IRIS-HEP
Analysis Systems Team

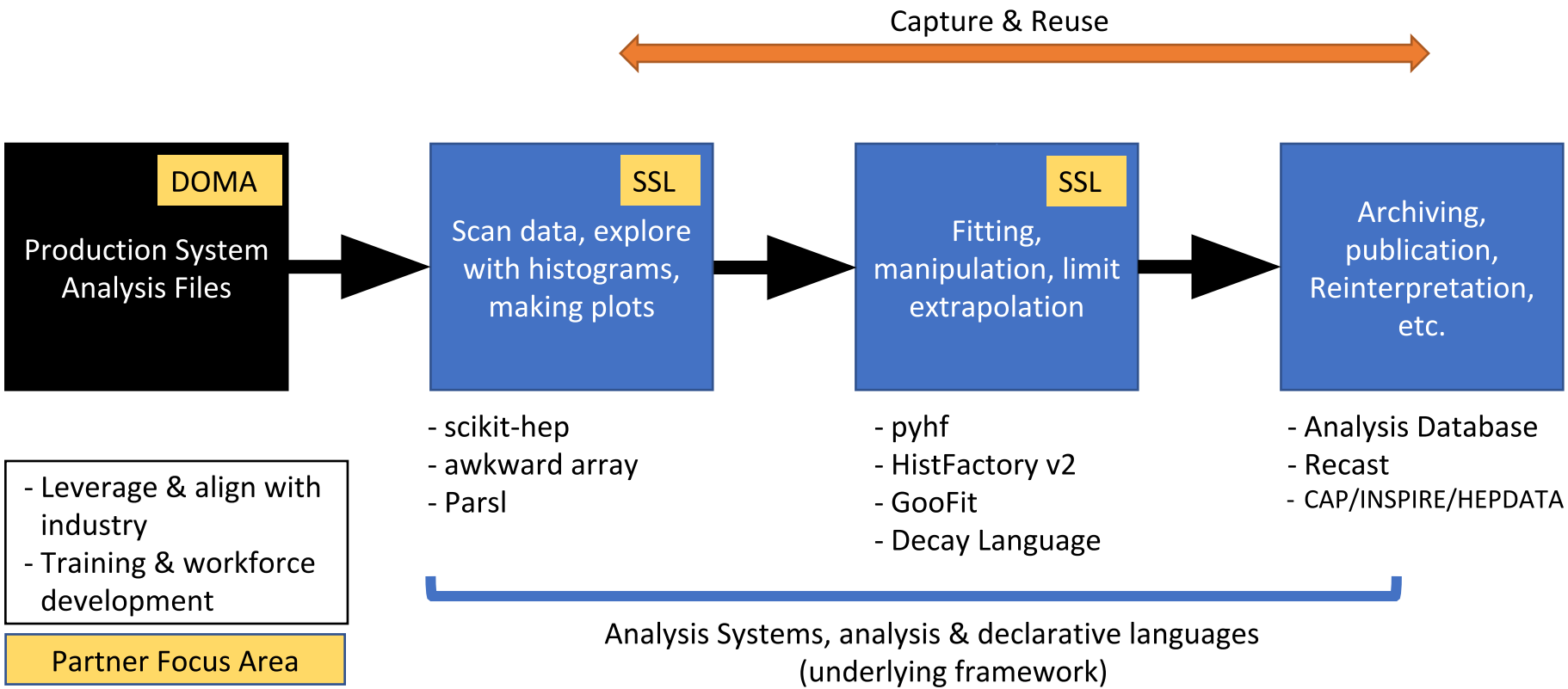
Kyle Cranmer for the team

Overall R&D goal for Analysis Systems

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

- creating greater functionality,
- reducing time-to-insight,
- lowering the barriers for smaller teams, and
- streamlining analysis preservation, reproducibility, and reuse.

Analysis Systems Scope



Context

- Compared to DOMA and IA (which has more targeted reco/trigger goals), the Analysis Systems group is dealing with more “greenfield” area where there is a very heterogeneous set of use cases and relevant components
  - Nature of AS tasks will be more exploratory
- The AS group is bringing together a few existing groups
  - DASPOS and capture/reproducibility/reuse components of DIANA
  - Scikit-hep and Jim P’s efforts on interoperability and query-based systems
  - High-performance statistical analysis tools (eg. GooFit, HistFactory, pyhf, etc.)
- And adding new connecting theme: declarative specifications

Analysis Systems Team (Directly supported by IRIS-HEP)

NYU: Kyle Cranmer, Johann Brehmer, Alex Held (starting August)

UIUC: Mark Neubaeur, Dan Katz, Ben Galwesky, Matthew Feickert (July)

UW: Gordon Watts, Mason Proffit, Emma Torro

Princeton: Jim Pivarski, Vassil Vassilev

Cincinnati: Mike Sokoloff, Tim Evans (leaving), TBD (starting)

Biweekly meetings, Slack discussions (open to others and partner projects)

Partnerships

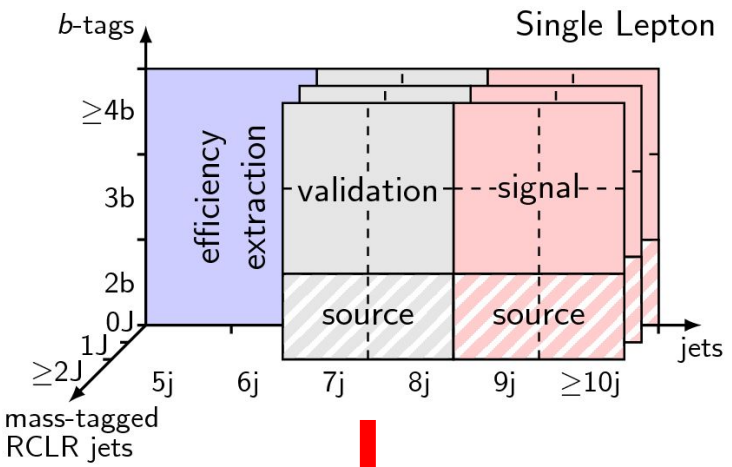
- **External**
  - Open Source Data Science Tools: Spark, Dask, Apache Arrow, pandas, Jupyter, …
  - Workflows, HPC, Cloud: Parsl, Common Workflow Language, Kubernetes, Singularity,
  - Statistics and ML-analysis tools: pytorch, tensorflow, mxnet, pyro, ONNX, ...
  - Industry ML: FAIR, DeepMind, Amazon, 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

A point of convergence

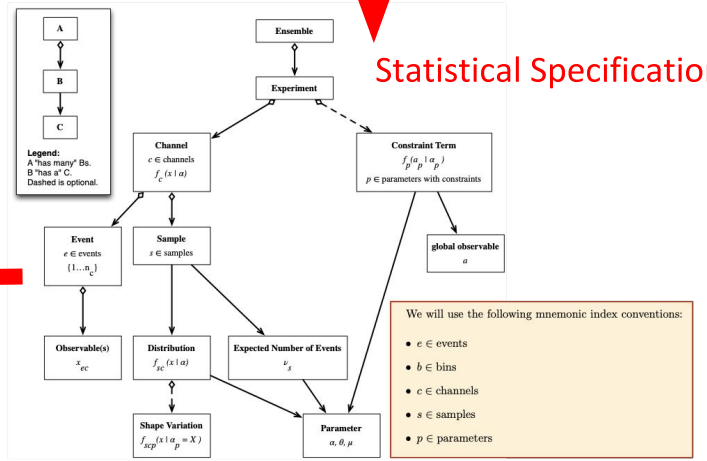
Several aspects of Analysis Systems converge in a typical physics plot:

- Specification of selection, histograms, systematics, etc.
- Data reduction (eg. run over data, fill histograms)
- Feed reduced data into specification for statistical tools
- Publishing results & derived data products
- Analysis preservation & reinterpretation

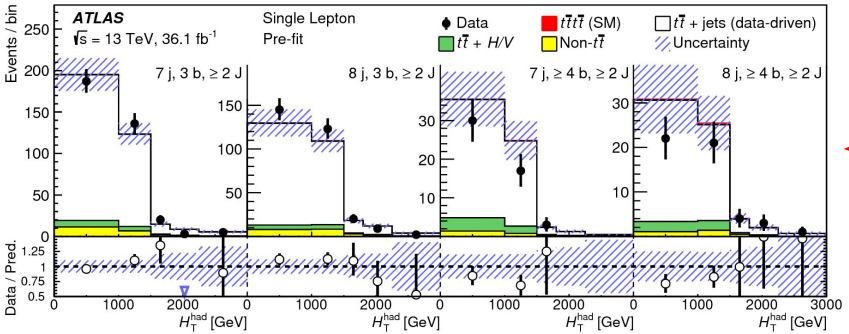
Selection Specification



Statistical Specification



Target Result



Focus area goals

- Establish declarative specifications for analysis tasks and workflows that will enable the technical development of analysis systems to be decoupled from the user- facing semantics of physics analysis.
- Leverage and align with developments from industry and the broader scientific software community to enhance sustainability of the analysis systems.
- Develop high-throughput, low-latency systems for analysis for HEP.
- Integrate analysis capture and reuse as first class concepts and capabilities into the analysis systems.

Upcoming Milestones & Deliverables in Y1Q3&4

- G2.2 List publicly-accessible repositories and other relevant documentation 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

Visibility & Engagement

2019 Joint HSF/OSG/WLCG Workshop HOW2019

18-22 March 2019
THOMAS JEFFERSON NATIONAL ACCELERATOR FACILITY
US/Eastern timezone

10-15 March 2019
Steinmatte conference center
Europe/Zurich timezone

Analysis Description Languages for the LHC

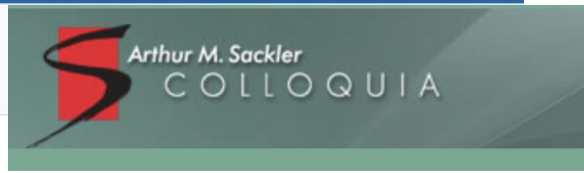
6-8 May 2019
Fermilab, Wilson Hall
America/Chicago timezone

(Re)interpreting the results of new physics searches at the LHC

2-4 April 2019
Imperial
Europe/London timezone

Data Analysis WG

Indico area for the HSF Data Analysis Working Group



The Science of Deep Learning

March 13 - 14, 2019
National Academy of Sciences, Washington, D.C.

Collaboration with Coffea

Will use analyses implemented in Coffea as one example of analysis use case with reference implementation

The Case for Columnar Analysis (a two-part series)

Nick Smith, on behalf of the Coffea team

Lindsey Gray, Matteo Cremonisi, Bo Jayatilaka, Oliver Gutsche, Nick Smith, Allison Hall, Kevin Pedro (FNAL); Andrew Melo (Vanderbilt); and others

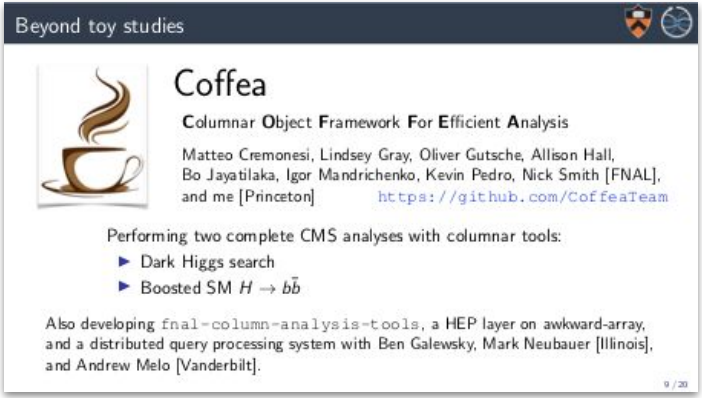
In collaboration with iris-hep members:

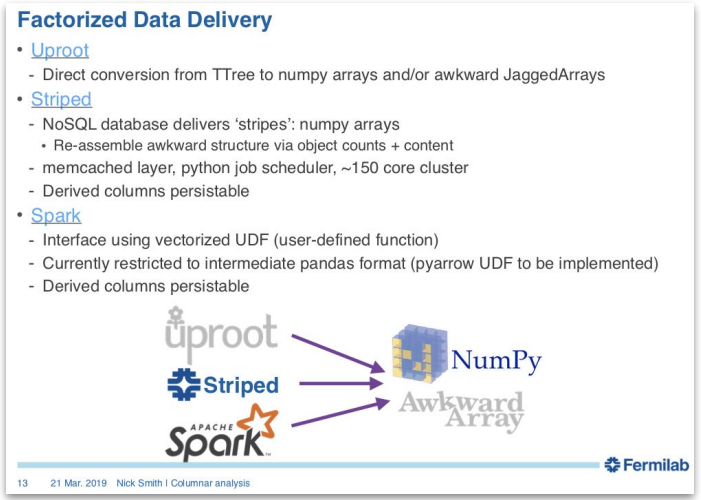
Jim Pivarski (Princeton); Ben Galewsky (NCSA); Mark Neubauer (UIUC)

HOW 2019

21 Mar. 2019

Data delivery system	*Z peak wall-time throughput*	*Subjective 'ease of use'*
uproot on laptop	*~ 100 kHz*	*5/5*
uproot + xrootd + multiprocessing	*~ 250 kHz @ 10 cores **	*5/5*
uproot + condor jobs	*Arbitrary*	*3/5*
striped system	*~ 10 MHz @ 100 cores*	*2/5*
Apache spark	*~ 1 MHz @ 100 cores ***	*4/5*





Cloud-native analysis

Using CMS open data & cloud infrastructure

Will use as an example analysis with reference implementation

Tuesday, May 21 • 17:20 - 17:40

Keynote: Reperforming a Nobel Prize Discovery on Kubernetes - Ricardo Rocha, Computing Engineer & Lukas Heinrich, Physicist, CERN

Sign up or log in to save this to your schedule and see who's attending!

https://sched.co/MRy Tweet Share

Back in 2012, CERN announced one of its most important achievements, the discovery of the Higgs boson leading to the 2013 Nobel Prize in Physics.

In this presentation, we will redo the data analysis that led to it, this time on top of Kubernetes, the new infrastructure stack growing in popularity in the laboratory.

The analysis submission will be done via a jupyter notebook into a small cluster on our private cloud, and both the application and the cluster itself will automatically scale out to exhaust resources - we will detail our setup and deployment decisions on the way. Then we will show how the work being done in the Multicluster SIG helps us define a set of placement and scheduling policies to scale out to external clouds.

The end result will be a physicist's dream: a histogram with a spike that back in 2012 indicated the discovery of a new particle.

HOME | ATTEND | SPONSORS | VENUE + TRAVEL | PROGRAM | REGISTER

KubeCon CloudNativeCon Europe 2019

May 20 – 23, 2019 | Fira Gran Via, Barcelona, Spain | #KubeCon #CloudNativeCon

Live streamed earlier today!

Speakers

Ricardo Rocha

Computing Engineer, CERN

Ricardo is a software engineer at CERN currently part of the CERN cloud team, focusing primarily on networking and container based deployments. Previously he helped develop and deploy several components of the Worldwide LHC Computing Grid, a network of ~200 collaborating sites around... Read More →

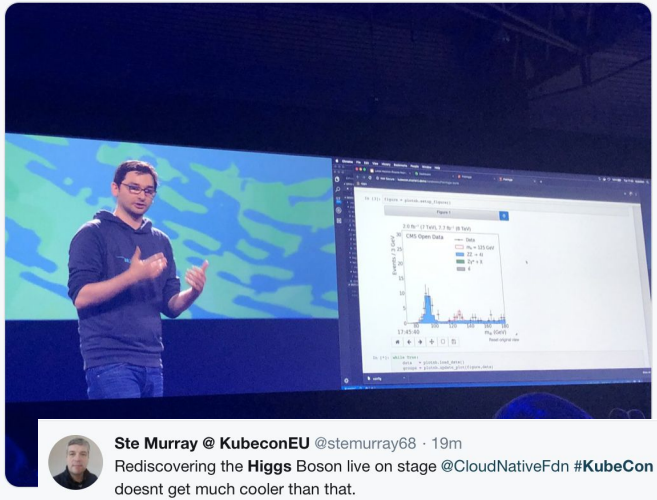
Lukas Heinrich

Physicist, CERN

Lukas Heinrich is a particle physicist working on the ATLAS Experiment. He focuses on introducing modern cloud computing tools to more systematically search for phenomena beyond the Standard Model of Particle Physics.

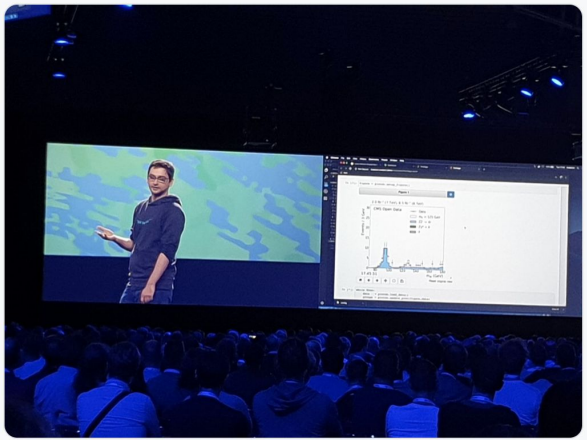
Brian Amedro @brianamedro · 11m

Really cool live demo running on GKE over 25,000 cores to (re) discover **Higgs** boson! Congrats to @ahcorporto @lukasheinrich_ and the community! **#KubeCon**

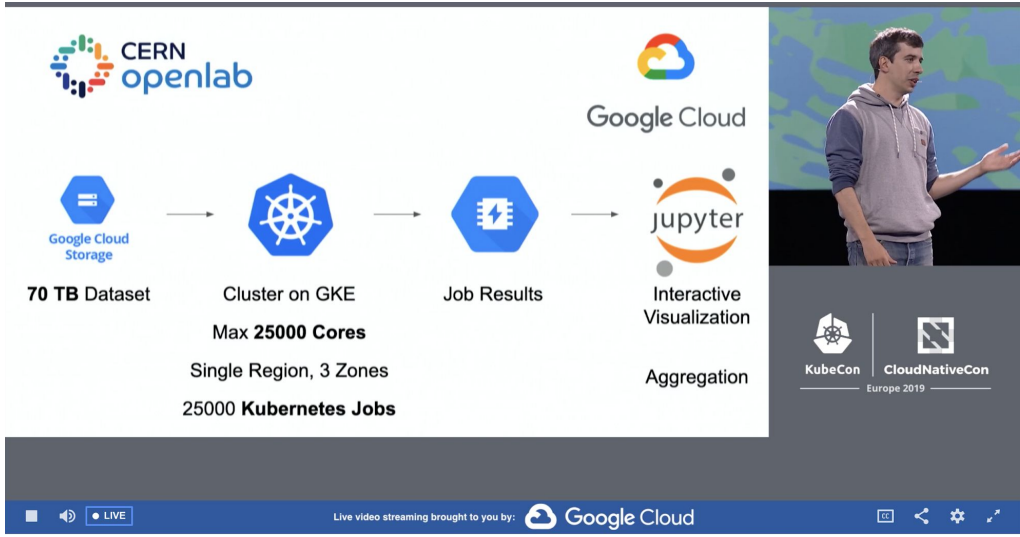


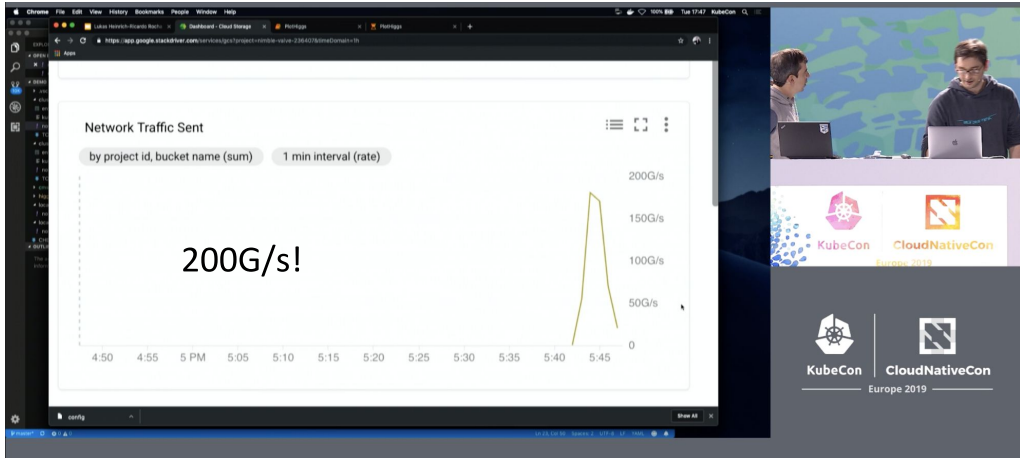
Ste Murray @ KubeconEU @stemurray68 · 19m

Rediscovering the **Higgs** Boson live on stage @CloudNativeFdn **#KubeCon** . It doesnt get much cooler than that.



Live streamed earlier today!

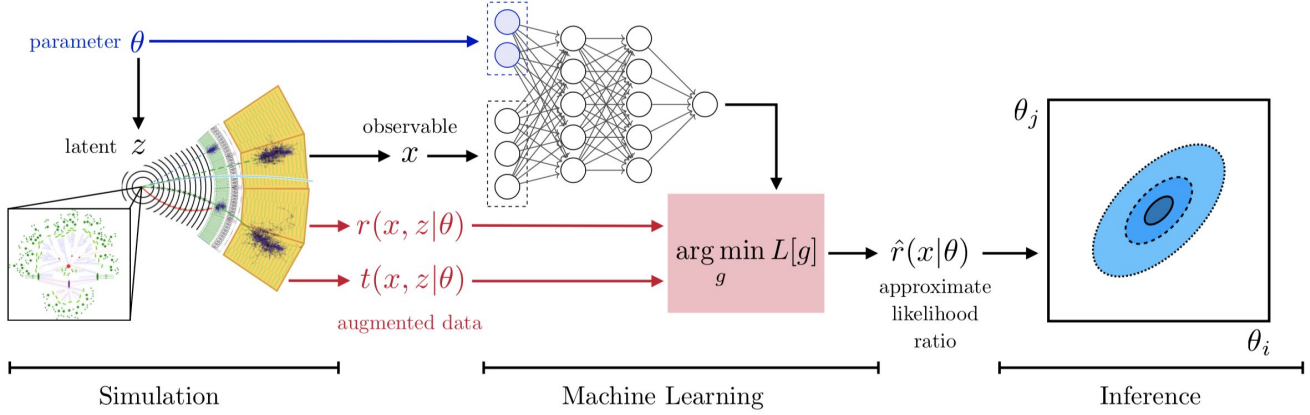


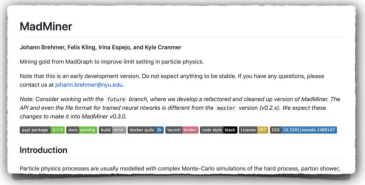


MadMiner

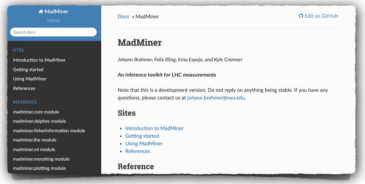
Tight integration of

- Simulation
- Machine Learning
- Statistical Inference

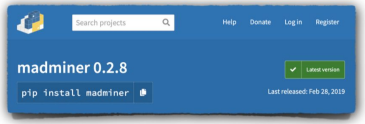




Repository and tutorials:
github.com/johannbrehmer/madminer



Documentation:
madminer.readthedocs.io

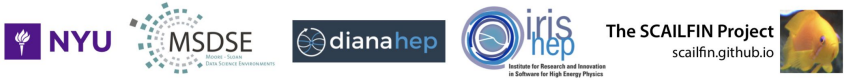


Installation:
`pip install madminer`

Deployment with Docker, yadage, REANA:
github.com/irinaespejo/workflow-madminer

34/40

Thanks to Kyle, Gilles, Felix, Irina, and Sam for material and inspiration for slides!



Slides from Johann Brehmer's Keynote talk at ACAT on Constraining Effective Field Theories with Machine Learning

Fitting Specification & Implementation

Leveraging accelerators (GPU, TPU) & ML frameworks

build passing License GPL v2

AMPGEN

AmpGen is a library and set of applications for fitting and generating multi-body particle decays using the isobar model. It developed out of the MINT project used in the fitting of three and four-body pseudoscalar decays by the CLEO-c and LHCb colloborations. The library can handle the fitting and generation of a wide variety of final states, including those involving fermions and photons, as well as polarised initial states.
Source code for the evaluation of amplitudes is dynamically generated by a custom engine, JIT compiled and dynamically linked to the user programme at runtime, which results in high flexibility and performance.

Acknowledgements

The development of this software has been supported by the National Science Foundation under grant PHY-1414736 and through a subcontract under Cooperative Agreement OAC-1836650. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the developers and do not necessarily reflect the views of the National Science Foundation.

University of CINCINNATI

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

NSF

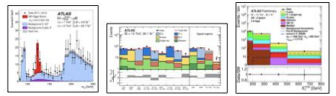
pyhf: auto-differentiable binned HEP likelihoods

Kyle Cranmer (NYU) Matthew Feickert (SMU)
Lukas Heinrich (CERN), Giordon Stark (UCSC)

HistFactory

declarative binned likelihoods

One of the most widely used statistical models in HEP for binned measurements and searches.



Standard Model SUSY Exotics

Mathematical Formulation:a parametrized p.d.f with parameters or interest (POI) and nuisance parameters

$$\mathcal{P}(n_c, a_c, a_p \mid \phi_p, \alpha_p, \gamma_b) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c \mid \nu_c) \prod_{e=1}^{n_c} f_c(x_e \mid \boldsymbol{\alpha}) \right] \cdot G(L_0 \mid \lambda, \Delta_L) \cdot \prod_{p \in \mathbf{S}+\mathbf{F}} f_p(a_p \mid \alpha_p) \quad (1)$$

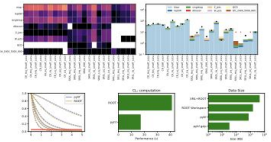
Primary Measurment:
- multiple disjoint "channels" (e.g. event observables), each with multiple bins
- Poisson with rate parameters as functions of poi and nuisance parameters

Auiliary Measurement
- constraint terms on modeled as "measurements" of auxiliary data

Performance

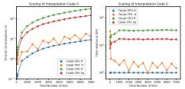
fast likelihood computation

efficient use of tensor computation makes pyhf fast.
Competitive with ROOT implementation - often faster.



Hardware Acceleration

For ML-library tensor backends Computational graph can be transparently placed on hardware accelerators: GPUs and TPUs for order of magnitude speed-up in computation.



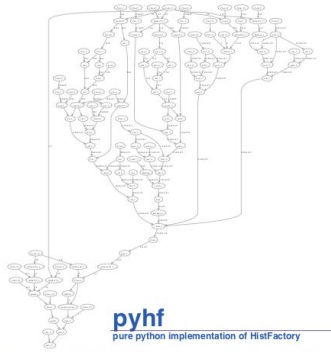
Sharable Likelihoods

built for reinterpretation

For reinterpretation of searches, the bulk of the likelihood is invariant. Only need to patch the likelihood. Use JSONPatch standard to inject the new signal.

```
[
  {"op": "replace", "path": "/channels/0/samples/0/data", "value": [5., 6.]}
]
```

```
$> pyhf cls original.json|jq .CLs_obs
0.05290115065118097
$> pyhf cls original.json --patch newsignal.json|jq .CLs_obs
0.3401578753020146
```



pyhf

pure python implementation of HistFactory

Implementation of HistFactory likelihood (1) as a computational graph of multi-dimensional array operations.

Use of array ("tensor") operations through a common API layer around high-performance tensor libraries: e.g.

NumPy TensorFlow PyTorch

Installation:

```
$> pip install pyhf
```

Example: simple number-counting experiment

Auto-Differentiation:

Tensor libraries from ML community provide exact gradients for use in minimization. $\frac{\partial \mathcal{L}}{\partial \mu}, \frac{\partial \mathcal{L}}{\partial \theta_i}$

Optimizers

pyhf likelihood are simple tensor-value python functions. Can use multiple minimization algorithms, such as `scipy.minimize` or MINUIT

JSON Format

The full likelihood can be expressed as a single JSON document

- easy archivability (HepData)
- easy sharing across network
- easy manipulation

```
$> curl http://url-to-json/workspace.json|pyhf cls
{
  "CLs_exp": [
    0.0026844885279398,
    0.014659866847622592,
    0.07807427329745232,
    0.2351618248553596,
    0.5738410837288444
  ],
  "CLs_obs": 0.052901150651180970
}
```

```
measurements:
  - {name: demo, config: {poi: mu}}
data:
  singlechannel: [51, 48]
channels:
  - name: singlechannel
    samples:
      - name: signal
        data: [12, 11]
        modifiers:
          - name: mu
            type: normfactor
            data: null
      - name: background
        data: [50,52]
        modifiers:
          - name: uncorr_bkguncrt
            type: shapesys
            data: [3,7]
```

Preservation & Reinterpretation

The field is at a tipping point, DIANA/DASPOS/IRIS-HEP contributions have been transformational.

Archiving Real ATLAS Analyses

Using **Industry Standard** Software Packaging to archive analysis:

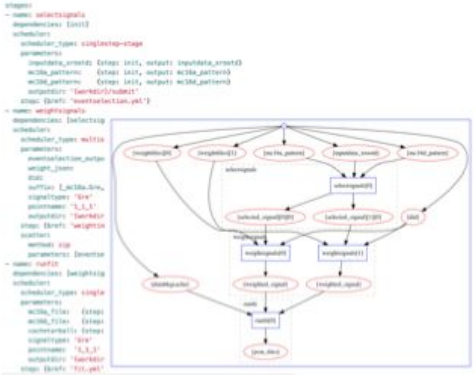
- Linux Containers ("Docker")
- Integrated into existing analysis infratructure (revision control, continous intergration, grid computing)

Plain-text JSON formats to capture commands and workflows

Close coordination with CERN Analysis Preservation / Reuse Projects

reana

CERN Analysis Preservation



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.

Blue Print (Y1 milestone G2.5)

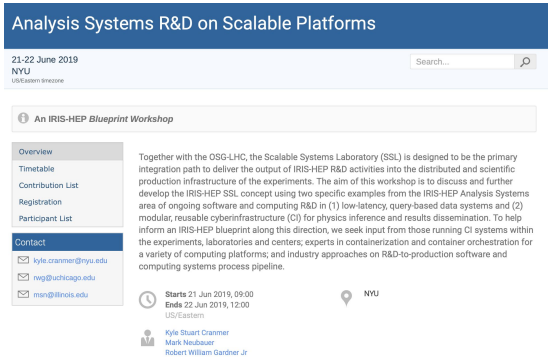
Major Goals

- Develop a near/medium-term AS requests for SSL, using the two specific AS R&D thrusts mentioned in the overview as concrete examples
- Summarize current status of progress on R&D in the two AS R&D thrusts
- Strengthen visibility of SSL and socialize the resource needs with Ops Programs, Labs and Centers and OTHERS
- Get informed by experts on latest developments in technologies and method relevant for SSL+AS

Desired Outcomes & Deliverables

- Produce a near and medium-term AS R&D area roadmap for SSL
- Produce a sharpened Plan for specific R&D in these AS areas
- Establish partnerships with those who have (or will have) resources appropriate for SSL+AS, who will then make it rain on the SSL
- Experts not directly connected to our work either validate our approaches or inform other directions that we might take

This will be first Blueprint workshop



https://indico.cern.ch/event/820946/

More details in Blueprint Dashboard

Analysis Systems Topical Meeting

https://indico.cern.ch/event/822074/

Will hold Analysis Systems topical meeting on June 19 & 20th at NYU

- Want to do it soon to prepare Y1 reporting and Y2 planning
- The ATLAS SW & Computing week is at NYU following week
- The AS-SSL blueprint is June 21 & 22
- Include partner projects: SCAILFIN, Coffea, HSF, etc.
- We just picked these dates, planning in earnest
  - Agenda
  - Finalize / sign off on several Y1 milestones & deliverables
  - Initiate discussions on Y2 milestones associated to roadmaps

Y2 planning

	Analysis Systems									
4	Description	Type (M/D)	Y1Q1	Y1Q2	Y1Q3	Y1Q4	Y2Q1	Y2Q2	Y2Q3	Y2Q4
5	Organize topical meetings, Analysis System group meetings, etc.									
6	List publicly-accessible repositories and other relevant documentation on the iris-hep.org website									
7	Collect and curate example analysis use cases with some existing reference implementation									
8	Survey of analysis systems efforts in the field to aid in planning for topical workshop									
9	Blueprint workshop coordinating resource needs for evaluating analysis systems coordinated by SSL with participation of operations program									
10	Develop initial specifications for user-facing interface to analysis system components									
11	Prototype awkward-array analyses in the scientific Python ecosystem									
12	Initial roadmap for ecosystem coherency									
13	Develop initial design for interface of analysis query system to the IDDS									
14	Translate analysis examples into new specifications, provide feedback, iterating as necessary									
15	Initial roadmap for high-level cyberinfrastructure components of analysis system									
16	Benchmarking and assessment of existing analysis systems									
17	Implement prototype query-based and cache-aware dispatch									
18	Establish analysis description database schema and integrate with archival tools like CAP, INSPIRE, HEPDATA, etc.									
19	GPU/accelerator-based implementation of statistical and other appropriate components									
20	Move prototypes of analysis system components to SSL									
21	Benchmarking and assessment of prototype analysis system components									

Scheduled for June

AS June Meeting Topic
(also relevant blueprint WS)

SSL Blueprint

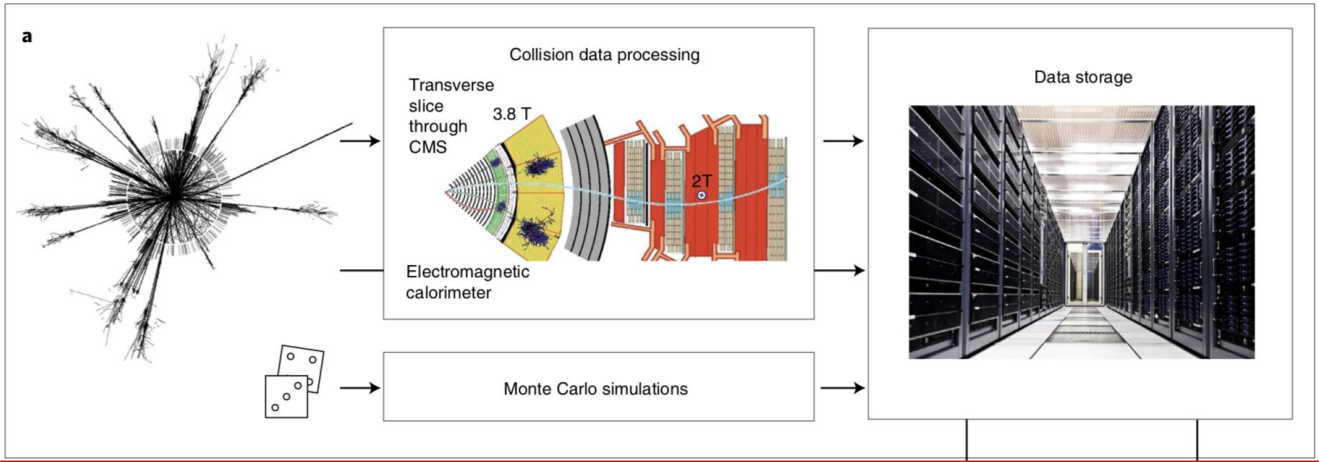
Conclusion

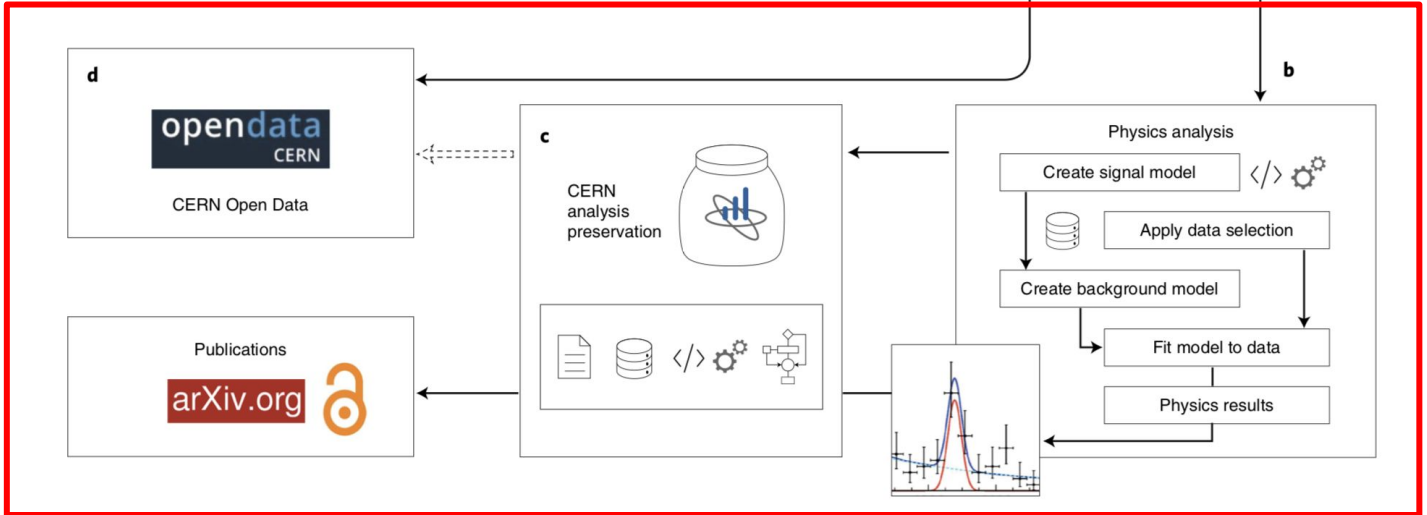
“just wanted to express my personal awe to you and your team working so hard on a bunch of wonderful projects. The talks delivered by Johann, Lukas and Gunes were excellent! In my personal opinion it was the best part of ACAT conference.”

- Andrey Ustyuzhanin (LHCb & Yandex School of Data Analysis)

- Making concrete progress in each focus area goal
- Establishing & strengthening connections among different components to build cohesive effort
- Good visibility and credibility as an intellectual hub

Backup





Scope of analysis systems

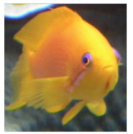
External Collaboration: SCAILFIN

- Not developing methodology, but implementing them in scalable distributed systems
- Theme = ML methods + Workflows and distributed systems
- Emphasis on use-cases that involve simulation+ML together and are iterative in nature (not static bulk processing)

Scalable Cyberinfrastructure for Artificial Intelligence and Likelihood-Free Inference (SCAILFIN) Project

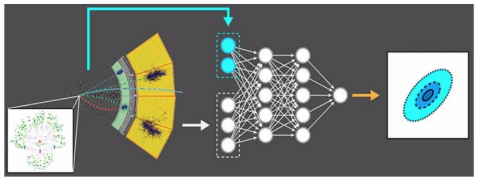
OAC-1841456, 1841471, 1841448

Kyle Cranmer (PI) and Heiko Mueller (co-PI), *New York University*
Michael Hildreth (PI), *University of Notre Dame*
Mark Neubauer (PI) and Daniel S. Katz (co-PI), *University of Illinois at Urbana-Champaign*



scailfin.github.io

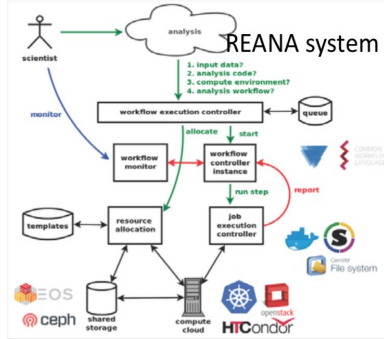
Main goal: to deploy artificial intelligence and likelihood-free inference (LFI) techniques and software using scalable cyberinfrastructure (CI) to be integrated into existing CI elements, such as the REANA system.



Primary science driver: Analysis of data from the Large Hadron Collider; technology is sufficiently generic to be widely applicable beyond LHC

Developing scalable CI components though SCAILFIN will catalyze convergent research because:

1) Abstract LFI problem formulation has already been demonstrated as the *lingua franca* for a diverse range of scientific problems
2) Current tools for many tasks are limited by a lack of scalability for data-intensive problems with computationally-intensive simulators
3) Tools will be designed to be scalable and immediately deployable on a diverse set of computing resources, including HPC
4) Integrating commonly-used workflow languages to drive an optimization of machine learning components and orchestrate large-scale workflows will lower the barrier-to-entry for researchers from other domains



Broader Impacts:

- Advance education of next generation of computationally-trained scientists, including students & postdocs
- Contribute to workforce development through 1) aligning HEP tools & practices with industry computing standards, and 2) training in high-value machine learning and data science skills
- Promote diversity, particularly at smaller universities, by lowering barriers for using advanced analysis methods

pyhf: python implementation of HistFactory (Cranmer)

HistFactory v1 specification implemented in ROOT used widely in ATLAS. Similar to CMS Combine for binned models. Now implemented in **pyhf**.

HistFactory Template

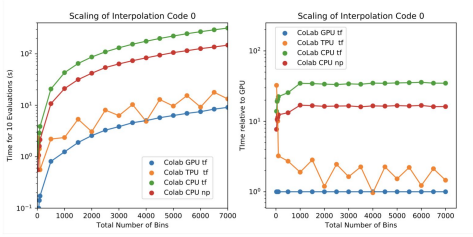
$$\mathcal{P}\left(n_c, x_e, a_p | \phi_p, \alpha_p, \gamma_b\right) = \prod_{c \in \text{channels}} \left[\text{Pois}\left(n_c | \nu_c\right) \prod_{e=1}^{n_c} f_c\left(x_e | \vec{\alpha}\right)\right] G\left(L_0 | \lambda, \Delta_L\right) \prod_{p \in \mathbb{S}+\Gamma} f_p\left(a_p | \alpha_p\right)$$

This is a mathematical representation! Nowhere is any software spec defined

Until now, the only implementation of HistFactory has been in RooStats+RooFit

- To start using HistFactory p.d.f.s first have to learn ROOT, RooFit, RooStats
  - Problem for our theory colleagues (generally don't want to)
- Possible issues with scaling I/O and memory for large models
  - Not multithreaded
- Difficult to interface with modern tools for minimization and computation of the p.d.f.
- Likelihood stored in the binary ROOT format
  - Challenge for long-term preservation (i.e. HEPData)
  - Why is a histogram needed for an array of numbers?

- First GPU and TPU results running on Google Colab show order of magnitude speedup for interpolators on large models



Machine Learning Frameworks for Computational Backends

All numerical operations implemented in **tensor backends** through an API of n-dimensional array operations

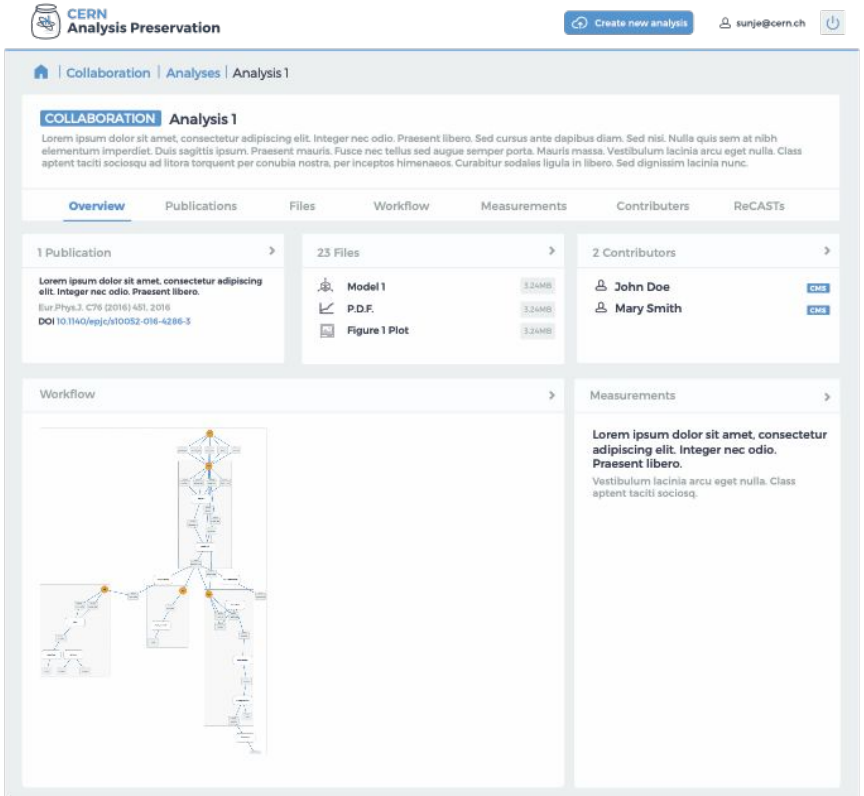
Allows for switching out the tensor backend with a single call

Using deep learning frameworks as computational backends allows for **exploitation of auto differentiation (autograd) and GPU acceleration**

As huge buy in from industry we benefit for free as these frameworks are **continually improved** by professional software engineers

M. Feickert Talk on pyhf at DIANA/HEP

CERN IT connections



reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.

COMMON WORKFLOW LANGUAGE

Scalable

Support for remote compute clouds.

kubernetes

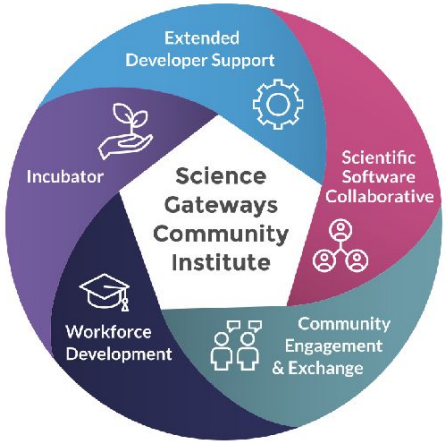
Reusable

Containerise once, reuse elsewhere. Cloud-native.

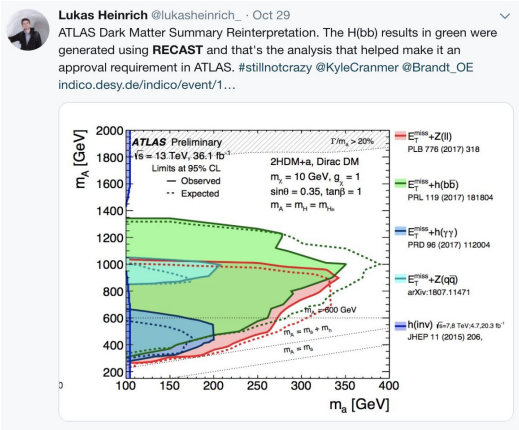
Free

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

Connections with Science Community Gateways Institute

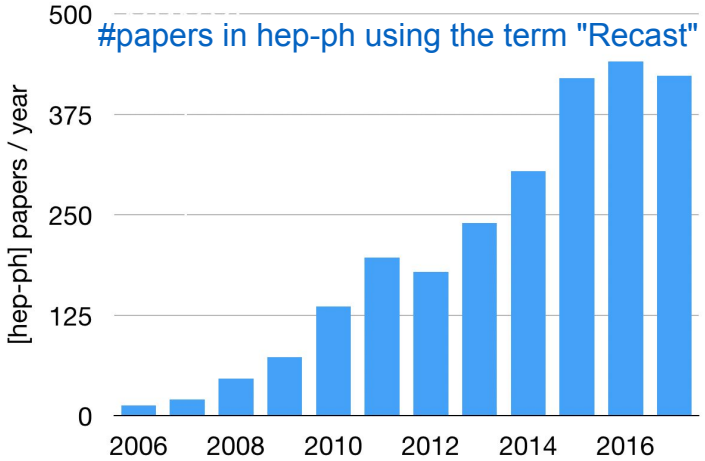


- Gateways are ideal for improving the Theory/Experiment interface
- Eg. Reinterpretation and "Recasting"

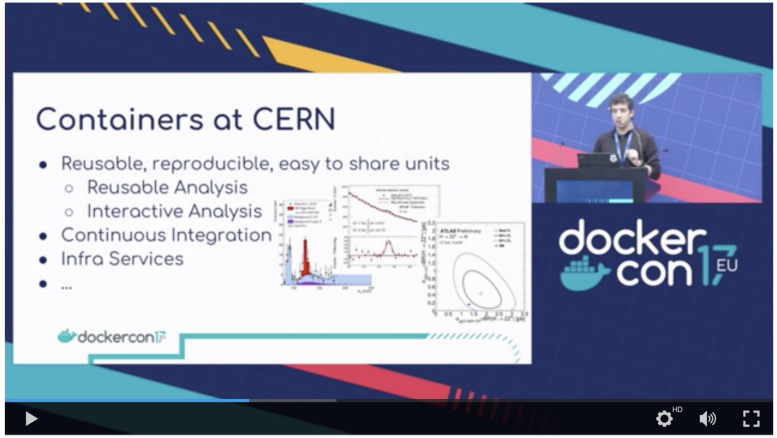


What is a Gateway?

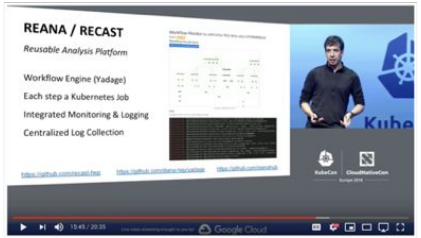
Science gateways allow science & engineering communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines. Read more on Wikipedia.



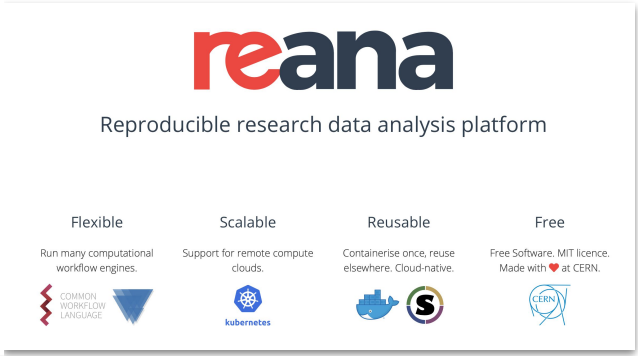
Connections with SSL



https://youtu.be/2PRGUOxL36M?t=15m43s



Containerd maintainer

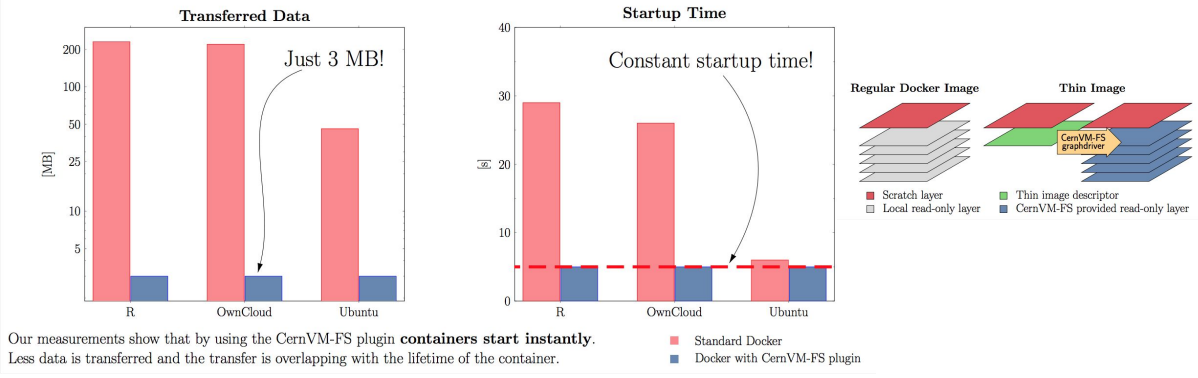


Integrating CERN Software Distribution with Docker

CVMFS is the CERNVM Filesystem, a scalable, reliable and low maintenance software distribution service. It offers a POSIX interface to a set of read-

Connections with OSG

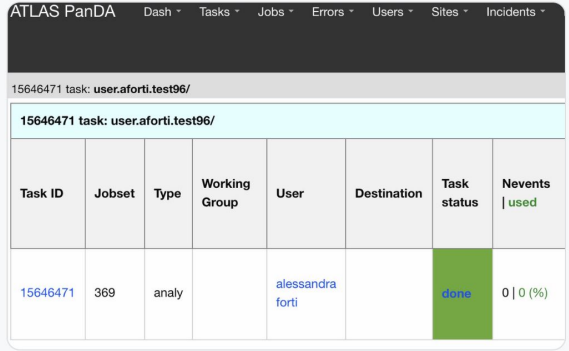
- Work to have interoperability between containerized end-user analysis jobs (that are natural with GitLab Continuous Integration and CAP/REANA) and GRID jobs.
- Common objective to solve image distribution at GRID scale (e.g: cvmfs containerd integration)



eg: See Blomer's ACAT Poster

Lukas Heinrich @lukasheinrich_ · Oct 11

1/ prun --containerImage docker://afortiorama/lukasanalysis --exec "/code/run_analysis.sh 404958 recast_sample 0.00122 /data/test.root /data/workdir 30.0" --outDS user.aforti.test96 --outputs hist-sample.root --site ANALY_MANC_TEST_SL7 --noBuild --inDS user.aforti:lukastest



1 3 11

Lukas Heinrich
@lukasheinrich_

Following

2/ today we ran the first job on the GRID that pulls user analysis code on top of the ATLAS analysis base images as a container image from Docker Hub (thanks Alessandra Forti!). This is a huge step to integrating reproducibility/reusability into standard LHC analysis workflows.

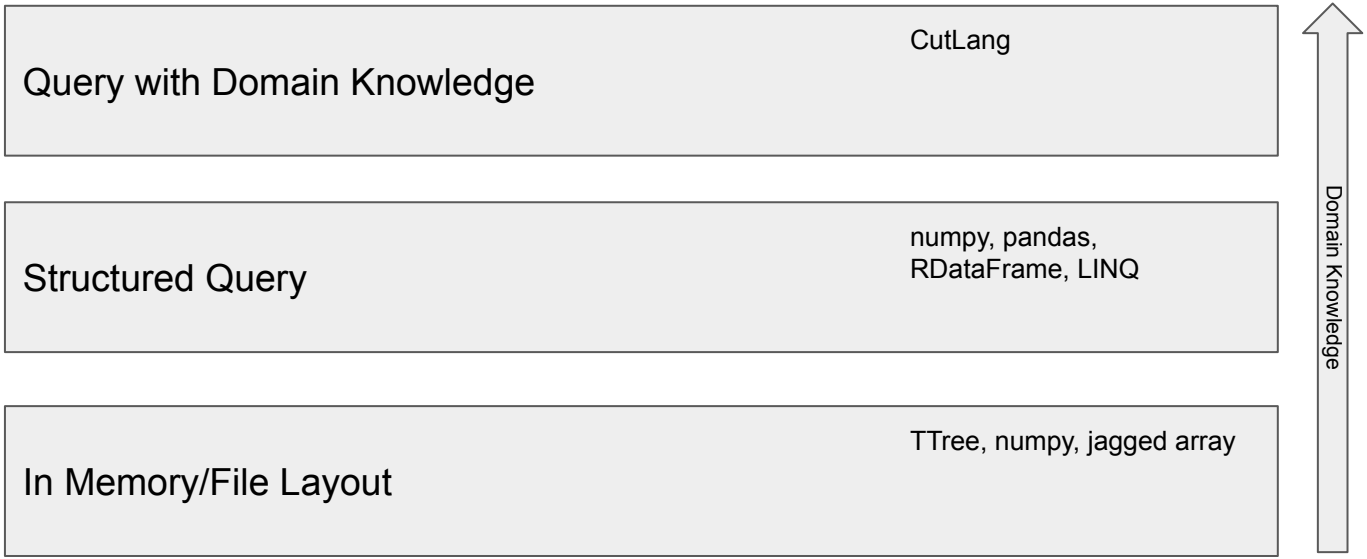
Lukas Heinrich @lukasheinrich_ · Oct 11

3/ We've been advocating for this for a long time as part of the RECAST project and analysis preservation efforts, so it's nice to see it come together. cc @KyleCranmer @ahcorporto

1 6

Components of the Analysis Language Hierarchy

Analysis languages translate the intent of the physicist into the code that does the work. They can be loosely arranged by how much domain knowledge they contain, from binary in memory/file formats that are very flexible to languages that are really only appropriate for a particular type of experiment (LHC collider, or perhaps a large nuclear experiment).



The electron is a first class object, specific to class of experiment.

Data model contains object definitions, data structure is part of the language, experiment agnostic

Data model contains all information, field and experiment agnostic