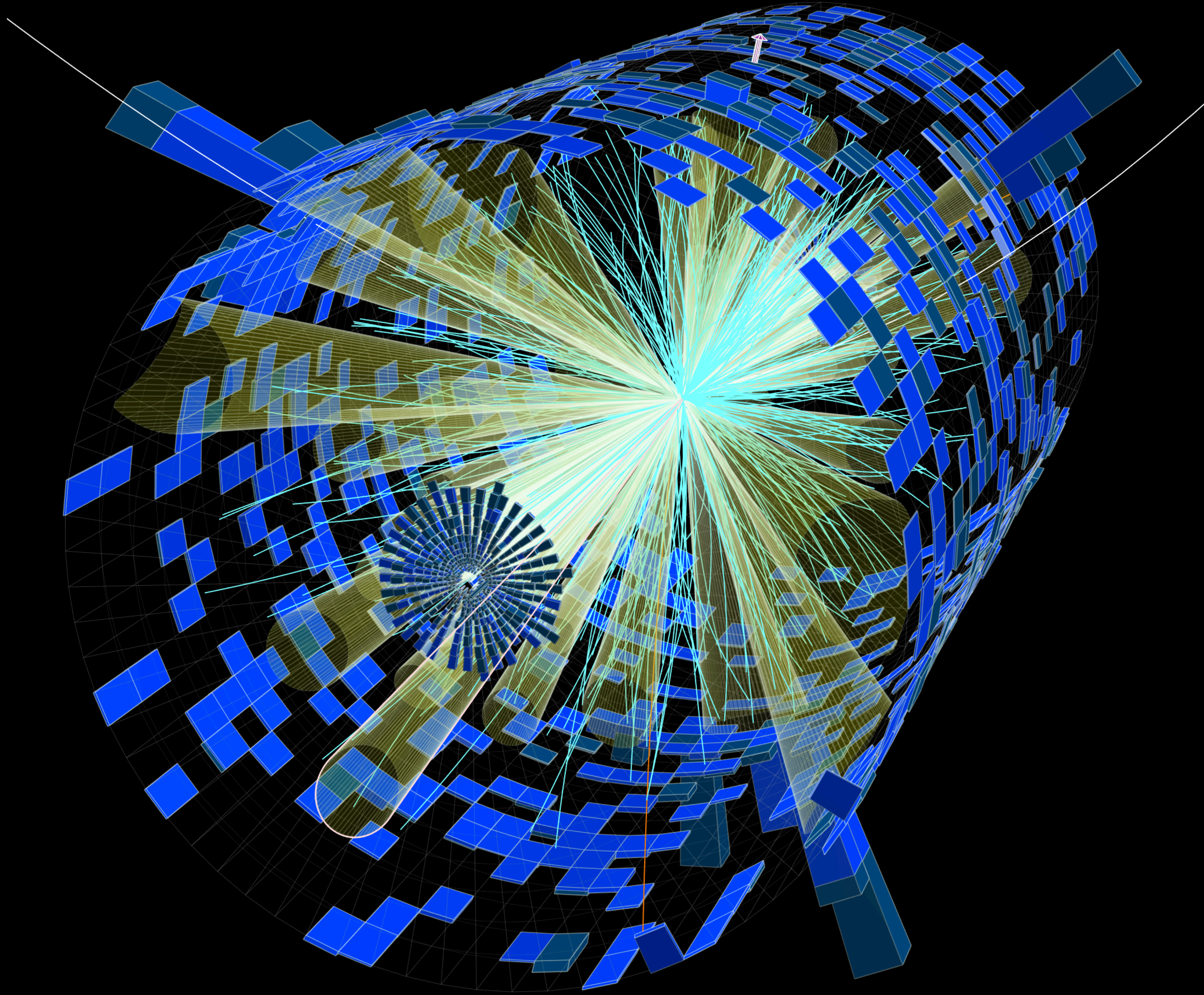
NYU Center for Data Science

Center for Cosmology and Particle Physics

REUSABLE WORKFLOWS, ACTIVE LEARNING, AND SIMULATION-BASED INFERENCE

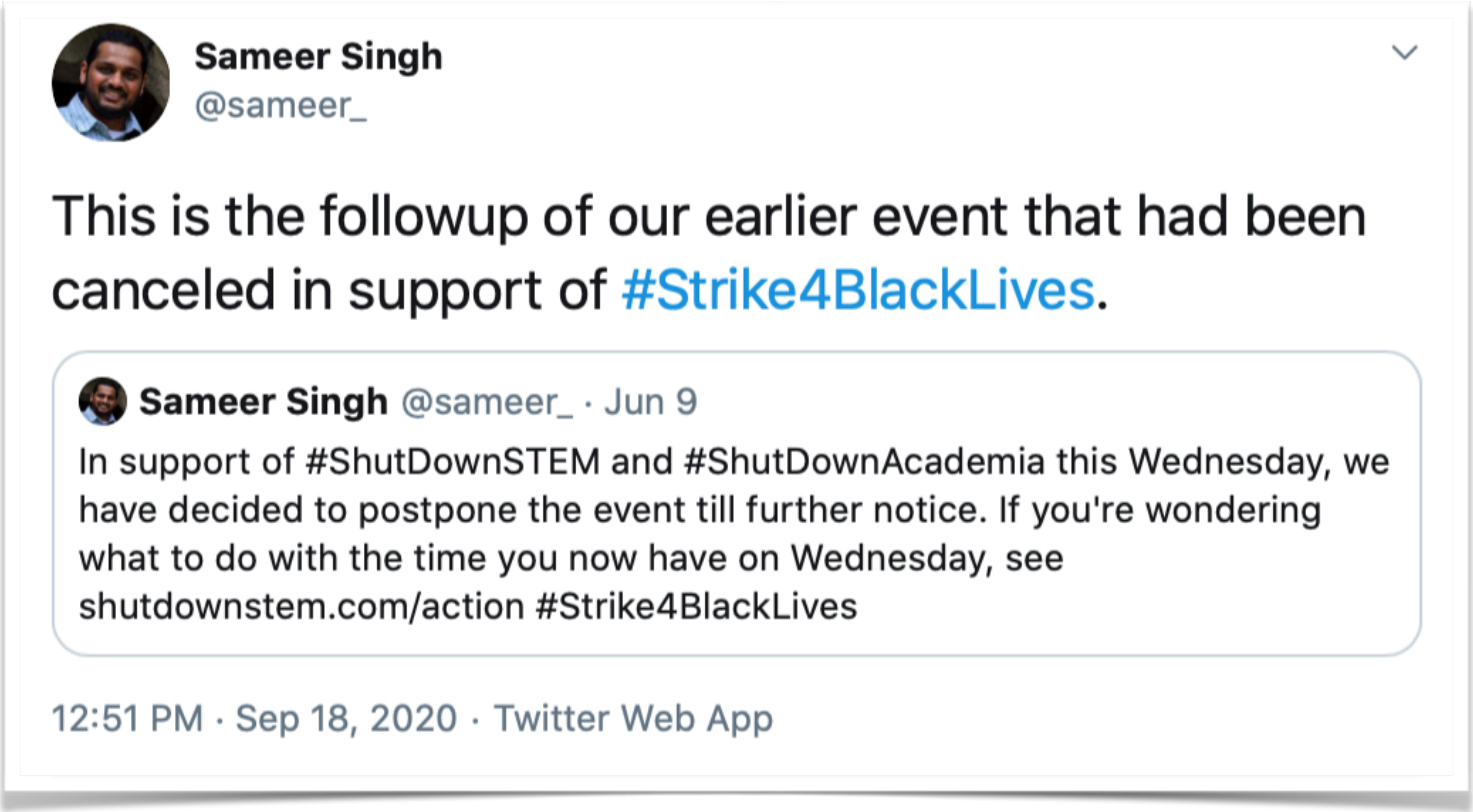
UCI SYMPOSIUM ON REPRODUCIBILITY IN MACHINE LEARNING



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

Thank you to the organizers for supporting #Strike4BlackLives

- The work is not done!



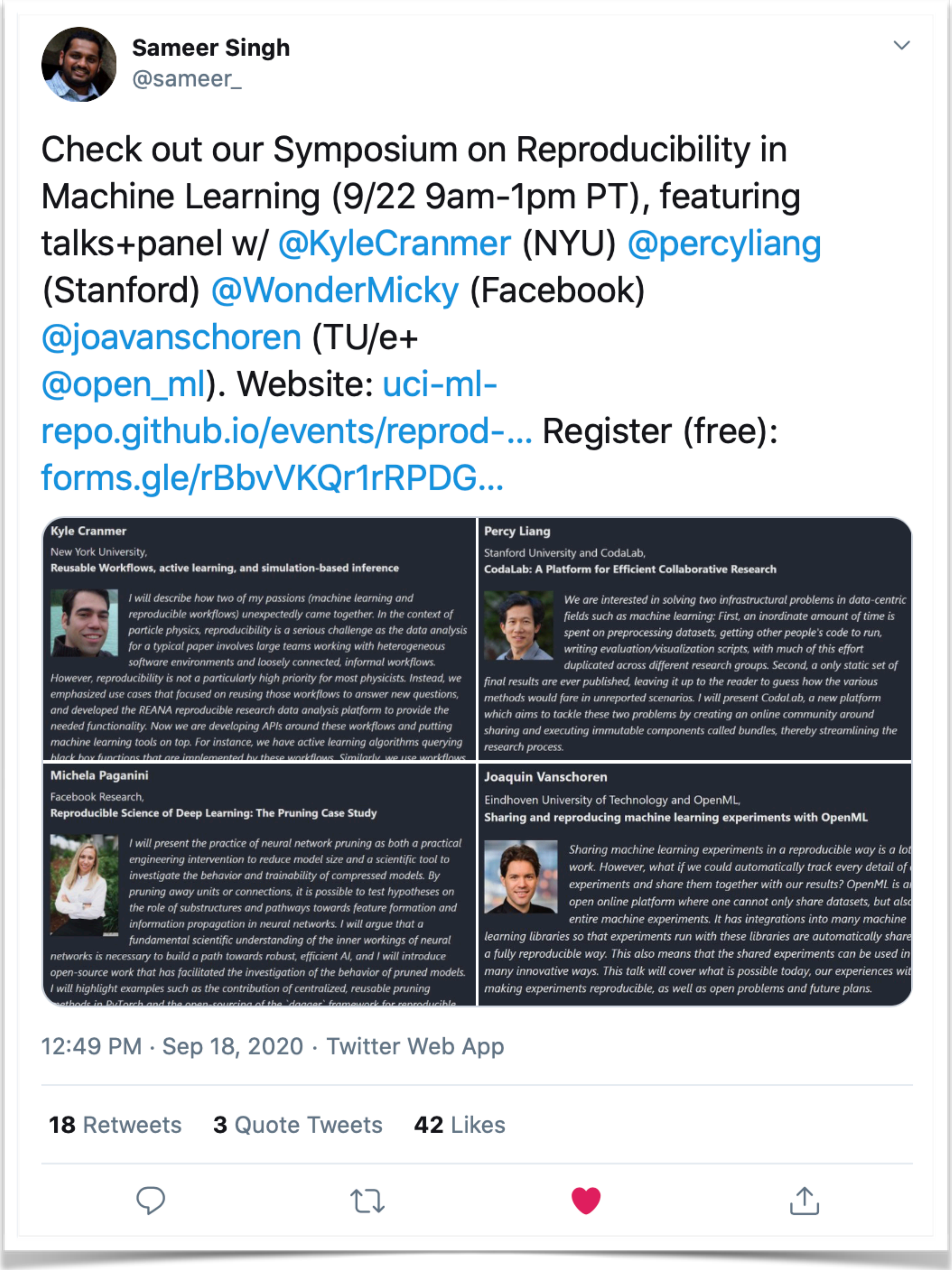
Sameer Singh
@sameer_

This is the followup of our earlier event that had been canceled in support of #Strike4BlackLives.

Sameer Singh @sameer_ · Jun 9

In support of #ShutDownSTEM and #ShutDownAcademia this Wednesday, we have decided to postpone the event till further notice. If you're wondering what to do with the time you now have on Wednesday, see shutdownstem.com/action #Strike4BlackLives

12:51 PM · Sep 18, 2020 · Twitter Web App



Sameer Singh
@sameer_

Check out our Symposium on Reproducibility in Machine Learning (9/22 9am-1pm PT), featuring talks+panel w/ @KyleCranmer (NYU) @percyliang (Stanford) @WonderMicky (Facebook) @joavanschoren (TU/e+ @open_ml). Website: uci-ml-repo.github.io/events/reprod-... Register (free): forms.gle/rBbvVKQr1rRPDG...

Kyle Cranmer

New York University,
Reusable Workflows, active learning, and simulation-based inference

I will describe how two of my passions (machine learning and reproducible workflows) unexpectedly came together. In the context of particle physics, reproducibility is a serious challenge as the data analysis for a typical paper involves large teams working with heterogeneous software environments and loosely connected, informal workflows. However, reproducibility is not a particularly high priority for most physicists. Instead, we emphasized use cases that focused on reusing those workflows to answer new questions, and developed the REANA reproducible research data analysis platform to provide the needed functionality. Now we are developing APIs around these workflows and putting machine learning tools on top. For instance, we have active learning algorithms querying black box functions that are implemented by these workflows. Similarly, we use workflows

Percy Liang

Stanford University and CodaLab,
CodaLab: A Platform for Efficient Collaborative Research

We are interested in solving two infrastructural problems in data-centric fields such as machine learning: First, an inordinate amount of time is spent on preprocessing datasets, getting other people's code to run, writing evaluation/visualization scripts, with much of this effort duplicated across different research groups. Second, a only static set of final results are ever published, leaving it up to the reader to guess how the various methods would fare in unreported scenarios. I will present CodaLab, a new platform which aims to tackle these two problems by creating an online community around sharing and executing immutable components called bundles, thereby streamlining the research process.

Michela Paganini

Facebook Research,
Reproducible Science of Deep Learning: The Pruning Case Study

I will present the practice of neural network pruning as both a practical engineering intervention to reduce model size and a scientific tool to investigate the behavior and trainability of compressed models. By pruning away units or connections, it is possible to test hypotheses on the role of substructures and pathways towards feature formation and information propagation in neural networks. I will argue that a fundamental scientific understanding of the inner workings of neural networks is necessary to build a path towards robust, efficient AI, and I will introduce open-source work that has facilitated the investigation of the behavior of pruned models. I will highlight examples such as the contribution of centralized, reusable pruning methods in PyTorch and the open-sourcing of the `dagger` framework for reproducible

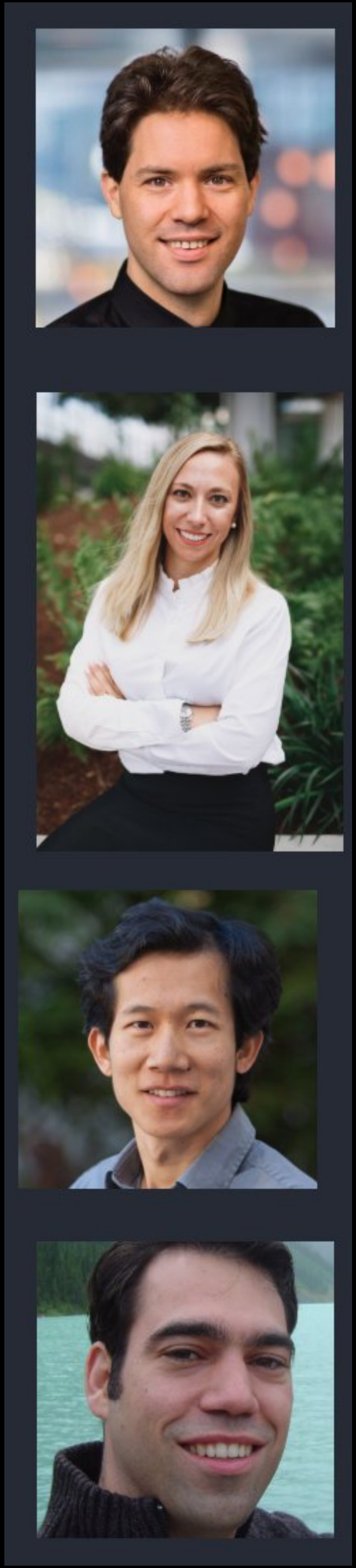
Joaquin Vanschoren

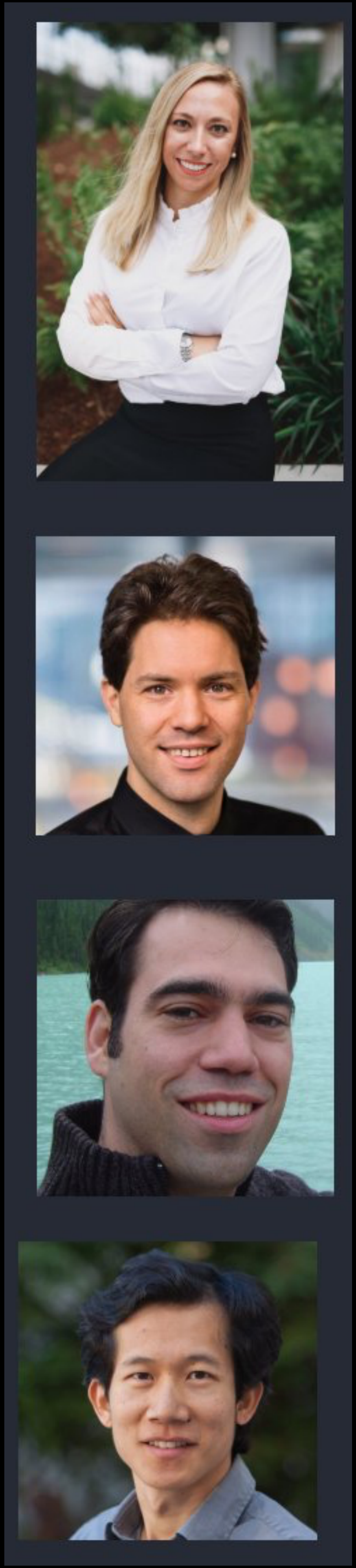
Eindhoven University of Technology and OpenML,
Sharing and reproducing machine learning experiments with OpenML

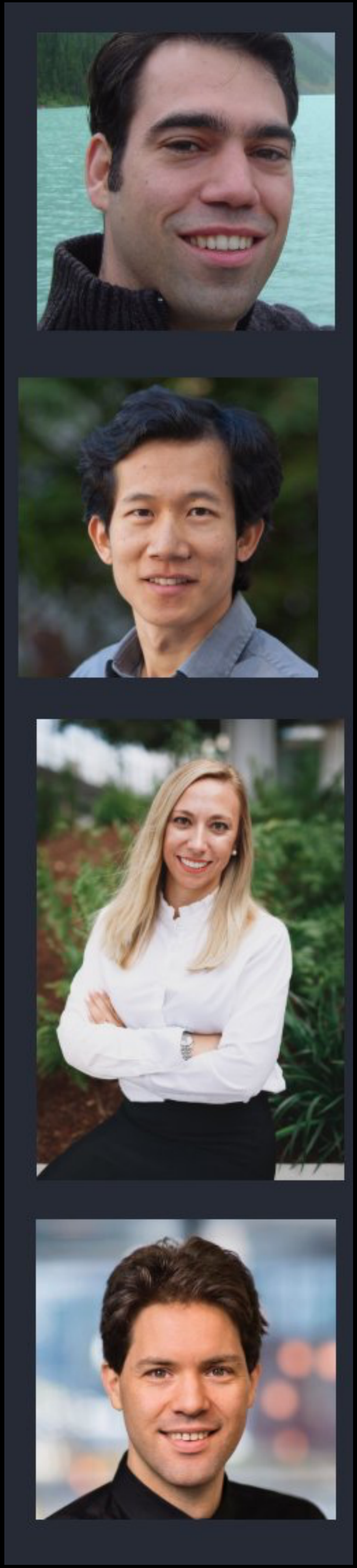
Sharing machine learning experiments in a reproducible way is a lot work. However, what if we could automatically track every detail of experiments and share them together with our results? OpenML is a open online platform where one cannot only share datasets, but also entire machine experiments. It has integrations into many machine learning libraries so that experiments run with these libraries are automatically share a fully reproducible way. This also means that the shared experiments can be used in many innovative ways. This talk will cover what is possible today, our experiences wit making experiments reproducible, as well as open problems and future plans.

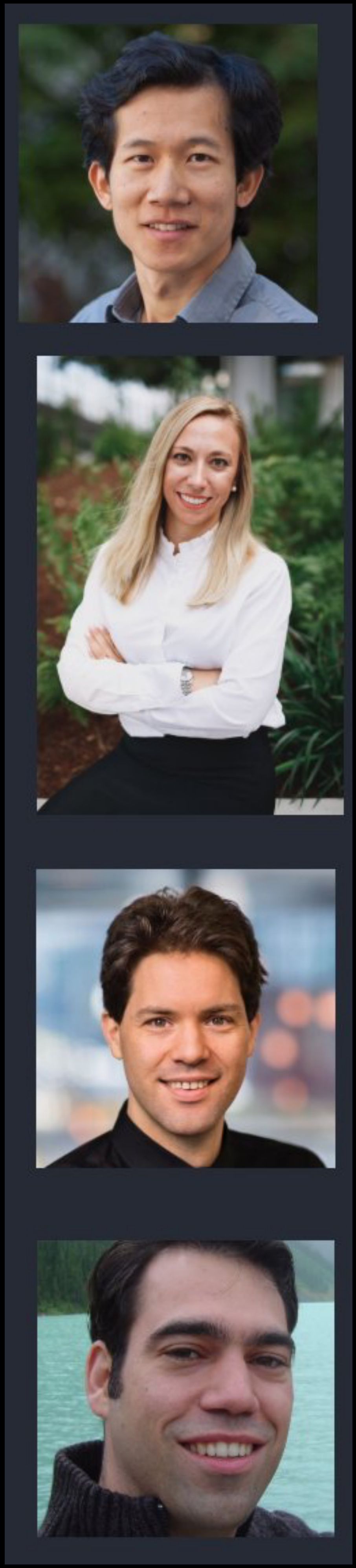
12:49 PM · Sep 18, 2020 · Twitter Web App

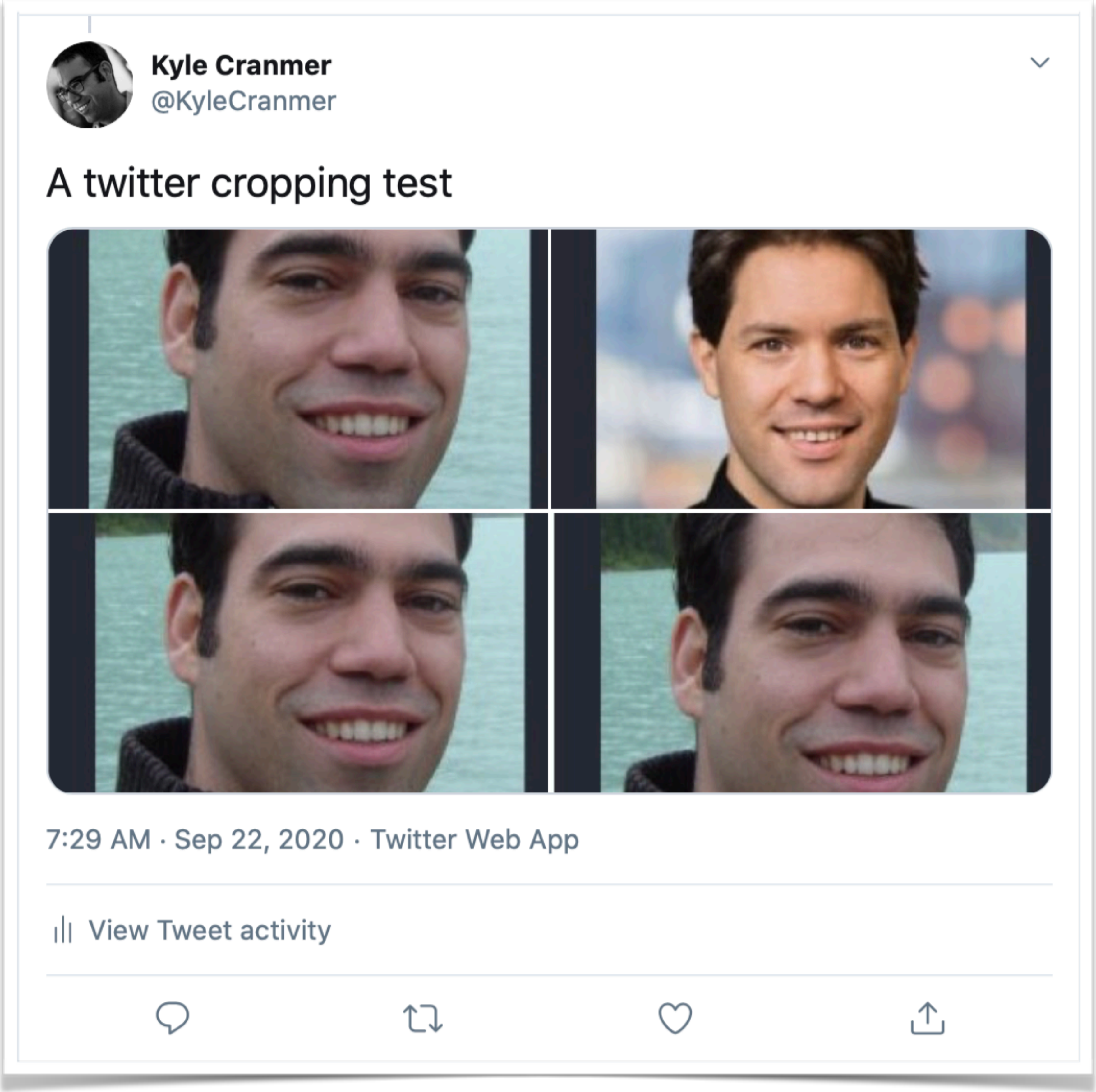
18 Retweets **3** Quote Tweets **42** Likes









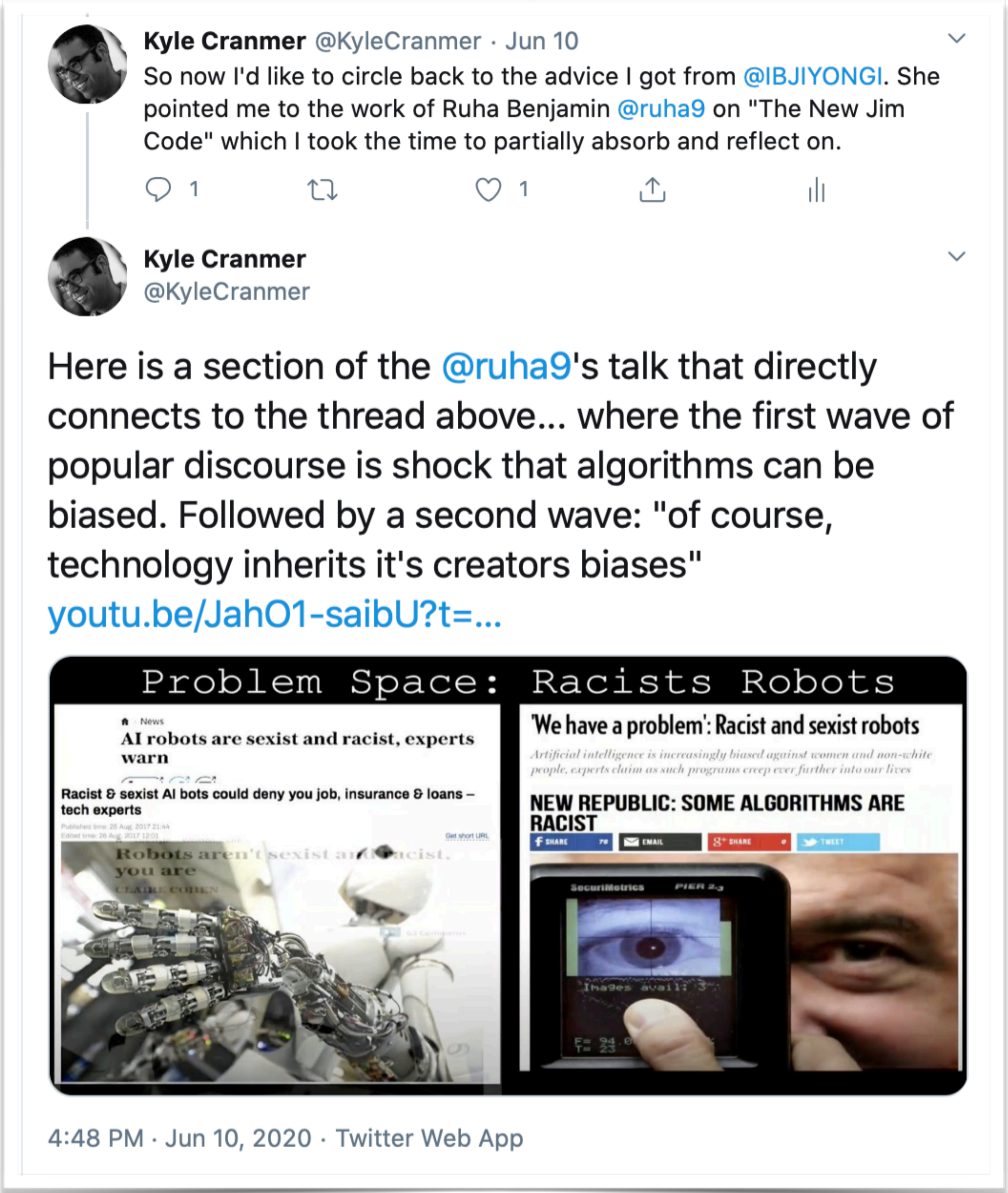


Kyle Cranmer
@KyleCranmer

A twitter cropping test

7:29 AM · Sep 22, 2020 · Twitter Web App

For more on the cropping issue, see: https://twitter.com/fhuszar/status/1307982313012236288



Kyle Cranmer @KyleCranmer · Jun 10

So now I'd like to circle back to the advice I got from @IBJIYONGI. She pointed me to the work of Ruha Benjamin @ruha9 on "The New Jim Code" which I took the time to partially absorb and reflect on.

Kyle Cranmer
@KyleCranmer

Here is a section of the @ruha9's talk that directly connects to the thread above... where the first wave of popular discourse is shock that algorithms can be biased. Followed by a second wave: "of course, technology inherits it's creators biases"
youtu.be/JahO1-saibU?t=...

Problem Space: Racists Robots

News
AI robots are sexist and racist, experts warn

Racist & sexist AI bots could deny you job, insurance & loans – tech experts

Robots aren't sexist and racist, you are

'We have a problem': Racist and sexist robots

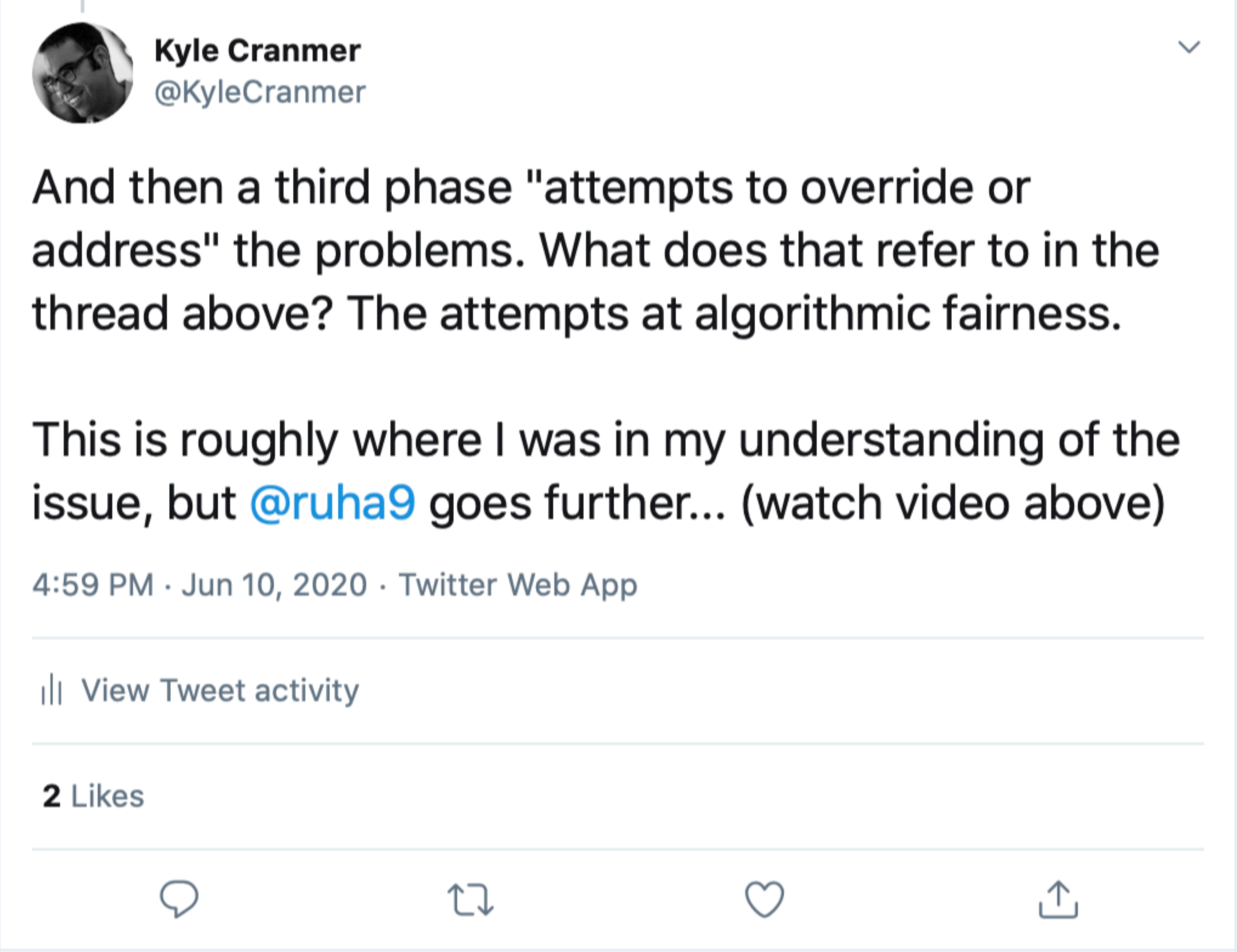
Artificial intelligence is increasingly biased against women and non-white people, experts claim as such programs creep ever further into our lives

NEW REPUBLIC: SOME ALGORITHMS ARE RACIST

4:48 PM · Jun 10, 2020 · Twitter Web App

Ruha Benjamin on "The New Jim Code? Race, Carceral Technoscience, and Liberatory Imagination"

https://www.youtube.com/watch?v=JahO1-saibU&feature=youtu.be&t=1323

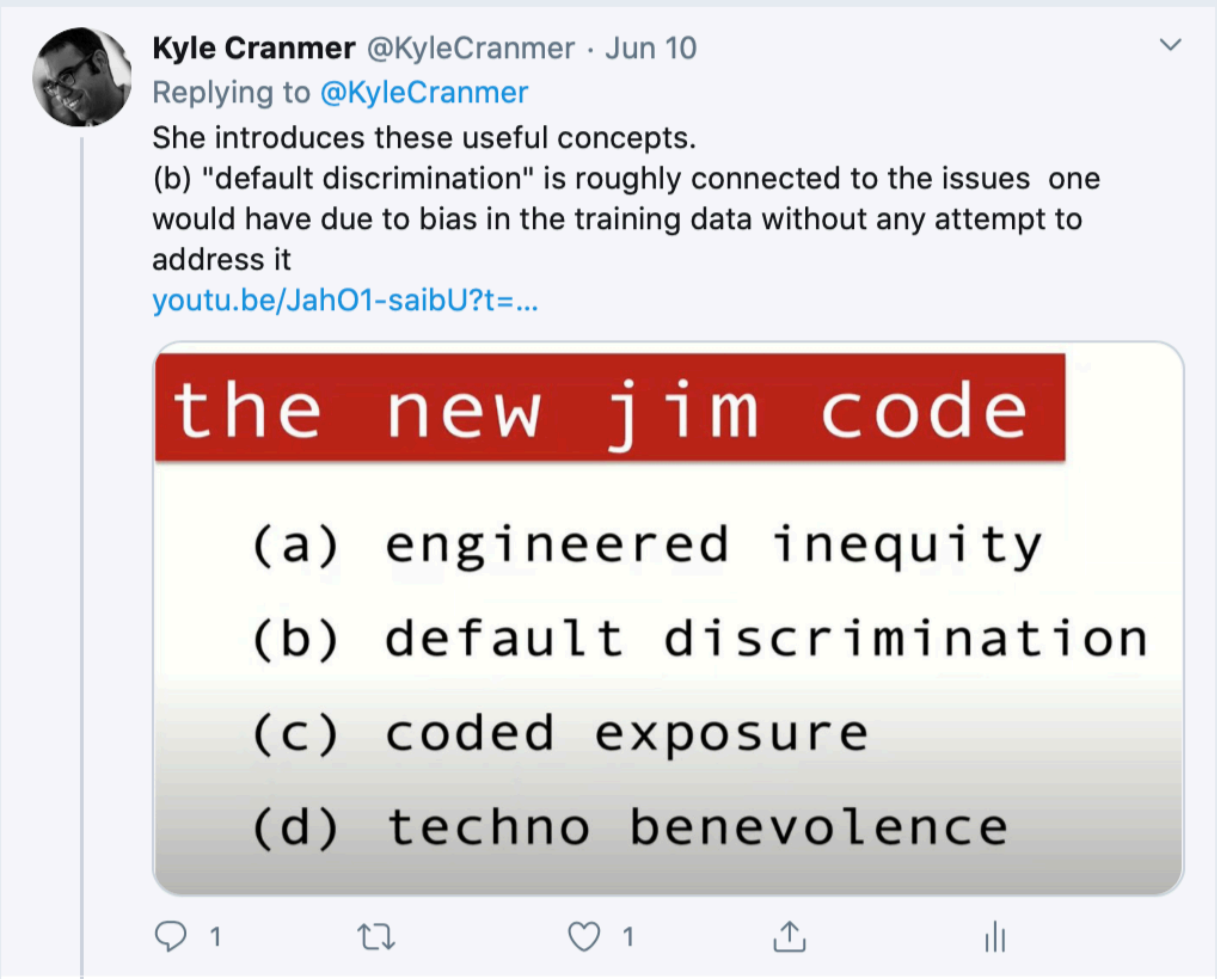


Kyle Cranmer
@KyleCranmer

And then a third phase "attempts to override or address" the problems. What does that refer to in the thread above? The attempts at algorithmic fairness.

This is roughly where I was in my understanding of the issue, but @ruha9 goes further... (watch video above)

4:59 PM · Jun 10, 2020 · Twitter Web App



Kyle Cranmer @KyleCranmer · Jun 10

Replying to @KyleCranmer

She introduces these useful concepts.
(b) "default discrimination" is roughly connected to the issues one would have due to bias in the training data without any attempt to address it
youtu.be/JahO1-saibU?t=...

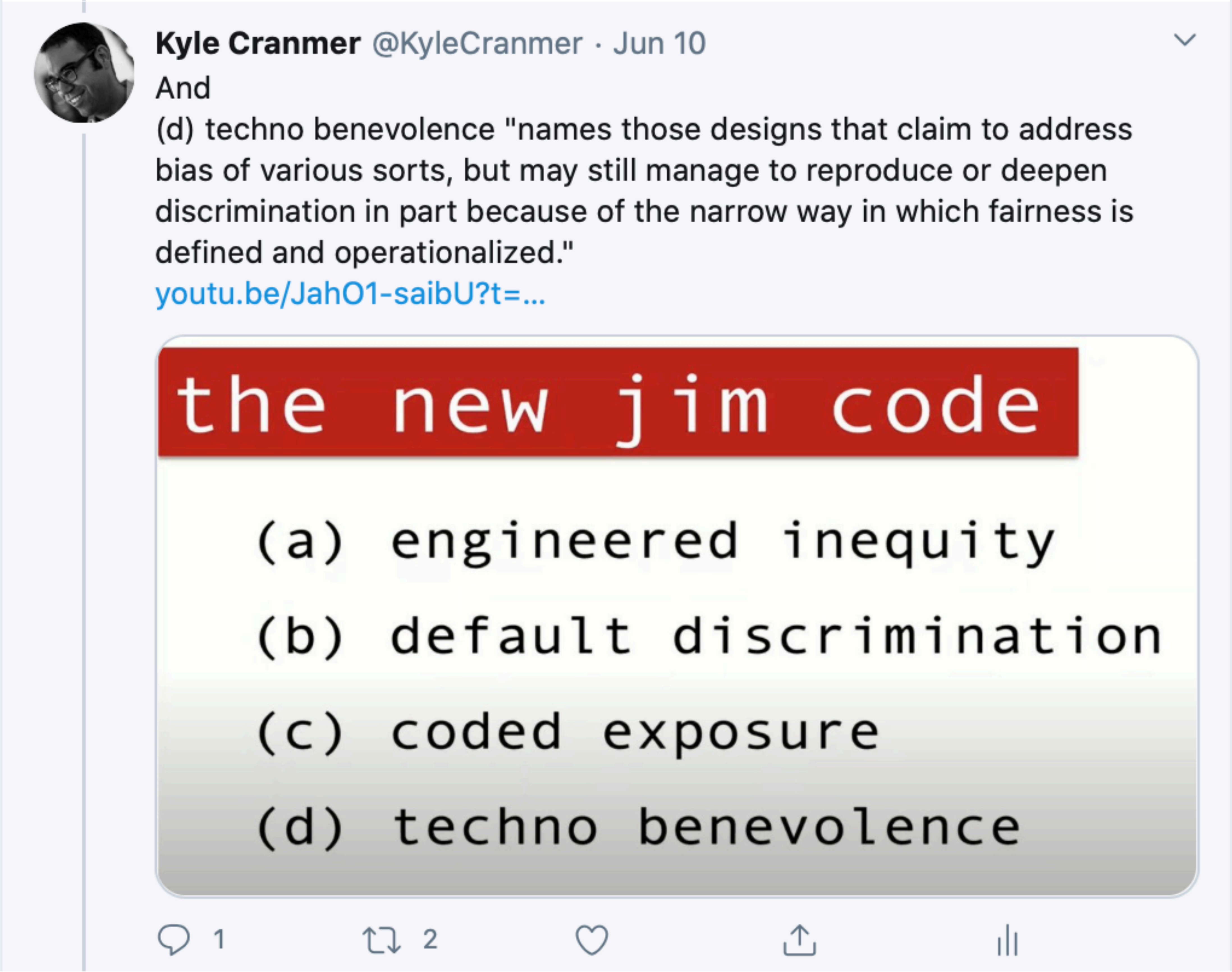
the new jim code

(a) engineered inequity

(b) default discrimination

(c) coded exposure

(d) techno benevolence



Kyle Cranmer @KyleCranmer · Jun 10

And
(d) techno benevolence "names those designs that claim to address bias of various sorts, but may still manage to reproduce or deepen discrimination in part because of the narrow way in which fairness is defined and operationalized."
youtu.be/JahO1-saibU?t=...

the new jim code

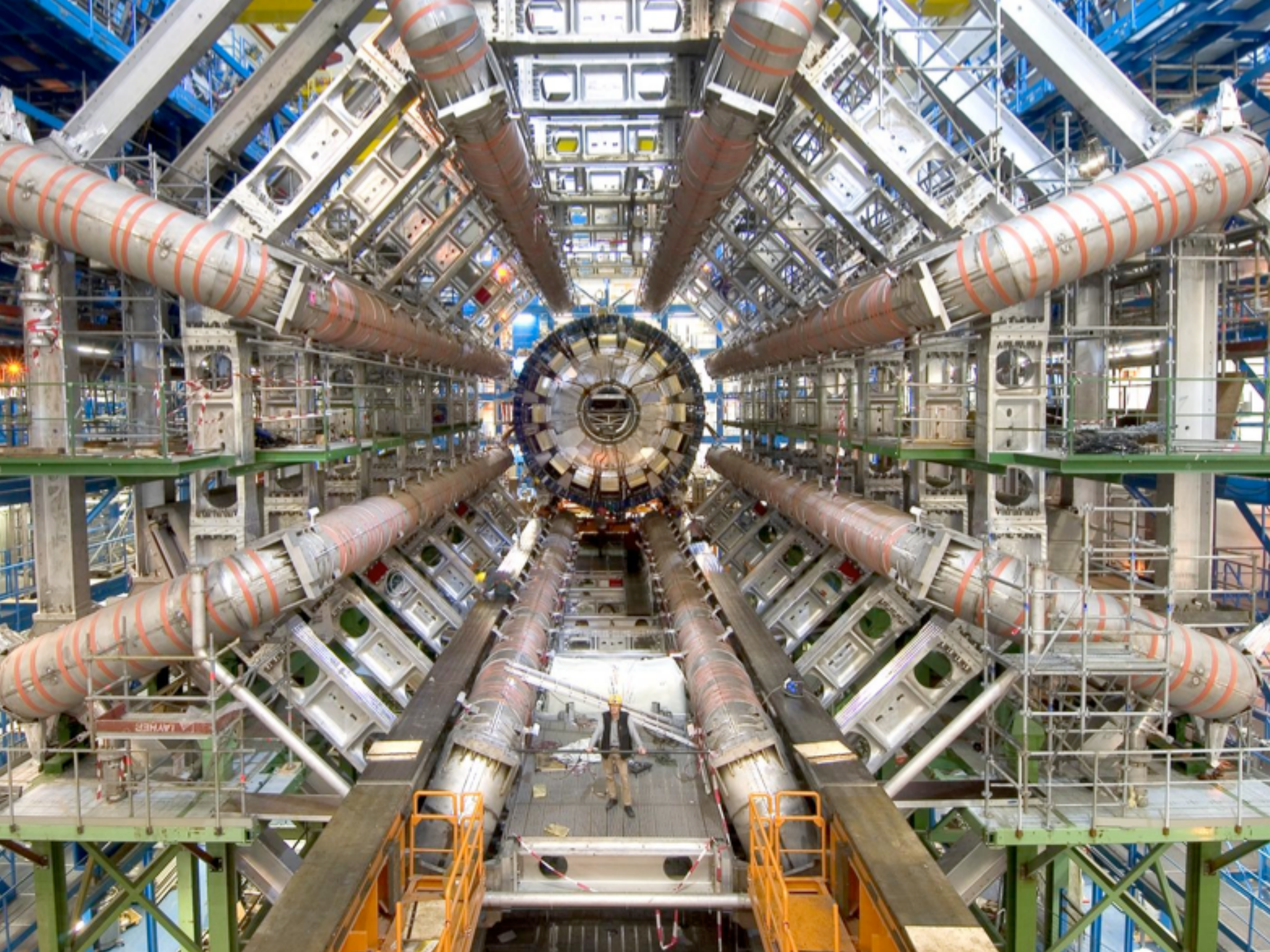
(a) engineered inequity

(b) default discrimination

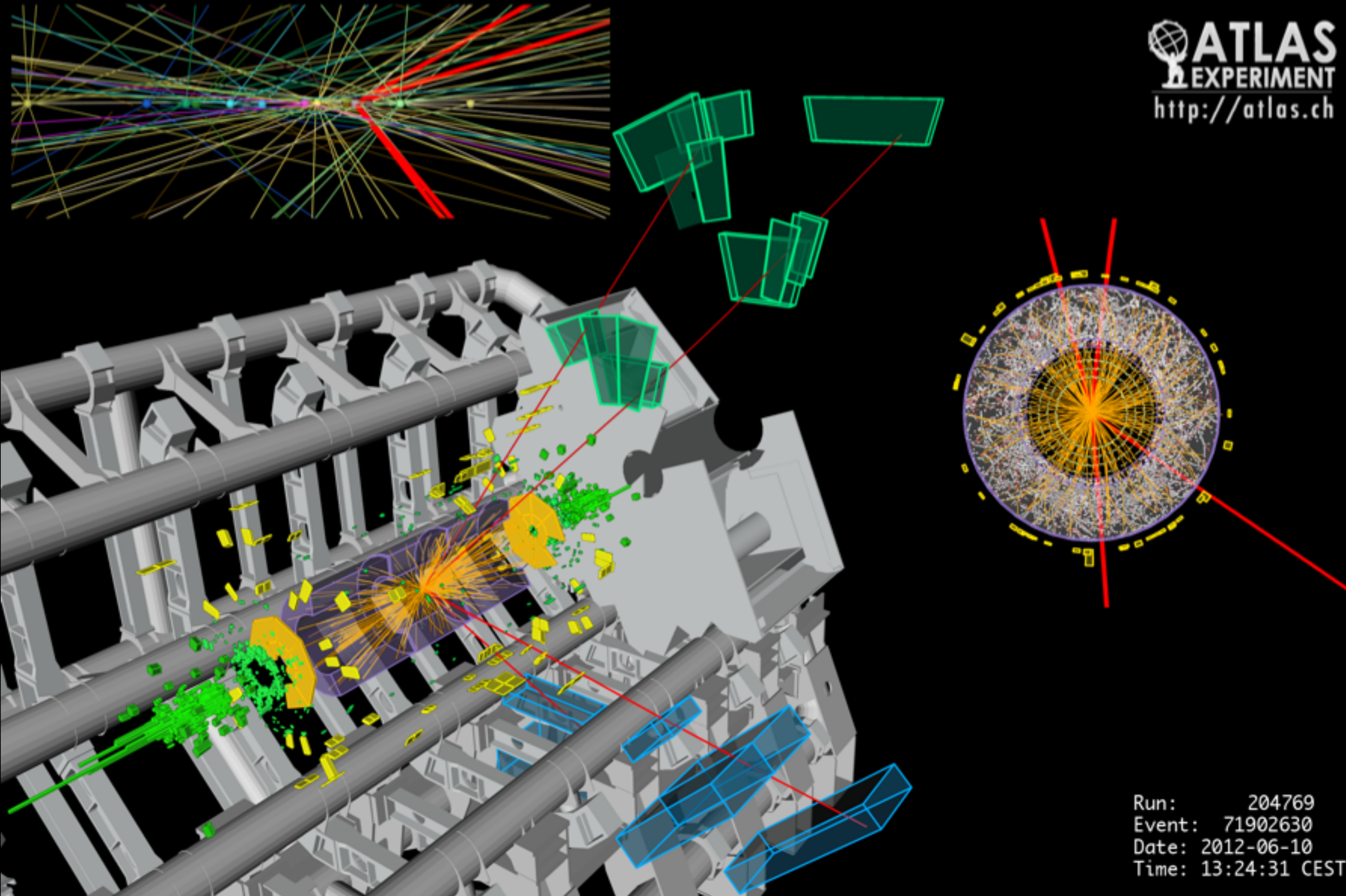
(c) coded exposure

(d) techno benevolence

Let us keep these issues in our minds
as we develop systems and technologies
that impact people



THE HIGGS BOSON



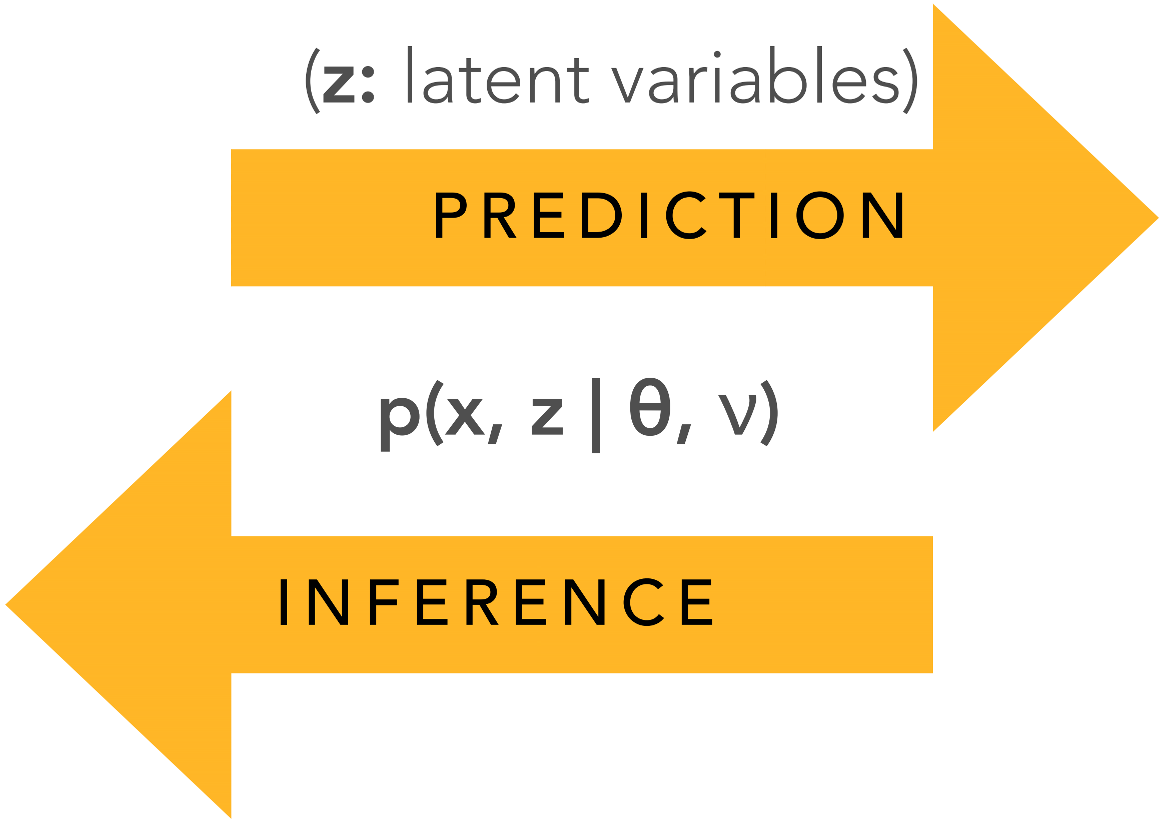
THE PLAYERS

θ
parameters of interest

ν
nuisance parameters

forward modeling
generation
simulation

(**z:** latent variables)



PREDICTION

$p(x, z \mid \theta, \nu)$

INFERENCE

inverse problem
measurement
parameter estimation

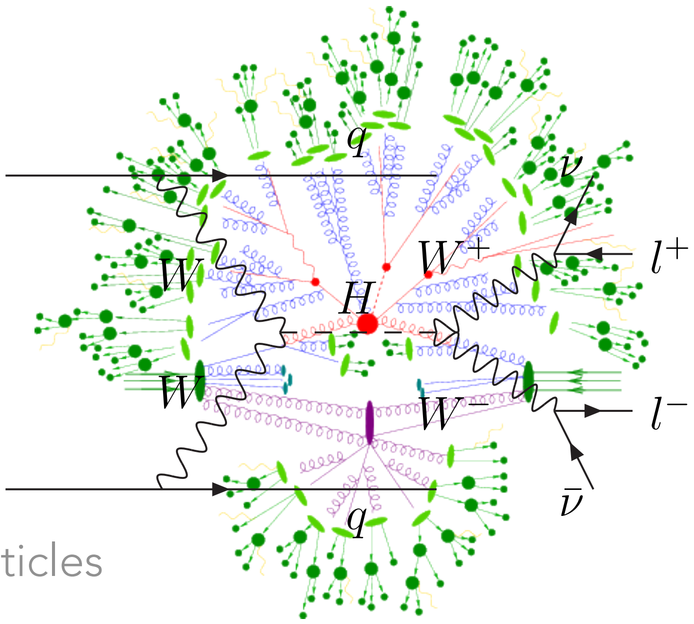
x
observed data
covariates
simulated data

THE FORWARD MODEL

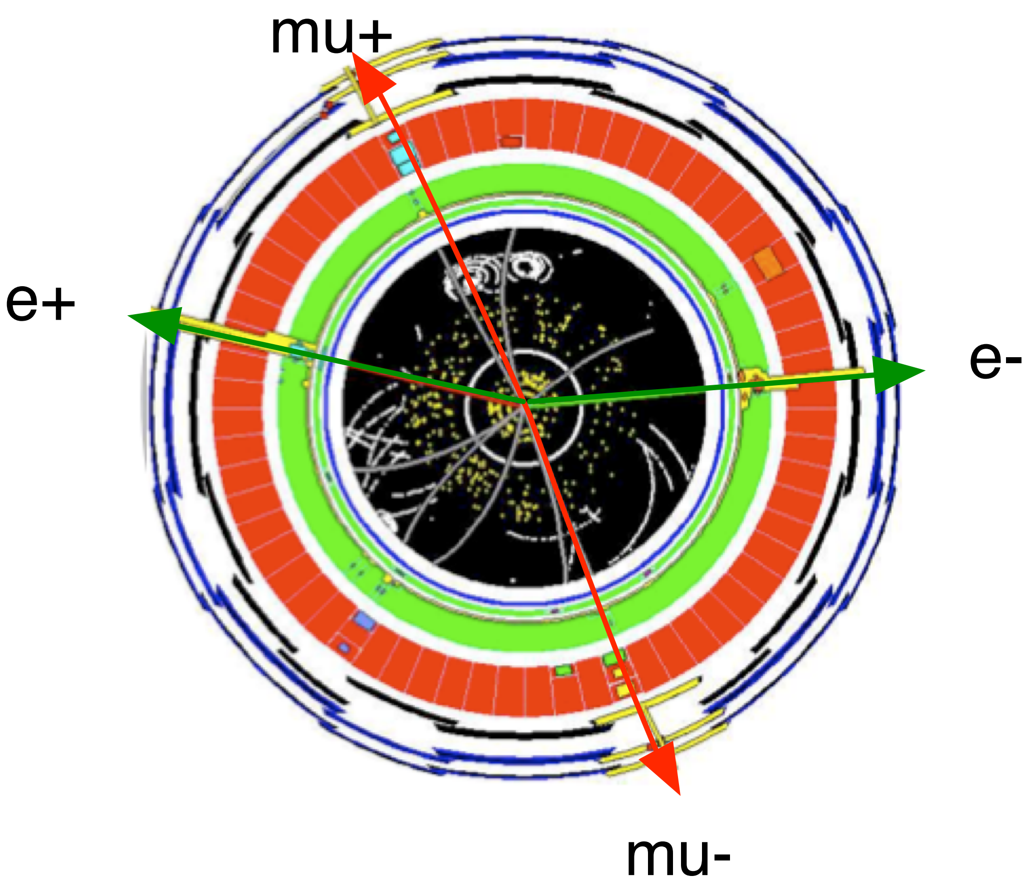
$$
\begin{aligned}
\mathcal{L}_{SM} = \quad & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
+ \quad & \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu}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

1) We begin with Quantum Field Theory

2) Theory gives detailed prediction for high-energy collisions



hierarchical: 2 → O(10) → O(100) particles



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

>100 million sensors

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

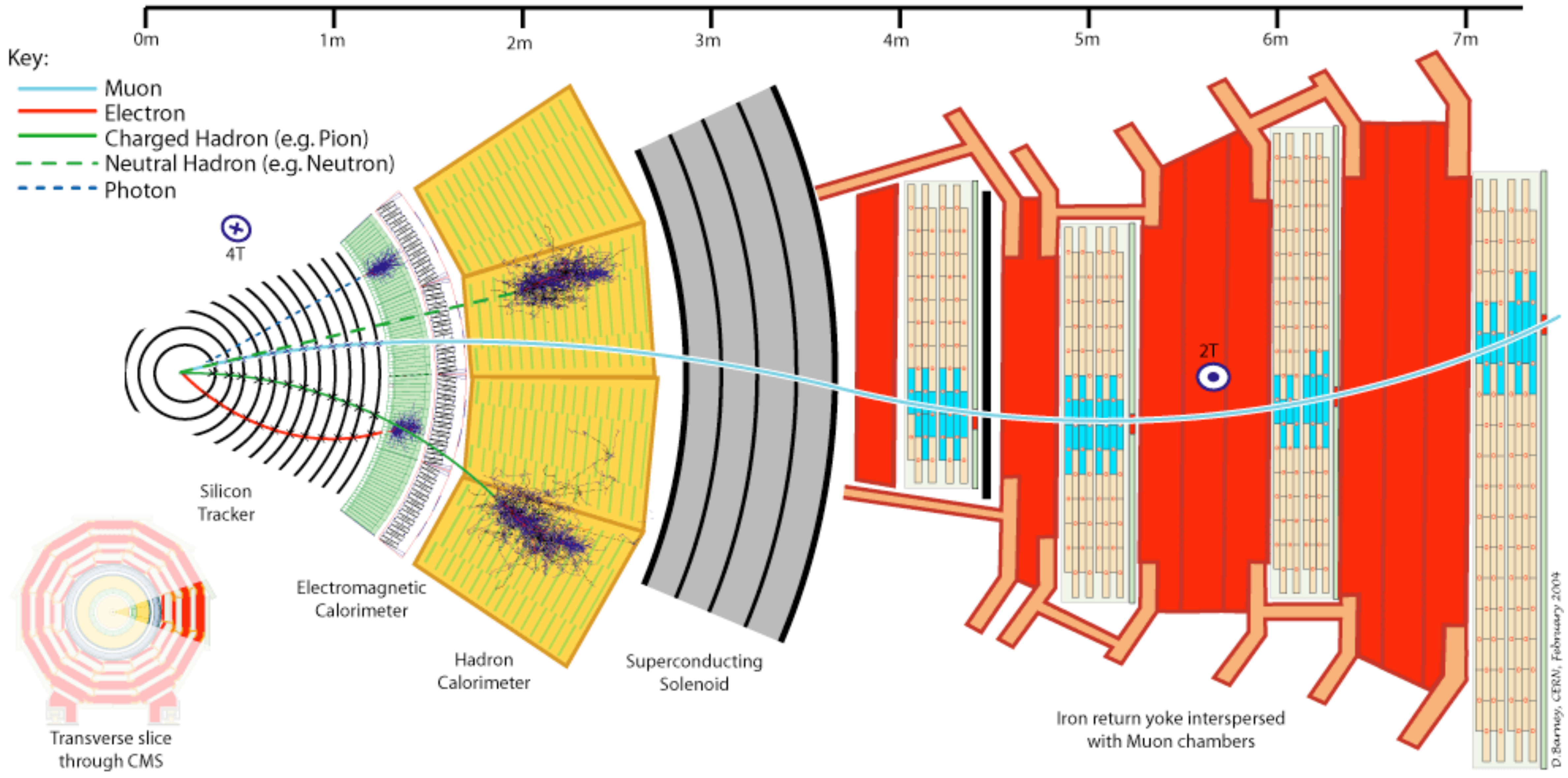
~10-30 features describe interesting part

DETECTOR SIMULATION

Conceptually: Prob(detector response | particles)

Implementation: Monte Carlo integration over micro-physics

Consequence: evaluation of the likelihood is intractable


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

DETECTOR SIMULATION

Conceptually: Prob(detector response | particles)

Implementation: Monte Carlo integration over micro-physics

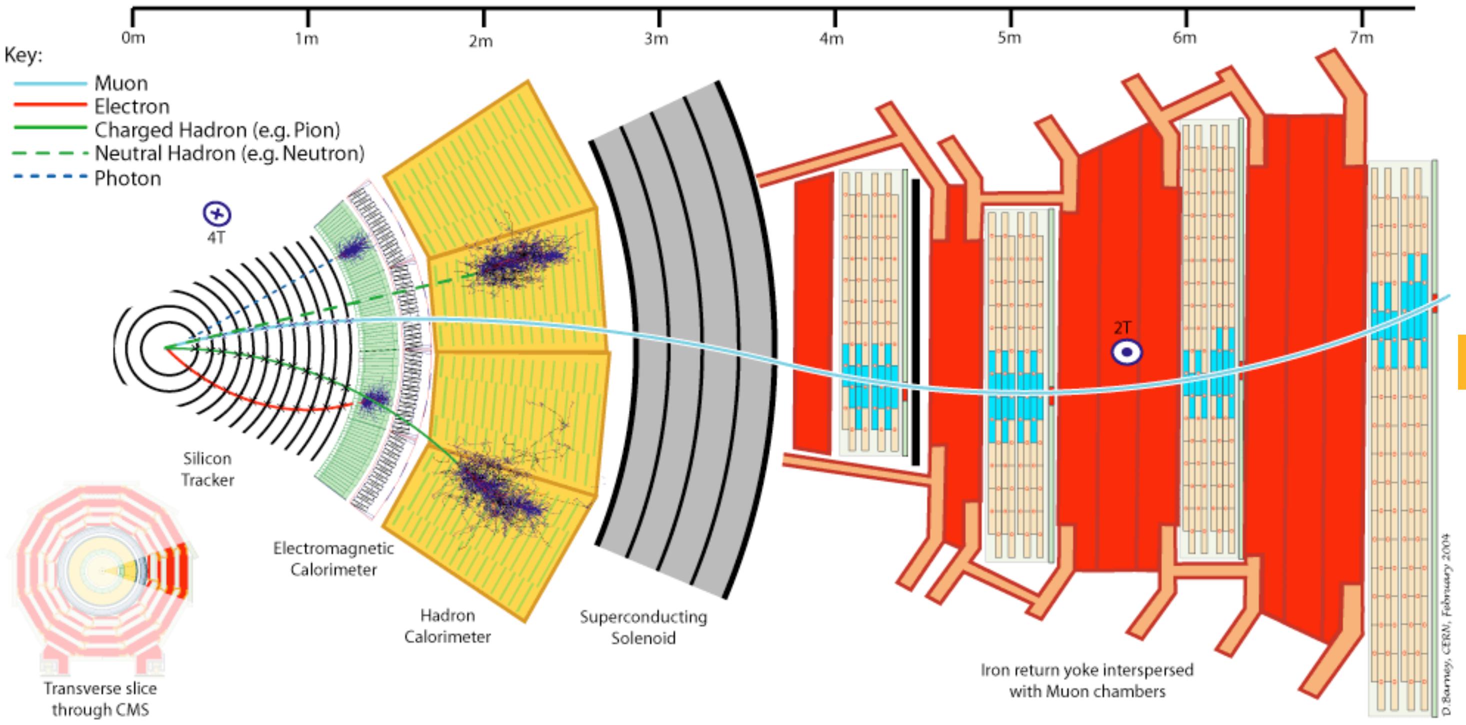
Consequence: evaluation of the likelihood is intractable

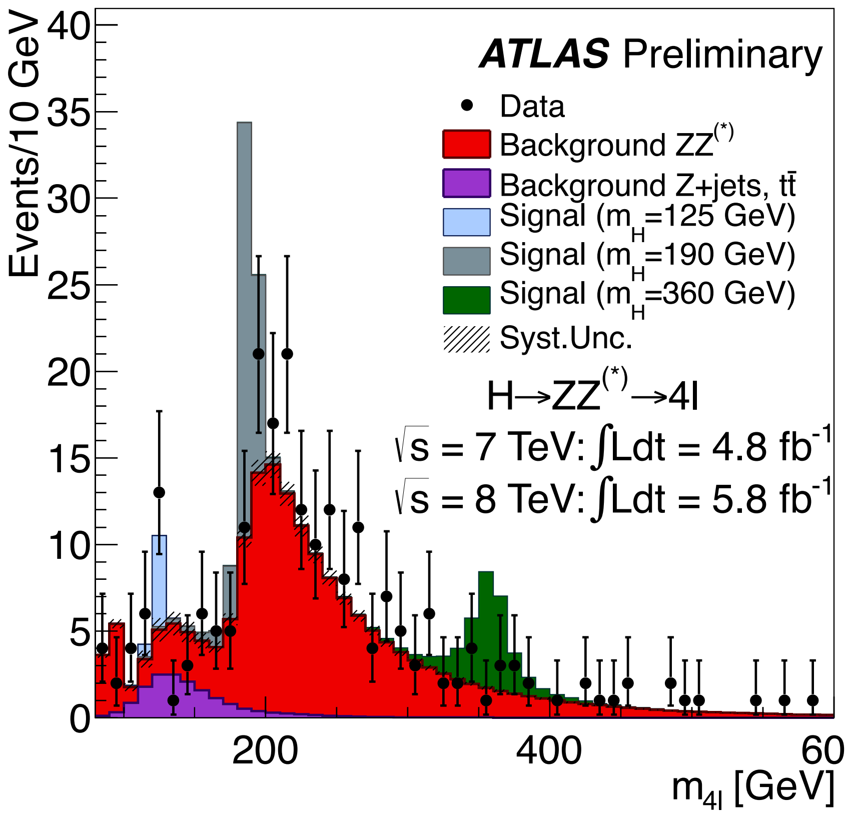
This motivates a new class of algorithms for what is called **likelihood-free inference (or simulation-based inference)**, which only require ability to generate samples from the simulation in the “forward mode”

10^8 SENSORS → 1 REAL-VALUED QUANTITY

Most measurements and searches for new particles at the LHC are based on the distribution of a single **summary statistic s(x)**

- choosing a summary statistic (feature engineering) is a task for a skilled physicist and tailored to the goal of measurement or new particle search
- likelihood p(s|θ) **approximated** using histograms (univariate density estimation)


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004



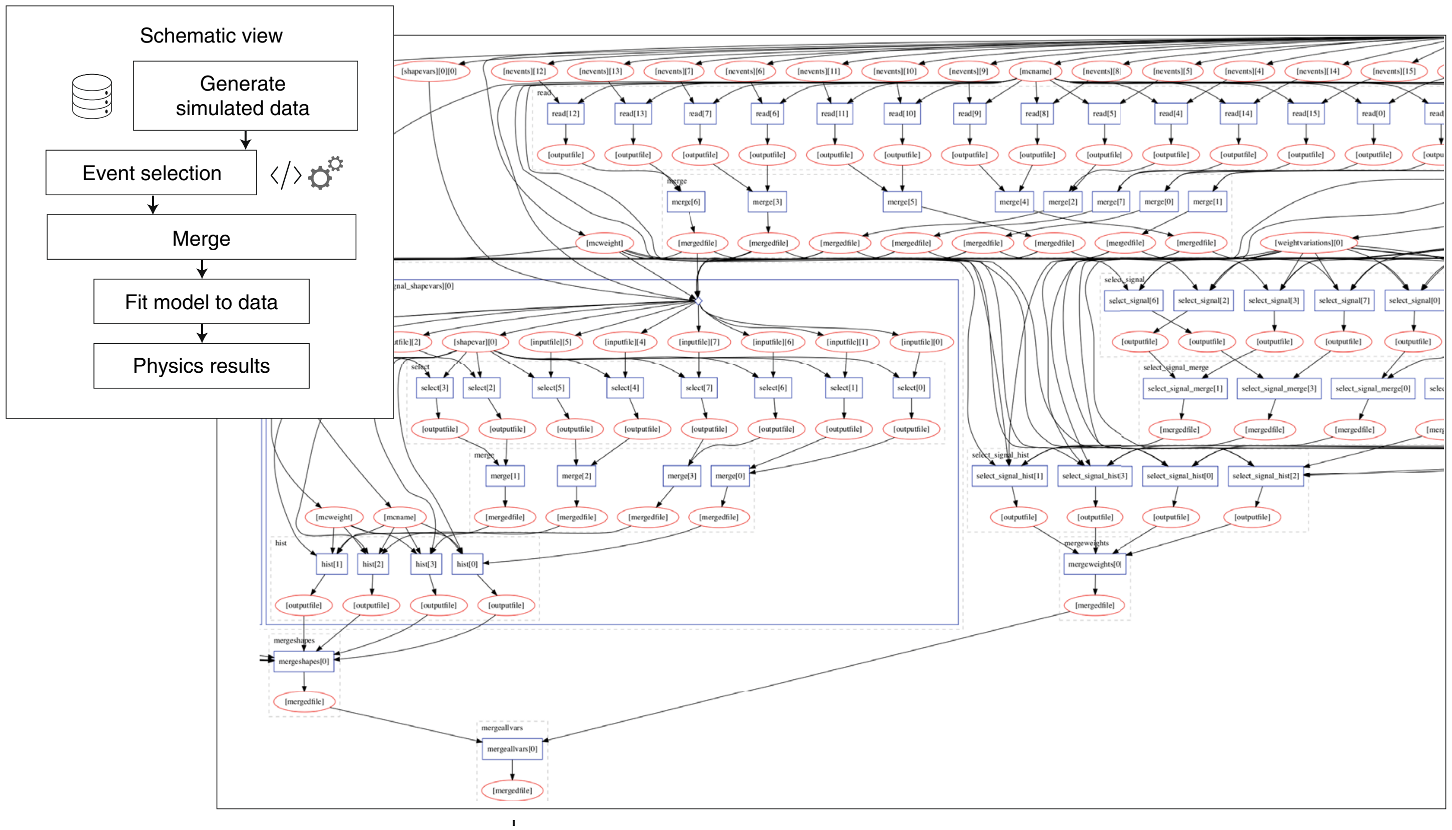
This doesn’t scale if x is high dimensional!

THUMBNAIL SKETCH OF ANALYSIS

In addition to defining a summary statistic, we define a complicated hierarchical filter (binary classifier) **1(x)** that operates on the high-dimensional data **x** to select data that targets a particular alternate hypothesis. most of the work!

- Created by team of people, usually an ad hoc computational "workflow" organized via email and meetings

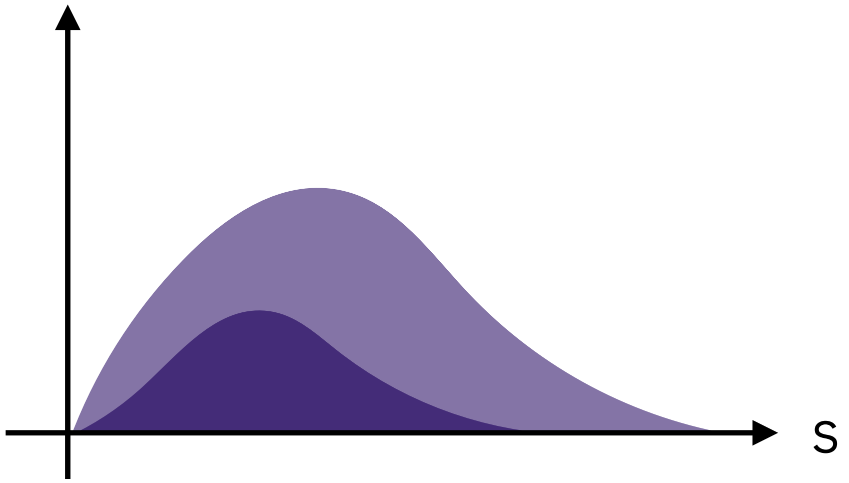
a



THUMBNAIL SKETCH OF ANALYSIS

In addition to defining a summary statistic, we define a complicated hierarchical filter (binary classifier) **1(x)** that operates on the high-dimensional data **x** to select data that targets a particular alternate hypothesis. most of the work!

- Then we run simulated collisions through the pipeline to make the prediction for the null or "background-only" hypothesis

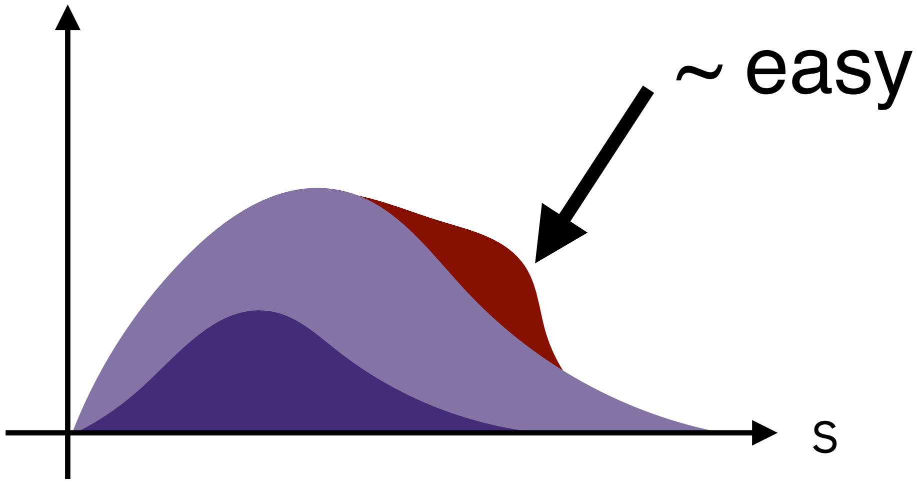


predicted distribution for the null

THUMBNAIL SKETCH OF ANALYSIS

In addition to defining a summary statistic, we define a complicated hierarchical filter (binary classifier) **1(x)** that operates on the high-dimensional data **x** to select data that targets a particular alternate hypothesis.

- Then we run simulated collisions for a hypothetical particle or interaction to make the prediction for alternate or "signal-plus-background" model (a mixture model)

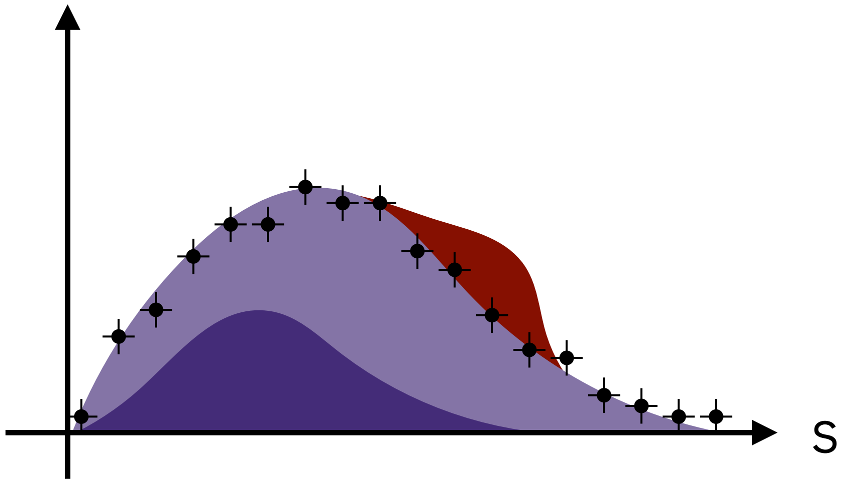


predicted distribution for the alternate in Model A

THUMBNAIL SKETCH OF ANALYSIS

In addition to defining a summary statistic, we define a complicated hierarchical filter (binary classifier) **1(x)** that operates on the high-dimensional data **x** to select data that targets a particular alternate hypothesis.

- Then we add the observed data

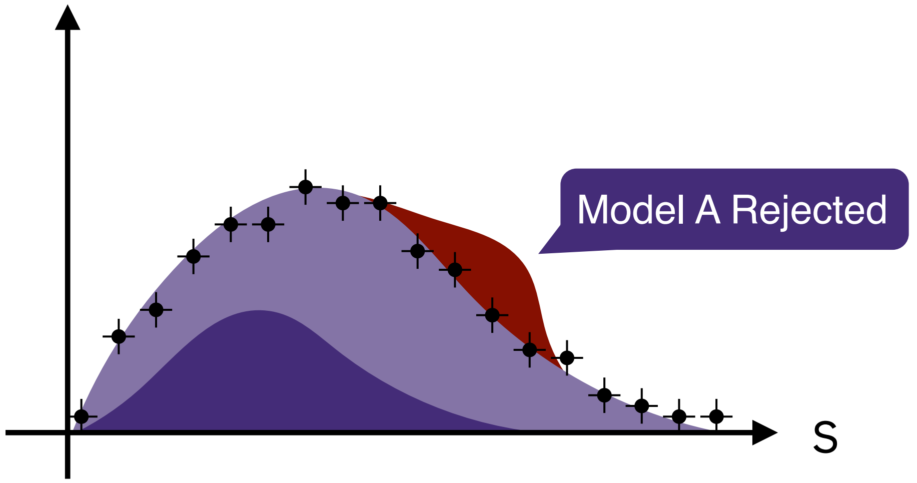


observed **data** + **predicted distribution for the alternate in Model A**

THUMBNAIL SKETCH OF ANALYSIS

In addition to defining a summary statistic, we define a complicated hierarchical filter (binary classifier) **1(x)** that operates on the high-dimensional data **x** to select data that targets a particular alternate hypothesis.

- And finally we test the hypothesis

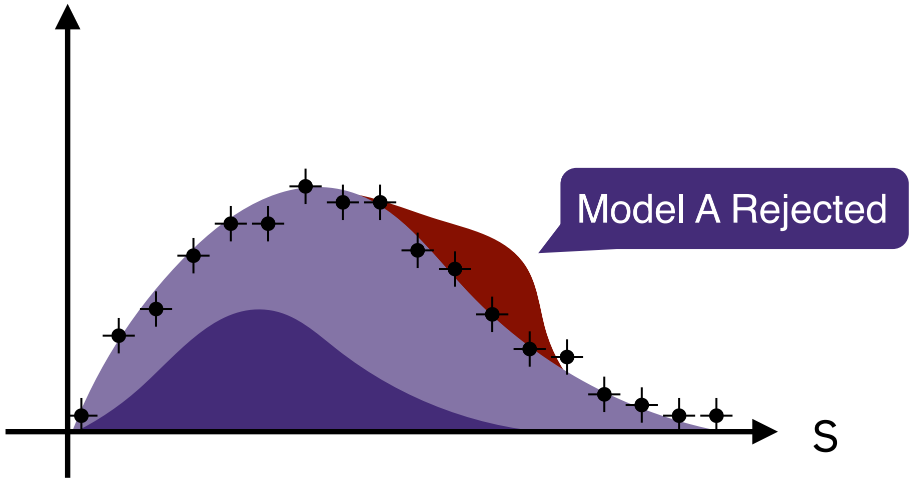


observed **data** + **predicted distribution for the alternate in Model A**

THUMBNAIL SKETCH OF ANALYSIS

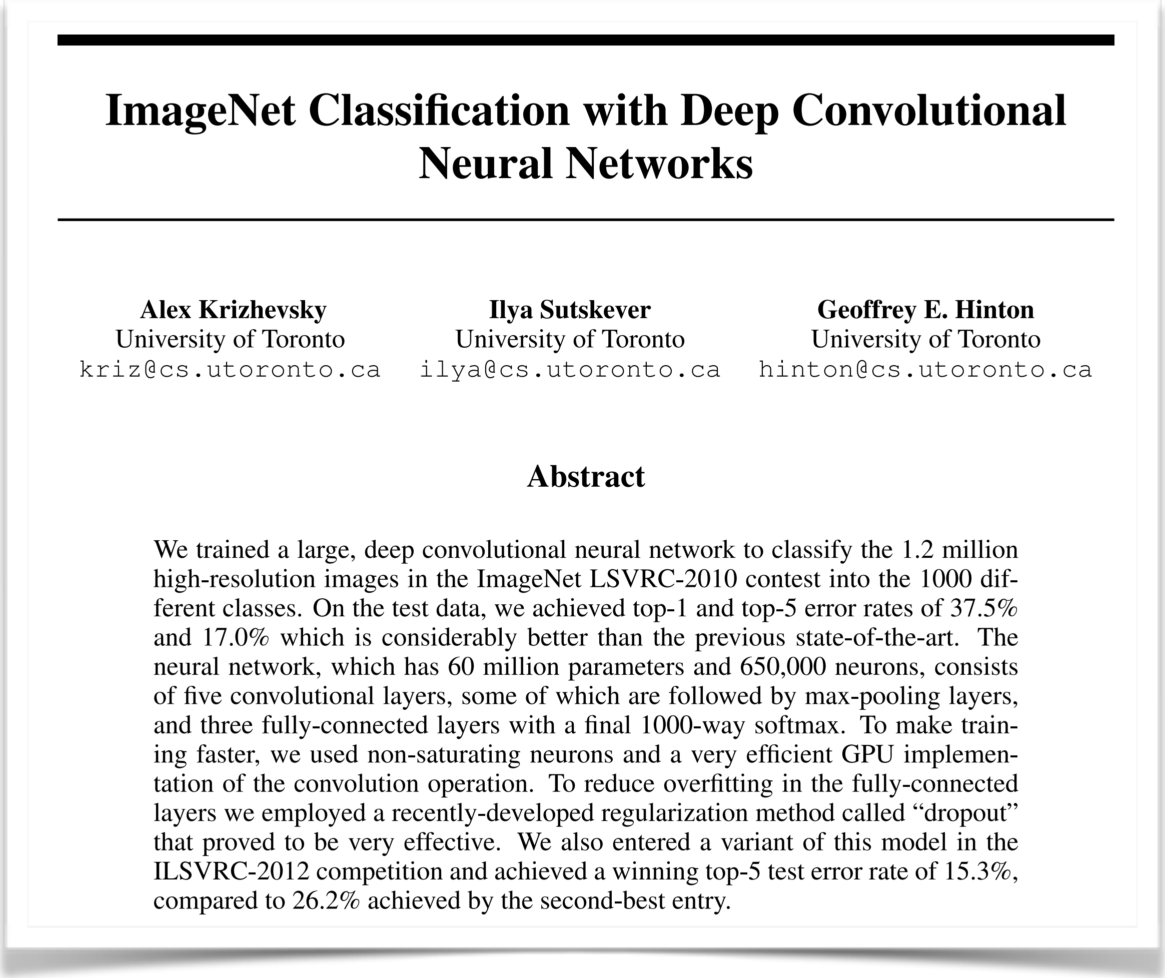
In addition to defining a summary statistic, we define a complicated hierarchical filter (binary classifier) **1(x)** that operates on the high-dimensional data **x** to select data that targets a particular alternate hypothesis.

- And we write a paper, graduate students graduate, code rots, and it would be difficult to reproduce



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

JULY 2012



ImageNet Classification with Deep Convolutional Neural Networks

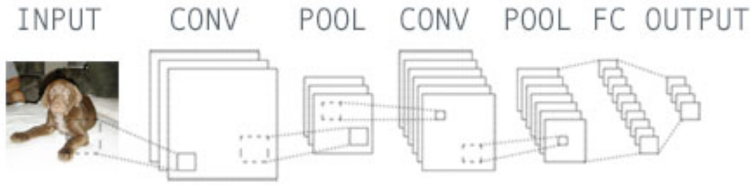
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

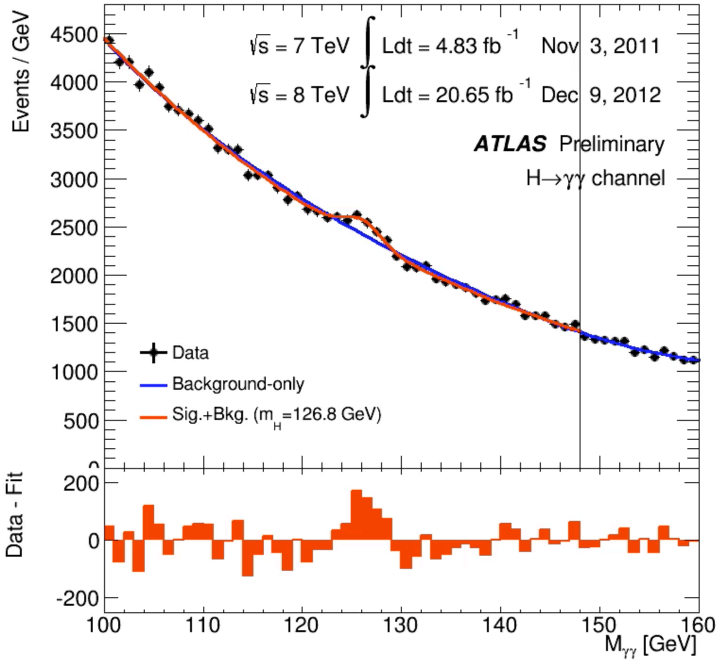
Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

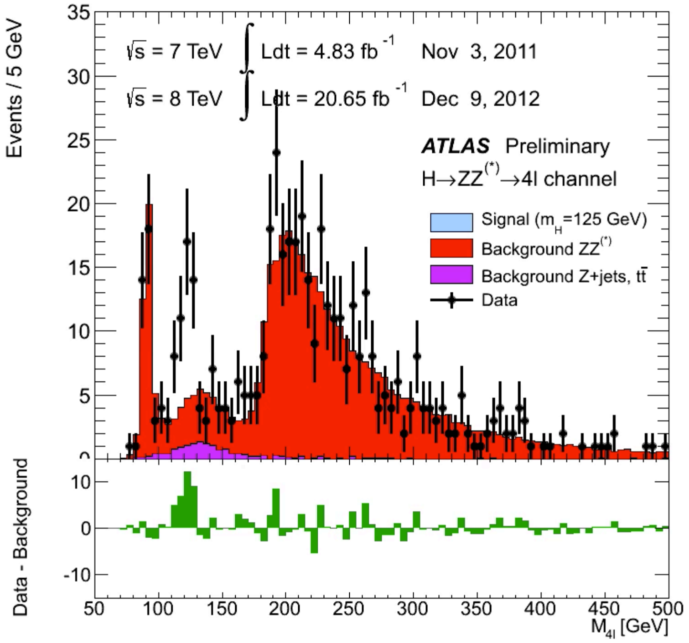
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

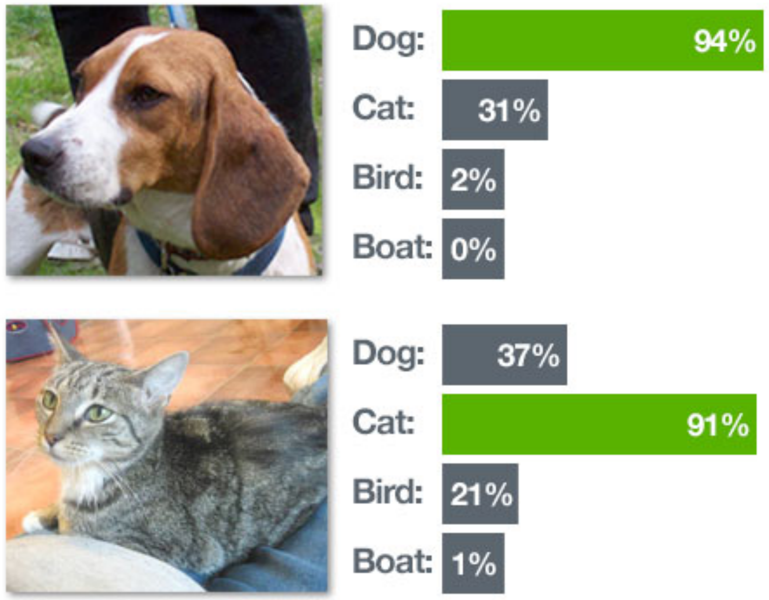
Abstract

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.









RESPONSES / POSSIBLE SOLUTIONS

1) **It's not a problem** now, the experiments are testing all the theoretical models that really matter, people just need to be patient.

> **BUT** maybe there will be some new idea later and we should make sure we **preserve the data** and tools to analyze it.

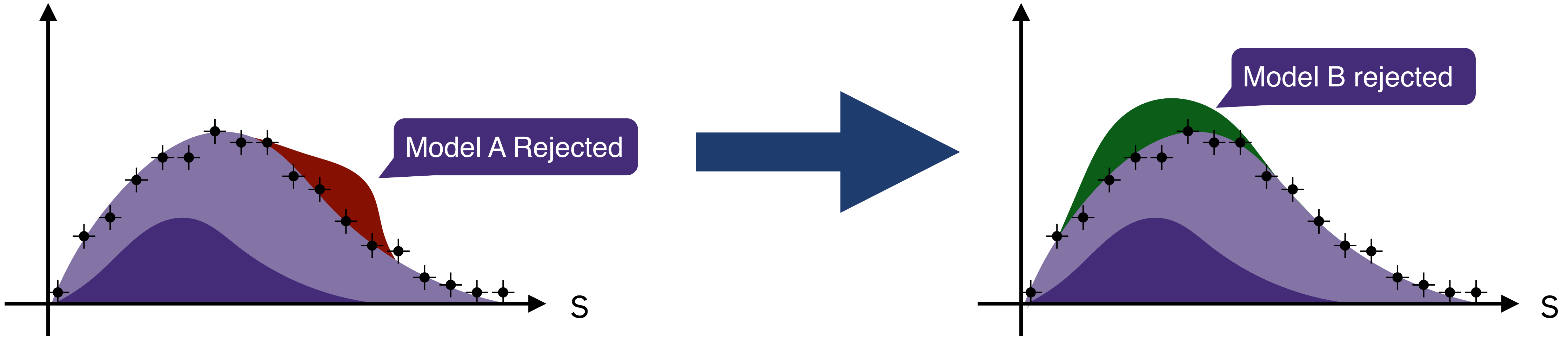
2) **It is a problem**, and the solution is to embrace **Open Data**

3) **It is a problem**, and there is a **technical solution**

REINTERPRETATION

If we can capture the definition of the summary **s(x)** and the event selection filter (binary classifier) **1(x)** 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



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

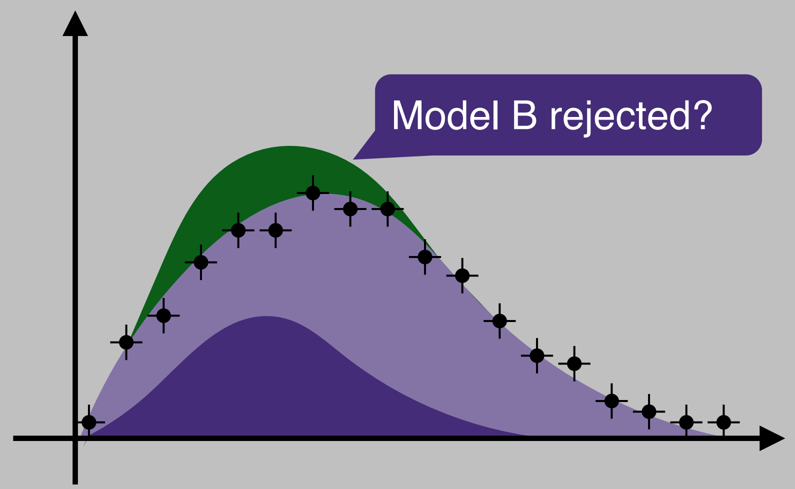
THEORY

SERVICE

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

Q

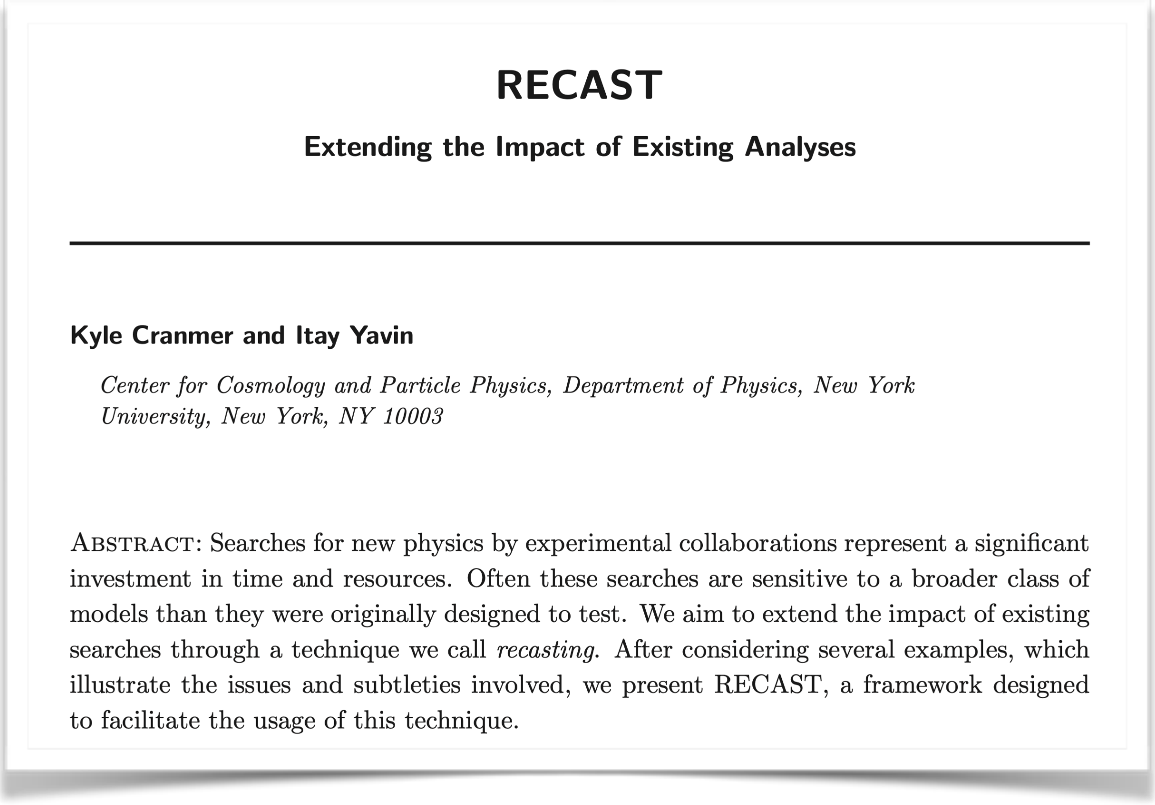
A



We proposed RECAST framework in Oct 2010

- Conservative narrative compared to "open data"
- Not totally general, it targets specific high-value use cases
- Not conservative enough for many. Lots of resistance
- People said it couldn't be done, our workflows are too complicated
- Hard to get effort to work on it.

recast



RECAST

Extending the Impact of Existing Analyses

Kyle Cranmer and Itay Yavin

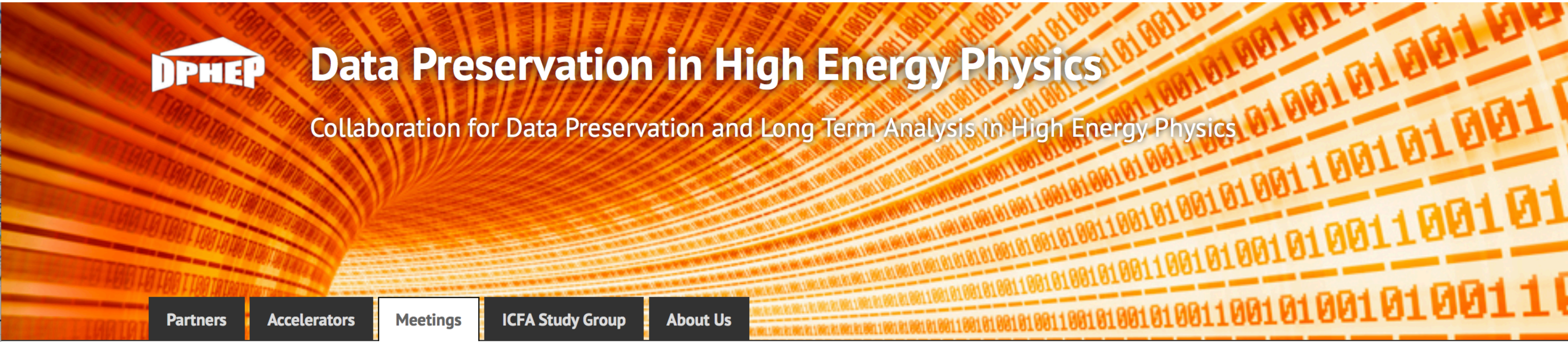
Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003

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.

Orig Proposal in 2010: [arXiv.org:1010.2506]

DATA PRESERVATION

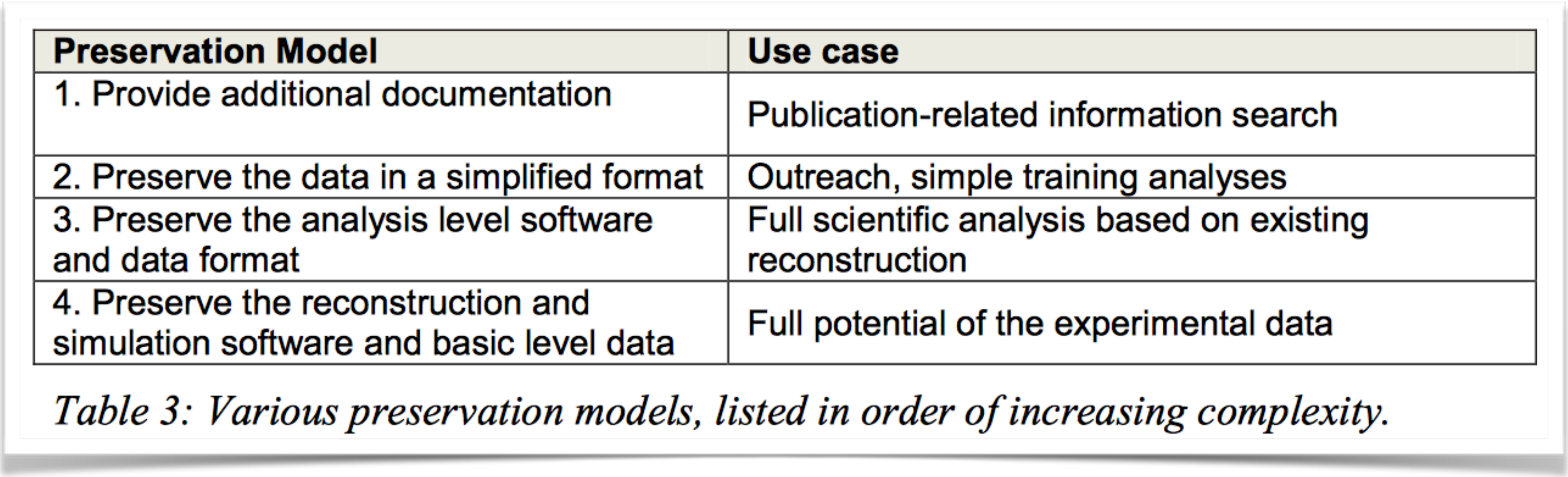
https://arxiv.org/abs/1205.4667



DPHEP

Data Preservation in High Energy Physics

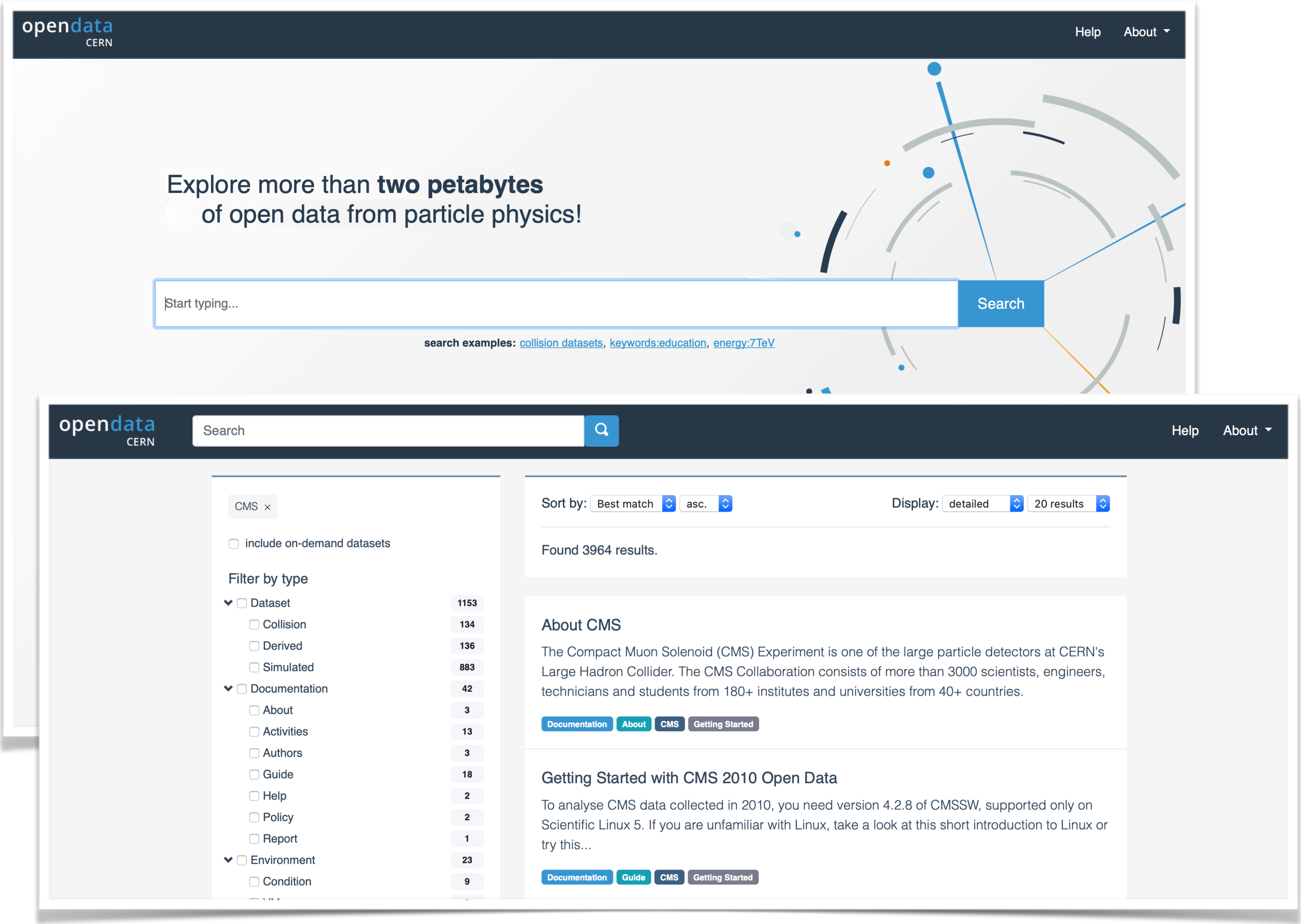
Collaboration for Data Preservation and Long Term Analysis in High Energy Physics



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

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

OPEN DATA



Explore more than **two petabytes** of open data from particle physics!

search examples: collision datasets, keywords:education, energy:7TeV

□ include on-demand datasets

Filter by type

□ Dataset	1153
□ Collision	134
□ Derived	136
□ Simulated	883
□ Documentation	42
□ About	3
□ Activities	13
□ Authors	3
□ Guide	18
□ Help	2
□ Policy	2
□ Report	1
□ Environment	23
□ Condition	9

Sort by: Best match, asc. Display: detailed, 20 results

Found 3964 results.

About CMS

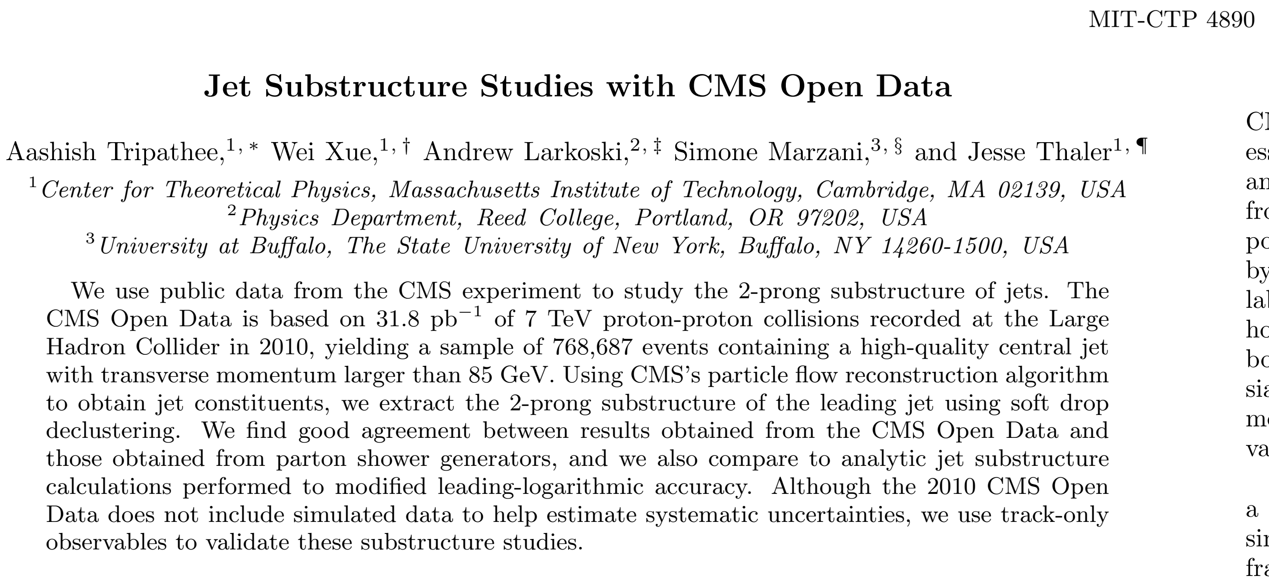
The Compact Muon Solenoid (CMS) Experiment is one of the large particle detectors at CERN's Large Hadron Collider. The CMS Collaboration consists of more than 3000 scientists, engineers, technicians and students from 180+ institutes and universities from 40+ countries.

Getting Started with CMS 2010 Open Data

To analyse CMS data collected in 2010, you need version 4.2.8 of CMSSW, supported only on Scientific Linux 5. If you are unfamiliar with Linux, take a look at this short introduction to Linux or try this...

A CASE STUDY WITH OPEN DATA

https://arxiv.org/abs/1704.05842



MIT-CTP 4890

Jet Substructure Studies with CMS Open Data

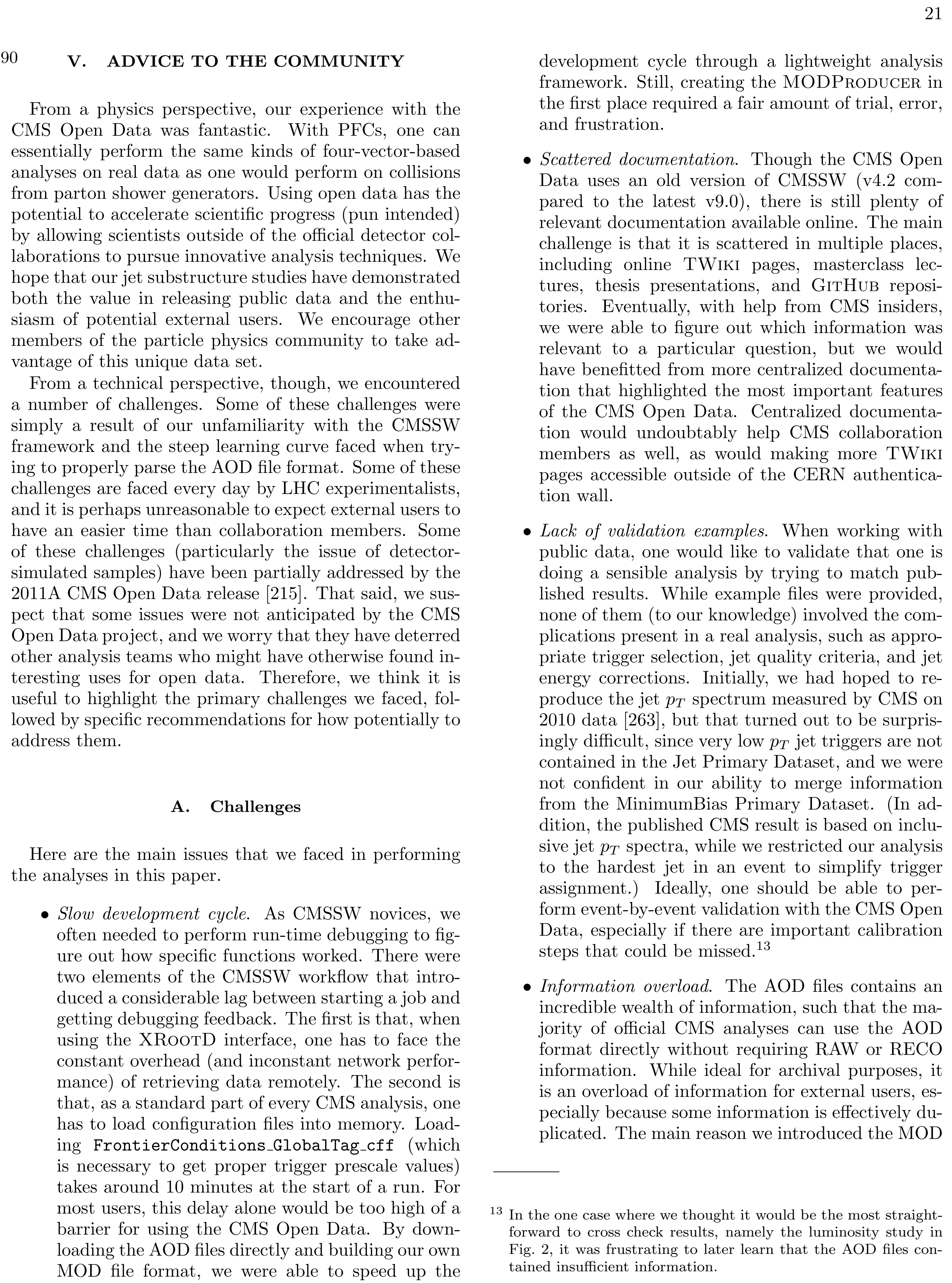
Aashish Tripathee,[1,*] Wei Xue,[1,†] Andrew Larkoski,[2,‡] Simone Marzani,[3,§] and Jesse Thaler[1,¶]

[1] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *Physics Department, Reed College, Portland, OR 97202, USA*
[3] *University at Buffalo, The State University of New York, Buffalo, NY 14260-1500, USA*

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

Challenges:

- Slow development cycle
- Scattered documentation
- Lack of validation examples
- Information Overload



V. ADVICE TO THE COMMUNITY

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

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

A. Challenges

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

- *Slow development cycle.* As CMSSW novices, we often needed to perform run-time debugging to figure out how specific functions worked. There were two elements of the CMSSW workflow that introduced a considerable lag between starting a job and getting debugging feedback. The first is that, when using the XRootD interface, one has to face the constant overhead (and inconstant network performance) of retrieving data remotely. The second is that, as a standard part of every CMS analysis, one has to load configuration files into memory. Loading `FrontierConditions_GlobalTag_cff` (which is necessary to get proper trigger prescale values) takes around 10 minutes at the start of a run. For most users, this delay alone would be too high of a barrier for using the CMS Open Data. By downloading the AOD files directly and building our own MOD file format, we were able to speed up the development cycle through a lightweight analysis framework. Still, creating the MODProducer in the first place required a fair amount of trial, error, and frustration.
- *Scattered documentation.* Though the CMS Open Data uses an old version of CMSSW (v4.2 compared to the latest v9.0), there is still plenty of relevant documentation available online. The main challenge is that it is scattered in multiple places, including online TWiki pages, masterclass lectures, thesis presentations, and GitHub repositories. Eventually, with help from CMS insiders, we were able to figure out which information was relevant to a particular question, but we would have benefitted from more centralized documentation that highlighted the most important features of the CMS Open Data. Centralized documentation would undoubtably help CMS collaboration members as well, as would making more TWiki pages accessible outside of the CERN authentication wall.
- *Lack of validation examples.* When working with public data, one would like to validate that one is doing a sensible analysis by trying to match published results. While example files were provided, none of them (to our knowledge) involved the complications present in a real analysis, such as appropriate trigger selection, jet quality criteria, and jet energy corrections. Initially, we had hoped to reproduce the jet p_T spectrum measured by CMS on 2010 data [263], but that turned out to be surprisingly difficult, since very low p_T jet triggers are not contained in the Jet Primary Dataset, and we were not confident in our ability to merge information from the MinimumBias Primary Dataset. (In addition, the published CMS result is based on inclusive jet p_T spectra, while we restricted our analysis to the hardest jet in an event to simplify trigger assignment.) Ideally, one should be able to perform event-by-event validation with the CMS Open Data, especially if there are important calibration steps that could be missed.[13]
- *Information overload.* The AOD files contains an incredible wealth of information, such that the majority of official CMS analyses can use the AOD format directly without requiring RAW or RECO information. While ideal for archival purposes, it is an overload of information for external users, especially because some information is effectively duplicated. The main reason we introduced the MOD

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

BUILD IT AND THEY WILL COME

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

- Containers & Cloud technology
- Small amount of support from NSF
- 9 years later …

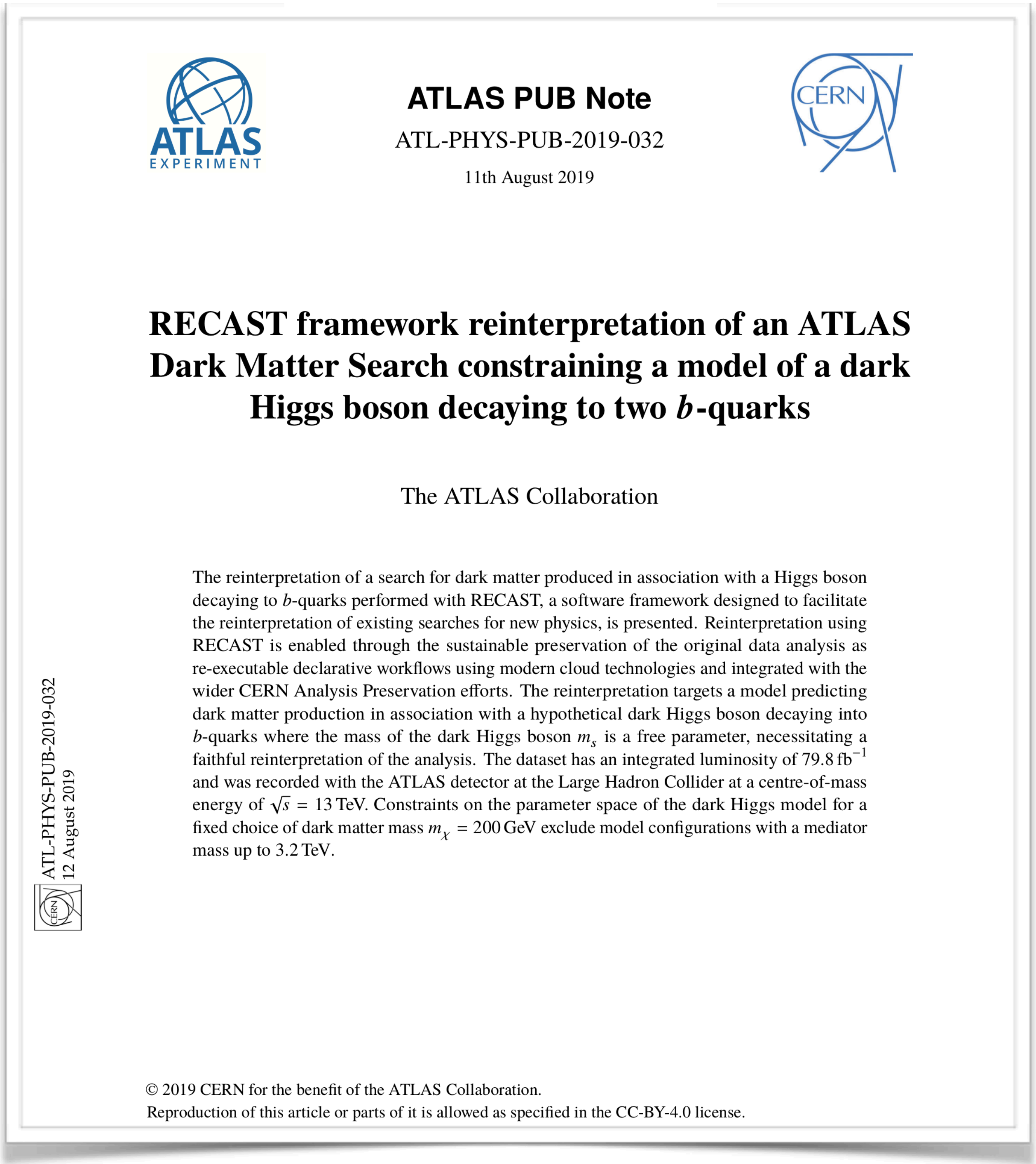
Reproducibility is a byproduct!

Reuse provides a forward-looking narrative, while reproducibility often perceived as backward-looking

Analysis Preservation distinct from reproducibility

recast

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

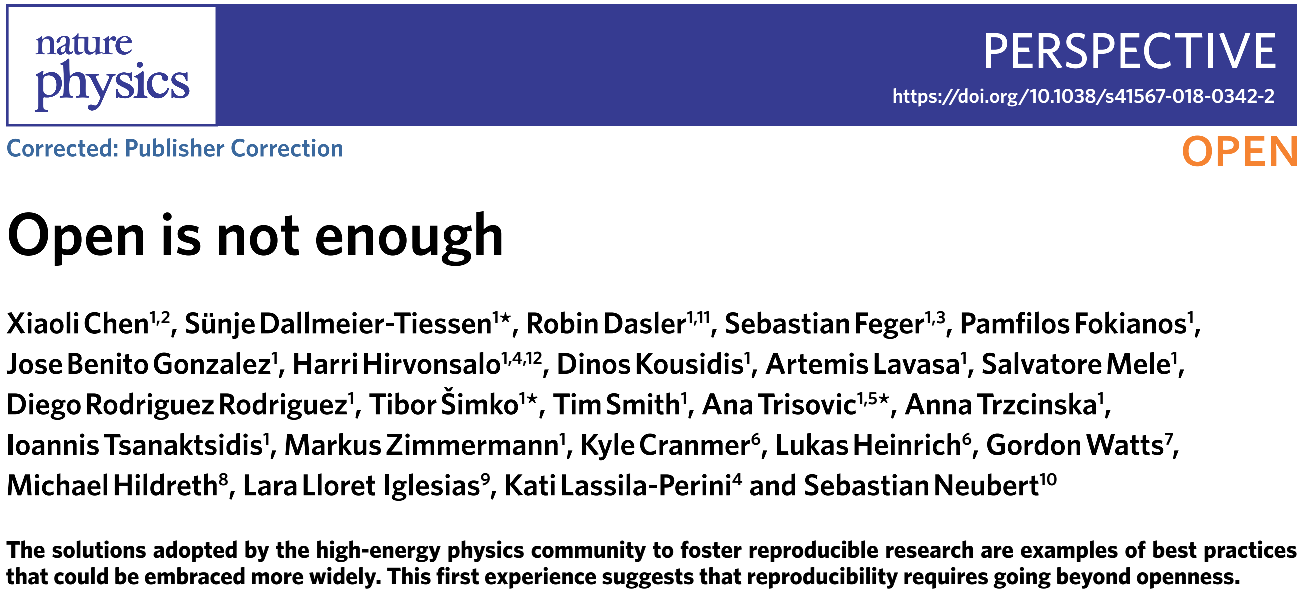
The ATLAS Collaboration

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 TeV. Constraints on the parameter space of the dark Higgs model for a fixed choice of dark matter mass m_χ = 200 GeV exclude model configurations with a mediator mass up to 3.2 TeV.

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

Orig Proposal in 2010: [arXiv.org:1010.2506]

SHIFTING FROM REPRODUCIBILITY TO REUSE



nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-018-0342-2

Corrected: Publisher Correction

OPEN

Open is not enough

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

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

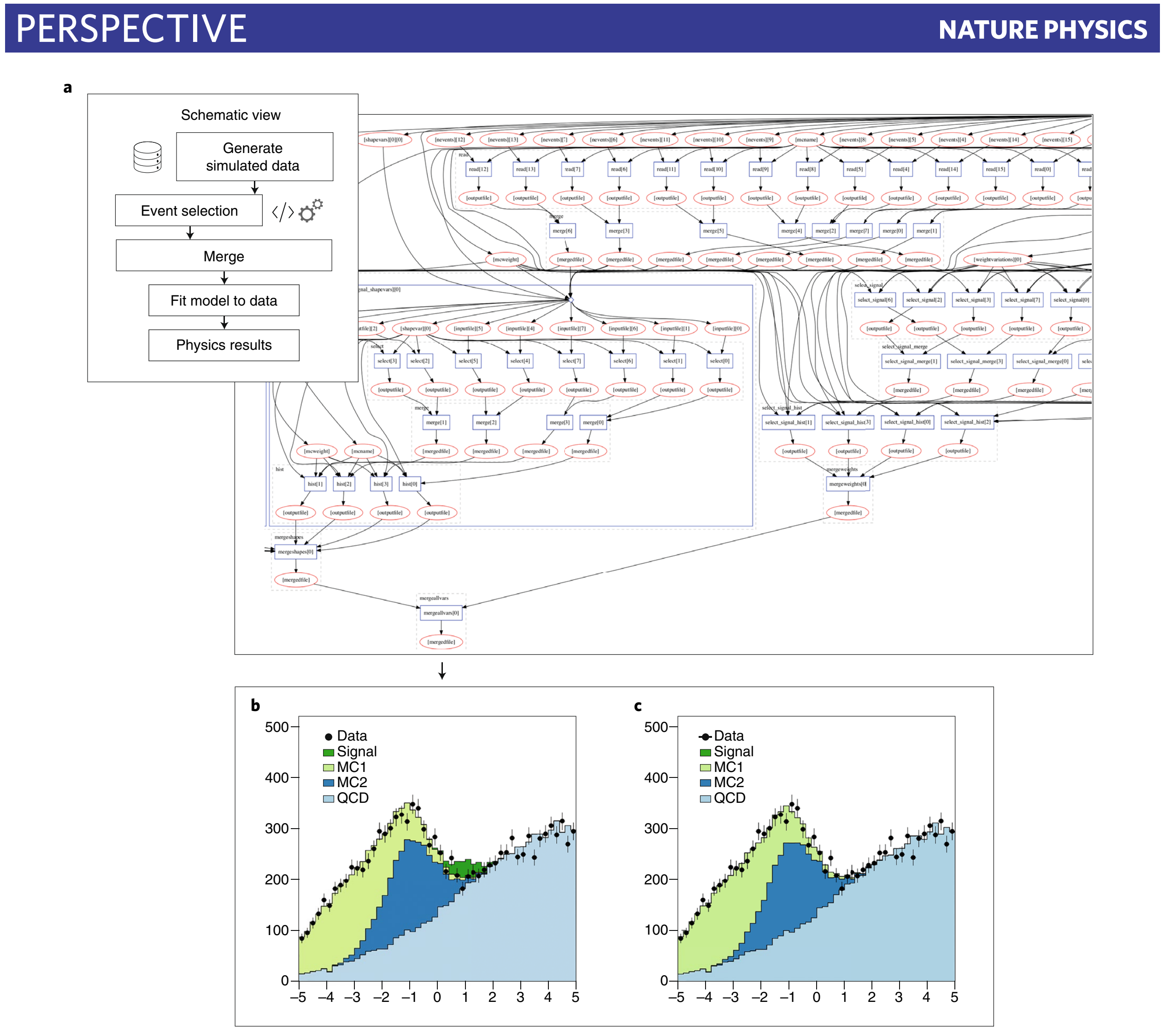


Fig. 2 | Example of a complex computational workflow on REANA mimicking a beyond the standard model (BSM) analysis . This figure shows an example where the experimental data is compared to the predictions of the standard model with an additional hypothesized signal component. The example permits one 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.

https://doi.org/10.1038/s41567-018-0342-2 http://reanahub.io

reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.

COMMON WORKFLOW LANGUAGE

Scalable

Support for remote compute clouds.

kubernetes

Reusable

Containerise once, reuse elsewhere. Cloud-native.

Free

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

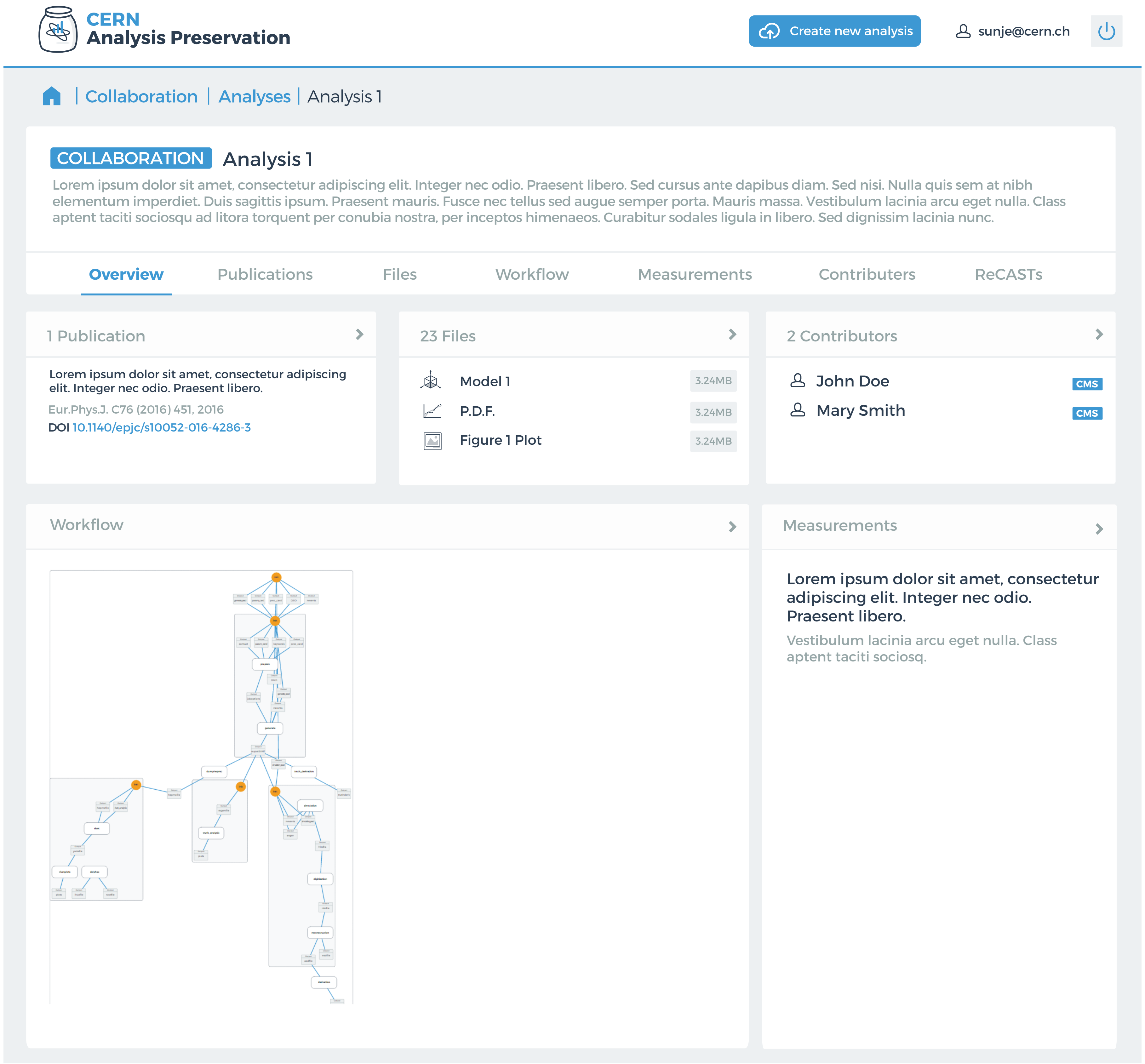
CERN

NSF

dianahep

iris hep
Institute for Research and Innovation in Software for High Energy Physics

The SCAILFIN Project
scailfin.github.io



Collaboration | Analyses | Analysis 1

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.

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

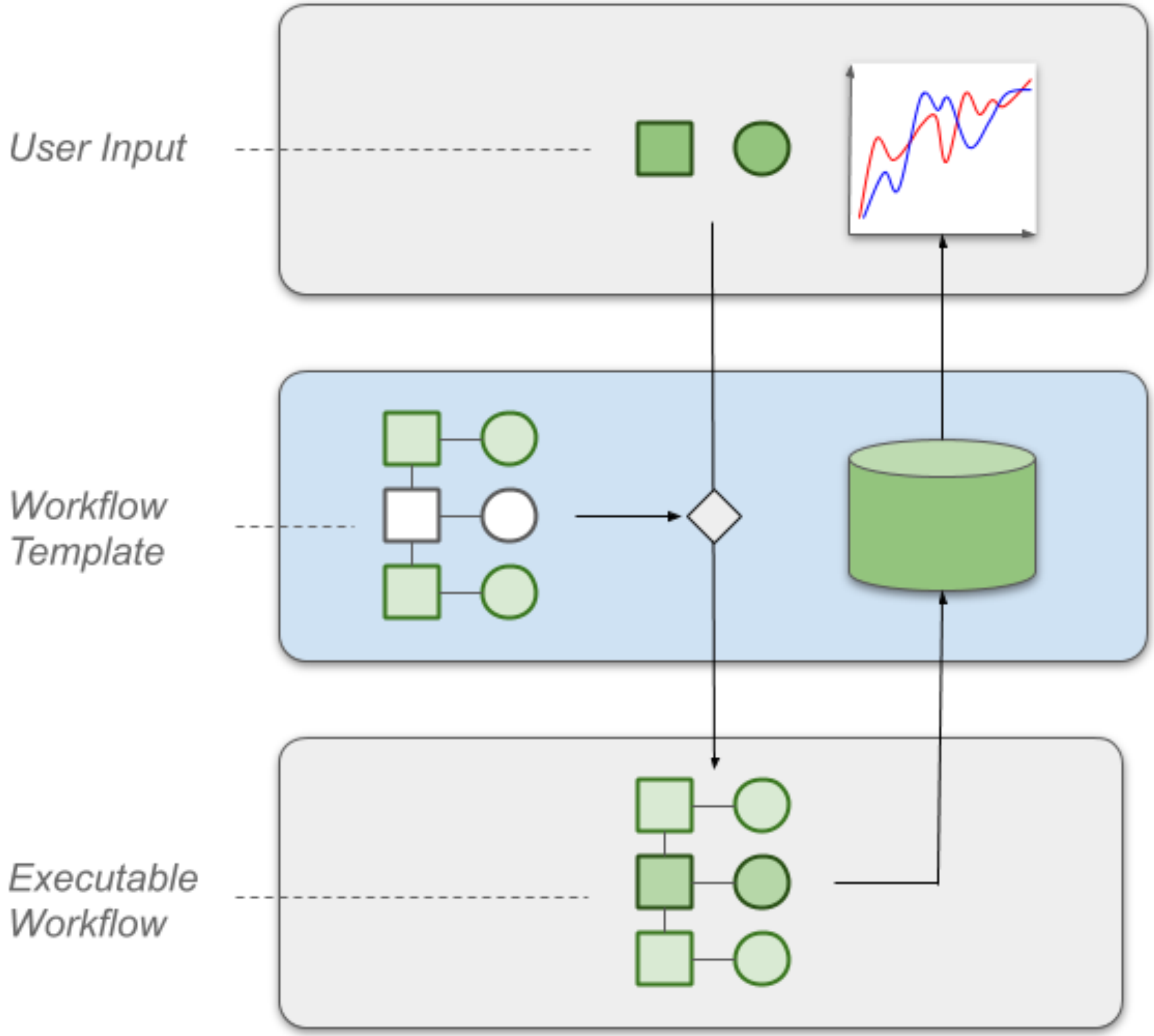
Measurements

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer nec odio. Praesent libero.

Vestibulum lacinia arcu eget nulla. Class aptent taciti sociosq.

CERN
Analysis Preservation

Generalize pattern: domain-specific User Interface around common database of workflows with some backend



OB

Reproducible
Open Benchmarks

flowServ

Back-End

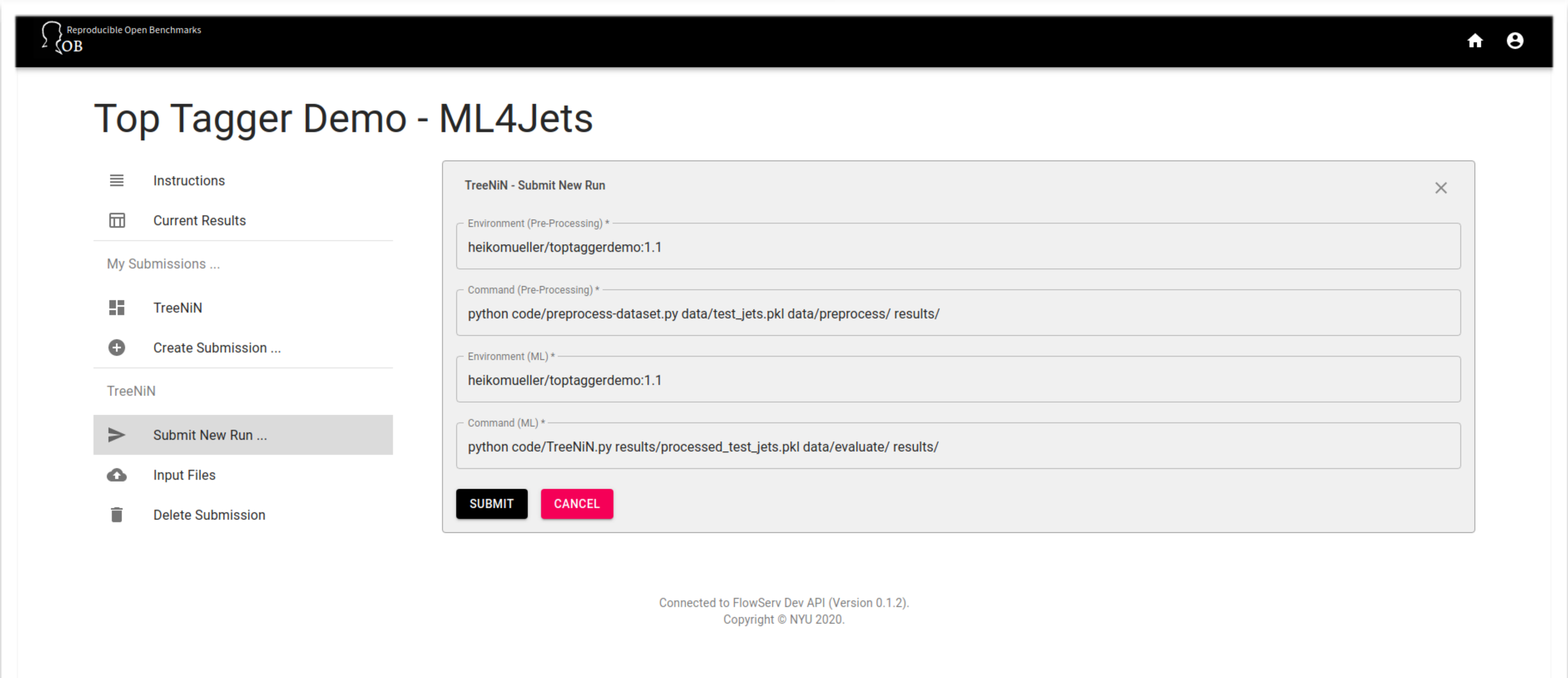
reana …

https://github.com/scailfin/flowserv-core

Reproducible Open Benchmarks

Web User Interface

Automatically renders forms for empty components of workflow



Top Tagger Demo - ML4Jets

TreeNiN - Submit New Run

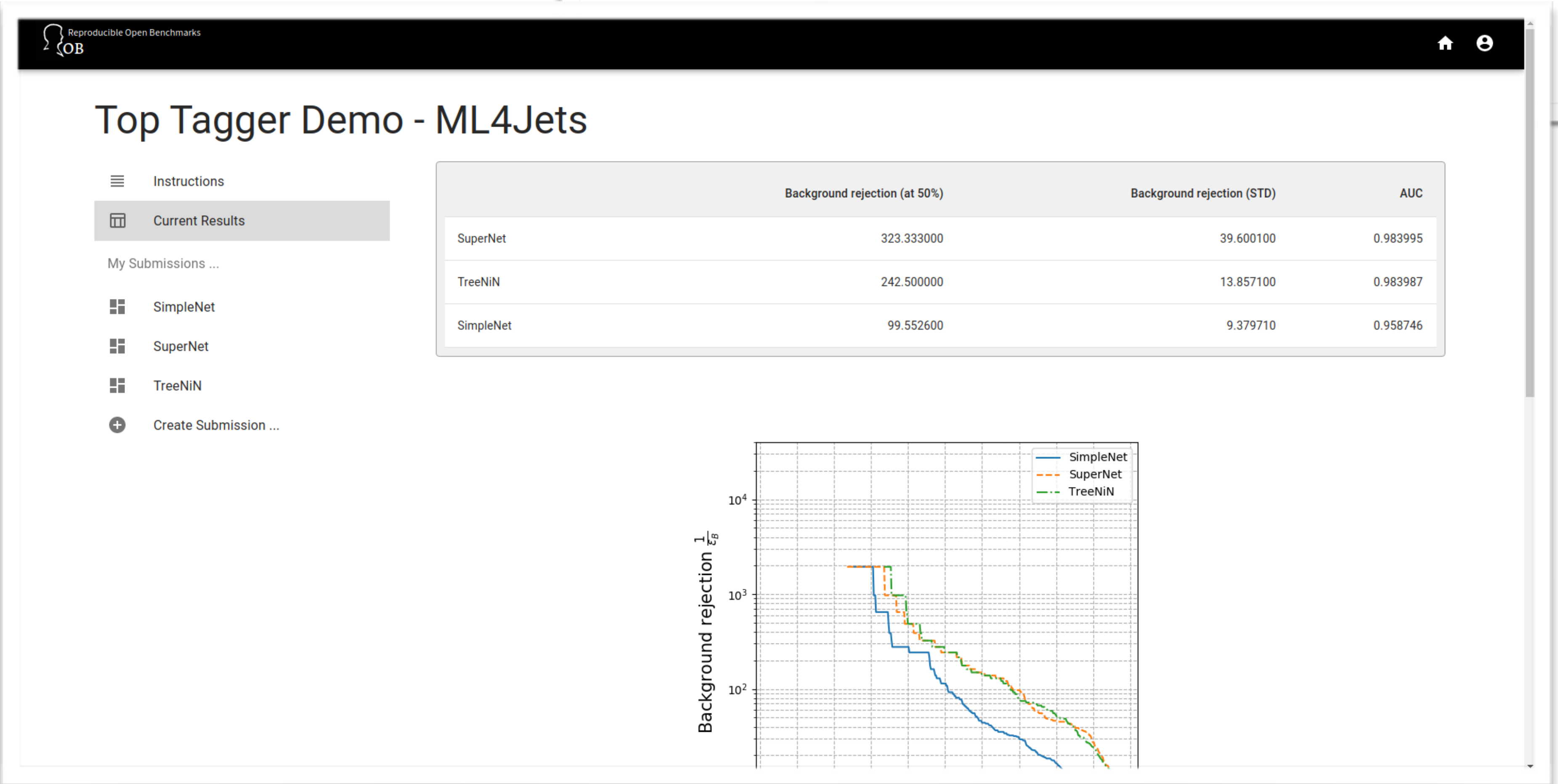
Environment (Pre-Processing) *: heikomueller/toptaggerdemo:1.1

Command (Pre-Processing) *: python code/preprocess-dataset.py data/test_jets.pkl data/preprocess/ results/

Environment (ML) *: heikomueller/toptaggerdemo:1.1

Command (ML) *: python code/TreeNiN.py results/processed_test_jets.pkl data/evaluate/ results/

Connected to FlowServ Dev API (Version 0.1.2).
Copyright © NYU 2020.



Top Tagger Demo - ML4Jets

	Background rejection (at 50%)	Background rejection (STD)	AUC
SuperNet	323.333000	39.600100	0.983995
TreeNiN	242.500000	13.857100	0.983987
SimpleNet	99.552600	9.379710	0.958746

The SCAILFIN Project

scailfin.github.io

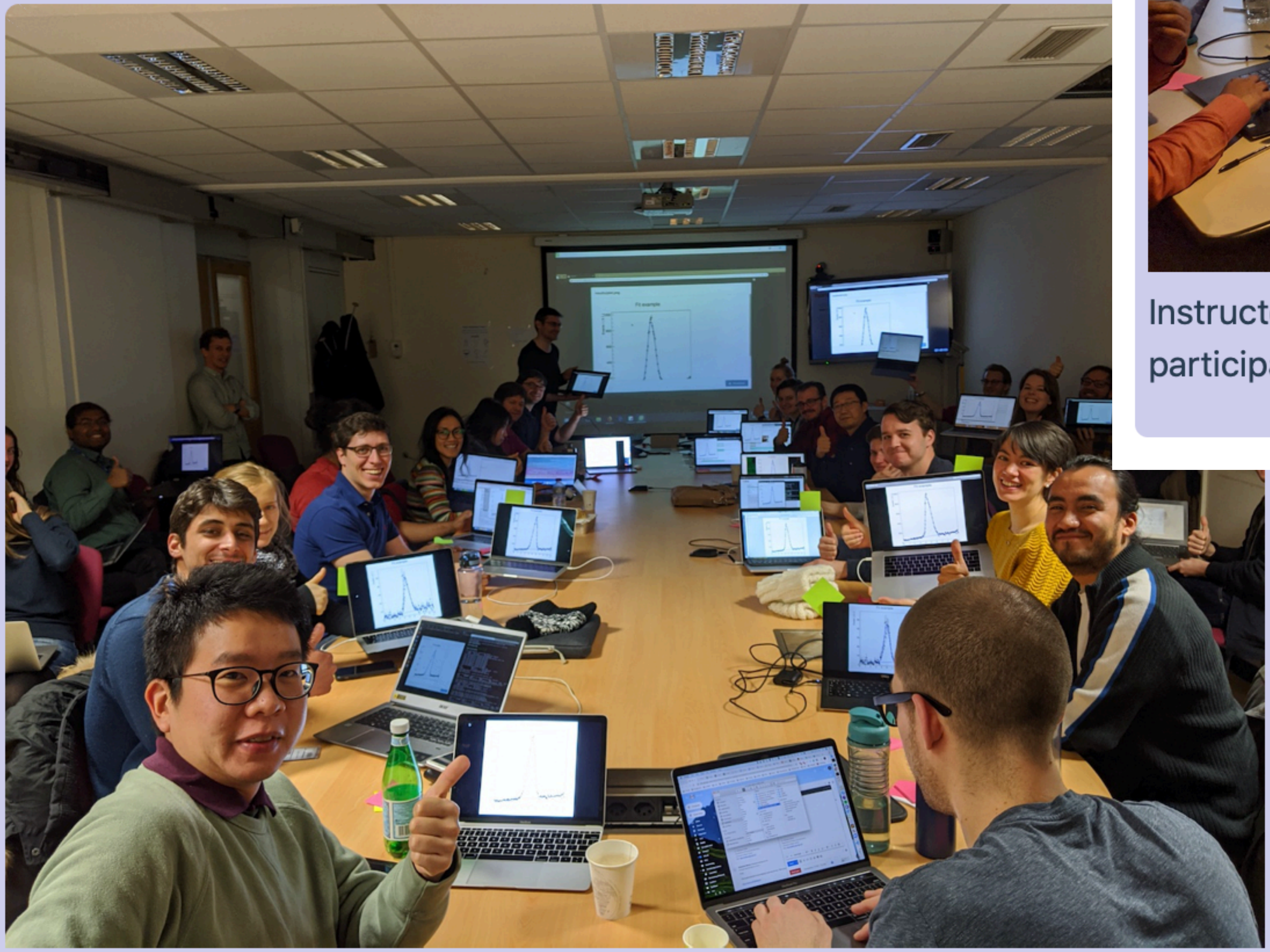
Heiko Müller

Sebastian Macaluso

TRAINING

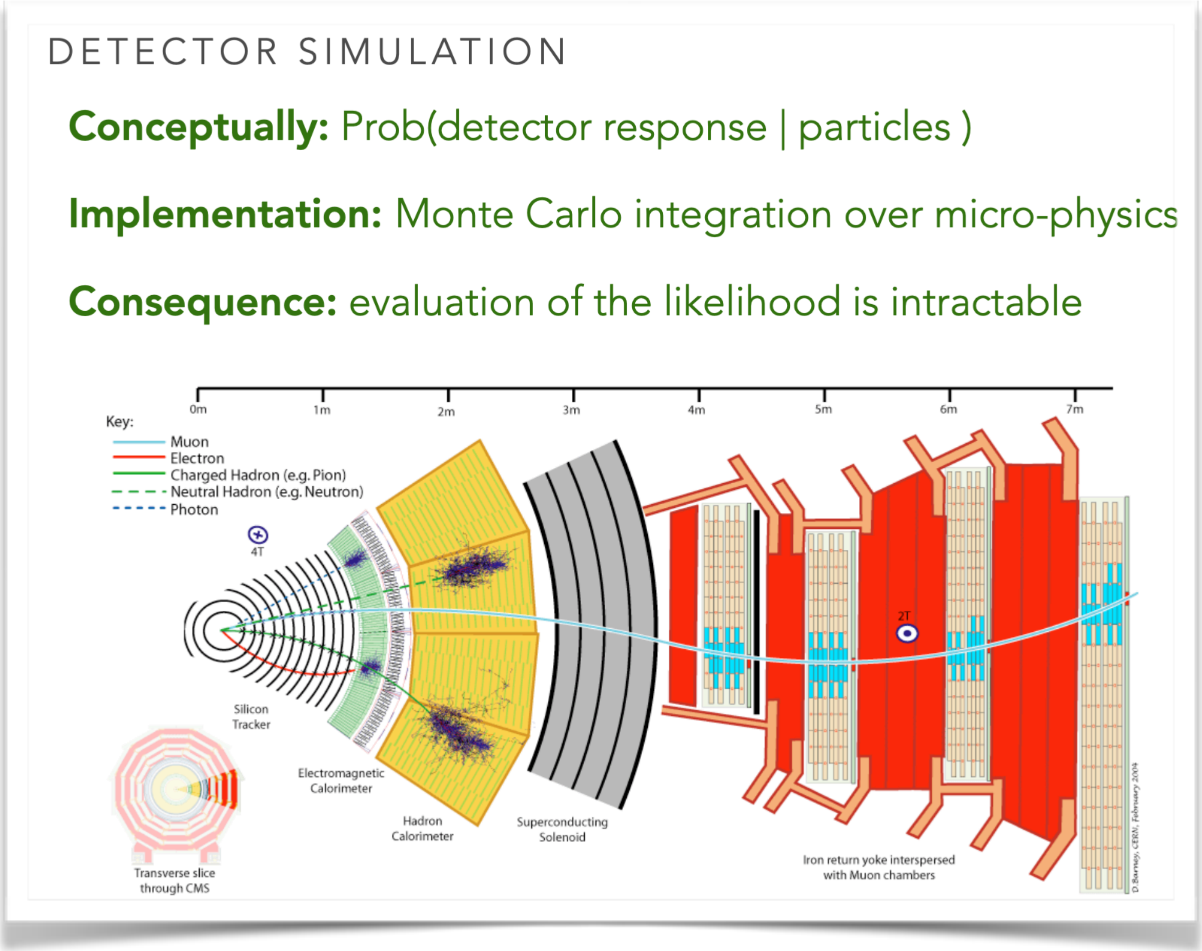
Encouraging response by the community

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

Rewind…



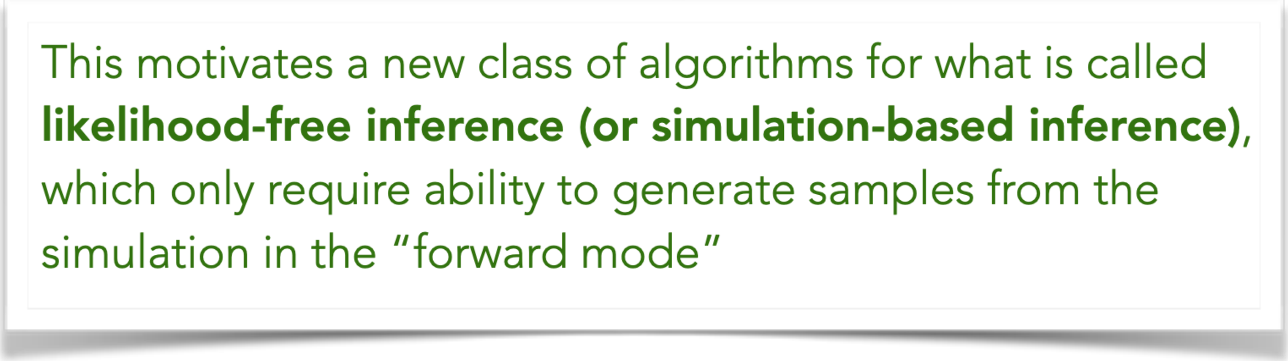
DETECTOR SIMULATION

Conceptually: Prob(detector response | particles)

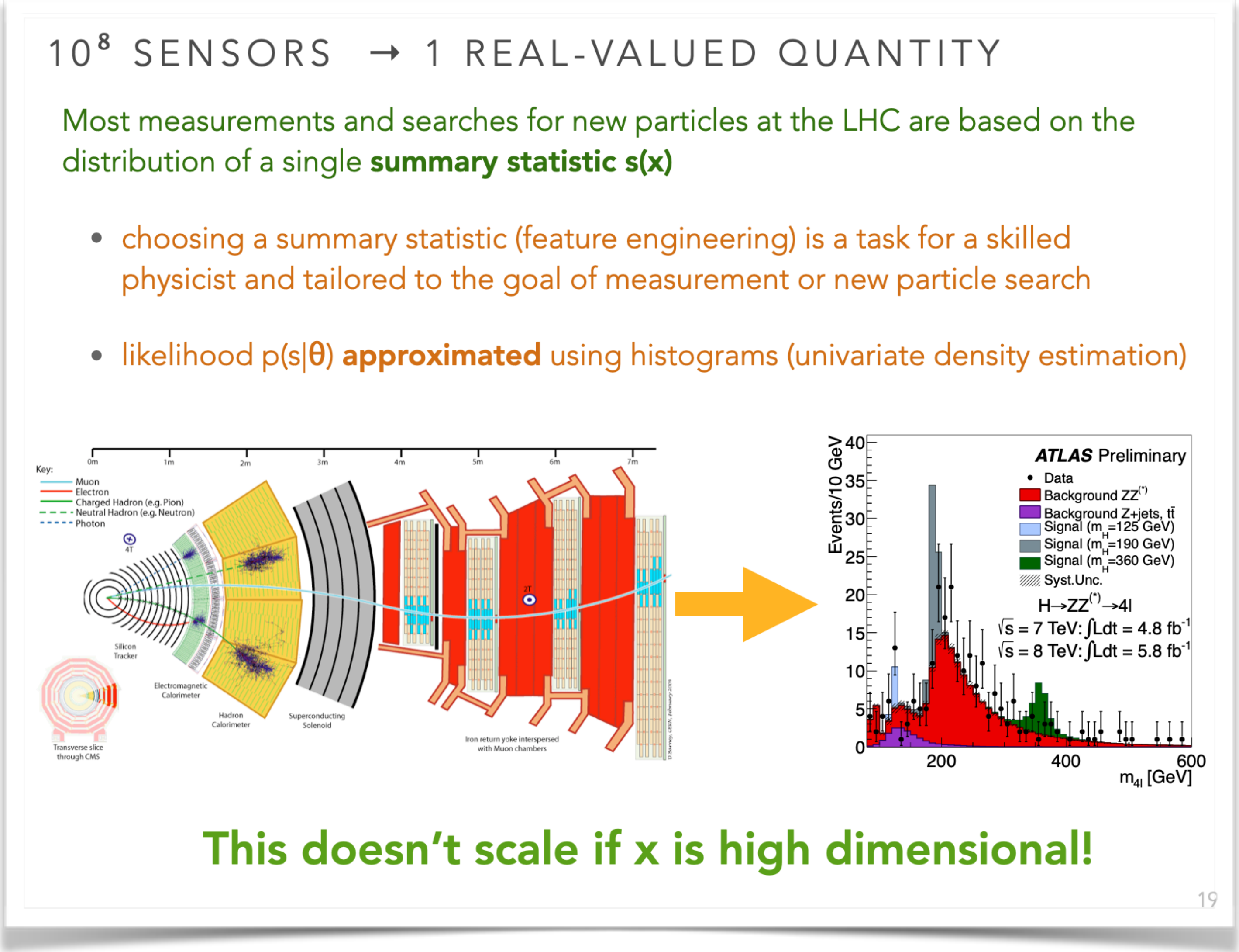
Implementation: Monte Carlo integration over micro-physics

Consequence: evaluation of the likelihood is intractable

Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
0m
1m
2m
3m
4m
5m
6m
7m
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D. Barney, CERN, February 2004



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 → 1 REAL-VALUED QUANTITY

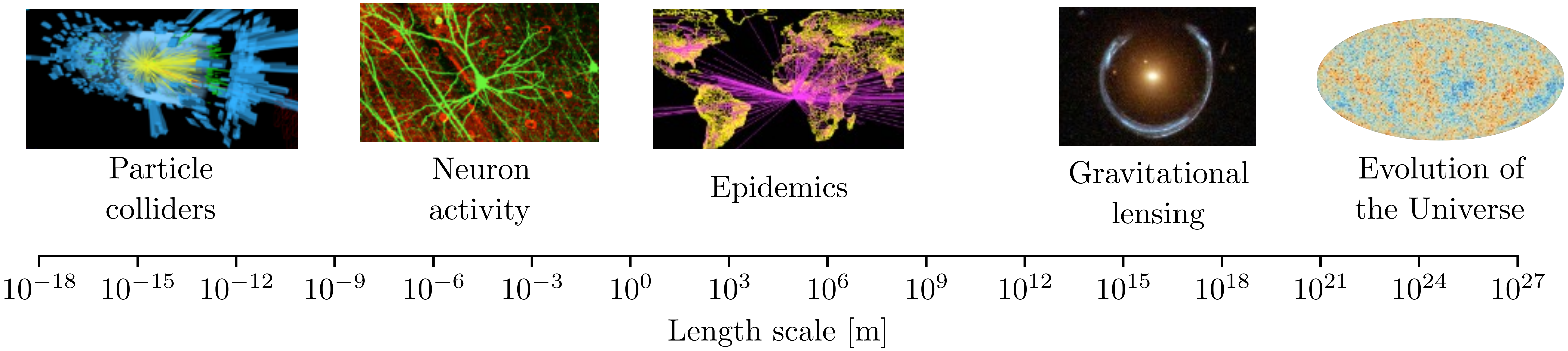
Most measurements and searches for new particles at the LHC are based on the distribution of a single **summary statistic s(x)**

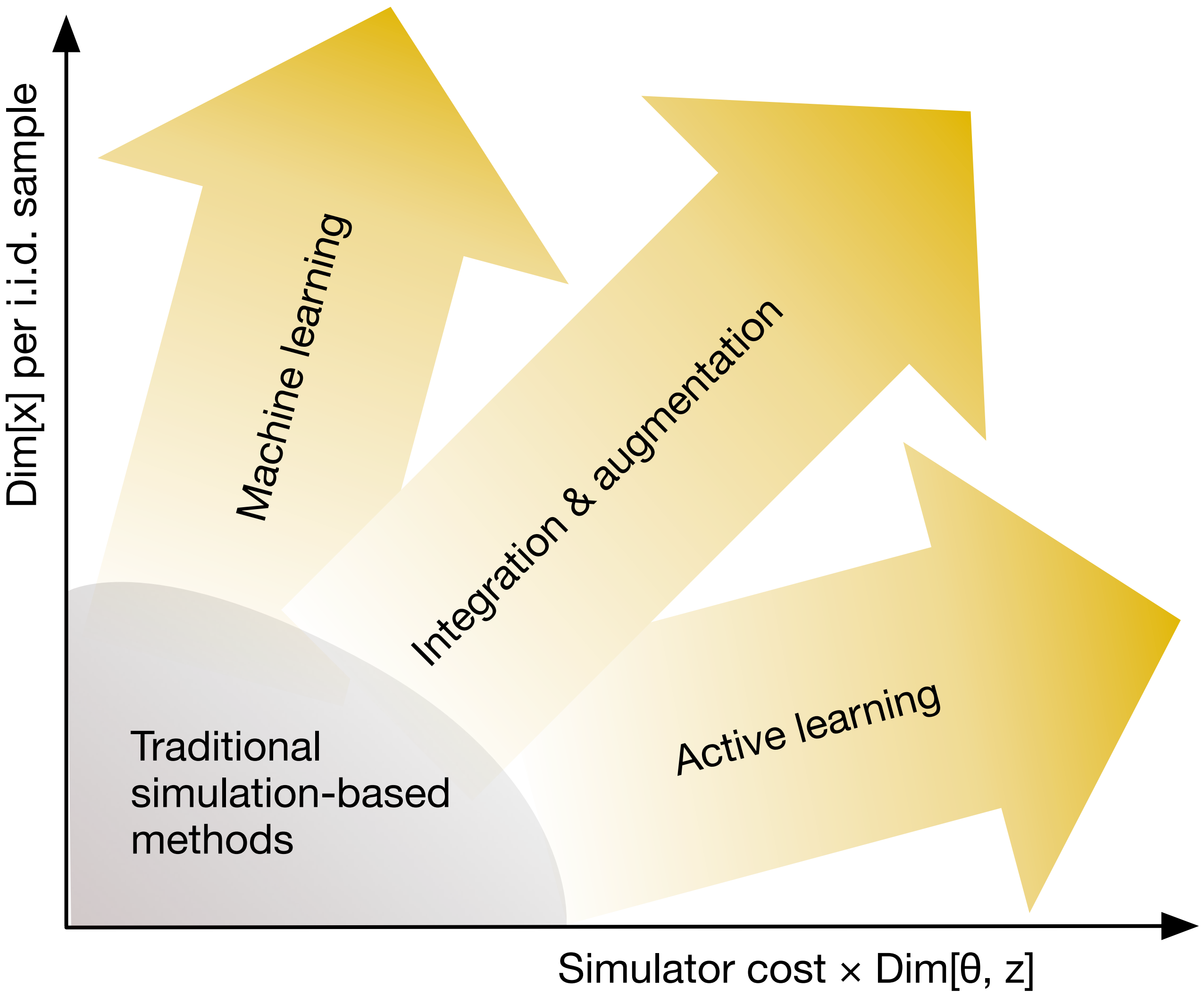
- choosing a summary statistic (feature engineering) is a task for a skilled physicist and tailored to the goal of measurement or new particle search
- likelihood $p(s|\theta)$ **approximated** using histograms (univariate density estimation)

This doesn't scale if x is high dimensional!

SIMULATION-BASED INFERENCE

K.C., J. Brehmer, G. Louppe [arXiv:1911.01429]





Dim[x] per i.i.d. sample
Machine learning
Integration & augmentation
Active learning
Traditional simulation-based methods
Simulator cost × Dim[θ, z]

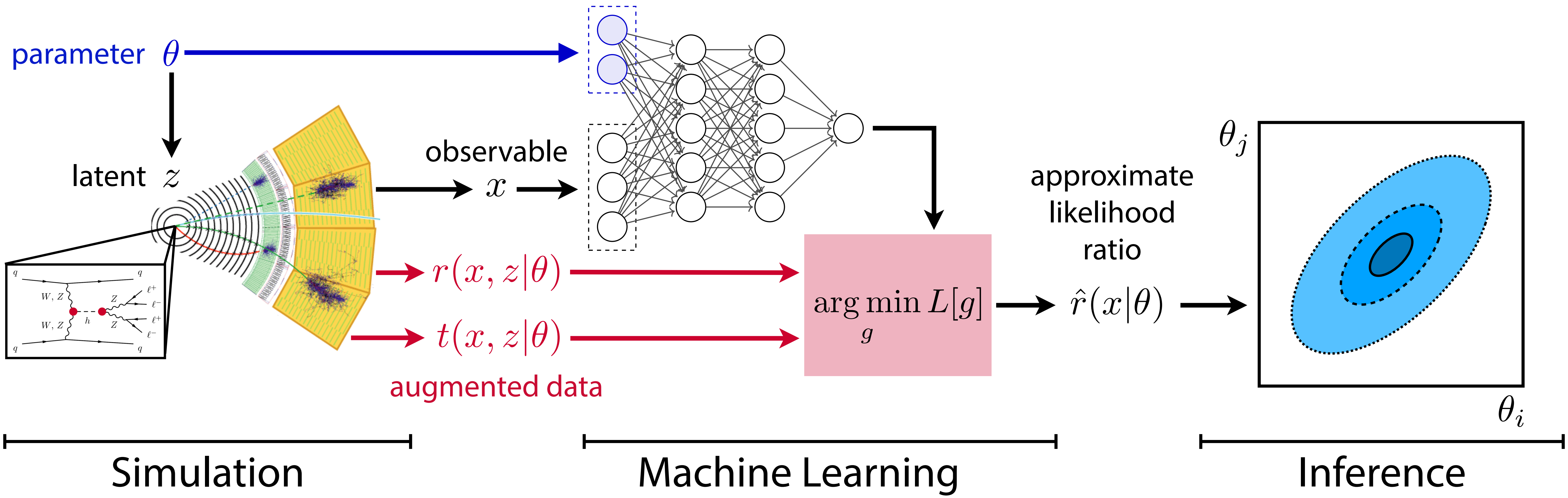
Many areas of science have simulations based on some well-motivated mechanistic model.

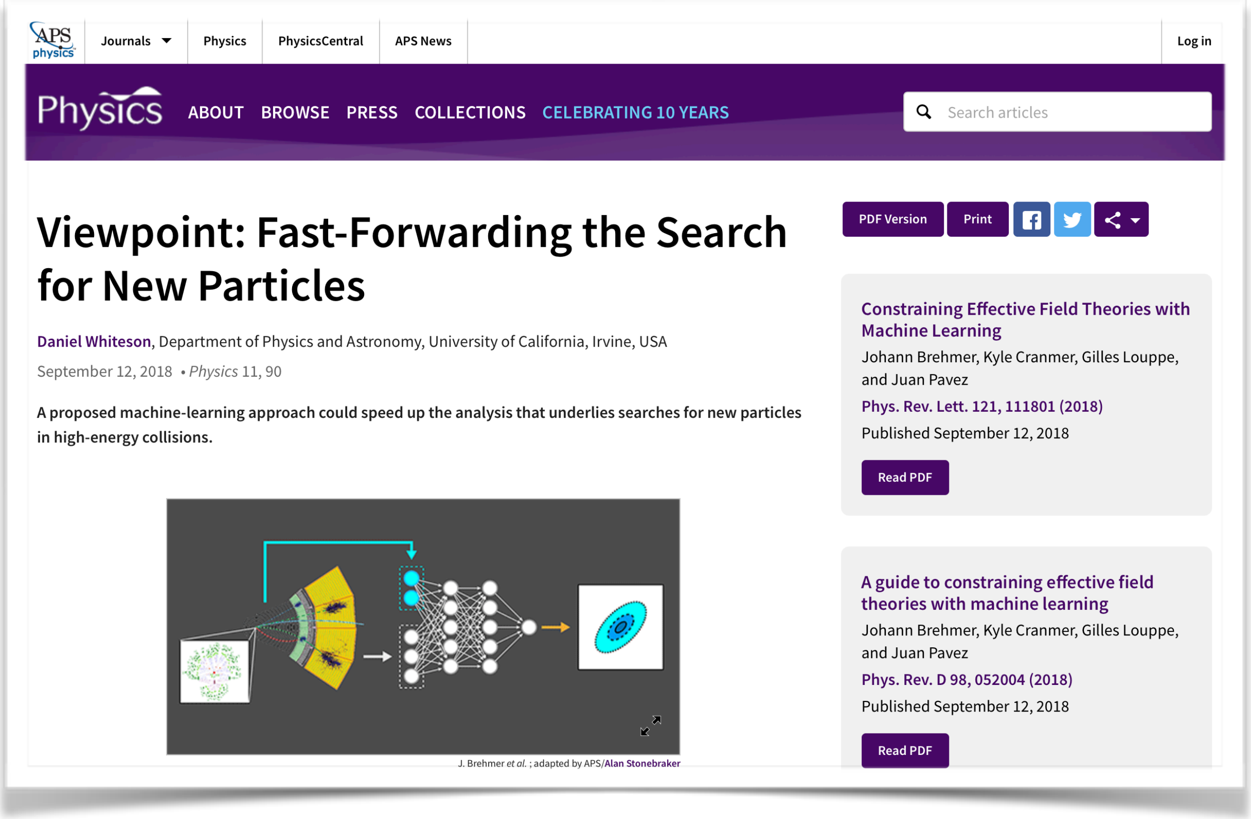
However, the aggregate effect of many interactions between these low-level components leads to an intractable inverse problem.

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

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

ML-POWERED SIMULATION-BASED INFERENCE





Viewpoint: Fast-Forwarding the Search for New Particles

Daniel Whiteson, Department of Physics and Astronomy, University of California, Irvine, USA

September 12, 2018 • *Physics* 11, 90

A proposed machine-learning approach could speed up the analysis that underlies searches for new particles in high-energy collisions.

J. Brehmer *et al.* ; adapted by APS/Alan Stonebraker

Constraining Effective Field Theories with Machine Learning

Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez

Phys. Rev. Lett. 121, 111801 (2018)

Published September 12, 2018

Read PDF

A guide to constraining effective field theories with machine learning

Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez

Phys. Rev. D 98, 052004 (2018)

Published September 12, 2018

Read PDF

Heiko Müller

Irina Espejo

Johann Brehmer

Sinclert Pérez

Felix Kling

arXiv:1805.12244
PRL, arXiv:1805.00013
PRD, arXiv:1805.00020
physics.aps.org/articles/v11/90

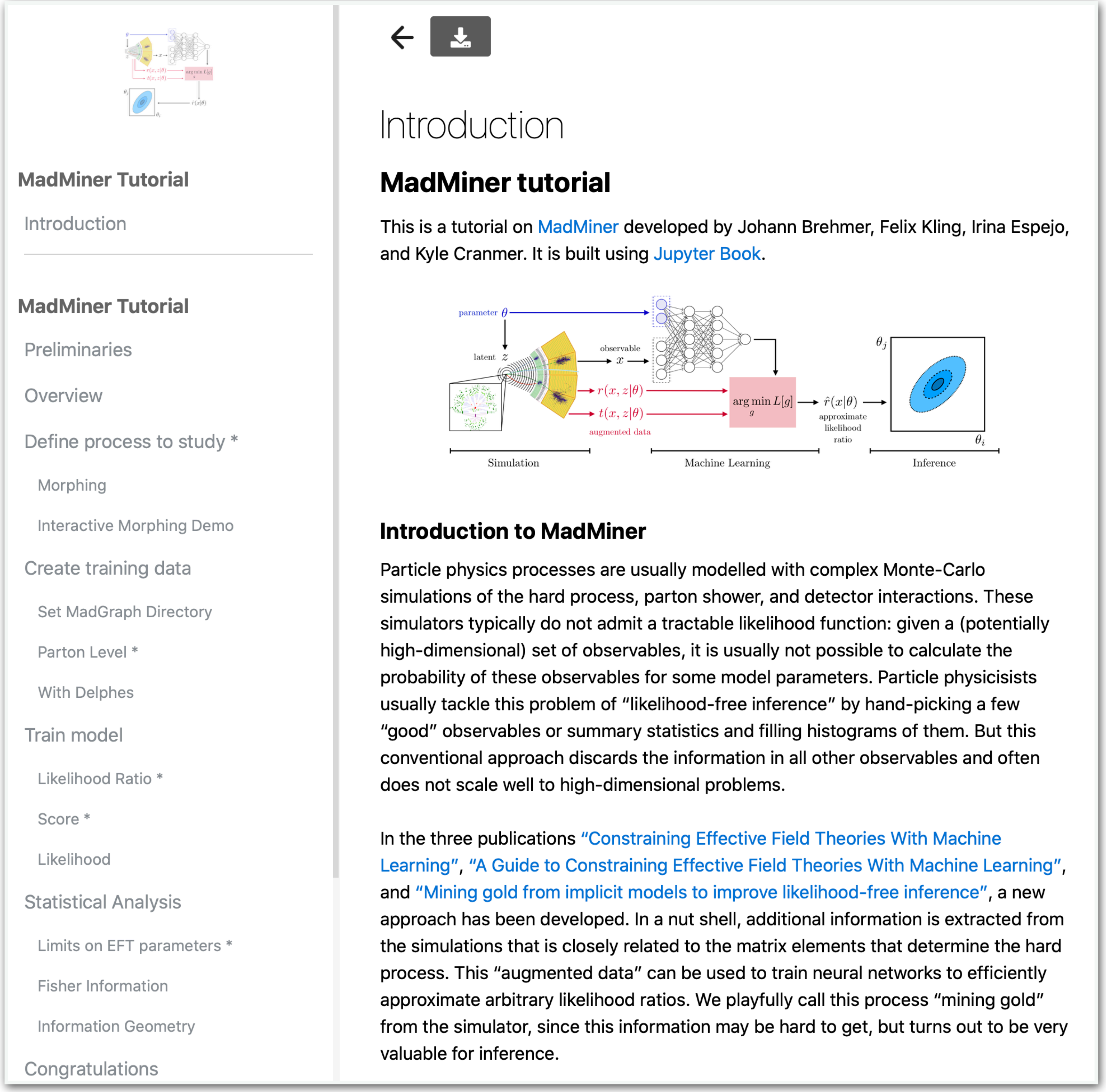
MADMINER

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

Domain-specific software for likelihood-free inference

Integrated into REANA workflow system

Tutorial in JupyterBooks, can run using Binder



Introduction

MadMiner tutorial

This is a tutorial on MadMiner developed by Johann Brehmer, Felix Kling, Irina Espejo, and Kyle Cranmer. It is built using Jupyter Book.

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 physicisists 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.

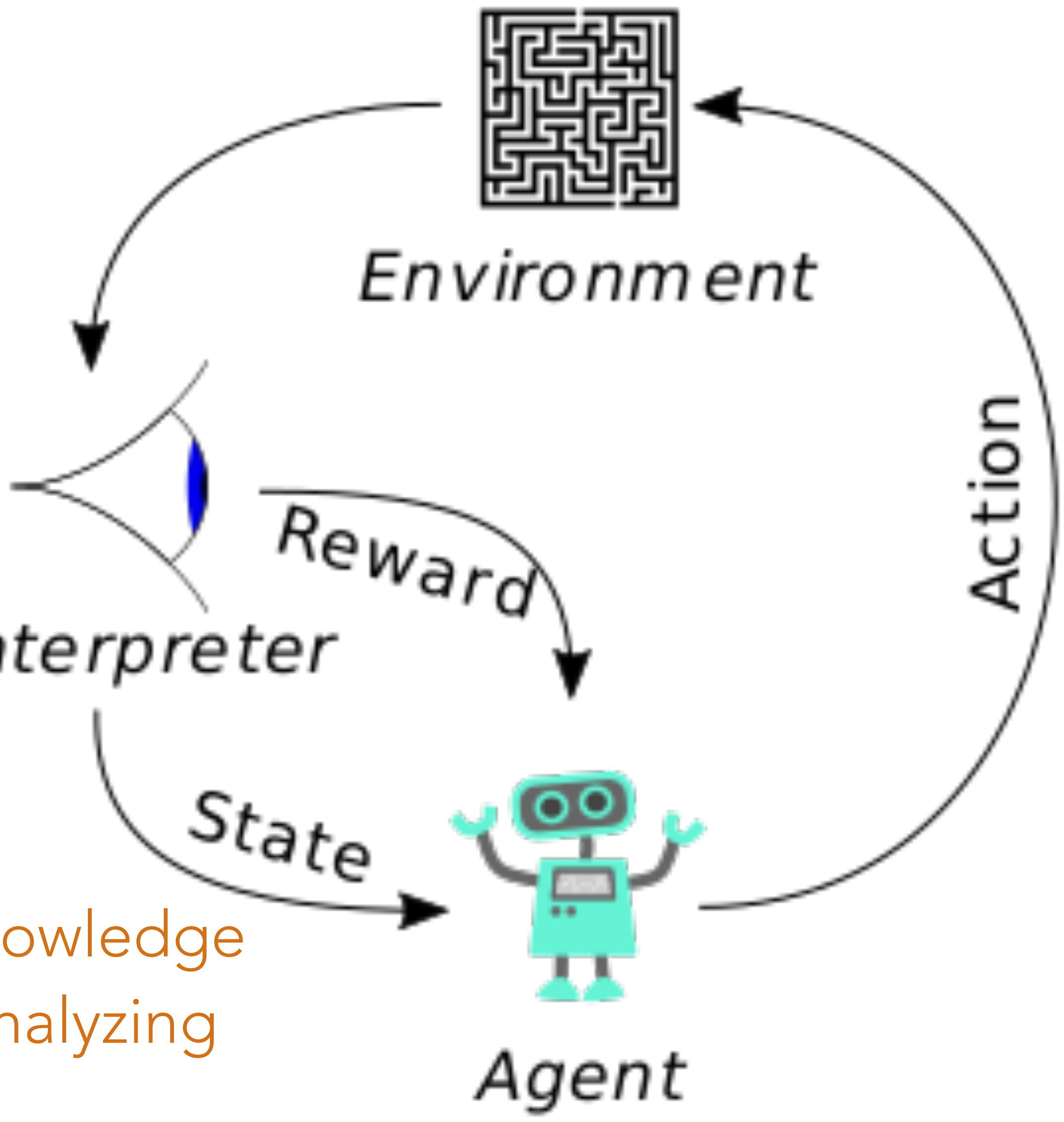
In the three publications “Constraining Effective Field Theories With Machine Learning”, “A Guide to Constraining Effective Field Theories With Machine Learning”, and “Mining gold from implicit models to improve likelihood-free inference”, a new approach has been developed. In a nut shell, additional 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.

REINFORCEMENT LEARNING & SCIENTIFIC METHOD

Scientist trying to decide what experiment to do next

perform experiment,
gather data

statistical analysis

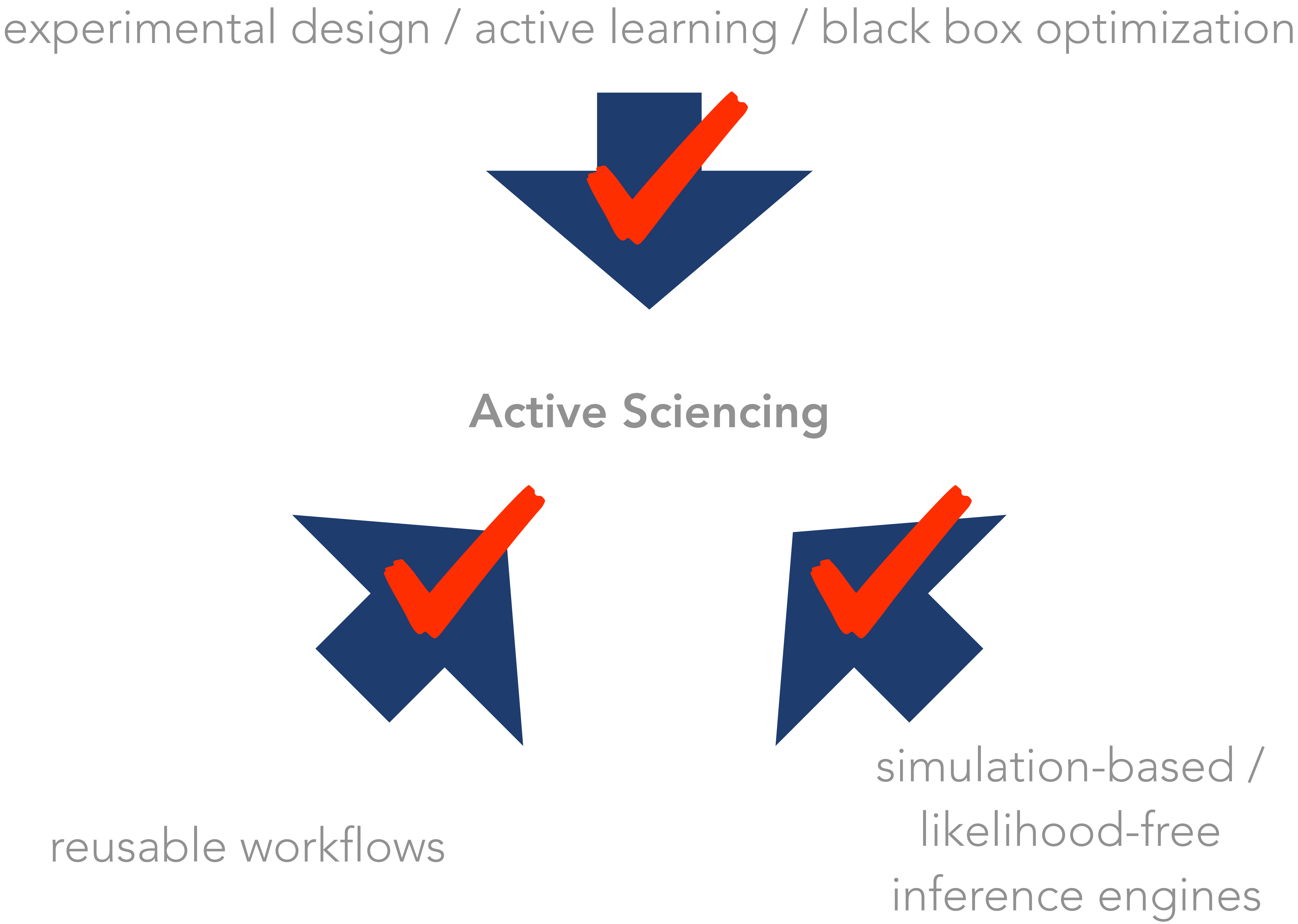


decide which
experiment to
perform

updated knowledge
based on analyzing
data

SYNTHESIS

https://github.com/cranmer/active_sciencing

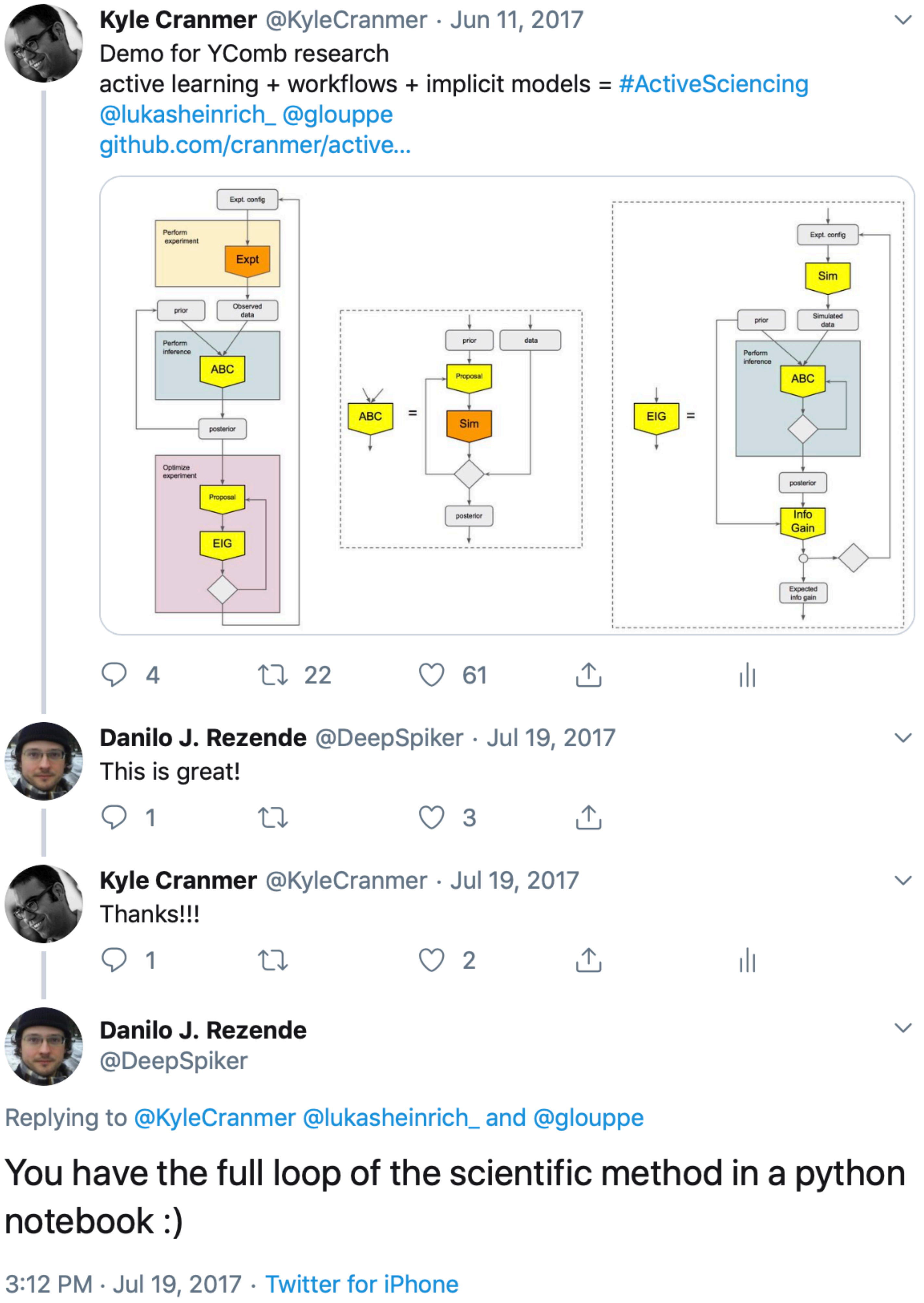


experimental design / active learning / black box optimization

Active Sciencing

reusable workflows

simulation-based /
likelihood-free
inference engines



Kyle Cranmer @KyleCranmer · Jun 11, 2017

Demo for YComb research
active learning + workflows + implicit models = #ActiveSciencing
@lukasheinrich_ @glouppe
github.com/cranmer/active...

Danilo J. Rezende @DeepSpiker · Jul 19, 2017

This is great!

Kyle Cranmer @KyleCranmer · Jul 19, 2017

Thanks!!!

Danilo J. Rezende
@DeepSpiker

Replying to @KyleCranmer @lukasheinrich_ and @glouppe

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

3:12 PM · Jul 19, 2017 · Twitter for iPhone

https://github.com/cranmer/active_sciencing

Reality check…

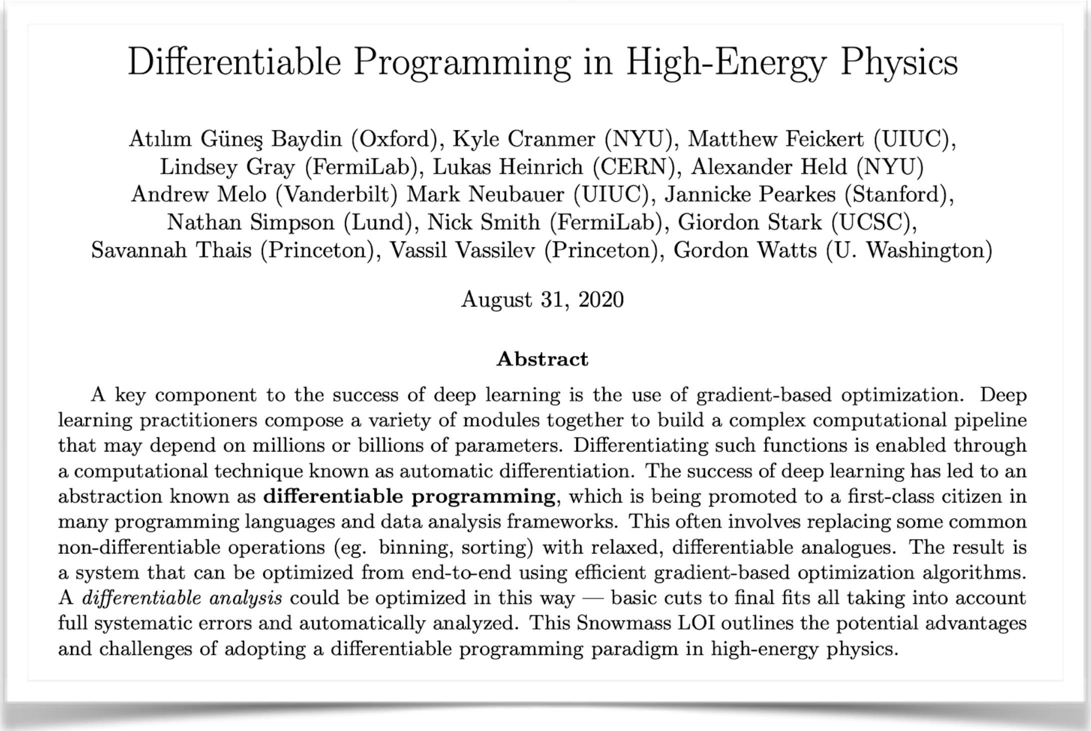
Keep in mind that
- the simulator model was specified
- the space of experimental configurations was well specified

Still it was hard enough!

Going to open world of experimental configurations and potential models much harder.

DIFFERENTIABLE WORKFLOWS

We are now working on **differentiable workflows** where gradients are passed through different processes workflow system



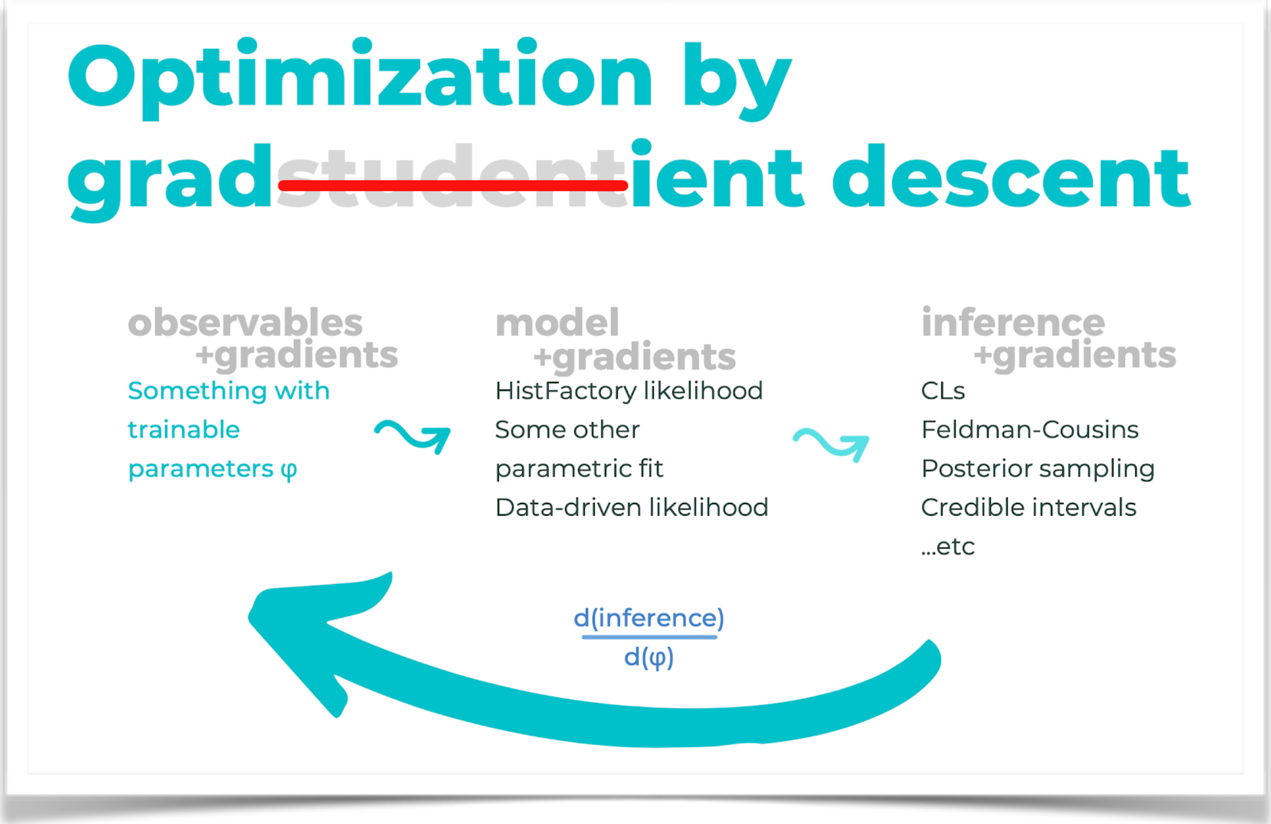
Differentiable Programming in High-Energy Physics

Atılım Güneş Baydin (Oxford), Kyle Cranmer (NYU), Matthew Feickert (UIUC), Lindsey Gray (FermiLab), Lukas Heinrich (CERN), Alexander Held (NYU) Andrew Melo (Vanderbilt) Mark Neubauer (UIUC), Jannicke Pearkes (Stanford), Nathan Simpson (Lund), Nick Smith (FermiLab), Giordon Stark (UCSC), Savannah Thais (Princeton), Vassil Vassilev (Princeton), Gordon Watts (U. Washington)

August 31, 2020

Abstract

A key component to the success of deep learning is the use of gradient-based optimization. Deep learning practitioners compose a variety of modules together to build a complex computational pipeline that may depend on millions or billions of parameters. Differentiating such functions is enabled through a computational technique known as automatic differentiation. The success of deep learning has led to an abstraction known as **differentiable programming**, which is being promoted to a first-class citizen in many programming languages and data analysis frameworks. This often involves replacing some common non-differentiable operations (eg. binning, sorting) with relaxed, differentiable analogues. The result is a system that can be optimized from end-to-end using efficient gradient-based optimization algorithms. A *differentiable analysis* could be optimized in this way — basic cuts to final fits all taking into account full systematic errors and automatically analyzed. This Snowmass LOI outlines the potential advantages and challenges of adopting a differentiable programming paradigm in high-energy physics.



Optimization by grad~~studen~~tient descent

observables +gradients

Something with trainable parameters φ

model +gradients

HistFactory likelihood

Some other parametric fit

Data-driven likelihood

inference +gradients

CLs

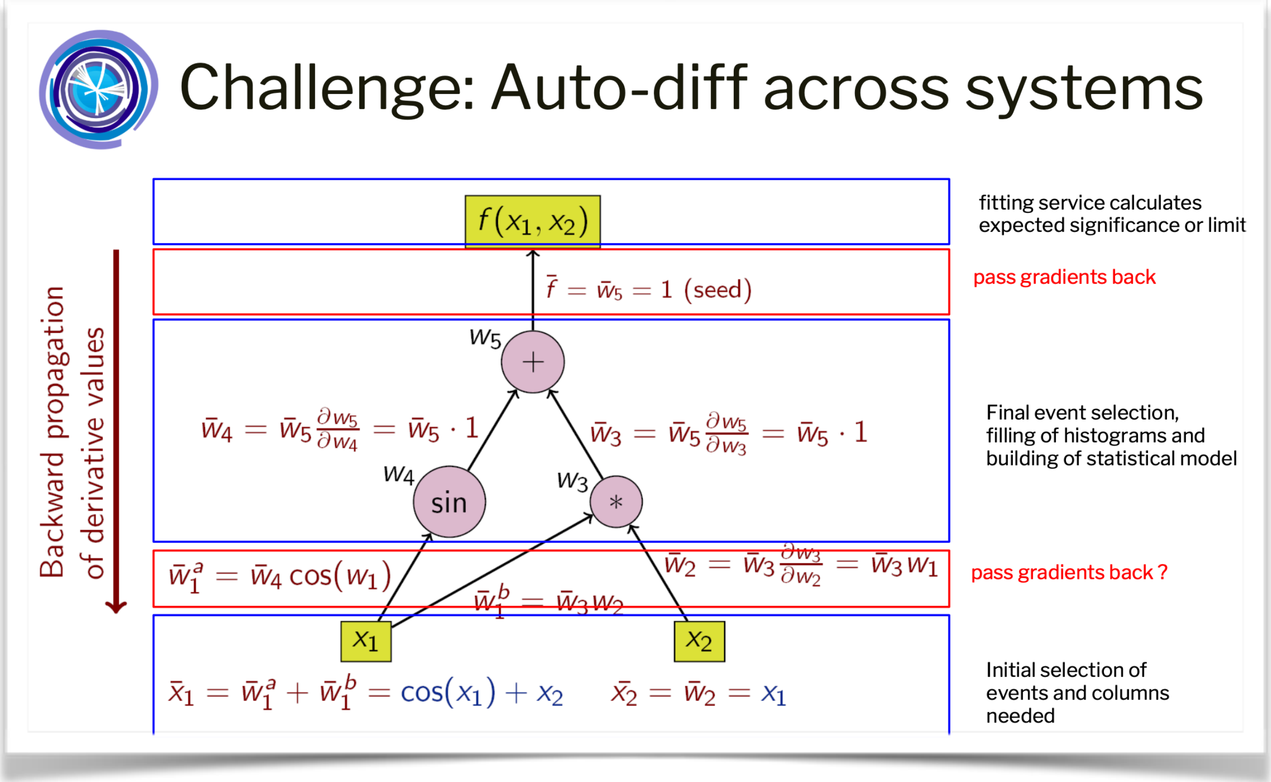
Feldman-Cousins

Posterior sampling

Credible intervals

...etc

$$\frac{d(\text{inference})}{d(\varphi)}$$



Challenge: Auto-diff across systems

Backward propagation of derivative values

$f(x_1, x_2)$

fitting service calculates expected significance or limit

$\bar{f} = \bar{w}_5 = 1$ (seed)

pass gradients back

w_5 +

$\bar{w}_4 = \bar{w}_5 \frac{\partial w_5}{\partial w_4} = \bar{w}_5 \cdot 1$

$\bar{w}_3 = \bar{w}_5 \frac{\partial w_5}{\partial w_3} = \bar{w}_5 \cdot 1$

w_4 sin

w_3 *

Final event selection, filling of histograms and building of statistical model

$\bar{w}_1^a = \bar{w}_4 \cos(w_1)$

$\bar{w}_1^b = \bar{w}_3 w_2$

$\bar{w}_2 = \bar{w}_3 \frac{\partial w_3}{\partial w_2} = \bar{w}_3 w_1$

pass gradients back ?

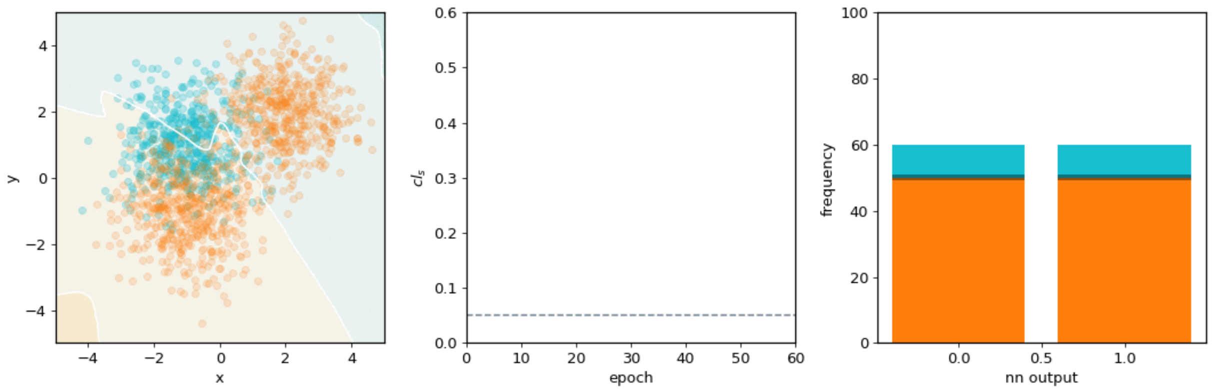
x_1 x_2

$\bar{x}_1 = \bar{w}_1^a + \bar{w}_1^b = \cos(x_1) + x_2$

$\bar{x}_2 = \bar{w}_2 = x_1$

Initial selection of events and columns needed

slide from Nathan Simpson: [link to talk]



CONCLUSIONS

Traditional narrative around reproducibility has been inefficient in changing practice

- Seen as backward-looking, inefficient, irrelevant

Target reuse and preservation

- Start with specific, high-value use-cases in community, then generalize around that
- Reproducibility a byproduct

GORDON AND BETTY MOORE FOUNDATION

Alfred P. Sloan FOUNDATION

MSDSE Moore - Sloan Data Science Environments

NSF

DASPOS

dianahep

iris hep Institute for Research and Innovation in Software for High Energy Physics

The SCAILFIN Project
scailfin.github.io

Backup

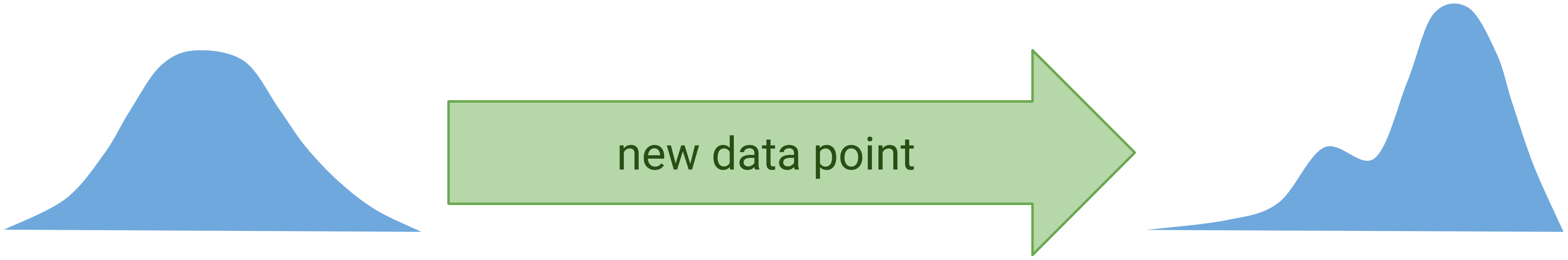
EXPECTED INFORMATION GAIN

Dustin Tran
Research Scientist at Google Brain
trandustin@google.com
Blog

Active Learning & Control

Given data points $\{x, y\}$, how to select the next data point to fit the model?

Ex. Select data points which maximize expected information gain. [Lindley et al. 1956; Mackay 1992; Houthooft et al. 2016]

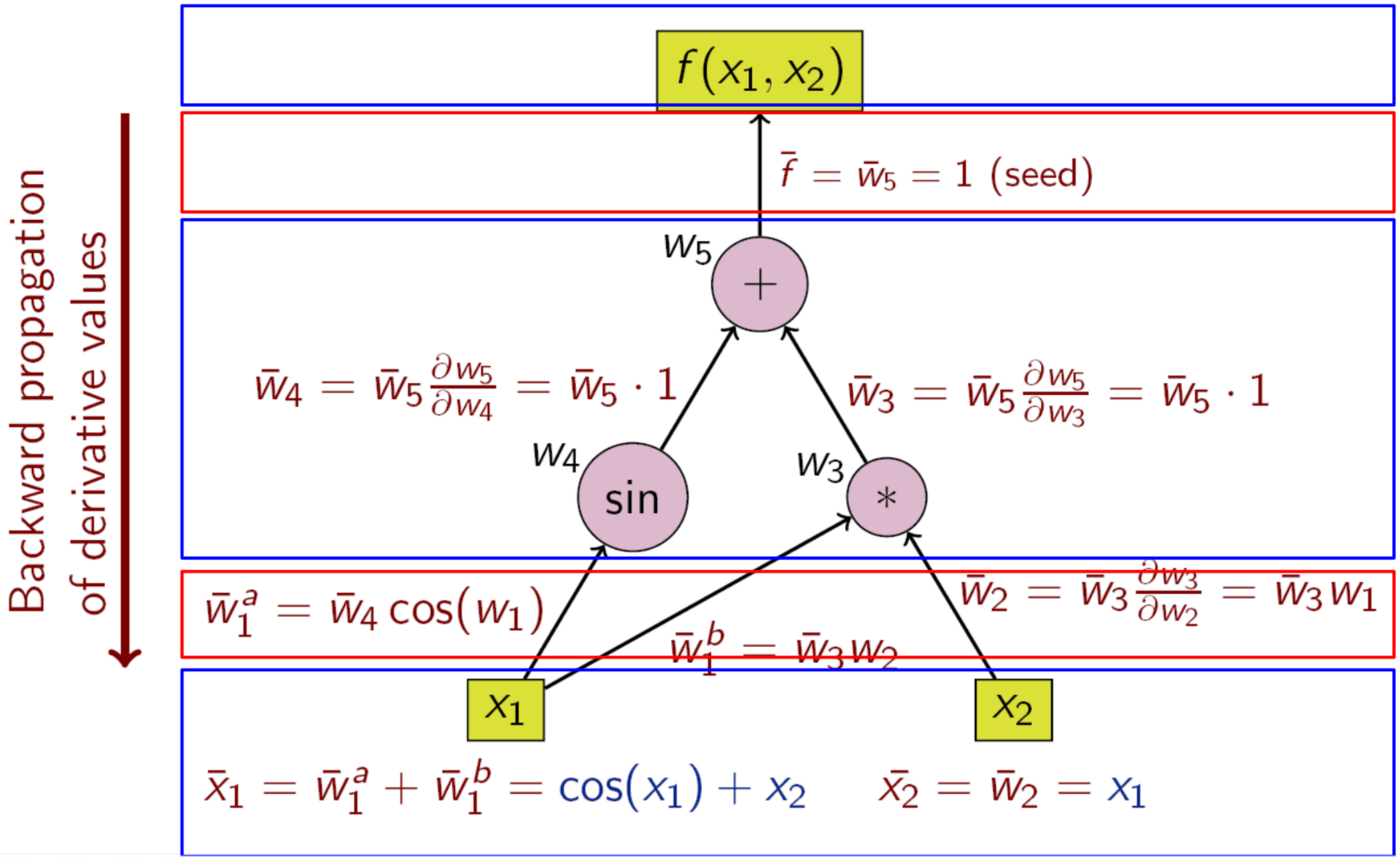


$$\arg\max_{\boldsymbol{x}} \mathrm{H}[\boldsymbol{\theta}|\mathcal{D}] - \mathbb{E}_{y\sim p(y|\boldsymbol{x}\mathcal{D})}\left[\mathrm{H}[\boldsymbol{\theta}|y, \boldsymbol{x}, \mathcal{D}]\right]$$

Uncertainty determines which **x** is most informative and, therefore, the model’s success.

[Hafner et al., 2019]

Challenge: Auto-diff across systems



fitting service calculates expected significance or limit

$\bar{f} = \bar{w}_5 = 1$ (seed)

pass gradients back

$\bar{w}_4 = \bar{w}_5 \frac{\partial w_5}{\partial w_4} = \bar{w}_5 \cdot 1$

$\bar{w}_3 = \bar{w}_5 \frac{\partial w_5}{\partial w_3} = \bar{w}_5 \cdot 1$

Final event selection, filling of histograms and building of statistical model

$\bar{w}_1^a = \bar{w}_4 \cos(w_1)$

$\bar{w}_1^b = \bar{w}_3 w_2$

$\bar{w}_2 = \bar{w}_3 \frac{\partial w_3}{\partial w_2} = \bar{w}_3 w_1$

pass gradients back ?

$\bar{x}_1 = \bar{w}_1^a + \bar{w}_1^b = \cos(x_1) + x_2$

$\bar{x}_2 = \bar{w}_2 = x_1$

Initial selection of events and columns needed