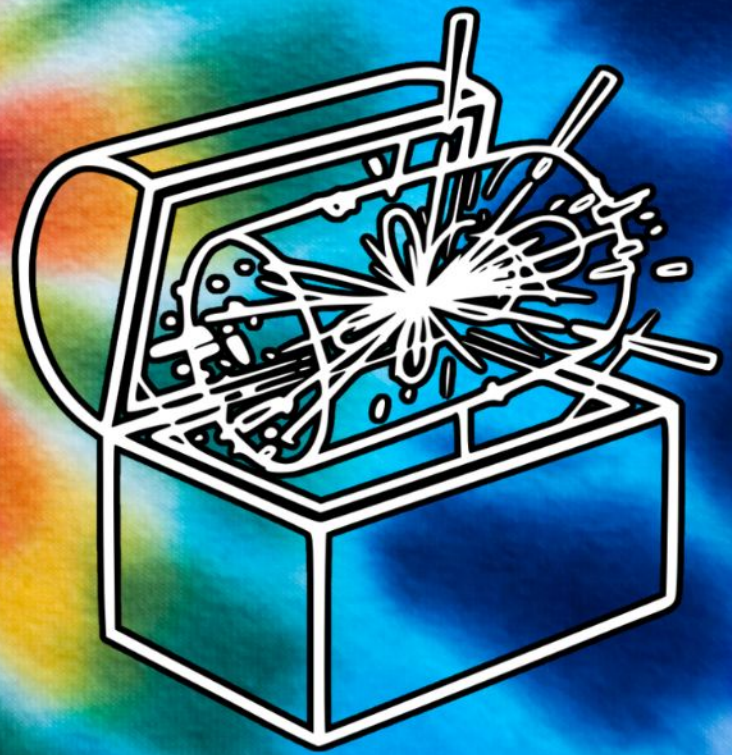
Welcome to the 1st Annual

TREASURE WORKSHOP



Viviana Cavaliere
Elizabeth Brost

April 27, 2026

Welcome to BNL

- Cool facts about BNL:
  - https://www.bnl.gov/newsroom/news.php?a=24923
  - In 1958, Brookhaven scientists built one of the first video games ever created, Tennis for Two, to entertain visitors at the Lab's annual visitors' day.
- The Omega Group got its name in the early 1970s to honor the discovery of the Ω^- - the baryon made of three valence strange quarks, first seen in BNL's 80-inch bubble chamber in 1964

TREASURE:

TOKENIZED TRACKS

Agenda

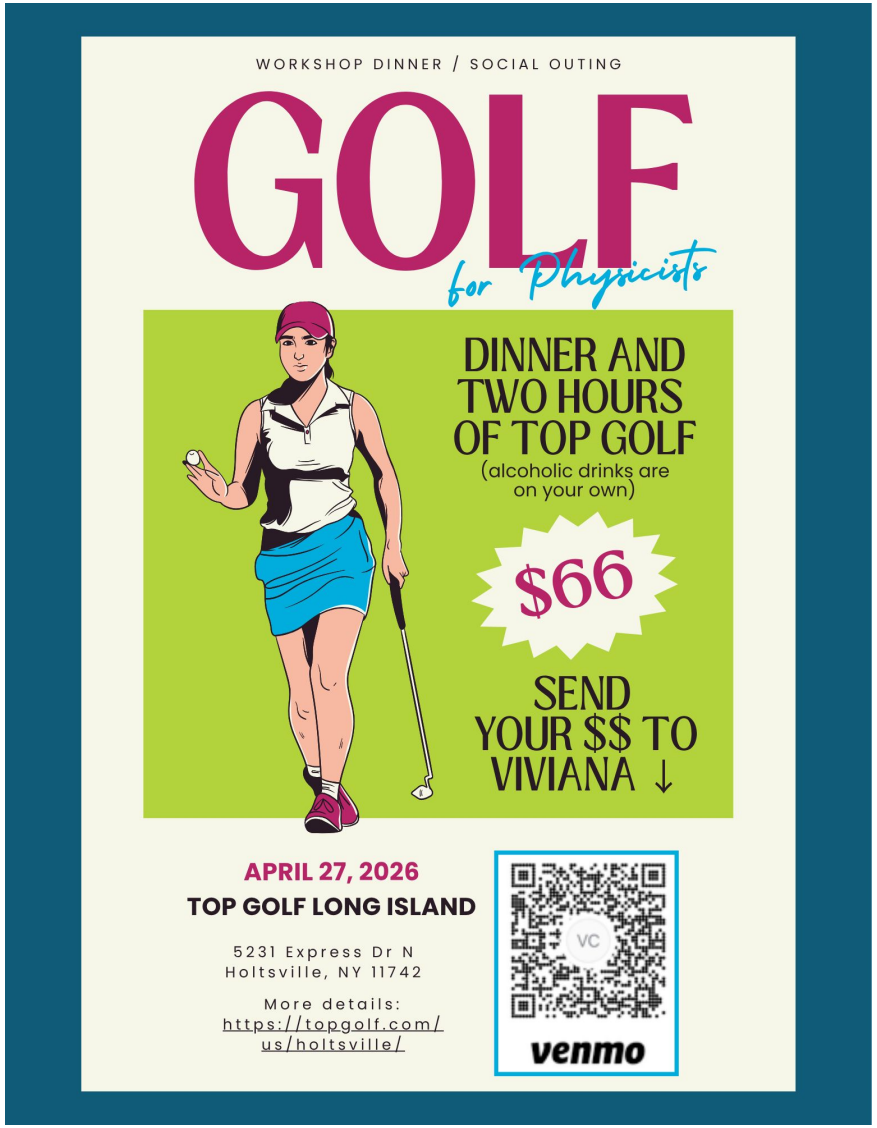
Monday	Tuesday	Wednesday
Breakfast 🍩	Breakfast 🍩	Breakfast 🍩
AmSC Context & Integration	**Topical Presentations**	**Topical Presentations**
☕☕☕☕☕☕☕ Coffee ☕☕☕☕☕☕☕	☕☕☕☕☕☕☕ Coffee ☕☕☕☕☕☕☕	☕☕☕☕☕☕☕ Coffee ☕☕☕☕☕☕☕
AmSC Context & Integration	**Foundation Models / AI in Trigger**	**Working time**
Group photo 📸 & Lunch	Lunch	Lunch
Open Data	**Working Time**	**Report from each work team & plan next steps**
☕☕☕☕☕☕☕ Coffee ☕☕☕☕☕☕☕	☕☕☕☕☕☕☕ Coffee ☕☕☕☕☕☕☕	✈️ travel home ✈️
Planning for Future Activities **Dinner + TopGolf!**	**BNL Physics Colloquium "An Open Future for ATLAS, HEP, and Global Software and Computing"**	

Brookhaven National Laboratory

Food and Dinner

- Food is reserved for registered participants so no colourful lanyard, no food

- Group outing / Dinner on Monday evening:
  - 6:30pm
  - TopGolf in Holtsville (link)
- If you need a ride to dinner or can offer a ride to dinner meet at 5:45 in the lounge (just in front of this room)



Brookhaven National Laboratory

WiFi

- Wifi is available through eduroam
- If you can’t use eduroam, it is possible to connect to the guest wifi through Corus

link

BNL Event Code of Conduct

- **All who participate in BSA-managed events, on site at Brookhaven Lab or off, must conduct themselves in an ethical manner, demonstrating respect through common courtesy, civility, and effective communication.**
  - BSA will not tolerate discrimination or harassment of any kind, including sexual harassment, bullying, intimidation, violence, threats of violence, retaliation, or other disruptive behavior.
  - Anyone asked to stop behavior deemed inappropriate must comply immediately.
  - If someone is in immediate risk of serious harm on site, dial 631.344.2222 from a mobile phone. From anywhere else, dial 911.
- **If inappropriate behavior is observed, notify the event coordinator as soon as possible.** Concerns may also be reported through EthicsPoint, which is hosted by an independent third party.
  - Consequences may include ejection from the event without a refund, if applicable. BSA reserves the right to notify appropriate authorities and organizations such as employers, institutions, or programs.

→ if you have any questions or concerns, see us or Omega group leader Michael Begel

Brookhaven National Laboratory

Workshop Venue

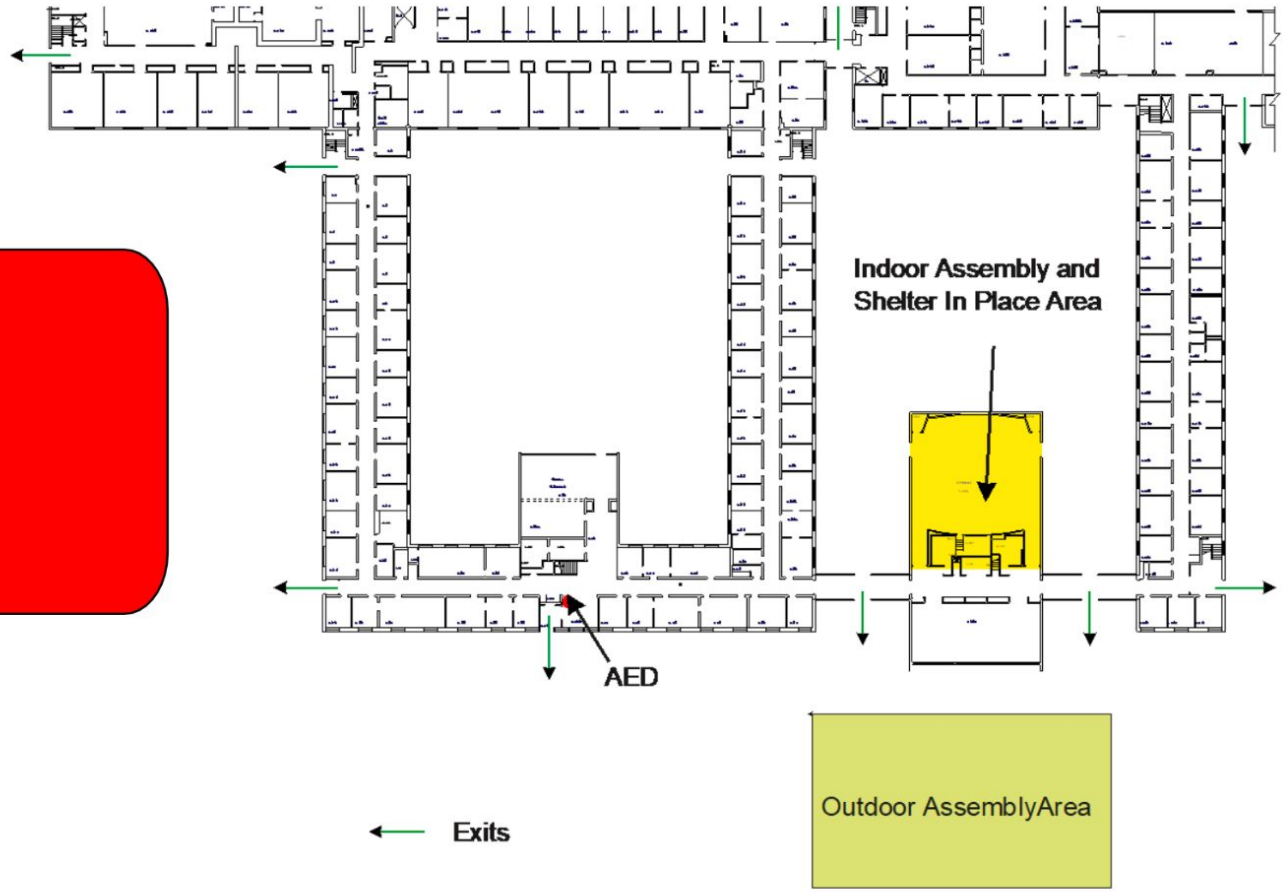
Front entrance (510 - Physics Building)

Brookhaven National Laboratory

In case of emergency

To Report a fire, spill, medical or other emergency,
DIAL EXT. 2222 or 911

If using a cell phone,
DIAL 631-344-2222
If a telephone is not available,
USE A FIRE ALARM BOX



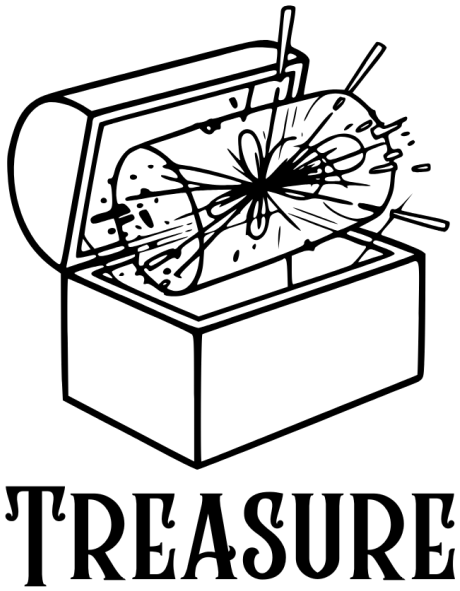
Brookhaven National Laboratory

Thanks to our admin team

Linda, Cia and especially Ivette! Thanks for everything!

What is TREASURE?

- *Tokenized Representations for Energy-frontier AI Searches via Understanding and REasoning*
  - DOE HEP American Science Cloud (AmSC) Intelligent Data Activities Pilot



Contacts:

Viviana Cavaliere (BNL, PI)

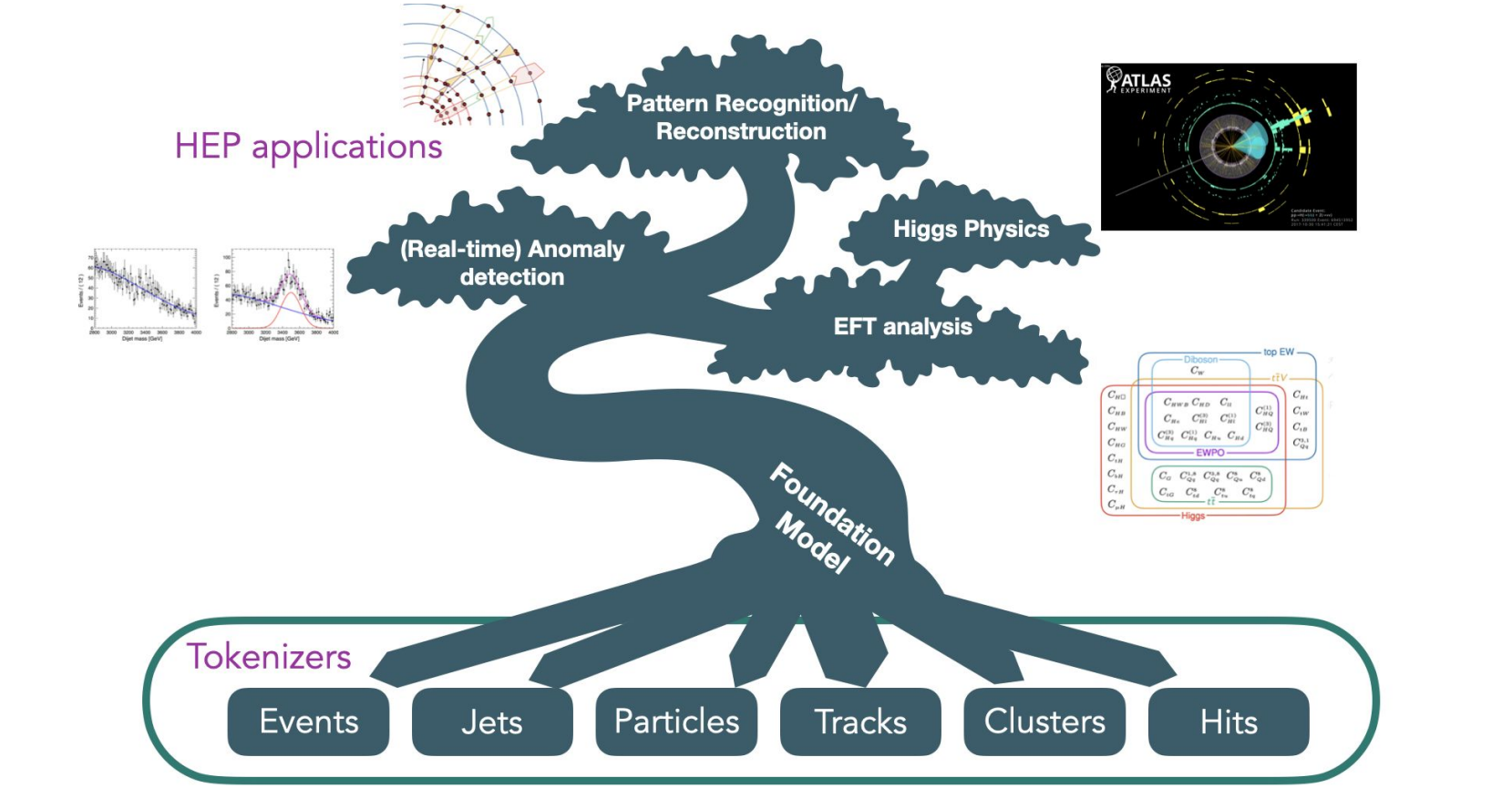
Paolo Calafiura (LBNL)

Walter Hopkins (ANL)

Michael Kagan (SLAC)

Kevin Pedro (FNAL)

What is Treasure?



- *Treasure prepares **collider data** so it can be effectively used by modern AI methods*
- It transforms heterogeneous, experiment-specific data into standardized, **tokenized**, AI-ready representations.
- **Builds Foundation Models** to learn across experiments, detectors, and eras of data.

Goal: unlock new discovery potential from both current and legacy high-energy physics datasets.

In Practice: the TREASURE Pipeline

Public

Public Framework to run different tokenization

Public

Experiment Data (ROOT)
Object-level Data (e.g., ATLAS PHYSLITE)

AI-READY (HDF5, PARQUET)
per-event arrays, masked, padded

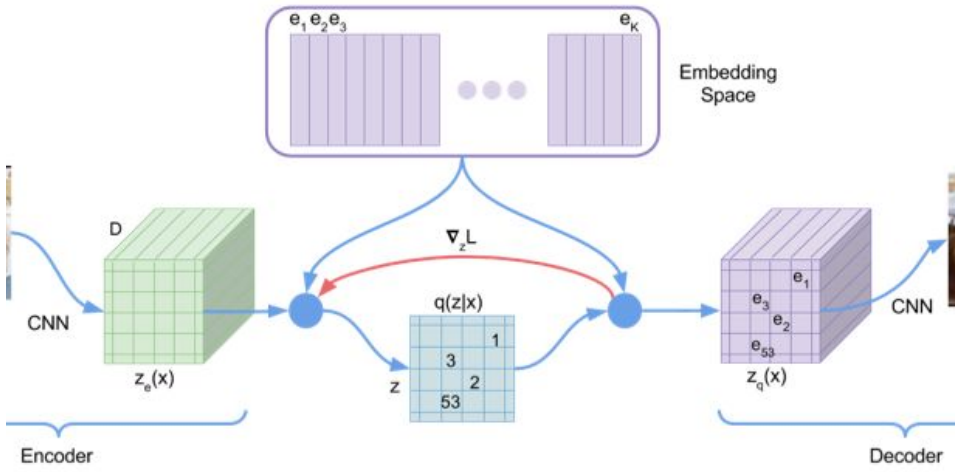
TOKENIZER
VQ-VAE / FSQ / PQ / RVQ per type

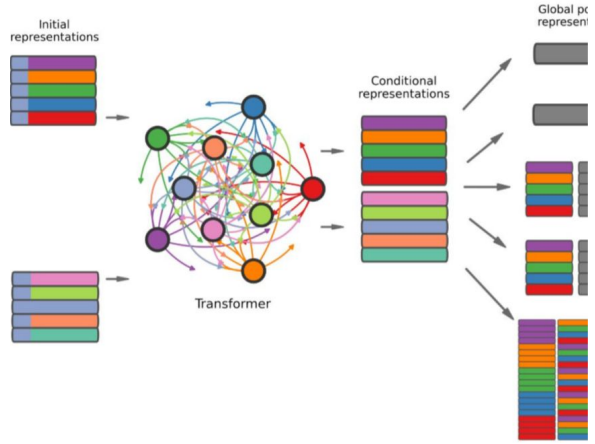
PARQUET
Integer token sequences

TRANSFORMER
Pretrain → fine-tune → transfer

https://gitlab.cern.ch/vcavalie/bnl-treasure

```
ATLAS DAOD_PHYSLITE to HDF5 Converter  (cross-experiment edition)
================================================================

Produces two layers in every HDF5 file:

  /common/    — variables defined identically across ATLAS and CMS.
                A CMS NanoAOD converter should fill the same keys.
                All energies/momenta in GeV, angles in radians.

  /atlas/     — ATLAS-specific variables kept for ATLAS-only studies.
                Do NOT use these in cross-experiment training without
                explicit experiment conditioning.

  /metadata   — HDF5 group attributes (no datasets).  Covers:
                  experiment, format_version, input_file, n_events,
                  git_hash (if available), pt_cuts, max_objects,
                  common_iso_cone, common_iso_pt_floor_gev,
                  common_iso_z0sintheta_cut_mm

Schema summary
--------------
common/electrons : pt, eta, phi, charge, trk_iso03, mask, n
common/muons     : pt, eta, phi, charge, trk_iso03, mask, n
common/taus      : pt, eta, phi, charge, is_1prong, mask, n
common/photons   : pt, eta, phi, trk_iso03, mask, n
common/jets      : pt, eta, phi, mass, n_trk, mask, n
common/tracks    : pt, eta, phi, d0, z0, mask, n
common/met       : pt, phi, sumet          (scalar per event)
common/event     : pvx, pvy, pvz, mu, experiment_id (0=ATLAS, 1=CMS),
                   is_simulation (1=MC, 0=data)

  mask  — boolean array shape (n_events, max_objects).  True = real object,
          False = zero-padding.  Required because events have variable object
```



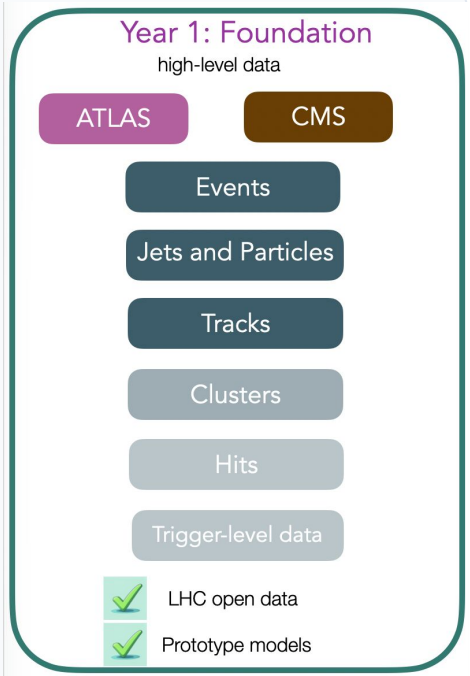


Project Timeline

Year 1

Foundations

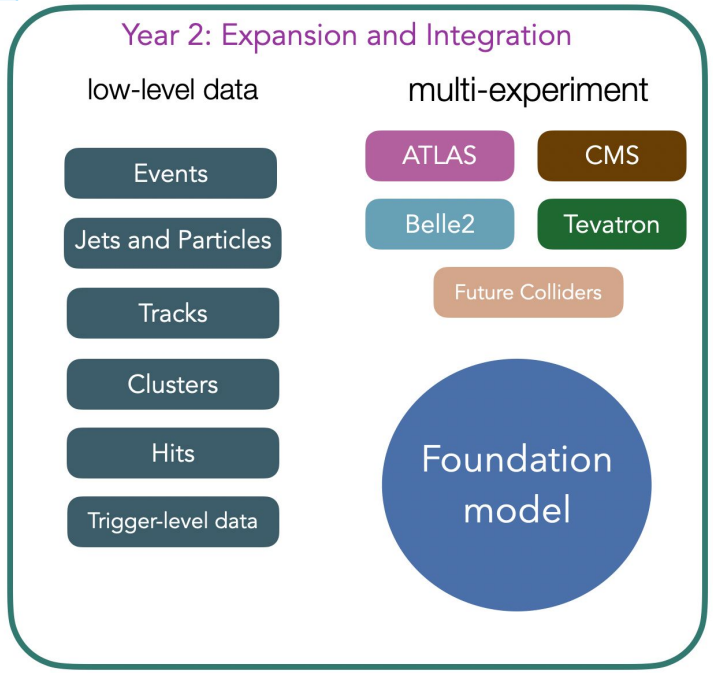
- AI-ready datasets from LHC Open Data
- Common data models, metadata, and tokenization schemes
- Assess AI-readiness and quantify tokenization impact
- **Demonstrate cross-experiment learning with prototype AI models**



Year 2

Expansion and Integration

- **Detector data:** hits, clusters, streaming compression
- **Beyond the LHC:** Tevatron (ppbar), Belle II (e+e-), future experiments
- **Scale up cross-experiment foundation model training**



Brookhaven National Laboratory · Fermilab · Berkeley Lab · SLAC National Accelerator Laboratory · Argonne National Laboratory

Goals of the workshop

- Have everyone seated in the same room
- Hear all the activities from the group
- Agree on a format at least for ATLAS and CMS
- Define a plan to request new open data from ATLAS and CMS with additional information
- Work on including event-level tokenization in the github library heptokens


Backup