Exploring and Analyzing Data with Splunk

splunk> a CISCO company

Document Usage Guidelines

- Instructor required; not intended to be a self-paced document
- Intended for enrolled students only
- Do not distribute

Before Taking This Course

To be successful, students must have a working understanding of these courses.

- Intro to Splunk
- Using Fields
- Scheduling Reports and Alerts
- Visualizations
- Working with Time
- Statistical Processing
- Comparing Values
- Result Modification
- Leveraging Lookups and Sub-searches
- Correlation Analysis
- Search Under the Hood
- Intro to Knowledge Objects
- Creating Field Extractions
- Search Optimization

Course Objectives

- Understand data science and machine learning
- Utilize SPL and visuals to better understand data and relationships
- Group together data points by behavior to uncover new trends
- Examine the similarity between text strings to identify common and rare occurrences
- Correlate data to examine the influence fields have on one another
- Create efficient high-level transactions
- Identify numerical and categorical anomalies
- Predict future values in a time series

Course Outline

- Topic 1: What is Data Science
- Topic 2: Exploratory Data Analysis
- Topic 3: Event Clustering
- Topic 4: Correlations and Transactions
- Topic 5: Anomaly Detection
- Topic 6: Forecasting

Topic 1: What is Data Science

Topic Objectives

- Goals of Data Science
- Understand the difference between AI and ML
- Examine how ML is being in used Splunk

What is Data Science?

Overlaps analytics: to extract **actionable insights** from data

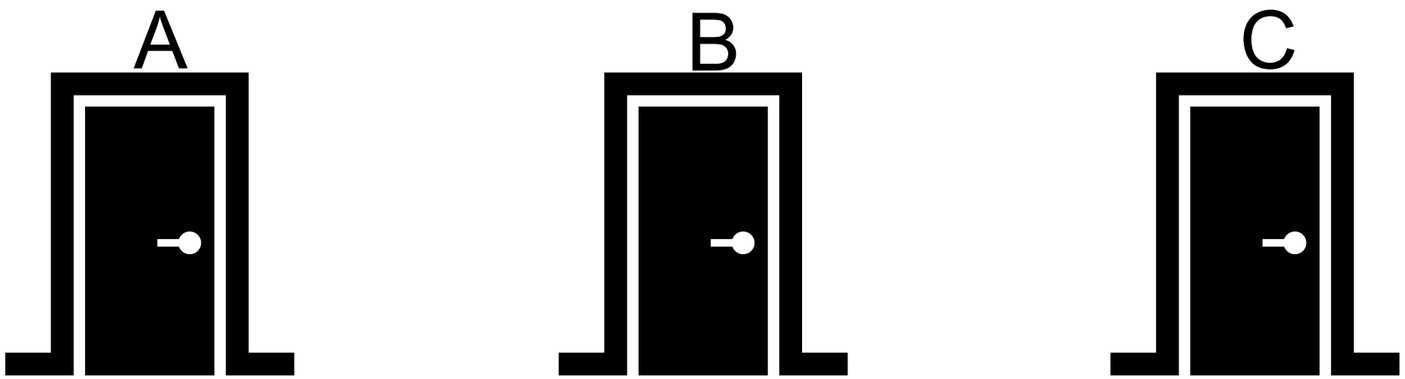
- Each data product is targeted to a specific persona
- Shifts between deductive (hypothesis-based) and inductive (pattern-based) reasoning

Explore the Data

- **First Rule of Data Science: Know your data.** Learn the shape of the data and look for unanticipated characteristics
  - Maximize insight into a data set
  - Uncover underlying structure
  - Test assumptions
- Visualize the data in multiple ways
  - Raw data (such as histograms)
  - Simple statistics such as mean, standard deviation, box plots, etc.
  - Adjust to maximize the brain's natural pattern-recognition abilities
- Target promising potential relationships

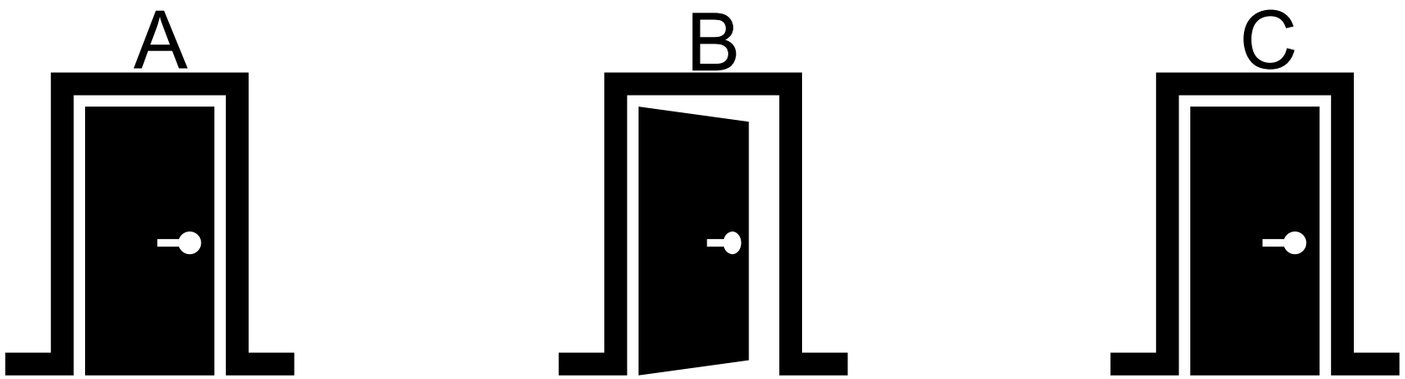
Statistics, Data Science, & Data Mining

- **Statistics:** a branch of math that provides theoretical and practical support for the use of tools and processes
  - Machine Learning: a process for generalizing from examples
- **Data Science:** using all available tools and processes to provide actionable insights to stakeholders in all organizational areas
- **Data Mining:** looking for useful information in large amounts of database data



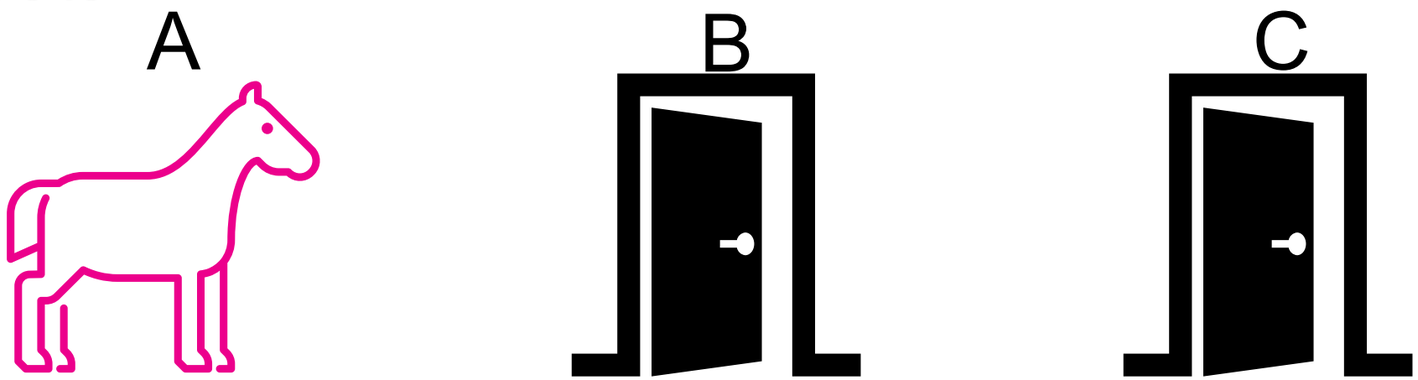
Statistics, Data Science, & Data Mining

- **Statistics:** a branch of math that provides theoretical and practical support for the use of tools and processes
  - Machine Learning: a process for generalizing from examples
- **Data Science:** using all available tools and processes to provide actionable insights to stakeholders in all organizational areas
- **Data Mining:** looking for useful information in large amounts of database data



Statistics, Data Science, & Data Mining

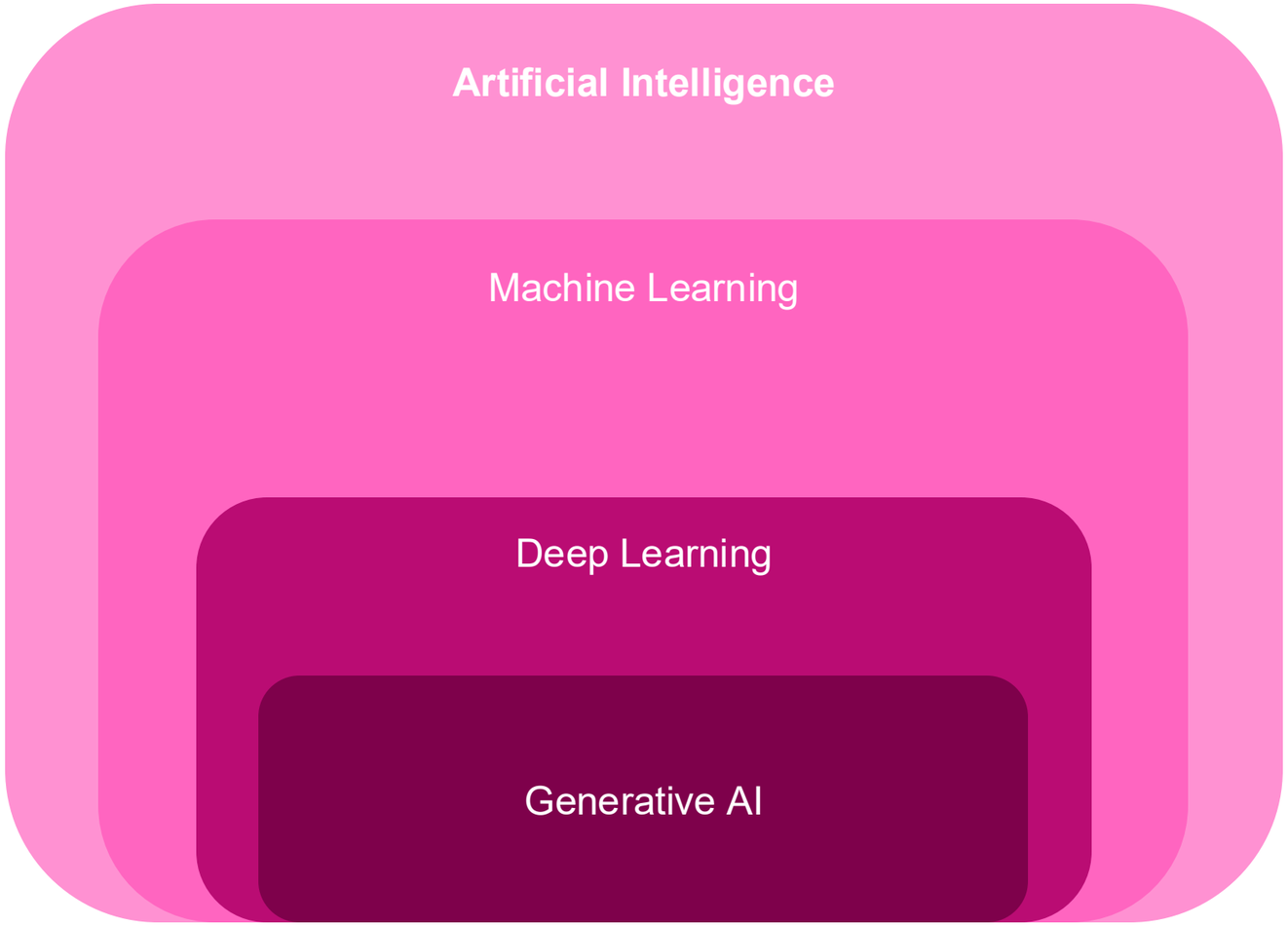
- **Statistics:** a branch of math that provides theoretical and practical support for the use of tools and processes
  - Machine Learning: a process for generalizing from examples
- **Data Science:** using all available tools and processes to provide actionable insights to stakeholders in all organizational areas
- **Data Mining:** looking for useful information in large amounts of database data



Knowing the Difference: AI vs. ML

Artificial Intelligence	Machine Learning
Development of systems capable of reasoning and problem-solving	Use of statistical models that enable systems to learn and make predictions on some data
May or may not depend on large amounts of data depending on the approach	Heavily dependent on data for training, validation, and testing of models
Typically has a degree of adaptability with self-learning mechanisms	Can be improved when introduced to new data, but requires retraining or fine-tuning implemented by a human
Wide range of applications with robotics, healthcare, finance, gaming, automation and more	Commonly used for predictive analytics, pattern recognition, natural language processing, and recommendation systems

Types of Artificial Intelligence



Generative AI sits within deep learning, which is a subset of machine learning, which sits within artificial intelligence.

Four Myths of AI and Machine Learning

1. AI is a magic wand
2. You need a Ph.D. to benefit from AI and ML
3. AI and ML will replace me
4. You need to have terabytes worth of data

Four Myths of AI and Machine Learning

AI is a magic wand

- AI enabled hair dryers and yoga pants may sound exciting, but there is a time and place for AI
- AI requires proper training to be useful
  - Spam filters must be trained on how to recognize a good email from a bad one
- Volume is important, but quality even more so
  - “garbage in, garbage out”
  - We need to train on the right data so the model can do its job
- When data is miscategorized, we have the opportunity to retrain and enhance the model

Four Myths of AI and Machine Learning

You need a Ph.D. to benefit from AI and ML

- Building an algorithm is complex, building a model is far more accessible
- There are countless opensource algorithms already available for handling different types of jobs
  - We need to recognize what algorithm is best for the job
  - We need to properly clean and prepare data for training
- No business or use case is too small to merit an investment in AI and ML
  - Identify anomalous security events
  - Forecast sales/demand
  - Analyze customer feed back and reviews

Four Myths of AI and Machine Learning

AI and ML will replace me

- McKinsey suggests that by 2030 there will be an additional 12 million occupational transitions
  - That doesn't mean 12 million fewer jobs
  - AI and ML still needs someone to develop, deploy, manage, and maintain it
- Generative AI (chatGPT) can enhance and streamline a lot of repetitive work
- AI can scour data, look for details that would take a us months to uncover, and we then need to check the the AI's results are on target
- AI doesn't have innate knowledge of business strategy, process, or implementation
  - We need human intervention and analysis to keep our AI relevant

Four Myths of AI and Machine Learning

You need to have terabytes worth of data

- The consensuses of the scientific community is that you need at least 50 data points to start working with
  - AI does thrive on large, accurate pools of data
  - If you have more, then excellent
- Much of this data is already in Splunk
  - This can include sales data, event logs, performance metrics
  - It just needs to be cleaned and prepared
- Not all data is needed, often we only train models on recent data that best represent the current world

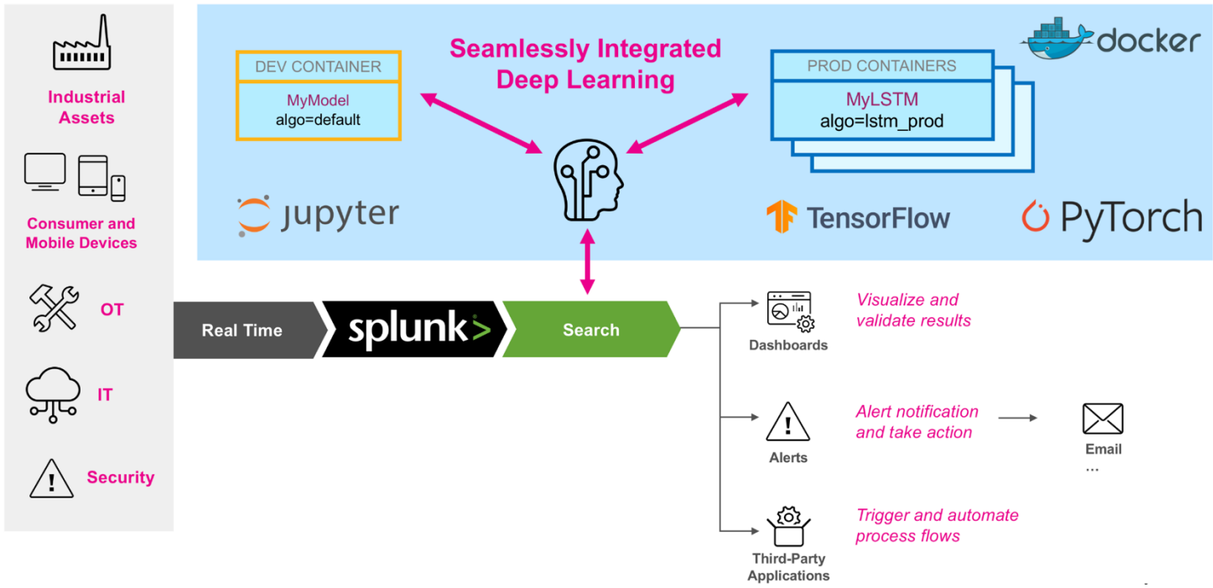
Machine Learning Toolkit

- Guided modeling Assistants that help build and test models with minimally SPL
- Packaged with over 30 common algorithms for clustering, anomaly detection, and prediction
- Ability to import your own Python based algorithms through the API
- Customs search commands and visualizations to build and explore your models
- Ability to upload externally trained models saved in an Open Neural Network Exchange (ONNX) format

https://splunkbase.splunk.com/app/2890

Splunk App for Data Science and Deep Learning

- Rapid model development workflows leveraging Jupyter Lab Notebooks
- Extension of the MLTK allowing you to better develop custom analytics with high computational workloads
- Connect your Search Head to container environments including Docker and Kubernetes with optional GPU support
- Access to pre-built containers including Golden Image GPU



https://splunkbase.splunk.com/app/4607

NLP Text Analytics

- Utilizes Python natural language processing libraries (NLTK 3.4.5) to analyze any text a user points to
- Expands upon the Machine Learning Toolkit's functionality
- Comes packaged with additional Splunk commands and ML algorithms
  - cleantext: tokenize, lemmenize, and normalize text
  - similatiry: measures the distance between texts and the number of steps required to make the text the same

https://splunkbase.splunk.com/app/4066

ML in Enterprise Security

- The MLTK's DensityFunction powers correlation search macros and workflow
  - Correlation searches do NOT group together related events
  - Correlation searches DO find numerical outliers within a dataset
- Correlation search workflow can take lookups or data models
  - The Splunk App for Common Information Model is often used as the provided data model
- The workflow allows you to define the time range, additional search filters, aggregate functions, and split-by fields

ML in IT Service Intelligence

- Anomaly Detection and Predictive – uses the MLTK to analyze system performance metrics, network traffic, and user behavior. Then uses adaptive thresholds to set alert levels
- Root Cause Analysis – Automated event correlation, incident prioritization, and integrations. This helps identify relationships between different fields and events so that IT teams can pinpoint the underlying cause
- Pattern Recognition – Identify recurring patterns and trends in IT data that may indicate specific types of issues or opportunities for improvement

Machine Learning Success Stories

Company	Product	Description
Google	PageRank (Search)	Exploit matrix factorization for good search results
Waymo	Self-Driving Cars	Take data from environment, fast vehicular response
American Express	Fraud Detection	Use historical purchase and vendor history to better identify and block fraudulent transactions
Facebook	News Feed	Use social network and “like” behavior to recommend content
T-Mobile	Messaging Service Outages	Use anomaly detection to quickly isolate incidents and minimize outages
Netflix	Movie Recommendation Engine	Use personal movie reviews to recommend new movies to watch
Quest Visual	WordLens	OCR language translation using phone’s built-in camera

Topic 2: Exploratory Data Analysis

Topic Objectives

- Use `bin` and `makecontinuous` to restructure and visualize data
- Examine field statistics with `fieldsummary`
- Transform fields with `eval` and `fillnull`
- Clean text with the `rex` and `cleantext` commands
- Apply boxplots and 3d scatterplots to visualize data

Analytics Framework

Analytics is a discipline based on people, processes, and technologies

Covered in this class

1. Ask a business-relevant question & gather requirements
2. Prepare data
3. Explore data
4. Model and validate data
5. Visualize, communicate results

Covered in Splunk for Analytics and Data Science

Exploratory Data Analysis

- Exploring data without first building a model
- Splunk
  - Helps you navigate and explore the data
  - Allows the relationships between the data to rise to the surface
- Once you find those relationships between the data:
  - You can tell stories about the data
  - The *stories* act as a proxy for the model

Preparing and Cleaning Data

Cleaning data is modifying the raw data to make it consistent and organized enough to input to your analytical algorithm

- Also called munging or wrangling
  - Sometimes creates some derived data, such as creating unique identifiers
  - Sometimes fills null or missing values
- Making data “tidy” makes it possible to fit a model
  - Variables or features should be the columns
  - Samples or data points should be the rows
- 80% is a common estimate of data science time spent munging

Anscombe's Quartet

- Statistical analysis on it own can be misleading and not provide an accurate description
- Francis Anscombe designed four data sets with identical descriptive statistics
  - mean, variance, correlation
- When graphed these datasets have very different underlying patterns and relationships
- Correlation coefficients are meant to be used when data is linear

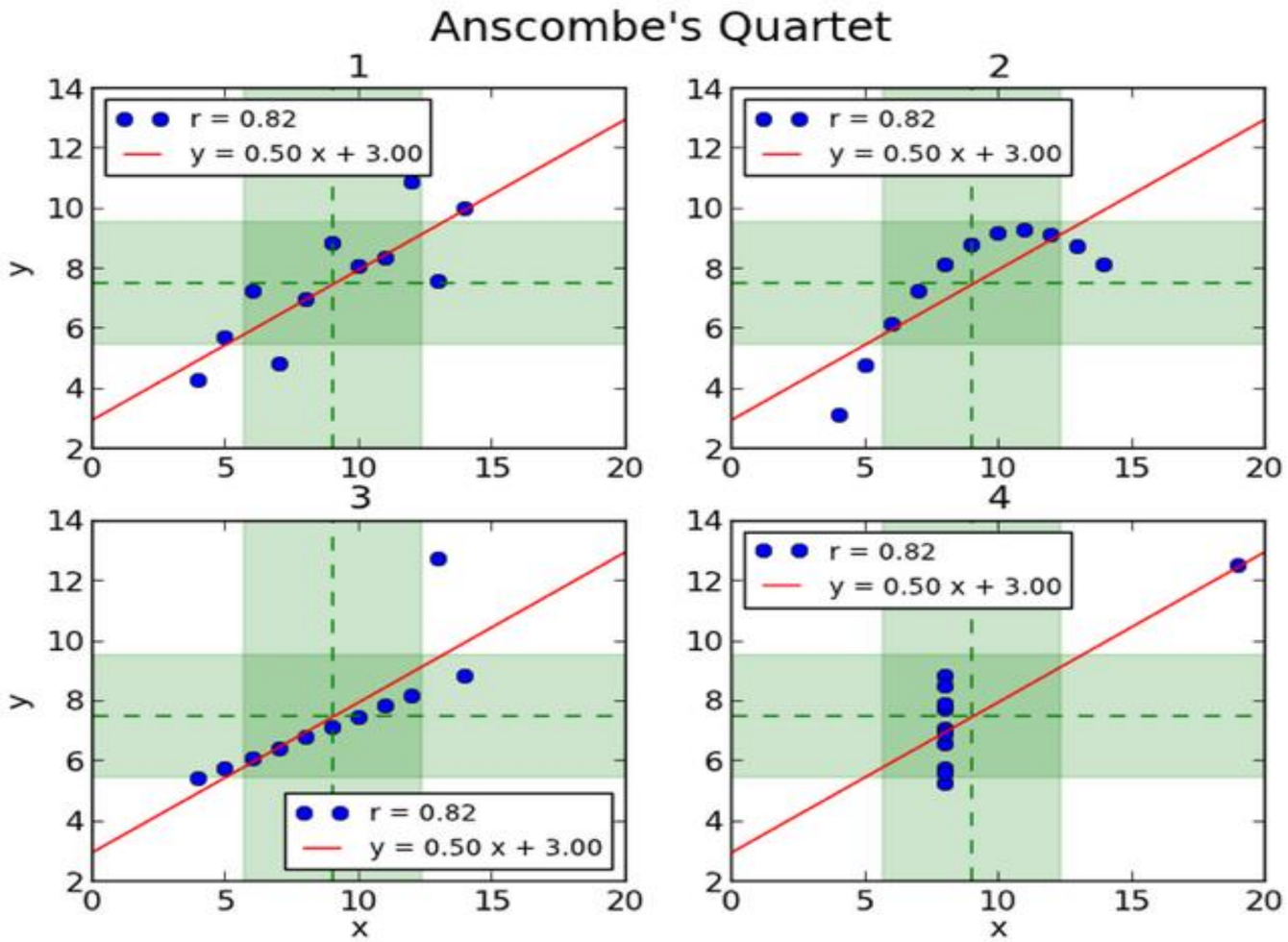
data set 1		data set 2		data set 3		data set 4	
X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

.81 .81 .81 .81

Correlation Coefficient

Wikipedia : dataMind.co

Anscombe’s Quartet



`bin` Command (`bucket`)

- Puts continuous numerical values into discrete sets, or bins
- Adjusts value of a numerical field so all items in a set have the same value range
  - The **`span`** argument of **`timechart`** also creates bins

```
bin [<bin-options>...]
<field> [AS <newfield>]
```

- Options
  - **`bins`** maximum number of bins (**`bins=20`**)
  - **`span`** sets the size for each bin (**`span=250`**)
    - If **`span`** creates more sets than the max specified by **`bins`**, **`bins`** is ignored
  - **`minspan`** specifies the smallest span size to use for each bin
  - **`<start-end>`** specifies minimum and maximum for numerical bins
    - When no span is used

bin Command Examples

```
sourcetype=access_combined
| stats sum(price) as totalSales by product_name
| bin totalSales bins=10
| stats list(product_name) as product_name by totalSales
| eval totalSales = "$".totalSales
```

totalSales	product_name
$0-1000000	Curling 2014 Fire Resistance Suit of Provolone Holy Blade of Gouda Manganiello Bros. Tee Puppies vs. Zombies World of Cheese Tee
$1000000-2000000	Benign Space Debris Final Sequel Mediocre Kingdoms SIM Cubicle
$2000000-3000000	Manganiello Bros. Orvil the Wolverine World of Cheese
$3000000-4000000	Dream Crusher

```
sourcetype=access_combined
| bin _time span=1h
| stats sum(bytes) as totalBytes
by _time, host
| eval totalBytes = round(totalBytes/(1024*1024),2)." MB"
| xyseries _time, host, totalBytes
```

_time	www1	www2	www3
2018-06-08 04:00	0.20 MB	0.18 MB	0.17 MB
2018-06-08 05:00	0.18 MB	0.31 MB	0.12 MB
2018-06-08 06:00	0.21 MB	0.14 MB	0.26 MB
2018-06-08 07:00	0.14 MB	0.21 MB	0.14 MB
2018-06-08 08:00	0.11 MB	0.20 MB	0.18 MB
2018-06-08 09:00	0.10 MB	0.16 MB	0.21 MB

`makecontinuous` Command

```
makecontinuous [<field>] <bin-options>
```

- Makes a field on the x-axis numerically continuous
  - Where no data exists, adds empty bins
  - Where there is data, quantifies the periods
  - Use **`chart`** or **`timechart`** to invoke this new x-axis value
  - Auto sorts the data

140-150	2
150-160	2
160-170	2
170-180	2
180-190	1
220-230	1
240-250	2

160-170	2
170-180	2
180-190	1
190-200	
200-210	
210-220	
220-230	1
230-240	
240-250	2
250-260	1

`fieldsummary` Command

- Summary stats for all or a subset of fields in your search results
- Summary information is displayed as a results table including:

`field`	field name in the event
`count`	number of events/results with that field
`distinct_count`	number of unique values in the field*
`is_exact`	whether or not the field is exact
`max`	if numeric, the maximum of its value
`mean`	if numeric, the mean of its values
`min`	if numeric, the minimum of its values
`numeric_count`	count of numeric values in the field (excludes NULL values)
`stdev`	if numeric, the standard deviation of its values
`values`	distinct values of the field and count of each value

`fieldsummary` Syntax & Example

```
fieldsummary [maxvals=<num>] [<wc-field-list>]
```

- Optional arguments
  - **`maxvals`** max distinct values to return for each field (default: 100)
  - **`wc-field-list`** field(s) that can include fields with wildcards

```
sourcetype=sendmail_syslog | fieldsummary maxvals=3
```

field	count	distinct_count	is_exact	max	mean	min	numeric_count	stdev	values
change_type	0	0	1				0		[]
class	676	1	1	0	0	0	676	0	[{"value":"0","count":676}]
ctladdr	618	8	0				0		[{"value":"britany@mailsv1.splunk.com (665/666)","count":178}, {"value":"hammer@mailsv1.splunk.com (967/967)","count":126}, {"value":"madonna@mailsv1.splunk.com (662/663)","count":124}]
daemon	672	1	1				0		[{"value":"MTA","count":672}]

`autoregress` Command

- Will clone a field and offset the values by **p** number of events/rows
- Can link a previous field value with the next event in a series
- Allows us to calculate changes over time
- A **p** attribute allows us to offset the value my multiple rows

```
sourcetype=access_combined | timechart sum(price) as sales | autoregress sales | eval diff = sales - sales_p1
```

_time	sales	diff	sales_p1
2024-04-23 20:00:00	1141.34		
2024-04-23 20:30:00	1313.45	172.11	1141.34
2024-04-23 21:00:00	719.59	-593.86	1313.45
2024-04-23 21:30:00	1187.52	467.93	719.59
2024-04-23 22:00:00	1035.53	-151.99	1187.52
2024-04-23 22:30:00	880.59	-154.94	1035.53
2024-04-23 23:00:00	1000.55	119.96	880.59

fillnull & filldown Commands

fillnull [<field-list> value=<string>]

- Fills in null values with a zero or user defined value

```
| timechart sum(price) by product_name |fillnull value=NA
```

_time	Benign Space Debris	Curling 2014	Dream Crusher	Final Sequel
2024-04-23 20:00:00	NA	39.98	NA	NA
2024-04-23 20:30:00	NA	NA	79.98	49.98
2024-04-23 21:00:00	NA	NA	39.99	NA
2024-04-23 21:30:00	NA	NA	NA	NA
2024-04-23 22:00:00	NA	NA	NA	49.98

filldown [<field-list>]

Fills in null values using the first, previous, non-null value.

- Null fields can remain if there is no prior event with a value

```
| timechart sum(price) by product_name |filldown
```

_time	Benign Space Debris	Curling 2014	Dream Crusher	Final Sequel
2024-04-23 20:00:00		39.98		
2024-04-23 20:30:00		39.98	79.98	49.98
2024-04-23 21:00:00		39.98	39.99	49.98
2024-04-23 21:30:00		39.98	39.99	49.98
2024-04-23 22:00:00		39.98	39.99	49.98
2024-04-23 22:30:00	24.99	39.98	39.99	49.98

Converting categorical fields into binary

- With eval we can break apart a categorical field into a series of binary fields, This is also known as **one-hot encoding**
- For each categorical value, a new field will be created set to 1 or null
- `fillnull` should be used to complete the conversion to binary fields

```
sourcetype=access_combined action=*
| table action
| eval is_{action}=1
| fillnull
```

action	is_addtocart	is_changequantity	is_purchase	is_remove	is_view
view	0	0	0	0	1
view	0	0	0	0	1
purchase	0	0	1	0	0
view	0	0	0	0	1
addtocart	1	0	0	0	0
purchase	0	0	1	0	0
purchase	0	0	1	0	0

rex Command

```
rex field=<text> mode=sed “<string>”
```

- The **rex** command has a sed mode
- **field** is set to the **_raw** field by default
- **<string>** is used to replace strings (s) or substitute characters (y)
  - The syntax for using sed to replace (s) text in your data is: **"s/<regex>/<replacement>/<flags>”**
  - The syntax for using sed to substitute characters is: **"y/<string1>/<string2>/"**

`rex` Command Example

```
| inputlookup airline_tweets.csv | fields text | eval orig_text = text
| rex field=text mode=sed "s/\W/ /g" | eval text = lower(text)
```

text	orig_text
virginamerica what dhepburn said	@VirginAmerica What @dhepburn said.
virginamerica plus you ve added commercials to the experience tacky	@VirginAmerica plus you've added commercials to the experience... tacky.
virginamerica i didn t today must mean i need to take another trip	@VirginAmerica I didn't today... Must mean I need to take another trip!
virginamerica it s really aggressive to blast obnoxious entertainment in your guests faces amp they have little recourse	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse
virginamerica and it s a really big bad thing about it	@VirginAmerica and it's a really big bad thing about it

`cleantext` Command (NLP Text Analytics)

```
| cleantext textfield=<field> keep_orig=<bool> remove_urls=<bool> remove_stopwords=<bool> base_word=<bool>
```

- Tokenize and normalize text
- Different options result in better but slower cleaning
- Returns results in a multi-value field
- **`textfield`** is the only required argument

- `keep_orig = false/true`
  - maintain a copy of the original text
- `remove_urls = true/false`
  - before cleaning, remove HTML links
- `remove_stopwords = true/false`
  - remove common English words
- `base_word = true/false`
  - turns on lemmatization or stemming
- `base_type = lemma/lemma_pos/stem`
  - lemma treats every word as a noun
  - lemma_pos is slower but more precise
  - stem attempts to strip suffixes from words

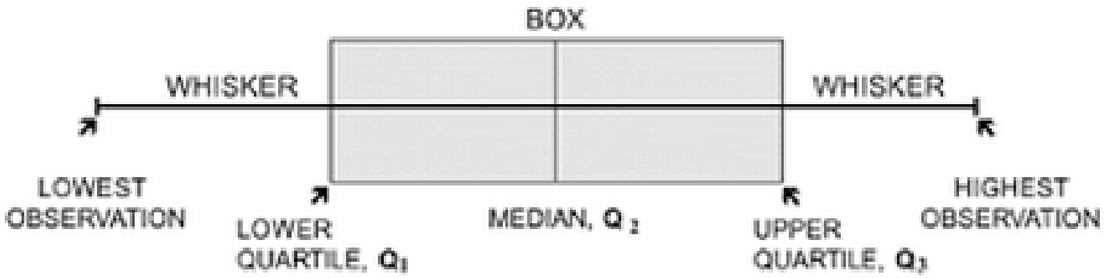
`cleantext` Example

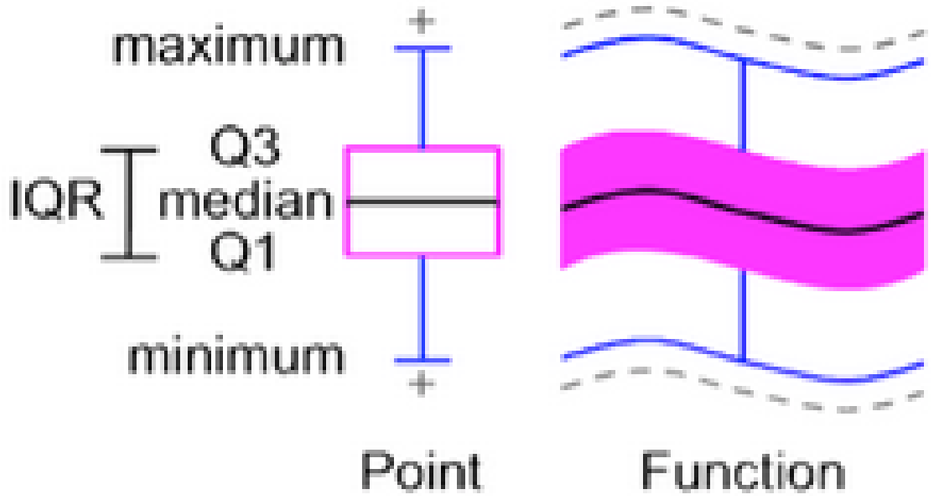
```
| inputlookup airline_tweets.csv | fields text
| cleantext textfield=text keep_orig=true remove_stopwords=false base_word=true base_type=lemma_pos
```

text	orig_text	pos_tuple	pos_tag
virginamerica	@VirginAmerica What @dhepburn said.	["virginamerica","NNP"]	NNP
what		["what","WP"]	WP
dhepburn		["dhepburn","NN"]	NN
say		["say","VBD"]	VBD
virginamerica	@VirginAmerica plus you've added commercials to the experience... tacky.	["virginamerica","NNP"]	NNP
plus		["plus","CC"]	CC
you		["you","PRP"]	PRP
add		["add","VBN"]	VBN
commercial		["commercial","NNS"]	NNS
to		["to","TO"]	TO
the		["the","DT"]	DT
experience		["experience","NN"]	NN
tacky		["tacky","NN"]	NN

MLTK `boxplot` Visualization

- A box and whisker plot
  - x-axis: category of comparison
  - center of boxes = median
  - most data is with the box spanning the 25th (Q1) and 75th (Q3) percentile
    - Also known as the interquartile range (IQR)
  - lowest and highest observations as indicated
    - often viewed as outliers





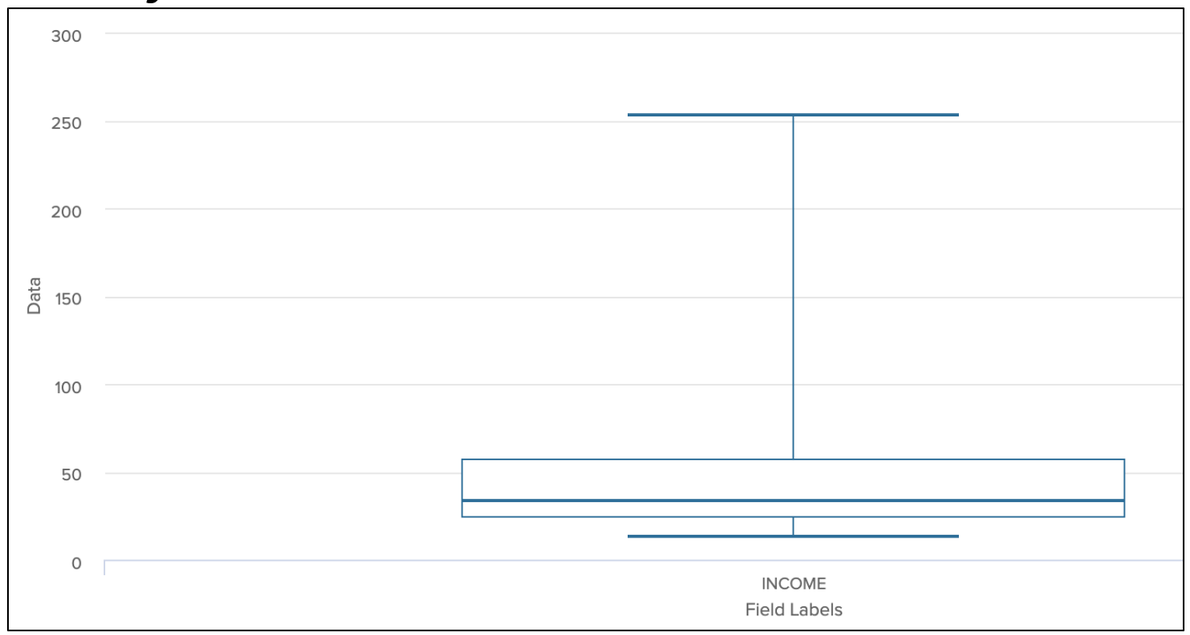
MLTK `boxplot` Macro

- A prebuilt `` `boxplot` `` macro allows for easier creation of the visual
- Otherwise, we can manually run a stats command to calculate the required values
- Visualization expects five rows corresponding to min, max, median, lower quartile and upper quartile, in any order
  - `exactperc25` is the lower quartile
  - `exactperc75` is the upper quartile

```
| inputlookup default.csv | fields INCOME | `boxplot`
```

`` `boxplot` `` macro expanded

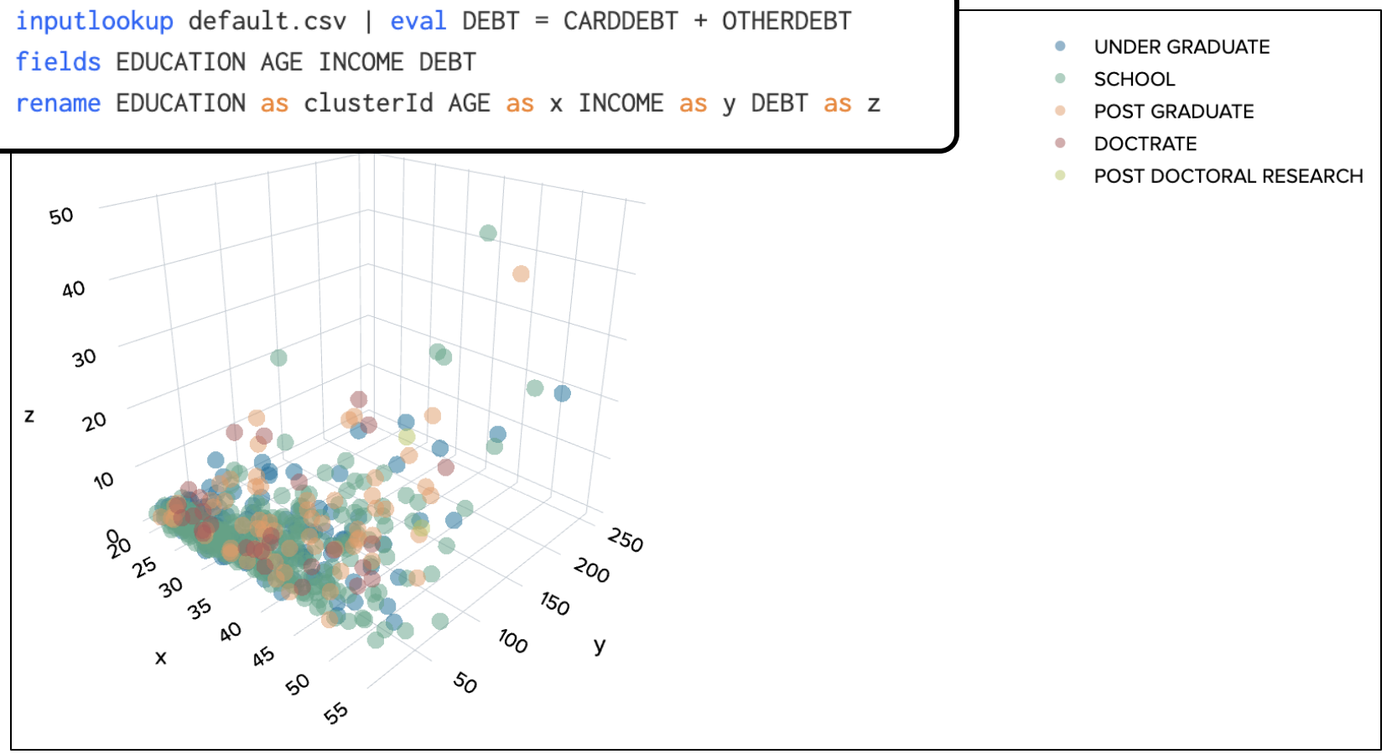
```
| untable _x field_name value
| stats min exactperc25 median exactperc75 max by field_name
| untable field_name calculations value
| xyseries calculations field_name value
| eval calculations = rtrim(calculations, "(value)")
```



MLTK `3d scatterplot` Visualization

- Look for clusters of similar data points
- Drilldown to identify singular data points
- Fields must be renamed to **`clusterId`**, **`x`**, **`y`**, and **`z`**

```
| inputlookup default.csv | eval DEBT = CARDDEBT + OTHERDEBT
| fields EDUCATION AGE INCOME DEBT
| rename EDUCATION as clusterId AGE as x INCOME as y DEBT as z
```



Topic 2 Lab

- Description: Clean and visualize data
- Duration: 20 minutes
- Tasks:
  - Examine basic stats with `fieldsummary`
  - Make a histogram visualization
  - Visualize order of office badge ins
  - Clean text based data

Topic 3: Event Clustering

Topic Objectives

- Take a behavioral based approach to cluster data
- Cluster numerical fields using the **`kmeans`** command
- Cluster based of string similarity with the **`cluster`** command
- Find patterns in text-based clusters

Rule Based Clustering

Rules (a priori)

- Easy to define using existing, discrete values in fields
- Achieved using a **by** clause in the **stats** or **chart** commands

...| stats count by state

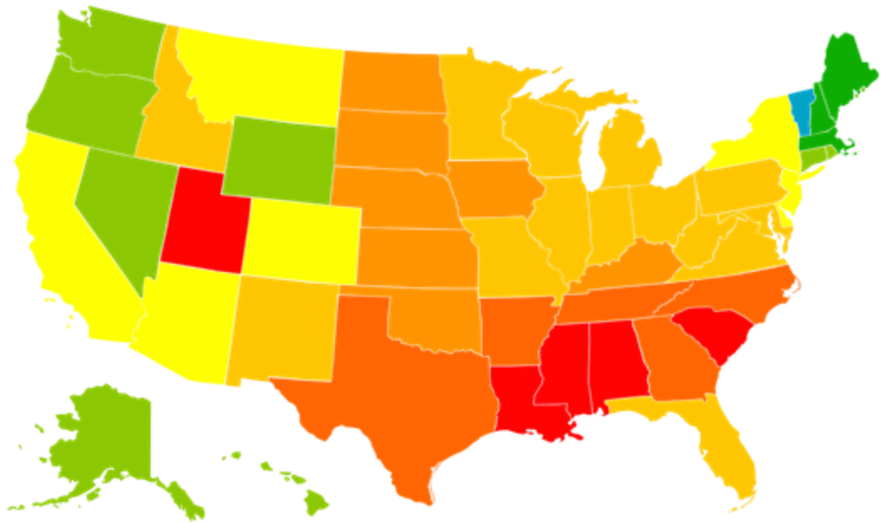
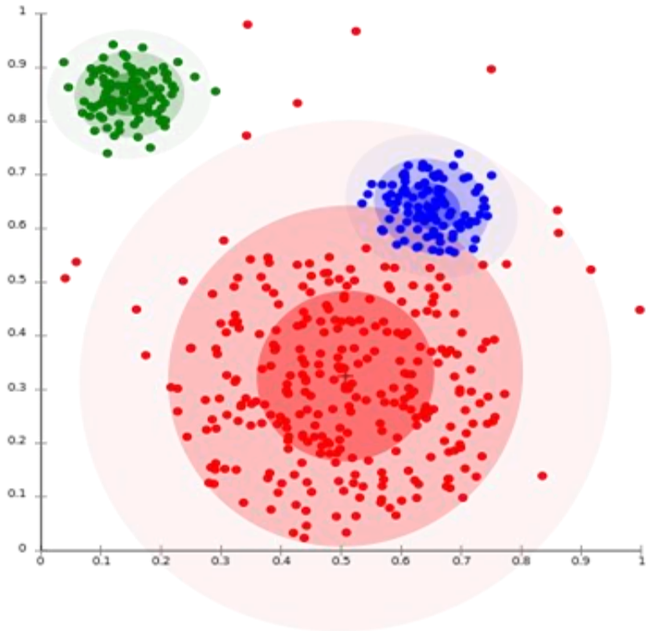


Image: U.S. segmentation. Data: church or synagogue attendance
Source:
http://en.wikipedia.org/wiki/Christianity_in_the_United_States#mediaviewer/File:Church_or_synagogue_attendance_by_state_GFDL.svg
Attribution: Creative Commons, Falcorian

Behavior Based Clustering

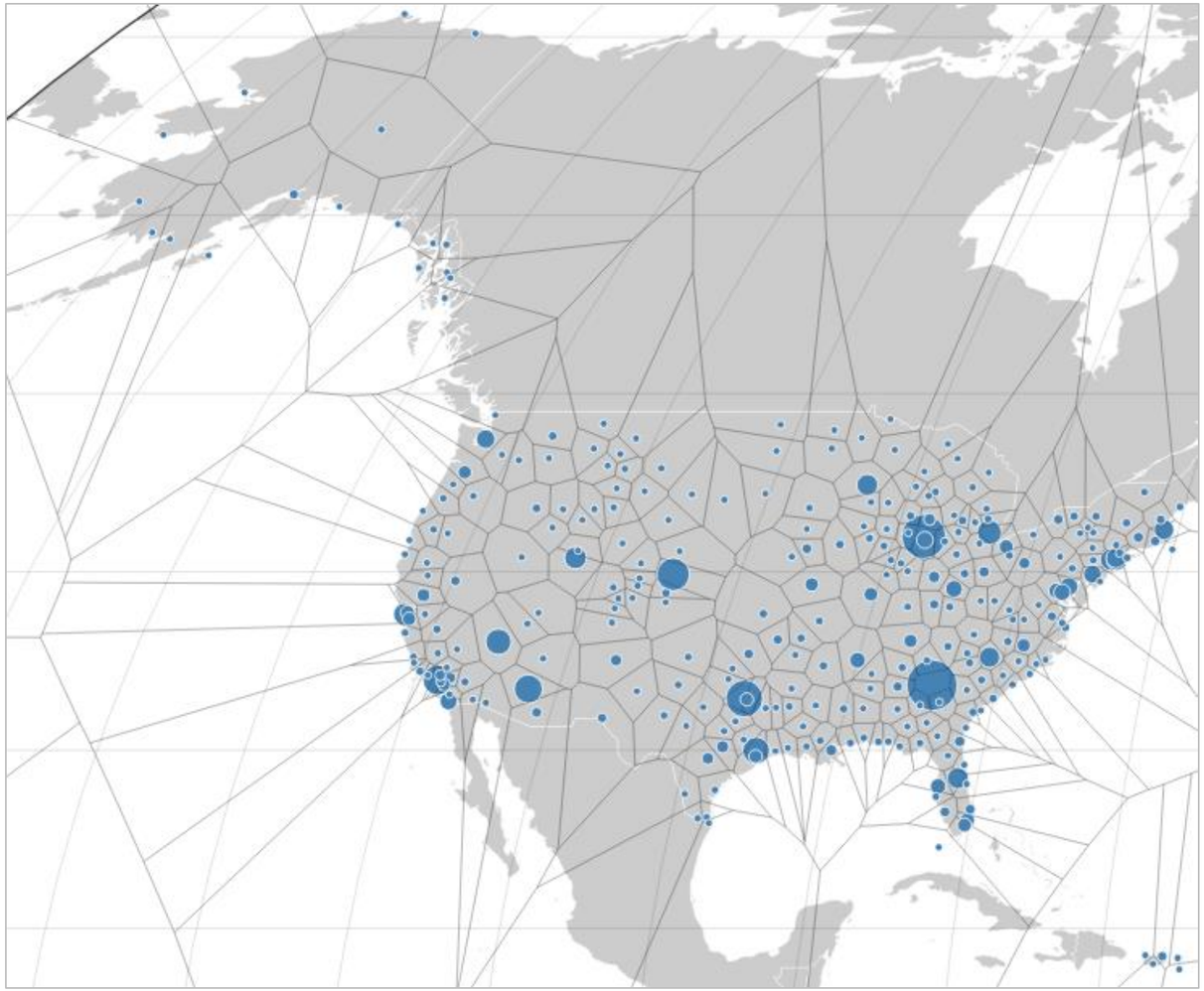
- Behavior (post hoc) or similarity-based grouping
  - Discover new groupings or combinations in the data
  - Actual groups are not known until clustering methods are applied
  - Groups are formed based on proximity to a central point or similarity
- Examples:
  - Window event log statuses
  - Server performance
  - Customer spending



`kmeans` Command

- Divides data into k clusters
- Each data point belongs to the cluster with the nearest mean (centroid)
- Partitioned into Voronoi cells
- Numerical fields common to both results sets are used

```
kmeans <fields-list>
[k=<num>] [options]
```



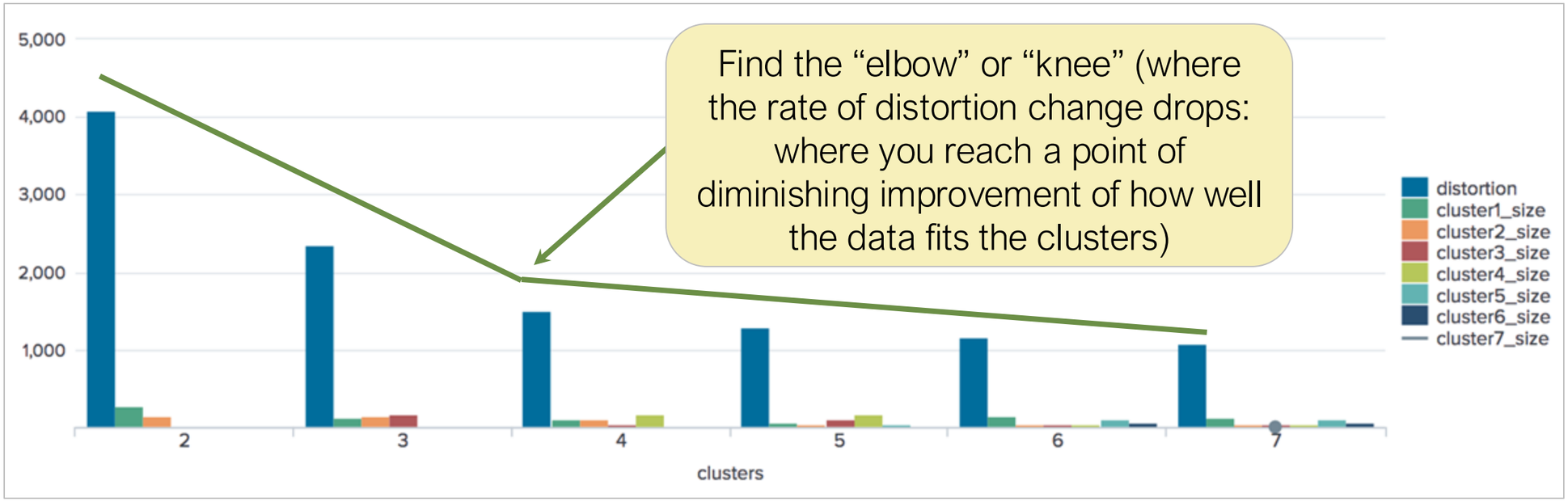
`kmeans` Command Options

- `k`: number of clusters to use or range of numbers (default: 2)
  - `range` clusters for each cluster count in the range, include the size of the clusters and a `'distortion'` field: how well the data fits those clusters
- `reps`: repeats of `kmeans` using random starting clusters (default:10)
- `iters`: maximum number of iterations allowed (default: 10,000)
- `t`: algorithm convergence tolerance (default: 0)
- `cnumfield`: names the field to annotate results (default: `CLUSTERNUM`)
- `dt` is the distance metric to use (default: `sqeuclidean`)
  - l1, l1norm, and cb: cityblock
  - l2, l2norm, and sq: sqeuclidean

Finding a value for k

Instead of a number, use a range for `kmeans` to examine options and optimize the number of market segments

```
sourcetype=access_combined action=purchase
| stats sum(price) as order_total by JSESSIONID
| kmeans k=2-7 order_total
```



Find the “elbow” or “knee” (where the rate of distortion change drops: where you reach a point of diminishing improvement of how well the data fits the clusters)

`kmeans` Command Example

- Do the clusters actually relate to anything?

```
sourcetype=access_combined action=purchase AND JSESSIONID=* AND product_name=* earliest=-7d@d
| fields price JSESSIONID product_name
| stats list(product_name) as product_name sum(price) as spending by JSESSIONID
| kmeans k=4 spending
| chart count by product_name CLUSTERNUM
```

product_name	1	2	3	4
Benign Space Debris	116	32	43	0
Curling 2014	136	39	40	0
Dream Crusher	0	102	231	0
Final Sequel	208	54	66	0
Fire Resistance Suit of Provolone	76	17	46	233
Holy Blade of Gouda	59	15	39	186
Manganiello Bros.	0	96	242	0
Manganiello Bros. Tee	36	12	57	197
Mediocre Kingdoms	235	62	92	0
Orvil the Wolverine	0	83	166	0
Puppies vs. Zombies	35	9	33	176
SIM Cubicle	241	58	55	0
World of Cheese	252	70	86	0
World of Cheese Tee	35	12	49	167

`cluster` Command

```
cluster [field=<field>] [options]
```

Groups events based on patterns it detects in event text streams

- Based on cosine similarity of the **`_raw`** field by default
  - Can be changed using **`field=<field>`** (i.e. **`field=action`**)
- Breaks fields into terms, computes the vector between events
- Creates 2 new fields named **`cluster_label, cluster_count`**
  - Default field names can be changed using **`labelfield=<field>`** and **`countfield=<field>`**

`cluster` Command Options

- `match=termlist` **(default)** breaks fields into words
  - Terms must appear in the same order to be considered matching
- `match=termset`
  - Terms can be in any order to be considered matching
- `match=ngramset`
  - Examines trigrams (3-character substrings) of characters
  - Most useful for short non-textual fields, like punct
- `labelonly=false` **(default)** one event from each cluster is returned
  - `labelonly=true` all events are returned
- `showcount=false` **(default)** hides the `cluster_count` field
- `t=.8` **(default)** adjusts threshold for cluster sensitivity

Cluster Scoring: `Termlist` vs. `Termset`

Termlist

Intersection of A and B = 1

$$\text{cosine}_{AB} = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}}$$

Number of terms in A Number of terms in B

EVENT	TEXT	# OF TERMS
A	INFO user generated content blocked by filter	7
B	WARN system generated alert from filter	6

$$= \frac{1}{\sqrt{7 \times 6}} \approx 0.1543$$

Termset

Intersection of A and B = 2

$$\text{cosine}_{AB} = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}}$$

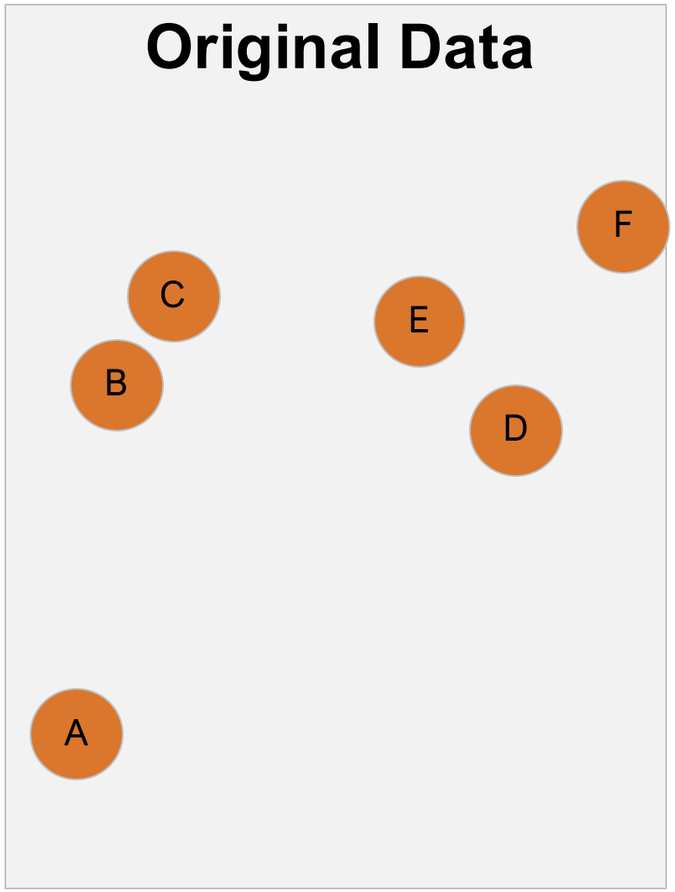
Number of terms in A Number of terms in B

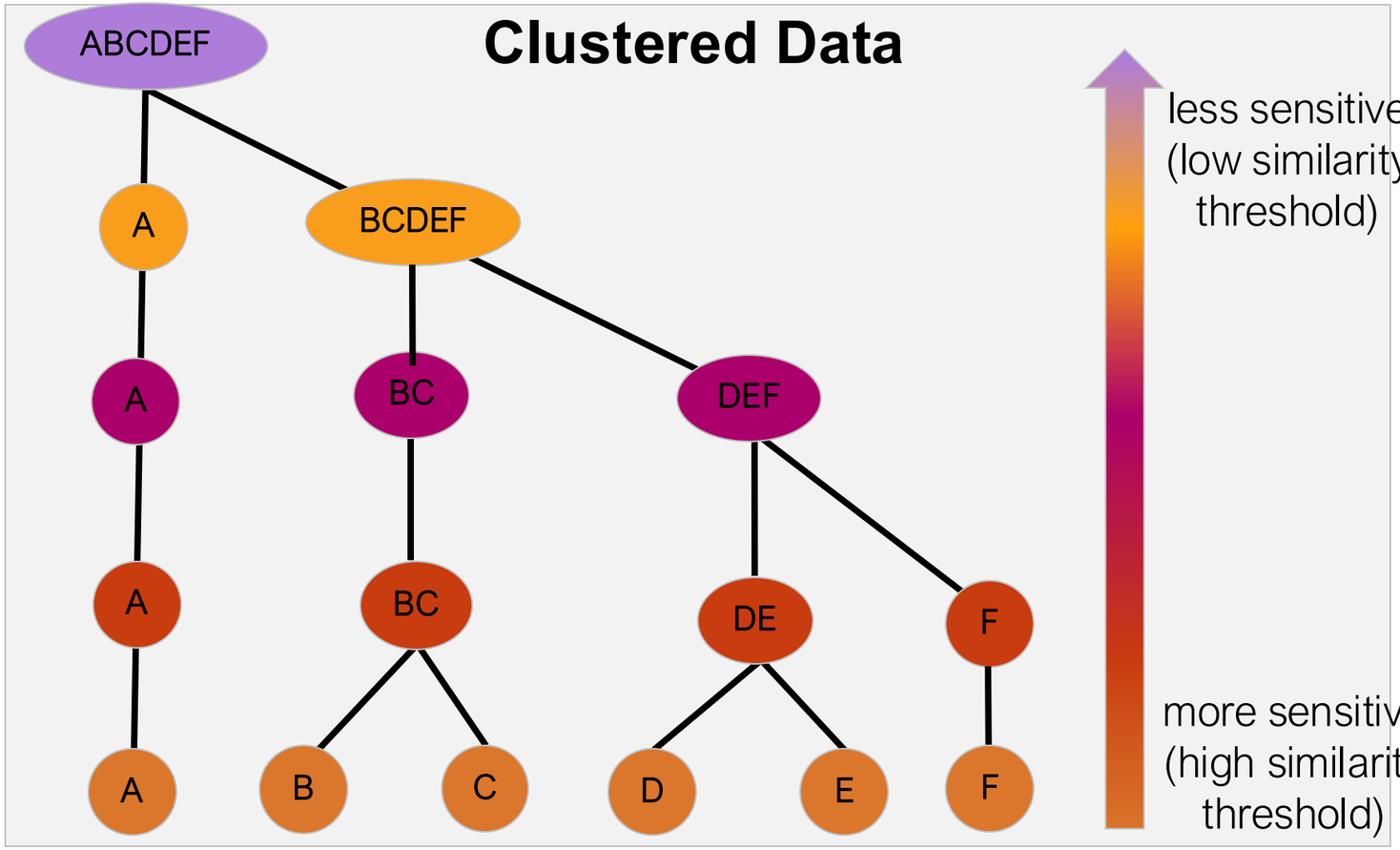
EVENT	TEXT	# OF TERMS
A	INFO user generated content blocked by filter	7
B	WARN system generated alert from filter	6

$$= \frac{2}{\sqrt{7 \times 6}} \approx 0.3086$$

`cluster` Command (Agglomerative)

Iteratively groups events based on a minimum similarity threshold





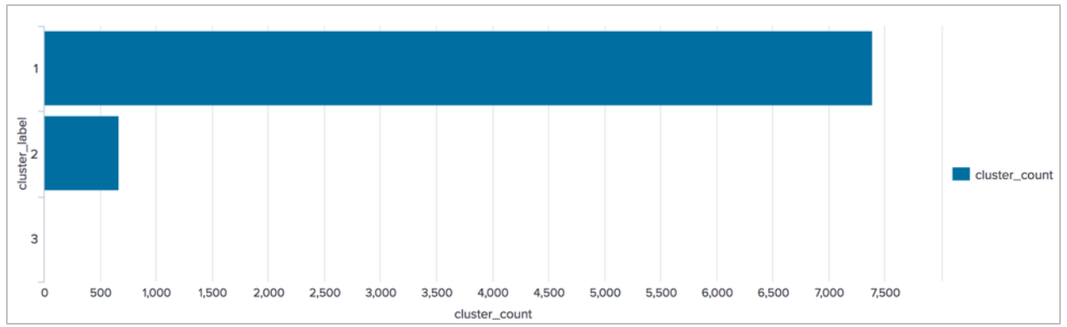
`n-gram` for Clustering

Moving window of character strings of length n (below, n=3)

Original data: `..._-_-_[//:::]_"_/.?=&=-&=--&=__."___"://.."_"/._`

`cluster` Command Example

```
sourcetype=cisco_esa
| cluster t=0.2 showcount=t
| table cluster_label cluster_count _raw
| sort -cluster_count
```



cluster_label	cluster_count	_raw
1	7338	Sat Jun 09 17:06:41 2018 Info: MID 245078 queued for delivery
2	700	Fri Jun 08 23:58:43 2018 Info: New SMTP ICID 743953 interface Management (192.168.3.120) address 206.176.229.254 reverse dns host ironport.mineralore.com verified yes
3	2	Fri Jun 08 13:42:29 2018 Warning: Received an invalid DNS Response: rcode=ServFail data="'@\\xba \\x80\\x02\\x00\\x01\\x00\\x00\\x00\\x00\\x00\\x00\\x0278\\x0269\\x03152\\x0285\\x07in addr\\x04arpa \\x00\\x00\\x0c\\x00\\x01'" to IP 193.0.0.193 looking up 78.69.152.85.in addr.arpa

`cluster` Command for Anomaly Detection

- Small group anomaly: 2 successful logins from a terminated user
  - For small groups, sort *ascending* by **`cluster_count`**
- Large group anomaly: a DDoS attack of 1000s of similar events
  - For large groups, sort *descending* by **`cluster_count`**

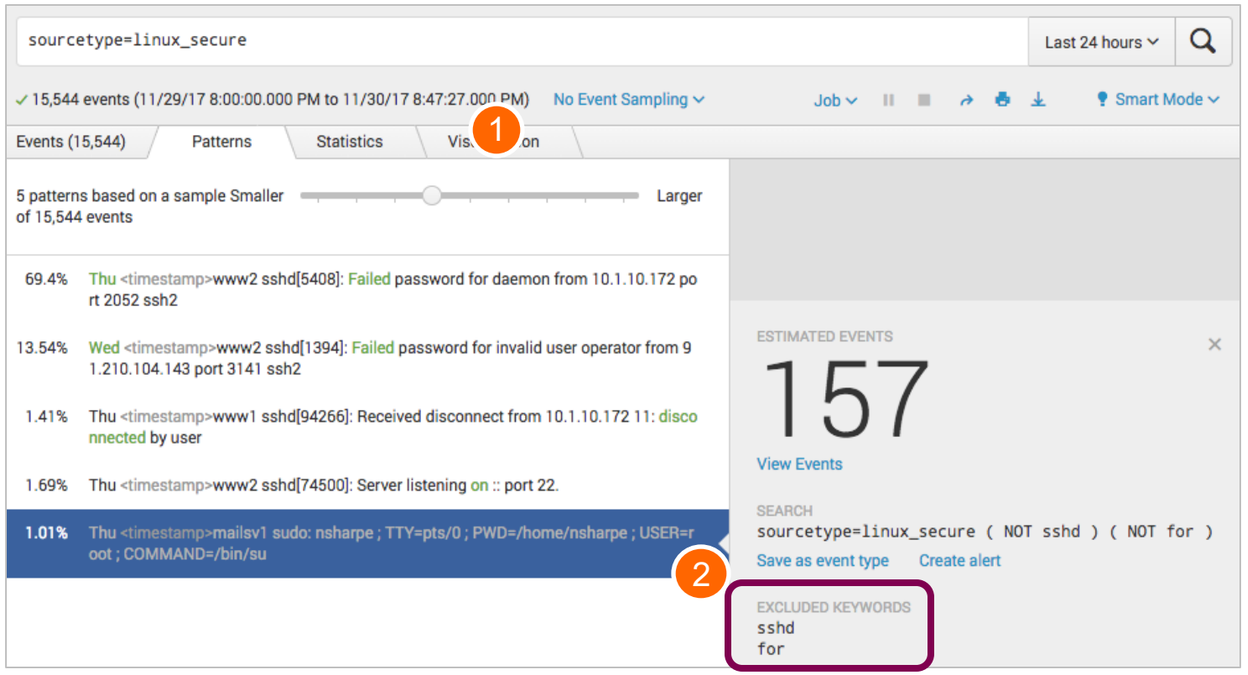
```
source="/opt/log/ecommsv1/sales_entries.log"
| cluster showcount=t t=0.7 labelonly=t
| table _time, cluster_count, cluster_label, _raw
| dedup 1 cluster_label
| sort cluster_count, cluster_label -_time
```

_time	cluster_count	cluster_label	_raw
2019-07-20 23:42:03	20	7	Sat Jul 20 2019 23:42:03 ErrorCode=224
2019-07-21 04:14:16	93	6	Sun Jul 21 2019 04:14:16 ErrorCode=205
2019-07-21 04:42:16	2820	1	Sun Jul 21 2019 04:42:16 Sent to Accounting System 101809

Event Pattern Detection

- Some patterns are rare and difficult to find in the Events tab
- Event pattern detection is the **`cluster`** command in Splunk UI
- Use the **Pattern** tab as a first step to exploring your data

1. Drag the slider to view events
   - More generically (Larger)
   - or
   - More specifically (Smaller)
2. Click the smallest cluster to view which keywords were used



findkeywords Command Example

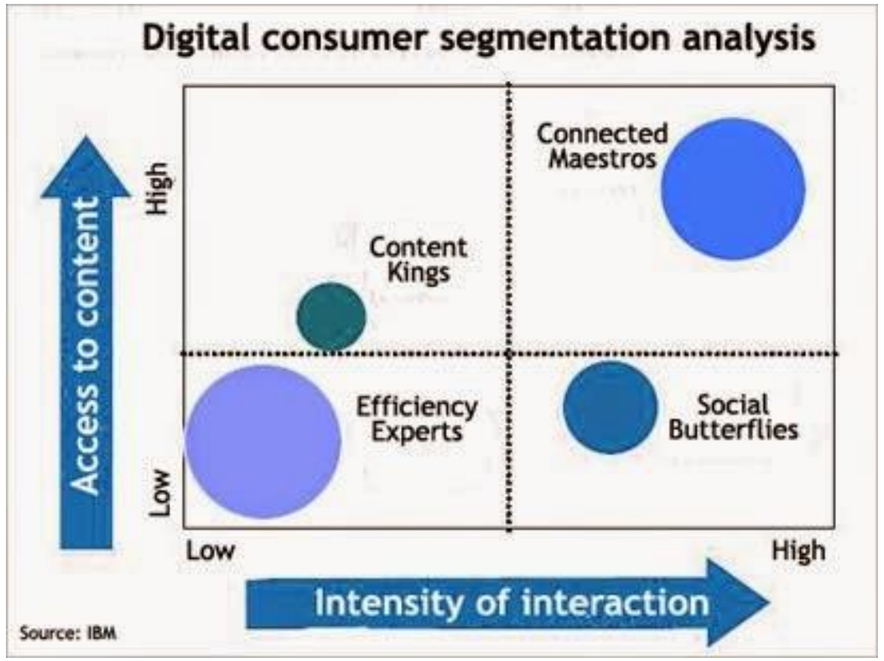
```
sourcetype=linux_secure
| cluster showcount=t t=0.5 labelonly=t
| findkeywords labelfield=cluster_label
```

confidence	eventTypeable	excludeKeywords	groupID	includeKeywords	numInInputGroup	numMatched	percentInInputGroup	percentMatched	sampleEvent	search
0.461636	1		8	invalid	21843	29883	0.436860	0.597660	Fri Sep 30 2016 21:12:39 www2 sshd[1509]: Failed password for invalid user mysql from 10.3.10.46 port 1948 ssh2	search sourcetype=linux_secure invalid
0.000000	1		1	ssh2	13804	43487	0.276080	0.869740	Sat Oct 29 2016 17:19:30 www2 sshd[5213]: Failed password for games from 67.133.102.54 port 1818 ssh2	search sourcetype=linux_secure ssh2
0.000000	1		6	invalid	7840	29883	0.156800	0.597660	Sat Oct 29 2016 16:12:12 www3 sshd[2742]: Failed password for invalid user jean-luc from 147.213.138.201 port 1703 ssh2	search sourcetype=linux_secure invalid
1.000000	1		2	uid	3483	3877	0.069660	0.077540	Sat Oct 29 2016 17:18:52 www2 sshd[33426]: pam_unix(sshd:session): session opened for user myuan by (uid=0)	search sourcetype=linux_secure uid
1.000000	1		4	on	912	912	0.018240	0.018240	Sat Oct 29 2016 17:16:25 www3 sshd[5409]: Server listening on :: port 22.	search sourcetype=linux_secure on
0.220833	1	sshd	7		888	1395	0.017760	0.027900	Fri Oct 28 2016 16:30:15 mailsv1 su: pam_unix(su:session): session closed for user root	search sourcetype=linux_secure (NOT sshd)
1.000000	1		3	disconnected	768	768	0.015360	0.015360	Sat Oct 29 2016 17:17:19 www3 sshd[60933]: Received disconnect from 10.3.10.46 11: disconnected by user	search sourcetype=linux_secure disconnected
1.000000	1	sshd for	5		462	462	0.009240	0.009240	Sat Oct 29 2016 16:58:51 www1 sudo: nsharpe ; TTY=pts/0 ; PWD=/home/nsharpe ; USER=root ; COMMAND=/bin/su	search sourcetype=linux_secure (NOT sshd) (NOT for)

Cluster Labeling Process

- Leverage existing labels in the data before using ML
- Human input is key to the success of this process

1. Find clusters
   - Use clustering commands, ML models
2. Validate (is there coherence?)
   - Assign labels and make meaningful names
3. Target strategies to customer preferences
   - Geographic
   - Performance
   - Behavioral



Topic 3 Lab

- Description: Create different clusters using numerical and string fields
- Duration: 25 minutes
- Tasks:
  - Examine distortion across different cluster sizes
  - Visualize cluster centroids in a 3d scatter plot visual
  - Cluster router data based on string similarity
  - Use the patterns tab and `findkeywords` command to examine important words within a cluster

Topic 4: Correlations and Transactions

Topic Objectives

- Define correlation and co-occurrence
- Use SPL correlation commands
- Use the statistical tests from the Machine Learning Toolkit to correlate fields
- Use streamstats and chart commands to correlate data

Correlation vs. Co-occurrence

- Correlation is a statistical measurement
- Indicates how changes in one variable can influence other variables
- Correlation does not imply causation
- Typically used to analyze numerical data

- Co-occurrence is a measurement of frequency or probability
- In natural language processing, refers to the frequency words or phrases appear together in a dataset
- In data analysis, used to identify patterns between variables

`associate` Command

```
associate [<options>]
```

- Can knowing the value of a field predict the value of another?
  - Calculates change in entropy based on field values
  - Entropy is a measurement of disorder, randomness, or uncertainty
- Fields outputted to a table:
  - Analyzed fields
    - `Reference_Key, Reference_Value, and Target_Key`
  - Calculated fields
    - `Unconditional_Entropy, Conditional_Entropy, and Entropy_Improvement`
  - Summary fields
    - `Description, Support`

https://www.statisticshowto.com/shannon-entropy

`associate` Command Options

If you specify fields, only specified fields are used; others are not

- **`supcnt`** minimum number of times that the "reference key=reference value" combination must appear
  - Default is 100
- **`supfreq`** minimum frequency of "reference key=reference value" combination as a fraction of the number of total events
  - Default is 0.1
- **`improv`** minimum entropy improvement for the "target key" outputted field
  - Default is 0.5

`associate` Command Example

```
| inputlookup phishing.csv
| associate
```

Reference_Key	Reference_Value	Target_Key	Support	Unconditional_Entropy	Conditional_Entropy	Entropy_Improvement	Top_Conditional_Value	Description
Favicon	-1	popUpWidnow	18.57%	0.708	0.193	0.515388	-1 (19.33% -> 97.03%)	When 'Favicon' has the value '-1', the entropy of 'popUpWidnow' decreases from 0.708 to 0.193.
Favicon	1	popUpWidnow	81.43%	0.708	0.119	0.589348	1 (80.67% -> 98.39%)	When 'Favicon' has the value '1', the entropy of 'popUpWidnow' decreases from 0.708 to 0.119.
Favicon	1	port	81.43%	0.573	0.040	0.533061	1 (86.41% -> 99.57%)	When 'Favicon' has the value '1', the entropy of 'port' decreases from 0.573 to 0.040.
Prefix_Suffix	1	Result	13.25%	0.991	0.000	0.990624	1 (55.69% -> 100.00%)	When 'Prefix_Suffix' has the value '1', the entropy of 'Result' decreases from 0.991 to 0.000.
Prefix_Suffix	1	SSLfinal_State	13.25%	1.329	0.365	0.964834	1 (57.27% -> 93.04%)	When 'Prefix_Suffix' has the value '1', the entropy of 'SSLfinal_State' decreases from 1.329 to 0.365.
Result	-1	Prefix_Suffix	44.31%	0.564	0.000	0.564307	-1 (86.75% -> 100.00%)	When 'Result' has the value '-1', the entropy of 'Prefix_Suffix' decreases from 0.564 to 0.000.
Result	1	SSLfinal_State	55.69%	1.329	0.442	0.887088	1 (57.27% -> 91.44%)	When 'Result' has the value '1', the entropy of 'SSLfinal_State' decreases from 1.329 to 0.442.
Result	1	URL_of_Anchor	55.69%	1.508	1.000	0.508669	0 (48.28% -> 62.29%)	When 'Result' has the value '1', the entropy of 'URL_of_Anchor' decreases from 1.508 to 1.000.

`analyzefields` Command

```
analyzefields classfield=<field>
```

Analyze all **numerical** fields to see how well each might predict the value of a field you choose (the target classified)

Finds the fields that best predict the value of a field you specify

- Assumes normal distribution
- `classfield` signifies the field whose value you want to predict
- The field should be a categorical field
  - Does not need to be binary, but binary problems are easier to solve

`analyzefields` Command Results

Returns a table with 5 columns

- **`Field`** a field from your search results that may predict your `classfield` field
- **`Count`** number of times that `field` occurs in your search results
- **`Cocur`** ratio of field and `classfield` both occurring (1=all events)
- **`Acc`** how accurately `field` value predicts `classfield` value (accurate predictions divided by total number of events)
- **`Balacc`** non weighted average of accuracies (mean of true positive and true negative rates)

analyzefields Command Example

```
sourcetype=cisco_wsa_squid
| eval violation = if(usage="violation", 1, 0)
| analyzefields classfield=violation
```

field	count	cocur	acc	balacc
Department				
Email				
First_Name				
Last_Name				
Username				
bytes_in	36727	1.000000	0.998720	0.998720
change_type				
date_hour	36727	1.000000	1.000000	1.000000
date_mday	36727	1.000000	1.000000	1.000000
date_minute	36727	1.000000	1.000000	1.000000
date_second	36727	1.000000	1.000000	1.000000
date_year	36727	1.000000	1.000000	1.000000
date_zone	36727	1.000000	1.000000	1.000000

`correlate` Command

View co-occurrence between fields (not values) in a matrix format

- Values indicate the percentage of times the two fields occur in the same events
- Processes data only for the first 1,000 fields
  - Controlled by **`Maxfields`** – set in **limits.conf**
  - Increasing default may have memory / CPU costs

```
sourcetype=sendmail_syslog | correlate
```

RowField	class	ctladdr	daemon	date_hour	date_mday	date_minute	date_month	date_second	date_wday	date_year	date_zone	delay	dsn
class	1.00	0.00	0.99	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.00	0.00
ctladdr	0.00	1.00	0.00	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.86	0.86
daemon	0.99	0.00	1.00	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.00	0.00
date_hour	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_mday	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_minute	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_month	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_second	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_wday	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_year	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
date_zone	0.29	0.27	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.31	0.31
delay	0.00	0.86	0.00	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	1.00	1.00

`score` command (Splunk MLTK)

```
score <method> <field1> against <field2>
```

- The score command runs a variety of of statistical functions and tests
  - `pearsonr`
  - `spearmanr`
- Can describe the relationship between two fields
- All methods follow the same preprocessing steps
  1. All rows containing NAN values are removed
  2. An error message displays if any categorical fields are used

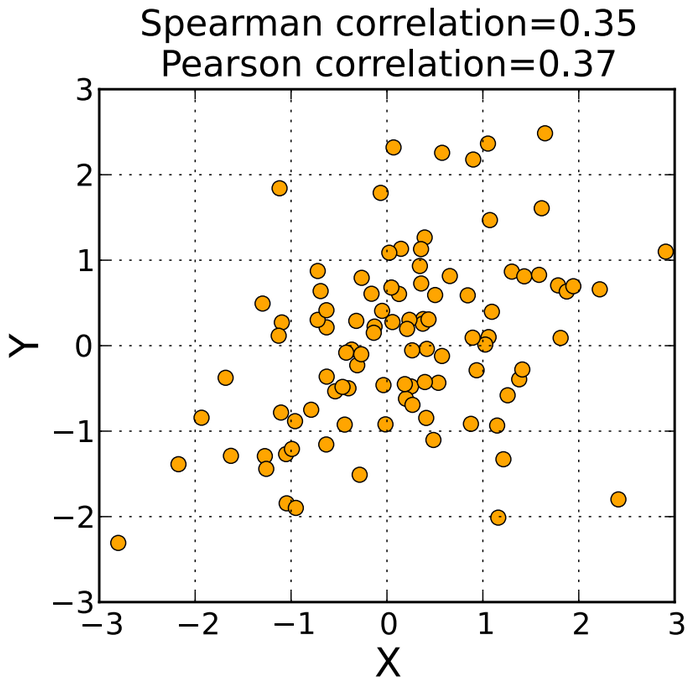
`pearsonr` method

- Statistical measure of the linear relationship between two continuous variables
- Coefficient ranges from -1 to 1, indicating a negative to positive linear relationship
- Assumes that the relationship between variables is linear, variables are normally distributed, and there are no outliers
- With a linear relationship, the change in one variable is proportional to the change in the other variable

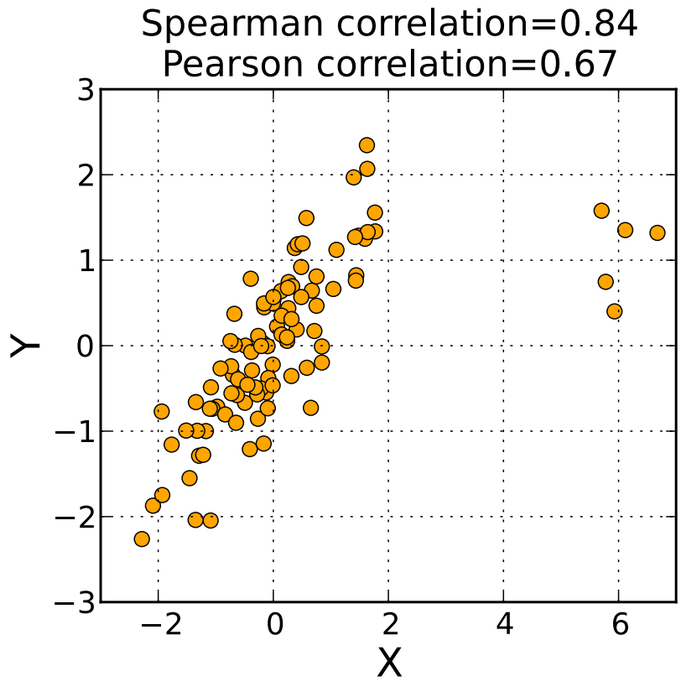
`spearmanr` method

- Statistical measurement of the strength and direction of monotonic relationships
- Values are ordered from lowest to highest
- Two variables have a monotonic relationship when changes in one variable cause the other to consistently increase or decrease
- These changes do not need to be linear
- The direction of change is consistent, even if the rate of change varies across different parts of the relationship

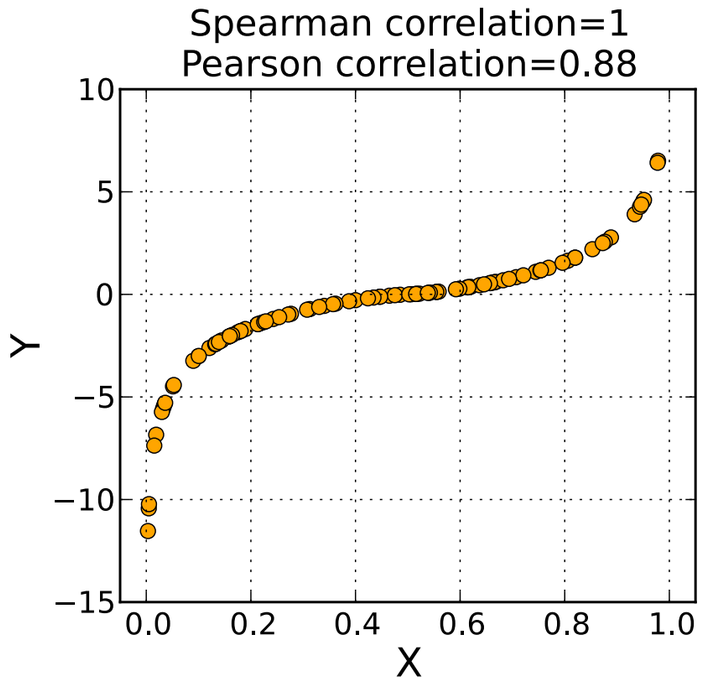
`score` Command Examples



Data is evenly distributed and there are no prominent outliers



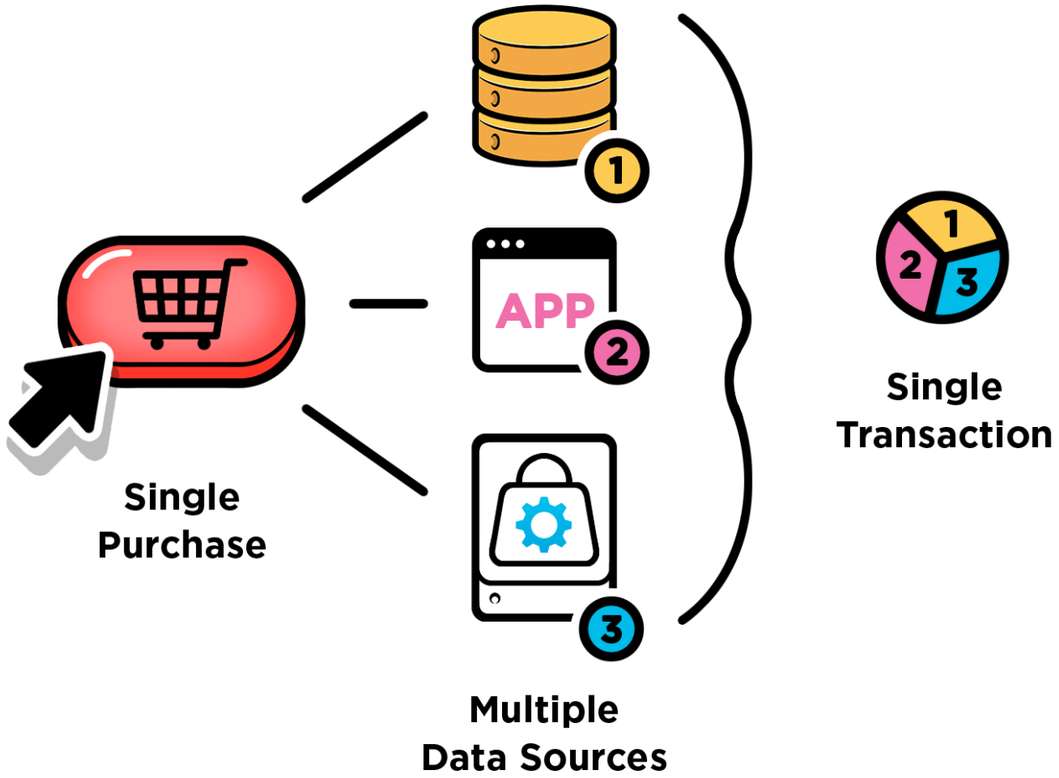
Spearman is less sensitive to strong outliers



For every data point, as X increases so does Y. Even if the relationship is not linear.

What are Transactions

- A group of related events from one or many different sources
- These events may share the same value for a field or be generated by the source
- Can be used to tell a story from beginning to end
- Give greater insight to the order actions or operations occurred



Transactional Analysis Use Cases

- Transactional Analysis: the study of any group of conceptually-related events which spans time and describes behavior
- Transactions are “high-level” events:
  - DDoS attack from unknown IP ranges
  - John Smith purchased a product on the website
- Transactions span IT & business data sources:
  - “Purchase transaction” involves web-server logs, e-commerce data, product and customer lookups
- Business users care about transactions
  - Basic unit of economic activity

`transaction` Command Review

```
transaction (<field>|<field-list>) [options]
```

- Groups events that share one or more fields
  - Group events on values from `<field>`
  - Group events based on shared values from `<field-list>`
- Determine event grouping behavior using `options`:
  - Ranges of time: `maxspan`, `maxpause`
  - Maximum number of events in a transaction: `maxevents`
  - Text contained in the first/last events: `startswith`, `endswith`
- Adds `duration`, `eventcount`, and `closed_txn` fields to events

Transaction Considerations

- The **`transaction`** command is resource intensive
- Only use **`transaction`** if you:
  - Need to group events on values from multiple fields
  - Need to define event grouping on start/end values or segment on time
  - Want to keep raw data associated with each event
- Otherwise use **`stats`** command
  - Faster and more efficient
  - Can perform calculations
  - Can group events based on a single field value (e.g. **`by src_ip`**)

Concept: Meta Transactions

- Transactions-of-transactions that represent the entirety of each customer's experiences
  - Do a transaction on transaction events by customer id
  - Build a meta transaction of all of a customer’s experiences
  - `ServiceTicketHistory -` full customer experience
- With a *collection* of these, you can do statistics
  - Find all customers who had a positive experience with your product

Event	Score
Search for site	9,0
Search products	4.6
Compare products	3.2
Add to wish list	7.4
Place in cart	9.1
Purchase	8.8
Receive shipment	9.3
Review product	7.9
Recommend product	9.6
File a warranty claim	2.1
Receive replacement	4.8

How to Construct Meta Transactions

- Goal: attach KPI of interest to high-level customer experiences
  - Construct transactions
  - Assemble transactions into higher-level transactions
  - They may be ongoing / may not have finished yet
- Avoid running `transaction` on top of `transaction`
  - Can be prohibitively slow
- Use alternatives (like `stats list` or `stats values`)

Modeling order of operations

1. In this example we wish to model the order of actions taken by unique Visitors to a web site
2. We only need the _time, JSESSIONID, and action fields
3. Use fields and table to remove all the unneeded information
4. Use reverse to put the events in chronological order

```
index=web action=*
| fields action JSESSIONID _time
| table _time JSESSIONID action
| reverse
```

_time	JSESSIONID	action
2024-04-24 17:37:01	SD2SL10FF4ADFF4960	view
2024-04-24 17:37:07	SD2SL10FF4ADFF4960	addtocart
2024-04-24 17:37:48	SD2SL10FF4ADFF4960	addtocart
2024-04-24 17:37:50	SD2SL10FF4ADFF4960	purchase
2024-04-24 17:40:31	SD7SL9FF9ADFF4964	addtocart
2024-04-24 17:40:42	SD7SL9FF9ADFF4964	view
2024-04-24 17:41:02	SD7SL9FF9ADFF4964	addtocart
2024-04-24 17:41:15	SD7SL9FF9ADFF4964	addtocart

Modeling order of operations

5. Use streamstats to create an order field to denote which event came 1st, 2nd, 3rd, etc…
6. Use chart to transform the table so that each row is a user and the column indicate the order

```
...
| streamstats count as order
by JSESSIONID
| chart values(action) over
JSESSIONID by order limit=0
```

JSESSIONID	1	10	11	12	13	14	15	16	17	18	2	3	4
SD0SL10FF2ADFF4960	purchase										remove	view	changequantity
SD0SL1FF1ADFF4962	addtocart	view									addtocart	purchase	purchase
SD0SL1FF3ADFF4964	view										changequantity		
SD0SL2FF4ADFF4966	view										addtocart		
SD0SL2FF5ADFF4958	addtocart	purchase									remove	view	addtocart
SD0SL2FF6ADFF4952	view										remove	addtocart	addtocart
SD0SL2FF6ADFF4966	changequantity										addtocart	purchase	purchase

Modeling order of operations

7. Use fillnull to fill in the blanks
8. Use stats to count number of unique action chains up to the fifth action
9. We can go beyond the fifth action by adding more orders to the stats by clause

```
...
| fillnull value=NULL
| stats count by 1, 2, 3, 4, 5
```

1	2	3	4	5	count
addtocart	NULL	NULL	NULL	NULL	16
addtocart	addtocart	addtocart	NULL	NULL	1
addtocart	addtocart	addtocart	addtocart	addtocart	1
addtocart	addtocart	addtocart	purchase	purchase	1
addtocart	addtocart	addtocart	view	view	1
addtocart	addtocart	purchase	purchase	NULL	1
addtocart	addtocart	purchase	purchase	addtocart	4
addtocart	addtocart	purchase	purchase	remove	1
addtocart	addtocart	purchase	purchase	view	3

Explore the performance difference

```
index=web action=*
| fields action JSESSIONID
| transaction mvlist=true JSESSIONID keepevicted=true
| table action JSESSIONID
| eval JSESSIONID=mvdedup(JSESSIONID)
| mvexpand action
| streamstats count as order by JSESSIONID
| xyseries order JSESSIONID action
| fillnull value=NULL
| sort order
| transpose 0 header_field=order
| stats count by 1, 2, 3, 4, 5
```

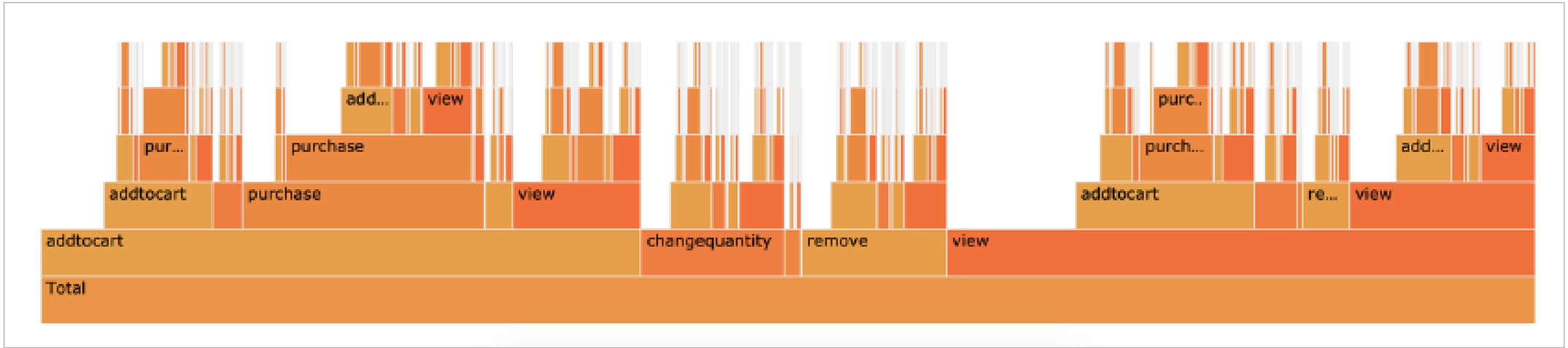
scanning **156,666** events in **2.593** seconds

```
index=web action=*
| fields action JSESSIONID _time
| table _time JSESSIONID action
| reverse
| streamstats count as order by JSESSIONID
| chart values(action) over JSESSIONID by order limit=0
| fillnull value=NULL
| stats count by 1, 2, 3, 4, 5
```

scanning **156,666** events in **1.663** seconds

Aggregate & Visualize

- Higher levels built from lower levels
  https://splunkbase.splunk.com/app/3468/
  - ***Economic transactions:*** understood from low-level machine events
  - ***Customer experiences:*** from mid-level economic transactions
- High levels affect the lower levels, too
  - If customers are happy, they make more raw events (web clicks, etc.)



Topic 4 Lab

- Description: Examine field correlations and create a high-level transaction
- Duration: 20 minutes
- Tasks:
  - Identify changes in entropy, field associations, and linear relationships
  - Create a high-level transaction showing badge access history

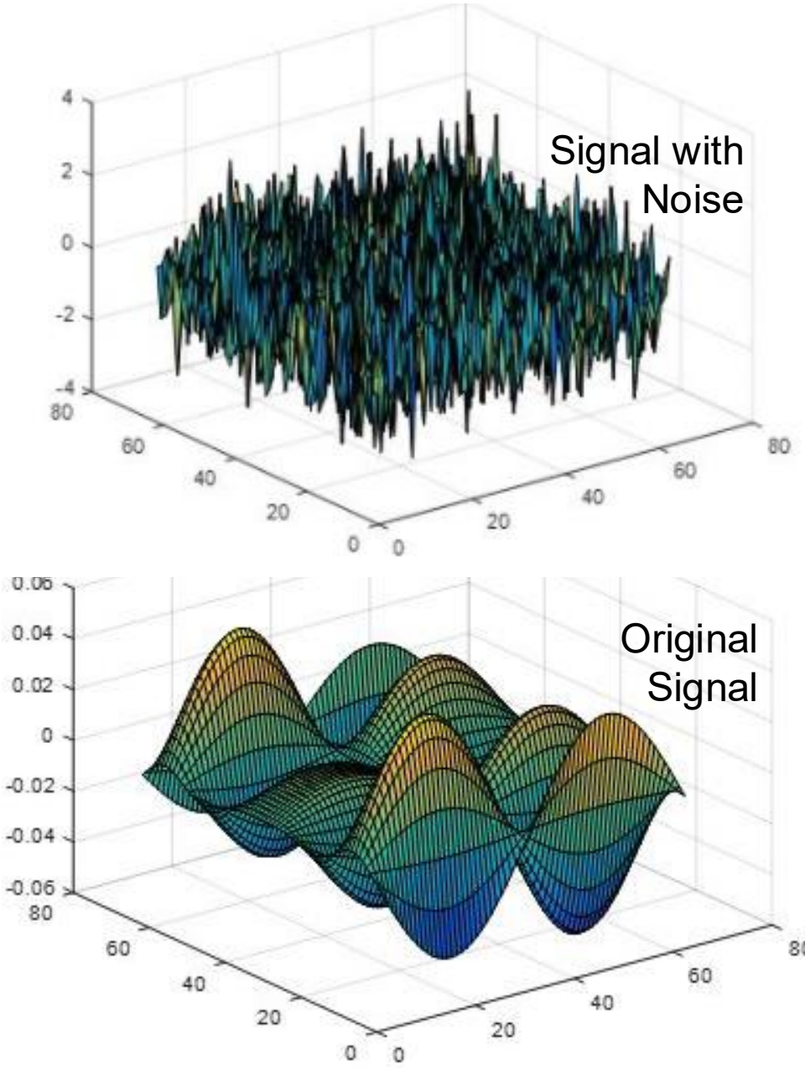
Topic 5: Anomaly Detection

Topic Objectives

- Define Statistical Outliers
- Use Add-hoc methods of numerical anomaly detection
- Find numerical or categorical anomalies with the `AnomalyDetection` command

Anomaly Detection with ML Overview

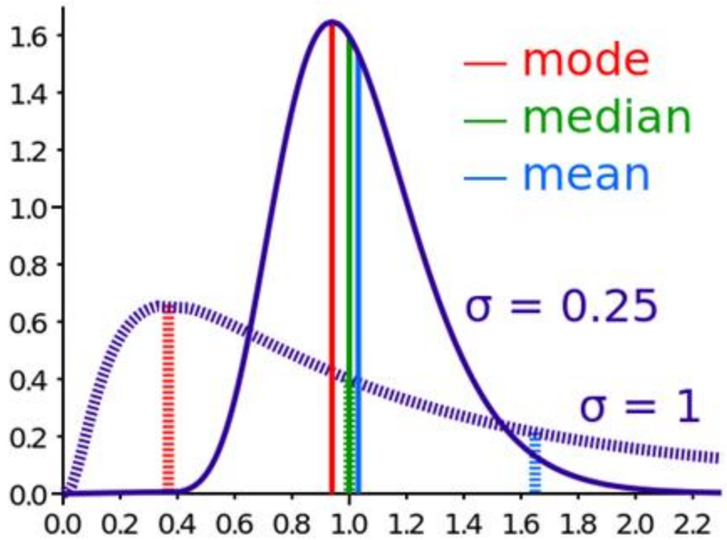
- Deviation from expected behavior of the data in an event or sequence
  - Keep outliers for reports and alerts
  - Remove outliers if they skew analysis
- ML helps model expected behaviors and identify large deviations
  - Assign anomaly scores to events and patterns
  - Estimate probabilities of rare events
- Create a baseline data image (current state of the system)



https://upload.wikimedia.org/wikipedia/en/5/5d/Sample_BEMD_Simulation_results_for_a_noisy_signal.jpg

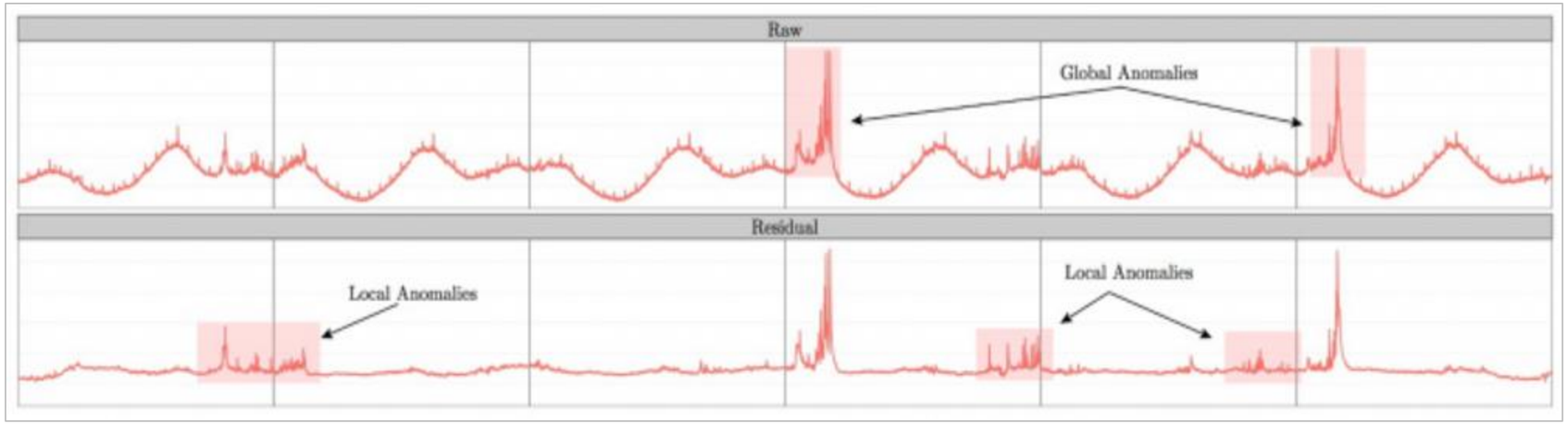
Statistical Outliers for Anomaly Detection

- Finding statistical outliers is a first step to effective anomaly detection
- A statistical outlier is any event which is far from some measure of centrality
  - Some measures of centrality
    - Mean: the average value
    - Median: the typical value
    - Mode: the most frequent value
  - Statistical outliers come in many different flavors
    - Non-average: far from the mean
    - Non-typical: far from the median
    - Non-popular: small count relative to the mode



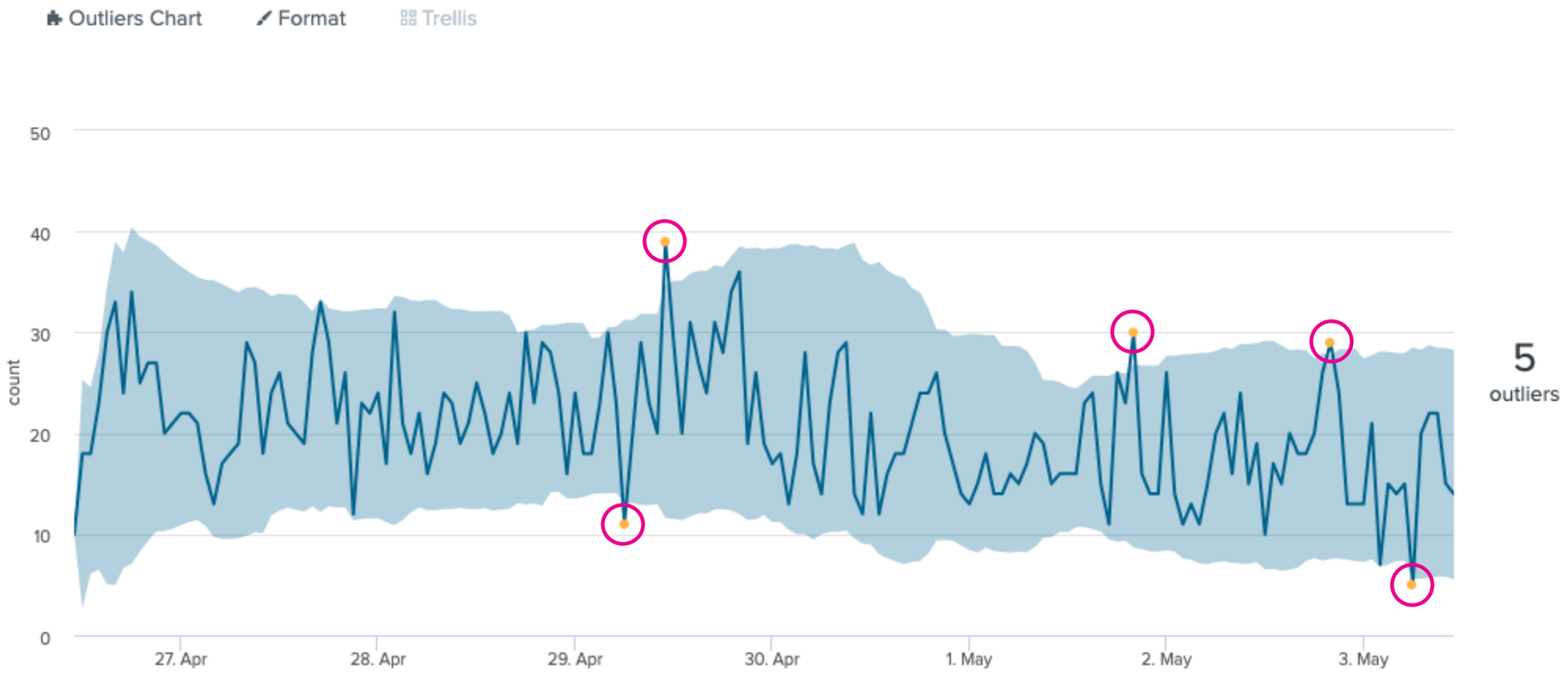
Types of Anomaly Detection

- Local anomalies: **`streamstats`** with standard deviation and mean absolute deviation
- Global anomalies: **`eventstats`** with interquartile range



Outliers Chart Visualization from MLTK

```
... | table _time, count, lower_bound, upper_bound
```

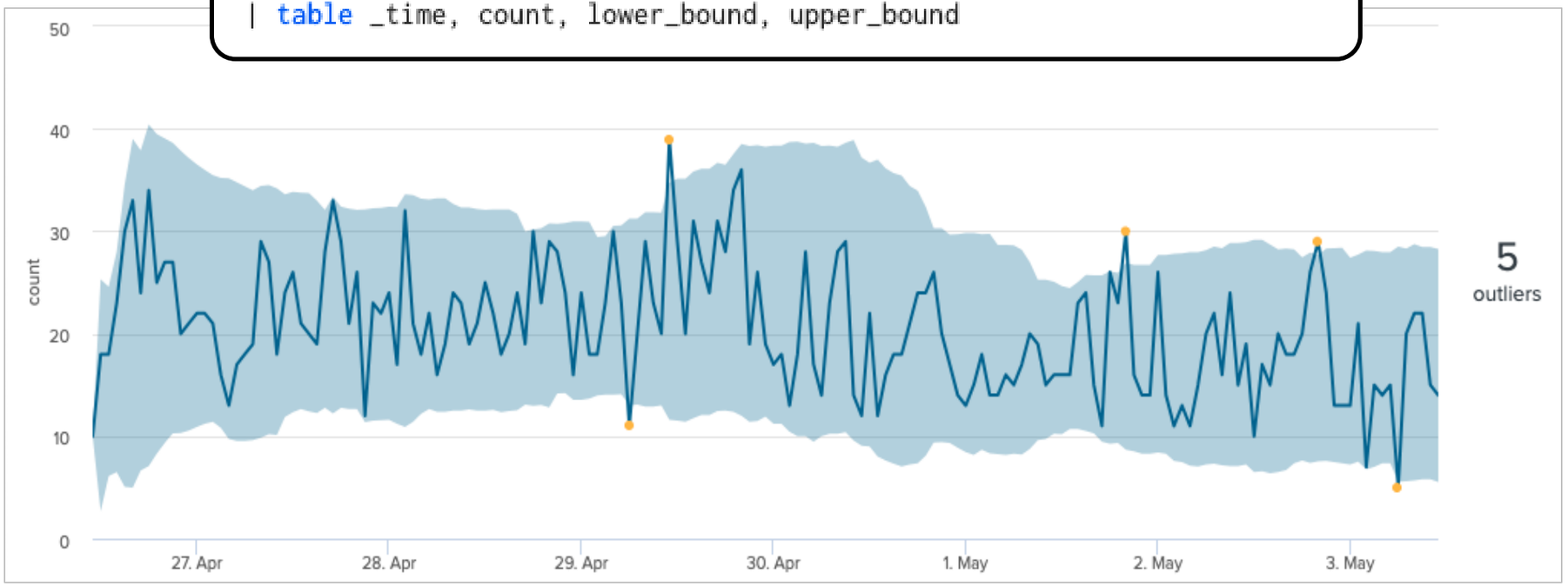


Methods for Numerical Outlier Detection

- Standard Deviation
  - Best used when data is normally distributed
- Median Absolute Deviation
  - Best used when data is not normal or standard deviations are not effective
- Interquartile Range
  - Best used when focusing on the middle 50% of data
- All three methods can be used to find local or global anomalies

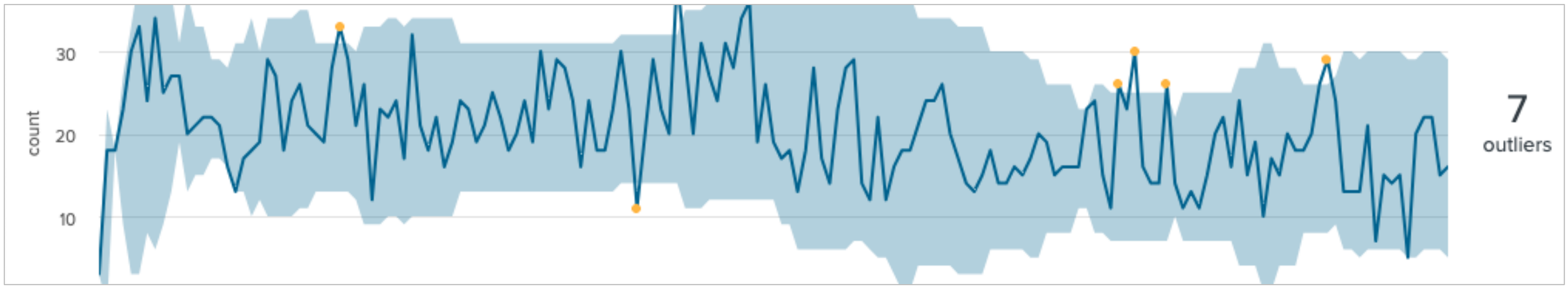
Standard Deviation Method

```
sourcetype = access_combined status>399 earliest=-1w
| timechart span=1h count
| streamstats window=24 avg(count) as avg, stdev(count) as stdev
| eval multiplier = 2
| eval lower_bound = avg - (stdev * multiplier)
| eval upper_bound = avg + (stdev * multiplier)
| eval outlier = if(count < lower_bound OR count > upper_bound, 1, 0)
| table _time, count, lower_bound, upper_bound
```



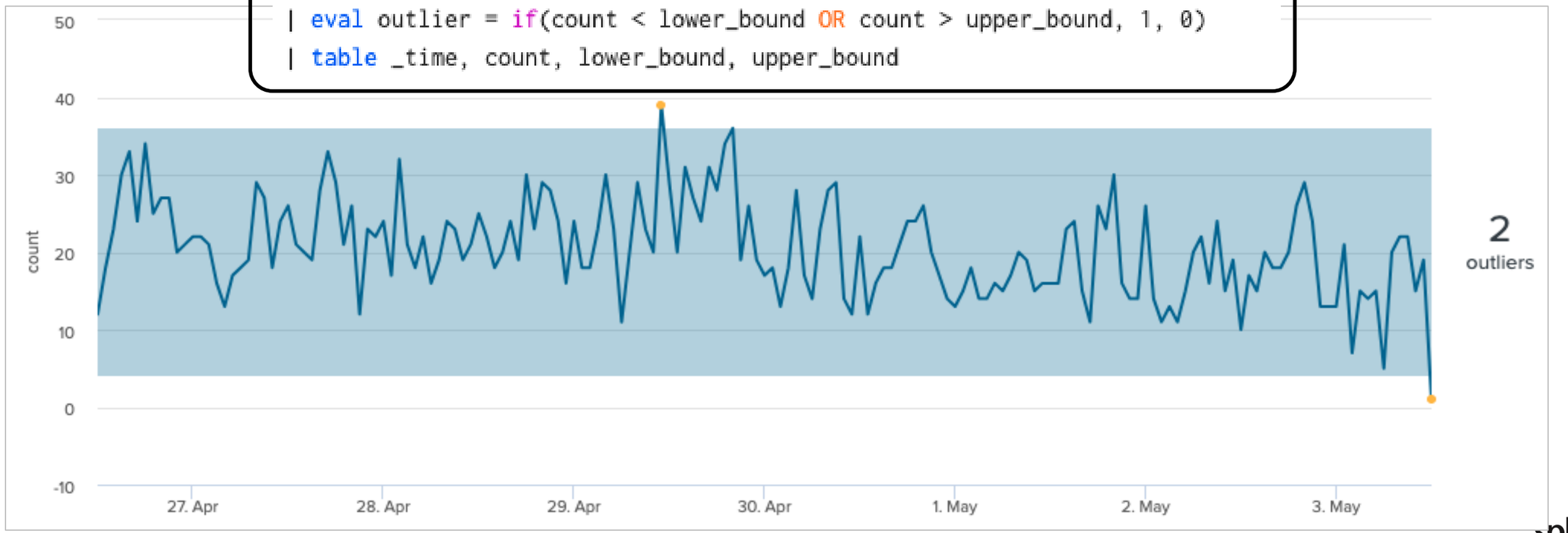
Median Absolute Deviation

```
sourcetype=access_combined status>399 earliest=-1w
| timechart span=1h count
| streamstats window=24 median(count) as median
| eval abs_dev = abs(count - median)
| streamstats window=24 median(abs_dev) as median_abs_dev
| eval lower_bound = median - (median_abs_dev * 3)
| eval upper_bound = median + (median_abs_dev * 3)
| eval outlier = if(count < lower_bound OR count > upper_bound, 1, 0)
| table _time, count, lower_bound, upper_bound
```



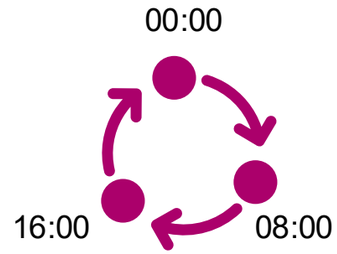
Interquartile Range Method (IQR)

```
sourcetype=access_combined status>399 earliest=-1w
| timechart span=1h count
| eventstats median(count) as median, p25(count) as p25, p75(count) as p75
| eval IQR = p75 - p25
| eval multiplier = 2
| eval lower_bound = median - (IQR * multiplier)
| eval upper_bound = median + (IQR * multiplier)
| eval outlier = if(count < lower_bound OR count > upper_bound, 1, 0)
| table _time, count, lower_bound, upper_bound
```



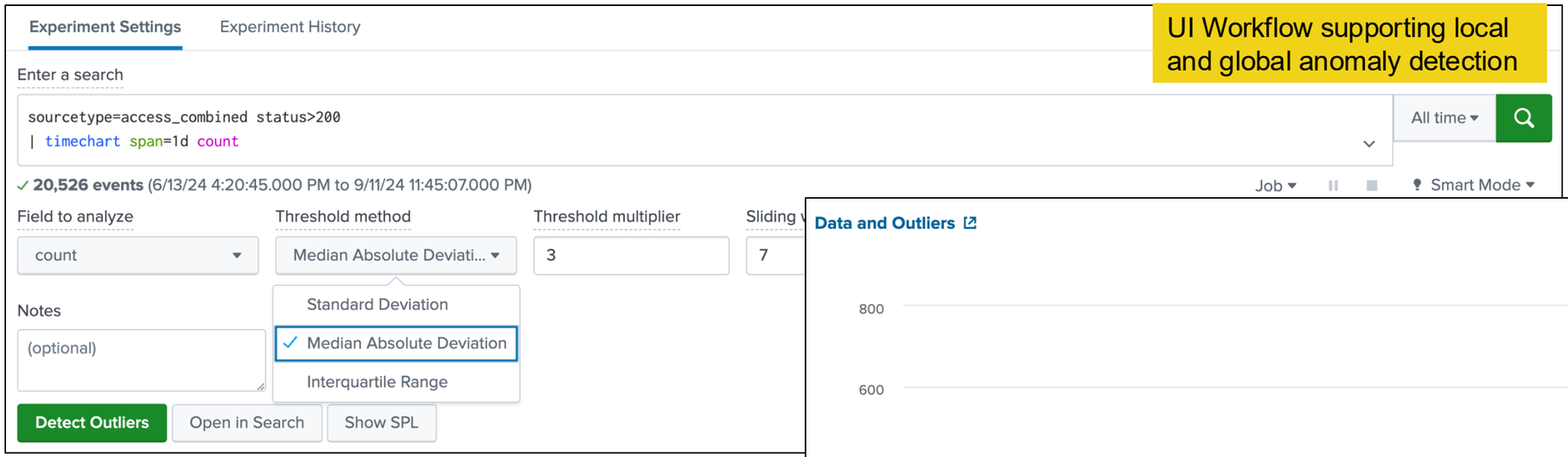
Windows and Multipliers

- The **`window`** used with **`streamstats`** refers to the period of the timeseries
- The period indicates the repeating cycle of the series
  - **`timechart`** with the **`span=1h`** would have a period of 24

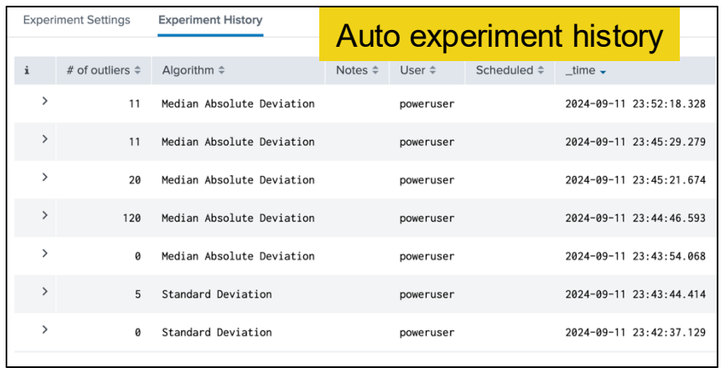


- The multiplier will move the upper and lower bounds further or closer to the original data
  - In may programing tools Standard Deviation uses a multiplier of **2** by default
  - Median Absolute Deviation uses a multiplier of **3** by default
  - Interquartile Range use a multiplier of **1.54** by default
- These are typical multipliers to start with and then adjusted as needed

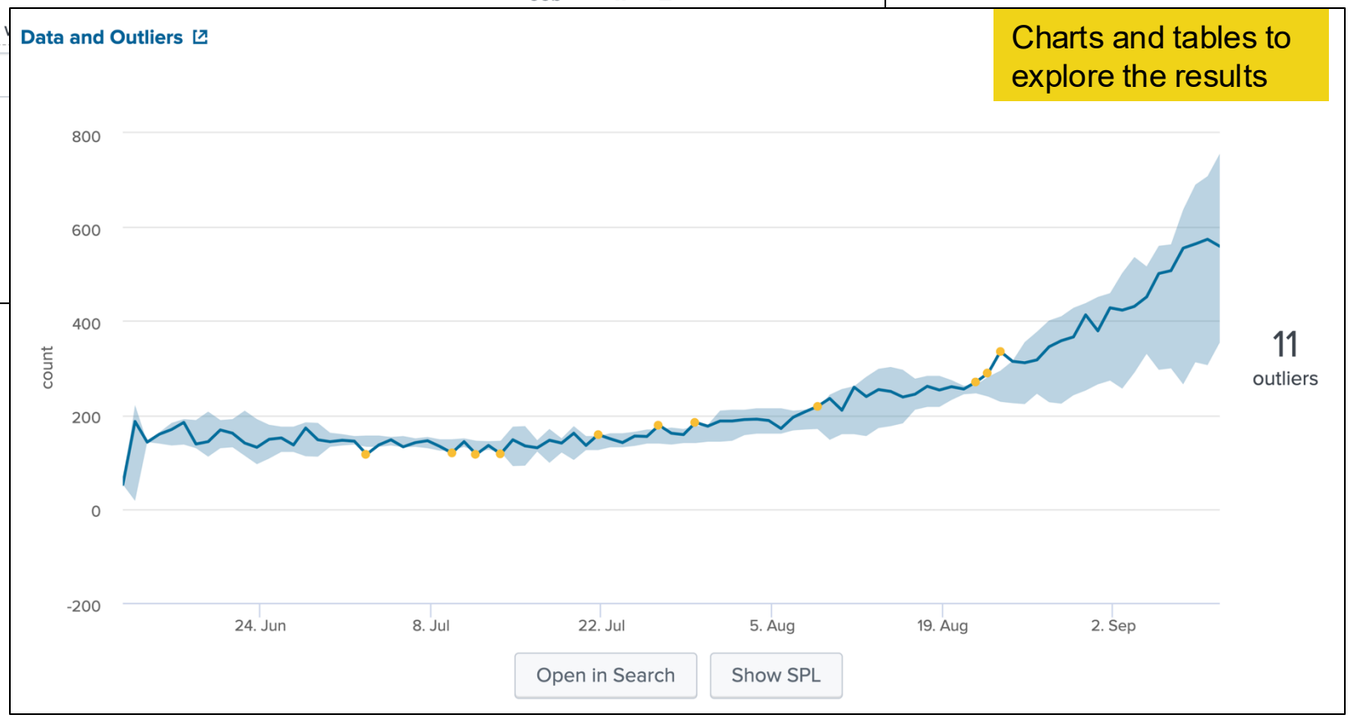
Detect Numeric Outliers Experiments (MLTK)



UI Workflow supporting local and global anomaly detection



Auto experiment history



Charts and tables to explore the results

`anomalydetection` Command

```
anomalydetection [<method-option>] [<action-option>]
 [<pthresh-option>] [<cutoff-option>] [<field-list>]
```

- `method=histogram` and `method=zscore` support numerical and categorical anomaly detection
- `method=IQR` only supports numerical anomaly detection
- `pthresh=<num>`
  - Only used with `method=zscore`
  - Adjusts the boundaries for anomaly detection
- `cutoff=true` (default)
  - Only used with `method=histogram`
  - If set to `true`, will limit the number of anomalies found to global

anomalydetection method=zscore

Examines each provided field independently. The same values will be flagged as anomalies regardless of what fields are provided to examine.

If numeric

$$\texttt{Anomaly\_Score\_Num(x)} = p(x) \sim \mathcal{N}(\mu, \sigma)$$

p(x) is the two-tailed probability of seeing such a value in a normal distribution with mean mu and standard deviation sigma

If categorical or < 100 distinct values

1.If frequency(x) < pThresh, it counts as anomalous

$$frequency(x) = \frac{count(x)}{dc(x) \times count}$$

2.Find the average frequency for non-anomalous values

$$avg_{freq} = \frac{count(x \neq anomalous)}{dc(x \neq anomalous) \times count}$$

3.Calculate the anomaly score:

$$\texttt{Anomaly\_Score\_Cat(x)} = \frac{frequency(x)}{avg_{freq}(x)}$$

anomalydetection method=histogram

1. Computes a probability for each event, default
   - Product of the frequencies of each individual field value in the event
2. Detects unusually small probabilities
   - Categorical fields
     - Frequency of X is the number of times X occurs divided by the total number of events
   - Numerical fields
     - Builds a histogram for all the values
     - Computes the frequency of a value X as the size of the bin that contains X divided by the number of events
3. Uses all provided fields in conjunction with one another to find anomalies
   - Very different results can be had when different fields are provided

anomalydetection method=IQR

Examines each provided field independently. The same values will be flagged as anomalies regardless of what fields are provided to examine.

- `action=remove` removes event containing outlying numerical value
  - abbreviation `rm`
- `action=transform`: truncates outlying value to the outlier threshold
  - default
  - abbreviation `tf`
  - If `mark=true`, the transform action prefixes the value with "000"

```
| inputlookup supermarket.csv
| anomalydetection method=iqr action=transform mark=true
```

price	product_id	quantity
6.356	p338	00012
0007.677	p343	1
0007.677	p348	1
0.991	p355	00012
0.909	p356	00012
1.032	p359	4
1.048	p363	00012

anomalydetection Command Example

```
| inputlookup supermarket.csv
| anomalydetection action=filter
```

Last 24 hours

✓ 9 results (4/25/24 9:00:00.000 PM to 4/26/24 9:29:08.000 PM) No Event Sampling Job Smart Mode

Events Patterns **Statistics (9)** Visualization

20 Per Page Format Preview

customer_id	distance	log_event_prob	max_freq	price	probable_cause	probable_cause_freq	product_id	quantity	shop_id
u62	2681.15450705	-36.1423	0.35050	0.172	price	0.00000	p325	9	s2
u72	248.699619449	-36.6300	0.35050	0.172	price	0.00000	p325	9	s4
u137	961.408935339	-35.6178	0.35050	0.172	price	0.00000	p325	5	s3
u137	961.408935339	-35.4043	0.61832	16.92	quantity	0.00004	p4029	106	s3
u166	2695.99047282	-35.1999	0.35050	0.172	price	0.00000	p325	1	s4
u176	1708.6583926	-35.5145	0.35050	0.172	price	0.00000	p325	7	s1
u196	4803.59241518	-36.0327	0.35050	51.306	price	0.00001	p439	1	s1
u214	1529.69862056	-35.9738	0.35050	0.172	price	0.00000	p325	6	s3
u231	583.590151221	-36.2727	0.35050	0.172	price	0.00000	p325	18	s2

Topic 5 Lab

- Description: Find anomalies in numerical and categorical fields
- Duration: 35 minutes
- Tasks:
  - Use different methods of anomaly detection to find outliers in a numerical field
  - Compare global to local anomalies
  - Find categorical outliers

Topic 6: Forecasting

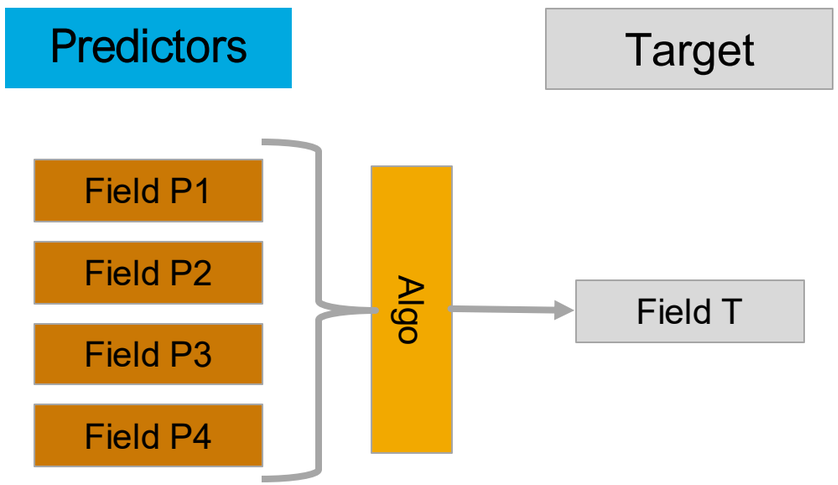
Topic Objectives

- Define forecasting use cases
- Use the **`predict`** command to forecast future timeseries
- Determine boundaries for future outliers in a time series

Predict or Forecast?

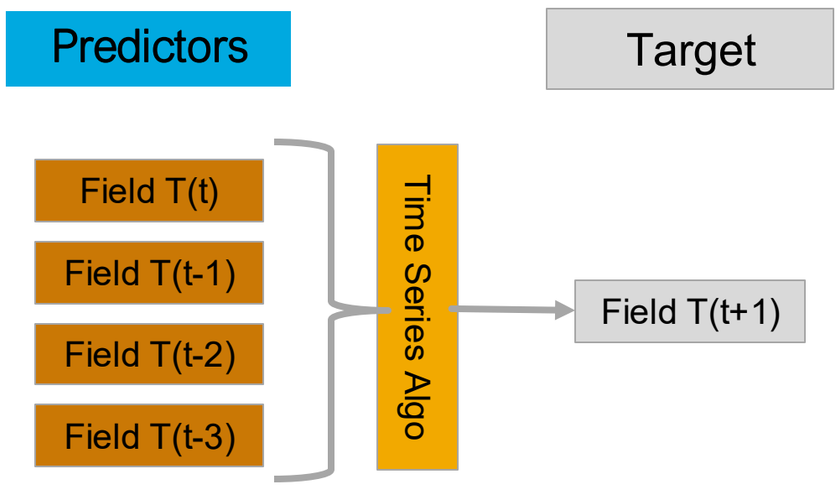
Forecast: projection of a future value or trend based on historic data

Predict (Regression)



The Cisco stock is influenced by interest rates, global economic conditions, etc.
(an estimate **based on the values of other independent fields**)

Forecast



The Cisco stock is cyclical, and every July stock price in the future will be similar to the July stock in the past (plus or minus trending)

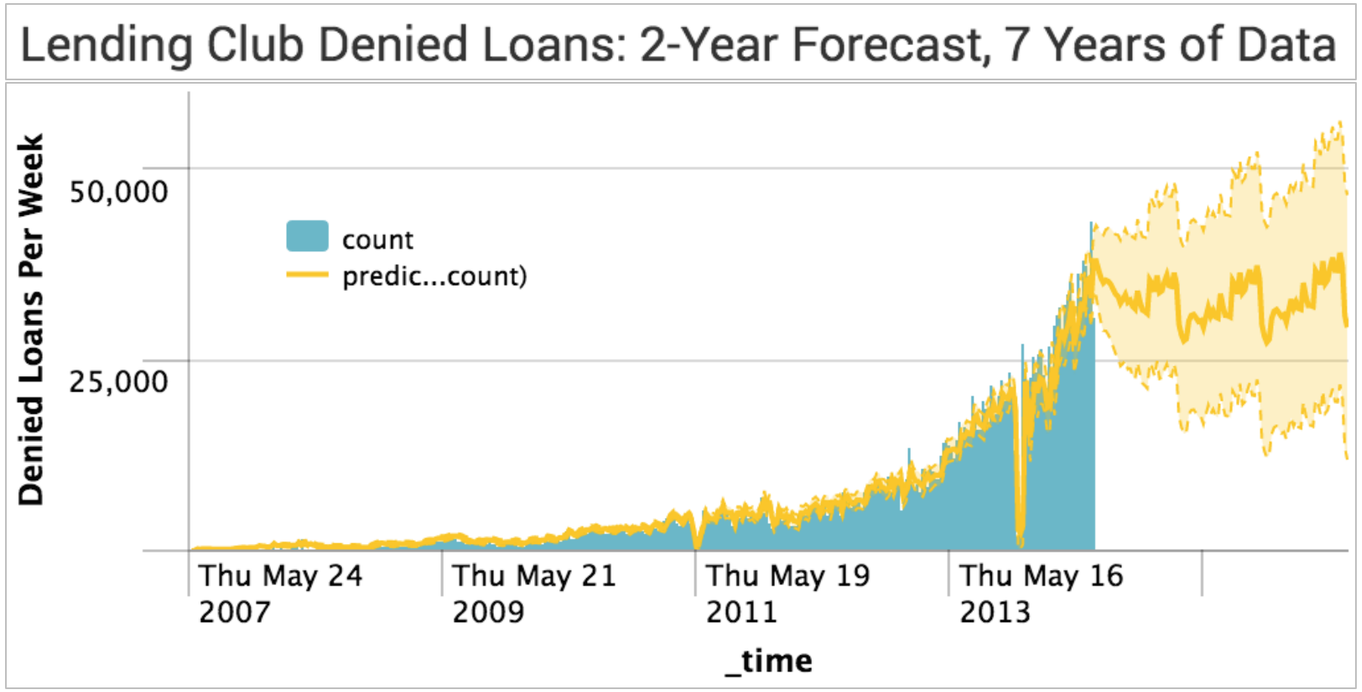
Estimation and Forecasting Use Cases

1. Business problem: capacity planning (IT)
   - Forecast web traffic and more
     - To estimate server usage in the future to improve resource allocation
2. Security threats (Security)
   - Identify patterns of anomalous behavior in network, firewall and ecommerce data
     - To target threats and block bad actors from disrupting the system
3. Customer conversion rates (Marketing)
   - Identify patterns in customer behavior
     - To target strategies to those market segments to maximize conversion rates

`predict` Command

- Forecasts trajectories of time series
  - Uses Kalman filter to identify seasonal trends
  - Gives a confidence interval as a buffer around the trend
- Uses lots of past data
- Includes low & high-frequency trends
- Can work with data that has missing values
  - Will calculate the best estimate for the missing value

```
| timechart span=7d count
| predict count future_timespan=104
```



`predict` Optional Arguments

```
predict <variable_to_predict> [AS <newfield_name>] [<predict_option>]
```

- **`algorithm=<string>`** method of forecasting being applied, **`default=LLP5`**
- **`future_timespan=<#>`** length of prediction, not used for LLB
- **`holdback=<#>`** will prevent the last points in the series from being used to train the model
- **`upper<int>`** and **`lower<int>=<field>`** **`<int>`** Specifies the percentage for the confidence interval. **`<field>`** specifies the custom output fieldname. **<int>** is 0-100; Defaults to **`lower95`** and **`upper95`**
- **`period=<num>`** length of a recurring cycle

`predict` Command Algorithm Options

Algorithm (Variations on the Kalman filter)	**Code**	**Univariate**	**Bivariate**	**Trends**	**Seasonality**
Local Level	LL	Yes	No	No	No
Seasonal Local Level	LLP	Yes	No	No	Yes
Local Level Trend	LLT	Yes	No	Yes	No
Bivariate Local Level	BiLL	No	Yes	No	No
Combo of LLT LLP	LLP5	Prediction and confidence interval based on weighted averages of LLT and LLP			

- The confidence interval does not cover 100% of the predictions
- The confidence interval describes probabilities

`predict` Command Algorithm Options

LL -- Local level

- Calculates a historical average, then adjusts it for noise. Will use that value as the forecast.
- The forecast will be a flat line.

LLT -- Local level trend

- A univariate model with trend, but no seasonality.
- Will forecast future values with a linear increasing or decreasing trend.
- Useful for high level predictions.

LLP -- Seasonal local level

- A univariate model with trend and seasonality.
- Will attempt to forecast the peaks and valleys.
- The number of data points must be at least twice length of the recurring cycle.
- If no period is set, this algorithm tries to calculate it. LLP returns an error message if the data is not periodic.

`predict` Command Algorithm Options

LLP5 -- Combines LLT and LLP models for its prediction.

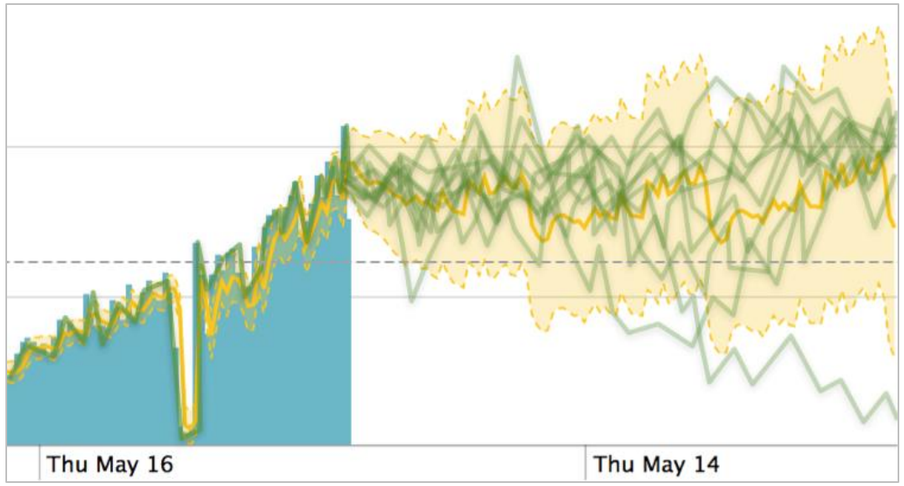
- If the time series is periodic, LLP5 computes two predictions, one using LLT and the other using LLP.
- Takes a weighted average of the two values and outputs that as the prediction.
- Capture more complex patterns and dynamics in the time series data.

BiLL -- Bivariate local level

- A bivariate model that predicts both time series simultaneously.
- The method captures the dynamics and relationship between the two fields.
- Used when there is a need to model the fields jointly.
- Like LL, it will produce a static forecast.

Interpreting the `predict` Command

- Interpreting **`predict`** is subtle
  - Probabilities (of most likely paths) concentrate around best-fit curve
  - Avoid choosing a single “best prediction”
- Uncertainty envelope (95% confidence)
  - With high probability, “true future path” stays *mostly* within
  - <=5% of the time it doesn’t

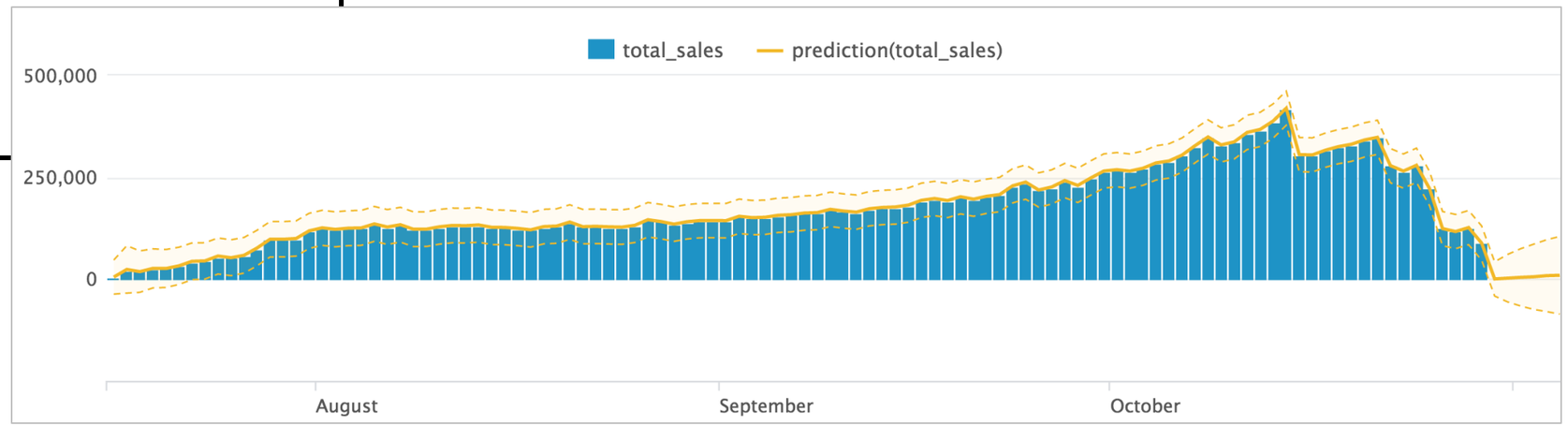


`predict` Command Example

Minimum of 2 - 4 cycles/seasons/periods of data

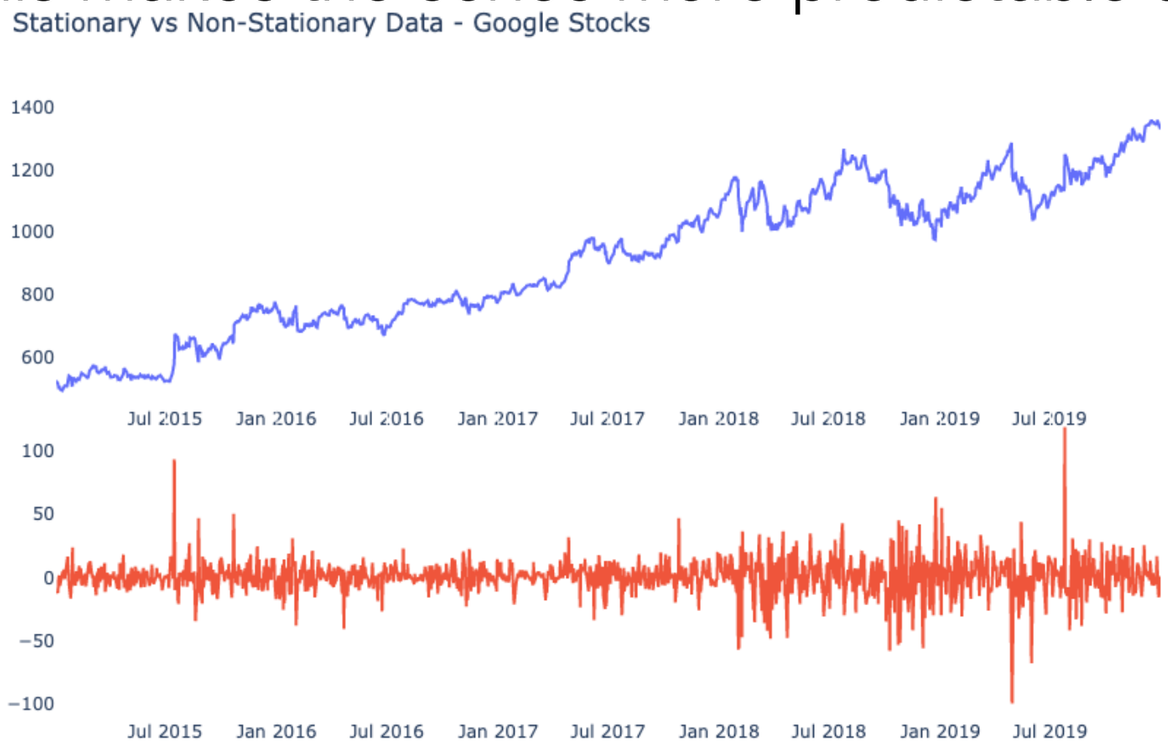
- To forecast quarterly, include at least 2 - 4 previous quarters
- Predict counts every time segment in your span as 1 data point
- More data? Go farther back in time and/or choose a finer span

```
sourcetype=access_combined
| timechart span=1d sum(price)
as total_sales
| predict total_sales
```



Time Series Stationarity

- A stationary time series has statistical properties (mean, variance, and autocorrelation) that remain constant over time
- This makes the series more predictable and easier to forecast.



Price of Google Stock over time

Change in Price of Google Stock over time

Statistical Tests: Scoring Methods

Test	Overview
Adfuller	Determine if the data is trending or not, even noisy time series datasets • If null hypothesis is accepted: data is nonstationary with a degree of certainty based on parameters set • If the null hypothesis is rejected, the data is stationary
KPSS	Determine if the data is level or trend stationary • If null hypothesis is accepted: data is stationary • If the null hypothesis is rejected, the data is nonstationary with a degree of certainty based on parameters set

If the p-value returned by the scoring methods is less then .05, it suggests differencing may be required to make the timeseries stationary. This would require calculating the difference between consecutive values and forecasting this difference.

https://docs.splunk.com/Documentation/MLApp/Latest/User/ScoreCommand#Statistical_testing_.28statstest.29

Topic 6 Lab

- Description: Compare different algorithms for forecasting
- Duration: 30 minutes
- Tasks:
  - Compare different options for the **`predict`** command
  - Use one field to help support predictions for a second field

Wrap Up

- Clean data for analytics
- Visualize data with different graphs and charts
- Cluster numerical data with **`kmeans`**
- Find string based clusters
- Identify field relationships with correlation tools
- Create a high-level transaction
- Find global and local numerical outliers
- Identity categorical anomalies
- Use the **`predict`** command to forecast timeseries data

Community

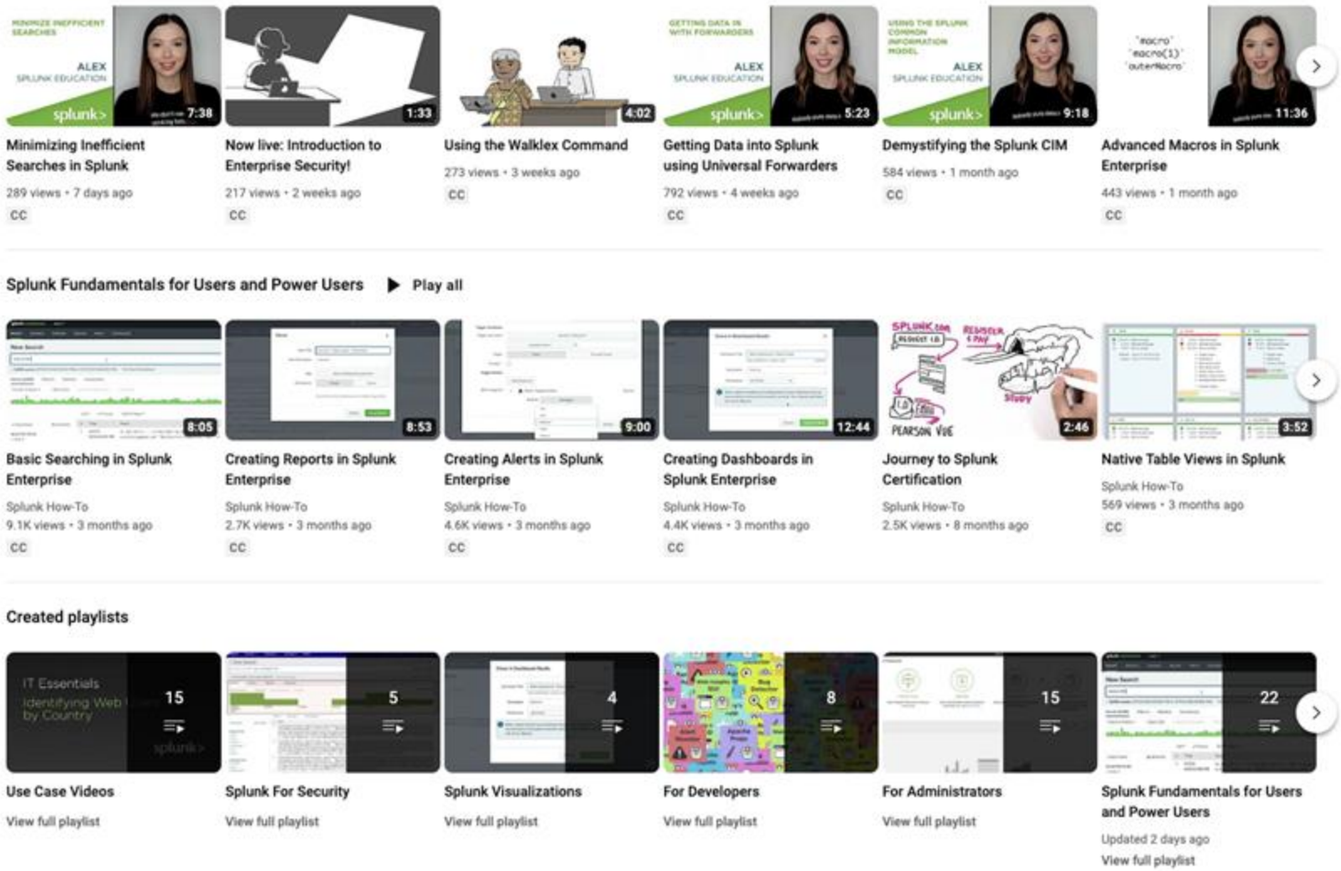
- Splunk Community Portal – community.splunk.com
  - Answers
  - Discussions
  - Splunk Trust
  - User Groups
  - Ideas
- Splunk Blogs – splunk.com/blog/
- Splunk Base – splunkbase.com
  - Apps
  - Curated Collections
- Splunk Docs on Twitter – twitter.com/splunkdocs
- Splunk Dev on Twitter – twitter.com/splunkdev
- Splunk on Slack – splk.it/slack
- .conf – conf.splunk.com

Support

- Knowledge Base – Search knowledge base, answers, and docs to troubleshoot your issue
- splunk>dev – Documentation for developers
- Splunk Docs – Product, best practices, and tools documentation for all Splunk products
- Splunk Lantern – Actionable guidance by experts
- Create a case – Support for critical issues
- Contact Us – Find region-specific support
  - (855) SPLUNK.S or (855) 775.8657
  - Not in the US? Find your local office
- System Status – Cloud Services, Observability Cloud, Splunk On-Call, Synthetic Monitoring
- Splunk Product Security – Critical Security Alerts, Quarterly Security Patches, and 3rd Party Bulletins

Splunk How-To Channel

Free, short videos on a variety of Splunk topics: splk.it/How-To



Learning Paths

Search Expert – Recommended Courses

Free eLearning courses are highlighted in blue and courses with an * are present in both learning paths.

- What is Splunk *
- Introduction to Splunk *
- Using Fields *
- Scheduling Reports and Alerts
- Visualizations
- Statistical Processing
- Working with Time
- Comparing Values
- Result Modification
- Leveraging Lookups and Subsearches
- Correlation Analysis
- Search Under the Hood
- Multivalue Fields
- Search Optimization *

Learning Paths

Knowledge Manager – Recommended Courses

Free eLearning courses are highlighted in blue and courses with an * are present in both learning paths.

- What is Splunk *
- Introduction to Splunk *
- Using Fields *
- Introduction to Knowledge Objects
- Creating Knowledge Objects
- Creating Field Extractions
- Enriching Data with Lookups
- Data Models
- Introduction to Dashboards
- Dynamic Dashboards
- Using Choropleth
- Search Optimization *

Splunk Mobile

- Free app available to all Splunk Cloud and Splunk Enterprise customers
- Analyze data and receive actionable alerts on-the-go with mobile-friendly dashboards
- iOS and Android
- See the Product Brief
- Download for iOS splk.it/ios

Thank You

splunk>
a CISCO company