HPS

Julian Kunkel, Sadegh Keshtkar, Stefanie Mühlhausen, Hendrik Nolte

A Computer Science Perspective on Data-Driven Healthcare

Outline

1 Introduction

2 KISSKI

3 ML in Medicine

4 Secure HPC

5 Data Management

6 Conclusions

Data-Driven Healthcare - Personal Perspective

- Big data analytics and AI models
  - Models for prediction/prescription aiding patients, expert systems for doctors
  - May want to run many models at the same time to get full picture/opinions
  - Personal wish: personalized health monitor
    a-priori indicator, e.g., Likely to get a stroke the next week, do X
    Make my health quantitatively objective to me - like a HUD in a 3D shooter
- **More data** will lead to better analysis - more instruments/sources captured
  - Clinical measurements, imaging data (MRI), other instruments
  - Subjective assessment, medical reports
  - Sensors in house, cooking devices, smart wear, ...
  - Incorporate external influences, e.g. weather, air, ...
- We could also aim to simulate a human ... with HPC and AI
  - Use physical/chemical models to predict system behavior / overall health
  - Multi-scale, from genetic information, statistics, to e.g. cardiovascular systems

Simulation and Big Data Analytics

Digital Twin for a human and personalized medicine

- Conduct what if simulations on yourself
- Simulate 1000 virtual humans with 50 years of smoking or vegan food

Compute and data requirements will rise

- Storing/analyzing a single MRI data set e.g. 250 MiB
- Storing time series, e.g., annual MRI scans of regions, statistics
- Data for Million people is necessary to built comprehensive models

Ethical questions

- How much would your health be worth for you?
  - Pay X% of income for +Y years of live time? Healthier life?
- How much can society afford for the health treatment?
- For me: it’s not ethical to not salvage and exploit the data treasure

Importance of Data Management and Visualization

Concerns of users

- Employing AI analysis into workloads - key for data-driven science
- Utilizing data analysis (e.g., via visual analytics - interactively)
- Ensuring reproducibility - confidence
- Documenting procedures (lab notebook)
- Performant data access
- Managing large number of data sets
  organization, data lifecycle management
- Usability - ease of use, tight integration of compute + storage
- Data sharing (with FAIR principles)
- Complex system software stack and landscape

AI Workflow Challenges

- Single model: Repeat many steps (until accuracy/loss is good enough)
  - Create a random subset of the data
  - Compute “error” of the data on the current model
  - Update weights in the model based on the error
- Challenge: 100k very small files, latency matters

  Need to package data together into suitable format/container
- Challenge: 100k large files, causing bandwidth issues

  Need smarter data selection and storage systems
- Challenge: creating big models with 100M parameters

  Need a scalable infrastructure to train models
- Challenge: Too little data available (e.g. rare disease)

Outline

1 Introduction

2 KISSKI

3 ML in Medicine

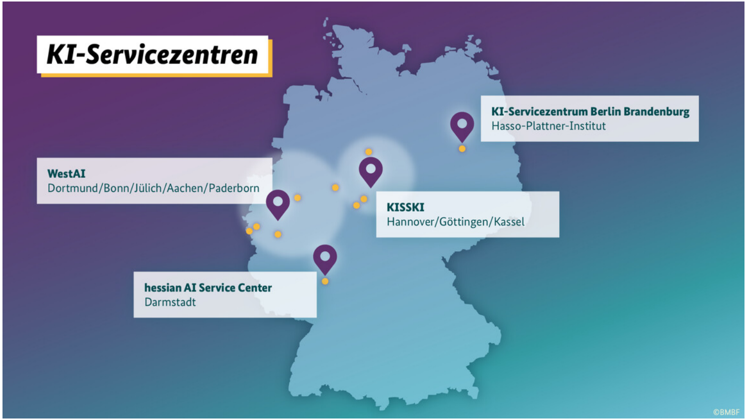
4 Secure HPC

5 Data Management

6 Conclusions

The four BMBF-funded, national AI service centres

The KISSKI - AI Service Centre


KI-Servicezentren
KI-Servicezentrum Berlin Brandenburg
Hasso-Plattner-Institut
WestAI
Dortmund/Bonn/Jülich/Aachen/Paderborn
KISSKI
Hannover/Göttingen/Kassel
hessian AI Service Center
Darmstadt
©BMBF

https://www.bmbf.de/bmbf/shareddocs/kurzmeldungen/de/2022/11/foerderung-von-4-ki-zentren-gestartet.html

AI Service Centre for Sensitive and Critical Infrastructures

The KISSKI - AI Service Centre

KISSKI
KI-Servicezentrum für sensible
und kritische Infrastrukturen

`https://kisski.de`

- Aim of the research project: Exploring how to establish an AI service centre
  - Meeting the **requirements** of critical infrastructures:
    Security, privacy, reliability
  - Services for **pilot projects** throughout Germany:
    Accessible AI infrastructure and expertise,
    Freely usable by researchers and industry, incl. Start-Ups & SMEs
  - **Research** to further improve the services:
    Scalability, **data management**, portability
- Aim at continuing the service centre after the funding period
- Funded with 19 million euros for 3 years

Projekt Partners

The KISSKI - AI Service Centre

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

GWDG
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen

aQua
ZUKUNFT DURCH QUALITÄT

CIDAS
Campus-Institut Data Science

Fraunhofer
IEE

UNIVERSITÄTSMEDIZIN
GÖTTINGEN UMG

Leibniz
Universität
Hannover

Entry-level Consulting for Energy and Health Sectors

Target group

- Companies (of any size) and research institutes from the fields of health and energy without previous experience in data-driven solutions and business models.

Your requirements

- Thorough survey of use case and needs
- Identification of initial possibilities for using the existing data
- Referral to suitable consultants from the consortium for further elaboration

Our offer

We offer entry-level consulting as well as support for companies and research institutes in the fields of health and energy that do not yet have practical experience with the design and implementation of data-driven solutions and business models. We focus on the application areas of health and energy, and define the desired use case as precisely as possible in a joint dialogue. To this end, we discuss the available data sources and models based on them in relation to the use case in terms of applicability and benefit. After consultation and a positive assessment, further KISSKI services (infrastructure, consulting or development) are applied and the project is forwarded to the appropriate expert within the KISSKI consortium.

Requirements

- Basic understanding of your own data structure(s).
- (At best) basic idea about the target image of the use case

Success stories

FAQ Support Buchen

Service type

Consulting

Contact person

Till Ole Diesterhöft
Felix Kegel

Planned start date

2023 Q3

© yelantsevv – stock.adobe.com

© vegefox.com – stock.adobe.com

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Outline

1 Introduction

2 KISSKI

3 ML in Medicine

4 Secure HPC

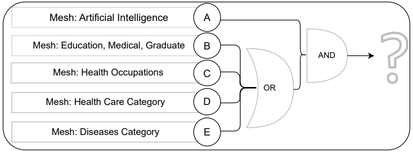
5 Data Management

6 Conclusions

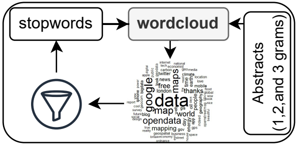
Surveying Usage of ML in Medicine

- We wanted to get an overview of ML/AI usage
  - Which diseases, models, data
- DB PubMed.gov to gather large data set (46k abstracts)
- Using AI methods to analyze the data set
  - Iterative filtering text based on methods
  - Using Wordclouds to investigate results

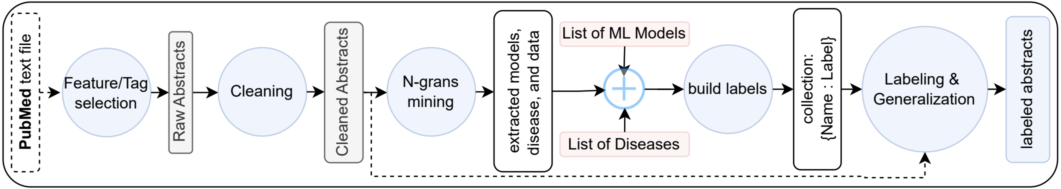
Survey Strategy


Mesh: Artificial Intelligence
A
Mesh: Education, Medical, Graduate
B
Mesh: Health Occupations
C
Mesh: Health Care Category
D
Mesh: Diseases Category
E
OR
AND
?

General Query



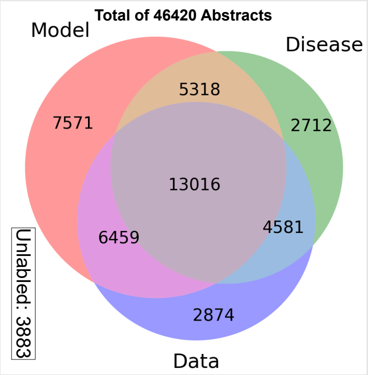
Iterative wordcloud filtering



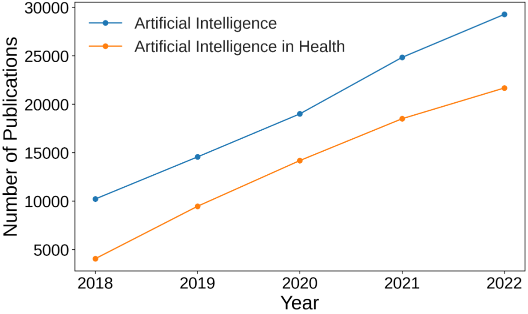
Steps applied on PubMed data to generated abstracts with labeled words.

Resulting Data Sets

- Defined categories: Model, Data, and Diseases



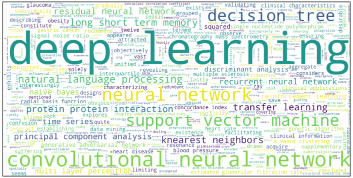
Venn Diagram of Abstracts Labeled by Model, Disease, and Data



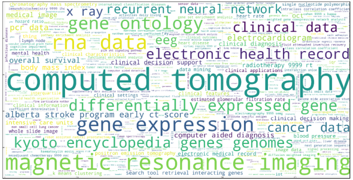
The number of Publications in Pubmed for AI vs AI in Medicine (2018-2023)

Wordcloud visualization of top ML models, data, and diseases

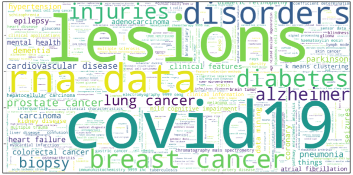
- **Top Models:** Deep learning, Convolutional neural network, Regression, Neural network, Decision tree, Support vector machine, Boosting, Natural language processing, Long short term memory, k-nearest neighbors.
- **Top Data:** Computed tomography, Magnetic resonance imaging, RNA data, Gene expression, Gene ontology, Deferentially expressed gene, Kyoto encyclopedia genes genomes, Electronic health records, clinical data, cancer data.
- **Top Disease:** COVID19, Lesions, Breast cancer, Diabetes, Injuries, Alzheimer, Disorders, Biopsy, Lung cancer, Prostate caner.



Wordcloud visualization of top ML models.



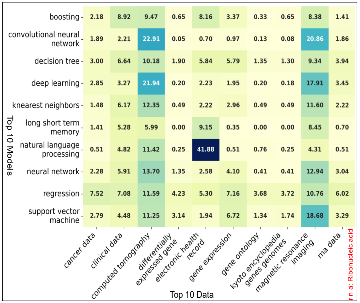
Wordcloud visualization of top Data.



Wordcloud visualization of top Diseases.

Correlation Analysis: Top Models vs Data and Diseases

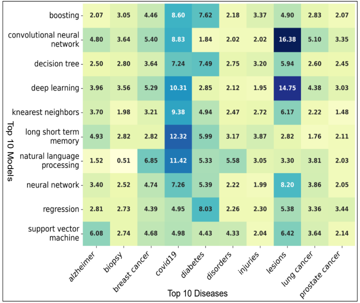
- **Structured:** 'Electronic Health Records' and 'NLP'
- **Image:** 'MRI' and 'CT Scans' with 'CNN', 'DL'.
- **Gene and RNA:** A constant preference across all models.
- **Unstructured:** 'Clinical Data' with 'Boosting' and 'Regression'.



Top 10 Models	cancer data	clinical data	computed tomography	differentially expressed gene	electronic health record	gene expression	gene ontology	kyoto encyclopedia genes genomes	magnetic resonance imaging	rna data
boosting	2.18	8.92	9.47	0.65	8.16	3.37	0.33	0.65	8.38	1.41
convolutional neural network	1.89	2.21	22.91	0.05	0.70	0.97	0.13	0.08	20.86	1.86
decision tree	3.00	6.64	10.18	1.90	5.84	5.79	1.35	1.30	9.34	3.94
deep learning	2.85	3.27	21.94	0.20	2.23	1.95	0.20	0.18	17.91	3.45
knearest neighbors	1.48	6.17	12.35	0.49	2.22	2.96	0.49	0.49	11.60	2.22
long short term memory	1.41	5.28	5.99	0.00	9.15	0.35	0.00	0.00	8.45	0.70
natural language processing	0.51	4.82	11.42	0.25	41.88	0.51	0.76	0.25	4.31	0.51
neural network	2.28	5.91	13.70	1.35	2.58	4.10	0.41	0.41	12.94	3.04
regression	7.52	7.08	11.59	4.23	5.30	7.16	3.68	3.72	10.76	6.02
support vector machine	2.79	4.48	11.25	3.14	1.94	6.72	1.34	1.74	18.68	3.29

Correlation Heatmap: Top Models and Data (2018-2022).

- **High Correlation:** Lesions: CNN and DL. Covid19: LSTM and NLP.
- **Medium Correlation:** Diabetes, Cancers, Alzheimer: all models
- **Minimum correlation:** Injuries and biopsy: all models
- **Unique phenomenon:** 'Covid19 high association with all ML models: exploration process.

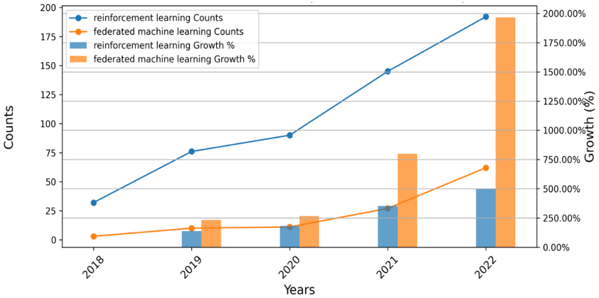


Top 10 Models	alzheimer	biopsy	breast cancer	covid19	diabetes	disorders	injuries	lesions	lung cancer	prostate cancer
boosting	2.07	3.05	4.46	8.60	7.62	2.18	3.37	4.90	2.83	2.07
convolutional neural network	4.80	3.64	5.40	8.83	1.84	2.02	2.02	16.38	5.10	3.35
decision tree	2.50	2.80	3.64	7.24	7.49	2.75	3.20	5.94	2.60	2.45
deep learning	3.96	3.56	5.29	10.31	2.85	2.12	1.95	14.75	4.38	3.03
knearest neighbors	3.70	1.98	3.21	9.38	4.94	2.47	2.72	6.17	2.22	1.48
long short term memory	4.93	2.82	2.82	12.32	5.99	3.17	3.87	2.82	1.76	2.11
natural language processing	1.52	0.51	6.85	11.42	5.33	5.58	3.05	3.30	3.81	2.03
neural network	3.40	2.52	4.74	7.26	5.39	2.22	1.99	8.20	3.86	2.05
regression	2.81	2.73	4.39	4.95	8.03	2.26	2.30	5.38	3.36	3.44
support vector machine	6.08	2.74	4.68	4.98	4.43	4.33	2.04	6.42	3.64	2.14

Correlation Heatmap: Top Models and Diseases (2018-2022).

The Potential Impact of Federated learning and Reinforcement learning in Healthcare

- The trend data from 2018 to 2022 shows that FL and RL have grown in popularity in healthcare.
- RL: from 32 in 2018 to 193 in 2022 (a more than 6-fold rise)
- FL: from 3 in 2018 to 62 in 2022 (a 20-fold increase).
- The annual growth of two approaches shows that they will be soon among the most popular techniques in healthcare
- The growth of privacy concerns leads to more interests toward FL in near future.



The number of yearly utilized FL and RL models (Line plot), and their yearly growth with respect to 2019 (Bar plot).

Outline

1 Introduction

2 KISSKI

3 ML in Medicine

4 Secure HPC

5 Data Management

6 Conclusions

Motivation

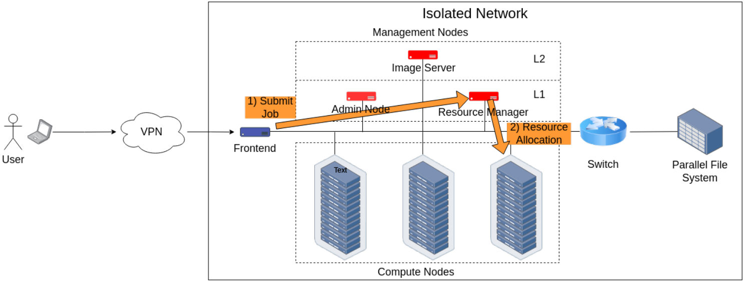
- As motivated, there are lots of opportunities to use processed health data
- However: Health data is **highly protected** by GDPR
- Consequence: Service providers must ensure full data sovereignty to users
  - No unauthorized data access from
    - other users,
    - attackers with root privileges,
    - and even admins
- To fulfill these high privacy requirements, GWDG offers a special service:
  - **SecureHPC**: A special, isolated Batch partition to process data of Risk Class D

Attack Scenarios

- Users on HPC systems get direct access to the host operating system
  - Therefore attackers can immediately exploit any local vulnerabilities
- Software dependencies, e.g. Lustre client, Slurm, often prevent fast updates
- Thus: Basic scenario is an attacker with root privileges
  - Root has immediate access to all files, shared, global and local filesystems
  - Root can manipulate the software stack, local OS or global module system
  - Network manipulations, ingestion of management packages

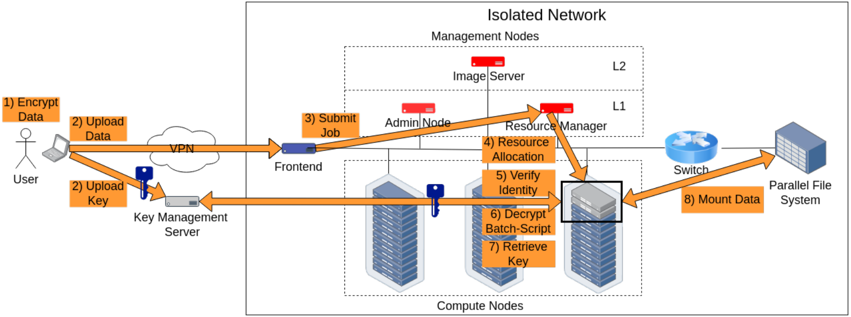
Unsecure/Normal Job Execution

- Users log in to the frontend nodes, authentication is only done at this time
- Root can directly mimic other users, e.g. sudo `-u ...`
- Slurm relies on signed messages, i.e., the munge key
  - Munge key is only protected by Linux permissions
- **Can't provide secure partition**



SecureHPC Overview

- Users encrypt **everything** on a secure, local system,
  - including: data, software, and batch script
- Batch script has to have a detached signature
- On especially isolated compute nodes Slurmd Prolog checks signature (2FA)
- Out-of-band transmission of keys between client and isolated nodes

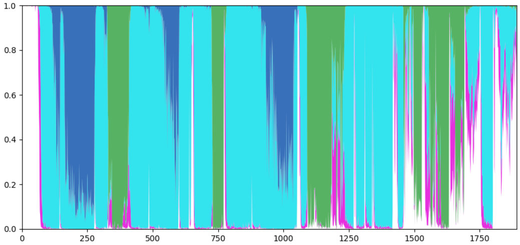


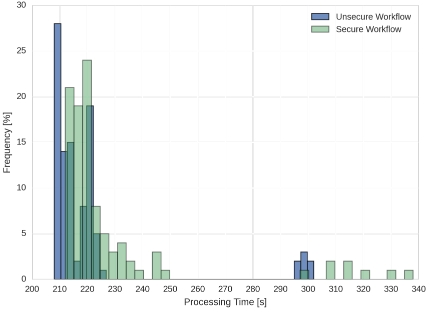
Security Assessment

- Privilege Escalation
  - Data is encrypted
  - Root can’t submit jobs with false uid due to 2FA on the secured node
  - ssh to secure nodes is not possible
- IP-Spoofing
  - If a KMS token gets leaked, it can only be used from an IP of a secure node
  - Spoofed requests will cause an response only to valid **secure nodes**
  - Will trigger security auditing systems
- Private key of secure node
  - Can’t be immediately exploited
  - Consequences are monitored

Use Case: Classification of Sleep Stages

- Analysis of nocturnal polysomnography data (EEG,EOG,EMG,ECG)
- Normally these data have to be analyzed manually
  - Data set is chunked into 30s and classified according to the depth of sleep
- We used the trained model “Stanford Stages” to automate this process
  - 15s intervals were classified
  - awake, light sleep (stages 1/2), deep sleep (stages 3/4), REM, unscored





Use Case: MRI Image Processing

- Quantitative, volumetric analysis of brain tissue
- Can detect even an early onset of dementia
- In production in clinical workflow at UMG, provides classification

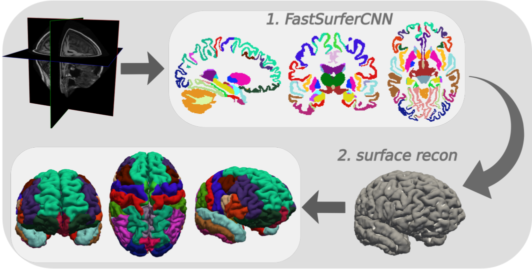


Abbildung: https://github.com/Deep-MI/FastSurfer

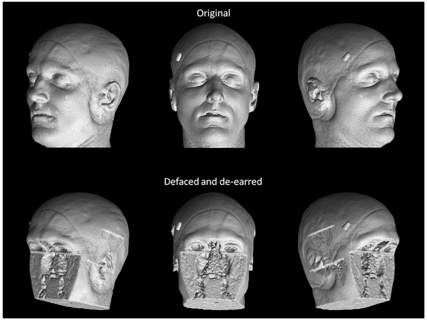
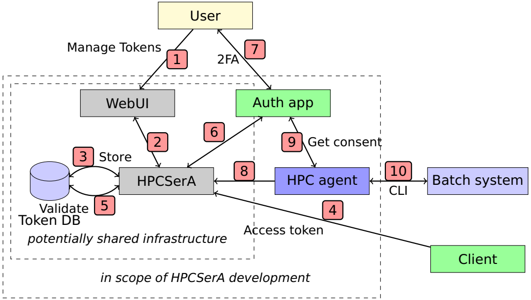


Abbildung: Theyers, Athena E., et al. "Multisite comparison of MRI defacing software across multiple cohorts."Frontiers in psychiatry 12

Solution for Remote HPC Access - with HPCSerA

- How to connect a service (e.g. Webpage) to HPC securely?
- A secure ReST API which offers a FaaS interface for HPC systems
- Fine-grained token-based scopes defined to isolate endpoints
- This enables the inclusion of DMS for HPC workloads
  - No private key has to be uploaded to some external server
  - Instead OAuth flow is triggered to provide access to the ressource



Outline

1 Introduction

2 KISSKI

3 ML in Medicine

4 Secure HPC

5 Data Management

6 Conclusions

Strategy: moving HPC data management to the next level

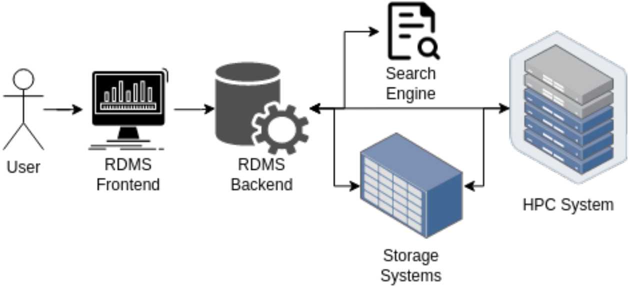
- Require users to include data management plans for non-trivial HPC usage
  - Requirements for experimental description increase with storage costs
  - Data center/system decides how to map steps to available storage resources
- Providing a data centric (DMS-centric) interaction paradigm
  - Build on top of FAIR principles
  - Web frontend to manage projects, data and HPC workloads
  - Integration into Python (Lab notebooks)
  - Tracking data products during HPC jobs and ensuring compliance
  - Integration with data repositories for sharing (DOI, PID)
- Providing semantic storage
  - Do not mix purpose and requirements (e.g. performance)
  - Ease procurement by knowing the intended usage of the storage

Data Management Plan

- Scientists define an experimental description at the beginning, containing
  - a high-level workflow description linking data with tasks
  - a data management plan for all input/output data
- Moving from a DMP as an abstract plan towards an enforced entity
  - Requires a machine-readable DMP, where users can specify
    - the data flow
    - the data sets
    - access and backup policies
    - the data life cycle
    - IO intensity (if known)
  - Each task, e.g. Slurm job, has to be linked to a workflow step

DMS-Centric Interaction Paradigms

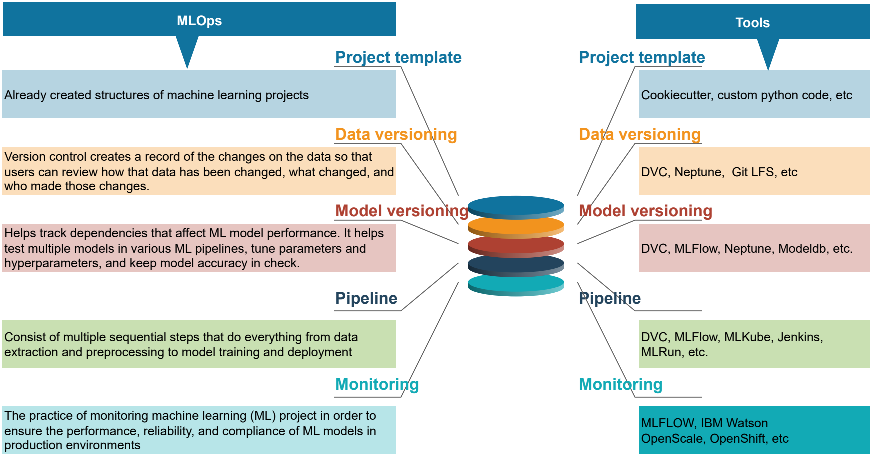
- Web frontend is used to
  - query and select input data
  - define a compute task
  - and submit it to the HPC system
- Users expect efficient and transparent data transfers
  - Data placement should also be transparently handled



Data management in Federated Learning

- Key aspects of data management in federated learning
  - **Secure Aggregation:** Secure aggregation techniques are used to protect the privacy of the updates sent from different devices.
  - **Data Sampling and Distribution:** Selecting clients in a way that maintains diversity, generocity, and performance.
  - **Model Synchronization:** Managing the synchronization of models across distributed devices or servers.
  - **Communication Efficiency:** Reducing communication overhead using techniques like compression, quantization, and client selection.
  - **Quality Control and Bias Mitigation:** Applying of techniques such as weighting updates based on the reliability of devices.
  - **Fault Tolerance and Robustness:** Implementing mechanisms to handle device failures, dropouts, or unreliable connections.

Machine learning operations (MLOPs) and Tools



MLOps

Project template

Already created structures of machine learning projects

Data versioning

Version control creates a record of the changes on the data so that users can review how that data has been changed, what changed, and who made those changes.

Model versioning

Helps track dependencies that affect ML model performance. It helps test multiple models in various ML pipelines, tune parameters and hyperparameters, and keep model accuracy in check.

Pipeline

Consist of multiple sequential steps that do everything from data extraction and preprocessing to model training and deployment

Monitoring

The practice of monitoring machine learning (ML) project in order to ensure the performance, reliability, and compliance of ML models in production environments

Tools

Project template

Cookiecutter, custom python code, etc

Data versioning

DVC, Neptune, Git LFS, etc

Model versioning

DVC, MLFlow, Neptune, Modeldb, etc.

Pipeline

DVC, MLFlow, MLKube, Jenkins, MLRun, etc.

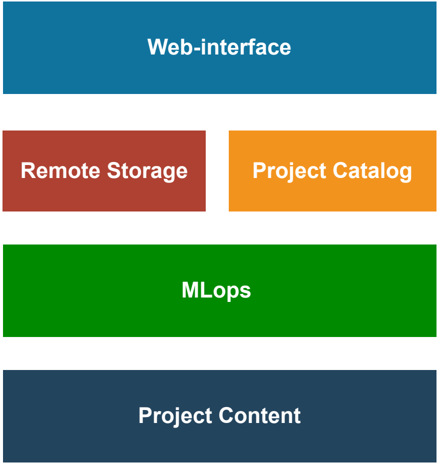
Monitoring

MLFLOW, IBM Watson OpenScale, OpenShift, etc

MLops principles apply to federated learning

- How MLops principles apply to federated learning
  - **Infrastructure Management:** Provisioning and maintaining resources across the decentralized devices or servers participating in the learning process.
  - **Versioning and Experiment Tracking:** MLops principles help in keeping track of different model versions, iterations.
  - **Automated Pipelines:** MLops facilitates the creation of such pipelines to streamline the entire process.
  - **Monitoring and Logging:** Monitoring the performance of individual devices, tracking model convergence, and logging system metrics.
  - **Model Deployment and Updates:** This involves managing updates, handling compatibility issues, and monitoring the performance of deployed models in real-time.
  - **Scaling and Resource Optimization:** This includes optimizing resource usage, managing compute resources, and handling scalability challenges.

GWDG ML Project management plan



Web-interface

Remote Storage

Project Catalog

MLops

Project Content

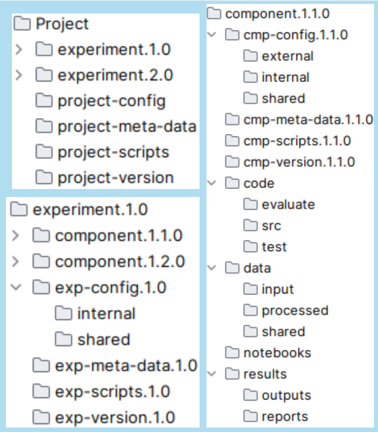
Project Registeration

1. Register the data to Remote Storage
2. Create/select the project template using Web-interface
3. Register the project and data to Catalog

Activities

Clone the project to local/remote machine
Load the raw data
Version the data
Version the Features
Version the Model
Switch the versions
Synchronize the modification between the project content and the catalog
Version the meta-data, config, and current project structure
Catalog the vesioning informations (model,data, and project)
Catalog the project structure
Catalog the project config files
Add/delete/modify experiment/component/file
Add/delete/modify scripts
Add/delete/modify meta-data (project, experiment, or component)
Add/delete/modify config file (project, experiment, or components)

Example Project Structure



```
Project
> experiment.1.0
> experiment.2.0
  project-config
  project-meta-data
  project-scripts
  project-version

experiment.1.0
> component.1.1.0
> component.1.2.0
v exp-config.1.0
    internal
    shared
  exp-meta-data.1.0
  exp-scripts.1.0
  exp-version.1.0

component.1.1.0
v cmp-config.1.1.0
    external
    internal
    shared
  cmp-meta-data.1.1.0
  cmp-scripts.1.1.0
  cmp-version.1.1.0
v code
    evaluate
    src
    test
v data
    input
    processed
    shared
  notebooks
v results
    outputs
    reports
```

Project-config

project structure
project scripts
project version information and respective scripts
Experiments shared config and respective scripts

Experiment-config

The internal config stores the information about how to handle the experiment
The shared config stores information about the objects that is going to be shared with other experiments of the project

Component-config

The internal config stores the information about how to handel the component
The shared config stores information about the objects that may be shared with other components of the current experiment
The external config stores the information about the objects that is going to be shared with other experiments of the project

Outline

1 Introduction

2 KISSKI

3 ML in Medicine

4 Secure HPC

5 Data Management

6 Conclusions

Conclusions

- Data-Driven healthcare bears high potential
- The ai-service center KISSKI supports researchers/industry
  Services and infrastructure are bookable via the webpage!
- Systematic literature research shows AI methods are on the rise
  Human analysis is barely possible (about 20k AI med pubs in 2022)
- Data management concepts are needed to handle data quantities/volume
  We are researching strategies to provide best-practices and blueprints
- Secure HPC is a workflow meeting highest security requirements