Heterogenous Data and Large Representation Models in Science WS

AI for science, science for AI

PANEL DISCUSSION

Jordi Inglada
French Space Agency (CNES, Toulouse), Center for Spatial Biosphere Studies (CESBIO, Toulouse)

Jonathan Gair
Max Planck Institute for Gravitational Physics - Albert Einstein Institute - (AEI Potsdam, Germany).

François Lanusse
(Polymathic AI)

Sylvain Caillou
(L2IT, IN2P3, CNRS/UT3)

Daniel Murnane
(NBI, Copenhagen University)

Anna Hallin
(Institute of Experimental Physics, Universität Hamburg)

David Roussel
Data scientist, Data Architect at Airbus Flight Tests

AI for science, science for AI

Foundation models in Science	30’
Heterogenous Data and Multimodal Representation Learning	30’
Inverse Problem - Likely hood free Simulation based approach	30’

Foundation models in Science

AI for science, science for AI

- What is a foundation model?
- Is a foundation model without a LLMs is still a foundation model ?
- What’s the point of foundation models ?
- Is leaving out any physic intuition lazy or smart ? (Symmetries …)
- What are the benefits / risks (ecological, automated science, reliability, etc)?
- Getting from simulation to real data ?
- What are the key considerations in preprocessing large datasets for use with large models?
- How do the data sources (human vs. instrument data vs simulated data) impact the design and application of foundation models in science or engineering?
- Considering the limitations of LLMs in science or engineering, such as hallucination and biases, How might Retrieval-Augmented Generation (RAG) address these challenges?

Heterogenous Data and Multimodal Representation Learning

AI for science, science for AI

- What can we say about the process and challenges of aligning latent spaces in multimodal learning?
- Question of mix of expert / modality dedicated models
- What are the architectural constraints when designing models for hetero/ multimodal learning, and how can these be overcome?
- How to mix Neural Networks encoding with Symbolic Engine ?

Inverse Problem - Likely hood free Simulation based approach

AI for science, science for AI

- What are the advantages and limits of using likelihood-free ML, simulation-based approaches in solving inverse problems?
- What do you see is the best tool to solve IP
  - How does causality give us tools for IPs?
  - How do generative models relate to solutions for IPs?
- IP can be treated as an unfolding problem, which mixes detector effects, simulation quality AND the possibility of new physics. How can we think about each of these sources of uncertainties?
- Active learning could be very useful in IP: Can we learn which parts of parameter space are the most interesting to simulate?
- Inference seems to be very expensive (e.g. reversible jump MCMC). What are some approaches for moving (amortizing) this cost into the training process?