



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.

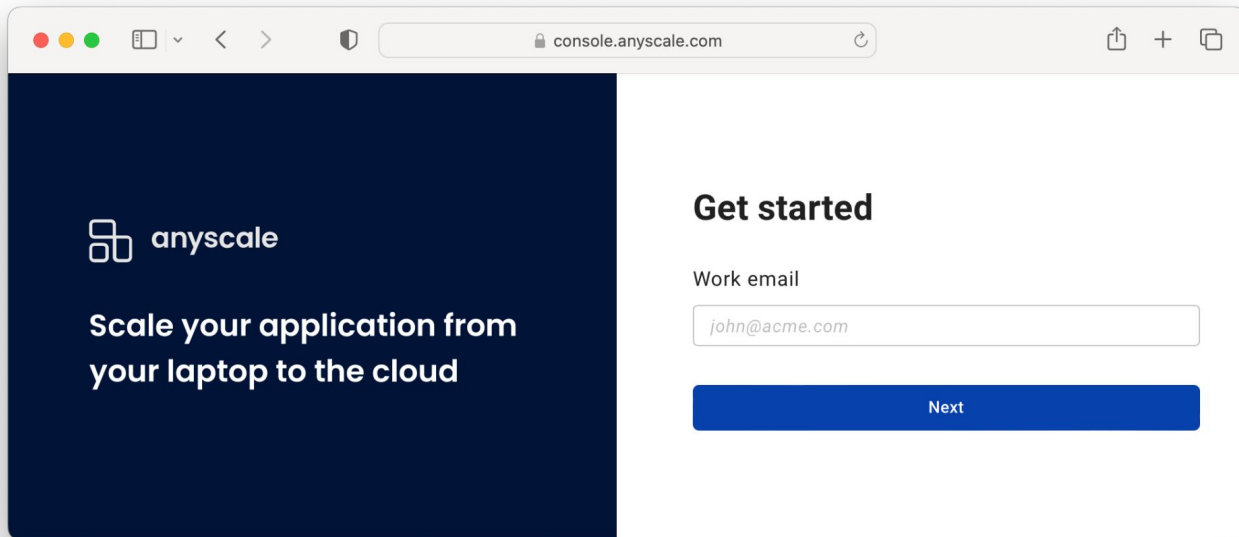


Accessing Anyscale clusters.

- 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



The screenshot shows a web browser window with the address bar displaying 'console.anyscale.com'. The page is split into two main sections. The left section has a dark blue background and contains the Anyscale logo, the text 'anyscale', and the slogan 'Scale your application from your laptop to the cloud'. The right section has a white background and is titled 'Get started'. It contains a 'Work email' label, a text input field with the placeholder 'john@acme.com', and a blue 'Next' button.

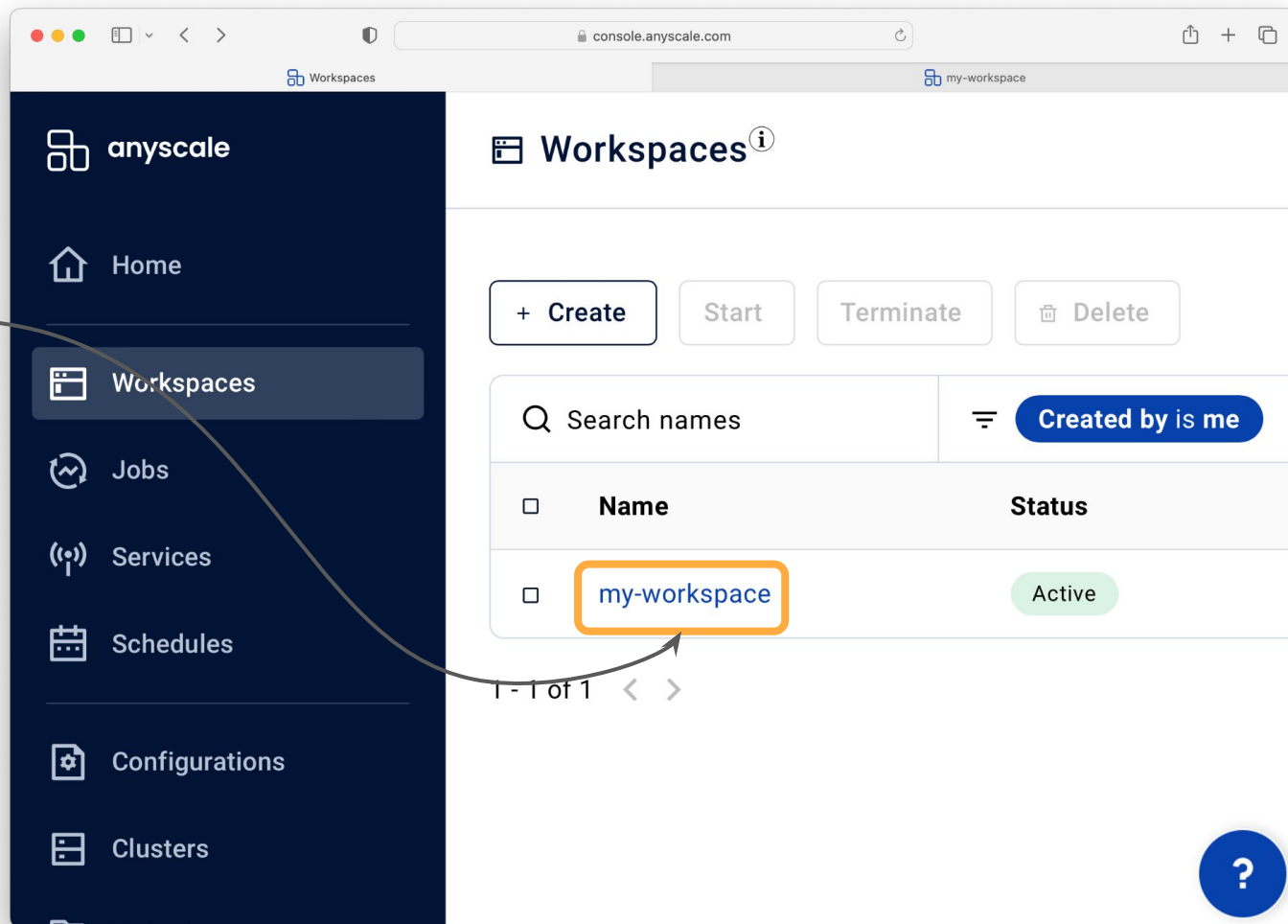
Enter the
**unique
credentials**
sent to your
email!

1. Select Workspaces



The screenshot shows the Anyscale console interface. On the left is a dark blue sidebar with the 'anyscale' logo at the top. Below the logo are several menu items: 'Home' (with a house icon), 'Workspaces' (with a document icon and highlighted by an orange border), 'Jobs' (with a circular arrow icon), 'Services' (with a radio tower icon), 'Schedules' (with a calendar icon), 'Configurations' (with a gear icon), 'Clusters' (with a document icon), and 'Projects' (with a folder icon). On the right is the main content area, which has a light gray header with a 'Home' link and a house icon. Below the header is a section titled 'Examples to get started'. This section contains two cards. The first card is titled 'Introduction to Anyscale & Ray' and has a 'Launch' button and a dropdown arrow. Below the title is a paragraph of text: '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.' At the bottom of this card are two buttons: 'Ray task' and 'Anyscale Job'. The second card is titled 'Many Model Training' and has a 'Launch' button and a dropdown arrow. A blue circle with a white question mark is overlaid on the dropdown arrow of the 'Many Model Training' card. The browser's address bar at the top shows 'console.anyscale.com'.

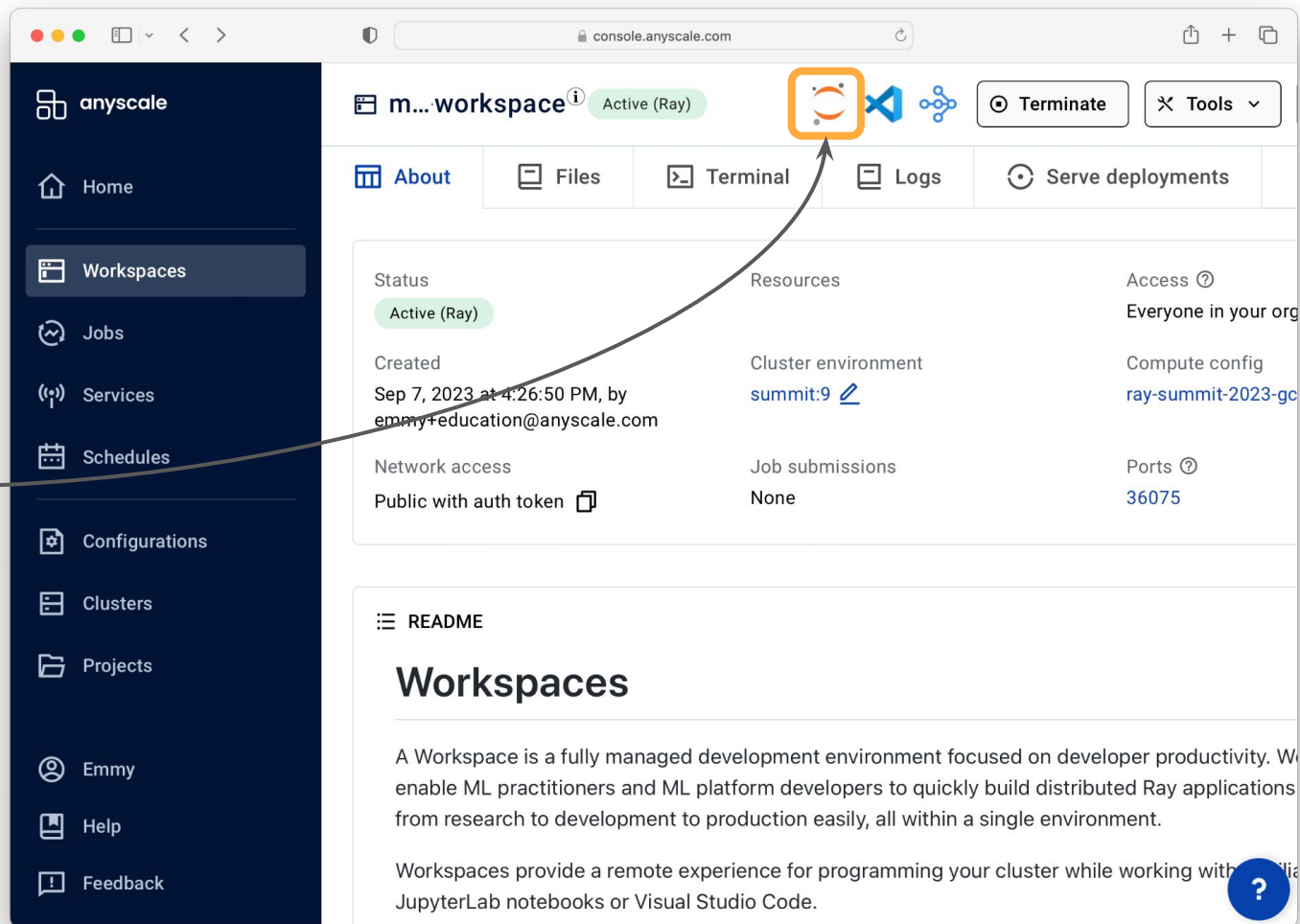
2. Select Your Workspace



The screenshot displays the Anyscale console interface. The left sidebar contains a navigation menu with the following items: Home, Workspaces (highlighted), Jobs, Services, Schedules, Configurations, and Clusters. The main content area is titled 'Workspaces' and includes a '+ Create' button, 'Start', 'Terminate', and 'Delete' buttons. Below these is a search bar labeled 'Search names' and a filter button labeled 'Created by is me'. A table lists the workspaces with columns 'Name' and 'Status'. The table contains one entry: 'my-workspace' with a status of 'Active'. The 'my-workspace' entry is highlighted with an orange box. At the bottom of the table, it says '1 - 1 of 1' with navigation arrows. A blue help button with a question mark is in the bottom right corner.

Name	Status
my-workspace	Active

3. Click on Jupyter icon



The screenshot shows the Anyscale console interface. On the left is a dark blue sidebar with navigation links: Home, Workspaces (selected), Jobs, Services, Schedules, Configurations, Clusters, Projects, Emmy, Help, and Feedback. The main content area displays details for a workspace named 'm...workspace' which is 'Active (Ray)'. At the top of this area, there is a row of icons: a Jupyter icon (highlighted with an orange square), a VS Code icon, and a Ray icon. To the right of these icons are buttons for 'Terminate' and 'Tools'. Below the icons is a tabbed interface with 'About', 'Files', 'Terminal', 'Logs', and 'Serve deployments'. The 'About' tab is active, showing metadata like 'Created Sep 7, 2023 at 4:26:50 PM, by emmy+education@anyscale.com' and 'Cluster environment summit:9'. A curved arrow points from the text '3. Click on Jupyter icon' to the Jupyter icon. A blue circle with a question mark is in the bottom right corner.

anyscale

Home

Workspaces

Jobs

Services

Schedules

Configurations

Clusters

Projects

Emmy

Help

Feedback

m...workspace ⁱ Active (Ray)

Terminate Tools

About Files Terminal Logs Serve deployments

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

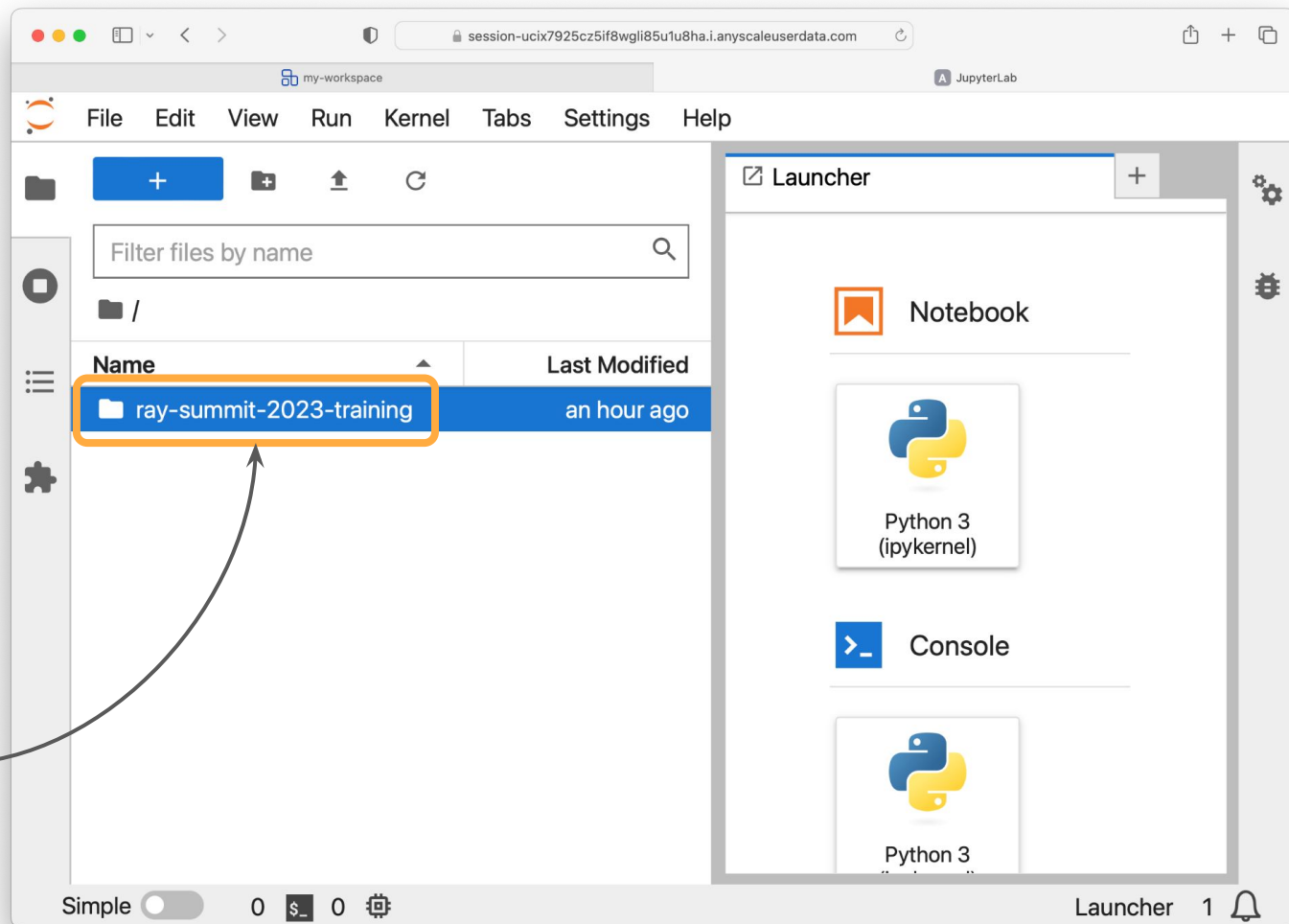
README

Workspaces

A Workspace is a fully managed development environment focused on developer productivity. We 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.

