User Experience Modeling with the Splunk Machine Learning Toolkit

Dr. Kenneth Tupper
Performance Engineering Lead | Paychex

splunk> .conf19

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans 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 may differ materially. 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, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

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 functionalities described or to include any such feature or functionality in a future release.

Can User Experience Data be Combined with Client Demographic Data to Build Predictive Models of User Experience?

Predictive models of user experience can be used to gain insight into areas of an application that are the best candidates for performance improvements

Agenda

1) **Paychex Inc.**

2) **Collecting User Experience Data at Paychex**

3) **Types of Client Demographic Data for Modeling User Experience**

4) **MLTK Examples**
   - What client demographics most affect pay run processing times
   - Predict pay run time as a function of important client demographics
     - Predict pay run tasks most susceptible to client demographics
   - Detect outliers in pay run processing times for further analysis
   - Predict pay run processing time as a function of date
   - Capacity Analysis: Predict memory as a function of date
   - Clustering to identify unique clients

About Paychex

Payroll – HR – Retirement – Insurance

Paychex, Inc. (NASDAQ:PAYX) is a leading provider of integrated human capital management solutions for payroll, benefits, human resources, and insurance services. By combining its innovative software-as-a-service technology and mobility platform with dedicated, personal service, Paychex empowers small- and medium-sized business owners to focus on the growth and management of their business. Backed by more than 45 years of industry expertise, Paychex serves approximately 670,000 payroll clients as of May 31, 2019 across more than 100 locations in the U.S. and Europe, and pays one out of every 12 American private sector employees. Learn more about Paychex by visiting paychex.com and stay connected on Twitter and LinkedIn.

splunk> .conf19

Market Leadership

Payroll

We'll take care of payroll while you take care of your business.

- Top 2 in both small and mid-market
- ~12M people paid
- Pay 1 in 12 American private sector workers

HR

We're here to guide you through every HR challenge.

- 1.5M worksite employees
- 52K clients
- Inc.com Best HR Outsourcing for SMBs

Benefits

Create a benefits package and leave the management to us.

- #1 recordkeeper in the U.S.*
- 87K clients served
- 1M participants

* By number of plans; listing by *PLANSPONSOR* magazine

Insurance

Protect what you've worked hard to build.

- 20th largest insurance agency in the U.S.**
- $2.4B in premiums paid
- 125K clients

** Listing by *Business Insurance* magazine

Collecting User Experience Data

Traceability/Marks and Measures

splunk> .conf19

Collecting User Experience Data

Solution to collect and correlate data

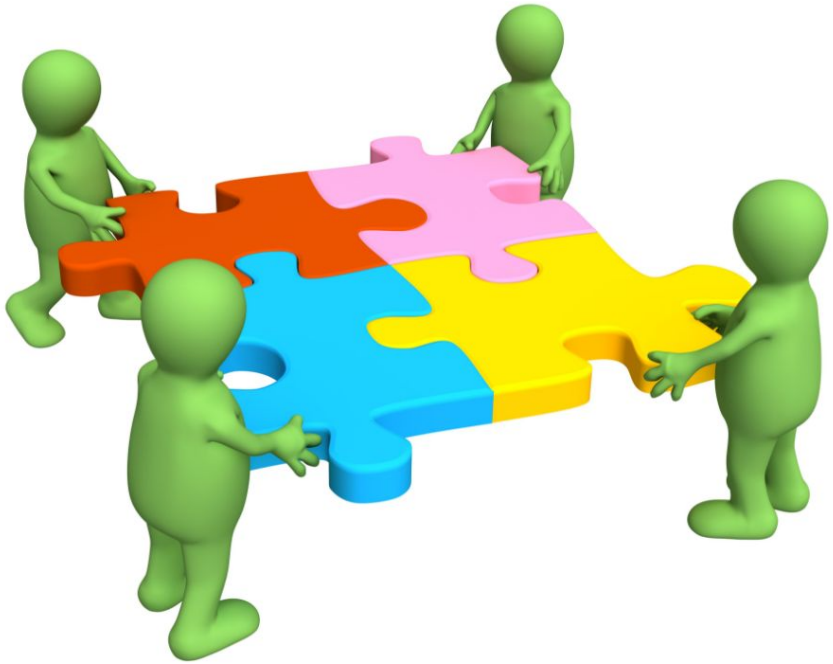
Traceability

- A standard for transaction logging in critical applications
  - Enables us to stitch together events
  - Improve problem resolution and MTTR
- Implemented across application and technology stacks

Marks and Measures

- 4 primary targets for marks and measures
  - Web application tier
  - Business service tier
  - Database tier
  - Network appliances

Routers, proxies, etc…



Without Splunk, correlating this data would take days

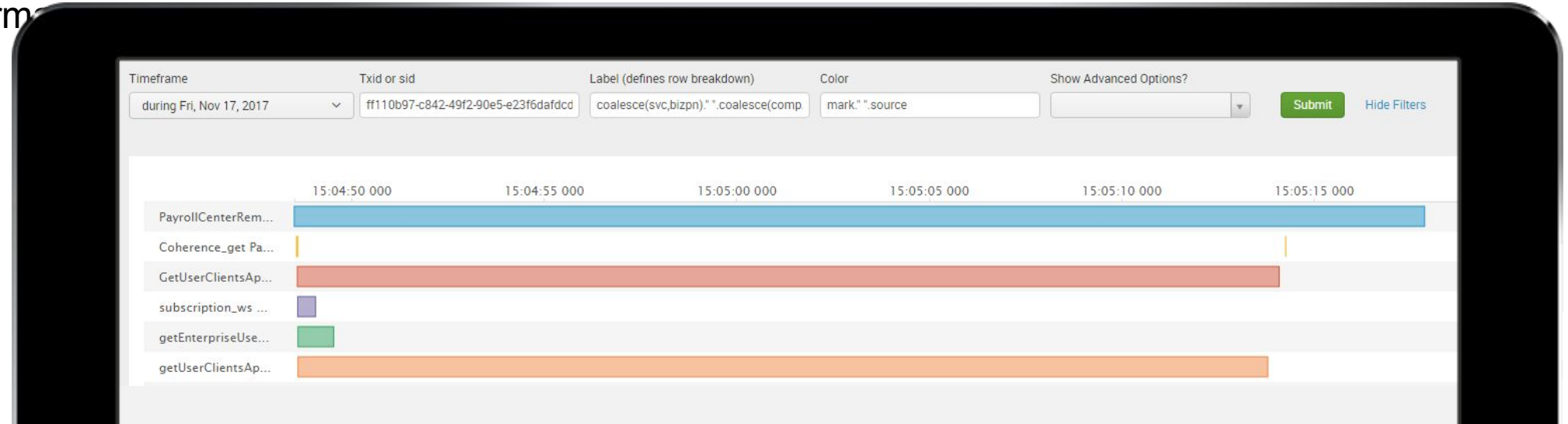
User Experience Reporting

How the dashboards work

The transaction ID can be used to look further into the application stack to see where time is spent at each layer of the application

For pay run processing we can see the transaction durations from the UI tier down to the individual SQL queries

Comparing transaction times for releases gives us a view into how code changes affect application perform



Building Predictive Models of User Experience

The Splunk Machine Learning Toolkit

splunk> .conf19

Client demographics

What types of data is available to build predictive models?

Non-temporal client demographics:

- Number of active employees
- Total employees in calendar year
- Company-level taxes (number of state/local)
- Pay components (holiday, overtime, etc.)
- Other (Organization units/Locations/Position)

Temporal client demographics:

- Total number of yearly pay runs (weekly, biweekly, etc.)
- Total number of YTD checks

What client demographics most affect pay run processing times?

Using the Field Selector to identify important demographics

```
| inputlookup payRunTimes.csv
```

```
| fit StandardScaler *
with_mean=true with_std=true
```

```
| fit FieldSelector "SS_duration"
from * type=numeric mode=k_best
param=2
```

Lookup tables used to improve performance

Standard Scalar is used to normalize the data

Field Selector is used over PCA to identify the individual demographics that most affect pay run times

Mode=k_best with param=2 will give the top two demographics that can be used to correlate the data

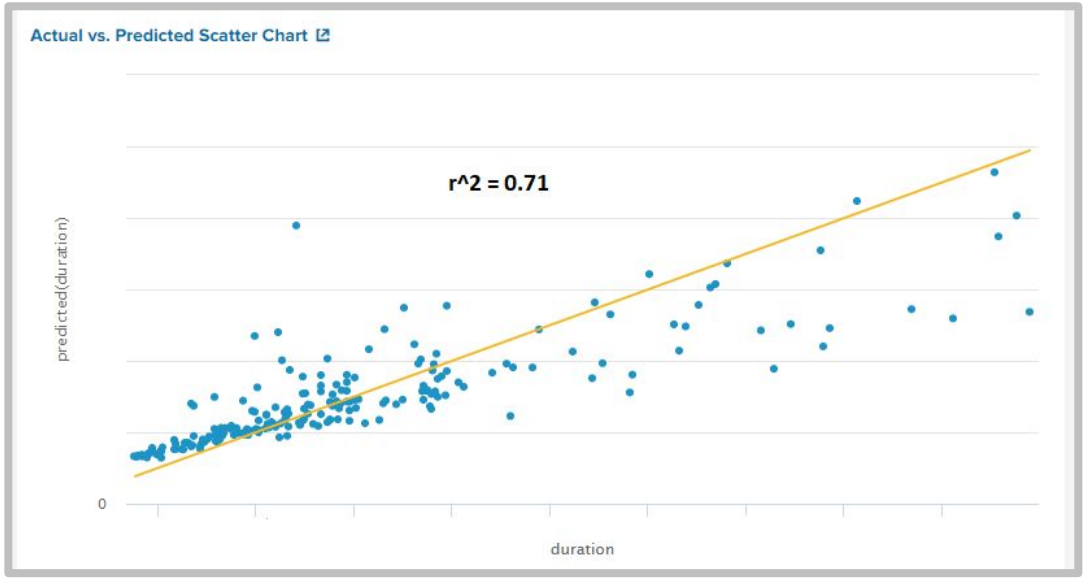
Predict pay run time as a function of important demographics

Linear regression of pay run times versus important demographics

```
| inputlookup payRunTimes.csv
```

```
| fit StandardScaler *
with_mean=true with_std=true
```

```
| fit LinearRegression
fit_intercept=true "SS_duration"
from "SS_Dem1" "SS_Dem2" into
"model_payRunTime"
```



Linear regression shows the demographics under-predict pay run times at high durations

What pay run tasks are most affected by client demographics

Using the MLTK to find subtasks most impacted by demographics

Same query as previous slide is runs for each pay run task

Tasks with a higher slope will be more impacted by an increase of the demographic in question

This information can be used to direct development efforts for application performance improvements

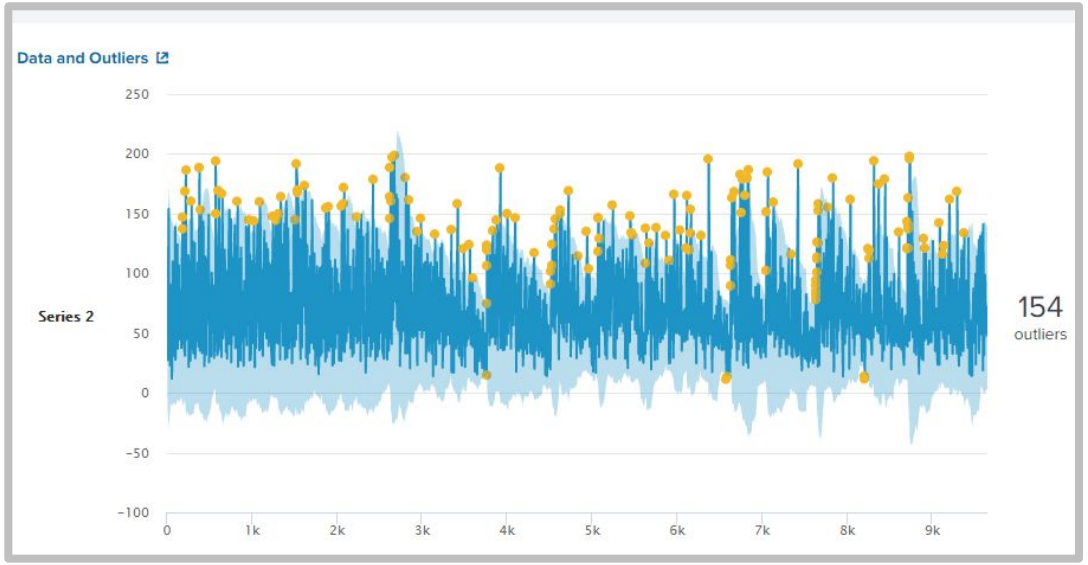
How did you get the linear fit on your chart?

```
| chart values(duration1) as d1 values(duration2) as d2 by xValue
| eval X2= xValue* xValue | eval Y2=d2*d2 | eval XY= xValue*d2
| table xValue d1 d2 X2 Y2 XY | sort xValue
| eventstats count as numevents sum(xValue) as sumX sum(d2) as sumY   sum(XY)
as sumXY  sum(X2) as sumX2 sum(Y2) as sumY2
| eval slope=((numevents*sumXY)-(sumX*sumY))/((numevents*sumX2)-   (sumX*sumX))
| eval yintercept= (sumY-(slope*sumX))/numevents
| eval linearFit=(yintercept + (slope* xValue))
| eval R=((numevents*sumXY) - (sumX*sumY))/sqrt(((numevents*sumX2)-
(sumX*sumX))*((numevents*sumY2)-(sumY*sumY)))
| eval R2=R*R

| chart values(d1) As Task1 values(d2) As Task2 values(linearFit) As "Linear
Fit" by xValue
```

Detect outliers in pay run processing times

Collecting information to further improve predictive models

```
| inputlookup payRunTimes.csv

| eventstats avg("duration") as avg
stdev("duration") as stdev

| eval
lowerBound=(avg-stdev*exact(2)),
upperBound=(avg+stdev*exact(2))

| eval isOutlier=if('duration' <
lowerBound   OR 'duration' >
upperBound, 1, 0)
```

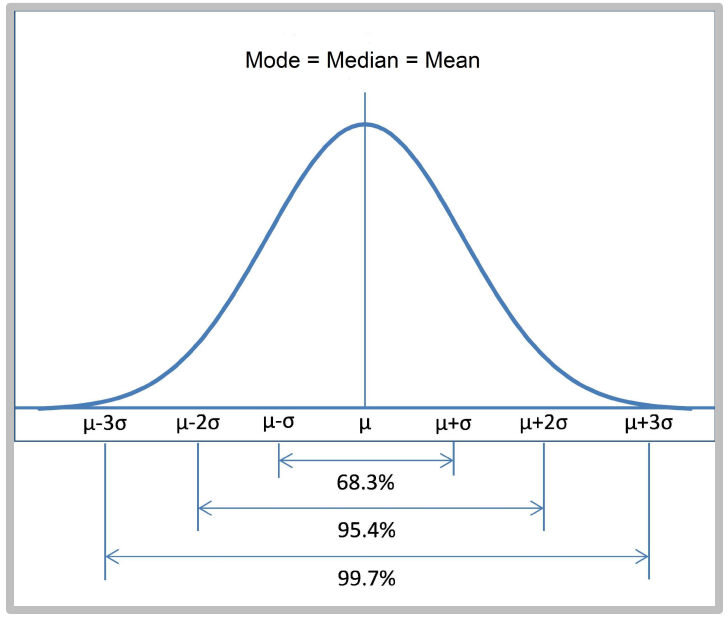


The above analysis assume that your data can be approximated by a Gaussian (Normal) distribution

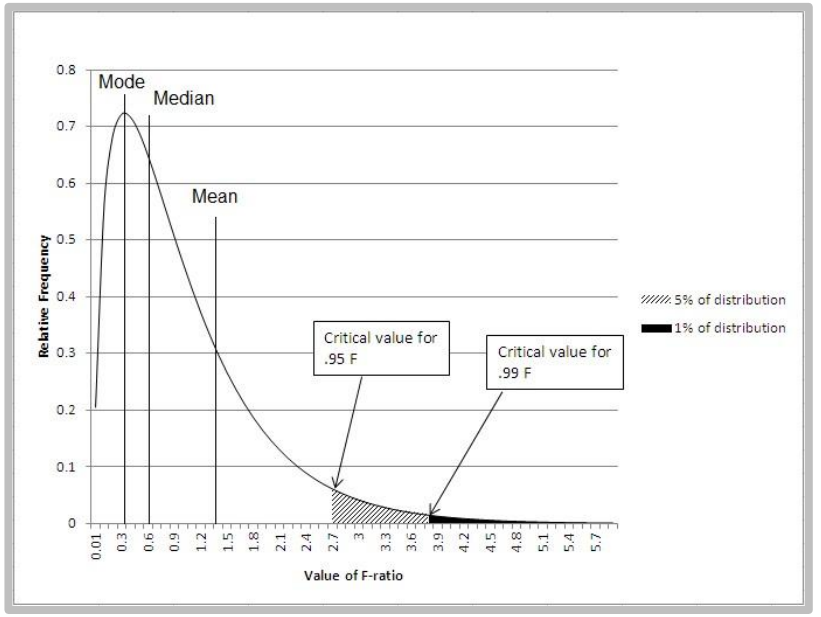
Working with non-Gaussian Data

Most performance-related data is not a normal distribution

Normal Distribution:



F Distribution:



To work with non-Gaussian distributions, you can use

| fit DensityFunction y by “x”

Pay run processing times as a function of date

Working with temporal data

More data exists for Year-to-Date calculations as more pay runs are executed

- Database partitioning helped to improve YTD calculations by only including data for the current calendar year in queries

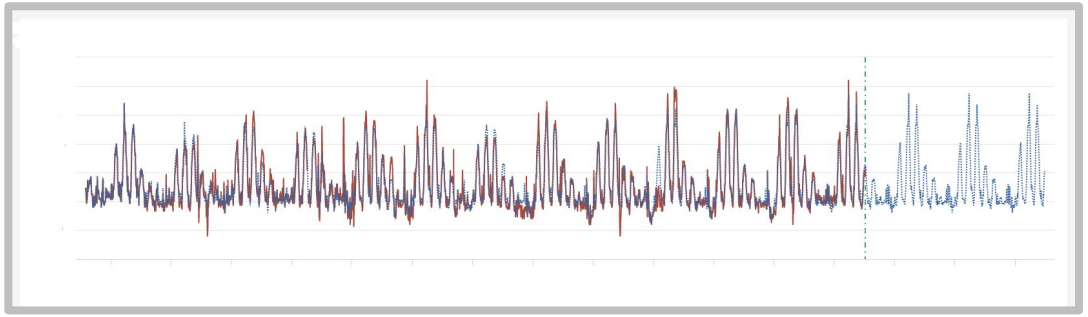
Can perform the same type of linear fit for pay run tasks as a function of date to see which tasks are most impacted by accumulated data

- Identified two reports where SQL tuning resulted in performance improvements for certain calculations

Capacity Analysis: Predict memory usage as a function of date

Can the effect of YTD calculations on memory usage be predicted

```
| inputlookup java_mem.csv

| eval _time=strptime(_time,
"%Y-%m-%d %H:%M:%S")

| timechart span=1h avg(p95Mem)
as aMem

| predict "aMem" as prediction
algorithm=LLP holdback=0
future_timespan=500 upper0=upper0
lower0=lower0

| `forecastviz(500, 0, "aMem",
0)`
```



As of version 4.3 can use a State Space Forecasting (Smart Forecasting) to look at periodic (temporal) data

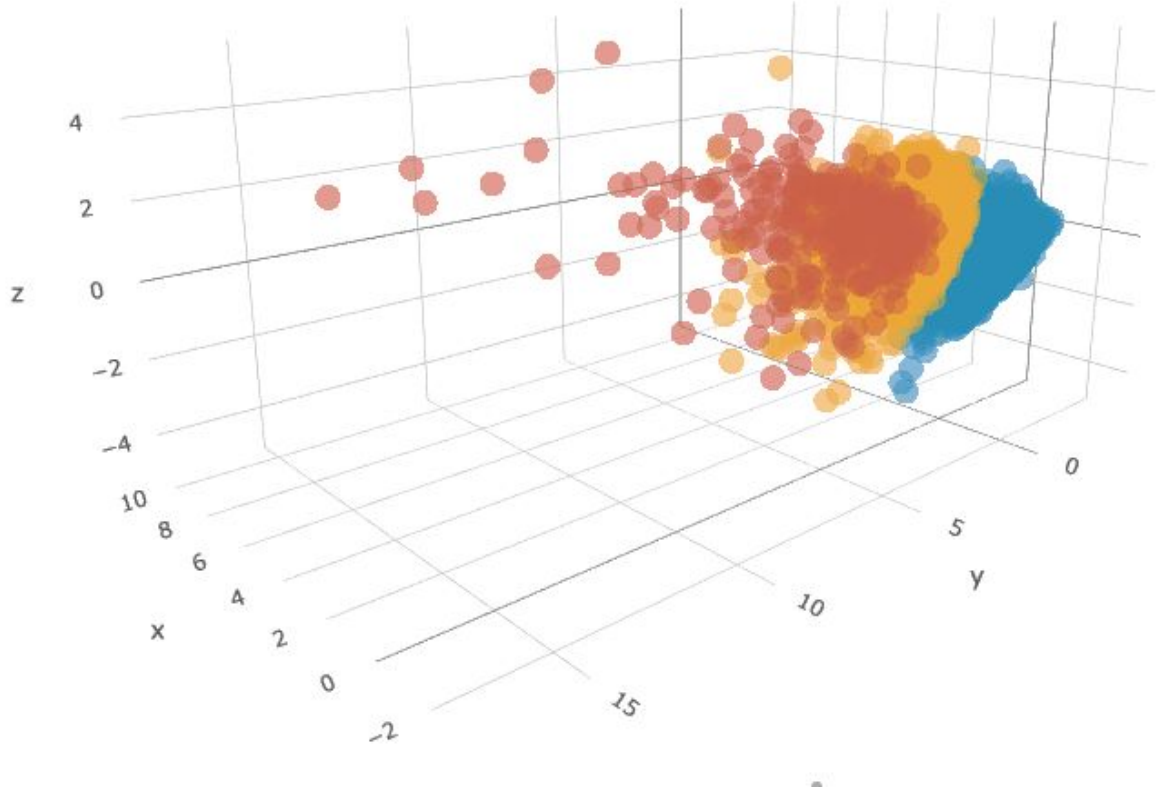
Clustering to identify unique clients

Building better models

```
| inputlookup payRunTimes.csv

| fit StandardScaler *
with_mean=true   with_std=true

| fit PCA * k=3

| fit KMeans k=3 "PC_1" "PC_2"
"PC_3"
```

To use the 3D Scatterplot tool, you also need to rename your columns

```
| rename cluster as clusterId PC_1
as x PC_2 as y PC_3 as z
```



Key Takeaways

Predicting User Experience

1. The Machine Learning Toolkit can be used to build predictive models of user experience based on client demographics
2. These predictive models can be used to pinpoint areas for improvement in applications
3. The same methods can be used to look at time-based trends in performance data

Q&A

Dr. Ken Tupper | Paychex

splunk> .conf19

.conf19

splunk>

Thank You !

Go to the .conf19 mobile app to

RATE THIS SESSION