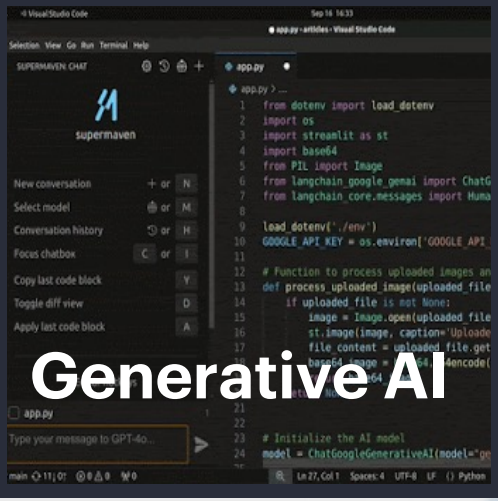
Zielgerichtet KI im Mittelstand nutzen

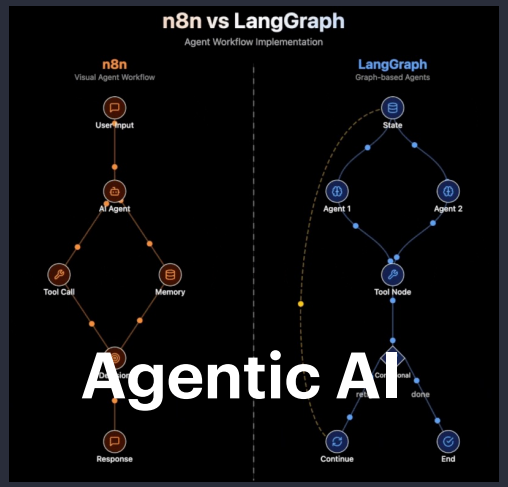
Jeremy Massow

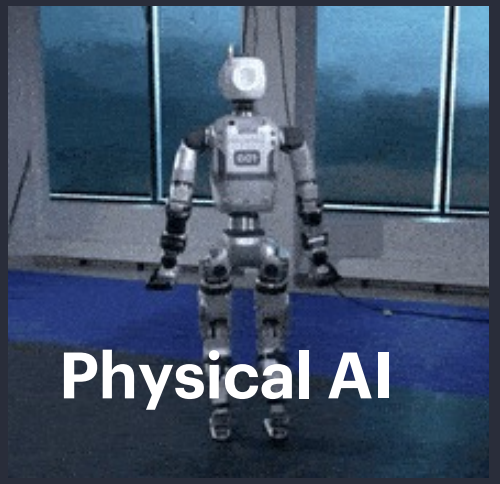
April 2026

Die drei Bereiche für

AI-GETRIEBENE TRANSFORMATION







FOMO

[Fear of Missing Out]

- Wettbewerbsdruck
- Hype-getriebene Dynamik
- Ecosystem lock-in
- Kurzfristige Erfolge

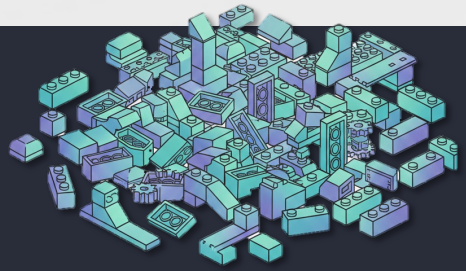
FOMU

[Fear of Messing Up]

- Integrationsrisiken
- Unsicherheit zum ROI
- Regulatorische und ethische Bedenken
- Überlastung

AI am Beispiel eines

LEGO Raumschiffs



oder

oder

DIY

Benötigt viele Bausteine, umfassendes Wissen über Lego & Star-Wars, Zeit, Trial & Error, maximale Komplexität und Risiko

Buy the kit

Bausteine und Pläne sind beinhaltet, jedoch 1.949 Einzelteile und 239 Seiten Anleitung. Signifikante Zeit und Aufwände notwendig

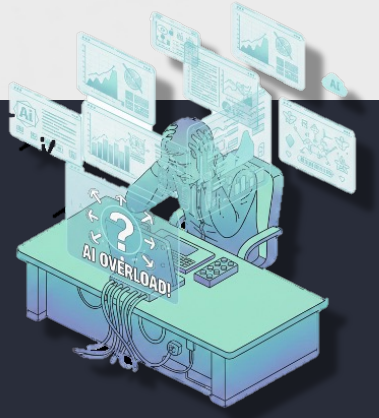
Buy the finished product

Alles ist spielfertig vorbereitet und einsatzbereit.

AI Implementierung

Der Grad an Komplexität

Laut MIT-Studie steigt die Erfolgsquote auf bis zu 67% wenn Unternehmen auf externe / vorgefertigte Lösungen statt Eigenentwicklungen setzen



DIY

Erfordert umfassende Infrastrukturkenntnisse, umfassende Softwarekenntnisse, Zeit. maximale Komplexität und Risiko

oder

Buy the kit

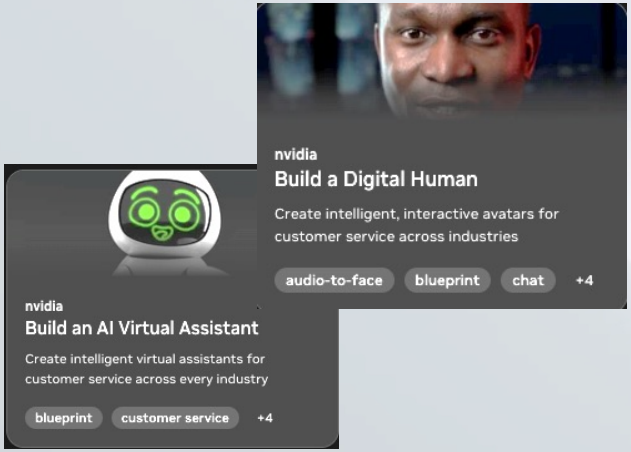
Vorkonfigurierte Infrastruktur mit teilweise installierter Software. Noch umfangreiche Arbeiten erforderlich, um Lösung einsatzbereit zu machen.

oder

Buy the finished product

Vollständig integrierte Lösung, mit minimalem Einrichtungsaufwand.. Ideal, wenn Sie Ergebnisse statt Komplexität wünschen.

Unleash AI
ISV Partner Program



NVIDIA Blueprints

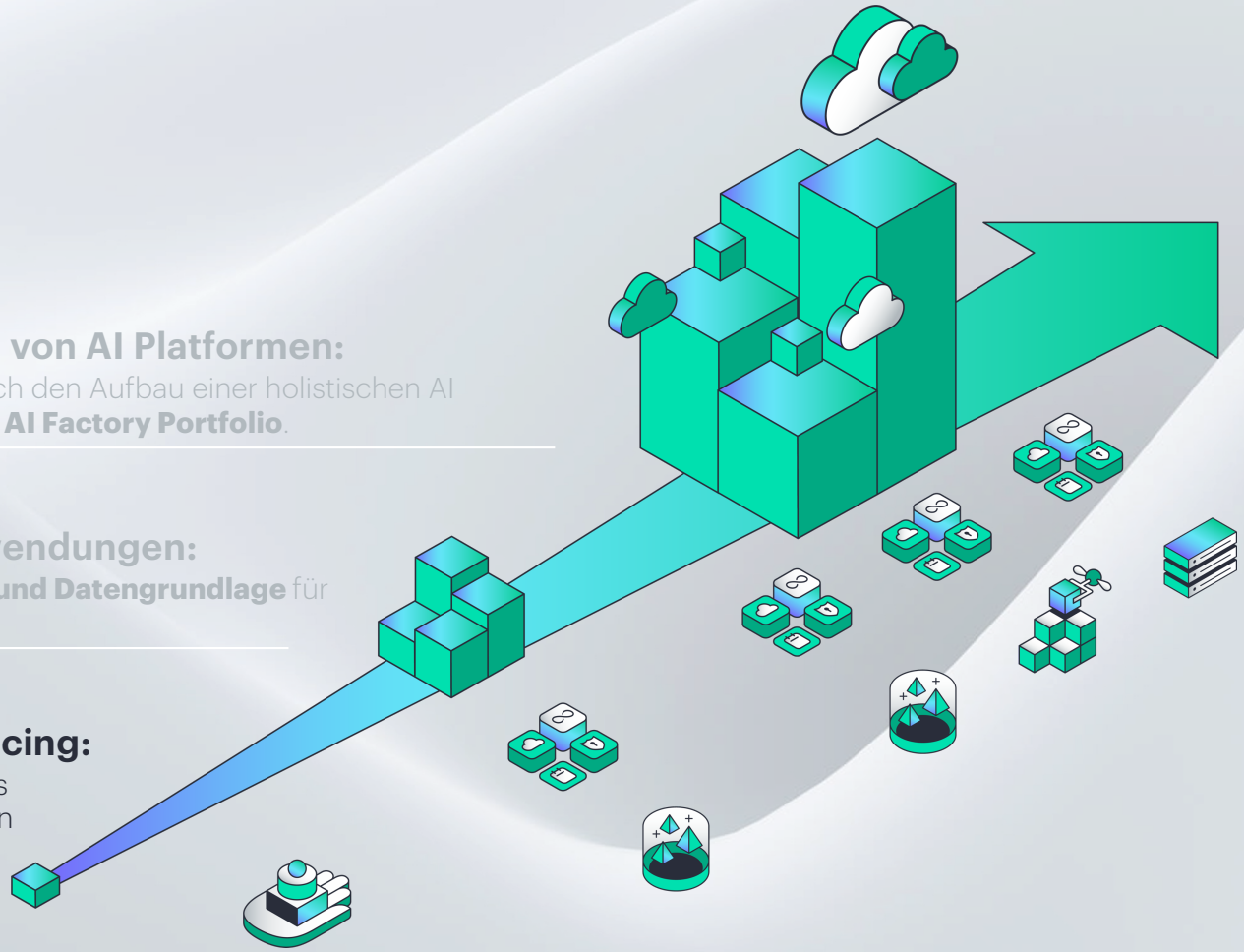
AI21labs, AIBLE, Adaptive ML, Aleph Alpha, Ansys, big.id next, blackshark.ai, clarifai
ClearML, Cohesity, Contextual AI, CrowdStrike, crewai, DagsHub, dataiku, DataQI
DAXA, Deepgram, Epic IO, Fortanix, GAMBIT, Iterate.ai, ITERNAL, KAMIWAZA
Lighton, nara logics, Neuroscale AI, NEXUS-MD, OPEN INNOVATION, opentext, Personal AI, ProHawk
PROTOPIA, QUANTPI, Reka, securiti, Sickbay, smart spatial, tabnine, teradata.
træfiklabs, UNIQUE, VAIDIO, cinity, LEGION

In 3 Leveln zur AI-Mastery

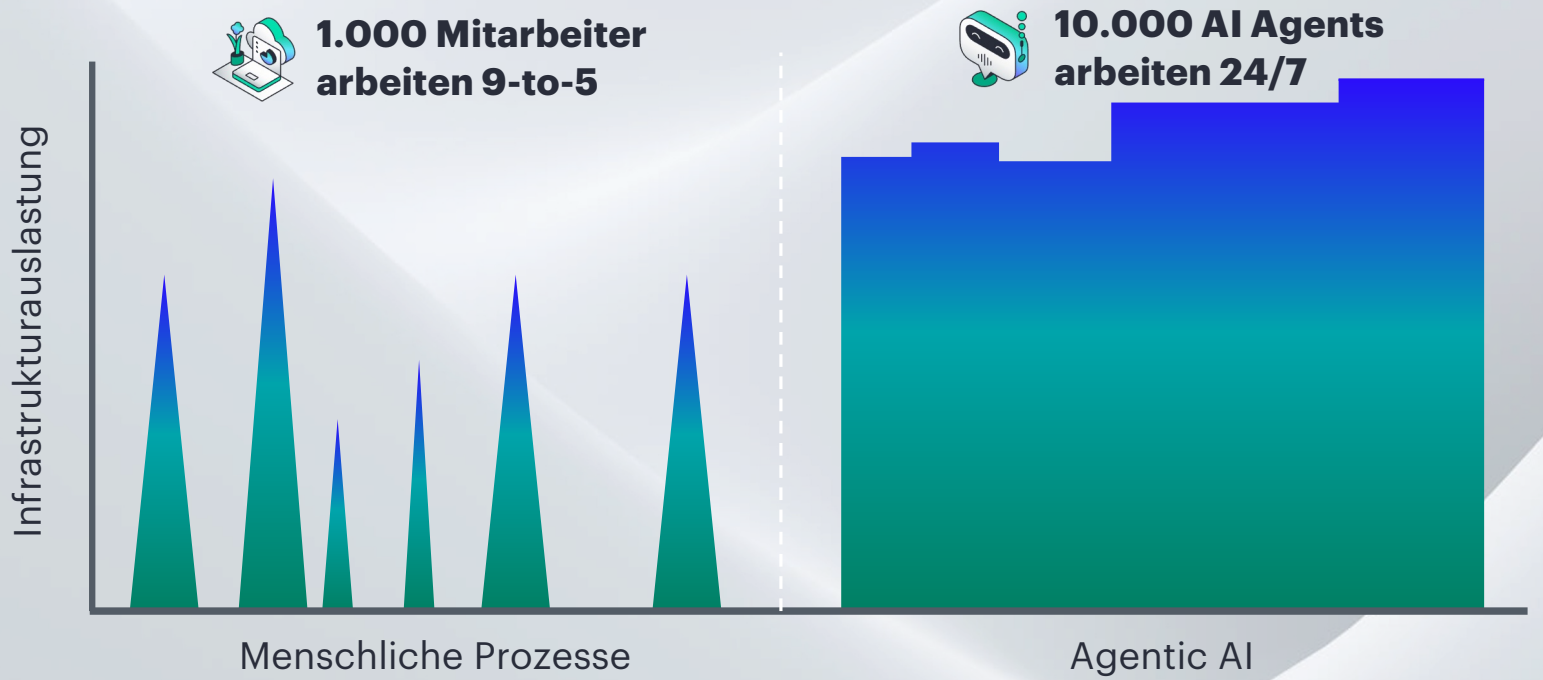
3. **AI Mastery durch die Nutzung von AI Platformen:**
Zentralisierung der AI Kompetenzen durch den Aufbau einer holistischen AI Platform wie **Private Cloud AI und dem AI Factory Portfolio**.

2. **Networking & Data für AI Anwendungen:**
Schaffung einer **AI-fähigen Netzwerk und Datengrundlage** für zukünftige AI Anwendungen.

1. **Rechenzentrum für AI Inferencing:**
Nutzung des KI-fähigen Server Portfolios **inkl. 3rd party SW** und Anforderungen an Rechenzentren.



Inferencing außerhalb der “Business Hours”



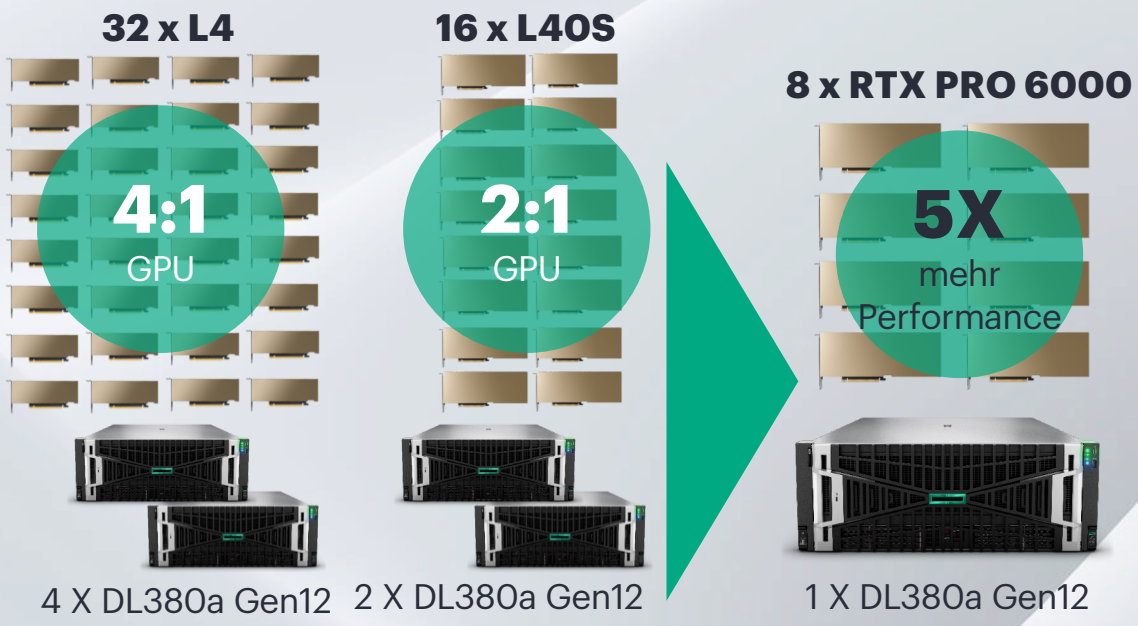
GPU Performance Sprünge sind massiv

Bis zu:

5X mehr Performance

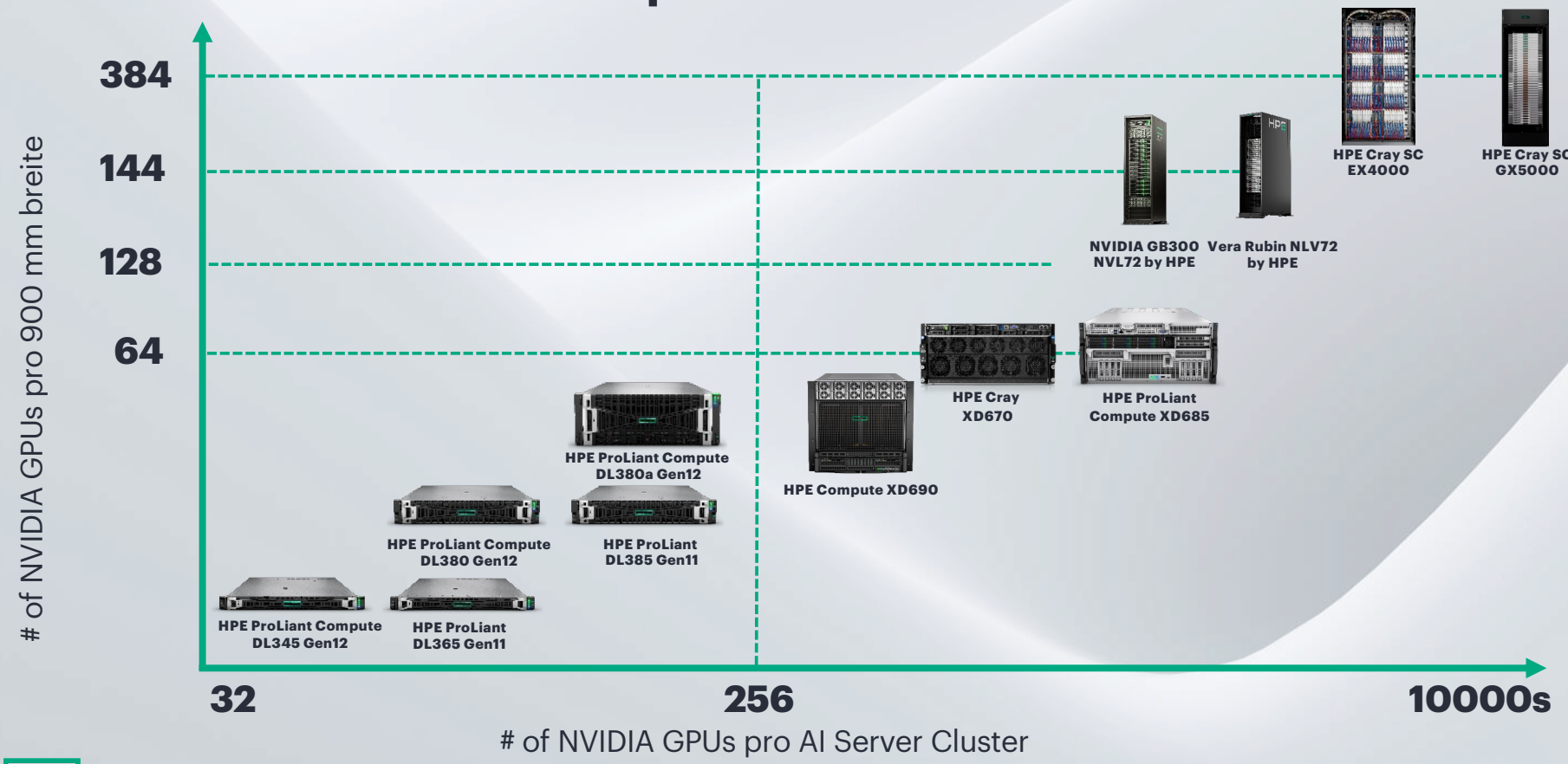
48% niedrigere Kosten

25% Energieeinsparung



Kostenersparnis	**41%**	**48%**
Energieersparnis		**25%**

Nicht nur das breiteste Portfolio am Markt, sondern auch die größte NVIDA GPU Dichte pro Rack!



Performancedichte ist gut aber, kann das RZ mithalten?

„In mein Rack gehen nur 10 kW“

„Meine Kühlung ist darauf nicht ausgelegt“

„Mein RZ ist so ineffizient“

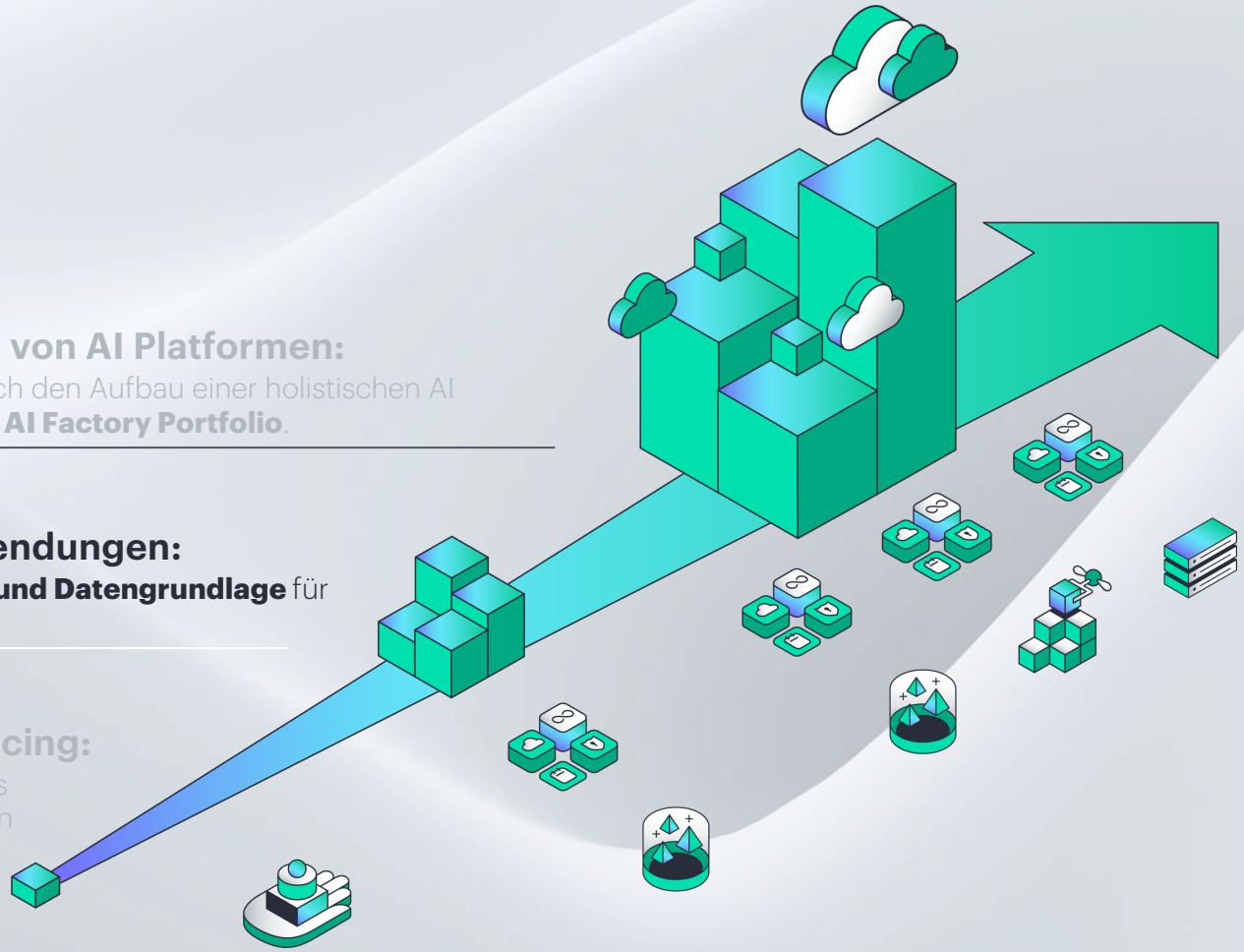
„Mein RZ kann keine Wasserkühlung“

In 3 Leveln zur AI-Mastery

3. **AI Mastery durch die Nutzung von AI Platformen:**
Zentralisierung der AI Kompetenzen durch den Aufbau einer holistischen AI Platform wie **Private Cloud AI und dem AI Factory Portfolio**.

2. **Netzwerk & Daten für AI Anwendungen:**
Schaffung einer **AI-fähigen Netzwerk und Datengrundlage** für zukünftige AI Anwendungen.

1. **Rechenzentrum für AI Inferencing:**
Nutzung des KI-fähigen Server Portfolios **inkl. 3rd party SW** und Anforderungen an Rechenzentren.




Official Team Partner
HPE
AMG
PETRONAS
FORMULA ONE TEAM
Snapdragon
SOLERA
AMD
TeamViewer
PETRONAS
AMG
IWC
signify
PIRELLI
P ZERO

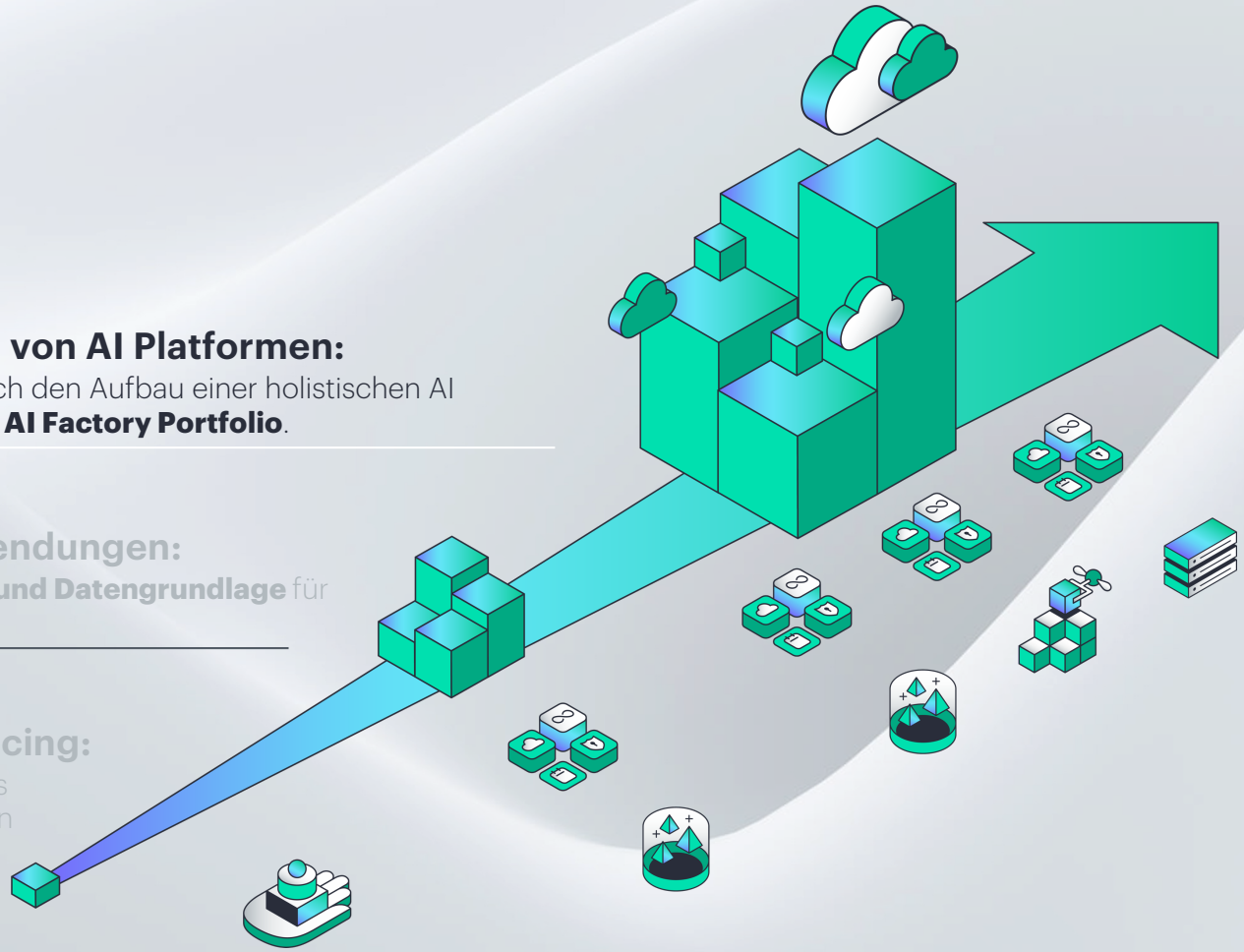
Daten speichern reicht nicht, sie brauchen Kontext

In 3 Leveln zur AI-Mastery

3. **AI Mastery durch die Nutzung von AI Platformen:**
Zentralisierung der AI Kompetenzen durch den Aufbau einer holistischen AI Platform wie **Private Cloud AI und dem AI Factory Portfolio**.

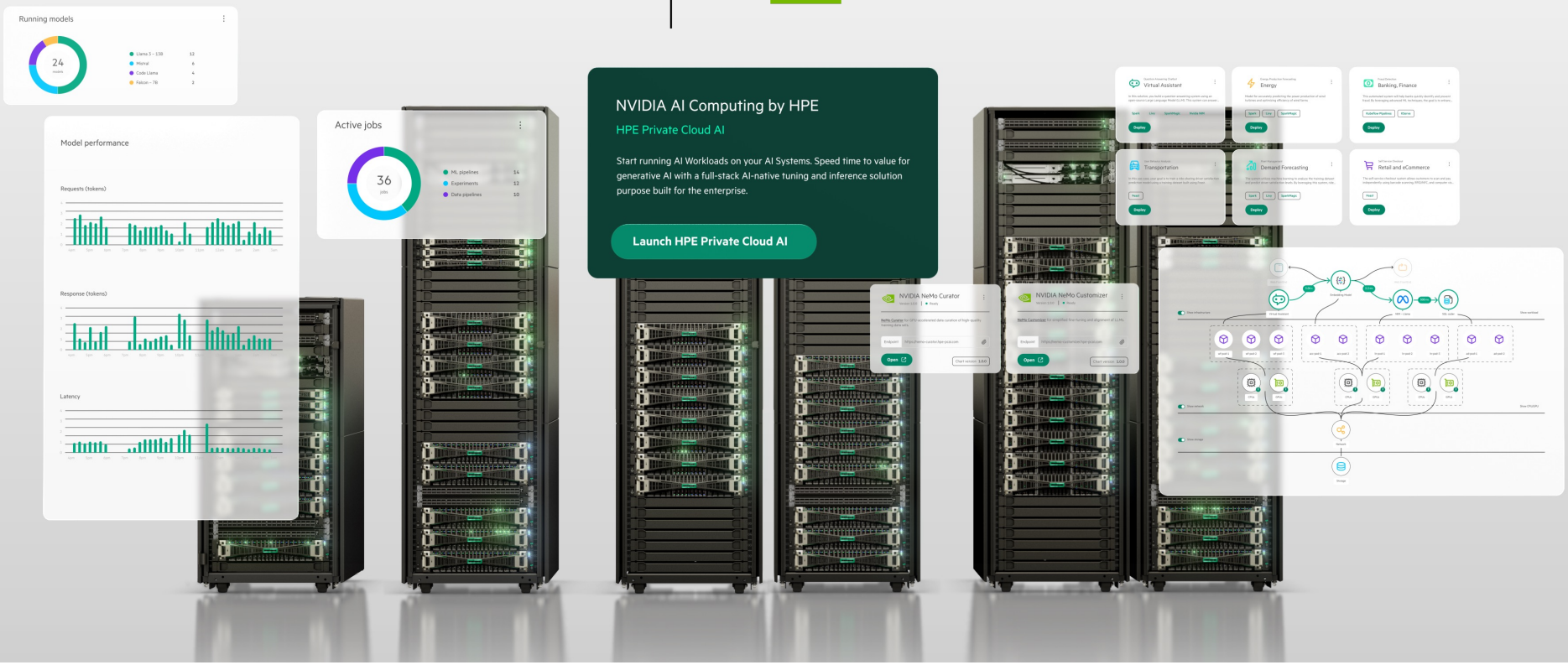
2. **Netzwerk & Daten für AI Anwendungen:**
Schaffung einer **AI-fähigen Netzwerk und Datengrundlage** für zukünftige AI Anwendungen.

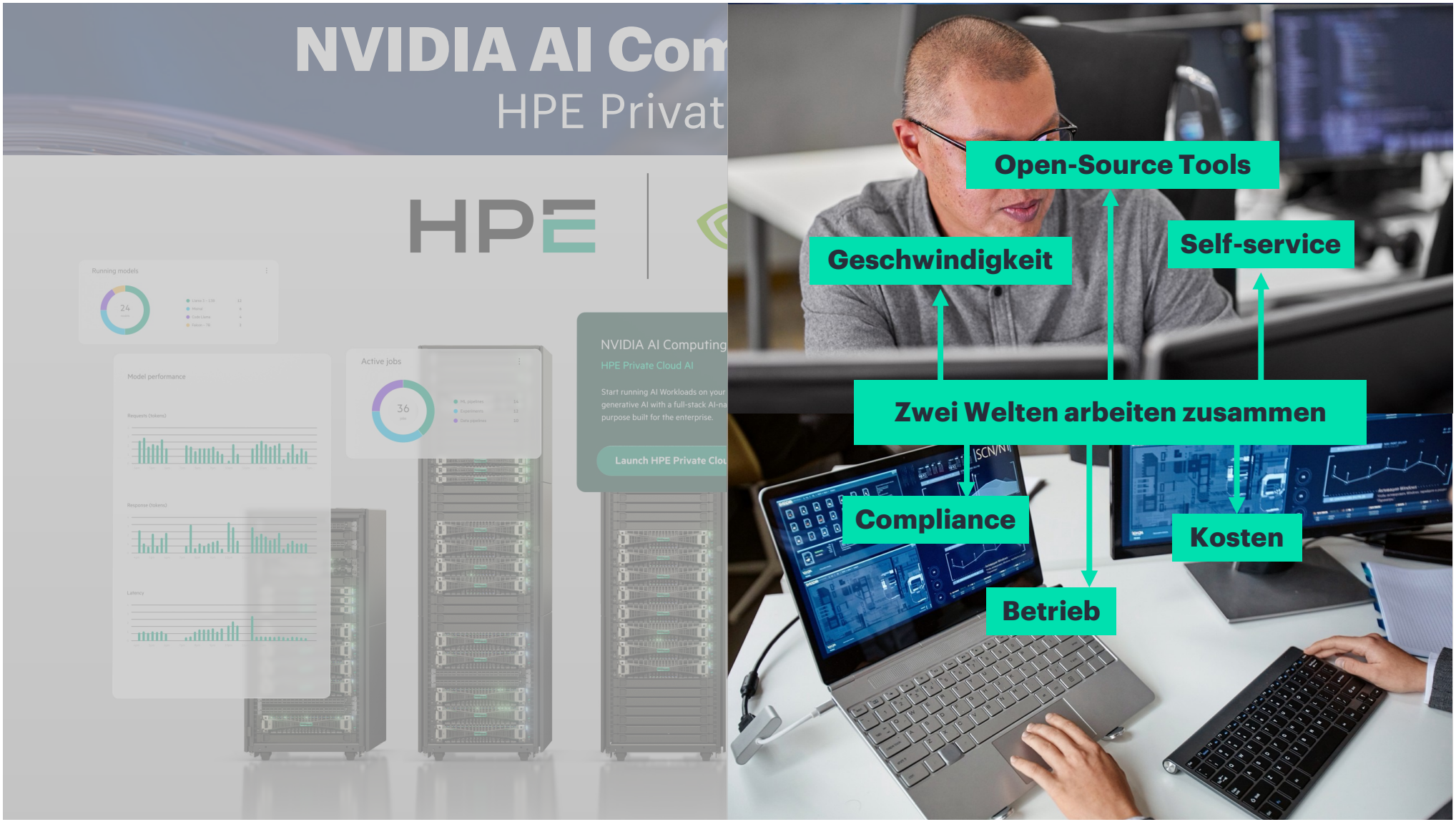
1. **Rechenzentrum für AI Inferencing:**
Nutzung des KI-fähigen Server Portfolios **inkl. 3rd party SW** und Anforderungen an Rechenzentren.



NVIDIA AI Computing by HPE

HPE Private Cloud AI




NVIDIA AI Com
HPE Privat
HPE
Running models
24
Model performance
Requests (tokens)
Response (tokens)
Latency
Active jobs
36
NVIDIA AI Computing
HPE Private Cloud AI
Start running AI Workloads on your
generative AI with a full-stack AI-na
purpose built for the enterprise.
Launch HPE Private Clou
Open-Source Tools
Geschwindigkeit
Self-service
Zwei Welten arbeiten zusammen
Compliance
Kosten
Betrieb

NVIDIA AI Computing by HPE

HPE Private Cloud AI

Bestandteile

KI-Modelle	Community + Partner + Custom NVIDIA NIM	
KI-Software	NVIDIA AI Enterprise	HPE AI Essentials
Infrastruktur	NVIDIA GPU/GPU NVIDIA Spectrum-X	HPE AI Servers HPE File Storage
Services	HPE	3rd Party Partners

Administration

Bereitstellung, Orchestrierung und Verwaltung von Workloads

Überwachung und E2E-Sicherheit für Anwendungen und Infrastruktur

HPE Private Cloud Control Plane

VM-Laufzeiten und Kubernetes-Cluster-Management

Einrichtung, Inventarisierung, IT- und Asset-Lebenszyklusmanagement

Größen

Entwickler | Klein | Mittel | Groß | Extra Groß

Erweiterungs Racks

Verwaltung

HPE GreenLake Cloud

NVIDIA AI Computing by HPE

HPE Private Cloud AI

Bestandteile

KI-Modelle	Community + Partner + Custom NVIDIA NIM	
KI-Software	NVIDIA AI Enterprise	HPE AI Essentials
Infrastruktur	NVIDIA GPU/GPU NVIDIA Spectrum-X	HPE AI Servers HPE File Storage
Services	HPE	3rd Party Partners

Administration

Bereitstellung, Orchestrierung und Verwaltung von Workloads

Überwachung und E2E-Sicherheit für Anwendungen und Infrastruktur

HPE Private Cloud Control Plane

VM-Laufzeiten und Kubernetes-Cluster-Management

Einrichtung, Inventarisierung, IT- und Asset-Lebenszyklusmanagement

Größen

Entwickler | Klein | Mittel | Groß | Extra Groß

Erweiterungs Racks

Verwaltung

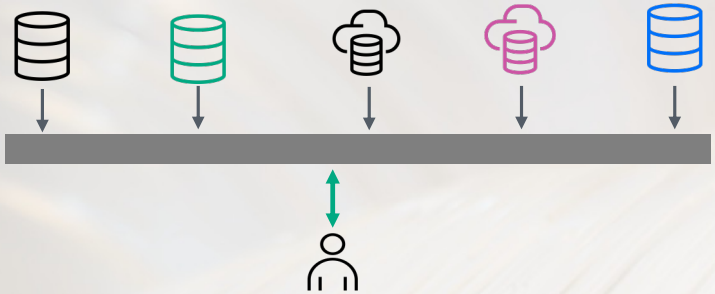
HPE GreenLake Cloud

HPE AI Essentials

Kernkomponenten

Data Engineering Tools
Airflow | Presto | Data connectors

Daten aus **verschiedenen** Quellen **sammeln**, **aufbereiten** und **bereitstellen**



Analytics & Visualization Tools
Livy | Superset | Apache Spark

Datenmengen **analysieren** und interaktive **Diagramme**, **Dashboards** oder **Reports** anzeigen

Data Science Tools
Kubeflow | Mlflow | Feast | Ray

Optimieren, **trainieren** und **versionieren** von KI-Modellen

HPE | NVIDIA

Danke!

Jeremy.massow@hpe.com