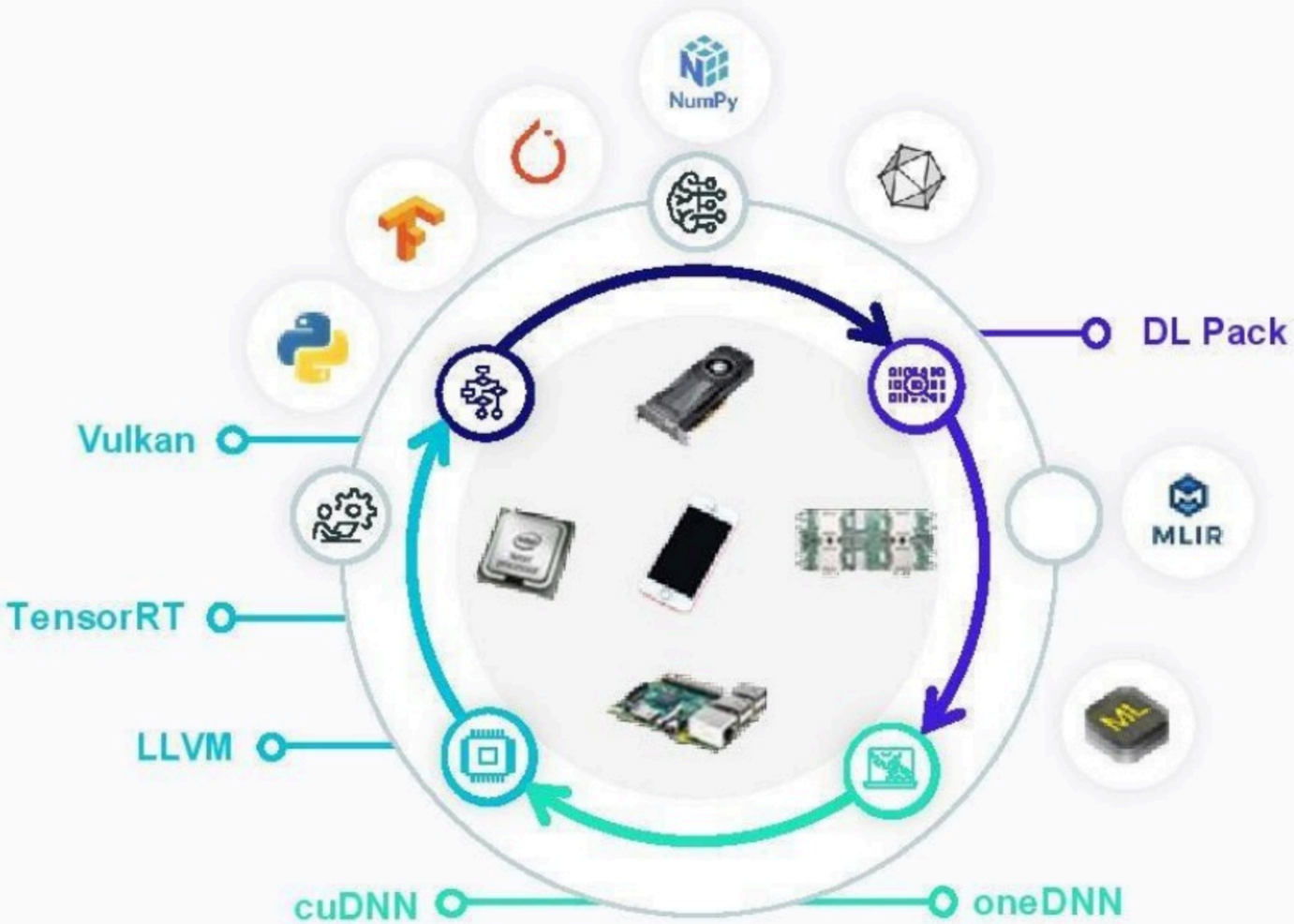
Machine Learning Deployment Platform

OctoML

OctoML founded in mid-2019 in Seattle, WA

- 85+ employees
- $47M via Seed/A/B from Madrona, Amplify and Addition
- Core product: OctoML SaaS Platform

Apache TVM Open Source Ecosystem Growth



Apache TVM OSS project was launched out of University of Washington by OctoML co-founders

645 lifetime contributors and counting. Support for all major HW. Many production deployments. Strong community and ecosystem integration.

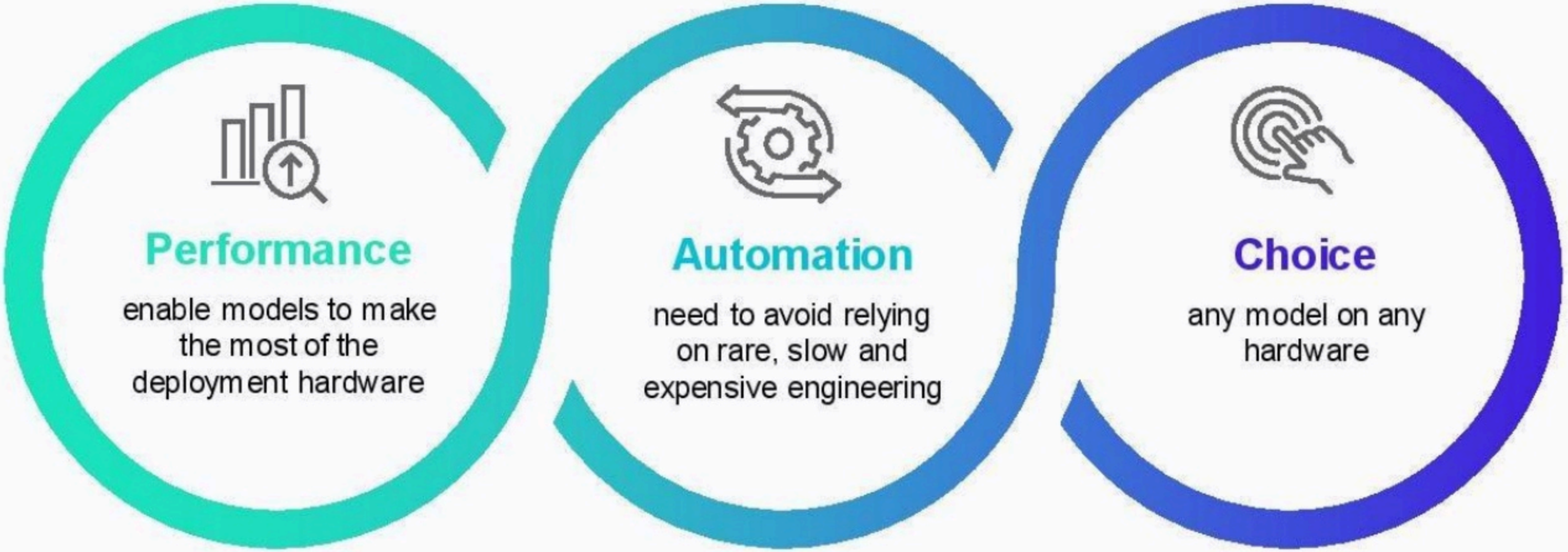
OctoML

Why OctoML Exists

- **Offer sustainable and accessible AI used thoughtfully to improve lives.**
- Catalyze Apache TVM's ecosystem growth and build a platform to enable anyone to easily deploy ML models on any hardware at peak performance.

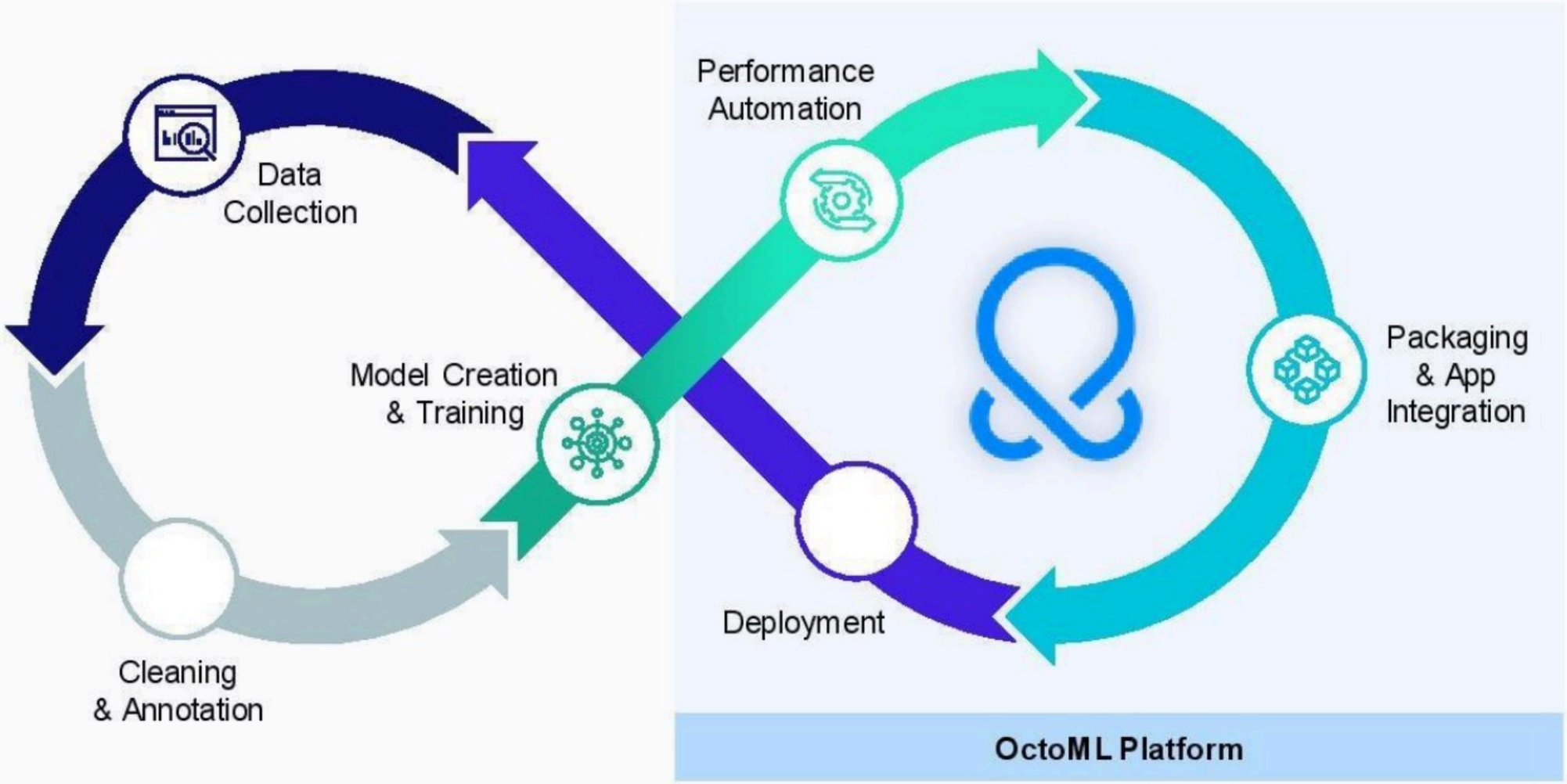
OctoML

What We Need to Get There



OctoML

OctoML Platform in the MLOps Flow



OctoML

Performance is Critical to ML Innovation

AI/ML Impact is Limited by Efficiency

Capacity growing fast to keep up with demand.

Carbon footprint is humongous by comparison.

Recommendation model at scale - **170** american homes for a full year!
Or plant **7.5k** trees to offset.

This, to me, is personal.

Source: Carole-Jean Wu et al., "Sustainable AI: Environmental Implications, Challenges and Opportunities"

OctoML

2x
Faster

1/2
Energy Use

1/2
Infrastructure

1/2
Cost

Imagine your code running at a global scale in data centers

Imagine the batteries needed to power a sea of devices scattered around the globe

Think of the planet and happy trees!

Performance is critical

OctoML

Achieving Performance is Hard...

ML Model = Code + Data

- Very sensitive to deployment HW and infrastructure
- Stresses memory, compute/communication resources
- It can trade-off accuracy for lower resource usage

To optimize for performance, ML engineers need to know

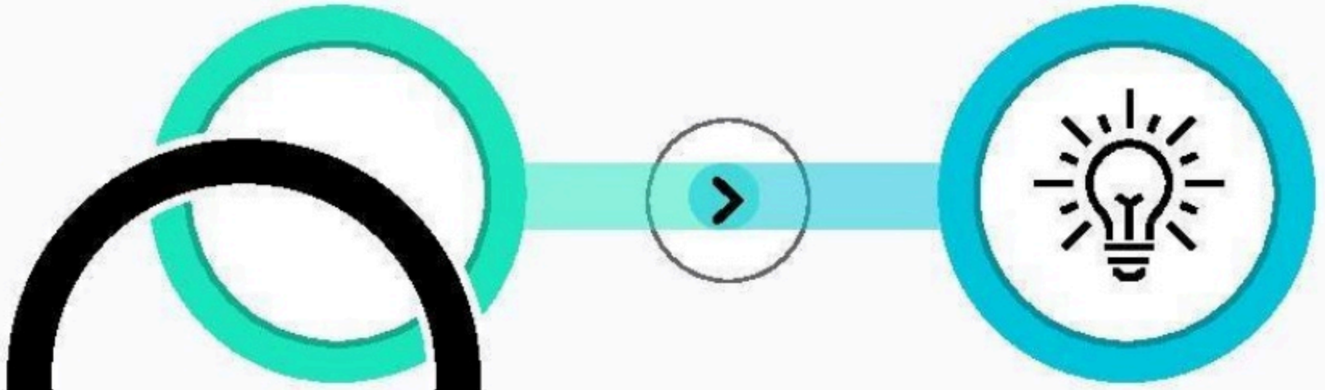
- Machine learning model details
- Optimization techniques (locality optimizations, code scheduling, quantization, etc)
- Compilers and even computer architecture

OctoML

…Yet More People Need to Be Able to Do It

Not enough people know how to optimize and deploy models

- It can't stay that way or it will be in the hands of too few players
- Need to make performance in ML accessible to a wide rang[illegible] practitioners



Automation is Key to Scaling ML deployment

Hand Engineering for ML Deployment Doesn't Scale

Getting a model to production can take months! And given the growth trends in AI/ML, there are not enough engineers to keep up with demand.

What needs to be automated?

- **Tuning model implementation on all deployment hardware options**
- **Choosing what hardware is best for deployment at scale**
  Best model throughput/watt? Or best throughput/$? Or latency?
- **Packaging model for easy deployment**

OctoML

Performance Automation with Apache TVM

ML model

TVM: ML-based optimizations to obviate need for hand-tuning for target HW

Optimized code specific to target HW

Thrives on diversity: models, frameworks, types of optimizations, HW targets.

OctoML

Case Study: Apple M1