

6. Deployment

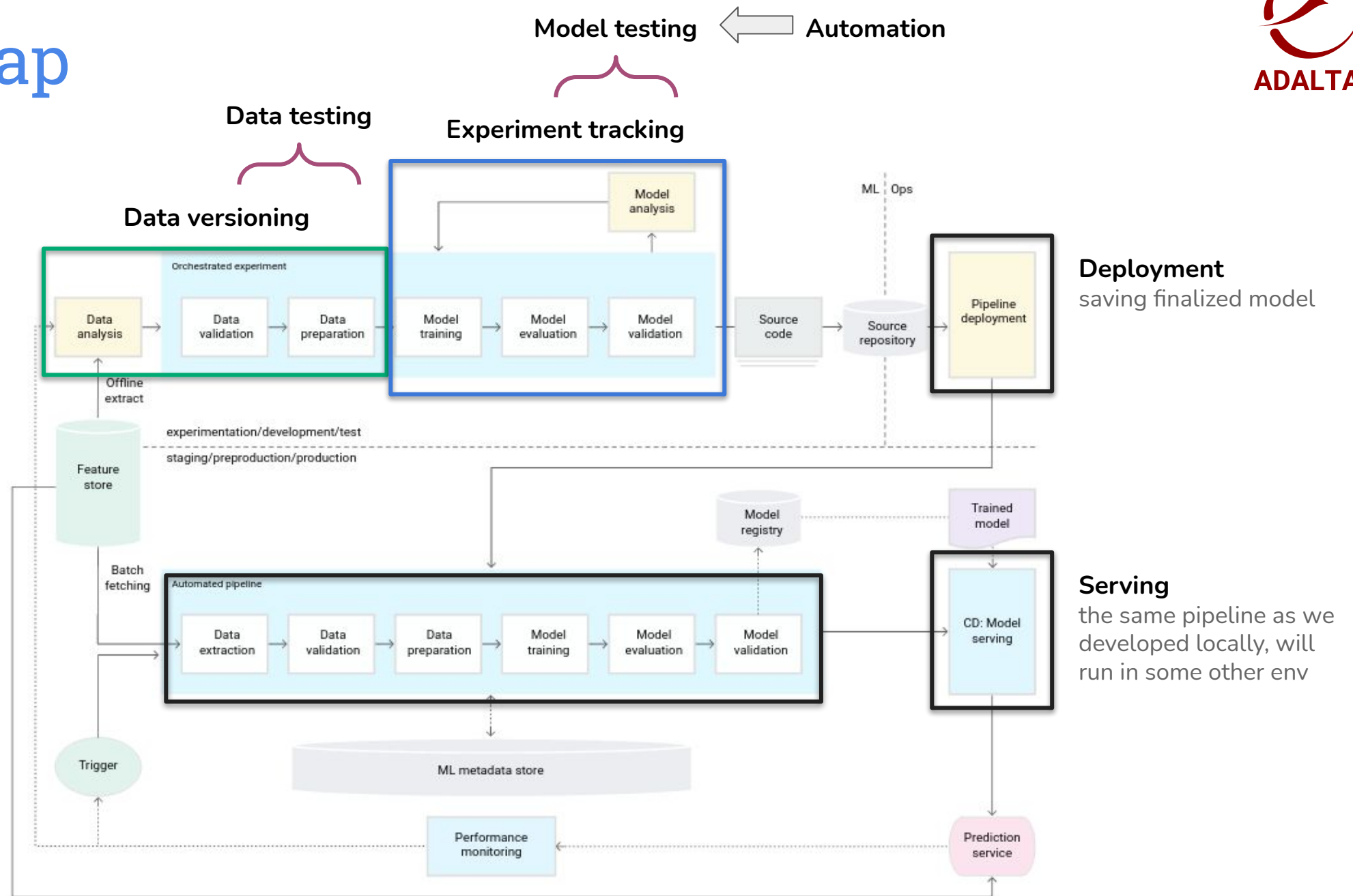
Exposing the model to the users

When we deploy:



- End-users can (finally) use it for inference
- We start generating value from all the hard work

Roadmap



Vocabulary

- **Delivery** - a step before deployment. The model is all ready, just waiting to be deployed.
- **Deployment** - integrating a new model into existing environment. It doesn't need to be in interaction with end-users.
- **Serving** - a model put in disposition to end-users.
- **Release** - when we say a version of a service is released, we mean that it is responsible for serving production traffic. In verb form, releasing is the process of moving production traffic to the new version.
- **Release in Place (Or deploy == release)** - When your team's shipping process involves pushing a new version of your software onto a server running the old version and re-starting the process, you're releasing in place.

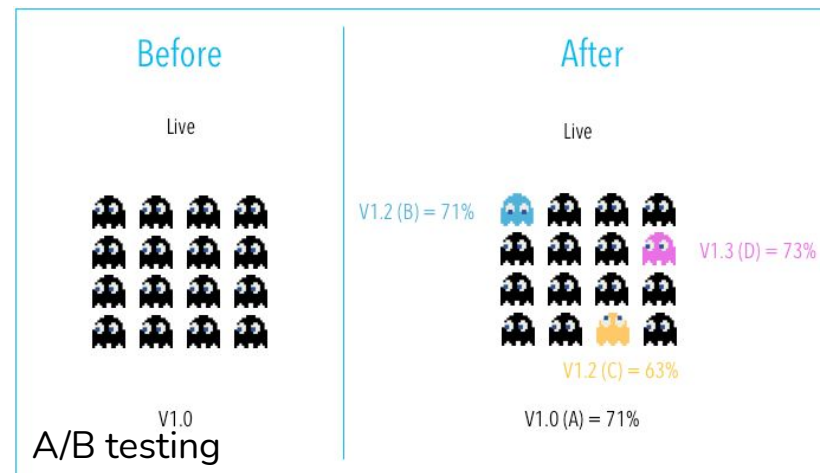
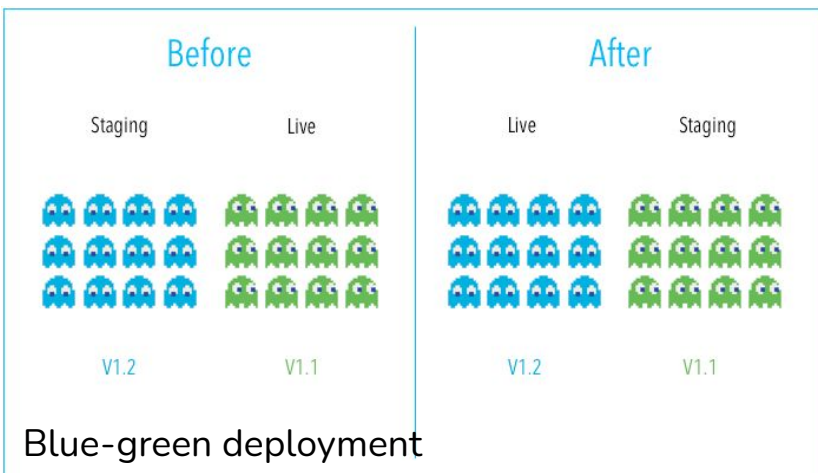
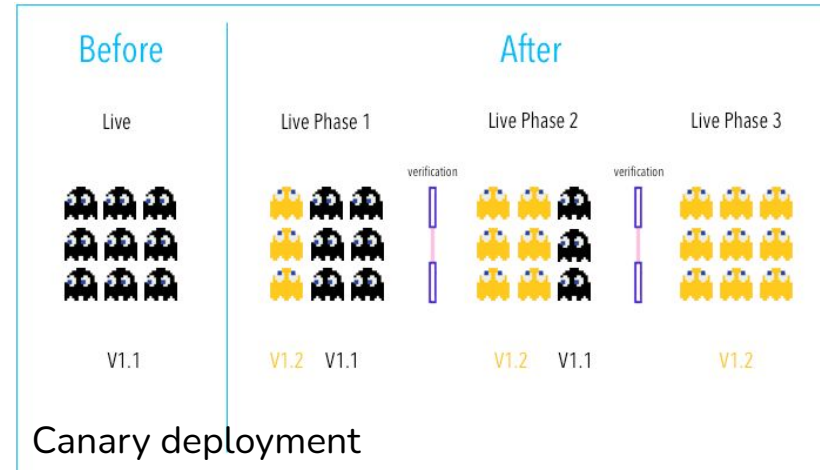
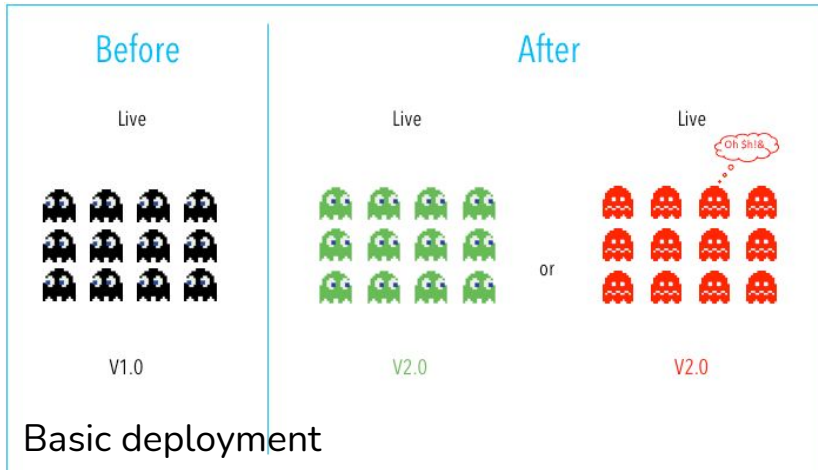
Deployment targets

- Microservices with a REST API to serve online predictions.
- An embedded model to an edge or mobile device.
- Part of a batch prediction system.

Latency constraints

- Batch predictions
- Real-time predictions

Deployment strategies



What do we have for now?

- tested and versioned source code
 - data pre-processing
 - modeling
- => we need to deploy both (the product is the entire pipeline, not only the model)

What do we need (what will we deploy)?

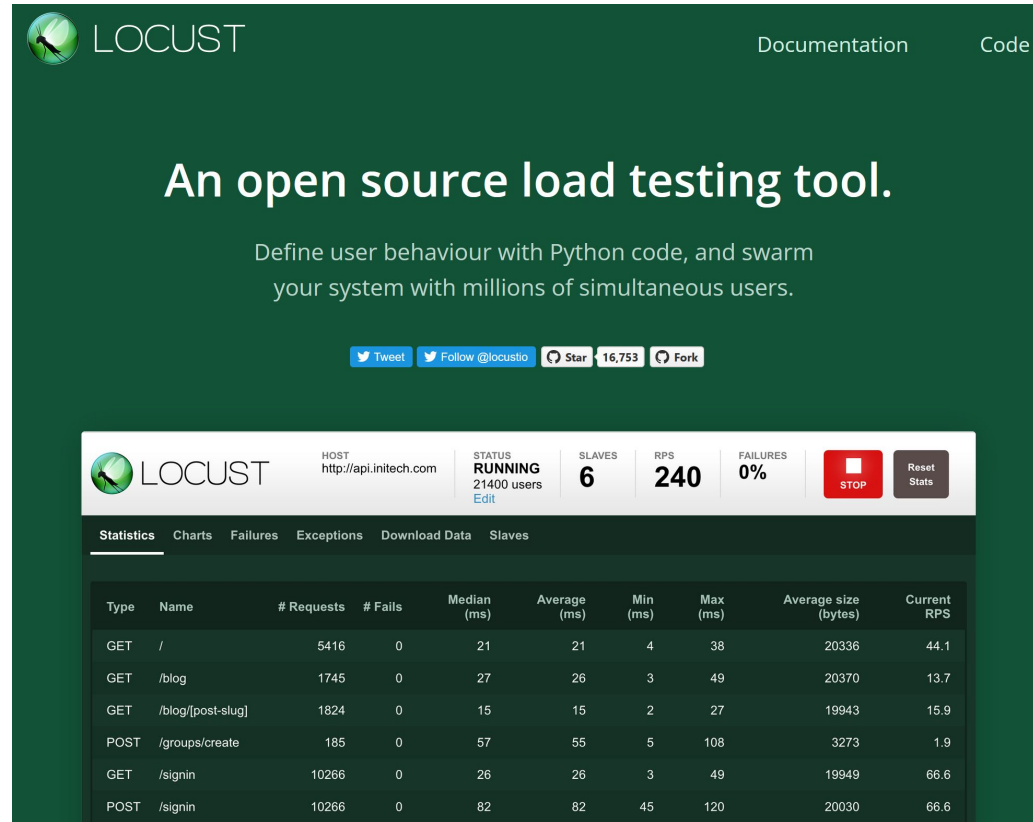
- 'packed' model (.pkl)...
- ...turned into an application
- description of the environment (list of libraries)
- server (Heroku)

What will we do?

- Turn the source code into .pkl
- Create Flask app
- Create environment definition file (.yaml)
- Deploy the app on Heroku cluster
- Objective:
 - get to know the steps between the source code and functional (deployed) model
- Result:
 - serving model, accessible through internet

What could be done next?

- <https://locust.io/>



The screenshot displays the Locust web interface. At the top, there's a navigation bar with the Locust logo, "LOCUST", and links for "Documentation" and "Code". Below this, a large green banner contains the text "An open source load testing tool." and "Define user behaviour with Python code, and swarm your system with millions of simultaneous users." Social media links for Twitter, GitHub, and others are present. The main content area shows a summary of the current test: "HOST: http://api.initech.com", "STATUS: RUNNING", "21400 users", "6 SLAVES", "240 RPS", and "0% FAILURES". A "STOP" button and a "Reset Stats" button are also visible. Below this, a "Statistics" tab is active, showing a table of request metrics.

| Type | Name | # Requests | # Fails | Median (ms) | Average (ms) | Min (ms) | Max (ms) | Average size (bytes) | Current RPS |
|------|-------------------|------------|---------|-------------|--------------|----------|----------|----------------------|-------------|
| GET | / | 5416 | 0 | 21 | 21 | 4 | 38 | 20336 | 44.1 |
| GET | /blog | 1745 | 0 | 27 | 26 | 3 | 49 | 20370 | 13.7 |
| GET | /blog/[post-slug] | 1824 | 0 | 15 | 15 | 2 | 27 | 19943 | 15.9 |
| POST | /groups/create | 185 | 0 | 57 | 55 | 5 | 108 | 3273 | 1.9 |
| GET | /signin | 10266 | 0 | 26 | 26 | 3 | 49 | 19949 | 66.6 |
| POST | /signin | 10266 | 0 | 82 | 82 | 45 | 120 | 20030 | 66.6 |