Helmholtz Analytics Framework (HAF)

Markus Götz, KIT

TRIUMF Data Science and Quantum Computing Workshop

Vancouver June, 28th 2018

HELMHOLTZ
Analytics Framework

Concept

- Helmholtz incubator: Scientific Big Data Analysis
- Co-design approach (i.e. domain scientists + data analysts)

Funding

- ~6m Euro over three years
- 23 FTEs, 2/3 domain sciences, 1/3 generic methods

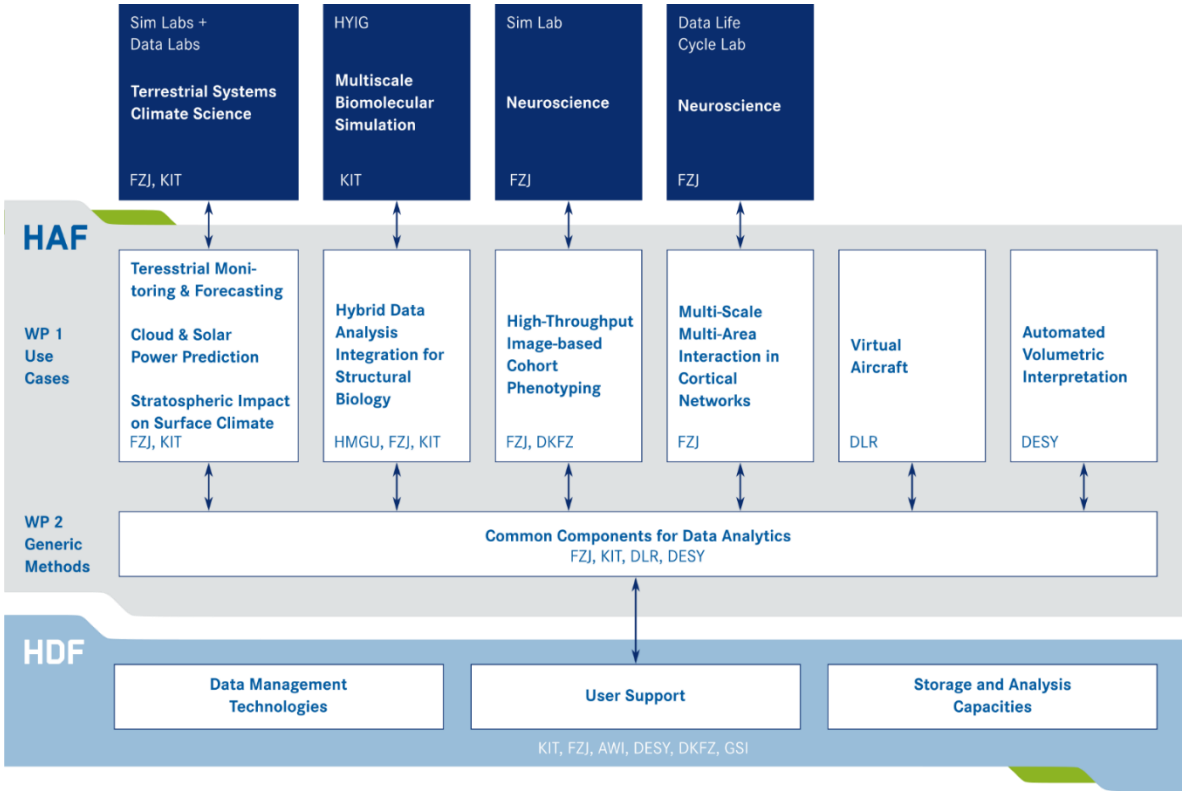
Mission

1. Data analysis methods exchange network and implementation
2. Software for large-scale analysis
3. Closer collaboration in Helmholtz

Project Partners

- Deutsches Elektronen Synchrotron (DESY)
- Deutsches Zentrum für Luft- und Raumfahrt (DLR)
- Deutschen Zentrum für Krebsforschung (DKFZ)
- Forschungszentrum Jülich (FZJ)
- Helmholtz Zentrum München (HMGU)
- Karlsruher Institut für Technologie (KIT)

Project Structure



Use Cases

Terrestrial Modelling and Forecasting

Scientific Problem

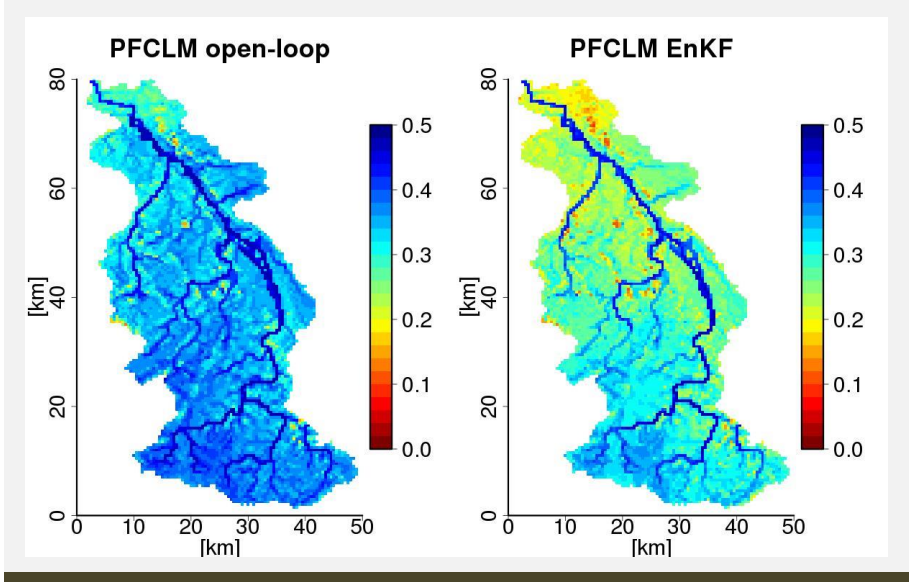
- Earth system modelling
- Simulation of ground, water and atmosphere
- Inclusion of measurement data (e.g. satellite)

Methods

- Data assimilation

Challenges

- Volume problem
- Square assimilation matrices O(10^8) elements



Use Cases

Cloud and Solar Prediction

Scientific Problem

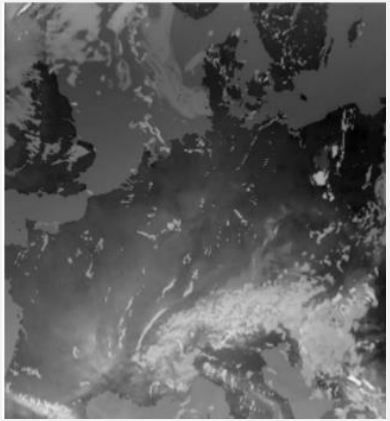
- Ensemble simulation of cloud cover
- Prediction of solar radiation for energy sector

Methods

- Ensemble state selection
- Non-gaussian assimilation filters

Challenges

- Small training data with large feature amount



Use Cases

Stratospheric Impact on Surface Climate

Scientific Problem

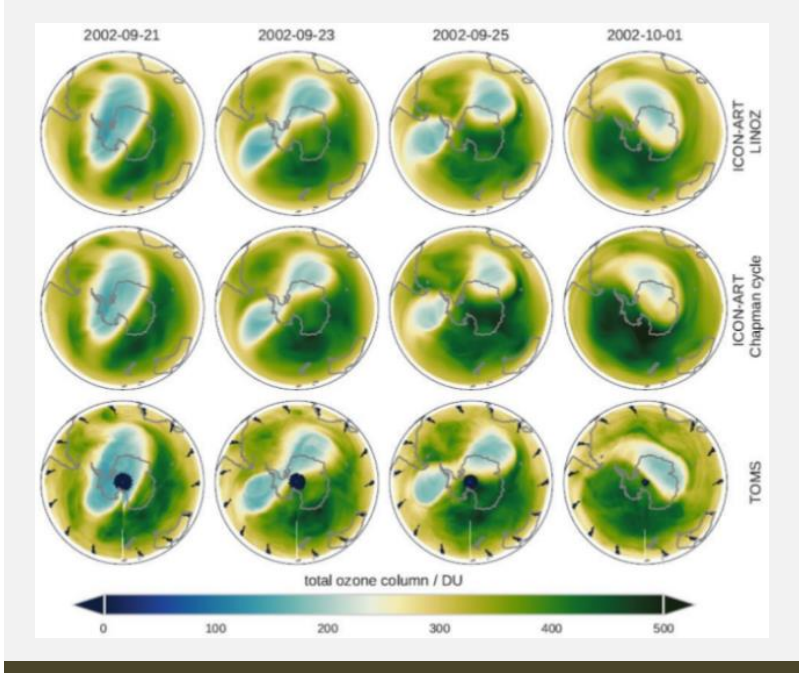
- Ensemble ozone layer simulation
- Hindcast approach

Methods

- Ensemble state estimation
- Deep neural network regression

Challenges

- Legacy data resolution/conversion problem



Use Cases

Structural Biology

Scientific Problem

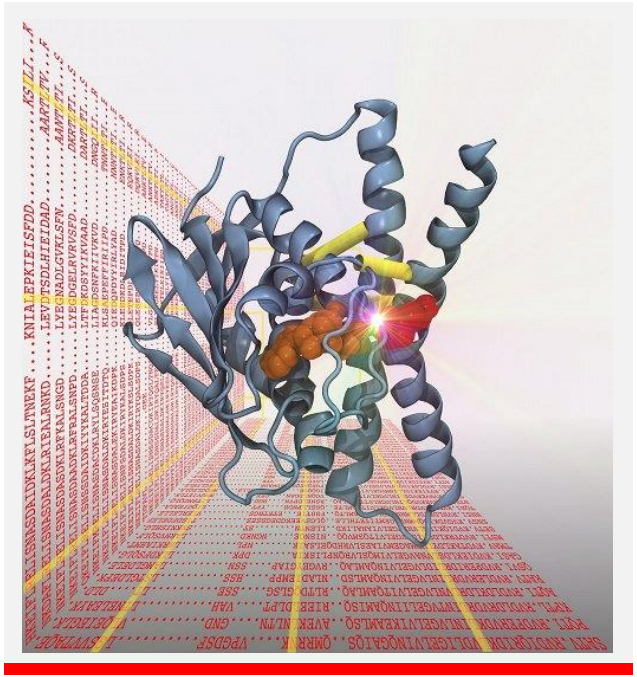
- Protein and RNA structure prediction
- Direct coupling analysis

Methods

- Scatter experiments
- Molecular simulation
- Sequence prediction using NLP techniques

Challenges

- Volume of sequencing/image scatter/simulation data
- Core research on sparse algebra in ML



Use Cases

Image-based Cohort Phenotyping

Problem

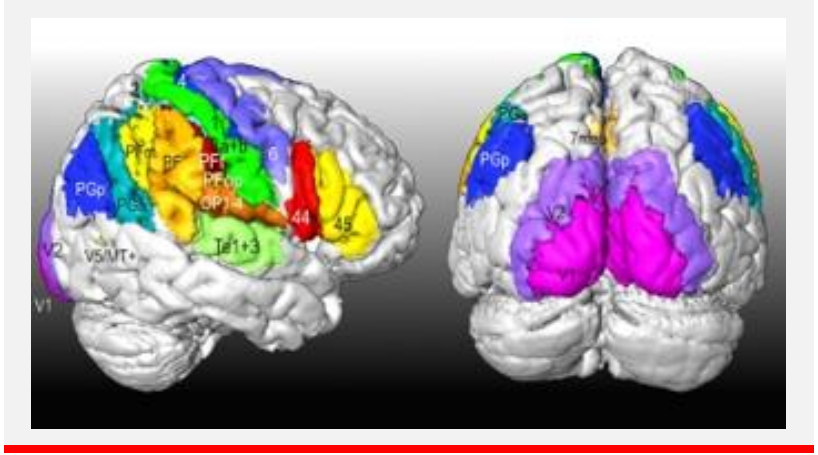
- Identification of functional areas in brain
- Effect of environment on brain development

Methods

- Image processing of brain slices
- Volumetric reconstruction

Challenges

- Data volume (entire cohort ~20 PB)



Use Cases

Interaction in Cortical Networks

Scientific Problem

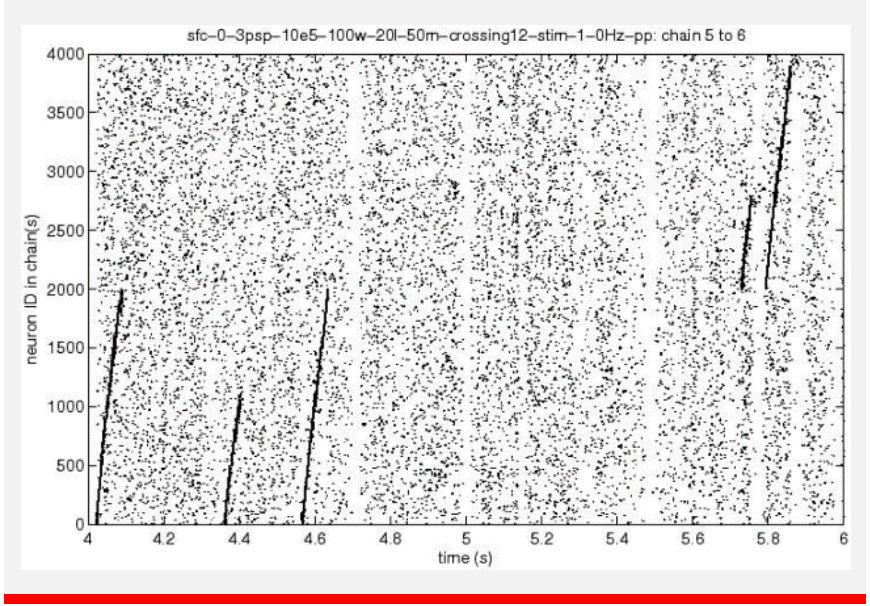
- Determine cortical interactions of neurons
- Identify connected, spiking neuron „networks“

Methods

- Market-basket analysis
- Sequence analysis with auto-encoders

Challenges

- Large search space (human: 10^10 neurons)
- Dimensionality reduction

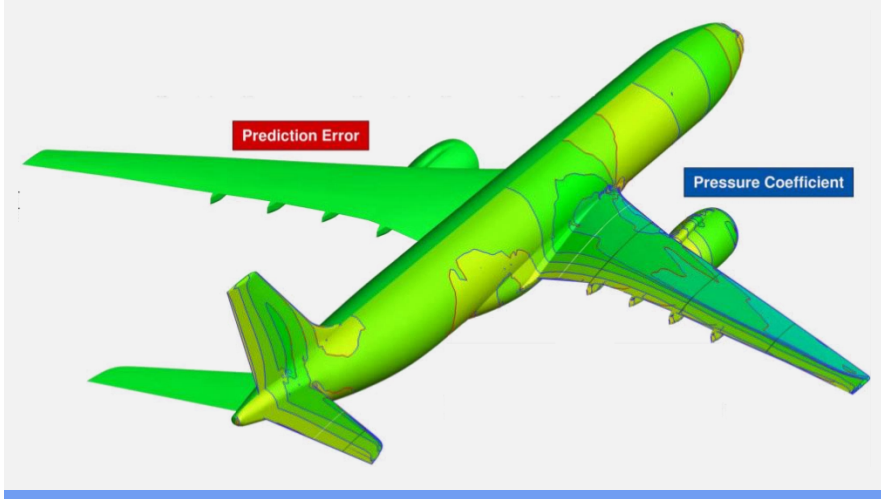


Use Cases

Virtual Aircraft

Scientific Problem

- Aircraft design (especially sheer and pressure)
- Simulation of designs expensive
- Probabilistic candidate determination



Methods

- Dimensionality reduction
- Clustering and classification (neural network driven)
- (Gradient-) Kriging

Challenges

- Verification of found solution

Use Cases

Volumetric Interpretation

Scientific Problem

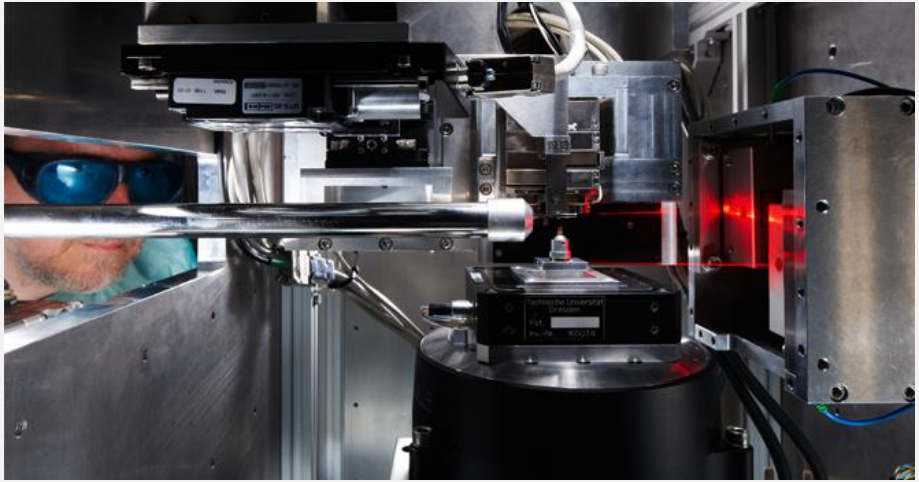
- Photon scatter experiments
- Trigger system for fast filtering decision

Methods

- Convolutional neural networks
- Classical image vision

Challenges

- Data volume
- Velocity of data – short decision time



Some Observations…

- Different domains, similar problems
- Data volume is main challenge (for our problems)
- Parallelization is major entry barrier into Big Data analysis
- Mixed computation mode
  - HPC – simulation/data analysis model
  - HTC – data parallel farming

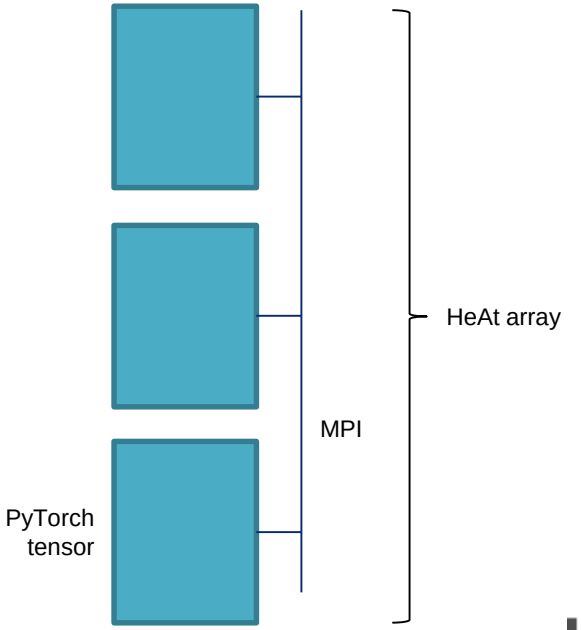
HeAT – Helmholtz Analytics Toolkit

Vision: Distributed Numpy++

- **Multi-dimensional, distributed arrays**
- Look-and-feel of numpy
- Parallelization and distributed computing
  - Vectorization
  - GPUs
  - MPI communication
- Automatic differentation for neural networks
- Internally tensor engine: PyTorch

Try it! … soon

- https://github.com/helmholtz-analytics/heat



HeAT

Conclusion

Summary

- Helmholtz Incubator
- Scientific big data analysis
- Highly multi-disciplinary

Outlook

- Active development on HeAT
- Porting of first complete data analysis pipelines

HELMHOLTZ
Analytics Framework