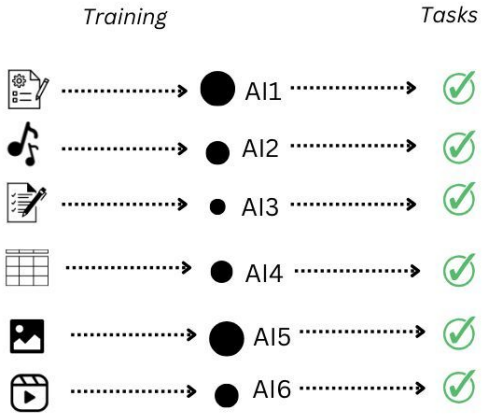
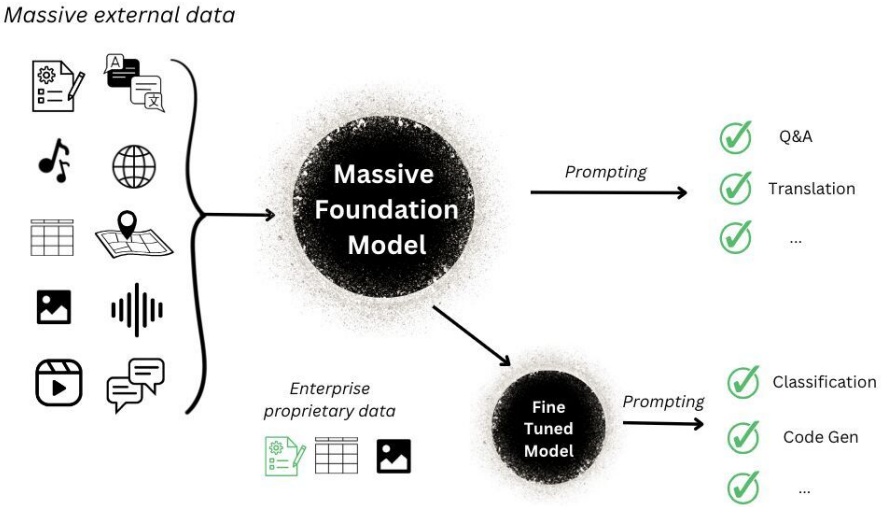
Größer, schneller, nachhaltiger?
Wie KI-Designs auf verschiedener Hardware effizienter werden können

Traditional ML



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

Foundation Models

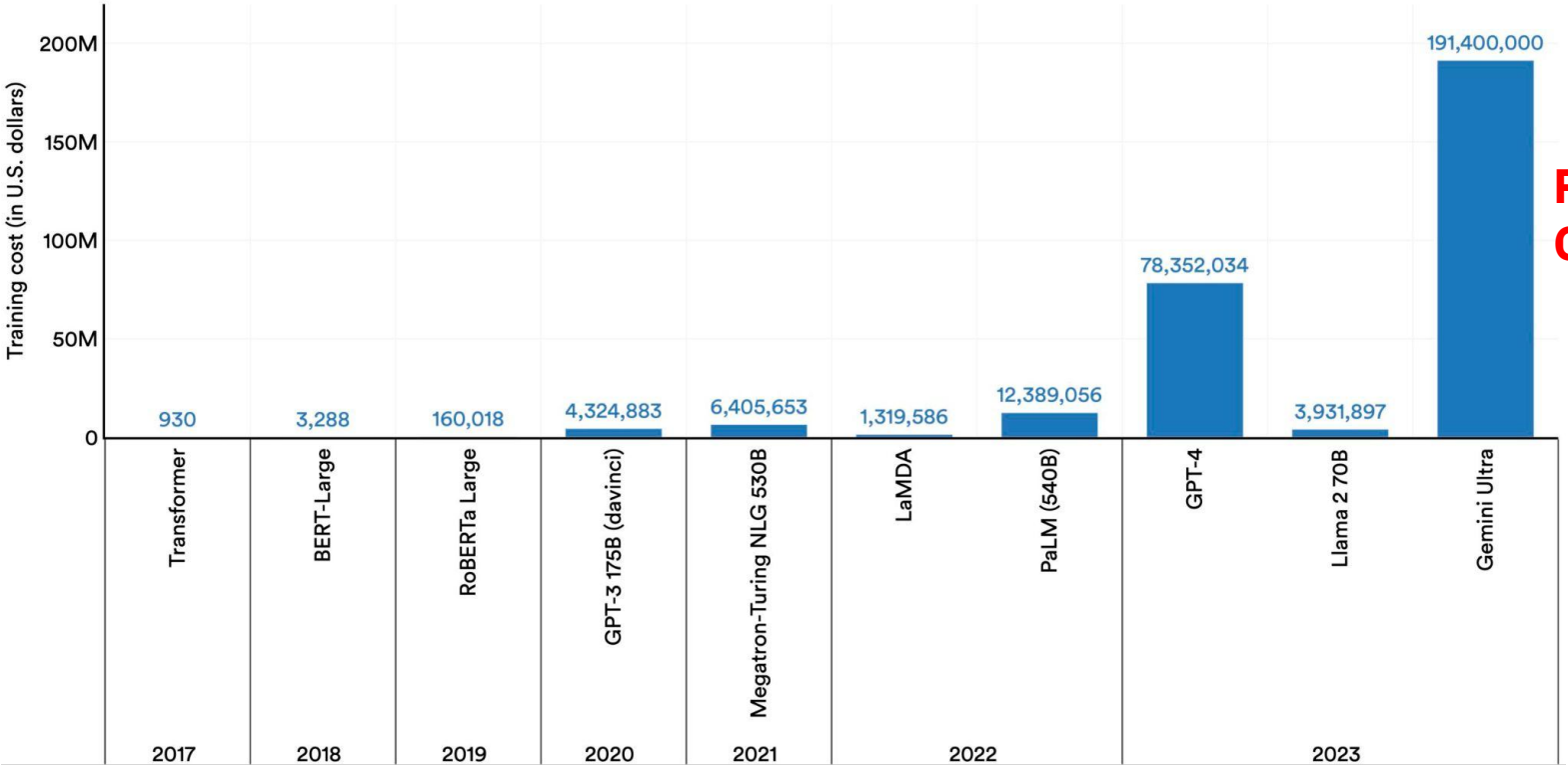


- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

Credits: [Armand Ruiz| Conor Kelly]

Estimated training cost of select AI models, 2017–23

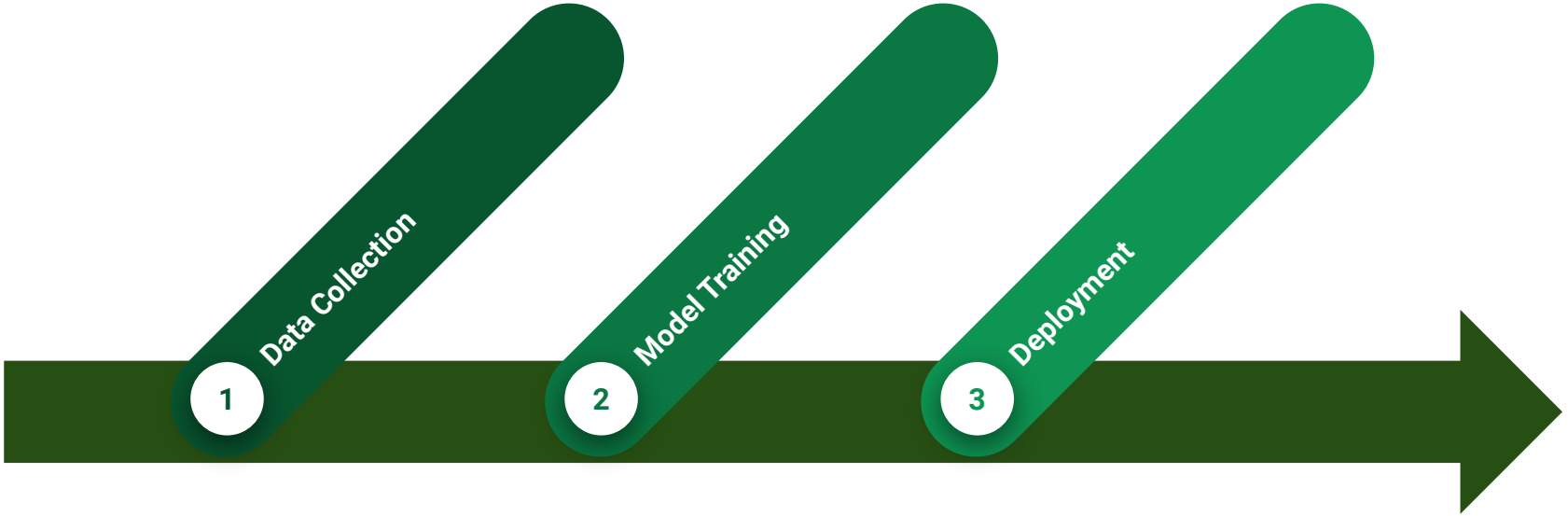
Source: Epoch, 2023 | Chart: 2024 AI Index report



Power Consumption!

Credits [Stanford AI Index]

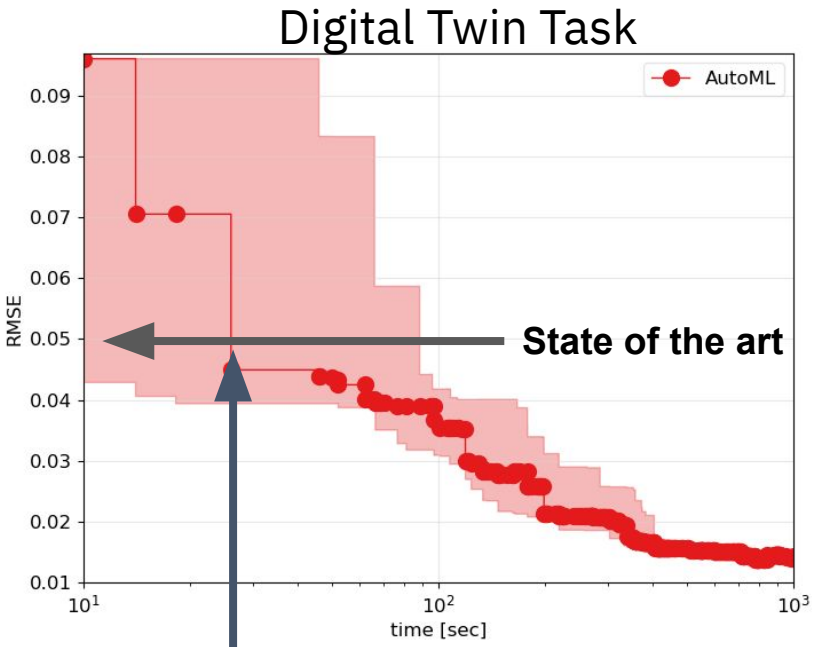
Old School Machine Learning



Why AutoML?

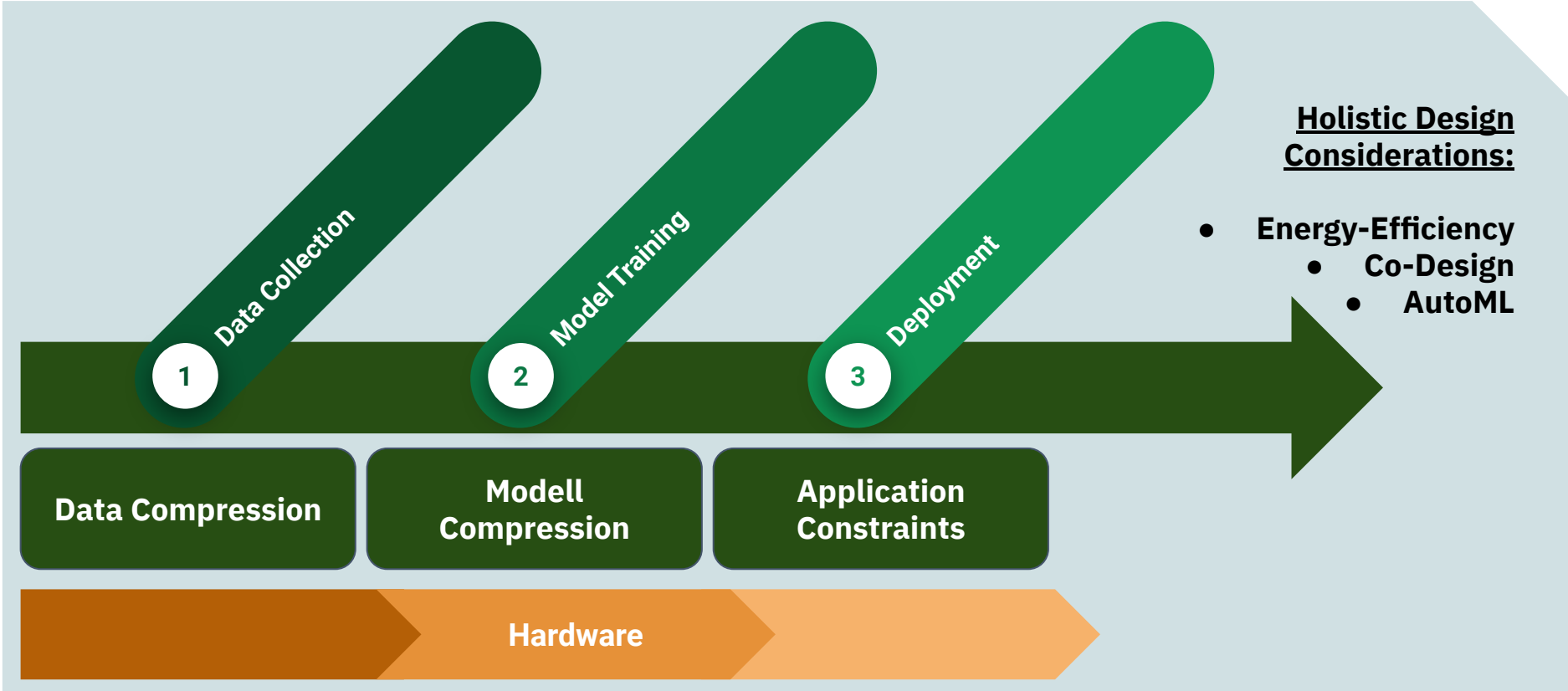
AutoML enables

- 🚀 More **efficient** research and development of ML applications
  → AutoML has been shown to outperform humans on subproblems
- 🧮 More **systematic** research and development of ML applications
  → no (human) bias or unsystematic evaluation
- 📋 More **reproducible** research
  → since it is systematic!
- 🦾 **Broader use** of ML methods
  → less required ML expert knowledge
  → not only limited to computer scientists



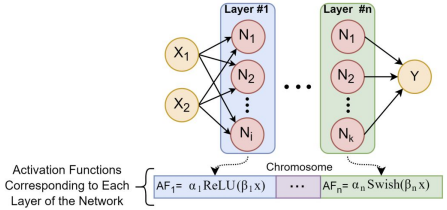
Better than state of the art in less than 30sec!

Efficient Machine Learning

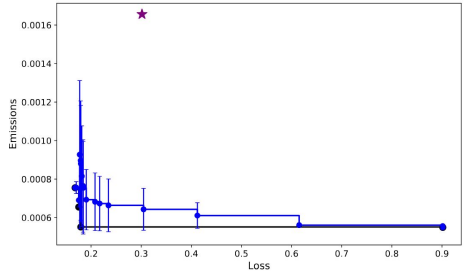


AutoML for Energy-Efficiency AI-Design

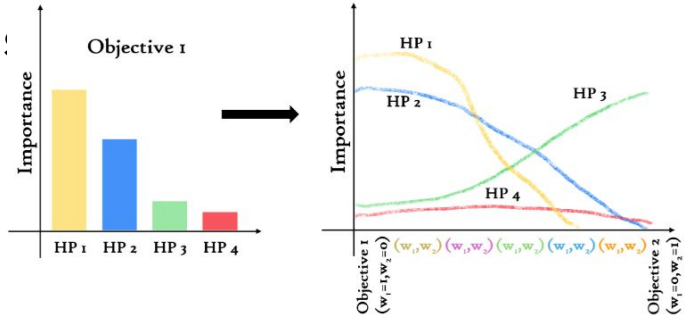
- **Level 1:** Adjusting the Network for Pruning
  [Loni et al. AutoML'23]
  - Automatically Optimizing the activation function for pruned networks



- **Level 2:** Automatically Designed Shift Networks
  [Hennig and Lindauer. PML4LRS Workshop @ ICLR'24]
  - Multi-fidelity and Multi-Objective Optimization
  - Pareto front of new state-of-the-art models



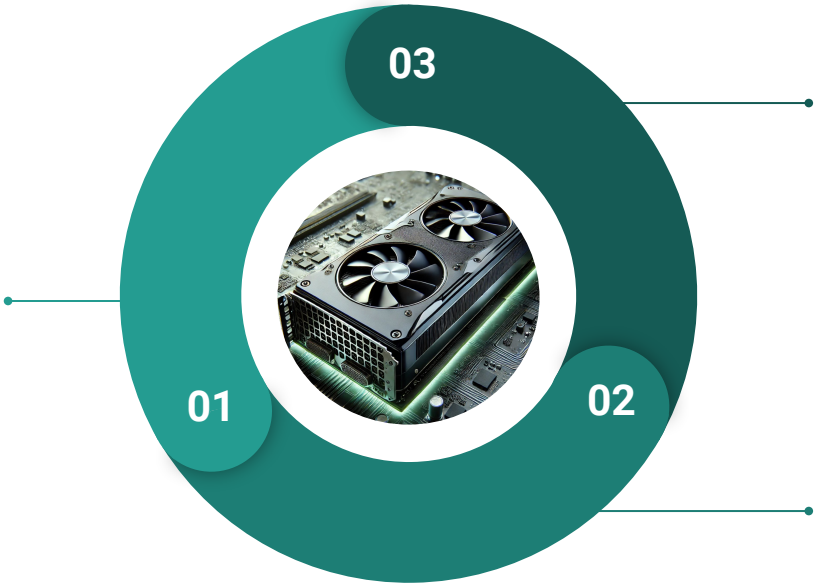
- **Level 3:** Analysis of important design decisions
  [Theodorakoloulos et al. ECAI'24]
  - Extending ablation analysis and fANOVA for multi-objective HPO



Design Loop of AI

Sampling of an AI-Design

- Manual search
- Random search
- Bayesian optimization
- Genetic algorithms
- ...



Evaluations

- Resampling strategy
- Data subsamples
- Proxy metrics
- ...

Training of an AI Model

- Early stopping,
- Learning curve extrapolation
- Scaling laws
- Mutli-Fidelity
- ...

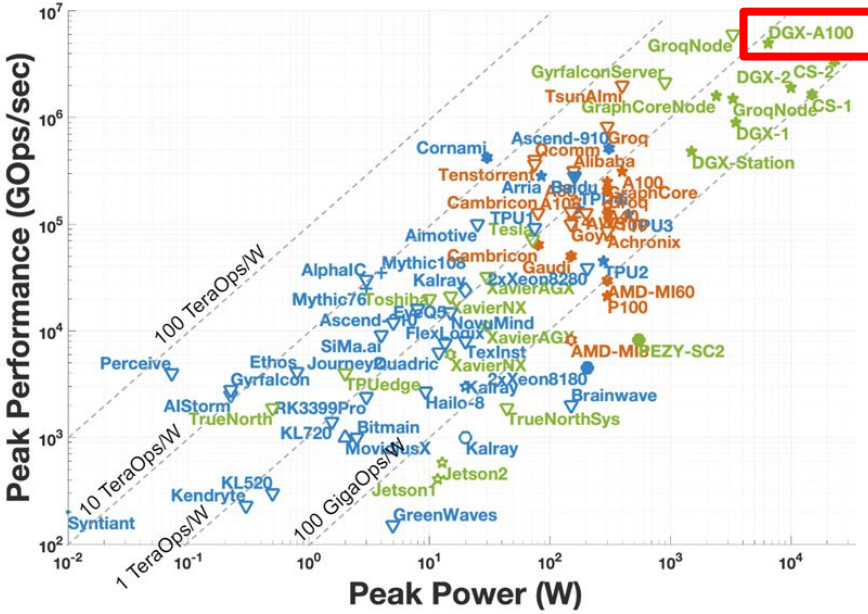
Overview - Current AI hardware

Slide inspired by Matthias Lüders

Complex AI algorithms need powerful hardware

- Various application areas for AI hardware
- Different hardware-types for AI computation
  - GPUs - Flexible, powerful, relatively high power consumption
  - FPGAs - High Flexible, good for reconfigurable architectures
  - NPUs - Low Flexibility, highest efficiency for NNs

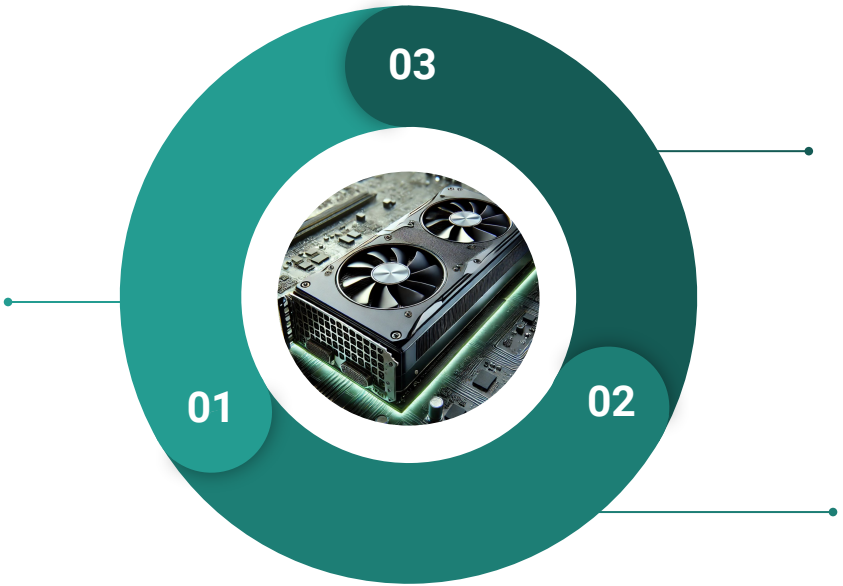
⇒ Important operators are nevertheless not efficient in hardware (for example, Softmax).



Albert Reuther et al., “AI and ML Accelerator Survey and Trends”, 2022

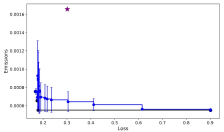
Design Loop of AI with Hardware in Mind?

Sampling of an AI-Design



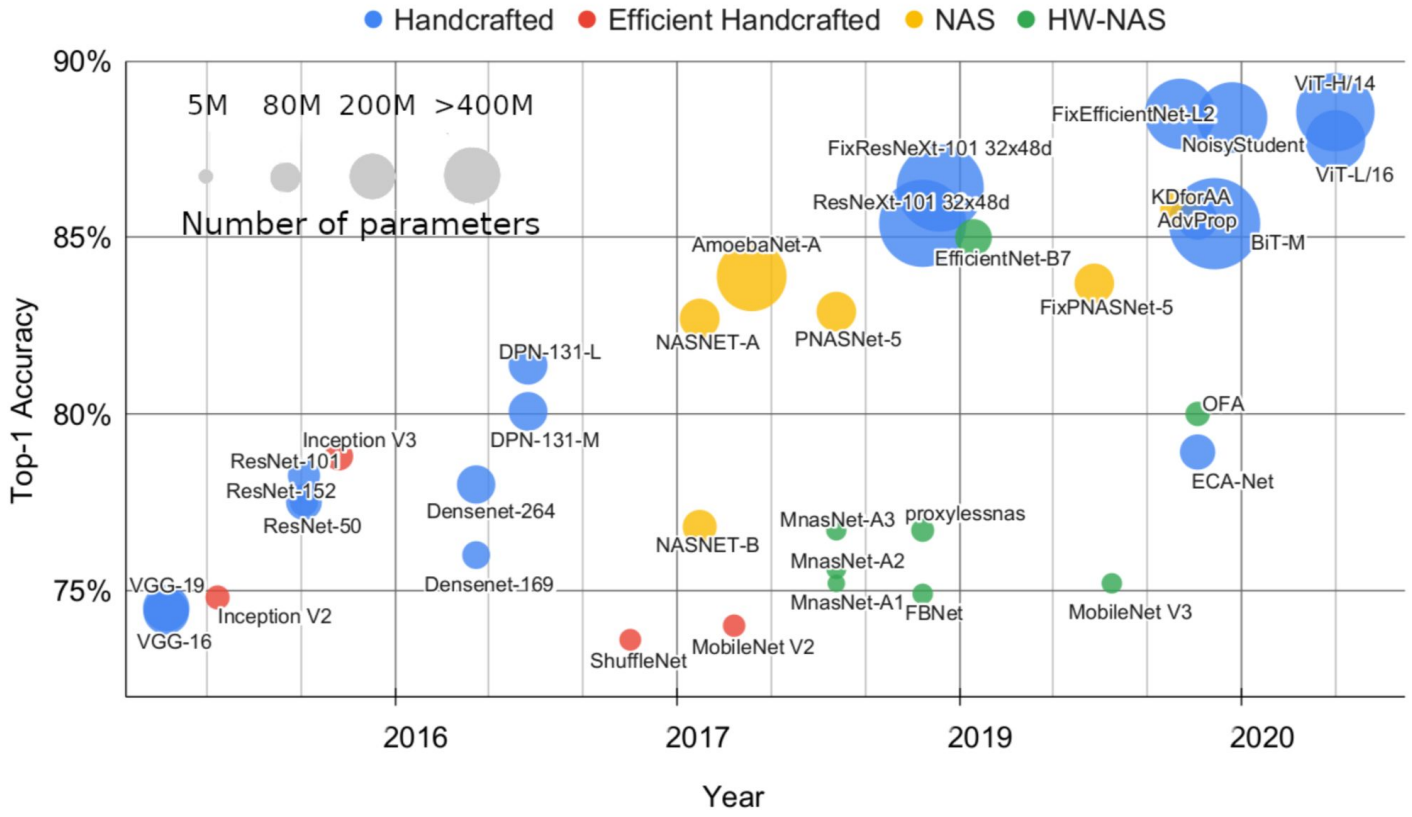
Evaluations

- Constraints on hardware capabilities
- Multiple evaluation metrics
  → Pareto fronts to check for possible trade-offs



Training of an AI Model

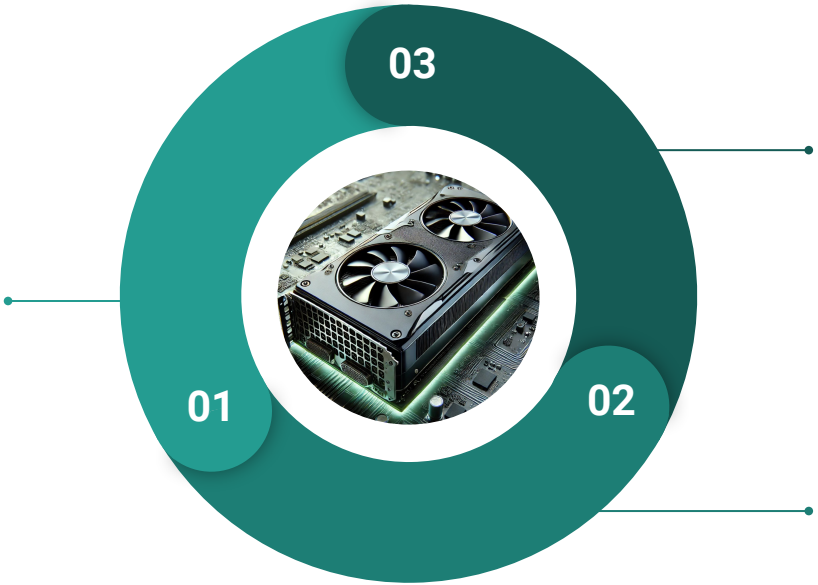
CNN-Models on Imagenet [Benmeziane et al. 2021]



Design Loop of AI with Hardware in Mind?

Sampling of an AI-Design

- **Being aware of the target hardware**
- **Consideration of training on one hardware and deploying on another**
- **Cost-aware operators in architectures**
- **Expecting that energy-mix and costs may change in the future**



Evaluations

- Constraints on hardware capabilities
- Multiple evaluation metrics
  → Pareto fronts to check for possible trade-offs

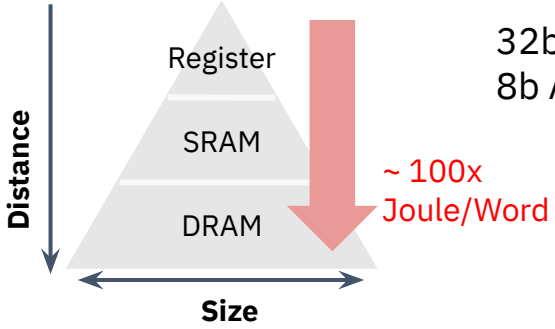
Training of an AI Model

Optimization of AI algorithms for AI Hardware

Slide inspired by Matthias Lüders

Increase energy efficiency and performance for real time processing

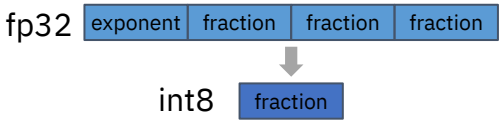
- Memory: Lower number of accesses, Local memory
- Computation: Reduce number of operations, Reduce bitwidth

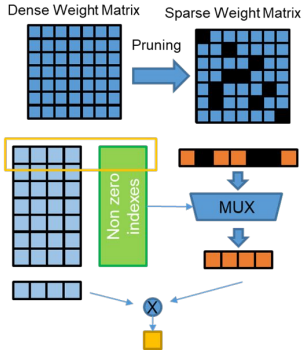


32b Add
8b Add
~ 4x
Joule/OP

Optimization-Techniques for AI algorithms

- **Quantization**: Reduce bit-width
- **Pruning**: Reduce number of weights
- **Layer Fusion**: Reduce number of layer
- **Conv-Types**: Using e.g. transposed conv

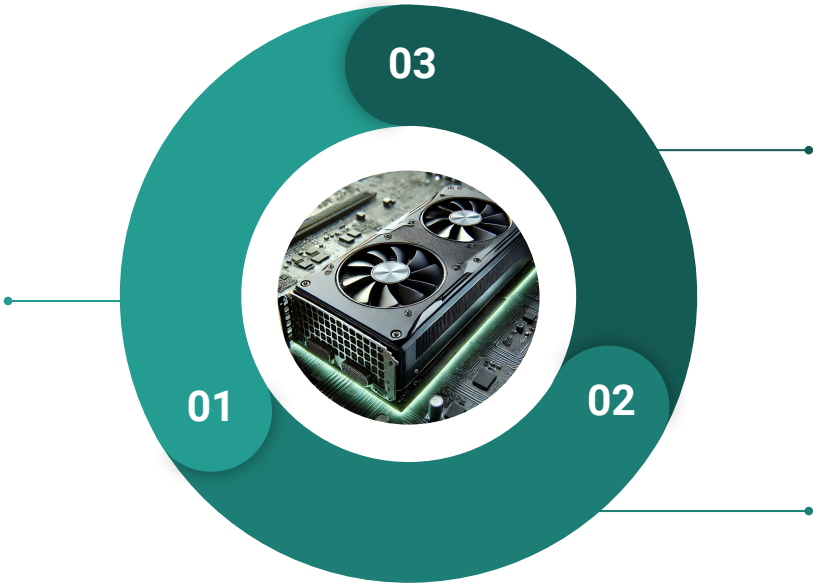




Design Loop of AI with Hardware in Mind?

Sampling of an AI-Design

- Being aware of the target hardware
- Consideration of training on one hardware and deploying on another
- Cost-aware operators in architectures
- Expecting that energy-mix and costs may change in the future



Evaluations

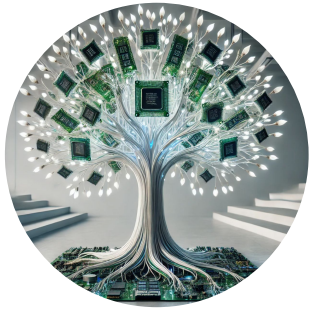
- Constraints on hardware capabilities
- Multiple Evaluation metrics
  → Pareto-Fronts to check for possible trade-offs

Training of an AI Model

- **On which hardware plattform?**
- **Training for how long?**

Chip People

Too slow in developing new AI chips?



Too little understanding?

AI People

KI wächst!

Hält Hardware mit?

Wir müssen reden!

Find Us

AutoML.org

automl-org

automl

@AutoML_org

LUH|AI

luh-ai

LUH-AI

@luh-ai

Funded by:

Backup Slides

AI @ L3S: Excellent research in Europe, innovation in Niedersachsen

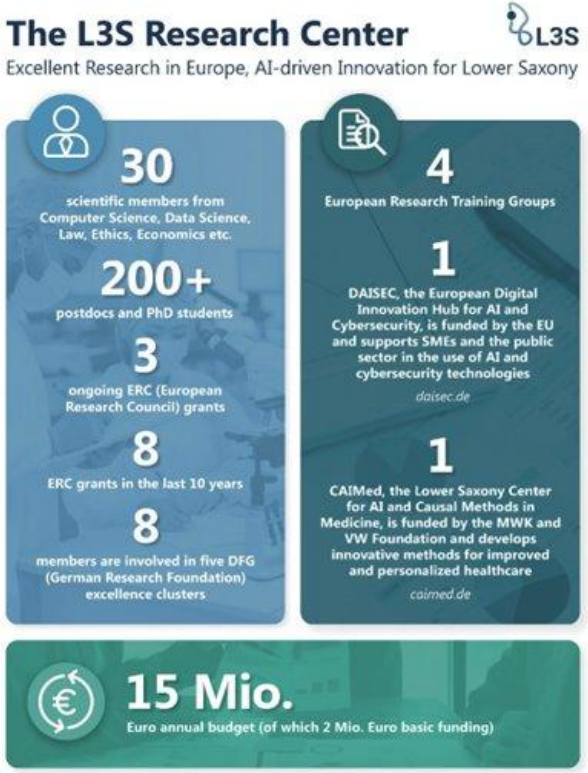
Research: Intelligent, reliable and responsible systems

- 4 European Research Training Groups (ITNs NoBIAS, KnowGraphs, ...)
- 8 ERC grants in the last 10 years (AutoML, ScienceGraph, Cont4Med, ...)
- 25+ L3S members, mainly from Hanover and Brunswick
- 200+ postdocs and PhD students

Innovation and research in interdisciplinary groups

- Intelligent production
- Intelligent mobility
- Personalized medicine
- Digital education
- as well as biodiversity, media startups, quantum computing ...

- 15 million € annual budget (of which 2 million € basic funding)
- thus the largest AI research center in Lower Saxony



What else can Green AutoML do for us?

[Tornede et al. JAIR'23]

Energy-efficient AutoML

Data compression,
Zero-cost AutoML,
multi-fidelity,
intelligent stopping, …

Searching for Energy-Efficient Models

Model size constraint,
Energy-aware objective functions,
Energy efficient architectures,
Model compression, …

+

AutoML for Sustainability

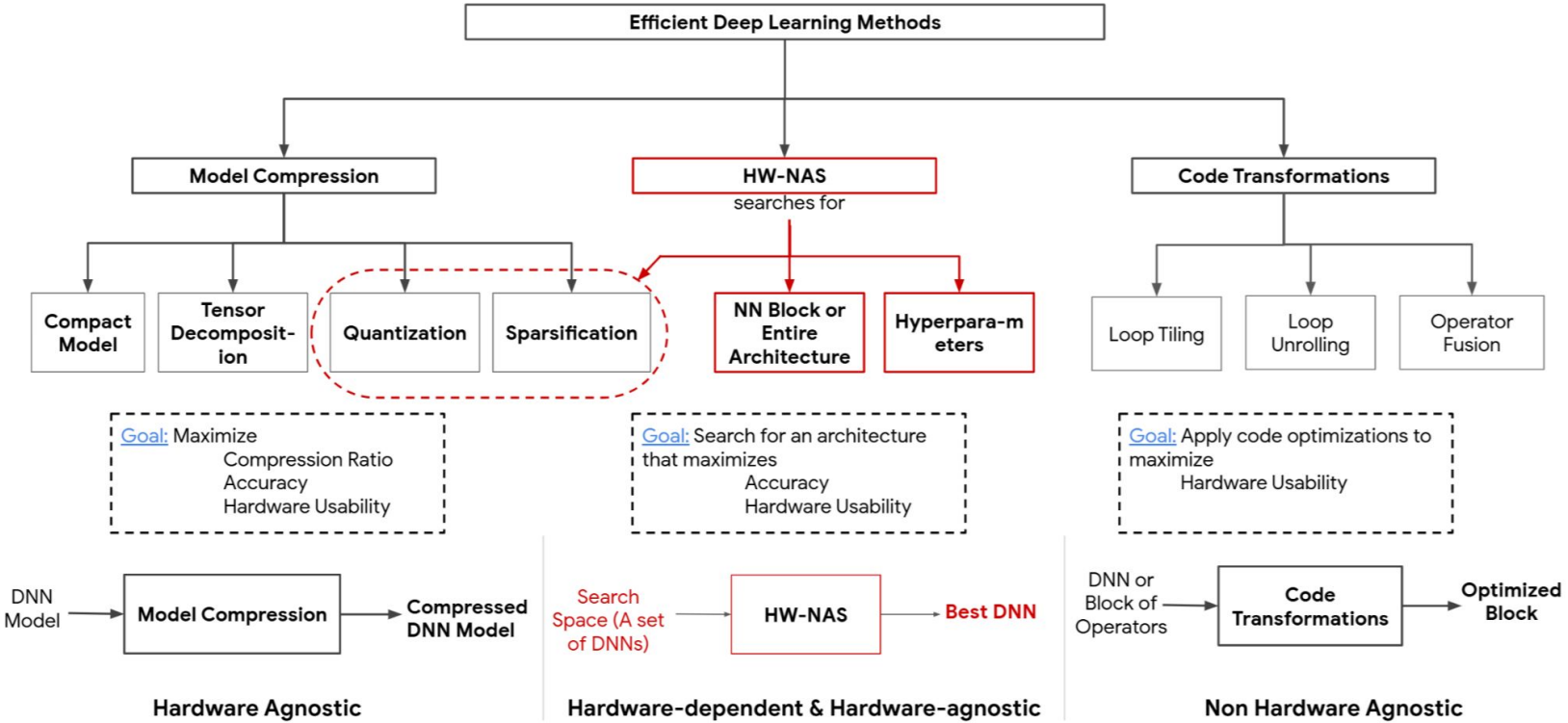
Plastic Litter Detection,
Green Assisted Driving,
Predictive Maintenance, …

Create Attention

Tracking emissions,
awareness among students,
researchers, industry partners, …

Overview of “Efficient” Deep Learning Techniques

[Benmeziane et al. 2021]



Is AutoML Green?

[Neutatz et al. EDBT'25. To appear]

AutoML has even 3 different stages:
(i) Development of the AutoML tool,
(ii) Finding the best ML model (incl training)
(iii) Inference based on the trained model

Insights:

- Energy efficiency depends on stage and package
- We can invest more time into the package development to get better efficiency later on