Advanced Machine Learning in SPL with the Machine Learning Toolkit

Jacob Leverich

Software Engineer, Splunk

.conf2016

Disclaimer

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 the 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.

Who am I?

- Splunker for 2 years, based in San Francisco
- Engineering lead for…
  - ML Toolkit and Showcase App
  - ITSI Anomaly Detection and Adaptive Thresholding features
  - Splunk custom search command interface
- Initial author of fit/apply commands in ML Toolkit
- Die-hard Longhorns fan

Agenda

- Machine Learning + Splunk
- ML-SPL: Machine Learning in SPL
  - What it is
  - How it works
- Overview of Algorithms and Analytics available in ML-SPL
- Tips for Feature Engineering in SPL
- Wrap up

Machine Learning + Splunk

.conf2016

splunk>

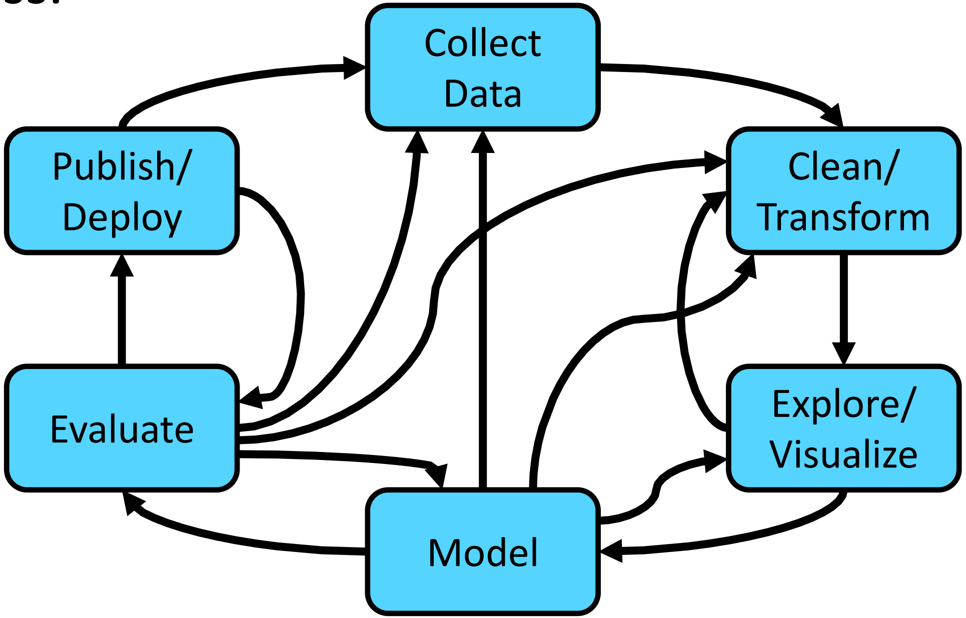
Machine Learning is Not Magic

- … it's a process.

- The process starts with a question:
  - How many requests do I expect in the next hour?
  - How likely is this hard drive to fail in the near future?
  - Am I being hacked?
    - Is it unexpected for Joe to login to the bastion host at 2am?

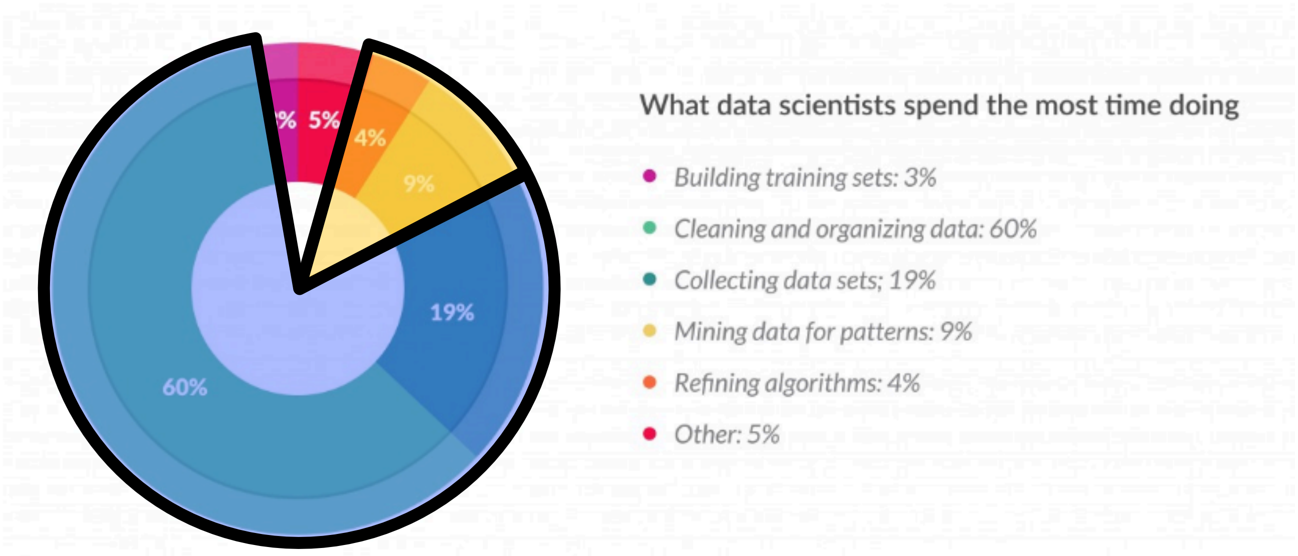
Machine Learning is Not Magic

- … it’s a process.

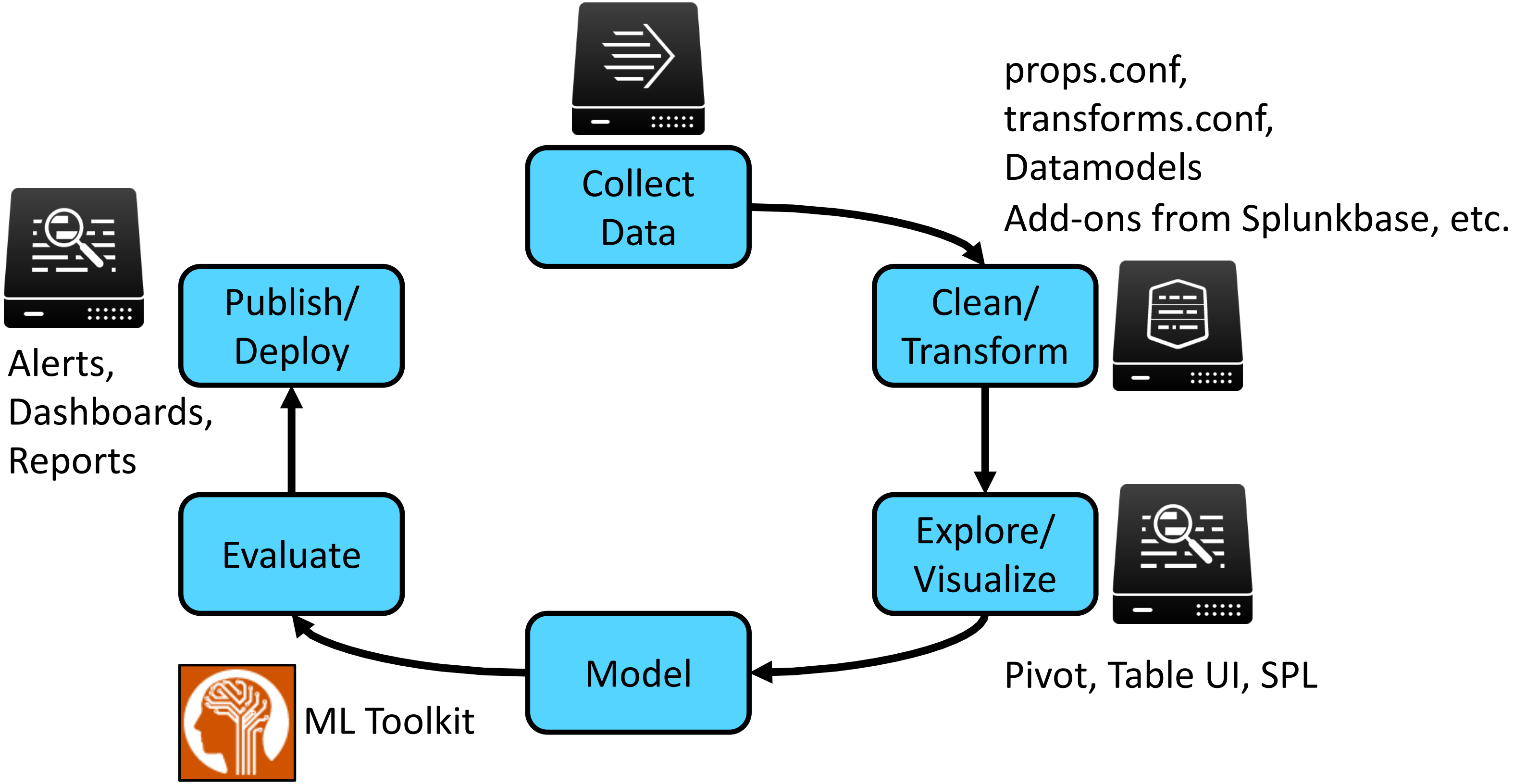


"Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says", Forbes Mar 23, 2016

***Data preparation** accounts for about 80% of the work of data scientists*



Splunk for Data Preparation



ML-SPL: Machine Learning in SPL

.conf2016

splunk>

ML-SPL: What is it?

- A suite of SPL search commands specifically for Machine Learning:
  - fit
  - apply
  - summary
  - listmodels
  - deletemodel
  - sample
- Implemented using modules from the Python for Scientific Computing add-on for Splunk:
  - scikit-learn, numpy, pandas, statsmodels, scipy

ML-SPL Commands: A "grammar" for ML

- Fit (i.e. train) a model from search results

```
… | fit <ALGORITHM> <TARGET> from <VARIABLES …>
          <PARAMETERS> into <MODEL>
```

- Apply a model to obtain predictions from (new) search results

```
… | apply <MODEL>
```

- Inspect the model built by <ALGORITHM> (e.g. display coefficients)

```
| summary <MODEL>
```

ML-SPL Commands: **fit**

optional

```
… | fit <ALGORITHM> <TARGET> from <VARIABLES …>
        <PARAMETERS> into <MODEL>
```

Examples:

```
… | fit LinearRegression
        system_temp from cpu_load fan_rpm
        into temp_model

… | fit KMeans k=10
        downloads purchases posts days_active visits_per_day
        into user_behavior_clusters

… | fit LinearRegression
        petal_length from species
```

Search | Splunk 6.4.1

127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20iris.csv%0A%7C%20fit%20Lin...

splunk> App: ML Toolkit and Show.. Administrator Messages Settings Activity Help Find

Search Showcase Assistants Docs ML Toolkit and Showcase

New Search

Save As Close

```
| inputlookup iris.csv
| fit LinearRegression petal_length from species
| table species petal_length predicted(petal_length)
```

All time

✓ 150 results (before 7/27/16 1:12:42.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (150) Visualization

Line Chart Format

species	petal_length	predicted(petal_length)
Iris Setosa	1.4	1.46606445312
Iris Setosa	1.4	1.46606445312
Iris Setosa	1.3	1.46606445312
Iris Setosa	1.5	1.46606445312
Iris Setosa	1.4	1.46606445312
Iris Setosa	1.7	1.46606445312
Iris Setosa	1.4	1.46606445312
Iris Setosa	1.5	1.46606445312

`fit`: How It Works

1. Discard fields that are null for all search results.
2. Discard non-numeric fields with >100 distinct values.
3. Discard search results with any null fields.
4. Convert non-numeric fields to binary indicator variables (i.e. “dummy coding”).
5. Convert to a numeric matrix and hand over to `<ALGORITHM>`.
6. Compute predictions for all search results.
7. Save the learned model.

`fit`: How It Works

```
… | fit LogisticRegression field_A from field_*
```

1. Discard fields that are null for all search results.

Target	Explanatory Variables…			
field_A	**field_B**	**field_C**	**field_D**	**field_E**
ok	41		red	172.24.16.5
ok	32		green	192.168.0.2
FRAUD	1		blue	10.6.6.6
ok	43			171.64.72.1
	2		blue	192.168.0.2

`fit`: How It Works

```
… | fit LogisticRegression field_A from field_*
```

2. Discard non-numeric fields with >100 distinct values.

Target	Explanatory Variables…		
field_A	**field_B**	**field_D**	**field_E**
ok	41	red	172.24.16.5
ok	32	green	192.168.0.2
FRAUD	1	blue	10.6.6.6
ok	43		171.64.72.1
	2	blue	192.168.0.2

`fit`: How It Works

```
… | fit LogisticRegression field_A from field_*
```

3. Discard search results with any null fields.

Target — Explanatory Variables…

field_A	field_B	field_D
ok	41	red
ok	32	green
FRAUD	1	blue
ok	43	
	2	blue

`fit`: How It Works

```
… | fit LogisticRegression field_A from field_*
```

4. Convert non-numeric fields to binary indicator variables.

Target — Explanatory Variables…

field_A	field_B	field_D=red	…=green	…=blue
ok	41	1	0	0
ok	32	0	1	0
FRAUD	1	0	0	1

`fit`: How It Works

```
… | fit LogisticRegression field_A from field_*
```

5. Convert to a numeric matrix and hand over to `<ALGORITHM>`.

y = [1, 1, 0]

X = [[41, 1, 0, 0],
[32, 0, 1, 0],
[1, 0, 0, 1]]

e.g. for Logistic Regression:

$$\hat{y} = \frac{1}{1 + e^{-(\theta^T x)}}$$

Find θ using maximum likelihood estimation.

Model inference generally delegated to scikit-learn and statsmodels.
(e.g. `sklearn.linear_model.LogisticRegression`)

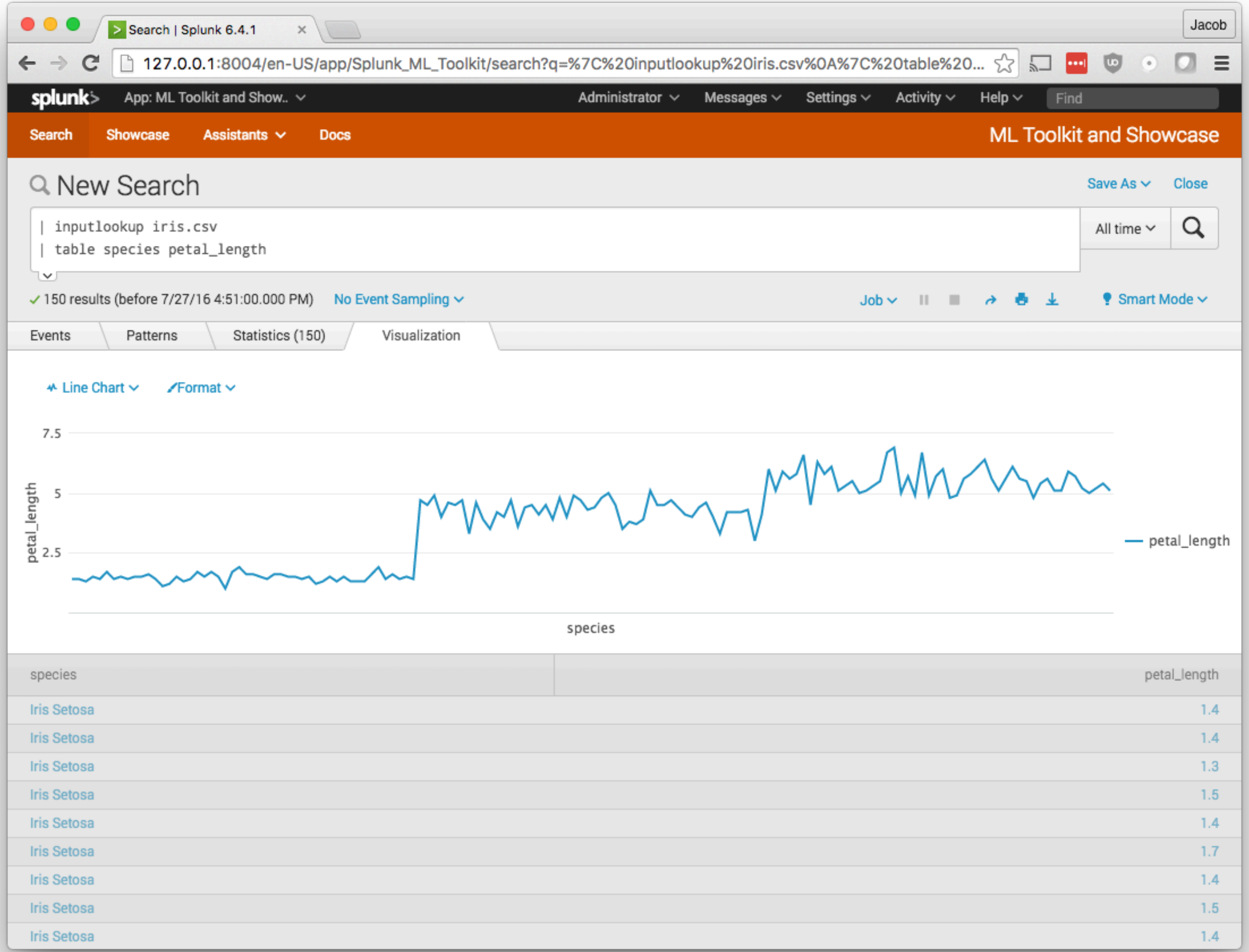
`fit`: How It Works

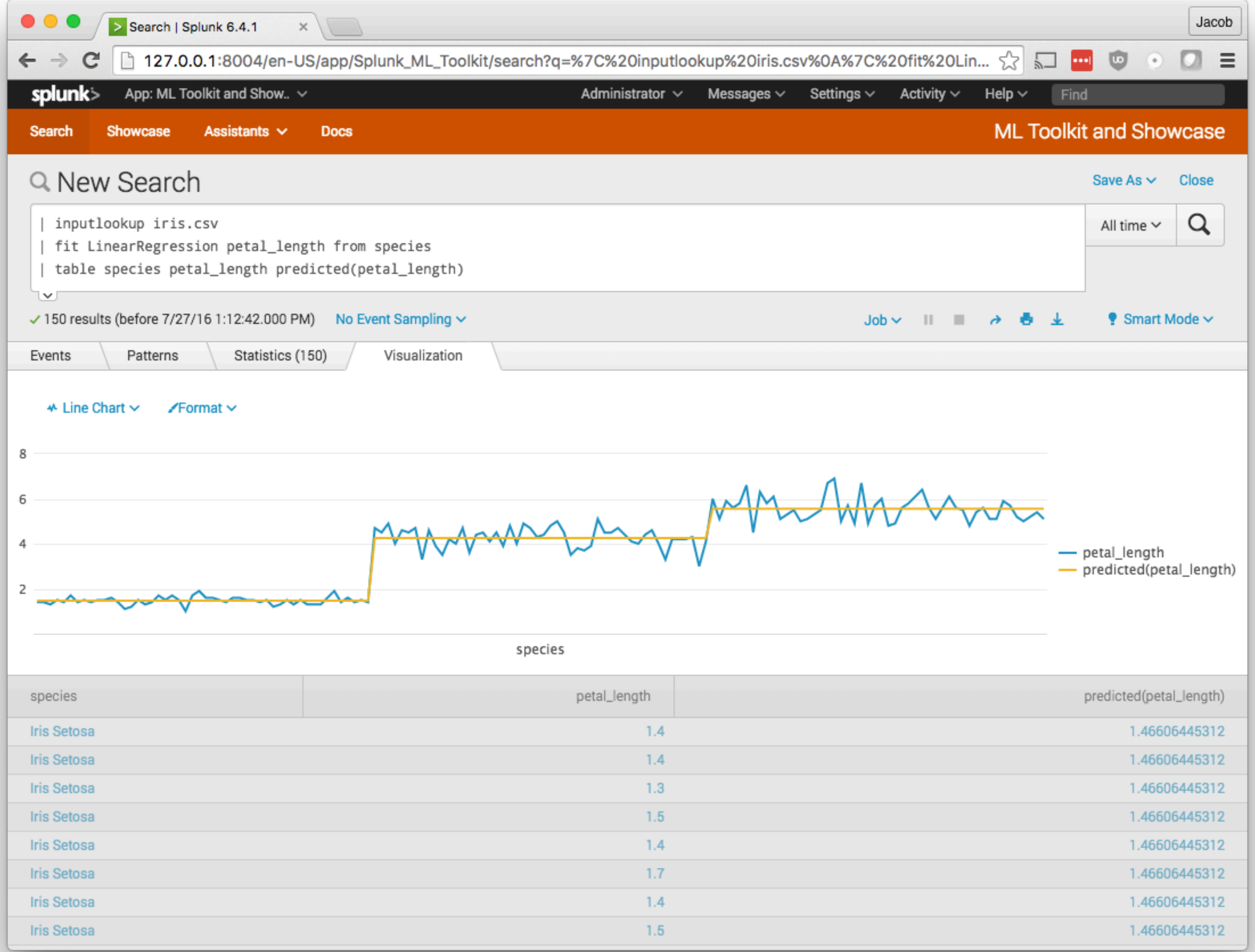
```
… | fit LogisticRegression field_A from field_*
```

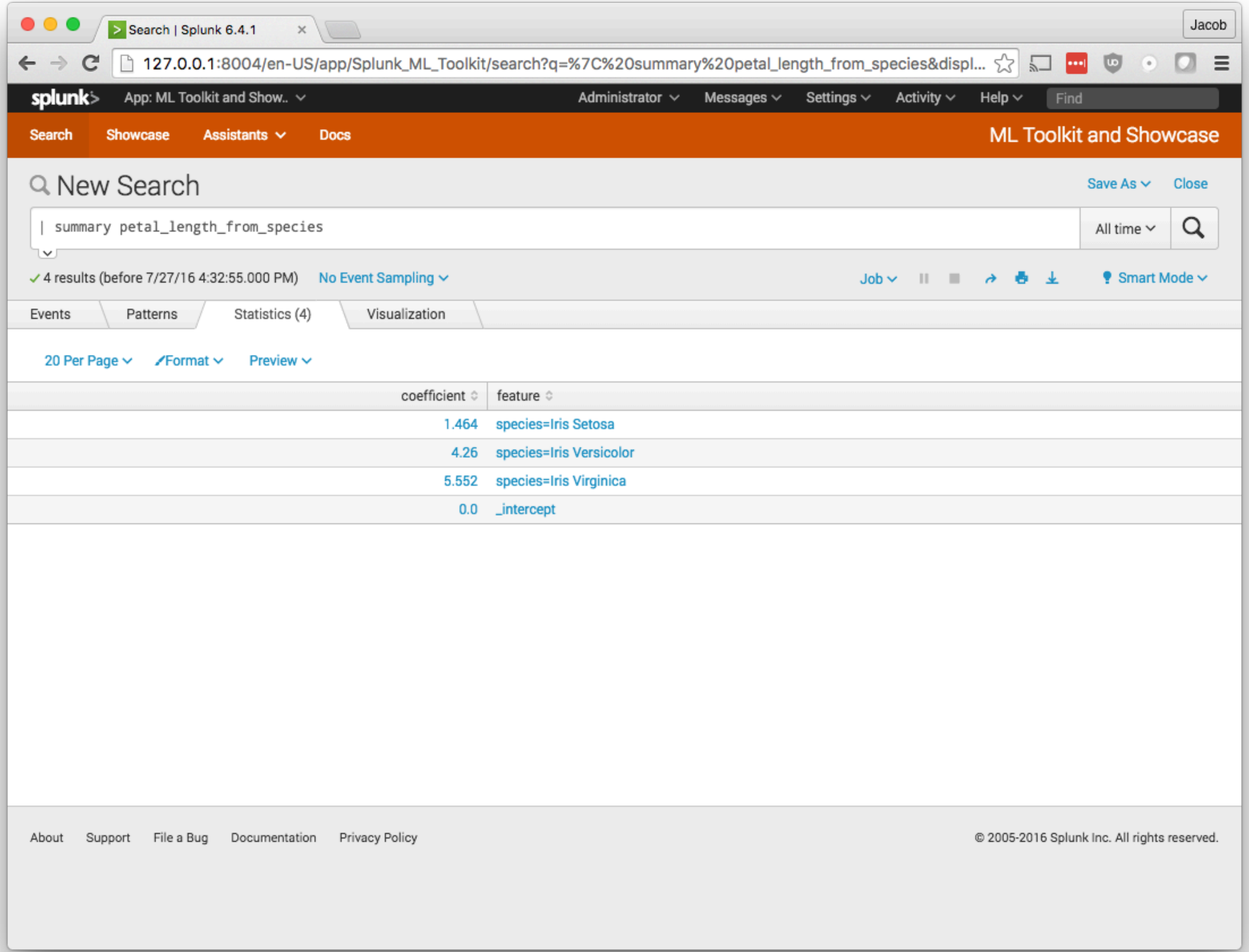
6. Compute predictions for all search results.

Target — Explanatory Variables… — Prediction

field_A	field_B	field_C	field_D	field_E	predicted(field_A)
ok	41		red	172.24.16.5	ok
ok	32		green	192.168.0.2	ok
FRAUD	1		blue	10.6.6.6	FRAUD
ok	43			171.64.72.1	ok
	2		blue	192.168.0.2	FRAUD


Search | Splunk 6.4.1
Jacob
127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20iris.csv%0A%7C%20table%20...
App: ML Toolkit and Show..
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Docs
ML Toolkit and Showcase
New Search
Save As
Close
| inputlookup iris.csv
| table species petal_length
All time
✓ 150 results (before 7/27/16 4:51:00.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (150)
Visualization
Line Chart
Format
7.5
5
2.5
petal_length
species
petal_length
species
petal_length
Iris Setosa 1.4
Iris Setosa 1.4
Iris Setosa 1.3
Iris Setosa 1.5
Iris Setosa 1.4
Iris Setosa 1.7
Iris Setosa 1.4
Iris Setosa 1.5
Iris Setosa 1.4


Search | Splunk 6.4.1
Jacob
127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20iris.csv%0A%7C%20fit%20Lin...
App: ML Toolkit and Show..
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Docs
ML Toolkit and Showcase
New Search
Save As
Close
| inputlookup iris.csv
| fit LinearRegression petal_length from species
| table species petal_length predicted(petal_length)
All time
150 results (before 7/27/16 1:12:42.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (150)
Visualization
Line Chart
Format
8
6
4
2
species
petal_length
predicted(petal_length)
species
petal_length
predicted(petal_length)
Iris Setosa 1.4 1.46606445312
Iris Setosa 1.4 1.46606445312
Iris Setosa 1.3 1.46606445312
Iris Setosa 1.5 1.46606445312
Iris Setosa 1.4 1.46606445312
Iris Setosa 1.7 1.46606445312
Iris Setosa 1.4 1.46606445312
Iris Setosa 1.5 1.46606445312


Search | Splunk 6.4.1
Jacob
127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20summary%20petal_length_from_species&displ...
App: ML Toolkit and Show..
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Docs
ML Toolkit and Showcase
New Search
Save As
Close
| summary petal_length_from_species
All time
4 results (before 7/27/16 4:32:55.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (4)
Visualization
20 Per Page
Format
Preview
coefficient
feature
1.464
species=Iris Setosa
4.26
species=Iris Versicolor
5.552
species=Iris Virginica
0.0
_intercept
About
Support
File a Bug
Documentation
Privacy Policy
© 2005-2016 Splunk Inc. All rights reserved.

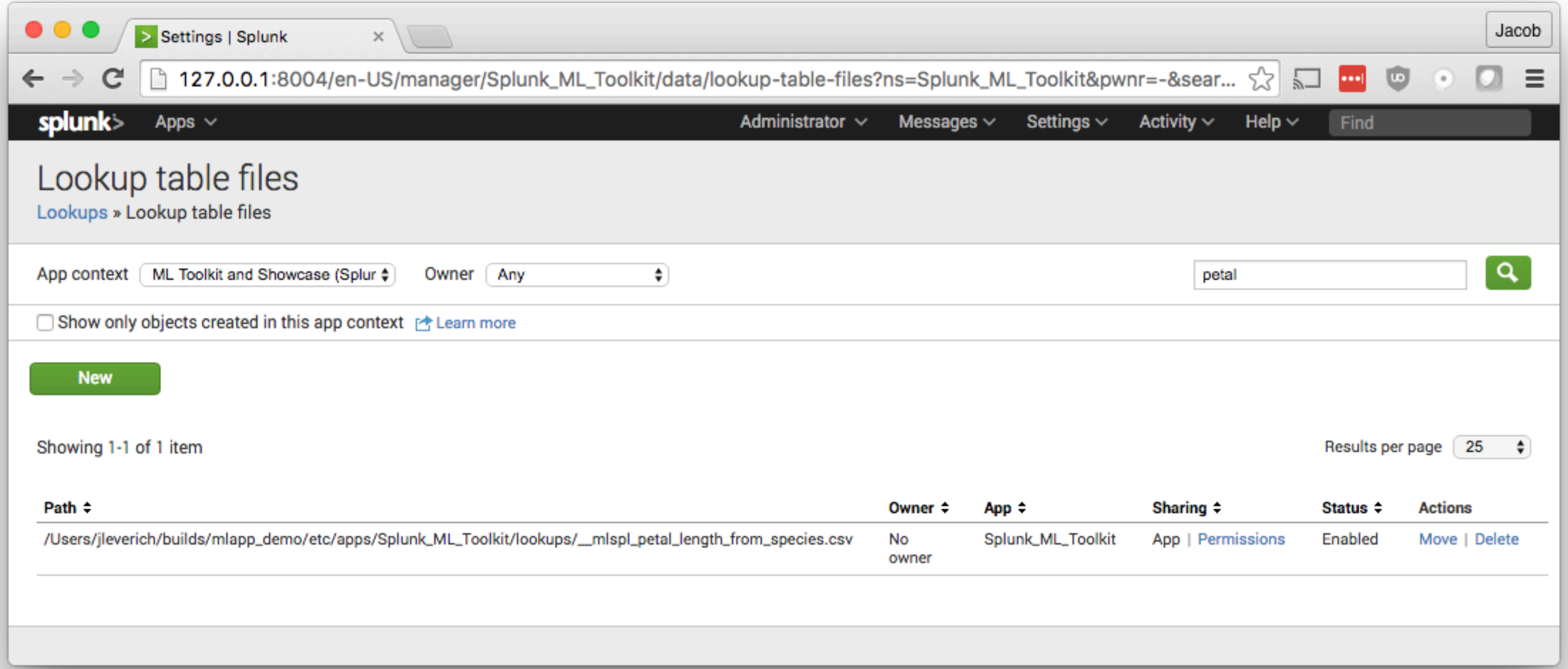
`fit`: How It Works

```
… | fit LogisticRegression field_A from field_* into logreg_model
```

7. Save the learned model.

Serialize model settings, coefficients, etc. into a Splunk lookup table.

- Replicated amongst members of Search Head Cluster.
- Automatically distributed to Indexers with search bundle.


Settings | Splunk
Jacob
127.0.0.1:8004/en-US/manager/Splunk_ML_Toolkit/data/lookup-table-files?ns=Splunk_ML_Toolkit&pwnr=-&sear...
Apps
Administrator
Messages
Settings
Activity
Help
Find
Lookup table files
Lookups » Lookup table files
App context
ML Toolkit and Showcase (Splur
Owner
Any
petal
Show only objects created in this app context
Learn more
New
Showing 1-1 of 1 item
Results per page
25
Path
Owner
App
Sharing
Status
Actions
/Users/jleverich/builds/mlapp_demo/etc/apps/Splunk_ML_Toolkit/lookups/__mlspl_petal_length_from_species.csv
No owner
Splunk_ML_Toolkit
App | Permissions
Enabled
Move | Delete

`fit`: Properties

- Each event is an "example" for the learning algorithm.
- Resilient to missing values. *(but be careful!)*
- Automatically handles categorical (e.g. non-numeric) fields.
- ***SAVES ITS WORK***:
  - Learned model can be applied to ***new, unseen*** data with the **`apply`** command.

`fit`: Scalability

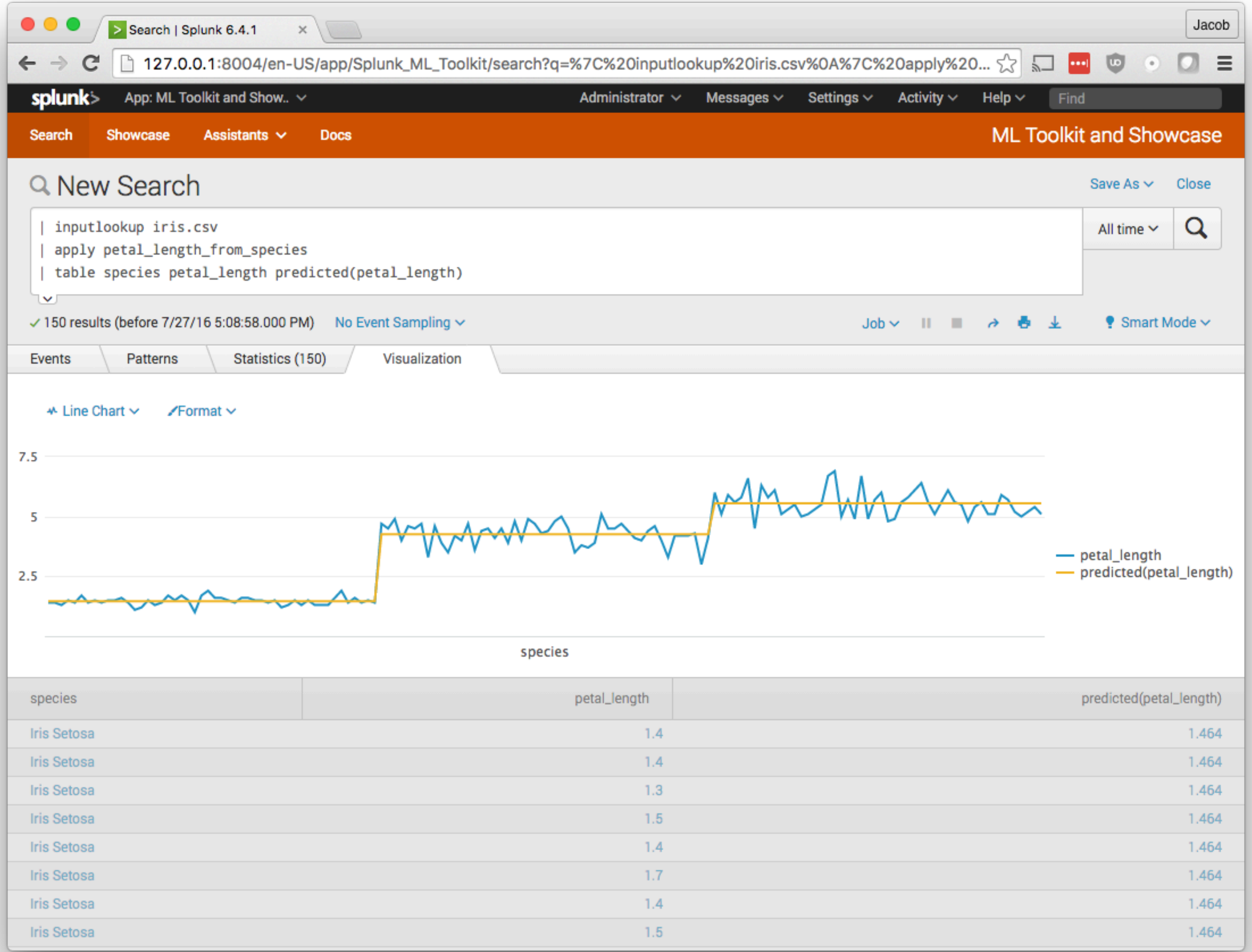
- Some algorithms are inherently ***not scalable***.
  - e.g. Kernel-based Support Vector Machines is $O(N^3)$
- Input is sampled using ***reservoir sampling***.
  - Per-algorithm sample reservoir size, typically 100,000 events
  - Configurable in **`mlspl.conf`**.
- Some algorithms support ***incremental fitting***, e.g.: SGDRegressor, SGDClassifier, NaiveBayes
  - Use "`partial_fit=t`" option with **`fit`** command.
  - No sampling, no event limit!
- For the most part, you don't need to care.

ML-SPL Commands: **apply**

```
… | apply <MODEL>
```

Examples:

```
… | apply temp_model
… | apply user_behavior_clusters
… | apply petal_length_from_species
```


Search | Splunk 6.4.1
127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20iris.csv%0A%7C%20apply%20...
App: ML Toolkit and Show..
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Docs
ML Toolkit and Showcase
New Search
Save As
Close
| inputlookup iris.csv
| apply petal_length_from_species
| table species petal_length predicted(petal_length)
All time
150 results (before 7/27/16 5:08:58.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (150)
Visualization
Line Chart
Format
7.5
5
2.5
species
petal_length
predicted(petal_length)
species
petal_length
predicted(petal_length)
Iris Setosa
1.4
1.464
Iris Setosa
1.4
1.464
Iris Setosa
1.3
1.464
Iris Setosa
1.5
1.464
Iris Setosa
1.4
1.464
Iris Setosa
1.7
1.464
Iris Setosa
1.4
1.464
Iris Setosa
1.5
1.464

`apply`: How It Works

1. Load the learned model.
2. Discard fields that are null for all search results.
3. Discard non-numeric fields with >100 distinct values.
4. Convert non-numeric fields to binary indicator variables (i.e. “dummy coding”).
5. Discard variables not in the learned model.
6. Fill missing fields with 0’s.
7. Convert to a numeric matrix and hand over to `<ALGORITHM>`.
8. Compute predictions for all search results.

`apply`: How It Works

```
… | apply fraud_model
```

4. Convert non-numeric fields to binary indicator variables.

Target / Explanatory Variables…

field_A	field_B	field_D=red	…=green	…=blue	…=yellow
ok	41	1	0	0	0
ok	32	0	1	0	0
FRAUD	1	0	0	1	0
	41	0	0	0	1

`apply`: How It Works

```
… | apply fraud_model
```

5. Discard variables not in the learned model.

Target	Explanatory Variables…				
field_A	**field_B**	**field_D=red**	**…=green**	**…=blue**	**…=yellow**
ok	41	1	0	0	0
ok	32	0	1	0	0
FRAUD	1	0	0	1	0
	41	0	0	0	1

`apply`: How It Works

```
… | apply fraud_model
```

5. Convert to a numeric matrix and hand over to `<ALGORITHM>`.

y = [1, 1, 0, 1, ?]

X = [[41, 1, 0, 0],
[32, 0, 1, 0],
[1, 0, 0, 1],
[41, 0, 0, 0]]

e.g. for Logistic Regression:

$$\hat{y} = \frac{1}{1 + e^{-(\theta^T x)}}$$

Compute $\hat{y}$ using θ found by **`fit`** command.

`apply`: How It Works

```
… | apply fraud_model
```

7. Compute predictions for all search results.

Target	Explanatory Variables…				Prediction
field_A	**field_B**	**field_C**	**field_D**	**field_E**	**predicted(field_A)**
ok	41		red	172.24.16.5	ok
ok	32		green	192.168.0.2	ok
FRAUD	1		blue	10.6.6.6	FRAUD
ok	43			171.64.72.1	ok
	41		yellow	192.168.0.2	ok

`apply`: Properties

- Learned models can be applied to ***new, unseen*** data.

  `| fit` is to `| apply`

  as

  `| outputlookup` is to `| lookup`

- Resilient to missing values. *(but, again, be careful!)*

- Automatically handles categorical (e.g. non-numeric) fields.

`apply`: Scalability

- No limits.

- When possible, executes at the Indexing tier.
  - Fully parallelized; harness the CPU power of your Indexing Cluster.
  - Must set "**`streaming_apply = true`**" in **`mlspl.conf`**.

ML-SPL Commands: **summary**

```
… | summary <MODEL>
```

Examples:

```
… | summary temp_model
… | summary user_behavior_clusters
… | summary petal_length_from_species
```

Search | Splunk 6.4.1

127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20iris.csv%0A%7C%20fit%20Lo...

splunk> App: ML Toolkit and Show.. Administrator Messages Settings Activity Help Find

Search Showcase Assistants Docs **ML Toolkit and Showcase**

New Search

Save As Close

```
| inputlookup iris.csv
| fit LogisticRegression species from petal_length petal_width sepal_length sepal_width into species_model
| sample 15
```

All time

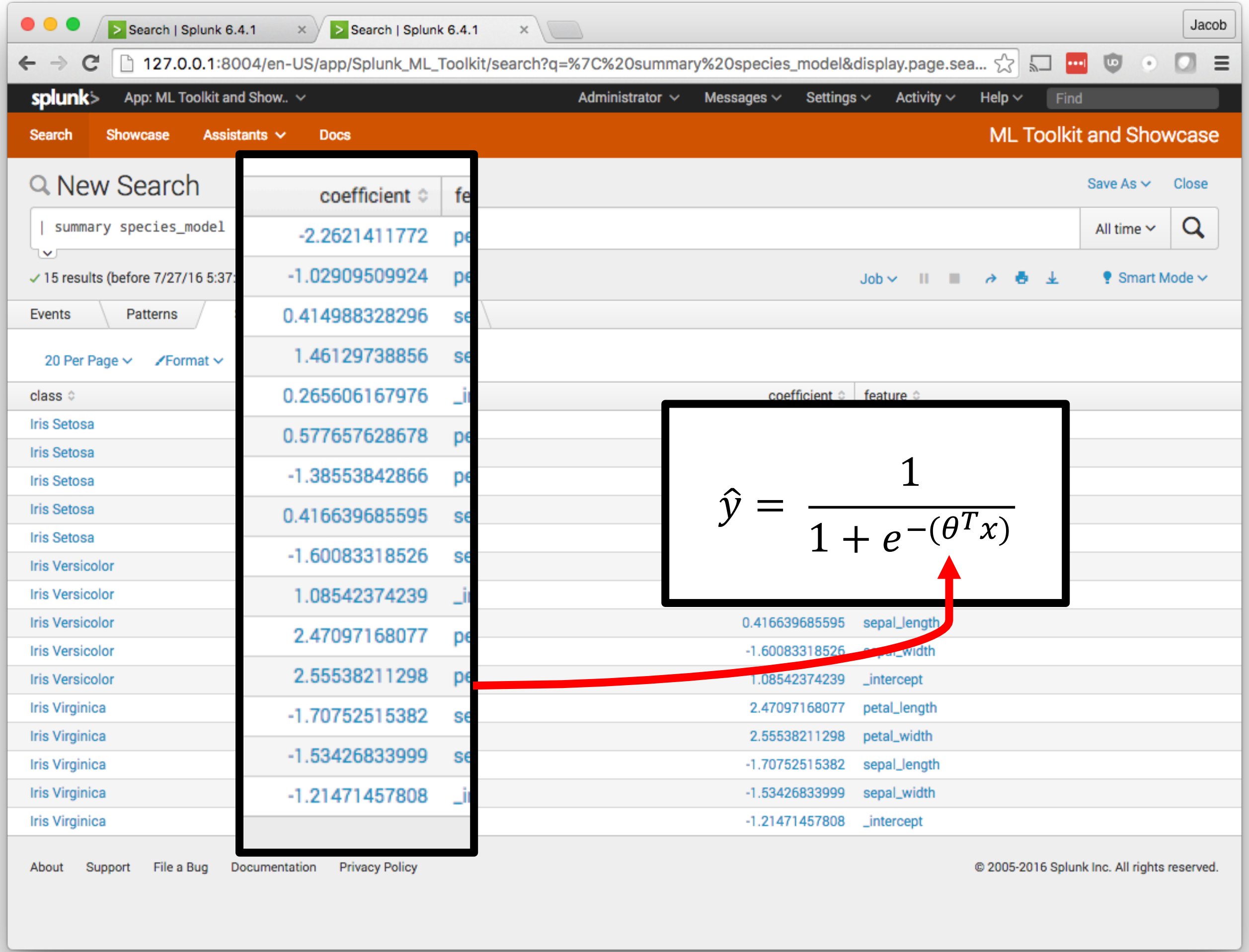
✓ 15 results (before 7/27/16 5:34:12.000 PM) No Event Sampling Job Smart Mode

Events Patterns Statistics (15) Visualization

20 Per Page Format Preview

petal_length	petal_width	predicted(species)	sepal_length	sepal_width	species
1.3	0.2	Iris Setosa	4.7	3.2	Iris Setosa
1.5	0.3	Iris Setosa	5.1	3.8	Iris Setosa
1.2	0.2	Iris Setosa	5.0	3.2	Iris Setosa
1.4	0.3	Iris Setosa	4.8	3.0	Iris Setosa
1.5	0.2	Iris Setosa	5.3	3.7	Iris Setosa
4.7	1.4	Iris Versicolor	6.1	2.9	Iris Versicolor
3.6	1.3	Iris Versicolor	5.6	2.9	Iris Versicolor
3.7	1.0	Iris Versicolor	5.5	2.4	Iris Versicolor
5.5	1.8	Iris Virginica	6.5	3.0	Iris Virginica
5.0	1.5	Iris Virginica	6.0	2.2	Iris Virginica
5.7	2.1	Iris Virginica	6.7	3.3	Iris Virginica
5.6	1.4	Iris Virginica	6.1	2.6	Iris Virginica
5.4	2.1	Iris Virginica	6.9	3.1	Iris Virginica
5.9	2.3	Iris Virginica	6.8	3.2	Iris Virginica
5.2	2.0	Iris Virginica	6.5	3.0	Iris Virginica

About Support File a Bug Documentation Privacy Policy


Search | Splunk 6.4.1
Search | Splunk 6.4.1
Jacob
127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20summary%20species_model&display.page.sea...
App: ML Toolkit and Show..
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Docs
ML Toolkit and Showcase
New Search
Save As
Close
| summary species_model
All time
15 results (before 7/27/16 5:37
Job
Smart Mode
Events
Patterns
20 Per Page
Format
class
coefficient
feature
Iris Setosa
Iris Versicolor
Iris Virginica
-2.2621411772
-1.02909509924
0.414988328296
1.46129738856
0.265606167976
0.577657628678
-1.38553842866
0.416639685595
-1.60083318526
1.08542374239
2.47097168077
2.55538211298
-1.70752515382
-1.53426833999
-1.21471457808
0.416639685595 sepal_length
-1.60083318526 sepal_width
1.08542374239 _intercept
2.47097168077 petal_length
2.55538211298 petal_width
-1.70752515382 sepal_length
-1.53426833999 sepal_width
-1.21471457808 _intercept
$\hat{y} = \frac{1}{1 + e^{-(\theta^T x)}}$
About
Support
File a Bug
Documentation
Privacy Policy
© 2005-2016 Splunk Inc. All rights reserved.

Algorithms and Analytics in ML-SPL

Regression Algorithms
(e.g. predict numeric fields)

- `LinearRegression`
  - … including Lasso, Ridge, ElasticNet
- `KernelRidge`
- `DecisionTreeRegressor`
- `RandomForestRegressor`
- `SGDRegressor`

- All implemented with sklearn models.

Classification Algorithms
(e.g. predict categorical fields)

- `LogisticRegression`
- `DecisionTreeClassifier`
- `RandomForestClassifier`
- `SGDClassifier`
- `SVM`
- Naïve Bayes
  - Including `BernoulliNB` and `GuassianNB`

Clustering Algorithms
(e.g. group like with like)

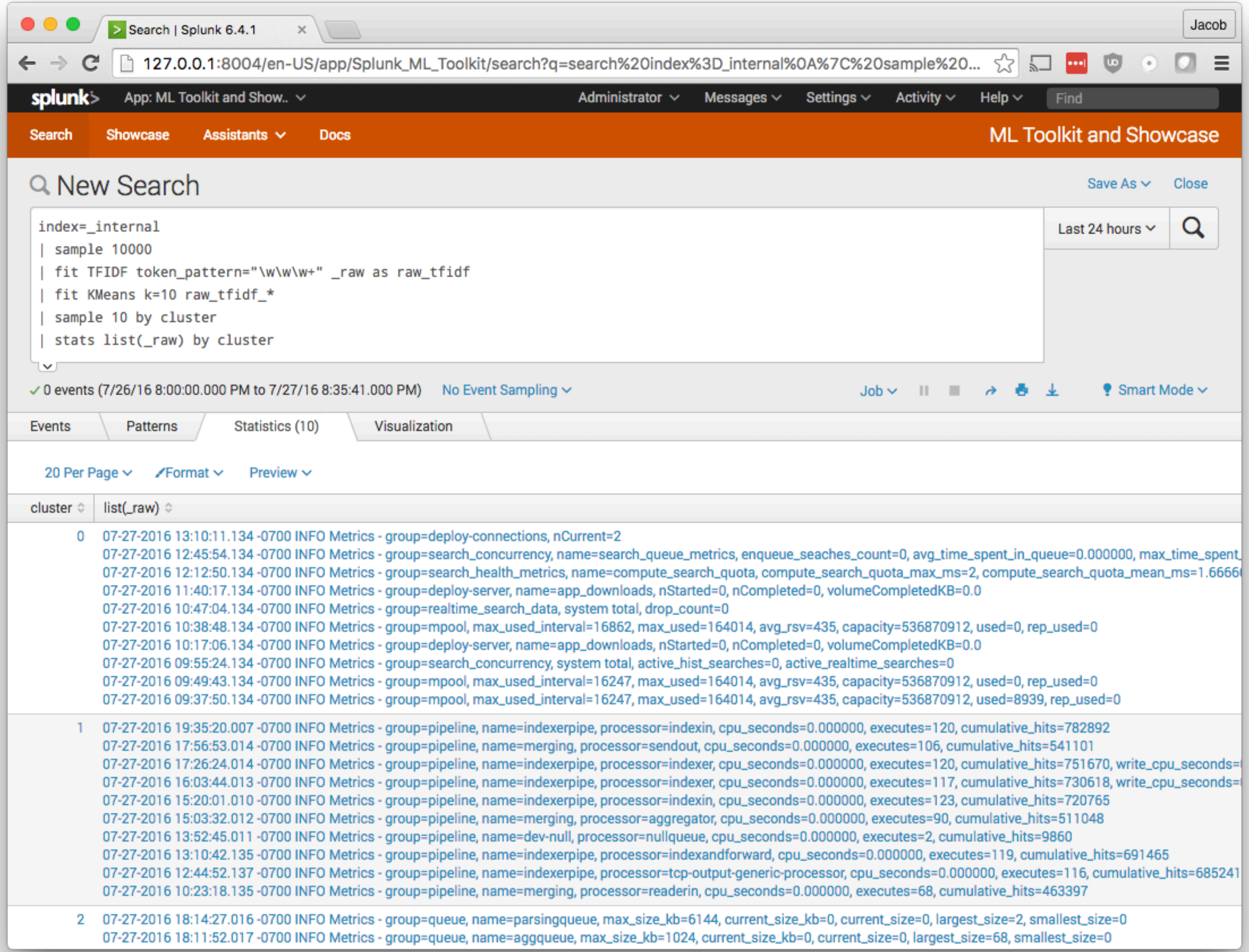
- `KMeans`
- `DBSCAN`
- `Birch`
- `SpectralClustering`

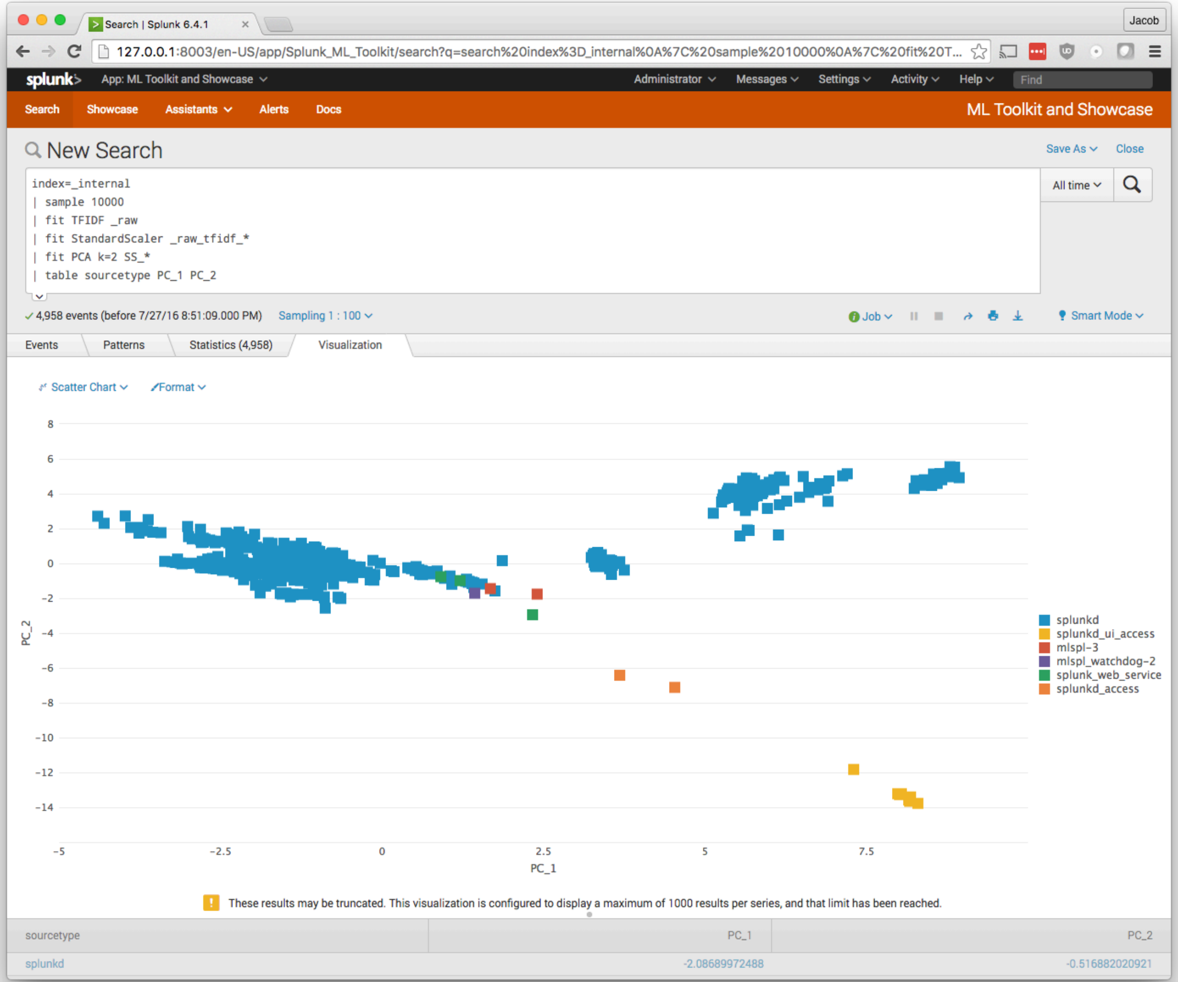
Feature Engineering Algorithms (e.g. data pre-processing)

- `TFIDF` (term-frequency x inverse document-frequency)
  - Transform free-form text into numeric fields
- `StandardScaler` (i.e. normalization)
- `FieldSelector` (i.e. choose K best features for regression/classification)
- PCA and `KernelPCA`

“Pipeline” Multiple Algorithms

- Example: Text Analytics
  - `TFIDF` to transform free-form messages into numeric fields, followed by…
    - `KMeans` to group similar messages
    - `BernoulliNB` to classify messages (e.g. according to sentiment)
    - `PCA` to visualize distribution of messages
  - **`… | fit TFIDF message | fit Kmeans message_tfidf_* | ...`**
- Analogous to Pipeline concept from sklearn or Spark MLLib


Search | Splunk 6.4.1
Jacob
127.0.0.1:8004/en-US/app/Splunk_ML_Toolkit/search?q=search%20index%3D_internal%0A%7C%20sample%20...
App: ML Toolkit and Show..
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Docs
ML Toolkit and Showcase
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 10 by cluster
| stats list(_raw) by cluster
Last 24 hours
✓ 0 events (7/26/16 8:00:00.000 PM to 7/27/16 8:35:41.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (10)
Visualization
20 Per Page
Format
Preview
cluster
list(_raw)
0
07-27-2016 13:10:11.134 -0700 INFO Metrics - group=deploy-connections, nCurrent=2
07-27-2016 12:45:54.134 -0700 INFO Metrics - group=search_concurrency, name=search_queue_metrics, enqueue_seaches_count=0, avg_time_spent_in_queue=0.000000, max_time_spent_
07-27-2016 12:12:50.134 -0700 INFO Metrics - group=search_health_metrics, name=compute_search_quota, compute_search_quota_max_ms=2, compute_search_quota_mean_ms=1.6666
07-27-2016 11:40:17.134 -0700 INFO Metrics - group=deploy-server, name=app_downloads, nStarted=0, nCompleted=0, volumeCompletedKB=0.0
07-27-2016 10:47:04.134 -0700 INFO Metrics - group=realtime_search_data, system total, drop_count=0
07-27-2016 10:38:48.134 -0700 INFO Metrics - group=mpool, max_used_interval=16862, max_used=164014, avg_rsv=435, capacity=536870912, used=0, rep_used=0
07-27-2016 10:17:06.134 -0700 INFO Metrics - group=deploy-server, name=app_downloads, nStarted=0, nCompleted=0, volumeCompletedKB=0.0
07-27-2016 09:55:24.134 -0700 INFO Metrics - group=search_concurrency, system total, active_hist_searches=0, active_realtime_searches=0
07-27-2016 09:49:43.134 -0700 INFO Metrics - group=mpool, max_used_interval=16247, max_used=164014, avg_rsv=435, capacity=536870912, used=0, rep_used=0
07-27-2016 09:37:50.134 -0700 INFO Metrics - group=mpool, max_used_interval=16247, max_used=164014, avg_rsv=435, capacity=536870912, used=8939, rep_used=0
1
07-27-2016 19:35:20.007 -0700 INFO Metrics - group=pipeline, name=indexerpipe, processor=indexin, cpu_seconds=0.000000, executes=120, cumulative_hits=782892
07-27-2016 17:56:53.014 -0700 INFO Metrics - group=pipeline, name=merging, processor=sendout, cpu_seconds=0.000000, executes=106, cumulative_hits=541101
07-27-2016 17:26:24.014 -0700 INFO Metrics - group=pipeline, name=indexerpipe, processor=indexer, cpu_seconds=0.000000, executes=120, cumulative_hits=751670, write_cpu_seconds=
07-27-2016 16:03:44.013 -0700 INFO Metrics - group=pipeline, name=indexerpipe, processor=indexer, cpu_seconds=0.000000, executes=117, cumulative_hits=730618, write_cpu_seconds=
07-27-2016 15:20:01.010 -0700 INFO Metrics - group=pipeline, name=indexerpipe, processor=indexin, cpu_seconds=0.000000, executes=123, cumulative_hits=720765
07-27-2016 15:03:32.012 -0700 INFO Metrics - group=pipeline, name=merging, processor=aggregator, cpu_seconds=0.000000, executes=90, cumulative_hits=511048
07-27-2016 13:52:45.011 -0700 INFO Metrics - group=pipeline, name=dev-null, processor=nullqueue, cpu_seconds=0.000000, executes=2, cumulative_hits=9860
07-27-2016 13:10:42.135 -0700 INFO Metrics - group=pipeline, name=indexerpipe, processor=indexandforward, cpu_seconds=0.000000, executes=119, cumulative_hits=691465
07-27-2016 12:44:52.137 -0700 INFO Metrics - group=pipeline, name=indexerpipe, processor=tcp-output-generic-processor, cpu_seconds=0.000000, executes=116, cumulative_hits=685241
07-27-2016 10:23:18.135 -0700 INFO Metrics - group=pipeline, name=merging, processor=readerin, cpu_seconds=0.000000, executes=68, cumulative_hits=463397
2
07-27-2016 18:14:27.016 -0700 INFO Metrics - group=queue, name=parsingqueue, max_size_kb=6144, current_size_kb=0, current_size=0, largest_size=2, smallest_size=0
07-27-2016 18:11:52.017 -0700 INFO Metrics - group=queue, name=aggqueue, max_size_kb=1024, current_size_kb=0, current_size=0, largest_size=68, smallest_size=0


Search | Splunk 6.4.1
127.0.0.1:8003/en-US/app/Splunk_ML_Toolkit/search?q=search%20index%3D_internal%0A%7C%20sample%2010000%0A%7C%20fit%20T...
App: ML Toolkit and Showcase
Administrator Messages Settings Activity Help
Search Showcase Assistants Alerts Docs
ML Toolkit and Showcase
New Search
Save As Close
index=_internal
| sample 10000
| fit TFIDF _raw
| fit StandardScaler _raw_tfidf_*
| fit PCA k=2 SS_*
| table sourcetype PC_1 PC_2
All time
✓4,958 events (before 7/27/16 8:51:09.000 PM) Sampling 1 : 100
Job
Smart Mode
Events Patterns Statistics (4,958) Visualization
Scatter Chart
Format
PC_2
PC_1
splunkd
splunkd_ui_access
mlspl-3
mlspl_watchdog-2
splunk_web_service
splunkd_access
These results may be truncated. This visualization is configured to display a maximum of 1000 results per series, and that limit has been reached.
sourcetype PC_1 PC_2
splunkd -2.08689972488 -0.516882020921

“Pipeline” Multiple Algorithms

- ML-SPL analytics are ***stackable***.
- Very advanced ML use-cases are succinctly expressible.

Tips for Feature Engineering

.conf2016

splunk>

Tips for Feature Engineering

- Work on aggregates, not raw events.
  - DO NOT use fit on 1,000,000,000 events. DO use stats.
- Use eval to compute new features.
- Use streamstats to construct leading indicators.
- …

Work on aggregates, not raw events

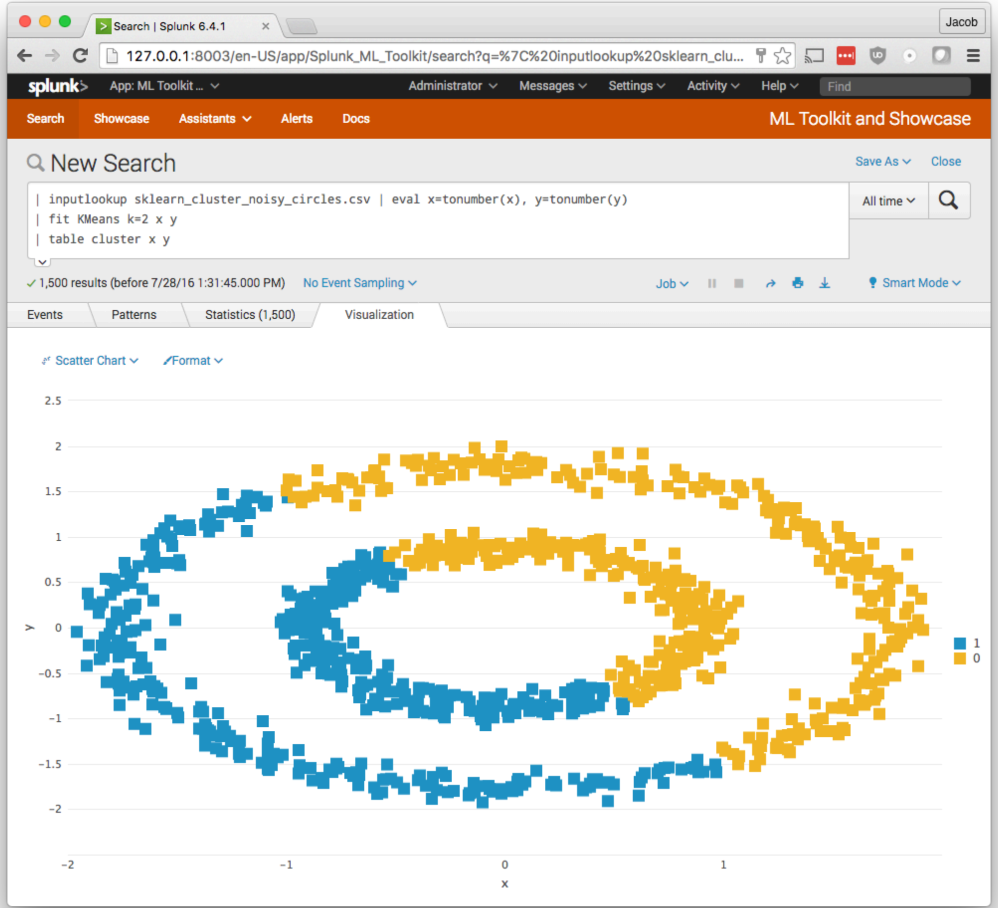
```
… | fit KMeans k=10
              downloads purchases posts days_active visits_per_day
              into user_behavior_clusters
```

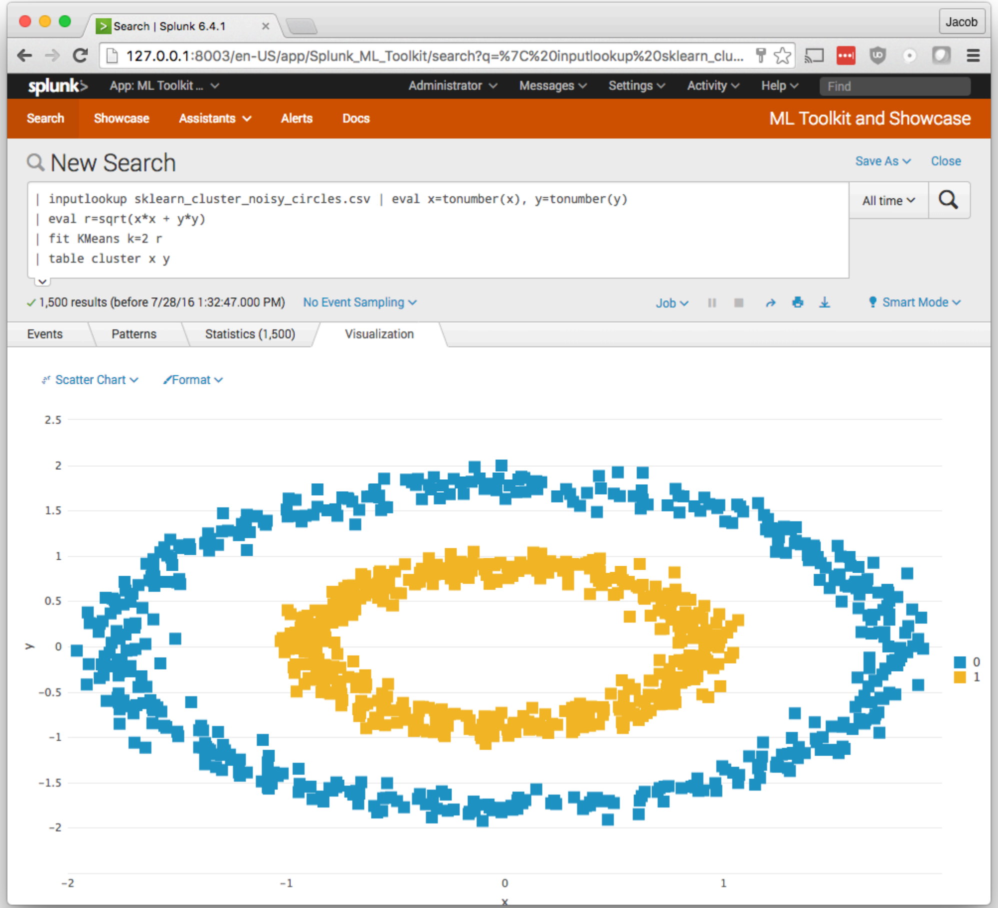
- Use **`stats`** and lookup tables to construct features:

```
index=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 …
```

Use **`eval`** to compute new features

- Coerce numbers into categories by prepending a string:
  - `… | eval region_id = “Region ” + region_id | …`
- Model interactions between features:
  - `… | eval X_factor = importance * urgency | …`
  - Use + for categorical fields, * for numeric
- Make non-linear features out of numeric values:
  - `… | eval temperature = pow(temperature,2) | …`
  - `… | eval latency = log(latency) | …`


Search | Splunk 6.4.1
Jacob
127.0.0.1:8003/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20sklearn_clu...
App: ML Toolkit ...
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Alerts
Docs
ML Toolkit and Showcase
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
All time
✓ 1,500 results (before 7/28/16 1:31:45.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (1,500)
Visualization
Scatter Chart
Format
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
y
-2
-1
0
1
x
1
0


Search | Splunk 6.4.1
Jacob
127.0.0.1:8003/en-US/app/Splunk_ML_Toolkit/search?q=%7C%20inputlookup%20sklearn_clu...
splunk>
App: ML Toolkit ...
Administrator
Messages
Settings
Activity
Help
Find
Search
Showcase
Assistants
Alerts
Docs
ML Toolkit and Showcase
New Search
Save As
Close
| inputlookup sklearn_cluster_noisy_circles.csv | eval x=tonumber(x), y=tonumber(y)
| eval r=sqrt(x*x + y*y)
| fit KMeans k=2 r
| table cluster x y
All time
✓ 1,500 results (before 7/28/16 1:32:47.000 PM)
No Event Sampling
Job
Smart Mode
Events
Patterns
Statistics (1,500)
Visualization
Scatter Chart
Format
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y
−2
−1
0
1
x
0
1

Use **streamstats** for leading indicators

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

Wrap-up

What did we cover?

- Machine Learning + Splunk
- ML-SPL: Machine Learning in SPL
  – What it is
  – How it works
- Overview of Algorithms and Analytics available in ML-SPL
- Tips for Feature Engineering in SPL

What Now?

- Install the ML Toolkit from Splunkbase!
  - http://tiny.cc/splunkmlapp
- Don’t miss Manish Sainani’s or Adam Oliner’s talks!

- Product Manager: Manish Sainani <msainani@splunk.com>
- Field Expert: Andrew Stein <astein@splunk.com>
- Me: Jacob Leverich <jleverich@splunk.com>

THANK YOU

.conf2016

splunk>

`fit`: Misc. details

- Multi-class classification problems typically modeled as "one-vs-rest"
- Some algorithms do NOT support saved models, e.g.:
  - DBSCAN and SpectralClustering

ML-SPL Commands

- `fit <ALGORITHM> <TARGET> from <VARIABLES …> <PARAMETERS> into <MODEL>`
  - Fit (i.e. train) a model from search results
- `apply <MODEL>`
  - Apply a model to obtain predictions from (new) search results
- `summary <MODEL>`
  - Inspect the model inferred by <ALGORITHM> (e.g., display coefficients)

Slide Title