Integrating Splunk with AWS services:

Using Redshift, Elastic Map Reduce (EMR), Amazon Machine Learning & S3 to gain actionable insights via predictive analytics via Splunk

Patrick Shumate

Solutions Architect, Amazon Web Services

Olivier de Garrigues

Professional Services , Splunk

amazon web services

Disclaimer

During the course of this presentation, we may make forward looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not, be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Who we are?

- **Patrick Shumate**
- Virginia, US
- Solutions Architect
- Big data/ Security SME

Olivier de Garrigues

London, UK

Professional Services

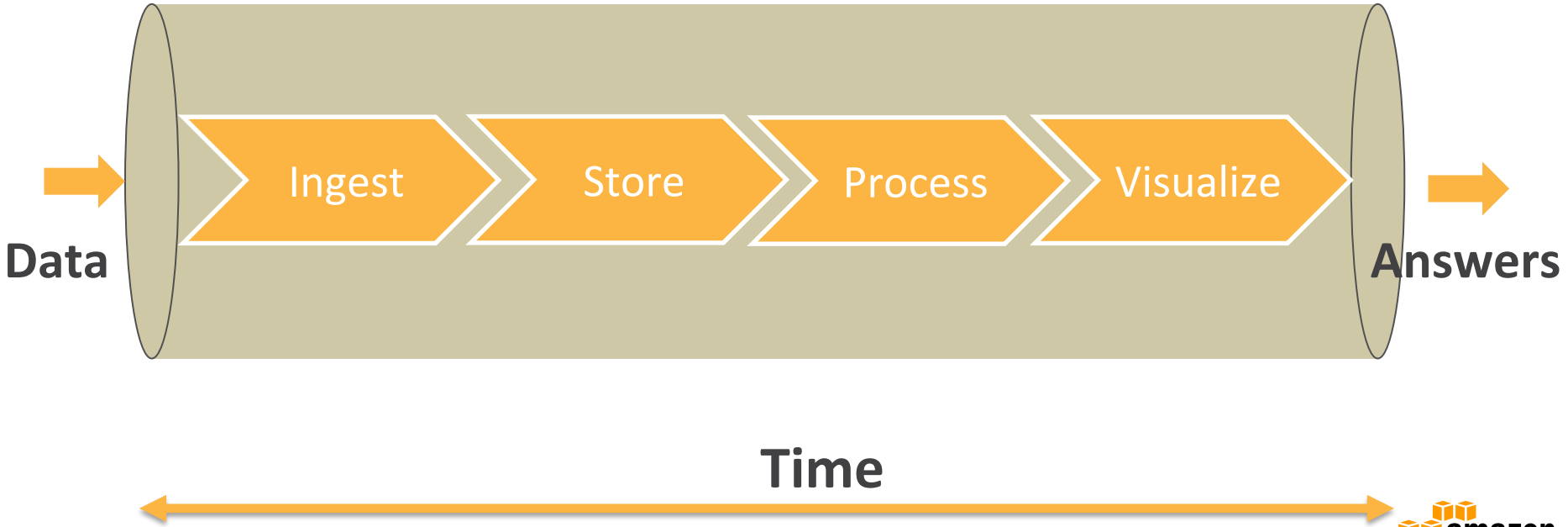
Analytics SME

Agenda

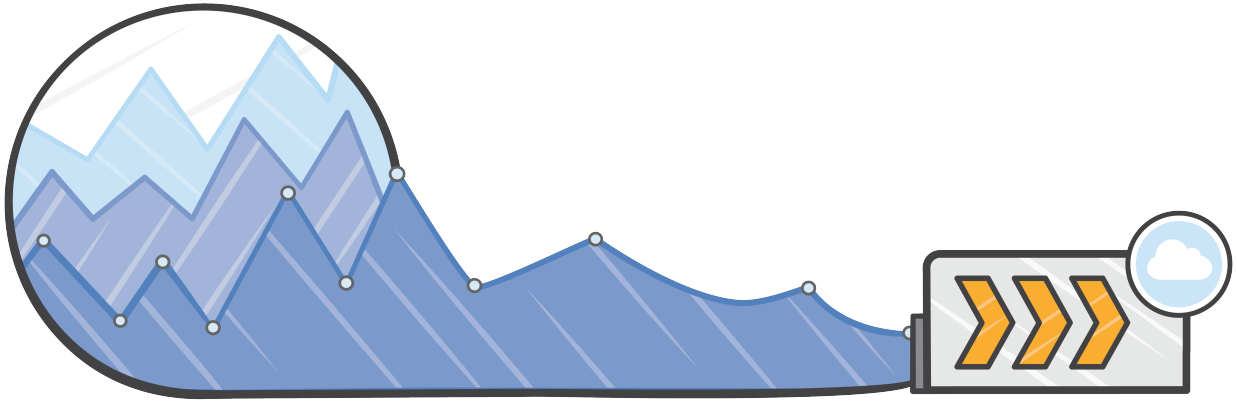
- Overview of Services
- Integration Demo of RedShift and Amazon ML
- Resources to help you get started
- Q&A

What Tools Should I Use?

Simplify Big Data Processing



Ingest: The act of collecting and storing data



amazon web services

Amazon
Kinesis

Real-time processing of streaming data

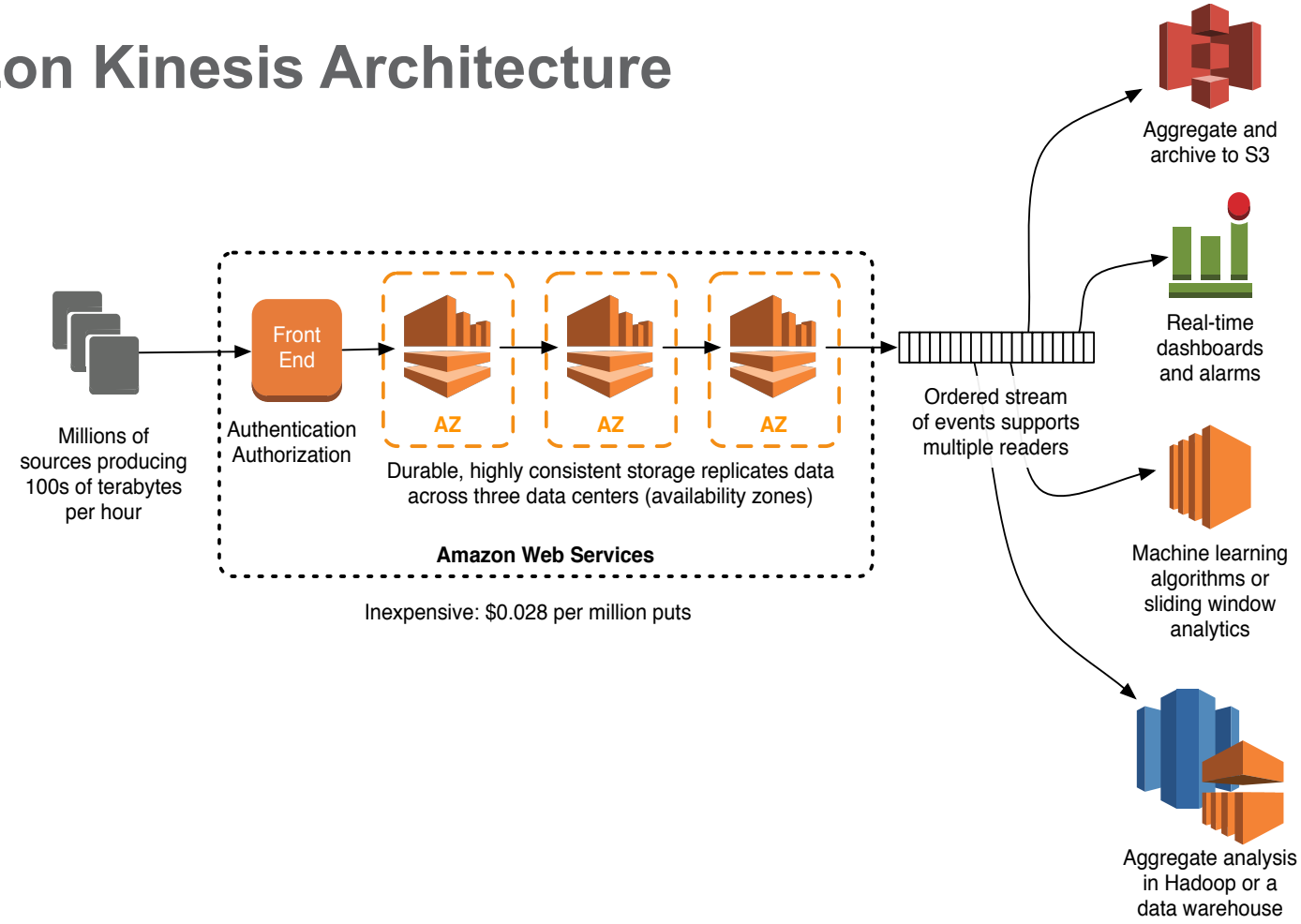
High throughput

Elastic

Easy to use

Connectors for EMR, S3, Redshift, DynamoDB

Amazon Kinesis Architecture



Amazon Kinesis Modular Input

Utilizes Modular Inputs to index from the stream

amazon web services

Integrates AWS Lambda for collection via HTTP

Amazon Kinesis

Amazon DynamoDB

Amazon S3

and other Amazon services

amazon web services

Storage

amazon web services

Amazon
S3

Store anything

Object storage

Scalable

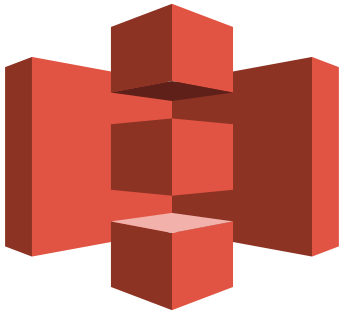
Designed for 99.999999999% durability

amazon web services

Why is Amazon S3 good for Big Data?

- No limit on the number of Objects
- Object size up to 5TB
- Central data storage for all systems
- High bandwidth
- 99.999999999% durability
- Versioning, Lifecycle Policies
- Glacier Integration

Aggregate All Data in S3 Surrounded by a collection of the right tools



Amazon S3

EMR

Kinesis

Data Pipeline

Redshift

DynamoDB

RDS

splunk>

Storm

Spark

Spark Streaming

amazon web services

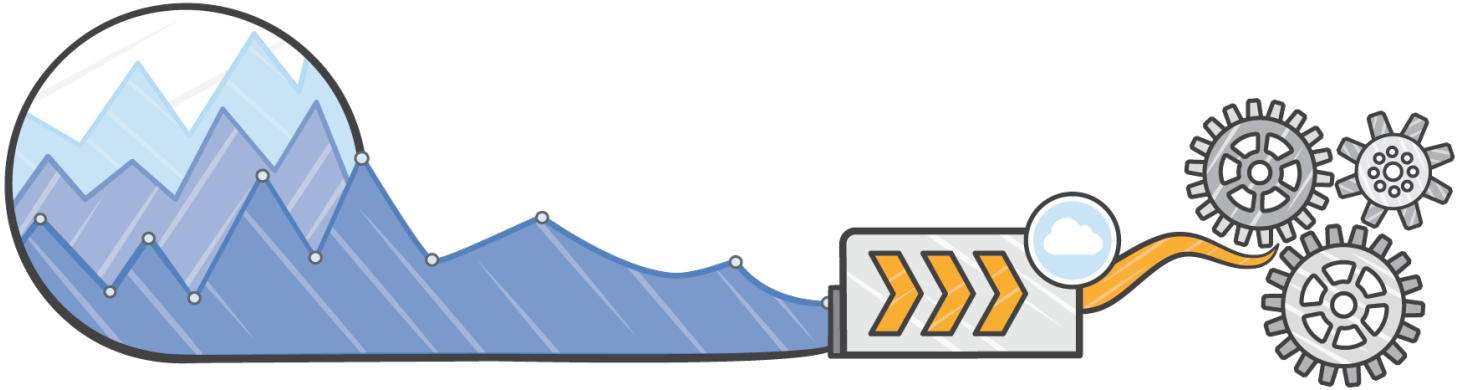
Splunk Add-on for Amazon Web Services

Collect events, alerts, performance metrics configuration snapshots, and billing information from the AWS CloudTrail, CloudWatch, and Config services

gather log data from generic S3 buckets

amazon web services

Process and Analyze



amazon web services

Amazon
Elastic
MapReduce

Hadoop/HDFS clusters

Hunk, Hive, Pig, Impala, HBase

Easy to use; fully managed

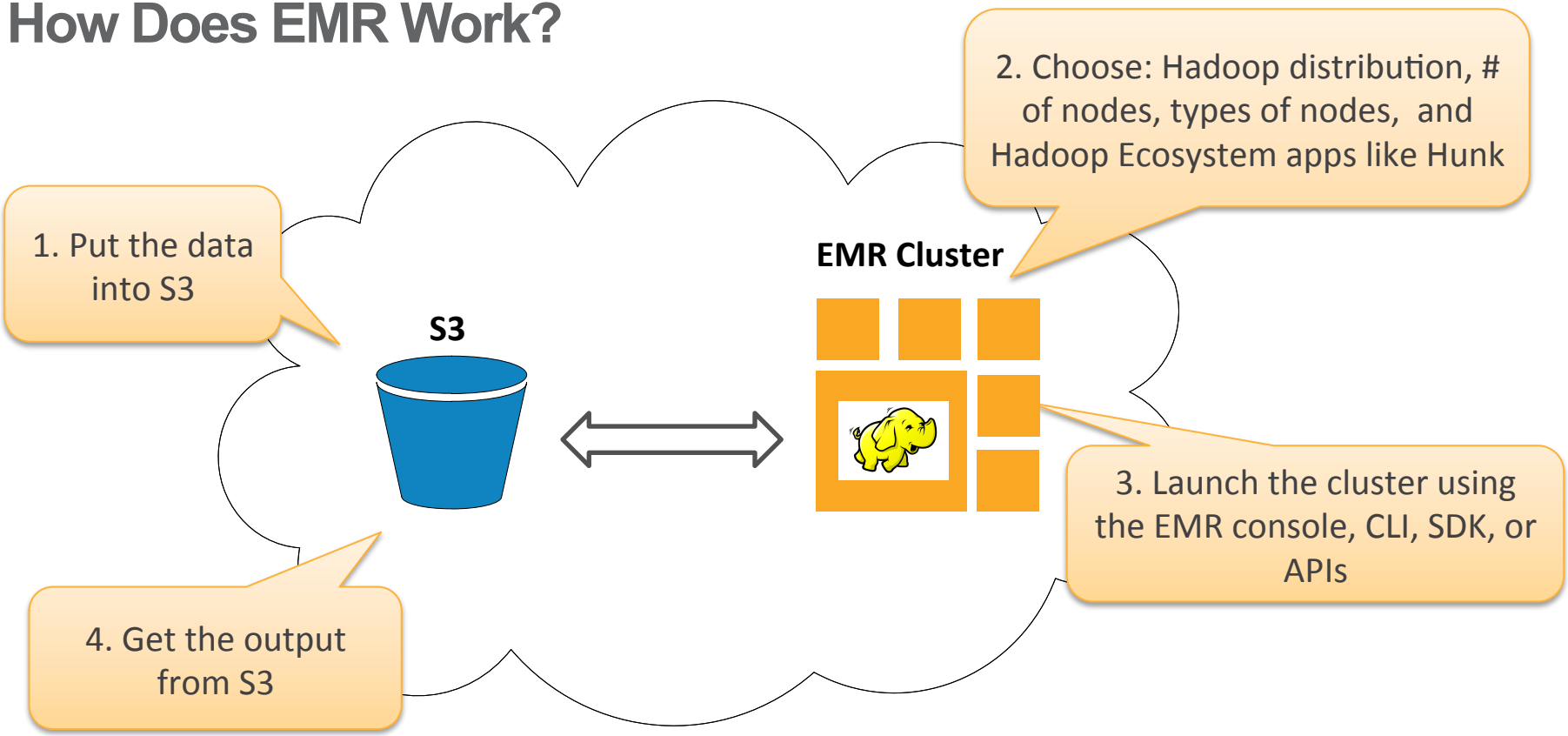
On-demand and spot pricing

Tight integration with S3,

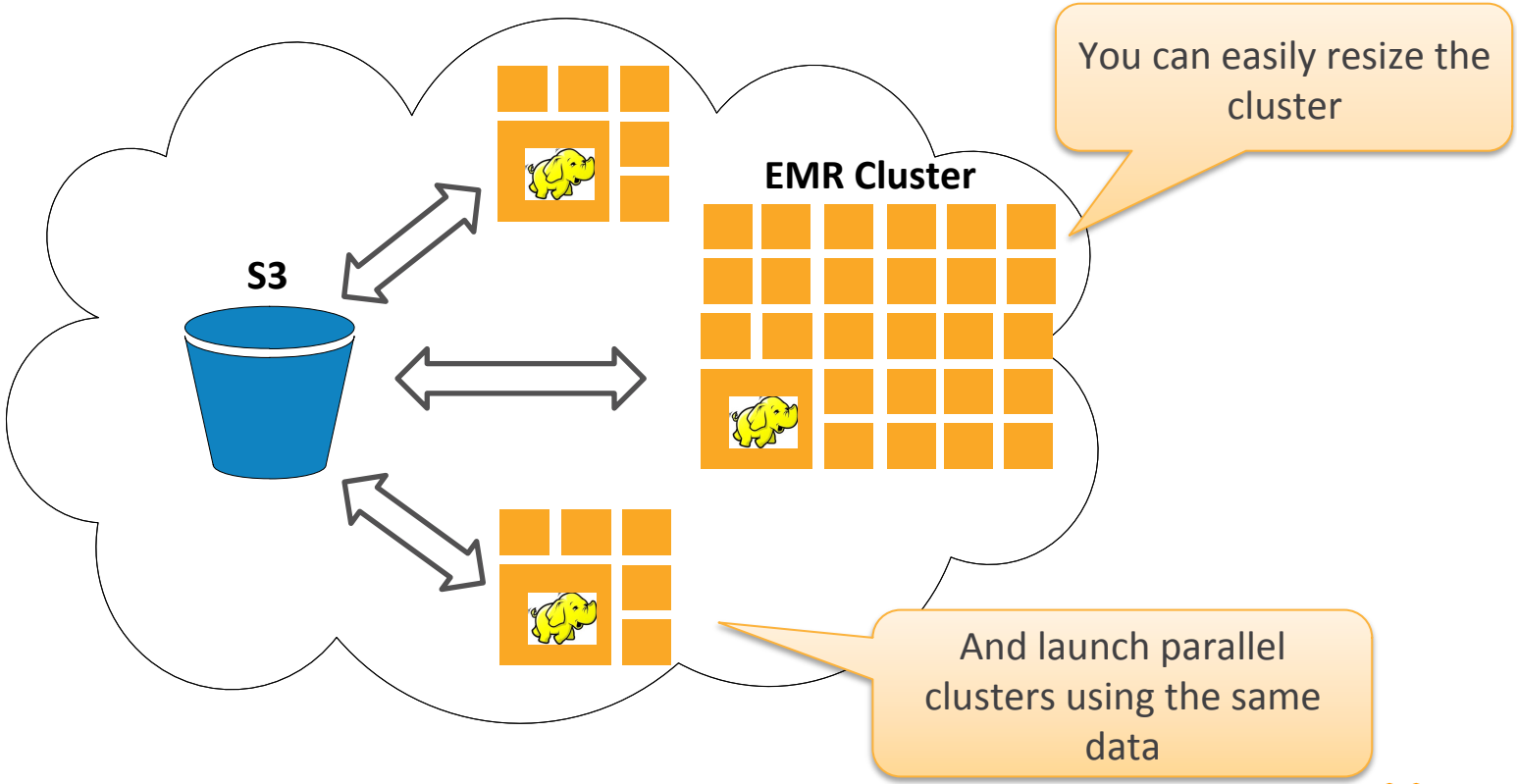
DynamoDB, and Kinesis

amazon web services

How Does EMR Work?



How Does EMR Work?



amazon web services

Hunk®

Full-Featured, Integrated Analytics
Fast to Deploy and Drive Value
Interactive Search
Report Acceleration
Results Preview
Very cost-effective

amazon web services

Amazon
Redshift

Columnar data warehouse

ANSI SQL compatible

Massively parallel

Petabyte scale

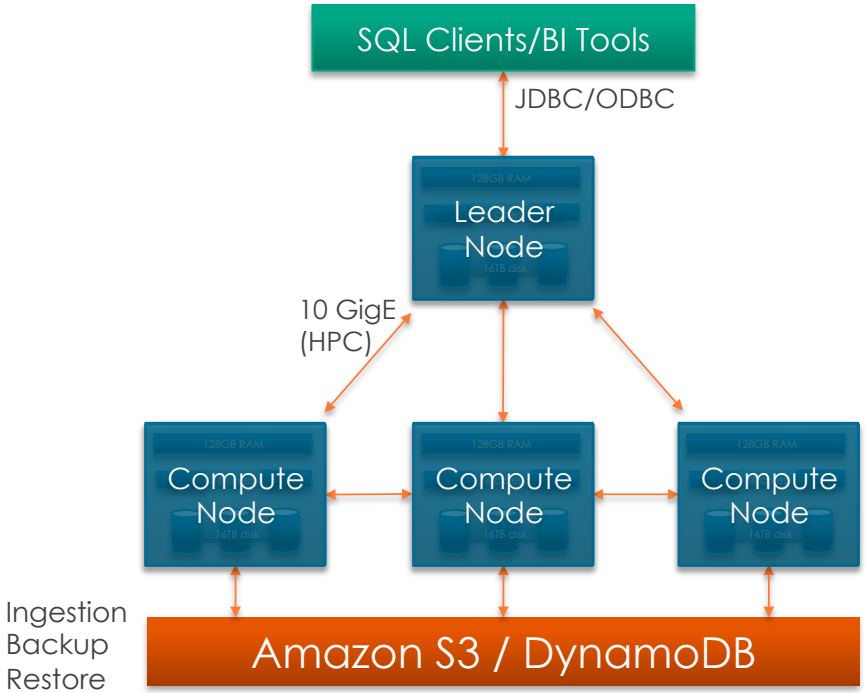
Fully-managed

Very cost-effective

amazon web services

Amazon Redshift architecture

- Leader Node
  - SQL endpoint
  - Stores metadata
  - Coordinates query execution
- Compute Nodes
  - Local, columnar storage
  - Execute queries in parallel
  - Load, backup, restore via Amazon S3
  - Parallel load from Amazon DynamoDB
- Hardware optimized for data processing
- Two hardware platforms
  - DW1: HDD; scale from 2TB to 1.6PB
  - DW2: SSD; scale from 160GB to 256TB



amazon web services

Amazon Redshift dramatically reduces I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

ID	Age	State	Amount
123	20	CA	500
345	25	WA	250
678	40	FL	125
957	37	WA	375

- With row storage you do unnecessary I/O
- To get total amount, you have to read everything

Amazon Redshift dramatically reduces I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

ID	Age	State	Amount
123	20	CA	500
345	25	WA	250
678	40	FL	125
957	37	WA	375

- With column storage, you only read the data you need

Amazon Redshift dramatically reduces I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

```
analyze compression listing;

  Table  |     Column      | Encoding
---------+-----------------+----------
 listing | listid          | delta
 listing | sellerid        | delta32k
 listing | eventid         | delta32k
 listing | dateid          | bytedict
 listing | numtickets      | bytedict
 listing | priceperticket  | delta32k
 listing | totalprice      | mostly32
 listing | listtime        | raw
```

- Columnar compression saves space & reduces I/O
- Amazon Redshift analyzes and compresses your data

Amazon Redshift dramatically reduces I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

- Track of the minimum and maximum value for each block
- Skip over blocks that don't contain the data needed for a given query
- Minimize unnecessary I/O

Amazon Redshift dramatically reduces I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

- Use direct-attached storage to maximize throughput
- Hardware optimized for high performance data processing
- Large block sizes to make the most of each read
- Amazon Redshift manages durability for you

Database
Connector 2

Inputs to import structured data

Outputs export data insights to a legacy database

Lookups add meaningful information to events

Amazon
Machine
Learning

robust, cloud-based

easy for developers of all skill levels to use

Wizard driven process to build models

API for automation

Batch and real time predictions

Easy to use and developer-friendly

Use the intuitive, powerful service console to build and explore your initial models

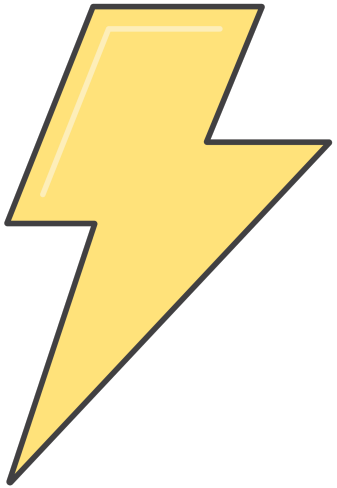
- Data retrieval
- Model training, quality evaluation, fine-tuning
- Deployment and management

Automate model lifecycle with fully featured APIs and SDKs

- Java, Python, .NET, JavaScript, Ruby, PHP

Easily create smart iOS and Android applications with AWS Mobile SDK

Powerful machine learning technology



Based on Amazon's battle-hardened internal systems

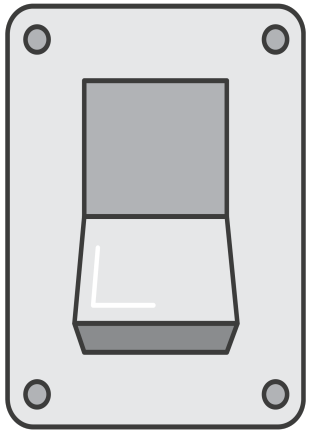
Not just the algorithms:

- Smart data transformations
- Input data and model quality alerts
- Built-in industry best practices

Grows with your needs

- Train on up to 100 GB of data
- Generate billions of predictions
- Obtain predictions in batches or real-time

Fully managed model and prediction services



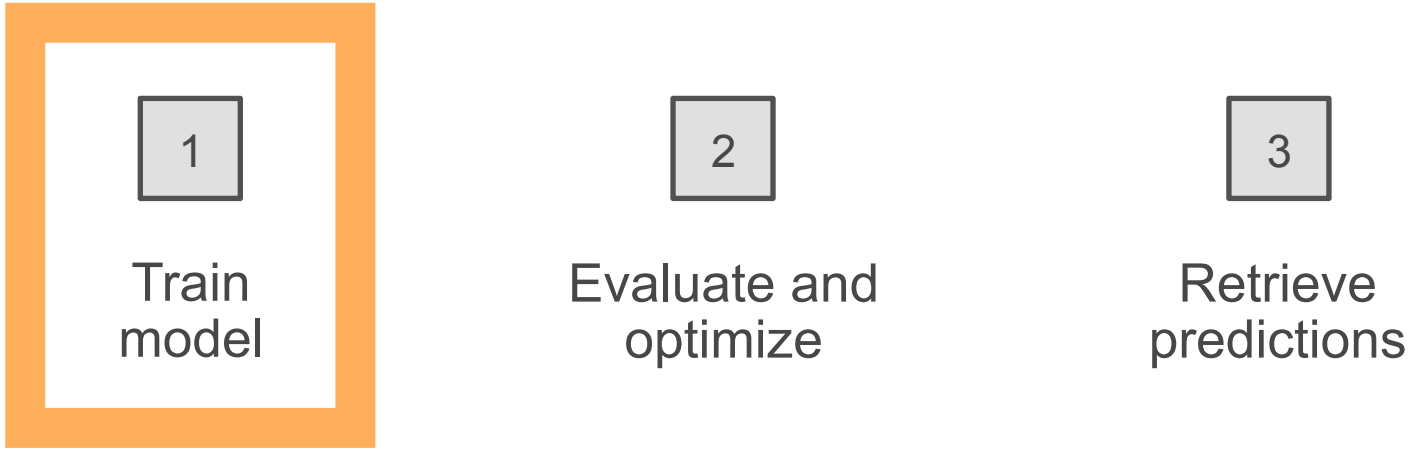
End-to-end service, with no servers to provision and manage

One-click production model deployment

Programmatically query model metadata to enable automatic retraining workflows

Monitor prediction usage patterns with Amazon CloudWatch metrics

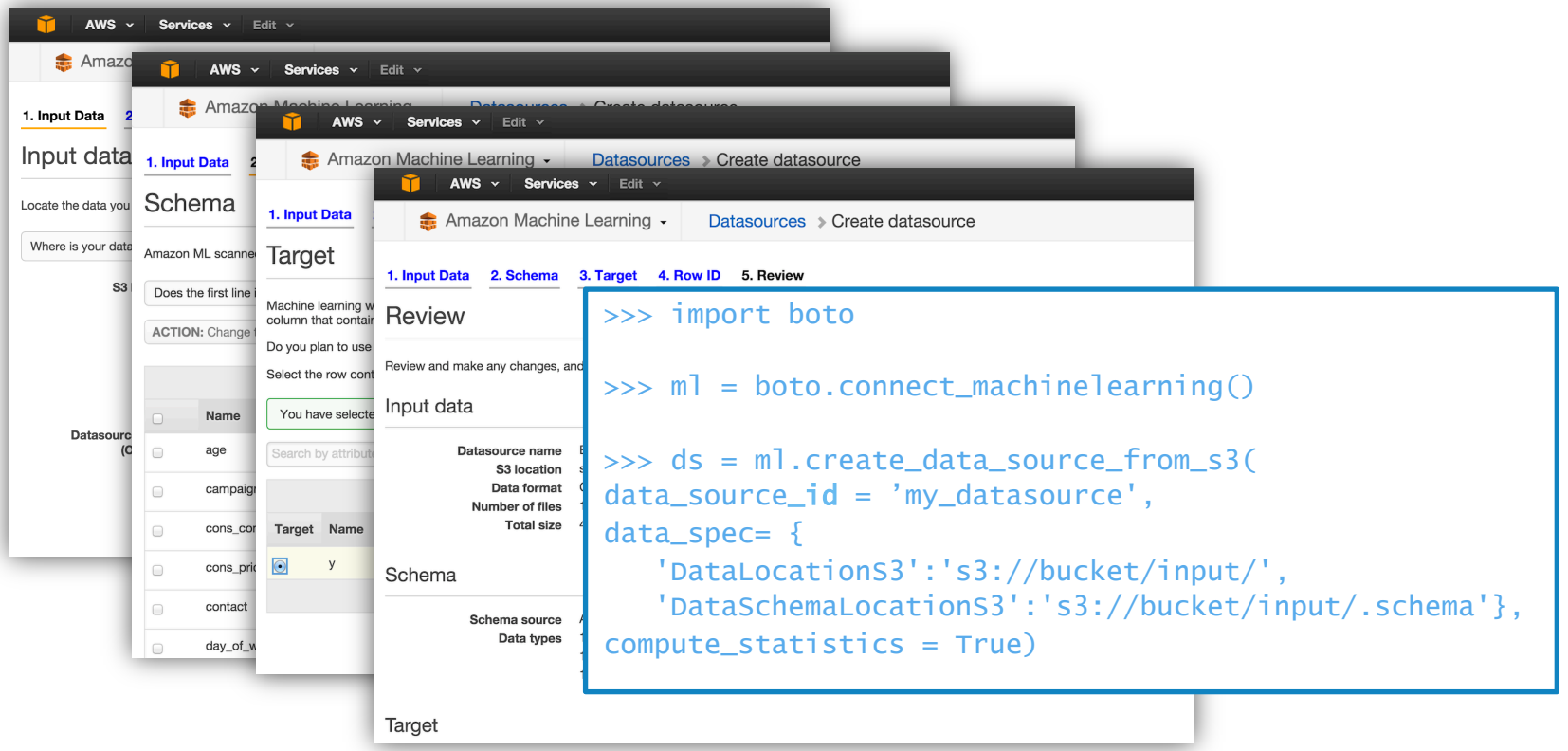
Building smart applications with Amazon ML



- Create a Datasource object pointing to your data
- Explore and understand your data
- Transform data and train your model

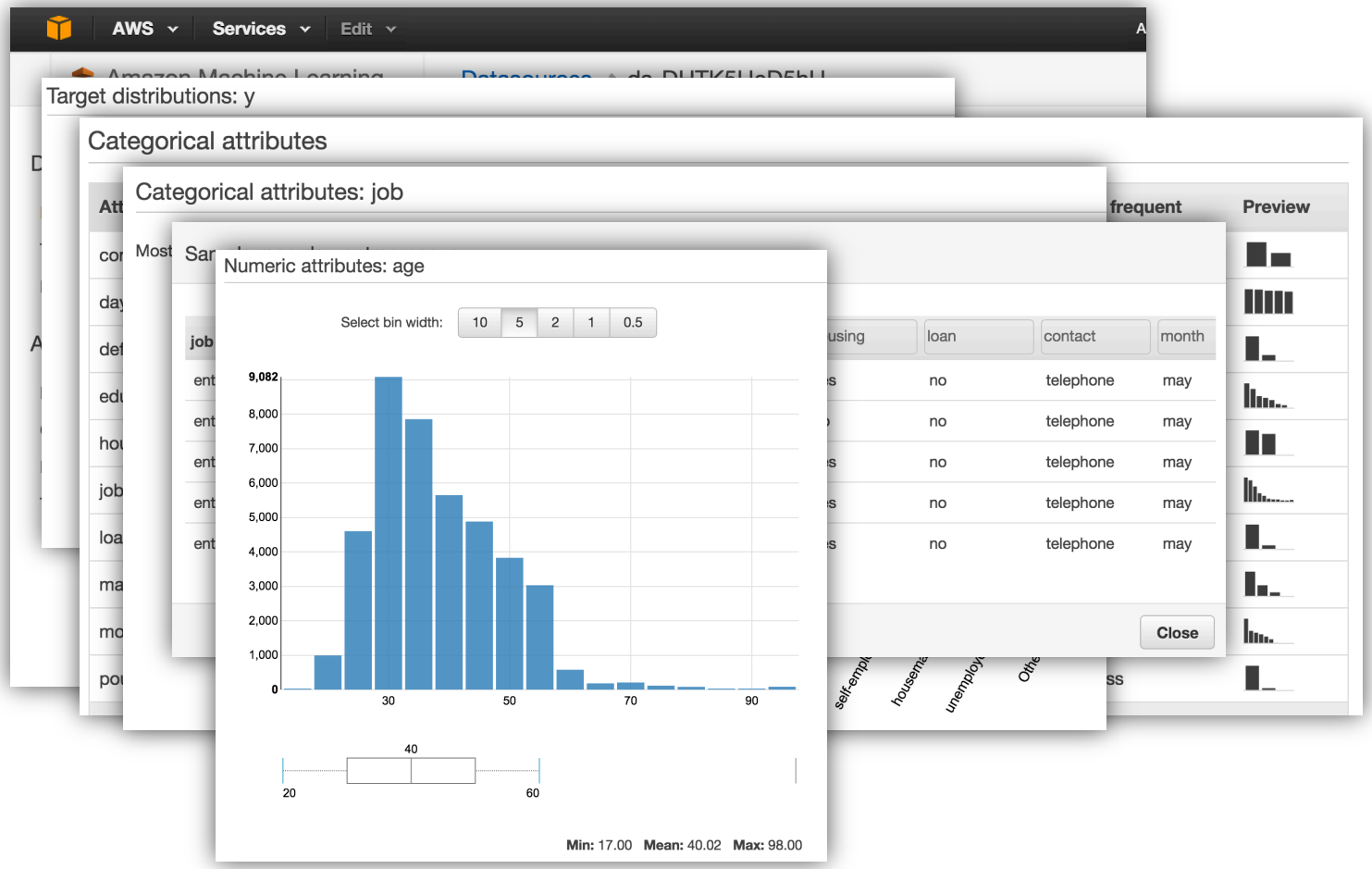
amazon web services

Create a Datasource object

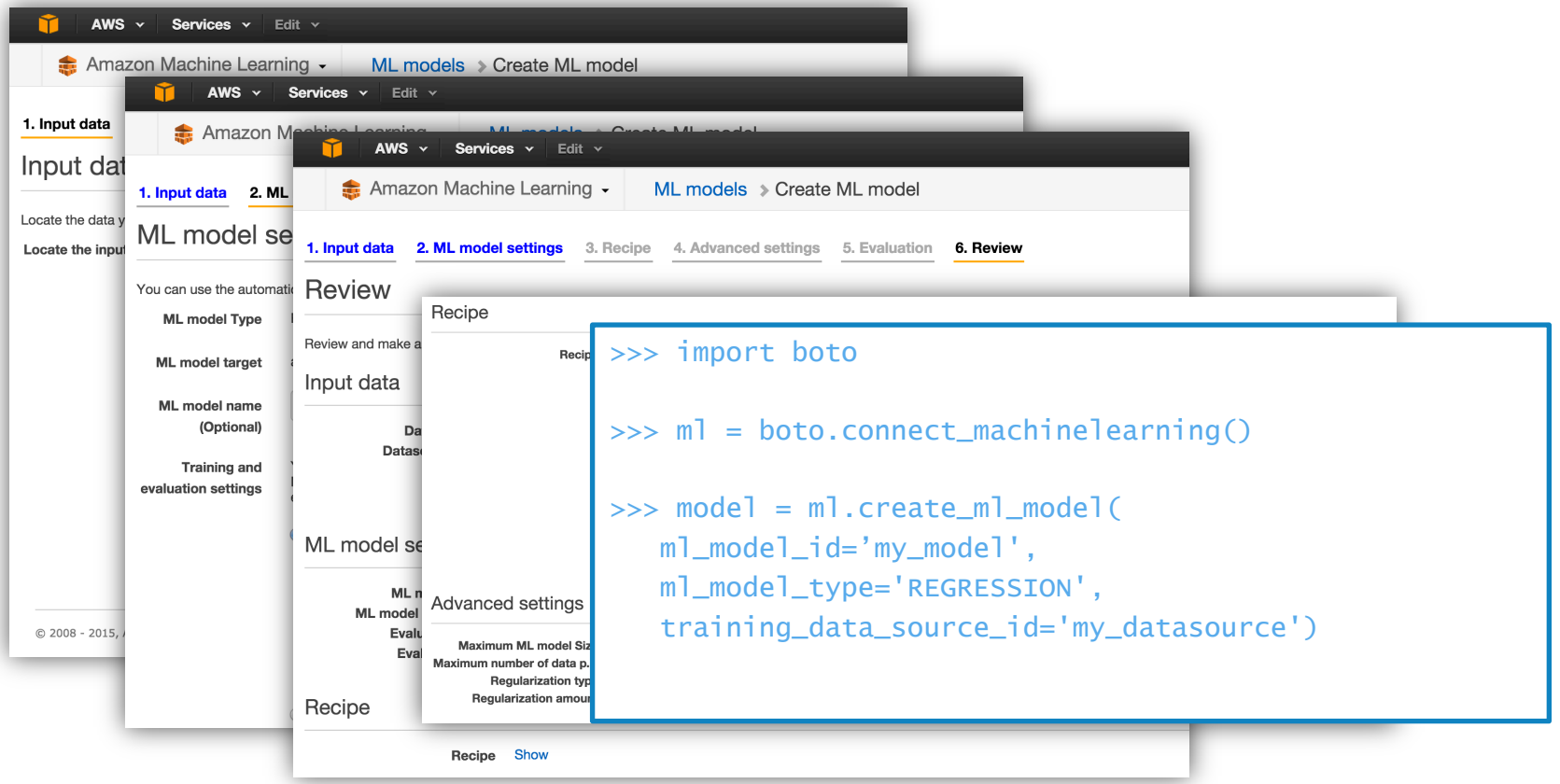


amazon web services

Explore and understand your data

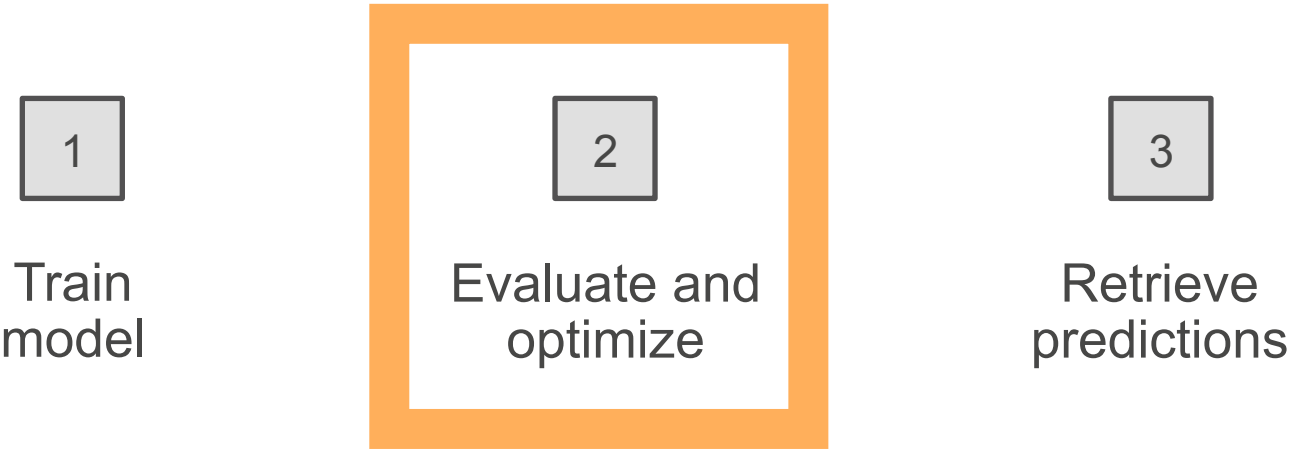


Train your model



amazon web services

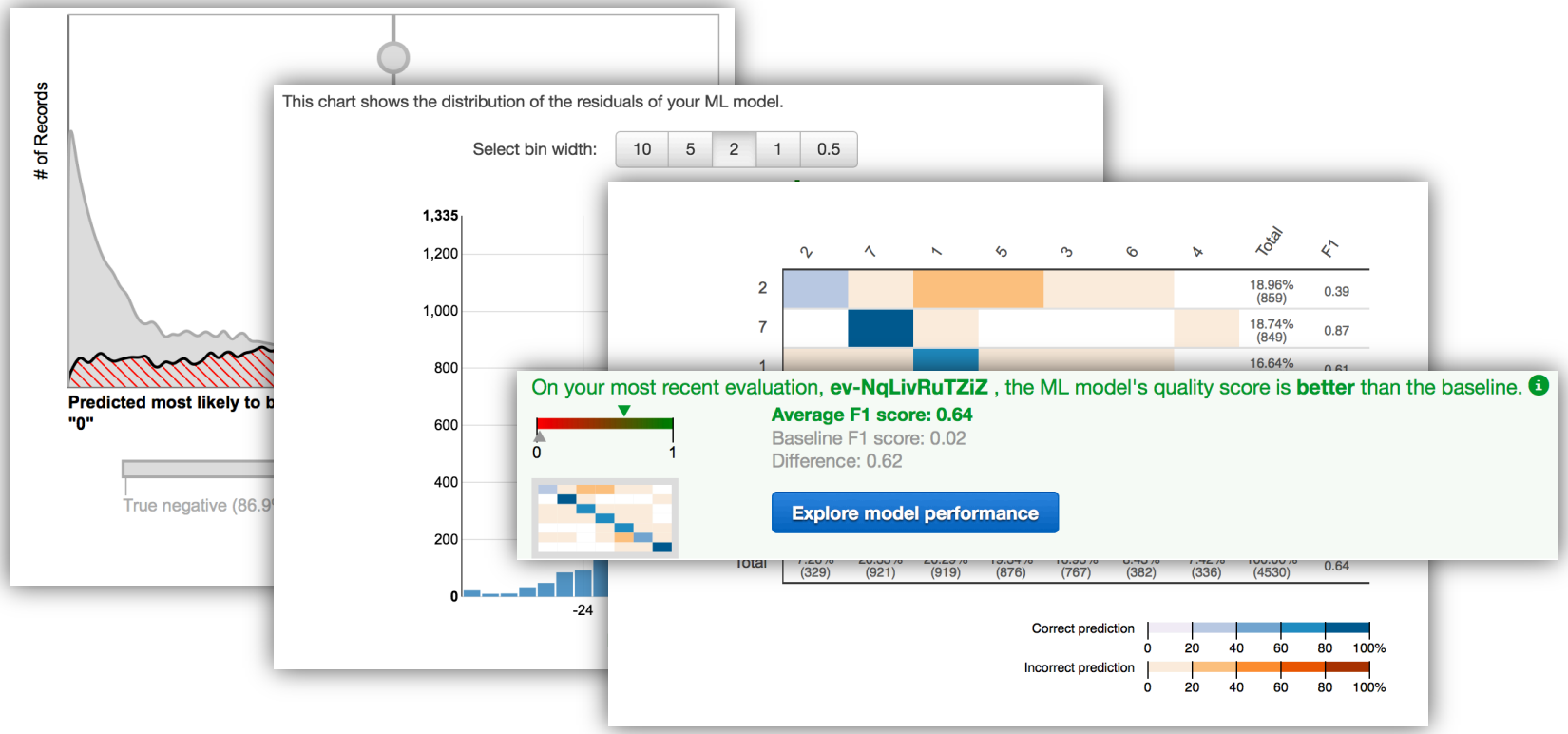
Building smart applications with Amazon ML



- Understand model quality
- Adjust model interpretation

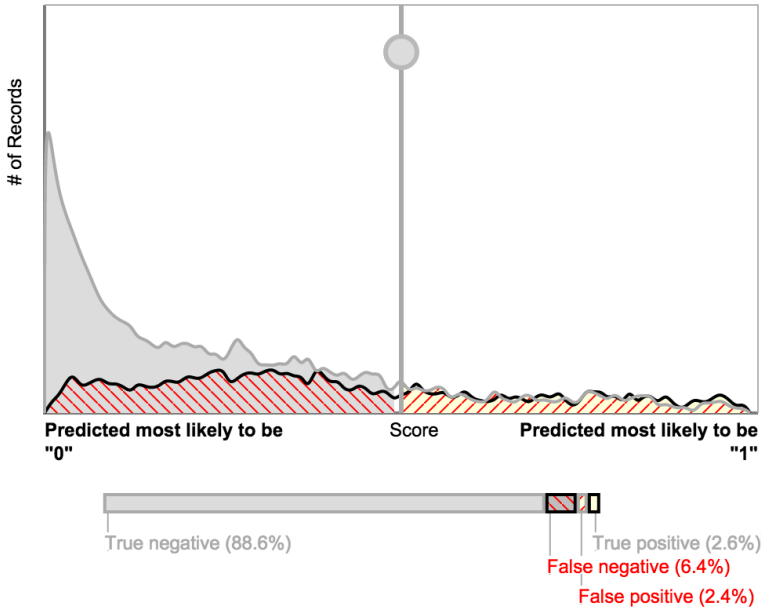
amazon web services

Explore model quality



amazon web services

Fine-tune model interpretation



Trade-off based on score threshold 0.5

Reset score threshold (0.5)

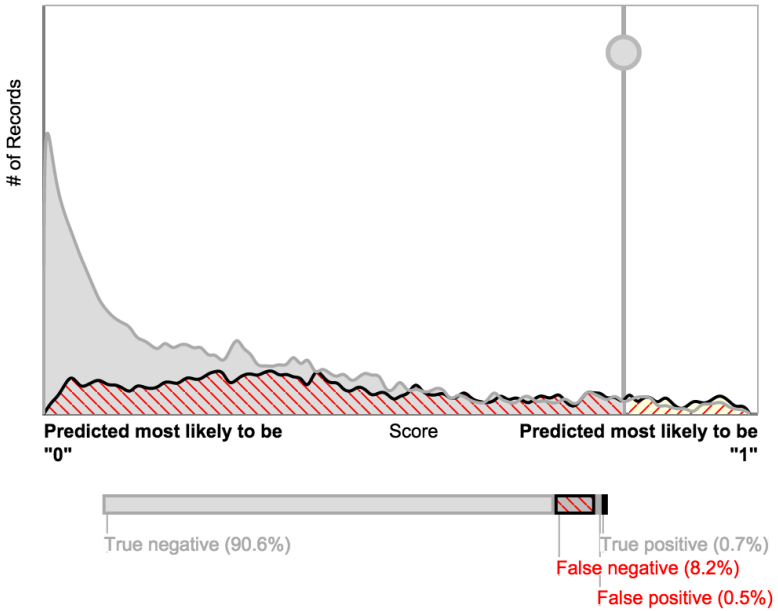
- **91% are correct**
  607 true positive
  20,933 true negative
- **9% are errors**
  567 false positive
  1,507 false negative

- 5% of the records are predicted as "1"
- 95% of the records are predicted as "0"

Save score threshold at 0.50

› Advance metrics

Fine-tune model interpretation



Trade-off based on score threshold 0.81

Reset score threshold (0.5)

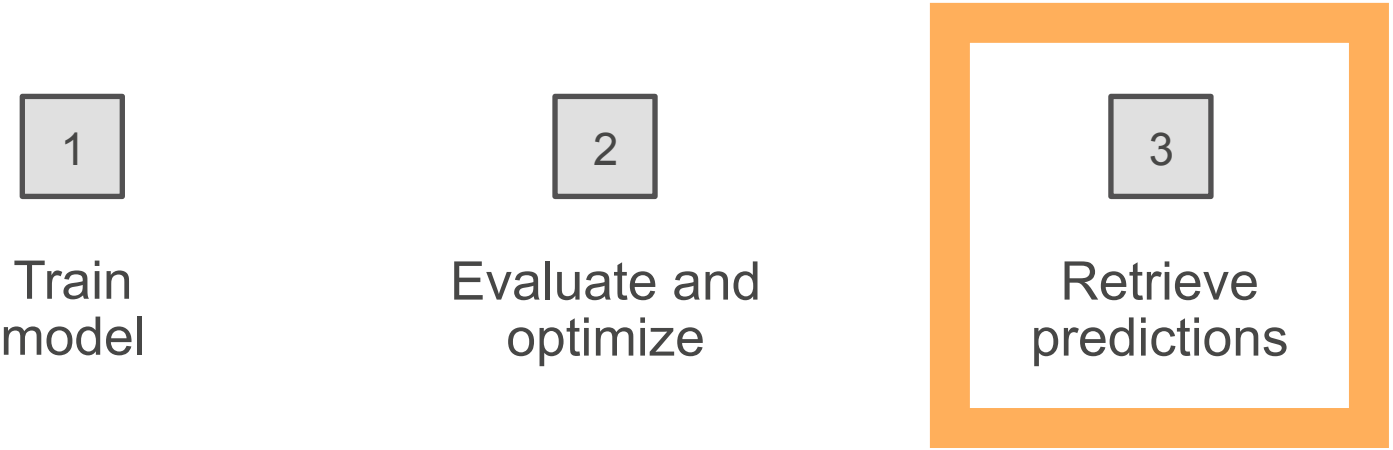
- **91% are correct**
  166 true positive
  21,385 true negative
- **9% are errors**
  115 false positive
  1,948 false negative

- 1% of the records are predicted as "1"
- 99% of the records are predicted as "0"

Save score threshold at 0.81

› Advance metrics

Building smart applications with Amazon ML



- Batch predictions
- Real-time predictions

amazon web services

Visualize (demo)

SHOUTOUT

- Leveraging Splunk to manage your AWS environment: Tuesday 22, 4:15-5PM
- Hunk and EMR: Big Data Analytics on AWS: Tuesday 22, 5:15-6PM
- Deploying Splunk on Amazon Web Services: Wednesday 23, 5:15-6PM
- **Splunking AWS for End-to-end Visibility: Thursday 24, 10-10:45AM**

amazon web services

SPECIAL THANKS

ML team: Manish Sainani, Fred Zhang, Shang Cai
AWS/Splunk coordination: Tony Bolander, Dritan Bitincka

amazon web services

Thank You!

amazon web services