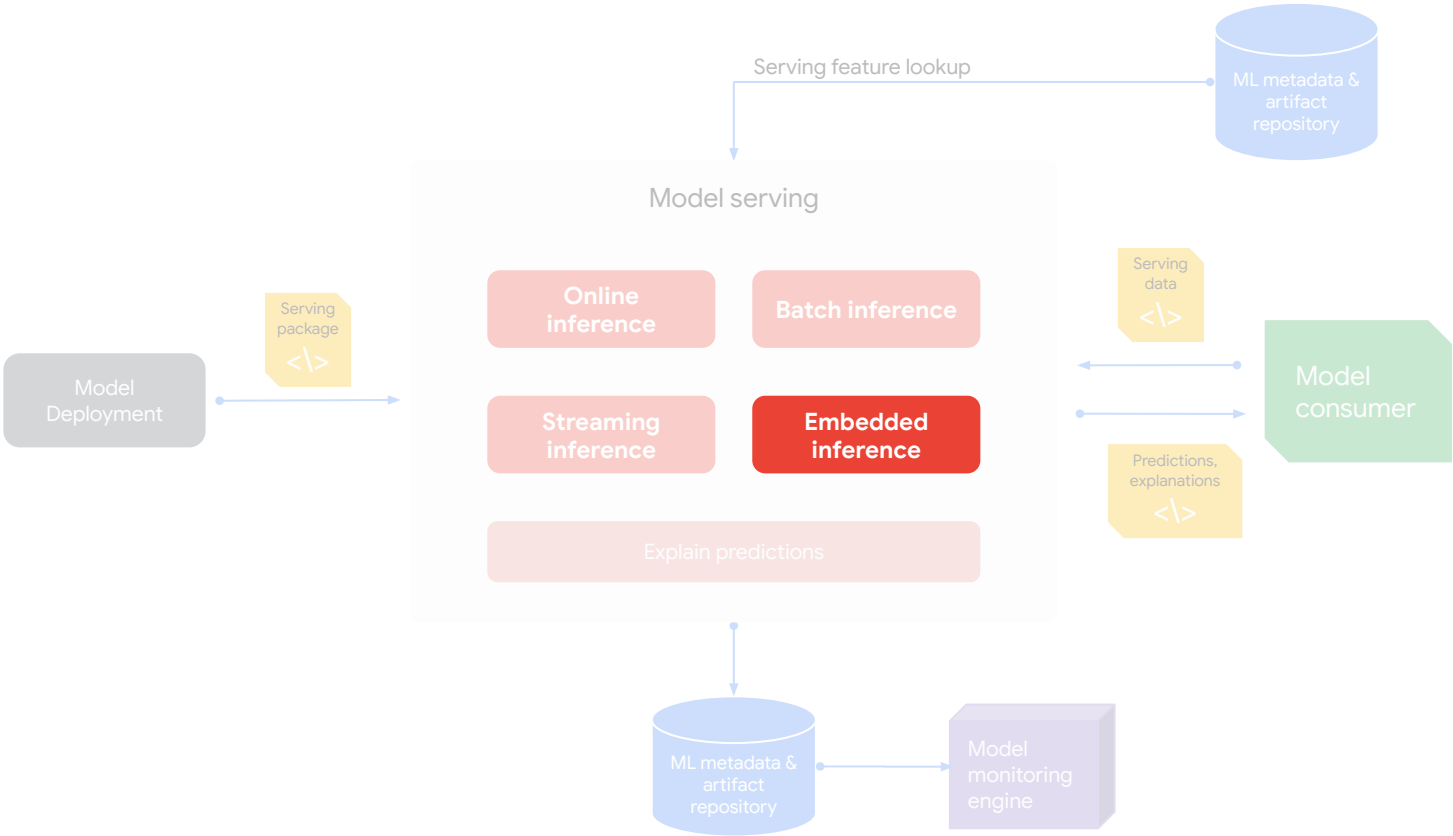
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Prediction Serving Scenarios: Embedded

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

MLOps: Prediction Serving



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

The MLOps **Personas**

ML Engineer

ML Researcher

Data Scientist

Data Engineer

Software Engineer

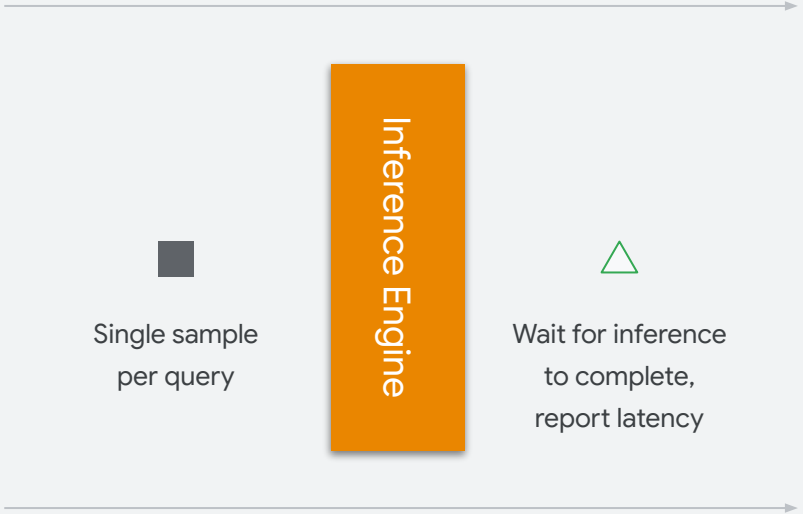
DevOps

Business Analyst

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: **What is it?**

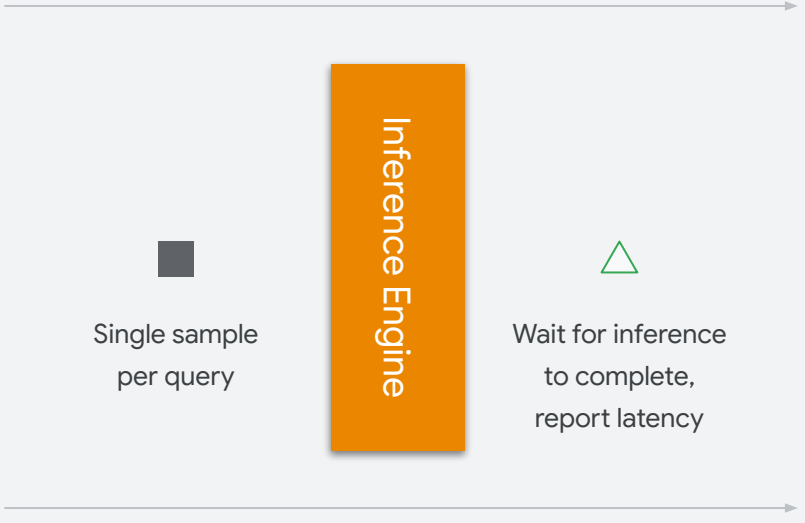
- Predict **on-demand**



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: **What is it?**

- Predict **on-demand**
- Online inference in near real time for *low-frequency* **singleton** requests



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: When is it useful?

- Smartphone camera
- TinyML use cases
- ...

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

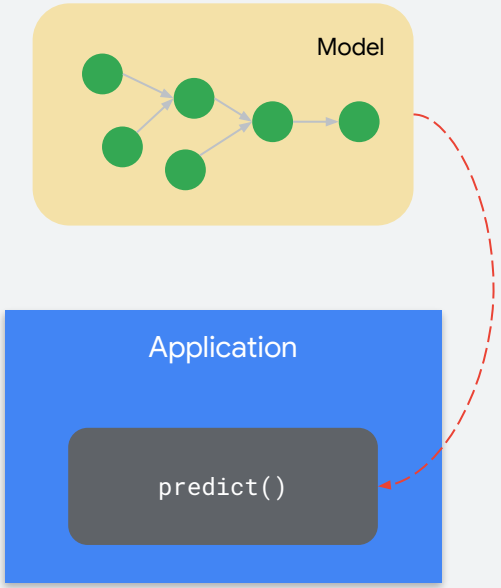
Embedded Inference: When is it useful?

- Smartphone camera
- TinyML use cases
- ...

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: How it works?

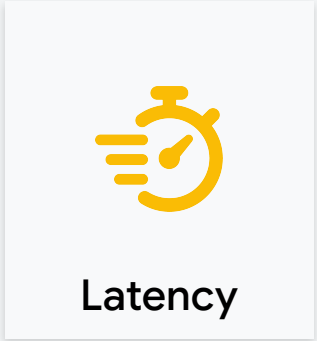
Model is packaged into the application for easy deployment at the endpoint device



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: **What metrics?**

- Single-stream
- Latency metric



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: Pros & Cons

Pros

+ Can make a **on-demand** predictions on items

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: Pros & Cons

Pros

- \+ Can make a **on-demand** predictions on items
- \+ Great for
  - \+ **Bandwidth**
  - \+ **Latency**
  - \+ **Energy-efficiency**
  - \+ **Reliability**
  - \+ **Privacy**

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: Pros & Cons

Pros

+ Can make a **on-demand** predictions on items
+ Great for
  + **Bandwidth**
  + **Latency**
  + **Energy-efficiency**
  + **Reliability**
  + **Privacy**

Cons

- Compute **intensive**

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: Pros & Cons

Pros

+ Can make a **on-demand** predictions on items
+ Great for
  + **Bandwidth**
  + **Latency**
  + **Energy-efficiency**
  + **Reliability**
  + **Privacy**

Cons

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

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: Pros & Cons

Pros

+ Can make a **on-demand** predictions on items
+ Great for
  + **Bandwidth**
  + **Latency**
  + **Energy-efficiency**
  + **Reliability**
  + **Privacy**

Cons

- Compute **intensive**
- Latency **sensitive**—may limit model complexity
- Monitoring needs are more **important** than for the other types of scenarios

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embedded Inference: Pros & Cons

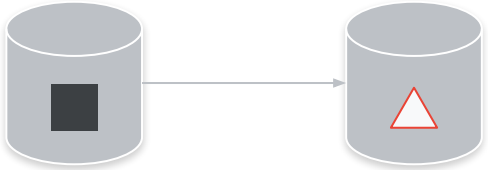
Pros

+ Can make a **on-demand** predictions on items
+ Great for
  + **Bandwidth**
  + **Latency**
  + **Energy-efficiency**
  + **Reliability**
  + **Privacy**

Cons

- Compute **intensive**
- Latency **sensitive**—may limit model complexity
- Monitoring needs are more **important** than for the other types of scenarios
- Embedded deployment makes scalability and flexibility **poor**

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Scenario	Metric
 **Batch inference** (e.g. photo sorting app)	**Throughput**
 **Online inference** (e.g. translation app)	**QPS** subject to latency bound
 **Streaming inference** (e.g. multiple camera driving assistance)	**Number streams** subject to latency bound
 **Embedded inference** (e.g. cell phone augmented vision)	**Latency**

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0