Module 12

Workshop 2: Scaling and Governing ML Systems across the Lifecycle

Zombie model

Imagine an old ML model at your company refuses to retire, it just keeps running forever like a zombie!

What funny or disastrous things might happen if this “zombie model” is left in production too long?

Type your ideas in the chat!

Module 12

Workshop 2: Scaling and Governing ML Systems across the Lifecycle

Agenda

- **Review:** Recap key concepts.
- **Practical exercise:** AutoDeliver lifecycle challenge.
- **Closing:** Wrap up and reflection.

Learning objectives

- Evaluate scalability requirements for ML systems by analysing data variety, data quality, and resource needs under varying loads.
- Design lifecycle governance plans that integrate change management workflows and logging practices for production ML systems.
- Develop decommissioning and archiving protocols that ensure compliance, traceability, and business continuity when retiring ML models.

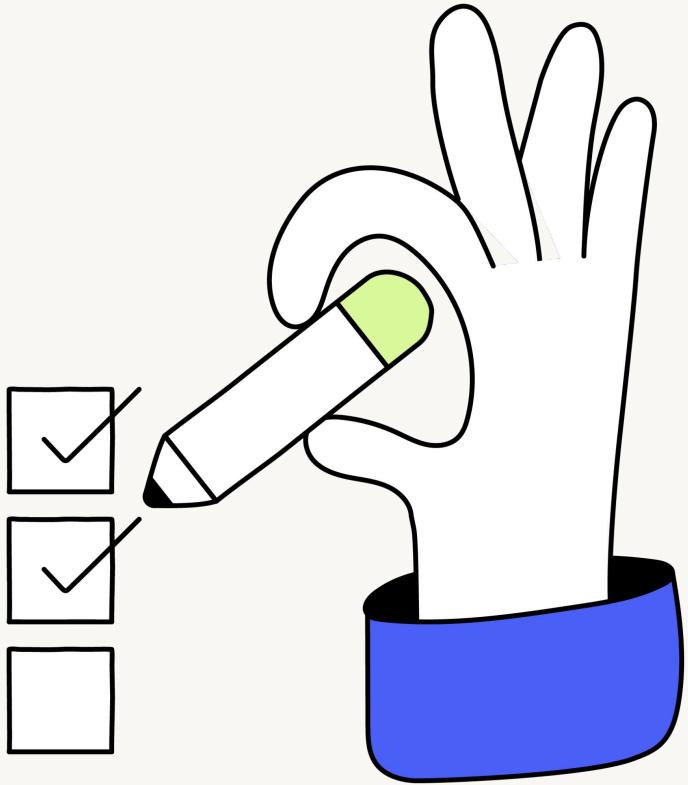
Async review: Units 2 and 3

Go to My Multiverse, Module 12
Workshop 2

Navigate to the **“Async review”** page

Review the summaries of Units 2 and 3, then complete the following Poll:

From scaling to retiring



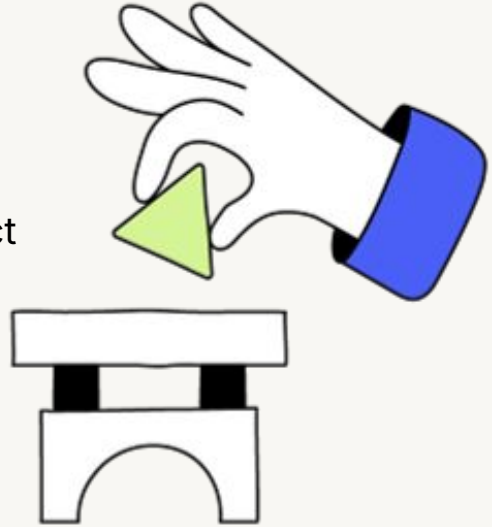
Poll debrief

- Scaling, governing, and retiring ML systems each come with unique challenges.

Which of these do you think is most difficult to manage in practice, and why?

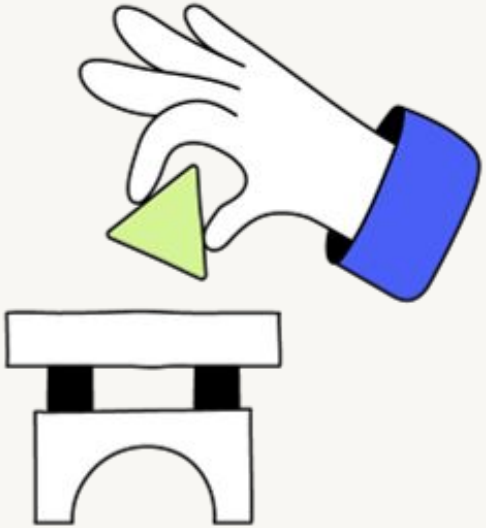
Scalability essentials

- **Throughput** and **latency** define system performance under load.
- **Scaling strategies:**
  - **Vertical scaling =** add more power to one machine.
  - **Horizontal scaling =** add more machines to share the load.
  - **Autoscaling =** automatic adjustment based on demand.
- Data variety and formats (structured, unstructured, multimodal) affect scalability and resource requirements.
- Data quality issues (missing values, schema changes, noisy inputs) increase compute costs and reduce efficiency.



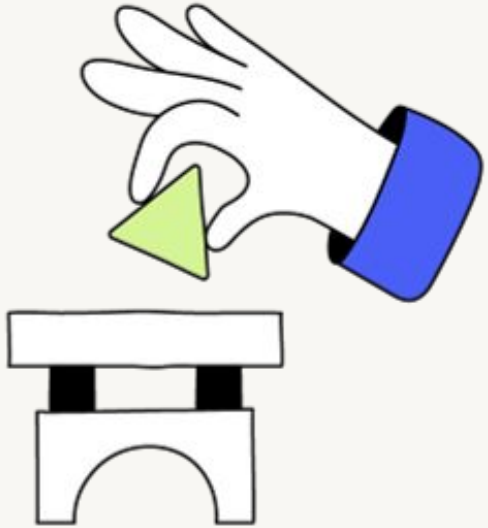
Supply chain and governance risks

- **Supply chain risks:** Vendor lock-in, reliance on third-party APIs, GPU shortages.
- **Change management:**
  - Version control and model registries.
  - Staged deployments: Shadow, canary, A/B testing.
- **Logging essentials:** Inference logs, metadata, audit trails for compliance and traceability.



Decommissioning protocols

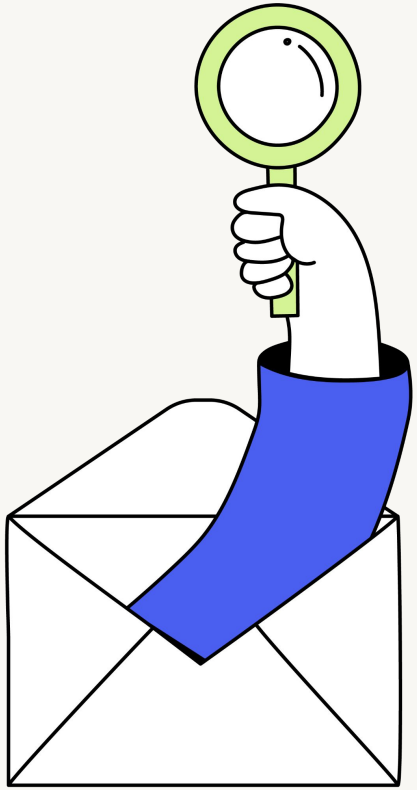
- **Dependency mapping:** Identify systems and stakeholders relying on the model.
- **Traffic diversion:** Safely shift users/services to newer models through staged rollouts.
- **De-provisioning:** Retire compute and storage resources once dependencies are cleared.
- **Archiving:** Retain models, datasets, logs, and documentation for compliance and traceability.
- **Communication:** Inform stakeholders of retirement timelines and mitigation plans to maintain business continuity.



Practical exercise: AutoDeliver lifecycle challenge

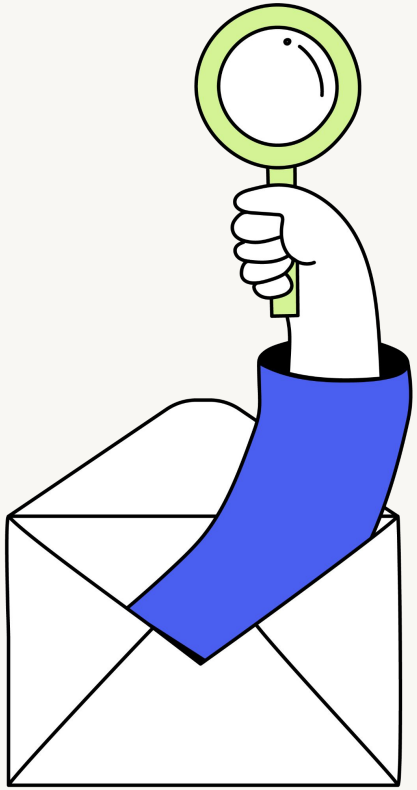
Practical exercise: Context

- **Company:** AutoDeliver, a logistics startup using ML for its autonomous delivery fleet.
- **Initiative:** Models manage navigation, traffic prediction, and routing to keep deliveries fast and efficient.
- **Challenge:**
  - Fleet growth is creating **scalability pressures**.
  - **Dependence on third-party** map APIs and cloud GPUs adds supply chain and governance risks.
  - A new **V2 navigation model** is planned, but older models are still active, raising retirement questions.
  - Regulators require **traceability**.



Practical exercise instructions

- Work in breakout rooms in small groups.
- Step into the role of AutoDeliver’s ML team.
- Find all instructions and resources on Ariel.



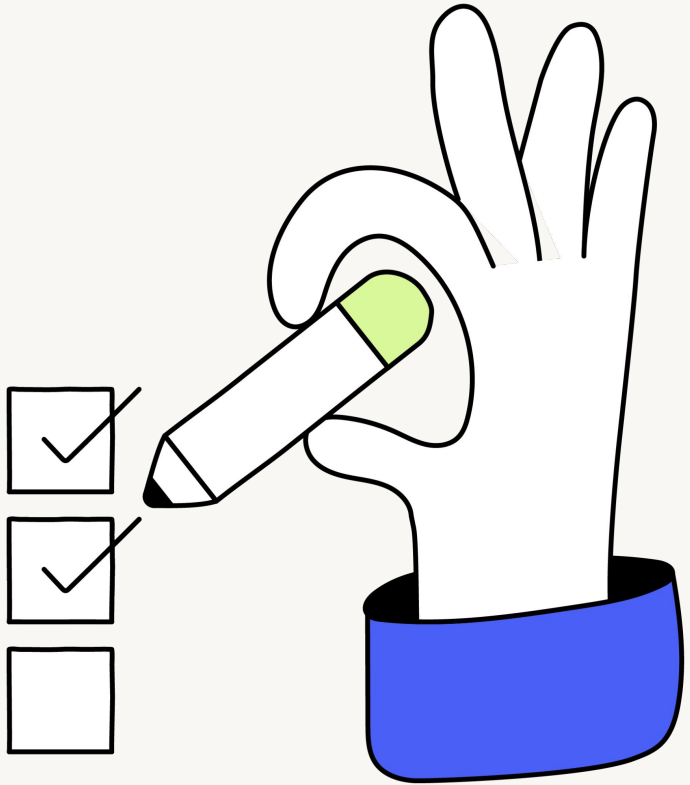
Access the skills application

Go to My Multiverse, Module 12 Workshop 2

Navigate to the **“Practical exercise”** page

Start working on the activity tasks

Be prepared to share in 20 minutes when we come back together



Group discussion debrief

Practical exercise debrief

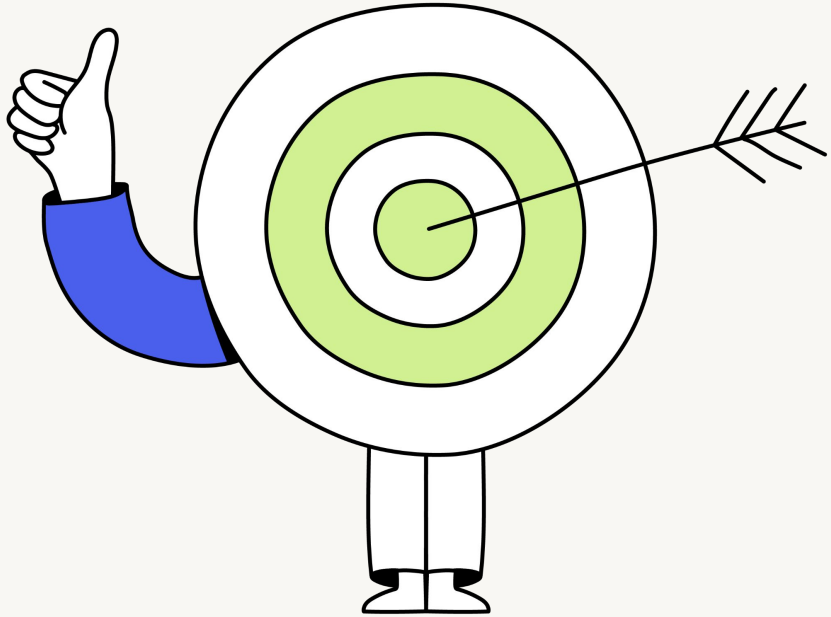
1. What is **one scalability risk** your group identified in AutoDeliver’s system?

2. What is **one governance practice** you would recommend to strengthen reliability and compliance?

3. How would you decide **when and how to retire an outdated model** without disrupting service?

Key takeaways

1. Scalability risks show up in throughput, latency, and resource strain.
2. Governance practices like logging, versioning, and staged deployments build reliability and compliance.
3. Retiring models responsibly requires dependency mapping, traffic diversion, and archiving.
4. Linking lifecycle decisions to business impact helps secure stakeholder support.



Applying key takeaways to your organisation

Applying key takeaways

1. How does your organisation currently **plan for scaling** ML systems as demand grows?

2. What **governance practices** are in place, and where are the gaps?

3. How does your team decide when and how to **retire ML models**, and what could make that process smoother?

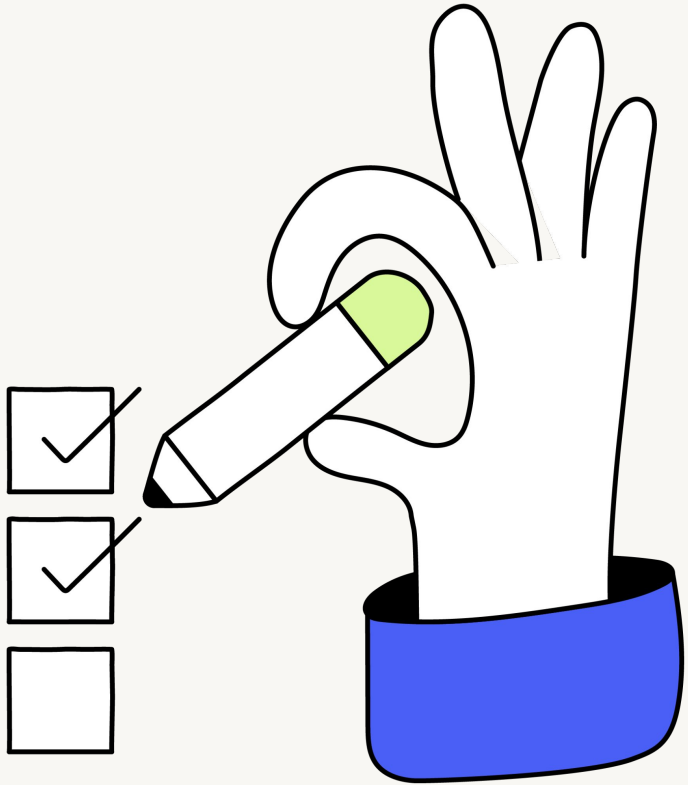
Share your plan to apply new skills

Go to My Multiverse, Module 12
Workshop 2

Navigate to the discussion activity page **“Apply new skills to your role”**

Take 5 minutes to share how you will apply this skill in your role

Take time this week to read what others have written and get new ideas!



Closing out

- **Next up:** Module Wrap-Up Workshop
- **If you have questions, you can:**
  - Chat with Atlas
  - Reach out to your Cohort Coach
  - Sign up for Tutoring
- **Log your OTJ hours**
- **Take the Workshop Survey** on My Multiverse

multiverse