

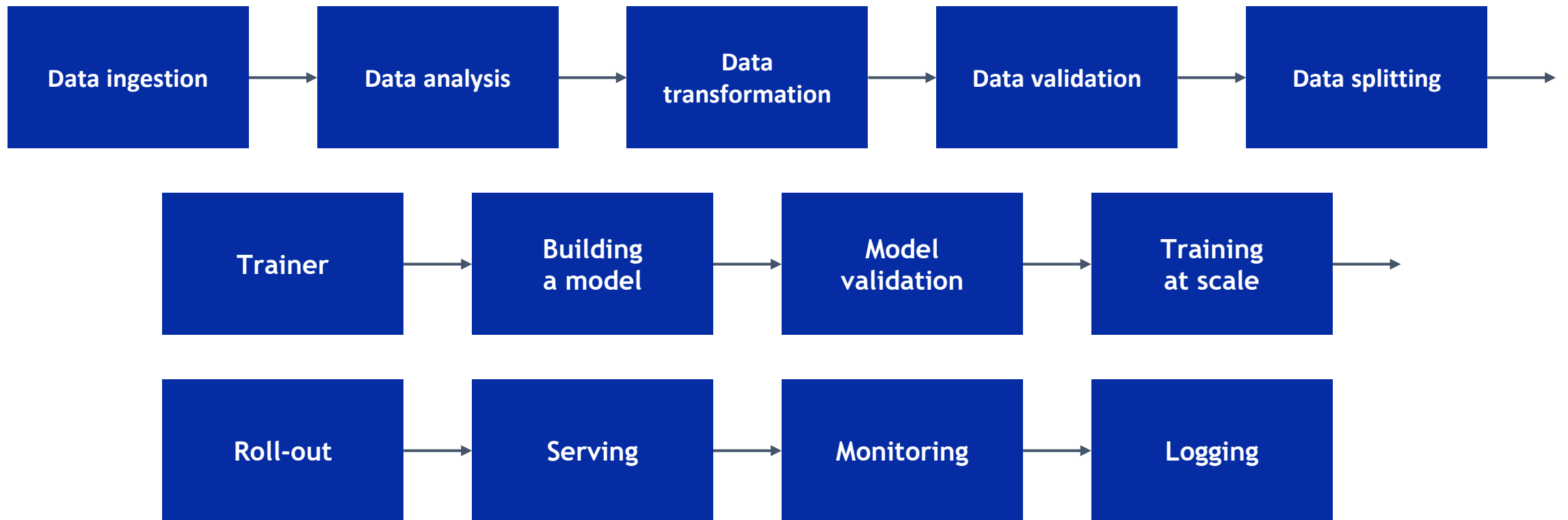


# DevOps for ML ("MLOps")

Priyanshi Singh  
Cloud Solution Architect  
Artificial Intelligence & Machine Learning

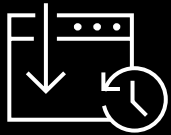


**Building a model**



# MLOps == How to bring ML to production

Bring together **people**, **process**, and **platform** to automate ML-infused software delivery & provide continuous value to our users.



## People

- Blend together the work of individual engineers in a repository.
- Each time you commit, your work is automatically built and tested, and bugs are detected faster.
- Code, data, models and training pipelines are shared to accelerate innovation.



## Process

- Provide templates to bootstrap your infrastructure and model development environment, expressed as code.
- Automate the entire process from code commit to production.



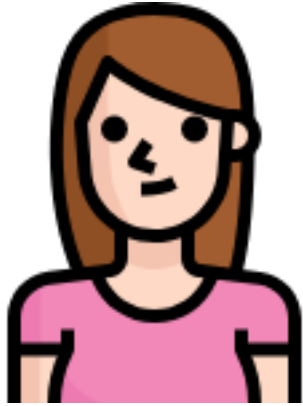
## Platform

- Safely deliver features to your customers as soon as they're ready.
- Monitor your pipelines, infrastructure and products in production and know when they aren't behaving as expected

**Ok, but, like, I'm  
a data scientist. ~~IDGAF~~  
I don't care  
about all that.**

**Yes You Do!**

# Cowboys and Ranchers Can Be Friends!



**Data Scientist**

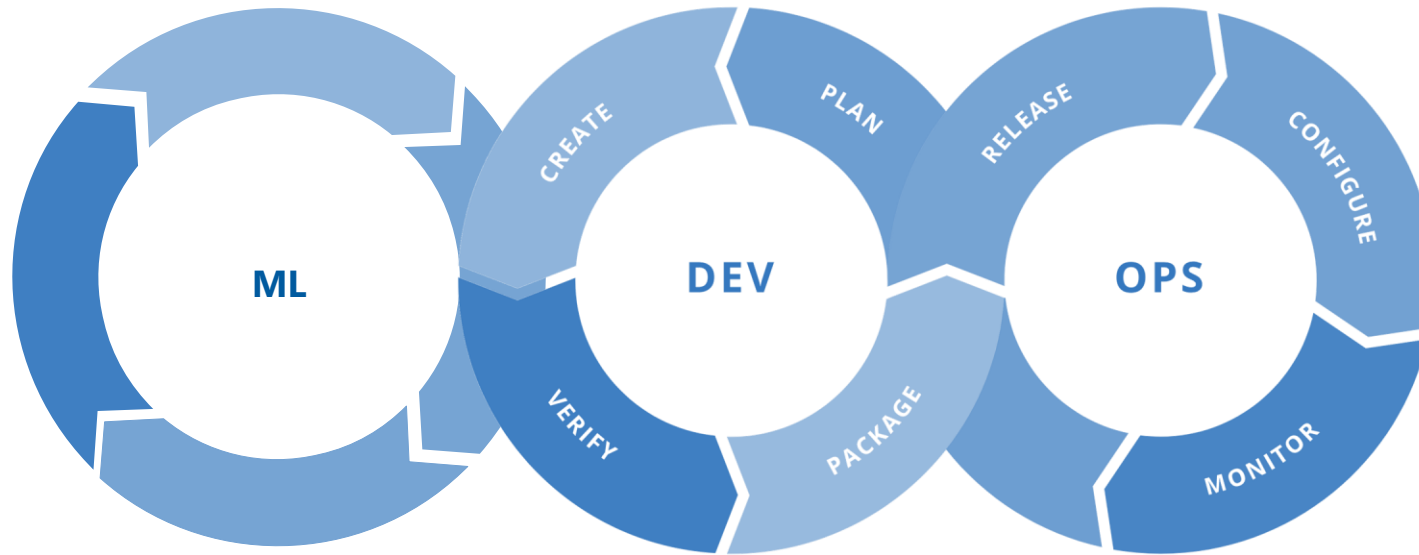
- Quick iteration
- Frameworks they understand
- Best of breed tools
- No management headaches
- Unlimited scale



**SRE/ML Engineers**

- Reuse of tooling and platforms
- Corporate compliance
- Observability
- Uptime

# MLOps = ML + DEV + OPS



## Experiment

Data Acquisition  
Business Understanding  
Initial Modeling

## Develop

Modeling + Testing  
Continuous Integration  
Continuous Deployment

## Operate

Continuous Delivery  
Data Feedback Loop  
System + Model Monitoring



# MLOps Benefits

## Automation / Observability

---

- Code drives **generation** and **deployments**
- Pipelines are **reproducible** and **verifiable**
- All artifacts can be **tagged** and **audited**

## Validation

---

- SWE best practices for **quality control**
- Offline comparisons of **model quality**
- Minimize **bias** and enable **explainability**

## Reproducibility /Auditability

---

- Controlled rollout capabilities
- Live comparison of **predicted vs. expected performance**
- Results fed back to watch for drift and improve model

**== VELOCITY and SECURITY (For ML)**

# There are many jobs & tools involved in production ML



Data Scientist

Azure Machine Learning  
GitHub  
TensorFlow, PyTorch, sklearn  
Azure Compute – CPU/GPU/FPGA



IT / Ops



Data Analyst



Business Owner



& many more...



Data Engineer

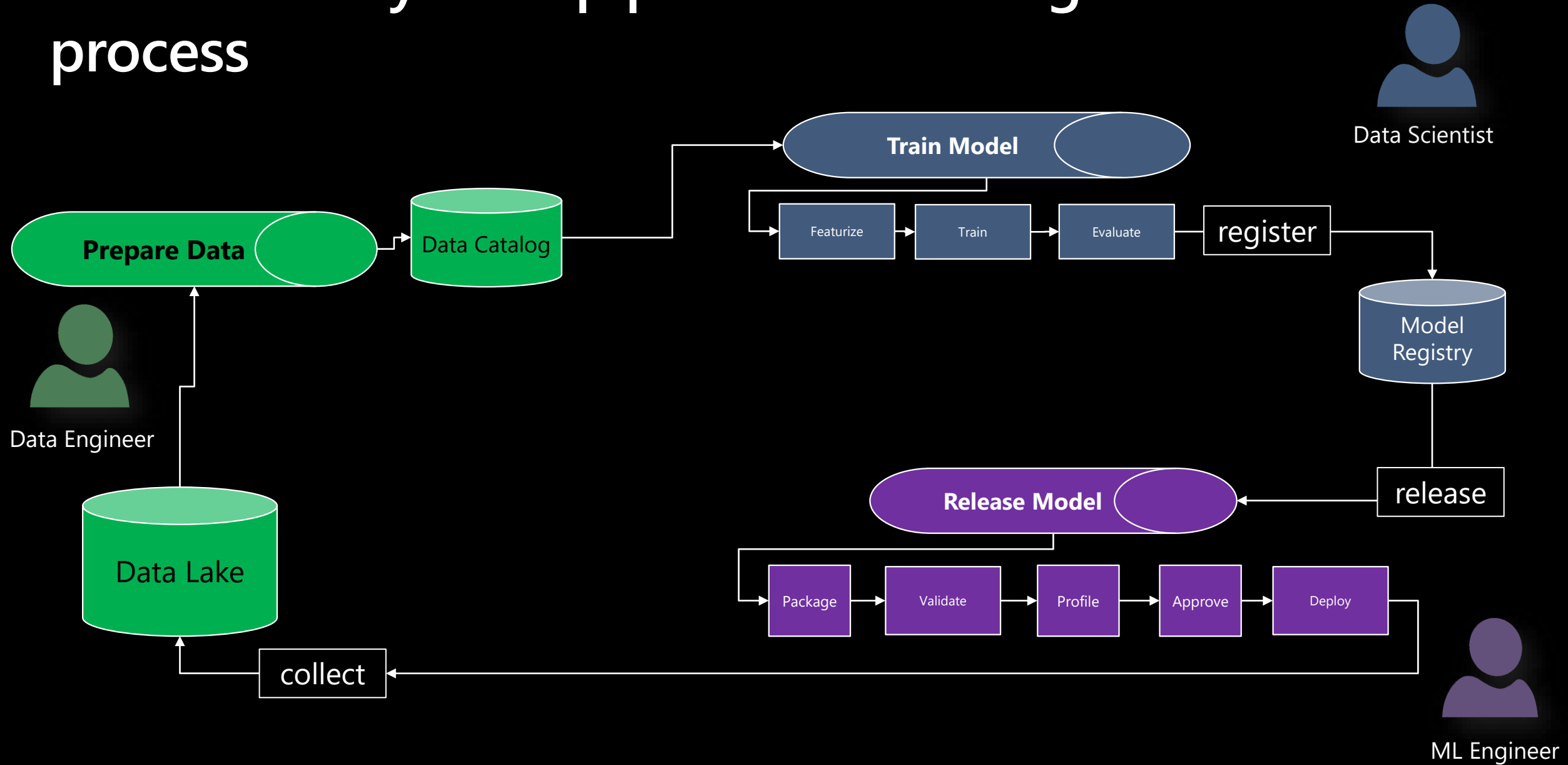
Azure Data Lake  
Azure Data Factory  
Azure DataBricks  
Azure SQL

Azure DevOps  
GitHub  
Azure Kubernetes Service  
Azure IoT Edge  
Azure Monitor



ML Engineer

# There is rarely “one pipeline” to manage the E2E process



# MLOps – Process Maturity Model

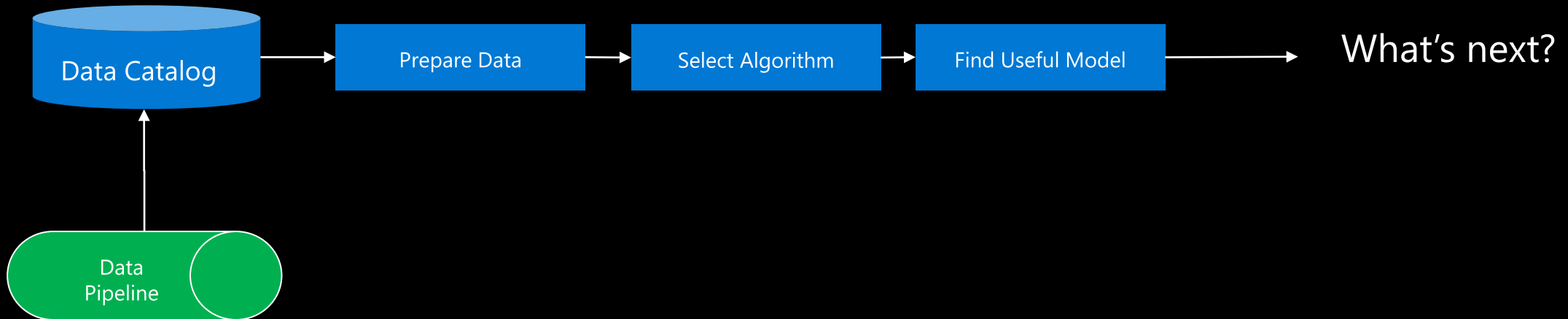
Maturity Level	People	Model Creation	Model Release	Application Integration	Technology
Level 1 - No MLOps	<ul style="list-style-type: none"> <li>Data Scientists - silo'd, not in regular comms with larger team</li> <li>Data Engineers - silo'd (if exists), not in regular comms with larger team</li> <li>Software Engineers - Silo'd, receive model "over the wall"</li> </ul>	<ul style="list-style-type: none"> <li>Data pipeline gathers data automatically</li> <li>Compute may or may not be managed</li> <li>Experiments are not predictably tracked</li> <li>End result may be a single file manually handed off (model), with inputs/outputs</li> </ul>	<ul style="list-style-type: none"> <li>Manual process</li> <li>Scoring script may be manually created well after experiments, likely version controlled</li> <li>Is handed off to Software Engineers</li> </ul>	<ul style="list-style-type: none"> <li>Basic integration tests exist for the model</li> <li>Heavily reliant on Data Scientist expertise to implement model</li> <li>Releases are automated</li> <li>Application code has unit tests</li> </ul>	<ul style="list-style-type: none"> <li>Automated Builds</li> <li>Automated Tests for Application code</li> <li>Manual model training</li> <li>No centralized tracking of model performance</li> </ul>
Level 2 - Automated Training	<ul style="list-style-type: none"> <li>Data Scientists - Working directly with Data Engineers to convert experimentation code into repeatable scripts/jobs</li> <li>Data Engineers - Working with Data Scientists</li> <li>Software Engineers - Silo'd, receive model "over the wall"</li> </ul>	<ul style="list-style-type: none"> <li>Data pipeline gathers data automatically</li> <li>Compute is managed</li> <li>Experiment results are tracked</li> <li>Both training code and resulting models are version controlled</li> </ul>	<ul style="list-style-type: none"> <li>Manual Release</li> <li>Scoring Script is version controlled with tests</li> <li>Release is managed by Software engineering team</li> </ul>	<ul style="list-style-type: none"> <li>Basic integration tests exist for the model</li> <li>Heavily reliant on Data Scientist expertise to implement model</li> <li>Application code has unit tests</li> </ul>	<ul style="list-style-type: none"> <li>Automated Builds</li> <li>Automated Tests for Application code</li> <li>Automated model training</li> <li>Centralized tracking of model training performance</li> <li>Model Management</li> </ul>
Level 3 - Automated Model Deployment	<ul style="list-style-type: none"> <li>Data Scientists - Working directly with Data Engineers to convert experimentation code into repeatable scripts/jobs</li> <li>Data Engineers - Working with Data Scientists and Software Engineers to manage inputs/outputs</li> <li>Software Engineers - Working with Data Engineers to automate model integration into application code</li> </ul>	<ul style="list-style-type: none"> <li>Data pipeline gathers data automatically</li> <li>Compute is managed</li> <li>Experiment results are tracked</li> <li>Both training code and resulting models are version controlled</li> </ul>	<ul style="list-style-type: none"> <li>Automatic Release</li> <li>Scoring Script is version controlled with tests</li> <li>Release is managed by CI/CD pipeline</li> </ul>	<ul style="list-style-type: none"> <li>Unit and Integration tests for each model release</li> <li>Less reliant on Data Scientist expertise to implement model</li> <li>Application code has unit/integration tests</li> </ul>	<ul style="list-style-type: none"> <li>Automated Builds</li> <li>Integrated A/B testing of model performance for deployment</li> <li>Automated Tests for All code</li> <li>Automated model training</li> <li>Centralized tracking of model training performance</li> <li>Model Management</li> </ul>
Level 4 - Automated Retraining (full MLOps)	<ul style="list-style-type: none"> <li>Data Scientists - Working directly with Data Engineers to convert experimentation code into repeatable scripts/jobs. Working with Software Engineers to identify markers for retraining</li> <li>Data Engineers - Working with Data Scientists and Software Engineers to manage inputs/outputs</li> <li>Software Engineers - Working with Data Engineers to automate model integration into application code. Implementing metrics gathering post-deployment</li> </ul>	<ul style="list-style-type: none"> <li>Data pipeline gathers data automatically</li> <li>Retraining triggered automatically based on production metrics</li> <li>Compute is managed</li> <li>Experiment results are tracked</li> <li>Both training code and resulting models are version controlled</li> </ul>	<ul style="list-style-type: none"> <li>Automatic Release</li> <li>Scoring Script is version controlled with tests</li> <li>Release is managed by CI/CD pipeline</li> </ul>	<ul style="list-style-type: none"> <li>Unit and Integration tests for each model release</li> <li>Less reliant on Data Scientist expertise to implement model</li> <li>Application code has unit/integration tests</li> </ul>	<ul style="list-style-type: none"> <li>Automated Builds</li> <li>Integrated A/B testing of model performance for deployment</li> <li>Automated Tests for All code</li> <li>Automated model training and testing</li> <li>Centralized tracking of model training performance</li> <li>Model Management</li> <li>Verbose, centralized metrics from deployed model</li> </ul>

# Level 1 – No MLOps

Interactive, exploratory, get to something useful.



Data Scientist

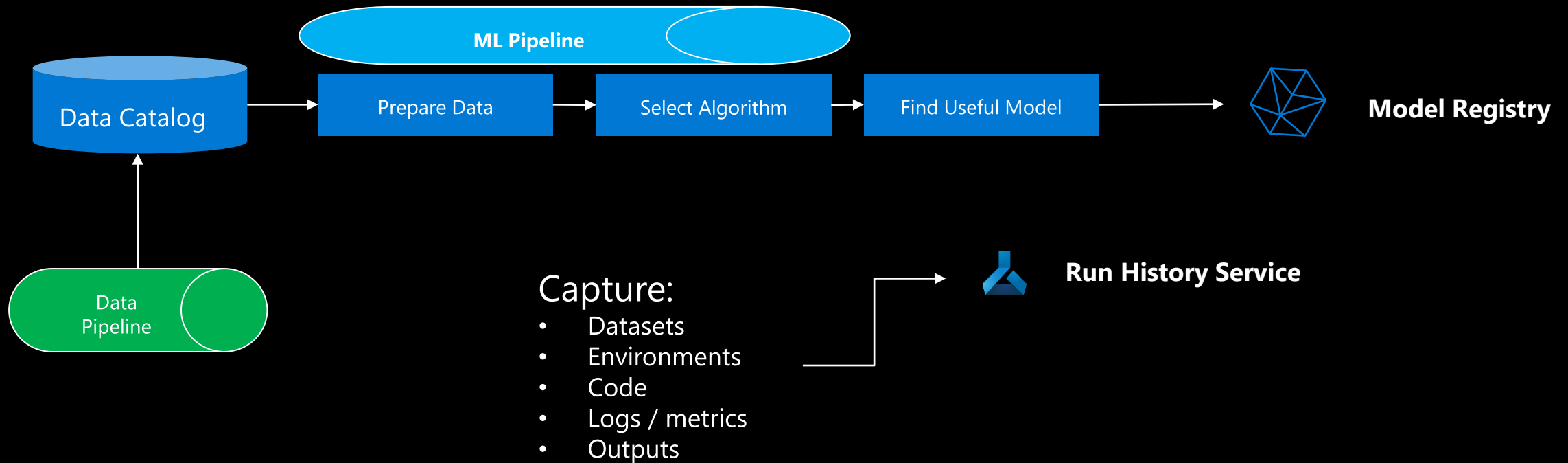


# Level 2 – Reproducible Model Training

Version code, data, ensure model can be recreated.



Data Scientist

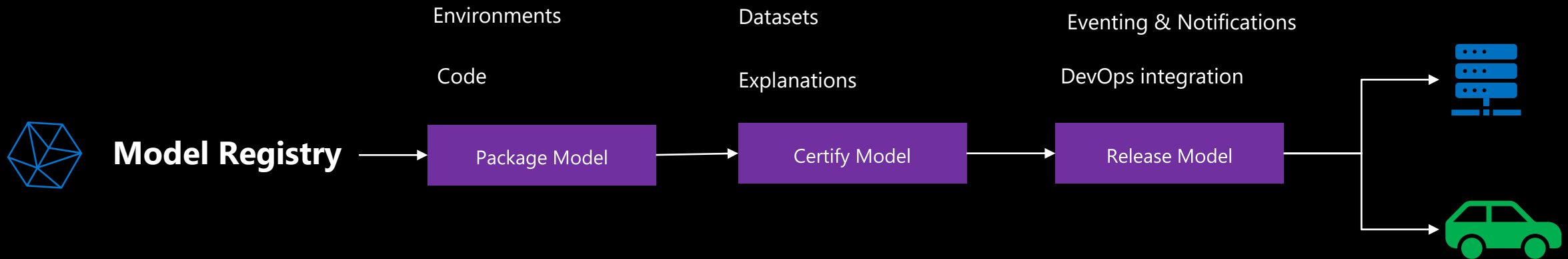


# Level 3 – Managed Model Operationalization

Package, certify, deploy

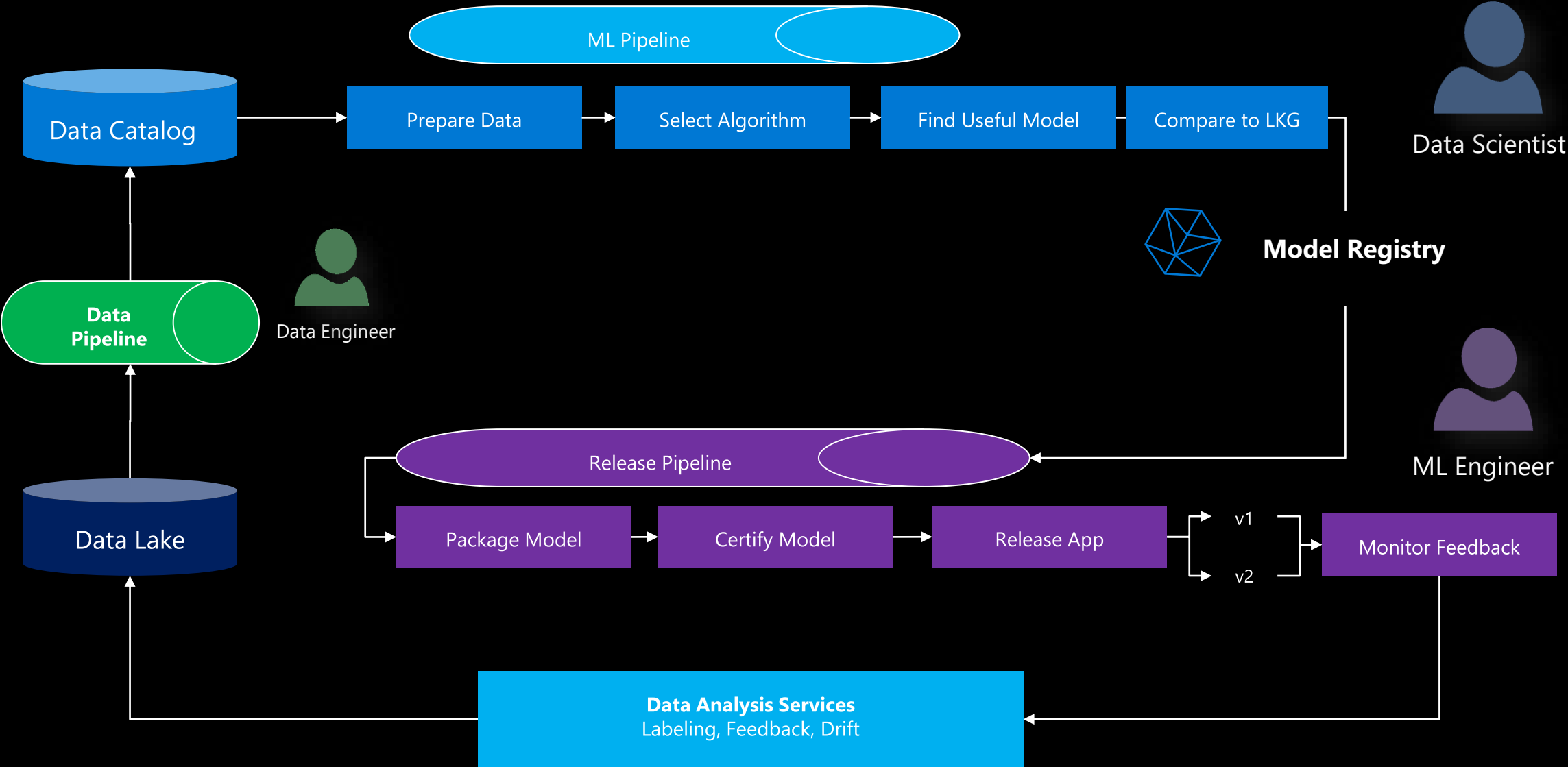


ML Engineer

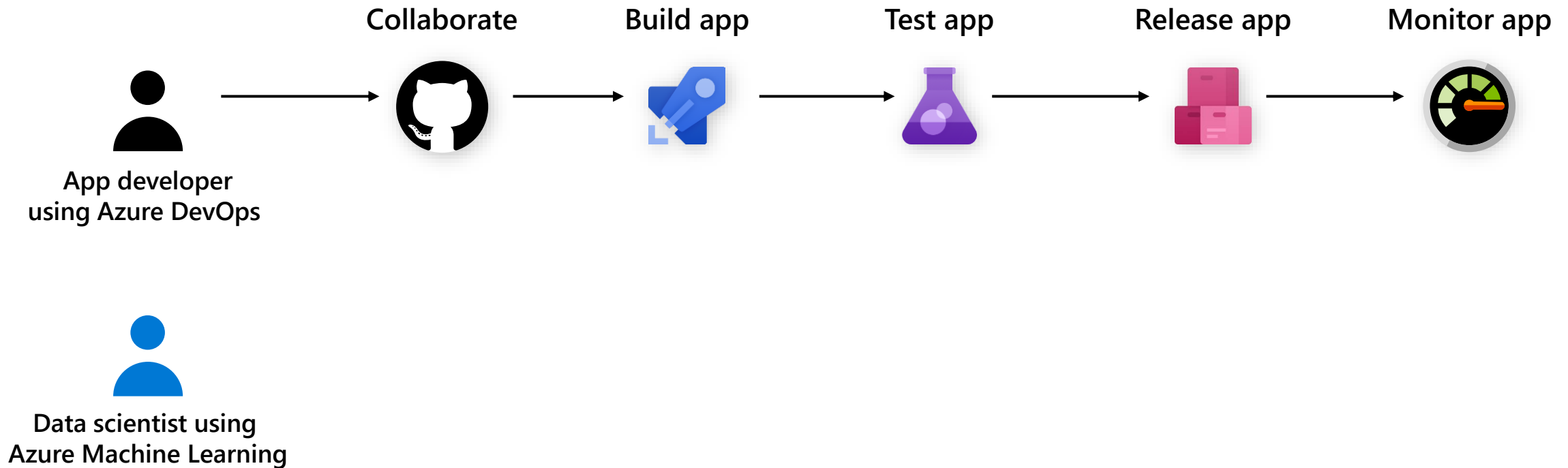




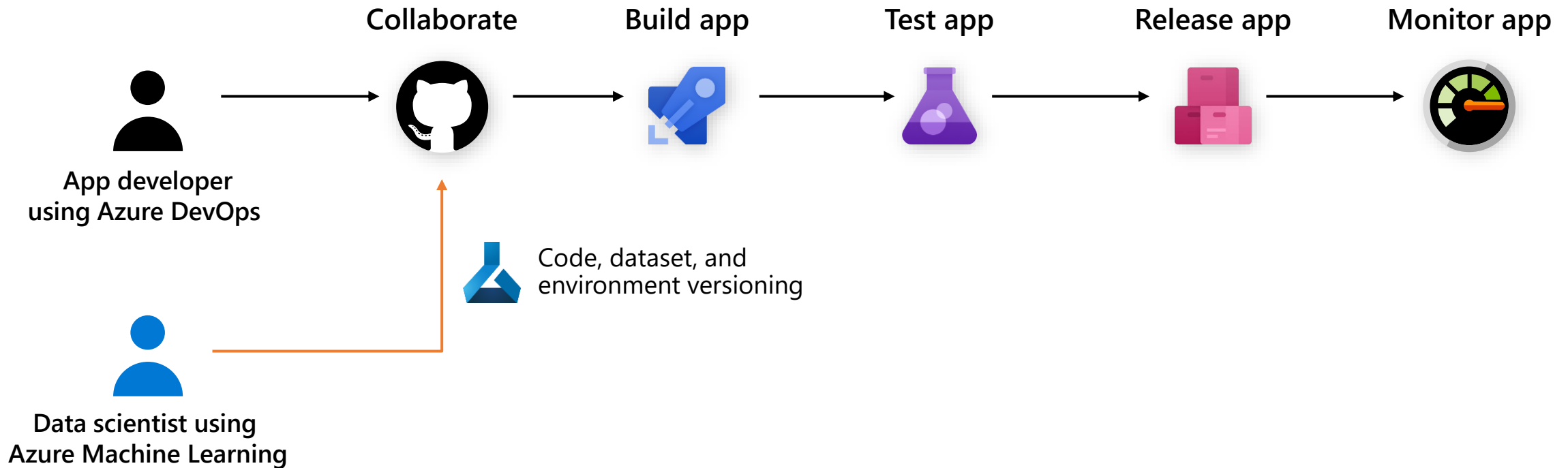
# Level 4 – Automated E2E ML Lifecycle



# MLOps Workflow



# MLOps Workflow



Model reproducibility



Model validation

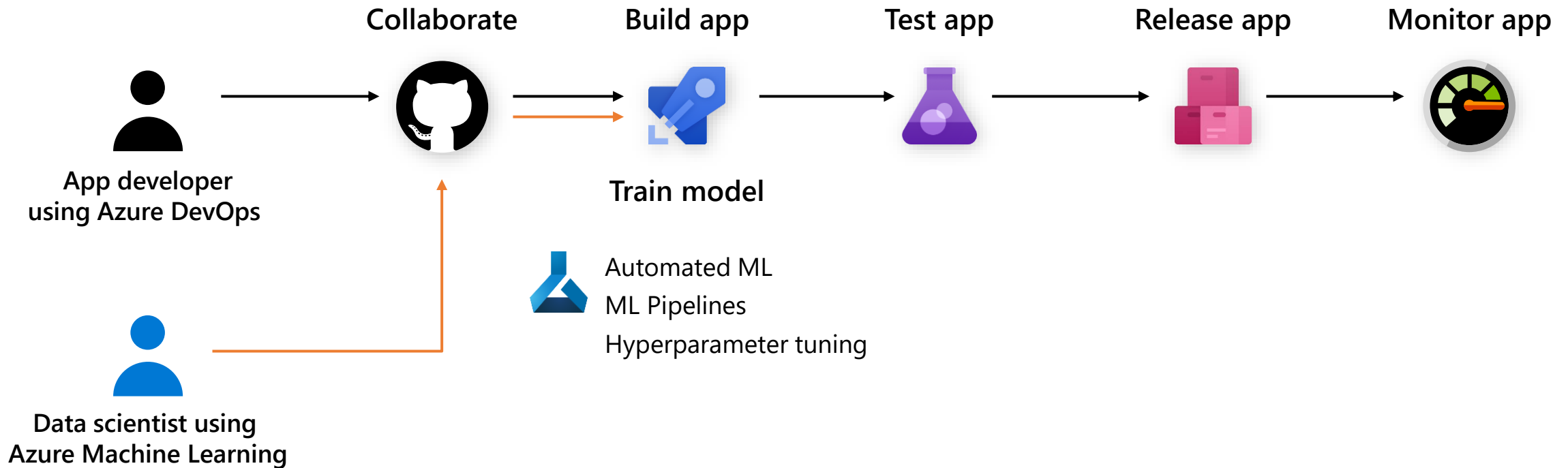


Model deployment



Model retraining

# MLOps Workflow



Model reproducibility



Model validation

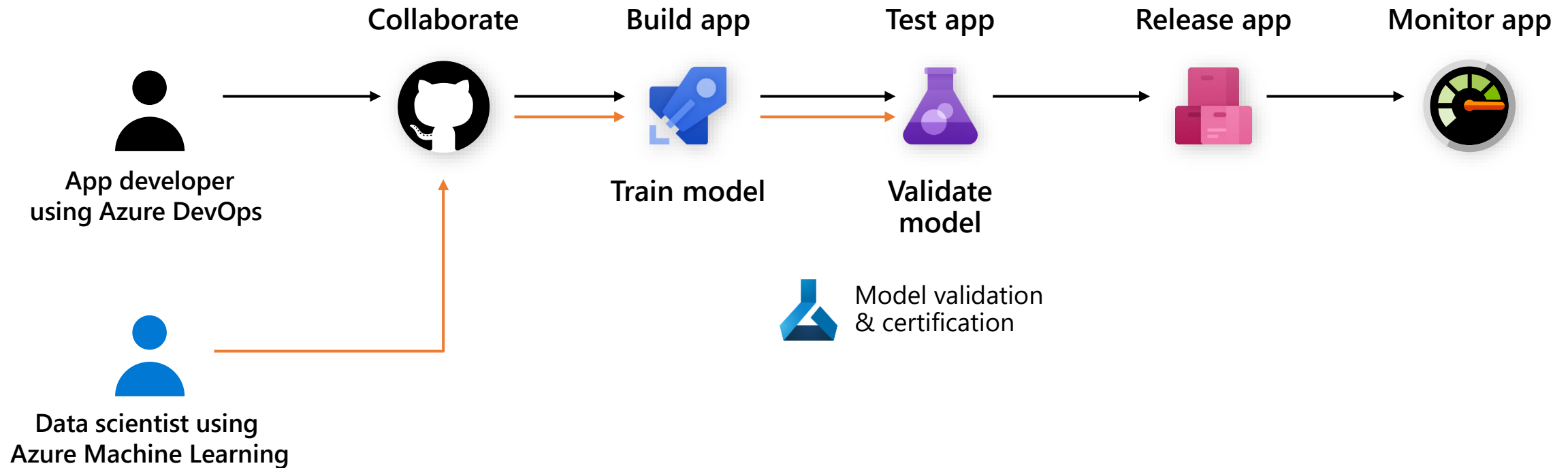


Model deployment



Model retraining

# MLOps Workflow



Model reproducibility



Model validation

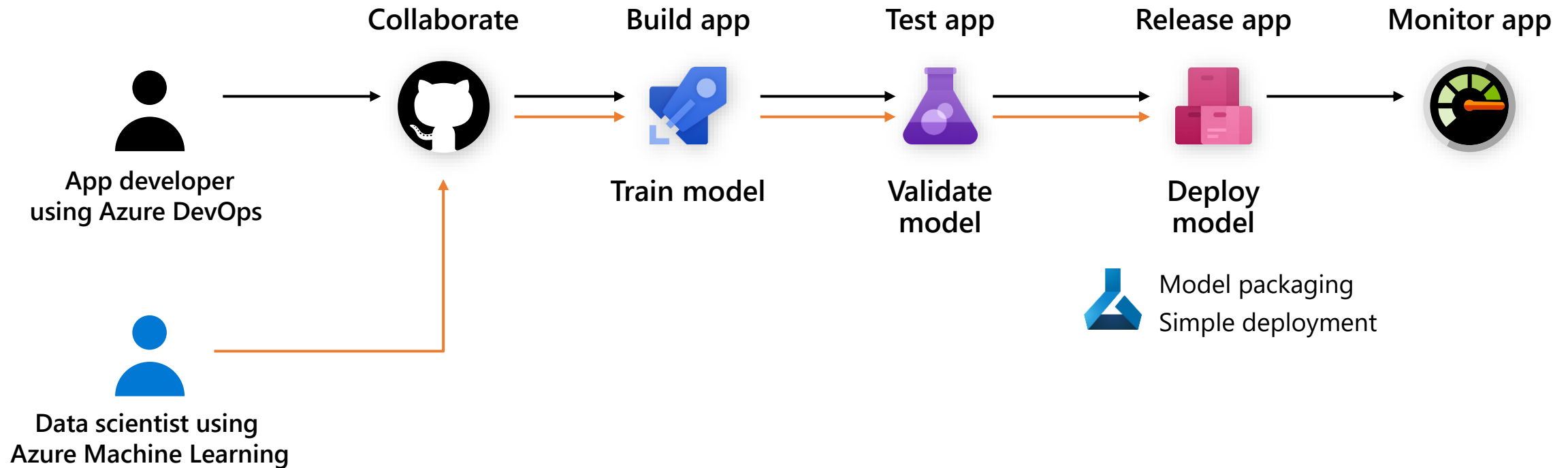


Model deployment



Model retraining

# MLOps Workflow



Model reproducibility



Model validation

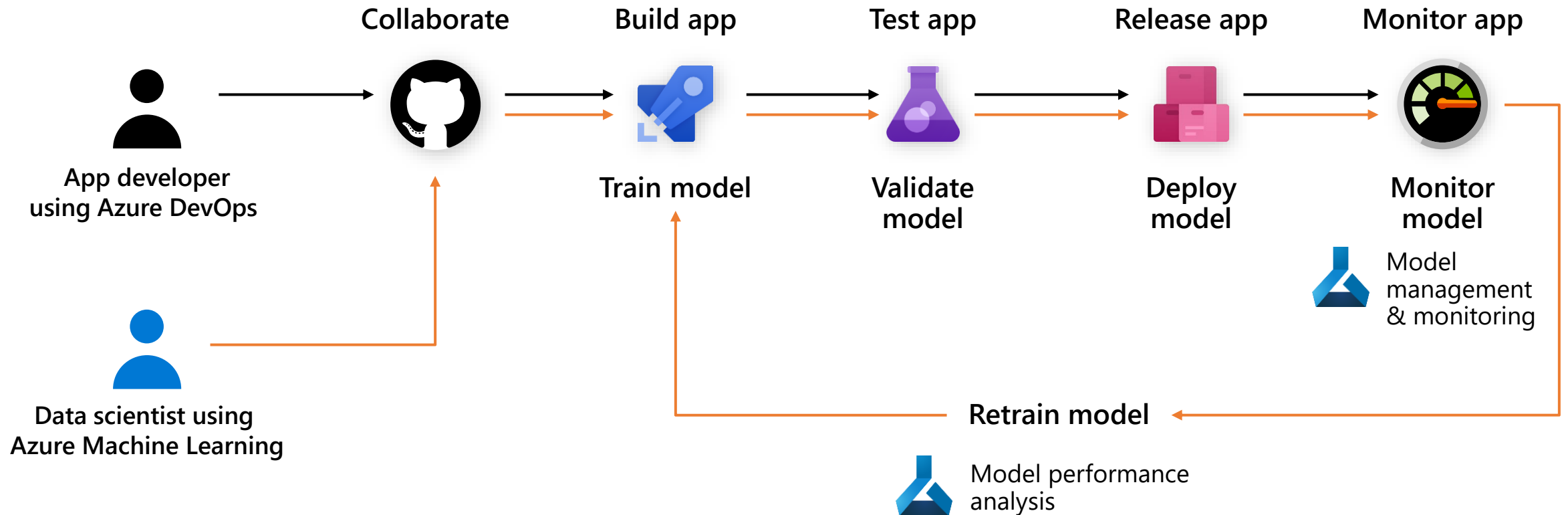


Model deployment

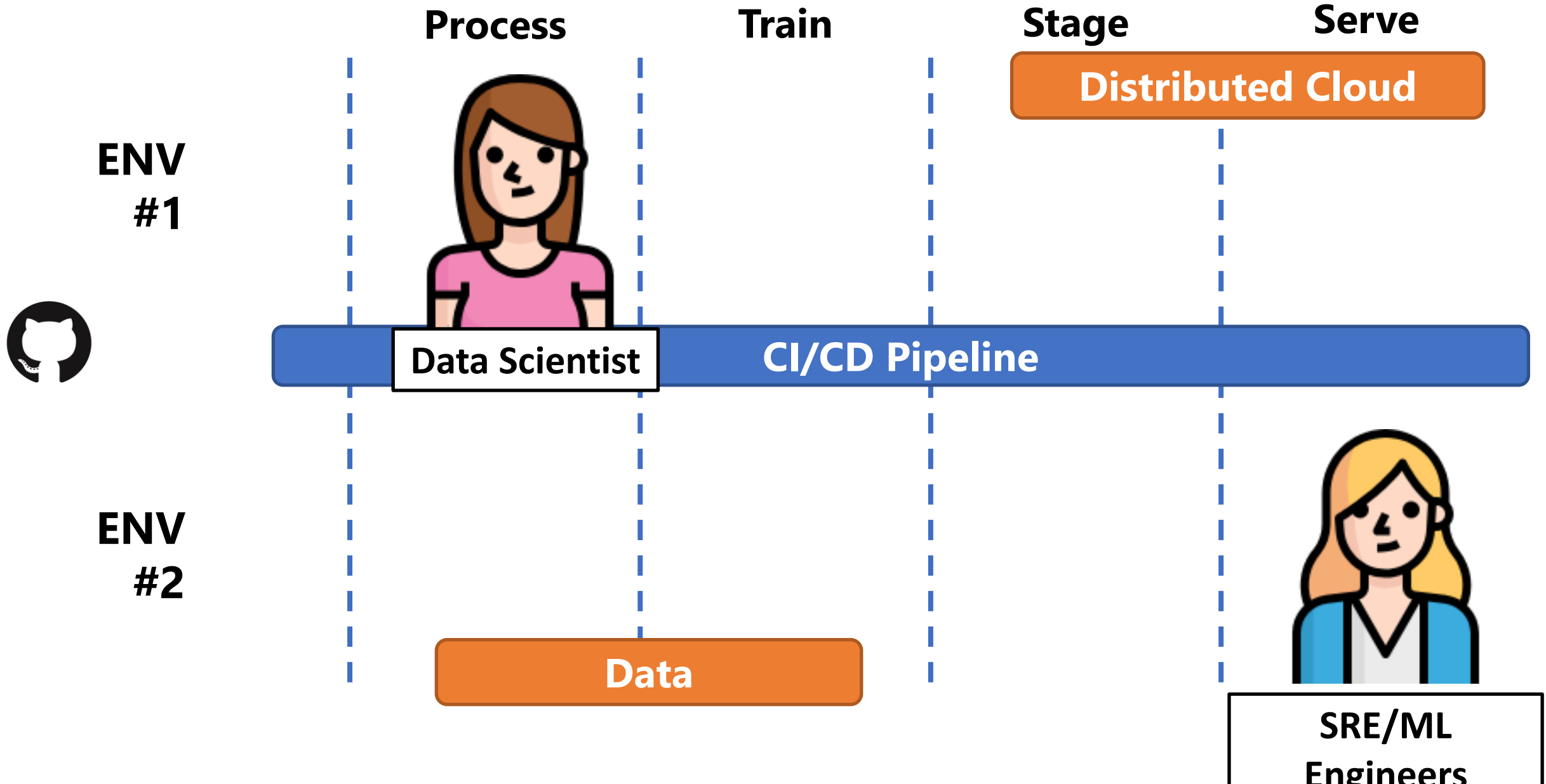


Model retraining

# MLOps Workflow

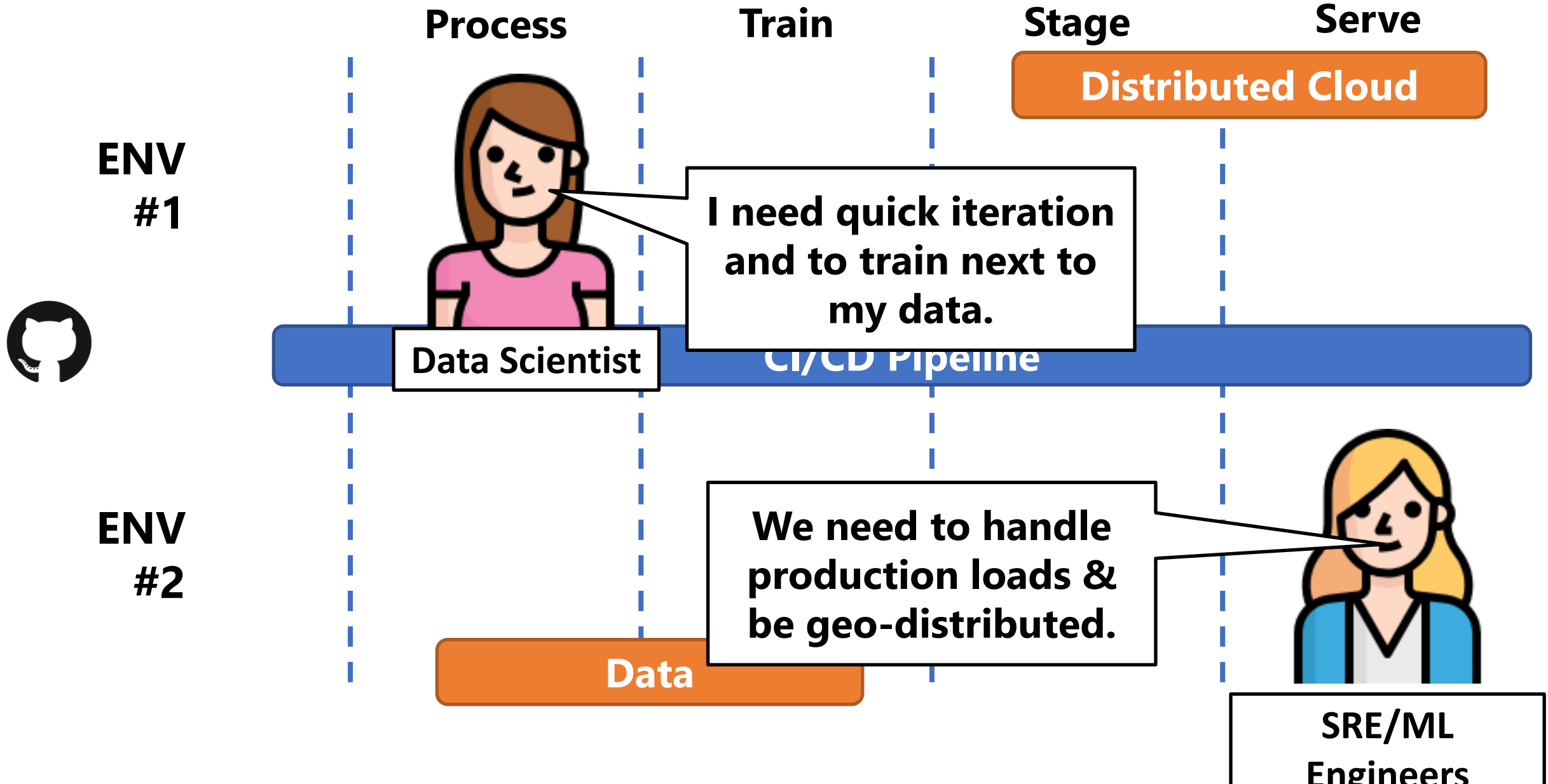


# Real World Multi-Cloud

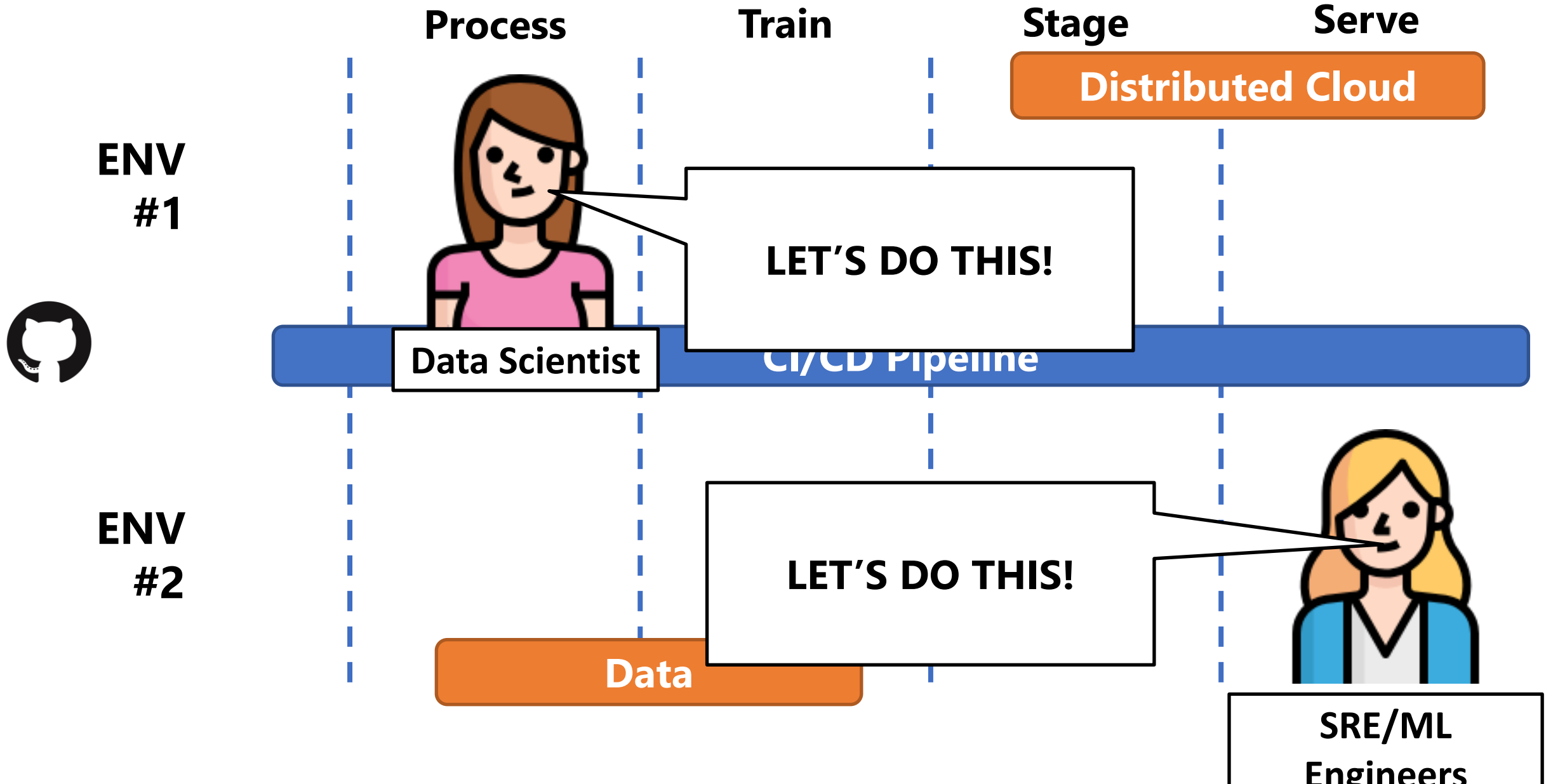




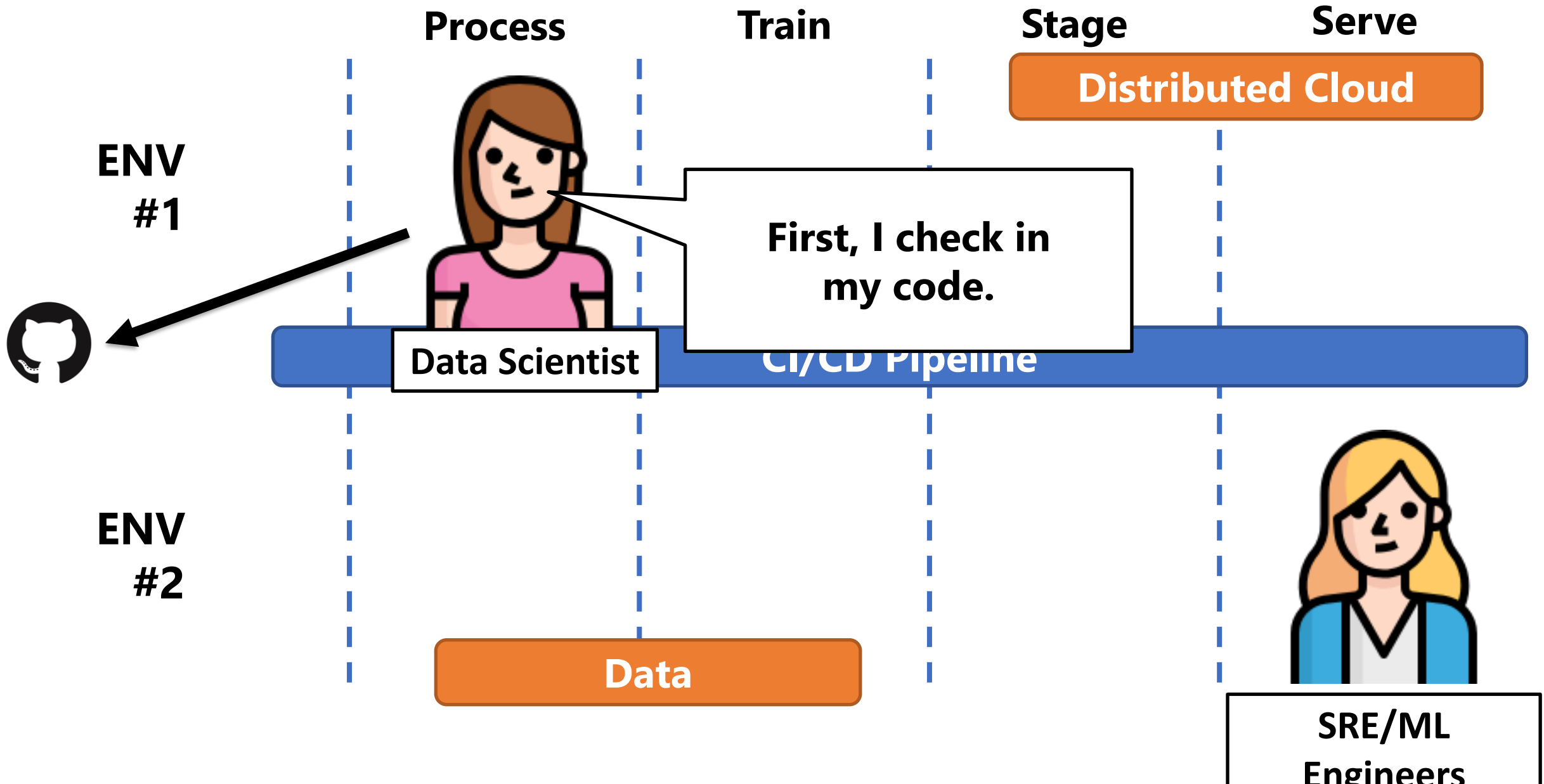
# Real World Multi-Cloud



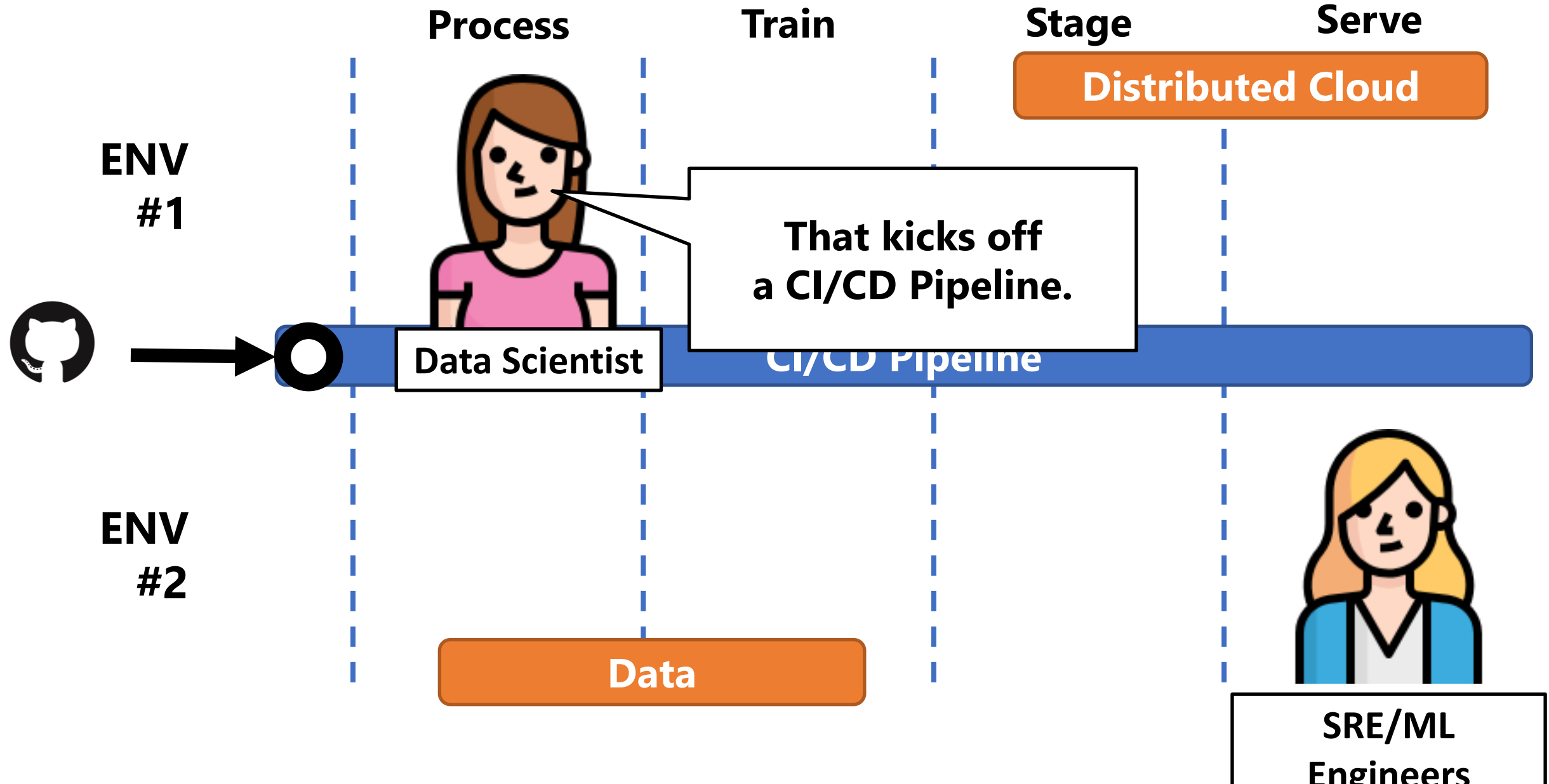
# Real World Multi-Cloud



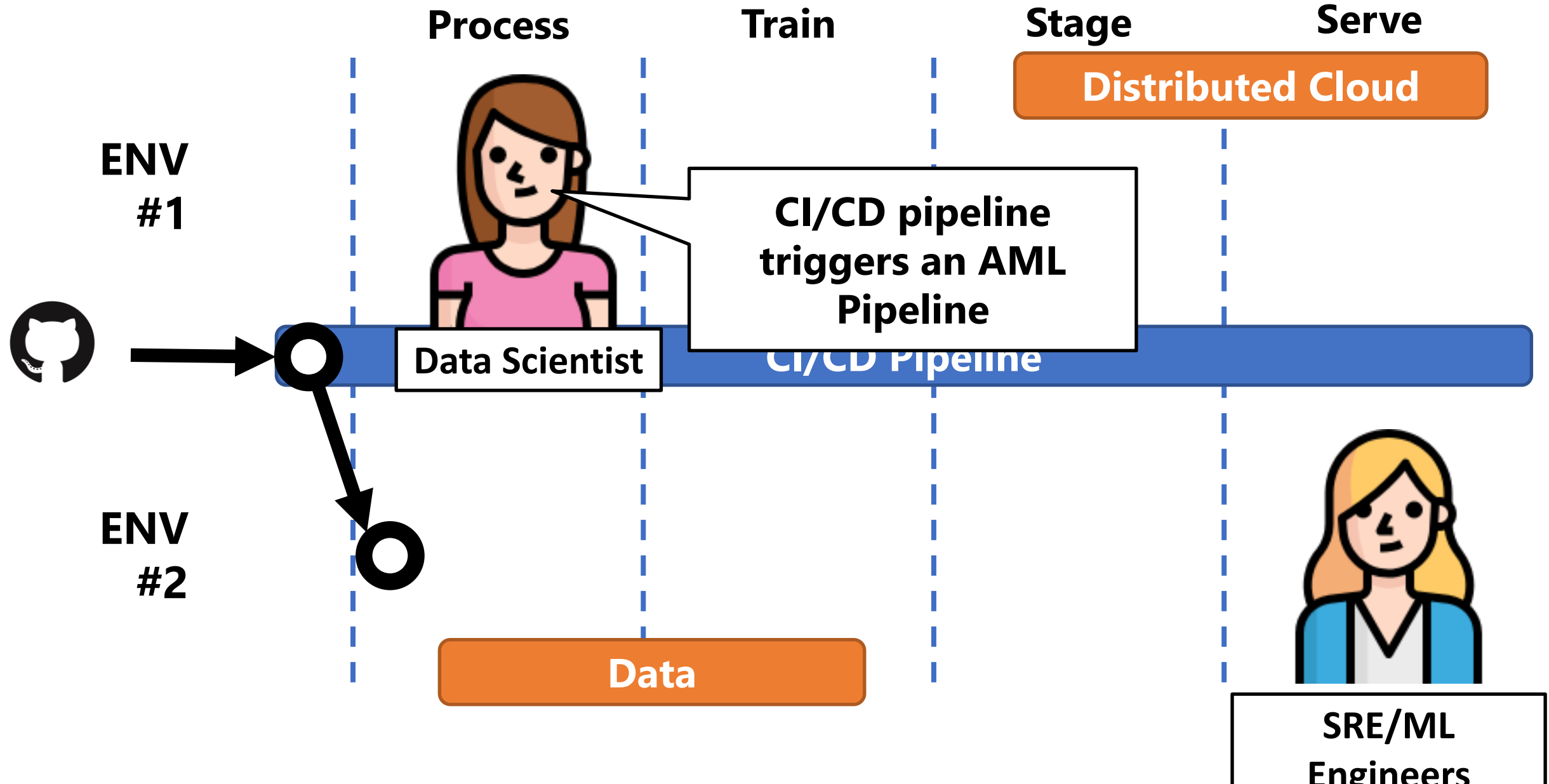
# Real World Multi-Cloud



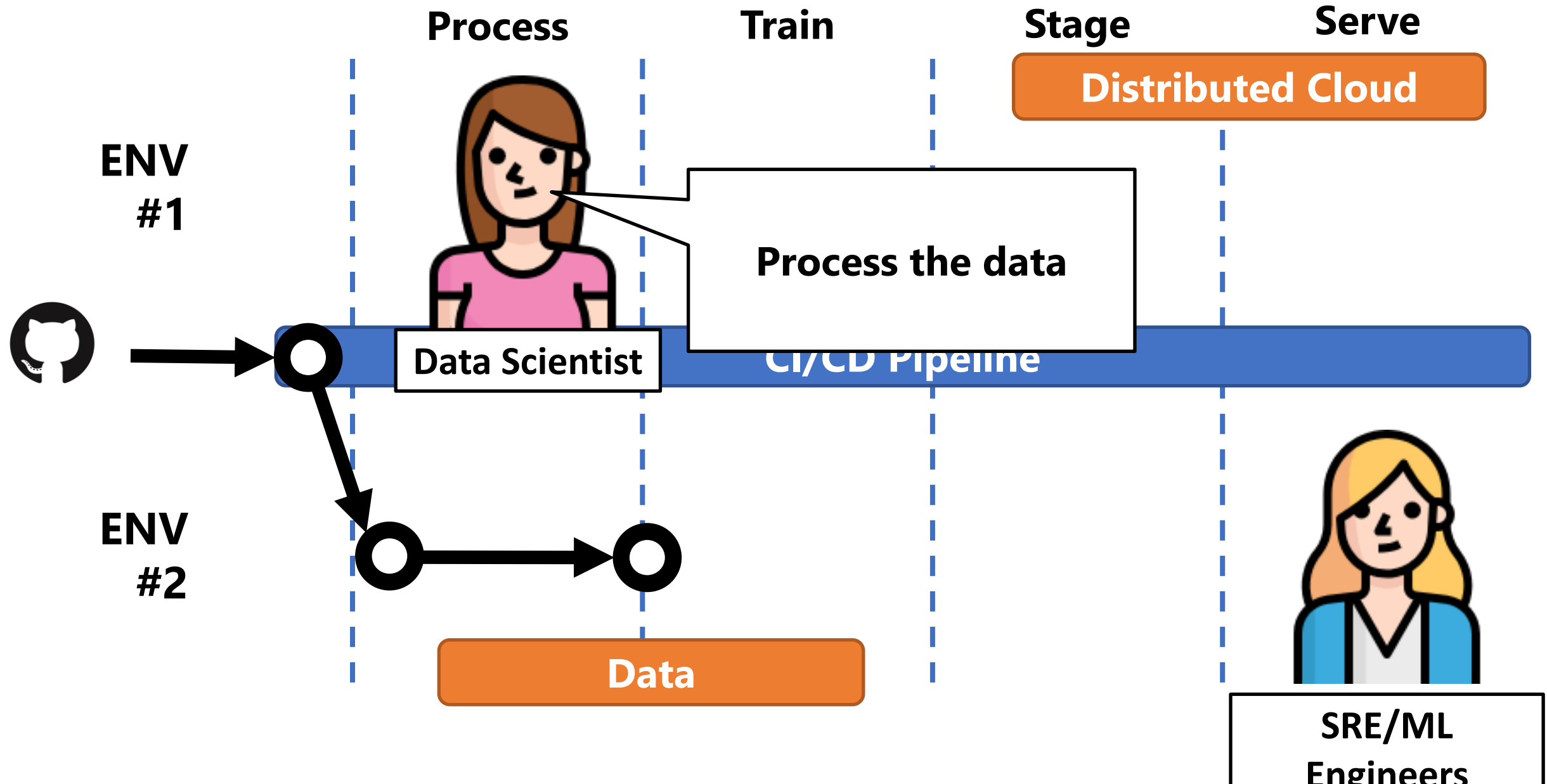
# Real World Multi-Cloud



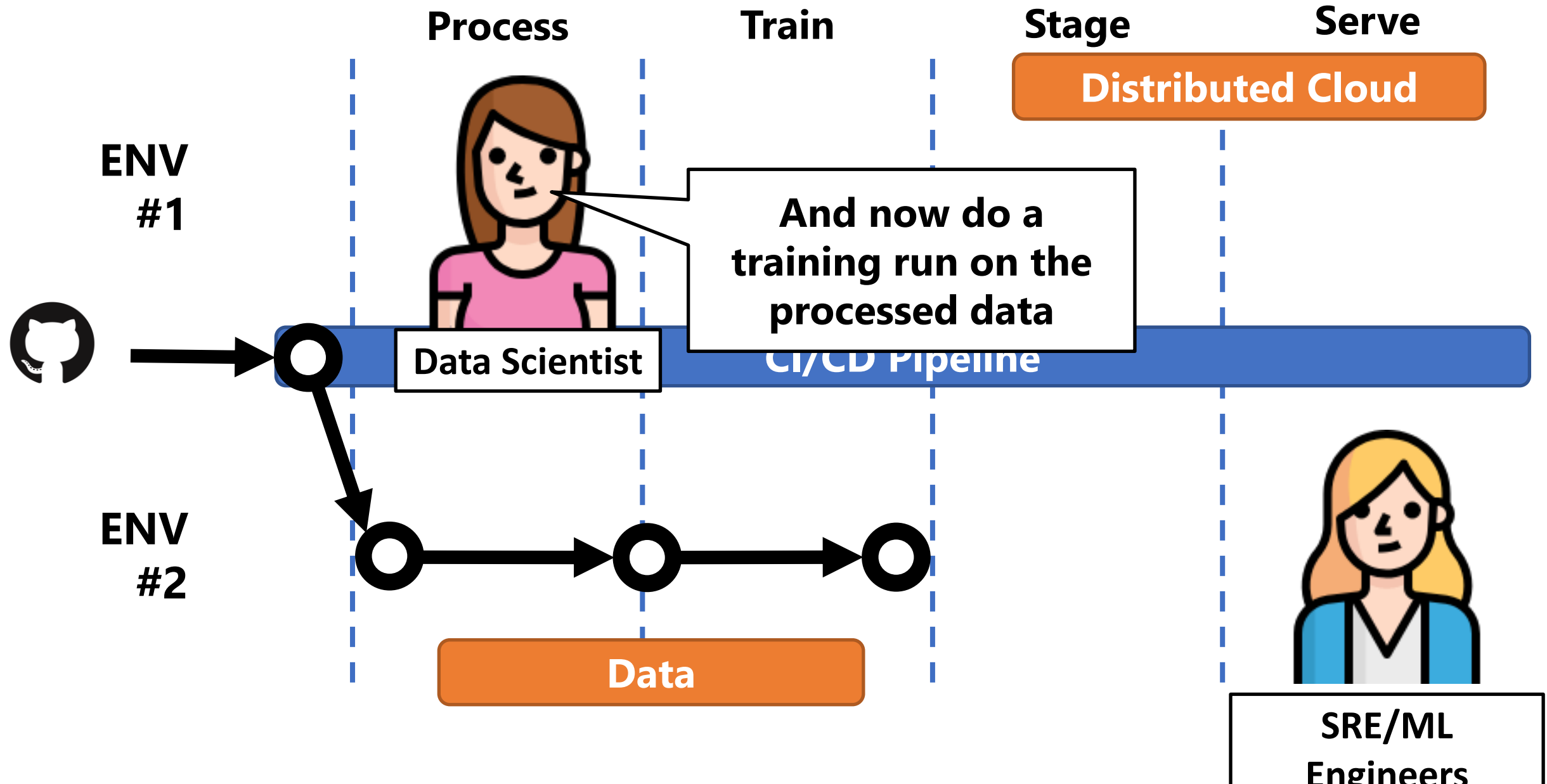
# Real World Multi-Cloud



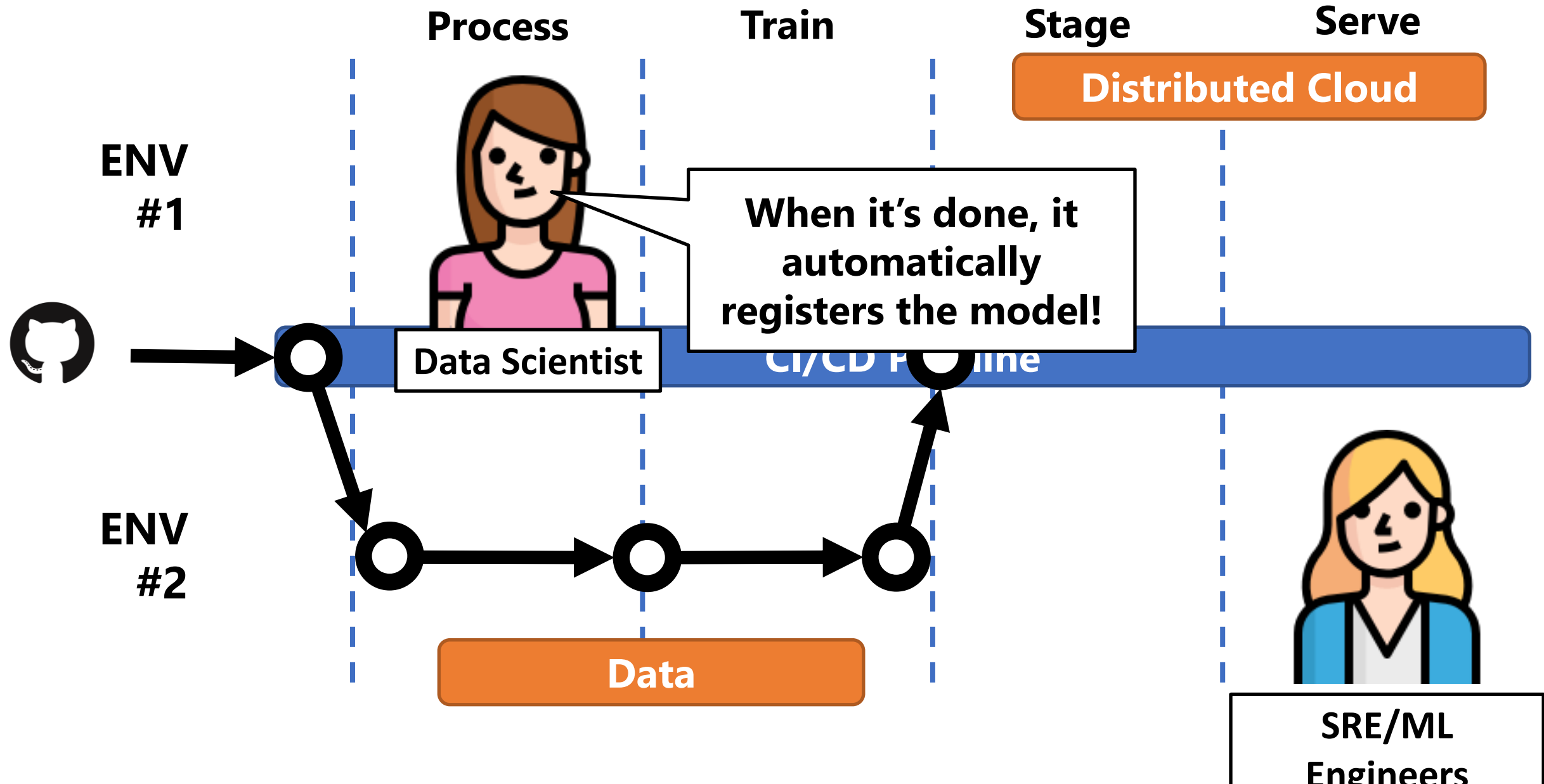
# Real World Multi-Cloud



# Real World Multi-Cloud

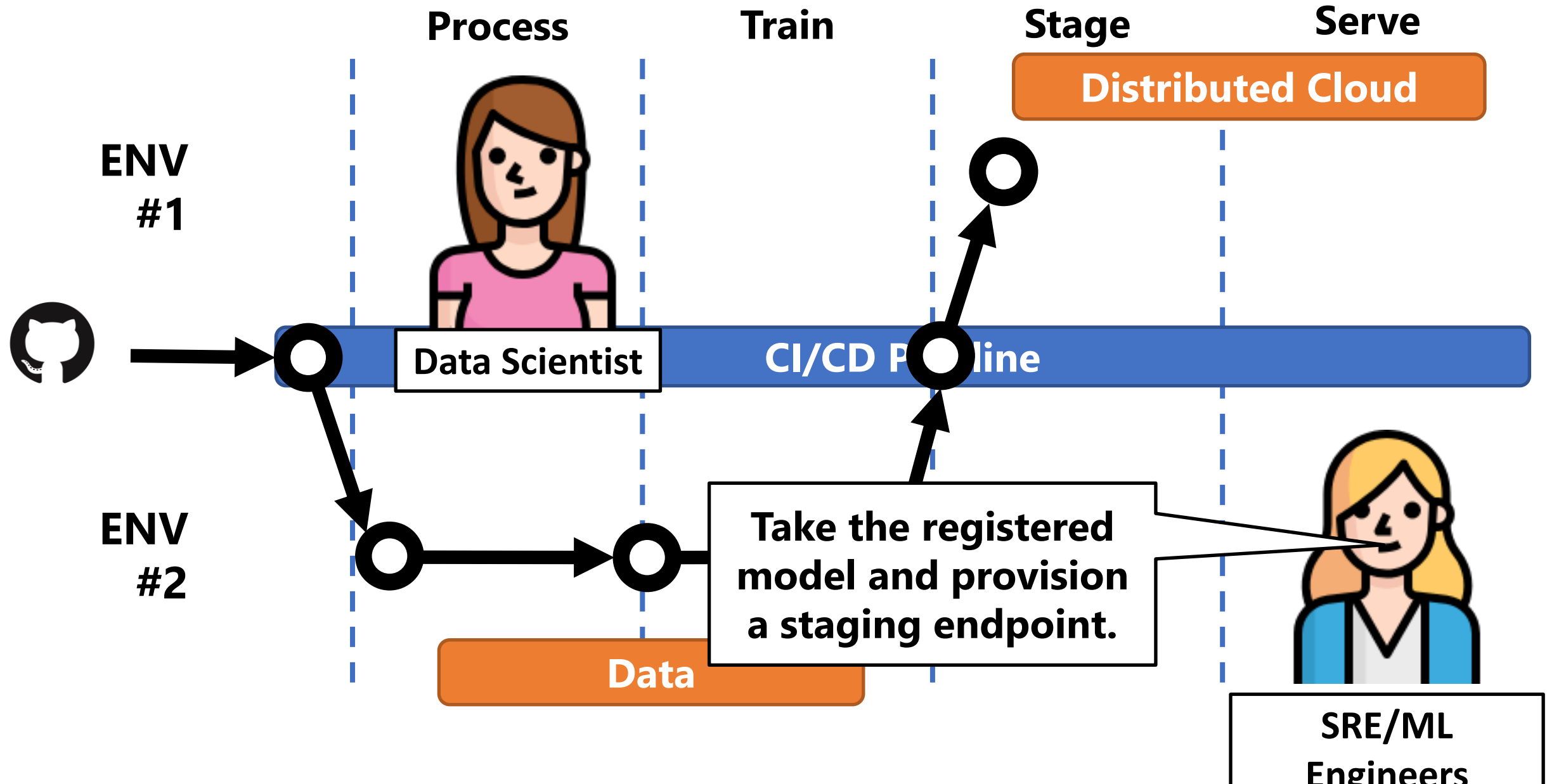


# Real World Multi-Cloud

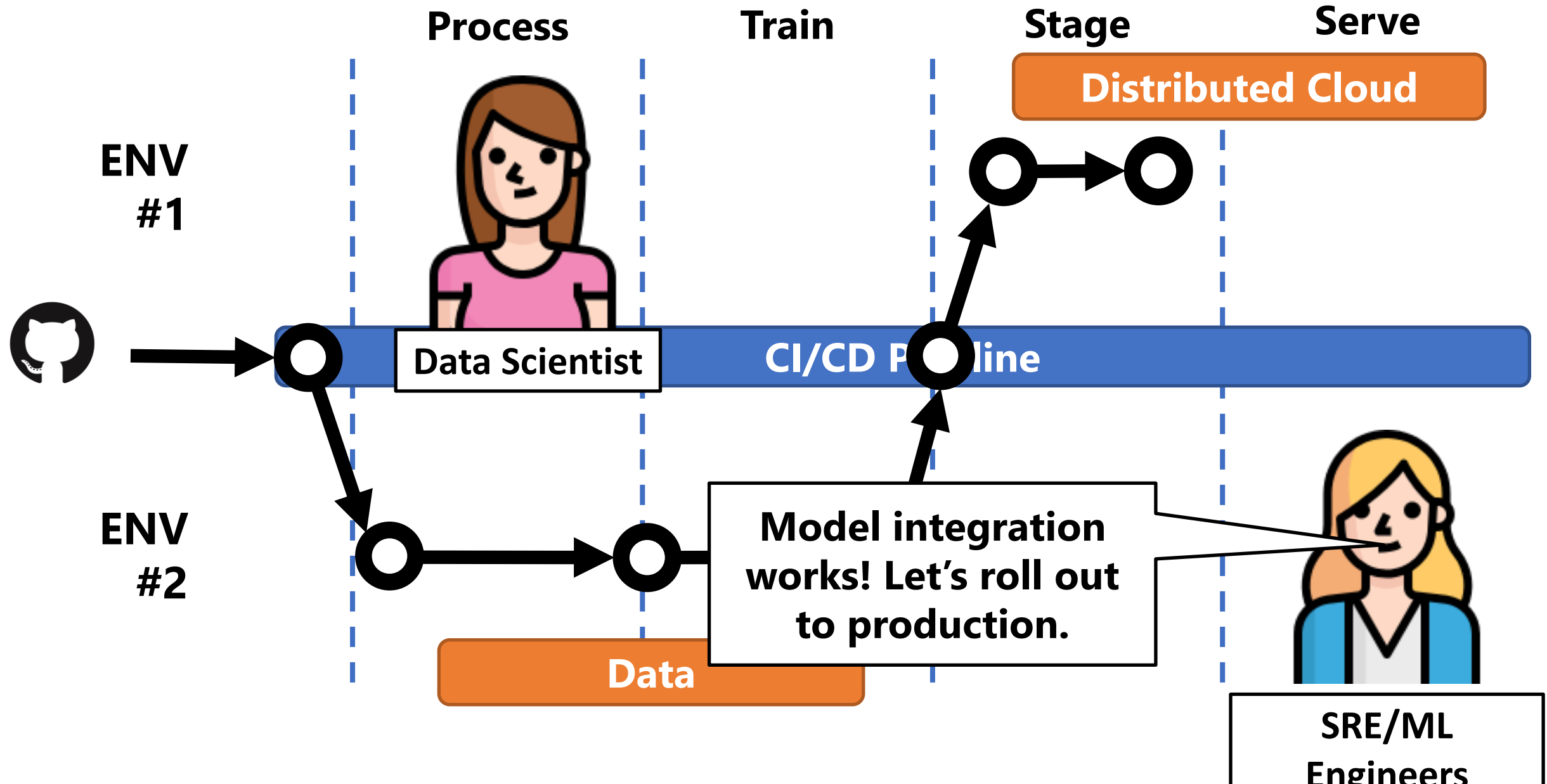




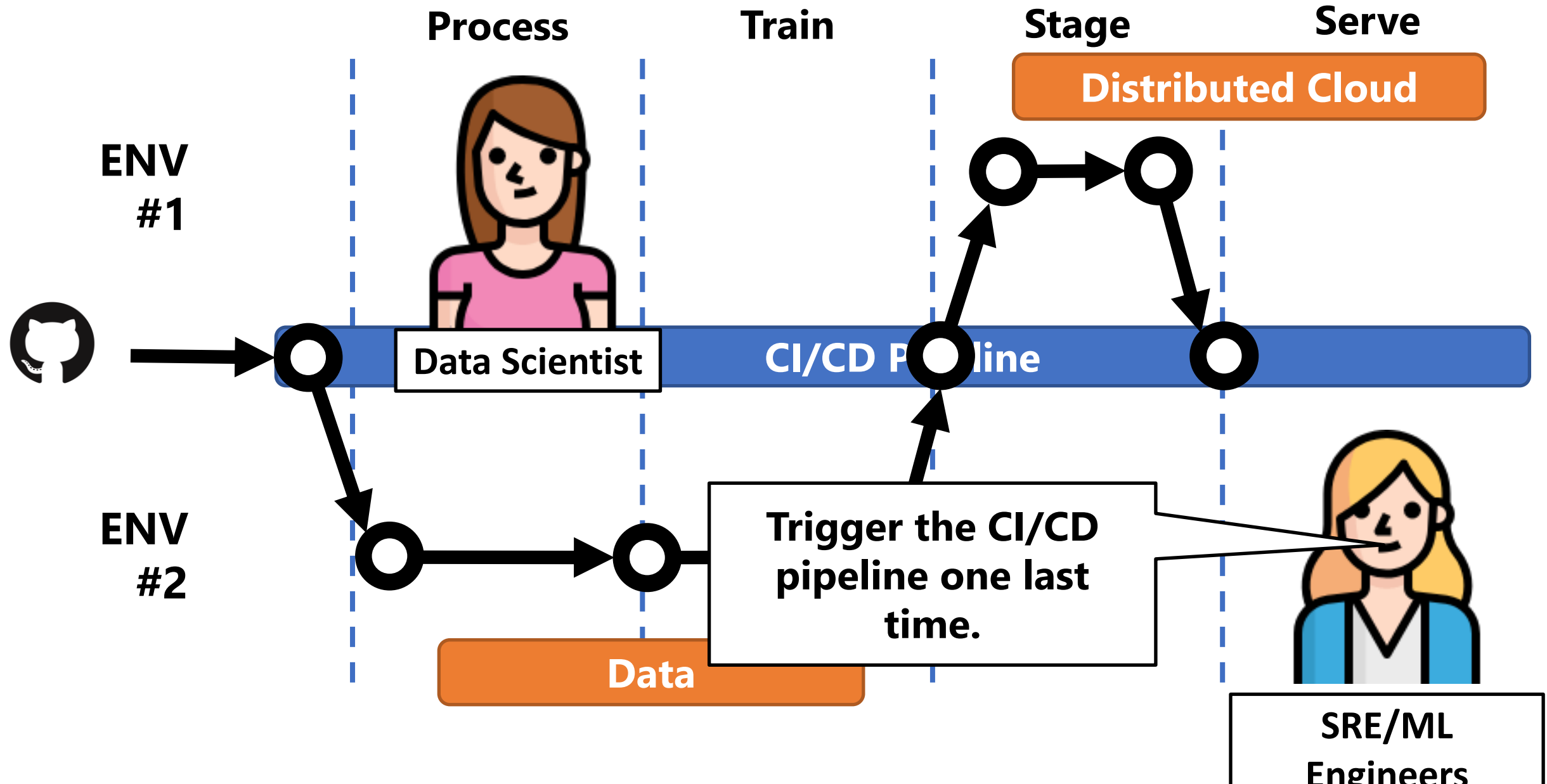
# Real World Multi-Cloud



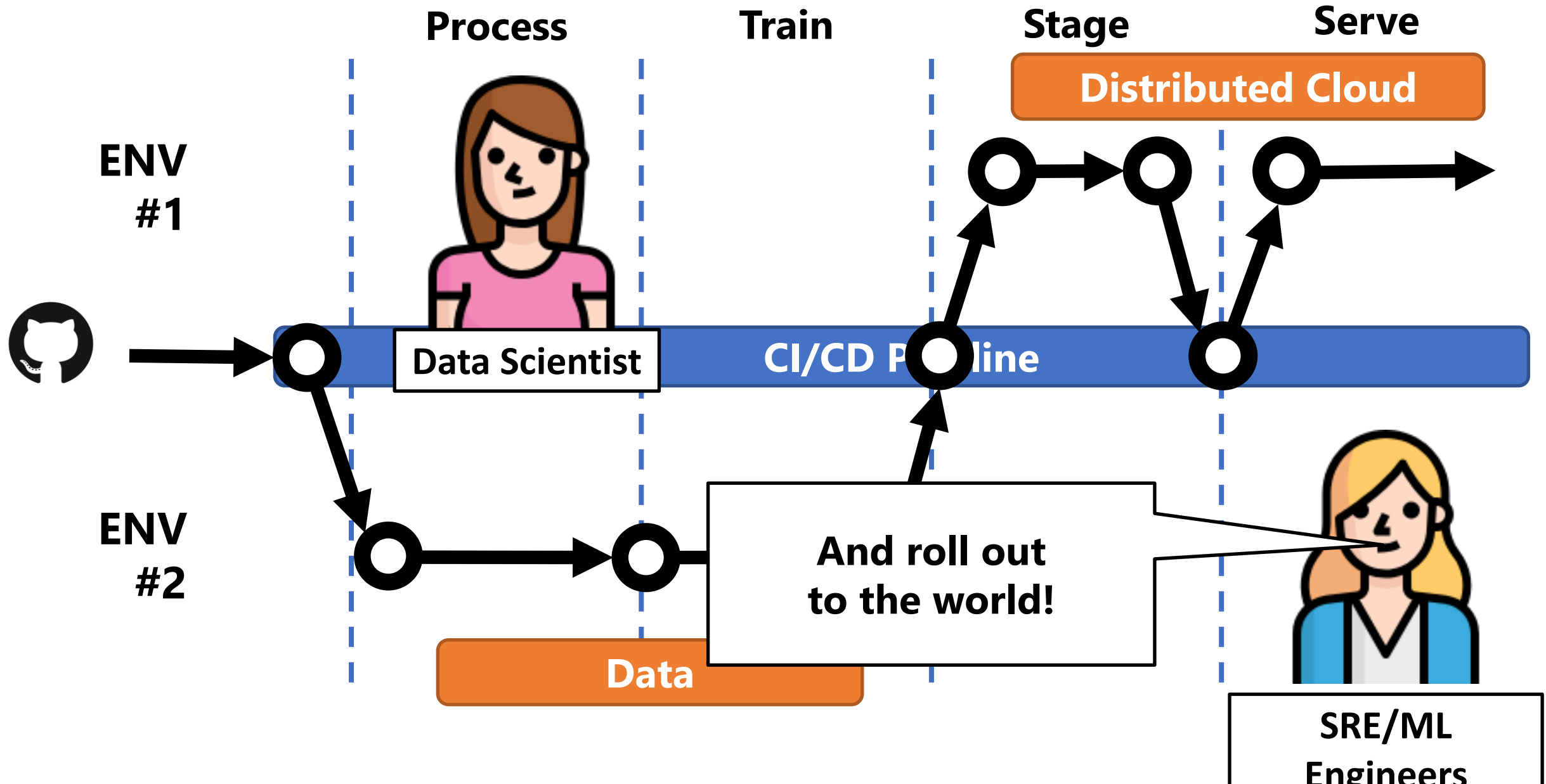
# Real World Multi-Cloud



# Real World Multi-Cloud



# Real World Multi-Cloud



# Real World Multi-Cloud

