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

L Heinrich, S Dallmeier-Tiessen¹, T Simko¹
4th Reinterpretation Workshop

May 15th 2018
CERN Analysis Preservation Portal

LHC-wide online portal to help experiments manage the preservation of their analysis throughout the entire analysis lifecycle

- manage meta-data: analysis team, searchable characteristics of analysis (final states, triggers used, etc...)
- high-value data assets that need preservation (Ntuples, final background estimates, statistical models, UFO models, etc..)
- source code preservation
- containerized software environments
- Computational Workflow Recipes
REANA — reusable analysis platform

Cloud-backed Computational Workflow platform, both for actively developing analyses as well as re-using analyses stored in CAP

Support:

- non-interactive “batch-like” workflows
- interactive workflows steered from a CLI interface or a browser
- designed to support multiple workflow engines
  - yadage
  - Common Workflow Language (CWL)
  - (nextflow? Parsl? snakemake?)
Putting it all together

- Ingest data
- Experiment internal collection
- Get analysis information
- Reproduce generic analysis results
- Run workflows in cloud
- Submit recast-specific re-executions

User can query and use preserved analyses

Dedicated UI for frequent re-use scenario

CERN Cloud Computing Infra

Ingest data

Get analysis information

Reproduce generic analysis results

Run workflows in cloud

Submit recast-specific re-executions

User can query and use preserved analyses

Dedicated UI for frequent re-use scenario

CERN Cloud Computing Infra
Running analyses on the Cloud

full-fidelity analyses need larger scale than e.g. CheckMate, Rivet, etc..

Computational Workflow
+ Container-based infrastructure
are a winning combination

(already widely used in bio-informatics
and genomics, but very applicable in
HEP — backend of REANA/RECAST)

Linux containers are ideal
packaging tool to archive
software environments and
distribute them across clouds

Experiments starting to adopt them
as part of their analysis workflow

R Rocha (CERN IT) presenting core REANA component
(distributed yadage) during a KubeCon 2018 keynote
Experience inside of ATLAS

ATLAS has been heavily investigating using containers as a part of the analysis model / analysis preservation strategy

• officially supported base software environments onto which user analysis code can be added + preserved as a single package
• preserving analyses (e.g. event selection) becomes a matter of a few-line text file
• is integrated in official documentation / recommendation on how to setup new ATLAS analysis code (e.g. for upcoming full Run-2 analyses)

Dockerfile 190 Bytes

FROM atlas/athanalysis:21.2.23
ADD . /xampp/XAMPPmonoH
WORKDIR /xampp/build
RUN source ~/release_setup.sh && \
    sudo chown -R atlas /xampp && \
    cmake .. /XAMPPmonoH && \
    make -j1

preserving a event selection
Experience inside of ATLAS

Successfully tested: Re-using analyses with minor modifications

- preserved analyses are not “black boxes”
- preservation format suitable such that modifications to analyses can be made at time of re-use

RPV-RPC project: Adding new systematics for displaced objects

Systematics on displaced objects

- Displaced jet energy scale
  - Studied jet response in MC as a function of radial decay length
  - Almost no variation in response up to ~1m in radial decay length
  - Response increases linearly up to 30% above the nominal value at 1.6 m
  - Assign deviation as uncertainty, conservative as data might follow the same trend

- Displaced jet b-tagging
  - Run tracking variations on signal MC, assign deviation from nominal as systematic
  - Variation of 20% on SRs with ≥ 4 b-tags, and a signal with 1 ns lifetime

- Displaced lepton systematics
  - No systematic, leptons require impact parameter cuts (z0 and d0 significance)
  - no acceptance of displaced leptons

- Missing energy uncertainties
  - Hard object systematics are propagated to MET
  - No uncertainty on soft term, tracks required to originate from primary vertex
  - No uncertainty on MET trigger, no variation seen on trigger efficiency turn-on
Experience inside of ATLAS

Reinterpreting mono-H searches (work in progress)

- needed to drop certain signal regions, as simulation was not reliable there
- re-use the prepackage code, but introduce additional steps in the workflow to handle details

Use case: MonoH(bb) for 2HDM+a

Analysis: Exotics Search for H(125) -> bb + MET
Goal: re-run analysis for fancy new signal model: 2HDM+a
→ use existing yadage preservation with new input samples
But...
→ Some samples are ATLFAST-II, and part of analysis is not AFII-safe
→ Some samples need to be merged to give physical distribution
... requiring modifications of the existing workflow

Experiment-internal preservation of analysis allows us to carefully re-use specific pieces of an analysis and adapt where necessary.

→ hard to do outside of the experiments

L Henkelmann
Experience inside of ATLAS

Within ATLAS, we are building infrastructure to allow to easily re-run analyses*, in which they were not originally involved.

```bash
$> yad submit workflow.yml \
  gitlab-cern:prieck/monoHbb16_preservation:yadage \
  -o fit/limits.txt -o fit/plots \
  -f pars.yml
```

* when preserved using latest recommendations

specify which analysis
specify desired outputs
specify new inputs (new signal files, cross-sections)
Developments in CMS (Clemens Lange, CERN)

Investigating similar strategies as ATLAS: e.g. preserve analyses using experiment-provided software images. Use Kubernetes-based workflow engines to steer the analysis. Looking to integrate with REANA

Preservation options

> Docker containers:
  - fully self-contained images (15-17 GB) + 1 GB of conditions database per CMSSW release
  - light-weight image mounting CVMFS (via FUSE) inside container (~700 MB) — slow
  - light-weight image mapping CVMFS (via GraphFS) as volume (~700 MB) — fast (not yet in repository)
  - repository: https://github.com/clelange/cmssw-docker
  - these containers are already being used for analysis development by individual analysts

> Kubernetes:
  - possibility to pass kerberos token and VOMS proxy to Kubernetes jobs — fully authenticated
  - example runs data set query and determines cross section for a given simulation sample (requires CMS VOMS membership)
  - repository: https://github.com/clelange/cmssw-k8s
  - currently exploring different workflow tools for chaining
Integrating alternative analysis implementations

REANA / RECAST are purposely agnostic to how the analysis is implemented (→ needed to support full analyses)

Makes it easy to integrate other analysis catalogues. Currently have

1. SimpleAnalysis (ATLAS-internal, truth-level implementations)
2. Rivet (using the official Rivet Docker Images)
3. CheckMate (current using custom image, but helping CM authors to automate release of image together with HepForge → GitHub transition)
4. MadAnalysis: not yet :( but interested to make this work
REANA Integration / Status

Much of ATLAS-internal / RECAST-internal infrastructure will be offloaded to REANA.

- REANA 0.2.0 released in April 2018
  - support Common Workflow Language (CWL) in addition to yadage
  - improved command lines tools, monitoring

- Work in progress
  - more workflow engines, for easy on-boarding of analyzers
  - support more container runtimes (singularity, etc…)
  - seamless integration into CAP
  - *lots* of documentation. It’s a new paradigm in HEP, so need good docs to make user’s feel at home.
REANA Examples:

REANA demo example of LHCb D→ππμμ analysis

REANA example

About

This example is based on analysis-case-study.

Mass fit results

The result of this analysis are the following plots in various dimuon mass ranges. We studied the three body decay high dimuon and low dimuon mass range, and we did not observe any signal.

<table>
<thead>
<tr>
<th>Dimuon resonances</th>
<th>Dimuon mass range (MeV)</th>
<th>Plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three body decay (low dimuon)</td>
<td>250 - 525</td>
<td>low_dimuon_signal.pdf</td>
</tr>
<tr>
<td>η</td>
<td>525 - 565</td>
<td>eta.pdf</td>
</tr>
<tr>
<td>ρ, ω</td>
<td>565 - 850</td>
<td>rho_omega.pdf</td>
</tr>
<tr>
<td>φ</td>
<td>850 - 1250</td>
<td>phi.pdf</td>
</tr>
<tr>
<td>Three body (high dimuon)</td>
<td>1250 - 2000</td>
<td>high_dimuon_signal.pdf</td>
</tr>
</tbody>
</table>

Invariant mass of the two muons
CAP Status

• CAP/REANA workshop on June 20 (please join if you are part of the experiments!)
• Lots of development leading up to workshop
  • new UI, powerful search, many data ingestion pipes from expts
  • integrating with REANA
  • code, s/w, data archiving
CERN Analysis Preservation

http://analysispreservation.cern.ch
http://github.com/cernanalysispreservation
analysis-preservation-support@cern.ch

REANA

http://reanahub.io
http://github.com/reanahub
reanahub
info@reanahub.io

RECAST

http://recast.cern.ch