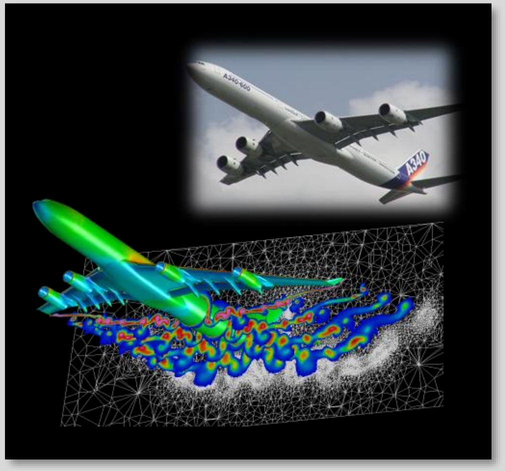


# High Performance Computing

ADVANCED SCIENTIFIC COMPUTING


**Ph.D. Student Chadi Barakat**

Simulation Data Lab – Health and Medicine

School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland

Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany

@ProfDrMorrisRiedel @Morris Riedel @MorrisRiedel @MorrisRiedel

https://www.youtube.com/channel/UCWC4VKHmL4NZgFfKoHtANKg morris@hi.is


LECTURE 11

# HPC Applications in Health and Neurosciences

March 15, 2021

Online Lecture

EUROPEAN OPEN SCIENCE CLOUD · EOSC NORDIC · EuroHPC Joint Undertaking · ADMIRE · EURO

UNIVERSITY OF ICELAND SCHOOL OF ENGINEERING AND NATURAL SCIENCES FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

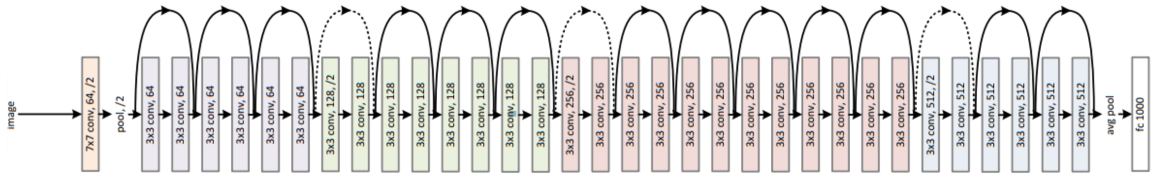
DEEP Projects

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

# Review of Practical Lecture 10.2 – Parallel & Scalable Machine and Deep Learning

## Deep Learning via RESNET-50 Architecture – A Case for interconnecting GPUs

- Classification of land cover in scenes in Remote Sensing
  - Very suitable for parallelization via distributed training on multi GPUs

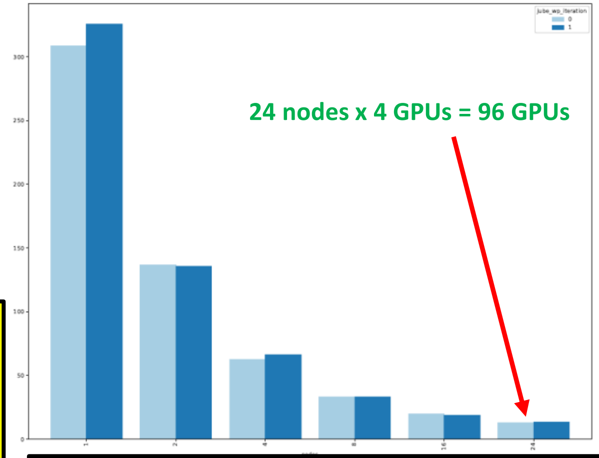



*[20] RESNET*

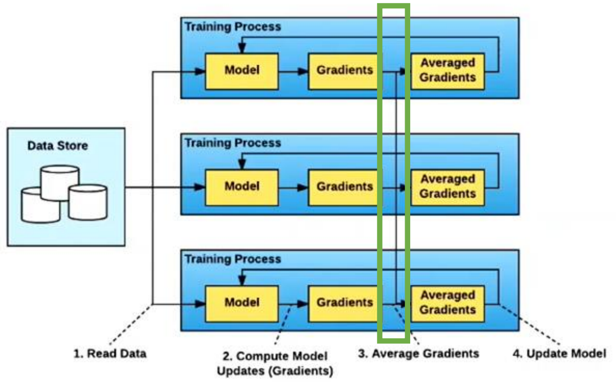
- **RESNET-50 is a known neural network architecture that has established a strong baseline in terms of accuracy**
- **The computational complexity of training the RESNET-50 architecture relies in the fact that is has ~ 25.6 millions of trainable parameters**
- **RESNET-50 still represents a good trade-off between accuracy, depth and number of parameters**
- **The setups of RESNET-50 makes it very suitable for parallelization via distributed training on multi GPUs**

## Distributed Training with Multi GPU Usage using Horovod

A partition of the JUWELS system has 56 compute nodes, each with 4 NVIDIA V100 GPUs (equipped with 16 GB of memory)


Horovod distributed training via MPI_Allreduce()




*[5] Horovod*

- **Horovod is a distributed training framework used in combination with low-level deep learning frameworks like Tensorflow**
- **Horovod uses MPI for inter-process communication, e.g., MPI_Allreduce()**
- **Distributed training using data parallelism approach means: (1) Gradients for different batches of data are calculated separately on each node; (2) But averaged across nodes to apply consistent updated to the deep learning model in each node**

# Outline of the Course

1. High Performance Computing
2. Parallel Programming with MPI
3. Parallelization Fundamentals
4. Advanced MPI Techniques
5. Parallel Algorithms & Data Structures
6. Parallel Programming with OpenMP
7. Hybrid Programming & Patterns
8. Debugging & Profiling & Performance Analysis
9. Accelerators & Graphical Processing Units
10. Parallel & Scalable Machine & Deep Learning
11. HPC in Health & Neurosciences
12. Computational Fluid Dynamics & Finite Elements
13. Systems Biology & Bioinformatics
14. Molecular Systems & Material Sciences
15. Terrestrial Systems & Climate
16. Epilogue

+ additional invited lectures by experts & practical lectures for our hands-on assignments in context

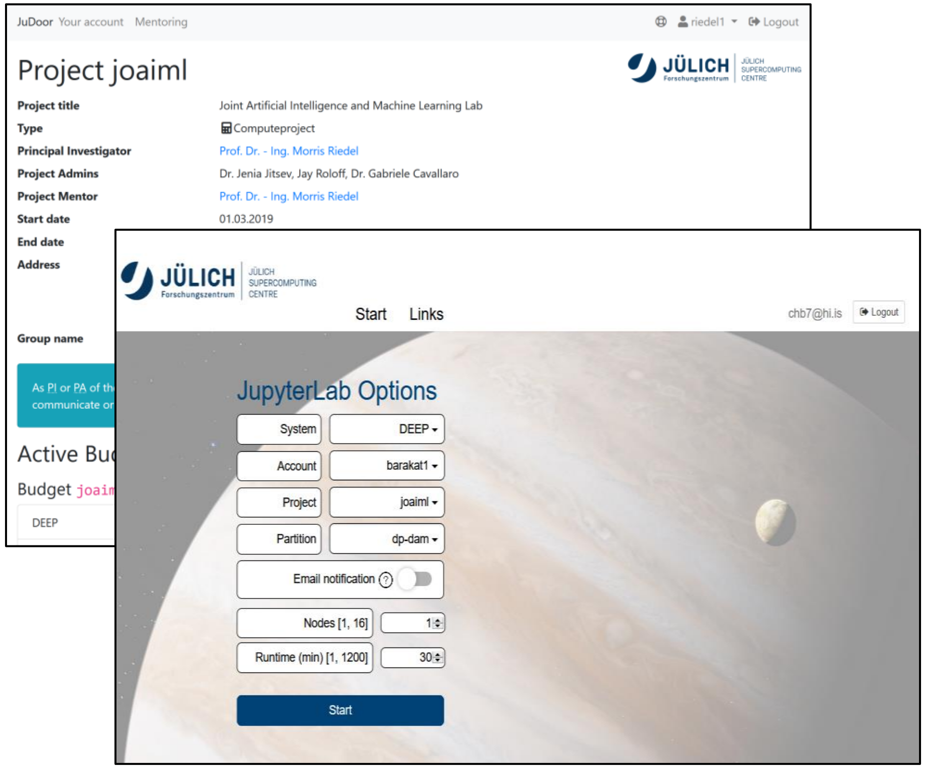
- Practical Topics
- Theoretical / Conceptual Topics

# Lecture Outline

- Jupyter-JSC: Bringing HPC to the Browser
  - Access to HPC Resources
  - Hardware Resources
- Applications of HPC in Healthcare
  - Different Approaches for Different Types of Data
  - RNN vs. CNN
  - LSTMs and GRUs
  - Sequence Data Use Case: ARDS
  - Image Processing Use Case: Covid-Net
  - Systems Biology and Biological Signalling Pathways
- HPC in Neuroscience
  - Neuroscience Expectation vs. Reality
  - Applications of Image Processing in Neuroscience
  - Neuroscience Technology Framework Use Case: The Big Brain Project

# Access to HPC Resources

- Access granted through an online account on JuDoor, and through specific projects with registered HPC budgets.
- JupyterLab: a browser-based modular development environment.
- Jupyter-JSC: a JupyterLab implementation with integrated access to HPC resources:
  - Expandable hardware model.
  - Pre-installed Machine Learning modules.
  - Out of the box GPU integration.
  - Access to remote storage clusters.




*https://judoor.fz-juelich.de/*

*https://jupyter-jsc.fz-juelich.de/*

# HPC Hardware Resources

- **Data Analytics Module (DAM)**
  - Specific requirements for data science & analytics frameworks
  - 16 nodes with 2x Intel Xeon Cascade Lake; 24 cores
  - 1x NVIDIA V100 GPU / node
  - 1x Intel STRATIX10 FPGA PCIe3 / node
  - 384 GB DDR4 memory / node
  - 2 TB non-volatile memore / node

- **DAM Prototype**
  - 3 x 4 GPUs Tesla Volta V100
  - Slurm scheduling system


*https://www.deep-projects.eu/*


*JUWELS Supercomputer*

# Icelandic HPC Community – Simulation & Data Lab Health & Medicine

IHPC

Jötunn | Garðar – decommissioned | History | Community | All Community Experts | Support | Acknowledgements | Hafðu samband – Contact us

Simulation and Data Lab Health and Medicine

## General information

The Simulation and Data Lab Health and Medicine (SimDataLab HM) aims to shed light on novel data analysis approaches in the medical field with extra focus on the application of High Performance Computing (HPC) architectures in the processing of patient medical data, as well as diagnosis and treatment assistance. The SimDataLab HM works in cooperation with the Juelich Supercomputing Centre (JSC) of Forschungszentrum Juelich (FZJ) – Juelich, Germany as part of the SMITH consortium's Algorithmic Surveillance of ICU Patients (ASIC) use case.

Prof. Dr. – Ing. Morris Riedel

Chadi Barakat

Seeking for new members from health & medicine experts that leverage HPC

[22] IHPC SimDataLab Health & Medicine Web Page

UKH Universitätsklinikum Halle (Saale)

Universitätsmedizin Essen

Universitätsmedizin Rostock

UNIKLINIK RWTH AACHEN

UKE Hamburg

UKD Universitätsklinikum Düsseldorf

ukb universitätsklinikumbonn

Universitätsklinikum Leipzig

SMITH Smart Medical Information Technology for Healthcare

[23] Alfred Winter, M. Riedel et al., 'Smart Medical Information Technology for Healthcare (SMITH): Data Integration based on Interoperability Standards', Journal of Methods of Information in Medicine, 2018

[25] SMITH Project Web Page

JOURNAL OF MEDICAL INTERNET RESEARCH | Maassen et al

Original Paper

Future Medical Artificial Intelligence Application Requirements and Expectations of Physicians in German University Hospitals: Web-Based Survey

Oliver Maassen[1,2], MSc; Sebastian Fritsch[1,2,3], MD; Julia Palm[2,4], MSc; Saskia Deffge[1,2], MSc; Julian Kunze[1,2], MD; Gernot Marx[1,2], MD, Prof Dr, FRCA; Morris Riedel[2,3,5], Prof Dr; Andreas Schuppert[2,6], Prof Dr; Johannes Bickenbach[1,2], MD, Prof Dr

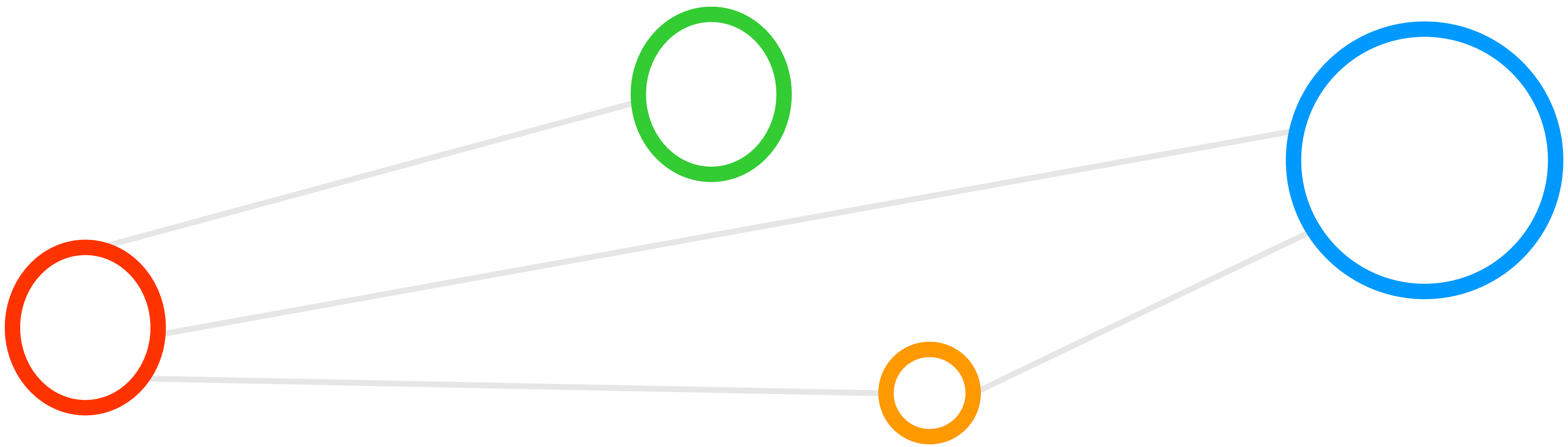
[1]Department of Intensive Care Medicine, University Hospital RWTH Aachen, Aachen, Germany
[2]SMITH Consortium of the German Medical Informatics Initiative, Leipzig, Germany
[3]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[4]Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Jena, Germany
[5]School of Natural Sciences and Engineering, University of Iceland, Reykjavik, Iceland
[6]Institute for Computational Biomedicine II, University Hospital RWTH Aachen, Aachen, Germany

[24] O.Maassen et al., Future Medical Artificial Intelligence Application Requirements and Expectations of Physicians in German University Hostpitals: Web-based Survey, Journal of Medical Internet Research, 2021

relatively low HPC & AI usage still, strict regulations for AI

data silos: no data sharing, GDPR & reiterating clinical studies

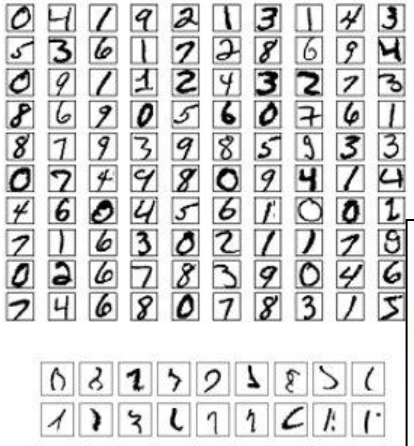
# Applications of HPC in Healthcare



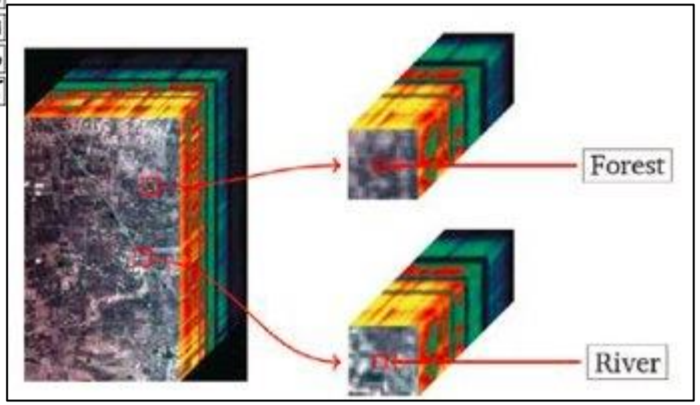
# Different Approaches for Different Data Types

- Input data for machine learning applications come in different types:
  - Numerical Data
  - Image Data (single- or multi-dimensional)
  - Sequence Data (text, sound, seismic waves, physiological signals)
- Specific Machine Learning approaches have been developed to take advantage of each data type.
- One ML approach that is effective on one datatype may not be effective for another.


***Sequence Data***



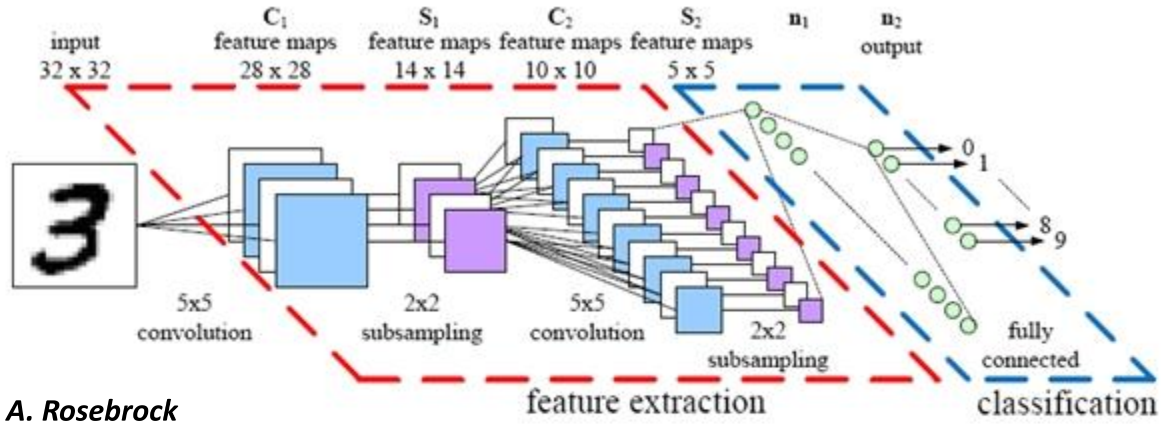
***MNIST Dataset***




***Hyperspectral Data***

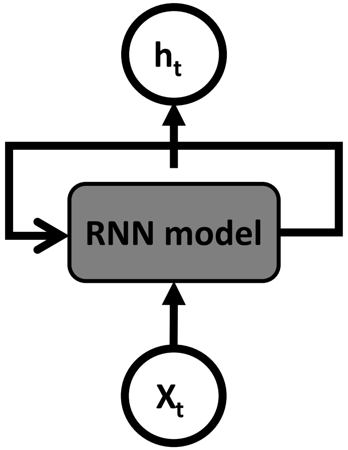
# Recurrent Neural Networks (RNN)

- Convolutional Neural Networks (CNNs)
  - Example: remote sensing application domain, hyperspectral datasets
  - Neural network key property: exploit spatial geometry of inputs
  - Approach: Apply convolution & pooling (height x width x feature) dimensions



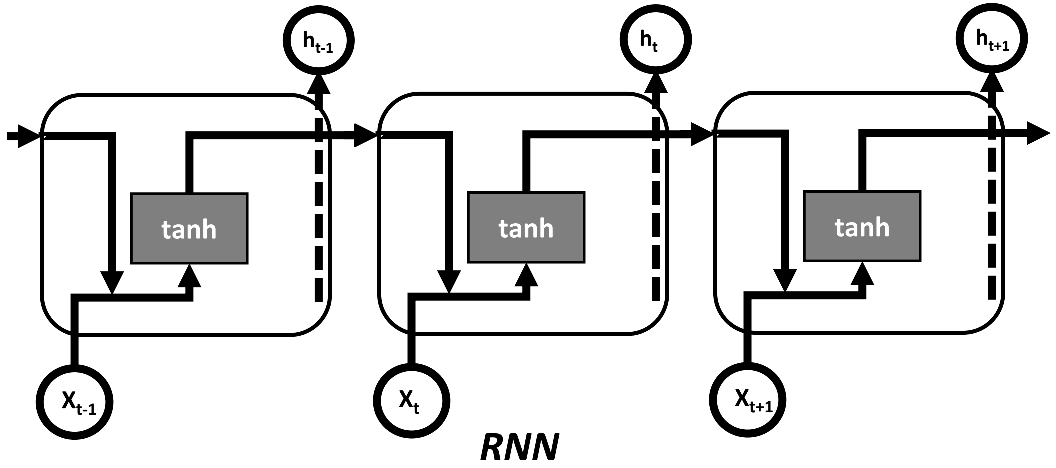

*[7] A. Rosebrock*

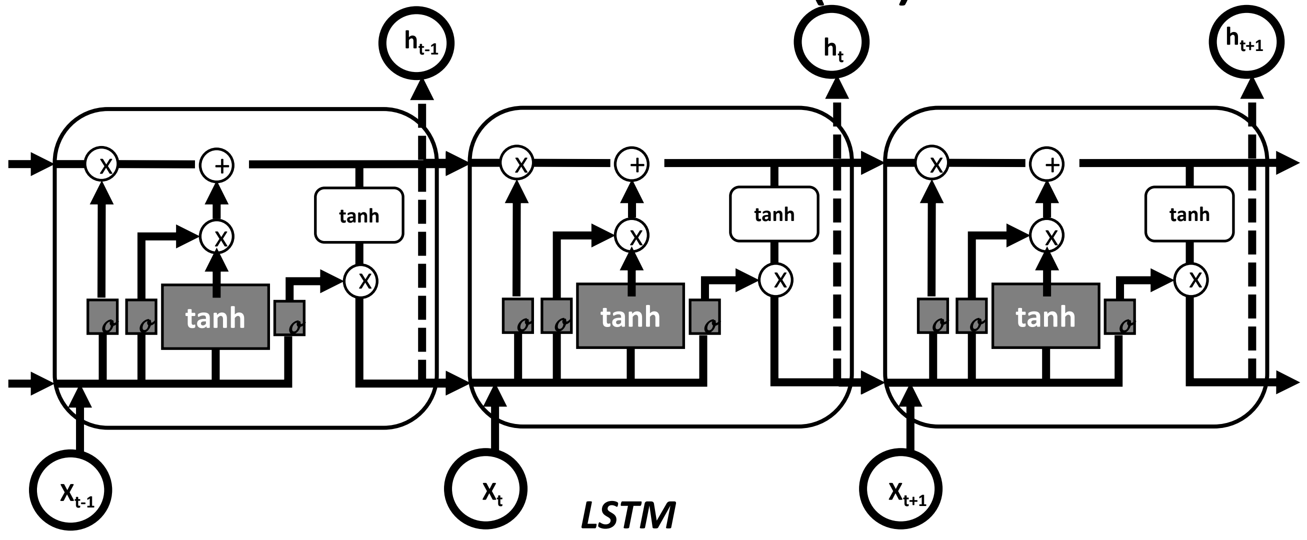
- RNNs
  - Examples: texts, speech, time series datasets
  - Neural network key property: exploit sequential nature of inputs
  - Approach: Train a graph of 'RNN cells' & each cell performs the same operation on every element in the given sequence

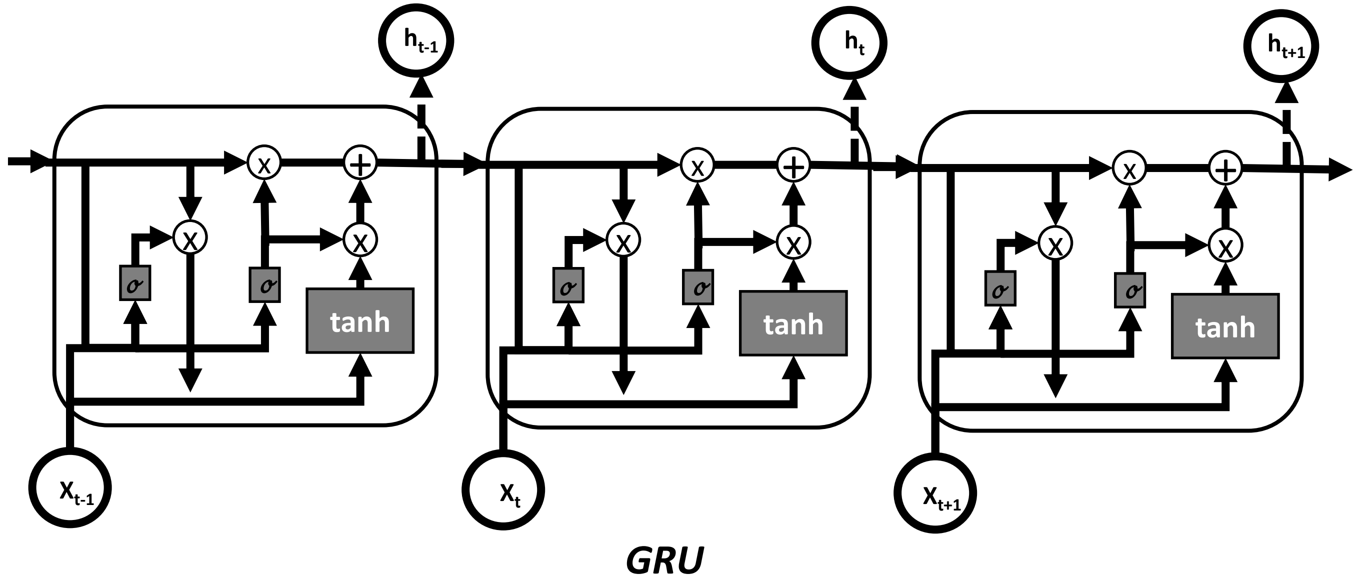



- **RNNs are used to create sequence models whereby the occurrence of an element in the sequence (e.g. text, speech, time series) is dependent on the elements that appeared before it**

# Advanced Applications of RNNs (1)



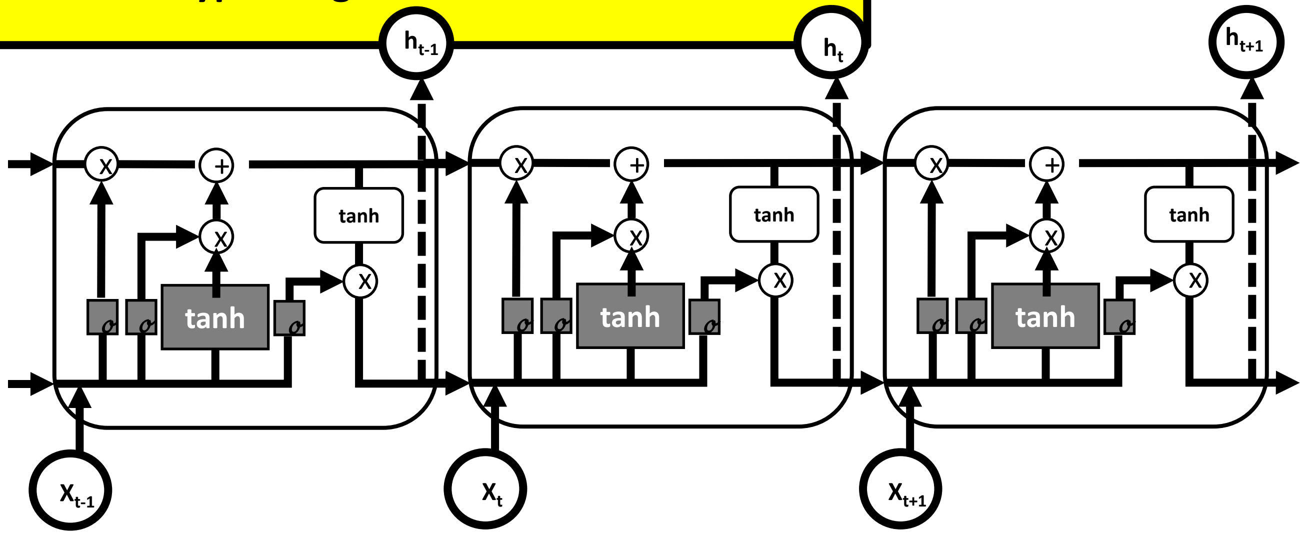



# Advanced Applications of RNNs (2)

- Long Short Term Memory (LSTM) networks are a special kind of Recurrent Neural Networks (RNNs) that learn long-term dependencies in data by remembering information for long periods of time.
- LSTM introduce a ‘memory cell‘ structure into the underlying basic RNN architecture using four key elements: an input gate, a neuron with self-current connection, a forget gate, and an output gate
- The data in the LSTM memory cell flows straight down the chain with some linear interactions (x,+)
- The cell state $s_t$ can be different at each of the LSTM model steps & modified with gate structures
- Linear interactions of the cell state are pointwise multiplication (x) and pointwise addition (+)
- In order to protect and control the cell state $s_t$ three different types of gates exist in the structure

- Gated Recurrent Units (GRUs) are a simpler version of LSTMs that offer comparable performance with reduced computational cost.



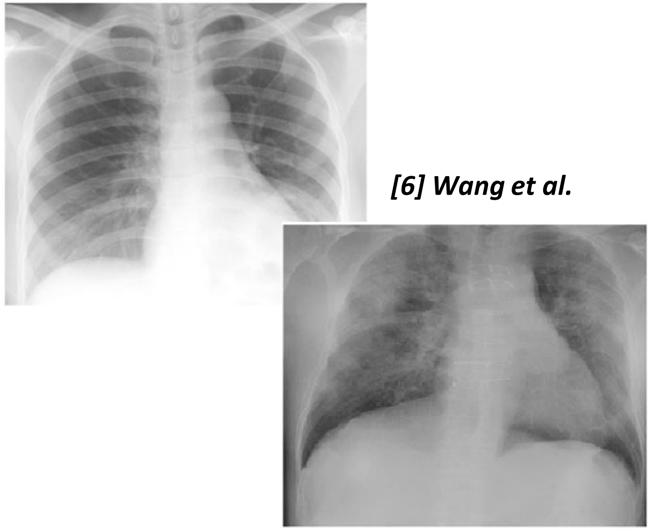
# Medical Timeseries Data Analysis

- Sequence data, by definition, entails that the order of the data is important, but also that future values depend on past values.
- Medical data (heart rate, blood oxygen levels, drug concentrations…) is a specific example where it is extremely beneficial to be able to draw predictions and diagnosis from timeseries data.
- These timeseries are usually quite long, spanning days or weeks, and far too difficult to be analysed.
- Digitisation made it possible to store this data in Electronic Health Records (EHRs), thus building large databases full of information to be mined and to develop diagnosis and treatment methods.

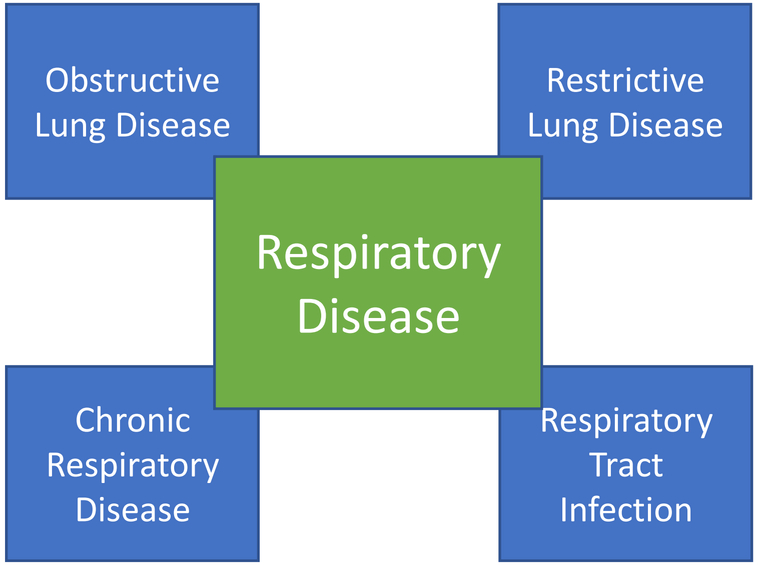

*https://www.philips.de/healthcare/*

# Respiratory Disease

- Respiratory diseases can have several causes including trauma, viral or bacterial infections, etc.
- Their effects are wide-ranging, including blood acidosis as oxygen levels in blood drop, increased heart rate, decrease blood pressure and a cascade of events that can lead to multi-organ failure.
- Treatment usually begins with mechanical ventilation which also causes stress to the lungs potentially causing injury and subsequently collapsed compartments or thrombosis.
- ICU staff usually have protocols to deal with lung injury but they are very subjective, and can vary within the same institute, and from one hospital to another.



*[6] Wang et al.*



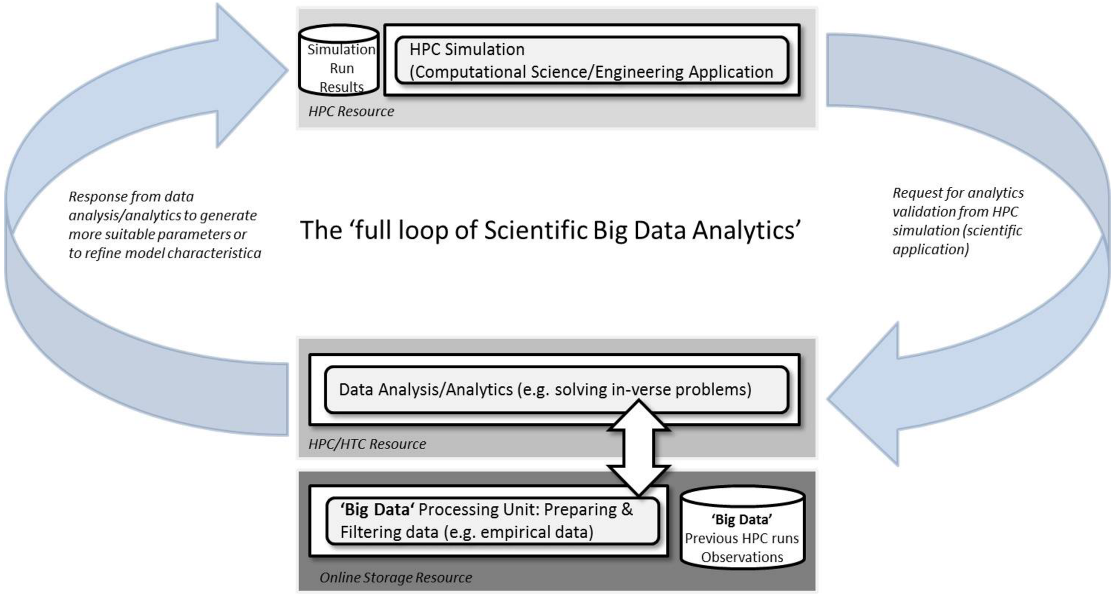
# Use Case: SMITH and ARDS

- Acute Respiratory Distress Syndrome (ARDS) is a rare condition that affects ICU patients, but has a high mortality rate.
- There is consensus on how to diagnose the condition, but not how to treat it.
- This is one of the use cases of the Smart Medical Information Technology for Healthcare (SMITH) consortium grouping major research institution in Germany.
- The aim is to develop algorithms that can efficiently and accurately diagnose the onset of ARDS, and potentially provide suggestions for treatment.

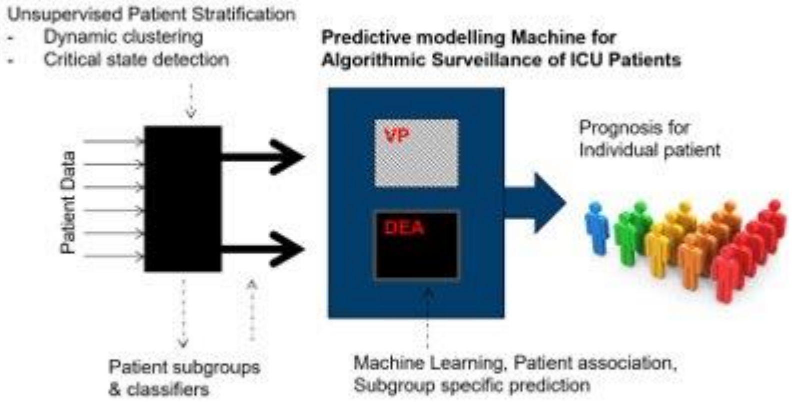
# Where HPC Comes into the Equation

- For a Mechanistic model to be able to diagnose and treat, it needs to learn and that requires data:
  - The largescale analysis, cleaning, and preparation of data requires well adapted resources for storage and processing.
  - Running multiple simulations with each patient's data to understand how small changes in parameters affect their overall state take up a lot of time.
  - Exporting this data and using it to train a numerical model requires efficient processing resources.
- Solution: use available HPC Resources!

- **Simulation and analysis have to go hand in hand in order to build proper modelling techniques.**
- **HPC simplifies and accelerates the process.**

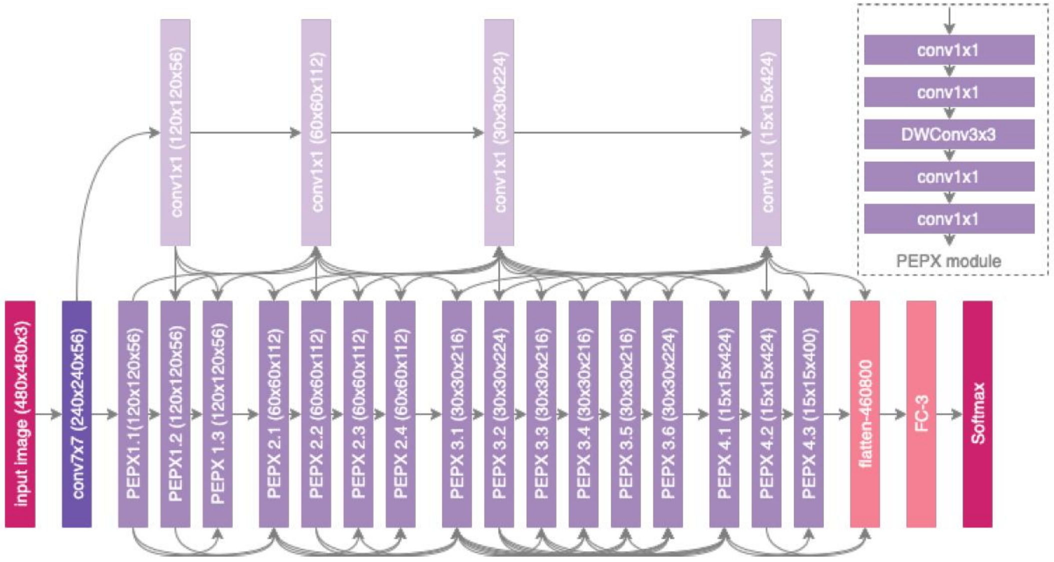



***[8] Th. Lippert & M. Riedel et al.***



# Use Case: COVID-NET

- Given the Covid-19 pandemic, a lot of research was done in order to provide effective screening of Covid-infected patients.
- COVID-NET is a CNN trained on a database of Chest X-Ray (CXR) images in order to distinguish between healthy, pneumonia, and Covid-19 patients.
- The results were compared to VGG-19 and ResNet-50 and highlighted the effectiveness of COVID-NET.




*[6] Wang et al.*

***Healthy Patient***

***Covid-19 Patient***

# Where HPC Comes into the Equation

- The original CXR database contained 13.975 images divided as show in Table 1.
- Proposed project was to confirm the initial results, and then apply transfer learning by training the network on new data from a different dataset.
- Data provided by E-Healthline consists of 1.066 images for training and 4.115 images for testing.
- Partial use of the data to test out the approach shown in Table 2.
- Data storage and training done using the resources available on the DEEP cluster and eventually on the JUWELS cluster (Fastest Supercomputer in Europe).

***Table 1: COVIDx Database Image Distribution***

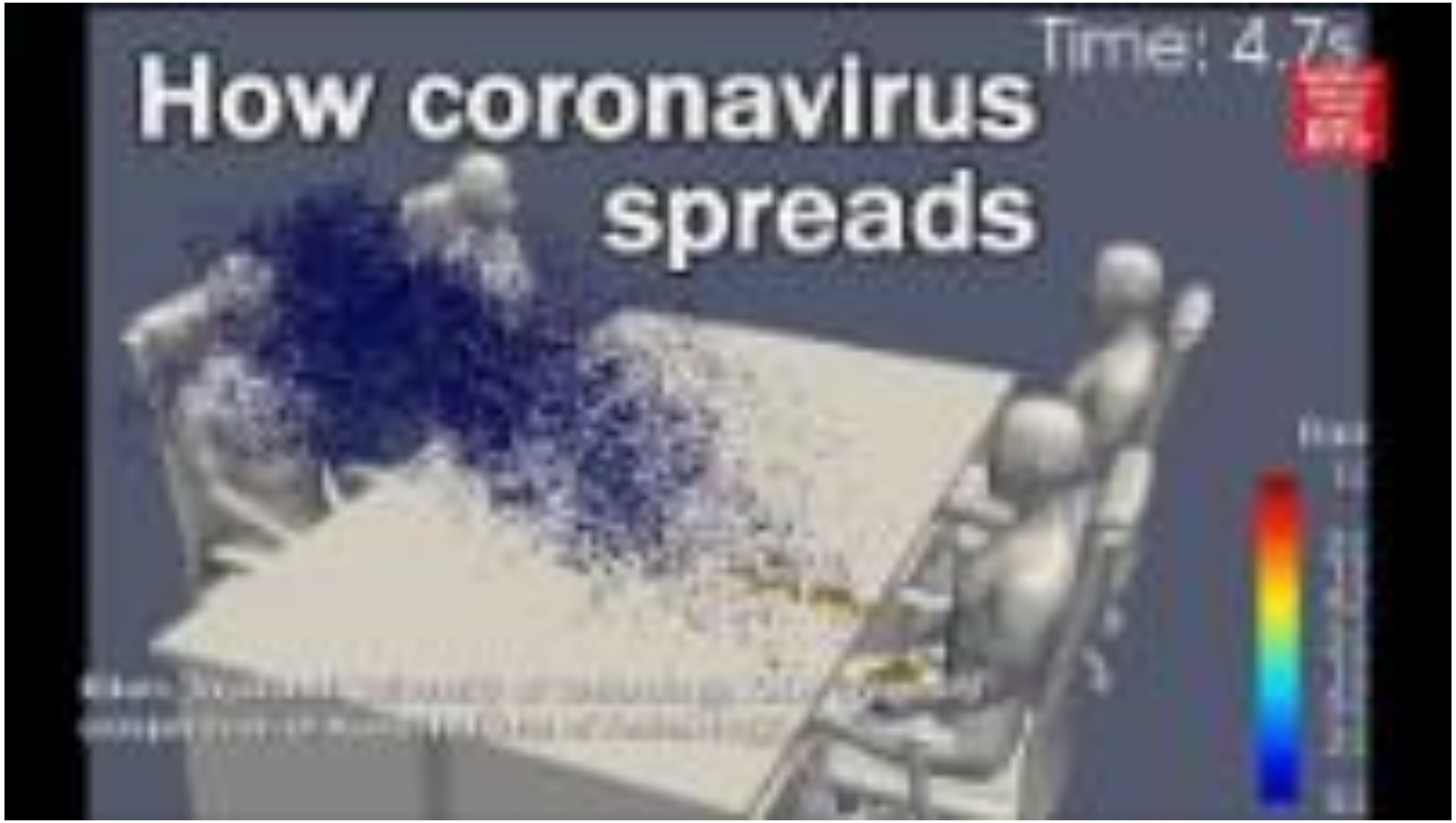
| | Healthy | Pneumonia | Covid-19 |
|---|---|---|---|
| # of Images | 8.066 | 5.538 | 358 |

- **Transfer Learning: applying a pretrained network on a new problem that has new data.**

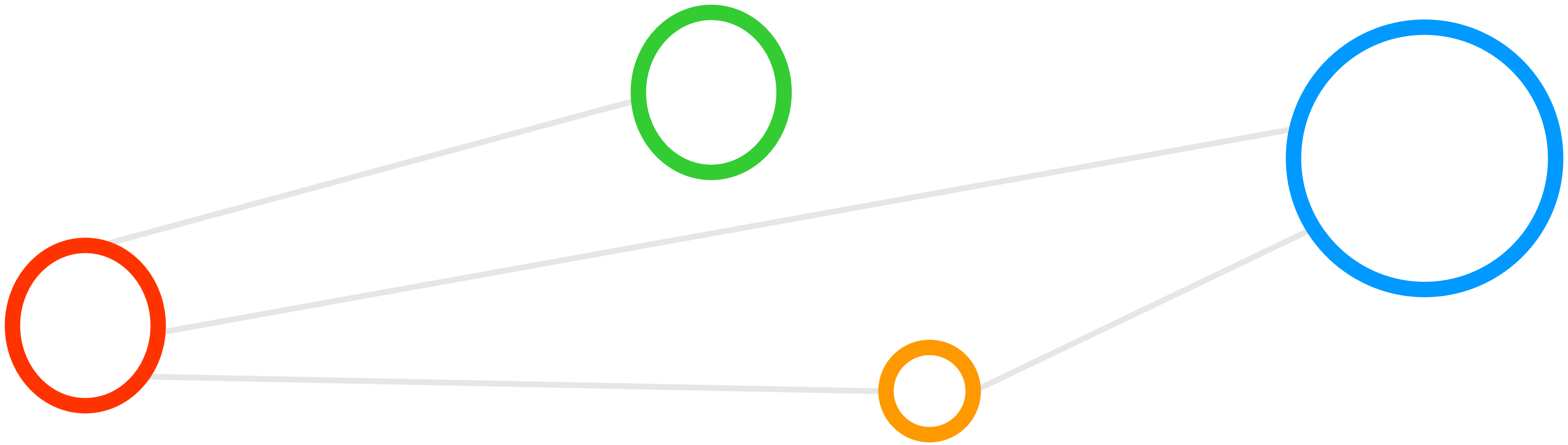
***Table 2: E-Healthline Database Image Distribution***

| | Healthy | Pneumonia | Covid-19 |
|---|---|---|---|
| Train | 198 | 21 | 85 |
| Test | 1.700 | 97 | 101 |
| Total | 1.898 | 118 | 186 |

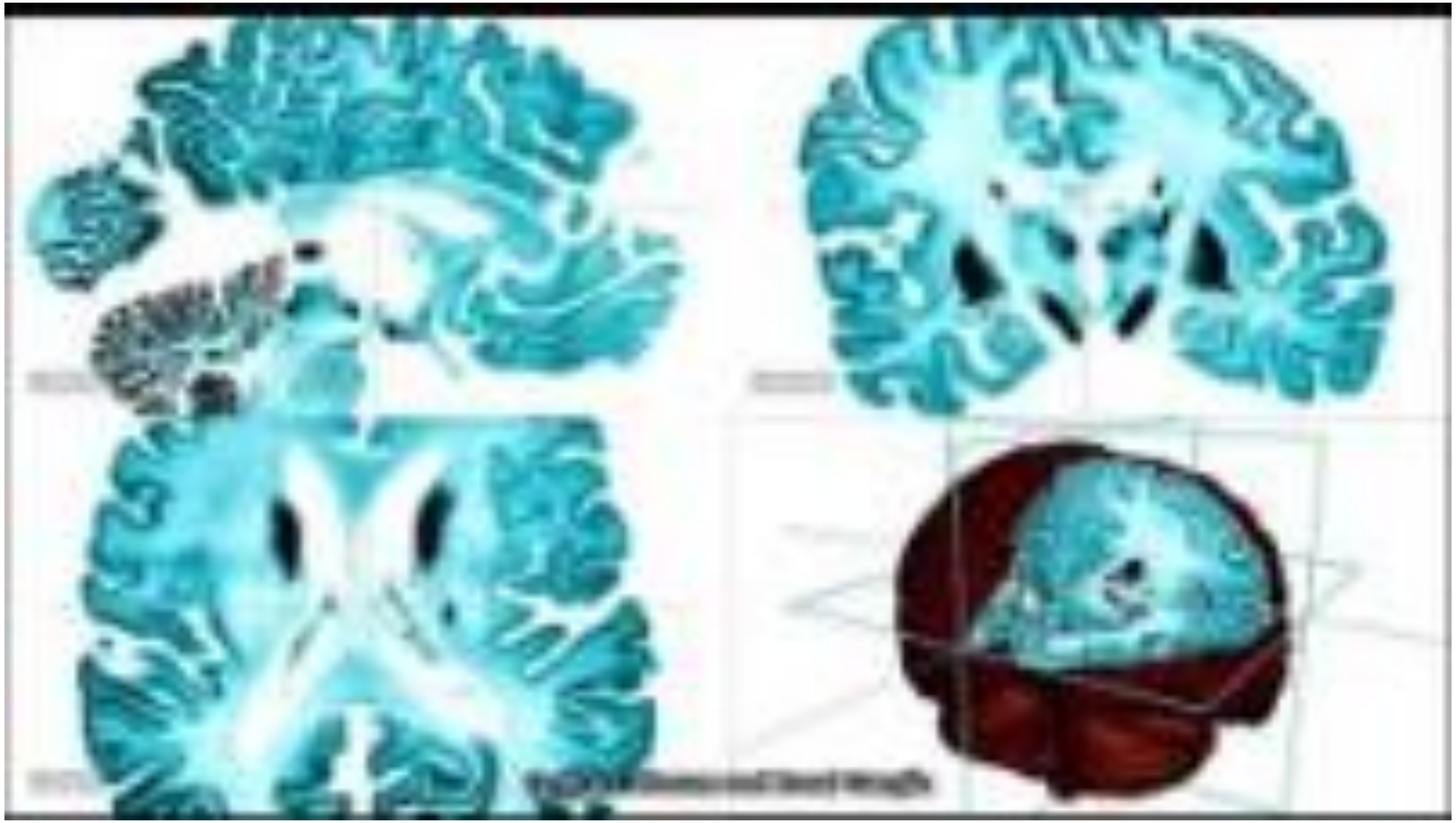
# Current Real World Applications



# HPC in Neurosciences

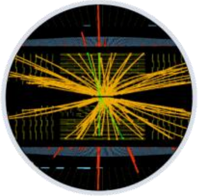


# The Face of Neuroscience



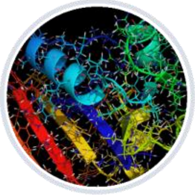
# Impacts of Artificial Intelligence in HPC Design

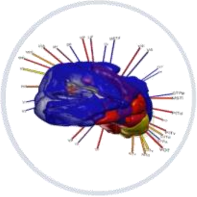
## Co-Design via Requirements from Machine/Deep Learning Applications & Innovative Simulation Sciences


High Energy Physics

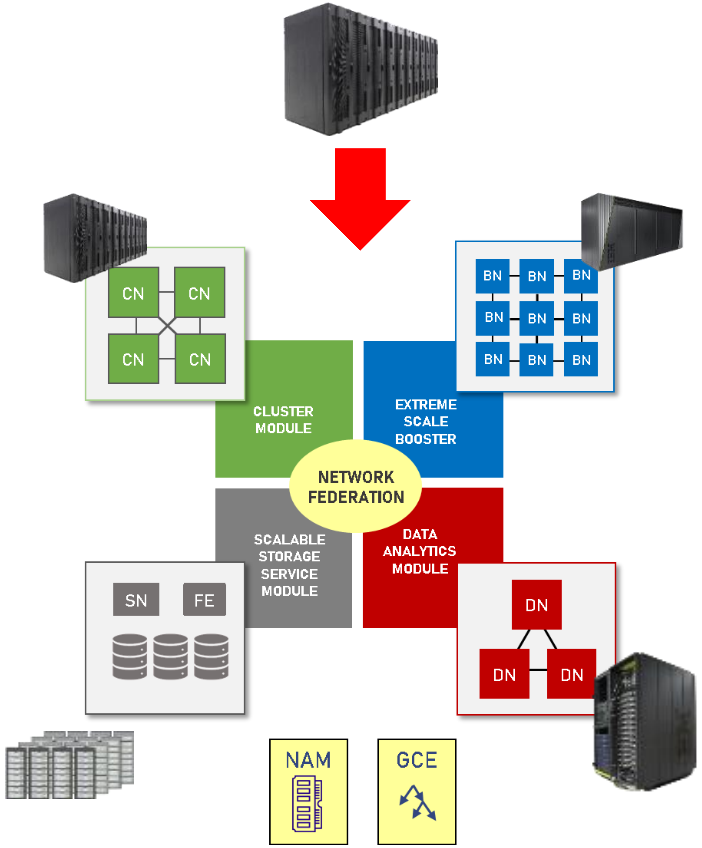

Earth Science

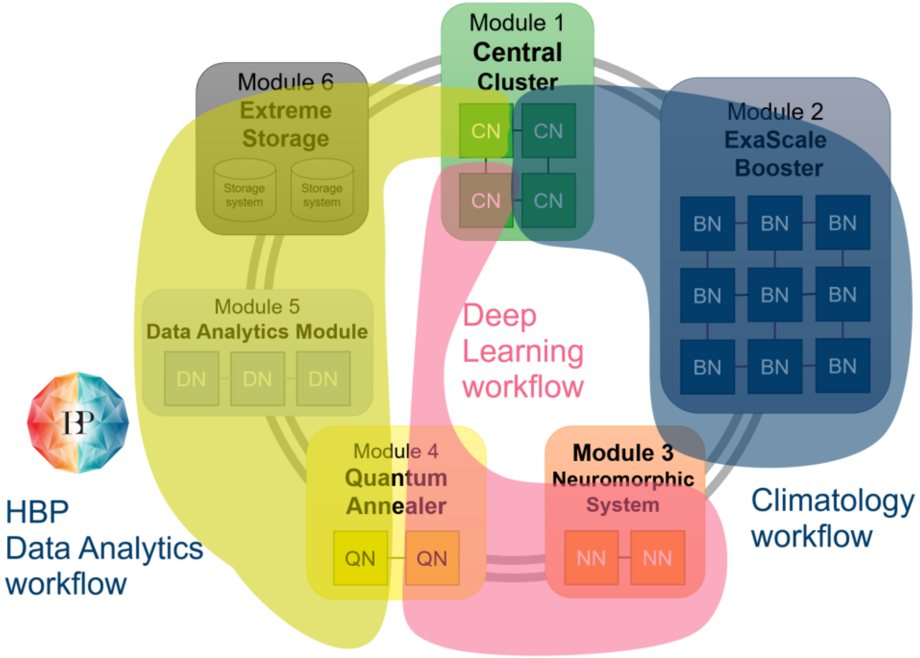

Space Weather


Molecular Dynamics


Neuroscience


Radio Astronomy



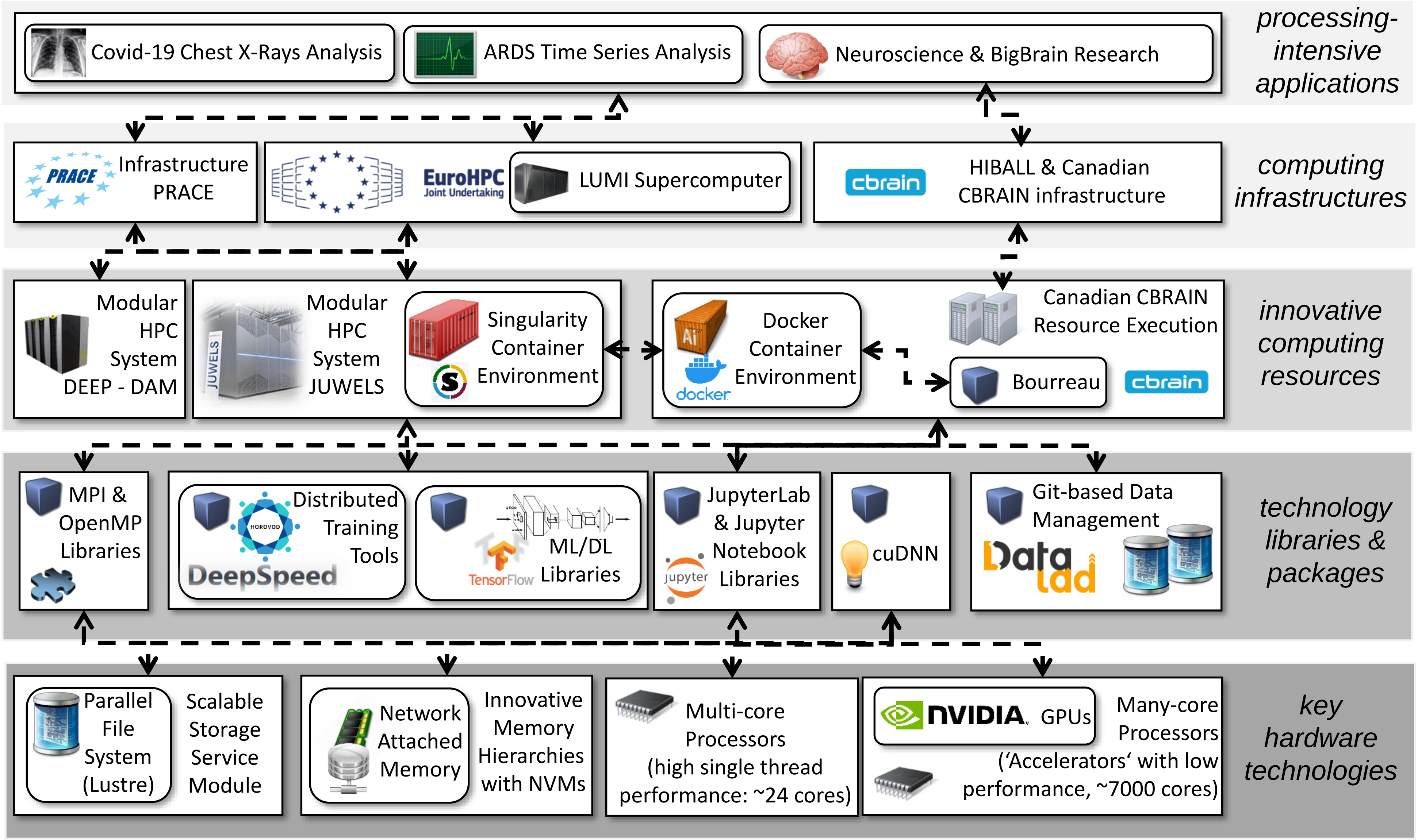




DEEP Projects

*https://www.deep-projects.eu/*

- **Neuroscience was one of the core concepts of the HPC design.**

# Data Science Platform for Health



# How a Data Science Platform Comes Together

- Before we use HPC for Neuroscience applications, we need to set up the infrastructure for data storage, management, and processing.
- Before employing compute and storage resources, we need to establish protocols for data and software communication and versioning.
- Before that, we need to have the proper software modules for the task at hand.
- **But first and foremost, we need to have the proper hardware resources that can load this software, communicate the data, and perform the required tasks.**

*processing-intensive applications*

*computing infrastructures*

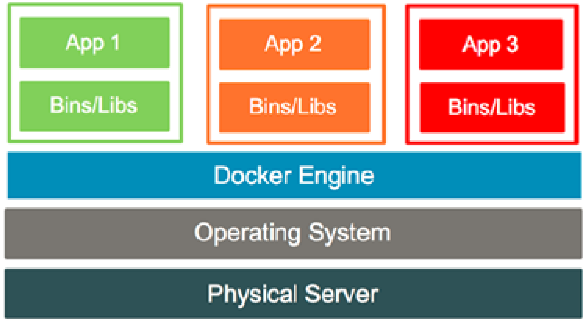
*innovative computing resources*

*technology libraries & packages*

*key hardware technologies*

# Container Management – Docker Example

- Open-Source 'containerization of software' tool
  - Docker container enable a software to be ready-to-run
  - 'Container images' contain everything that is needed to run: source code, runtime, system tools, specific libraries, data, etc.
  - Enables flexibility and portability on where the application software is able to run ('towards a standard')
  - Basis for specific offerings (e.g. Ubercloud & Engineering)


docker

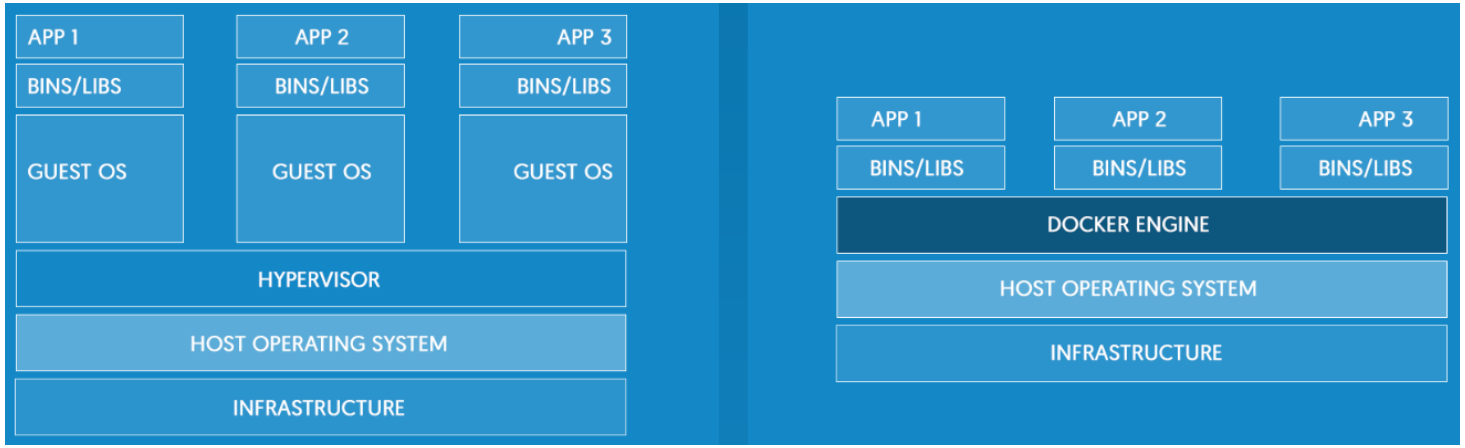
*[9] Docker Web page*

*[10] Ubercloud CFD*

- **The core idea of Docker is to provide a software container with all required software elements that guarantees that the application within the software container will always run the same way, regardless of the environment it is running in or which cloud infrastructure is used underneath**
- **Docker is an open-source tool that automates the deployment of applications within so-called software containers that can be bundled as a Docker image and broadly used in Clouds today**
- **Docker enables an easier migration for applications from one cloud to another**

# Virtualization vs. Container Approaches


- Virtualization (high storage footprint)
  - VMs include application, binaries, libraries and an entire guest operating system (≈tens of GBs)
- Containers (low storage footprint, vendor-lock free)
  - Include application, all dependencies, runs isolated, but share kernel of the operating system with other containers (vendor independence)




*[9] Docker Web page*

*[10] Ubercloud CFD*

# Software and Data Versioning – DataLad Example

## Discover Data

DataLad has built-in support for **metadata** extraction and **search**. With only a few steps, you can search through a large collection of readily available datasets and immediately download them. See more...

## Consume Data

DataLad offers direct **access to individual files** — great when you only need a few files from some large datasets for an analysis. Files in a dataset can be distributed across multiple download sources with tailored permissions to match your **data privacy** needs. See more...

DataLad can help with small or large-scale data management

https://www.datalad.org/

## Publish Data

DataLad supports sharing datasets with the **public or just some colleagues** on platforms that you are using already — **no need for a central service**. You have complete freedom to share your work in multiple platforms simultaneously (your own server, DropBox, GitHub, etc.) without losing track. See more...

## Reproducibility

DataLad provides **joint management of analysis code *and* data**. This enables you to comprehensively track the exact state of any analysis inputs that produced your results — across the entire lifetime of a project, and across multiple datasets. See more...

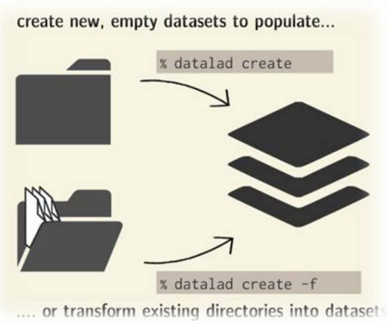
## Data Portal

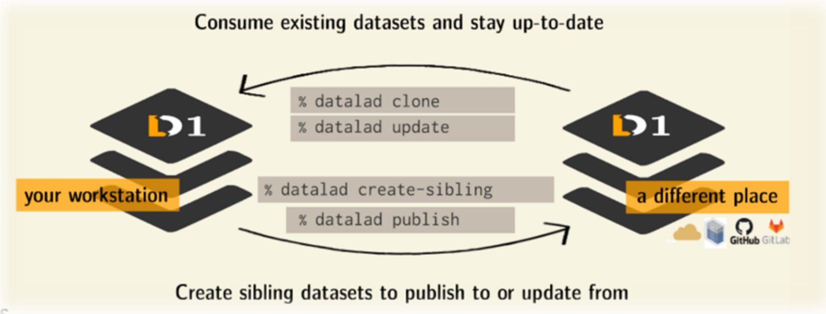
The DataLad project operates a crawler that regularly indexes datasets from scientific data portals such as OpenNeuro and CRCNS, making them trivial to acquire and work with using DataLad. Take a look at the available datasets.

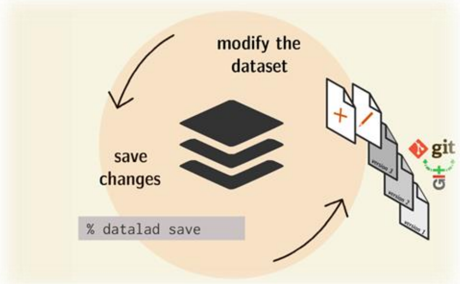
➢ **Great source: DataLad itself offers a massive amount of information about the technology and its usage: handbook.datalad.org**

# DataLad – How it Works

- Free open-source data management tool





https://www.datalad.org/

- Usable via command-line
  or for developers also via APIs in Python

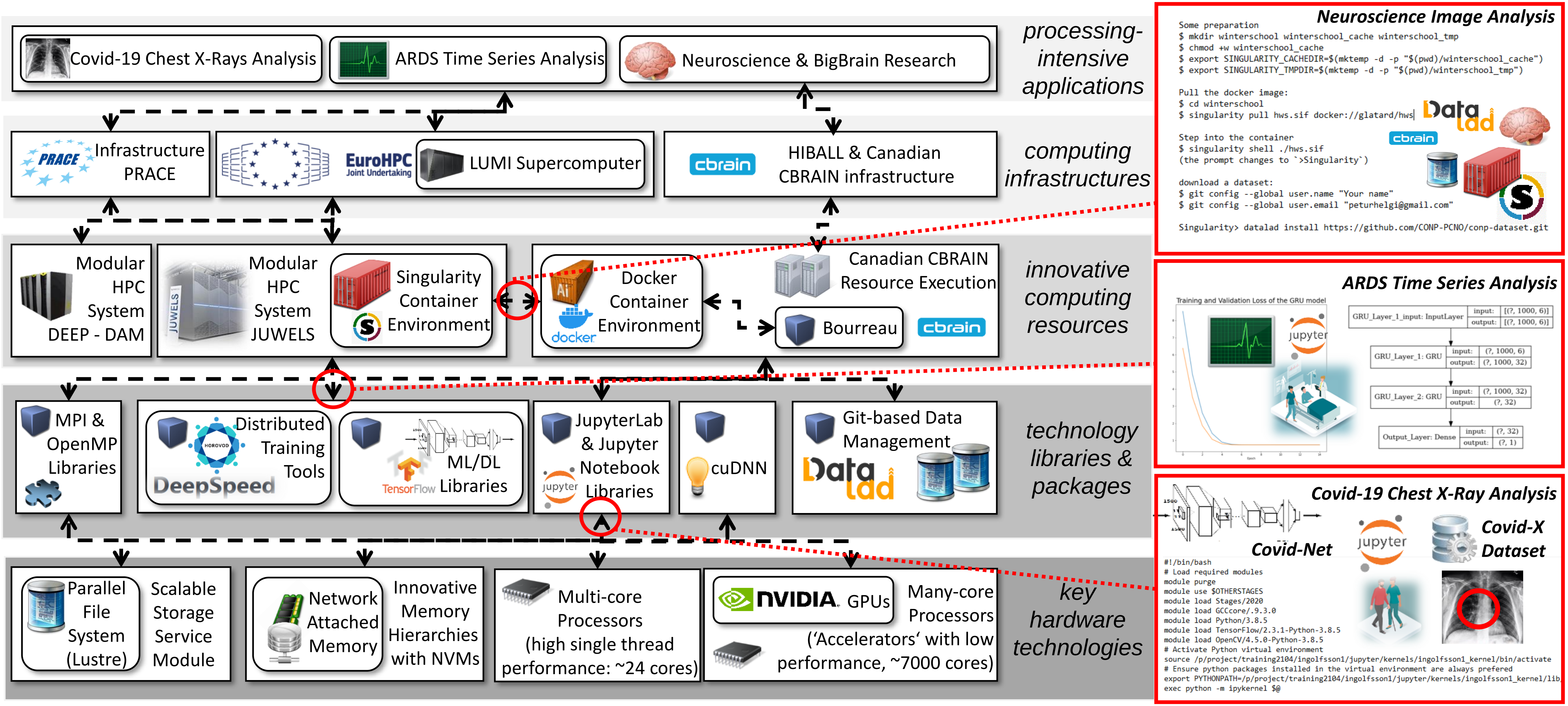
```
datalad create mydataset
```

```
import datalad.api as dl
dl.create(path="mydataset")
```

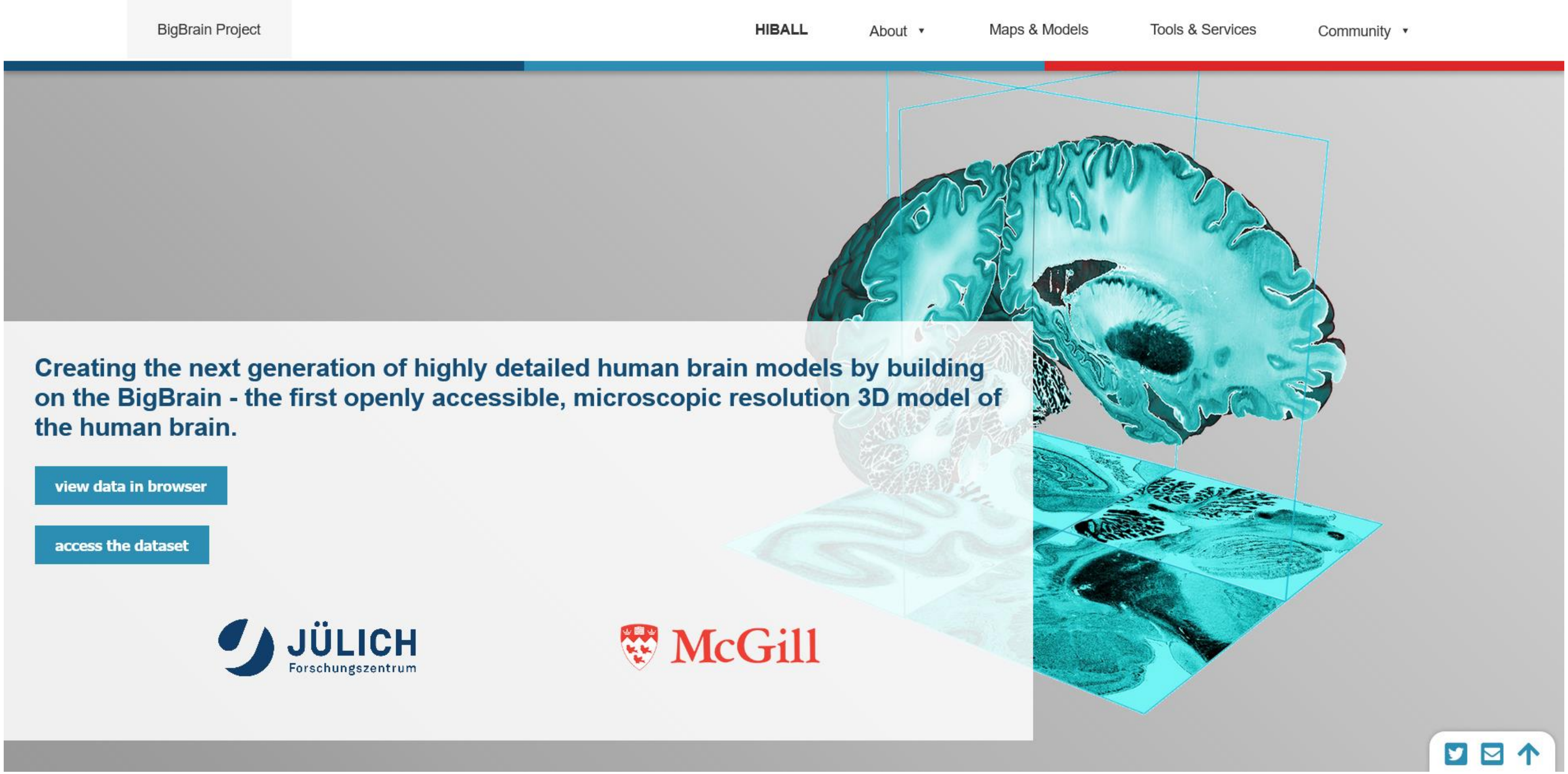
- **Built on top of Git & Git-annex versioning tool**
- **Version-controlling arbitrarily large content: version control data & software alongside to code**
- **Transport mechanisms for sharing & obtaining data: consume & collaborative on data analysis like software**
- **Computationally reproducible data analysis: track & share provenance of all digital objects**
- **Interesting for large datasets (e.g. CFD simulations)**

➢ **Great source: DataLad itself offers a massive amount of information about the technology and its usage: handbook.datalad.org**

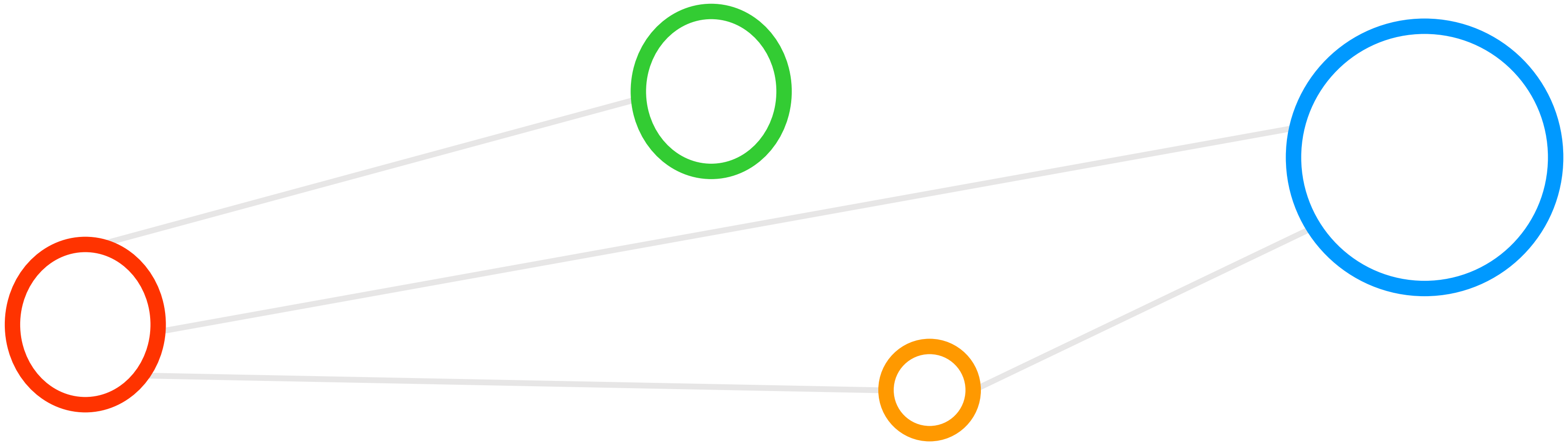
# How it all Comes Together on the Platform



# The Big Brain Project



# Lecture Bibliography



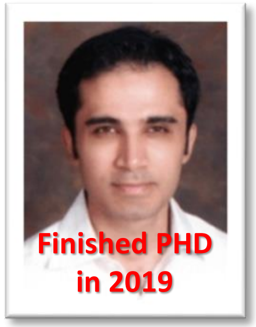
# Lecture Bibliography (1)


- [1] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *29th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, June 2016, doi:10.1109/CVPR.2016.90
- [2] Python
Online: https://www.python.org/
- [3] Keras
Online: https://keras.io/api/applications/
- [4] F. Chollet, “Transfer Learning and Fine-Tuning”
Online: https://keras.io/guides/transfer_learning/
- [5] Cheng, A.C, Lin, C.H., Juan, D.C., InstaNAS: Instance-aware Neural Architecture Search, Online:
https://arxiv.org/abs/1811.10201
- [6] L. Wang, A. Wong, 'COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images’
Online: https://arxiv.org/abs/2003.09871
- [7] A. Rosebrock, ‘Get off the deep learning bandwagon and get some perspective‘,
Online: http://www.pyimagesearch.com/2014/06/09/get-deep-learning-bandwagon-get-perspective/
- [8] Th. Lippert, D. Mallmann, M. Riedel, ‘Scientific Big Data Analytics by HPC’, *Publication Series of the John von Neumann Institute for Computing (NIC) NIC Series 48*, 417, ISBN 978-3-95806-109-5, pp. 1 - 10, 2016
- [9] Docker Web page – What is a Container, Online:
https://www.docker.com/resources/what-container
- [10] LinkedIn – UberCloud Compendium of 39 Case Studies in Computational Fluid Dynamics, Online:
https://www.linkedin.com/pulse/2020-ubercloud-compendium-39-case-studies-fluid-wolfgang-gentzsch/
- [11] DataLad
Online: https://www.datalad.org/
- [12] Deep Projects
Online: https://www.deep-projects.eu/
- [13] Big Brain Project
Online: https://bigbrainproject.org/index.html
- [14] Icelandic HPC Simulation and Data Lab Health & Medicine, Online:
https://ihpc.is/simulation-and-data-lab-health-and-medicine/
- [15] Alfred Winter, A. Schuppert, M. Riedel et al., ‘Smart Medical Information Technology for Healthcare (SMITH) – Data Integration based on Interoperability Standards’, Journal of Methods of Information in Medicine, 57 (S 01), e92-e105, 2018, Online:
https://www.researchgate.net/publication/326465733_Smart_Medical_Information_Technology_for_Healthcare_SMITH
- [16] O.Maassen et al., Future Medical Artificial Intelligence Application Requirements and Expectations of Physicians in German University Hostpitals: Web-based Survey, Journal of Medical Internet Research, 2021, Online:
https://www.researchgate.net/publication/349343535_Future_Medical_Artificial_Intelligence_Application_Requirements_and_Expectations_of_Physicians_in_German_University_Hospitals_Web-Based_Survey
- [17] Alfred Winter, A. Schuppert, M. Riedel et al., ‘Smart Medical Information Technology for Healthcare (SMITH) – Data Integration based on Interoperability Standards’, Journal of Methods of Information in Medicine, 57 (S 01), e92-e105, 2018, Online:
https://www.researchgate.net/publication/326465733_Smart_Medical_Information_Technology_for_Healthcare_SMITH

# Acknowledgements – High Productivity Data Processing Research Group


PD Dr.
G. Cavallaro


Senior PhD
Student A.S. Memon




Senior PhD
Student M.S. Memon

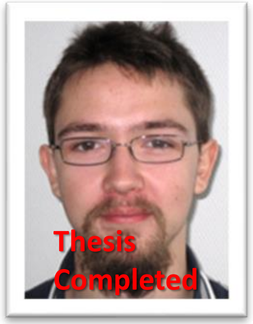

PhD Student
E. Erlingsson


PhD Student
S. Bakarat

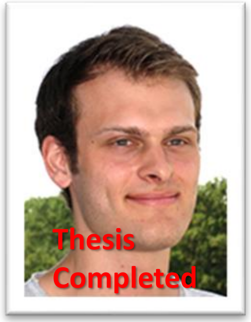

PhD Student
R. Sedona



Dr. M. Goetz
(now KIT)




MSc M.
Richerzhagen
(now other division)




MSc
P. Glock
(now INM-1)


MSc
C. Bodenstein
(now
Soccerwatch.tv)

MSc Student
G.S. Guðmundsson
(Landsverkjun)


This research group has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 763558 (DEEP-EST EU Project)


services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
year
Computational
Test
E.g
Computing
disciplines
results
Approaches
construction
Fusion
Enable
like
Cross-Disciplinary
methods
big
nodes
PB
community
Environment
Modelling
different
EU
Science
High
centers
Key
scientific
important
files
model
international
etc
DKRZ
Use
storage
basis
Infrastructure
simulations
ENES
technologies
formats
brain
physical
engineering
References
increasing
via
EUDAT
processing
IPCC
Common
TB
Big
Project
Data
MapReduce
create
computing
Using
Methods
European
SMAQ
databases
codes
GB
Supercomputing
Work
images
analysis
Large
Summary
data-intensive
Service
Scientific
levels
often
DLCL
stored
Resources
analyze
external
use
JSC
computational
Level
Research
Matters
Concrete
performance
research
e.g
science
HPC
Juelich
system
datasets
acquisition
large-scale
compute
Landscape
management
System
Climate
modelling
AR5
Hardware
distributed
cores
access
hundreds
Nutshell
Services
Technologies
Earth
Structure
current
network
DLCLs
Understanding
structures
directory
General
Germany
Simulation
project
Provide
NASA
Energy
systems
day
Processing
Health