Generative AI + Ray

Fine-tuning and Deploying Stable Diffusion

Emmy Li, Kourosh Hakhamaneshi, Justin Yu

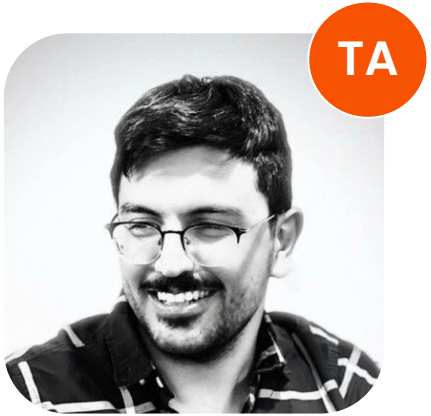
RAY SUMMIT 2023

👋 Welcome!

We're happy to have you here.

Meet the team!

Emmy



Kourosh

Justin

📋 The Plan

Here’s what to expect today.

Today's agenda.

1:00pm (20 min)	**Talk:** Ray for Production-Grade GenAI
1:20pm (70 min)	**Coding Lab:** Fine-tuning Stable Diffusion with Ray Data and Train
2:30pm (15 min)	Coffee Break
2:45pm (60 min)	**Coding Lab:** Serving Stable Diffusion with Ray Serve
3:45pm (15 min)	**Talk:** Resources for Further Exploration

✅ Tech check.

Participating via app.sli.do

- Join with code **#ray-genai**
- Ask questions.
  - Pose your own and upvote others.
  - TAs will be answering questions on a rolling basis.

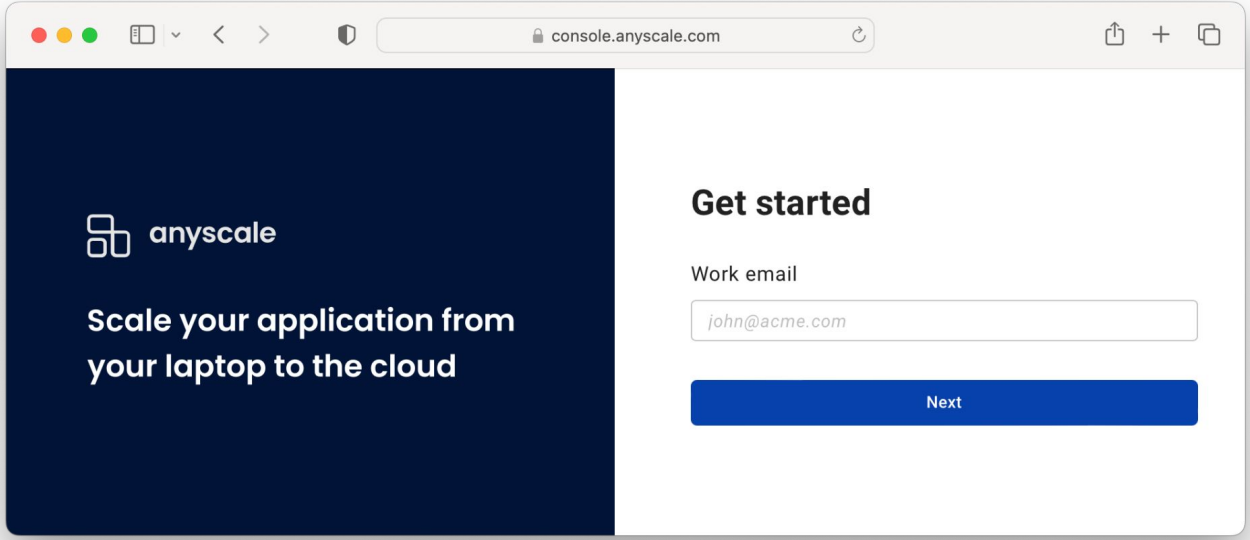
✅ Tech check.

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



Enter the **unique credentials** sent to your email!

1. Select Workspaces

console.anyscale.com

anyscale

Home

Workspaces

Jobs

Services

Schedules

Configurations

Clusters

Projects

Home

Examples to get started

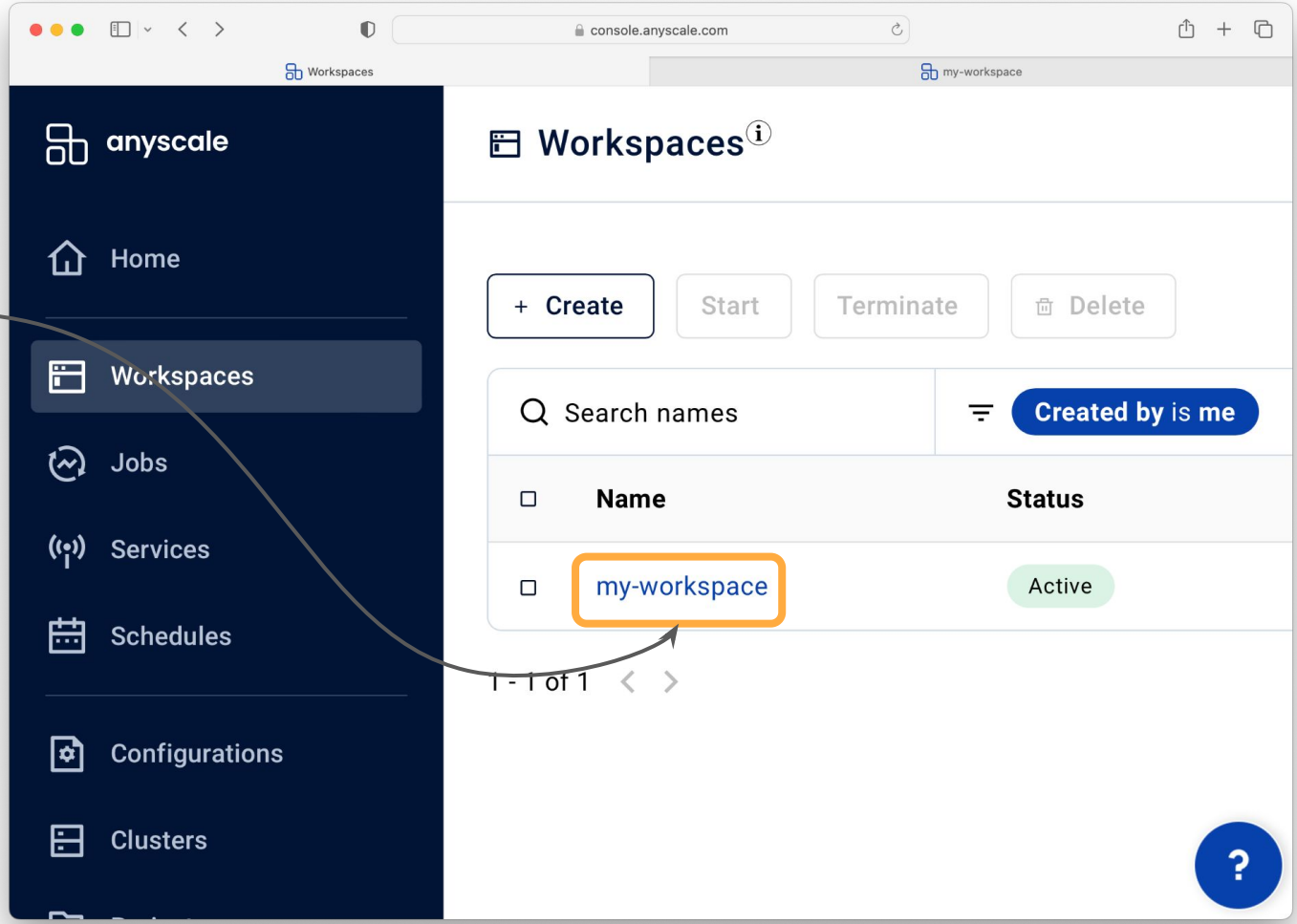
Introduction to Anyscale & Ray Launch

Learn about Anyscale and Ray in this introductory tutorial. This template runs a simple Ray program on a distributed Ray cluster then deploys an Anyscale Job based on the Ray program.

Ray task | Anyscale Job

Many Model Training Launch

2. Select Your Workspace



3. Click on Jupyter icon

console.anyscale.com

anyscale

Home

Workspaces

Jobs

Services

Schedules

Configurations

Clusters

Projects

Emmy

Help

Feedback

m...workspace Active (Ray)

Terminate

Tools

About | Files | Terminal | Logs | Serve deployments

Status
Active (Ray)

Resources

Access
Everyone in your org

Created
Sep 7, 2023 at 4:26:50 PM, by emmy+education@anyscale.com

Cluster environment
summit:9

Compute config
ray-summit-2023-gc

Network access
Public with auth token

Job submissions
None

Ports
36075

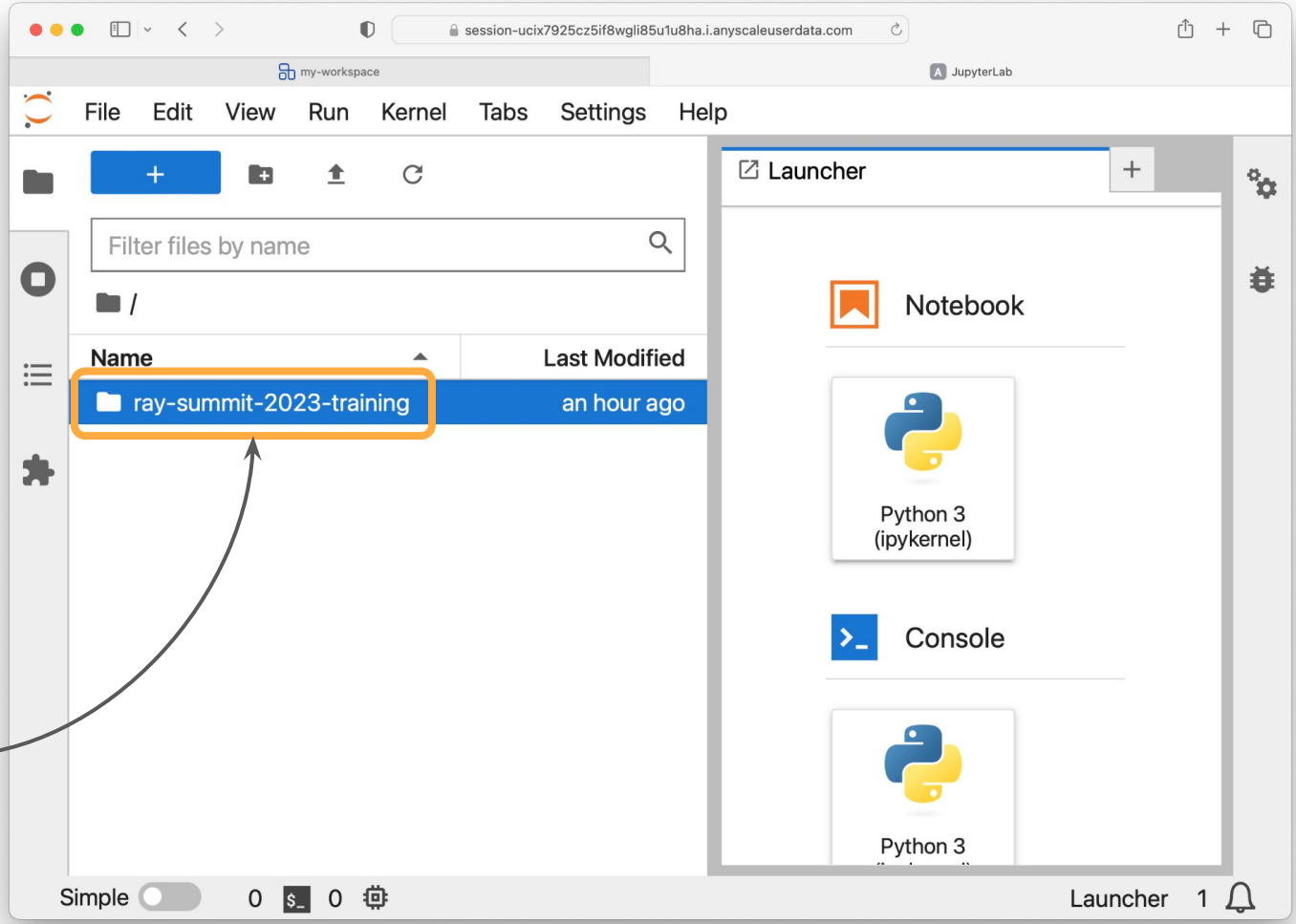
README

Workspaces

A Workspace is a fully managed development environment focused on developer productivity. W enable ML practitioners and ML platform developers to quickly build distributed Ray applications from research to development to production easily, all within a single environment.

Workspaces provide a remote experience for programming your cluster while working with JupyterLab notebooks or Visual Studio Code.

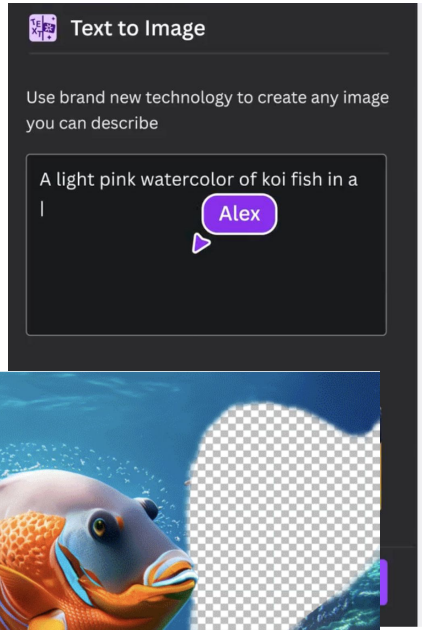
4. Find the content for your class here.

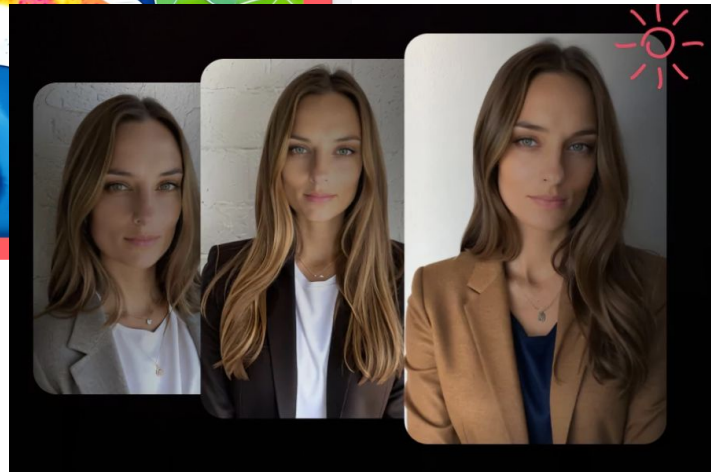


From local to cloud

An introduction to Ray and Anyscale.

Potential use cases.





🏭 Let’s move to production!

✅ **Tested thoroughly** on your local machine.

✅ Refactored from notebooks to a **reusable, encapsulated** format.

✅ Hit the **quality** and **latency** benchmarks we’re okay with.

What could go wrong?

😠 Everything that went wrong.

Infrastructure

❌ Deployment strategy

Which cloud, how much storage, how much compute

❌ Load balancing

Making sure no surge in traffic breaks the entire system.

❌ Fault tolerance

Dealing with disaster and building in redundancy.

😡 Everything that went wrong.

Maintenance

❌ Monitoring and logging

Inspecting performance, error tracking, metrics.

❌ Continual learning

Swapping in new data, model, and prompt versions.

❌ Dependency management

Ensuring consistent execution of complicated LLM systems.

Everything that went wrong.

Cost

❌ Scaling

Orchestrating large-scale deployments that adjust to traffic.

❌ Resource management

Precise resource allocation, using spot instances, batching

❌ Proprietary vs. OSS models

Pay through the teeth or go the self-hosted route.

😭 Everything that went wrong.

Trap Doors

❌ Security and privacy

Working with sensitive data, breaches, unauthorized access.

❌ Ethics and bias mitigation

Monitoring a non-deterministic app for problematic content.

❌ Inflexibility

Painting yourself into a corner with choices you made.

🌟 The wishlist.

✨ **Easy scaling and reliability**

"I got into this for ML, not for infrastructure management."

✨ **Efficiency and performance**

Built-in optimizations and ability to control when needed.

✨ **Extensibility**

Flexible integrations with other frameworks, clouds, and tools.

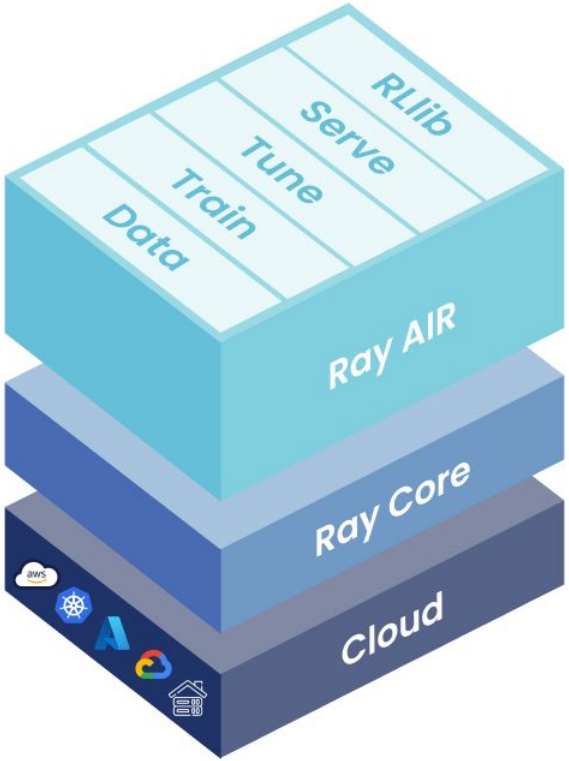
✨ **Observability tooling**

Inspect the infrastructure and ML application layers.

✨ **Intuitive cost control**

Clarity into $$$-eating resources and inefficiencies.

The Ray AI Libraries



enables simple scaling of AI workloads

enables scalable apps to be built in pure Python

The GenAI Primer

A briefing on what we're doing with Stable Diffusion.

🖼️ Dreambooth

Background

- Few-shot fine-tuning for Stable Diffusion
- Allows for personalized models

Your goal

- Distributed fine-tuning
- Serving generative models at scale

Let’s make our way over to the notebooks!

☕ Time for a Break!

15 minutes.

More Resources

For further exploration with Ray, Anyscale, and GenAI.

🍎 Today we learned...

🚂 Overview of Ray AI Libraries

Getting acquainted with each library for distributed ML.

🔬 Hugging Face ➡️ Ray

Converting a vision transformer to run distributed.

🔍 Exploring Ray Train, Data, Serve

Constructing an end-to-end ML pipeline with Ray.

🍏 Sneak Peek: Self-Paced Ray & Anyscale Education

📧 Online at training.anyscale.com

🔍 Preview special technical content releases from the whole team!

📑 Fill out the survey.

🔗 Go to bit.ly/ray-summit-feedback

Reading list.

Self-Paced Ray & Anyscale Education

Access bonus notebooks and scripts about Ray.

Ray documentation

API references and user guides.

Anyscale Blogs

Real world use cases and announcements.

YouTube Tutorials

Video walkthroughs about learning LLMs with Ray.

Upcoming events

Bay Area AI + Ray Summit Happy Hour

Today at 5:00p.m.

Cap off an exciting conference with lightning talks, new friends, and good times!

bit.ly/bayai_ray_meetup

👩‍💻 Connect with the community.

👋 Join the community

Attend events, *subscribe to newsletter*, *follow on Twitter*.

🙋‍♀️ Get support

Join Ray Slack, *ask questions on forum*, *open an issue*.

🔭 Contribute to Ray

Read contributor guide, *create a pull request*.

Thank you!

We hope to meet again.