HEPData

Graeme Watt (IPPP Durham)
Belle II Data Preservation Workshop, 7th October 2022

https://hepdata.net

Email: info@hepdata.net
Forum: hepdata-forum.cern.ch

Follow @HEPData

Code: https://github.com/HEPData

What is HEPData?

- Unique *open-access* repository for high-level **data** from almost 10,000 **HEP** ("`hep-ex`" or "`nucl-ex`") papers.

- **FAIR** data: **F**indable, **A**ccessible, **I**nteroperable, **R**eusable.

- *Complementary* to other HEP information providers, e.g. INSPIRE-HEP (literature), PDG (particle properties), CERN Open Data (event-level data), Zenodo (files).

- Based in **I**nstitute for **P**article **P**hysics **P**henomenology (**IPPP**) at Durham University (UK), going back to 1970s.

- Transition in 2017 to hepdata.net site, hosted at CERN. Partnership with CERN Scientific Information Service. J. Phys.: Conf. Ser. 898 102006 [`arXiv:1704.05473`]

Search from inspirehep.net

- New <u>INSPIRE</u> (<u>blog</u>) shares many features with HEPData.

- <u>INSPIRE</u> publication records link to <u>hepdata.net</u> records.

- Search <u>INSPIRE</u> for <u>publication records with HEPData</u>.

- LHC publications with HEPData (<u>Jupyter Notebook</u>):

LHC publications with HEPData records (2022-10-05)
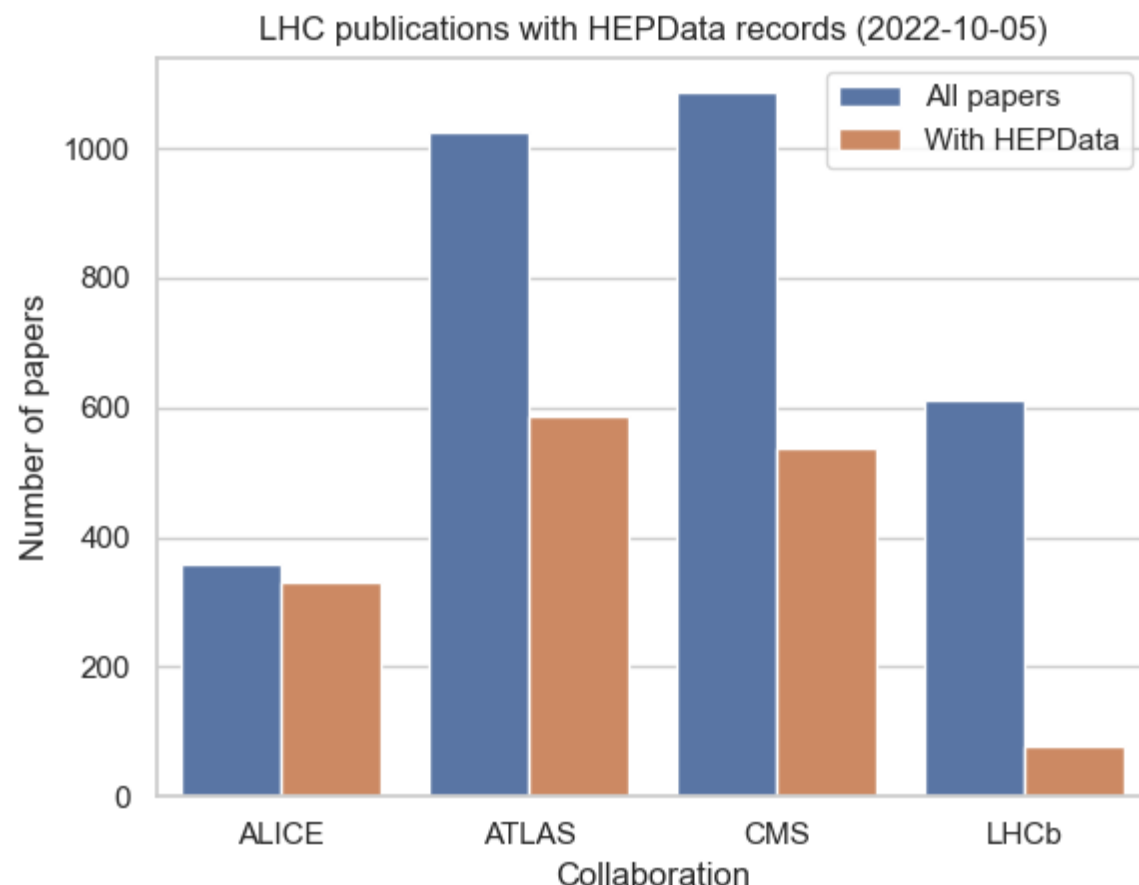
INSPIRE search query:
- `hep-ex` or `nucl-ex`
- Published in a journal
- Not conference paper

ALICE: **92%**
ATLAS: **57%**
CMS: **49%**
LHCb: **13%**

Belle: 31/533 = **6%**
Belle-II: 1/9 = **11%**

Search from hepdata.net

Connection to <u>Rivet</u> analyses

● <u>HEPData</u> exports data in <u>YODA</u> format for <u>Rivet</u>.

Record from <u>hepdata.net</u>

- <u>DOIs</u> minted for whole (versioned) record and each table.

Data output formats

YAML: native
HEPData format.
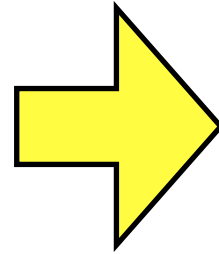
submission.yaml
+ YAML data files for each table
+ optional resource files

- JSON: JavaScript Object Notation.

- CSV: comma-separated values.

- YODA: for inclusion in a Rivet analysis.

- ROOT: binary `.root` file.

`https://www.hepdata.net/record/ins1860766?format=csv`

- Programmatic access to download data in different formats.

- `format={original,yaml,json,csv,yoda,root}`

- Additional (optional) arguments for `version` and `table`.

- NEW `curl` record w/o `format` returns Schema.org JSON-LD.

G.Watt   7

Modes of data entry

1. Manually harvested from "`hep-ex`" / "`nucl-ex`". HEPData staff extracted tables from `.tex` source.

2. Data points directly submitted by experiments.

   - Pre-2014: no guidelines on preferred format.

   - Early 2014: encourage standard "input" format.

   - Late 2014: introduce online submission system.

   - Early 2017: allow submissions from <u>hepdata.net</u>.

Mode 1. now phased out in favour of mode 2.

Submission system on hepdata.net

https://hepdata.net/submission

Coordinator
e.g. convener of physics group

Uploader
e.g. primary author of physics analysis

Reviewer
e.g. supervisor of a primary author

START Initiate new paper with INSPIRE ID or title. Invitation sent by email

Upload text files in the YAML format (+ optional supplementary files). Alternatively, .oldhepdata file.

Receives email. Checks each table.

END Make record public

Add INSPIRE ID

No   Yes

INSPIRE ID added?

OK?

No   Yes

Mark as "Attention Required" (with optional comments)

Mark all tables as "Passed" review.
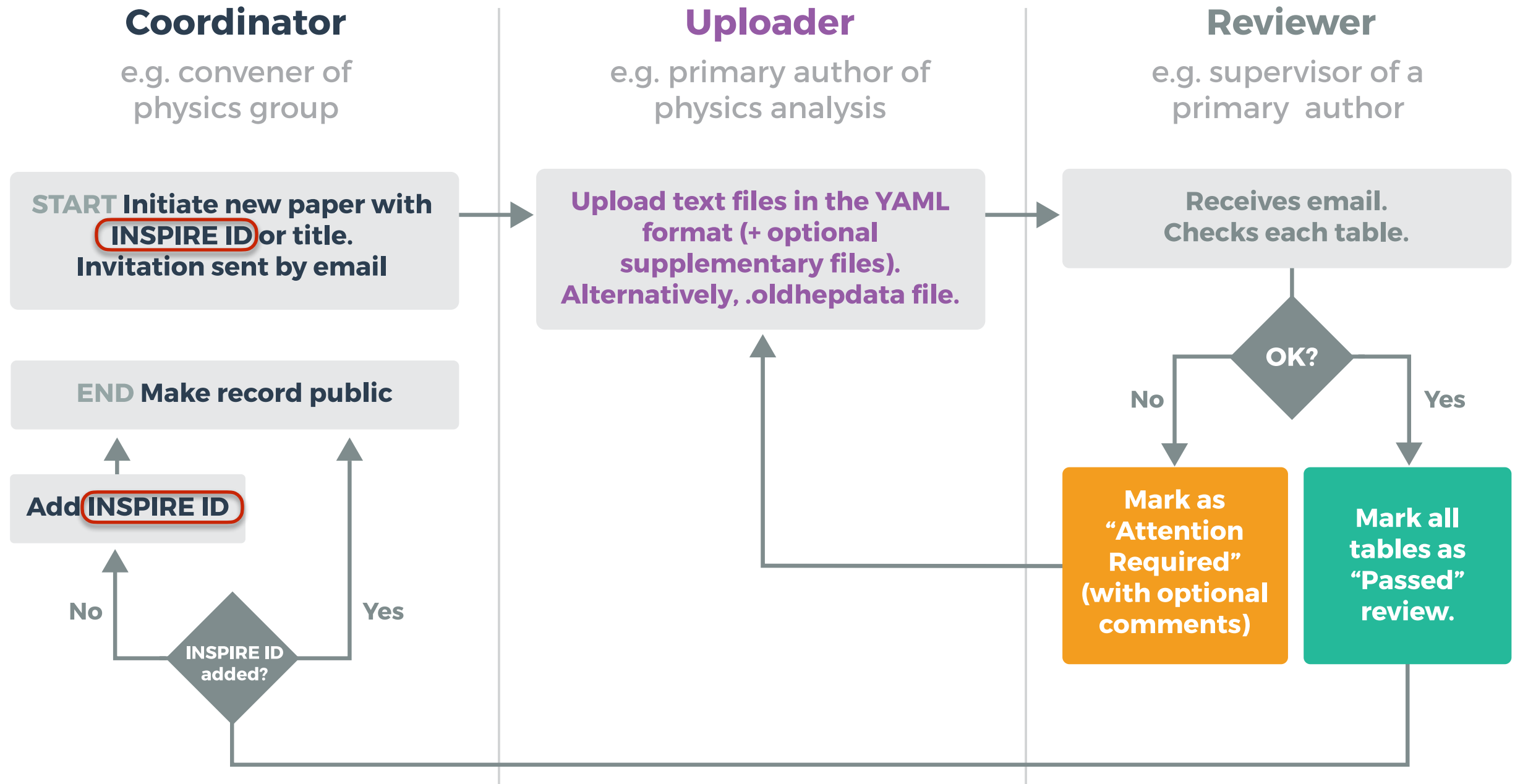
- Also a <u>Sandbox</u> for any user to test uploads without special privileges.

Submission system usage (01/2017-)

All (2022-10-05)

Non-LHC (2022-10-05)
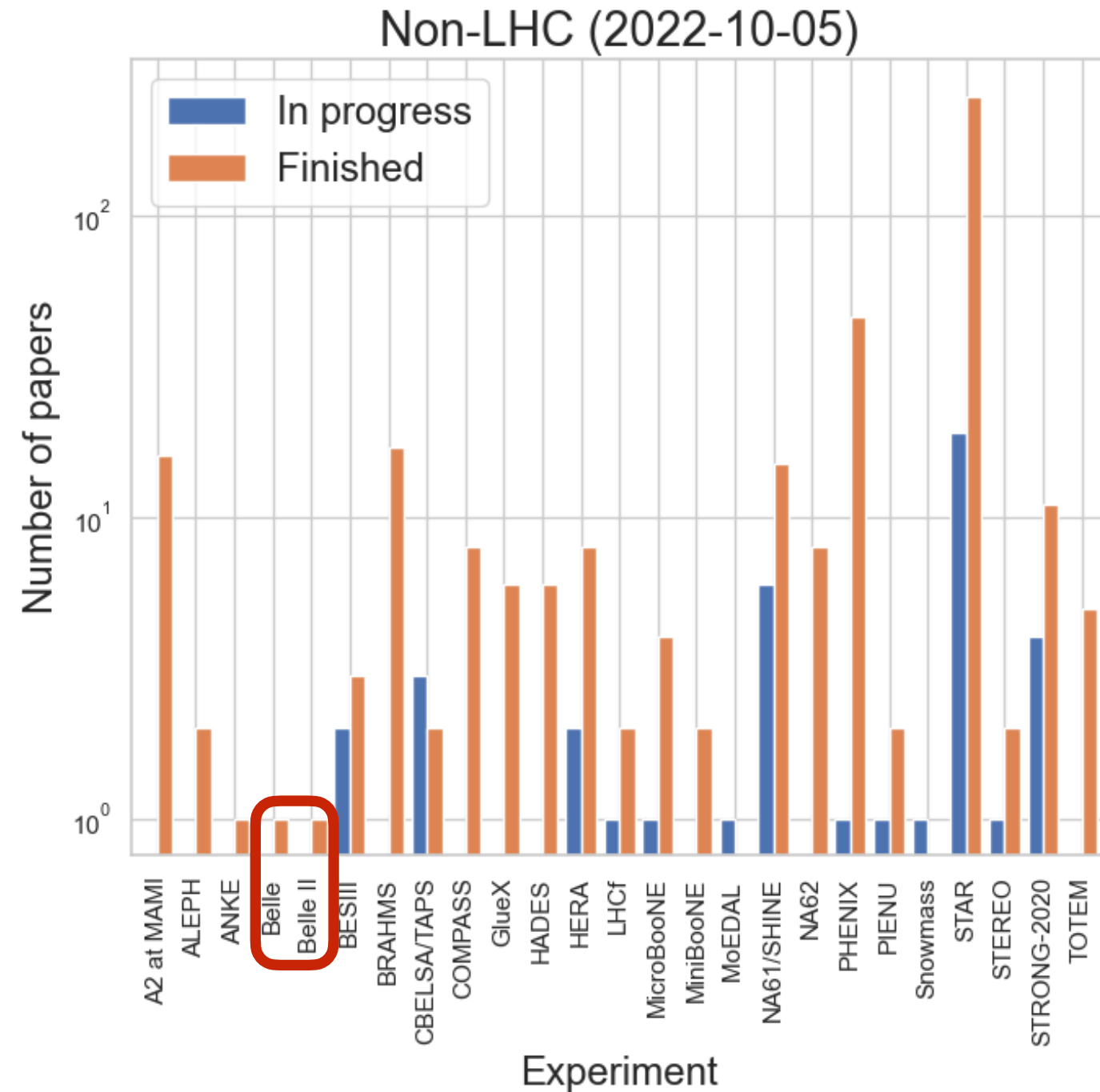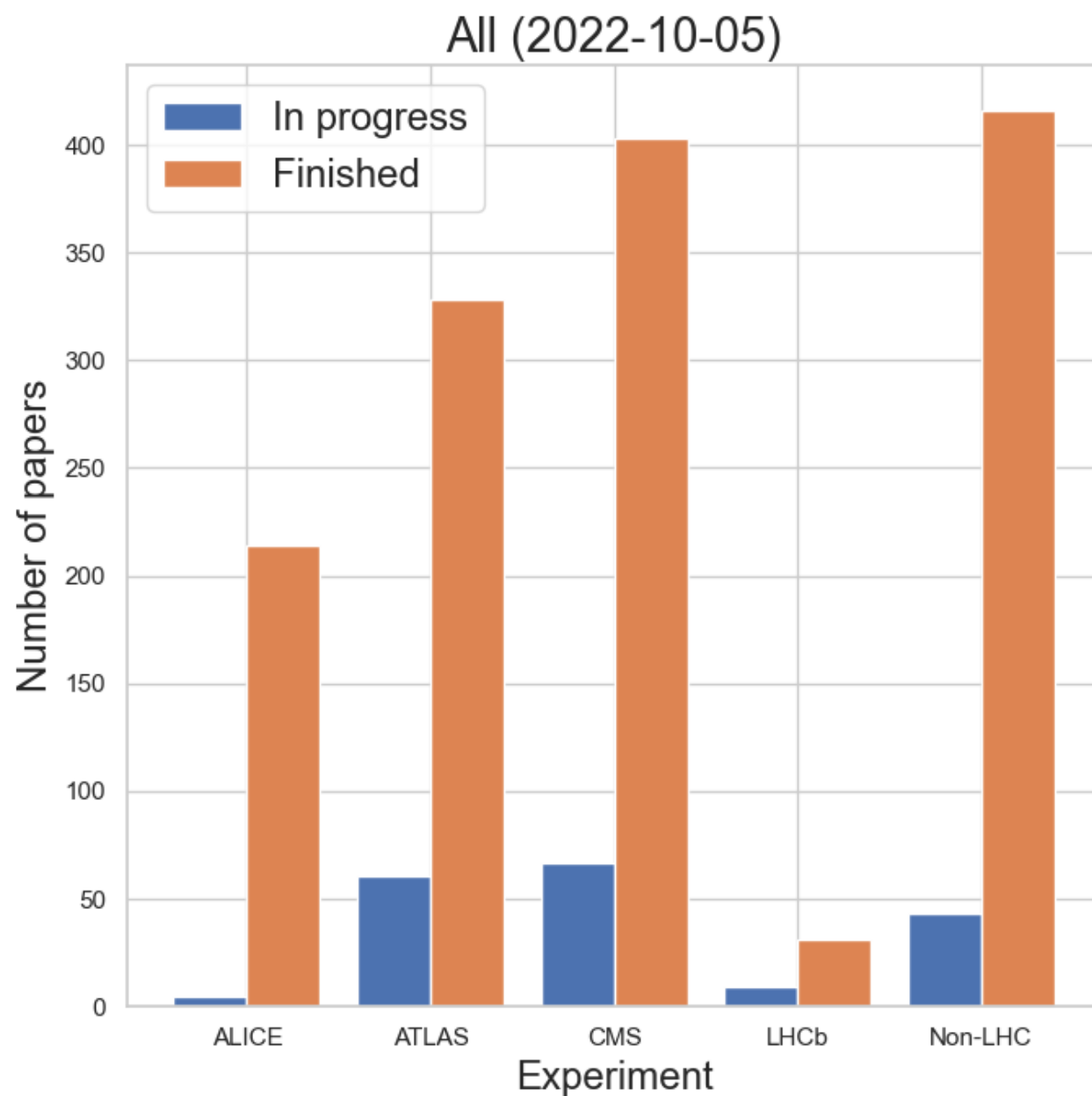
- Total of **1392** finished submissions, comprising ALICE (214), ATLAS (328), CMS (403), LHCb (31), Non-LHC (416).

Submissions per month from 2017 to 2022

hepdata.net submissions by experiments (2022-10-05)
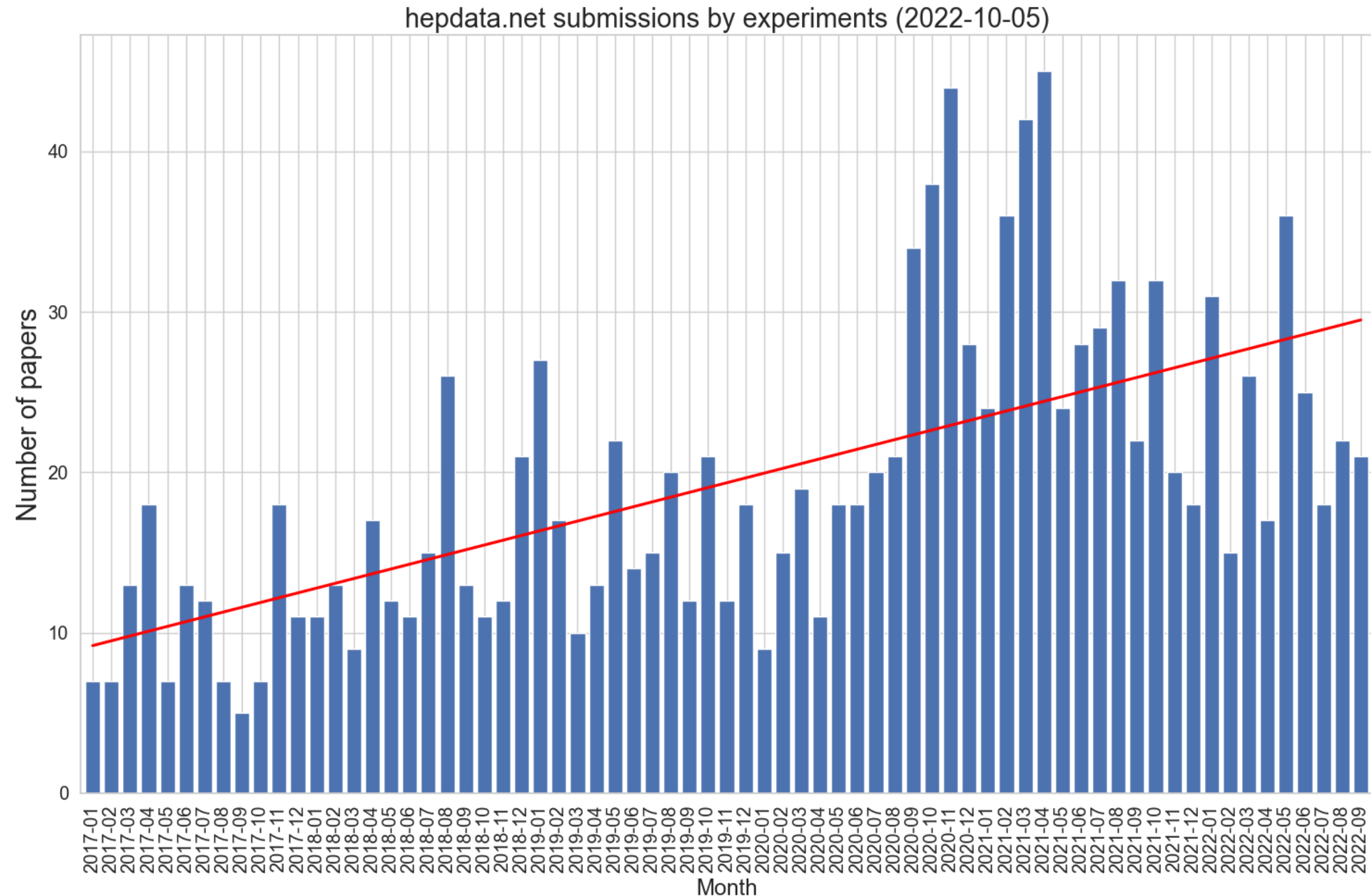
- Year-on-year growth: **125** in 2017, **171** in 2018, **201** in 2019, **275** in 2020, **352** in 2021.

`submission.yaml` file

- Separate <u>YAML</u> "document" for each data table, e.g.

```yaml
---
# This is Table 3.
name: "Table 3"
location: Data from Figure 8A
description: Normalized ZZ fiducial cross section (multiplied by 10^6 for
keywords: # used for searching, possibly multiple values for each keyword
- {name: reactions, values: [P P --> Z0 Z0 X]}
- {name: observables, values: [DSIG/DPT]}
- {name: cmenergies, values: [7000.0]}
- {name: phrases, values: [Inclusive, Single Differential Cross Section,
data_file: data3.yaml
additional_resources:
- {description: Image file, location: figFigure8A.png}
- {description: Thumbnail image file, location: thumb_figFigure8A.png}
```

- Validated against <u>`submission_schema.json`</u>

YAML data file

- "Independent" and "dependent" variables:

```
independent_variables:
- header: {name: Leading dilepton PT, units: GEV}
  values:
  - {low: 0, high: 60}
  - {low: 60, high: 100}
  - {low: 100, high: 200}
  - {low: 200, high: 600}
dependent_variables:
- header: {name: 10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DPT, units: GEV**-1}
  qualifiers:
  - {name: RE, value: P P --> Z0 < LEPTON+ LEPTON- > Z0 < LEPTON+ LEPTON- > X}
  - {name: SQRT(S), units: GEV, value: 7000}
  values:
  - value: 7000
    errors:
    - {symerror: 1100, label: stat}
    - {symerror: 79, label: 'sys,detector'}
    - {symerror: 15, label: 'sys,background'}
  - value: 9800
    errors:
    - {symerror: 1600, label: stat}
    - {symerror: 75, label: 'sys,detector'}
    - {symerror: 15, label: 'sys,background'}
  - value: 1600
    errors:
    - {symerror: 490, label: stat}
    - {symerror: 41, label: 'sys,detector'}
    - {symerror: 2, label: 'sys,background'}
  - value: 80
    errors:
    - {symerror: 60, label: stat}
    - {symerror: 2, label: 'sys,detector'}
    - {symerror: 0, label: 'sys,background'}
```
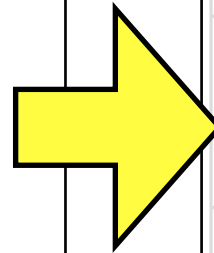
RE	P P --> Z0 < LEPTON+ LEPTON- > Z0 < LEPTON+ LEPTON- > X
SQRT(S)	7000.0 GeV
Leading dilepton PT [GEV]	10**6 * 1/SIG(fiducial) * D(SIG(fiducial))/DPT [GEV**-1]
0.0 - 60.0	7000.0 ±1100.0 stat ±79.0 sys,detector ±15.0 sys,background
60.0 - 100.0	9800.0 ±1600.0 stat ±75.0 sys,detector ±15.0 sys,background
100.0 - 200.0	1600.0 ±490.0 stat ±41.0 sys,detector ±2.0 sys,background
200.0 - 600.0	80.0 ±60.0 stat ±2.0 sys,detector

- Validated against `data_schema.json`

Covariance matrices

**Correlation/covariance matrices**   hepdata-submission.readthedocs.io

Correlation/covariance matrices can be encoded in a format with two independent variables (giving the bins) and one dependent variable (giving the covariance/correlation), e.g.

```
independent_variables:
- header: {name: PTjet, units: GeV}
  values:
  - {low: 25, high: 45}
  - {low: 45, high: 65}
  - {low: 45, high: 65}
  ...
- header: {name: PTjet, units: GeV}
  values:
  - {low: 25, high: 45}
  - {low: 25, high: 45}
  - {low: 45, high: 65}
  ...
dependent_variables:
- header: {name: Correlation}
  values:
  - {value: 1.0000}
  - {value: 0.8727}
  - {value: 1.0000}
  ...
```

ins1605749

CMS

ins1454404

LHCb
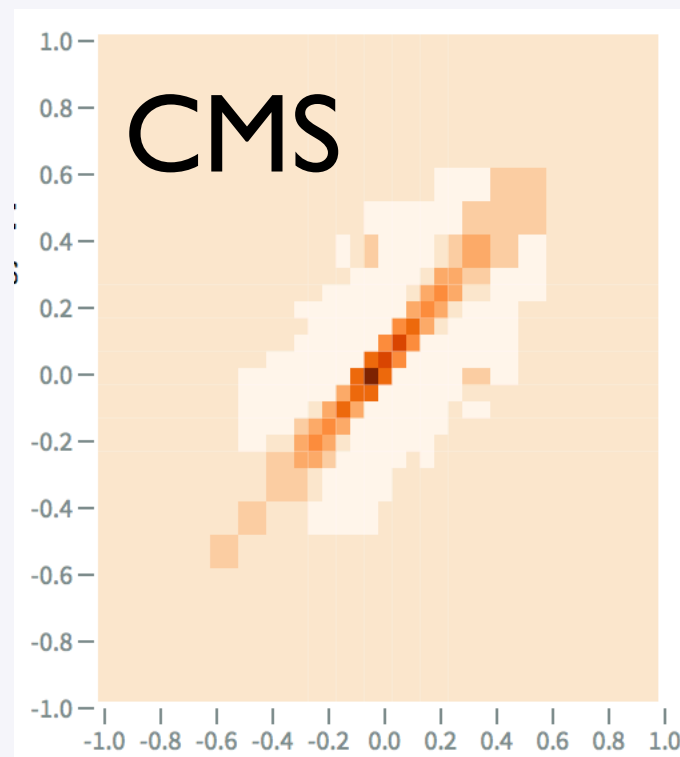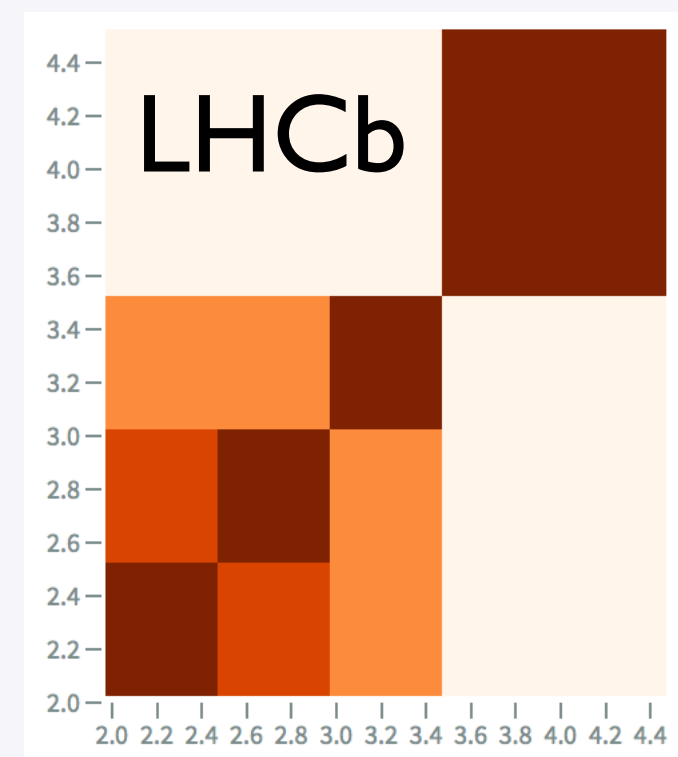
- Export to JSON, CSV, ROOT (TGraph2DErrors, TH2F), YODA (Scatter3D).

Submission documentation

- Documentation at hepdata-submission.readthedocs.io. Includes example Python scripts (simple, complicated).

- HEPData YAML files checked against JSON schema by validation code during submission. `hepdata-validate`

- hepdata_lib package by *Clemens Lange* and *Andreas Albert*. Library to read in text/ROOT and write HEPData YAML. https://github.com/HEPData/hepdata_lib

- Similar Python package by *Christian Holm Christensen*. https://gitlab.com/cholmcc/hepdata

- Also hepdata_maker tool by *Krzysztof Bożek* (ATLAS).

- Experiments often develop internal HEPData tools/docs.

hepdata-cli

- CLI and Python API for HEPData search/download/upload.

- Summer project in 2020 by <u>Giuseppe De Laurentis</u>.

- Install (in venv) with: `pip install hepdata-cli`

- Examples of usage:

```
hepdata-cli find 'collaborations:"Belle-II"' -i inspire
```

```
hepdata-cli fetch-names 1860766 -i inspire
```

```
hepdata-cli download 1860766 -f csv -i inspire
```

```
hepdata-cli upload /path/to/TestHEPSubmission.tar.gz -e
my@email.com -p $PASSWORD —r 123456 -i $INVITATION_COOKIE -s False
```
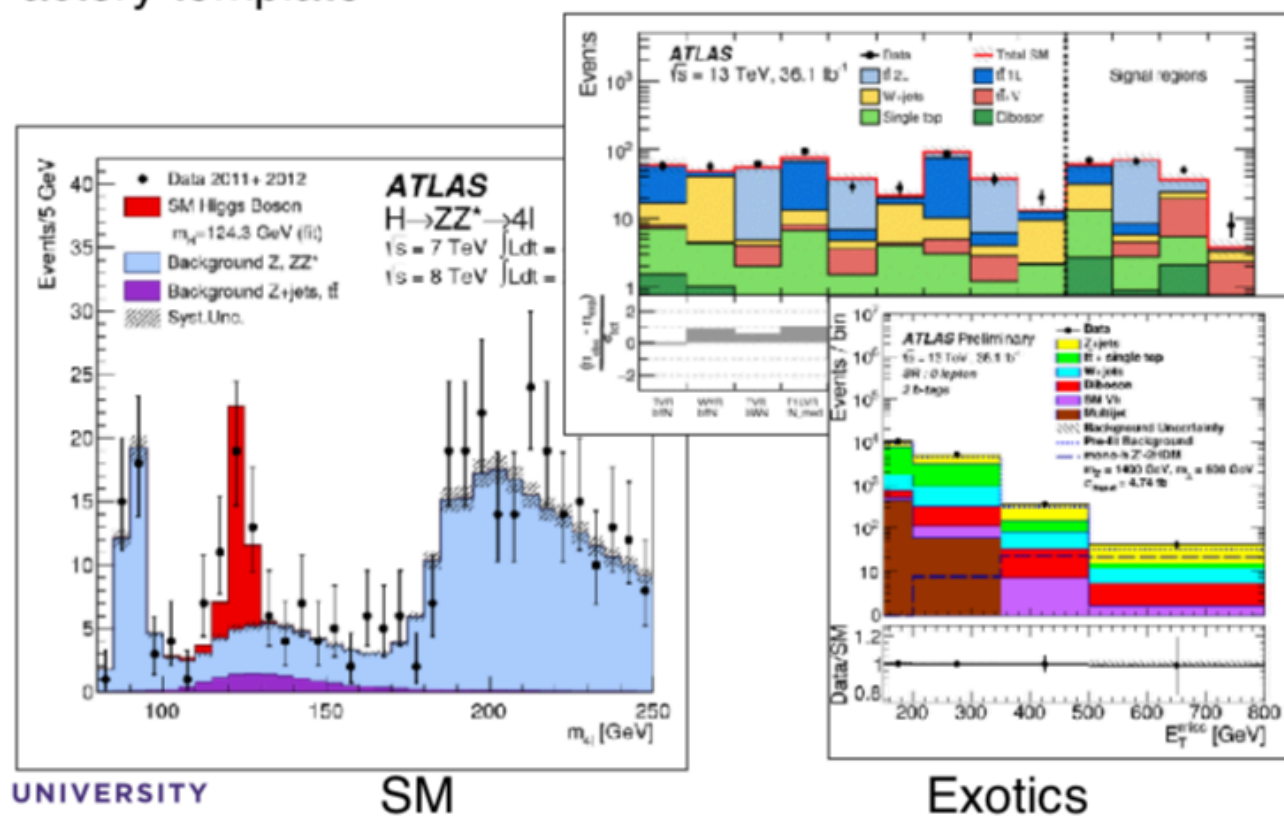
Code: https://github.com/HEPData/hepdata-cli

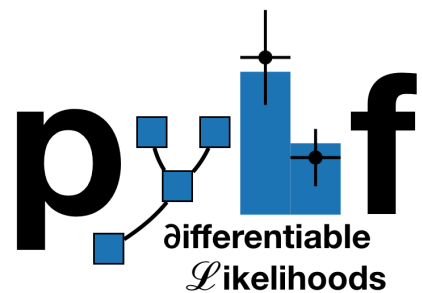# HistFactory

HistFactory

Lukas Heinrich

fundamentally a (quite flexible) p.d.f template to build **statistical models** from binned distributions and data.

~all binned, folded, frequentist statistical models in ATLAS are expressed using HistFactory template

NEW YORK UNIVERSITY

SM   SUSY   Exotics

- Based on simple ROOT histograms organised in an XML file.

- ROOT/XML workspace replaced by plain-text pyhf JSON.

pyhf likelihoods

• Separate DOIs now allocated for attached `pyhf` archive files.

Summary

Email: info@hepdata.net
Forum: hepdata-forum.cern.ch

- **HEPData** is *the* repository for publication-related HEP data.

- *Caveats:* design restricts size (~MB) and format (mostly tabular).

- Responsibility for data submission now held by experiments.

- Belle II Coordinators: `mgt-hepdata-coordinators@belle2.org`

- **NEW** A Belle and a Belle II record completed on 29th August 2022.

- Documentation and software tools available to assist upload.

- `hepdata-cli` search/download/upload from command line.

- **NEW** Enhanced linking to Rivet analyses and pyhf likelihood data.

Code: `https://github.com/HEPData`

Credit: Thanks to Alison Clarke for her software contributions from 11/2019 to 05/2022.

G. Watt