HiDA HELMHOLTZ Information & Data Science Academy

Visualisation of Numerical Simulation Data for Multiphase Flows at Industrial Scale.

Dr. Fabian Schlegel - Helmholz Zentrum Dresden Rossendorf

Dr. Fabian Schlegel - Helmholz Zentrum Dresden Rossendorf

The group is interested in developing new models and methods to simulate multiphase flows at industrial scales. Since a few years a morphology-adaptive two-fluid model for multiphase flows (OpenFOAM-Hybrid) is developed by the group for the open-source library OpenFOAM. The model is based on a 4-field approach and distinguishes into continuous phases that share a resolved interface and dispersed phases with statistically modelled interfaces.

https://hzdr.de/openfoam-hybrid

r.meller@hzdr.de

b.krull@hzdr.de

What is the project's research question?

Visualisation techniques with Paraview and Nvidia Optix for OpenFOAM Hybrid simulations, which combines resolved and statistically modelled interfaces.

What data will be worked on?

Numerical Simulation Data written by OpenFOAM: Velocity and pressure field information. Information about position of resolved interfaces. Probability information for modelled interfaces. The available simulation data includes, e.g., a single tray of a distillation column, and a cyclone separator.

What tasks will this project involve?

Load the required data in Paraview. Enable Nvidia Optix RayTracing Plugin. Define proper set of material parameters for water and air. Find a creative idea to visualize the probability information for the statistically resolved interfaces. Export images and videos.

What makes this project interesting to work on?

The project introduces to numerical methods for multiphase flows. It allows getting insight into one of the largest open-source projects for numerical simulations (OpenFOAM) and one of the most used visualisation tools (Paraview). The used data represents real world applications and problems from chemical and process engineering industry (distillation column, cyclone separator). Furthermore, state-of-the-art computer vision and imaging techniques will be used (NVIDA Optix on NVIDIA Quattro RTX 4000) to generate an illustrative showcase video.

What is the expected outcome?

Contribution to software, Contribution to social media (requires permission by HZDR)

What infrastructure, programs and tools will be used? Can they be used remotely?

OpenFOAM, Paraview, NVIDIA Optix

Is the project open source?

Yes

What skills are necessary for this project?

High-performance computing, Computer vision and image processing/analysis, Python.

Interested candidates should be at Bachelor or Master level.

Training machine learning force fields (MLFFs) for metal hydrogen systems.

Dr. Kai Sellschopp – Helmholtz Zentrum Hereon

Dr. Kai Sellschopp – Helmholtz Zentrum Hereon

The "Materials Design" department in the Institute for Hydrogen Technology at Hereon develops innovative functional materials based on metal hydrides and light metal hydride composites for applications in energy technology, such as for hydrogen storage. We combine computational methods with experimental techniques, to obtain a full understanding of all relevant processes in these materials from the atomic to the macro scale.

https://www.hereon.de/institutes/hydrogen_technology/materials_design/index.php.en

kai.sellschopp@hereon.de

What is the project's research question?

Which ML method is the best, i.e., most efficient, accurate, and practical, for training force fields for metal-hydrogen systems?

What data will be worked on?

The aim of a force field is to predict the energy, forces, and/or stresses of an atomic configuration. You will work on data that we obtain from density-functional theory (DFT) calculations, which are reasonably accurate for this task, but also computationally demanding. Depending on your background, additional sampling of data can be part of the project, or we will provide data sampled with molecular dynamics simulations based on DFT. Therefore, prior knowledge of materials modelling techniques is not required, but very basic materials science knowledge would be helpful for understanding the data.

What tasks will this project involve?

Analyse the training data, get accustomed to materials modelling data. Train ML-FFs for a metal-hydride system using different methods (e.g., kernel-based, or artificial neural network) and compare them. Closely work with people from the group responsible for creating the training data and for testing the quality of the ML-FFs in production. Present your results in group meetings and contribute to writing a manuscript.

What makes this project interesting to work on?

With this project, you will be integrated into the Materials Design department of the Institute of Hydrogen Technology at Hereon in Geesthacht. This gives you the chance to learn more about modern materials research, not only computationally, but also experimentally. Furthermore, your contribution will help us to develop hydrogen storage materials, which play an important role in a sustainable, environmentally friendly economy. Even though we do have some ideas on how to approach this project, you are very welcome to contribute and explore your own ideas as well. The scope of the project can be adjusted depending on your background and interests.

What is the expected outcome?

Co-authorship to research paper.

What infrastructure, programs and tools will be used? Can they be used remotely?

You will have access to our modern in-house high-performance-computing (HPC) cluster for training and testing the models. The datasets are created from materials simulations using a combination of the Vienna Ab-initio Simulation Package (VASP) and python scripts based on the Atomic Simulation Environment (ASE) package. For the ML part, you are free to use your code of choice. All these resources can be accessed remotely, if an in-person stay is not possible.

Is the project open source?

Will be made available with publication.

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Machine learning, Deep learning, High-performance computing, Python.

Interested candidates should be at Master, PhD, or Postdoc Level.

Improving the performance of the atomic-physics code atoMEC.

Dr. Attila Cangi – CASUS Center for Advanced Systems Understanding

Dr. Attila Cangi – CASUS Center for Advanced Systems Understanding

The "Matter under Extreme Conditions" department at the Center for Advanced Systems Understanding (CASUS) at Helmholtz-Zentrum Dresden-Rossendorf focuses on advancing physics-informed and data-driven numerical simulation techniques to achieve multi-scale modelling. Our goal is to connect fundamental quantum processes on the microscopic level with observables in materials science on larger scales. To accomplish this, we employ a combination of first-principles methods such as density functional theory, efficient reduced methods like average-atom models and quantum hydrodynamics, and large-scale molecular dynamics simulations. We integrate these techniques into data-driven workflows with machine-learning methods and develop novel machine-learning techniques inspired by quantum computing.

https://www.casus.science/research/matter-under-extreme-conditions/

a.cangi@hzdr.de

What is the project's research question?

The main question to be addressed is how to improve the performance — namely, the speed and the memory requirements — of the atoMEC code. The code should be extensively tested, particularly for edge cases, to ensure that any performance gains do not compromise accuracy or reliability.

What data will be worked on?

Mostly, the student will work with data generated by the code itself. AtoMEC is an average-atom code for studying materials under extreme conditions, meaning high material densities and/or temperatures. The data of interest is a mixture of physical observables, such as pressure or conductivity, and intermediate data, such as electronic wave functions. This intermediate data is often very large, and one of the aims of the project is to process and store this data more efficiently. Depending on how the project develops, the student may also work with data from density-functional theory codes, to benchmark the atoMEC results.

What tasks will this project involve?

Performance: changing the default numerical accuracy from double to single precision, benchmarking the change in accuracy as well as typical speed improvements. Performance: matrix diagonalization is the main computational bottleneck. Can the current parallelization be improved? Is it better to solve the PDEs in a different way than via matrix diagonalization? Performance/Usability: Under certain conditions (high temperatures and low densities), calculations are prohibitively expensive. Is there a way around this, for example, using a different method than matrix diagonalization? Usability: There are many parameters in each calculation, which most but not always have default values. Is there way

to pick suitable defaults for the parameters which the user must currently specify? Can we choose a better set of defaults than what is currently implemented?

What makes this project interesting to work on?

From a scientific perspective, average-atom codes are widely used to model inertial confinement fusion. There have recently been many exciting breakthroughs in inertial confinement fusion which makes nuclear fusion an increasingly promising method for the generation of abundant clean energy. The atoMEC code is still young and currently maintained by a small team of developers at CASUS, however, it is unique in that (as far as we know) it is the only open-source average-atom code under active development. The student therefore can make a significant impact on the atoMEC code, and as a result, the nuclear-fusion community more generally.

What is the expected outcome?

Contribution to software,

What infrastructure, programs and tools will be used? Can they be used remotely?

HPC cluster at HZDR, Python and its main scientific libraries. Tools: Git. All these can be used remotely. However, we highly encourage the student to visit us in person (at least for part of the project duration), to get the most from scientific and social interactions.

Is the project open source?

Yes

What skills are necessary for this project?

Scientific computation, data mining, Software development, Python

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

Synthesis of biologically stained images based on 3D greyscale tomography data using generative adversarial networks.

Dr. Berit Zeller-Plumhoff – Helmholtz Zentrum Hereon.

Dr. Berit Zeller-Plumhoff – Helmholtz Zentrum Hereon.

The department for Imaging and Data Science focusses on the use of high-resolution 3D micro computed tomography to image the degradation and osseointegration of biodegradable magnesium implants in bone. Advanced image processing techniques and deep learning are used to attain quantitative parameters from image data. Moreover, new techniques are developed to enable the correlation and/or synthesis of different characterization techniques.

https://www.hereon.de/institutes/metallic_biomaterials/imaging_and_data_science/index.php.en

berit.zeller-plumhoff@hereon.de

What is the project's research question?

Can we use GANs to predict the histological staining of tissue based on X-ray tomography data to extract biologically relevant information?

What data will be worked on?

We have a large number (>100) of 3D image volumes from micro computed tomography of rat bone samples implanted with four different implant types and corresponding 2D histology images. The images are registered, so that we have the exact slice in the (greyscale) tomographic volume which represents the (stained) 2D histology. This data will be used for training, testing and validation of the GAN. The aim is to train the GAN such that it can predict the staining for the whole 3D tomography volume. A prototype of this GAN trained on a limited number of samples is available.

What tasks will this project involve?

All data will be prepared prior to the start of the project. Understand the current prototype of GAN trained with 8 datasets. Normalize and pre-process (RGB to greyscale) the existing histology data. Check current hyperparameter setting and train/test/validate the existing GAN with the available data

What makes this project interesting to work on?

Both the data that is being used in this project, as well as the methodology itself are highly relevant research topics now. The image data used on this project stems from biodegradable magnesium implants in rat bone, as well as titanium and PEEK implants which were used as reference materials. Magnesium implants represent a powerful alternative to the more traditional implant materials. However, in performing animal experiments, we can only get a limited number of histological sections from any one sample, due to the destructive processing methods. By synthesizing a biologically

meaningful 3D histology stack, this major limitation will be lifted. Ultimately, this will be of significance also to other materials and studies in the biomedical field.

Additionally, to learn about the use of deep learning and GANs on real-life data is an important addition to otherwise artificial datasets, as the limitations and challenges of both the methodology and the underlying data become apparent.

What is the expected outcome?

Co-authorship to research paper,

What infrastructure, programs and tools will be used? Can they be used remotely?

All image data is currently stored on the storage system at Deutsches Elektronensynchrotron (DESY), in Hamburg, where the tomography data was obtained. The data will be processed on the DESY Maxwell computing cluster, where the required hardware and software resources are available. As such, the student will become familiar with Python, PyTorch and Slurm jobs. Additionally, other image processing software (Fiji/ImageJ, Avizo) can be used. The software and cluster can be used remotely once access has been granted. If on site, the student can participate in an imaging experiment, to gain additional understanding on how the data they work with was obtained.

Is the project open source?

The data is available upon request as it's too large to be placed on conventional repositories.

What skills are necessary for this project?

Machine learning, Deep learning, Computer vision and image processing/analysis, Python.

Interested candidates should be at Master, PhD, or Postdoc Level.

Making sense of real-world epidemiological data.

Dr. Stefanie Castell – Helmholtz Centre for Infection Research

Dr. Stefanie Castell – Helmholtz Centre for Infection Research

The team "Climate, Cohorts and PIA" as part of the Department for Epidemiology at the HZI focuses its research on digital epidemiology and real-world data from methodologically sound population-based cohorts. We develop an eResearch System as product owners ("Prospective Monitoring and Management - App", PIA: www.info-pia.de) and take part in the largest epidemiological study Germany has ever conducted, the NAKO Gesundheitsstudie (www.nako.de). We built up a sub cohort within NAKO for intensified digital syndromic surveillance of acute infectious diseases.

https://www.helmholtz-hzi.de/en/research/forschungsprojekte/view/projekt/detail/zifco/

stefanie.castell@helmholtz-hzi.de

What is the project's research question?

What are the pitfalls, challenges and solutions of processing and analysing real world epidemiological data?

What data will be worked on?

We provide epidemiological overservational data involving longitudinal data collection (e.g., digital reporting of symptoms of acute respiratory infections) and if available complex laboratory data from immune cells and plasma of our study participants, Due to their longitudinal nature, the multitude of questionnaires, and the real-world missingness patterns, the data are complex and sometimes challenging to disentangle for sound scientific use.

What tasks will this project involve?

The student will learn about epidemiological raw data, will process, and prepare data and conduct quality assurance and collection evaluation that provides stakeholders with the information they need to improve data collection in the future. If possible, the student can conduct time-to-event analyses for factors potentially linked to infection susceptibility. If interested the student can support the software development of PIA based on scrum:

www.info-pia.de

https://github.com/hzi-braunschweig/pia-system

https://gitlab.com/pia-eresearch-system/pia

What makes this project interesting to work on?

We provide real world observational data on a hot topic, i.e., respiratory infections. The projects are suited to connect to observational epidemiology and encounter the challenge we deal with in epidemiology in that we analyse messy data with methodological scientific rigour. The supervision by Stefanie Castell, a medical doctor and epidemiologist with

knowledge on software development as product owner, focuses on the meta-level, and is not suited for someone who needs support with R. Our team loves digital tools and has a strong and long-standing expertise in epidemiological cohort studies. We emphasise participant involvement and conduct citizen science projects; hence, we offer a link between down to earth public health-oriented science, complex data and software development.

What is the expected outcome?

Co-authorship to research paper, Contribution to software.

What infrastructure, programs and tools will be used? Can they be used remotely?

The student can work remotely. We use Zoom for meetings; jira, confluence and GitLab to structure our work; Rocket chat and Outlook to connect throughout the day, for data analysis, we work with R or Stata. If the student is present in the region in Germany (Brunswick/Hanover), he or she can visit our study centre to deepen his or her understanding of data collection processes.

Is the project open source?

The data are sensitive (medical research data) and are not open source. The software PIA, however, is free and open source.

What skills are necessary for this project?

Data analytics, statistics, R, Software development is optional, not necessary.

Interested candidates should be at Master, PhD, or Postdoc level.

Development and automation of a unique incubation system to simultaneously measure N2 as well as N2O and CO2 fluxes from plant-soil mesocosms.

Dr. Irina Yankelzon - Institut für Meteorologie und Klimaforschung Atmosphärische Umweltforschung

Dr. Irina Yankelzon - Institut für Meteorologie und Klimaforschung Atmosphärische Umweltforschung

Our division is focused on understanding the impacts of human activity on the biogeochemical cycles of carbon and nitrogen and the exchange processes between ecosystems, the atmosphere, and the hydrosphere. We use a combination of field studies, observational data, and process modelling to study these impacts at multiple scales.

https://www.imk-ifu.kit.edu/

irina.yankelzon@kit.edu

What is the project's research question?

Measuring plants effect on denitrification using innovative techniques.

What data will be worked on?

Our exchange student will be working on data collected from our new and special plant-soil mesocosm incubation system, which allows growth of plants in a N2-free atmosphere. The system continuously measures gas concentrations in the incubation vessels while they are being flushed with a He/O2 gas mixture, resulting in a large amount of data that needs to be converted into flux rates. This data will be used to study the effect of plants on denitrification processes and specifically to quantify N2, N2O, and CO2 emissions simultaneously.

What tasks will this project involve?

The main task is to develop software or script to process and analyse the large amounts of data generated by the mesocosm incubation system. While additional tasks are also possible, such as. Develop a database or interface to store, manage, and visualize the data collected from the mesocosm incubation system. Participating in the design, implementation and testing of new features for the incubation system. Assisting in the preparation of manuscript for publication in scientific journal

What makes this project interesting to work on?

This project offers a unique opportunity to work on a cutting-edge plant-soil mesocosm incubation system that allows for the growth of plants in a N2-free atmosphere. This system provides new insights into the denitrification process, which is important due to its environmental impact.

Automating the flux calculations will make the system even more powerful and highly precise. This is because the automation process will allow for more efficient and accurate data analysis, making the system more sensitive to slight changes in N gaseous emissions.

This system is one of its kind currently and automating the flux calculations will make it even more unique and valuable for understanding the denitrification mechanics and its environmental impact.

What is the expected outcome?

Co-authorship to research paper, Contribution to software

What infrastructure, programs and tools will be used? Can they be used remotely?

The main programming languages that will be used for this task are Python or R. Remote access to the data and tools is possible, but we highly recommend and encourage the exchange student to come to our lab and work with us together. This will not only provide the opportunity to see the system in person, but also allow for more efficient collaboration and problem-solving.

Is the project open source?

Yes

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Python, R

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

Context-aware self-supervised learning for reliable feature extraction of X-ray imaging.

Nico Hoffmann - Helmholtz-Zentrum Dresden-Rossendorf

Nico Hoffmann - Helmholtz-Zentrum Dresden-Rossendorf

We are researching data-driven methods for analysis of X-ray imaging data acquired at large-scale light sources like DESY or EuropeanXFEL. Loss of information during image acquisition renders the corresponding inverse imaging problem ill-posed (phase problem). We therefore deploy generative models (i.e., normalising flows and stable diffusion) that are solving these inverse imaging problems by recovering multiple solutions that correspond to an observation.

http://photon-ai-research.github.io/

n.hoffmann@hzdr.de

What is the project's research question?

Is self-supervised learning producing reliable and robust features for image analysis?

What data will be worked on?

You will be working on a large-scale COCO image dataset as well as 2d X-ray imaging data provided by HZDR/DESY.

What tasks will this project involve?

You will be implementing PyTorch datasets for COCO and X-ray imaging data. Then, you will be integrating data2vec into common workflows for image analysis based on two downstream tasks: 1) semantic segmentation (COCO) and 2) regression (X-ray data). Eventually, the performance of representations learnt by data2vec for solving these downstream tasks will be systematically assessed with respect to perturbations of the data.

What makes this project interesting to work on?

You will be experimenting with DataVec 2.0 of Facebook AI research, a visual feature learning scheme implemented in Python and PyTorch that is trained in a self-supervised fashion. Supervisory feedback for context-awareness of the feature encoder will be provided by instance-wise transformations such as 1) masking, 2) rotation, 3) intensity modulations. This is a state-of-the-art method that is used for very different applications and modalities ranging from images, speech to text. However, we will be focusing on images due to constraints in project time.

What is the expected outcome?

Co-authorship to research paper, Contribution to software,

What infrastructure, programs and tools will be used? Can they be used remotely?

You will be using HZDR's GPU cluster HEMERA where you get access to a partition with 7 nodes with 28 V100 GPUs. All coding will be done in Python and Pytorch. Distributed training, if appropriate, will be implemented via Horovod. Remote access is possible.

Is the project open source?

Yes

What skills are necessary for this project?

Deep learning, Computer vision and image processing/analysis, Python. PyTorch knowledge would be very helpful.

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

Exploring potential impacts of extreme weather events (e.g., heat waves and heat stress) on human health outcomes.

Dr. Monica Ionita-Scholz – Alfred Wegener Institute

Dr. Monica Ionita-Scholz – Alfred Wegener Institute

Among my group's major research interests is the intensity, frequency, severity, and duration of extreme events such as droughts, floods, heat and cold waves and their influence on socio-economic systems and public health. Studying extremes is an important task because they often have immediate impacts on society causing widespread adverse health outcomes and infrastructure destruction. We use statistical and climate modelling, data analysis and machine learning tools to understand which environmental extremes have the major negative impacts on human systems, and which protective factors (such as an extent of green residential spaces) could potentially moderate the effects of extreme weather events.

https://www.awi.de/en/about-us/service/expert-database/translate-to-english-monica-ionita.html

Monica.Ionita@awi.de

What is the project's research question?

One of the following research questions could be explored during the project: What are the associations between extreme weather events and adverse (negative) health outcomes? Which events could have the highest impact on health outcomes from historical perspective? Which protective factors could moderate the negative effects of extreme events?

What data will be worked on?

European and Middle East observational and modeling temperature time series (e.g. E-OBS (spatial resolution 0.1°x0.1°), CRU TS4.04c (spatial resolution: 0.5°x0.5°), TerraClimate (spatial resolution: 0.5°x0.5°), ERA5 (spatial resolution: 0.1° x 0.1°)), as well as daily meteorological and health indicators from different federal agencies in Germany and Israel.

What tasks will this project involve?

Statistical data analysis with the potential to build predictive models. Specifically: correlation and association analysis, empirical orthogonal function analysis, canonical correlation analysis, machine learning methods (e.g., transformers, convolutional layers, variational autoencoders)

What makes this project interesting to work on?

From a broader perspective, this interdisciplinary project gives participant(s) an opportunity to explore the outcomes of environmental and medical studies, thus linking climate change and societal issues. Evidence from epidemiological studies show that extreme heat and cold events, droughts, wildfires as well as floods in Europe have negative impacts on population

health. Furthermore, climate data reveals increased intensity, frequency, severity, and duration of such extreme events, showing that societies are becoming more vulnerable to the increasing climate risks. The project helps to perform an in-depth association and causal analysis of both climate and health data at a country and international levels. This provides participant(s) with the possibility to extend knowledge and methodological background from different research fields.

What is the expected outcome?

Co-authorship to research paper, Contribution to software. An in-depth statistical analysis of the relationship between different type of extreme events (e.g., heat waves, heat stress, cold spells), and the variability and trends of health indicators at country level.

What infrastructure, programs and tools will be used? Can they be used remotely?

External supercomputer (dkrz, julich) as well as AWI computer infrastructure. Python, R, GitLab/GitHub, Jupyter notebooks

Is the project open source?

Climate data is open source while medical data needs special access.

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Machine learning, High-performance computing, Python, R,

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

When Geomorphology meets Quantum Machine learning: a QSVM model for surface flows classification.

Dr. Hui Tang – GFZ Helmholtz Zentrum Potsdam.

Dr. Hui Tang – GFZ Helmholtz Zentrum Potsdam.

Our research group studies hazards and related surface processes across various environments, from mountain regions to coastal areas and even deep oceans, over different time scales. Our work has wide topics ranging from earthquakes and tsunamis, storms and hurricanes, landslides, and debris flow, to floods and paleo-flood. We use various tools and methods, including field surveys, remote sensing, environmental seismology methods, processes-based modelling, and data science methods, including machine learning to understand the physical processes behind all these natural hazards.

https://www.gfz-potsdam.de/en/section/earth-surface-process-modelling/overview/

htang@gfz-potsdam.de

What is the project's research question?

Landslides, debris flow, hyper concentrated flow, and floods are among the most dangerous natural hazards worldwide. One of the fundamental tasks for geomorphologists is to classify and identify which kinds of processes they observe in the field. The task is more challenging than it sounds, especially considering high-damage processes like debris flows and landslides. Meanwhile, multiple dimensionless numbers (e.g., Reynolds number and Einstein number) based on first-principal physics have been widely used to describe these natural flows. When we use these dimensionless numbers and datasets to classify the flow, we automatically face a long-standing challenge in machine learning (maybe one of the biggest challenges in data science): the curse of dimensionality. One of the expertise for quantum machine learning methods (e.g., QSVM) is to deal with such a high-dimensional dataset. Therefore, we ask the central research question in geomorphology and machine learning: can we objectively define the type of natural flows using the dimensionless number and Quantum machine learning methods?

What data will be worked on?

The applicant will work on a high-dimensional dataset for different surface flows. They are usually dimensionless in a physical sense based on the theoretical derivation and well-defined with uncertainties. Please see the following reference for details: Du, J., Zhou, G. G., Tang, H., Turowski, J. M., & Cui, K. F. E. (2023). Classification of Stream, Hyperconcentrated, and Debris Flow Using Dimensional Analysis and Machine Learning. Water Resources Research, 59, e2022WR033242. https://doi.org/10.1029/2022W

What tasks will this project involve?

The applicant will develop a high-dimensional quantum support vector machine model to classify the type of natural flow in the dataset. Based on our previous study, the applicant needs to design a simple, supervised quantum machine learning algorithm (e.g., QSVM). If

time allows, we will investigate the possibility of designing and training an unsupervised quantum machine learning algorithm such as the Quantum K-means model.

What makes this project interesting to work on?

As part of our effort to build Digital Twins in catchment scale, the applicant will get the chance to learn not only about basic geophysics and geomorphology knowledge about hazards but also about state-of-the-art machine learning methods (dimensionless learning and physics-informed machine learning) as well as quantum computations. The successful applicants also will have a chance to interact with a large data science community around Berlin and Potsdam. In addition to the six weeks of funding provided by HIDA, our group (Hazards and Surface Processes) will be happy to provide matched funding for the successful applicants (i.e., another six weeks of funding) based on performance to make sure the applicant has the chance to integrate fully with the research institute and group members.

What is the expected outcome?

Co-authorship to research paper, Contribution to software.

What infrastructure, programs and tools will be used? Can they be used remotely?

We will provide chances to use HPC at GFZ or Quantum computer at IBM, depending on the availability of the computer. He or she will mainly use IBM Quantum tools Qiskit and Jupiter notebook to develop the code. In principle, all infrastructure, programs, and tools can be developed and used remotely. But we have a preference to work on-site, as we believe learning how to interact with other members in research group is a crucial part of training a successful researcher.

Is the project open source?

Yes

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Machine learning, Deep learning, Software engineering, Python.

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

A Data Scraping Study of Polymers of Intrinsic Microporosity (PIM) Membranes: Exploring the Effect of the Monomer Ration on the Performance of PIM Membranes.

Dr. Volkan Filiz – Helmholtz Zentrum Hereon.

Dr. Volkan Filiz – Helmholtz Zentrum Hereon.

Clean drinking water for all people—this is one of the greatest future challenges. Innovative membrane methods can help with both water supply and wastewater disposal, thereby reducing the burden on the environment. We develop high-performance membranes that filter micropollutants and heavy metals from the water. We also design novel microporous polymers, for example, to desalinate seawater or to remove climate-damaging gases from the air. This work is supported by comprehensive digital modelling.

https://hereon.de/institutes/membrane_research/microporous_polymers/index.php.en

volkan.filiz@hereon.de

What is the project's research question?

Can we correlate the chemical structure and composition of PIM molecules to the performance of the membranes prepared from these molecules?

What data will be worked on?

Polymers of Intrinsic Microporosity (PIMs) are a class of materials that are characterized by a high density of small, uniformly distributed pores within their structure. These pores are typically in the range of around 1 nm in diameter, and they give PIMs unique properties such as high surface area (700−1000 m2 g−1), high gas adsorption capacity, and high chemical stability. PIMs are made by polymerizing monomers that contain different structures, such as cyclic ethers or cyclic esters, which create the micropores within the polymer structure. PIMs have a wide range of potential applications, including gas storage, catalysis, and separation science. Depending on the composition of the PIM molecule and the used monomers, properties such as permeability can vary tremendously. However, a good overview over existing PIM films and their performances is missing in the moment.

In this project, data scrapping from research websites will be used to generate an up-to-date database containing PIM film composition (different monomers and their ratio), measurement parameters (e.g. temperature), reference and performance parameters (permeabilities of different gases such as O2, N2, CO2, propene and methane). The data obtained will be of importance for material scientists as well as for further machine learning applications. Unfortunately, there is no very good overview of the gas separation properties coupled to the chemical structure of the membrane, nowadays. This is not only interesting for the scientific community, but also for an application of these membranes in the industry.

What tasks will this project involve?

Identifying the most useful websites for the project. Writing of the code for data scrapping. Preparation of the database containing chemical structure and composition of PIM membranes compared to their performances.

What makes this project interesting to work on?

The Hereon-Institute of Membrane Research provides an international and interdisciplinary work environment, which will be great experience for the student working on the project. Additionally, he or she can learn to apply data scrapping methods in a scientific project as well as gather new knowledge about material science. Therefore, the project is interesting for young data scientists, which want to apply their knowledge in material science. Also, the student can learn more about membranes used for various applications.

What is the expected outcome?

Co-authorship to research paper

What infrastructure, programs and tools will be used? Can they be used remotely?

Python will be used as coding language. Additionally, licenses for many scientific journals are available. A PC or laptop as well as office space will be provided by the Hereon. Additionally, we can arrange accommodation in one of our guesthouses. The work can be done remotely. However, onsite work is preferred.

Is the project open source?

No

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Databases, Python.

Interested candidates should be at Master, PhD, or Postdoc level.

Label-efficient deep learning of urinary tract infections

Dr. Artur Yakimovich - Helmholtz-Zentrum Dresden-Rossendorf

Dr. Artur Yakimovich - Helmholtz-Zentrum Dresden-Rossendorf

Recent advances in Machine Learning (ML) and Deep Learning (DL) are revolutionizing our abilities to analyse biomedical images and deepen our understanding of infection and disease. Among other host-pathogen interactions may be readily deciphered from microscopy data using convolutional neural networks. We work on developing the latest ML/DL and Computer Science methods to facilitate our understanding of Infection Biology and Disease Biology.

https://www.casus.science/casus/team/

a.yakimovich@hzdr.de

What is the project's research question?

Develop novel label-efficient annotation techniques for clinical brightfield microscopy of urine samples.

What data will be worked on?

Annotation of large clinical microscopy datasets is laborious and requires expert training. At the same time experts' time is often unattainable due to their primary roles as clinicians. This project will employ a clinical dataset of brightfield microscopy of patients' urine with a few annotated samples obtained by our collaborators at the Royal Free hospital in London. We aim to develop a diagnostic phenotype quantification workflow using label-efficient machine learning approaches.

What tasks will this project involve?

The tasks of this project will involve establishing the state-of-the-art self-supervised or weak-labelling learning method or architecture for object detection task in bright-filed microscopy. Comparing their performance on the clinical (or so-called "wild") dataset. Identification of the best performing method. Testing the method performance on an unseen dataset.

What makes this project interesting to work on?

Urinary tract infections (UTI) belong to the most common clinically relevant bacterial infections (Murray et al. 2021). 1 in 3 women worldwide will have at least one UTI by 24 years of age and 40 - 50% of women will experience one UTI during their lifetime with 44% experiencing recurrences. Improving deep learning methods without increasing the need for annotation efforts will have a direct impact on the clinical outcomes of UTI patients.

What is the expected outcome?

Co-authorship to research paper, Contribution to software

What infrastructure, programs and tools will be used? Can they be used remotely?

We will use HZDR Hemera GPU cluster to train the deep learning models used in this project. These tools can be used remotely. However, it will be essential to have the exchange student available for in-person work.

Is the project open source?

The dataset will be published by the time the project begins

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Deep learning, Machine learning, Parallel/distributed programming with GPUs, Computer vision and image processing/analysis, Software engineering, Python.

Interested candidates should be at Bachelor, Master, or PhD.

Massively parallel outlier detection in high-energy astrophysics.

Dr. Holger Stiele – Jülich Forschungszentrum.

Dr. Holger Stiele – Jülich Forschungszentrum.

Based at the Jülich Supercomputing Centre, the Sim Data Lab Astronomy and Astrophysics is a targeted research and support structure that provides an interface between the Supercomputer facilities in Jülich and the Astrophysics research communities. Our tasks include the support of Data Science Projects and High-Performance Computing Simulations for the Astrophysics Community.

https://www.fz-juelich.de/en/ias/jsc/about-us/structure/simulation-and-data-labs/sdl-astrophysics-and-astronomy

h.stiele@fz-juelich.de

What is the project's research question?

Anomaly detection in massive datasets is one of the most common problems in data science. In this project, we will address the detection of outliers in the massive Chandra Source Catalogue (see below). Identifying outliers in a large data set is a prerequisite to focus investigations on a smaller set of objects with potentially unexpected properties. Previous searches for outliers in the Chandra Source Catalogue (see Swarm et al. 2022) successfully apply machine learning algorithms but are constrained by the single-CPU memory limitation associated with libraries like scikit-learn. In this project, we will test the scalability of the memory-distributed tensor framework (e.g., Heat, see below) for outlier detection on Chandra data, with the goal of scaling out to even more massive datasets.

What data will be worked on?

The dataset we will work on is extracted from the Chandra Source Catalogue v.2 (CSC2). Chandra is NASA's big, multipurpose X-ray telescope that is in operation for more than two decades. CSC2 provides more than 500 measured or estimated properties for more than 300,000 X-ray sources. We will limit our study to sources with a high detection significance and to a limited subset of relevant summary properties of the different categories.

What tasks will this project involve?

The project involves applying machine learning algorithms available in the Heat framework (https://github.com/helmholtz-analytics/heat) to develop a pipeline for massively parallel outlier detection. You will work in close contact with the Heat dev team. Your code will be tested on a subset of the Chandra Source Catalogue to reproduce the findings of Swarm et al. (2022), and later applied to larger datasets. Supercomputing infrastructure will be available for testing.

What makes this project interesting to work on?

The ever-increasing amount of information makes it more and more difficult to find worthwhile targets of study in the sea of data. Identifying outliers in each set of data can help to focus investigations on a smaller set of objects that could prove as interesting because they do not follow the trends of the underlying population. In this project you will gain experience in machine learning for outlier detection on an HPC system, and basic parallel computing. You will be involved in the development/improvement of open-source software. You will also gain some basic experience in the properties of high-energy astrophysical objects.

What is the expected outcome?

Co-authorship to research paper, Contribution to software

What infrastructure, programs and tools will be used? Can they be used remotely?

Heat is written in Python, using NumPy and PyTorch functionalities, and MPI for parallel operations. https://github.com/helmholtz-analytics/heat. We use GitHub, Mattermost and video calls for developer discussions. We have access to the supercomputers at the Jülich Supercomputing Centre: https://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/supercomputers_node.html.

All the infrastructure can be used remotely.

Is the project open source?

Yes

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Machine learning, Python

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

SMARAGD (Sensors for measuring aerosols and reactive gases to deduce health effects)

Dr. Robert Wegener - Jülich Forschungszentrum.

Dr. Robert Wegener - Jülich Forschungszentrum.

Within our group, we study the emission, distribution, and degradation of pollutants. We use our mobile laboratory, high resolution atmospheric chemistry models and small sensors to monitor the distribution of pollutants in the urban environment. We develop new methods and operate the World Calibration Center for Nitrogen Oxides.

https://www.fz-juelich.de/en/iek/iek-8/research/reactive-trace-substances/energy-related-emissions

r.wegener@fz-juelich.de

What is the project's research question?

The aim of the project is to link the pollutant concentrations measured at citizens' homes with their health data.

What data will be worked on?

As part of this Citizen Science project, a network of around 100 sensors is being set up in the city of Cologne. The sensor units contain electrochemical sensors to measure CO, NO, NO2 and Ox as well as particulate matter (PM). Although these sensors can measure accurately, they are susceptible to interference that must be characterized and monitored. The sensors will be placed at citizens' homes to estimate individual pollution levels. Along with the sensors, participants will receive an app to report on their health status. Both datasets will be linked in the project. https://www.fz-juelich.de/en/iek/iek-8/projects/smaragd

What tasks will this project involve?

The student works with the data set, develops methods to ensure the quality of the data by comparing them to each other, locates failed sensors, derives the distribution of pollutants in the city, their individual daily cycles, and their variability.

The student visualizes the data, compares data sets from different sensors and sensor data with data from official monitoring sites. He or she derives individual citizen exposure to pollutants from the data, identify factors that determine pollutant concentrations, and investigate possible relationships between air quality and health data.

What makes this project interesting to work on?

The project is a joint project between the Forschungszentrum Jülich, where the sensors are developed and characterized, the Helmholtz Institute Munich, the Open Knowledge Lab in Cologne, and the Helmholtz Centre for Infection Research in Braunschweig, where the app with the health data questionnaire is maintained. The project is thus at the interface of epidemiological studies and atmospheric chemistry, and the student can gain experience in

both subjects. The citizen science approach is particularly interesting, as citizens are also encouraged to participate and share their experiences in the data analysis.

What is the expected outcome?

Co-authorship to research paper, data quality of the sensor network.

What infrastructure, programs and tools will be used? Can they be used remotely?

Air quality data is stored on an SQL server in Jülich. The data will be analysed with Python tools and IDL. The tools can be operated remotely.

Is the project open source?

Yes

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Computer vision and image processing/analysis, Python.

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

Personal exposure to airborne pollutants in three German cities.

Dr. Simonas Kecorius – Helmholtz Munich

Dr. Simonas Kecorius – Helmholtz Munich

The primary aim of the research unit is the development and employment of exposure assessment methods for applying in epidemiological studies on health effects of air pollutants. Main activities are focused on detailed physical and chemical characterization of ambient particles collected at an aerosol measurement stations and during the intensive field campaigns. The data is then used for investigating the health relevance of particulate matter.

https://www.helmholtz-munich.de/en/epi/research-groups/environmental-exposure-assessment

simonas.kecorius@helmholtz-muenchen.de

What is the project's research question?

What are the determinants of personal exposure to airborne pollutants in the city?

What data will be worked on?

The student will work on the air quality data set collected during intensive mobile measurement campaign in three German cities - Munich, Augsburg, and Regensburg. Specifically, data set comprises of aerosol particle and equivalent black carbon, as well gaseous pollutant number and mass concentrations, road traffic videos from onboard camera, and geo-spatial data (e.g., commuted routes; geo-location; road types; etc.). Additionally, meteorological and air quality data from long-term monitoring sites will also be available for the project.

What tasks will this project involve?

Data evaluation and quality assurance (initial data screening and preparation for analysis); Development and application Artificial Intelligence (AI; e.g. Neural Network) model to evaluate vehicular flow from onboard videos. Spatiotemporal analysis of airborne pollutant concentrations (creating pollution maps in geographic information system (GIS); categorizing data; etc). Visualization of the results in a form of plots and figures, as well as preparing a short report concluding the main findings.

What makes this project interesting to work on?

Impaired air quality due to vehicular emissions is one of the most important environmental factors contributing to premature deaths in Europe. Qualitative and quantitative exposure assessment of airborne pollutants is therefore of high importance for controlling air quality, reducing health risks, and improving life quality in general. Increasing our understanding about pollutant emission, dispersion, and its effects on personal exposure through the application of novel tools onto complex data sets is therefore both interesting and highly rewarding. During this project, exchange student will increase his/her competence in

complex environmental data analysis, work management, scientific writing and communication.

What is the expected outcome?

Co-authorship to research paper, Contribution to software, Material for the conference.

What infrastructure, programs and tools will be used? Can they be used remotely?

To successfully accomplish project tasks, the applicant is expected to be familiar with some geographic information system (GIS. E.g. qgis) software, programming language for statistical computing and graphics (e.g. R, Python, etc.), as well the concepts for machine learning. All required tools are freely available online and will be provided by the host institution.

Is the project open source?

Yes

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Machine learning, Deep learning, Computer vision and image processing/analysis, Geographic information systems, Python, R

Interested candidates should be at Bachelor, Master, PhD, or Postdoc level.

B2S3: Benchmark and Baselines for Semantic Segmentation of SAR Data

Dr. Ronny Hänsch – DLR Microwaves and Radar Institute

Dr. Ronny Hänsch – DLR Microwaves and Radar Institute

Synthetic Aperture Radar (SAR) has many advantages over optical systems, including being independent of daylight and being able to penetrate clouds. My institute is concerned with radar on different levels, including space-borne and air-borne SAR missions. My lab builds the air-borne SAR system F-SAR which acquires high-resolution, multi-frequency, fully-polarimetric SAR images and creates algorithms to process and analyse SAR data to its fullest potential.

https://www.dlr.de/hr/en/desktopdefault.aspx/tabid-2326/

rww.haensch@gmail.com

What is the project's research question?

How well do deep-learning based approaches general over remote sensing data acquired over different regions of the Earth?

What data will be worked on?

The Copernicus program offers free public access to several European Earth observation satellites including the Sentinel 1 (SAR) and Sentinel 2 (Multispectral) constellations. We will exploit these missions and leverage both, SAR as well as multispectral (i.e. optical) data.

What tasks will this project involve?

Get acquainted with sources to download Sentinel 1 and Sentinel 2 data (either directly via the Copernicus data hub or other providers). Perform standard pre-processing of the data via SNAP (i.e. geocoding). Annotate a few (<5) scenes to complement an existing dataset with common land cover classes. Train/evaluate deep neural networks for land cover classification on the data in various configuration to test generalization over geographic regions. Train/evaluate a self-supervised learning approach on the data to evaluate its potential regarding label-scarcity. Publish data, code, and results in a scientific publication.

What makes this project interesting to work on?

This project spans a large variety of highly relevant tasks providing deep insights into several hot research topics (curation of datasets, evaluation of machine learning, self-supervised learning) which would be impossible to cover in a tight time frame such as given in this exchange program. However, for each step there is already a lot of prior work that will be leveraged, keeping the actual workload at a minimum while maximizing the potential outcome. The student will learn something about SAR, i.e., an exciting sensor technology that is complementary to optical sensors, benchmarking, as well as deep learning and self-supervised learning.

What is the expected outcome?

Co-authorship to research paper, Contribution to software. Co-authorship in a dataset.

What infrastructure, programs and tools will be used? Can they be used remotely?

Copernicus open data hub (data access), SNAP (pre-processing), PyTorch/Tensorflow (deep learning). None of them is proprietary, all of them can be used locally (or even remotely).

Is the project open source?

Yes

What skills are necessary for this project?

Machine learning, Deep learning, Computer vision and image processing/analysis.

Interested candidates should be at Master or PhD level.

Perovskite PV Device & Module Database: Slot-die Coated Devices

Prof. Dr. Eva Unger – Helmholtz Zentrum Berlin

Prof. Dr. Eva Unger – Helmholtz Zentrum Berlin

We work on scalable solution-based deposition methods for halide perovskite solar cells making emerging PV module prototypes in close collaboration with Industry. We have as a team also been actively setting up a literature database for perovskite solar cells that we launched in 2022 in collaboration with the Israeli Start up Materials Zone (more information regarding this project can be found here: https://www.perovskitedatabase.com). We are now looking to expand our collaborative activities with Materials Zone to develop standardized data entry protocols with respect to industrially relevant PV prototype fabrication processes.

https://www.helmholtz-berlin.de/forschung/oe/se/hybrid-materials/index_en.html

eva.unger@helmholtz-berlin.de

What is the project's research question?

Structure-Property-Performance relationships in "historic" device data produced by slot-die coating in the HySPRINT Innovation Lab.

What data will be worked on?

The goal is to consolidate data available in the published literature, which has been collected on the PerovskiteDatabase (www.perovskitedatabase.com) and experimental solar cell data for solar cells and modules made by slot-die coating in our research laboratory.

What tasks will this project involve?

Define a standardized data entry protocol specifically for "slot-die coated" "Perovskite solar cells" with a particular focus on larger area "modules" for Materials Zone Data management platform. Extract/Transfer data from the PerovskiteDatabase (www.perovskitedatabase.com), into experimental data management environment of Materials Zone. Assist in the entry of experimental data into the database generated by team members. If capable, by making use of TextMining approaches to extract data from recent publications. If capable, make use of AI tools to establish relationship between ink composition, processing conditions, device structure, and devices and module performance

What makes this project interesting to work on?

Contribution to developing data management, sharing and dissemination platforms both in the OpenScience domain but also complementary with platforms used in commercial PV manufacturing sector.

What is the expected outcome?

Co-authorship to research paper, Contribution to software

What infrastructure, programs and tools will be used? Can they be used remotely?

Materials Zone data management platform - can be used remotely.

Is the project open source?

Yes

What skills are necessary for this project?

Data analytics, statistics, Scientific computation, data mining, Machine learning, Databases, Python.

Interested candidates should be at Master or PhD level.

Developing sonification techniques for climate data analysis, interpretation, and communication.

Dr. Vera Fofonova - Alfred Wegener Institute

Dr. Vera Fofonova - Alfred Wegener Institute

I am a theoretical mathematician, and climate model developer with the focus of my research in numerical ocean modelling, and hydrodynamics. Apart from the main subject of research I am developing algorithms that could help communicate scientific data through musical signals. In other words, I aim to develop neural network models that will enable sonification (making sound out of external resources) of any ocean and climate data sets.

https://www.awi.de/ueber-uns/organisation/mitarbeiter/detailseite/vera-fofonova.html

vera.fofonova@awi.de

What is the project's research question?

Our project explores the possibilities to use AI technologies to link scientific data and music composition. In particular, the project aims to explore and further develop sonification techniques that would help analyse, interpret and communicate scientific data (e.g. climate model outputs).

What data will be worked on?

NetCDF climate model outputs (.nc), csv files of converted climate data, audio files in MIDI format.

What tasks will this project involve?

The primary goal of the project is to build an encoder-decoder structure that is often used to solve “translation” problems with machine learning (ML) algorithms, e.g., German-to-english, text-to-image, image-to-text, music-to-image. Both the encoder and the decoder part can be ML algorithms. Depending on the development stage participant(s) could be involved in one of the following project phases:

Phase 1. Building an encoder.

The first part of the project focuses on building an Encoder that will obtain a lower-dimensional representation from high-dimensional input data (dimensionality reducible algorithms), e.g. 4D fields from climate model output. Here, participant(s) would apply dimensionality reducible algorithms (self-supervised) to multi-dimensional climate data in order to obtain the outputs that can be used as inputs in a sound generating software (e.g. MAX).

Phase 2. Building a decoder

Mapping the extracted in Phase 1 time series to audio features, so-called “data sonification” and exploring data composition by ML algorithms (involves training and labelling data sets).

What makes this project interesting to work on?

AI technologies are increasingly commonplace in academic inquiry and an emergent area of interest in contemporary music and sound-art practice. Music may be a valuable resource in capturing and transforming scientific information, whereas scientific data can become a background for creative works and stage performances. This project suggests a foundation for new creative works and helps convert complex scientific language into a form of music as a lingua franca to reach a wider audience. Project outputs will give new insights, approaches and processes for scientists and artists working together. The project involves collaboration with a contemporary musician from Edinburg (Michael Begg) and data scientists from Helmholtz AI network.

What is the expected outcome?

Contribution to software

What infrastructure, programs and tools will be used? Can they be used remotely?

Computing resources for the Helmholtz AI community (HAIRCORE); Python and related to python modules and programs.

Is the project open source?

Yes

What skills are necessary for this project?

Machine learning, Deep learning, Software development, Python, any additional programming languages will be of an advantage.

Interested candidates should be at Master, PhD or, Postdoc level.

Artificial Intelligence for Ophthalmology Medical Support of Astronauts.

Dr. Claudia Stern and Dr. Scott Ritter – DLR Klinische Luft und Raumfahrtmedizin.

Dr. Claudia Stern and Dr. Scott Ritter – DLR Klinische Luft und Raumfahrtmedizin.

Long duration human spaceflight missions create medical support challenges for eye changes, which can occur in nearly two-thirds of astronauts. To address these challenges, we are developing artificial intelligence applications to support crew members in monitoring their eyes. These applications have the potential to be used for crew medical support aboard the International Space Station, and beyond.

https://www.dlr.de/me/de/desktopdefault.aspx/tabid-1768

claudia.stern@dlr.de

scott.ritter@dlr.de

What is the project's research question?

Can artificial intelligence applications be used to provide crew medical support for ophthalmology during long duration spaceflight missions?

What data will be worked on?

Medical image and video data collected from the human eye at various sites.

What tasks will this project involve?

Supporting the development of artificial intelligence models, using computer vision, convolutional neural networks, regression, classification, and object detection.

What makes this project interesting to work on?

The project would be interesting to work on because it has the potential to be used during future human spaceflight missions and may help address medical concerns for exploration-class human spaceflight missions.

What is the expected outcome?

Co-authorship to research paper, Contribution to software.

What infrastructure, programs and tools will be used? Can they be used remotely?

For the machine learning component of our research, Python, convolutional neural networks, Tensorflow, GPU servers, and computer vision tools are used to conduct our analyses. To collect the raw image and video data used in our analysis, we use ophthalmology imaging tools (e.g., for fundoscopy and optical coherence tomography (OCT)) commonly used in clinical practice worldwide. You would require access to a development environment (e.g., VSCode, Pycharm, Colab), understanding of and adherence

to data security and ethics standards, and a modern smartphone/tablet. Work could be done remotely.

Is the project open source?

No

What skills are necessary for this project?

Data analytics, statistics, Machine learning, Deep learning, Computer vision and image processing/analysis, Software development, Python.

Interested candidates should be at Master, PhD, or Postdoc level. We are looking for a person with experience in applying computer vision or machine learning to images and video data. Experience working with human medical data would be valuable.