OctoML

DevOps Agility for Machine Learning Deployment

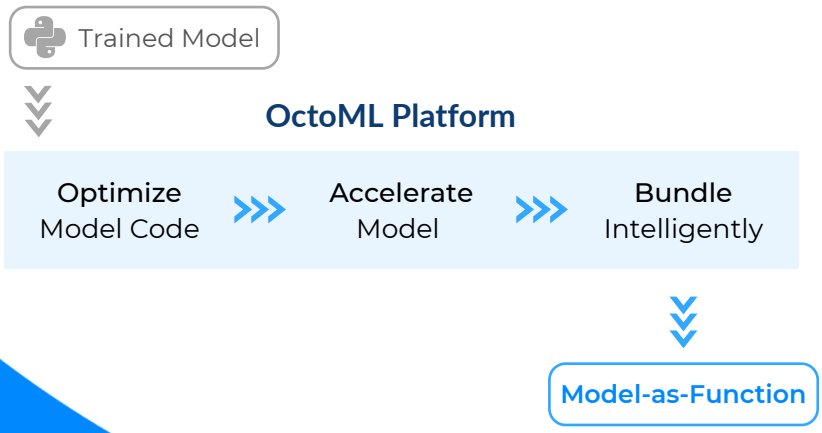
Get More Models to Production

OctoML is on a mission to make AI more accessible and sustainable so it can be used thoughtfully to improve lives. OctoML automates the model deployment process so companies can create smarter applications, faster.

Automate Deployment to Any Hardware

OctoML transforms machine learning models into flexible, hardware-independent, production-ready software functions that can run in the cloud or at the edge. The SaaS platform generates an instantly accessible model-as-function with a stable API that bridges the gap between developers and ops.

It integrates with existing DevOps workflows that don't require special ML expertise so you can bring all your resources to the table, not leave them on the sidelines. Customers shrink the model deployment times from weeks to hours.



- 130 Employees
- Offices in Seattle and San Francisco
- Founded 2019

World-Class Hardware Partnerships

Collaboration with AMD, ARM, Qualcomm, and NVIDIA helps OctoML accelerate deployment of ML applications for a range of devices, including multi-cloud standard CPUs and GPUs, edge GPUs and CPUs, with new devices coming all the time.

ML Automation Pioneers

OctoML was founded by the creators of Apache TVM, an open-source ML stack for performance and portability -- a key part of the architecture of popular consumer devices like Amazon Alexa.

Unique SaaS Architecture

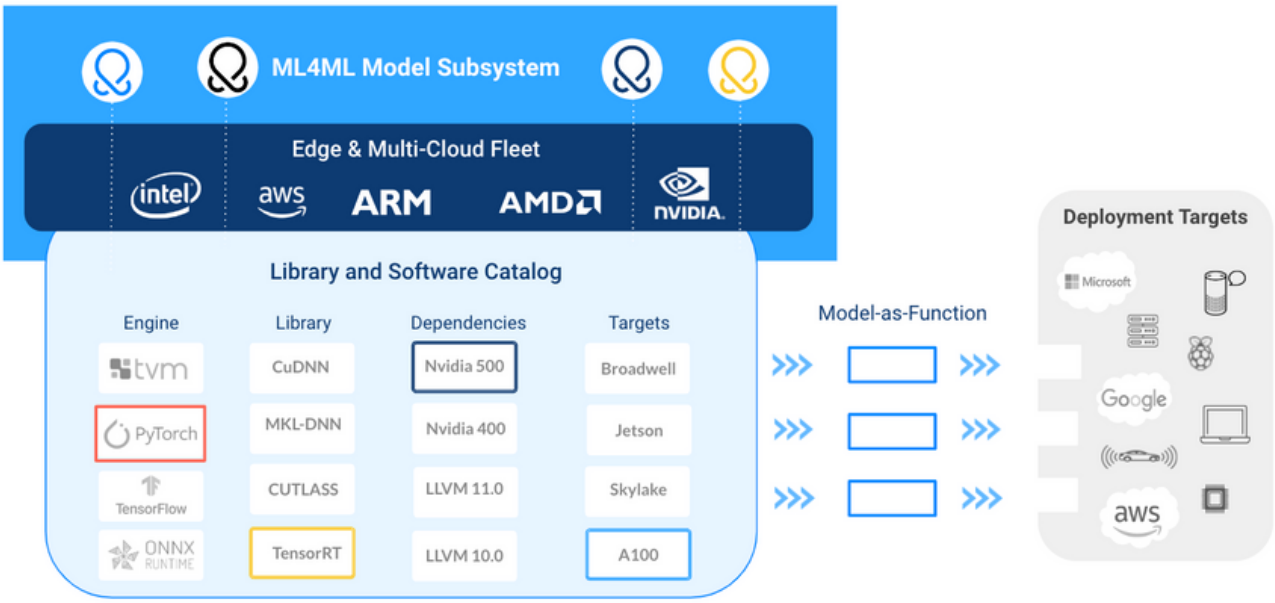
Automation engine powered by 'Machine Learning for Machine Learning.' Trained on a catalogue of software libraries, cloud and device fleets to auto-select the best possible deployment path.

Put Your DevOps Workflows to Work

ML for ML automates the hardest parts of model deployment

The Machine Learning for Machine Learning subsystem *optimizes* model code, *accelerates* the model to meet cost and speed SLAs, and intelligently *bundles* a complete cloud-deployment recipe.

OctoML is continually training against the industry's largest catalog of software libraries, ML engines and ML dependencies in combination with a fleet of edge and cloud targets to hone and deliver better results with each deployment.



Microsoft arm Google wipro woven planet

Azure aws vmware NVIDIA AMD Qualcomm

> OctoML enabled us to deliver a better and faster experience for our customers by delivering 2x performance improvement on our deep learning models.

-Wilson Yu, Head of Advanced Technologies Group, AMD

Partnership Contact: Juan-Antonio Carballo, VP of Business Development
Website: www.octoml.ai