Machine Learning as a Service (MLaaS)

Jake Carlson, Ian Johnson

What is Machine Learning

We’re going to treat ML as a black box for this talk. Let’s think of it as a type of computing that is prohibitively expensive to run on-device in many situations.

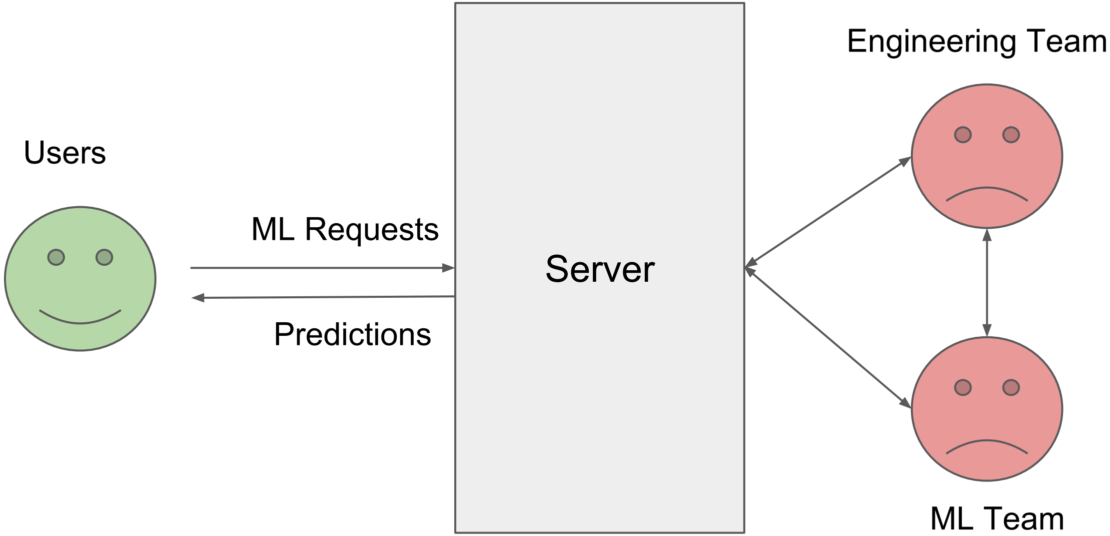
Enterprise Machine Learning

Two sets of people working on server-side ML tools:

1) ***Engineering teams*** build platform for ML models to run on
2) ***Data/ML teams*** research and train models

In a monolithic server application design, this really doesn’t work well.

Monolithic ML Server Architecture (cont.)



Issues

How can the data science team push out new models without involving the engineering team?

How can the engineering team update the serving mechanism without involving the ML team?

Decoupled systems are favorable to monolithic systems in situations like these.

MLaaS Architecture

Load Balancer:

- Sends inference requests to runners

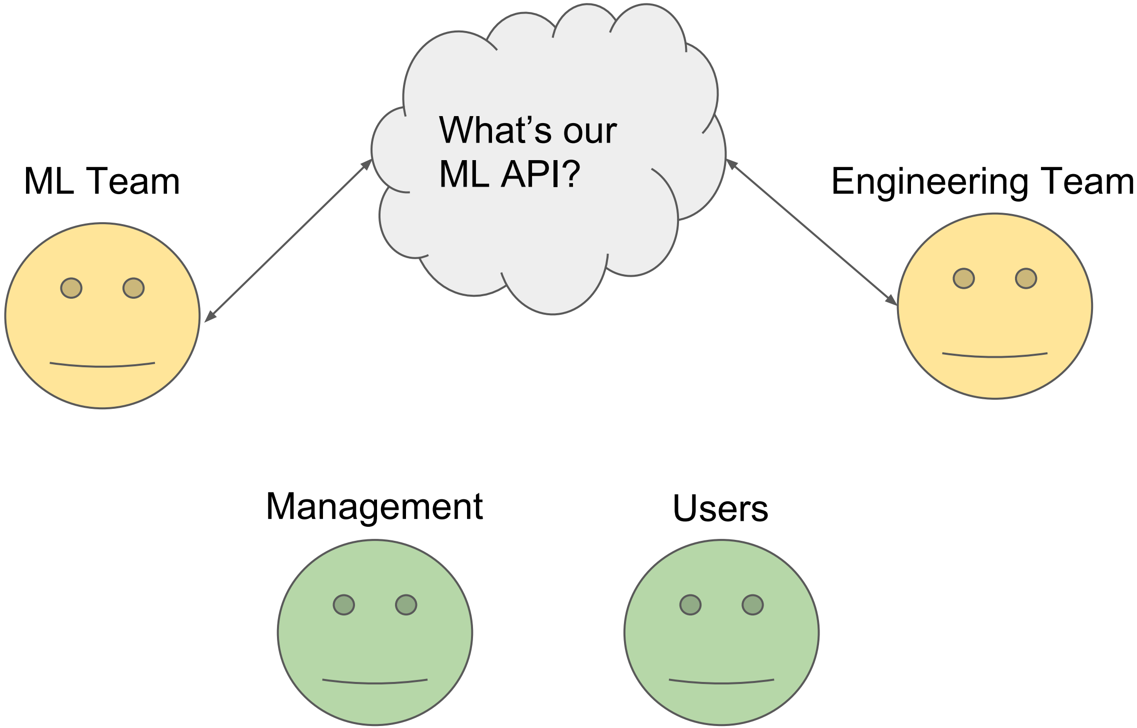
Inference Runner:

- Pulls model from Model Server
- Performs the expensive computations of running the ML model
- Can run on specialized hardware for ML inference (think cloud TPU/GPU)

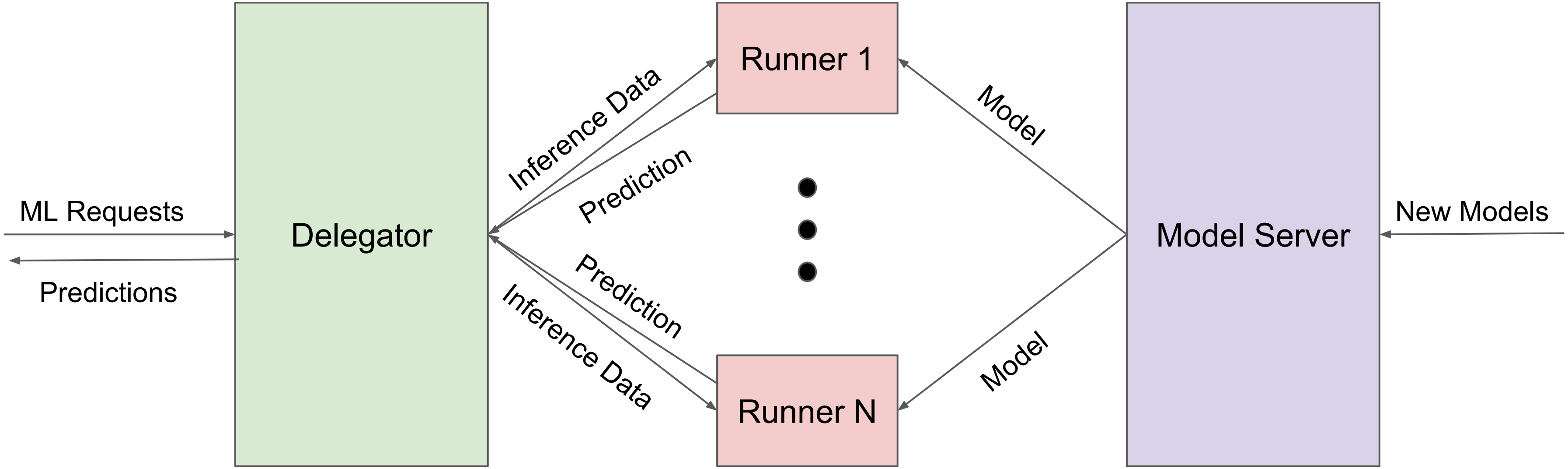
Model Server:

- Accepts new models from data team
- Sends models to runners

Communication Between Engineering and ML Teams



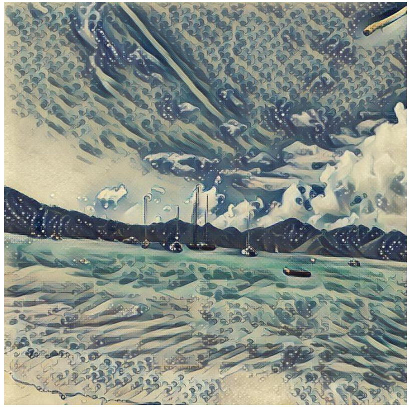
MLaaS Architecture (cont.)



Demo

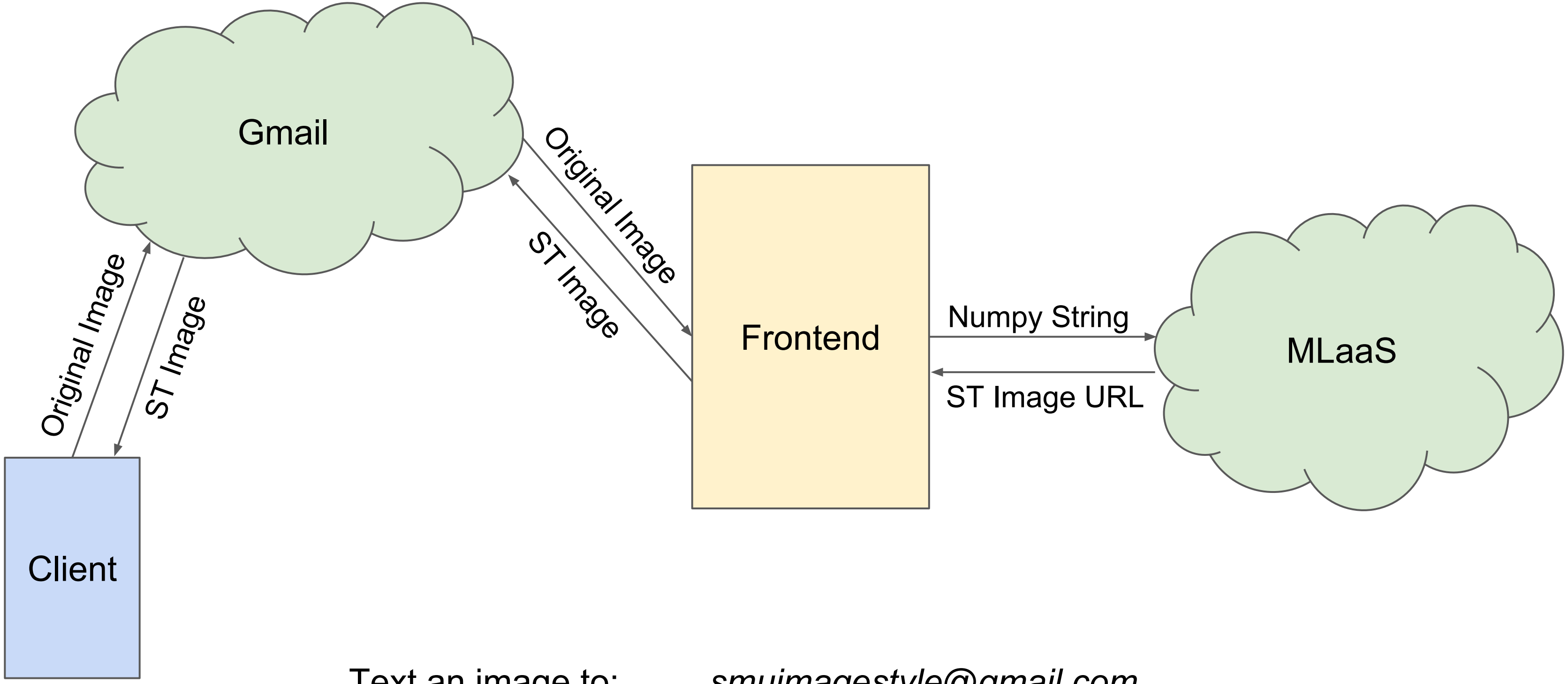
Reference Implementation - Style Transfer

冨嶽三十六景 神奈川沖浪裏



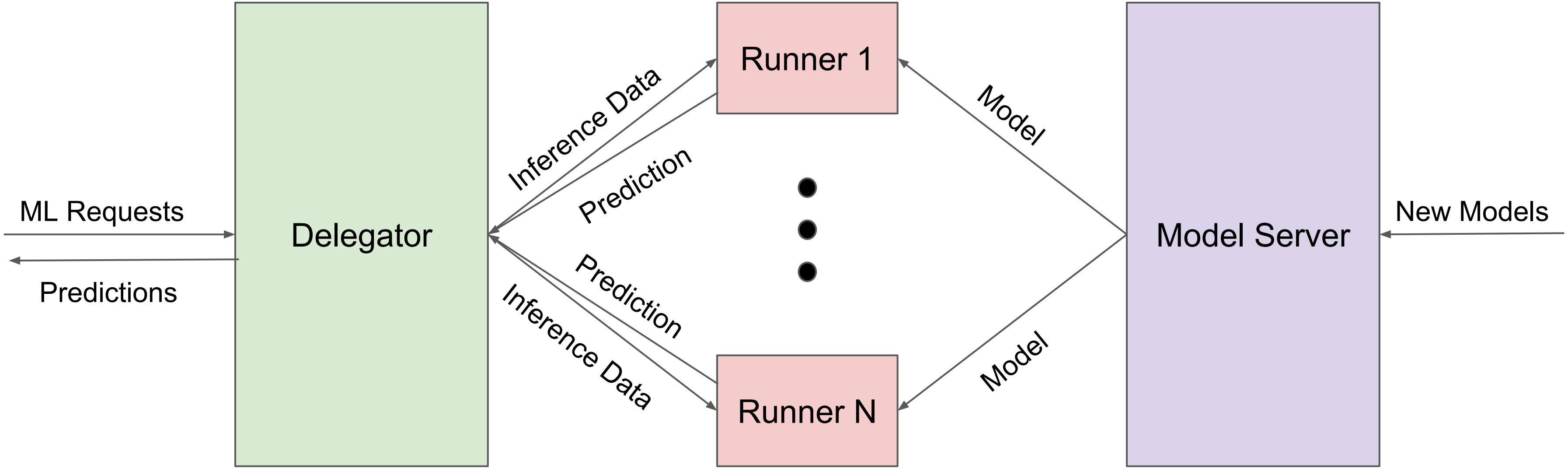
Text an image to: *smuimagestyle@gmail.com*

Reference Implementation - Architecture



Text an image to: *smuimagestyle@gmail.com*

MLaaS Architecture (cont.)



Text an image to: *smuimagestyle@gmail.com*

Strengths

- Easy to incorporate into existing monolith or microservice platforms
- New models can be used instantly
- No involvement from engineering team to deploy new models

Weaknesses

- Models server needs a database for the server to be stateless
- Starting and stopping runners requires additional configuration (Kubernetes)
- Inference data gets rerouted several times

References

- Tornado server framework - Dr. Eric Larson:
  - https://github.com/SMU-MSLC/tornado_bare
- Style Transfer Models - ImageStyle:
  - https://github.com/ImageStyle