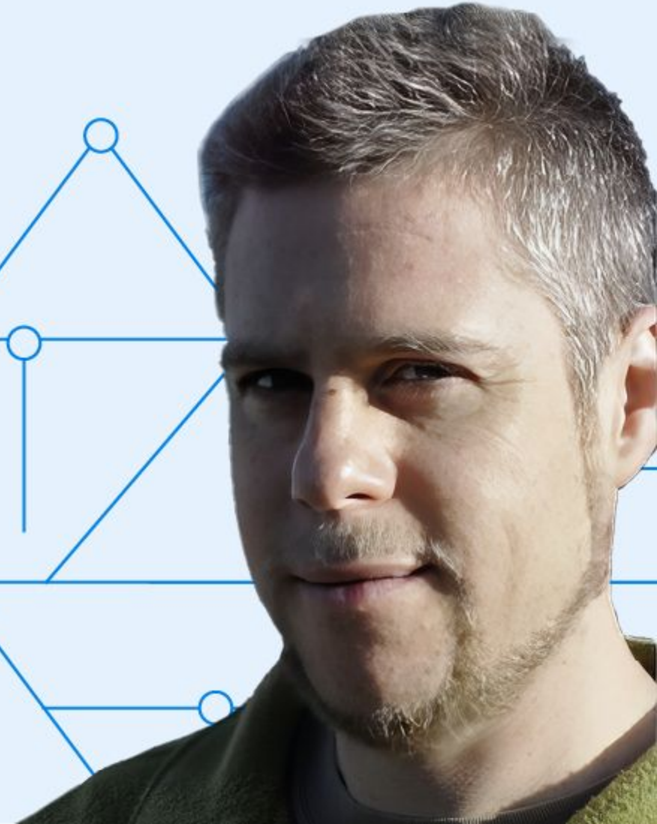
RAY SUMMIT 23

REAL-TIME ML-DEPLOYING ONLINE INFERENCE PIPELINES



presented by anyscale

🧑‍💻 Meet the TAs!

Ed

Akshay

Alexey

📋 The Plan

Here’s what to expect today.

📅 Today’s agenda.

- What is Ray Serve?
- Why use Ray and Ray Serve for scalable AI?
- Build complex ML applications with Ray + Serve
- Under the hood: features for powering production apps
- Architecture options, hands-on labs, and Q&A

✅ Tech check.

Participating via app.sli.do

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

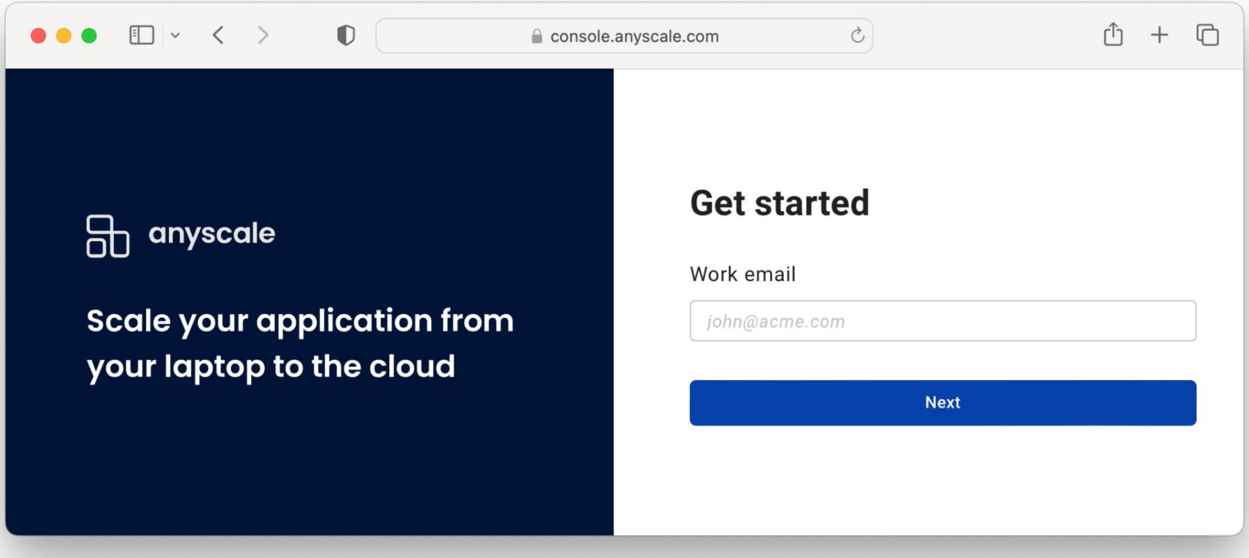
✅ Tech check.

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



console.anyscale.com

anyscale

Scale your application from your laptop to the cloud

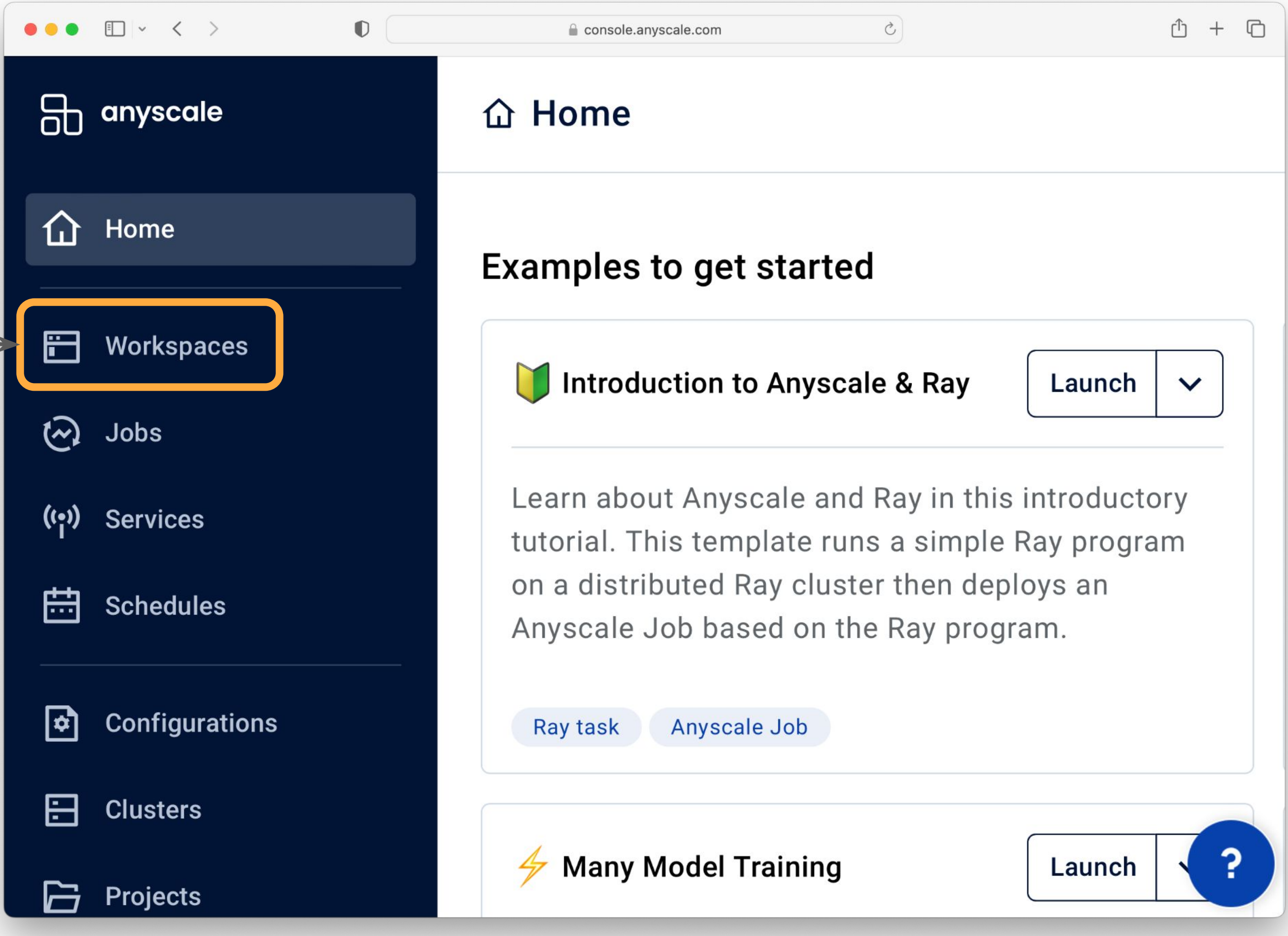
Get started

Work email

john@acme.com

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

1. Select Workspaces



Home

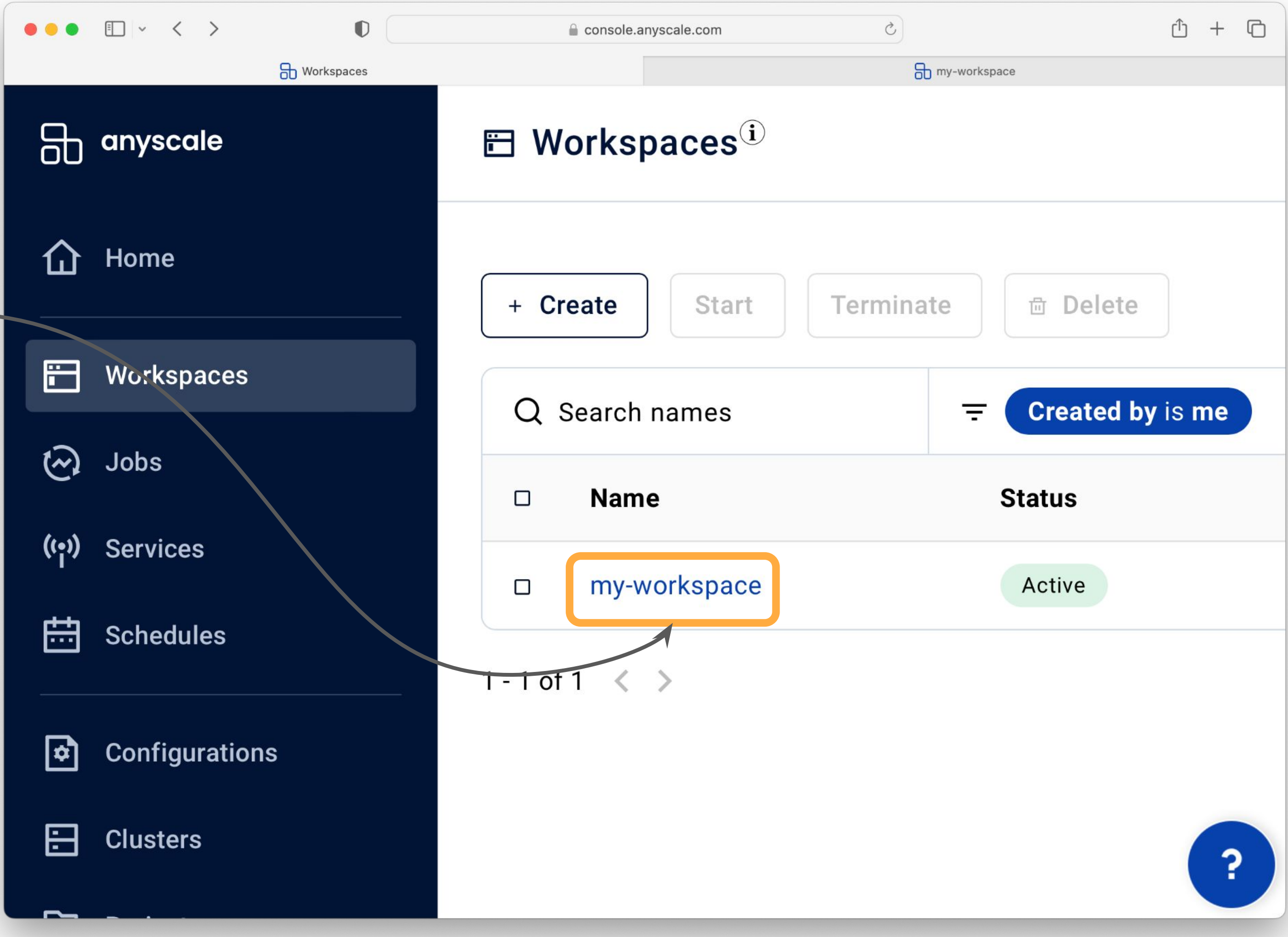
Examples to get started

Introduction to Anyscale & Ray

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.

Many Model Training

2. Select Your Workspace



console.anyscale.com

Workspaces

my-workspace

anyscale

- Home
- Workspaces
- Jobs
- Services
- Schedules
- Configurations
- Clusters

Workspaces

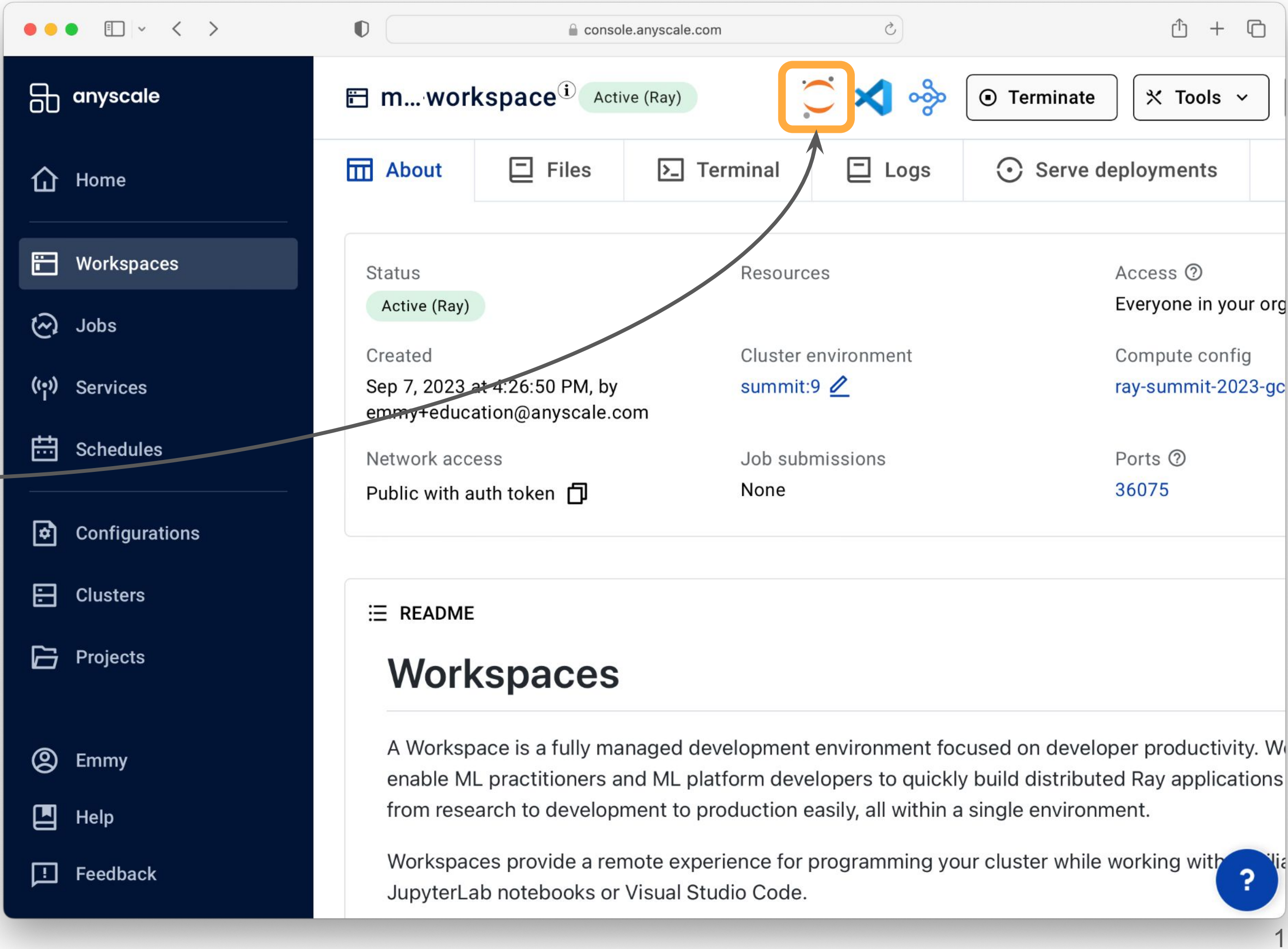
+ Create | Start | Terminate | Delete

Search names | Created by is me

□	Name	Status
□	my-workspace	Active

1 - 1 of 1

3. Click on Jupyter icon



Status
Active (Ray)

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

Network access
Public with auth token

Resources

Cluster environment
summit:9

Job submissions
None

Access
Everyone in your org

Compute config
ray-summit-2023-gc

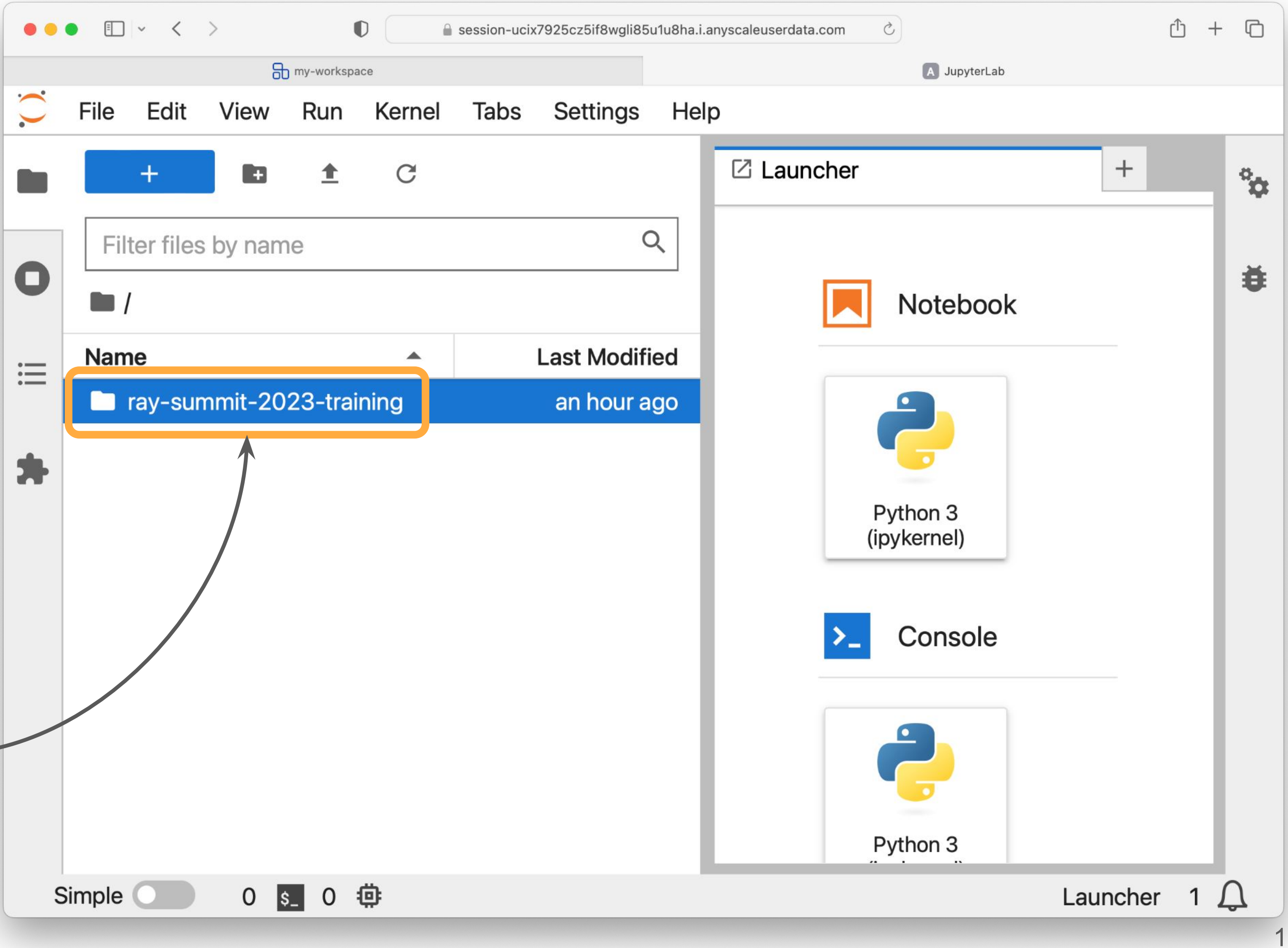
Ports
36075

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.

☕ Time for a Break!

15 minutes.

More Resources

For further exploration with Ray, Anyscale, and LLMs.

🍎 Today we learned…

- 🚂 What Ray Serve is and how it works
- 🔬 How to use Serve for production services
- 🔍 Why to choose Serve for AI-based apps

🍏 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.

Ray Education GitHub

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.