Helmholtz Analytics Framework

The Helmholtz Analytics Framework is a data science pilot project funded by the Helmholtz Association. Six Helmholtz centers will pursue a systematic development of domain-specific data analysis techniques in a co-design approach between domain scientists and information experts in order to strengthen the development of data sciences in the Helmholtz Association. Data analytics methods will be applied to challenging applications from a variety of scientific fields in order to demonstrate their potential in leading to scientific breakthroughs and new knowledge. In addition, the exchange of methods among the scientific areas will lead to their generalization.

The Helmholtz Analytics Framework (HAF) is complementary to the Helmholtz Data Federation (HDF) in that the developed libraries will be made available there first. The three-year project starts in October 2017 and receives funding of close to €3 million.

Scientific Big Data Analytics (SBDA) has become a major instrument of modern research for tackling scientific problems of highest data and computational complexity. SBDA deals with data retrieval, assimilation, integration, processing and federation on an unprecedented scale, made possible through leading-edge high-performance computing and data management technologies.

It is crucial that systematic development of domain-specific data analytics techniques will be carried out as a co-design activity between domain and infrastructure scientists. This happens within a set of highly demanding use cases spanning six Helmholtz centers—DESY, DKFZ, DLR, FZJ, HMGU, and KIT—spanning

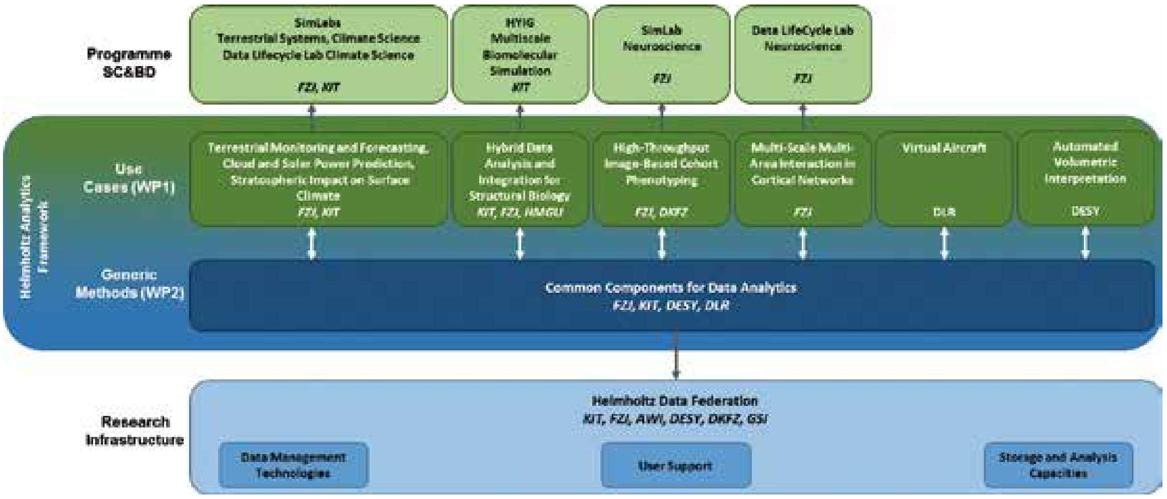
Programme SC&BD
SimLabs Terrestrial Systems, Climate Science Data Lifecycle Lab Climate Science FZJ, KIT
HYIG Multiscale Biomolecular Simulation KIT
SimLab Neuroscience FZJ
Data LifeCycle Lab Neuroscience FZJ
Helmholtz Analytics Framework
Use Cases (WP1)
Terrestrial Monitoring and Forecasting, Cloud and Solar Power Prediction, Stratospheric Impact on Surface Climate FZJ, KIT
Hybrid Data Analysis and Integration for Structural Biology KIT, FZJ, HMGU
High-Throughput Image-Based Cohort Phenotyping FZJ, DKFZ
Multi-Scale Multi-Area Interaction in Cortical Networks FZJ
Virtual Aircraft DLR
Automated Volumetric Interpretation DESY
Generic Methods (WP2)
Common Components for Data Analytics FZJ, KIT, DESY, DLR
Research Infrastructure
Helmholtz Data Federation KIT, FZJ, AWI, DESY, DKFZ, GSI
Data Management Technologies
User Support
Storage and Analysis Capacities

Fig. 1: Structure of the Helmholtz Analytics Framework project

five scientific domains: earth system modeling, structural biology, aeronautics and aerospace, medical imaging, and neuroscience. The exchange of techniques between the use cases will lead to generalizations and standardization to be made available to yet other fields and users.

The HAF will boost the development of the HDF, which is designed to be the hardware and support backbone for the entire Helmholtz Association and will address the dramatically increasing demands of science and engineering for transforming data into knowledge.

Thus, we start an exciting culture for future systematic developments of the HAF on top of the HDF.

Approach

The research strategy of the project is based on co-designing domain-specific data analytics techniques by domain scientists, together with data and computer scientists, evolving data analytics methods, developing the infrastructure, the HDF, with basic software systems and suitable interfaces to the application software. These activities are coherently derived from properly defined "use cases." The use cases are chosen such that they themselves target, in a complementary manner, scientific challenges with an important societal impact and a high potential for breakthroughs in their respective domains. Through this interdisciplinary cooperation, the HDF investments will be leveraged towards a full-system solution. It is an important goal of the project to translate specific methods developed within given use cases into generic tools and services. In a first step, they are made available to other use cases raising synergy within the project. Later, the methods will become beneficial to other fields.

Eight use cases from five scientific domains are participating in this project. The domains are Earth System Modeling, Structural Biology, Neuroscience, Aeronautics and Aerospace, and Research with Photons.

Earth system modeling

In the use case Terrestrial Monitoring and Forecasting, forecasts and projections of the terrestrial water and energy cycles constitute a scientific grand challenge due to the complexities involved and the socioeconomic relevance. Prominent examples include forecasts of weather, extreme events (floods, low-flows, droughts), water resources and long-term climate projections emerging as one of the major pillars in Earth system discovery including climate change research. Major SBDA methods encompass ensemble data assimilation technologies and genetic algorithms.

The proper forecasting of clouds in the use case Cloud and Solar Power Prediction is important for the short-term predictions of photovoltaic power, photo-chemically impaired air quality, and precipitation. The transfer of this spaceborne information in prognostic models, to result in a demonstrated beneficial effect on cloud evolution and prediction capabilities, is an unresolved issue. Major SBDA methods to be applied are supervised learning as well as parallel and scalable classification algorithms.

Recent model developments in meteorology allow more seamless approaches to modeling weather and climate in a unified framework for the use case Stratospheric Impact on Surface Climate. An application of these advances is a hind-cast assessment of well-observed winter seasons in the northern hemisphere. Each of these (retrospective) forecasts will consist of an ensemble of realizations subsequently compared to the "real world" in order to find the most realistic ensemble member and put the real development into context with the ensemble statistics. Simulation runs will produce a large volume of 5 dimensional data, requiring fast processing for building up successive analysis layers for individual winters and for comparing all available winters in a climatological context.

Structural biology

The use case Hybrid Data Analysis and Integration for Structural Biology deals with the determination of structural ensembles of biomolecular complexes required to understand their biological functions. Single experimental techniques cannot describe the complex conformational space and temporal dynamics, and thus the integration of many complementary data with advanced computational modeling is essential. The project vision is to develop the concepts and methods needed to integrate experimental data from NMR spectroscopy, single-particle cryo-electron microscopy, and co-evolution analysis of genetic sequences with molecular dynamics simulations. The required computational tools such as Bayesian modeling, enhanced sampling techniques, multidimensional statistical inference, feature extraction, and pattern recognition, will be developed within the algorithmic and technological framework of the HDF.

Neuroscience

Advanced medical research, like understanding of the brain or personalized medicine, are facing the challenge to understand the correlation and effect model between environmental or genetic influence and the observed resulting phenotypes (e.g. morphological structures, function, variability) in healthy or pathologic tissue. The use case High-Throughput Image-Based Cohort Phenotyping will involve neuroimaging as piloting image domain to establish time-efficient parallel processing on HPC clusters as well as highly robust but flexible processing pipelines, efficient data mining techniques, uncertainty management, sophisticated machine learning and inference approaches. Such analyses are not only of high value for systems neuroscience and medical science, but also could be generalized for other disciplines searching for causalities between image-based observations and underlying mechanisms.

The use case Multi-Scale Multi-Area Interaction in Cortical Networks employs parallelized data mining strategies paired with statistical Monte-Carlo approaches to evaluate signatures of correlated activity hidden in the high-dimensional ensemble dynamics recorded simultaneously from visual and motor brain areas in order to link neuronal interactions to behavior. There are two challenges to be tackled by this use case. Multi-dimensional correlation analysis methods of activity due to the combinatorial complexity, strong undersampling of the system, and non-stationarities that prohibit the use of analytic statistical tests lead to increased computational demands. In addition, the heterogeneity and complex structure of the various data streams, including rich metadata, require suitable informatics tools and protocols for the acquisition of metadata and provenance tracking of the analysis workflows.

Aeronautics and aerospace

The use case Virtual Aircraft employs reduced order models that extract relevant information from a limited set of large-scale high-fidelity simulations through elaborate result analysis methods to provide an attractive approach to reduce numerical complexity and computational cost while providing accurate answers. Data classification methods are of interest to gain more physical insight, e.g., to identify (aerodynamic) nonlinearities and to track how they evolve over the design space and flight envelope. The Virtual Aircraft use case will lead to SBDA techniques from other fields of research being evaluated for extracting a comprehensive digital description of an aircraft from a parallel workflow based on high-fidelity numerical simulations. The Virtual Aircraft use case will also contribute a wide range of methods for data fusion, surrogate and reduced-order modeling to the generic methods that can be applied to the use cases of other partners. The software and SBDA methods to set up the Virtual Aircraft can be developed in such a generic fashion that it will be possible to adapt them to other fields of research that deal with product virtualization.

Research with photons

SBDA techniques can be used for Automated Volumetric Interpretation of time-resolved imaging of materials and biological specimen to provide deep insight into dynamics of bacterial cells, composite materials, or living organisms, among others. Experiments are coming from X-ray imaging at synchrotrons or free-electron lasers. The quality of automated segmentation and interpretation algorithms will strongly increase with the amount of available data combined with SBDA techniques to harvest and mine prior information from similar experiments across facilities and disciplines. To maximize the sample size, we aim to exploit the vast amount of imaging data available in the Helmholtz Data Centers as well as the PaNdata collaboration, which includes almost all European Photon and Neutron sources, and also collaborations with various other light sources, particularly in the USA. The interpretation of 3D-data by volumetric segmentation and interpretation can greatly benefit from SBDA by harvesting and mining prior information from similar experiments across facilities and disciplines.

Work plan

The project has a duration of 36 months. During the initial phase, we will determine common methods and respective tools among the use cases. An initial set of common method areas, including stochastics, image analysis, supervised and unsupervised learning, has already been identified. During the second phase, the methods for mutual use in the participating use cases will be generalized and the tools will be adapted and rolled out on the HDF. It is expected that this will lead to cross-fertilization in the use of common methods. In the final phase, the common methods and tools will be made available for a wider audience. Care will be taken to make tools available not only among participating scientific domains, but also generically. This will include appropriate documentation of the methods as well as the tools that implement them and their installation and usage on the HDF.

Acknowledgement

The project is funded by the Helmholtz Association Initiative and Networking Fund under project number ZT-I-0003.

Written by Björn Hagemeier

Jülich Supercomputing Centre (JSC)

Contact: b.hagemeier@fz-juelich.de

Daniel Mallmann

Jülich Supercomputing Centre (JSC)

Contact: d.mallmann@fz-juelich.de

Achim Streit

Karlsruhe Institute of Technology, Steinbuch Centre for Computing

Contact: achim.streit@kit.edu