

Machine Learning

Splunk4Rookies | Primer

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 February 20, 2024 and its most recent report on Form 10-K filed on September 7, 2023, as well as the "Risk Factors" section of Splunk's most recent report on Form 10-Q filed with the SEC on November 28, 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, Splunk> and Turn Data Into Doing are trademarks and registered trademarks of Splunk LLC 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 AI Adoption Journey

Legend



Strategy

Planning



Implementation

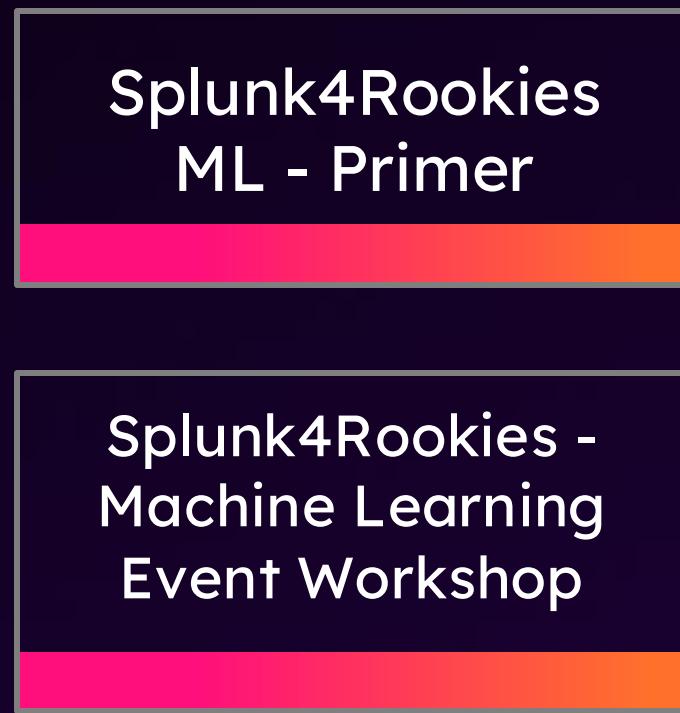


Review

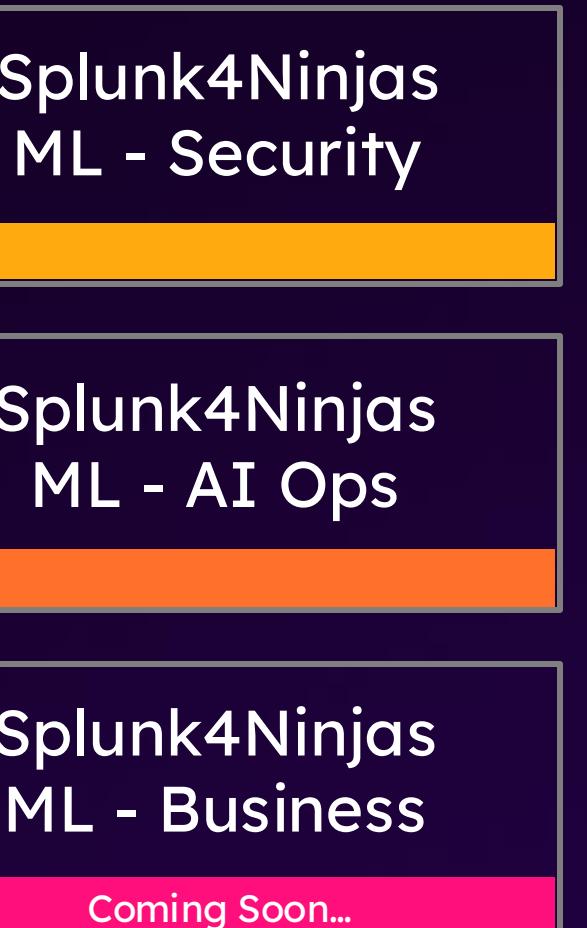
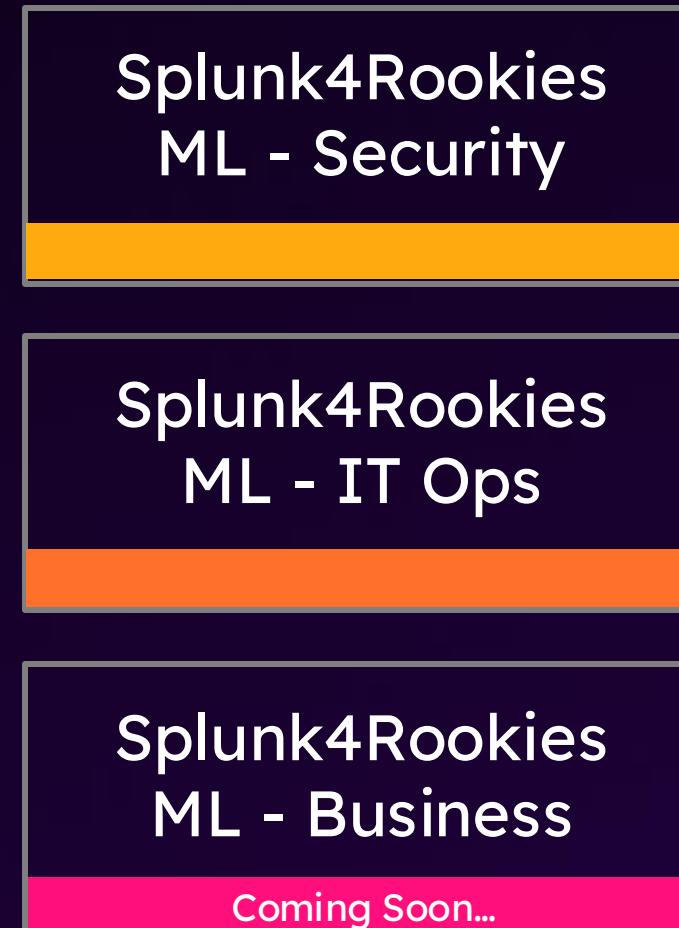


Hands-on Enablement

Level 0



Level 1



Level 2



Agenda

- ❑ What is Machine Learning
- ❑ How Splunk Drives Machine Learning
- ❑ Hands-on
- ❑ Wrap Up and Next Steps

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%

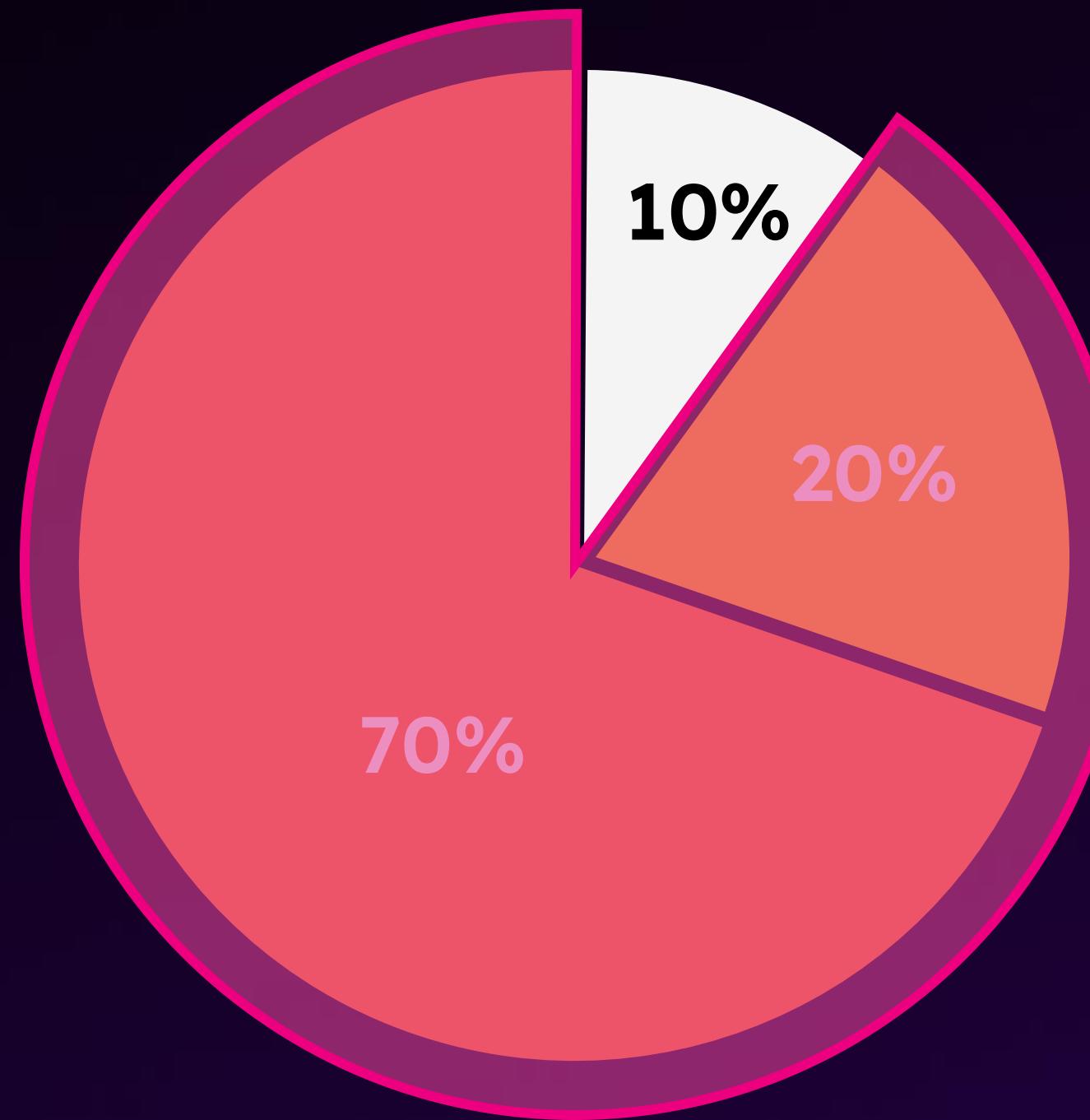
Failure rate for companies
which try to implement machine
learning from scratch, due
to **lack of expertise**

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/>

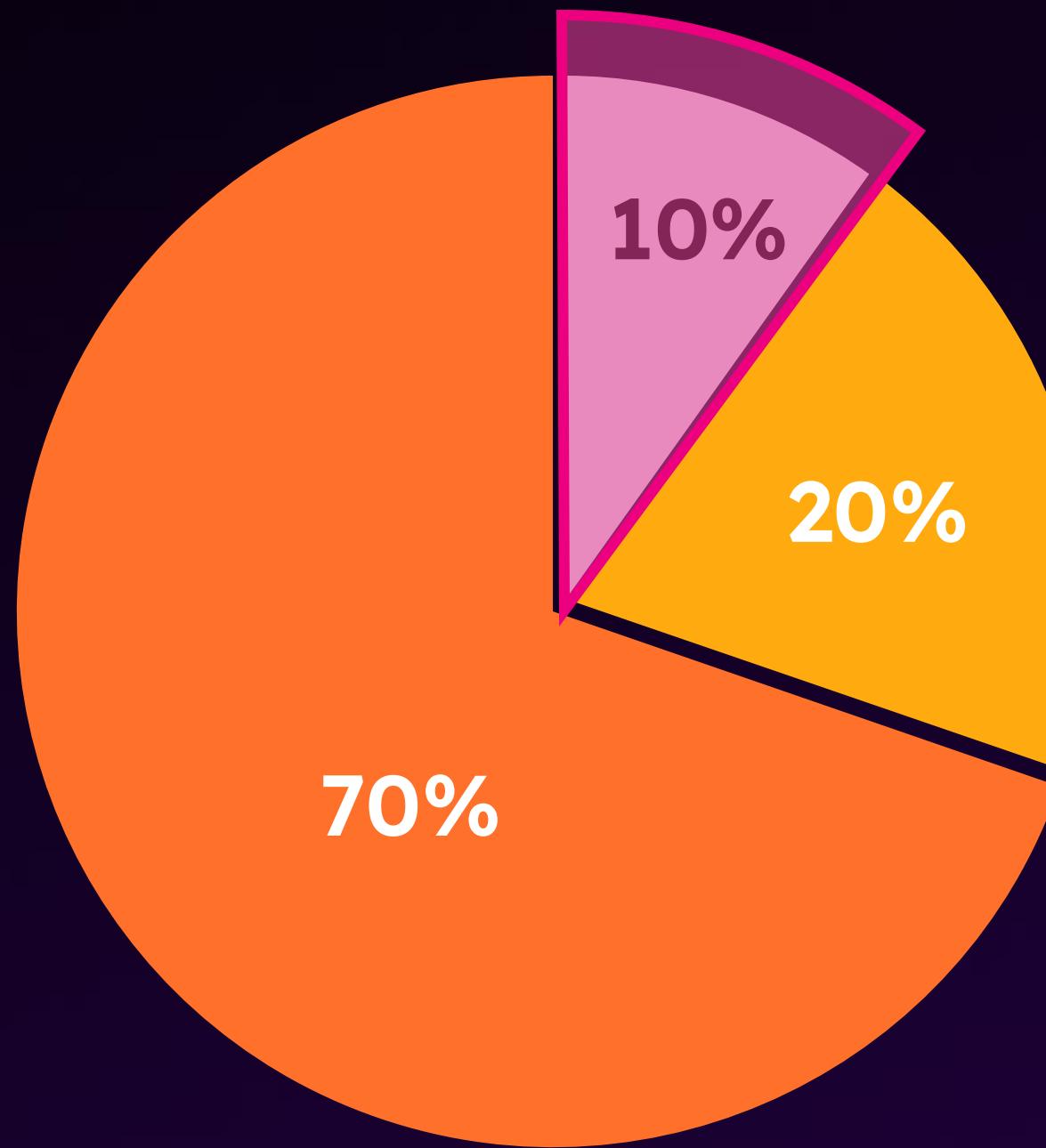
How Data Scientists Spend Their Time



- **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".

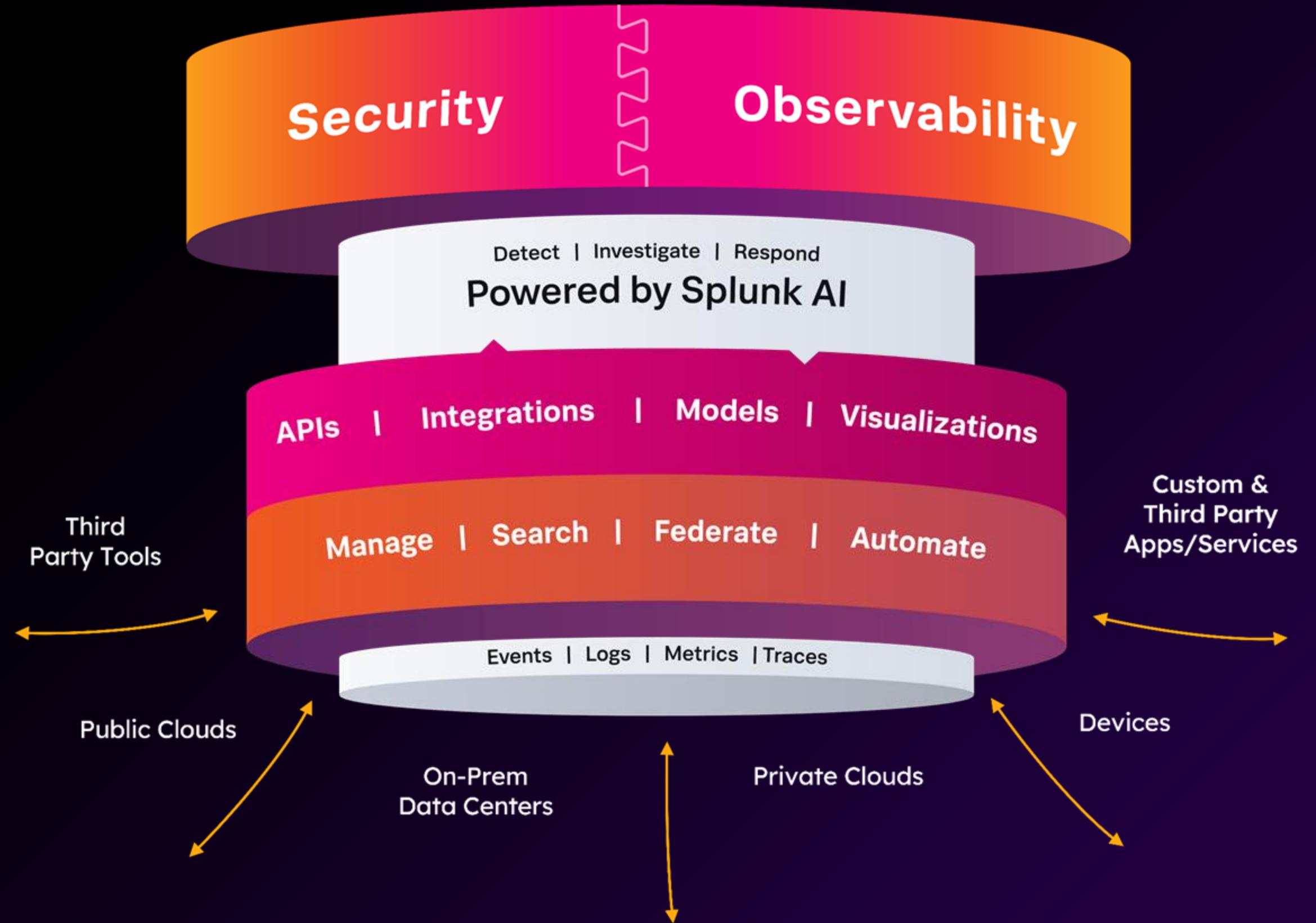
How Data Scientists Spend Their Time



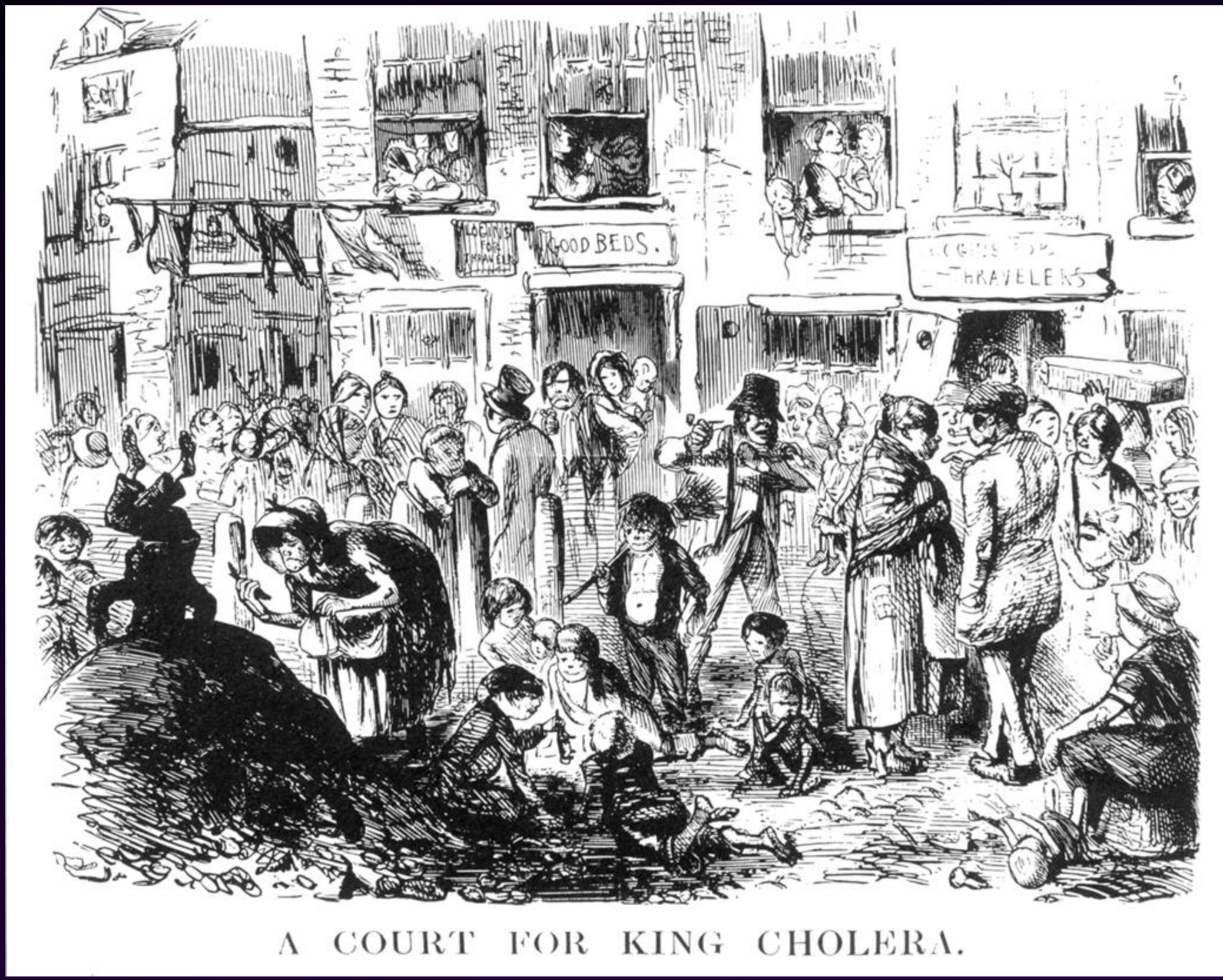
- **Data Engineering**
- **Machine Learning**
- **Other**

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

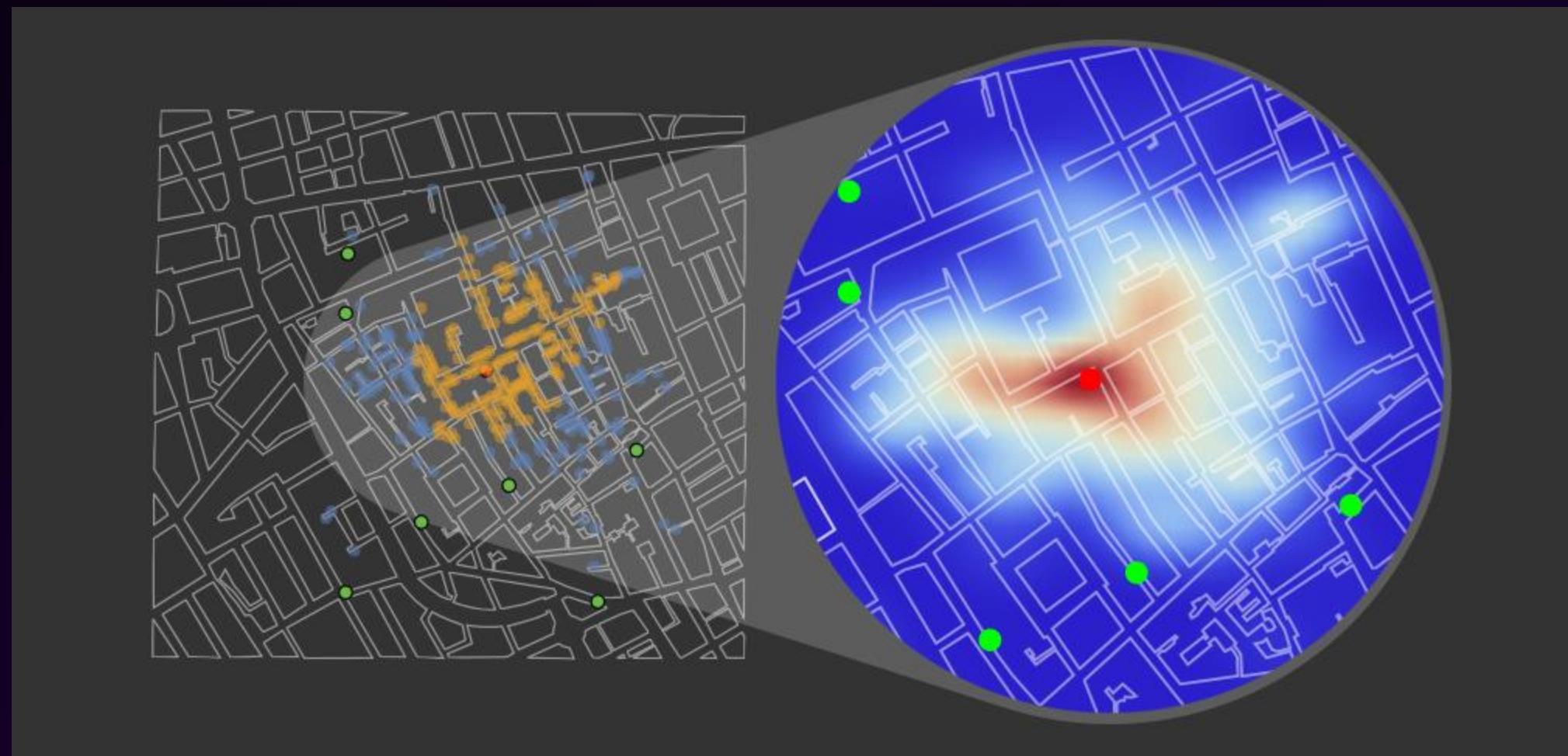
The Unified Security and Observability Platform



Why Data Science?



Clustering

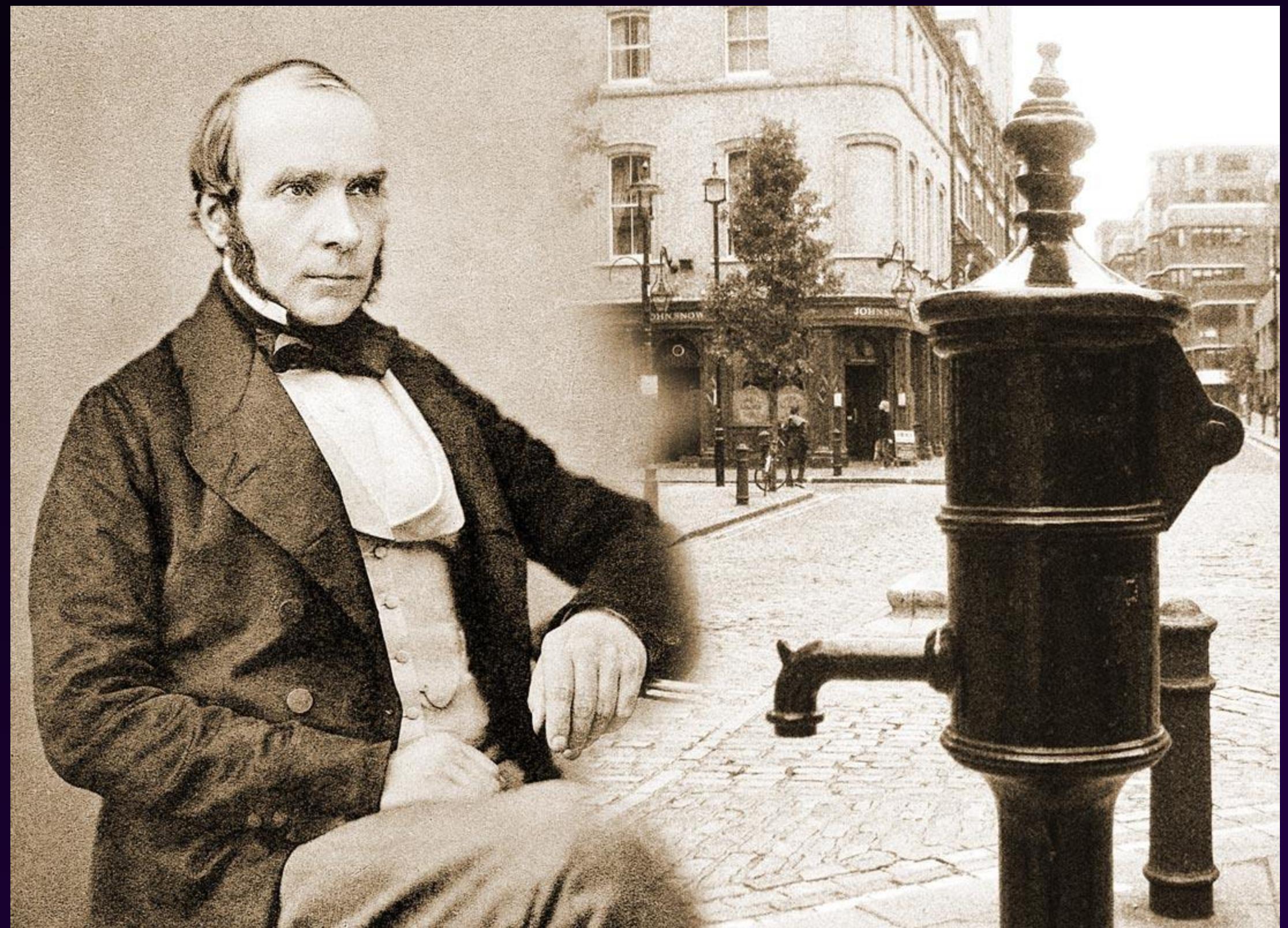


To prove his theory, he persuaded local officials to remove the pump handle. Soon after, the outbreak subsided. While the removal may not have halted the epidemic entirely — as some people had already fled or were using other water sources — it provided compelling evidence that cholera was waterborne, not airborne.

This was one of the earliest and most powerful examples of using **data visualization and spatial analysis** to influence public health policy. Snow's map wasn't just a visual — it told a **story**. It connected data to human lives and changed the course of epidemiology.

Today, Snow is considered one of the fathers of both **modern epidemiology and data science**. His work reminds us that data, when grounded in observation and mapped meaningfully, has the power to challenge assumptions, uncover hidden patterns, and save lives.

The source!



To support his theory, he persuaded local officials to remove the pump handle. Soon after, the outbreak subsided. While the removal may not have halted the epidemic entirely — as some people had already fled or were using other water sources — it provided compelling evidence that cholera was waterborne, not airborne.

This was one of the earliest and most powerful examples of using **data visualization and spatial analysis** to influence public health policy. Snow's map wasn't just a visual — it told a **story**. It connected data to human lives and changed the course of epidemiology.

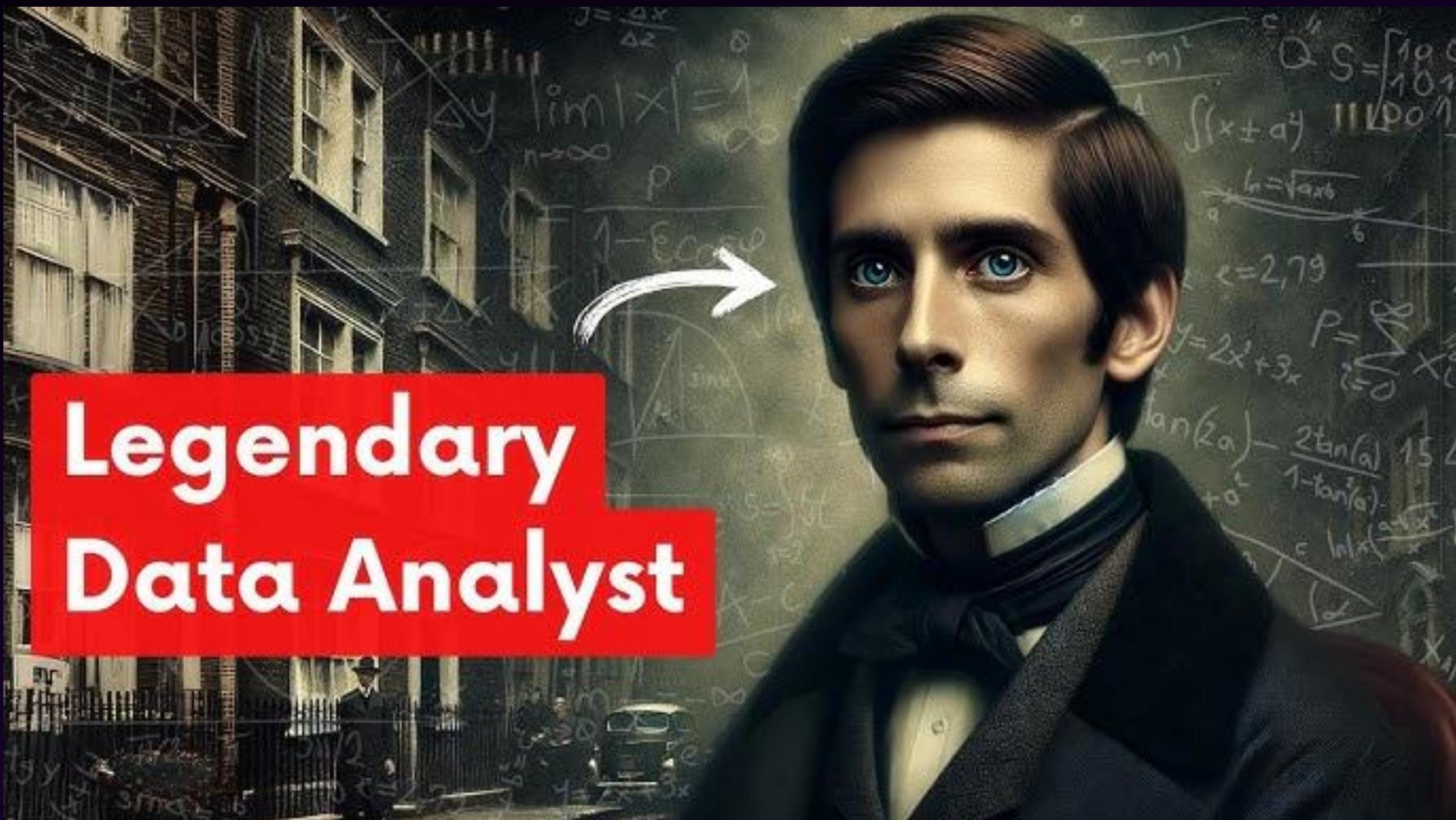
Today, Snow is considered one of the fathers of both **modern epidemiology and data science**. His work reminds us that data, when grounded in observation and mapped meaningfully, has the power to challenge assumptions, uncover hidden patterns, and save lives.



To support his theory, he persuaded local officials to remove the pump handle. Soon after, the outbreak subsided. While the removal may not have halted the epidemic entirely — as some people had already fled or were using other water sources — it provided compelling evidence that cholera was waterborne, not airborne.

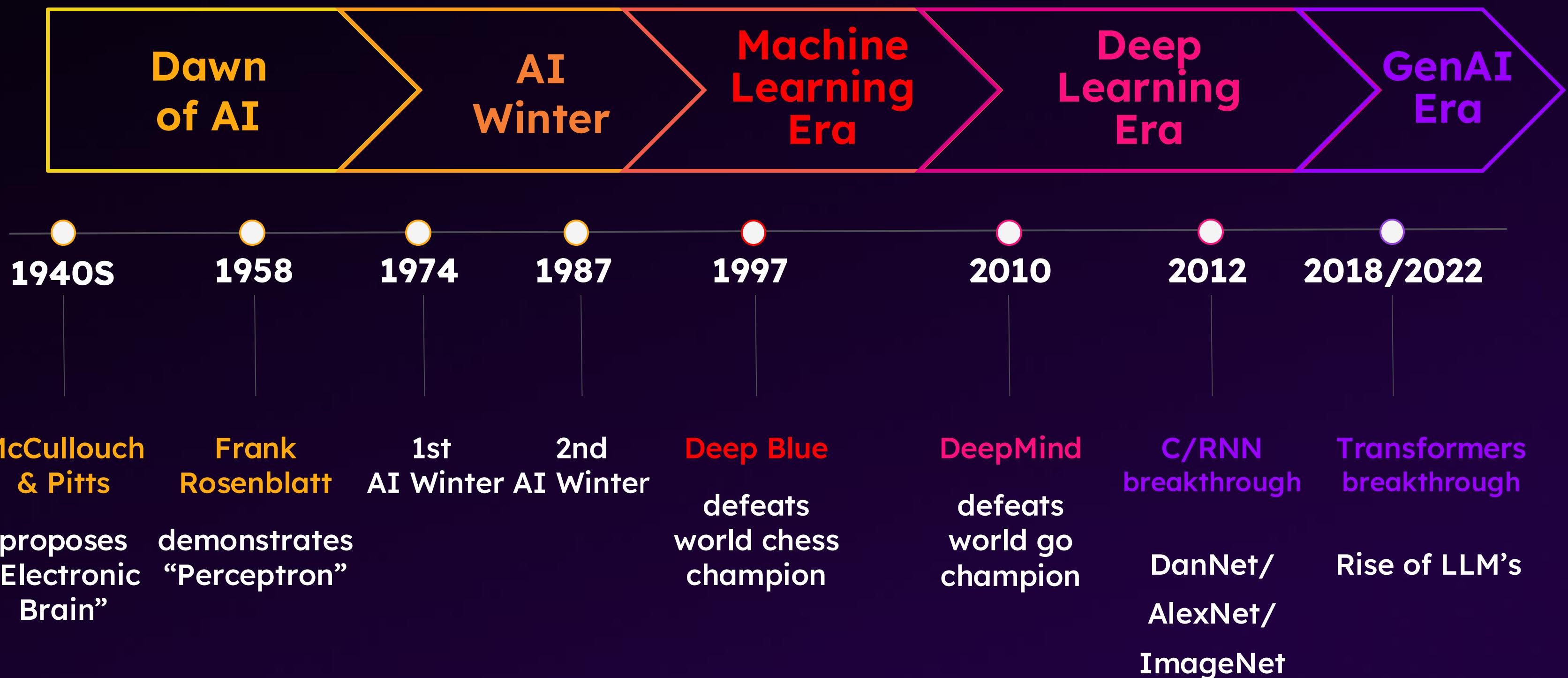
This was one of the earliest and most powerful examples of using **data visualization and spatial analysis** to influence public health policy. Snow's map wasn't just a visual — it told a **story**. It connected data to human lives and changed the course of epidemiology.

Today, Snow is considered one of the fathers of both **modern epidemiology and data science**. His work reminds us that data, when grounded in observation and mapped meaningfully, has the power to challenge assumptions, uncover hidden patterns, and save lives.



Legendary Data Analyst

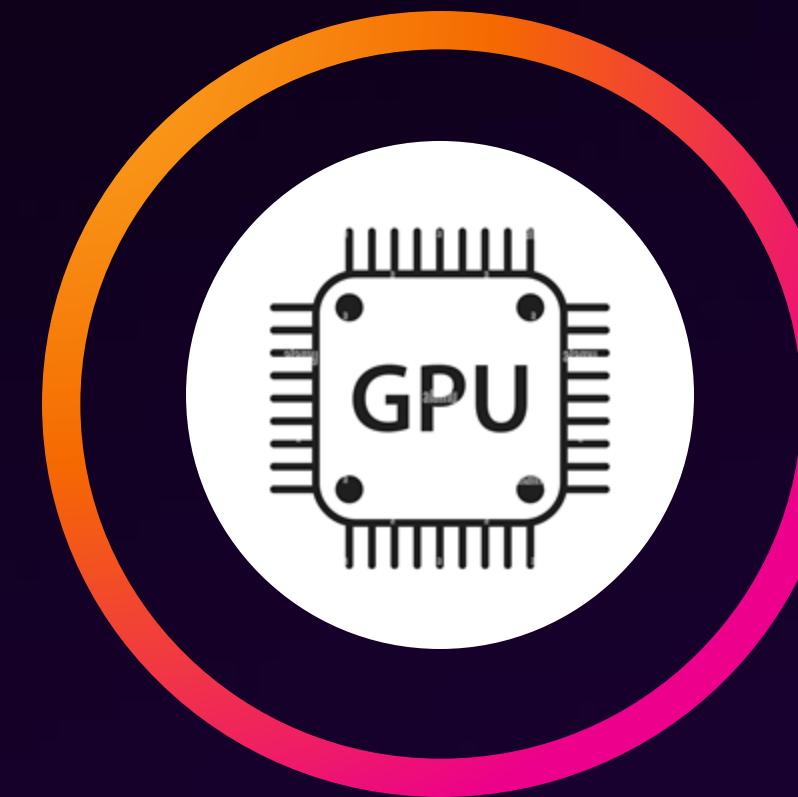
History of AI



What is driving the explosive *growth* of AI?



Algorithm

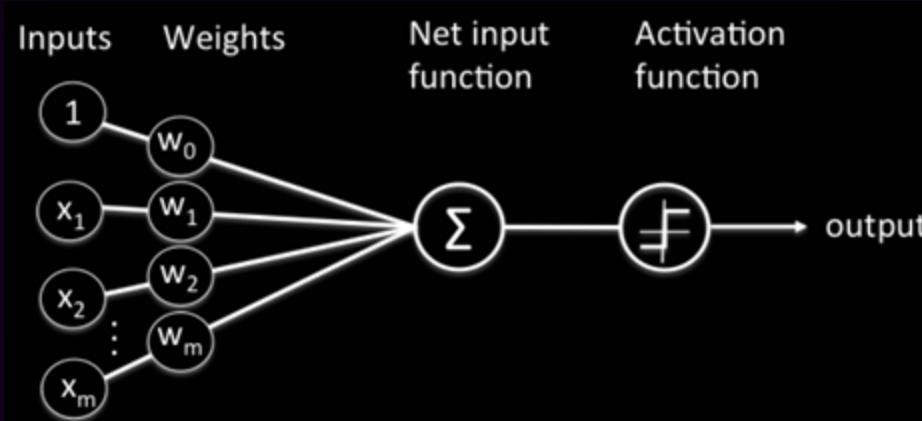


Compute

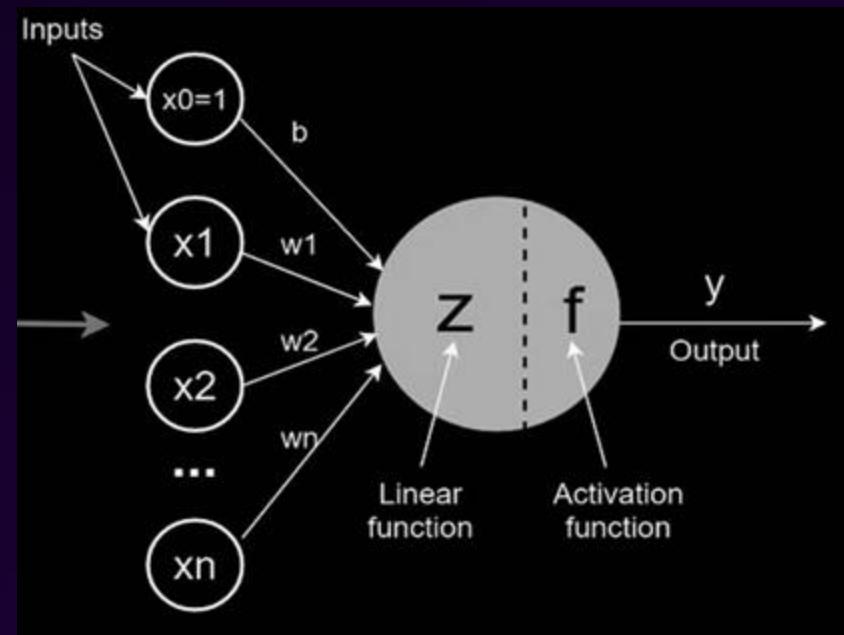


Data

What is driving the explosive growth of AI?

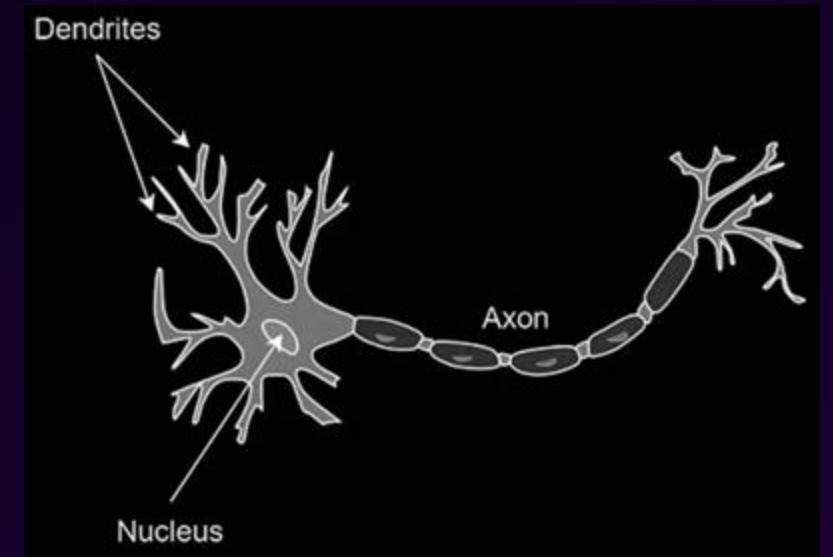


Perceptron



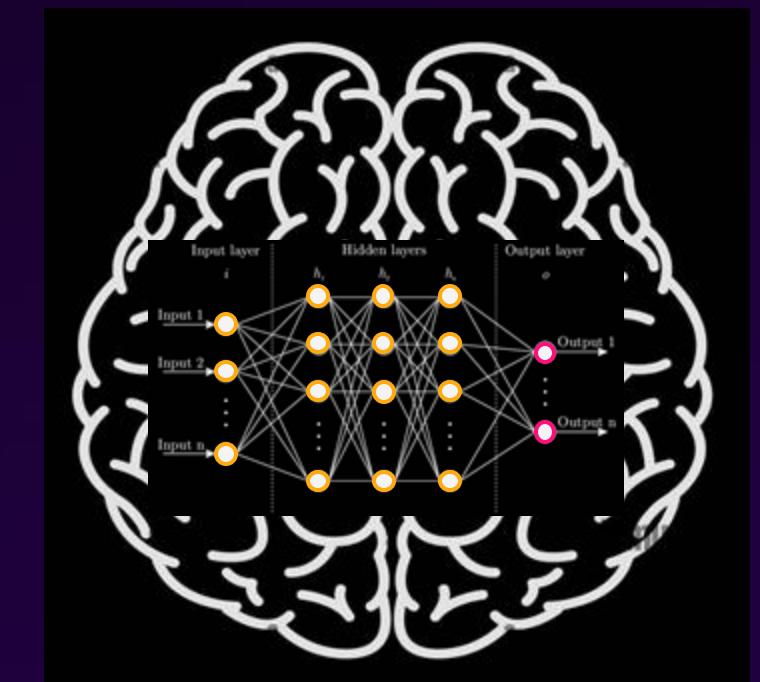
Artificial Neuron

1950's



Biological Neuron

2010's



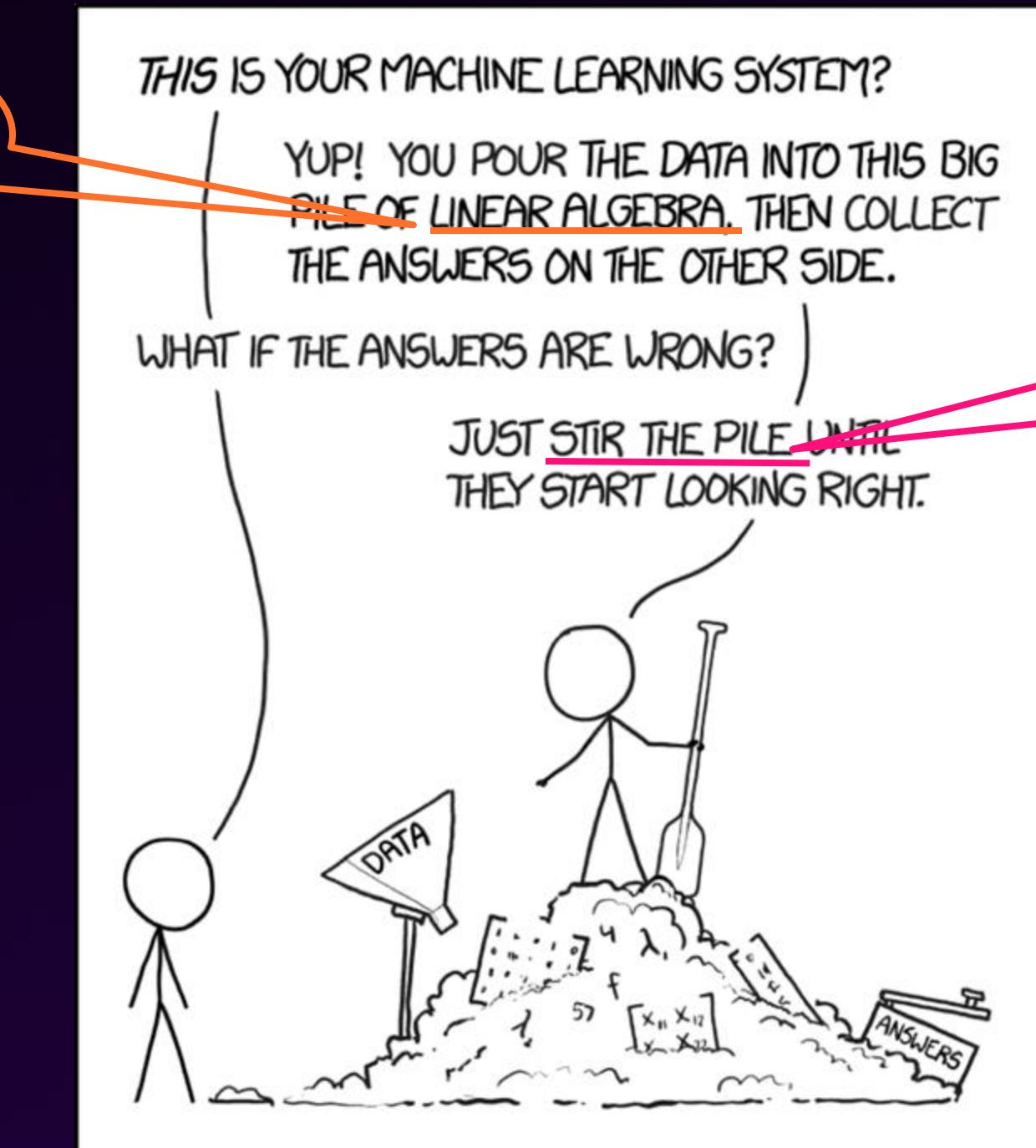
Deep Neural Network

What is driving the explosive growth of AI?



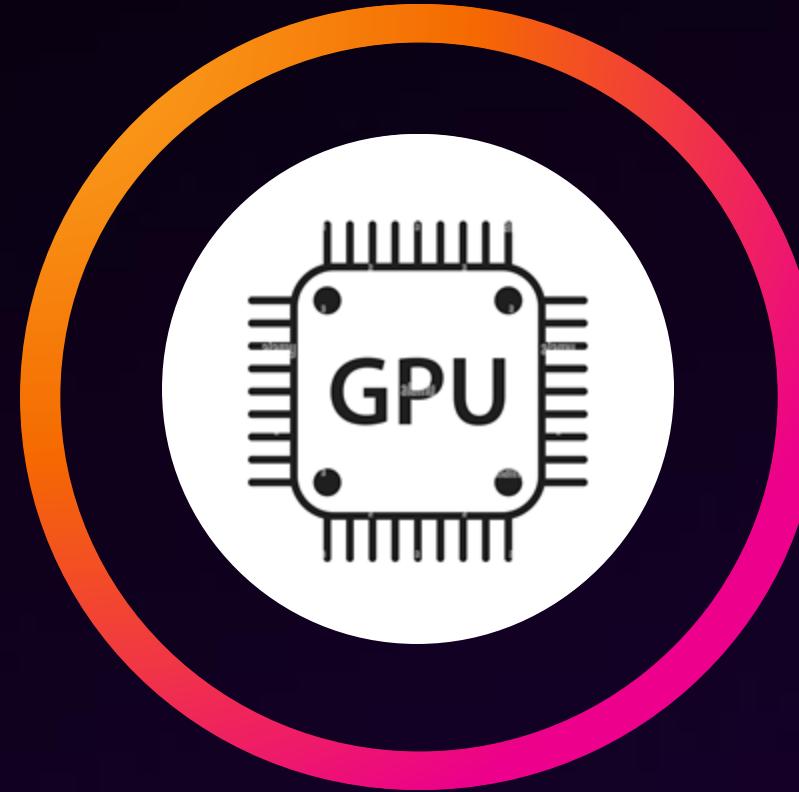
Algorithm

Matrix
Multiplication

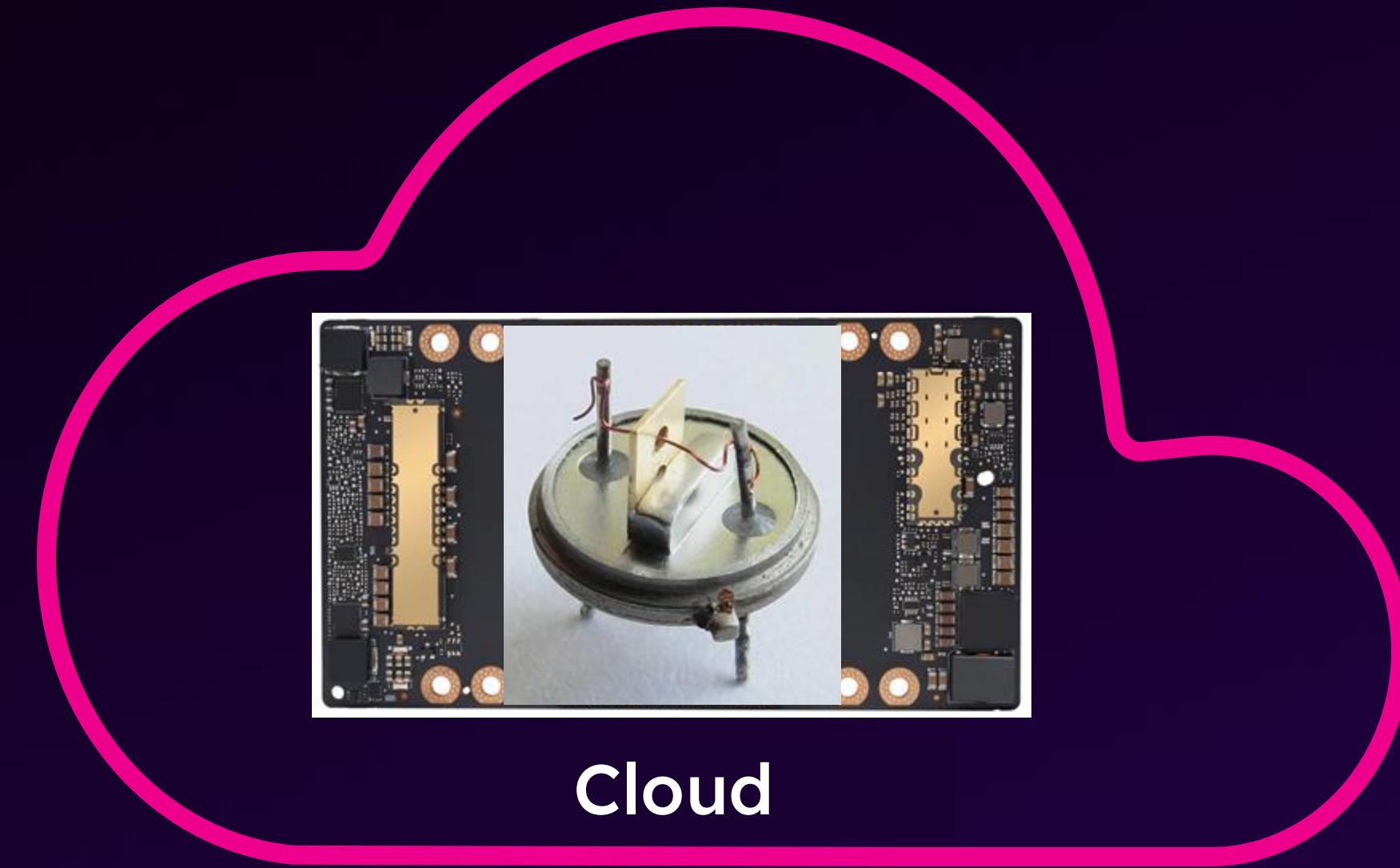


Backpropagation

What is driving the explosive growth of AI?



Compute



What is driving the explosive growth of AI?



Data



WIKIPEDIA
The Free Encyclopedia



Google



Different Levels of Teaching Machines



Artificial Intelligence

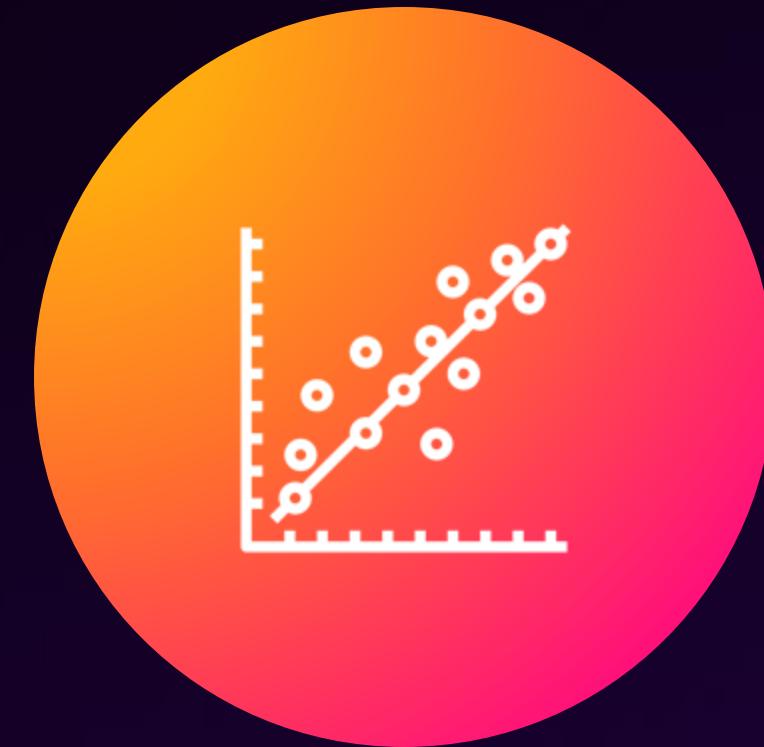
The broad study of teaching a computer
to process data and make decisions

Different Levels of Teaching Machines



Artificial Intelligence

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



Machine Learning

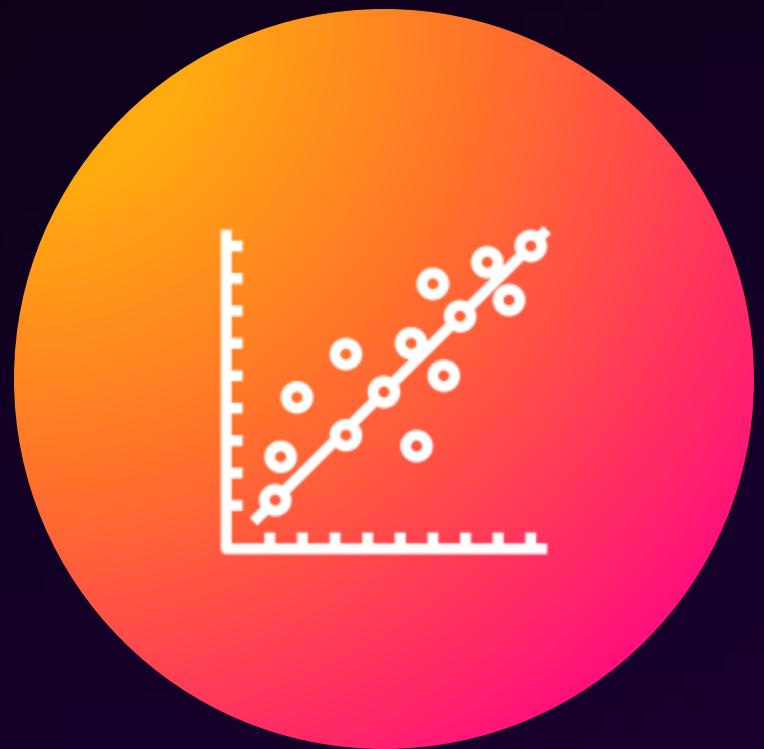
Subset of AI. Predictions and insight with minimal human interference

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

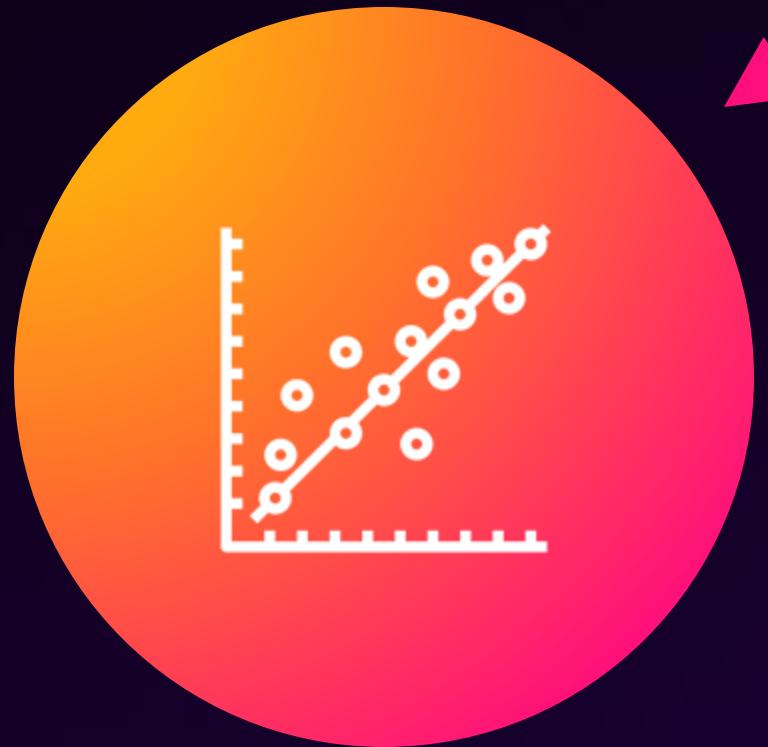
Subset of ML. Predictions via neural networks

Different Levels of Teaching Machines



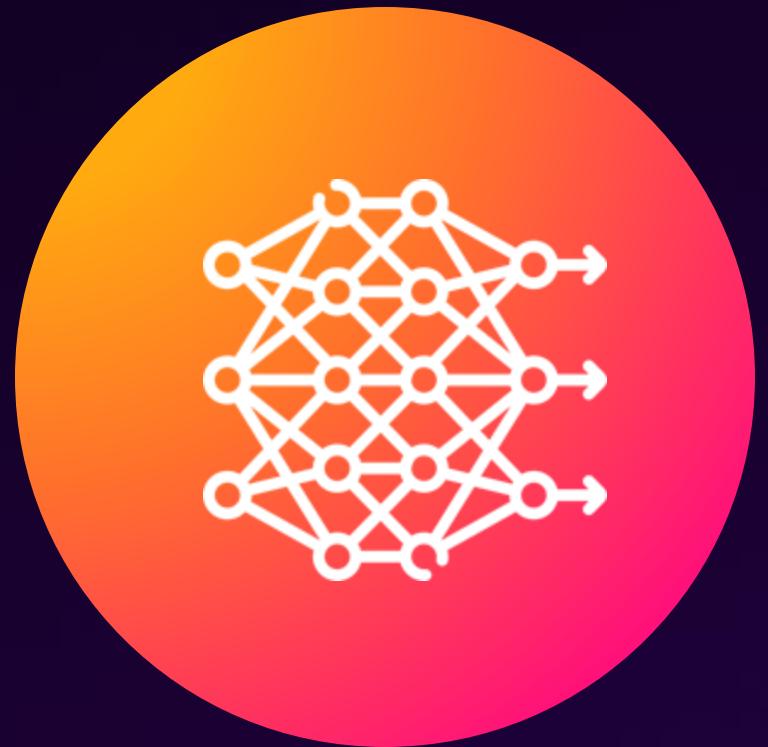
Artificial Intelligence

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

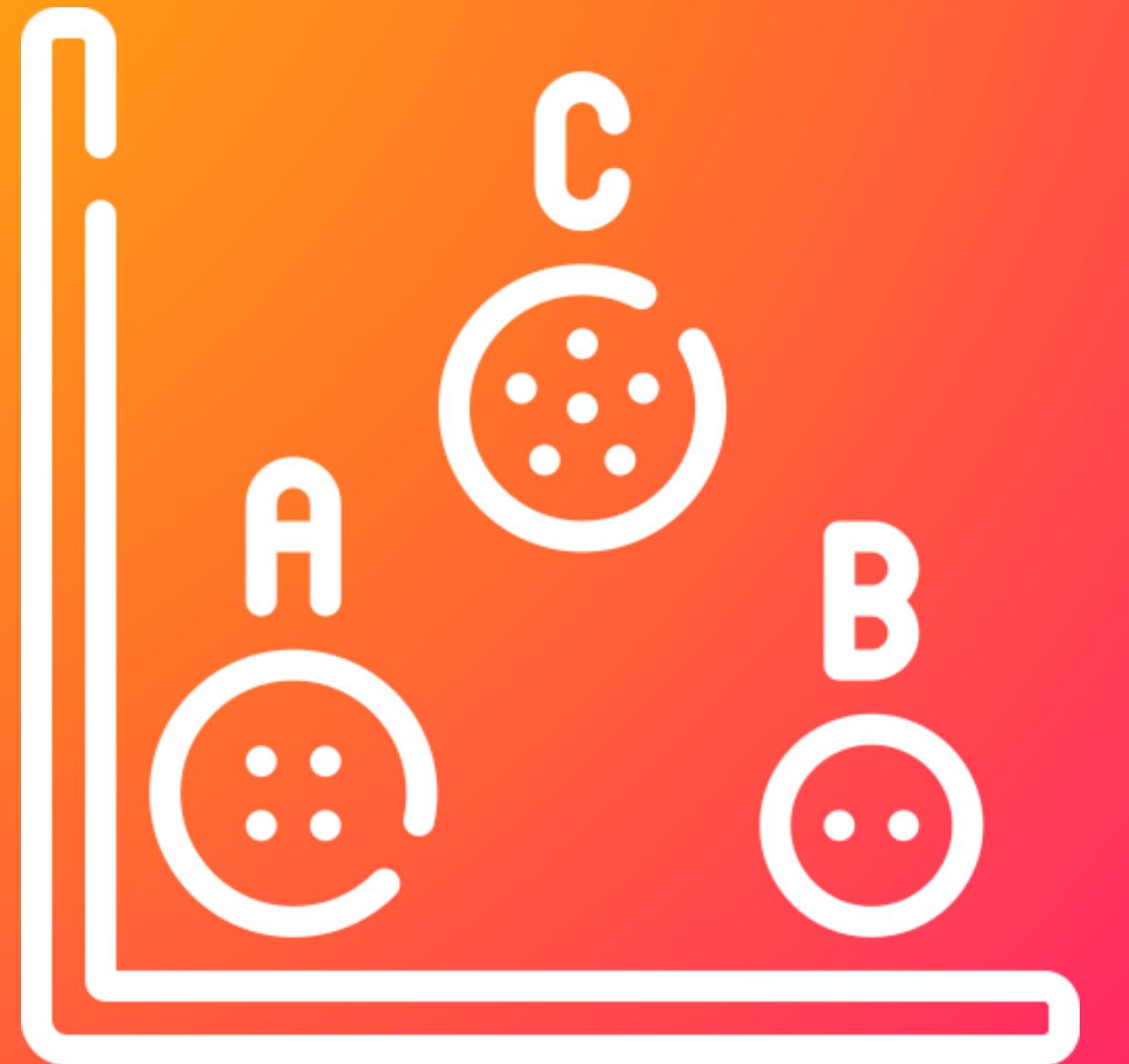
Predictive Algorithms

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



Source:Hajicon, flaticon.com

Splunk ML & AI

Where to find ML



CORE PLATFORM
SEARCH



PACKAGED (PREMIUM)
SOLUTIONS

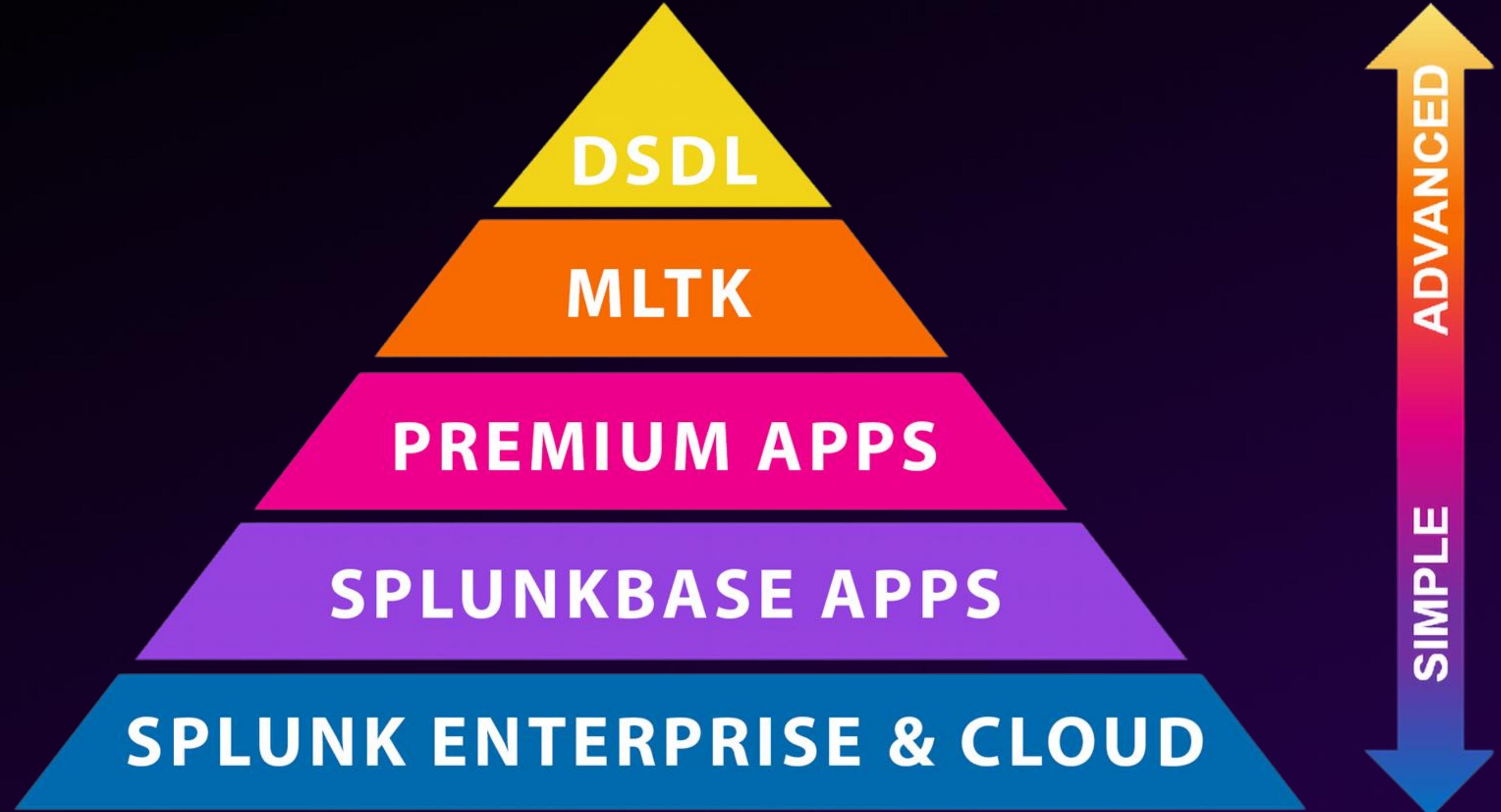


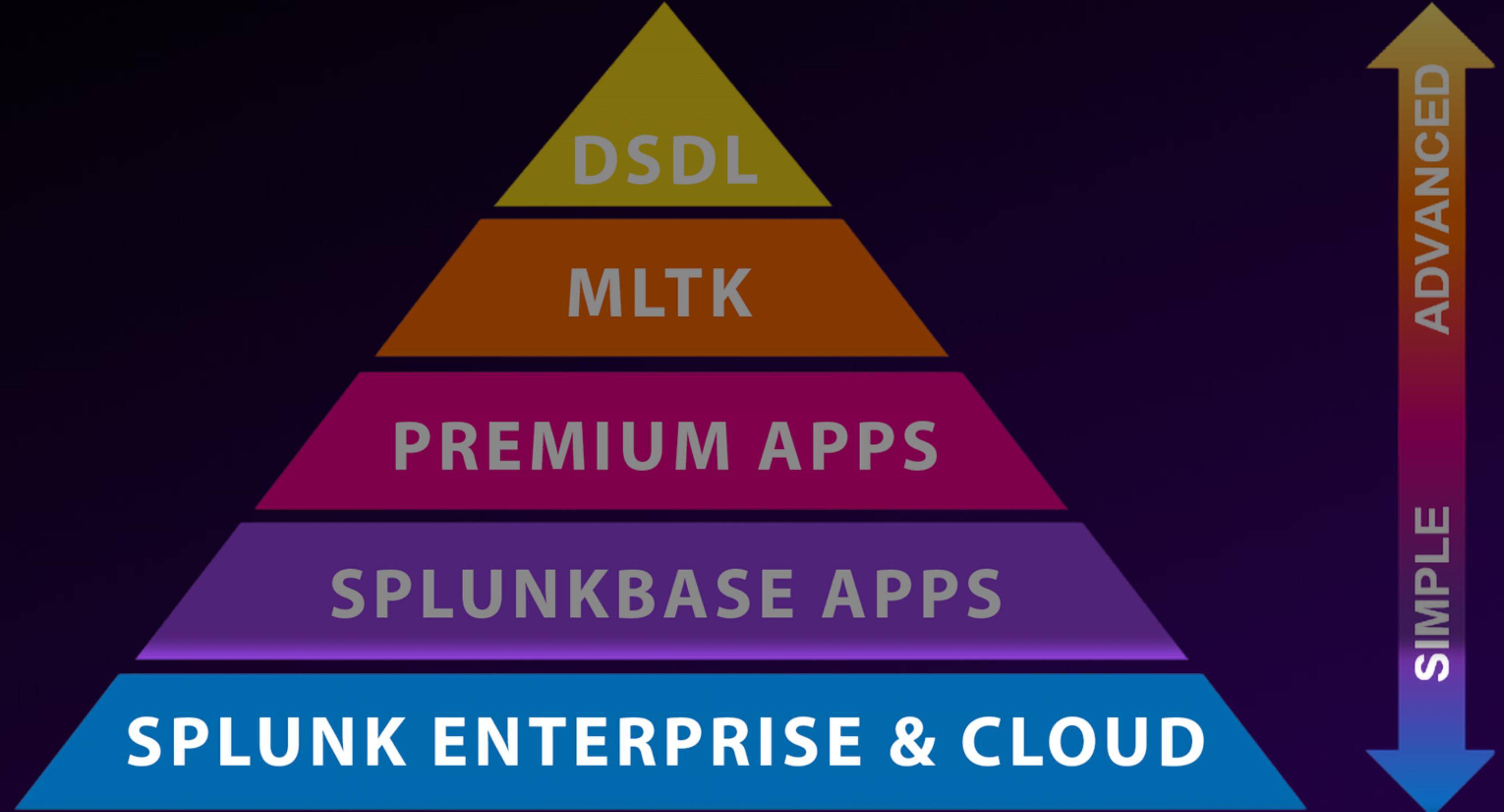
MACHINE LEARNING
TOOLKIT



SPLUNK APP FOR
DATA SCIENCE AND
DEEP LEARNING

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 lists of other commands such as abstract, accum, addcoltotals, addinfo, addtotals, analyzefields, anomalies, anomalousvalue, anomalydetection, append, appendcols, and appendpipe. The main content area starts with a 'Description' section: 'A transforming command that identifies anomalous events by computing a probability for each event and then detecting unusually small probabilities. The probability is defined as the product of the frequencies of each individual field value in the event.' It includes two bulleted lists: one for categorical fields (frequency of X divided by total events) and one for numerical fields (frequency of X as size of bin divided by number of events). Below this is a note about the anomalydetection command's capabilities and a callout to the Splunk Machine Learning Toolkit. The 'Syntax' section shows the command line: 'anomalydetection [<method-option>] [<action-option>] [<pthresh-option>] [<cutoff-option>] [<field-list>]'. The 'Optional arguments' section details the '<method-option>' argument, which can be 'histogram', 'zscore', or 'iqr'. The 'Description' for this argument states it selects the method of anomaly detection, with specific notes for 'zscore' and 'iqr'.

|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'. On the left, there's a sidebar with 'Search Reference' navigation, including sections like '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 't'. The main content area starts with a 'cluster' section, followed by 'Description', 'Syntax', 'Optional arguments', 'SLC options', and 't'. A right sidebar titled 'Previously Viewed' lists 'cluster' with sub-links for 'Description', 'Syntax', 'Usage', 'Examples', and 'See also'.

Splunk® Enterprise
Search Reference

Documentation / Splunk® Enterprise / Search Reference / cluster

cluster

Description

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.

The result of the `cluster` command appends two new fields to each event. You can specify what to name these fields with the `countfield` and `labelfield` parameters, which default to `cluster_count` and `cluster_label`. The `cluster_count` value is the number of events that are part of the cluster, or the cluster size. Each event in the cluster is assigned the `cluster_label` value of the cluster it belongs to. For example, if the search returns 10 clusters, then the clusters are labeled from 1 to 10.

Syntax

`cluster [slc-options]...`

Optional arguments

slc-options

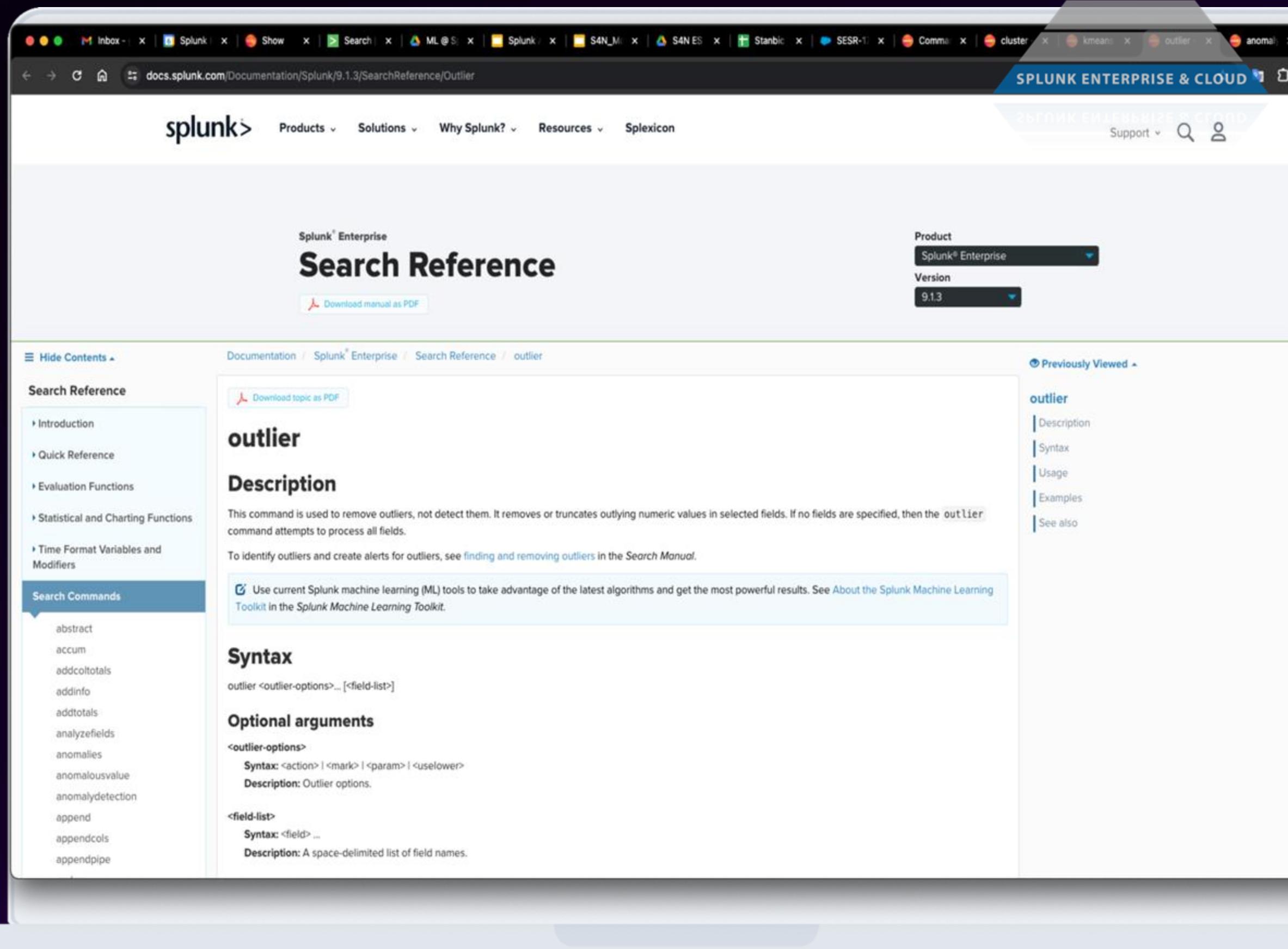
SLC options

t

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 lists various search commands, with 'outlier' highlighted. The main content area describes the 'outlier' command, its optional arguments, and syntax. A sidebar on the right provides links to 'outlier' documentation, syntax, usage, examples, and see also sections. The top navigation bar includes links for Products, Solutions, Why Splunk?, Resources, and Plexicon, along with a search bar and user profile icon.

Splunk® Enterprise
Search Reference

Download manual as PDF

Documentation / Splunk® Enterprise / Search Reference / outlier

Download topic as PDF

outlier

Description

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.

To identify outliers and create alerts for outliers, see [finding and removing outliers](#) in the [Search Manual](#).

Use current Splunk machine learning (ML) tools to take advantage of the latest algorithms and get the most powerful results. See [About the Splunk Machine Learning Toolkit](#) in the [Splunk Machine Learning Toolkit](#).

Syntax

`outlier <outlier-options>... [<field-list>]`

Optional arguments

`<outlier-options>`

Syntax: `<action> | <mark> | <param> | <uselower>`
Description: Outlier options.

`<field-list>`

Syntax: `<field> ...`
Description: A space-delimited list of field names.

Product: Splunk® Enterprise
Version: 9.1.3

Previously Viewed ▾

outlier

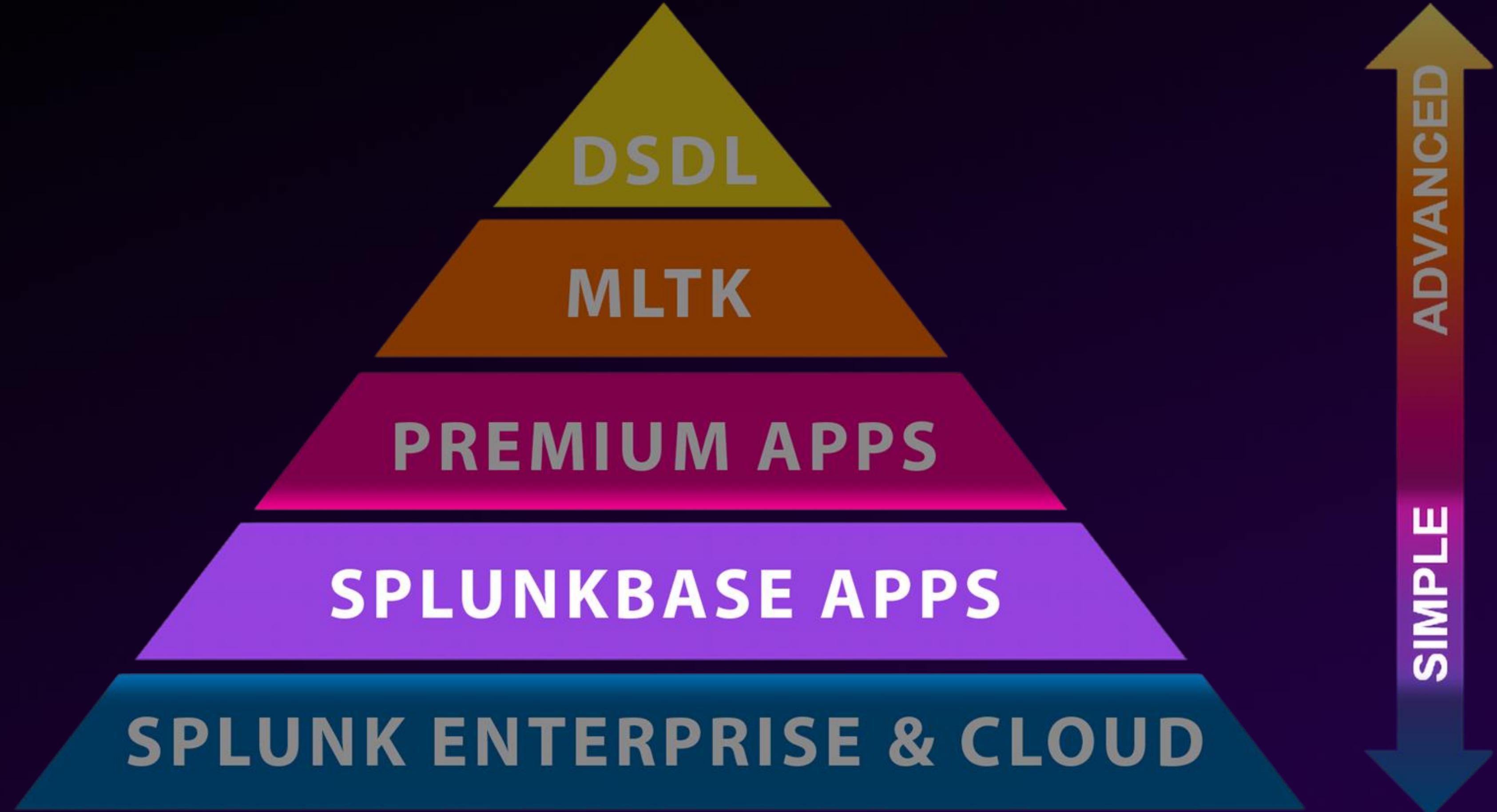
- Description
- Syntax
- Usage
- Examples
- See also

kmeans

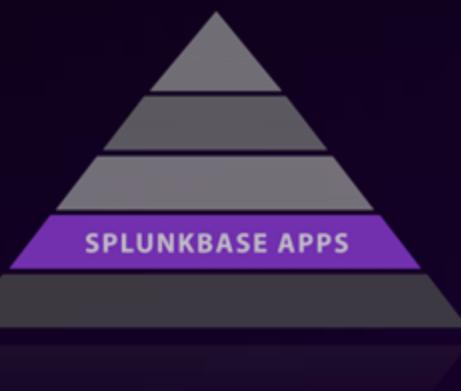
Partitions the events into k clusters, with each cluster defined by its mean value.

<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". On the left, there's a sidebar with a "Search Reference" section containing links to "Introduction", "Quick Reference", "Evaluation Functions", "Statistical and Charting Functions", and "Time Format Variables and Modifiers". Below this is a "Search Commands" section with a list of commands including `abstract`, `accum`, `addcoltotals`, `addinfo`, `addtotals`, `analyzefields`, `anomalies`, `anomalousvalue`, `anomalydetection`, `append`, `appendcols`, `appendpipe`, and `kmeans`. The `kmeans` command is highlighted with a blue background. The main content area starts with the **Description** section, which states: "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 this is the **Syntax** section: `kmeans [kmeans-options...][field-list]`. The **Required arguments** section says "None". The **Optional arguments** section includes the `field-list` and `kmeans-options`. The `field-list` is described with a syntax of `<field> ...`, a description of specifying a space-separated list of exact fields for joining, and a default of using all numerical fields if none are specified. The `kmeans-options` are described with a syntax of `<reps> | <iters> | <D> | <k> | <cnumfield> | <distype> | <showcentroid>`, and a description of options for the `kmeans` command.



Splunk AI Assistants everywhere



Our goal: Make everyone a Splunk expert!

AI Assistant for SPL

To detect the creation of a new local user in Windows, we can monitor the Windows Security log for new user accounts. The SPL query for this would be:

```
index=<index> source==*WinEventLog*:Security AND EventID=4704  
| dedup session_id  
| table _time, host, User, TargetUserName, Subject
```

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

AI Assistant
Geographically Improbable Access Detected

Source Host: workstation-478.internal.domain
Destination Host: server-932.internal.domain
Source IP Address: 192.168.10.47
Destination IP Address: 192.168.10.123
Protocol: SMB

Event Information
Files accessed and transferred: \\server-932\share\payroll\\server-932\share\config

Additional Context
1. The source workstation ('workstation-478') has been identified as a regular employee named 'John Doe'. It has established connections to this server.

AI Assistant in Observability Cloud

AI Assistant Chat 1

AZ Why is frontend having issues?

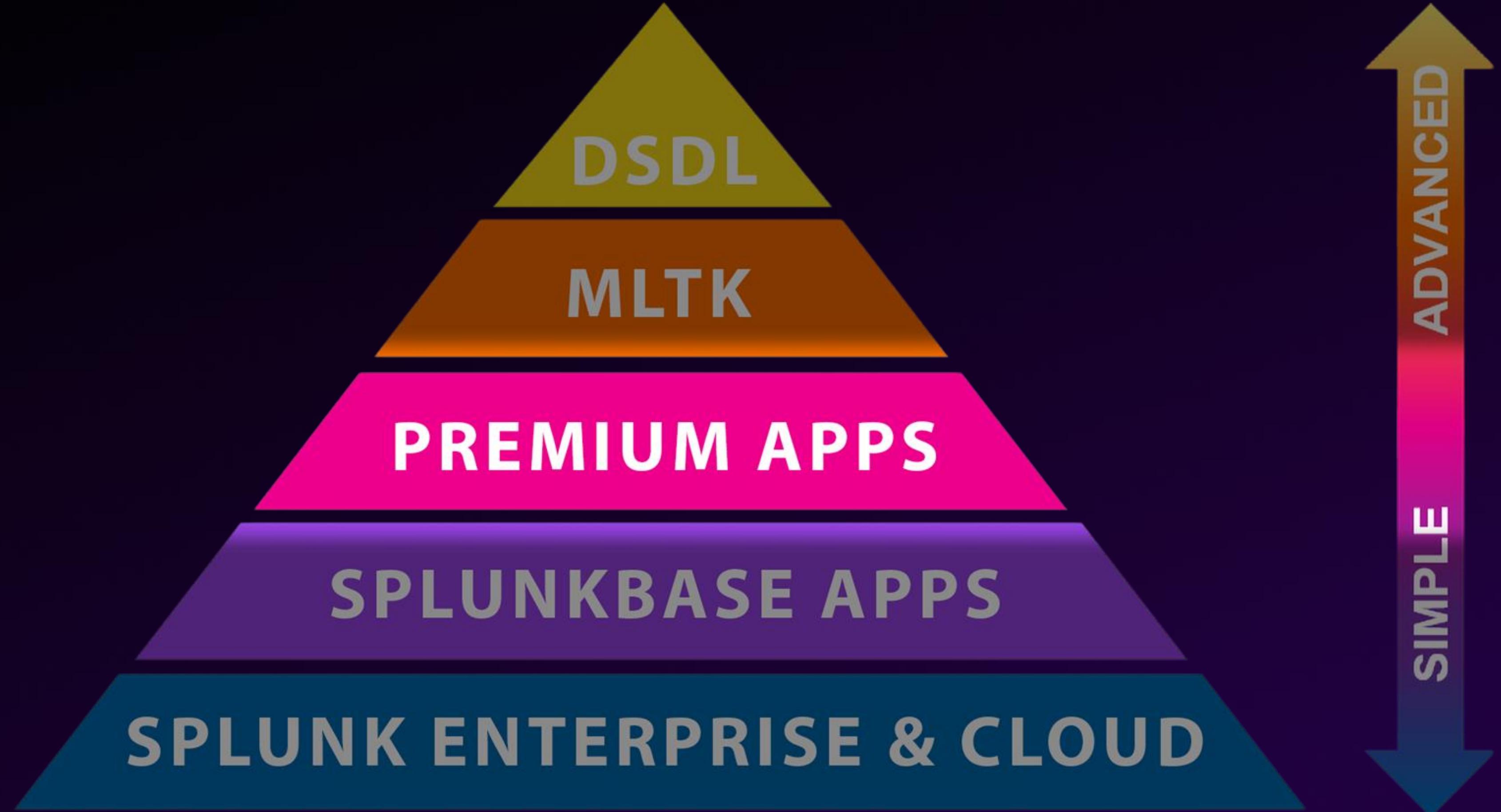
AI Assistant Here are some possible reasons:
it seems like frontend is having issues downstream dependency "paymentservice" is experiencing problems.

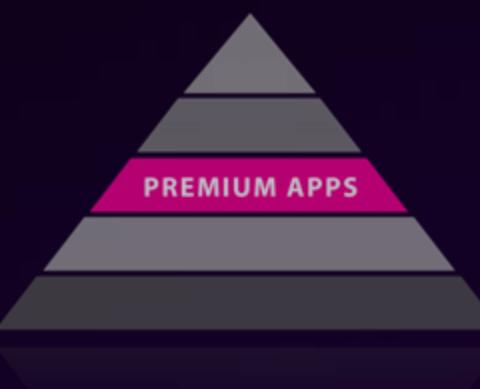
Checking the health map

Generally Available

Generally Available

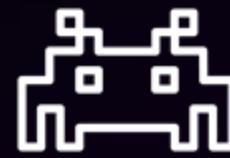
Generally Available





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

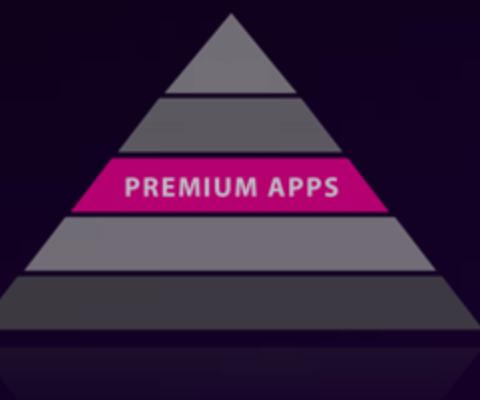


Release

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](#)

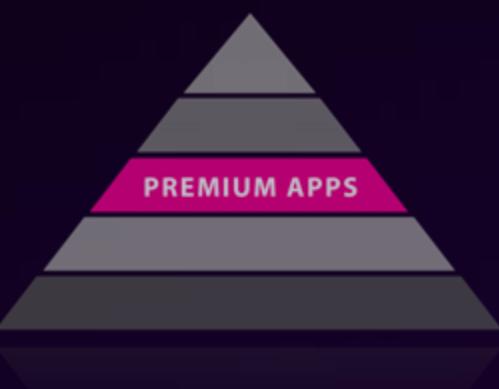
Try in Splunk Security Cloud Description This analytic identifies suspicious modification of registry to deface or change the wallpaper of...

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!

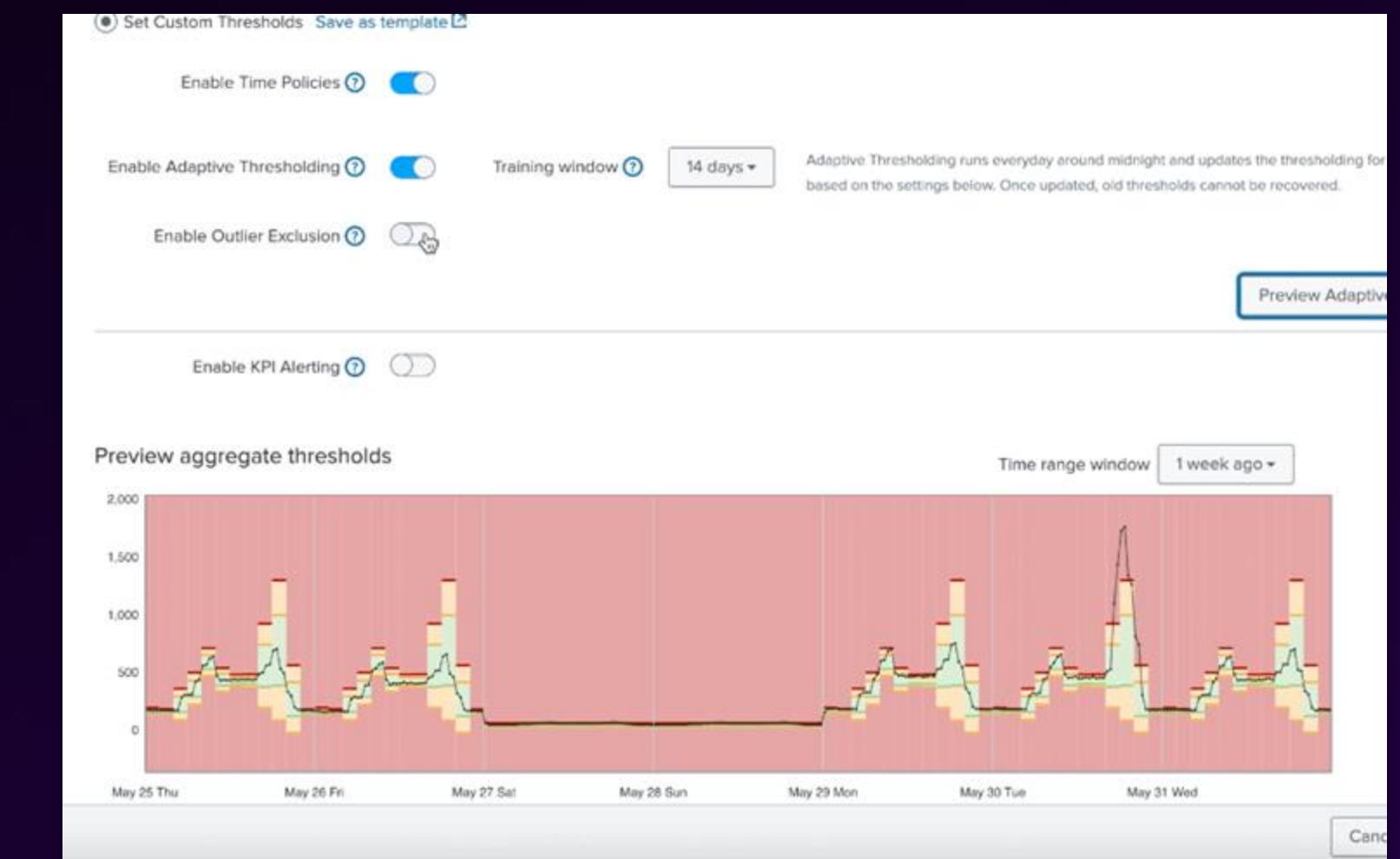


Splunk IT Service Intelligence (ITSI)

Splunk's AIOps Solution

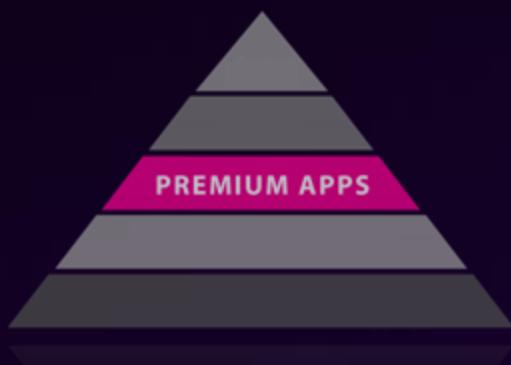
- Splunk ITSI applies machine learning to **proactively prevent outages** by correlating and reducing alerts, monitoring service health, and streamlining incident management.

- 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



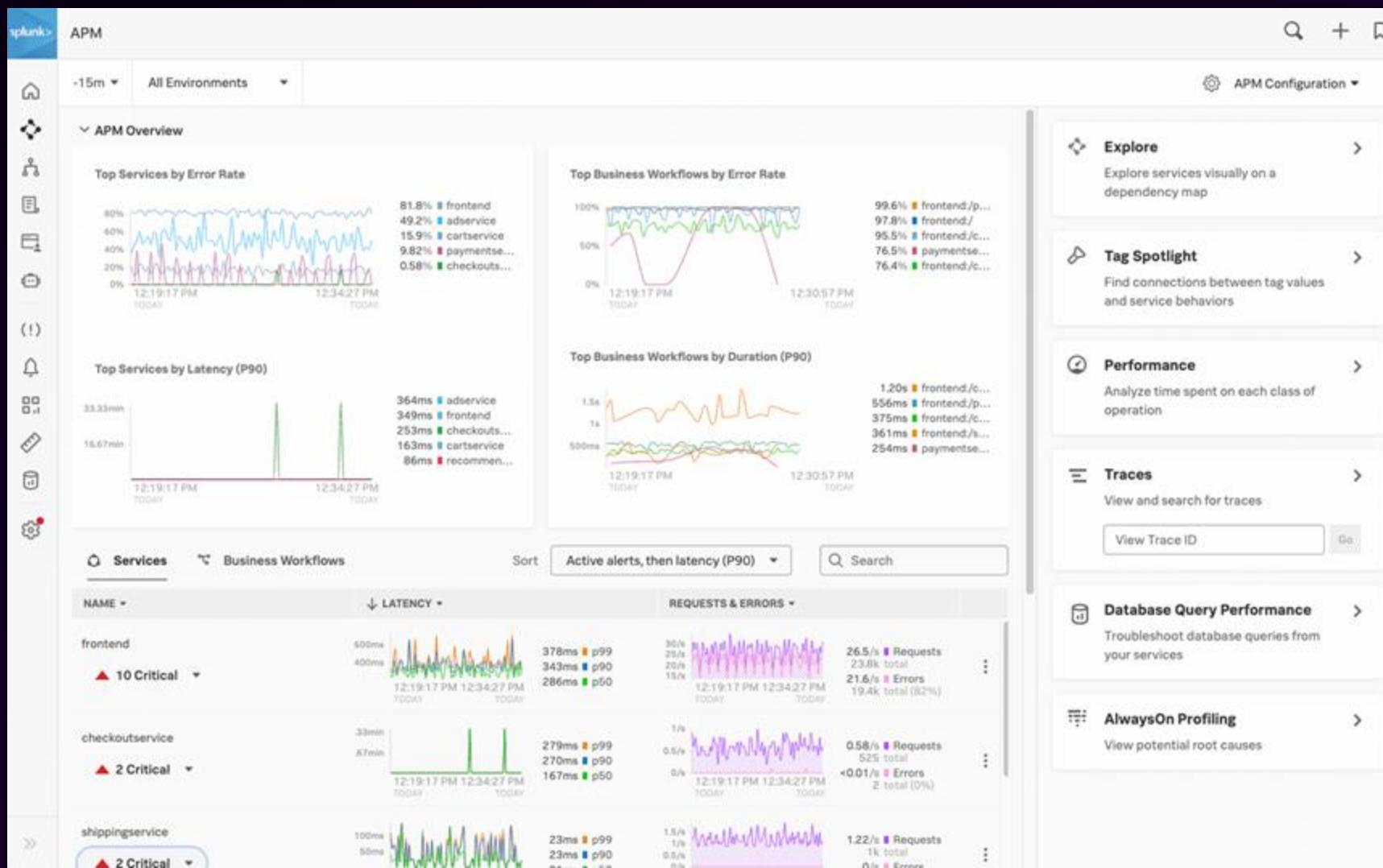
New updates!

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



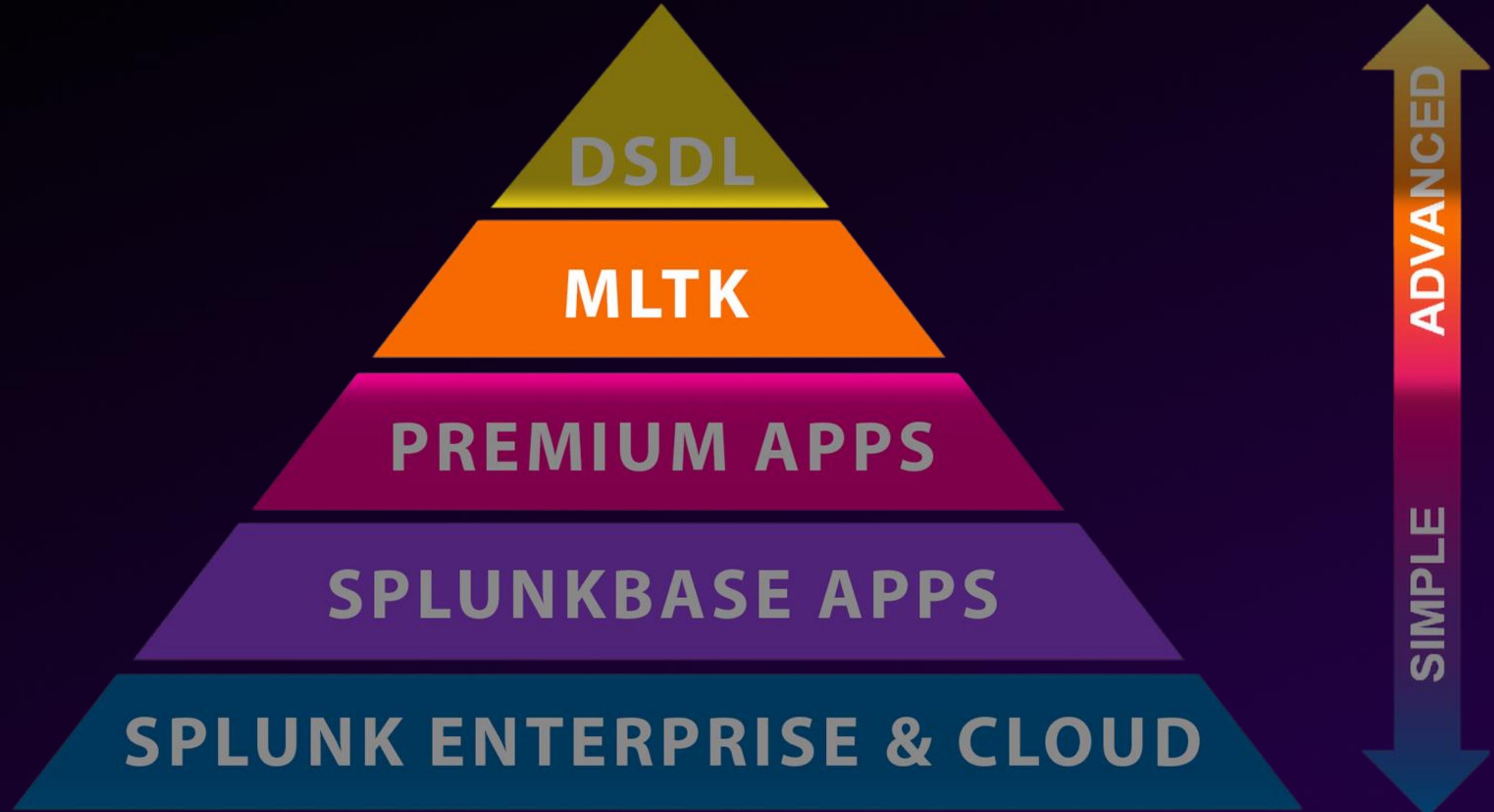
Application and Infrastructure Monitoring: Autodetect

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

Customize AI with AI Toolkit (AITK)

In Splunk Enterprise and Cloud Platform

- Experiment and model your Splunk data with guided assistant for the whole AI workflow.
- 50+ algorithms to choose from or bring your own model.
- Train and deploy with search commands and operationalize in real-time.

The screenshot shows the 'Showcase' tab selected in the top navigation bar of the ML Toolkit. Below it, there's a search bar and several navigation links: Experiments, Search, Models, Settings, Docs, and Video Tutorials. The main content area is titled 'Showcase' and features a welcome message: 'Welcome to the Machine Learning Toolkit Showcase. Watch and learn from interactive end-to-end examples using real datasets. Click on an example to see how it works.' There are two filter buttons: 'View examples by' with options 'ML Operation' (selected) and 'Industry'. The page is divided into three sections: 'Predict Fields' (with a blue line chart icon), 'Forecast Time Series' (with a blue line chart icon), and 'Detect Outliers' (with a blue line chart icon). Each section has a brief description and a '15 Examples Available' link.

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 see how it works.

View examples by ML Operation Industry

Predict Fields

View examples that predict the value of a numeric or categorical field using the values from other fields in the event.

15 Examples Available

Forecast Time Series

View examples that predict the next value in a sequence of time series data by using past time series data.

9 Examples Available

Detect Outliers

View examples that detect numeric and categorical outliers in the rest of the data. Identified outliers can be used to detect and possibly dangerous events.

Cluster Events

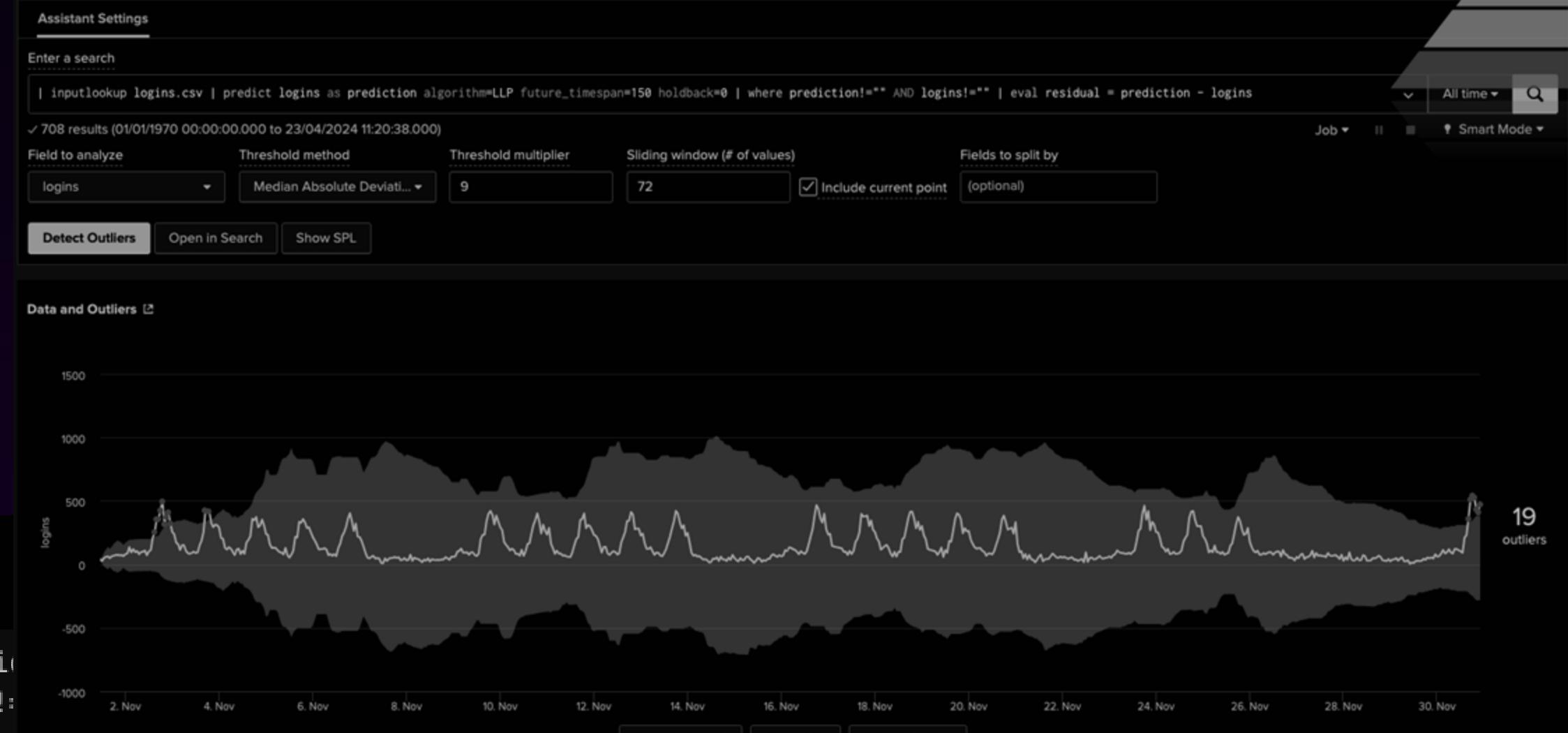
View examples that partition events with similar values on the values of those fields.

Plot the outliers

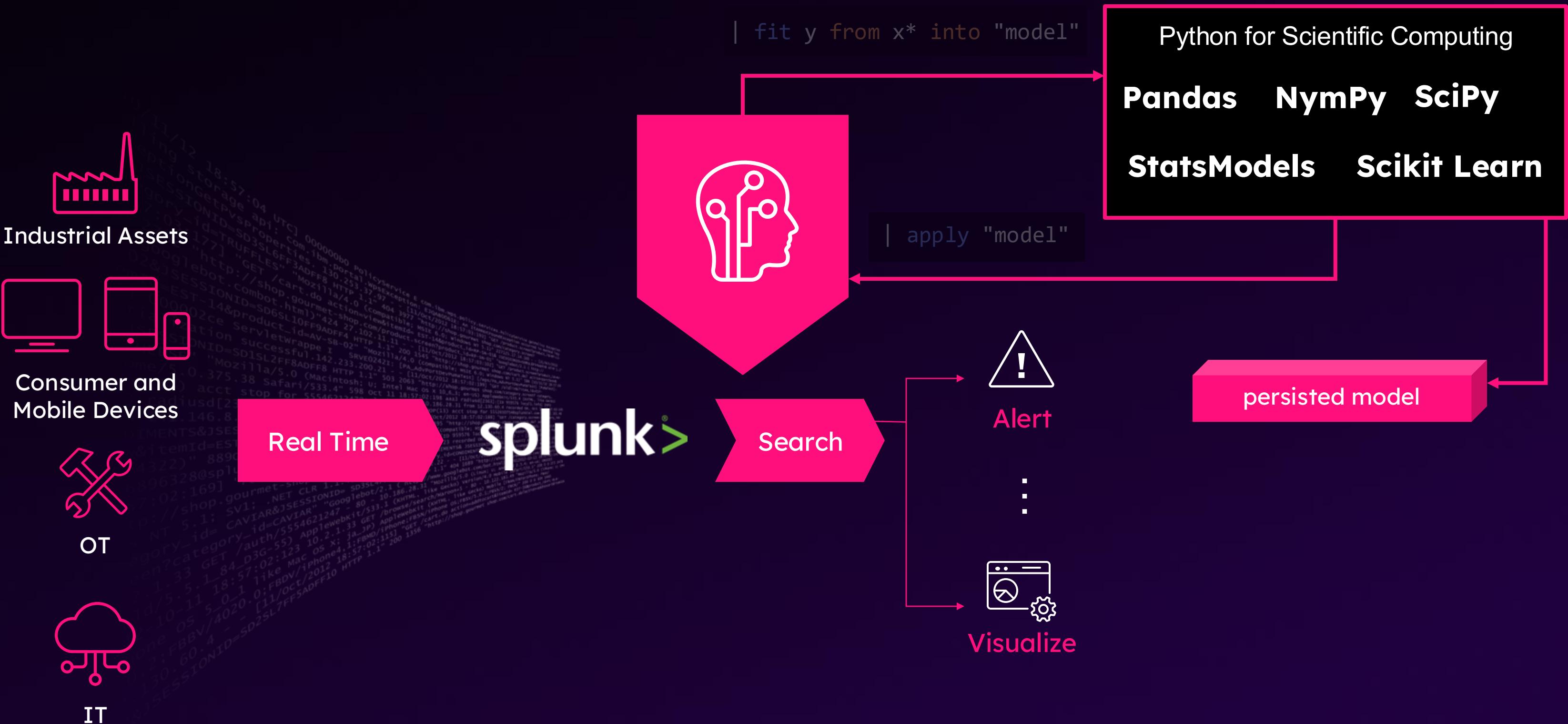
```
| inputlookup logins.csv | predict logins as prediction
future_timespan=150 holdback=0 | where prediction!=""
eval residual = prediction - logins

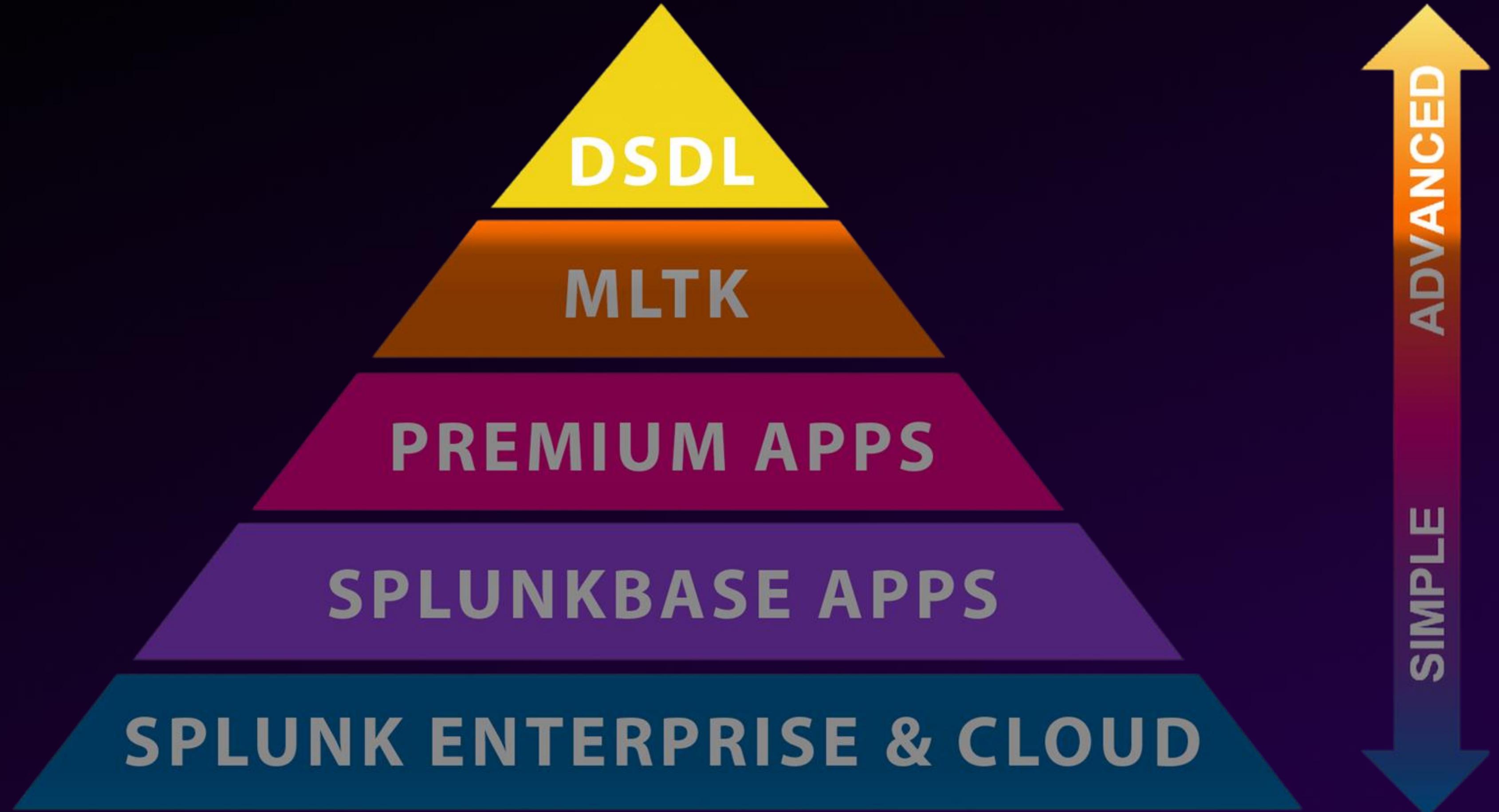
| streamstats window=72 current=true median("logins") as median          // calculate the median value using a sliding window
| eval absDev=(abs('logins'-median))                                         // calculate the absolute deviation of each value from the median
| streamstats window=72 current=true 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:**

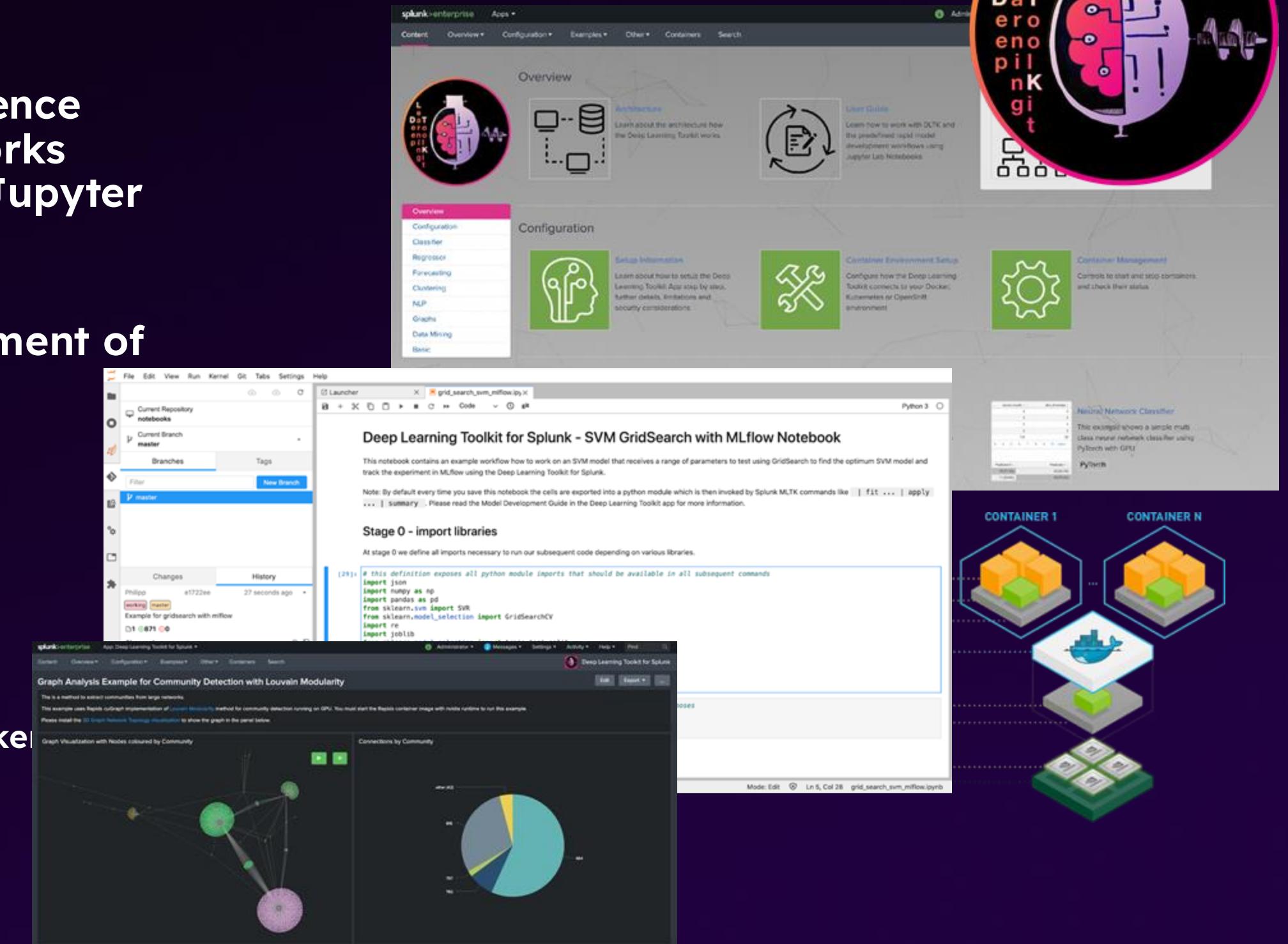
Speed up your data science projects with GPU accelerated containers

- **K8s Support:**

Scalable and HA with K8s deployment

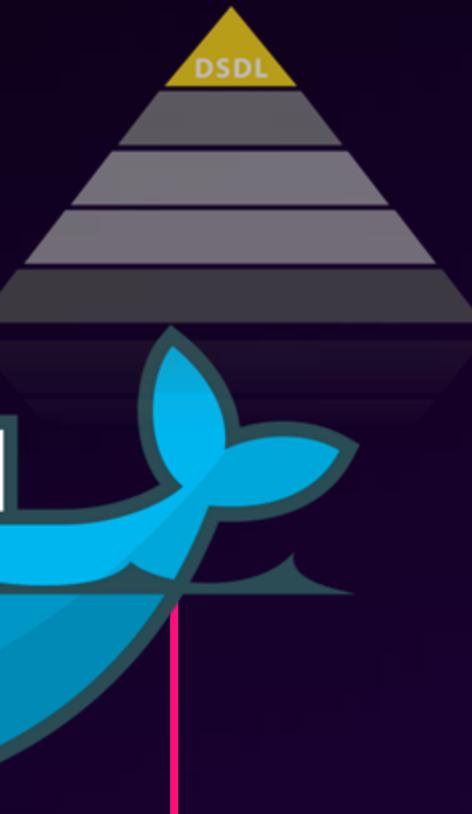
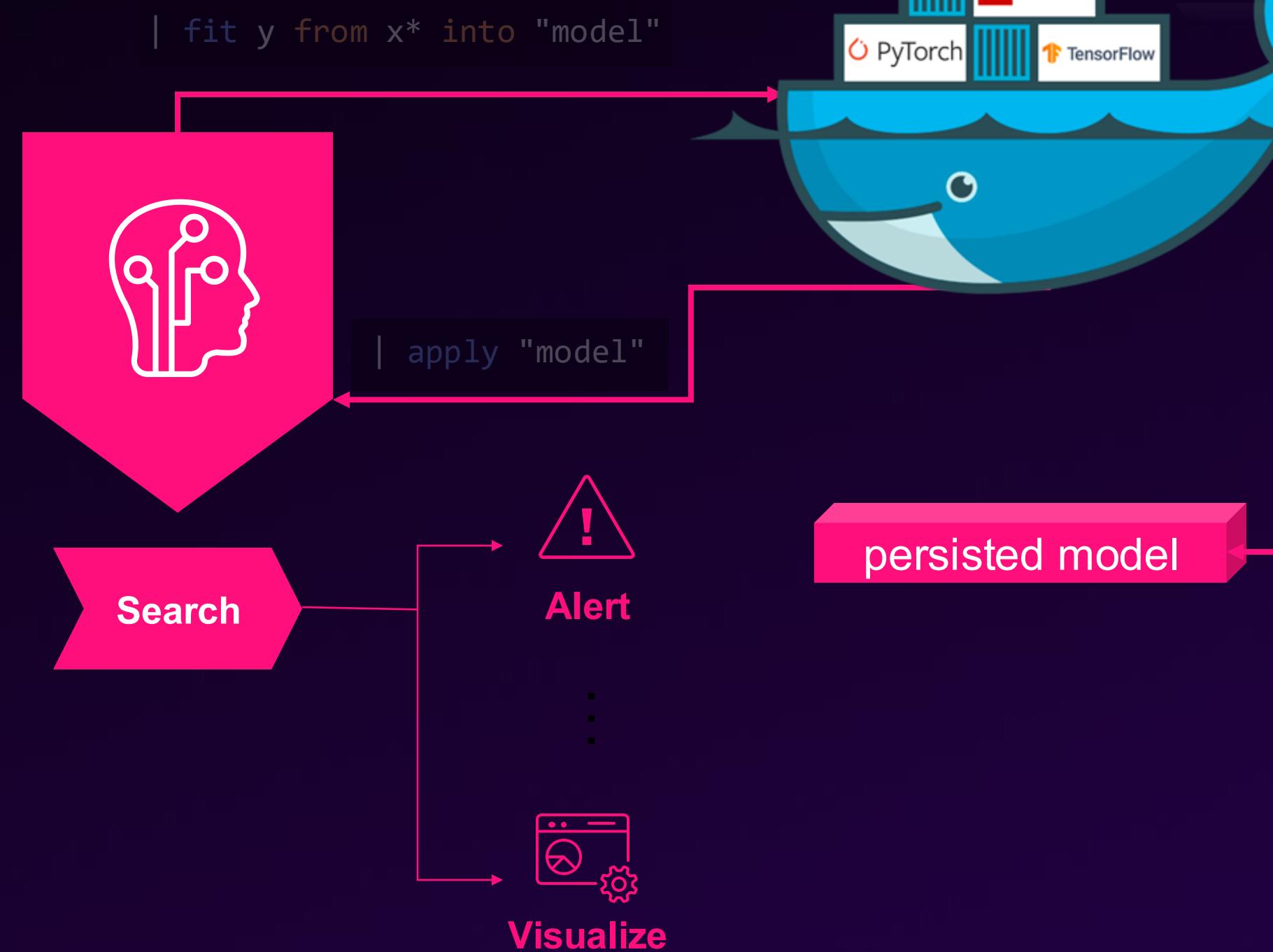
- **Open source for customization:**

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



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

DLTK for Splunk



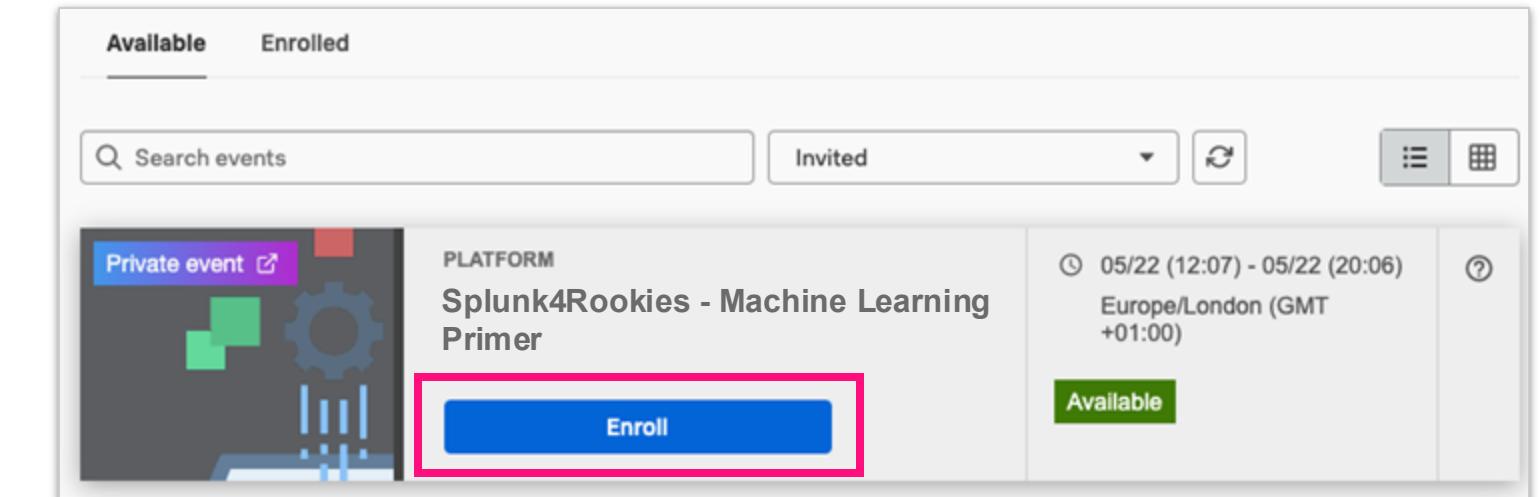


Enroll in Today's Workshop

Tasks

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

Goal



Enroll in today's event

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





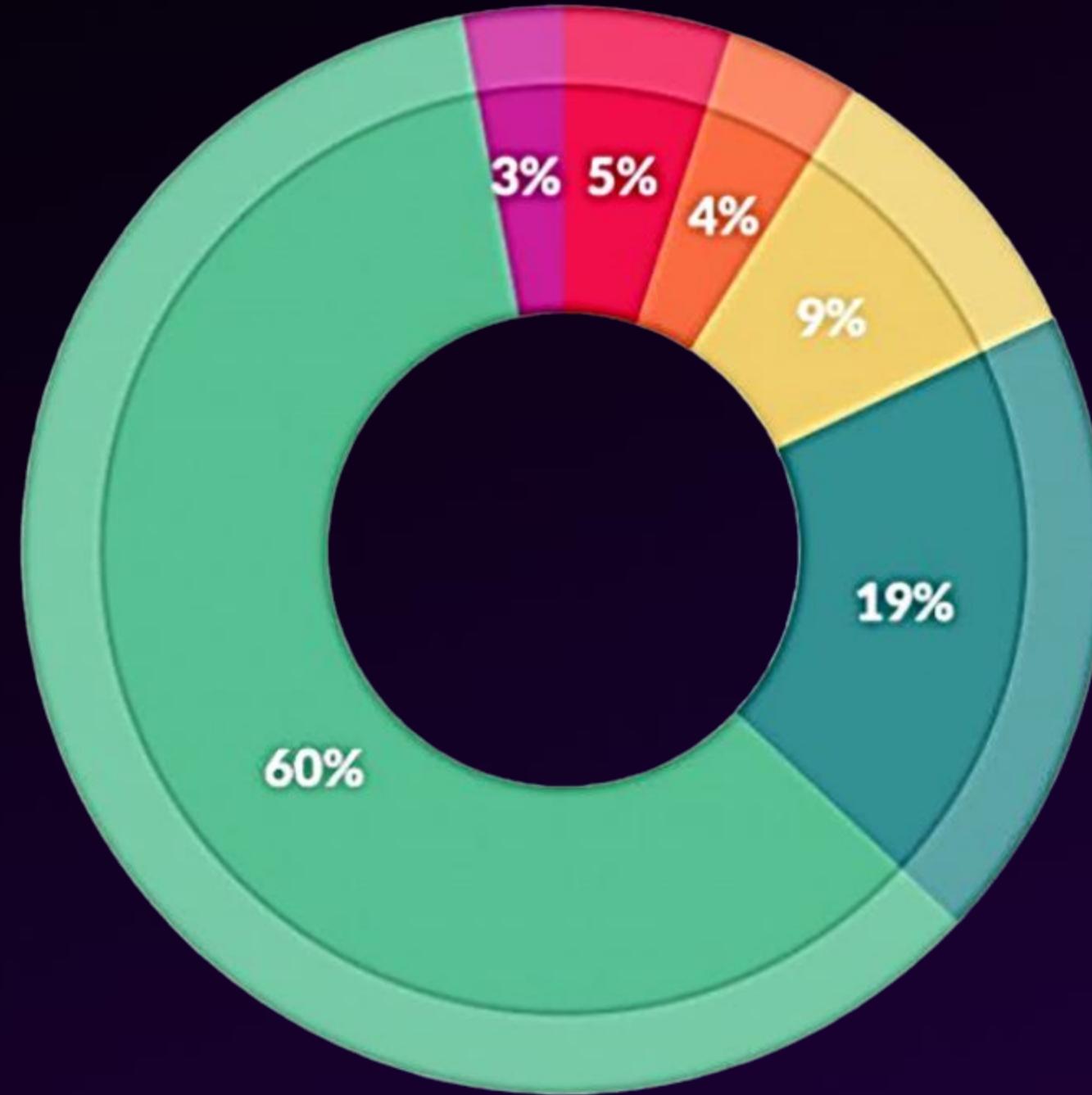
Cleaning Big Data, the Most Time-Consuming, Least Enjoyable Data Science Task

Forbes Survey

March 23

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

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		
Layer 2: Data		AGENT TOOL FRAMEWORKS					LangChain	Auto-gpt	FIXIE	LlamaIndex	
Layer 1: Compute + Foundation	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										
	OpenAI	ANTHROPIC	MISTRAL AI	contextual-ai	Hugging Face	Llama 2					
Layer 1: Compute + Foundation	GPU PROVIDERS										
	aws	Azure	Google Cloud	CoreWeave	Lambda	FOUNDRY	together.ai				

Custom ML with the Splunk Platform

Ecosystem

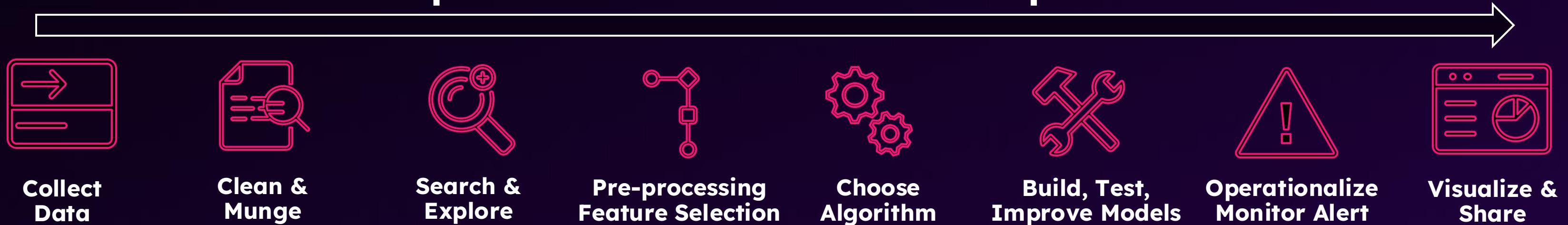
Splunk's App Ecosystem contains 1000's of free add-ons for getting data in, applying structure and visualizing your data giving you faster time to value.

MLTK

Splunk

The Machine Learning Toolkit delivers new SPL commands, custom visualizations, assistants, and examples to explore a variety of ml concepts. Splunk Enterprise is the mission-critical platform for indexing, searching, analyzing, alerting and visualizing machine data.

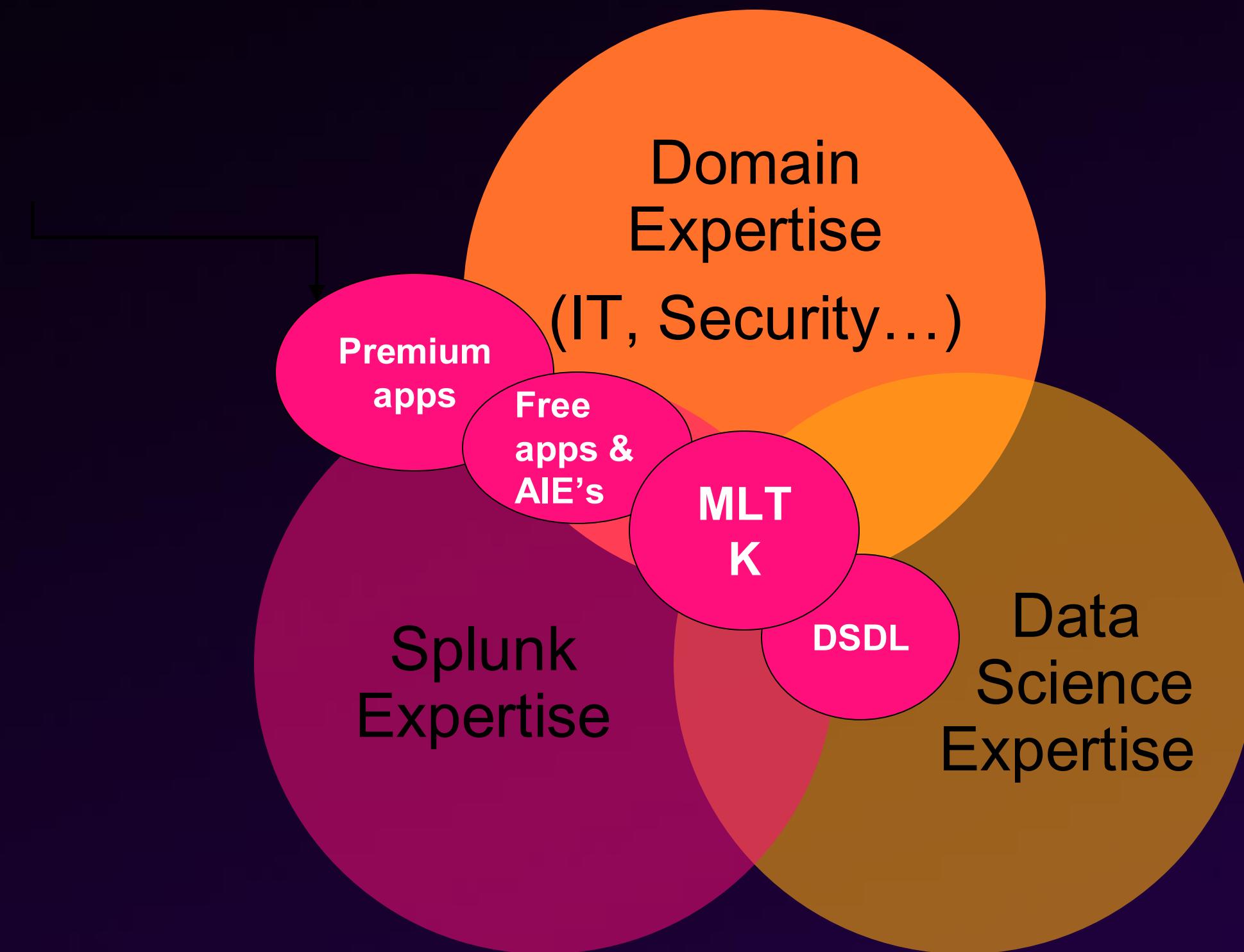
Operationalized Data Science Pipeline



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

splunk> Platform for Operational Intelligence

Splunk can help at every skill level



Numeric



Categorical



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

Missing data in a dataset can lead to:

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

	School ID	Name	Address	City	Subject	Marks	Rank	Grade
0	101.0	Alice	123 Main St	Los Angeles	Math	85.0	2	B
1	102.0	Bob	456 Oak Ave	New York	English	92.0	1	A
2	103.0	Charlie	789 Pine Ln	Houston	Science	78.0	4	C
3	NaN	David	101 Elm St	Los Angeles	Math	89.0	3	B
4	105.0	Eva	NaN	Miami	History	NaN	8	D
5	106.0	Frank	222 Maple Rd	NaN	Math	95.0	1	A
6	107.0	Grace	444 Cedar Blvd	Houston	Science	80.0	5	C
7	108.0	Henry	555 Birch Dr	New York	English	88.0	3	B

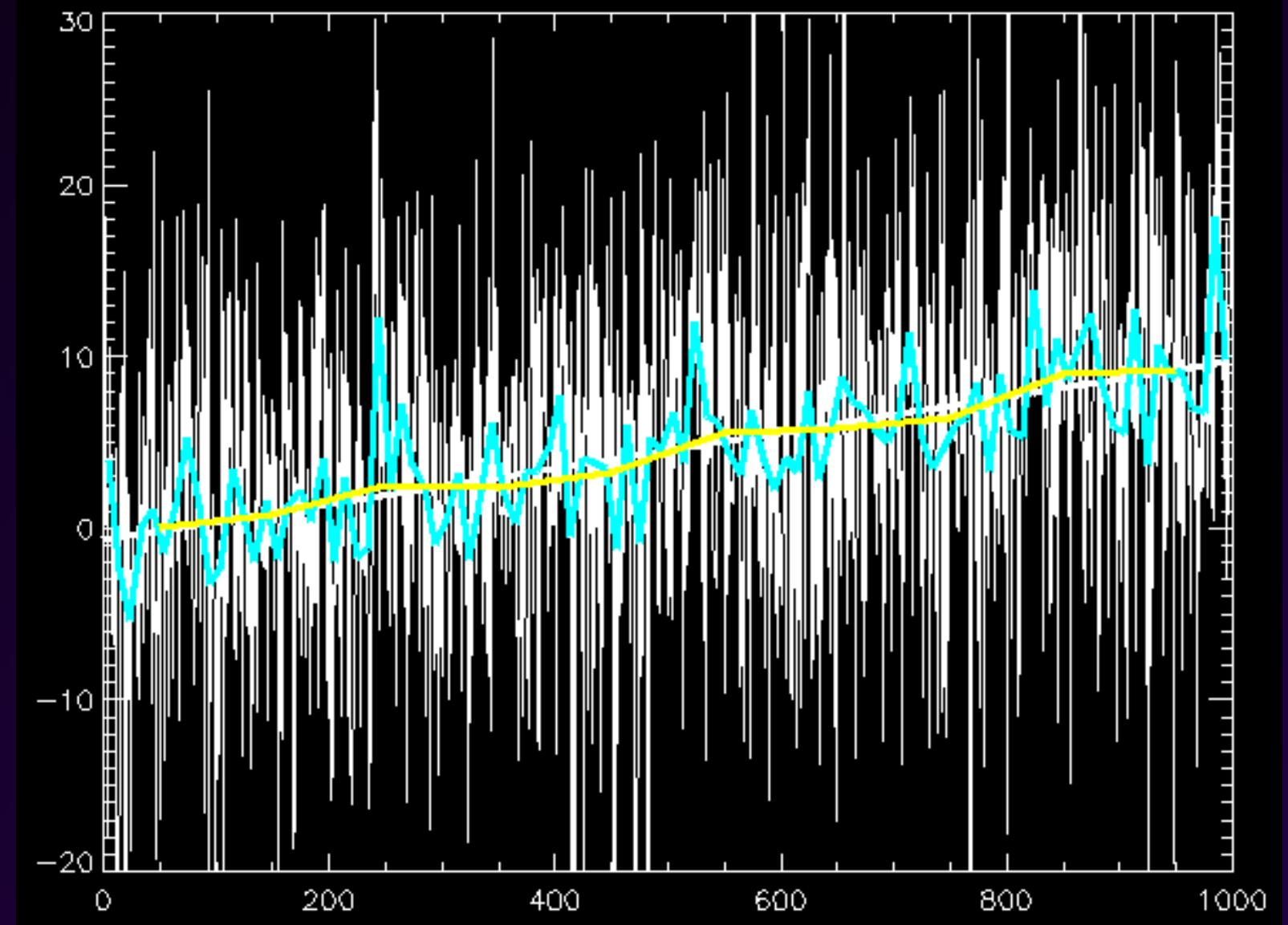
ML can not answer questions on which it does not have data

Noisy Data

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

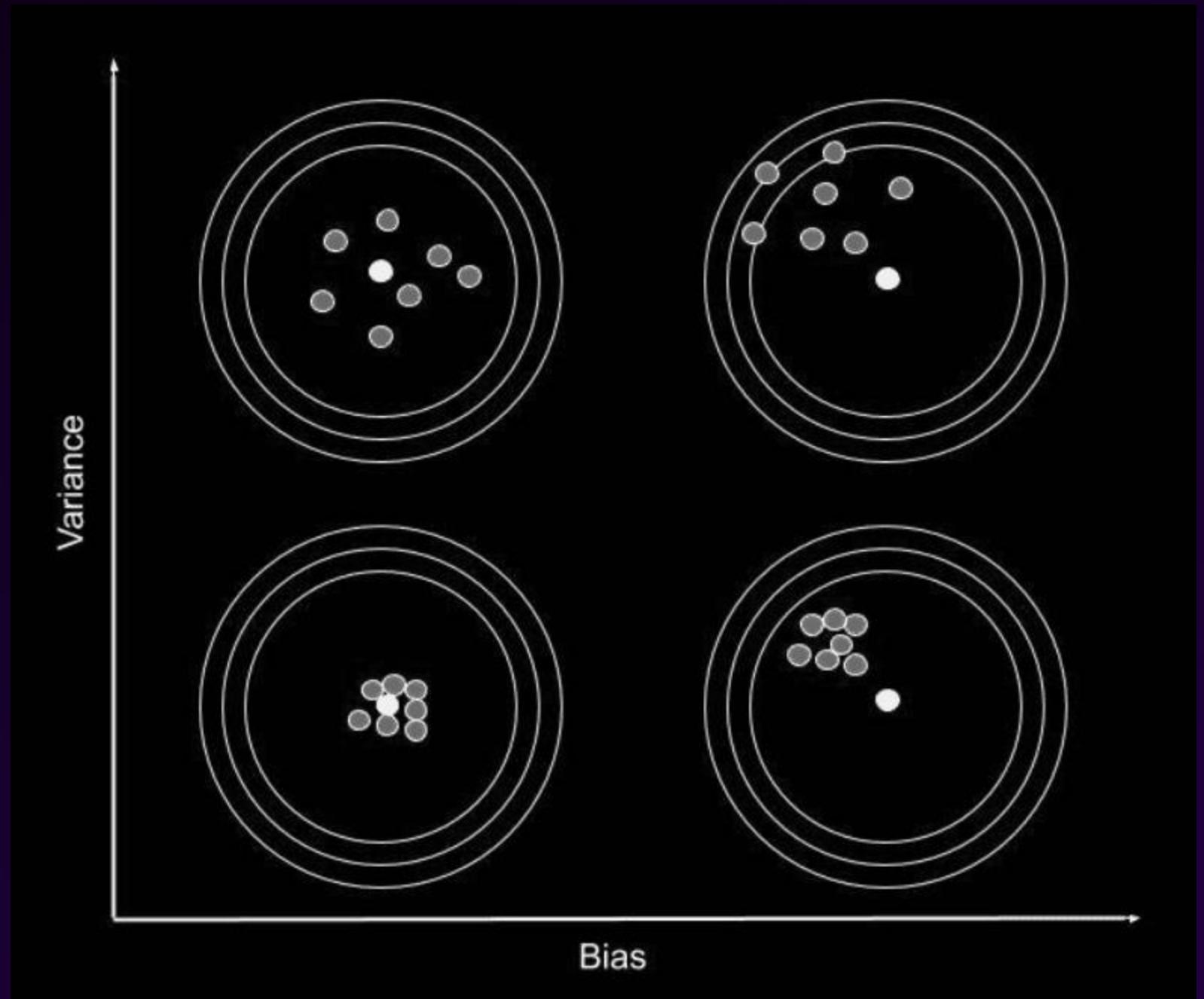
Noisy data in a dataset can lead to:

- 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.



Duplicate Data

Duplicate entries in a dataset can lead to:

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

	In	dob	gn	fn	is_duplicate
0	SMITH JR	01/03/68	F	WILLIAM	0
1	ROTHMEYER JR	01/03/68	F	WILLIAM	0
2	ASBY JR	01/03/68	F	WILLIAM	0
3	SALTER JR	01/03/68	F	WILLIAM	0
4	SALTER JR	01/03/68	F	WILLIAM	1

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.

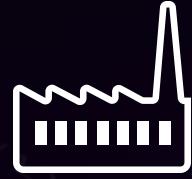
Outliers can lead to:

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



Data Collection

Gain access to previously unused data



Industrial Assets



Consumer and
Mobile Devices



OT



IT



Infrastructure data

Application usage and access data

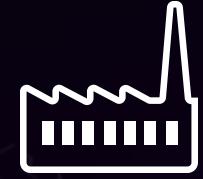
GenAI prompt and output data

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

splunk>

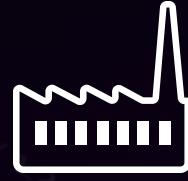
Scalable index and storage

Flexible offering models

Federated analytics

Data understanding

Gain insights to previously unused data



Industrial Assets



Consumer and
Mobile Devices



OT



IT

IT

Real Time

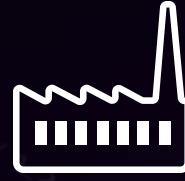


Machine data is messy and complex...

```
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 · User IP · 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_HFREZL
Mozilla/5.0 (Linux; Android 12.0.0; FR-fr;
SM-S901B Build/S908EXXU) AppleWebKit/537.36
AppleWebKit/537.36 Chrome/114.0.5735.131
Mobile Safari/537.36" 954
```

User IP

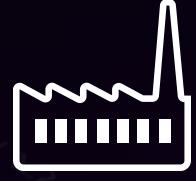
Product Viewed

Preferred Language

Device

Data understanding

Is it good data?



Industrial Assets



Consumer and
Mobile Devices



OT



IT



splunk>

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

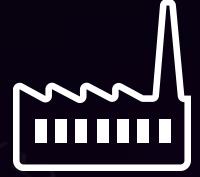
If your data set does not contain any of this you can start applying ML.

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

Filter



Industrial Assets



Consumer and
Mobile Devices



OT

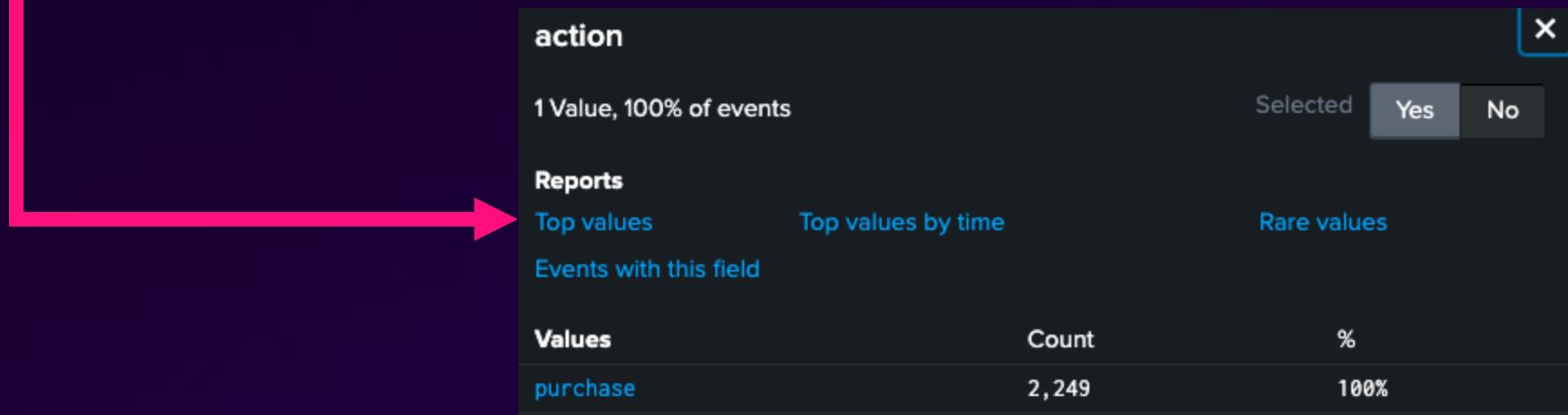


IT

Real Time

splunk>

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



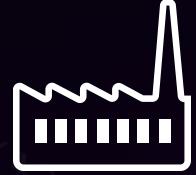
Data Preparation

Transform data with SPL



Filter

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



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

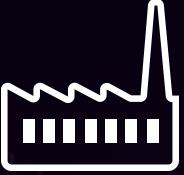
clientip	Country	City
0.134.99.193		
0.134.99.193		
0.134.99.193		
0.152.133.194		
0.152.133.194		
0.152.133.194		
1.12.191.128		
1.16.209.124		

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



Industrial Assets



Consumer and
Mobile Devices



OT



IT

Real Time

splunk>

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

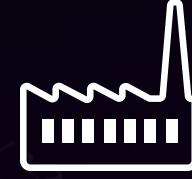
Filter

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
```

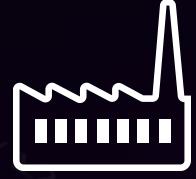
Aggregate

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

Filter

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
```

Aggregate

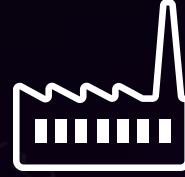
```
| fit algorithm from clientip into "model"
```

Train
"model"

Filter

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

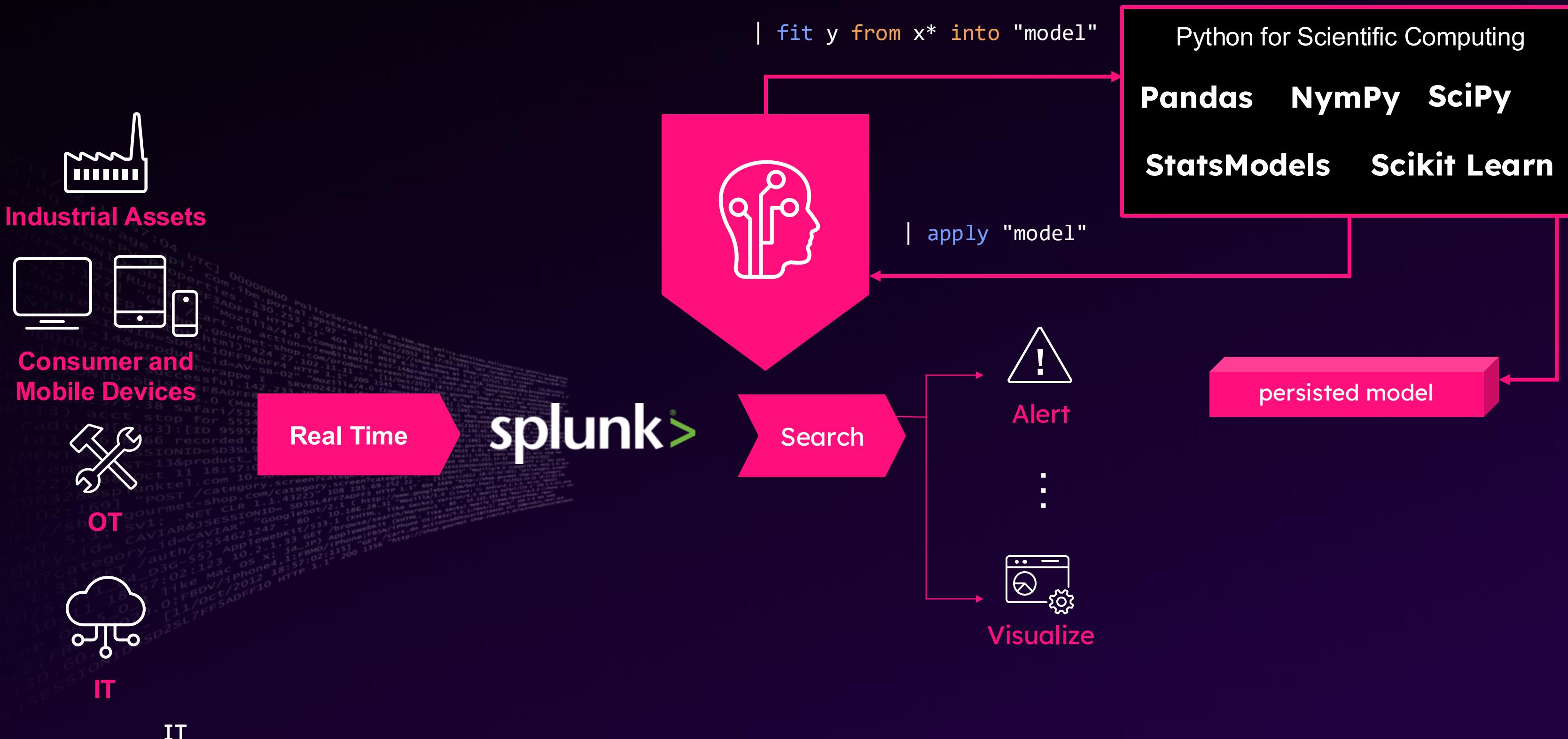
```
| fit algorithm from clientip into "model"
```

Train
"model"

```
| apply "model"
```

Run "model"

Model training and application



Hands-on

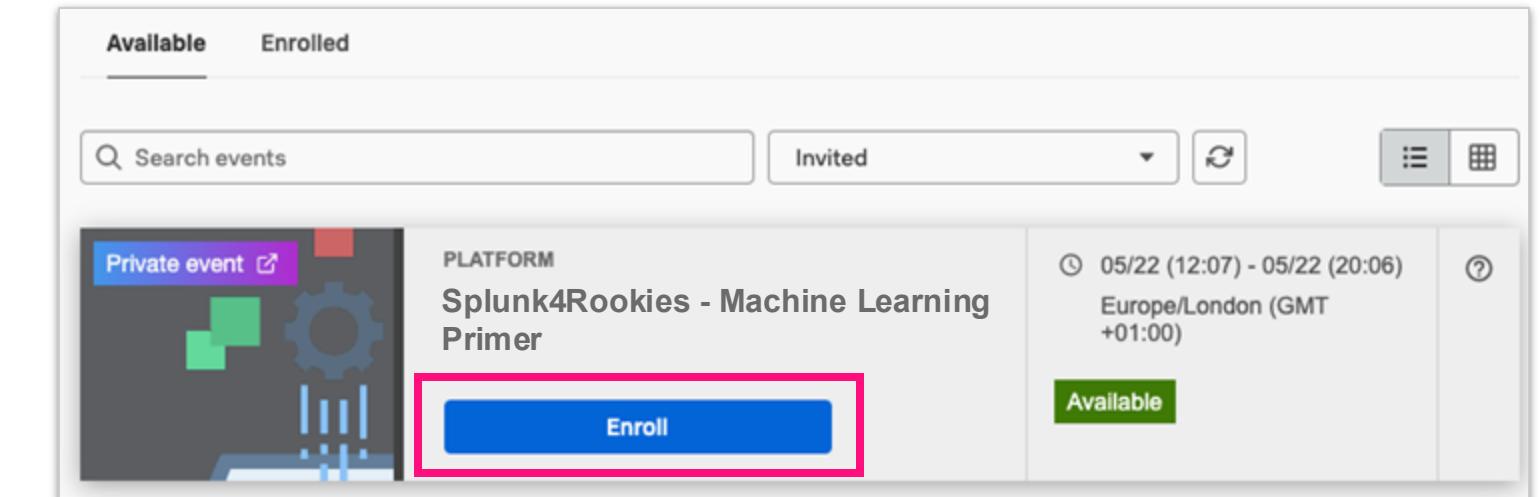


Enroll in Today's Workshop

Tasks

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

Goal



Enroll in today's event

Use Case: Find Anomalies in Supermarket Purchases



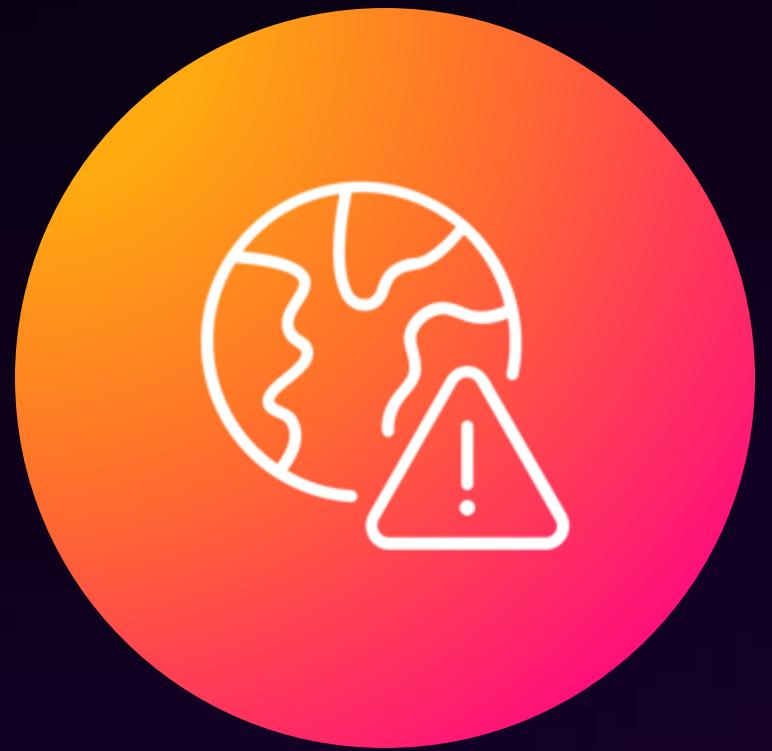
Outlier Detection Algorithms

Identify and analyze abnormal behavior in your data

Includes:

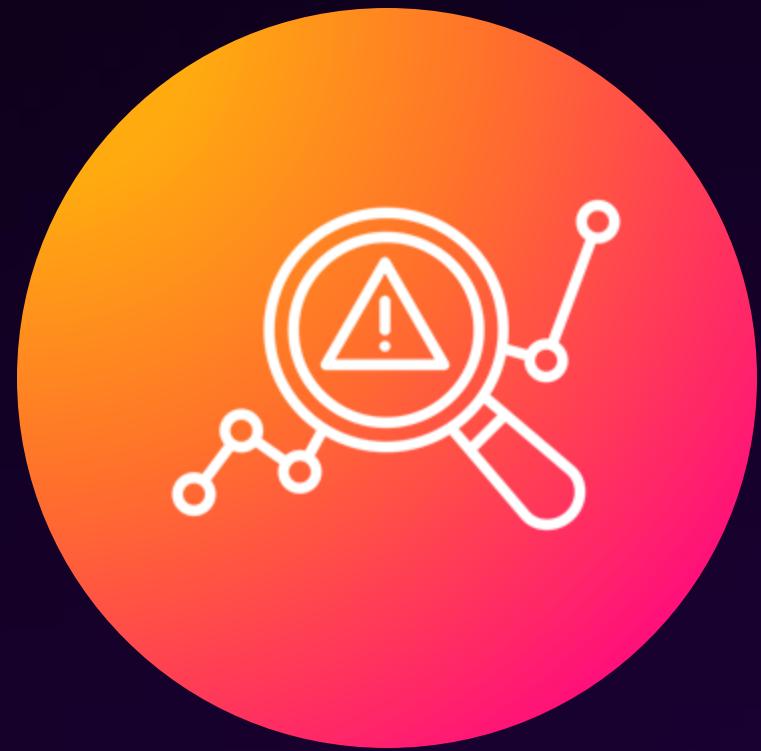
- Clustering
- Outlier Detection

Global



Data points different from
expected pattern, range, or norm

Contextual



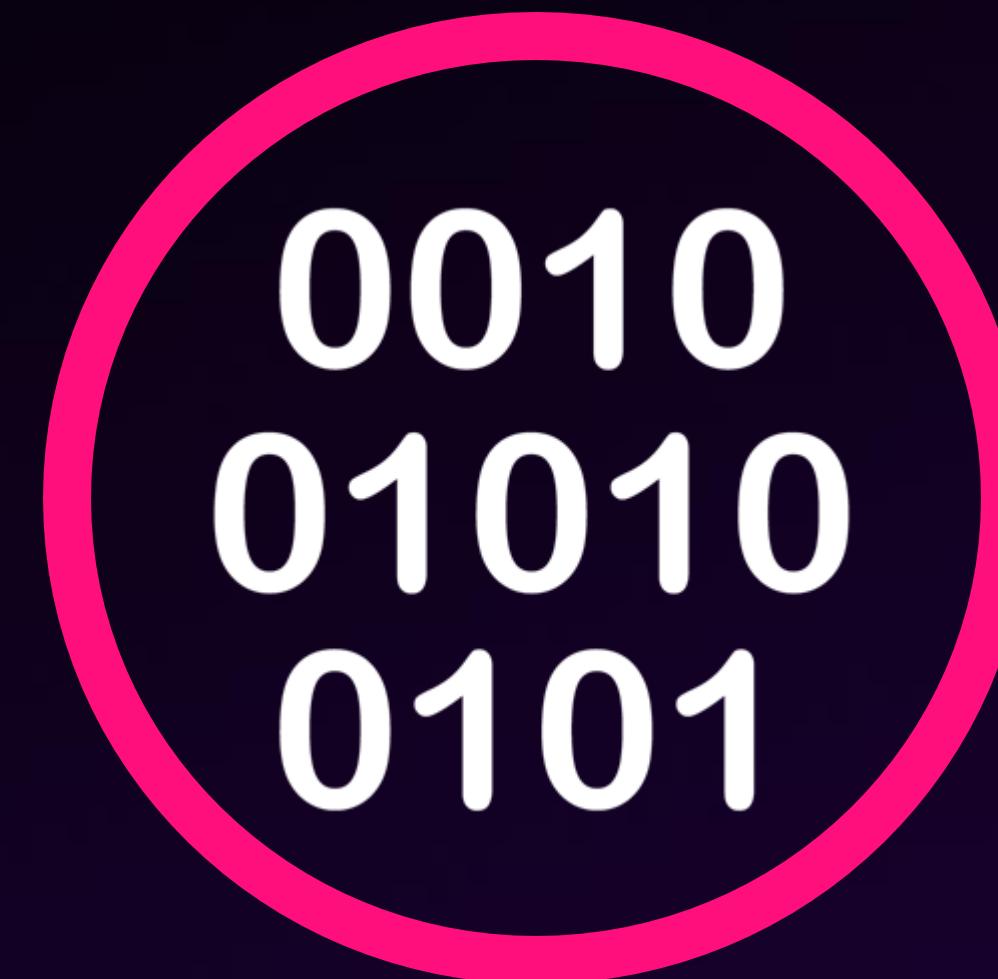
Are the results out
of context?

Collective



Looks normal with isolation
but stands out in a group

Numeric



Categorical



Included Algorithms

DensityFunction

| LocalOutlierFactor

| MultiVariateOutlierDetection

| One-Class SVM

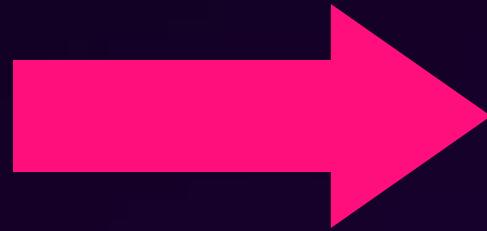
Categorical data to Numeric data?

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

Log Into [INSTANCE URL]

Exercise #1
Time: 10 minutes

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

Use Case: Forecast Monthly Sales

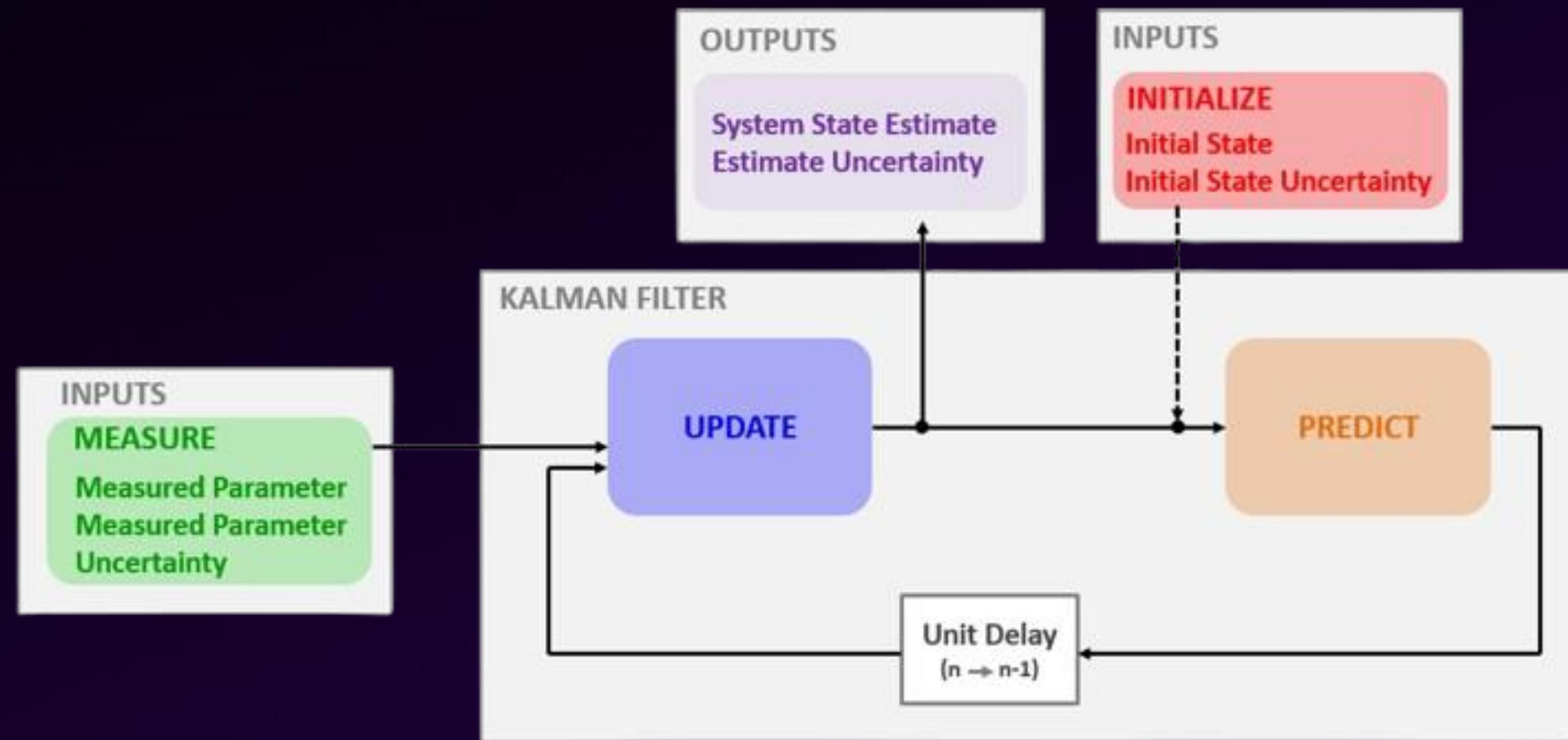
Forecasting Time Series

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

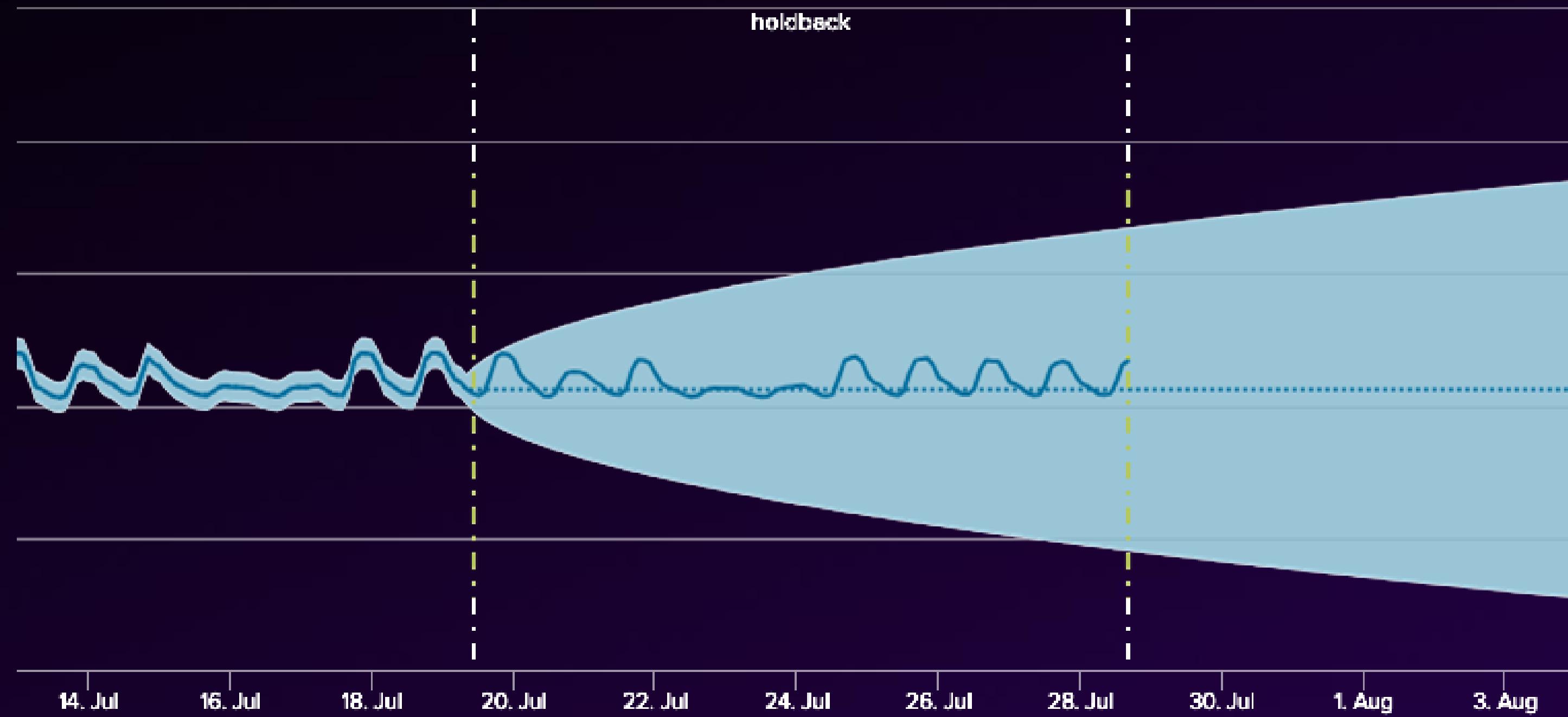


Source:Parzival'1997, flaticon.com

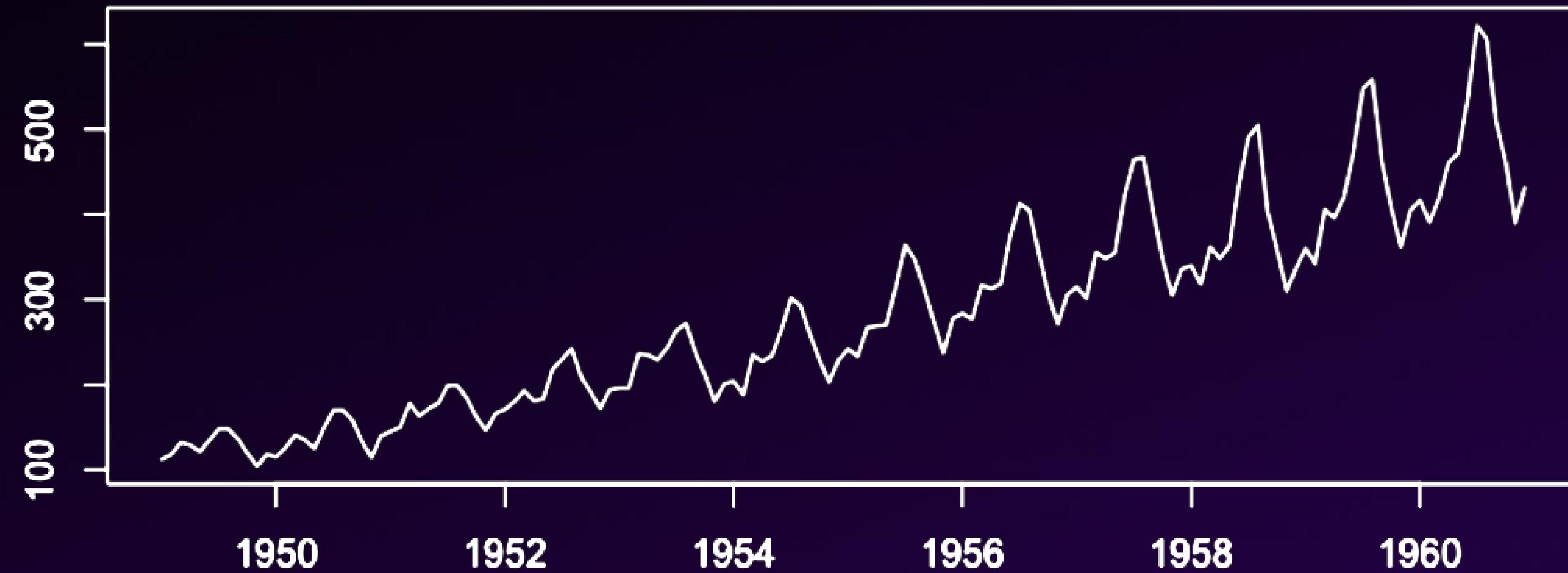
The Kalman Filter



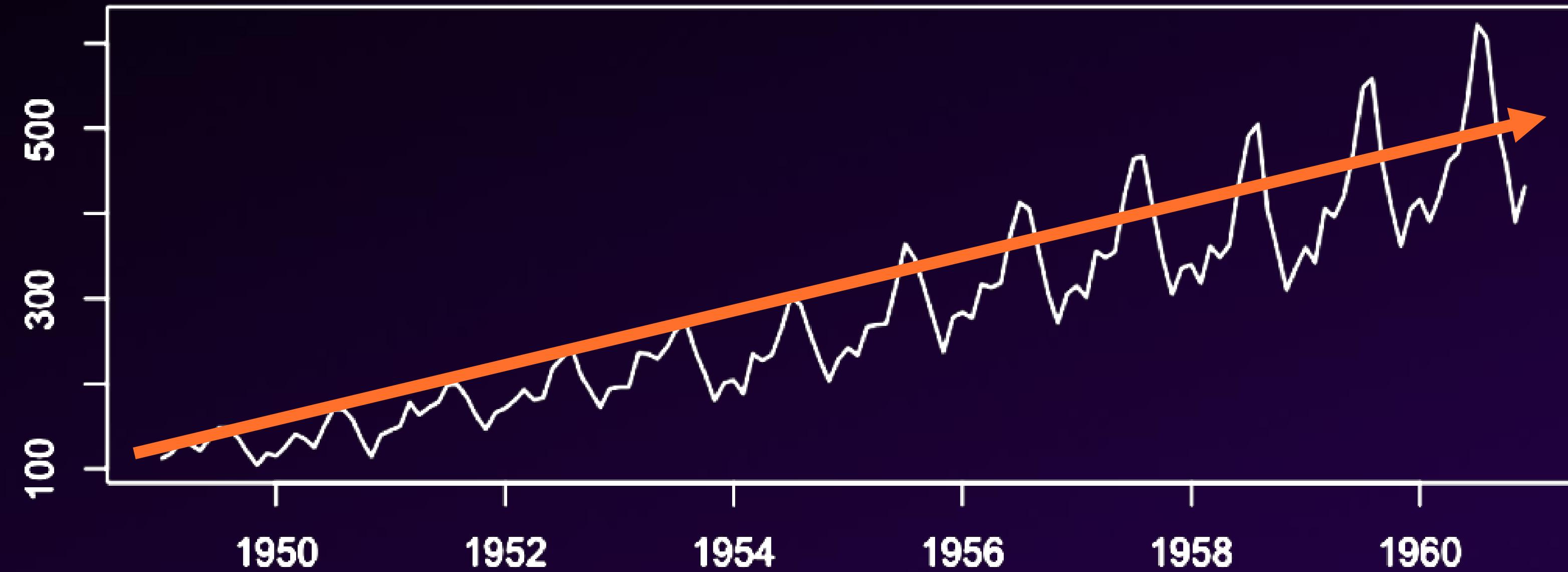
The Kalman Filter



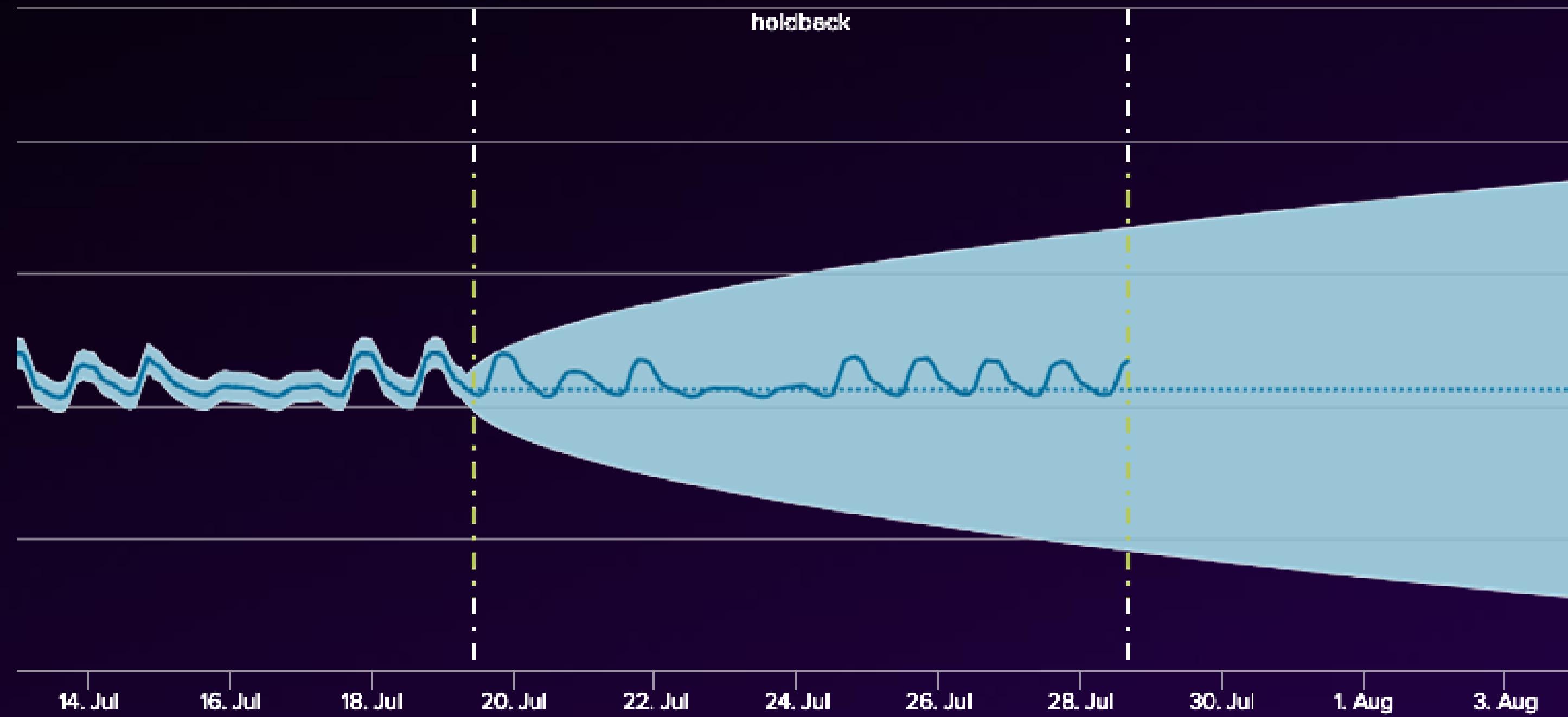
The Kalman Filter



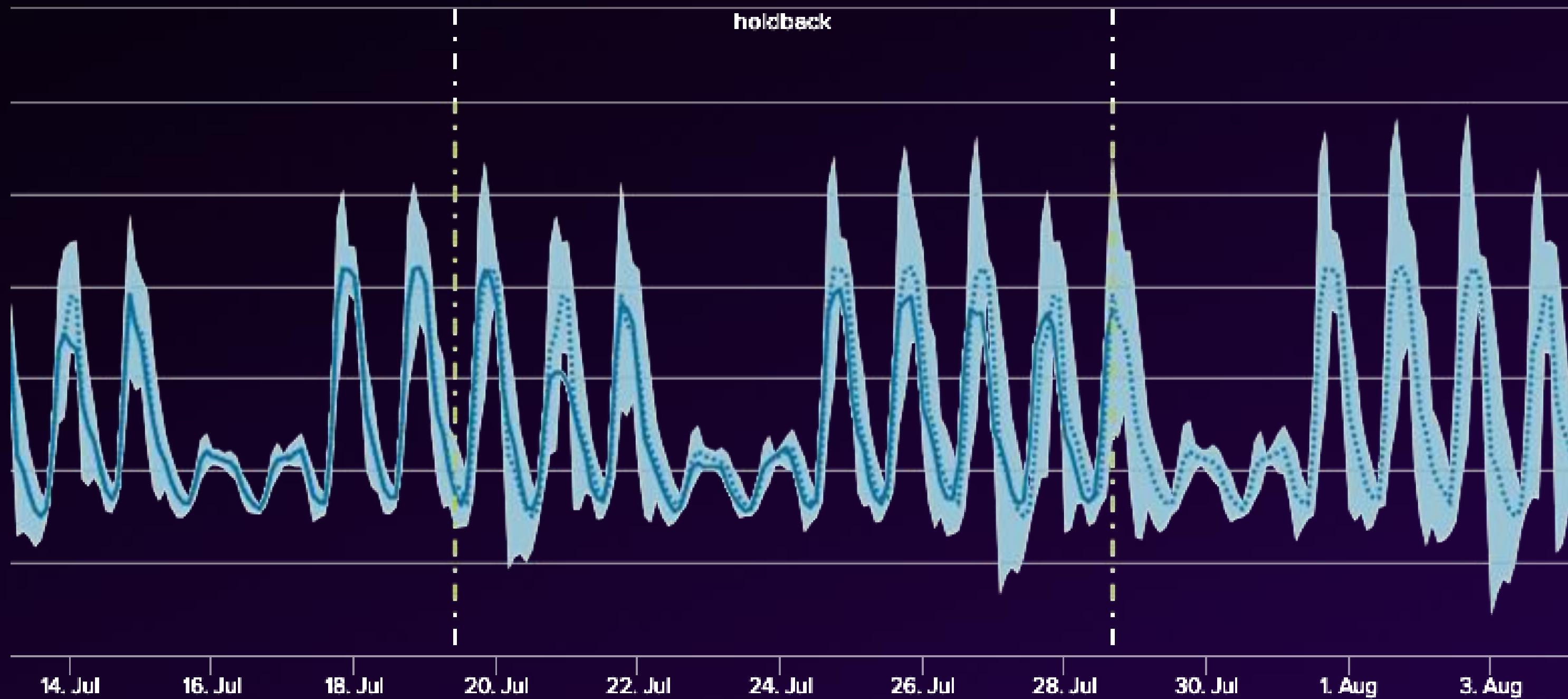
The Kalman Filter



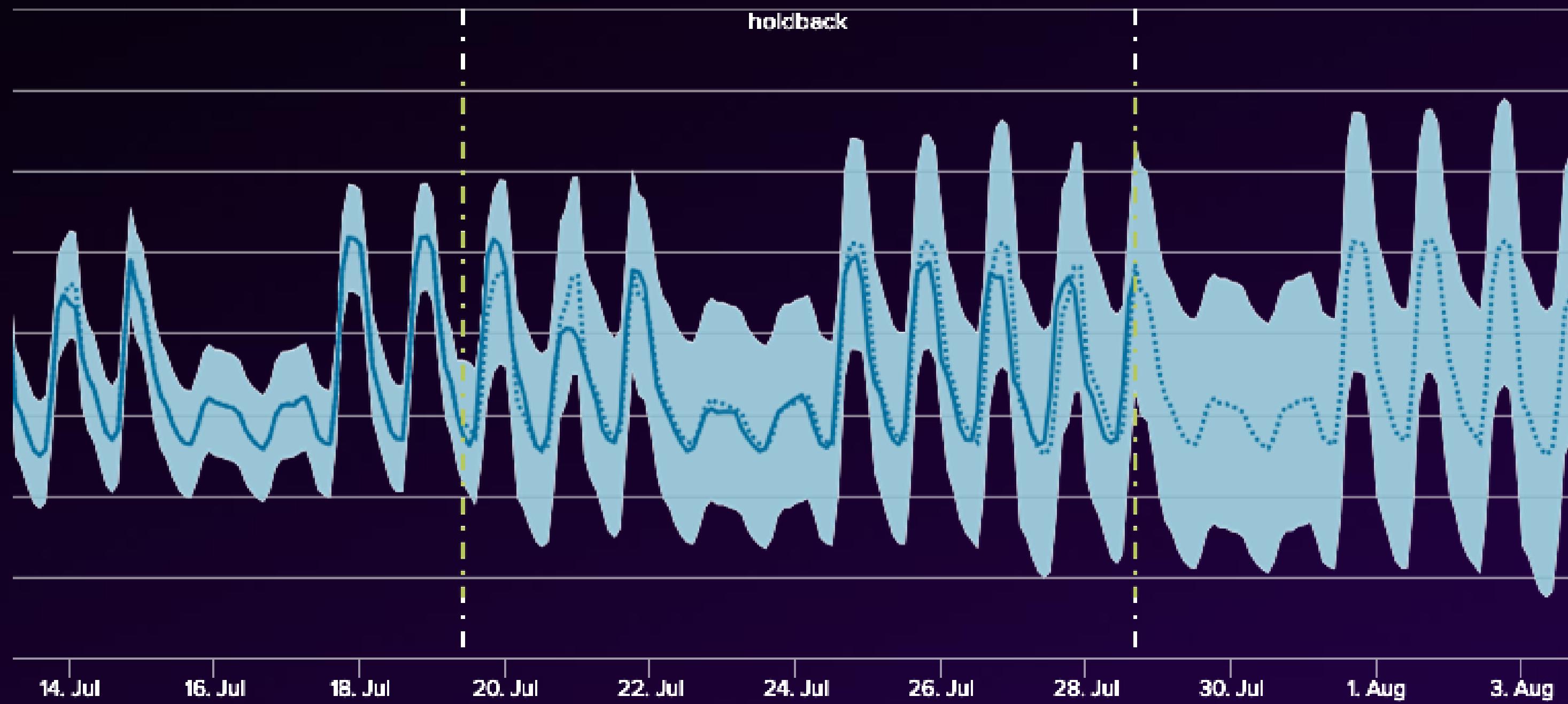
The Kalman Filter



The Kalman Filter



The Kalman Filter



Forecasting Time Series

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



Live Instance Demo

Log Into [INSTANCE URL]

Exercise #2

Time: 10 minutes

Summary

Top 4 most important things to remember about forecasting time series

1

Forecasting time series
is done using a
supervised learning
method

2

Models **assume historic
data as a baseline**,
and will self-correct
accordingly

3

Parameters have a
**large impact on
performance**. Tuning
each model is highly
recommended

4

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

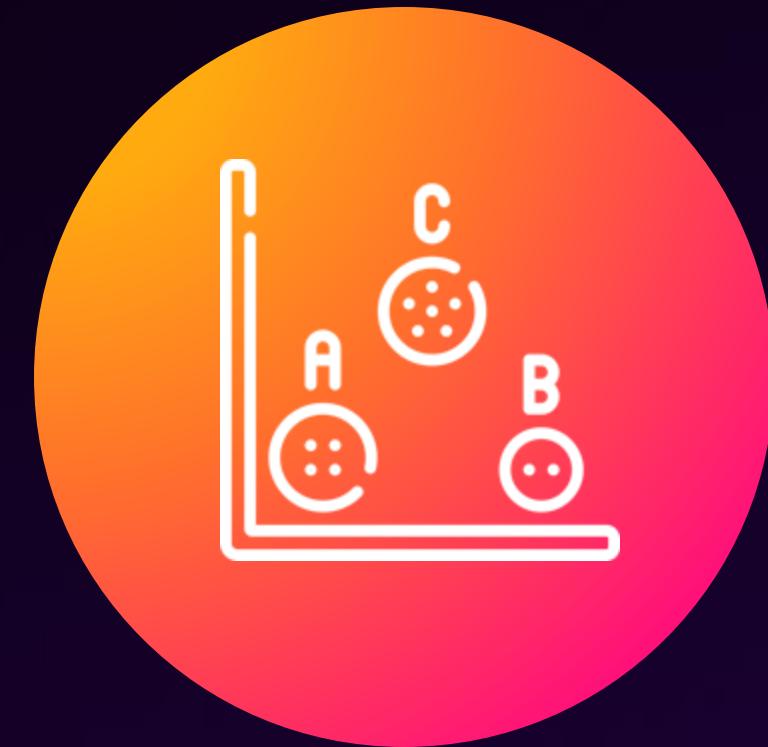
And we're done!

Tools in Your ML Toolkit Now



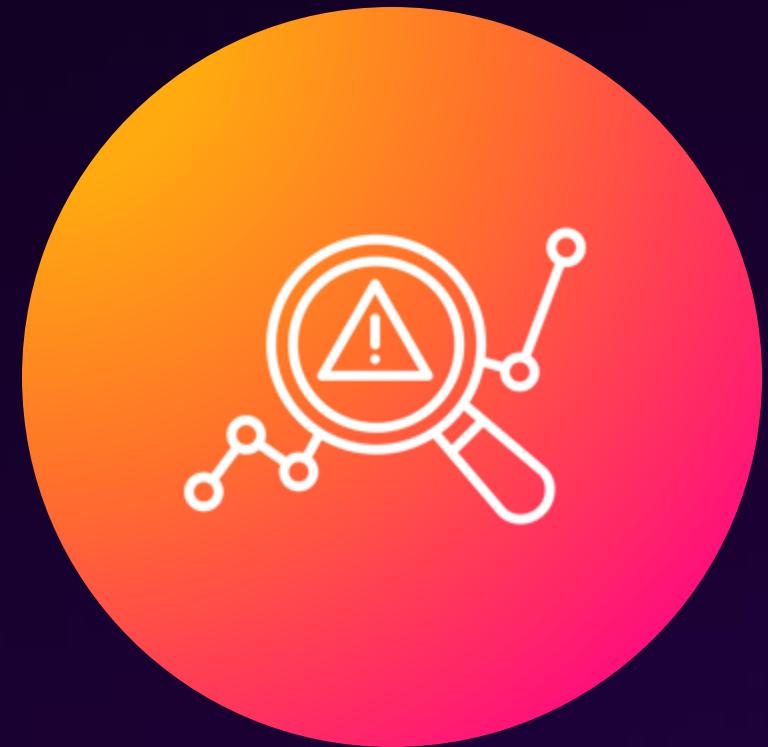
Prediction

Get ahead of issues that may happen in the future



Categorization

Uncover insights about your data to quickly respond in the present



Outlier Detection

Identify and analyze abnormal behavior in your data

Splunk Artificial Intelligence for Observability

[Link to landing page](#)

The table below shows how the following use cases map to the different techniques.

Use Case	Anomaly Detection	Predictive Analytics	Clustering
1. Forecasting Resource Utilization		✓	
2. Detecting Service Performance Issues	✓	✓	
3. User Experience Monitoring	✓	✓	
4. Noise Reduction			✓
5. Predicting Data Downtime in Splunk	✓	✓	
6. Predictive Maintenance	✓	✓	✓
7. Cell Tower Monitoring	✓		
8. Geohazards Monitoring	✓		

Security Use Cases Enhanced by AI and ML



[Link to landing page](#)

Use Case	Anomaly Detection	Predictive Analytics	Clustering	Graph Analytics	Generative AI
1. Identifying User Access Anomalies	✓				
2. Spotting Potential Insider Threats	✓		✓		
3. Detecting Domain Generation Algorithms (DGA)s		✓			
4. Finding Command Line Anomalies	✓	✓			
5. Using ML for Threat Hunting	✓		✓	✓	✓
6. Detecting Malicious Patterns of Network Traffic	✓				
7. Detecting Fraudulent Activity	✓		✓	✓	✓
8. Predicting Data Downtime in Splunk	✓	✓			
9. Demystifying Security Searches with the Splunk AI Assistant					✓

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