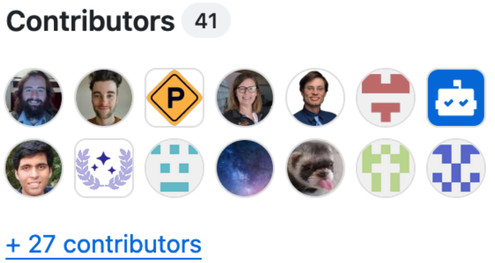
Columnar Data Analysis

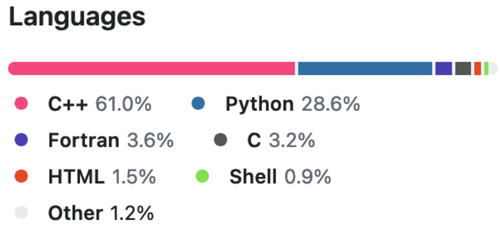
at CoDaS HEP 2024

Ianna Osborne, July 24 2024

Who I am

- Ianna Osborne (“ee-AN-uh”)
  - Research Software Engineer
  - Princeton and IRIS-HEP based at CERN
    - #3 contributor to AwkwardArray
- member of CMS experiment
  - #12 contributor to CMSSW (core software, geometry description, event display, simulation, etc.)
- background in Physics and Computer Science
- C++, Python, Julia
- Contact
  - linkedin.com/in/ianna-osborne-9982a342
  - ianna.osborne@cern.ch





PRINCETON UNIVERSITY

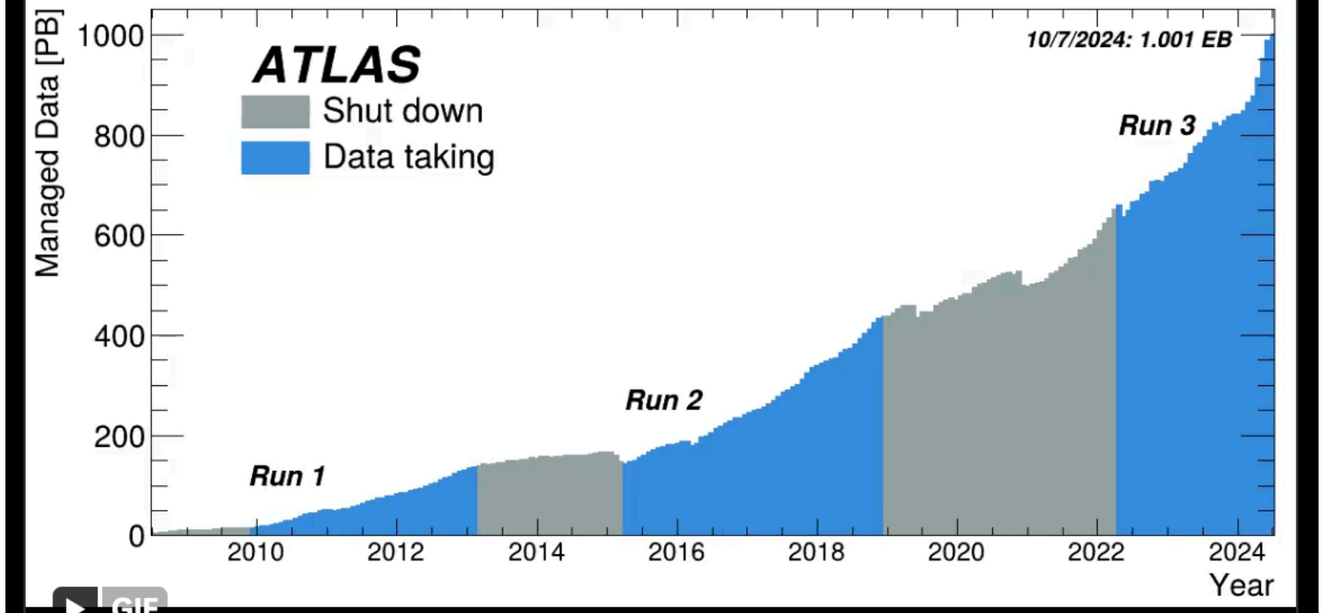
Motivation and Outline

Motivation

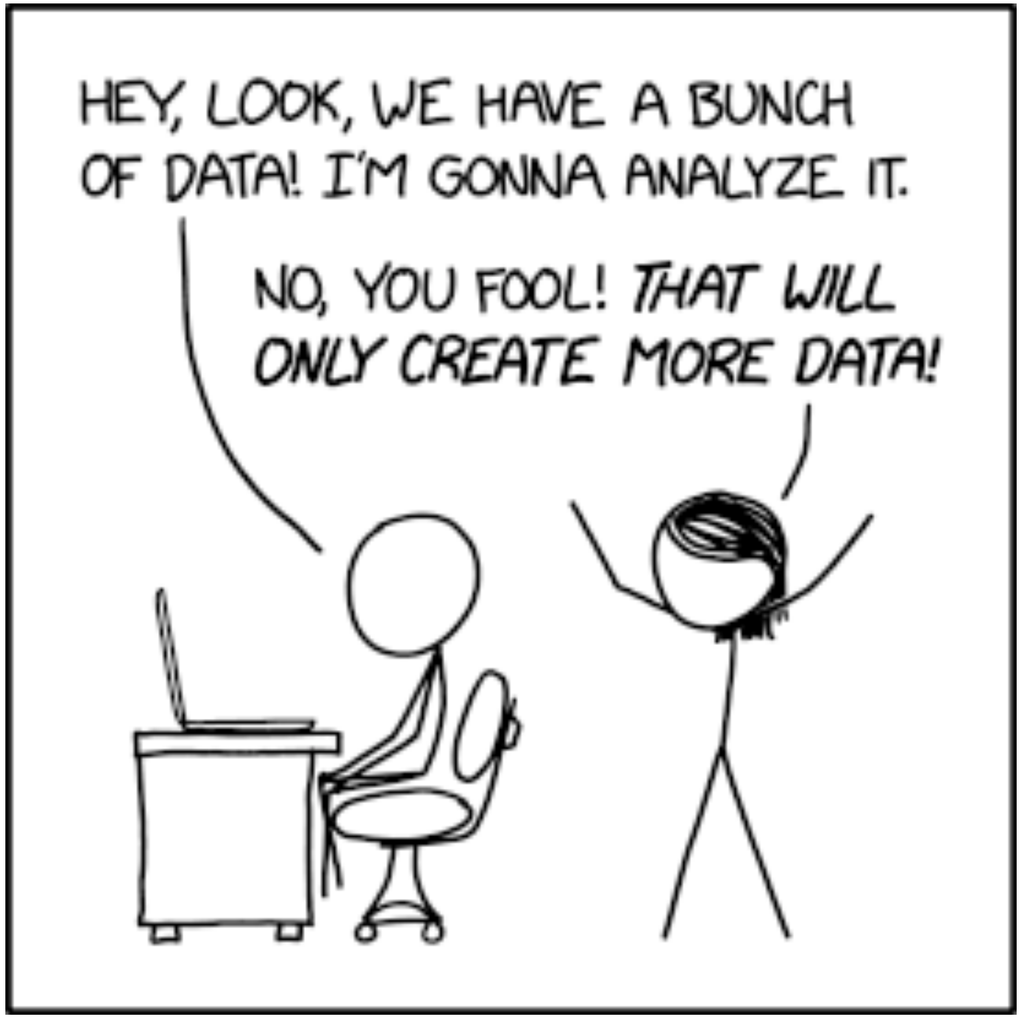
ATLAS Experiment @ATLASexperiment · Jul 12

New milestone for the ATLAS experiment at @CERN! We've hit over 1 exabyte (that's 1 billion billion bytes) of real and simulated data worldwide, managed using @RucioData.

That's a whole lotta data! #datascience



Motivation


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! THAT WILL ONLY CREATE MORE DATA!

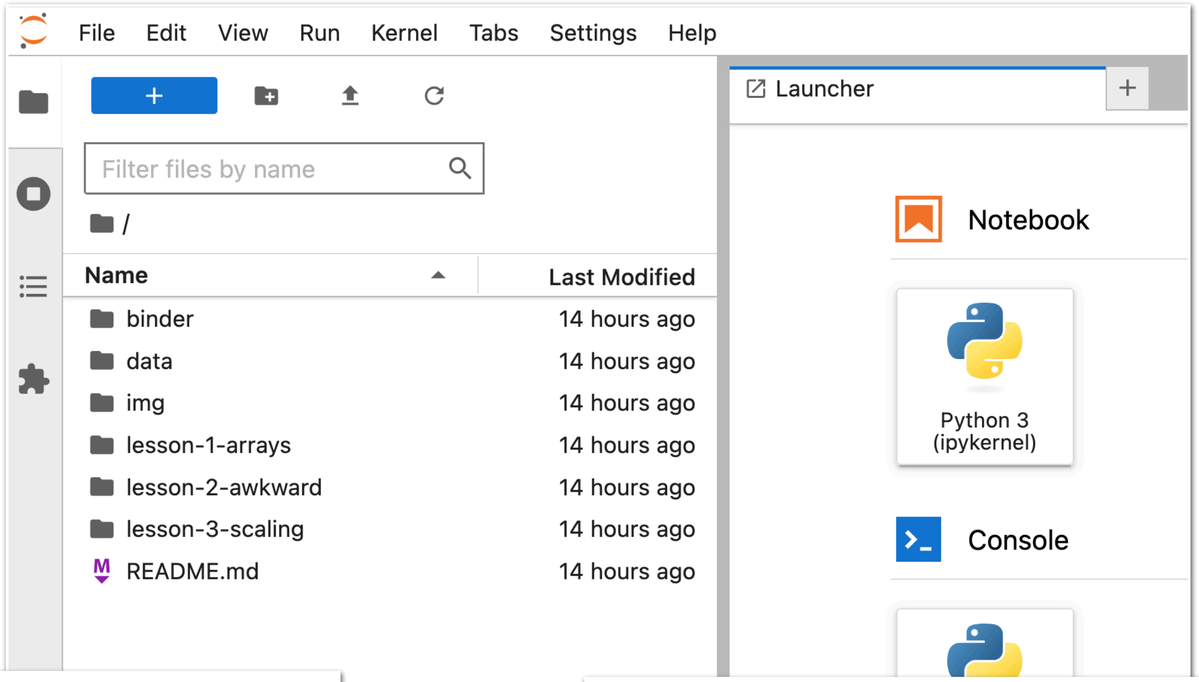
Outline

- Lesson 1: Array-oriented Programming
  - Programming paradigms
  - What is array-oriented programming good for?
  - NumPy as an example
- Lesson 2: Ragged and nested arrays
- Lesson 3: Vertical and horizontal scaling
  - about making Python faster
  - introducing Dask

How to participate

- No need to install anything on your computer

launch binder



/ lesson-1-arrays /

Name	Last Modified
lecture-slides.ipynb	14 hours ago
lecture-workbook.ipynb	14 hours ago
project-1-life.ipynb	14 hours ago
project-2-higgs.ipynb	14 hours ago
solutions-1-life.ipynb	14 hours ago
solutions-2-higgs.ipynb	14 hours ago

/ lesson-2-awkward /

Name	Last Modified
lecture-slides.ipynb	14 hours ago
lecture-workbook.ipynb	14 hours ago
project-1-taxi.ipynb	14 hours ago
project-2-higgs.ipynb	14 hours ago
solutions-1-taxi.ipynb	14 hours ago
solutions-2-higgs.ipynb	14 hours ago

/ lesson-3-scaling /

Name	Last Modified
lecture-slides.ipynb	14 hours ago
lecture-workbook.ipynb	14 hours ago
project-1-mandelbrot.ipynb	14 hours ago
project-2-higgs.ipynb	14 hours ago
solutions-1-mandelbrot.ipynb	14 hours ago
solutions-2-higgs.ipynb	14 hours ago

Takeaways

Why Array-oriented Programming?

- **Conciseness**: You can express complex operations on large datasets with just a few lines of code.
- **Performance**: Operations on entire arrays can be optimized by the underlying libraries to run faster than equivalent loops.
- **Simplicity**: It simplifies code by eliminating the need for explicit loops and conditionals.