© 2020 SPLUNK INC.

Machine Learning on Streams

Ram Sriharsha

Sr Director, Head of Machine Learning | Splunk

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.

Splunk, Splunk>, Data-to-Everything, D2E and Turn Data Into Doing are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names or trademarks belong to their respective owners. © 2020 Splunk Inc. All rights reserved

© 2020 SPLUNK INC.

Ram Sriharsha

Sr Director, Head of Machine Learning | Splunk

splunk> .conf20

© 2020 SPLUNK INC.

Agenda

1) A sample problem and its real-world use case

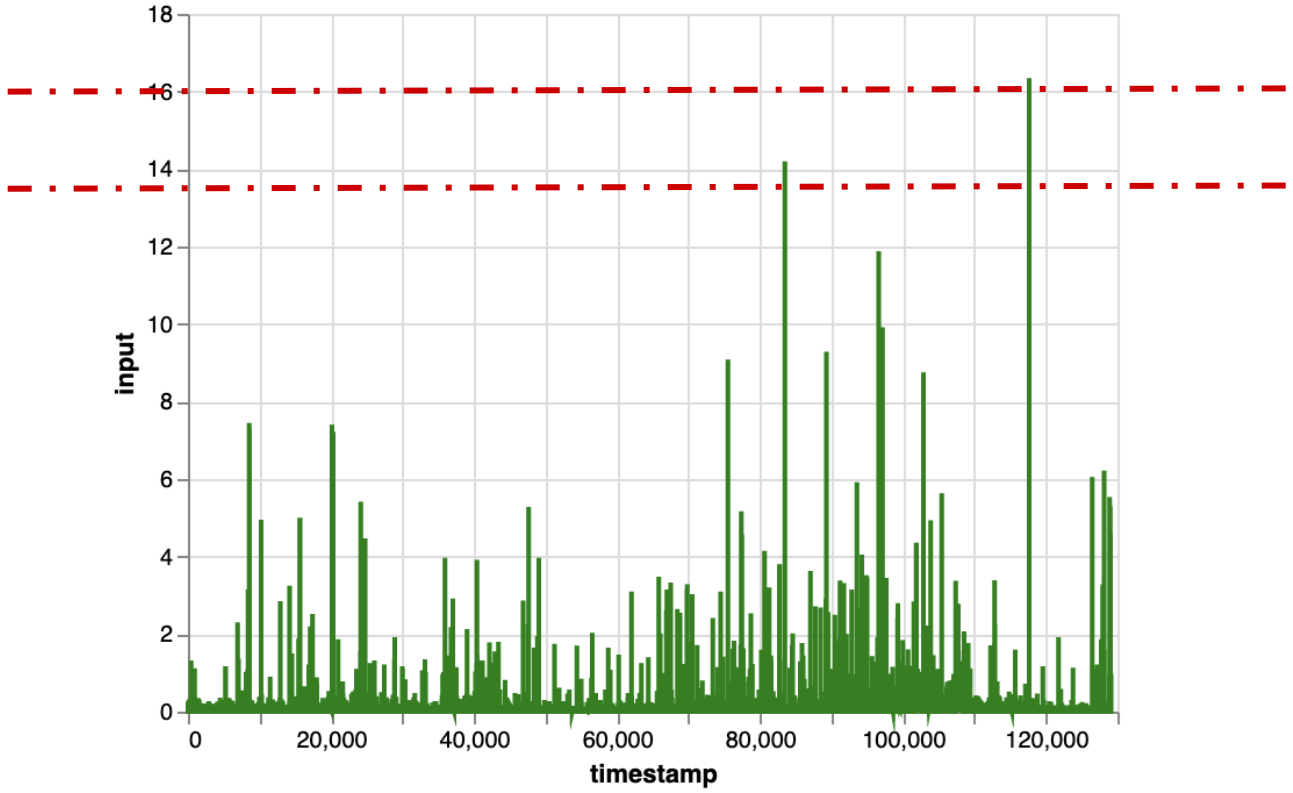
2) Solving for scale and operational Simplicity

3) Where are we going from here?

© 2020 SPLUNK INC.

A Sample Problem

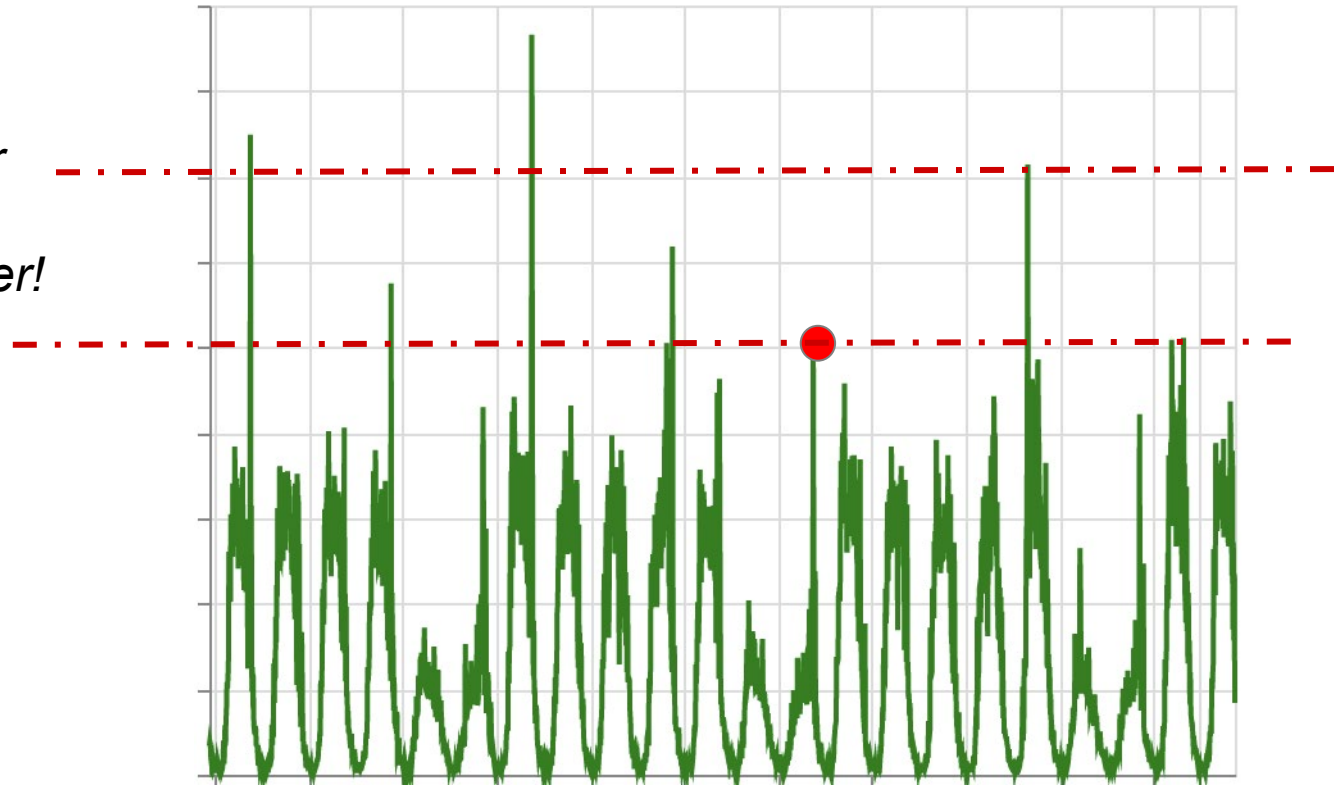
We have a time series. We would like to detect outliers in this time series
Naive approach = static thresholds.



© 2020 SPLUNK INC.

What if the Metric Has Seasonality?

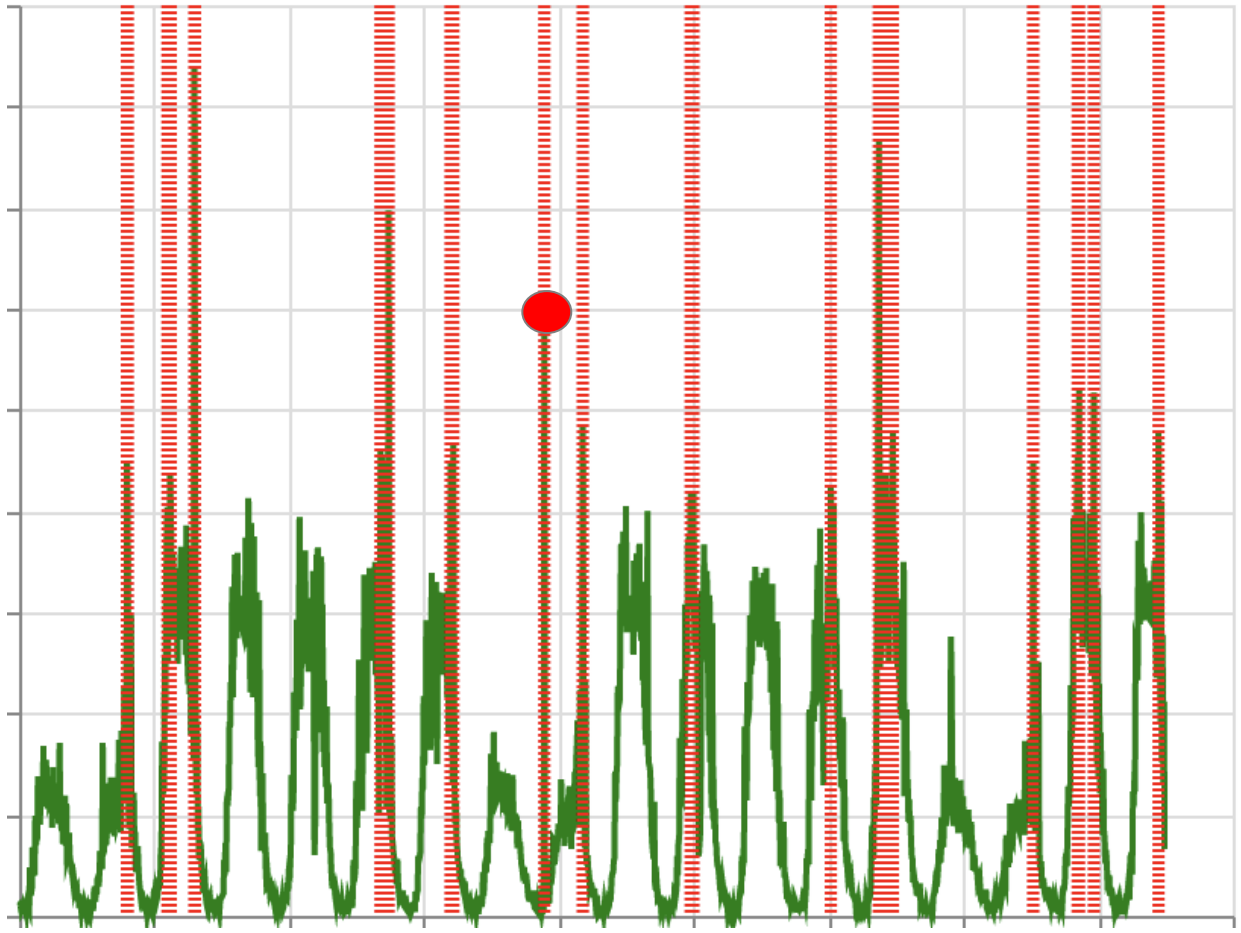
Static Threshold either has too much noise or doesn't catch the outlier!



© 2020 SPLUNK INC.

What if the Metric Has Seasonality?

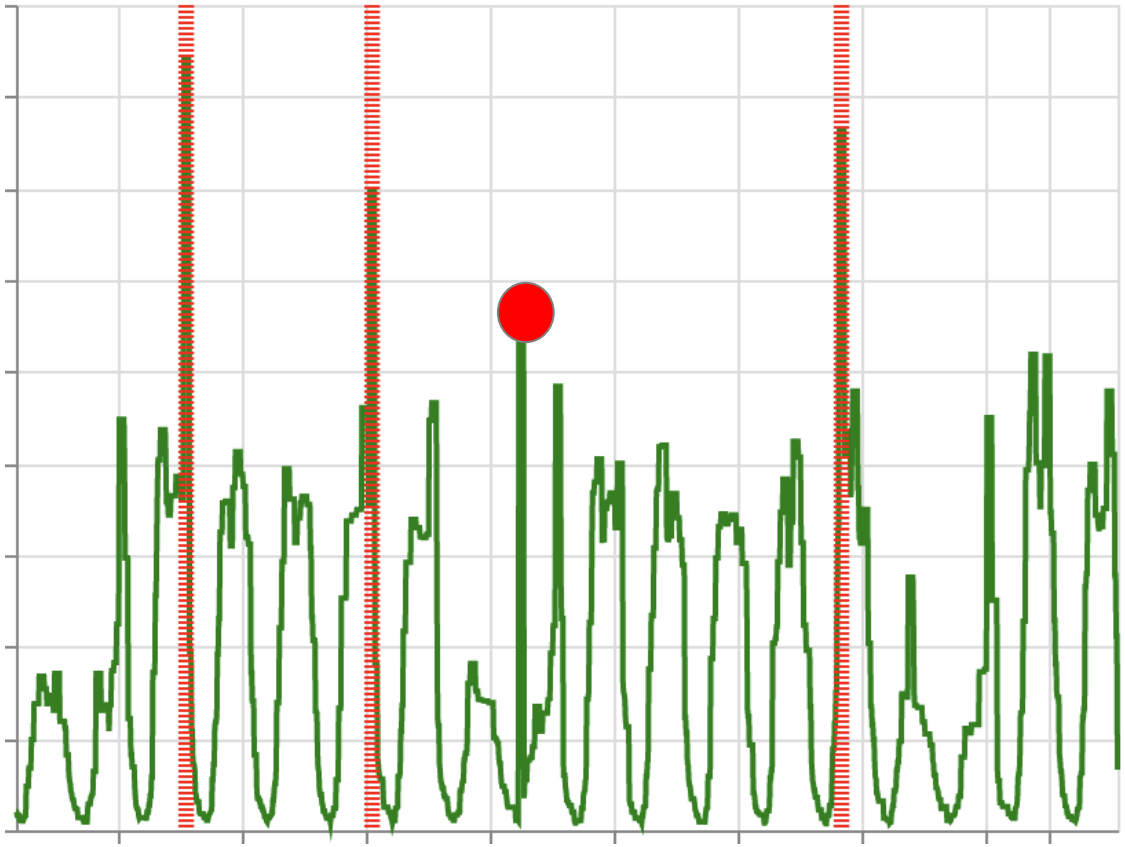
Even adaptive threshold doesn't fare much better!



© 2020 SPLUNK INC.

Maybe Preprocess?

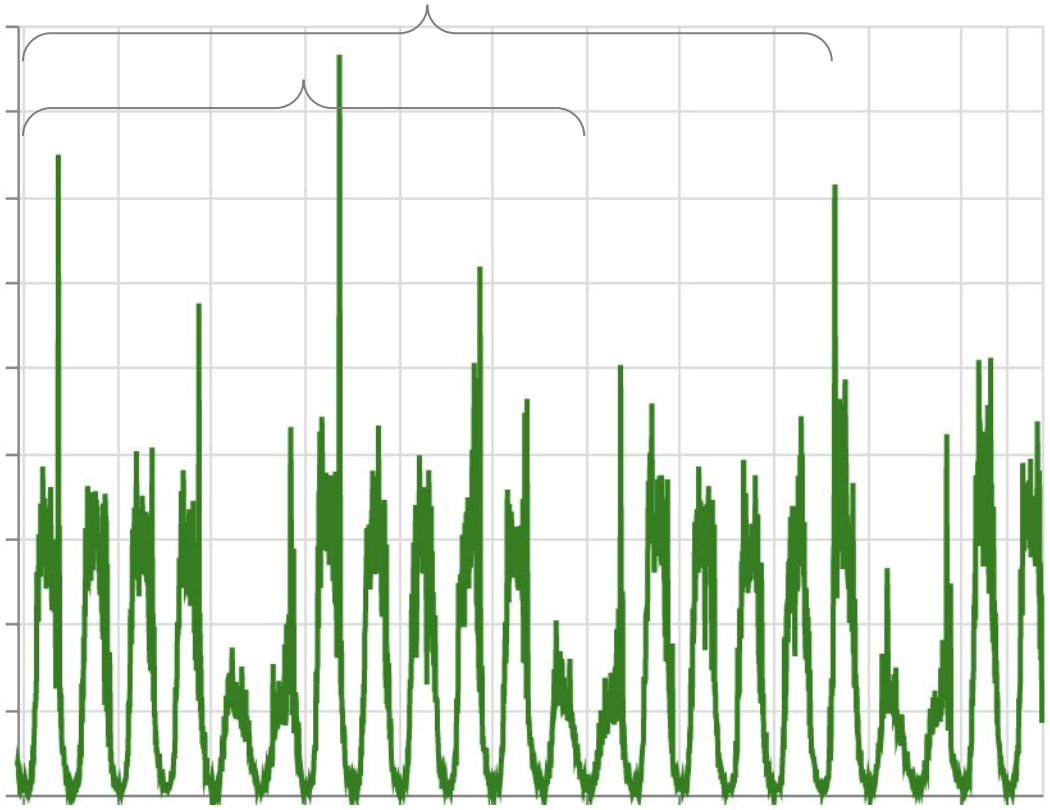
Compute 95th percentile over last two hours and apply adaptive thresholding -> YMMV



© 2020 SPLUNK INC.

A Better Solution

Learn a model of time series based on past,
use that to predict outliers in the present



© 2020 SPLUNK INC.

Many Ways to Do This!

ARMA/ ARIMA/ SARIMAX

- Decompose time series into Autoregressive Moving Avg
- Use Seasonal differencing and de-trending to make the signal stationary

STL

- Direct decomposition of time series into seasonal, trend and residual
- Loess smoothing

Probability Density Function

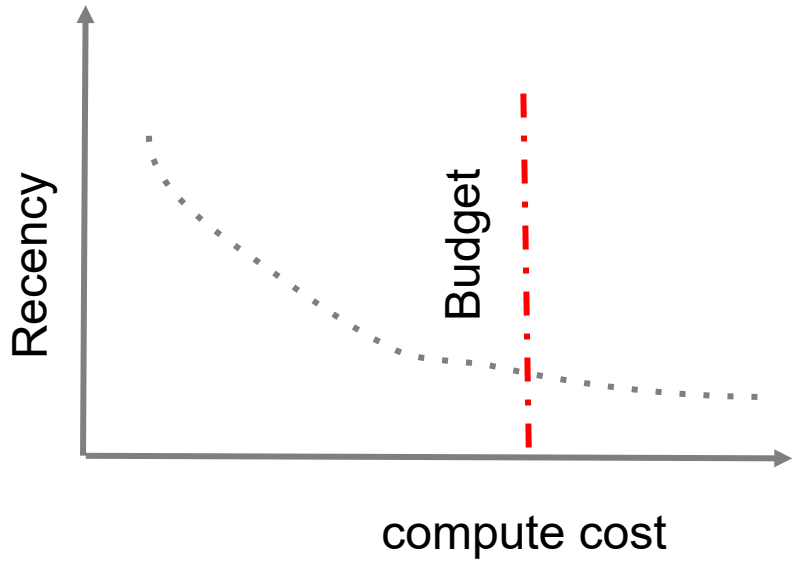
- Fit a kernel density function for a multiple of seasonality
- Tune for a small % of errors
- Use fitted function to predict outliers in new data

© 2020 SPLUNK INC.

They All Have Problems!

All these approaches are batch algorithms

- Retraining on each data point = prohibitive
- Typically trained on last X days/ weeks and recency of a day or week
- Frequent retraining >> more complex models



© 2020 SPLUNK INC.

Doing This at Scale!

We collect multiple petabytes of logs from thousands of applications real time daily. We metricize these logs which results up to a million metrics. We want to monitor these metrics for outliers

How do we even threshold a million metrics?

- Logistically impossible to even set up!

Do these metrics contain any seasonality? if so, how do we handle it?

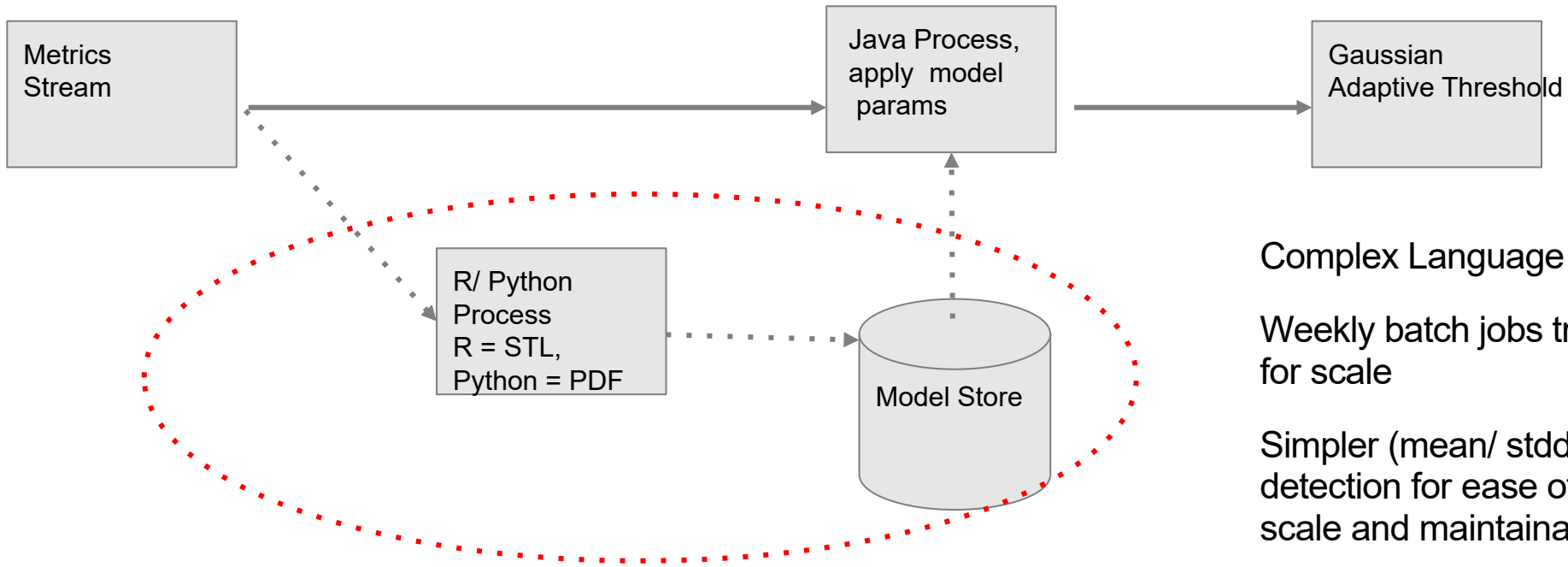
- Cannot possibly individually analyze each metric and model separately
- Current batch approaches are trained on a few weeks worth of data and run as weekly jobs => stale data!!

How do we deploy and monitor these systems?

- Even if we build O(100k) models, how do we deploy them, score on stream and monitor?

© 2020 SPLUNK INC.

How is This Done Today?



Complex Language Bindings

Weekly batch jobs trade off recency for scale

Simpler (mean/ stddev) outlier detection for ease of development, scale and maintainability

Customers need to manage state, scalability, operationalization etc.

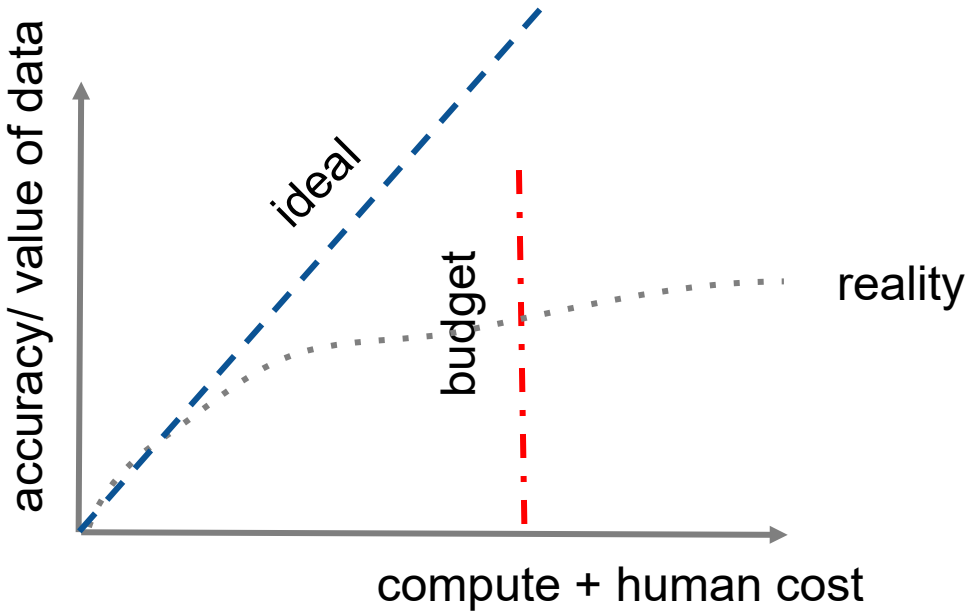
© 2020 SPLUNK INC.

These Challenges Generalize to Many ML Use Cases!

Traditional ML = batch

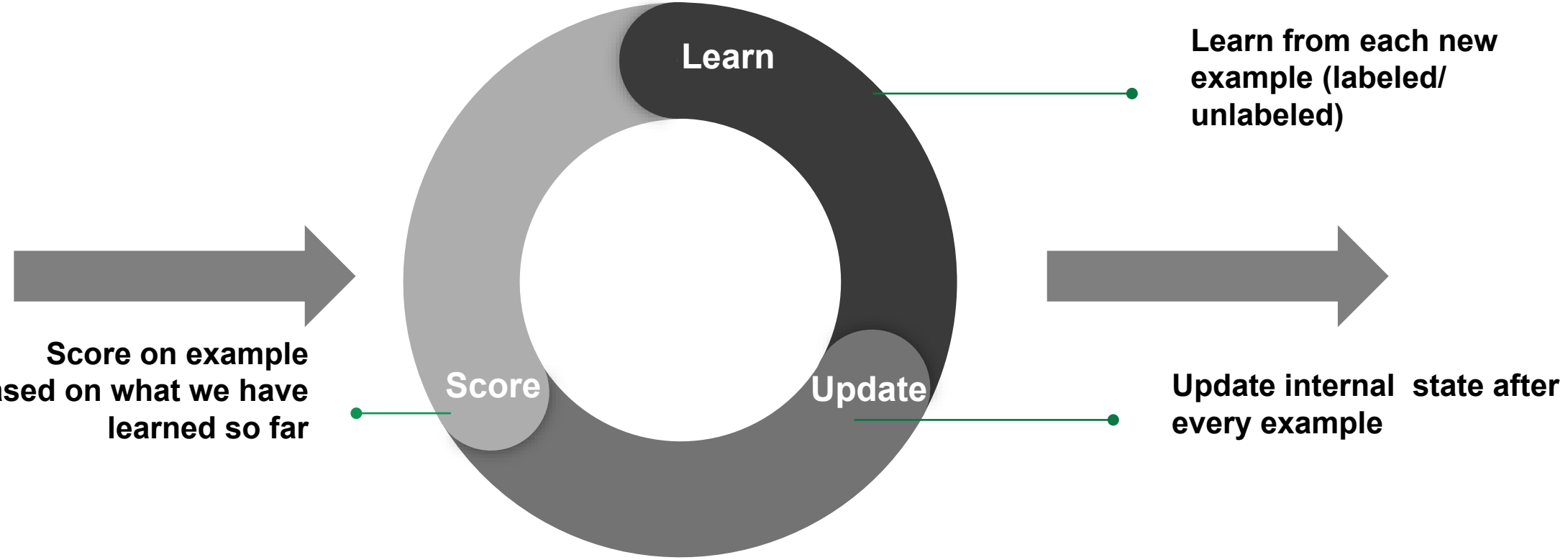
- Retraining on each data point = prohibitive
- Modeling / Feature Engineering cost does not scale as number of models grow!

Frequent retraining >> more complex models



© 2020 SPLUNK INC.

How to Get Closer to Ideal Tradeoff Curve?



We call them Unbounded learning systems!

© 2020 SPLUNK INC.

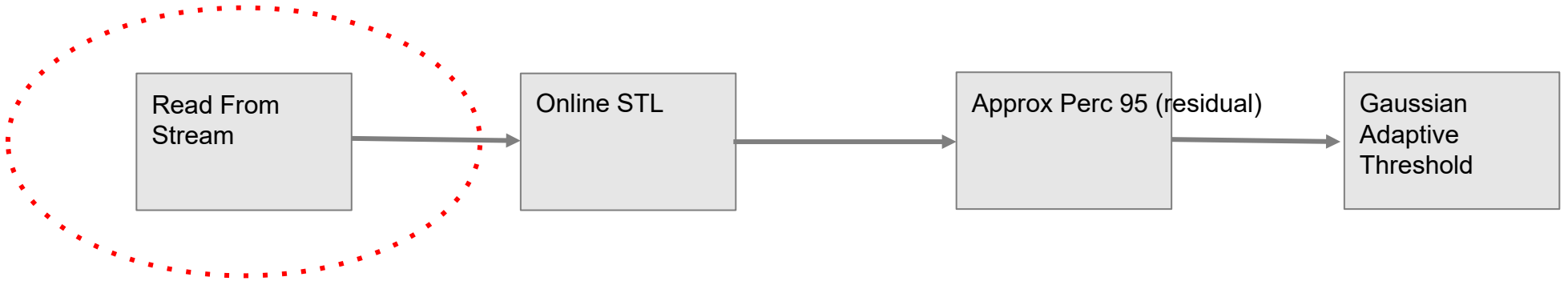
How Do We Solve the Toy Problem at Scale?


```
| from read_csv("s3://conf2020/errors.csv")
| eval body=_time
| extract_timestamp field="time" rules=[iso8601_timestamp()]
| stl value="input" seasonality=80
| eval key=concat([host, application)
| stats perc(residual, 0.95) as perc95 by key, span(time, 2h, 15min)
| eval timestamp=window_end
| adaptive_threshold value="perc95" entity="key";
```

- No Complex Language Bindings, simply use SPL!
- Models are always recent!
- More sophisticated adaptive thresholding including skewness adjustment possible
- We manage state, scalability, ops

© 2020 SPLUNK INC.

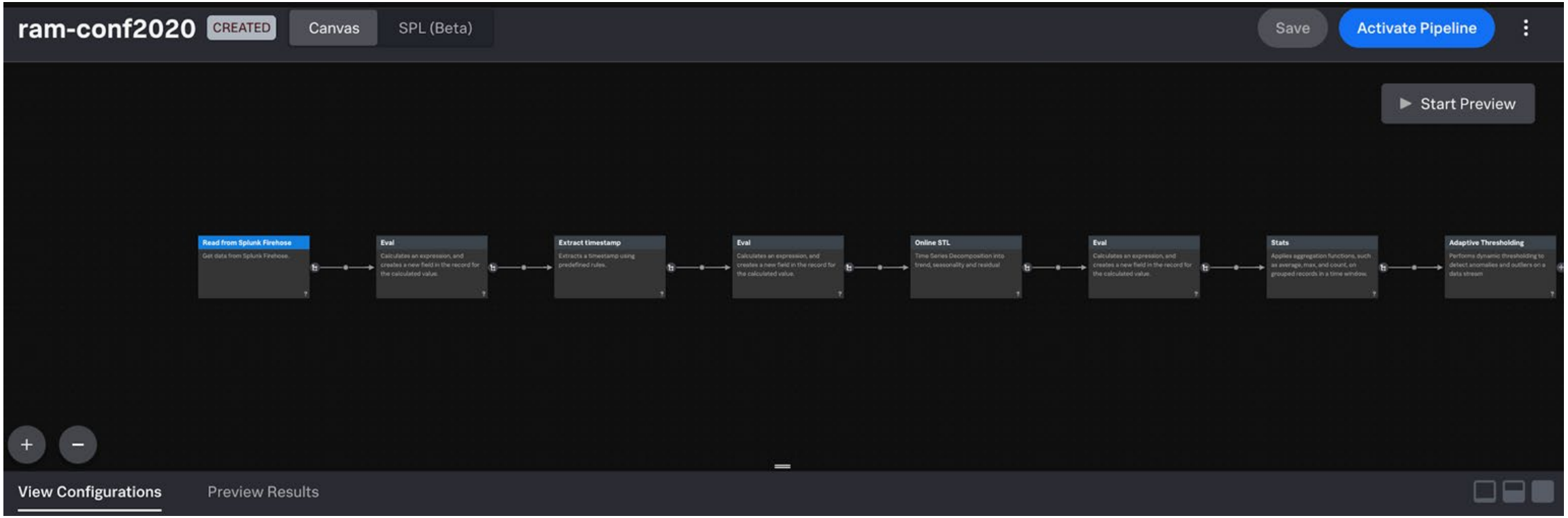
How Do You Deploy This in Production?




```
| from read_pulsar(topic=...)
| eval body=_time
| extract_timestamp field="time" rules=[iso8601_timestamp()]
| stl value="input" seasonality=80
| eval key=concat([host, application)
| stats perc(residual, 0.95) as perc95 by key, span(time, 2h, 15min)
| eval timestamp=window_end
| adaptive_threshold value="perc95" entity="key";
```

© 2020 SPLUNK INC.

How Do You Deploy This in Production?



© 2020 SPLUNK INC.

Scale

40k

records per sec per core

50ms

max latency of ML pipeline outside of stats, ...

1mm

cardinality

100x

throughput compared to batch pipelines for same accuracy

© 2020 SPLUNK INC.

Other Algorithms Today

Outlier Detection

Adaptive Thresholding, Sequential Outlier Detection, Conditional Anomaly Detection

Time Series

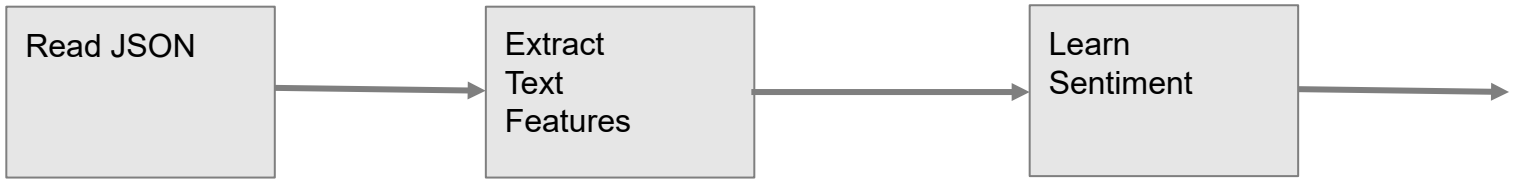
Driift Detection, Online STL

Statistics/ ML

Online Clustering, Approx Perc, Distinct Count, Frequent Items

© 2020 SPLUNK INC.

Another Example




```
| from read_json("s3://conf2020/reviews.json")
| eval  input=concat([reviewText," ",summary]),
    key=productCategory,
    label=if(overall >= 3.0, 1.0, -1.0)
| select input, label, key, reviewerID
| analyze_sentiment value="input";
```

- Sentiment Analysis learns from labeled examples
- Applies learning to unlabeled/ new examples
- Handles feature engineering as well as learning (eg. tokenization, word embedding and online learning)

© 2020 SPLUNK INC.

Design for Scale

Throw your data at the streaming algorithm and it will start to learn

- No pre-configuration needed, i.E self tuning
- Adapts to data
- Learns the distributions

Converges quickly

- Learning stabilizes soon

Unbounded in cardinality of models

- Learn as many models as you want , computational cost ~ linear in cardinality

Unbounded in data volume

Recency baked in - models are continually learning!

© 2020 SPLUNK INC.

Design Not Just For Scale

Drag and Drop streaming-ml pipelines

- Zero code experience
- One “click” deploy entire ML pipelines

SPL interoperability

- Use ML constructs within SPL to explore data, build pipelines

No downtime ML

- Upgrades to our algorithms do not require pipeline downtime
- No pesky version incompatibilities!

Scale, state and fault tolerance handled by us

- Do not have to re-engineer ML pipelines for production

Observability and Explain-ability of models built in

- Model metrics are available for the pipeline builder
- Explain-ability for various machine learning tasks to help triaging predictions

© 2020 SPLUNK INC.

Where are We Going With This?

Solve the problem of machine learning

- Learning in humans is constant and continuous. Why not in machines?

Make ML a seamless part of your pipeline

- just drop our ML in and let us do the hard work for you

Free you to focus on domain expertise and adding domain knowledge/ rules/ labels where needed

- leverage our labeling framework to provide weak labels as needed

Obtain insights while the data is in motion

Thank You

Please provide feedback via the

SESSION SURVEY