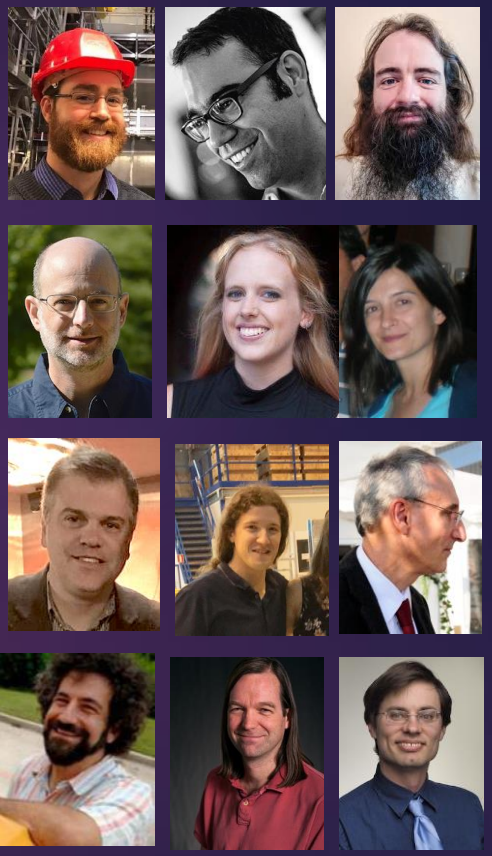
iris hep

Institute for Research & Innovation in Software for High Energy Physics



Analysis Systems

IRIS-HEP ADVISORY PANEL

2019-09-09

GORDON WATTS (UW/SEATTLE) FOR KYLE CRANMER

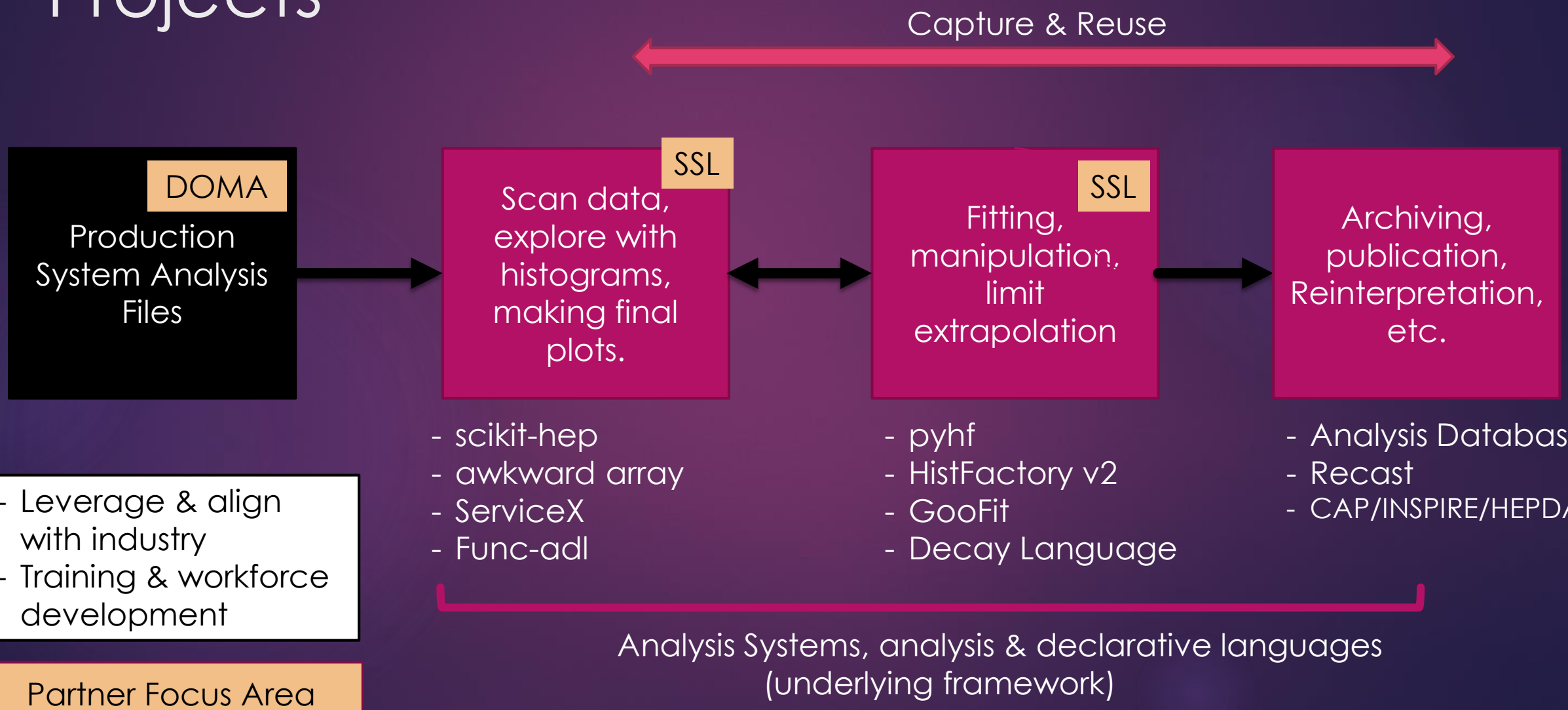
Analysis Systems Focus Area

"Realize the maximum scientific potential in the least time"

Develop sustainable analysis tools to extend the physics reach of the HL-LHC experiments

- Creating greater functionality
- Reudcing time-to-insight
- Lowering the barriers for smaller teams
- Streamlining analysis preservation, reproducibility, and reuse.

Analysis Systems Data Flow & Projects

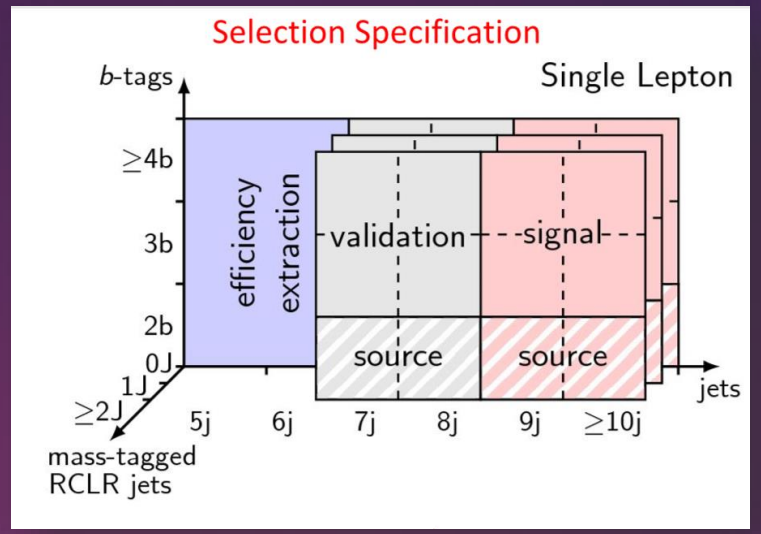


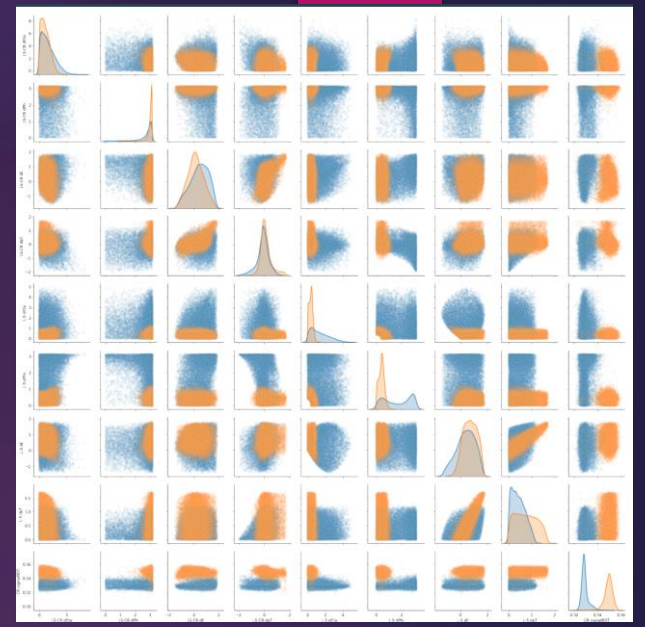
- Leverage & align with industry
- Training & workforce development

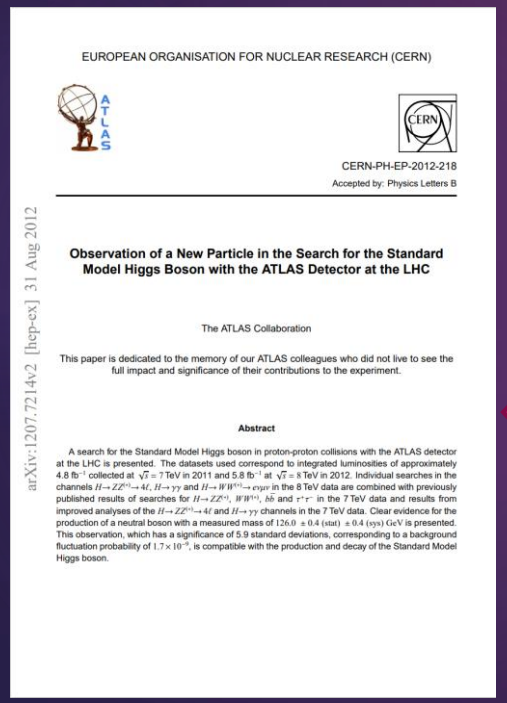
Partner Focus Area

Where would we like to be?

Production System Analysis Files







EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP-2012-218

Accepted by: Physics Letters B

arXiv:1207.7214v2 [hep-ex] 31 Aug 2012

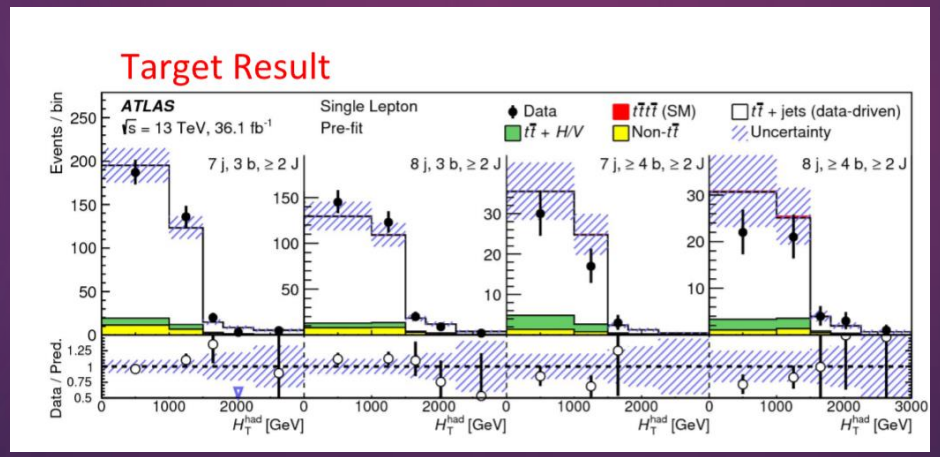
Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

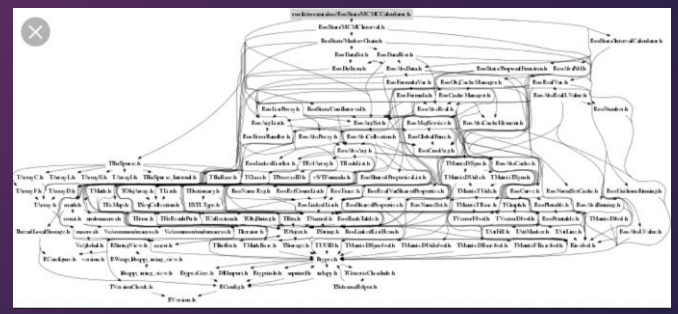
The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.





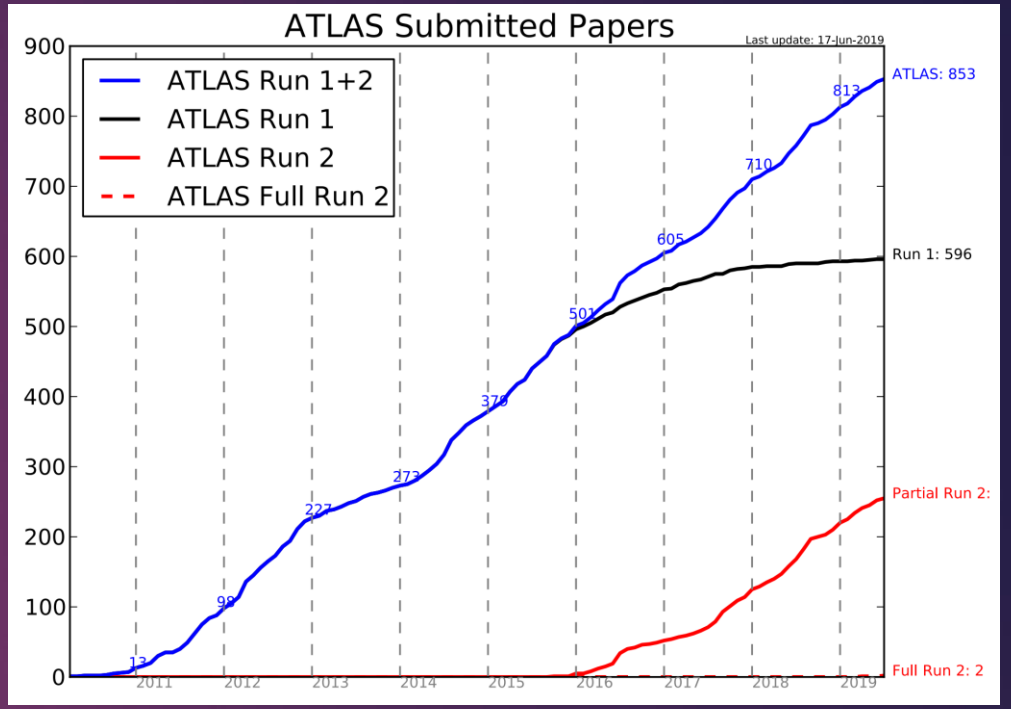
And preserve that workflow for future reuse


Alicia

Context

HEP Publishes lots of papers!

Yet this focus area is green!



- Heterogeneous set of use-cases
- Lots of components not designed with a single view in mind
- New advances in compute and ML mean new opportunities

➡ Many Analysis Systems tasks will be exploratory

Initial Focus

- Establish declarative specifications for analysis tasks and workflows.
  - Enable the technical development of analysis systems to be decoupled from the user-facing semantics of physics analysis.
- Leverage and align developments from industry and the broader software community
  - This should enhance the sustainability of the systems and libraries we have to develop
  - And avoid developing things that are already out there
- Develop high-throughput, low latency systems for analysis for HEP.
- Integrate analysis capture and reuse as first-class concepts and capabilities

The Team

Meetings every other week
Slack channel (open to everyone)

Partnerships

- **External**
  - Open Source Data Science Tools: Spark, Dask, Apache Arrow, pandas, Jupyter, …
  - Workflows, HPC, Cloud: Parsl, Common Workflow Language, Kubernetes, Singularity,
  - Statistics and ML-analysis tools: pytorch, tensorflow, mxnet, pyro, ONNX, ...
  - Industry ML: FAIR, DeepMind, Amazon, nVidia, ...
  - **SCAILFIN** (NSF grant: Workflows + Machine Learning: Hildreth, Cranmer, Neubaeur)
  - Astro. & Cosmo (via stats. & likelihood-free inference), Genomics (via workflows)
  - **CERN IT** via INSPIRE, HEPData, CAP, REANA, … and **CERN SFT** via **ROOT**
  - **Coffea team**
  - HSF analysis group
  - Scientific Gateways Institute
- **Internal**
  - DOMA iDDS
  - SSL
  - Sustainable Core
  - OSG

Tools for Handling Data

Released

ROOT on Conda — Make ROOT as easy to install as any python package

Released

urpoot
Awkward Array — Low level high speed support for hierarchical data: operates at the same level as numpy. And adaptors for loading data from ROOT Ttree's.

Func-adl — Functional declarative language for querying and reducing data to histograms or tables

ServiceX — Tight integration between DOMA, SSL, and Analysis Systems. A service to generate column-wise ntuples from binary formats.

Infrastructure Tools

Histogram Projects

Python interfaces to high performance histogram packages (e.g. boost.histogram) as well as a format converter.

Released

Scikit-HEP

Collection of python packages that together can make up an analysis environment for HEP analyses

Released

Particle
Decay Language

PDG tools and format converters for particles and decays – with pythonic interfaces for all that information

Released

awesome-hep

Awesome list of python packages any HEP analyzer might want in their toolbox.

Fitting Tools

Released

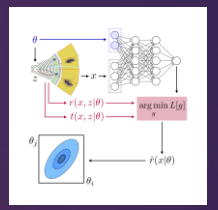
AMPGEN

Fitting and generation of multibody decays.

Released

MadMiner

Instrumenting MC generation and simulation to derive likelihoods for final state configurations given model parameters.



PPX

Common model description language for inference engines

Released

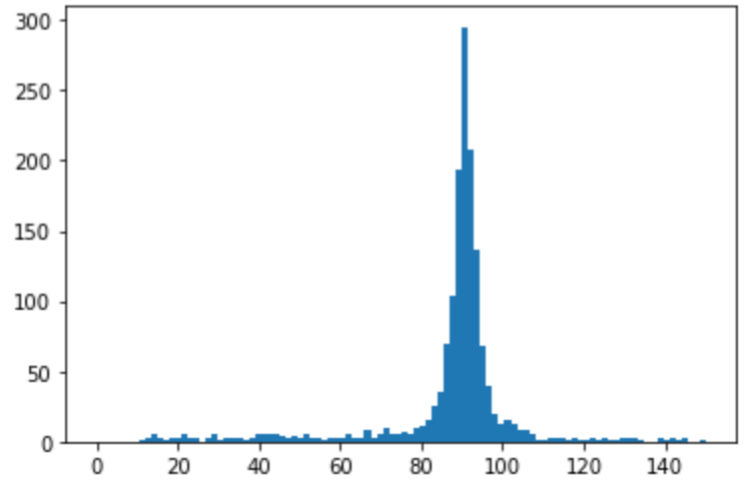
pyhf

Python implementation of the common HEP tool HistFactory, used to build PDF's from a model and data. Will use modern tools (Tensorflow/pytorch backend to use GPUs) to speed calculation.

Awkward Array & Friends

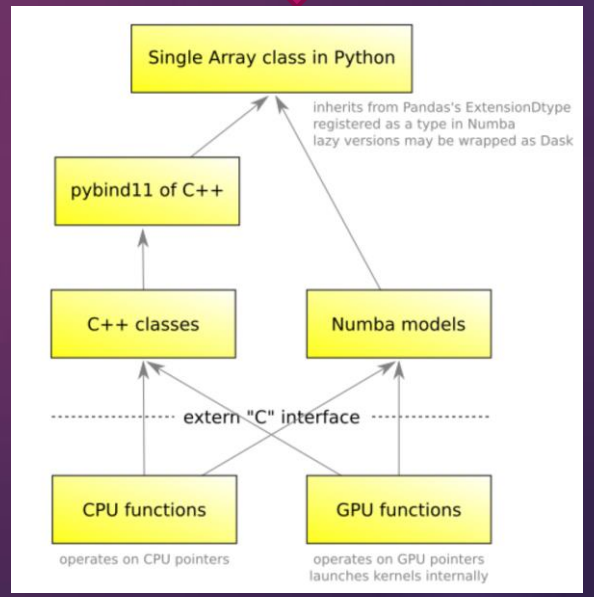
From the CoDaS School last month

```
[57]: (<JaggedArrayMethods [[TLorentzVector(-52.899, -11.655, -8.1
      ... [] [] []] at 0x79184ccb06d8>,
       <JaggedArrayMethods [[TLorentzVector(37.738, 0.69347, -11.3
      ... [] [] []] at 0x79184ccb0c88>)

[58]: # Compute the mass and plot.
      #
      # ("flatten" because Matplotlib needs a flat array, not a ja

      matplotlib.pyplot.hist((first + second).mass.flatten(), bins
```



Large interest in the community (presentations at all the conferences)

Student interest in processing HEP data with these python tools

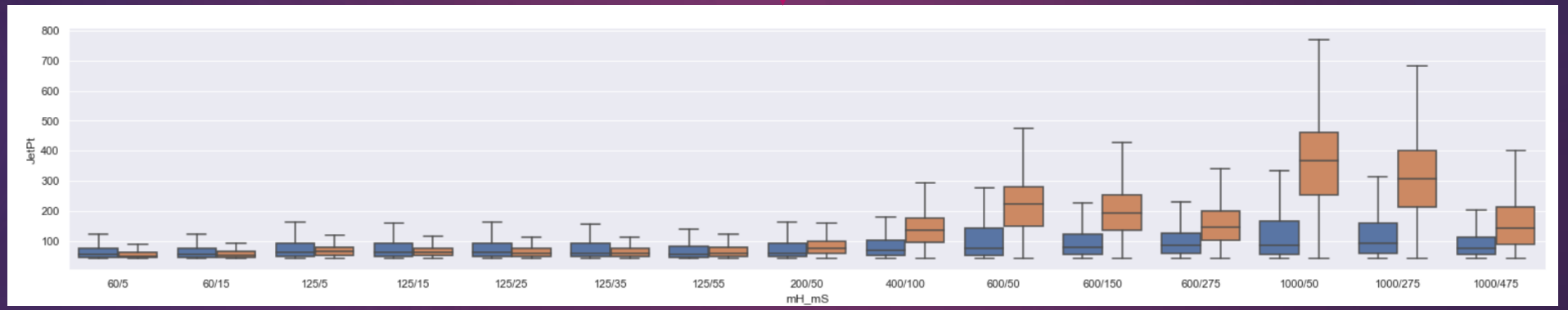


Full Analysis Benchmarks have shown where we can have improvements. V.Next planning well underway.

Testing On Larger Datasets

func-adl → Backend Server

- 57 xAOD MC Datasets (2.7 TB)
- Extract columns for training in ATLAS Analysis
- Using python tools to plot…



Quick check of MC quality and parameters with only small number of lines of python (see on github)

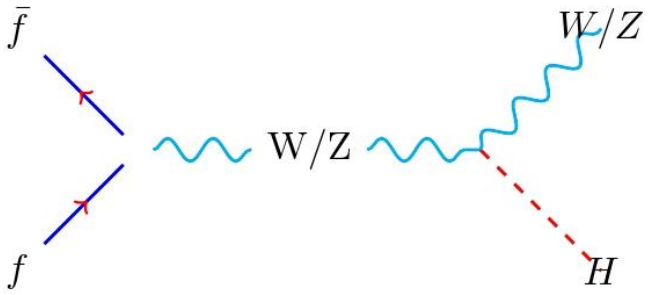
Benchmarking simplified template cross sections in WH production

Johann Brehmer, Sally Dawson, Samuel Homiller, Felix Kling, Tilman Plehn

[Is the Higgs really the SM Higgs?]

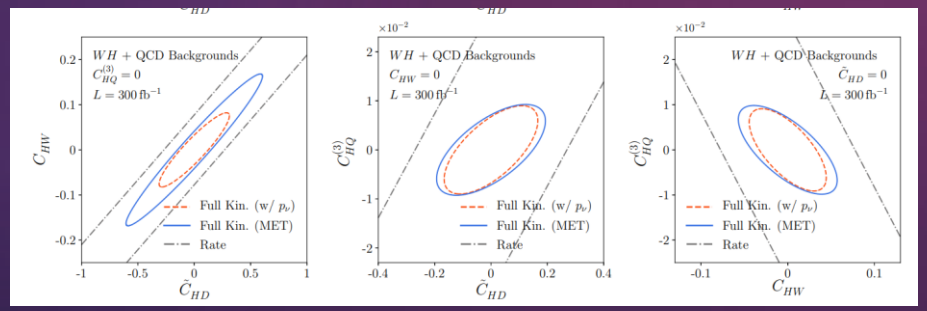
arxiv:1908.06980

Uses MadMiner to instrument Madgraph+Pythia and Delphes



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

What kinematic distributions are most sensitive to new physics (operators)?



Use MadMiner to compare the sensitivity of the various distributions.

- Matrix Element + ML to calculate likelihood
- Can take into account all final states, including ones involving neutrinos, etc.
- Automated framework

Impact

AS Presentations

- 05 Oct 2019 - The Interplay of Data Science and Particle Physics, Kyle Cranmer (New York University), The 6th IEEE International Conference on Data Science and Advanced Analytics
- 09 Sep 2019 - The interplay between physically motivated simulations and machine learning, Kyle Cranmer (New York University), Machine Learning for Physics and the Physics of Learning Long Program at IPAM
- 31 Jul 2019 - Overview and Future directions for ML in particle and astro physics, Kyle Cranmer (New York University), Hammers & Nails 2019
- 29 Jul 2019 - IRIS-HEP Tutorial: Fast columnar data analysis with data science tools, Jim Pivarski (Princeton University), Division of Particles and Fields (DPF) of the American Physical Society (APS)
- 23 Jul 2019 - Scientific Python Ecosystem: Columnar Data Analysis; Accelerating Python, Jim Pivarski (Princeton University), Third Computational and Data Science for High Energy Physics (CoDaS-HEP) School
- 22 Jul 2019 - IRIS-HEP View, Gordon Watts (University of Washington), Computing infrastructures for future data analysis
- 10 Jul 2019 - Motivation and requirements for awkward 1.0, Jim Pivarski (Princeton University), Analysis Systems Biweekly Meeting
- 09 Jul 2019 - pyhf: a pure Python statistical fitting library for High Energy Physics with tensors and autograd, Matthew Feickert (University of Illinois at Urbana-Champaign), 18th annual Scientific Computing with Python conference (SciPy 2019)
- 27 Jun 2019 - Analysis in Run 4, Gordon Watts (University of Washington), ATLAS Software and Computing Week
- 24 Jun 2019 - Delivery of columnar data to analysis systems, Marc Weinberg (University of Chicago), ATLAS Software & Computing Week #62
- 24 Jun 2019 - Future areas of focus for ML in particle physics, Kyle Cranmer (New York University), ATLAS Software and Computing Week
- 21 Jun 2019 - IRIS-HEP Blueprint Concepts and Process, Mark Neubauer (University of Illinois at Urbana-Champaign), Blueprint Meeting on Analysis Systems on Scalable Platforms
- 21 Jun 2019 - Analysis Systems Perspectives and Goals, Kyle Cranmer (New York University), Analysis Systems R&D on Scalable Platforms Blueprint meeting
- 19 Jun 2019 - SCAILFIN: Reproducible Open Benchmarks, Sebastian Macaluso (New York University), Analysis Systems Topical Meeting
- 19 Jun 2019 - Reinterpretation Roadmap, Kyle Cranmer (New York University), Analysis Systems Topical Meeting
- 19 Jun 2019 - MadMiner Update, Johann Brehmer (New York University), Analysis Systems Topical Meeting
- 19 Jun 2019 - Update on awkward-array, uproot, and related projects, Jim Pivarski (Princeton University), Analysis Systems Topical Workshop
- 19 Jun 2019 - SCAILFIN: Madminer deployment using REANA, Irina Espejo (New York University), Analysis Systems Topical Meeting
- 19 Jun 2019 - Histograms, Henry Schreiner (Princeton University), IRIS-HEP Analysis Systems Topical Workshop
- 19 Jun 2019 - AmpGen & Particle/DecayLanguage, Henry Schreiner (Princeton University), IRIS-HEP Analysis Systems Topical Workshop
- 19 Jun 2019 - Functional/Declarative Selection Languages, Gordon Watts (University of Washington), Analysis Systems Topical Workshop
- 19 Jun 2019 - ServiceX, Ben Galewsky (National Center for Supercomputing Applications), Analysis Systems Topical Workshop
- 19 Jun 2019 - Template Fits: HistFitter / TRexFitter, Alexander Held (New York University), Analysis Systems Topical Meeting
- 18 Jun 2019 - Uproot: accessing ROOT data in the scientific Python ecosystem, Jim Pivarski (Princeton University), 3rd CMS Machine Learning Workshop
- 14 Jun 2019 - Advances in Deep Learning motivated by Physics Problems, Kyle Cranmer (New York University), Theoretical Physics for Deep Learning
- 10 Jun 2019 - Numpy, Pandas, PyROOT, and Uproot, Jim Pivarski (Princeton University), U.S. ATLAS Software Training at Argonne National Lab
- 06 Jun 2019 - Constraining effective field theories with machine learning, Johann Brehmer (New York University), INFN Padova seminar
- 04 Jun 2019 - Deep Learning for Higgs Boson Identification and Searches for New Physics at the Large Hadron Collider, Mark Neubauer (University of Illinois at Urbana-Champaign), Blue Waters Symposium for Petascale Science and Beyond
- 29 May 2019 - The Primacy of Experiment, Kyle Cranmer (New York University), The Universe Speaks in Numbers
- 29 May 2019 - Columnar Analysis Tools HATS, Jim Pivarski (Princeton University), LPC HATS: Hands-on Training for CMS
- 28 May 2019 - Scientific Python and Uproot HATS, Jim Pivarski (Princeton University), LPC HATS: Hands-on Training for CMS
- 22 May 2019 - Pattern matching for decay trees, Jim Pivarski (Princeton University), IRIS-HEP Topical Meetings
- 09 May 2019 - Summary of the 'Analysis Description Languages for the LHC' workshop, Jim Pivarski (Princeton University), LPC Physics Forum
- 08 May 2019 - IRIS-HEP: A new software institute to prepare for the data from the High Luminosity Large Hadron Collider in the exabyte era, Mason Proffitt (University of Washington), Northwest Data Science Summit
- 08 May 2019 - Programming languages and particle physics, Jim Pivarski (Princeton University), Fermilab Colloquium
- 07 May 2019 - Thinking about Analysis Languages and Recent Progress, Gordon Watts (University of Washington), Analysis Description Languages
- 06 May 2019 - How to build your own language (hands-on demo), Jim Pivarski (Princeton University), Analysis Description Languages Workshop
- 01 May 2019 - Future areas of focus for ML in particle physics, Kyle Cranmer (New York University), Gotham City Physics X ML
- 18 Apr 2019 - Constraining effective field theories with machine learning, Johann Brehmer (New York University), Higgs and Effective Field Theory 2019
- 15 Apr 2019 - Future areas of focus for ML in particle physics, Kyle Cranmer (New York University), 3rd IML Machine Learning Workshop
- 15 Apr 2019 - Aghast, Jim Pivarski (Princeton University), IRIS-HEP Topical Meetings
- 15 Apr 2019 - boost-histogram and hist, Henry Schreiner (Princeton University), IRIS-HEP Topical Meeting
- 08 Apr 2019 - High-Performance Python and Interoperability with Compiled Code, Jim Pivarski (Princeton University), Princeton PICSciE mini-courses
- 05 Apr 2019 - Scalable Cyberinfrastructure for Artificial Intelligence and Likelihood-Free Inference, Mark Neubauer (University of Illinois at Urbana-Champaign), NSF Large Facilities Workshop
- 01 Apr 2019 - PyROOT, uproot, and awkward-arrays, Jim Pivarski (Princeton University), Software Carpentry at Fermilab
- 21 Mar 2019 - Conda: a complete reproducible ROOT environment in under 5 minutes, Henry Schreiner (Princeton University), 2019 Joint HSF/OSG/WLCG Workshop
- 18 Mar 2019 - Overview of Likelihood-Free Inference for Physics, Kyle Cranmer (New York University), Likelihood-Free Inference Workshop
- 18 Mar 2019 - 'Mining gold' from simulators to improve likelihood-free inference, Johann Brehmer (New York University), Likelihood-free inference workshop
- 14 Mar 2019 - Keynote: Constraining effective field theories with machine learning, Johann Brehmer (New York University), International Workshop on Advanced Computing and Analysis Techniques in Physics Research
- 14 Mar 2019 - Nested data structures in array and SIMD frameworks, Jim Pivarski (Princeton University), ACAT 2019
- 14 Mar 2019 - Beyond the Roadmap: HL-LHC HEP Software, Gordon Watts (University of Washington), ACAT 2019
- 11 Mar 2019 - Aligning the MATHUSA Test Stand Detector Using Tensorflow, Gordon Watts (University of Washington), ACAT 2019
- 28 Feb 2019 - Bringing together simulations, physics insight, and machine learning to constrain new physics, Johann Brehmer (New York University), Dark universe seminar
- 25 Feb 2019 - The C# LINQ Analysis Language, Gordon Watts (University of Washington), IRIS-HEP Topical Meeting on Analysis Description Languages
- 20 Feb 2019 - IRIS-HEP and ATLAS, Gordon Watts (University of Washington), US ATLAS IB Meeting
- 13 Feb 2019 - LINQ To ROOT, Gordon Watts (University of Washington), 1st DAWG Technology and Innovation Survey (HSF)
- 06 Feb 2019 - IRIS-HEP Analysis Systems, Kyle Cranmer (New York University), IRIS-HEP Steering Board Meeting
- 06 Feb 2019 - IRIS-HEP Steering Board Meeting #1, Gordon Watts (University of Washington), IRIS-HEP Steering Board Meeting
- 14 Jan 2019 - Meticulous measurements with matrix elements and machine learning, Johann Brehmer (New York University), ITS/CHEP joint seminar
- 11 Jan 2019 - Improving inference with matrix elements and machine learning, Johann Brehmer (New York University), HK IAS Program on High Energy Physics
- 29 Oct 2018 - pyhf: a pure Python implementation of HistFactory with tensors and autograd, Matthew Feickert (University of Illinois at Urbana-Champaign), DIANA/HEP Meeting
- 19 Oct 2018 - Design Roadmap for Future Collaborations, Mark Neubauer (University of Illinois at Urbana-Champaign), Deep Learning for Multimessenger Astrophysics Real-time Discovery at Scale
- 20 Sep 2018 - Learning to constrain new physics, Johann Brehmer (New York University), IPPP seminar
- 13 Sep 2018 - Learning to constrain new physics, Johann Brehmer (New York University), Pheno & Vino Seminar
- 27 Aug 2018 - Learning to constrain new physics, Johann Brehmer (New York University), Elementary particle seminar
- 23 Jul 2018 - Machine Learning to Probe a BSM Higgs Sector, Johann Brehmer (New York University), Higgs Hunting
- 27 Jun 2018 - Constraining Effective Theories with Machine Learning, Johann Brehmer (New York University), Theory seminar

Presentations range from internal, to trainings, to seminars, conference talks, and colloquia.

Publications are just getting started. Most in journals, some conference proceedings as well.

AS Publications

- Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning, J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe and K. Cranmer, arXiv 1909.02005 (04 Sep 2019).
- Benchmarking simplified template cross sections in WH production, J. Brehmer, S. Dawson, S. Homiller, F. Kling and T. Plehn, arXiv 1908.06980 (19 Aug 2019).
- MadMiner: Machine learning-based inference for particle physics, J. Brehmer, F. Kling, I. Espejo and K. Cranmer, arXiv 1907.10621 (24 Jul 2019).
- Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale, A. Baydin, L. Shao, W. Bhimji, L. Heinrich, L. Meadows et. al., arXiv 1907.03382 (07 Jul 2019).
- Effective LHC measurements with matrix elements and machine learning, J. Brehmer, K. Cranmer, I. Espejo, F. Kling, G. Louppe et. al., arXiv 1906.01578 (04 Jun 2019).
- Machine learning and the physical sciences, G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld et. al., arXiv 1903.10563 (25 Mar 2019).
- Open is not enough, X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos et. al., Nature Phys. 15 (2019) (15 Nov 2018).
- Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model, A. Baydin, L. Heinrich, W. Bhimji, L. Shao, S. Naderiparizi et. al., arXiv 1807.07706 (20 Jul 2018).

Some of the mliestones

G2.1	Organize topical meetings, Analysis System group meetings, etc.
G2.2	List publicly-accessible repositories and other relevant doc-umentation on the iris-hep.org website
G2.3	Collect and curate example analysis use cases with some existing reference implementation
G2.4	Survey of analysis systems efforts in the field to aid in planning for topical workshop
G2.5	Blueprint workshop coordinating resource needs for evaluating analysis systems coordinated by SSL with participation of operations program
G2.6	Develop initial specifications for user-facing interface to analysis system components
G2.7	Prototype awkward-array analyses in the scientific Python ecosystem

Coming Up

G2.8	Initial roadmap for ecosystem coherency
G2.9	Develop initial design for interface of analysis query system to the IDDS
G2.10	Translate analysis examples into new specifications, provide feedback, iterating as necessary
G2.11	Initial roadmap for high-level cyberinfrastructure components of analysis system
G2.12	Benchmarking and assessment of existing analysis systems

Conclusions & Observations

- Many packages are now actively worked on by IRIS-HEP Analysis System Members
  - Some have wide adoption (e..g uproot and awkward array)
- Everything is open source and almost all of it is on github
  - Will allow use and adoption no matter what direction IRIS-HEP chooses in the future
- Even the services we are working on are designed in the same vein
  - E.g. helm charts or similar
- Analysis Systems is now planning for the transition from Development to Implementation phase
  - We still have a significant amount of development and design work to do
  - This will likely continue in parallel with implementation