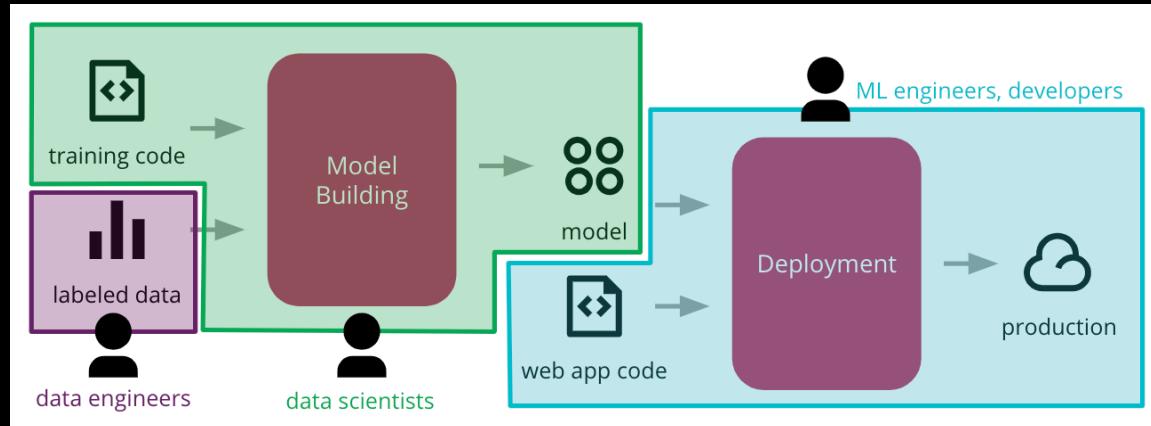


Continuous Delivery for ML

Jin Guo

Oct 27th, 2020



Silos -> delays and friction, quickly becomes stale and hard to update

Hard to reproduce and audit: different tools and workflow, hard to automate end-to-end.

Data pipeline

Model training pipeline

Deployment pipeline

Safely

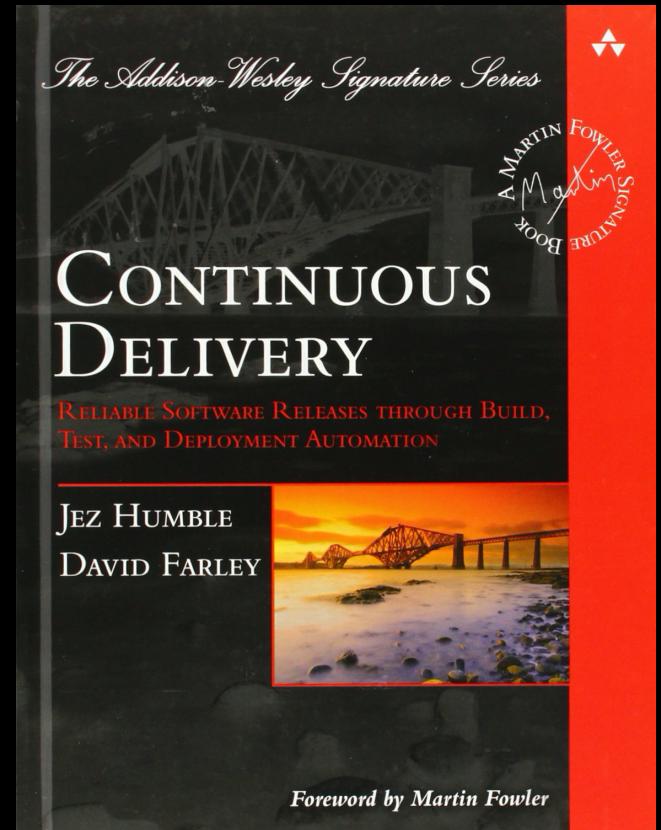
Quickly

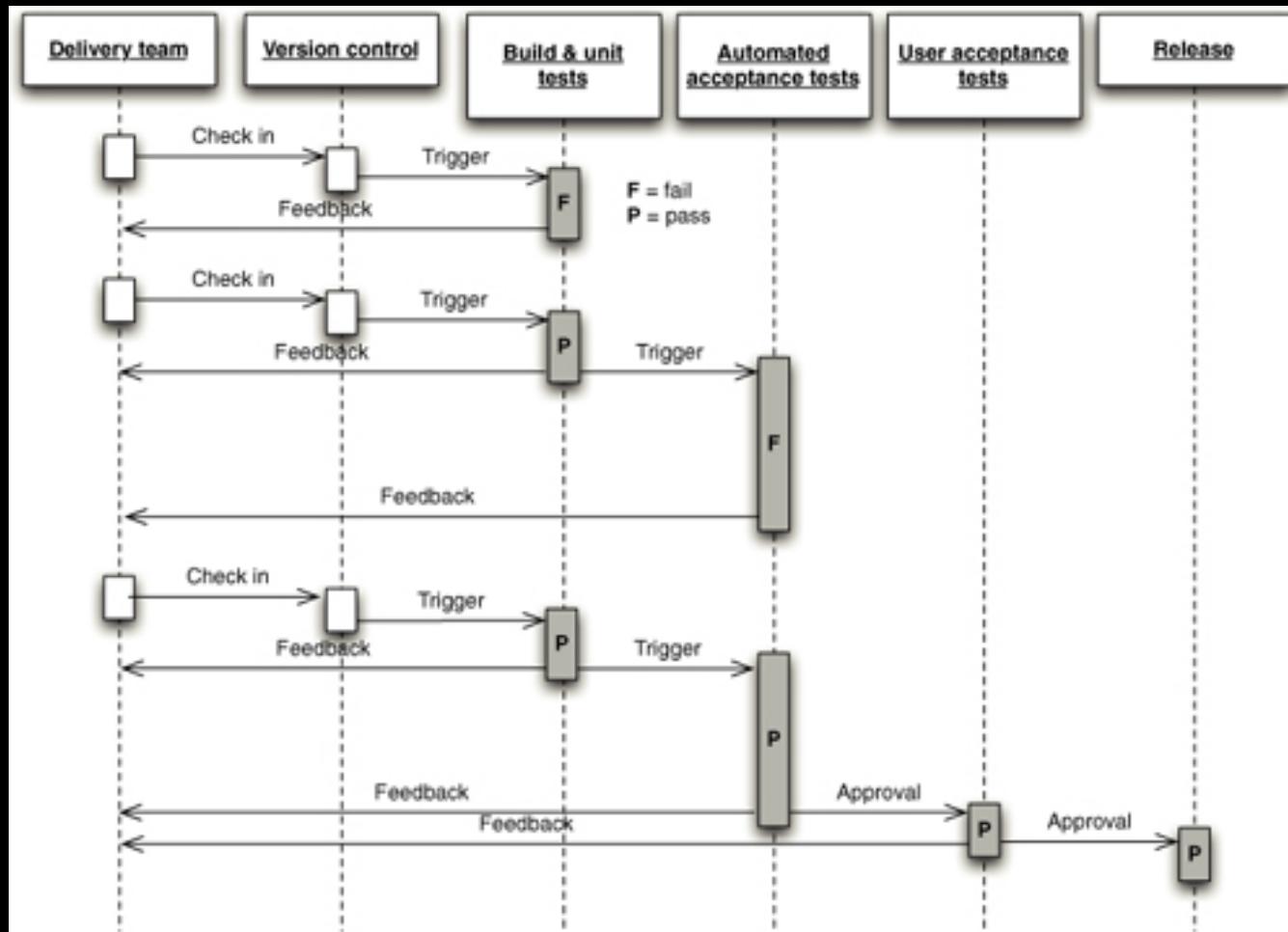
Sustainably

Continuous Delivery

A close, collaborative working relationship between development and deployment team

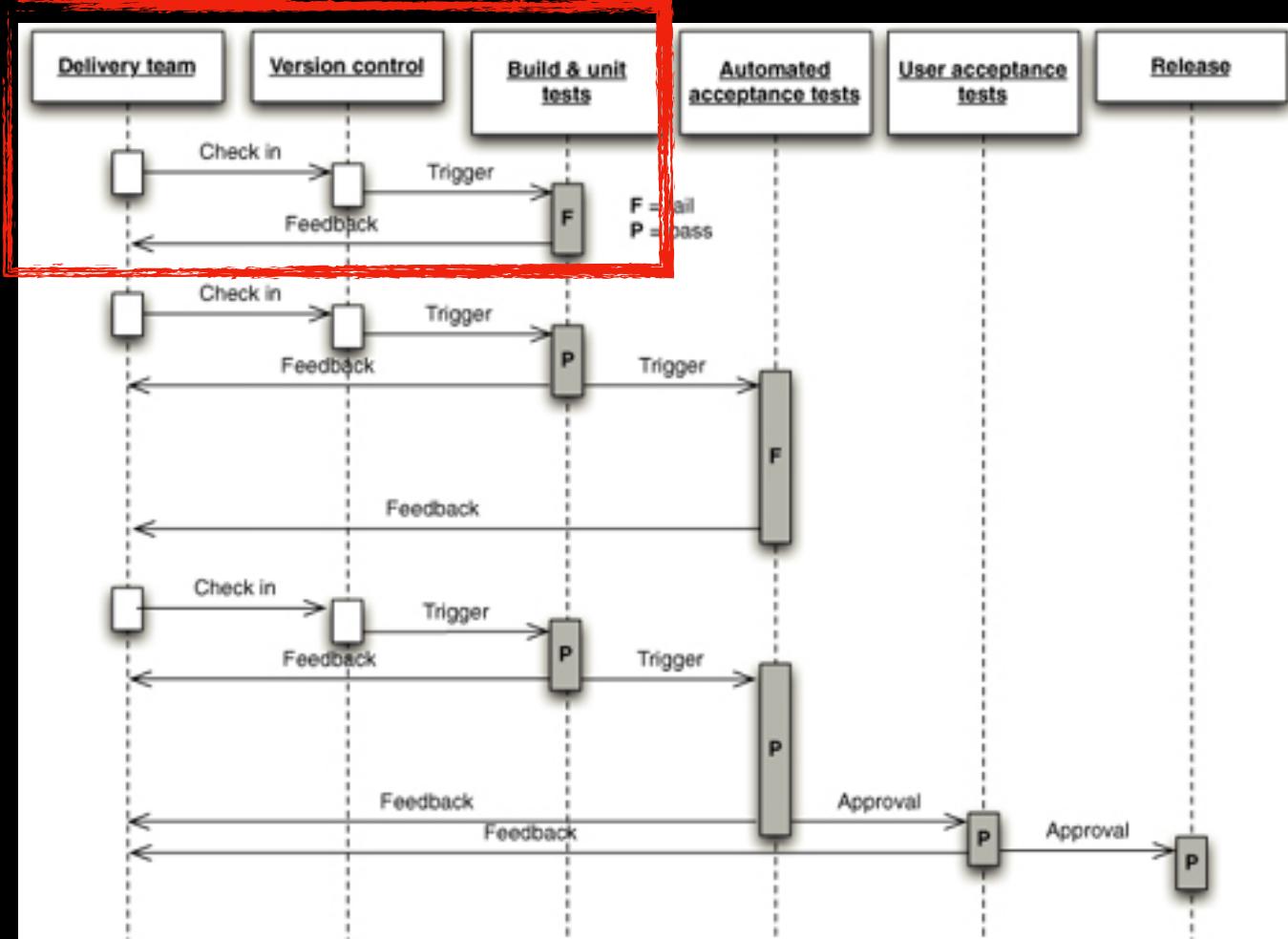
Extensive automation of all possible parts of the delivery process



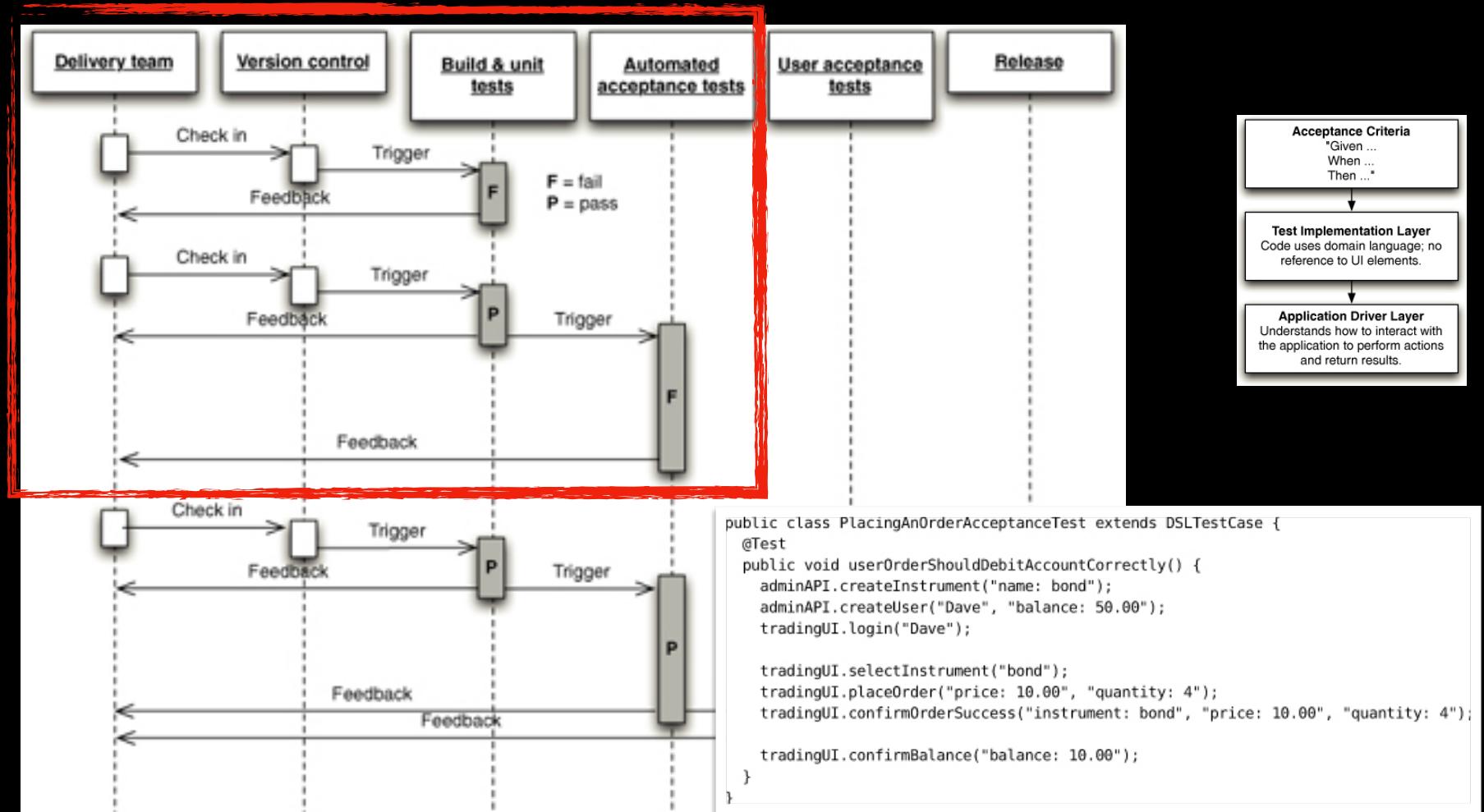


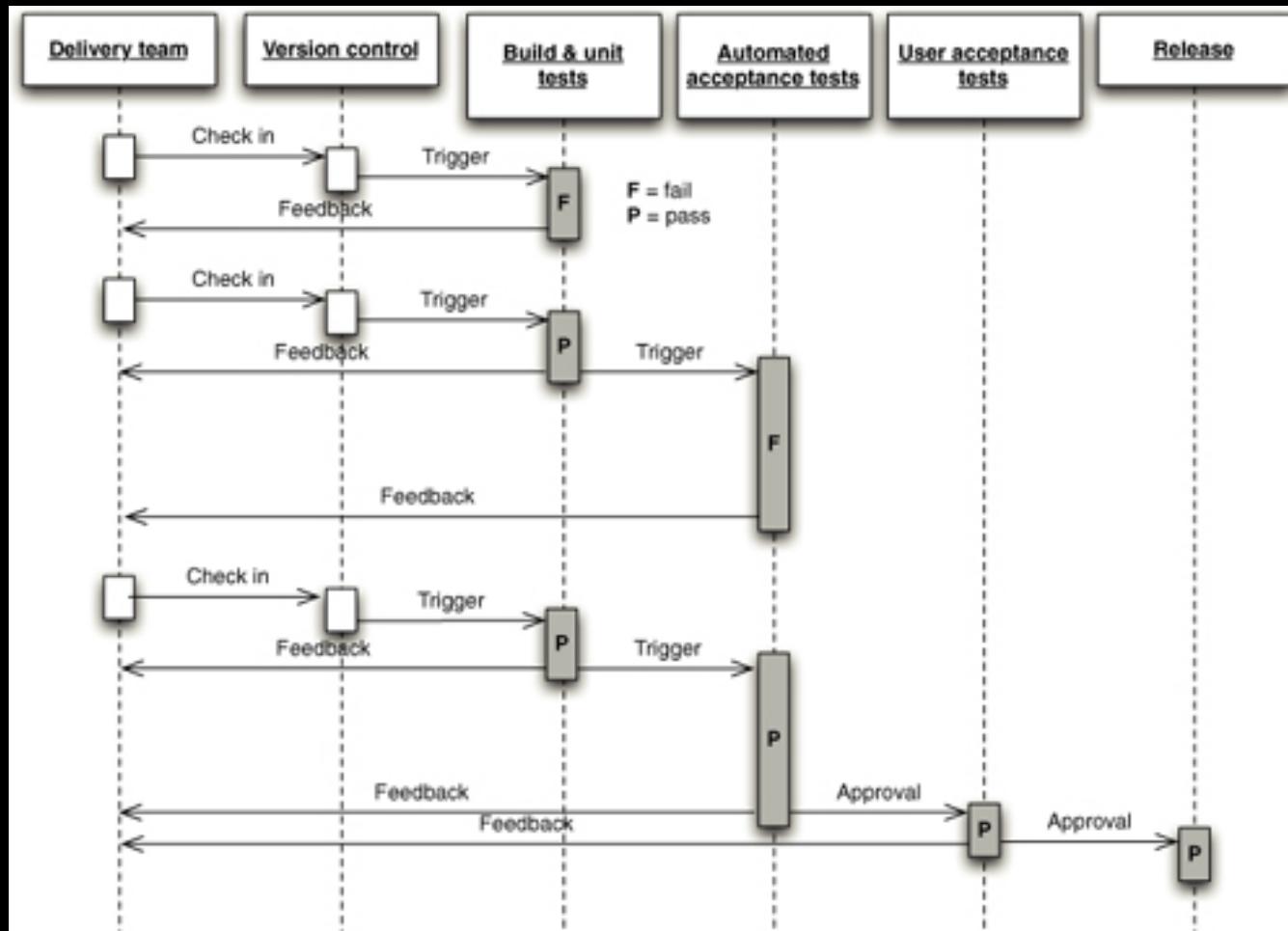
Continuous Delivery Reliable Software Releases through Build, Test, and Deployment Automation by Jez Humble and David Farley

Commit Stage



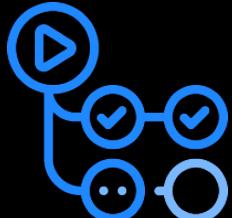
detecting, as fast as possible, the most common failures that changes to the system may introduce, and notifying the developers so they can fix the problems quickly.





Continuous Delivery Reliable Software Releases through Build, Test, and Deployment Automation by Jez Humble and David Farley

CI/CD Tools



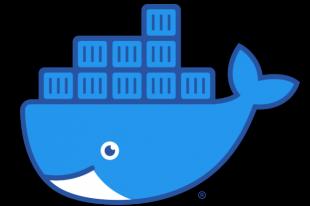
The image displays three screenshots illustrating CI/CD tools:

- Top Screenshot:** A modal titled "Where do you store your code?" lists integration options: Bitbucket Cloud, Bitbucket Server, GitHub, GitHub Enterprise, and Git.
- Middle Screenshot:** The Bitbucket Cloud CI/CD interface for a pull request. It shows a pipeline step named "Initialize" with a status of "Running". The pipeline consists of three stages: Initialize, Build, and Report.
- Bottom Screenshot:** The Jenkins dashboard for the repository "bitwise-jenkins / junit-plugin". It shows the history of five builds. The table includes columns for Status, Run, Commit, Branch, Message, Duration, and Completed. The most recent build (Run 5) is successful, while others show various errors or warnings.

Status	Run	Commit	Branch	Message	Duration	Completed
✓	3	b518058	PR-7	-	4m 51s	a minute ago
✓	1	63a7f47	master	-	5s	8 minutes ago
✗	1	86b2229	PR-7	-	48s	7 minutes ago
✗	6	d3d9a39	blog/blue-ocean-editor	-	1m 52s	12 minutes ago
✓	5	d3d9a39	blog/blue-ocean-editor	-	11m 8s	4 hours ago

ML Specific Concerns

Serve Model in Production



- **Embedded model:** treat the model artifact as a dependency that is built and packaged within the consuming application.
- **Model deployed as a separate service:** the model is wrapped in a service that can be deployed independently of the consuming applications.
- **Model published as data:** the model is treated and published independently, but the consuming application will ingest it as data at runtime.

ML Specific Concerns

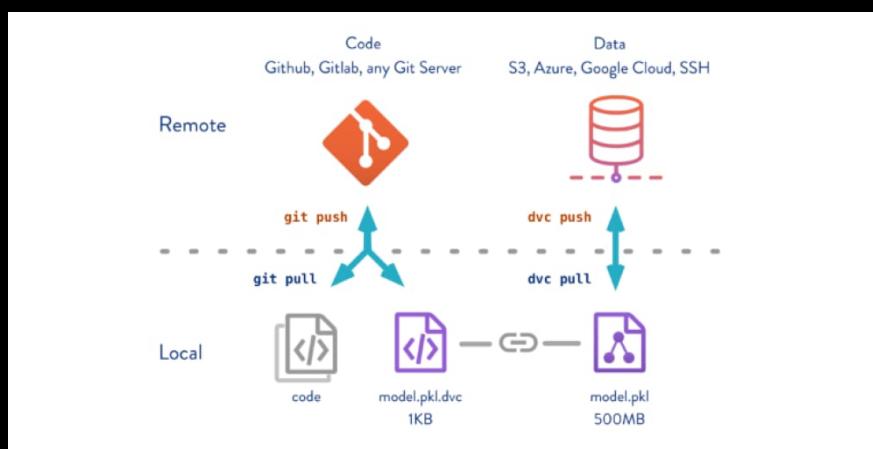
Deployment Options

- Multiple models: release more than one model performing the same task
- Shadow models: send the same production traffic to gather data on how the shadow model performs before promoting it, similar to Blue-green deployment
- Competing models: trying multiple versions of the model in production, similar to an A/B test
- Online learning models: need to version not only the training data, but also the production data that will impact the model's performance.

ML Specific Concerns

Experiment Management

- Which configuration works best for which dataset?



The screenshot shows the mlflow web interface with the title 'Listing Price Prediction'.

Experiment ID: 0 Artifact Location: /Users/matei/mlflow/demo/mlruns/0

Search Runs: metrics.R2 > 0.24 Filter Metrics: rmse, r2 Search

Filter Params: alpha, lr Clear

4 matching runs Compare Selected Download CSV

	Time	User	Source	Version	Parameters		Metrics		
					alpha	I1_ratio	MAE	R2	RMSE
□	17:37	matei	linear.py	3a1995	0.5	0.2	84.27	0.277	158.1
□	17:37	matei	linear.py	3a1995	0.2	0.5	84.08	0.264	159.6
□	17:37	matei	linear.py	3a1995	0.5	0.5	84.12	0.272	158.6
□	17:37	matei	linear.py	3a1995	0	0	84.49	0.249	161.2

ML Specific Concerns

Register Models

The screenshot shows the Databricks Model Registry interface for a registered model named "Airline_Delay_SparkML" at version 5. The left sidebar includes links for Home, Workspace, Recents, Data, Clusters, Jobs, Models (which is selected), and Search.

Key details on the page:

- Registered At: 2019-10-11 12:44:44
- Creator: clemens@demo.com
- Last Modified: 2019-10-14 12:19:32
- Source Run: Run 6151fe768a5e49d39076b07448e60d57
- Stage: Staging (highlighted in orange)

The Stage dropdown menu shows three options:

- Request transition to → None
- Request transition to → Production (highlighted in green)
- Request transition to → Archived

The Pending Requests section shows one pending request:

Request	Request by	Actions
Transition to → Production	matei@demo.com	Approve Reject

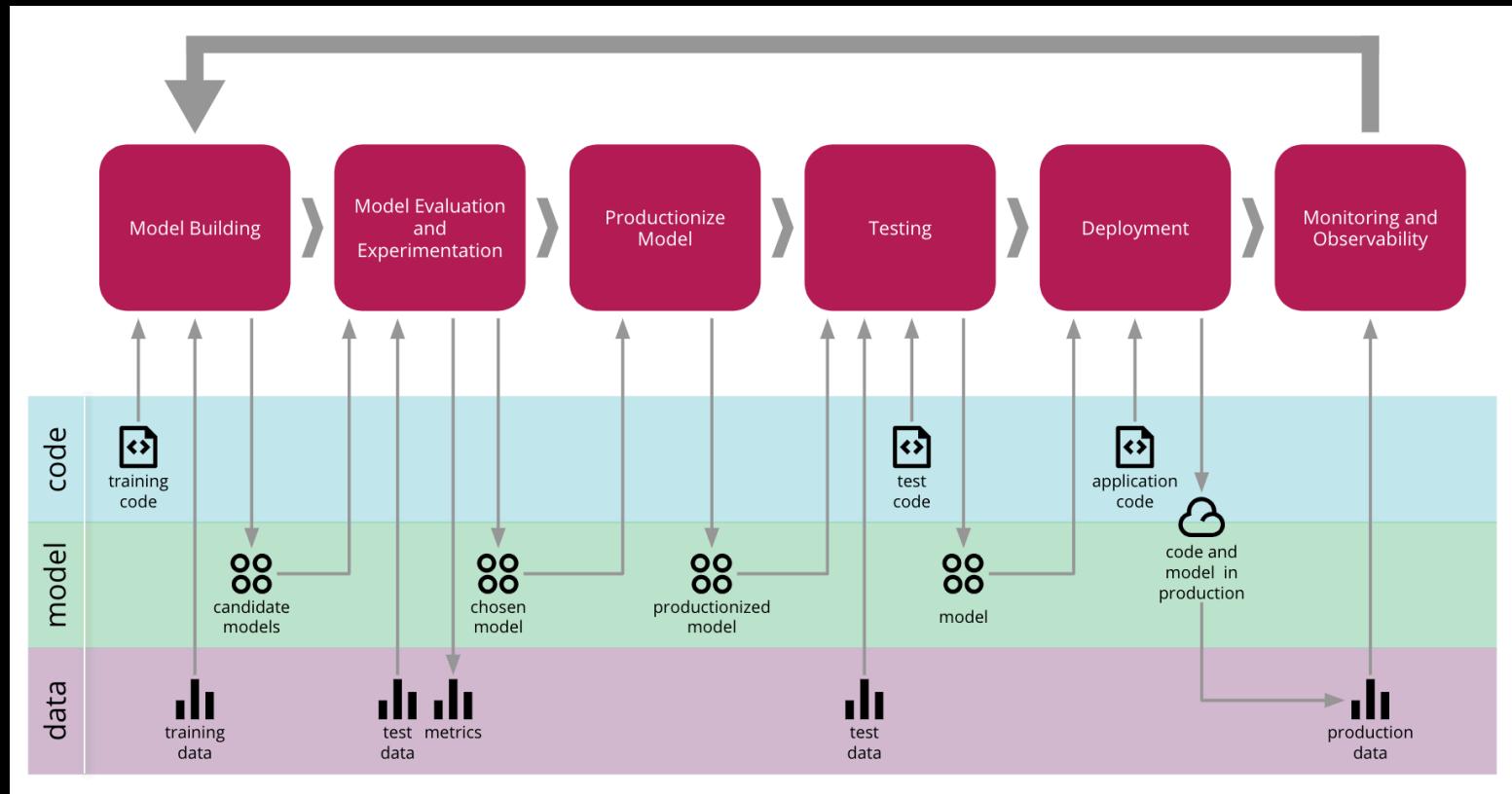
The Activities section lists recent events:

- clemens@demo.com rejected a stage transition → None 5 minutes ago (indicated by a red circle with a white X)
- matei@demo.com applied a stage transition None → Staging 4 minutes ago (indicated by a green circle with a checkmark)
- matei@demo.com requested a stage transition Staging → Production 4 minutes ago (indicated by a blue circle with a hand icon)

A note at the bottom says: "Tested this offline, looks good to launch!"

<https://databricks.com/blog/2019/10/17/introducing-the-mlflow-model-registry.html>

Primary Objective



<https://martinfowler.com/articles/cd4ml.html>