Welcome!

We’re happy to have you here.

Meet the team!

Kamil

Emmy

Xiaowei

Artur

Balaji

Yunxuan

The Plan

Here’s what to expect today.

Today's agenda.

Time	Description
10 min	Welcome + Accessing Your Cluster
15 min	Introduction to Ray and Anyscale
20 min	Anyscale Clusters + Workspaces
45 min	Hugging Face Fine-Tuning, Hyperparameter Tuning, and Batch Inference w/ Ray AIR
10 min	Anyscale Jobs
5 min	Summary + Survey

Course Logistics

How to access your cluster and other resources.

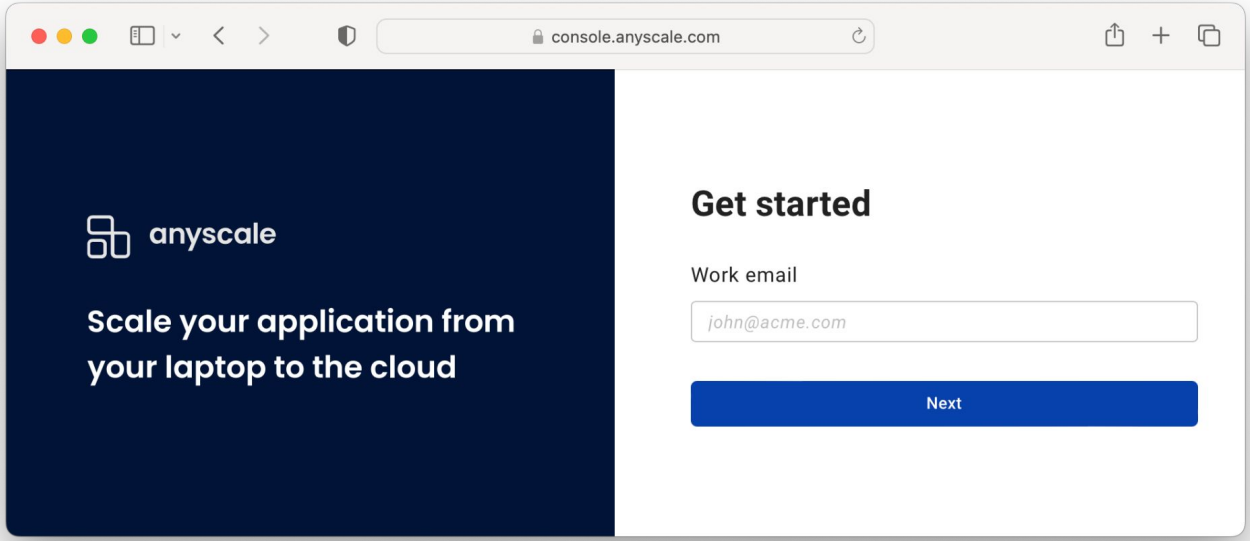
✅ Tech check.

 Accessing Anyscale clusters.

- All work will be in Anyscale provisioned clusters.
- Our GitHub repo will be mounted automatically.
- Access begins now.
  - Check your email for login information.
  - Step-by-step instructions to follow.

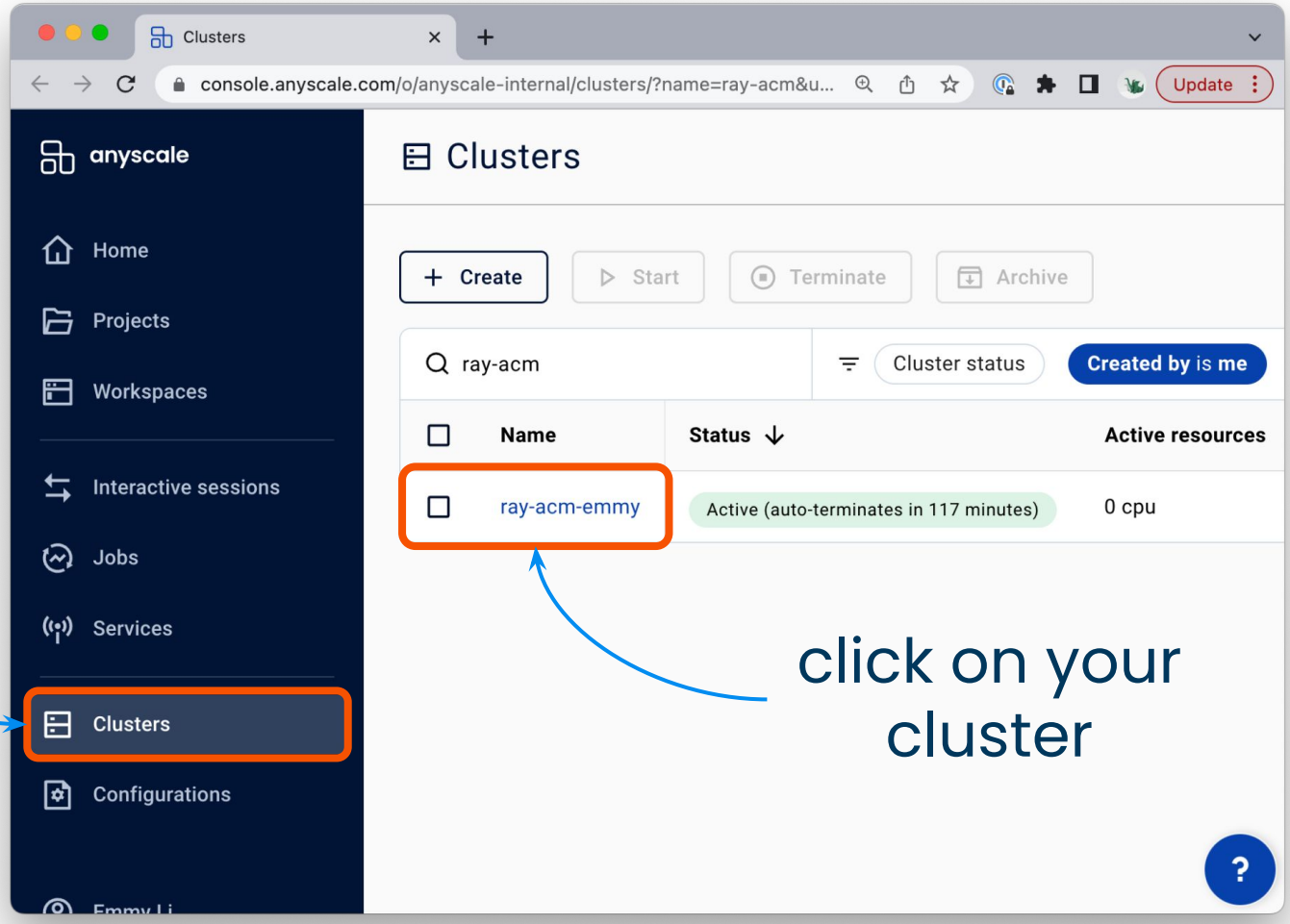
Anyscale Login

Link to Anyscale cluster: console.anyscale.com



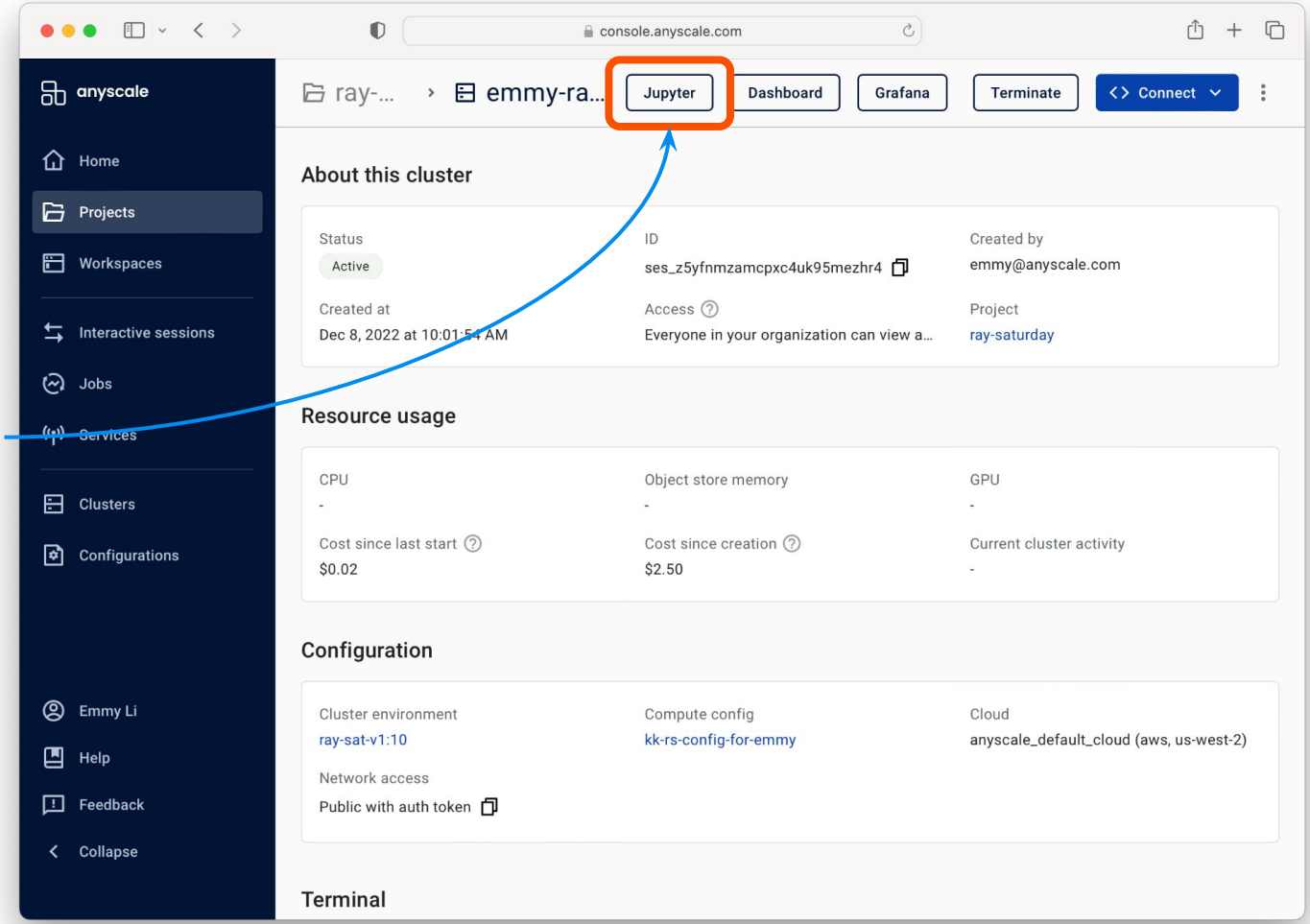
Check your **email** for your username and password!

select “Clusters”

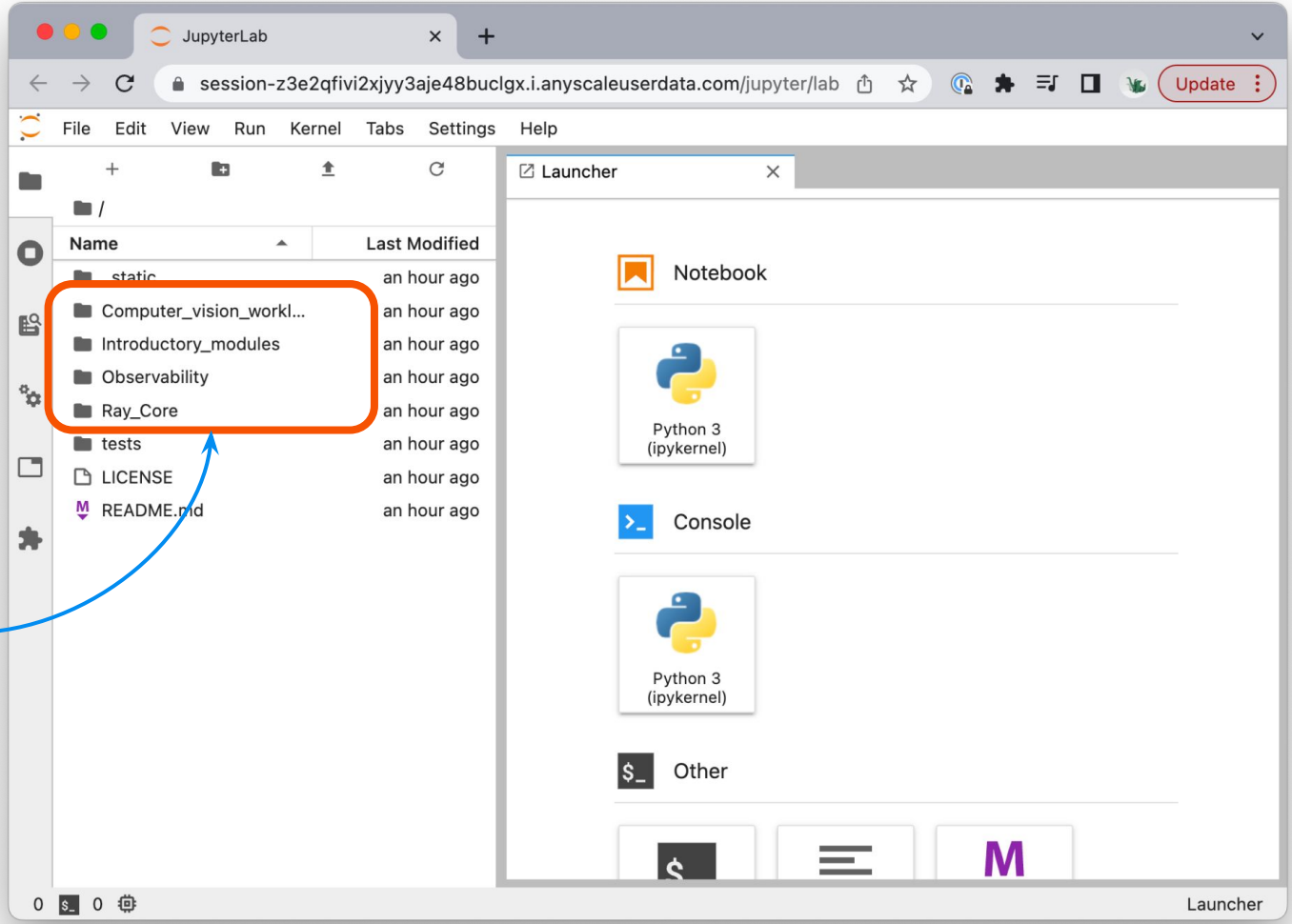


click on your cluster

"Start" the cluster, then select "Jupyter"

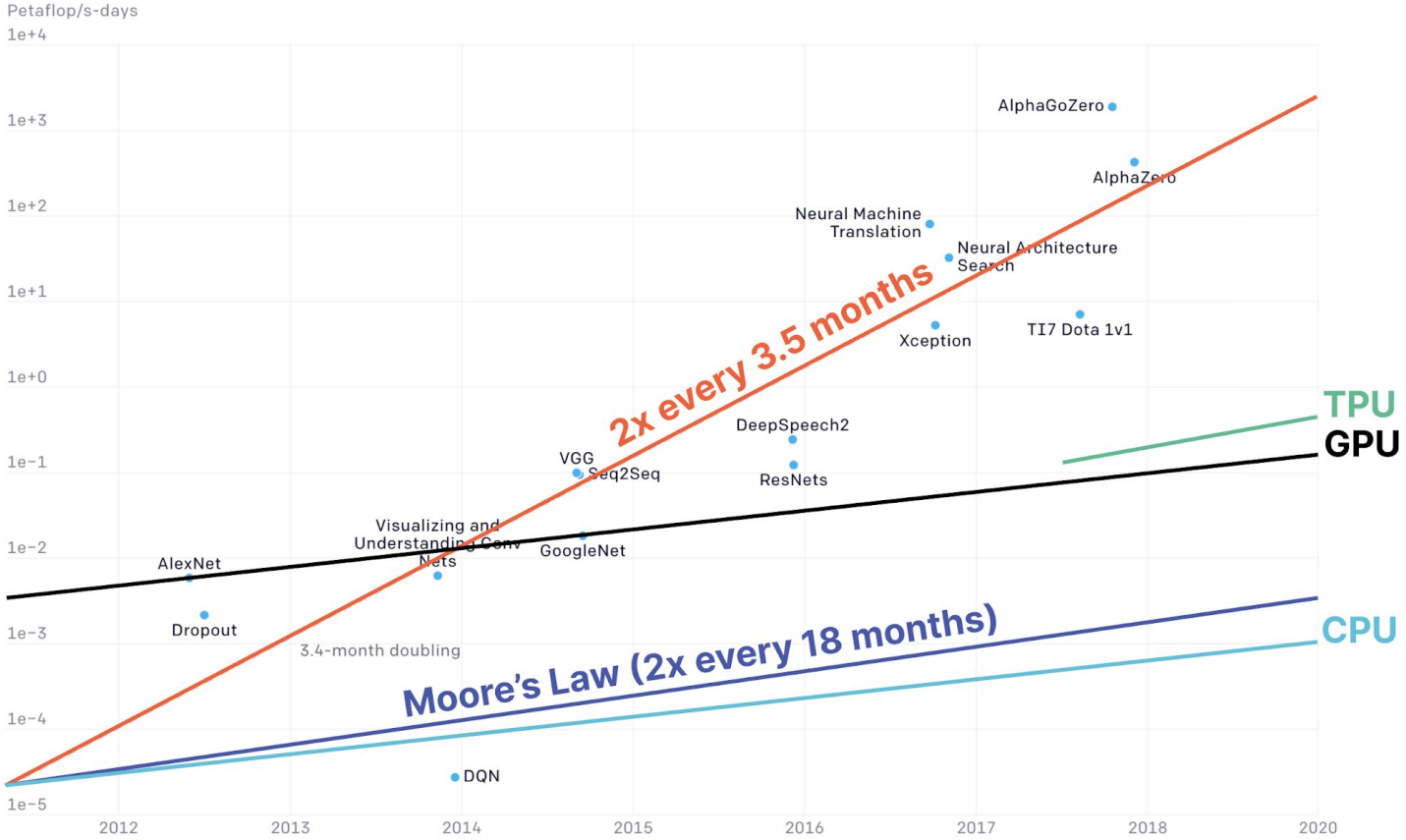


View modules here



The State of Ray

Distributed computing: a bit of context



RAY


Artificial
Intelligence
O'REILLY
intel Nervana

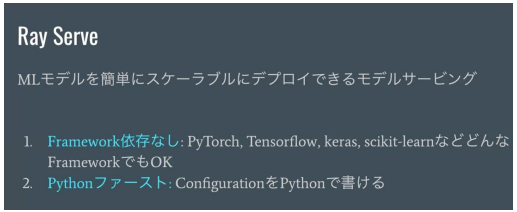

EXIT


Ray 2.0 is production ready

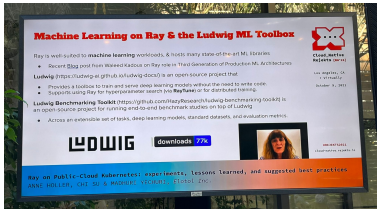
800+ Contributors

Ray: A global community

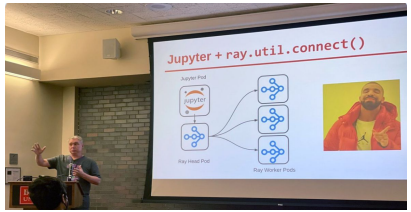
Colombia



Japan



Los Angeles



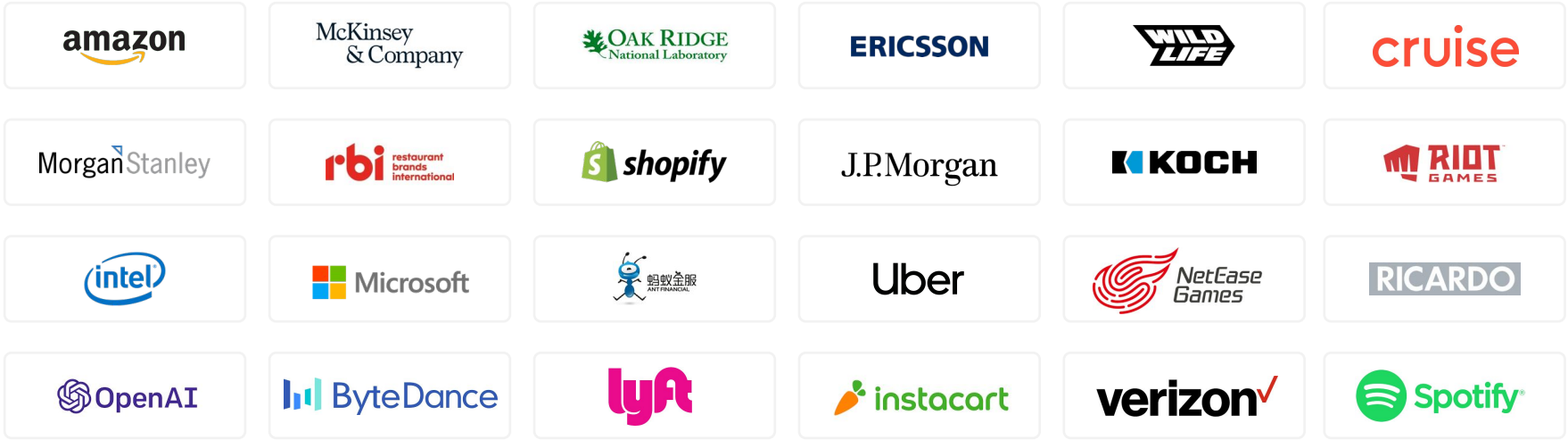
Boston

China

Toronto

... and many more!

Ray: Fastest Growing Scalable Compute Framework



25,000+
GitHub
stars

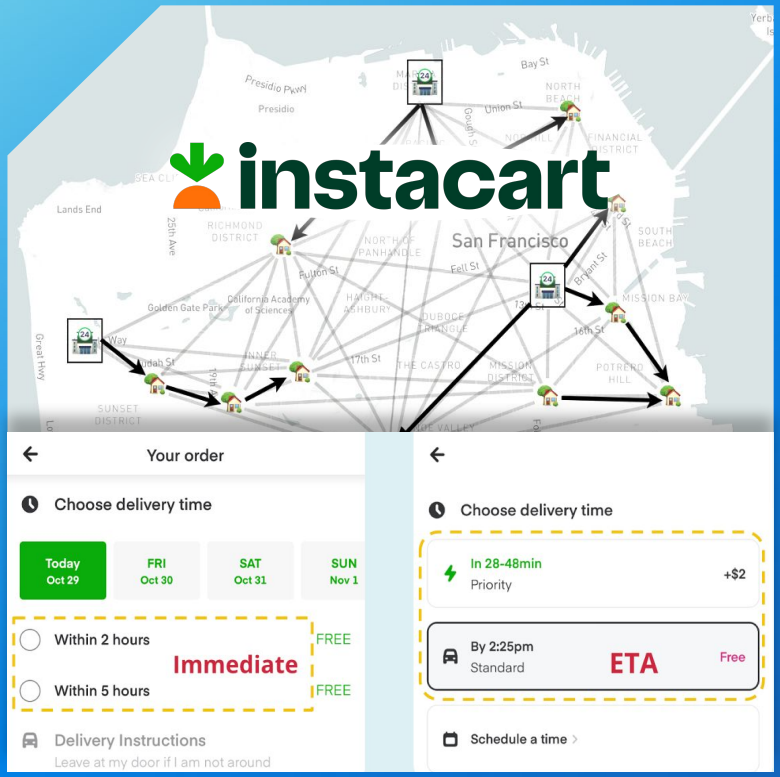
800+
Community
Contributors

5,000+
Repositories
Depend on Ray

1,000+
Organizations
Using Ray

ETA prediction, safety, maps, eats, marketplace

Deep learning, classical ML, hyperparameter tuning, data ingest



Order fulfillment, scheduling, delivery ETAs

Reduce training times for 12,000 models from days to hours

Designing sailboats in simulation, building them in the real world

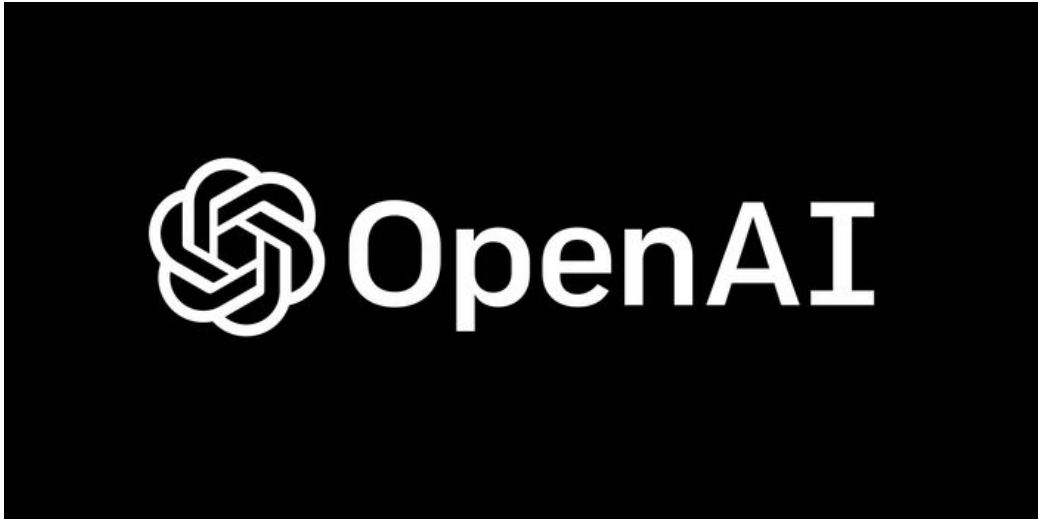
They won the America’s Cup

McKinsey & Company

OpenAI on Ray

“We use Ray to train our largest models including ChatGPT.”

Greg Brockman, Co-founder and President, OpenAI



LLM Pain Points

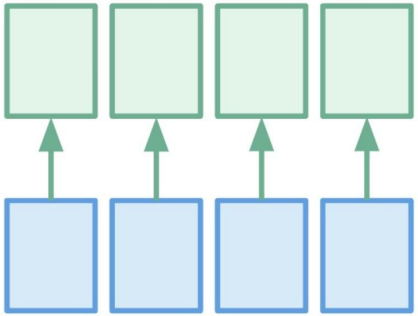
- Large-scale: need to run on many **expensive** GPUs 💰💰💰
- Existing training & inference solutions **don't work** (can't naively scale) 🤷‍♂️🤷‍♂️🤷‍♂️
- Technology is emerging - important to be able to **iterate quickly at scale** 🏃‍♂️💨

Ray: fitting solution

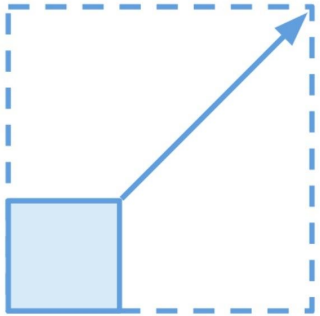
- **Spot instance** support for training.
- **Flexible scheduling** and resource allocation enables new scaling patterns.
- Very **fast development** model - re-run your Python script across a cluster.

Key Ray characteristics

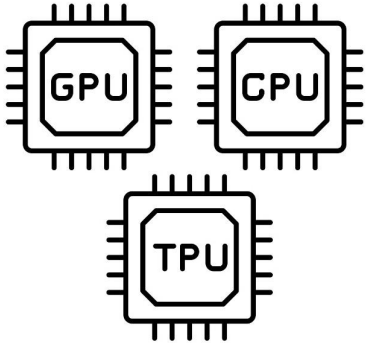
Python first approach



Simple and flexible API

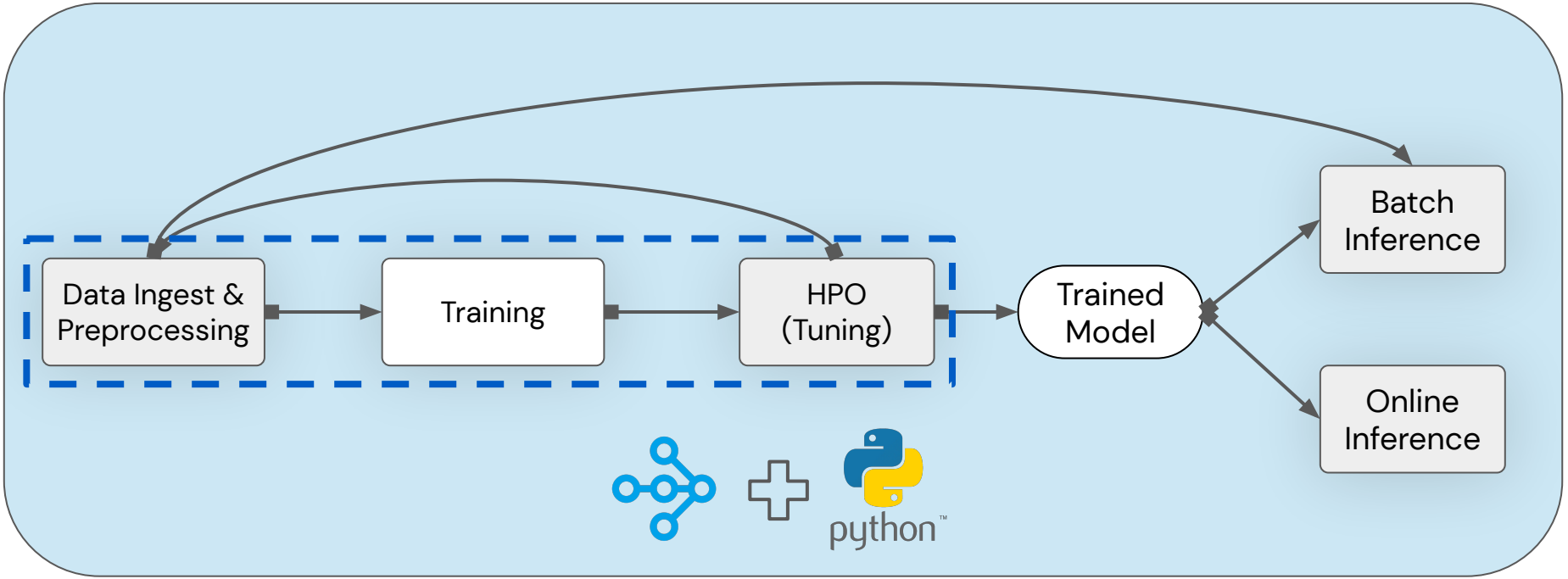


Scalability



Support for heterogeneous hardware

ML Application Lifecycle



Ray enables building end-to-end ML applications, all in Python

Ray AIR and integrations

RAY

About Anyscale

Founded by the creators of Ray

- Ion Stoica, Chairman & President (Co-Founder Databricks, Lead UC Berkeley RISELab and AMPLab, Founder Conviva)
- Robert Nishihara, CEO (Ray co-creator, Ph.D. Berkeley)
- Philipp Moritz, CTO (Ray co-creator, Ph.D. Berkeley)
- Michael Jordan (UC Berkeley RISELab, AMPLab)

Team

- Experts in Distributed Computing and AI
- Ex-Uber, Stripe, AWS, Databricks, Google Brain, Facebook

Investors & Board

- Investors: Andreessen Horowitz, NEA, Addition, Foundation, Intel Capital
- Board: Ben Horowitz, Pete Sonsini
- Raised $250M+

Three Trends

1. AI computational demands are exploding
2. The end of Moore's Law
3. AI is permeating every industry

Ray | Unified framework for scalable ML

Ray

Unified	Scalable	Open
• Common infrastructure for all distributed workloads • Libraries for training, serving, data loading • Extensible to future workloads	• Laptop to cloud with zero-code changes • Elastic scaling in production • Scale to support the most demanding workloads	• Open source • Integrates with the entire ML and Python ecosystem • Run anywhere, including on any cloud

Anyscale | The best way to run Ray

Anyscale

SCALABLE COMPUTE PLATFORM

Unified	Scalable	Open
• Common infrastructure for all distributed workloads • Libraries for training, serving, data loading • Extensible to future workloads	• Laptop to cloud with zero-code changes • Elastic scaling in production • Scale to support the most demanding workloads	• Run anywhere, including on any cloud • Integrates with the entire ML and Python ecosystem • Open source

Accelerate time to market with Anyscale

- Managed service
- Enterprise ready
- Dev-to-prod experience
- Built-in observability & operational tools



Conclusion

More applications need to **scale** (driven by AI).

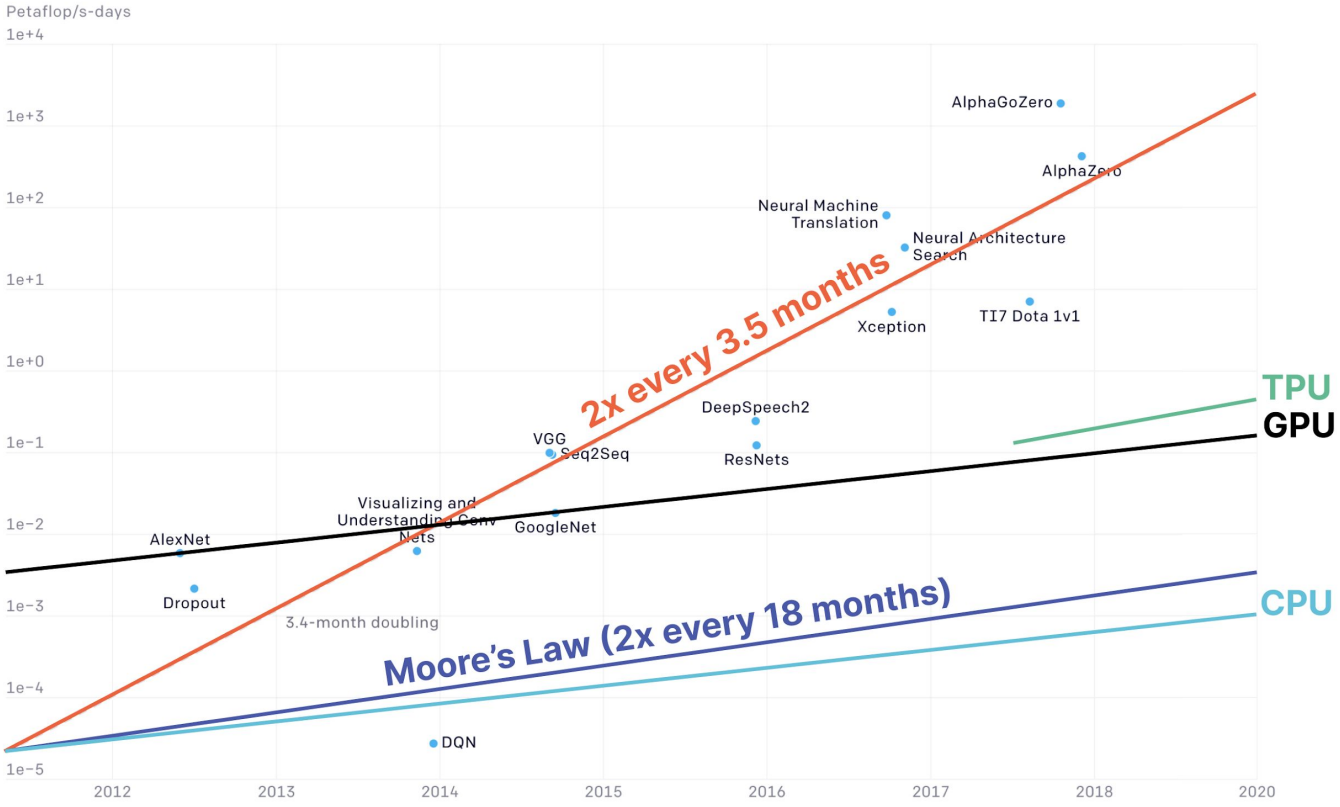
Ray: unified framework for scalable computing.

Anyscale: the best platform for Ray.

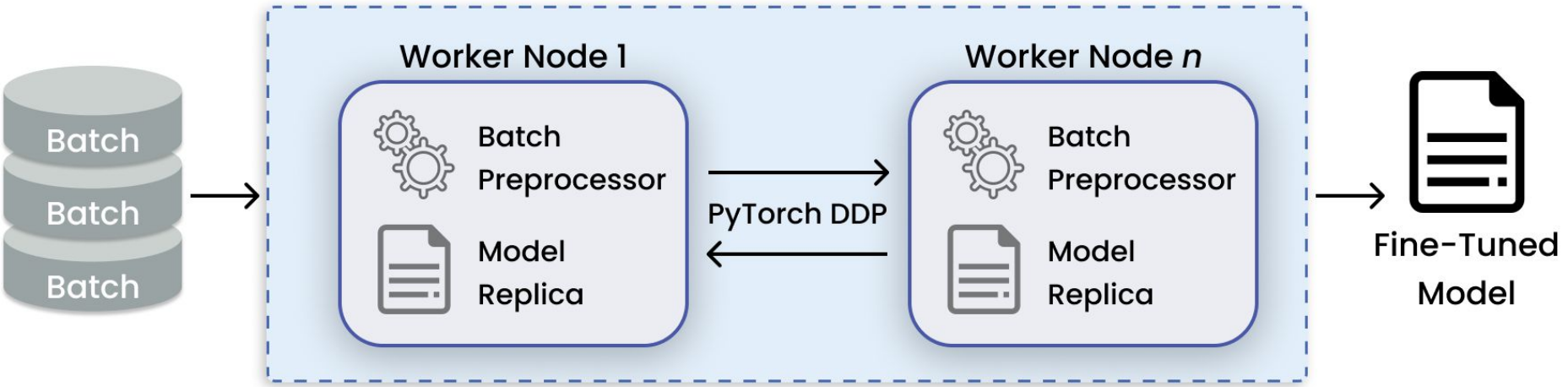
Distributed Model Fine-Tuning

Data parallelism across multiple machines.

Why distribute training?



Data parallelism



The task.

Instructions and demonstrations

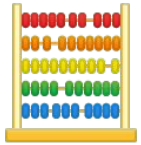
- `instruction` - A prompt and/or question.
- `input` - Additional context information.
- `output` - The generated response.

The data.

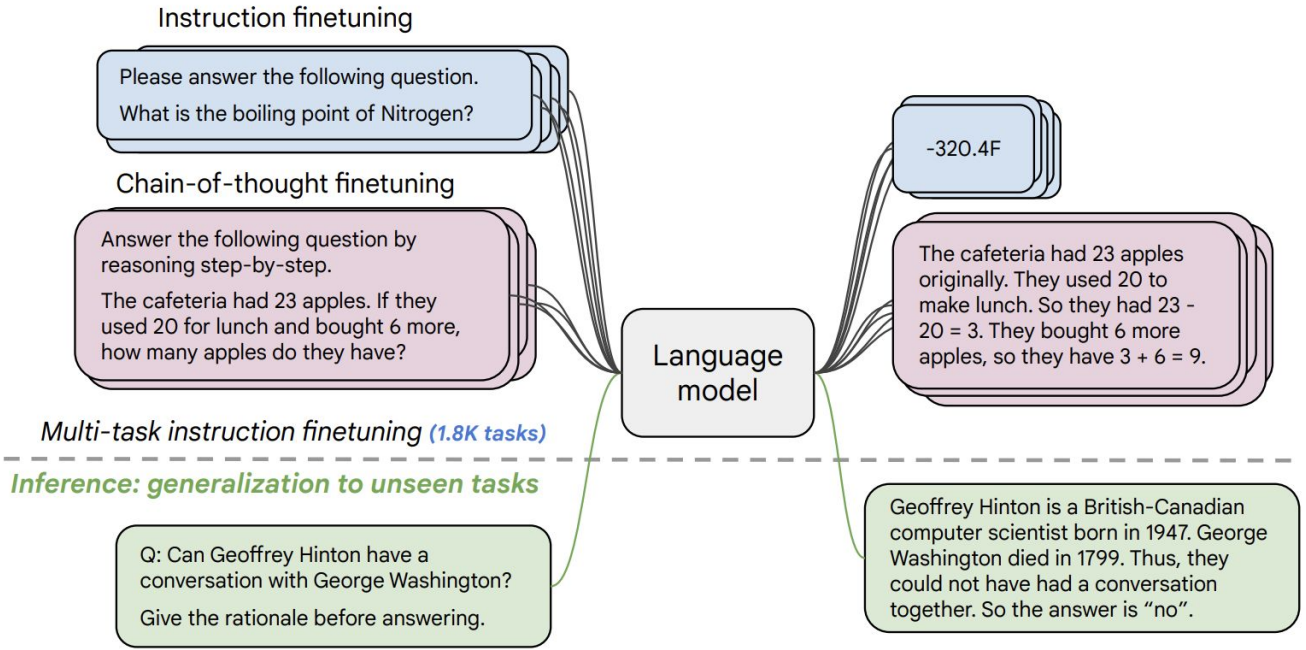
Alpaca

- 52k instructions and demonstrations.
- Generated by Open AI's `text-davinci-003`

instruction (string)	input (string)	output (string)
"Identify the odd one out."	"Twitter, Instagram, Telegram"	"Telegram"



The model.



FLAN-T5

 The goals.

 Fine-Tuning

Train FLAN-T5 on batches of Alpaca data.

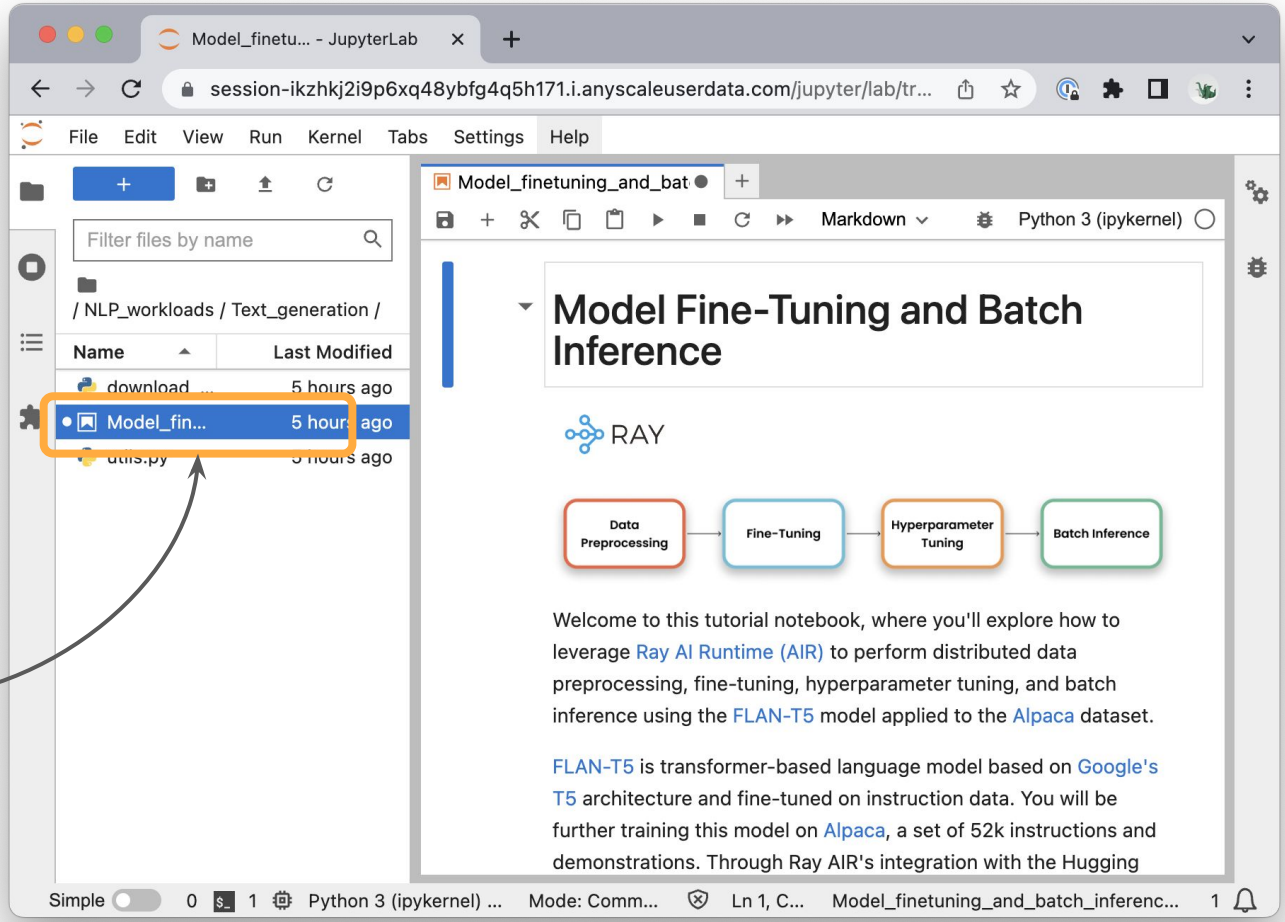
 Hyperparameter Tuning

Run multiple trial experiments with different hyperparameters.

 Batch Inference

Generate predictions on batches of data.

Meet me here!



NLP_workloads/Text_generation/Model_finetuning_and_batch_inference.ipynb

📄 Fill out the survey.

🔗 Go to bit.ly/anyscale-ray-feedback

🎟️ Ray Summit 2023 tickets

A random draw from survey submissions.