Getting Your Data Ready for Machine Learning

Kristal Curtis | Software Engineer, Machine Learning

Adam J. Oliner | Director of Engineering

.conf18 | Orlando, FL

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Machine Learning Wants a Matrix

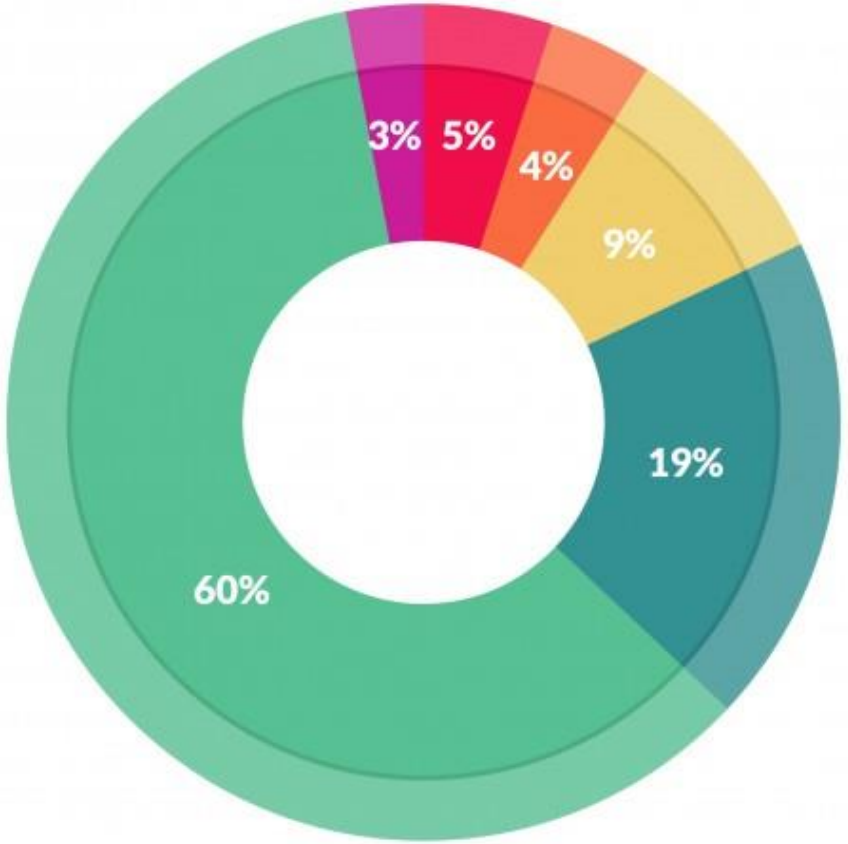
ideally with the minimal number of relevant features

What You Have

What You Want

Data Gathering and Prep

Source: CrowdFlower



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Want to learn more about MLTK? Check out the following session!

Using the Latest Features from the Splunk Machine Learning Toolkit to Create Your Own Custom Models

Speakers

Harsh Keswani, Product Manager, Machine Learning, Splunk
Adam J. Oliner, Director of Engineering, Splunk

Thursday, Oct 04, 12:15 p.m. - 1:00 p.m.

splunk> .conf18

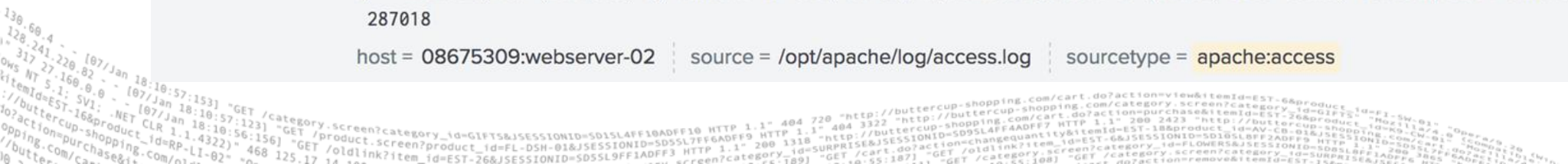
Data Cleaning Process

1. Semi-structured text
   a. Typical machine data
   b. Possibly some structure (e.g., key=value, JSON)
2. Tabular
   a. (row, column) = value
3. Matrix
   a. All numbers
   b. No missing values

Example

_raw

Time	Event
7/23/18 6:47:46.456 PM	176.207.200.6 www.jockorivercoffee.com - jfrazier13e 443 [23/Jul/2018 17:47:46:456442] "POST /cart?action=checkout&basket_contents=[{'POW2':'2'}]&JSESSIONID=86a52014-8ecc-11e8-b353-784f4371b6d6 HTTP 1.1" "?action=checkout&basket_contents=[{'POW2':'2'}]&JSESSIONID=86a52014-8ecc-11e8-b353-784f4371b6d6" 200 784 "http://www.jockorivercoffee.com/view?q=Morning_Warrior_Pod" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/534.24 (KHTML; like Gecko) Chrome/11.0.696.3 Safari/534.24" 100 444 275253 method=GIFTCARD saleamount=19.98 ErrorCode=99999 host = 08675309:webserver-03 | source = /opt/apache/log/access.log | sourcetype = apache:access
7/23/18 6:47:34.055 PM	132.84.138.75 www.jockorivercoffee.com - ifox15n 443 [23/Jul/2018 17:47:34:055102] "POST /view?product_name=Morning_Warrior_Instant&JSESSIONID=86a3c930-8ecc-11e8-b42d-784f4371b6d6 HTTP 1.1" "?product_name=Morning_Warrior_Instant&JSESSIONID=86a3c930-8ecc-11e8-b42d-784f4371b6d6" 200 606 "http://www.jockorivercoffee.com/search?q=Morning_Warrior_Instant" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US) AppleWebKit/533.17.8 (KHTML; like Gecko) Version/5.0.1 Safari/533.17.8" 113 450 344582 host = 08675309:webserver-04 | source = /opt/apache/log/access.log | sourcetype = apache:access
7/23/18 6:47:24.674 PM	176.207.200.6 www.jockorivercoffee.com - jfrazier13e 443 [23/Jul/2018 17:47:24:674640] "POST /cart?action=checkout&basket_contents=[{'POW2':'2'}]&JSESSIONID=86a52014-8ecc-11e8-b353-784f4371b6d6 HTTP 1.1" "?action=checkout&basket_contents=[{'POW2':'2'}]&JSESSIONID=86a52014-8ecc-11e8-b353-784f4371b6d6" 200 602 "http://www.jockorivercoffee.com/view?q=Morning_Warrior_Pod" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/534.24 (KHTML; like Gecko) Chrome/11.0.696.3 Safari/534.24" 88 523 316558 method=GIFTCARD host = 08675309:webserver-01 | source = /opt/apache/log/access.log | sourcetype = apache:access
7/23/18 6:47:22.641 PM	132.84.138.75 www.jockorivercoffee.com - ifox15n 443 [23/Jul/2018 17:47:22:641988] "GET /search?q=Vanilla&JSESSIONID=86a3c930-8ecc-11e8-b42d-784f4371b6d6 HTTP 1.1" "?q=Vanilla&JSESSIONID=86a3c930-8ecc-11e8-b42d-784f4371b6d6" 200 930 "http://www.jockorivercoffee.com/" "Mozilla/5.0 (Windows; U; Windows NT 5.2; en-US) AppleWebKit/533.17.8 (KHTML; like Gecko) Version/5.0.1 Safari/533.17.8" 103 529 287018 host = 08675309:webserver-02 | source = /opt/apache/log/access.log | sourcetype = apache:access



splunk> .conf18

Example

Messy Table

JSESSIONID	method_modified	request_size	response_size	response_time
95ca35e8-8ecc-11e8-9b3b-784f4371b6d6	GIFTCARD	0.09		2429370
95ca35e8-8ecc-11e8-9b3b-784f4371b6d6	GIFTCARD	0.084	511	2826820
95c5b8c5-8ecc-11e8-9d94-784f4371b6d6	GIFTCARD	0.09		2467710
952ad821-8ecc-11e8-8568-784f4371b6d6	CREDITCARD	0.082		2411460
952ad821-8ecc-11e8-8568-784f4371b6d6	CREDITCARD	0.115	559	3167550
952ad821-8ecc-11e8-8568-784f4371b6d6	CREDITCARD	0.103	500	2634080
952ad821-8ecc-11e8-8568-784f4371b6d6		0.081	522	3252010
9529c8e1-8ecc-11e8-a493-784f4371b6d6	GIFTCARD	0.098	501	3172080
9529c8e1-8ecc-11e8-a493-784f4371b6d6		0.114	520	3012260
95286c40-8ecc-11e8-a8bf-784f4371b6d6	GIFTCARD	0.118	514	3219490
95286c40-8ecc-11e8-a8bf-784f4371b6d6	GIFTCARD	0.094	581	2578870
95286c40-8ecc-11e8-a8bf-784f4371b6d6	GIFTCARD	0.088	494	3437350

splunk> .conf18

Example

Clean, Numeric Table

method_numeric	request_bytes	response_bytes	response_time	SS_request_bytes	SS_response_bytes	SS_response_time
1	117	484	28.633	1.52829611445	-0.273100937052	-5.6116822758
0	118	560	24.977	1.61486301221	1.04517158866	-5.68840180219
0	107	581	25.55	0.662627136916	1.40943110235	-5.67637765322
0	88	599	25.136	-0.982143920403	1.72165354265	-5.68506525824
1	94	521	26.524	-0.462742533881	0.368689634678	-5.65593869844
0	89	599	26.11	-0.89557702265	1.72165354265	-5.66462630345
0	84	582	29.068	-1.32841151142	1.42677679348	-5.60255399517
1	90	599	242.937	-0.809010124896	1.72165354265	-1.11460859337
0	84	511	282.682	-1.32841151142	0.195232723399	-0.280577527647
1	90	559	246.771	-0.809010124896	1.02782589754	-1.03415381652

splunk> .conf18

Outline

1. Identify the right data
2. Clean the data
3. Engineer features
4. Split the data

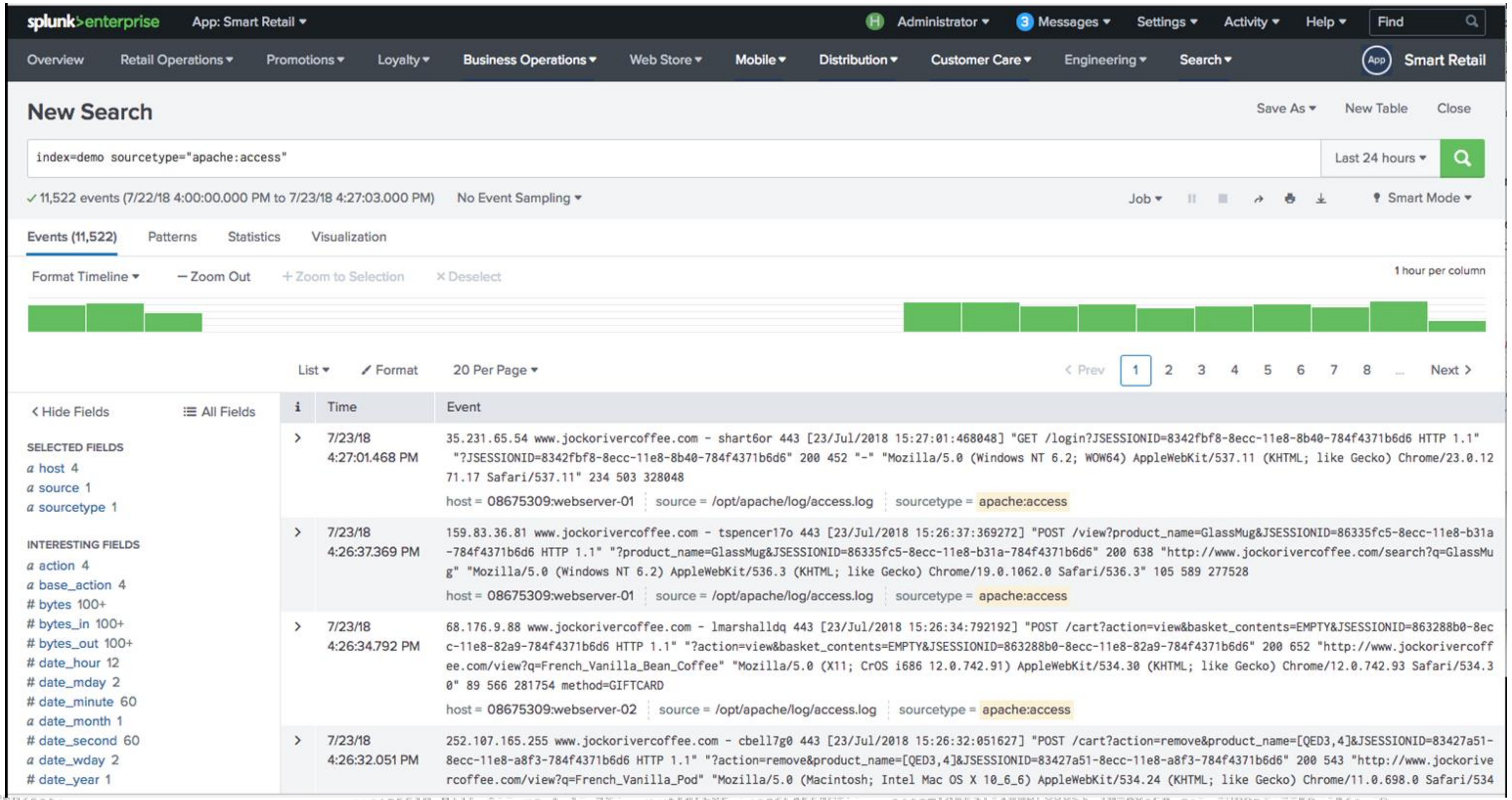
Identify the Data

relevance and sufficiency

splunk> .conf18

Example Dataset

- Donut shop
  - Point-of-sale system from brick & mortar store
  - Transaction data from online sales (coffee, gear)



Identify the Data

- Labels
  - Do you have examples of the task being performed correctly?
- Relevance
  - Would the information help a human perform the task?
- Completeness
  - Do you need to integrate multiple datasets?
- Weighting
  - Is older data less valid or useful than recent data?

Data Integration

- Joins are natural (cf. SQL)
- But, Splunk is not optimized for joins, so they are inefficient
- Preferable to use OR
- e.g., Correlating user logins & logouts by session ID

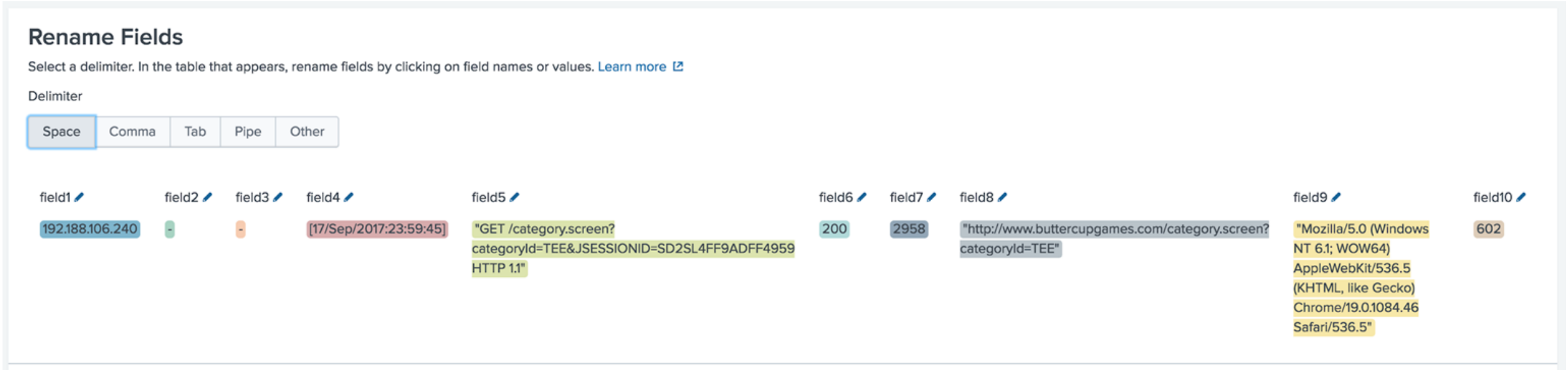
```
sourcetype=login | join sessionID [ search sourcetype=logout ]
```

vs.

```
sourcetype=login OR sourcetype=logout
```

Field Extraction

- Automatic
- TA
- Manual / IFX



Cleaning

tidying up what's messy or missing

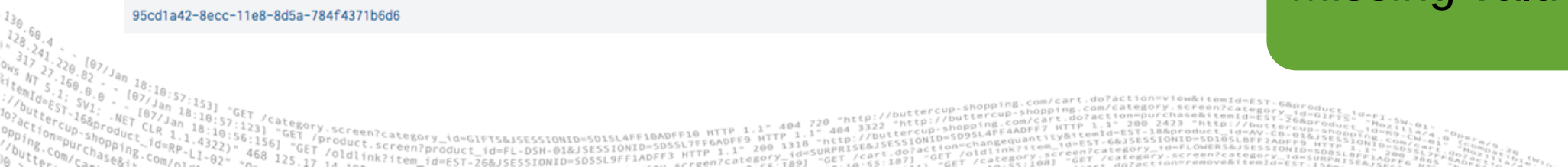
Why So Dirty?

- User Input
- Versioning
- Upstream Processing
- Dirty Source

Cleaning

JSESSIONID	bytes_in	bytes_in_modified
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	117	
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	118	118
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	107	107
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	88	88
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	94	
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	89	89
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	84	84
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	104	104
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	195	195
95ce6b68-8ecc-11e8-931b-784f4371b6d6	106	106
95ce6b68-8ecc-11e8-931b-784f4371b6d6	103	103
95ce6b68-8ecc-11e8-931b-784f4371b6d6	100	
95ce6b68-8ecc-11e8-931b-784f4371b6d6	81	81
95ce6b68-8ecc-11e8-931b-784f4371b6d6	213	213
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6		85
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6		
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6		
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6		173

Missing values



splunk> .conf18

Cleaning: fillnull

```
index=demo sourcetype="apache:access"
| search bytes_in=*
| eval bytes_in_modified = if (random() % 5 = 0, null, bytes_in)
| table JSESSIONID, bytes_in, bytes_in_modified
| fillnull value=0
```

All time

✓ 201,654 events (before 9/26/18 5:40:37.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (201,654) Visualization

20 Per Page Format Preview < Prev 1 2 3 4 5 6 7 8 ... Next >

JSESSIONID	bytes_in	bytes_in_modified
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	117	117
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	118	118
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	107	107
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	88	88
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	94	94
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	89	89
95cfd71e-8ecc-11e8-bb90-784f4371b6d6		0
95cfd71e-8ecc-11e8-bb90-784f4371b6d6		104
95cfd71e-8ecc-11e8-bb90-784f4371b6d6		195
95ce6b68-8ecc-11e8-931b-784f4371b6d6		106
95ce6b68-8ecc-11e8-931b-784f4371b6d6		103
95ce6b68-8ecc-11e8-931b-784f4371b6d6		100
95ce6b68-8ecc-11e8-931b-784f4371b6d6	81	81
95ce6b68-8ecc-11e8-931b-784f4371b6d6	213	213
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	85	0
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	106	106
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	80	80

Replace with 0

splunk> .conf18

Cleaning

- Missing values

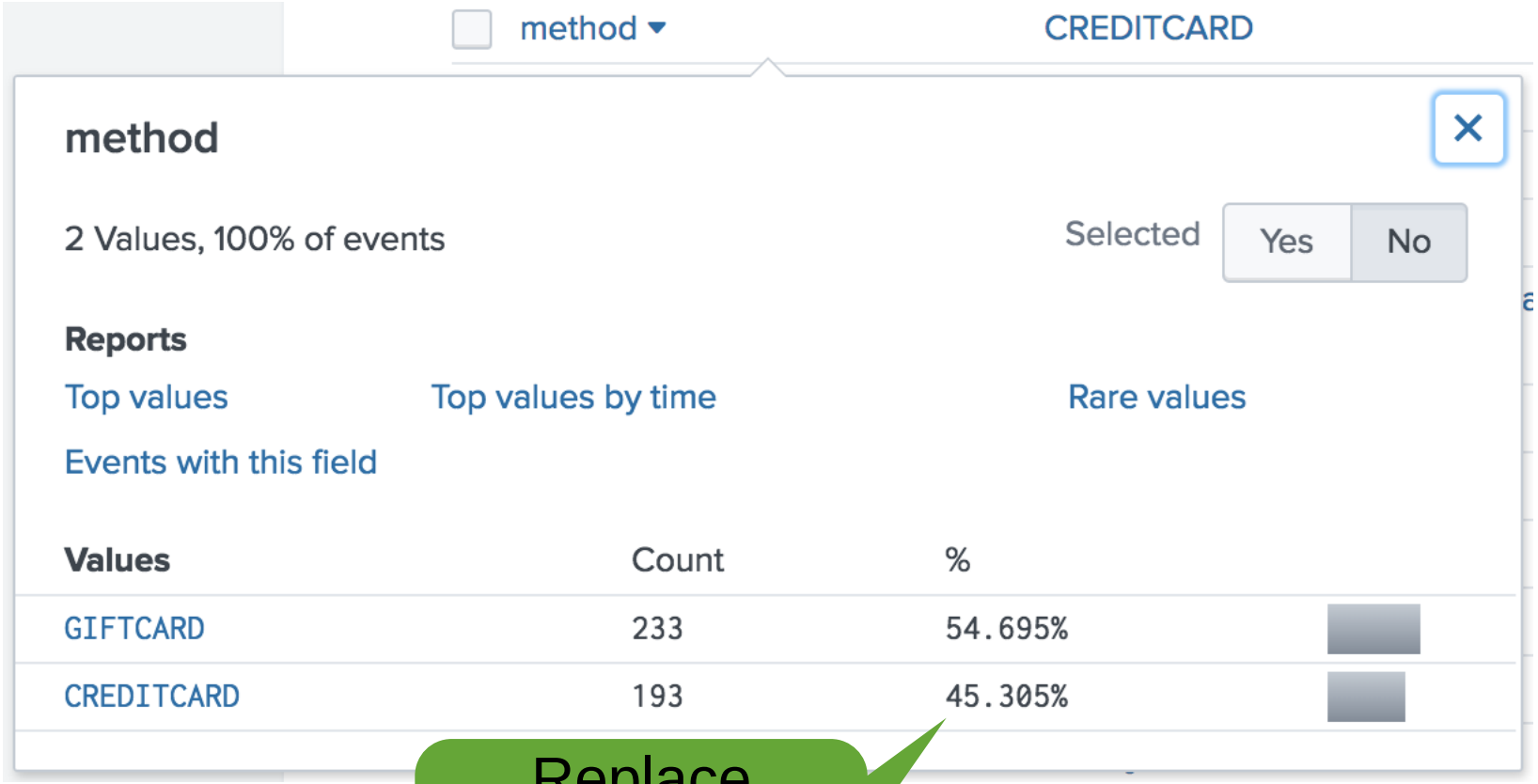
JSESSIONID	basket_contents	newSaleAmount
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	[{'QED1':'2'}]	
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	[{'QED1':'2'}]	36.2
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	[{'QED1':'2'}]	36.2
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	[{'QED1':'2'}]	36.2
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	[{'QED1':'2'}]	36.2
951d918a-8ecc-11e8-95c0-784f4371b6d6	[{'COF1A':'4'}]	35.96
9484aa23-8ecc-11e8-8719-784f4371b6d6	[{'COF1A':'2'}]	17.98
947a27c0-8ecc-11e8-8c39-784f4371b6d6	[{'COF1A':'1'}]	8.99
947a27c0-8ecc-11e8-8c39-784f4371b6d6	[{'COF1A':'1'}]	8.99
94775357-8ecc-11e8-9003-784f4371b6d6	[{'COF1':'1','CAP1':'1'}]	29.09
93db0138-8ecc-11e8-bd40-784f4371b6d6	[{'CAP1':'2'}]	40.2
93da7111-8ecc-11e8-b2ea-784f4371b6d6	[{'QED1':'15'}]	

Compute missing sale amount

splunk> .conf18

Cleaning

- Missing values



Replace according to this probability

splunk> .conf18

Cleaning

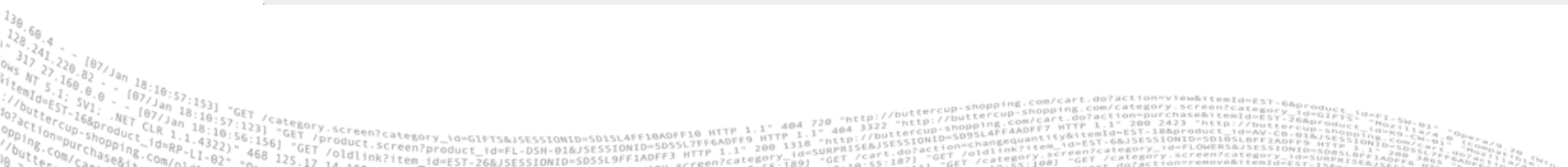
```
index=demo sourcetype="apache:access" method=*
| eval method_modified = if (random() % 4 = 0, null, method)
| table JSESSIONID, method, method_modified
| eval method_modified_filled = if (isnull(method_modified), if(random() % 100 < 54, "_GIFTCARD", "_CREDITCARD"), method_modified)
```

✓ 18,829 events (7/14/18 12:00:00.000 AM to 8/13/18 6:06:56.000 PM) No Event Sampling ▾ Job ▾

Events Patterns **Statistics (18,829)** Visualization

20 Per Page ▾ Format Preview ▾ < Prev 1 2 3 4 5 6

JSESSIONID	method	method_modified	method_modified_filled
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD	GIFTCARD	GIFTCARD
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD		_GIFTCARD
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD		_GIFTCARD
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD	GIFTCARD	GIFTCARD
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD	GIFTCARD	GIFTCARD
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD	GIFTCARD	GIFTCARD
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	GIFTCARD		_GIFTCARD
95ca35e8-8ecc-11e8-9b3b-784f4371b6d6	GIFTCARD	GIFTCARD	GIFTCARD
95ca35e8-8ecc-11e8-9b3b-784f4371b6d6	GIFTCARD		_GIFTCARD
95c5b8c5-8ecc-11e8-9d94-784f4371b6d6	GIFTCARD		_GIFTCARD
952ad821-8ecc-11e8-8568-784f4371b6d6	CREDITCARD		_CREDITCARD
952ad821-8ecc-11e8-8568-784f4371b6d6	CREDITCARD	CREDITCARD	CREDITCARD
952ad821-8ecc-11e8-8568-784f4371b6d6	CREDITCARD	CREDITCARD	CREDITCARD



splunk> .conf18

Cleaning

- Missing values

base_action	http_method	request_bytes_new	response_bytes_new
/cart	POST		484
/cart	POST	118	560
/cart	POST		
/cart	POST	88	599
/cart	POST	94	521
/cart	POST	89	
/cart	POST	84	582
/view	POST	104	577
/login	GET	195	432
/cart	POST	106	479
/search	GET	103	
/view	POST	100	533

Impute missing values

Impute with Model Predictions

It's ML all the way down

- Train a model to predict sometimes-missing fields from known fields

```
| fit RandomForestRegressor "request_bytes_new" from "http_method" "base_action" into rbn_model
```

- Use the model to fill in the missing values via eval

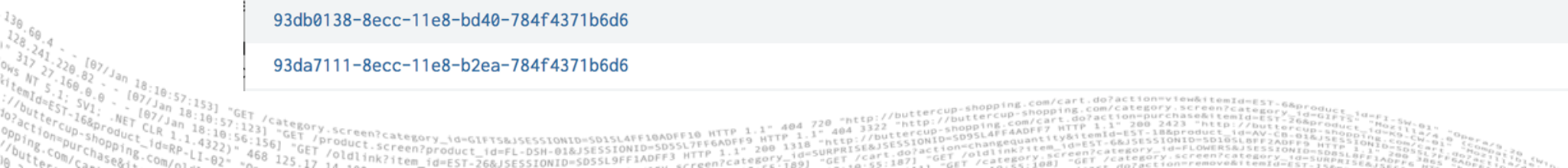
```
| apply rbn_model as "Imputed_request_bytes"
| eval request_bytes = if(isnull("request_bytes_new"), "Imputed_request_bytes", "request_bytes_new")
```

Cleaning

- Unit disagreement (bytes vs. kilobytes)

JSESSIONID	response_bytes_new
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	4840
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	5600
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	5810
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	5990
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	5210
9529c8e1-8ecc-11e8-a493-784f4371b6d6	5010
95286c40-8ecc-11e8-a8bf-784f4371b6d6	5140
95286c40-8ecc-11e8-a8bf-784f4371b6d6	5.67
951d918a-8ecc-11e8-95c0-784f4371b6d6	5100
9484aa23-8ecc-11e8-8719-784f4371b6d6	5.61
947a27c0-8ecc-11e8-8c39-784f4371b6d6	5510
947a27c0-8ecc-11e8-8c39-784f4371b6d6	5540
94775357-8ecc-11e8-9003-784f4371b6d6	4650
93db0138-8ecc-11e8-bd40-784f4371b6d6	4.20
93da7111-8ecc-11e8-b2ea-784f4371b6d6	4.47

Possible kilobytes



splunk> .conf18

Cleaning

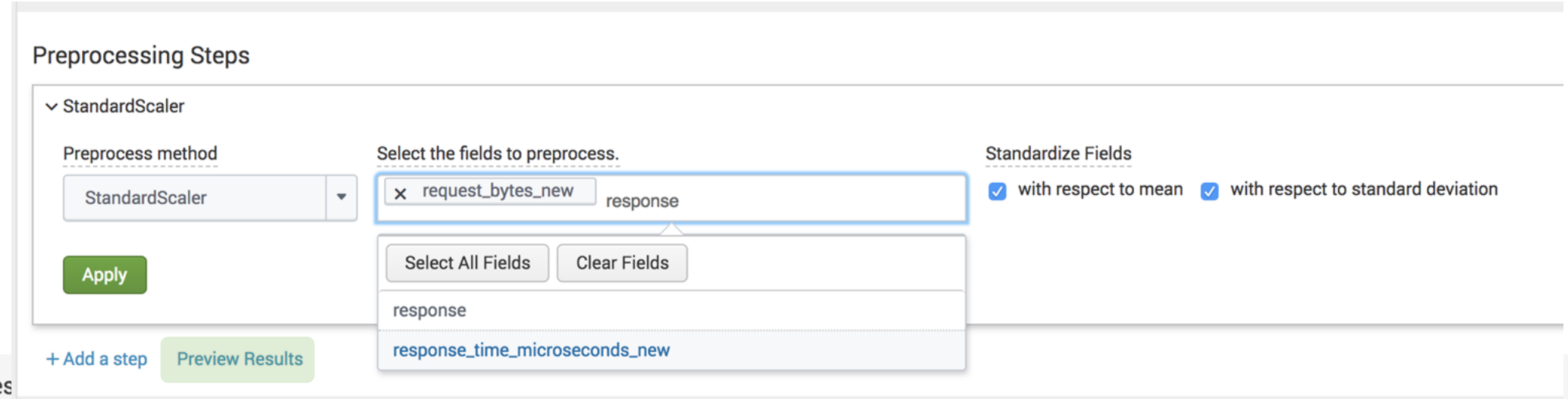
- Scale (0.15 vs. 1M)
  - Fix via StandardScaler in MLTK

JSESSIONID	request_bytes_new	response_time_microseconds_new
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.117	286330
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.118	249770
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.107	255500
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.088	251360
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.094	265240
9529c8e1-8ecc-11e8-a493-784f4371b6d6	0.098	3172080
95286c40-8ecc-11e8-a8bf-784f4371b6d6	0.118	3219490
95286c40-8ecc-11e8-a8bf-784f4371b6d6	0.094	2578870
951d918a-8ecc-11e8-95c0-784f4371b6d6	0.09	3253630
9484aa23-8ecc-11e8-8719-784f4371b6d6	0.098	2938310
947a27c0-8ecc-11e8-8c39-784f4371b6d6	0.113	2484670
947a27c0-8ecc-11e8-8c39-784f4371b6d6	0.087	3228460
94775357-8ecc-11e8-9003-784f4371b6d6	0.088	3262560

splunk> .conf18

Fix: StandardScaler

Preprocessing Step in MLTK



Preproces

JSESSIONID	request_bytes_new	response_time_microseconds_new	SS_request_bytes_new	SS_response_time_microseconds_new
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.117	286330	1.5435055898	-3.74666753942
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.118	249770	1.63035576601	-3.79991360299
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.107	255500	0.675003827612	-3.79156841852
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.088	251360	-0.975149520538	-3.79759792353
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	0.094	265240	-0.454048463228	-3.77738306132
9529c8e1-8ecc-11e8-a493-784f4371b6d6	0.098	3172080	-0.106647758354	0.456145041376
95286c40-8ecc-11e8-a8bf-784f4371b6d6	0.118	3219490	1.63035576601	0.525193068563
95286c40-8ecc-11e8-a8bf-784f4371b6d6	0.094	2578870	-0.454048463228	-0.407807293738
951d918a-8ecc-11e8-95c0-784f4371b6d6	0.09	3253630	-0.801449168101	0.574914638868
9484aa23-8ecc-11e8-8719-784f4371b6d6	0.098	2938310	-0.106647758354	0.115681904579

New Fields

Cleaning

- Numerical vs. Categorical vs. Numerical that should be Categorical

JSESSIONID	bytes_in	q	ErrorCode
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	117	Morning_Warrior_Pod	9999999
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	118	Morning_Warrior_Pod	9999999
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	107	Morning_Warrior_Pod	9999999
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	88	Morning_Warrior_Pod	9999999
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	94	Morning_Warrior_Pod	9999999
9529c8e1-8ecc-11e8-a493-784f4371b6d6	98	Morning_Warrior_Ground_Coffee	99999
95286c40-8ecc-11e8-a8bf-784f4371b6d6	118	Morning_Warrior_Cap	99999
95286c40-8ecc-11e8-a8bf-784f4371b6d6	94	Morning_Warrior_Cap	91512
951d918a-8ecc-11e8-95c0-784f4371b6d6	90	Morning_Warrior_Ground_Coffee	81535
9484aa23-8ecc-11e8-8719-784f4371b6d6	98	Morning_Warrior_Bean_Coffee	99999
947a27c0-8ecc-11e8-8c39-784f4371b6d6	113	Morning_Warrior_Bean_Coffee	99999
947a27c0-8ecc-11e8-8c39-784f4371b6d6	87	Morning_Warrior_Bean_Coffee	81715
94775357-8ecc-11e8-9003-784f4371b6d6		ning_Warrior_Ground_Coffee	99999

Obviously numeric

Should be categorical

Obviously categorical

splunk> .conf18

Cleaning

Numerical => Categorical

```
index=demo sourcetype="apache:access" saleamount="*"
| eval ErrorCodeCategorical = ErrorCode."_str"
| table JSESSIONID, ErrorCode, ErrorCodeCategorical
```

✓ 3,271 events (7/15/18 12:00:00.000 AM to 8/14/18 9:39:33.000 AM) No Event Sampling ▾ Job ▾

Events Patterns **Statistics (3,271)** Visualization

20 Per Page ▾ Format Preview ▾ < Prev 1 2

ErrorCode 100+

a ErrorCodeCategorical 100+

JSESSIONID	ErrorCode	ErrorCodeCategorical
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	9999999	9999999_str
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	9999999	9999999_str
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	9999999	9999999_str
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	9999999	9999999_str
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	9999999	9999999_str
9529c8e1-8ecc-11e8-a493-784f4371b6d6	99999	99999_str
95286c40-8ecc-11e8-a8bf-784f4371b6d6	99999	99999_str
95286c40-8ecc-11e8-a8bf-784f4371b6d6	91512	91512_str
951d918a-8ecc-11e8-95c0-784f4371b6d6	81535	81535_str
9484aa23-8ecc-11e8-8719-784f4371b6d6	99999	99999_str
947a27c0-8ecc-11e8-8c39-784f4371b6d6	99999	99999_str
947a27c0-8ecc-11e8-8c39-784f4371b6d6	81715	81715_str
94775357-8ecc-11e8-9003-784f4371b6d6	99999	99999_str

Cleaning

Categorical => Numerical

```
index=demo sourcetype="apache:access"
| eval severity = case(base_action == "/view", "CRITICAL", base_action == "/search", "WARN", base_action == "/cart", "INFO", true(), "DEBUG")
| eval severity_numerical = case(severity == "INFO", 1, severity == "DEBUG", 2, severity == "WARN", 3, severity == "CRITICAL", 4)
| table JSESSIONID, severity, severity_numerical
```

All time

✓ 201,654 events (before 9/26/18 6:11:59.000 PM) No Event Sampling

Job Smart Mode

Events Patterns Statistics (201,654) Visualization

20 Per Page Format Preview

< Prev 1 2 3 4 5 6 7 8 ... Next >

a severity 4

severity_numerical 4

JSESSIONID	severity	severity_numerical
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	INFO	1
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	CRITICAL	4
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	DEBUG	2
95ce6b68-8ecc-11e8-931b-784f4371b6d6	INFO	1
95ce6b68-8ecc-11e8-931b-784f4371b6d6	WARN	3
95ce6b68-8ecc-11e8-931b-784f4371b6d6	CRITICAL	4
95ce6b68-8ecc-11e8-931b-784f4371b6d6	WARN	3
95ce6b68-8ecc-11e8-931b-784f4371b6d6	DEBUG	2
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	INFO	1
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	WARN	3
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	WARN	3
95cd1a42-8ecc-11e8-8d5a-784f4371b6d6	DEBUG	2

Cleaning
Inconsistencies

```
index=demo sourcetype="apache:access" tag=*
| table JSESSIONID, tag
```

✓ 102,453 events (7/14/18 12:00:00.000 AM to 8/13/18 5:53:42.000 PM) No Event Sampling ▾ Job ▾

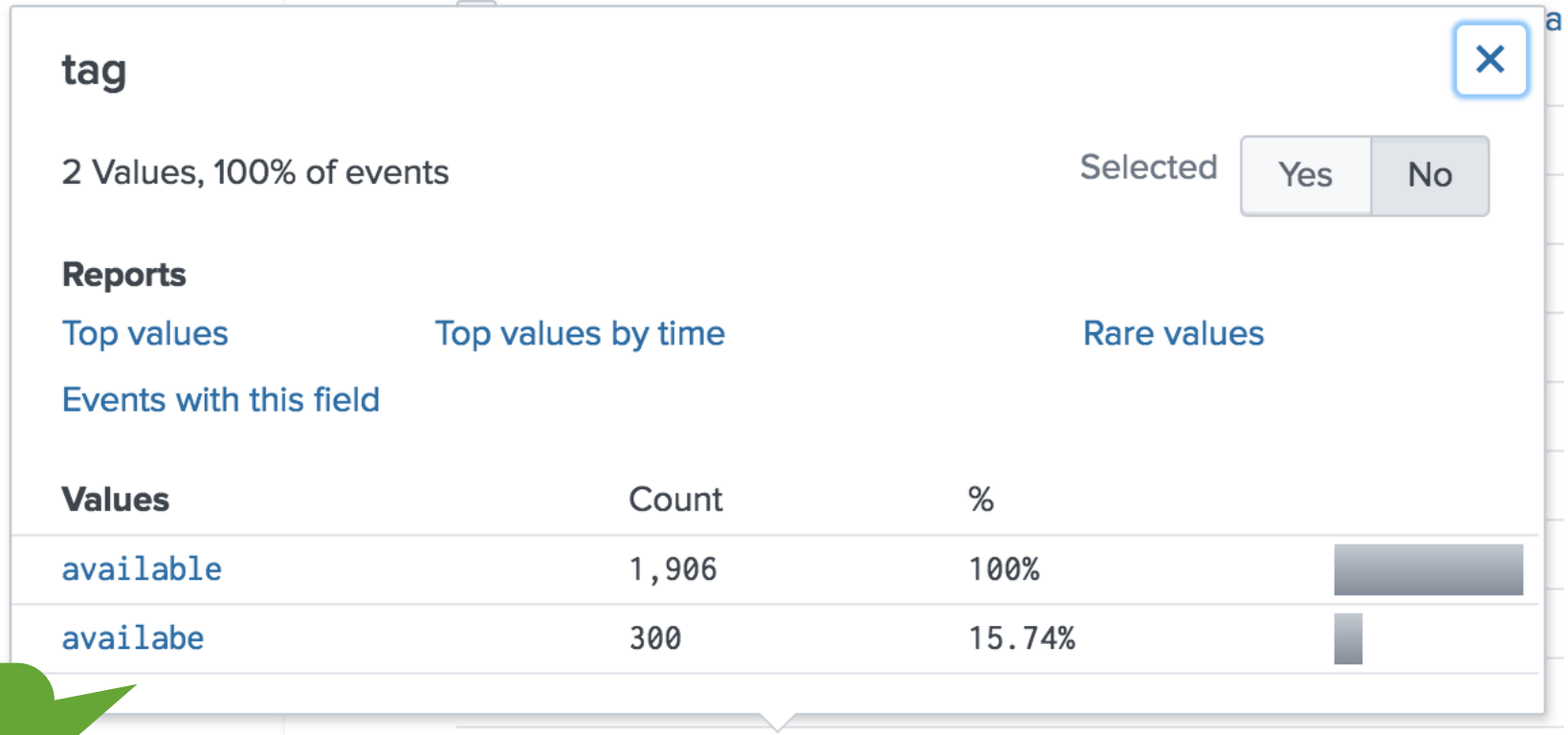
Events Patterns **Statistics (102,453)** Visualization

20 Per Page ▾ Format Preview ▾ ‹ Prev 1 2 3 4 5 6

JSESSIONID	tag
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95cfd71e-8ecc-11e8-bb90-784f4371b6d6	available
95ce6b68-8ecc-11e8-931b-784f4371b6d6	available
95ce6b68-8ecc-11e8-931b-784f4371b6d6	available
95ce6b68-8ecc-11e8-931b-784f4371b6d6	available
95ce6b68-8ecc-11e8-931b-784f4371b6d6	available
95ce6b68-8ecc-11e8-931b-784f4371b6d6	available

Cleaning

- Inconsistencies



Erroneous value

Cleaning
Inconsistencies

```
index=demo sourcetype="apache:access" tag=availabe
| table JSESSIONID, tag
```

✓ 8,398 events (7/14/18 12:00:00.000 AM to 8/13/18 5:55:45.000 PM) No Event Sampling ▾ Job ▾

Events Patterns **Statistics (8,398)** Visualization

20 Per Page ▾ Format Preview ▾ < Prev 1 2 3 4 5 6

JSESSIONID	tag
348e928c-8ecb-11e8-abf3-784f4371b6d6	availabe available
348e928c-8ecb-11e8-abf3-784f4371b6d6	availabe available
348e928c-8ecb-11e8-abf3-784f4371b6d6	availabe available
348e928c-8ecb-11e8-abf3-784f4371b6d6	availabe available
348e928c-8ecb-11e8-abf3-784f4371b6d6	availabe available
348e928c-8ecb-11e8-abf3-784f4371b6d6	availabe available
348e5197-8ecb-11e8-b515-784f4371b6d6	availabe available

splunk> .conf18

Feature Engineering

making the implicit, explicit

Feature Engineering: stats

- Work on aggregates, not raw events

```
index=detailed_activity_logs
| stats count by action user_id
| xyseries user_id action count
| fillnull
| lookup user_activity user_id OUTPUT days_active visits_per_day
| fit KMeans k=10 days_active visits_per_day
```

Feature Engineering: eval

- Model feature interactions

```
... | eval numeric_factor = importance * urgency | ...

... | eval categorical_factor = role + industry | ...
```

- Introduce nonlinearity

```
... | eval temperature_sq = pow(temperature, 2) | ...

... | eval latency_log = log(latency) | ...
```

New Search

Save As Close

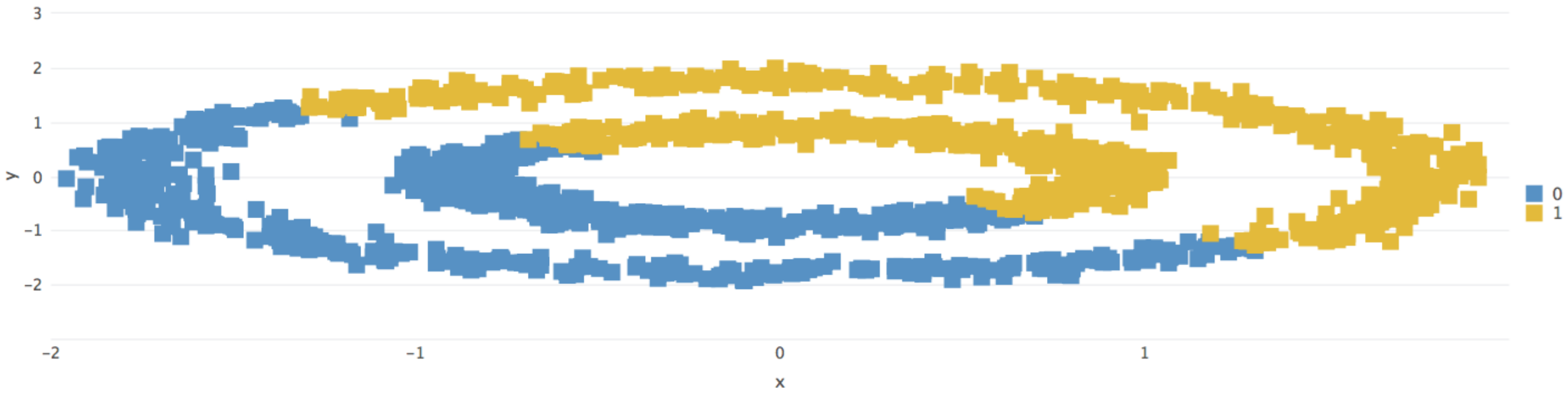
```
| inputlookup sklearn_cluster_noisy_circles.csv
| eval x=tonumber(x), y=tonumber(y)
| fit KMeans k=2 x y
| table cluster x y
```

Last 24 hours

✓ 1,500 results (5/10/18 5:00:00.000 PM to 5/11/18 5:25:51.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (1,500) Visualization

Scatter Chart Format Trellis



New Search

Save As Close

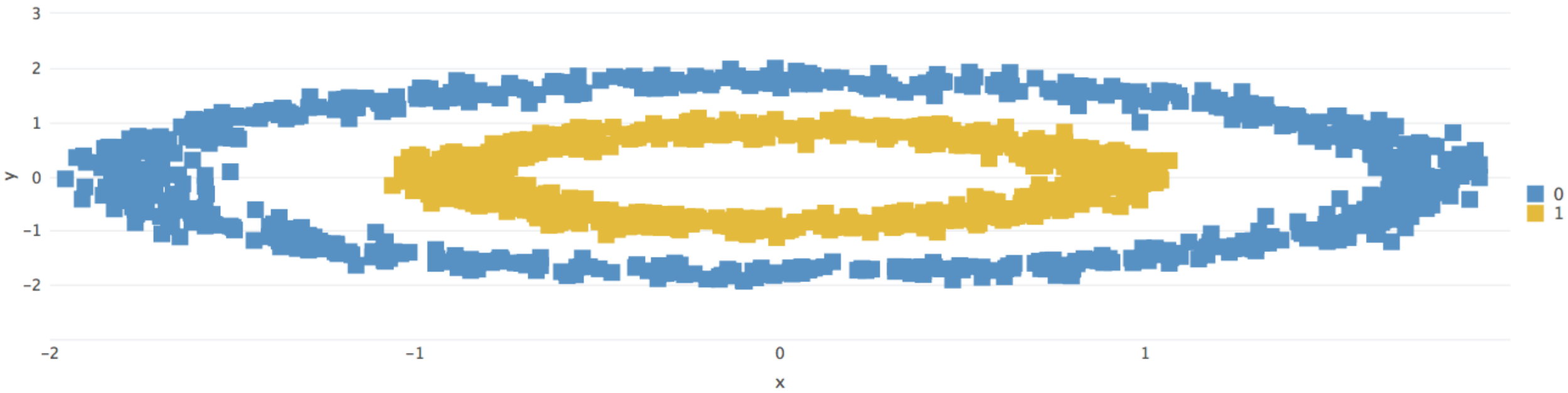
```
| inputlookup sklearn_cluster_noisy_circles.csv
| eval r=sqrt(x*x + y*y)
| eval x=tonumber(x), y=tonumber(y)
| fit KMeans k=2 r
| table cluster x y
```

Last 24 hours

✓ 1,500 results (5/10/18 5:00:00.000 PM to 5/11/18 5:27:08.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (1,500) Visualization

Scatter Chart Format Trellis



Feature Engineering: Text Analytics

- Transform free-form text to numeric fields with TFIDF
- Then…
  - Group similar text with KMeans
  - Classify text with BernoulliNB
  - Visualize the distribution of text with PCA

```
...
| fit TFIDF message
| fit KMeans message_ftidf_*
| ...
```

New Search

Save As Close

```
index=_internal
| sample 10000
| fit TFIDF token_pattern="\w\w\w+" _raw as raw_tfidf
| fit KMeans k=10 raw_tfidf_*
| sample 5 by cluster
| stats list(_raw) by cluster
```

Last 24 hours

✓ 0 events (5/10/18 5:00:00.000 PM to 5/11/18 5:02:31.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (10) Visualization

20 Per Page Format Preview

cluster	list(_raw)
9	05-11-2018 09:47:32.390 +0000 INFO Metrics - group=tpool, name=bundlereplthreadpool, qsize=0, workers=0, qwork_units=0 05-11-2018 08:53:17.390 +0000 INFO Metrics - group=tpool, name=indexertpool, qsize=0, workers=2, qwork_units=0 05-11-2018 04:14:48.390 +0000 INFO Metrics - group=tpool, name=bundlereplthreadpool, qsize=0, workers=0, qwork_units=0 05-10-2018 21:32:50.390 +0000 INFO Metrics - group=tpool, name=indexertpool, qsize=0, workers=2, qwork_units=0 05-10-2018 17:00:02.390 +0000 INFO Metrics - group=tpool, name=bundlereplthreadpool, qsize=0, workers=0, qwork_units=0
8	05-11-2018 10:02:31.390 +0000 INFO Metrics - group=per_source_thruput, series="/home/grana/splunk/var/log/introspection/resource_usage.log", kbps=0.3483515754958008, eps=0.7741986181393381, kb=10.798828125, ev=24, avg_age=0, max_age=0 05-11-2018 04:16:21.390 +0000 INFO Metrics - group=per_source_thruput, series="/home/grana/splunk/var/log/introspection/kvstore.log", kbps=0.30872193903563383, eps=0.19354975443213615, kb=9.5703125, ev=6, avg_age=0, max_age=0 05-11-2018 02:59:22.390 +0000 INFO Metrics - group=per_source_thruput, series="/home/grana/splunk/var/log/splunk/metrics.log", kbps=0.38870189744116457, eps=2.354829746131934, kb=12.0498046875, ev=73, avg_age=2.1780821917808217, max_age=3 05-11-2018 02:40:15.391 +0000 INFO Metrics - group=per_source_thruput, series="/home/grana/splunk/var/log/introspection/resource_usage.log", kbps=0.3581391199540706, eps=0.8386925789125848, kb=11.1025390625, ev=26, avg_age=0.07692307692307693, max_age=2 05-11-2018 00:17:08.391 +0000 INFO Metrics - group=per_source_thruput, series="/home/grana/splunk/var/log/splunk/splunkd_access.log", kbps=0.08719772691231611, eps=0.6774204911569529, kb=2.703125, ev=21, avg_age=2.4285714285714284, max_age=3
7	05-11-2018 16:01:05.390 +0000 INFO Metrics - group=executor, name=cachemgr_up, jobs_added=0, jobs_finished=0, current_size=0, smallest_size=0, largest_size=0, max_size=0 05-11-2018 14:30:09.390 +0000 INFO Metrics - group=executor, name=cachemgr_down, jobs_added=0, jobs_finished=0, current_size=0, smallest_size=0, largest_size=0, max_size=0 05-11-2018 06:31:12.390 +0000 INFO Metrics - group=executor, name=cachemgr_up, jobs_added=0, jobs_finished=0, current_size=0, smallest_size=0, largest_size=0, max_size=0

New Search

Save As Close

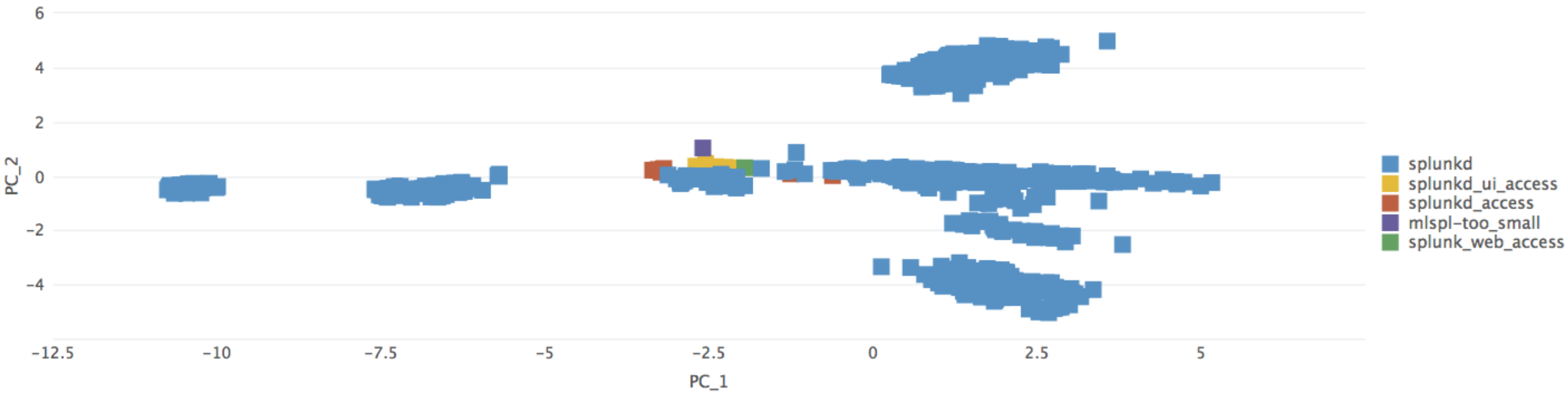
```
index=_internal
| sample 1000
| fit TFIDF _raw
| fit StandardScaler _raw_tfidf_*
| fit PCA k=2 SS_*
| table sourcetype PC_1 PC_2
```

Last 24 hours

✓ 1,000 events (5/10/18 5:00:00.000 PM to 5/11/18 5:08:05.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (1,000) Visualization

Scatter Chart Format Trellis



3D Scatter Plot at https://splunkbase.splunk.com/app/3138/

splunk> listen to your data

New Search

Save As Close

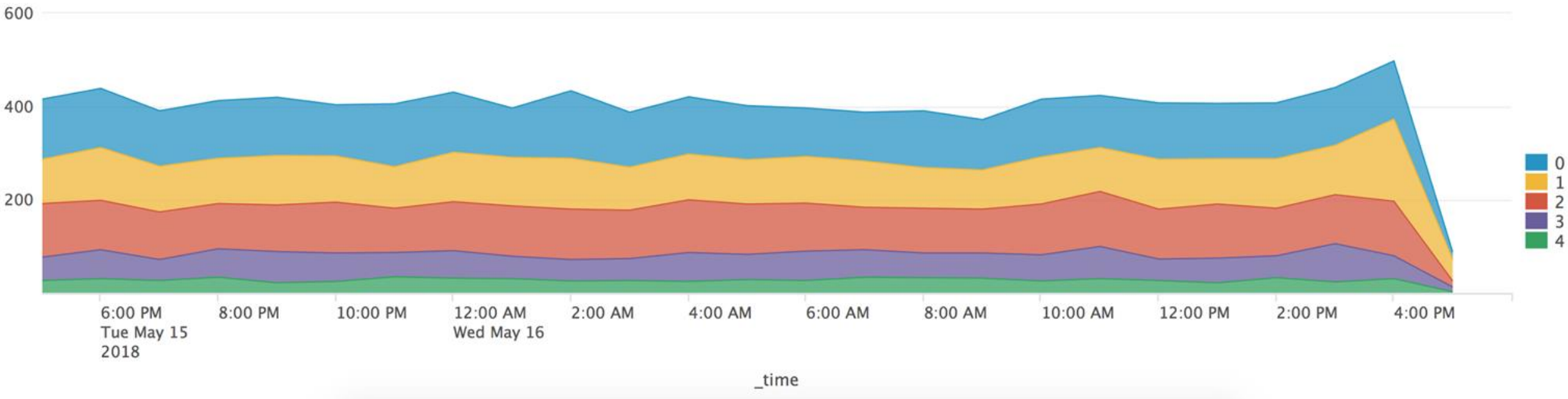
```
index=_internal
| sample 10000
| fit TFIDF token_pattern="\w\w\w+" _raw as raw_tfidf
| fit KMeans k=5 raw_tfidf_*
| timechart span=1h count by cluster
```

Last 24 hours

✓ 10,000 events (5/15/18 5:00:00.000 PM to 5/16/18 5:07:24.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (25) Visualization

Area Chart Format Trellis



splunk> listen to your data

Feature Engineering: Leading Indicators

… | reverse | streamstats … | reverse | …

- All features need to be in a single event before passing to fit
  - We can use streamstats to look backwards
  - But what if we want to predict the future?

```
index=application_log OR index=tickets
| timechart span=1d count(failure) as FAILS count("Change Request") as CHANGES
| reverse
| streamstats window=3 sum(FAILS) current=f as FAILS_NEXT_3DAYS
| reverse
| fit LinearRegression FAILS_NEXT_3DAYS from CHANGES into FAILS_PREDICTION_MODEL
```

Quantization

- Quantization (_time to hour, minute, etc.)

_time	date_month	date_mday	date_hour	date_minute
2018-07-23 08:04:04.145	july	23	7	4
2018-07-23 08:03:30.732	july	23	7	3
2018-07-23 08:02:56.494	july	23	7	2
2018-07-23 08:02:23.530	july	23	7	2
2018-07-23 08:02:04.977	july	23	7	2
2018-07-23 17:22:21.612	july	23	16	22
2018-07-23 17:07:19.575	july	23	16	7
2018-07-23 17:07:05.927	july	23	16	7
2018-07-23 16:24:03.719	july	23	15	24
2018-07-23 16:18:48.566	july	23	15	18
2018-07-23 15:29:59.962	july	23	14	29
2018-07-23 15:29:41.542	july	23	14	29
2018-07-23 15:16:15.206	july	23	14	16

splunk> .conf18

Data Splitting

don’t peek at the answers

splunk> .conf18

Data Splitting

- Training / Test
- Class imbalance
- Cross-validation

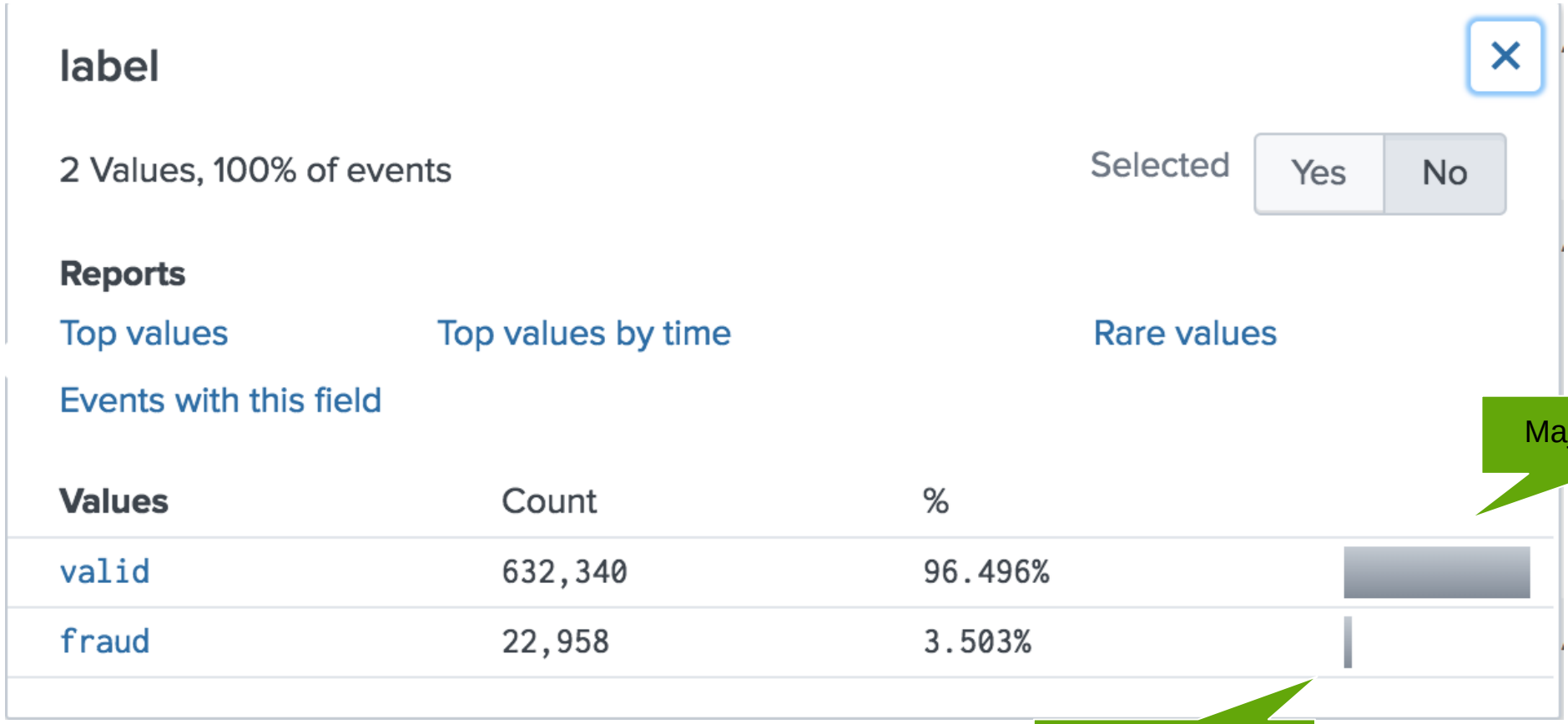
Class Imbalance

Definition

- In classification / anomaly detection, you have classes you are trying to predict, e.g.,
  - Valid / Fraud
  - Normal / Anomaly
  - Success / Failure
- Class imbalance refers to wildly different class frequencies
  - e.g., 99.999% valid; 0.001% fraud
- If we don't handle it correctly, our model will have trouble learning about the minority classes
  - e.g., no examples of the rare class appear in the training data

Class Imbalance

Detection



Class Imbalance

Stratified Sampling: Training Set

```
| search label="valid"
| sample partitions=100 seed=1001
| where partition_number <= 70
| outputlookup training_valid.csv
```

Training Proportion

```
| search label="fraud"
| sample partitions=100 seed=1001
| where partition_number <= 70
| outputlookup training_fraud.csv
```

_time	label	partition_number
2018-07-23 18:28:20.960	valid	5
2018-07-23 18:28:18.960	valid	9
2018-07-23 18:28:17.960	valid	40
2018-07-23 18:28:16.960	valid	37
2018-07-23 18:28:16.960	valid	64
2018-07-23 18:28:15.960	valid	40
2018-07-23 18:28:15.960	valid	57
2018-07-23 18:28:14.960	valid	60
2018-07-23 18:28:14.960	valid	5
2018-07-23 18:28:13.960	valid	70
2018-07-23 18:28:11.960	valid	65

Class Imbalance

Stratified Sampling: Test Set

```
| search label="valid"
| sample partitions=100 seed=1001
| where partition_number > 70
| outputlookup test_valid.csv
```

```
| search label="fraud"
| sample partitions=100 seed=1001
| where partition_number > 70
| outputlookup test_fraud.csv
```

_time	label	partition_number
2018-07-23 19:28:06.960	fraud	79
2018-07-23 19:28:05.960	fraud	76
2018-07-23 19:28:05.960	fraud	86
2018-07-23 19:28:04.960	fraud	90
2018-07-23 19:28:04.960	fraud	78
2018-07-23 19:28:02.960	fraud	73
2018-07-23 19:28:01.960	fraud	92
2018-07-23 19:28:01.960	fraud	71
2018-07-23 19:28:01.960	fraud	73
2018-07-23 19:28:00.960	fraud	89
2018-07-23 19:28:00.960	fraud	92

splunk> .conf18

Class Imbalance

Stratified Sampling: Merging the samples

- Merge training samples from majority & minority classes:

```
| inputlookup training_valid.csv | append [ inputlookup training_fraud.csv ]
```

- Merge test samples from majority & minority classes:

```
| inputlookup test_valid.csv | append [ inputlookup test_fraud.csv ]
```

Class Imbalance

Downsampling

Values	Count	%	
valid	632,340	96.496%	
fraud	22,958	3.503%	

Downsample to the smallest class:

```
| sample 22958 by label
```

Cross Validation

- 2-Fold:

```
index=demo sourcetype="wifi"
| head 10000
| eval label = if (n=-81, "valid", "fraud")
| sample partitions=2 seed=42
| appendpipe [ where partition_number=0 | fit LogisticRegression label from eventtype date_hour s state into validOrFraud ]
| where partition_number=1
| apply validOrFraud
```

- K-Fold (e.g., K = 5):

```
index=demo sourcetype="wifi"
| head 10000
| eval label = if (n=-81, "valid", "fraud")
| sample partitions=5 seed=42
| appendpipe [ where partition_number != 2 | fit LogisticRegression label from eventtype date_hour s state into validOrFraud ]
| where partition_number=2
| apply validOrFraud as second_pa
```

Train on partitions 0, 1, 3, 4

Infer on partition 2

Built-in K-fold Cross Validation

New in MLTK 4.0

- Specify the number of folds with kfold_cv=
- Validation metrics generated according to the type of model

```
| inputlookup iris.csv | fit LogisticRegression species from * kfold_cv=3
```

✓ 3 results (9/19/18 9:00:00.000 AM to 9/20/18 9:29:02.000 AM) No Event Sampling

Events | Patterns | Statistics (3) | Visualization

20 Per Page | Format | Preview

accuracy	f1_weighted	precision_weighted
0.960784313725	0.960648148148	0.96491228(
0.921568627451	0.921568627451	0.92156862
0.958333333333	0.958169934641	0.96296296

Summary

- Machine learning wants clean data (usually a matrix)
- Machine data begins life dirty
  - Relevant data scattered across sources
  - Missing values, incorrect values, wonky formats
  - Class imbalance
- Splunk Enterprise and MLTK provide many tools for preparing data for ML
  - Field extraction, joins, and lookups
  - Fit algorithms like StandardScaler and FieldSelector
  - Commands like eval, sample, and fillnull

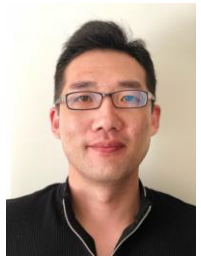
Want to know more?

Check out these upcoming ML sessions.

FN1364 - Using Spark and MLLib for Large Scale Machine Learning With Splunk Machine Learning Toolkit

(Thursday, Oct 04, 11:00 a.m. - 11:45 a.m.)

Lin Ma, Principal Software Engineer, Splunk
Fred Zhang, Principal Data Scientist, Splunk




FN1409 - Thank You for Sharing: Expanding Machine Learning using Splunk MLTK GitHub Collaboration

(Thursday, Oct 04, 11:00 a.m. - 11:45 a.m.)

Gyanendra Rana, Senior Product Manager, Splunk
Nathan Worsham, IS Security Administrator, Pinnacol Assurance


FN1478 - Exciting, To-Be-Announced Platform Session

Wednesday, Oct 03, 4:30 p.m. - 5:15 p.m.

Phillipp Drieger, Staff Machine Learning Architect , Splunk

FND1248 - Discriminatory Algorithms and Biased Data: Is the Future of Machine Learning Doomed?

Thursday, Oct 04, 1:30 p.m. - 2:15 p.m.

Sarah Moir, Program Manager, Splunk
Celeste Tretto, Data Scientist, Splunk


splunk> .conf18