inovex

Data Science from Whiteboard to Production: End-to-End Image Captioning

TUM Data Innovation Lab

February 18, 2020

Co-Mentor: Michael Rauchensteiner
Project Lead: Dr. Ricardo Acevedo Cabra
Supervisor: Prof. Dr. Massimo Fornasier

Our Team

Dr. Robert Pesch

Sebastian Blank

Julia Kronburger

Oliver Borchert

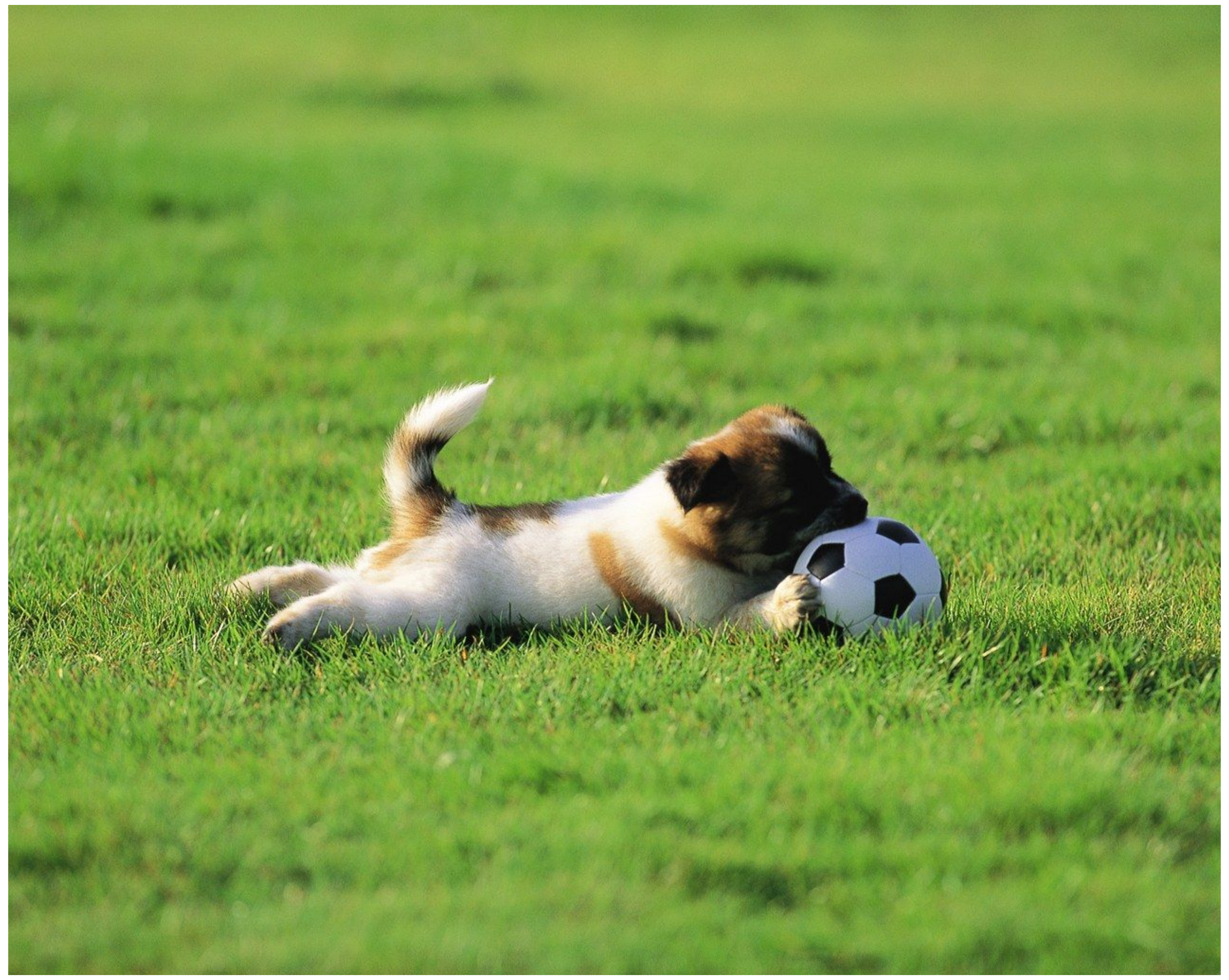
Nika Dogonadze

Mürüvvet Hasanbasoglu

Murtaza Raza

Anastasia Stamatouli

What is Image Captioning?



A dog is playing with
a ball on the grass

Computer Vision + Text Generation

From Whiteboard to Production

Data Acquisition

MSCoco

Flickr30k

Flickr8k

Model Development

Vanilla Encoder-Decoder

Visual Attention

Beam Search

Model Evaluation

BLEU

METEOR

ROUGE

CIDEr

Deployment

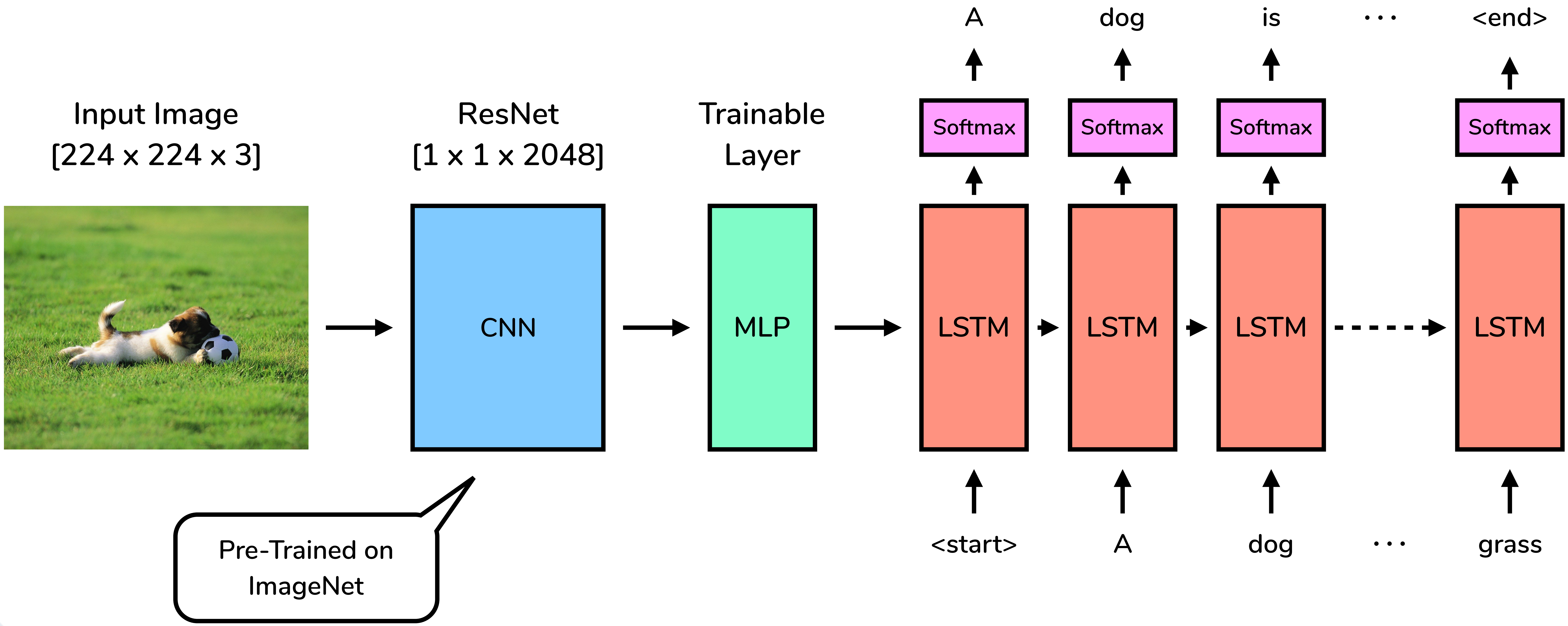
Architecture

Continuous Integration

Monitoring

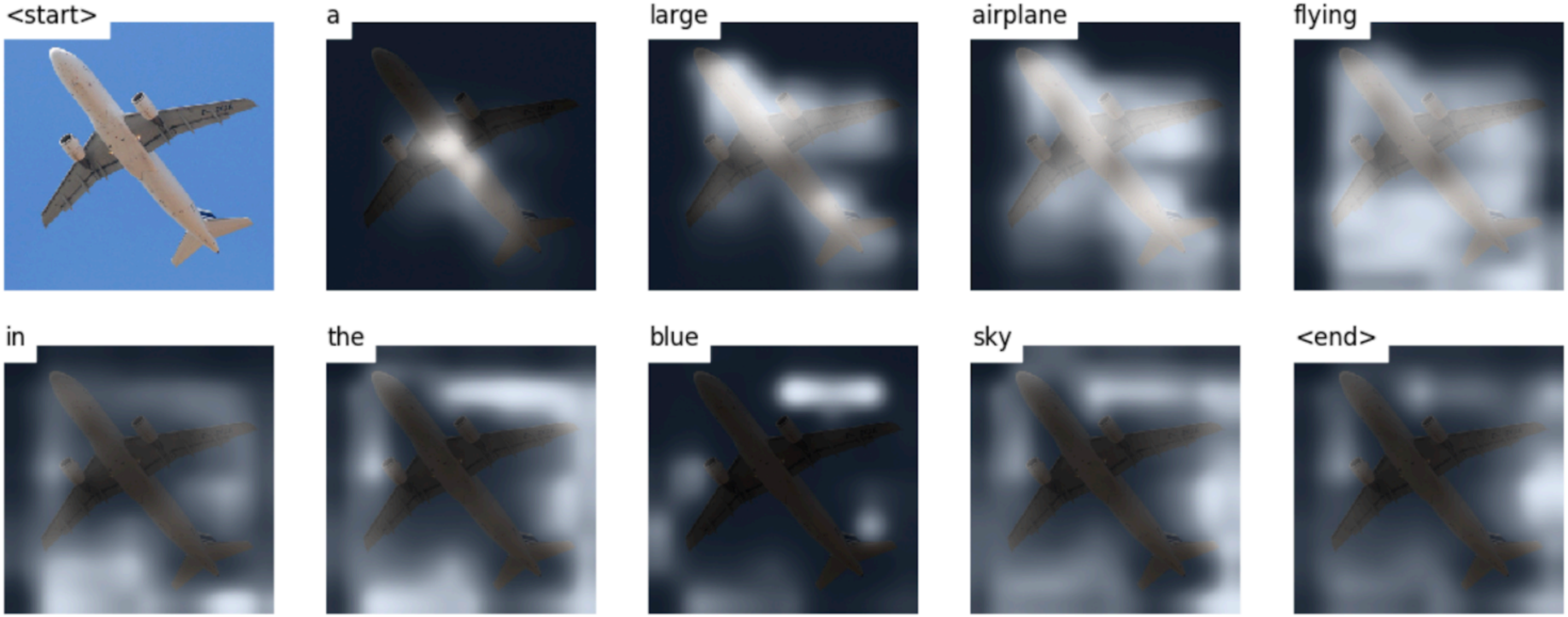
Model Development and Evaluation

Vanilla Encoder-Decoder

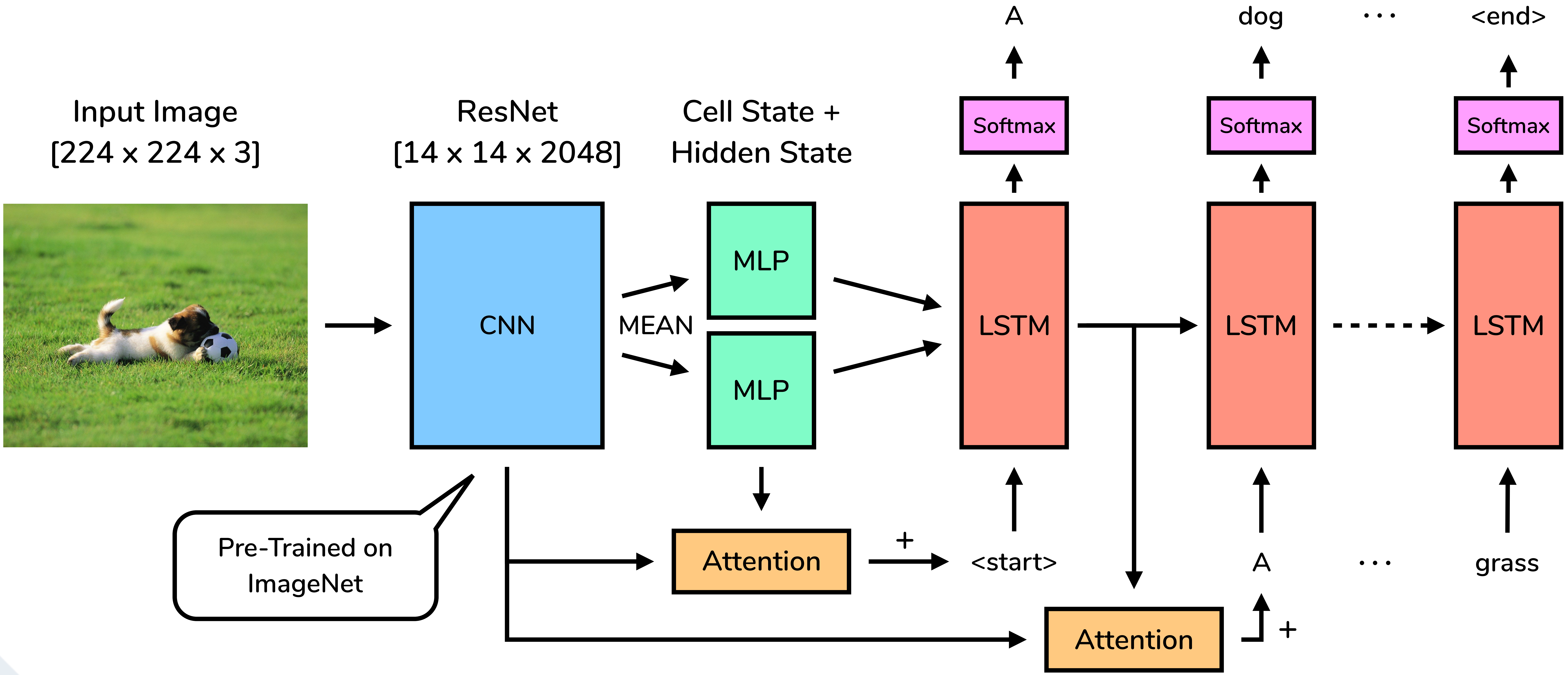


Pre-Trained on ImageNet

Visual Attention



Attentive Encoder-Decoder



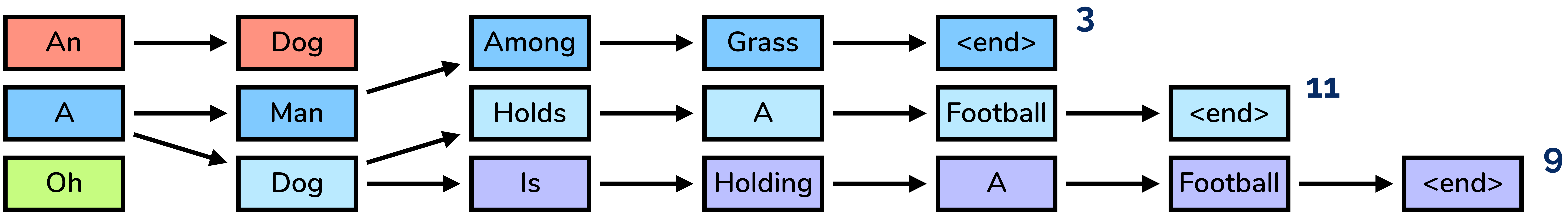
Beam Search

Greedy Search

Sequence of Most Likely States

Beam Search

Most Likely Sequence of States



An → Dog

A → Man

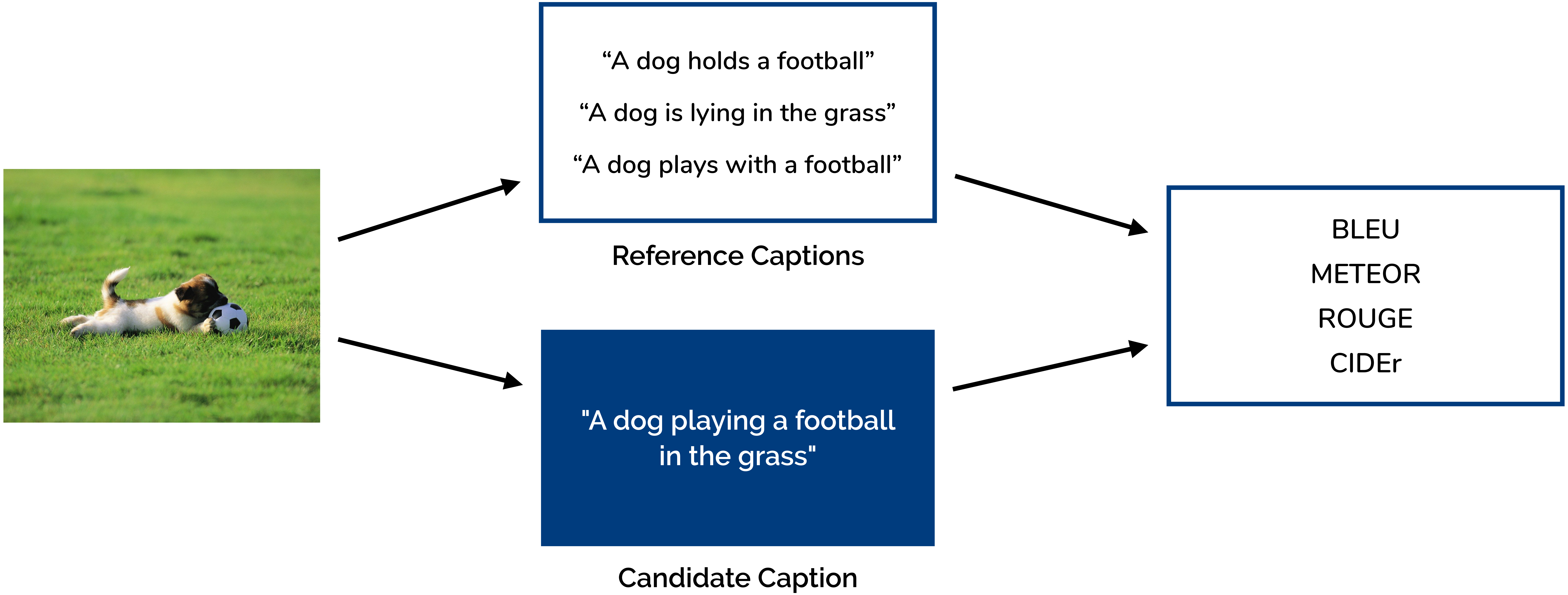
Oh

Dog → Is → Holding → A → Football → <end> 9

Among → Grass → <end> 3

Holds → A → Football → <end> 11

Evaluation Metrics



“A dog holds a football”

“A dog is lying in the grass”

“A dog plays with a football”

Reference Captions

"A dog playing a football in the grass"

Candidate Caption

BLEU
METEOR
ROUGE
CIDEr

Evaluation Metrics

"A dog holds a football"

"A dog is lying in the grass"

"A dog plays with a football"

Reference Captions

BLEU

Exact n-gram matches

METEOR

Exact n-gram + word stem + paraphrase + synonym matches

ROUGE

Longest common subsequence

CIDEr

TF-IDF weighted n-gram matches

Training Results

	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
Vanilla Encoder-Decoder	41.4	11.0	19.6	43.7	35.3
Attentive Encoder-Decoder	70.8	24.0	22.7	52.9	48.3
+ Beam Search	**72.7**	**30.3**	**24.1**	**54.4**	**87.6**
Show, Attend & Tell [1]	70.7	24.3	23.9	—	—

Models trained and evaluated on MSCOCO

[1] Xu, K. et. al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” (ICML 2015)

Experiment and Model Management

Evaluation Framework

Data Loading Interface

mongoDB®

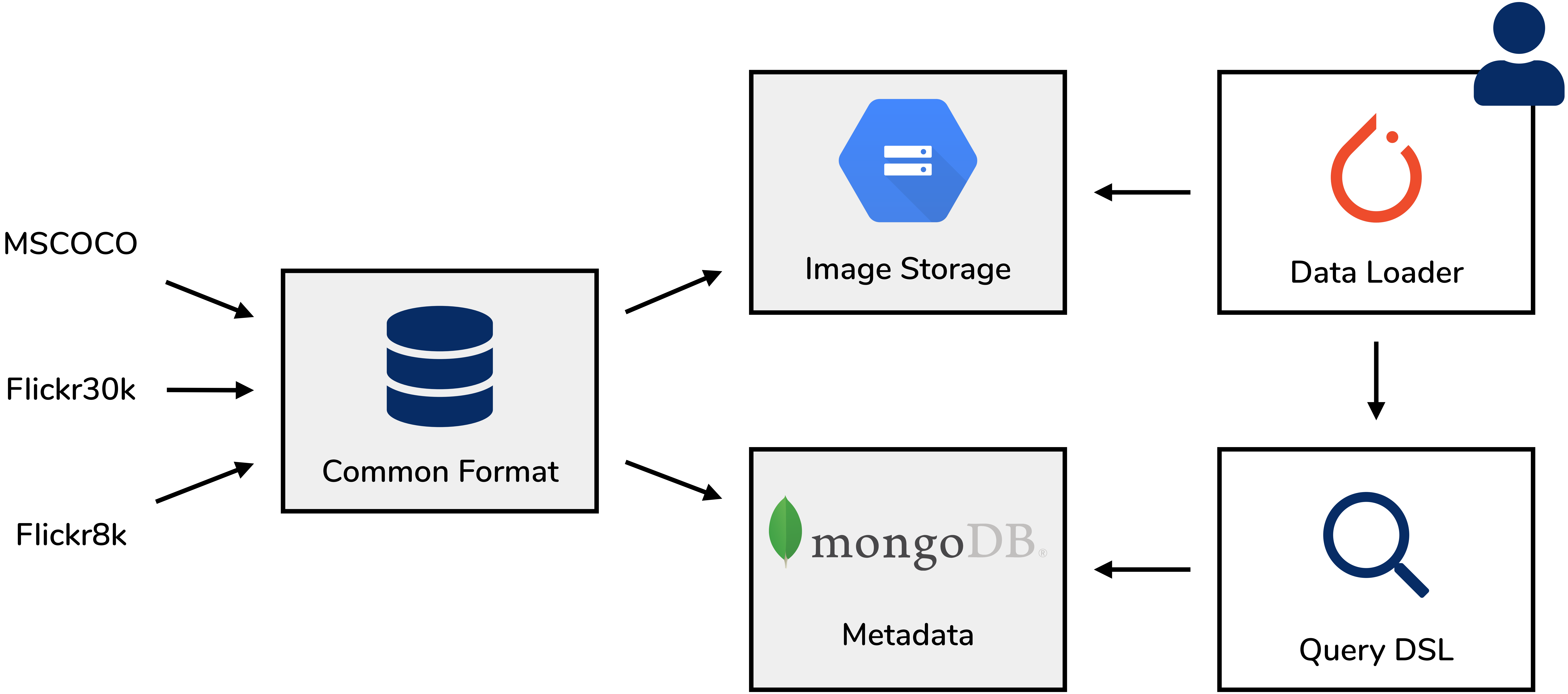
Training Interface

PyTorch

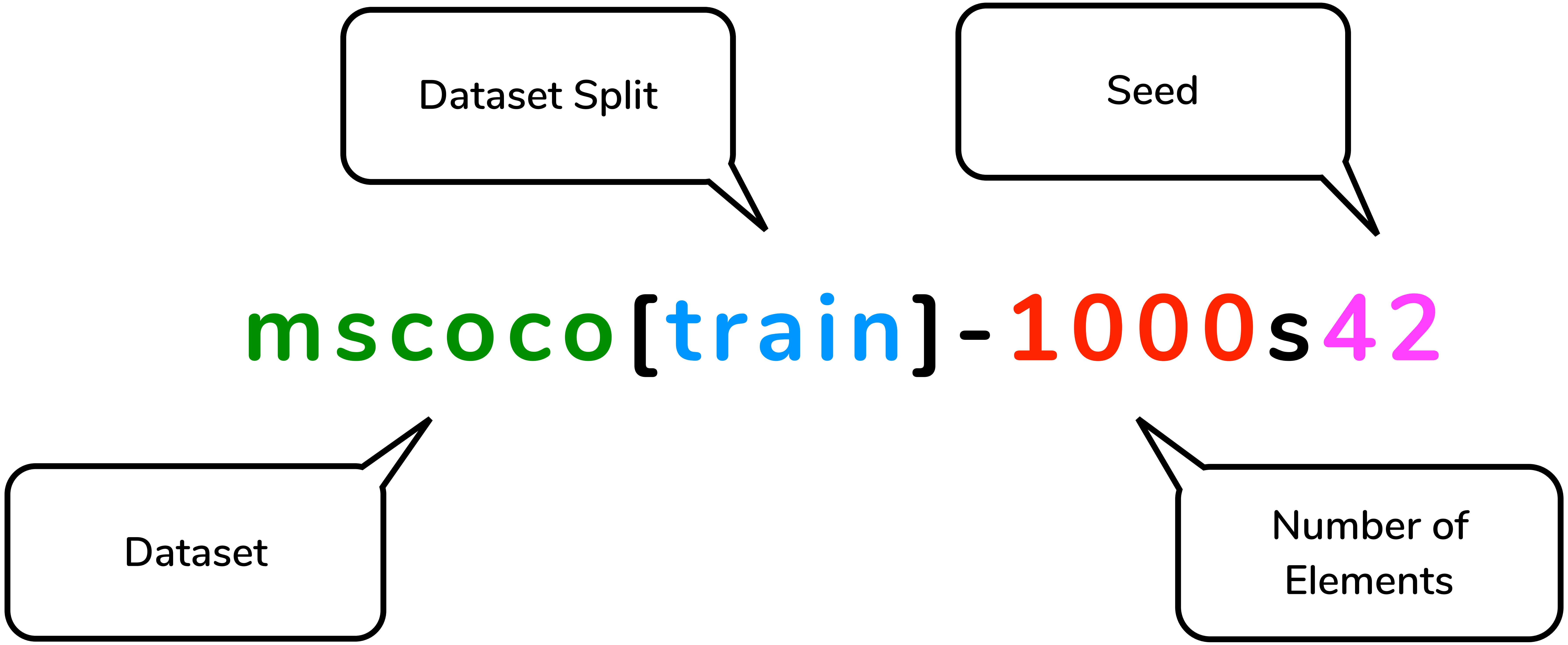
Experiment Tracking

mlflow

Data Loading Interface



Query DSL



Dataset Split

Seed

mscoco[train]-1000s42

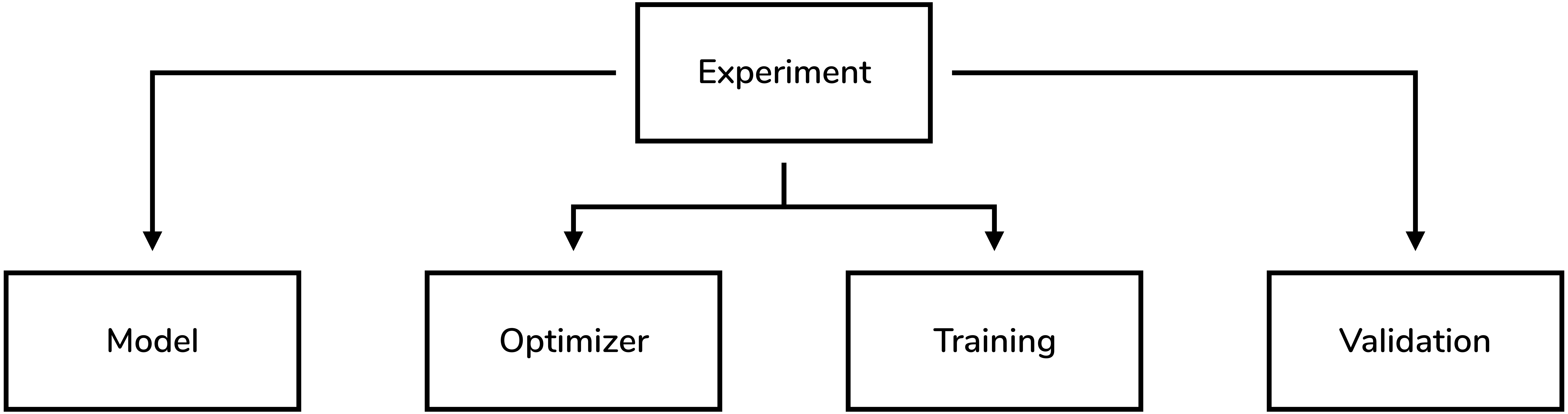
Dataset

Number of Elements

Training Interface

Training Script

YAML Configuration



Experiment

Model

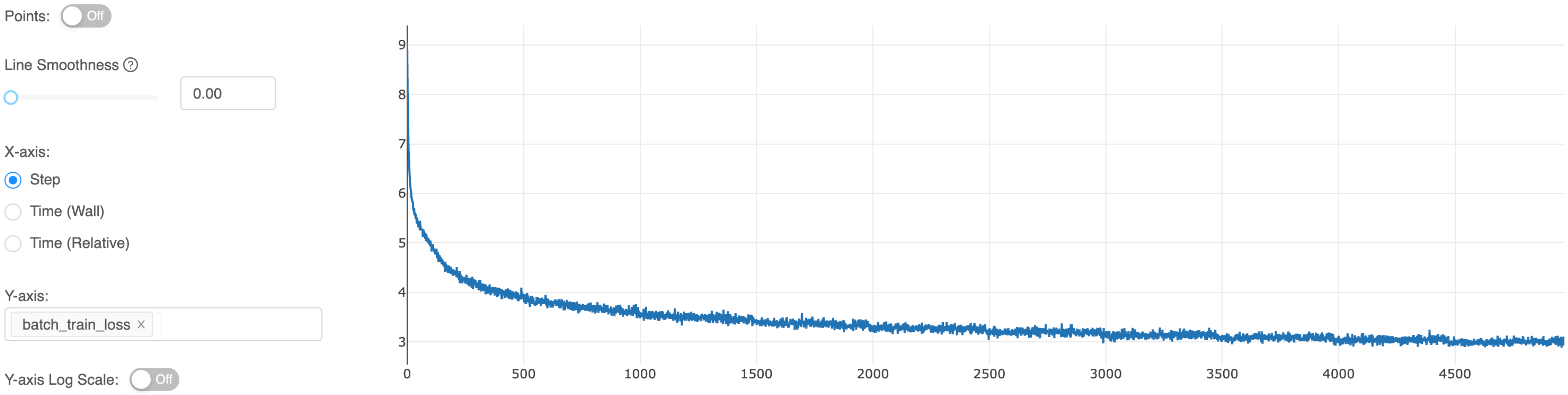
Optimizer

Training

Validation

Experiment Tracking

mlflow



Points: Off

Line Smoothness ⓘ 0.00

X-axis:

● Step

○ Time (Wall)

○ Time (Relative)

Y-axis: batch_train_loss ×

Y-axis Log Scale: Off

Deployment

Production-Grade Deployment

Reliability

Security

Logging

Accurate Results

Continuous Model Improvement

Friendly UI

Responsive UI

Fault Tolerance

High Availability

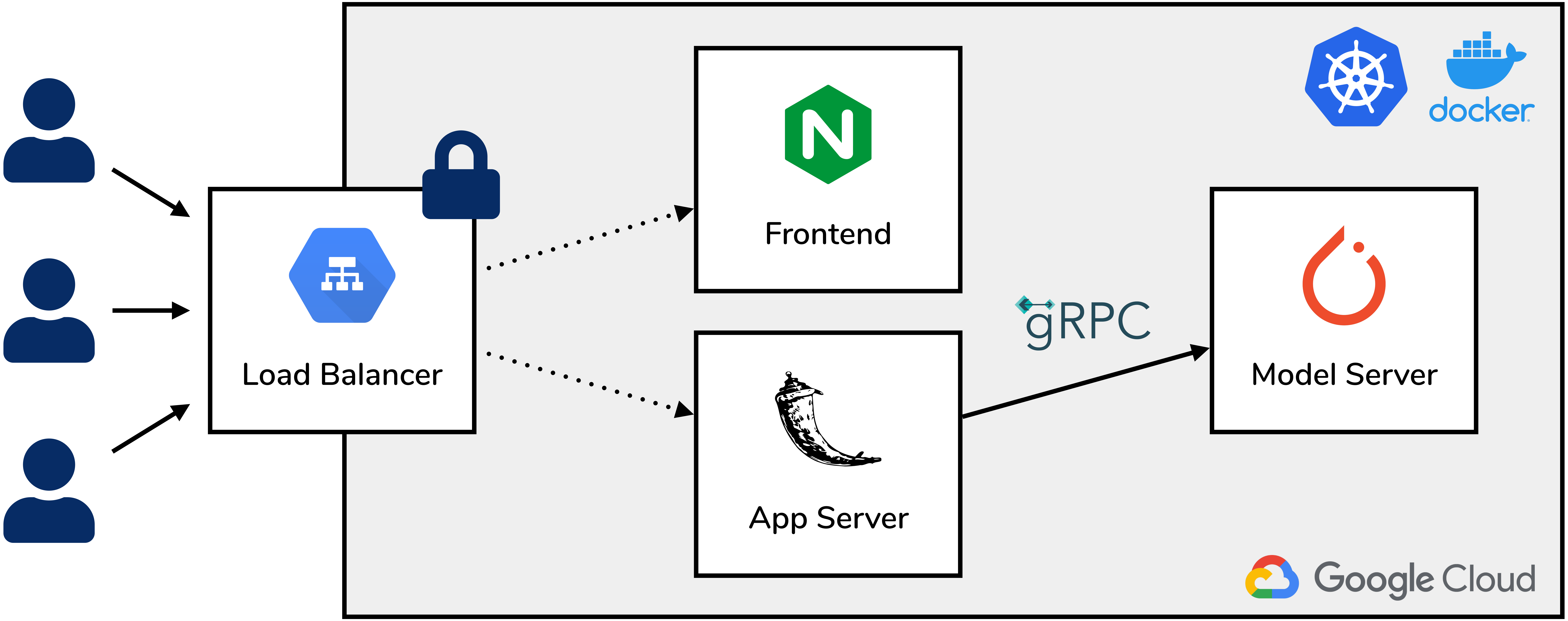
Monitoring

Auto-Scalability

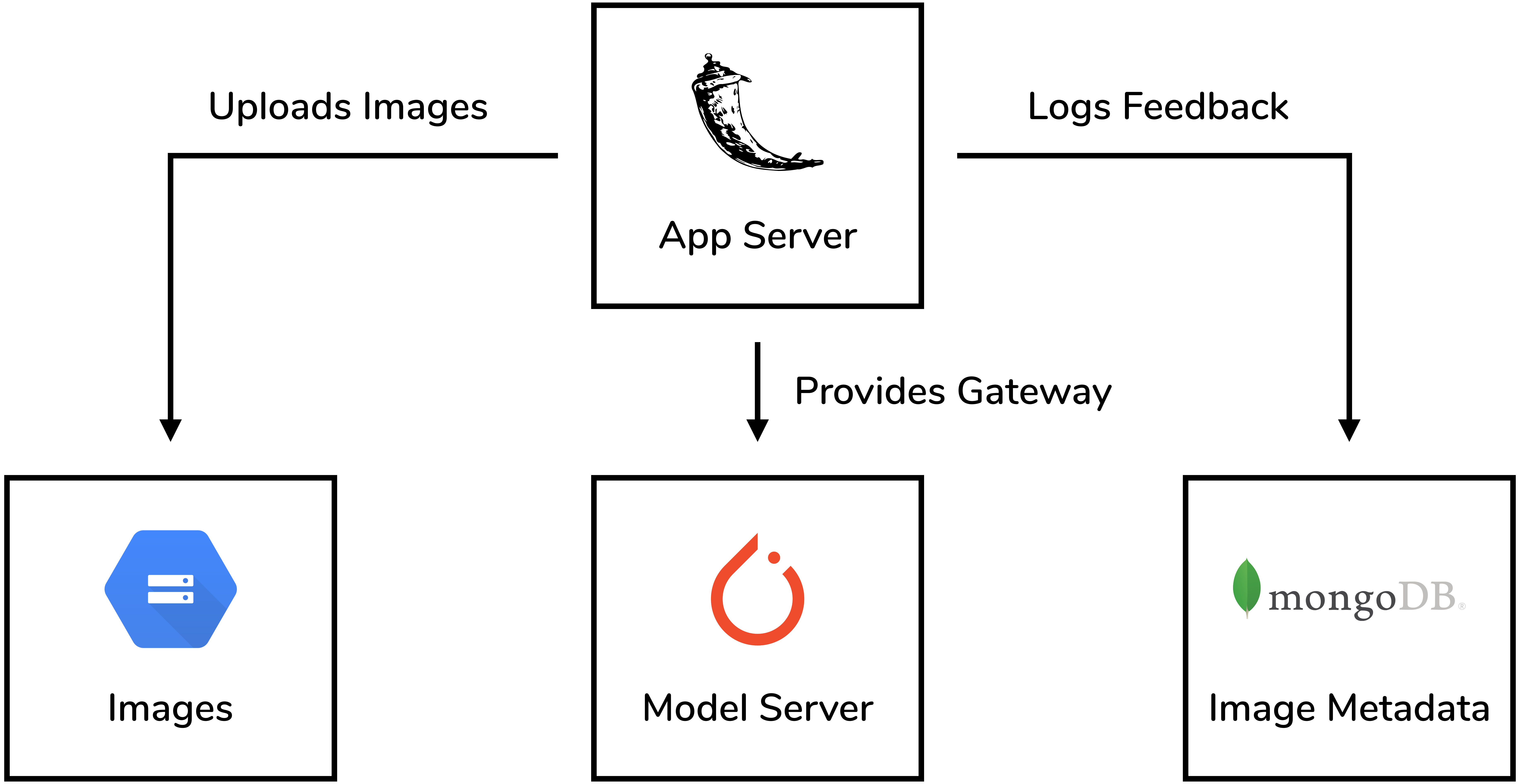
REST API

Fast Response Times

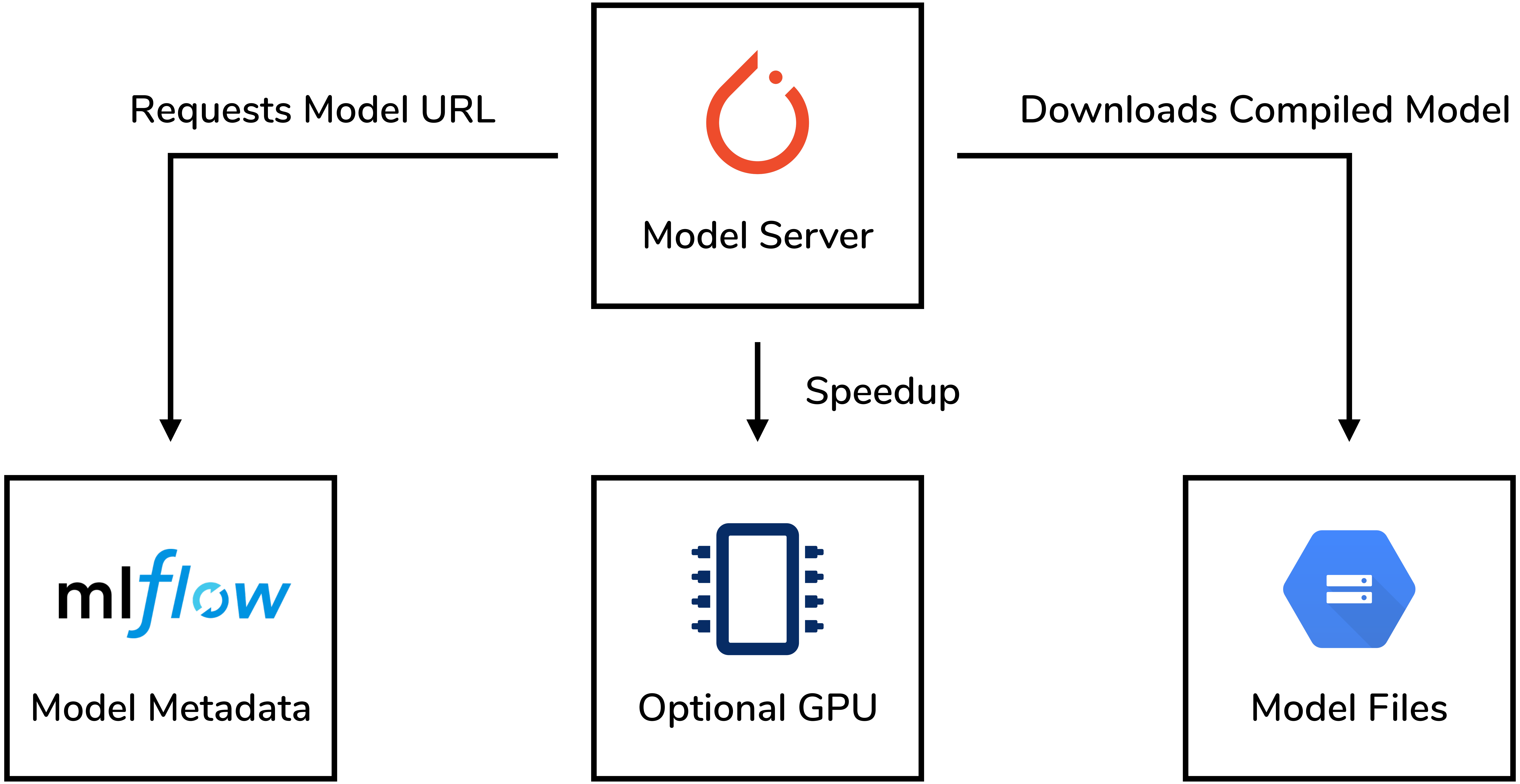
System Architecture



Application Server



Model Server



Requests Model URL

Model Server

Downloads Compiled Model

Speedup

mlflow

Model Metadata

Optional GPU

Model Files

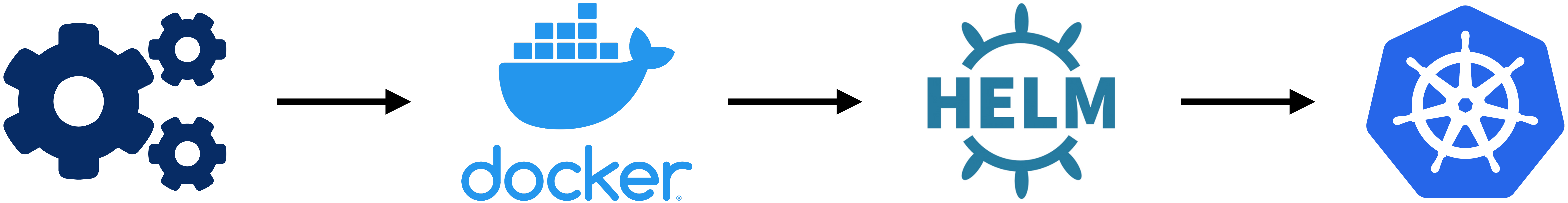
Autoscaling

Demand-Based Number of Replicas

Independent Scaling

Automatic Load Balancing

Continuous Integration & Deployment



Run Automated Tests

Build & Tag Docker Image

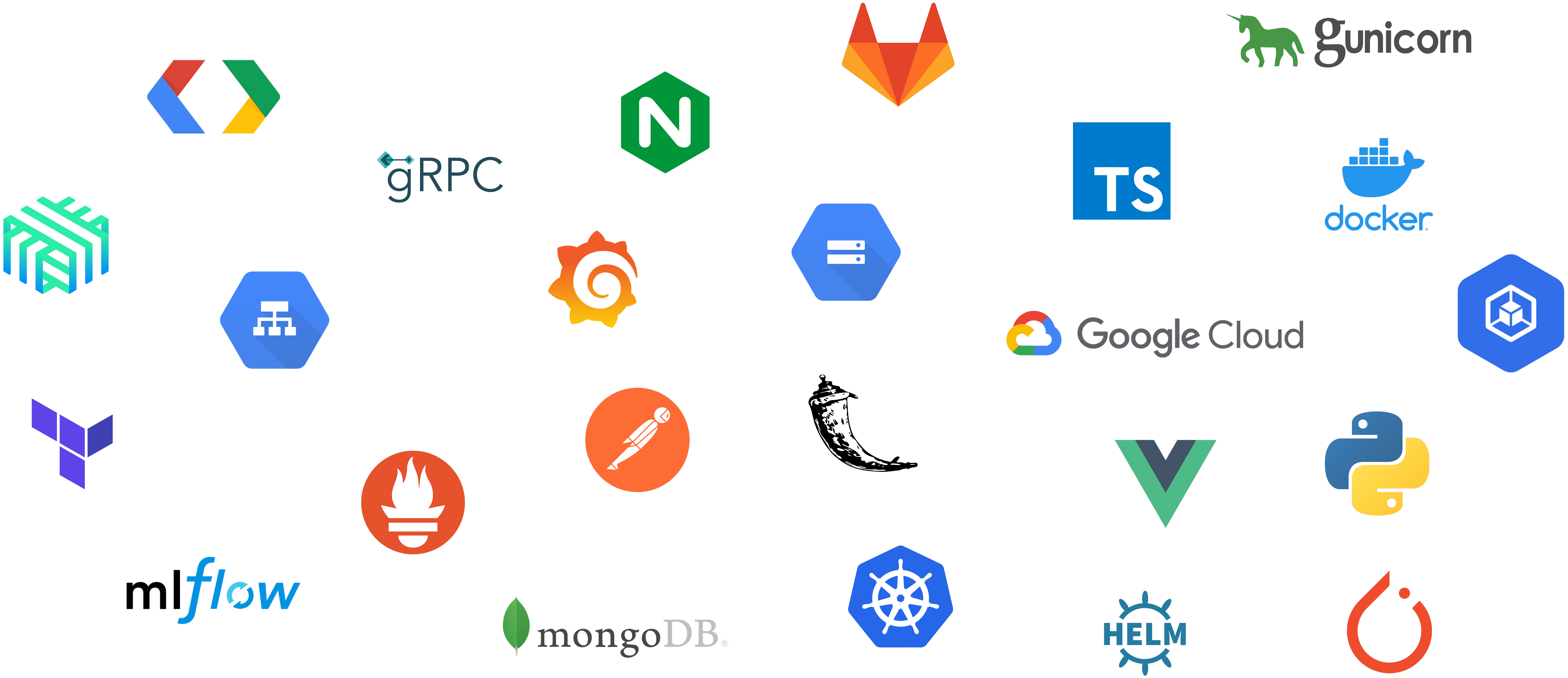
Bundle Required Resources

Deploy Resources to Production

Monitoring

Linkerd + Prometheus + Grafana

Tools & Technologies



Thanks for your attention!

Let’s see it in action…

https://dilab.inovex.de/