
BIG DATA
ANALYTICS
Final presentation
25.02.2019
TU Berlin

# Agenda

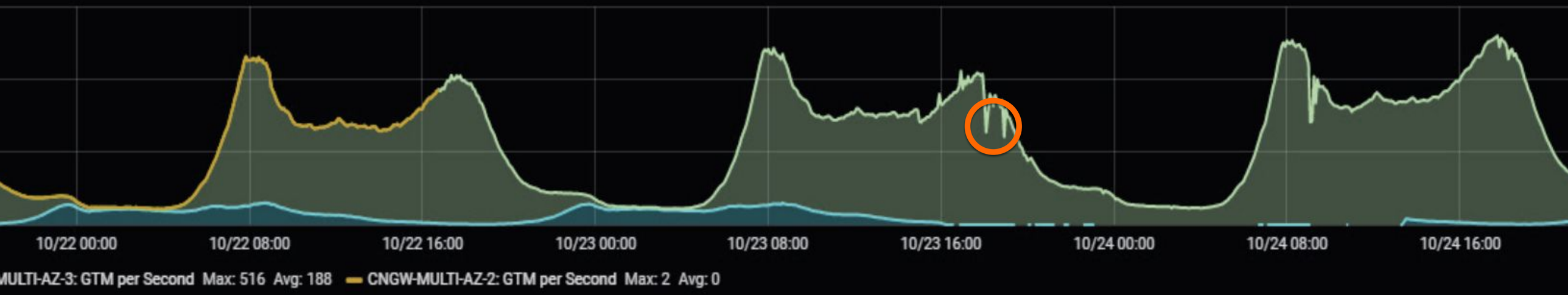
1. Use case presentation
2. System architecture
3. Game plan & Challenges
4. Data preprocessing
5. Anomaly detection models explanation and comparison
6. Prediction models explanation and comparison
7. Notification functions
8. Recap and Recommandations


RTTI
BMW
8 Mio.
ConnectedDrive Cars
P
P
P

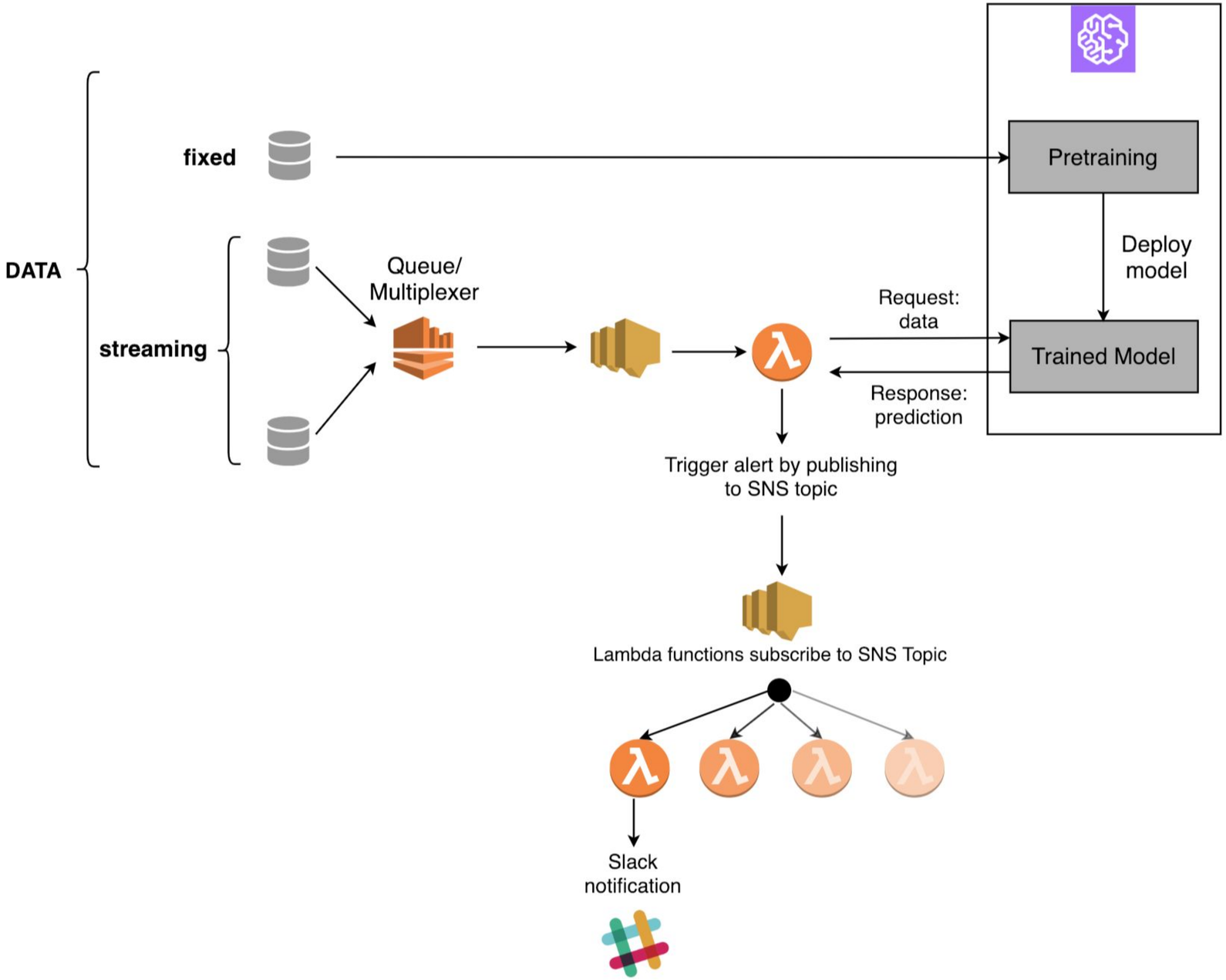
# Use Case Presentation

Anomaly detection on time series data

(requests per second)



# System architecture



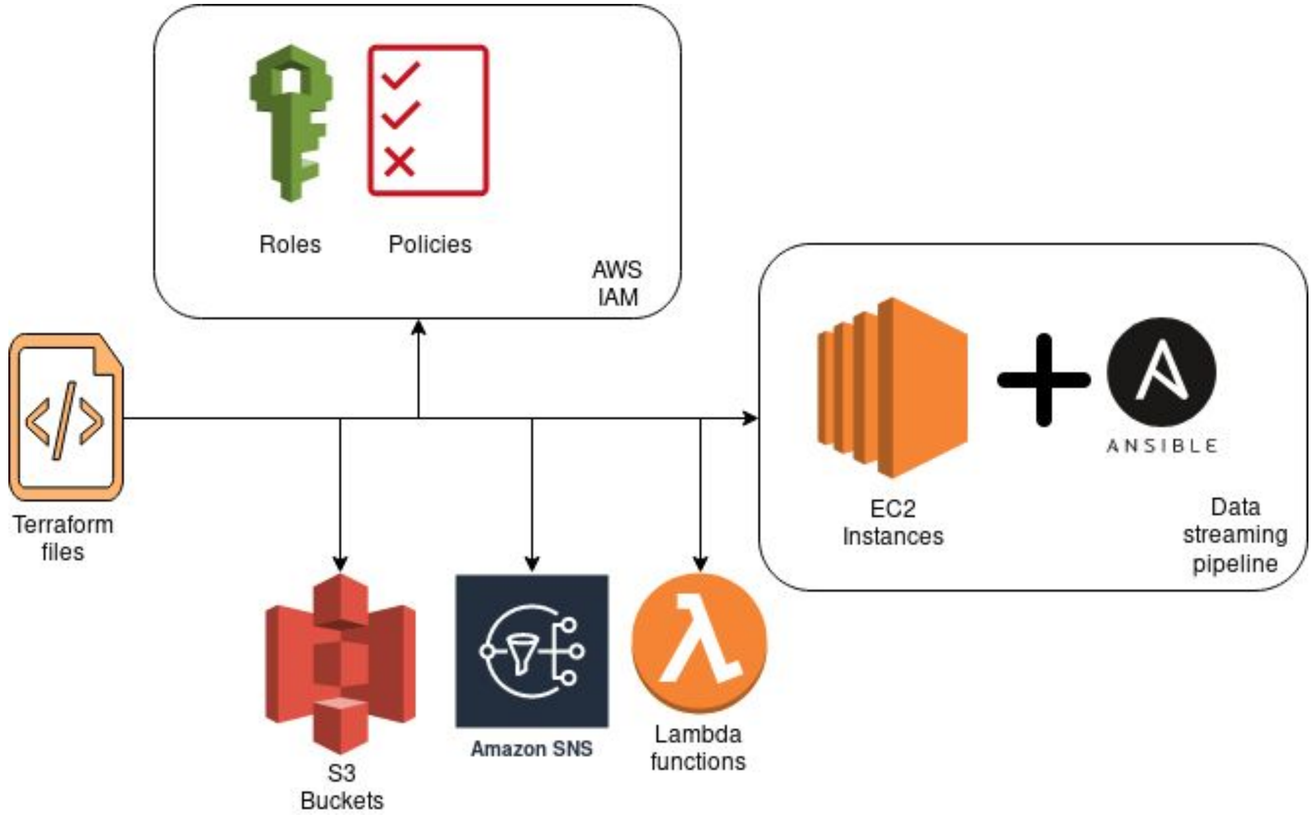
# Deployment

Infrastructure managed with Terraform

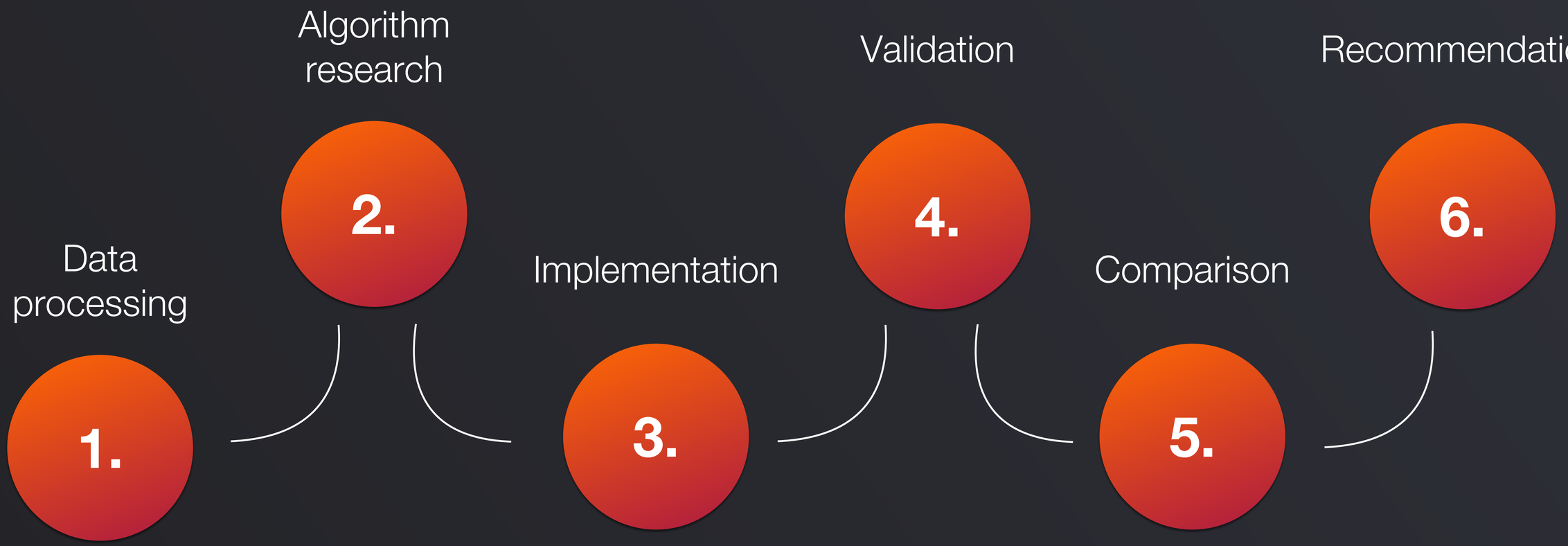
Provisioning with Ansible

Goal: get rid of most of the setup complexity

BUT some features are not supported yet



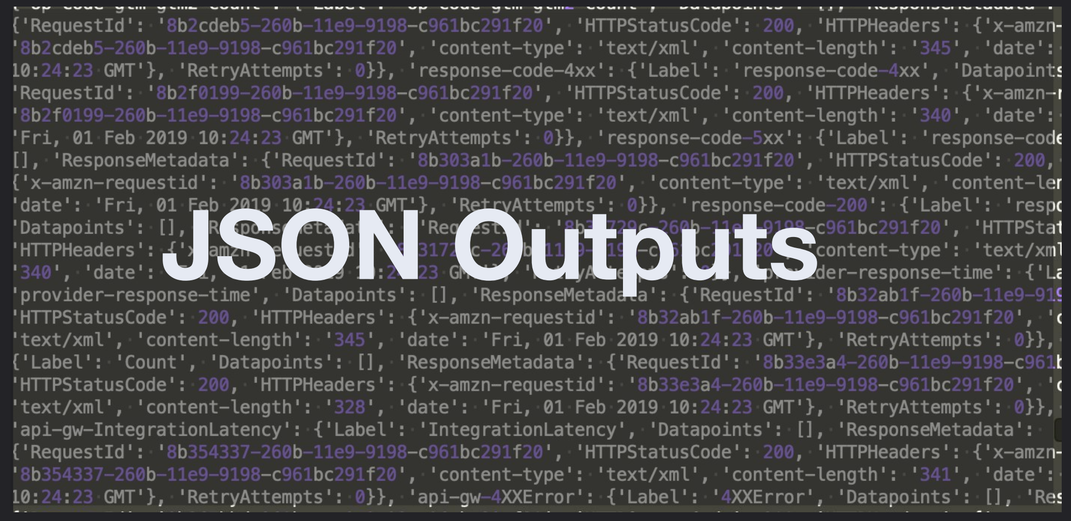
# Game plan


Data processing
1.
Algorithm research
2.
Implementation
3.
Validation
4.
Comparison
5.
Recommendation
6.

| Challenges | Solutions |
| --- | --- |
| 1. Understanding and preprocessing data | 1. Create 1 min and 5 min granularity buckets |
| 2. Faking a streaming data pipeline | 2. Feeding portions of data once every 15 min |
| 3. Lack of extensive AWS knowledge | 3. Organizing knowledge sharing sessions |
| 4. Exploring Machine Intelligence solutions | 4. Trying and comparing a couple algorithms |
| 5. Small amount of data | 5. Generating new data with our prediction models |

# ◦◦◦ Data processing

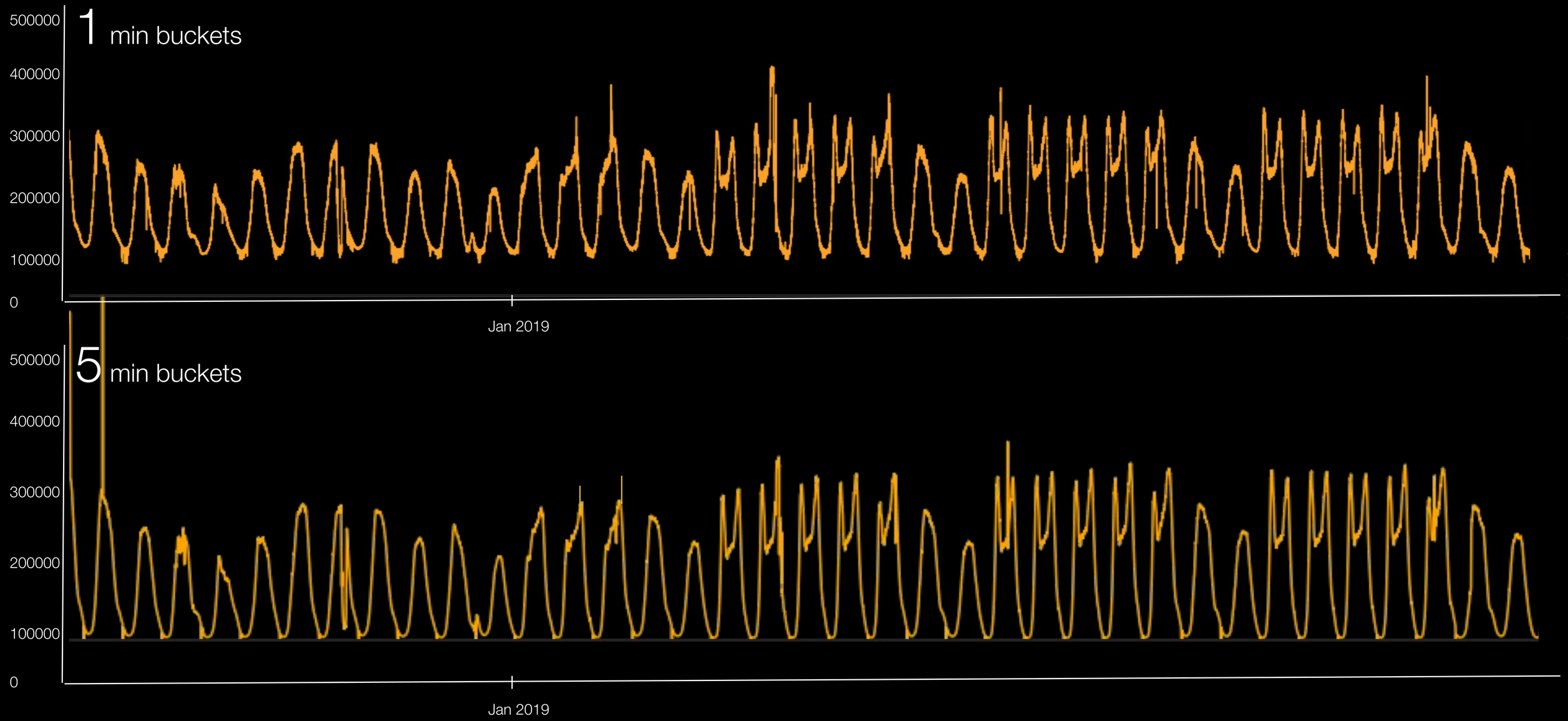

1) Aggregate logs into
**1 min** buckets



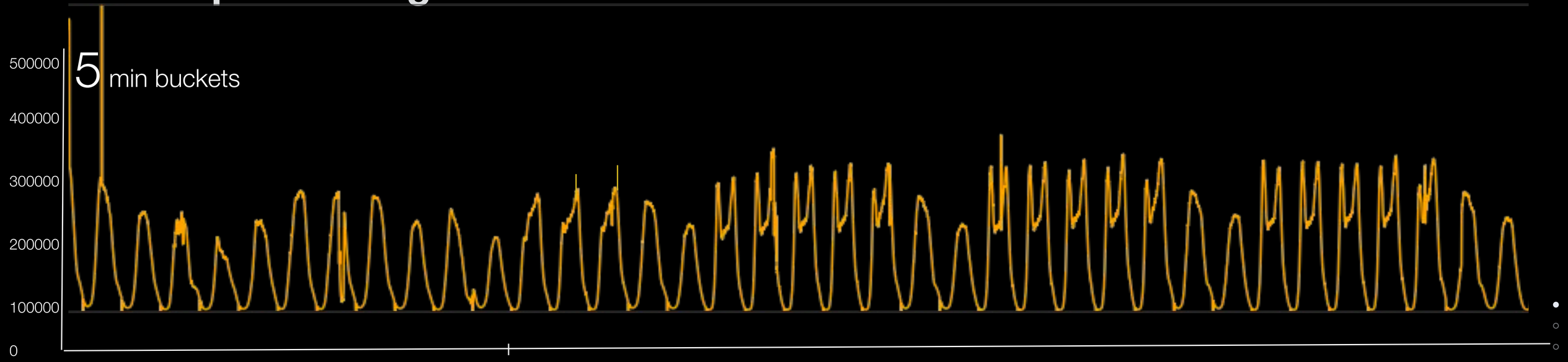

1) Extract the
**Datetime & Sum**

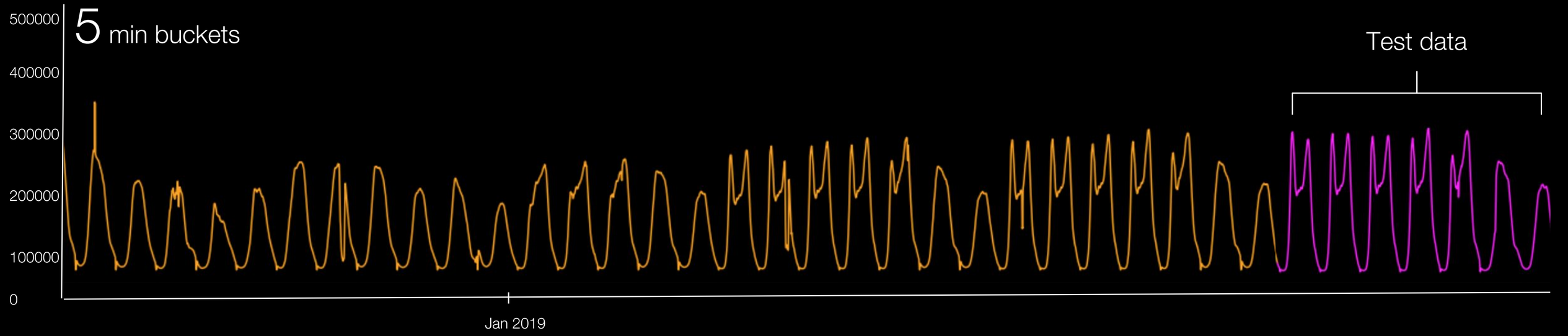
2) Aggregate results into
**5 min** buckets

# Data visualisation



# Data processing



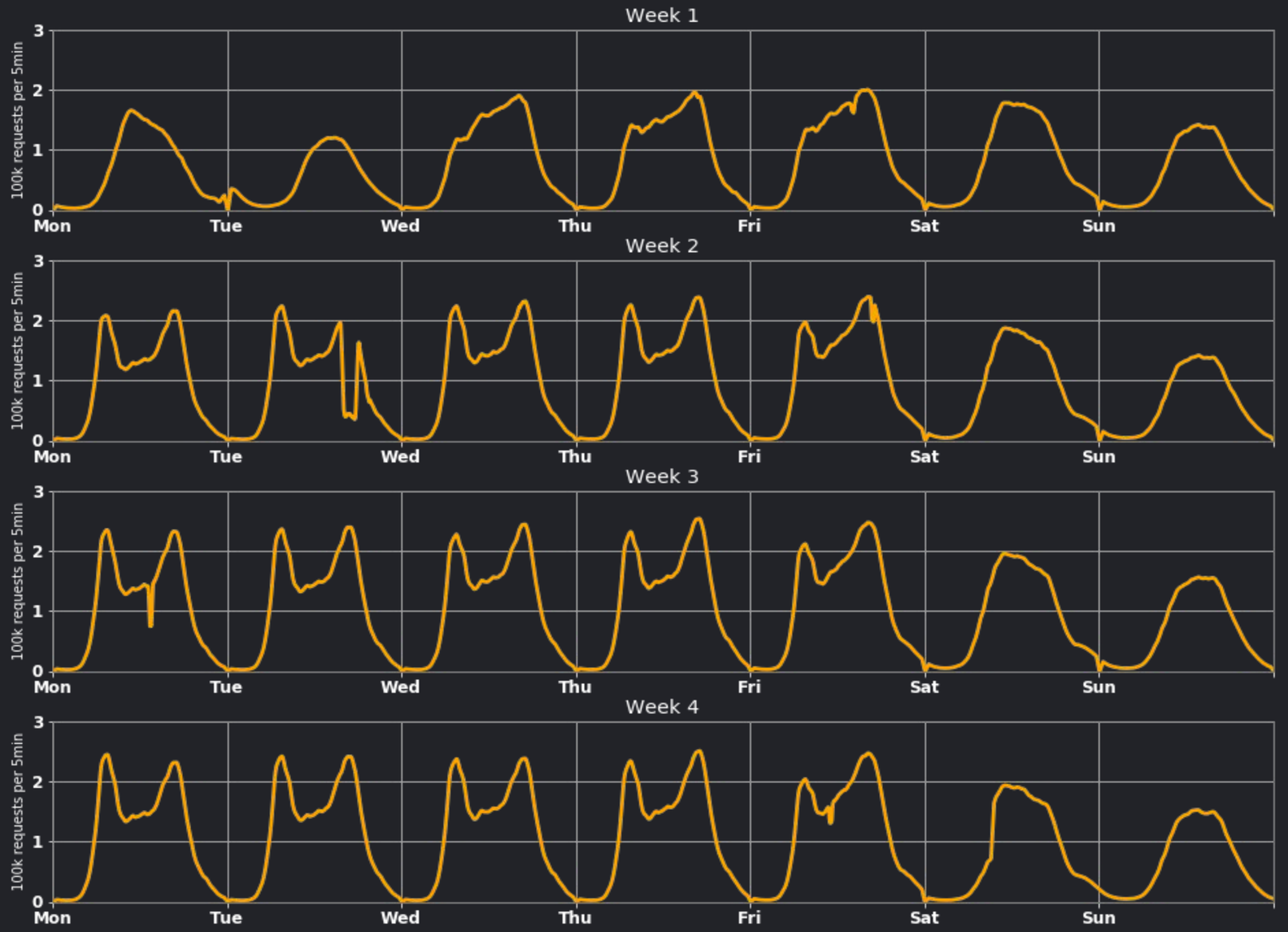




Anomaly detection
models

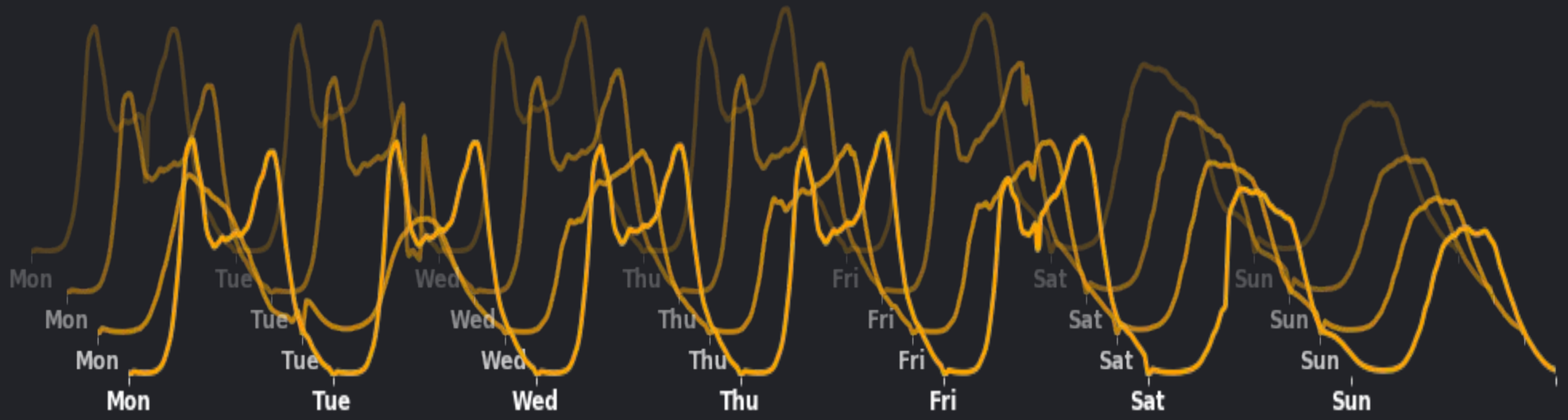
# 1 - Mean Predictor

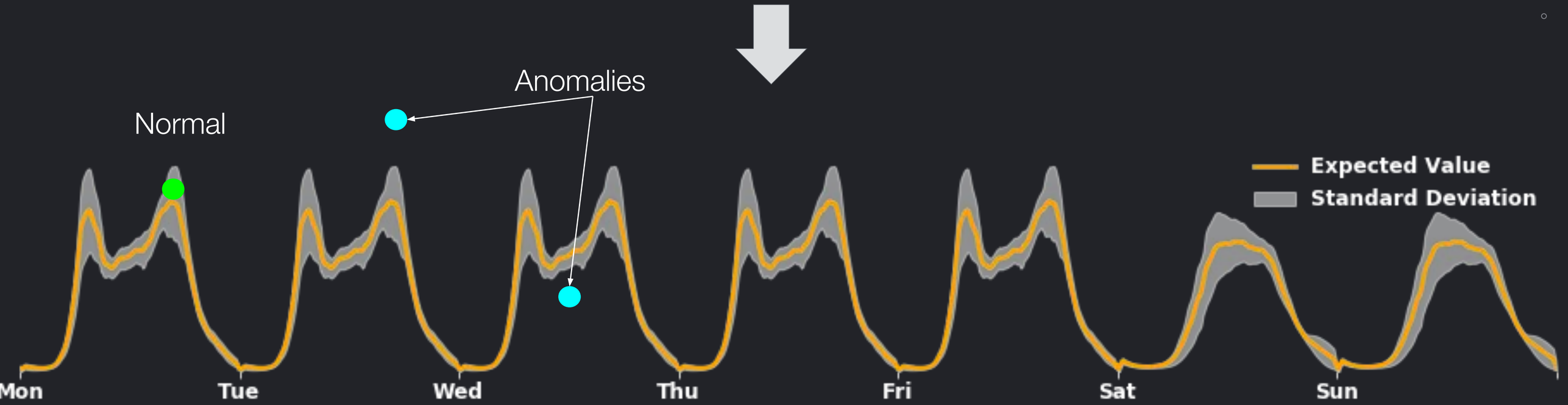
# Data visualization





# Mean predictor



# Mean predictor



## PROS

Requires very little data to train

Trains very fast

Requires very little computation power

Robust against outliers

Can be used both for forecasting as well as anomaly detection

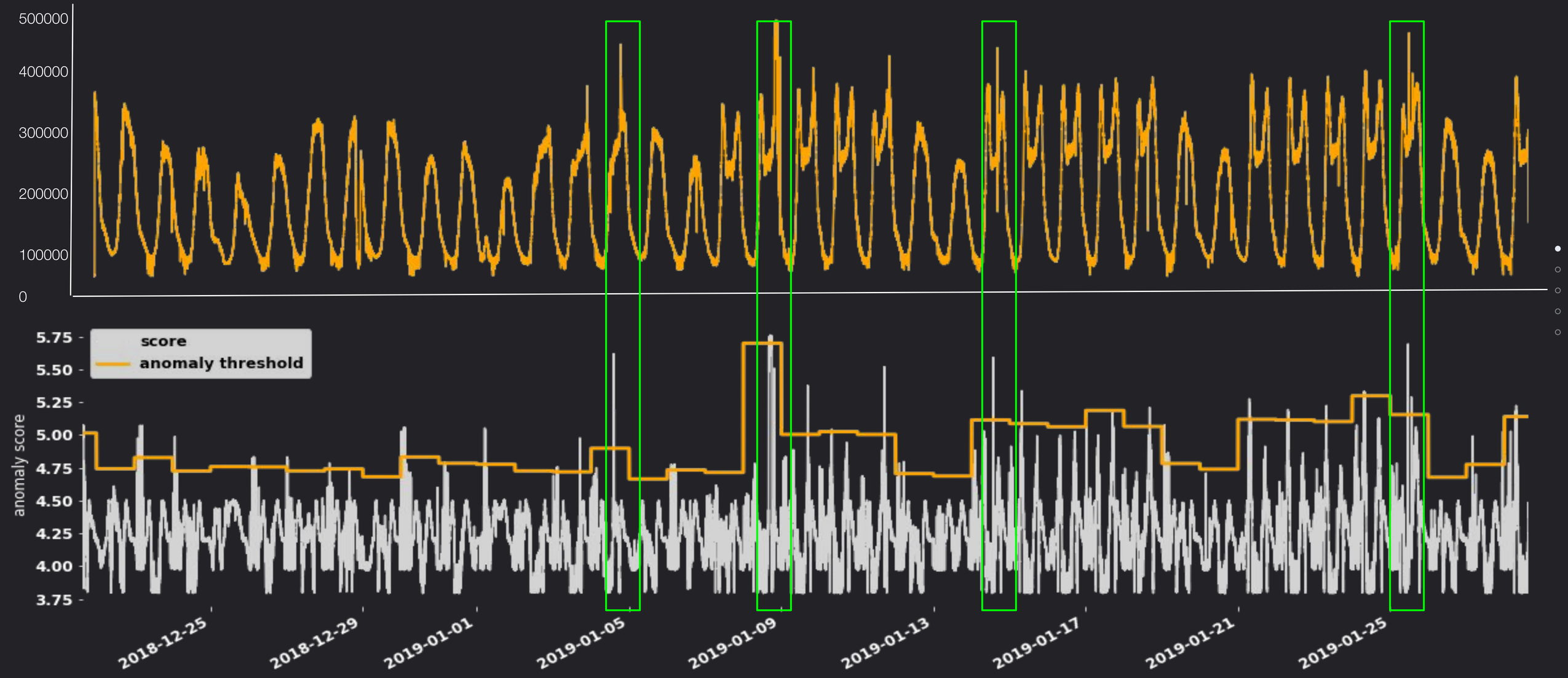
## CONS

Can only take into account a single source of information

Not an Out-Of-The-Box solution from AWS

# 2- Random Cut Forest

# Data visualization



# Random Cut Forest



## PROS

Good performance for long-lived data streams

Short training time

Relatively easy to create and deploy the model (Out-Of-The-Box solution from Amazon)

Has a stream-friendly version

## CONS

Supports streaming only on kinesis streams

Requires huge amounts of data

# Technical decisions

**We have:**

- A lack of long-term data (FlowLogs collection started in December)
- A unique time series to analyze

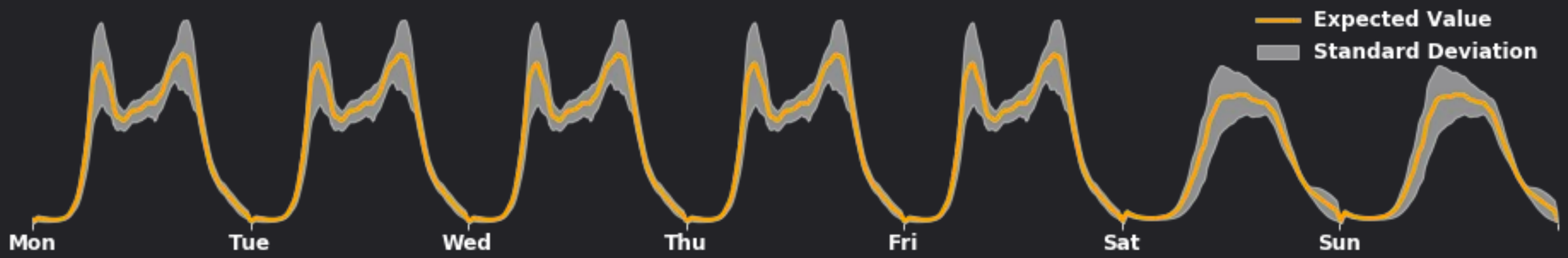
**We need:**

- A tweakable solution
- Short train times

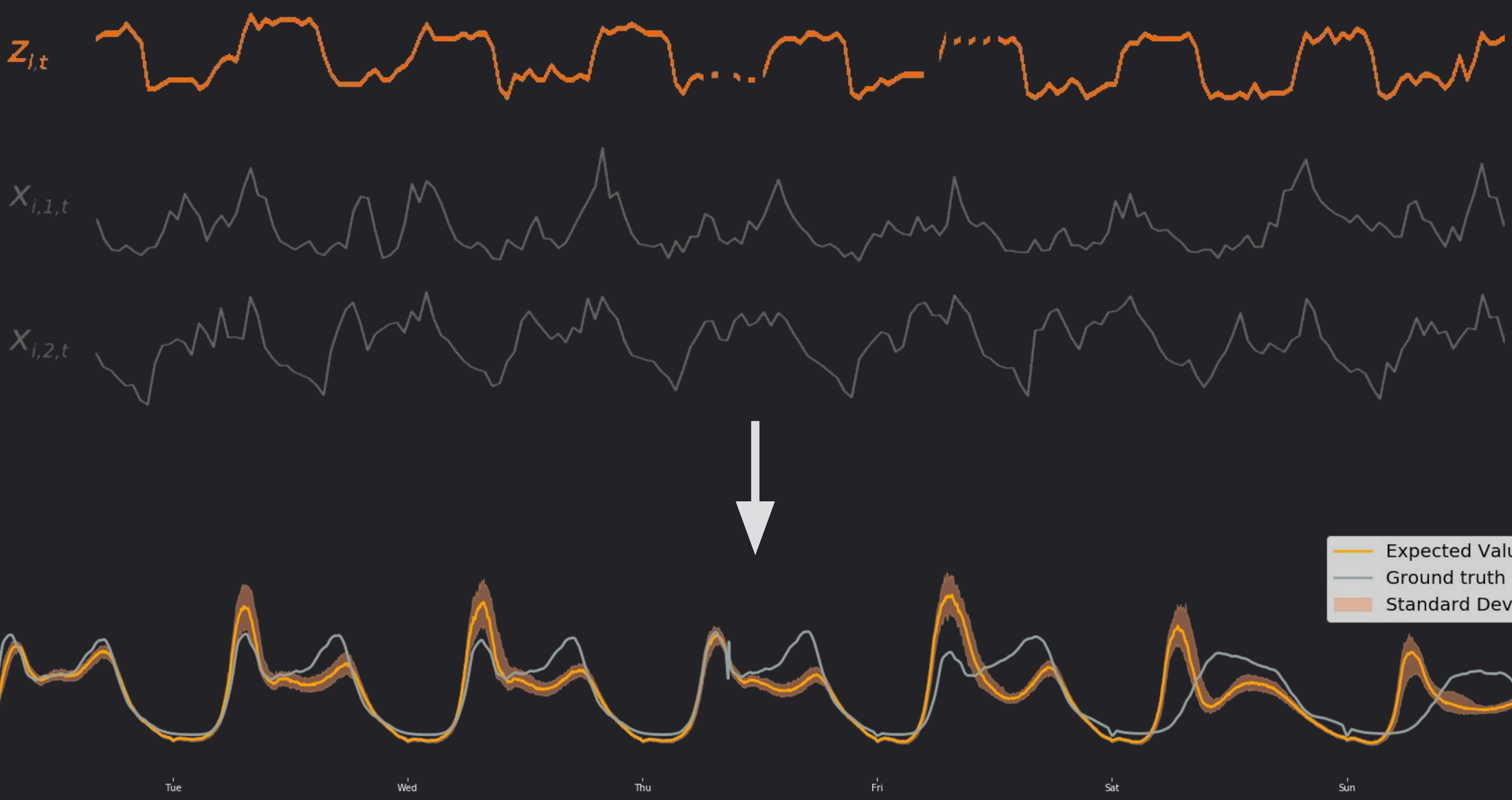
→ **Explains why we built our demo pipeline with Mean Predictor**


Prediction models

# Mean Predictor



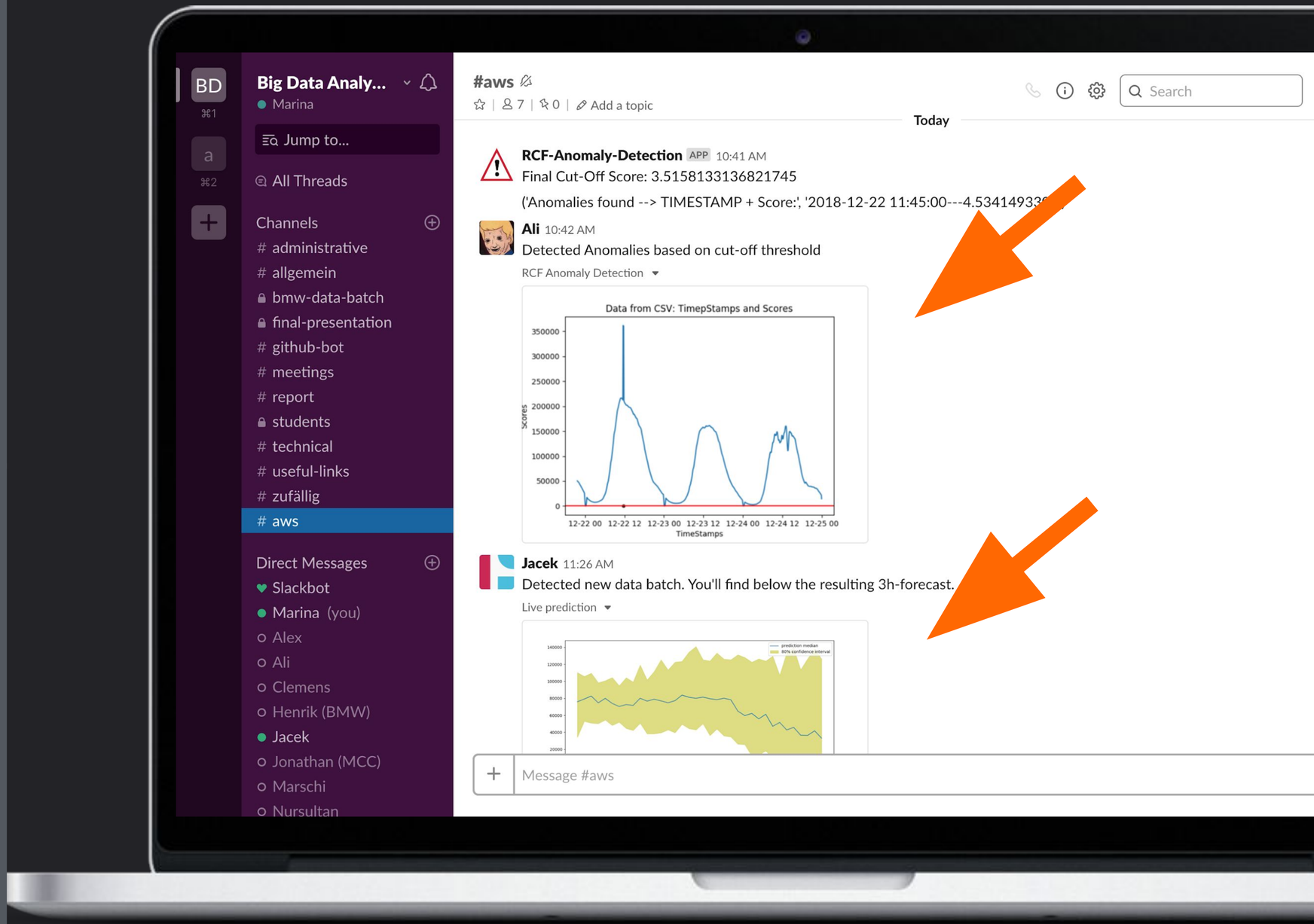
# DeepAR


$z_{i,t}$
$x_{i,1,t}$
$x_{i,2,t}$
Expected Value
Ground truth
Standard Deviation
Mon
Tue
Wed
Thu
Fri
Sat
Sun

# Notification functions

Anomaly detection notification

3 hour forecast notification



# RECOMMENDATIONS

Implement the Mean Predictor Model

Switch to RCF and DeepAR

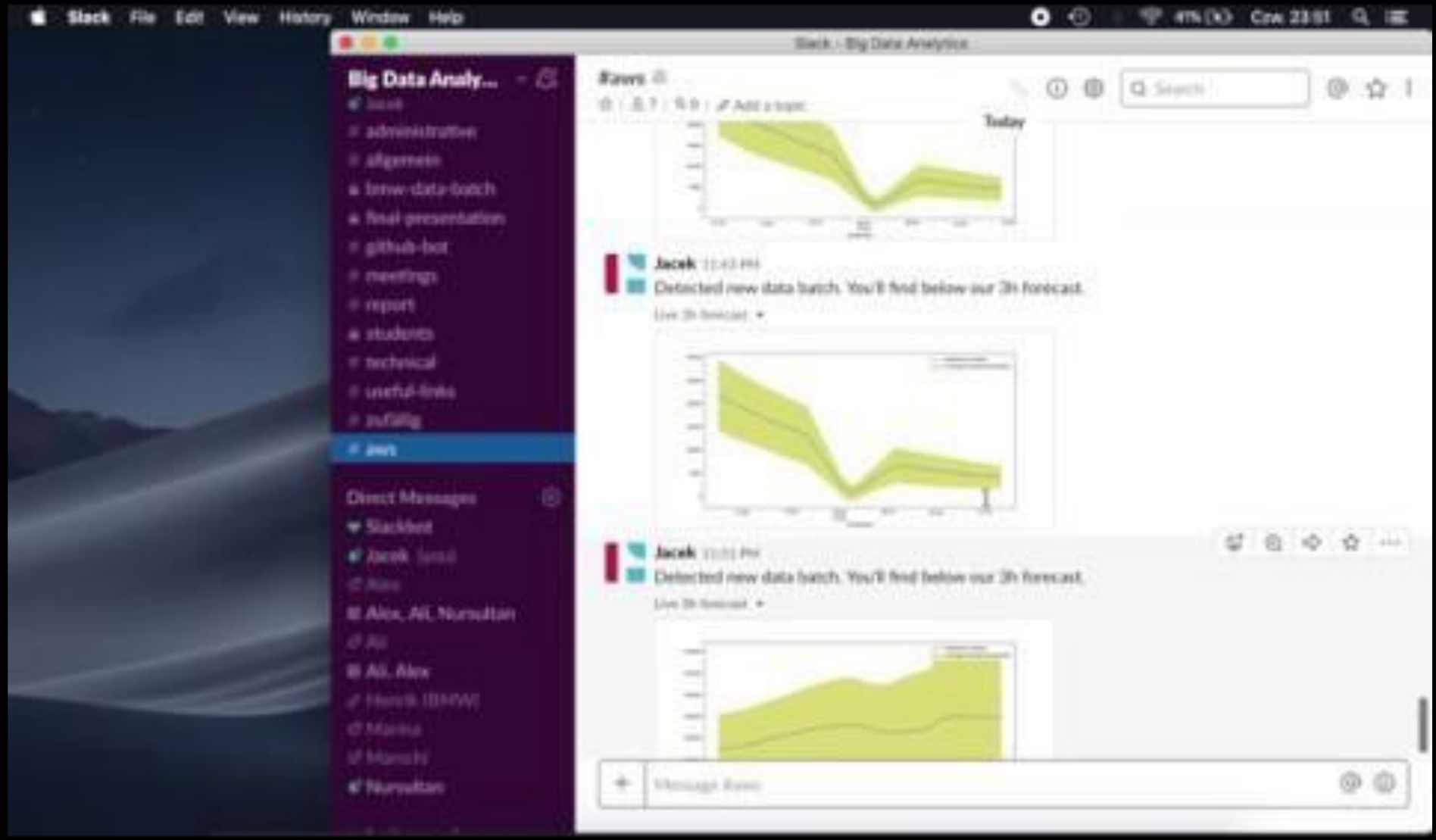
Attach geolocation to data points

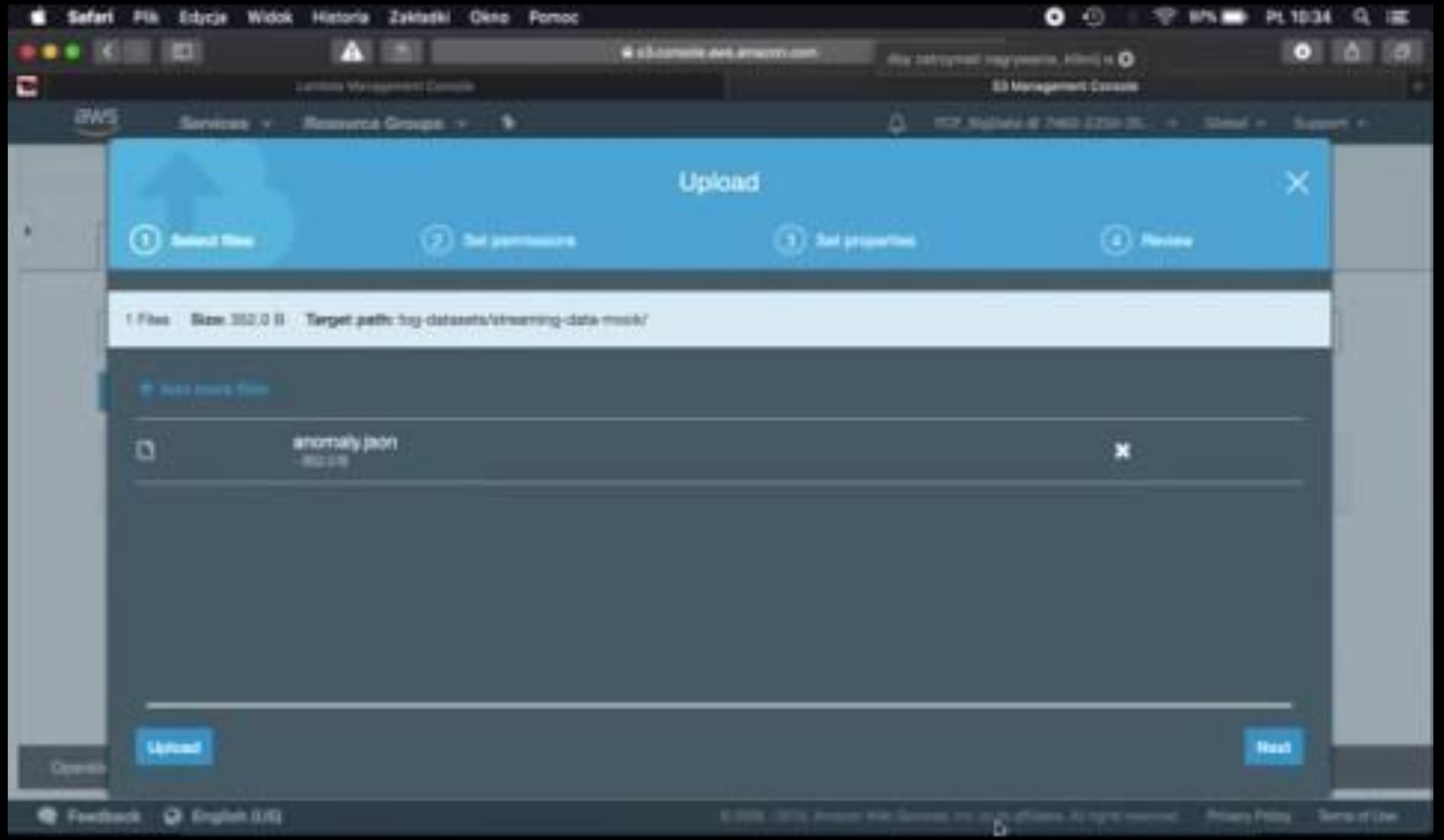
Gather 2-3 years of data

Look for seasonal trends

Explore the holidays periods

# Thank you for your attention


Slack File Edit View History Window Help
Slack - Big Data Analytics
Big Data Analy...
administration
allgemein
bmw-data-batch
final-presentation
github-bot
meetings
report
students
technical
useful-links
zufällig
aws
Direct Messages
Slackbot
Alex, Ali, Nursultan
Ali, Alex
Nursultan
#aws
Add a topic
Today
Jacek
Detected new data batch. You'll find below our 3h forecast.
Live 3h forecast
Jacek
Detected new data batch. You'll find below our 3h forecast.
Live 3h forecast
Message #aws


Safari Plik Edycja Widok Historia Zakładki Okno Pomoc
Upload
anomaly.json
Upload
Next

# RCF Kinesis architecture