## 01 Scholars in the loop? Bridging the gap between distinctive knowledge of small disciplines and training data for AI

**MATH PI:** Dr. Danah Tonne, Steinbuch Centre for Computing (SCC), Data Exploitation Methods (SCC-DEM)

**SEE PI:** Dr. Charlotte Debus, Steinbuch Centre for Computing (SCC), Junior Research Group Robust and Efficient AI (SCC-RAI), Germaine Götzelmann, Steinbuch Centre for Computing (SCC), Data Exploitation Methods (SCC-DEM)

Department(s): Informatics (Computer Science)

Type of position: 75% FTE, E13 TV-L

Recently, critical assessments of AI have pointed out that machine learning approaches are acting as a magnifier glass on social biases, revealing critical blind spots and imbalances in the underlying data labels. Small research disciplines, on the other hand, are especially prone to deal with edge cases and filling in blind spots regarding human culture and knowledge. They are important correction factors to address and ultimately adjust cultural biases, skewed views and information gaps in a postcolonial world as well as providing well-needed scholarly information about areas prone to misinformation, fake news and pseudoscience activities. To be able to substantiate their findings by the evaluation of large data collections new methodical approaches like Artificial Intelligence (AI) have been introduced by the research field Digital Humanities.

In small research disciplines, scholarly digital annotation has gained track in recent years, capturing fast amounts of unique knowledge in tedious manual or semiautomatic processes. But the exchange formats and publications of results mainly address human knowledge sharing and cannot be directly transformed into training data. Machine learning projects quite often have to start from scratch in regards to data labeling, making their training data potentially shallow and sparse.

This makes the research process both unsustainable and less likely to succeed. State-of-the-art solutions to address those data labelling challenges are involving activation of a broader community with diversified world views in form of crowd sourcing concepts as well as interactive labeling (IL) and ‘human in the loop’ (HITL) approaches. However, one important puzzle piece for resolving the stated challenges is facilitating the in-depth knowledge already produced by research projects for data labeling purposes.

The goal of this dissertation project is to identify the main obstacles and to develop solutions to facilitate the research data flow between more ‘traditional’ projects and much needed high quality training data. The focus lies on selected use cases from small (humanities) research fields with close connection to domain experts, e.g. philological, medieval, or religious studies and the possibility to adequately survey their needs and constraints.

**Requirements for this position:**

- Solid background in either Computer Science or related fields like Mathematics, Physics, Electrical Engineering or Digital Humanities
- Software development and basic programming language, e.g. Python, C/C++
- Prior experiences in data science and machine learning, and corresponding software frameworks (e.g. PyTorch) is advantageous
- High interest in interdisciplinary research

## 02 Advanced Parallelization Techniques for data-driven Uncertainty Quantification

**MATH PI:** TT-Prof. Dr. Sebastian Krumscheid, Steinbuch Centre for Computing (SCC), Junior Research Group Uncertainty Quantification (SCC-UQ) & Institute for Applied and Numerical Mathematics (IANM)

**SEE PI:** Dr. Linus Seelinger, YIG Prep Pro Fellow

Department(s): Mathematics

Type of position: 75% FTE, E13 TV-L

In the field of natural sciences, the integration of Bayesian inference and Uncertainty Quantification (UQ) shows great potential in improving our understanding of complex systems and enhancing the reliability of scientific predictions. However, UQ often faces challenges as it requires a large number of simulation runs, which necessitates advanced UQ algorithms and high-performance computing (HPC). Unfortunately, intricate data dependencies and technical obstacles hinder parallel computing for inverse (i.e., data-driven) UQ. To address this problem, UM-Bridge introduces a novel software architecture to UQ, which solves these technical challenges and lays the foundation for developing sophisticated parallelization strategies at a relatively low compute cost. Additionally, it allows for the seamless application of these strategies to various scientific models.

Consequently, UM-Bridge enables comprehensive study of advanced parallelization techniques for UQ with application to realistic large-scale models. The objectives of this project are twofold. Firstly, we aim to develop new parallelization methods for inverse UQ that will enable efficient Bayesian inversion on modern HPC systems. We will consider approaches such as hierarchical Markov chain Monte Carlo (MCMC) methods combined with sample pooling techniques obtained from fast, yet approximate surrogate models and parallel, multiple-try MCMC proposal methods. Secondly, we will adapt these new methods for use in Earth System Sciences, specifically for data-driven modeling in compute-intensive Earth system models. This will be done collaboratively in the context of Simulation and Data Life Cycle Labs at the Steinbuch Centre for Computing.

**Requirements for this position:**

- Solid knowledge of probability theory and statistics, as well as numerical analysis and
- mathematical modelling.
- Experience with uncertainty quantification techniques is advantageous.
- Good programming skills, including experience with parallel programming.

## 03 Mathematical Foundations of Bayesian Neural Networks

**MATH PI:** TT-Prof. Dr. Sebastian Krumscheid, Steinbuch Centre for Computing (SCC), Junior Research Group Uncertainty Quantification (SCC-UQ) & Institute for Applied and Numerical Mathematics (IANM)

**SEE PI:** Dr. Charlotte Debus, Steinbuch Centre for Computing (SCC), Junior Research Group Robust and Efficient AI (SCC-RAI)

Department(s): Mathematics or Informatics (Computer Science)

Type of position: 75% FTE, TV-L E13

With the increasing application of machine learning (ML) methods, the robustness of such data-driven methods becomes a central aspect. Modern ML models must not only be able to deliver unprecedented prediction accuracy but are also required to deliver an estimate of the uncertainty of that prediction. Assessing the possible error margin on a prediction is essential in applying ML models to critical infrastructures, such as electricity resource planning from renewable energy sources.

For deep learning (DL), Bayesian Neural Networks (BNN) provide a promising approach to quantifying the inherent data uncertainty and that of the ML model itself, which arises from the optimization process. However, currently available theoretical approaches and their practical implementations of BNNs need to be improved, particularly regarding computational efficiency and the accurate description of uncertainties.

Addressing these shortcomings is the aim of this doctoral project. Specifically, the overarching goal is to expand the mathematical theory behind variational inference underpinning Bayesian neural networks to provide accurate and computationally efficient model uncertainties. Situated at the intersection of mathematics and computer science, this doctoral project combines statistical methods and Bayesian theory with state-of-the-art deep learning approaches. The methods developed in the context of this project will be evaluated on the use-case of predicting photovoltaic electricity generation, which is relevant for optimal scheduling of electricity allocation.

**Requirements for this position:**

- A degree (M.Sc. or equivalent) in computer science, mathematics or another related field, e.g. physics or engineering.
- Basic knowledge of and initial experience with machine learning methods, preferably in Deep Learning
- Basic knowledge of applied mathematics, including numerical analysis, statistics, and Bayesian inference
- Solid programming skills in any scientific programming language, such as Python, C/C++

## 04 Replacement of physical PSC simulation by machine learninq in an Earth system model

**MATH PI:** Dr. Ole Kirner, Steinbuch Centre for Computing (SCC), Scientific Computing & Mathematics (SCC-SCM)

**SEE PI:** Dr. Jörg Meyer, Steinbuch Centre for Computing (SCC), Data Analytics, Access and Applications (SCC-D3A)

Department(s): Informatics (Computer Science) or Physics

Type of position: 75% FTE, E13 TV-L

Polar stratospheric clouds (PSCs) exist in winter in the lower/middle atmosphere and are responsible for ozone depletion in the polar spring and the resulting ozone hole.

The goal of the doctoral research is to show that the physical simulation of PSCs within an earth system model can be replaced by an Al model. It will be investigated and evaluated if this enables a reqlistic simulation of the PSCs and thus improves the performance of the PSC modul as part of the earth system model ICON-ART.

Tasks of the thesis include:

- Performance analysis of the ICON-ART model code at different resolutions on High Performance Computing (HPC) systems
- Creation of a concept for replacing the PSC simulation with a suitable Al model (such as Transformer, LSTM, CNN) including the identification of suitable features dimension reduction, and hyperparameter tuning
- Implementation and evaluation of the procedure and investigation of suitable metrics
- Evaluation of the Al model integrated into ICON-ART (including parallelization)

**Requirements for this position:**

- Completed studies (master) in computer science, mathematics or physics
- Programming skills (e.g. Fortran, C++, Python)
- Ability to work and publish in a targeted and scientific manner.
- Good communication and presentation skills and willingness and ability to work a team
- Good writing and oral communication skills in English

**Optional requirements:**

- Knowledge of current deep learning frameworks (e.g. PyTorch or Tensorflow)
- Experience in working with climate/earth system models

## 06 Trainability of data driven quantum models

**MATH PI:** Dr. Leonid Chaichenets, Steinbuch Centre for Computing (SCC), Scientific Computing & Mathematics (SCC-SCM)

**SEE PI:** Dr.-Ing. Eileen Kühn, Steinbuch Centre for Computing (SCC), Data Analytics, Access and Applications (SCC-D3A)

Department(s): Informatics (Computer Science) or Mathematics

Type of position: 75% FTE, TV-L E13

In the field of quantum machine learning many ansätze for designing quantum circuits that make up a trainable quantum model are influenced by heuristics but also by current challenges of quantum computers such as noise and size. One influential paper presents a collection of potential hardware-efficient building blocks for quantum circuits analyzing those regarding trainability and their efficiency to benefit from the available problem space of a quantum computer. Typical scientific questions thus involve deciding for one of the building blocks, and the number of repetitions required to solve the underlying problem. To answer these questions several experiments are required. This contrasts with current developments in geometric quantum machine learning, exploiting symmetries in data to be encoded as part of the quantum circuit. Based on these symmetries that effectively limit the search space, the learning process can become more efficient in terms of computing resources, but also in terms of time. Further questions about the choice of ansatz or number of repetitions can be omitted in the best case.

In this project, varying datasets shall be reviewed with regard to their features and correlations to identify a new set of building blocks to minimize the number of experiments required while respecting the challenges of today’s quantum computers.

**Requirements for this position:**

- Knowledge of machine learning frameworks (e.g. PyTorch, TensorFlow)
- Good programming skills (e.g. Python)