

Generative AI + Ray

Fine-tuning and Deploying Stable Diffusion

Emmy Li, Kourosh Hakhamaneshi, Justin Yu





Welcome!

We're happy to have you here.





Meet the team!



Emmy



TA

Kourosh



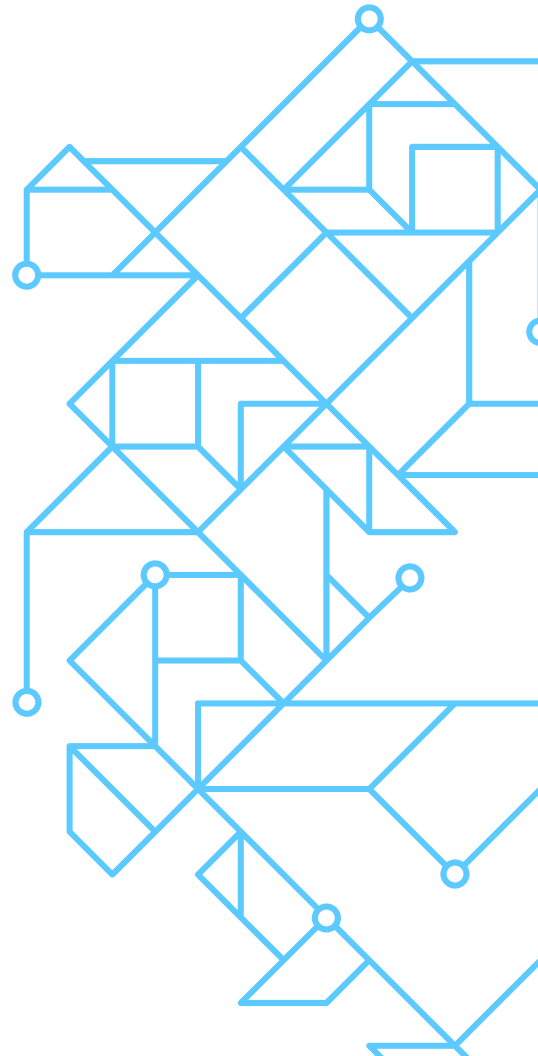
TA

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.

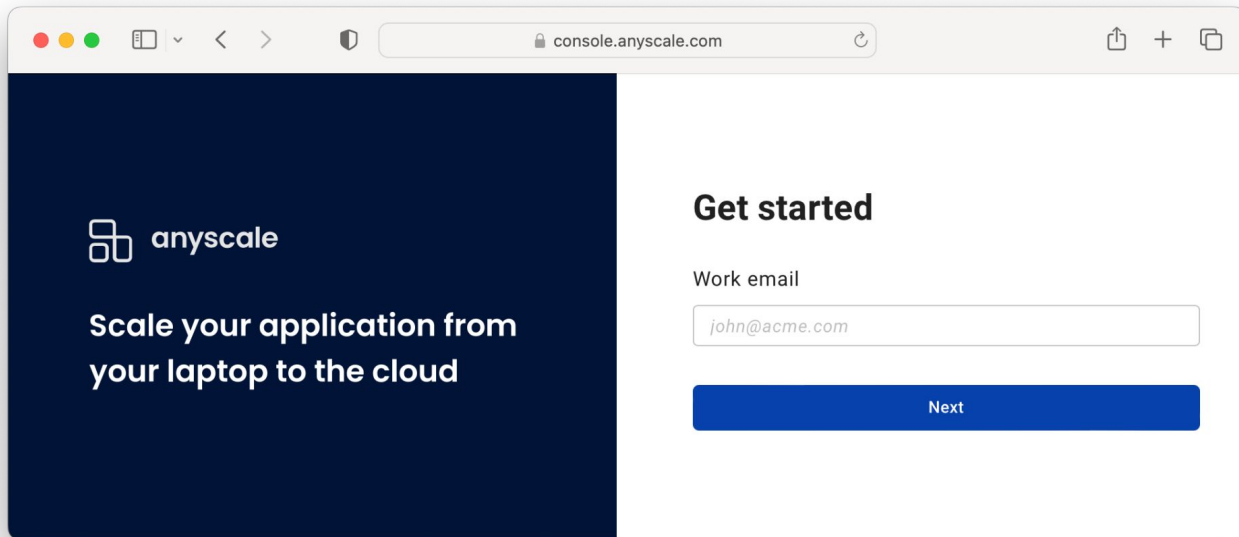


Accessing Anyscale clusters.

- 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.any scale.com



The screenshot shows a web browser window with the address bar displaying 'console.any scale.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.

anyscale

Scale your application from
your laptop to the cloud

Get started

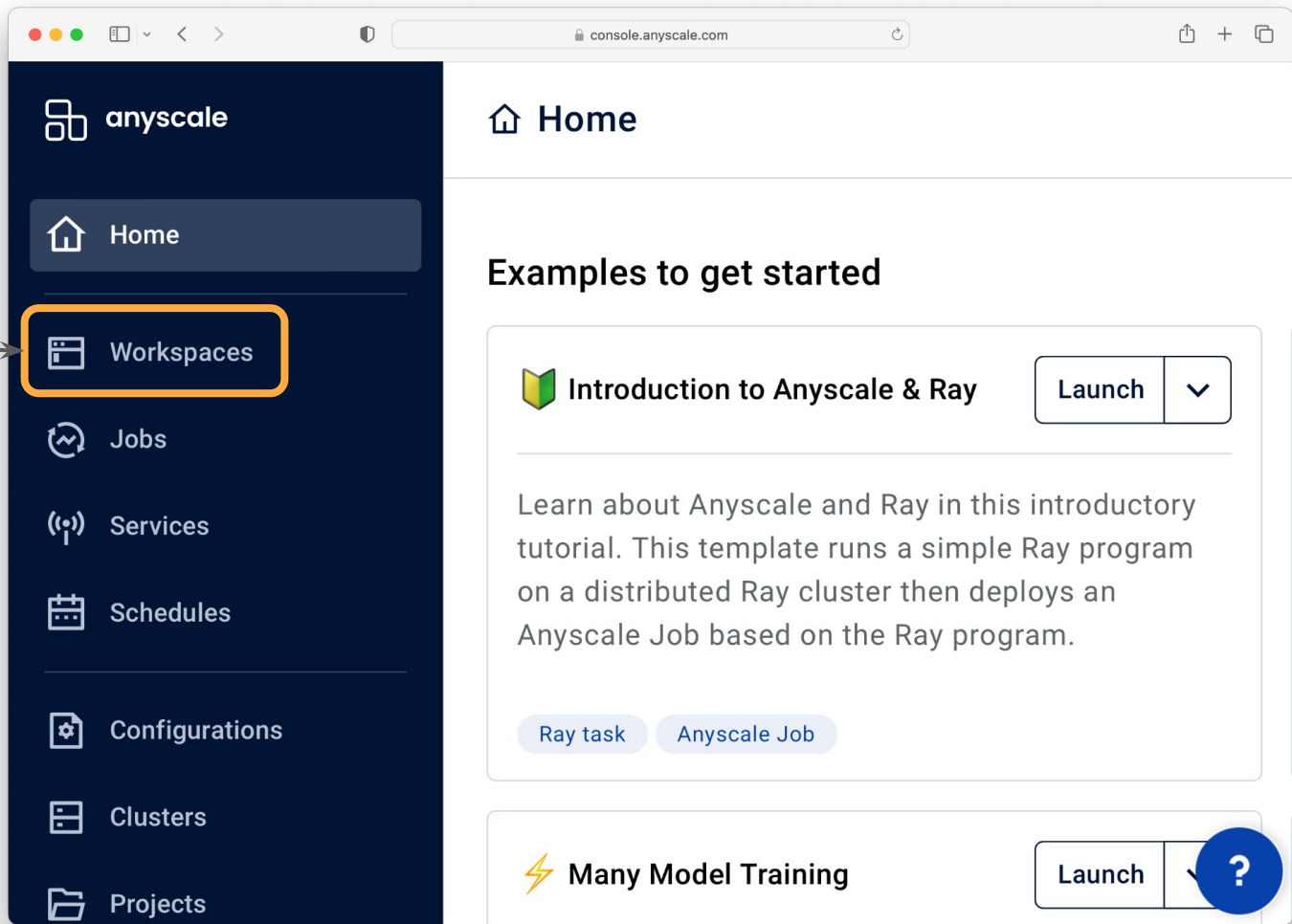
Work email

john@acme.com

Next

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

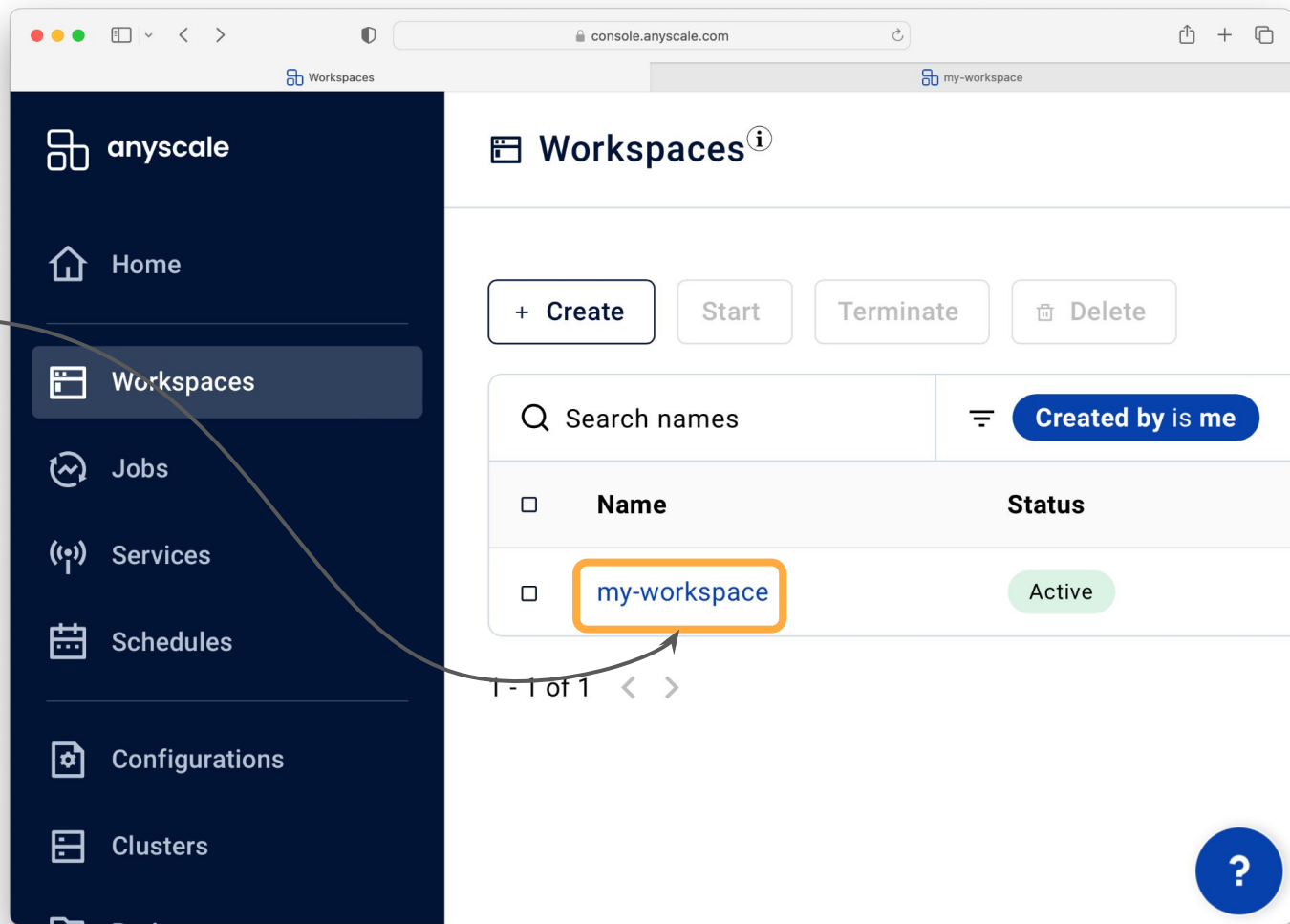
 Introduction to Anyscale & Ray Launch ▼

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 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 is titled 'm... workspace' and shows it is 'Active (Ray)'. Below this are tabs for About, Files, Terminal, Logs, and Serve deployments. A table displays workspace details: 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), and Access (Everyone in your org). A 'Jupyter' icon (a circle with a dot and a line) is highlighted with an orange box in the top right of the workspace header. An arrow points from the text '3. Click on Jupyter icon' to this icon. Other icons for VS Code and a share icon are also visible. Buttons for 'Terminate' and 'Tools' are in the top right. A 'README' section titled 'Workspaces' is at the bottom, explaining that a workspace is a fully managed development environment.

anyscale

Home

Workspaces

Jobs

Services

Schedules

Configurations

Clusters

Projects

Emmy

Help

Feedback

m... workspace ⁱ Active (Ray)

About Files Terminal Logs Serve deployments

Terminate Tools

Status	Resources	Access [?]
Active (Ray)		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

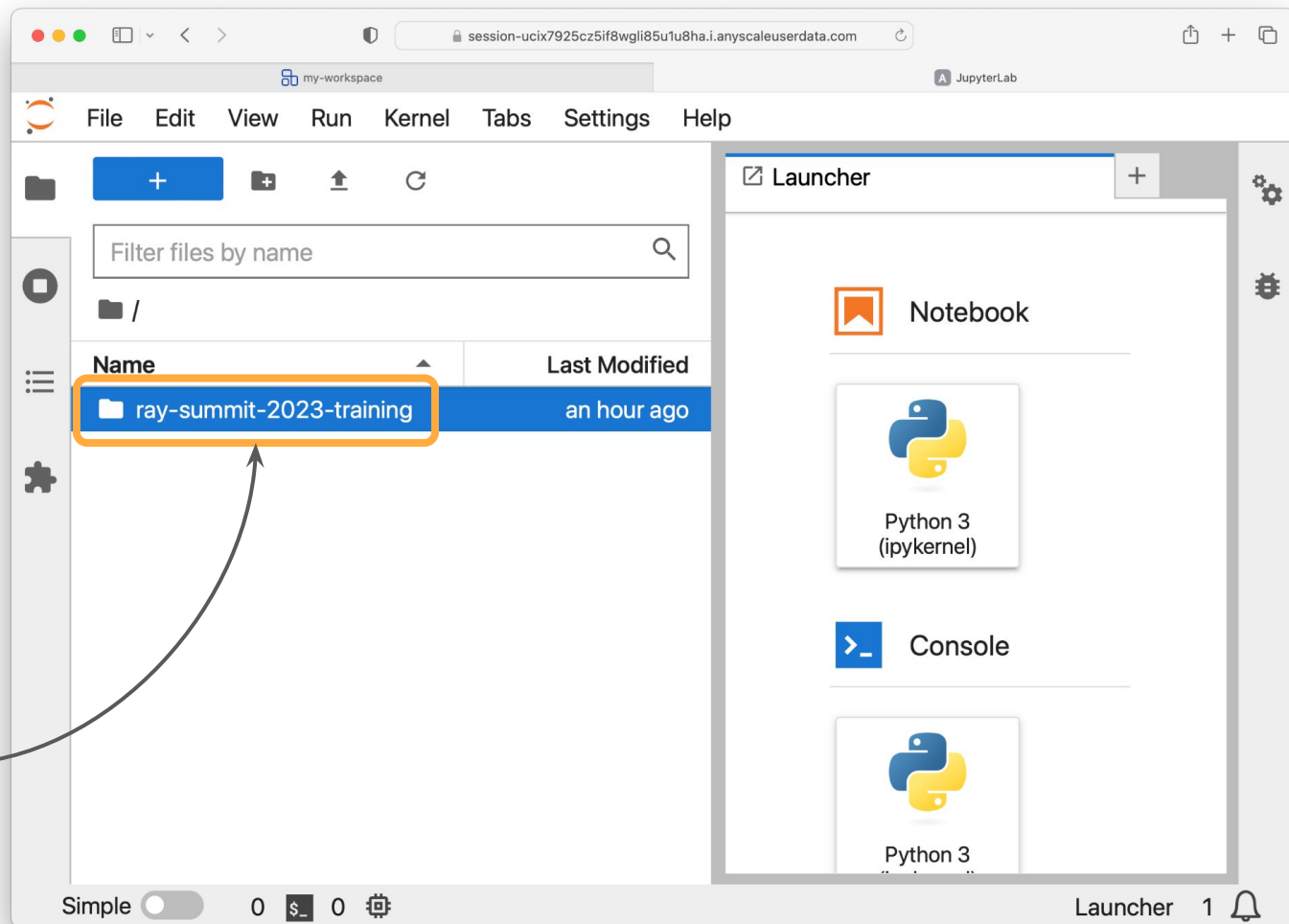
README

Workspaces

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



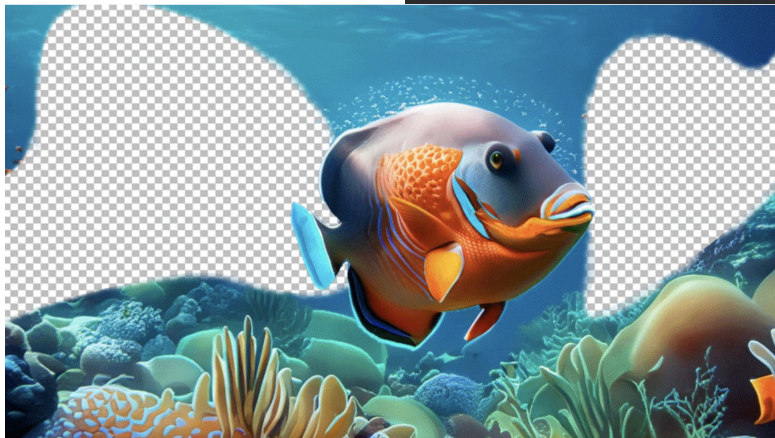
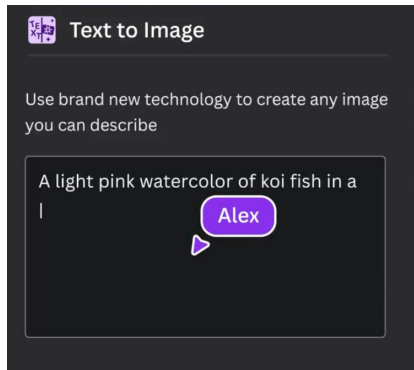
From local to cloud

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

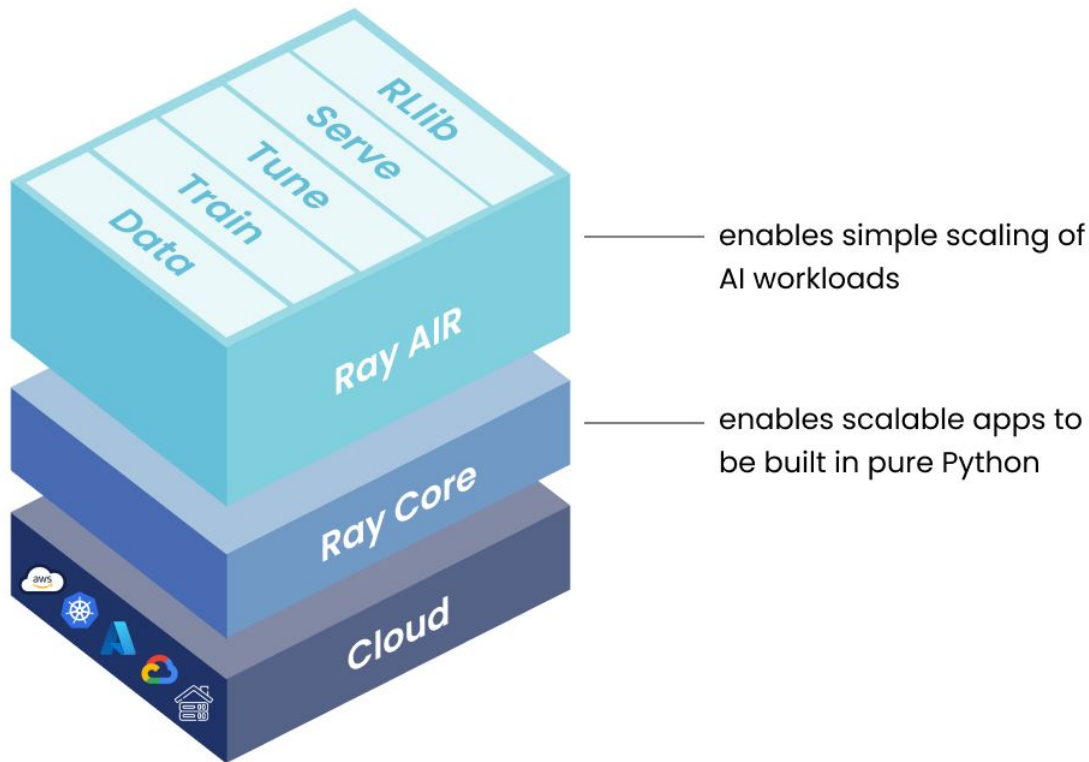
Inspect the infrastructure and ML application layers.

Intuitive cost control

Clarity into \$\$\$-eating resources and inefficiencies.



The Ray AI Libraries



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.

