs_obs": 0.5856790659557595
}
```

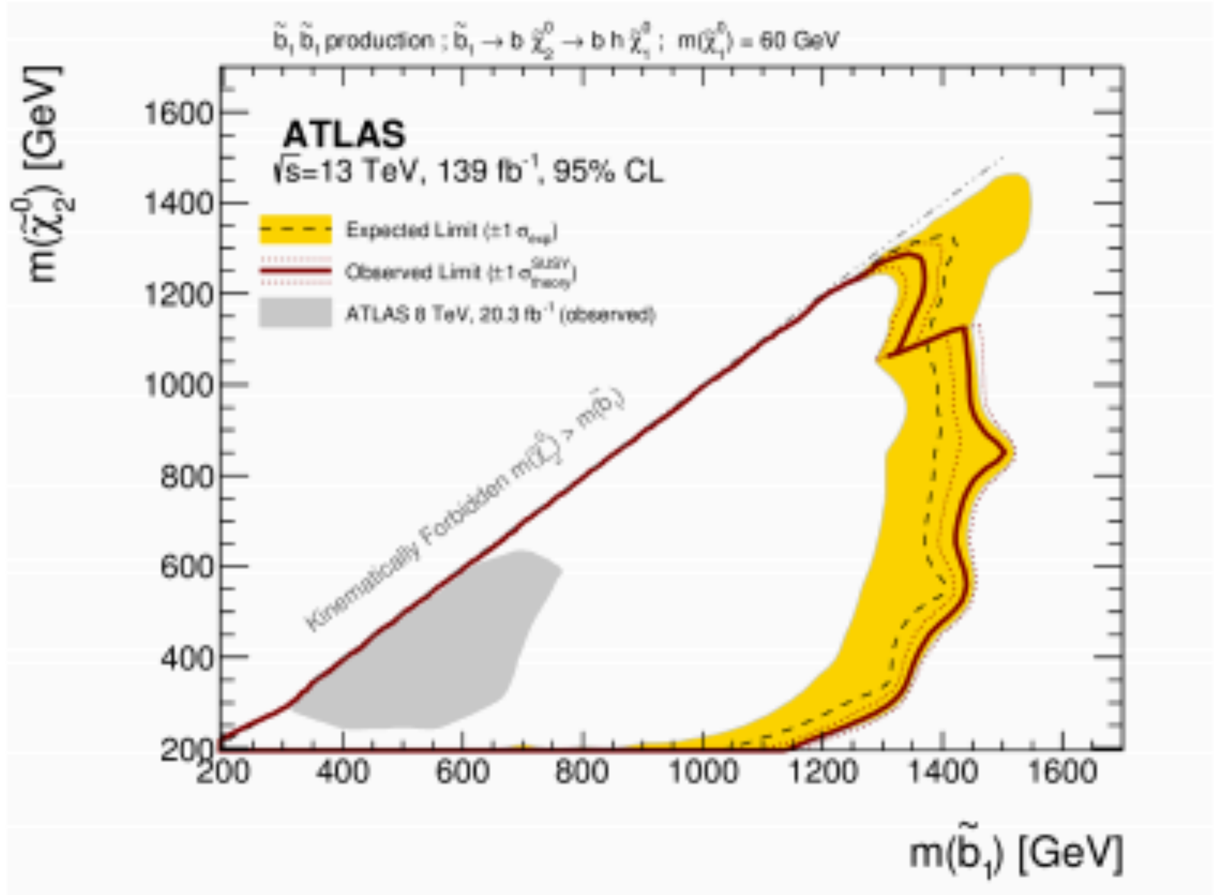



With full likelihood functions, theorists can calculate how well their theories fit the data collected by the detector “at a completely different level of reliability and precision," says Kraml.

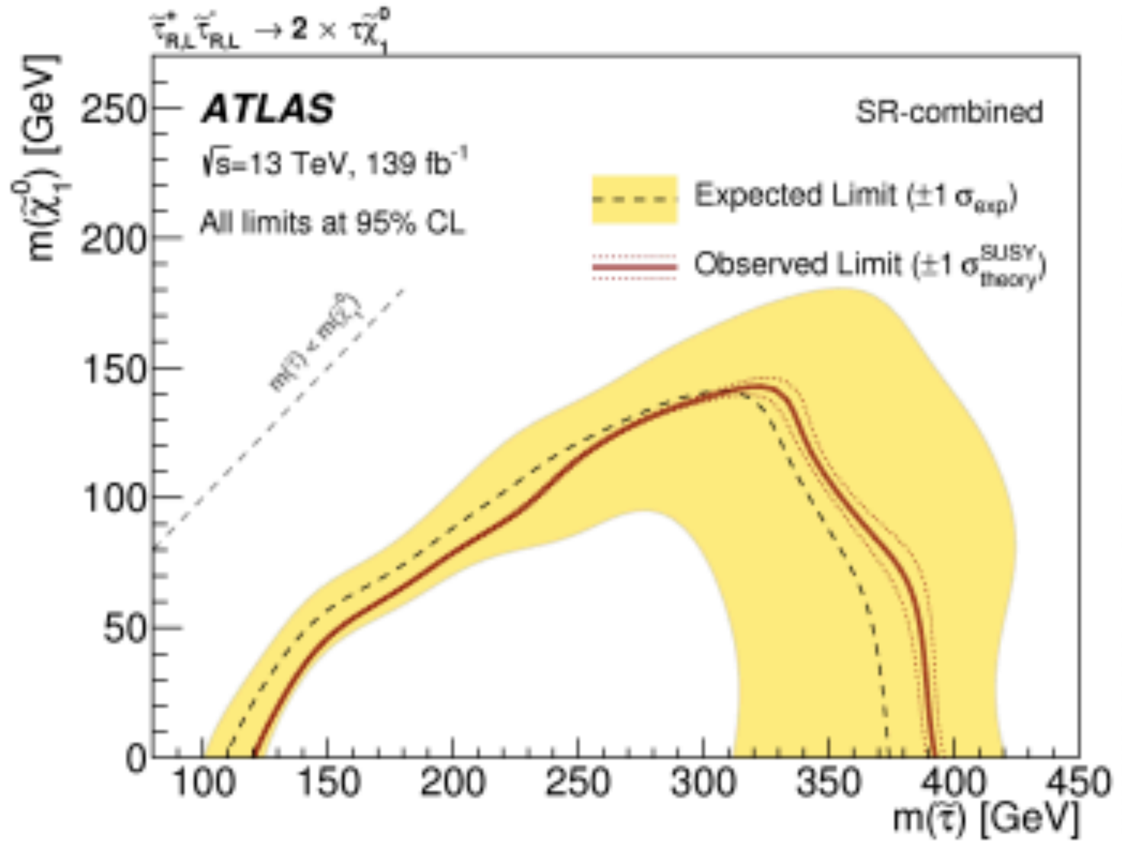
To understand just how much more sophisticated and complex the analysis becomes, she says, consider the difference between a simple song and a full orchestral symphony.

[Symmetry Mag]

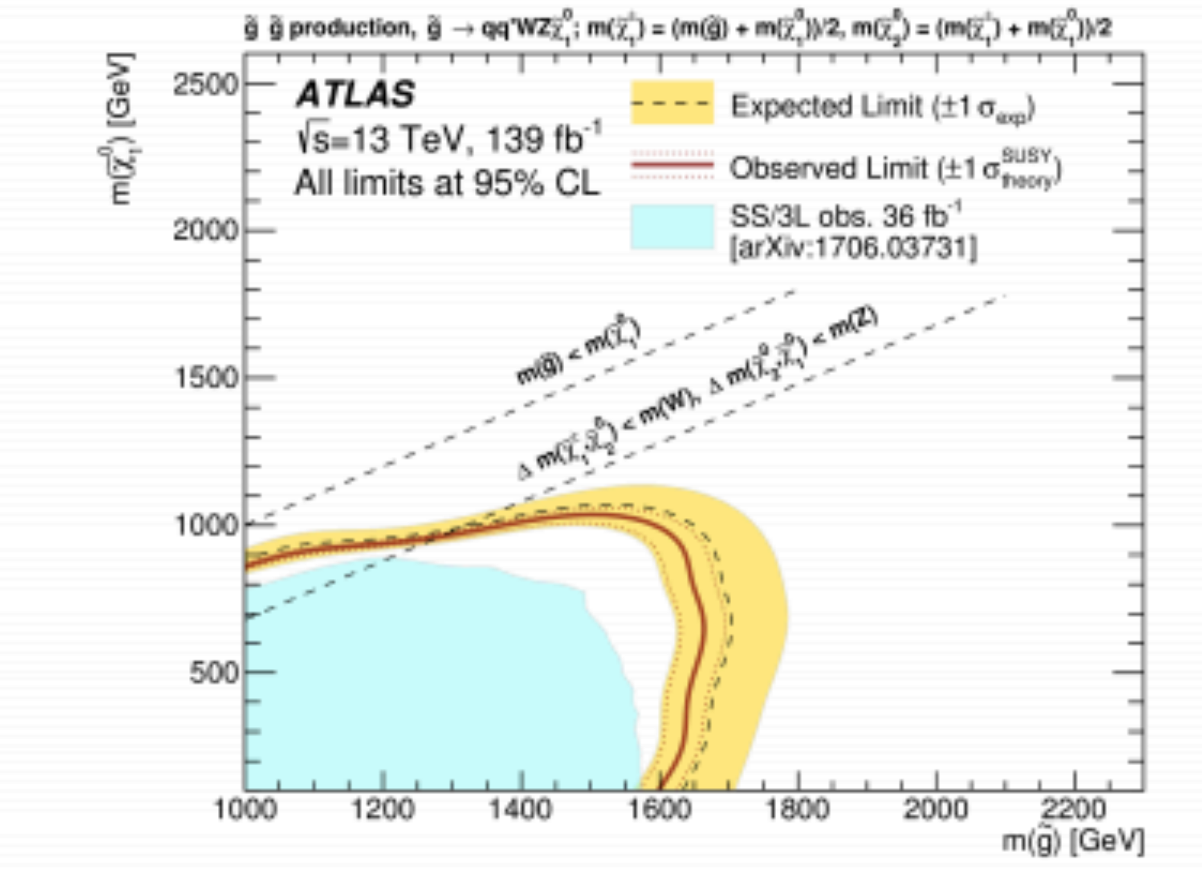
DOI 10.17182/hepdata.89408.v2



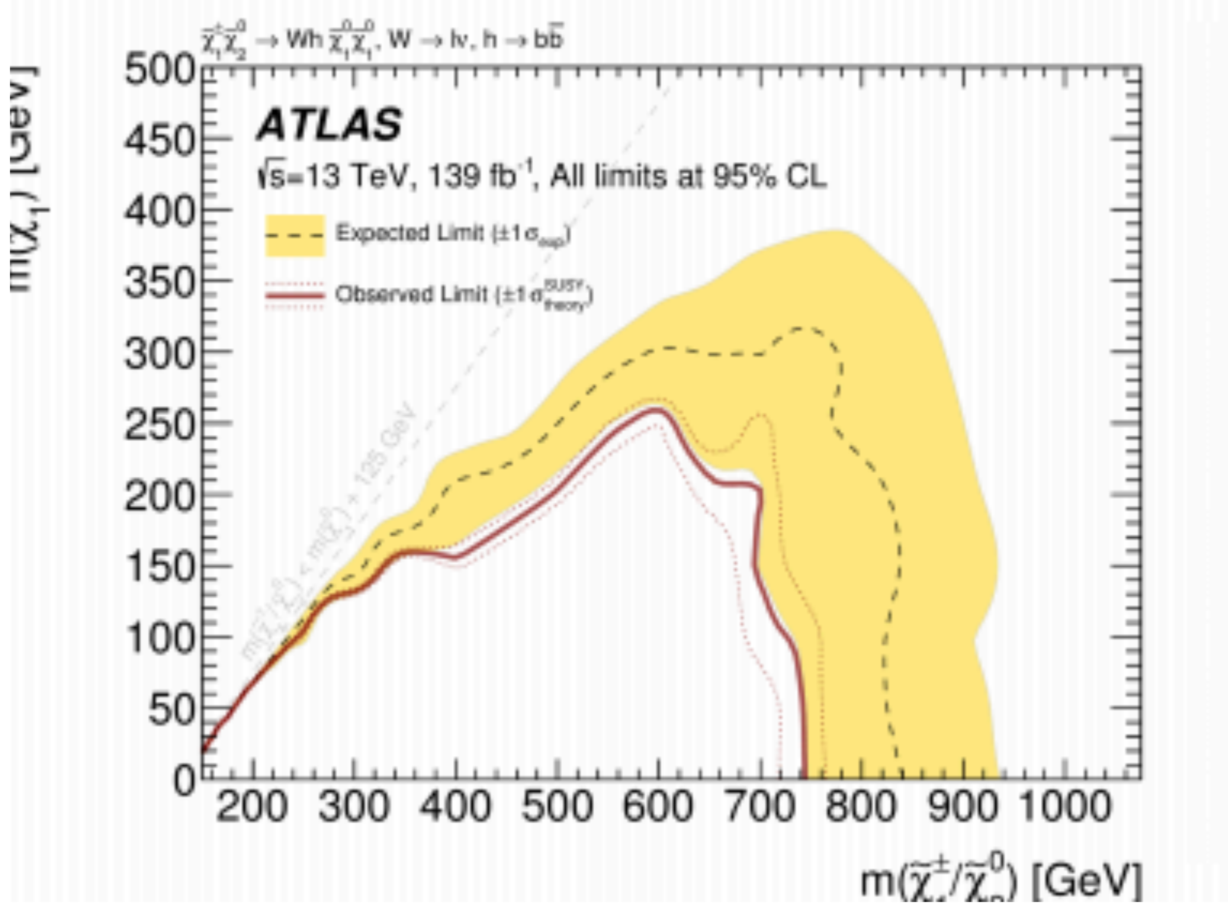

DOI 10.17182/hepdata.92006.v2



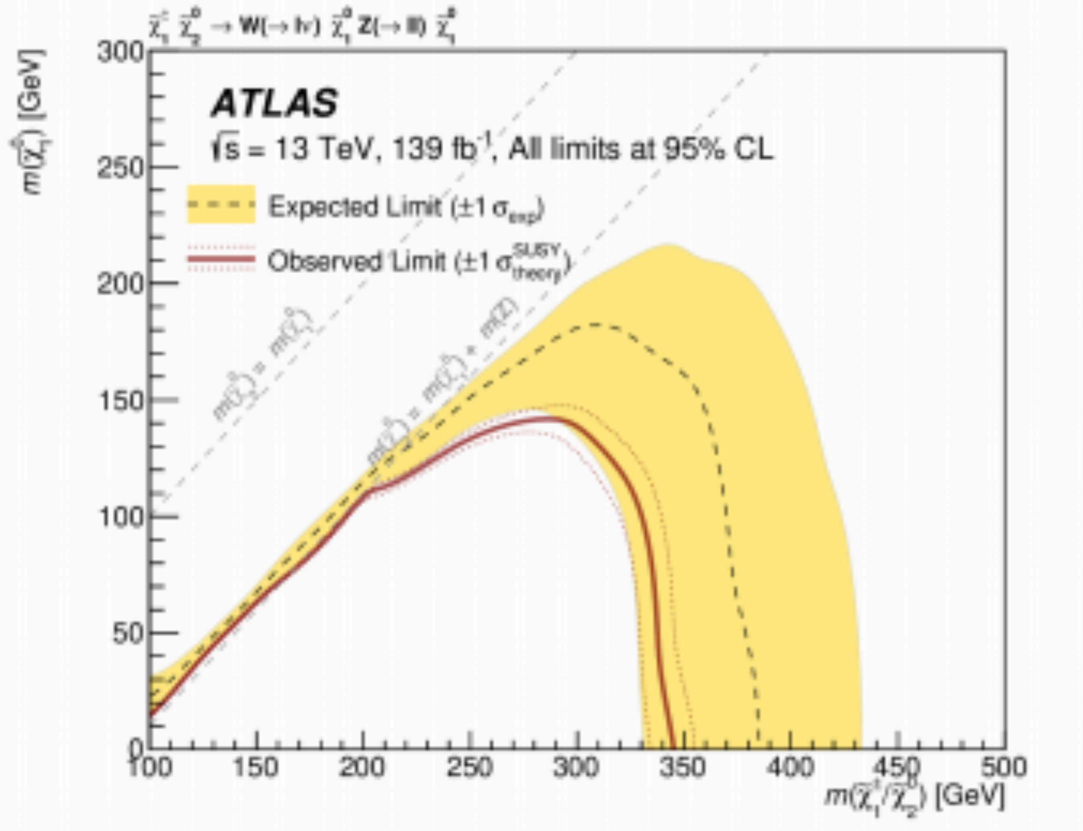

DOI 10.17182/hepdata.91214.v3



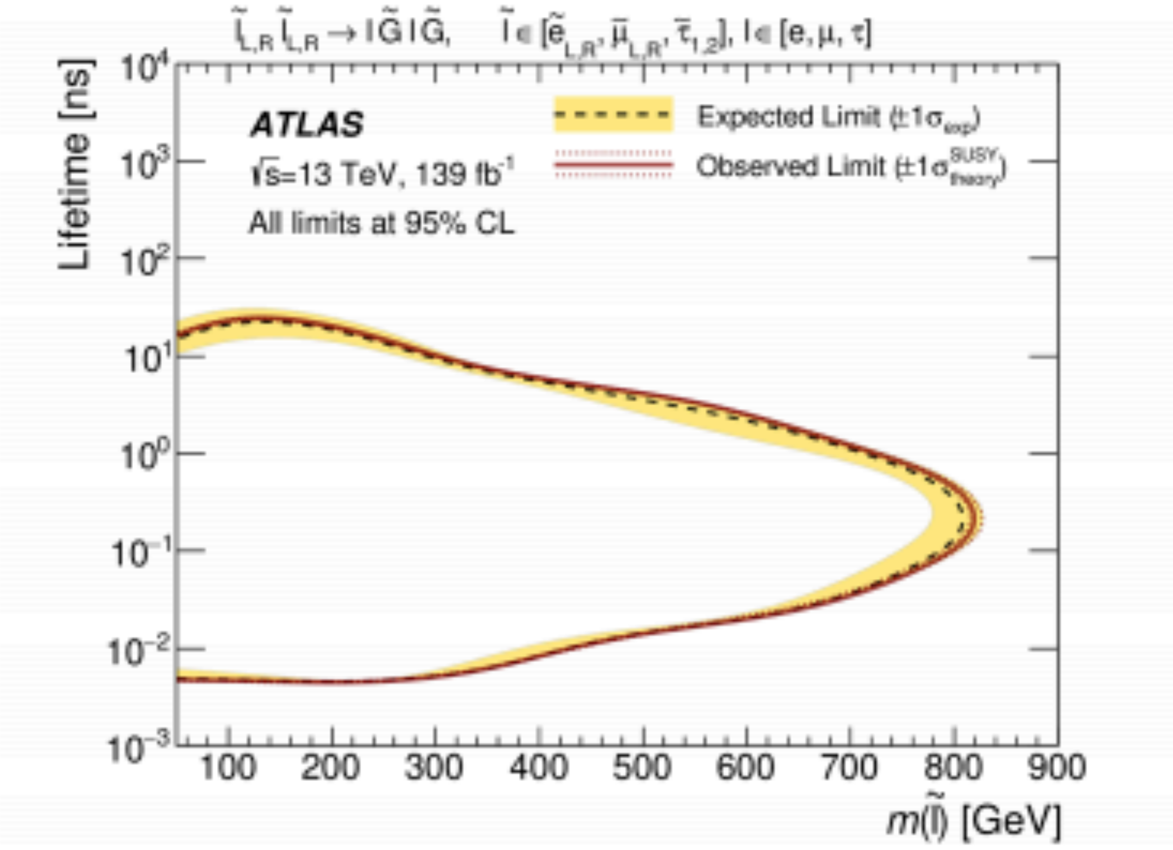

DOI 10.17182/hepdata.90607.v3



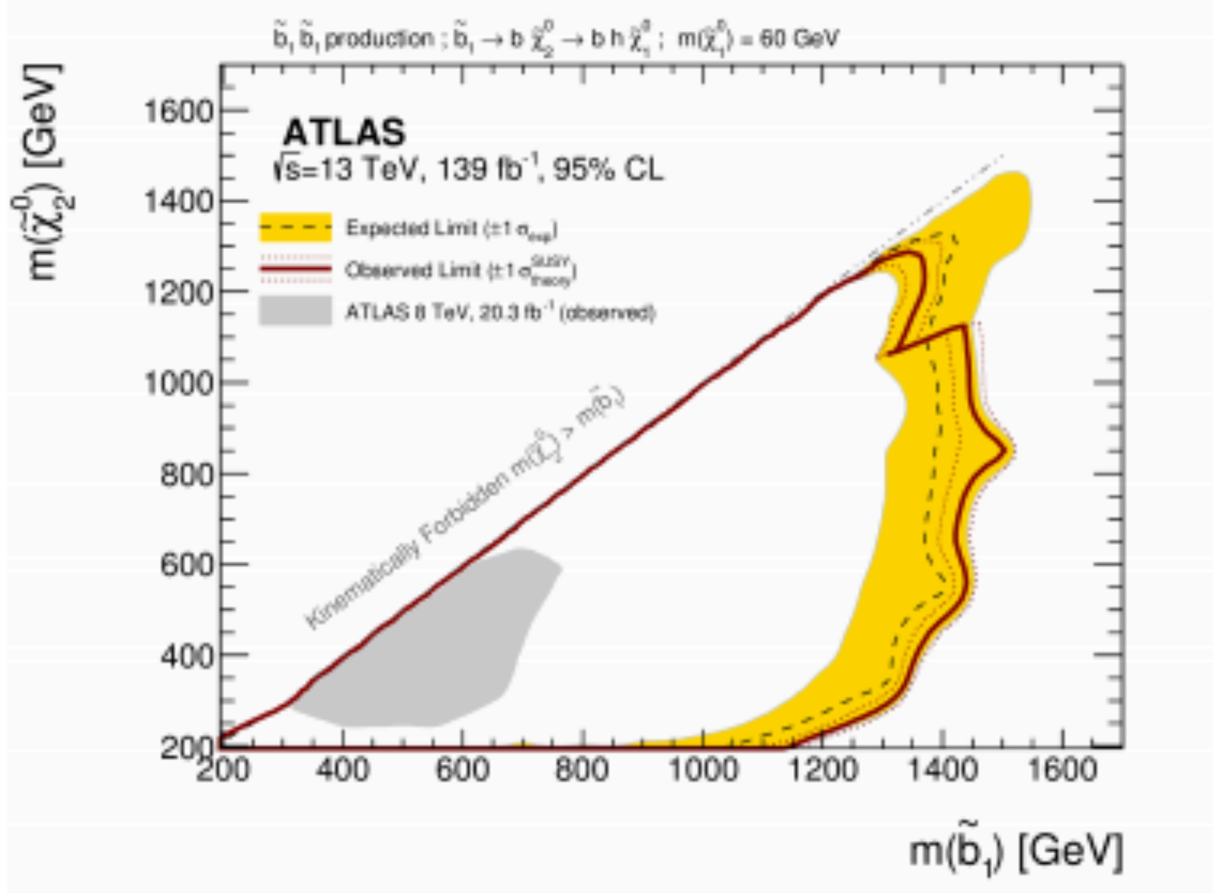

DOI 10.17182/hepdata.91127.v2




DOI 10.17182/hepdata.98796.v2



DOI 10.17182/hepdata.89408.v2







# Let's reproduce this result!

## Two separate measurements:

1. 3 Signal region + 6 Control regions
   59 source of system. uncertainties
   8 Source of Background events

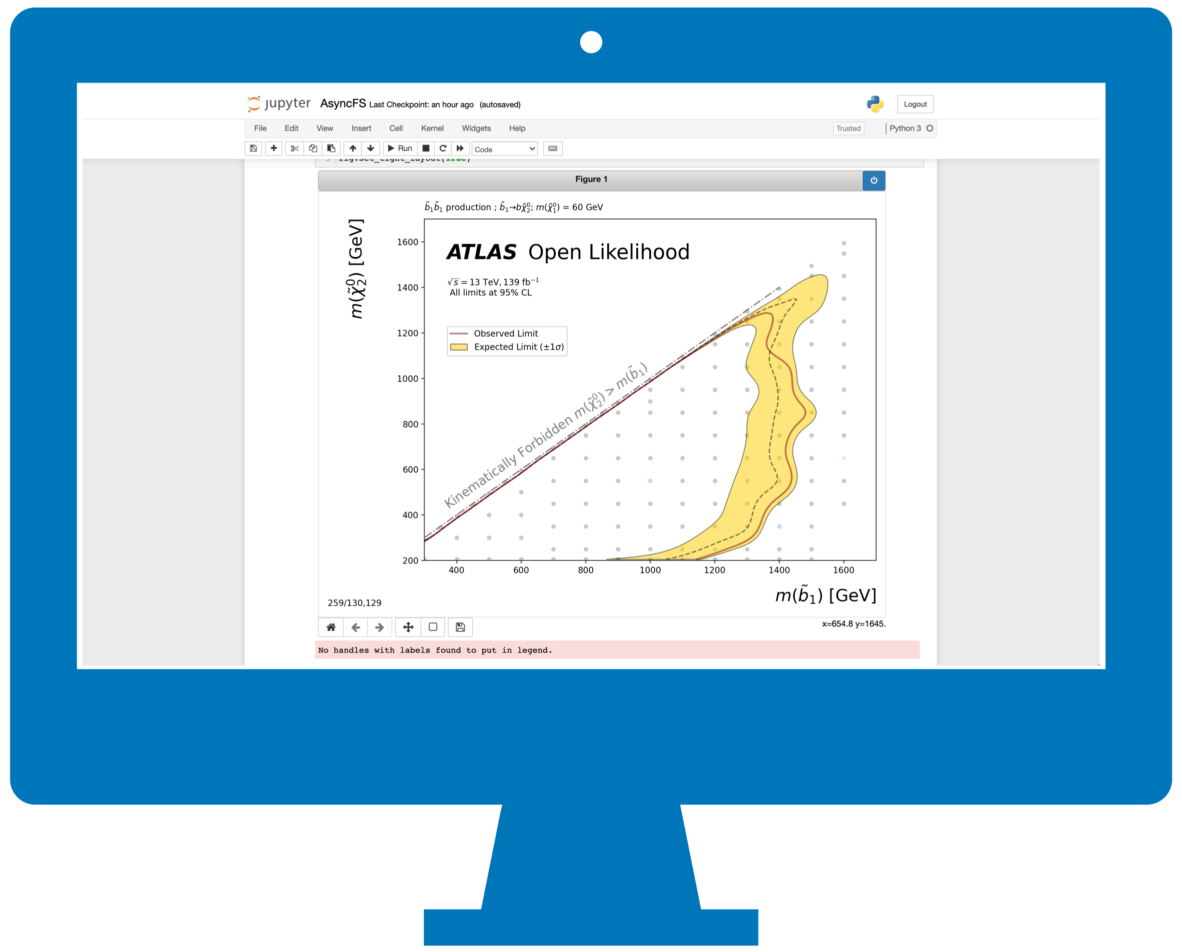
2. 4 Signal region + 5 Control regions
   66 source of system. uncertainties
   8 Source of Background events

**Theory: 130 SUSY models**

**Measurements overlap:**
**take most sensitive result**

# Demo: reproduce full contour (260 CLs) for simplified SUSY model using cloud APIs

- **auto-scaling, distributed statistics**




User Session

$$L(\theta, \theta_{\text{nuis}}) = p(x_{\text{obs}} \,|\, \mu s(\theta) + b(\theta_{\text{nuis}}))$$

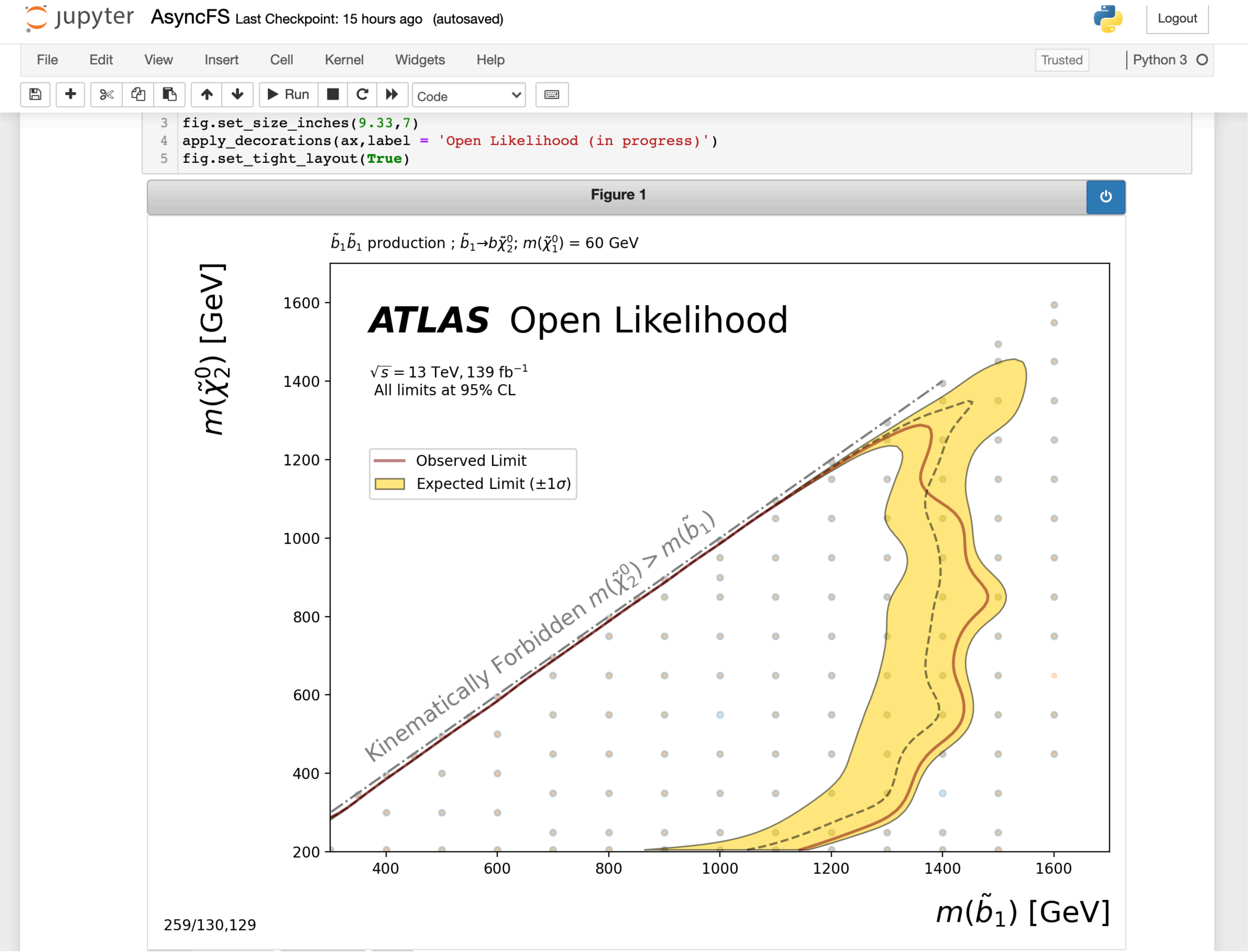
send likelihood

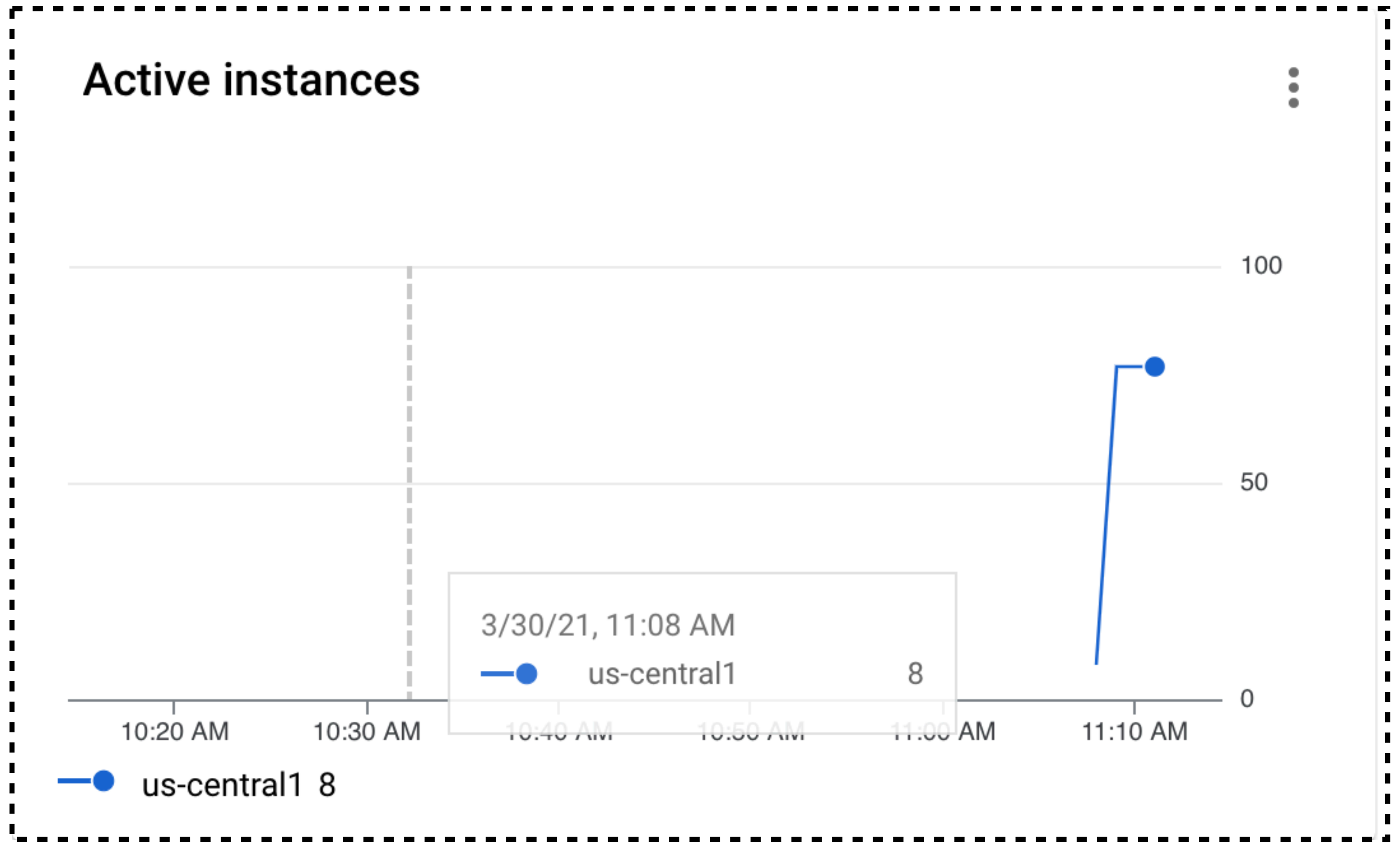
get results

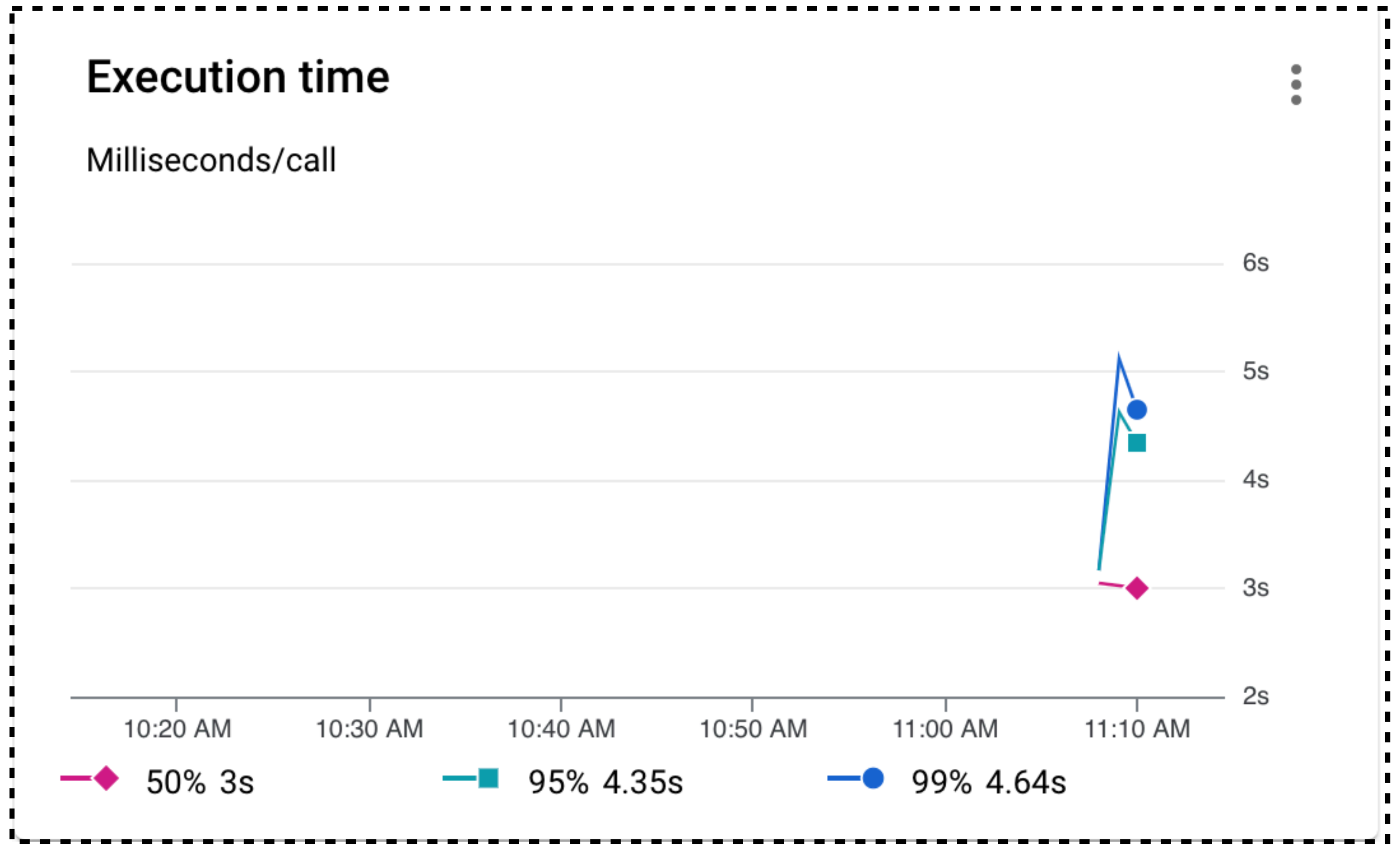
$$CL_s(\theta)$$


Cloud Fitting Service

# Demo


jupyter AsyncFS Last Checkpoint: 15 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
fig.set_size_inches(9.33,7)
apply_decorations(ax,label = 'Open Likelihood (in progress)')
fig.set_tight_layout(True)
Figure 1
$\tilde{b}_1\tilde{b}_1$ production ; $\tilde{b}_1 \to b\tilde{\chi}^0_2$; $m(\tilde{\chi}^0_1)$ = 60 GeV
ATLAS Open Likelihood
$\sqrt{s}$ = 13 TeV, 139 fb$^{-1}$
All limits at 95% CL
Observed Limit
Expected Limit ($\pm 1\sigma$)
Kinematically Forbidden $m(\tilde{\chi}^0_2) > m(\tilde{b}_1)$
$m(\tilde{\chi}^0_2)$ [GeV]
$m(\tilde{b}_1)$ [GeV]
259/130,129


Active instances
3/30/21, 11:08 AM
us-central1 8
10:20 AM
10:30 AM
11:10 AM
100
50
0
us-central1 8


Execution time
Milliseconds/call
6s
5s
4s
3s
2s
10:20 AM
10:30 AM
10:40 AM
10:50 AM
11:00 AM
11:10 AM
50% 3s
95% 4.35s
99% 4.64s

(Cost: 0.05$)

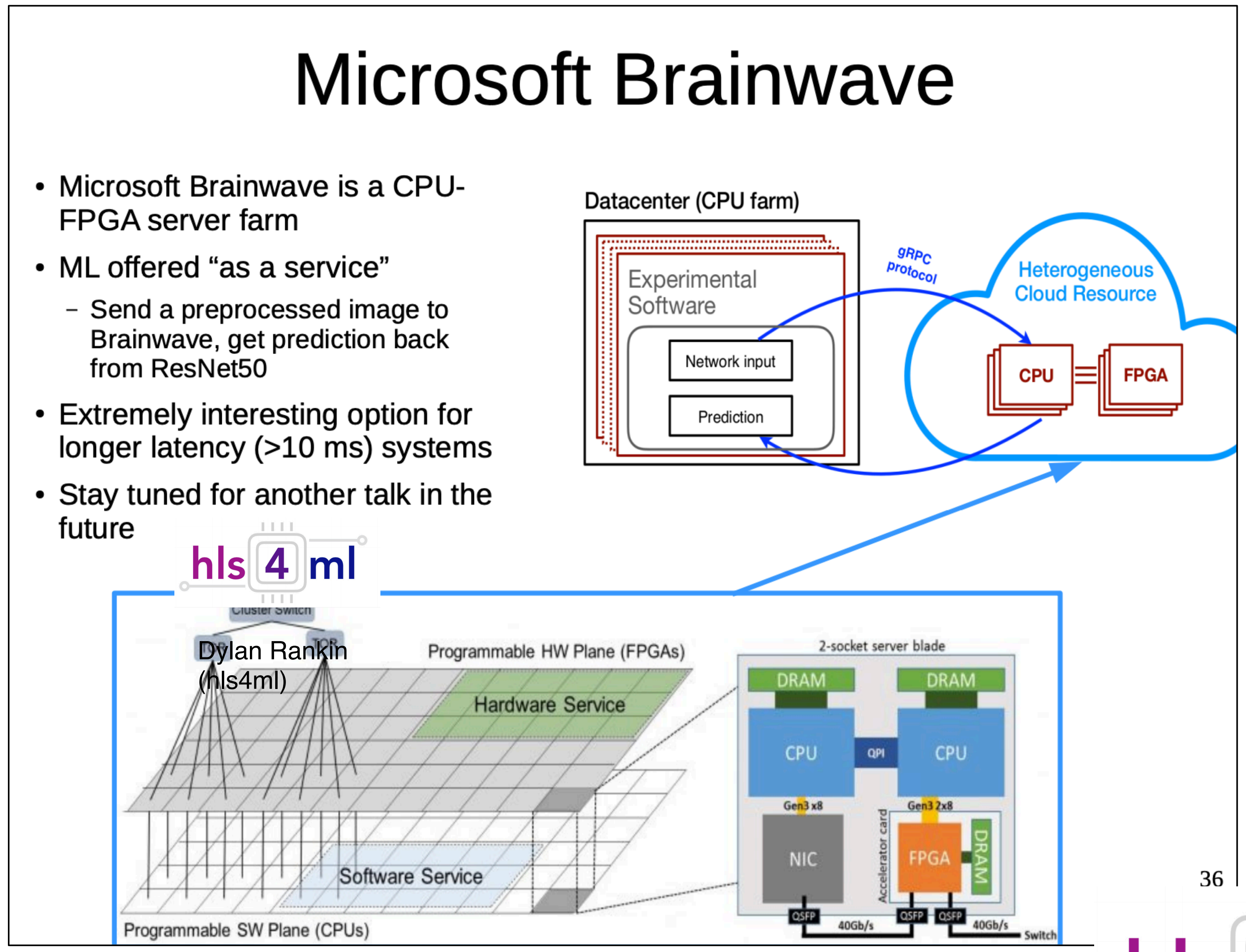
# pyhf - with functions as a service

## As with Higgs example:

- Open Data provides access to LHC data: **but now a more refined product.**
- can offload heavy computation to cloud - cost only per CPUh
  scale to hundreds of (private) cores within seconds
- Here: "Fitting-as-a-service": user only needs a URL
  can optimize hardware for stat. analysis.

## Related: ML-as-service

- Investigated for fast FPGA-based inference (Trigger, Reco, ...)




[J.Ngadiuba, D. Rankin]

# Hold on, why so restrictive?

**(just reproducing things is boring)**

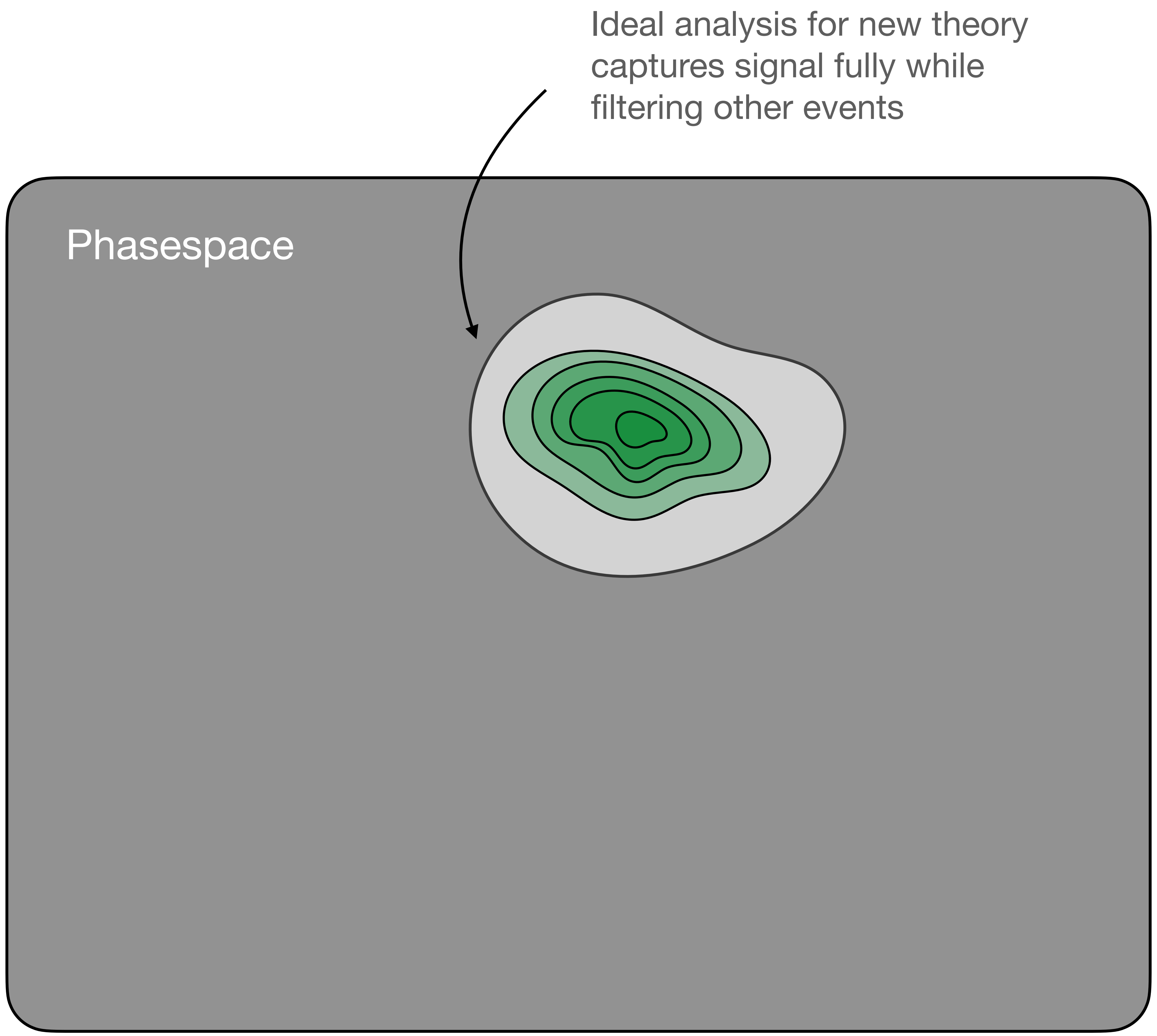
# Reinterpretation

Did we over-correct from Open Data? Open Likelihood is great for

- reproducing results
- combination with other measurements wrt. to same theory ("global fits")

→ but not enough to study new theories

# Reinterpretation




**Main Concern creating new analyses internally & externally (w/ Open Data)**

**Human resources for**

- **analysis design**
- **background estimation**
- **systematic analysis**

**Can we maybe reuse the analysis we already have?**

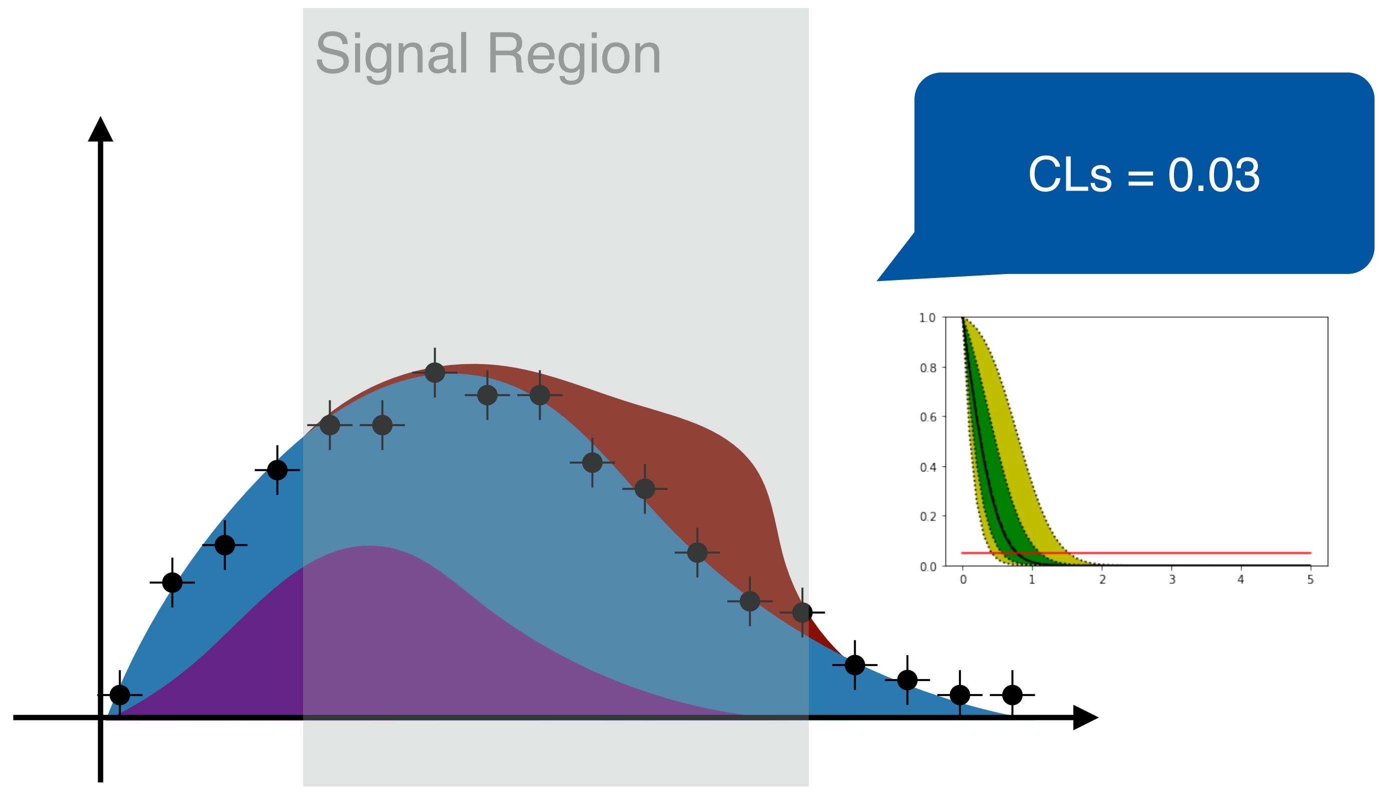
# Reinterpretation

**Increasing coverage of data space as LHC progresses.**

**It's likely that a analysis already did all the hard work for us**

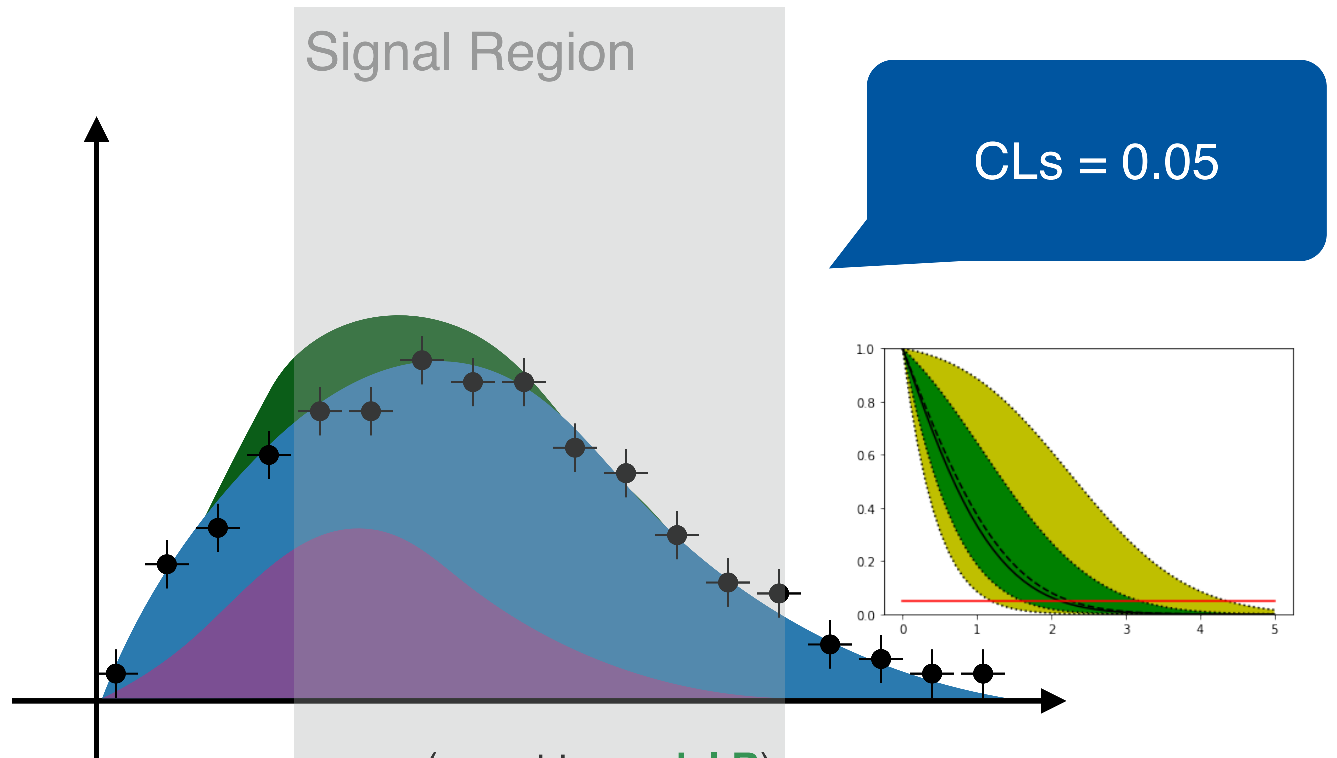
**Not optimized for new theory but can be quite sensitive**

# Reinterpretation



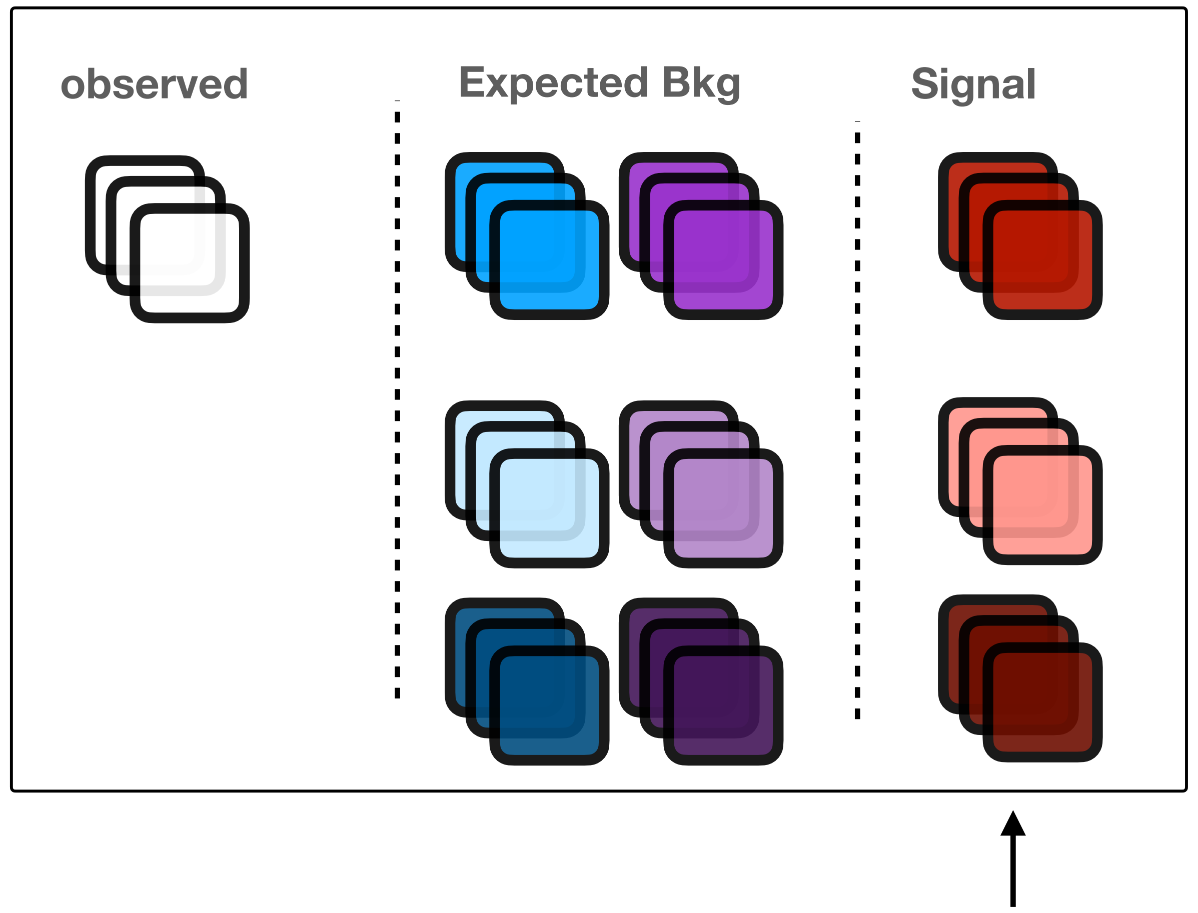

(optimized for **model A**)

RECAST:
- remove old signal
- compute new signal
- inject into likelihood
- run stat. analysis

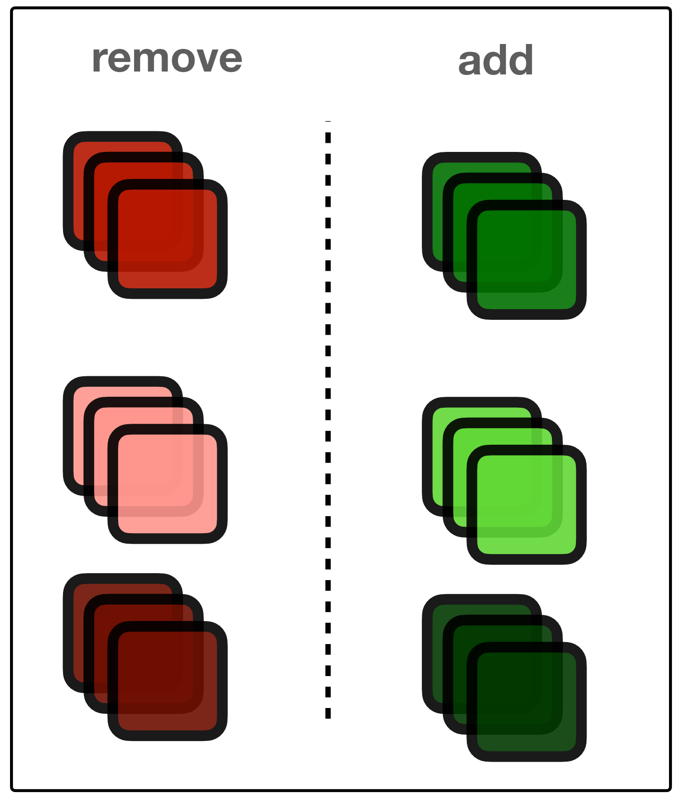



(recast to **model B**)

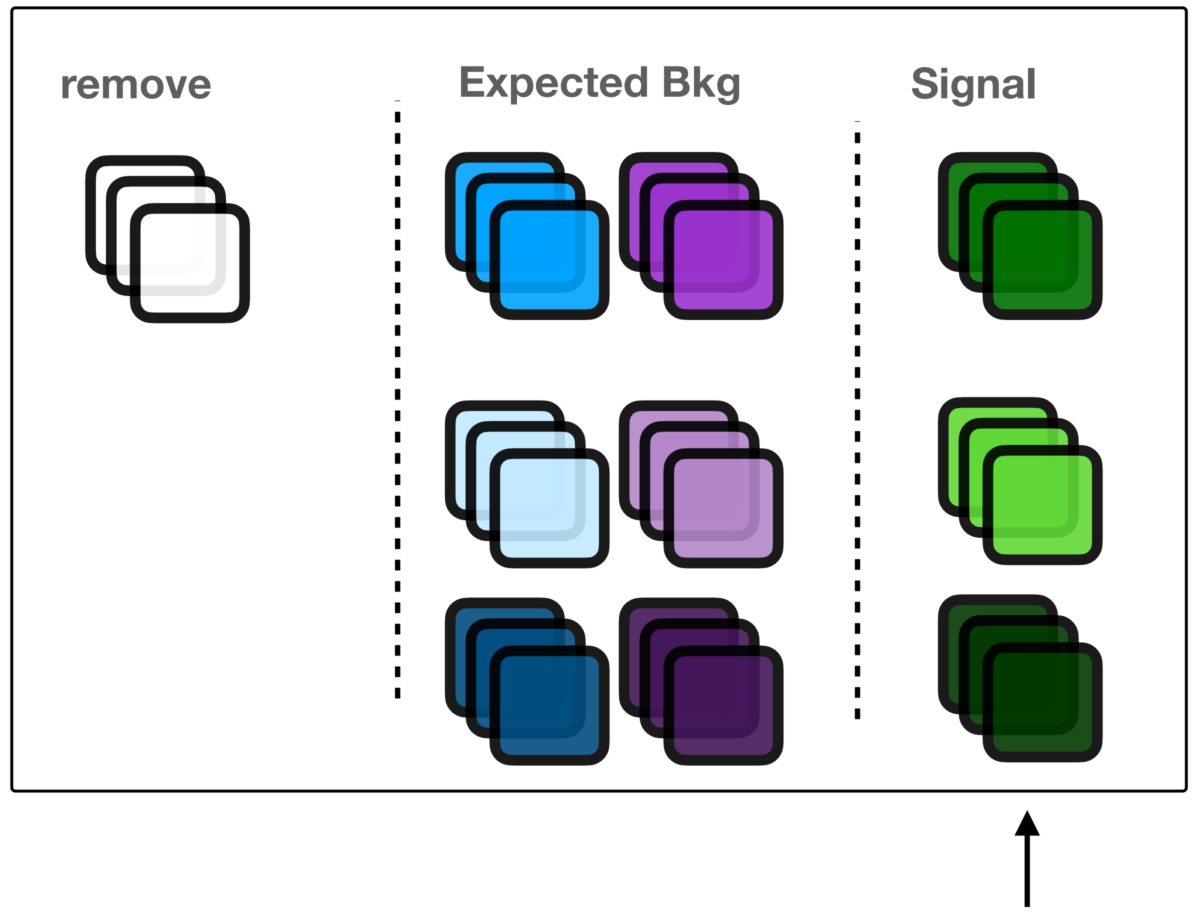
"Original Likelihood"




"Likelihood Patch"
from RECAST



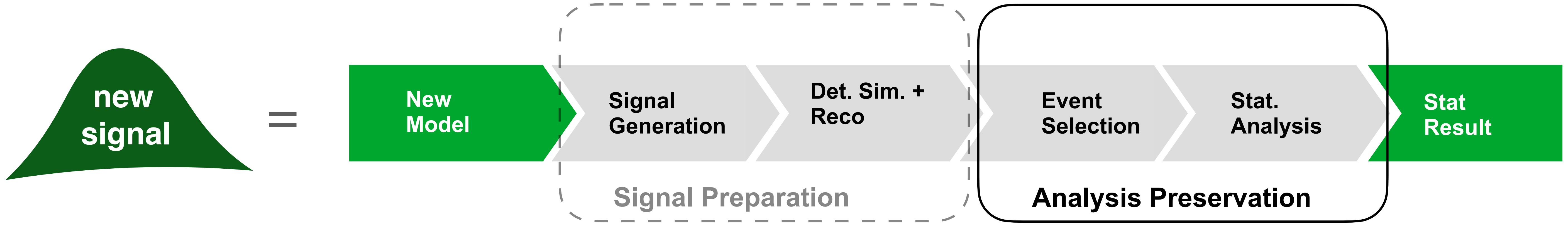

"Recasted Likelihood"

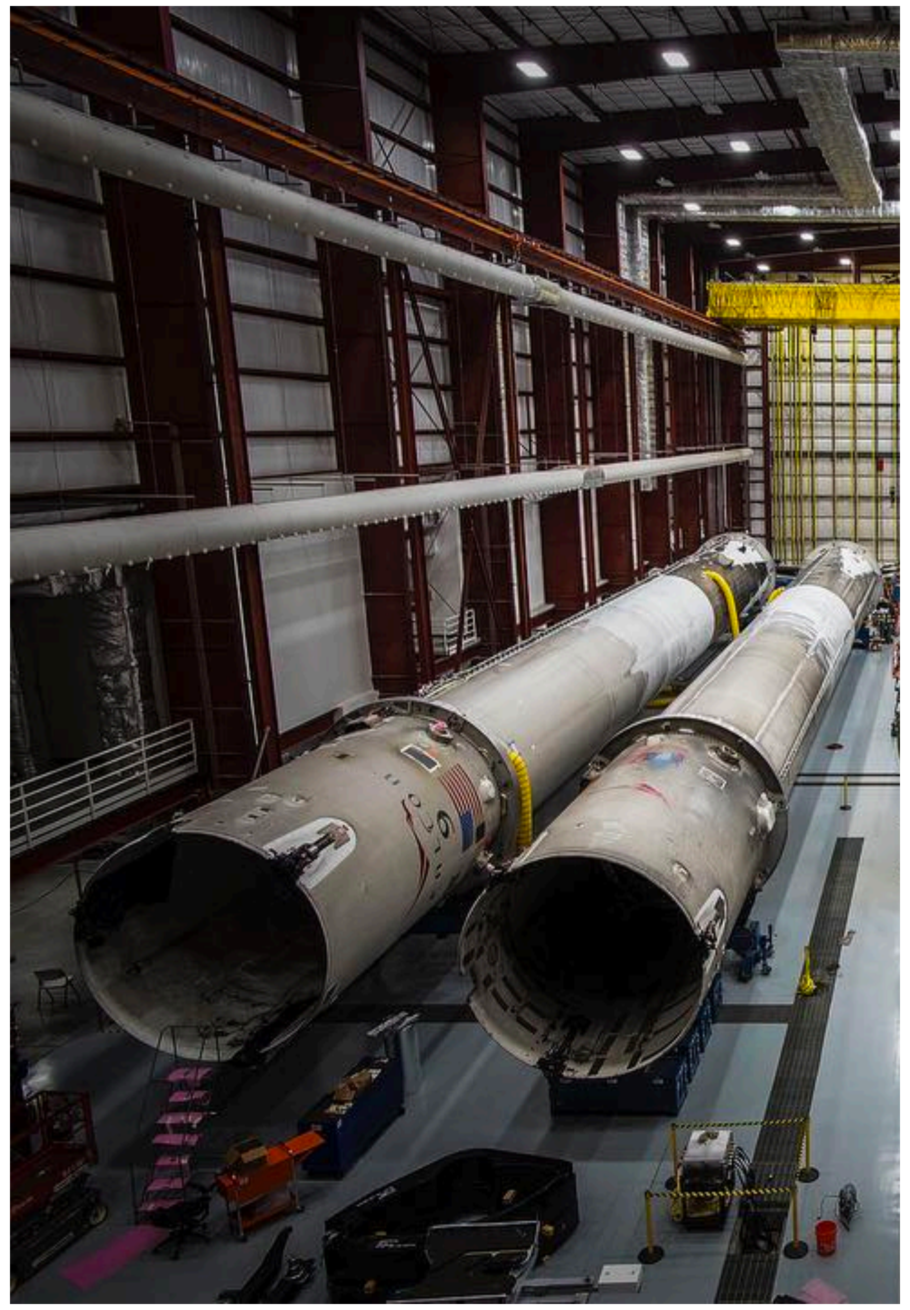


# More Physics through Analysis Preservation

## Simple/Obvious Idea: whats the problem?

- Need to be able to compute new signal component
- Need an archived, but operational version of the analysis
  - ready to be re-run if needed
- "reuse" much more powerful than "reproducibility"

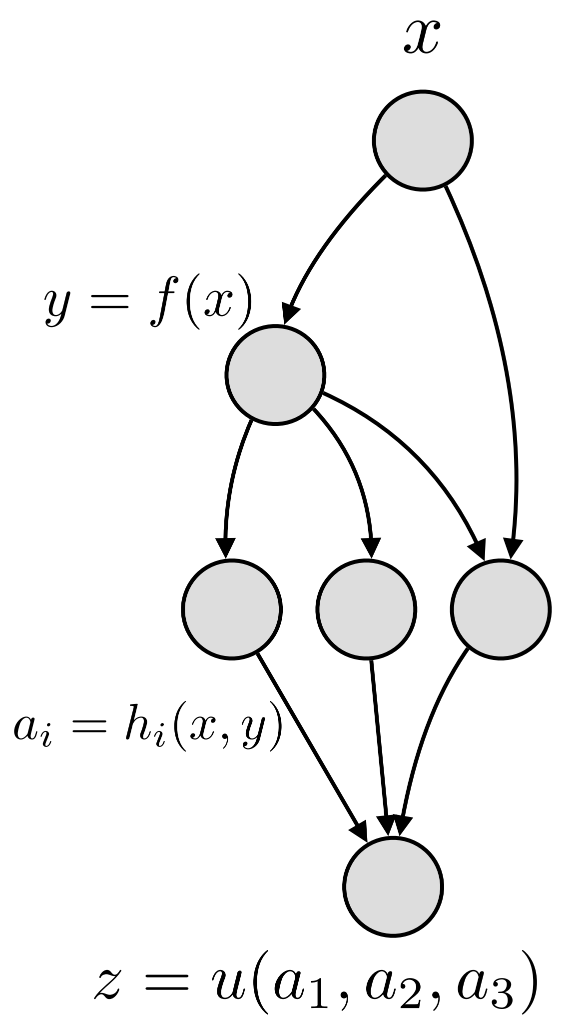


# More Physics through Analysis Preservation

## Analysis Preservation

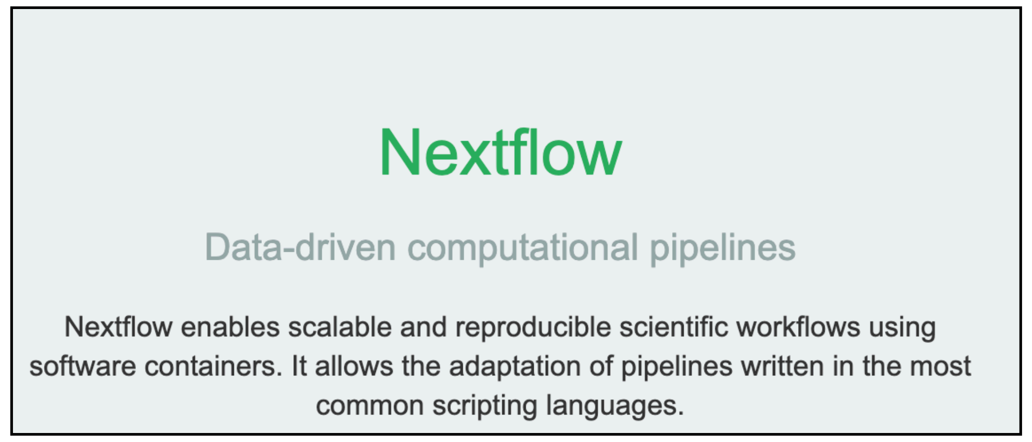
| capture software | capture scripts | capture workflow |
|---|---|---|
| package software itself in system-indep format (all dep.) | templates how to use the software | correct sequence of steps for analysis pipeline |



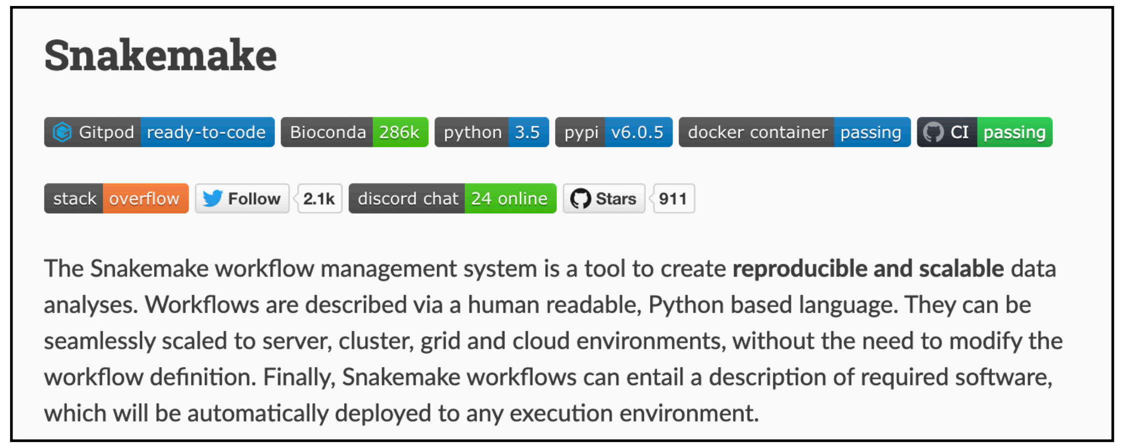

## In ATLAS, we provide common infrastructure to make this easy

- automation for software archiving via containers
- workflow language to capture physics logic
- requirement now for across BSM program

## Similar efforts across many data-intensive sciences:

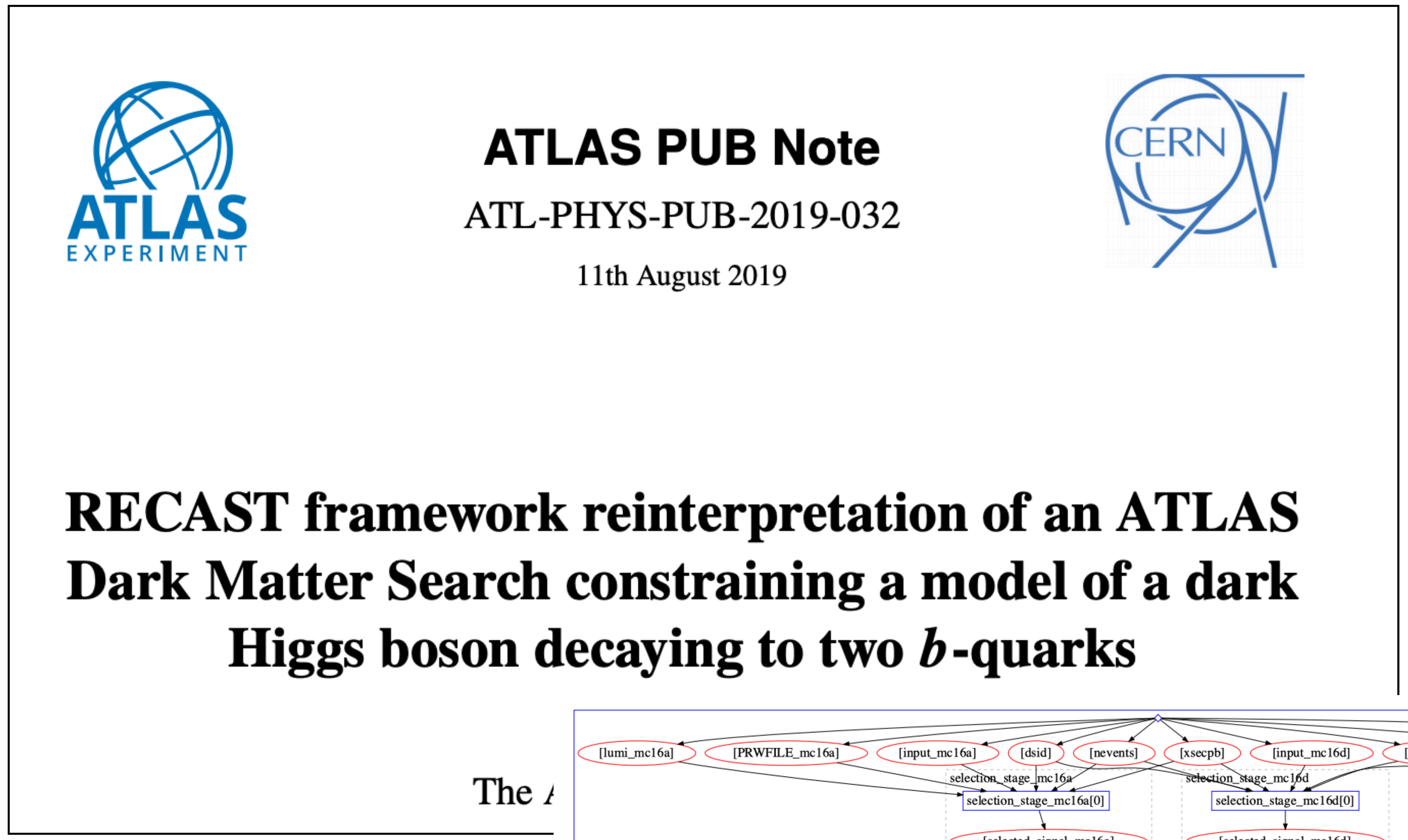




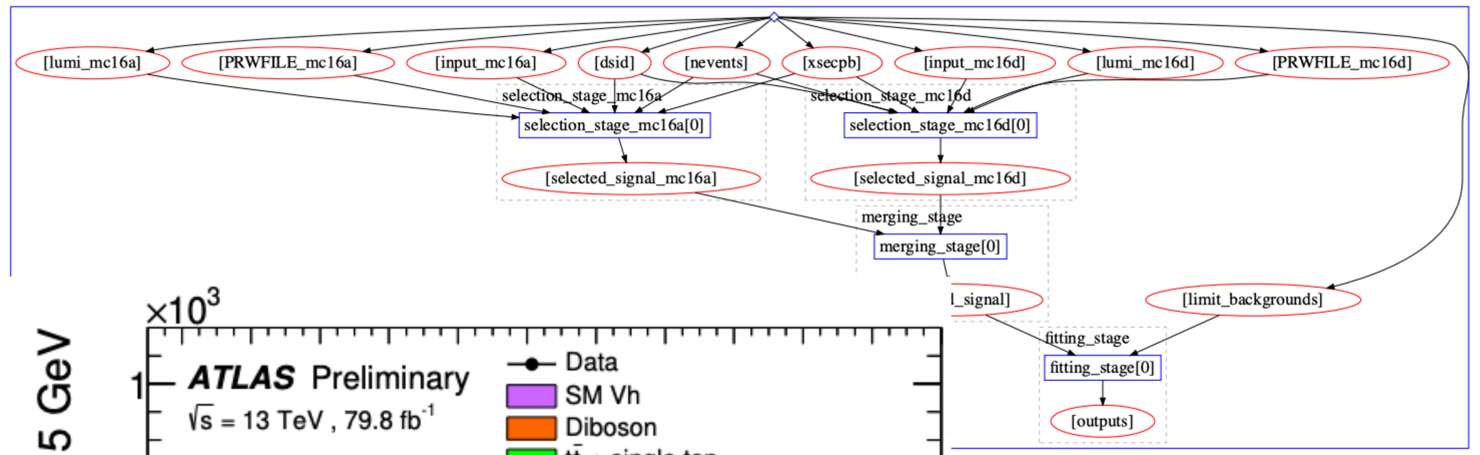


# More Physics through Analysis Preservation

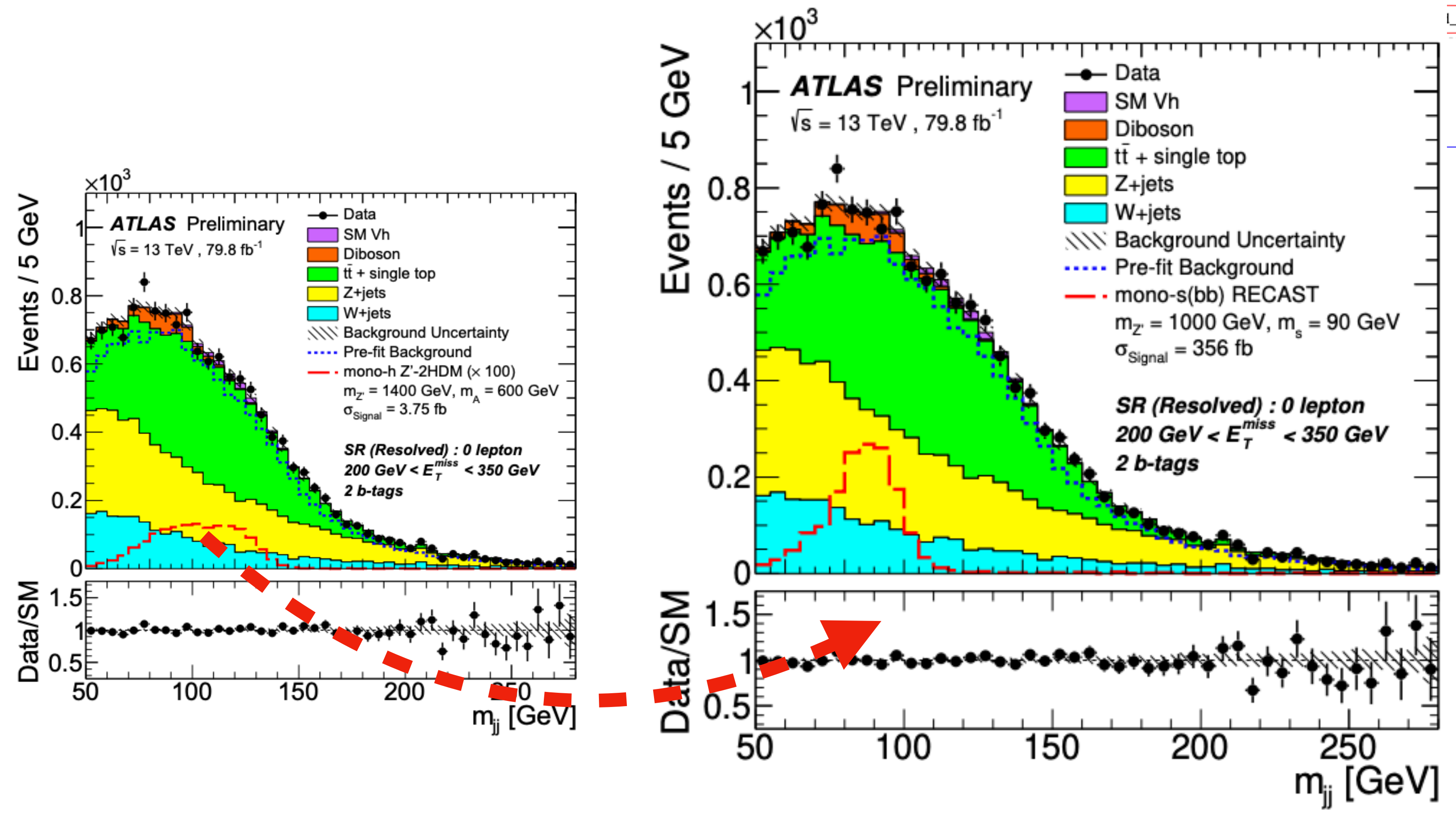
## From Cartoon to reality:

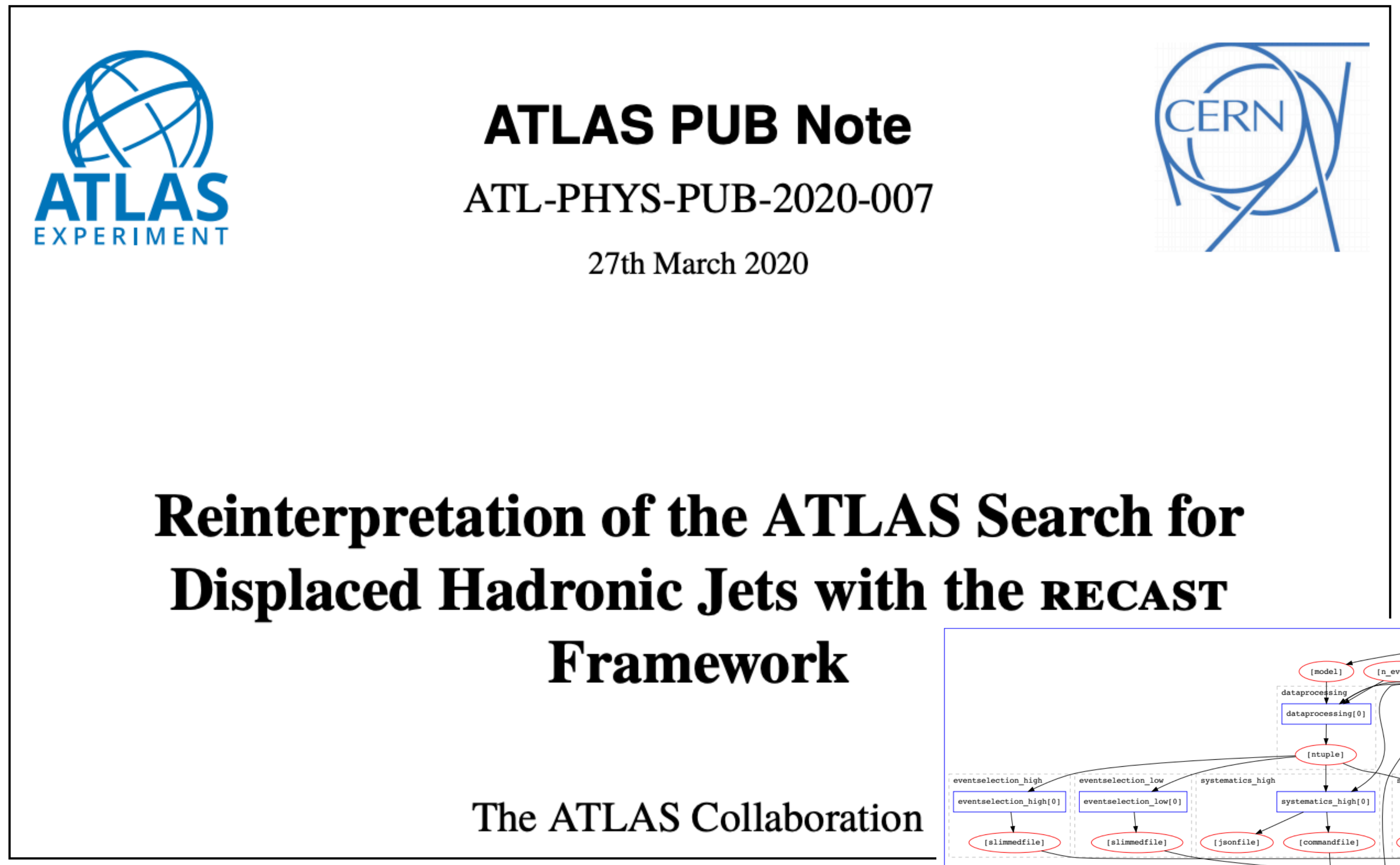

ATLAS PUB Note
ATL-PHYS-PUB-2019-032
11th August 2019

**RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two *b*-quarks**



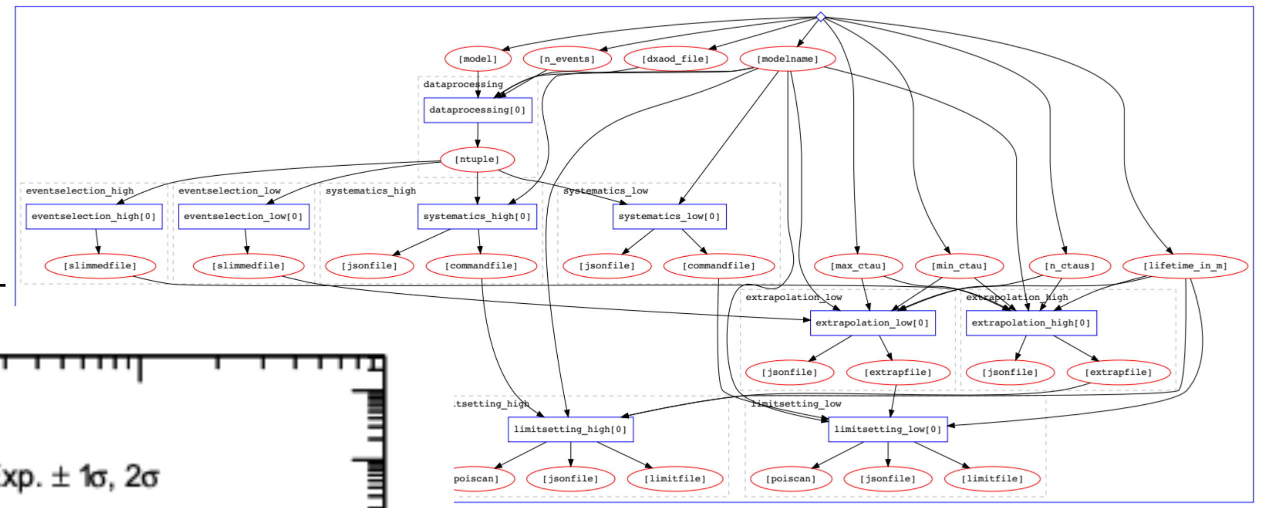
[P. Gadow (DESY) et al]



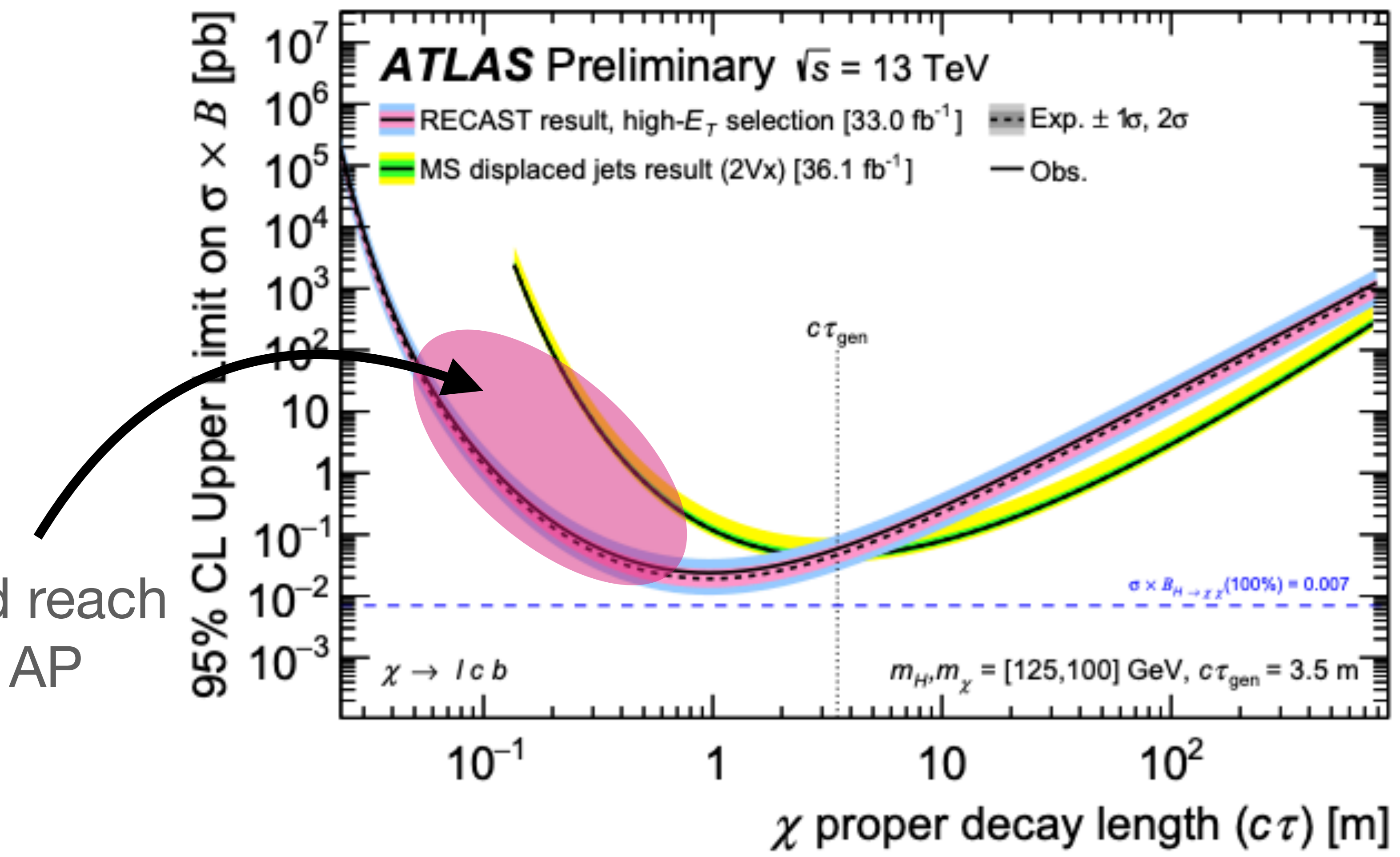


ATLAS PUB Note
ATL-PHYS-PUB-2020-007
27th March 2020

**Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework**

The ATLAS Collaboration



expanded reach thanks to AP

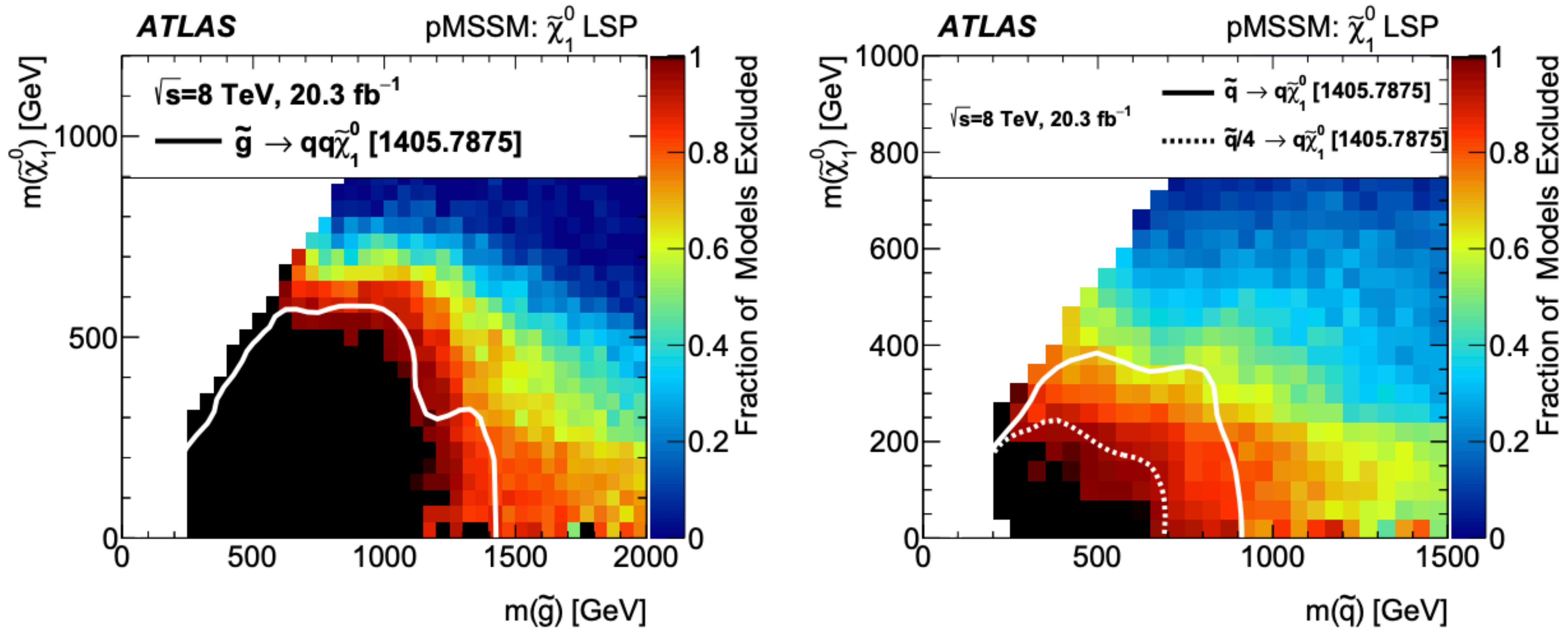


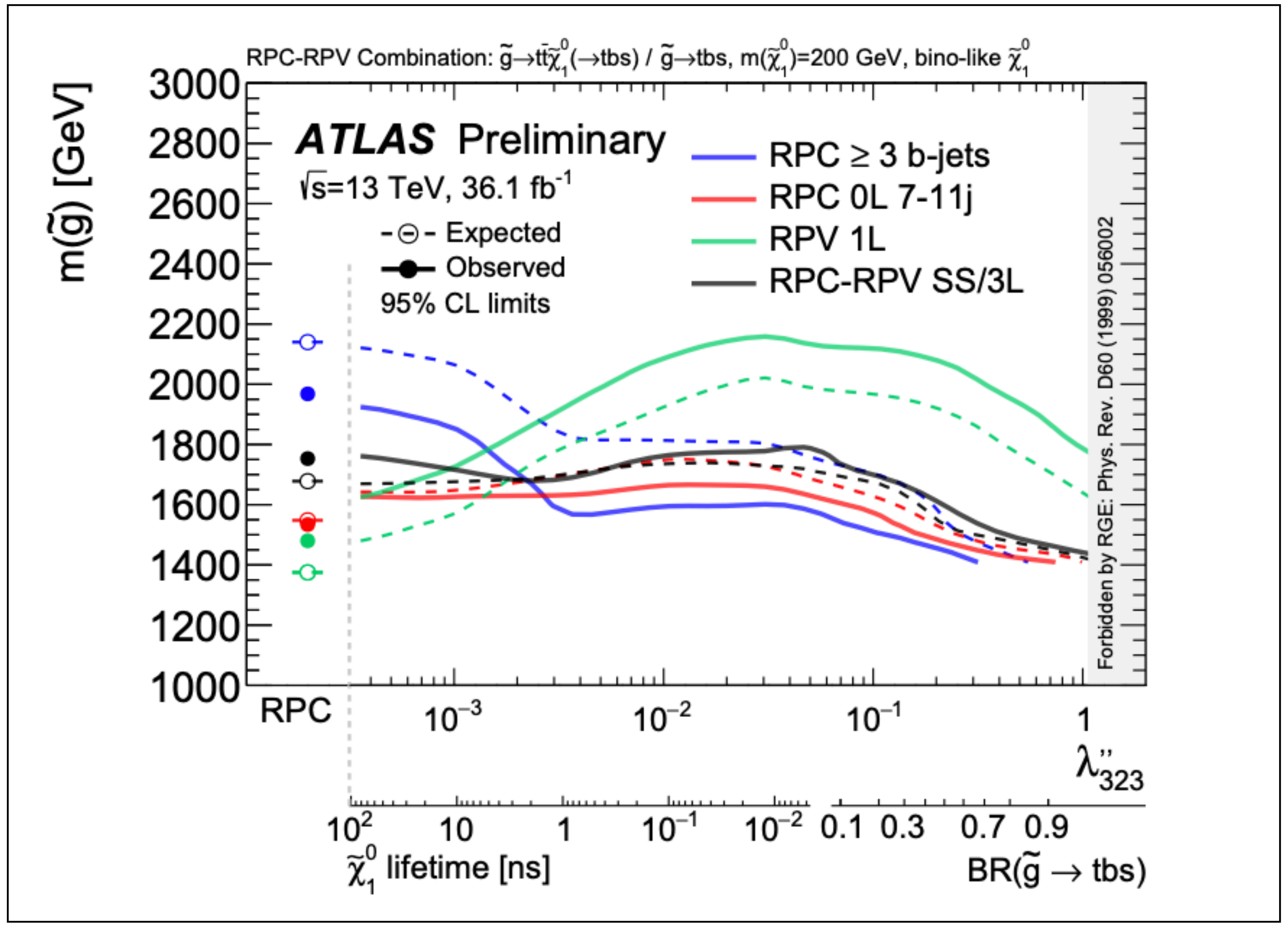
# More Physics through Analysis Preservation

## Used in cross-cutting summary analyses w/ multiple analyses

- **Dark Matter models at the LHC**
- **SUSY reinterpretations**

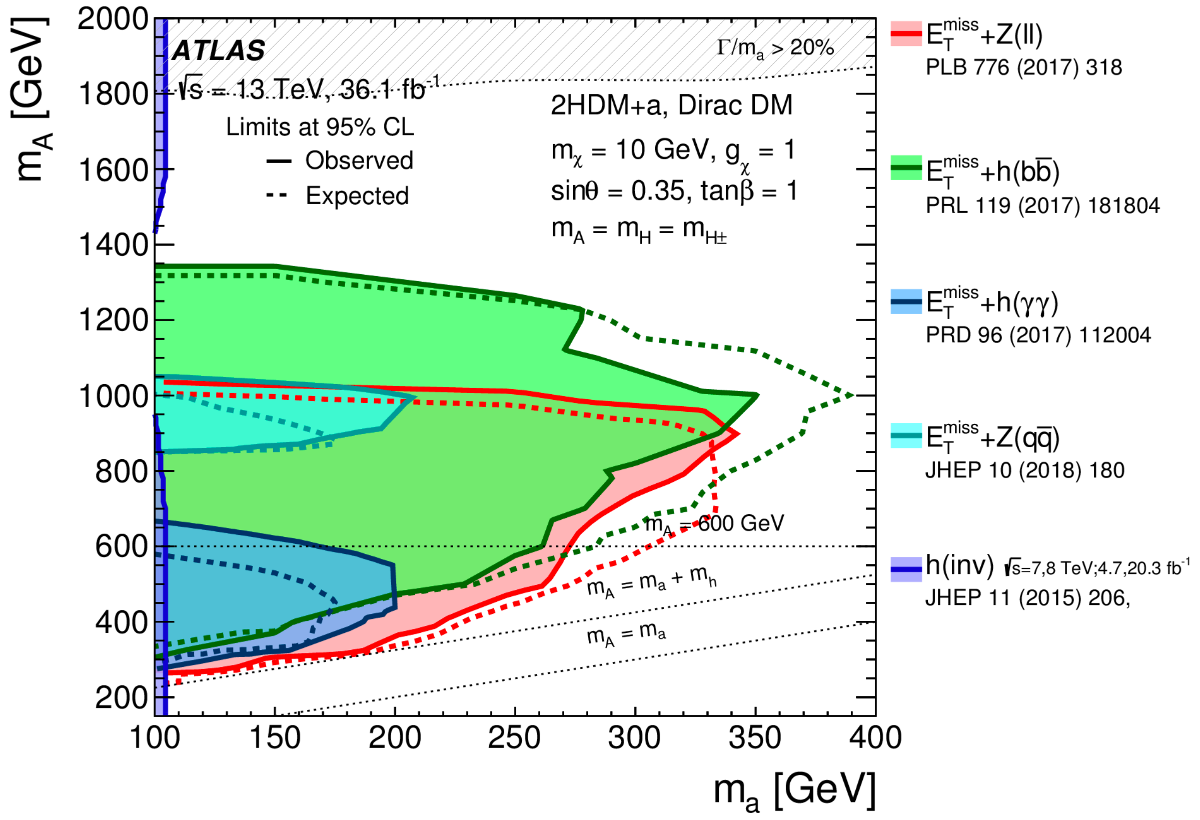
pMSSM arxiv:1508.06608




RPV SUSY

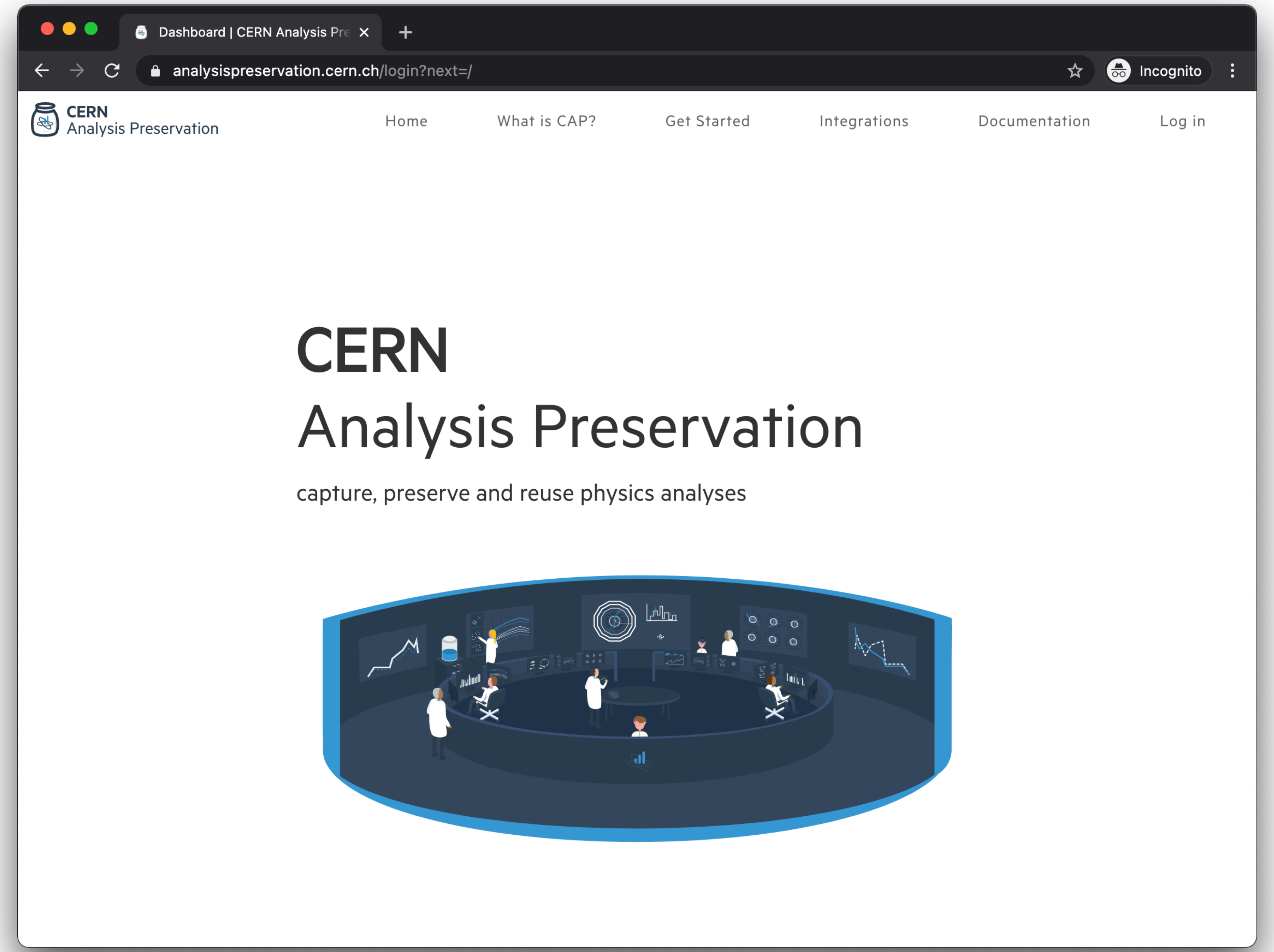
Mediator DM arxiv:1903.01400



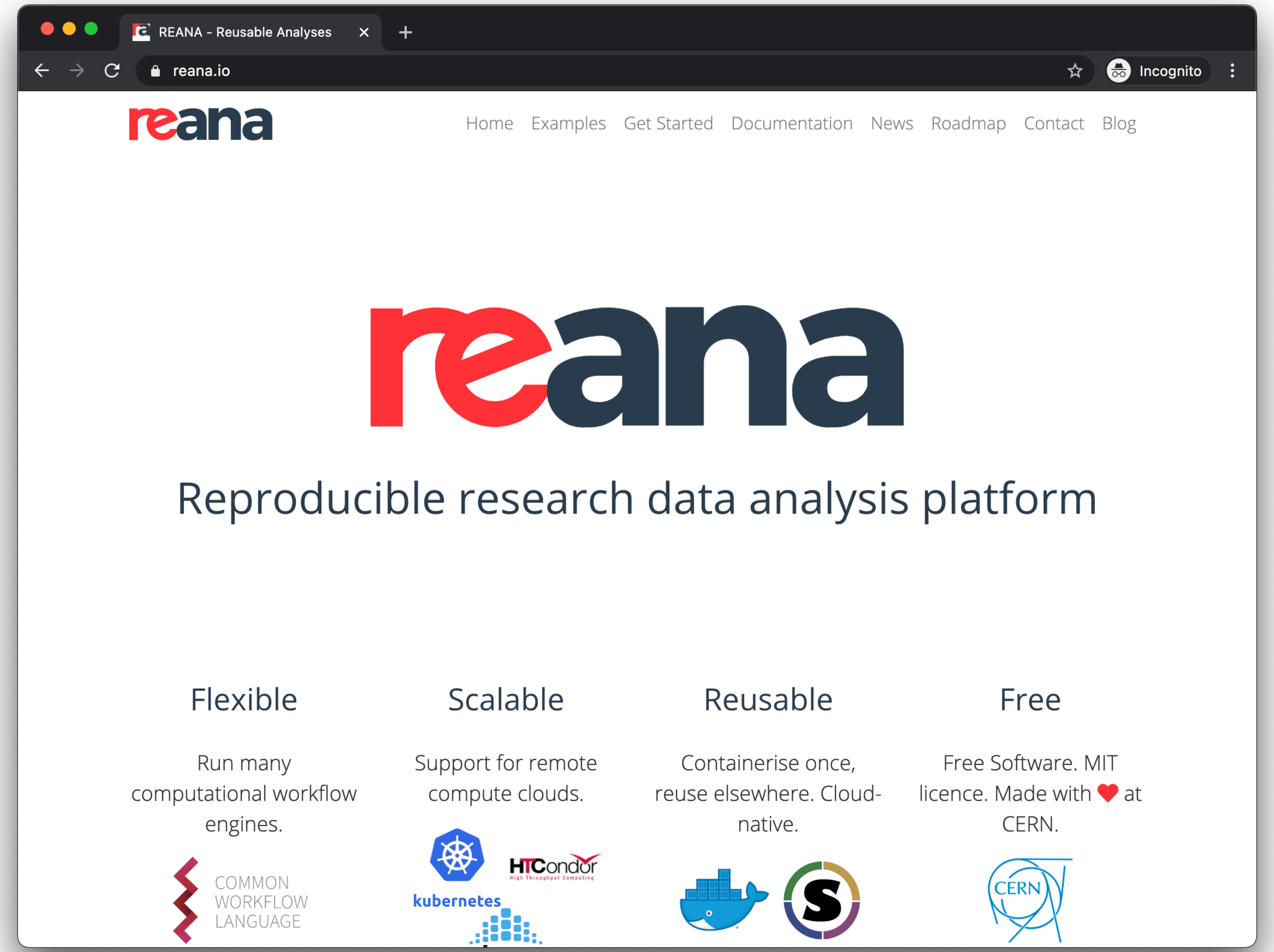

[K. Behr, P. Pani (DESY) et al]

# Institutional Support

## CERN investing in infrastructure to systematically archive and re-run analyses on using cloud technologies




Archive

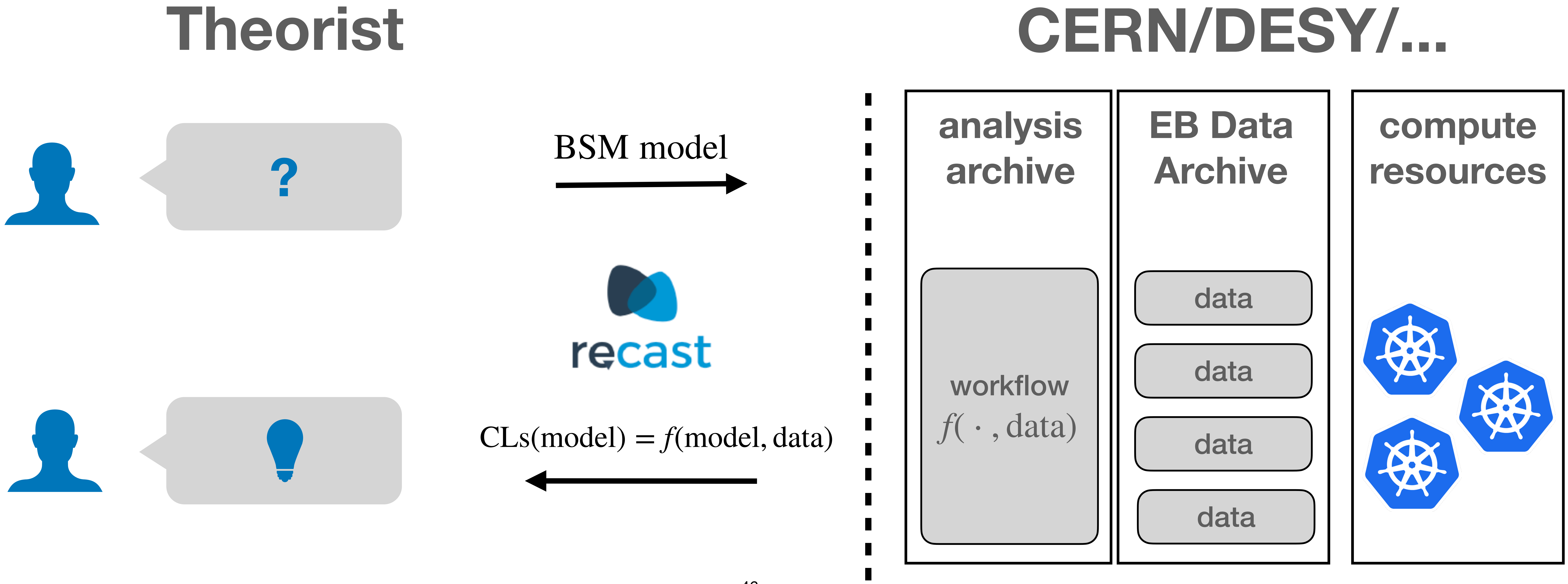



Re-run

# Future

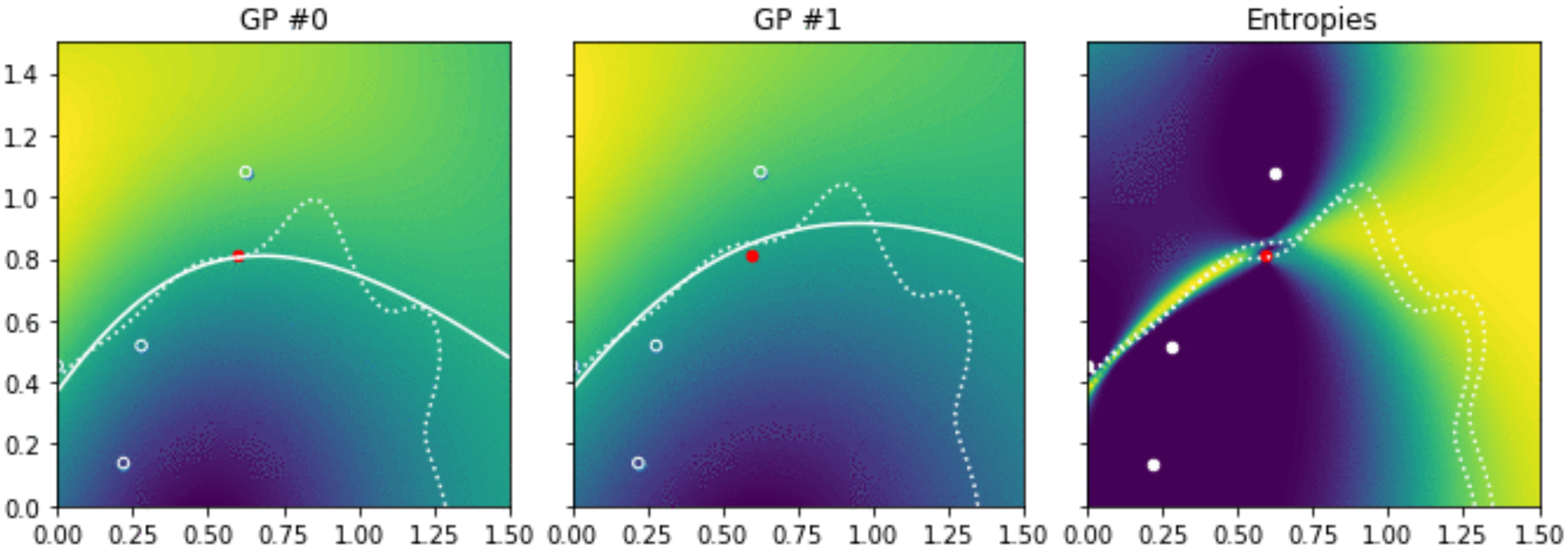
## Once we gain experience: Open Data at a higher abstraction level

- allow external researches to query LHC wrt. to new theories



# Future

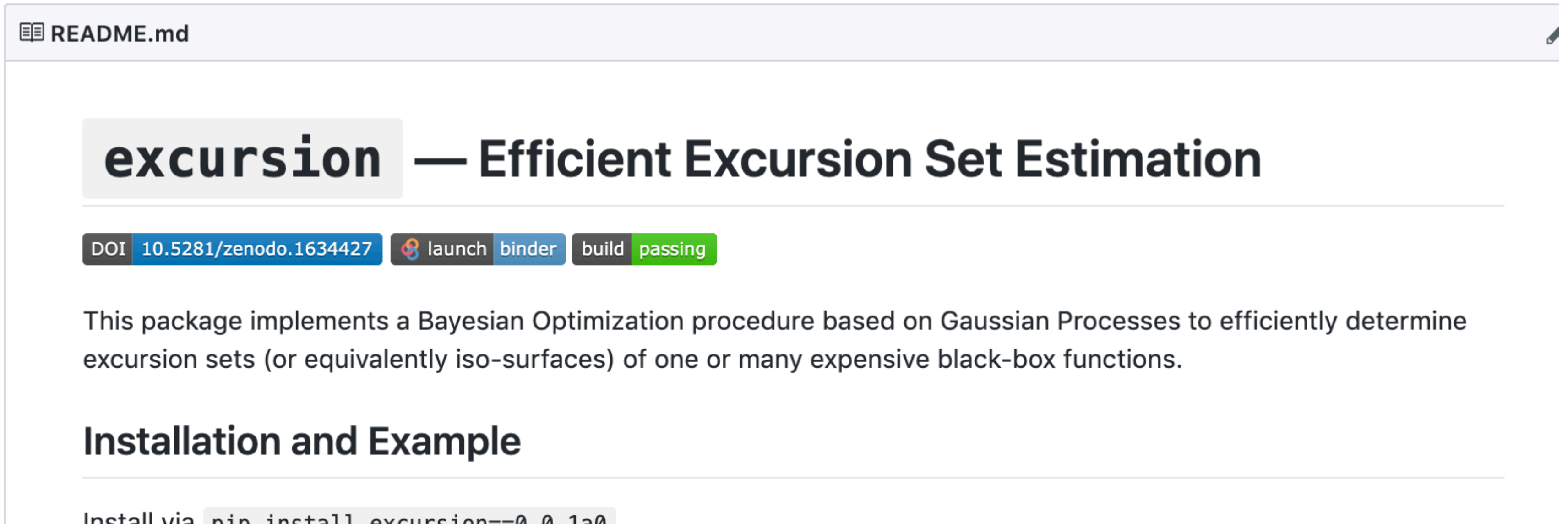
## Theory space is large, LHC is not sensitive everywhere.
## With streamlined reinterpretation, use ML to identify interesting models




[LH, G. Louppe]

## Now being tested in ATLAS - stay tuned!

**[P. Rieck, P. Gadow, J. v. Ahnen, I. Espejo +]**


README.md

## `excursion` — Efficient Excursion Set Estimation

DOI 10.5281/zenodo.1634427 | launch binder | build passing

This package implements a Bayesian Optimization procedure based on Gaussian Processes to efficiently determine excursion sets (or equivalently iso-surfaces) of one or many expensive black-box functions.

### Installation and Example

# Outlook

**HEP & other big science have unique challenges due to scale**

**Recent trends in IT, Data Science, ML bring us new tools**

- **possibility to fundamentally rethink how we approach analysis**
- **foundation of HL-LHC analysis & computing is defined now (join!)**
  **→ PB-scale interactive analysis?**
- **Technology can drive physics reach**

**Open Data, Reproducibility: How can we fill the buzzwords with life?**

- **LHC-wide policy: growing community & ecosystem of external tools**
- **Open Likelihoods: release the best info we have**
- **RECAST: reuse vs just reproduce**