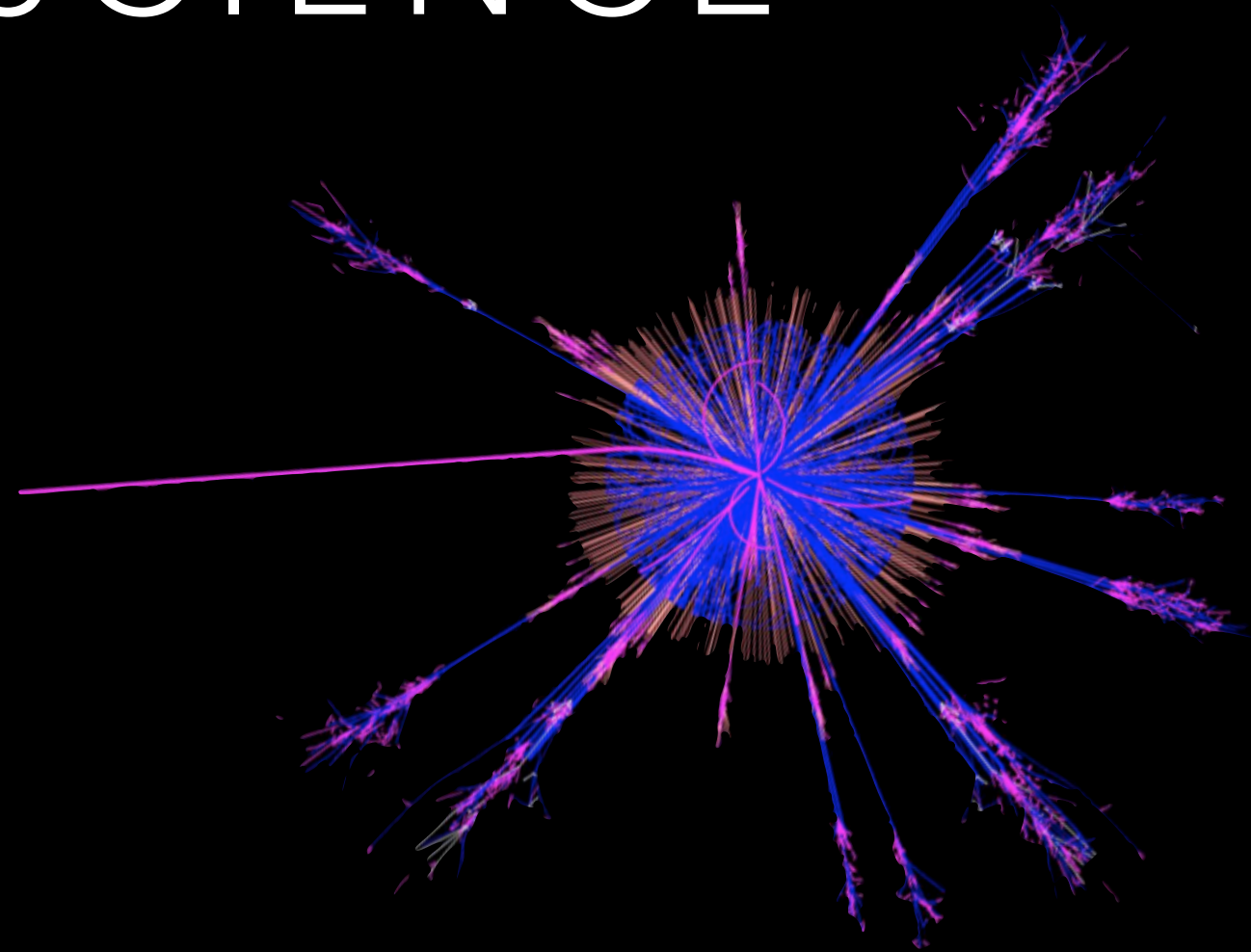




CYBERINFRASTRUCTURE FOR  
**HIGH-LEVEL SCIENCE**  
**GOALS**

**@KyleCranmer**  
New York University  
Department of Physics  
Center for Data Science

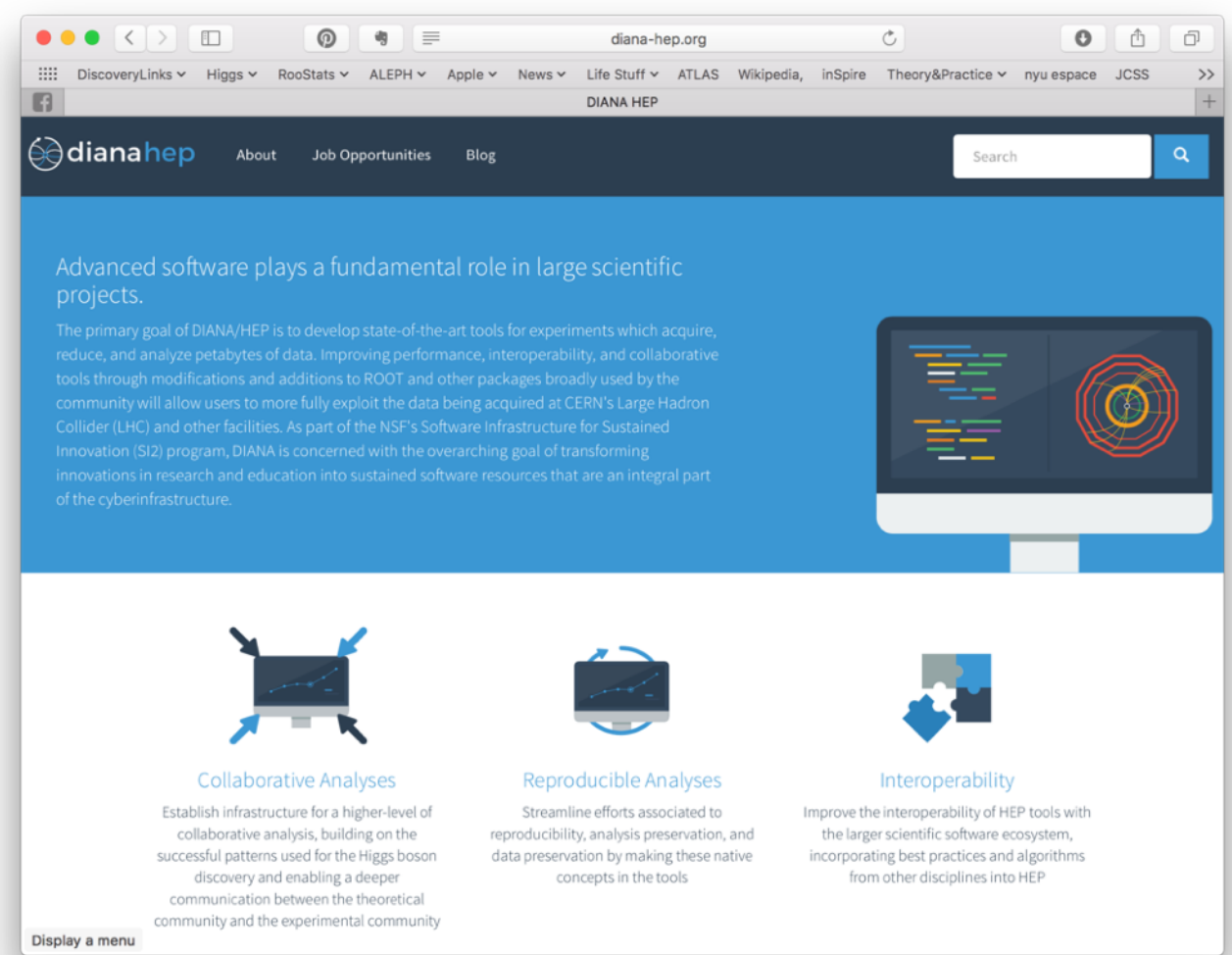


# Infrastructure Components

# DASPOS AND DIANA

DASPOS and DIANA are two large projects funded by the U.S. National Science Foundation focusing on issues around software and data for high energy physics.

We are working closely with CERN Analysis Preservation (CAP) portal, INSPIRE, and HEPData to build infrastructure for High Energy Physics



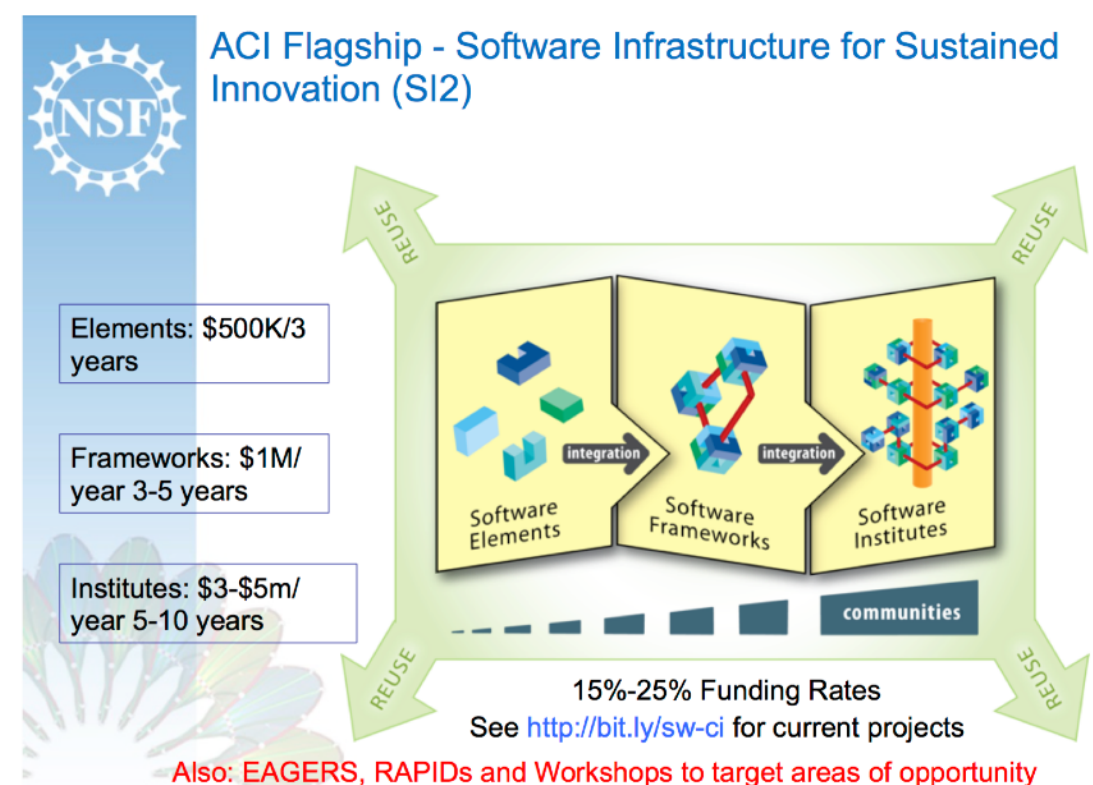
The SI2 program includes four classes of awards:

1. **Scientific Software Elements (SSE)**: SSE awards are Software Elements. They target small groups that will create and deploy robust software elements for which there is a demonstrated need that will advance one or more significant areas of science and engineering.

2. **Scientific Software Integration (SSI)**: SSI awards are Software Frameworks. They target larger, interdisciplinary teams organized around the development and application of common software infrastructure aimed at solving common research problems. SSI awards will result in sustainable community software frameworks serving a diverse community. ← DIANA is an SSI

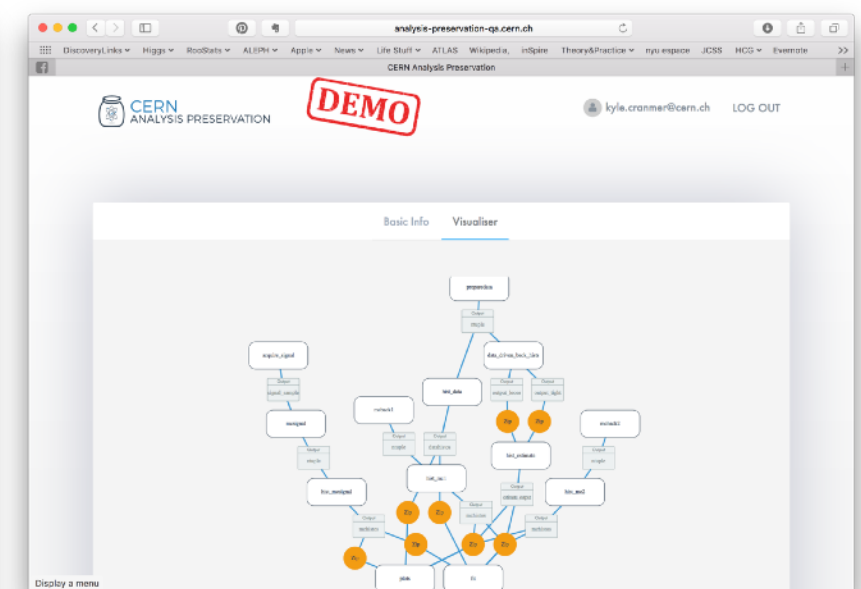
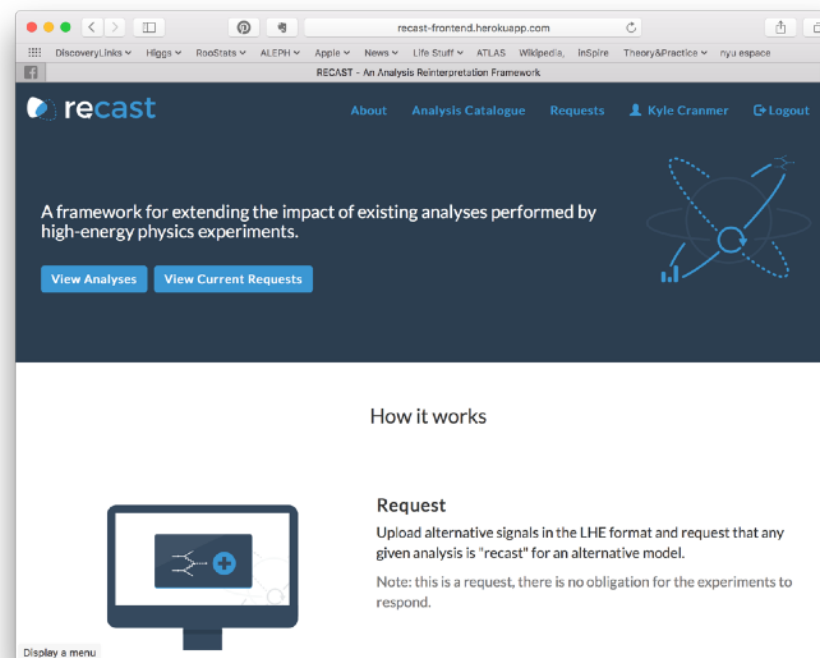
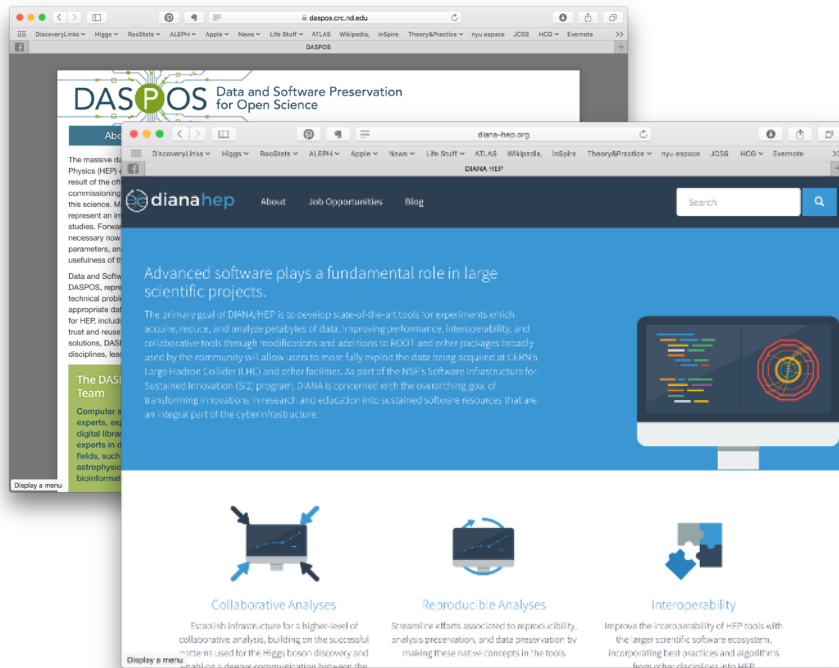
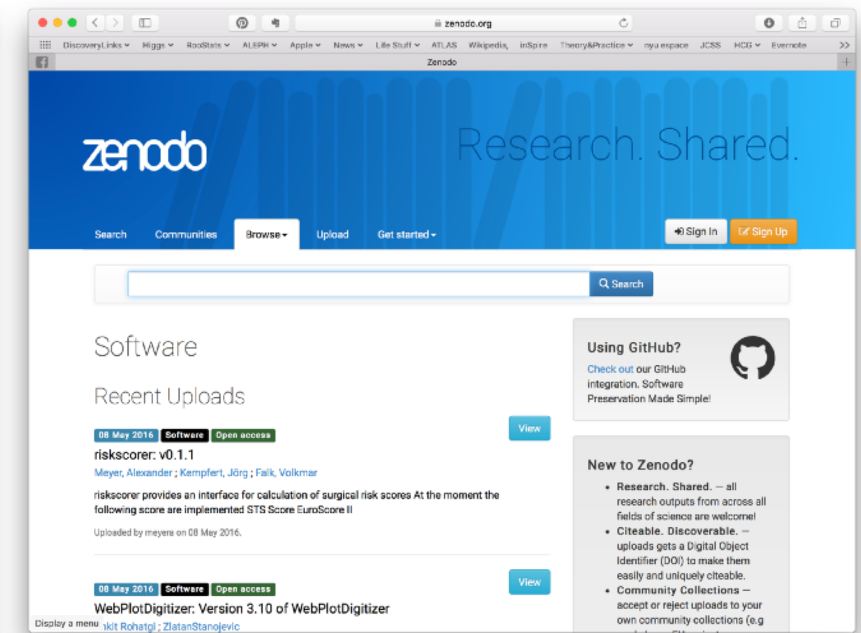
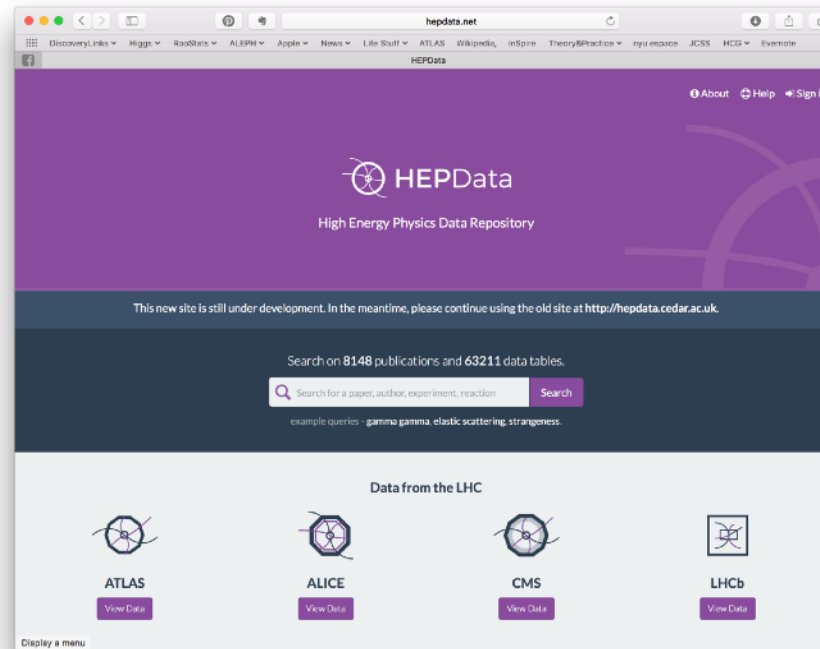
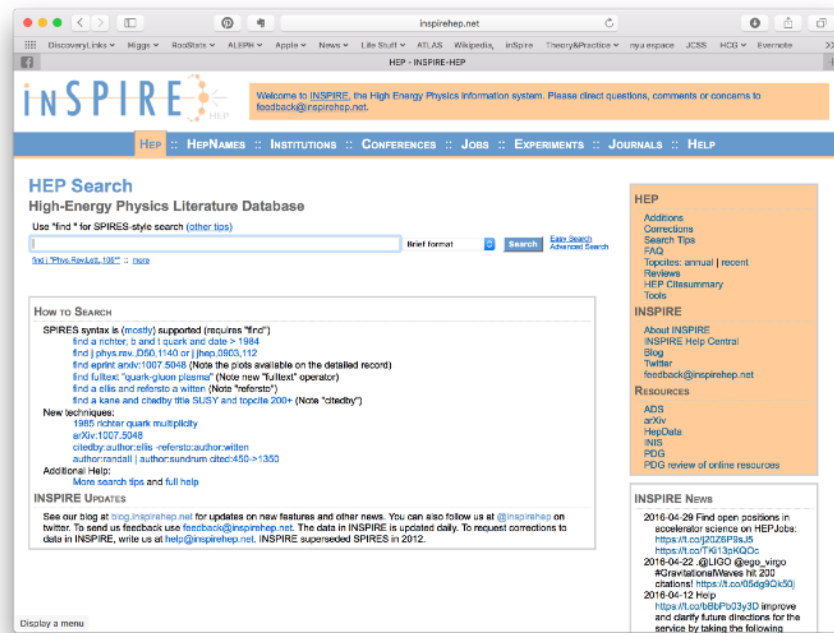
3. **Scientific Software Innovation Institutes (S2I2)**: S2I2 awards are Software Institutes. They focus on the establishment of long-term hubs of excellence in software infrastructure and technologies that will serve a research community of substantial size and disciplinary breadth. ← opens door to this

4. **Reuse**: In addition, SI2 provides support through a variety of mechanisms (including co-funding and supplements) to proposals from other programs that include, as an explicit outcome, reuse of software. Proposals that integrate with previously developed software, either by reference or inclusion, are encouraged. Proposals developing new software with an explicitly open design for reuse may also be considered. The purpose of the Reuse class is to stimulate connections within the broader software ecosystem. The class of reuse awards is currently being developed.





# INFRASTRUCTURE FOR DATA AND ANALYSIS PRESERVATION



Search HEPdata

Search

Reset search

Max results

Sort by

Reverse order

Showing 25 of 8194 results

## Date



## Collaboration

ATLAS	246
CDF	205
ZEUS	167
CMS	162
H1	142

[Next 5](#)
[Show All](#)

## Phrases

Exclusive	4187
Cross Section	3843
Integrated Cross Section	3843
Inclusive	3833
Single Differential Cross Section	2773

[Next 5](#)
[Show All](#)

## Reactions

pp --> pp	314
-----------	-----

### Production of $K^*(892)^0$ and $\phi(1020)$ in p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV

Adam, Jaroslav ; Adamova, Dagmar ; Aggarwal, Madan Mohan ; *et al.* The ALICE collaboration.

No Journal Information, 2016.

[Inspire Record 1418181](#)
[DOI 10.17182/hepdata.72807](#)

The production of  $K^*(892)^0$  and  $\phi(1020)$  mesons has been measured in p-Pb collisions at  $\sqrt{s_{NN}} = 5.02$  TeV.  $K^{*0}$  and  $\phi$  are reconstructed via their decay into charged hadrons with the ALICE detector in the rapidity range  $-0.5 < y < 0$ . The transverse momentum spectra, measured as a function of the multiplicity, have  $p_T$  range from 0 to 15 GeV/c for  $K^{*0}$  and from 0.3 to 21 GeV/c for  $\phi$ . Integrated yields, mean transverse momenta and particle...

30 data tables

Table 1	Average charged particle pseudo-rapidity density, $\langle dN_{ch}/d\eta_{lab} \rangle$ , measured at mid-rapidity in visible cross section event classes and average number of colliding nucleons, $\langle N_{coll} \rangle$ . Multiplicity classes are defined using the VOA estimator; values for $\langle dN_{ch}/d\eta_{lab} \rangle$ are corrected for vertexing and trigger efficiency. Since statistical uncertainties are negligible, only total systematic uncertainties are reported.
Table 2	$p_T$ -differential yield of $(K^{*0} + \overline{K^{*0}})/2$ in p-Pb collisions with centre-of-mass energy/nucleon=5.02 TeV (NSD). Additional systematic error: +- 3.1% (normalization).
Table 3	$p_T$ -differential yield of $(K^{*0} + \overline{K^{*0}})/2$ in p-Pb collisions with centre-of-mass energy/nucleon=5.02 TeV (0-20% multiplicity class).
More...	

### INVESTIGATION OF INCLUSIVE PROCESSES $\pi^- A \rightarrow \pi^- X$ AND $\pi^- A \rightarrow p \text{ (backwards)} X$ AT 40-GeV/c

Abrosimov, A.T. ; Albini, E. ; Antipov, V.V. ; *et al.*

Conference Paper, 2016.

[Inspire Record 209961](#)
[DOI 10.17182/hepdata.39782](#)

None

3 data tables

Table 1	No description provided.
---------	--------------------------

## Interactive Plotting Library

[Hide Publication Information Information](#)



momentum for...

### Table 9

Data from Auxiliary Material  
10.17182/hepdata.72205.v1/t9  
Extrapolated charged-particle multiplicity distributions in proton-proton collisions at a centre-of-mass energy of 13000 GeV for events with the number of...

### Table 10

Data from Auxiliary Material  
10.17182/hepdata.72205.v1/t10  
Extrapolated average transverse momentum in proton-proton collisions at a centre-of-mass energy of 13000 GeV as a function of the number...

### Table 11

Data from F 5A  
10.17182/hepdata.72205.v1/t11  
Charged-particle multiplicities in proton-proton collisions at a centre-of-mass energy of 13000 GeV as a function of pseudorapidity for events with...

### Table 12

Data from F 5B

### Table 10

Extrapolated average transverse momentum in proton-proton collisions at a centre-of-mass energy of 13000 GeV as a function of the number of charged particles in the event for events with the number of charged particles  $\geq 1$  having transverse momentum  $> 500$  MeV and absolute(pseudorapidity)  $< 2.5$ .

[10.17182/hepdata.72205.v1/t10](http://dx.doi.org/10.17182/hepdata.72205.v1/t10)

#### observables

☒ PT

#### phrases

☒ Inclusive

☒ Proton-Proton Scattering

#### reactions

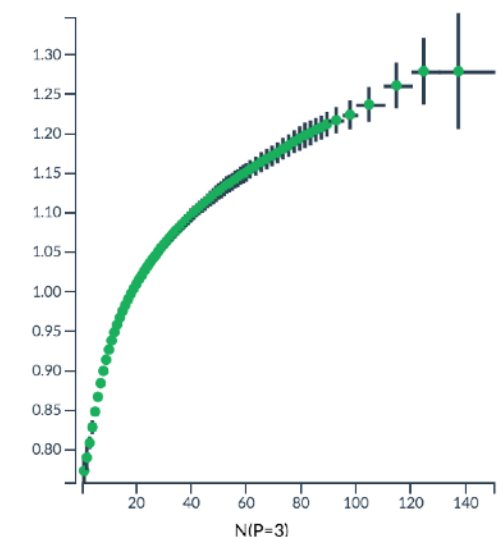
☒ P P --> CHARGED X

Showing 50 of 81 values

[Show All 81 values](#)

ETARAP(P=3)	-2.5-2.5
Extrapolated to include strange baryons	
N(P=3)	$\geq 1$
PT(P=3)	$> 500$ MEV
RE	P P --> CHARGED X
SQRTS(S)	13000.0 GeV
N(P=3)	MEAN(NAME=PT(P=3)) [GEV]
0.50 - 1.50	0.7737 $\pm 0.0008$ stat $\pm 0.0155$ sys
1.50 - 2.50	0.7904 $\pm 0.0007$ stat $\pm 0.0158$ sys
2.50 - 3.50	0.809 $\pm 0.001$ stat $\pm 0.008$ sys
3.50 - 4.50	0.8289 $\pm 0.0006$ stat $\pm 0.0084$ sys

### Visualize


 Sum errors ☒ Log Scale (X) ☐ Log Scale (Y) ☐

**Abstract (data abstract)**  
CERN-LHC. Measurements of charged particle distributions in proton-proton collisions at a centre-of-mass energy of 13 TeV. A data sample of nearly 9 million events recorded by the ATLAS detector during a special LHC fill with low beam currents, and thus giving a low expected mean number of interactions, is used. The charged-particle multiplicity, its dependence on transverse momentum and pseudorapidity and the dependence of the mean transverse momentum on the charged-particle multiplicity are presented. The measurements are performed with charged particles with transverse momentum greater than 500 MeV and absolute pseudorapidity less than 2.5, in events with at least one charged particle satisfying these kinematic requirements

# Open Data & Preservation





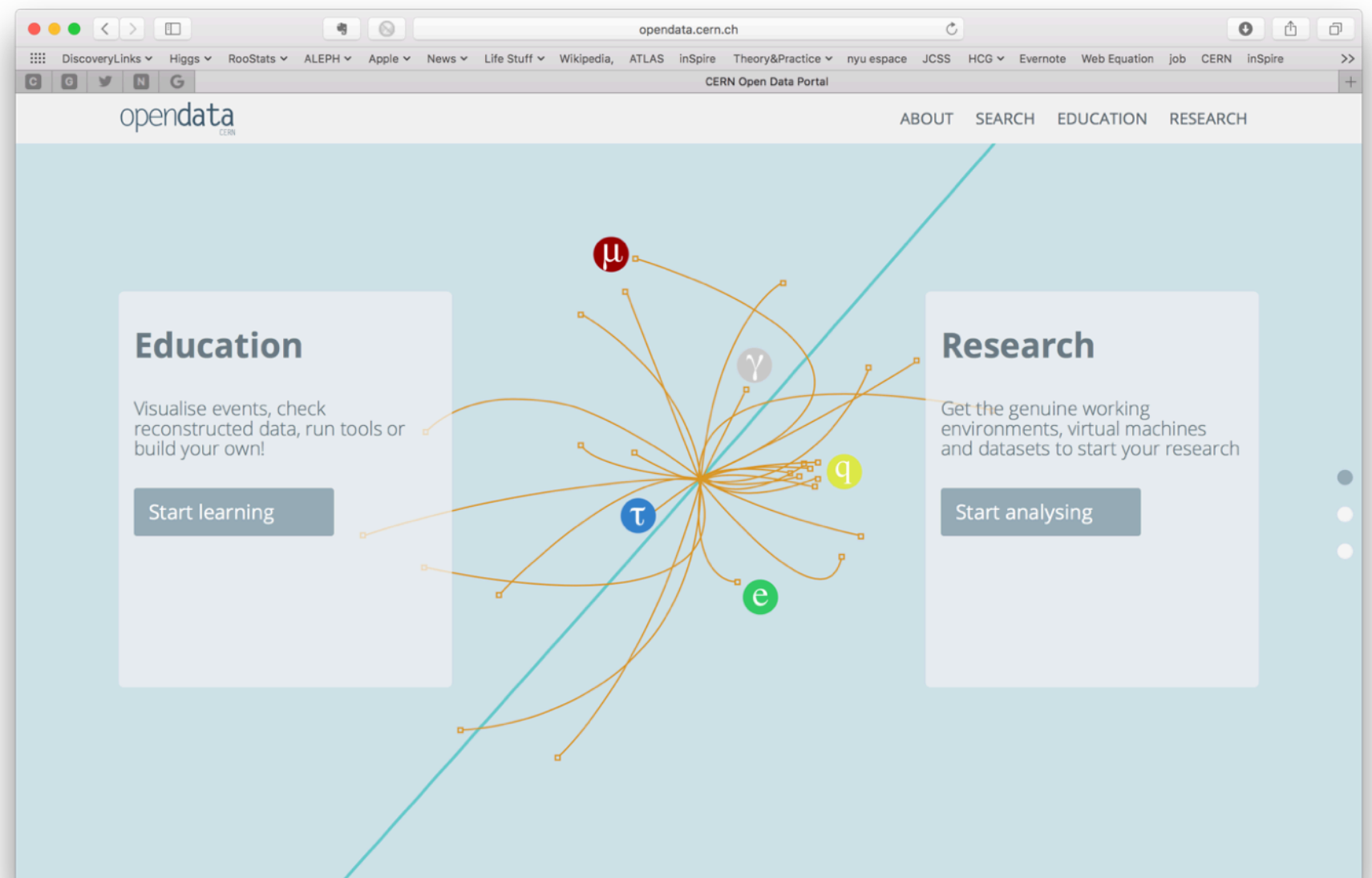
# Data Preservation in High Energy Physics

Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

Partners Accelerators Meetings ICFA Stud

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Table 3: Various preservation models, listed in order of increasing complexity.





## Jet Substructure Studies with CMS Open Data

Aashish Tripathee,<sup>1,\*</sup> Wei Xue,<sup>1,†</sup> Andrew Larkoski,<sup>2,‡</sup> Simone Marzani,<sup>3,§</sup> and Jesse Thaler<sup>1,¶</sup>

<sup>1</sup>*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

<sup>2</sup>*Physics Department, Reed College, Portland, OR 97202, USA*

<sup>3</sup>*University at Buffalo, The State University of New York, Buffalo, NY 14260-1500, USA*

We use public data from the CMS experiment to study the 2-prong substructure of jets. The CMS Open Data is based on  $31.8 \text{ pb}^{-1}$  of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768,687 events containing a high-quality central jet with transverse momentum larger than 85 GeV. Using CMS's particle flow reconstruction algorithm to obtain jet constituents, we extract the 2-prong substructure of the leading jet using soft drop declustering. We find good agreement between results obtained from the CMS Open Data and those obtained from parton shower generators, and we also compare to analytic jet substructure calculations performed to modified leading-logarithmic accuracy. Although the 2010 CMS Open Data does not include simulated data to help estimate systematic uncertainties, we use track-only observables to validate these substructure studies.

From a physics perspective, our experience with the CMS Open Data was fantastic. With PFCs, one can essentially perform the same kinds of four-vector-based analyses on real data as one would perform on collisions from parton shower generators. Using open data has the potential to accelerate scientific progress (pun intended) by allowing scientists outside of the official detector collaborations to pursue innovative analysis techniques. We hope that our jet substructure studies have demonstrated both the value in releasing public data and the enthusiasm of potential external users. We encourage other members of the particle physics community to take advantage of this unique data set.

From a technical perspective, though, we encountered a number of challenges. Some of these challenges were simply a result of our unfamiliarity with the CMSSW framework and the steep learning curve faced when trying to properly parse the AOD file format. Some of these challenges are faced every day by LHC experimentalists, and it is perhaps unreasonable to expect external users to have an easier time than collaboration members. Some of these challenges (particularly the issue of detector-simulated samples) have been partially addressed by the 2011A CMS Open Data release [215]. That said, we suspect that some issues were not anticipated by the CMS Open Data project, and we worry that they have deterred other analysis teams who might have otherwise found interesting uses for open data. Therefore, we think it is useful to highlight the primary challenges we faced, followed by specific recommendations for how potentially to address them.

### A. Challenges

Here are the main issues that we faced in performing the analyses in this paper.

- *Slow development cycle.* As CMSSW novices, we often needed to perform run-time debugging to figure out how specific functions worked. There were two elements of the CMSSW workflow that introduced a considerable lag between starting a job and getting debugging feedback. The first is that, when using the XROOTD interface, one has to face the constant overhead (and inconstant network performance) of retrieving data remotely. The second is that, as a standard part of every CMS analysis, one has to load configuration files into memory. Loading `FrontierConditions_GlobalTag.cff` (which is necessary to get proper trigger prescale values) takes around 10 minutes at the start of a run. For most users, this delay alone would be too high of a barrier for using the CMS Open Data. By downloading the AOD files directly and building our own MOD file format, we were able to speed up the

development cycle through a lightweight analysis framework. Still, creating the MODPRODUCER in the first place required a fair amount of trial, error, and frustration.

- *Scattered documentation.* Though the CMS Open Data uses an old version of CMSSW (v4.2 compared to the latest v9.0), there is still plenty of relevant documentation available online. The main challenge is that it is scattered in multiple places, including online TWIKI pages, masterclass lectures, thesis presentations, and GITHUB repositories. Eventually, with help from CMS insiders, we were able to figure out which information was relevant to a particular question, but we would have benefitted from more centralized documentation that highlighted the most important features of the CMS Open Data. Centralized documentation would undoubtedly help CMS collaboration members as well, as would making more TWIKI pages accessible outside of the CERN authentication wall.
- *Lack of validation examples.* When working with public data, one would like to validate that one is doing a sensible analysis by trying to match published results. While example files were provided, none of them (to our knowledge) involved the complications present in a real analysis, such as appropriate trigger selection, jet quality criteria, and jet energy corrections. Initially, we had hoped to reproduce the jet  $p_T$  spectrum measured by CMS on 2010 data [263], but that turned out to be surprisingly difficult, since very low  $p_T$  jet triggers are not contained in the Jet Primary Dataset, and we were not confident in our ability to merge information from the MinimumBias Primary Dataset. (In addition, the published CMS result is based on inclusive jet  $p_T$  spectra, while we restricted our analysis to the hardest jet in an event to simplify trigger assignment.) Ideally, one should be able to perform event-by-event validation with the CMS Open Data, especially if there are important calibration steps that could be missed.<sup>13</sup>
- *Information overload.* The AOD files contains an incredible wealth of information, such that the majority of official CMS analyses can use the AOD format directly without requiring RAW or RECO information. While ideal for archival purposes, it is an overload of information for external users, especially because some information is effectively duplicated. The main reason we introduced the MOD

<sup>13</sup> In the one case where we thought it would be the most straightforward to cross check results, namely the luminosity study in Fig. 2, it was frustrating to later learn that the AOD files contained insufficient information.

# Excerpts from Symposium

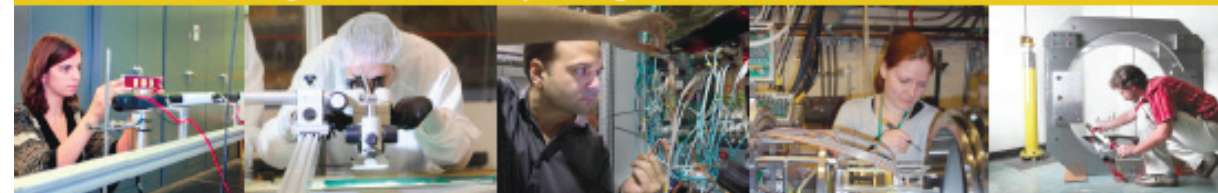
# NUCLEAR PHYSICS IN A DECADE

**DONALD GEESAMAN**  
Argonne National Laboratory

Future Trends in Nuclear Physics Computing  
May 2, 2017

REACHING FOR THE HORIZON

The Site of the Wright Brothers' First Airplane Flight



The 2015  
**LONG RANGE PLAN**  
for **NUCLEAR SCIENCE**

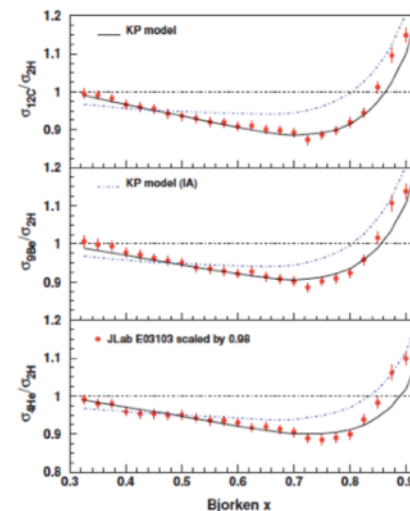


## IN MOST CASES, IT WILL BE JOINT PROGRESS BETWEEN THEORY AND EXPERIMENT THAT MOVES US FORWARD, NOT IN ONE SIDE ALONE

Does the structure of the proton and the forces between nucleons change in nuclear matter (beyond n-body force effects)?

This has been an active subject since the dawn of the EMC effect. The problem is we have just had “toy” models that fit systematics for one observable but gave us little insight into others.

In ten years I expect lattice QCD calculations of nuclei to be able to properly ask this question and tell us what measurements will confirm it. I think we will have the measurements in hand.



## SUMMARY

- It will be joint progress and theory and experiment that moves us forward, not in one side alone
- Continued rapid progress in Nuclear Physics Computing is imperative.
  - Challenge in handling the raw data
  - Challenge in analyzing the data
  - Challenge in simulating real detectors
- Theory challenge in calculating the physics right
  - Nuclear structure
  - Hadron structure
  - Astrophysical environments

## GLOBAL COLLABORATION

All of this is in the context of more globalization of the research effort.

There will be an emphasis on

- Common Tools
- Common Structures
- Remote collaboration

but on diverse architectures and with diverse short range goals.





# Synergies of Computing and the Next Generation of Nuclear Physics Experiments

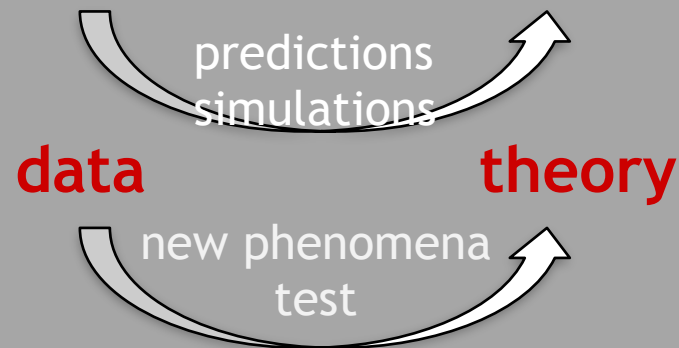
Rolf Ent (Jefferson Lab)

With acknowledgement to Amber Boehnlein, Markus Diefenthaler and Graham Heyes for their help



# Interplay of data and theory

## Feedback loop between data and theory



## Comparison to:

- analytical calculations
- Monte Carlo (MC) simulations
- Lattice-QCD calculations

## Data-theory comparison: relies on

- open access to data-theory tools
- standardization of data-theory tools
- comparison tools for quick turnaround

## MC event generator:

- faithful representation of QCD dynamics
- based on QCD factorization and evolution equations

## Usage by experimentalists:

- detector corrections
- analysis prototyping
- comparing to theory

## Usage by theoreticians:

- easy off-the-shelf state-of-the-art tool that looks like data
- validate against and investigate theoretical improvements

# Analysis environments

## Developments of analysis environments:

- new projects starting (JLab 12 GeV) and on the horizon (EIC)
- likely explosion of data even at the small nuclear experiments
- think about the **next generation(s) of analysis environments** that will **maximize** the **science output**

**LHC experiments:** tremendous success in achieving their analysis goals and producing results in timely manners

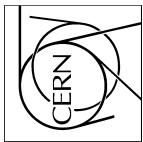
## Lesson learned at LHC experiments:

- as the complexity and size of the experiments grew
- the complexity of analysis environment grew
- time dealing with the analysis infrastructure grew

### Anecdote from LHC

a typical LHC student or post-doc spends up to 50 % of his/her time dealing with computing issues

Targeting the theory/experiment interface



## Level-1. Published results

All scientific output is published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN.

Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, ...) are stored in appropriate repositories such as [HEPDATA\[2\]](#). ATLAS also strives to make additional material related to the paper available that allows a reinterpretation of the data in the context of new theoretical models. For example, an extended encapsulation of the analysis is often provided for measurements in the framework of RIVET [3]. For searches information on signal acceptances is also made available to allow reinterpretation of these searches in the context of models developed by theorists after the publication. ATLAS is also exploring how to provide the capability for reinterpretation of searches in the future via a service such as RECAST [4]. RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.



# U.S. Particle Physics: Building for Discovery

*U.S. Particle Physics Strategy*

*Education and Outreach Site*

Five intertwined Science Drivers provide compelling lines of inquiry that show great promise for discovery.



*Use the Higgs boson as a new tool for discovery.*



*Pursue the physics associated with neutrino mass.*



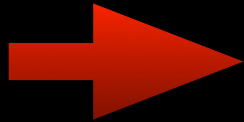
*Identify the new physics of dark matter.*



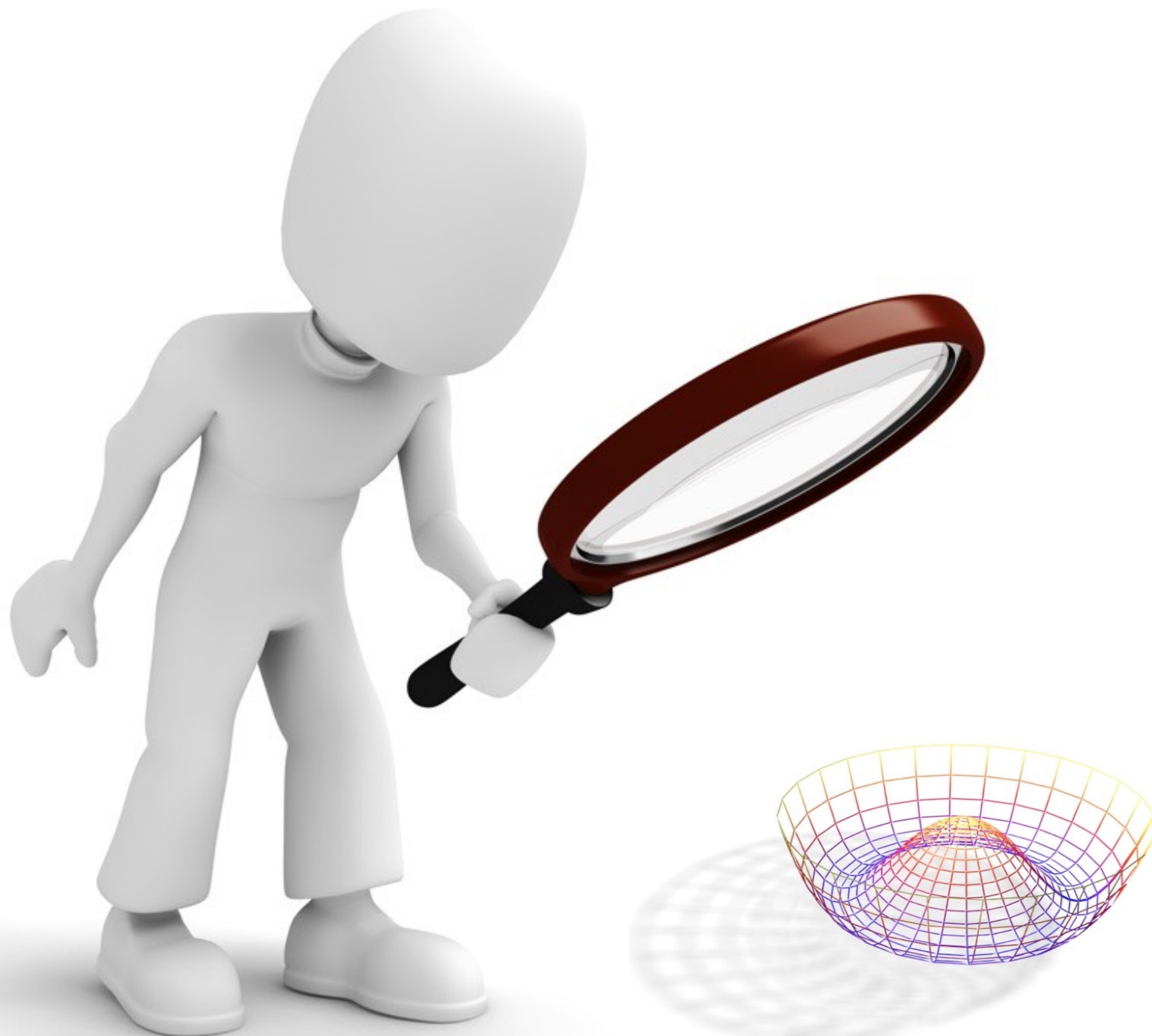
*Understand cosmic acceleration: dark energy and inflation.*



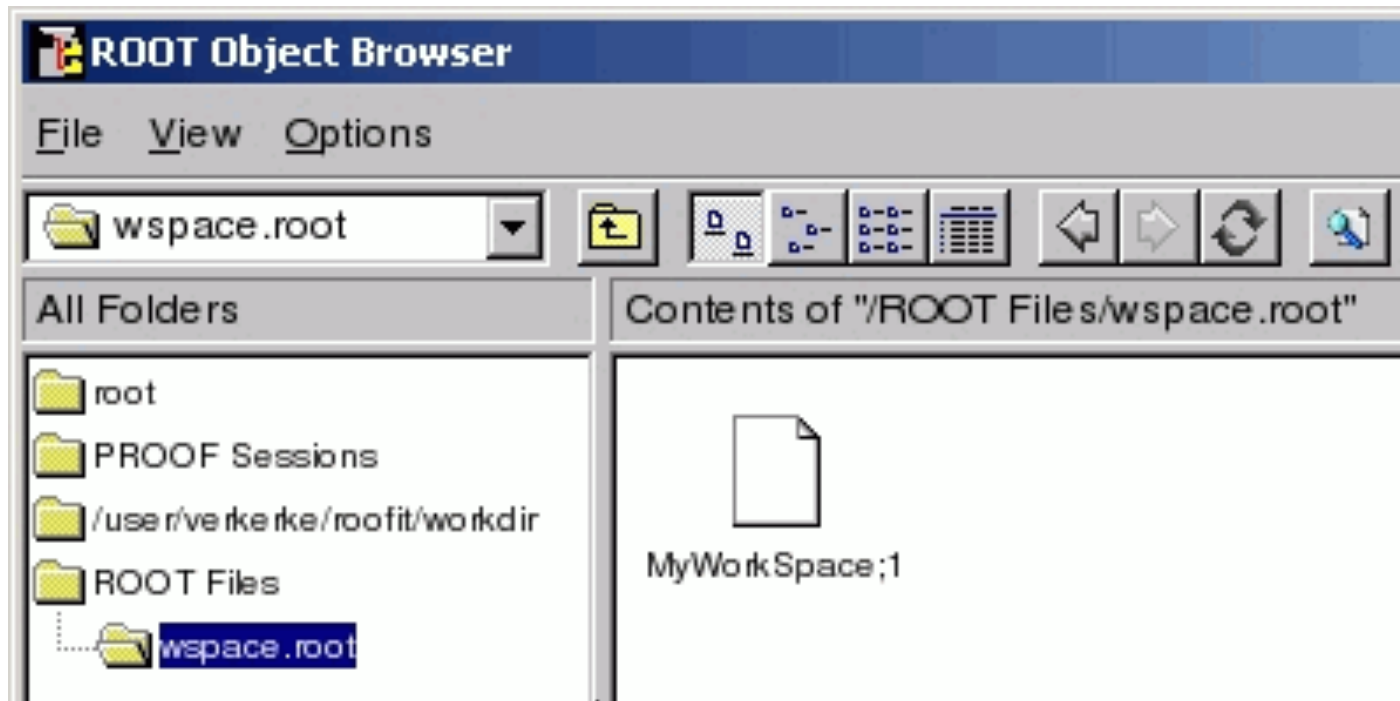
*Explore the unknown: new particles, interactions, and physical principles.*



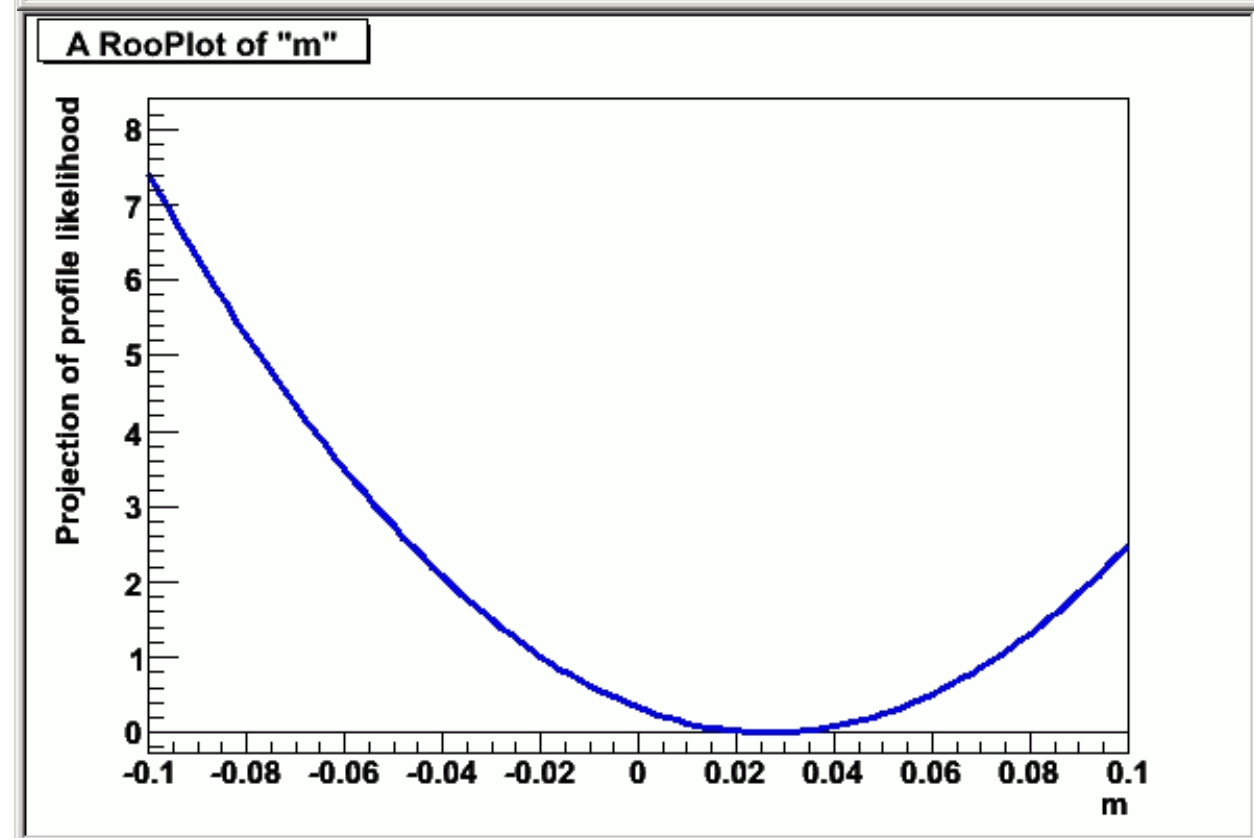
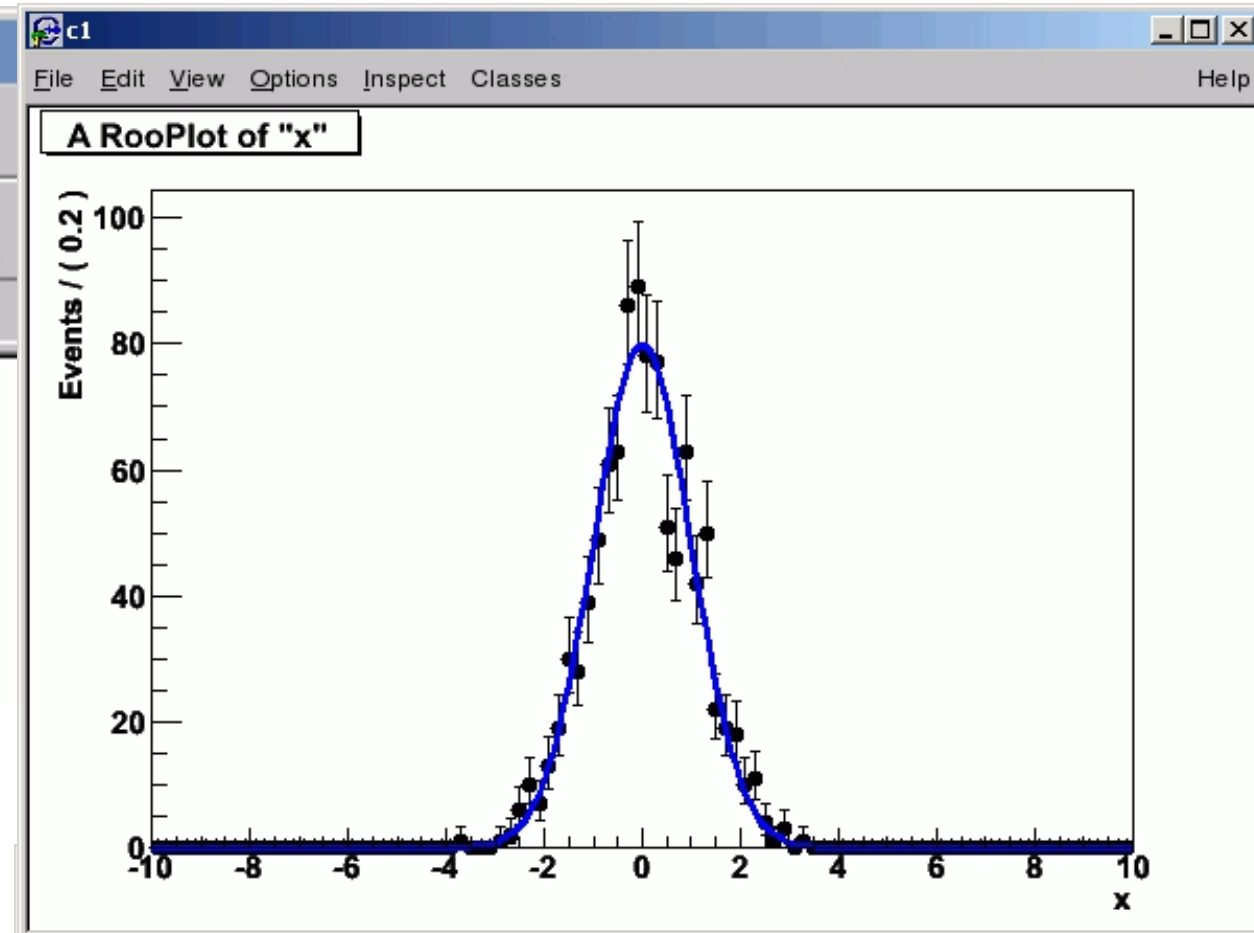




# DIGITAL PUBLISHING STATISTICAL MODELS

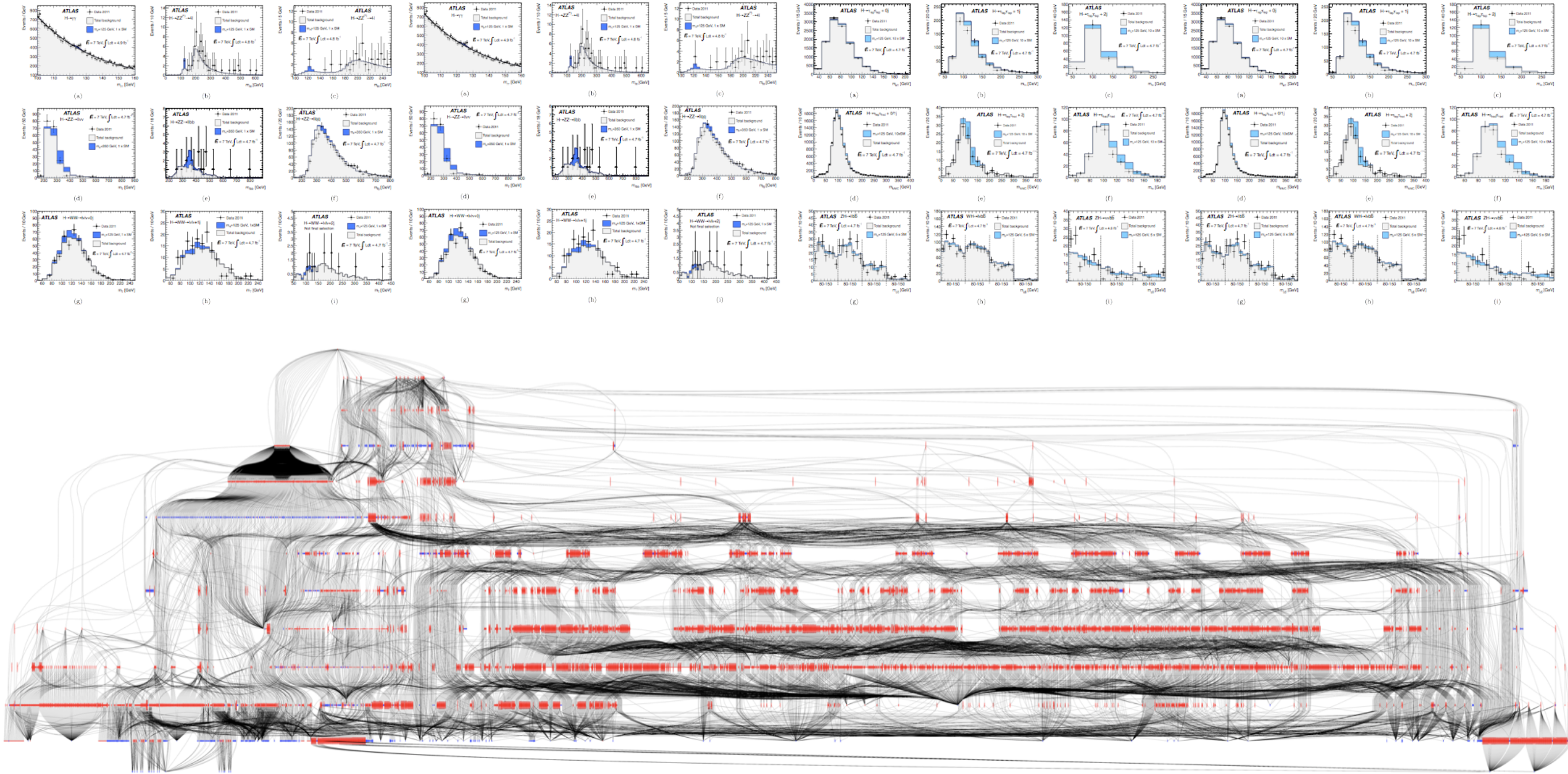


RooFit workspaces & HistFactory



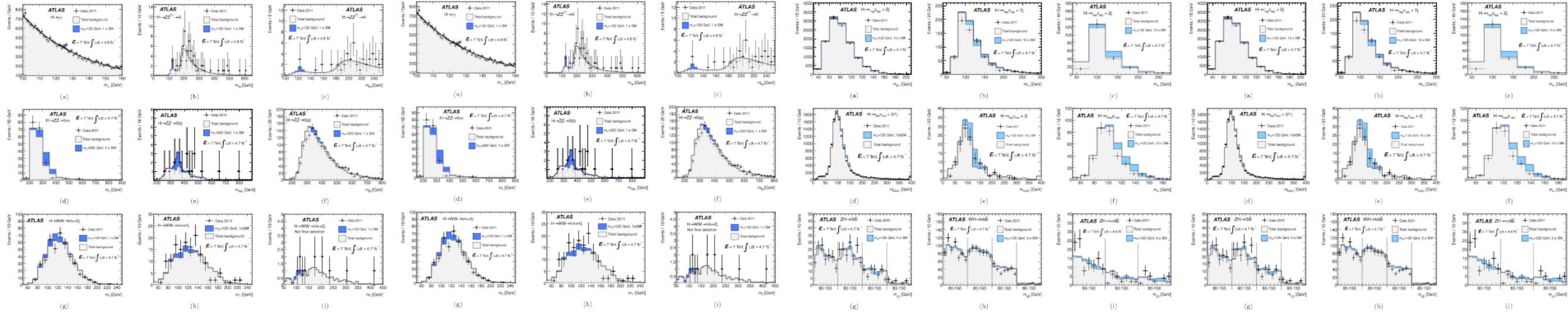


# COLLABORATIVE STATISTICAL MODELING

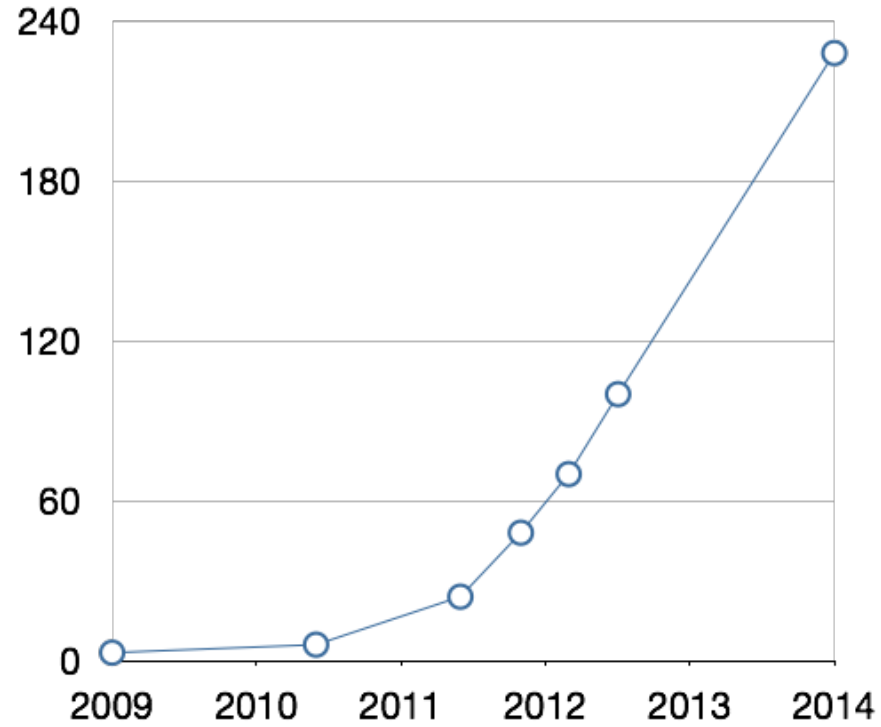


$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[ \text{Pois}(n_c | \nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce} | \boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathcal{S}} f_p(a_p | \alpha_p)$$

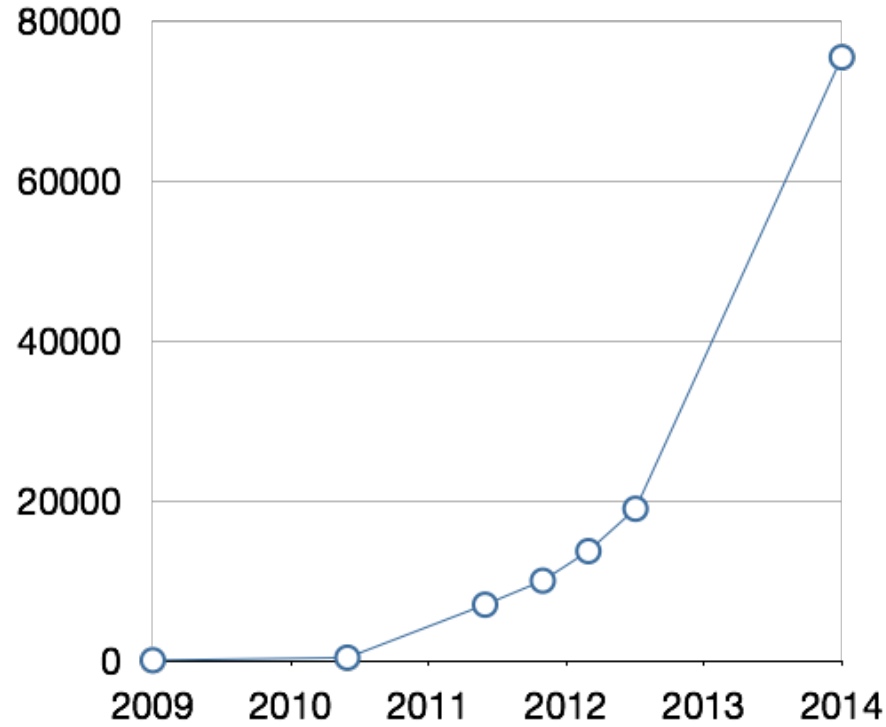
# COLLABORATIVE STATISTICAL MODELING



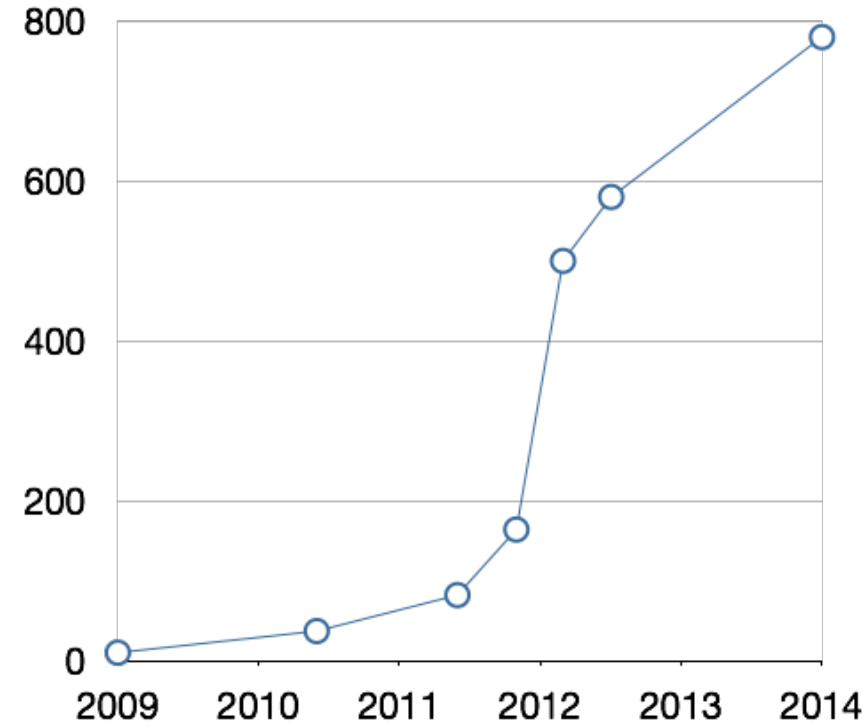
**Number of Datasets Combined**



**Number of Model Components**

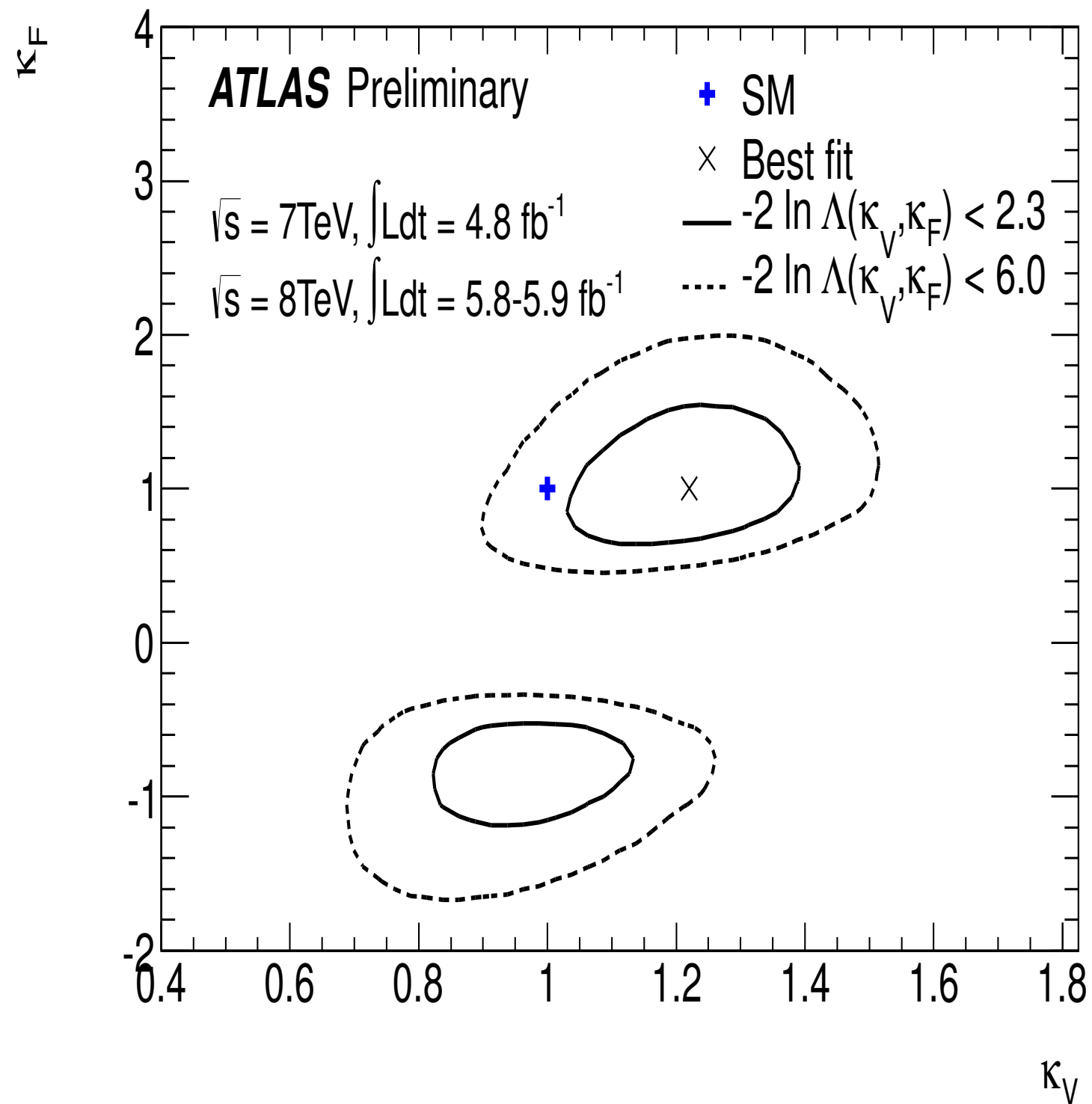


**Number of Parameters in Likelihood**



# REPRODUCIBILITY PROBLEM

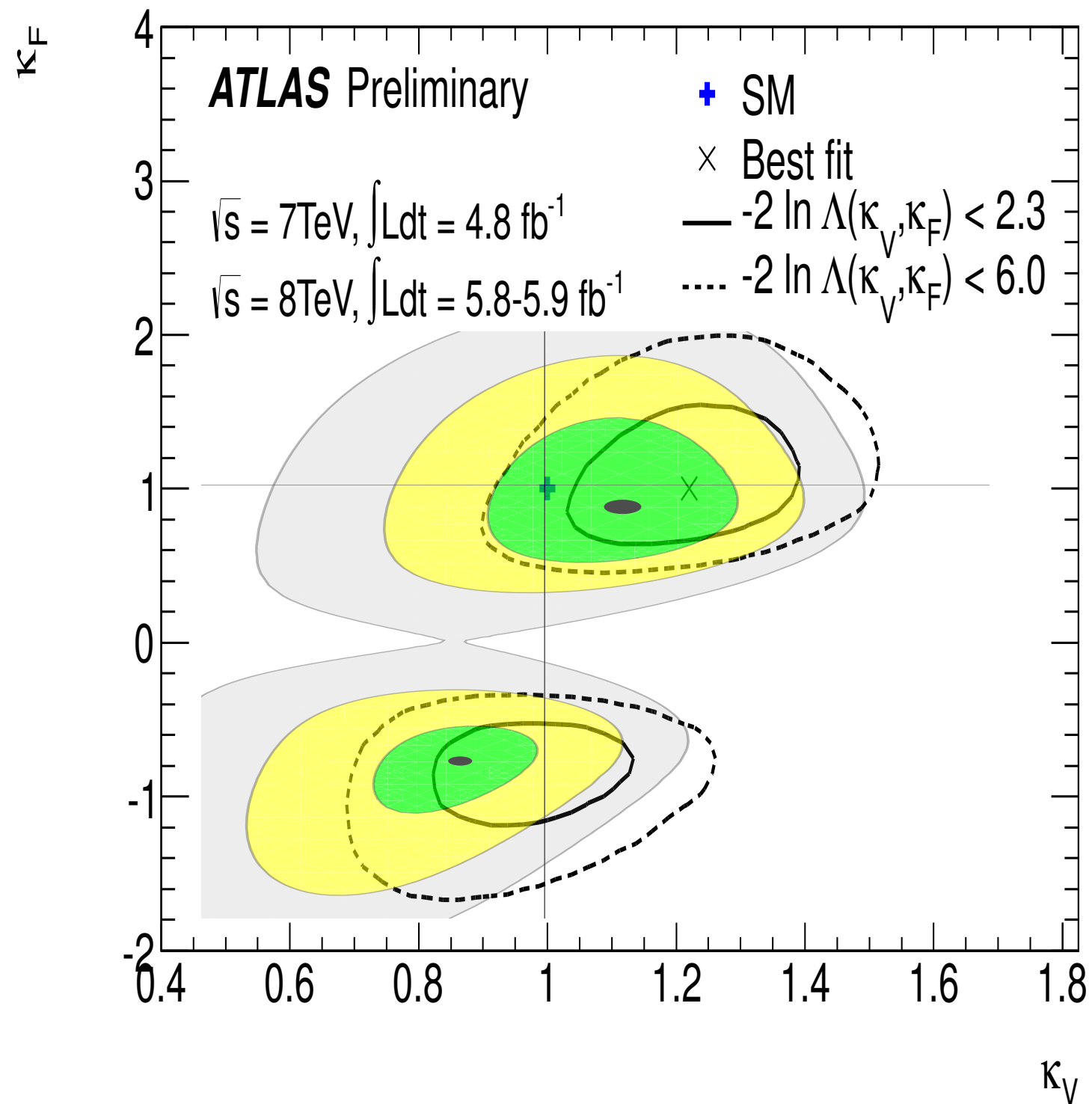
*Not possible for others to reproduce results from paper.*





# REPRODUCIBILITY PROBLEM

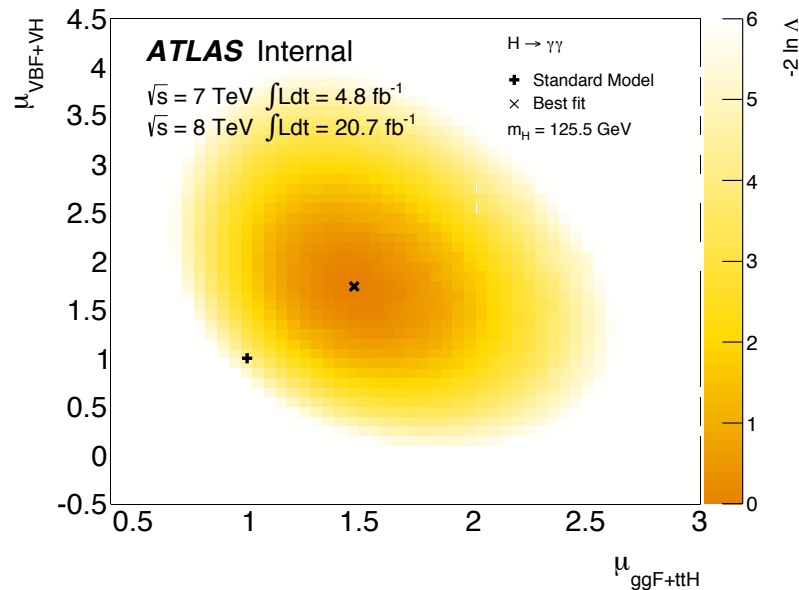
*Not possible for others to reproduce results from paper.*



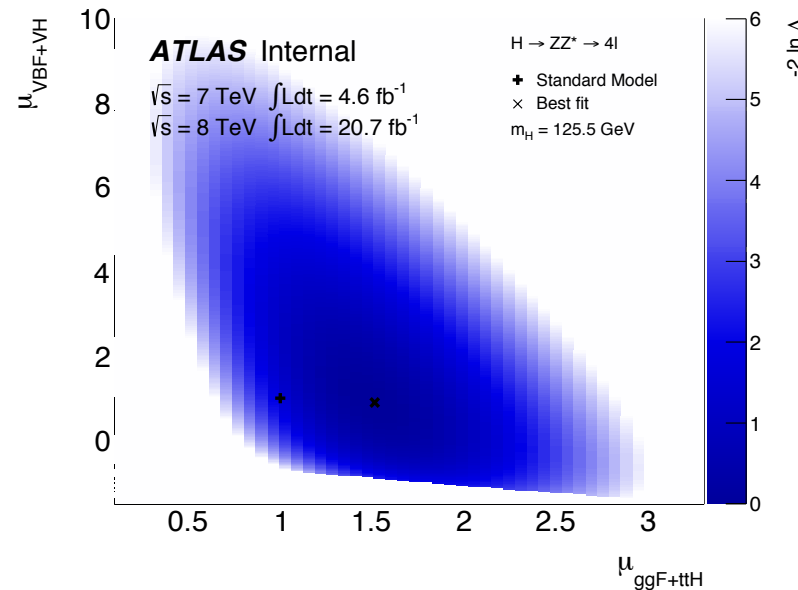
# WHAT INFO AND HOW TO RETRIEVE IT

Through collaboration with theoretical community, we were able to identify a targeted form of data sharing that balanced generality &

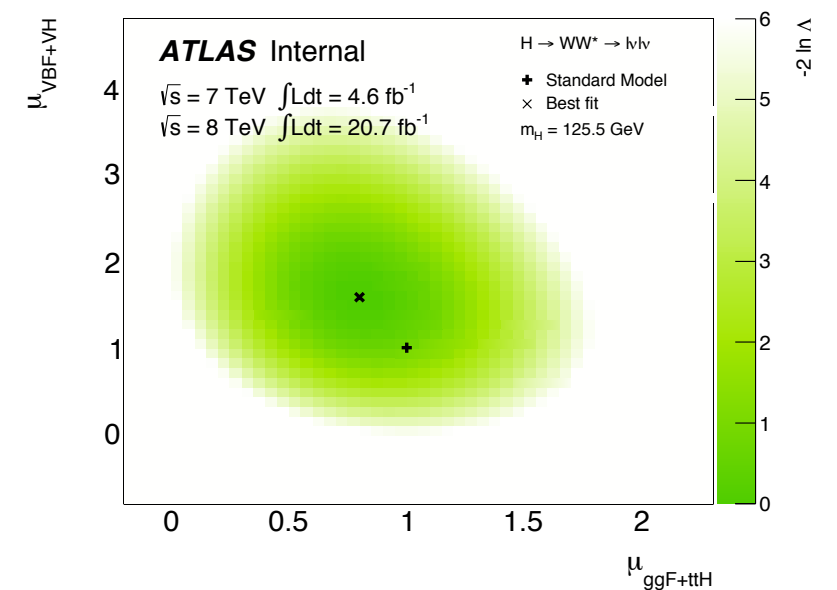
<http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44>



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



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



These data are directly linked to the paper in INSPIRE and have been cited:



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

HEP :: HEPNames :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HEPData

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>

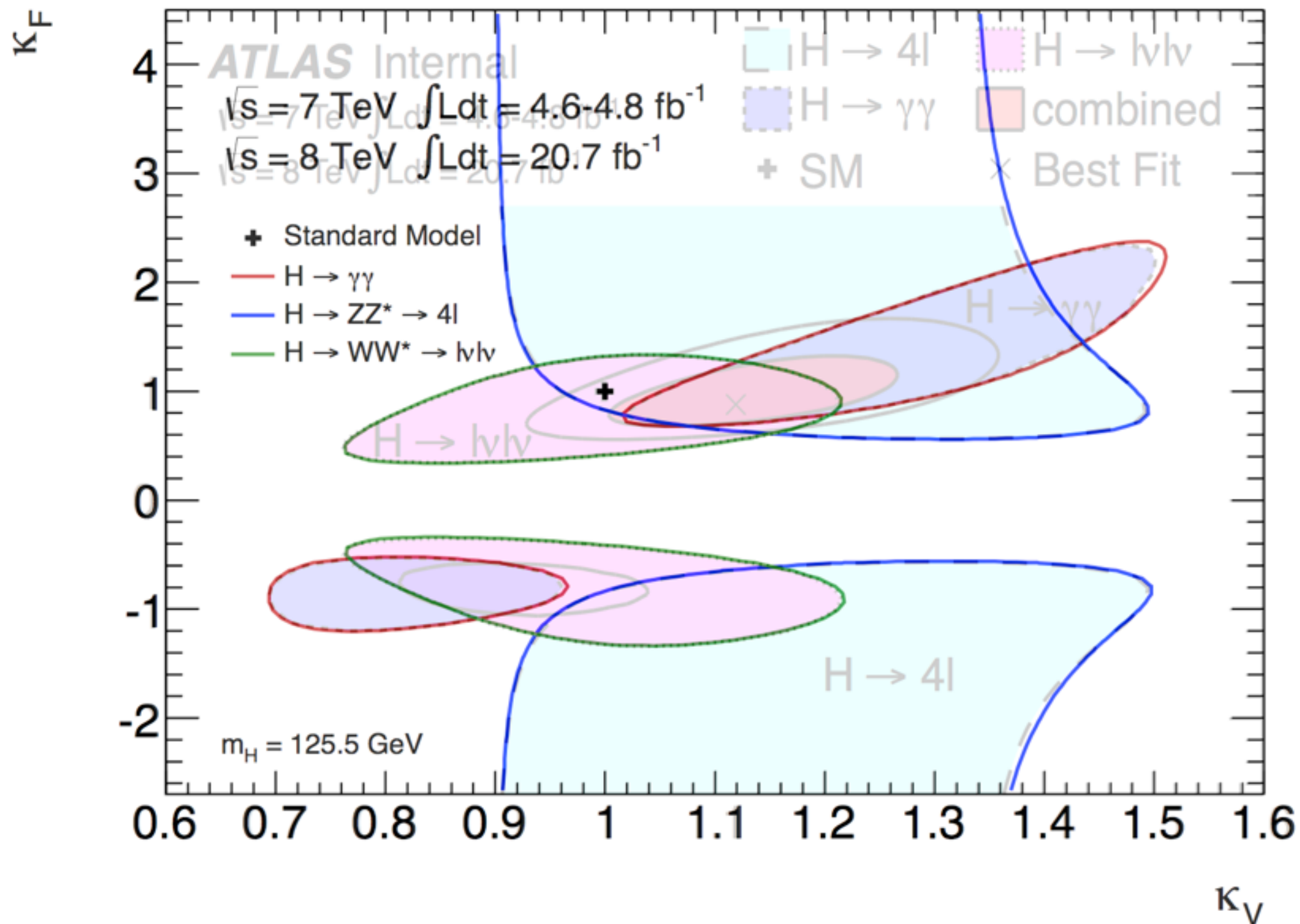


Blogged by 3  
Tweeted by 6

[Click for more details](#)

# LIKELIHOODS ON HEPDATA

*Reproducing derived results from original paper!*



Reinterpretation & Reusability



# FOLDING VS. UNFOLDING



Rivet  
Unfolding


Accuracy meets precision  
MC event simulation in a decade

Stefan Höche

SLAC National Accelerator Laboratory

Future Trends in Nuclear Physics Computing  
Jefferson Lab, 05/02/2017

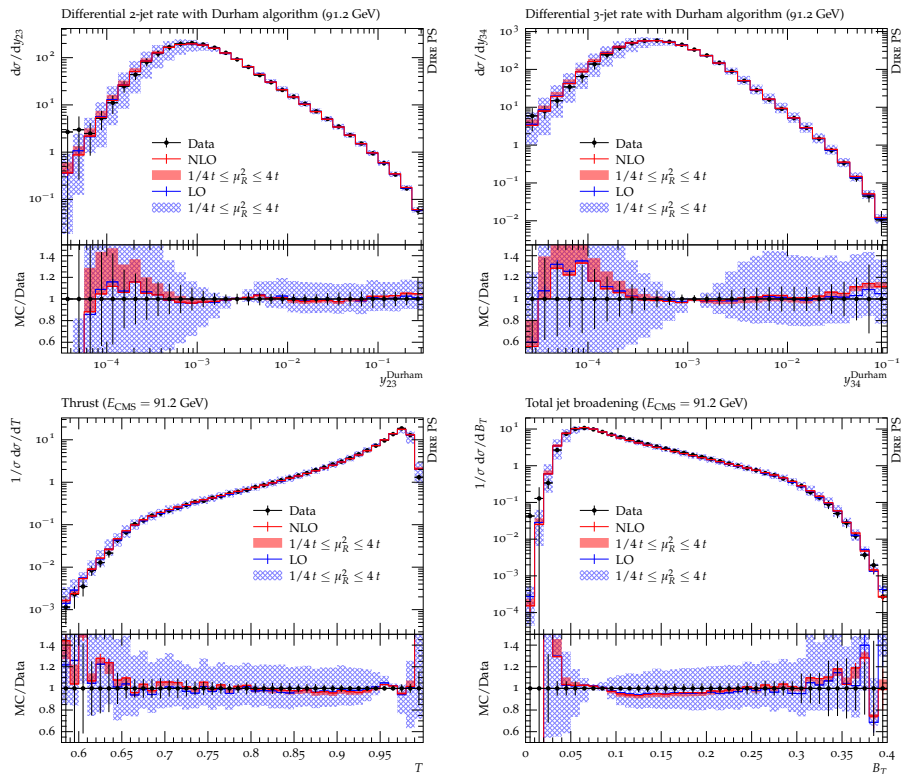




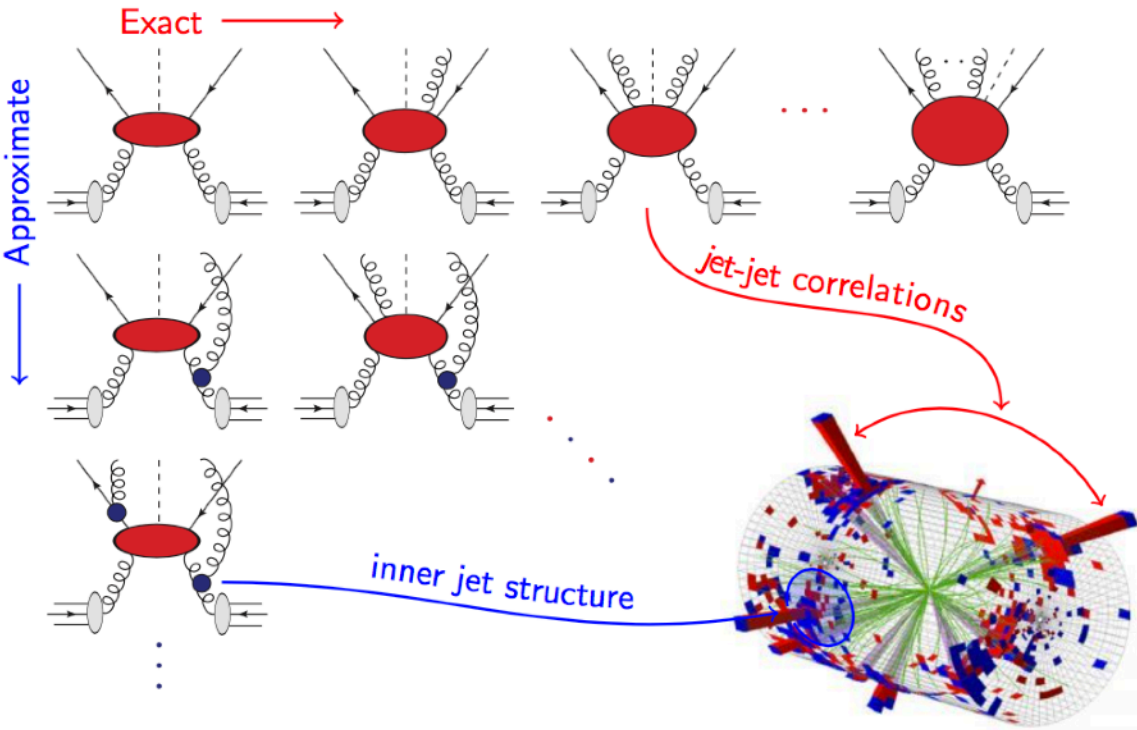
Recast  
Folding

## First phenomenological predictions

[Krauss,Prestel,SH] arXiv:1705.tonight



## Parton-shower matching & merging







# U.S. Particle Physics: Building for Discovery

*U.S. Particle Physics Strategy*

*Education and Outreach Site*

Five intertwined Science Drivers provide compelling lines of inquiry that show great promise for discovery.



*Use the Higgs boson as a new tool for discovery.*



*Pursue the physics associated with neutrino mass.*



*Identify the new physics of dark matter.*



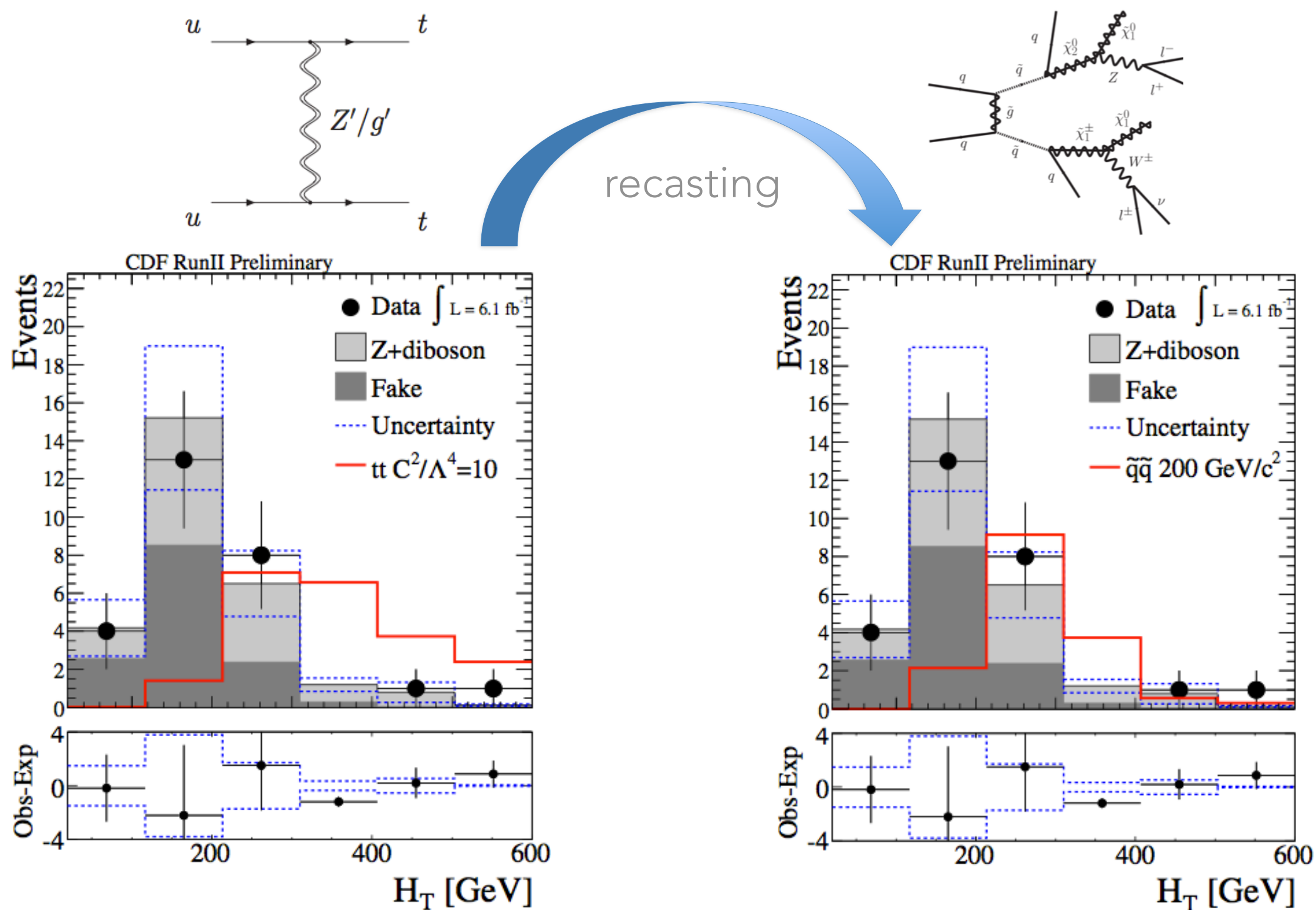
*Understand cosmic acceleration: dark energy and inflation.*



*Explore the unknown: new particles, interactions, and physical principles.*



# RECASTING / REINTERPRETATION



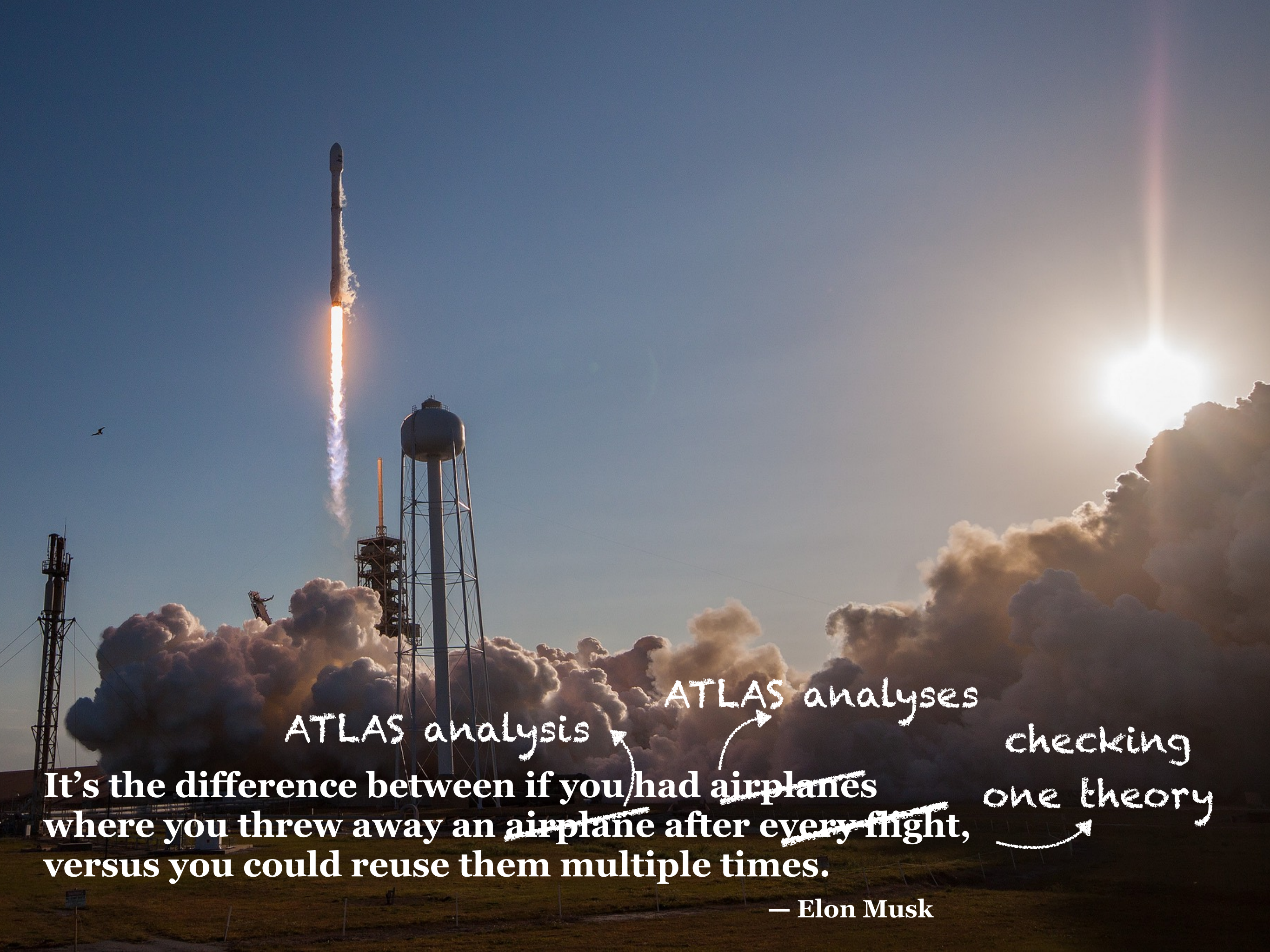




**It's the difference between if you had airplanes  
where you threw away an airplane after every flight,  
versus you could reuse them multiple times.**

**— Elon Musk**





ATLAS analysis

ATLAS analyses

checking  
one theory

It's the difference between if you had airplanes  
where you threw away an airplane after every flight,  
versus you could reuse them multiple times.

— Elon Musk



# DEMAND

>50 requests from theorists within ~2 months

RECAST [beta]

Feedback

Login

Register

Home

Analyses Catalog


Requests

About

Developers

News

Help



Search

About this site

RECAST is a framework for extending the impact of existing analyses performed by high-energy physics experiments.

1. Anyone can add *analyses* to the Analysis Catalog

2. Anyone can upload alternative signals in the LHE format and *request* that any given analysis is "recast" for their alternative model (Note: this is a request, there is no obligation for the experiments to respond.)

3. Anyone can *subscribe* to an analysis to be informed of activity associated with the analysis

4. Experimentalists can accept the request, process these alternative signals with the full simulation, reconstruction, and analysis selection. If they are authorized by their collaboration, then they can respond with an authoritative *result* for the selection efficiency and cross-section limits for the alternative signal. Note, anyone can provide a non-authoritative result, for instance one based on fast simulation.

Latest Requests

Request	Analysis	Model	Status
1603.0045	Search for massive supersymmetric particles decaying to many jets using the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV	asdafa	Incomplete
1505.0044	Search for massive supersymmetric particles decaying to many jets using the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV	stealth supersymmetry	Incomplete
1408.0043	Search for direct third-generation squark pair production in final states with missing transverse momentum and two b-jets in $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector	3-body decay of sbottom into b-quark and two invisible final states	Active
1408.0042	Search for direct top-squark pair production in final states with two leptons in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector	3-body decay of stop into top and two invisible final states	Active
1408.0041	Search for direct pair production of the top squark in all-hadronic final states in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector	3-body decay of stop into top and two invisible final states	Active
	Search for direct third-generation squark pair	3-body decay of	



# PHENO RECASTING SOFTWARE

- Several tools being developed by phenomenologists to address the need for an organized approach to recasting (but using unofficial and/or approximate methods).


Sabine Kraml

arXiv:1407.3278

- ATOM
- FastLim
- MadAnalysis
- Gambit
- SModelS
- XQCAT
- CheckMate
- unofficial contributions to Rivet

- As I'll show, it is possible to interface RECAST infrastructure with these unofficial pheno recasting tools.

**Towards a public analysis database**



We think it would be of great value for the whole community to have a database of LHC analyses based on fast simulation.

→ we propose to create such a database using the MadAnalysis 5 framework

- Validated analysis codes, easy to check and to use for everybody.
- Can serve for the interpretation of the LHC results in a large variety of models.
- Convenient way of documentation; helps long-term preservation of the analyses performed by ATLAS and CMS.
- Modular approach, easy to extend, everybody who implements and validates an existing ATLAS or CMS analysis can publish it within this framework.
- Provides feedback to the experiments about documentation and use of their results. (The ease with which an experimental analysis can be implemented and validated may actually serve as a useful check for the experimental collaborations for the quality of their documentation.)

Towards a public analysis database ... Aug 21, 2014 6

## Forum on the Interpretation of the LHC Results for BSM studies

The quest for new physics beyond the Standard Model is arguably the driving topic for Run 2 of the LHC. Indeed, the LHC collaborations are pursuing searches for new physics in a vast variety of channels. While the collaborations typically provide themselves interpretations of their results, for instance in terms of simplified models, **the full understanding of the implications of these searches requires the interpretation of the experimental results in the context of all kinds of theoretical models.** This is a very active field, with close theory-experiment interaction and with several public tools being developed.

With this forum, we want to provide a platform for continued discussion of topics related to the BSM (re)interpretation of LHC data, including the development of the necessary **public** [RecastingTools](#) and related infrastructure.

If you have questions or want to contribute, contact Sabine Kraml, [sabine.kraml@gmail.com](mailto:sabine.kraml@gmail.com), or any of the topical contacts given below.

## Meetings

### Meetings of this forum

- [2nd workshop](#), 12-14 Dec 2016 at CERN
  - [Agenda](#) | [introduction](#) | [final discussion](#) | [WorkshopSummaryNotes](#)
- **Kick-off workshop: (Re)interpreting the results of new physics searches at the LHC**, 15-17 June 2016 at CERN
  - [Agenda](#) | [general discussion](#) | [KickoffSummaryNotes](#)

### Other workshops, potentially interesting for our forum

- The [Les Houches PhysTev2017 workshop](#) will have a strong activity on interpreting LHC results; the BSM session in LH is taking place 14-23 June 2017.
- 2nd [LHC Long-Lived Particle workshop](#), CERN, 24-26 April 2017, "to address the status and future of beyond-the-Standard Model LLP searches at the ATLAS, CMS, and LHCb experiments, as well as auxiliary LHC detectors and projects".
- 6th edition of the workshop "[Implications of LHCb measurements and future prospects](#)", CERN, 12-14 October 2016. NB participation is restricted to the members of the LHCb Collaboration, and of interested theorists.

## Mailing list

- CERN e-group: [info-LHC-interpretation@cern.ch](mailto:info-LHC-interpretation@cern.ch)
- To subscribe, go to <https://simba3.web.cern.ch/simba3/SelfSubscription.aspx?groupName=info-lhc-interpretation>



# A FLEXIBLE WORKFLOW MODEL

1) read/export json

2) interactive shell

3) projection  $(3 \times 3 \times 3) \rightarrow 1 \times 3$

4) dimension reduction

$3 \times 4 \times 5 \times 6 \rightarrow 4 \times 5 \times 6$

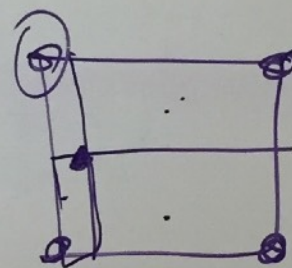
5) rebin one dimension

$3 \times 4 \times 5 \times 6 \rightarrow 3 \times 2 \times 5 \times 6$

6) interpolation

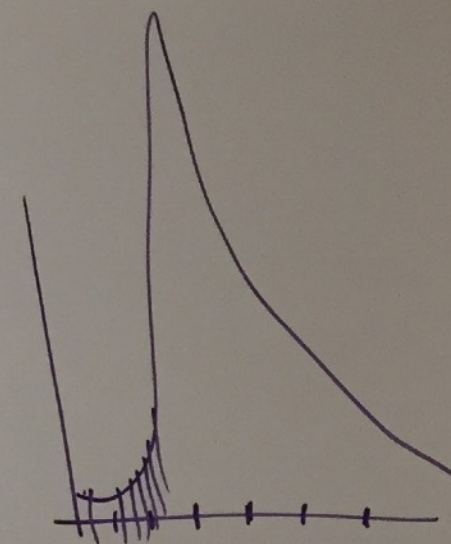
7) merging

8) visualization



TMD  
SF  
X-sec

SF1  
SF2



program

9) random generator  $(x_1, \dots, x_n)$

10) event generator  $e_i, \bar{e}_i$

$x, y, z, p_T, \phi, \eta$   
 $x, y, z, M_{\pi\pi}, p_T, \phi, \eta$

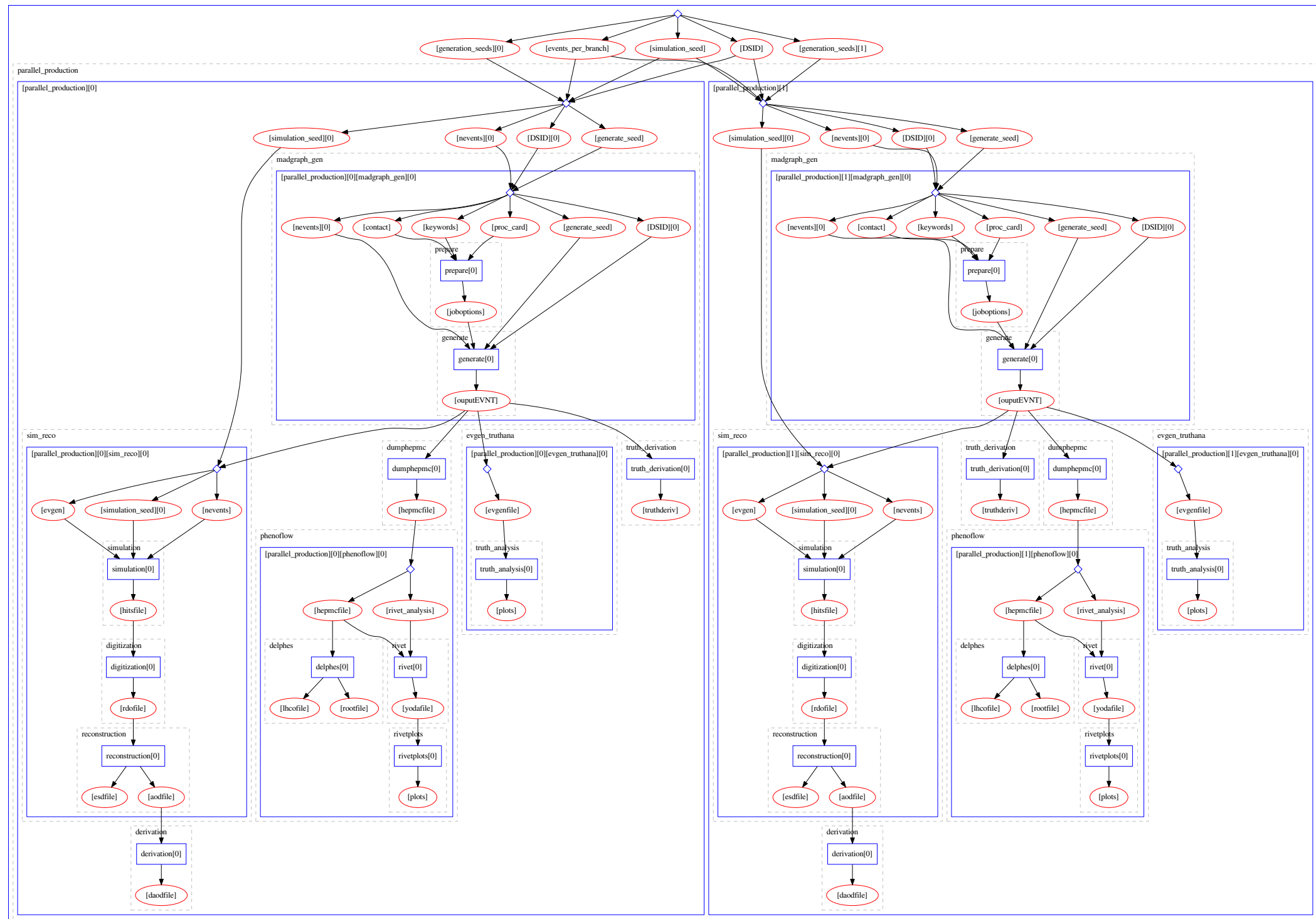
produce  
 $xs = ms.get("SF1", Q^2)$   
 $ms.get("XSES").setBin(Q^2, y, x, pT, xs)$

$Q^2, y, x, pT, xs$

```
main { mypioncalc
init() {
  ms.load("SF1")
  ms.load("SF2")
  ms.load("SF3")
  ms.create(7, ...)
}
```

# A FLEXIBLE WORKFLOW MODEL

A workflow composed of sub-workflows that run Rivet, Delphes, and ATLAS analyses in parallel on the same input

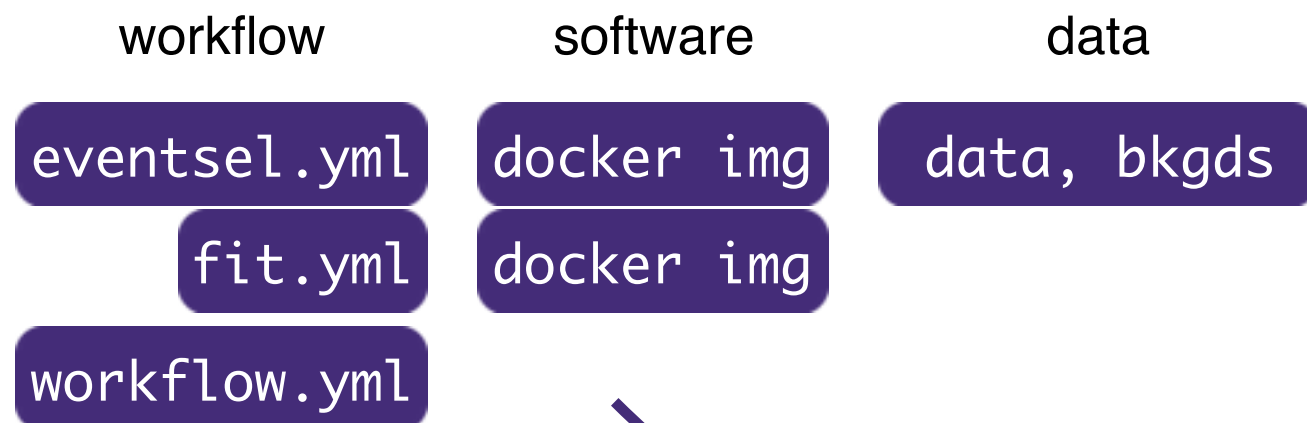




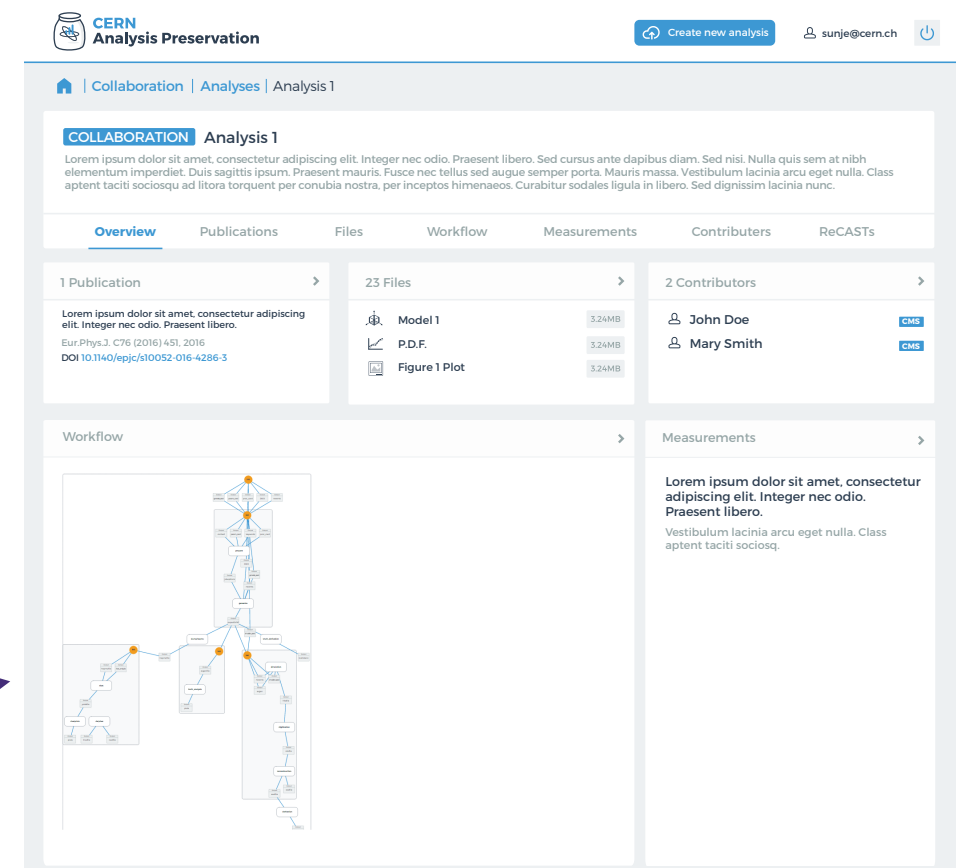
## Technical Solution:

Workflow (i.e. logic which steps to run in which order: reconstruction → analysis → fit)

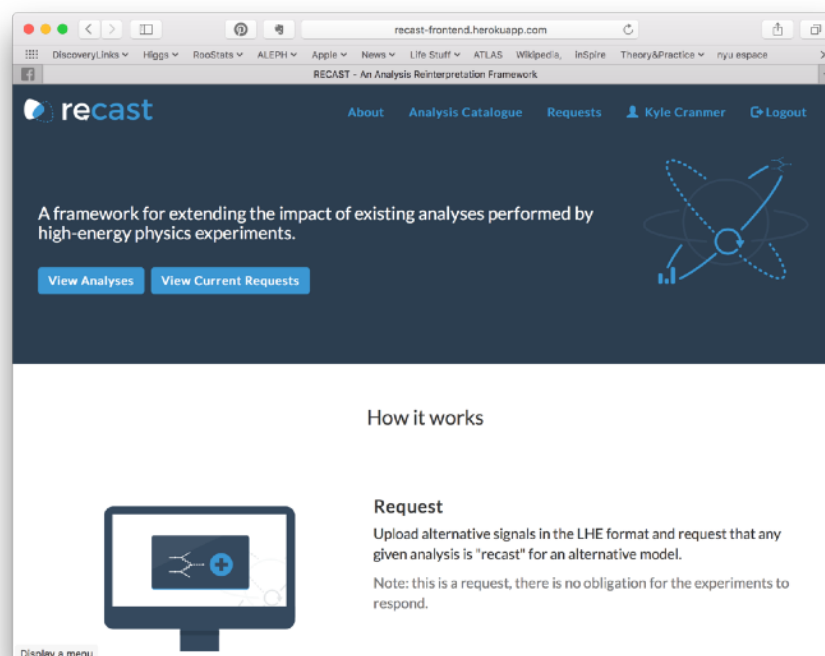
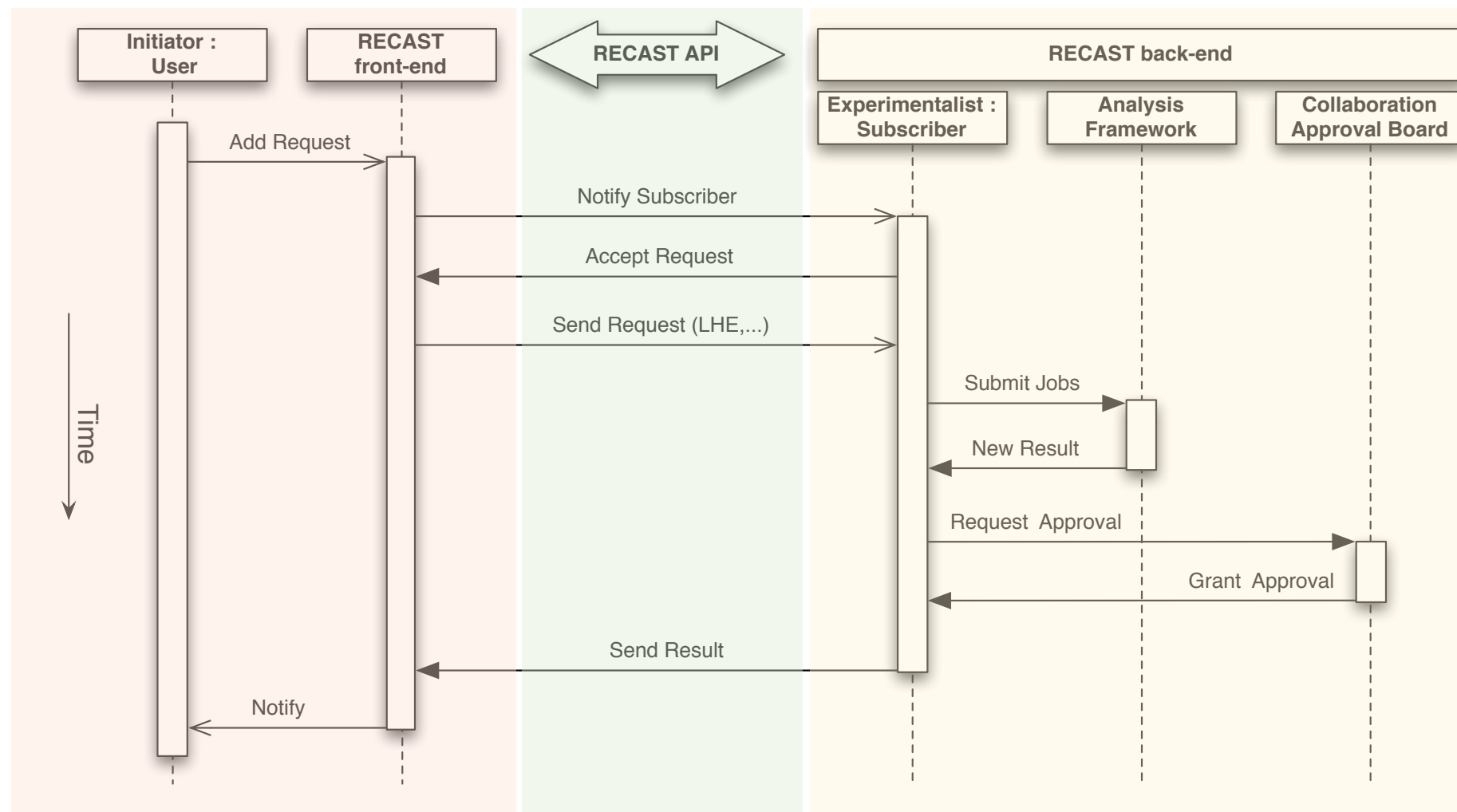
- in easy to write / read text based format (YAML)
- generic workflow language “**yadage**” based on graphs. No assumption on how you run your analysis. Should be able to accommodate your workflows.
- integrated into CERN Analysis Preservation.
- re-run workflow using tool that interprets info stored in CAP



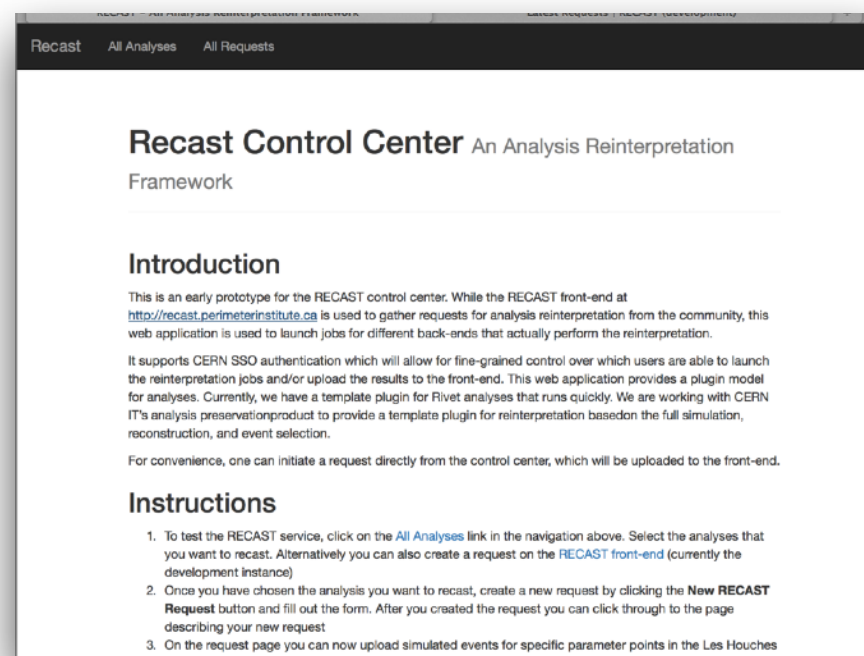
import analysis  
workflow



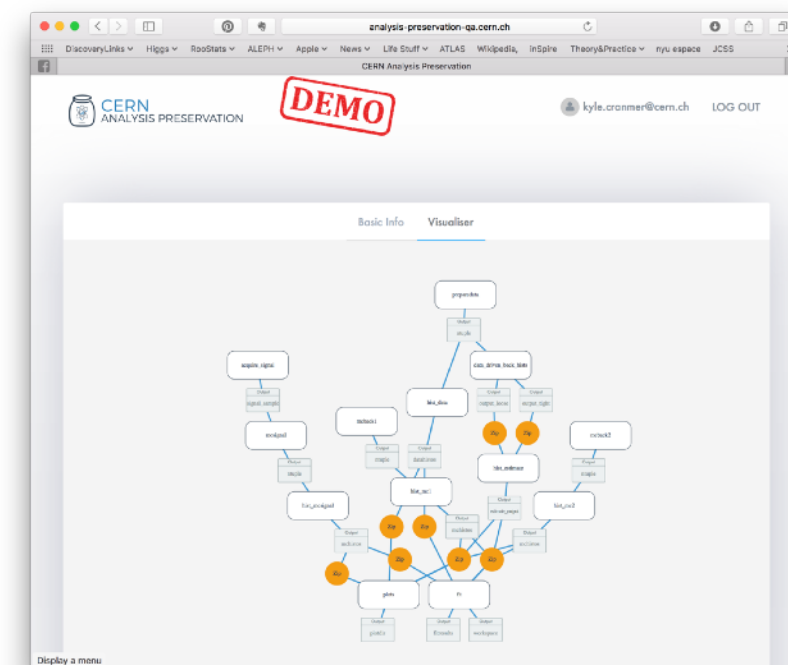




Front-End: public facing  
collects requests



Control Center: not public, uses CERN auth., oversees processing of jobs on back-end



CERN Analysis Preservation:  
Stores workflows, provides back-end  
computing resources

# front-end (open)

Home » Analyses Catalog » Demo with working rivet-based back-end » List Requests » test for UCI

View Edit Edit Contact Requester Show Results Devel

**1. request initiated**

**Analysis:** Demo with working rivet-based back-end

**Status:** Completed

**Requester:** lheinric

**Recast Audience:** all

**Model Name:** CMSSM

**Selected Subscriber(s):** lheinric, cranmer

Mon, 02/02/2015 - 14:26 - Activated  
Wed, 02/04/2015 - 03:06 - Completed

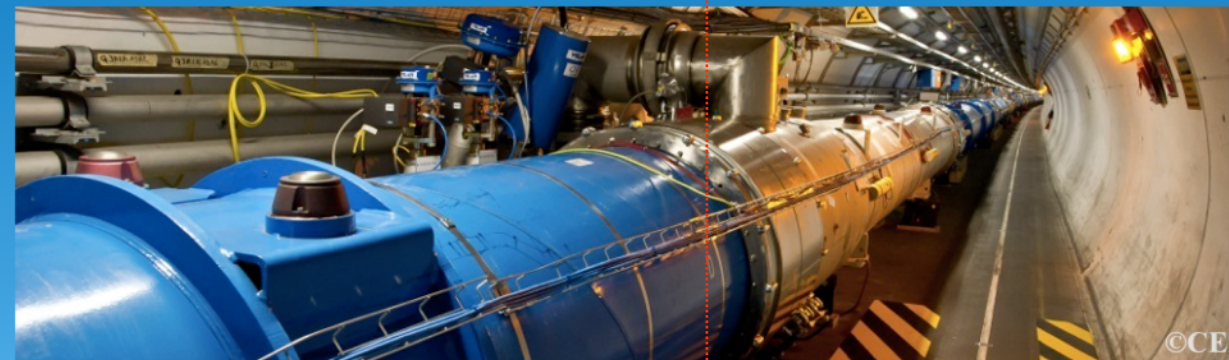
**Request Description and Potential**

**Reason for request:**  
because we can

**Additional Information:**  
No information available

**5. response public**

Home Analyses Catalog Requests My Subscriptions About Developers News Help



Home » Analyses Catalog » Demo with working rivet-based back-end » List Requests » test-upload-2 » Show Results » Recast Response for Request #test-upload-2

**Recast Response for Request #test-upload-2**

View Edit Devel

Submitted by lheinric on Sun, 01/18/2015 - 09:54

**Request:** test-upload-2

**ROOT file with TH1:** 20150118095414b5872abo-1a2b-10a4-c154-5cead413be8f.zip

**Status:** Completed

# control center (closed to experiment)

**Recast Request** test for UCI

**Request Details**

**analysis** Demo with working rivet-based back-end

**status** 1

**model-type** None

**uuid** 4cdc558c-8f4a-eab4-fdbf-cd91a34db4b2

**new-model-information** None

**title** test for UCI

**predefined-model** CMSSM

**reason-for-request** because we can

**requestor** lheinric

**audience** None

**subscribers** lheinric

**additional-information** None

**4. upload response**

+Add Parameter Point Upload to RECAST

Parameter	Description	Number of Events	Cross-Section
parameter-0	test for UCI	1000	20

**2. process request**

process results

**3. review results**

**Results for request** 4cdc558c-8f4a-eab4-fdbf-cd91a34db4b2 - parameter-0

**Efficiency**

0.18272727272727274

**Plots**

> MET: 

> PhotonPt: 

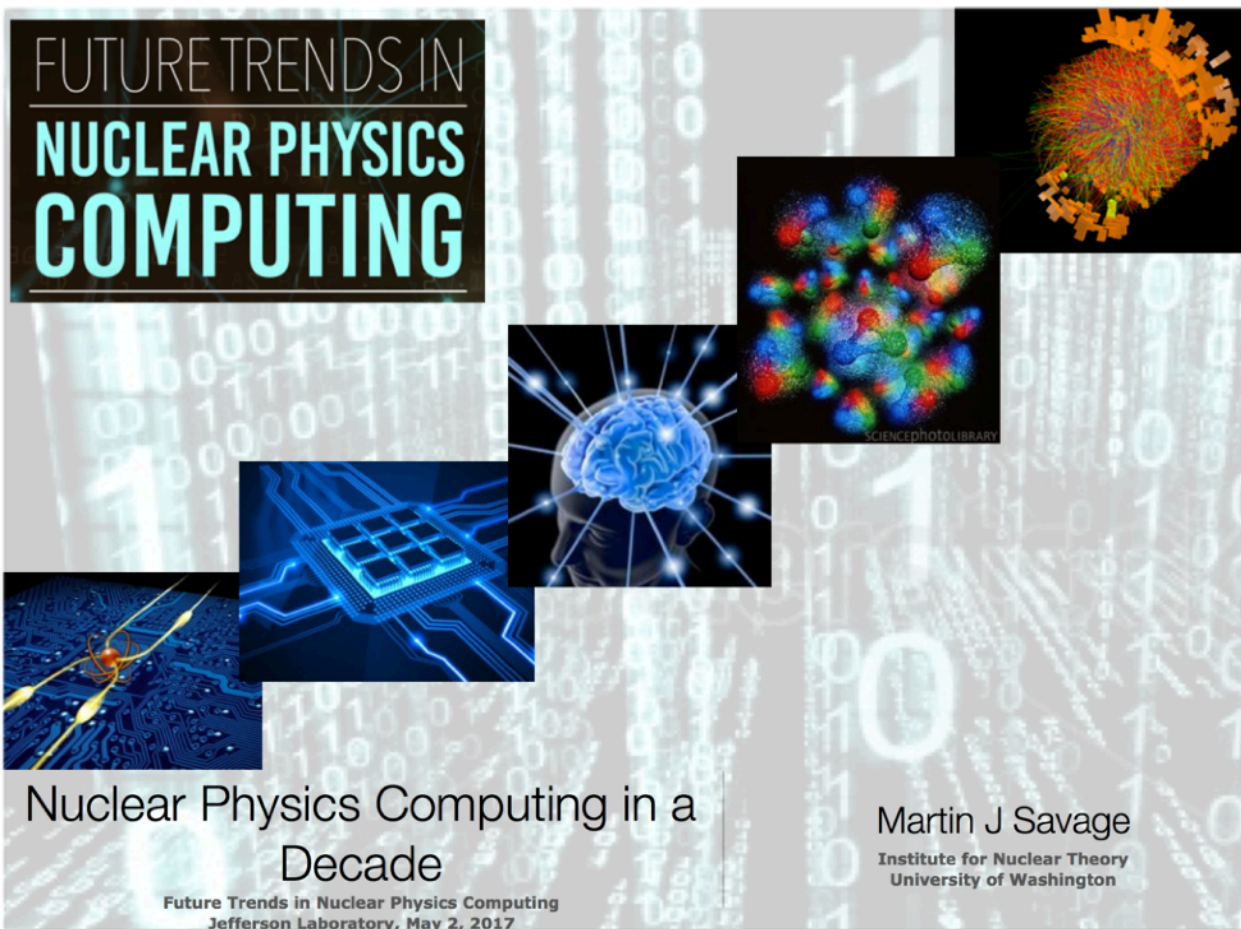
> Cutflow: 

> PhotonEta: 

# Abstracting the Problem for the Machine Learning Community



# FUTURE TRENDS IN NUCLEAR PHYSICS COMPUTING



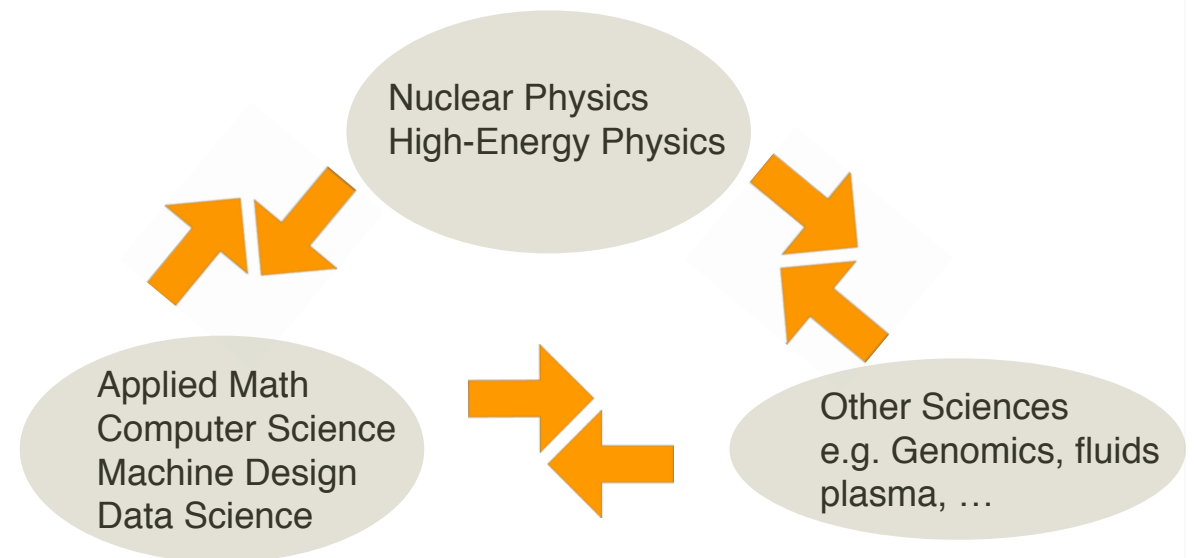
Nuclear Physics Computing in a Decade

Future Trends in Nuclear Physics Computing  
Jefferson Laboratory, May 2, 2017

Martin J Savage  
Institute for Nuclear Theory  
University of Washington

## FUTURE TRENDS IN NUCLEAR PHYSICS COMPUTING

### Algorithms From an Idea, through Development, to Production - Critical Element



## FUTURE TRENDS IN NUCLEAR PHYSICS COMPUTING

### New (disruptive?) Technologies

#### Quantum Computing

- NP has some computations that require exponential time on a classical computer - quantum computer?
- Microsoft, Google, IBM indicate ~50 qubit QC's up within a year
- NP could start exploring ....



#### Quantum Testbeds Stakeholder Workshop

Mayflower Hotel  
1127 Connecticut Avenue NW  
Washington, DC 20036  
February 14 - 16, 2017



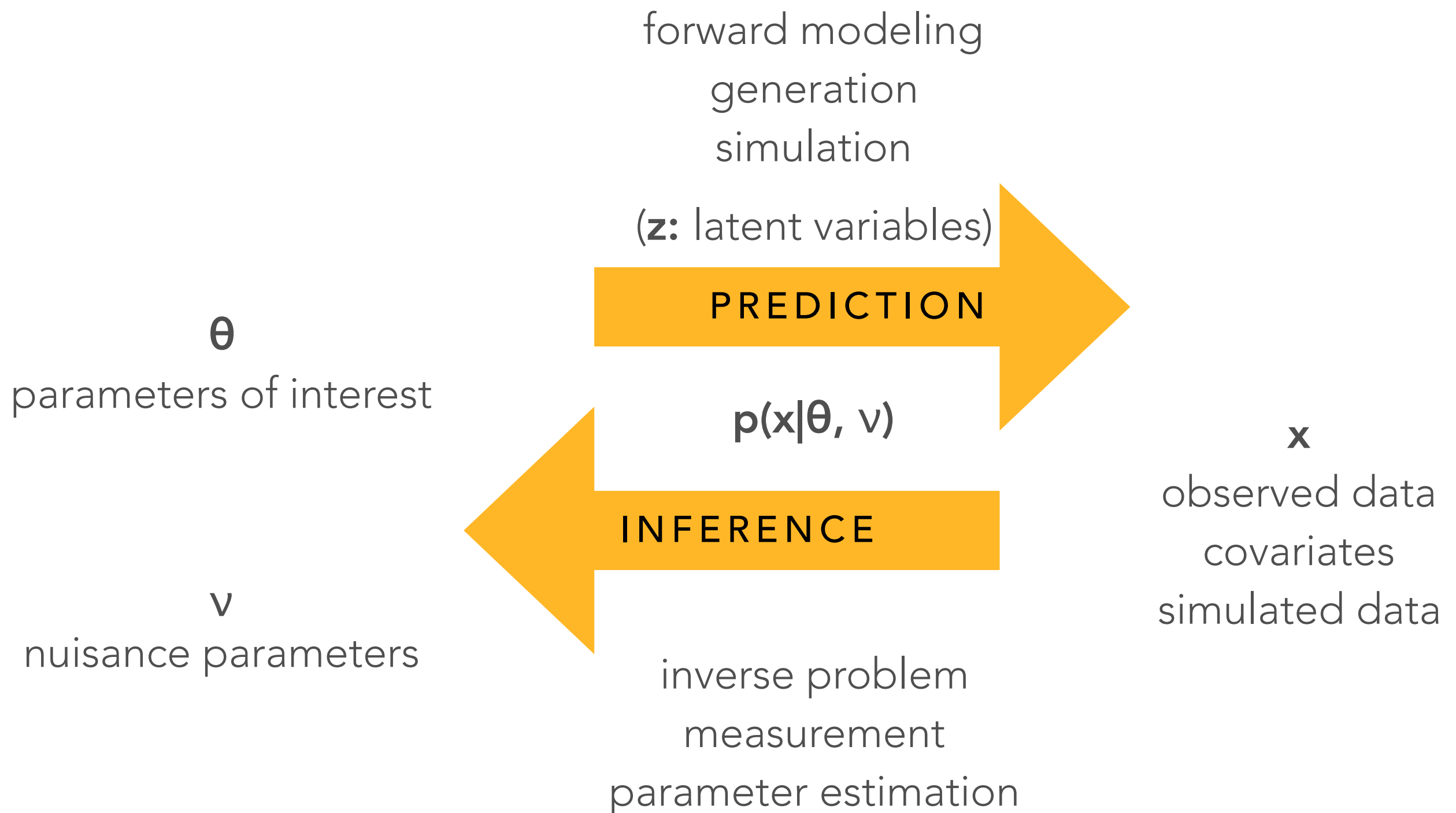
#### Machine Learning

- Already in use in HEP expt at some level
- Sophisticated pattern recognition - correlations in data
- Nonlinear regression on steroids+epo
- Some NP engagement, and much more, would likely be valuable





# THE PLAYERS



# THE FORWARD MODEL

1) We begin with Quantum Field Theory

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

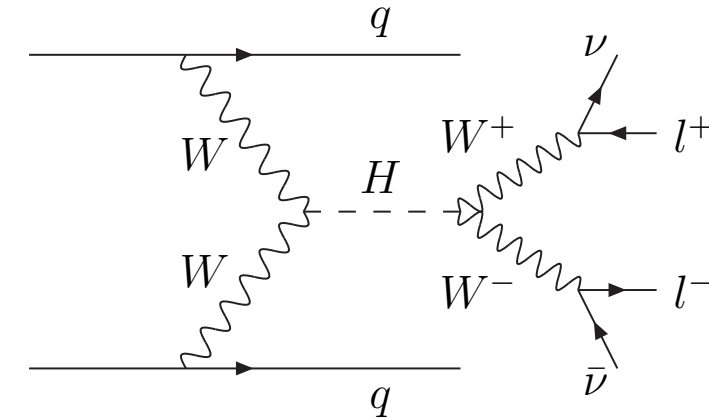
# THE FORWARD MODEL

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

1) We begin with Quantum Field Theory

2) Theory gives detailed prediction for high-energy collisions

hierarchical:  $2 \rightarrow \mathcal{O}(10) \rightarrow \mathcal{O}(100)$  particles



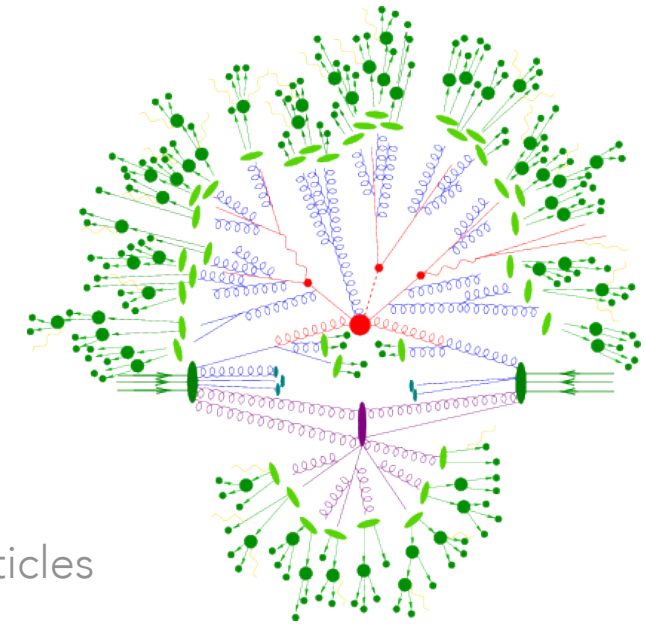
# THE FORWARD MODEL

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

1) We begin with Quantum Field Theory

2) Theory gives detailed prediction for high-energy collisions

hierarchical:  $2 \rightarrow \text{O}(10) \rightarrow \text{O}(100)$  particles



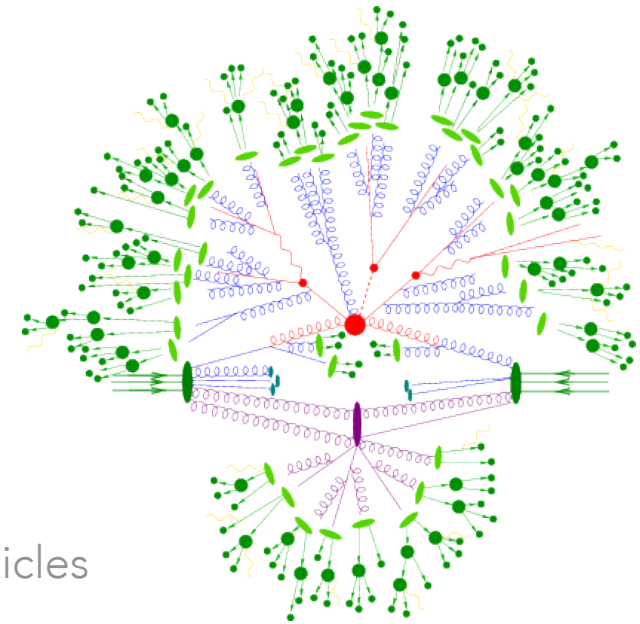
# THE FORWARD MODEL

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

1) We begin with Quantum Field Theory

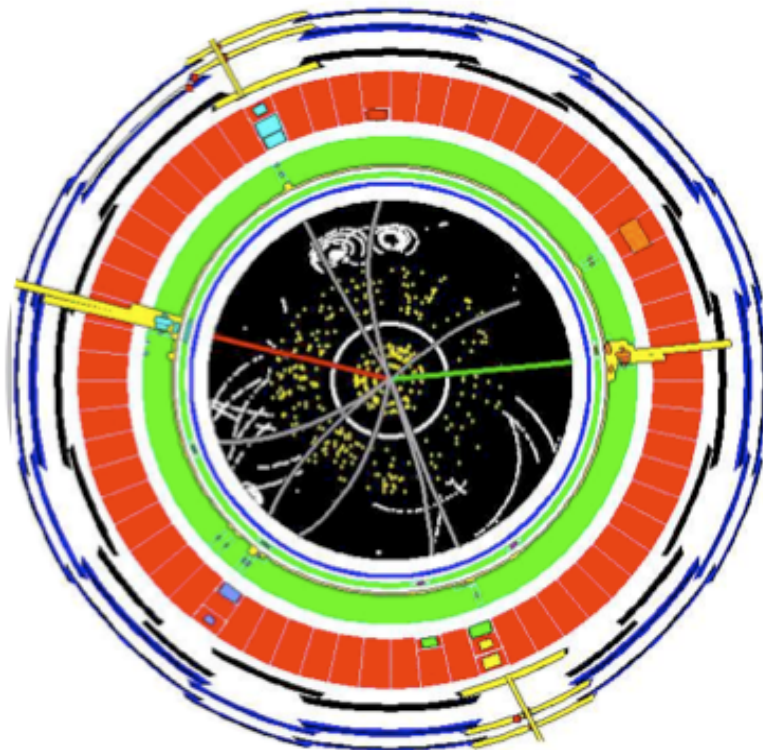
2) Theory gives detailed prediction for high-energy collisions

hierarchical:  $2 \rightarrow \mathcal{O}(10) \rightarrow \mathcal{O}(100)$  particles



3) The interaction of outgoing particles with the detector is simulated.

>100 million sensors





# THE FORWARD MODEL

$$\mathcal{L}_{SM} =$$

$$\underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2}g'Y B_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'Y B_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

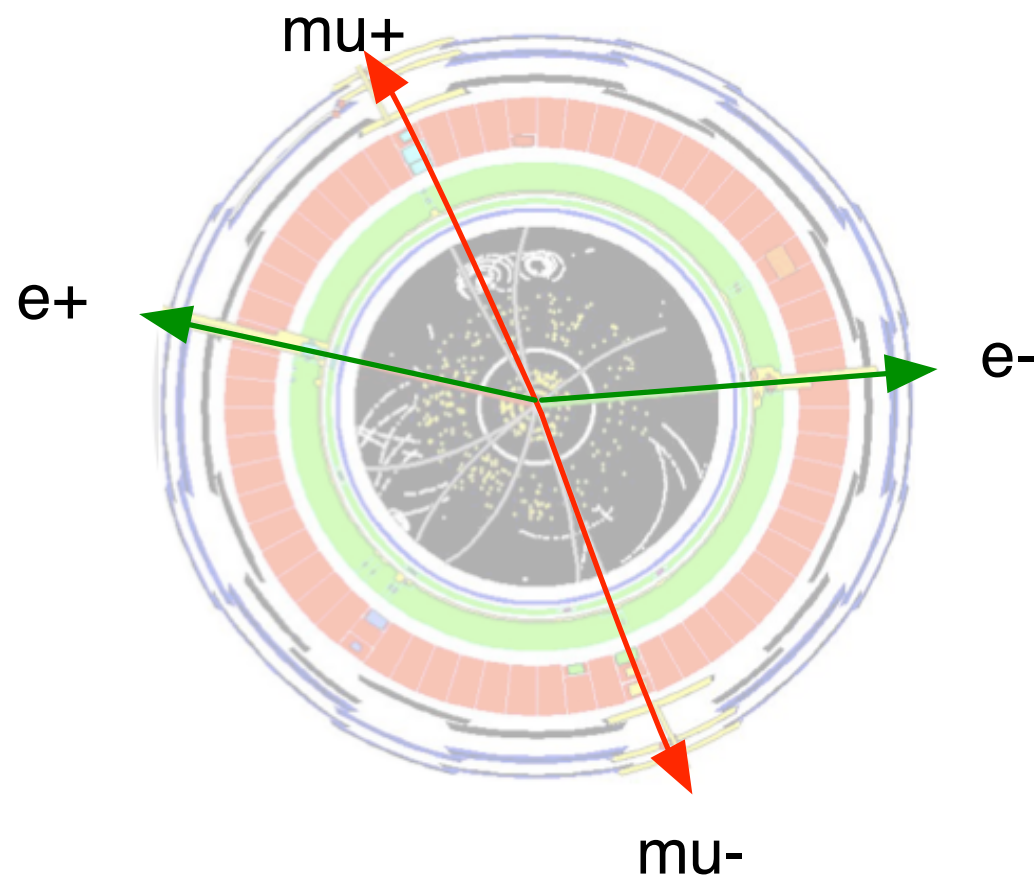
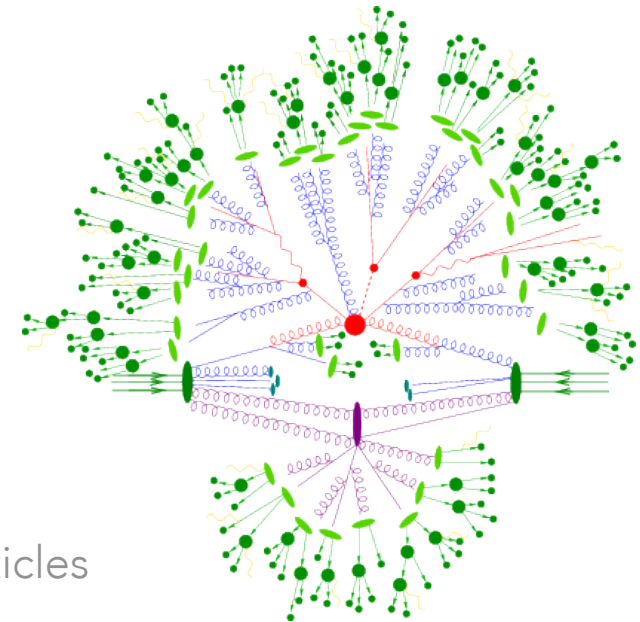
$$+ \underbrace{\frac{1}{2} |(i\partial_\mu - \frac{1}{2}g\boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2}g'Y B_\mu)\phi|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

1) We begin with Quantum Field Theory

2) Theory gives detailed prediction for high-energy collisions

hierarchical:  $2 \rightarrow \mathcal{O}(10) \rightarrow \mathcal{O}(100)$  particles



3) The interaction of outgoing particles with the detector is simulated.

>100 million sensors

4) Finally, we run particle identification and feature extraction algorithms on the simulated data as if they were from real collisions.

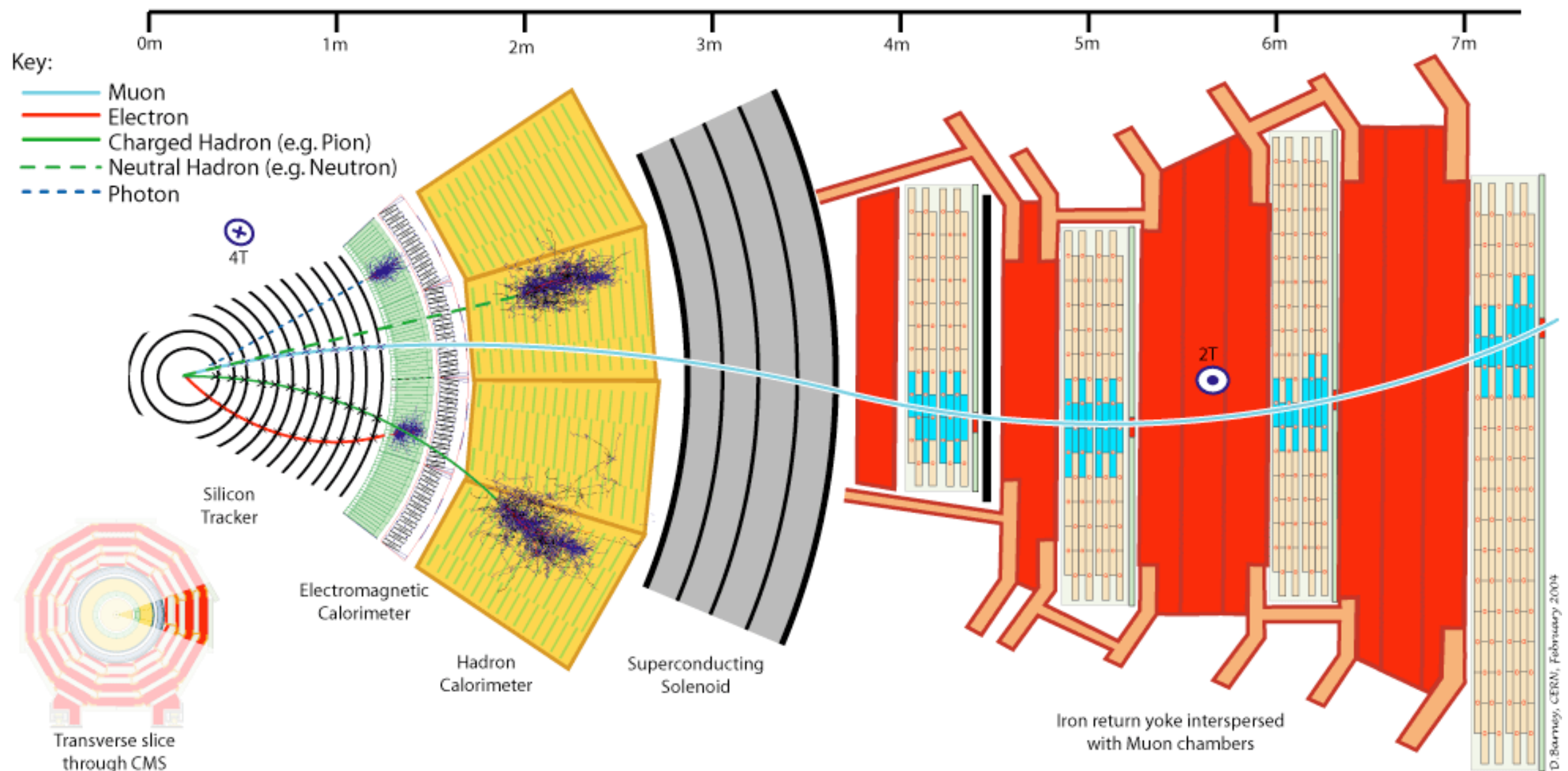
~10-30 features describe interesting part

# DETECTOR SIMULATION

**Conceptually:**  $\text{Prob}(\text{detector response} \mid \text{particles})$

**Implementation:** Monte Carlo integration over micro-physics

**Consequence:** evaluation of the likelihood is intractable



# DETECTOR SIMULATION

**Conceptually:**  $\text{Prob}(\text{detector response} \mid \text{particles})$

**Implementation:** Monte Carlo integration over micro-physics

**Consequence:** evaluation of the likelihood is intractable

This motivates a new class of algorithms for what is called **likelihood-free inference**, which only require ability to generate samples from the simulation in the “forward mode”

# A COMMON THEME

## ABC

resources on approximate  
Bayesian computational  
methods

 Search

Home

## Home

This website keeps track of developments in approximate Bayesian computation (ABC) (a.k.a. likelihood-free), a class of computational statistical methods for Bayesian inference under intractable likelihoods. The site is meant to be a resource both for biologists and statisticians who want to learn more about ABC and related methods. Recent publications are under Publications 2012. A comprehensive list of publications can be found under Literature. If you are unfamiliar with ABC methods see the Introduction. Navigate using the menu to learn more.

[ABC in Montreal](#)

[ABC in Montreal \(2014\)](#)

## ABC in Montreal

Approximate Bayesian computation (ABC) or likelihood-free (LF) methods have developed mostly beyond the radar of the machine learning community, but are important tools for a large and diverse segment of the scientific community. This is particularly true for systems and population biology, computational neuroscience, computer vision, healthcare sciences, but also many others.

Interaction between the ABC and machine learning community has recently started and contributed to important advances. In general, however, there is still significant room for more intense interaction and collaboration. Our workshop aims at being a place for this to happen.



# NIPS 2016

BARCELONA · SPAIN · DECEMBER 5 - 10, 2016 | <http://nips.cc/>

## TUTORIALS

**Deep Reinforcement Learning Through Policy Optimization**  
Pieter Abbeel (OpenAI, UC Berkeley) and John Schulman (OpenAI)

**Large-scale Optimization: Beyond Stochastic Gradient Descent and Convexity**  
Francis Bach (INRIA, ENS) and Suvrit Sra (MIT)

**Variational Inference: Foundations and Modern Methods**  
David Blei (Columbia), Shakir Mohamed (Google DeepMind) and Rajesh Ranganath (Princeton)

**Natural Language Processing for Computational Social Science**  
Cristian Danescu-Niculescu-Mizil (Cornell) and Lillian Lee (Cornell)

**Generative Adversarial Networks**  
Ian Goodfellow (OpenAI)

**Theory and Algorithms for Forecasting Non-stationary Time Series**  
Vitaly Kuznetsov (Google) and Mehryar Mohri (Courant Institute, Google Research)

**Deep Learning for Building AI Systems**  
Andrew Ng (Baidu, Stanford University)

**ML Foundations and Methods for Precision Medicine and Healthcare**  
Suchi Saria (Johns Hopkins) and Peter Schulam (Johns Hopkins)

**Crowdsourcing: Beyond Label Generation**  
Jenn Wortman Vaughan (Microsoft Research)

## INVITED SPEAKERS

**Reproducible Research: the Case of the Human Microbiome**  
Susan Holmes (Stanford University)

**Dynamic Legged Robots**  
Marc Raibert (Boston Dynamics)

**Intelligent Biosphere**  
Drew Purves (Google DeepMind)

**Predictive Learning**  
Yann LeCun (Facebook and New York University)

**Machine Learning and Likelihood-Free Inference in Particle Physics**  
Kyle Cranmer (New York University)

**Learning About the Brain: Neuroimaging and Beyond**  
Irina Rish (IBM T.J. Watson Research Center)

**Engineering Principles From Stable And Developing Brains**  
Saket Navlakha (The Salk Institute for Biological Studies)

## SYMPOSIA

**Recurrent Neural Networks and other Machines that Learn Algorithms**  
Alex Graves (Google DeepMind)  
Juergen Schmidhuber (IDSIA)  
Rupesh Srivastava (IDSIA)  
Sepp Hochreiter (Johannes Kepler University)

**Deep Learning**  
Navdeep Jaitly (Google)  
Roger Grosse (University of Toronto)  
Yann LeCun (New York University & Facebook)

**Machine Learning and the Law**  
Adrian Weller (Cambridge, Alan Turing Inst.)  
Conrad McDonnell (Gray's Inn Tax Chambers)  
Jatinder Singh (University of Cambridge)  
Thomas Grant (University of Cambridge)

## ORGANIZING COMMITTEE

**General Chairs:**  
Daniel D Lee (University of Pennsylvania)  
Masashi Sugiyama (The University of Tokyo)

**Program Chairs:**  
Ulrike von Luxburg (University of Tübingen)  
Isabelle Guyon (Clopinet)

**Tutorials Chair:**  
Joelle Pineau (McGill University)  
Hanna Wallach (Microsoft)

**Workshop Chairs:**  
Ralf Herbrich (Amazon)

**Demonstration Chair:**  
Raia Hadsell (Google DeepMind)

**Publications Chair & Electronic Proceedings Chair:**  
Roman Garnett (Washington University)

**Program Managers:**  
Krikamol Muandet (Mahidol University and MPI)  
Rohit Babbar, Behzad Tabibian (MPI for Intelligent Systems)

## PROGRAM COMMITTEE

Emmanuel Abbe, Princeton Univ.  
Alakh Agarwal, Microsoft  
Anima Anandkumar, UC Irvine  
Chloé-Agathe Azencott, MINES ParisTech  
Shai Ben-David, Univ. of Waterloo  
Alina Beygelzimer, Yahoo Research  
Jeff Bilmes, Univ. of Washington, Seattle  
Giles Blanchard, Univ. of Potsdam  
Matthew Blaschko, KU Leuven  
Tamara Broderick, MIT

Sebastian Bubeck, Princeton  
Alexandra Carpentier, Univ. Potsdam  
Miguel Carreira-Perpinán, UC Merced  
Kamalika Chaudhuri, UC San Diego  
Gal Chechik, Google, Bar-Ilan Univ.  
Kyunghyun Cho, New York Univ.  
Aaron Courville, Univ. of Montreal  
Koby Crammer, Technion  
Florence d'Alché-Buc, Telecom ParisTech  
Amak Dalalyan, ENSAE ParisTech  
Marc Deisenroth, Imperial College London  
Francesco Dinuzzo, Amazon

Finale Doshi-Velez, Harvard  
Ran El-Yaniv, Technion  
Hugo Jair Escalante, INAC  
Sergio Escalera, Univ. of Barcelona  
Maryam Fazeli, Univ. of Washington  
Aasa Feragen, Univ. of Copenhagen  
Rob Fergus, New York Univ.  
Koby Crammer, Technion  
Florence d'Alché-Buc, Telecom ParisTech  
Amak Dalalyan, ENSAE ParisTech  
Marc Deisenroth, Imperial College London  
Francesco Dinuzzo, Amazon

Lise Getoor, UC Santa Cruz  
Mark Girolami, Imperial College London  
Amir Globerson, Tel Aviv Univ.  
Yoav Goldberg, Bar-Ilan Univ.  
Manuel Gomez, Max Planck Institute  
Yves Grandvalet, Univ. of Compiègne & CNRS  
Moritz Grosse-Wenstrup, MPI  
Zaid Harchaoui, Univ. of Washington  
Moritz Hardt, Google  
Matthias Hein, Saarland Univ.  
Philipp Hennig, MPI IS Tübingen  
Frank Hutter, Univ. of Freiburg

Prateek Jain, Microsoft Research  
Navdeep Jaitly, Google Brain  
Stefanie Jegelka, MIT  
Samuel Kaski, Aalto Univ.  
Koray Kavukcuoglu, Google DeepMind  
Jens Kober, TU Delft  
Samory Kpotufe, Princeton Univ.  
Sanjiv Kumar, Google Research  
James Kwok, Hong Kong Univ.  
Simon Lacoste-Julien, U. of Montreal  
Christoph Lampert, IST Austria  
Hugo Larochelle, Twitter

François Laviolette, L'Université Laval  
Honglak Lee, Univ. of Michigan  
Christoph Lippert, Human Longevity  
Po-Ling Lo, UW-Madison  
Phil Long, Sentient Technologies  
Jakob Macke, Caesar Bonn  
Julien Mairal, Inria  
Shie Mannor, Technion  
Marta Mella, Univ. of Washington  
Claire Monteleoni, Google  
Washington Univ.  
Remi Munos, Google DeepMind

Guillaume Obozinski, Ecole Paris  
Cheng Soon Ong, Data61 and ANU  
Francesco Orabona, Stony Brook U.  
Fernando Perez-Cruz, Universidad Carlos III de Madrid, Bell Labs (Nokia)  
Jonathan Piliow, Princeton Univ.  
Doina Precup, McGill Montreal  
Allan Rakotomamonjy, Univ. of Rouen  
Manuel Rodriguez, Max Planck Inst.  
Romer Rosales, LinkedIn  
Lorenzo Rosasco, U. of Genova, MIT  
Sivan Sabato, Ben-Gurion Univ.

Mehreen Saeed, FAST, Univ. of CES  
Ruslan Salakhutdinov, CMU  
Purnamrita Sarkar, Univ. T. Austin  
Fei Sha, USC  
Omid Shariq Weizmann, Inst. of Science  
Jonathan Shiels, Google Brain  
David Sonntag, New York Univ.  
Suvrit Sra, MIT  
Karthik Sridharan, Cornell Univ.  
Bharath Sriperumbudur, Pennsylvania State Univ.  
Erik Sudderth, Brown Univ.

Csaba Szepesvári, Univ. of Alberta  
Graham Taylor, Univ. of Guelph  
Ambuj Tewari, Univ. of Michigan  
Ruth Utner, MPI Tübingen  
Benjamin Van Roy, Stanford  
Jean-Philippe Vert, MINES ParisTech  
Bob Williamson, Data61 and ANU  
Jennifer Wortman, Vaughan  
Microsoft Research  
Lin Xiao, Microsoft Research  
Kun Zhang, CMU





# Conclusions

# CONCLUSIONS

A large portion of the scientific process involves interaction between theory and experiment

- The scientific process does not end when the results of an analysis are published in a paper!
- this requires a different type of infrastructure
- often neglected or addressed post-facto

I recommend targeted approach to theory/experiment interface

- An open-ended approach often gets bogged down
- start by identifying problems of scientific value with limited scope

Reusable & composable workflows are incredibly helpful and recent technology from industry makes it possible



Backup

# EXAMPLE RECAST → HEPDATA / ZENODO

After re-running analysis on new physics model, experiments might want to push result of new interpretation to HEPData. Technically we can do this with Zenodo. Discussing with HEPData and INSPIRE to have API connection to upload result. Both are based on Invenio, so should be easy.

- this allows for new results to get a DOI and be associated with the original analysis publication

The screenshot shows a web browser window displaying a Zenodo record. The URL is <https://sandbox.zenodo.org/record/84#.VUESk9NVhBc>. The record is titled "recast request response 3ee4bfde-739b-c844-99bc-f00b130e1ee3" and was created by "Heinrich, Lukas" on 29 April 2015. It is marked as a "Dataset" with "Embargoed access". The record is embargoed, with files available as "Open Access" after 01 January 2016. The DOI is 10.5072/zenodo.84, and the license is Creative Commons CCZero. The record is uploaded by lukasheinrich on 29 April 2015.

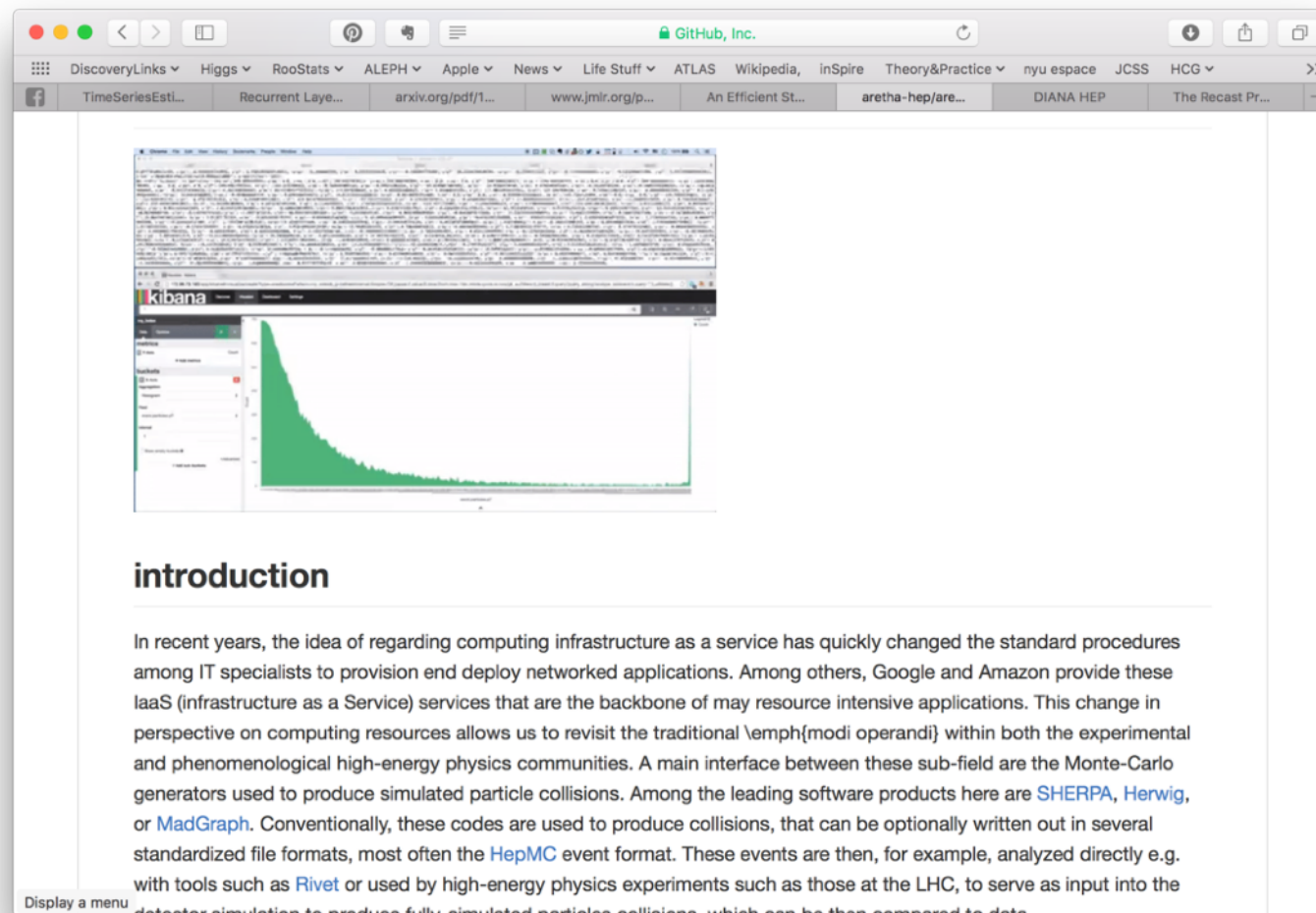
The main content area shows a preview of a document. The document is titled "Cutflow" and contains a plot of the cutflow for a Rivet analysis. The x-axis is labeled "Step" and includes the following steps: all,  $n_\gamma = 2$  crack,  $\eta_{max}$ ,  $m_{\gamma\gamma}$ ,  $p_{T,min}$ ,  $p_{T,1}$ ,  $p_{T,2}$ ,  $p_{T,\gamma\gamma}$ , and  $E_T^{miss}$ . The y-axis is labeled "Cutflow" and has a scale of  $10^3$ . The plot shows a red line representing the cutflow, which starts at a high value and decreases stepwise as the cuts are applied, ending at a lower value for the final cut.

On the right side of the page, there is a "Share" section with social media icons (Twitter, Facebook, Email, etc.) and a "Cite as" section with the citation: Heinrich, Lukas. (2015). recast request response 3ee4bfde-739b-c844-99bc-f00b130e1ee3. Zenodo. 10.5072/zenodo.84. Below the citation is a dropdown menu to "Select citation style...". At the bottom right, there is an "Export" section with links to BibTeX, DataCite, DC, EndNote, NLM, RefWorks, MARC, and MARCXML.

# "NETFLIX FOR MONTE CARLO"

Lukas has prototyped a web service called Aretha that encapsulates Monte Carlo tools and wraps them as a web service.

- Specific version of "cards" configuring Monte Carlo generator
- specific installation (stored in a docker container) that ensures version of generator and other dependencies (compiler etc.)



<https://github.com/aretha-hep/aretha-doc>

- ideally, give DOIs to the generator cards and docker container
- can generate more consistent MC on demand



# LIKELIHOOD FREE INFERENCE

# DeepMind and OpenAI releasing simulators to train advanced machine learning models

Similarly, HEP could / should release simulation tools and analysis workflows to produce training data for ML.



Also, new “likelihood-free” inference techniques (like Approximate Bayesian Computation) require running the simulation




*A sample of Universe game environments played by human demonstrators.*


# Create standalone simulation tools to facilitate collaboration between HEP and machine learning community


By [Pierre Baldi](#), [Peter Sadowski](#), [Daniel Whiteson](#), [Christian Lorenz Müller](#), [Michael Williams](#), [Lukas Heinrich](#), [Steven Schramm](#), [Maurizio Pierini](#), [Sergei Gleyzer](#), [Amir Farbin](#), [jean-roch vlimant](#), [Tim Head](#), [Juan Pavez](#), [Peter Elmer](#), [Balázs Kégl](#), [Andrey Ustyuzhanin](#), [Vladimir Gligorov](#), [Gilles Louppe](#), [Kyle Cranmer](#)


Kyle Cranmer · [Sign out](#)

## Actions


 1 vote

 Hide

 Collect

 Share

29

 Tweet

9

## Authors

[Pierre Baldi](#), [Peter Sadowski](#), [Daniel Whiteson](#), [Christian Lorenz Müller](#), [Michael Williams](#), [Lukas Heinrich](#), [Steven Schramm](#), [Maurizio Pierini](#), [Sergei Gleyzer](#), [Amir Farbin](#), [jean-roch vlimant](#), [Tim Head](#), [Juan Pavez](#), [Peter Elmer](#), [Balázs Kégl](#), [Andrey Ustyuzhanin](#), [Vladimir Gligorov](#), [Gilles Louppe](#), [Kyle Cranmer](#)

## Metadata

DOI [10.5281/zenodo.46864](#)

Published: 26 Feb, 2016



[dslhc](#) [machinelearning](#) [datascience](#) [open data](#) [simulation](#)

Discussions at recent workshops have made it clear that one of the key barriers to collaboration between high energy physics and the machine learning community is access to training data. Recent successes in data sharing through the [HiggsML](#) and [Flavours of Physics](#) Kaggle challenges have borne much fruit, but required significant effort to coordinate.

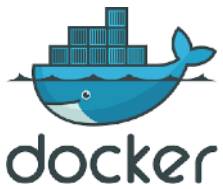
While static simulated datasets are useful for challenges, in the course of investigating new machine learning techniques it is advantageous to be able to generate training data on demand (e.g. Refs. [1](#), [2](#), [3](#)).

Therefore we recommend efforts be made to produce the ingredients required to facilitate such collaboration:

- Specific challenges for HEP experiments should be fully specified such that minimal domain-specific knowledge is required to attack them.
- Stand-alone simulators should be made open source. They should be developed to be easy to use without domain-specific expertise, while still being representative of real experimental challenges. Such a simulation will permit non-HEP researchers to generate realistic HEP datasets for training and testing. These simulators could range from truth-level simulation of a hard scattering to fast simulation like [Delphes](#), to full [GEANT4](#) simulation of sensor arrays.
- Performance metrics (objective functions) and operational constraints should be defined to evaluate proposed solutions.



# ENCAPSULATING THE SIMULATION



<https://github.com/lukasheinrich/weinberg-test>

## README.md

### Run HEP workflows from the web.

by [Kyle Cranmer](#) and [Lukas Heinrich](#)

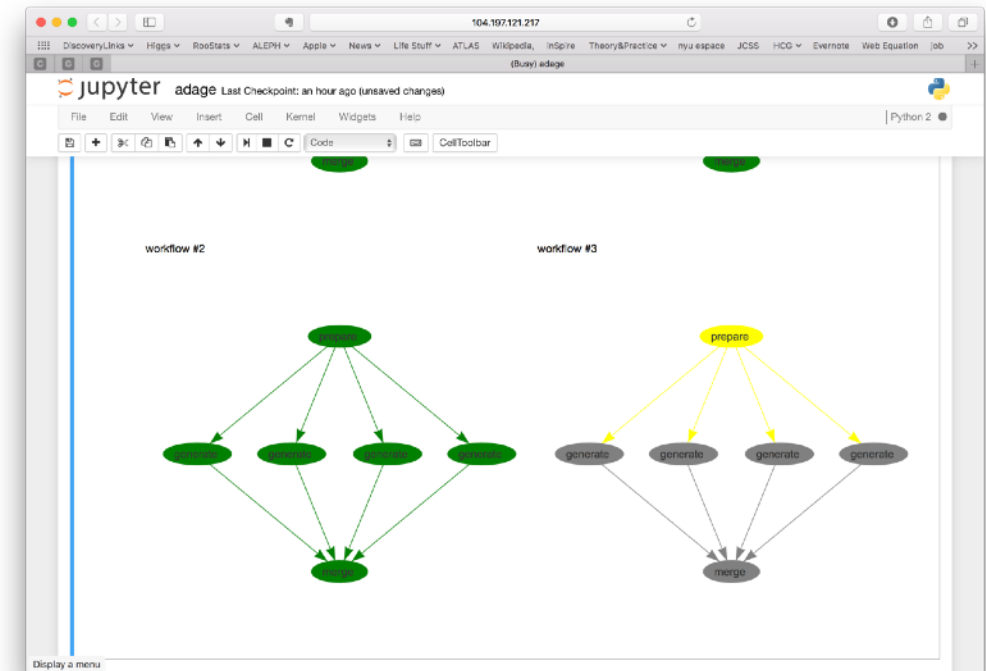
An example notebook on how to generate simulated high energy physics collision events using the generator package MadGraph. Simulated datasets obtained from this notebook can then be used to train and evaluate the performance of generative models for physics.

### Usage:

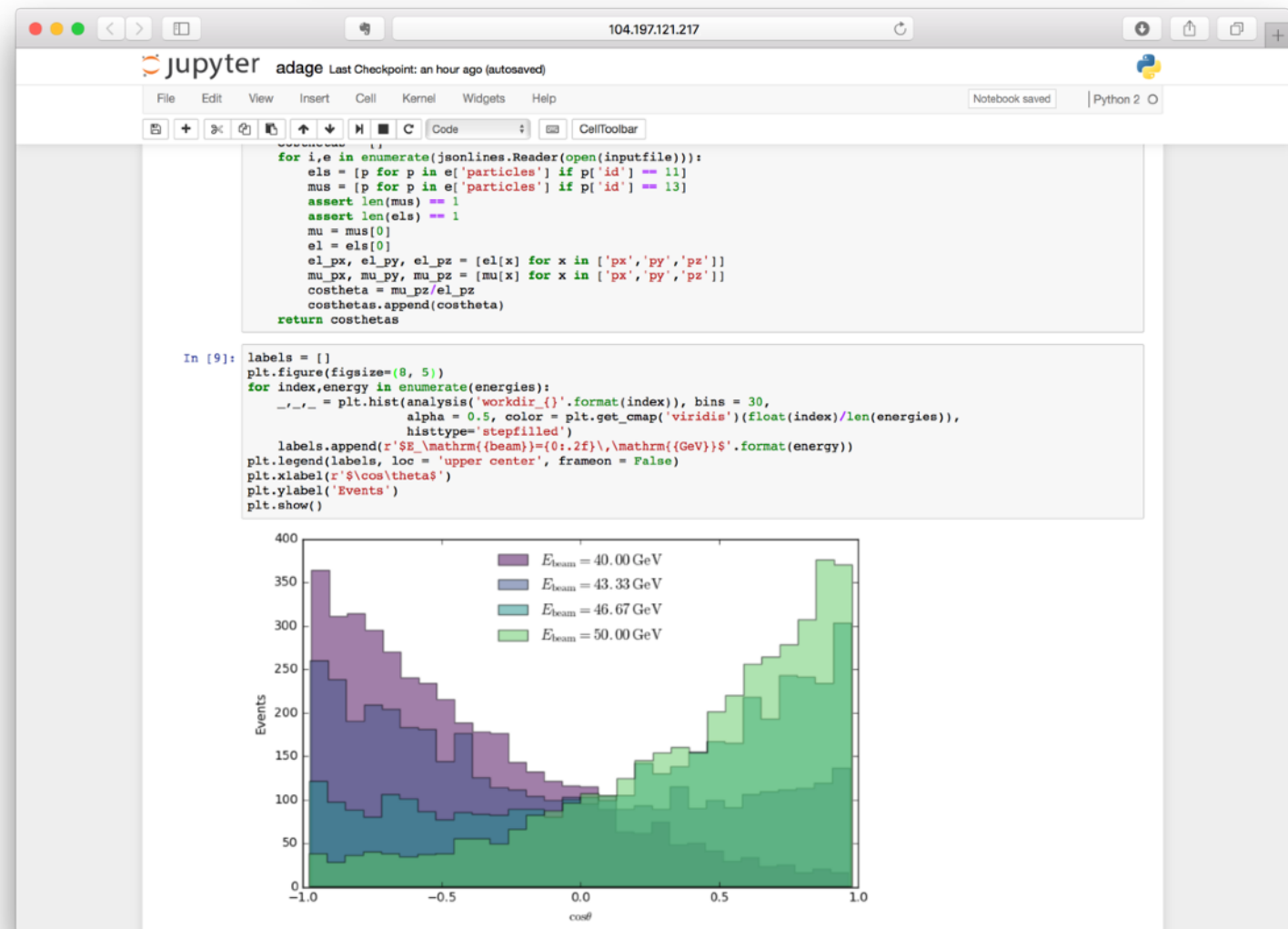
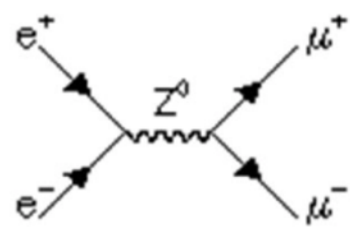
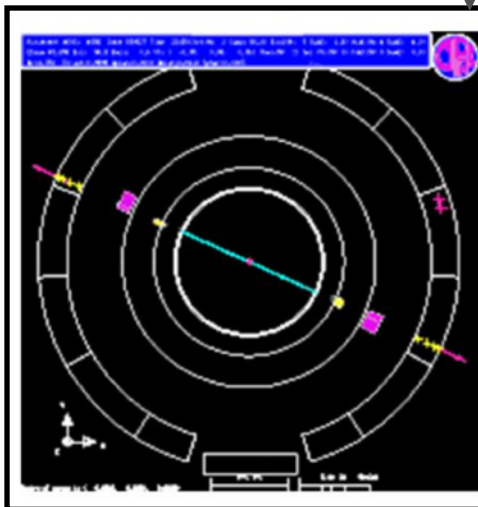
This repository has been equipped with a Dockerfile to encapsulate its software environment. It can be used with the [mybinder](#) service to launch an ephemeral jupyter notebook server to run the notebook.

Click on the below badge and open the notebook `adage.ipynb`.

launch [binder](#)



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





## Analysis as a function mapping data and models to results

$$\text{result} = f_{\text{analysis}}(\text{data} | \text{model})$$

observable distributions,  
confidence intervals  
on model parameters

reconstruction, event  
selection, stat. evaluation

collision data from LHC detector

model hypothesis  
(SM, many SUSY models,  
etc..)



# Analysis Preservation: two-step process

Modern HEP analysis:

- Multiple steps/code-bases, possibly developed by independent teams, with differing software requirements.  
Example: one team developing the event selection, another team developing the statistical analysis

Need to capture:

## 1. Individual processing steps

- code bases
- software environments
- identify binaries, scripts in code base
- templates how to run binaries (semantic description of arguments, naming etc..)
- description of step output, what are the relevant data fragments

## 2. How to connect these steps

- How to wire individual steps together
- What outputs of which steps, are used as inputs for other steps, ...

**Goal:** capture all this with least amount of work for analysis teams, preferably *while analysis is being developed*. *Should not take more than a few days*



# How to preserve $f_{\text{analysis}}(\cdot)$ ?

## 1. Problem: Preserve Individual Processing Steps

(Example: Run Detector Simulation + Reconstruction on MC events)

### Example:

```
process:
  process_type: 'string-interpolated-cmd'
  cmd: 'DelphesHepMC {delphes_card} {outputroot} {inputhepmc}'
publisher:
  publisher_type: 'frompar-pub'
  outputmap:
    rootfile: outputroot
environment:
  environment_type: 'docker-encapsulated'
  image: lukasheinrich/root-delphes
```

### python package: “packtivity”

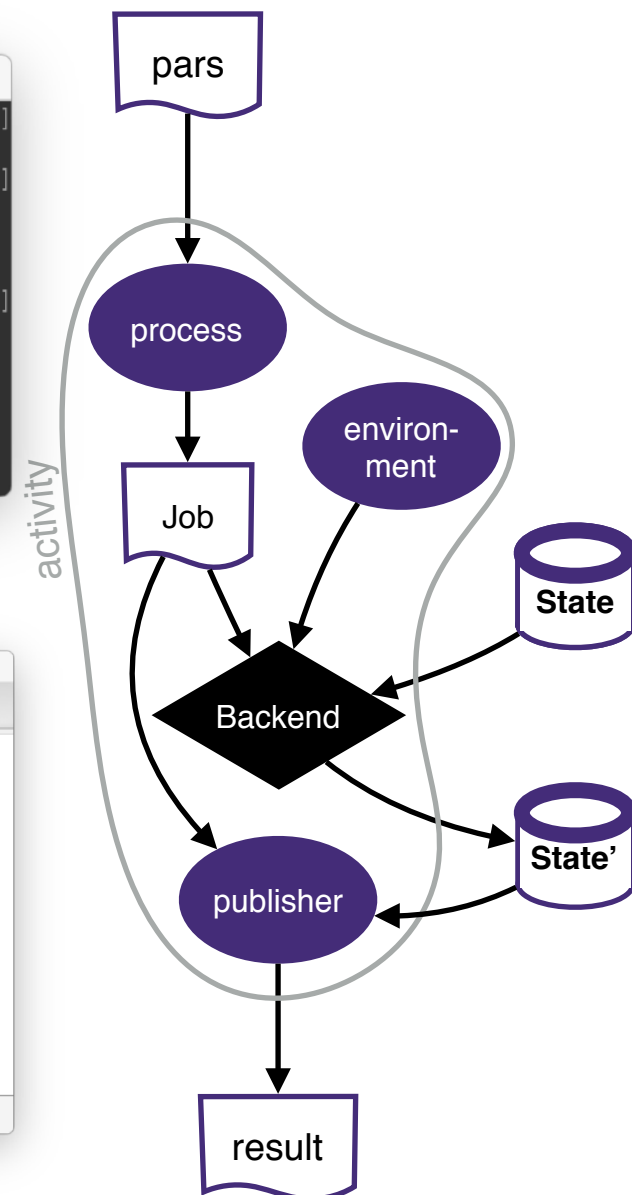
- executes packtivities according to JSON spec for given parameters
- cli tool and python bindings
- multi-host / remote execution ready via e.g. Docker Swarm

### CLI tool

```
172-27-219-223 — -zsh — 64x12
[$> ls
delphes.yml input.hepmc pars.yml
[$> pygmentize -g pars.yml
delphes_card: 'delphes/cards/delphes_card_ATLAS.tcl'
inputhepmc: '{workdir}/input.hepmc'
outputroot: '{workdir}/out.root'
[$> packtivity-run delphes.yml pars.yml
{'rootfile': '/Users/lukas/chep2016/out.root'} (prepublished)
[$> ]
```

### python bindings

```
pack.py — /Users/lukas/chep2016
pack.py
1 import os
2 import capschemas
3 from packtivity import packtivity
4 from packtivity.statecontexts.poxisfs_context import make_new_context
5 packtivity_description = capschemas.load(
6     'delphes.yml', os.curdir,
7     'packtivity/packtivity-schema')
8 pars = {
9     delphes_card: 'delphes/cards/delphes_card_ATLAS.tcl',
10    inputhepmc: '{workdir}/input.hepmc',
11    outputroot: '{workdir}/out.root'
12 }
13 packtivity(packtivity_description, parameters, make_new_context(os.curdir))
```



# How to preserve $f_{\text{analysis}}(\cdot)$ ?

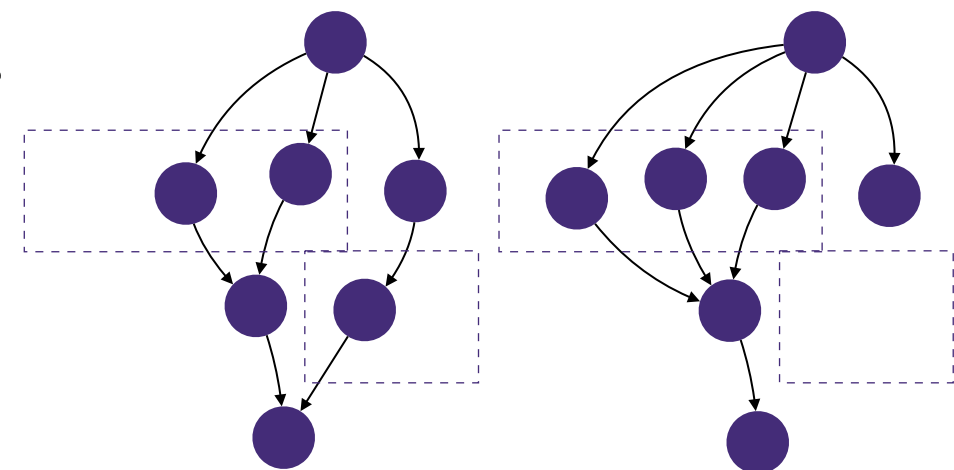
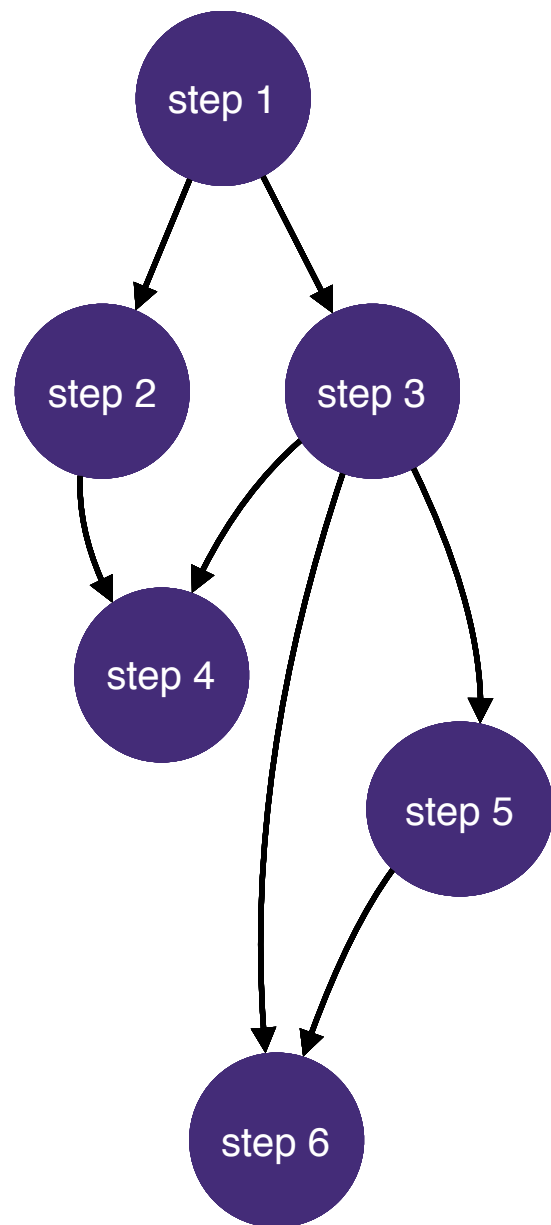
## 2. Problem: Preserve Parametrized Workflow

Natural Data Model: *directed acyclic graphs (DAGs)*

- **nodes**: individual steps
- **edges**: dependency relations

Two place where parametrization enter:

1. individual steps parametrized: covered by “packtivities”  
graph topology may *depend on the parameters* of the analysis and only emerge during run-time
2. Examples:
  - variable number of created files during execution,
  - conditional choices (if/else)/flags do enable/disable steps, e.g. run systematics / not



Par. Set 1

Par. Set 2





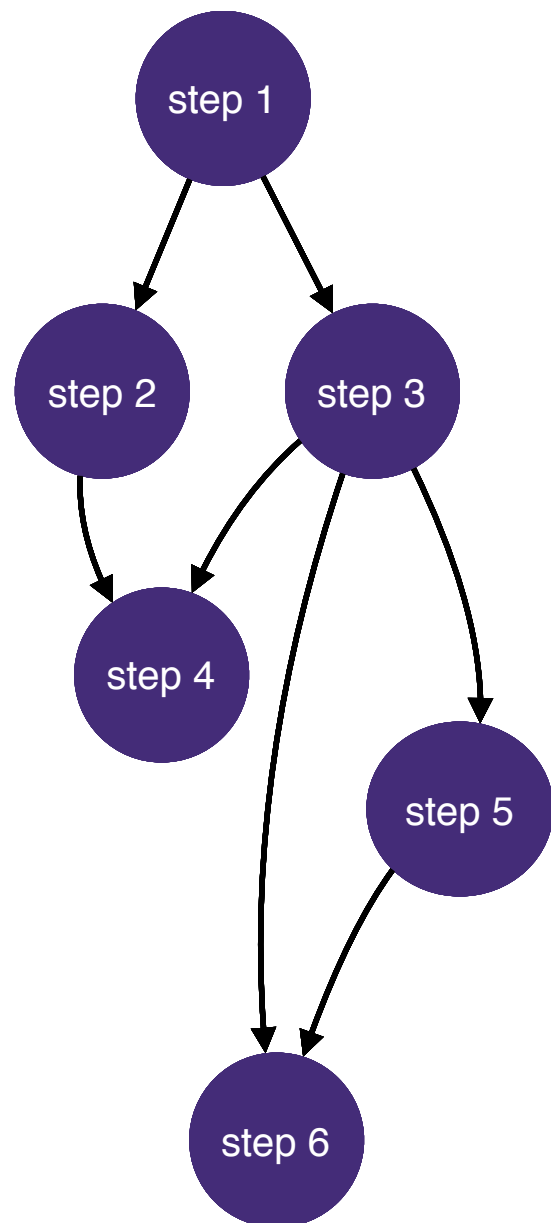
# How to preserve $f_{\text{analysis}}(\cdot)$ ?

## 2. Problem: Preserve Parametrized Workflow

**Therefore:** Sequentially build up graph, as sufficient information becomes available, using a number of stages that add nodes and edges

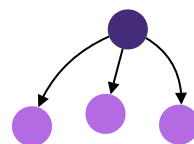
**To capture analysis workflow, capture the stages.**

**Example:  
Parametrized  
Map-Reduce**



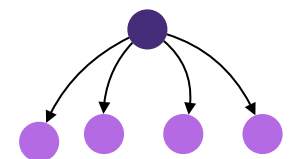
**Stage 1:**

unknown number of files. e.g.  
download & unpack archive with a  
priori unknown # of files



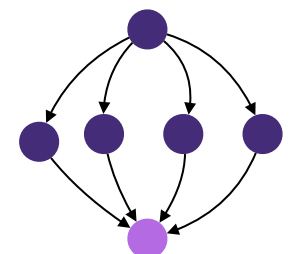
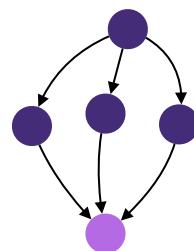
**Stage 2:**

for each file in the archive, add node  
to process it  
(only possible after first node done)



**Stage 3:**

add a node that merges results of  
the map nodes  
node/edge can be added before  
execution of map nodes



Par. Set 1

Par. Set 2

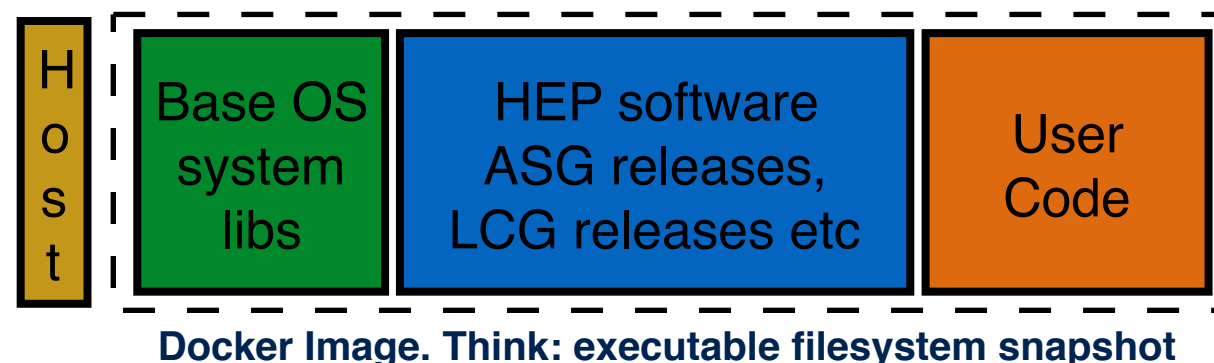


## Technical Solution:

Preserve Software using industry standard Linux Containers (Docker)

- industry backed (Google, Amazon, ...) solution for reproducible software environments. Like a VM, but boots in milliseconds.
- complete freedom for analysis team on software choices. Makes no assumption on how you run your code (“snapshot your work directory”)
- can capture using a script or in an interactive session :

```
lxplus> docker run ... #start snapshot session
container> svn co ...
container> make ...
lxplus> docker commit ... #save snapshot of workdir
```



Many people contributing now. Contributions from CERN, DASPOS, DIANA, GitHub, Moore-Sloan Data Science Environment at NYU, Notre Dame, Nebraska, ...

Using **yadage** and **packtivity** JSON schemas developed by Lukas Heinrich and described in draft DASPOS technical report for packaging realistic LHC analyses

CERN Analysis Portal (CAP) is able to store and serve up analysis workflows stored in this format.

New front-end webpage thanks to Christian Bora (Nebraska, DASPOS) and Eamonn Maguire (CERN)

