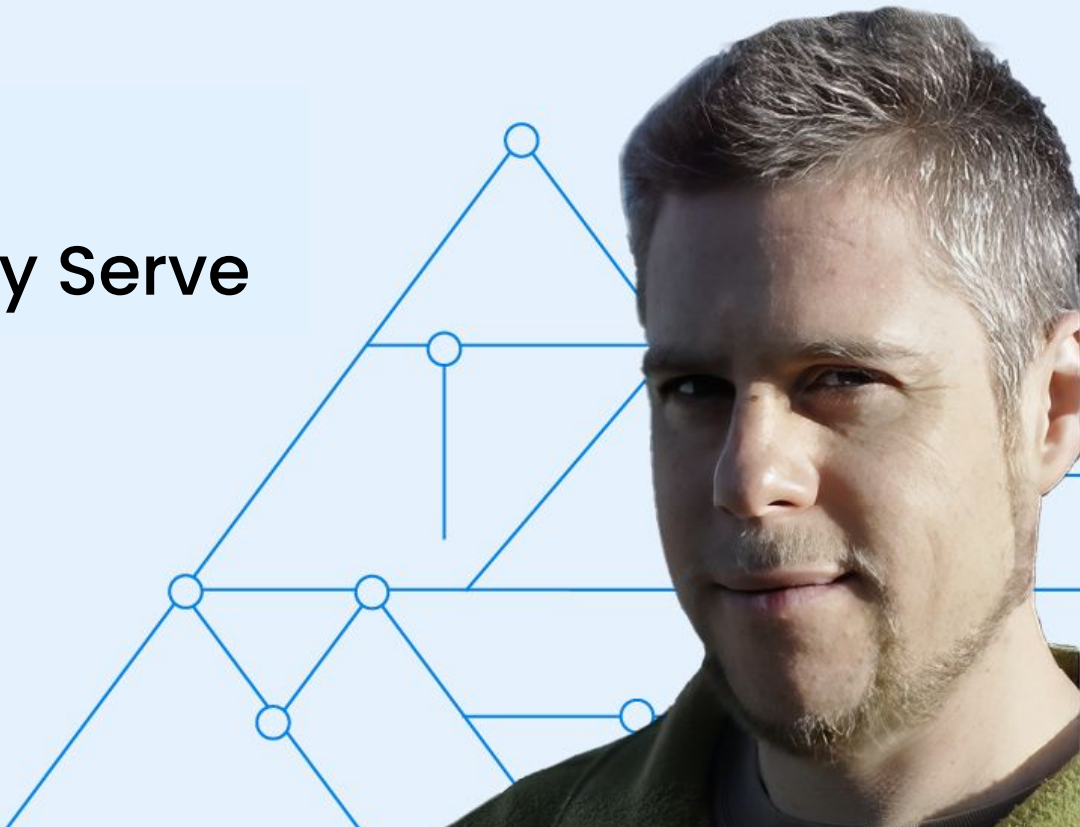




Productionizing AI + LLM Apps with Ray Serve

Adam Breindel
Anyscale

adamb@anyscale.com
[@adbreind](https://twitter.com/adbreind)





Meet the tutorial team!



Marwan

marwan@anyscale.com



Adam

adamb@anyscale.com



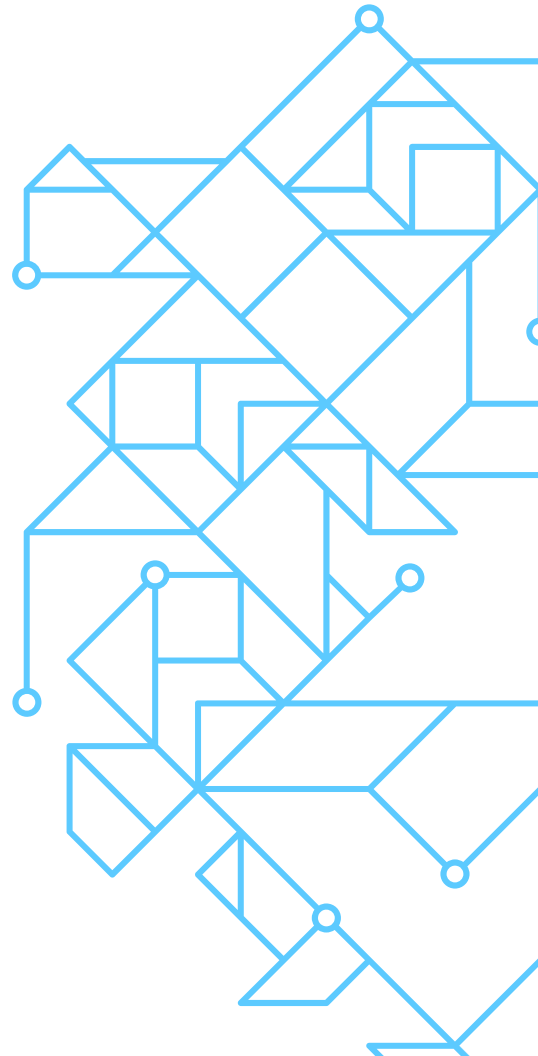
Kamil

kamil@anyscale.com



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.

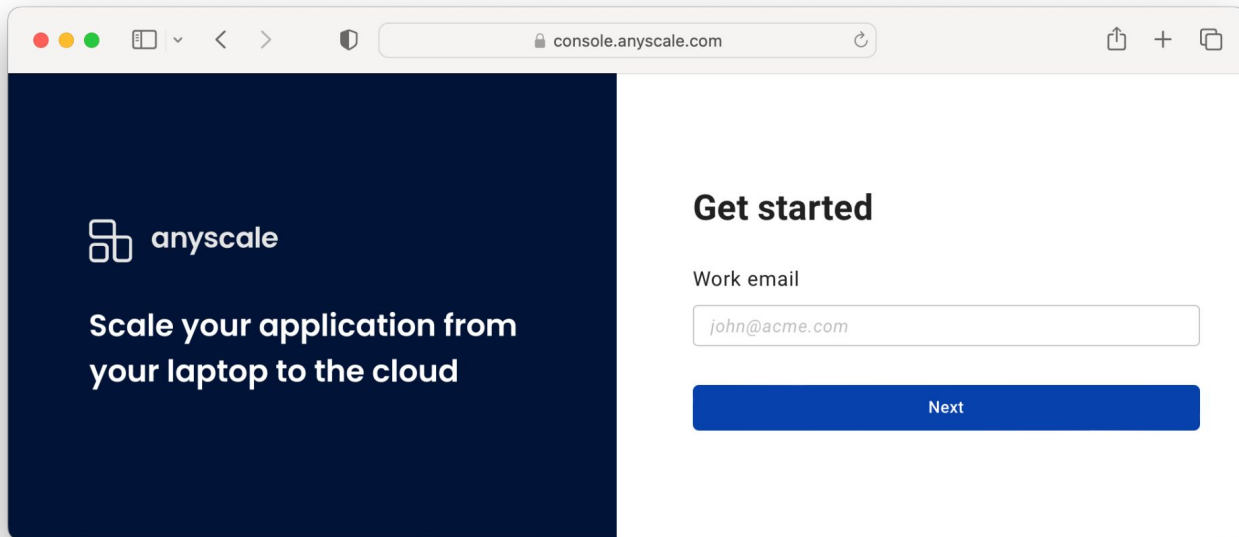


Accessing Anyscale clusters.

- 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

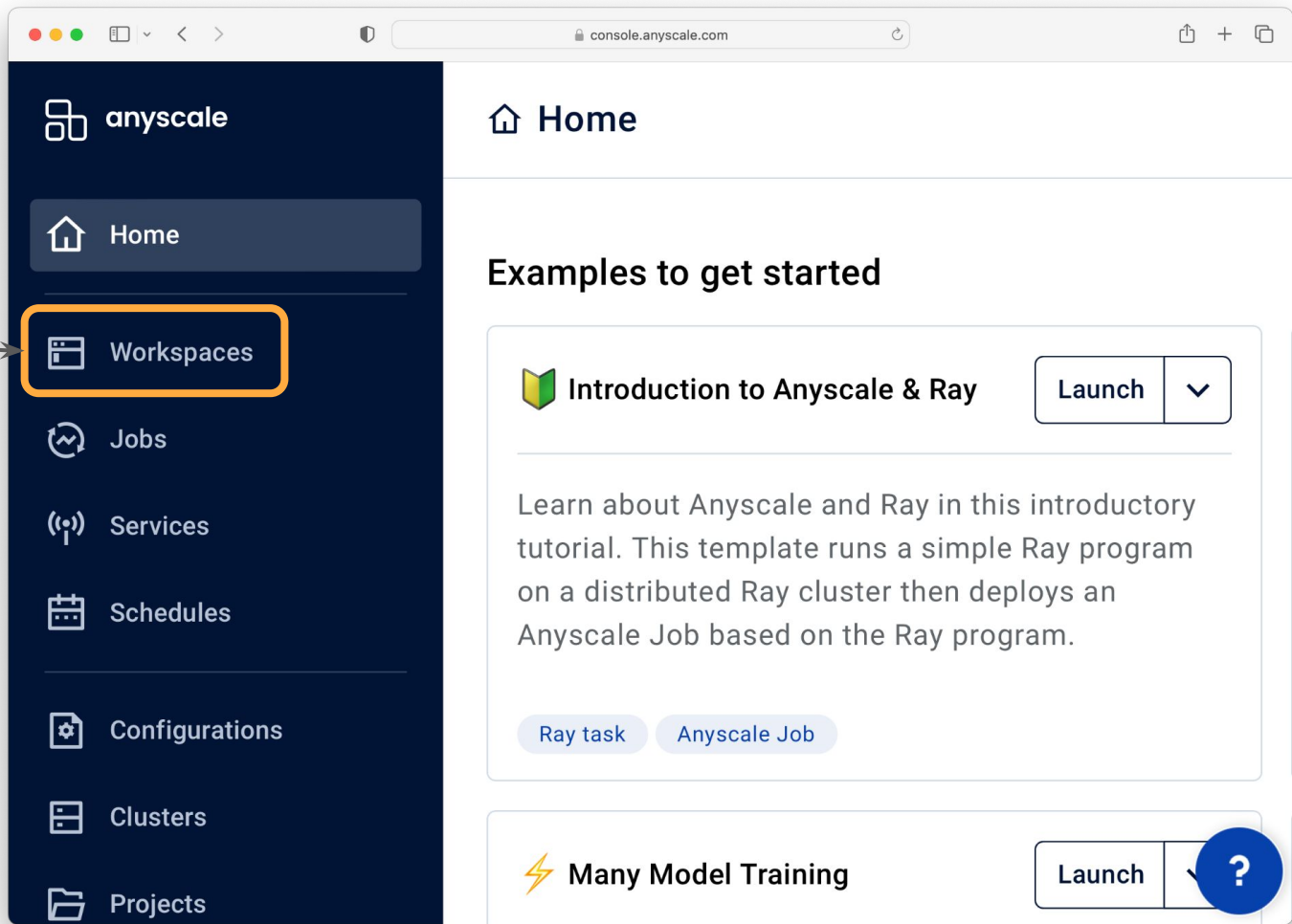
Link to Anyscale cluster: console.anyscale.com



A screenshot of a web browser window showing the Anyscale console login page. The browser's address bar displays 'console.anyscale.com'. The page is split into two main sections. The left section has a dark blue background and features 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



anyscale

Home

Workspaces

Jobs

Services

Schedules


Configurations

Clusters

Projects


Home

Examples to get started

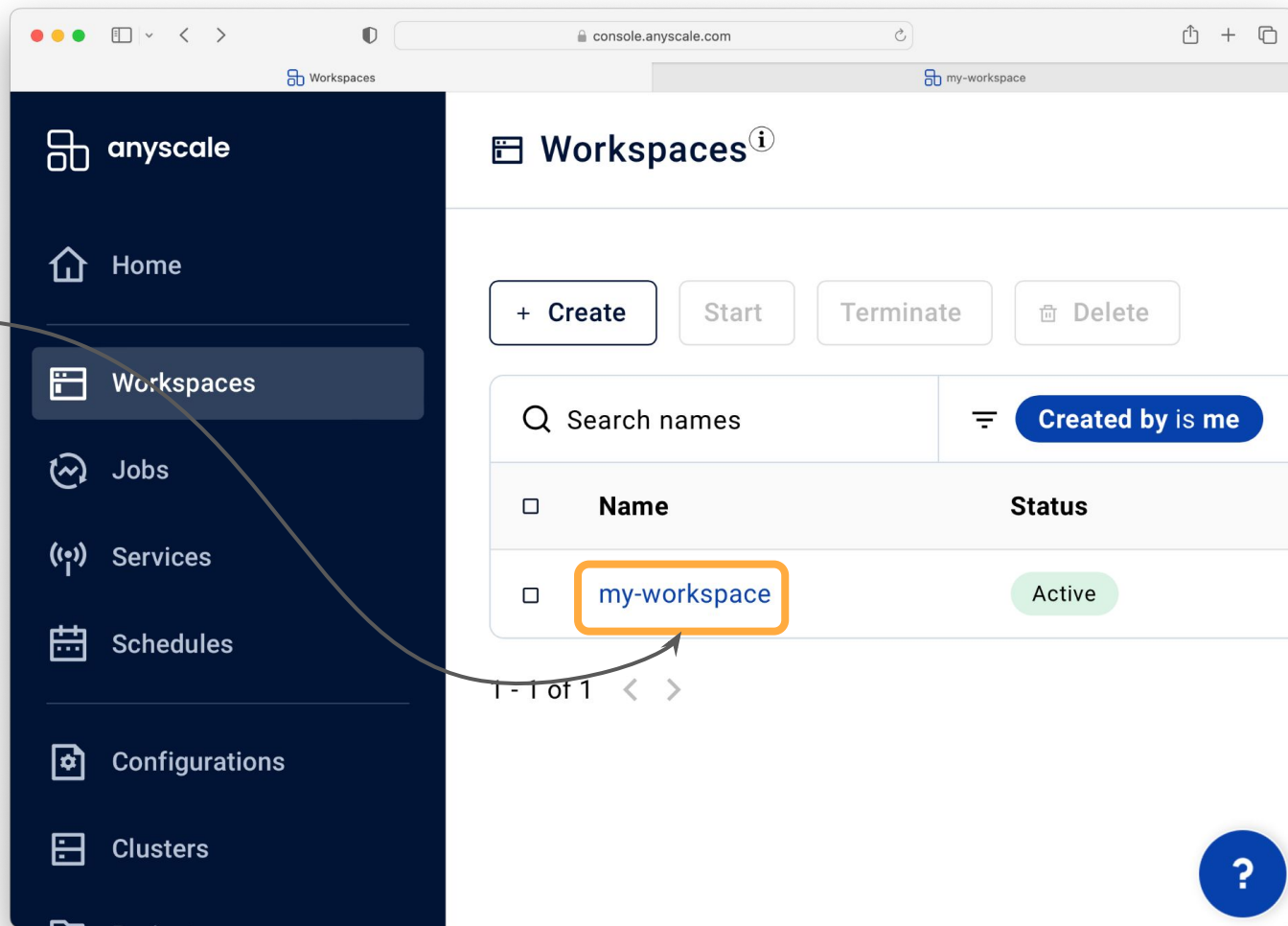
 Introduction to Anyscale & Ray Launch ▼

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 ?

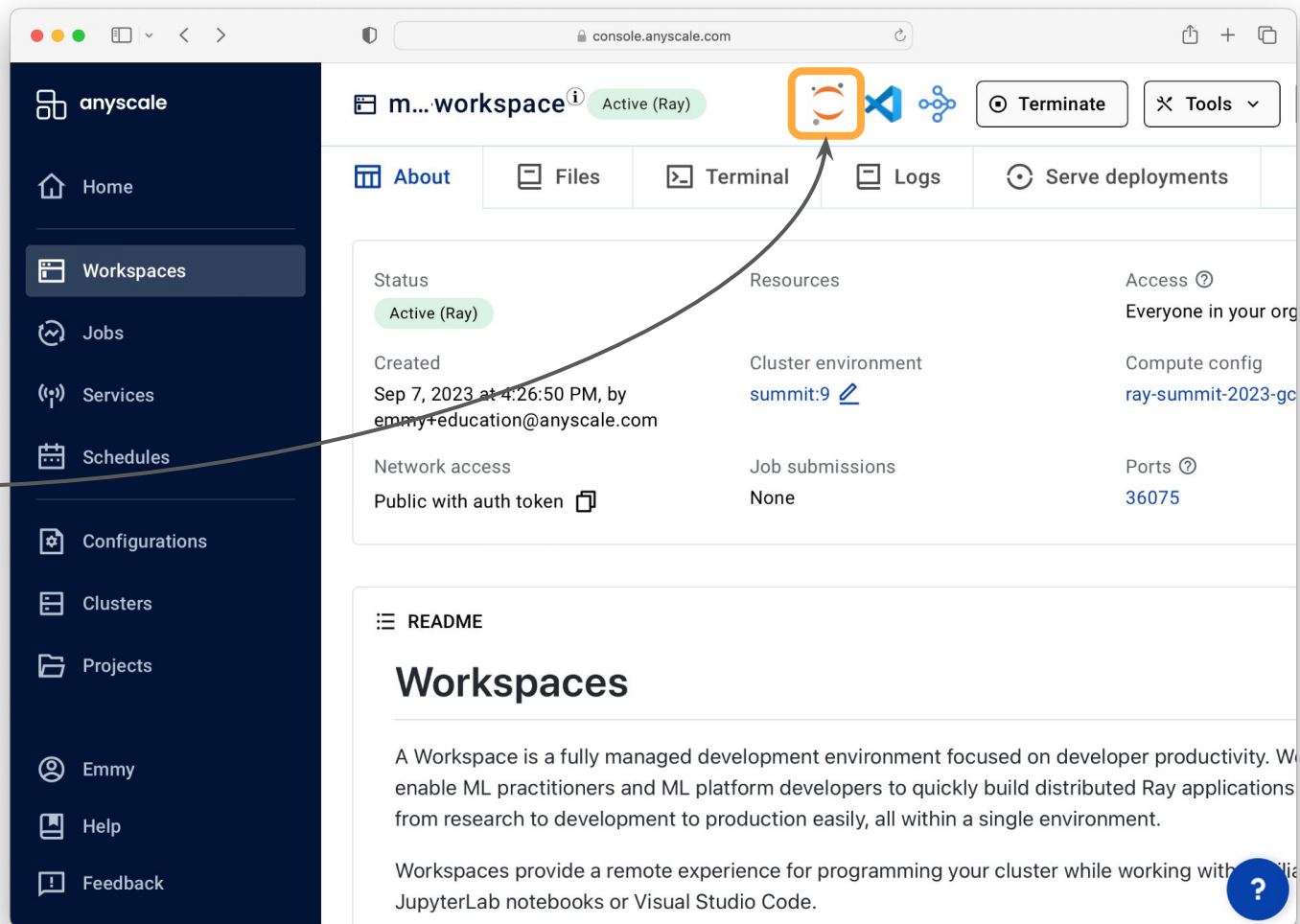
2. Select Your Workspace



The screenshot shows the Anyscale console interface. The left sidebar contains the following menu items: Home, Workspaces (highlighted), Jobs, Services, Schedules, Configurations, and Clusters. The main content area is titled 'Workspaces' and includes buttons for '+ Create', 'Start', 'Terminate', and 'Delete'. Below these buttons 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'. An orange box highlights the 'my-workspace' entry, and a curved arrow points from the 'Workspaces' menu item in the sidebar to this entry. The bottom of the table shows '1 - 1 of 1' and navigation arrows. A blue help button with a question mark is in the bottom right corner.

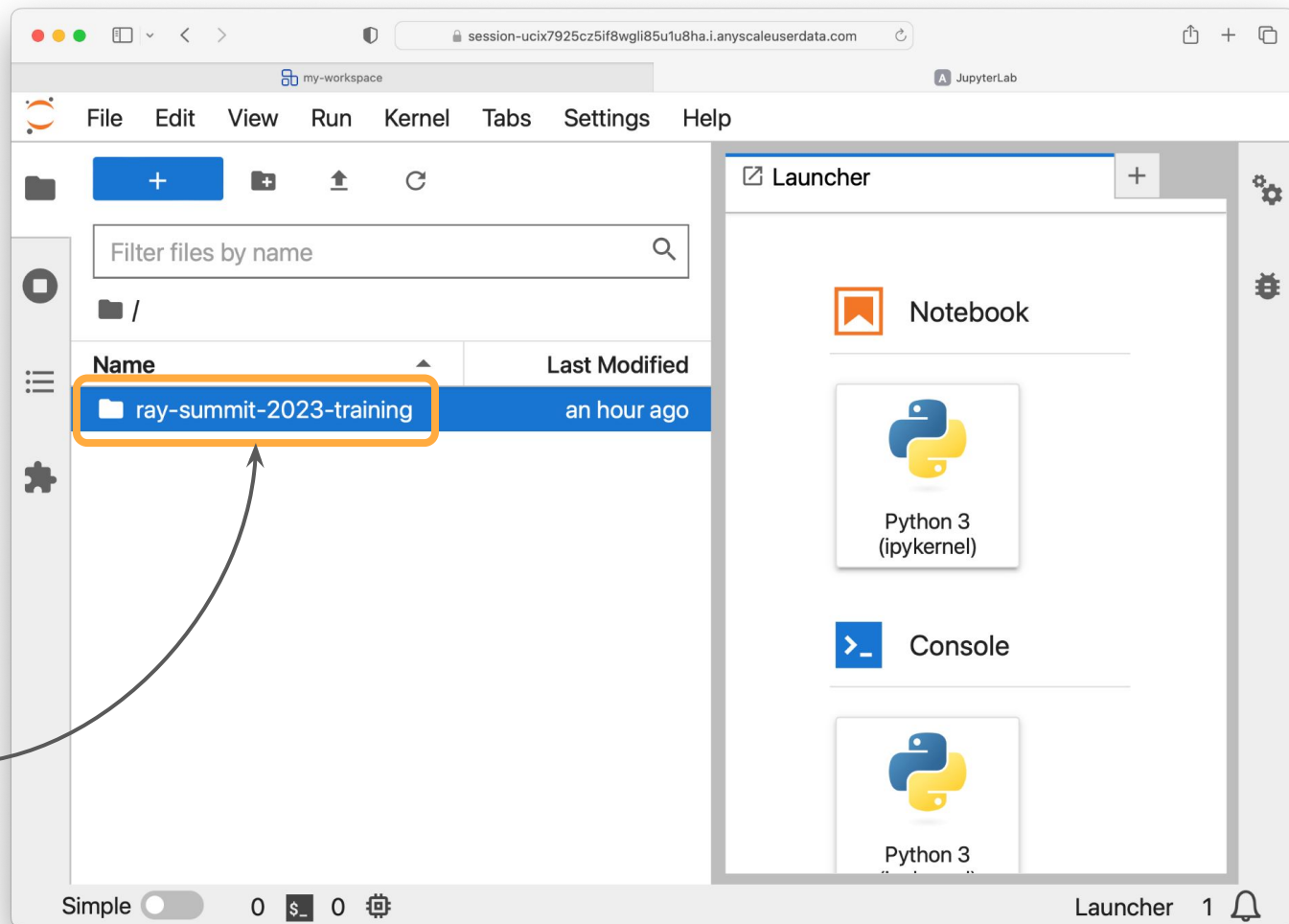
Name	Status
my-workspace	Active

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 Visual Studio 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 'Network access' (Public with auth token). A table below lists 'Resources' with columns for 'Status' (Active (Ray)), 'Cluster environment' (summit:9), 'Access' (Everyone in your org), 'Compute config' (ray-summit-2023-gc), 'Job submissions' (None), and 'Ports' (36075). At the bottom, there is a 'README' section titled 'Workspaces' with a description of the workspace environment. A blue question mark icon is visible in the bottom right corner of the console.

4. Find the content for your class here.





**Time for a
Break!**

15 minutes.



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

More Resources

For further exploration with
Ray, Anyscale, and LLMs.





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.



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.

