# INNOVATE

AI/ML EDITION

24 February 2022

# Integrate easy-to-use machine learning into your analytics workloads

Niladri Bhattacharya

Senior Analytics Specialist Solution Architect
AWS

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agenda

1. Machine learning for all personas across different stages of analytics pipeline
2. Machine learning in data integration
3. Machine learning to optimize search & log analytics
4. Machine learning on structured data using SQL
5. Machine learning in BI and visualization

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

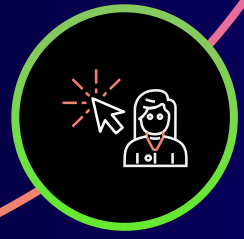
# Personas that use AWS analytics services

Data engineers/database developers

Data analysts

Data scientists

BI professionals



Application developers

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Typical phases in an analytics pipeline

## Right tools for the right job

### Ingest

AWS Glue

Extract, Transform & Load

### Store & catalog

Amazon Simple Storage Service (Amazon S3)

Data lake storage

AWS Glue Data Catalog

Data Catalog

### Analyze

Amazon Redshift

MPP Data Warehouse

Amazon OpenSearch

Search and Log Analytics

Amazon QuickSight

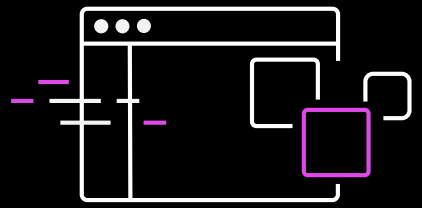
BI and Visualization

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

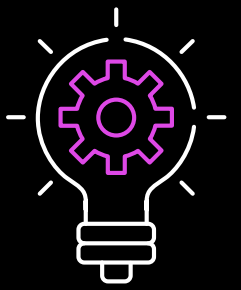
# Benefits of ML

Transform customer experience

Improve business operations



Better and faster decision-making



Innovate product or service

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Machine learning in data integration – AWS Glue

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS Glue: Components

**Data catalog**

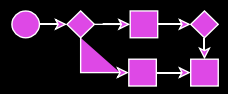
- Hive Metastore compatible with enhanced functionality
- Crawlers automatically extracts metadata and creates tables
- Integrated with Amazon Athena, Amazon EMR and many more

**Job authoring**

- Auto-generates ETL (extract, transform, and load) code, UI based ETL development
- Build on open frameworks – Python/Scala and Apache Spark
- Developer-centric – editing, debugging and sharing

**Job execution**

- Run jobs on a serverless Spark platform
- Provides flexible scheduling , job monitoring and alerting



**Job workflow**

- Orchestrate triggers, crawlers and jobs
- Author & monitor entire flows and integrated alerting

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS Glue FindMatches ML Transform

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# “FindMatches” ML Transform target problems

## Data Integration

*Finding the relationships between multiple datasets, even when those datasets do not share a common & reliable identifier. E.g. find related products, places, suppliers, customers, and more.*

## Deduplication

*Classifying multiple rows in a dataset as “referring to the same thing” (duplicates). Although the rows may not have exact matches. E.g. Improve fraud detection. Identifying duplicate customer accounts, determining when a newly created account is (or might be) a match for a previously known fraudulent user.*

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

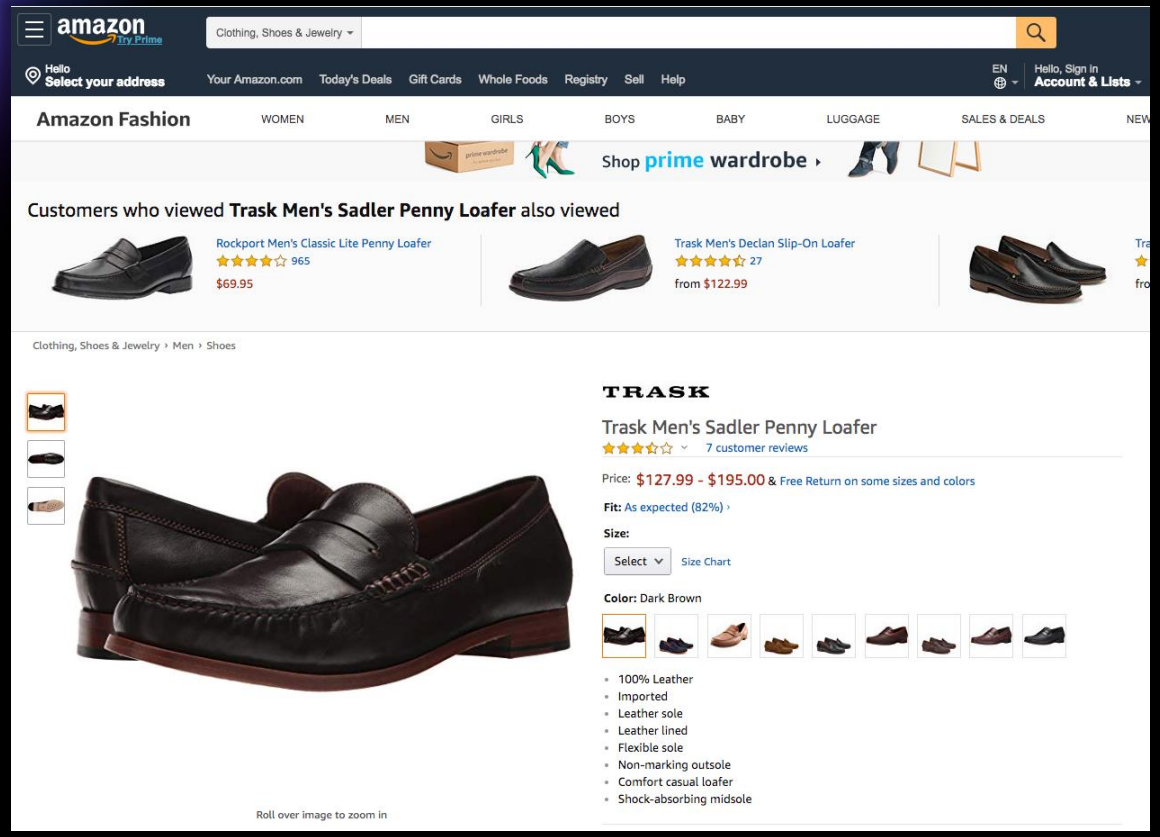
# Identity resolution

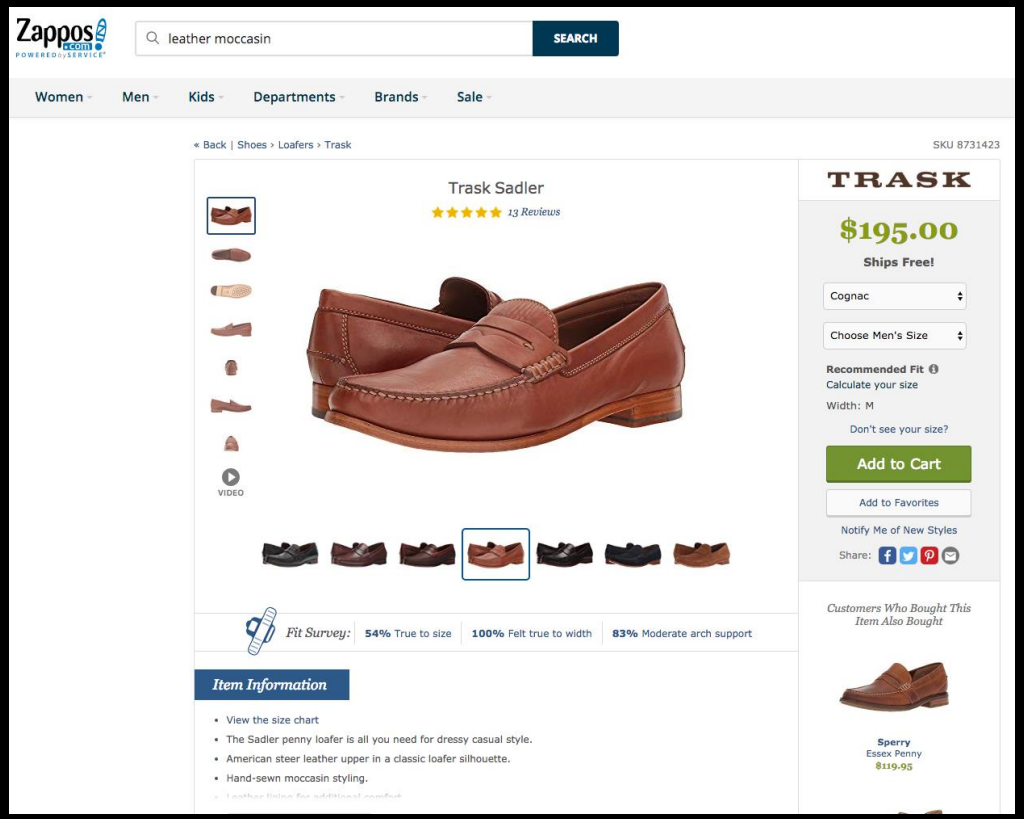
Name:- Niladri Shekhar Bhattacharya
Postal Code:- 791
DOB:- 1/1/1980
Email:- niladri1234@zzzz.com
Contact:- 6599999999

Name:- Niladri S Bhattacharya
Postal Code:- 1000791
DOB:- 1/12/1980
Email:- niladrish1234@zzzz.com
Contact:- 99999999

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Solving matching problems




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How to use “FindMatches” ML Transform in AWS Glue

6 easy steps

Step 1: Catalog your datasets

Step 2: Add a ML Transform

Step 3: Teach the ML Transform

Step 4: Estimate the quality and tune the outcome

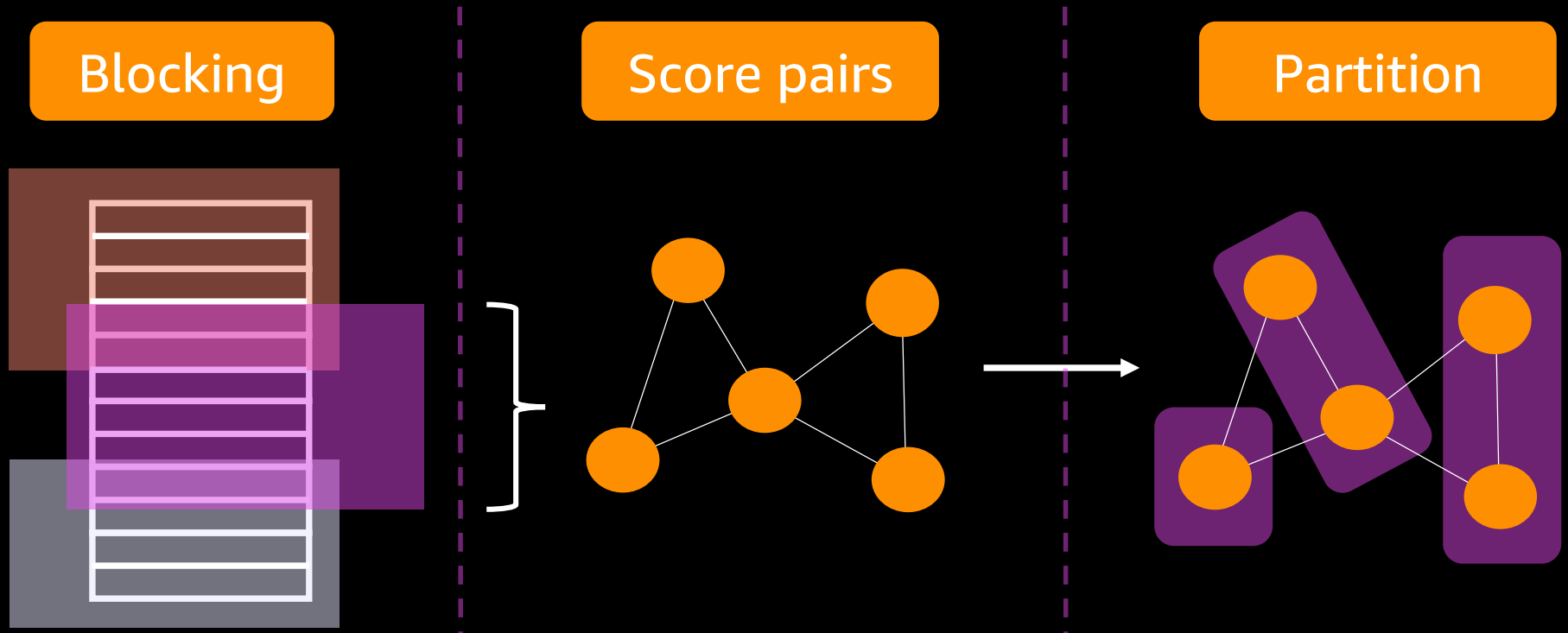
Step 5: Add and run a job with the ML Transform

Step 6: Verify output data

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Fuzzy de-duplication – under the hood

Naïve: looks at all pairs ($N^2$) vs **State-of-the-art**




| **Tune via:** | cost/accuracy slider | positive and negative examples | precision/recall slider |
|---|---|---|---|

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Machine learning to optimize search & log analytics

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Trending use cases for Amazon OpenSearch

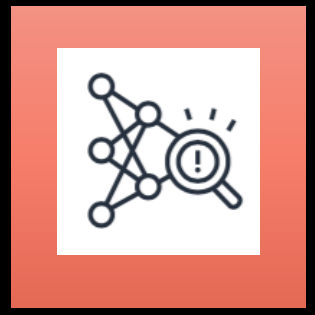
## Time series

Fraud detection

Intrusion detection

System health monitoring

## Search



Product recommendations

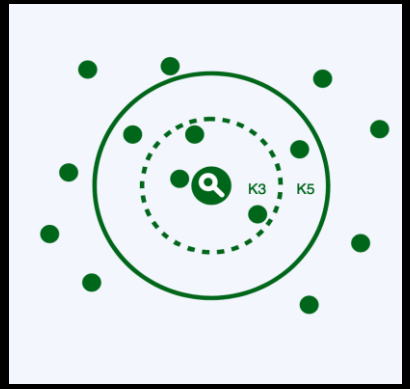
Personalization

Image search

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Machine learning innovations in Amazon OpenSearch

## K-Nearest Neighbor (k-NN) for search




- Find nearest matches in vectorized space in milliseconds
- Updates in real time
- Range of distance techniques including: Euclidean, cosine similarity, hamming distance and FAISS

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# K-Nearest Neighbor (k-NN) for search

## 3 Different methods

- **Approximate k-NN:-** Sacrifice indexing speed and search accuracy in return for performance benefits such as lower latency, smaller memory footprints and more scalable search.

- **Script Score k-NN:-** Brute force, exact k-NN search over “knn_vector” fields or fields that can represent binary objects.

- **Painless extensions:-** Similar to k-NN Script Score but adds the distance functions painless extensions, supports pre-filtering.

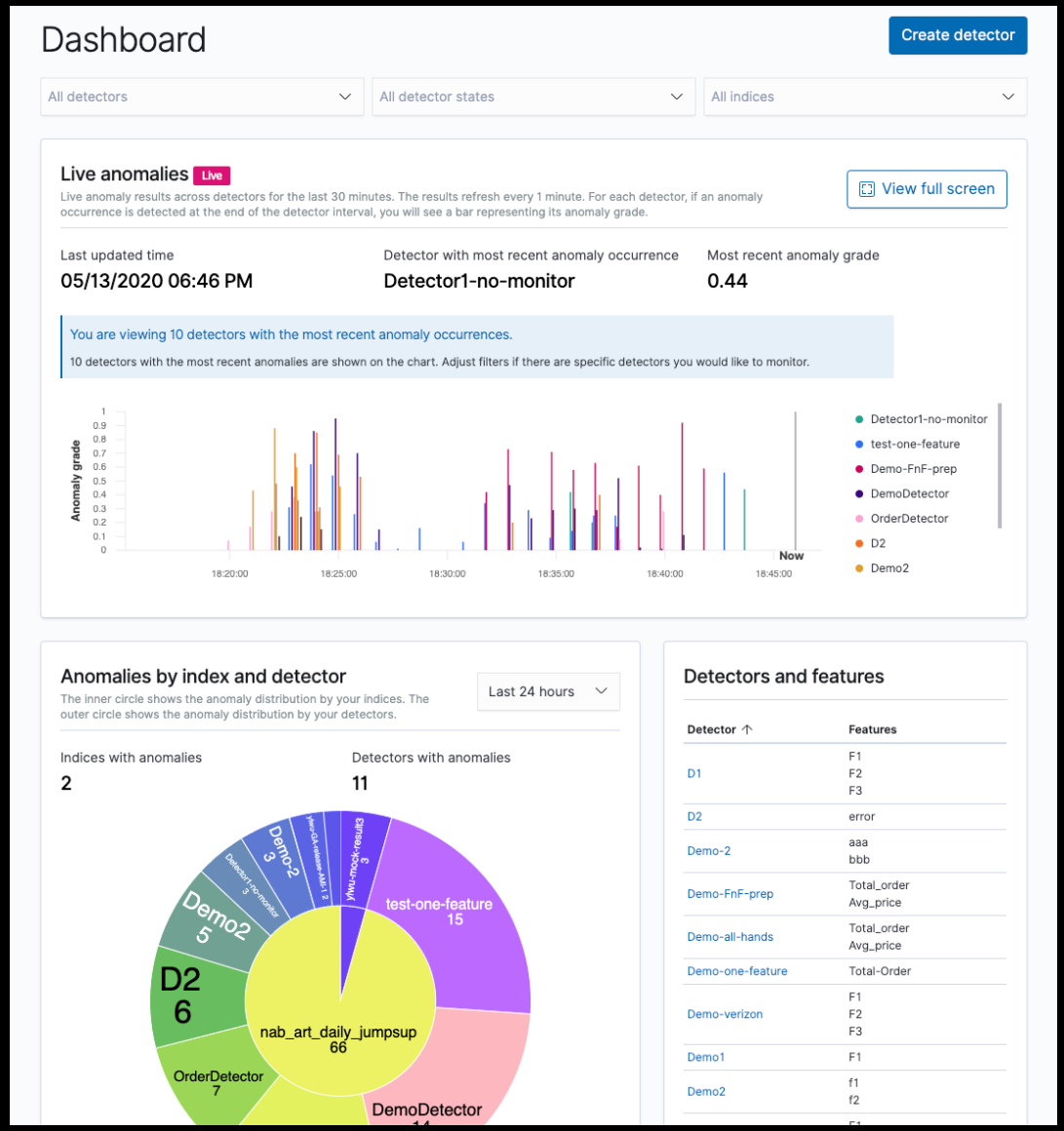
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Machine learning innovations in OpenSearch

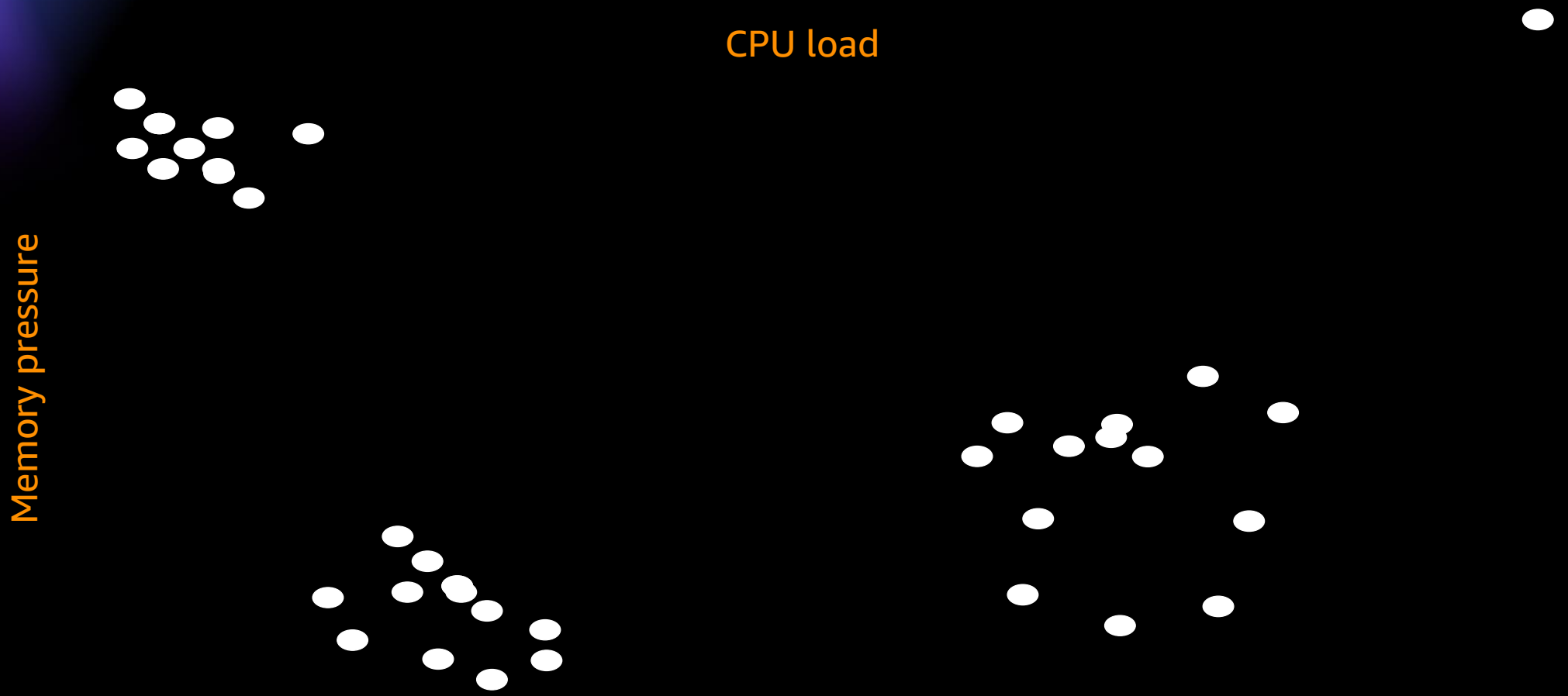
## Anomaly detection for time series

- Detect and mitigate issues faster with anomaly detection
- Performant at scale
  - Machine learning models are distributed and processed across nodes
- Easy to use
  - No machine learning expertise required

Based on Random Cut Forest (RCF)*

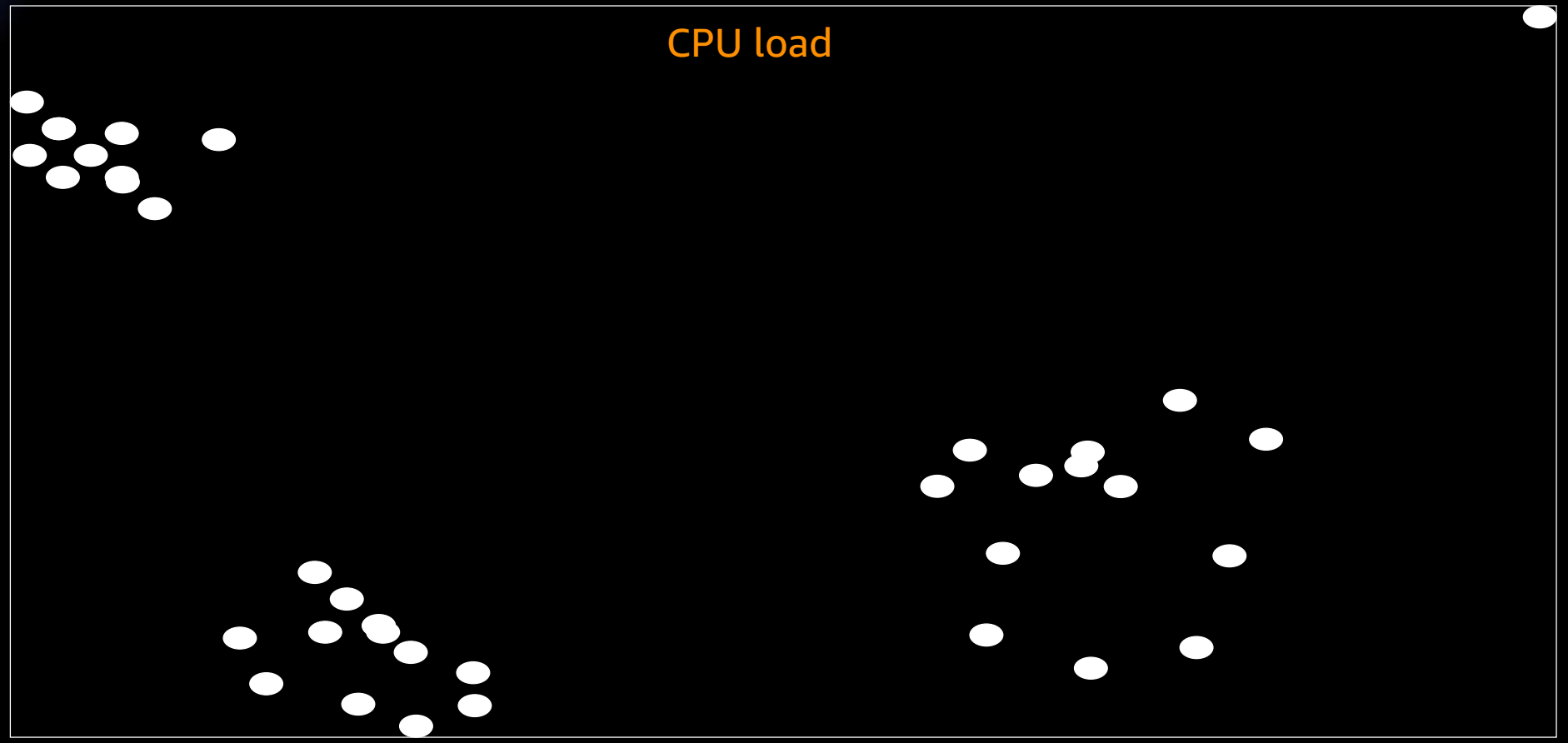


# Random Cut Tree: A sketch



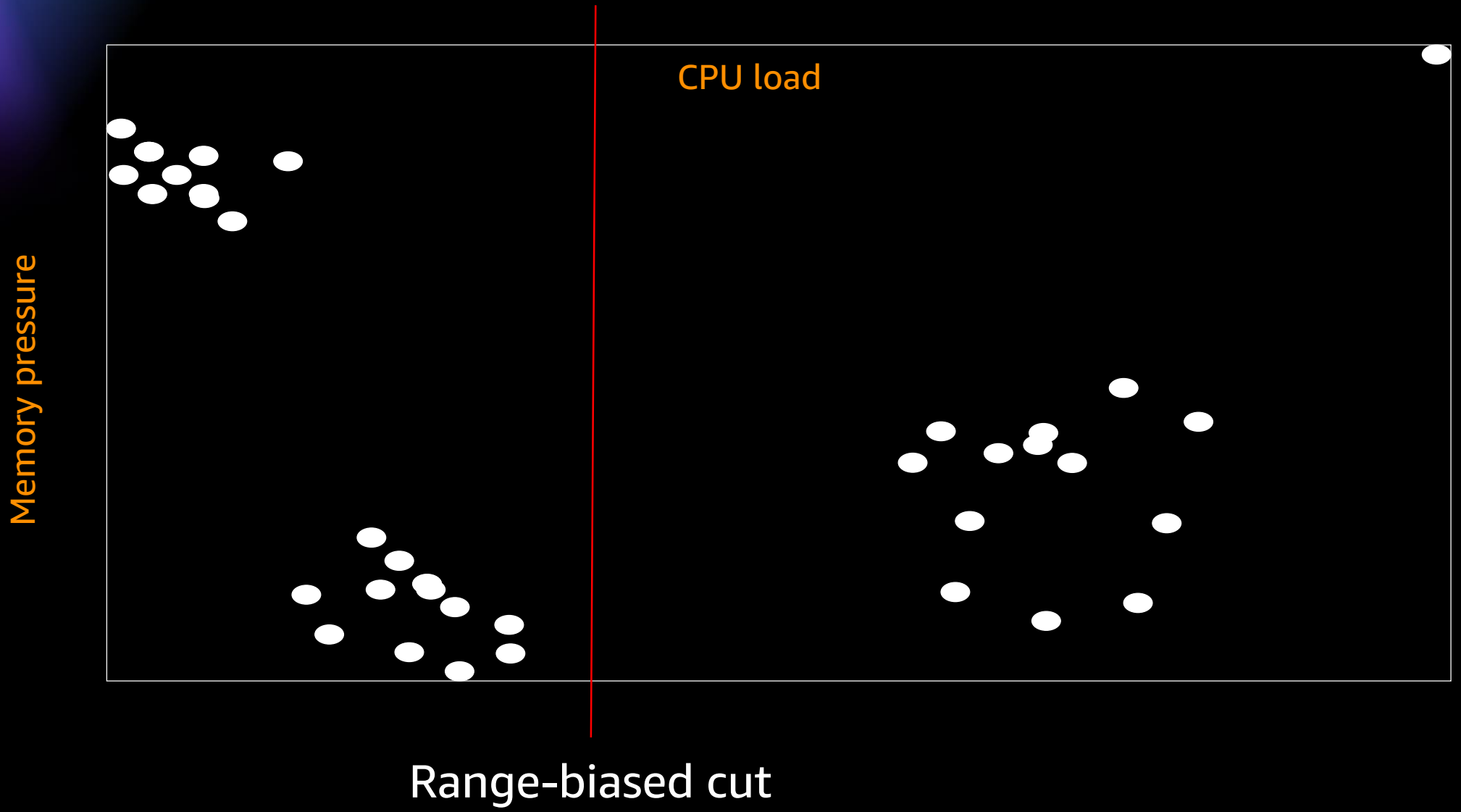

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Random Cut Tree: A sketch



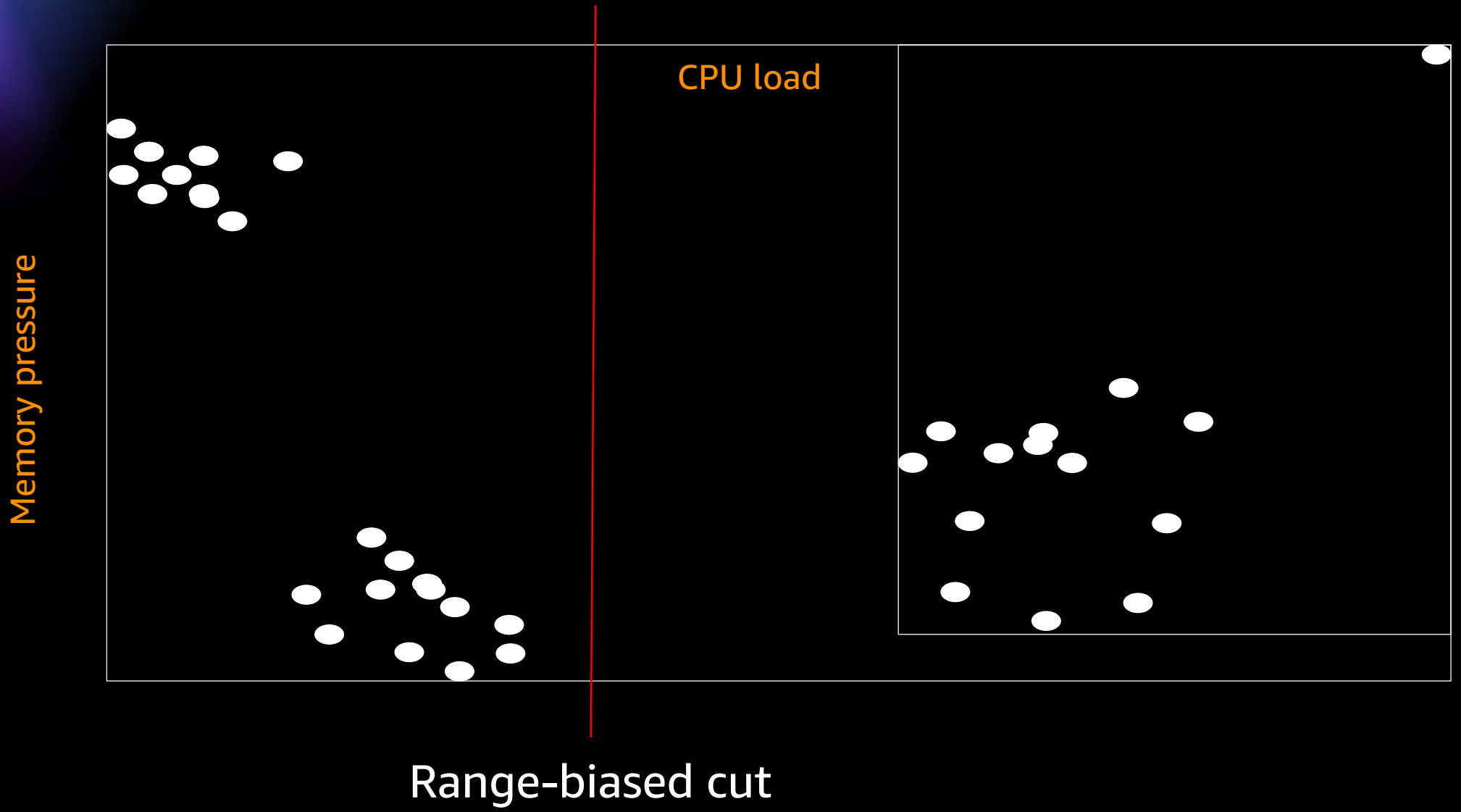

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Random Cut Tree: A sketch



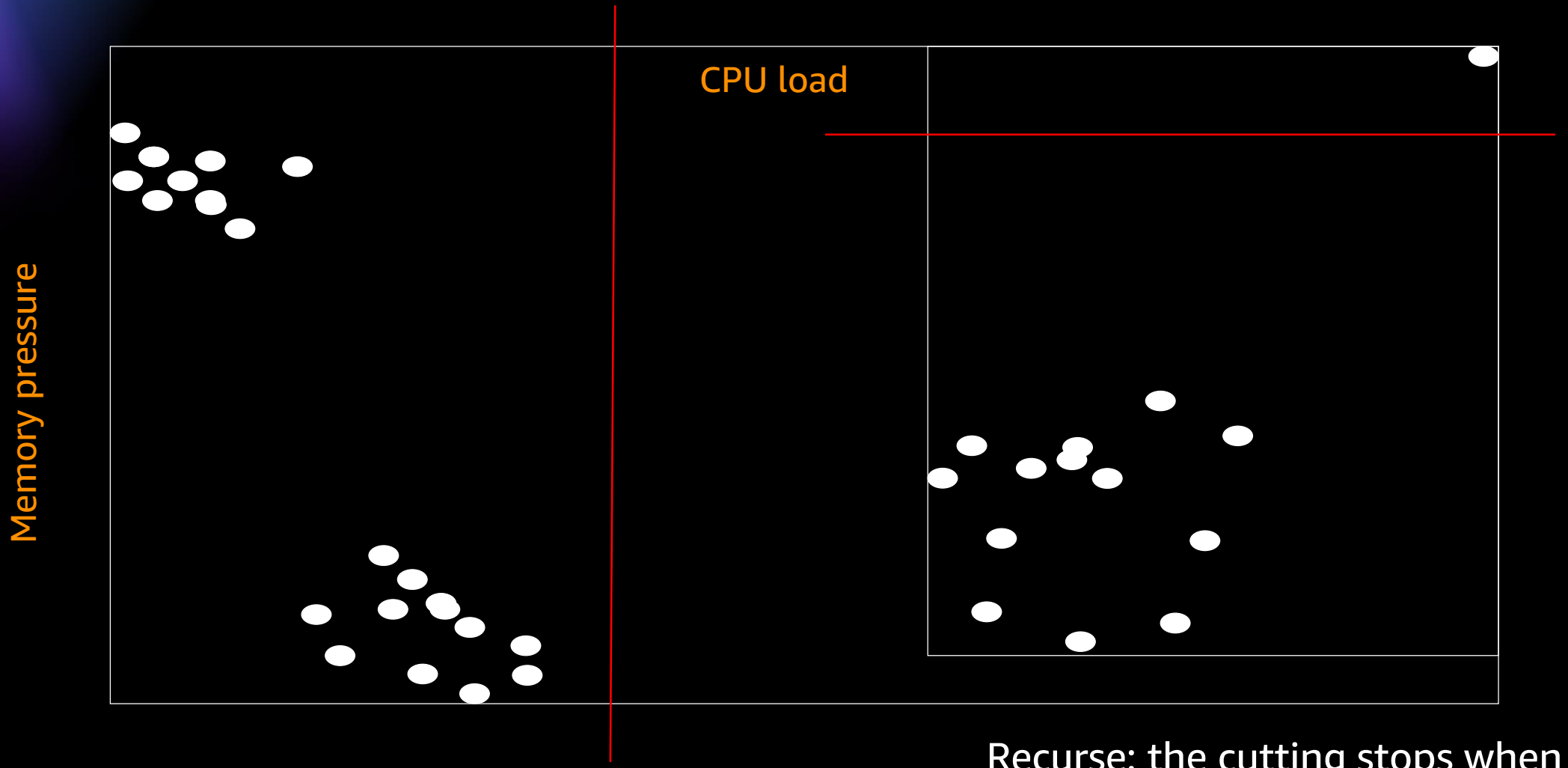

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Random Cut Tree: A sketch




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Random Cut Tree: A sketch



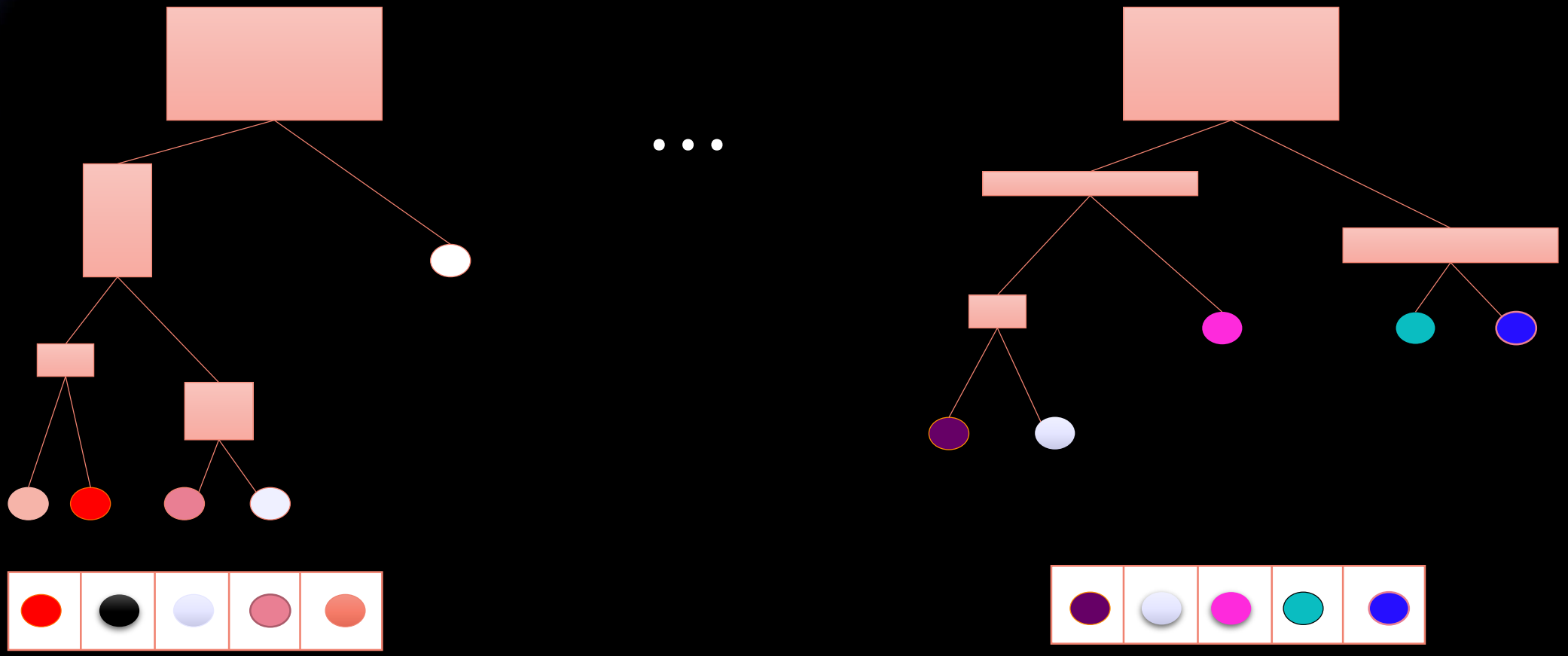

Range-biased cut

Recurse: the cutting stops when each point is isolated

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Random Cut Forest (RCF)



- Each tree built on a random sample
- Streaming: add and delete doesn't require rebuilding trees
- Anomaly score: how much tree description grows

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# High cardinality anomaly detection

**OpenSearch 1.1**

- Support for **1 million** entities
- Underpinned by improved RCF algorithm (RCF 2.0)
- Supports an additional **group-by** category
- One model per entity
- (e.g., entity = host id * error type)
- Use cases: identifying individual hosts with high error count by error types, isolating suspicious users accessing sensitive information by Amazon S3 buckets

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Machine learning on structured data using SQL

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon Redshift ML

## Easily create and train ML models using SQL queries with Amazon SageMaker

Use case: product recommendations, fraud prevention, reduce customer churn

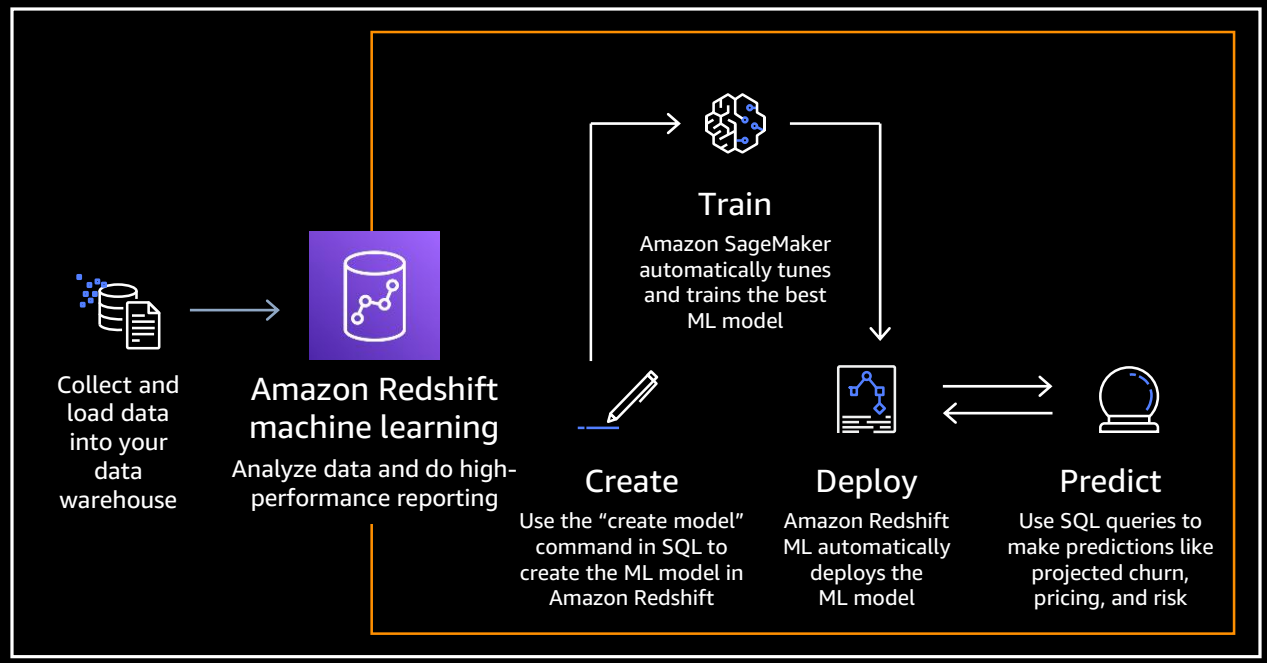
Train and create ML models using SQL

Automatic pre-processing, creation, training, deployment of your model

Deploy inference models locally in Amazon Redshift

Supports ability to bring your SageMaker models for either in-database or remote inference

Supervised and unsupervised trainings




```
CREATE MODEL customer_churn
FROM (SELECT c.age, c.zip, c.monthly_spend, c.monthly_cases,
c.active FROM customer_info_table c)
TARGET c.active
FUNCTION predict_customer_churn
…;
```

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Machine learning in BI and visualization

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon QuickSight

First analytics platform built for the cloud

## Insights on demand

Ask questions using NL

Anomaly Detection and Forecasting

Bring your own model from Amazon SageMaker

## Auto scaling and serverless

Deploy globally to 100k's of users without provisioning servers

Built-in High Availability

AWS does all the work

## Deeply integrated with AWS services

Secure, private access to AWS data

Integrated S3 data lake permissions

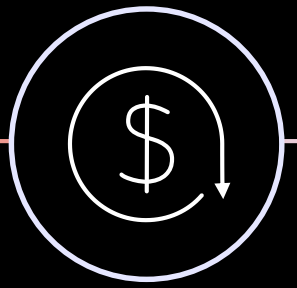
## Embeddable

Share insights with external parties

Rich set of APIs and SDKs

Multi-tenant & secure

## Low Cost



Cost effective at any scale

Usage-based & fixed-price options

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Augment your data with machine learning

NEW

## Auto narratives

Summarize your business metrics in plain language

## ML predictions

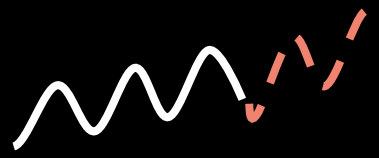
Visualize and build predictive dashboards with Amazon SageMaker models

## Q

Ask questions using natural language

## Anomaly detection

Discover unexpected trends and outliers against millions of business metrics



## Forecasting

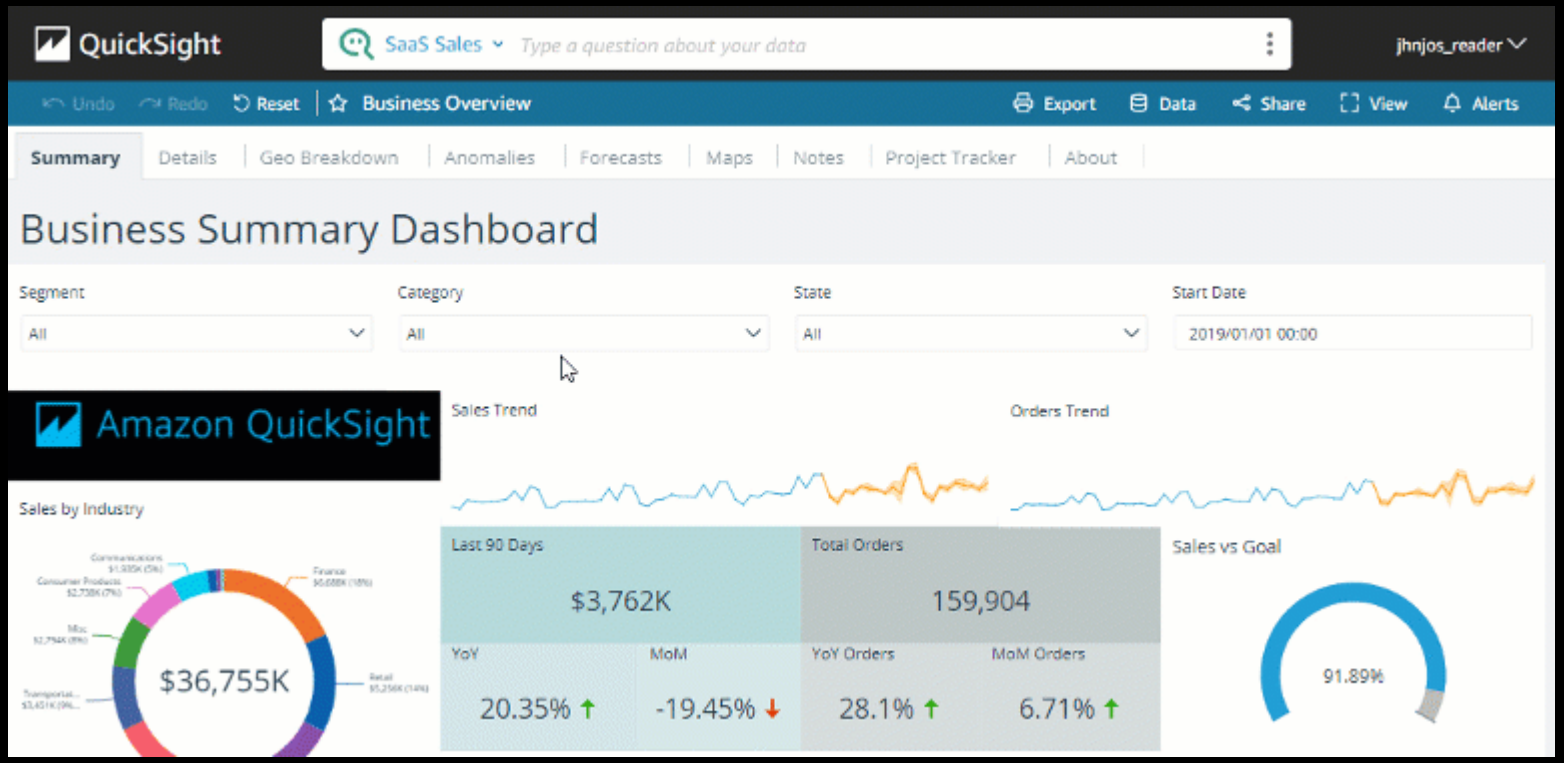
Machine learning forecasting with point-and-click simplicity

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon QuickSight Q

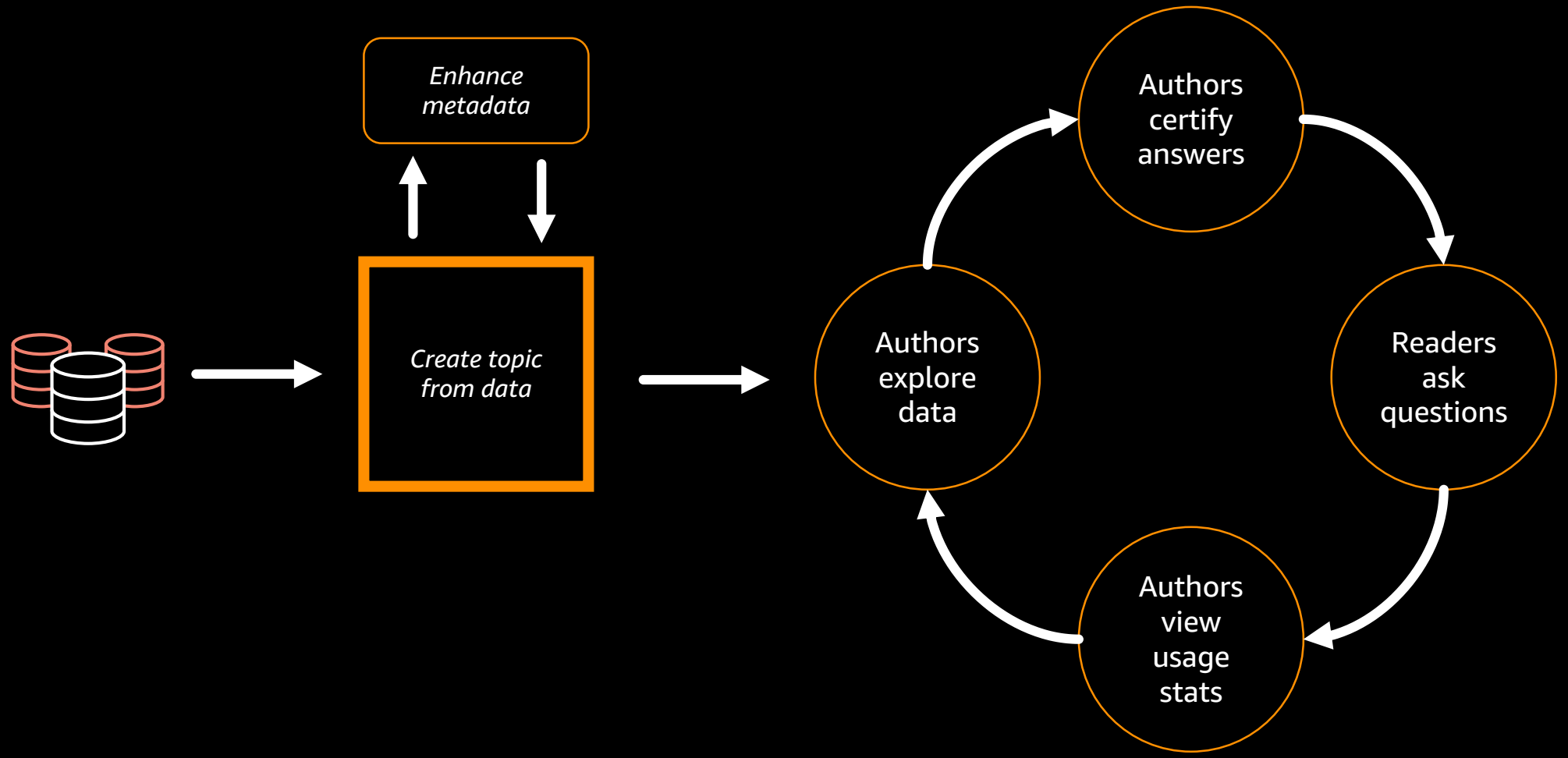
- Ask natural language questions about all your data.
- Get instant answers with visualization
- Interact with answers
- Feedback




**Go beyond what's presented on the dashboard**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Amazon QuickSight Q process




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How to get the most out of Amazon Quicksight Q

## Creating effective topics

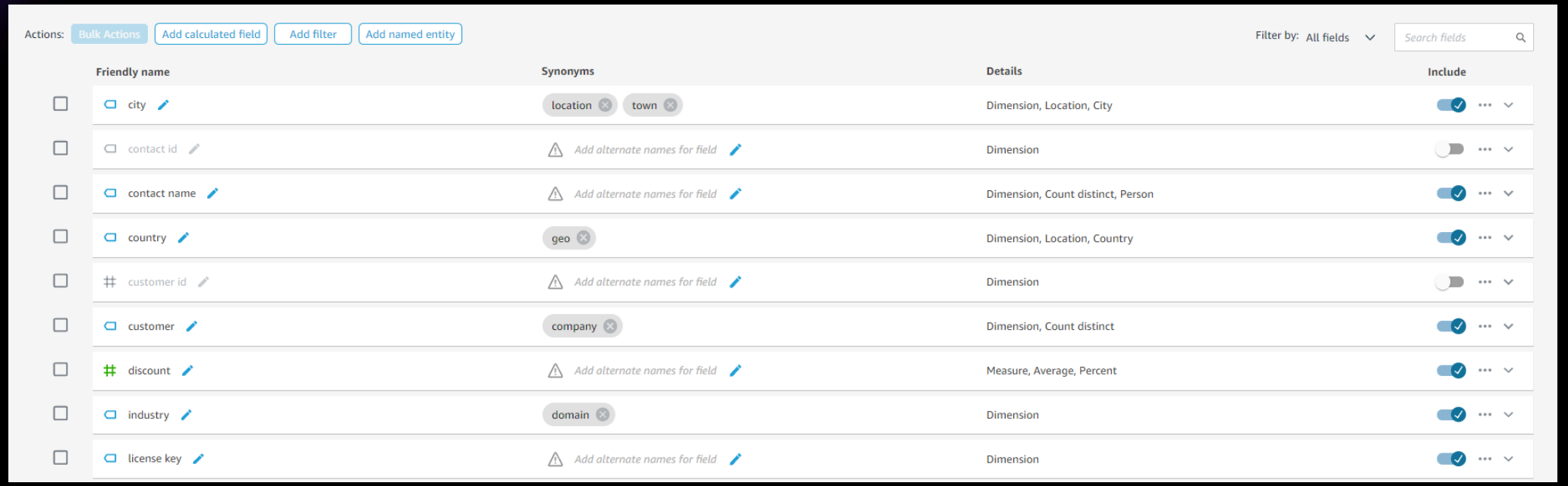
1. Give datasets friendly names and descriptions:- E.g. rename D_CUST_DLY_ORD_DTL to customer daily order details
2. Tell Q how to use date fields in your datasets
3. Exclude unused fields
4. Rename fields to be natural-language-friendly
5. Add synonyms to fields and field values
6. Tell Q more about your fields. E.g. Assign field roles, set field aggregations, formatting field values and semantic types

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How to get the most out of Amazon Quicksight Q

## Creating effective topics

Curate topics with business language and semantics




**Add domain-specific or organizational language to Q**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Recap

1. Solving data integration and de-duplication problems using AWS Glue FindMatches ML Transform
2. Similarity search using k-NN and anomaly detection using random cut forest in Amazon OpenSearch
3. Amazon Redshift ML
4. ML capabilities in Amazon QuickSight especially Amazon QuickSight Q

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Visit the AI & Machine Learning resource hub for more resources

Dive deeper into these resources, get inspired and learn how you can use AI and machine learning to accelerate your business outcomes.

- The machine learning journey e-book
- 7 leading machine learning use cases e-book
- A strategic playbook for data, analytics, and machine learning e-book
  Accelerate machine learning innovation with the right cloud services & infrastructure e-book
- Choosing the right compute infrastructure for machine learning e-book
- Improving service and reducing costs in contact centers e-book
- Why ML is essential in your fight against online fraud e-book
- … and more!

https://bit.ly/3mwi59V

**Visit resource hub**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# AWS Machine Learning (ML) Training and Certification

## AWS is how you build machine learning skills

Courses built on the curriculum leveraged by Amazon's own teams. Learn from the experts at AWS.

[aws.training/machinelearning](aws.training/machinelearning)

## Flexibility to learn your way

Learn online with on-demand digital courses or live with virtual instructor-led training, plus hands-on labs and opportunities for practical application.

[explore.skillbuilder.aws/learn](explore.skillbuilder.aws/learn)

## Validate your expertise

Demonstrate expertise in building, training, tuning, and deploying machine learning models with an industry-recognized credential.

[aws.amazon.com/certification](aws.amazon.com/certification)

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you for attending AWS Innovate – AI/ML Edition

We hope you found it interesting! A kind reminder to **complete the survey.**
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

aws-apj-marketing@amazon.com

twitter.com/AWSCloud

facebook.com/AmazonWebServices

youtube.com/user/AmazonWebServices

slideshare.net/AmazonWebServices

twitch.tv/aws

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!

Niladri Bhattacharya

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.