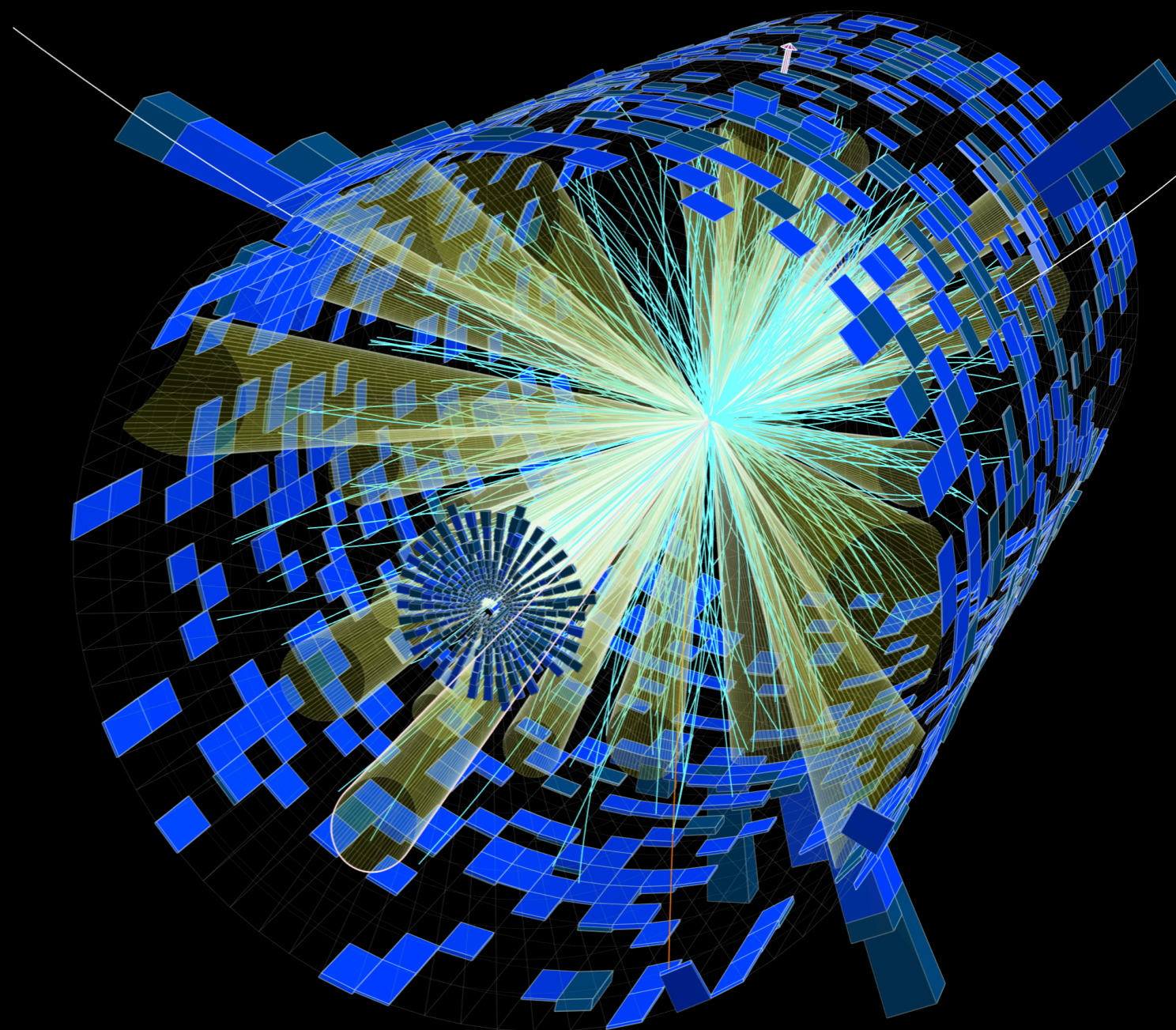




LIKELIHOOD PUBLISHING, RECAST, AND SIMULATION-BASED INFERENCE



@KyleCranmer

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

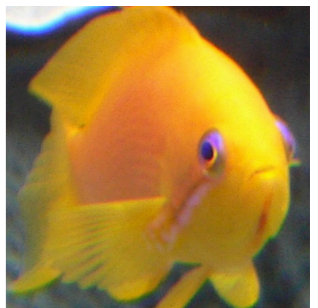
SUPPORT



United States – Israel
Binational Science Foundation



The SCAILFIN Project
scailfin.github.io

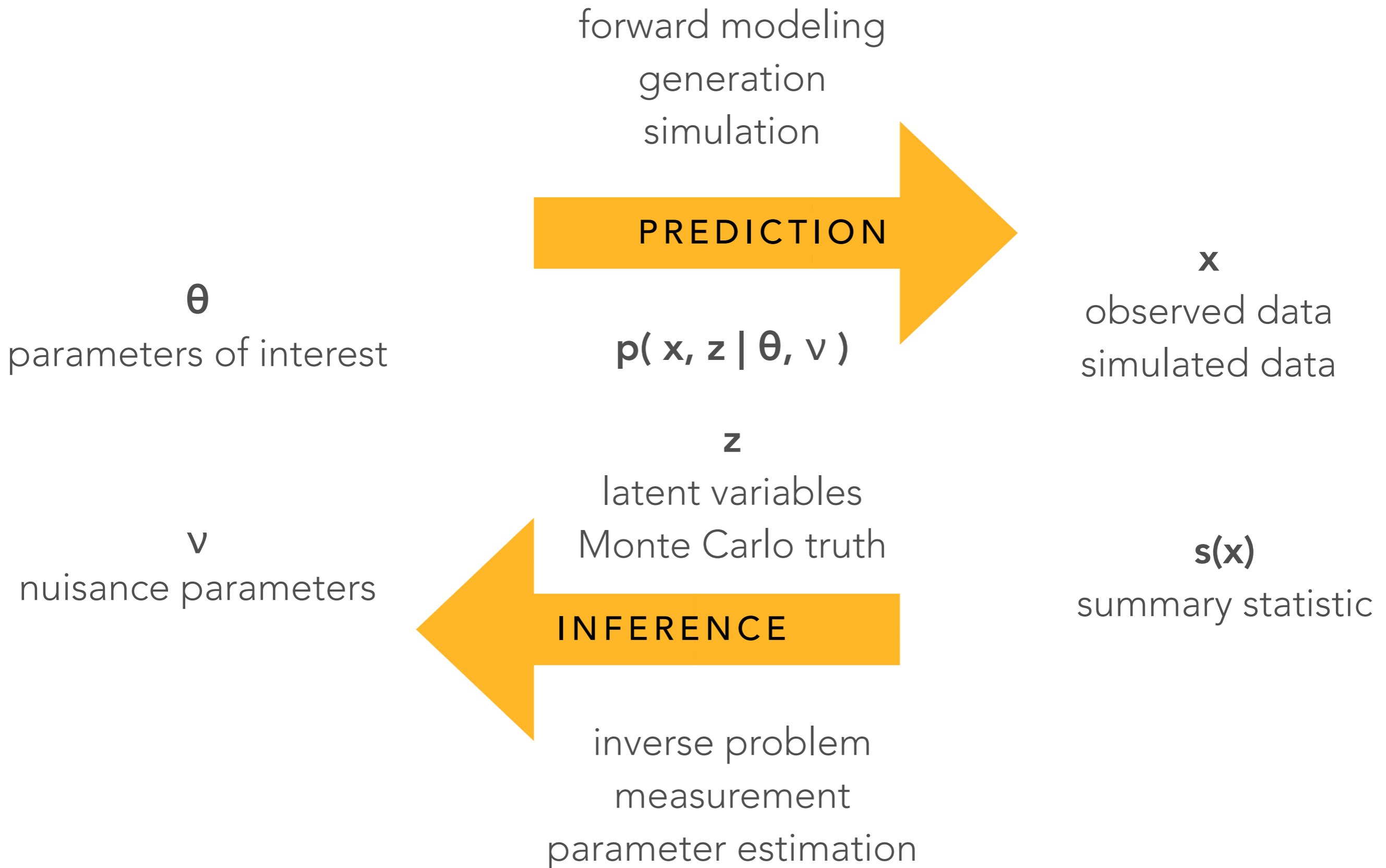


OUTLINE

3 developments addressing fundamental issues in the analysis of particle physics data:

- 1) combining or interpreting results of experiments, where the likelihood of the data given some physically meaningful parameters is the key object.
 - recent progress in the 20-year quest to publish these likelihoods.
- 2) reinterpretation of results that cannot be addressed by the likelihood alone as it requires processing alternative signal hypotheses through the analysis pipeline.
 - recent progress in RECAST, which was proposed 10 years ago.
- 3) a challenge at the heart of analyzing HEP data: our predictions are based on simulations, but the likelihood for the simulator is intractable
 - how machine learning is pushing the frontier of simulation-based inference and how it can greatly enhance the sensitivity of measurements at the LHC (eg. for constraining effective field theories).

STATISTICAL FRAMING



1) Likelihood Publishing

+ Some History of PhyStat
and LHC Statistics

THE FIRST PHYSTAT

It was 20 years ago!

- I was there and just getting started in HEP and statistics
- Thanks Louis!



Louis Lyons of Oxford, co-convenor of the workshop on confidence limits.

<https://cds.cern.ch/record/411537?ln=en>

CERN 2000-005
30 May 2000

SUE 2000 86

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

WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17-18 January 2000

CERN LIBRARIES, GENEVA



P00037096

PROCEEDINGS

Editors: F. James, L. Lyons, Y. Perrin

GENEVA
2000

AN OBSERVATION

The emphasis of the PhyStat series and academic training lectures on statistics has typically been on statistical methods

- hypothesis tests, confidence intervals, Bayes vs. Frequentist etc.

These are important topics, but most are well defined statistical procedures with well defined properties

- The statistical model $p(\text{data} | \text{theory})$ (aka “likelihood function”) is the input to almost all of these statistical procedures.
- This allows us to decouple modeling of the data from debates about statistical procedures
- We should focus less on statistical procedures and more on how we model the data.

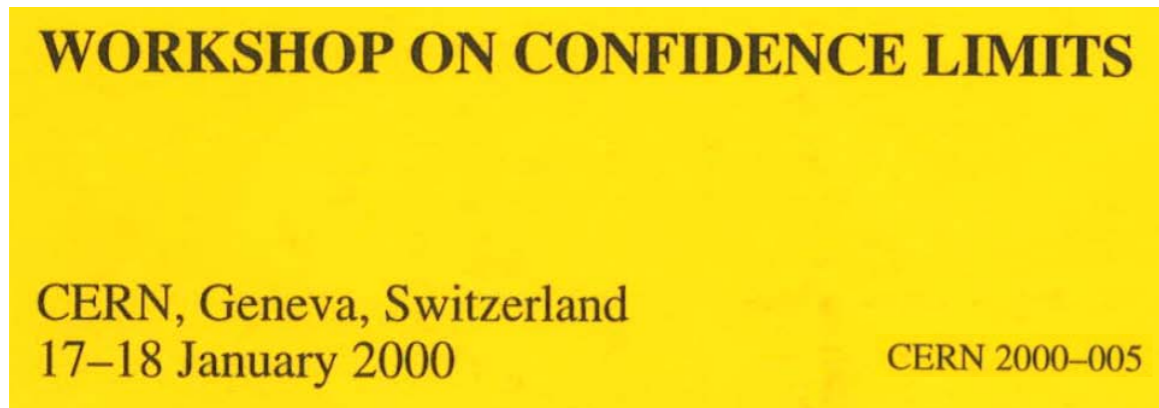
TERMINOLOGY

Given a **probability model** $p(X | \theta)$ and a data x_0

- The **likelihood function** is a function of the parameter θ , and the value is given by $L(\theta) = p(X = x_0 | \theta)$
- But $L(\theta)$ doesn't describe the distribution in X
- Technically the likelihood function doesn't have enough information to generate synthetic data (toy Monte Carlo), which is needed for most frequentist statistical procedures

Colloquially, the term likelihood function is used in HEP often when we mean the full probability model $p(X | \theta)$

Origins I: The First “Statistics in HEP” conference



Massimo Corradi

Does everybody agree on this statement, to publish likelihoods?

Louis Lyons

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

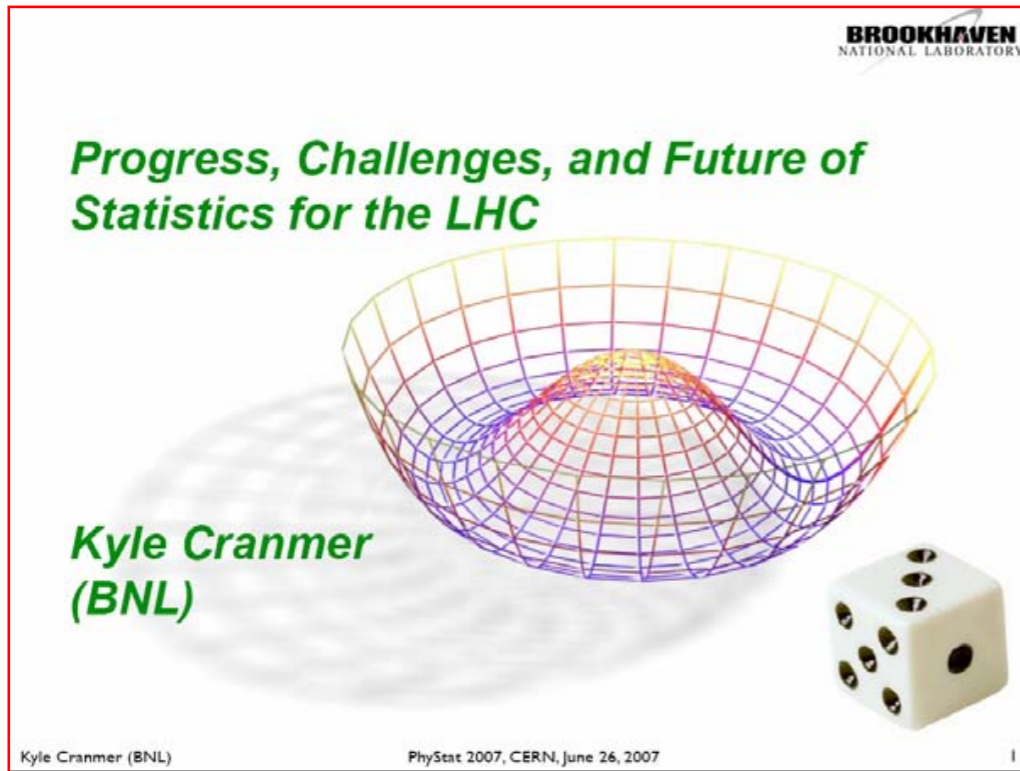
...[Fred James wants to be able to calculate coverage, Don Groom wants to be able to calculate goodness of fit]...

Cousins

I thought the point of unanimity was that publishing the likelihood function was a *necessary* condition, not a sufficient condition.

But a practical problem remained: How to communicate multi-D likelihood?

From Bob Cousins Slides: <http://indico.cern.ch/conferenceDisplay.py?confId=100458>



ROOT Statistical Software



Lorenzo Moneta (CERN, PH-SFT)
 on behalf of the ROOT Math Work Package
 (R. Brun, A. Kreshuk, E. Offermann + many others contributors)

Statistics software for the LHC

Wouter Verkerke



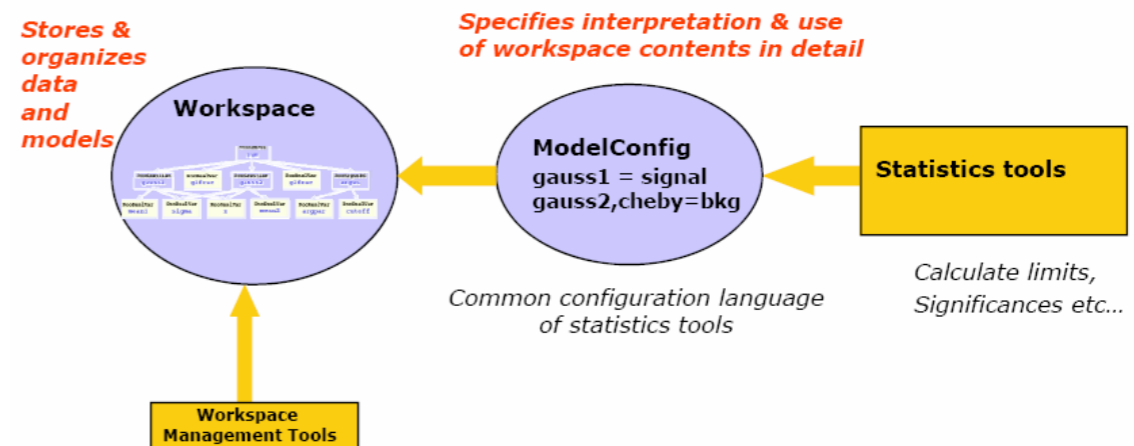
The Workspace as publication

- Now have functional **RooWorkspace** class that can contain
 - Probability density functions and its components
 - (Multiple) Datasets
 - Supporting interpretation information (**RooModelConfig**)
 - **Can be stored in file with regular ROOT persistence**
- **Ultimate publication of analysis...**
 - Full likelihood available for Bayesian analysis
 - Probability density function available for Frequentist analysis
 - Information can be easily extracted, combined etc...
 - Common format for sharing, combining of various physics results



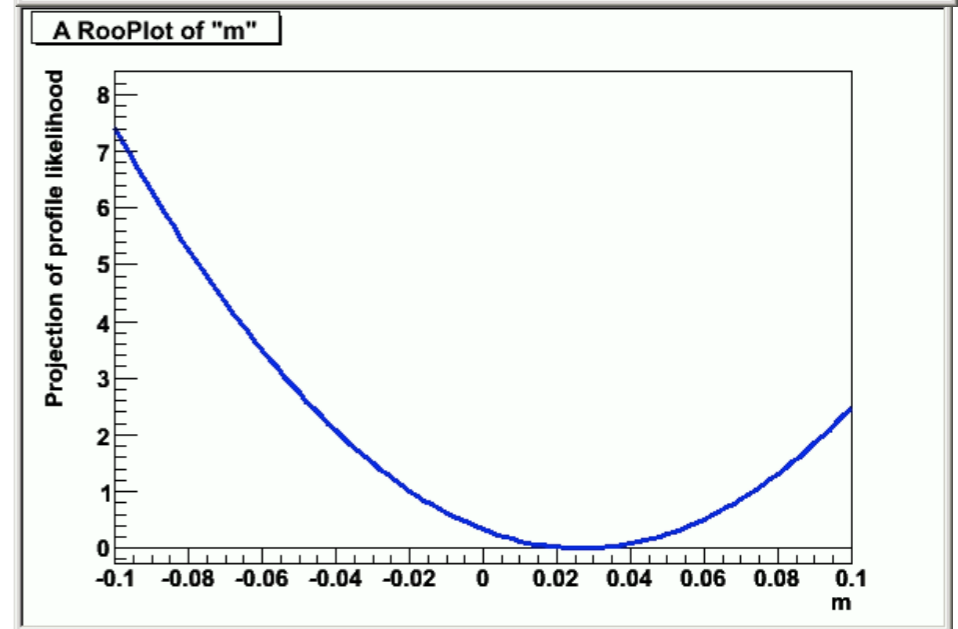
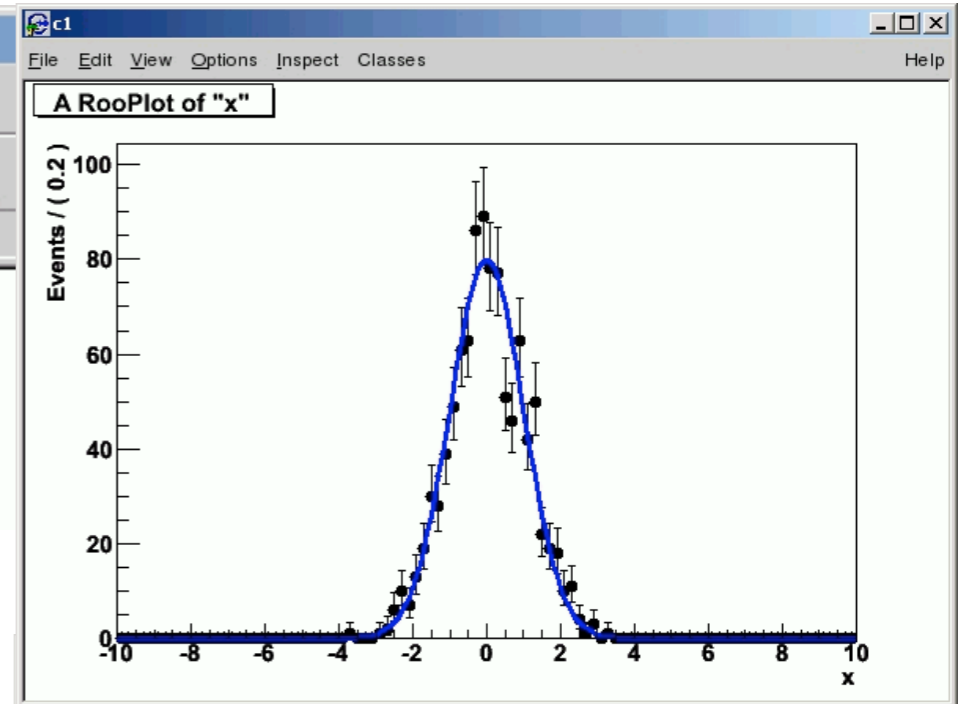
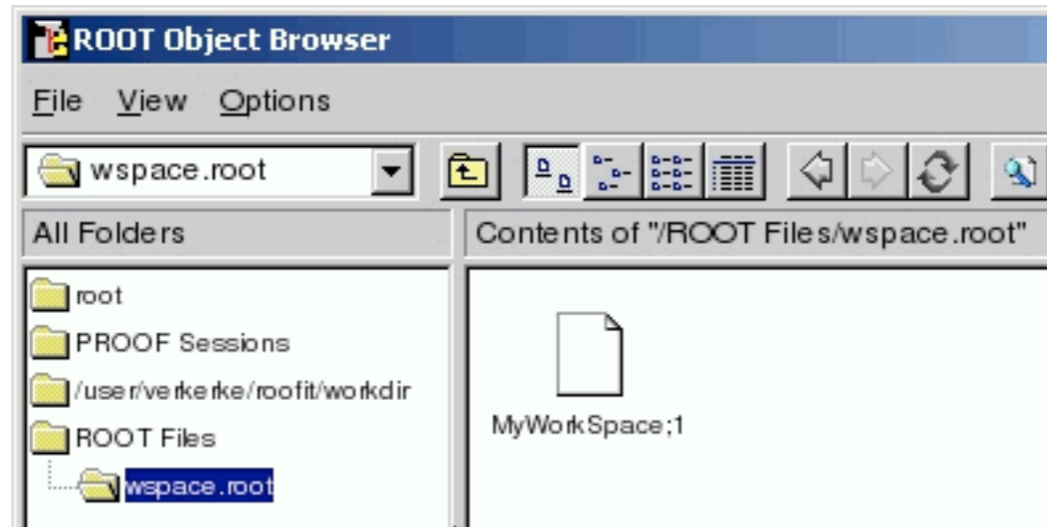
Framework design & RooFit adaptations

- Have had more meetings last 3 months to review RooFit lessons from BaBar
 - Kyle, Amir Farbin (ex-Babar), Frank Wrinklmeyer (ex-Babar), WV
 - Design for **WorkSpace** and **ModelConfig** concept in RooFit to interface with statistics tools



NOT JUST THE LIKELIHOOD, THE FULL MODEL

Example of Digital Publishing



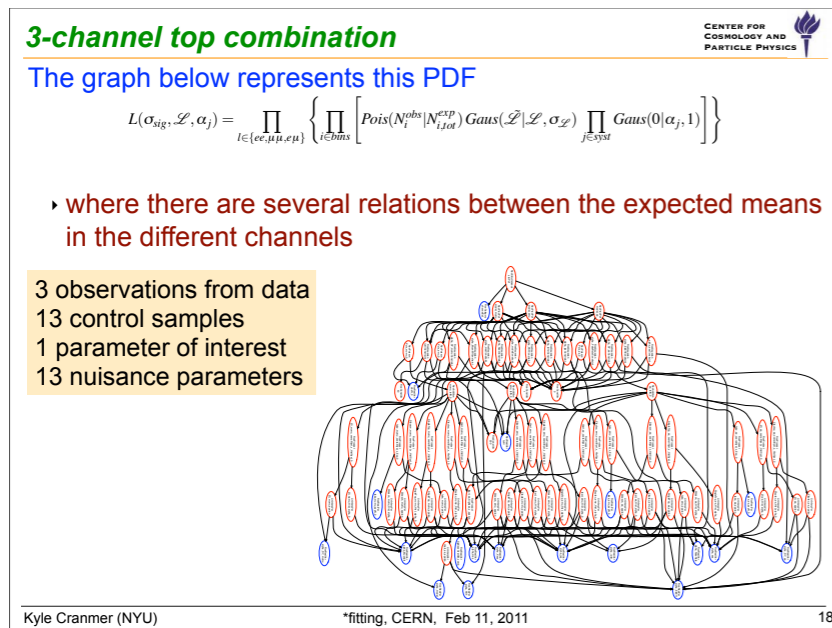
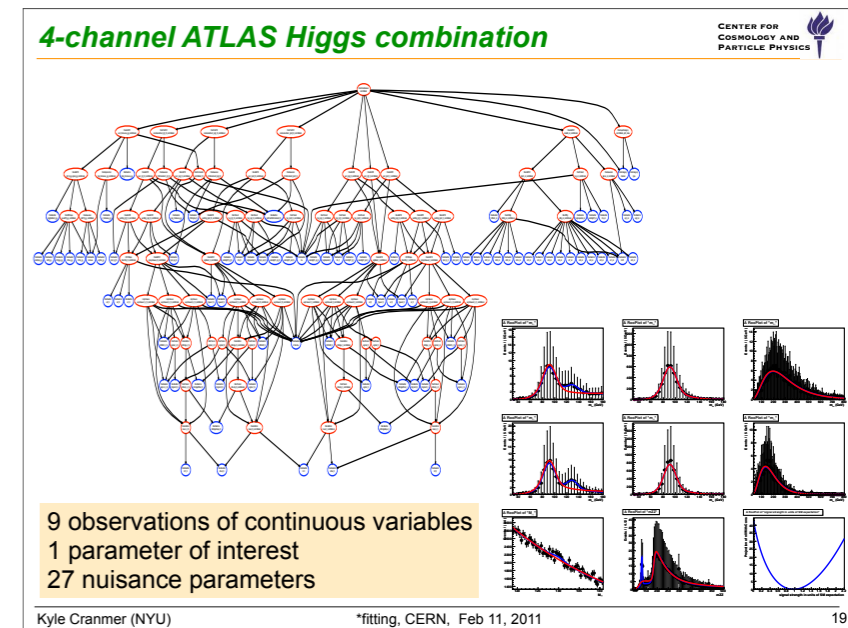
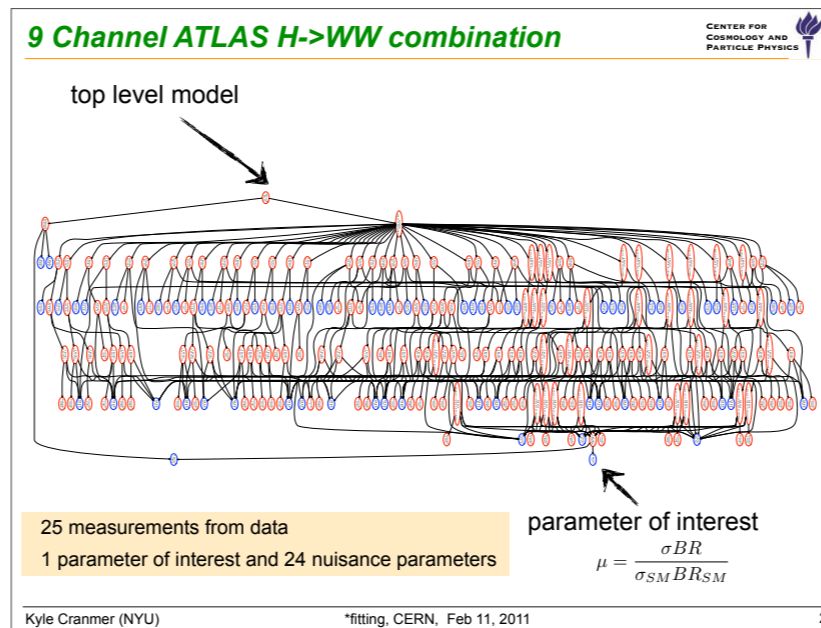
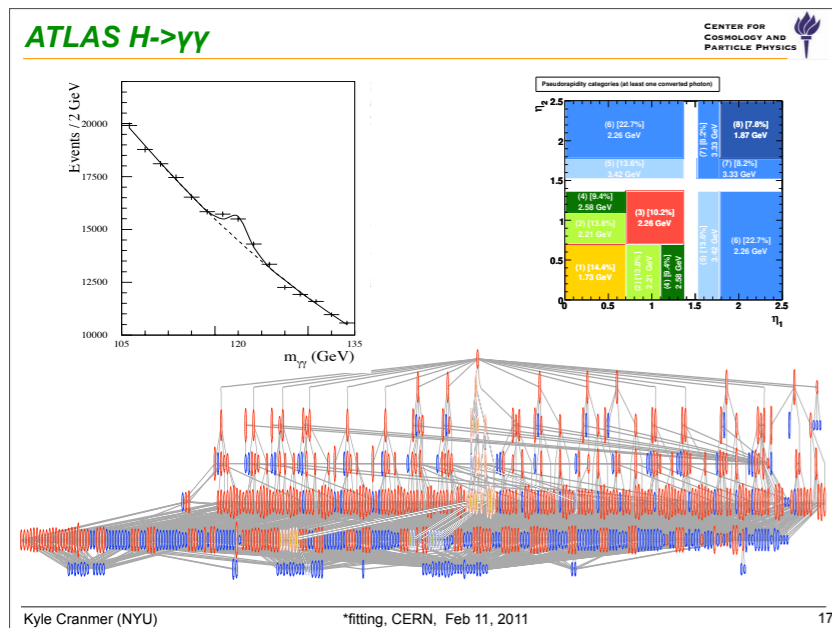
Wouter recently demonstrated the ability to save the function $L(x|\theta_r, \theta_s)$ in a Root file with minimal data necessary to reproduce likelihood function.

Can also evaluate integrals over x necessary for Neyman construction!

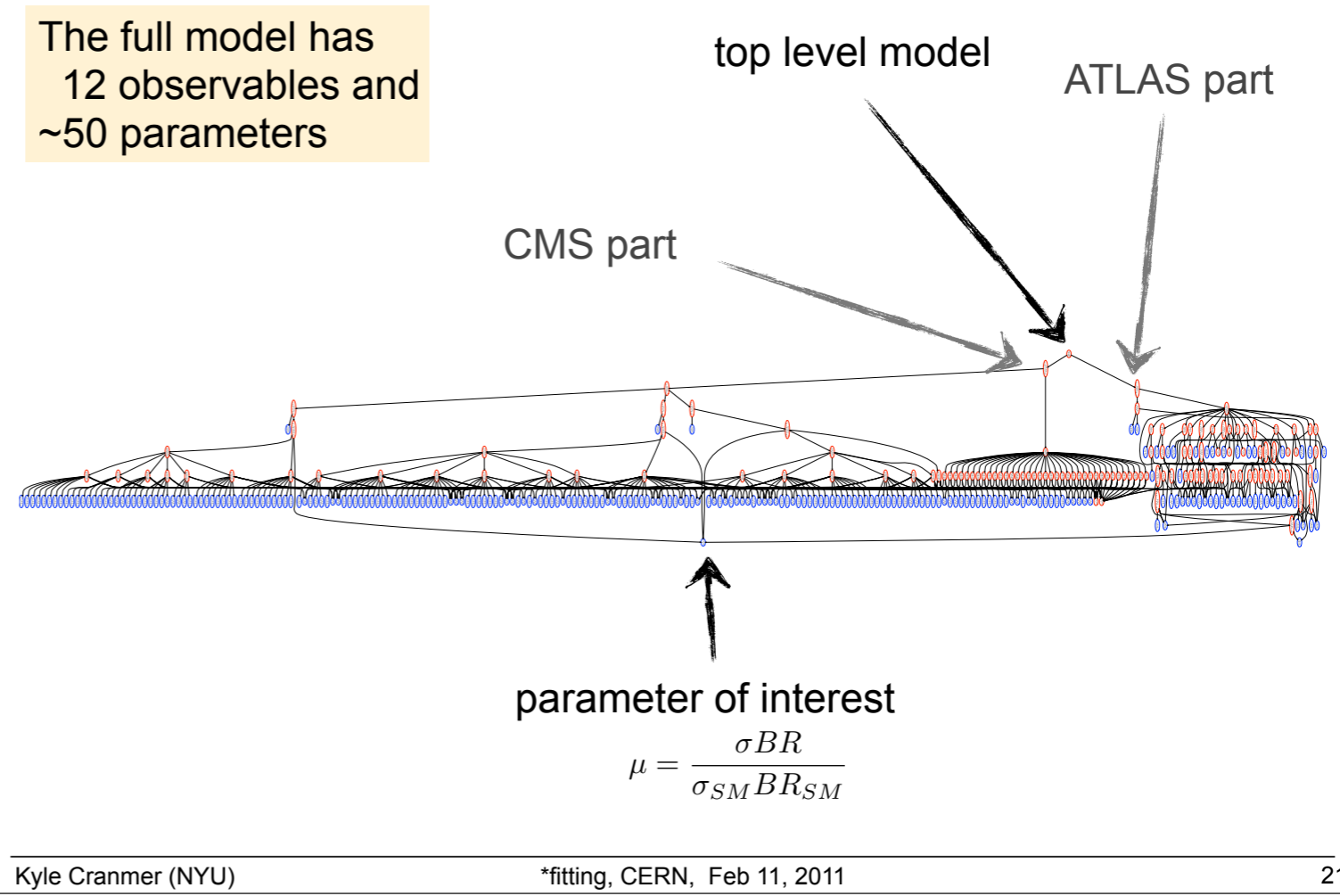
Need this for combinations, we should publish them to some repository!

Full statistical model $p(X|\theta)$ → can generate toys

EARLY LHC EXAMPLES (2011)



Visualization of the ATLAS+CMS Workspace



Global BSM fits and LHC data

10-11 February 2011
CERN
Europe/Zurich timezone

Search...

Overview

Timetable

Registration

List of registrants

The aims of this workshop include:

- to review the progress of the tools for global fits of BSM models
- to propose benchmarks for the parameterization of specific classes of models, in order to facilitate and standardize the representation of the results of the experimental searches at the LHC, and their use in the fitting codes
- to liaise with the "simplified models" approaches, as discussed e.g. in the "Characterization of new physics at the LHC" meetings
- to provide an update of the work carried out within the DESY SUSY/BSM Fit Working Group

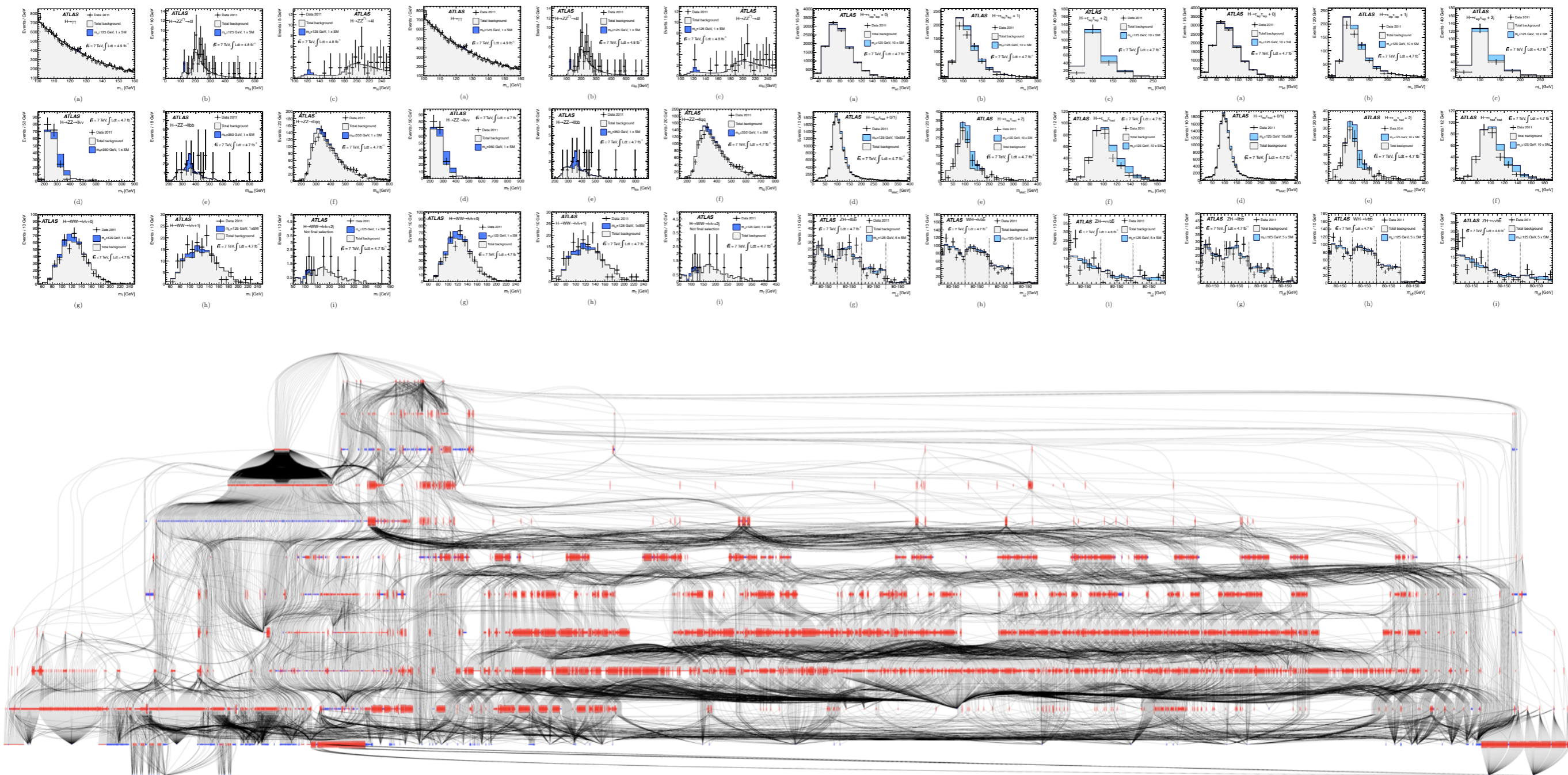
Information on accommodation, access to CERN and laptop registration is available from <http://lpc.web.cern.ch/LPCC/index.php?page=visit>

Starts 10 Feb 2011, 08:00
Ends 11 Feb 2011, 18:00
Europe/Zurich

CERN
TH Theory Conference Room

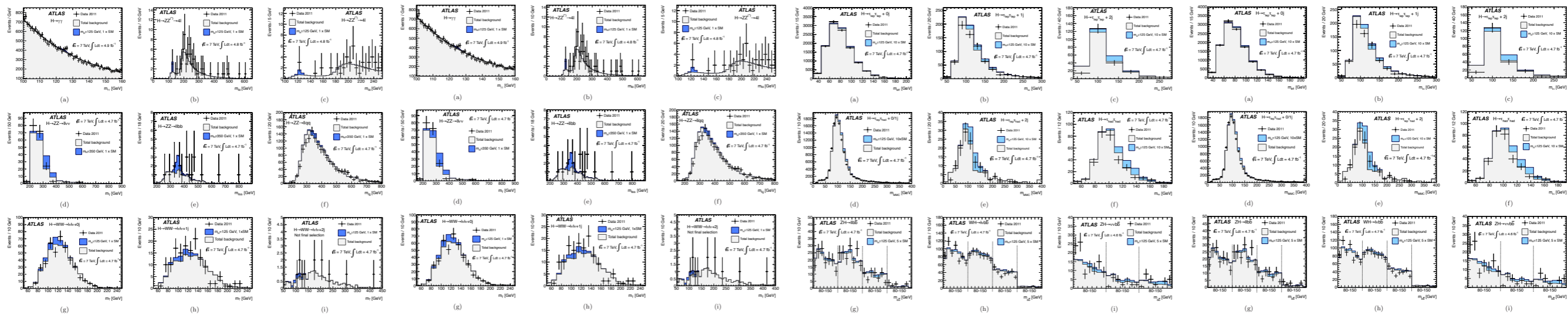
Michelangelo Mangano

Collaborative Statistical Modeling

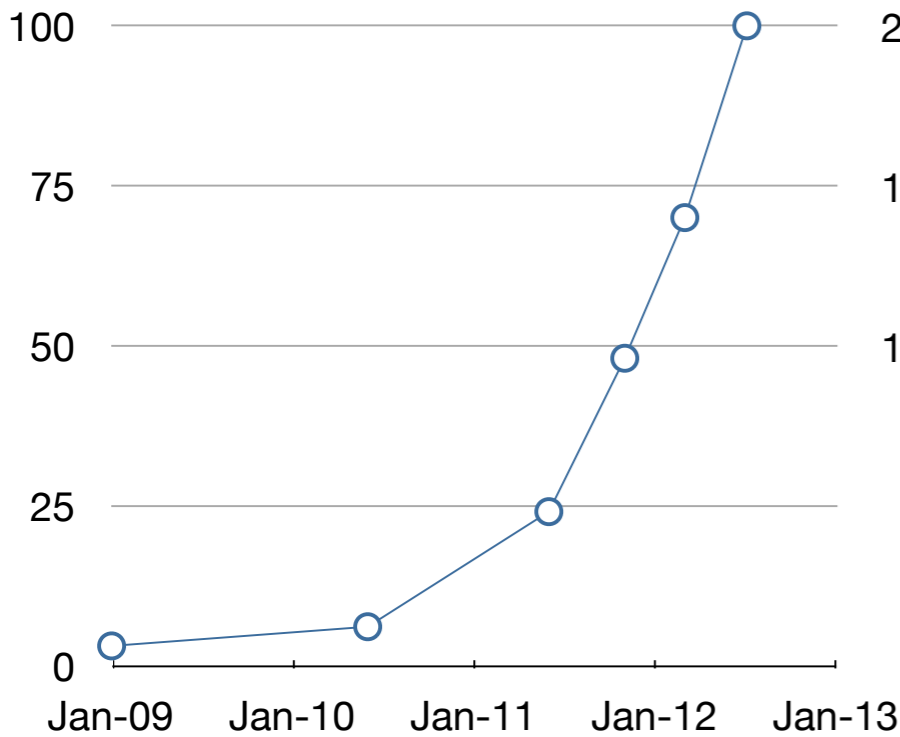


$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G} | \alpha) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c | \nu_c(\alpha)) \prod_{e=1}^{n_c} f_c(x_{ce} | \alpha) \right] \cdot \prod_{p \in \mathcal{S}} f_p(a_p | \alpha_p)$$

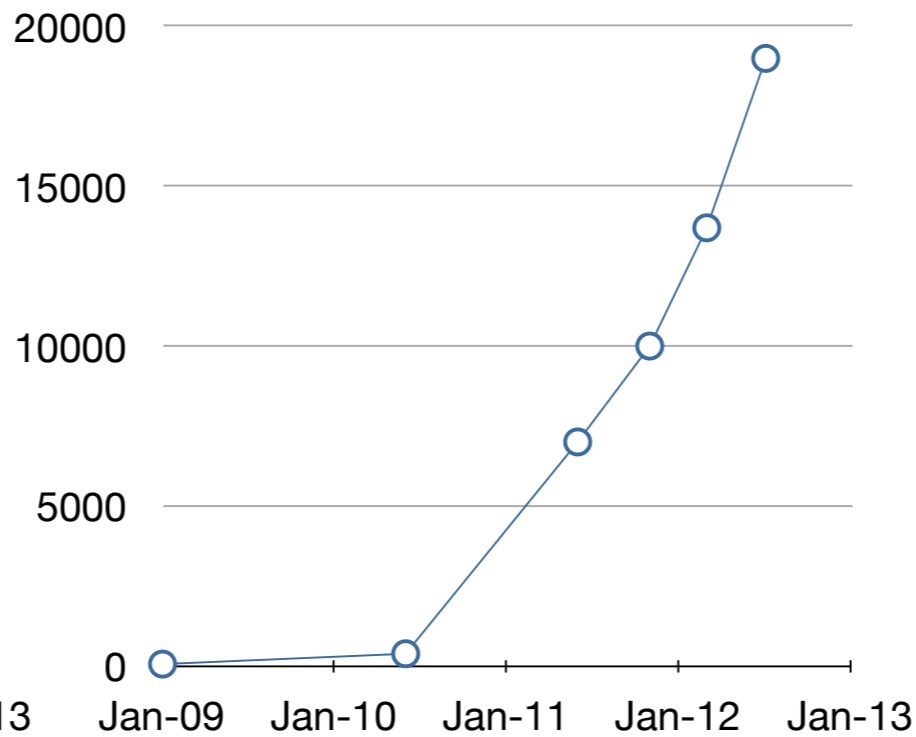
Collaborative Statistical Modeling



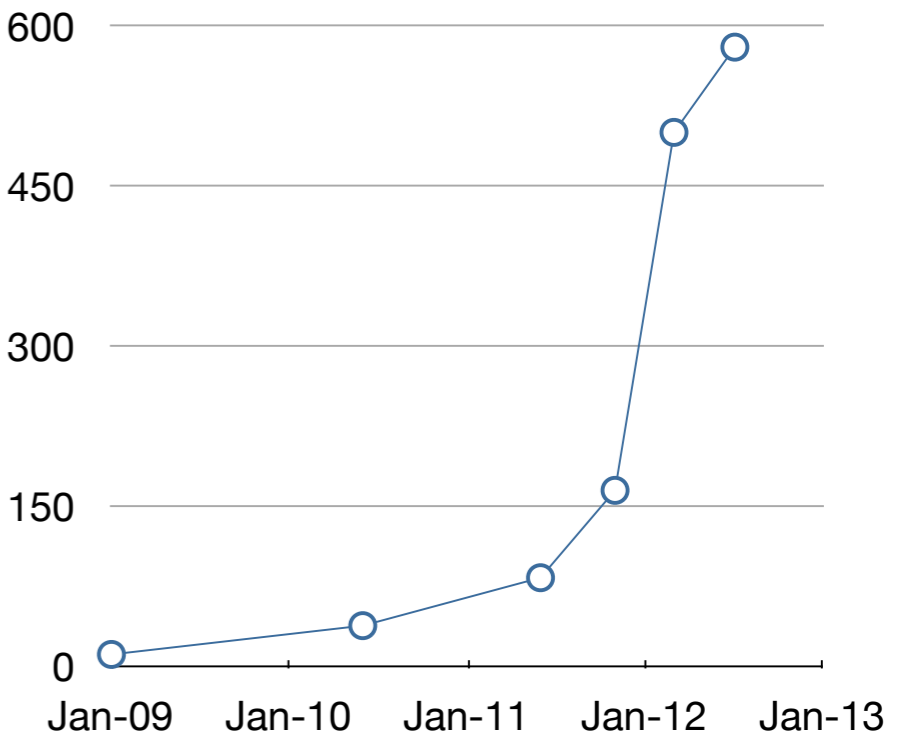
Number of Datasets Combined



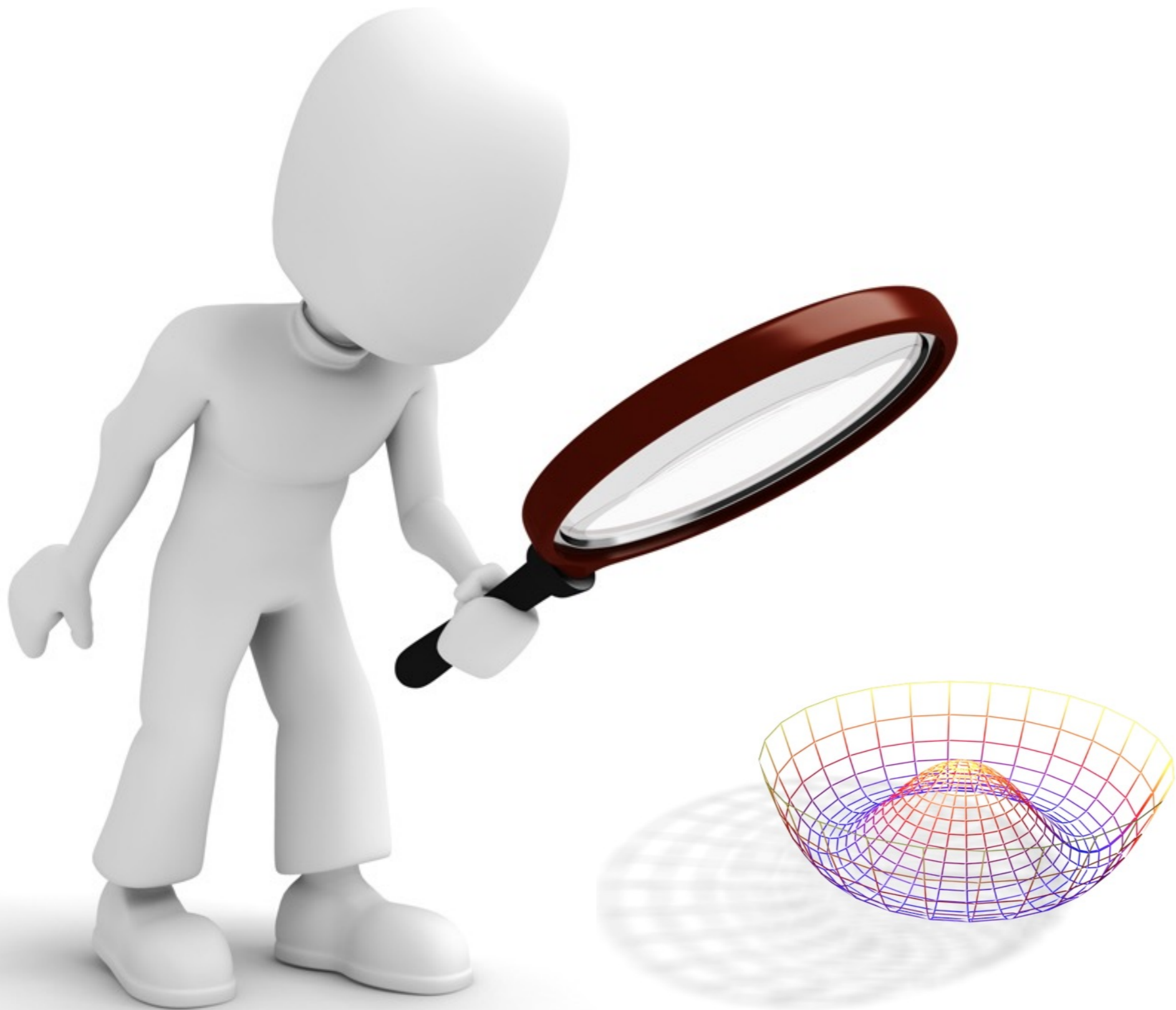
Number of Model Components



Number of Parameters in Likelihood

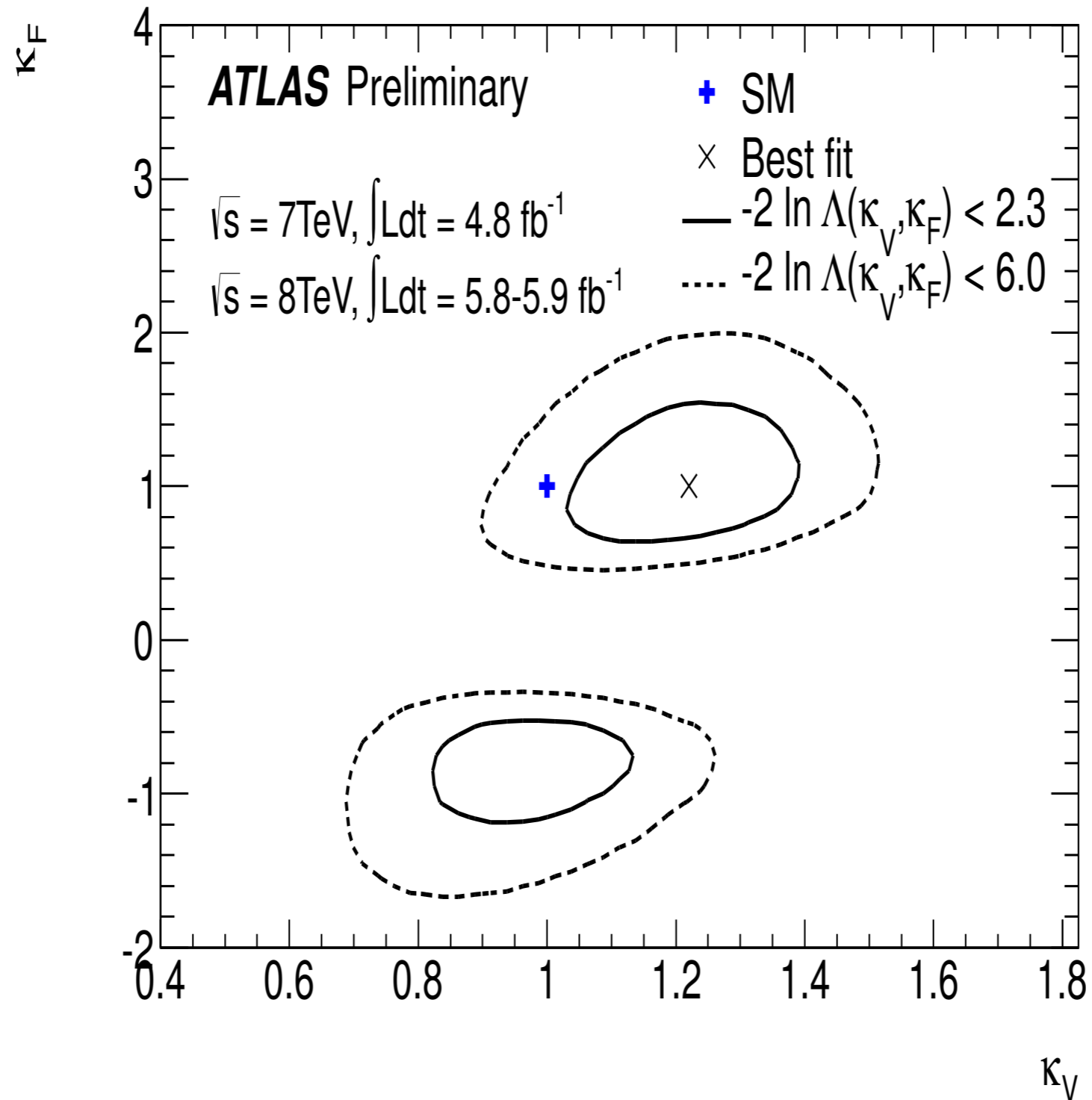






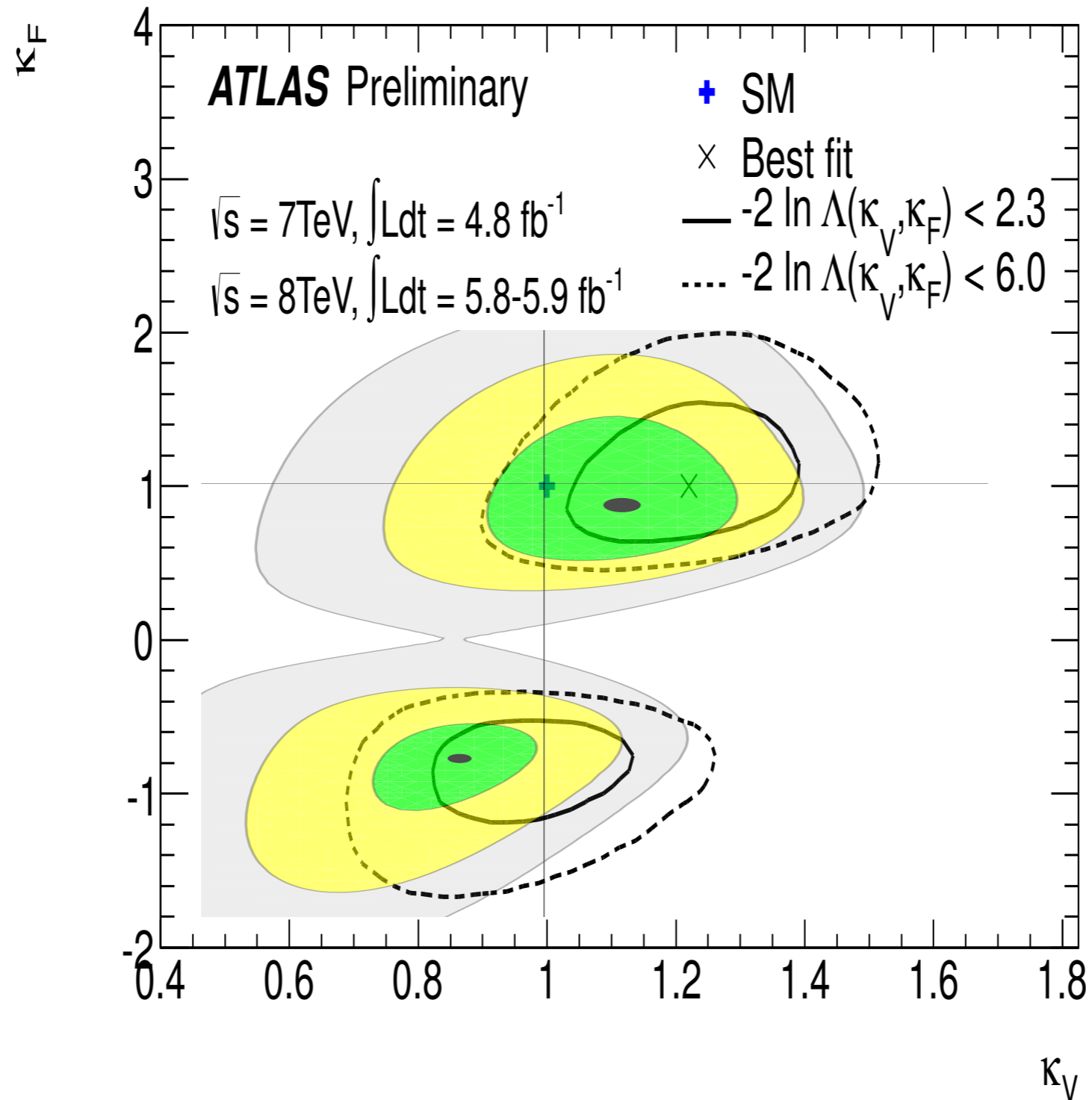
REPRODUCIBILITY PROBLEM

Not possible for others to reproduce results from paper.



REPRODUCIBILITY PROBLEM

Not possible for others to reproduce results from paper.



S. Kraml¹, B.C. Allanach², M. Mangano³, H.B. Prosper⁴, S. Sekmen^{3,4} (editors), C. Balazs⁵, A. Barr⁶, P. Bechtle⁷, G. Belanger⁸, A. Belyaev^{9,10}, K. Benslama¹¹, M. Campanelli¹², K. Cranmer¹³, A. De Roeck³, M.J. Dolan¹⁴, T. Eifert¹⁵, J.R. Ellis^{16,3}, M. Felcini¹⁷, B. Fuks¹⁸, D. Guadagnoli^{8,19}, J.F. Gunion²⁰, S. Heinemeyer¹⁷, J. Hewett¹⁵, A. Ismail¹⁵, M. Kadastik²¹, M. Krämer²², J. Lykken²³, F. Mahmoudi^{3,24}, S.P. Martin^{25,26,27}, T. Rizzo¹⁵, T. Robens²⁸, M. Tytgat²⁹, A. Weiler³⁰

Why public likelihoods

- The statistical model of an experimental analysis provides the complete mathematical description of that analysis

$p(o|\alpha)$ relating the observed quantities o to the parameters α

- Given the likelihood, all the standard statistical approaches are available for extracting information from it
- Essential information for any detailed interpretation of experimental results
 - = determining the compatibility of the observations with theoretical predictions

Les Houches Recommendations (2012)

3b: When feasible, **provide a mathematical description of the final likelihood** function in which experimental data and parameters are clearly distinguished, either in the publication or the auxiliary information. Limits of validity should always be clearly specified.

3c: Additionally **provide a digitized implementation of the likelihood** that is consistent with the mathematical description.

[arXiv:1203.2489](https://arxiv.org/abs/1203.2489)



Searches for new physics: recommendations for the presentation of LHC results

13 February 2012

CERN

Europe/Zurich timezone



Overview

Timetable

Registration

Participant List

During the Les Houches 2011 workshop, discussions started to define a set of recommendations for the presentation of LHC results on searches for new physics, aimed at providing a more efficient flow of scientific information between the experimental collaborations and the rest of the high energy physics community, and facilitating the interpretation of the results in a wide class of models. This discussion evolved into a draft document, available for download from this page. The goal of this meeting is to review this draft and present it to the experiments for discussion and eventual endorsement.

EVO connection will be available: the link will appear on the agenda page 30' before the start of the meeting. **Please register even if you participate only by EVO**



Starts 13 Feb 2012, 09:00

Ends 13 Feb 2012, 19:00

Europe/Zurich



[Michelangelo Mangano](#)

[Sabine Kraml](#)



CERN

TH Theory Conference Room



Paper

[Final paper, arXiv:1203.2489](#)

[LH proceedings contribution](#)

Likelihoods for the LHC Searches

21-23 January 2013

CERN

Europe/Zurich timezone



Overview

Timetable

Registration

Participant List

The primary goal of this 3-day workshop is to educate the LHC community about the scientific utility of likelihoods. We shall do so by describing and discussing several real-world examples of the use of likelihoods, including a one-day in-depth examination of likelihoods in the Higgs boson studies by ATLAS and CMS.

The workshop will start with two pedagogical lectures that introduce likelihood concepts and terminology. These lectures are followed, in the afternoon of Day 1, by a moderated discussion that focuses on the concepts and issues raised in the lectures. Day 1 ends with several presentations that illustrate the use of likelihoods in Higgs and Beyond the Standard Model (BSM) research. The goal here is to get feedback from researchers who have used Higgs and BSM results in their work.

Given the importance of the work on the Higgs boson, we shall devote the second day of the workshop to the thorough deconstruction of likelihood usage in the Higgs boson work by ATLAS and CMS. The goal is to shed a bright light on the many details, and assumptions, that underlie the likelihoods used in the recently published results.

The final day of the workshop covers the use of likelihoods in BSM work and ends with an examination and discussion of the concrete steps needed to make the publication of likelihoods by the LHC community systematic and routine.



Starts 21 Jan 2013, 09:00

Ends 23 Jan 2013, 18:00

Europe/Zurich



CERN

4/3-006 - TH Conference Room

[Go to map](#)



[Harrison Prosper](#)

[Kyle Stuart Cranmer](#)

[Sezen Sekmen](#)

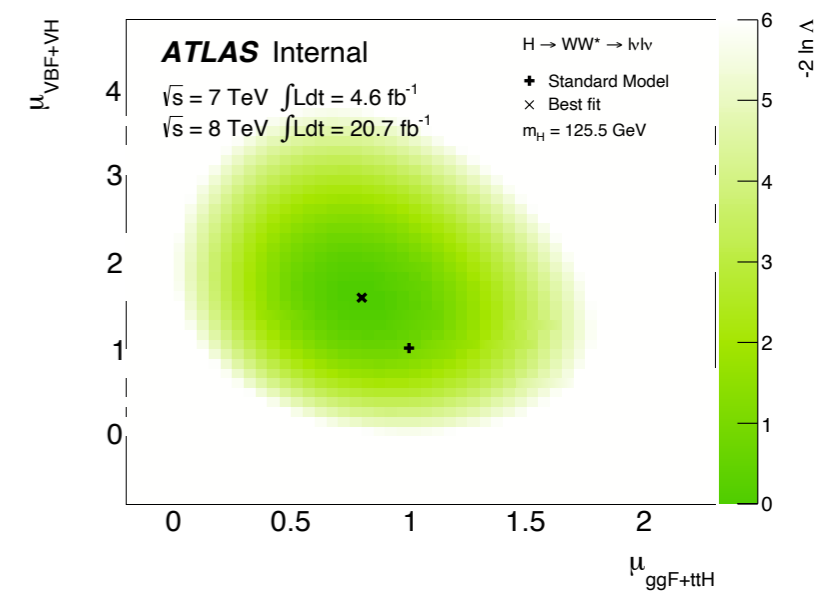
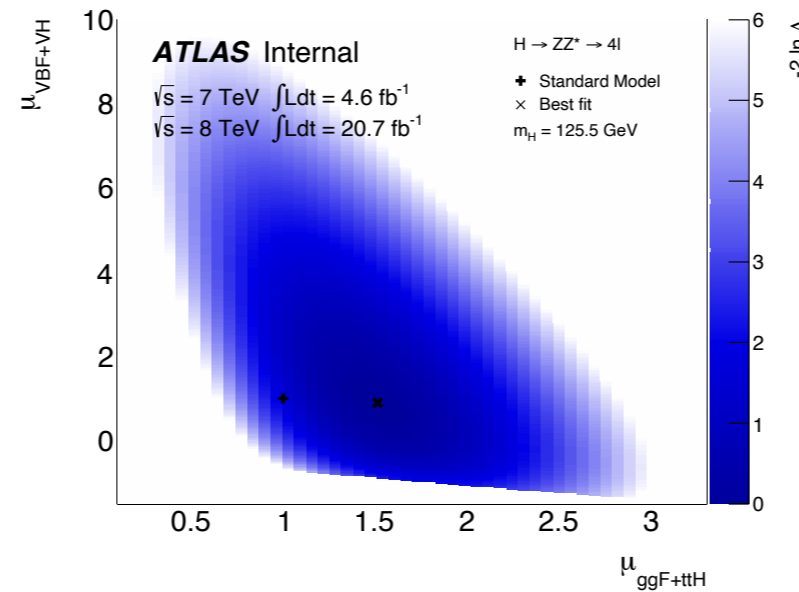
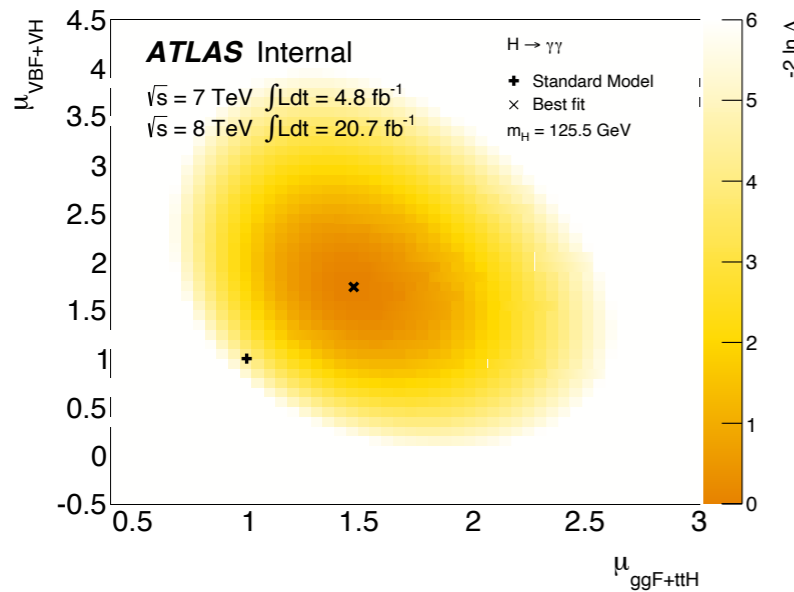
LIKELIHOOD SCANS

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

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

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

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



Information References (121) Citations (128) Files Plots **HepData**

Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

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

Jul 4, 2013 - 32 pages

Phys.Lett. B726 (2013) 88-119



Blogged by 3
Tweeted by 6

[Click for more details](#)

Information Citations (7) Files

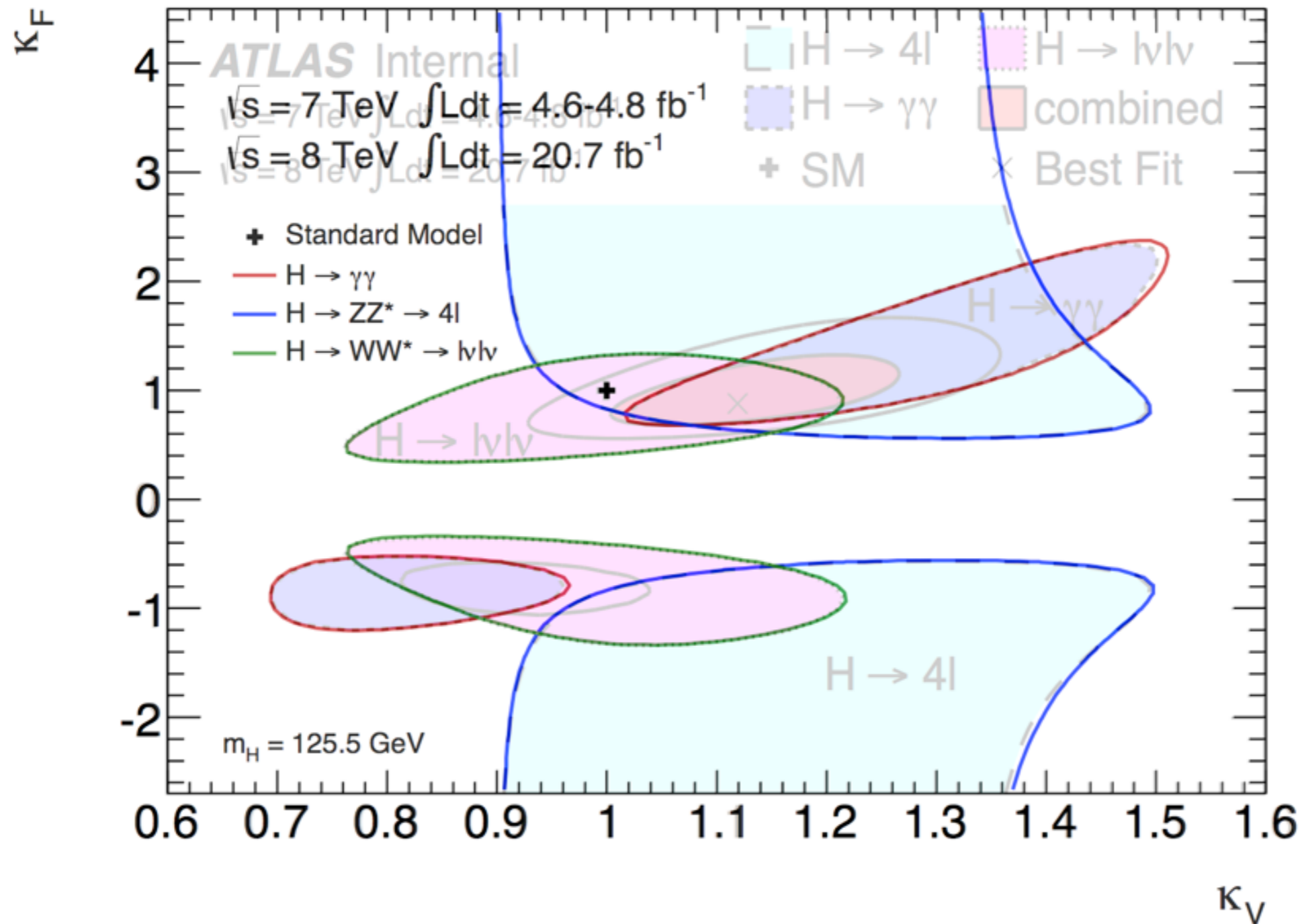
Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

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

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

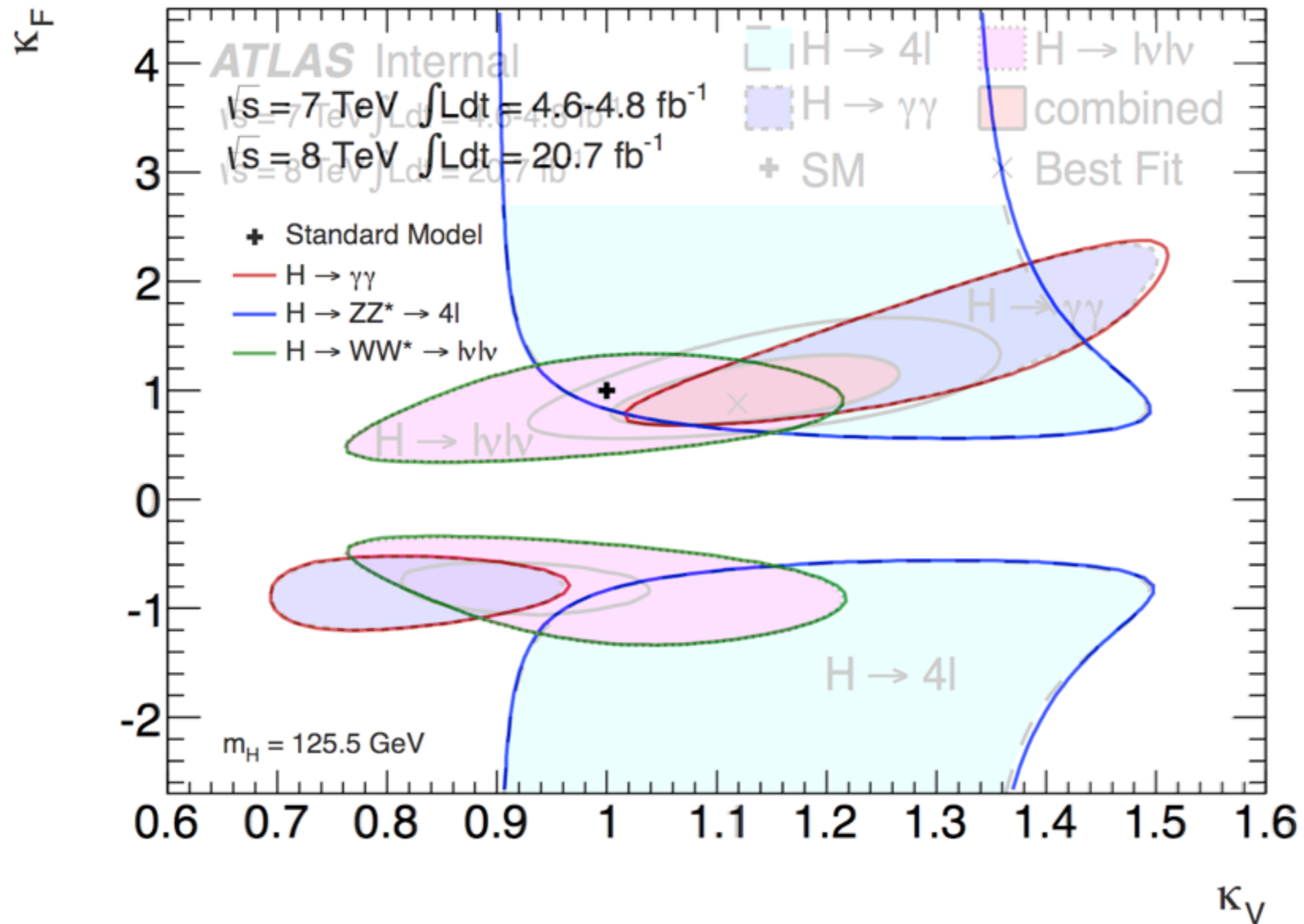
LIKELIHOODS SCANS

Reproducing derived results from original paper!



LIKELIHOODS SCANS

Reproducing derived results from original paper!



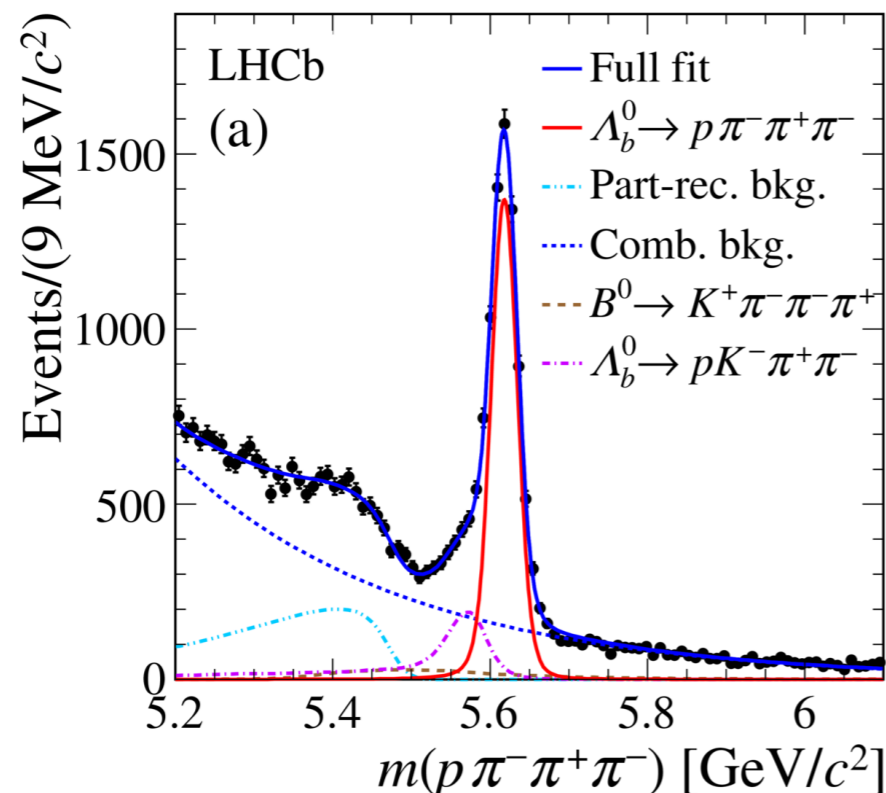
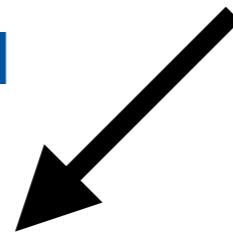
But still simplified likelihood scans, not the full statistical model

OPEN WORLD

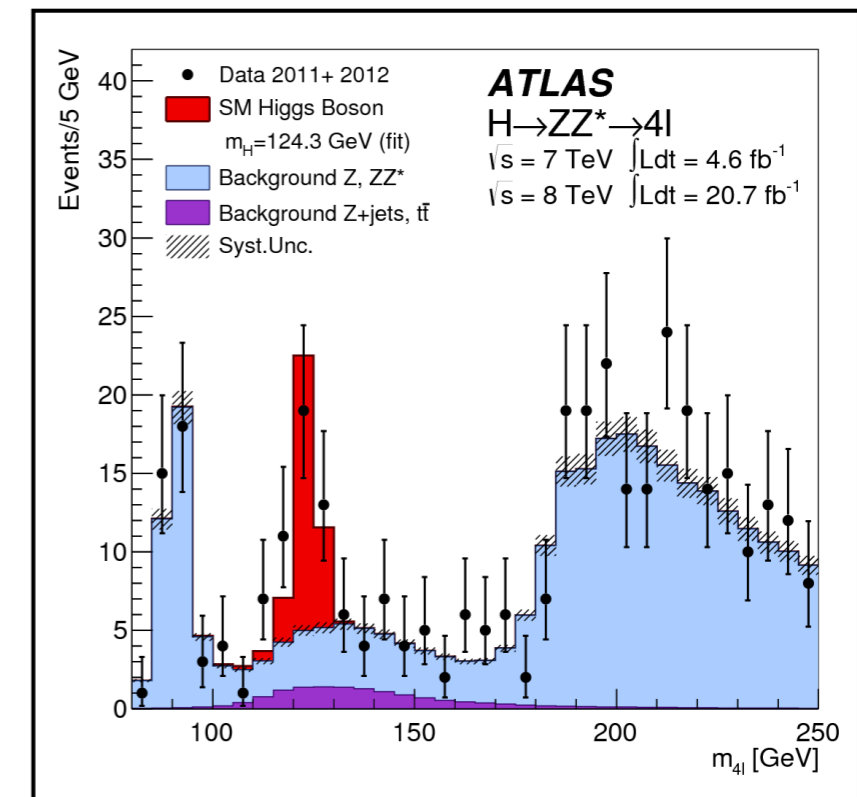
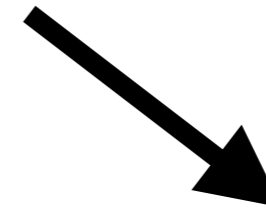
The RooWorkspace was designed to be able to store any type of statistical model → source of many complications

broadly we have two classes of analyses: **binned** and **unbinned**

unbinned



binned



HistFactory tool that ships with ROOT targeting binned analyses

- ▶ XML files organize the histograms
- ▶ conventions define model exactly
 - [CERN-OPEN-2012-016](https://arxiv.org/abs/1207.1332)
- ▶ command line tool creates likelihood

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

```
<!DOCTYPE Channel SYSTEM 'HistFactorySchema.dtd'>

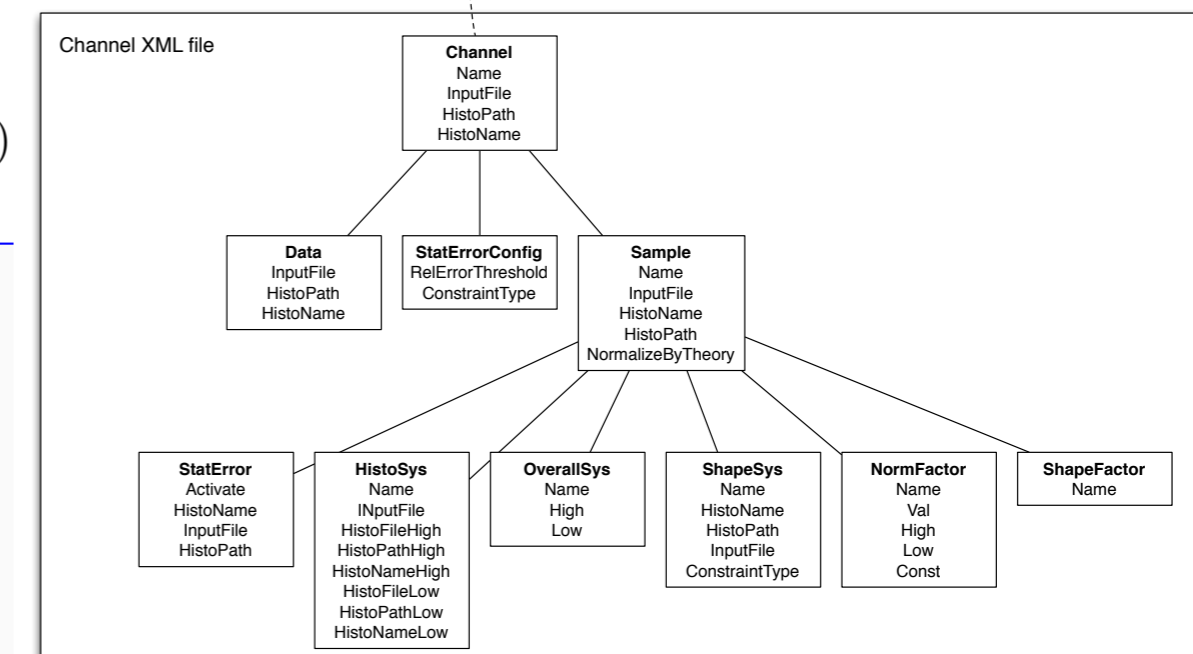
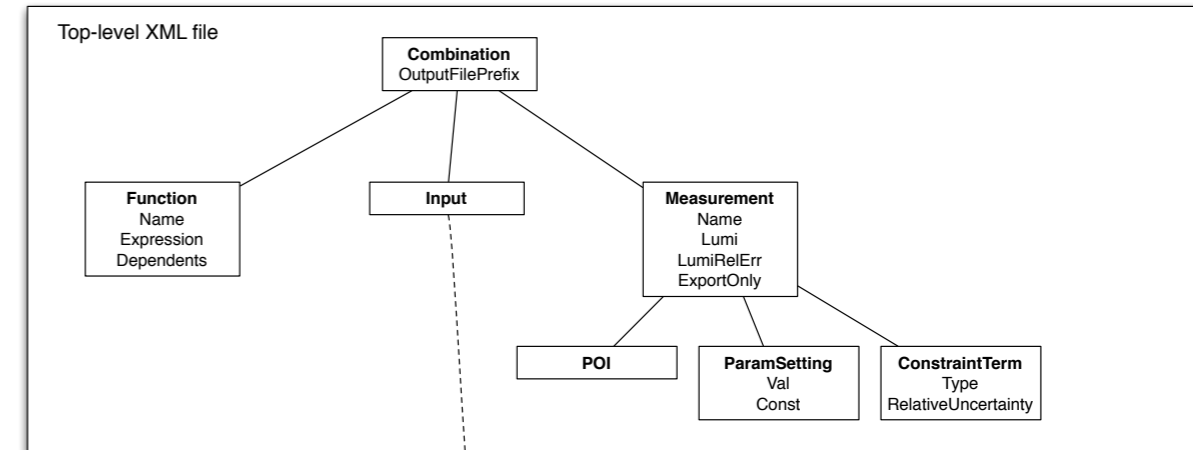
<Channel Name="A" InputFile="./data/ABCD.root" >
  <Data HistoName="A_data" HistoPath="" />

  <!-- This is the signal (eg. mu)-->
  <Sample Name="A_signal" HistoPath="" HistoName="unit_histogram">
    <!-- now mu is number of events-->
    <NormFactor Name="mu" Val="1" Low="0" High="200" />
    <OverallSys Name="syst1" High="1.01" Low="0.99" />
  </Sample>

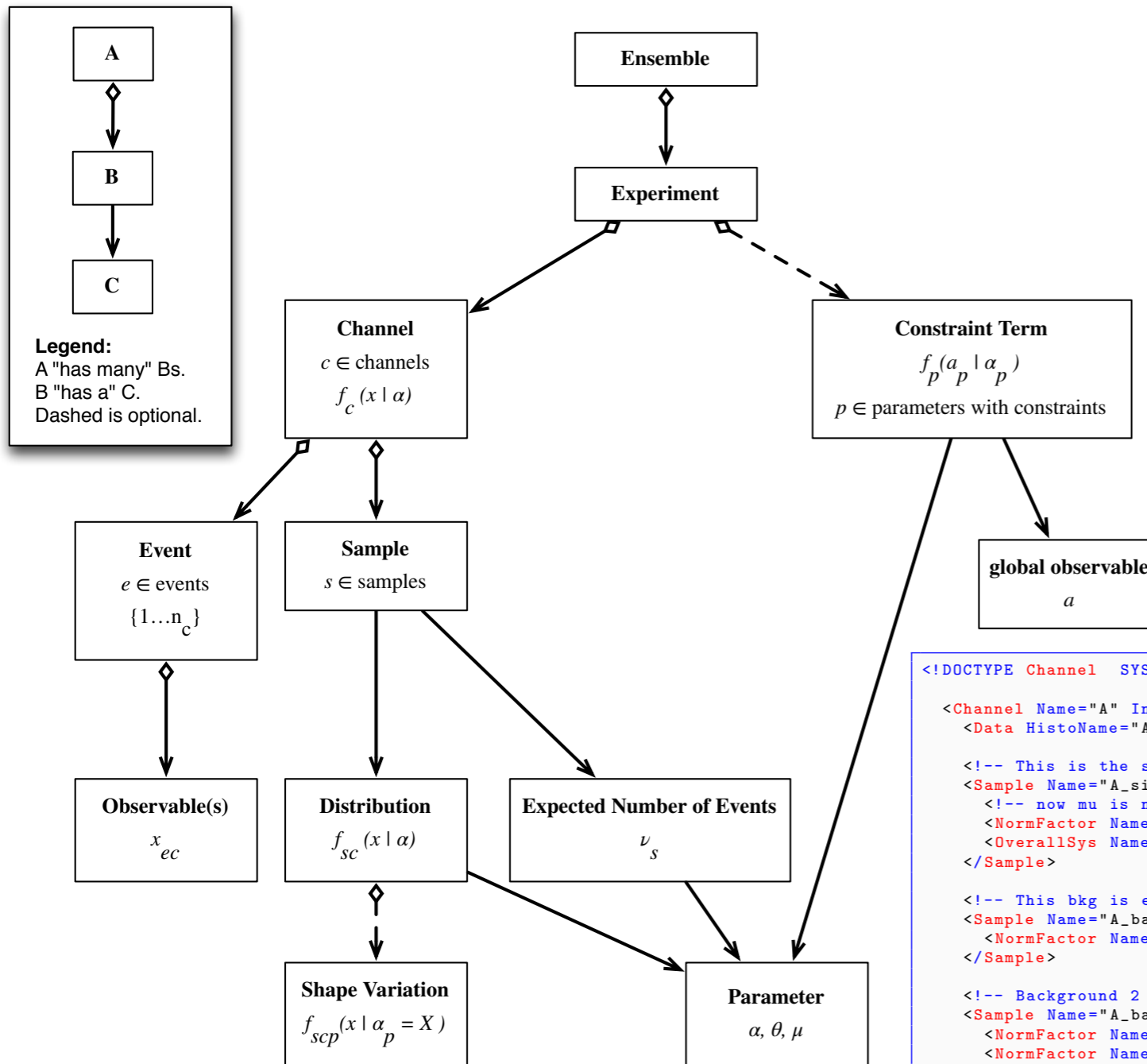
  <!-- This bkg is estimated from MC (eg. mu_A^K) -->
  <Sample Name="A_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram" >
    <NormFactor Name="mu_K_A" Val="100" Low="0" High="200" />
  </Sample>

  <!-- Background 2 is completely Data-Driven -->
  <Sample Name="A_backgroundDD" HistoPath="" NormalizeByTheory="False" HistoName="unit_histogram" >
    <NormFactor Name="mu_D_U" Val="100" Low="24500" High="26000" />
    <NormFactor Name="etaB" Val="1" Low="0." High="0.02" Const="False" />
    <NormFactor Name="etaC" Val="1" Low="0." High="0.3" Const="False" />
    <!-- NormFactor and ShapeFactor same for a 1-bin histogram. But we can name NormFactor-->
  </Sample>

</Channel>
```



XML SPECIFICATION



Provides machine-readable semantics for histograms used to build statistical models

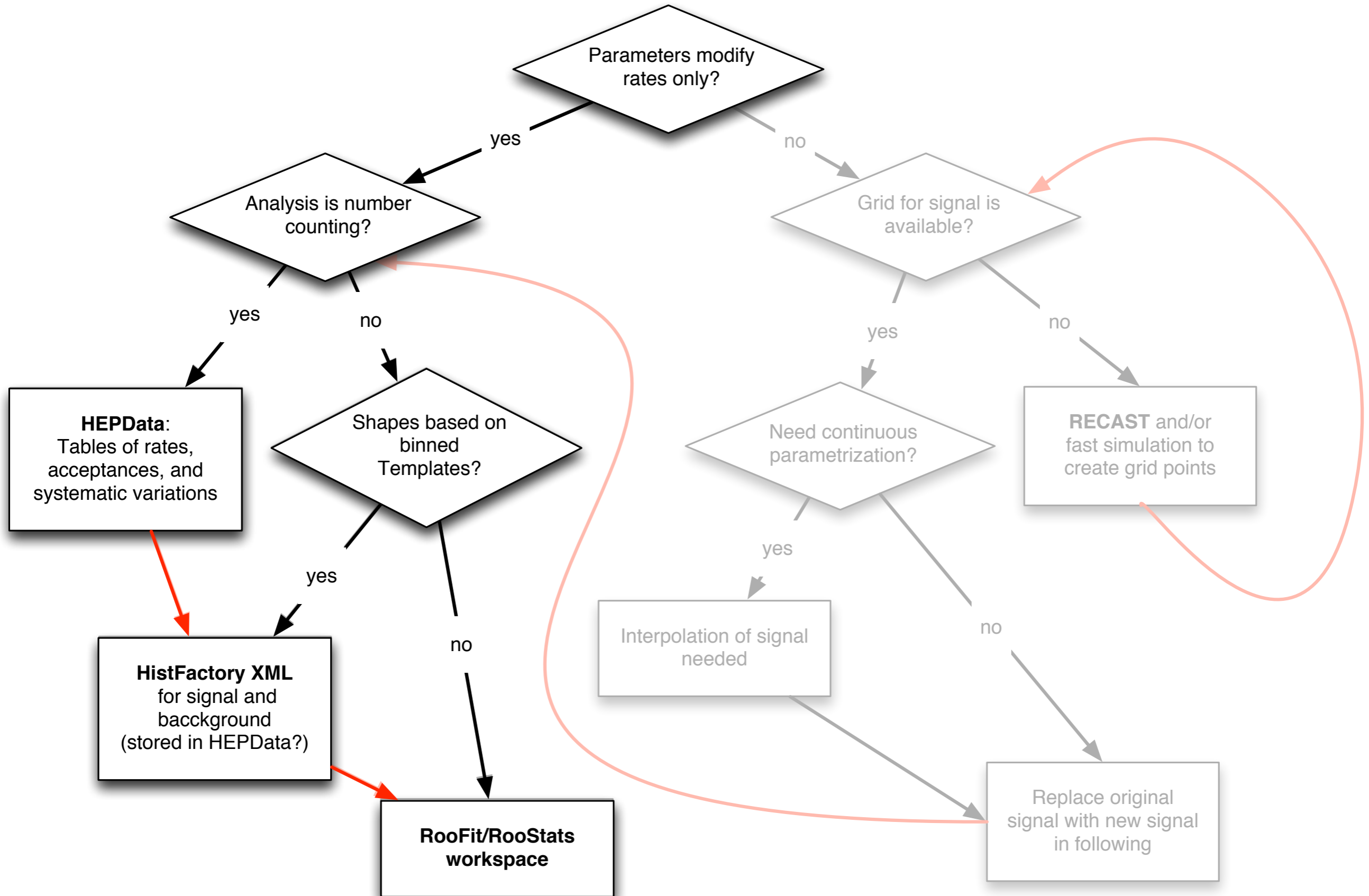
► Declarative specification defines model exactly

- [CERN-OPEN-2012-016](#)

```
<!DOCTYPE Channel SYSTEM 'HistFactorySchema.dtd'>
<Channel Name="A" InputFile="./data/ABCD.root" >
  <Data HistoName="A_data" HistoPath="" />
  <!-- This is the signal (eg. mu)-->
  <Sample Name="A_signal" HistoPath="" HistoName="unit_histogram">
    <!-- now mu is number of events-->
    <NormFactor Name="mu" Val="1" Low="0" High="200" />
    <OverallSys Name="syst1" High="1.01" Low="0.99" />
  </Sample>
  <!-- This bkg is estimated from MC (eg. mu_A^K) -->
  <Sample Name="A_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram" >
    <NormFactor Name="mu_K_A" Val="100" Low="0" High="200" />
  </Sample>
  <!-- Background 2 is completely Data-Driven -->
  <Sample Name="A_backgroundDD" HistoPath="" NormalizeByTheory="False" HistoName="unit_histogram" >
    <NormFactor Name="mu_D_U" Val="100" Low="24500" High="26000" />
    <NormFactor Name="etaB" Val="1" Low="0." High="0.02" Const="False" />
    <NormFactor Name="etaC" Val="1" Low="0." High="0.3" Const="False" />
    <!-- NormFactor and ShapeFactor same for a 1-bin histogram. But we can name NormFactor-->
  </Sample>
</Channel>
```

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

A Reinterpretation Roadmap



AAHEP7 Information Provider Summit

1-3 April 2014
Stony Brook University
US/Eastern timezone



Overview

Timetable

Participants

Slides

Meeting Location

Accommodations

Travel Information

The **7th Summit of Information Providers in Astronomy, Astrophysics and High Energy Physics** will be hosted by APS and held at Stony Brook University. Attendees will include representatives of INSPIRE, ADS, arXiv, journal publishers, and others involved in providing access to information in these fields.

This meeting follows on from 6 previous meetings in the series:

- 2012, CERN: <http://indico.cern.ch/conferenceDisplay.py?confId=209651>
- 2011, Cornell: <https://indico.cern.ch/conferenceDisplay.py?confId=128826>
- 2010, Harvard: <http://conf.adsabs.harvard.edu/AAHEP4/>
- 2009, Fermilab: <http://indico.fnal.gov/conferenceProgram.py?confId=2473>
- 2008, DESY: <https://indico.desy.de/conferenceDisplay.py?confId=800>
- 2007, SLAC: <https://indico.cern.ch/conferenceDisplay.py?confId=11611>

By Invitation Only

Support

✉ [Annette.Holtkamp@cer...](mailto:Annette.Holtkamp@cern.ch)

✉ [Thorsten.Schwander@c...](mailto:Thorsten.Schwander@cern.ch)

✉ [bhecker@slac.stanford....](mailto:bhecker@slac.stanford.edu)



Starts 1 Apr 2014, 09:00
Ends 3 Apr 2014, 17:00
US/Eastern



Stony Brook University
Wang Center, Lecture Hall and Chapel
Stony Brook, New York, US



[Annette Holtkamp](#)
[Bernard Louis Hecker](#)
[Hans-Thorsten Schwander](#)
[Mark Doyle](#)



📁 **notes**

📄 test_document.txt

CYBERINFRASTRUCTURE

The screenshot shows the INSPIRE website interface. At the top, there's a navigation bar with links like 'HEP', 'HEPNAMES', 'INSTITUTIONS', etc. Below that is a search bar with the text 'HEP Search High-Energy Physics Literature Database'. A sidebar on the right contains 'HEP' related links and 'INSPIRE News'.

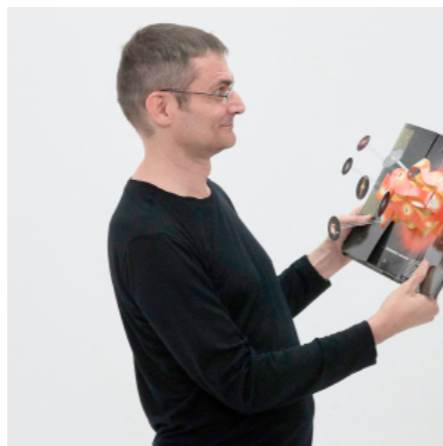
The screenshot shows the HEPData website. It features a purple header with the HEPData logo and the text 'High Energy Physics Data Repository'. Below the header, there's a search bar and a section titled 'Data from the LHC' with icons for ATLAS, ALICE, CMS, and LHCb.

The screenshot shows the Zenodo website. It has a blue header with the Zenodo logo and the tagline 'Research. Shared.'. Below the header, there's a search bar and a section for 'Software' with 'Recent Uploads'.

The screenshot shows the INVENIO website. It features a blue header with the INVENIO logo and the tagline 'Powering Open Science'. Below the header, there's a section for 'Framework' and 'RDM' (Research Data Management) with icons for 'ILS' (Integrated Library System).

The screenshot shows the RECAST website. It features a blue header with the RECAST logo and the tagline 'A framework for extending the impact of existing analyses performed by high-energy physics experiments.'. Below the header, there's a section for 'Analyses' with a search bar and a list of analyses.

The screenshot shows the CERN Analysis Preservation website. It features a blue header with the CERN logo and the tagline 'ANALYSIS PRESERVATION'. Below the header, there's a section for 'Visualiser' showing a complex network diagram of analysis data.



Tibor Simko



Sünje Dallmeier-Tiessen



HEPData: a repository for high energy physics data

Eamonn Maguire¹, Lukas Heinrich² and Graeme Watt³

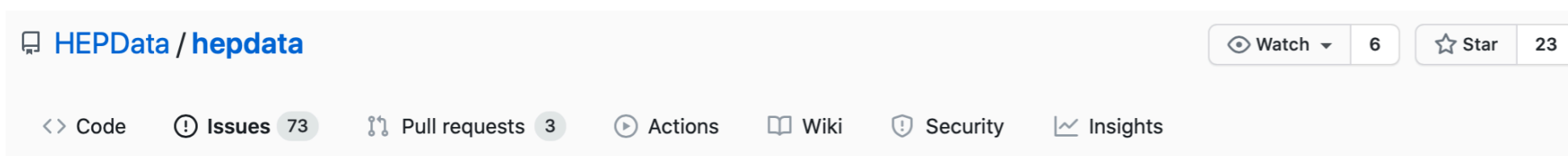
¹ CERN, Geneva, Switzerland

² Department of Physics, New York University, New York, USA

³ IPPP, Department of Physics, Durham University, Durham, UK

In future we plan to support a mixed YAML/ROOT input format where metadata is provided in YAML files (as before), but numerical values are extracted from ROOT objects and converted to the standard YAML format. HistFactory [7] is a framework used in many ATLAS studies for statistical analysis (such as determining exclusion contours). It encodes the full likelihood (including systematic uncertainties) of a measurement using semantic XML and histograms stored in ROOT files. Some preliminary work has been done to extract HEPData tables in the standard YAML format directly from a HistFactory configuration. Furthermore, work has begun on expanding the set of natively supported data types beyond a simple table to allow for richer datasets such as HistFactory configurations or simplified likelihoods [8]. The archival of such likelihood data in a lossless format could then be used by various reinterpretation packages.

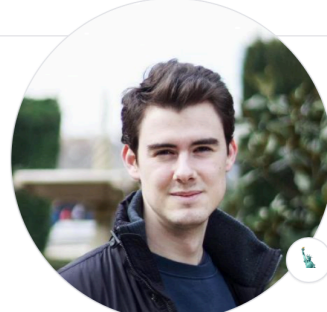
CHEP 2016: 10.1088/1742-6596/898/10/102006



submission: provide native support for individual "pyhf" JSON files #164



Alison Clarke



Sinclert Pérez

REIMPLEMENTING HISTFACTORY

Around 2016 we realized that there was a huge opportunity to leverage machine learning frameworks like TensorFlow for probability models.

- GPUs and automatic differentiation!

In 2017 we began porting HistFactory to pure python: **pyhf**



Lukas Heinrich
CERN

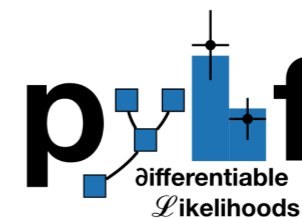


Matthew Feickert
University of Illinois
Urbana-Champaign



Gordon Stark
UCSC SCIPP

www.scikit-hep.org/pyhf



DOI 10.5281/zenodo.1169739 Scikit-HEP Project

See Matthew Feickert's SciPy 2020 talk

pyhf team decided to use the widely-used JSON format to store the models

- Independence from binary formats and ROOT dependencies

Probabilistic programming frameworks

RooFit serves us well, but shows limits in terms of scalability.

Using a data flow graph framework, RooFit would be **distributed**, **GPU-enabled** and automatically **differentiable**.

Feasibility? Certainly **within reach**! As illustrated by our tentative proof-of-concepts **carl.distributions** [Gilles Louppe] and **tensorprob** [Igor Babuschkin, now at DeepMind]. See also Edward.

Edward A library for probabilistic modeling, inference, and criticism.

Edward is a Python library for probabilistic modeling, inference, and criticism. It is a testbed for fast experimentation and research with probabilistic models, ranging from classical hierarchical models on small data sets to complex deep probabilistic models on large data sets. Edward fuses three fields: Bayesian statistics and machine learning, deep learning, and probabilistic programming.

It supports modeling with

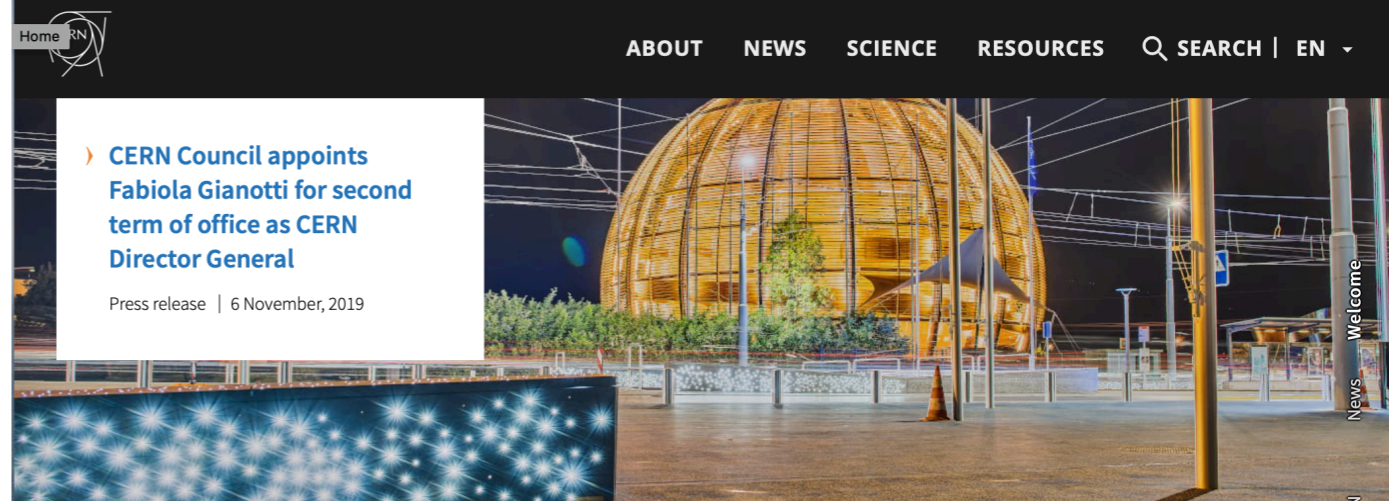
Dustin Tran
Ph.D. Student
Columbia University
dustin@cs.columbia.edu (@dustintran)
<http://dustintran.com>

Matthew Feickert
High Energy Physics Ph.D. Candidate
Southern Methodist University
matthew.feickert@cern.ch or mfeickert@smu.edu
GitHub: matthewfeickert @HEPfeickert

MAKING IT STANDARD

20 years later: community embraces publishing likelihoods as a standard

- Moved to JSON format



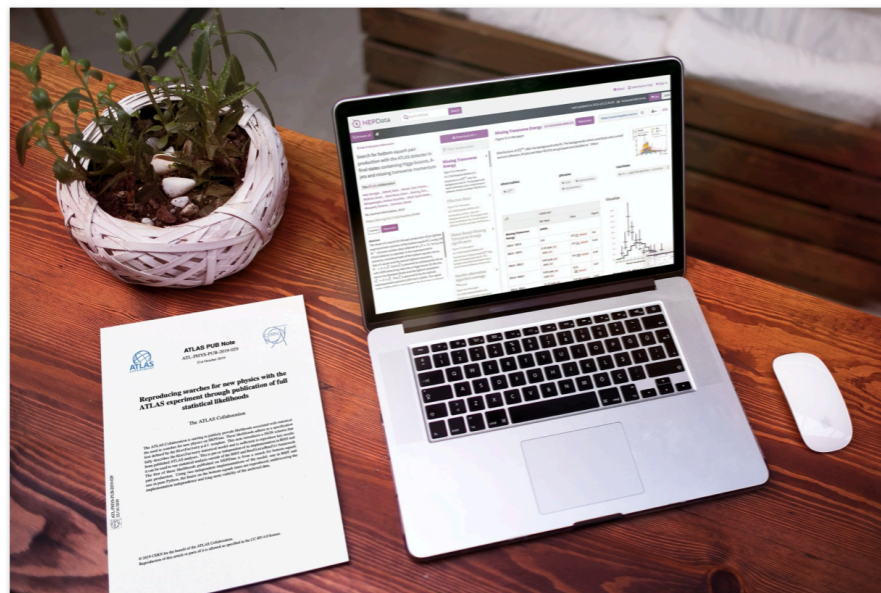
LATEST NEWS



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

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

9 JANUARY, 2020 | By Katarina Anthony



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

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

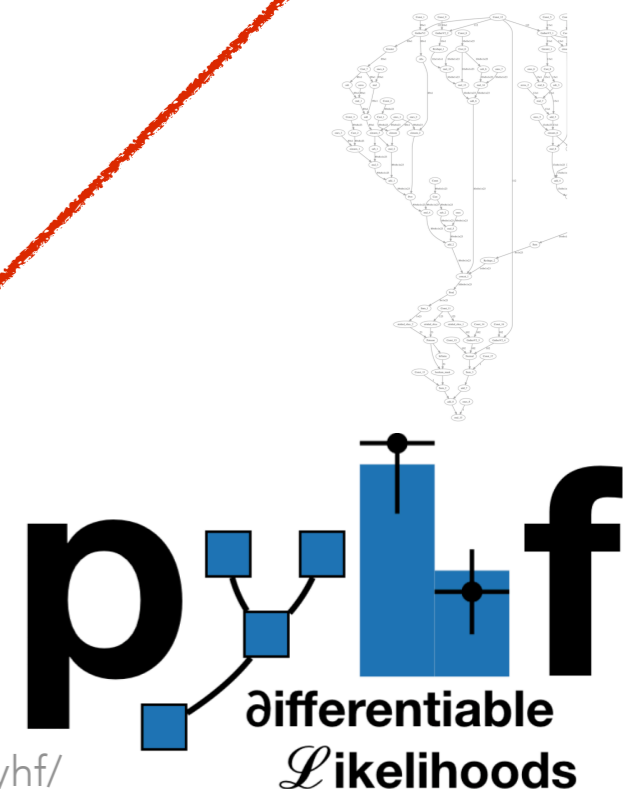
Display a menu y from the ATLAS collaboration, with the first open release of full analysis likelihoods

Related Articles



[View all news >](#)

<https://scikit-hep.org/pyhf/>





Now: full likelihoods !!

ATL-PHYS-PUB-2019-029 (05 Aug 2019)

- Plain-text serialisation of HistFactory workspaces, JSON format
 - Provides background estimates, changes under systematic variations, and observed data counts at the same fidelity as used in the experiment.



gz File

Archive of full likelihoods in the HistFactory JSON format described in ATL-PHYS-PUB-2019-029. Provided are 3 statistical models labeled RegionA, RegionB and RegionC respectively each in their own sub-directory. For each model the background-only model is found in the file named 'BkgOnly.json'. For each model a set of patches for various signal points is provided.

Download

	Description	Modification	Constraint Term c_{χ}	Input
constrained	Uncorrelated Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Pois}(r_b = \sigma_b^{-2} \rho_b = \sigma_b^{-2} \gamma_b)$	σ_b
	Correlated Shape	$\Delta_{scb}(\alpha) = f_p(\alpha \Delta_{scb, \alpha=-1}, \Delta_{scb, \alpha=1})$	$\text{Gaus}(a=0 \alpha, \sigma=1)$	$\Delta_{scb, \alpha=\pm 1}$
	Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p(\alpha \kappa_{scb, \alpha=-1}, \kappa_{scb, \alpha=1})$	$\text{Gaus}(a=0 \alpha, \sigma=1)$	$\kappa_{scb, \alpha=\pm 1}$
	MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}(a_{\gamma_b} = 1 \gamma_b, \delta_b)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
	Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}(l = \lambda_0 \lambda, \sigma_\lambda)$	$\lambda_0, \sigma_\lambda$
free	Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
	Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

Rate modifications defined in HistFactory for bin b , sample s , channel c .

- Usage: RooFit, **pyhf**
- Target: long-term data/analysis preservation, reinterpretation purposes

So far available for 4/12 SUSY analyses with 139 fb⁻¹

SUSY-2018-31 (1908.03122)	multi-b sbottom: 2b+2H(bb)
SUSY-2018-04 (1911.06660)	stau search, 2 hadr. taus
SUSY-2019-08 (1909.09226)	1 lept. + H(bb), EW-ino
SUSY-2018-06 (1912.08479)	3 lept. EW-ino



Reinterpretation Forum Report 2020

“.... In fact, many of the data products discussed here, such as [signal/background yields and correlations](#), are used by the various external reinterpretation packages to [construct likelihoods](#). Whilst extremely useful, the likelihoods constructed from these products are however always [only an approximation](#) to the true underlying experimental likelihood. The reinterpretation workflow can be greatly facilitated and rendered much more precise if the original likelihood of the analysis is published in full. [We strongly encourage the movement towards the publication of full experimental likelihoods wherever possible.](#)”

“ATLAS has recently started to do this using a JSON serialisation of the likelihood [...] The provision of this full likelihood information is much appreciated and we hope that it will become a standard, as it **greatly improves the quality of any reinterpretation.**”

Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2
arXiv:2003.07868, SciPost Phys. 9, 022 (2020)

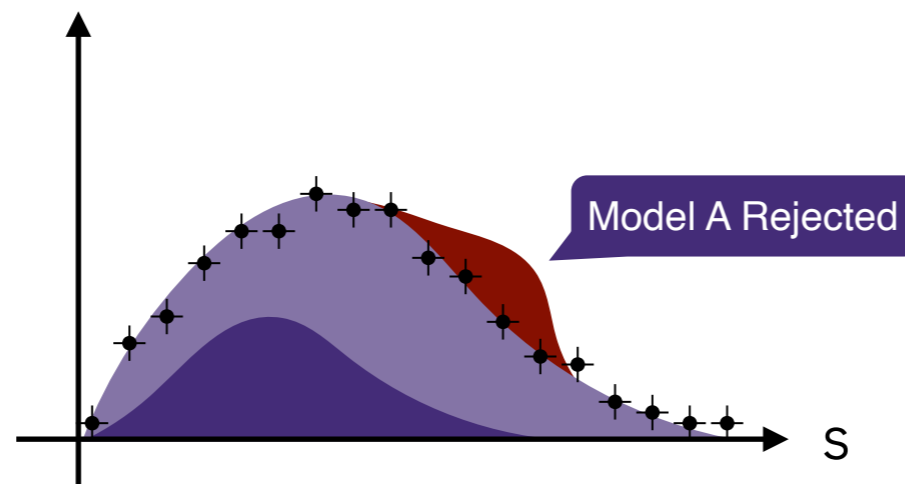
2) Reinterpretation

THUMBNAIL SKETCH OF ANALYSIS

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- We build a statistical model $p(s \mid \text{Model A}, \theta_A)$
- Then we test the hypothesis and write a paper



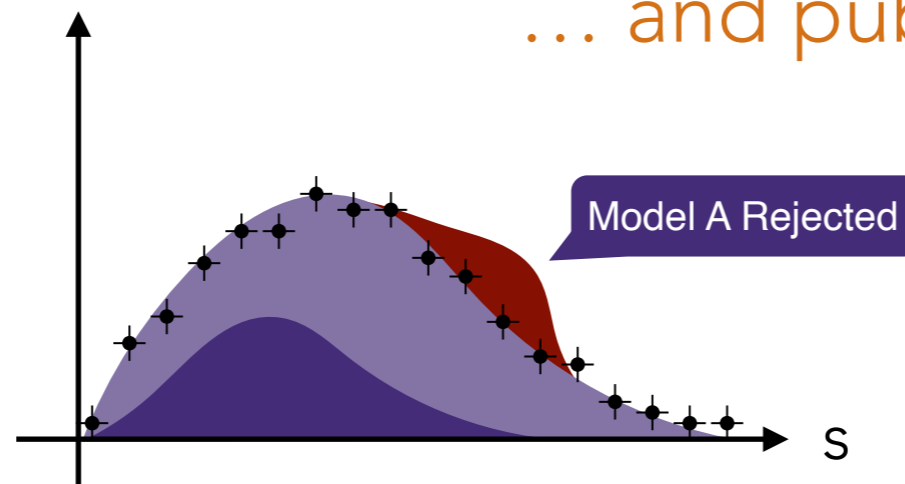
observed **data** + predicted distribution for the alternate in **Model A**

THUMBNAIL SKETCH OF ANALYSIS

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic \mathbf{s} that can distinguish between different hypotheses we are considering.

- We build a statistical model $p(s | \text{Model A}, \theta_A)$
- Then we test the hypothesis and write a paper
... and publish $p(s | \text{Model A}, \theta_A)$!



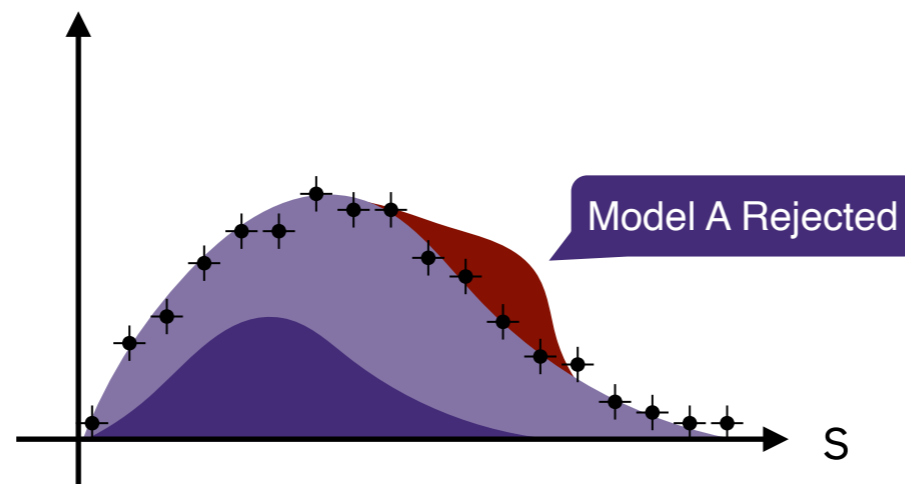
observed **data** + predicted distribution for the alternate in **Model A**

THUMBNAIL SKETCH OF HEP ANALYSIS

We select a small subset of the collision events relevant for testing the hypotheses we are considering.

And we design a summary statistic s that can distinguish between different hypotheses we are considering.

- ... and graduate students graduate, analysis code rots, and it would be difficult to reproduce or reuse this work



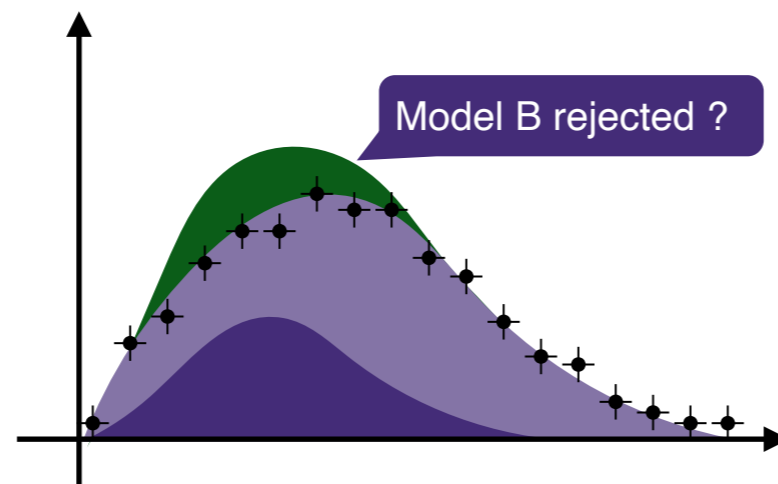
observed **data** + predicted distribution for the alternate in **Model A**

REINTERPRETATION

The statistical model $p(s | \text{Model A}, \theta_A)$ is great for combinations and studies within Model A

But isn't useful for answering questions about Model B

- the efficiency, acceptance, and distribution $p(s | \text{Model B}, \theta_B)$ for the new signal will be different

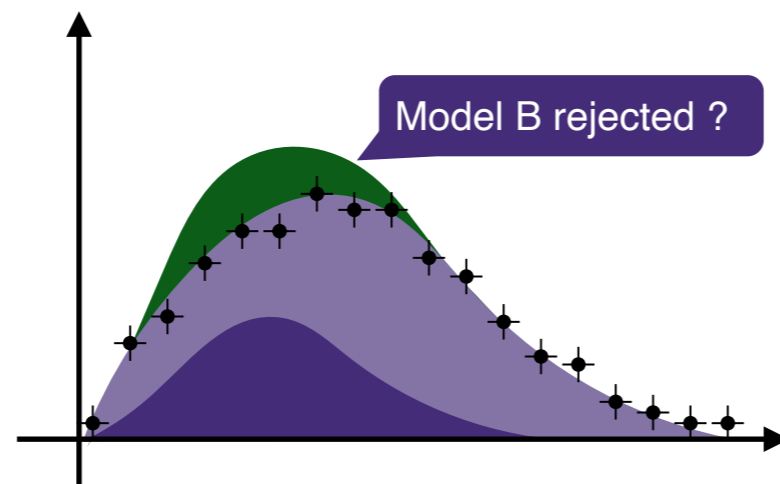


observed **data** + predicted distribution for the alternate in **Model B**

REINTERPRETATION

Typically HEP experiments feel compelled to design an entirely new analysis pipeline targeting **Model B**, but

- that is very time consuming and labor intensive, and
- sometimes **Model A** and **Model B** have a lot in common, and the original analysis will also be sensitive to **Model B**

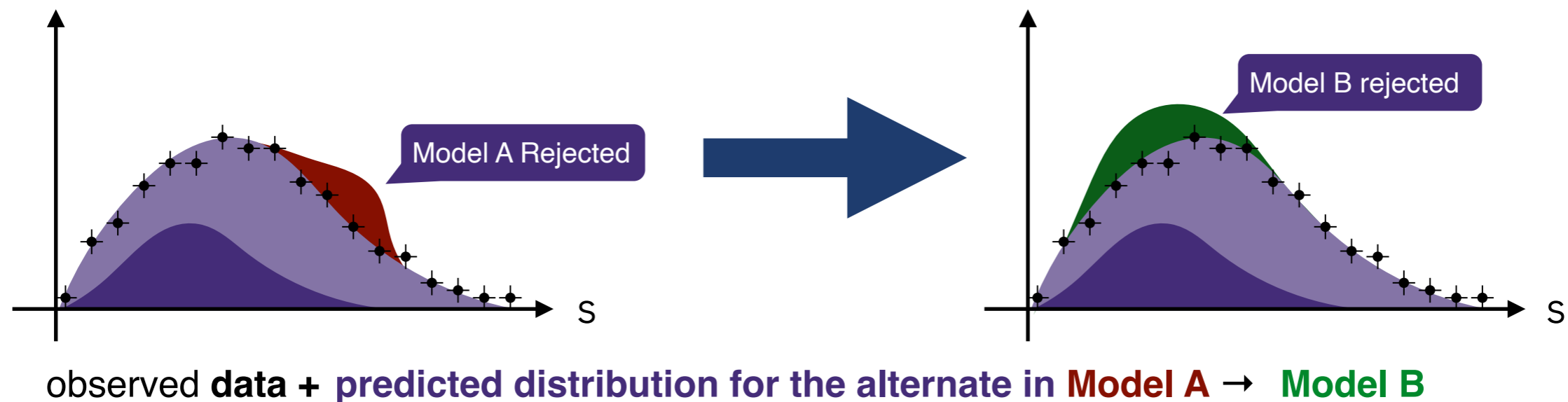


observed **data** + predicted distribution for the alternate in **Model B**

RECASTING

If we can capture the definition of the summary $\mathbf{s}(\mathbf{x})$ and the event selection, then we can reuse the existing analysis (prediction for the null and observation in the data)

- We just need to run simulated events for **Model B** through the pipeline and test the new signal+background alternate hypothesis



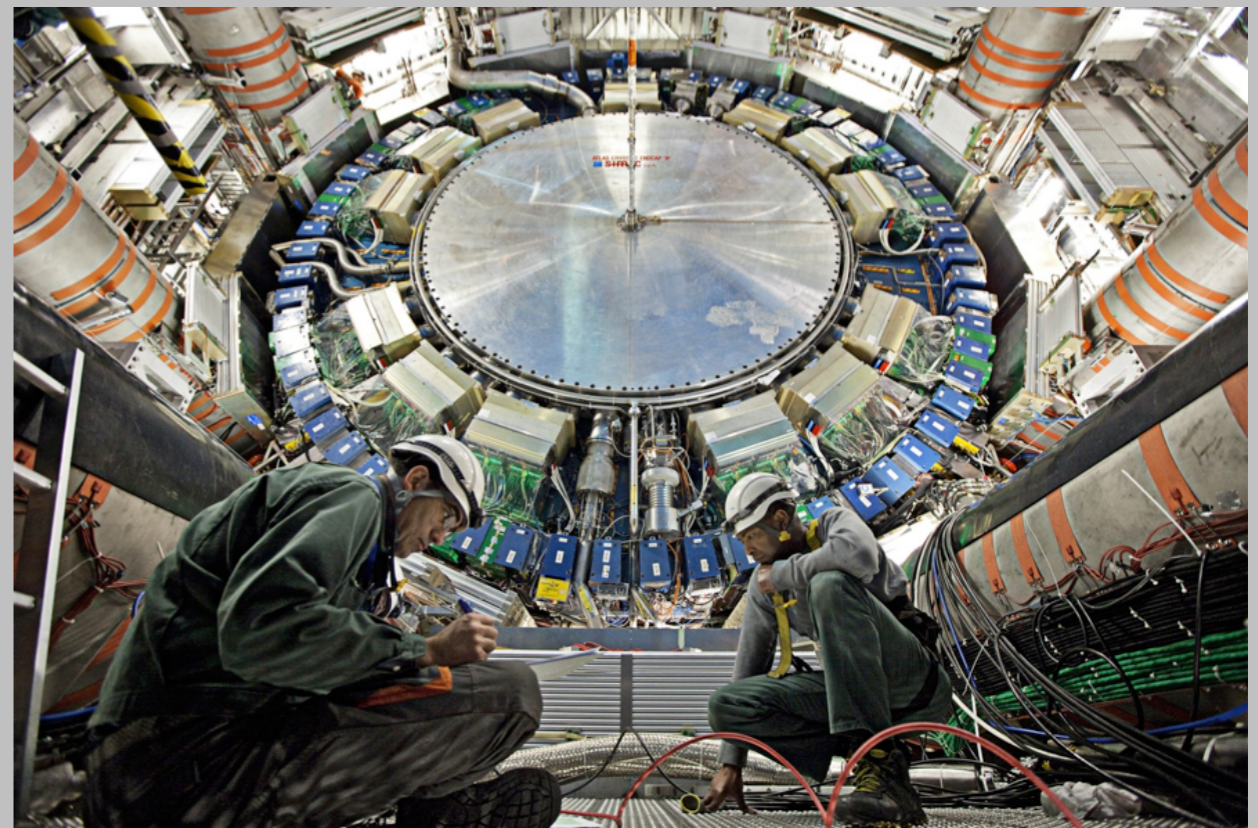
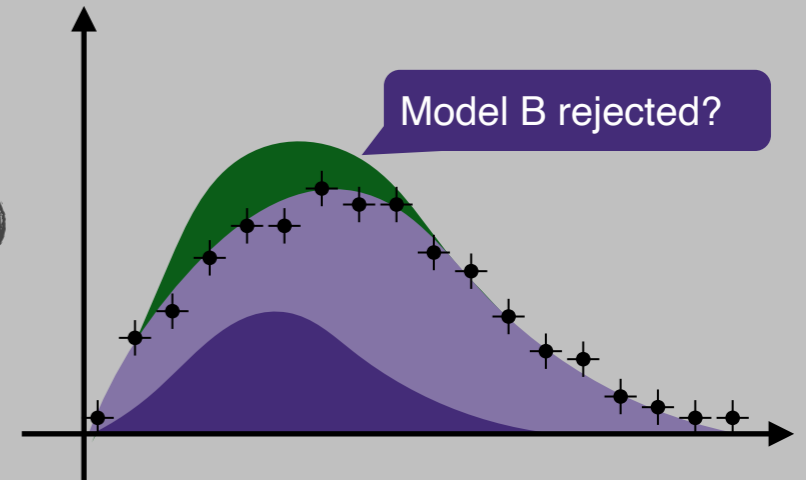
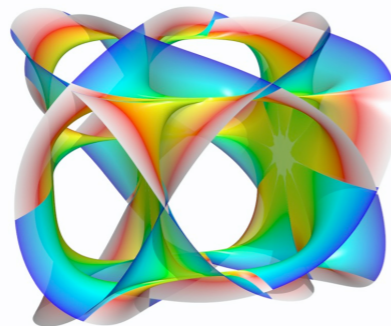
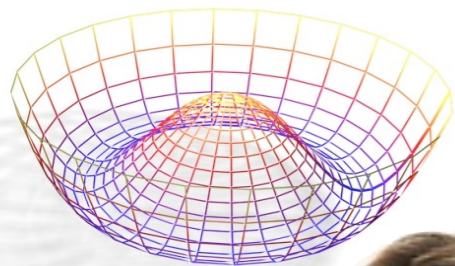
THEORY

SERVICE

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

Q

A



RECAST

We proposed RECAST framework in Oct 2010

- People said it couldn't be done, our workflows are too complicated



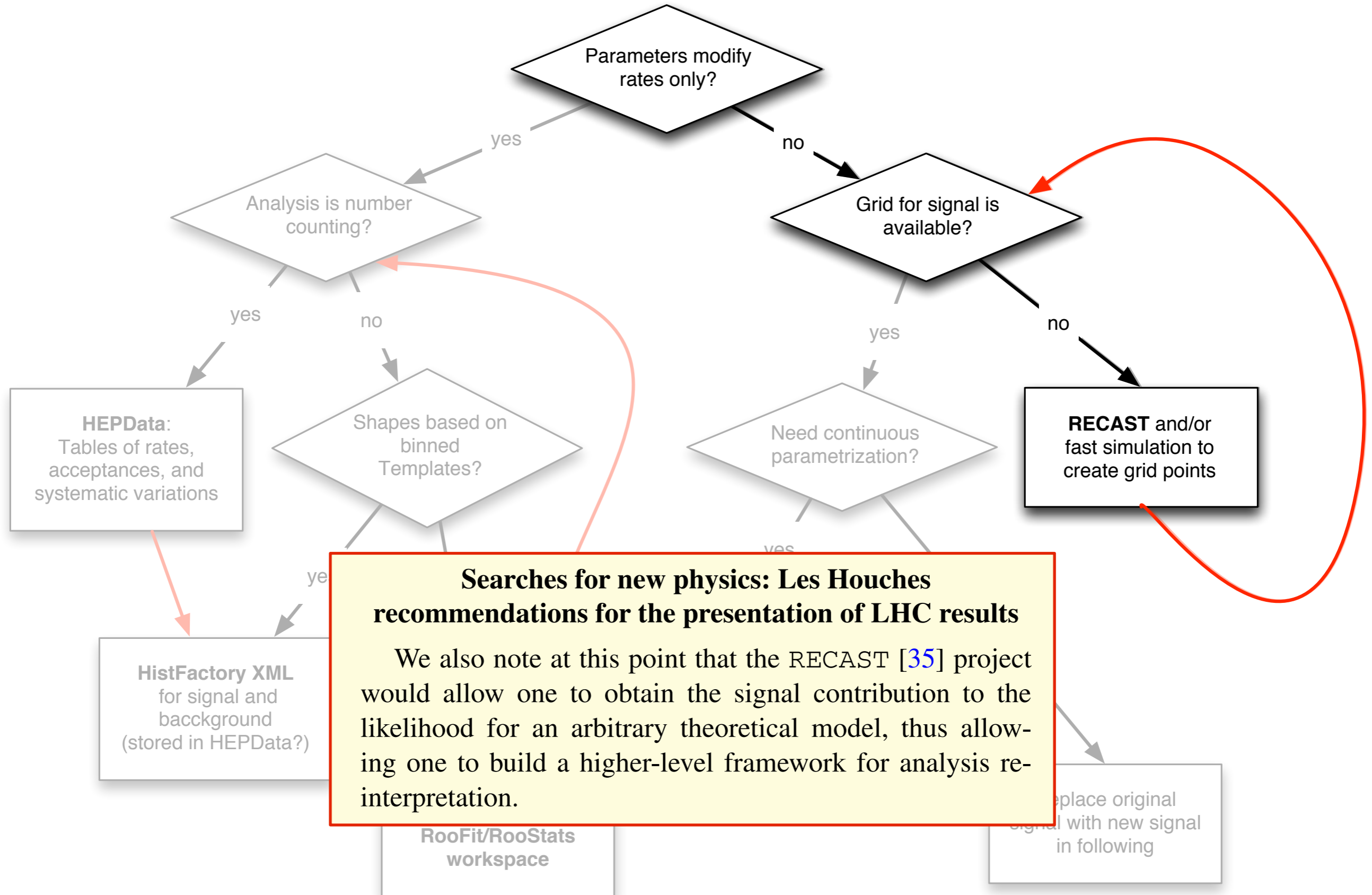
RECAST

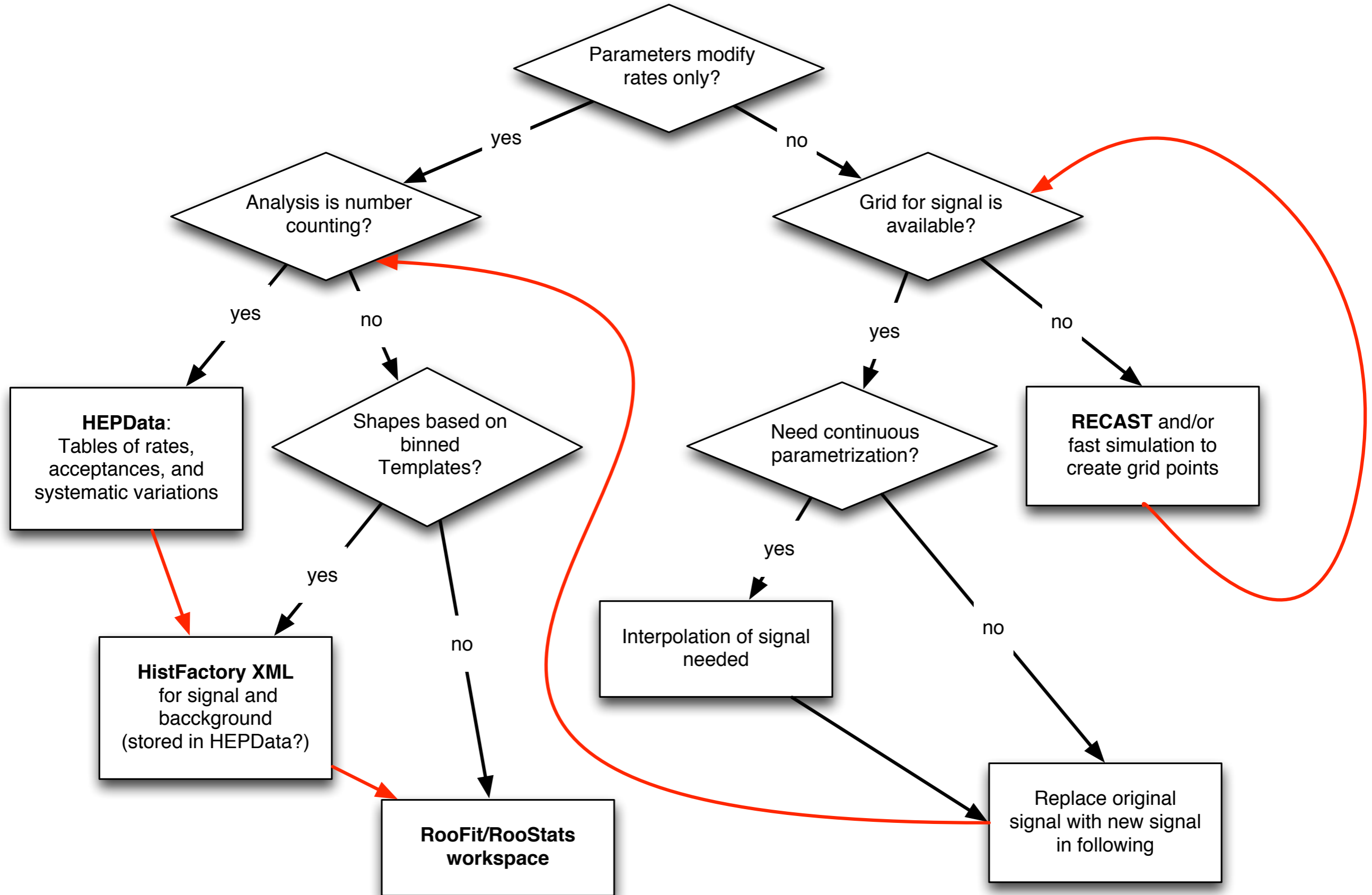
Extending the Impact of Existing Analyses

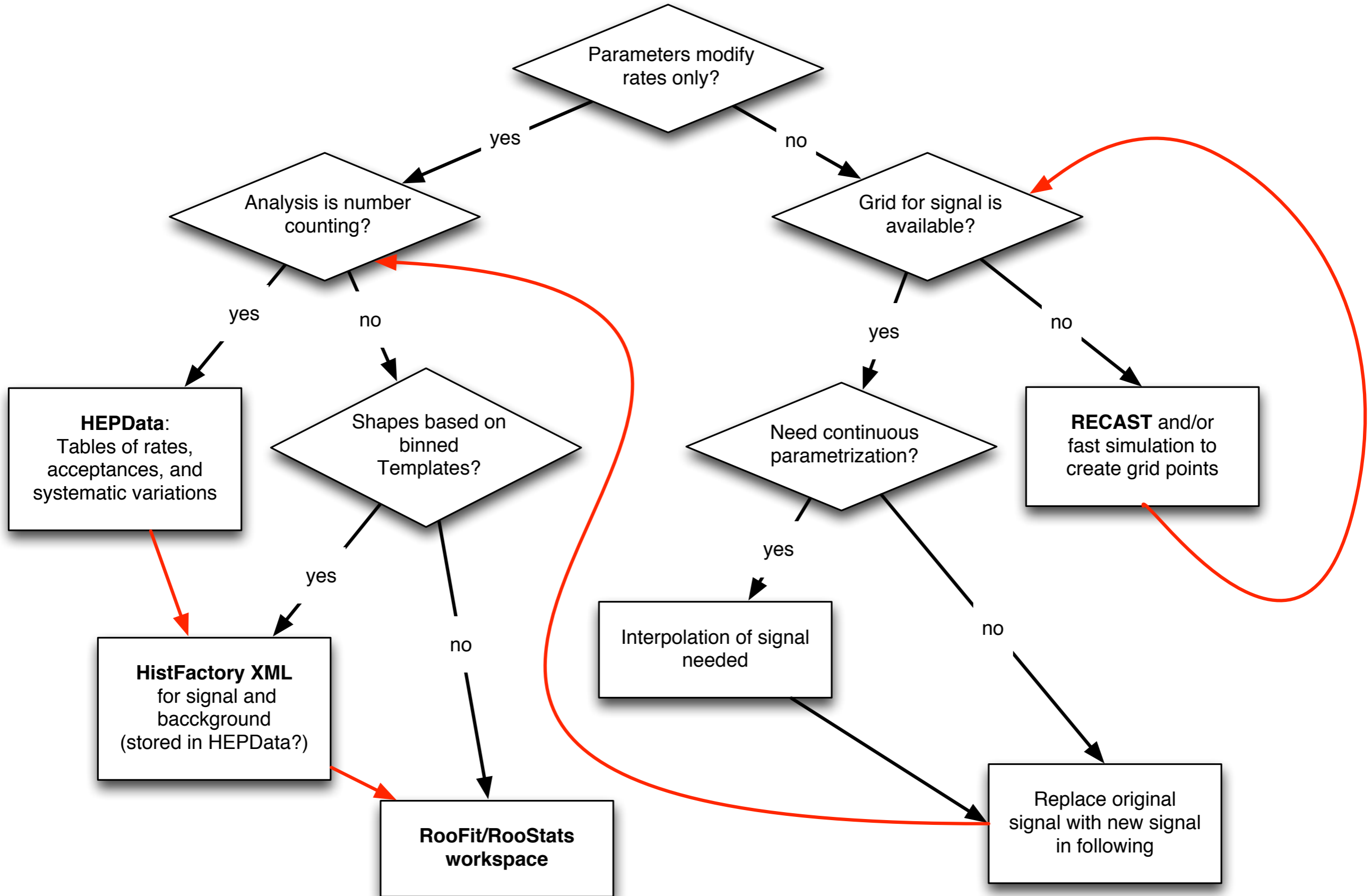
Kyle Cranmer and Itay Yavin

Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003

ABSTRACT: Searches for new physics by experimental collaborations represent a significant investment in time and resources. Often these searches are sensitive to a broader class of models than they were originally designed to test. We aim to extend the impact of existing searches through a technique we call *recasting*. After considering several examples, which illustrate the issues and subtleties involved, we present RECAST, a framework designed to facilitate the usage of this technique.







Higgs discovery!
RECAST efforts stalled for a couple of years



Data Science @ LHC 2015 Workshop

9-13 November 2015
CERN

Europe/Zurich timezone

There is a [live webcast](#) for this event.



Overview

[Program](#)

[Reading materials](#)

[Speaker List](#)

[Registration](#)

[Participant List](#)

[Videoconference Rooms](#)

[Poster](#)

[Network Connection
Request Forms](#)

[CERN ACCESS INFO](#)

[Orientation](#)

The LHC experiments have been producing the largest amount of complex data. 100TB/s of real-time data analyses and analyses of 100 EB of data are anticipated and planned for. The field of data science beyond statistical methods has been producing advanced, intelligent methods for data analysis, pattern recognition and model inference. This workshop will engage the two communities towards cross exchanges and applications that can forge accelerated progress in big basic science questions.

Some of the topics that will be addressed include cutting edge pattern recognition methods for elementary particle identification; intelligent detectors that learn from their failures and self-adjust to increase their performance efficiency; fast reconstruction of charged particle tracks; high-rate event selection algorithms that learn to select rare physics processes; advanced data techniques that can guide discovery and other challenges that can profit from advanced computational methods and resources.

The workshop includes plenary presentations, tutorials and hands-on hackathon-type of ML exercises as well as directed and undirected discussion and brainstorming time.

Subscribe to the [participants mailing list](#) for discussions on the topic and announcements before and during the workshop by sending email to: HEP-data-science+subscribe@googlegroups.com

Follow the workshop official account [@DataScienceLHC](#). Feel free to tweet using the recommended hash tag **#DSLHC15**

2015: FROM METADATA TO WORKFLOWS

CERN Analysis Preservation

- “closed counterpart” to CERN Open Data that captures the complexity of
 - The data
 - The processing steps
 - Code involved
 - Documentation, Physics information
 - Peer review, QAi.e. all the information contributing to the research claim/presentation/publication to enable future reuse



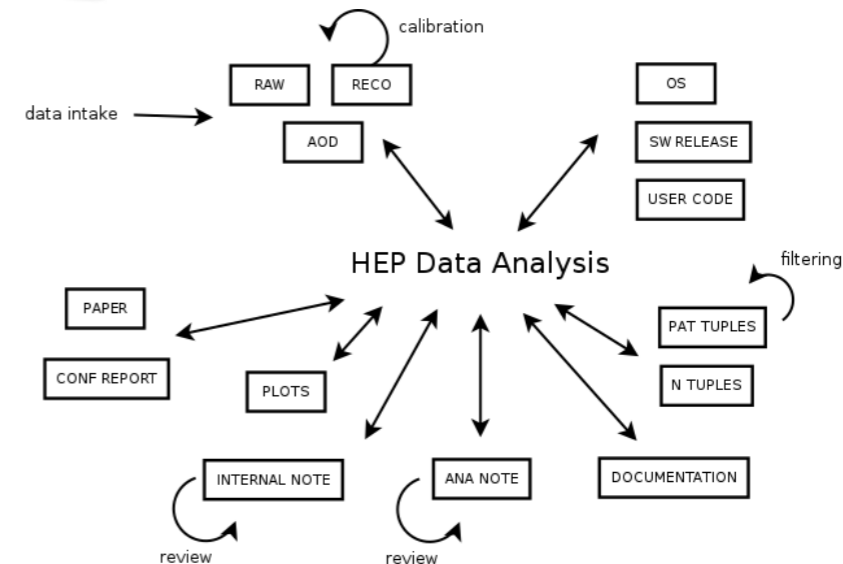
Open Data @CERN

Sünje Dallmeier-Tiessen
for many others in GS-SIS and IT-CIS



opendata

Preserve an analysis



Lukas and I began working on tools to preserve HEP workflows



Lukas Heinrich

Yadage and Packtivity – analysis preservation using parametrized workflows

Kyle Cranmer¹ and Lukas Heinrich¹

¹ Department of Physics, New York University, New York, USA

E-mail: lukas.heinrich@cern.ch

Abstract. Preserving data analyses produced by the collaborations at LHC in a parametrized fashion is crucial in order to maintain reproducibility and re-usability. We argue for a declarative description in terms of individual processing steps – “packtivities” – linked through a dynamic directed acyclic graph (DAG) and present an initial set of JSON schemas for such a description and an implementation – “yadage” – capable of executing workflows of analysis preserved via Linux containers.

COLLABORATION Analysis 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer nec odio. Praesent libero. Sed cursus ante dapibus diam. Sed nisi. Nulla quis sem at nibh elementum imperdiet. Duis sagittis ipsum. Praesent mauris. Fusce nec tellus sed augue semper porta. Mauris massa. Vestibulum lacinia arcu eget nulla. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Curabitur sodales ligula in libero. Sed dignissim lacinia nunc.

Overview

[Publications](#)

[Files](#)

[Workflow](#)

[Measurements](#)

[Contributors](#)

[ReCASTs](#)

1 Publication >

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer nec odio. Praesent libero.
Eur.Phys.J. C76 (2016) 451, 2016
DOI [10.1140/epjc/s10052-016-4286-3](#)

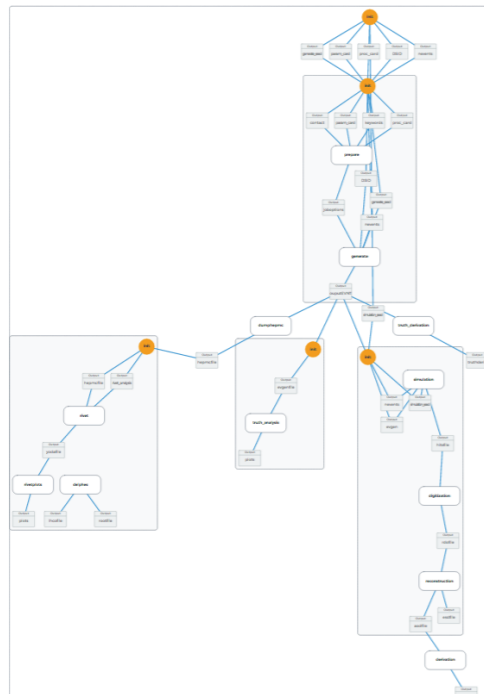
23 Files >

 Model 1	3.24MB
 P.D.F.	3.24MB
 Figure 1 Plot	3.24MB

2 Contributors >

-  **John Doe** CMS
-  **Mary Smith** CMS

Workflow >



Measurements >

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer nec odio. Praesent libero.
Vestibulum lacinia arcu eget nulla. Class aptent taciti sociosq.



reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.



Scalable

Support for remote compute clouds.



Reusable

Containerise once, reuse elsewhere. Cloud-native.



Free

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



The SCALFIN Project
scailfin.github.io



SHIFTING FROM REPRODUCIBILITY TO REUSE

Open is not enough

Xiaoli Chen^{1,2}, Sünje Dallmeier-Tiessen^{1*}, Robin Dasler^{1,11}, Sebastian Feger^{1,3}, Pamfilos Fokianos¹, Jose Benito Gonzalez¹, Harri Hirvonsalo^{1,4,12}, Dinos Kousidis¹, Artemis Lavasa¹, Salvatore Mele¹, Diego Rodriguez Rodriguez¹, Tibor Šimko^{1*}, Tim Smith¹, Ana Trisovic^{1,5*}, Anna Trzcinska¹, Ioannis Tsanaktsidis¹, Markus Zimmermann¹, Kyle Cranmer⁶, Lukas Heinrich⁶, Gordon Watts⁷, Michael Hildreth⁸, Lara Lloret Iglesias⁹, Kati Lassila-Perini⁴ and Sebastian Neubert¹⁰

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

- Reuse provides a forward-looking narrative, while reproducibility often perceived as backward-looking
- Reproducibility is a byproduct!
- Analysis Preservation distinct from reproducibility

- Helps with onboarding
- Empowers reuse, remixing, reproducibility
- Improves efficiency & equity

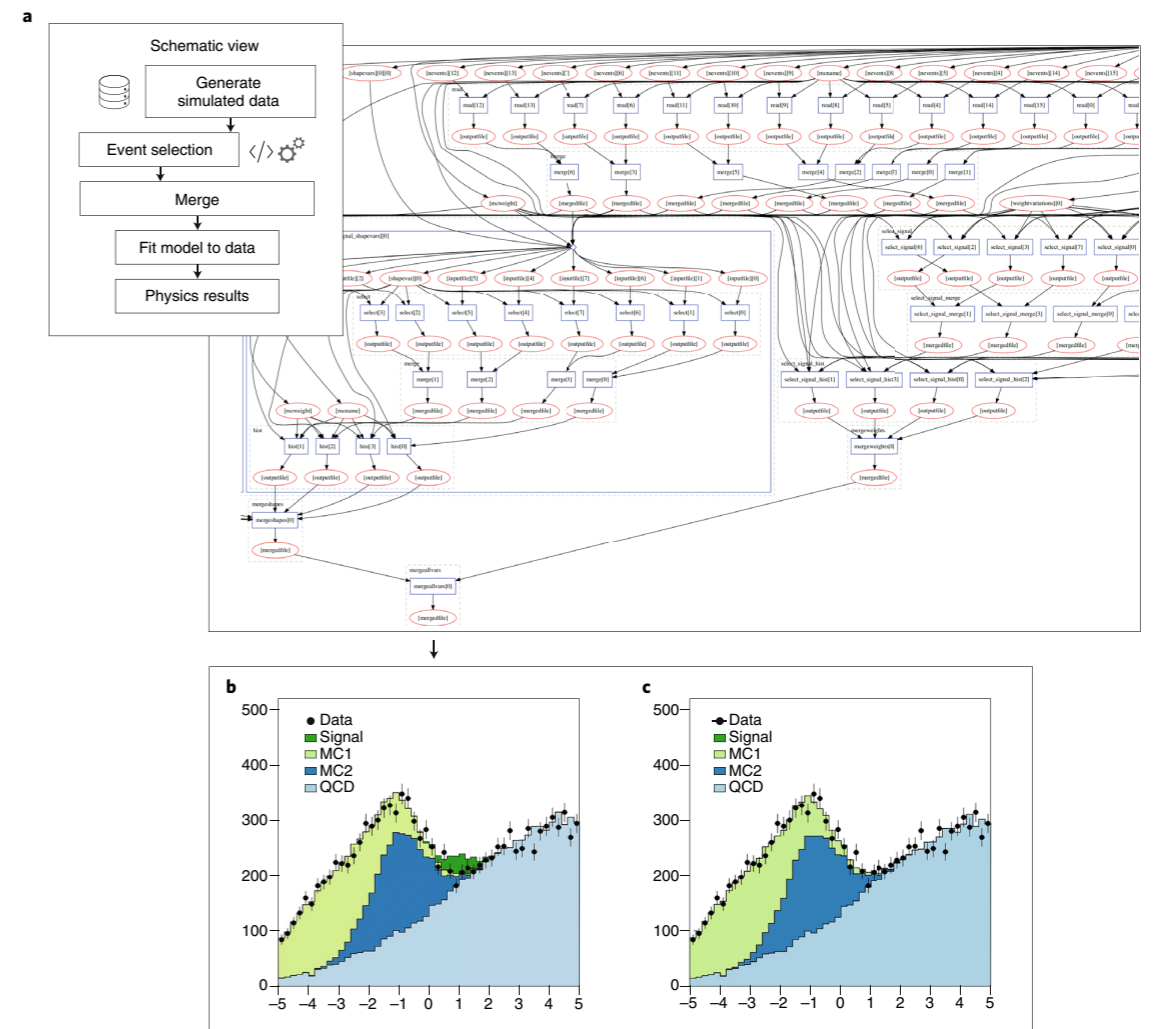


Fig. 2 | Example of a complex computational workflow on REANA mimicking a beyond the standard model (BSM) analysis. This figure shows an example where the experimental data is compared to the predictions of the standard model with an additional hypothesized signal component. The example permits one to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

RECAST IN ACTION

ATLAS has started using RECAST to reinterpret SUSY and exotics searches

- Exotic signals that require the full simulation



Lukas Heinrich



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

The ATLAS Collaboration

The reinterpretation of a search for dark matter produced in association with a Higgs boson decaying to b -quarks performed with RECAST, a software framework designed to facilitate the reinterpretation of existing searches for new physics, is presented. Reinterpretation using RECAST is enabled through the sustainable preservation of the original data analysis as re-executable declarative workflows using modern cloud technologies and integrated with the wider CERN Analysis Preservation efforts. The reinterpretation targets a model predicting dark matter production in association with a hypothetical dark Higgs boson decaying into b -quarks where the mass of the dark Higgs boson m_s is a free parameter, necessitating a faithful reinterpretation of the analysis. The dataset has an integrated luminosity of 79.8 fb^{-1} and was recorded with the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy of $\sqrt{s} = 13 \text{ TeV}$. Constraints on the parameter space of the dark Higgs model for a fixed choice of dark matter mass $m_\chi = 200 \text{ GeV}$ exclude model configurations with a mediator mass up to 3.2 TeV .

ATL-PHYS-PUB-2019-032
12 August 2019



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

A recent ATLAS search for displaced jets in the hadronic calorimeter is preserved in RECAST and thereafter used to constrain three new physics models not studied in the original work. A Stealth SUSY model and a Higgs-portal baryogenesis model, both predicting long-lived particles and therefore displaced decays, are probed for proper decay lengths between a few cm and 500 m. A dark sector model predicting Higgs and heavy boson decays to collimated hadrons via long-lived dark photons is also probed. The cross-section times branching ratio for the Higgs channel is constrained between a few millimetres and a few metres, while for a heavier 800 GeV boson the constraints extend from tenths of a millimetre to a few tens of metres. The original data analysis workflow was completely captured using virtualisation techniques, allowing for an accurate and efficient reinterpretation of the published result in terms of new signal models following the RECAST protocol.

ATL-PHYS-PUB-2020-007
28/03/2020



TRAINING

Encouraging response by the community



Instructors Danika MacDonnel and Giordon Stark working with participants. Photo Credit: Samuel Meehan.



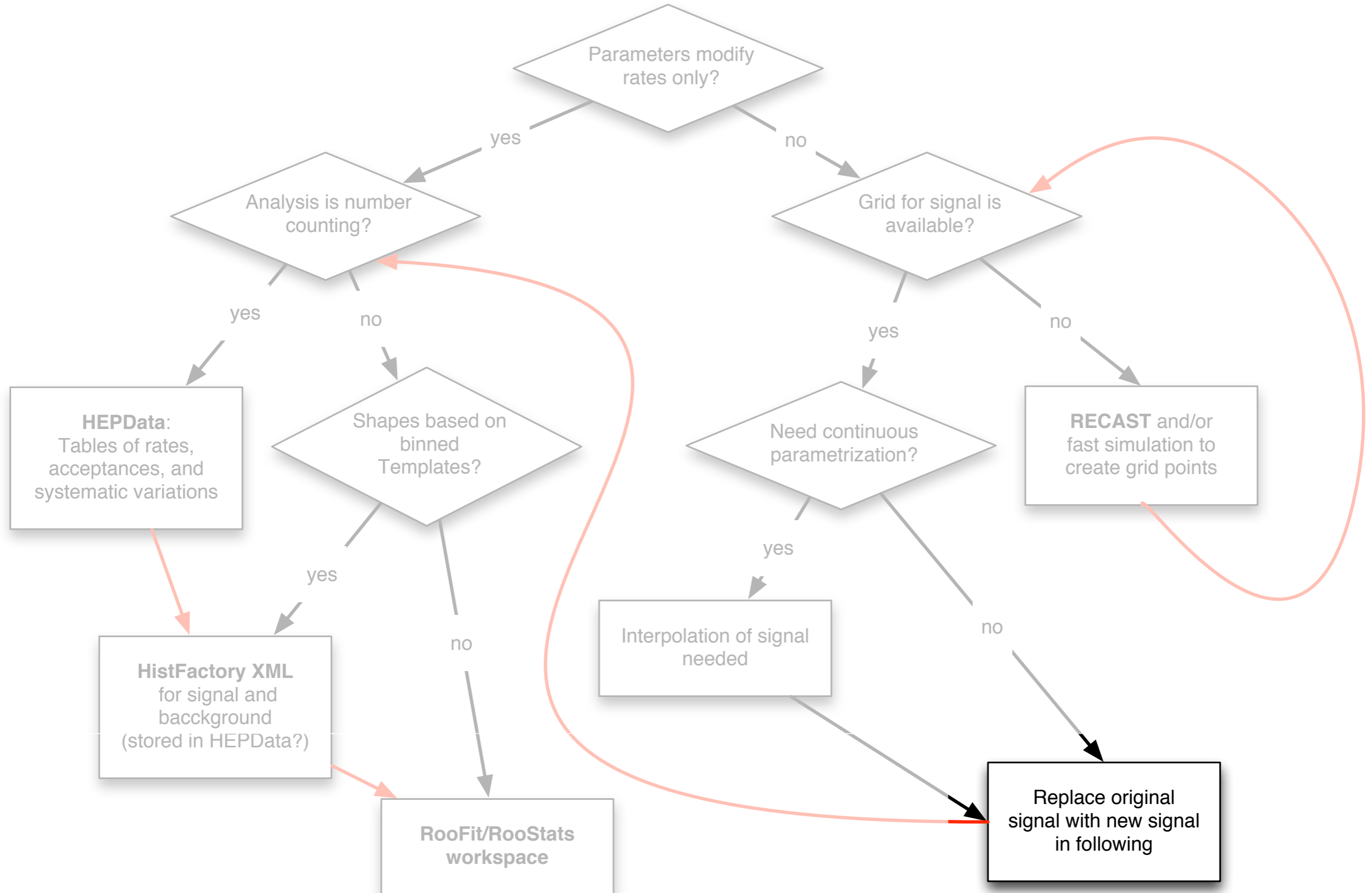
Participants in [Analysis Preservation Bootcamp](#) showing off their ability to reproduce an LHC analysis. Photo Credit: Samuel Meehan



Leonora Vesterbacka

Likelihood Publishing + RECAST =







Lukas Heinrich
CERN



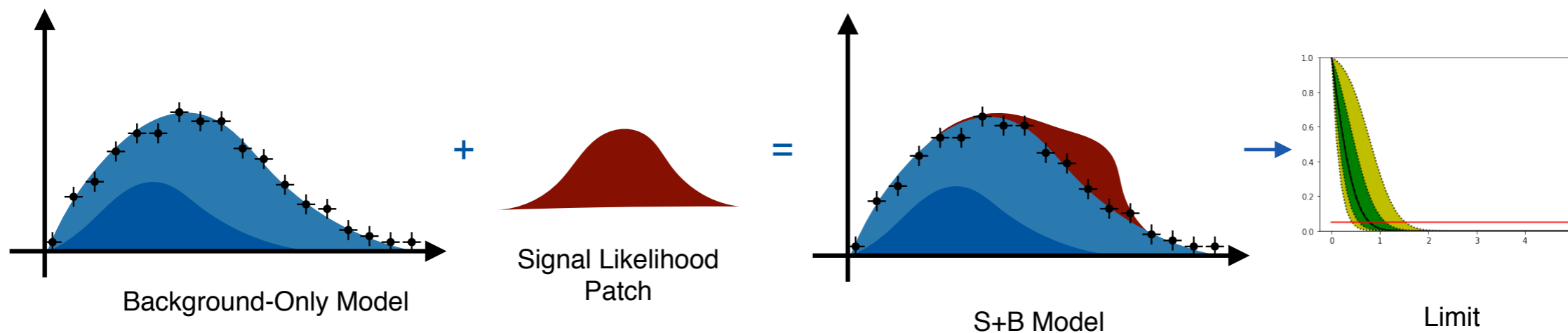
Matthew Feickert
University of Illinois
Urbana-Champaign

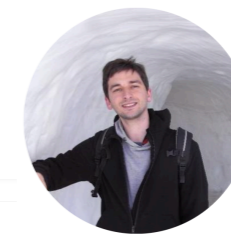


Giordon Stark
UCSC SCIPP

Can think of reinterpretation as a "patching operation"

- take base likelihood model
- **(remove signal if necessary)**
- **add new signal**
- run inference





Lukas Heinrich
CERN



Matthew Feickert
University of Illinois
Urbana-Champaign



Giordon Stark
UCSC SCIPP

Reinterpretation with Likelihood Preservation

Idea: upload to hepdata:

- base likelihood JSON (no signal)
- original signal patches

Reinterpreters can generate new patches (according to a new signal) and combine with the published base likelihood

Additional Publication Resources

filter

Common Resources 4

- Missing Transverse Energy 2
- Effective Mass 2
- Object Based Missing Transverse Energy significance 2
- MaxMin alternative algorithm average m_{hcand} 2
- Leading jet p_T 2
- MaxMin algorithm m_{hcand} 2
- Efficiency_SRA_M_m60 2
- Acceptance_SRC_28 2
- Acceptance_SRC_26 2
- Acceptance_SRC_24 2
- Acceptance_SRA_M_dm130 2
- Acceptance_SRB 2
- Acceptance_SRA_L_dm130 2
- Acceptance_SRC_incl 2

</>

External Link

Web page with auxiliary material

View Resource

📄

C++ File

Truth code to compute acceptance for all signal regions using the SimpleAnalysis framework

Download

📄

gz File

Archive of full likelihoods in the HistFactory JSON format described in ATL-PHYS-PUB-2019-029. Provided are 3 statistical models labeled RegionA, RegionB and RegionC respectively each in their own sub-directory. For each model the background-only model is found in the file named 'BkgOnly.json'. For each model a set of patches for various signal points is provided.

Download

📄

gz File

slha files for the 3 baseline signal points used in the analysis for regions A,B,C

Download

```

├─ README.md
├─ RegionA
│   └─ BkgOnly.json
│   └─ patch.sbottom_1000_131_1.json
│   └─ patch.sbottom_1000_205_60.json
│   └─ patch.sbottom_1000_230_100.json
│   └─ patch.sbottom_1000_250_60.json
│   └─ patch.sbottom_1000_330_200.json
│   └─ patch.sbottom_1000_350_60.json
│   └─ patch.sbottom_1000_430_300.json
└─
                    
```

USAGE IN PHONO RECASTING TOOLS

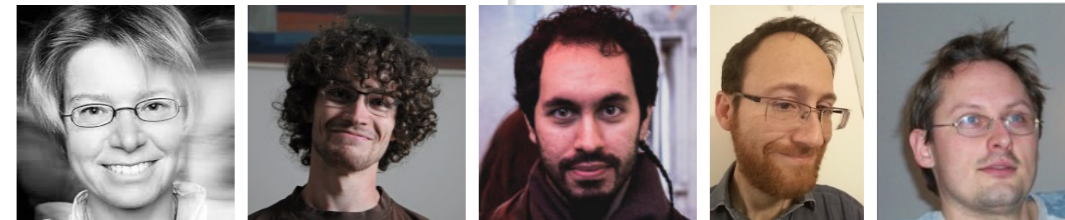
G. Alguero, J. Araz, B. Fuks, SK, W. Waltenberger
Contr. 15, LH 2019 BSM WG report, arXiv:2002.12220

Usage in MadAnalysis 5

MA5-pyhf interface established within the LH PhysTeV 2019 workshop

- the relevant JSON files must be located in the same analysis folder as the recast code (done automatically at the time of the PAD installation).
- The `analysis.info` file must include new `<pyhf>` elements specifying the names of the JSON files together with the corresponding channels (ensembles of SRs) and the regions they include, as defined in the JSON files.

```
<pyhf id="RegionA">
  <name>atlas_susy_2018_031_SRA.json</name>
  <regions>
    <channel name="SR_meff">SRA_L SRA_M SRA_H</channel>
    <channel name="VRtt_meff"> </channel>
    <channel name="CRtt_meff"> </channel>
  </regions>
</pyhf>
```



S. Kraml - Feedback on use of public likelihoods - 24 Sep 2020

SModelS-pyhf interface

Gaël Alguero, SK, Wolfgang Waltenberger,
arXiv:2009.01809

- Available from SModelS v1.2.4 onward (**released Sep. 3rd, 2020**)
- The interfacing of pyhf to SModelS consists of two parts:
 - addition of an independent module `tools/pyhfInterface.py`
 - changes brought to `experiment/datasetObj.py`
- Can be turned on/off by setting

```
combineSR = True/False
```

in the `parameters.ini` file *)

PyhfData class:

Storing and handling of the information related to the JSON files and input signal predictions.

Collects information in the workspaces such as the number of SRs, and the paths to the SR samples where the BSM predictions are to be written.

The VRs and CRs are assumed not to contribute and removed from the workspaces.

PyhfUpperLimitComputer class:

For inferring the upper limits given the PyhfData information

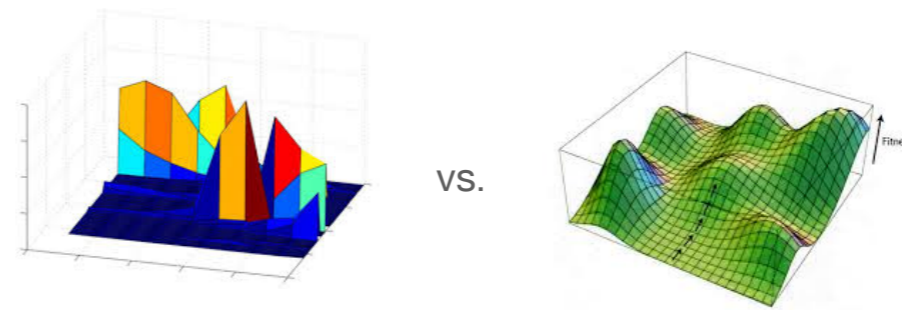
*) The same flag also turns on the SR combination in the simplified likelihood approach for CMS efficiency map results, for which a covariance matrix is available.



Going further

Besides allowing us to better reproduce the official limits of each analysis, the full likelihoods

- will **greatly improve global fits**
- offer interesting possibilities to **explore cross-analysis correlations**
 - Systematic naming of nuisances?
- Both is also very useful for projects like the **Protomodel Builder**
(cf talk by W. Waltenberger on June 4)
- Differentiability will allow for gradient-based methods in the future
- Lots to do on the pheno side, we are not yet using the full potential of full likelihoods.



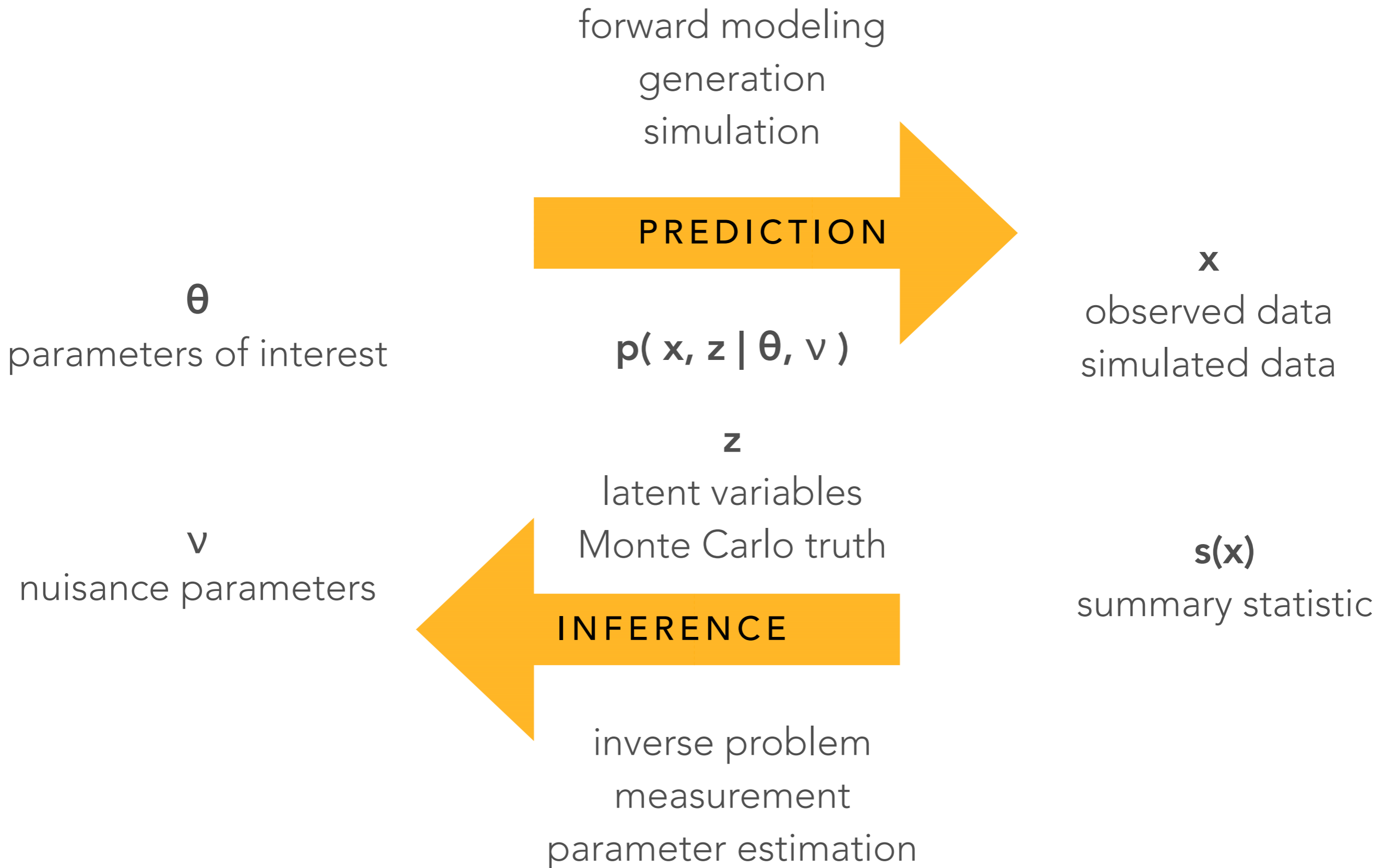
Part 3

A close-up shot of Morpheus from the movie The Matrix. He is bald, has a serious expression, and is wearing his signature black sunglasses. The background is a blurred outdoor setting.

WHAT IF I TOLD YOU

YOU CAN'T EVALUATE THE LIKELIHOOD

STATISTICAL FRAMING



PARTICLE PHYSICS

Parameters
of interest

Theory
parameters

θ



Evolution

PARTICLE PHYSICS

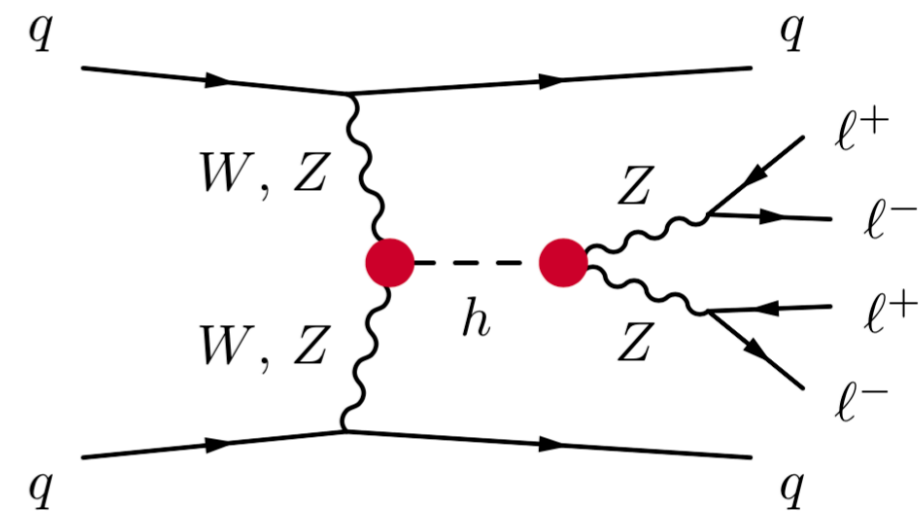
Latent variables

Parameters of interest

Parton-level momenta

Theory parameters

$$z_p \longleftarrow \theta$$



PARTICLE PHYSICS

Latent variables

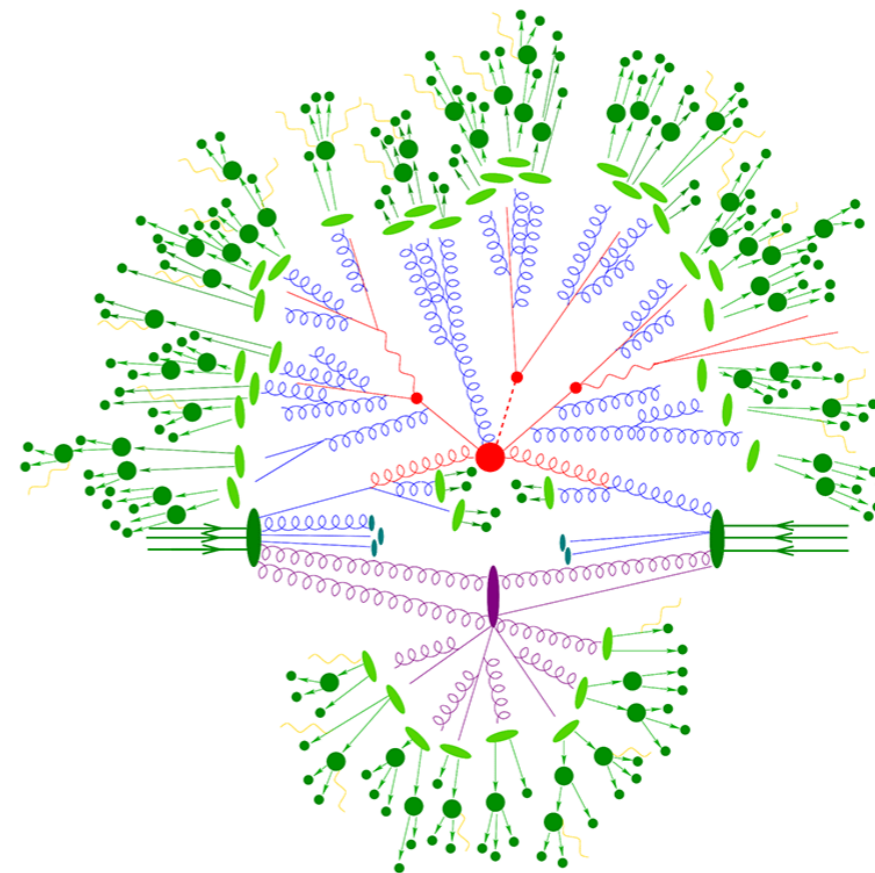
Parameters
of interest

Shower
splittings

Parton-level
momenta

Theory
parameters

z_s ← z_p ← θ

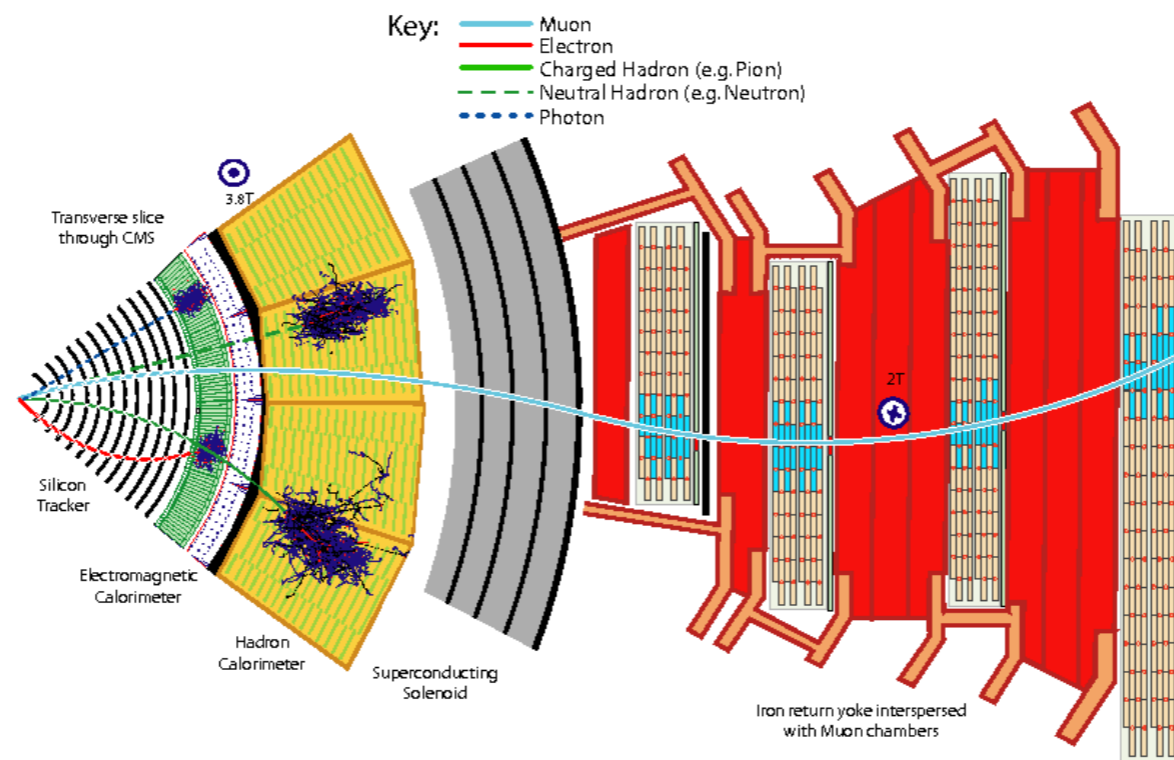
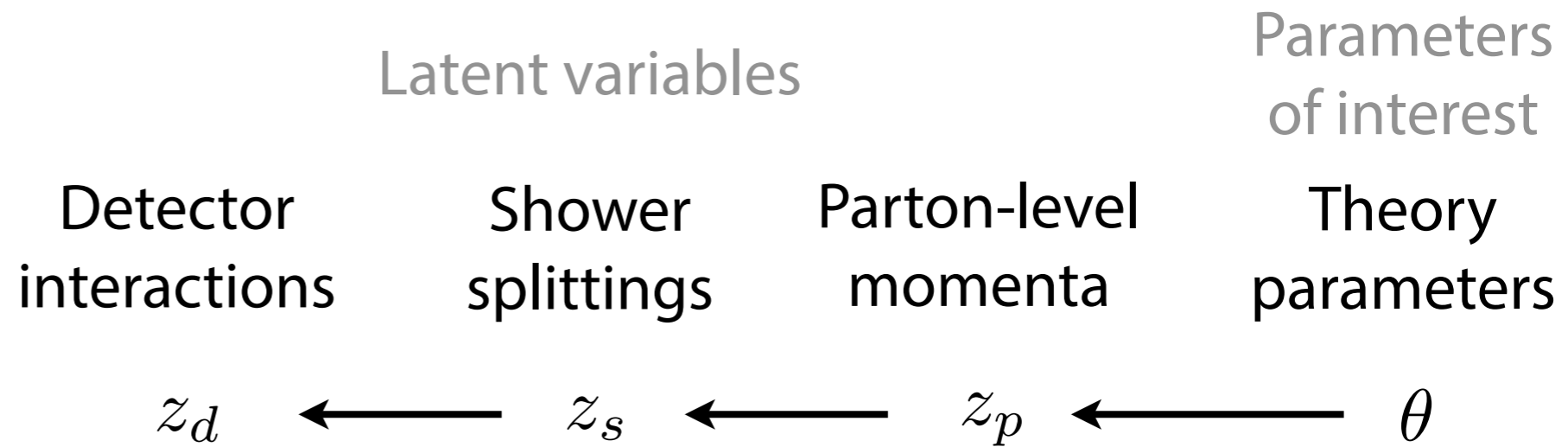


[Source: F. Krauss]



Evolution

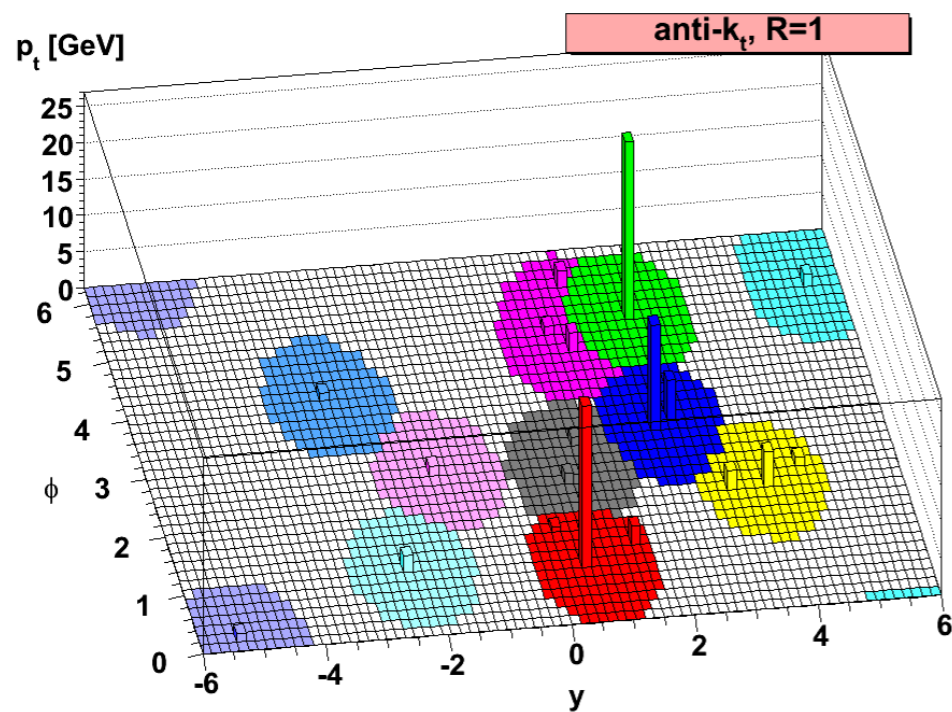
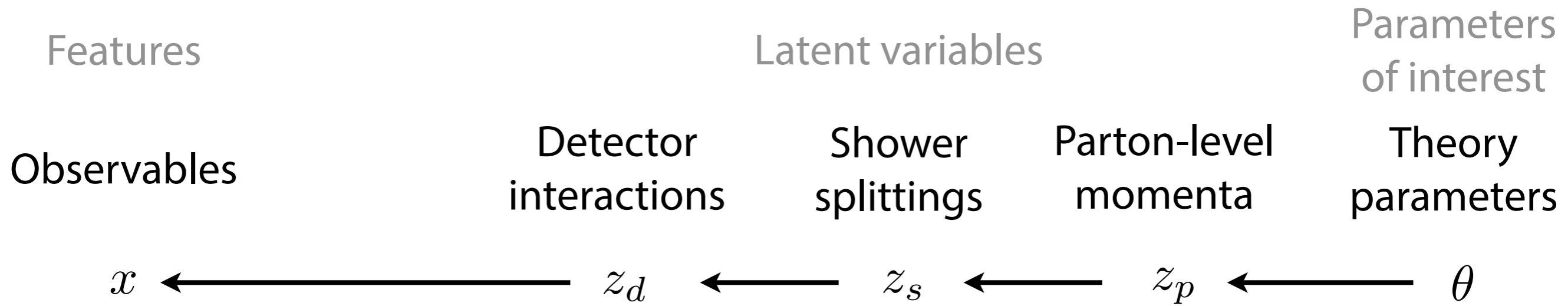
PARTICLE PHYSICS



[Source: CMS]



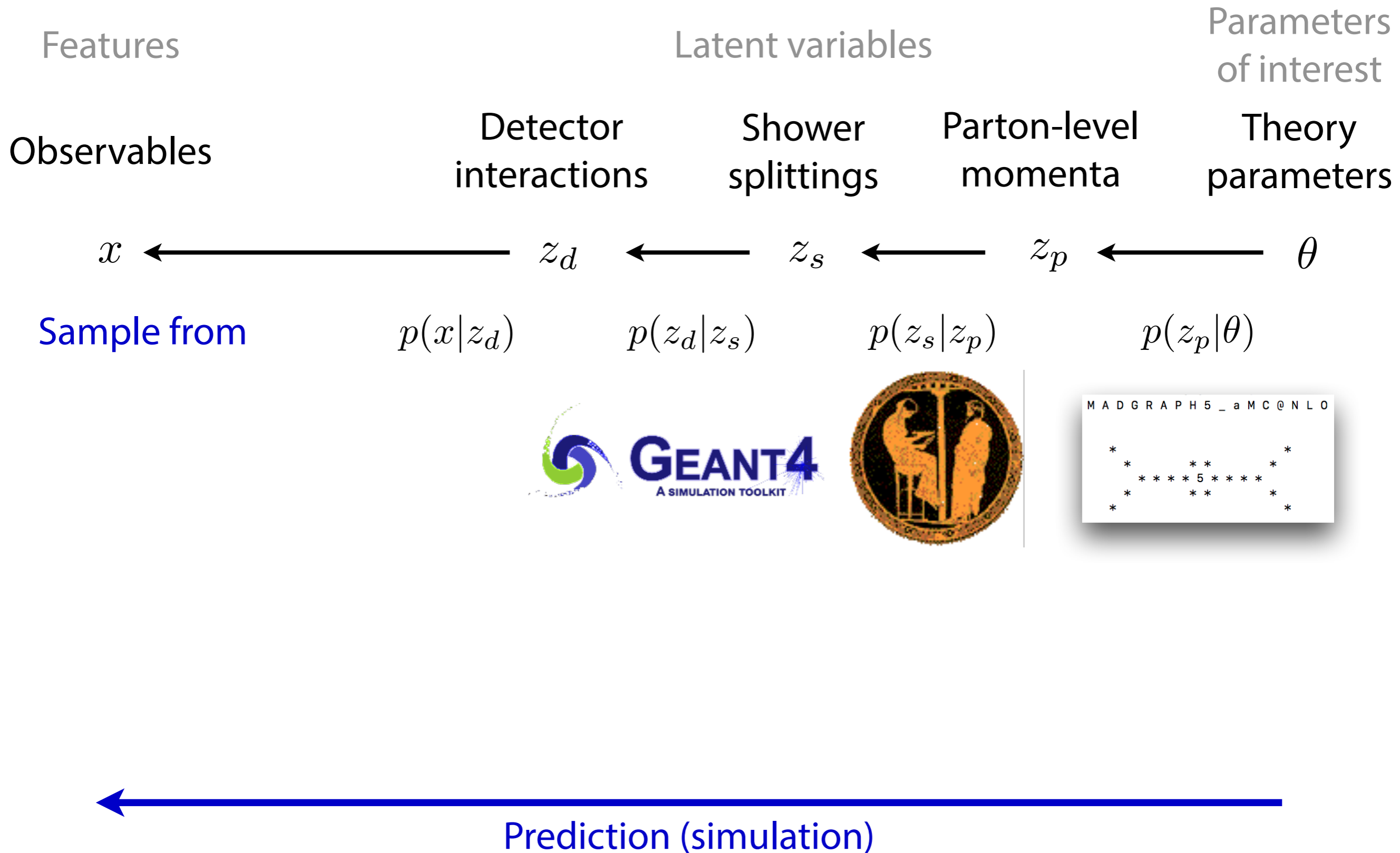
PARTICLE PHYSICS



[Source: M. Cacciari,
G. Salam, G. Soyez 0802.1189]

← Evolution

PARTICLE PHYSICS



PARTICLE PHYSICS

Features

Latent variables

Parameters
of interest

Observables

Detector
interactions

Shower
splittings

Parton-level
momenta

Theory
parameters

$x \longleftarrow z_d \longleftarrow z_s \longleftarrow z_p \longleftarrow \theta$

$$p(x|\theta) = \int dz_d \int dz_s \int dz_p p(x|z_d)$$

$$p(z_d|z_s)$$

$$p(z_s|z_p)$$

$$p(z_p|\theta)$$



Inference

PARTICLE PHYSICS

Features

Latent variables

Parameters of interest

Observables

Detector interactions

Shower splittings

Parton-level momenta

Theory parameters



$$p(x|\theta) = \int dz_d \int dz_s \int dz_p p(x|z_d)$$

Infeasible to calculate the integral over this enormous space

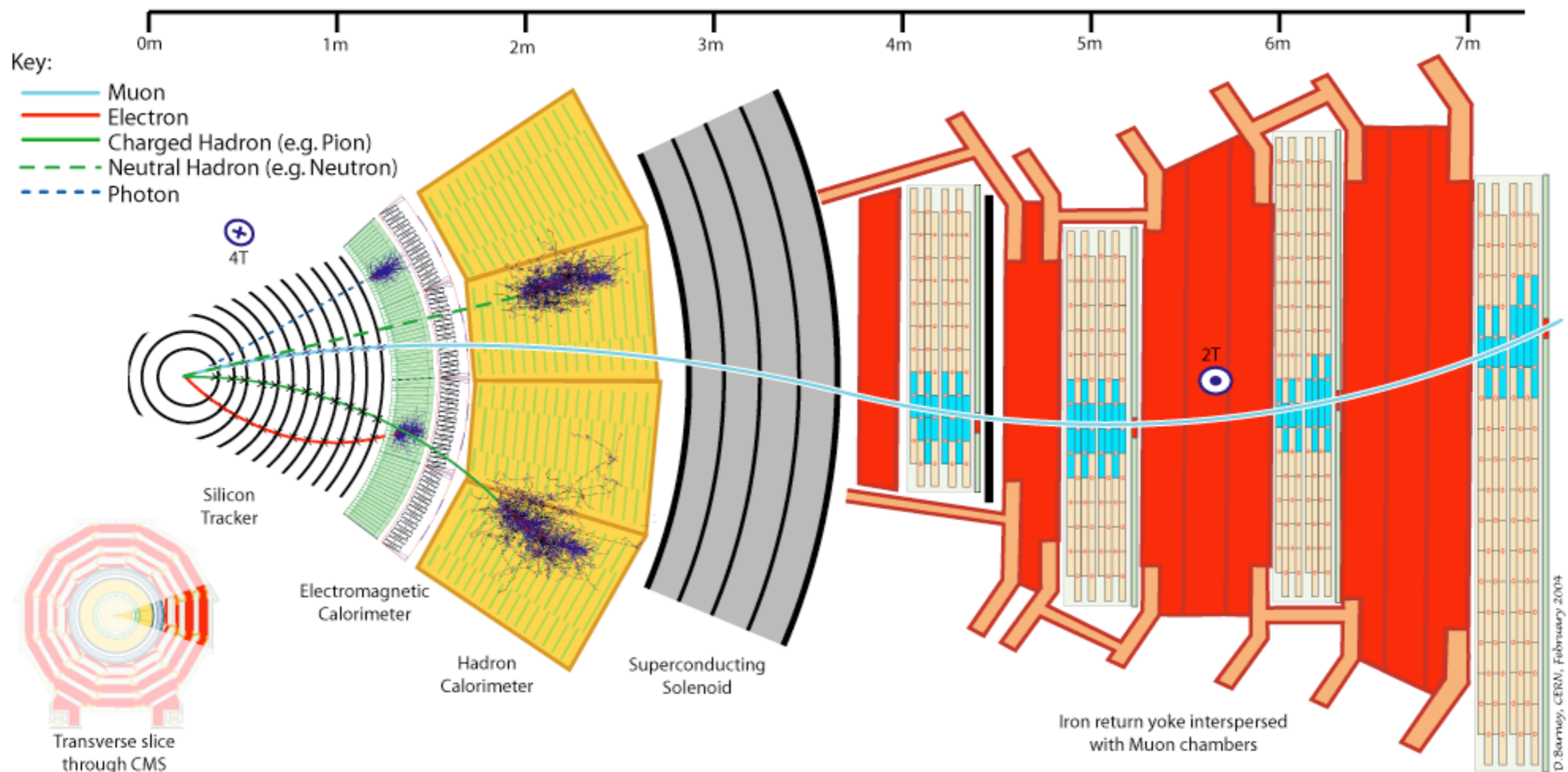
Inference

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

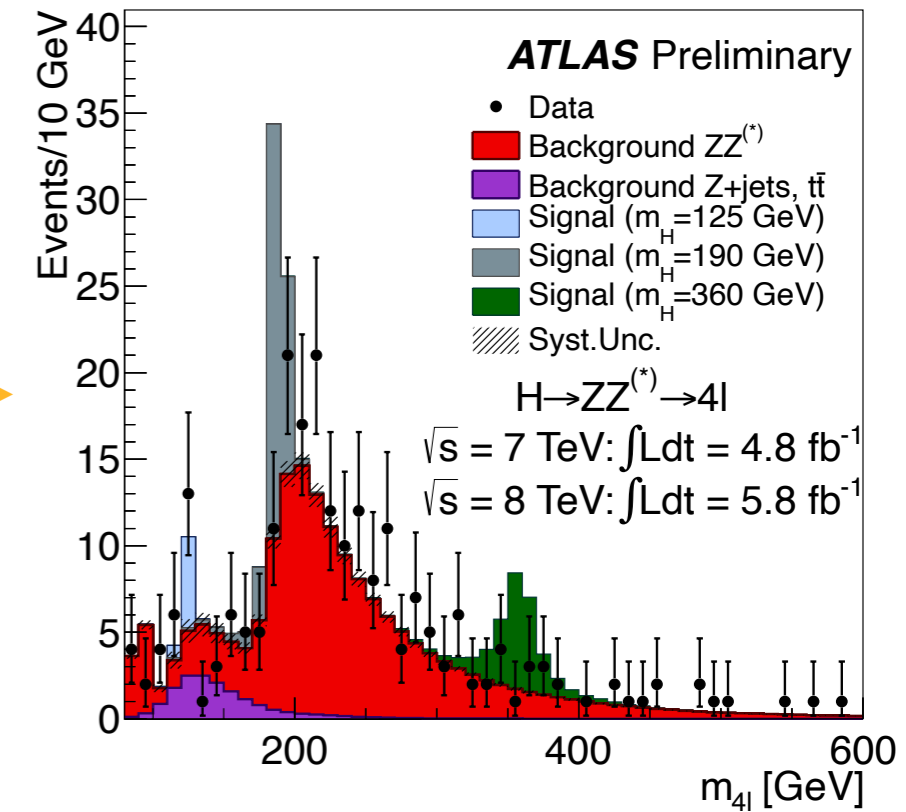
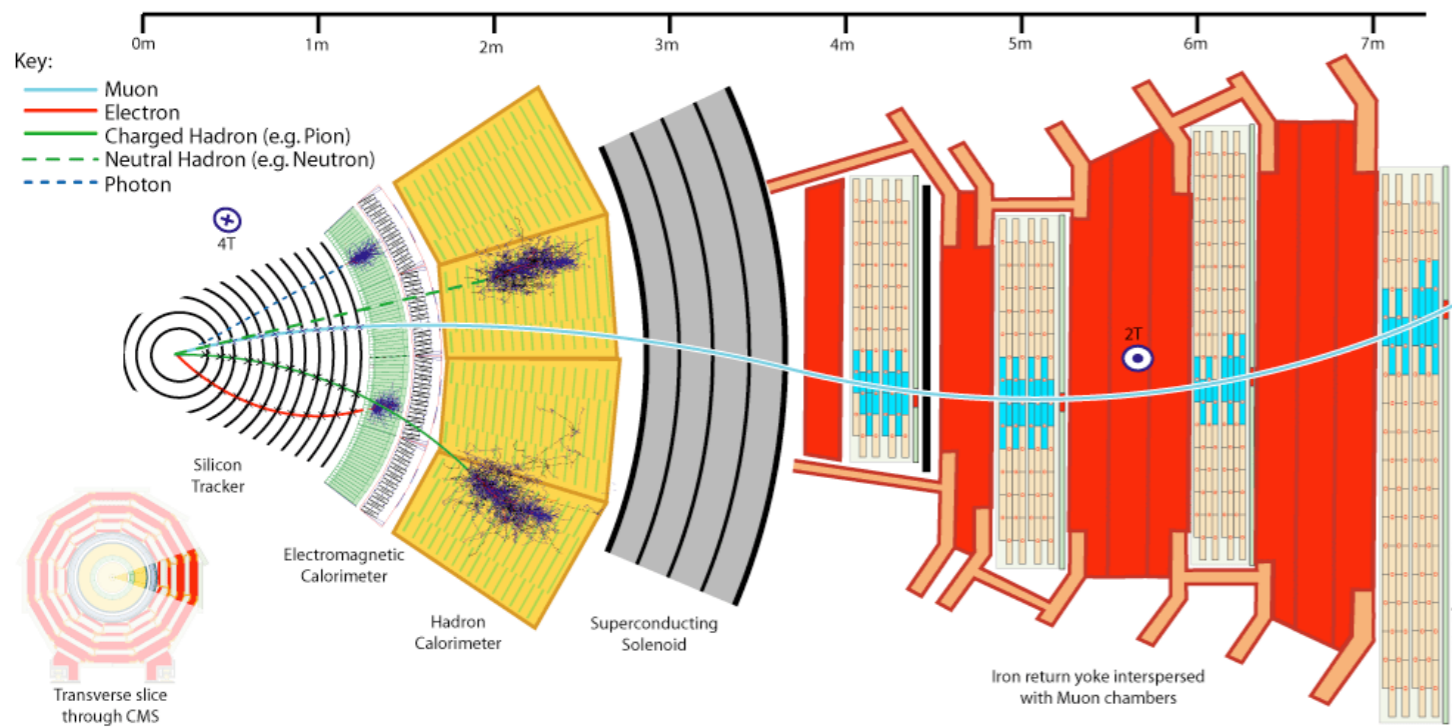
Consequence: evaluation of the likelihood is intractable

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

10^8 SENSORS \rightarrow 1 REAL-VALUED QUANTITY

Most measurements and searches for new particles at the LHC are based on the distribution of a single variable / observable / feature / summary statistic $\mathbf{s}(\mathbf{x})$

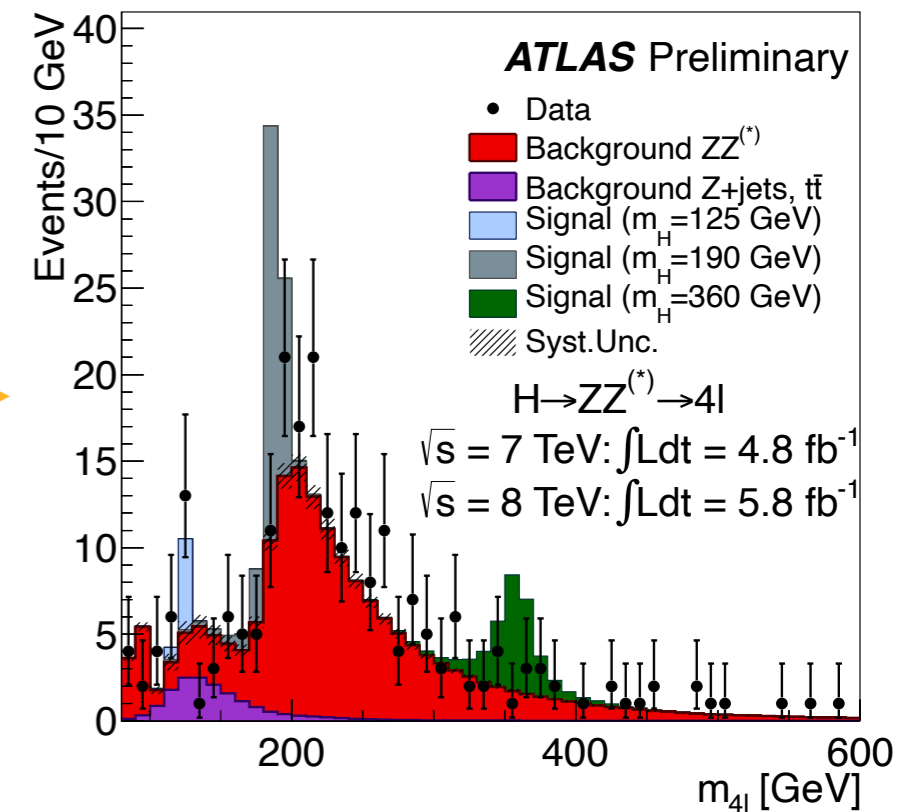
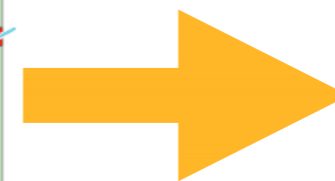
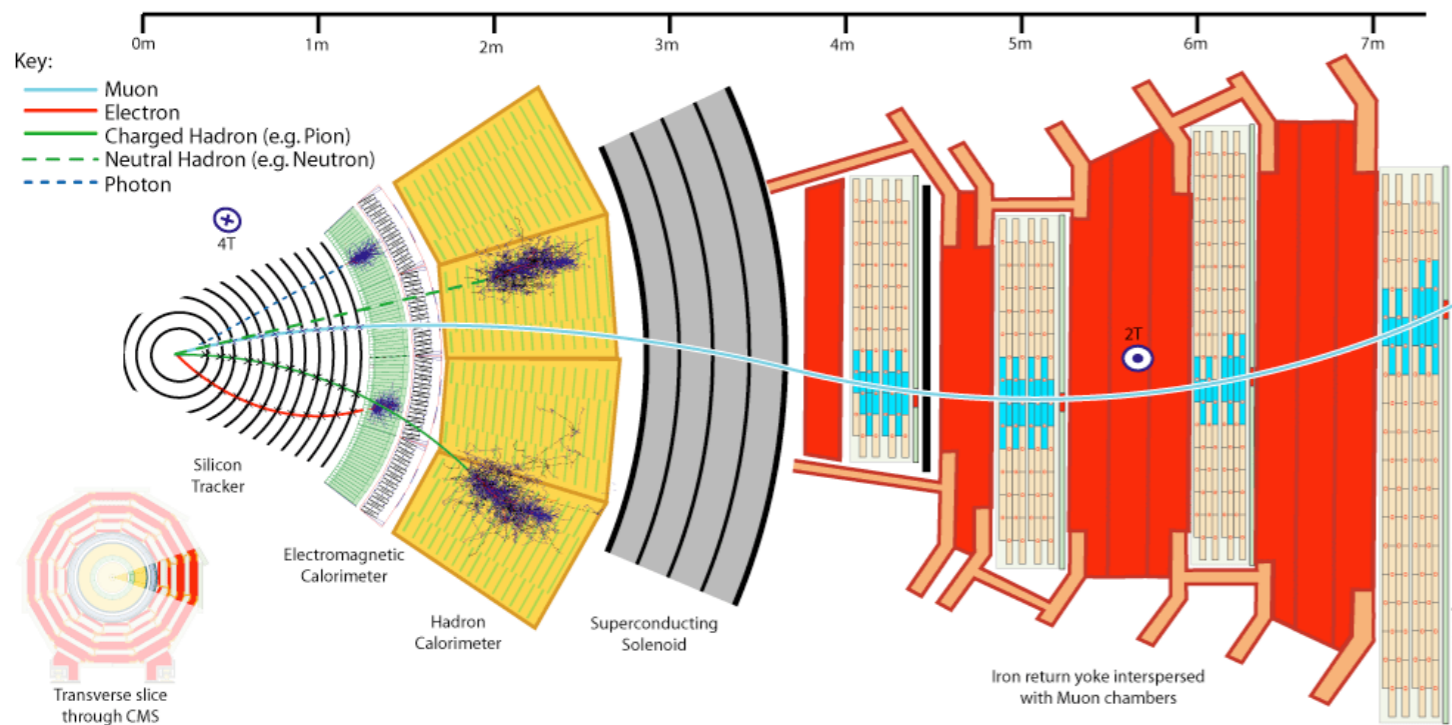
- designing a good observable / summary statistic $\mathbf{s}(\mathbf{x})$ is a task for a skilled physicist and tailored to the goal of measurement or new particle search
- likelihood $p(\mathbf{s}|\theta)$ **approximated** using histograms (univariate density estimation)



10^8 SENSORS \rightarrow 1 REAL-VALUED QUANTITY

Most measurements and searches for new particles at the LHC are based on the distribution of a single variable / observable / feature / summary statistic $\mathbf{s}(\mathbf{x})$

- designing a good observable / summary statistic $\mathbf{s}(\mathbf{x})$ is a task for a skilled physicist and tailored to the goal of measurement or new particle search
- likelihood $p(\mathbf{s}|\theta)$ **approximated** using histograms (univariate density estimation)



This doesn't scale if \mathbf{s} is high dimensional!

THE CRUX, AN INTRACTABLE INTEGRAL

Monte Carlo
Sampling

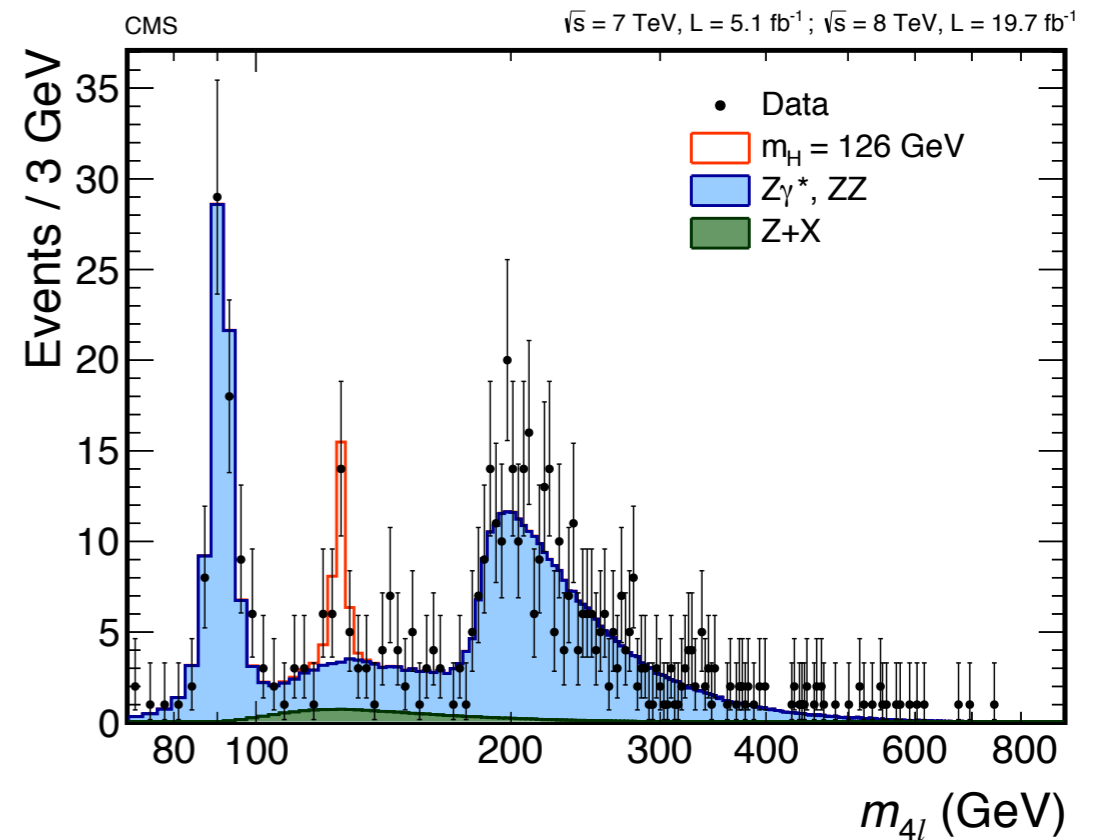
observed

what happened
inside simulation

$$p(s | \theta) = \int dz dx p(s(x), z | \theta)$$

$\hat{p}(s | \theta)$

↑
histogram
approximation



A COMMON THEME, A COMMON LANGUAGE

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.

Markov chain Monte Carlo without likelihoods

Paul Marjoram*, John Molitor*, Vincent Plagnol†, and Simon Tavaré†‡

*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and †Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Communicated by Michael S. Waterman, University of Southern California, Los Angeles, CA, October 24, 2003 (received for review June 20, 2003)

Many stochastic simulation approaches for generating observations from a posterior distribution depend on knowing a likelihood function. However, for many complex probability models, such likelihoods are either impossible or computationally prohibitive to obtain. Here we present a Markov chain Monte Carlo method for generating observations from a posterior distribution without the use of likelihoods. It can also be used in frequentist applications, in particular for maximum-likelihood estimation. The approach is illustrated by an example of ancestral inference in population genetics. A number of open problems are highlighted in the discussion.

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data \mathcal{D} generated from a model \mathcal{M} determined by parameters θ , the prior density of which is denoted by $\pi(\theta)$. We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is $f(\theta|\mathcal{D})$, which is given by

$$f(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)/\mathbb{P}(\mathcal{D}), \quad [1]$$

where $\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)d\theta$ is the normalizing constant.

In most scientific contexts, explicit formulae for such posterior densities are few and far between, and we usually resort to stochastic simulation to generate observations from f . Perhaps the simplest approach for this is the rejection method:

- A1. Generate θ from $\pi(\cdot)$.
- A2. Accept θ with probability $h = \mathbb{P}(\mathcal{D}|\theta)$; return to A1.

of ε therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given ρ and ε , we are generating independent and identically distributed observations from $f(\theta|\rho(\mathcal{D}, \mathcal{D}') \leq \varepsilon)$.

When \mathcal{D} is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of \mathcal{D}' with \mathcal{D} can be made by using lower-dimensional summaries of the data. The motivation for this approach is that if the set of statistics $S = (S_1, \dots, S_p)$ is sufficient for θ , in that $\mathbb{P}(\mathcal{D}|S, \theta)$ is independent of θ , then $f(\theta|\mathcal{D}) = f(\theta|S)$. The normalizing constant $\mathbb{P}(S)$ is typically larger than $\mathbb{P}(\mathcal{D})$, resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about θ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data \mathcal{D} summarized by S :

- D1. Generate θ from $\pi(\cdot)$.
- D2. Simulate \mathcal{D}' from stochastic model \mathcal{M} with parameter θ , and compute the corresponding statistics S' .
- D3. Calculate the distance $\rho(S, S')$ between S and S' .
- D4. Accept θ if $\rho \leq \varepsilon$, and return to D1.

There are several advantages to these rejection methods, among them the fact that they are usually easy to code, they generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-

Markov chain Monte Carlo without likelihoods

Paul Marjoram*, John Molitor*, Vincent Plagnol†, and Simon Tavaré†‡

*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and †Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Communicated by Michael S. Waterman, University of Southern California, Los Angeles, CA, October 24, 2003 (received for review June 20, 2003)

Many stochastic simulation approaches for generating observations from a posterior distribution depend on knowing a likelihood function. However, for many complex probability models, such likelihoods are either impossible or computationally prohibitive to obtain. Here we present a Markov chain Monte Carlo method for generating observations from a posterior distribution without the use of likelihoods. It can also be used in frequentist applications, in particular for maximum-likelihood estimation. The approach is illustrated by an example of ancestral inference in population genetics. A number of open problems are highlighted in the discussion.

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data \mathcal{D} generated from a model \mathcal{M} determined by parameters θ , the prior density of which is denoted by $\pi(\theta)$. We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is $f(\theta|\mathcal{D})$, which is given by

of ε therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given ρ and ε , we are generating independent and identically distributed observations from $f(\theta|\rho(\mathcal{D}, \mathcal{D}') \leq \varepsilon)$.

When \mathcal{D} is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of \mathcal{D}' with \mathcal{D} can be made by using lower-dimensional summaries of the data. The motivation for this approach is that if the set of statistics $S = (S_1, \dots, S_p)$ is sufficient for θ , in that $\mathbb{P}(\mathcal{D}|S, \theta)$ is independent of θ , then $f(\theta|\mathcal{D}) = f(\theta|S)$. The normalizing constant $\mathbb{P}(S)$ is typically larger than $\mathbb{P}(\mathcal{D})$, resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about θ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data \mathcal{D} summarized by S :

When \mathcal{D} is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of \mathcal{D}' with \mathcal{D} can be made by using lower-dimensional summaries of the data. The

- A1. Generate θ from $\pi(\cdot)$.
- A2. Accept θ with probability $h = \mathbb{P}(\mathcal{D}|\theta)$; return to A1.

generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-

Markov chain Monte Carlo without likelihoods

Paul Marjoram*, John Molitor*, Vincent Plagnol†, and Simon Tavaré†‡

*Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, and †Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Communicated by Michael S. Waterman, University of Southern California, Los Angeles, CA, October 24, 2003 (received for review June 20, 2003)

Many stochastic simulation approaches for generating observations from a posterior distribution depend on knowing a likelihood function. However, for many complex probability models, such likelihoods are either impossible or computationally prohibitive to obtain. Here we present a Markov chain Monte Carlo method for generating observations from a posterior distribution without the use of likelihoods. It can also be used in frequentist applications, in particular for maximum-likelihood estimation. The approach is illustrated by an example of ancestral inference in population genetics. A number of open problems are highlighted in the discussion.

One of the basic problems in Bayesian statistics is the computation of posterior distributions. We imagine data \mathcal{D} generated from a model \mathcal{M} determined by parameters θ , the prior density of which is denoted by $\pi(\theta)$. We assume unless otherwise stated that the data are discrete. The posterior distribution of interest is $f(\theta|\mathcal{D})$, which is given by

of ε therefore reflects a tension between computability and accuracy. The method is still honest in that, for a given ρ and ε , we are generating independent and identically distributed observations from $f(\theta|\rho(\mathcal{D}, \mathcal{D}') \leq \varepsilon)$.

When \mathcal{D} is high-dimensional or continuous, this approach can be impractical as well, and then the comparison of \mathcal{D}' with \mathcal{D} can be made by using lower-dimensional summaries of the data. The motivation for this approach is that if the set of statistics $S = (S_1, \dots, S_p)$ is sufficient for θ , in that $\mathbb{P}(\mathcal{D}|S, \theta)$ is independent of θ , then $f(\theta|\mathcal{D}) = f(\theta|S)$. The normalizing constant $\mathbb{P}(S)$ is typically larger than $\mathbb{P}(\mathcal{D})$, resulting in more acceptances. In practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach. Thus we seek to use knowledge of the particular problem at hand to suggest summary statistics that capture information about θ . With these statistics in hand, we have the following approximate Bayesian computation scheme for data \mathcal{D} summarized by S :

practice it will be hard, if not impossible, to identify a suitable set of sufficient statistics, and we then might resort to a more heuristic approach.

- A1. Generate θ from $\pi(\cdot)$.
- A2. Accept θ with probability $h = \mathbb{P}(\mathcal{D}|\theta)$; return to A1.

generate independent observations (and thus can use embarrassingly parallel computation), and they readily provide estimates of Bayes factors that can be used for model com-



ImageNet Classification with Deep Convolutional Neural Networks

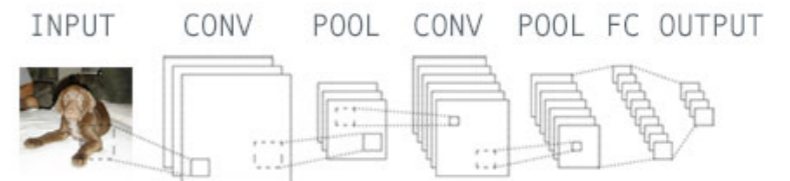
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

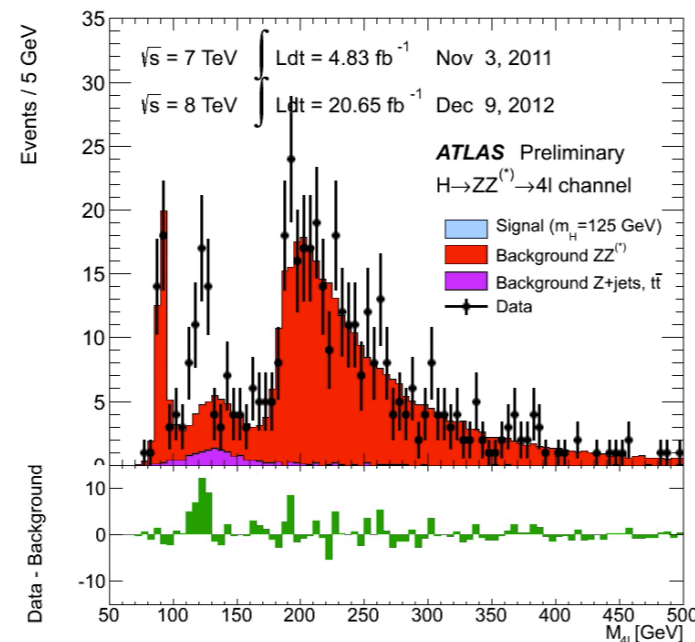
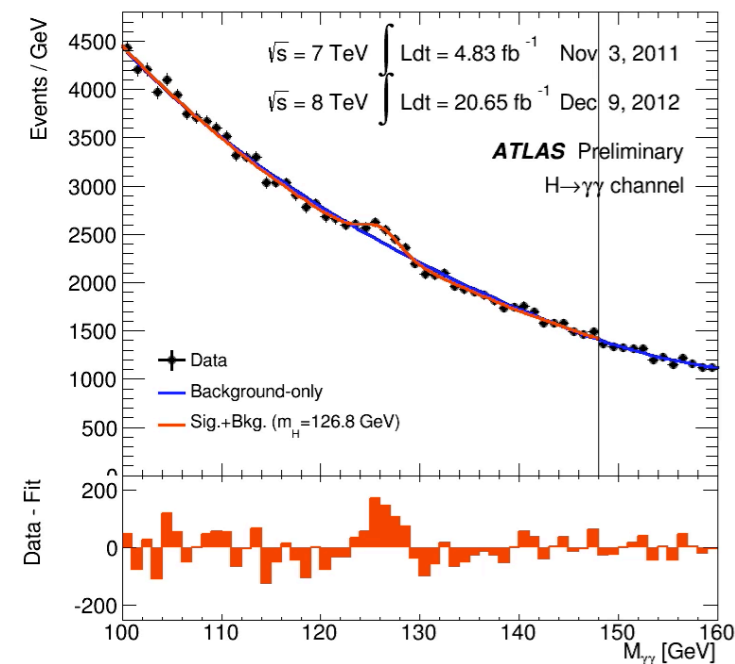
We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.



Dog:	94%
Cat:	31%
Bird:	2%
Boat:	0%



Dog:	37%
Cat:	91%
Bird:	21%
Boat:	1%





ImageNet Classification with Deep Convolutional Neural Networks

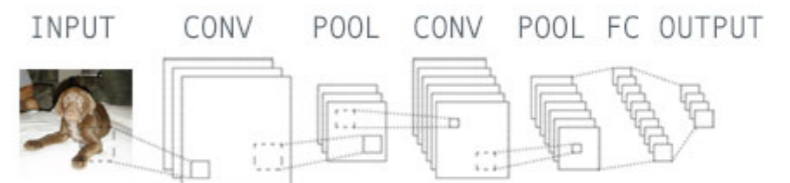
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

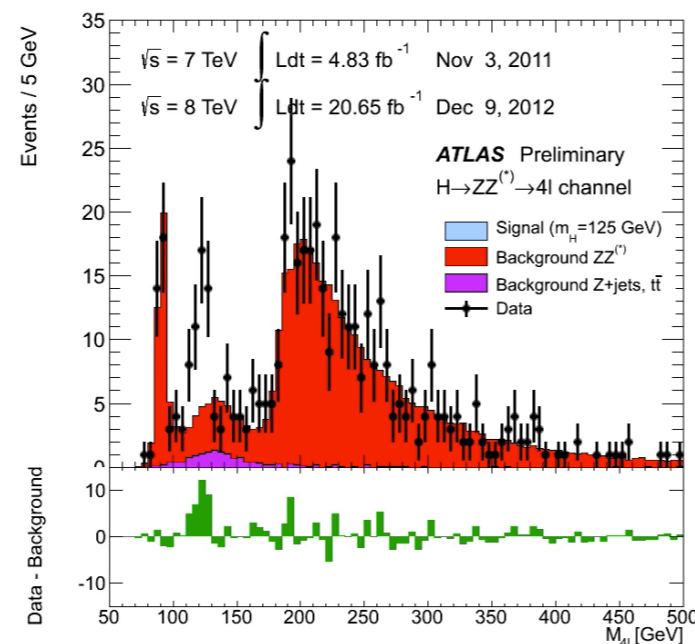
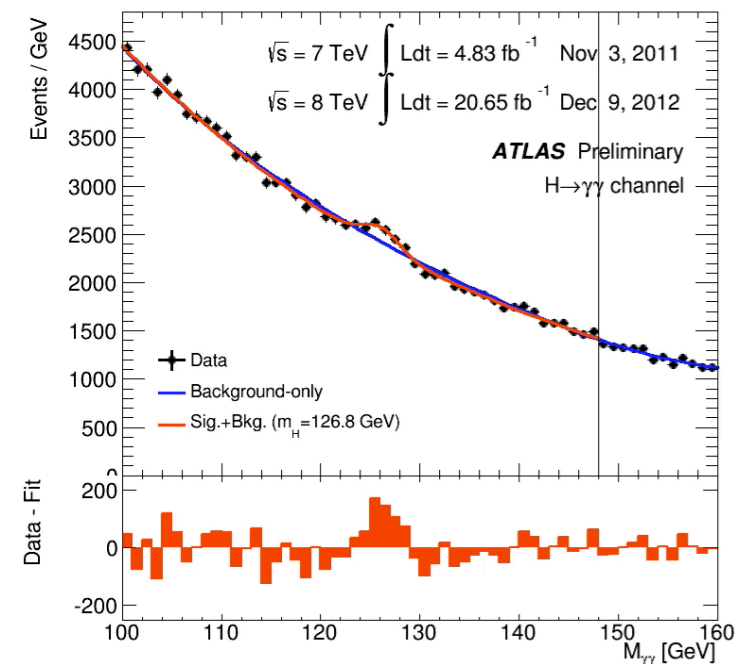
We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.



Dog:	94%
Cat:	31%
Bird:	2%
Boat:	0%



Dog:	37%
Cat:	91%
Bird:	21%
Boat:	1%



ICML 2017 Workshop on Implicit Models

Workshop Aims

Probabilistic models are an important tool in machine learning. They form the basis for models that generate realistic data, uncover hidden structure, and make predictions. Traditionally, probabilistic models in machine learning have focused on prescribed models. Prescribed models specify a joint density over observed and hidden variables that can be easily evaluated. The requirement of a tractable density simplifies their learning but limits their flexibility --- several real world phenomena are better described by simulators that do not admit a tractable density. Probabilistic models defined only via the simulations they produce are called implicit models.

Arguably starting with generative adversarial networks, research on implicit models in machine learning has exploded in recent years. This workshop's aim is to foster a discussion around the recent developments and future directions of implicit models.

Implicit models have many applications. They are used in ecology where models simulate animal populations over time; they are used in phylogeny, where simulations produce hypothetical ancestry trees; they are used in physics to generate particle simulations for high energy processes. Recently, implicit models have been used to improve the state-of-the-art in image and content generation. Part of the workshop's focus is to discuss the commonalities among applications of implicit models.

Of particular interest at this workshop is to unite fields that work on implicit models. For example:

- **Generative adversarial networks** (a NIPS 2016 workshop) are implicit models with an adversarial training scheme.
- Recent advances in **variational inference** (a NIPS 2015 and 2016 workshop) have leveraged implicit models for more accurate approximations.
- **Approximate Bayesian computation** (a NIPS 2015 workshop) focuses on posterior inference for models with implicit likelihoods.
- Learning implicit models is deeply connected to **two sample testing, density ratio and density difference** estimation.

We hope to bring together these different views on implicit models, identifying their core challenges and combining their innovations.

How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos

By CADE METZ and KEITH COLLINS JAN. 2, 2018



This one is computer-generated



This one is also computer-generated

TWO REVIEWS

We have written two reviews:



Gilles Louppe



Johann Brehmer

- One focuses on particle physics, one is abstracted

The frontier of simulation-based inference

Kyle Cranmer^{a,b,1}, Johann Brehmer^{a,b}, and Gilles Louppe^c

^aCenter for Cosmology and Particle Physics, New York University, USA; ^bCenter for Data Science, New York University, USA; ^cMontefiore Institute, University of Liège, Belgium

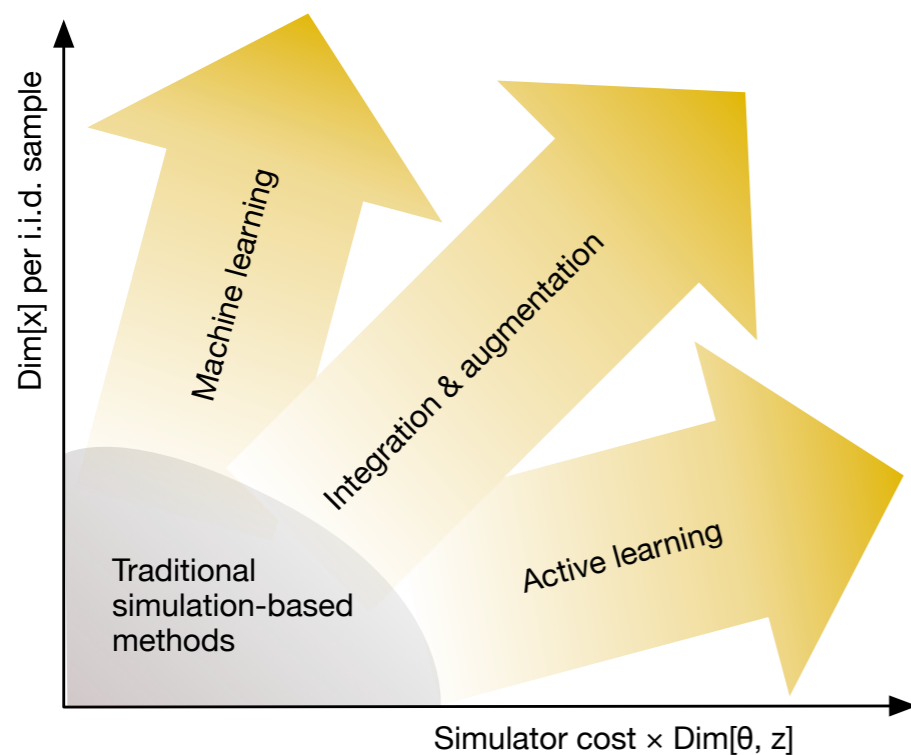
April 3, 2020

Simulation-based inference methods for particle physics

Johann Brehmer and Kyle Cranmer

New York University, New York, NY, 10003

Our predictions for particle physics processes are realized in a chain of complex simulators. They allow us to generate high-fidelity simulated data, but they are not well-suited for inference on the theory parameters with observed data. We explain why the likelihood function of high-dimensional LHC data cannot be explicitly evaluated, why this matters for data analysis, and reframe what the field has traditionally done to circumvent this problem. We then review new simulation-based inference methods that let us directly analyze high-dimensional data by combining machine learning techniques and information from the simulator. Initial studies indicate that these techniques have the potential to substantially improve the precision of LHC measurements. Finally, we discuss probabilistic programming, an emerging paradigm that lets us extend inference to the latent process of the simulator.



Published in Proceedings of the National Academy of Sciences

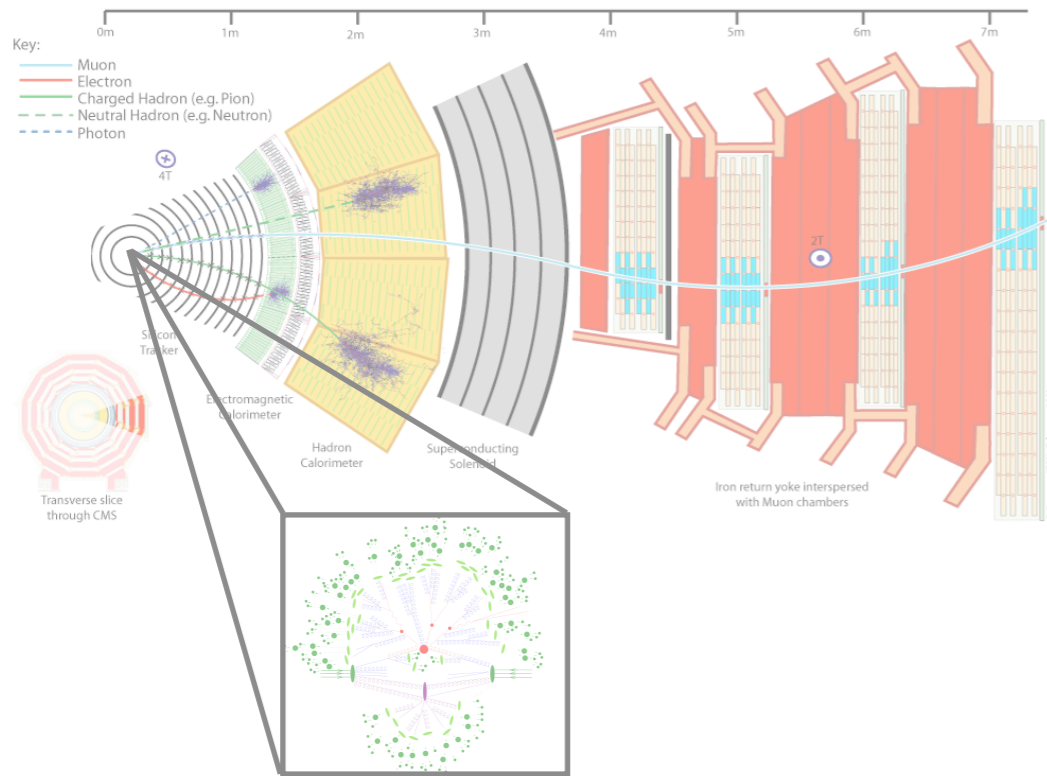
[PNAS (2020), [arXiv:1911.01429](https://arxiv.org/abs/1911.01429)]

This morning!

<https://arxiv.org/abs/2010.06439>

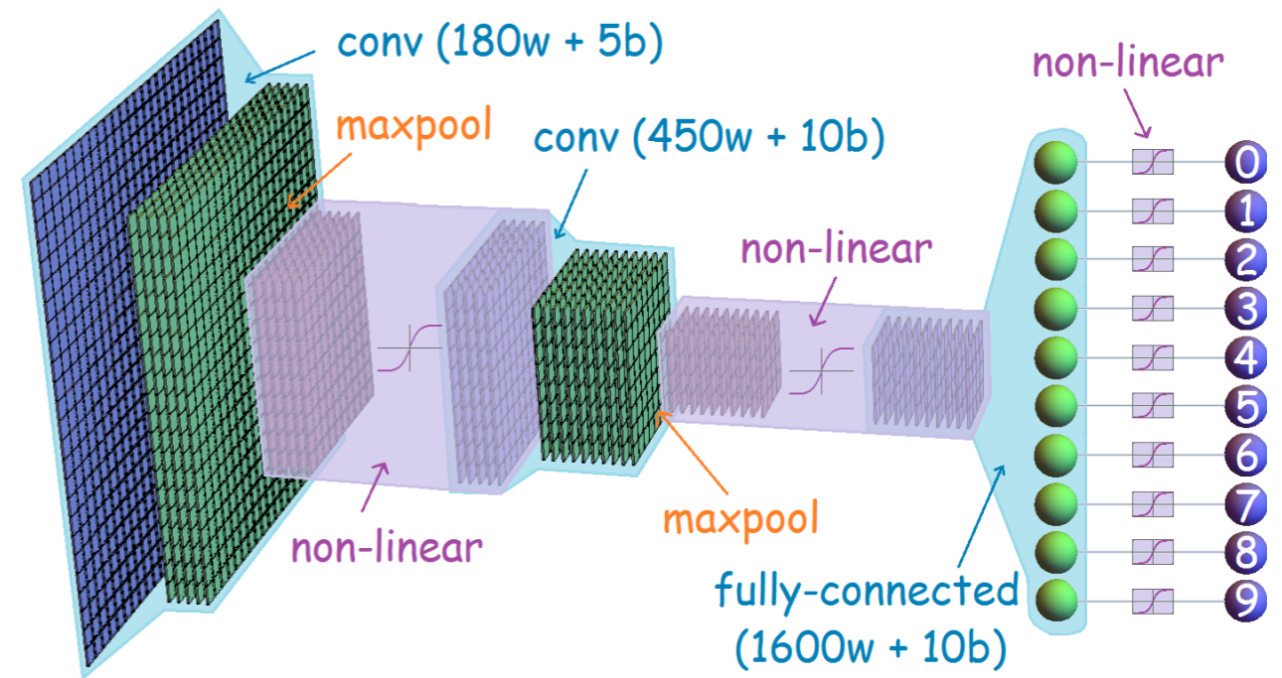
TWO APPROACHES TO SIMULATION-BASED INFERENCE

Use simulator
(much more efficiently)



- Approximate Bayesian Computation (ABC)
- Probabilistic Programming
- Adversarial Variational Optimization (AVO)

Learn simulator
(with deep learning)



- Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAE)
- Likelihood ratio from classifiers (CARL)
- Autogressive models, Normalizing Flows

SIMULATION-BASED INFERENCE

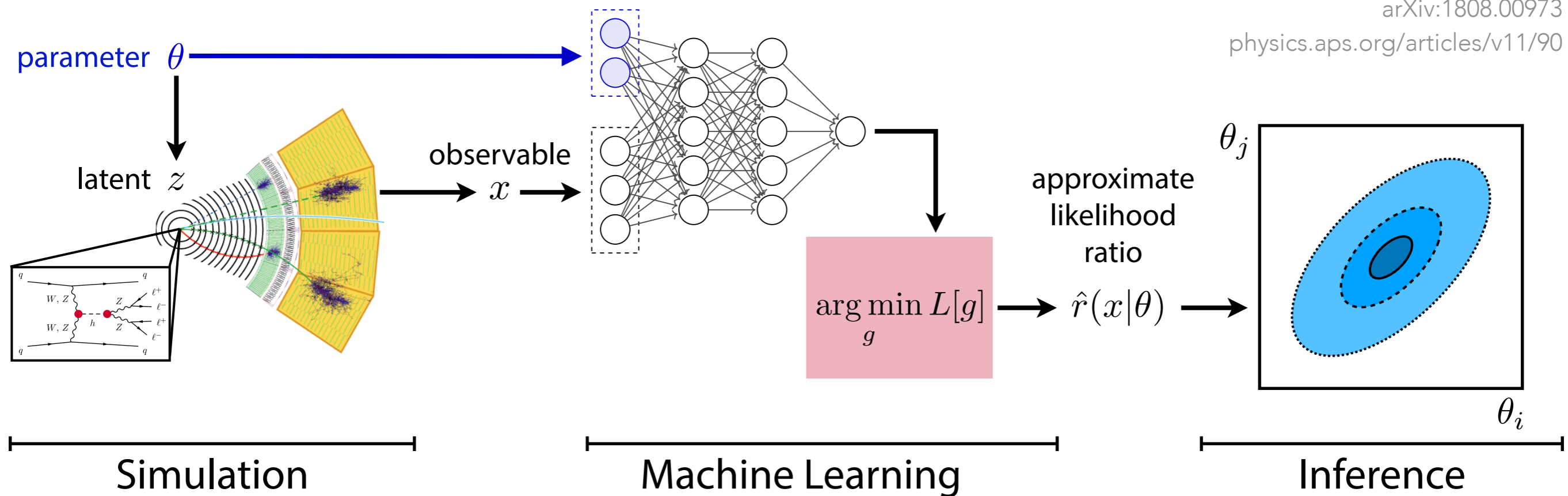
arXiv:1805.12244

PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

arXiv:1808.00973

physics.aps.org/articles/v11/90



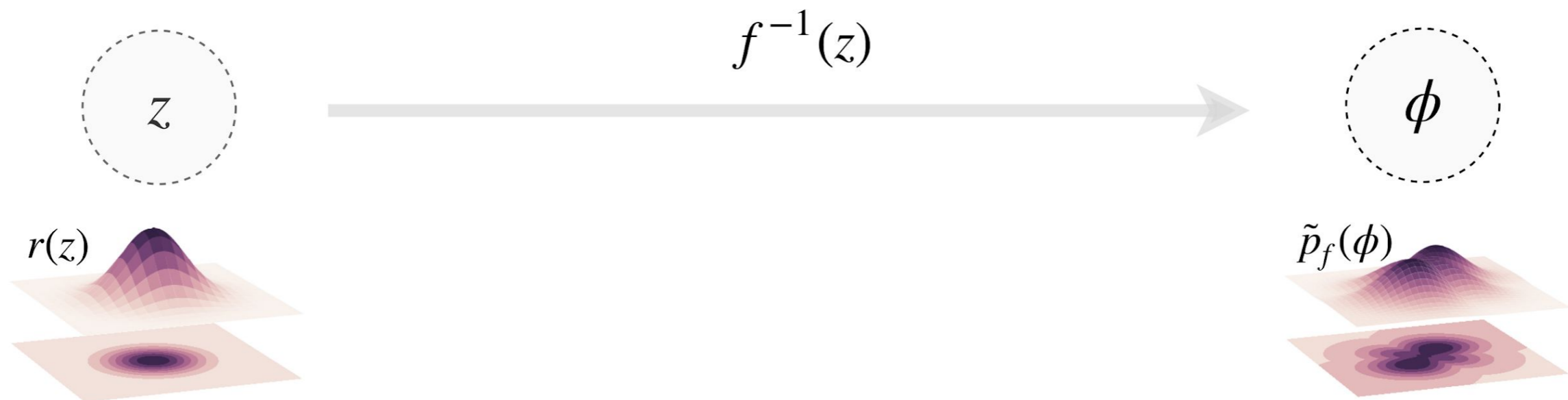
The **surrogate for the likelihood (ratio)** used for inference

A 2-stage process:

1. learning surrogate (**amortized**)
2. Inference on parameters of simulator

Using a change-of-variables, produce a distribution approximating what you want.

[Rezende & Mohamed 1505.05770] + early work by Estaban Tabak, Cristina Turner, Eric Vanden-Eijnden [2010, 2013]

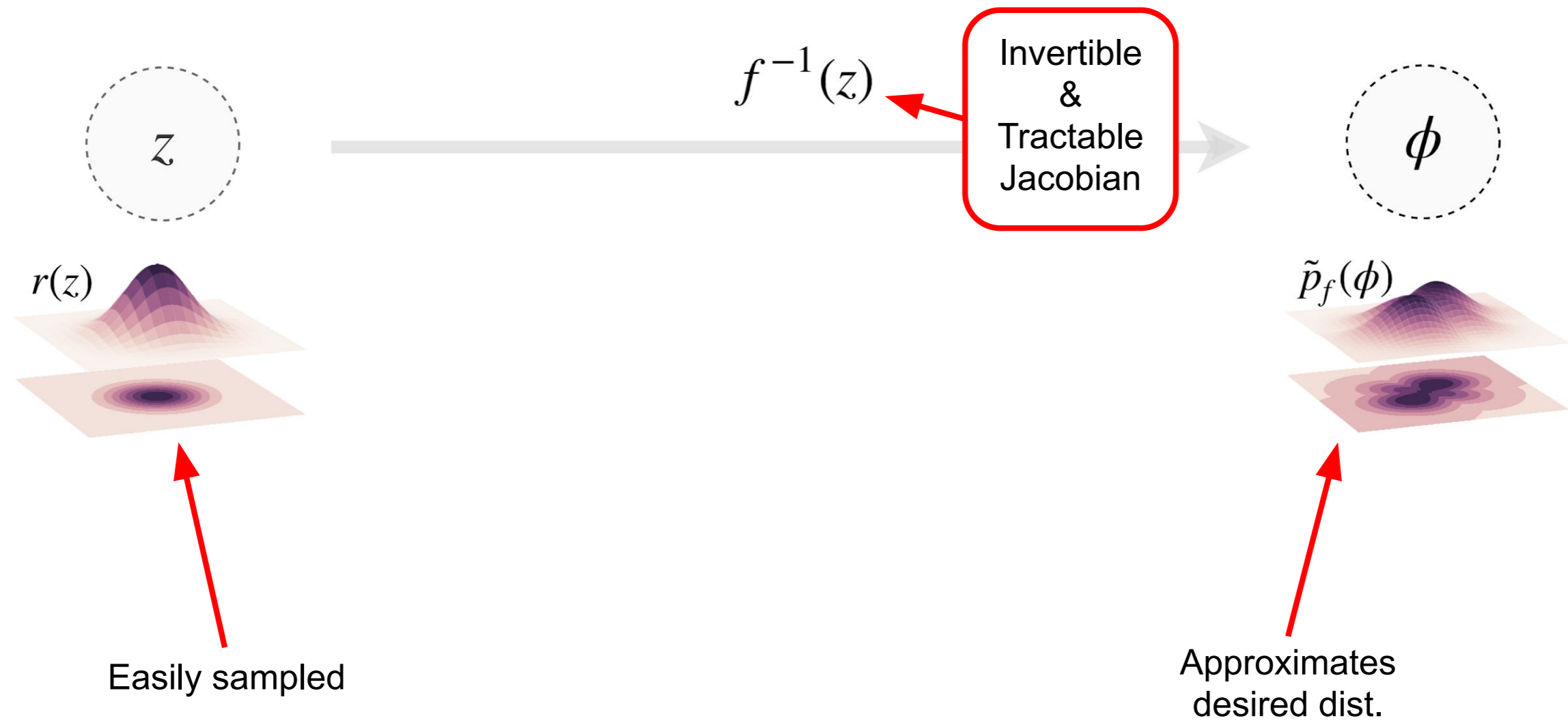


NORMALIZING FLOWS

Using a change-of-variables, produce a distribution approximating what you want.

[Rezende & Mohamed 1505.05770]

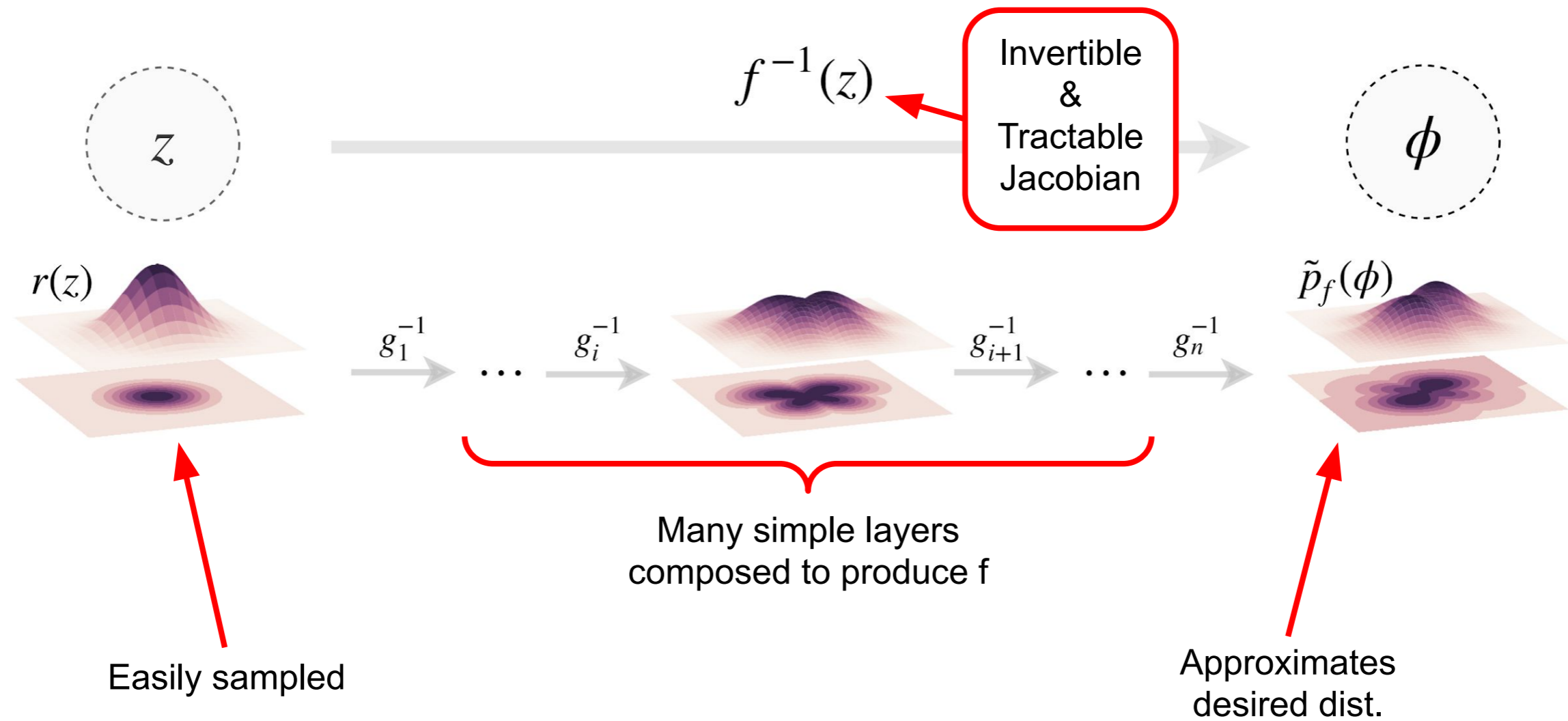
$$\tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z)$$



Construct complex functions via composition of simple functions

[Dinh et al. 1605.08803]

$$\tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z)$$



LIKELIHOOD RATIO TRICK

- **binary classifier**: find function $s(x)$ that minimizes **loss**:

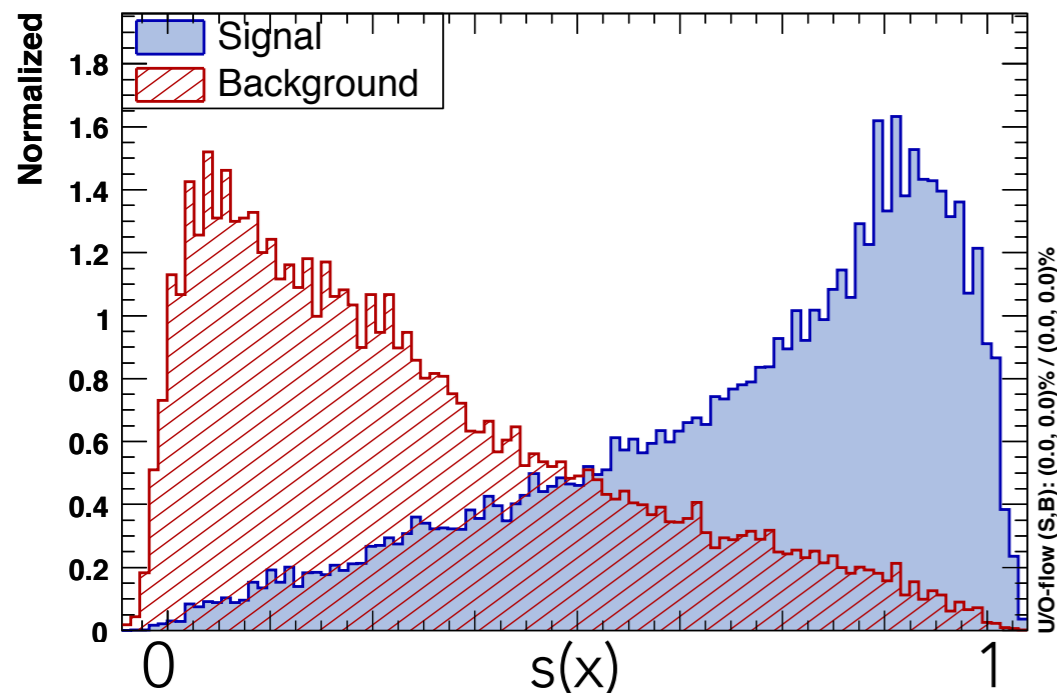
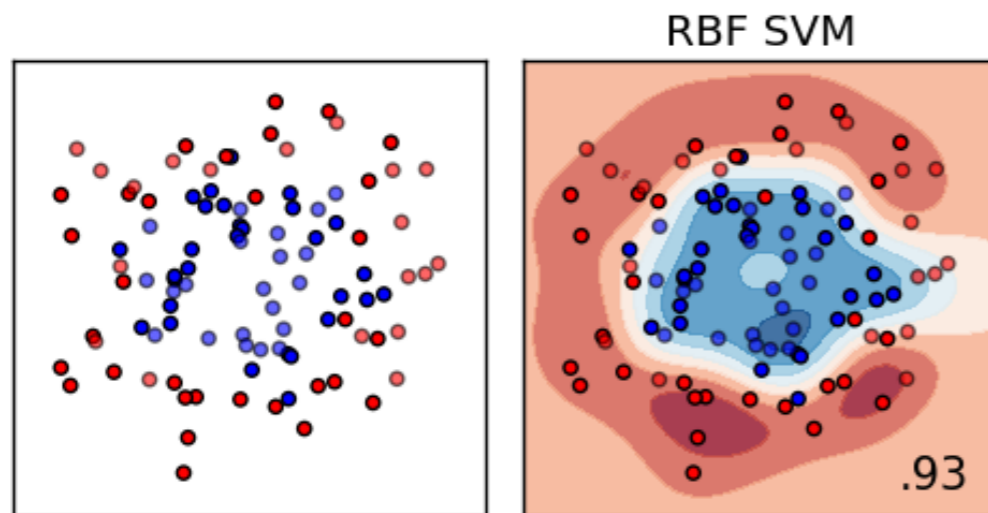
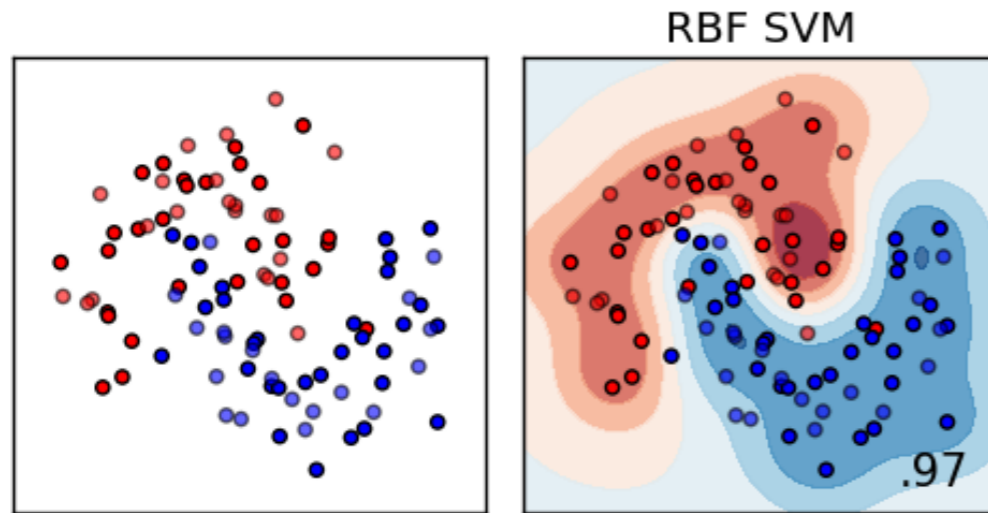
$$L[s] = \int p(x|H_0) (0 - s(x))^2 dx + \int p(x|H_1) (1 - s(x))^2 dx$$

- i.e. approximate the optimal classifier

$$s(x) = \frac{p(x|H_1)}{p(x|H_0) + p(x|H_1)}$$

- which is 1-to-1 with the likelihood ratio

$$\frac{p(x|H_1)}{p(x|H_0)}$$



LIKELIHOOD RATIO TRICK

- **binary classifier**: find function $s(x)$ that minimizes **loss**:

$$L[s] = \int p(x|H_0) (0 - s(x))^2 dx + \int p(x|H_1) (1 - s(x))^2 dx$$

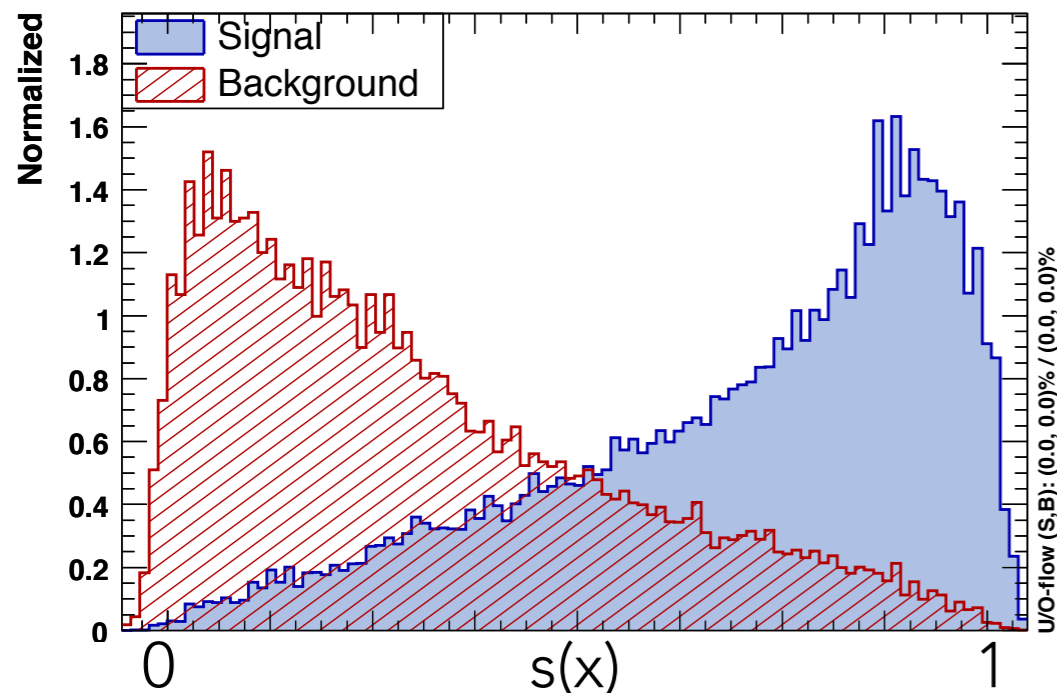
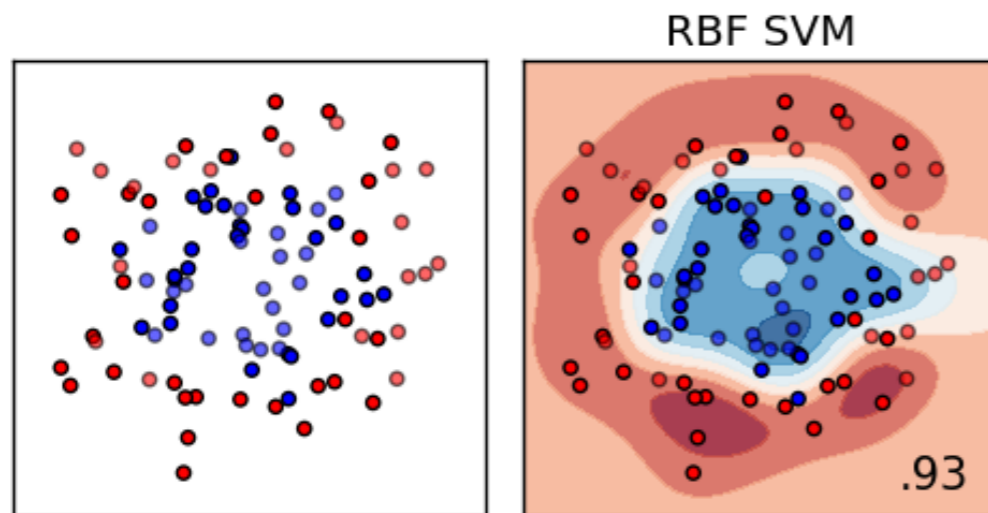
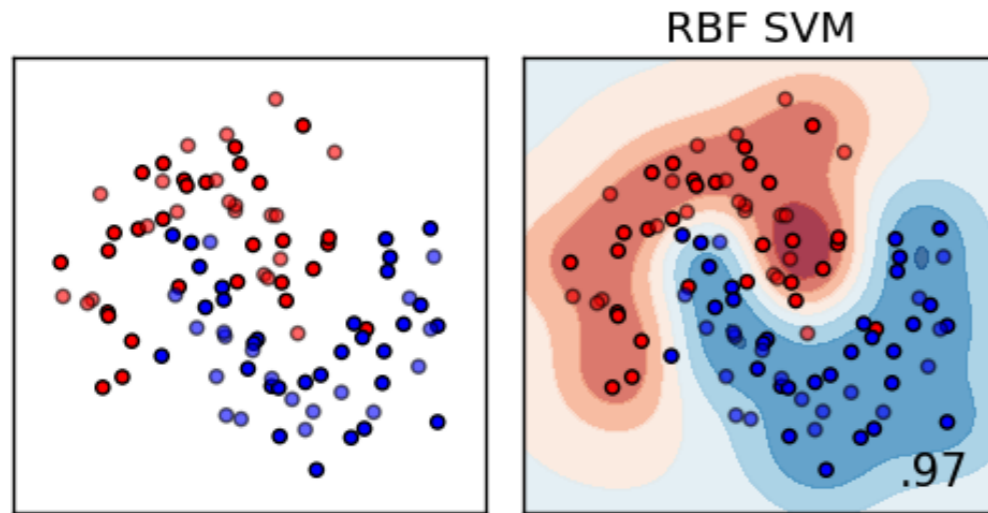
$$\approx \frac{1}{N} \sum_{i=1}^N (y_i - s(x_i))^2$$

- i.e. approximate the optimal classifier

$$s(x) = \frac{p(x|H_1)}{p(x|H_0) + p(x|H_1)}$$

- which is 1-to-1 with the likelihood ratio

$$\frac{p(x|H_1)}{p(x|H_0)}$$

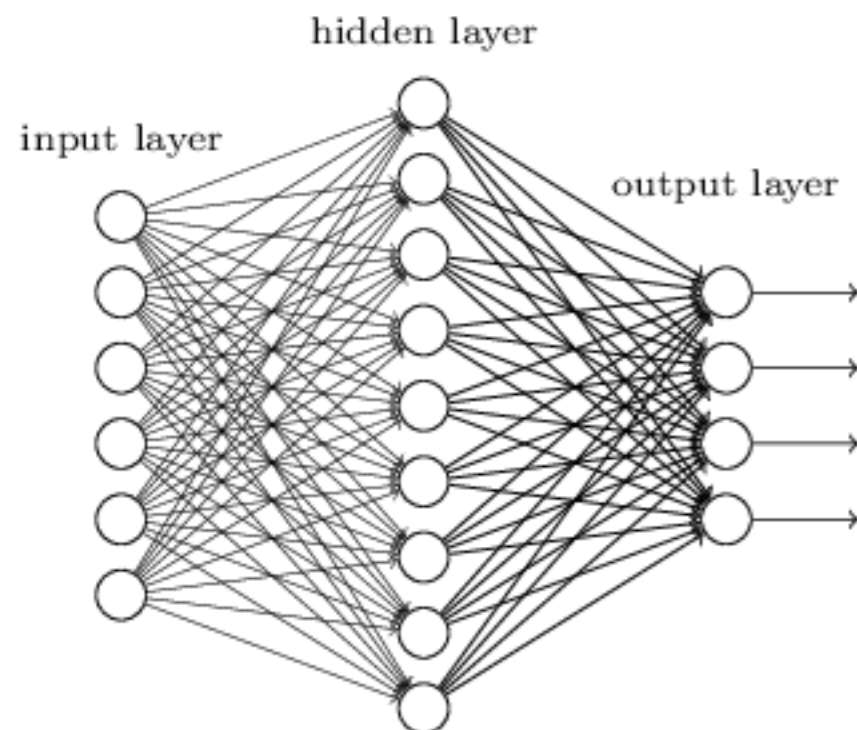


NN = A HIGHLY FLEXIBLE FAMILY OF FUNCTIONS

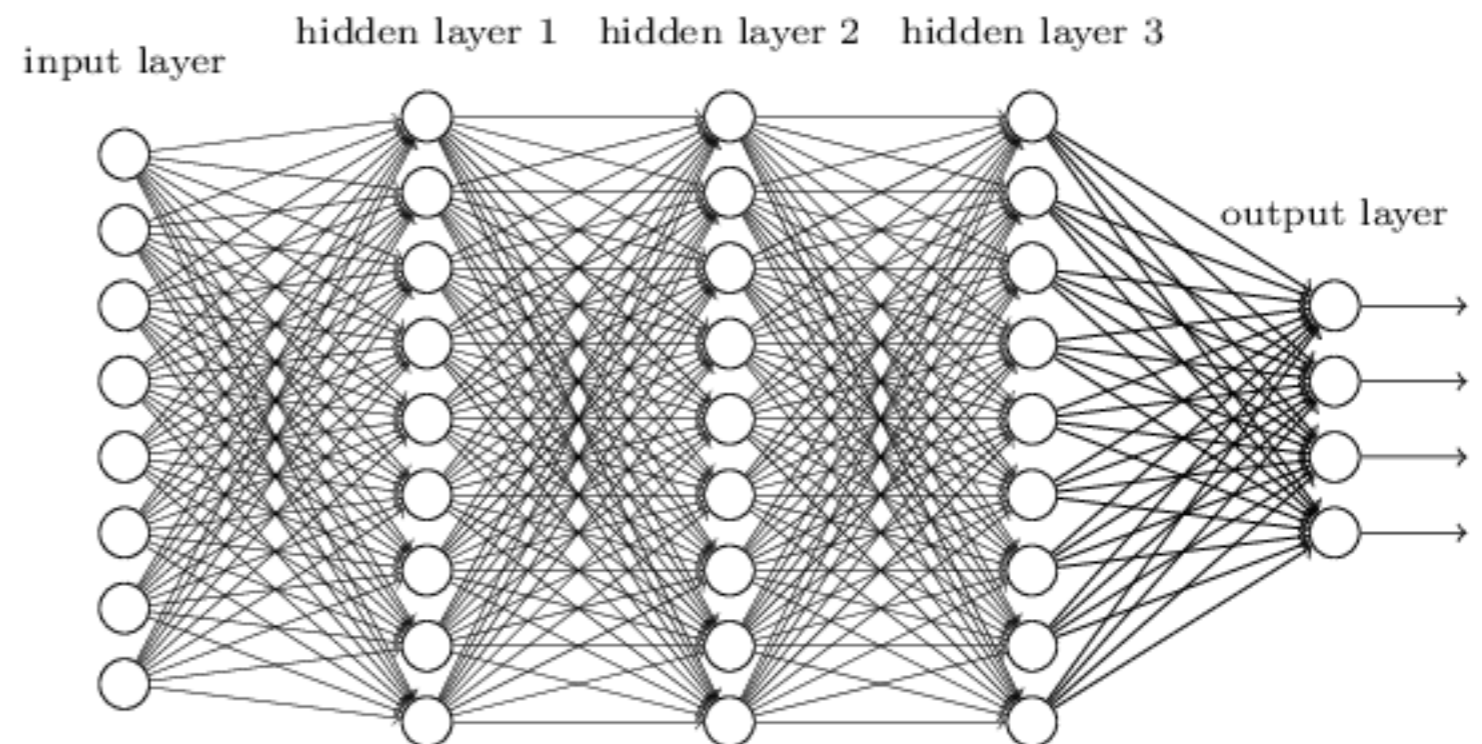
In calculus of variations, the optimization is over all functions: $\hat{s} = \operatorname{argmin}_s L[s]$

- In applied calculus of variations, we consider a highly flexible family of functions s_ϕ and optimize
- Think of neural networks as a highly flexible family of functions
- Machine learning also includes non-convex optimization algorithms that are effective even with millions of parameters!

' Shallow neural network



Deep neural network



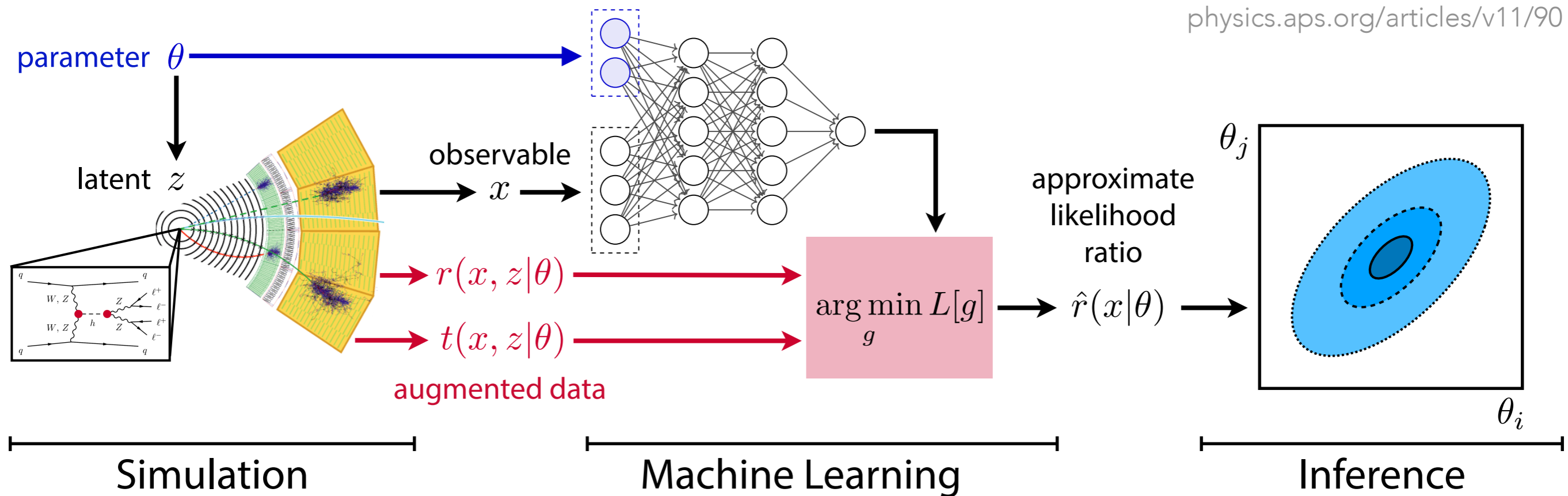
LEARNING THE LIKELIHOOD RATIO

arXiv:1805.12244

PRL, arXiv:1805.00013

PRD, arXiv:1805.00020

physics.aps.org/articles/v11/90

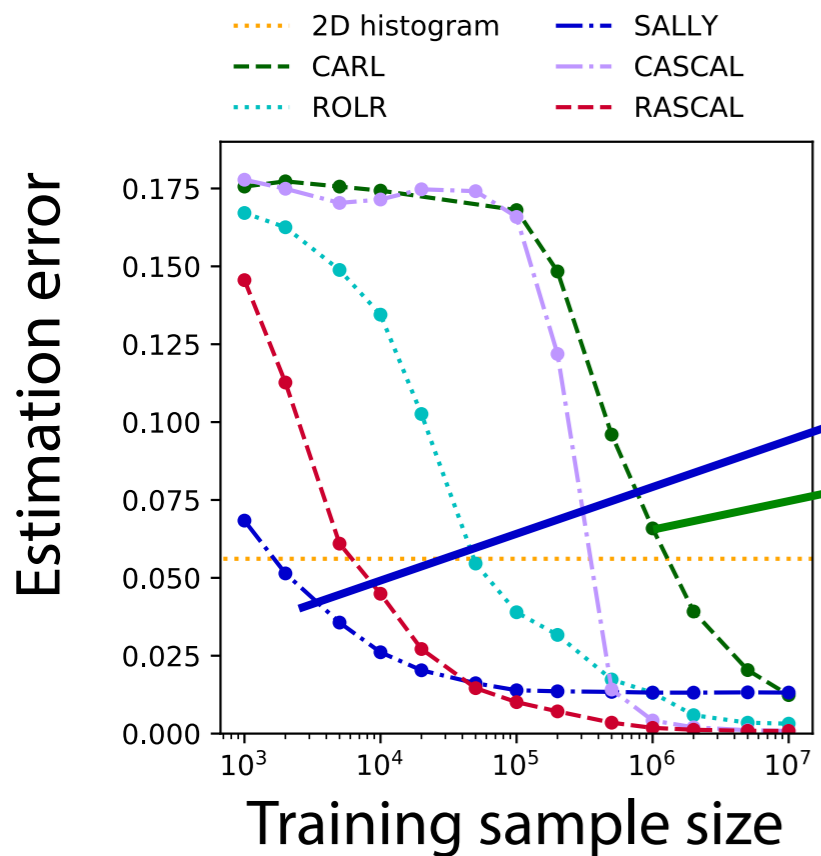
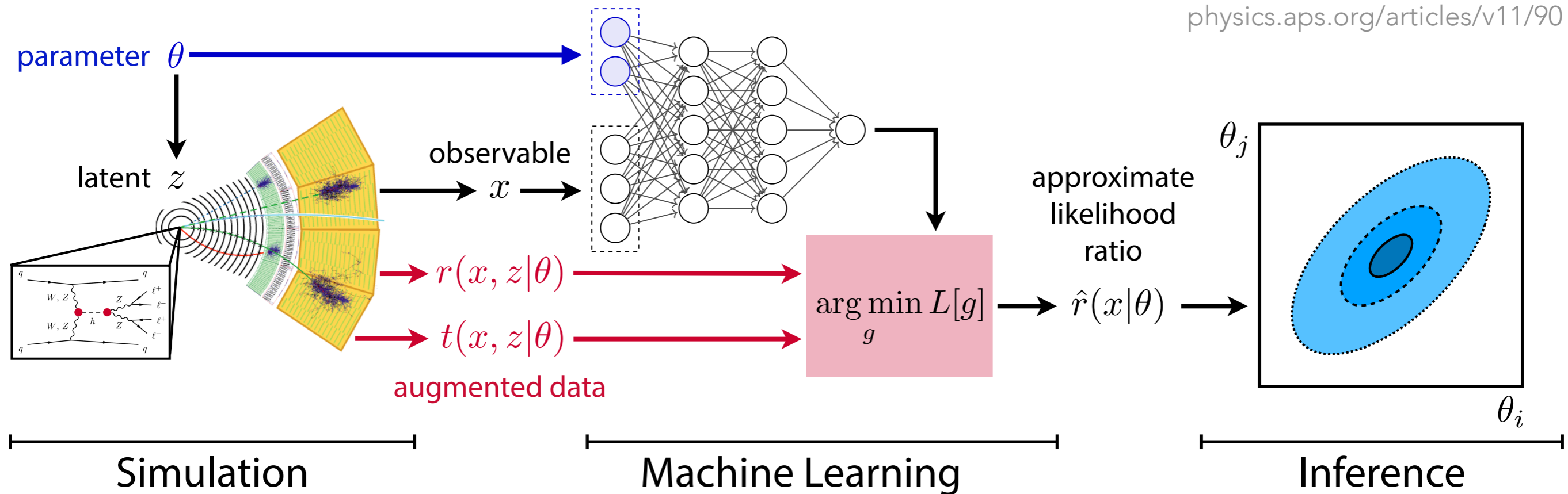


Recently, we realized we can **extract more from the simulator**.

We can use **augmented data** to improve training

(connections to reinforcement learning)

LEARNING THE LIKELIHOOD RATIO



New techniques require less data than without augmented data

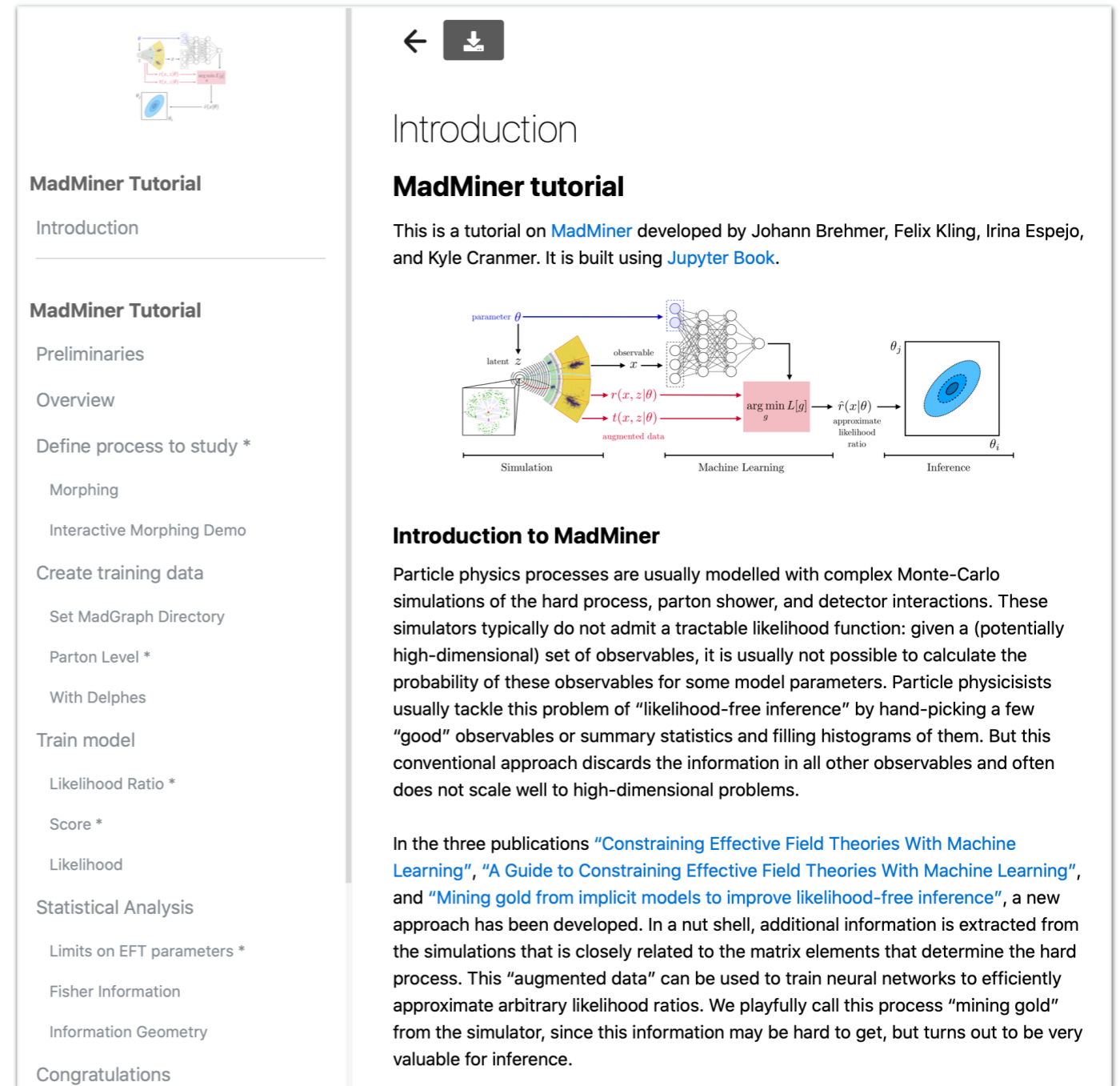
Traditional binned analysis

MADMINER

<https://cranmer.github.io/madminer-tutorial/>

MadMiner is a tool that streamlines these simulation-based inference techniques

- Works with any MadGraph model file
- Automates the “gold mining” process that makes training much more sample efficient
- Useful for experimental work or phenomenology studies



The screenshot shows the MadMiner tutorial website. On the left is a navigation menu with the following items: MadMiner Tutorial, Introduction, Preliminaries, Overview, Define process to study *, Morphing, Interactive Morphing Demo, Create training data, Set MadGraph Directory, Parton Level *, With Delphes, Train model, Likelihood Ratio *, Score *, Likelihood, Statistical Analysis, Limits on EFT parameters *, Fisher Information, Information Geometry, and Congratulations. The main content area is titled 'Introduction' and contains the text: 'This is a tutorial on MadMiner developed by Johann Brehmer, Felix Kling, Irina Espejo, and Kyle Cranmer. It is built using Jupyter Book.' Below the text is a diagram illustrating the MadMiner workflow. The diagram is divided into three stages: Simulation, Machine Learning, and Inference. In the Simulation stage, a parameter θ is used to generate a latent variable z , which is then used to produce an observable x . In the Machine Learning stage, the observable x is used to train a neural network, which outputs an approximate likelihood ratio $\hat{r}(x|\theta)$. In the Inference stage, the approximate likelihood ratio is used to estimate the parameters θ_i .

Introduction to MadMiner

Particle physics processes are usually modelled with complex Monte-Carlo simulations of the hard process, parton shower, and detector interactions. These simulators typically do not admit a tractable likelihood function: given a (potentially high-dimensional) set of observables, it is usually not possible to calculate the probability of these observables for some model parameters. Particle physicists usually tackle this problem of “likelihood-free inference” by hand-picking a few “good” observables or summary statistics and filling histograms of them. But this conventional approach discards the information in all other observables and often does not scale well to high-dimensional problems.

In the three publications “Constraining Effective Field Theories With Machine Learning”, “A Guide to Constraining Effective Field Theories With Machine Learning”, and “Mining gold from implicit models to improve likelihood-free inference”, a new approach has been developed. In a nut shell, additional information is extracted from the simulations that is closely related to the matrix elements that determine the hard process. This “augmented data” can be used to train neural networks to efficiently approximate arbitrary likelihood ratios. We playfully call this process “mining gold” from the simulator, since this information may be hard to get, but turns out to be very valuable for inference.



Gilles Louppe



Johann Brehmer



Felix Kling



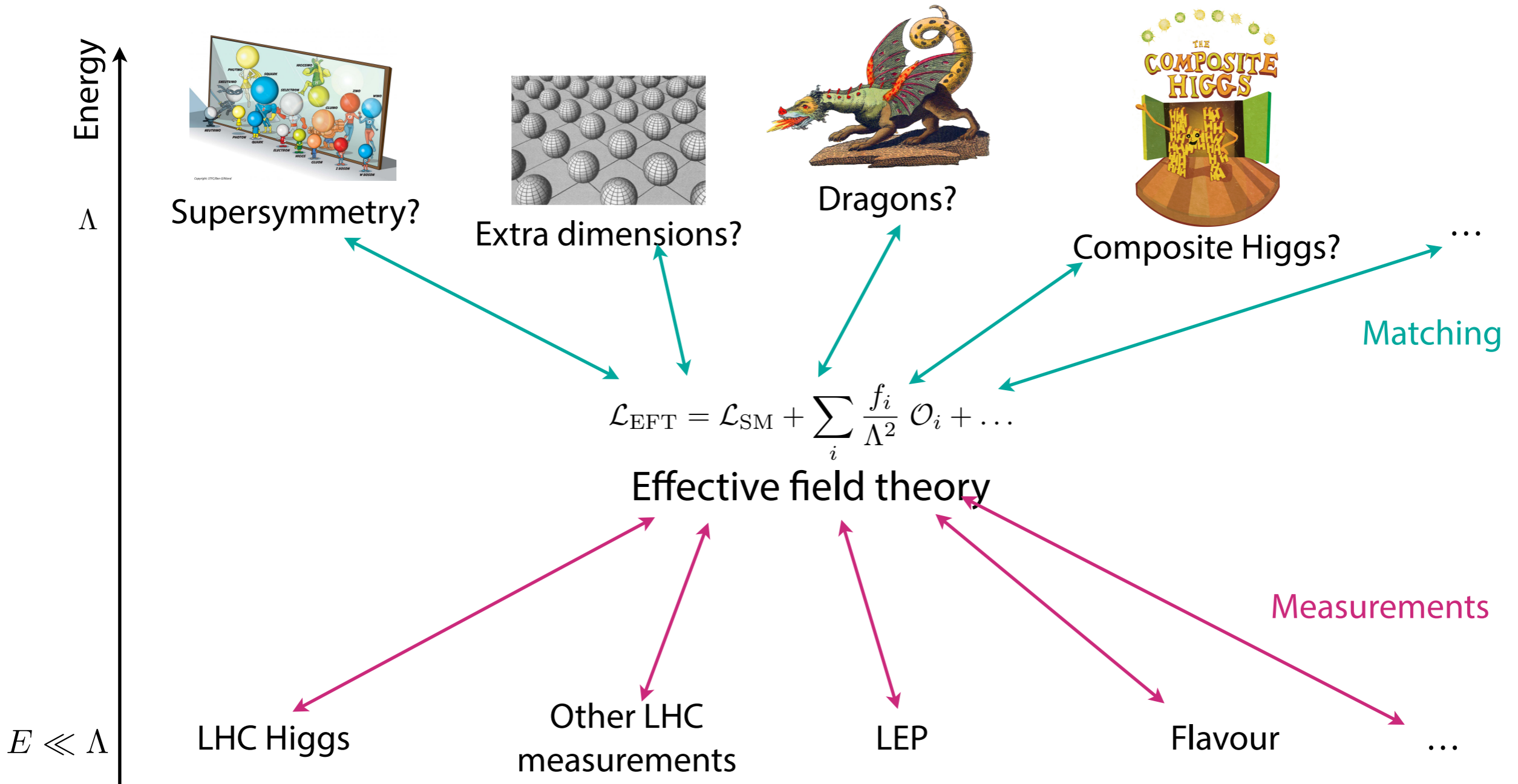
Irina Espejo



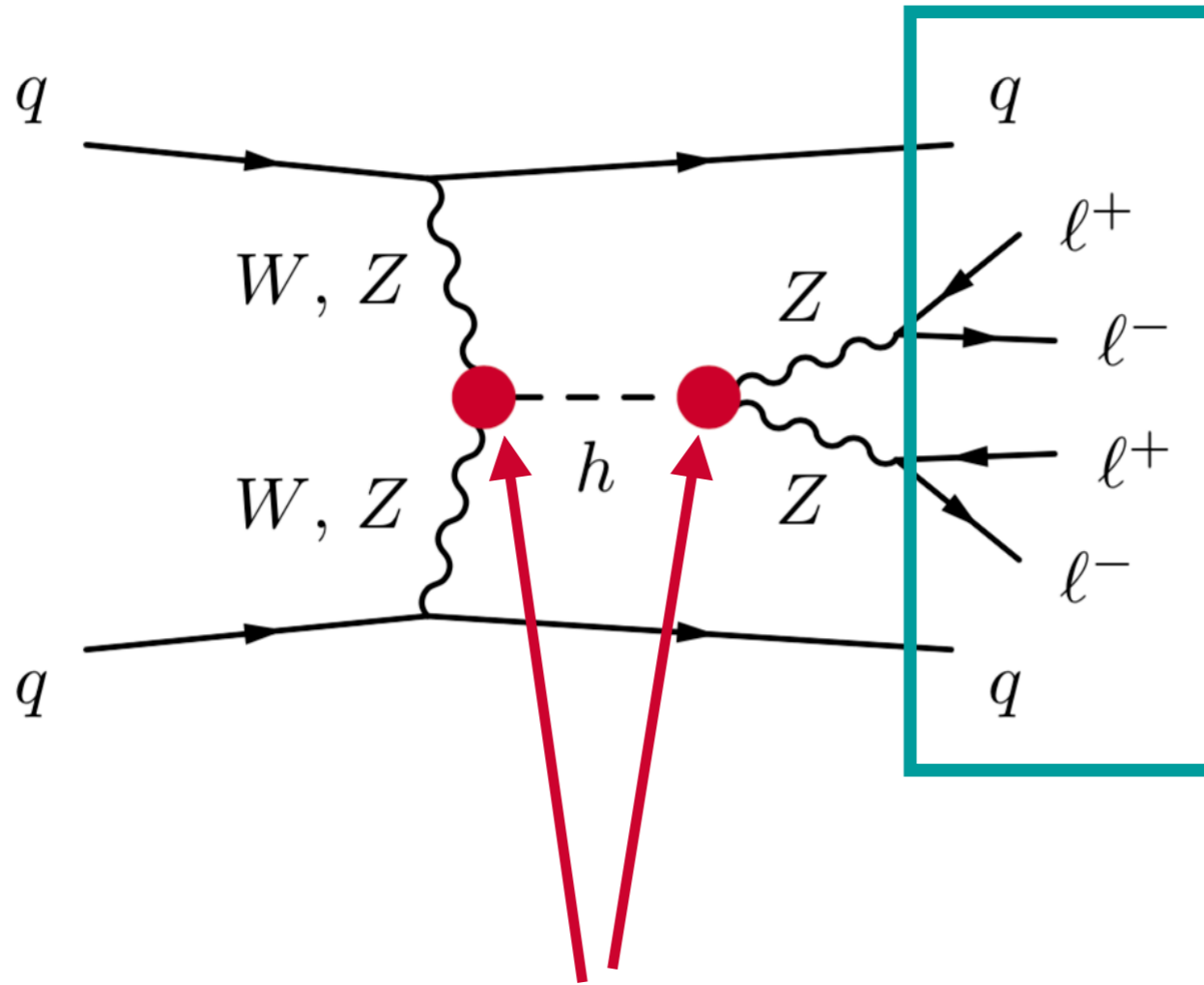
Sinclert Pérez

EFFECTIVE FIELD THEORY

[STFC / Ben Gililand, Sean Carroll, Friedrich Justin Bertuch 1806, symmetry]



IMPACT ON STUDIES OF THE HIGGS BOSON



16-Dim phase space,
but we will use a
42-Dim observable \mathbf{x}
with redundant
information

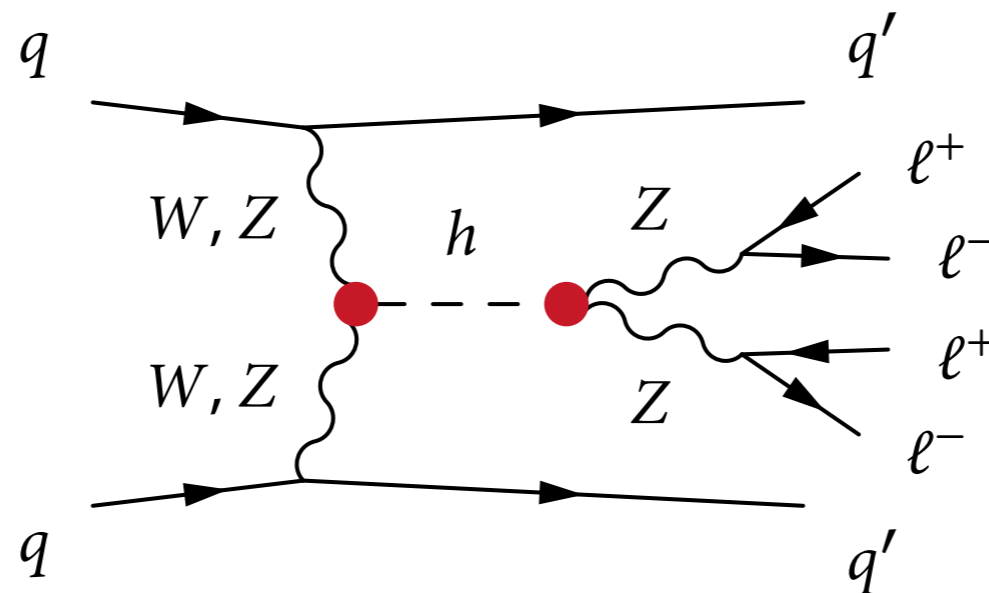
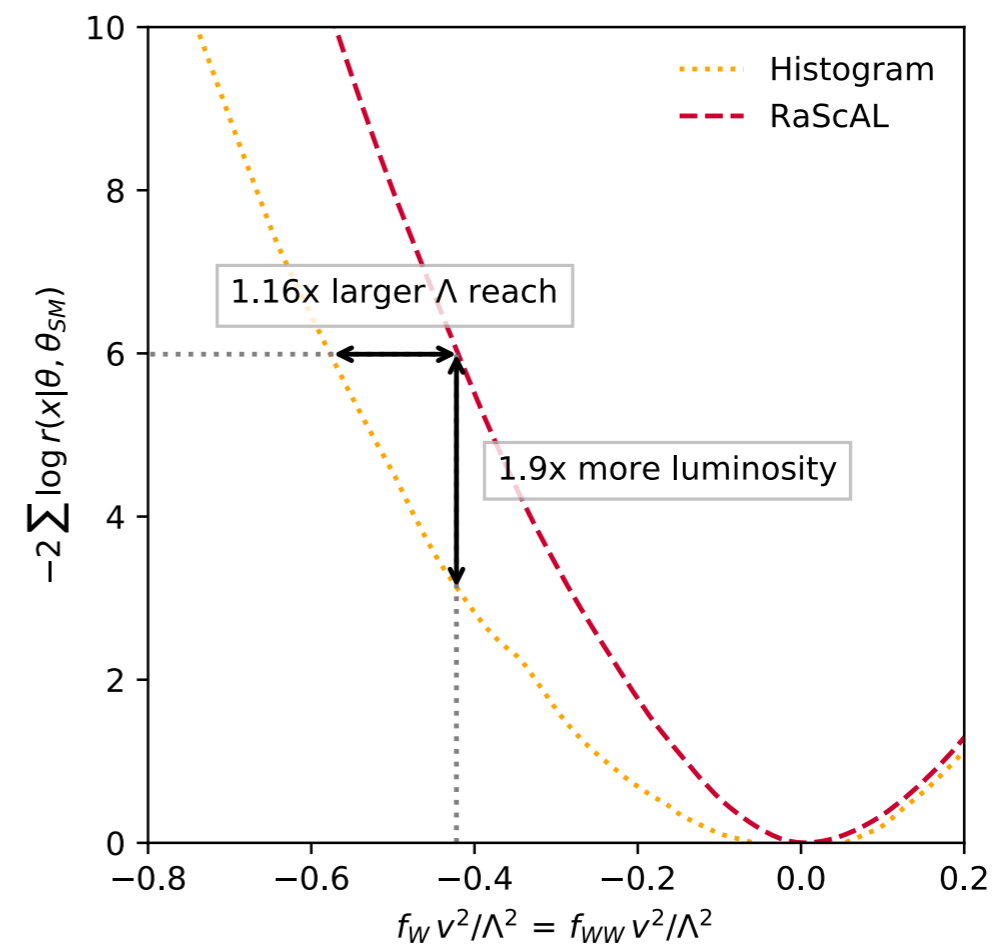
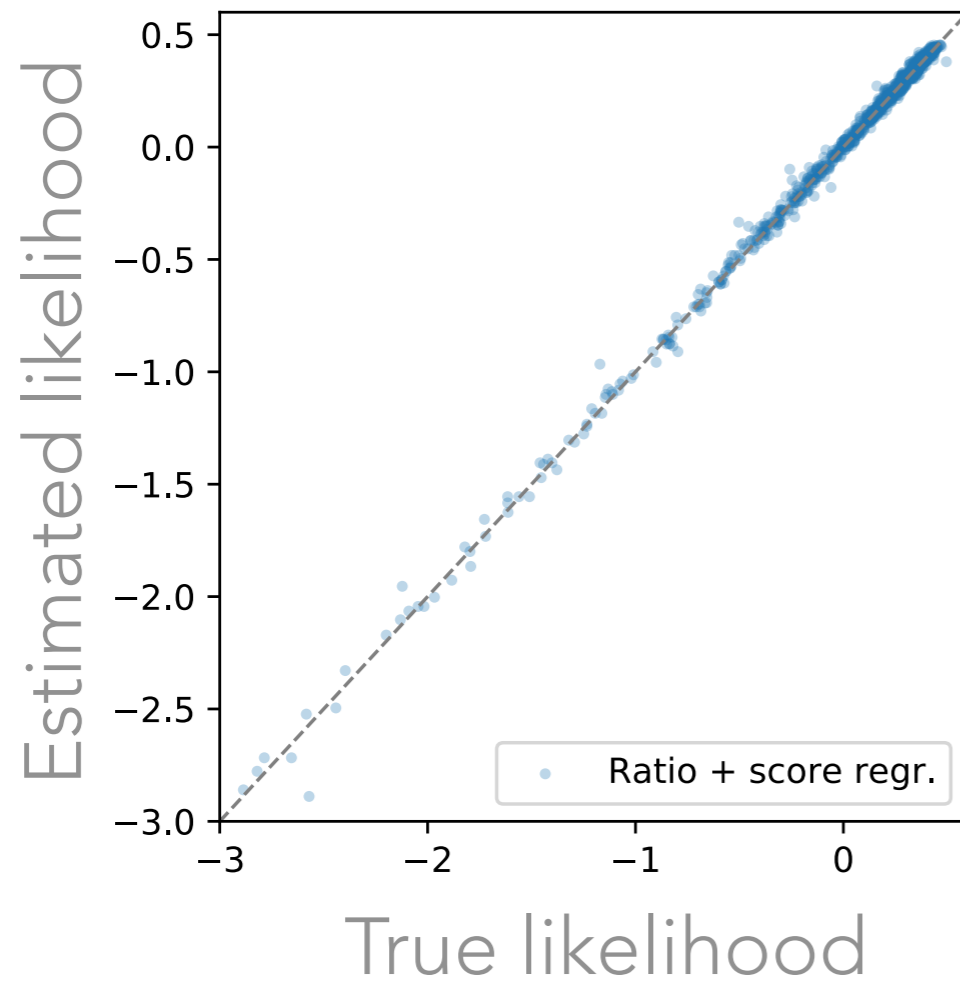
Exciting new physics might hide here!

We parameterize it with two EFT coefficients:

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \boxed{\frac{f_W}{\Lambda^2}} \underbrace{\frac{ig}{2} (D^\mu \phi)^\dagger \sigma^a D^\nu \phi W_{\mu\nu}^a}_{\mathcal{O}_W} - \boxed{\frac{f_{WW}}{\Lambda^2}} \underbrace{\frac{g^2}{4} (\phi^\dagger \phi) W_{\mu\nu}^a W^{\mu\nu a}}_{\mathcal{O}_{WW}}$$

IMPACT ON STUDIES OF THE HIGGS BOSON

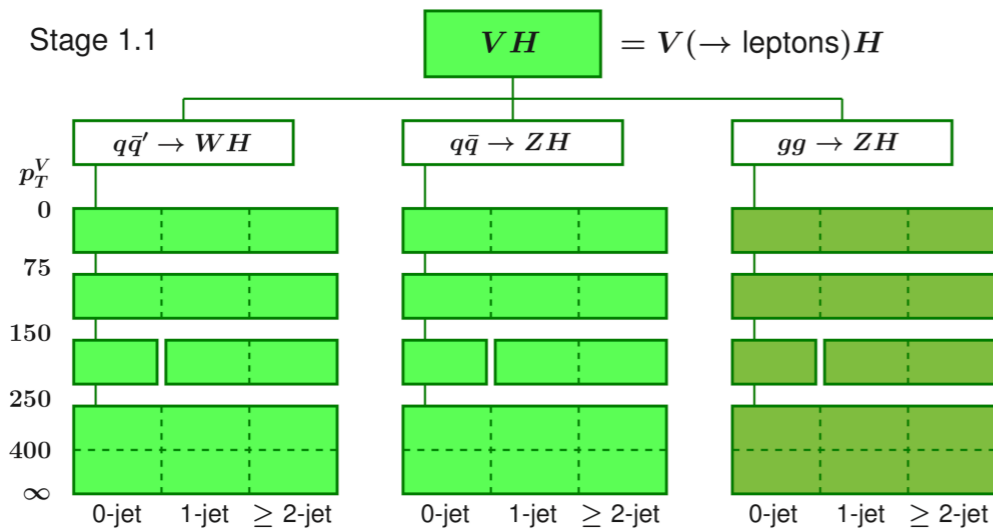
(based on a 42-Dim observation \mathbf{x})



MORE COMPARISONS TO BASELINE APPROACH @ LHC

- **Baseline: Simplified Template Cross-Sections (STXS)**
define observable regions that try to capture as much information on new physics as possible

[N. Berger et al. 1906.02754; HXSWG YR4]



- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger \phi) \Box (\phi^\dagger \phi) - \frac{1}{4} (\phi^\dagger D^\mu \phi)^* (\phi^\dagger D_\mu \phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger \phi W_{\mu\nu}^a W^{\mu\nu a}$$

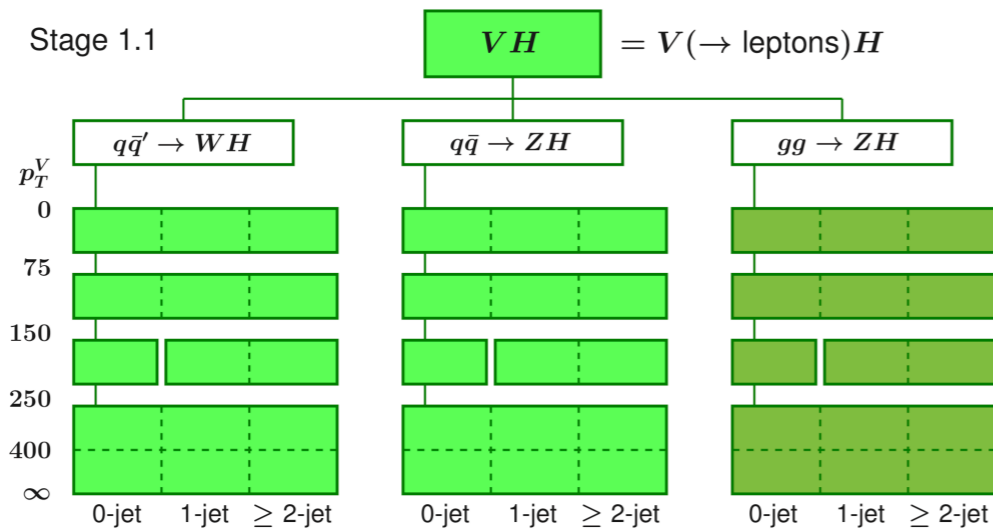
$$\mathcal{O}_{Hq}^{(3)} = (\phi^\dagger i \overleftrightarrow{D}_\mu^a \phi) (\bar{Q}_L \sigma^a \gamma^\mu Q_L),$$

can we extract from. $pp \rightarrow WH \rightarrow \ell\nu b\bar{b}$

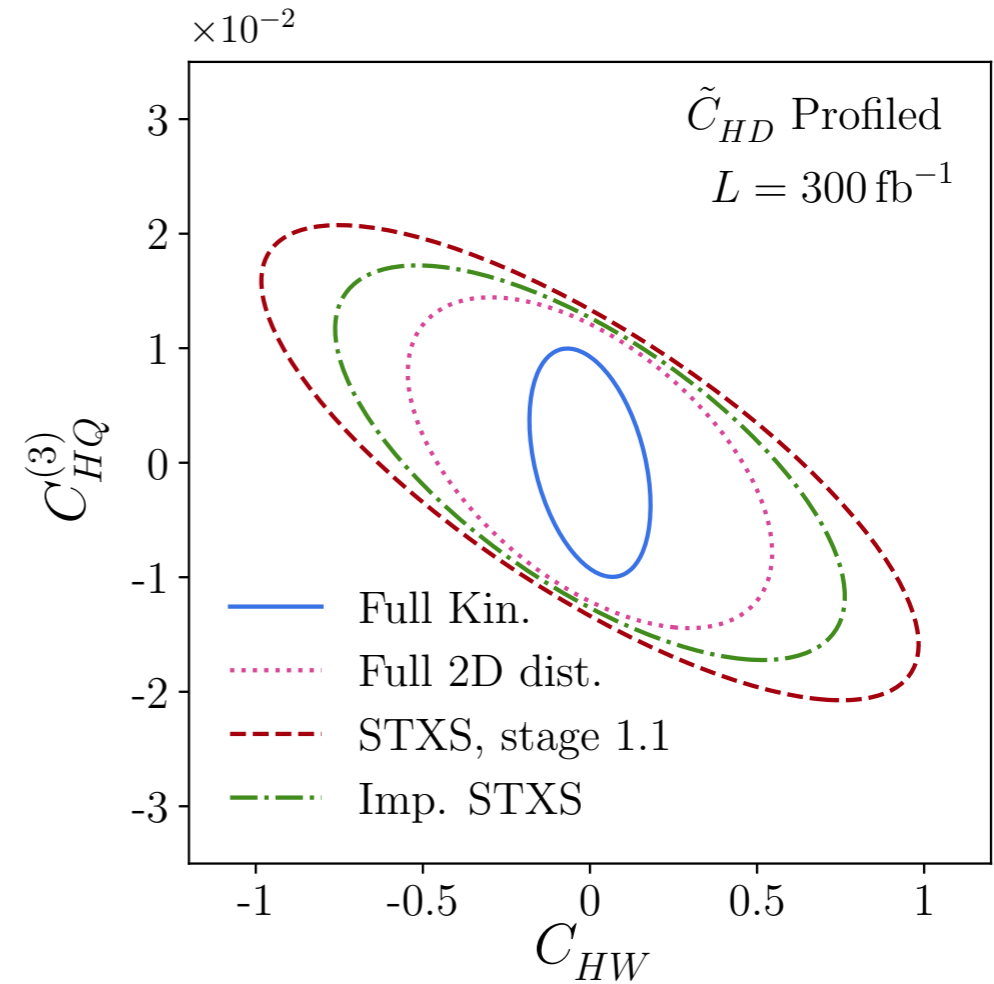
MORE COMPARISONS TO BASELINE APPROACH @ LHC

- Baseline: Simplified Template Cross-Sections (**STXS**) define observable regions that try to capture as much information on new physics as possible

[N. Berger et al. 1906.02754; HXSWG YR4]



- Results: **STXS** are sensitive to operators, adding a few more regions improve them, but a **multivariate analysis** is *much* stronger!



- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger \phi) \Box (\phi^\dagger \phi) - \frac{1}{4} (\phi^\dagger D^\mu \phi)^* (\phi^\dagger D_\mu \phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger \phi W_{\mu\nu}^a W^{\mu\nu a}$$

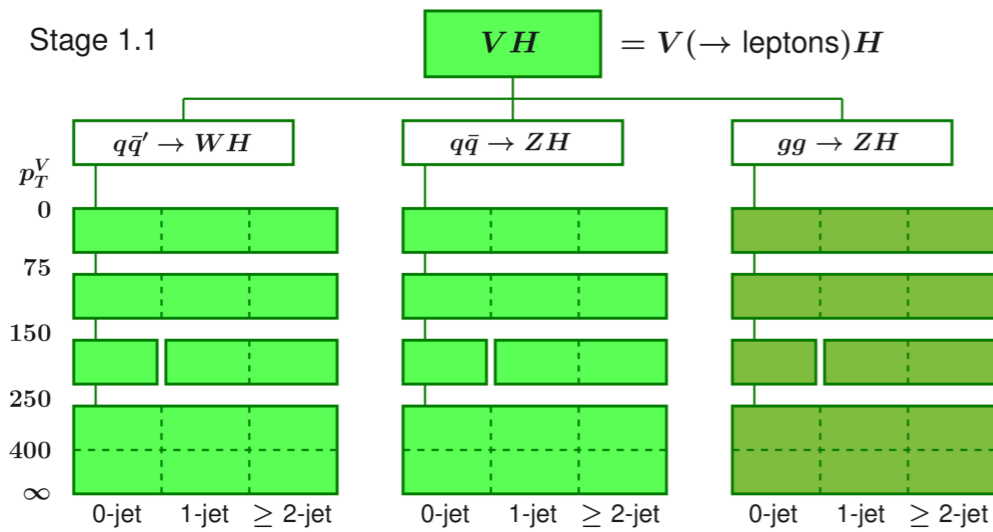
$$\mathcal{O}_{Hq}^{(3)} = (\phi^\dagger i \overleftrightarrow{D}_\mu^a \phi) (\bar{Q}_L \sigma^a \gamma^\mu Q_L),$$

can we extract from. $pp \rightarrow WH \rightarrow \ell\nu b\bar{b}$

MORE COMPARISONS TO BASELINE APPROACH @ LHC

- Baseline: Simplified Template Cross-Sections (**STXS**) define observable regions that try to capture as much information on new physics as possible

[N. Berger et al. 1906.02754; HXSWG YR4]



- Results: **STXS** are sensitive to operators, adding a few more regions improve them, but a **multivariate analysis** is *much* stronger!

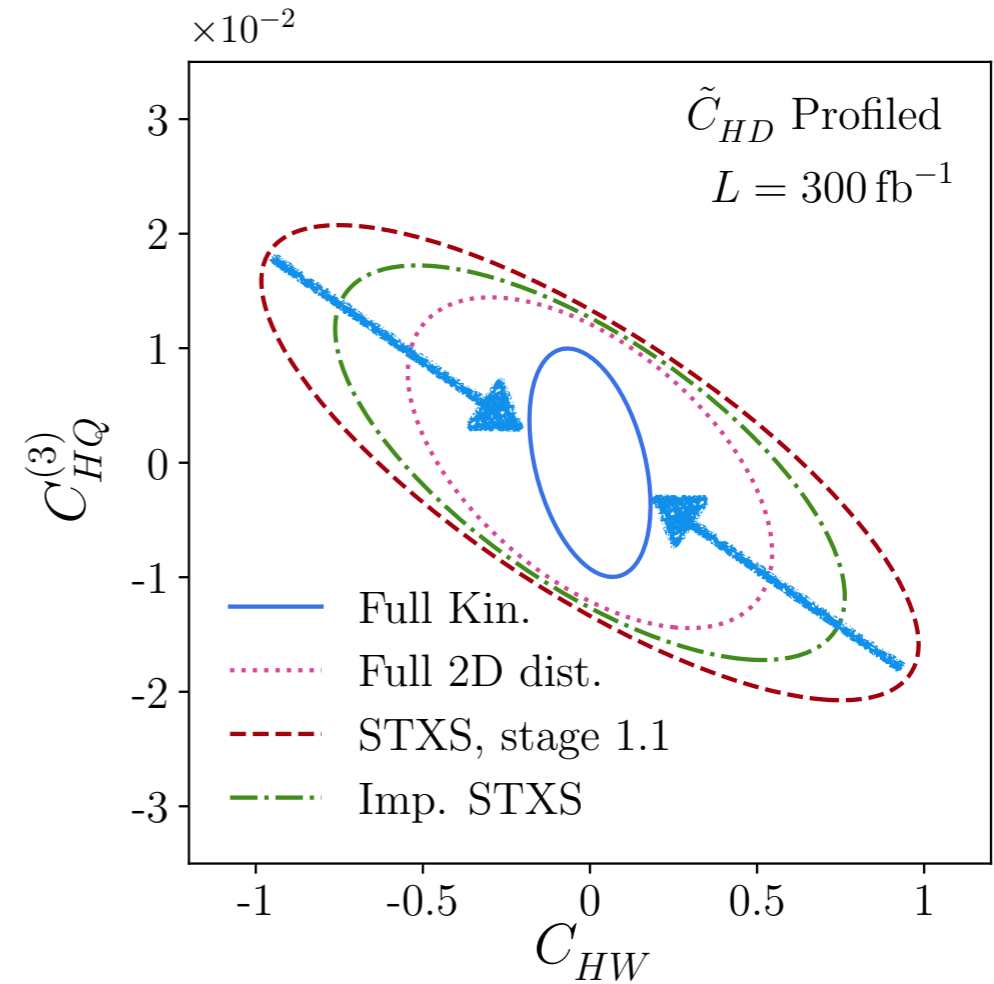
- Let's check! How much information on

$$\tilde{\mathcal{O}}_{HD} = \mathcal{O}_{H\Box} - \frac{\mathcal{O}_{HD}}{4} = (\phi^\dagger \phi) \Box (\phi^\dagger \phi) - \frac{1}{4} (\phi^\dagger D^\mu \phi)^* (\phi^\dagger D_\mu \phi)$$

$$\mathcal{O}_{HW} = \phi^\dagger \phi W_{\mu\nu}^a W^{\mu\nu a}$$

$$\mathcal{O}_{Hq}^{(3)} = (\phi^\dagger i \overleftrightarrow{D}_\mu^a \phi) (\bar{Q}_L \sigma^a \gamma^\mu Q_L),$$

can we extract from. $pp \rightarrow WH \rightarrow \ell\nu b\bar{b}$



FULLY LEPTONIC WZ

Recently, a variant of these models proposed by Chen, Glioti, Panico, Wulzer and applied to fully leptonic WZ

- “Quadratic classifier” (QC) outperforms binned analysis and approaches exact matrix element information

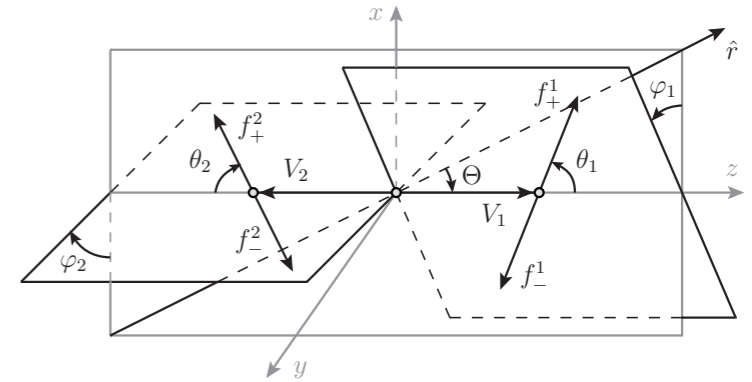
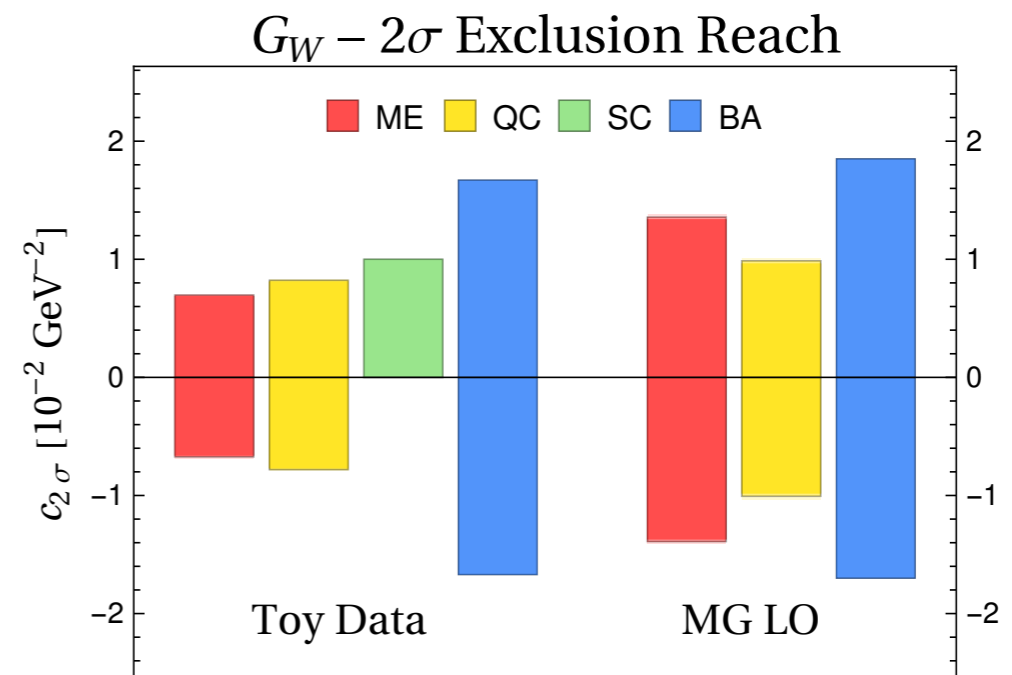
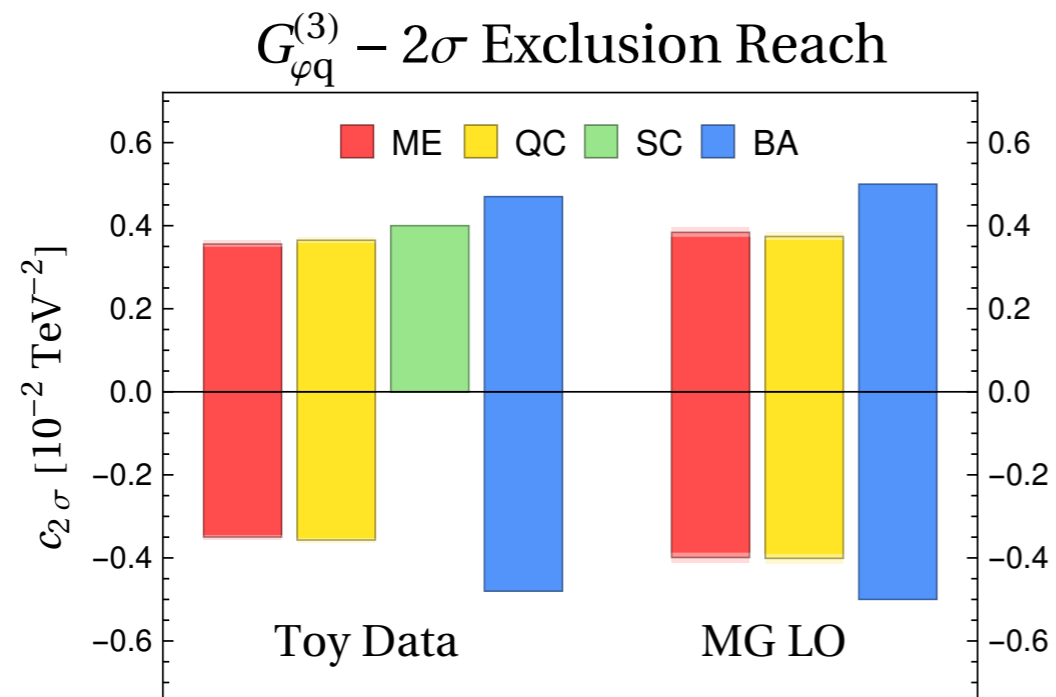


Figure 1: The kinematical variables in the “special” coordinate frame [28].

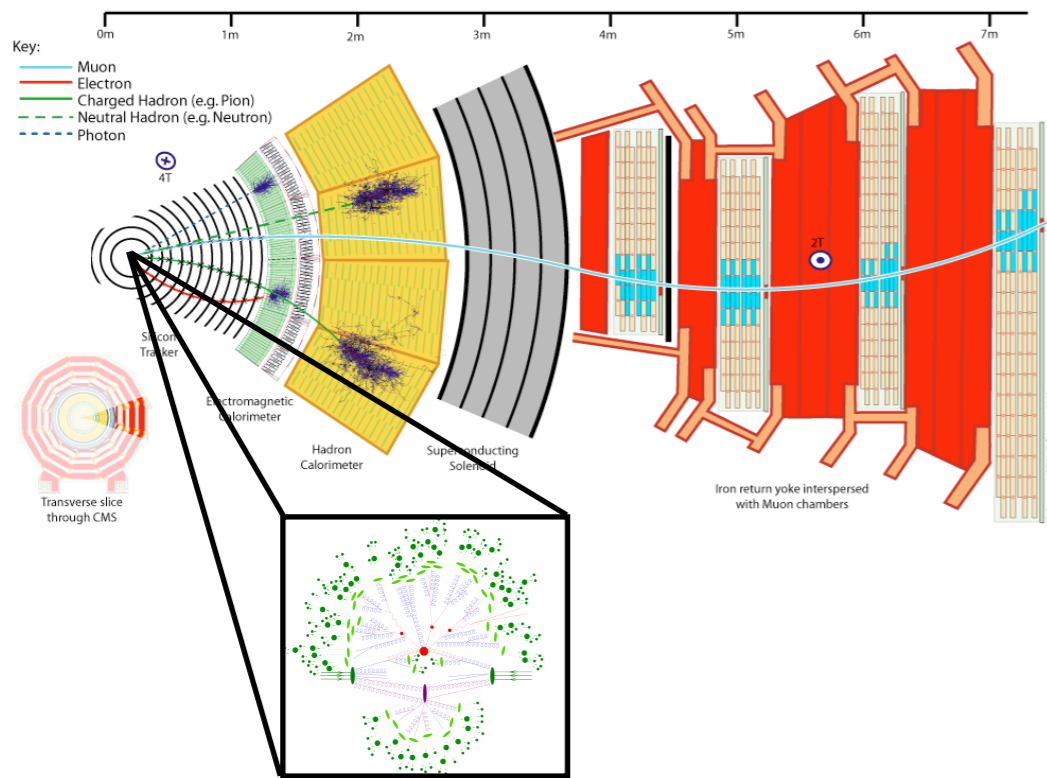
$$\mathcal{O}_{\varphi q}^{(3)} = G_{\varphi q}^{(3)} (\bar{Q}_L \sigma^a \gamma^\mu Q_L) (iH^\dagger \overleftrightarrow{D}_\mu H),$$

$$\mathcal{O}_W = G_W \varepsilon_{abc} W_\mu^{a\nu} W_\nu^{b\rho} W_\rho^{c\mu}.$$



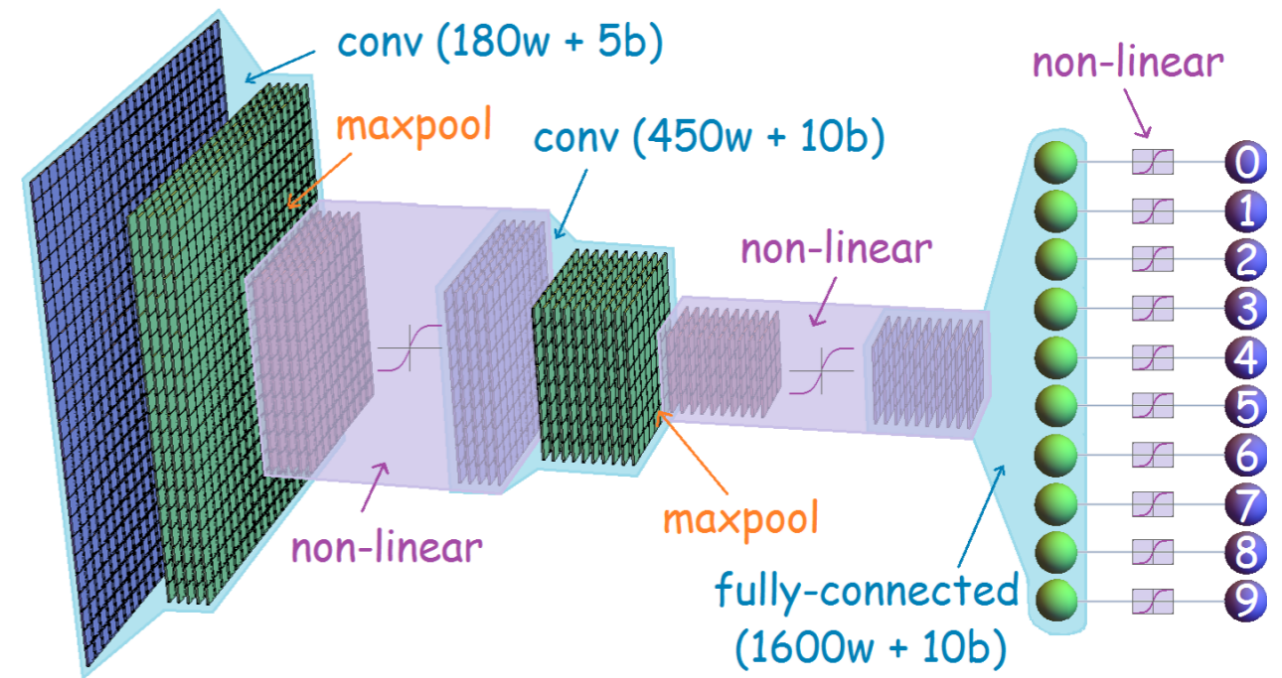
TWO APPROACHES TO SIMULATION-BASED INFERENCE

Use simulator
(much more efficiently)



- Approximate Bayesian Computation (ABC)
- Probabilistic Programming
- Adversarial Variational Optimization (AVO)

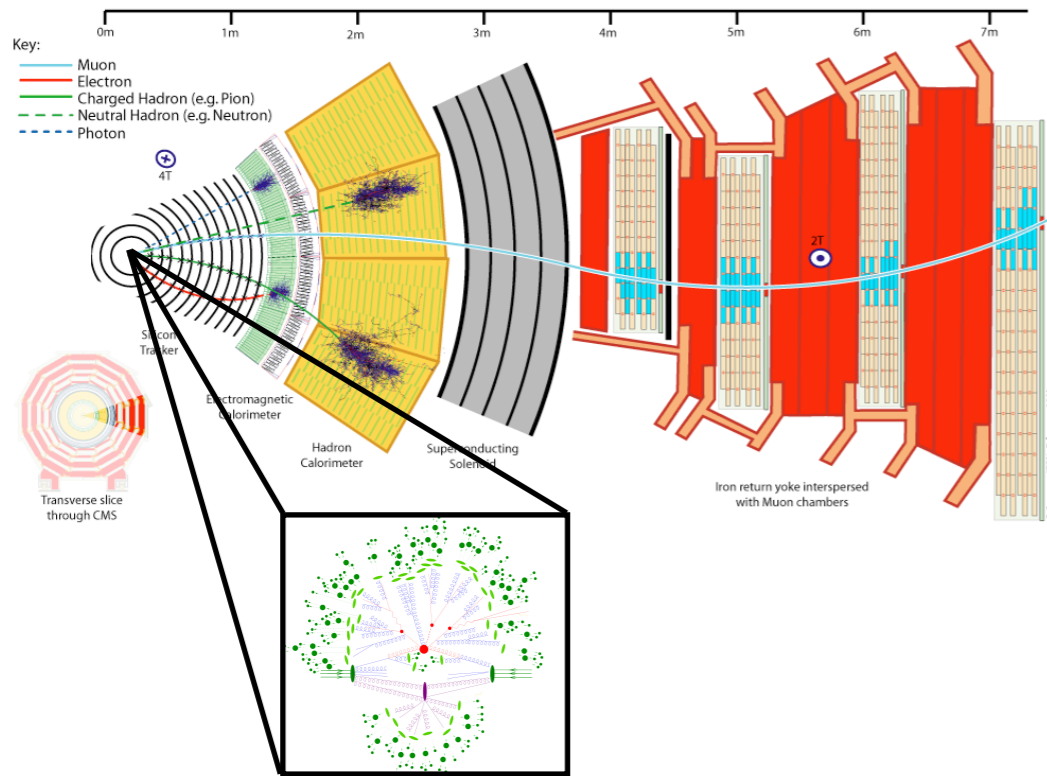
Learn simulator
(with deep learning)



- Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAE)
- Likelihood ratio from classifiers (CARL)
- Autogressive models, Normalizing Flows

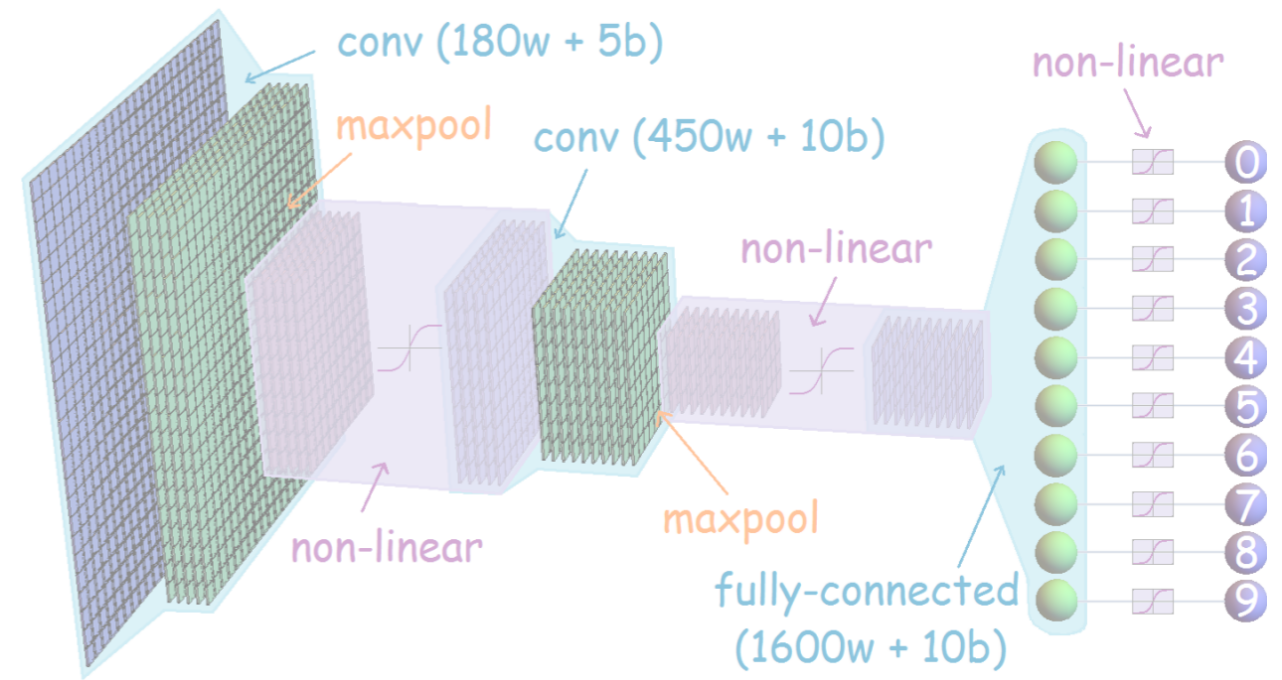
TWO APPROACHES TO SIMULATION-BASED INFERENCE

Use simulator
(much more efficiently)



- Approximate Bayesian Computation (ABC)
- Probabilistic Programming
- Adversarial Variational Optimization (AVO)

Learn simulator
(with deep learning)



- Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAE)
- Likelihood ratio from classifiers (CARL)
- Autogressive models, Normalizing Flows

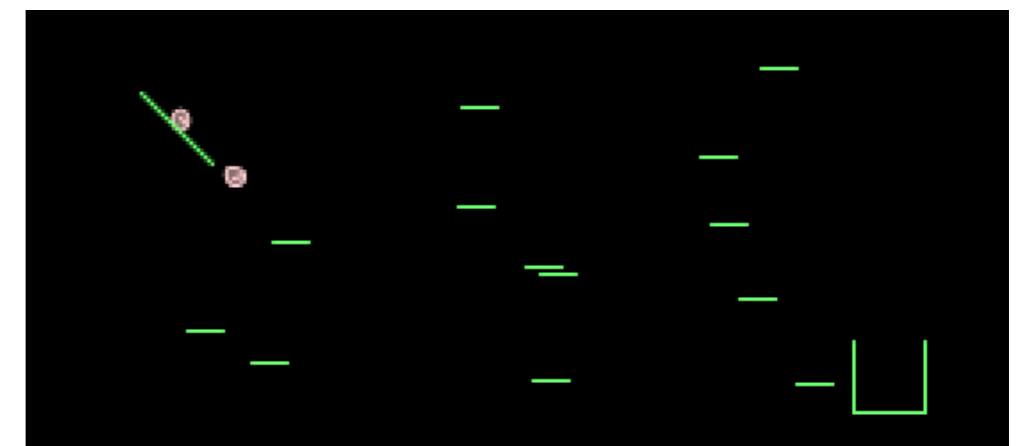
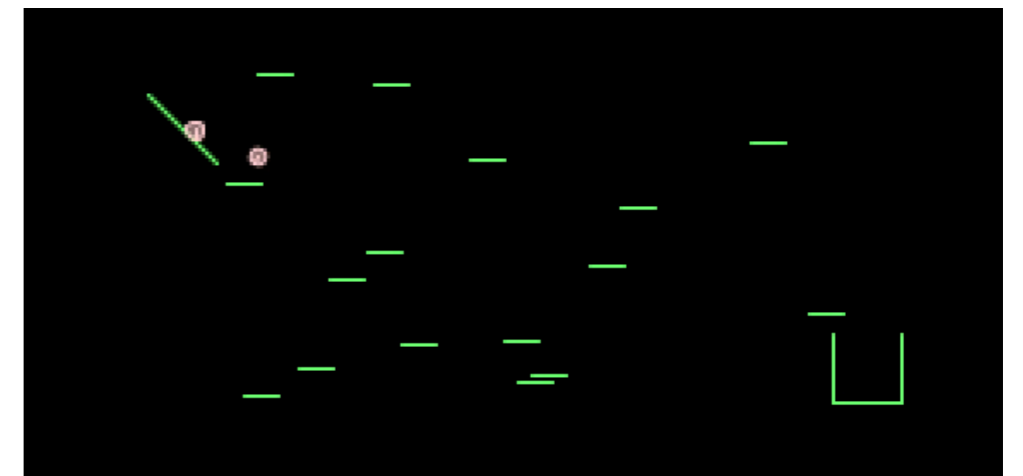
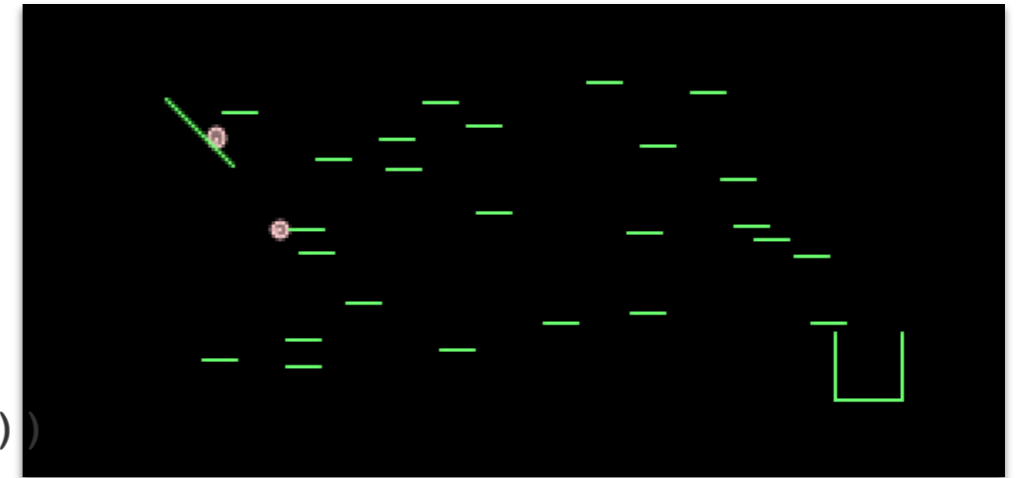
PROBABILISTIC PROGRAMMING EXAMPLE

```
(defquery arrange-bumpers []
  (let [number-of-bumpers (sample (poisson 20))
        bumpydist (uniform-continuous 0 10)
        bumpxdist (uniform-continuous -5 14)
        bumper-positions (repeatedly
                           number-of-bumpers
                           #(vector (sample bumpxdist)
                                   (sample bumpydist))))

        ;; code to simulate the world
        world (create-world bumper-positions)
        end-world (simulate-world world)
        balls (:balls end-world)

        ;; how many balls entered the box?
        num-balls-in-box (balls-in-box end-world)]

    {:balls balls
     :num-balls-in-box num-balls-in-box
     :bumper-positions bumper-positions}))
```



3 examples generated from simulator

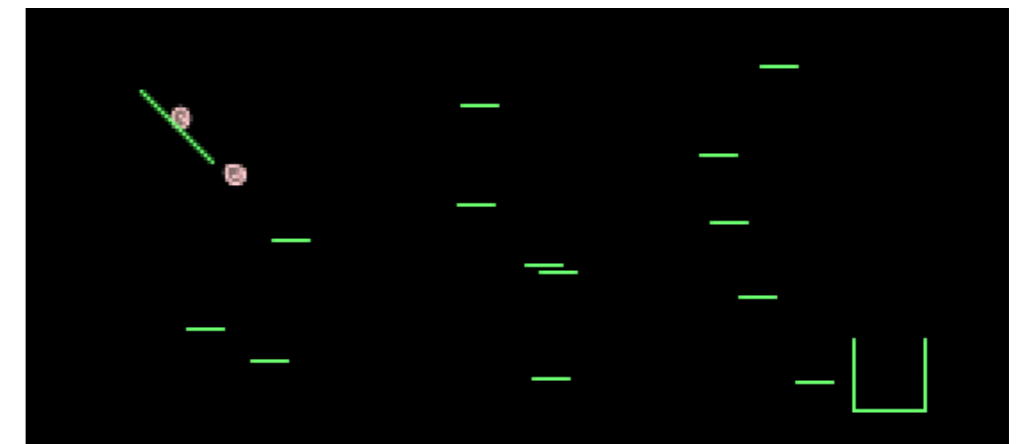
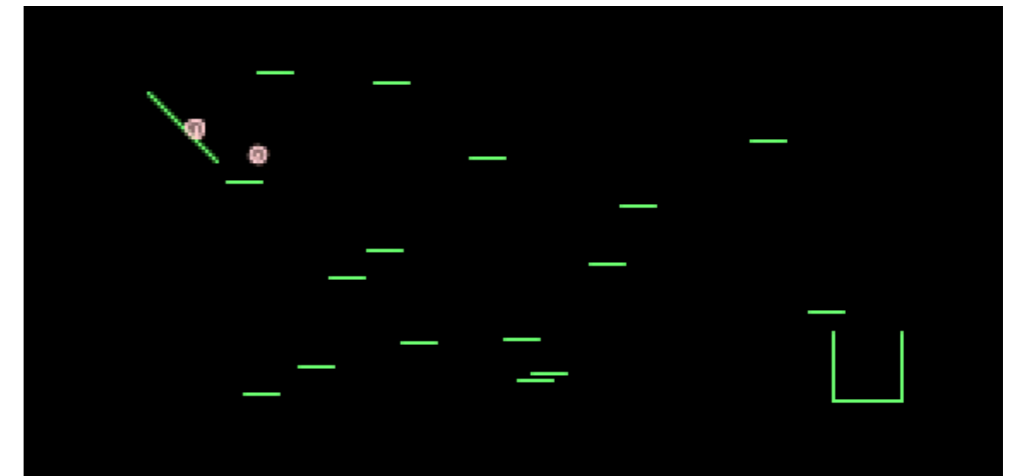
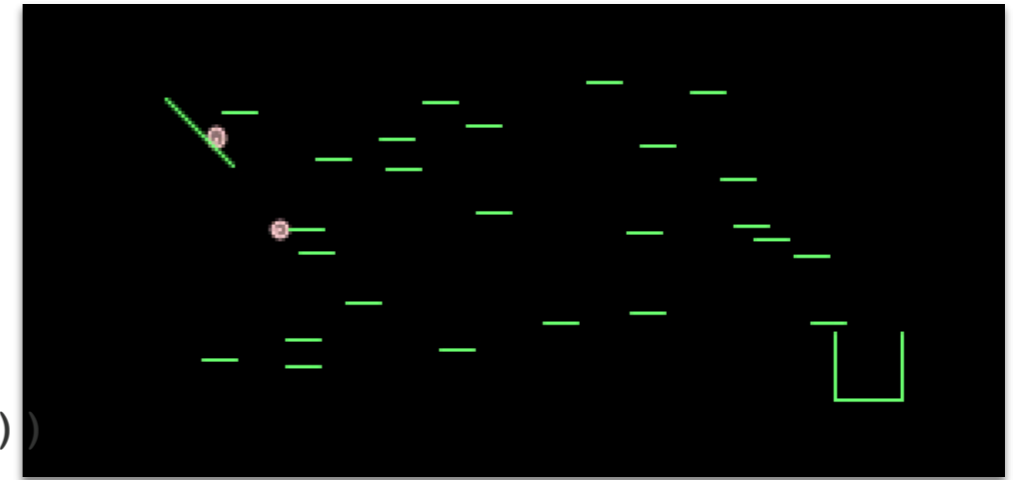
PROBABILISTIC PROGRAMMING EXAMPLE

```
(defquery arrange-bumpers []
  (let [number-of-bumpers (sample (poisson 20))
        bumpydist (uniform-continuous 0 10)
        bumpxdist (uniform-continuous -5 14)
        bumper-positions (repeatedly
                           number-of-bumpers
                           #(vector (sample bumpxdist)
                                   (sample bumpydist))))

        ;; code to simulate the world
        world (create-world bumper-positions)
        end-world (simulate-world world)
        balls (:balls end-world)

        ;; how many balls entered the box?
        num-balls-in-box (balls-in-box end-world)]

    {:balls balls
     :num-balls-in-box num-balls-in-box
     :bumper-positions bumper-positions}))
```



3 examples generated from simulator

PROBABILISTIC PROGRAMMING EXAMPLE

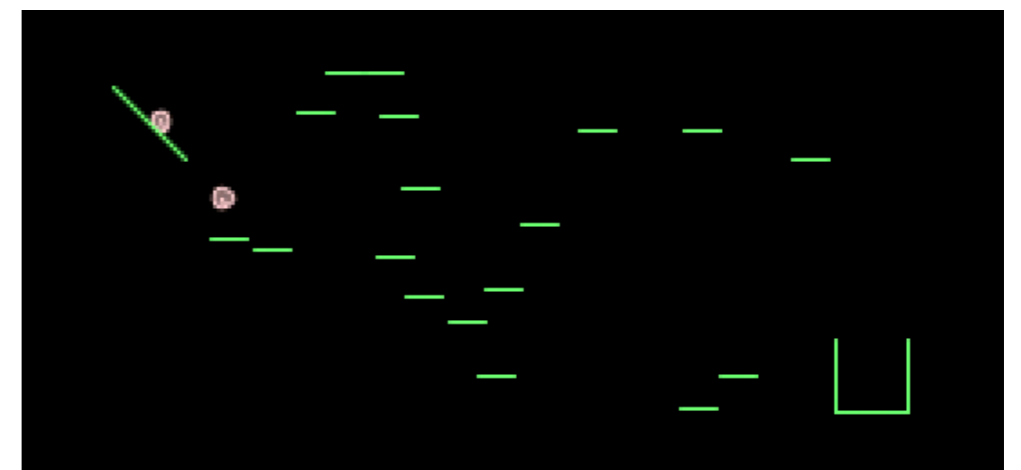
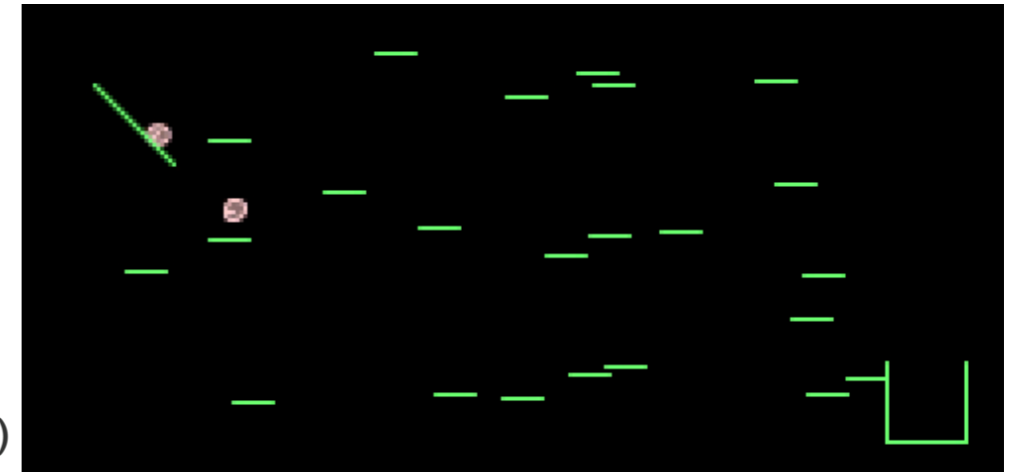
```
(defquery arrange-bumpers []
  (let [number-of-bumpers (sample (poisson 20))
        bumpydist (uniform-continuous 0 10)
        bumpxdist (uniform-continuous -5 14)
        bumper-positions (repeatedly
                          number-of-bumpers
                          #(vector (sample bumpxdist)
                                   (sample bumpydist)))]

    ;; code to simulate the world
    world (create-world bumper-positions)
    end-world (simulate-world world)
    balls (:balls end-world)

    ;; how many balls entered the box?
    num-balls-in-box (balls-in-box end-world)

    obs-dist (normal 4 0.1)]

  (observe obs-dist num-balls-in-box))
```



3 examples generated from simulator
conditioned on ~20% of balls land in box

PROBABILISTIC PROGRAMMING EXAMPLE

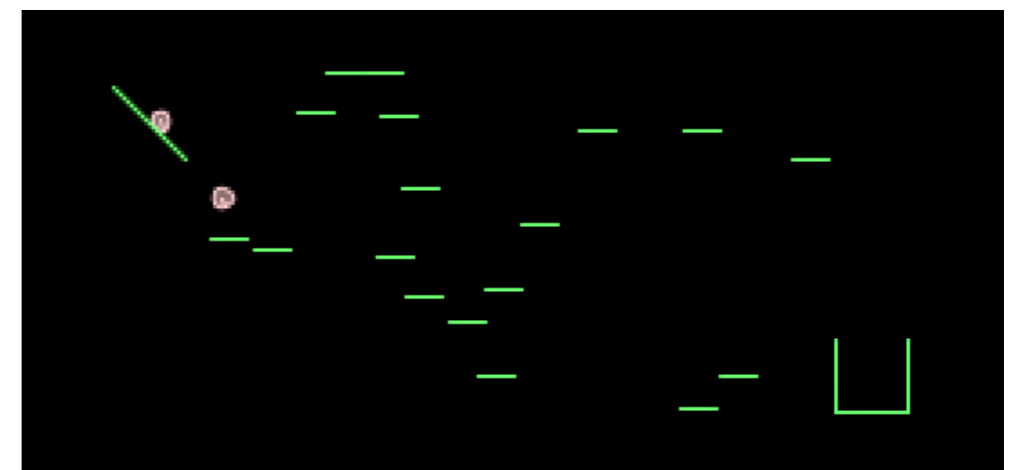
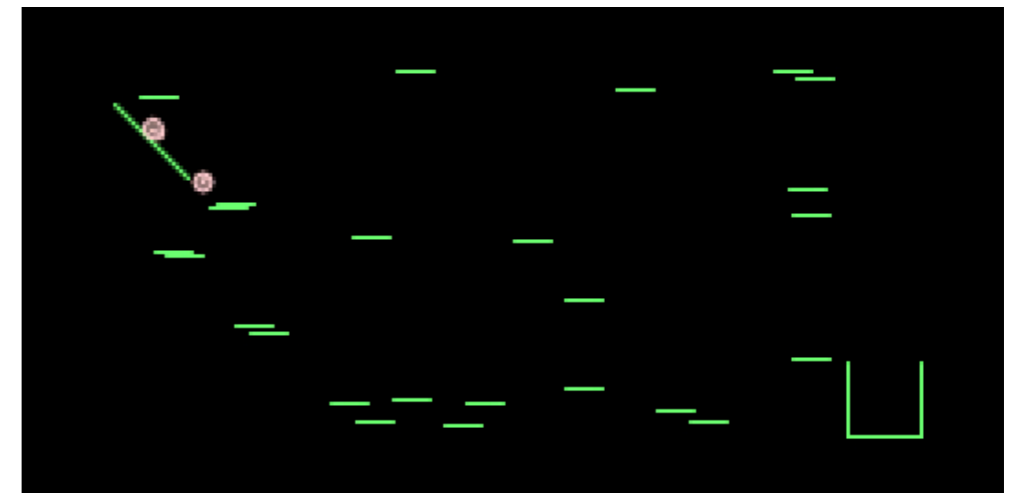
```
(defquery arrange-bumpers []
  (let [number-of-bumpers (sample (poisson 20))
        bumpydist (uniform-continuous 0 10)
        bumpxdist (uniform-continuous -5 14)
        bumper-positions (repeatedly
                          number-of-bumpers
                          #(vector (sample bumpxdist)
                                  (sample bumpydist)))]

    ;; code to simulate the world
    world (create-world bumper-positions)
    end-world (simulate-world world)
    balls (:balls end-world)

    ;; how many balls entered the box?
    num-balls-in-box (balls-in-box end-world)

    obs-dist (normal 4 0.1)])

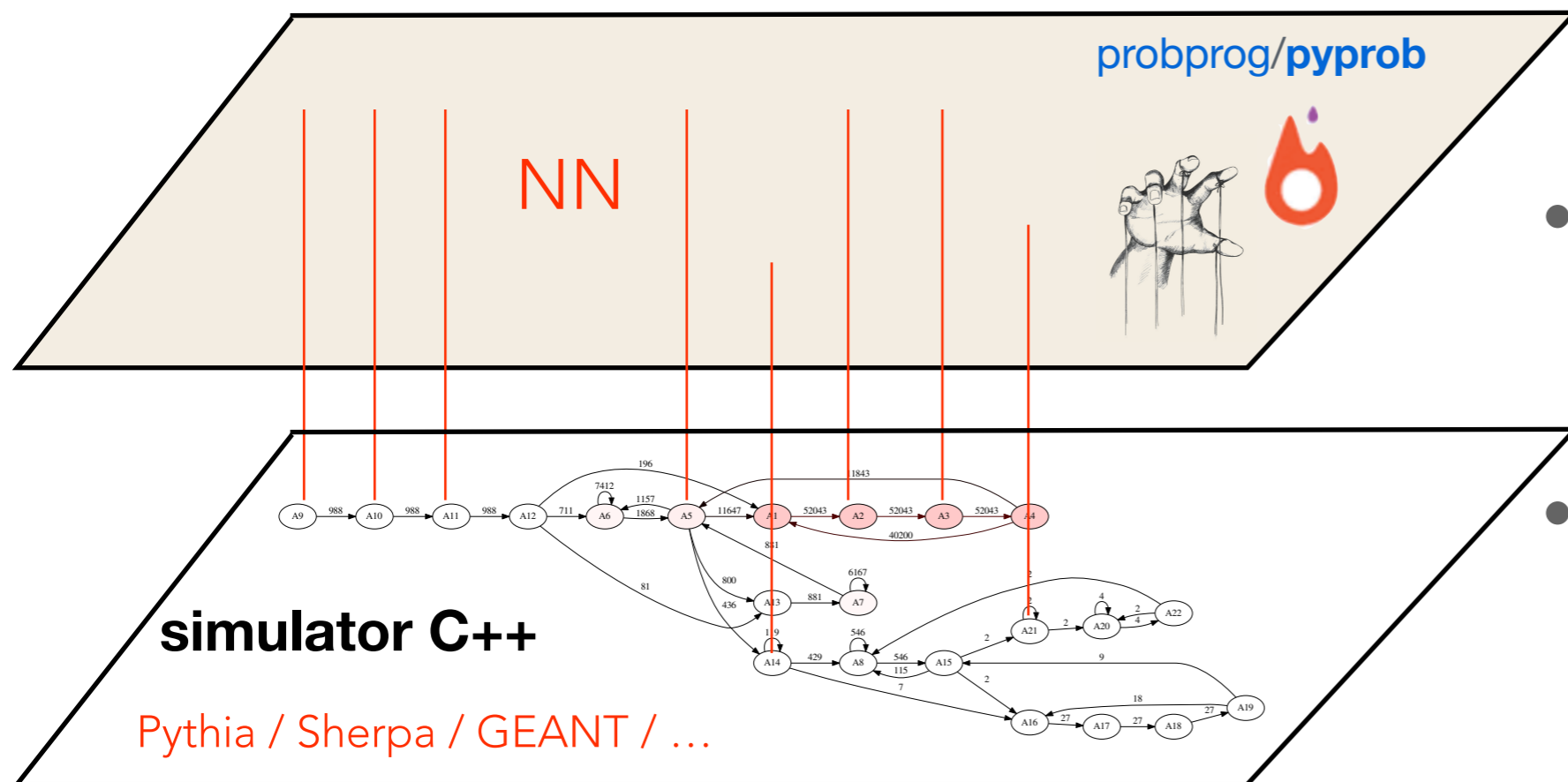
(observe obs-dist num-balls-in-box)
```



3 examples generated from simulator
conditioned on ~20% of balls land in box

PROBABILISTIC PROGRAMMING

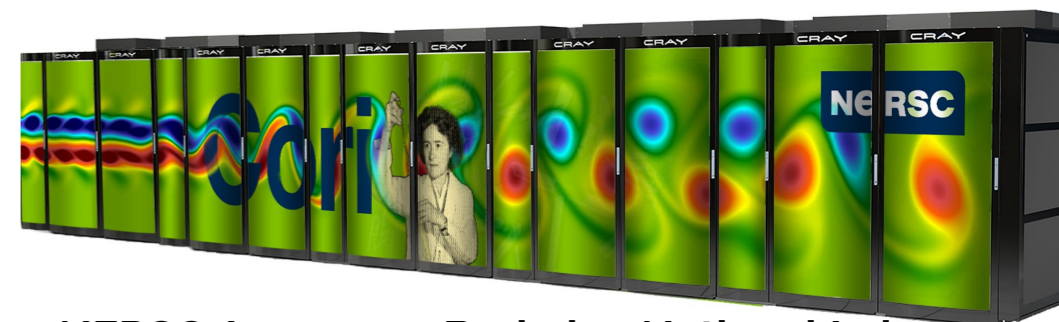
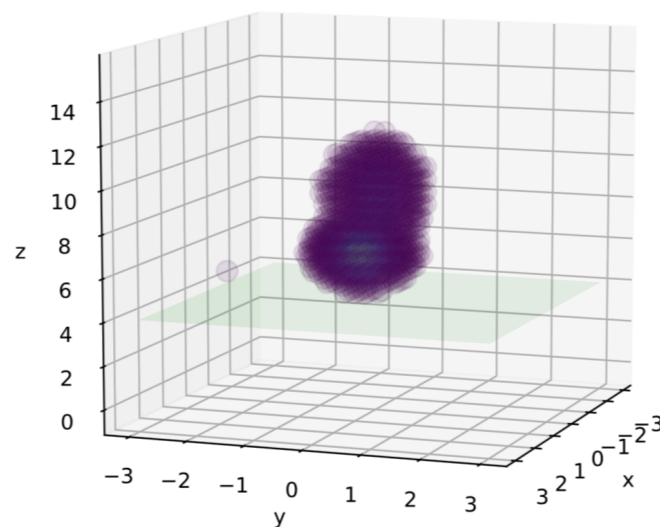
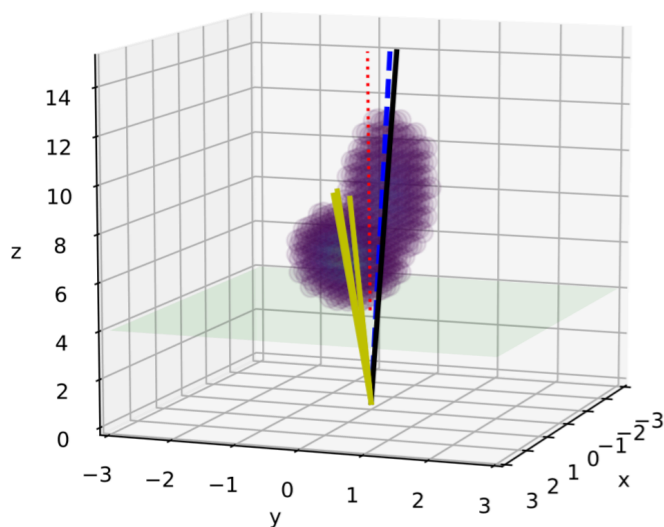
Idea: hijack the random number generators and perform a very fancy type of importance sampling or MCMC



- **Augment** real-world scientific simulator (C++)
- Use ML-powered inference engine

Observation

Mean Simulated Observation

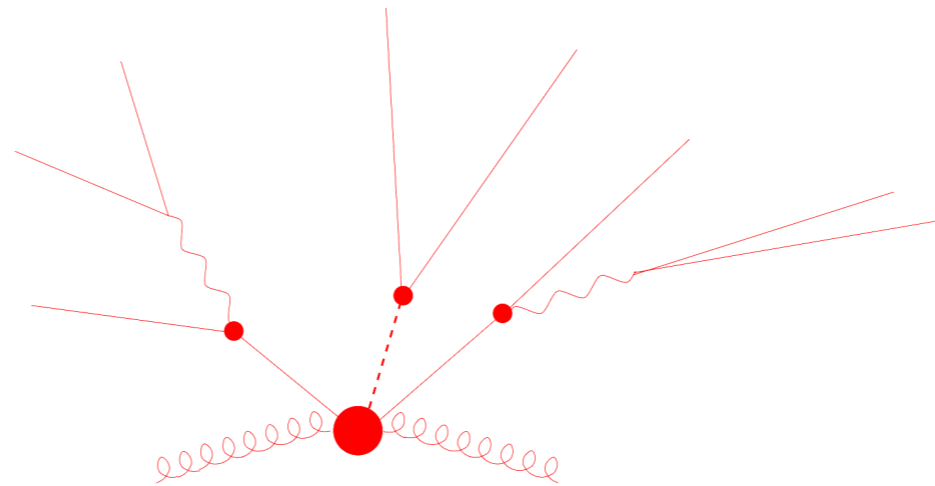


NERSC, Lawrence Berkeley National Lab

A PROBABILISTIC MODEL FOR JETS

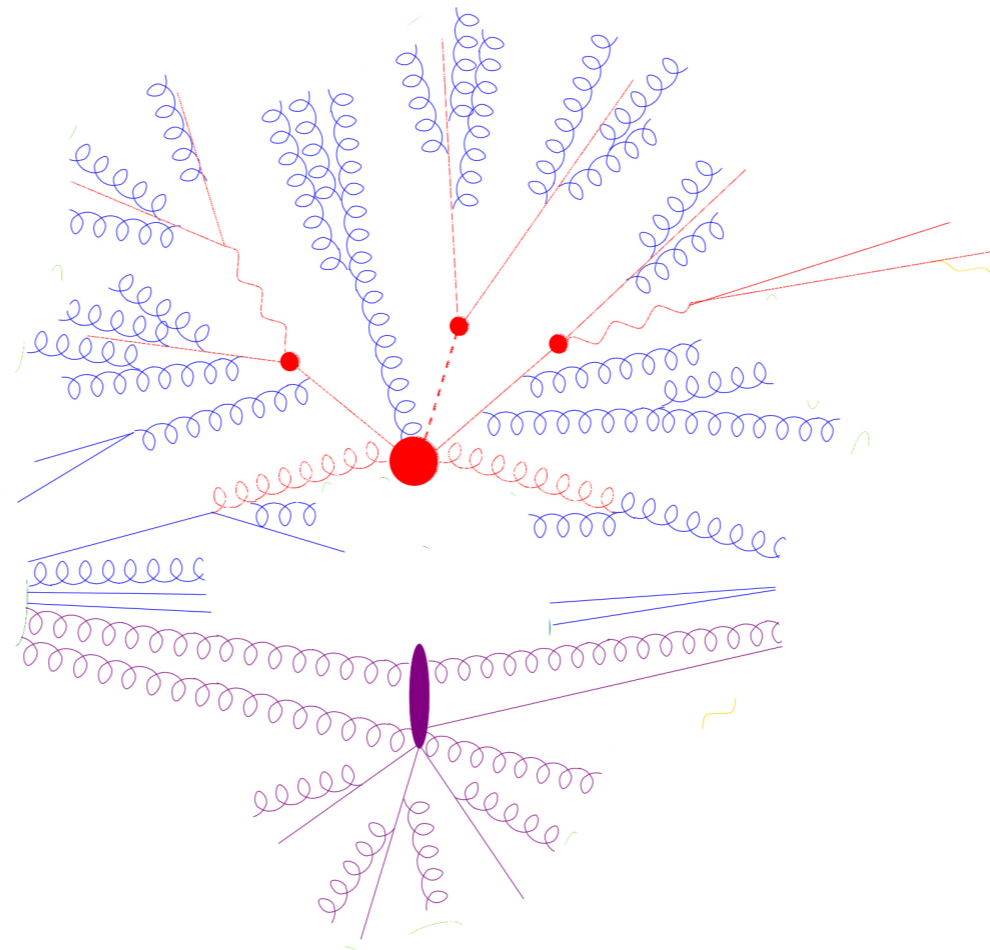
$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$

A PROBABILISTIC MODEL FOR JETS



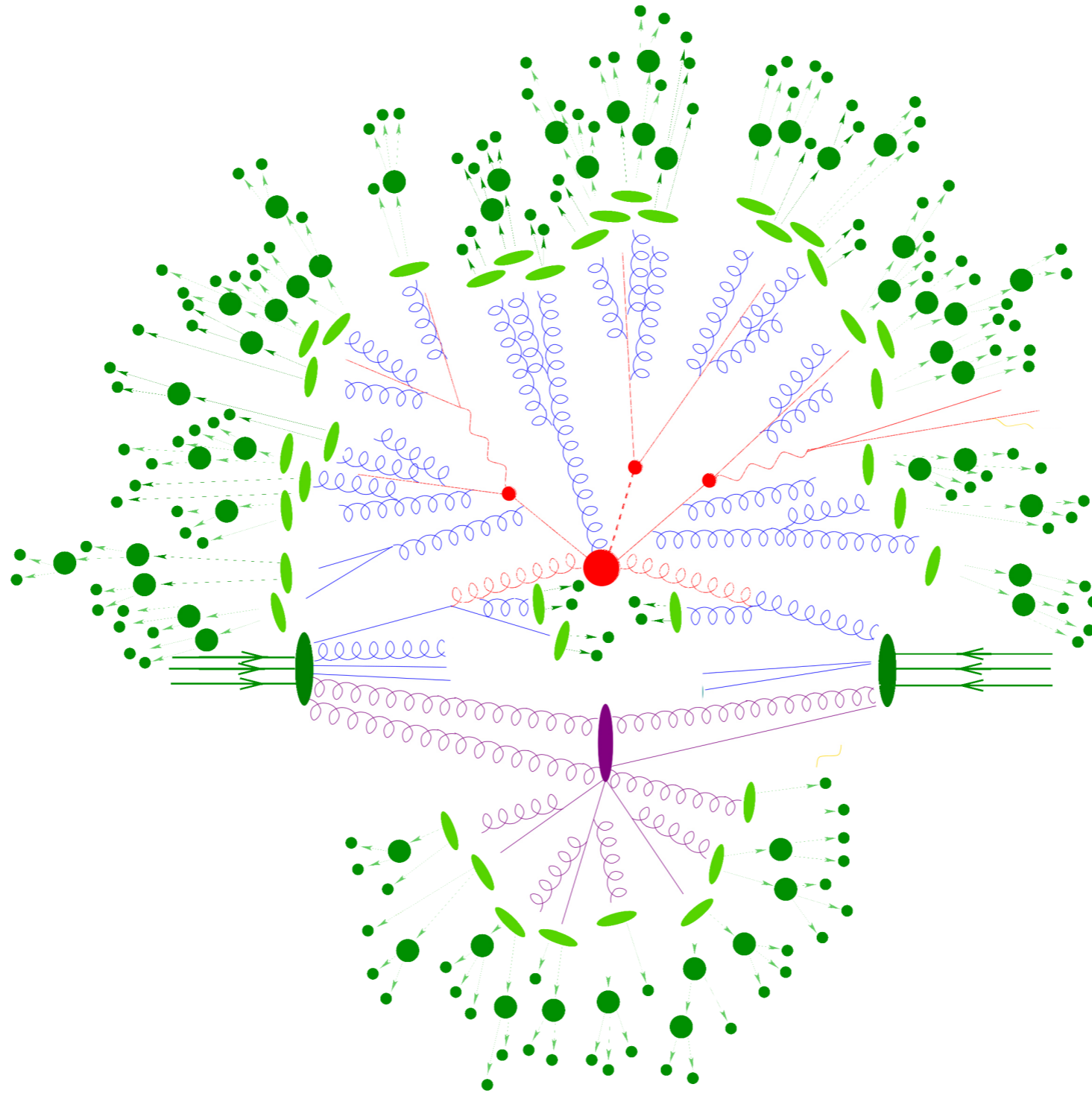
$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$

A PROBABILISTIC MODEL FOR JETS



$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$

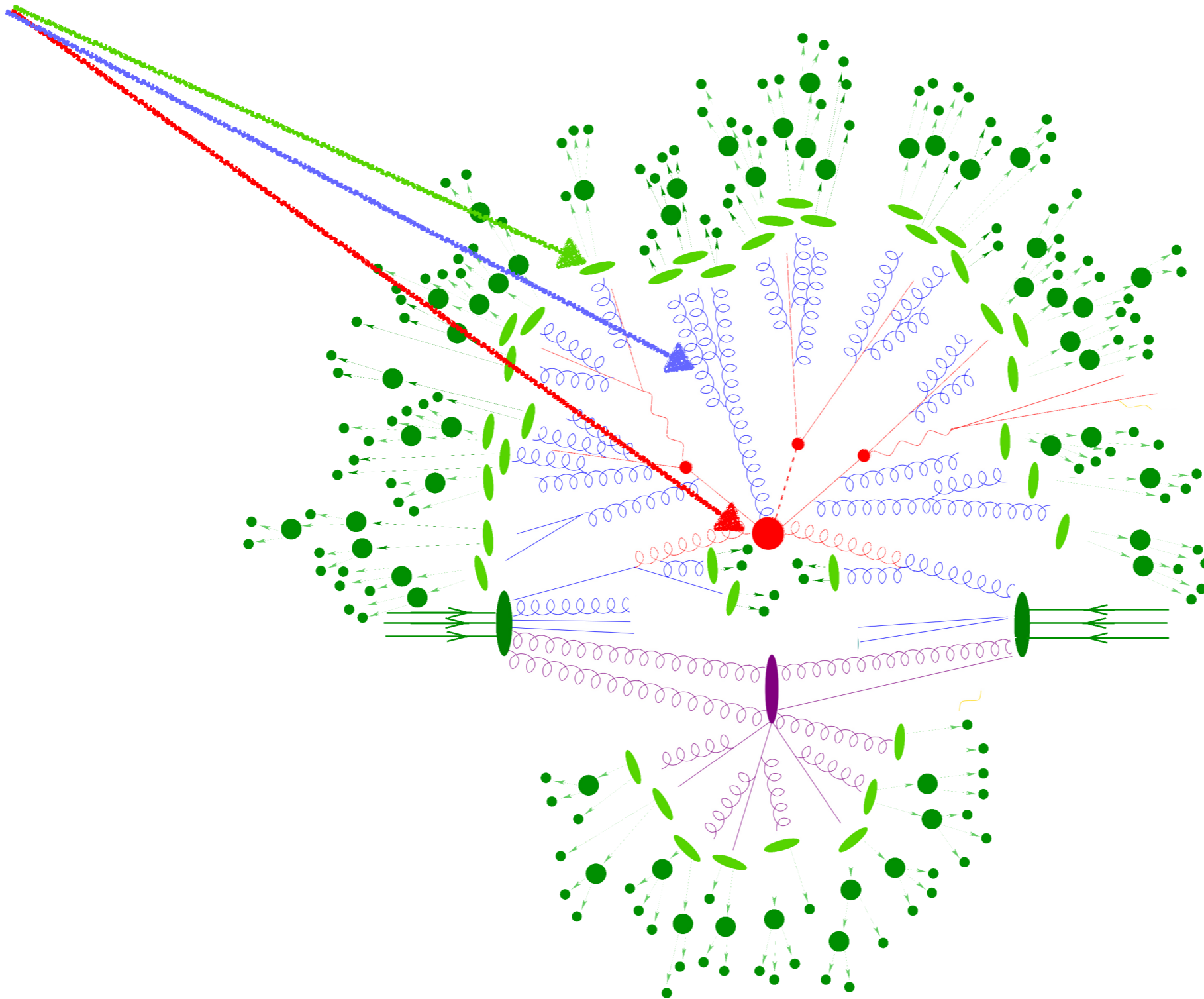
A PROBABILISTIC MODEL FOR JETS



$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$

A PROBABILISTIC MODEL FOR JETS

Evolution of the tree is **latent z**

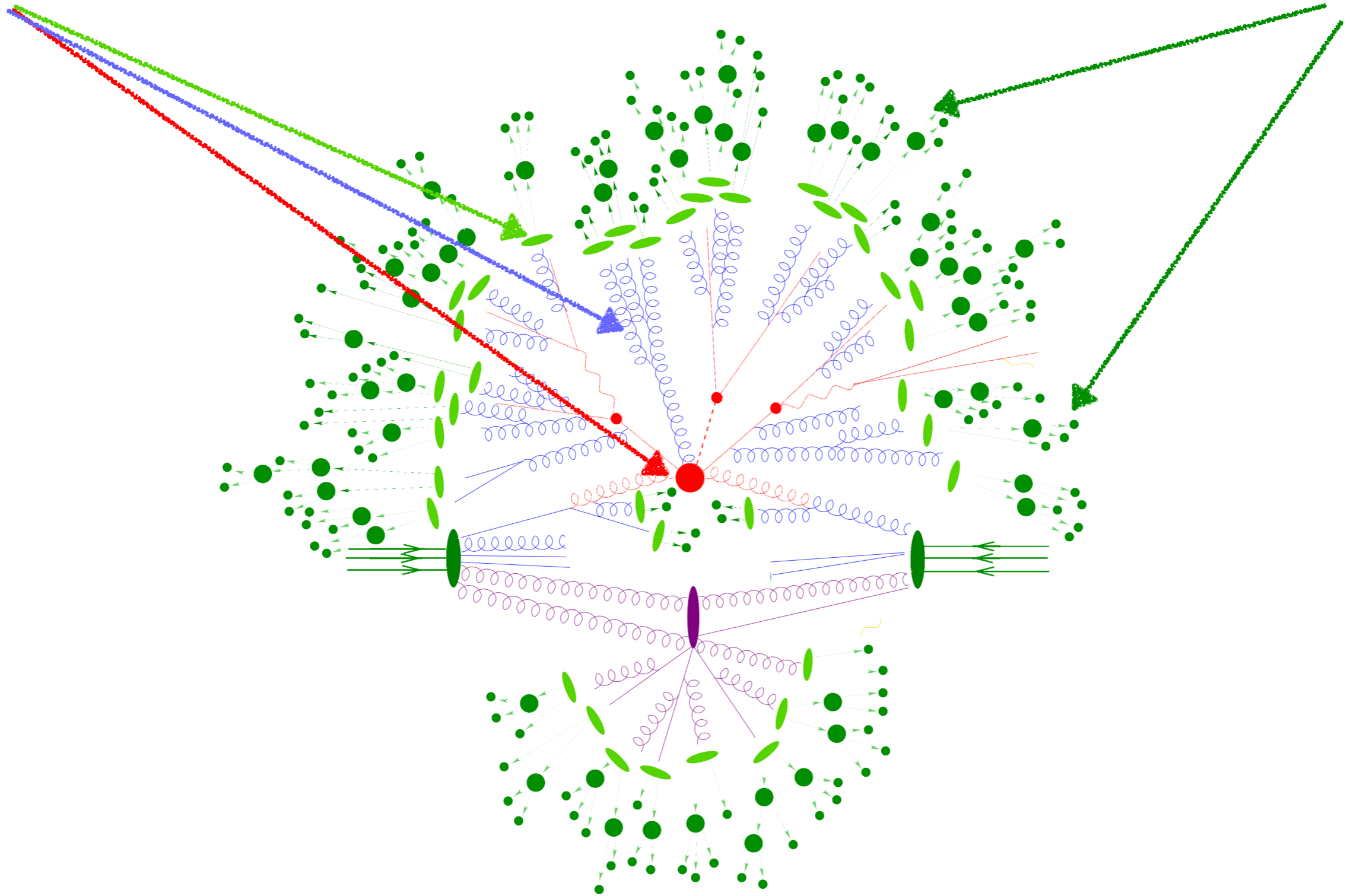


$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$

A PROBABILISTIC MODEL FOR JETS

Evolution of the tree is **latent z**

We only **observe** the **leaves x**



$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$

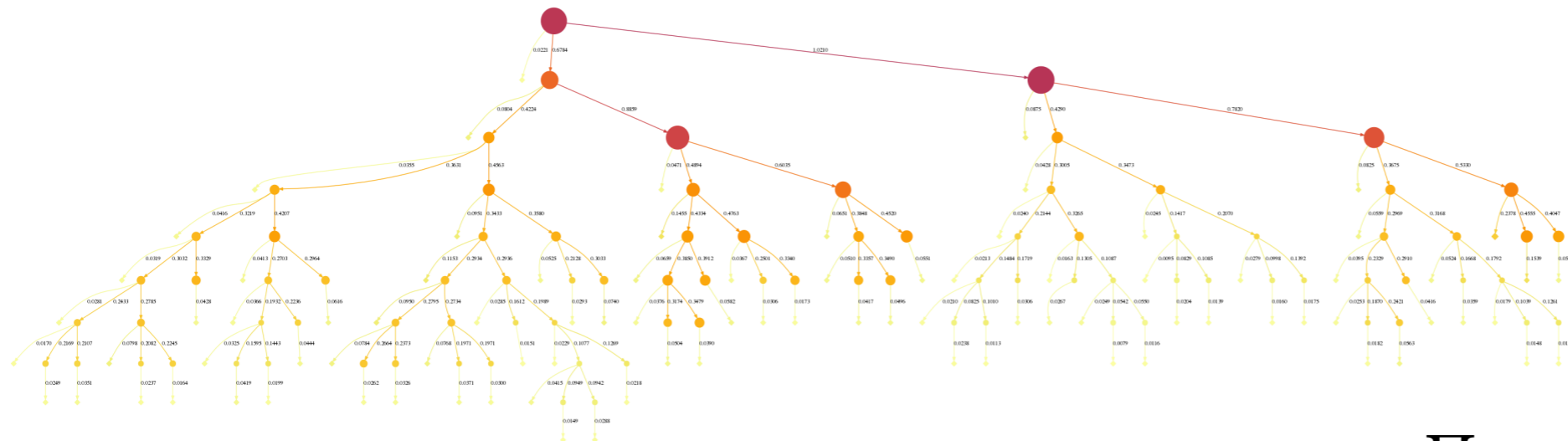
PROB. PROG. APPLIED TO JETS



Matt Dnevich

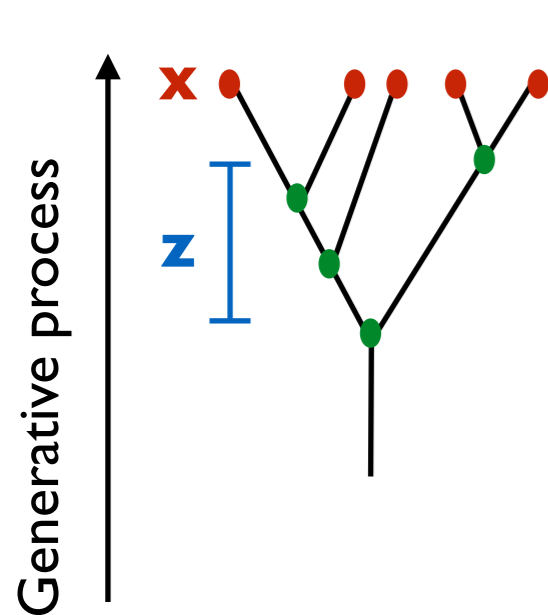
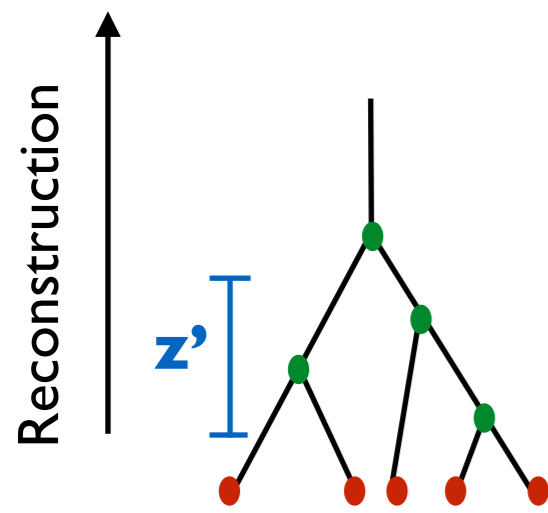


Sebastian Macaluso

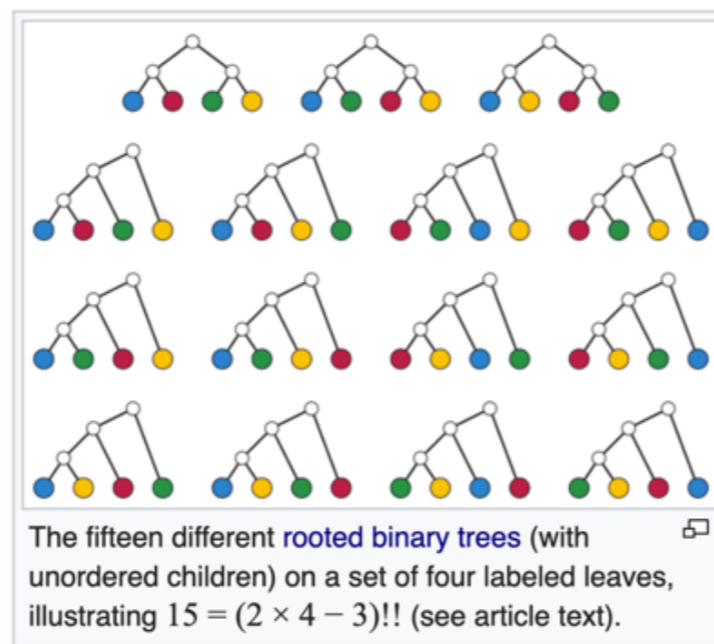


y=0, y_pred=0.1529

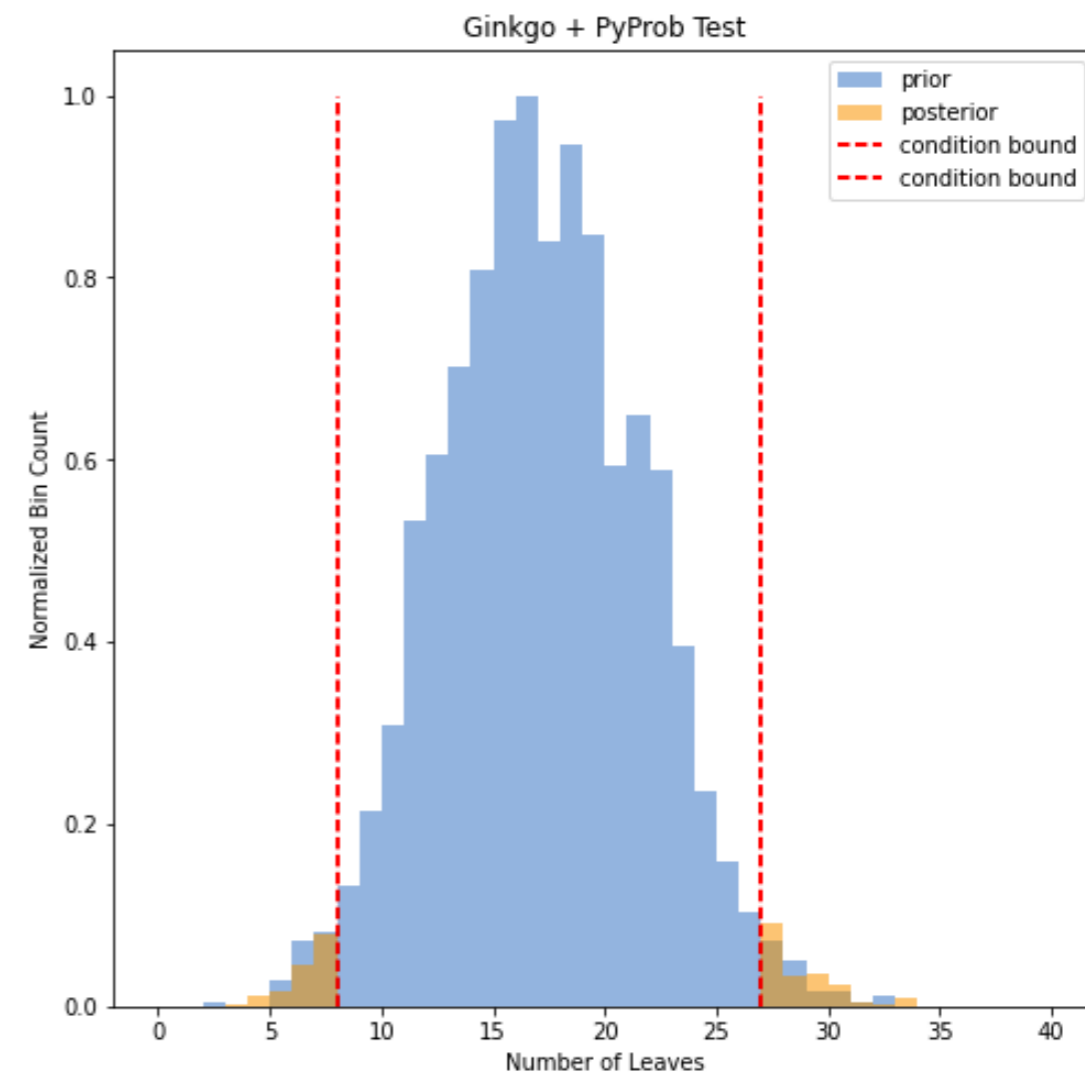
$$p(x, z|\theta) = \prod_j p(x_j|z_{\text{parent}(x_j)}, \theta) \prod_i p(z_i|z_{\text{parent}(z_i)}, \theta)$$



# of leaves	Approx. # of trees
4	15
5	100
7	10 k
9	2 M
11	600 M



https://en.wikipedia.org/wiki/Double_factorial



Simulation-Based Inference for Global Health Decisions

Christian Schroeder de Witt¹ Bradley Gram-Hansen¹ Nantas Nardelli¹
 Andrew Gambardella¹ Rob Zinkov¹ Puneet Dokania¹ N. Siddharth¹
 Ana Belen Espinosa-Gonzalez² Ara Darzi² Philip Torr¹ Atılım Güneş Baydin¹

Abstract

The COVID-19 pandemic has highlighted the importance of in-silico epidemiological modelling in predicting the dynamics of infectious diseases to inform health policy and decision makers about suitable prevention and containment strategies. Work in this setting involves solving challenging inference and control problems in individual-based models of ever increasing complexity. Here we discuss recent breakthroughs in machine learning, specifically in simulation-based inference, and explore its potential as a novel venue for model calibration to support the design and evaluation of public health interventions. To further stimulate research, we are developing software interfaces that turn two cornerstone COVID-19 and malaria epidemiology models (CovidSim¹ and OpenMalaria²) into probabilistic programs, enabling efficient interpretable Bayesian inference within those simulators.

1 Introduction

Machine learning has a growing role in increasing health service access and efficiency, particularly in resource-constrained settings, making it a valuable tool for the global health community [39, 54]. Moreover, the COVID-19 pandemic [55] has underlined the importance of epidemiological modelling and computer simulation in informing the design and implementation of public health interventions at an unprecedented scale [18]. For many endemic diseases (e.g., malaria), in-silico optimisation of multi-modal intervention portfolios—from mass vaccination to bed nets—is well established [47]. Analogous modelling for COVID-19 interventions, including social distancing [20], is mostly unexplored, yet subject to intense public interest [32].

The adoption of health informatics in worldwide health systems (e.g., OpenMRS [33], mHealth [1]) enables access to

¹Department of Engineering Science, University of Oxford, UK ²Department of Surgery and Cancer, Imperial College London, UK. Correspondence to: Christian Schroeder de Witt <cs@robots.ox.ac.uk>.

¹<https://github.com/mrc-ide/covid-sim/>
²<https://github.com/SwissTPH/openmalaria>

abundant patient-level and aggregated health data [54]. This is fomenting the development of comprehensive modelling and simulation to support the design of health interventions and policies, and to guide decision-making in a variety of health system domains [22, 49]. For example, simulations have provided valuable insight to deal with public health problems such as tobacco consumption in New Zealand [50], and diabetes and obesity in the US [58]. They have been used to explore policy options such as those in maternal and antenatal care in Uganda [44], and applied to evaluate health reform scenarios such as predicting changes in access to primary care services in Portugal [21]. Their applicability in informing the design of cancer screening programmes has been also discussed [42, 23]. Recently, simulations have informed the response to the COVID-19 outbreak [19].

The process of informing health interventions and policies through simulations generally involves two steps:

Model calibration The extent to which a simulator can reliably inform real-world prediction and planning is bounded by both model discrepancy [13] and how well the model has been calibrated to empirical data [3].

Optimising decision-making Identifying optimal multi-modal intervention strategies and corresponding risks and uncertainties requires searching through potentially vast parameter spaces, which, due to the computational cost of running large simulators (e.g., in some epidemiological studies), usually cannot be exhaustively evaluated [46].

Despite their fundamental importance, model discrepancy and calibration of public-health simulators are frequently only informally addressed, or left undocumented [48, 40]. This may be partially explained by the fact that, while numerous methods for formal sensitivity and uncertainty analysis exist [28], they in general do not scale to complex simulators with more than a few dozen parameters [38]. Similarly, evidence-based decision-making is usually optimised by comparing outcomes on a small number of hand-crafted scenarios and intervention strategies [46].

2 Epidemiology simulations and inference

Among the simplest mathematical epidemiology models are deterministic *compartmental models* that partition individu-

Hijacking Malaria Simulators with Probabilistic Programming

Bradley J. Gram-Hansen^{*1} Christian Schröder de Witt^{*1}
 Tom Rainforth² Philip H.S. Torr¹ Yee Whye Teh² Atılım Güneş Baydin¹

Abstract

Epidemiology simulations have become a fundamental tool in the fight against the epidemics of various infectious diseases like AIDS and malaria. However, the complicated and stochastic nature of these simulators can mean their output is difficult to interpret, which reduces their usefulness to policymakers. In this paper, we introduce an approach that allows one to treat a large class of population-based epidemiology simulators as probabilistic generative models. This is achieved by *hijacking* the internal random number generator calls, through the use of an universal probabilistic programming system (PPS). In contrast to other methods, our approach can be easily retrofitted to simulators written in popular industrial programming frameworks. We demonstrate that our method can be used for interpretable introspection and inference, thus shedding light on black-box simulators. This reinstates much needed trust between policymakers and evidence-based methods.

1. Introduction

Ending the epidemics of AIDS, tuberculosis, malaria and other infectious diseases by 2030 is a key target within the Good Health & Well-Being section of the UN Sustainable Development Goals (UN, 2017; 2018). However, despite decades of substantial international efforts, these diseases kill hundreds of million people a year. For example, malaria still annually kills about a quarter of a million children under the age of 5 in Africa alone.

To reach the WHO’s target of reducing malaria incidence and mortality rates by at least 90% by 2030, policymakers are increasingly turning to evidence-based methods, thus oftentimes relying on computational simulations (WHO,

^{*}Equal contribution ¹Department of Engineering Science, University of Oxford, UK ²Department of Statistics, University of Oxford, UK. Correspondence to: Bradley J. Gram-Hansen <bradley@robots.ox.ac.uk>.

Appearing at the *International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, United States, 2019.

2015). These simulations allow policymakers to infer critical information on disease dynamics and make predictions about the impacts of policies before they are rolled out. This frequently increases the effectiveness of interventions and thus ultimately saves resources, or even lives. For example, it has been shown that mass vaccination may be largely ineffective in regions of large transmission rates, but may play a crucial role in areas of low transmission (Cameron et al., 2015).

Malaria epidemiology is governed by a complex set of drivers, few of which can be understood in isolation (Cameron et al., 2015; Autino et al.; Smith et al., 2008; Bershteyn et al., 2018). These include within-host dynamics, population-specific traits and even local geography. Comprehensive modeling of all of these components remains challenging, particularly in a region-specific context. Computational epidemiology simulators have to reflect these complexities and are usually stochastic in nature. This can make simulation output highly non-trivial to interpret, particularly when trying to draw desired inferences coupled with observed data (Mwendera et al.; Ferris et al.).

In this paper, we introduce a novel method that allows one to shed light on the inner workings of a large class of population-based stochastic simulators. We achieve this by extending the work of Baydin et al. (2018) by interpreting such population-based simulators as probabilistic generative models within the framework of universal probabilistic programming (UPP) (Le et al., 2017). To this end, we *hijack* existing simulators by overriding their internal random number generators. Specifically, by replacing the existing low-level random number generator in a simulator with a call to a purpose-built UPP “controller”, which can thus control, track and manipulate the stochasticity of the simulator.

This allows for a variety of tasks to be performed on the hijacked simulator, such as running inference (by conditioning the values of certain draws and manipulating others), uncovering stochastic structure, and automatically producing result summaries, such as establishing the probability of different program paths/traces. By providing a common abstraction framework for different simulators, our approach further allows for easy and direct comparison between re-

DISCUSSION

Inference is always done within the context of a model

- If the model is mis-specified it will affect inference
- Here the model is the simulator, or the surrogate for the simulator
 - **One hand:** simulators usually include more effects than traditional approaches
 - **Other hand:** more chances for method to focus on aspects that are poorly modeled

Humans are good at designing robust summary statistics that are not sensitive to mis-modeled features in the data

- Now there are numerous approaches to build this into the training of ML models (related to domain adaptation, algorithmic fairness, pivotal quantities, profiling, etc.) eg. uBoost by J. Stevens, M. Williams, "learning to pivot" by KC, Louppe, Kagan

These methods **do not address hypothesis generation.**

- They are not designed to discover new laws of nature.

CONCLUSIONS

Machine Learning can help us get more out of our simulators

- it can provide effective statistical models for emergent phenomena that are tied back to the low-level microscopic reductionist model

Our understanding of how to leverage our prior physics knowledge while letting machine learning do what it's good at is maturing.

- build in robustness to systematic uncertainties
- ability to inject and extract physics knowledge from models
- exploit symmetries, hierarchical structure of data

Harnessing the full potential of these techniques requires augmenting existing simulators.



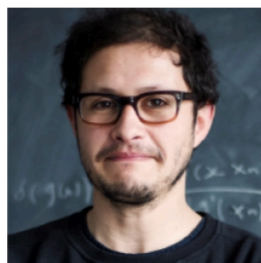
Machine Learning and the Physical Sciences

Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS)

December 11, 2020



Atılım Güneş Baydin
University of Oxford



Juan Felipe Carrasquilla
Vector Institute /
University of Waterloo



Adji Bousso Dieng
Columbia University



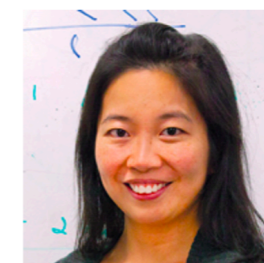
Karthik Kashinath
NERSC, Berkeley Lab



Anima Anandkumar
Caltech / NVIDIA



Kyle Cranmer
New York University



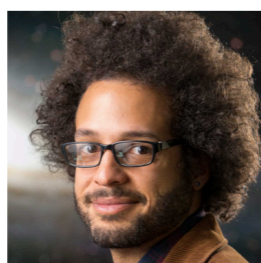
Shirley Ho
Flatiron / Princeton /
Carnegie Mellon



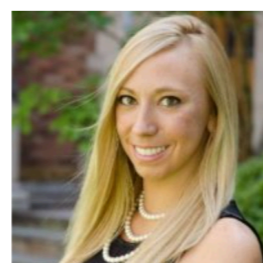
Prabhat
NERSC, Berkeley Lab



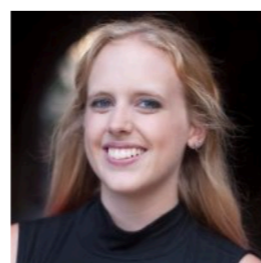
Gilles Louppe
University of Liège



Brian Nord
Fermilab



Michela Paganini
Facebook AI Research



Savannah Thais
Princeton University /
IRIS-HEP



Lenka Zdeborova
Institut de Physique
Théorique

References

Information Geometry for Higgs EFT (no machine learning)

Brehmer, Kling, Plehn, Cranmer

“Better Higgs Measurements Through Information Geometry”

[PRD, 1612.05261]

Brehmer, Kling, Plehn, Tait

“Better Higgs-CP Tests Through Information Geometry”

[PRD, 1712.02350]

Core References for ML-based approach

JB, K. Cranmer, G. Louppe, J. Pavez:

“Constraining Effective Field Theories with machine learning”

[PRL, 1805.00013]

JB, K. Cranmer, G. Louppe, J. Pavez:

“A guide to constraining Effective Field Theories with machine learning”

[PRD, 1805.00020]

JB, G. Louppe, J. Pavez, K. Cranmer:

“Mining gold from implicit models to improve likelihood-free inference”

[PNAS, 1805.12244]

Follow-up with incremental improvements

M. Stoye, JB, K. Cranmer, G. Louppe, J. Pavez:

“Likelihood-free inference with an improved cross-entropy estimator”

[NeurIPS workshop, 1808.00973]

Opinionated review of simulation-based inference

K. Cranmer, JB, G. Louppe:

“The frontier of simulation-based inference”

[1911.01429]

Do It Yourself (for LHC physics)

JB, F. Kling, I. Espejo, K. Cranmer:

“MadMiner: Machine learning—based inference for particle physics”

[CSBS, 1907.10621, <https://github.com/diana-hep/madminer>]

Strong lensing

JB, S. Mishra-Sharma, J. Hermans, G. Louppe, K. Cranmer

“Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning”

[ApJ, 1909.02005]

LHC HXSWG YR4 STXS

JB, S. Dawson, S. Homiller, F. Kling, T. Plehn:

“Benchmarking simplified template cross sections in WH production”

[JHEP, 1908.06980]

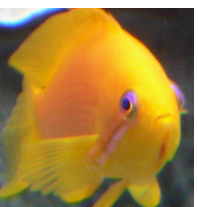
ACKNOWLEDGEMENTS



Johann Brehmer Gilles Louppe Juan Pavez Markus Stoye Felix Kling Irina Espejo Tilman Plehn Sam Homiller Sally Dawson Sid Mishra-Sharma Zubair Bhatti



The SCALFIN Project
scalfn.github.io



Dark Matter Examples

SMALL-SCALE STRUCTURE: A PROBE OF DM PARTICLE PROPERTIES



Sid Mishra-Sharma



Johann Brehmer

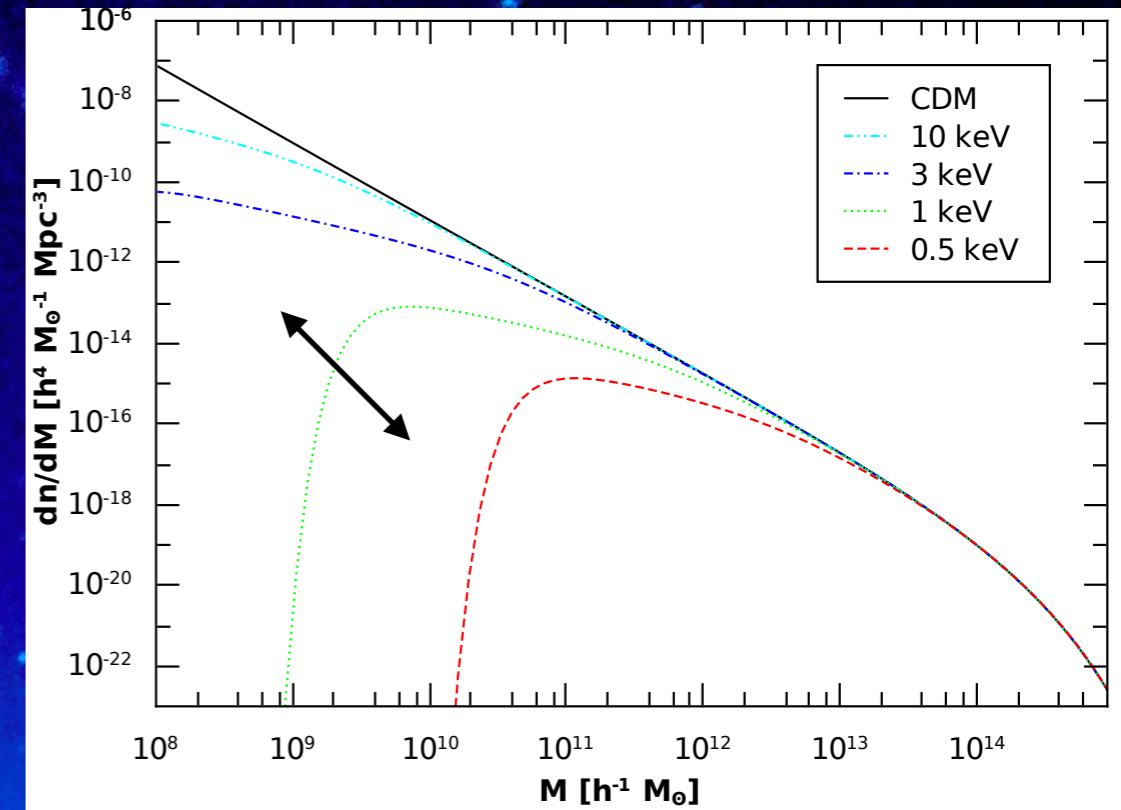


Gilles Louppe

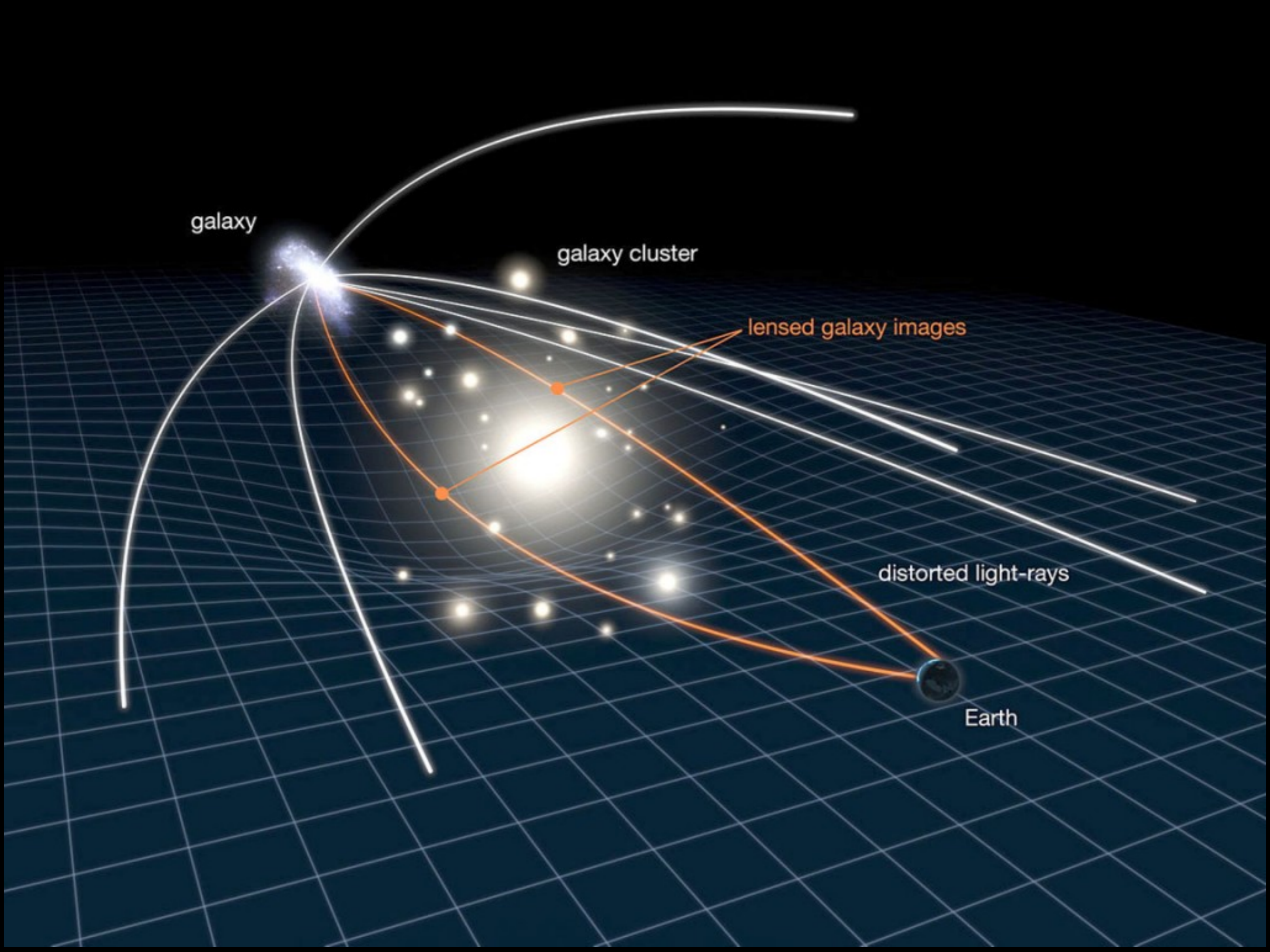


Joeri Hermans

Abundance of DM subhalos vs mass:



[R. Dunstan et al 1109.6291]



galaxy

galaxy cluster

lensed galaxy images

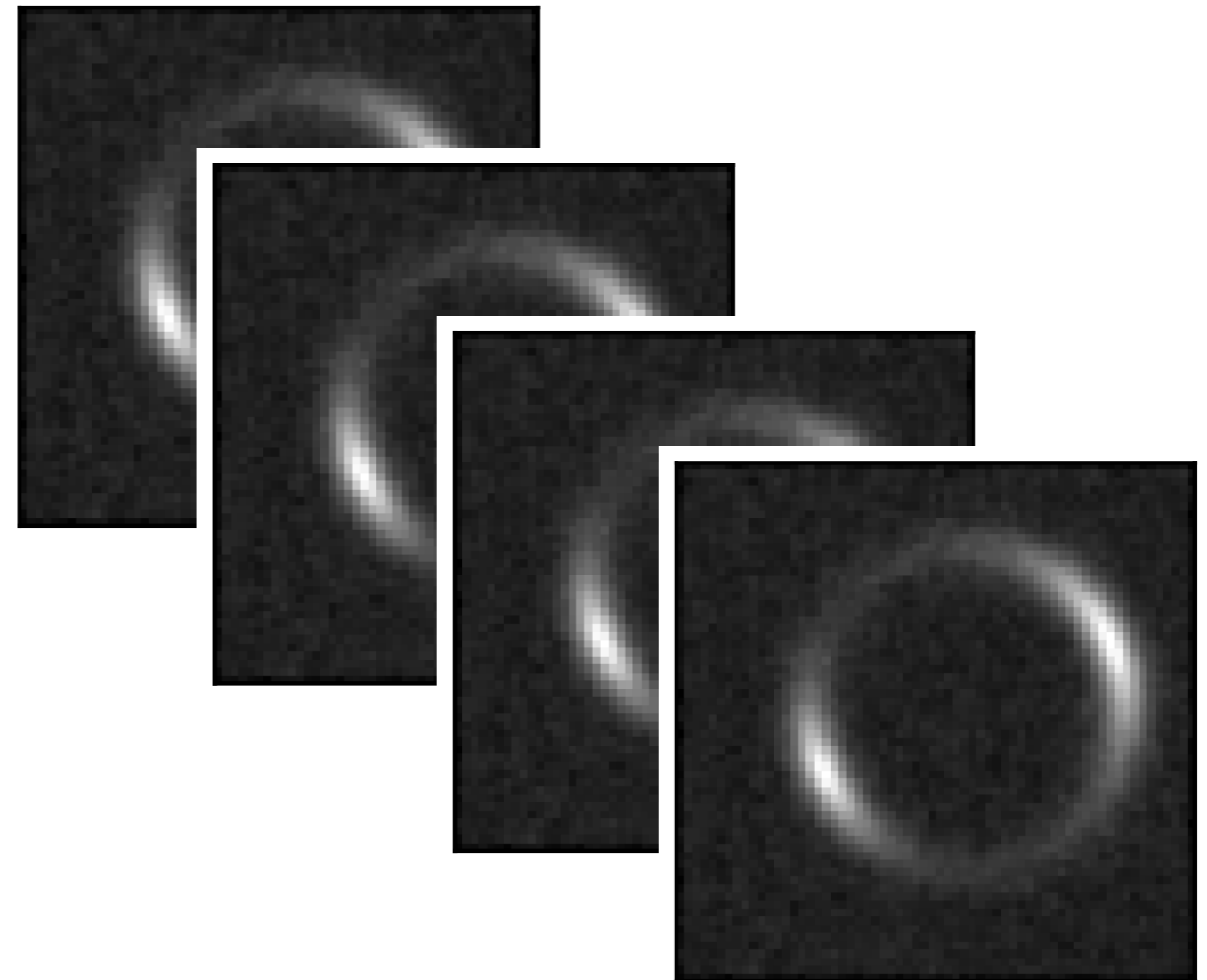
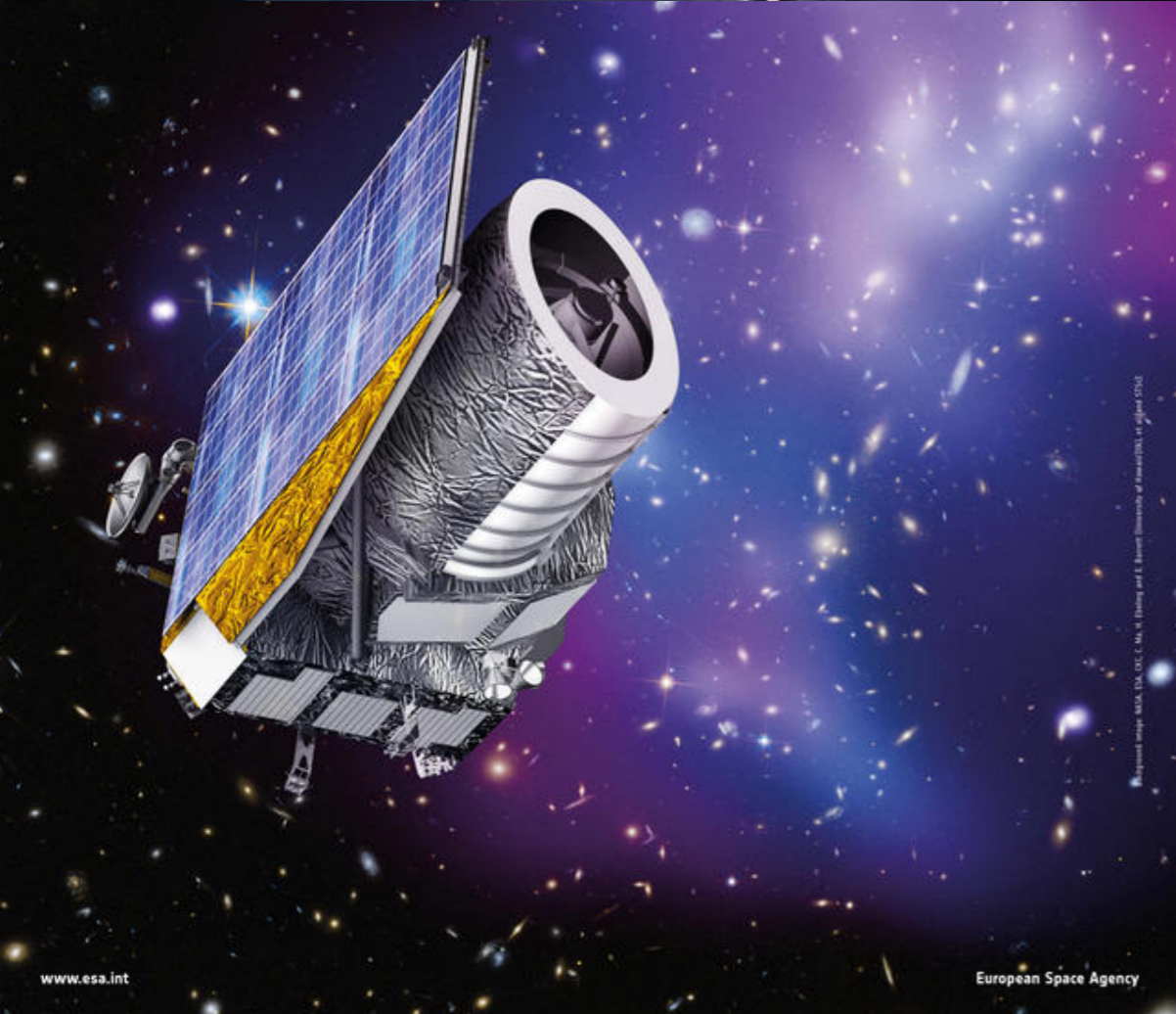
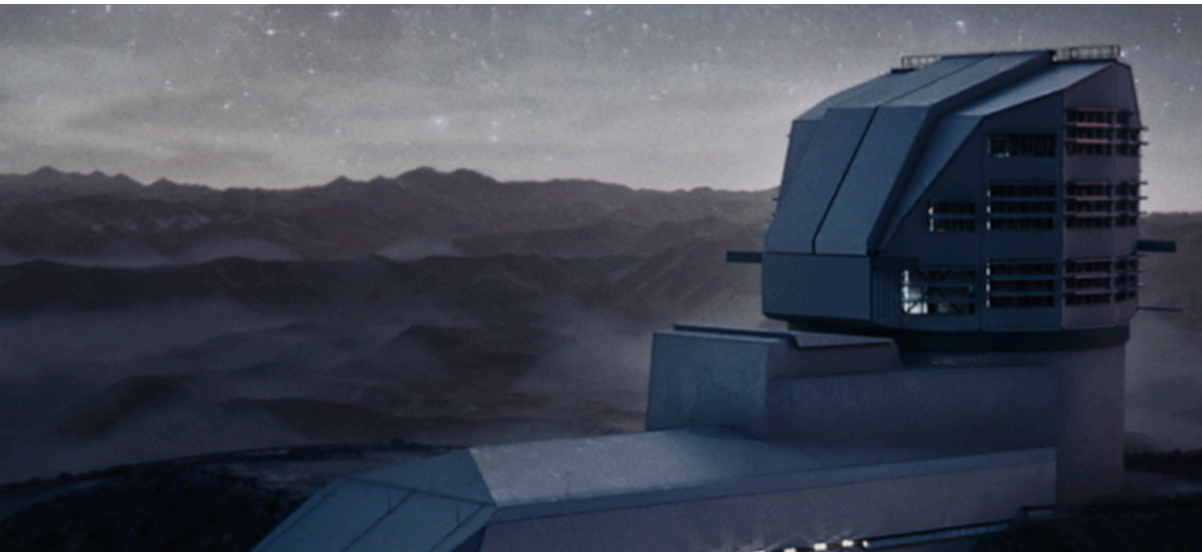
distorted light-rays

Earth



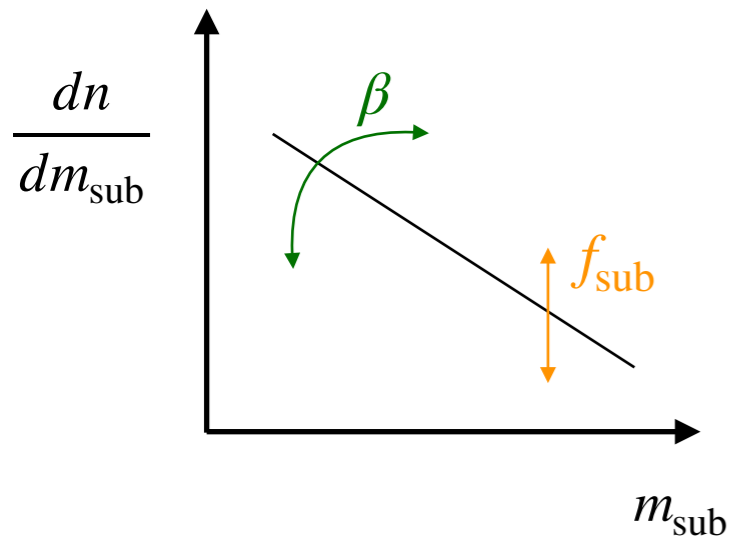
SCALABLE INFERENCE FOR SMALL SUBHALOS

Future surveys (LSST, Euclid) are expected to deliver large samples of galaxy-galaxy strong lenses [Collett et al 1507.02657]



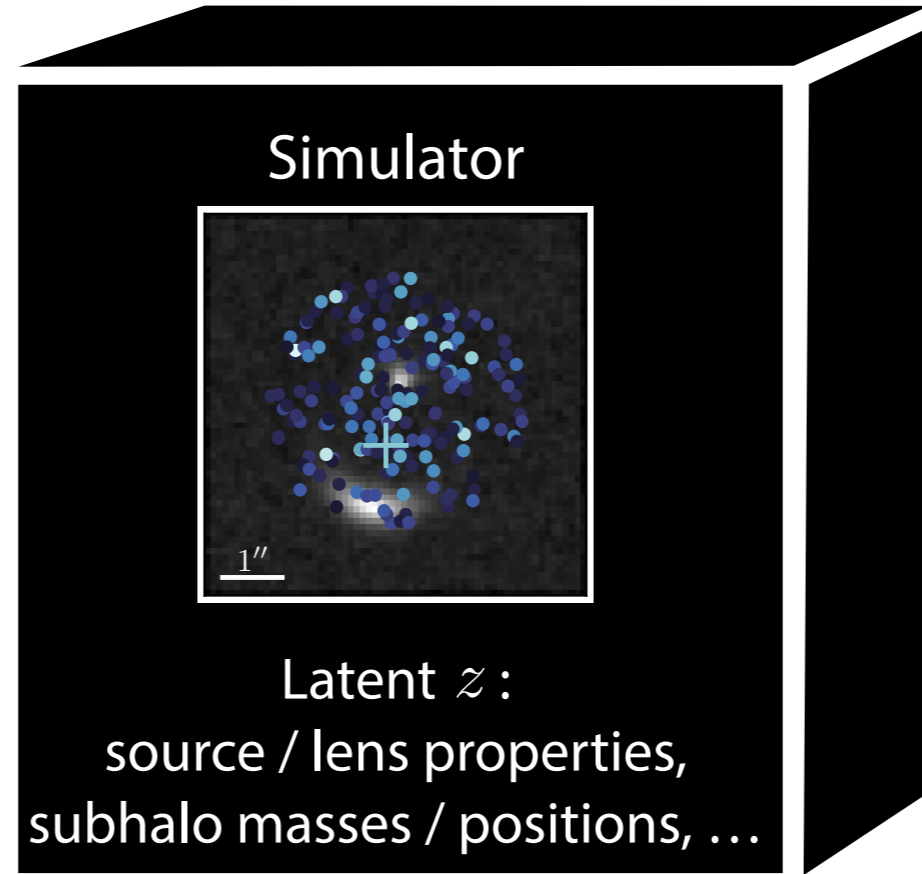
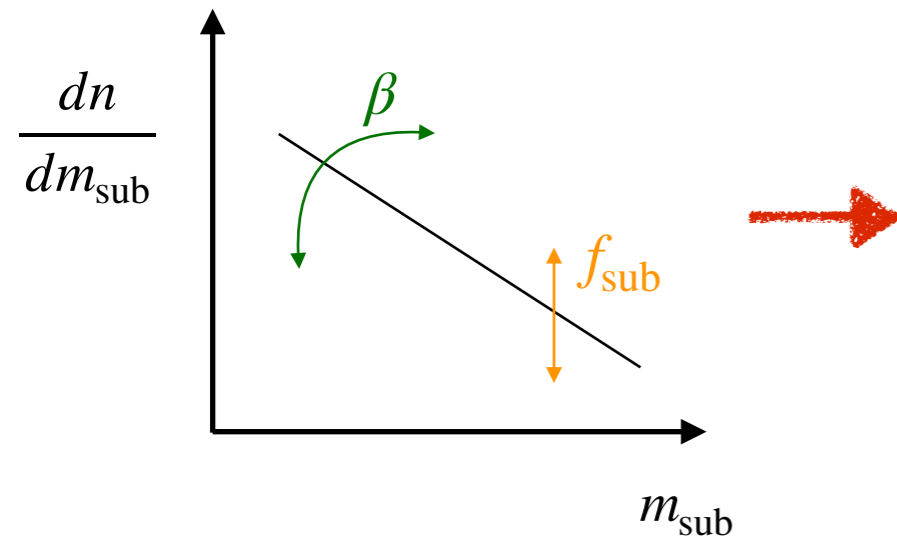
SIMULATION-BASED INFERENCE FOR STRONG LENSING

2 parameters $\theta = (\beta, f_{\text{sub}})$

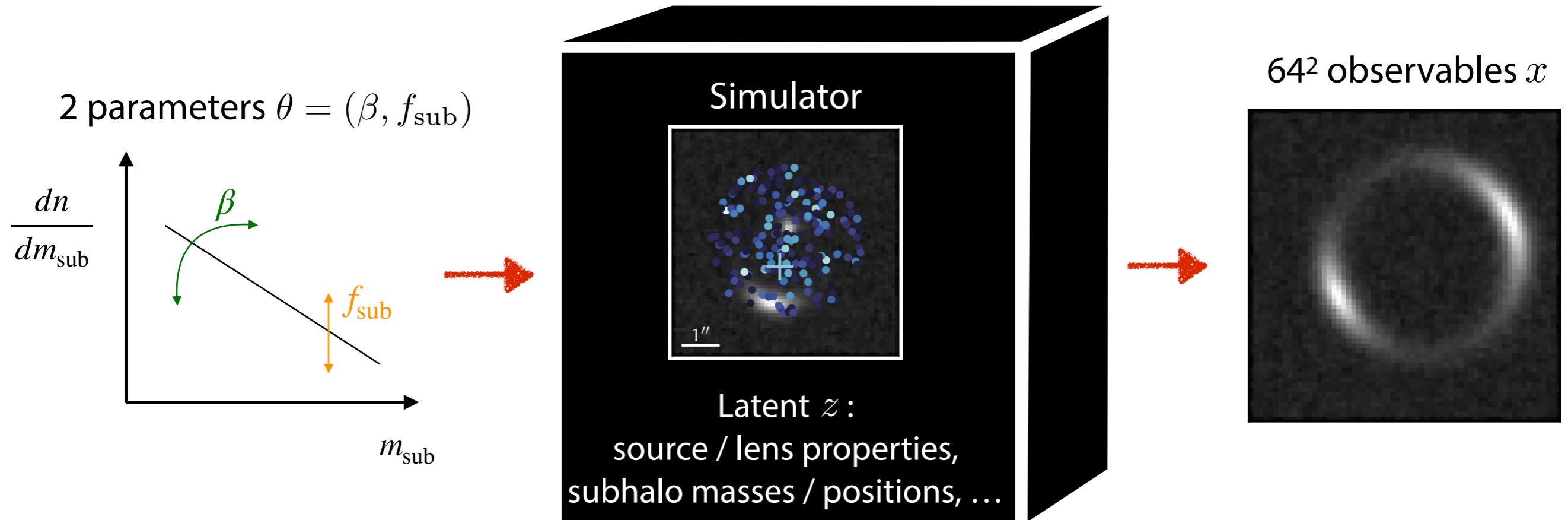


SIMULATION-BASED INFERENCE FOR STRONG LENSING

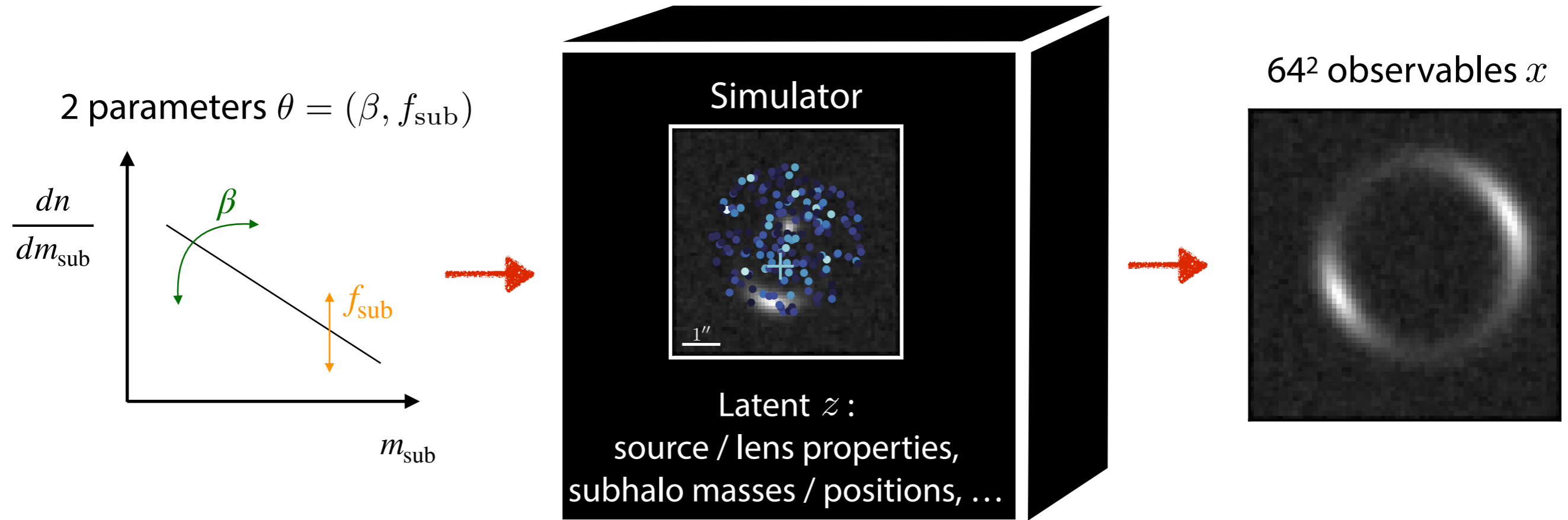
2 parameters $\theta = (\beta, f_{\text{sub}})$



SIMULATION-BASED INFERENCE FOR STRONG LENSING

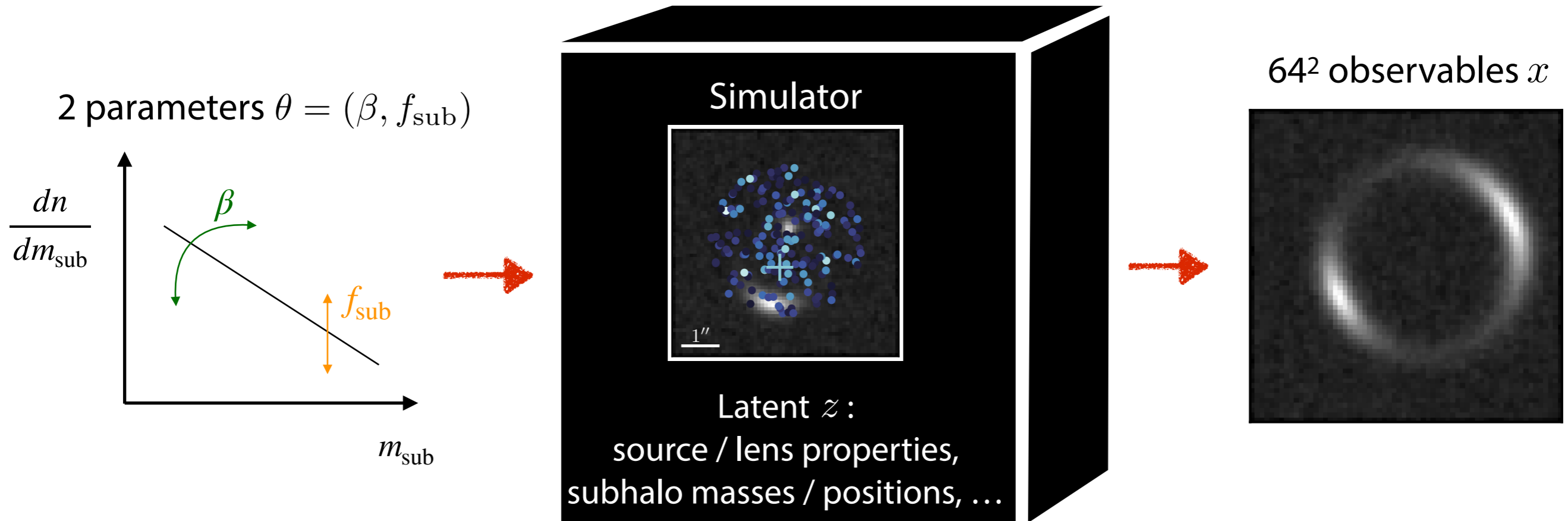


SIMULATION-BASED INFERENCE FOR STRONG LENSING



Prediction: We construct a simulator that can sample $x \sim p(x|\theta)$

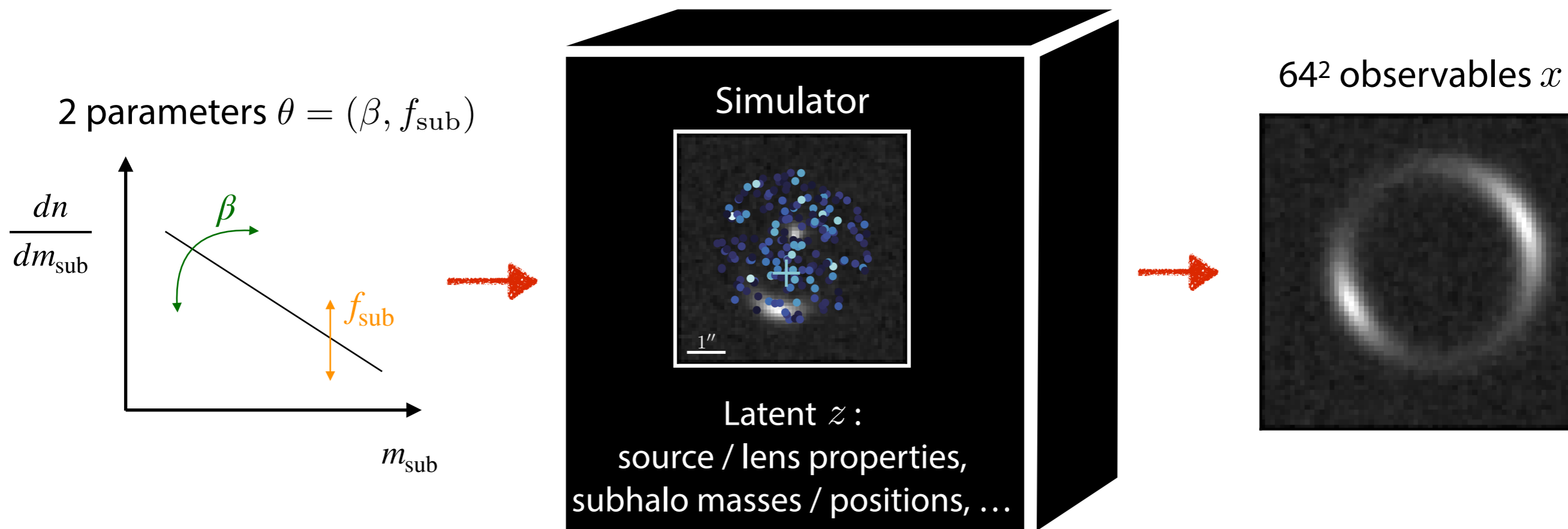
SIMULATION-BASED INFERENCE FOR STRONG LENSING



Prediction: We construct a simulator that can sample $x \sim p(x|\theta)$

Inference: We train neural likelihood ratio estimators $\hat{r}(x|\theta)$

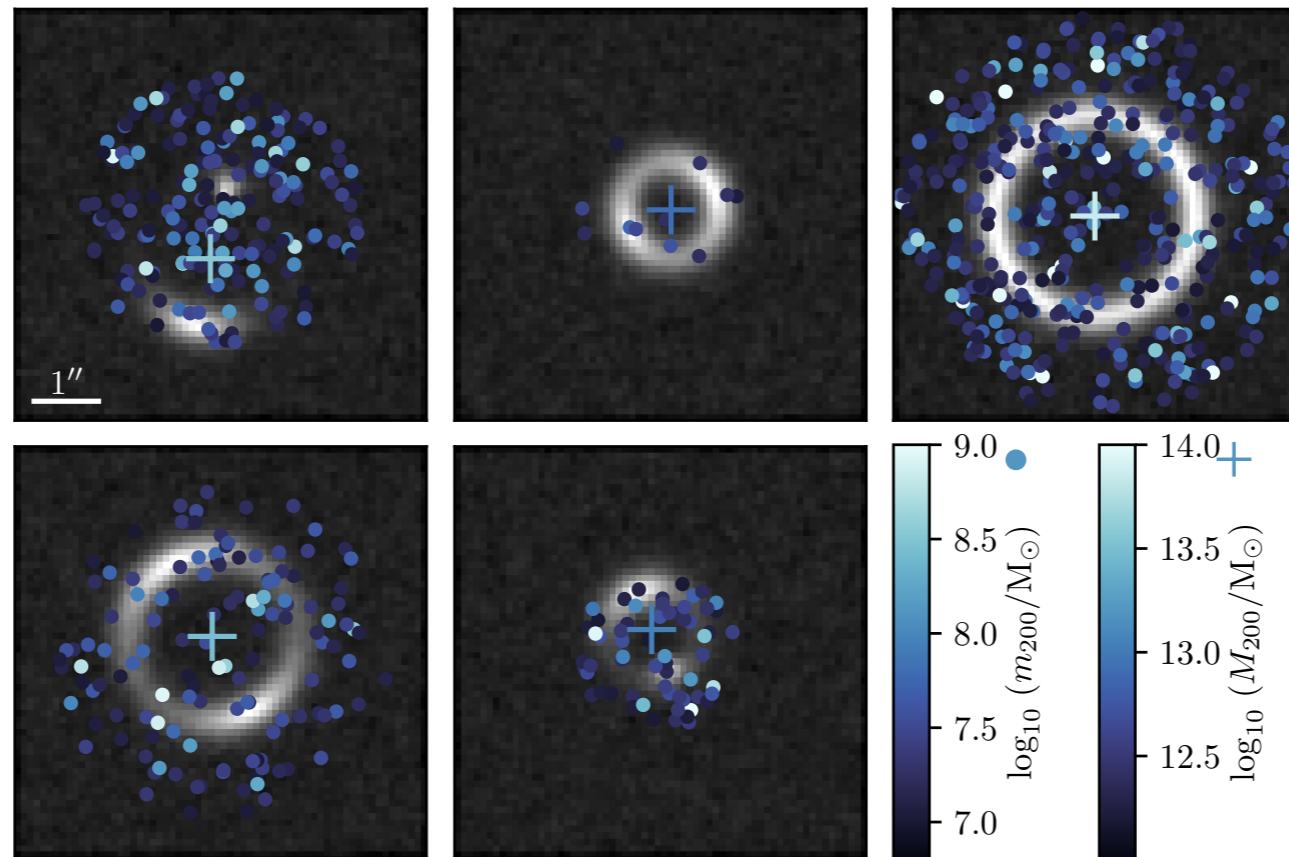
SIMULATION-BASED INFERENCE FOR STRONG LENSING



- ⇒ Need inference technique that
- scales to many lenses (fast evaluation)
 - captures subtle effects in high-dimensional image data
 - can deal with a large number of subhalos (latent variables)

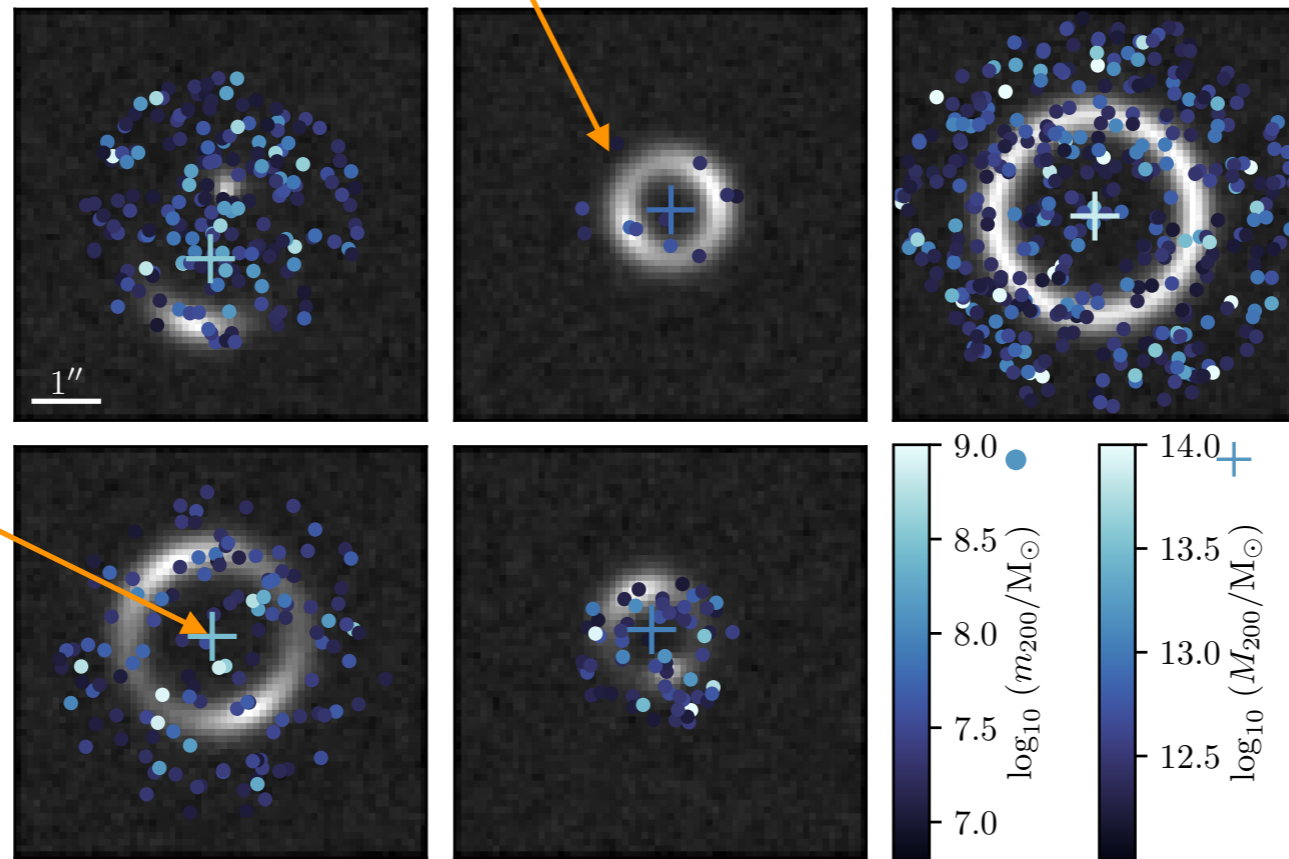
PROOF-OF-PRINCIPLE SIMULATOR

[following T. Collett 1507.02657]



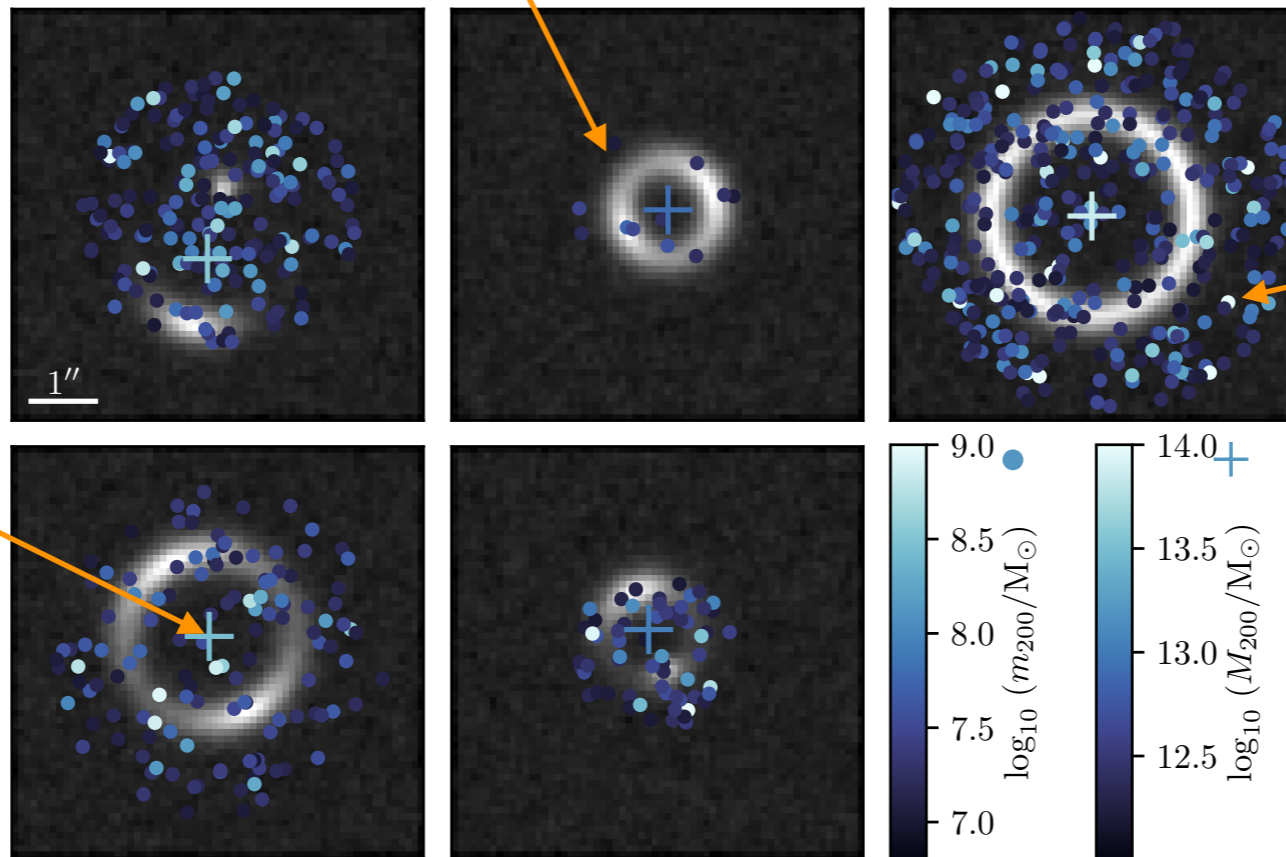
Spherical lensing host galaxies at redshift $\sim 0.5 \dots 1$

Extended galaxy sources at redshift 1.5

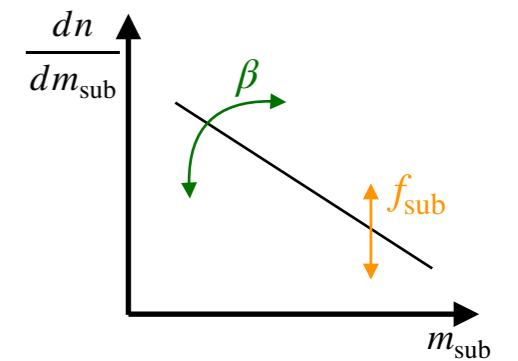


Spherical lensing host galaxies at redshift $\sim 0.5 \dots 1$

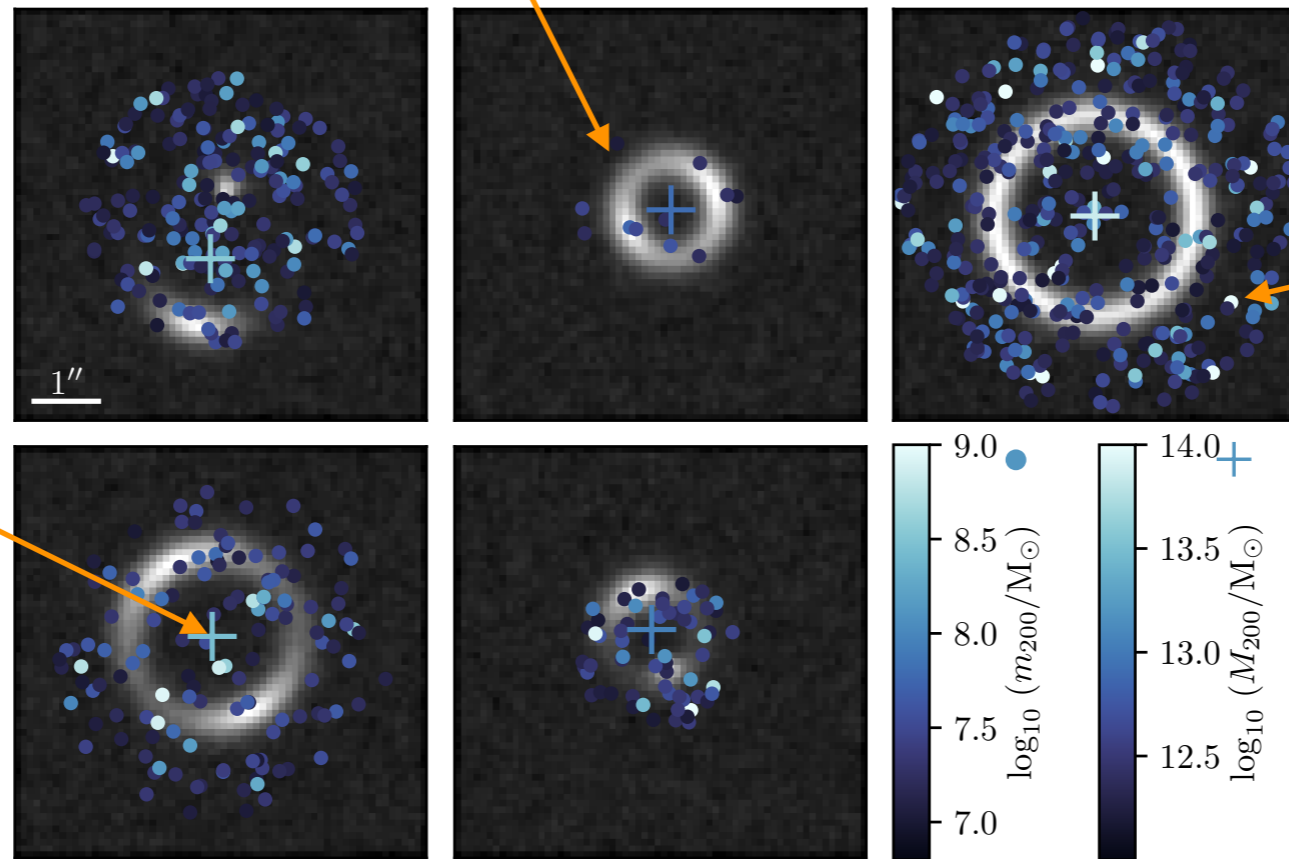
Extended galaxy sources at redshift 1.5



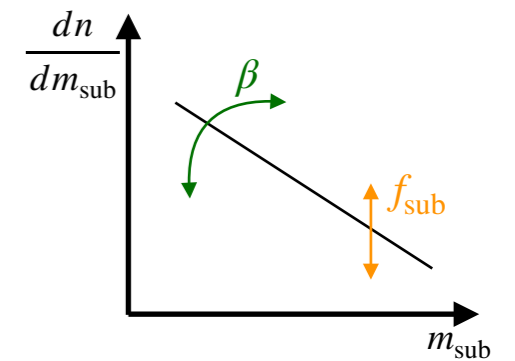
Subhalos follow mass distribution with two parameters



Spherical lensing host galaxies at redshift $\sim 0.5 \dots 1$



Subhalos follow mass distribution with two parameters



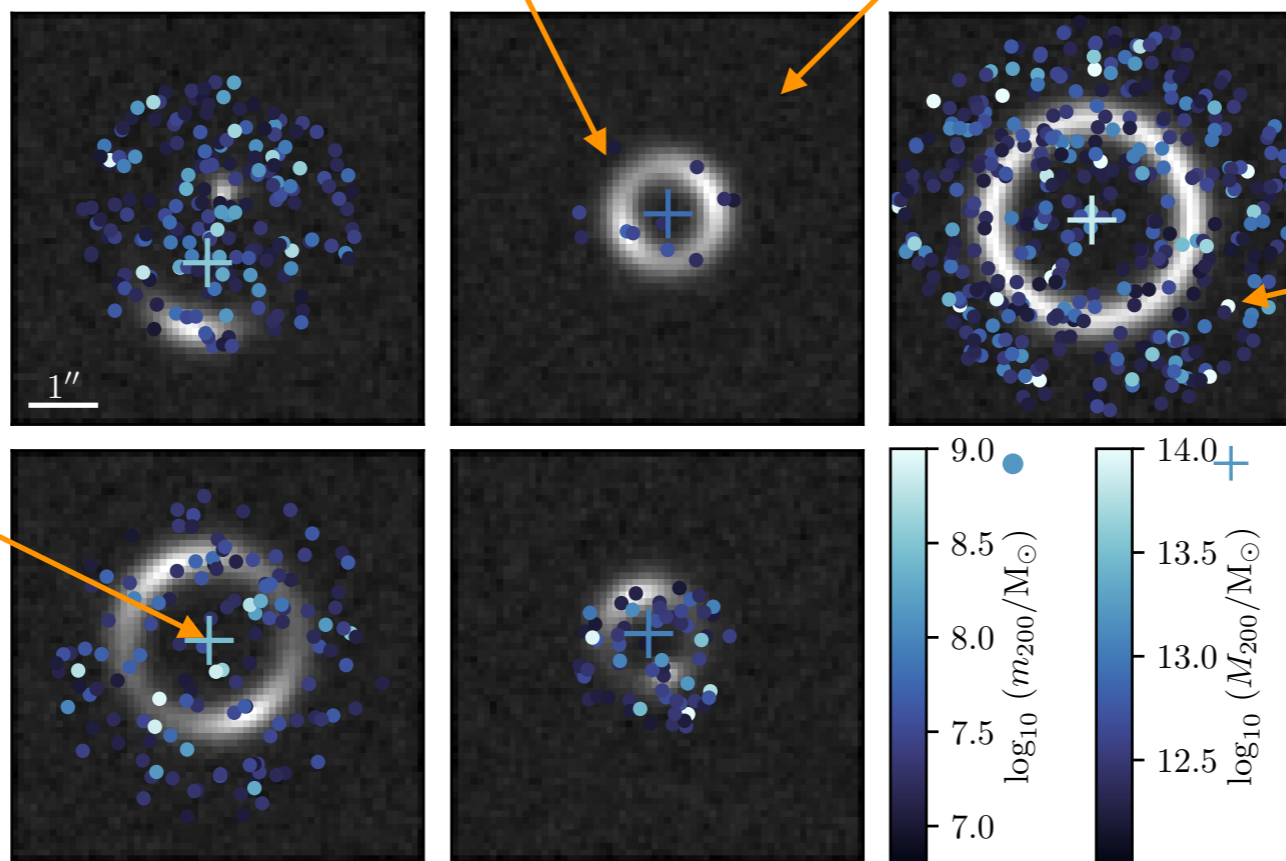
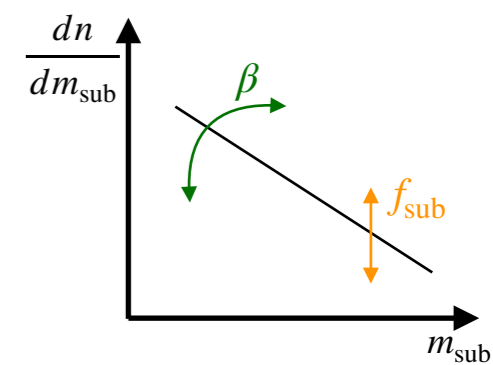
Extended galaxy sources at redshift 1.5

Euclid detector model
(single optical band,
 64^2 pixel, Gaussian PSF)

Spherical lensing host galaxies at redshift $\sim 0.5 \dots 1$

Poisson fluctuations

Subhalos follow mass distribution with two parameters



Extended galaxy sources at redshift 1.5

Euclid detector model
(single optical band,
 64^2 pixel, Gaussian PSF)

DARK MATTER

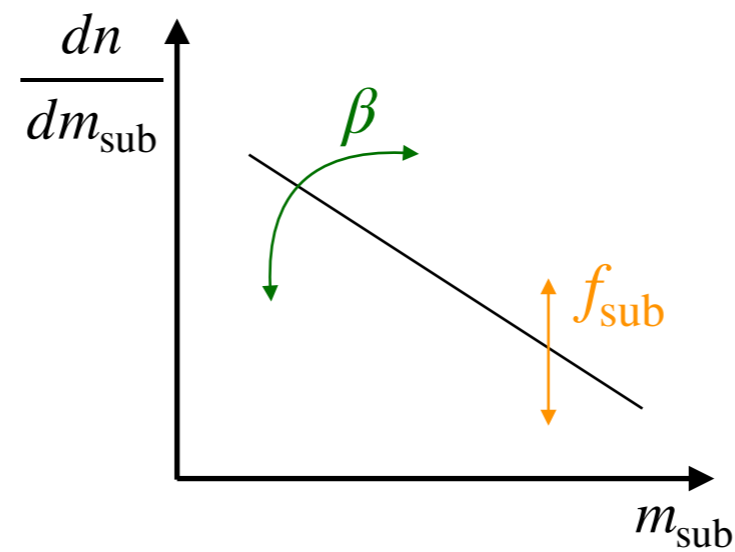
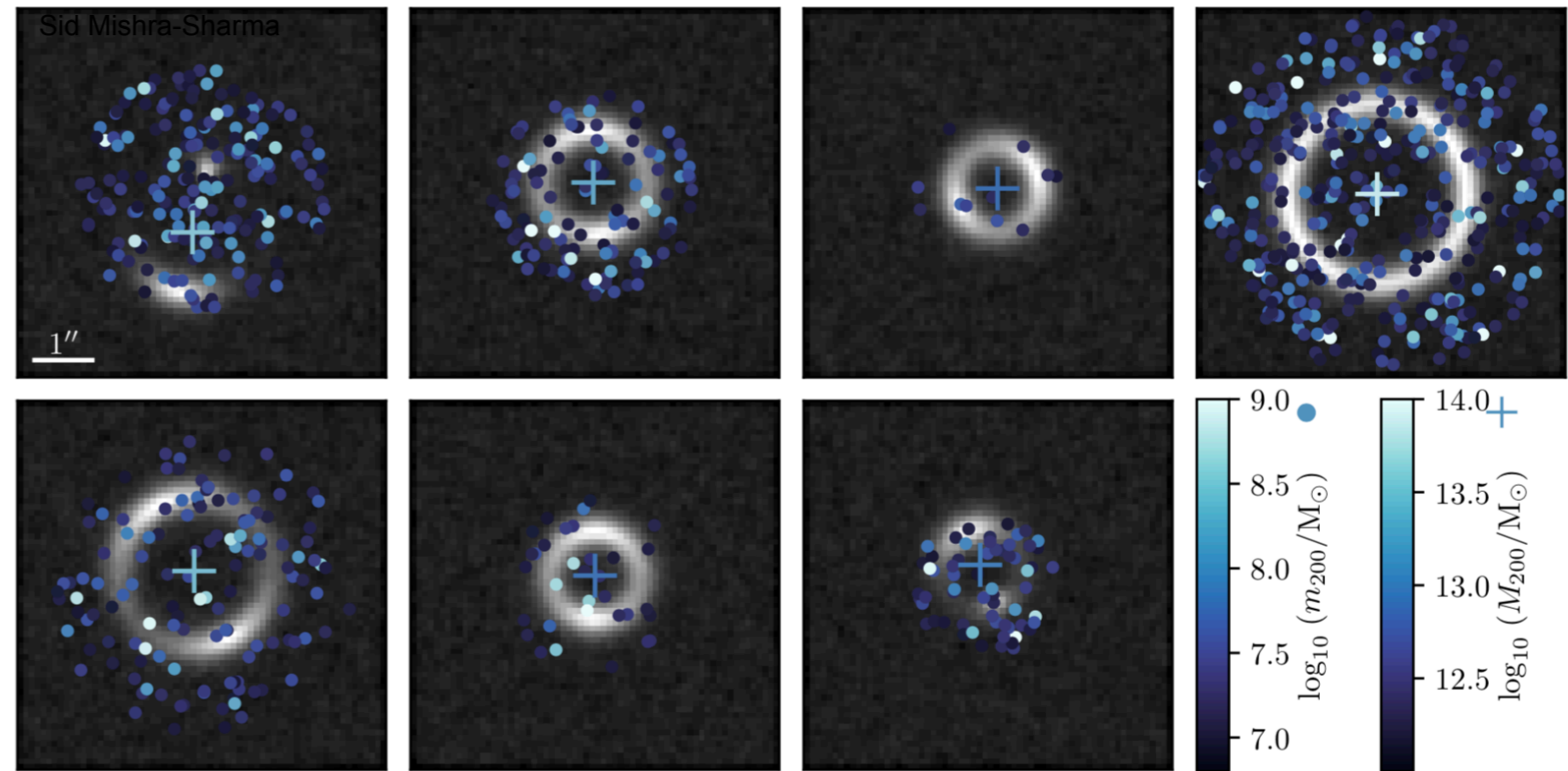
8

BREHMER AND MISHRA-SHARMA ET AL.

Latent space Z :

Number of dark matter sub halos and their mass and location lead to complex latent space for each image.

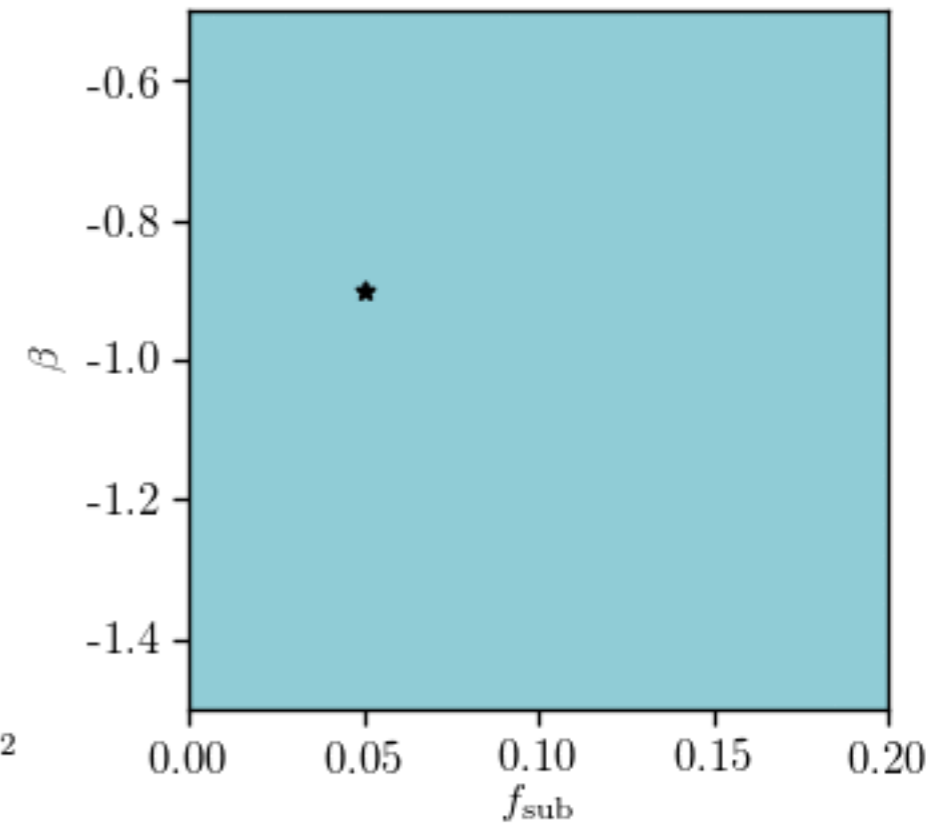
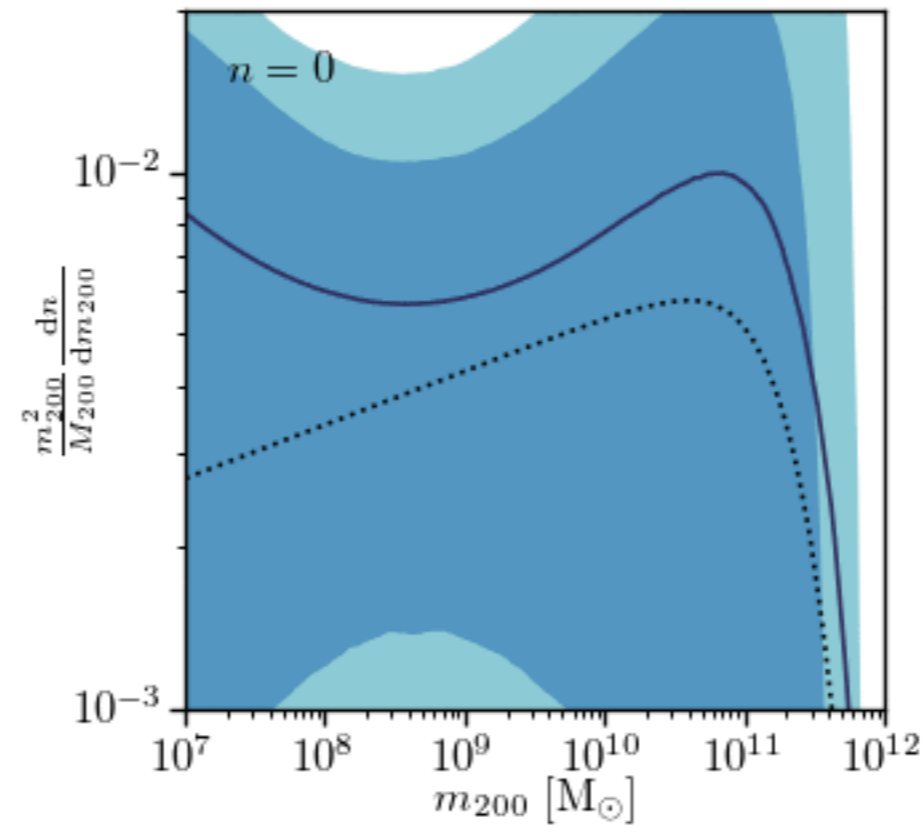
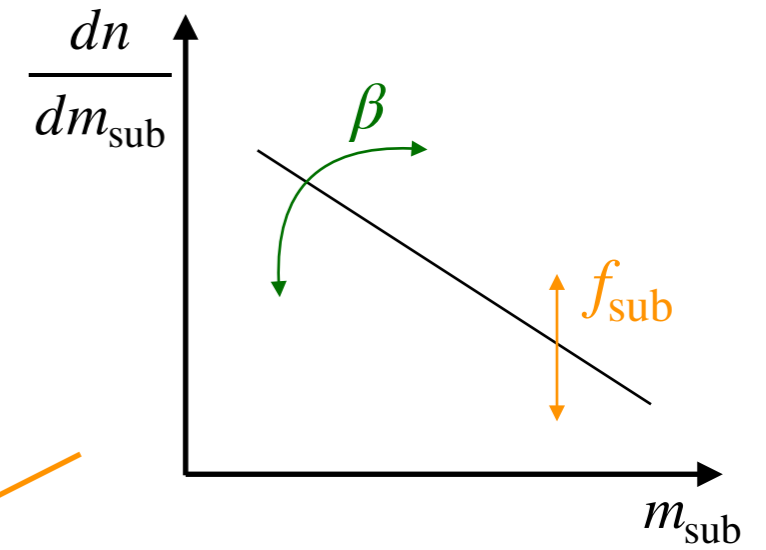
Goal is inference on θ at the population-level



RESULTS!

Watch how knowledge of the subhalos mass distribution improves as data comes in. See posterior for two parameters concentrate around true value used to generate mock data.

(plotted slightly differently in middle panel)



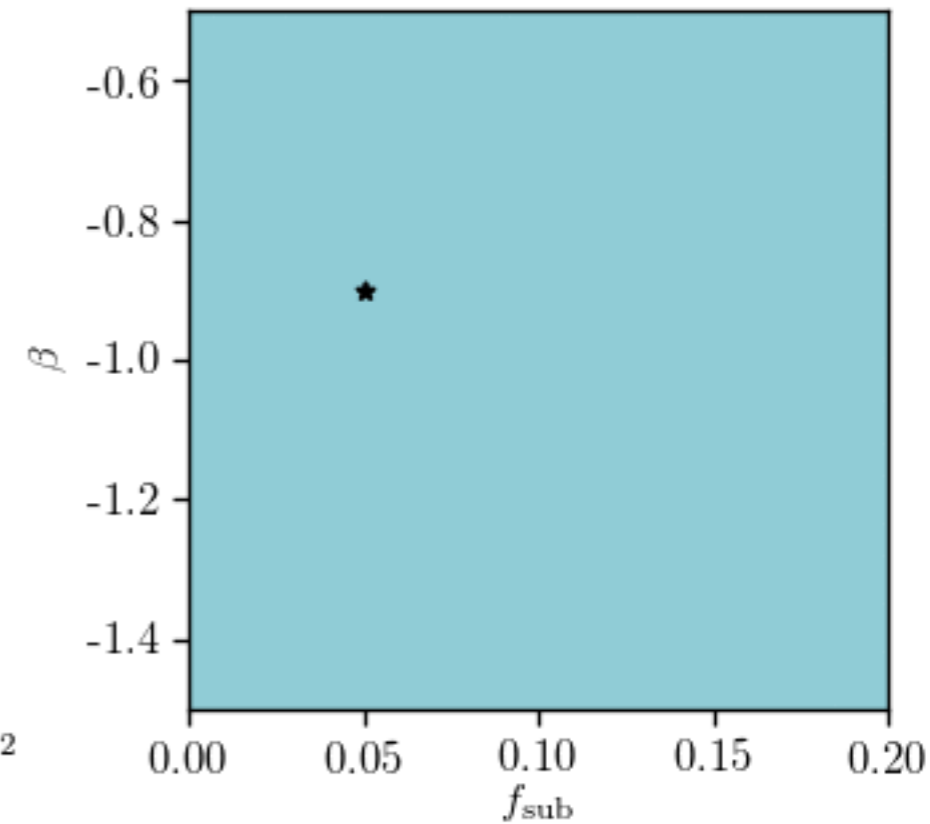
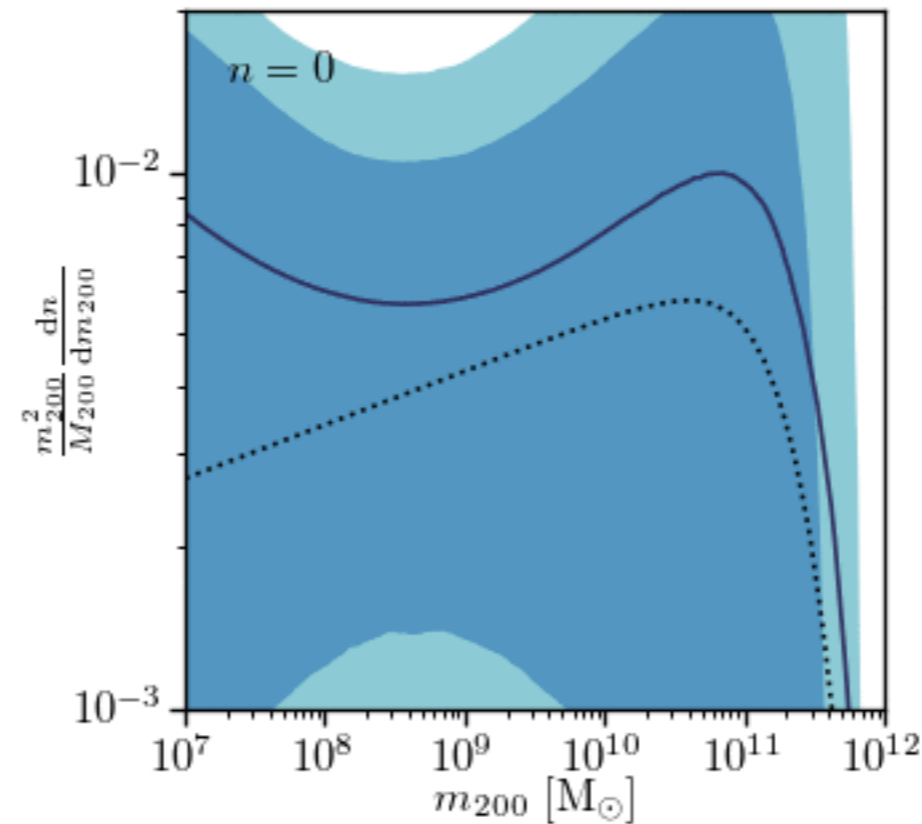
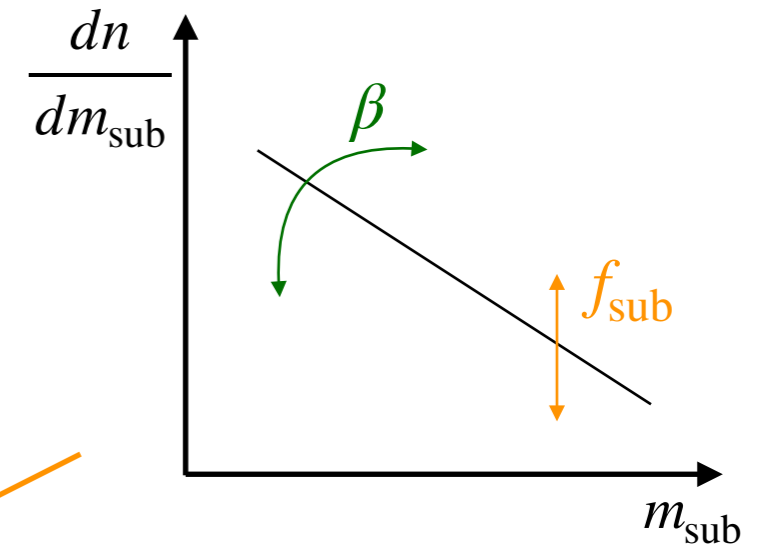
- f_{sub} , defined as the fraction of the total dark matter halo mass contained in bound substructure in a given mass range
- The halo virial mass M_{200} describes the total mass contained with the virial radius r_{200} , defined as the radius within which the mean density is 200 times the critical density of the universe



RESULTS!

Watch how knowledge of the subhalos mass distribution improves as data comes in. See posterior for two parameters concentrate around true value used to generate mock data.

(plotted slightly differently in middle panel)



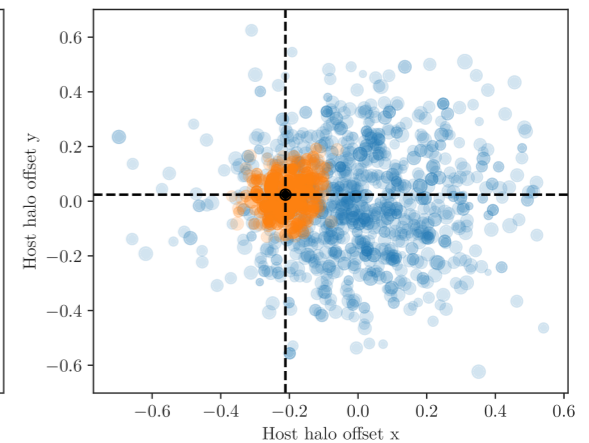
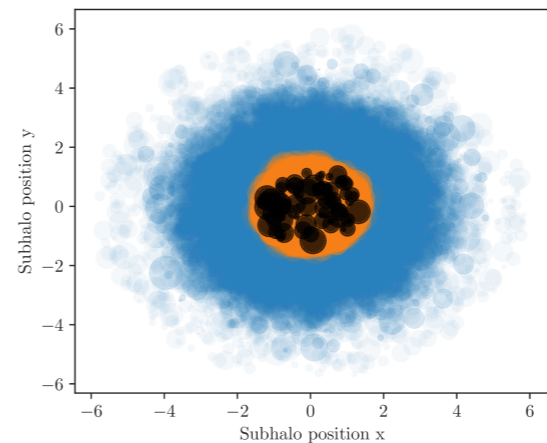
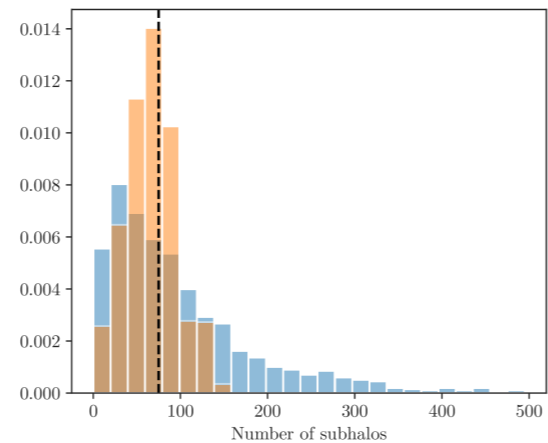
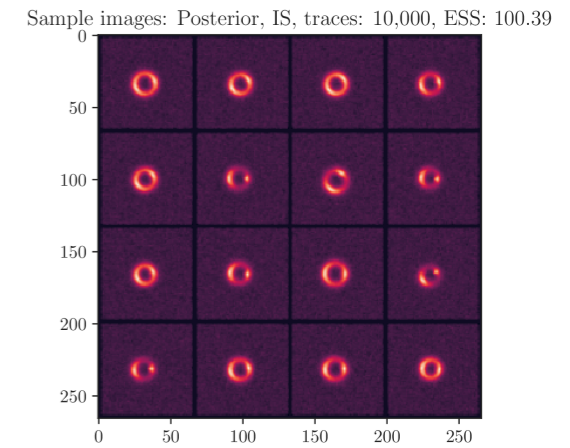
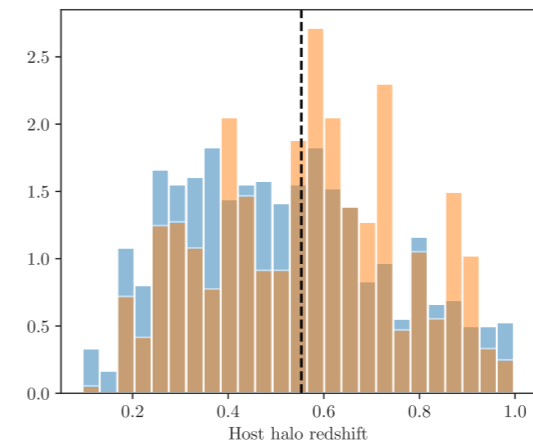
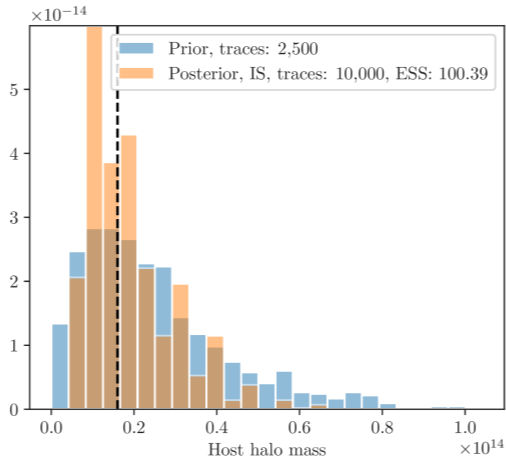
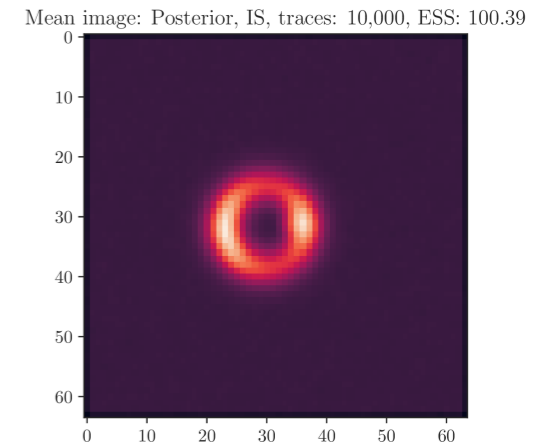
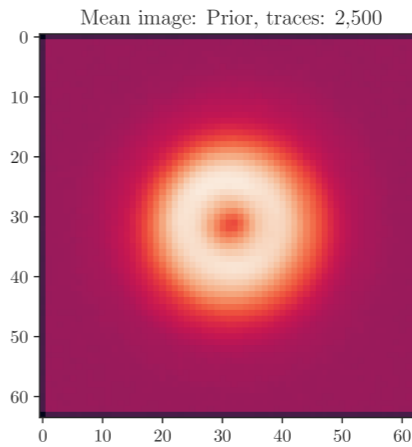
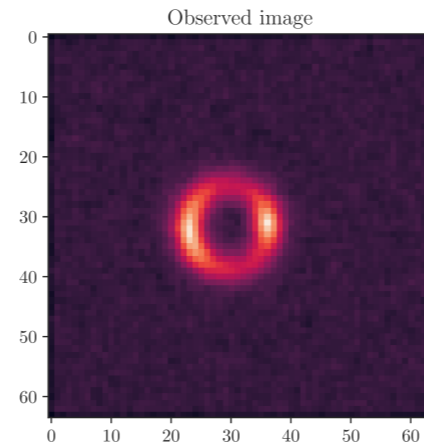
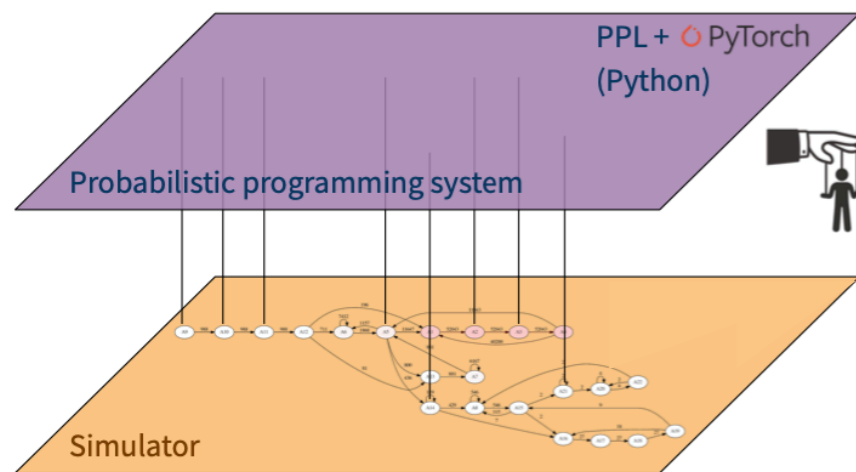
- f_{sub} , defined as the fraction of the total dark matter halo mass contained in bound substructure in a given mass range
- The halo virial mass M_{200} describes the total mass contained with the virial radius r_{200} , defined as the radius within which the mean density is 200 times the critical density of the universe

PROBABILISTIC PROGRAMMING FOR LENSING

PRELIMINARY!

Here we use probabilistic programming to infer the latent variables: the details of sub halo for a particular image

- prior
- posterior given observed image



Sid Mishra-Sharma



Johann Brehmer



Andreas Munk



Atılım Güneş Baydin