DEEP LEARNING ON HPC SYSTEMS

Dr. Peter Labus,

Competence Center for High Performance Computing, Fraunhofer ITWM, Kaiserslautern.

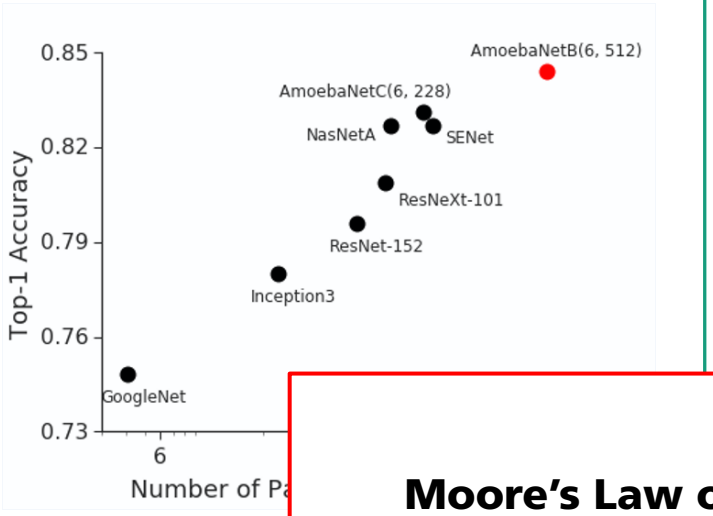
Fraunhofer Center Machine Learning.

9th HPC Status conference Gauß-Allianz

October 17, 2019, Paderborn

Increasing AI model complexity leads to increasing compute demand

increasing AI model complexity

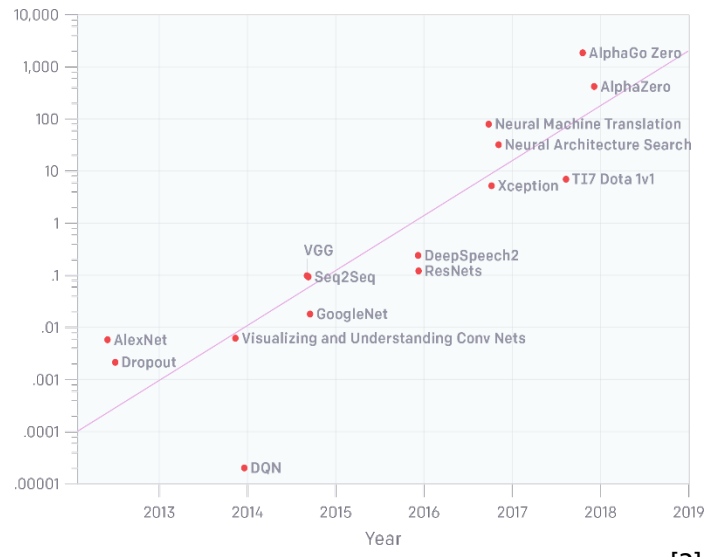


leads to

Moore's Law of AI:

exponential growth with a 3.5 month-doubling time

increasing compute demand



[2]

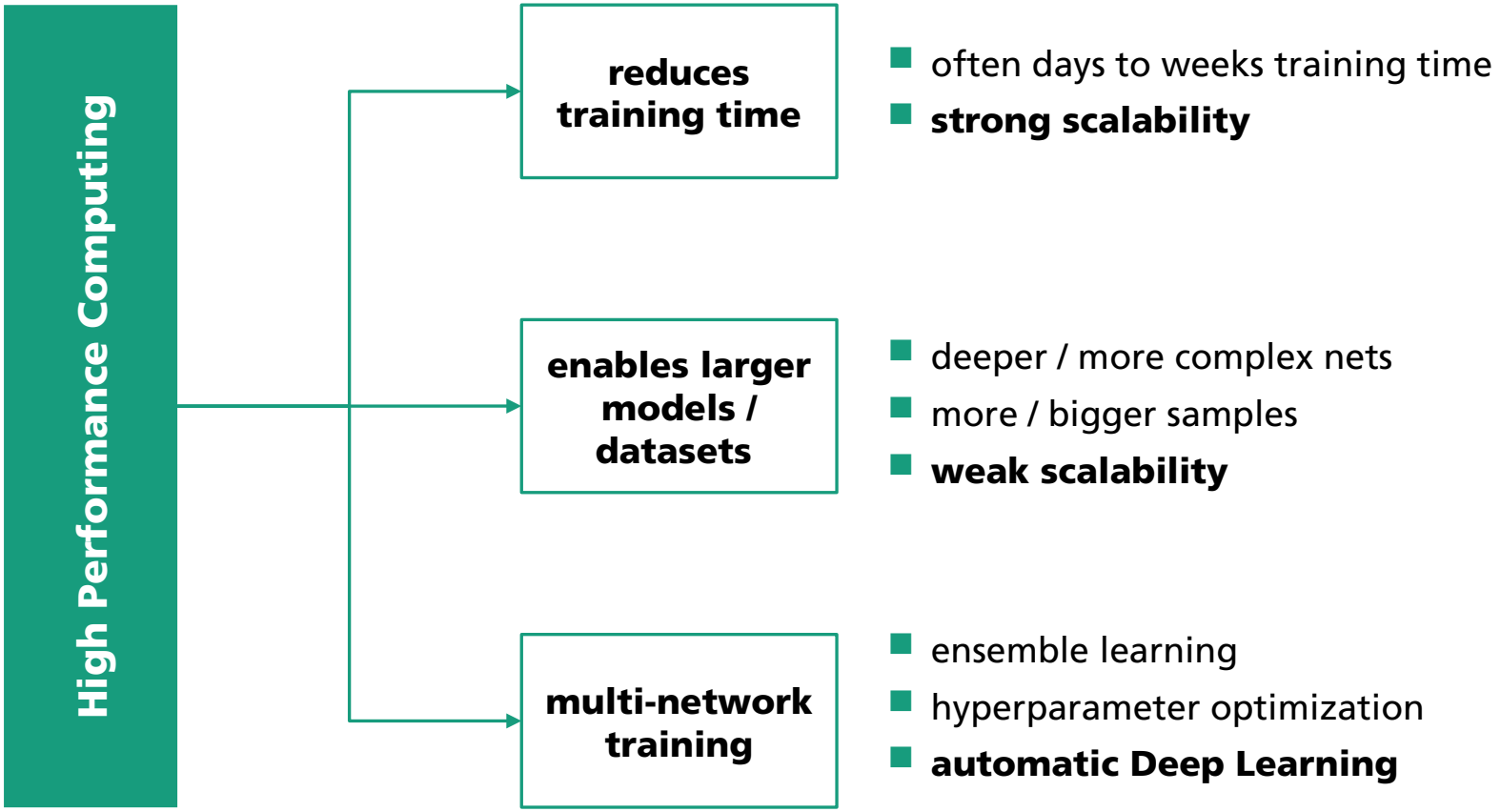
[1] Huang, Yanping, et al. "Gpipe: Efficient training of giant neural networks using pipeline parallelism."
[2] https://openai.com/blog/ai-and-compute/

Fraunhofer

AGENDA

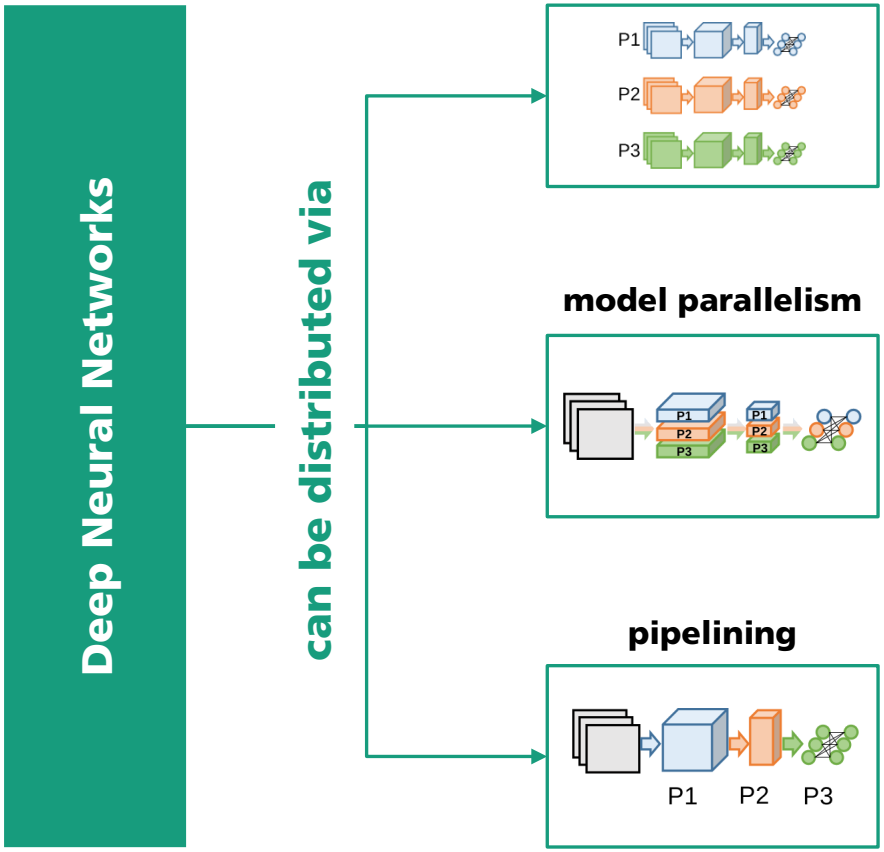
1. Deep Learning & High Performance Computing
2. Scaling Deep Neural Network Training
3. Novel algorithms & visualization
4. Automatic Deep Learning

HPC reduces training time, enables larger models and datasets & multi-network training



Fraunhofer

DNNs can be distributed via data parallelism, model parallelism & pipelining



- example: almost everything
- ResNet-50 scaling records [4]
- **best strong scalability**

- example: **AlexNet**
- enables weak scaling
- hard to implement

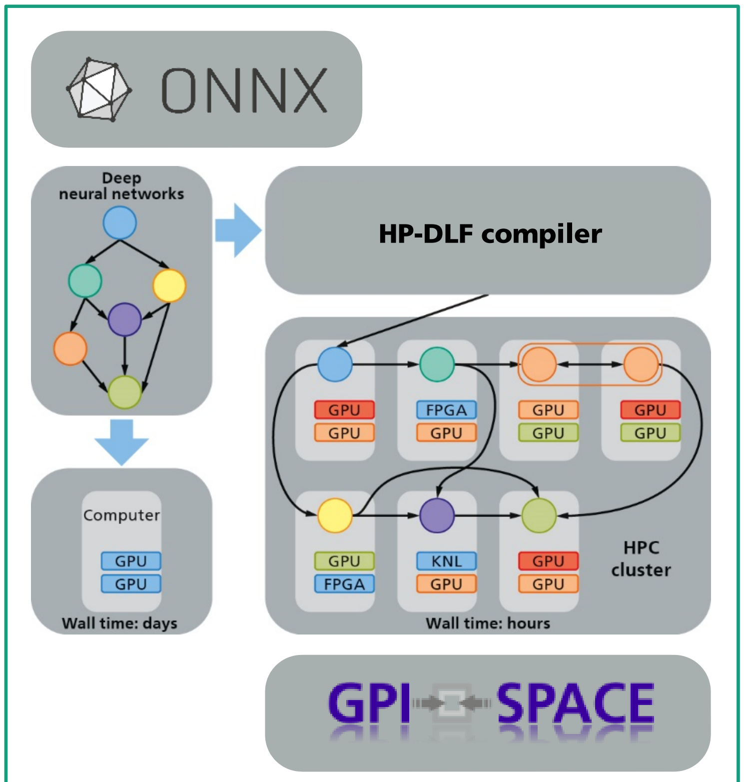
- examples: **AmoebaNet, Transformer models**
- works best for bigger models

[3] Ben-Nun, Tal, and Torsten Hoefler. "Demystifying parallel and distributed deep learning" ACM Computing Surveys (CSUR) 52.4 (2019):65.
[4] Mikami, Hiroaki, et al. "Imagenet/resnet-50 training in 224 seconds." arXiv preprint arXiv:1811.05233 (2018). C

Fraunhofer

Initial design of HP-DLF led to a number of issues

initial design of HP-DLF



led to

a number of issues

- ONNX as frontend

no training interface (solver loss functions, da processing), ot wide upported, user need learn ne interf

- CE

virtua es, need more fi

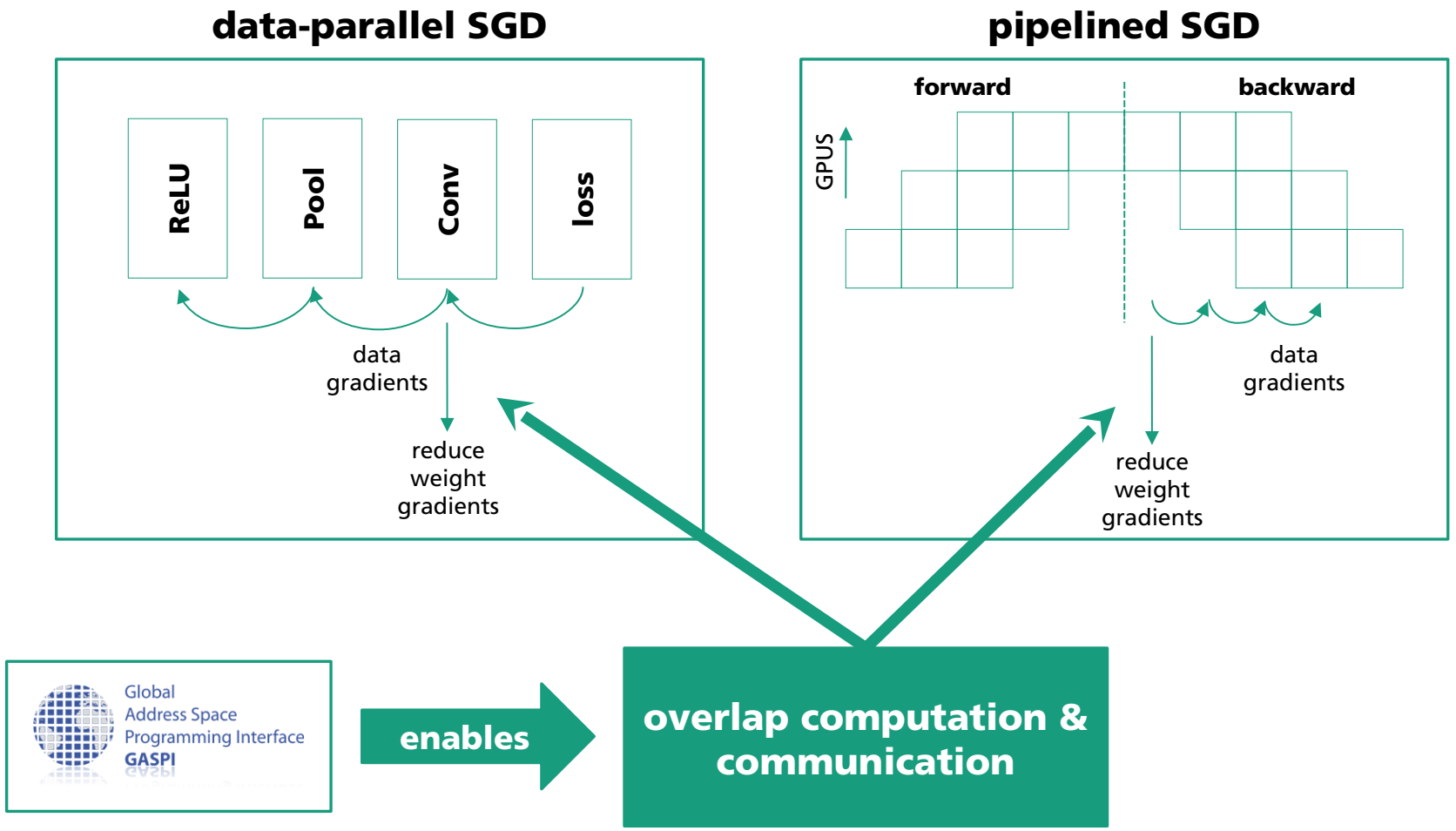
- frame ork eve ome

lots of developme overh ad for user interface, data pr cessing, implementation wrappers, …

Build scalable communication library on top of TensorFlow

Fraunhofer

GASPI communication library enables the overlap of computation & communication

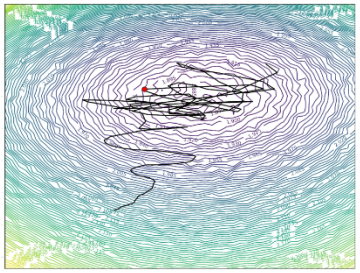


[5] Huang, Yanping, et al. "Gpipe: Efficient training of giant neural networks using pipeline parallelism." arXiv preprint arXiv:1811.06965 (2018).

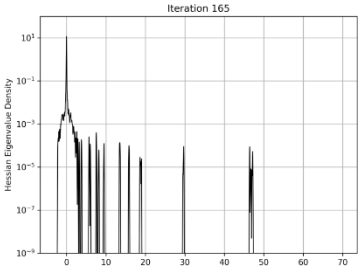
Fraunhofer

New visualization techniques help study novel optimizers

New visualization techniques



2D/3D loss surfaces & trajectories



Hessian eigenvalue spectra

Our contributions: [6]

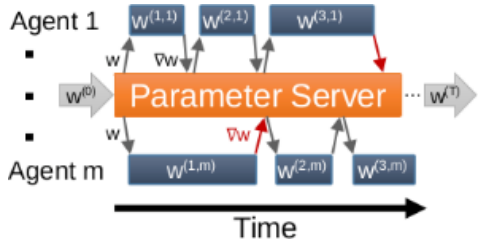
- projection onto EVs of Hessian
- scalable visualization impl.
- scalable & efficient Lanczos

github.com/cc-hpc-itwm/GradVis

help to study

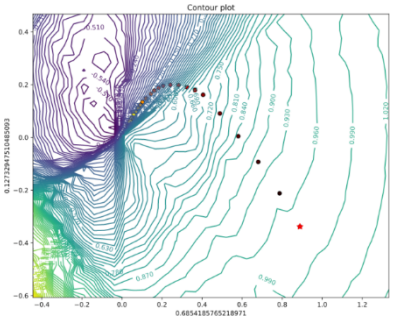
Novel optimizers

Asynchronous SGD [7]



improve scaling by reduction of message frequency

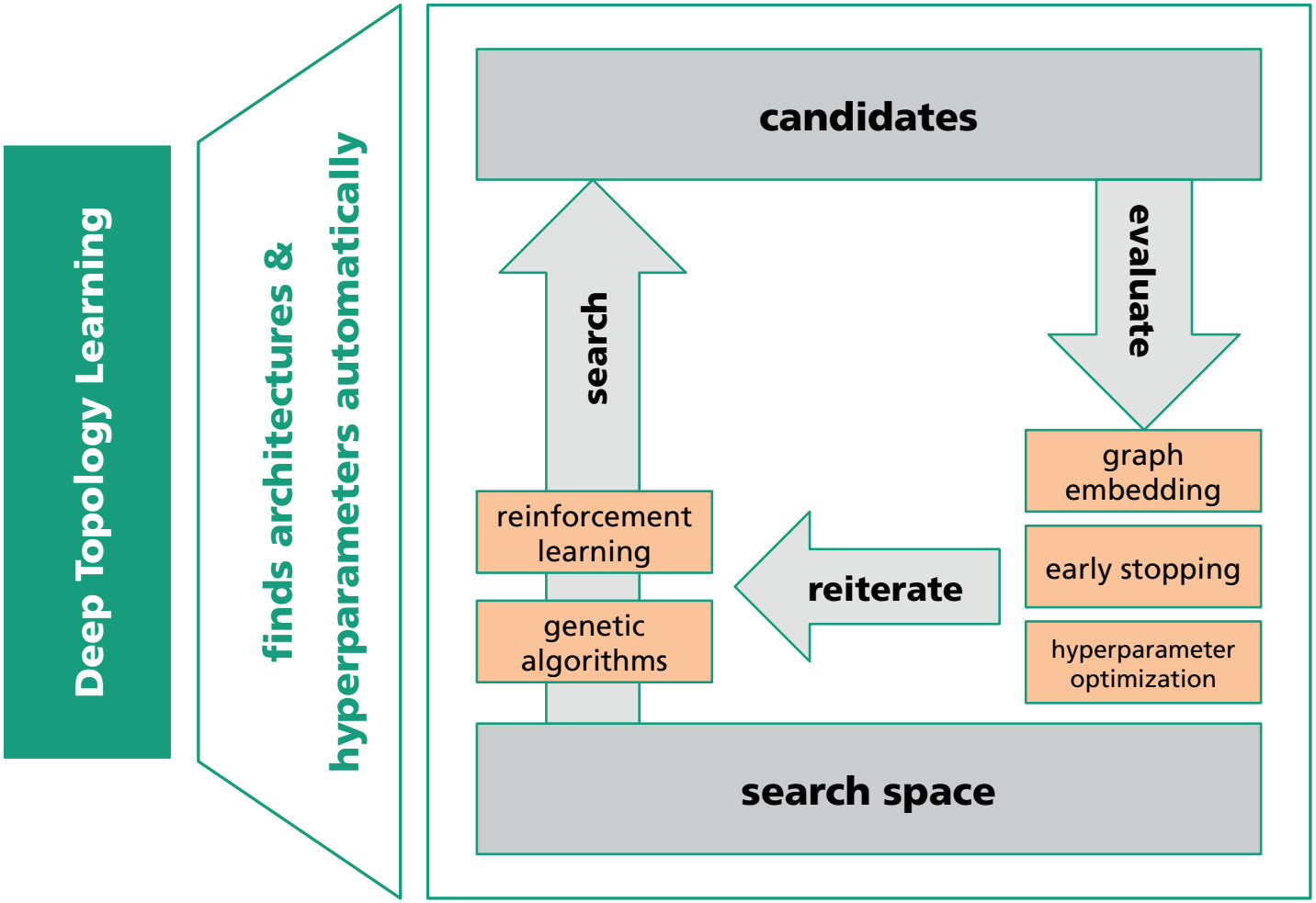
ResOpt [8]



improve scaling by reduction of message sizes

[6] Chatzimichailidis, A., et al. "GradVis: Visualization and Second Order […]" arXiv preprint arXiv:1909.12108 (2019).
[7] Keuper, J. et al. "Asynchronous parallel stochastic gradient descent […] Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments. ACM, 2015.
[8] Loroch, D. "ResOpt […]" to appear

Deep Topology Learning finds DNN architectures & hyperparameters automatically



Bundesministerium für Bildung und Forschung

UNI FREIBURG

Fraunhofer ITWM

Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

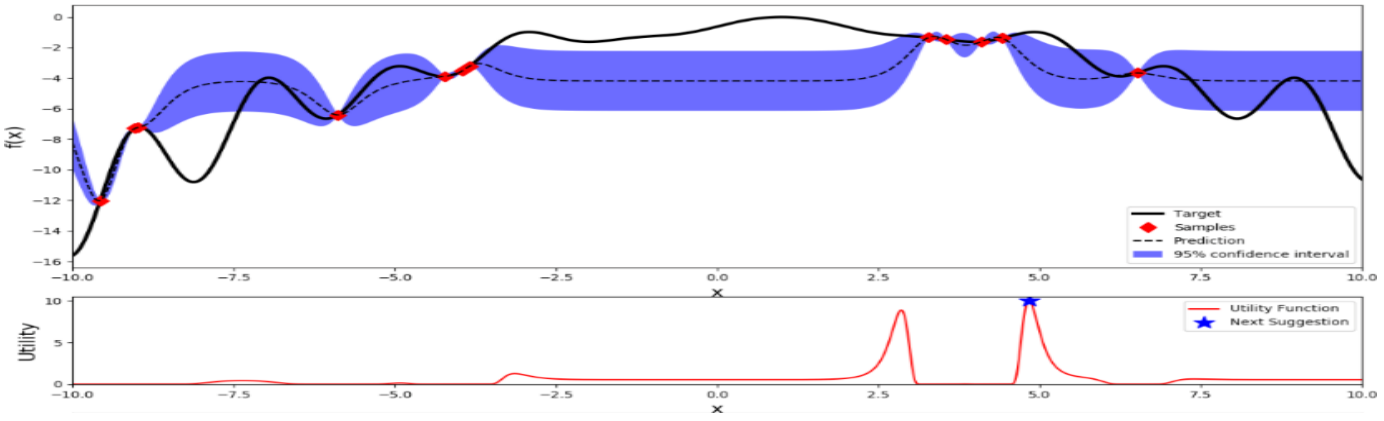
PSIORI

UNIVERSITÄT MANNHEIM

Fraunhofer

Lazy Gaussian Processes enable scalable Bayesian Optimization

Lazy Gaussian Processes



Build covariance matrix iteratively: $K_{n+1} = \begin{pmatrix} K_n & \mathbf{p} \\ \mathbf{p}^T & c \end{pmatrix}$

enables

Scalable Bayesian Opt

Matrix inversion via Cholesky decomposition:

improvement from O(n^3) to O(n^2)

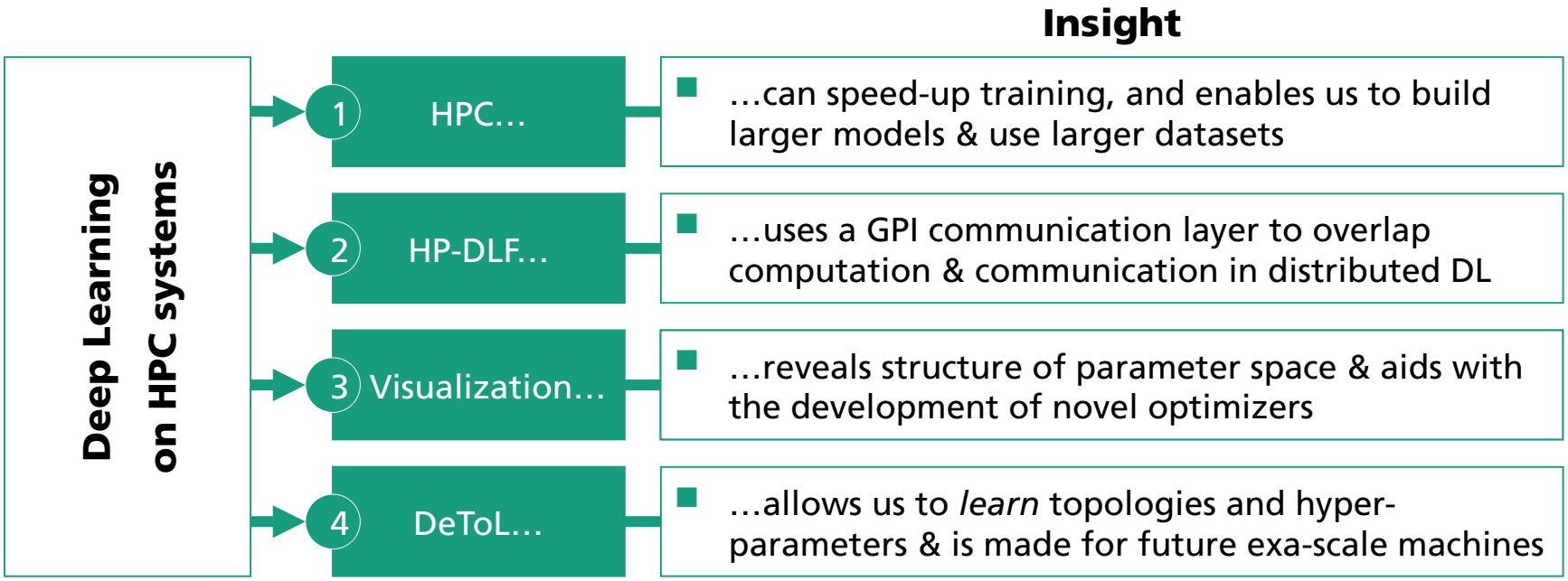
Naïve Cholesky decomposition	
Iteration	Accuracy
16	0.74
26	0.75
43	0.77
50	0.78
176	0.79

Optimized Cholesky decomposition	
Iteration	Accuracy
1	0.11
16	0.51
21	0.77
35	0.79
61	**0.80**

[6] Ram, Raju, et. al. "Scalable Hyperparameter Optimization with Lazy Gaussian Processes" to appear in Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments. ACM, 2019.

Fraunhofer

Summary: Deep Learning on HPC systems



Fraunhofer