splunk> .conf2017

Building ML Solutions using MLTK

Repeatable ML Workflows

Andrew Stein | Analytical Architect
Iman Makaremi | Senior Data Scientist

September 2017 | Washington, DC

Forward-Looking Statements

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 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.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners.

Agenda

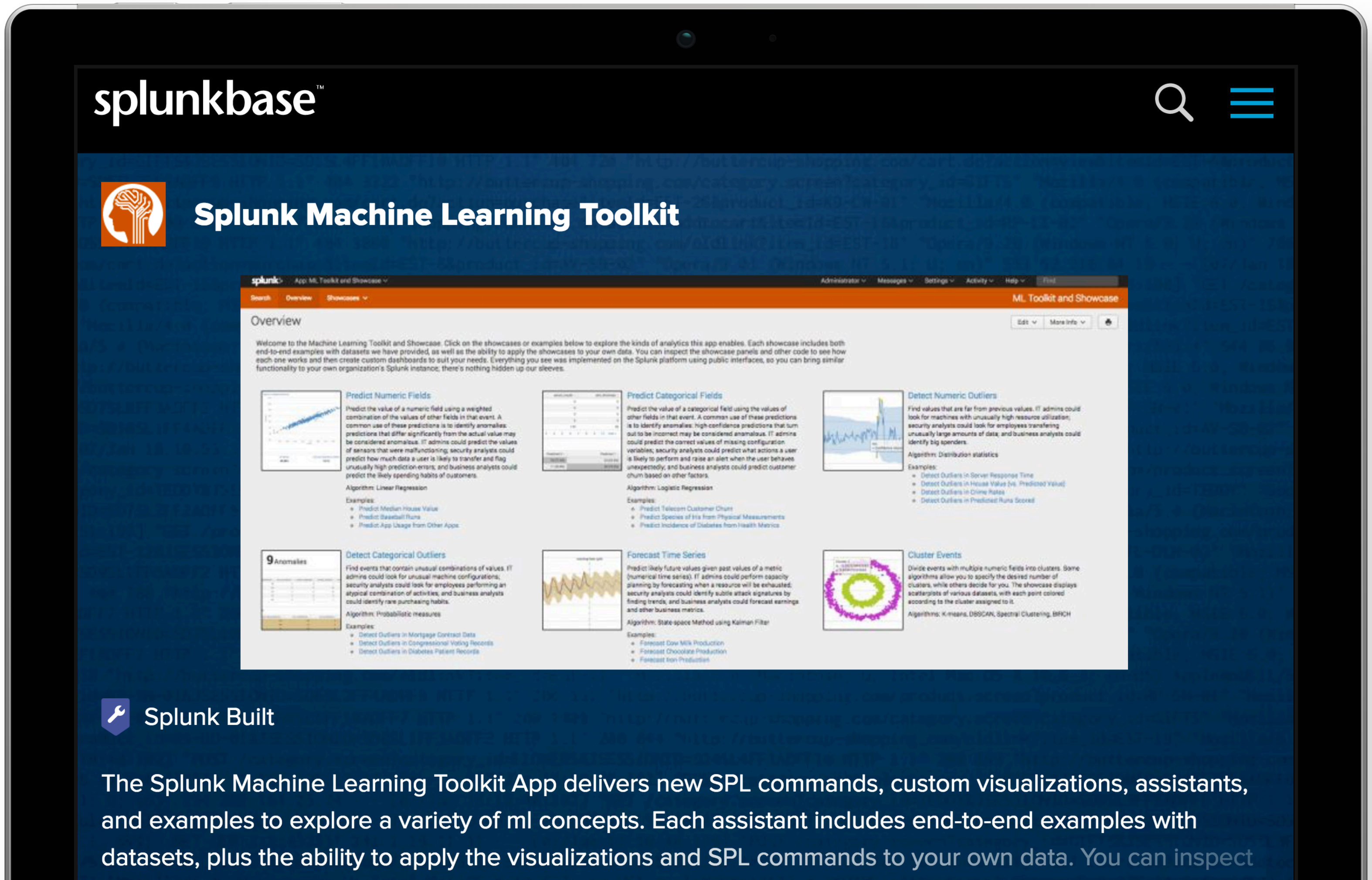
- MLTK? What?!
- Assistant : Numeric Outlier Detection
- Assistant : Predict Numeric Fields
- Assistant : Predict Categorical Fields

What?

MLTK

What?

Splunk has a Machine Learning Toolkit App!



Splunk Machine Learning Toolkit

Overview

Welcome to the Machine Learning Toolkit and Showcase. Click on the showcases or examples below to explore the kinds of analytics this app enables. Each showcase includes both end-to-end examples with datasets we have provided, as well as the ability to apply the showcases to your own data. You can inspect the showcase panels and other code to see how each one works and then create custom dashboards to suit your needs. Everything you see was implemented on the Splunk platform using public interfaces, so you can bring similar functionality to your own organization's Splunk instance; there's nothing hidden up our sleeves.

Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous. IT admins could predict the values of sensors that were malfunctioning; security analysts could predict how much data a user is likely to transfer and flag unusually high prediction errors; and business analysts could predict the likely spending habits of customers.

Algorithm: Linear Regression

Examples:
- Predict Median House Value
- Predict Baseball Runs
- Predict App Usage from Other Apps

Predict Categorical Fields

Predict the value of a categorical field using the values of other fields in that event. A common use of these predictions is to identify anomalies: high-confidence predictions that turn out to be incorrect may be considered anomalous. IT admins could predict the correct values of missing configuration variables; security analysts could predict what actions a user is likely to perform and raise an alert when the user behaves unexpectedly; and business analysts could predict customer churn based on other factors.

Algorithm: Logistic Regression

Examples:
- Predict Telecom Customer Churn
- Predict Species of Iris from Physical Measurements
- Predict Incidence of Diabetes from Health Metrics

Detect Numeric Outliers

Find values that are far from previous values. IT admins could look for machines with unusually high resource utilization; security analysts could look for employees transfering unusually large amounts of data; and business analysts could identify big spenders.

Algorithm: Distribution statistics

Examples:
- Detect Outliers in Server Response Time
- Detect Outliers in House Value (vs. Predicted Value)
- Detect Outliers in Crime Rates
- Detect Outliers in Predicted Runs Scored

Detect Categorical Outliers

Find events that contain unusual combinations of values. IT admins could look for unusual machine configurations; security analysts could look for employees performing an atypical combination of activities; and business analysts could identify rare purchasing habits.

Algorithm: Probabilistic measures

Examples:
- Detect Outliers in Mortgage Contract Data
- Detect Outliers in Congressional Voting Records
- Detect Outliers in Diabetes Patient Records

Forecast Time Series

Predict likely future values given past values of a metric (numerical time series). IT admins could perform capacity planning by forecasting when a resource will be exhausted; security analysts could identify subtle attack signatures by finding trends; and business analysts could forecast earnings and other business metrics.

Algorithm: State-space Method using Kalman Filter

Examples:
- Forecast Cow Milk Production
- Forecast Chocolate Production
- Forecast Iron Production

Cluster Events

Divide events with multiple numeric fields into clusters. Some algorithms allow you to specify the desired number of clusters, while others decide for you. The showcase displays scatterplots of various datasets, with each point colored according to the cluster assigned to it.

Algorithms: K-means, DBSCAN, Spectral Clustering, BIRCH

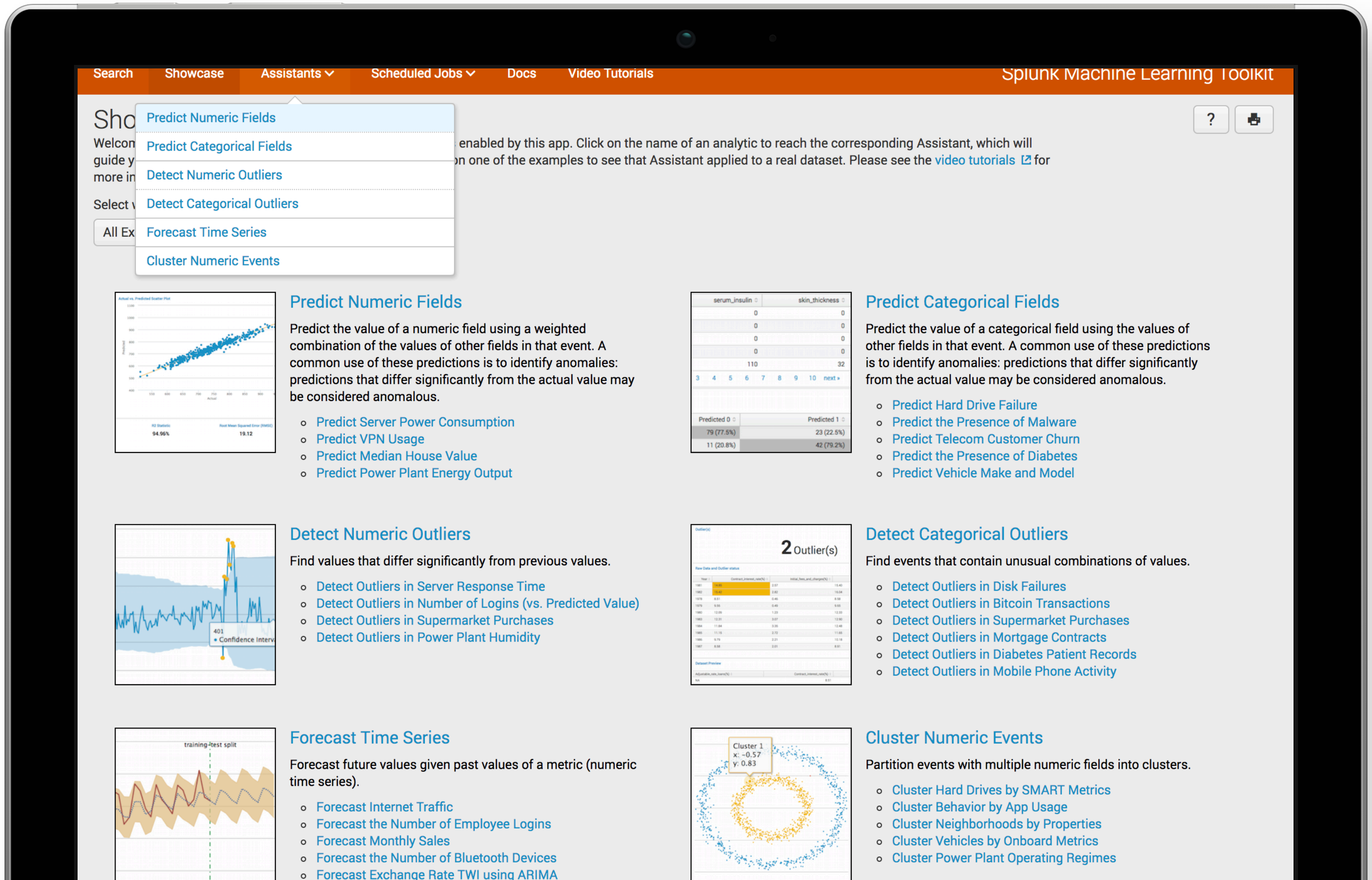
Splunk Built

The Splunk Machine Learning Toolkit App delivers new SPL commands, custom visualizations, assistants, and examples to explore a variety of ml concepts. Each assistant includes end-to-end examples with datasets, plus the ability to apply the visualizations and SPL commands to your own data. You can inspect

- What is Splunkbase
- What is the App
- Where can I go to learn more

What?

The Machine Learning Toolkit has Assistants!



Sho

Welcon … enabled by this app. Click on the name of an analytic to reach the corresponding Assistant, which will guide y… on one of the examples to see that Assistant applied to a real dataset. Please see the video tutorials for more in…

Select

Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

- Predict Server Power Consumption
- Predict VPN Usage
- Predict Median House Value
- Predict Power Plant Energy Output

Predict Categorical Fields

Predict the value of a categorical field using the values of other fields in that event. A common use of these predictions is to identify anomalies: predictions that differ significantly from the actual value may be considered anomalous.

- Predict Hard Drive Failure
- Predict the Presence of Malware
- Predict Telecom Customer Churn
- Predict the Presence of Diabetes
- Predict Vehicle Make and Model

Detect Numeric Outliers

Find values that differ significantly from previous values.

- Detect Outliers in Server Response Time
- Detect Outliers in Number of Logins (vs. Predicted Value)
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Power Plant Humidity

Detect Categorical Outliers

Find events that contain unusual combinations of values.

- Detect Outliers in Disk Failures
- Detect Outliers in Bitcoin Transactions
- Detect Outliers in Supermarket Purchases
- Detect Outliers in Mortgage Contracts
- Detect Outliers in Diabetes Patient Records
- Detect Outliers in Mobile Phone Activity

Forecast Time Series

Forecast future values given past values of a metric (numeric time series).

- Forecast Internet Traffic
- Forecast the Number of Employee Logins
- Forecast Monthly Sales
- Forecast the Number of Bluetooth Devices
- Forecast Exchange Rate TWI using ARIMA

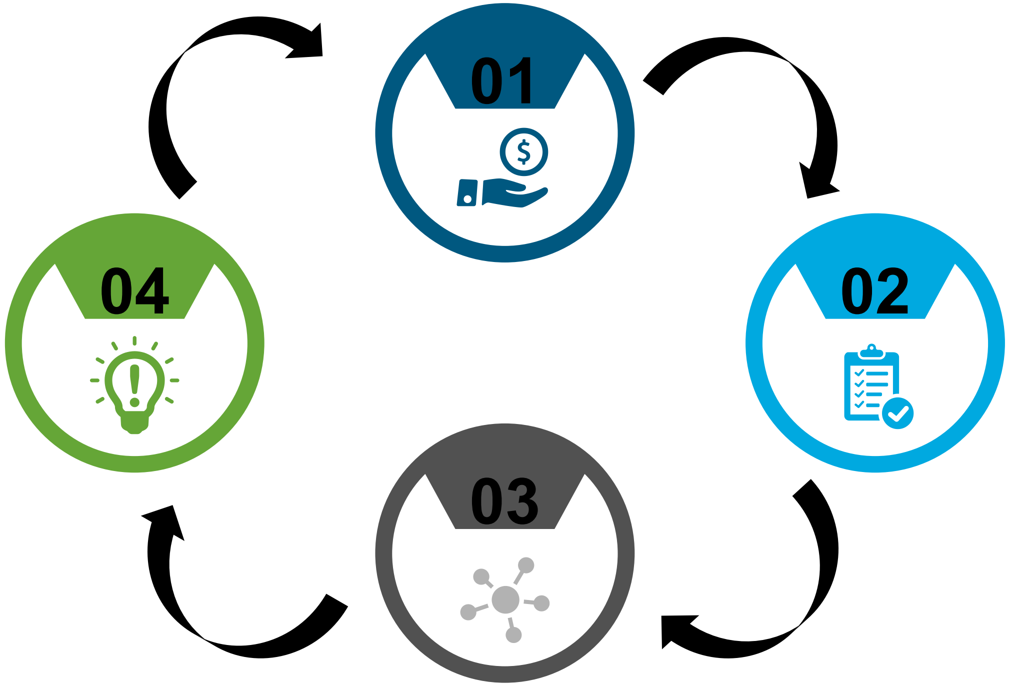
Cluster Numeric Events

Partition events with multiple numeric fields into clusters.

- Cluster Hard Drives by SMART Metrics
- Cluster Behavior by App Usage
- Cluster Neighborhoods by Properties
- Cluster Vehicles by Onboard Metrics
- Cluster Power Plant Operating Regimes

- What is an Assistant

A Successful Machine Learning Process



Data Driven Decisions

Acquire Data

Generate Insights

Publish Insights

Assistant: Numeric Outlier Detection

Assistant: Detect Numeric Outlier

One workflow, multiple use cases

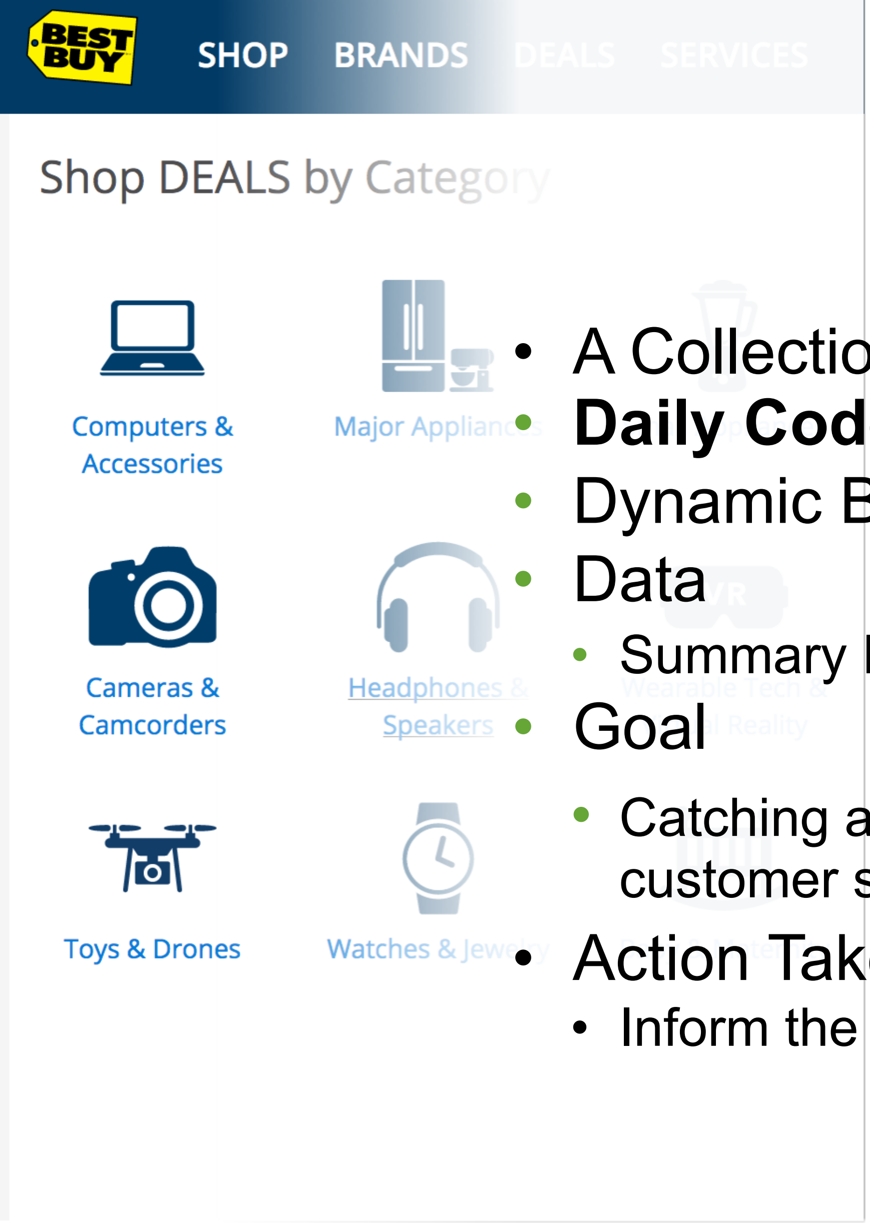
- Site Reliability Engineer

Reduced Time to Action by 50%

- Design Specialist

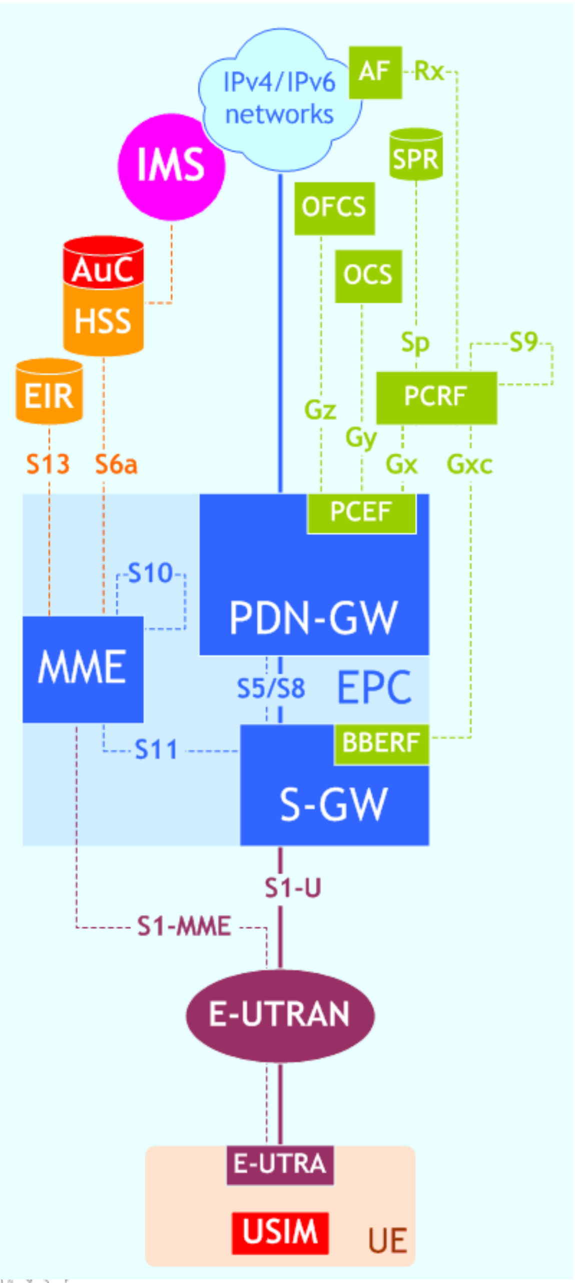
16 Million Dollars Saved

Website Accessibility



- A Collection of **Several Applications**
- **Daily Code Releases** for each Application
- Dynamic Behaviour of Internet Browsing
- Data
  - Summary Indexes
- Goal
  - Catching a problematic code release in a timely manner is important for customer satisfaction.
- Action Taken
  - Inform the release team about the problems

Mobility 3GPP Core KPI



Data

- The 3GPP Core receives transactions from each subscriber **to maintain connection**.
- The KPI captures **the behavior of TELUS's network.**

Goal

- Monitor this **dynamic** KPI and alert on **performance degradation**.

Action Taken

- Radio engineers informed about problems as soon as something occurred.

TELUS

the future is friendly®

splunk> .conf2017

Assistant : Detect Numeric Outliers

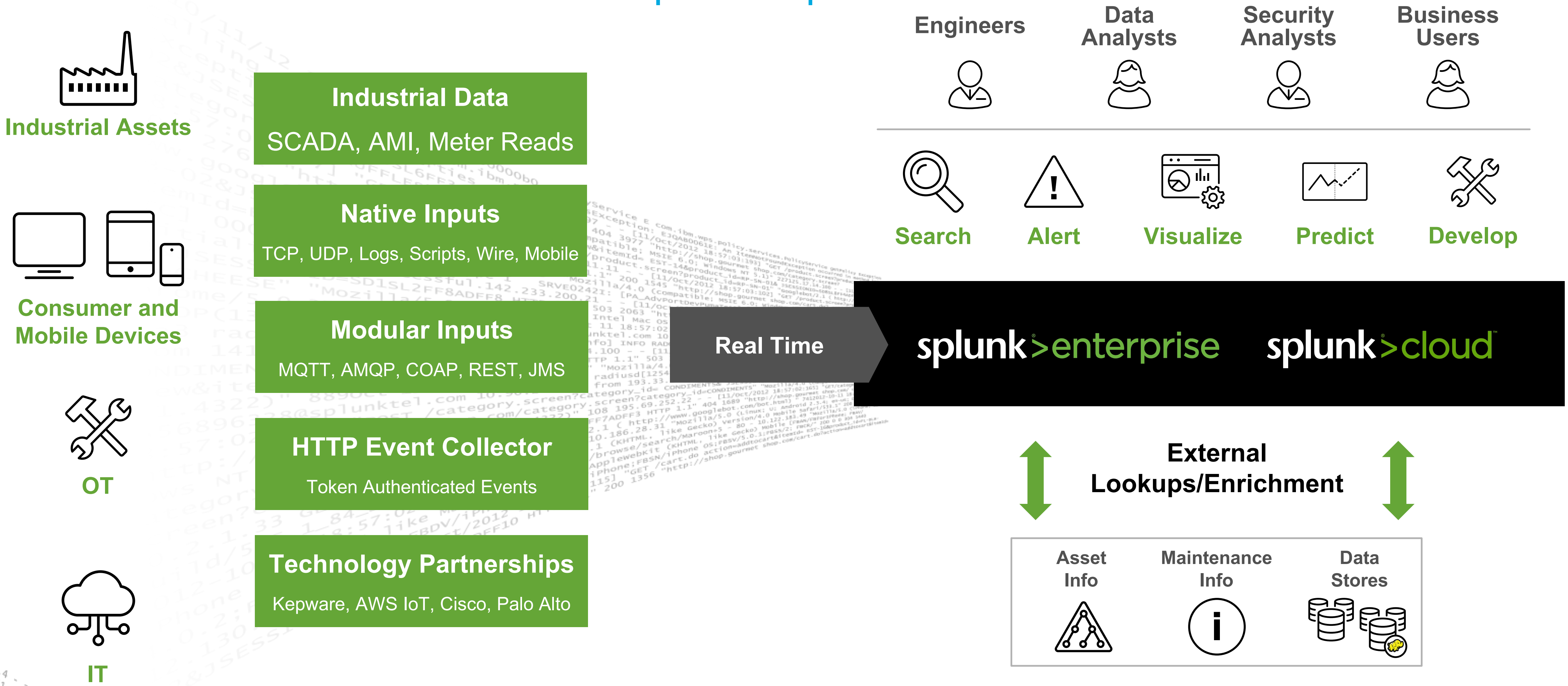
Step 1 : Data Driven Decisions

- Define your Problem!

  - I want **detect and alert** on metric(s) that deviate **significantly** from their **past behaviour**.

Assistant : Detect Numeric Outliers

Step 2 : Acquire Data



Industrial Assets

Consumer and Mobile Devices

OT

IT

Industrial Data
SCADA, AMI, Meter Reads

Native Inputs
TCP, UDP, Logs, Scripts, Wire, Mobile

Modular Inputs
MQTT, AMQP, COAP, REST, JMS

HTTP Event Collector
Token Authenticated Events

Technology Partnerships
Kepware, AWS IoT, Cisco, Palo Alto

Real Time

splunk>enterprise splunk>cloud

Engineers | Data Analysts | Security Analysts | Business Users

Search | Alert | Visualize | Predict | Develop

External Lookups/Enrichment

Asset Info | Maintenance Info | Data Stores

Assistant : Detect Numeric Outliers

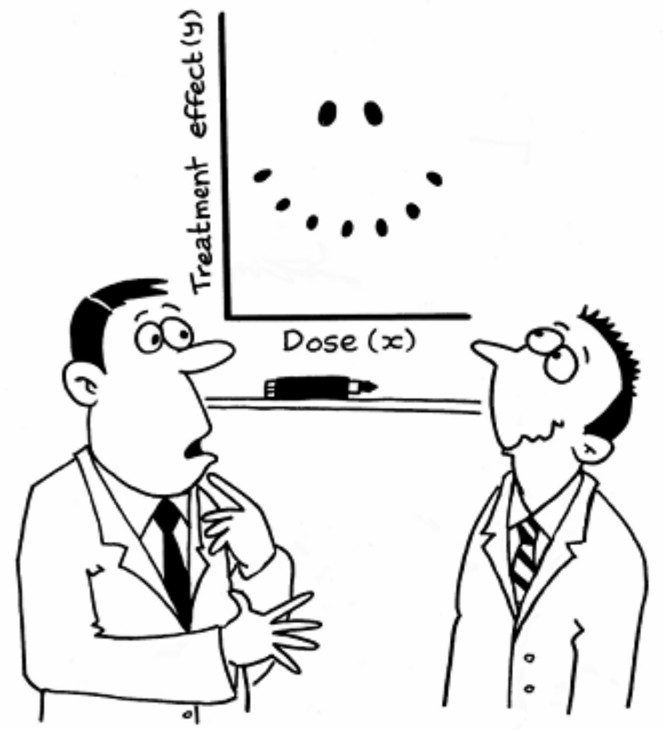
Step 3 : Generate Insights

- Place search in Detect Numeric Outliers Assistant
- Validate results using the visualizations
- Create alerts
- One unified workflow.
- Go customize to your hearts content

Assistant : Detect Numeric Outliers

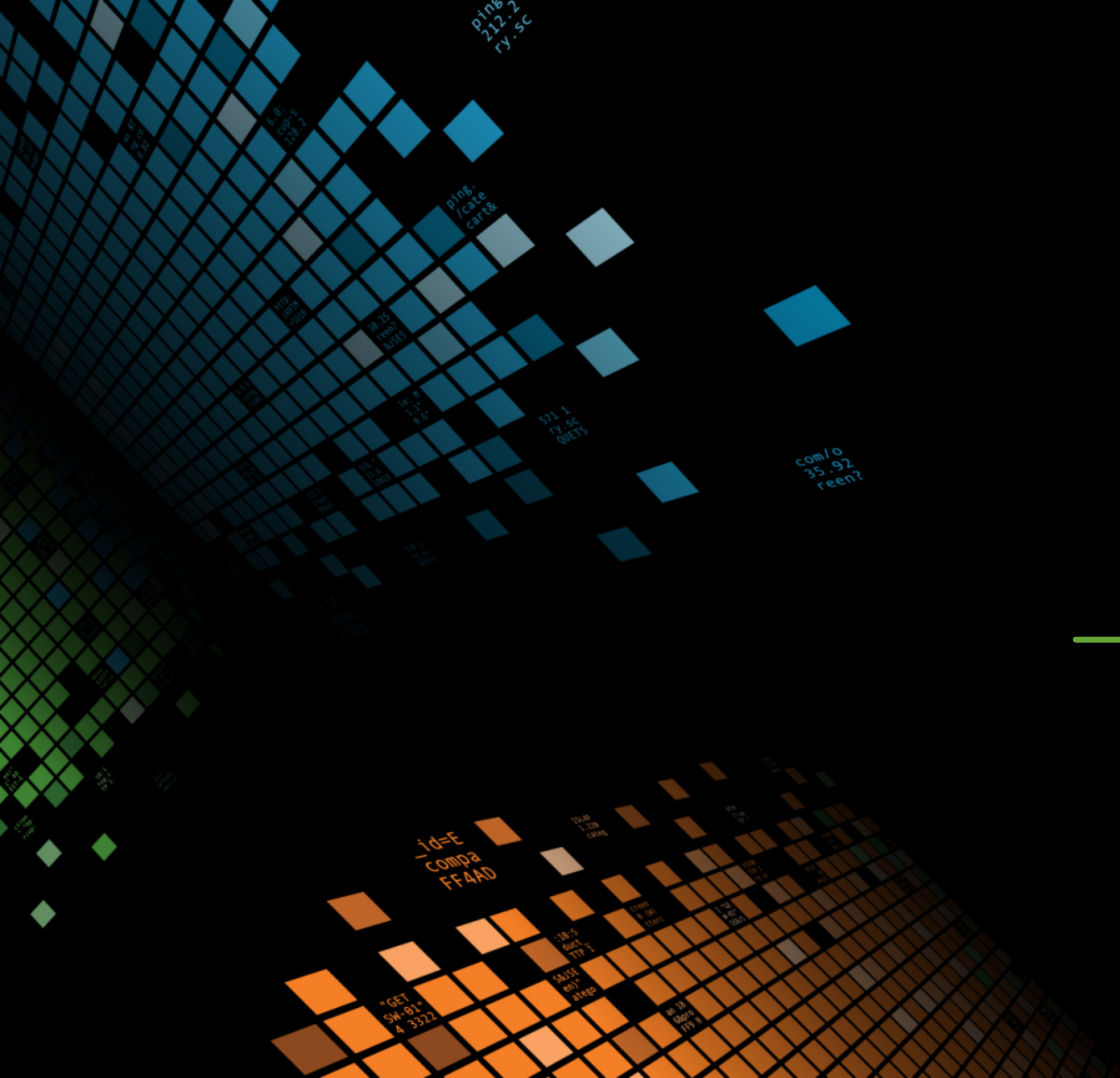
Step 4 : Publish Insights

- Reuse searches and visualizations
- Built your own report
- Schedule outlier detection searches



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

J.B. Landers ©



To the Toolkit Example!

Assistant: Predict Numeric Field

Assistant: Predict Numeric Field

One workflow, multiple use cases

BEST BUY

TELUS

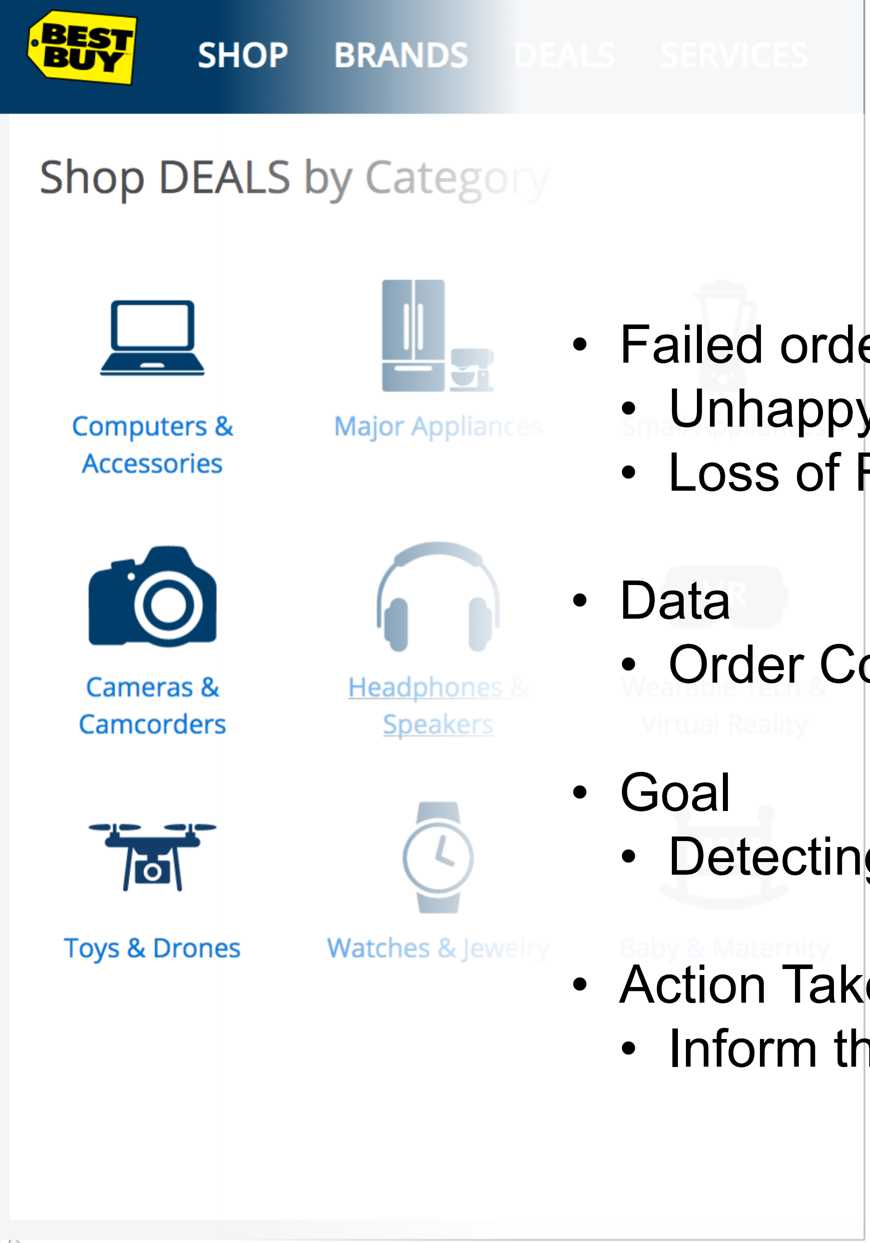
- Site Reliability Engineer

- Senior Design Specialist

Reduce Sales Loss

Improved Cell Tower Performance
Reduced Troubleshooting Time

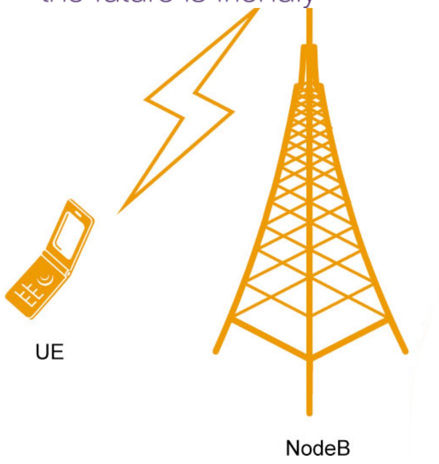
Aggregated Order Behavior



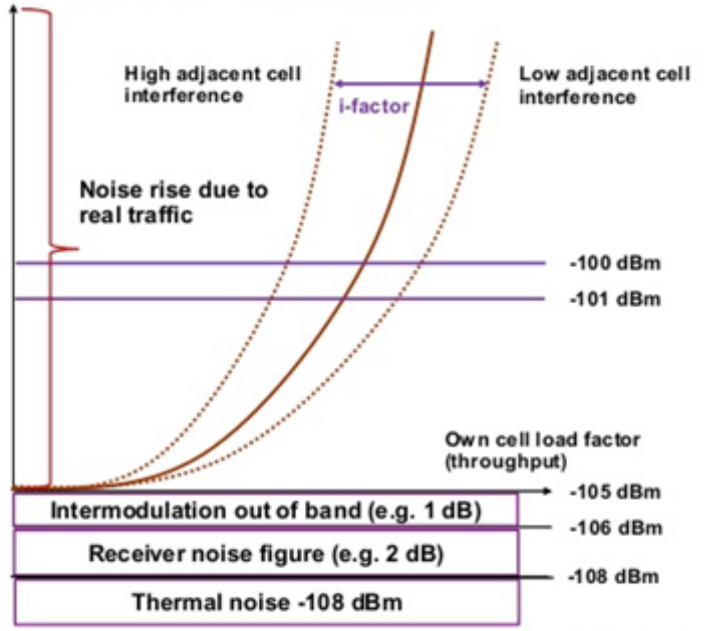
- Failed orders result in
  - Unhappy Customer
  - Loss of Revenue
- Data
  - Order Counts by HTTP Response Codes
- Goal
  - Detecting deviation in relationship with the response codes
- Action Taken
  - Inform the release team about the problems

TELUS
the future is friendly®

Radio Access Network Interference



- Interference
  - The level of noise within the frequency band in cell towers
- Uplink Interference Impacting Factors
  - Number of Subscribers
  - Connection Types
  - Radio Conditions
- Data
  - Uplink Rate/min for Each Cell Tower
  - Number of Subscribers
- Goal
  - Find underperforming cells and the characteristics of their problem.
- Action Taken
  - Reconfigure Underperforming Cells
  - Identify non-standard devices on the Network



splunk> .conf2017

Assistant : Predict Numeric Field

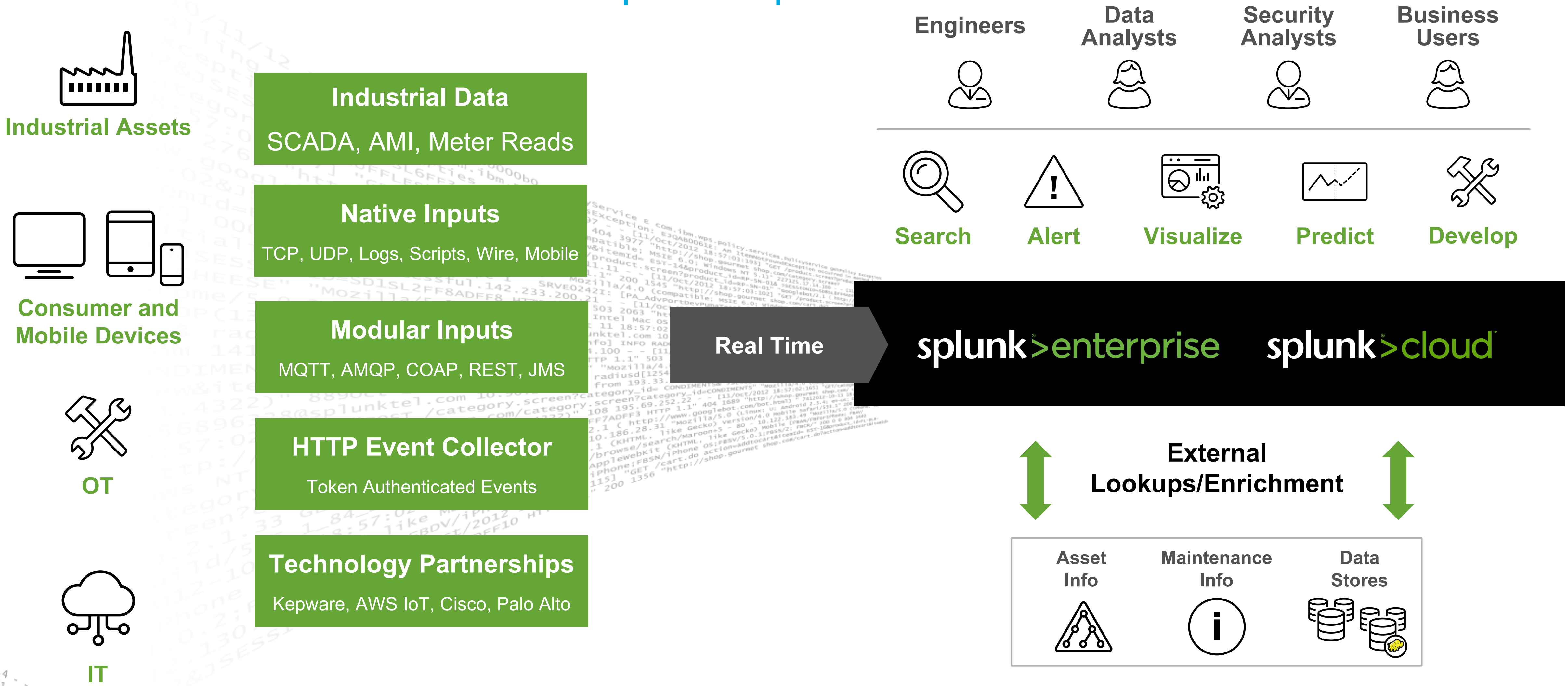
Step 1 : Data Driven Decisions

- Define your Problem!

  - I want to **predict a numeric field** given **multiple other fields**.
  - I want to **predict** the **future**.

Assistant : Predict Numeric Field

Step 2 : Acquire Data



Industrial Assets

Consumer and Mobile Devices

OT

IT

Industrial Data
SCADA, AMI, Meter Reads

Native Inputs
TCP, UDP, Logs, Scripts, Wire, Mobile

Modular Inputs
MQTT, AMQP, COAP, REST, JMS

HTTP Event Collector
Token Authenticated Events

Technology Partnerships
Kepware, AWS IoT, Cisco, Palo Alto

Real Time

splunk>enterprise splunk>cloud

Engineers | Data Analysts | Security Analysts | Business Users

Search | Alert | Visualize | Predict | Develop

External Lookups/Enrichment

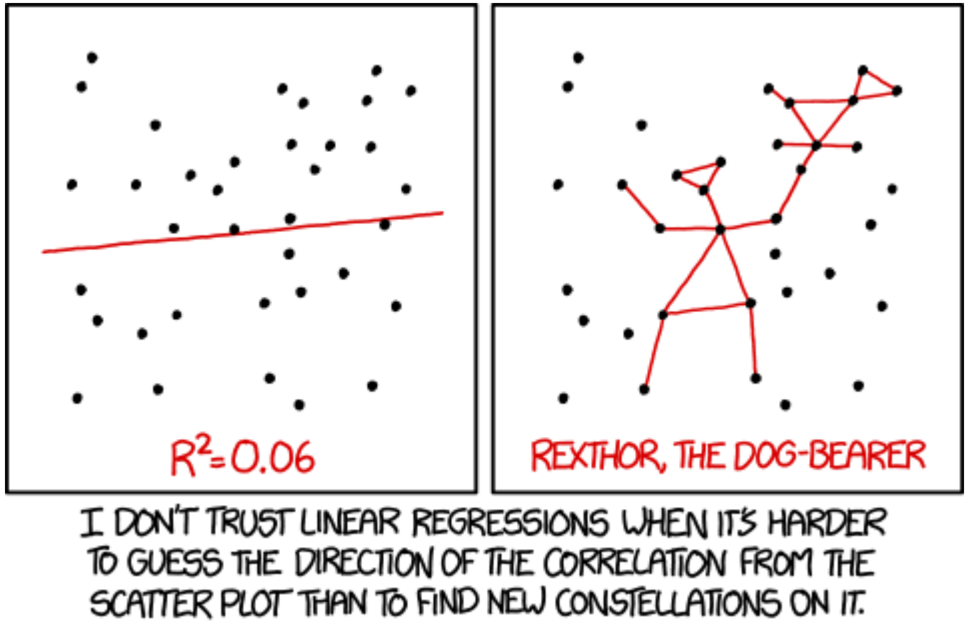
Asset Info | Maintenance Info | Data Stores

splunk> .conf2017

Assistant : Predict Numeric Field

Step 3 : Generate Insights

- Preprocessing Step
- 7 Algorithms
- Model Analysis
- Customized Visualizations
- Create Alerts

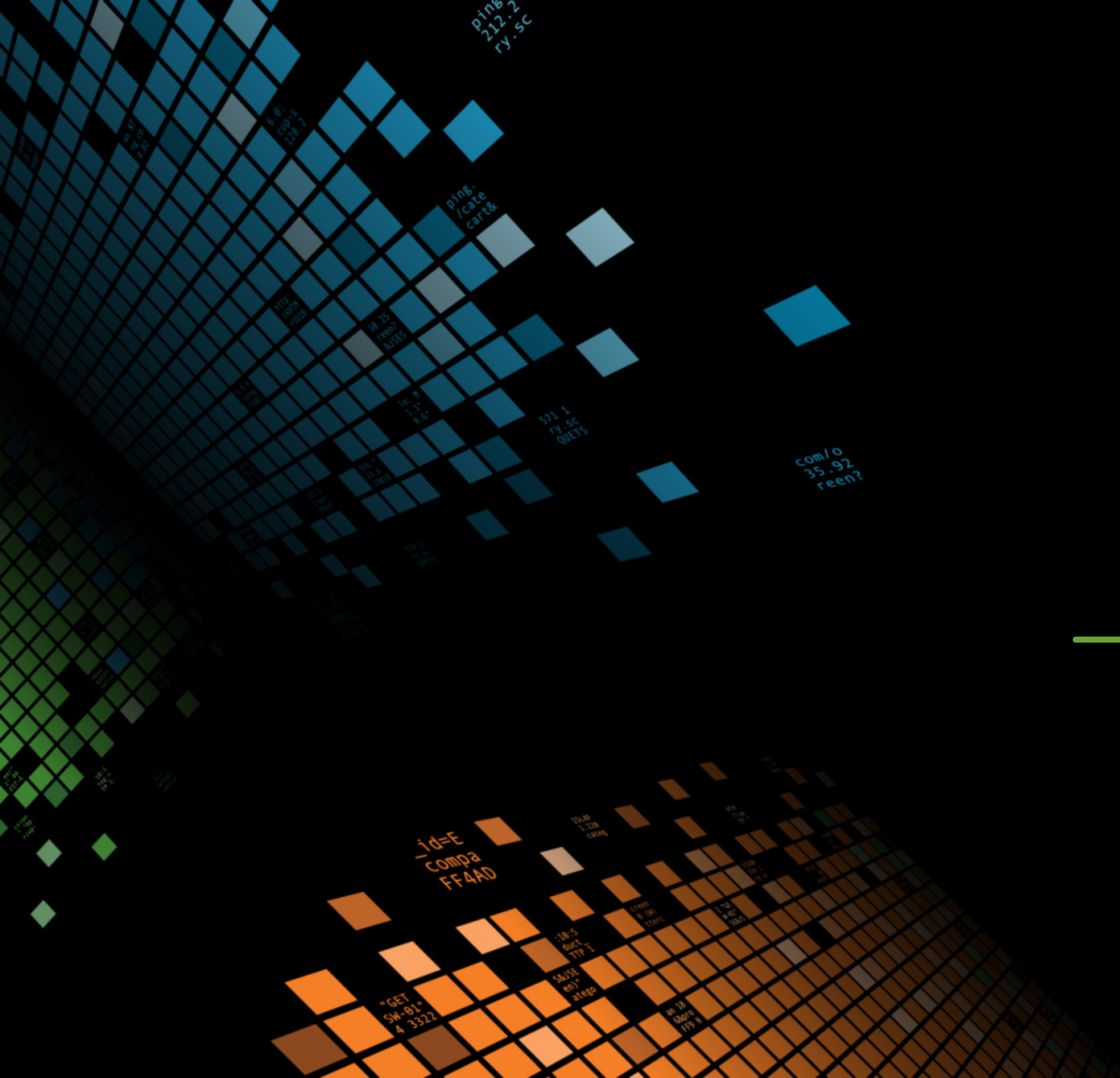

R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

https://xkcd.com/1725/

Assistant : Predict Numeric Field

Step 4 : Publish Insights

- Consume Predicted Field as an Alert
  - Directly from the Assistant, or
  - In another Search Bar
- Schedule Model Training
  - Batch Training
  - Partial Training
- Alert on Model Deviation



To the Toolkit Example!

Assistant: Predict Categorical Field

Assistant: Predict Categorical Field

One workflow, multiple use cases

Zillow®

- Wizard

Work in Progress

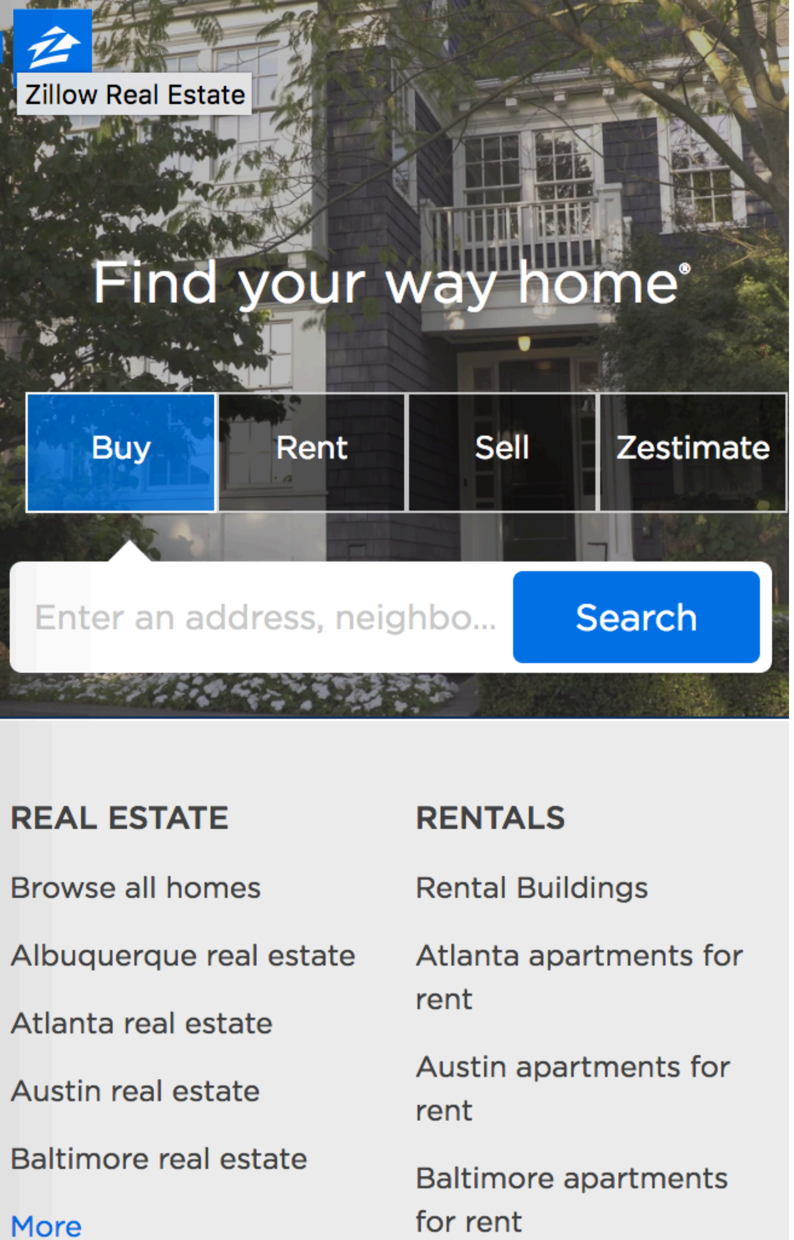
FERGUSON®

- Senior Design Specialist

At least 15% increase in expected sales adoption

Sales Monitoring

- Enterprise Adoption of Sales Analytics
- Data
  - Sales data
- Goal
  - Diminish Churn
  - Tailor the Customer Experience
- Action Taken
  - Align Sales Initiatives With Marketing Initiatives



REAL ESTATE

- Browse all homes
- Albuquerque real estate
- Atlanta real estate
- Austin real estate
- Baltimore real estate
- More

RENTALS

- Rental Buildings
- Atlanta apartments for rent
- Austin apartments for rent
- Baltimore apartments for rent

Bot Detection

- Good and bad bots scrapping
- Bad bots go unnoticed
- A Liability Issue
- Data
  - Browsing Log
- Goal
  - Detect Bad Bots
- Action Taken
  - Ban or reduce access for malicious scrapping

Assistant : Predict Categorical Field

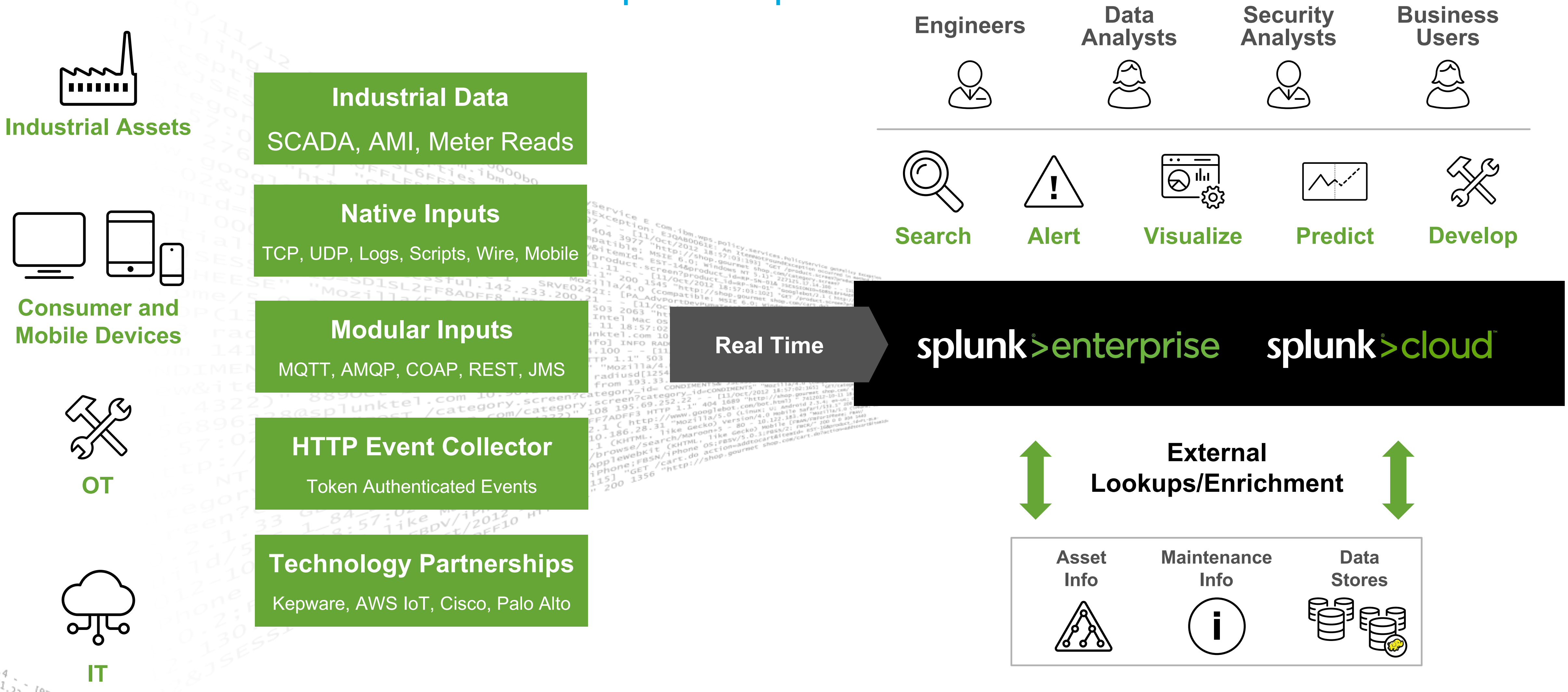
Step 1 : Data Driven Decisions

- Define your Problem!

  - I want to **predict a categorical field** given **multiple other fields**.
    - Good or Bad Bot
    - Malware or Not
    - Churn or Not
    - Face Recognition
    - Cat or Dog or …

Assistant: Predict Categorical Field

Step 2 : Acquire Data



Industrial Assets

Consumer and Mobile Devices

OT

IT

Industrial Data
SCADA, AMI, Meter Reads

Native Inputs
TCP, UDP, Logs, Scripts, Wire, Mobile

Modular Inputs
MQTT, AMQP, COAP, REST, JMS

HTTP Event Collector
Token Authenticated Events

Technology Partnerships
Kepware, AWS IoT, Cisco, Palo Alto

Real Time

splunk>enterprise splunk>cloud

Engineers | Data Analysts | Security Analysts | Business Users

Search | Alert | Visualize | Predict | Develop

External Lookups/Enrichment

Asset Info | Maintenance Info | Data Stores

splunk> .conf2017

Assistant : Predict Categorical Field

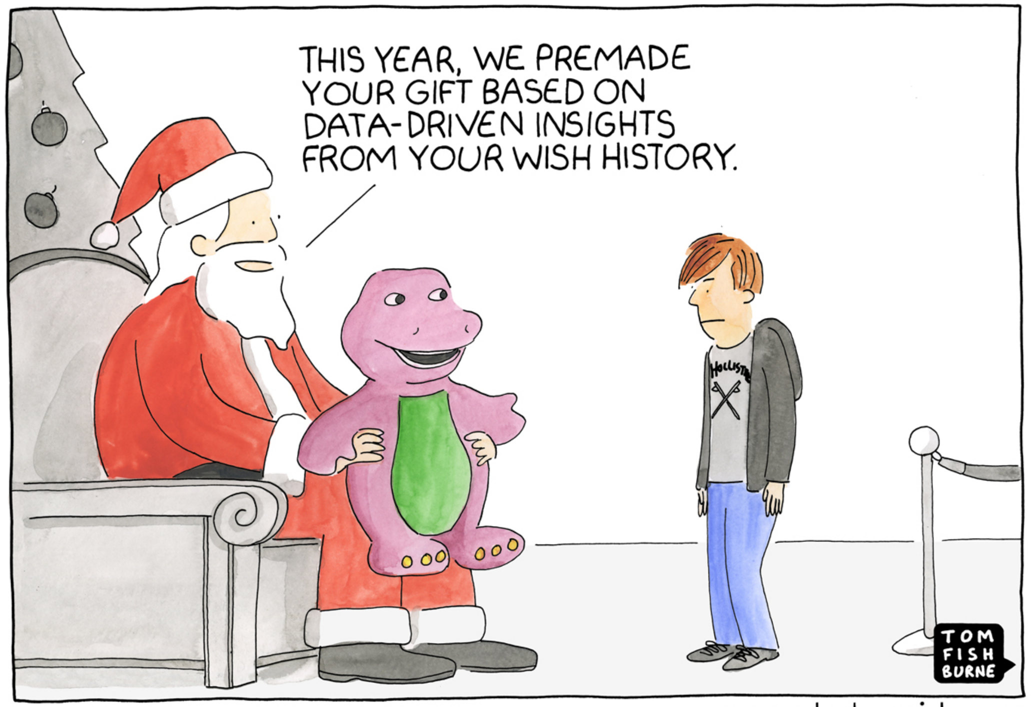
Step 3 : Generate Insights

- Preprocessing Step
- 6 Algorithms
- Model Analysis
- Custom Visualizations
- Create Alerts

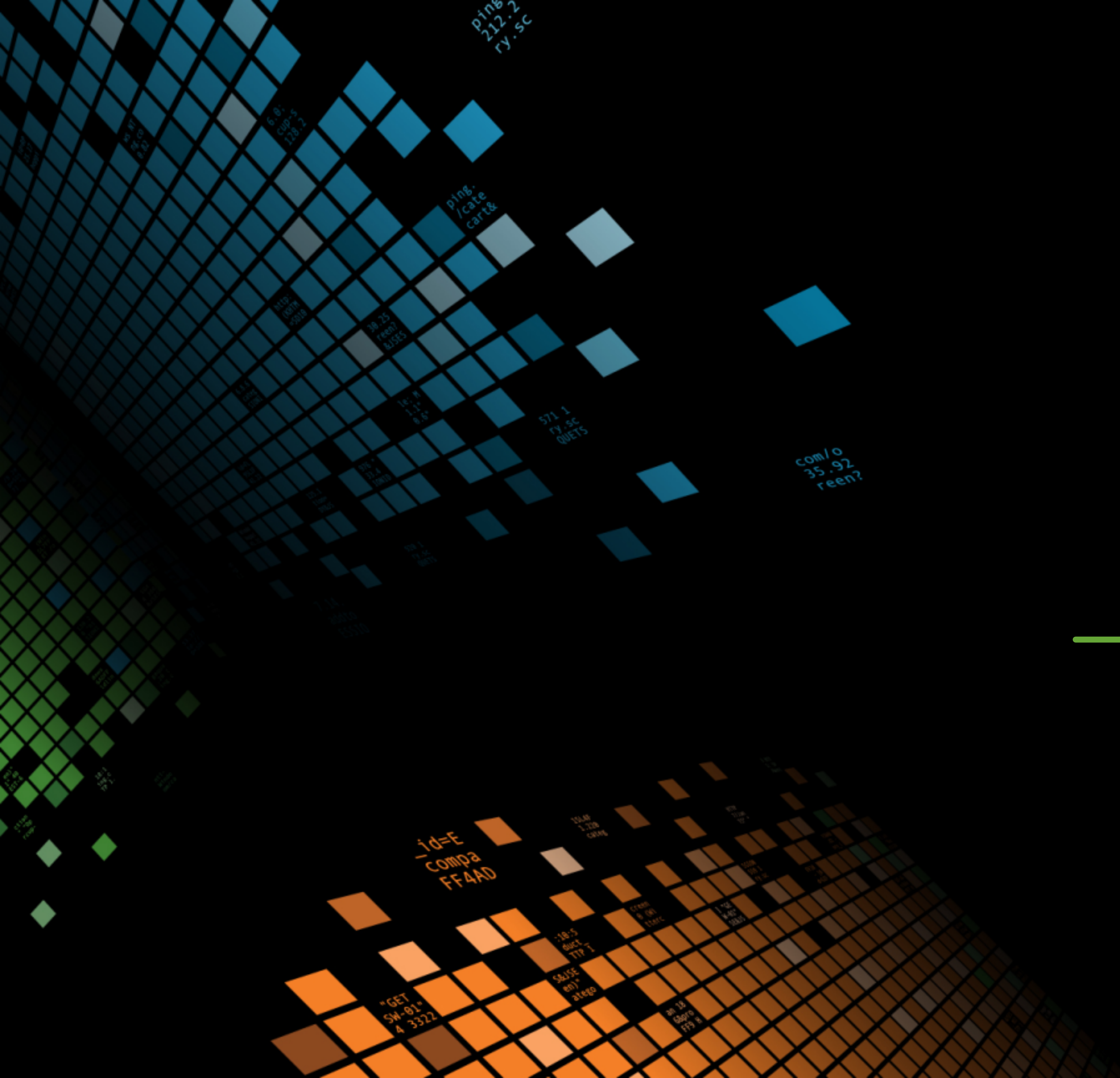
Assistant : Predict Categorical Field

Step 4 : Publish Insights

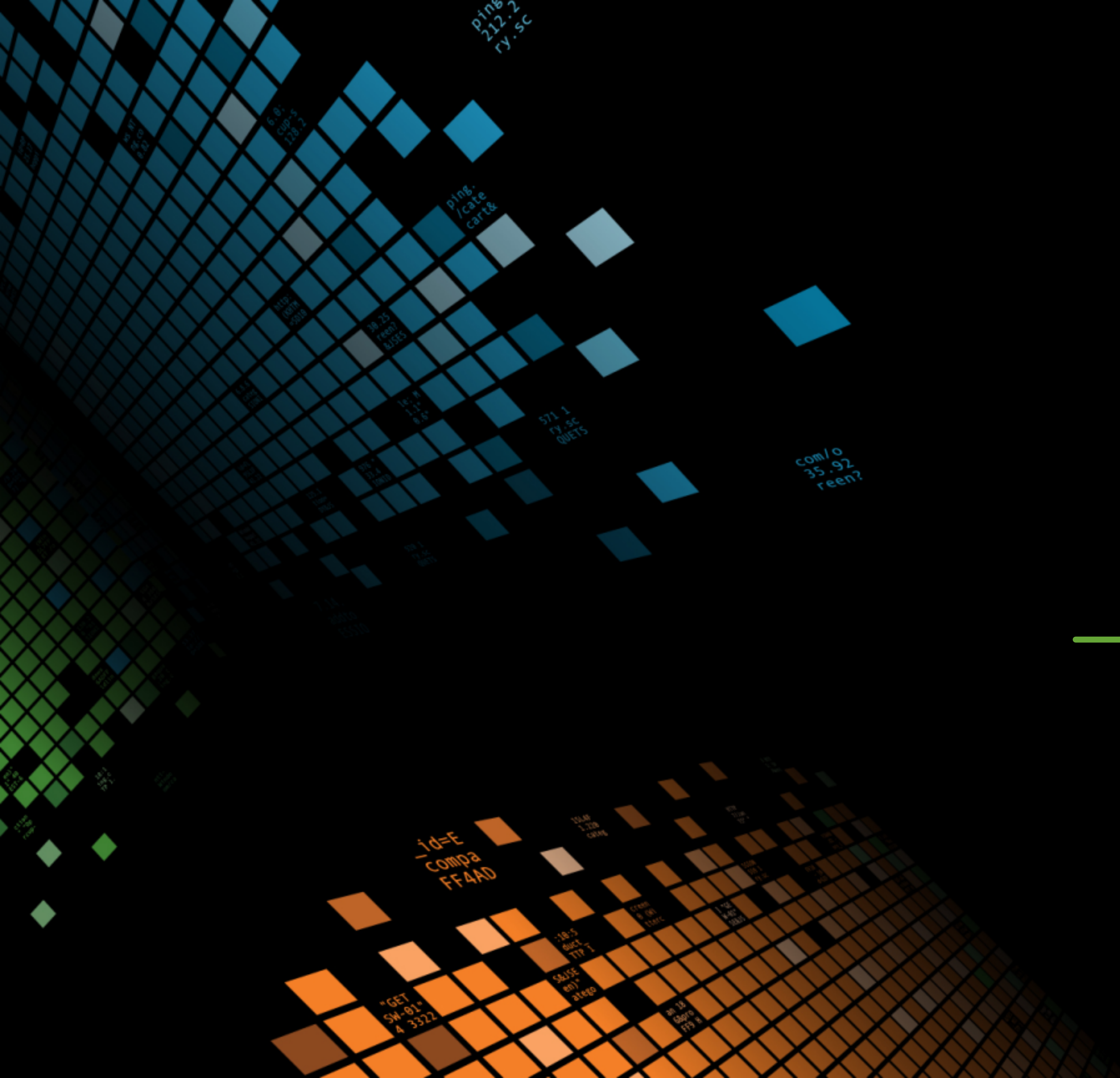
- Consume Predicted Field as an Alert
  - Directly from the Assistant, or
  - In another Search Bar
- Schedule Model Training
  - Batch Training
  - Partial Training
- Alert on Model Deviation


THIS YEAR, WE PREMADE YOUR GIFT BASED ON DATA-DRIVEN INSIGHTS FROM YOUR WISH HISTORY.
TOM FISH BURNE

© marketoonist.com



To the Toolkit Example!



Q&A

Go See These Talks

- Advanced Machine Learning Using the Extensible ML API
  - *Alexander Johnson & Zidong Yang*
- Automation of Event Correlation and Clustering With Built-In Machine Learning Algorithms in Splunk IT Service Intelligence (ITSI)
  - *Vineetha Bettaiah & Ross Lazerowitz*
- Prioritizing Anomalies Using the Machine Learning Toolkit
  - *Harsh Keswani*
- Splunk Machine Learning Capabilities and Condition-Based Maintenance: Train Doors on the German Public Rail Transport System
  - *Henning Brandt & Daniel Pal*