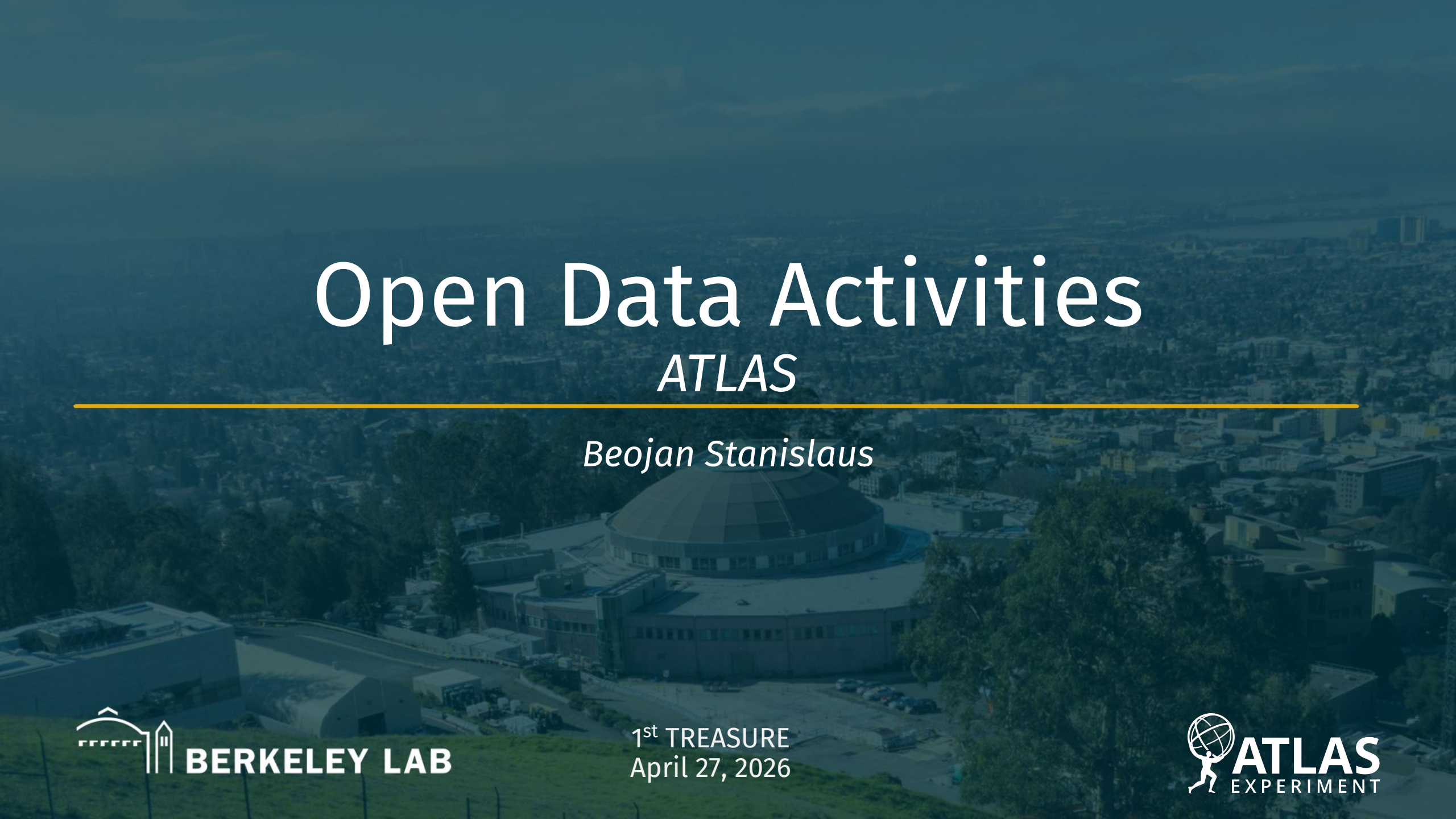
Open Data Activities

ATLAS

Beojan Stanislaus



BERKELEY LAB

1st TREASURE
April 27, 2026

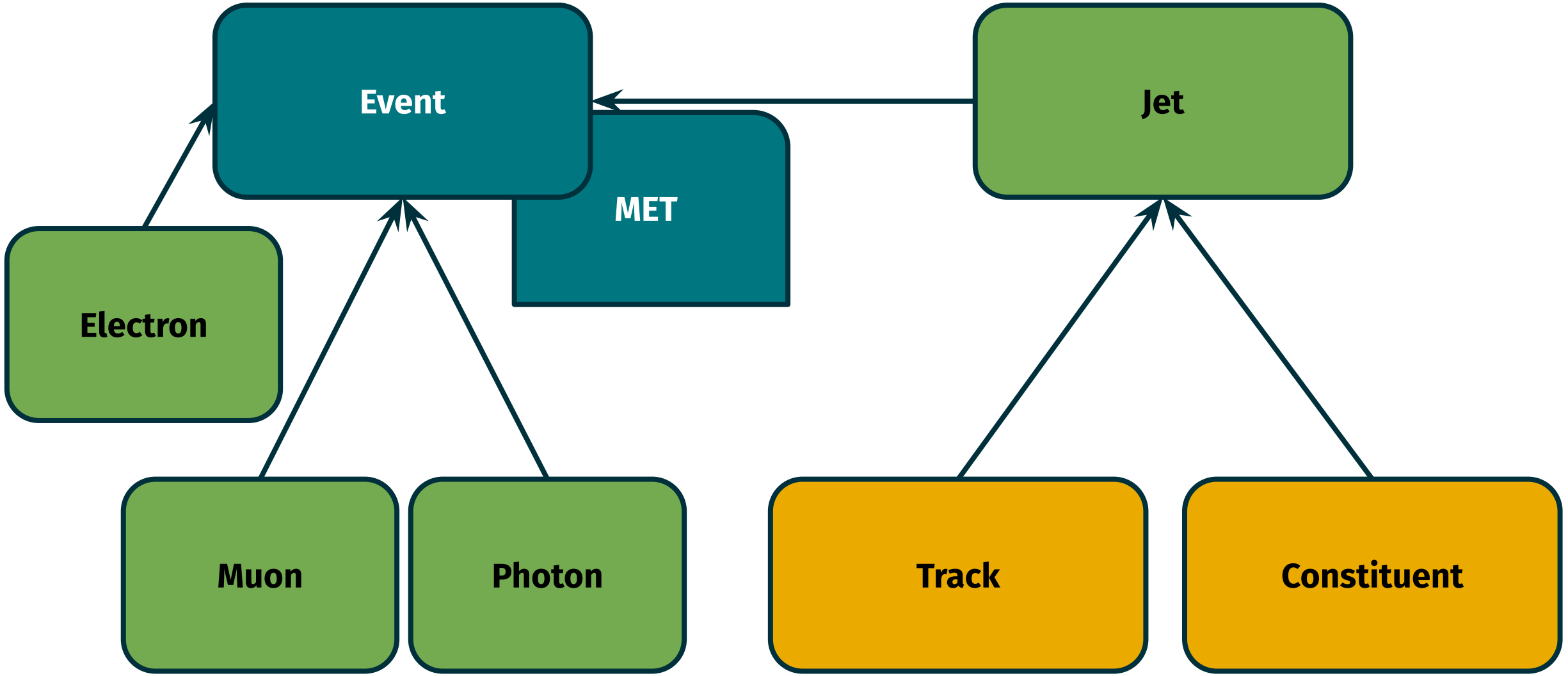
ATLAS EXPERIMENT

Open Data in TREASURE

One of our goals is to standardize collider open data into a common format

1. Common to various colliders with DOE participation
   - At least ATLAS and CMS
2. Comprehensive enough to enable a variety of AI use-cases
3. Input to our tokenization activities

Particle Cloud Data Format



Implementation for ATLAS

- Implemented a new “DAOD” format for TREASURE by cobbling together PHYSLITE and parts of FTAG1LITE
- Uproot / Awkward n-tuple maker on top of this makes PCDF in Parquet format
- Revealed deficiencies in PCDF format

Made request for 10 million events of already released MC in this format.

Event

EventIndex **[Primary Key]**

Detector Index [Link to some other DB?]

Simulation tag

Run number

Event number

~~Primary Vertex x, y, z~~ Beam Position x, y, z

(Auxiliary Info ObjectID)

MET

MET_phi

Change to attach MET to event

Objects (Electron, Muon, Photon)

index	**[Primary Key]**
eventIndex	[Link to Event]
pt	
eta	
phi	
charge (electron and muon only)	
truth	

Jet (Large and Small *R*)

index	**[Primary Key]**
eventIndex	[Link to Event]
pt	
eta	
phi	
m	
truth / flavor (large-R / small-R)	

Constituents (Particle Flow)

index [Primary Key]

jetIndex

eventIndex

pt

eta

phi

m

- *For now, R = 0.4 only*
- *Read both charged and neutral constituents from PHYSLITE and follow jet → constituent element links*
- *NB: ElementLinks are not pleasant to use from Uproot*

Tracks

index **[Primary Key]**

eventIndex

jetIndex

q_p

theta

phi

d0

z0

Can later calculate a pt, eta, phi representation

Discussion

File Format

- Use of indices as cross-references allows tables to be flat
- Enables storage as Parquet files with metadata such as column names and statistics
- Right software (e.g. Polars) can group by the indices to generate *n*-dimensional HDF5 tables

Storage Location

- American Science Cloud resource?
  - What is available?
- HPDF
- Fermi Data Platform
- CERN Open Data Portal
  - Possibility is there for ATLAS and CMS

Other Options

- Hugging Face
- Cloud, e.g. S3

Transformation and Compute

- Plan is to store one format and transform on the fly to what users want
  - e.g. User-preferred HDF5 schema
- Puts constraint on storage location

Low Level Detector Information

- Would need to use same eventIndex
- Everything else doesn't need to be the same format
  - Ideally, just needs to fit into on-the-fly transformation framework