Relevance of Transformer Models in CoE RAISE

Prof. Dr. – Ing. Morris Riedel

School of Engineering & Natural Sciences, University of Iceland, Iceland

Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany

2022-09-26, RAISE CoE Training Transformer Models, Online

IHPC National Competence Center (NCC) for HPC & AI in Iceland

@ProfDrMorrisRiedel   @Morris Riedel   @MorrisRiedel   @MorrisRiedel

https://www.youtube.com/channel/UCWC4VKHmL4NZgFfKoHtANKg   morris@hi.is
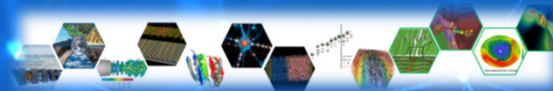
CoE RAISE Web Page & More Information

https://www.coe-raise.eu
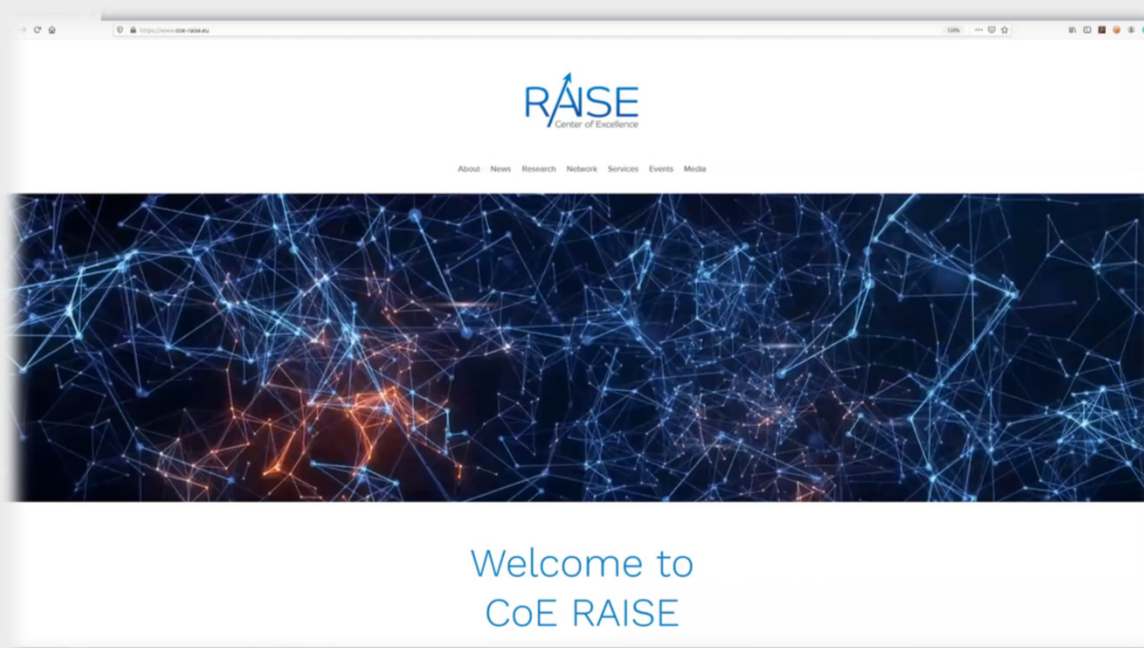
CoE RAISE – Motivation & Approach

Simulation / Experiment [3]

data generation

predict / replace / accelerate

Big Data

Surrogate / Model

input

$+$ 10^{18}

AI technologies   Exascale

training
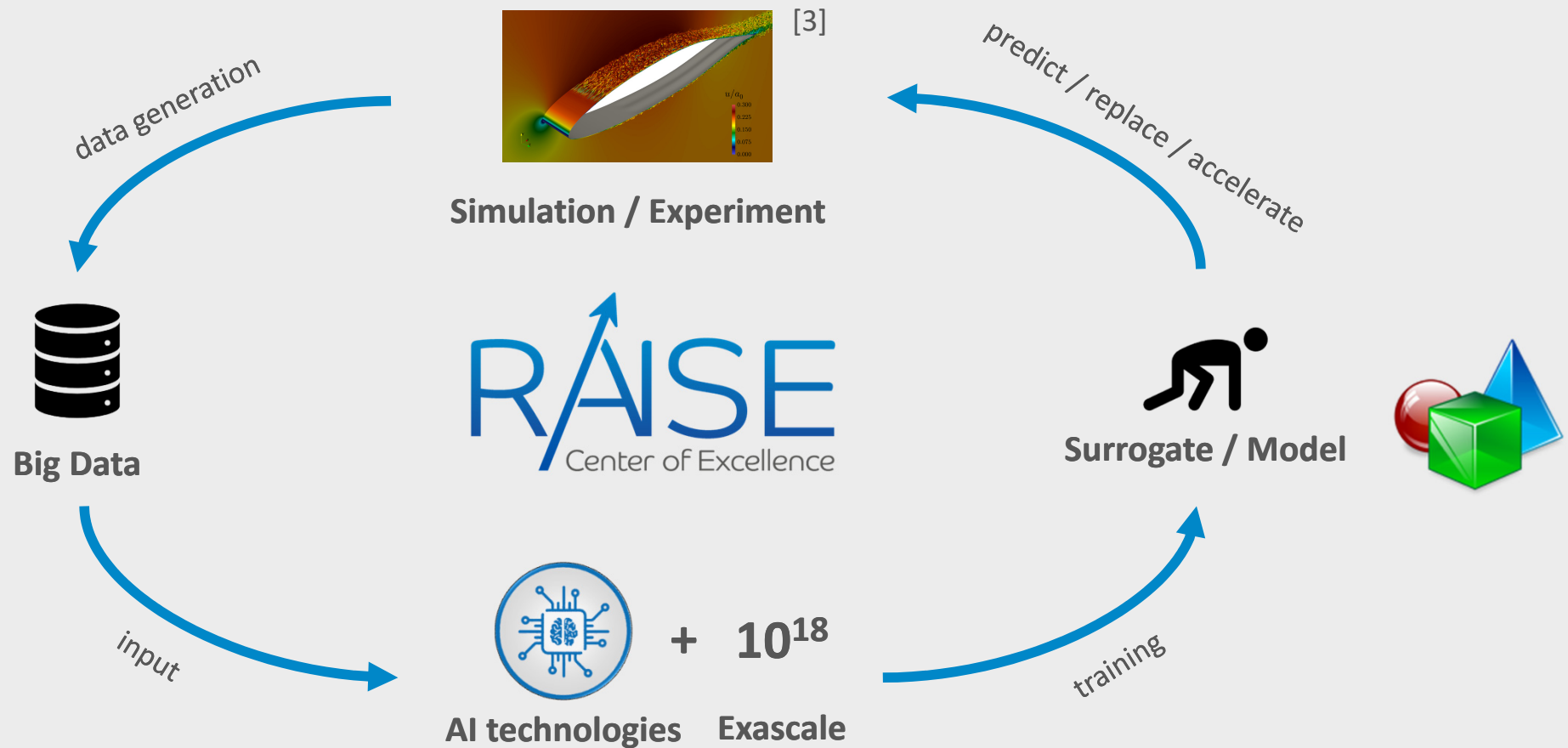
Use Cases in CoE RAISE

➤ Two kinds of use cases:

turbulent flow · clean energy · reactive flows · engine design · coating

compute-driven use-cases → AI at Exascale

fundamental physics · seismic imaging · manufacturing · sound engineering

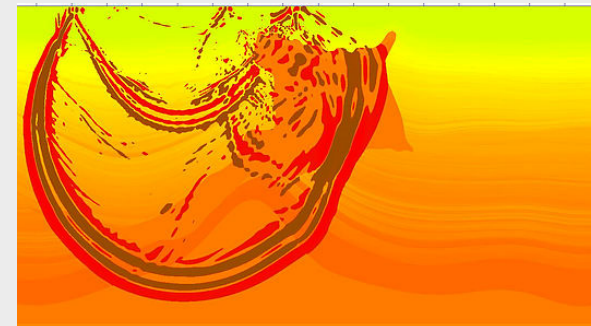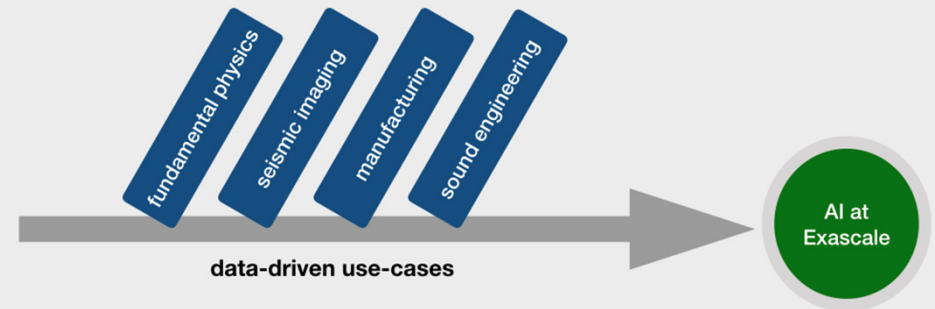**data-driven use-cases** → AI at Exascale
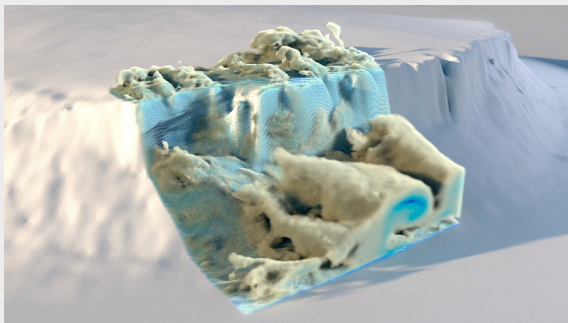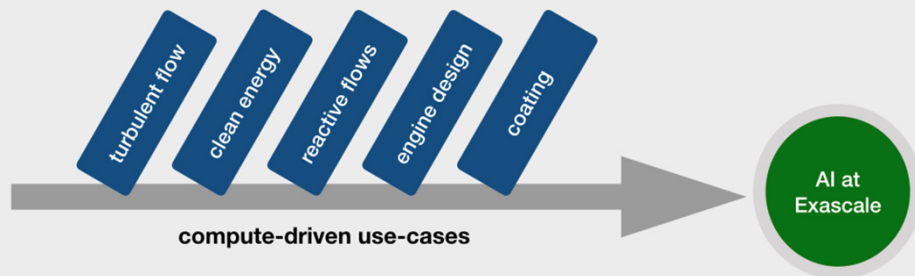
Example from use case "AI for wind farm layout": Turbulence generated by a cliff on Bolund Island, Denmark.

Example from use case " Seismic imaging with remote sensing - oil and gas exploration and well maintenance": Snapshot from a wavefield.

Compute- and Data-driven Use Cases – Data & Modeling

AI for turbulent boundary layers

AI for wind farm layout optimization

AI for data-driven models in reacting flows

Smart models for next-generation aircraft engine design

AI for wetting hydrodynamics

Event reconstruction and classification at the CERN HL-LHC

Seismic imaging with remote sensing - oil and gas exploration and maintenance

Defect-free metal additive manufacturing

Sound Engineering

Mesoscale ~ 1µm

×1000

×1000

Macroscale ~ 1mm

u/a_0

WP2 – AI- & HPC-Cross Methods at Exascale in a nutshell

- WP3 (Compute-Driven Use-Cases towards Exascale)
- WP4 (Data-Driven Use-Cases towards Exascale
- Developments in these WPs will be supported by the cross-linking activities of WP2
  - E.g. scaling machine & deep learning codes with frameworks like Horovod/Deepspeed
  - E.g. introduction to new AI methods such as Long-Short Term Memory (Time series)
  - E.g. data augmentation approaches
  - E.g. benchmarking HPC machines and offer also pre-trained AI algorithms (i.e., transfer learning)
  - E.g. offer neural architecture search methods for hyperparameter – tuning in semi-automatic way
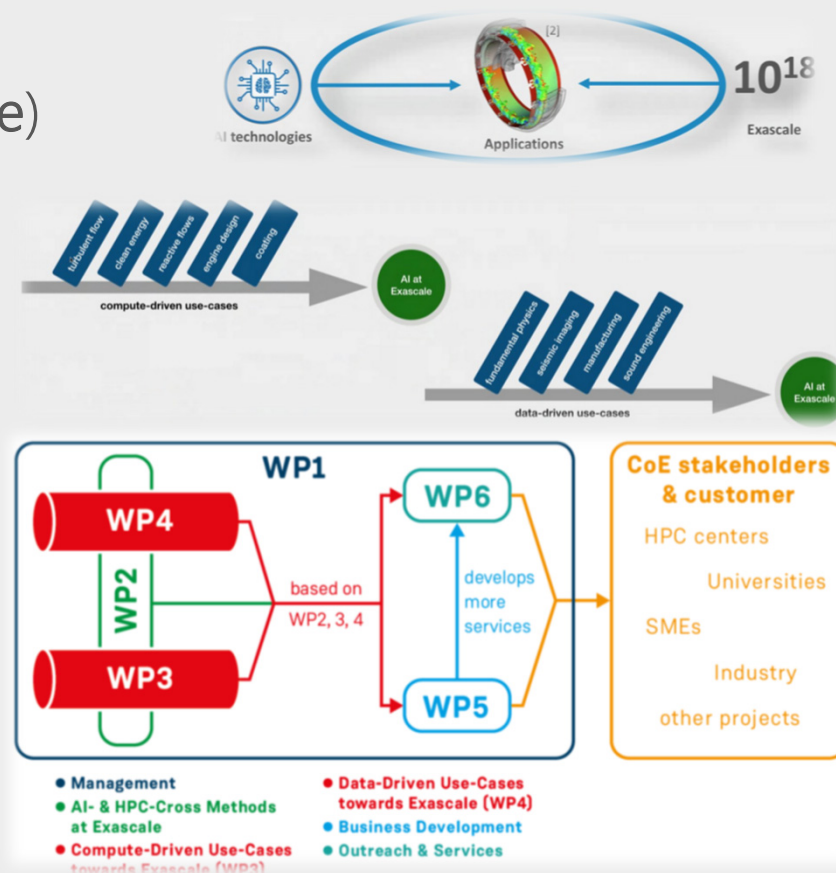
Towards AI & HPC at Exascale with CoE RAISE Results
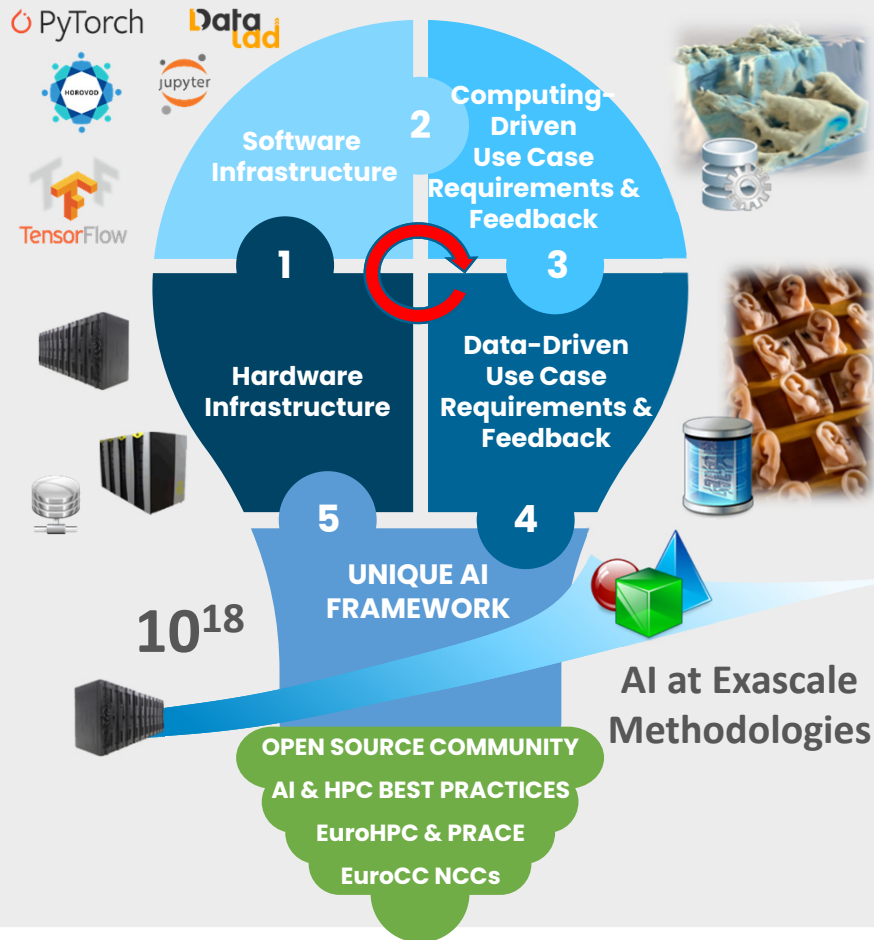
Hardware Infrastructure

Prepare & Document available production systems at partners' HPC centers

Examples: JUWELS (JUELICH), LUMI (UoICELAND), DEEP Modular Prototypes, JUNIQ (JUELICH), etc.

Software Infrastructure

Prepare & Document available open source tools & libraries for HPC & AI useful for implementing use cases

Examples: DeepSpeed and/or Horovod for interconnecting N GPUs for a scalable deep learning jobs

Computing-driven Use Cases Requirements & Feedback

Use cases with emphasize on computing bring in co-design information about AI framework & hardware

Examples: Use feedback that TensorFlow does not work nicely, so WP2 works with use cases on pyTorch

Data-driven Use Cases Requirements & Feedback

Use cases with emphasize on data bring in co-design information about AI framework & hardware

Examples: Deployment blueprint by using AI training on cluster module & inference/testing on booster

→ UNIQUE AI FRAMEWORK

Living design document & software framework blueprint for HPC & AI also with pretrained AI models

Selected Techniques to Identify Cross-Methods for HPC & AI

- Fact Sheets
  - Foster initial understanding
  - Living document & each Fact Sheet per WP3/WP4 Use Case
  - *(Experience from many other EU projects)*
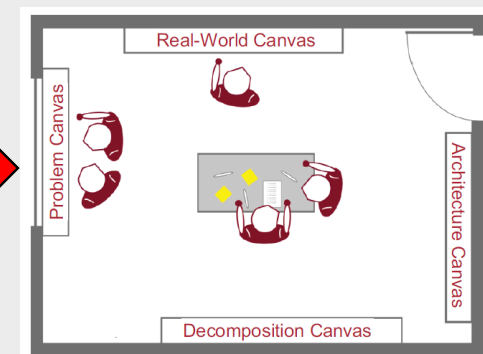
- Selected Contents
  - Short Application Introduction
  - Clarify Primary Contacts
  - Codes/Libraries/Executables
  - HPC System Usage Details
  - Specific Platforms & 'where is what data'?
  - Machine/Deep Learning Approaches of Interest
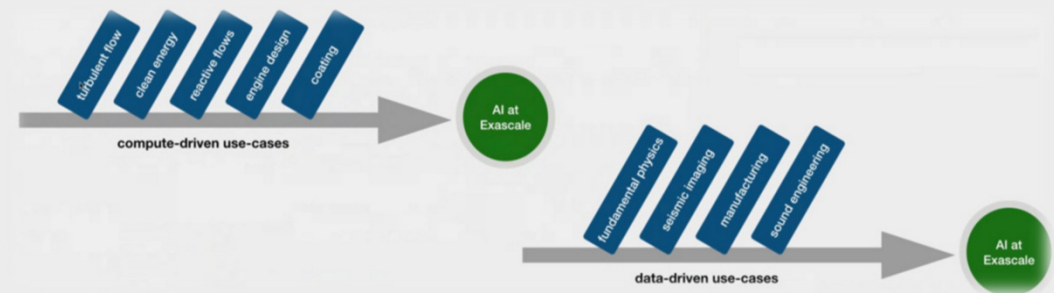
initial steps

Driven by
**Prof. Matthias Book &
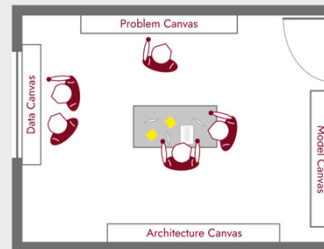Prof. Helmut Neukirchen**

interaction room process

HPC Systems Engineering in the Interaction Room Seminar

- CoR RAISE Interaction Room Process
  - Supports the proper software engineering design of the unique AI framework blueprint
  - Expecting to work with WP3 & WP4 experts in an open minded way
  - Process guided by Software Engineering Expert **Prof. Dr. Matthias Book** (University of Iceland)
  - Supported by Software Engineering & testing expert **Prof. Dr. Helmut Neukirchen** (University of Iceland)
  - CoE RAISE @ YouTube: https://www.youtube.com/channel/UCAdIZ-v6cWwGdapwYxdN7dg
  - Methology as one CoE RAISE outcome
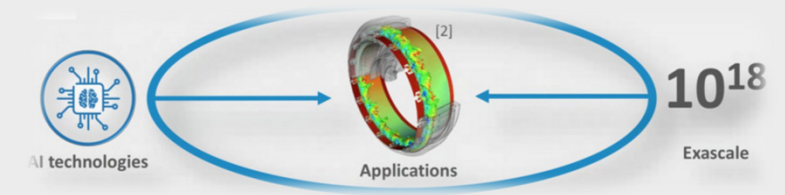
Book, M., **Riedel, M.**, **Neukirchen, H.,** Goetz, M.: **Facilitating Collaboration in High-Performance Computing Projects with an Interaction Room**, in conference proceedings of the 4th ACM SIGPLAN International Workshop on Software Engineering for Parallel Systems (SEPS 2017), October 22-27, 2017, Vancouver, Canada

CoE RAISE Unique AI Software Framework for Exascale

processing-intensive applications

- A — Compute-Intensive CoE RAISE Use Cases
- B — Data-Intensive CoE RAISE Use Cases
- C — Other Exascale HPC & AI Community Use Cases

Reference Architecture elements of CoE RAISE unique AI framework for Exascale HPC & AI Methods

- D — Secure Sheel Access (SSH) using batch submits to scale-up distributed training
- E — Interactive Jupyter notebooks with JupyterLab sharing of datasets and scripts for rapid DL model prototyping
- F — API to specify required models in ONNX format also enabling re-usability of existing AI models
- G — API to share and re-use of AI models of the MLFlow community platform, tools, data, and AI models
- H — Slim Facade Pattern to encapsulate concrete use case implementations
- I — Wrapper-scripts for using concrete scalability-proven AI frameworks
- J — HPC Systems library config & library availability checks

software infrastructure

- K — Basic Science & AI Libraries
- R — **NEW** Hyperparameter RAYtune Tuner
- L — TensorFlow & PyTorch Basic DL / ML libraries
- M — Horovod / PyTorch-DDG Distributed Deep learning

hardware infrastructure

- N — Modular HPC System prototype DEEP
- O — Modular HPC System JUWELS
- **NEW** Apptainer Container Environment
- P — **NEW** Apptainer Container Environment
- Q — BSC-CNS HPC System Mare Nostrum

✓ RQ1, RQ2, RQ4, RQ5

❖ Parts of the framework layout plan is to provide Kernels for Jupyter notebooks with correct version setups of modules for specific HPC Systems

✓ RQ3, RQ6 **NEW**

❖ Parts of the framework layout plan is to provide lightweight & abstract Python APIs building on ONNX enabling exchange with MLFlow, OpenML, ClearML, etc.

✓ RQ1, RQ2

❖ Parts of the framework layout plan is to provide a lightweight Python API that abstracts from low level versioning of AI packages (with proven scalability) and is harmonized with different available HPC system module versions
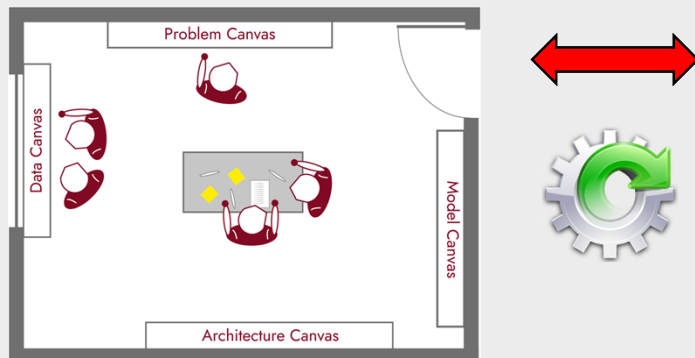
DRAFT

Continously Updating

Legend:

Tangible outputs of RAISE WP2 as part of the unique AI framework layout

✓ RQ6, RQ7

❖ Part of the framework layout plan is to provide containers in Apptainer with prepackaged datasets and required software stacks needed for AI models

✓ Interaction Rooms

- ✓ Update of Matrix
- ✓ Components relatively constant & common
- ✓ Methods change & new methods added (e.g., Transformer Models)

Use Case / Details	AE CAE	PINN	ANNs ANN	ANNs RBF-ANN	CNN U-Net	CNN RES NET	NO FNO	GNN MLPF	GNN GAT	RNN LSTM	RNN GRU	GAN WGAN	TF MViT	TF ViViT	TF Swin	SVM	RF
AI for turbulent boundary layers	X	X	X									X					
AI for wind farm layout optimization				X											X		
AI for data-driven models in reacting flows					X				X								
Smart models for next generation aircraft engine design					X				X								
AI for wetting hydrodynamics	X	X					X				X						
Event reconstruction and classification at the CERN HL-LHC use case								X									
Seismic imaging with remote sensing for energy applications	X	X			X	X				X	X				X	X	X
Detect-free metal additive manufacturing	X		X									X	X	X	X		
Sound Engineering	X		X														

drive. enable. innovate.