White Paper for the NSF Workshop on Smart Cyberinfrastructure, Washington, USA, February 25-27, 2020

Prof. Dr. – Ing. Morris Riedel
Head of Research Group, Juelich Supercomputing Centre, Forschungszentrum Juelich, Juelich, Germany
Professor of AI, School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
Web: www.morrisriedel.de – Twitter: @MorrisRiedel

Selected European Perspectives and Trends as Top 5 Recommendations for Smart Cyberinfrastructures for AI in the Future

Executive Summary
There is a wide variety of activities of developing ‘Cyberinfrastructures in Europe‘ (i.e., rather known as Research infrastructures) in several application domains such as those driven forward by the European Strategy Forum for Research Infrastructures (ESFRI) [1] that are primarily based on domain-specific use of data, computing, and tools. A key European Artificial Intelligence (AI) – driven effort that can not directly be considered as a ‘Cyberinfrastructure‘, but is rather a ‘AI on-demand platform‘ (AI4EU) [2] shares several ambitious goals with Cyberinfrastructures. This ‘top-down‘ initiated platform is currently in a very early stage of development but aims to inform the AI community about AI news, new AI tools, emerging AI services, and enable the sharing of AI techniques and algorithms between users of the platform and AI-related EU projects. In addition to European activities also national AI activities emerge like the Helmholtz AI initiative [11] funded by the Helmholtz Association in Germany. At the same time, we observe that ‘European Cyberinfrastructures‘ primarily focussed on computing and storage like the Partnership for Advanced Computing in Europe (PRACE) [3] or the European Grid Initiative (EGI) [4] work on satisfying an increasing number of requests related to AI applications. For example, this includes annual PRACE training courses such as ‘Parallel and Scalable Machine Learning‘ [5] or ‘Introduction to Deep Learning Model‘ [6]. The feedback of users in these training courses and the below outlined research activities contributed to lessons learned summarized as top five recommendations below for NSF w.r.t. development of smart cyberinfrastructures for AI in the future.

Despite the momentum in deep learning, users of Cyberinfrastructures in Europe such as PRACE also increasingly use traditional feature engineering approaches like parallel component trees [15], or machine learning techniques like parallel and scalable Support Vector Machines (SVM) [12], or traditional data mining methods like parallel and scalable Density-based spatial clustering of applications with noise (DBSCAN) [13]. Especially for application areas with less labeled samples, this makes sense while we observe more and more availability of data (i.e., often quoted ‘big data‘) but that does not necessarily mean that ‘big data‘ is of high quality or perfectly labeled to be used with machine learning techniques. For cases where large quantities of data are available, Cyberinfrastructures support distributed deep learning training at scale, for example by using Horovod for scaling across many GPUs for AI applications like in remote sensing [14]. Internally, Horovod is using concepts of the Message Passing Interface (MPI). Also, users demand increased number of open datasets for AI applications.

Given the fact that many users of Cyberinfrastructures continue to use it with simulation sciences based on numerical methods based on known physical laws, the increase in using AI workloads with new frameworks becomes a challenge for Cyberinfrastructures leading to more required heterogeneity in the Cyberinfrastructures. Examples include also the use of Apache Spark for AI workloads like using autoencoders [17] or the joint use of AI in physics-informed ML applications [21]. Resource providers need to support a broader set of workloads that investigations in the European DEEP series of projects [16] reveal leading to a modular supercomputing architecture (MSA) [22] implemented in the Juelich Supercomputing Centre (JSC). The modularity will be key for Cyberinfrastructures to support, not only to support traditional simulation sciences and more emerging AI workloads but also to integrate innovative computing aspects (e.g., quantum computing, neuromorphic computing, etc.). We have already started to use an SVM based on the quantum annealing approach [19] to solve an inherent optimization problem in remote sensing applications [18]. Hence, Quantum computing is in reach and need to be part of future ‘modular‘ concepts of Cyberinfrastructures. Another example of the need for modularity are MSA-enabled reinforcement learning environments that, for example, may help to deal with the large complexity of finding the right hyper-parameters in deep learning networks.

Recommendation 1: Enable Application Enabling Process after Cyberinfrastructure Training
After performing Cyberinfrastructure & AI training (i.e., lessons learned from PRACE training), users often raise demand for so-called 'application-enabling' by high level support teams (HLSTs) that can be summarized as supporting AI users in several weeks of adopting new AI techniques within their application domains using Cyberinfrastructures; training and offering 'some core-hours' alone is not enough to ensure an uptake of the Cyberinfrastructure capabilities and offered services & tools.

Recommendation 2: Provide Easy Access and Sharing of Scripts & Data with JupyterLab
PRACE training courses for AI as described above (but also for other topics like Python, MPI, etc.) are increasingly carried out with Jupyter notebooks and JupyterLab and this environment is available also for researchers at the Juelich Supercomputing Centre (JSC) and was presented at ISC2019 last year [10]. Cyberinfrastructures need to support these technologies more broadly to enable easier collaboration in research through the sharing of scripts, Jupyter kernels, or datasets for AI applications. This also lowers the technical boundary for AI end-users that are not familiar with concepts like SSH and could be considered for integration into existing Science Gateway activities in the US. Challenges for Cyberinfrastructure providers need to be tackled like supporting interactive access instead of the traditional batch operations and implement Jupyter kernels on various coherent versions of deep learning tools (e.g., specific versions of Tensorflow/Keras) and underlying hardware libraries (e.g., CUDDN).

Recommendation 3: Contribute with Cyberinfrastructures to International AI Activities
There are excellent examples of international collaboration activities such as the Joint Laboratory for Extreme-Scale Computing (JLESC) [7] or Big Data & Extreme-Scale Computing (BDEC) activities [8] that both are not primarily focussed on AI and rather focus towards developments in Exascale. Hence, in contrast to the AI4EU platform, the Joint Artificial Intelligence and Machine Learning Lab (JoAIML) [9] was created as a spin-off from JLESC as Bottom-Up activity to encourage the exchange of AI-based scripts and datasets via GitLab [9]. Although JoAIML is currently driven forward by the Helmholtz AI at JSC, the University of Iceland (UoIceland), and the National Center for Supercomputing Applications (NCSA), JoAIML is considered to be open to other international members (e.g., other partners from US, Japan, UK, etc.). More recently, discussions emerge how JoAIML can become broader and more visible (e.g., own Web page) for Cyberinfrastructure users in the future.

Recommendation 4: Offer Parallel & Scalable Tools & Techniques at Scale
Despite the fact that Deep Learning tools (e.g., TensorFlow, Keras, PyTorch, etc.) and distributed training frameworks (e.g., using Horovod) are required, it makes sense also to offer also traditional machine learning algorithms as parallel and scalable codes in the Cyberinfrastructure. Not all applications really need cutting-edge deep learning models like RESNET-50, for example, when labeled dataset samples are limited that is observed quite often in science and engineering. For example, SVMs, as desribed above, are robust methods that may lead to better results in some application areas. Reasons are grounded in statistical learning theory and that the number of parameters for SVMs is not much (in contrast to deep learning models) thus enabling better generalization and less risk in overfitting.

Recommendation 5: Modularity of the Cyberinfrastructure with distinct Services & Resources
The modularity of the Cyberinfrastructure will be key for AI applications in the future and is here inspired by the MSA as described above. While computational-intensive model training can be done by using cluster modules (CM) with high single-thread performance CPUs or specific GPUs, less computational-intensive AI model testing and inference can be performed on scalable booster modules (BM). Apache Spark stacks with AI algorithms can be supported by Data Analytics Modules (DAMs) contributing to the fact that for specific AI application problems it is more efficient to use specific devices. Like accelerators are used now for deep learning problems, it can be possible that a quantum device module can be used for the complex solving of optimization problems in AI algorithms. Equally interesting, future resources and services may offer neuromorphic devices or more use of containers (i.e., Docker, Singularity, etc.) for AI applications. Other distinct services of the Cyberinfrastructure should support end-users in a systematic fashion by enabling (instance-aware) Neural Architecture Search (NAS), for example, based on reinforcement learning environments as described in [20]. Finding the right hyper-parameter and parameter-tuning is very computational-intensive, overlaps with AutoML techniques, and is a key problem to be solved for AI communities by Cyberinfrastructures of the future.

References

[1] European Strategy Forum on Research Infrastructures (ESFRI), Online:
https://www.esfri.eu/
[2] AI on Demand Platform for Europe (AI4EU), Online:
https://www.ai4eu.eu/
[3] Partnership for Advanced Computing in Europe (PRACE), Online:
http://www.prace-ri.eu/
[4] European Grid Initiative (EGI), Online:
https://www.egi.eu/
[5] PRACE Tutorial: Parallel and Scalable Learning – Introduction, Online:
http://www.morrisriedel.de/prace-tutorial-parallel-and-scalable-machine-learning-introduction
[6] PRACE Tutorial: Introduction to Deep Learning Models, Online:
http://www.morrisriedel.de/introduction-to-deep-learning-models
[7] Joint Laboratory for Extreme-Scale Computing (JLESC), Online:
https://jlesc.github.io/
[8] Big Data & Extreme-Scale Computing (BDEC), Online:
https://www.exascale.org/bdec/
[9] Joint Artificial Intelligence and Machine Learning Lab (JoAIML), Online:
https://gitlab.version.fz-juelich.de/JoAIML_Lab/workspace/material
[10] Goebbert, J.H., Kreuzer, T., Grosch, A., Lintermann, A., **Riedel, M.**: **Enabling Interactive Supercomputing at JSC Lessons Learned**, in conference proceedings of the International Conference on High-Performance Computing, Springer, Lecture Notes in Computer Science (LNCS) Vol. 11203, June 24-28, 2019, Frankfurt, Germany, Online:
https://www.researchgate.net/publication/330621591_Enabling_Interactive_Supercomputing_at_JSC_Lessons_Learned_ISC_High_Performance_2018_International_Workshops_FrankfurtMain_Germany_June_28_2018_Revised_Selected_Papers
[11] Helmholtz AI Initiative of the Helmholtz Association, Online:
https://www.helmholtz.ai/
[12] Cavallaro, G., **Riedel, M.**, Richerzhagen, M., Benediktsson, J., Plaza, A.: **On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods**, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), Vol 9 (10), pp. 1-13, 2015, Online:
https://www.researchgate.net/publication/282524415_On_Understanding_Big_Data_Impacts_in_Remotely_Sensed_Image_Classification_Using_Support_Vector_Machine_Methods
[13] Goetz, M., Bodenstein, C., **Riedel, M.**: **HPDBSCAN – Highly Parallel DBSCAN**, in conference proceedings of ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC 2015), Machine Learning in HPC Environments (MLHPC 2015) Workshop, November 15-20, 2015, Austin, Texas, USA, Online:
https://www.researchgate.net/publication/301463871_HPDBSCAN_highly_parallel_DBSCAN
[14] Sedona, R., Cavallaro, G., Jitsev, J., Strube, A., **Riedel, M.**, Benediktsson, J.A.: **Remote Sensing Big Data Classification with High Performance Distributed Deep Learning**, Journal of Remote Sensing, Multidisciplinary Digital Publishing Institute (MDPI), Special Issue on Analysis of Big Data in Remote Sensing, 2019, Online:
https://www.researchgate.net/publication/338077024_Remote_Sensing_Big_Data_Classification_with_High_Performance_Distributed_Deep_Learning
[15] Goetz, M., Cavallaro, G., Geraud, T., Book, M., **Riedel, M.**: **Parallel Computation of Component Trees on Distributed Memory Machines**, IEEE Transactions on Parallel and Distributed Systems (TPDS), Vol 29 (11), 2018, Online:
https://www.researchgate.net/publication/325212343_Parallel_Computation_of_Component_Trees_on_Distributed_Memory_Machines
[16] DEEP Series of EU Projects, Online:
https://www.deep-projects.eu/
[17] Haut, J.M., Gallardo, J.A., Paoletti, M.E., Cavallaro, G., Plaza, J., Plaza, A., **Riedel, M.**: **Cloud Deep Networks for Hyperspectral Image Analysis**, IEEE Transactions on Geoscience and Remote Sensing, PP(99):1-17, 2019, Online:
https://www.researchgate.net/publication/335181248_Cloud_Deep_Networks_for_Hyperspectral_Image_Analysis
[18] G. Cavallaro, D. Willsch, M. Willsch, K. Michielsen, and **M. Riedel**: **Approching Remote Sensing Image Classification with Ensembles of Support Vector Machines on the D-Wave Quantum Annealer**, in the IEEE International Geoscience and Remote Sensing Symposium, 2020 (submitted)
[19] D. Willsch, M. Willsch, H. De Raedt and K. Michielsen, **'Support Vector Machines on the D-Wave Quantum Annealer'** 2019, Online:
http://dx.doi.org/10.1016/j.cpc.2019.107006
[20] **Morris Riedel**, ,'**Neural Architecture Search with Reinforcement Learning**' Invited Talk, 5th International Summer School on Big Data and Machine Learning, Technical University of Dresden, Dresden, Germany, Online:
http://www.morrisriedel.de/neural-architecture-search-with-reinforcement-learning
[21] Mathis Bode, Michael Gauding, Zehu Lian, Dominik Denker, Marco Davidovic, Konstantin Kleinheinz, Jenia Jitsev, Heinz Pitsch, '**Using Physics-Informed Super-Resolution Generative Adversarial Networks for Subgrid Modeling in Turbulent Reactive Flows**', Online:
https://www.researchgate.net/publication/337560077_Using_Physics-Informed_Super-Resolution_Generative_Adversarial_Networks_for_Subgrid_Modeling_in_Turbulent_Reactive_Flows
[22] E. Suarez, N. Eicker, Th. Lippert, ,'**Modular Supercomputing Architecture: From Idea to Production**', Book, Contemporary High Performance Computing, Online:
https://www.researchgate.net/publication/334264090_Modular_Supercomputing_Architecture_From_Idea_to_Production