Karlsruhe Institute of Technology

[1]Institute for Automation and Applied Informatics (IAI)
[2]Institute of Biological and Chemical Systems (IBCS)
[3]Steinbuch Centre for Computing (SCC)
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

A Computational Workflow for Interdisciplinary Deep Learning Projects utilizing bwHPC Infrastructure

Marcel P. Schilling[*,1], **Oliver Neumann**[1], **Tim Scherr**[1], **Haijun Cui**[2], **Anna A. Popova**[2], **Pavel A. Levkin**[2], **Markus Götz**[3], **Markus Reischl**[1]
[*]marcel.schilling@kit.edu

Motivation

- **Use case**
  Interdisciplinary Deep Learning (DL) projects
- **Objective**
  Analyze domain data in accurate, smart, and efficient fashion
- **Challenges** (cf. Figure 1)
  - Complex and interdisciplinary development process within DL, e.g., logging, hyperparameter search, definition of objectives
  - Solution for data management required
  - Need of concepts w.r.t. computing since constrains in local resources
- **State of the art**
  Individual partial solutions → comprehensive workflow desired

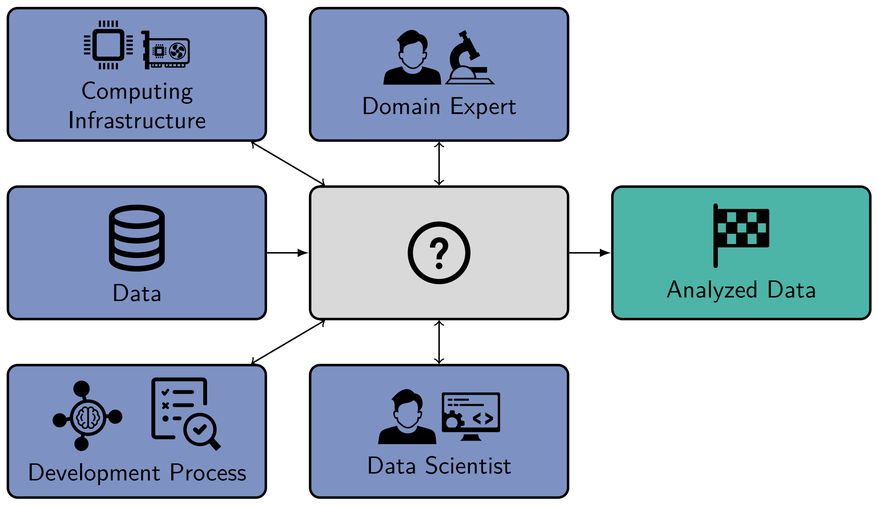


Figure 1: Problem area.

Workflow

- Composition of **Planning/Development**, **Dispatching**, and **Evaluation** (cf. Figure 2)
- Applicable in development and reduced in application stage

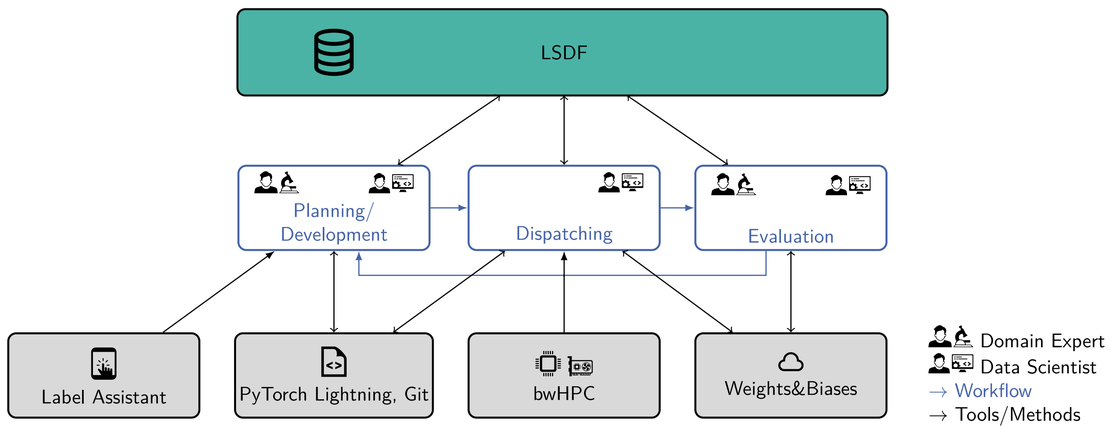


Figure 2: Workflow proposal.

Planning/Development

- Sharing data via LSDF [1]
- Label Assistant to enhance data annotation
- Alignment of processing requirements with algorithms
- PyTorch Lightning [2] to simplify DL implementations
- Git for versioning and deployment to bwHPC infrastructure

Dispatching

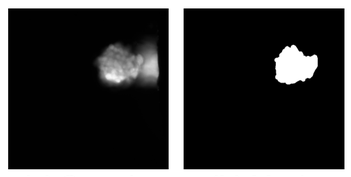
- **High Performance Computing (HPC)** for data parallel GPU training using PyTorch Lightning (SLURM support)
- **High Throughput Computing (HTC)** for parallelization of experiments (e.g. hyperparameter search, ablation studies) utilizing Weights & Biases sweep approach [3]
- LSDF versioning for large DL parameter files

Evaluation

- Interactive web-based logging via Weights & Biases [3]
- LSDF as alternative for big data results

Results

- Binary spheroid segmentation [3] high-throughput Droplet Microarray experiment (cf. Figure 3)
- Smart and flexible logging via Weights & Biases (cf. Figure 4)
- Hybrid HPC-HTC approach computation boost (cf. Table 1)
- Reduced requirements for local resources



(a) Sample **(b)** Mask

Figure 3: DMA spheroid segmentation [4].

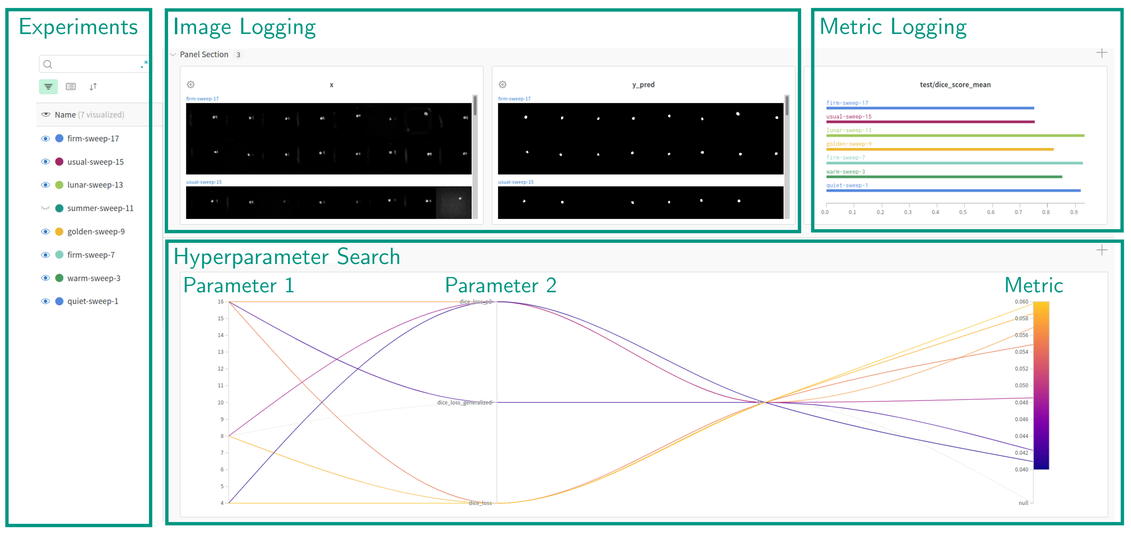


Figure 4: Weights & Biases logging: Exemplary logging possibilities [3].

Conclusions and Outlook

- Workflow as a template for DL projects linking tools/methods with solutions for flexible HPC/HTC computing and data storage
- Reduce additional overhead in DL projects
- **Outlook:** (i) Data version control [5], (ii) data submission system with auto processing, (iii) benchmark PyTorch Lightning HPC training with DASO [6].

Table 1: Run-time comparison: Average training time per epoch t_{epoch} and processing time t_{image} for different setups. DDP/DP denote the data parallelism.

Metric	CPU	1 GPU	2 GPUs (DP)	2 GPUs (DDP)	Human[†]
∅ t_{epoch} in s	332.83	7.21	5.88	**4.72**	-
∅ t_{image} in s	0.31	**0.003**	**0.003**	**0.003**	180

[†]Time for segmentation needed by an expert (lower bound obtained via a survey).

References

[1] T. Jejkal et al. LAMBDA – the LSDF execution framework for data intensive applications. In: PDP 2012, 2012.
[2] W. Falcon. PyTorch lightning. 2020, Available: https://github.com/PyTorchLightning/pytorch-lightning.
[3] L. Biewald. Experiment tracking with weights and biases. 2020, Available: https://wandb.ai/site.
[4] A. A. Popova et al. Facile one step formation and screening of tumor spheroids using droplet-microarray platform. Small, 19515(25), 2019.
[5] D. Petrov et al. Data Version Control. 2021, Available: https://dvc.org/.
[6] D. Coquelin et al. Accelerating neural network training with distributed asynchronous and selective optimization (DASO). arXiv:2104.05588 [cs.LG], 2021.

Acknowledgments

Funding: KIT Future Fields project "Screening Platform for Personalized Oncology";
Computational resource: bwUniCluster (framework program bwHPC); **Support:** DFG (Heisenbergprofessur Projektnummer: 406232485, LE 2936/9-1), Helmholtz Program "Materials Systems Engineering"