# *Potential for AutoML from PUNCH*

Gregor Kasieczka
(gregor.kasieczka@uni-hamburg.de)
Twitter: @GregorKasieczka

NFDI Workshop, 5.4.2022

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

PIER
Partnership of
Universität Hamburg and DESY

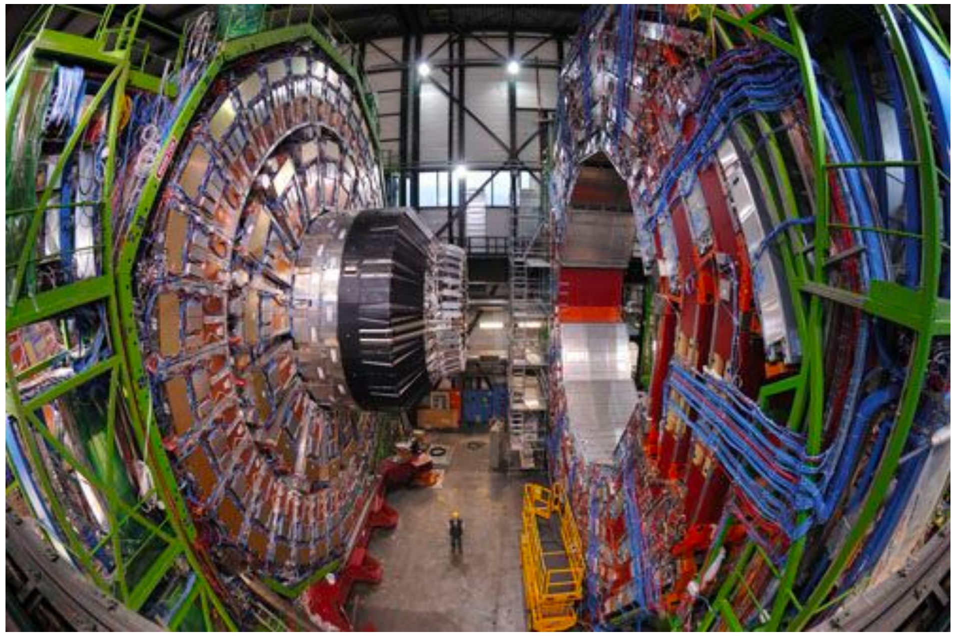
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

Bundesministerium
für Bildung
und Forschung

# Motivation

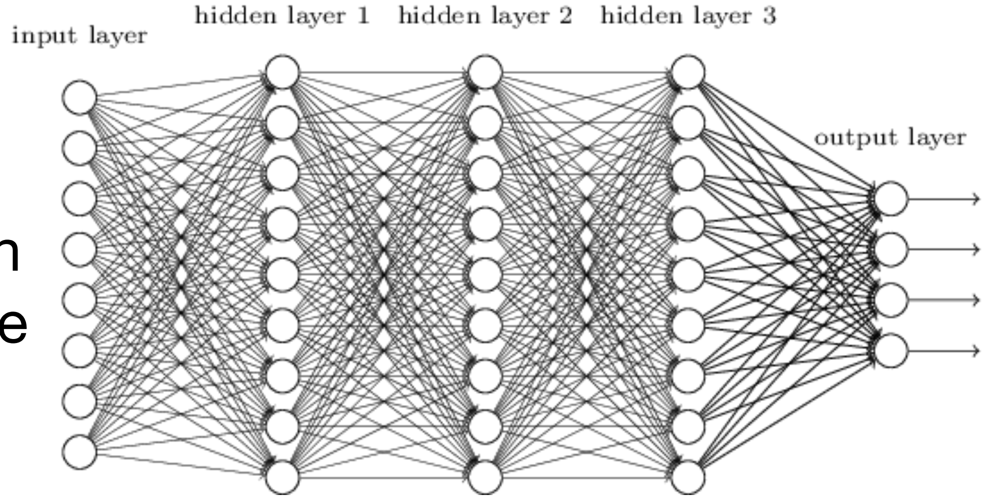
**Common themes:**

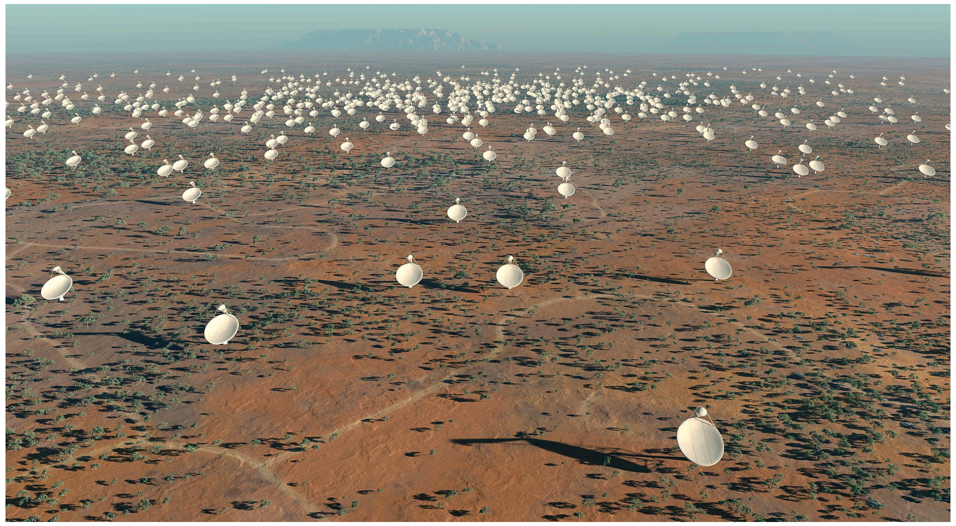
- Large volumes of experimental and simulation data
- Interesting structures
(symmetries, causality, compositeness)
- Underlying mechanistic model…
- …but complex noise (e.g. detector response)
- High-quality labelled synthetic training data
- Well understood uncertainties of predictions



***CMS experiment at CERN***
*40 million proton-proton collisions / second*
*~1 PB/s non-zero-suppressed raw input data*
*~1 GB/s stored*

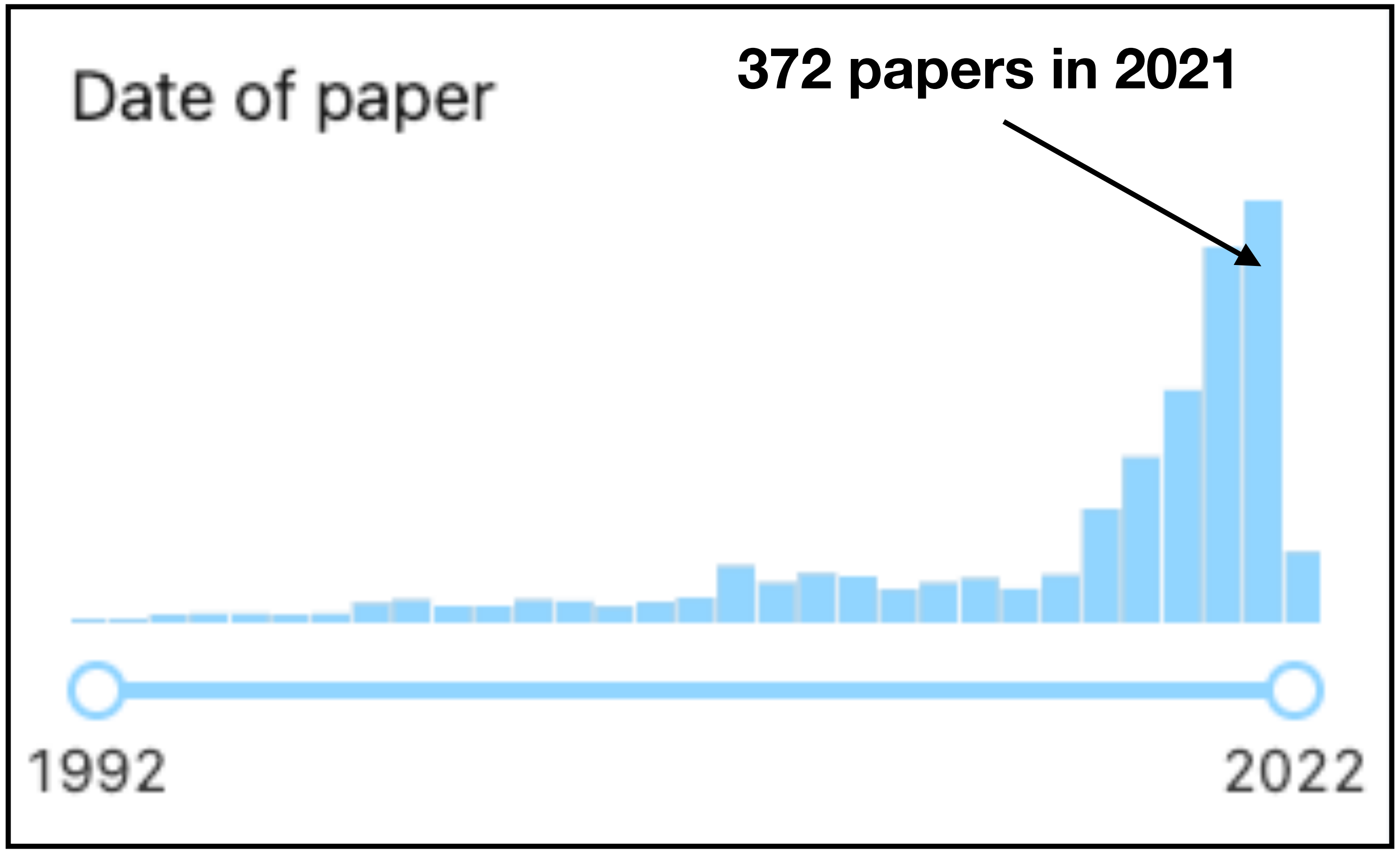
**Note**: Many machine learning developments in physics center around the use of (deep) neural networks



***SKA radio telescope***
*Operation from 2027 onwards*
*Expect ~1 EB/day raw input data*

# Activity



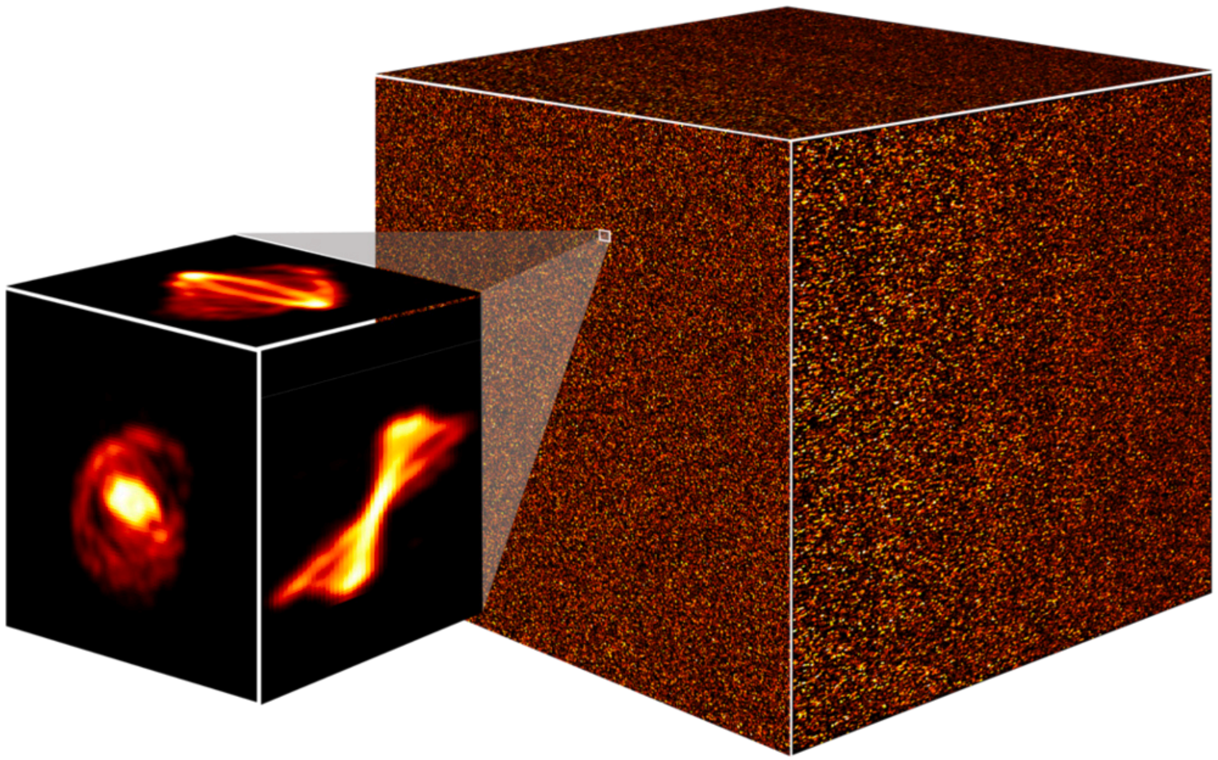

**Inspire Search:**

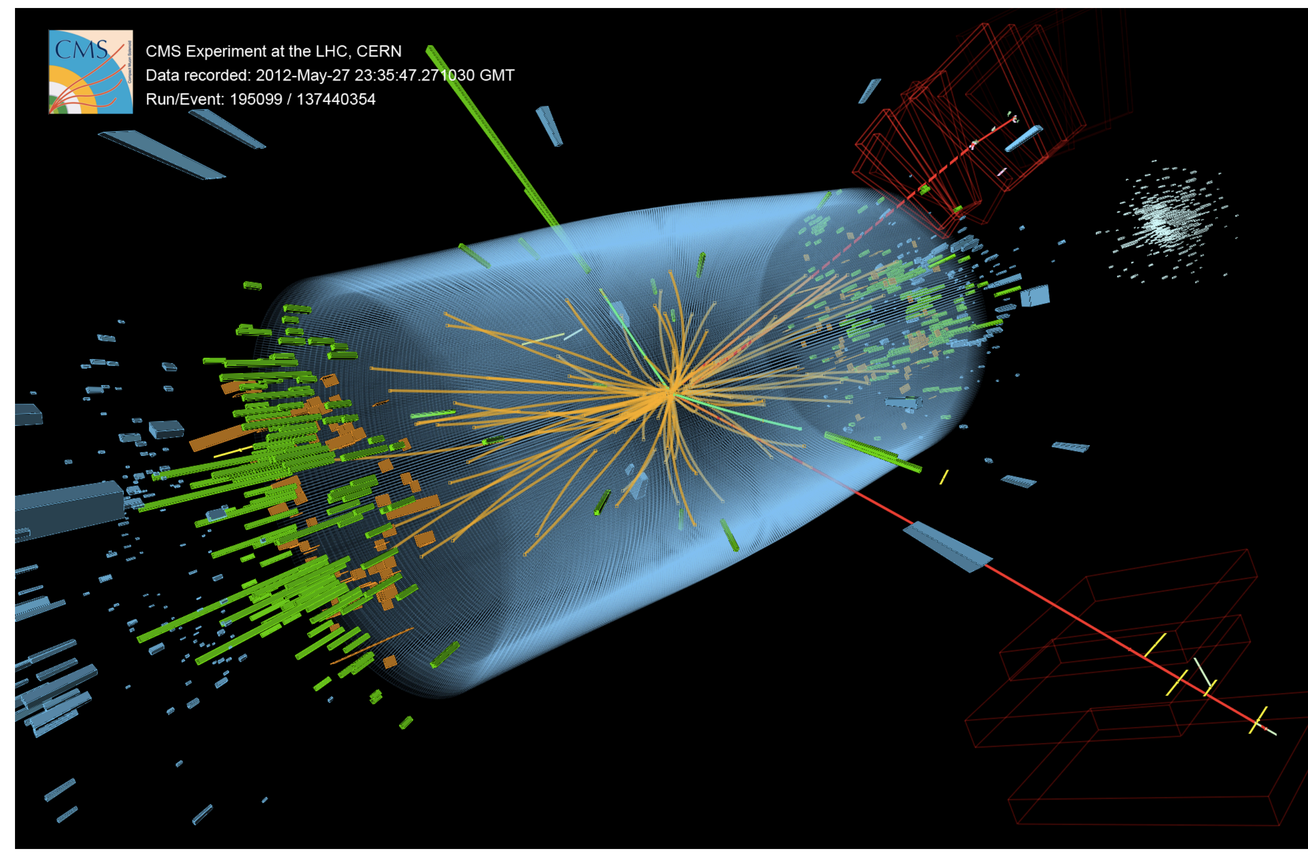
*("machine learning" or "deep learning" or neural) and (hep-ex or hep-ph or hep-th)*

**Very active and wide-spread adaptation of modern machine learning in PUNCH domains.**

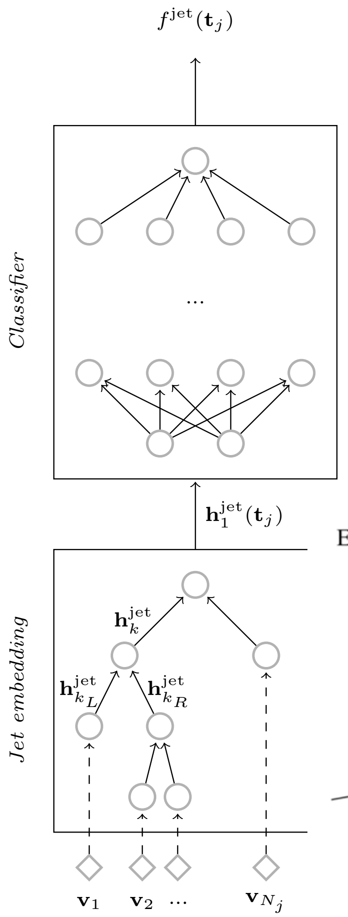
**See e.g. arxiv:2112.03769 for a short review.**

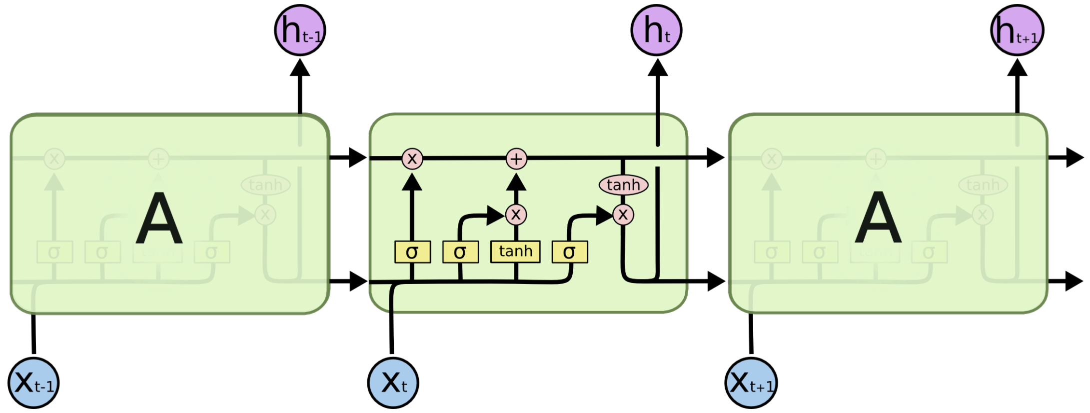
# Diversity of data types…

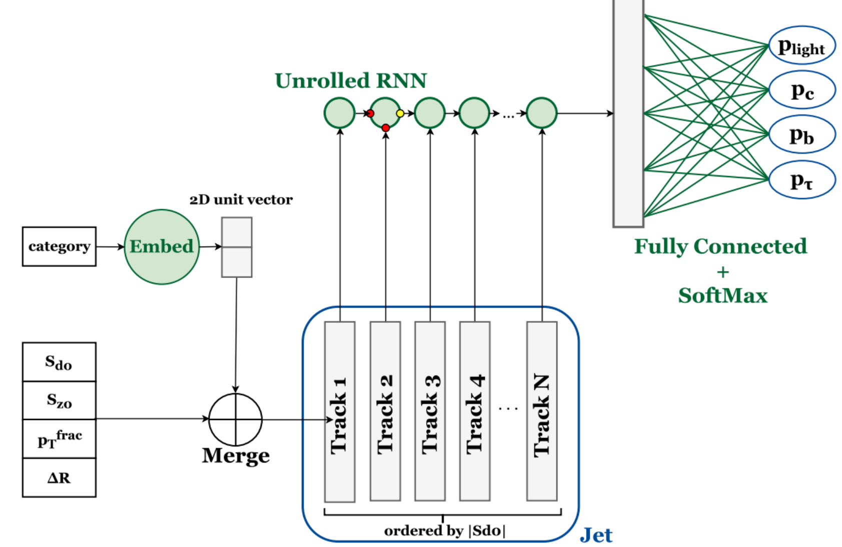


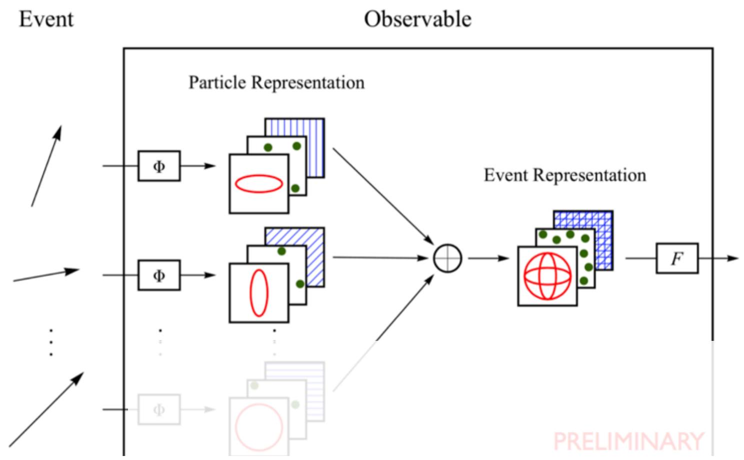


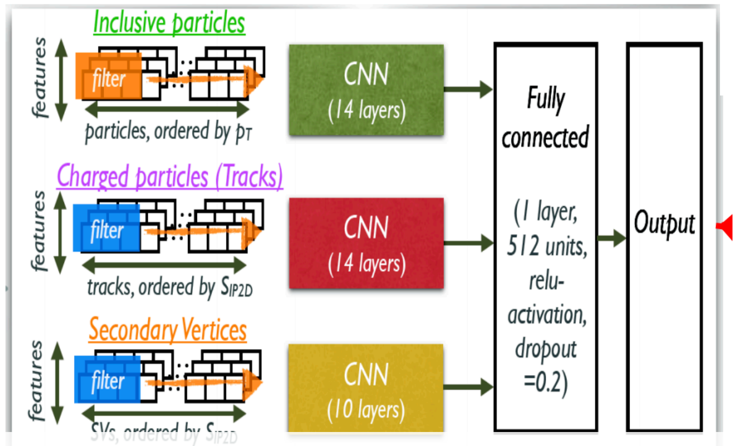

# ...and approaches

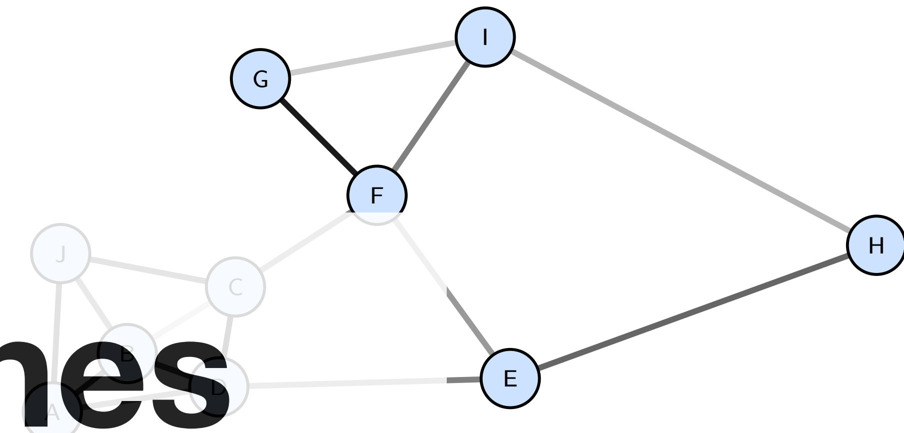


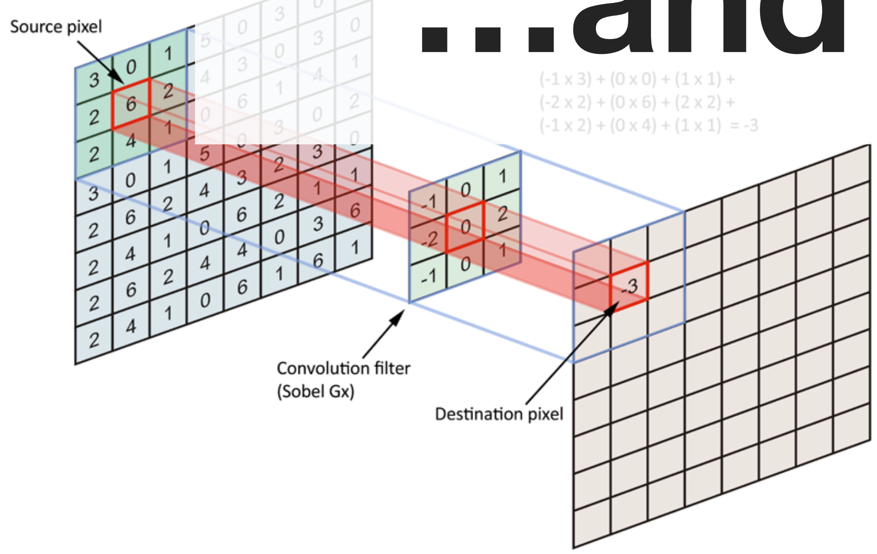



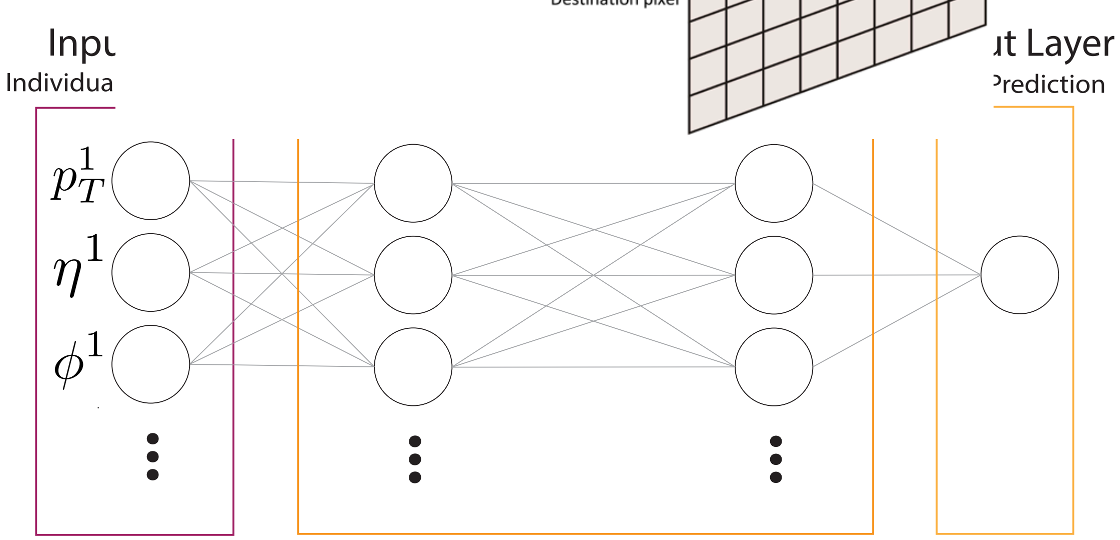



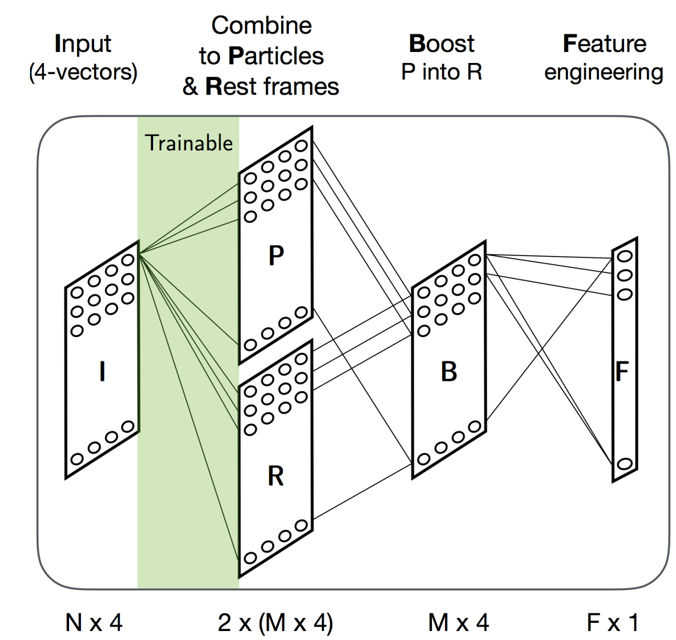

$$k_{\mu,i} = \begin{pmatrix} E_0 & E_1 & \dots & E_N \\ p_{x,0} & p_{x,1} & \dots & p_{x,N} \\ p_{y,0} & p_{y,1} & \dots & p_{y,N} \\ p_{z,0} & p_{z,1} & \dots & p_{z,N} \end{pmatrix}$$

Combination Layer (**CoLa**): create linear combinations:

$$k_{\mu,i} \xrightarrow{\text{CoLa}} \widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$$

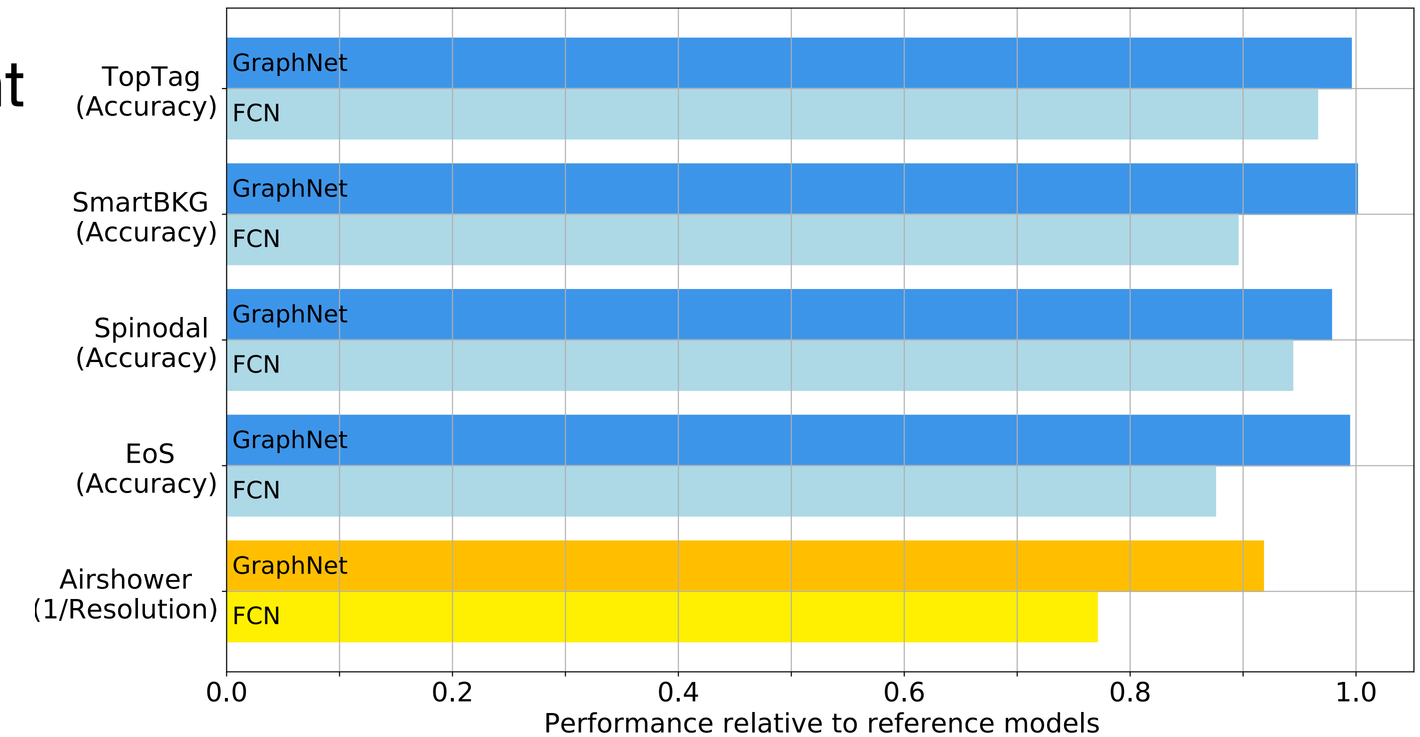
Lorentz Layer (**LoLa**): Use resulting matrix to extract physics features.
Main assumption is the Minkowski metric

$$\widetilde{k}_{\mu,i} \to \sum_j (\widetilde{k}_i - \widetilde{k}_j)_\mu (\widetilde{k}_i - \widetilde{k}_j)_\nu \eta^{\mu\nu} B_{ij}$$



# Finding common ground

- Built collection of datasets from different scientific domains (particle, hadron&nuclei, astro-particle) (Open Data, of course)
  - Focus on supervised learning tasks (classification/regression for simplicity)
  - (Open for more additions)
- Developed graph-based model achieving state-of-the-art performance on all datasets out-of-the-box




**See arXiv:2107.00656 and [https://github.com/erum-data-idt/pd4ml](https://github.com/erum-data-idt/pd4ml)**

| | Task | Examples (train/test/validation) | Structure | Dimension |
|---|---|---|---|---|
| Top Tagging Landscape | Class. | 1.2M/400k/400k | Four vectors | 200 particles, 4 features/particle |
| Smart Backgrounds | Class. | 157k/39k/84k | Decay Graph | 100 particles, 9 features/particle |
| Spinodal or Not | Class. | 16.3k/4k/8.7k | 2D Histogram | 20x20 histogram of pion spectra |
| EoS | Class. | 121k/25k/54k | 2D Histogram | 24x24 histogram of pion spectra |
| Air Showers | Regr. | 56k/30k/14k | 81 1D Traces | 81 stations, 80 signal bins + timing |

# Vision going forward

**Machine Learning pipeline:**

Input in custom format (experiment, theory framework)

↓

Convert to ML dataformat

↓

Data pre-processing

↓

Feature engineering

↓

Model/architecture selection

↓

Hyperparameter tuning

↓

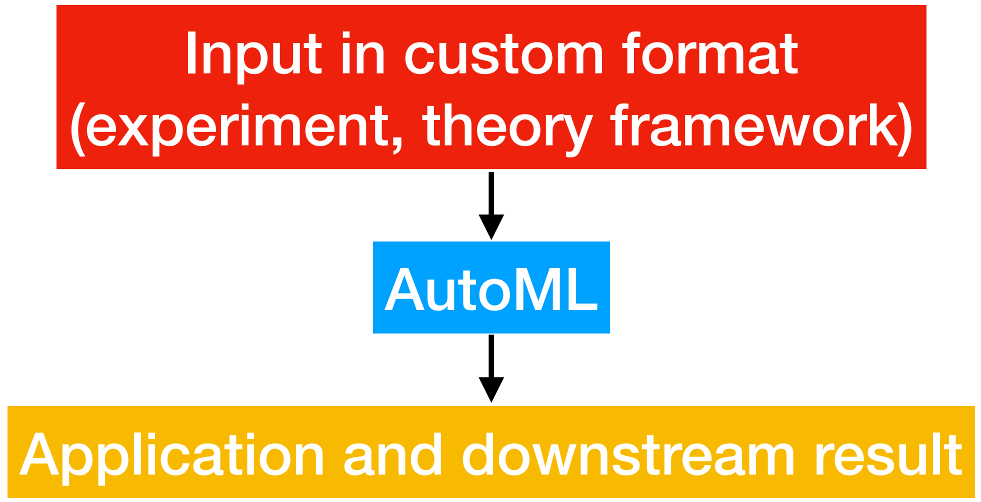
Application and downstream result

- Large part of research time spent on these tasks
- Similar for many of our applications
- Automate these: AutoML
- *(the goal is not to produce the best model for each scenario but provide a reliable baseline with minimal user effort)*

# Some thoughts:

- AutoML is a (much) bigger research effort outside of our domain
- Do not compete with, find a way to build a fully automated toolchain
- Make sure it works for all our examples (limit to supervised tasks)
- Build our domain expertise into the automated model selection
- Possibility to extend to data from areas (e.h. consortia present today?)

**Goal:**



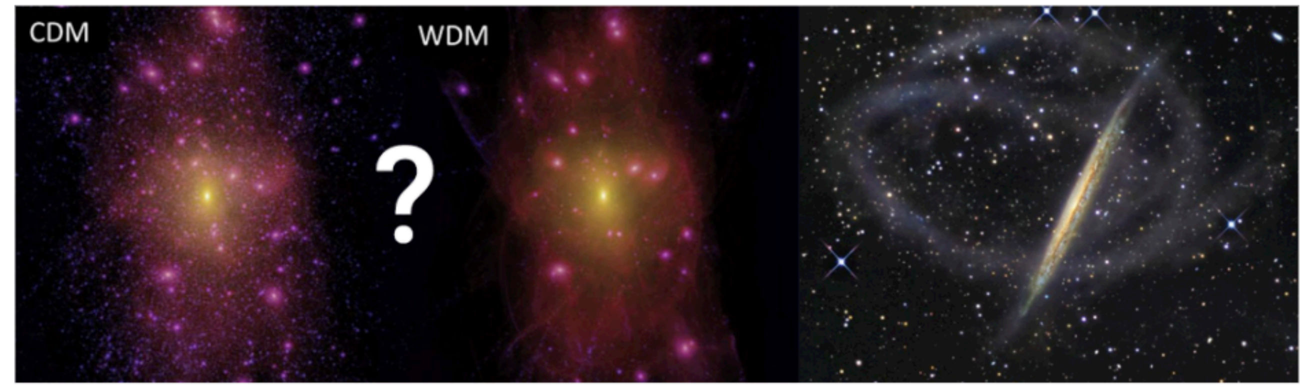
# Concrete steps

- **Data**
  - Can work with PD4ML collection as a start; integrate with PUNCH science data portal
- **Algorithms** / Meta-learning and automated Deep Learning
  - Initially focus on hyperparameter search, in particular:
    - SMAC3: Bayesian optimization to probe configuration space
    - Hyperband: preempt computations whose configurations are wasteful
- **Infrastructure** / Technical aspects
  - Deliver as a wrapper of a workflow to maximize applicability
  - Initial prototype based on https://docs.ray.io/en/latest/tune/index.html or methods from the Freiburg-Hannover group
  - Targeted features:
    - Workflow and resource definitions directly in Python
    - Distributed training and hyperparameter optimization
    - Use cross-platform compute infrastructure (e.g., Kubernetes, Slurm Workload Manager).



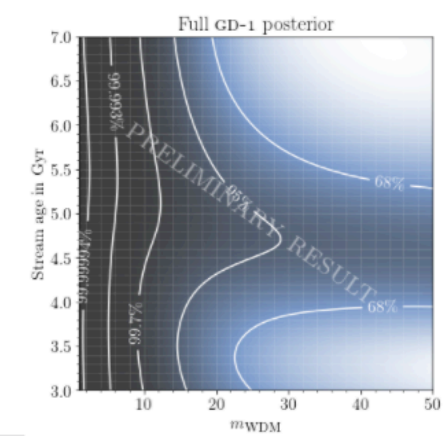
***Many thanks to Joeri Hermans!***

# Example use-case: Automated Simulation-Based (Bayesian) Inference




- User provides:
  - A prior p(ϑ)
  - A simulation model that accepts ϑ ~ p(ϑ) to produce x ~ p(x | ϑ)
  - A set of observations (optional)

| Empirical coverage probability | | | |
|---|---|---|---|
| Architecture | 68% CR | 95% CR | 99.7% CR |
| $\hat{r}(x\|\vartheta)$ with $\vartheta \triangleq (m_{\mathrm{WDM}})$ | | | |
| MLP | 0.704 $\pm 0.004$ | 0.972 $\pm 0.002$ | 0.999 $\pm 0.000$ |
| MLP-BN | 0.706 $\pm 0.003$ | 0.970 $\pm 0.001$ | 0.999 $\pm 0.000$ |
| RESNET-18 | 0.687 $\pm 0.004$ | 0.955 $\pm 0.002$ | 0.998 $\pm 0.000$ |
| RESNET-18-BN | 0.693 $\pm 0.004$ | 0.966 $\pm 0.002$ | 0.999 $\pm 0.000$ |
| RESNET-50 | 0.689 $\pm 0.006$ | 0.967 $\pm 0.001$ | 0.998 $\pm 0.000$ |
| RESNET-50-BN | 0.698 $\pm 0.004$ | 0.969 $\pm 0.001$ | 0.999 $\pm 0.000$ |




- User gets:
  - An ensemble of trained posterior estimators (in ONNX)
  - Statistical diagnostics (expected coverage for all confidence levels, SBC)
  - A pre-generated Jupyter notebook with all results (including constraints on ϑ)



***Many thanks to Joeri Hermans!***

# Closing

- Wide and growing use of machine learning in fundamental sciences
- Large potential from automating parts of this workflow
- Interesting challenges on the algorithmic and technical side
- Will carry out these developments on the PUNCH side, very open towards other application domains!

***Thank you!***