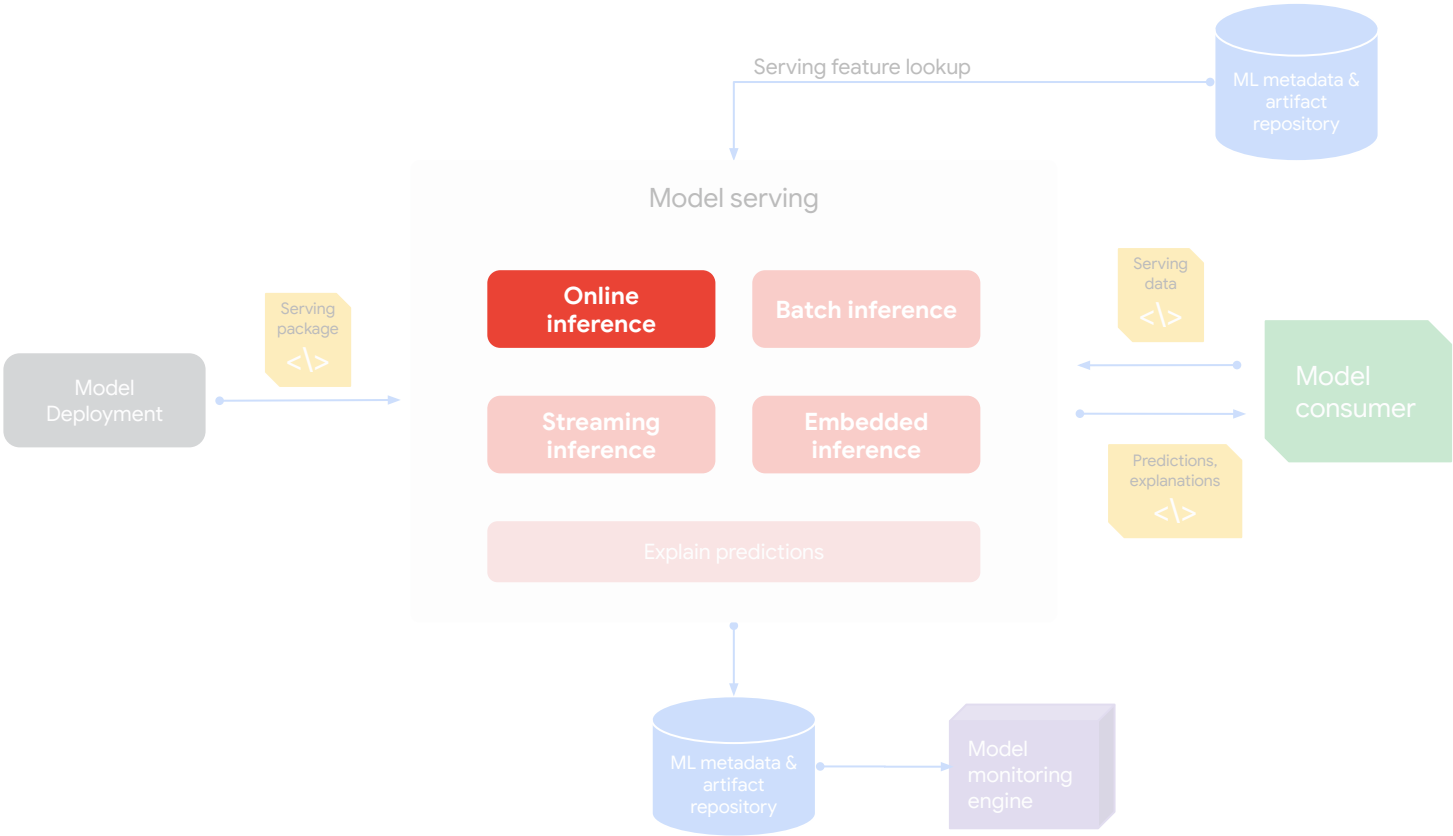
Prediction Serving Scenarios: Online

MLOps: Prediction Serving



The MLOps **Personas**

ML Engineer

ML Researcher

Data Scientist

Data Engineer

Software Engineer

DevOps

Business Analyst

Online Inference: What is it?

Online inference in near real time for high-frequency singleton requests coming into the service

Online Inference: **What is it?**

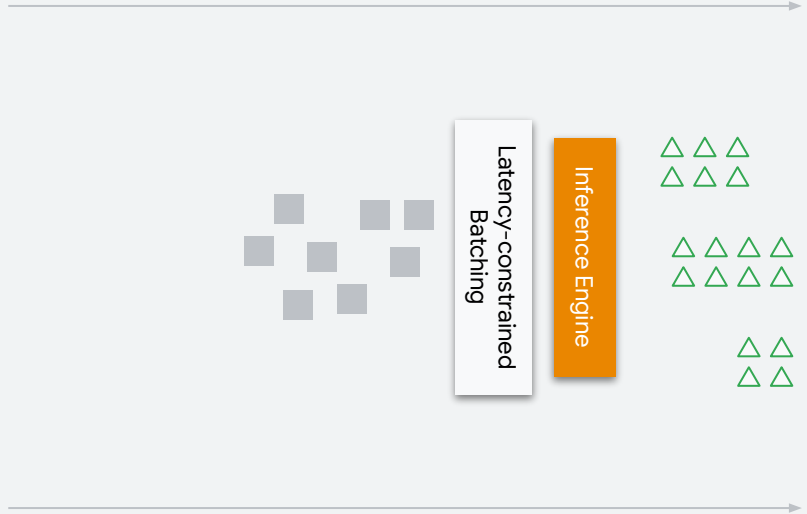
Online inference in near real time for high-frequency singleton requests coming into the service

Predict on-demand

Online Inference: **What is it?**

Online inference in near real time for high-frequency singleton requests coming into the service

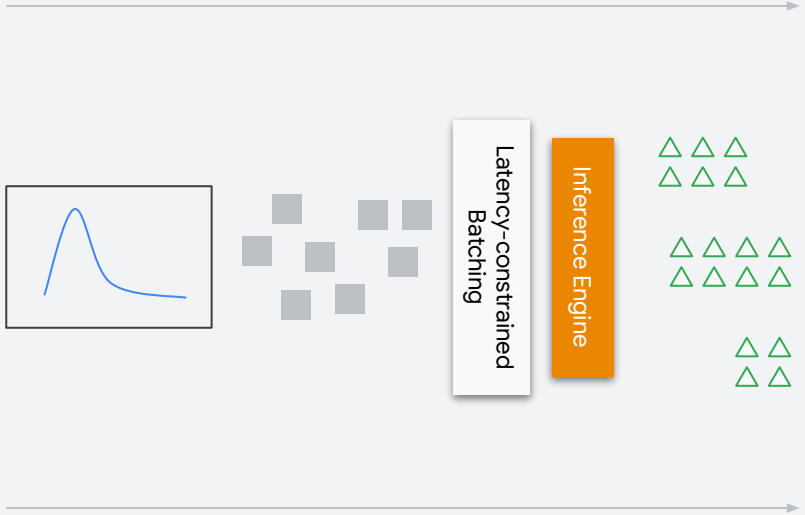
Predict on-demand



Online Inference: **What is it?**

Online inference in near real time for high-frequency singleton requests coming into the service

Predict on-demand



Online Inference: When is it useful?

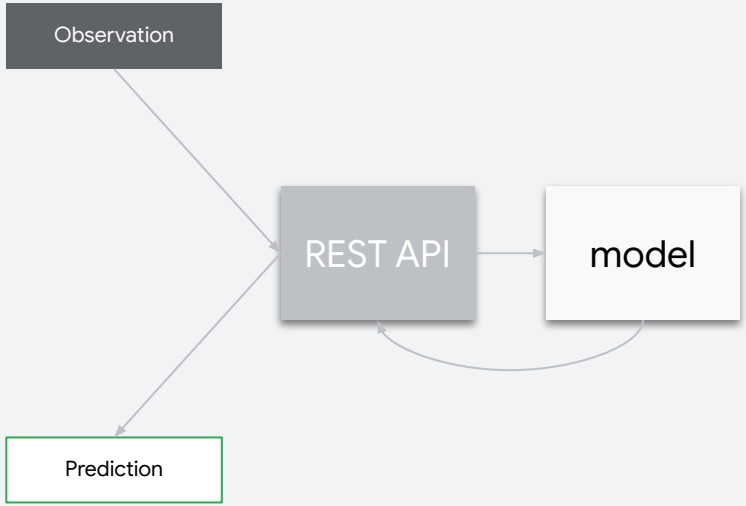
- Food delivery times
- Estimated arrival time
- Predictive maintenance
- ...

Online Inference: When is it useful?

- Food delivery times
- Estimated arrival time
- Predictive maintenance
- ...

Online Inference: How it works?

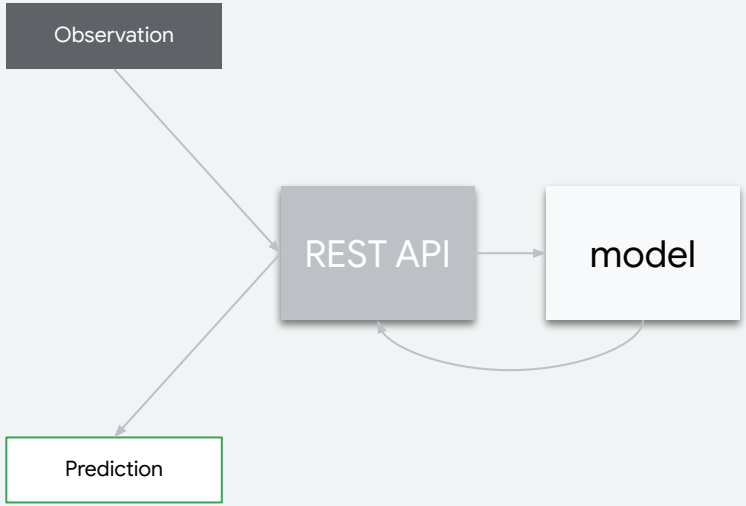
Typically exposed through a REST API via a client-server network architecture and needs fairly good knowledge of the system



Online Inference: How it works?

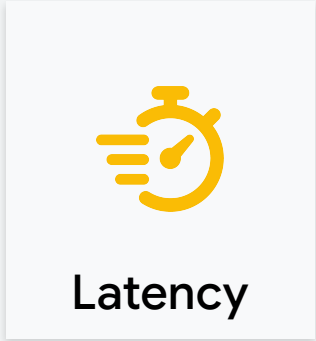
Typically exposed through a REST API via a client-server network architecture and needs fairly good knowledge of the system

Takes the whole end-to-end inference life-cycle into account to deliver good latency responses



Online Inference: What metrics?

- Single-stream
- Latency & throughput metric
  - E.g., bounds 100 ms
  - E.g., thousands of queries per second
  - Queries-per-second (QPS)



Online Inference: Pros & Cons

Pros

\+ Can make on-the-fly predictions on **new** requests

Online Inference: Pros & Cons

Pros

+ Can make on-the-fly predictions on **new** requests
+ Great for the **long-tail**

Online Inference: Pros & Cons

Pros

+ Can make on-the-fly predictions on **new** requests
+ Great for the **long-tail**
+ **Pay-per-use** service

Online Inference: Pros & Cons

Pros

+ Can make on-the-fly predictions on **new** requests
+ Great for the **long-tail**
+ **Pay-per-use** service

Cons

- **Compute intensive**

Online Inference: Pros & Cons

Pros

+ Can make on-the-fly predictions on **new** requests
+ Great for the **long-tail**
+ **Pay-per-use** service

Cons

- **Compute intensive**
- **Latency sensitive**—may limit model complexity

Online Inference: Pros & Cons

Pros

+ Can make on-the-fly predictions on **new** requests
+ Great for the **long-tail**
+ **Pay-per-use** service

Cons

- **Compute intensive**
- **Latency sensitive**—may limit model complexity
- **Monitoring** needs are more demanding and required

Online Inference: Pros & Cons

Pros

+ Can make on-the-fly predictions on **new** requests
+ Great for the **long-tail**
+ **Pay-per-use** service

Cons

- **Compute intensive**
- **Latency sensitive**—may limit model complexity
- **Monitoring** needs are more demanding and required

Scenario	Metric
Batch inference (e.g. photo sorting app)	Throughput
Online inference (e.g. translation app)	QPS subject to latency bound