Scalable Infrastructure for Malware Labeling and Analysis

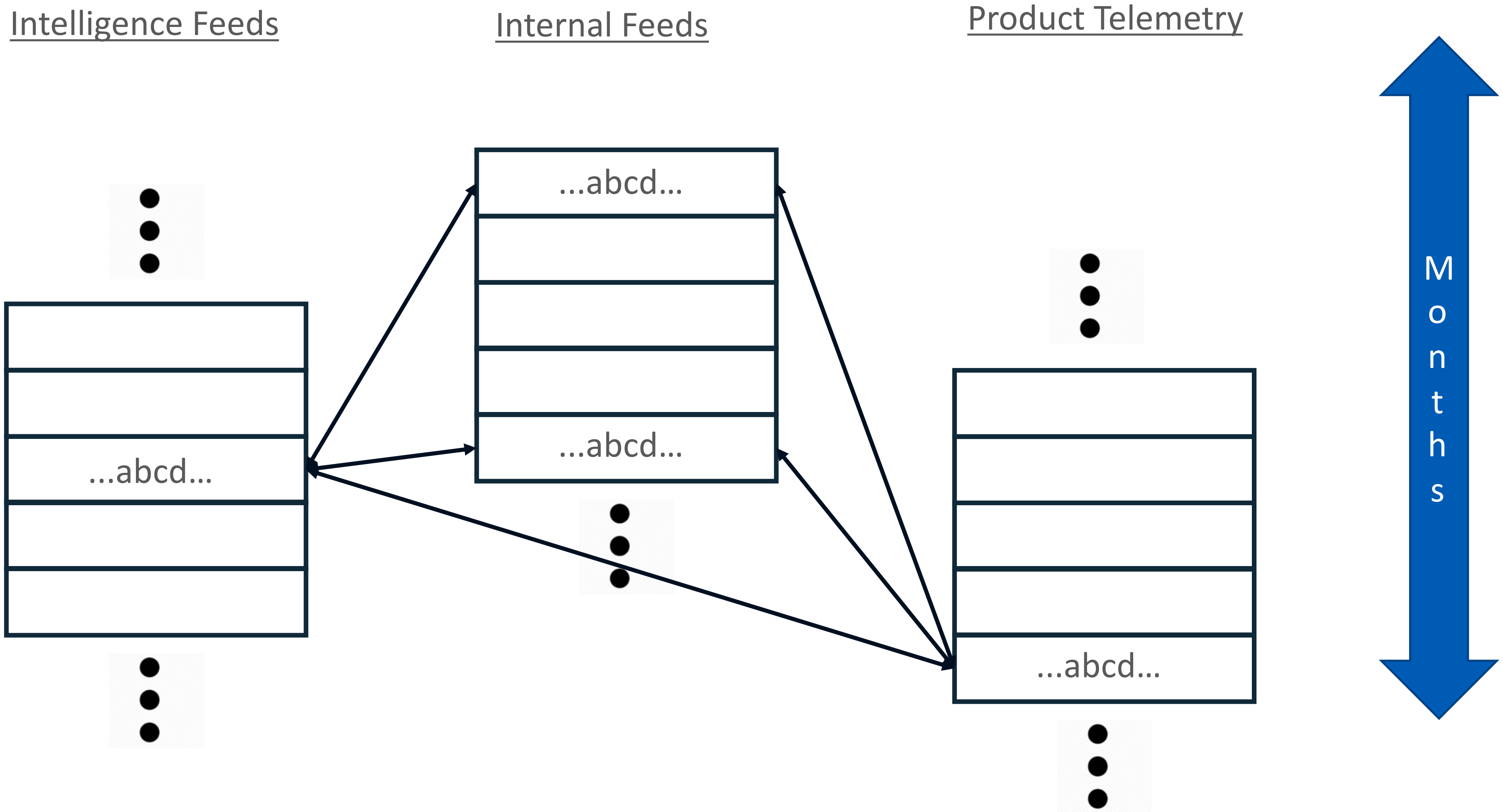
Konstantin Berlin

Director, Sophos AI

October 8, 2019

SOPHOS

Problem in a nutshell



Complexities

- Numerious Feeds
  - Multiple products
  - External intelligence feeds
  - Analyst feedback
- Data Size
  - Raw data is huge
  - Billions of events per day
  - Information distributed across multiple feeds over months
- Labeling
  - **Labels change constantly**
  - Complex logic
  - Constantly refined
- Validating/Monitoring
  - New files must be constantly scored
  - New model release requires rescoring of all files quickly
  - Need to roll back state to time of each event
- GDPR
  - Raw data distributed across multiple regions

Key AWS Technologies

SQS

- Fully managed message queue
- Autoscaling
- 14 day retention
- Multiple retries with delay
- Recovery from incomplete operation

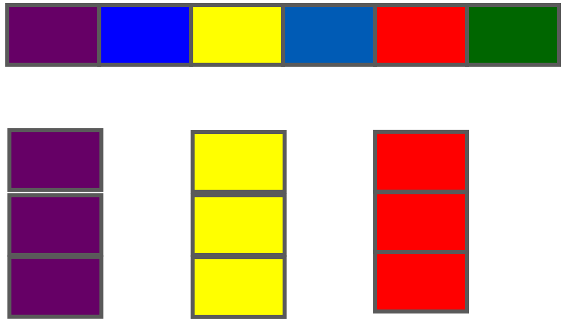
S3

- Cheap blob storage
- Automated cold storage
- Sends changes to SQS

Spot Autoscaling Cluster

- Cheaper than Lambda
- Easy to initialize complex environment, including GPU inference
- Scaling based on SQS and CloudWatch properties

Redshift

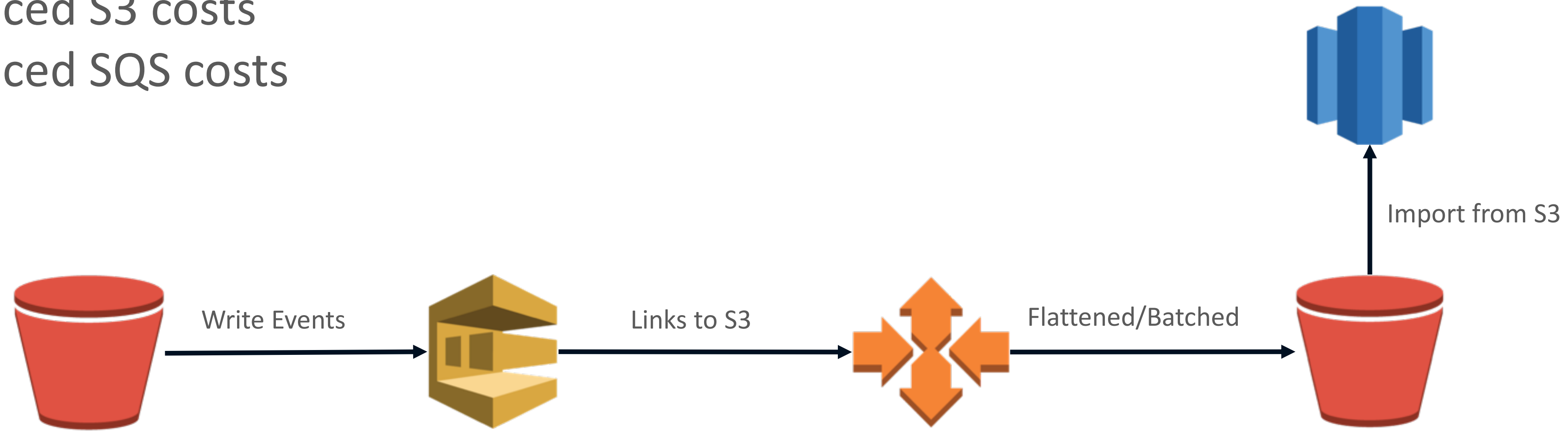


- Column oriented distributed DB
- Large write capacity
- Very high compression level (cheap storage)
- Ok to have wide tables

Basic Paradigms

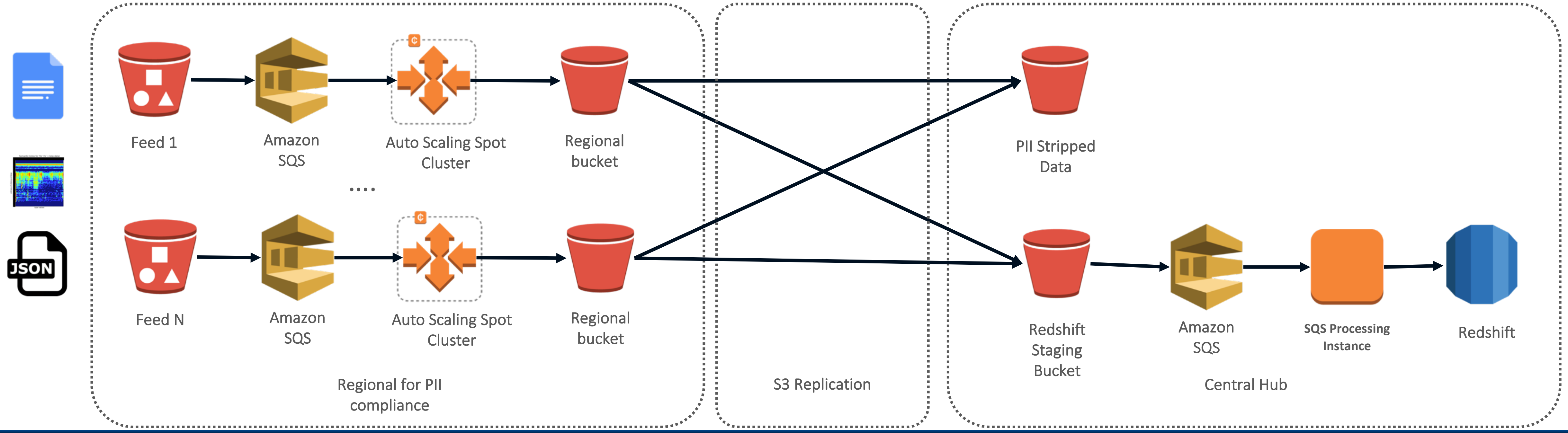
- "Data lake first"
  - **No data goes into a database only**
  - Easy replay if something goes wrong
  - Easy to change databases
  - Easy data sharing across groups
- Aggressive Batching
  - Minimizes number of events
  - Reduced S3 costs
  - Reduced SQS costs

- Fully Managed, When Possible
  - Let engineers work on more important problems
  - Keeps up with latest and greatest

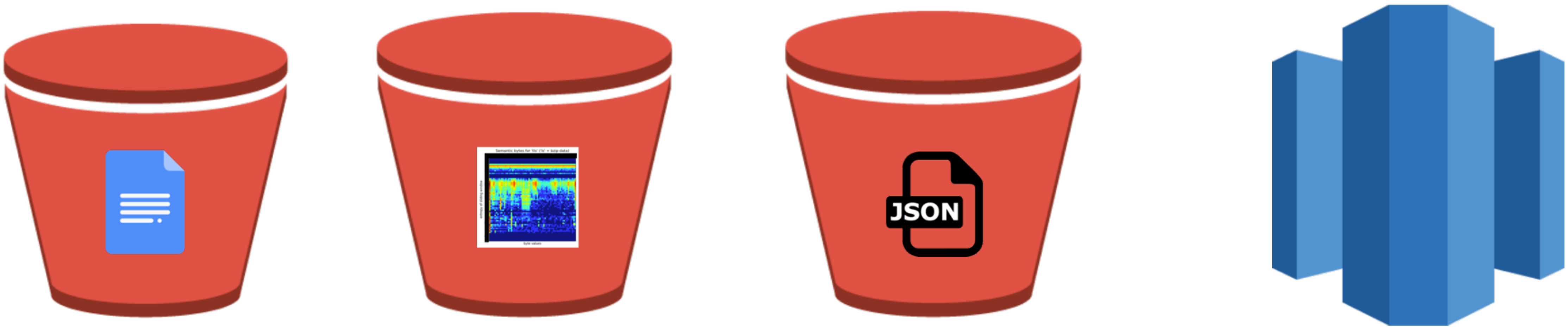


Data Ingestion (Telemetry, VT, Model Scores)

- Minimize Cost
  - Spot instances, batching, S3 replication across GDPR regions
- Minimize Maintenance
  - Managed services, minimum components, automatic recovery via SQS and S3
- Resilience and Scaling
  - Autoscaling in all components, supports 10-100x data bursts for backpopulation



How Does Storage Look?



~1 PB
Intelligent Tiering

Artifacts

~1 TB

Feature Vectors

~100 TB

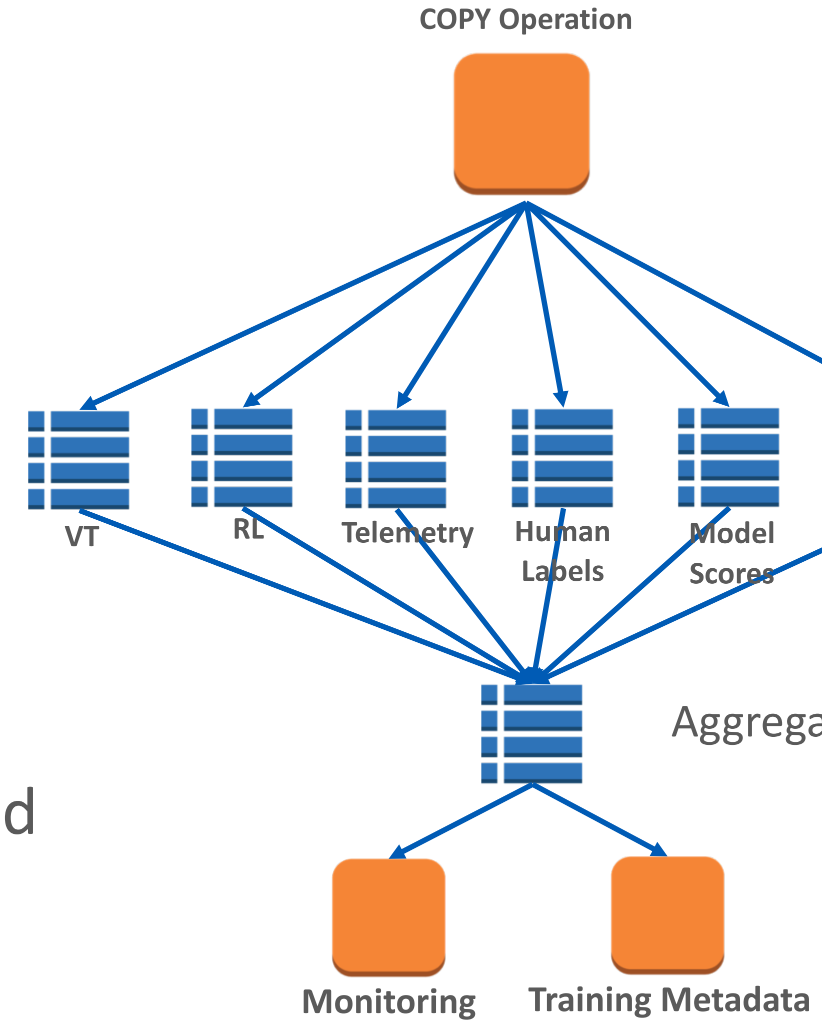
Raw Metadata

~25 TB

Indexed Metadata

Metadata Aggregation and Correlation

- Columnar distributed storage
  - Wide tables
    - Keep as much data as you can
  - Most queries need few columns
    - Ex. Label, Prediction
- Timestamp everything!
- Daily joins between all sources
  - Keeps only first seen and last seen
  - sha256 as distribution key
  - sha256 + timestamp as sort key
- Constant vacuuming in background
- Weekly cleaning of duplicates and older data



Column oriented distributed database

Historic Data

Aggregated Information (overnight)

Usage

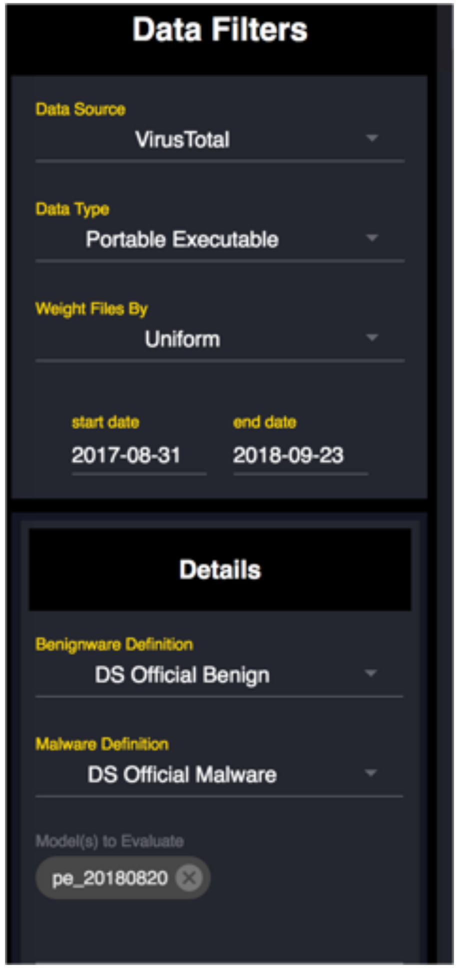
Redshift Use Cases

SOPHOS

Improve ML Training

- Labeling
  - Join across multiple source to form labels
  - Instantly relabel all artifacts
- Training metadata
  - Redshift unload to S3
  - Complex queries define arbitrary training labels
  - Export of 100M+ rows takes minutes
  - SQL define training and validation data for all models
- Fill gaps using smart queries
  - Implement active learning strategies
  - Find missing data and fill it

Dashboard Monitoring (Performance and Issues)



Data Filters

Data Source
VirusTotal

Data Type
Portable Executable

Weight Files By
Uniform

start date
2017-08-31

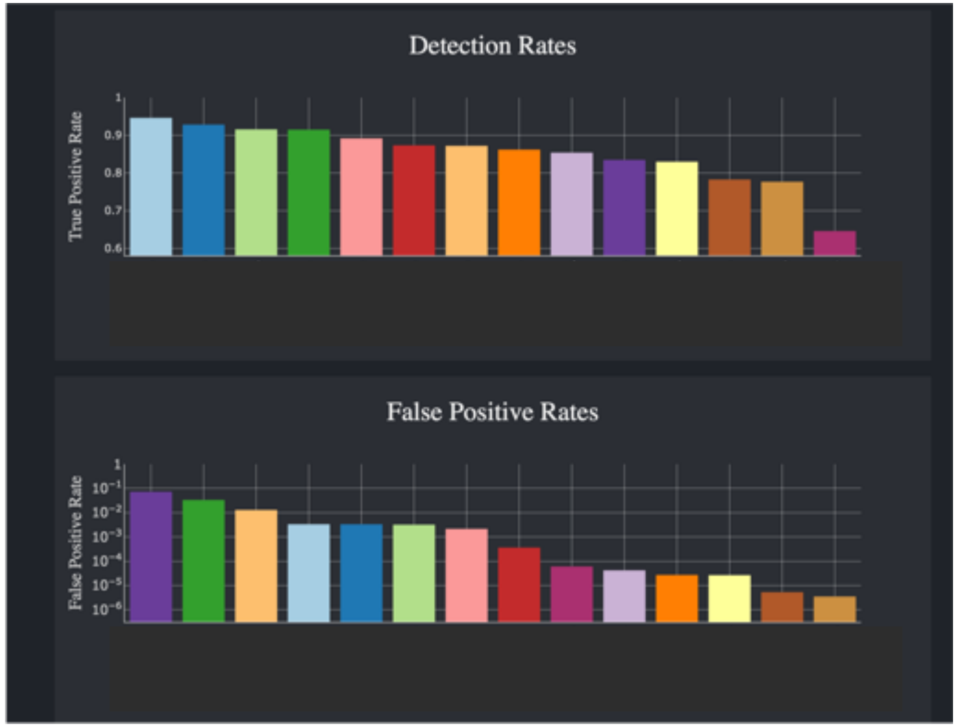
end date
2018-09-23

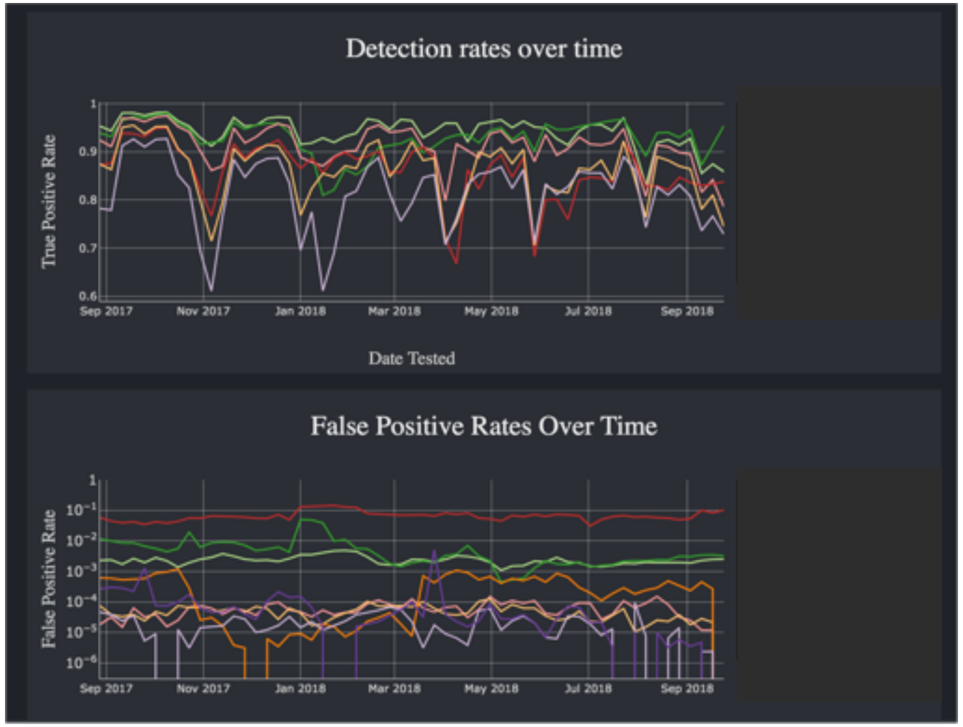
Details

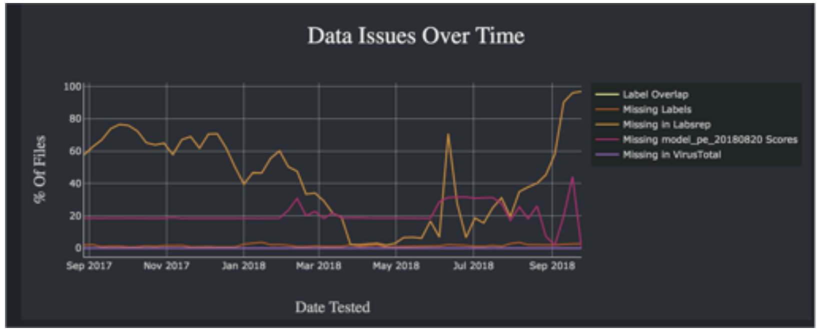
Benignware Definition
DS Official Benign

Malware Definition
DS Official Malware

Model(s) to Evaluate
pe_20180820







Questions?

SOPHOS