Pre-training strategy using real particle collision data for event classification in collider physics

Tomoe Kishimoto

Computing Research Center, KEK

tomoe.kishimoto@kek.jp

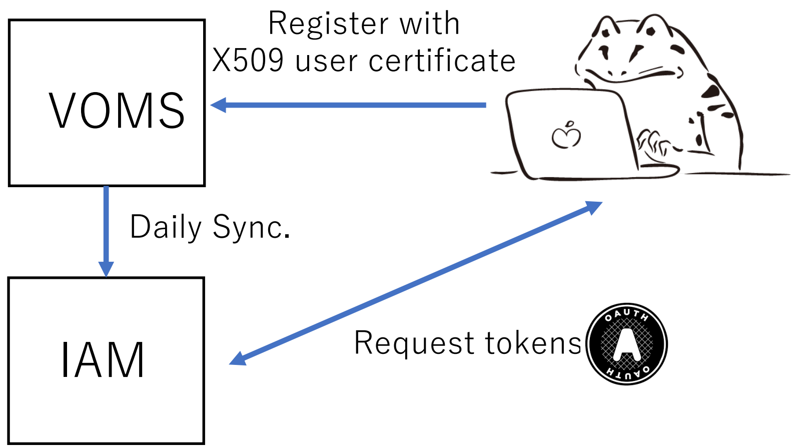
Ref: arXiv:2312.06909

加速器だから見える世界。

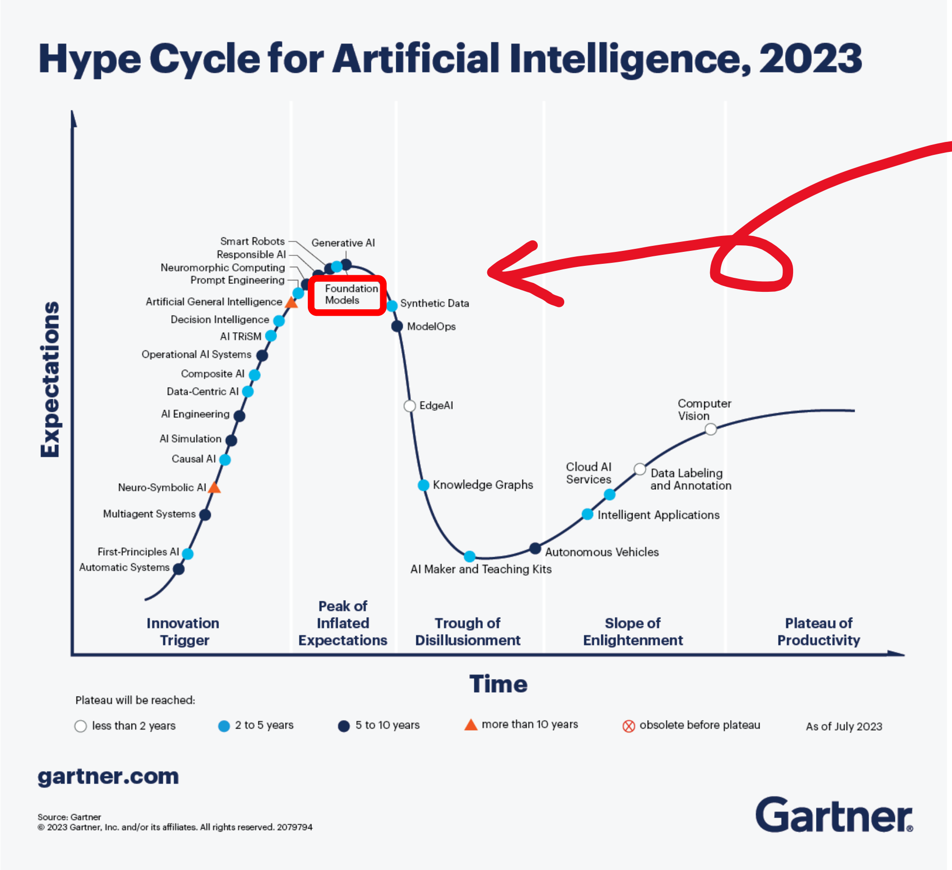
KEK

My recent activities

- Migration from VOMS to IAM
  - IAM instances have been deployed for BelleII group
  - Integration of Japanese ID federation (GakuNin)
- Deep learning for batch job scheduler
  - Paper was accepted for JSSPP 2023 conference, and published
  - https://link.springer.com/chapter/10.1007/978-3-031-43943-8_7
- Deep learning for physics analysis in collider physics



Introduction



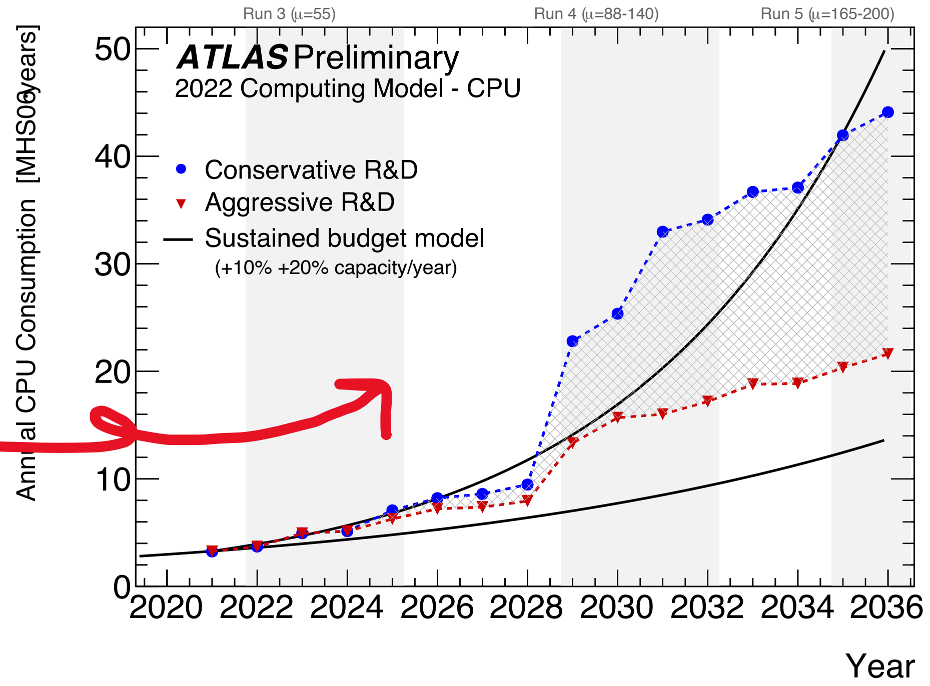
[Gartner.com](https://www.gartner.com)

- "Foundation models" was one of the keywords for AI technology in 2023
  - Pre-training using a large amount of unlabeled data
  - Fine-tuning for a target application (transfer learning)

→ Q: Is the concept of foundation models beneficial to collider physics

Sustainability

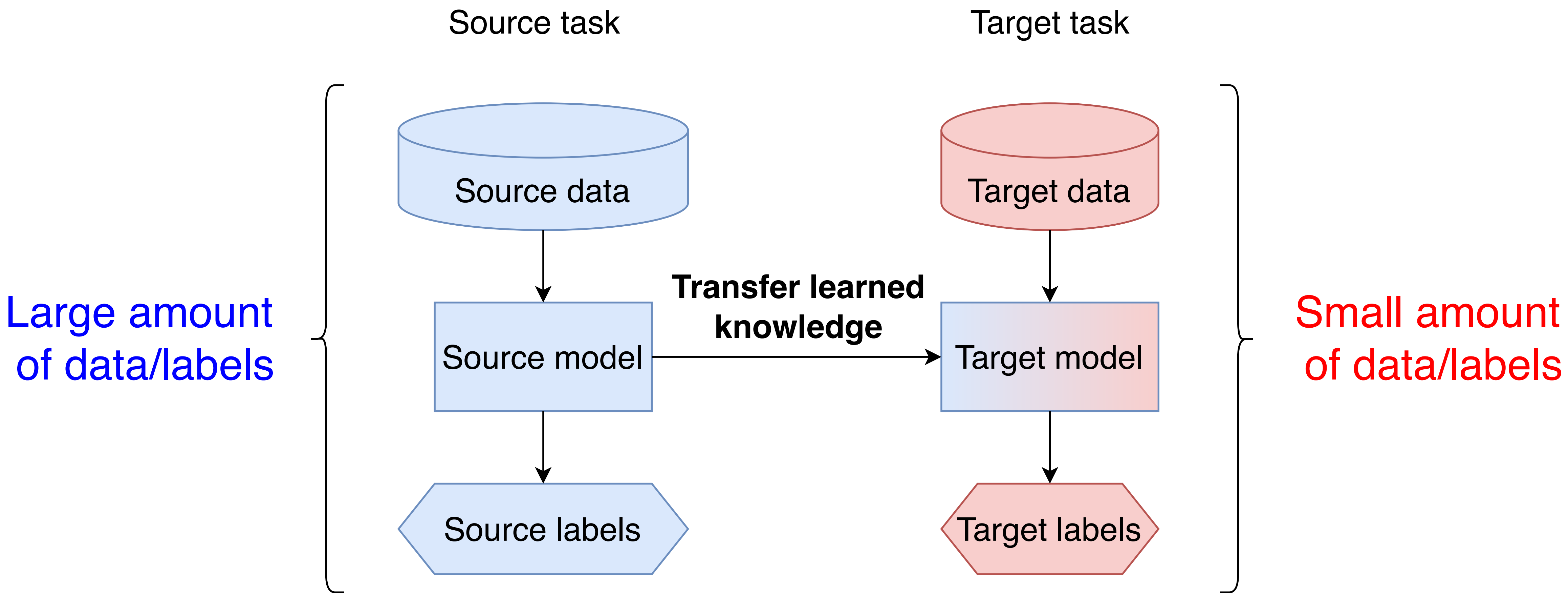
- Deep Learning (DL) requires a large amount of training data
    - In HEP, training data are typically generated by Monte Carlo (MC) simulations ← Computationally expensive
- Electric power consumption, Green computing



→ Maximizing DL performance with a limited amount of data is a key concept

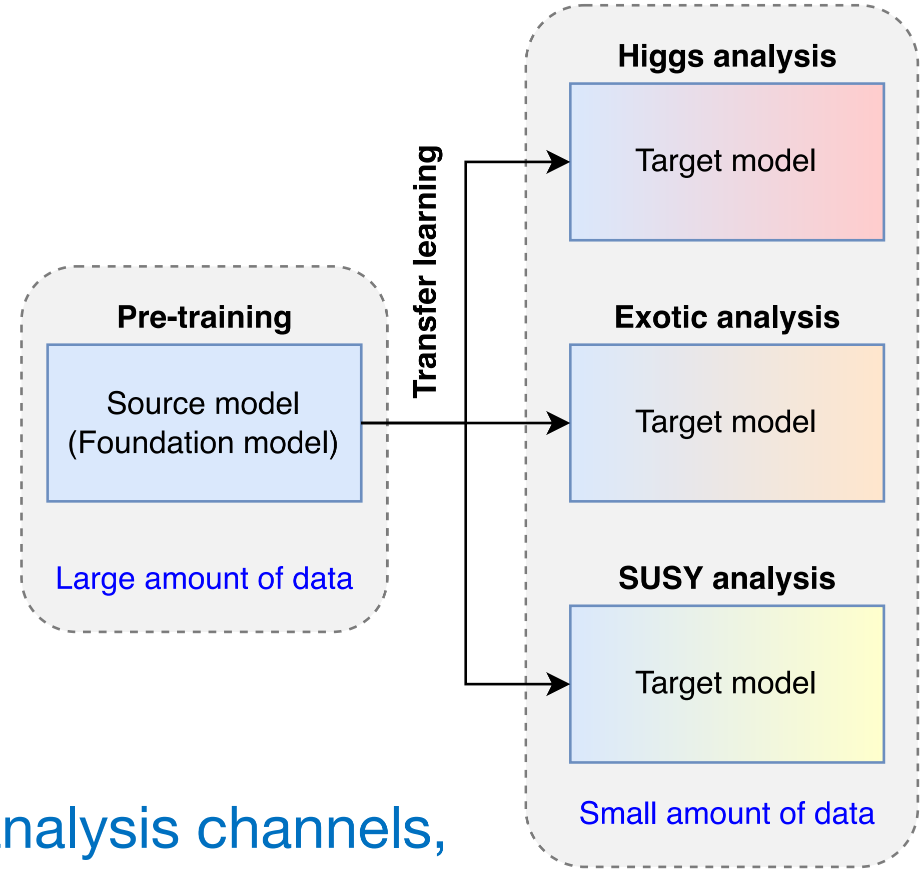
加速器だから見える世界。
KEK

Transfer learning



Use case of physics analysis

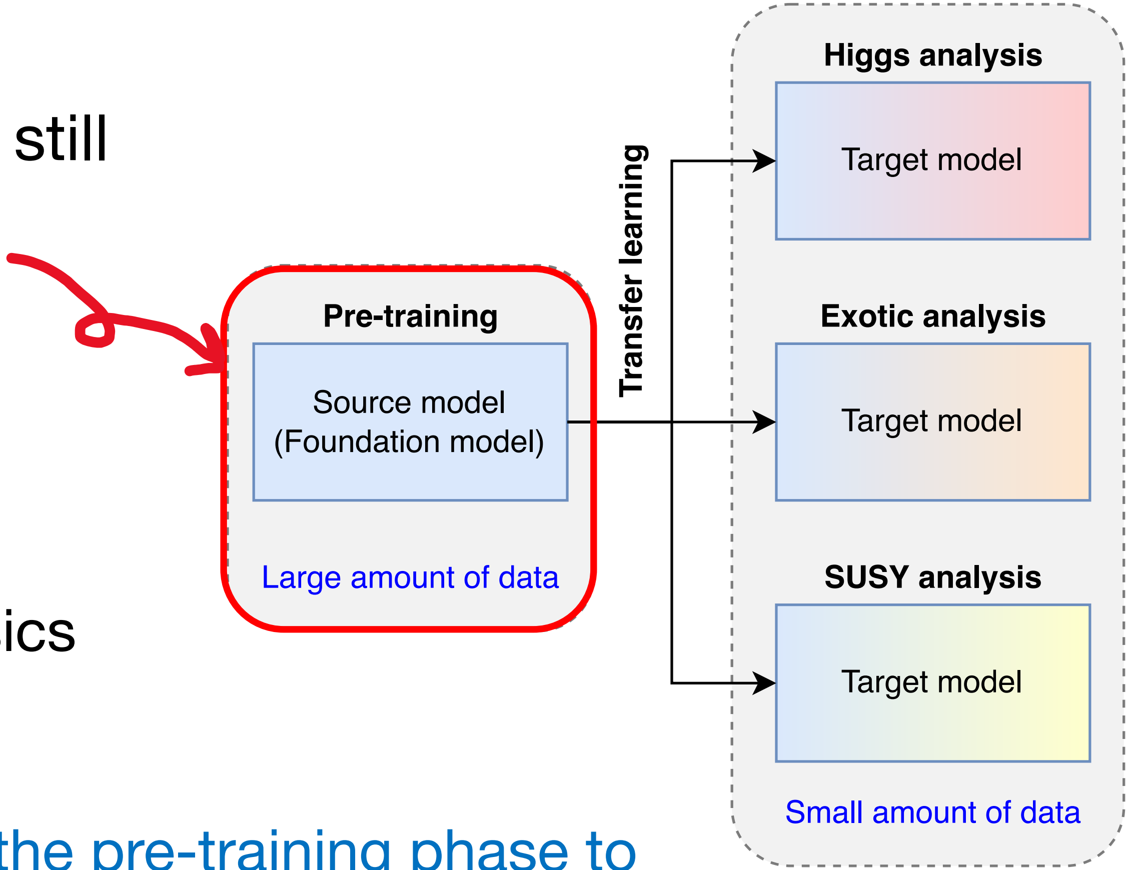
- Many analysis channels in collider physics
  - Higgs, Exotic, SUSY, etc
  - Currently, dedicated DL models are trained from scratch for each channel ← Large amount of training data (MC) for each channel



→ If transfer learning can be applied to different analysis channels, computing resources for MC simulations and DL training are saved

Limitation of idea

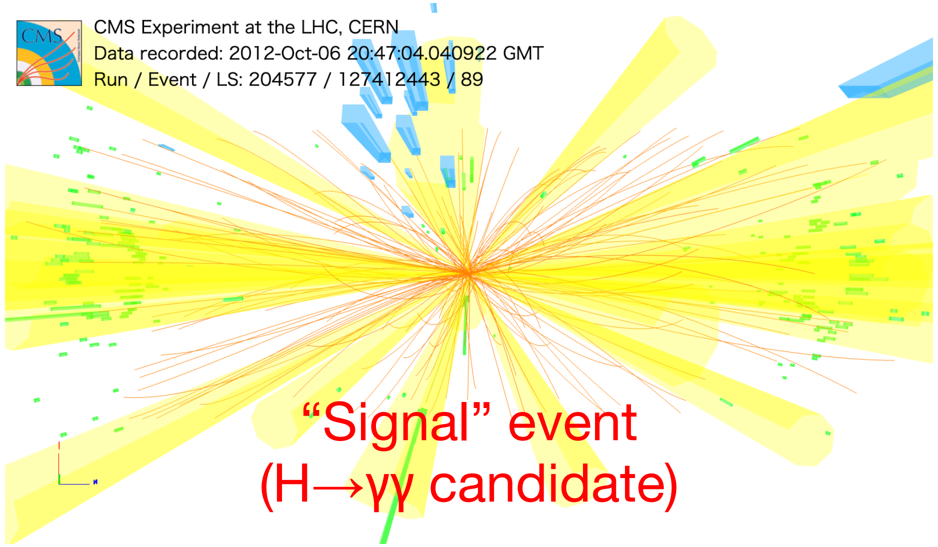
1. Large amount of MC simulations is still required for the pre-training phase
2. Choice of physics process of MC simulations is arbitrary
   - ➢ Transfer learning shows better performance between similar physics processes (Ref: PoS(ISGC2022)016)

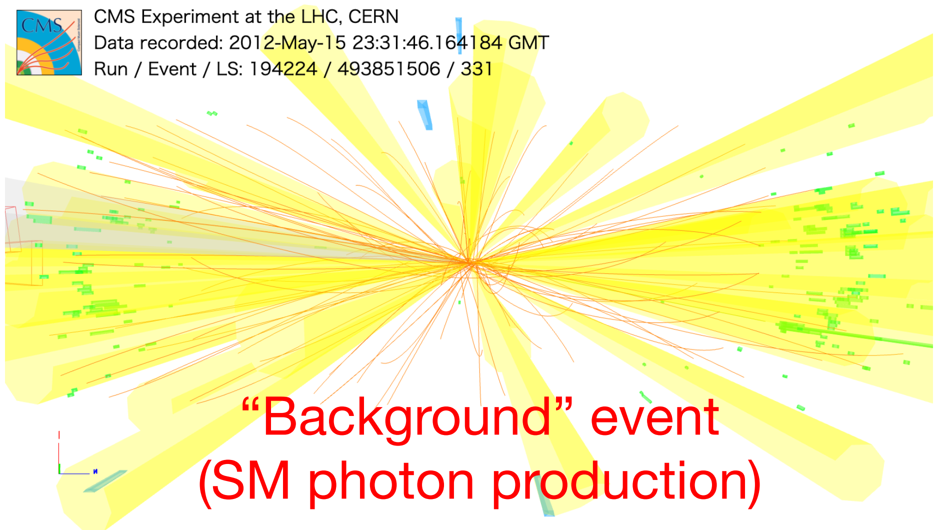


→ Real particle collision data are used in the pre-training phase to overcome these limitations

Event classification

- The concept was examined using “event classification” problem
  - A typical problem in HEP, signal event vs. background event

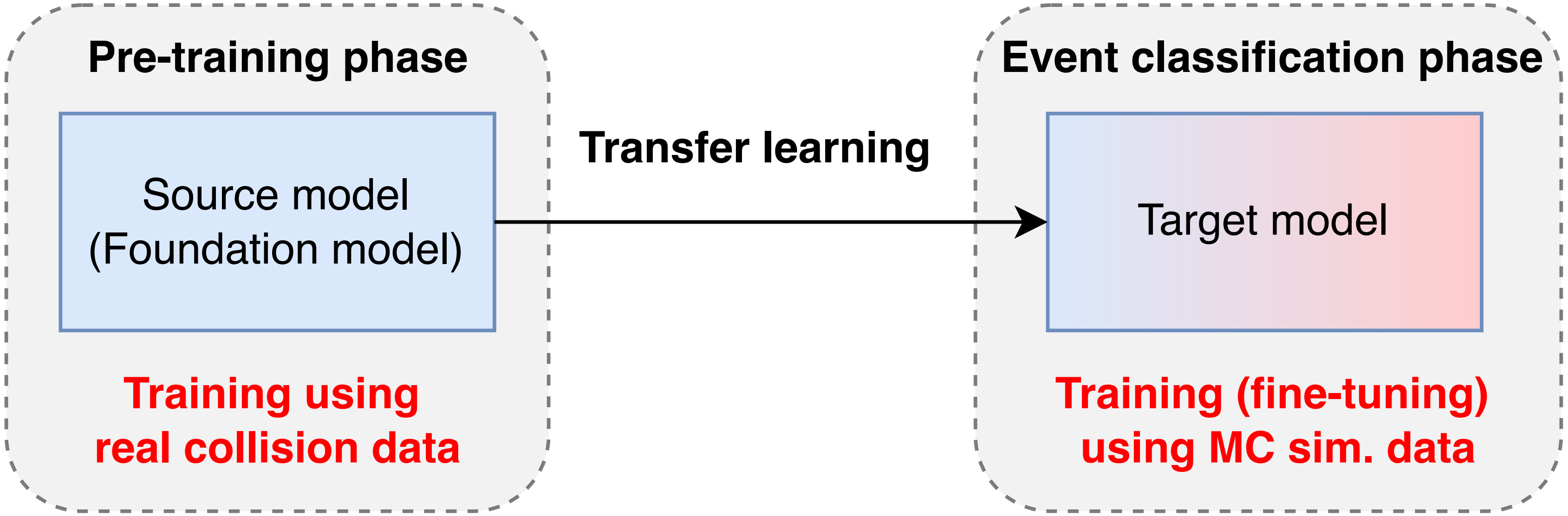




→ Reconstructed particles (objects) are the basic information for the classification

Event classification

➢ Two phases of the training:

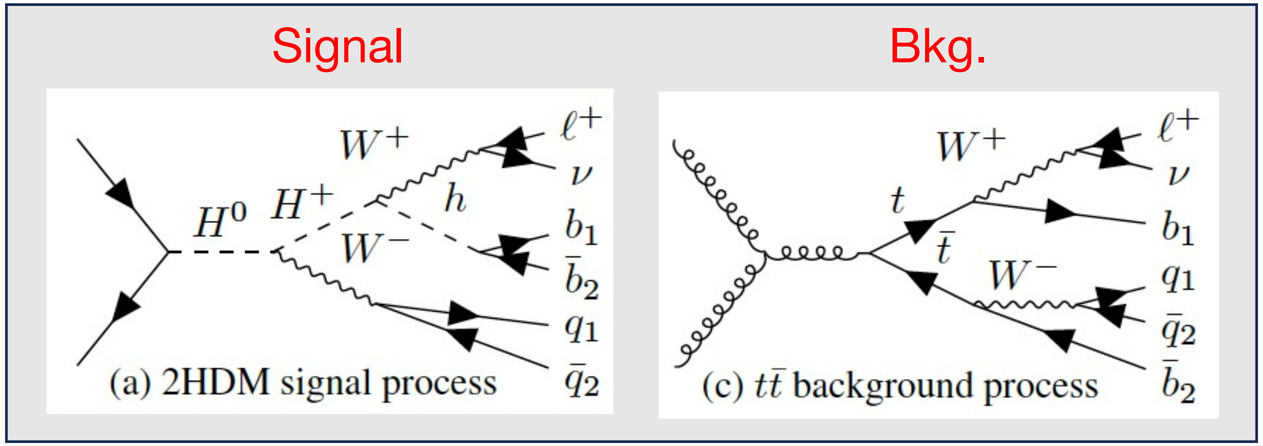


→ Event classification performance (AUC) is compared with and without the pre-training phase

加速器だから見える世界。
KEK

Datasets

- Pre-training phase:
  - [CMS 13TeV opendata]
  - Pre-selection: (at least 1 lepton) + (at least 2 b-jets) + (at least 2 light-jets)
  - ~ 1M events are available after the pre-selection
- Event classification phase:
  - 2HDM vs. ttbar
  - Madgraph + Pythia8 + Delphes (CMS card)

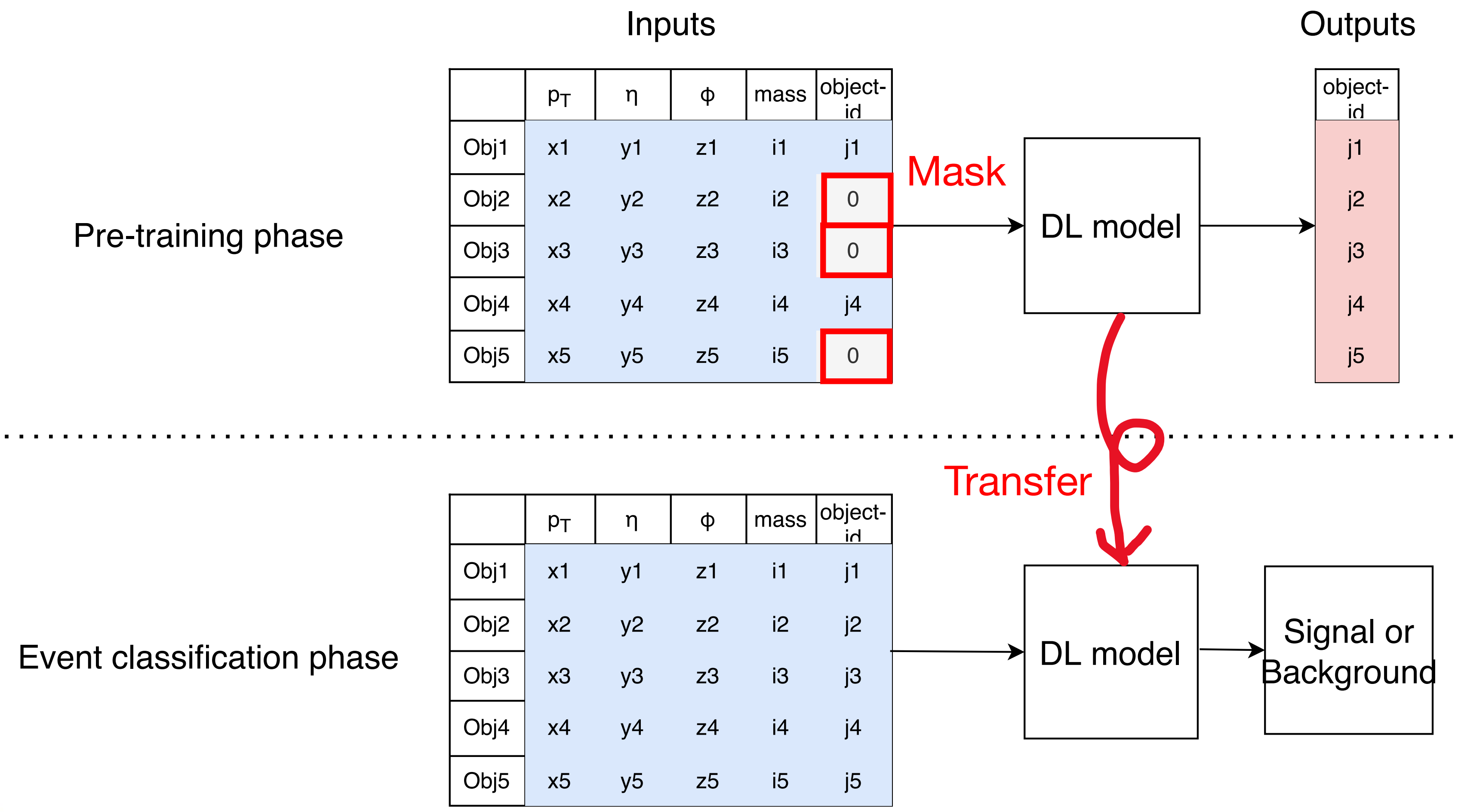


(a) 2HDM signal process
(c) $t\bar{t}$ background process

Pre-training strategy

- Only low-level features of each object (4-vector, object-id) are used as inputs
- Self-supervised learning is employed to handle the unlabeled real data
- Strategy:
  - Object-id (lepton, b-jet, light-jet, or MET) is randomly masked by zeros when preparing a mini-batch
    → DL model is trained to predict masked object-ids as a multi-label classification
  - All input features, including object-id, are used in the target event classification

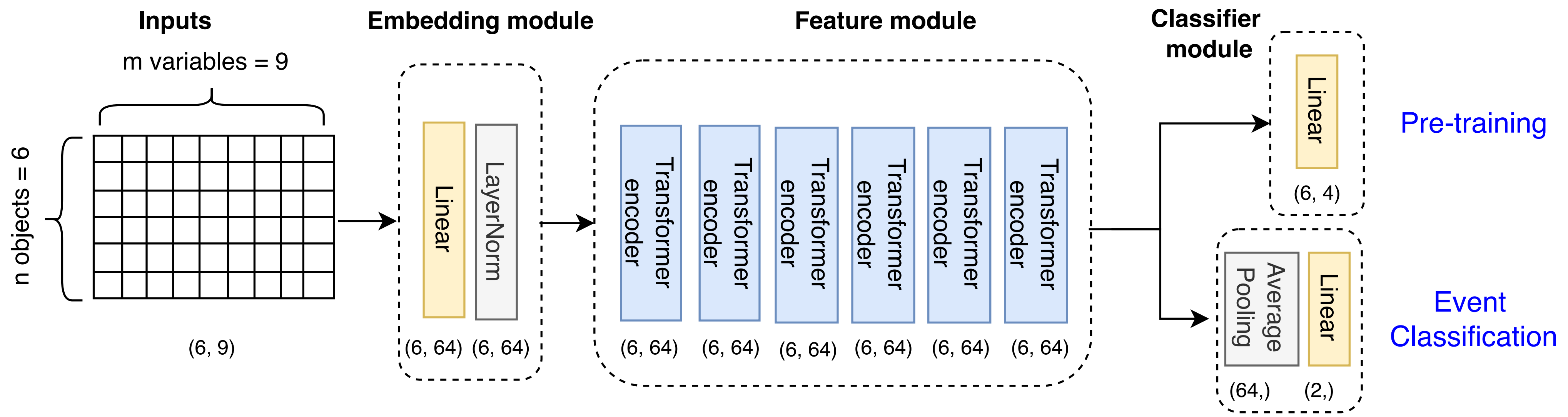
Pre-training strategy



→ Random masks increase prediction pattern (data augmentation)

DL model

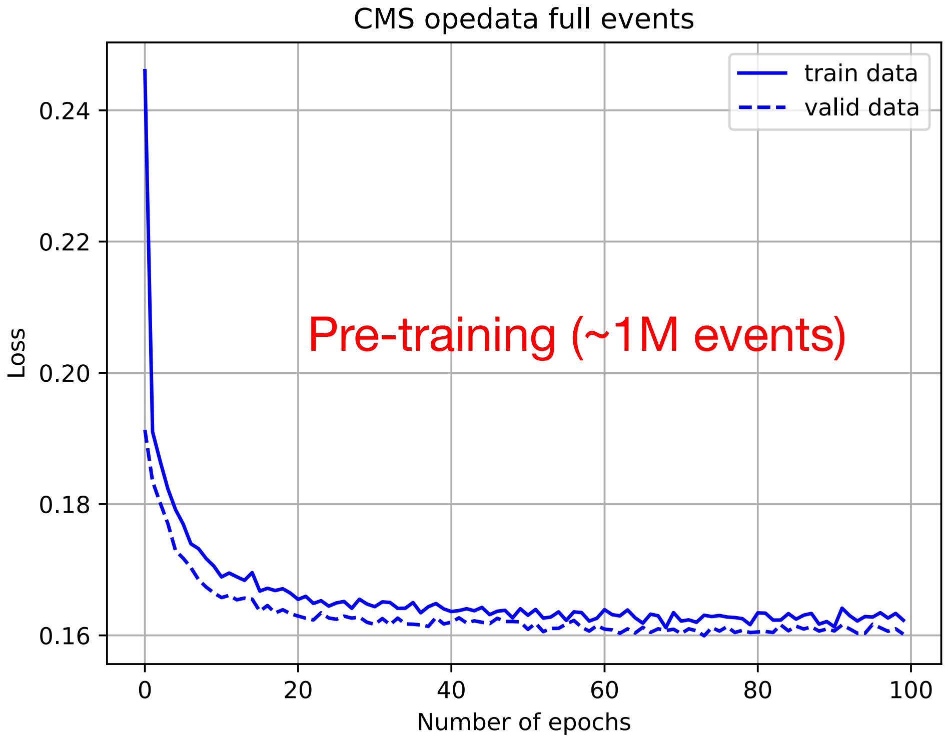
- Transformer encoder is employed:
  - ~1.7M trainable parameters



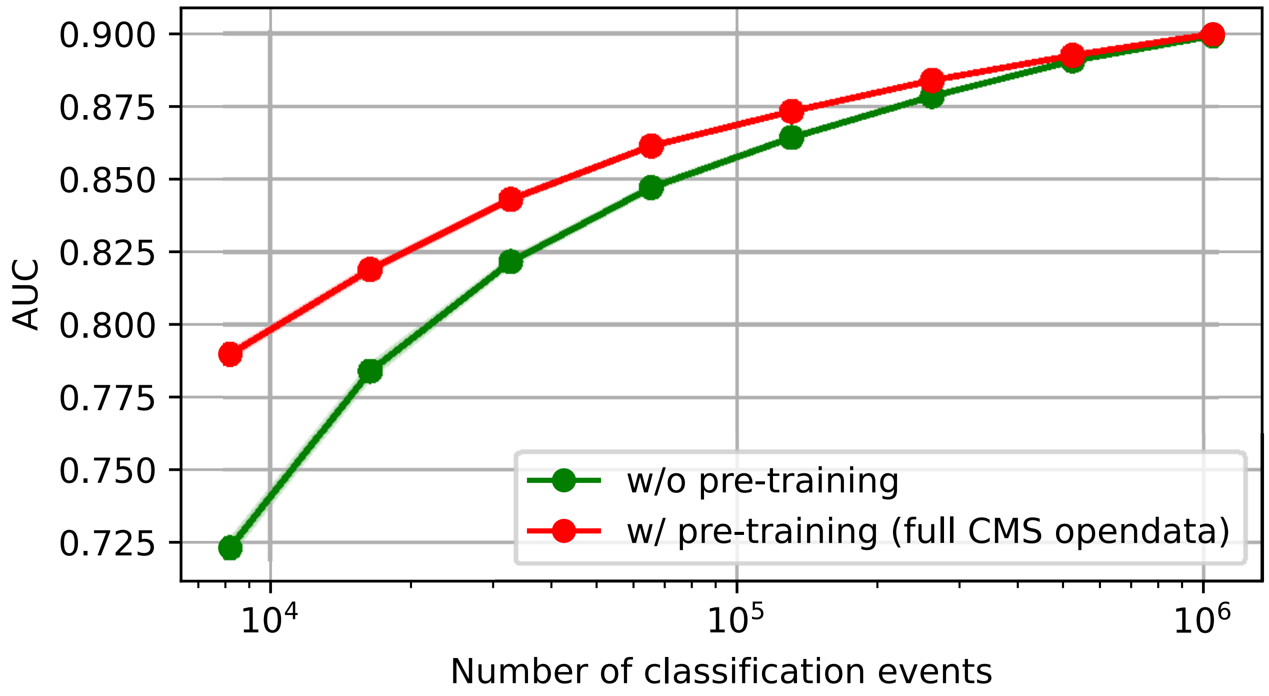
→ Weight parameters of embedding and feature modules are transferred and fine-tuned
→ Classifier module is always trained from scratch

Training details

- Basically, the same setting between the pre-training and event classification phases:
  - SGD optimizer:
    - Learning rate: 10^{-2}-10^{-4} (CosineAnnealingLR)
  - Batch size: 1024, Epochs: 100
  - Cross entropy loss:
    - Pre-training: lepton, b-jet, l-jet, or MET
    - Event classification: 2HDM or ttbar
- NVIDIA A100: ~90 batches/s



AUC of event classification



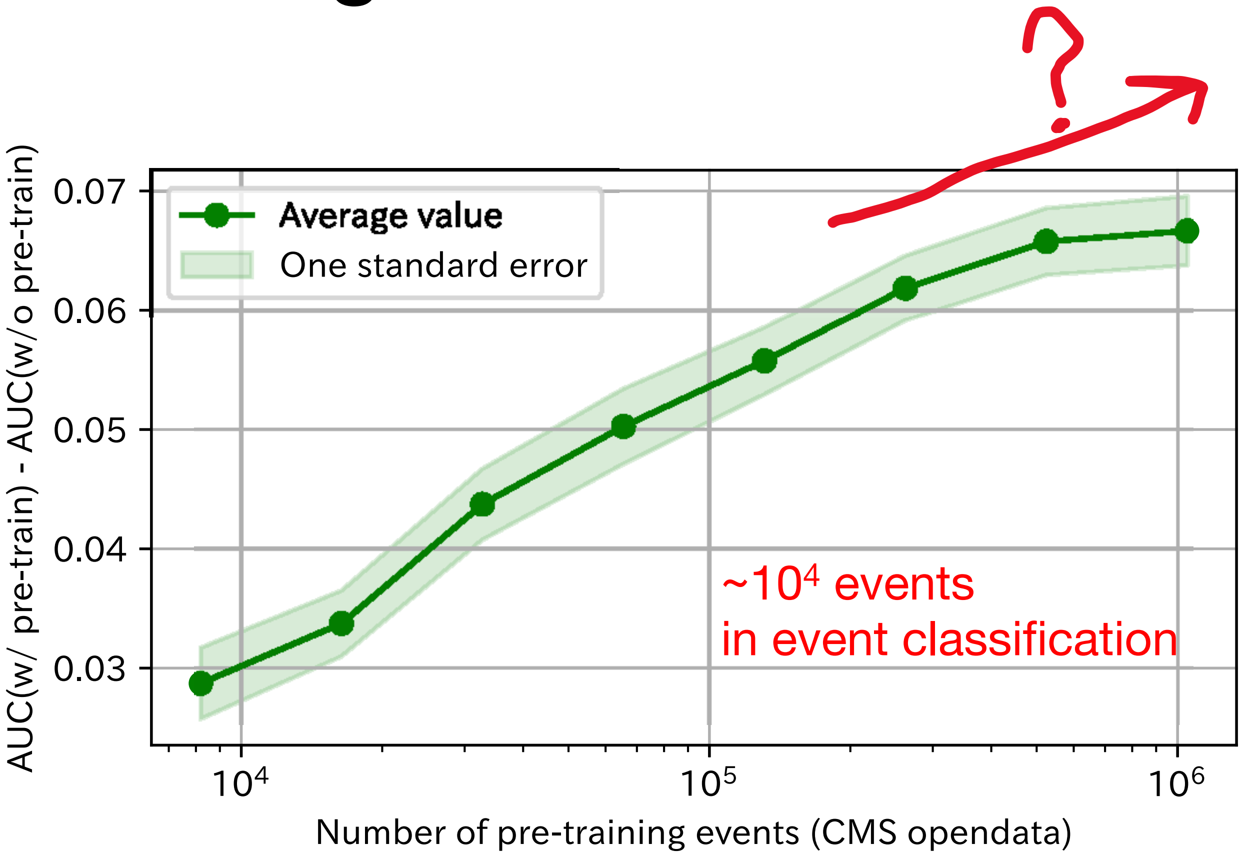
Small data

Large data

- ➢ Significant improvement when # of events in event classification is small (~10^4)
  - ➢ Performances converged when # of events increased to ~10^6

← Expected behavior of the transfer learning

Scaling raw



- Currently, event classification performance improves by increasing events in the pre-training phase
- (One training with 10^{10} events will require (A100 x 8) x 15 days)

加速器だから見える世界。 KEK

Limitations of our experiments

- The scaling behavior encourages a pre-training with a larger data
  - However, the number of events in the CMS open data itself is limited

→ Discussions with ATLAS colleagues are ongoing

ICEPP
The University of Tokyo

- We should adapt the pre-trained model to different signal events to evaluate the generalization of the model
- We also need to evaluate the foundation model’s impact on reducing computing costs

Summary

- Focusing on transfer learning techniques and studying their applications to collider physics
  - Motivated by reduction of computing resources for future experiments
- Transfer learning: Self-supervised learning using real data → Event classification
  - Significant improvements when the # of events in event classification is small
  - The scaling behavior encourages pre-training with a larger data

Input variables