High-Performance
Computing Center
Stuttgart

Special Track: HPC/AI Convergence

The 1st International Conference on AI-based Systems and Services AISyS 2024

Dennis Hoppe

Special Track: HPC/AI Convergence

Chairs

- Prof. Dr-Ing. Michael **Resch**
  - Director of the High-Performance Computing Center Stuttgart (HLRS)
- Dennis **Hoppe**, M.Sc.
  - Head of Converged Computing at HLRS

Moderator

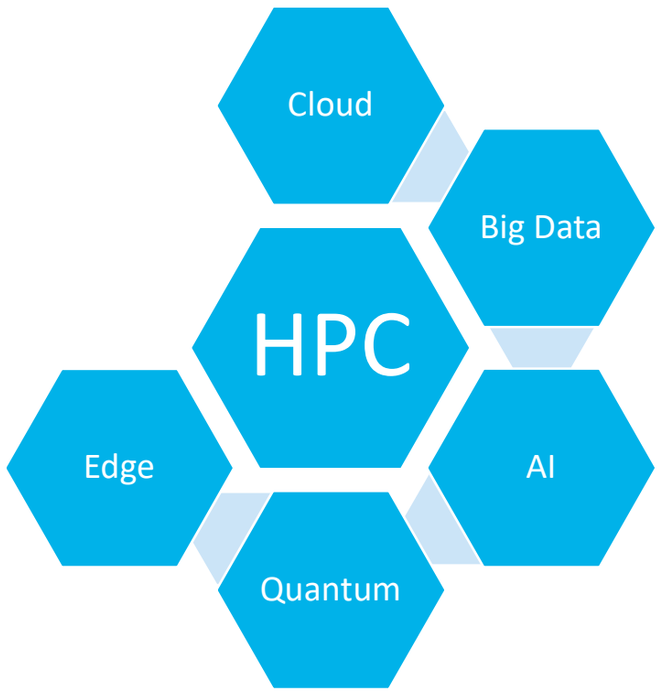
- Rishabh **Saxena**, M.Sc.
  - Researcher at HLRS in the Converged Computing department

Enabling Next-Generation HPC Solutions

H L R S

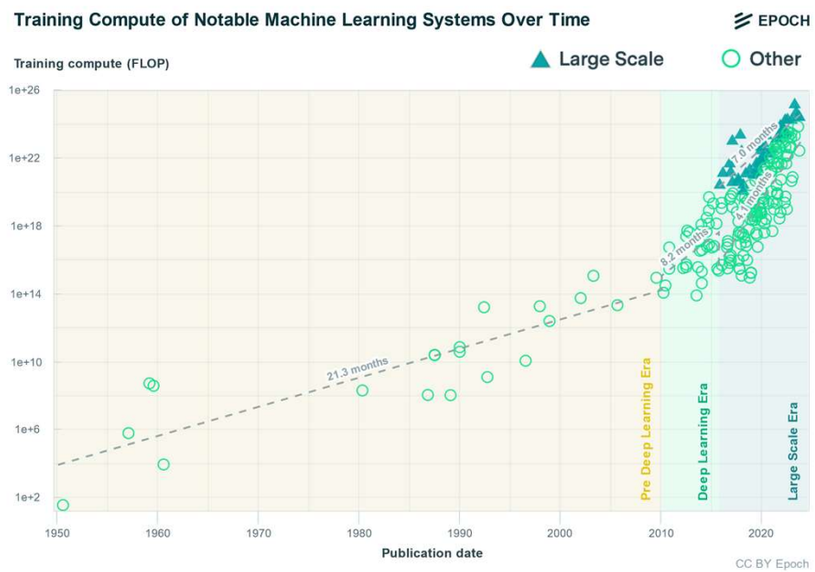
- **Transformation of HPC** through emerging technologies and methodologies
  - We are entering an **era of workflows**
  - Spans the entire compute continuum
    - Cloud, AI, Quantum, Edge, …?

- **Key objectives**
  - Exploring **emerging technologies**
  - Identifying **synergies**
  - Evaluating **hybrid workflows**
  - Driving **seamless integration**
  - Integrate the **user**



Why does AI need HPC?

- Three factors **drive AI innovation**:
  - algorithmic innovation, data, and FLOPs available



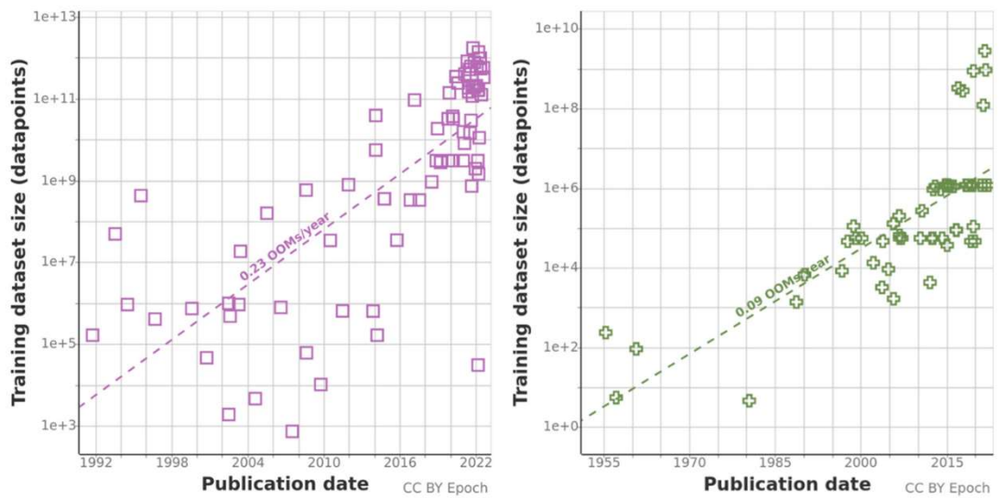


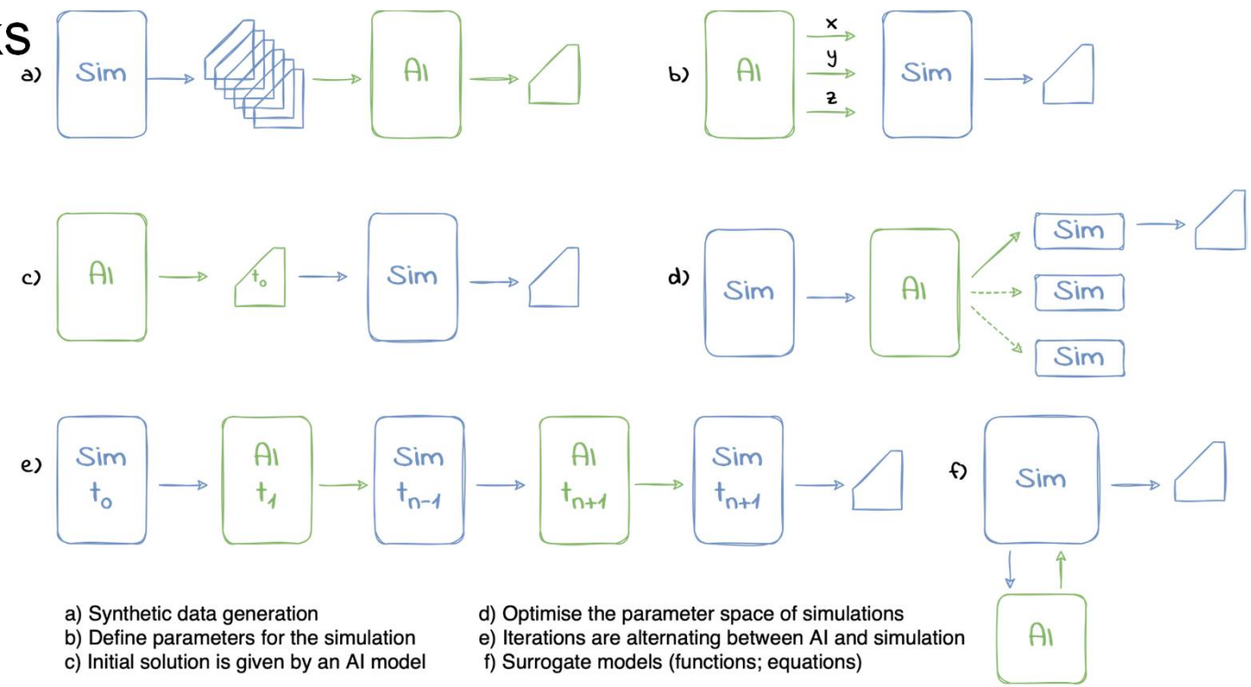
Figure 1: Training datasets for language (left) and vision (right).

[1] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, 'Compute Trends Across Three Eras of Machine Learning'.
[2] Pablo Villalobos and Anson Ho (2022), "Trends in Training Dataset Sizes". *Published online at epochai.org.*

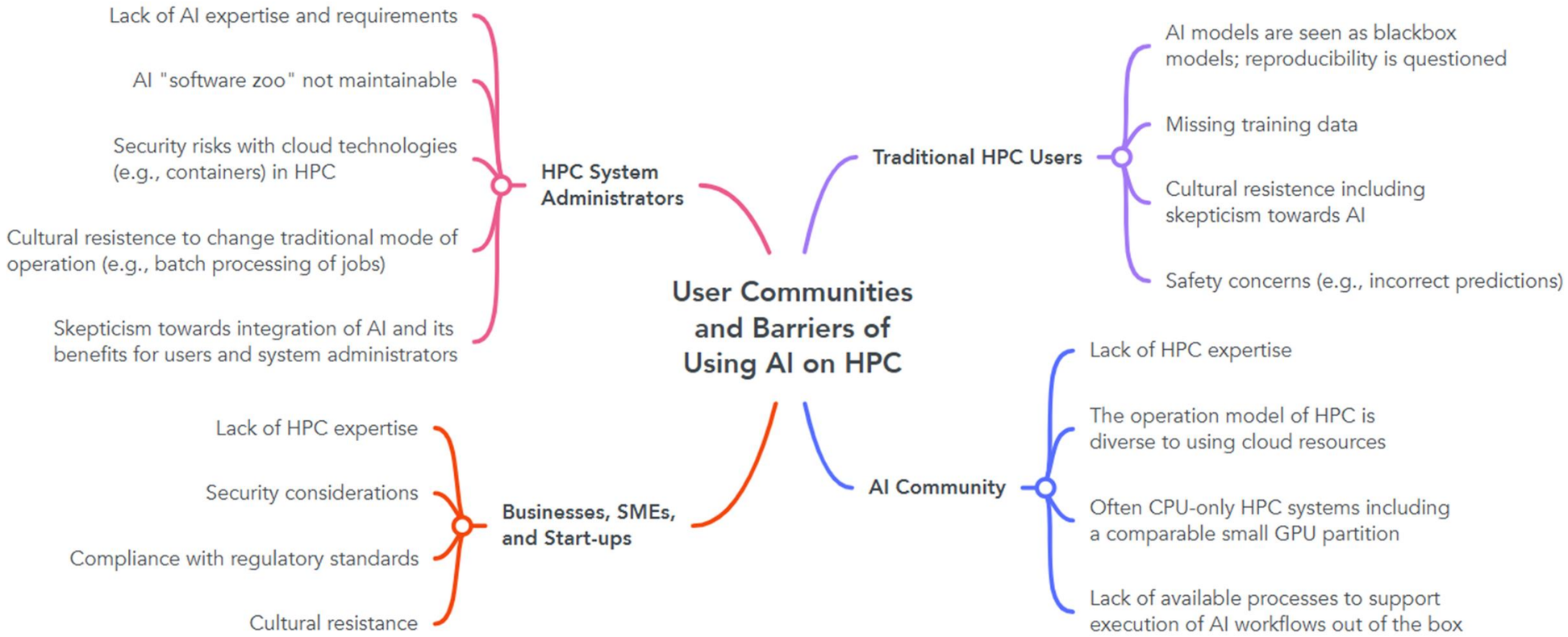
Why does HPC need AI?

H L R S

- **Hybrid HPC/AI workflows** extend classical simulations by AI methodologies to improve accuracy and/or speed-up simulations
- Physics-Informed Neural Networks
- Operational Data Analytics
- AI-driven Cybersecurity
- … and more



Challenges while Adopting AI on HPC



Special Track Contributions

H L R S

AI for Fluid Dynamics

- *An Advanced Surrogate Model Approach for Enhancing Fluid Dynamics Simulations*
  S. Kavane et al., Friedrich-Alexander-Universität Erlangen (FAU)

AI for Global System Sciences

- *AI for Global Challenges: Case Studies in Urban Solar Exposure and Wildfire Management;* G. Filandrianos et al., National Technical University of Athens (NTUA)

AI for Cybersecurity

- *Cybersecurity Concerns of AI Applications on High-Performance Computing Systems*
  R. Saxena et al., HLRS, University of Stuttgart

Future Directions and Challenges

H L R S

- **Sustainable AI** → Foundational Models become too large [1]
  - Algorithmic
    - Research towards smaller models (less data, less parameters) → SLM and MLMs
    - Mixed precision training to improve the energy efficiency of training and inference [2]
  - Hardware
    - Establish good practices to report energy costs along with each AI model
    - Using more energy-efficient hardware (e.g., TPUs, ASICS, …)
- **Data Trends** → Running out of high-quality data
  - Study suggests we might be out of high-quality data for training by 2026 [2]
  - Compute might no longer be the bottleneck, but data availability

[1] McDonald, Joseph, et al. "Great power, great responsibility: Recommendations for reducing energy for training language models." *arXiv preprint arXiv:2205.09646* (2022).
[2] Dörrich, M., Fan, M., & Kist, A. M. (2023). Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks. IEEE Access, 11, 57627-57634.
[3] Source: https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

Thank you!

Dennis Hoppe
Head of Converged Computing

High Performance Computing Center (HLRS)
University of Stuttgart
Nobelstr. 19
D-70569 Stuttgart, Germany

skype: dennis.hoppe

phone: +49-711-685-60300
fax: +49-711-685-65832
web: www.hlrs.de