intel. Tiber™ Enterprise Solutions

# Simplify and Secure Your AI and Cloud Journey

Sneha Chattopadhyay & Avishay Sebban

July 2024

# Simplify AI

Accelerate and scale AI on managed, cost-effective infrastructure with Intel® Tiber™ Developer Cloud and AI Studio

Access cutting-edge AI-ready hardware

Comprehensive AI software stack to speed time to market

Automate the AI model lifecycle

Intel® Tiber™ Portfolio of Business Solutions

# Intel® Tiber™ Developer Cloud

## Boutique AI Cloud Services

### Build & Deploy AI at Scale

Develop models, applications & solutions, deploy production workloads at scale. Deliver **10-100X** more performance using common tools.

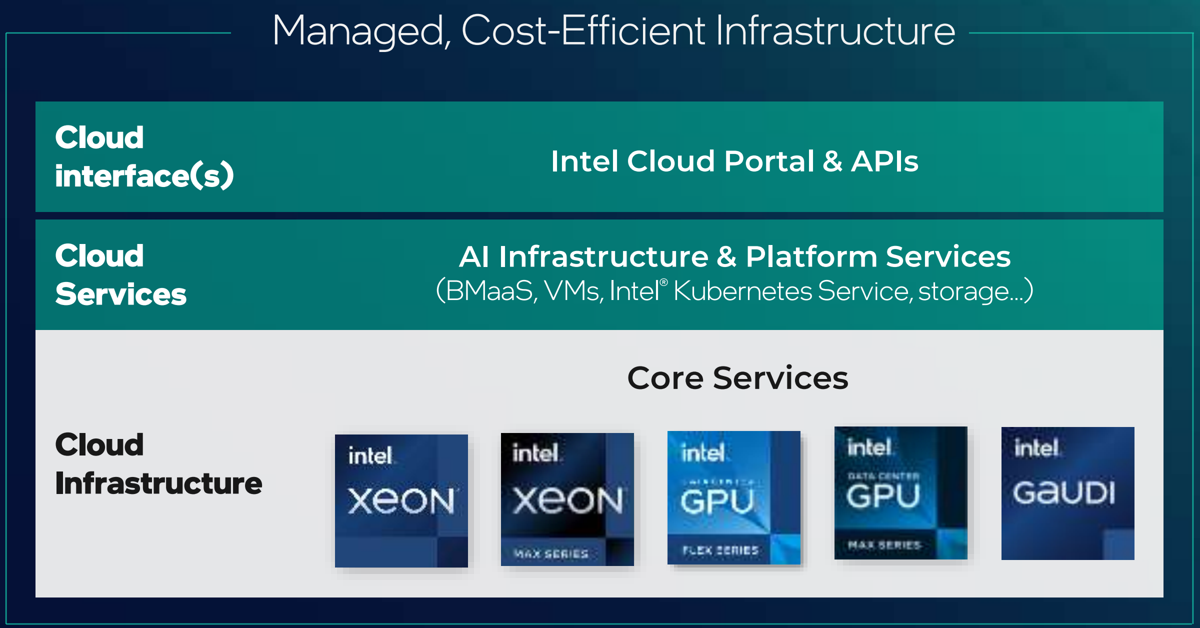
### Maximize AI Compute Resources

Choose the best accelerator for every use case for optimal price-performance.

Customers gained **up to 400%** cost savings for select AI workloads vs. on-prem or another CSP.[2]

### Open Software, Open Platform Advantage

Provides choice in hardware. Supports a wide range of optimized models, frameworks & tools. Compatible with open source LLMs & GenAI products.



Managed, Cost-Efficient Infrastructure

**Cloud interface(s)** — Intel Cloud Portal & APIs

**Cloud Services** — AI Infrastructure & Platform Services (BMaaS, VMs, Intel® Kubernetes Service, storage...)

**Cloud Infrastructure** — Core Services

Used by developers, companies & partners

Intel® Tiber™ Portfolio of Business Solutions

1. Performance varies by use, configuration, & other factors. Performance results are based on testing as of dates shown in configurations & may not reflect all publicly available updates. Learn more at www.Intel.com/PerformanceIndex & intel.com/content/www/us/en/developer/articles/technical/software-ai-accelerators-ai-performance-boost-for-free.html.
2. Prediction Guard case study, CIO.com Seekr article. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

# Trusted by Thousands of AI Experts & Devs

## Intel® Tiber™ Developer Cloud

Evaluate architecture, determine best accelerator for specific need

Deep learning/training on single node to large multi-node clusters

Develop & optimize AI applications

Build & optimize AI models

Run, deploy AI training & inference production workloads at scale

Education: AI, genAI, MLOps certification, oneAPI programming

**Achieve Significant Cost Savings & Scale**

Intel® Tiber™ Portfolio of Business Solutions seekr

# Building Trustworthy LLMs for Evaluating & Generating Content at Scale

## Solution

Trusted AI content evaluation with an LLM platform, development toolset & content analysis/scoring tool

## Pain Points

Networking, AI workloads required immense compute capacity, high infrastructure & energy costs

## Technology Foundation

Transitioned from colo provider GPU/CPU systems & on-prem Nvidia GPU A100 systems to Intel® Tiber™ Developer Cloud

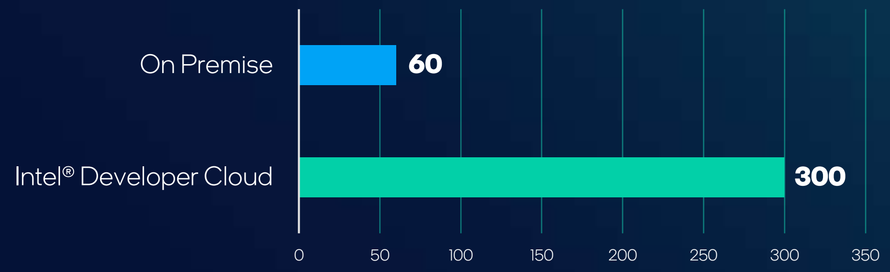
- Intel® Gaudi® 2 AI accelerators for audio scoring inference
- Intel® Max Series GPU + Intel-optimized PyTorch for audio transcription & diarization
- 4th & 5th Intel Xeon processors
- Hugging Face Optimum Habana & Habana PyTorch APIs, Intel® Kubernetes Service

Learn more: Seekr case study | Blog | CIO.com article[1] | press release

## Business Value

- **2**X inference volume, **50**% faster inference
- **20**% faster AI training
- **40**% to **400**% cost savings for select AI workloads

## Seekr LLM Inference – Requests per Second

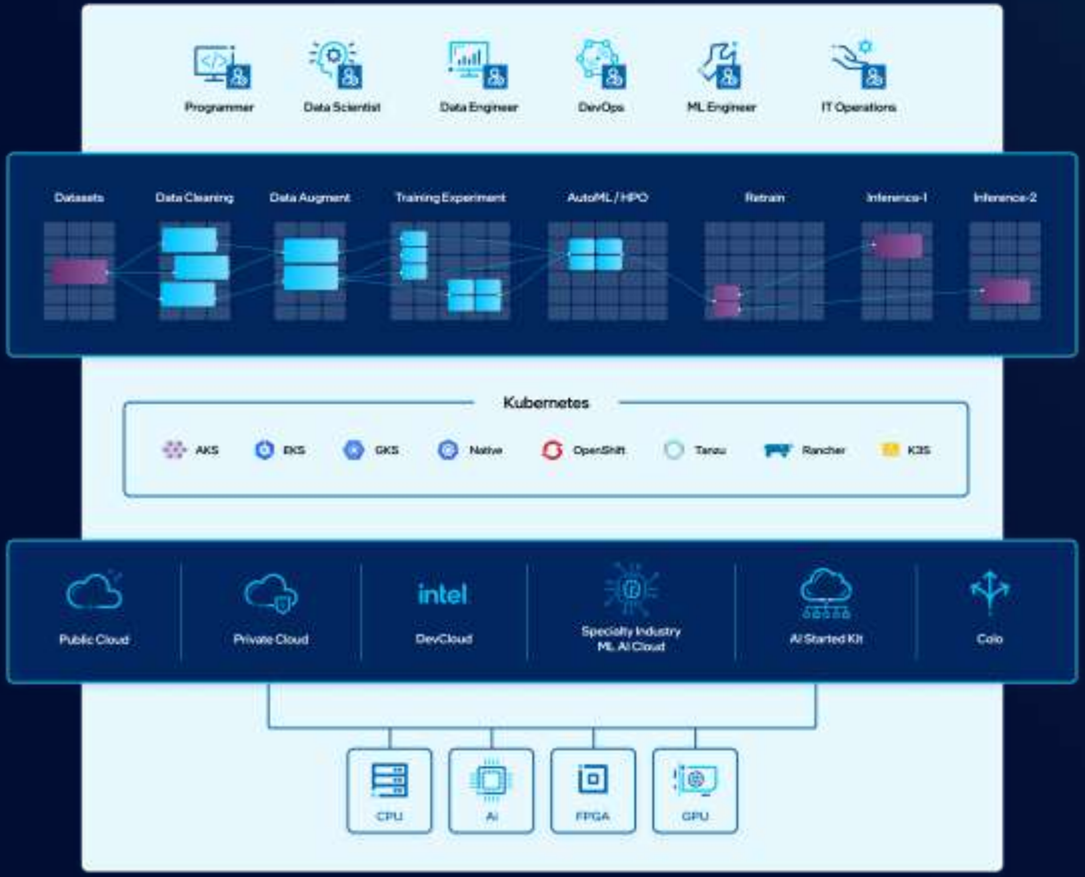


Intel® Tiber™ Portfolio of Business Solutions

Benchmark Source: Seekr. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

# Intel® Tiber™ AI Studio (formerly: cnvrg.io)

An end-to-end platform for getting better models into production, faster



Comprehensive dashboard

Software package manager

Automated pipelines

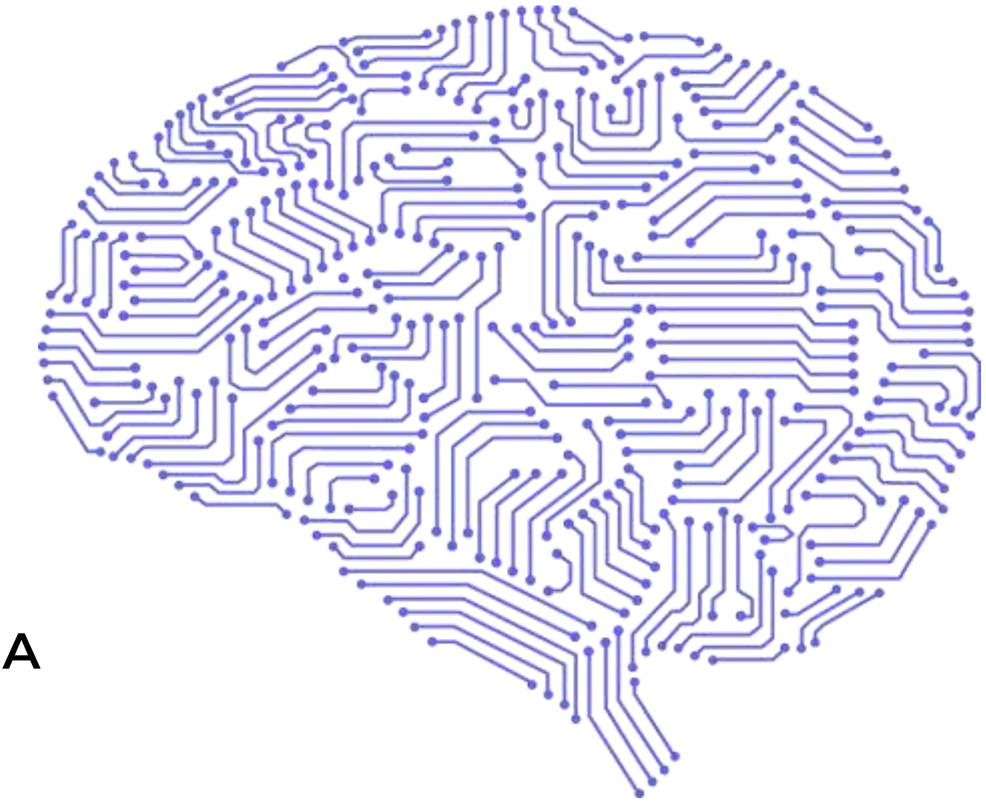
Native orchestration and scheduling

Agile compute & storage

cnvrg.io

# Continuous training and deployment of AI

Avishay Sebban, Solution Architect Lead EMEA



# Getting value from AI is hard

AI projects take too long to deliver value, if they're delivered at all

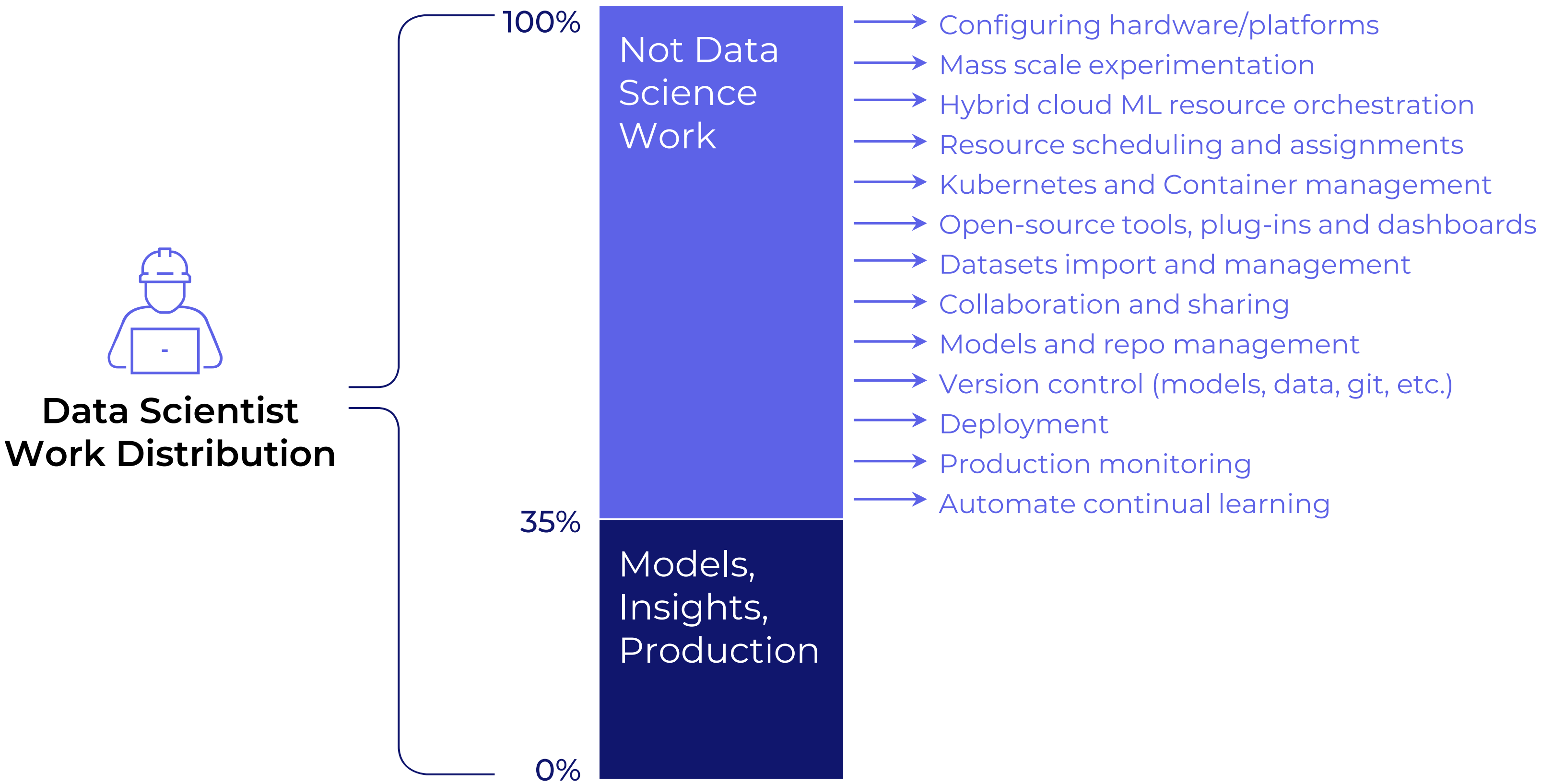
According to Gartner research, a bare majority of AI projects eventually move beyond the lab into production

**It takes an average of**

**9 months**

**to get out of the lab**

# Data scientists spend too much time on incidental tasks



Data Scientist Work Distribution

- Configuring hardware/platforms
- Mass scale experimentation
- Hybrid cloud ML resource orchestration
- Resource scheduling and assignments
- Kubernetes and Container management
- Open-source tools, plug-ins and dashboards
- Datasets import and management
- Collaboration and sharing
- Models and repo management
- Version control (models, data, git, etc.)
- Deployment
- Production monitoring
- Automate continual learning

50%
OPEX
WASTE

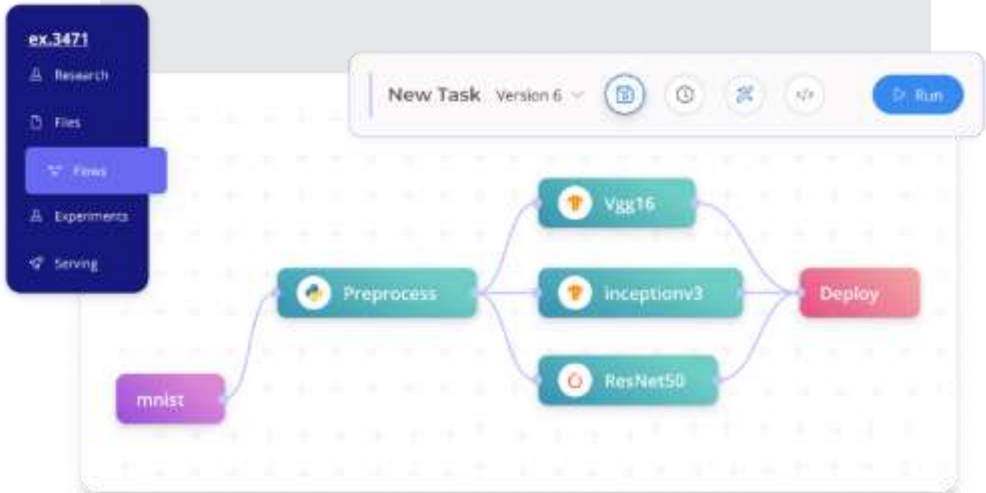
2X
**SLOWER** TIME
TO MARKET

30%
REVENUES
**LOST**

# cnvrg.io Overview

## Built by data scientists for developers of AI applications

- A platform to automate the **continuous training and deployment** of AI and ML models.
- Manages the **entire lifecycle**: data preprocessing, experimentation, training, testing, versioning, deployment, monitoring, and automatic retraining.
- Enables developers to train and deploy on **any infrastructure at scale**
- cnvrg.io **Metacloud** is the cnvrg.io platform offered as a **managed service**

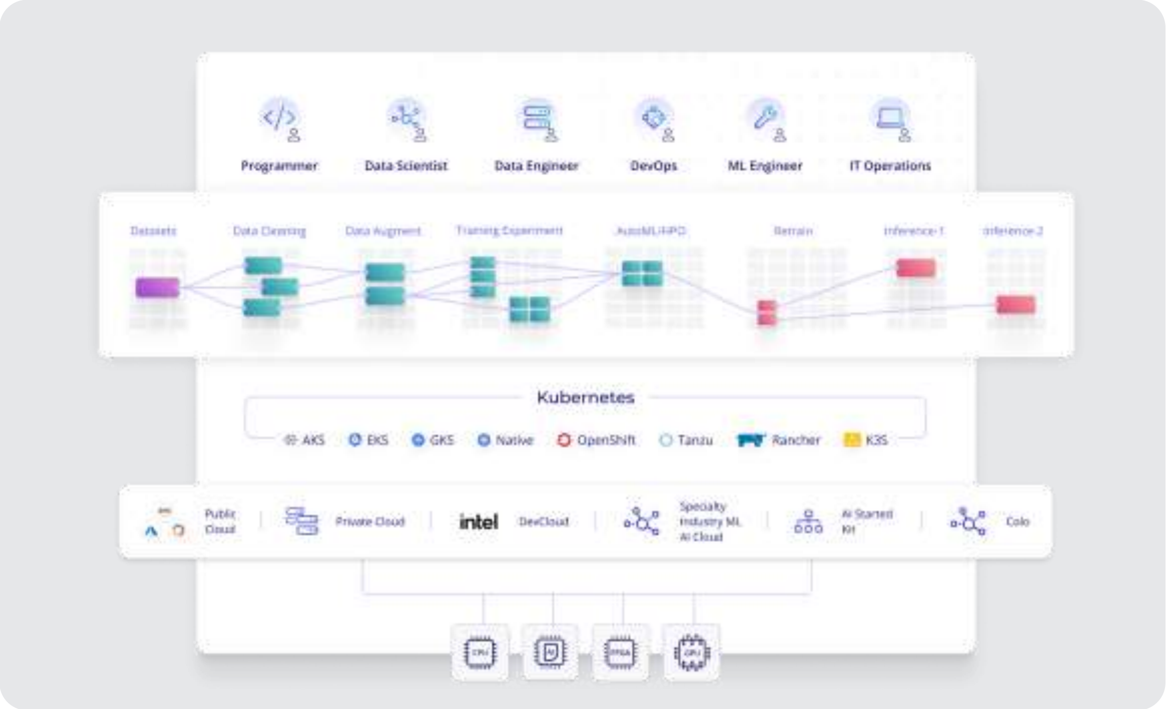


### Benefits

- Up to **10x increase in productivity**
- Up to **5x faster model training**
- Up to **50% increase in compute utilization**

# cnvrg.io: Operating System for AI

## Everything needed to build and deploy AI on any infrastructure



**Control Plane**
Management layer for datasets, model code, jobs, model performance, cluster and resource statistics

**AI Library**
Package manager for algorithms and data components, with Git integration for adding your own repositories

**Pipelines**
Drag-and-drop interface for building end-to-end ML pipelines

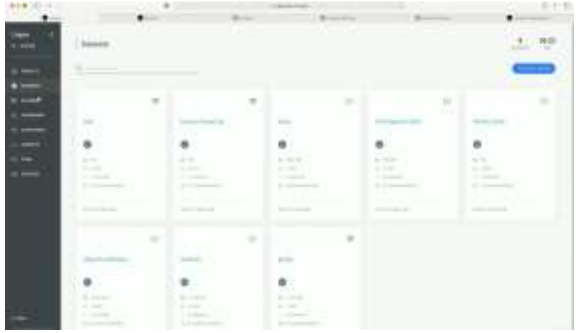
**Orchestration and Scheduling**
Kubernetes-based meta-scheduler for orchestration, scheduling, and scaling across clusters

**Compute and Storage**
Connect your own compute and storage, or choose partner-provided resources from our marketplace

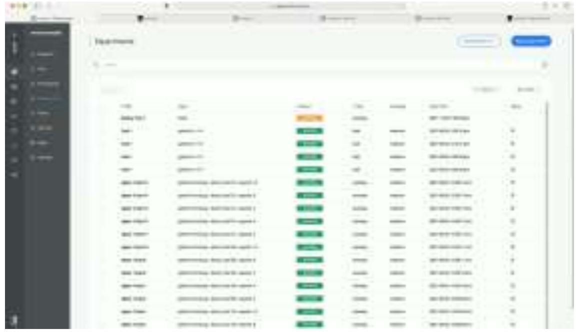
# cnvrg simplifies ML workflows from end to end

1 Create projects and workspaces

2 Connect data

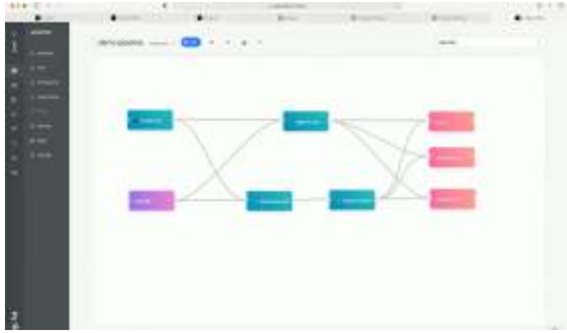


3 Manage experiments



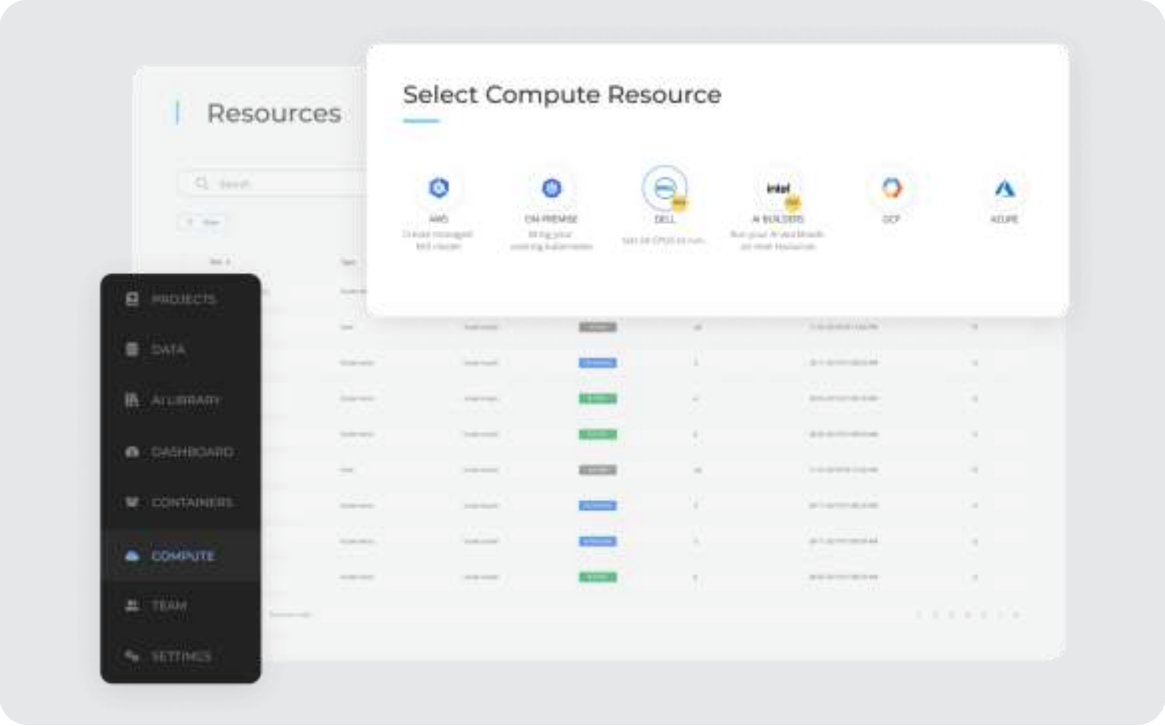
4 Create and re-use models

5 Drag-and-drop ML pipelines



6 Deploy and monitor models/clusters

# cnvrg.io Marketplace:
## AI infrastructure with cloud-native simplicity



- Self-service selection of partner and OEM compute/storage
- Choose CPUs, GPUs, Gaudi and tensors, AI accelerators, storage in customizable sizes
- Choose partner clouds, or cloud-like deployment on-prem
- No more waiting for infra and ops teams to stand up resources
- Menu-driven, point and click configuration settings
- MetaGPU Capabilities

# cnvrg.io works with all popular K8s distributions

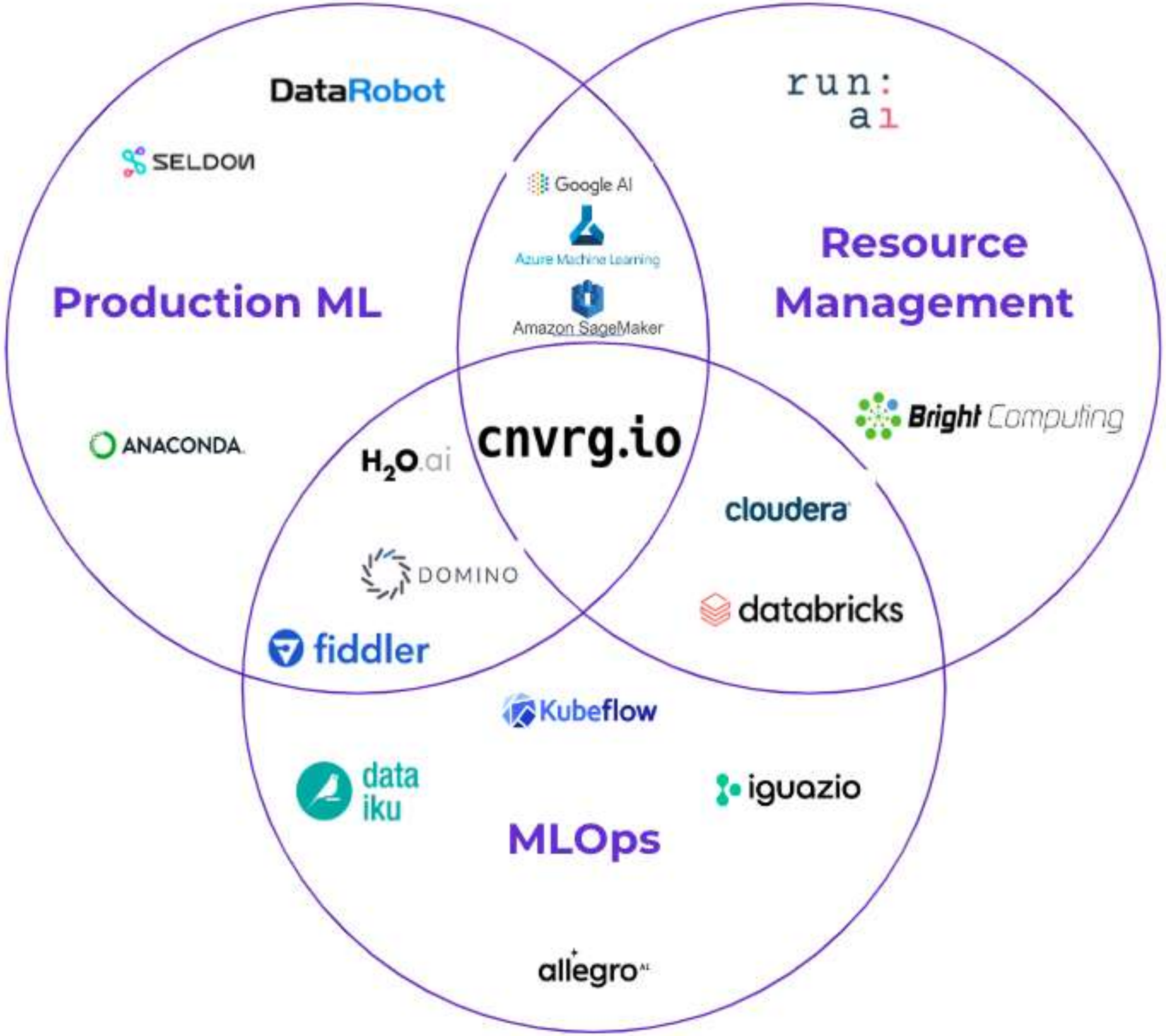
All major CNCF-compliant K8s distributions, including

- OpenShift
- VMware
- EKS, AKS, GKS

cnvrg.io control plane contains all of the MLOps logic – installs as a K8s operator on any supported distribution

K8s taints and tolerations restrict pods to appropriate clusters, e.g., GPU jobs only running on GPU-enabled clusters

# Competitive Landscape



Thank you

cnvrg.io