**Why it's needed**

Particle physicists have always needed big datasets of nested, variable-sized data.

Figure from a 45-year old physics-software manual:

We'd draw similar figures today!

Traditionally, this problem was solved by making data analysis use **FORTRAN/C++**.

**Awkward Array**

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

Arrays are **dynamically typed**, but operations on them are **compiled and fast**.

Coincides with NumPy when arrays are regular; generalizes when they're not.

**What it does**

Jagged broadcasting of NumPy “ufuncs”

```
[[0, 0], [1, [0, 0]]] +
[[[0, 0], [1, [0, 0]]]]
```

Advanced indexing

```
>>> events[1, 'muons', [0, 1]]
>>> events[events['muons', 'pt'] > 50]
```

Combinatorics

```
[[0, 0], [1, [0, 0]]] ⊗
[[[0, 0], [0, 0]], [[0, 0], [0, 0]]]
```

Reshaping for plotting and machine learning

```
[[0, 0], [1, [0, 0]]] →
[[[0, 0], [0, 0]], [[0, 0], [0, 0]]]
```

Jagged reducers

```
[1, 2, 4], [[8, 16]] → 7, 0, 24
```

**ROOT & Arrow/Parquet I/O**

Originally intended as an array type for ROOT files, Awkward Arrays are convertible to/from Apache Arrow and Parquet (sometimes zero-copy).

**Who uses it?**

Mostly physicists, but a few geneticists and data scientists have expressed interest.

**How it works**

Arrays and their operations are **columnar**.

Consider these lists of particle objects:

```
[[Muon(1, 0.412, 0.801), Muon(2.76, -0.124, 0.504), Muon(1.18, -0.119, 0.921)],
 [Muon(27, 1.246, 0.911)],
 [Muon(72, -0.267, 0.951)],
 [Muon(59, -0.754, 0.241), Muon(71, 0.181, 0.620)]]
```

We represent them in columnar arrays, contiguously by field:

**Interface with Numba**

Awkward Arrays can be arguments and return values in Numba’s JIT-compiled functions, enabling for-loop logic at the speed of compiled code.

...with Pandas

Awkward Arrays can be columns of a Pandas DataFrame.

...NumExpr, Autograd, ...

Jagged broadcasting is applied to all elementwise array calculations.

**Why now?**

Python/NumPy is rapidly becoming a standard language for data analysis in particle physics.

**But...**

NumPy does not work on nested, variable-sized data!