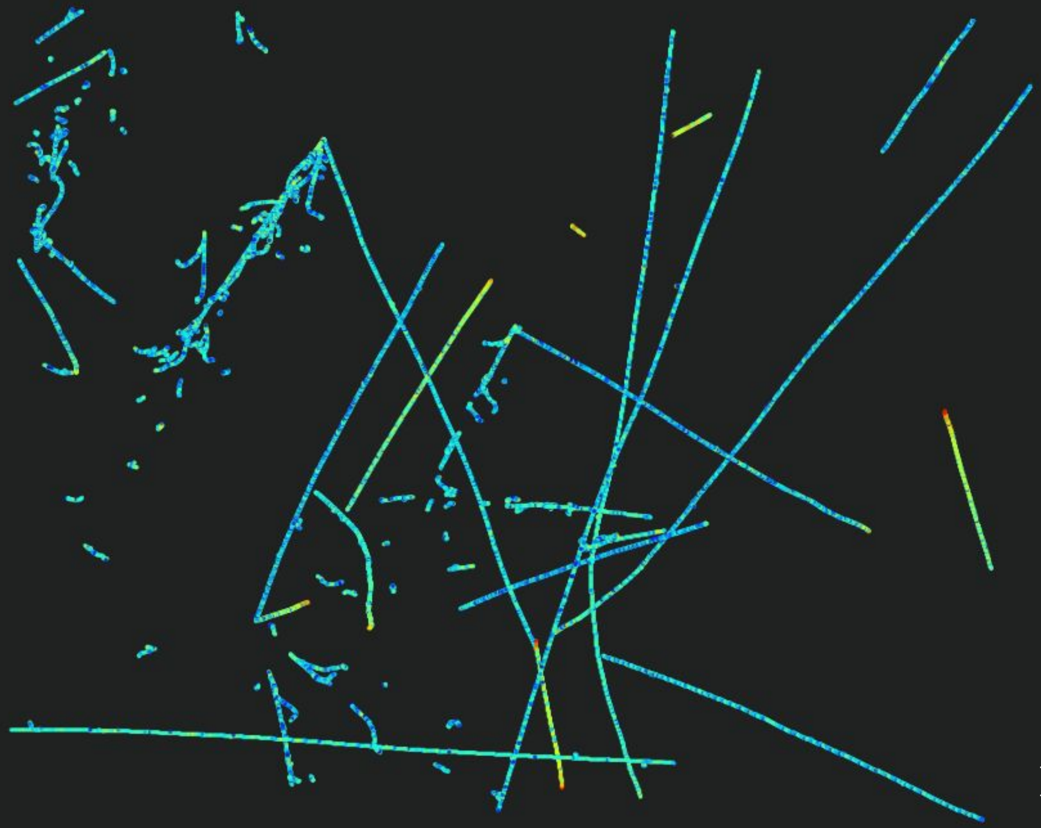
Open data, R&D, and Olympics



Goals to address

What models are there?

How do they differ/work?

Strengths? Weakness?

Which to use for my data?

Is this one the best?

Demonstrated on what data?

Can I reproduce your study?

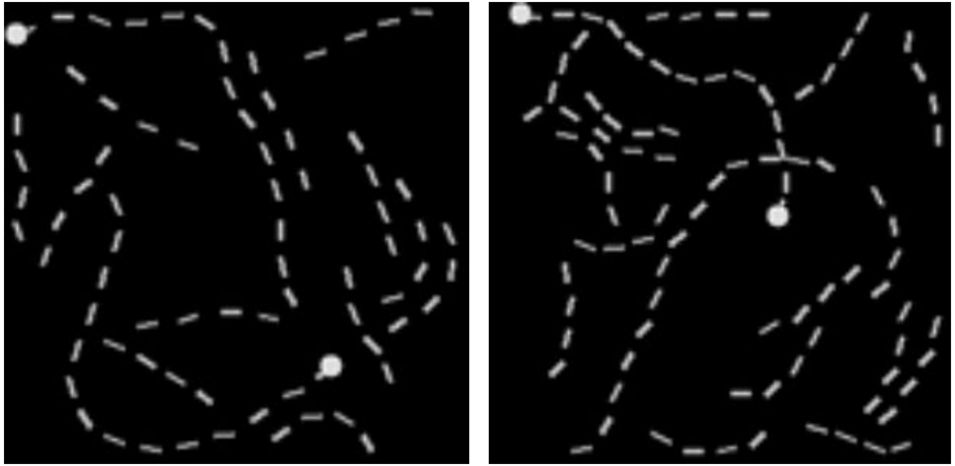
Kazuhiro Terao @ SLAC/Stanford

June 25th 2025 @ NPML @ ETH

Public AI/ML Dataset

- Key research challenges (guidance)
- Enable open, reproducible R&D
- Common data+metric = fair comparison
- Builds a **community** and **standards**

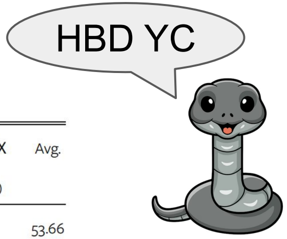
Public AI/ML Dataset

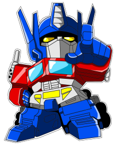


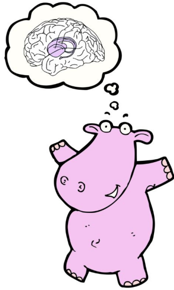
E.g. Long-range Arena Dataset

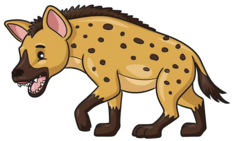
Designed to challenge Transformer's critical bottleneck = computational scalability associated with poor performance for a long range sequence.

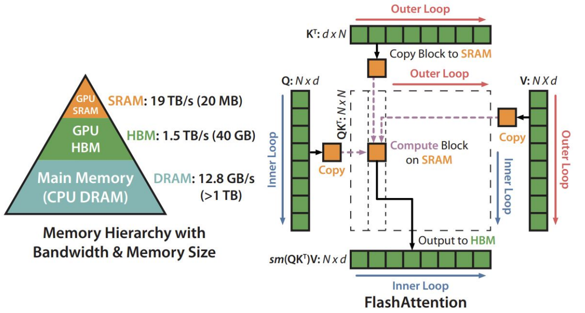
⇒ even simple dataset can revolutionize!

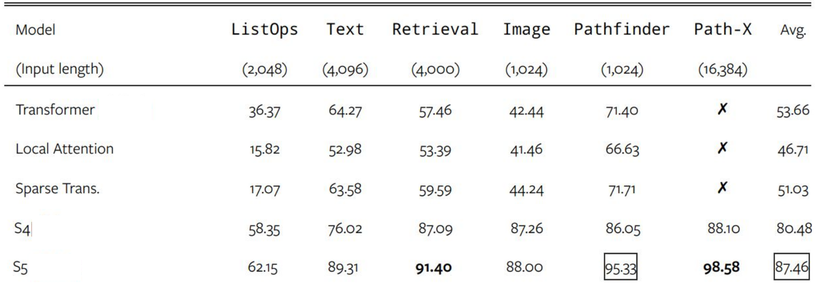










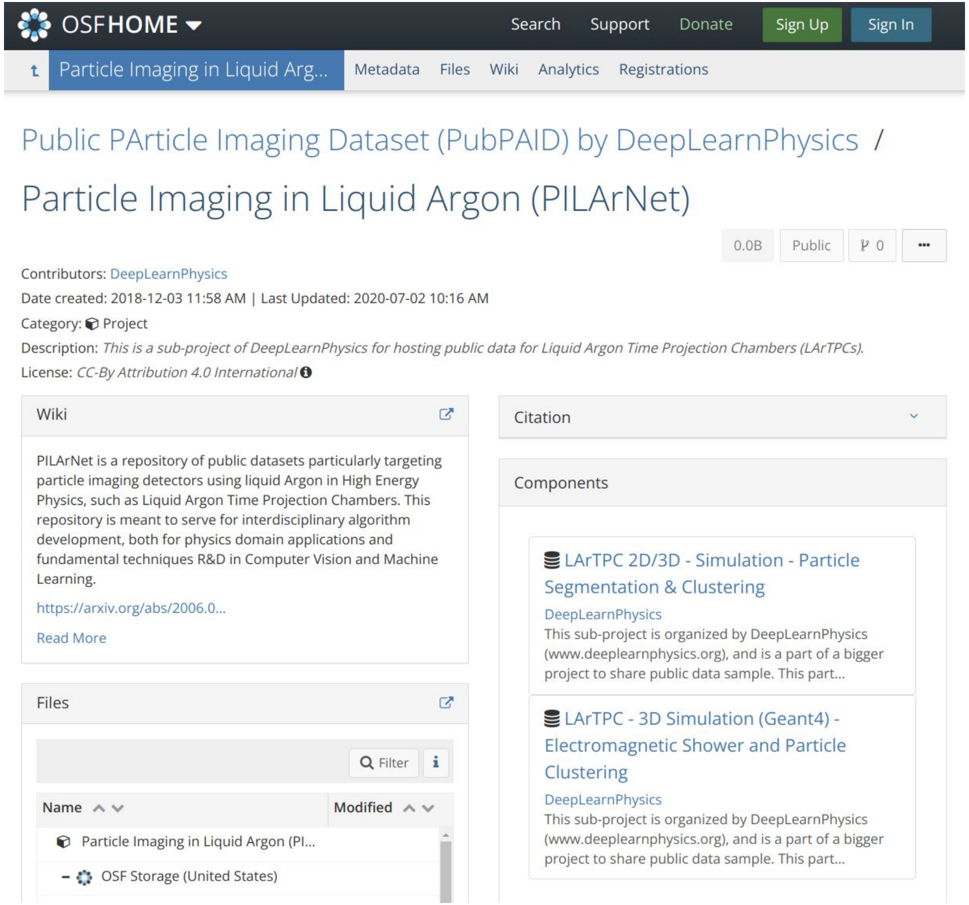


Model (Input length)	ListOps (2,048)	Text (4,096)	Retrieval (4,000)	Image (1,024)	Pathfinder (1,024)	Path-X (16,384)	Avg.
Transformer	36.37	64.27	57.46	42.44	71.40	✗	53.66
Local Attention	15.82	52.98	53.39	41.46	66.63	✗	46.71
Sparse Trans.	17.07	63.58	59.59	44.24	71.71	✗	51.03
S4	58.35	76.02	87.09	87.26	86.05	88.10	80.48
S5	62.15	89.31	**91.40**	88.00	95.33	**98.58**	87.46

Public Scientific AI/ML Dataset

Science domains have unique challenges

- Data space
  - ○ Generally sparse, locally dense images
  - ○ Extremely long sequences
  - ○ Rich and specific metadata
  - ○ Multiple data modalities



Public PArticle Imaging Dataset (PubPAID) by DeepLearnPhysics /

Particle Imaging in Liquid Argon (PILArNet)

Contributors: DeepLearnPhysics

Date created: 2018-12-03 11:58 AM | Last Updated: 2020-07-02 10:16 AM

Category: Project

Description: *This is a sub-project of DeepLearnPhysics for hosting public data for Liquid Argon Time Projection Chambers (LArTPCs).*

License: *CC-By Attribution 4.0 International*

Wiki

PILArNet is a repository of public datasets particularly targeting particle imaging detectors using liquid Argon in High Energy Physics, such as Liquid Argon Time Projection Chambers. This repository is meant to serve for interdisciplinary algorithm development, both for physics domain applications and fundamental techniques R&D in Computer Vision and Machine Learning.

https://arxiv.org/abs/2006.0...

Read More

Files

Name	Modified
Particle Imaging in Liquid Argon (PI...	
– OSF Storage (United States)	

Citation

Components

LArTPC 2D/3D - Simulation - Particle Segmentation & Clustering

DeepLearnPhysics

This sub-project is organized by DeepLearnPhysics (www.deeplearnphysics.org), and is a part of a bigger project to share public data sample. This part...

LArTPC - 3D Simulation (Geant4) - Electromagnetic Shower and Particle Clustering

DeepLearnPhysics

This sub-project is organized by DeepLearnPhysics (www.deeplearnphysics.org), and is a part of a bigger project to share public data sample. This part...

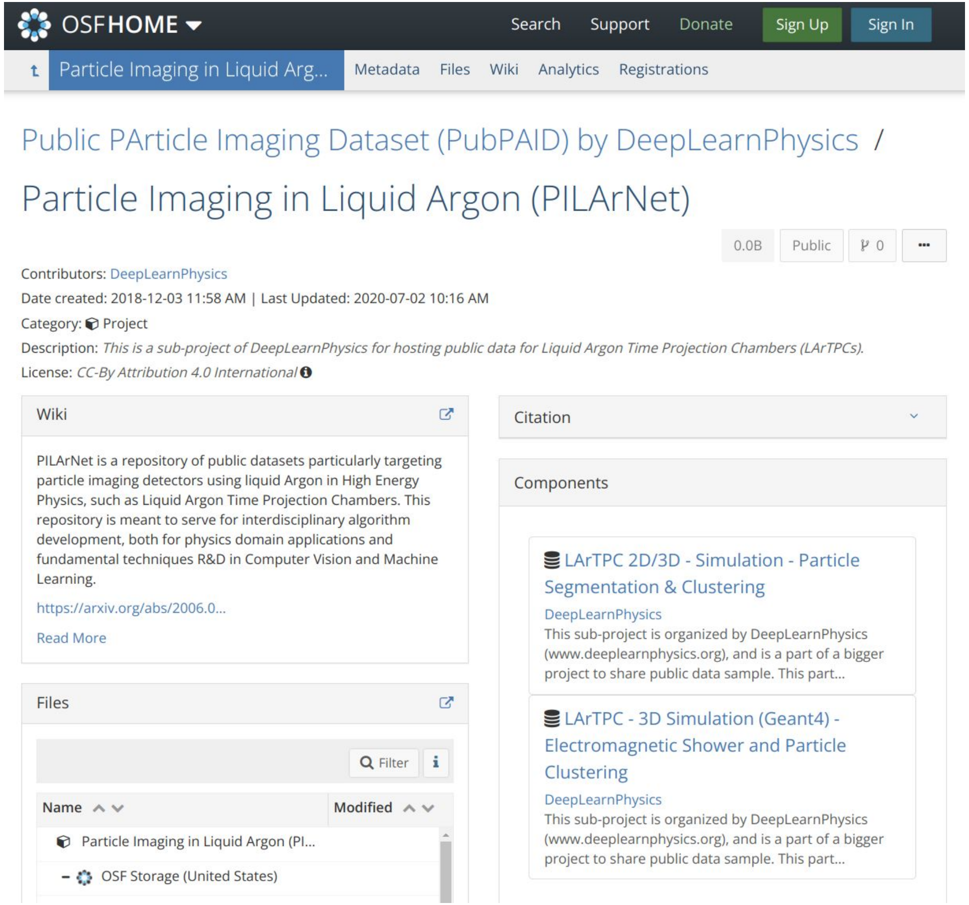
Public Scientific AI/ML Dataset

Science domains have unique challenges

- Data space
    - Generally sparse, locally dense images
    - Extremely long sequences
    - Rich and specific metadata
    - Multiple data modalities

- Science domain
    - Invariance / conservation laws
    - Causation and correlations
    - Anomaly detection
    - Uncertainty/precision requirements

Effort to consolidate key research challenges, organize datasets, and build NPML research community w/ standards.



Public PArticle Imaging Dataset (PubPAID) by DeepLearnPhysics /

Particle Imaging in Liquid Argon (PILArNet)

Contributors: DeepLearnPhysics

Date created: 2018-12-03 11:58 AM | Last Updated: 2020-07-02 10:16 AM

Category: Project

Description: *This is a sub-project of DeepLearnPhysics for hosting public data for Liquid Argon Time Projection Chambers (LArTPCs).*

License: *CC-By Attribution 4.0 International*

Wiki

PILArNet is a repository of public datasets particularly targeting particle imaging detectors using liquid Argon in High Energy Physics, such as Liquid Argon Time Projection Chambers. This repository is meant to serve for interdisciplinary algorithm development, both for physics domain applications and fundamental techniques R&D in Computer Vision and Machine Learning.

https://arxiv.org/abs/2006.0...

Read More

Files

Particle Imaging in Liquid Argon (PI...

OSF Storage (United States)

Citation

Components

LArTPC 2D/3D - Simulation - Particle Segmentation & Clustering

DeepLearnPhysics

This sub-project is organized by DeepLearnPhysics (www.deeplearnphysics.org), and is a part of a bigger project to share public data sample. This part...

LArTPC - 3D Simulation (Geant4) - Electromagnetic Shower and Particle Clustering

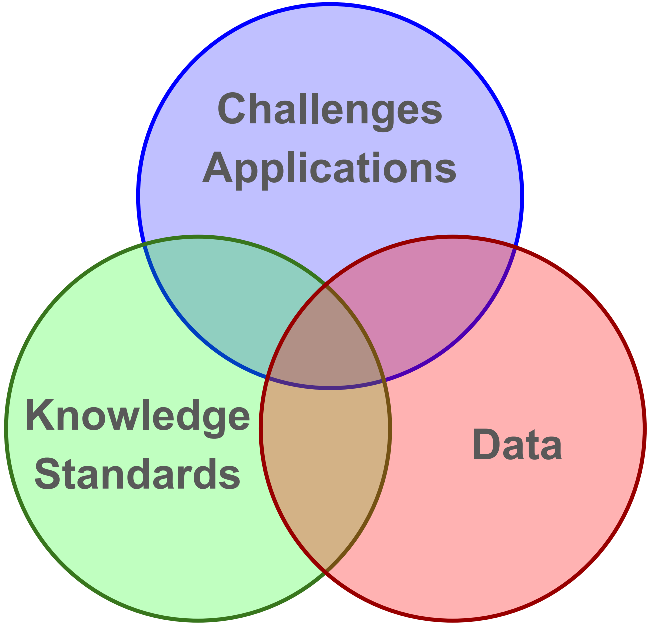
DeepLearnPhysics

This sub-project is organized by DeepLearnPhysics (www.deeplearnphysics.org), and is a part of a bigger project to share public data sample. This part...

Public Scientific AI/ML Data Portal

Data portal consists of 3 cores

- Data!
  - Garbage-in-garbage-out remains very true
  - Quality, big data + metadata
  - Many data modalities (e.g. enable CLIP)
- Scientific challenges / application categories
  - classification/regression, denoising, tomography, etc …
  - object reconstruction, particle flow, SBI, etc. …
- Knowledge base and standards
  - Suitable model architecture and optimization methods
    - Depend on data, application, computational resources
    - Every solution choice should have principle justification
      - Not because "everyone else uses so we tried"



Building Neutrino Open Data Portal

- **Data curation**
  - Identify a contributor (e.g. experiment collab.)
  - Consolidate AI/ML + science research challenges
  - Consolidate data format, tools, and documentation
  - Develop the baseline AI/ML model
  - Curate, upload, and publish the dataset

Will support with researchers with dedicated time.

- **Technical resources**
  - large storage space with public data access
  - Website and connection to scientific compute

≅ 1PB storage with public access at SLAC + looking for other resources (e.g. NERSC)

- **Organization**
  - Organize events, interface w/ requests
  - Advisory committee
    - Categorization for AI/ML, data, challenges
    - Prioritization of challenges

Building Neutrino Open Data Portal

- **Data curation**
  - Identify a contributor (e.g. experiment collab.)
  - Consolidate AI/ML + science research challenges
  - Consolidate data format, tools, and documentation
  - Develop the baseline AI/ML model
  - Curate, upload, and publish the dataset

Will support with researchers with dedicated time.

- **Technical resources**
  - large storage space with public data access
  - Website and connection to scientific compute

≅ 1PB storage with public access at SLAC + looking for other resources (e.g. NERSC)

- **Organization**
  - Organize events, interface w/ requests
  - Advisory committee
    - Categorization for AI/ML, data, challenges
    - Prioritization of challenges

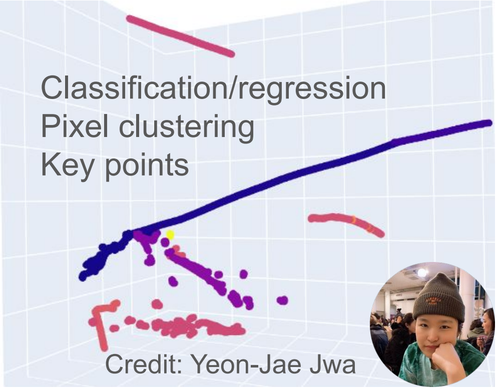
Looking for volunteers with interest. Please come and join.

Last Slide

Neutrino data portal

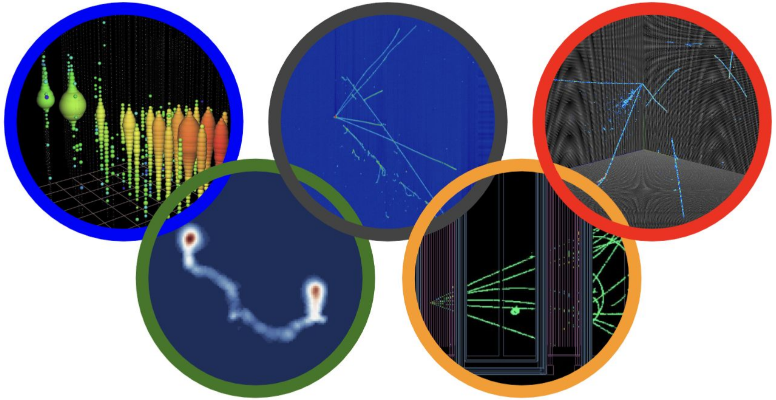
- Same dataset for fair comparison of approaches
  ⇒ build common knowledge base and standards
  ⇒ reproducible research + reusable tools
- Identify common and high priority research challenges
- Develop open collaboration space with AI/ML challenges unique in science

Initial data set: 1M images, 1024^3 image,10k-100k signal voxels/img



Classification/regression
Pixel clustering
Key points

Credit: Yeon-Jae Jwa



Events!

- **NPML Olympic**:
  - 1-2 weeks of hackathon to develop AI/ML techniques for a research category
- **Neutrino AI/ML school**
  - Go over datasets and developed AI models to learn principles and real world applications