splunk> .conf2017

Continuing Collaboration Between IT Operations + Research

The Impact of Student Achievement Predictions to Operational Prediction…and back again

Matt Bernacki | College of Education Faculty, University of Nevada, Las Vegas
Cyndi Backstrom | IT Operations, University of Nevada, Las Vegas

September 25-28, 2017 | Washington, DC

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.

About Us

- Dr. Matthew Bernacki
  - Educational Researcher
  - Studies student motivation, behavior, and self-regulation of learning with technology
  - Learning Science lead for the Research Project

UNLV College of EDUCATION

MALT·lab

Research on Metacognition and Motivation in Advanced Learning Technologies

NSF DRL #1420491

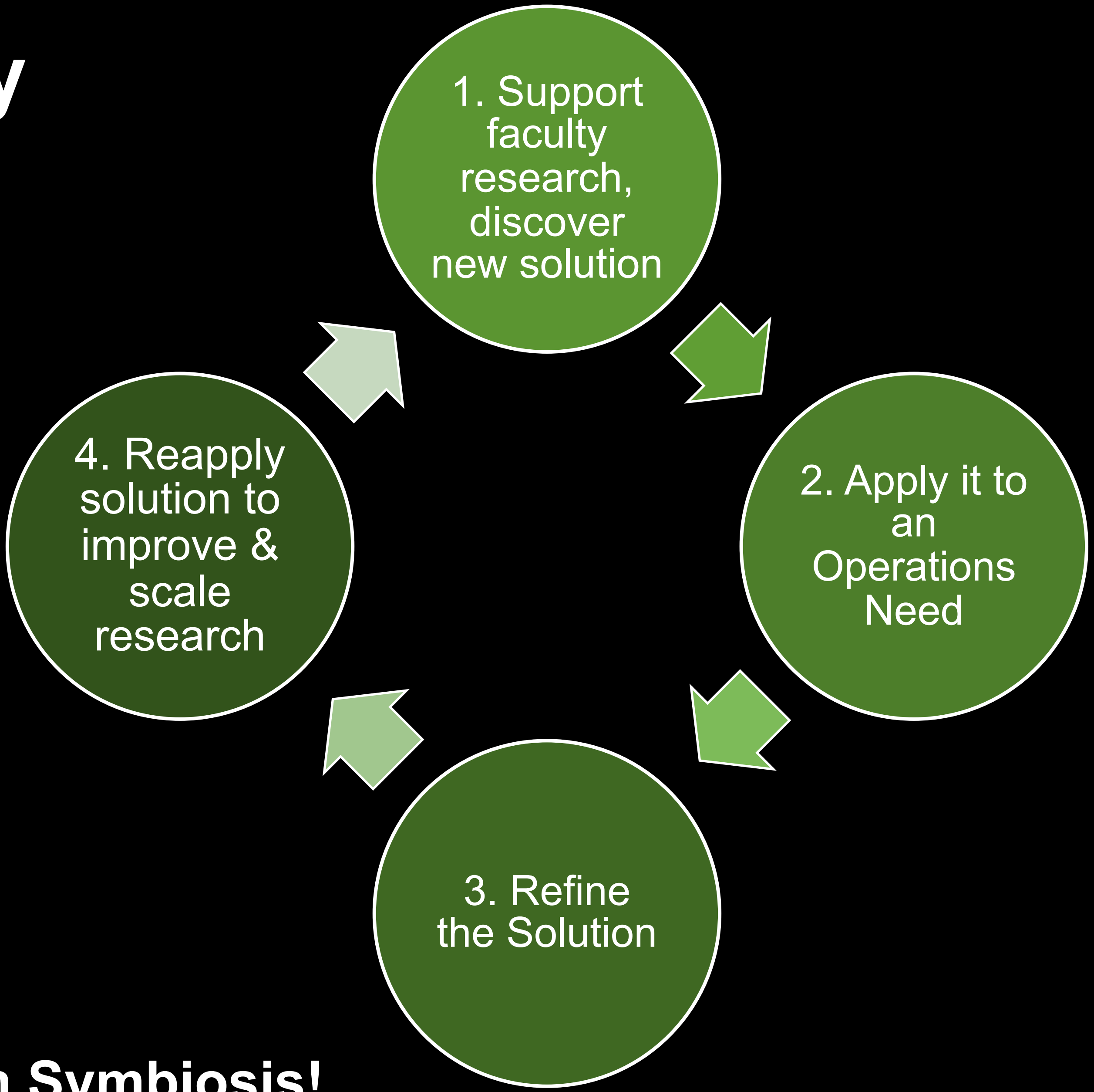
- Cyndi Backstrom
  - Splunk Support
  - Data Modeling Lead for the Research Project
  - Emerging MLTK user
    - 1 week of training in February, 2017
    - Increasing use… and lots of trial and error

UNLV Office of INFORMATION TECHNOLOGY

Operations Center

Chapters of Today's Story

- Splunk 2016 .conf recap
- Research Updates
- Applying Research-Derived Knowledge to Improve Operations
- Ops+MLTK Expertise Back to Research



Operations & Research Symbiosis!

The .conf 2016 Recap

Research Context

~ 29,000 Students (24,000 Undergraduates)

Minority Serving Institution (MSI)

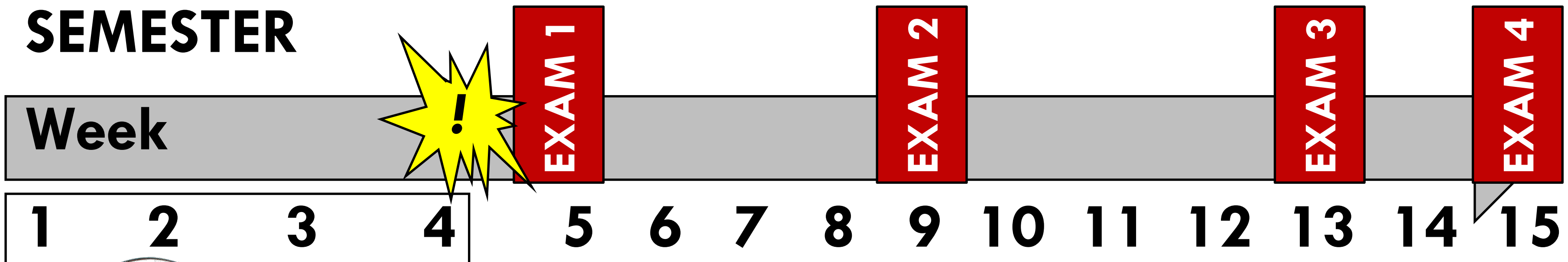
- *Hispanic Serving Institution (HSI)*
- *Asian, Native American & Pacific Islander Serving Institution (ANAPISI)*

Majority first generation & Title 1 HS graduation

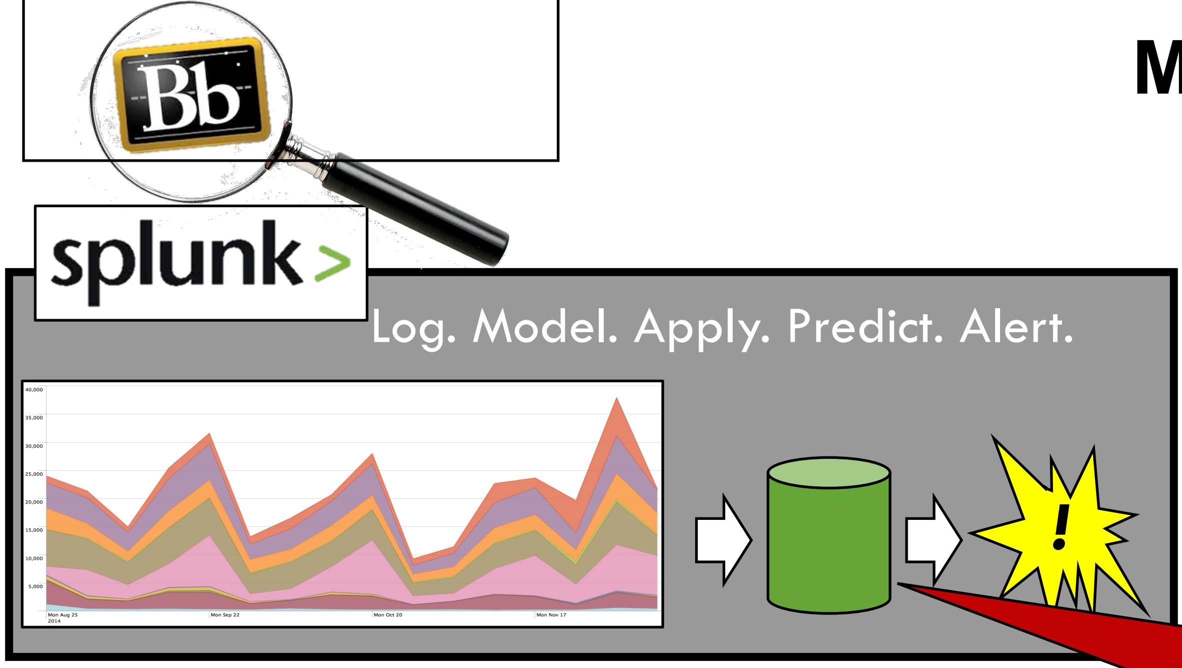
The .conf 2016 Story

Research+Operations: A Love Story

- Project Goals
  1. Work with STEM instructors to digitize and host materials they use in large lecture courses
  2. Use Splunk to build data models to trace student learning with digital LMS-hosted resources
  3. Use student traces + grades to develop prediction models that identify those who will struggle
  4. Program the prediction model into Splunk; provide alerts to students before they begin to fail



Model Building (Summer & Fall 2015)



N: 334 Fall 2014 bio students

Data: week 1-4 LMS events

Criteria to Predict:
Earning of a "C or worse"
Goal: Identify those who'll need to retake a class to progress in their STEM major

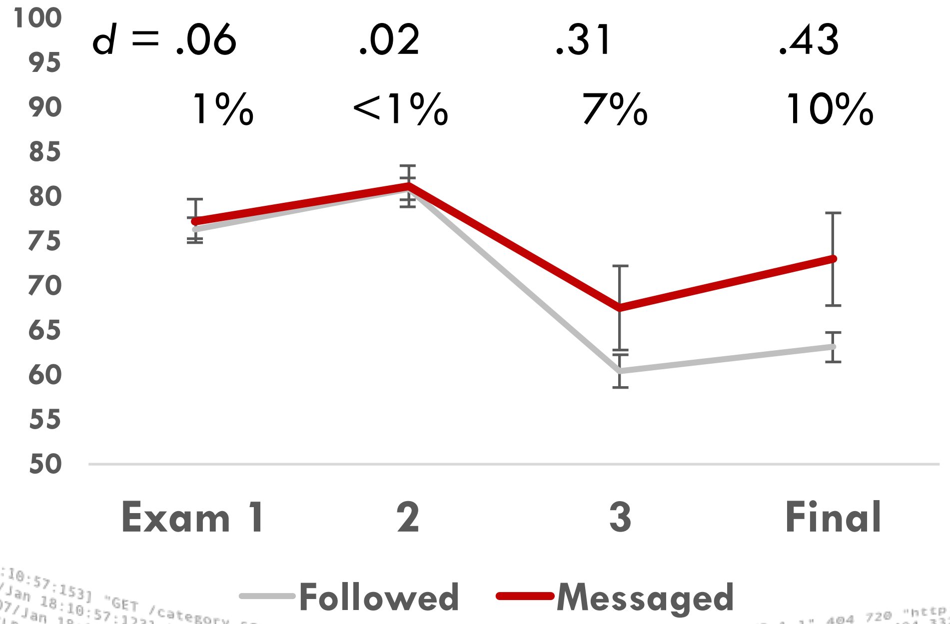
- Events as occurrence & frequency/week
  (Item level and resource type level)
- Forward Selection Logistic Regression Model
  (best possible model)
- 10-fold, leave one out cross validation (prevent overfit)

Study 1 **Results**

Effects on Exam scores

Messaged vs. **Follow**

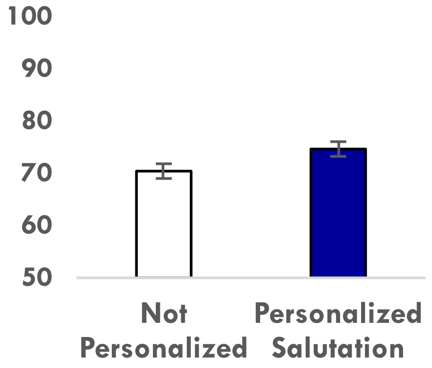
- No immediate effect…
- … but over time, messaged students increase their gains



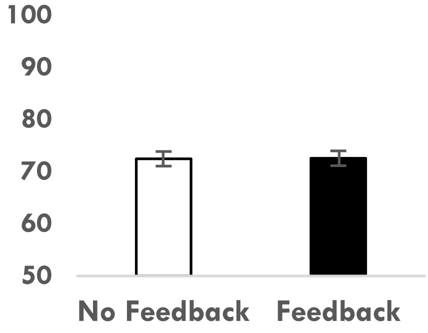
Does the content of the message matter?

- Oversampled (80%) to test message features:
  - **Personalized Salutation**
  - **Negative Feedback**
  - No impact on student responsiveness… …but Impacts on performance

Personalization made a difference (d = .28)



Feedback did not (d = .01)



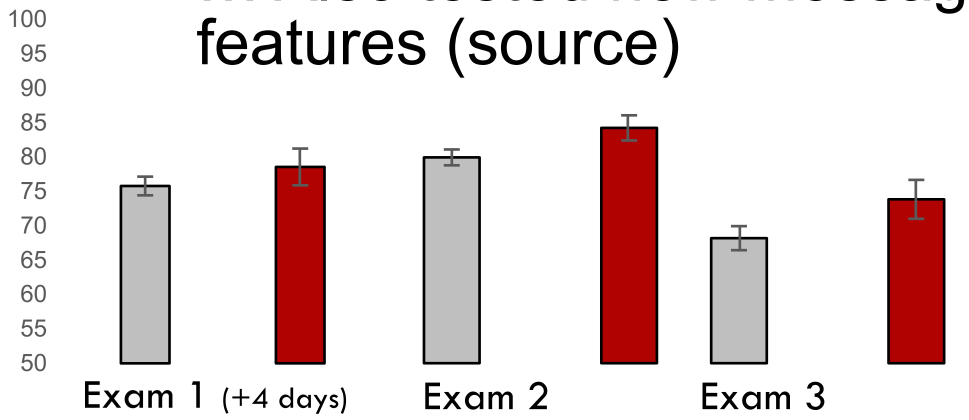
Research Updates

Refinement & Extension: Study 2 & 3

Biology #2

- Refit the prediction model using 2 semesters of data
  - Similar accuracy, less likely overfit
- Personalized message, no feedback

… Also tested new message features (source)

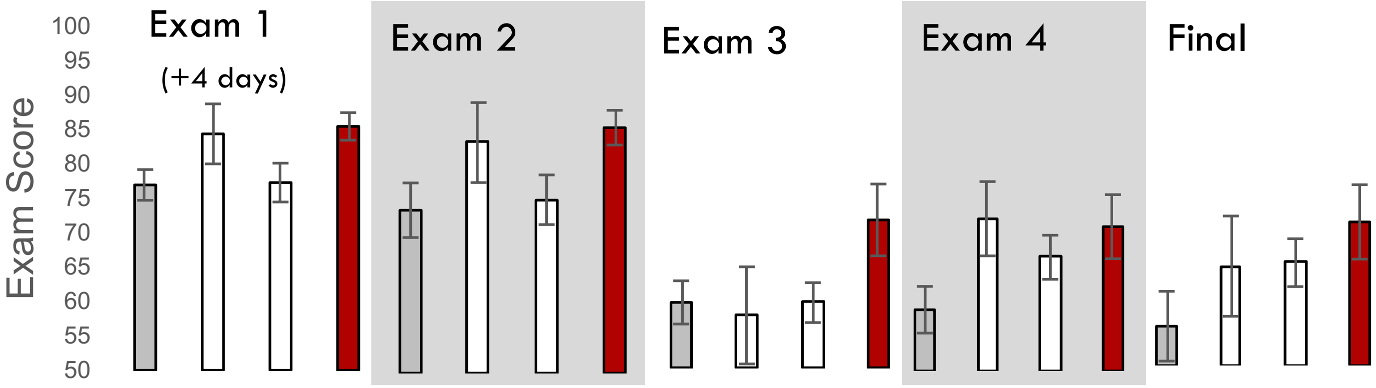


d=.20 (3%) *d*=.40 (4%) *d*=.34 (6%)

445 students identified!

And Calculus!

- Replicated prediction modeling method
- Messaged Day 1 of Week 4 (Exam on Friday [Day 5])
- Create Math specific advice page
  - Learning strategies re: problem solving



(9%) (12%) (12%) (12%) (15%)

+94!

Applying Research-Derived Knowledge to Improve Operations

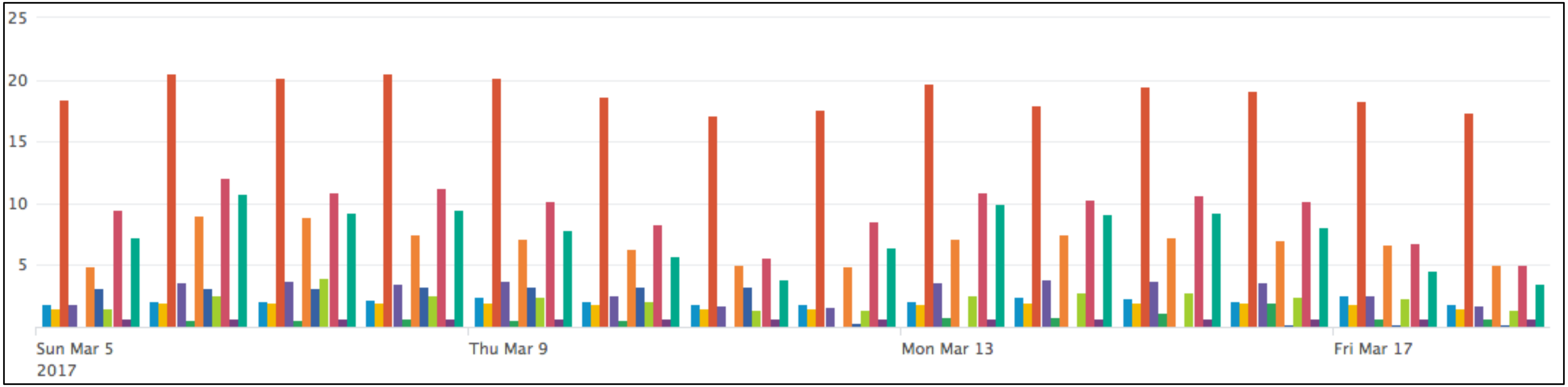
How to make your day better!

Pivot Research to Operations

- Data integrity
  - Data interruptions
  - Incomplete database import
  - Duplicate data
- Operations passive to active

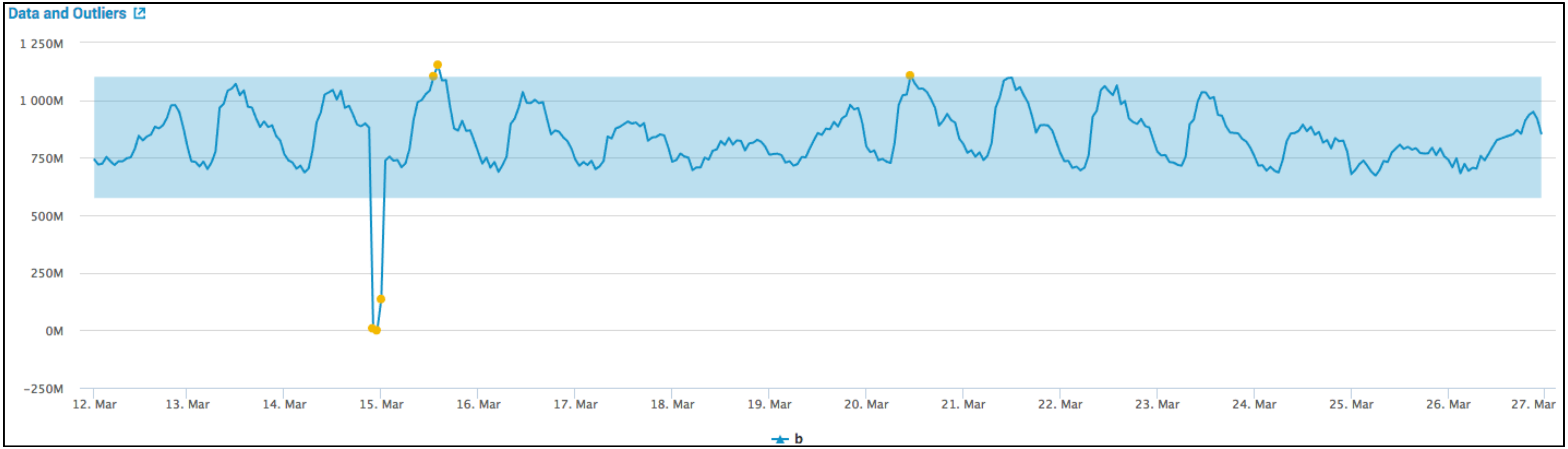
Data Interruptions

- License usage per index over two weeks
- Can you find the data interruption?



Data Interruptions - Found

- License usage for one index over two weeks
- Can you find the data interruption?

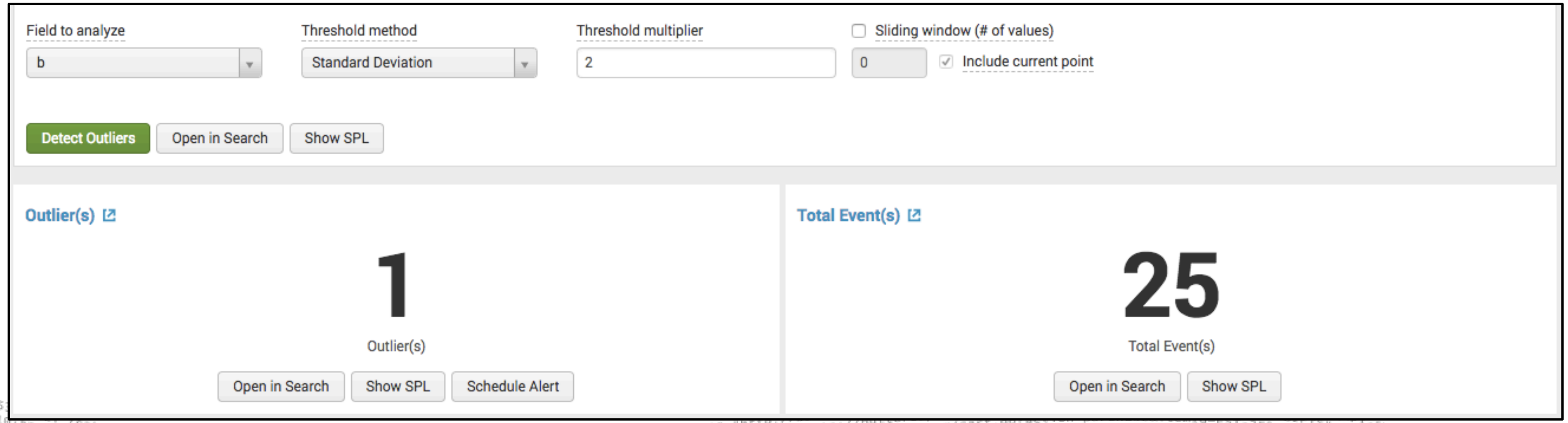


Data Interruptions - Search

- Base search:

```
index=_internal source=*license_usage.log type="Usage" idx=nde_fwsm-dc b=*
| bin _time span=1h
| stats sum(b) as b by _time
| makecontinuous _time span=1h
| fillnull value=0
```

- MLTK – Assistants - Detect Numeric Outliers - Standard deviation



Data Interruptions - Operations

- Operations solution:
  - Report on all indexes
  - Send alert if an issue is found

_time	idx	b	lowerBound	upperBound	isOutlier	avg	stdev
2017-08-26	bos_footprints-temp	106	1343.792713	5029.457287	1	3186.625000	1842.832287
2017-08-27	bos_footprints-temp	424	1343.792713	5029.457287	1	3186.625000	1842.832287
2017-08-21	clt_sql	796302	1988671.026462	7252669.473538	1	4620670.250000	2631999.223538
2017-08-27	clt_sql	8738932	1988671.026462	7252669.473538	1	4620670.250000	2631999.223538
2017-08-21	clt_syslog	79673	203568.344490	624971.155510	1	414269.750000	210701.405510
2017-08-27	clt_syslog	725701	203568.344490	624971.155510	1	414269.750000	210701.405510
2017-08-21	cs_footprints	1424038	2101208.419191	3594605.330809	1	2847906.875000	746698.455809
2017-08-23	cs_footprints	3736188	2101208.419191	3594605.330809	1	2847906.875000	746698.455809
2017-08-21	dev-sea_idm	13456525	20808328.157096	35989818.342904	1	28399073.250000	7590745.092904
2017-08-28	dev-sea_idm	19274577	20808328.157096	35989818.342904	1	28399073.250000	7590745.092904

Data Interruptions - Operations

- Operations solution:

Generate list of indexes

```
index=_internal source=*license_usage.log type="Usage" idx=* b=*
| stats count(result_count) by idx
```

Calculate outliers per index

```
| map maxsearches=25 search="search index=_internal source=*license_usage.log type="Usage" idx=$idx$ b=*
| bin _time span=1d
| stats sum(b) as b by _time, idx
| makecontinuous _time span=1h
| fillnull value=0
| eval b=round(b,0)
| eventstats avg(b) as avg stdev(b) as stdev by idx
| eval lowerBound=if((avg-stdev*1)<0,(0),(avg-stdev*1)), upperBound=(avg+stdev*1)
| eval isOutlier=if('b' < lowerBound OR 'b' > upperBound, 1, 0)
| where isOutlier=1
| table _time, idx, b, lowerBound, upperBound, isOutlier, avg, stdev
| sort idx, _time"
```

Send alert

```
| eval alert_send=if(_time=(relative_time(now(),"-1d@d")),"send","no send")
| search alert_send="send"
```

Operations - Data Integrity

- Same approach to resolve other known issues:
  - Incomplete database import:
    - Normal is 39,451 vs 1,000
  - Duplicate data:
    - Syslog being feed is being indexed twice
  - MORE: Sourcetypes, Saved Searches (lookup builds), Alerts, Notifications, Help Requests, etc.

Operations - Data Integrity

- Future:
  - Comparing like events
    - Cyclical events (start of the semester)
    - Monday to Monday
  - Adding metadata:
    - Static lower bounds
    - Alert priority

Pivot Operations to Research

- Recommendation:
  - Outreach to other groups
  - Different projects may provide insight into providing solutions for yourself

Ops + MLTK Expertise Back To Research

splunk> .conf2017

Circling MLTK Knowledge Back to Improve and Scale Research

- The Research Solution (i.e., our business as usual)
  Not scalable! Lots to clean up…
  - ***Messes***
    - Data models that need to be tidied
    - Lookups with many contributors, poor documentation
  - ***Inefficiencies***
    - Data models rely on semester specific metadata; requires rebuilding of lookups, reports each semester
    - Prediction modeling happens ***offline***, apart from data model

Current Problems With Offline Prediction Modeling

- To model, **data** fields need to be
  - Selected into 1+ report(s)
  - Frozen into a static table of predictors
  - Exported (and per FERPA, deidentified)
- When **modeling** offline
  - Some prediction algorithms have quirks (and poor documentation)
  - Processing power limits the size of your predictor set
- To build the model back into Splunk for **predicting** student success
  - Rebuilding is work intensive, repetitive, and human-driven

Solutions Provided By MLTK

- Data grab
  - MLTK can use an SPL interface to conduct modeling based reports that are live, editable
- Model Building
  - Algorithms are known, plentiful
  - Processing power is immense; optimal models can be identified quickly
- Applying Prediction Models
  - No rebuilding required; can clone data models and point and the new source

Goal 1: Replication of the Offline Solution in MLTK

OFFLINE

SELF CONTAINED IN SPLUNK + MLTK

Prepare the test data in Splunk. Export as .csv

Apply the Prediction Algorithm

Logistic Regression

With Forward Selection

Cross validate and Confirm the solution.

Program predictors into a Splunk report; apply to new data model

Intervene!

Predi A or B vs. C or worse

Use o ly most predictive behaviors

Split t e sample into 10 parts

Ensu it fits all groups

Simila levels of accuracy

Simila set of predictors

Goal 1: Replication of the Offline Solution in MLTK

OFFLINE

Prepare the test data in Splunk. Export as .csv

Apply the Prediction Algorithm

Logistic Regression

With Forward Selection

Cross validate! Confirm the solution!

Program predictors into a Splunk report; Apply to new data model

Intervene!

SELF CONTAINED IN SPLUNK + MLTK

No need. We can do our whole workflow in Splunk now!

Logistic Regression is available out of the box.

Forward Selection can be added from a python library and wrapped into the Splunk MLTK App.

Cross validation isn't included out of the box…
… but it can be written right in search!:

* SEE APPENDIX

Predictors, accuracy metrics are similar.

Success! We can now model right in Splunk, improve our models as new data are available, and update our predictor sets to make more precise predictions and

Intervene with confidence

Goal 2: Use MLTK to Improve the Approach!

The workflow: Pre-Splunk

In Splunk MLTK

- SPL anyone can read and reference:
  - MLTK
    - |fit FieldSelector type=categorical param=10 Grade from *
    - |fit LogisticRegression Grade from fs_* into model_a
    - | fit SVM Grade from fs_* into model_b
    - | fit RandomizedLogisticRegression Grade from fs_* into model_c
  - Consume immediately as a report/dashboard/alert

Goal 3: Spread the Solution, Improve All Students' Success…

- Soon!: An APP (available from Splunkbase… or GitHub?) Stay tuned…

1. Prepare your data
   - What is student success? (identify your outcome to predict)
   - What do you have on hand to predict it? (prepare your reports)
2. Apply the SPL for prediction and cross-validation
3. Check your accuracy metrics
   - Do you successfully predict the outcome for your target population?
4. Build reports for those predictors, sum them and identify students in need.
5. Help them out!

Questions?

CONTACT

matt.bernacki@unlv.edu • cyndi.backstrom@unlv.edu

MORE

faculty.unlv.edu/wpmu/bernacki/

splunk> .conf2017

Thank You

Don't forget to rate this session in the .conf2017 mobile app

APPENDIX

MACHINE LEARNING TOOL KIT

SPL FOR CROSS VALIDATION

Cross Validation Informally in SPL

Step 1 : Create your models with one partition holdout randomly

From the Desk of Alexander Johnson

```
| makeresults count=10
| streamstats count| rename comment as "0-indexed partition_numbers require us to subtract 1”
| eval count = count – 1
| map maxsearches=10 search="
| inputlookup airline_tweets.csv where airline_sentiment_confidence > 0.8
| fields airline_sentiment text
| sample partitions=10 seed=42
| search partition_number != $count$
| fit TFIDF text stop_words=english into vectorizer_$count$
 | fit LogisticRegression airline_sentiment from text_tfidf* into lr_$count$ "
```

Cross Validation Informally in SPL

Step 2 : Score your models on the holdouts

From the Desk of Alexander Johnson

```
| makeresults count=10
| streamstats count| rename comment as "0-indexed partition_numbers require us to subtract 1”
| eval count = count – 1
| map maxsearches=10 search="
| inputlookup airline_tweets.csv where airline_sentiment_confidence > 0.8
| fields airline_sentiment text
| sample partitions=10 seed=42
| search partition_number = $count$
| apply vectorizer_$count$
| apply lr_$count$ as p
| `classificationstatistics(airline_sentiment, p)`"
```