

splunk®> 4 rookies

Machine Learning for Security



Forward-looking statements

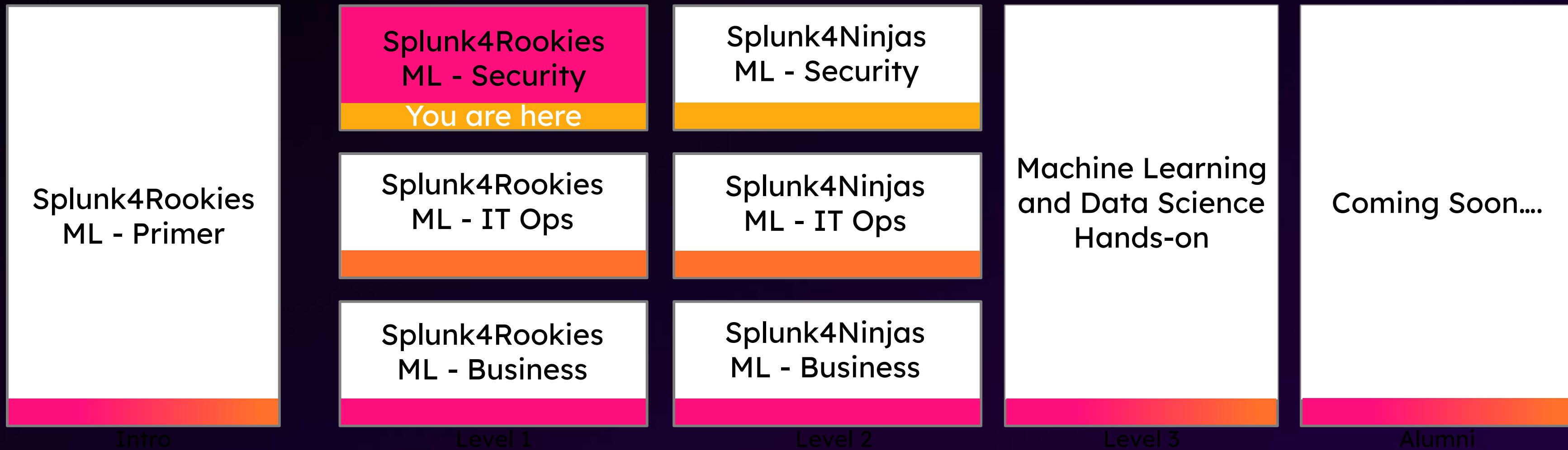
This presentation may contain forward-looking statements that are subject to the safe harbors created under the Securities Act of 1933, as amended, and the Securities Exchange Act of 1934, as amended. All statements other than statements of historical facts are statements that could be deemed forward-looking statements. These statements are based on current expectations, estimates, forecasts, and projections about the industries in which we operate and the beliefs and assumptions of our management based on the information currently available to us. Words such as "expects," "anticipates," "targets," "goals," "projects," "intends," "plans," "believes," "momentum," "seeks," "estimates," "continues," "endeavors," "strives," "may," variations of such words, and similar expressions are intended to identify such forward-looking statements. In addition, any statements that refer to (1) our goals, commitments, and programs; (2) our business plans, initiatives, and objectives; and (3) our assumptions and expectations, including our expectations regarding our financial performance, products, technology, strategy, customers, markets, acquisitions and investments are forward-looking statements. These forward-looking statements are not guarantees of future performance and involve significant risks, uncertainties and other factors that may cause our actual results, performance or achievements to be materially different from results, performance or achievements expressed or implied by the forward-looking statements contained in this presentation. Readers are cautioned that these forward-looking statements are only predictions and are subject to risks, uncertainties, and assumptions that are difficult to predict, including those identified in the "Risk Factors" section of Cisco's most recent report on Form 10-Q filed on May 21, 2024 and its most recent report on Form 10-K filed on September 7, 2023. The forward-looking statements made in this presentation are made as of the time and date of this presentation. If reviewed after the initial presentation, even if made available by Cisco or Splunk, on Cisco or Splunk's website or otherwise, it may not contain current or accurate information. Cisco and Splunk undertake no obligation to revise or update any forward-looking statements for any reason, except as required by law.

In addition, any information about new products, features, functionality or 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 or be relied upon in making a purchasing decision. We undertake no commitment, promise or obligation either to develop the features or functionalities described, in beta or in preview (used interchangeably), or to include any such feature or functionality in a future release. The development, release, and timing of any features or functionality described for our products remains at our sole discretion.

Splunk and Splunk® 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. © 2025 SPLUNK LLC All rights reserved.



Splunk Machine Learning Workshops



Boss of Machine Learning (Coming Soon)

Introductory

Advanced

Legend

Security

IT Ops

Business

All

Leading Initiatives Driving ML Adoption

200%

Increase in proactive detection of security and performance issues, significantly reducing downtime

2.1x

More likely to have automated processes for alerts, helping operationalize data at scale

\$365k/hour

On average saved from costly outages, helping organizations protect against revenue loss

Sources:

Harvard Business Review - Artificial Intelligence for the Real World Digital Enterprise Journal Report: The Roadmap to Becoming a Top Performing Organization in Managing IT Operations

Obstacles Blocking ML Adoption

1.8x

Increase in data and events
to process every two years,
creating challenges in
handling data volume

1 in 2

Companies increase the number
of data silos, leading to difficulties
integrating ML in isolated systems

79%

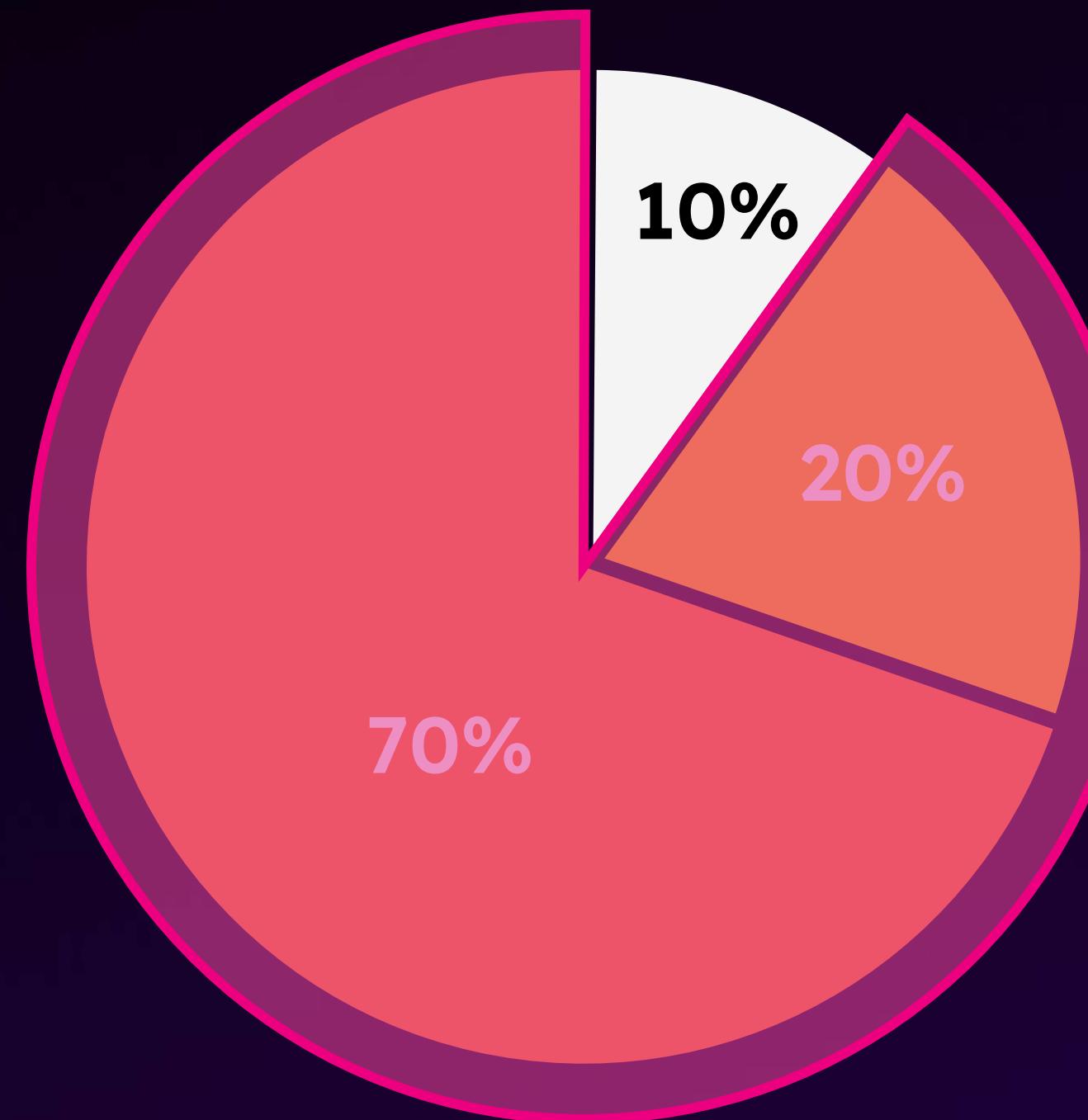
Failure rate for companies
which try to implement machine
learning from scratch, due
to lack of expertise

Sources:

Harvard Business Review - Artificial Intelligence for the Real World
Digital Enterprise Journal Report: The Roadmap to Becoming a Top Performing Organization in
Managing IT Operations

<https://www.xplm.com/news/press/industry-study-2023-companies-cannot-control-their-data-silos/>

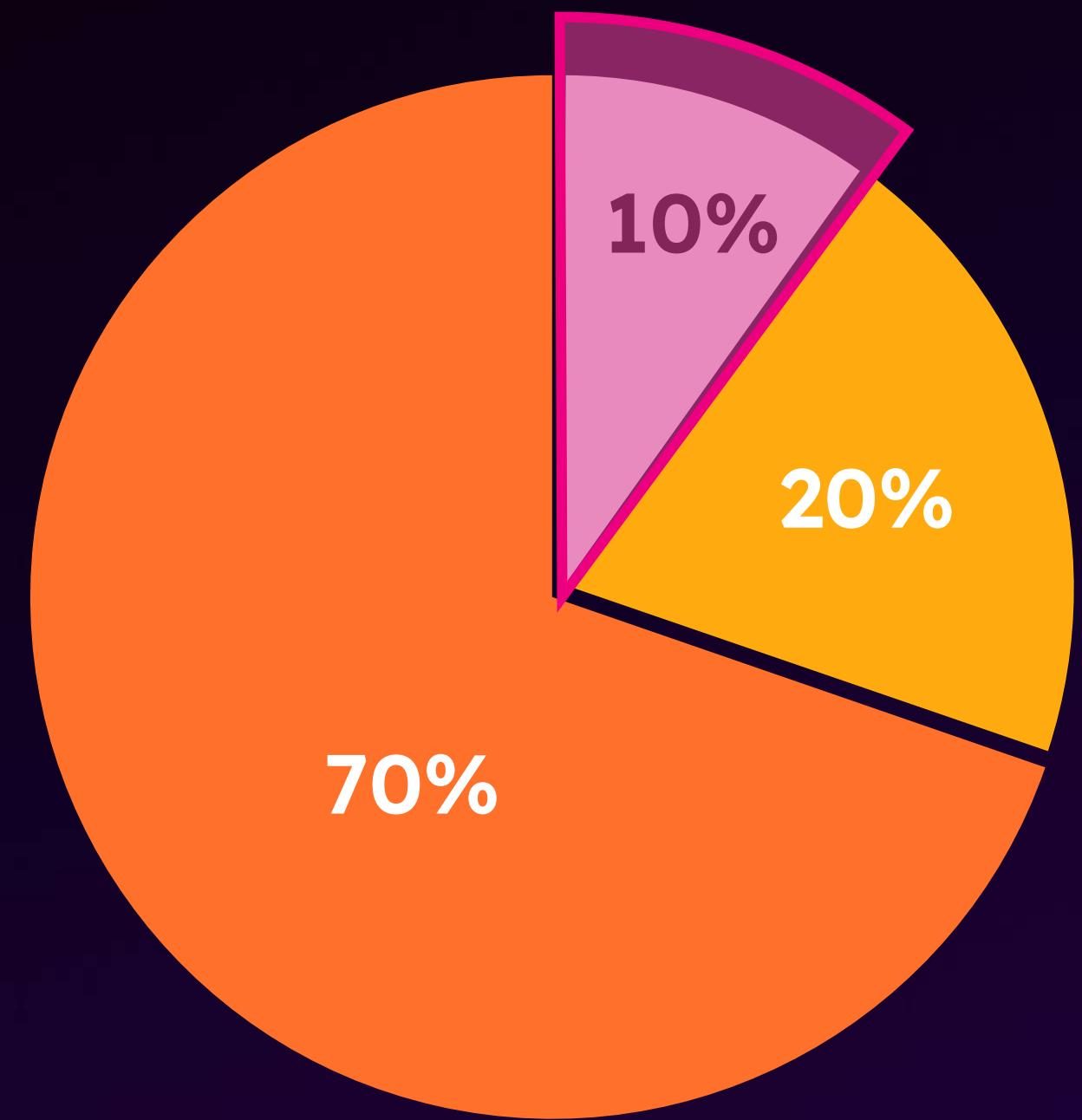
How Data Scientists Spend Their Time



- Data Engineering
- Machine Learning
- Other

Data originated from "[Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says](#)", Forbes Mar 23, 2016".

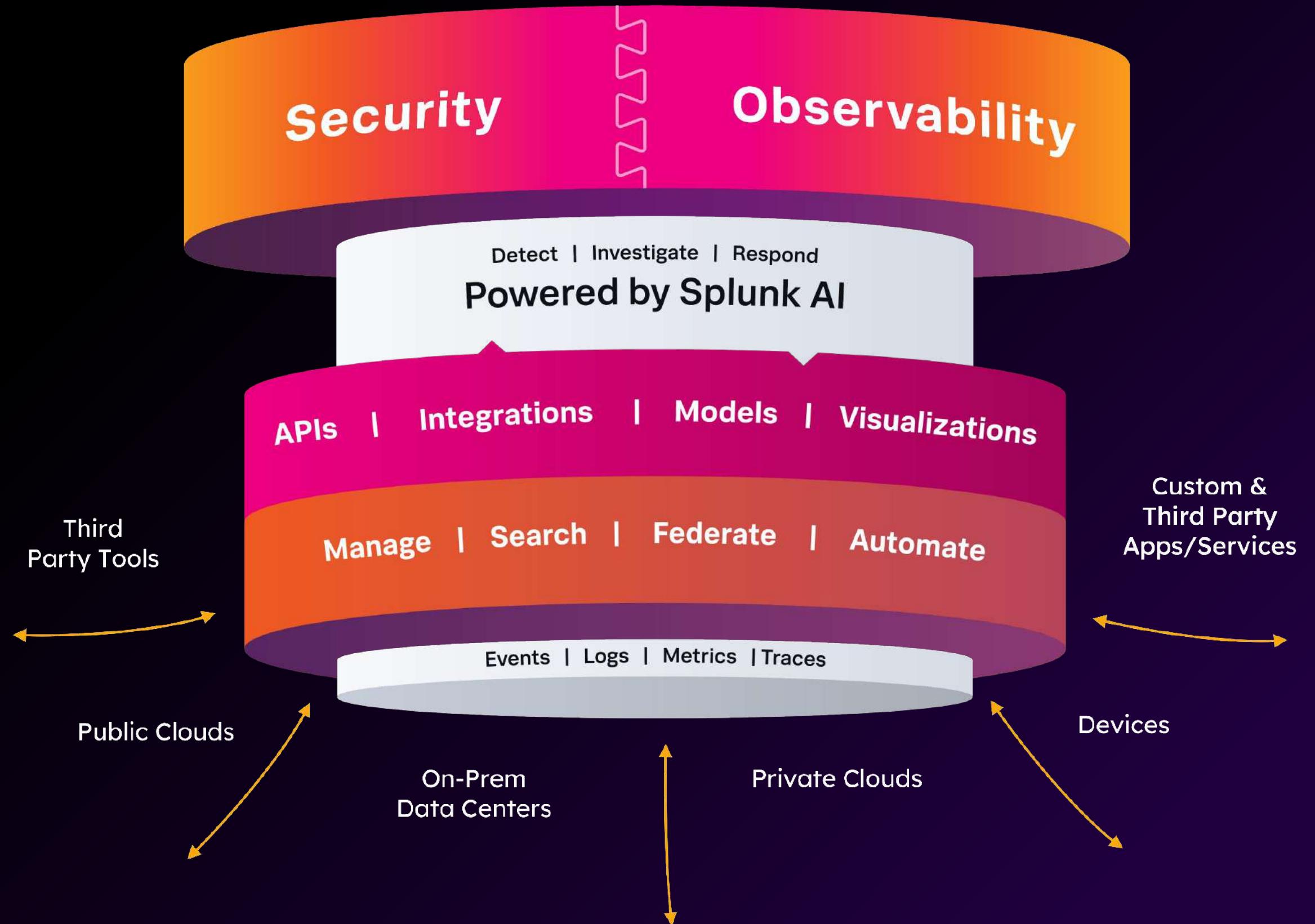
How Data Scientists Spend Their Time



- Data Engineering
- Machine Learning
- Other

Data originated from "[Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says](#)", Forbes Mar 23, 2016".

The Unified Security and Observability Platform



Cisco + Splunk power and protect the AI revolution.

Infrastructure for AI

Networking and compute solutions for training AI models at scale.

Data for AI

Massive breadth and depth of data across domains is the foundation for AI models.

Observability for AI

Observability supports trustworthy, performant, and reliable AI deployment.

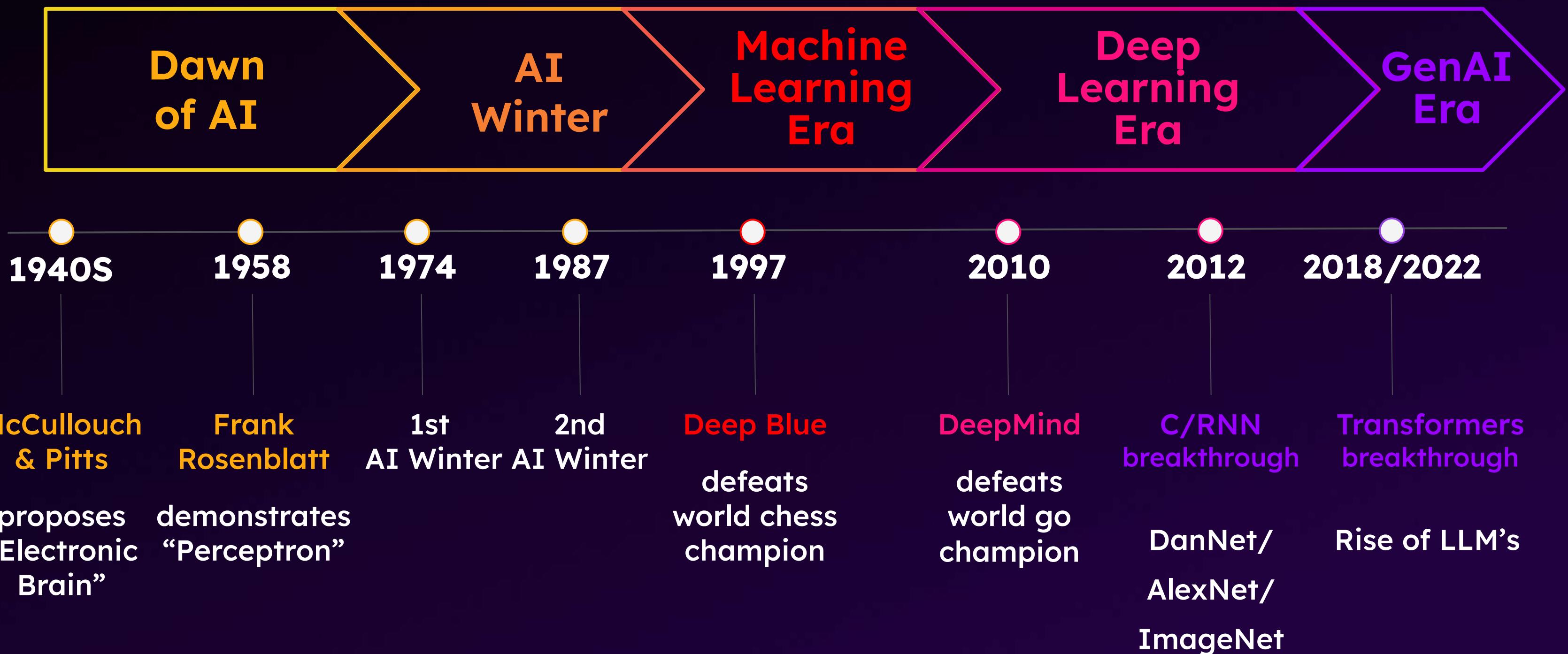
AI for Security

AI solutions improve efficiency, efficacy, and economics of defending against security threats.

Security for AI

Cisco's Identity Intelligence and User Protection and Splunk's analytics help protect enterprises in their use of AI from within.

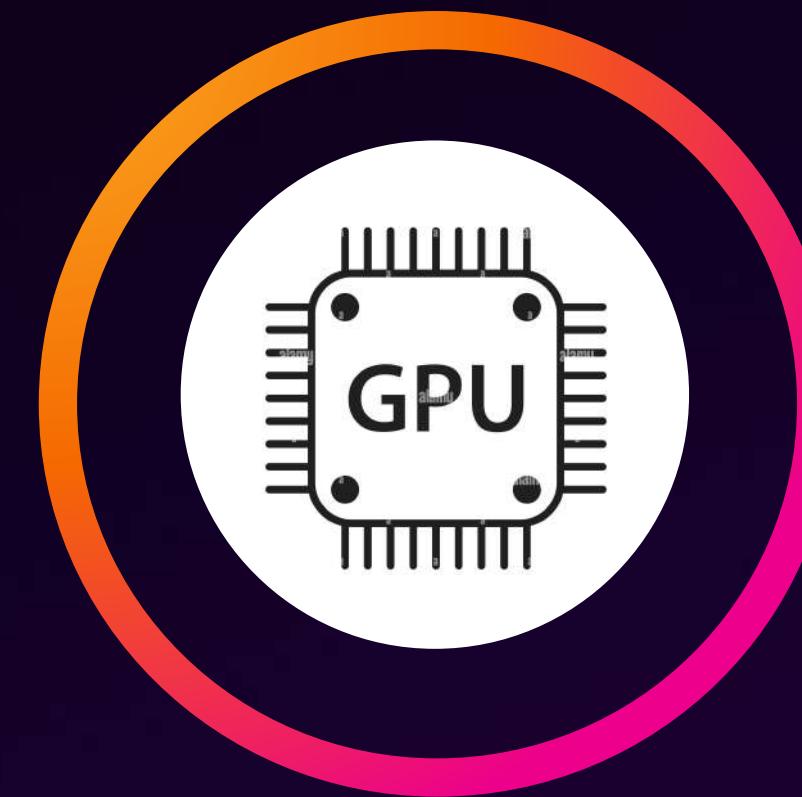
History of AI



What is driving the explosive growth of AI?



Algorithm

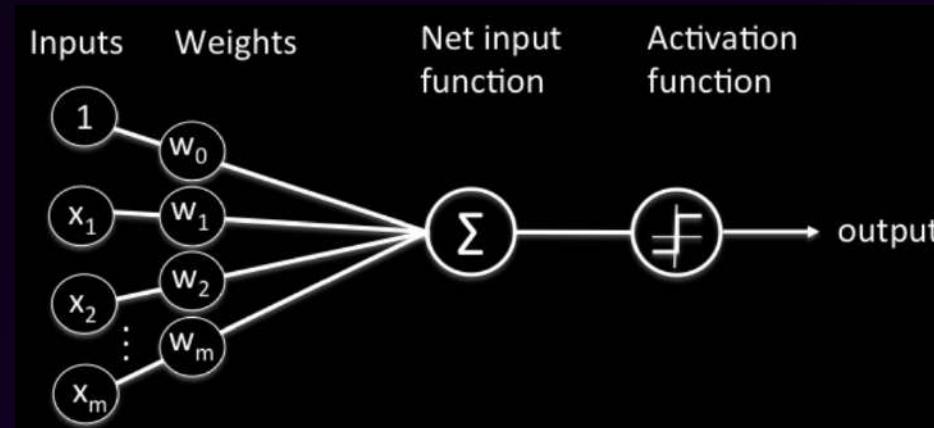


Compute

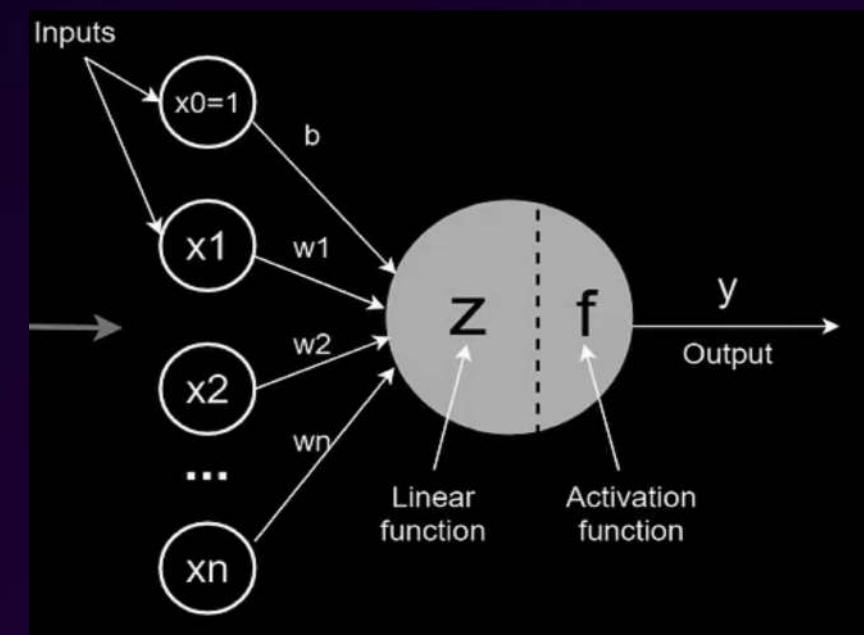
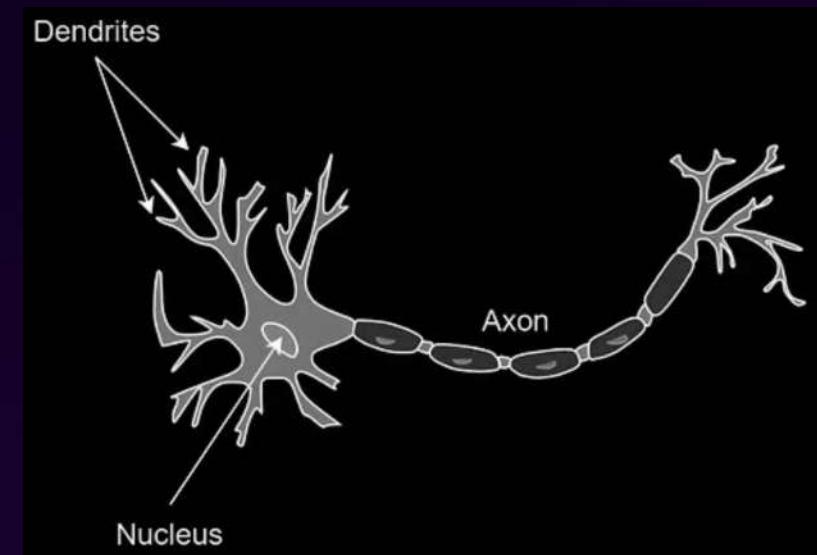


Data

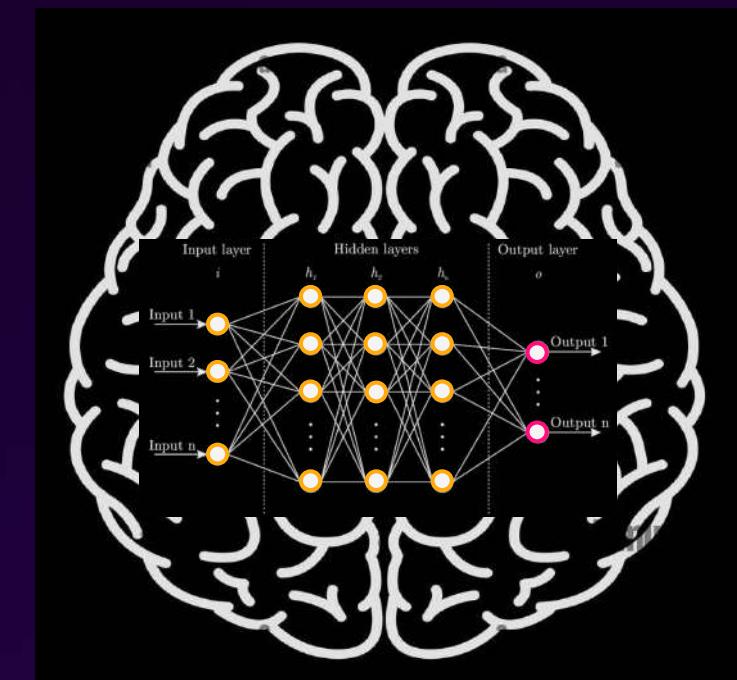
What is driving the explosive growth of AI?



1950's



2010's

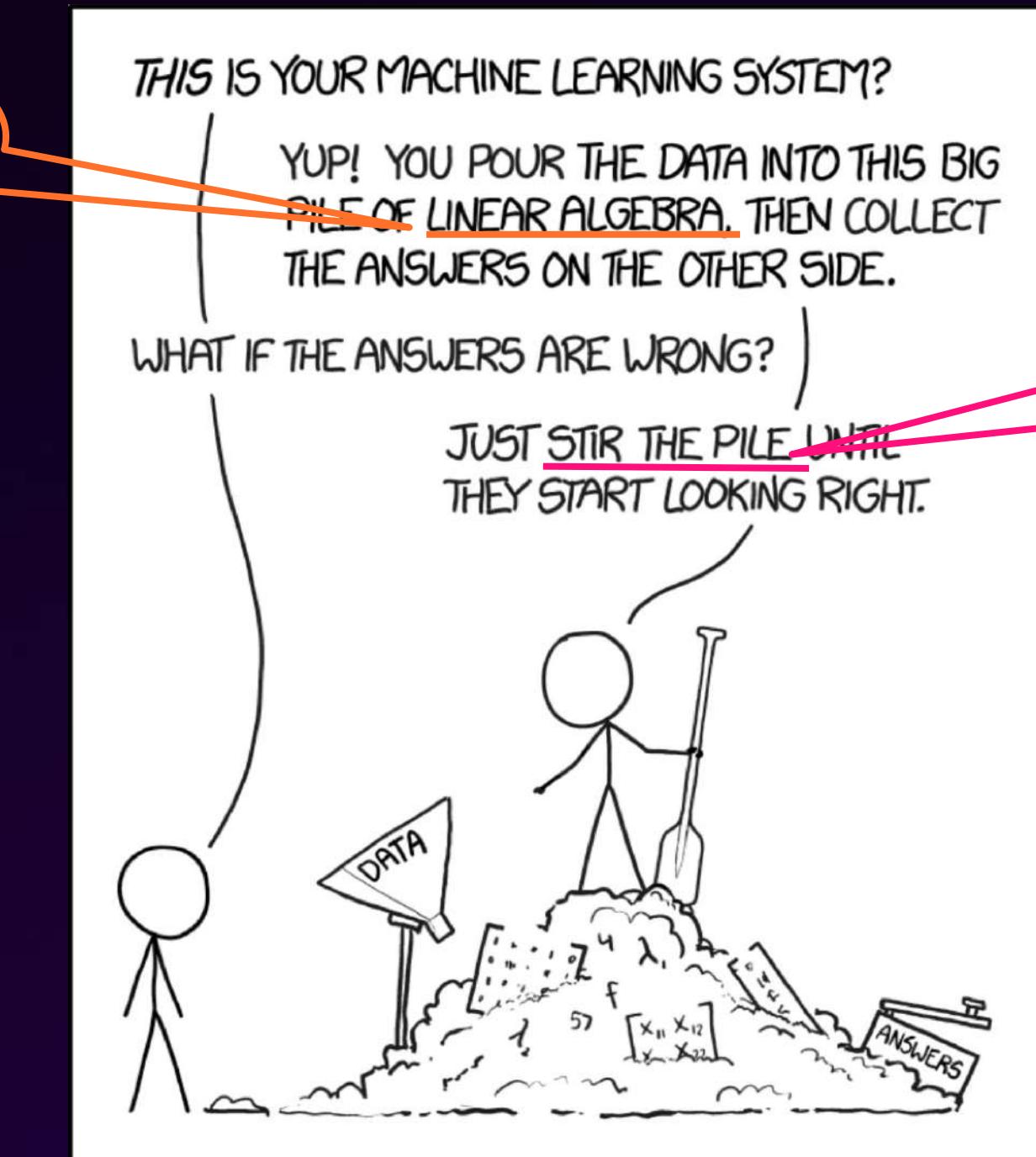


What is driving the explosive growth of AI?



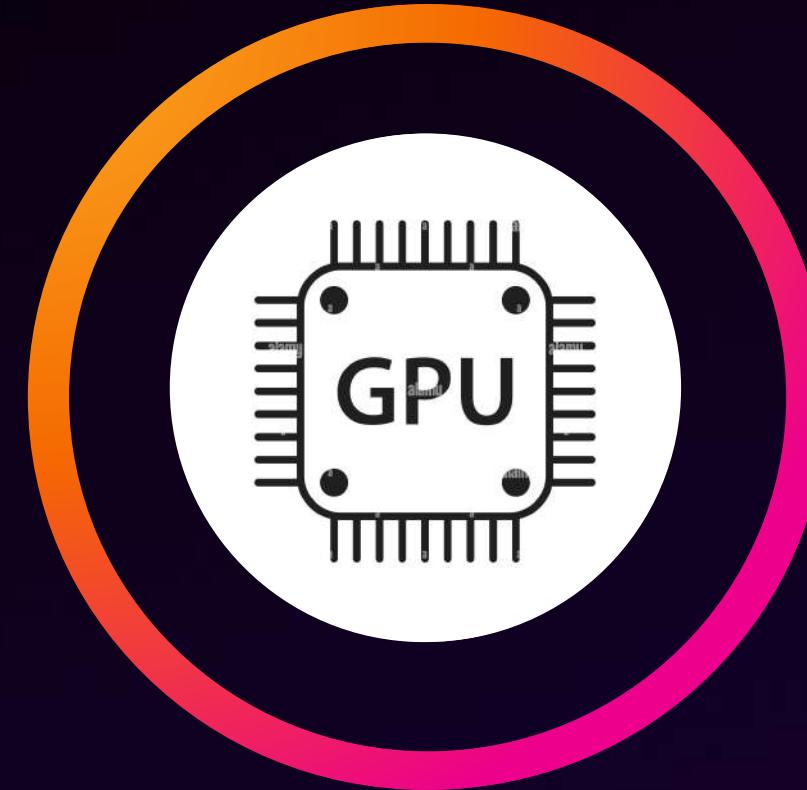
Algorithm

Matrix
Multiplication

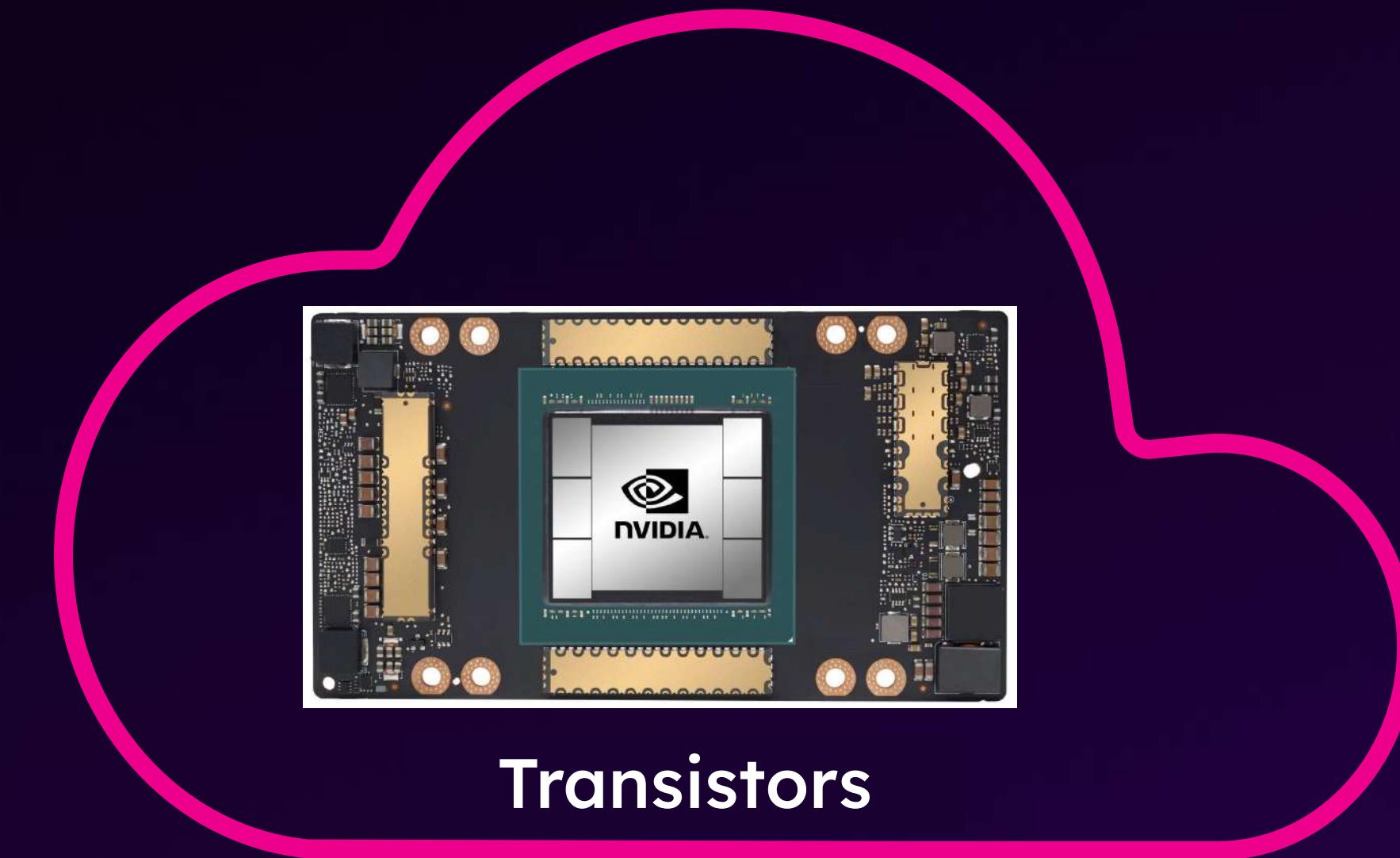


Backpropagation

What is driving the explosive growth of AI?



Compute

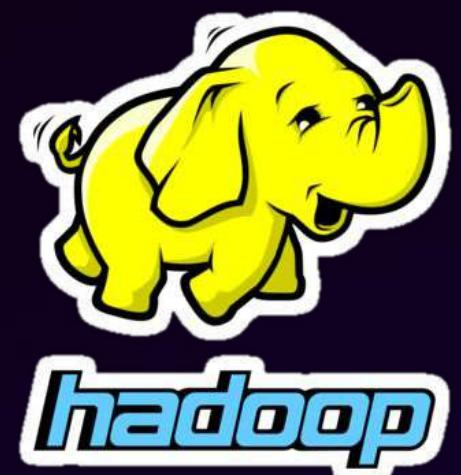


Transistors

What is driving the explosive growth of AI?



Data



WIKIPEDIA
The Free Encyclopedia



Google



Github

Different Levels of Teaching Machines



Artificial Intelligence

The broad study of teaching a computer
to process data and make decisions

Different Levels of Teaching Machines



Artificial Intelligence

The broad study of teaching a computer to process data and make decisions



Machine Learning

Subset of AI. Predictions and insight with minimal human interference

Different Levels of Teaching Machines



Artificial Intelligence

The broad study of teaching a computer to process data and make decisions



Machine Learning

Subset of AI. Predictions and insight with minimal human interference



Deep Learning

Subset of ML. Predictions via neural networks

Different Levels of Teaching Machines



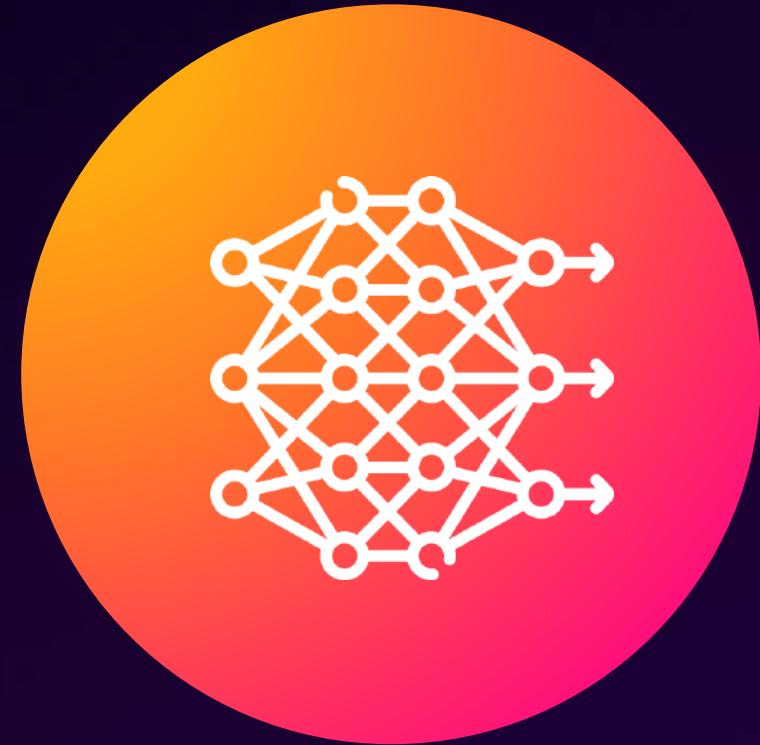
Artificial Intelligence

The broad study of teaching a computer to process data and make decisions



Machine Learning

Subset of AI. Predictions and insight with minimal human interference



Deep Learning

Subset of ML. Predictions via neural networks

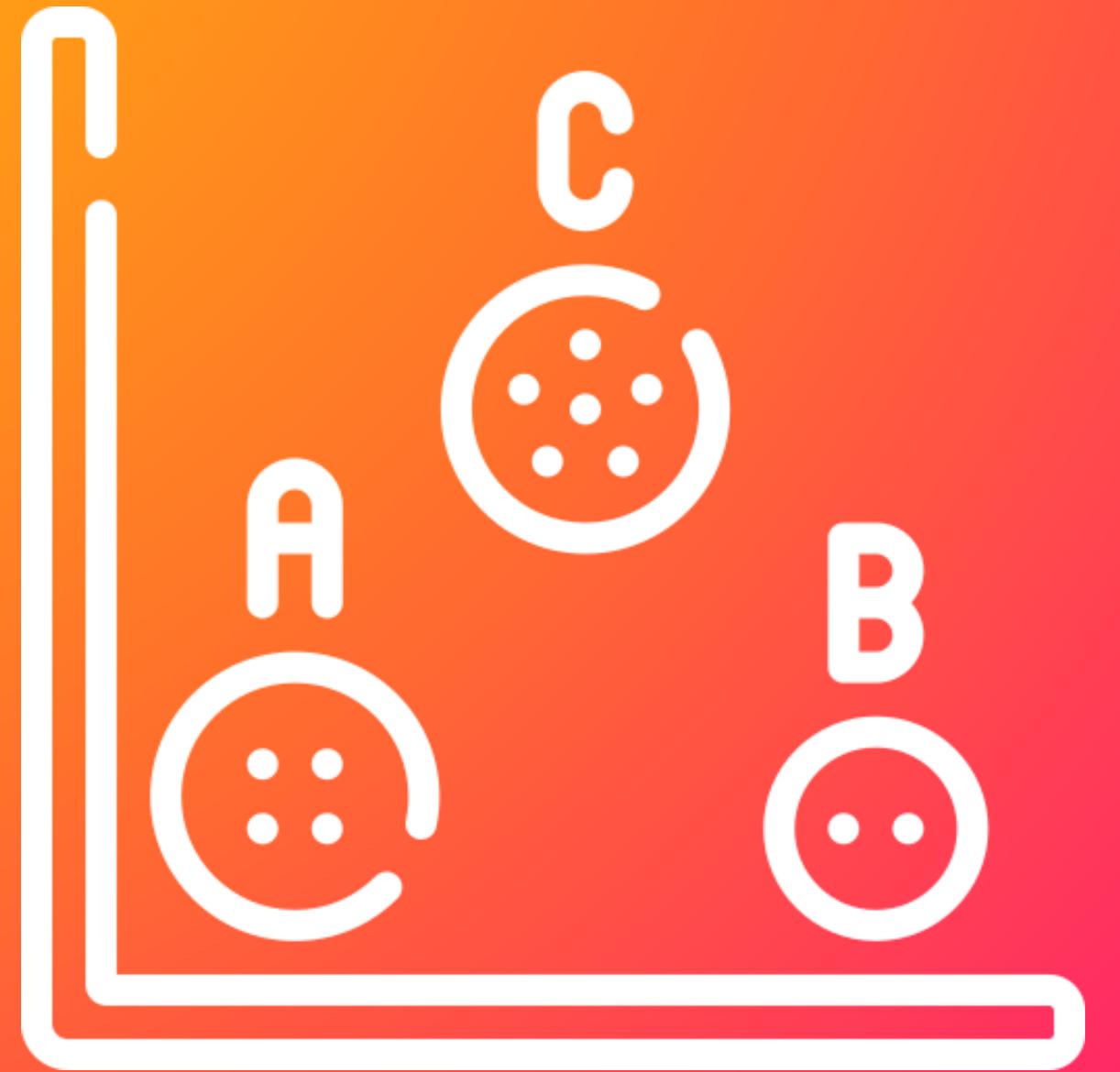
Predictive Algorithms

Methods that help you get ahead of issues that may happen in the future

Includes:

- Numerical Regression
- Categorical Regression
- Time Series Forecasting





Categorization Algorithms

Uncover insights about
your data to quickly
respond in the present

Includes:

- Categorical Regression
- Clustering

Outlier Detection Algorithms

Identify and analyze
abnormal behavior in
your data

Includes:

- Clustering
- Outlier Detection



Easy to Operationalize

Every Search Can Use
Machine Learning



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk®

Search

Alert



Webhook



Email



Tickets



Third-Party
Applications



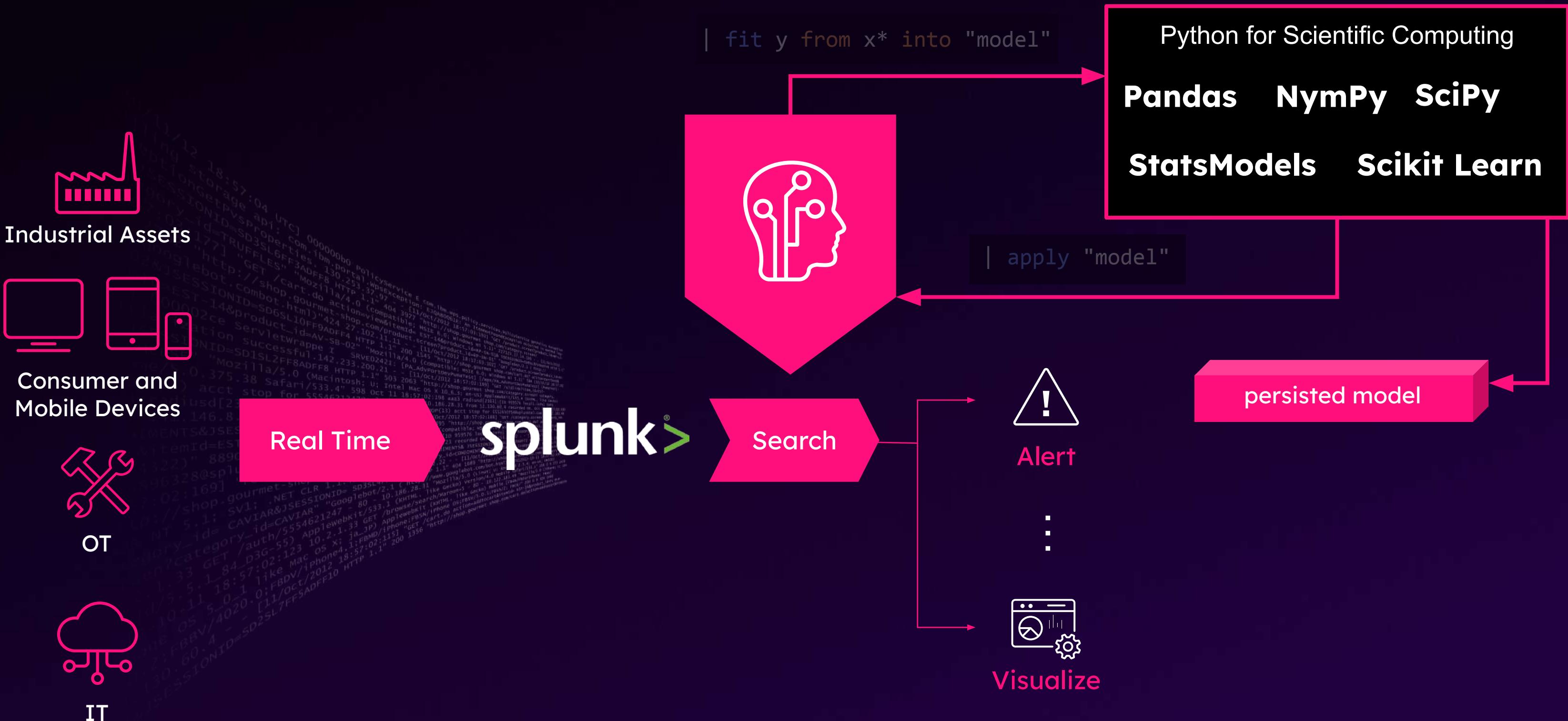
Smartphones
and Devices

File a ticket

Trigger process flow

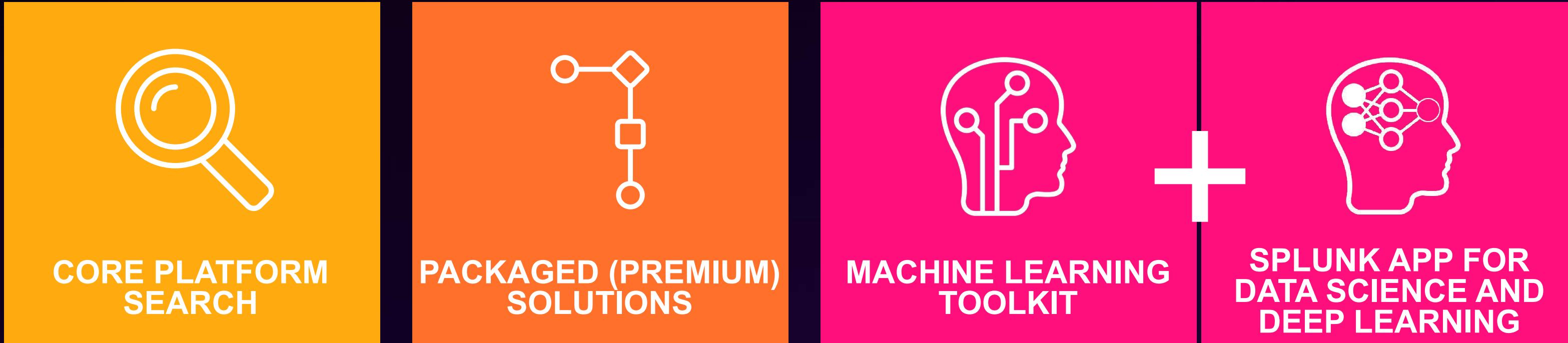
Send a text

Model Longevity

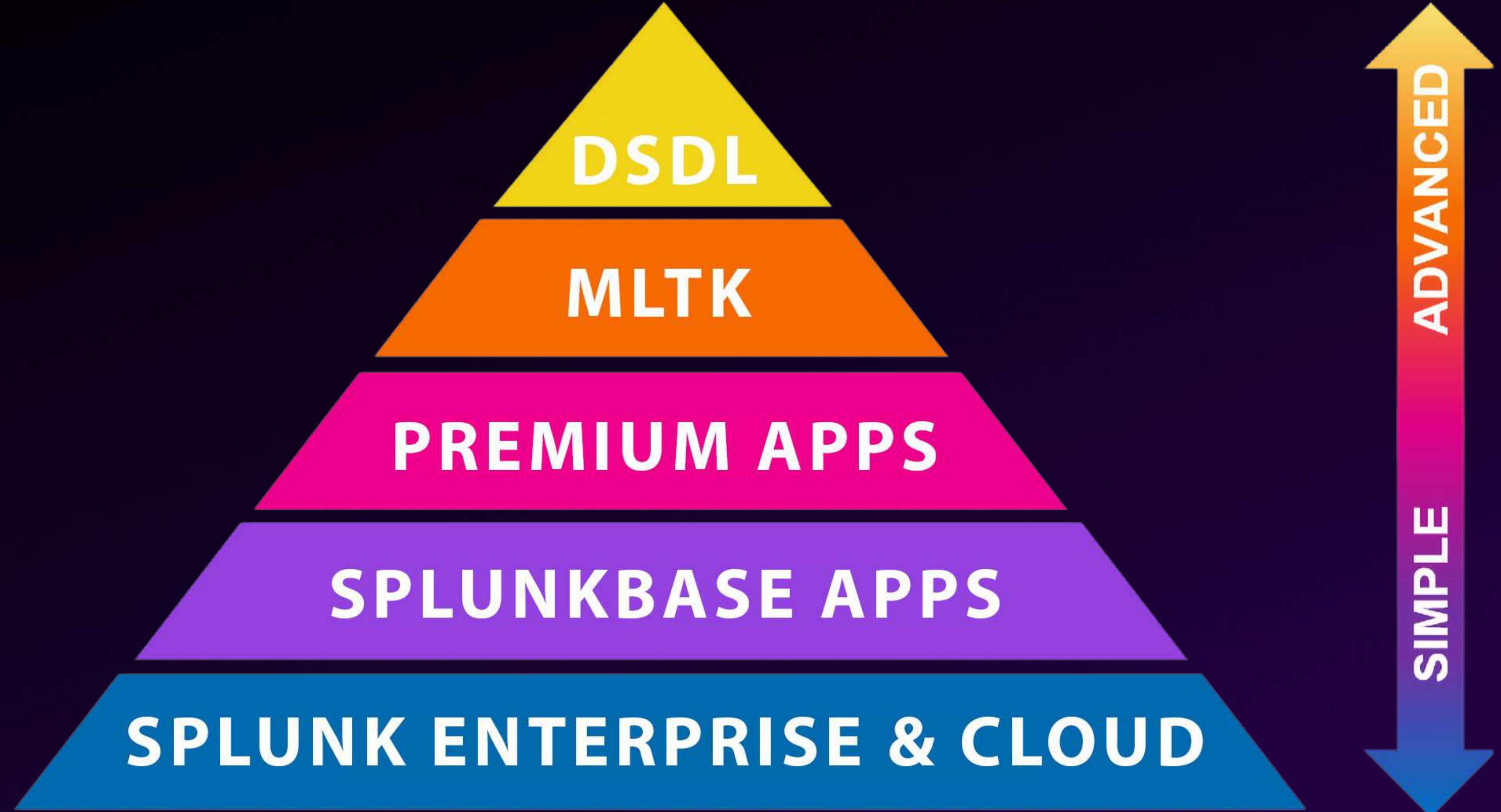


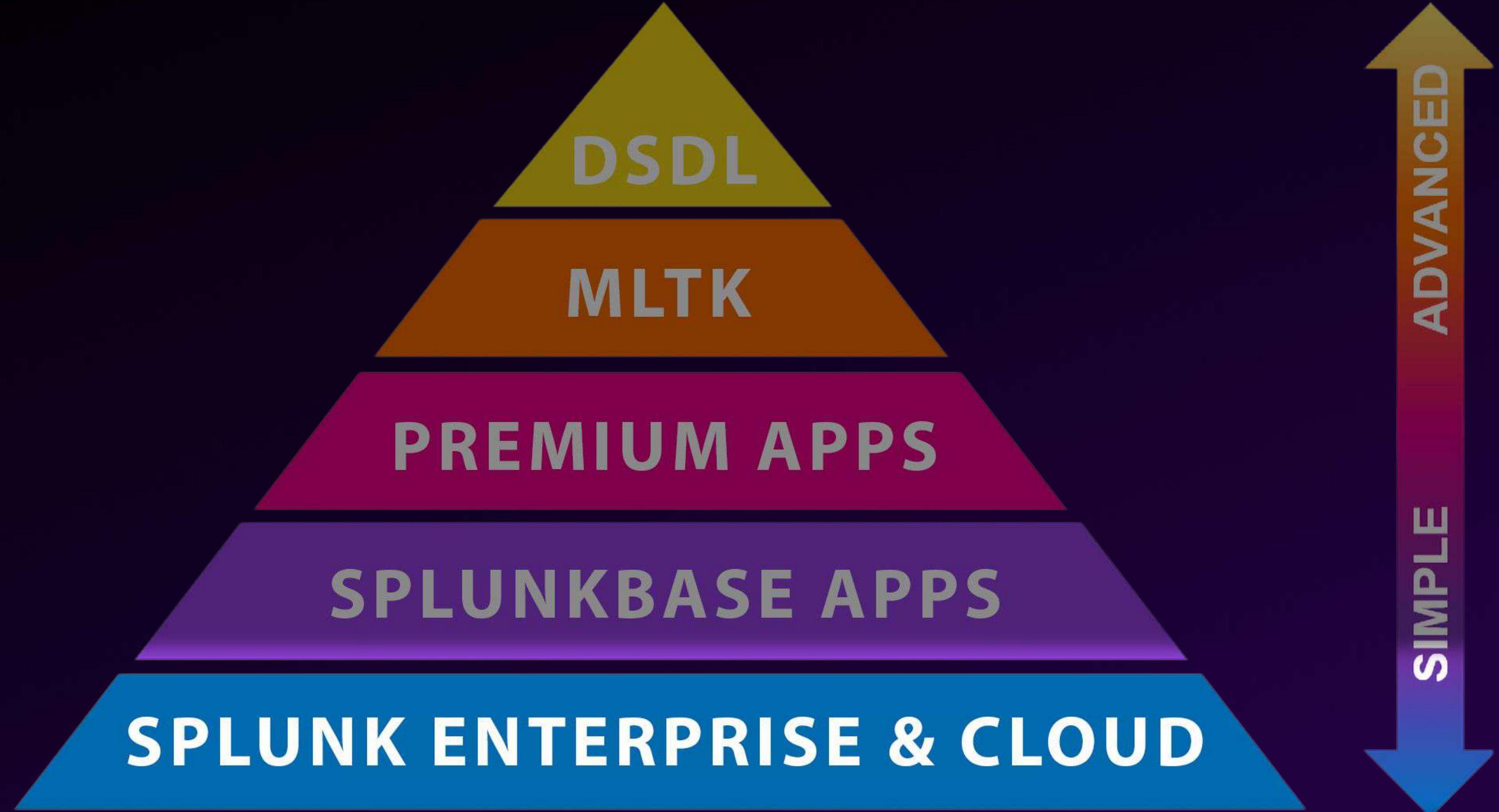
Splunk ML & AI

Where to find ML



splunk® Platform for Operational Intelligence





anomalydetection

A transforming command that identifies anomalous events by computing a probability for each event and then detecting unusually small probabilities.

<https://docs.splunk.com/Documentation/Splunk/latest/SearchReference/Anomalydetection>

The screenshot shows a web browser displaying the Splunk Search Reference documentation for the 'anomalydetection' command. The URL in the address bar is docs.splunk.com/Documentation/Splunk/9.1.3/SearchReference/Anomalydetection. The page title is 'Search Reference' under 'Splunk® Enterprise'. The left sidebar contains a navigation menu with sections like 'Search Reference', 'Search Commands' (which is currently selected), and various other search-related commands. The main content area starts with a 'Description' section, which defines the command as a transforming command that identifies anomalous events by computing a probability for each event and then detecting unusually small probabilities. It also mentions that the probability is defined as the product of the frequencies of each individual field value in the event. Below this is a note about using current Splunk machine learning (ML) tools. Further down are sections for 'Syntax' (containing the command line syntax) and 'Optional arguments' (with details on the '<method-option>' argument). A sidebar on the right provides links to 'Previously Viewed' documents and a detailed sidebar for the 'anomalydetection' command itself, including 'Description', 'Syntax', 'Usage', 'Examples', and 'See also' sections.

|cluster

The cluster command groups events together based on how similar they are to each other.

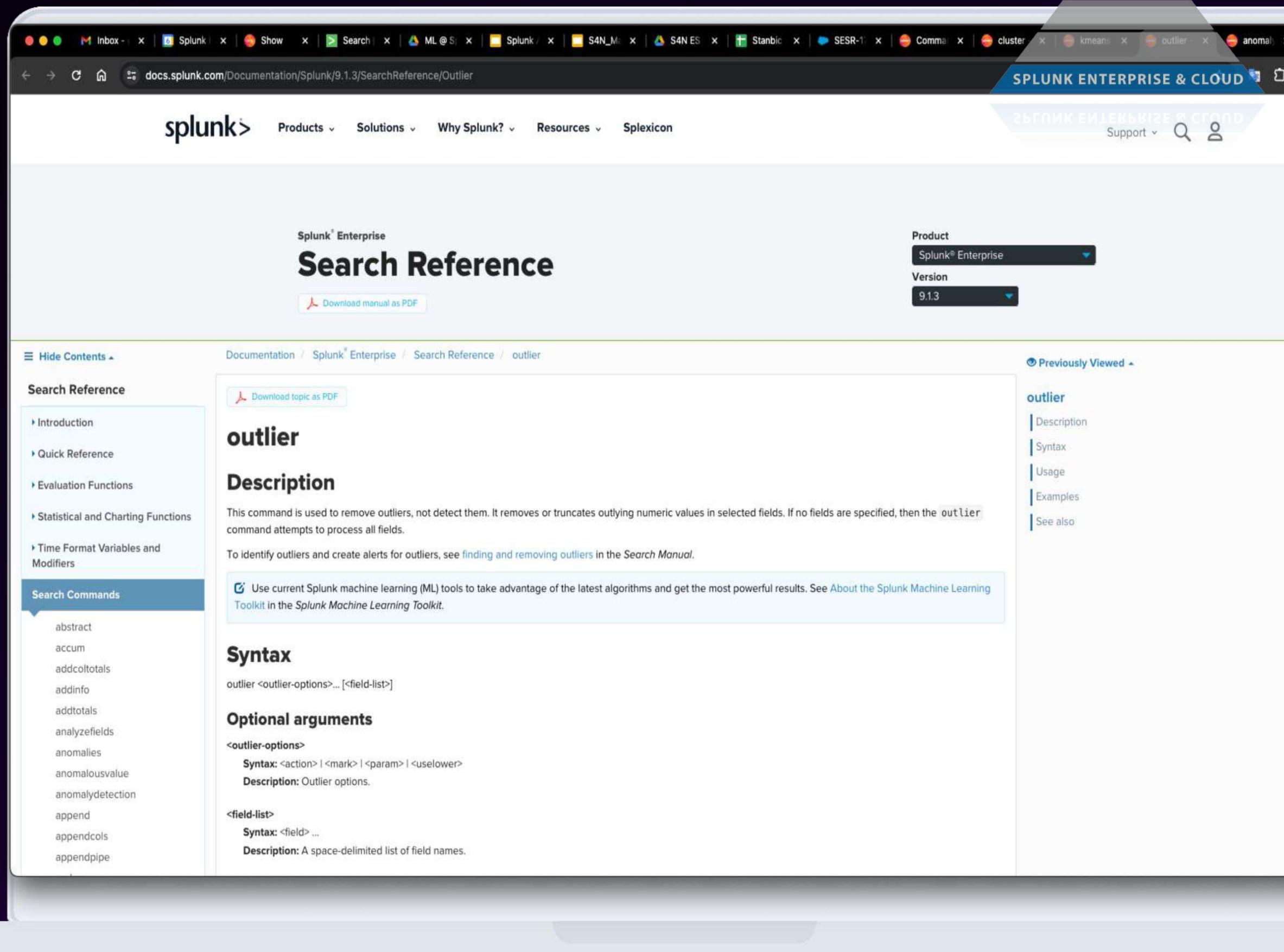
<https://docs.splunk.com/Documentation/Splunk/latest/SearchReference/Cluster>

The screenshot shows a web browser displaying the Splunk Search Reference page for the 'cluster' command. The URL in the address bar is docs.splunk.com/Documentation/Splunk/9.1.3/SearchReference/Cluster. The page title is 'Search Reference' under 'Splunk® Enterprise'. The left sidebar has a 'Search Reference' section with links to 'Introduction', 'Quick Reference', 'Evaluation Functions', 'Statistical and Charting Functions', 'Time Format Variables and Modifiers', and 'Search Commands'. The 'Search Commands' section is expanded, showing a list of commands including 'abstract', 'accum', 'addcoltotals', 'addinfo', 'addtotals', 'analyzefields', 'anomalies', 'anomalousvalue', 'anomalydetection', 'append', 'appendcols', 'appendpipe', and 'cluster'. The 'cluster' command is highlighted with a blue background. The main content area starts with a 'Description' section: 'The `cluster` command groups events together based on how similar they are to each other. Unless you specify a different field, `cluster` groups events based on the contents of the `_raw` field. The default grouping method is to break down the events into terms (`match=termlist`) and compute the vector between events. Set a higher threshold value for `t`, if you want the command to be more discriminating about which events are grouped together.' It then describes the result of the command, mentioning new fields like `countfield` and `labelfield`, and provides examples of cluster sizes from 1 to 10. Below this is a 'Syntax' section: 'cluster [slc-options]...'. Under 'Optional arguments', there is a 'slc-options' section with a 'Syntax' of '`t=<num>`' and a 'Description' stating it sets the cluster threshold. The right sidebar shows a 'Previously Viewed' list with 'cluster' selected, and a detailed sidebar for 'cluster' with sections for 'Description', 'Syntax', 'Usage', 'Examples', and 'See also'.

outlier

This command is used to remove outliers, not detect them. It removes or truncates outlying numeric values in selected fields.

<https://docs.splunk.com/Documentation/Splunk/latest/SearchReference/Outlier>



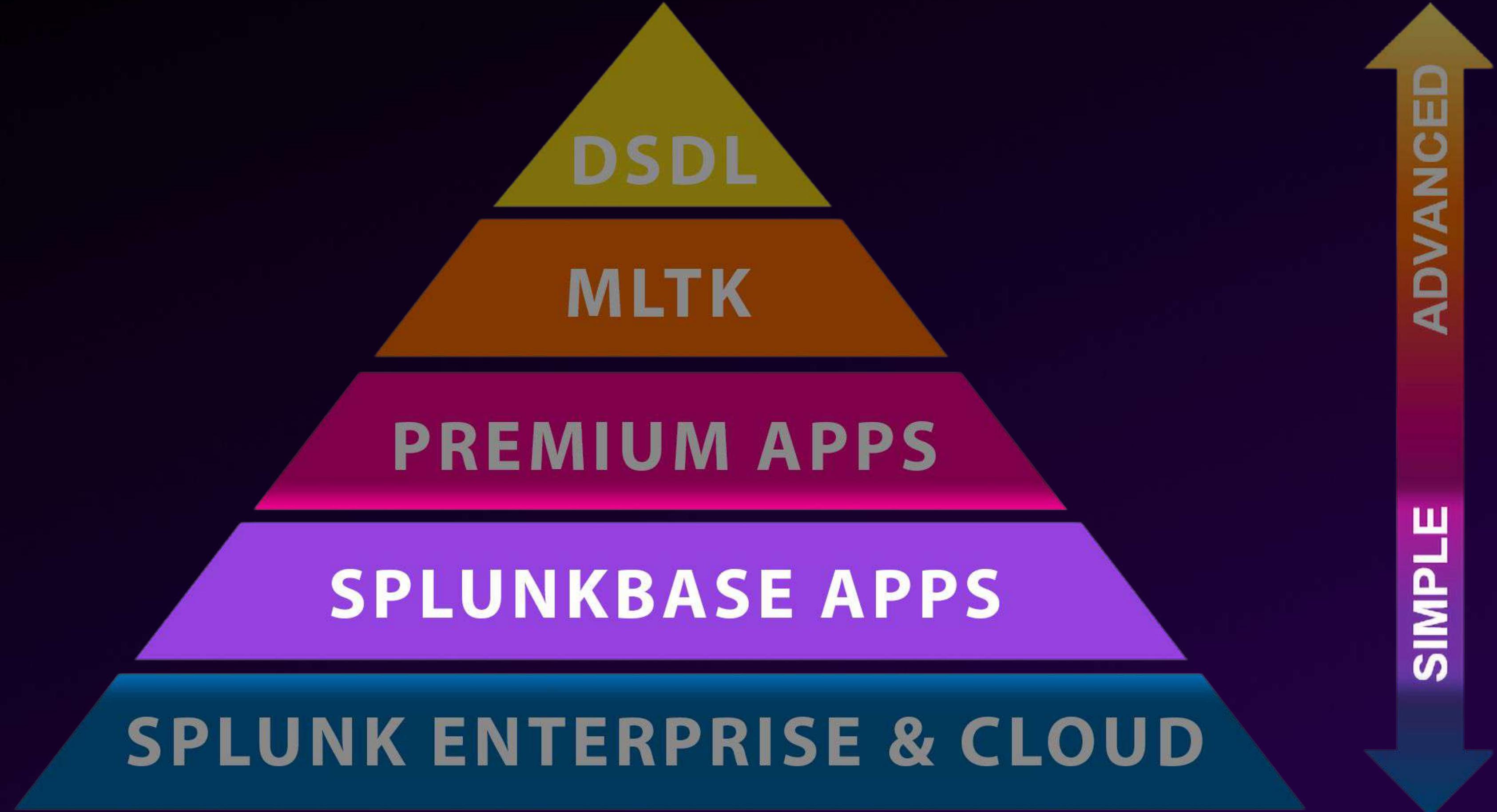
The screenshot shows a web browser displaying the Splunk Search Reference page for the 'outlier' command. The URL in the address bar is docs.splunk.com/Documentation/Splunk/9.1.3/SearchReference/Outlier. The page title is 'Search Reference' under 'Splunk® Enterprise'. The left sidebar has a 'Search Reference' section with links to Introduction, Quick Reference, Evaluation Functions, Statistical and Charting Functions, Time Format Variables and Modifiers, and a 'Search Commands' section which is expanded to show 'outlier' and other commands like abstract, accum, addcoltotals, addinfo, addtotals, analyzefields, anomalies, anomalousvalue, anomalydetection, append, appendcols, and appendpipe. The main content area starts with a 'Description' section: 'This command is used to remove outliers, not detect them. It removes or truncates outlying numeric values in selected fields. If no fields are specified, then the outlier command attempts to process all fields.' It also mentions using current Splunk machine learning (ML) tools. Below that is a 'Syntax' section with the command structure: 'outlier <outlier-options>... [<field-list>]'. There are also sections for 'Optional arguments' and '<field-list>' with their respective syntax and descriptions.

kmeans

Partitions the events into k clusters, with each cluster defined by its mean value.

<https://docs.splunk.com/Documentation/Splunk/latest/SearchReference/Kmeans>

The screenshot shows a web browser displaying the Splunk Search Reference page for the `kmeans` command. The URL in the address bar is `docs.splunk.com/Documentation/Splunk/9.1.3/SearchReference/Kmeans`. The page title is "Search Reference" under "Splunk® Enterprise". The left sidebar has a "Search Reference" section with links to "Introduction", "Quick Reference", "Evaluation Functions", "Statistical and Charting Functions", "Time Format Variables and Modifiers", and a "Search Commands" section which is currently active, showing a list of commands including `abstract`, `accum`, `addcoltotals`, `addinfo`, `addtotals`, `analyzefields`, `anomalies`, `anomalousvalue`, `anomalydetection`, `append`, `appendcols`, and `appendpipe`. The main content area starts with a "Description" section: "Partitions the events into k clusters, with each cluster defined by its mean value. Each event belongs to the cluster with the nearest mean value. Performs k-means clustering on the list of fields that you specify. If no fields are specified, performs the clustering on all numeric fields. Events in the same cluster are moved next to each other. You have the option to display the cluster number for each event." Below it is a "Syntax" section: `kmeans [kmeans-options...] [field-list]`. The "Required arguments" section states "None". The "Optional arguments" section includes a "field-list" section with a "Syntax" of `<field> ...`, a "Description" of "Specify a space separated list of the exact fields to use for the join.", and a "Default" of "If no fields are specified, uses all numerical fields that are common to both result sets. Skips events with non-numerical fields.". Finally, there is a "kmeans-options" section with a "Syntax" of `<reps> | <iters> | <t> | <k> | <cnumfield> | <distype> | <showcentroid>` and a "Description" of "Options for the `kmeans` command."



- Smart Alerts Assistant for Splunk (beta):

Smart Alerts Assistant for Splunk captures user feedback to improve prioritization and uses machine learning to output a recommended, stack-ranked priority for each alert.

- Splunk App for Anomaly Detection:

Finds anomalies in time series datasets and provides an end-to-end workflow to manage and operationalize anomaly detection tasks.

COLLECTION

Getting Started with AI

These Machine Learning and AI powered apps and assistants give you the power of Machine Learning for common use cases with just a couple of clicks.

Intelligent Experiences

These ML-powered apps and assistants give you the power of ML for common use cases with just a couple of clicks.

A screenshot of the Smart Alerts Assistant for Splunk (beta) app card. The card features a purple rounded square icon with a white 'sa' logo. To the right of the icon, the app's name is displayed in large, bold, black font, followed by '(beta)' in parentheses. Below the name, the developer, 'By Splunk Works', is shown in a smaller, gray font. A detailed description of the app is provided below the developer information, starting with an asterisk and ending with '...'. At the bottom of the card, there are two sections: 'PLATFORM' with 'Splunk Enterprise' listed and 'RATING' with a five-star rating icon.

The image shows a dark-themed mobile app store listing for the "Splunk App for Anomaly Detection". It features a blue square icon with a white "AD" logo. The title "Splunk App for Anomaly Detection" is displayed in large white font, followed by "By Splunk LLC". A detailed description below reads: "The Splunk App for Anomaly Detection finds anomalies in time series datasets and provides an end-to-end workflow t...". At the bottom, it shows the platform "Splunk Enterprise, Splunk Cloud" and a rating of 5 stars based on 5 reviews.

An app card for the "Splunk AI Assistant for SPL (Cloud Version)" app. The card features a red square icon with the letters "SAIA" in white. The title "Splunk AI Assistant for SPL (Cloud Version)" is displayed in large, bold, black font. Below the title, it says "By Splunk LLC". A detailed description of the app's function is provided: "Splunk AI Assistant for SPL offers bi-directional translation between natural language (NL) and Splunk Search...". The card also includes sections for "PLATFORM" (Splunk Enterprise, Splunk Cloud) and "RATING" (4 stars from 2 reviews). At the bottom, there is a yellow headphones icon followed by the text "SPLUNK SUPPORTED APP".

AI Assistant for SPL

(Cloud only)

Increased efficiency
for users

- ▶ Get your job done faster using natural language.
- ▶ Drill down to deeper insights.
- ▶ Accelerate your learning journey.

The screenshot shows a user interface for an AI-powered search tool. At the top, there are three buttons: "Write SPL" (highlighted in blue), "Explain SPL", and "Tell me about...". Below this, a user profile icon and a question are displayed: "What disk has the least amount of freespace in WinHostMon?". A red box highlights the AI response area. The AI replies: "To answer this question, we will use the following SPL query:" followed by a complex SPL command. At the bottom of this box are "Explain SPL", "Copy", and "Open in Search" buttons. Below the main box, a section titled "This query will..." lists four steps: 1. Filter the WinHostMon data from the specified index. 2. Calculate the free space percentage, total space in GB, and free space in GB for each disk. 3. Group the results by host and find the minimum values for each of these fields. 4. Display the results in a table format with the host, minimum total space in GB, minimum free space in GB, and minimum free space percentage.

What disk has the least amount of freespace in WinHostMon?

To answer this question, we will use the following SPL query:

```
index="[" sourcetype="WinHostMon" Type=Disk
| eval FreeSpacePct=round(FreeSpaceKB/TotalSpaceKB*100)
| eval TotalSpaceGB=round(TotalSpaceKB/1024/1024)
| eval FreeSpaceGB=round(FreeSpaceKB/1024/1024)
| stats min(FreeSpacePct) as MinFreeSpacePct, min(TotalSpaceGB) as MinTotalSpaceGB, min(FreeSpaceGB) as
MinFreeSpaceGB by host
| table host, MinTotalSpaceGB, MinFreeSpaceGB, MinFreeSpacePct
```

This query will:

1. Filter the WinHostMon data from the specified index.
2. Calculate the free space percentage, total space in GB, and free space in GB for each disk.
3. Group the results by host and find the minimum values for each of these fields.
4. Display the results in a table format with the host, minimum total space in GB, minimum free space in GB, and minimum free space percentage.



Splunk AI Assistants everywhere

SPLUNKBASE APPS

Our goal: Make everyone a Splunk expert!

AI Assistant for SPL

The screenshot shows a user interface for building Splunk queries. A modal window is open, prompting the user to "Write SPL to detect the creation of a new local user in windows". Below the prompt, it says: "To detect the creation of a new local user in Windows, we can monitor the Windows Security log for new user account creation events. The SPL query for this would be:"

```
index=<index> source=*WinEventLog*:Security AND EventID=4624 | dedup session_id | table _time, host, User, TargetUserName, Subject
```

Below the query, there's an "Explain SPL" button and a note about the fields returned:

This query will return a table with the following fields:
- _time: The time of the event
- host: The host where the event occurred
- User: The user who created the new user account
- Name: The name of the new user account

AI Assistant in Enterprise Security

The screenshot shows the Splunk Cloud interface with the AI Assistant integrated. A specific alert titled "Geographically Improbable Access Detected" is highlighted. The details show a connection from "workstation-478.internal.domain" to "server-932.internal.domain" over SMB protocol. The "Event Information" section lists "Files accessed and transferred" as "\\\server-932\\share\\payroll" and "\\\server-932\\share\\config". The "Additional Context" section provides context about the source host being a regular employee named "jdoe".

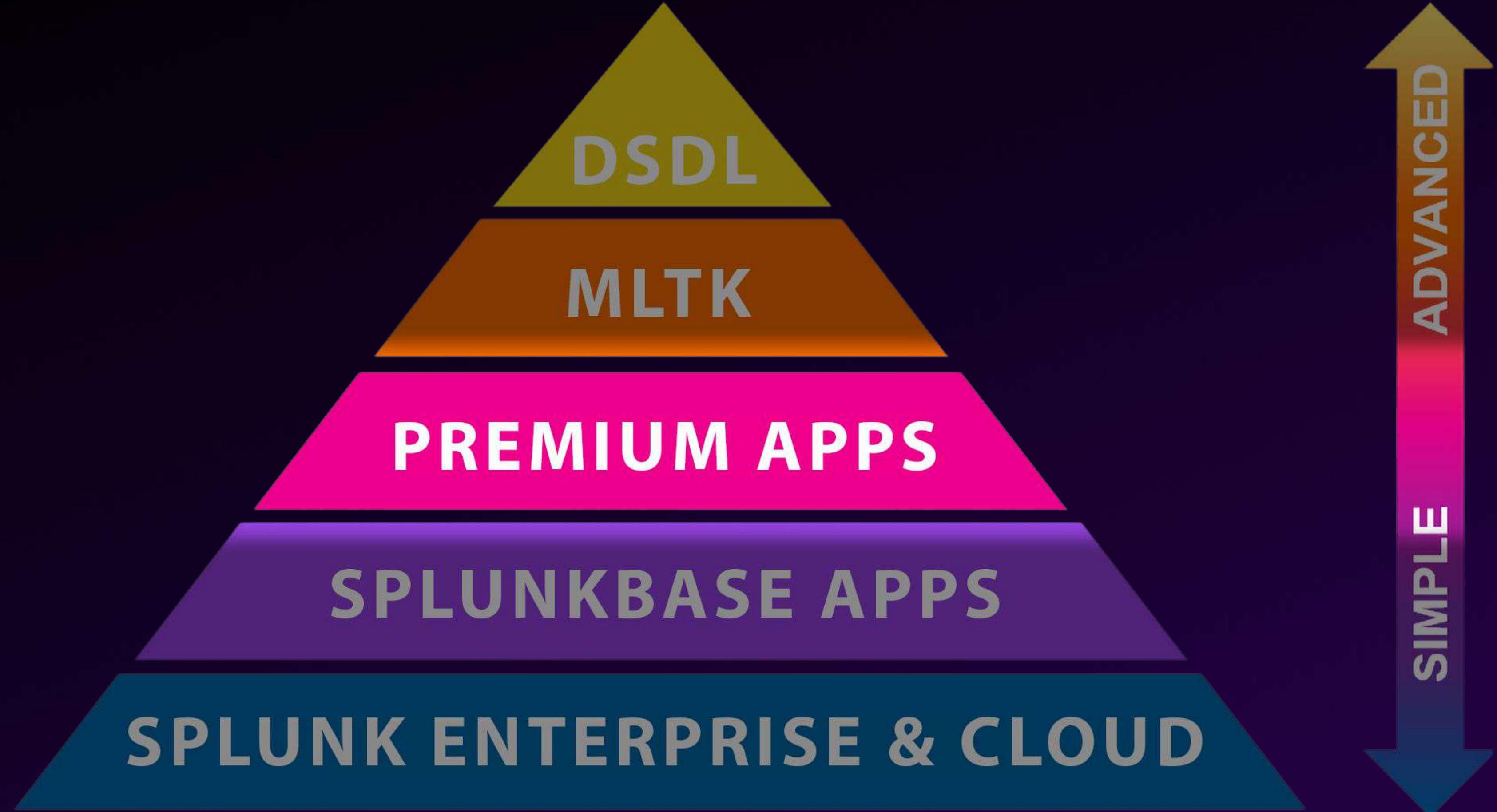
AI Assistant in Observability Cloud

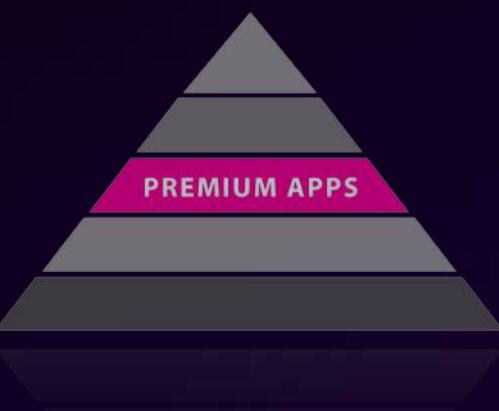
The screenshot shows the Splunk Observability Cloud interface with the AI Assistant. On the left, a dependency map for the "frontend" service shows various components and their latencies. On the right, a chat interface is shown with a message from "AZ" asking why the "frontend" is having issues. The AI Assistant responds with "Here are some possible reasons: it seems like frontend is having issues downstream dependency "paymentservice" experiencing problems." Below the map, a message says "Checking the health map like the one above".

Available now

Private Preview

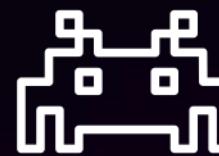
Available now





Splunk Enterprise Security

with ML-Powered Content Updates from the Splunk Machine Learning for Security Team



Study Threats

Identify emerging threats and understand how they operate



Create Datasets

Collect data and use Splunk to parse the data and identify patterns that can be used to detect the threat



Build ML-Powered Detections

Build a model based on data in order to make predictions or decisions; enable systems to learn from data, identify patterns, and make decisions with minimal human intervention; and craft rules or queries designed to identify specific activity associated with threats



Test Detections

Run queries against a dataset that simulates attacker behavior to improve accuracy and reduce false positives

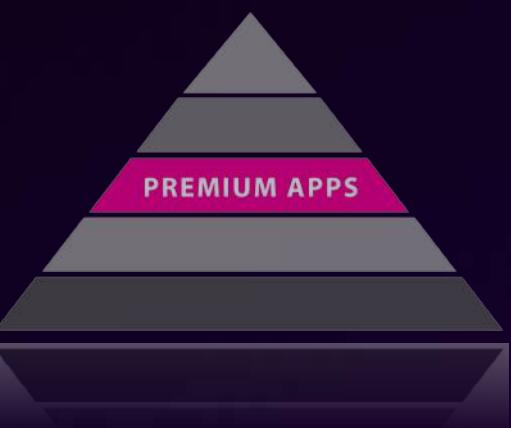


Release

Package detections to deliver timely and effective protections against emerging threats to Splunk customers

ESCU

Enterprise Security Content Updates



Security Content

- Detections
- Analytic Stories
- Playbooks
- Blog
- About
-

machine learning

78 Result(s) found

[Splunk Command and Scripting Interpreter](#)

[Risky SPL MLTK](#)

Try in Splunk Security Cloud Description This detection utilizes machine learning model named "risky_command_abuse" trained from "Splunk Command and Scripting..."

[Potentially malicious code on commandline](#)

Try in Splunk Security Cloud Description The following analytic uses a pretrained machine learning text classifier to detect potentially malicious...

[Azure Active Directory High Risk Sign-in](#)

Try in Splunk Security Cloud Description The following analytic triggers on a high risk sign-in against Azure Active Directory identified...

[Modification Of Wallpaper](#)

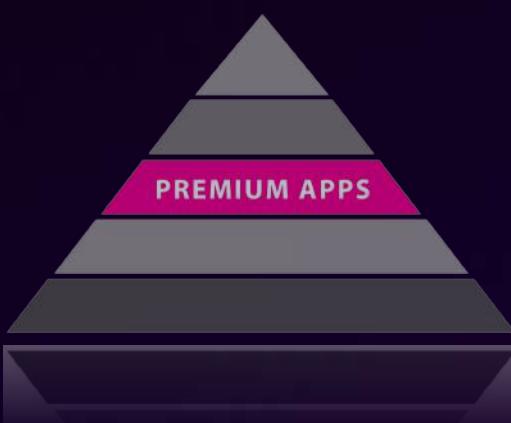
Try in Splunk Security Cloud Description This analytic identifies suspicious modification of registry to deface or change the wallpaper of...

Splunk ES Content Update

The Splunk ES Content Update (ESCU) app delivers pre-packaged Security Content. ESCU provides regular Security Content updates to help security practitioners address ongoing time-sensitive threats, attack methods, and other security issues. Security Content consists of tactics,...

Built by [Splunk Inc.](#)

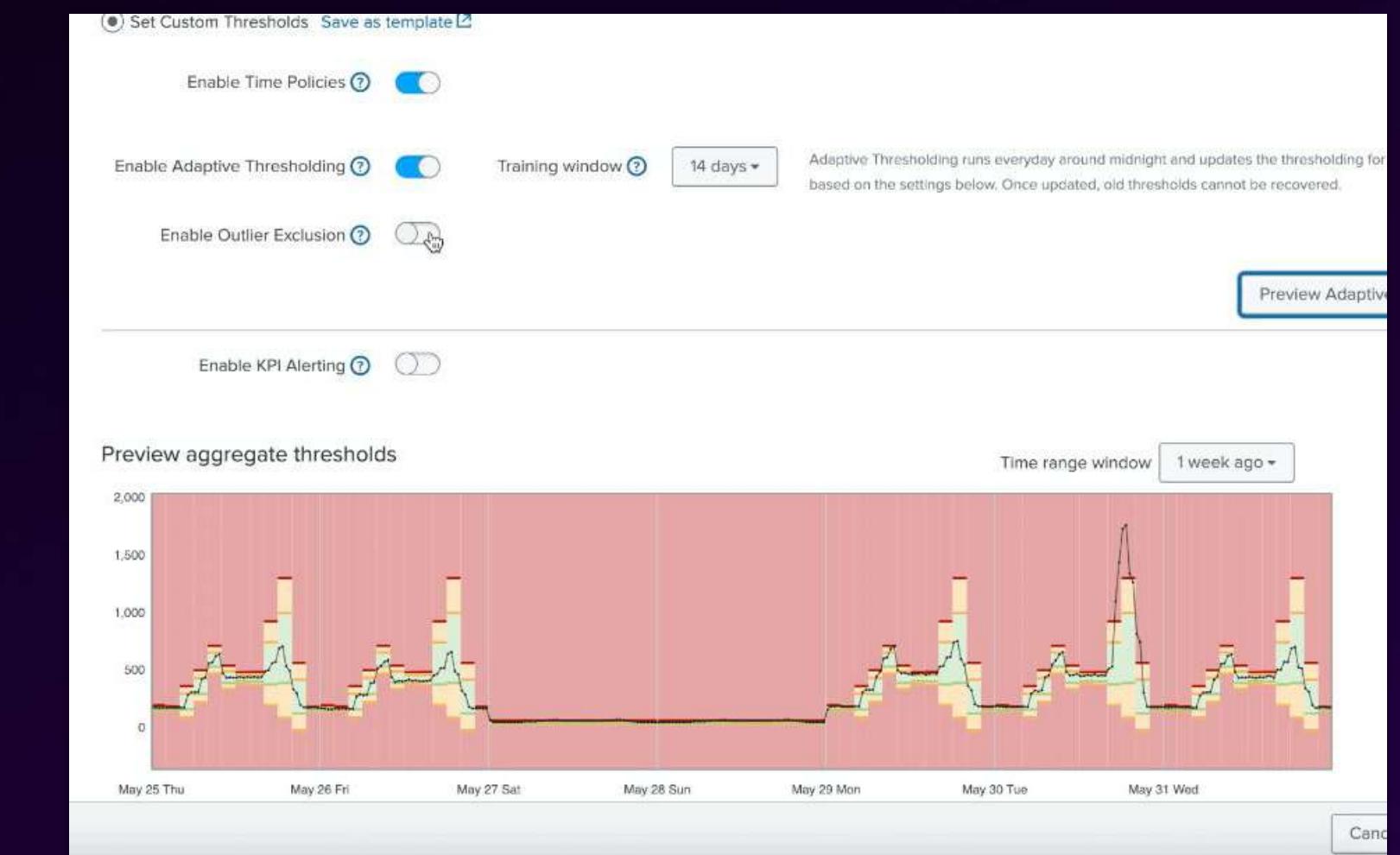
Welcome to Enterprise Security Content Updates (ESCU), brought to you by the Splunk Security Research Team!



Splunk IT Service Intelligence (ITSI)

- Splunk's **AIOps** Solution
- Splunk ITSI applies machine learning to **proactively prevent outages** by correlating and reducing alerts, monitoring service health, and streamlining incident management.

- Clustering & aggregation to reduce alert noise
- Adaptive (dynamic) thresholds incorporate seasonality
- Anomaly and outlier detection
- Actionable additional context
- Assisted root cause investigation
- Predict service health to prevent outages



New updates!

- Outlier Exclusion in Adaptive Thresholds
- ML-Assisted Thresholding (*Preview*)

07 MLOps Incident ▾

MLOps Dashboard for running an ML environment

Global Time Range

Last 24 hours

splunk> MLOps Dashboard

83

Overall Health Score

55

Predictive Health Score

Sales KPIs

Predicted Sales vs. Actual Sales

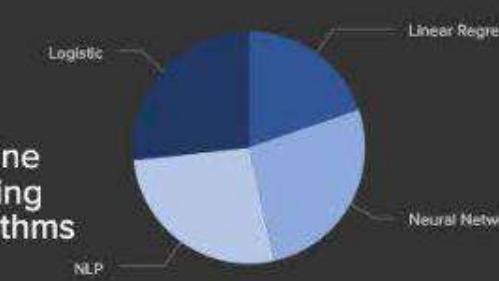


Machine Learning

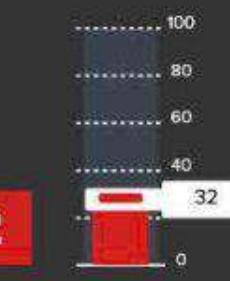
ML Models

Trained Models **46**Retrained Today **22**

Machine Learning Algorithms



ML Scores

MSE **2,115**RMSE **46.0**R² **0.32**

Inventory Deficits

182

Overstocks

0

Applications

64**97****94****97****97**

Security

99

External Threats

15

Internal Threats

4

Data Collection

Outliers

4

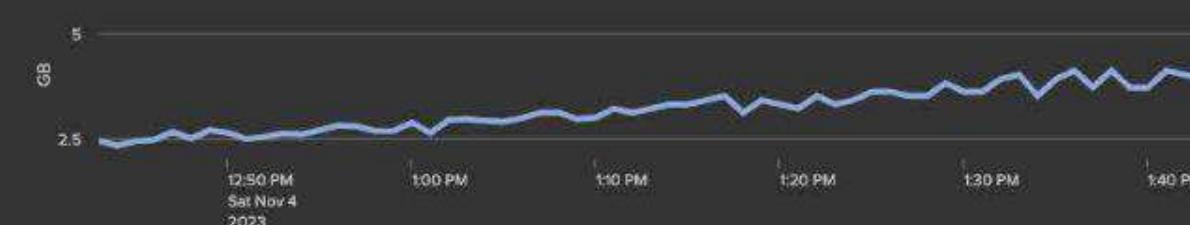
Missing Events

886

Missing Fields

65,036

Data Volume last hour

3.90GB

Infrastructure

Cloud



Network



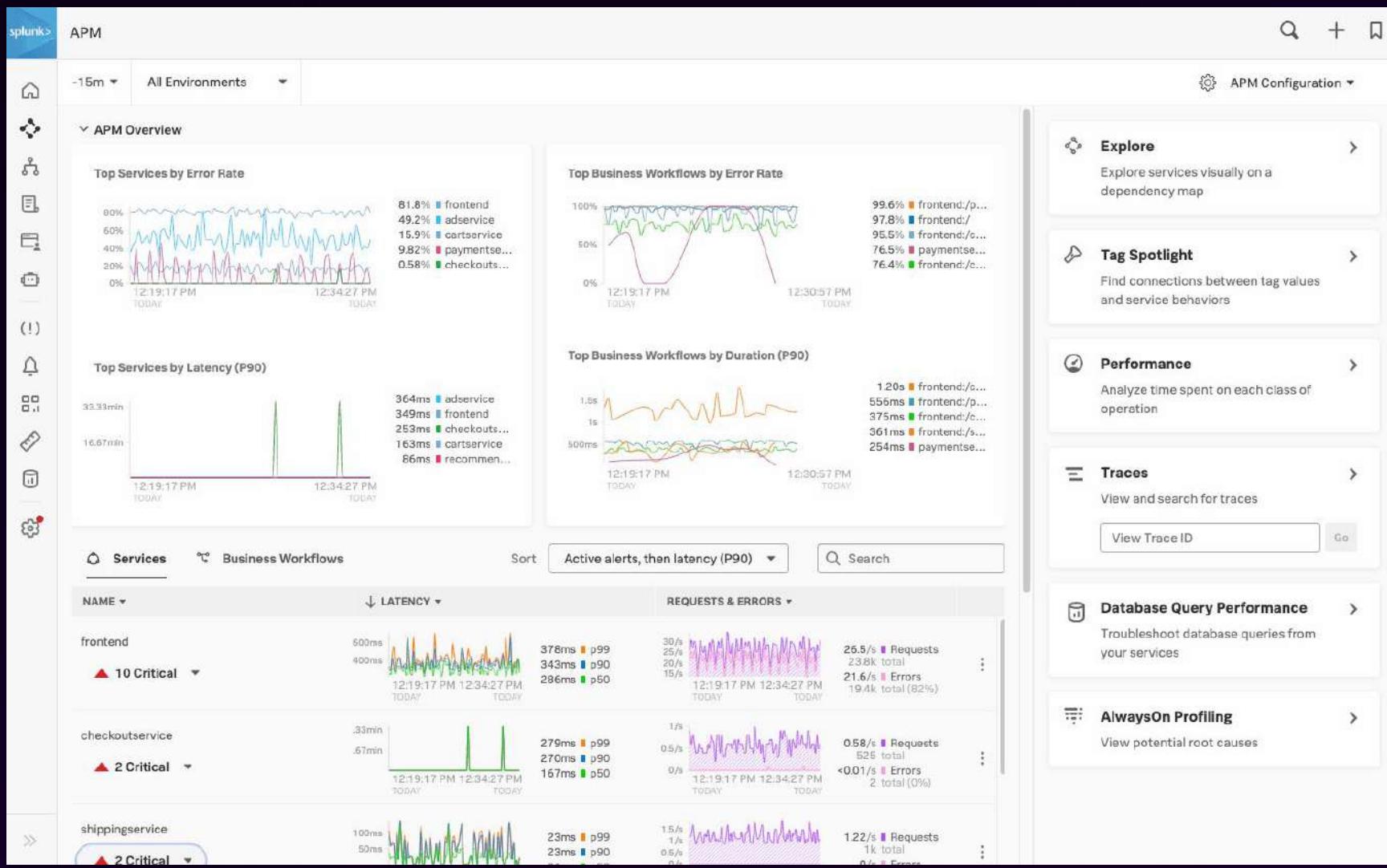
On-premise





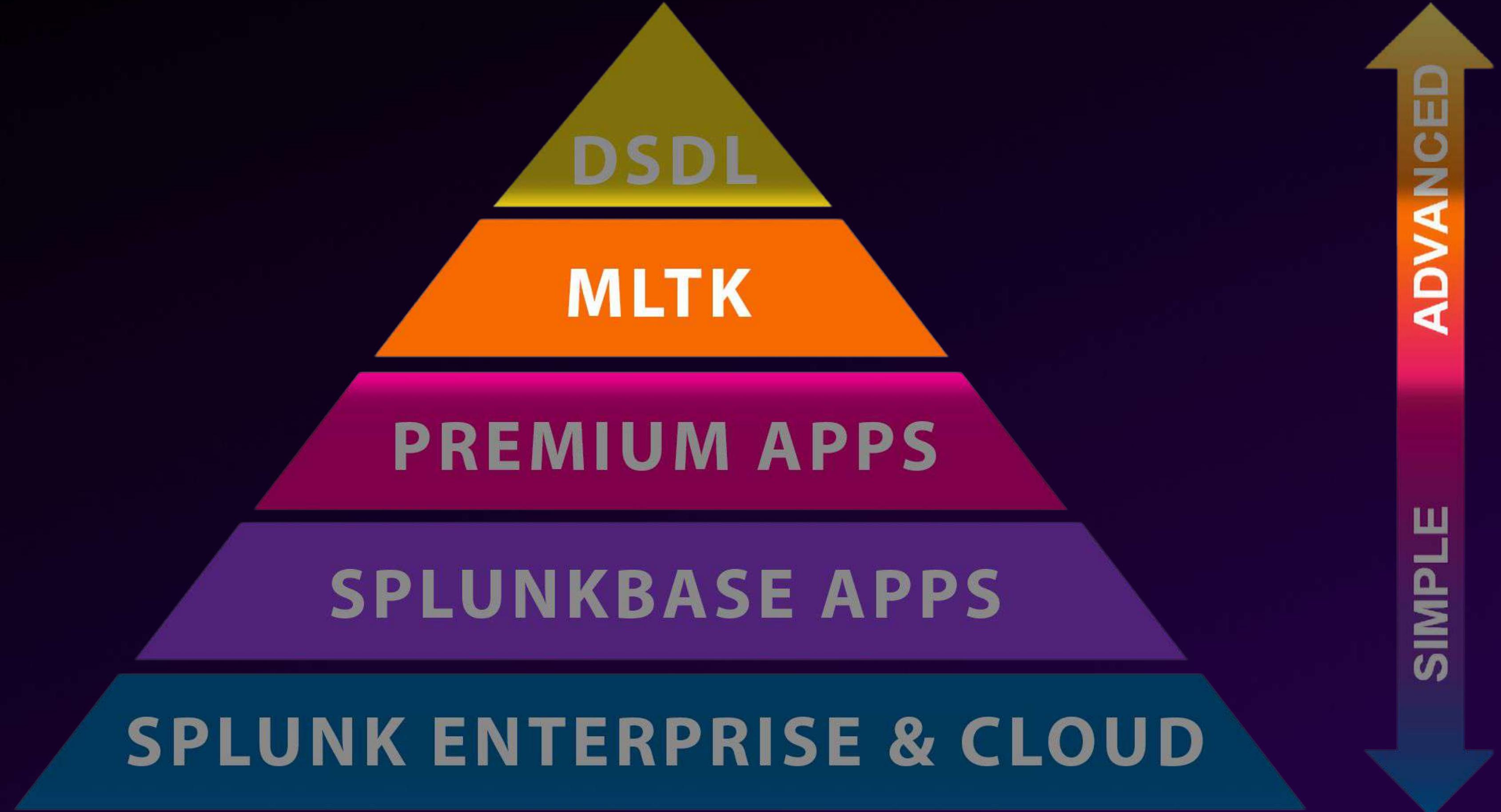
Application and Infrastructure Monitoring: Autodetect

More accurate and efficient alerting for your infrastructure and services



Use machine learning to **improve accuracy** and **reduce manual effort** across infrastructure and service alerting

- Establish performance baselines across every service
- Automate alerting by receiving recommendations for the biggest spikes in latency, errors, and resource utilization
- Easily customize alert thresholds and subscribe to notifications for specific services or teams



Python for Scientific Computing (PSC)

Extend the Splunk Platform Python runtime with AI specific libraries



Run complex AI/ML based analytics

in the Splunk Platform, with a broad range of supported Open Source Python libraries



Splunk Machine Learning Toolkit (MLTK)

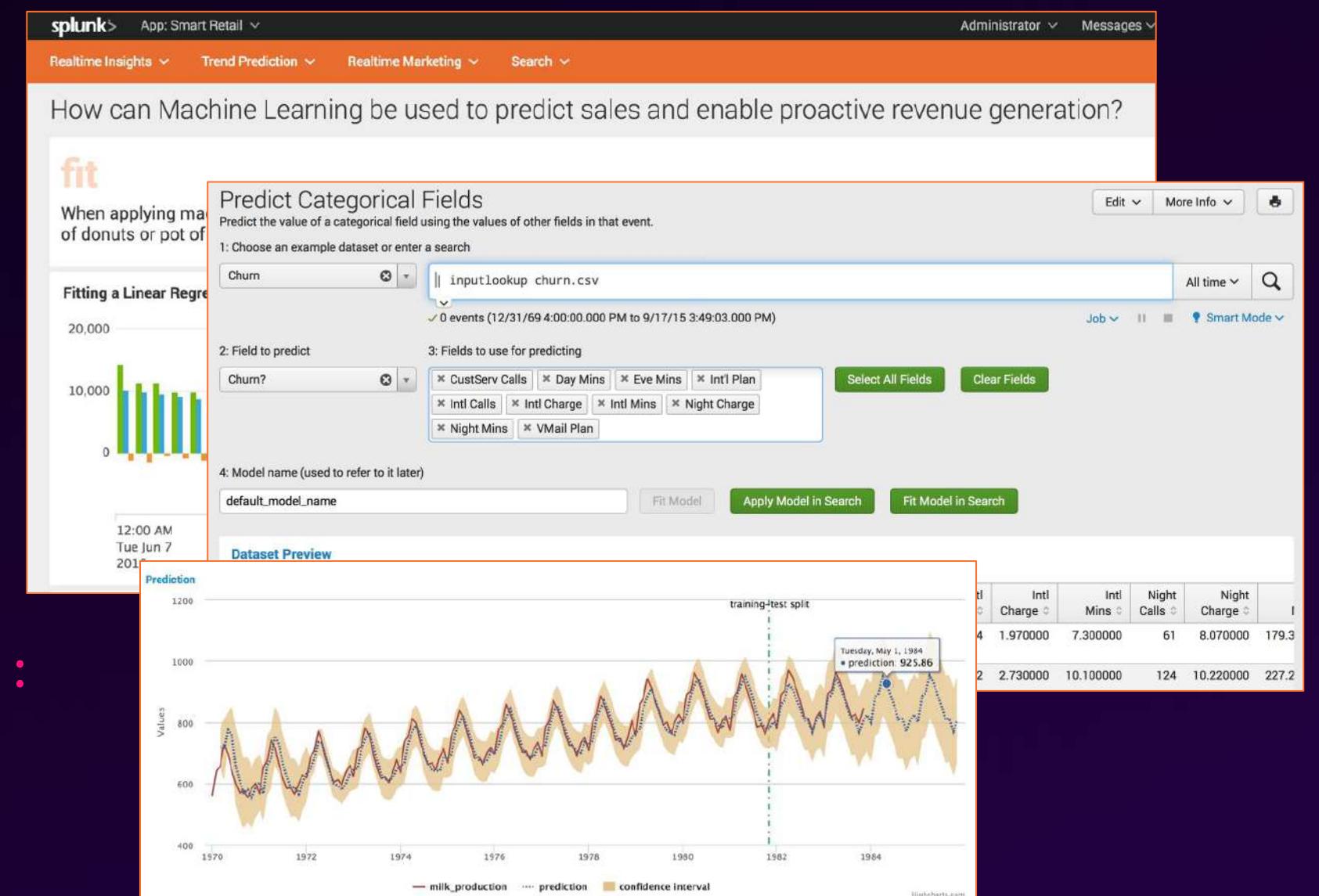
Extends Splunk platform functions and provides a guided modeling environment

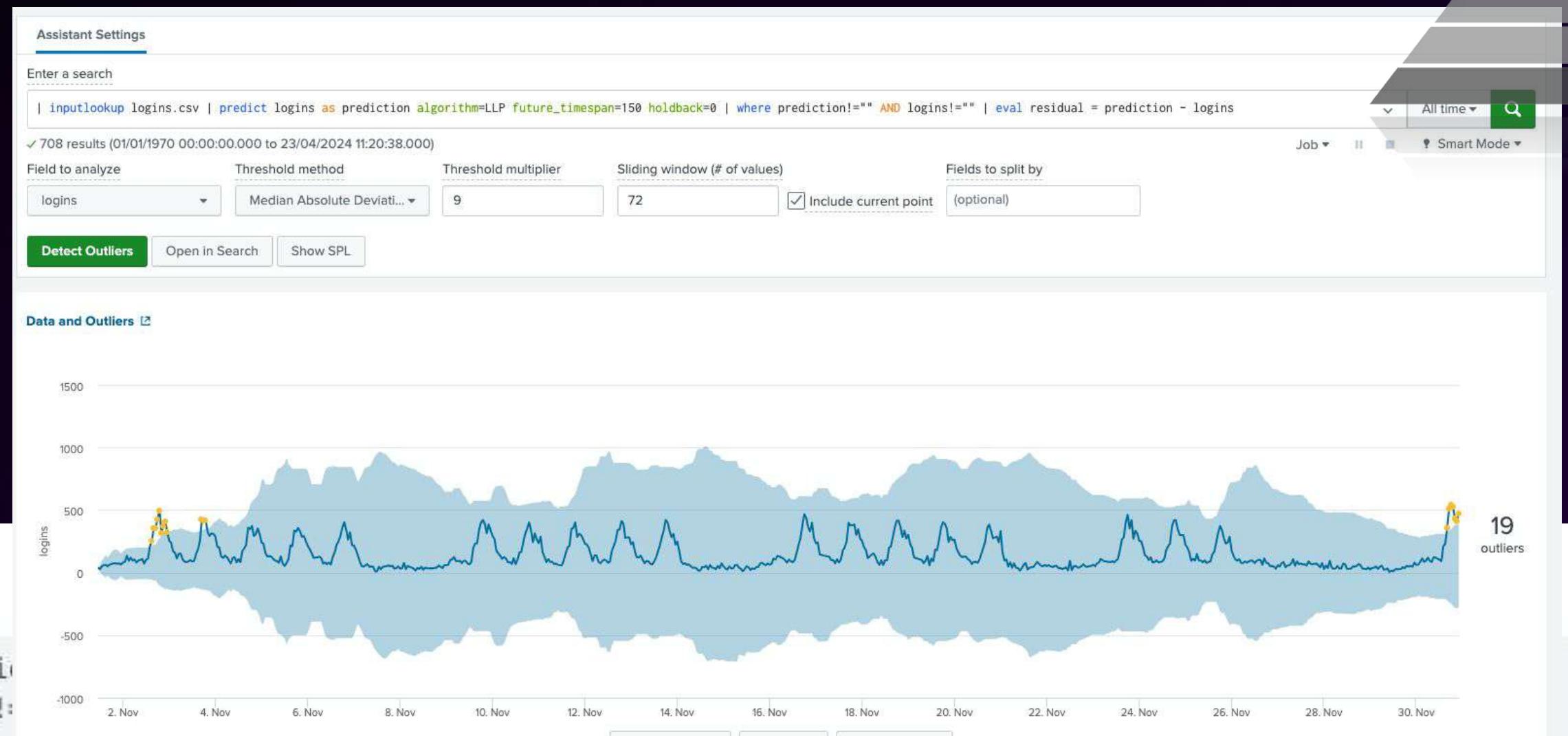
Built for the Citizen Data Scientist

- **Experiments and Assistants:**
 - Guided model building, testing, and deployment
 - for common objectives
- **Algorithms:**
 - 80+ standard algorithms

Extensible to operationalize any use case

- **Python for Scientific Computing Library:**
 - Access to 300+ open source algorithms
- **Splunk App for Data Science and Deep Learning (DSDL) :**
 - Supports NN and GPU accelerated machine learning
- **ML-SPL API:**
 - Import any open-source or proprietary algorithm



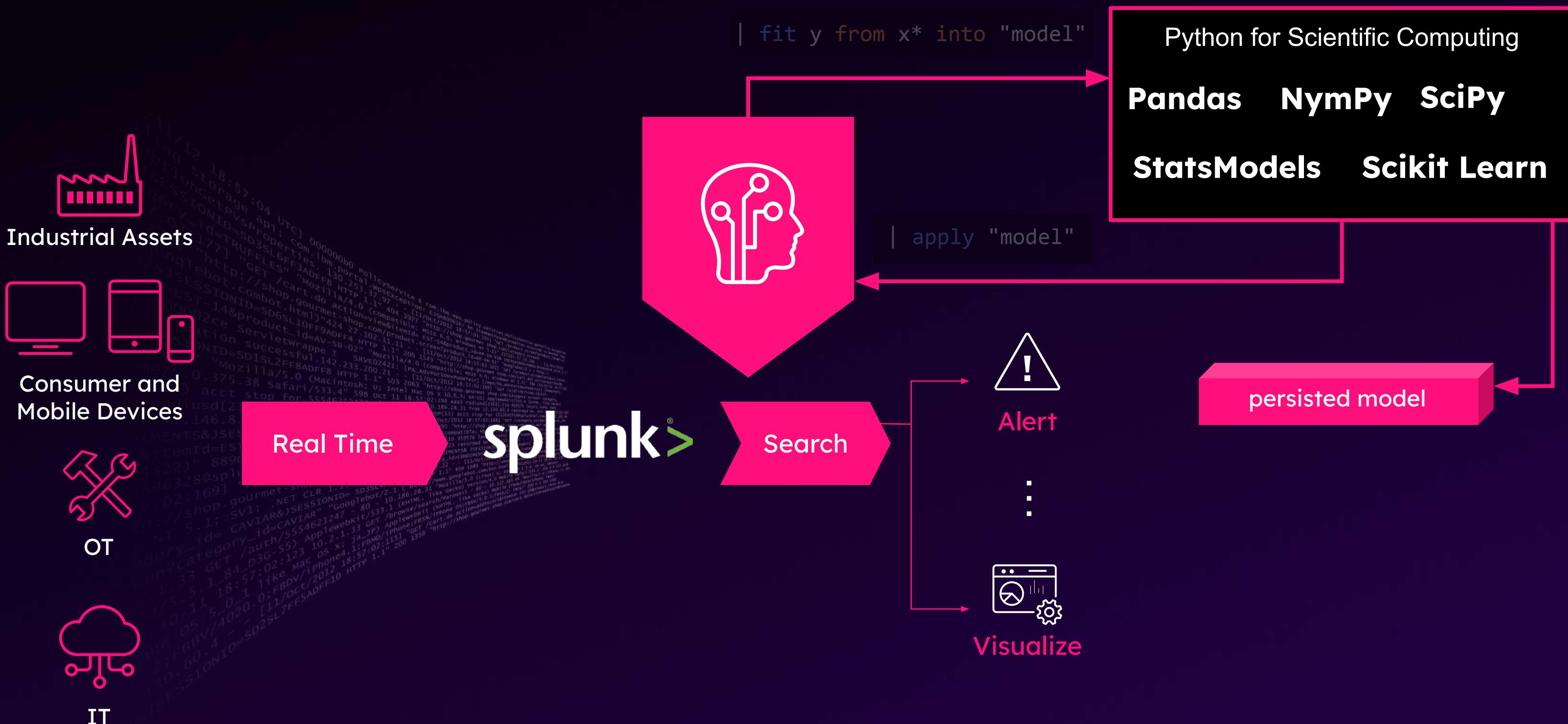


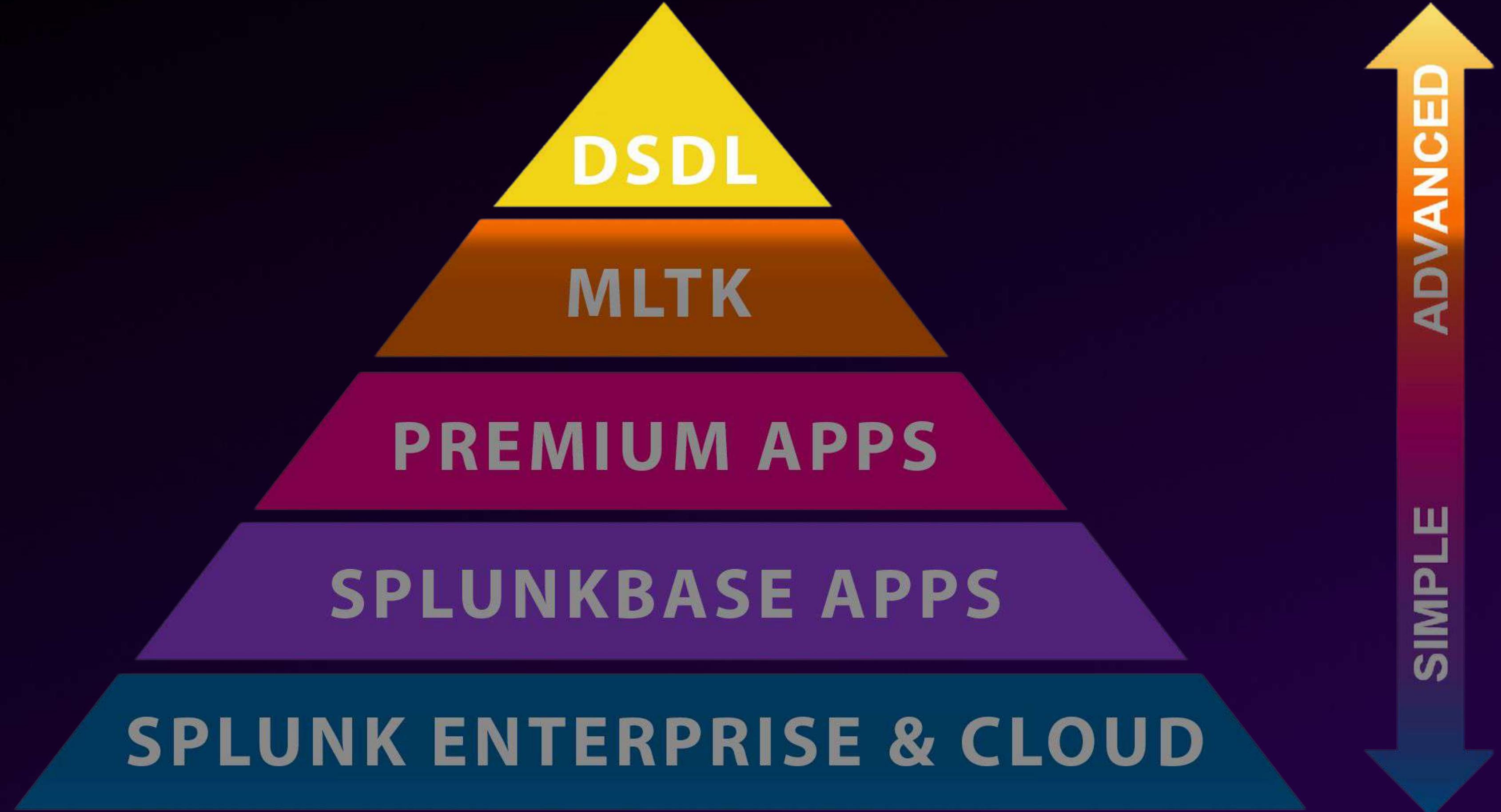
Plot the outliers

```
| inputlookup logins.csv | predict logins as prediction
future_timespan=150 holdback=0 | where prediction!=""
eval residual = prediction - logins

| streamstats window=72 current=true median("logins") as median          // calculate the median value using a sliding window
| eval absDev=(abs('logins'-median))                                         // calculate the absolute deviation of each value from the median
| streamstats window=72 current=true median(absDev) as medianAbsDev         // use the same sliding window to compute the median absolute deviation
| eval lowerBound=(median-medianAbsDev*exact(9)), upperBound=              // calculate the bounds as a multiple of the median absolute deviation
(median+medianAbsDev*exact(9))
| eval isOutlier=if('logins' < lowerBound OR 'logins' > upperBound, 1, 0)   // mark values outside the bounds as outliers
| fields _time, "logins", lowerBound, upperBound, isOutlier, *               // format the columns to be in the order expected by the Outliers Plot visualization
```

Model Longevity





Splunk App for Data Science and Deep Learning

aka. Deep Learning Toolkit for Splunk (DLTK)

Built for Data Scientists

- **Frameworks:**

Freely available app for advanced data science projects using any open-source AI frameworks contains PyTorch, Tensorflow 2.0, SpaCey, Jupyter Notebook, & lot more

- **Code Examples:**

Guided model building, testing, and deployment of deep learning frameworks

- **GPU Support:**

Speed up your data science projects with GPU accelerated containers

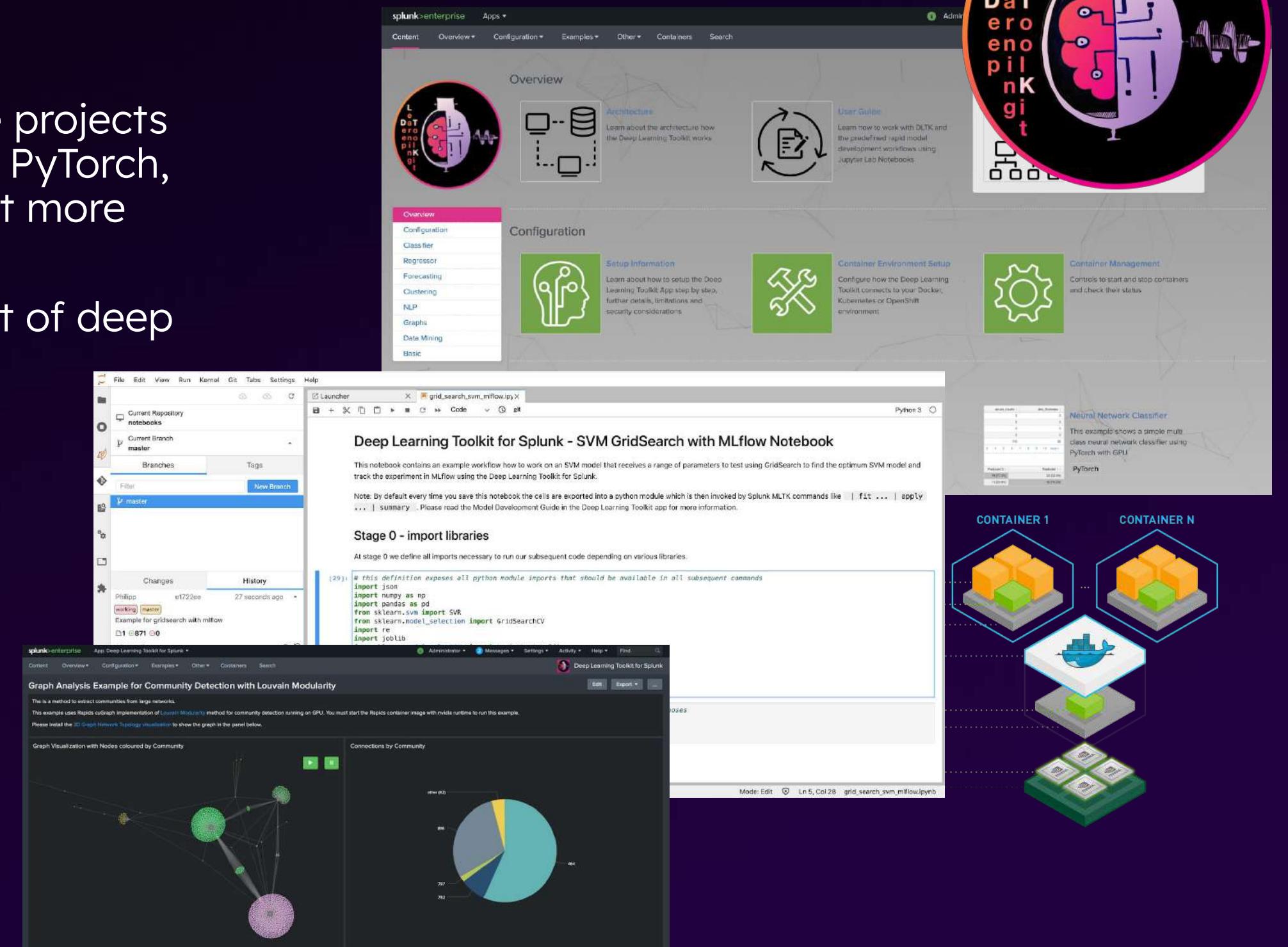
- **K8s Support:**

Scalable and HA with K8s deployment

- **Open source for customization:**

<https://github.com/splunk/splunk-mltk-container-docker>

This is now becoming a "must-have" for every citizen data scientists



DLTK for Splunk



Industrial Assets



Consumer and
Mobile Devices



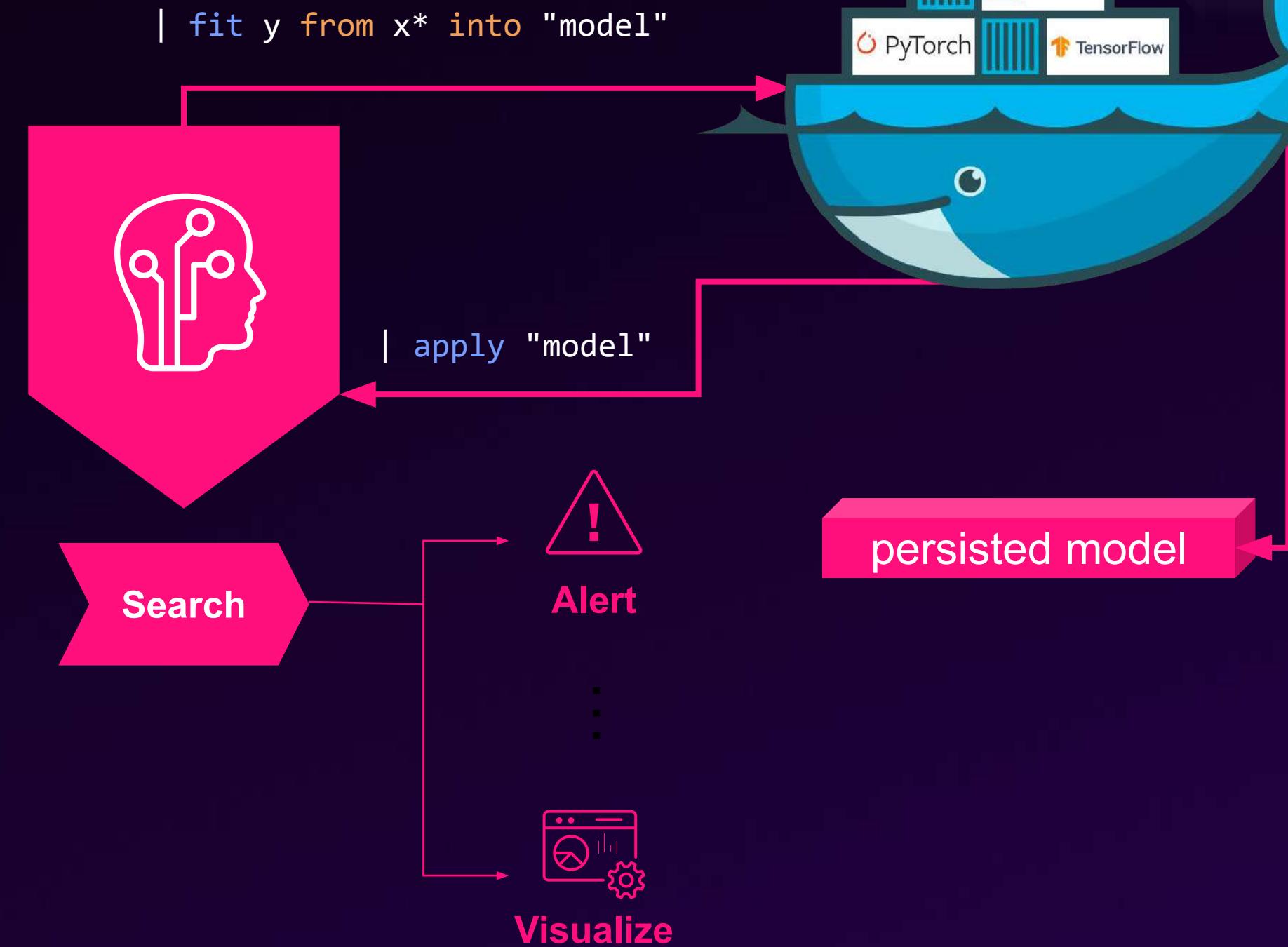
OT



IT

Real Time

splunk®



Road to SUCCESS

What steps will help you to make your ML project a success.

Includes:

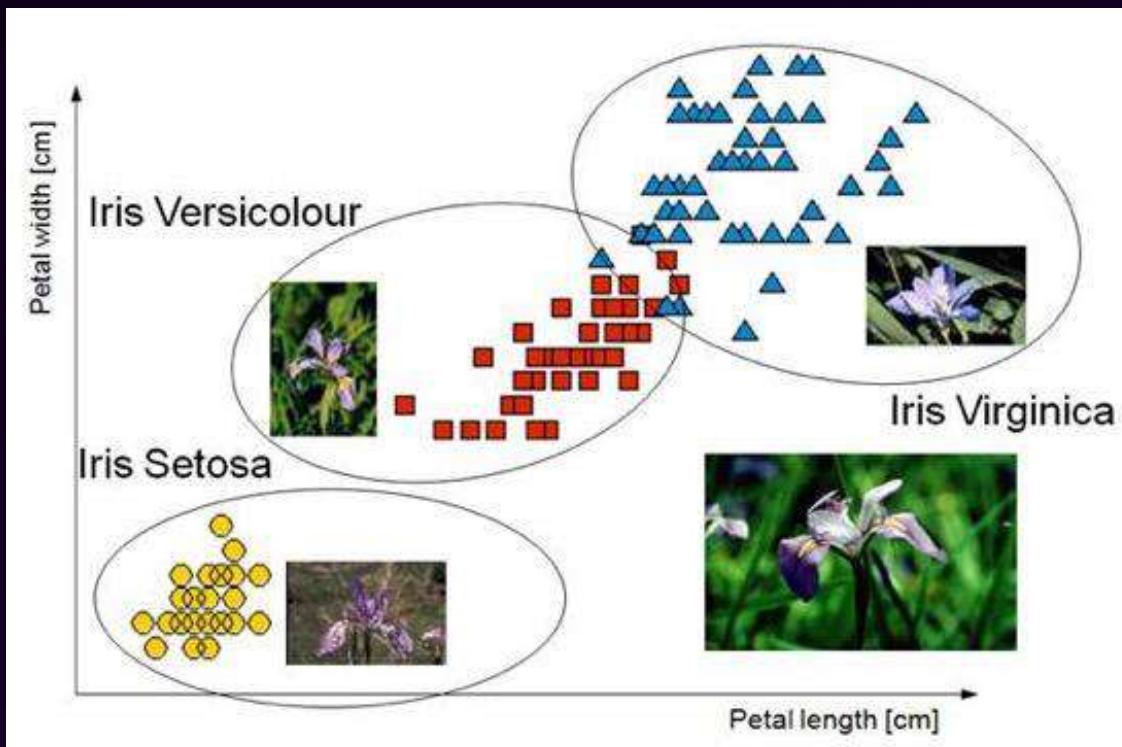
- Understanding your data
- Cleaning/munging your data
- Operationalizing the data set



Types of Machine Learning

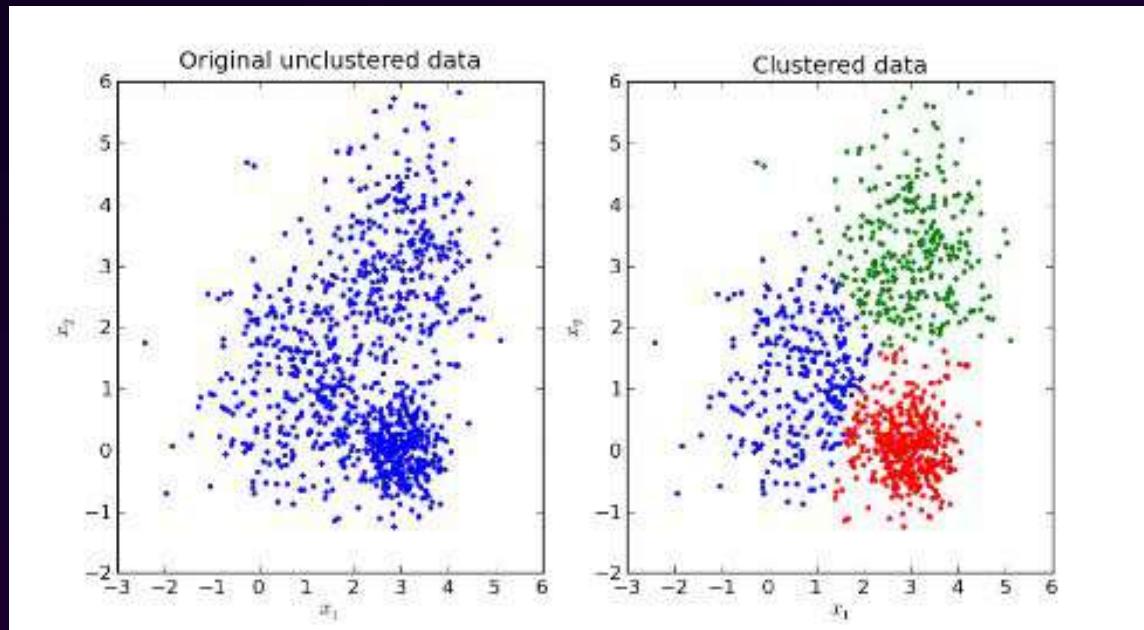
Supervised Learning (labeled data)

- ▶ Regression
- ▶ Classification



Unsupervised Learning (unlabeled data)

- ▶ Clustering
- ▶ Anomaly Detection



Mixed Models (with reinforcement or feedback)

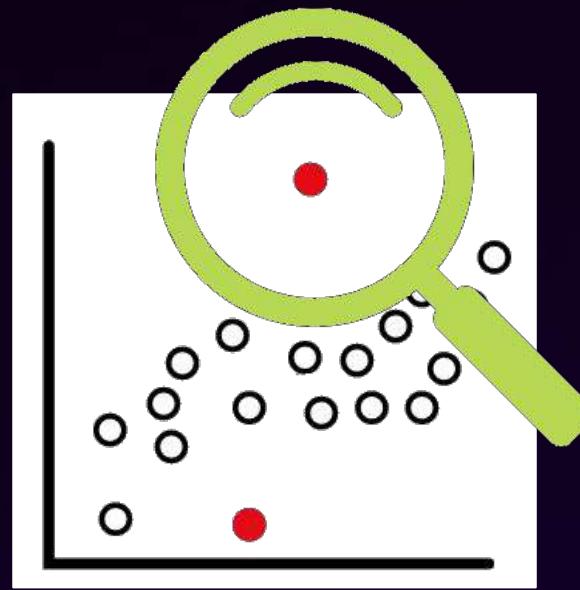
- ▶ Human in the Loop
- ▶ Autonomous Systems



Where to start

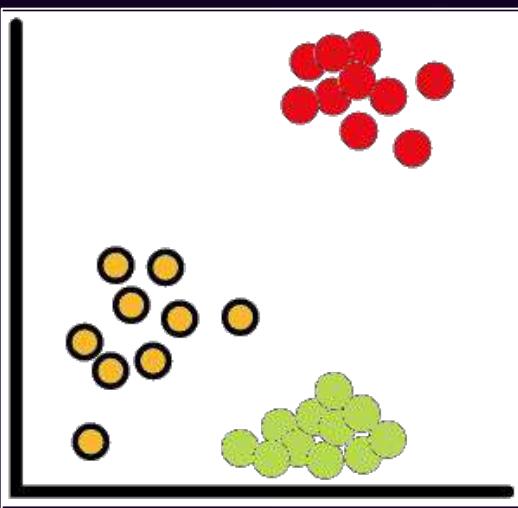


Anomaly detection



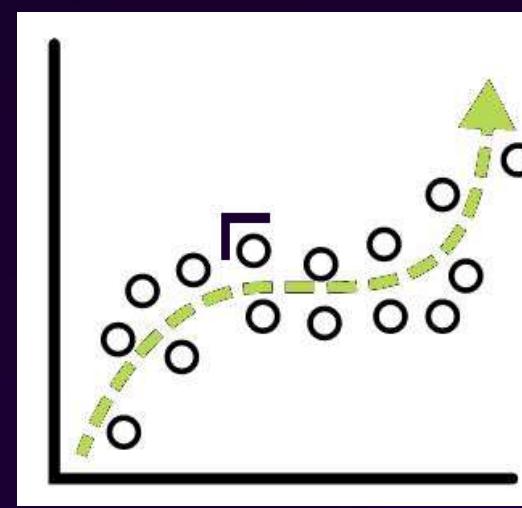
- ▶ Deviation from past behavior
- ▶ Deviation from peers
(aka Multivariate AD or Cohesive AD)
- ▶ Unusual change in features

Clustering



- ▶ Identify peer groups
- ▶ Event Correlation
- ▶ Reduce alert noise
- ▶ Behavioral Analytics

Predictive Analytics



- ▶ Predict Service Health Score/Churn
- ▶ Predicting Events
- ▶ Trend Forecasting
- ▶ Detecting influencing entities
- ▶ Early warning of failure

Data understanding

Is it good data?

Machine data is complex but to get answers of your data, it is important to understand your data.

A clean and usable data set does not have:

- Missing data
- Noisy data
- Bias
- Duplicate data
- outliers

If your data set does not contain any of this you can start applying ML.

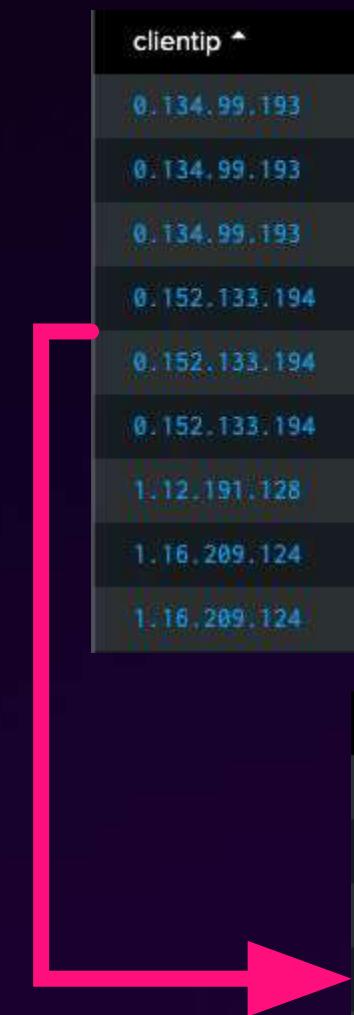
Otherwise, it's time to gather and clean!

Missing Data

ML can not answer questions on which it does not have data

Missing data in a dataset can lead to:

- Reduced model performance
- Corrupted relationships
- Loss of information
- Bias in models



clientip	Country	City
0.134.99.193		
0.134.99.193		
0.134.99.193		
0.152.133.194		
0.152.133.194		
0.152.133.194		
1.12.191.128	China	Haidian (Haidian Qu)
1.16.209.124	Australia	South Brisbane
1.16.209.124	Australia	South Brisbane

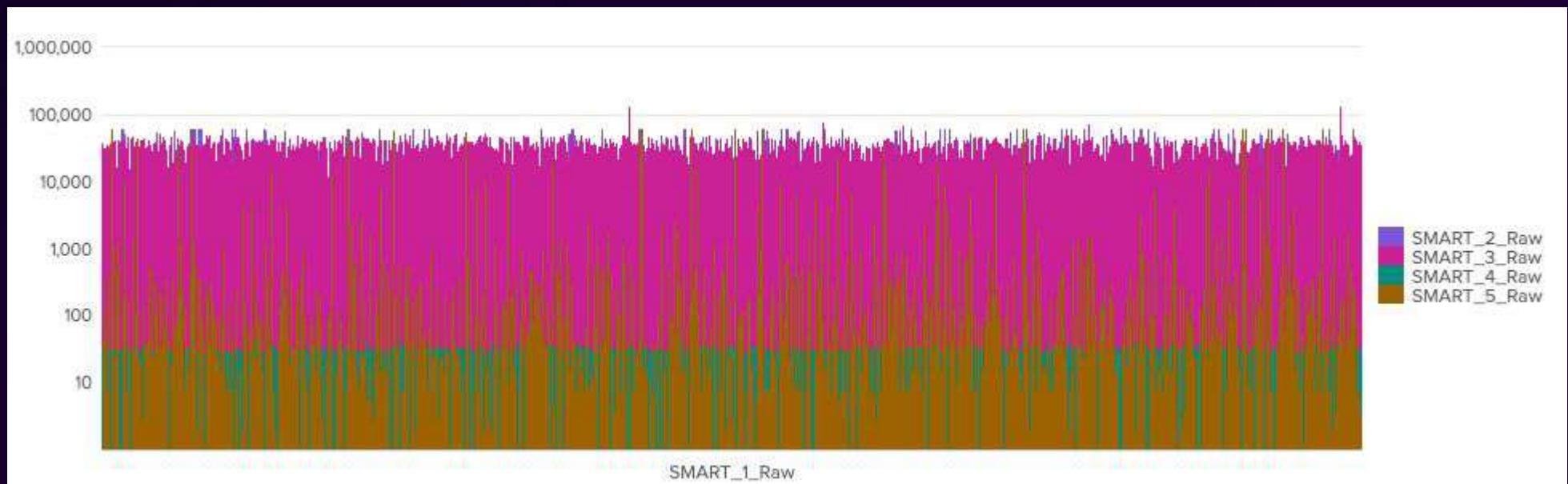
clientip	Country	City
0.134.99.193	unknown	unknown
0.134.99.193	unknown	unknown
0.134.99.193	unknown	unknown
0.152.133.194	unknown	unknown
0.152.133.194	unknown	unknown
0.152.133.194	unknown	unknown
1.12.191.128	China	Haidian (Haidian Qu)
1.16.209.124	Australia	South Brisbane

Noisy Data

Noisy data consists of incorrect, irrelevant, or random variations in the dataset that obscure meaningful patterns and degrade model performance.

Noisy data in a dataset can lead to:

- Confused Model Training
- Reduced Model Accuracy
- Increased Overfitting Risk

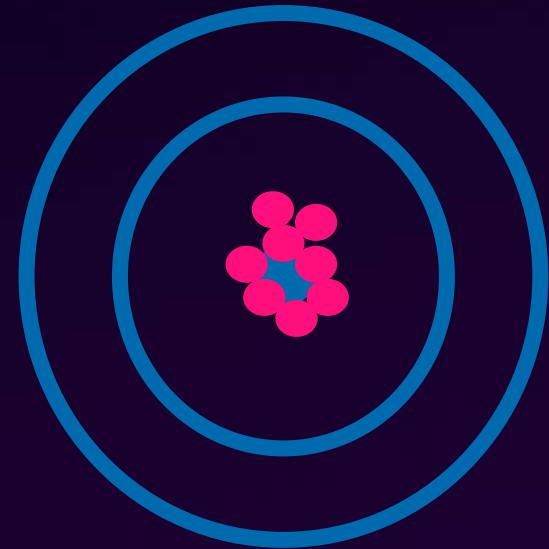


Bias

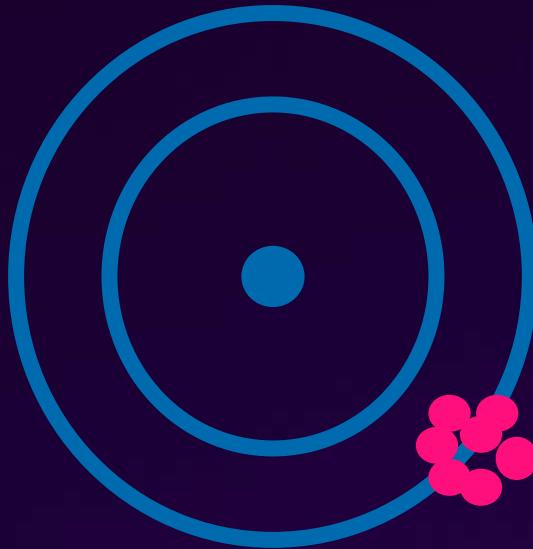
Bias in machine learning refers to systematic errors or assumptions in a model that lead to unfair or inaccurate outcomes by favoring certain patterns, groups, or perspectives over others.

Bias in a dataset can lead to:

- **Unfair Predictions**
- **Discriminatory Outcomes**
- **Misleading Insights**
- **Reduced Trust in the Model**



Low Bias



High Bias

Duplicate Data

Duplicate data refers to repeated records in a dataset that can obscure true patterns, compromise data integrity, and lead to misleading or redundant analysis.

Year	Purchase_price	Term_to_maturity
1978	51.00	24.70
1978	51.00	24.70
1979	59.90	25.00
1979	59.90	25.00
1980	63.00	25.40
1980	63.00	25.40
1981	68.90	25.10
1981	68.90	25.10

Duplicate entries in a dataset can lead to:

- Distorted Model Training
- Biased Results
- Reduced Model Generalization
- Inefficient Use of Resources

```
| inputlookup mortgage_loan_ny_duplicates.csv  
| dedup Year
```

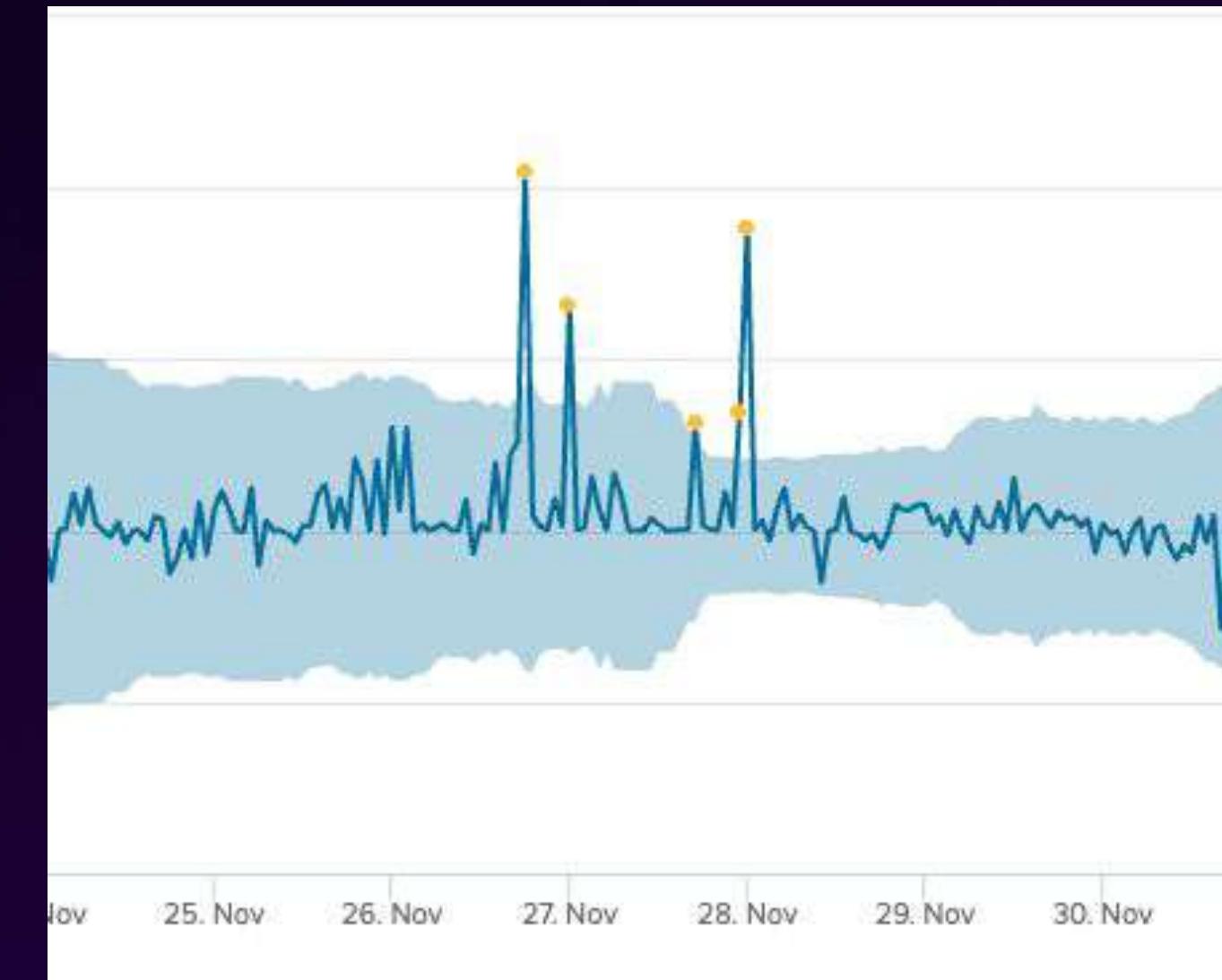
Year	Purchase_price	Term_to_maturity
1978	51.00	24.70
1979	59.90	25.00
1980	63.00	25.40
1981	68.90	25.10
1982	76.70	24.40
1983	86.40	25.20
1984	91.60	26.20

Outliers

Outliers are data points that significantly differ from the majority of the dataset, often representing extreme values or errors that can distort analysis and model performance.

Outliers can lead to:

- Skewed Model Training
- Distorted Metrics
- Reduced Model Generalization
- Increased Preprocessing Efforts



Modern AI Stack: The Emerging Building Blocks for GenAI

Layer 4: Observability	OBSERVABILITY, EVALUATION, SECURITY									
	Helicone	AgentOps	Humanloop	Credal.ai	CALYPSO AI	truera	eppo	BRAINTRUST	Patronus AI	splunk
Layer 3: Deployment	PROMPT MANAGEMENT ORCHESTRATION									
	vellum	LangSmith			Martian	orkes	Radiant			
	AGENT TOOL FRAMEWORKS									
	LangChain	Auto-gpt	FIXIE	LlamaIndex						
Layer 2: Data	DATA PRE-PROCESSING					ETL + DATA PIPELINES				
	gable	datologyai	Cleanlab			UNSTRUCTURED	NOMIC	Lexy	Indexify	
	DATABASES (VECTOR, DB, METADATA STORE, CONTEXT CACHE)									
	databricks	upstash	Pinecone	NEON	WarpStream	momento				
Layer 1: Compute + Foundation	MODEL DEPLOYMENT + INFERENCE							FINETUNING + RLHF		
	baseten	Modal	Replicate	clarifai	Substrate	fireworks.ai	LAMINI	Predibase	arcee.ai	
	FOUNDATION MODELS							TRAINING		
	OpenAI	ANTHROPIC	MISTRAL AI	contextual-ai	Hugging Face	Llama 2	Modular	Lightning AI	OctoML	
	GPU PROVIDERS									
	aws	Azure	Google Cloud	CoreWeave	Lambda	FOUNDRY	together.ai			

Custom ML with the Splunk Platform

Ecosystem

Splunk's App Ecosystem contains 1000's of free add-ons for getting data in, applying structure and visualizing your data giving you faster time to value.

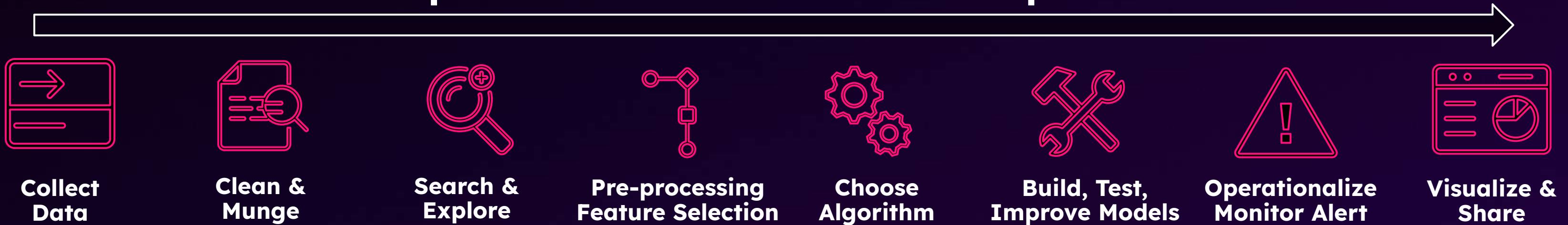
MLTK

The Machine Learning Toolkit delivers new SPL commands, custom visualizations, assistants, and examples to explore a variety of ml concepts.

Splunk

Splunk Enterprise is the mission-critical platform for indexing, searching, analyzing, alerting and visualizing machine data.

Operationalized Data Science Pipeline



Ecosystem	Ecosystem	Ecosystem	MLTK	MLTK	MLTK	MLTK	Ecosystem
Splunk	Splunk	Splunk	Splunk	Splunk	Splunk	Splunk	Splunk

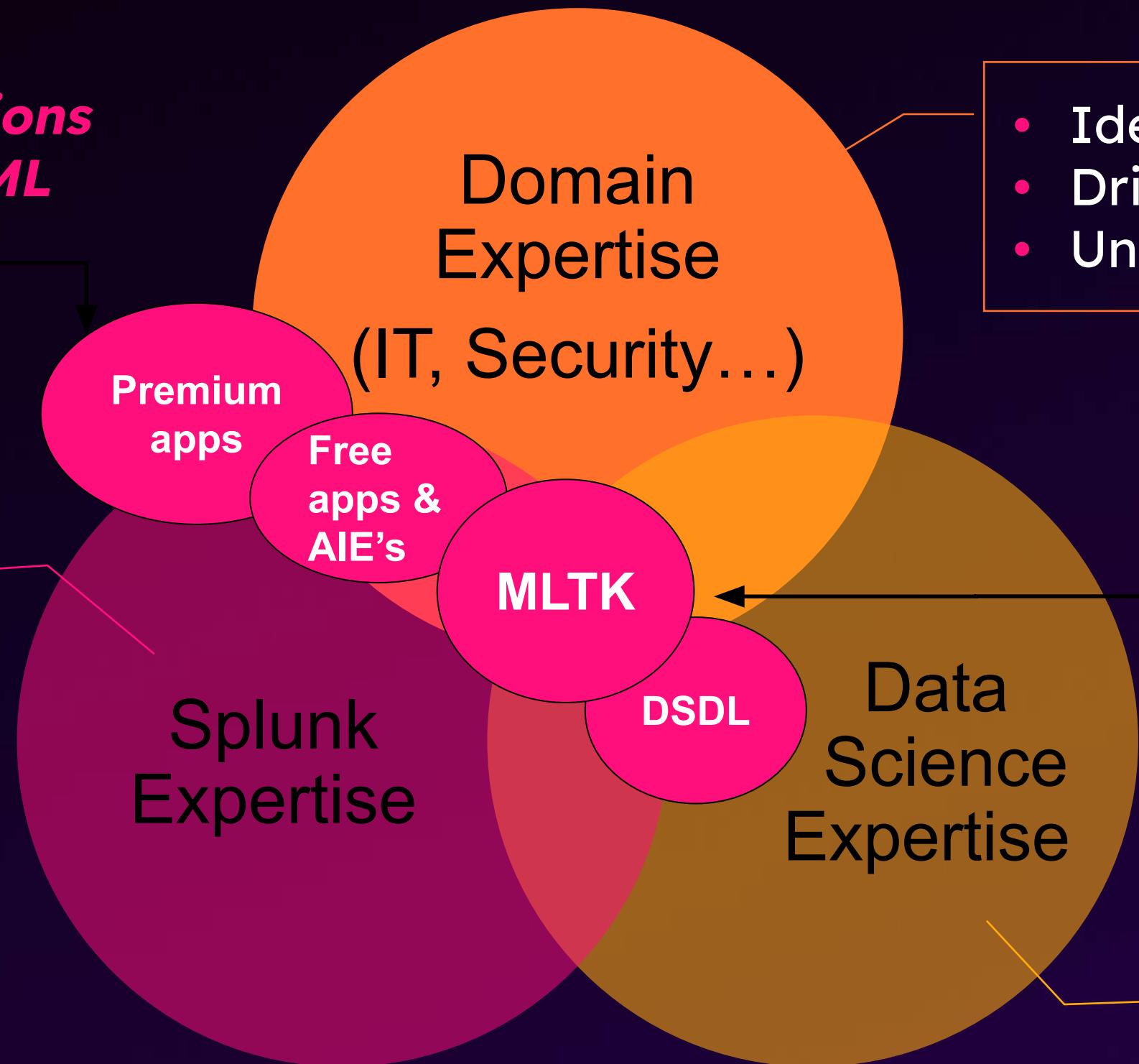
splunk> Platform for Operational Intelligence

Skill Areas for Machine Learning @ Splunk

Premium and AIE solutions provide out of the box ML capabilities.

- Searching
- Reporting
- Alerting
- Workflow

2.



- Identify use cases
- Drive decisions
- Understanding of business impact

1.

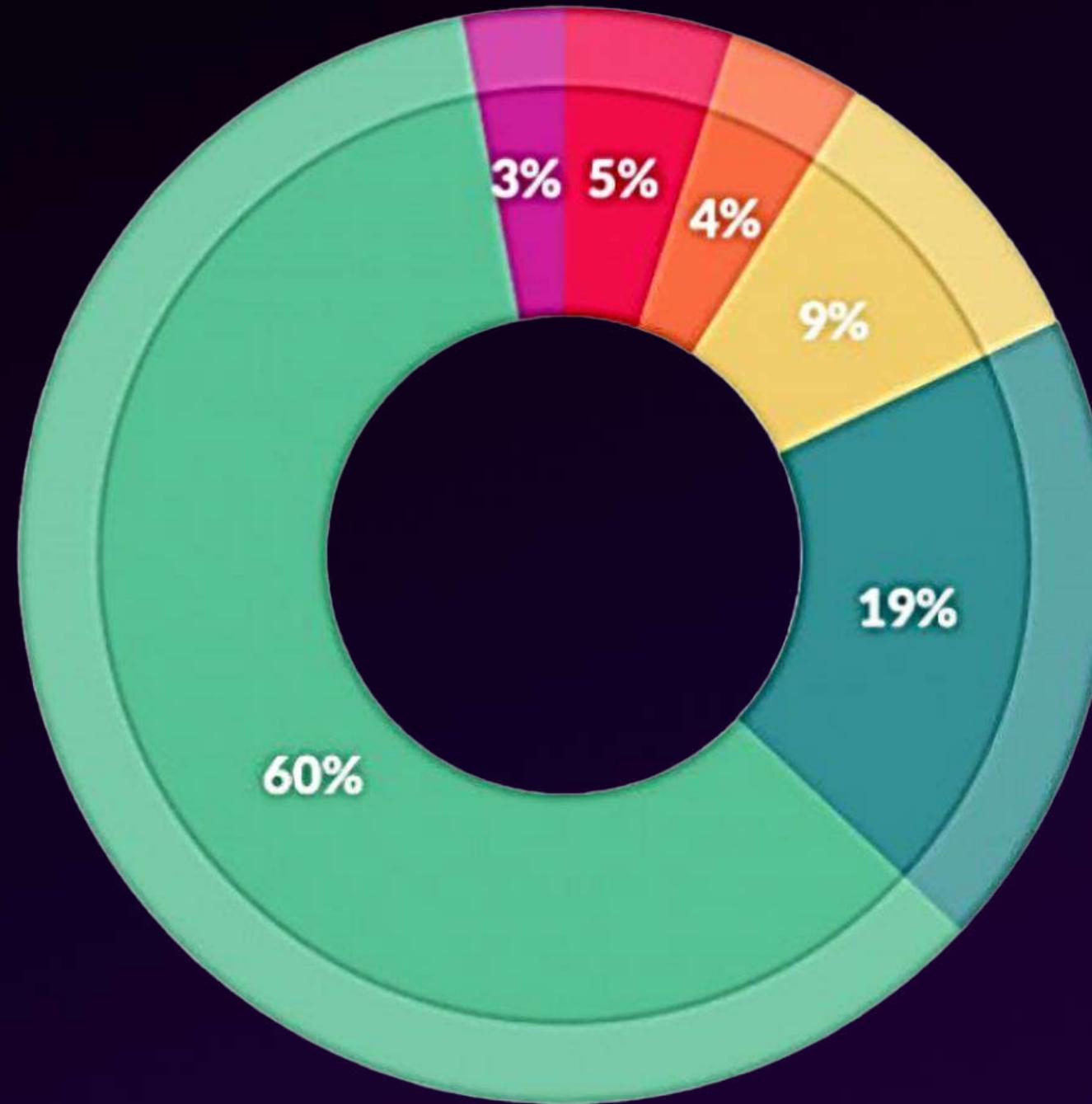
Splunk ML Toolkit facilitates and simplifies via examples & guidance

- Statistics/math background
- Algorithm selection
- Model building

3.

What Data Scientists Really Do

Data Preparation accounts for about 80% of the work of data scientists



Building training sets

3%

Cleaning and organizing data

60%

Collecting data sets

19%

Mining data for patterns

9%

Refining algorithms

4%

Other

5%

“Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says”, Forbes Mar 23, 2016



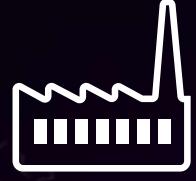
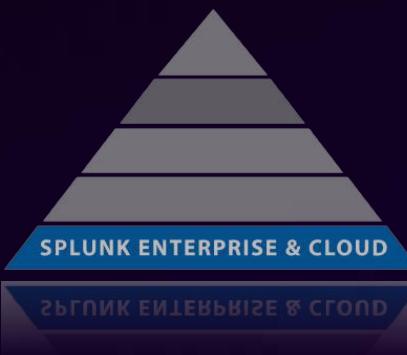
Cleaning Big Data, the Most Time-Consuming, Least Enjoyable Data Science Task

Forbes Survey

March 23

Data Collection

Gain access to previously unused data



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time



Infrastructure data

Application usage and access data

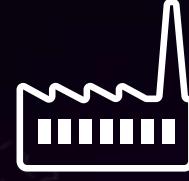
GenAI prompt and output data

Help desk and ticketing system data

IoT and OT data

At Scale

Store, retain, and search data at unprecedented scale



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time



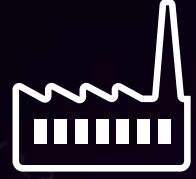
Scalable index and storage

Flexible offering models

Federated analytics

Data understanding

Gain insights to previously unused data



Industrial Assets



Consumer and
Mobile Devices



OT



IT

IT

Real Time

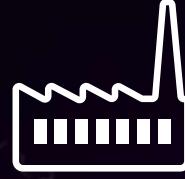
splunk>

Machine data is messy and complex...

```
10.2.1.35 64.66.0.20 - - [17/Jan/2024  
16:21:51] "GET  
/product.screen?product_id=CC-P3-BELKIN-  
SILBLKIPH5&JSESSIONID=SD5SL6FF1ADFF9 HTTP  
1.1" 503 865  
"http://shop.splunktel.com/product.screen?pr  
oduct_id=CC-P3-BELKIN-BLK_BT0OTH_HFREE"  
"Mozilla/5.0 (Linux; Android 12.0.0; FR-fr;  
SM-S901B Build/S908EXXU2BVJA)  
AppleWebKit/537.36 Chrome/114.0.5735.131  
Mobile Safari/537.36" 954
```

Data understanding

Gain insights to previously unused data



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

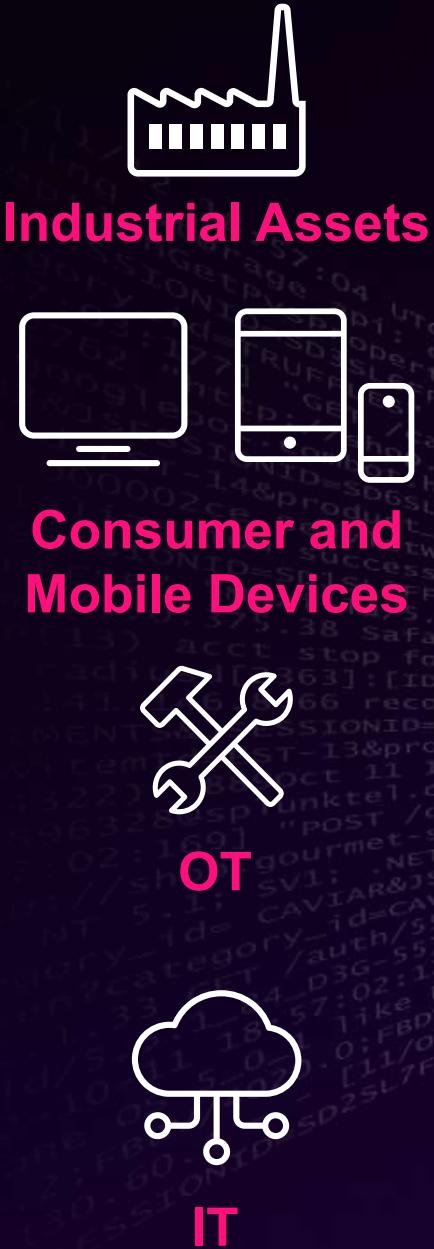
splunk>

Machine data is ~~messy and complex~~ valuable!

```
10.2.1.35 64.66.0.20 - - [17/Jan/2024
16:21:51] "GET
/product.screen?product_id=CC-P3-BELKIN-
SILBLKIPH5&JSESSIONID=SD5SL6FF1ADFF9 HTTP
1.1" 503 865
"http://shop.splunktel.com/product.screen?pr
oduct_id=CC-P3-BELKIN-BLK_BT0OTH_HFREE"
Mozilla/5.0 (Linux; Android 12.0.0; FR-fr;
SM-S901B Build/S908EXXU2BVJA)
AppleWebKit/537.36 Chrome/114.0.5735.131
Mobile Safari/537.36" 954
```

Data understanding

Gain insights to previously unused data



Machine data · and complex valuable!

User IP

```
10.2.1.35 64.66.0.20 - - [17/Jan/2024
16:21:51] "GET
/product.screen?product_id=CC-P3-BELKIN-
SILBLKIPH5&JSESSIONID=SD5SL6FF1ADFF9 HTTP
1.1" 503 865
"http://shop.splunktel.com/product.screen?pro
duct_id=CC-P3-BELKIN-BLK_BT0OTH_HFRE
Mozilla/5.0 (Linux; Android 12 9.0; FR-fr;
SM-S901B Build/S908EXXU21A)
AppleWebKit/537.36 Chrome/114.0.5735.131
Mobile Safari/537.36" 954
```

Product Viewed

Preferred Language

Device

Data understanding

Is it good data?



Machine data is complex but to get answers of your data, it is important to understand your data.

A clean and usable data set does not have:

- **Missing data**
- **Noisy data**
- **Bias**
- **Duplicate data**
- **outliers**

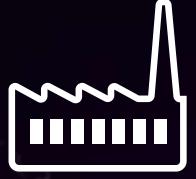
If your data set does not contain any of this you can start applying ML.

Otherwise, it's time to gather and clean!

Data Access

Gain access to previously unused data

```
index="oidemo" sourcetype="access_combined"
```



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

Data Preparation

Filter data with SPL



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

index="oidemo" sourcetype="access_combined" action=purchase

action

5 Values, 44.612% of events

Selected Yes No

Reports

Top values Top values by time Rare values

Events with this field

Values	Count	%
purchase	2,243	20.709%
addtocart	2,172	20.054%
view	2,168	20.017%
remove	2,161	19.952%
changequantity	2,087	19.269%

action

1 Value, 100% of events

Selected Yes No

Reports

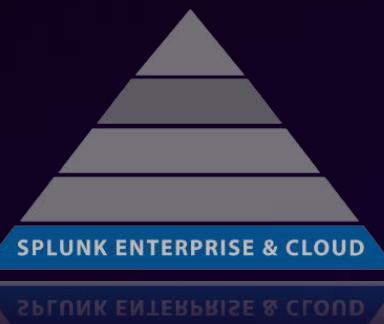
Top values Top values by time Rare values

Events with this field

Values	Count	%
purchase	2,249	100%

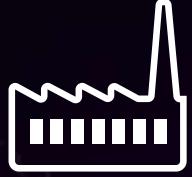
Data Preparation

Transform data with SPL



Filter

```
index sourcetype="access_combined" action=purchase  
| iplocation clientip
```



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

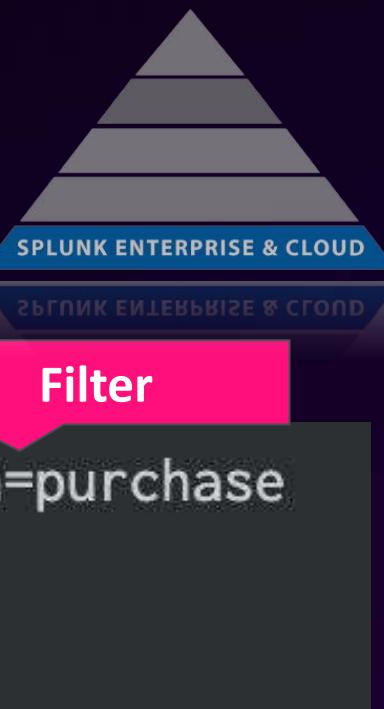
splunk>

clientip	Country	City
0.134.99.193		
0.134.99.193		
0.134.99.193		
0.152.133.194		
0.152.133.194		
0.152.133.194		
1.12.191.128		
1.16.209.124		

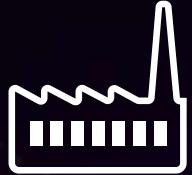
clientip	Country	City
0.134.99.193		
0.134.99.193		
0.134.99.193		
0.152.133.194		
0.152.133.194		
0.152.133.194		
1.12.191.128	China	Haidian (Haidian Qu)
1.16.209.124	Australia	South Brisbane
1.16.209.124	Australia	South Brisbane

Data Preparation

Clean data with SPL



Filter



Industrial Assets



Consumer and
Mobile Devices



OT



IT

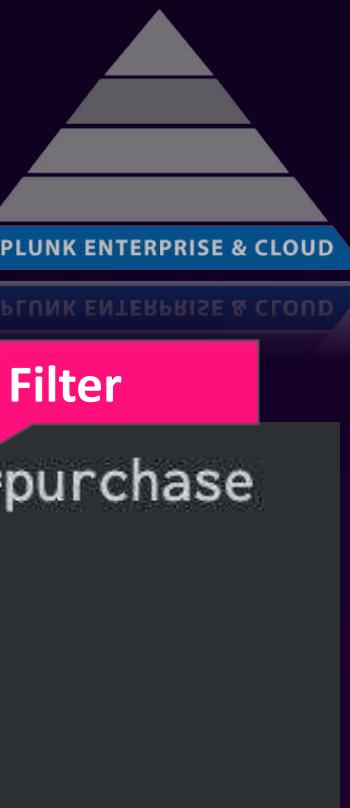
Real Time

splunk>

```
index sourcetype="access_combined" action=purchase  
| iplocation clientip  
| fillnull value="Unknown" Country
```

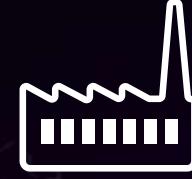
clientip	Country	City
0.134.99.193		
0.134.99.193		
0.134.99.193		
0.152.133.194		
0.152.133.194		
0.152.133.194		
1.12.191.128	China	Haidian (Haidian Qu)
1.16.209.124	Australia	South Brisbane
1.16.209.124	Australia	South Brisbane

clientip	Country	City
0.134.99.193	unknown	unknown
0.134.99.193	unknown	unknown
0.134.99.193	unknown	unknown
0.152.133.194	unknown	unknown
0.152.133.194	unknown	unknown
0.152.133.194	unknown	unknown
1.12.191.128	China	Haidian (Haidian Qu)
1.16.209.124	Australia	South Brisbane



Data Preparation

Aggregate data with SPL



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

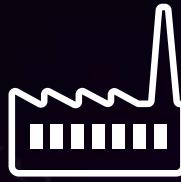
```
index sourcetype="access_combined" action=purchase  
| iplocation clientip  
| fillnull value="Unknown" Country  
| stats count by product device Country
```

Aggregate

product	device	Country	count
Bubble_Wrap	iPhone	United States	11
Man_Candle-Bacon	iPhone	United States	10
Canned_Unicorn_Meat	Windows	United States	9

Training ML model

Fit model



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

```
index sourcetype="access_combined" action=purchase
| iplocation clientip
| fillnull value="Unknown" Country
| stats count by product device Country
```

Filter

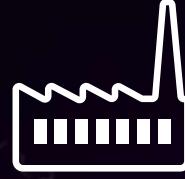
| fit algorithm from clientip into "model"

Aggregate

Train "model"

Productionalize ML model

Apply model



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

```
index sourcetype="access_combined" action=purchase
| iplocation clientip
| fillnull value="Unknown" Country
| stats count by product device Country
```

Filter

Transform

Clean

Aggregate

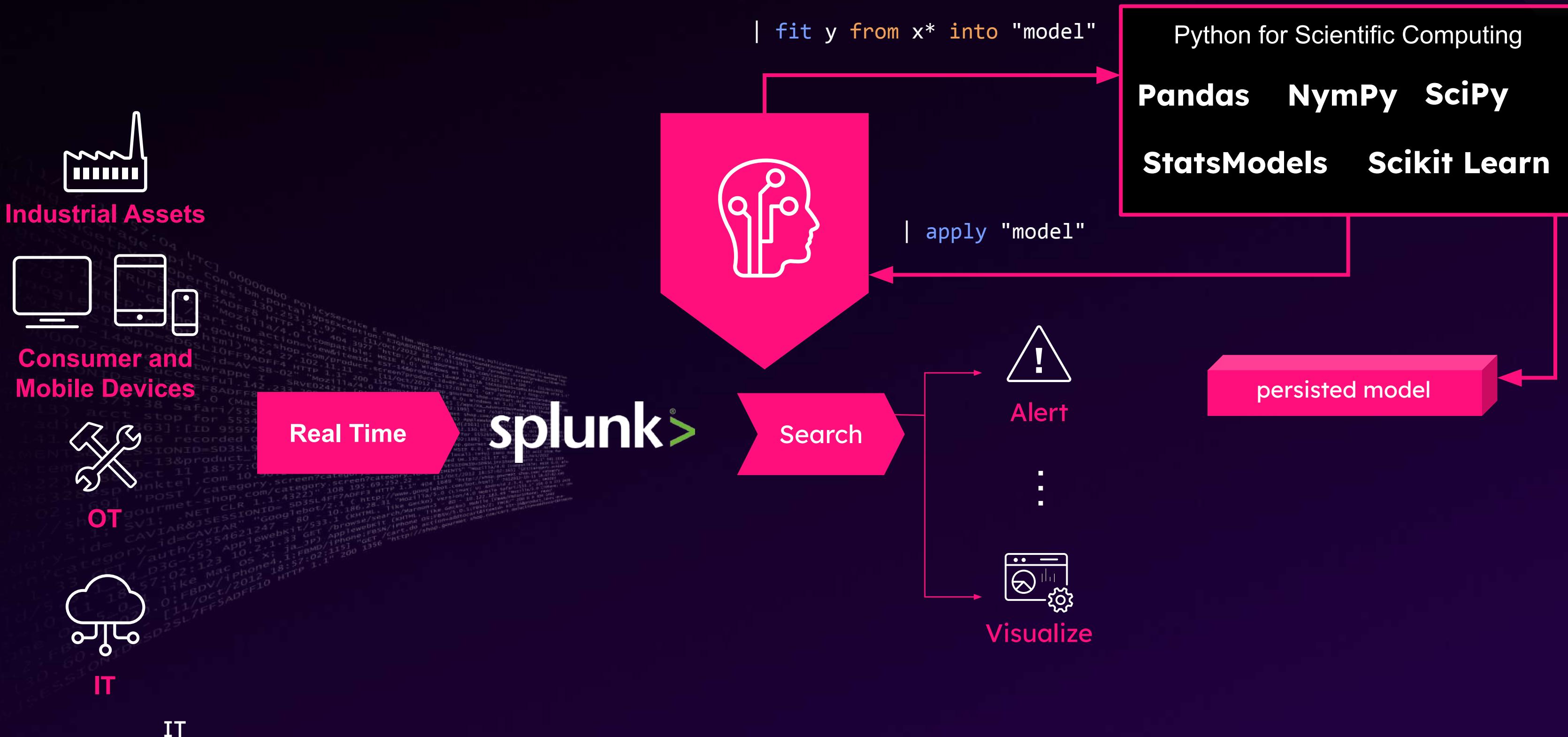
```
| fit algorithm from clientip into "model"
```

Train "model"

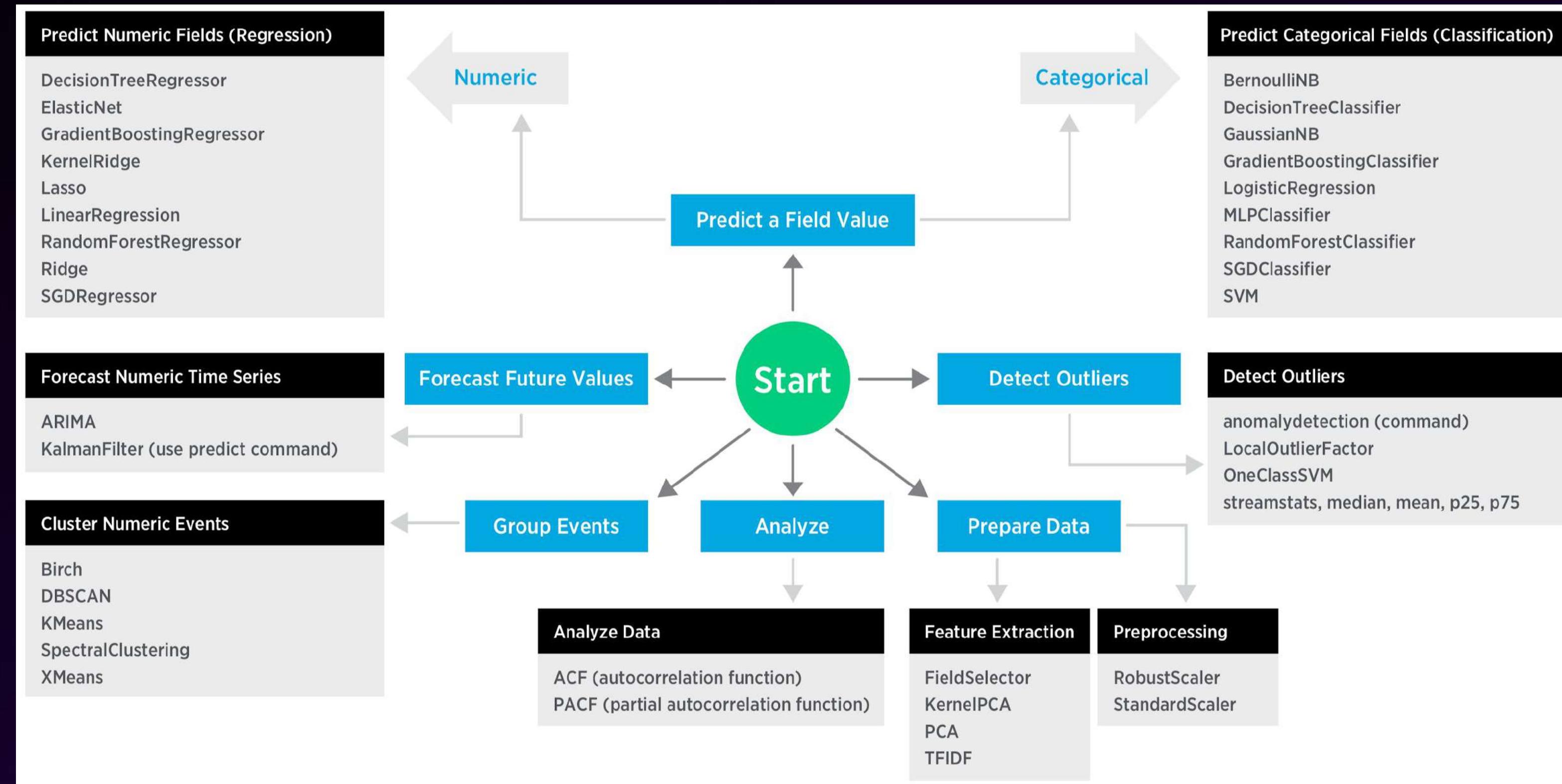
```
| apply "model"
```

Run "model"

Model training and application



Quick Overview of Algorithms in MLTK



<https://www.splunk.com/pdfs/solution-guides/machine-learning-quick-ref-guide.pdf>

Search Commands for MLTK

MLTK implements new custom search commands

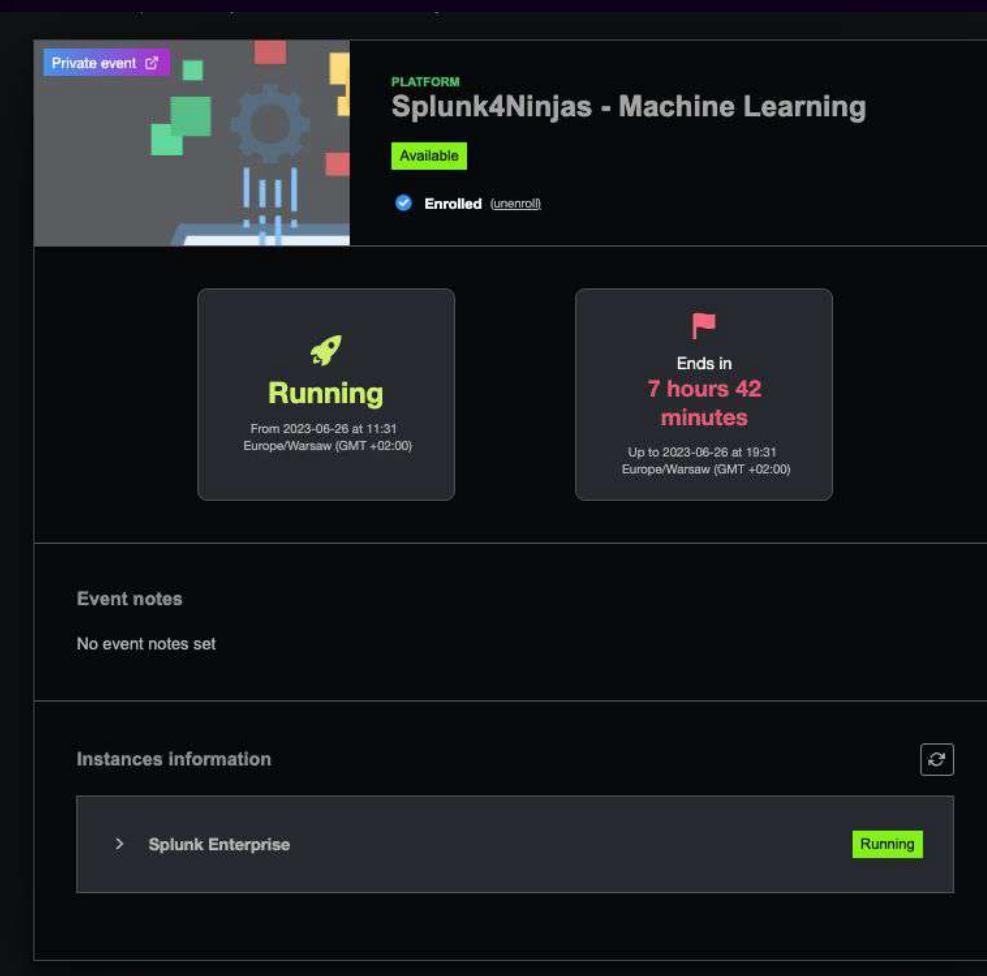
ML-SPL command name	Description
fit	Fit and apply a machine learning model to search results.
apply	Apply a machine learning model that was learned using the <code>fit</code> command.
summary	Return a summary of a machine learning model that was learned using the <code>fit</code> command.
listmodels	Return a list of machine learning models that were learned using the <code>fit</code> command.
deletemodel	Delete a machine learning model that was learned using the <code>fit</code> command.
sample	Randomly sample or partition events.
score	Run statistical tests to validate model outcomes.

<https://docs.splunk.com/Documentation/MLApp/latest/User/Customsearchcommands>

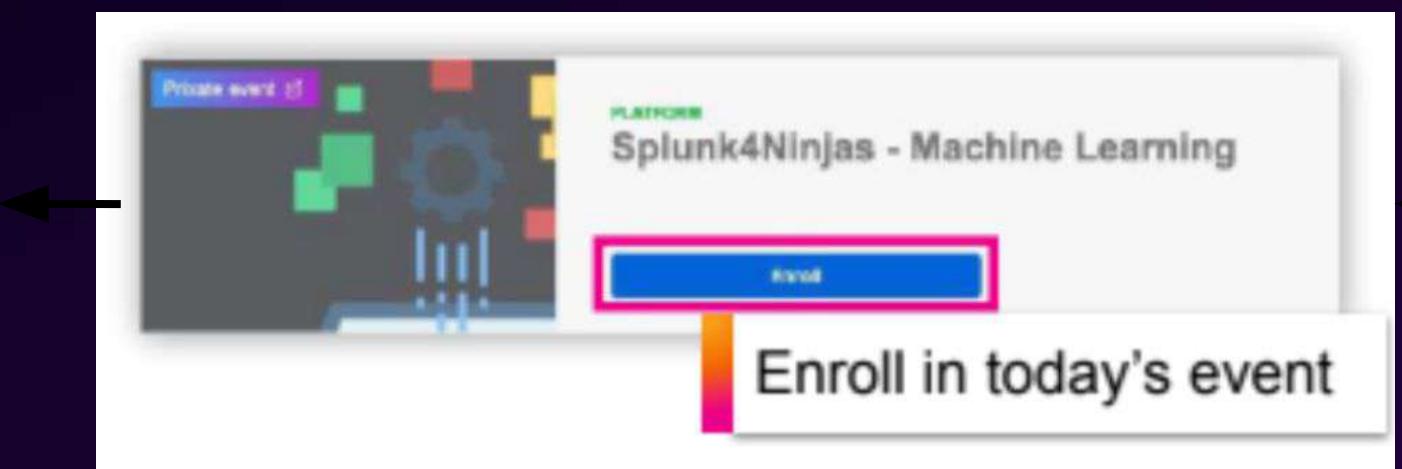
Registration

1. Register fo a splunk.com account if you don't have one yet:
https://www.splunk.com/en_us/sign-up

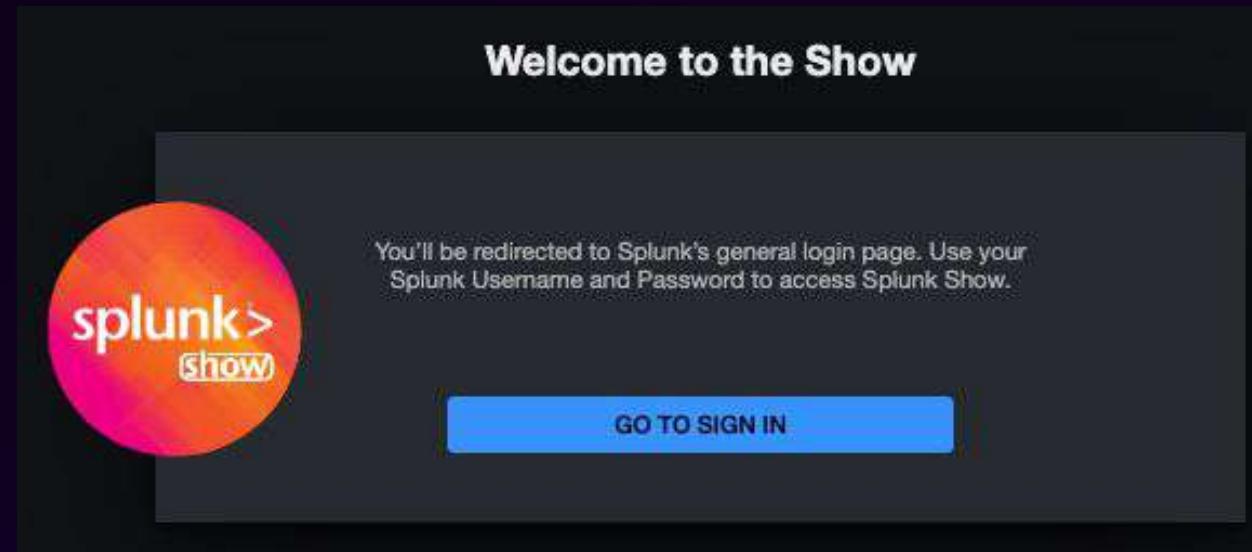
2. Enroll in the Splunk Show workshop event:
<https://show.splunk.com/event/<eventID>>



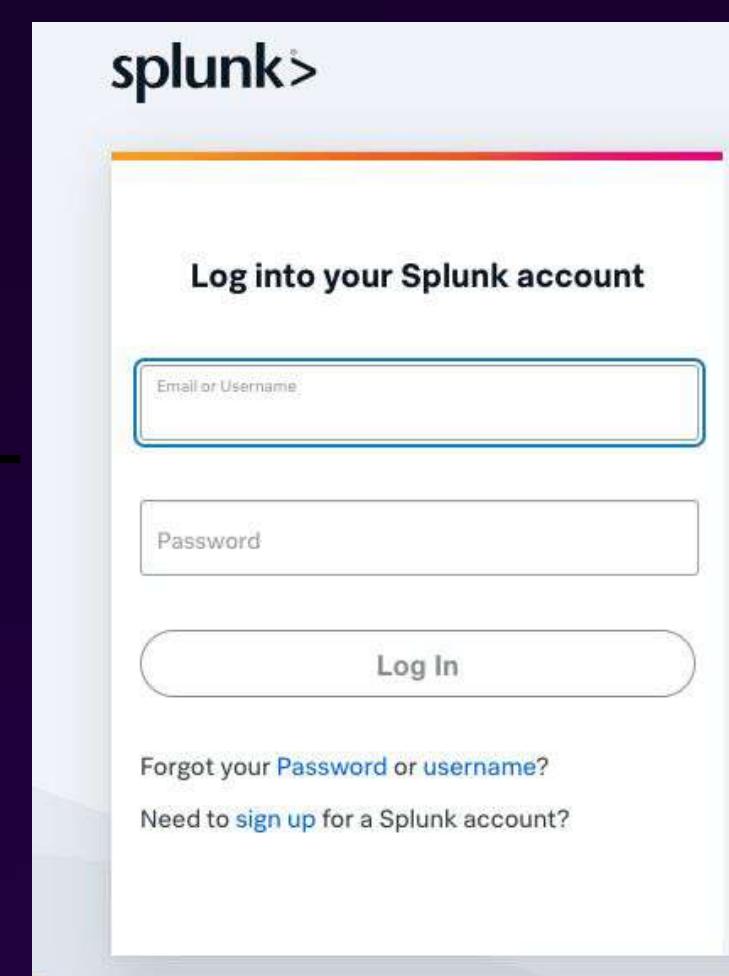
This screenshot shows the event details for 'Splunk4Ninjas - Machine Learning'. It includes a thumbnail image, the platform name, and status information. The event is marked as 'Running' from 2023-06-26 at 11:31 Europe/Warsaw (GMT +02:00) and ends in 7 hours 42 minutes on 2023-06-26 at 19:31 Europe/Warsaw (GMT +02:00). The status is 'Enrolled' with a link to 'unenroll'. Below this, there's a section for 'Event notes' stating 'No event notes set'. Under 'Instances information', it lists 'Splunk Enterprise' as running.



This screenshot shows the enrollment page for the same event. It features the event thumbnail, platform name, and a large blue 'Enroll' button. Below the button, a call-to-action says 'Enroll in today's event'.



This screenshot shows the 'Welcome to the Show' page. It features the Splunk Show logo and a message: 'You'll be redirected to Splunk's general login page. Use your Splunk Username and Password to access Splunk Show.' A blue 'GO TO SIGN IN' button is present.



This screenshot shows the 'Log into your Splunk account' page. It has fields for 'Email or Username' and 'Password', and a 'Log In' button. Below the form, links for 'Forgot your Password or username?' and 'Need to sign up for a Splunk account?' are provided.

Security Use Cases Enhanced by AI and ML

splunk>



Security Use Cases Enhanced by AI and ML

Over the last few years, organizations have had to cope with different types of disruption on a global scale, investing in digital resilience to ward off potential threats and eliminate business risk. This ability to predict incidents before they occur is a key area for driving value with artificial intelligence (AI) and machine learning (ML), helping companies prevent downtime in the face of an outage or attack.

Download your complimentary copy of **Security Use Cases Enhanced by AI and ML** to learn more about artificial intelligence and machine learning, including:



- ✓ How to implement AI and ML in Splunk
- ✓ Examples of business challenges and use cases
- ✓ Customer case studies and documentation

Download Now

Business Email

REQUIRED

Continue



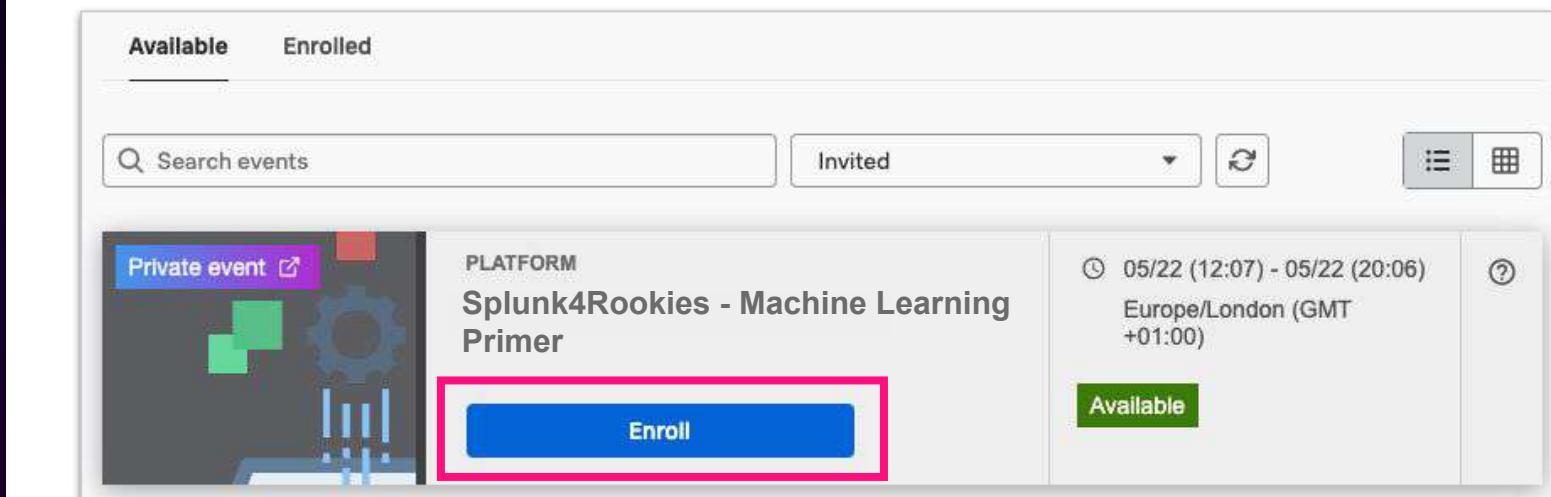


Enroll in Today's Workshop

Tasks

1. Get a [splunk.com account if you don't have one yet:](https://splk.it/SignUp)
<https://splk.it/SignUp>
2. Enroll in the Splunk Show workshop event:
<https://show.splunk.com/event/<eventID>>
3. Download a copy of today's slide deck:
<https://splk.it/S4RML-Sec-Attendee>

Goal



Enroll in today's event

Use case: Detect User behavior outliers

1. Identifying user access anomalies

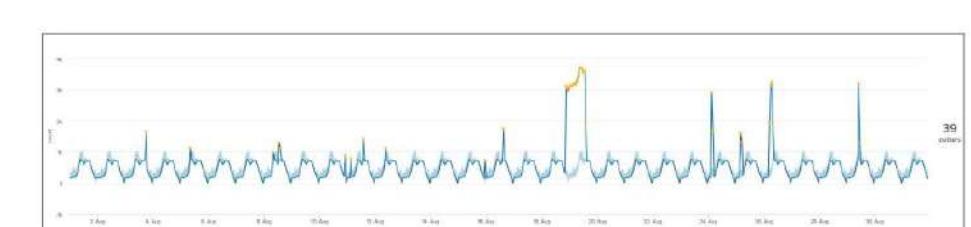
Business challenge

Research suggests that compromised credentials provide the entry point for **71% of cyber attacks against businesses**. Without recognizing what normal user behavior looks like in an organization, catching potential compromises can be difficult. Furthermore, baselining user behavior can be challenging when different departments, geographies, teams or facilities have different working patterns. If organizations are unable to quickly spot compromises and baseline user behavior, then threat actors will leverage those handicaps to easily gain access to systems undetected.

Splunk's approach

We have out-of-the-box capabilities in Splunk Enterprise Security that are designed to detect potential user compromises using ML, such as detecting potential brute force attempts by baselining failed logon attempts. Additionally, our UBA product profiles and baselines user access data to detect potential compromises.

For customers who wish to design custom rules for detecting unusual account behavior, the MLTK provides a range of algorithms for generating baselines from historic data that can be used to detect deviations from the baseline.



Value

Increase detection efficiency: Use of ML to identify user access anomalies can help to identify potential compromise to the business and mitigate risk.

Reduce manual processing: Use of ML can increase efficiency by programmatically identifying anomalies, reducing time spent analyzing data manually.



Ministry of Energy
www.energy.gov.il

TLP: White | בלחימת | CERT-IL

סיבר ישראל | מערך הסייבר הלאומי

19 אוגוסט 01
בנ"ד תמוד תשע"ט
מספר: ב-ט-943

חשיפת קמפיין תקיפה מתמשך בנגד ארגונים בישראל

תקציר



1. בחקירות שהתקיימו במהלך החודשים האחרונים, זוהתה קבוצת תקיפה הפעלת מול ארגונים רבים בישראל, בין היתר ב�ג'ור האנרגיה, אקדמיה, חברות IT וחברות Hosting.

2. מהחקרה עולה כי קבוצת התקיפה פעלת מול מספר סוגים של תשתיות מחשב וביניהם – שירותי WEB, מסקייני גישה מרוחק (RDP / VPN), שירותי DNS ושירותי OWA.

3. מטרת מסמך זה היא לפרט את שיטות הפעולה, הכלים והתשתיות בהם נעשה שימוש והמלצות טכניות ליהיו, מנעה וחיסימה.

Cyberattacks Don't Stop at Critical Infrastructure

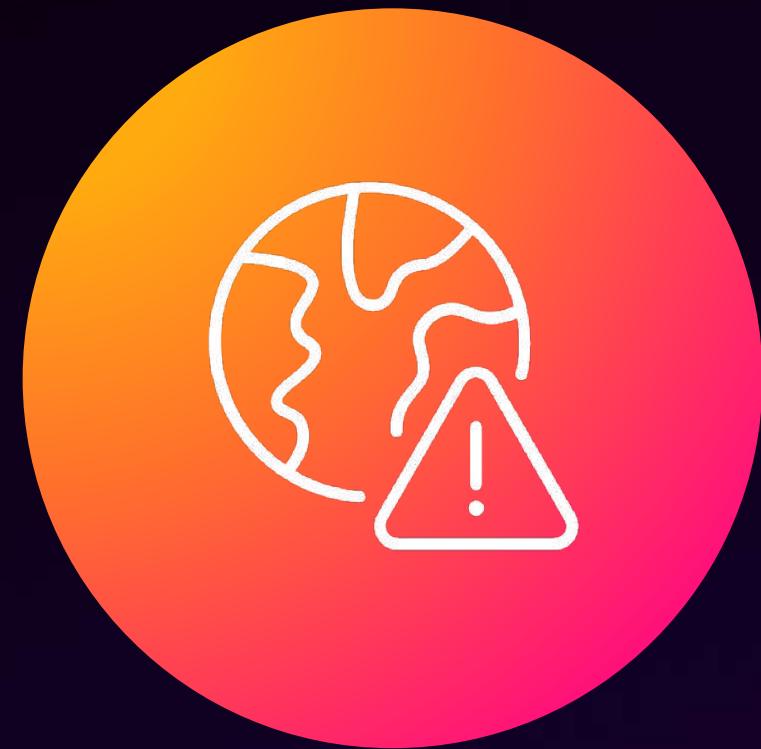
Just recently a threat activity group targeted many organizations in Israel in the **energy sector**, academia and hosting companies. **IT infrastructure, web servers, RDP/VPN used for remote access and more were under attack.**



Outlier Detection Algorithms

Identify and analyze
abnormal behavior in
your data

Global



Data points different from
expected pattern, range, or norm

Contextual



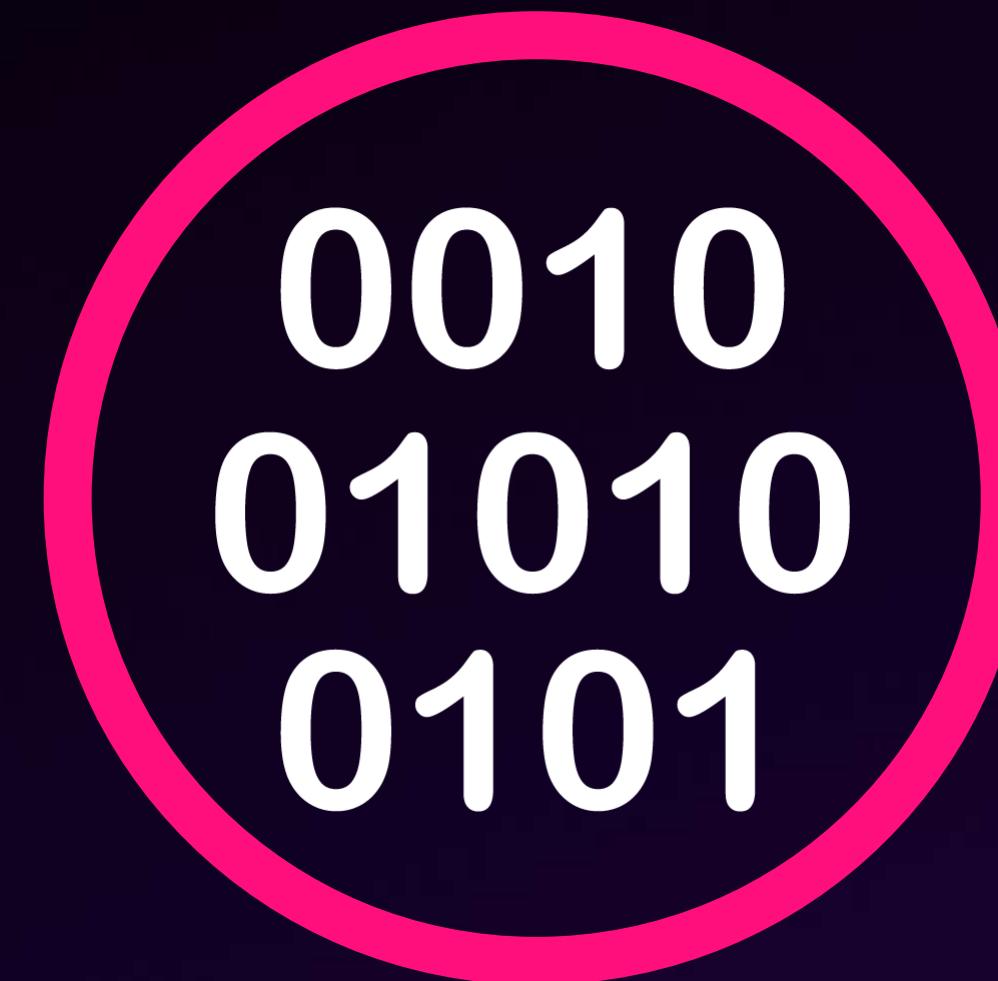
Are the results out
of context?

Collective



Looks normal with isolation
but stands out in a group

Numeric



Categorical



Included Algorithms

DensityFunction

| LocalOutlierFactor

| MultiVariateOutlierDetection

| One-Class SVM

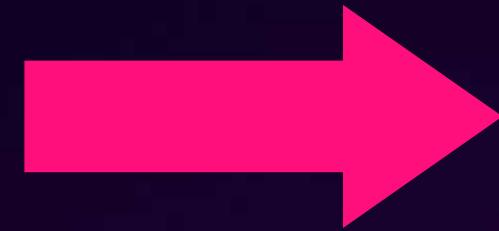
Categorical data to Numeric data?

device
server01
server02
server03



Categorical data to Numeric data?

device
server01
server02
server03



device	server01	server02	server03
server01	1	0	0
server02	0	1	0
server03	0	0	1

Outlier Detection Algorithms

Identify and analyze
abnormal behavior in
your data

Includes:

- Clustering
- Outlier Detection



Live Instance Demo

Summary

Top 4 most important things to remember about outlier detection

1



Outlier detection is a way of analyzing your data for **historical baseline outliers**

2



Models **assume historic data input represents normal data**

3



Encoding is necessary for categorical outlier detection

4



Choice of outlier algorithm may rely on a **subject matter expert** of the data

Exercise 1

Detect User behavior outliers

Time: 15 minutes

Use Case: Predict presence of Malware

6. Detecting malicious patterns of network traffic

Business challenge

Network security monitoring is an important function within a SOC and is used to detect potential infiltration, exfiltration and lateral movement by malicious actors. However, understanding and baselining what normal network behavior looks like in an organization can be challenging, with potentially thousands of enterprise endpoints and API calls creating noise that attackers can use as a hideout.

Splunk's approach

Splunk Enterprise Security contains out-of-the-box correlation searches that utilize the MLTK to detect unusual volumes of network activity or substantial increases in port activity. Additionally, the Splunk UBA product ships with many ML-enabled use cases for network security monitoring, such as potential data exfiltration identification.

In addition to these products that provide ready-to-use content, the MLTK can be used to produce analytics that detect unusual patterns of network traffic.

Value

Increase detection efficiency: The ability to baseline typical network behavior and identify times when there is anomalous activity on the network can help detect a range of potentially malicious behavior, such as DDoS attacks, botnet activity or the presence of malware.

Reduce manual processing: Furthermore, behavioral-based detections can improve analyst efficiency, reducing the amount of time that security analysts spend manually triaging network behavior after an alert triggers.

Case studies and further information

Siemens

As described above, Siemens partnered with Splunk to develop a set of ML-based security detections to augment their SOC. Siemens was able to baseline proxy communications to identify outliers in outbound web proxy communications. This baseline allowed Siemens to detect potential malware communications, in particular flagging HTTP tunneling of SSH traffic.

To develop this use case, Siemens identified eight features from their proxy logs in which they wanted to identify outliers, including bytes in, bytes out, and the number of distinct IP addresses visited by a given source IP over hourly intervals. For each of these eight features, a model was trained using the MLTK's DensityFunction to baseline expected behavior for each feature. From these baselines,

outliers are detected for each feature. Every hour, an anomaly score is created for each source IP based on the number of outliers across all eight features. Creating a set of drill downs from the anomaly detection allowed analysts quickly to triage potentially malicious web proxy traffic.

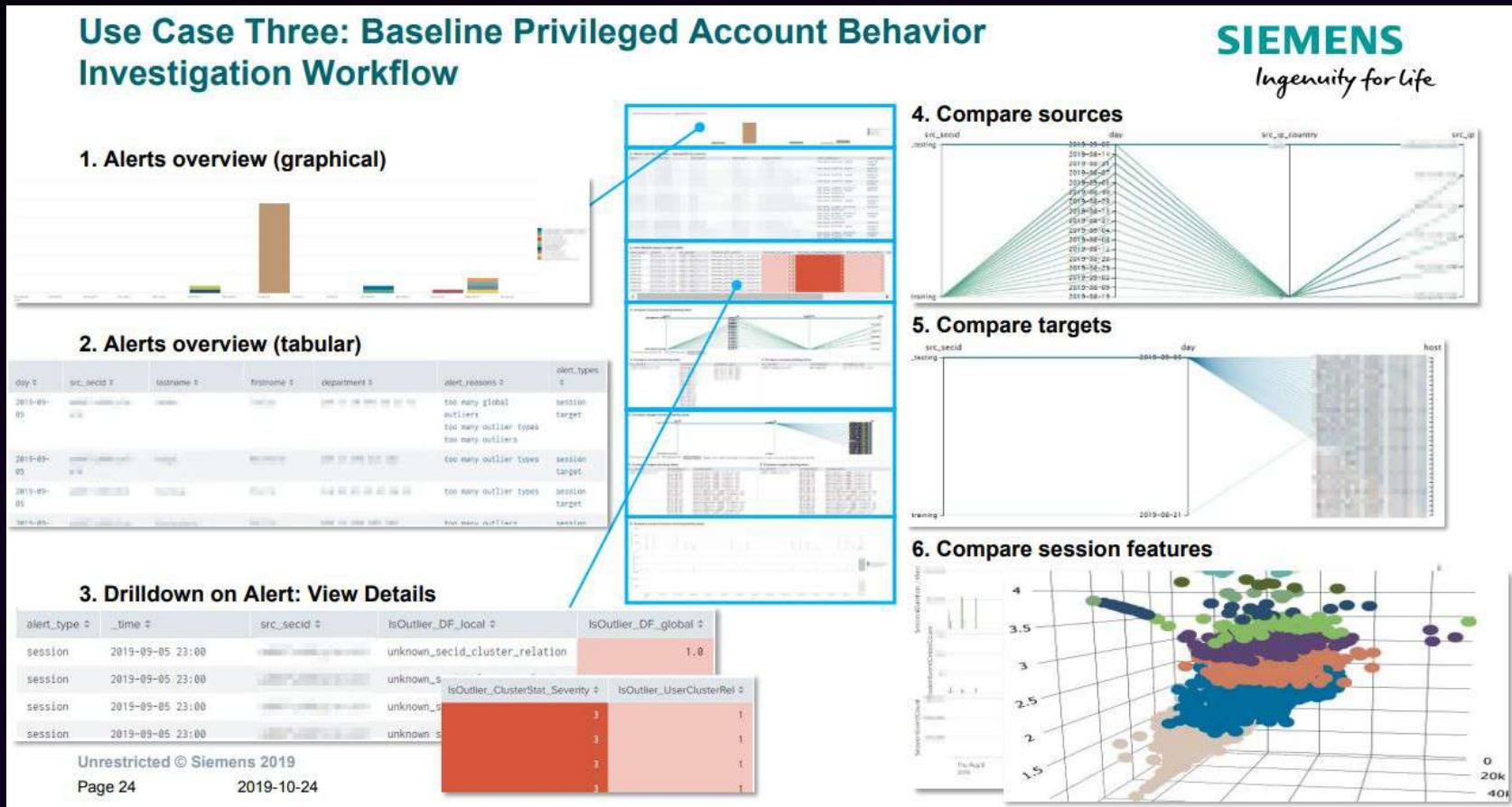
Find out more about Siemens experience with the MLTK [here](#).

Further information

Additionally, Splunk has a simple guide that describes how to detect network traffic anomalies based on the amount of data being transferred between source and destination IP pairs [here](#). Provided that all the correct apps are installed and some network traffic data exists in a Splunk instance, this how-to guide should take a few hours to implement and test.

Additionally, a number of Splunkbase apps are available that provide workflows for detecting unusual network behavior. Examples include the Botnet App for Splunk, which uses an open source data set and a guided workflow to help users to train a set of classification models for detecting potential botnet activity on their network. This app can be found [here](#).

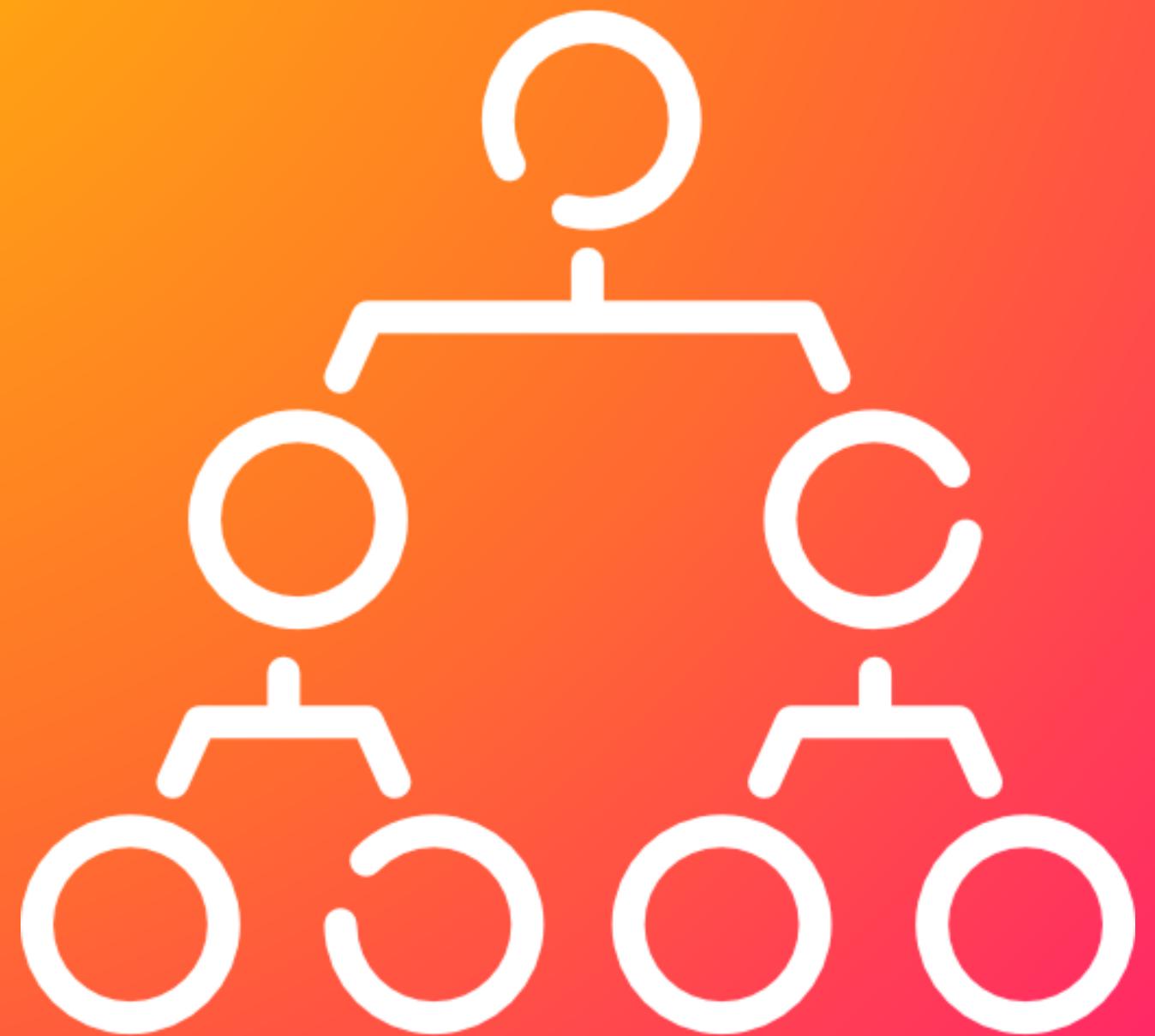
Splunk at Siemens EAGLE: Since 12 years In 4 locations



8 Petabytes storage
Net-zero carbon footprint
IT Ops and Security
70 indexes
3,200 data sources
210 data types

Categorical Prediction

“A method that lets you quickly, easily, and sustainably gain insight into your data by predicting its categorical features”



How Does Categorical Prediction Work?

How Does Categorical Prediction Work?

alert group	count	src	alert value	Predicate
T-Shirt Co.	15	checkoutsvc	107	False positive
BTCP Digital	17	btcup_checkout	375	True positive
BTCP Digital	3	payment_svc	89	True positive

How Does Categorical Prediction Work?

alert group	count	src	alert value	Predicate
T-Shirt Co.	15	checkoutsvc	107	False positive
BTCP Digital	17	btcup_checkout	375	True positive
BTCP Digital	3	payment_svc	89	True positive

How Does Categorical Prediction Work?

alert group	count	src	alert value	Predicate
T-Shirt Co.	15	checkoutsvc	107	False positive
BTcup Digital	17	btcup_checkout	375	True positive
BTcup Digital	3	payment_svc	89	True positive

How Does Categorical Prediction Work?

alert group	count	src	alert value	Predicate
T-Shirt Co.	15	checkoutsvc	107	False positive
BTCP Digital	17	btcup_checkout	375	True positive
BTCP Digital	3	payment_svc	89	True positive

alert group	count	src	alert value	Predicate (predicted)
T-Shirt Co.	17	checkoutsvc	100	False positive
O11y Cloud	16	browsercheck	89	False positive

How Does Categorical Prediction Work?

alert group	count	src	alert value	Predicate
T-Shirt Co.	15	checkoutsvc	107	False positive
BTCP Digital	17	btcup_checkout	375	True positive
BTCP Digital	3	payment_svc	89	True positive

alert group	count	src	alert value	Predicate (predicted)
T-Shirt Co.	17	checkoutsvc	100	False positive
O11y Cloud	16	browsercheck	89	True positive?

How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL

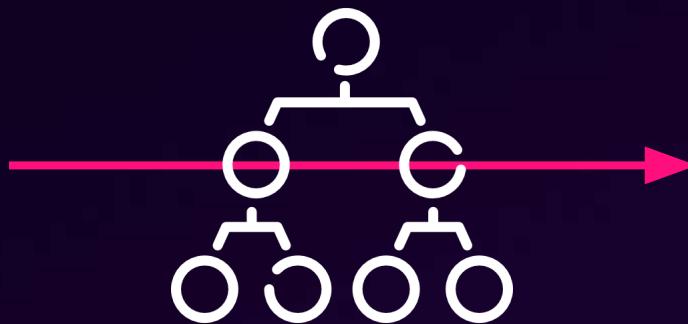
How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL

New Data

Feature 1	...	Feature n	Severity (predicted)
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL



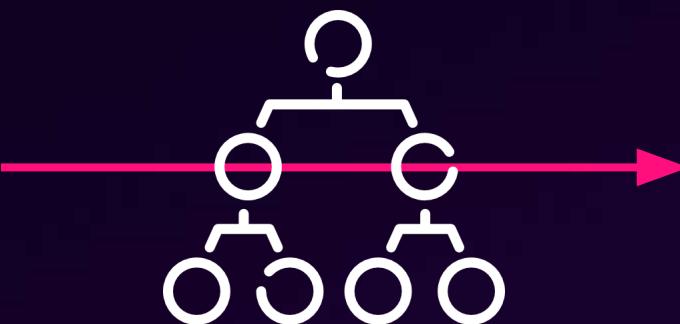
How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL

New Data

Feature 1	...	Feature n	Severity (predicted)	Severity (actual)
[data]	...	[data]	CRITICAL	MAJOR
[data]	...	[data]	CRITICAL	LOW
[data]	...	[data]	CRITICAL	MINOR
[data]	...	[data]	CRITICAL	CRITICAL
[data]	...	[data]	CRITICAL	LOW
[data]	...	[data]	CRITICAL	MINOR
[data]	...	[data]	CRITICAL	MAJOR
[data]	...	[data]	CRITICAL	MAJOR
[data]	...	[data]	CRITICAL	MINOR



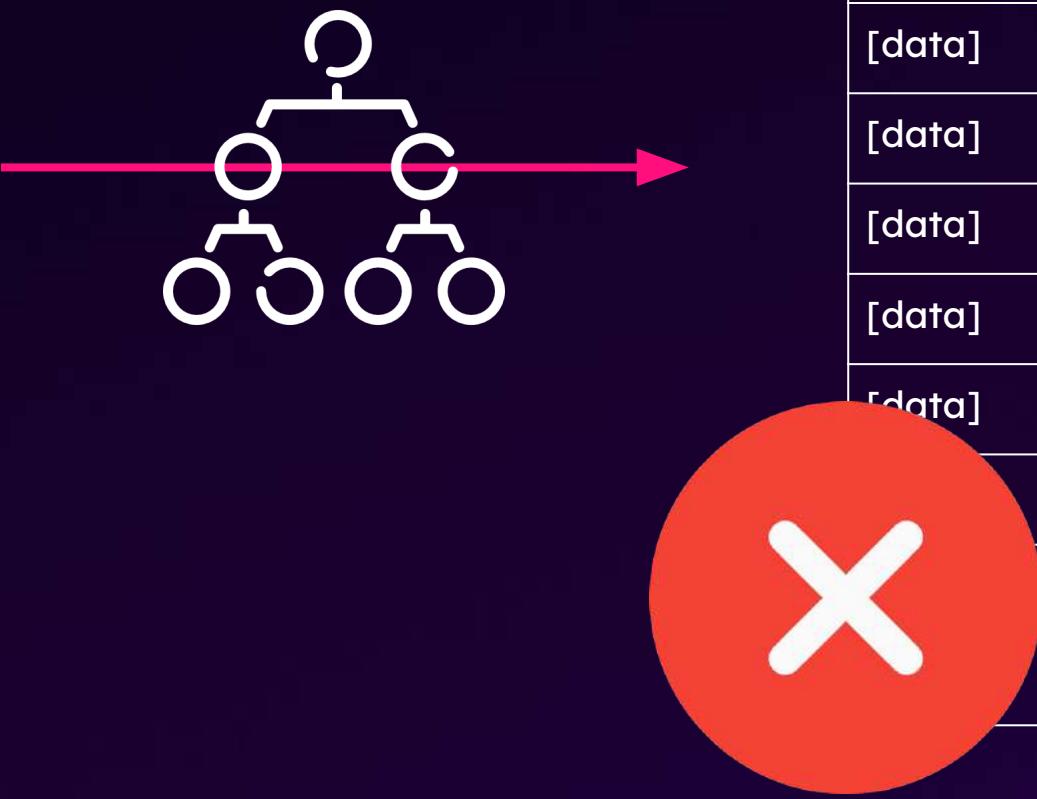
How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL

New Data

Feature 1	...	Feature n	Severity (predicted)	Severity (actual)
[data]	...	[data]	CRITICAL	MAJOR
[data]	...	[data]	CRITICAL	LOW
[data]	...	[data]	CRITICAL	MINOR
[data]	...	[data]	CRITICAL	CRITICAL
[data]	...	[data]	CRITICAL	LOW
[data]	...	[data]	CRITICAL	MINOR
...	...	[data]	CRITICAL	MAJOR
...	...	[data]	CRITICAL	MAJOR
...	...	[data]	CRITICAL	MINOR



How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	MAJOR
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MINOR

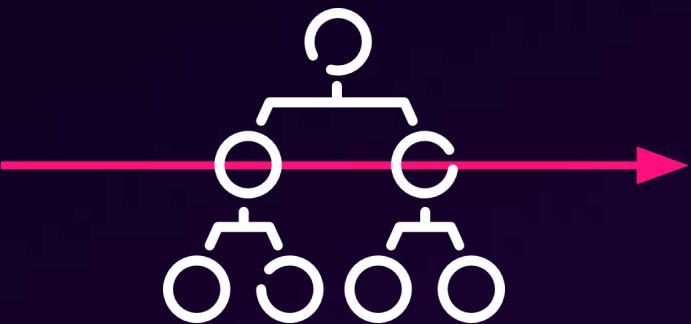
How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	MAJOR
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MINOR

New Data

Feature 1	...	Feature n	Severity (predicted)
[data]	...	[data]	LOW
[data]	...	[data]	MAJOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MINOR
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	LOW



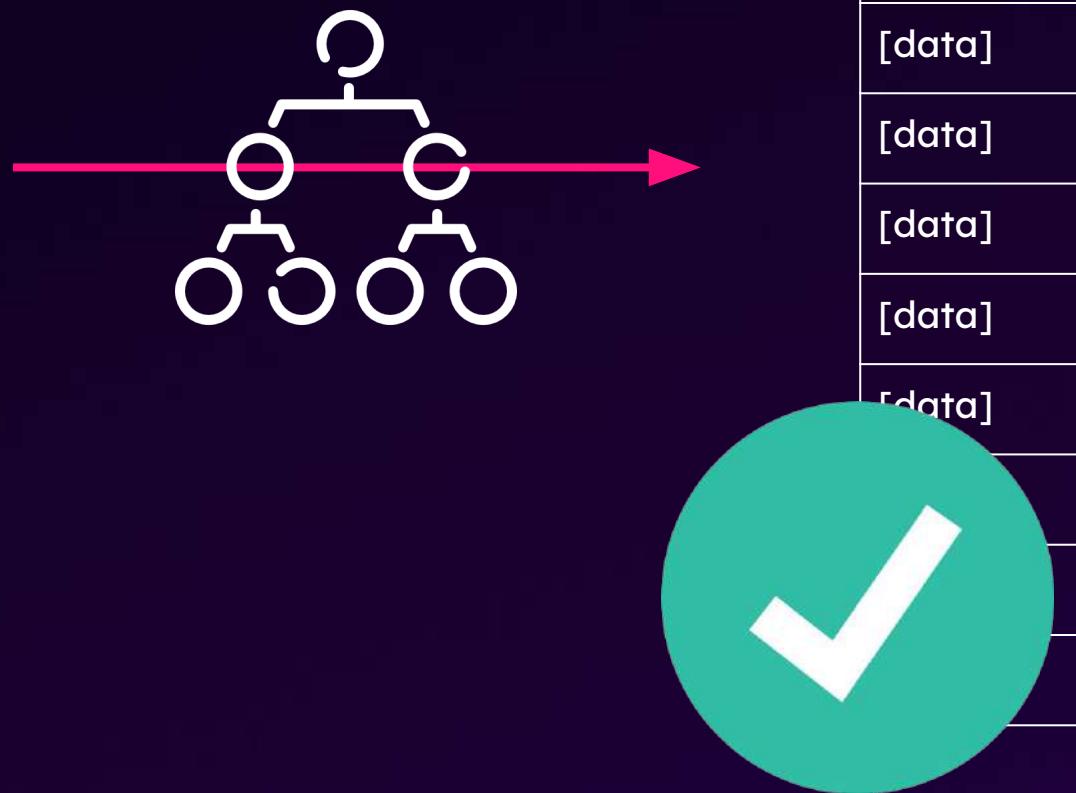
How Does Categorical Prediction Work?

Training Data

Feature 1	...	Feature n	Severity
[data]	...	[data]	MAJOR
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	CRITICAL
[data]	...	[data]	LOW
[data]	...	[data]	MINOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MINOR

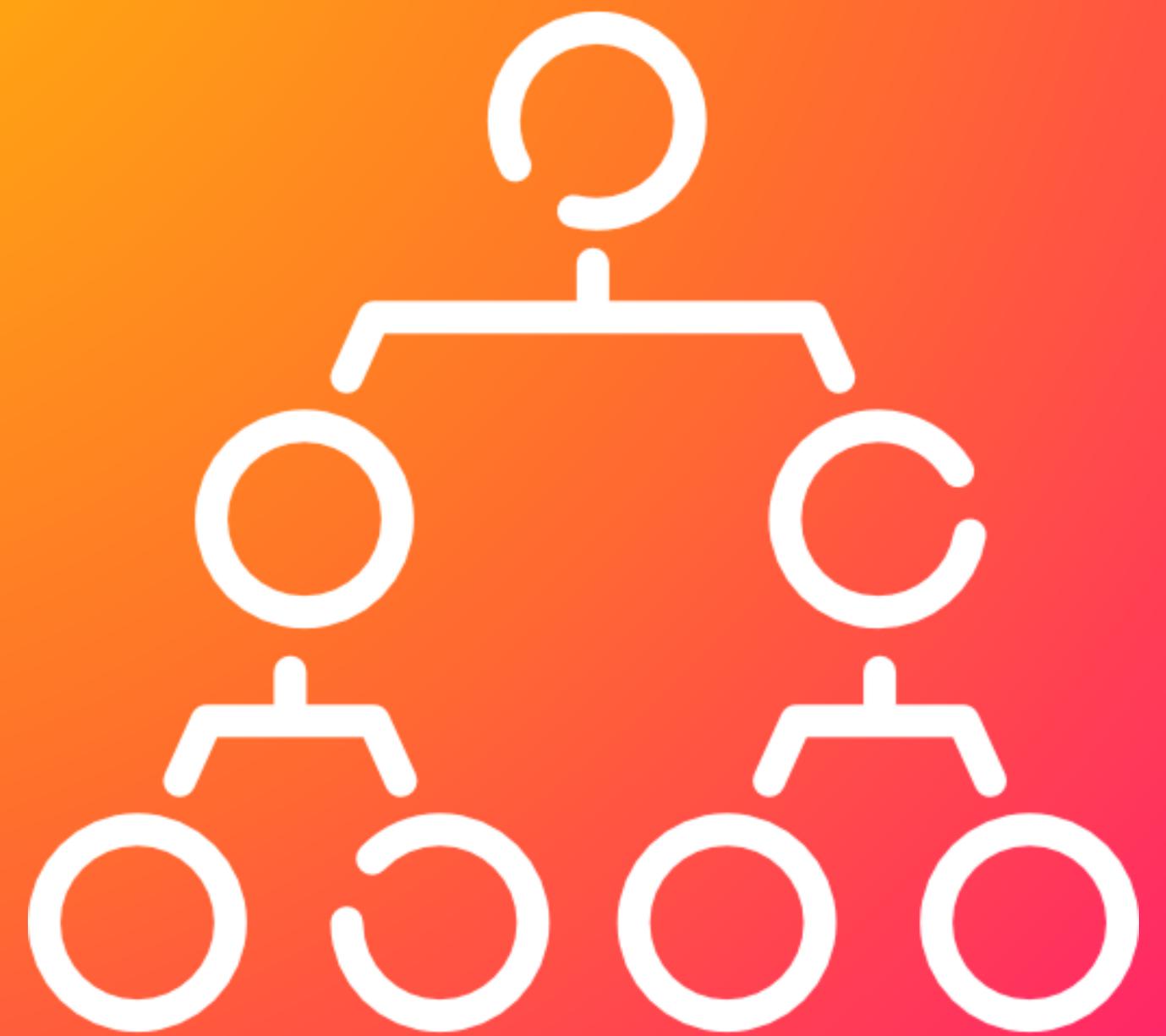
New Data

Feature 1	...	Feature n	Severity (predicted)
[data]	...	[data]	LOW
[data]	...	[data]	MAJOR
[data]	...	[data]	MAJOR
[data]	...	[data]	MINOR
[data]	...	[data]	CRITICAL
[data]	...	[data]	CRITICAL
...	[data]	...	LOW
...	[data]	...	MINOR
...	[data]	...	LOW



Categorical Prediction

“A method that lets you quickly, easily, and sustainably gain insight into your data by predicting its categorical features”



Live Instance Demo

Exercise 2

Predict presence of Malware

Time: 15 minutes



Showcase

Welcome to the Machine Learning Toolkit Showcase. Watch and learn from interactive end-to-end examples using real datasets. Click on an example to pre-populate the Assistant with the sample dataset and its settings. Inspect the Search Processing Language as well as the underlying code of these examples to see how it all works.

View examples by

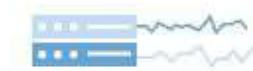
ML Operation

Industry

Filter Examples



IT



View examples geared towards IT administrators using data sets that include server power consumption, hard drive failures, server response time, and VPN usage.

15 Examples Available

Security



View examples geared towards security analysts using data sets that include employee logins, VPN usage, Bitcoin transactions, and app usage.

7 Examples Available

Business Analytics



View examples geared towards business analysts using data sets that include house values, sales, purchases, and customer churn.

11 Examples Available

IoT



View examples geared towards analysts and IoT administrators using data sets that include power plant energy output and humidity, vehicle make and model, and mobile phone activity.

8 Examples Available

Featured Examples

Cluster Behavior by App Usage

This example uses the Cluster Numeric Events Assistant, a preprocessing step using the Standard Scaler method, and the Spectral Clustering algorithm to cluster on four different fields.

Cluster Numeric Fields

Detect Outliers in Bitcoin Transactions

This example uses the Detect Categorical Outliers Assistant on three fields of data that include user and value.

Detect Categorical Outliers

Predict the Presence of Vulnerabilities

This example uses the Smart Prediction Assistant and the AutoPrediction algorithm to predict vulnerabilities in firewall data from fields in the data including bytes received, bytes sent, packets received, and packets sent.

Smart Prediction

Predict VPN Usage

This example uses the Predict Numeric Fields Assistant and the Linear Regression algorithm to predict remote access to the VPN from four other fields.

Predict Numeric Fields

Detect Outliers in Number of Logins (vs. Predicted Value)

This example uses the Detect Numeric Outliers Assistant and threshold method of Median Absolute Deviation to look for outliers in login information.

Detect Numeric Outliers

Predict the Presence of Malware

This example uses the Predict Categorical Fields Assistant and the Logistic Regression algorithm to predict malware from fields including bytes received, bytes sent, and destination port.

Predict Categorical Fields

How is the prediction working by default?

The screenshot shows a machine learning configuration interface. At the top, the algorithm is set to "LogisticRegression" and the field to predict is "used_by_malware". The fields used for prediction include "bytes_sent, bytes_re..." (7). The training/test split is set at 70 / 30. A "Fit Intercept" checkbox is checked, and the model is saved as "example_malware". Below these are buttons for "Fit Model", "Schedule Training", "Open in Search", and "Show S". The "Prediction Results" section displays a table comparing "used_by_malware" (actual) and "predicted(used_by_malware)" (predicted) for various entries. The table includes columns for "bytes_sent" and "bytes_received".

used_by_malware	predicted(used_by_malware)
yes	yes
no	no
yes	yes
no	no
no	yes
no	no

Fields to use for predicting:
bytes_sent, bytes_re... (7)
bytes_received
bytes_sent
dest_port
has_known_vulnerability
packets_received
packets_sent
src_port
dst_ip
receive_time
serial_number
session_id
src_ip

Split for training / test: 70 / 30

Fit Intercept:

Save the model as: example_malware

Fit Model | Schedule Training | Open in Search | Show S

Prediction Results

used_by_malware	predicted(used_by_malware)	bytes_sent	bytes_received
yes	yes	75	
no	no	1872	
yes	yes	1331	
no	no	1758	
no	yes	817	
no	no	81	

Currently using the LogisticRegression algorithm. Do we want to change it?

We are predicting if this will be used by malware.
Base on what? ↓

- number of bytes (sent and received)
- ports (source and destination)
- packets (sent and received)
- has_known_vulnerability

We are splitting our data - 70% for learning and 30% for testing.

Do you have additional fields that could contribute to make the prediction more reliable?

How do we validate our inputs?



The input values should be adjusted until Precision, Recall, Accuracy and F1 are almost 1.00. The values in the confusion matrix should be on the 95~99.9% range.

Why shouldn't these values be 1.00, or 100%?

This could mean that our data is not varied enough, that we have a wrong split of data (for example, 100/0), ...

The future results of our algorithm might become unreliable if data sample is considered to be 100% certain.

Summary

Top 4 most important things to remember about categorical prediction

1

Predicting categorical fields is done using a **supervised learning** method which uses **labeled data**

2

Models **assumes there exists a pattern** determining existing categories

3

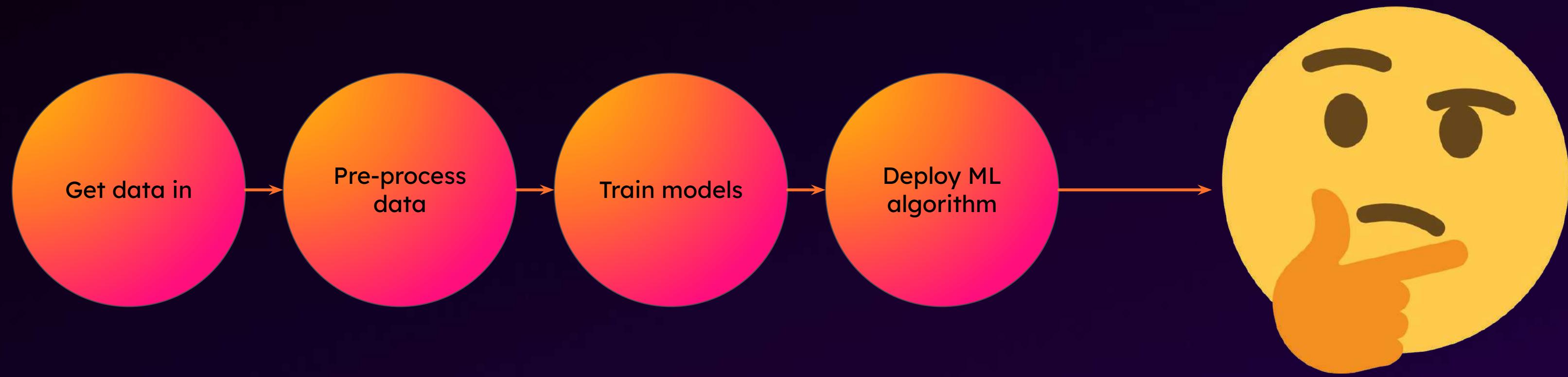
Scaling data prior to training is **often necessary**

4

Choice of categorical prediction algorithm(s) may rely on a **subject matter expert** of the data

Use Case: Operationalize ML in ES

What do we do with the detections?



Create a detection in ES

Create detection

Test and validate

Deploy

Create detection logic and underlying search, specify detection details

Test and validate detection against historical data, in a test environment, etc

Deploy detection to production environment, and start receiving alerts

Live Instance Demo

Exercise 3

Operationalize ML in ES

Time: 15 minutes

And we're done!

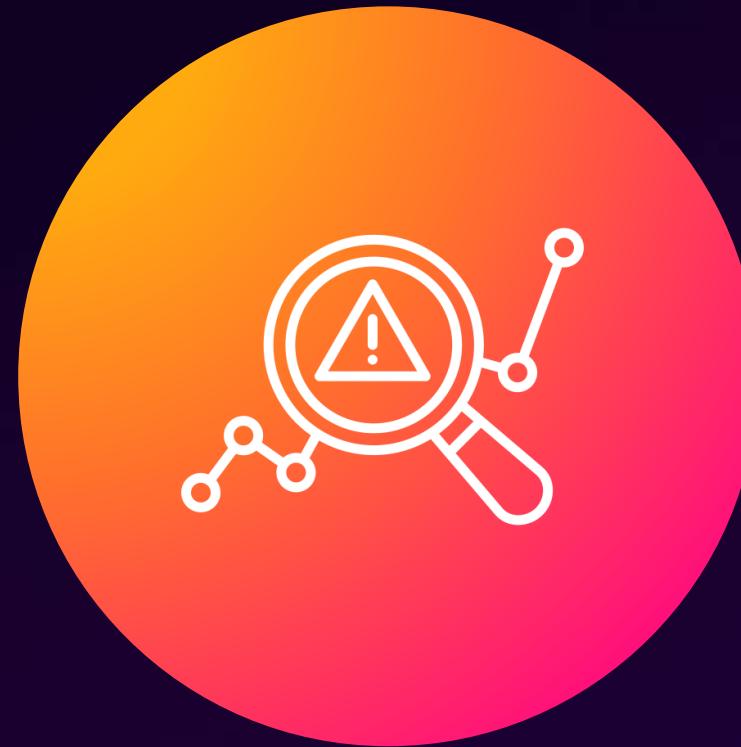
Summary

Tools in Your ML Toolkit Now



Prediction

Get ahead of issues that may happen in the future



Outlier Detection

Identify and analyze abnormal behavior in your data



Numeric Prediction

“A method that lets you model and predict how a metric will behave based on changes in the environment”

Clustering

“A method that organizes a set of numeric data points in a way that objects in the same cluster are more similar to each other than those in other clusters”

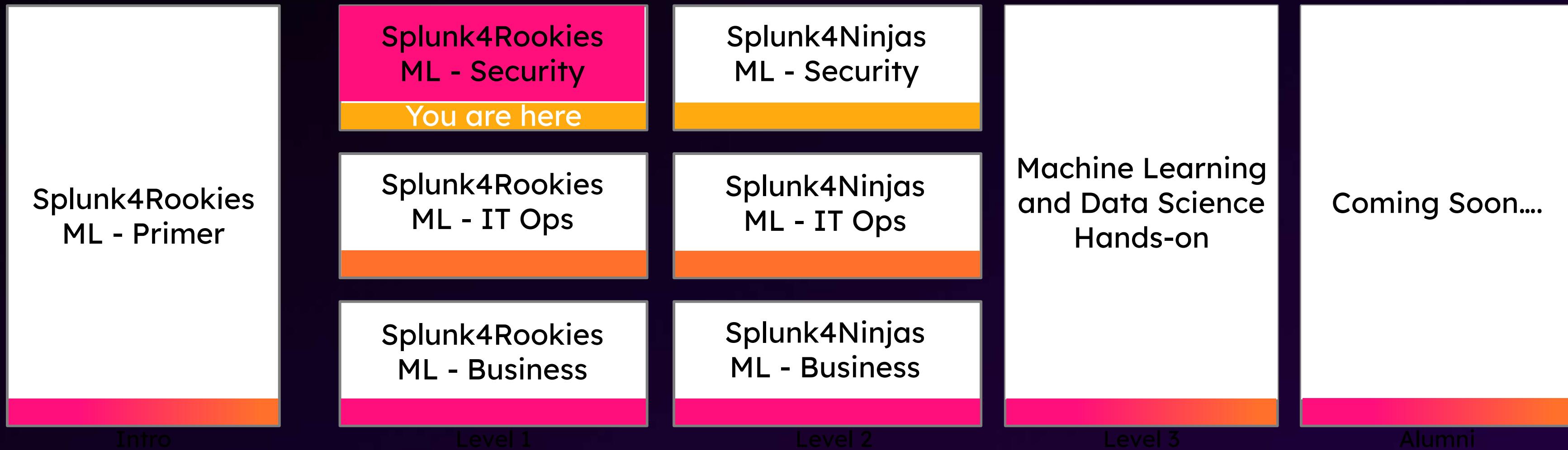


Forecasting Time Series

“Using historical data to identify patterns, which are then used to forecast how your data might behave in the future”



Splunk Machine Learning Workshops



Boss of Machine Learning (Coming Soon)

Introductory

Advanced

Legend

- Security
- IT Ops
- Business
- All

Additional Resources

Getting started

- View some of our [webinars](#)
- Check out our YouTube [playlist](#)
- Check out the blog on [MLTK 5.4 release](#)
- Try out some of our starter blogs, such as [Cyclical Statistical Forecasts and Anomalies, part 1](#)
- Try our new [MLTK Deep Dives](#)

Increasing complexity

- Try [part 4 or 6](#) of the Cyclical Statistical Forecasts and Anomalies series
- Brush up on how MLTK works with our [comprehensive documentation](#)
- Get familiar with the [Workshop Guide](#)

More advanced

- The [Analytics and Data Science course](#)
- Try out the [Anomalies Are Like a Gallon of Neapolitan Ice Cream - Part 1](#)
- Try out [part 5](#) of the cyclical statistical forecasts and anomalies series
- Try the [ML-SPL API](#)

Thank you!

Q&A