AWS re:Invent

FSI308

Detect market data anomalies using Amazon SageMaker

Avinash Nidumbur

Sr. Data Platform Architect, Global Financial Services

Amazon Web Services

Agenda

Machine learning (ML) in Financial Services

Anomaly detection with Amazon SageMaker

Stock trading volume anomaly detection

Demo: Data preparation & training

Demo: Inference analysis & benchmarks

ML in Financial Services

AWS re:Invent

aws

AI/ML creates the next edge for financial institutions

Financial institutions are increasingly investing in AI/ML, thanks in part to the availability of cost-effective, easy-to-use, and scalable AI/ML cloud services

Compliance, surveillance, and fraud detection

- Credit card/account fraud detection
- Sales practices/ transaction surveillance

Document processing

- Common financial instrument taxonomy
- Contract ingestion and analytics

Pricing and product recommendation

- Loan/insurance underwriting
- Sales/recommendations of financial products
- Credit assessments

Trading & analytics

- Portfolio management/ robo-advising
- Sentiment/news analysis
- Image analysis
- Grid-computing scheduling

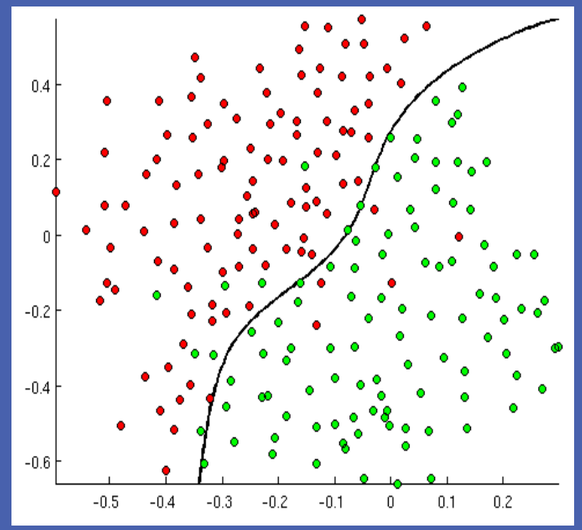
Customer experience

- Enhanced customer service through mobile apps and chatbots
- Call center optimization

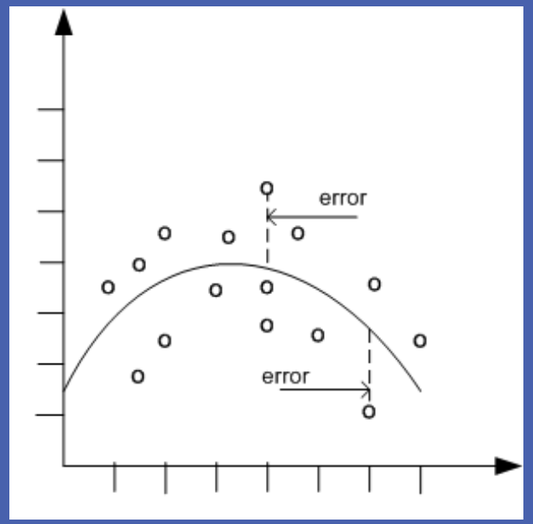
Core processing → Client facing

Some common classes of ML problems

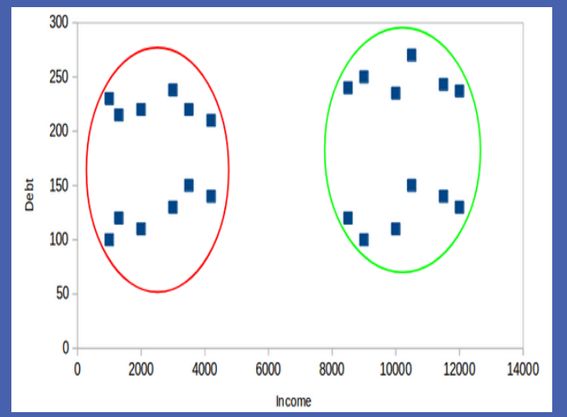
Classification



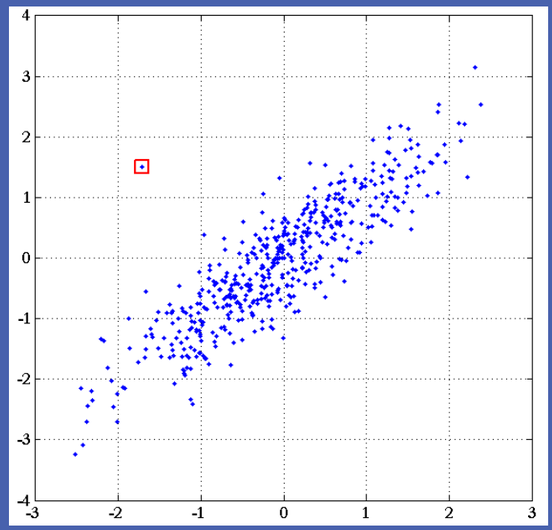
Regression



Clustering



Anomaly detection



Anomaly detection

An *anomaly* is an observation that diverges from otherwise well-structured or patterned data

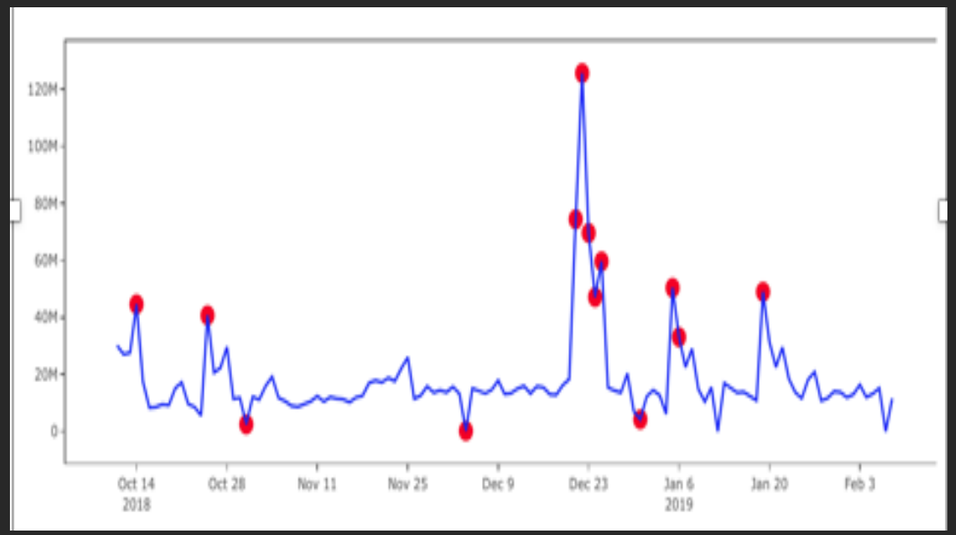
- Can manifest as unexpected spikes in time series data, breaks in periodicity, or unclassifiable data points

Types of anomalies

- Point anomaly
- Contextual anomaly
- Collective anomaly

Detection of anomalies

- Establish thresholds – Analytics pattern
- Unsupervised algorithm models – ML pattern



Source:
https://towardsdatascience.com/anomaly-detection-with-isolation-forest-visualization

Anomaly detection: Financial market use cases

- Transaction fraud
- Anti-money laundering
- Identity theft and fake account registration
- Risk modeling
- Account takeover
- Promotion abuse
- Customer behavior analytics
- Cybersecurity

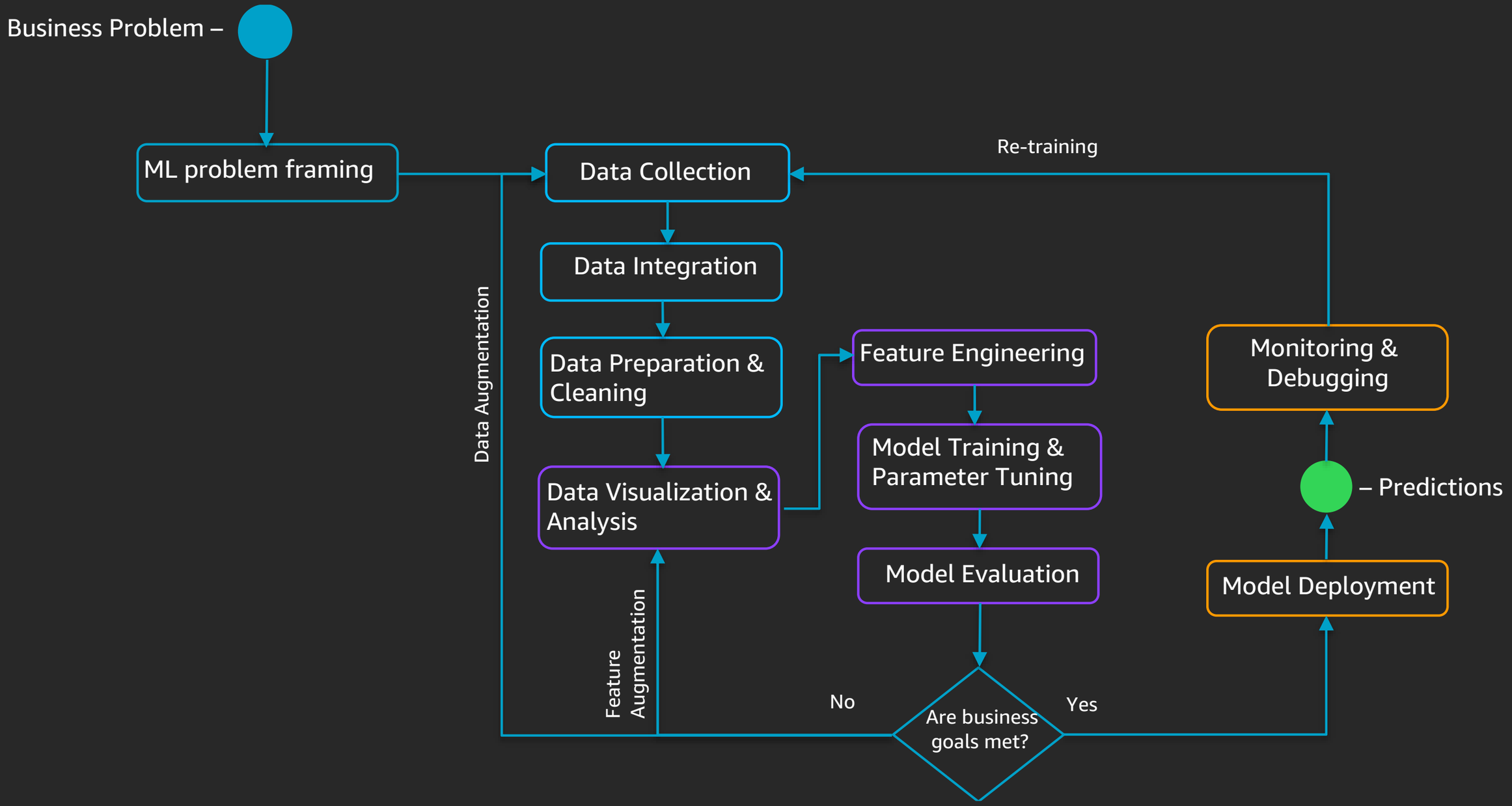
Source: Anomaly Detection in Finance

Anomaly detection with Amazon SageMaker

AWS re:Invent

aws

ML process

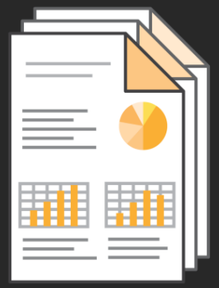

Business Problem –
ML problem framing
Data Collection
Re-training
Data Integration
Data Augmentation
Data Preparation & Cleaning
Feature Engineering
Monitoring & Debugging
Data Visualization & Analysis
Model Training & Parameter Tuning
– Predictions
Model Evaluation
Model Deployment
Feature Augmentation
No
Are business goals met?
Yes

Amazon SageMaker components

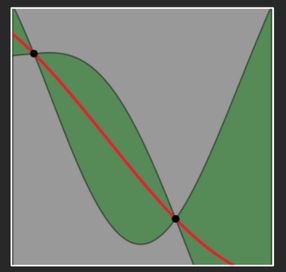
Jupyter

Hosted notebook

Training service



Built-in algorithms



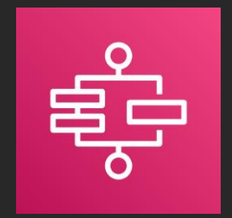
Hyperparameter optimization

Hosting

Console

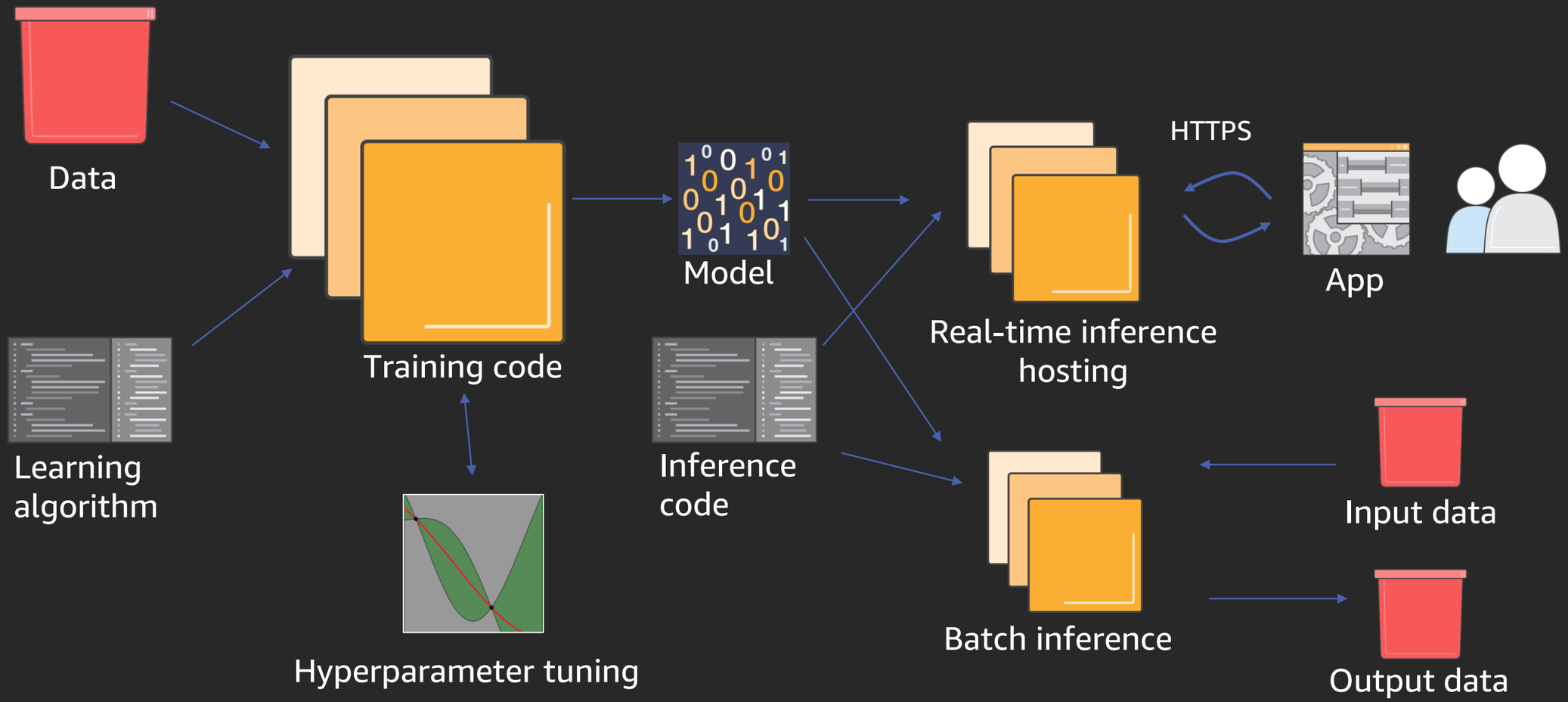
Python SDK

Spark SDK



AWS Step Functions

Amazon SageMaker training and inference flow



Anomaly detection using Random Cut Forest algorithm

ICML 2016

"Robust Random Cut Forest Based Anomaly Detection on Streams"
[Guha, Mishra, Roy, Schrijvers]

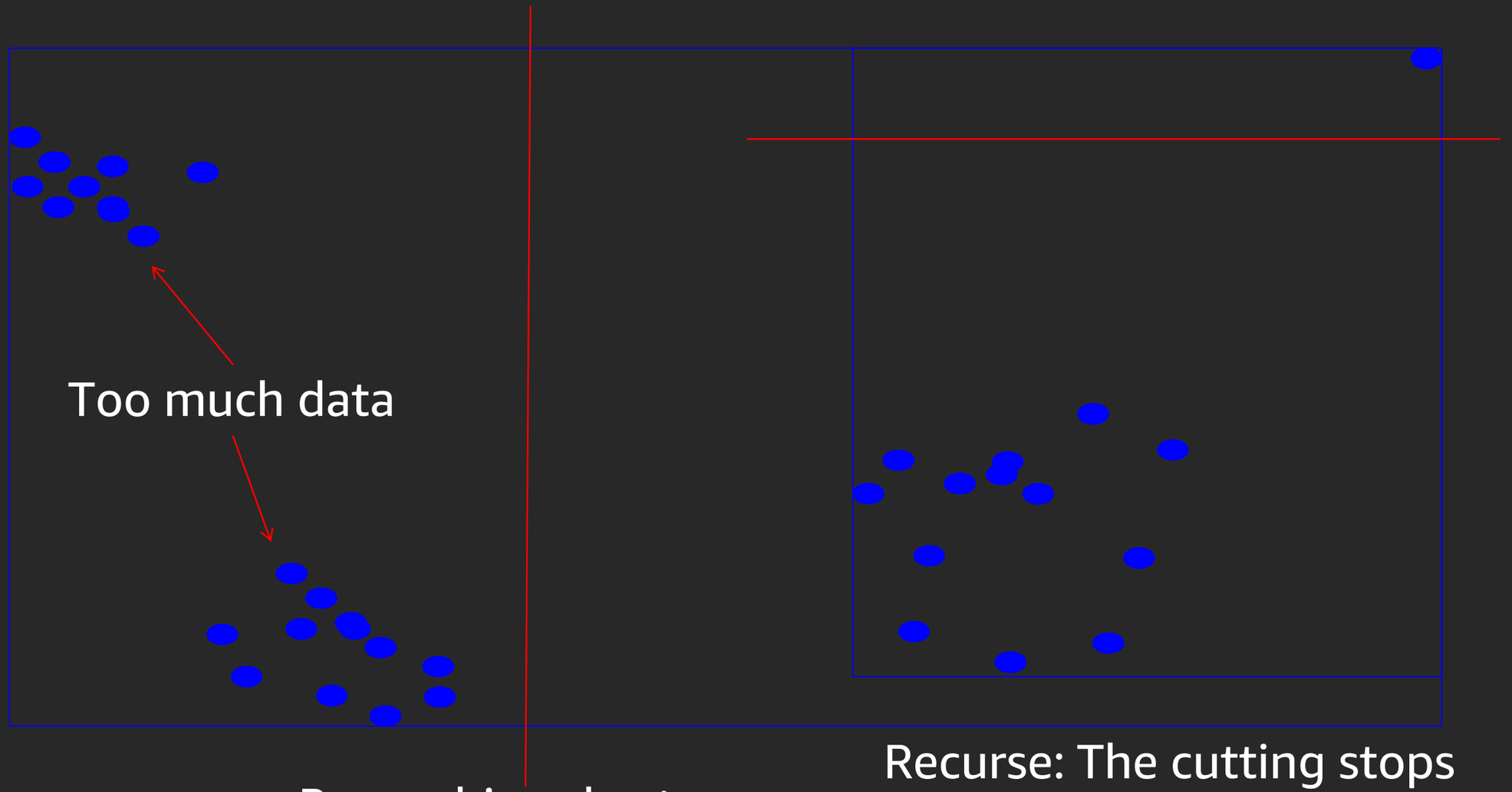
Amazon Kinesis

http://docs.aws.amazon.com/kinesisanalytics/latest/dev/app-anomaly-detection.html

Amazon SageMaker

https://docs.aws.amazon.com/sagemaker/latest/dg/randomcutforest.html

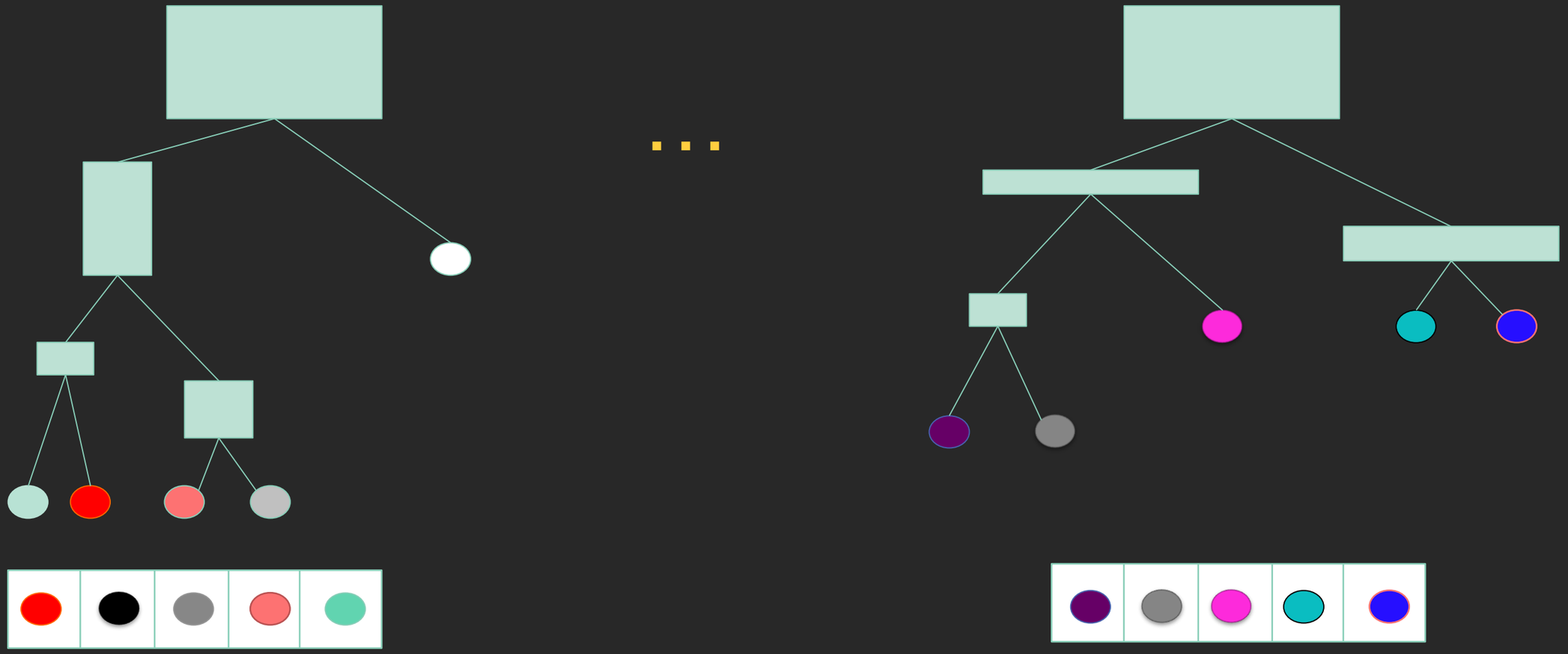
Random cut tree



Range-biased cut

Recurse: The cutting stops when each point is isolated

Random Cut Forest

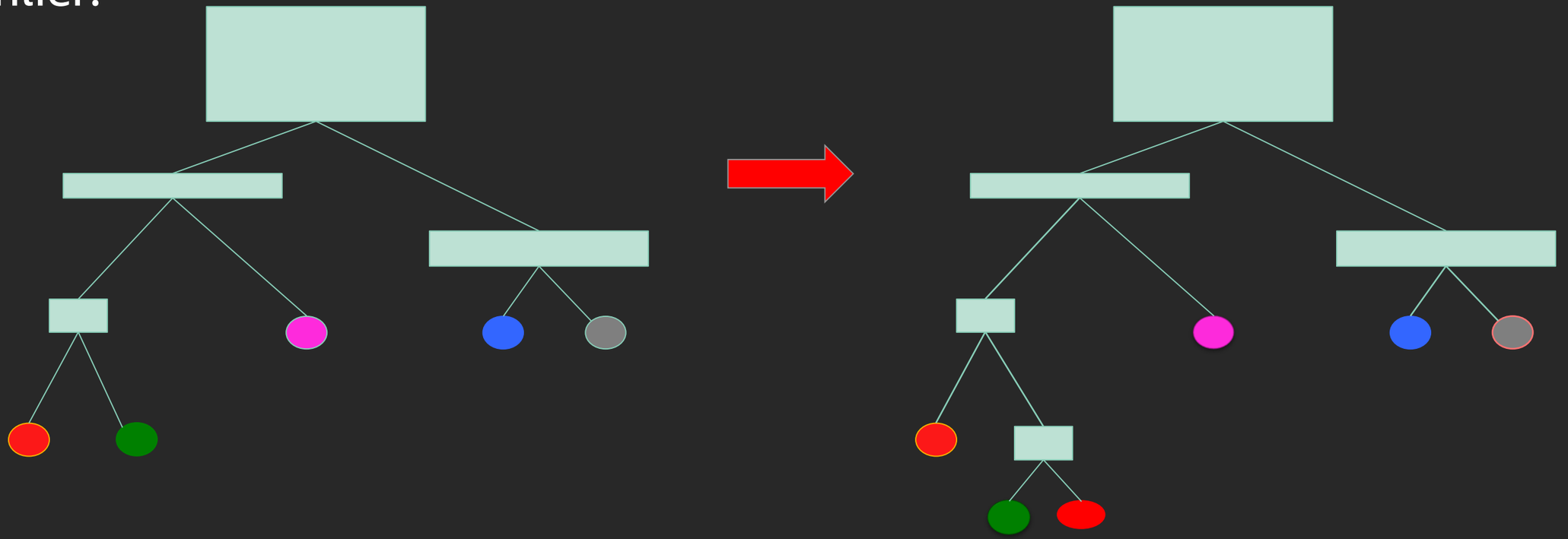


Each tree built on a random sample

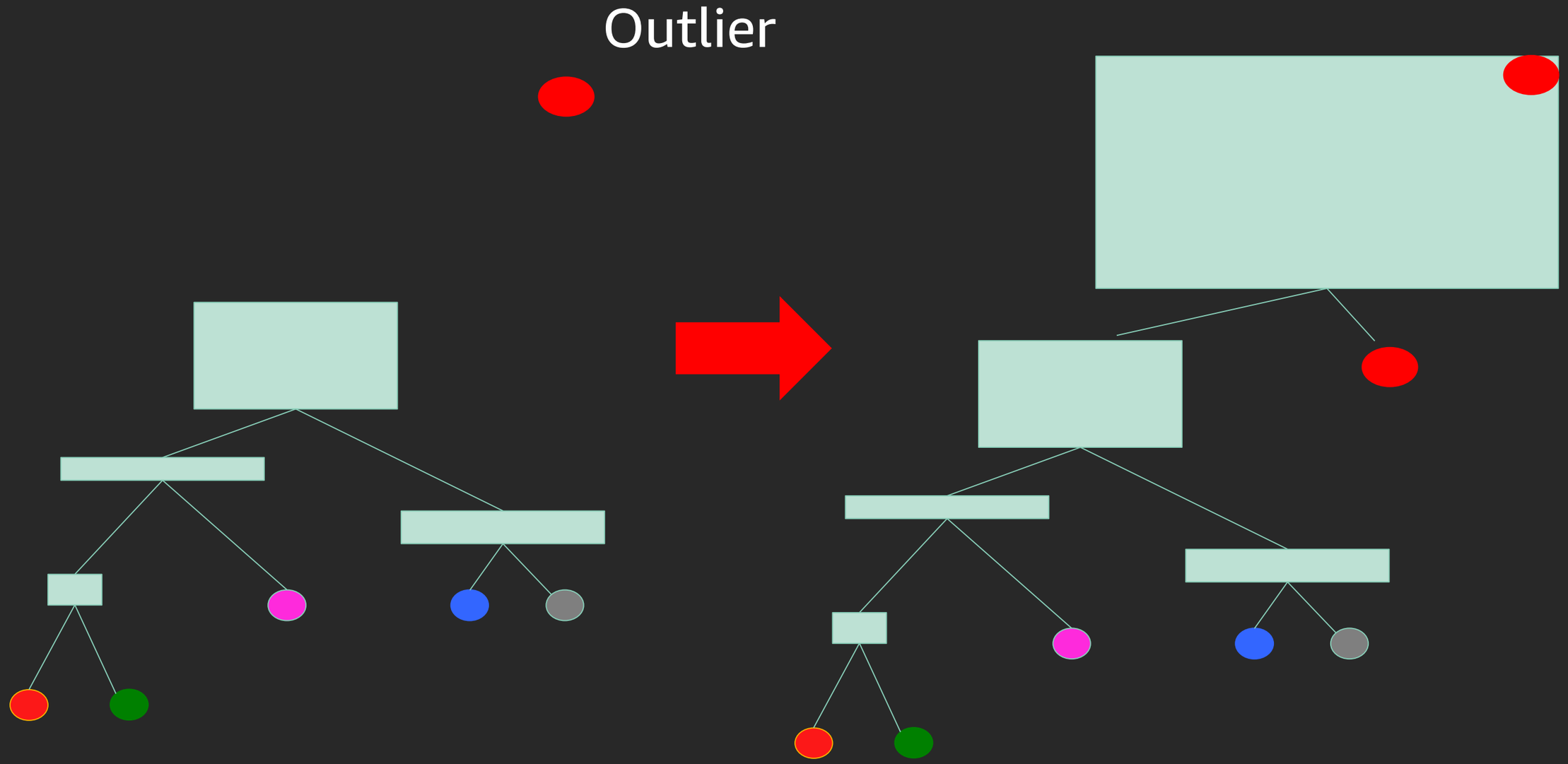
Anomaly score: Displacement

The anomaly score assigned to an input data point is inversely proportional to its average depth across the forest (= sum of path lengths from root to leaves = description length)

Inlier:



Anomaly score: Displacement



Random Cut Forest in Amazon SageMaker

- Works in batch mode
- Does not learn continuously by itself
- Shingling done manually during data preprocessing
- Accuracy can be measured if anomalies are known in training data
- Hyperparameter options
    - The number of features in the dataset (feature_dim)
    - A list of metrics used to score a labeled test dataset (eval_metrics)
    - Number of samples given to each tree from the training set (num_samples_per tree)
    - Number of trees in the forest (num_trees)

Stock trading volume anomaly detection

AWS re:Invent

aws

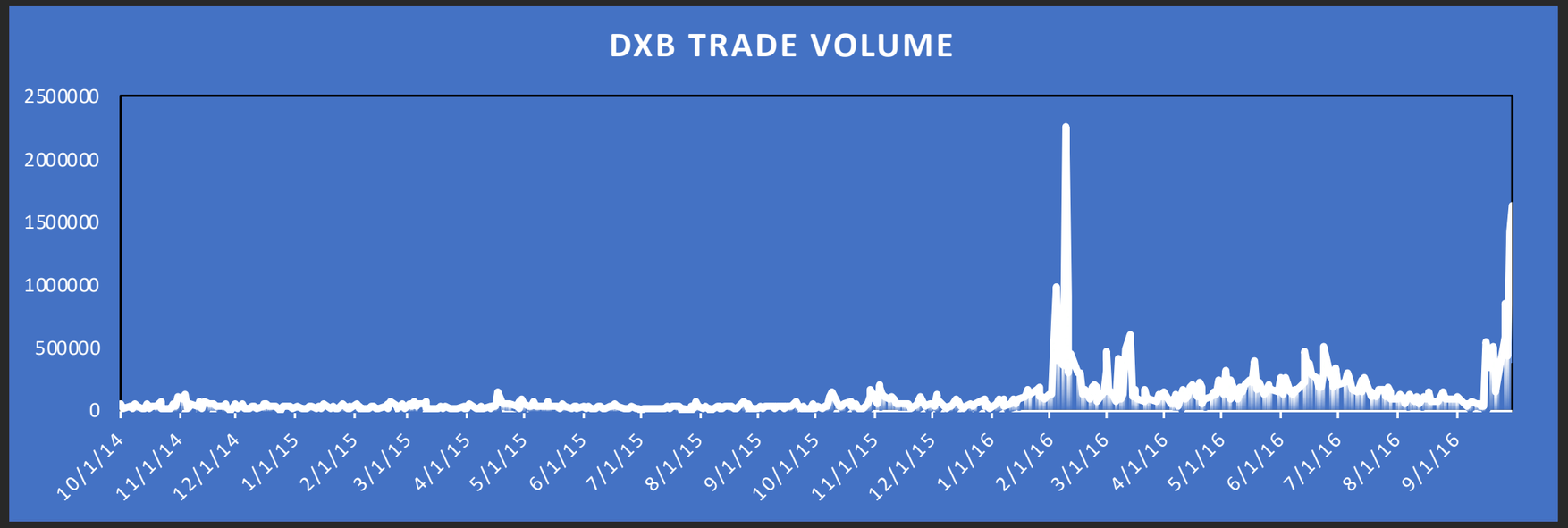
Stock trading volume analysis

- Monitoring stock trading volume
  - Indicator of a share’s performance
  - High trading volumes could be indicators of corporate events or market sentiments
  - Analysis could be used to predict stock price movement or determine trading strategies

- Dataset
  - 5 years of trading volume (daily aggregation) of publicly traded bank stock sourced from public datasets
  - 40/60 split for training/inference
  - Feature engineered to retain trading date and volume

- Unsupervised modeling using Random Cut Forest

Public bank trade volume pattern



Training data –

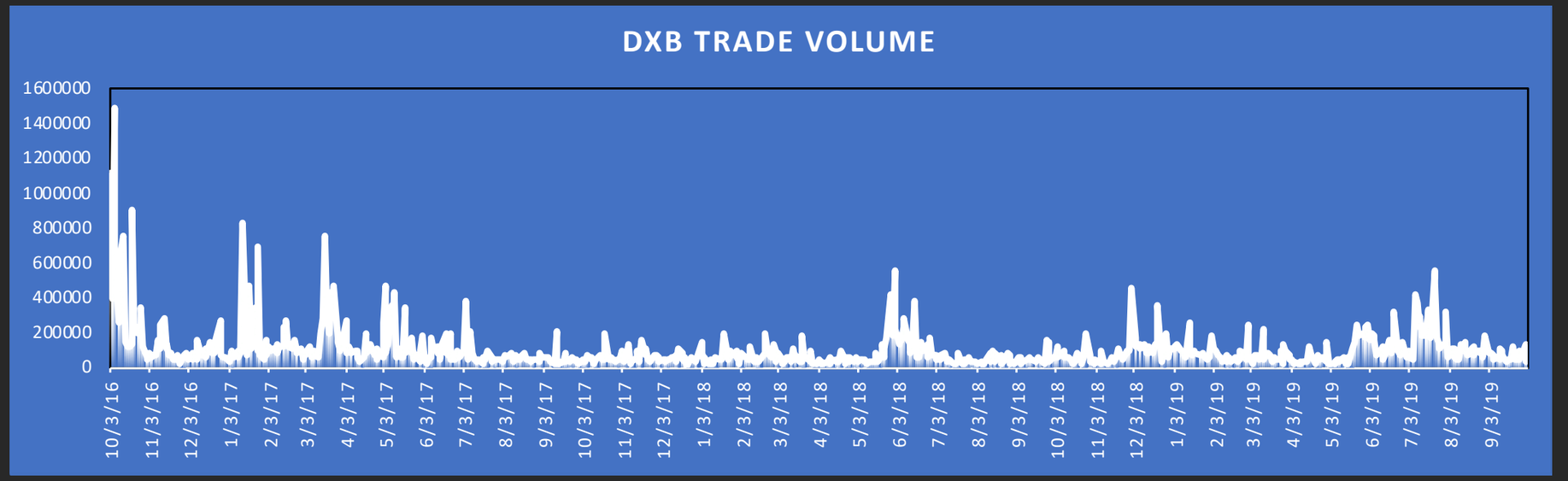
Volume aggregated daily

Sample size – 2 yrs
(10/1/14 – 09/30/16)

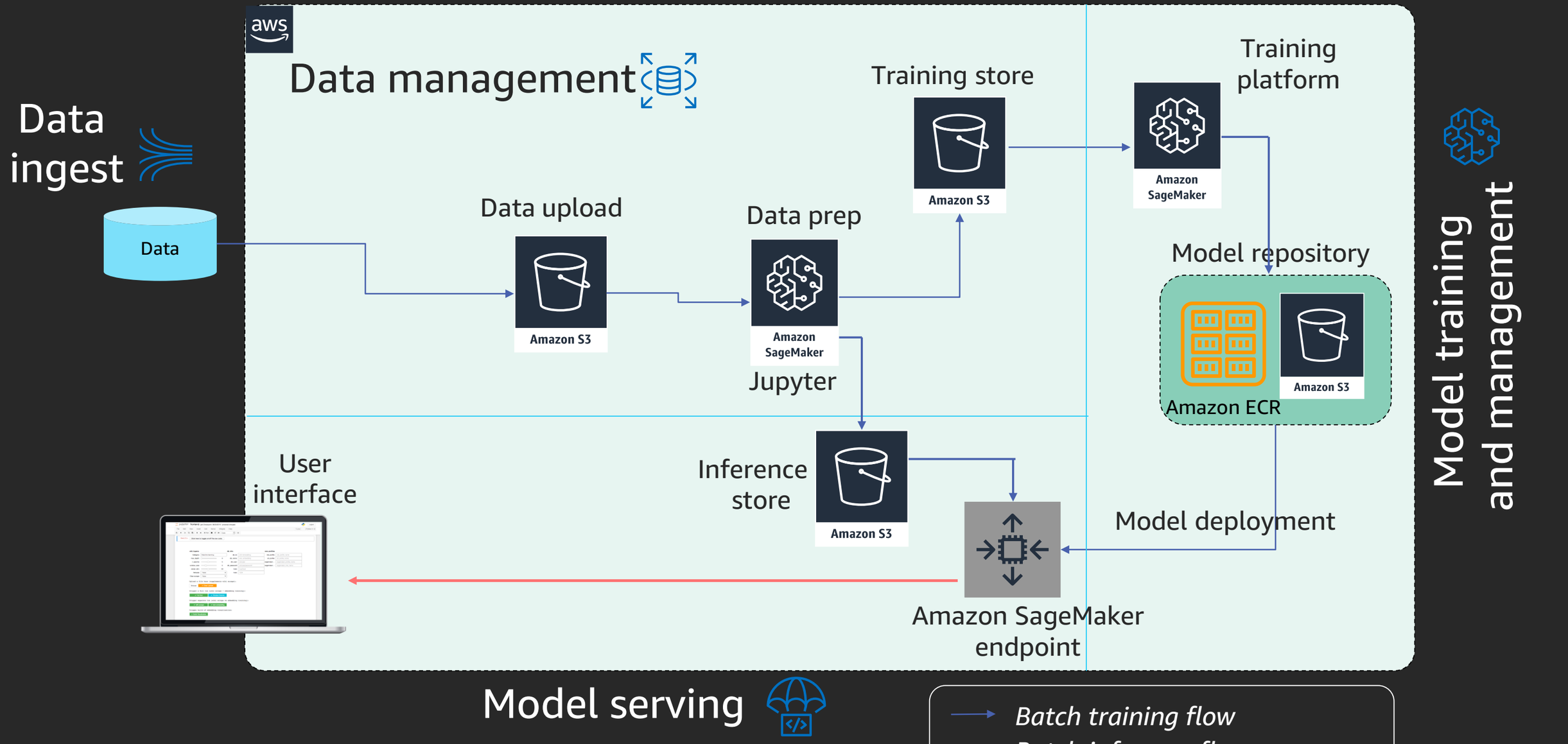
Inference data –

Volume aggregated daily

Sample size – 3 yrs
(10/3/16 – 10/02/19)



Stock trade volume anomaly detection: Architecture



Demo:

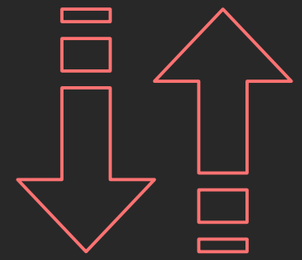
Data preparation using notebook
Random Cut Forest training

Amazon Elastic Inference

Reduce deep learning inference costs up to 75%

Lower inference costs

Match capacity to demand



Available between 1 to 32 TFLOPS

Key features

Integrated with Amazon Elastic Compute Cloud (Amazon EC2), Amazon SageMaker, and Amazon deep learning AMIs

Support for TensorFlow, Apache MXNet, ONNX, PyTorch

Single and mixed-precision operations

Model deployment: Best practices

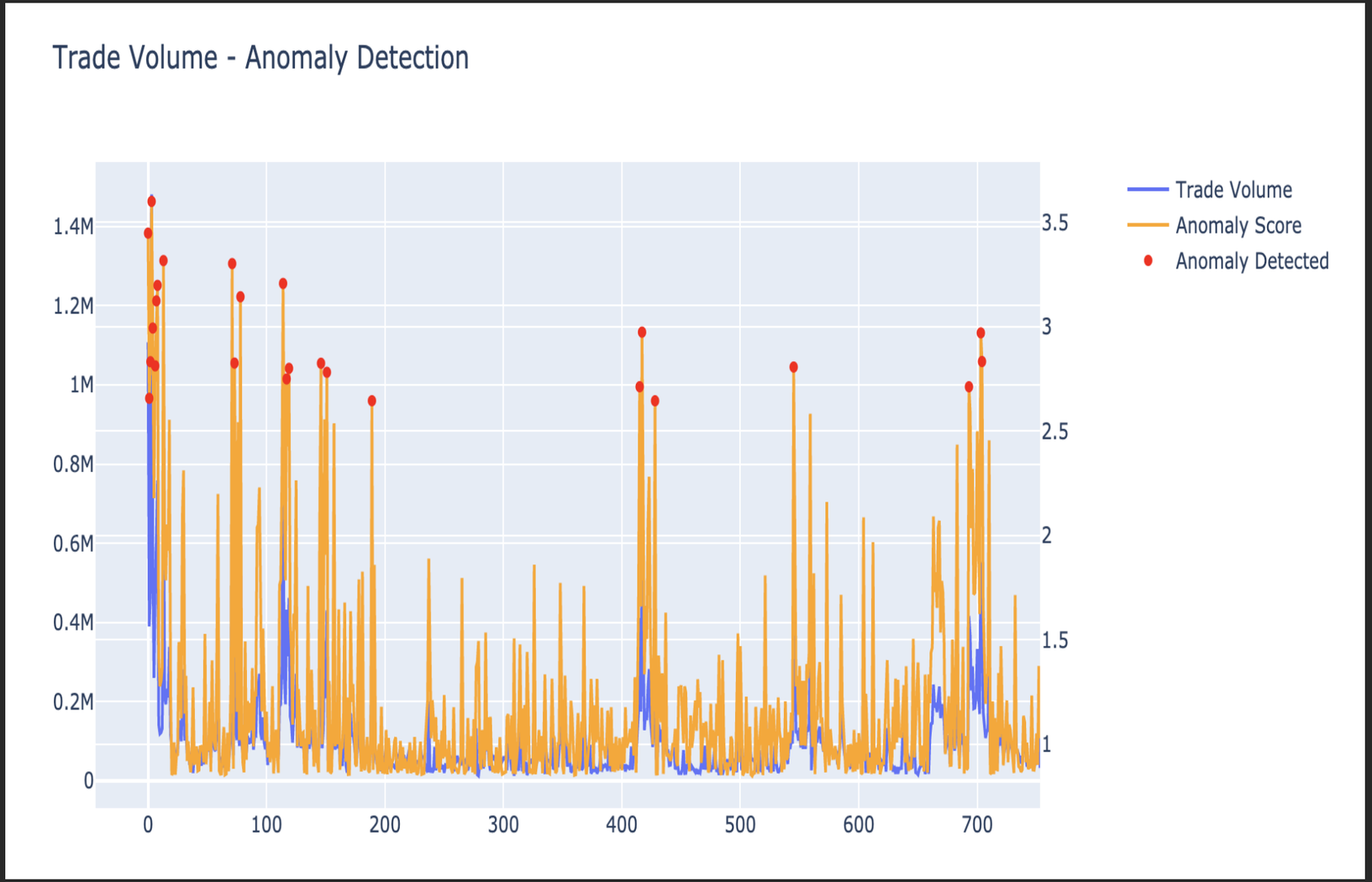
- Deploy to at least 2 Availability Zones, multiple regions for DR
- Use auto-scaling to accommodate traffic spikes
- Consider batch inference or edge deployment
- Use Elastic Inference accelerators to reduce inference costs
- Automate training and deployment pipeline, blue/green deployment
- Clearly define retraining triggers
- Enable model versioning, implement ML DevOps pipeline
- Monitor performance

Demo:

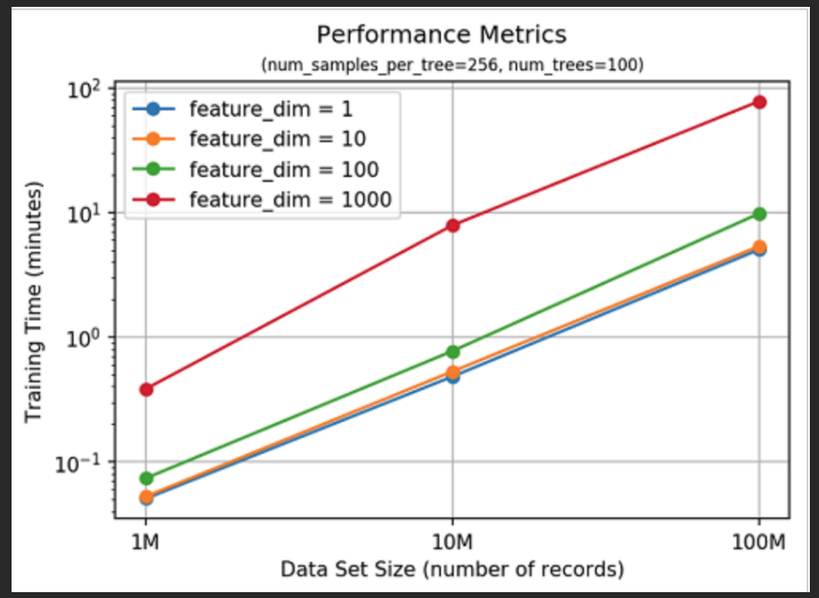
Create an endpoint
Inference analysis

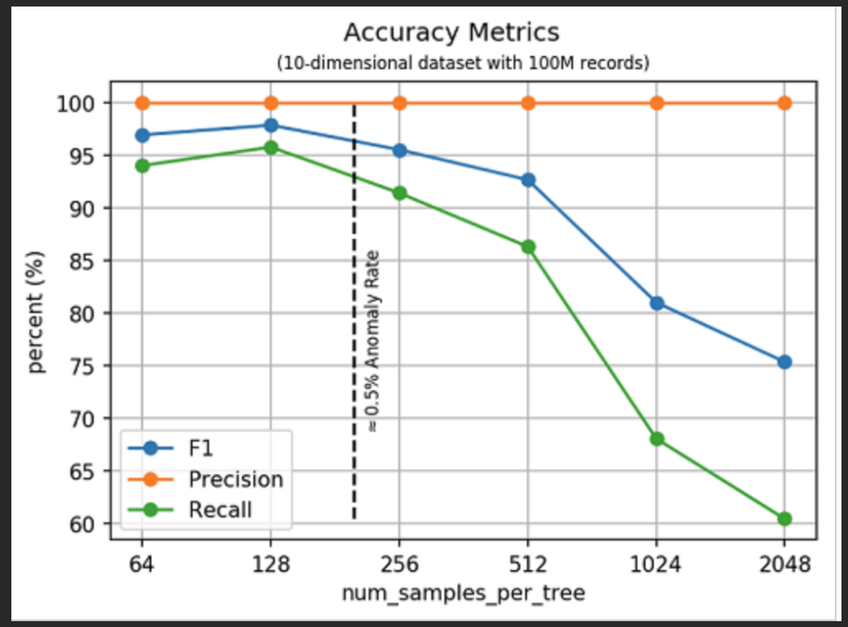
Anomaly detection: Inference analysis

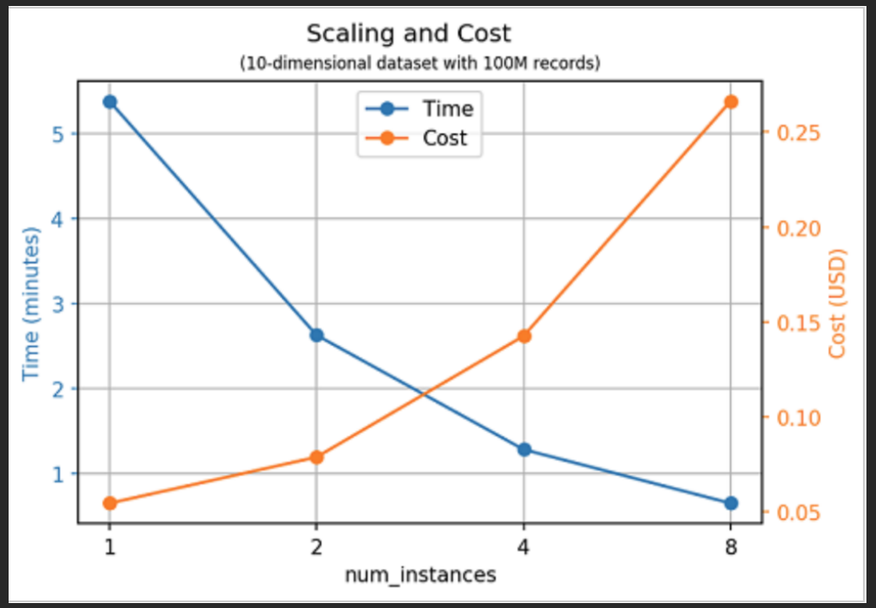
- Deployed Random Cut Forest model calculates *anomaly score* on each point in the dataset
- Mean score of the entire dataset is calculated
- Anomaly scores outside three standard deviations from the mean are considered anomalous
- Accuracy further improved by shingling



Amazon SageMaker Random Cut Forest: Benchmarks







Source: Amazon SageMaker Random Cut Forest algorithm for anomaly detection

Thank you!

AWS re:Invent

aws

Please complete the session survey in the mobile app.