*An array library for **nested**, **variable-sized data**, including arbitrary-length lists, records, mixed types, and missing data, using **NumPy-like idioms**.*

Awkward Array

Embracing Irregular Data Across Disciplines
Community Outreach

Ianna Osborne, Princeton University

*Arrays are **dynamically typed**, but operations on them are **compiled and fast**.*
Coincides with NumPy when arrays are regular; generalizes when they're not.

[NFS '25]
NUMFOCUS PROJECT SUMMIT

NUMFOCUS
OPEN CODE = BETTER SCIENCE

PHY-2323298

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

Example of an “awkward” array

```
array = ak.Array( [

    [{"x": 1.1, "y": [1]}, {"x": 2.2, "y": [1, 2]}, {"x": 3.3, "y": [1, 2, 3]}],

    [ ],

    [{"x": 4.4, "y": [1, 2, 3, 4]}, {"x": 5.5, "y": [1, 2, 3, 4, 5]}]

] )
```

Record structures with differently typed fields

Array of variable-length lists (“ragged” or “jagged” arrays)

Heterogeneous data (union/variant types)

Nested variable-length lists

Missing data

Numpy-like expression

```
output = np.square(array["y", ..., 1:])
```

Result

```
output.to_list()
[
    [[], [4], [4, 9]],
    [],
    [[4, 9, 16], [4, 9, 16, 25]]
]
```

(for a similar calculation 10 million times larger, single threaded)

343 ms ± 10.1 ms
576.0 MB of memory

Equivalent Python

```
output = []
for sublist in python_objects:
    tmp1 = []
    for record in sublist:
        tmp2 = []
        for number in record["y"][1:]:
            tmp2.append(number)
        tmp1.append(tmp2)
    output.append(tmp1)
```

3.03 s ± 532 ms
4–5 GB of memory

[NFS '25]
NUMFOCUS PROJECT SUMMIT

NUMFOCUS
OPEN CODE = BETTER SCIENCE

PHY-2323298

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

Why Awkward Array should be interesting (for you)

- Nested data is everywhere
- Open source thrives on shared insight
- We need each other’s edge cases

[NFS '25]
NUMFOCUS PROJECT SUMMIT

NUMFOCUS
OPEN CODE = BETTER SCIENCE

NSF
PHY-2323298

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

What we bring to the table

- Expressive, Numpy-like syntax for nested arrays
- Interoperability with Pandas, Arrow, and JAX
- Scalable performance for large datasets
- Open governance and community calls

[NFS '25] NUMFOCUS PROJECT SUMMIT

NUMFOCUS OPEN CODE = BETTER SCIENCE

PHY-2323298

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

Awkward Array Library Integrations

Awkward Array is a part of larger scientific Python ecosystem. It works well with other libraries.

Institute for Research & Innovation in Software for High Energy Physics

APACHE ARROW

Parquet

Lossless, bidirectional, zero-copy conversions between Awkward Array, Apache Arrow, Feather, and Parquet.

ROOT Data Analysis Framework

JIT-compiled as a C++ iterator in ROOT's RDataFrame workflow.

CuPy

Awkward Arrays on GPUs, backed by CuPy.

JIT-compile Awkward Arrays in C++ with cppyy. Header-only library to build Awkward Arrays in C++ and transfer them to Python.

Numba

Efficiently iterate over or build Awkward Arrays in code JIT-compiled with Numba, on CPUs and GPUs

pandas

polars

RAPIDS

Akimbo: work with Awkward Arrays as DataFrame columns

[ragged]

Manipulating ragged arrays as though they were **NumPy** or **CuPy** arrays, following the Array API specification.

TensorFlow

Bidirectional conversion between Awkward Array and TensorFlow RaggedTensor.

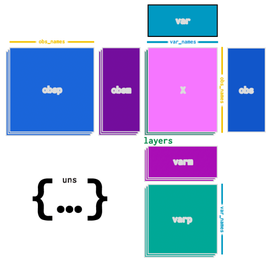
Auto-differentiate through expressions involving Awkward Arrays, using JAX

bluesky

Store Awkward Arrays in the Tiled database and slice them in the cloud.

DASK

dask-awkward is a high-level collection type for Dask, alongside Array and DataFrame.



Integrated into AnnData for single-cell genomics.

Read any Kaitai Struct format into Awkward Arrays.

Awkward Array.jl

Exchange data with the Julia language through a reimplementation of the Awkward Array memory layouts.

NSF

PHY-2323298

Where else might Awkward Array be useful?

- Astronomers? Ecologists? Linguists? Bioinformaticians?
- Open source developers and educators
- Anyone wrangling nested or hierarchical data

[NFS '25]
NUMFOCUS PROJECT SUMMIT

NUMFOCUS
OPEN CODE = BETTER SCIENCE

NSF
PHY-2323298

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

Let’s collaborate

- Join our community calls
- Share your use cases
- Help improve docs and examples
- Propose cross-domain tutorials

[NFS '25]
NUMFOCUS PROJECT SUMMIT

NUMFOCUS
OPEN CODE = BETTER SCIENCE

NSF
PHY-2323298

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

A shared future

- Ad vitam - for life, for legacy, for the communities we build

NSF
PHY-2323298
iris hep
Institute for Research & Innovation in Software for High Energy Physics