Data Science on Prometheus Metrics

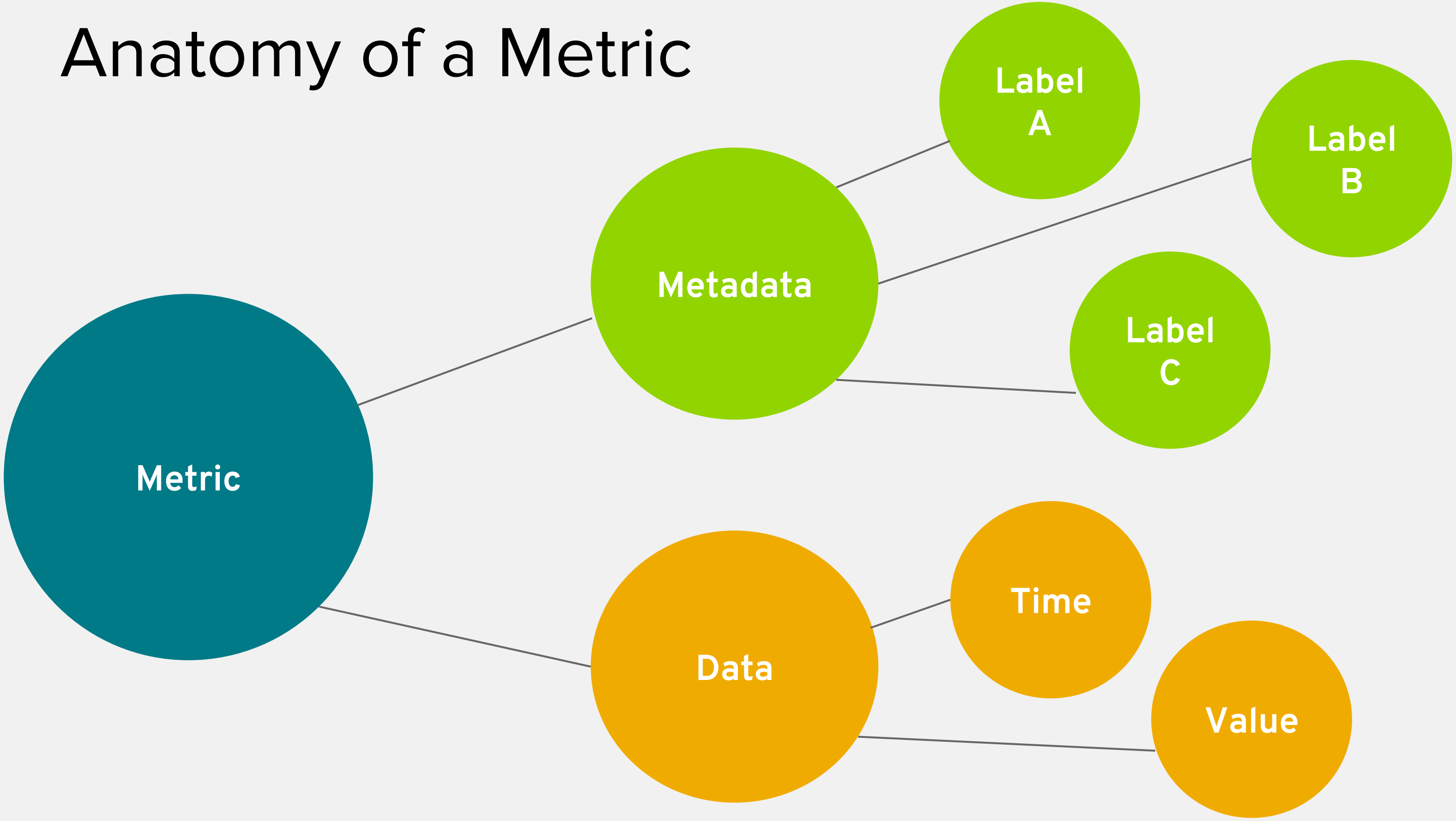
Natasha Frumkin
AIOPS Intern
August 21, 2018

Outline

1. Metadata Analysis
2. Time Series Forecasting
3. Model Comparison

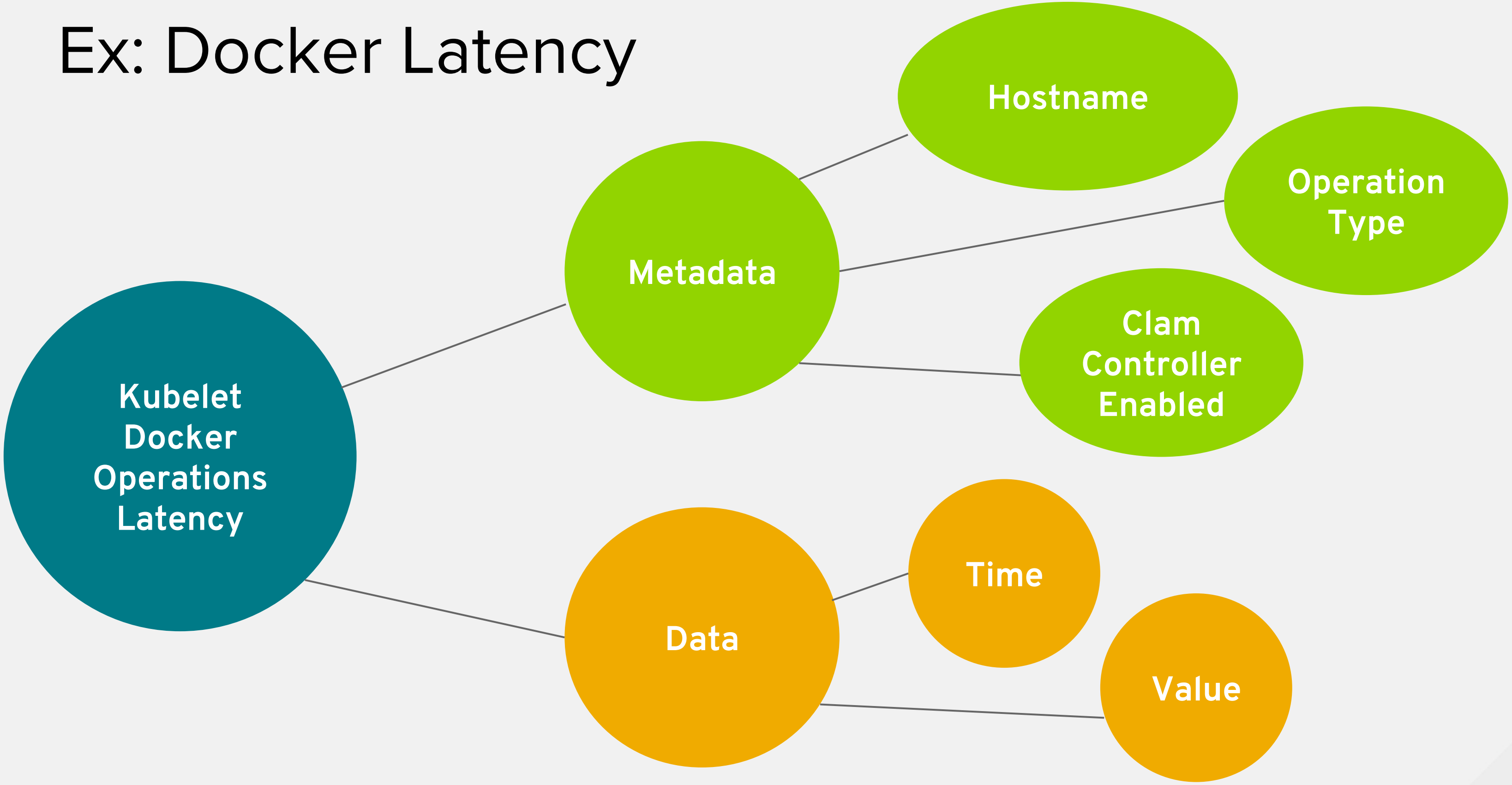
Metadata Analysis

Anatomy of a Metric

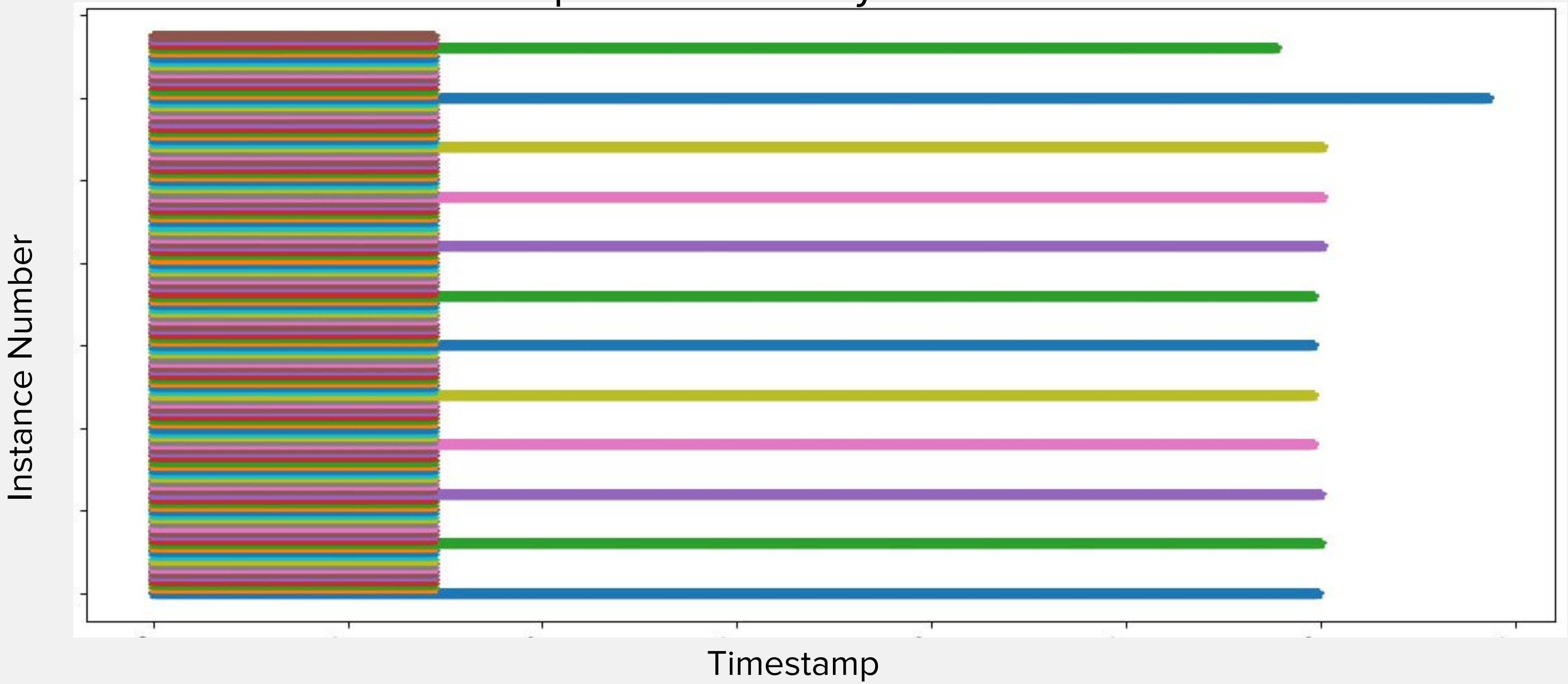


redhat.

Ex: Docker Latency

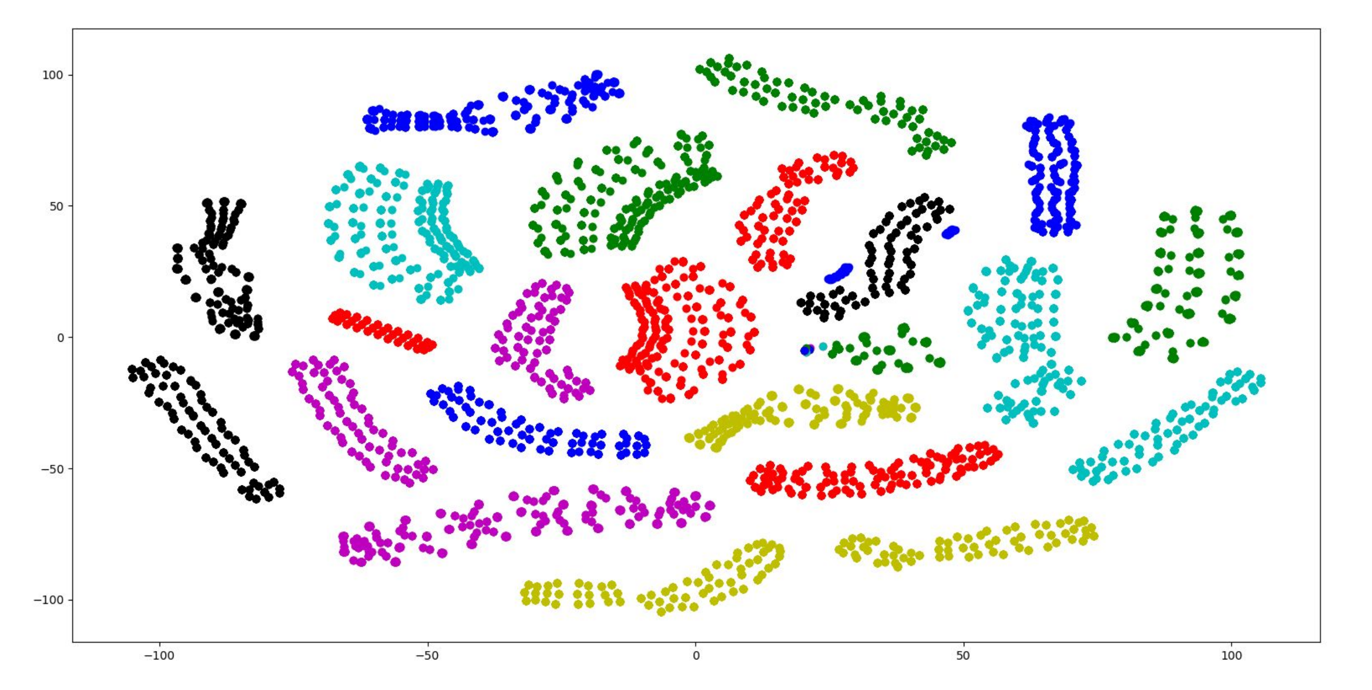


Metric: kubelet docker operation latency microseconds



0	ip-172-31-78-254.us-east-2.compute.internal
1	ip-172-31-73-251.us-east-2.compute.internal
2	ip-172-31-65-74.us-east-2.compute.internal
3	ip-172-31-74-247.us-east-2.compute.internal
4	ip-172-31-76-218.us-east-2.compute.internal
5	ip-172-31-75-193.us-east-2.compute.internal
6	ip-172-31-71-195.us-east-2.compute.internal
7	ip-172-31-69-53.us-east-2.compute.internal

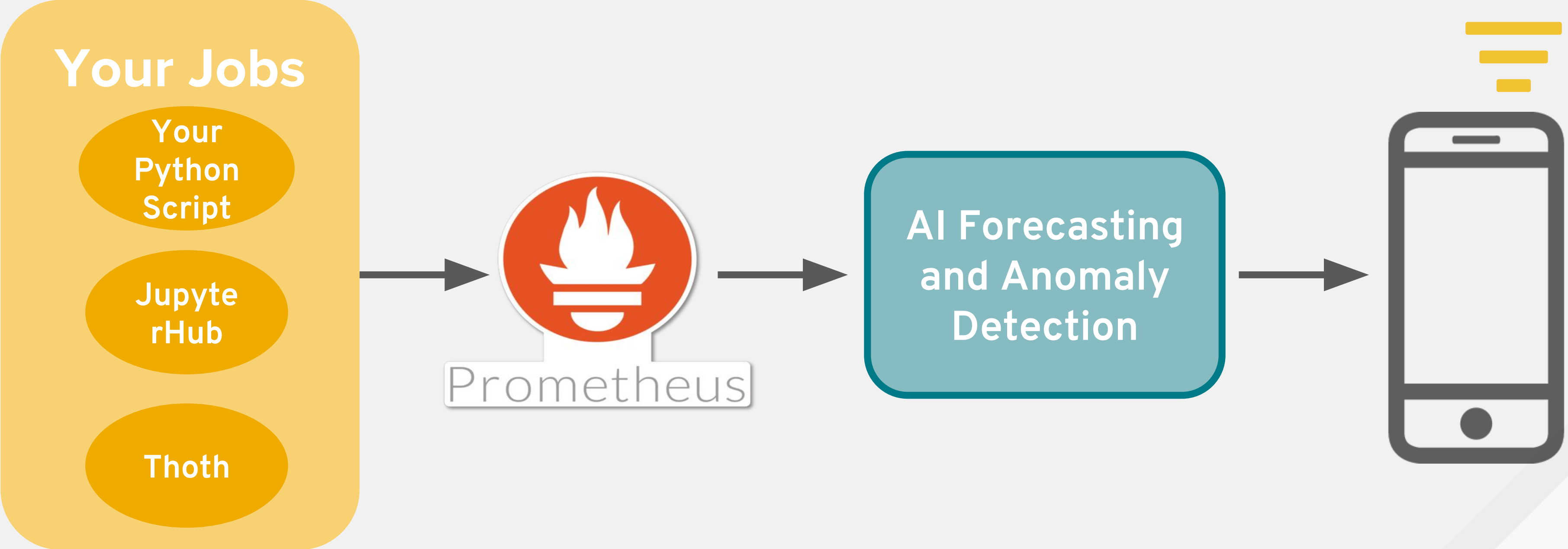
T-SNE Embedding of Metric Metadata



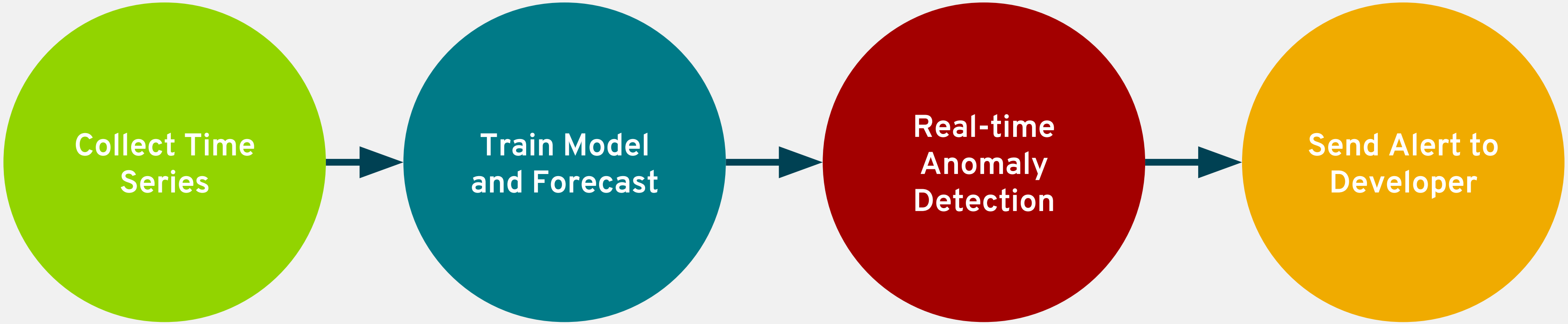
Time Series Forecasting and Anomaly Detection

redhat.

Goal: to provide automatic alerting to developers when there are anomalies in Prometheus data



The Data Transfer Pipeline



Data Processing

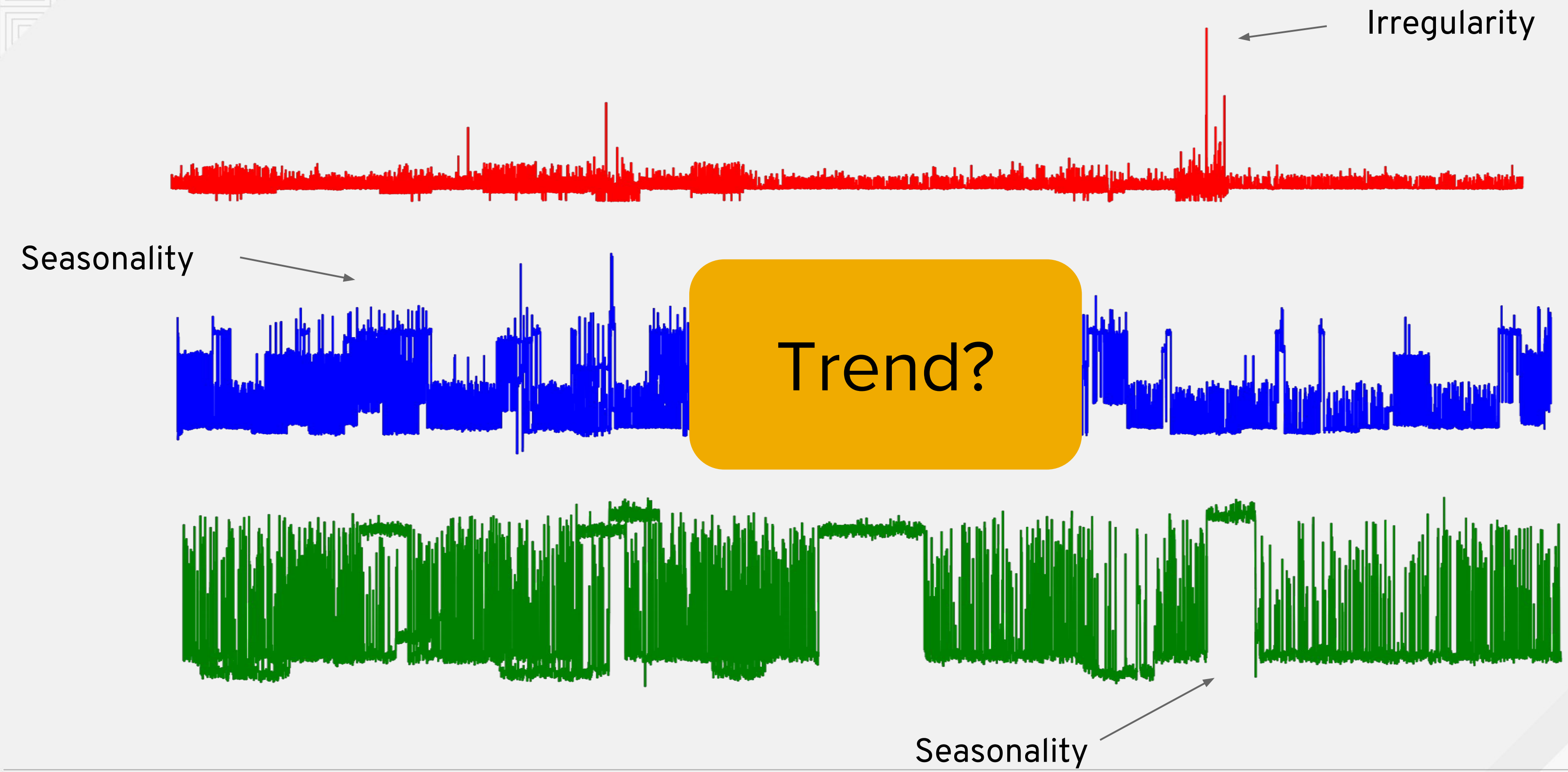
Bz2 File
Text file of Json packets

Pickle File
Dict of DataFrames
Key = str(metadata)
Value = Pandas DataFrame

{"metric": {"__name__": "kubelet_docker_operations_latency_microseconds", "beta_kubernetes_io_arch": "amd64", "beta_kubernetes_io_fluentd_ds_ready": "true", "beta_kubernetes_io_instance_type": "m4.xlarge", "beta_kubernetes_io_os": "linux", "clam_controller_enabled": "True", "failure_domain_beta_kubernetes_io_region": "us-east-2"}, "values": [[1518307199, "12844"], [1518308638, "13212"], [1518310077, "13830"], [1518311516, "13395"], [1518312955, "16546"], [1518314394, "15174"], [1518315833, "14455"], [1518317272, "12949"], [1518318711, "13439"], [1518320150, "14386"], [1518321589, "12447"], [1518323028, "15947"], [1518324467, "14893"], [1518325906, "14096"], [1518327345, "14735"], [1518328784, "12969"], [1518330223, "14067"], [1518331662, "16286"], [1518333101, "14008"], [1518334540, "12923"], [1518335979, "11888"], [1518337418, "12263"], [1518338857, "11751"], [1518340296, "13534"], [1518341735, "15522"], [1518343174, "14912"], [1518390661, "12235"], [1518392100, "14209"], [1518393539, "15757"]]}

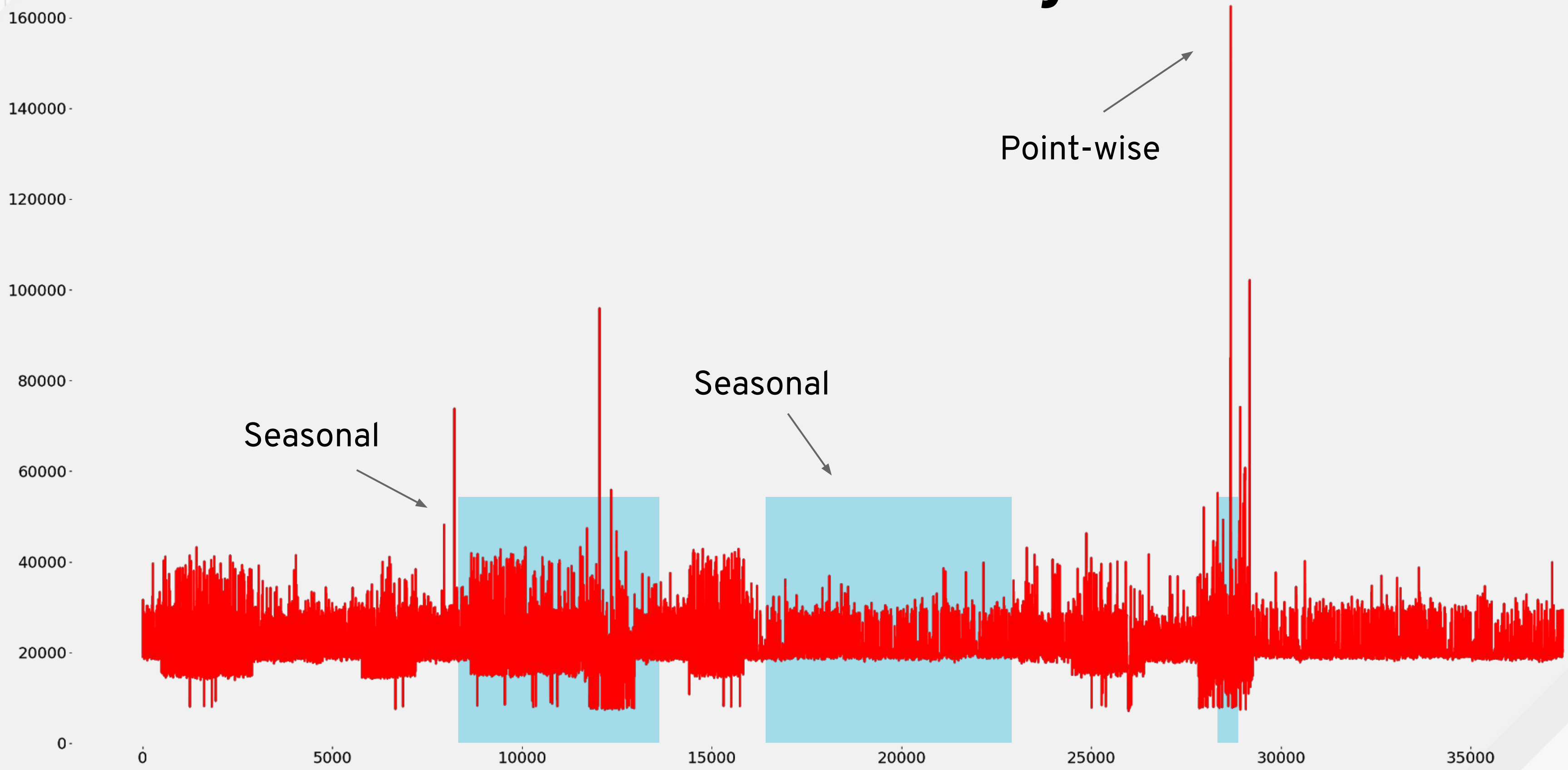
Timestamps	Values
03-03-18 12:30:15	32193
03-03-18 12:31:15	33210
03-03-18 12:32:15	32184

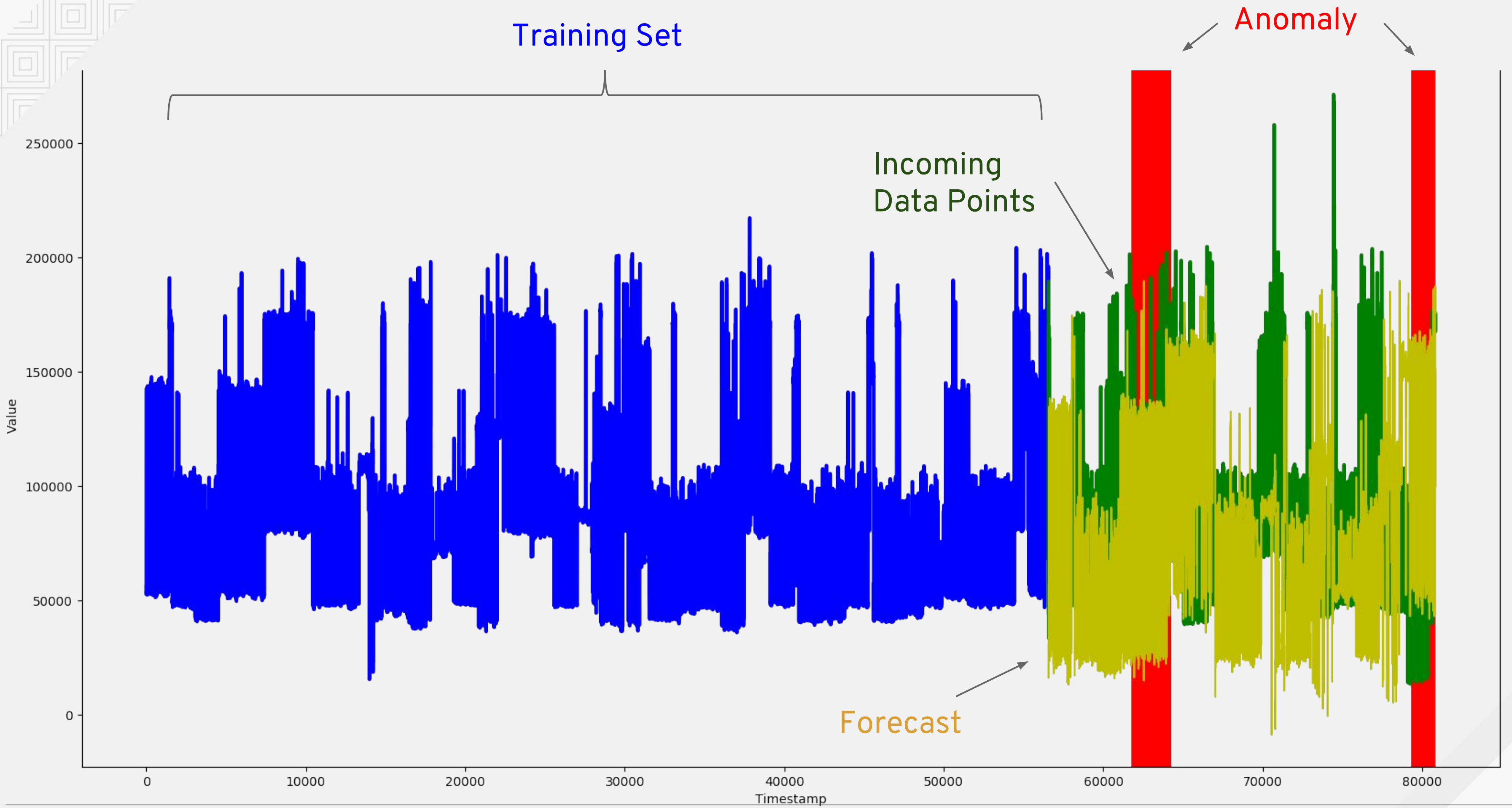
Metric: http request duration



Fourier Extrapolation

Problem: What is an anomaly?

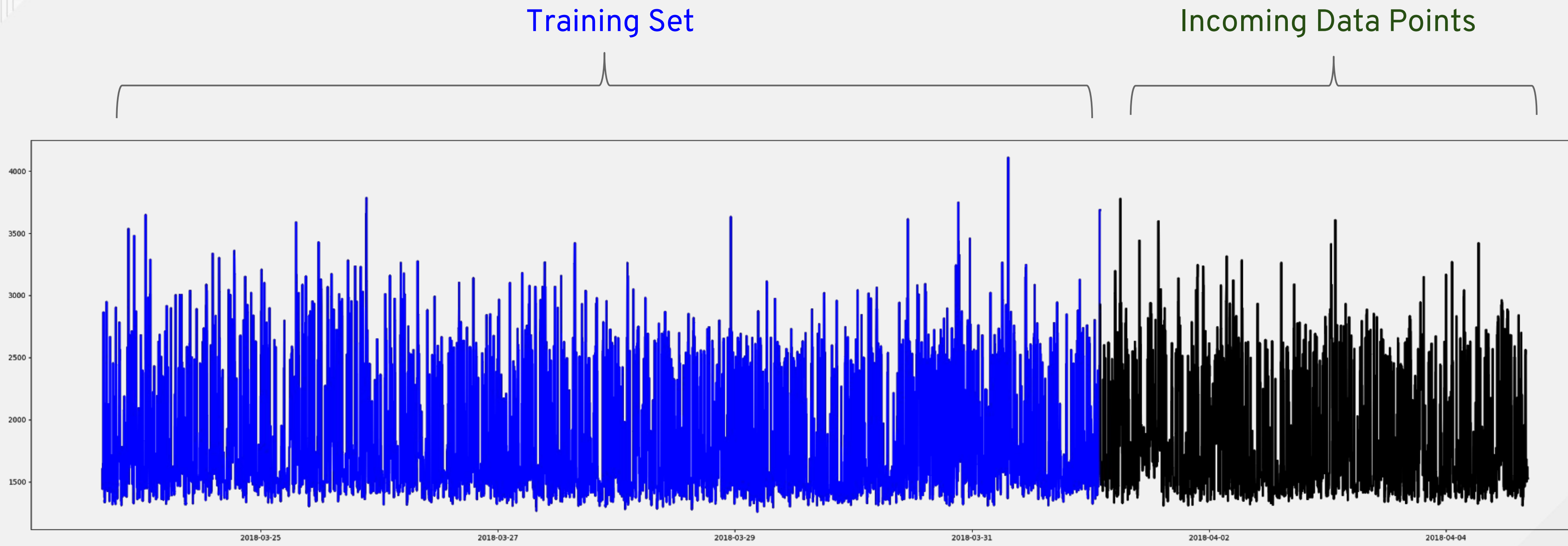


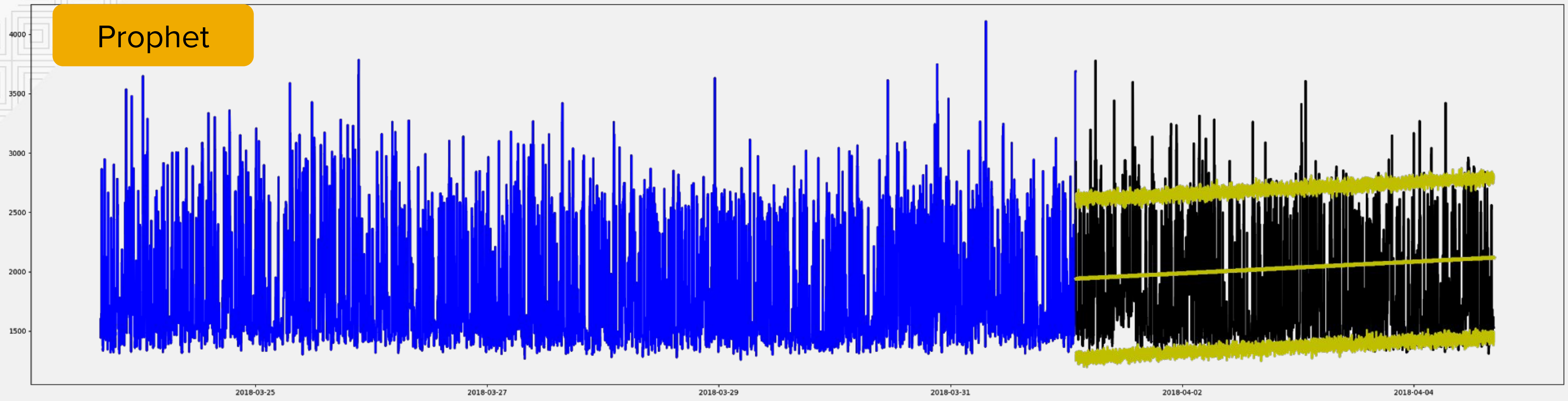

Training Set
Anomaly
Incoming
Data Points
Forecast
250000
200000
150000
100000
50000
0
Value
0
10000
20000
30000
40000
50000
60000
70000
80000
Timestamp

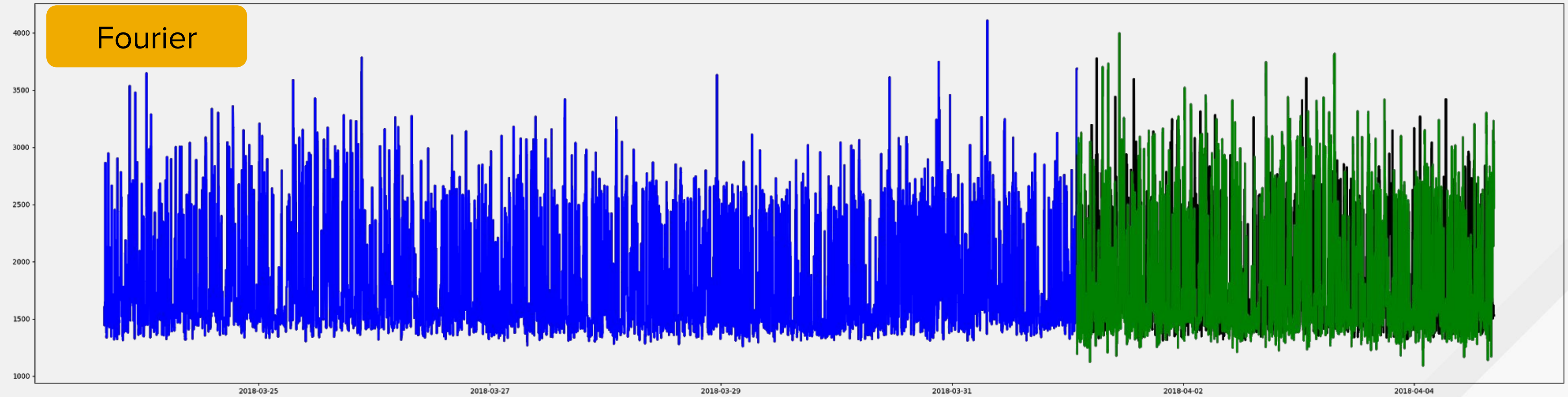
Forecast Comparison

Prophet vs. Fourier

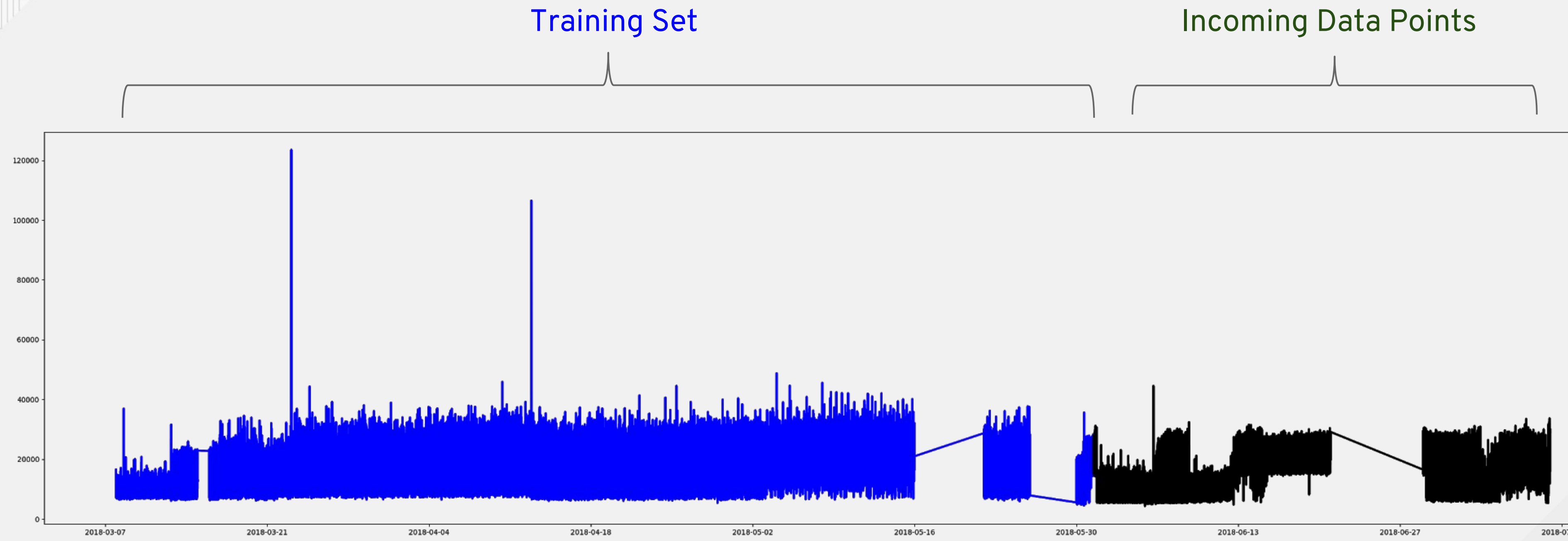
Metric: http request duration

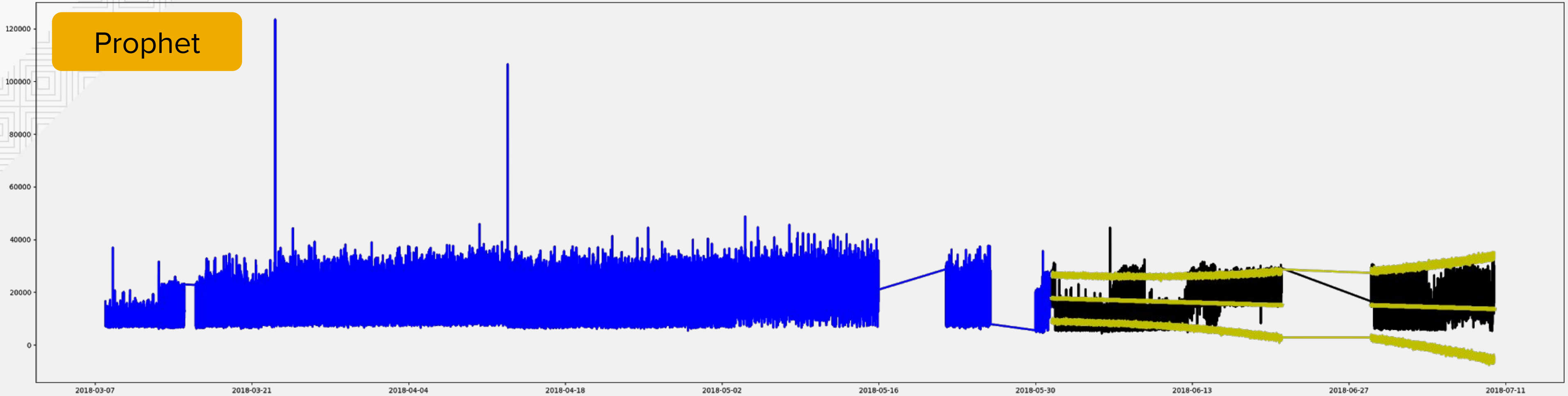


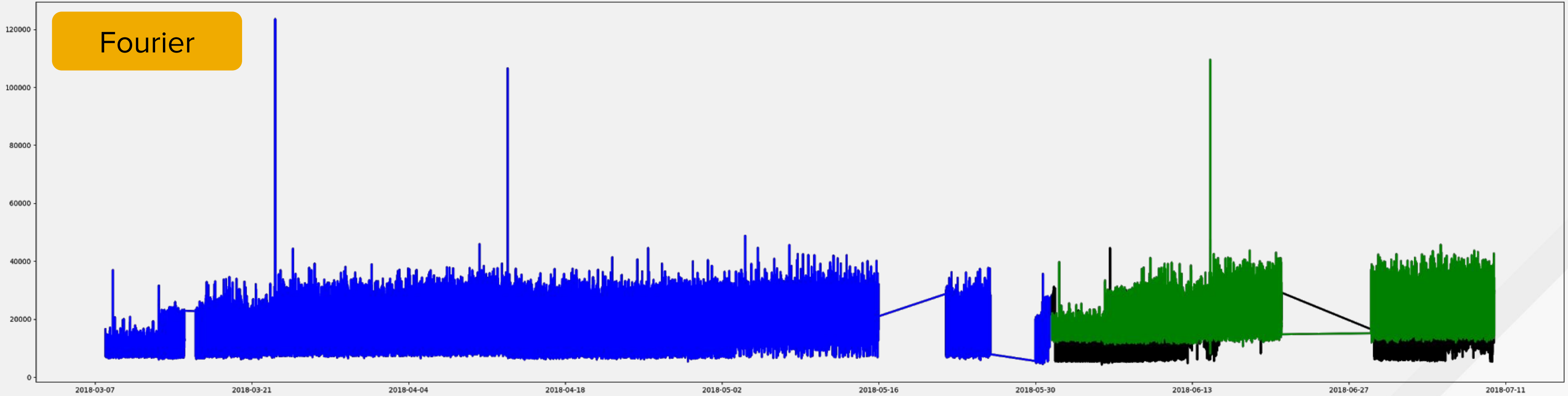

Prophet
4000
3500
3000
2500
2000
1500
2018-03-25
2018-03-27
2018-03-29
2018-03-31
2018-04-02
2018-04-04


Fourier
4000
3500
3000
2500
2000
1500
1000
2018-03-25
2018-03-27
2018-03-29
2018-03-31
2018-04-02
2018-04-04

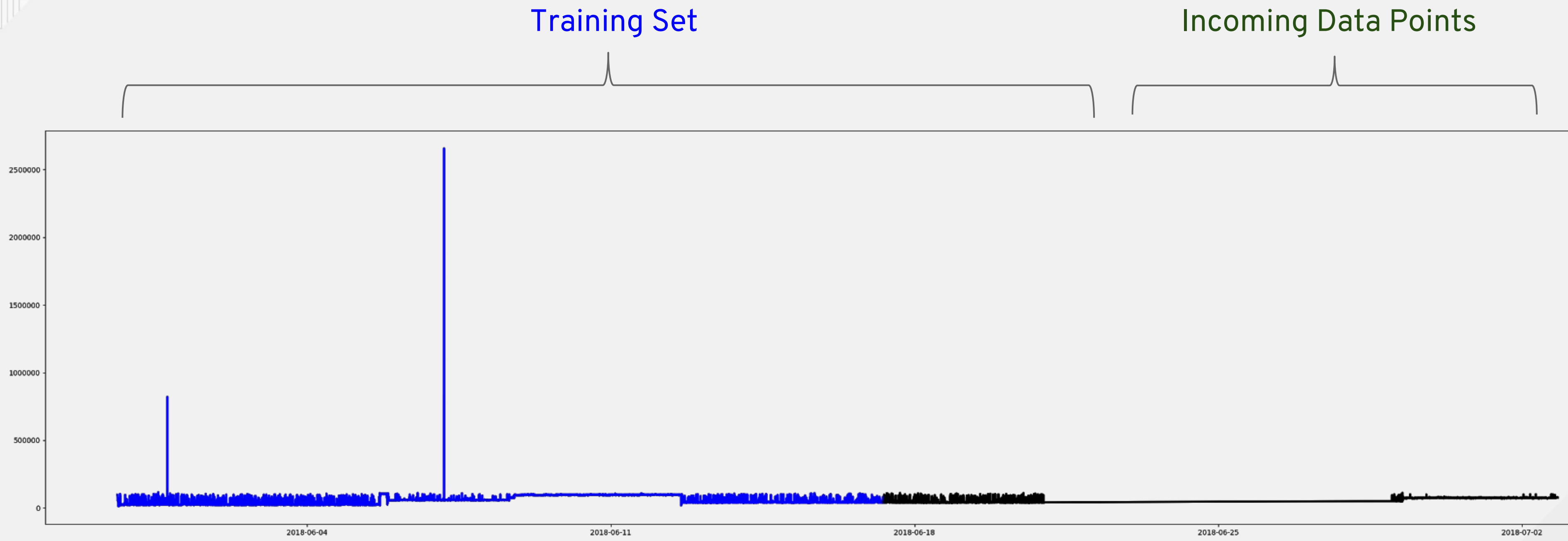
Metric: http request duration

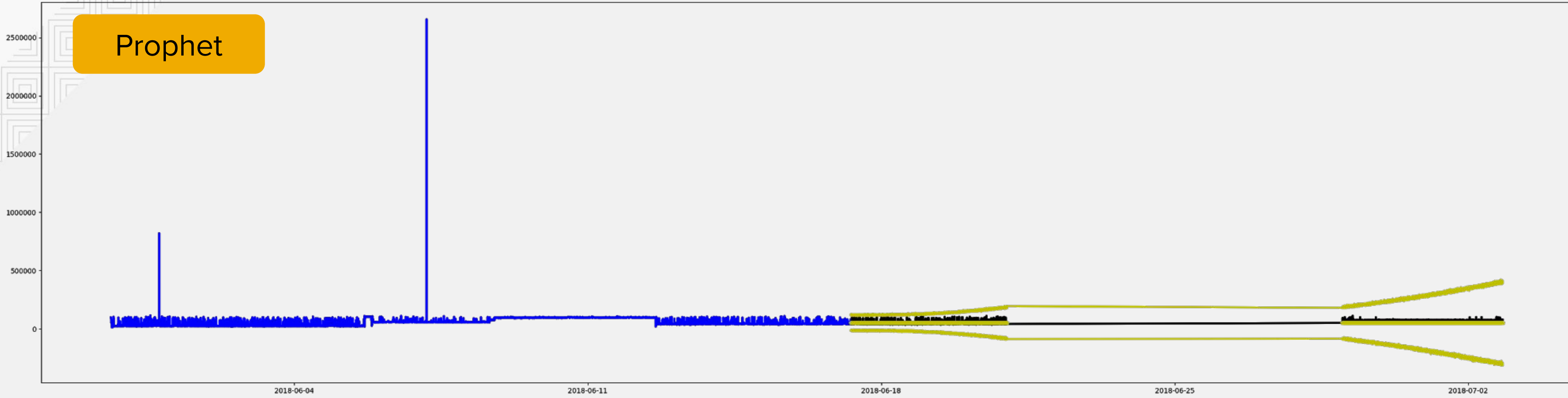



Prophet
120000
100000
80000
60000
40000
20000
0
2018-03-07
2018-03-21
2018-04-04
2018-04-18
2018-05-02
2018-05-16
2018-05-30
2018-06-13
2018-06-27
2018-07-11

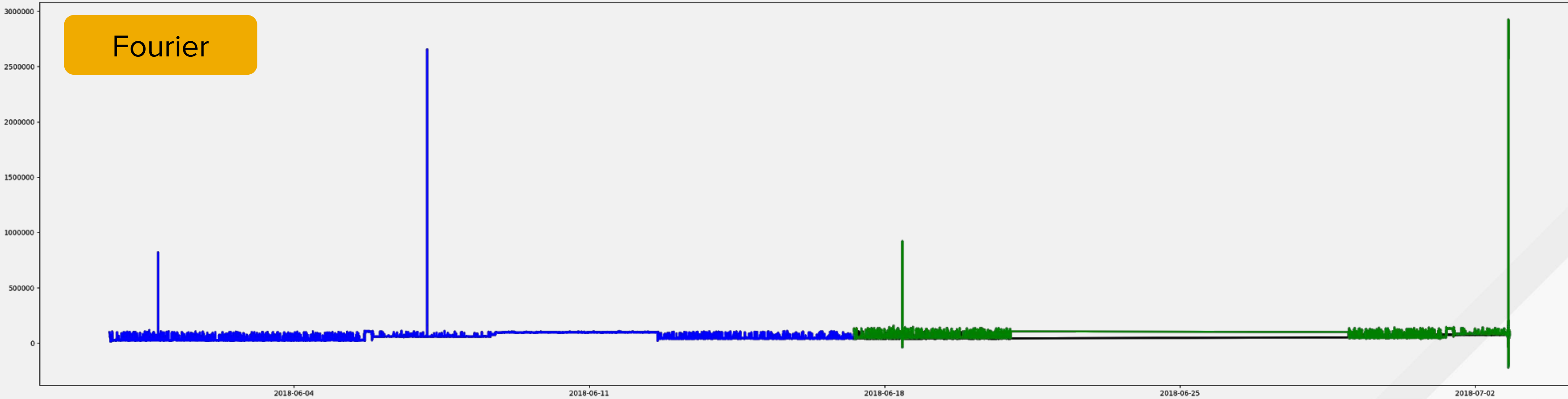

Fourier
120000
100000
80000
60000
40000
20000
0
2018-03-07
2018-03-21
2018-04-04
2018-04-18
2018-05-02
2018-05-16
2018-05-30
2018-06-13
2018-06-27
2018-07-11

Metric: http request duration



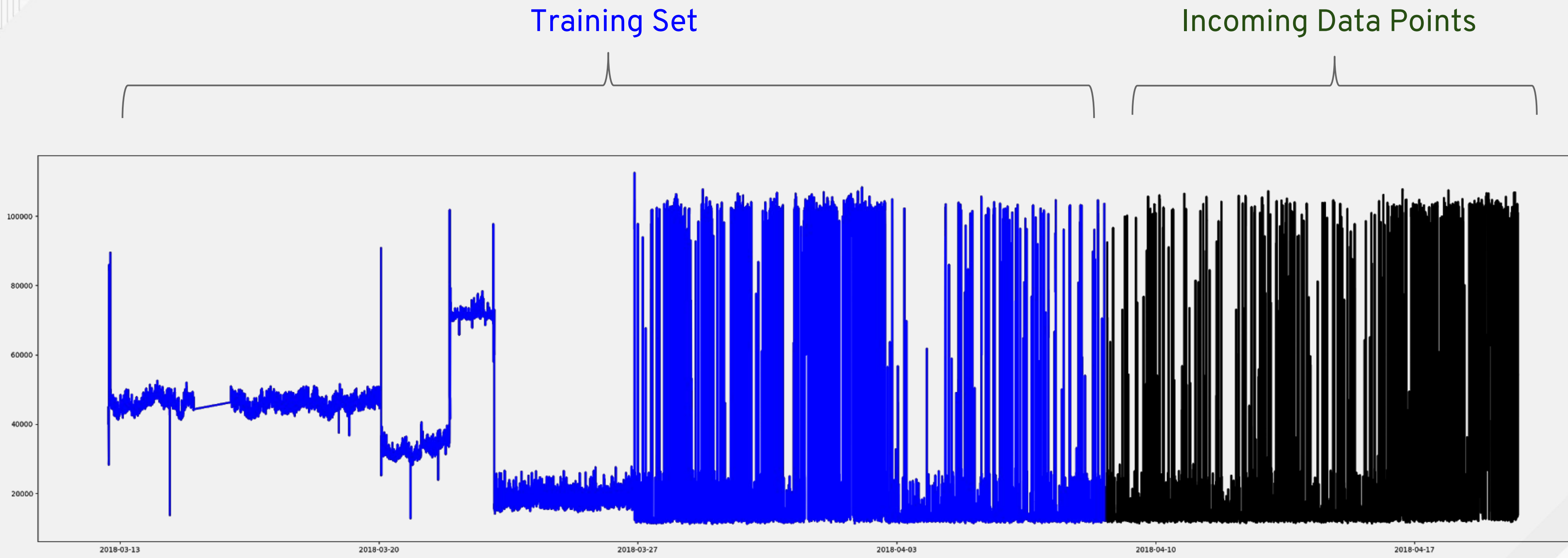


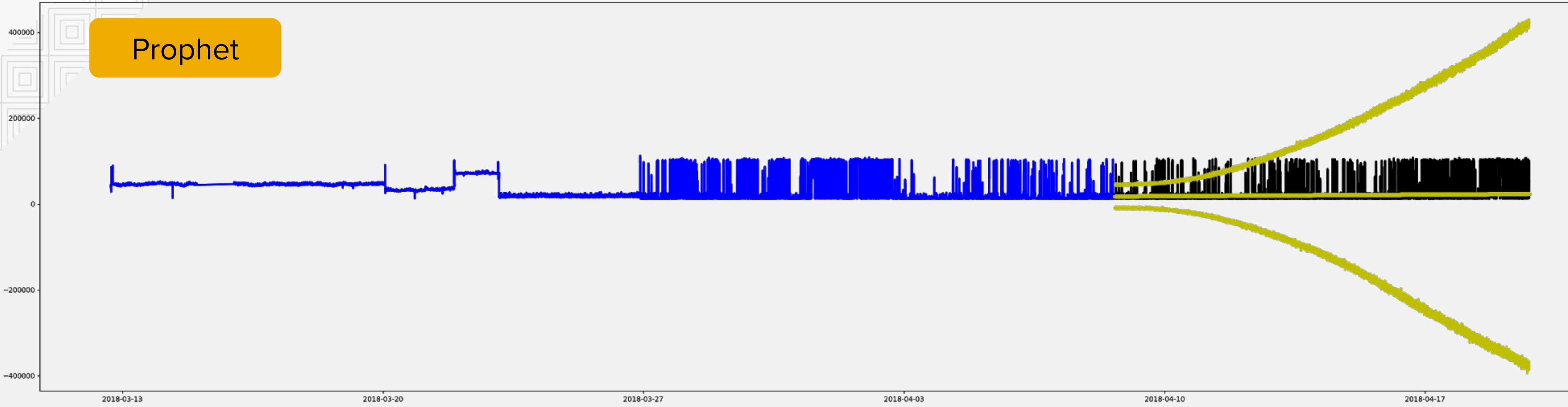
Prophet
2500000
2000000
1500000
1000000
500000
0
2018-06-04
2018-06-11
2018-06-18
2018-06-25
2018-07-02



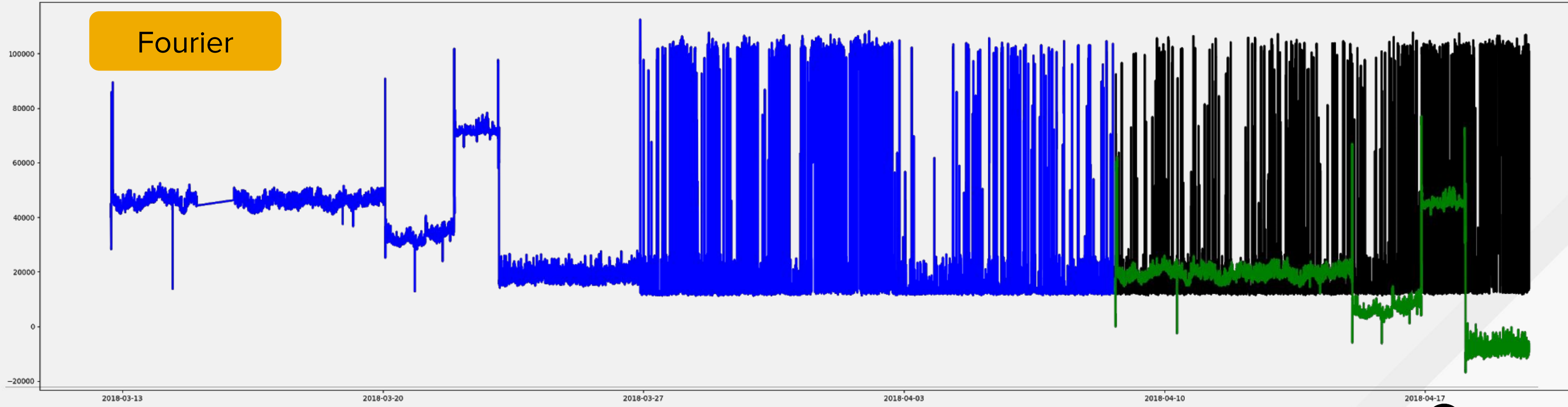
Fourier
3000000
2500000
2000000
1500000
1000000
500000
0
2018-06-04
2018-06-11
2018-06-18
2018-06-25
2018-07-02

Metric: http request duration

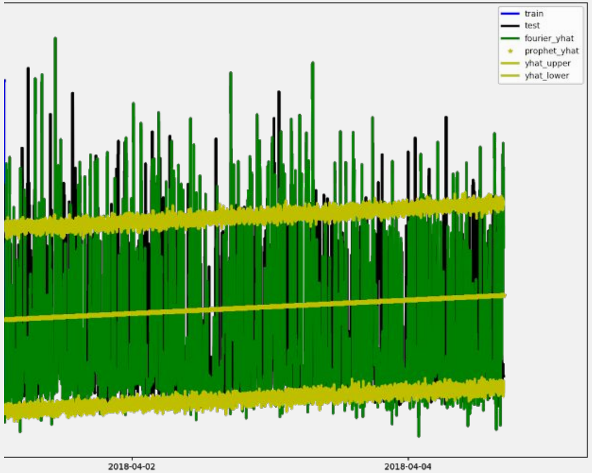


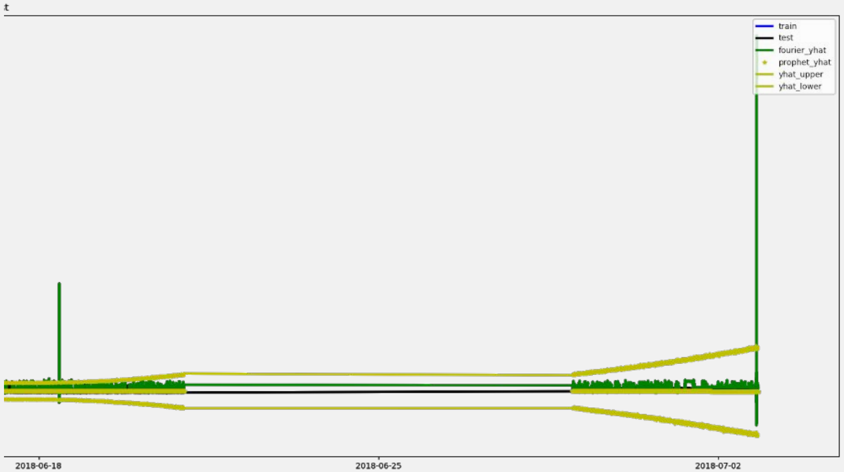


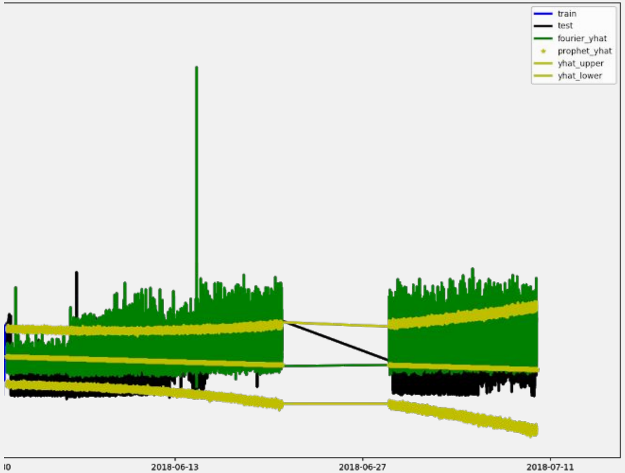
Prophet
400000
200000
0
−200000
−400000
2018-03-13
2018-03-20
2018-03-27
2018-04-03
2018-04-10
2018-04-17

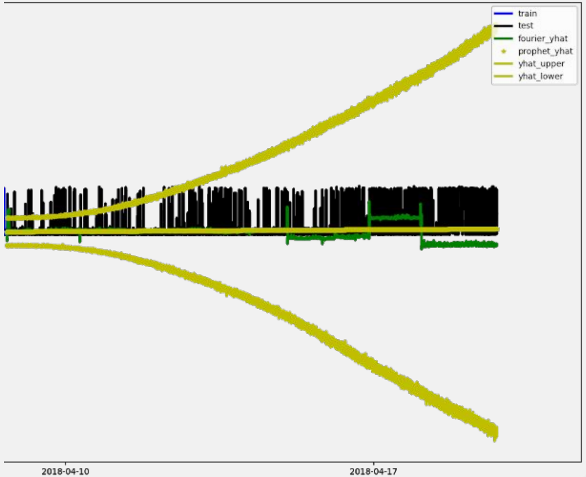


Fourier
100000
80000
60000
40000
20000
0
−20000
2018-03-13
2018-03-20
2018-03-27
2018-04-03
2018-04-10
2018-04-17


train
test
fourier_yhat
prophet_yhat
yhat_upper
yhat_lower
2018-04-02
2018-04-04


train
test
fourier_yhat
prophet_yhat
yhat_upper
yhat_lower
2018-06-18
2018-06-25
2018-07-02


train
test
fourier_yhat
prophet_yhat
yhat_upper
yhat_lower
2018-06-13
2018-06-27
2018-07-11


train
test
fourier_yhat
prophet_yhat
yhat_upper
yhat_lower
2018-04-10
2018-04-17

Summary of our Techniques

Exponential Smoothing

ARIMA models

Fourier Analysis

Prophet models

RNNs

training models

Thresholding

Gaussian Tail Probability

Accumulators

anomaly detection rules

redhat.

Main Takeaways

Metadata configurations are constantly changing.

Prophet vs. Fourier vs. RNNs Which features do we care about?

Anomaly detection requires finesse. Need to test parameters.

Next Steps

For which anomalies do we send alerts? Threshold needed.

Dive deeper into more complex models. Ensemble methods?

Scalability? Which time series do we choose to monitor?

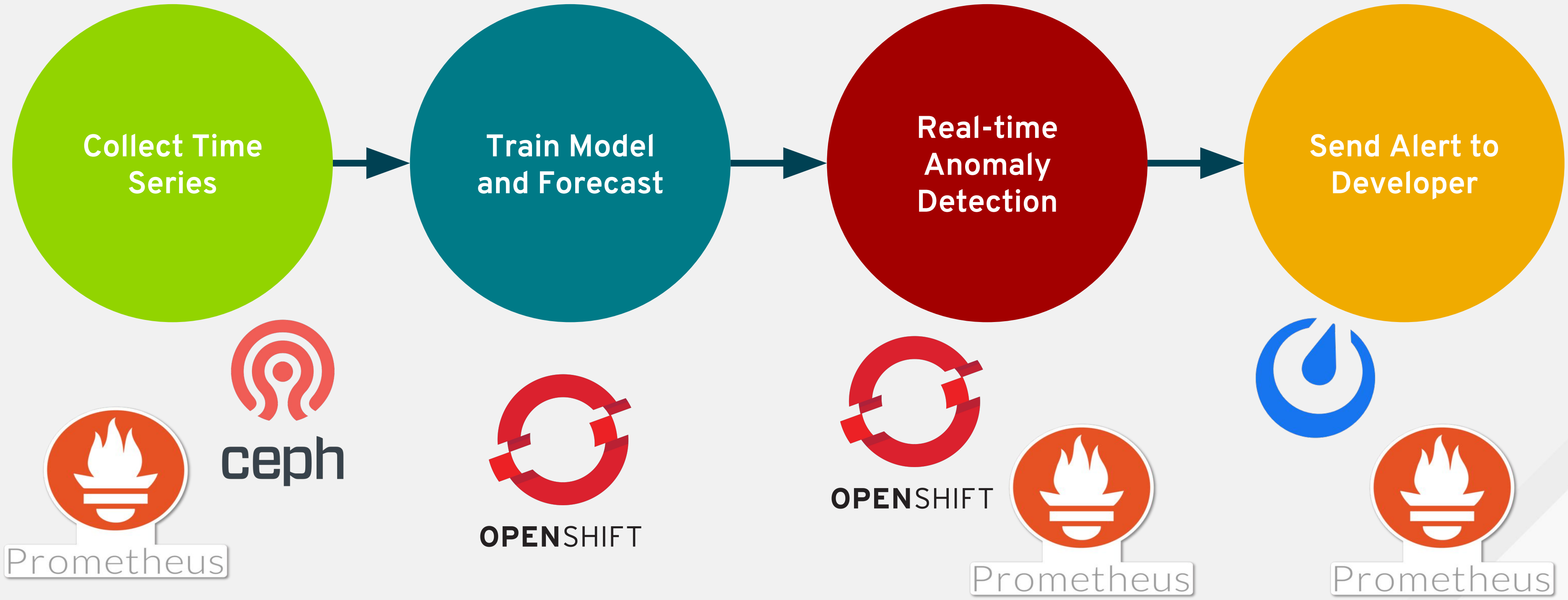
THANK YOU

Notebooks: Gitlab AICOE/jupyter-notebooks
Documentation and Scripts: github.com/nfrumkin/forecast-prometheus

Challenges with Prometheus Dataset

- **Data comes from multiple sources**
  - Need to explore correct time series filtering
- **Data has holidays and season**
  - Leverage known smoothing and decomposition techniques
- **Wide range of metric types and behavior**
  - Possibly apply different AD techniques for different series
- **Training Data has hidden anomalies and dropouts**
  - Find a way to accurately prepare historical data for training

The Data Transfer Pipeline



redhat.