SECURITY ANALYTICS PLATFORM

January 2020

THE TEAM

Gary Gabriel
Principal Security Developer

Mason Cheng
Principal Data Scientist

COMCAST

AGENDA

- Enterprise Data Lake Overview
- Architecture Design
- Data Processing and Job Orchestration
- Analytics Platform

COMCAST

ENTERPRISE DATA LAKE OVERVIEW

As new threats and attacks emerge, and the volume of data grows, so does the complexity of data. Solutions make use of various security tools while supplementing and extending its capabilities with a scalable and intelligence driven Data Lake.

DATA LAKE KEY COMPONENTS:

Apache Spark

Open source distributed framework for large-scale processing

Databricks Delta Lake

Open source storage layer for Spark that provides optimizations and ACID transactions

Apache Airflow

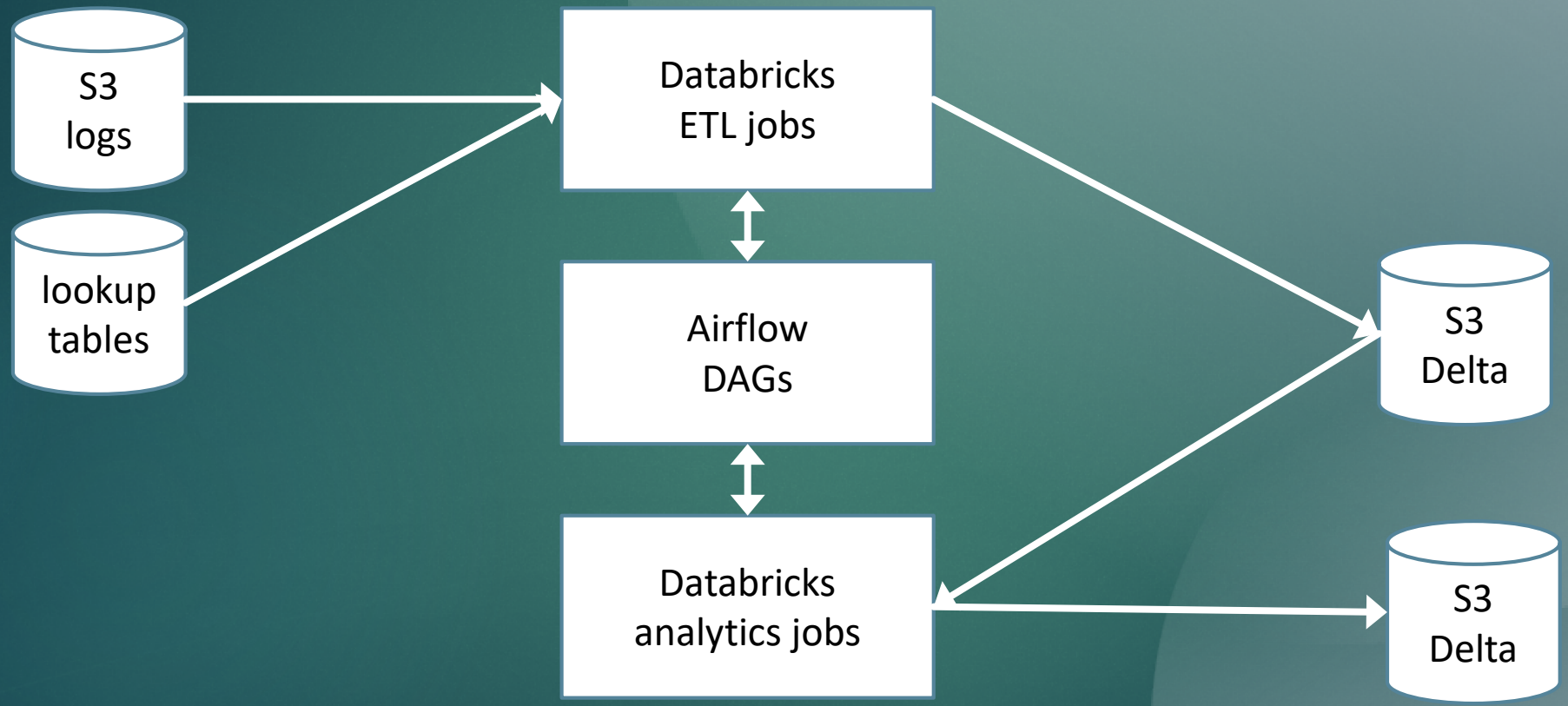
Open source workflow scheduler and monitoring solution

Amazon VPC, EC2, S3

Cloud compute platform and storage

COMCAST

ARCHITECTURE DESIGN



COMCAST

ENTERPRISE DATA LAKE ARCHITECTURE

PYTHON APACHE SPARK BATCH JOBS

- Hourly or daily
- Command line flags to control input/output

DATABRICKS DELTA AND CLUSTER ISOLATION

- Schema checking, transactions and optimizations
- One job per cluster

PARTITIONING

- Raw data is partitioned by hour (e.g. /y=2019/m=10/d=19/h=04/)
- Output tables partitioned by hour or day
- Partition overwrite by each script

AD-HOC CLUSTERS

- Interactive data exploration in notebooks
- Script Development

COMCAST

DATA PROCESSING (EXTRACT, TRANSFORM, LOAD)

LOG PARSING

- Terabytes of logs every day
- Many tools, many formats
- Regular expressions for parsing

ENRICHMENT

- Adding network zones
- Creating sessions from DHCP (Dynamic Host Configuration Protocol) and VPN (Virtual Private Network) logs
- Domain lookups – WHOIS, blacklists

LOOKUP DATA SOURCES

- Asset databases
- Active Directory

COMCAST

JOB ORCHESTRATION

CI – CONTINUOUS INTEGRATION

- Code reviews before merging
- Production git repository synchronized to S3

APACHE AIRFLOW

- Runs on EC2
- Uses DatabricksSubmitRunOperator

DAGS – DIRECTED ACYCLIC GRAPHS

- Define data dependencies
- Ensure scripts run in the correct order

BACKFILL

- Sometimes jobs fail
- Finds gaps in the data and reruns appropriate jobs

COMCAST

PERSONAS – PRIMARY USERS OF SECURITY ANALYTICS PLATFORM

THREAT HUNTERS

- Searching for unknown threats
- Use the security analytics platform to supplement the SIEM

INCIDENT RESPONDERS

- Handle escalated threats
- Forensic analysis and evidence

DATA SCIENCE MODELERS

- Data exploration and analysis
- Use historical data to model patterns of suspicious behaviors to identify outliers/potential risk

MANAGEMENT AND STAKEHOLDERS

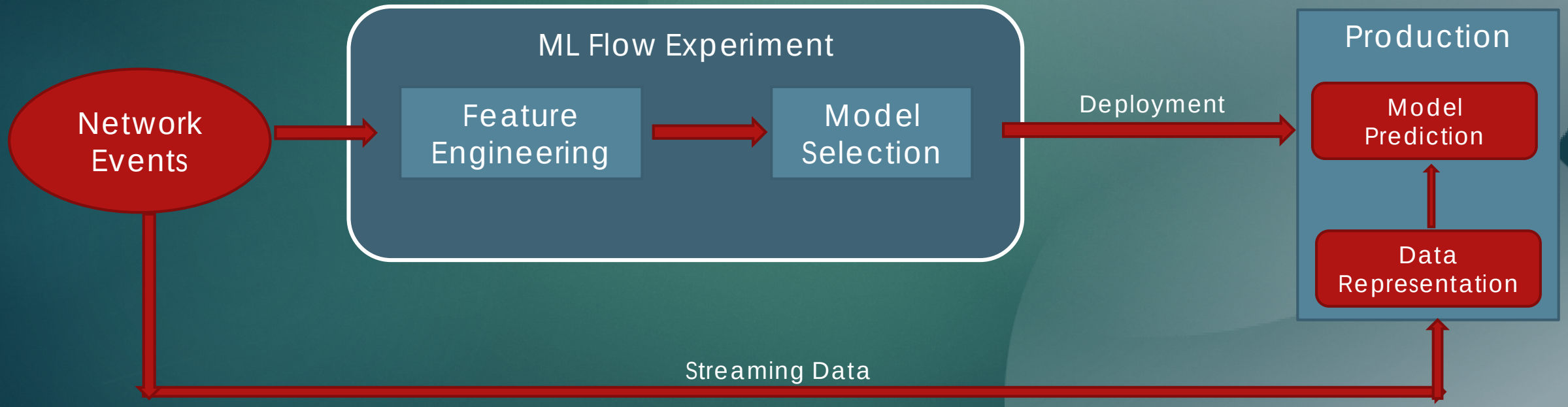
- Summarized KPI metrics
- Dashboard with self-service/drill down capabilities

COMCAST

USE OF MLFLOW FOR ML MODEL DEVELOPMENT

Data Science Modelers use MLflow to manage end-to-end ML model development.

- Run experiments with tracking for parameters, code versions, metrics and output
- Package and share ML code with others
- Manage and deploy models in downstream tools & production



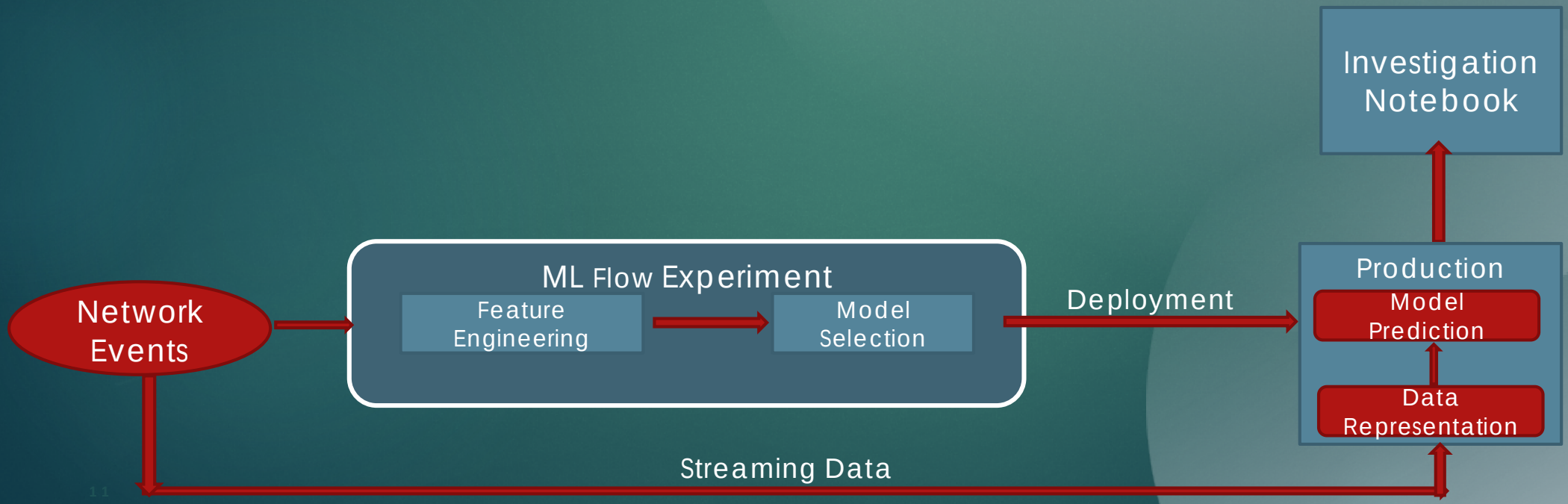
COMCAST

WORKFLOW-BASED INVESTIGATION NOTEBOOK

Create notebooks for workflows specific to alerts, threats, or use cases to enable efficient investigation.

Examples:

- Suspicious/C2 communication over HTTP
- DNS tunneling/exfiltration
- Privileged User Behavioral anomaly

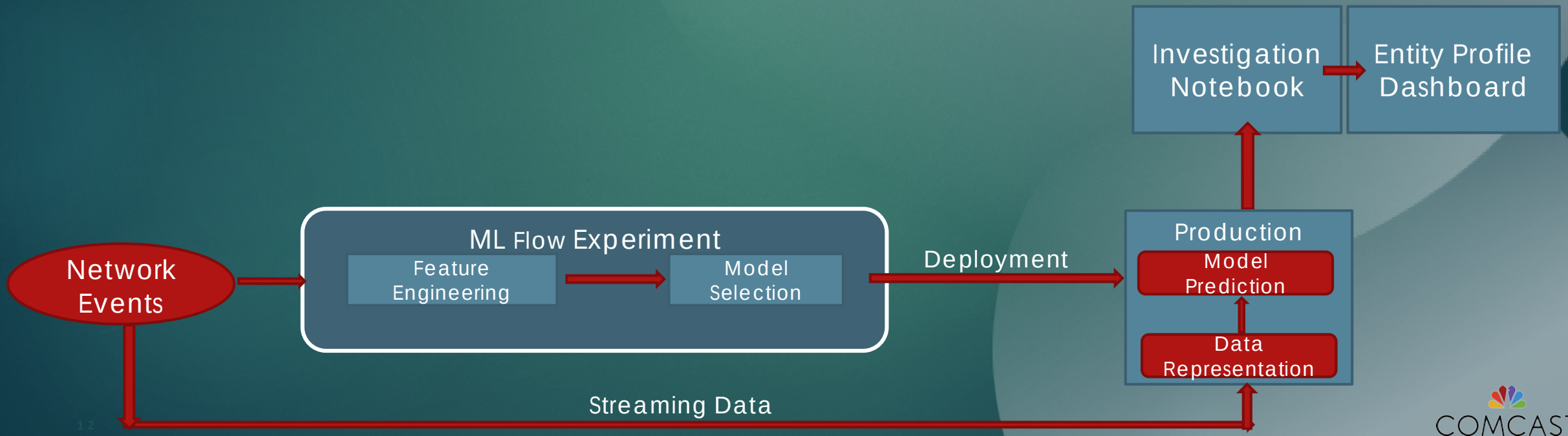


COMCAST

ENTITY PROFILE DASHBOARD

Enable Threat Hunters to look up profiles and perform drill downs to various events for a given entity.

- Users
- Devices (Endpoints/Assets)
- Threat Intelligence
- External Domains



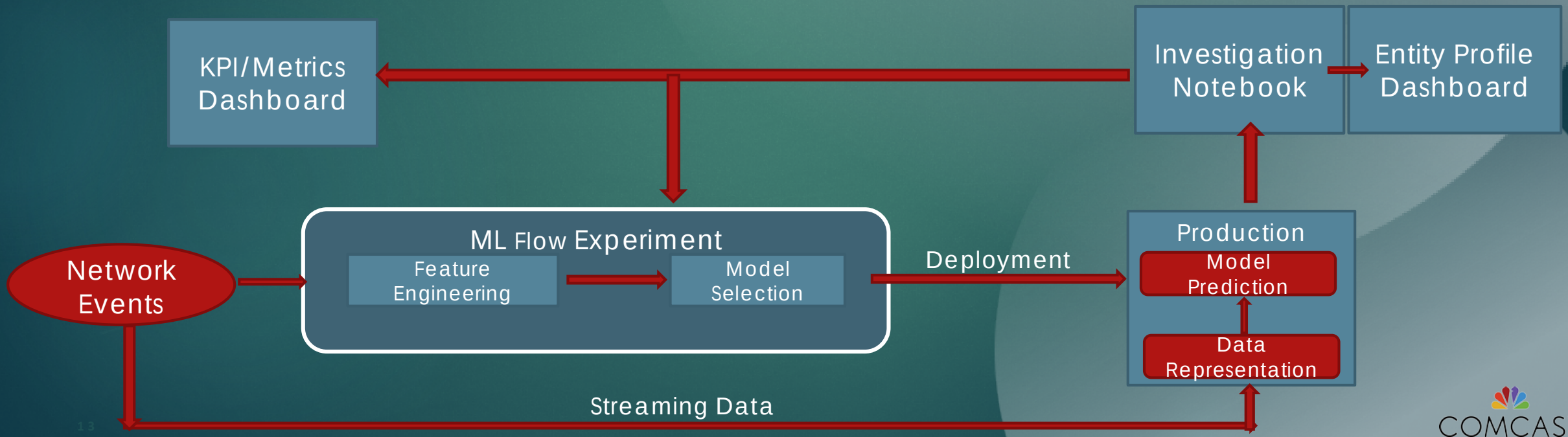
COMCAST

KPI/METRICS DASHBOARD

A single pane of glass for executives/managers and stakeholders to track and evaluate threat analytical model output & results in Tableau Dashboard.

Example KPIs:

- # of alerts generated by models
- # of users reported and business units they are associated with
- # of total traffic & users being scanned per day



COMCAST

SECURITY ANALYTICS PLATFORM

January 2020