A (VERY) Brief Introduction to Machine Learning for ITOA

Toufic Boubez, PhD

VP Engineering, Machine Learning

Splunk Inc.

.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.

Agenda

- Why Machine Learning?
- Overview of Machine Learning Usage
- Flavor of Statistical Learning
- Machine Learning and ITOA
- Key Takeaways
- Questions
- Answers (if we have time ☺)

Preamble

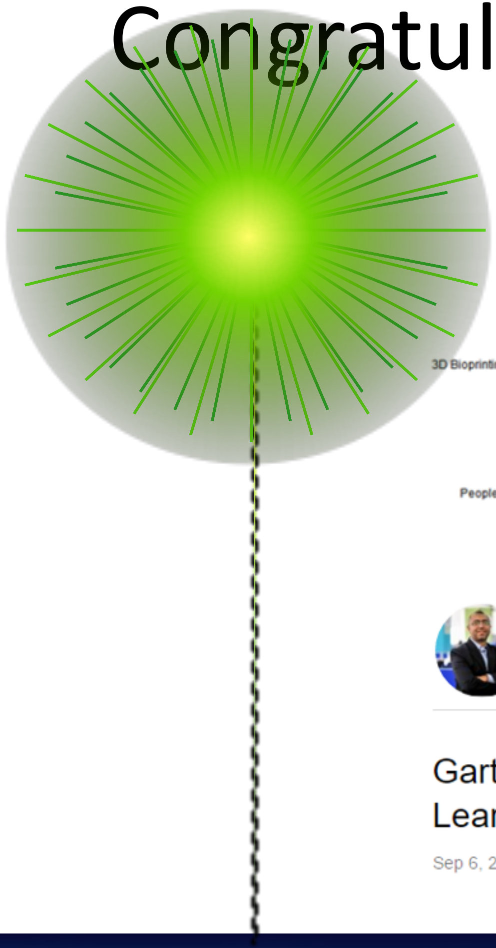
- NOT an advanced course in ML
- IANA Data Scientist! I’m just an engineer that needed to get stuff done!
- Note: all real data

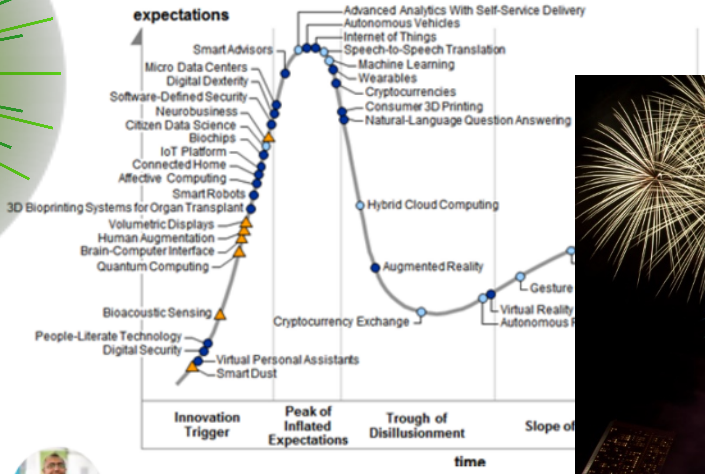
- Note to self: remember to SLOW DOWN
- Note to self: mention cats somewhere – everybody loves cats

About Me

- VP Engineering, Machine Learning, Splunk
- Co-Founder/CTO Metafor Software
- Co-Founder/CTO Layer 7 Technologies
- Co-Founder/CTO Saffron Technology
- IBM Chief Architect for SOA
- Co-Author, Co-Editor: WS-Trust, WS-SecureConversation, WS-Federation, WS-Policy

Congratulations Machine Learning!





Sherif Fathy
Providing Advisory on Analytics Best practices

Gartner 2015 Hype Cycle: Big Data is Out, Machine Learning is in

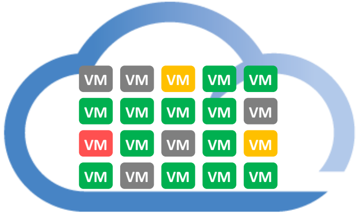
Sep 6, 2015 | 502 views | 18 Likes | 1 Comment

Why Machine Learning??

Evolution of Human Tools

The current IT situation

Fluid Infrastructure




October 17, 2012 2:11 pm

Knight Capital glitch loss hits $461m

By Arash Massoudi in New York

Amazon blames human error for Xmas Eve outage; Netflix vows

Update: Microsoft restores Outlook.com after three-day outage

Apologizes, promises it's taken steps to fix, but within hours

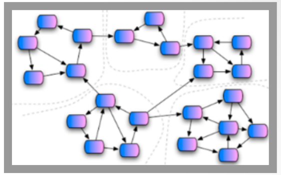
By Gregg Keizer

Google's 5-minute outage means $545,000 revenue loss, 40% drop in global website traffic

George Tinari | 18 August 2013 - 02:02 | 90 Comments | HOT!

Tweet 79 | Like 102

Distributed Applications



Continuous Deployment

Current State Of Affairs: #monitoringsucks

Measure Everything

- Collect 1000's of metrics and logs, most unused
- Analytics methods too simple, not correlated, doesn't help solve outages

Threshold = alert overload

- Too many false positives
- Hundreds of alerts a day, most ignored

IT operations has become a big data challenge

"The [traditional] tools present us with the raw data, and lots of it, but sufficient insight into the actual meaning buried in all that data is still remarkably scarce"

- Turn Big Data Inward With IT Analytics, Forrester Research

Wall of Charts™

The WoC side-effects: alert fatigue

“Alert fatigue is the single biggest problem we have right now … We need to be more intelligent about our alerts or we’ll all go insane.”

- John Vincent (#monitoringsucks)

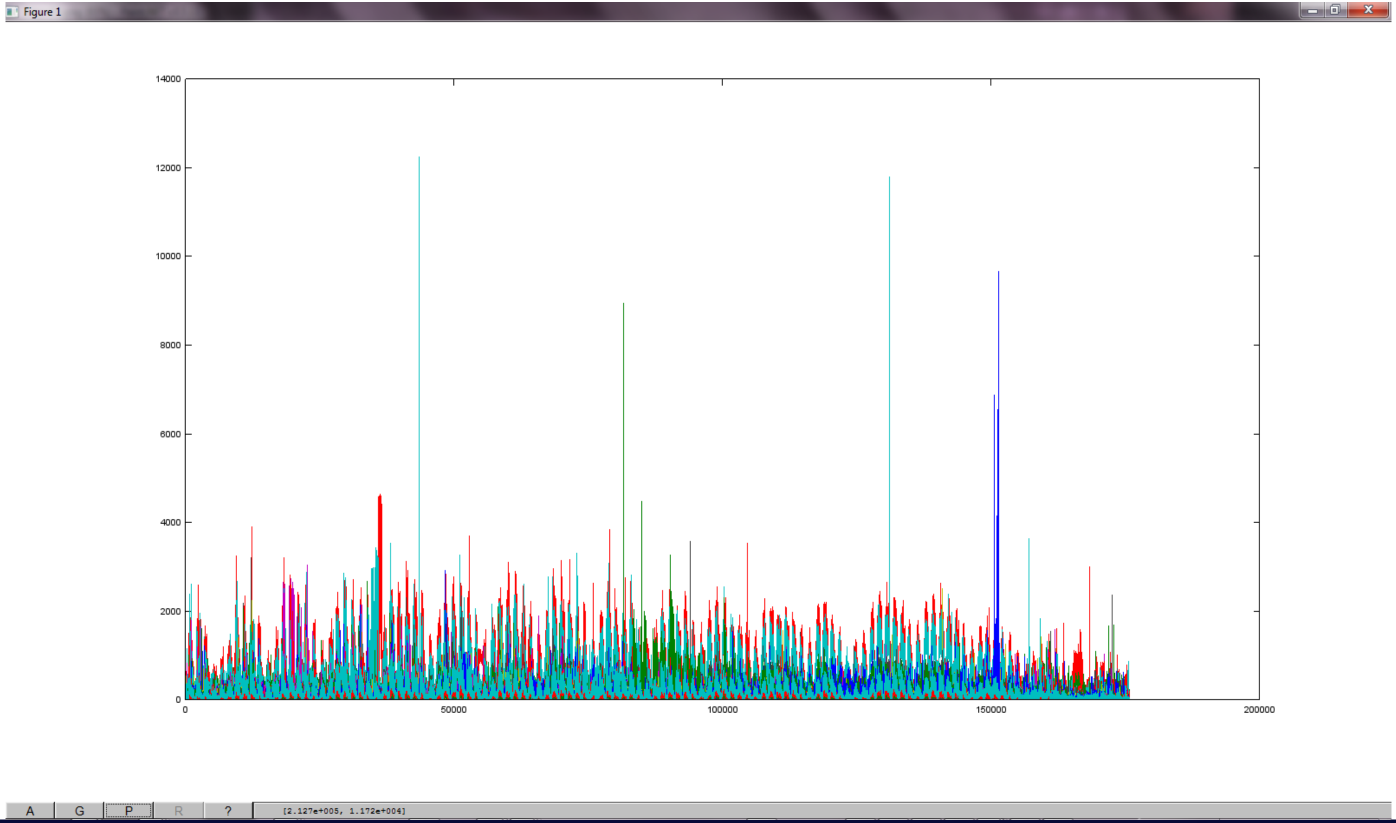
Watching screens cannot scale + it’s useless

Human brains are good at detecting patterns

Even subtle ones

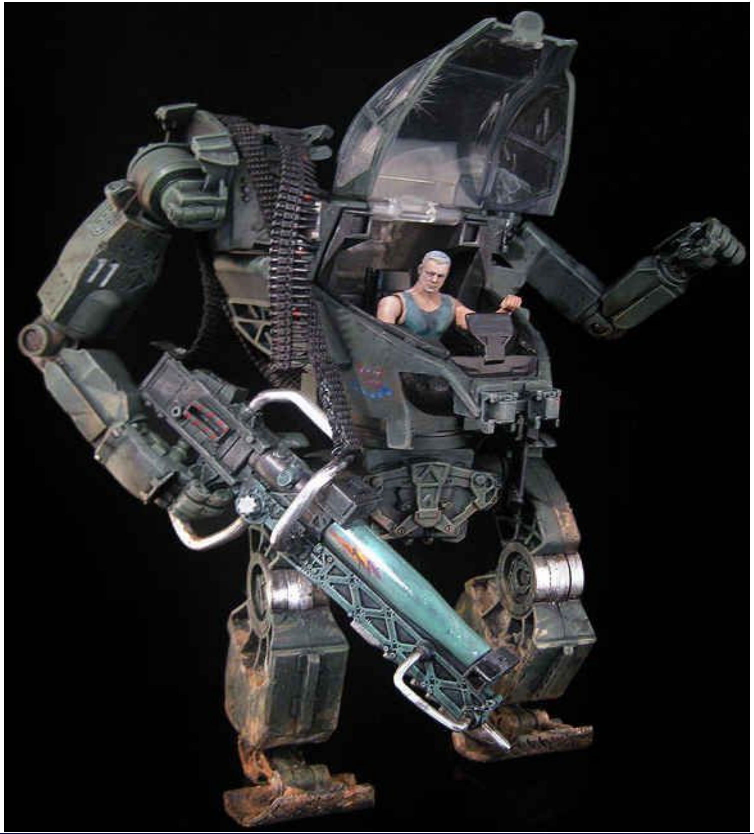
Computers suck at it

OTOH, humans get lost in volume and details

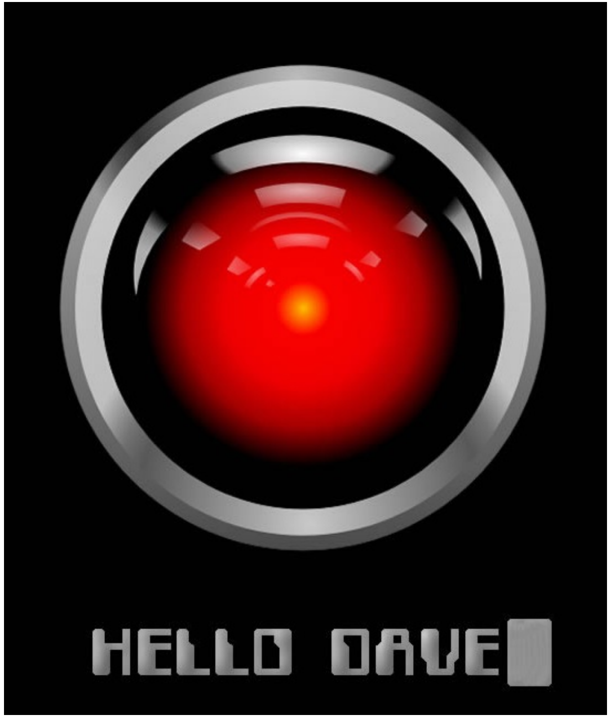


Current IT fire fighting situation

Need the cognitive equivalent of THIS!



But NOT necessarily turn things over completely to the machines!



Synergy? (I KNEW I could sneak that word in!)

- Challenge:
  - Can we have the machines do the high volume drudge work and allow the humans to exercise judgement and high level reasoning?

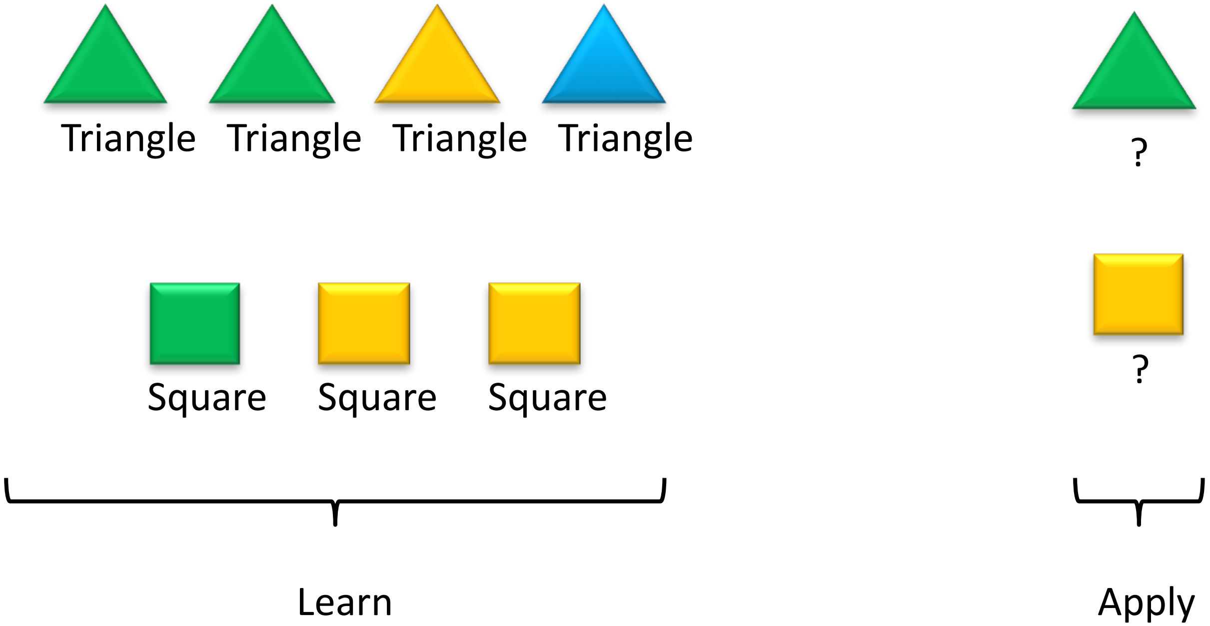
Enter Machine Learning!

What: “Field of study that gives computers the ability to learn without being explicitly programmed” – Arthur Samuel, 1959

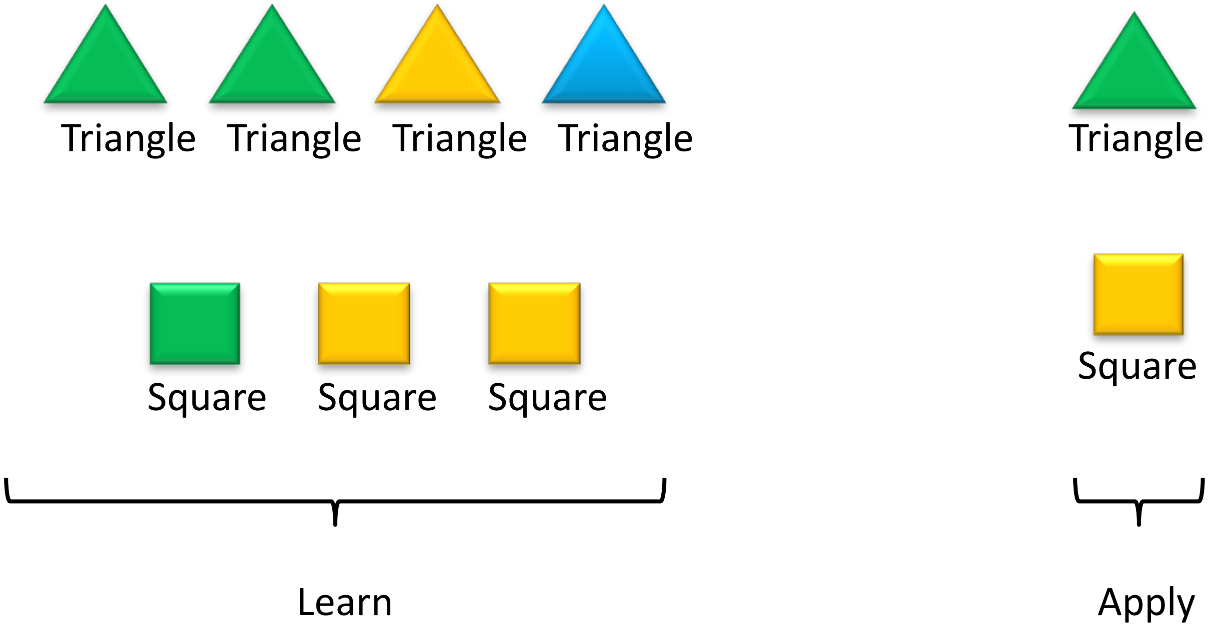
How: Generalizing (learning) from examples (data)

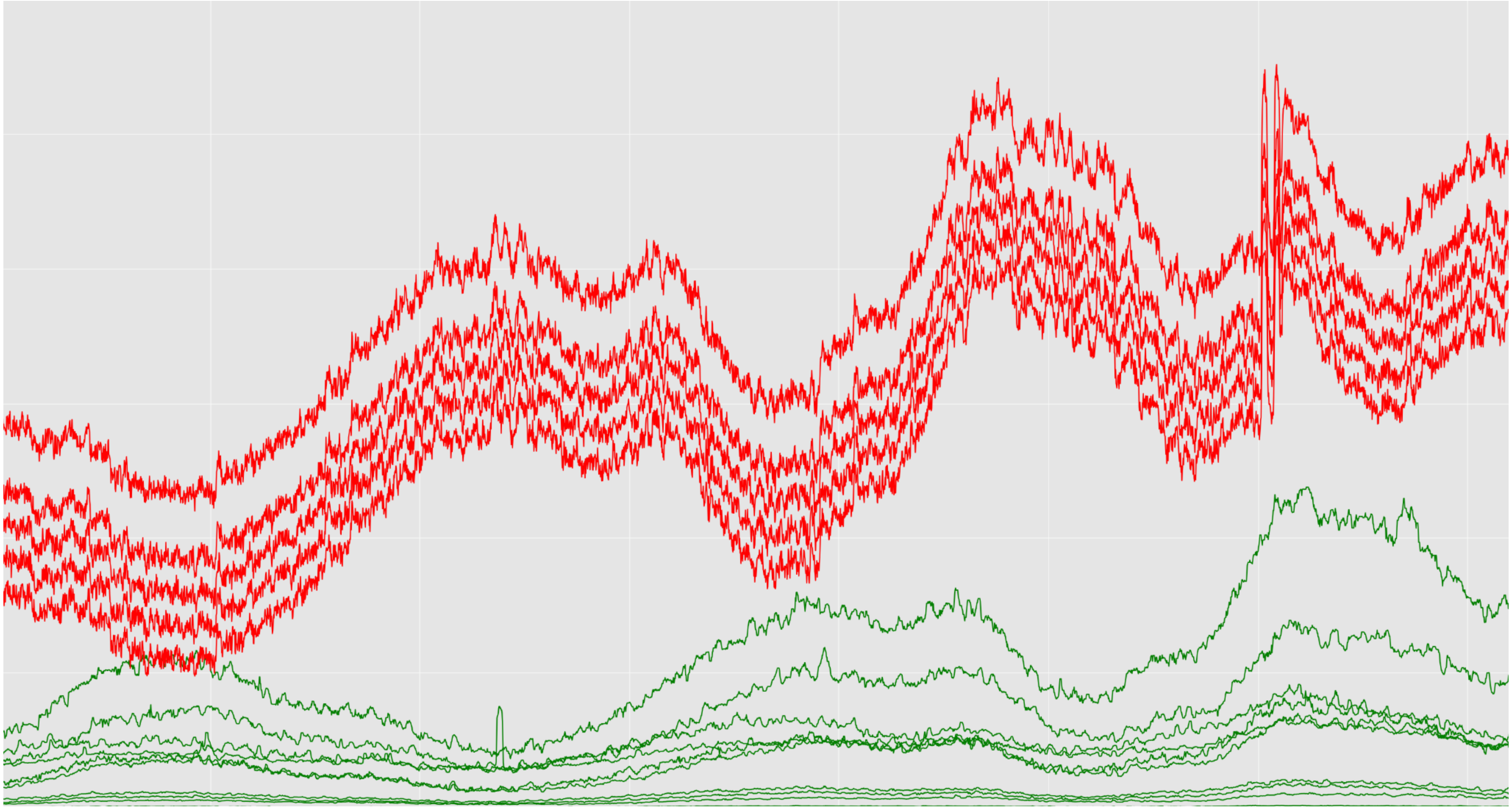
What is ML used for?

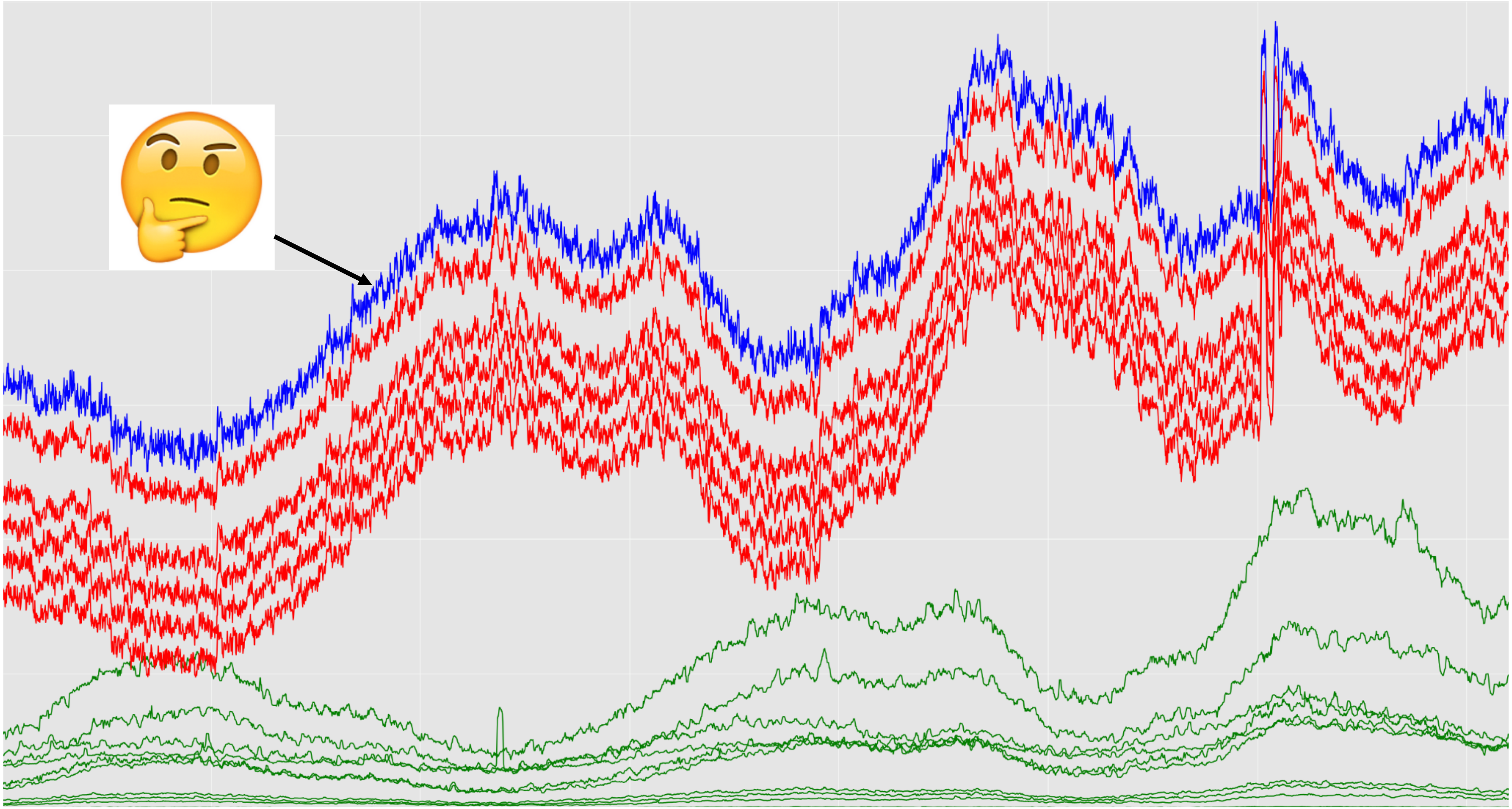
Classification: Applying labels



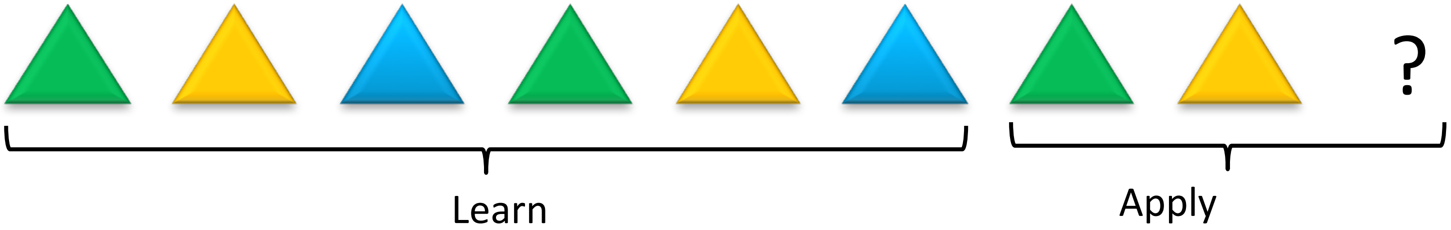
Classification: Applying labels



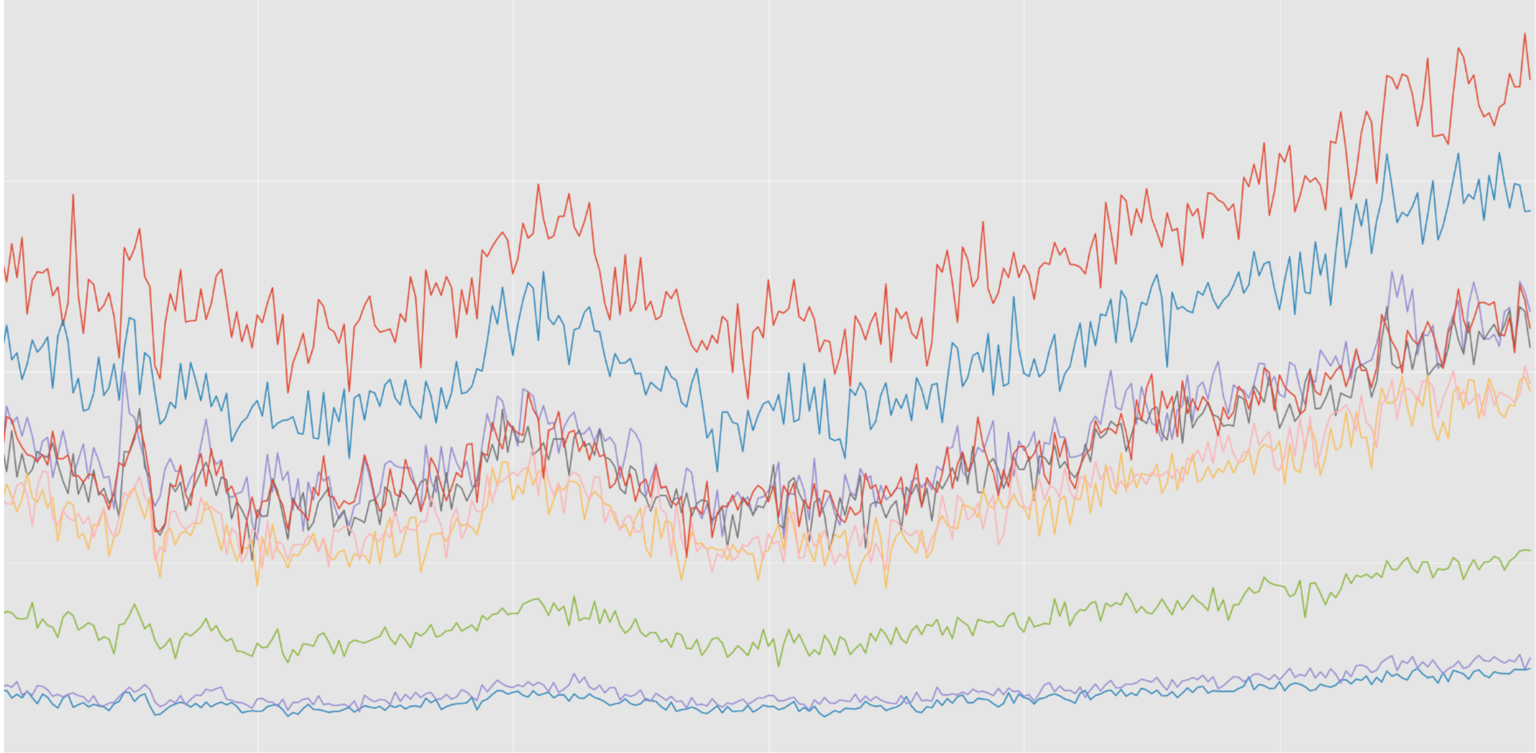




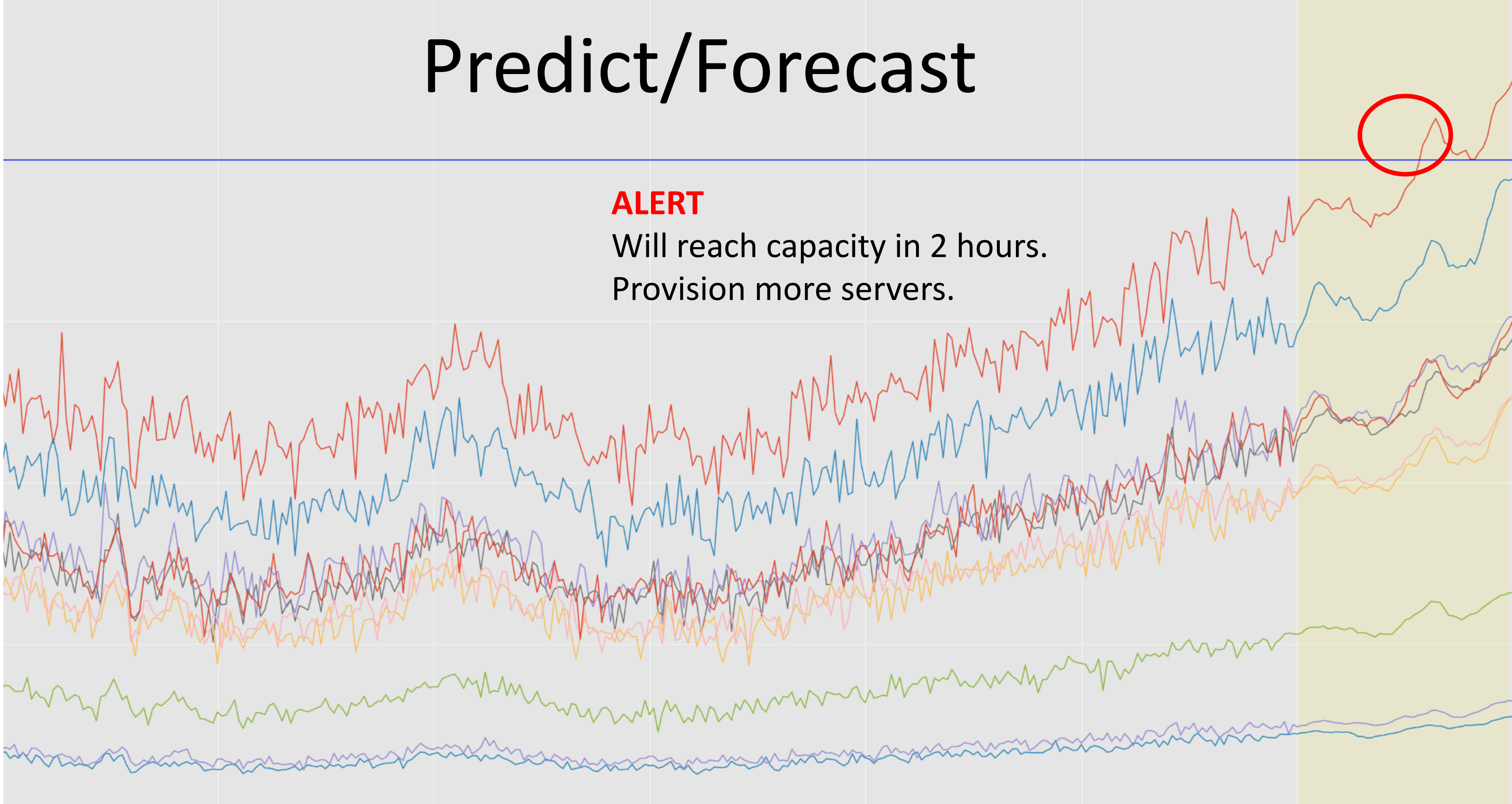
Predict/Forecast



Predict/Forecast

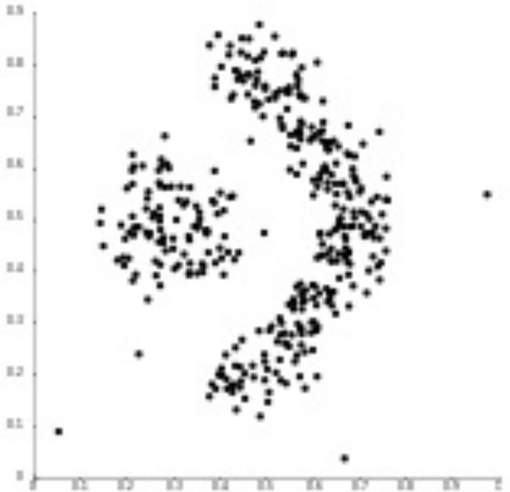


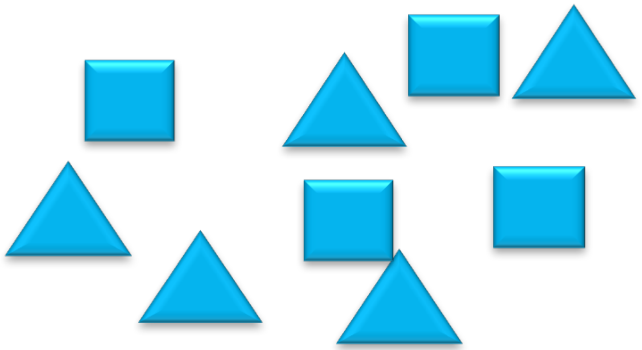
Predict/Forecast



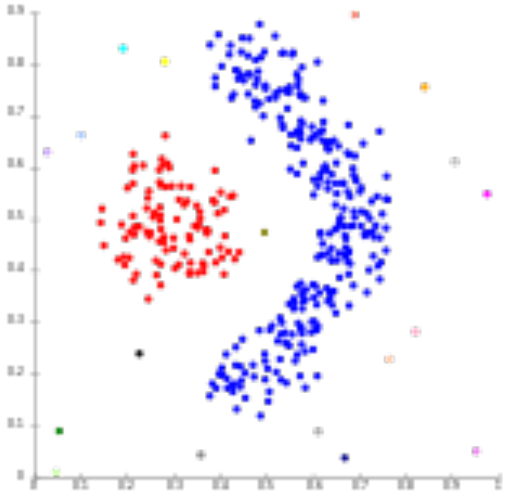
ALERT
Will reach capacity in 2 hours.
Provision more servers.

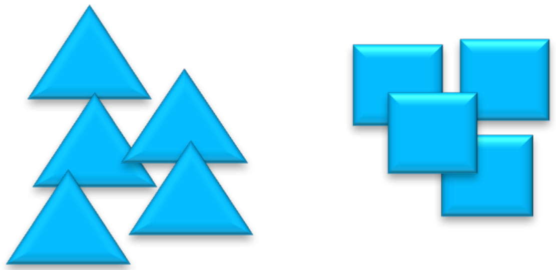
Clustering: Grouping similar things

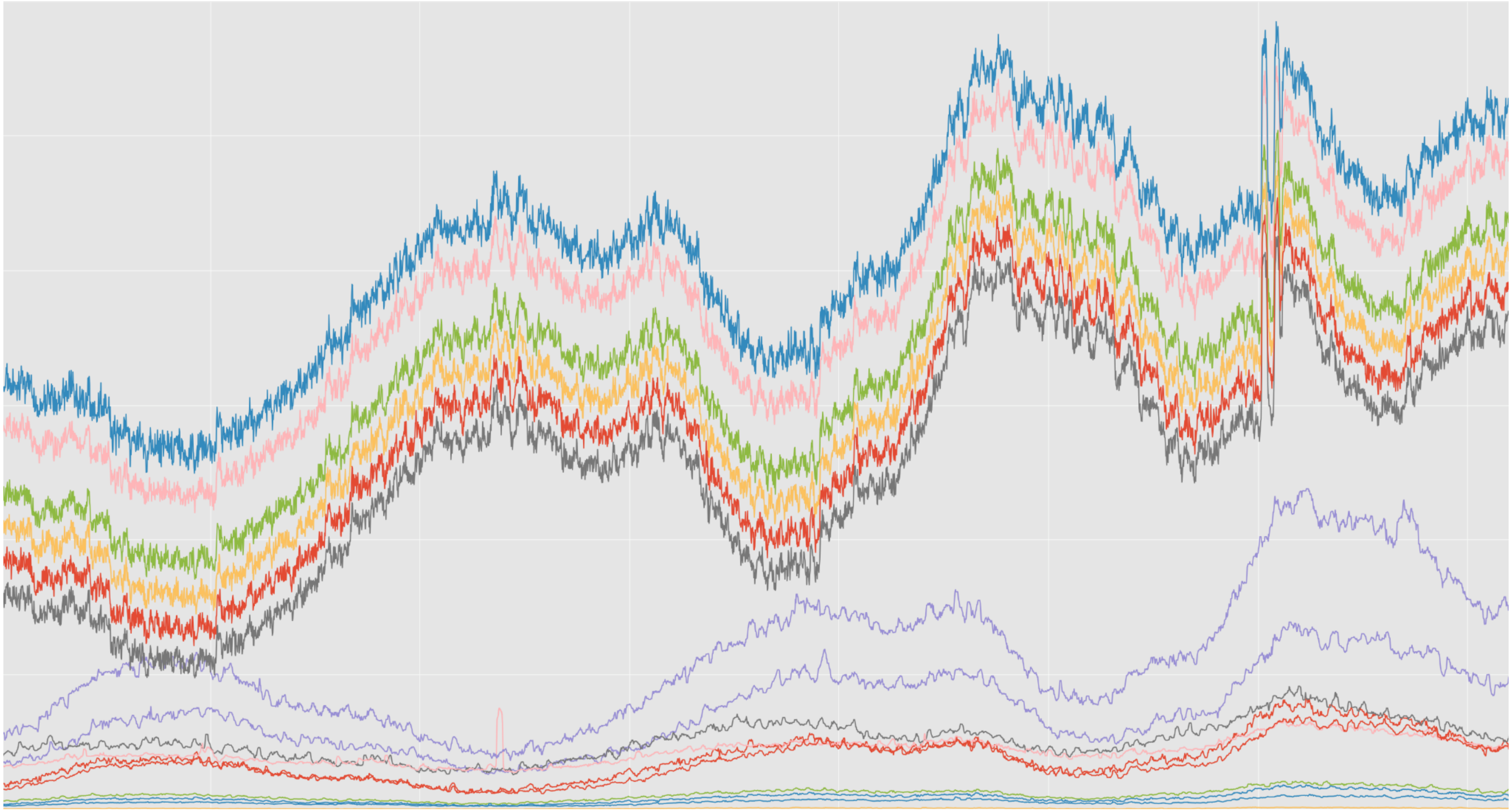


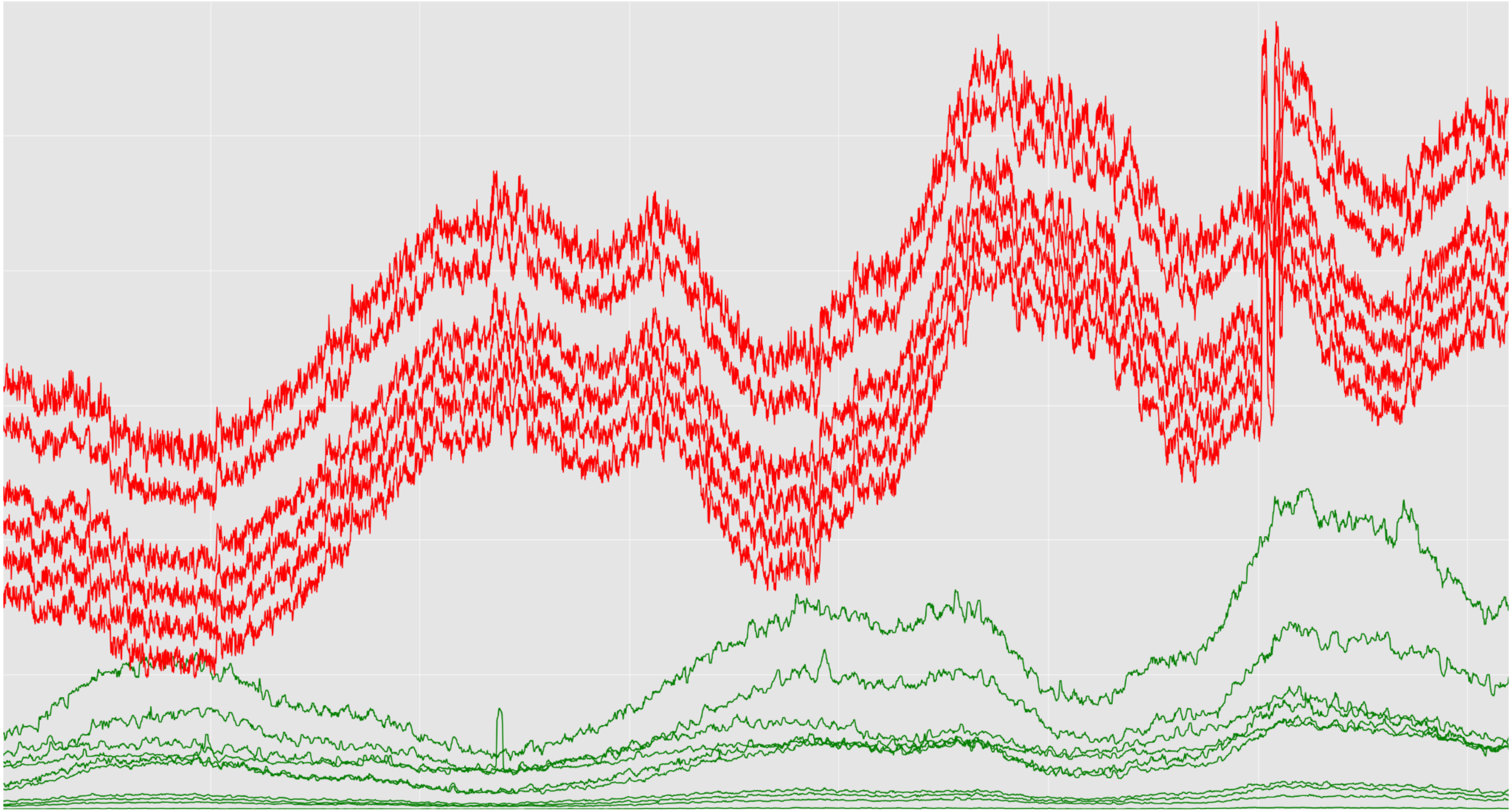


Clustering: Grouping similar things

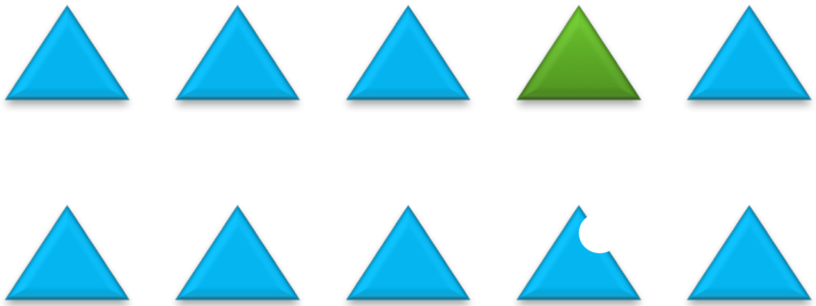


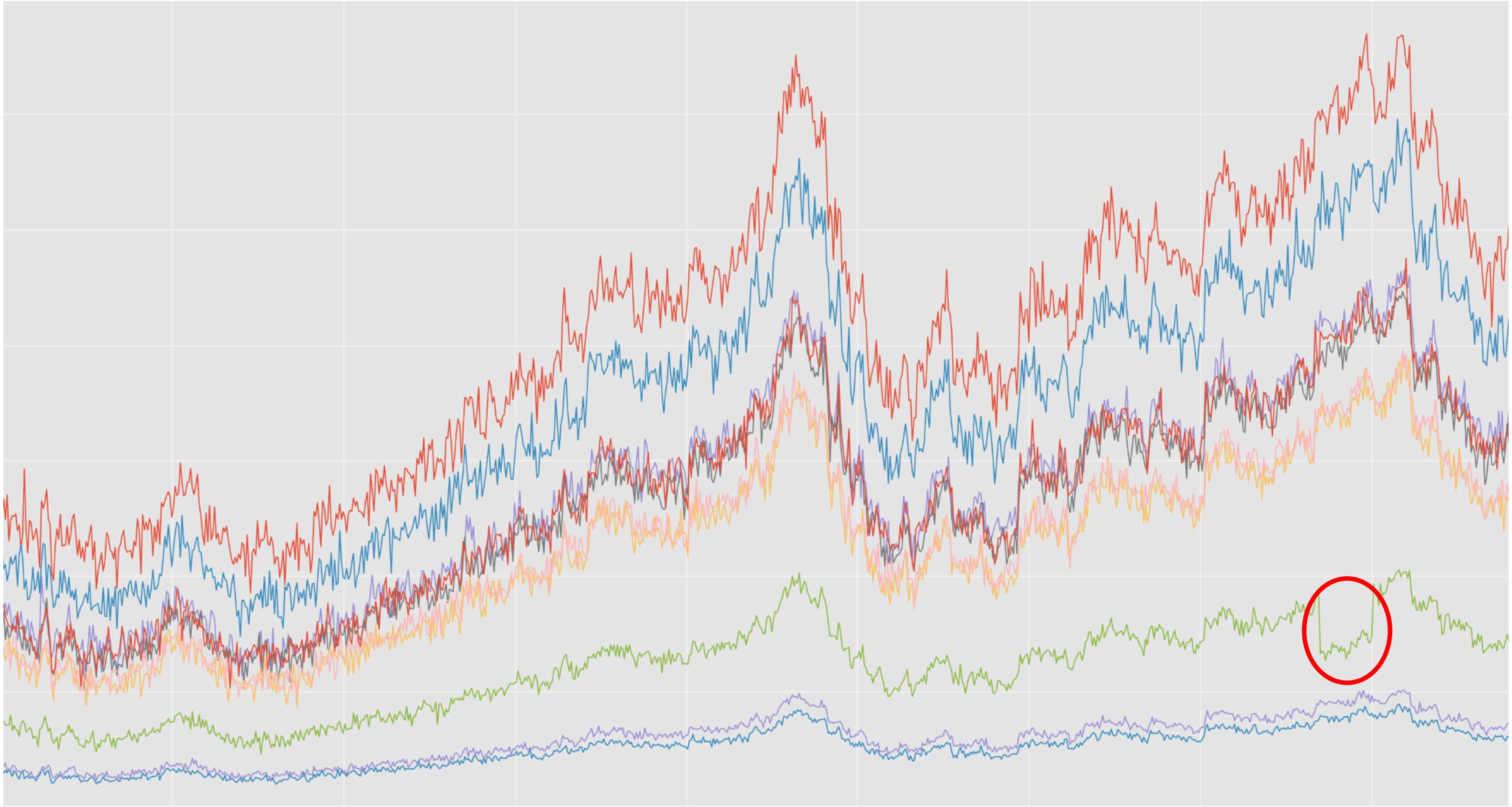






Anomaly Detection: Find unusual stuff





Real world commercial applications

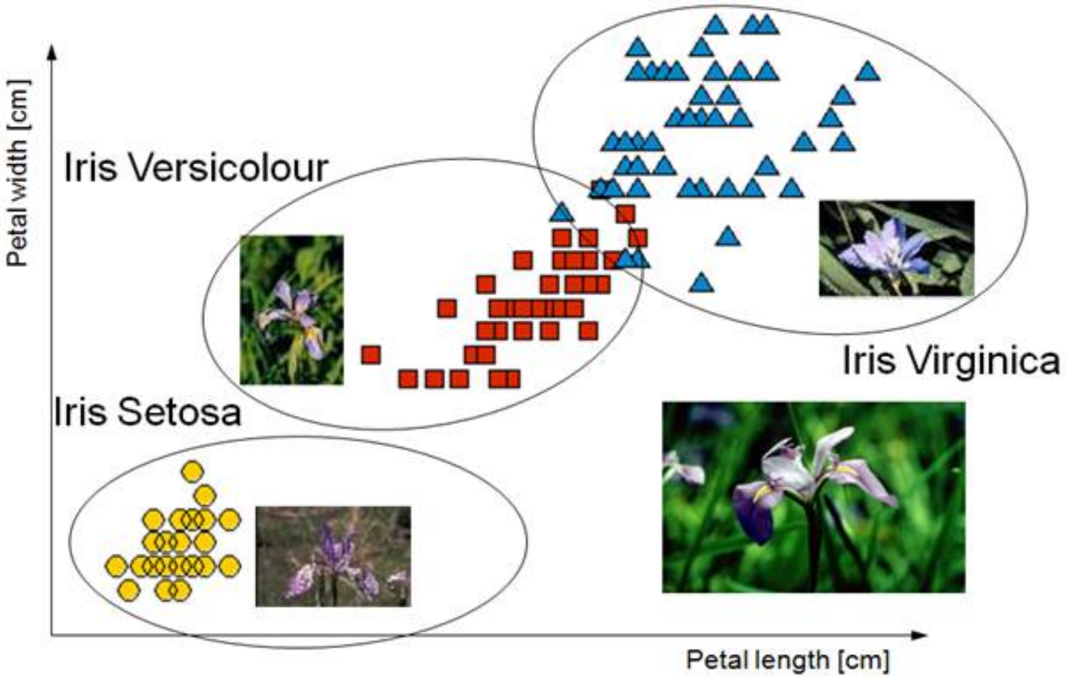
- Fraud: credit card fraud, spam, DLP
- Automated recognition: face, handwriting
- Capacity planning: product stocking, server provisioning
- Anomaly detection for security and IT Operations
- Product recommendations
- Customer segmentation
- Medical diagnoses
- …

Types of Learning

Supervised Learning

- In ML, Supervised Learning is the general set of techniques for inferring a model from a set of observations:
  - Observations in a Training Set are labelled with the desired outcomes (e.g. “normal vs. anomalous”, “normal vs. fraudulent”, “red/green/yellow”, etc)
  - As observations are fed into the learning system, it learns to differentiate by inferring a model based on these labels
  - Once sufficiently “trained”, the system is used in production on “real” unlabelled data and can label the new data based on the inferred model

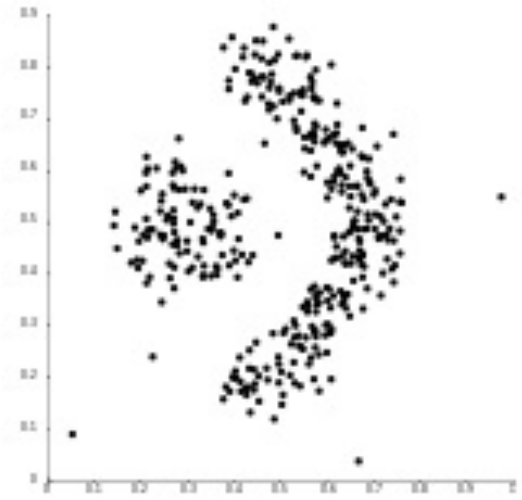
Supervised Learning example



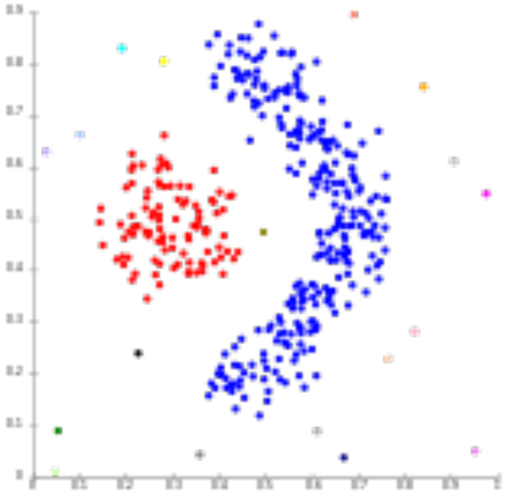
Unsupervised Learning

- In Unsupervised Learning, the system is tasked with inferring a model without having access to a set of labeled examples
  - Much harder in general
  - Well-suited to tasks where data labeling is not possible or practical: clustering, self-driving cars ☺

Unsupervised Learning example



Unsupervised Learning example



Reinforcement Learning

- System is rewarded (or punished) based on the outcomes it generates
  - Action leads to a change in the state of the world and generates an error score

Statistical Learning

- Machine Learning is not all about Neural Networks, Deep Learning,
- Large portion of ML in practice today is statistical in nature:
  - Linear regression, logistic regression
  - Three-sigma
  - Kolmogorov-Smirnov test
  - Holt-Winters and exponential smoothing
  - K-means, k-nearest neighbors
  - Support Vector Machines
  - Random trees, random forests
  - …

Flavor of Statistical ML:

Three Things to Remember for Anomaly Detection

.conf2016

splunk>

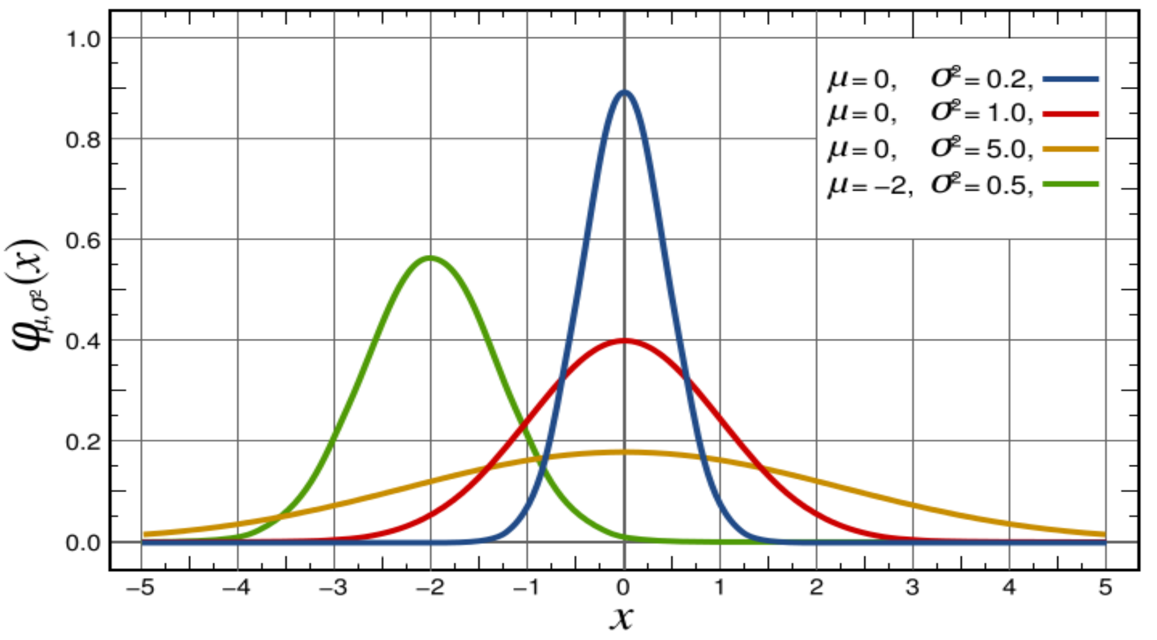
Thing 1:
Your data is NOT necessarily Gaussian

.conf2016

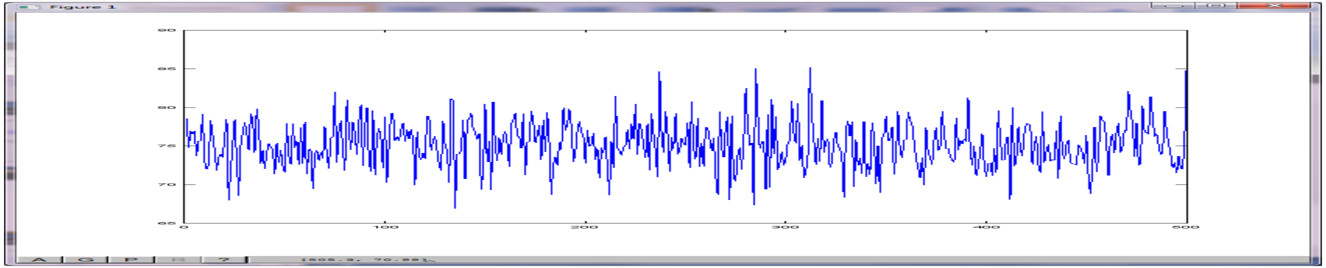
splunk>

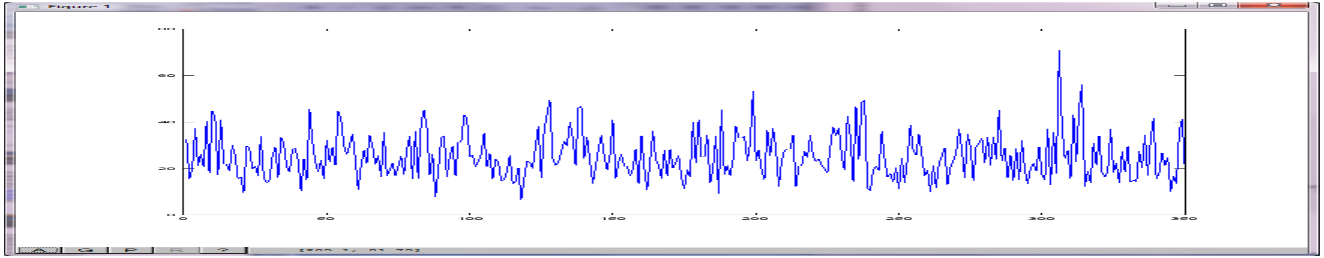
Gaussian or Normal distribution

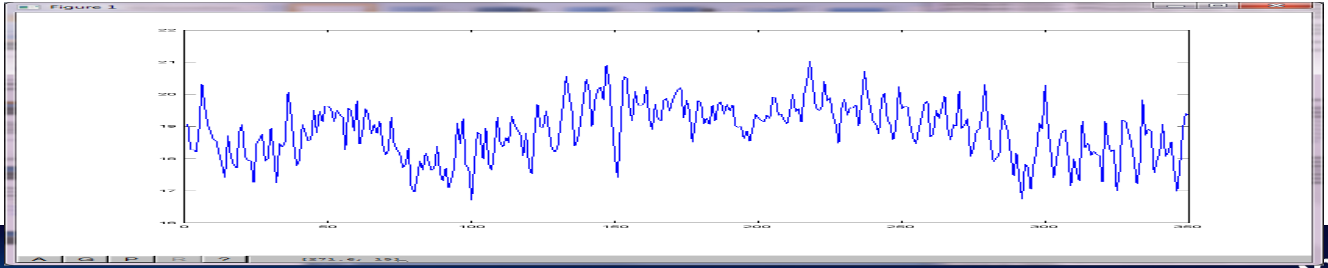
- Bell-shaped distribution



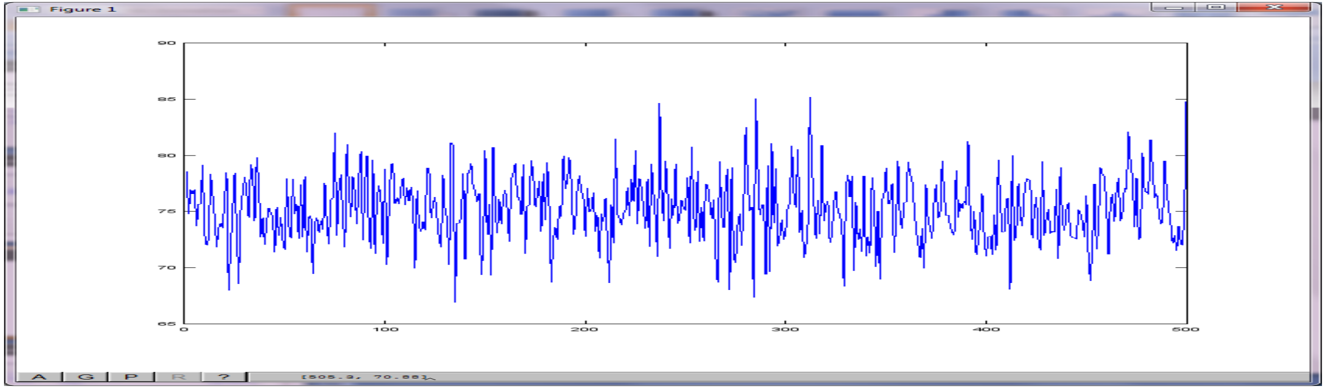
Can you tell?

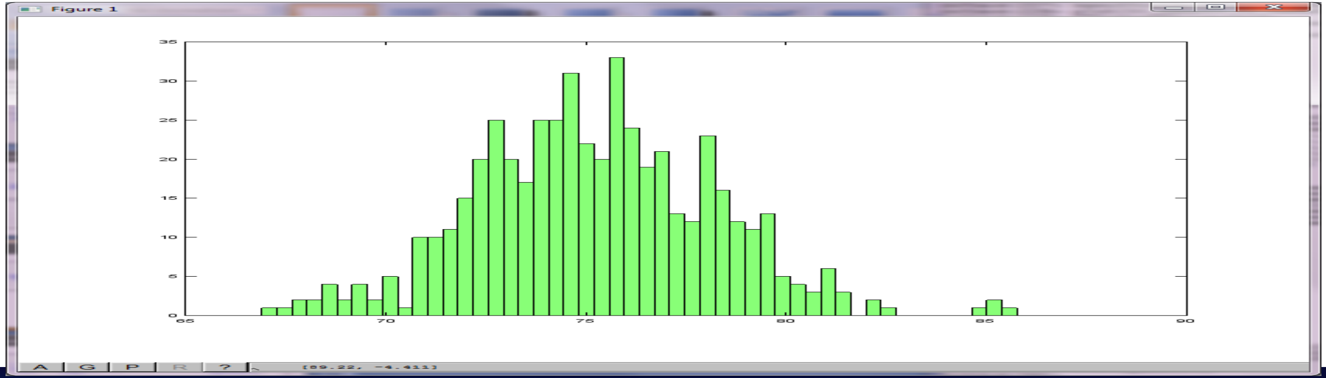






THIS is normal





This isn't

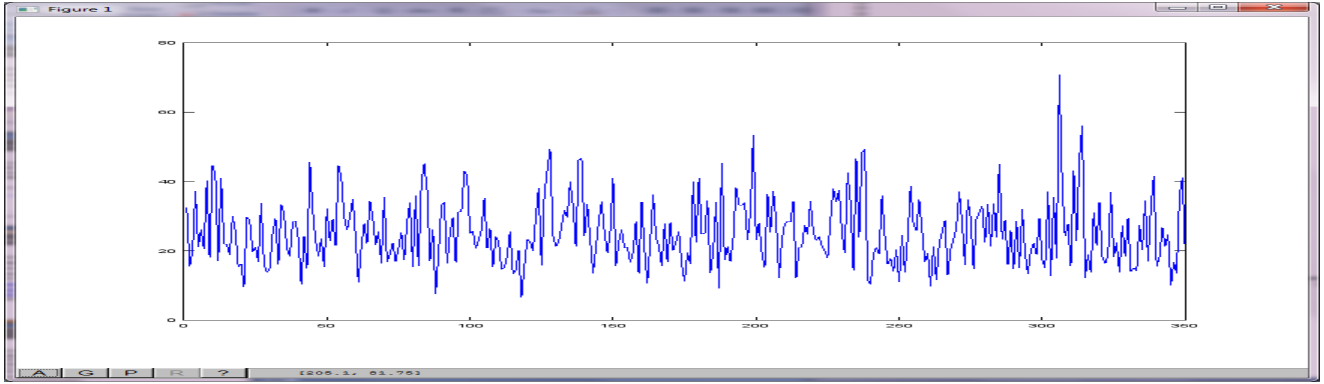


Figure 1
80
60
40
20
0
0
50
100
150
200
250
300
350

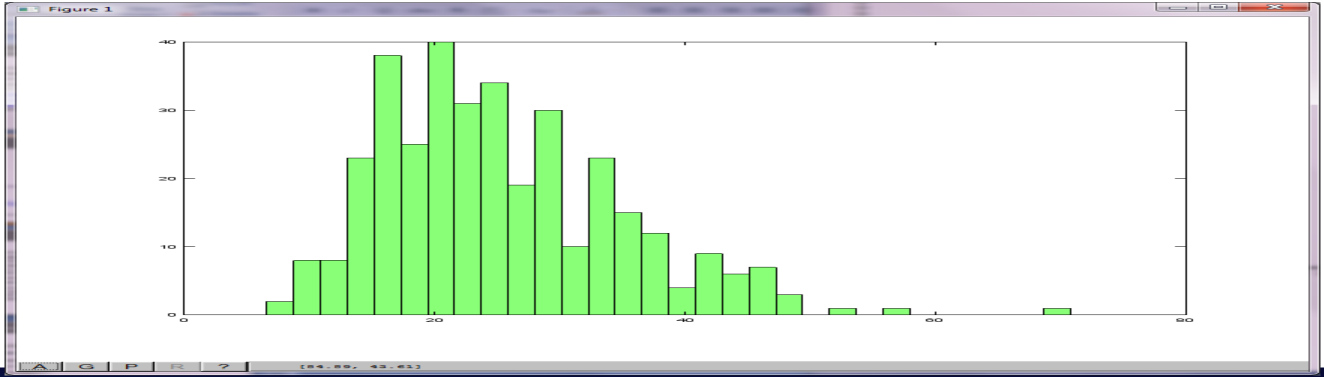
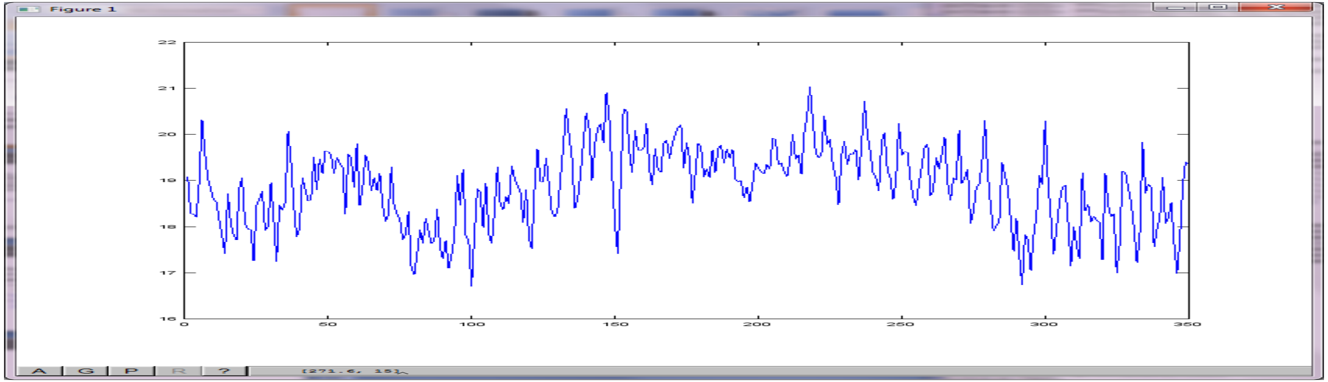
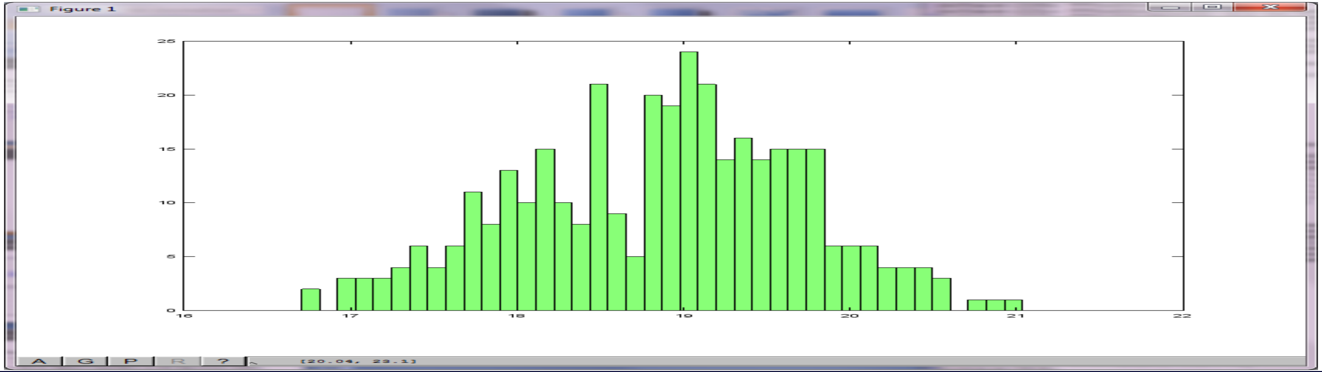


Figure 1
40
30
20
10
0
0
20
40
60
80

Neither is this





Normal distributions are really useful

- I can make powerful predictions because of the statistical properties of the data
- I can easily compare different metrics since they have similar statistical properties
- There is a HUGE body of statistical work on parametric techniques for normally distributed data

Normally distributed vs Not

Normally distributed

- Most naturally occurring processes
- Population height, IQ distributions (present company excepted of course)
- Widget sizes, weights in manufacturing
- …

Not

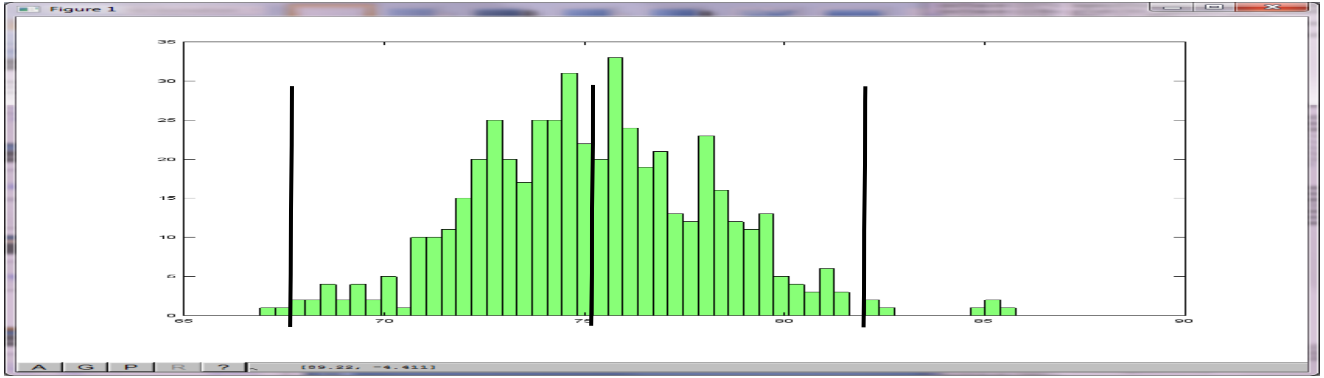
- A LOT of your data

Why is that important?

- Most analytics tools are based on two assumptions:
  1. **Data is normally distributed with a useful and usable mean and standard deviation**
  2. Data is probabilistically “stationary”

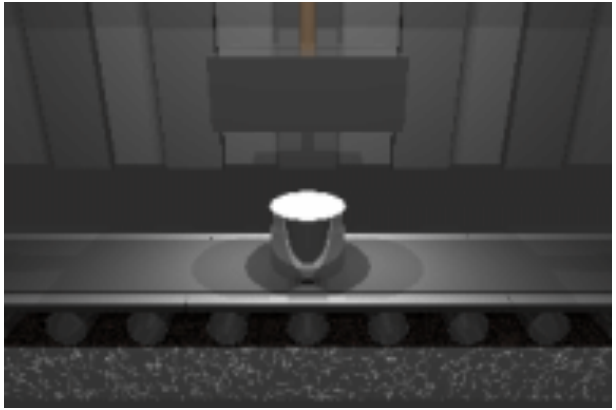
Example: Three-Sigma Rule

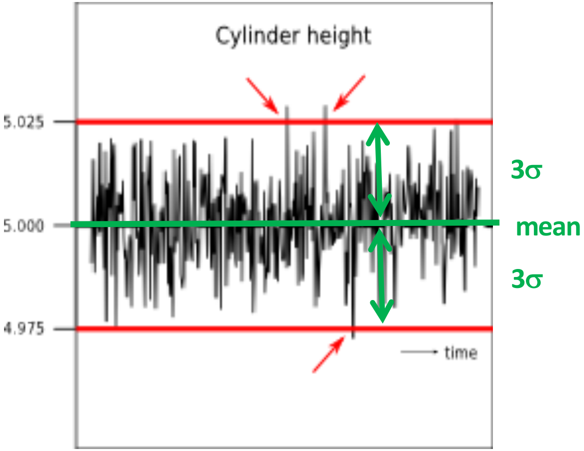
- Three-sigma rule
  - ~68% of the values lie within 1 std deviation of the mean
  - ~95% of the values lie within 2 std deviations
  - **99.73% of the values lie within 3 std deviations: anything else is considered an outlier**



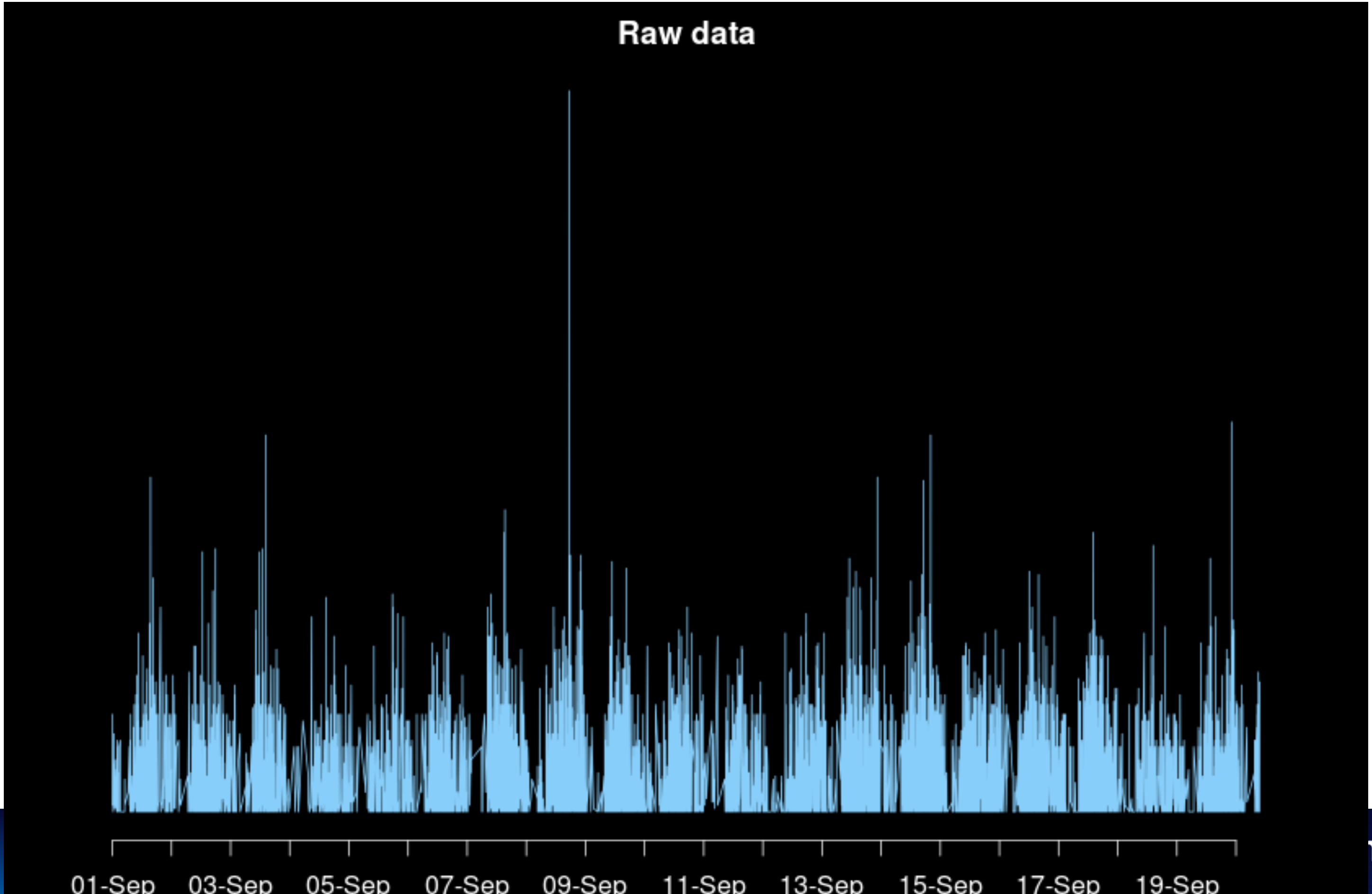
Aaahhhh

- The mysterious red lines explained

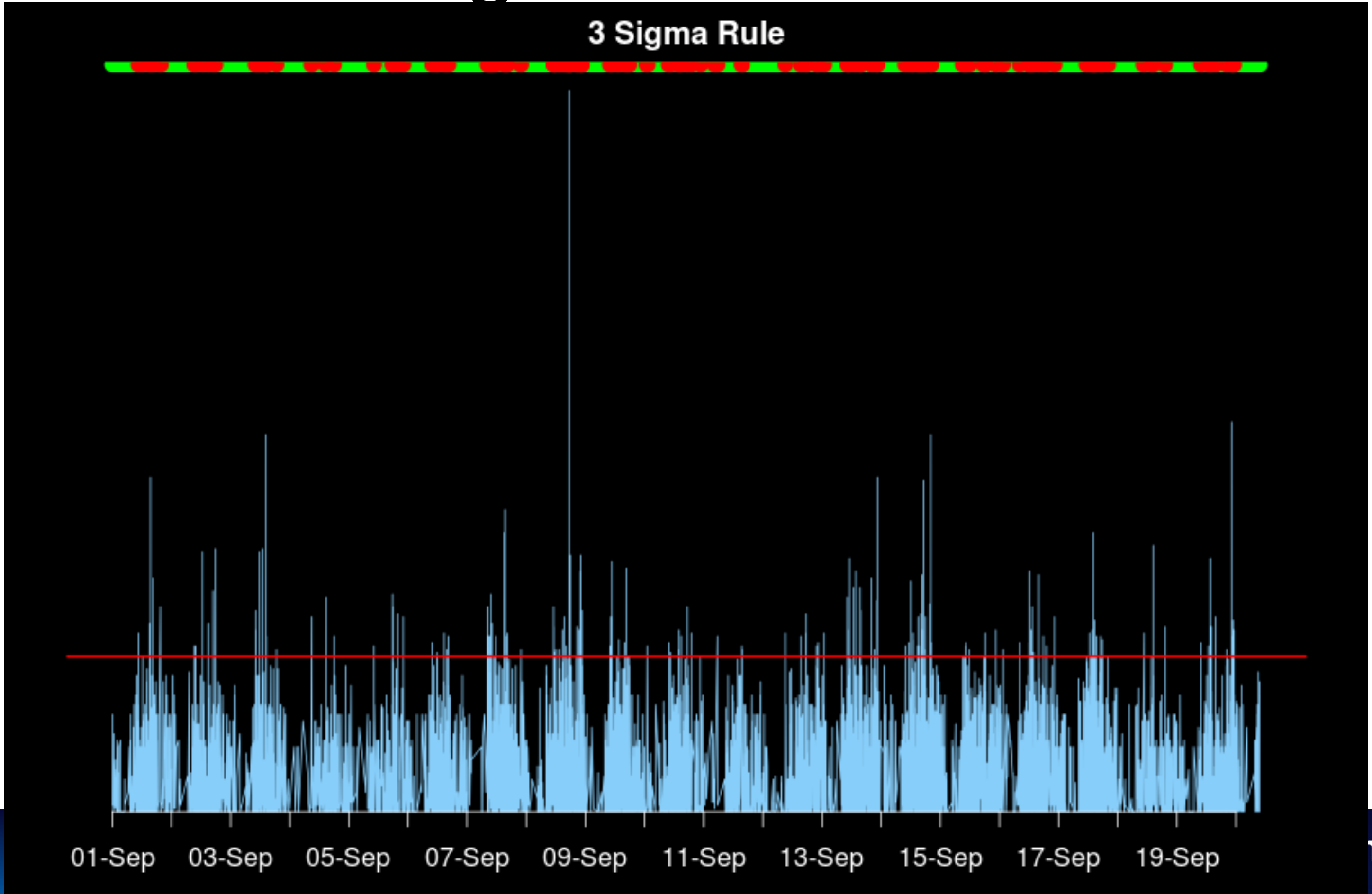




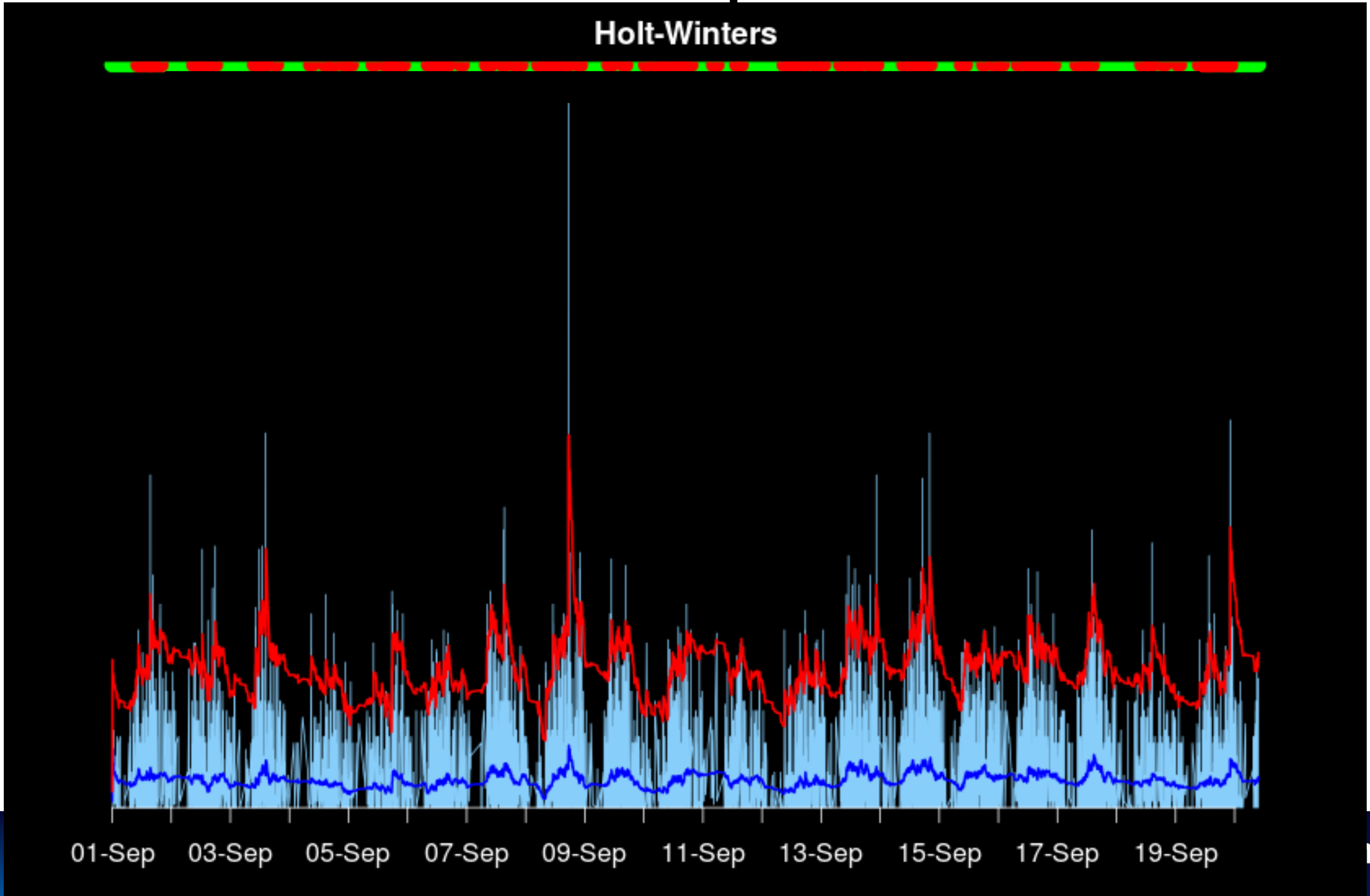
Doesn't work because THIS



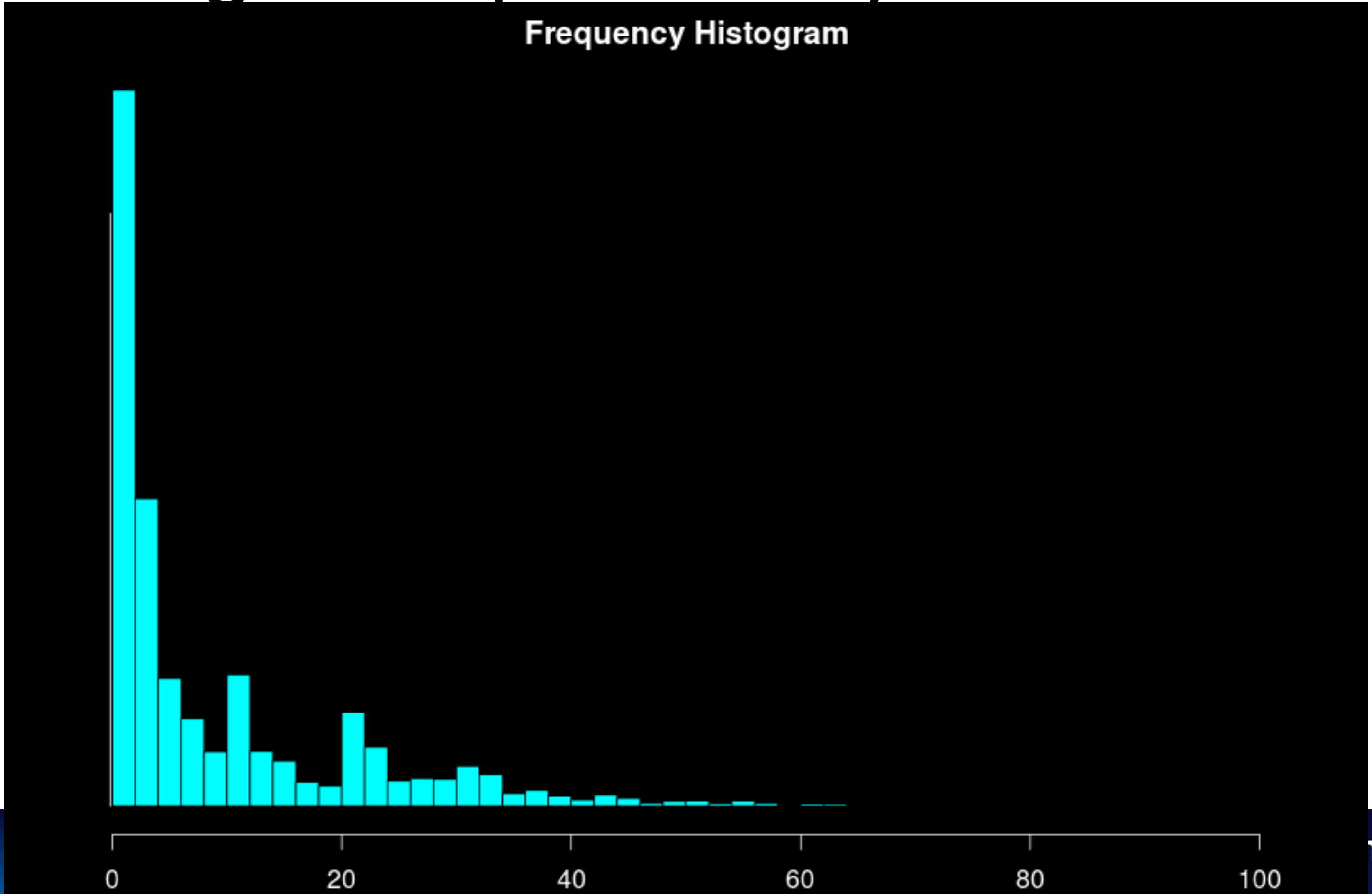
3-sigma rule alerts



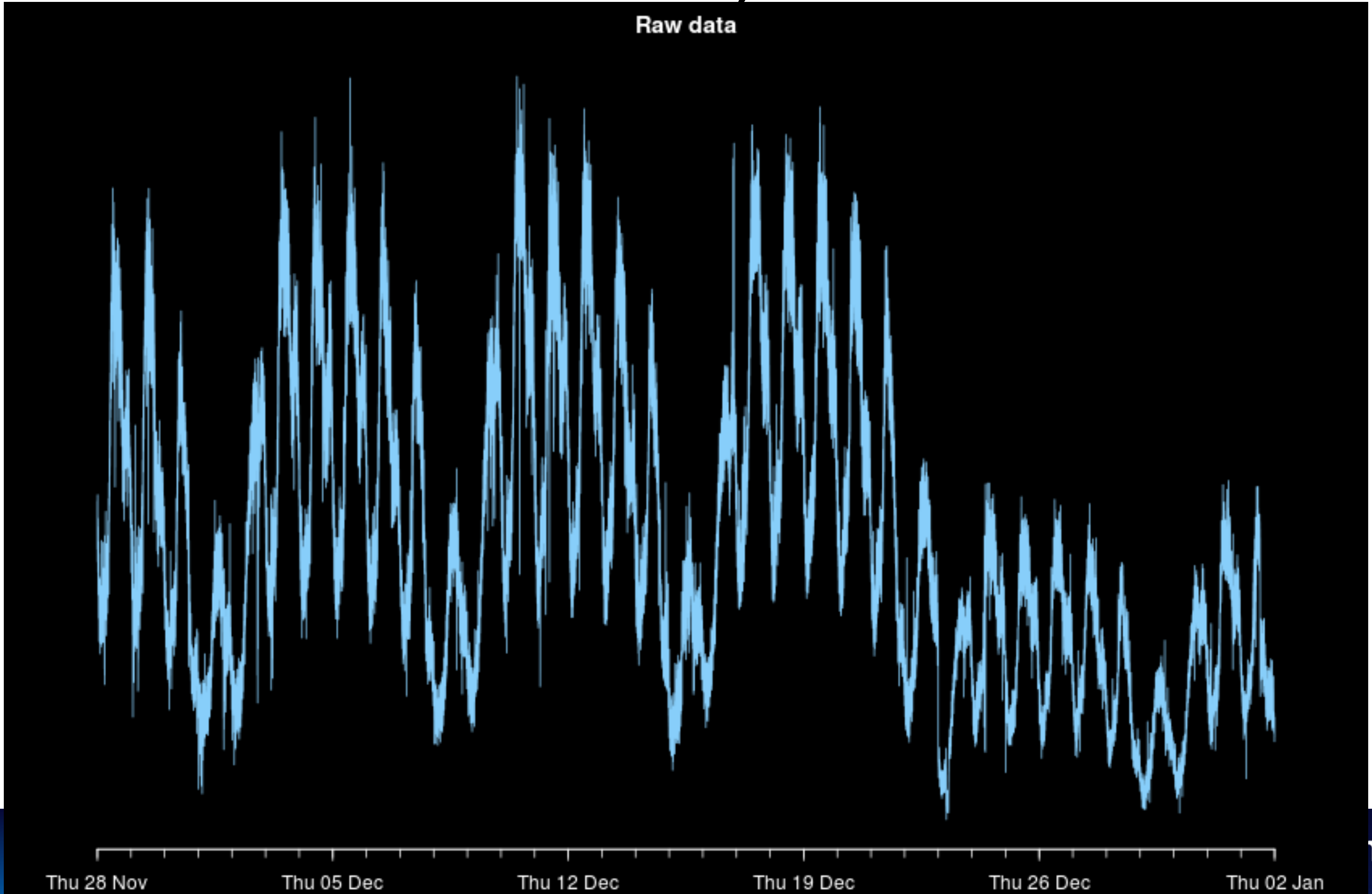
Holt-Winters predictions



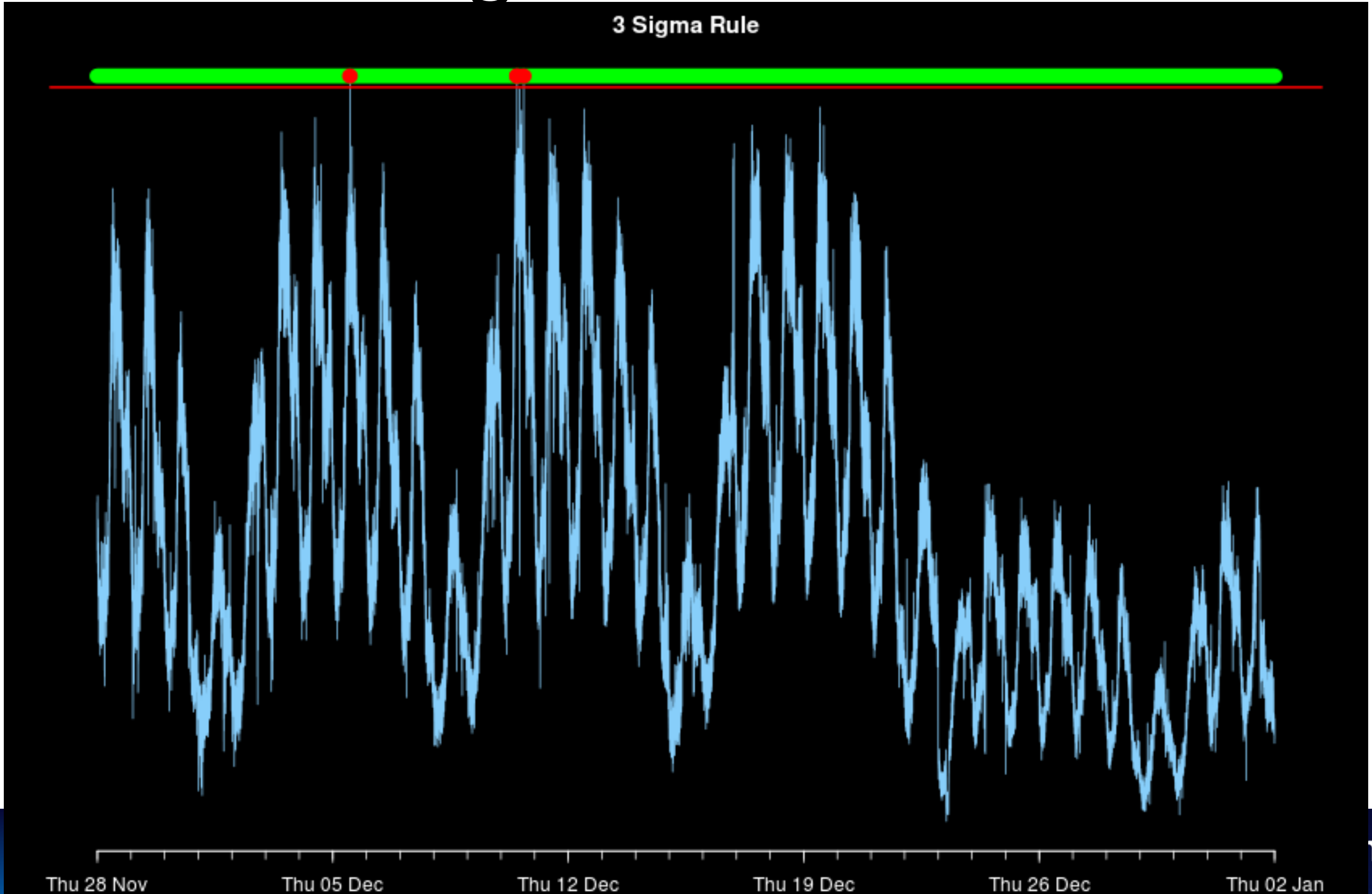
Histogram – probability distribution



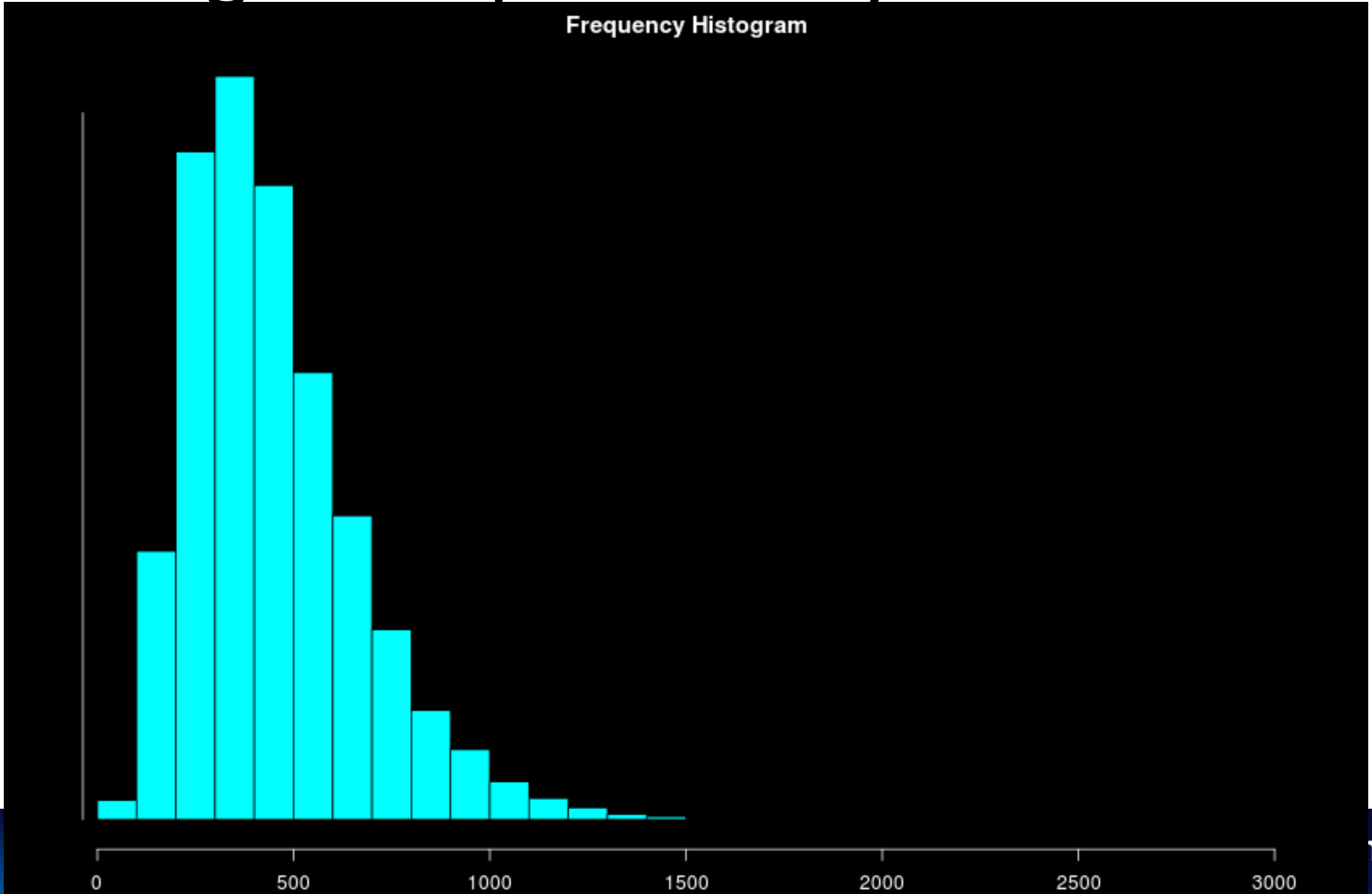
Or worse, THIS!



3-sigma rule alerts



Histogram – probability distribution



Thing 2
Saying *Kolmogorov-Smirnov*
is a great way to impress everyone

.conf2016

splunk>

Why is that important?

- Seriously!?
- Ok, actually non-parametric techniques that make no assumptions about normality or any other probability distribution are ***crucial*** in your effort to understand what's going on in your systems

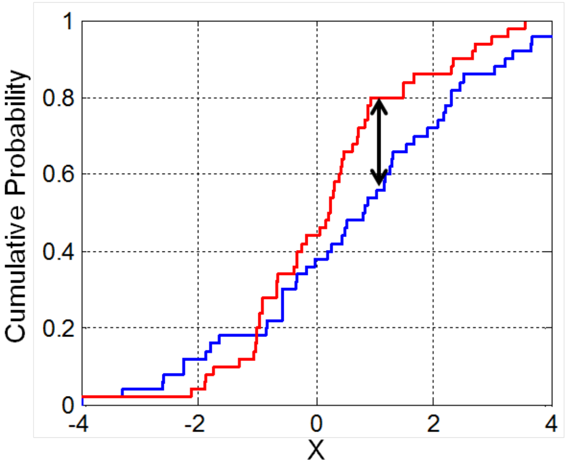
Parametric vs Non-Parametric Learning

- Parametric learning:
  - Finite, manageable number of parameters
  - Makes strong assumptions about the data (e.g. Gaussian distribution)
  - Example: Linear Regression

- Non-Parametric:
  - Large (or infinite) number of parameters
  - No assumptions about the underlying characteristics of the data
  - Example: Kolmogorov-Smirnov

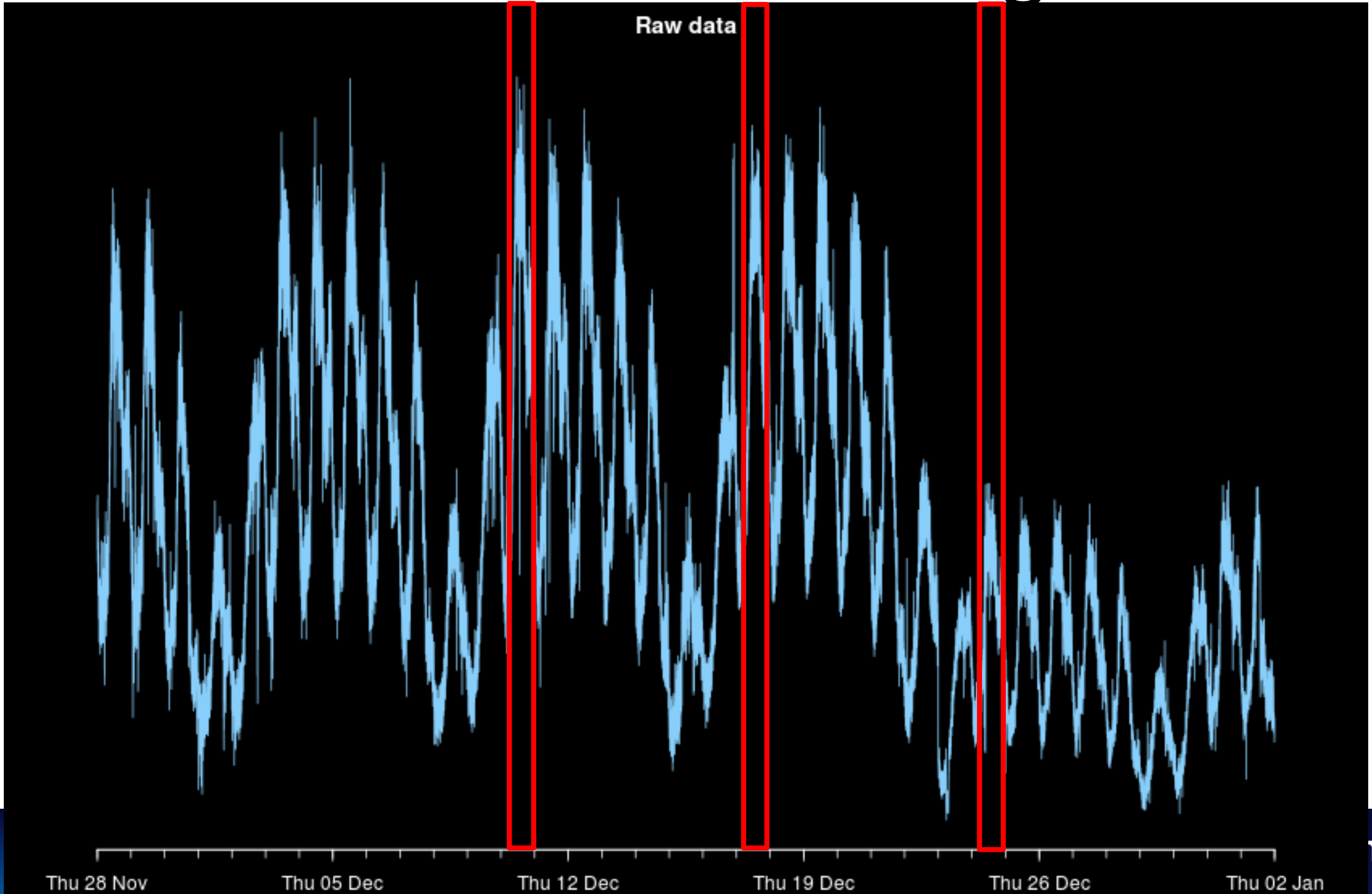
The Kolmogorov-Smirnov test

- *Non-parametric* test
  - Compare two probability distributions
  - Makes no assumptions (e.g. Gaussian) about the distributions of the samples
  - Measures maximum distance between *cumulative* distributions
  - Can be used to compare periodic/seasonal metric periods (e.g. day-to-day or week-to-week)

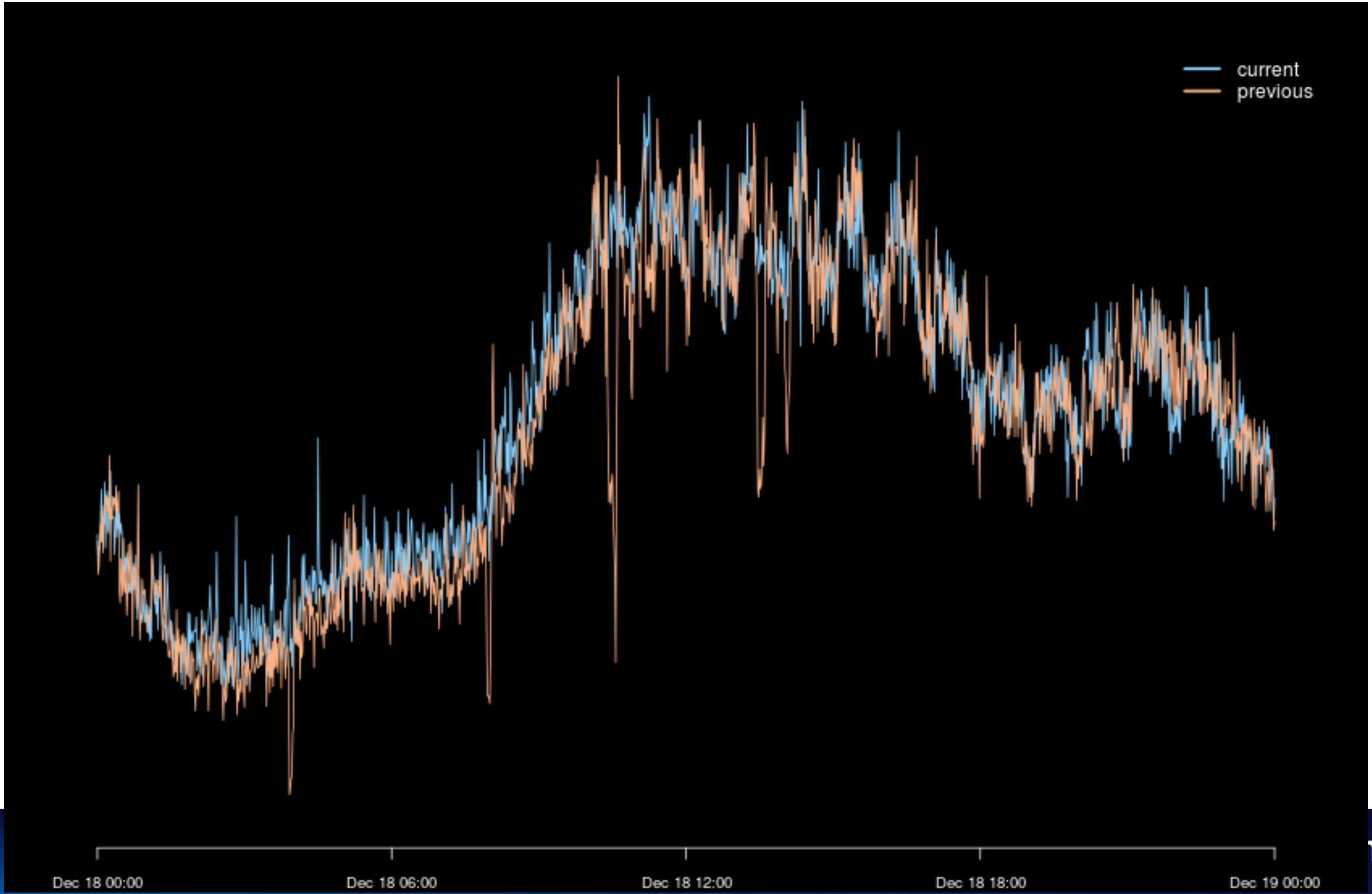


http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

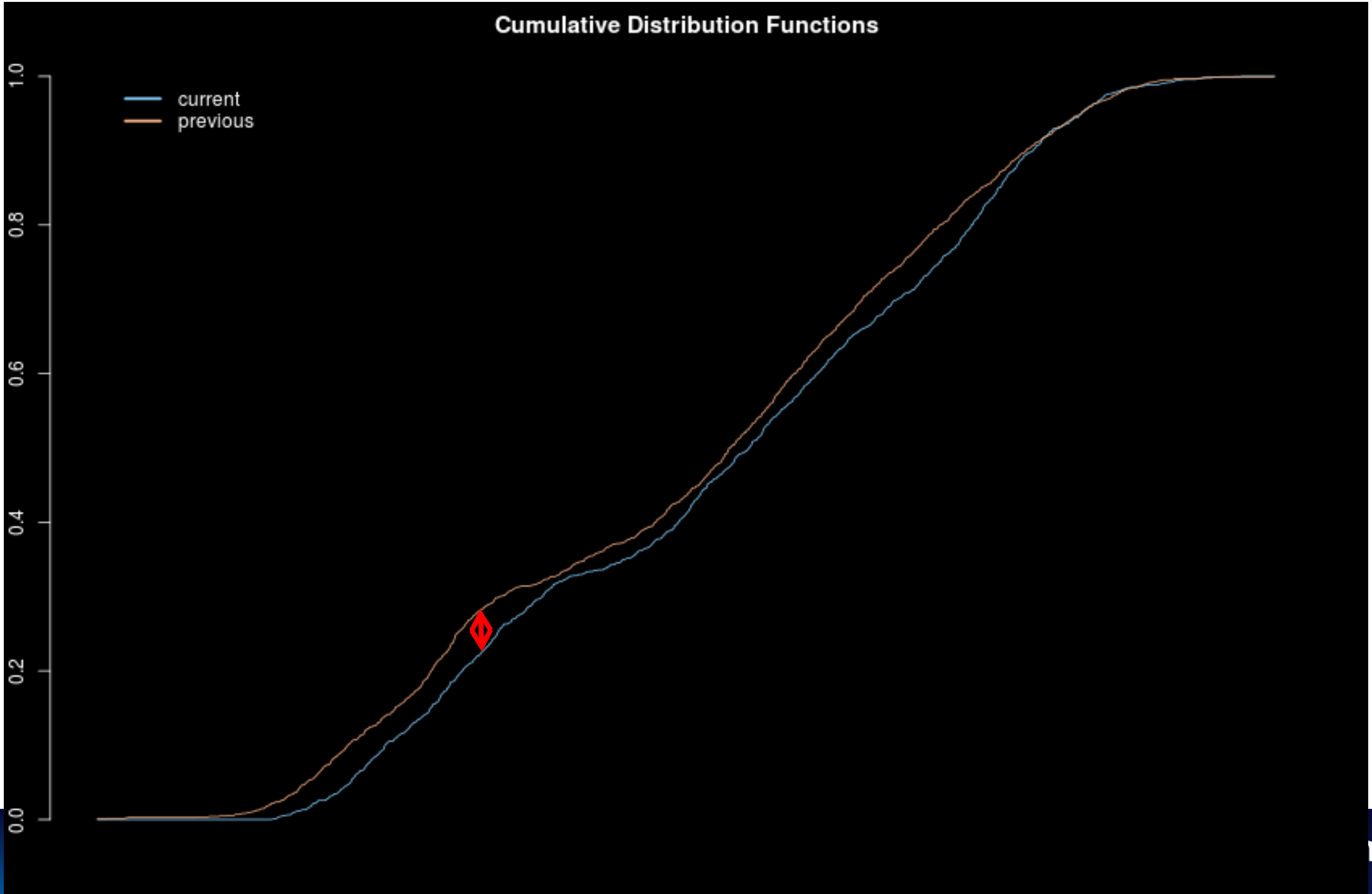
KS with windowing



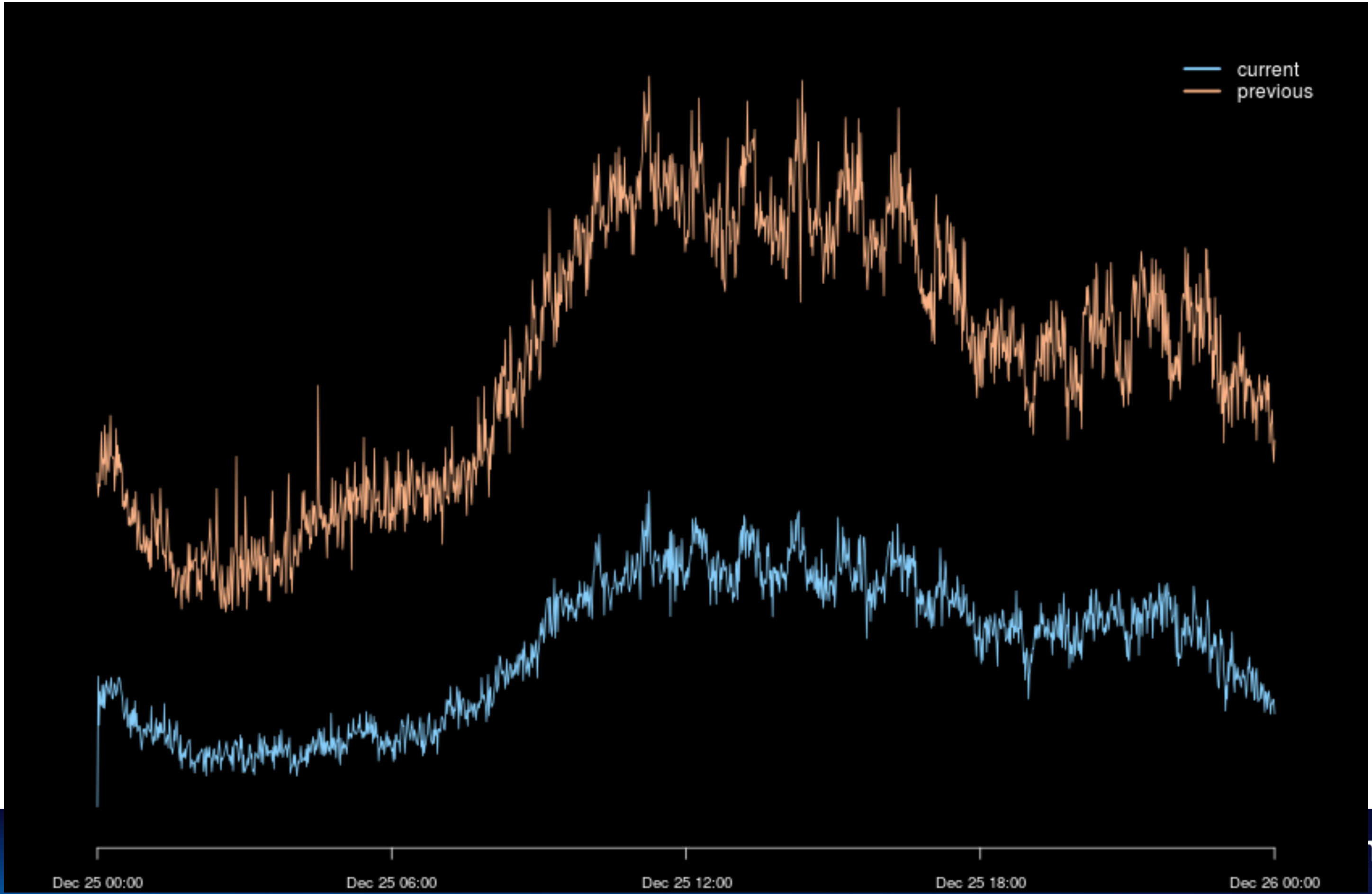
Data from similar windows



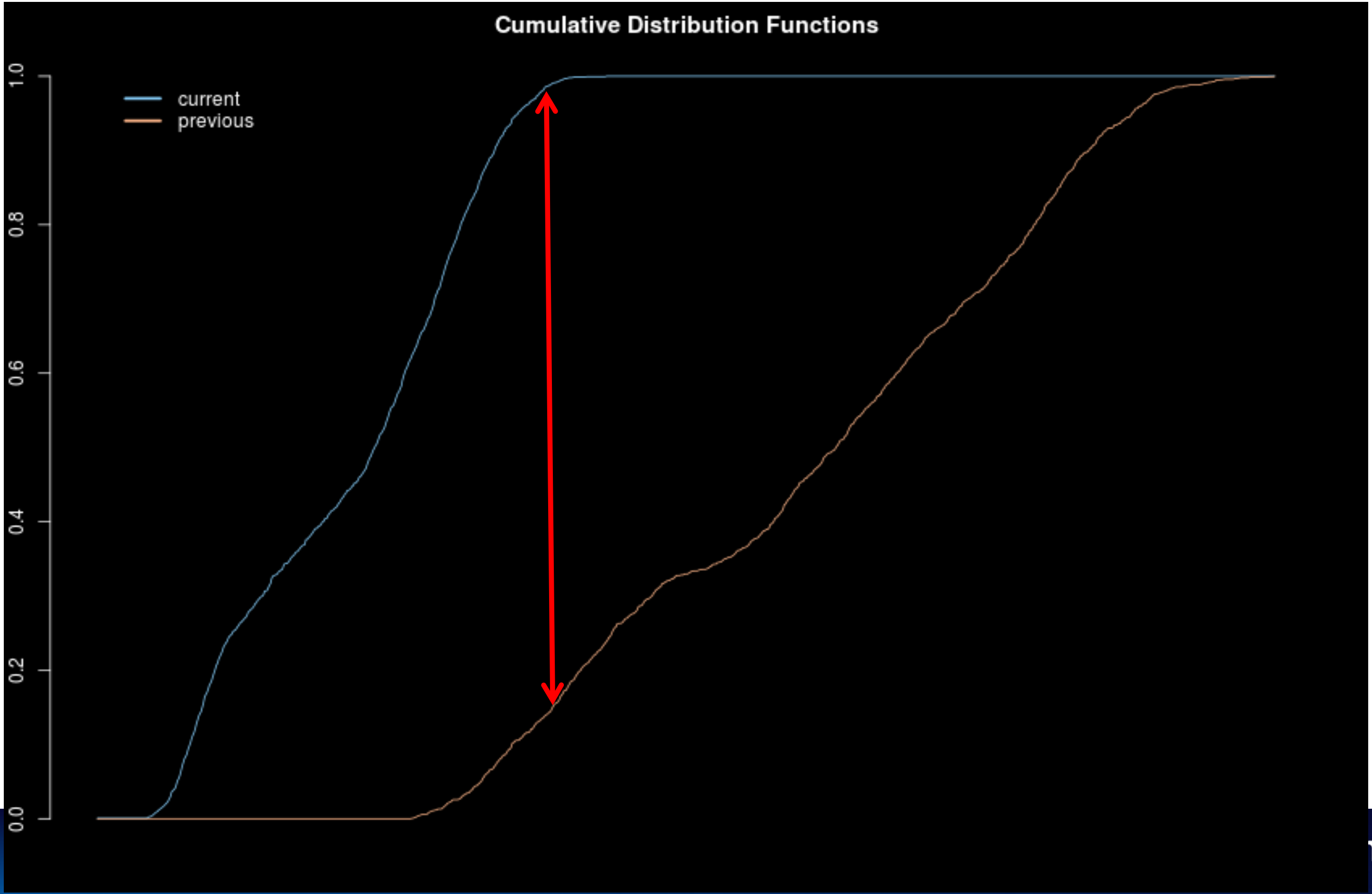
Cumulative distribution for those windows



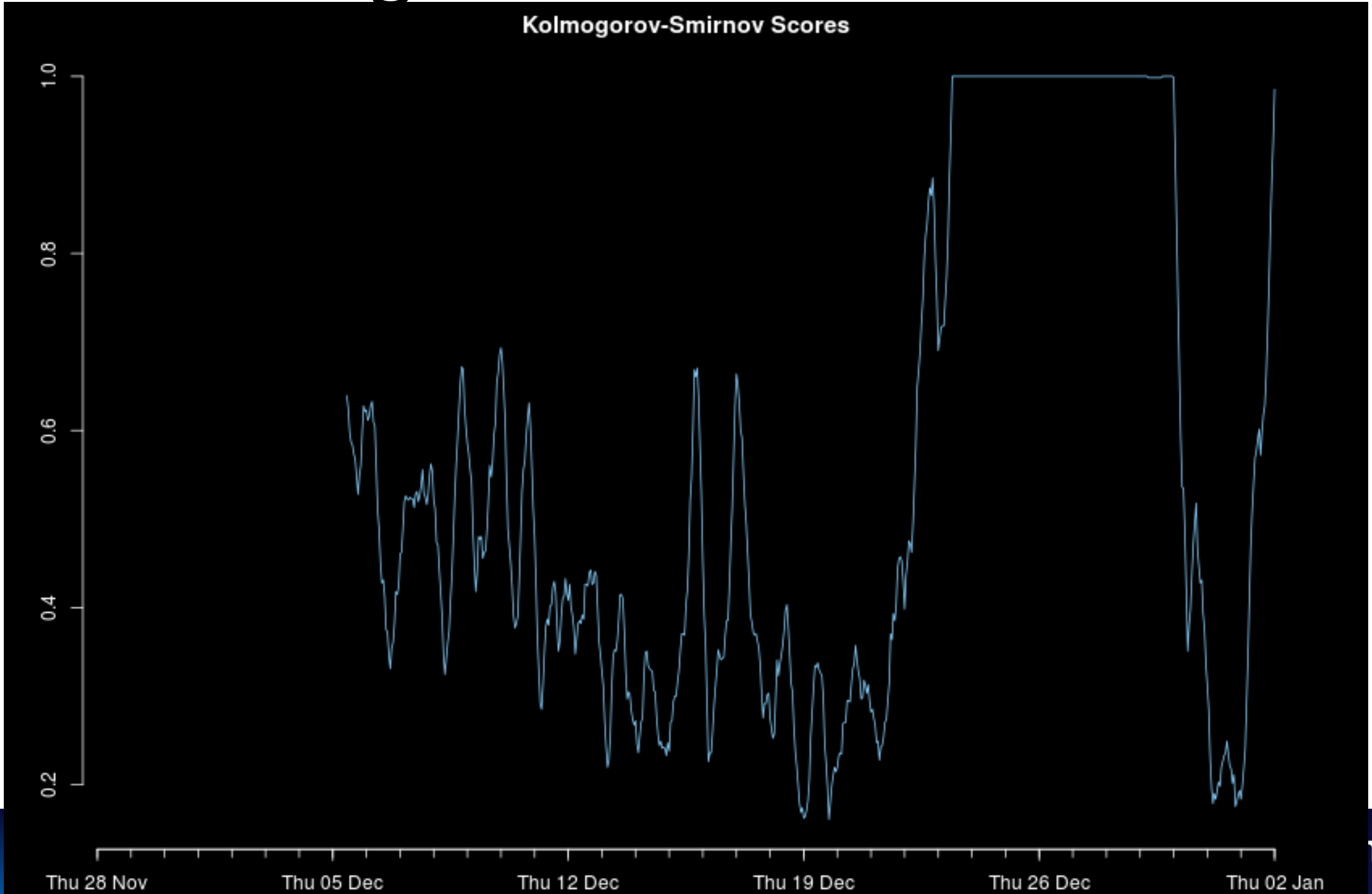
Data from dissimilar windows



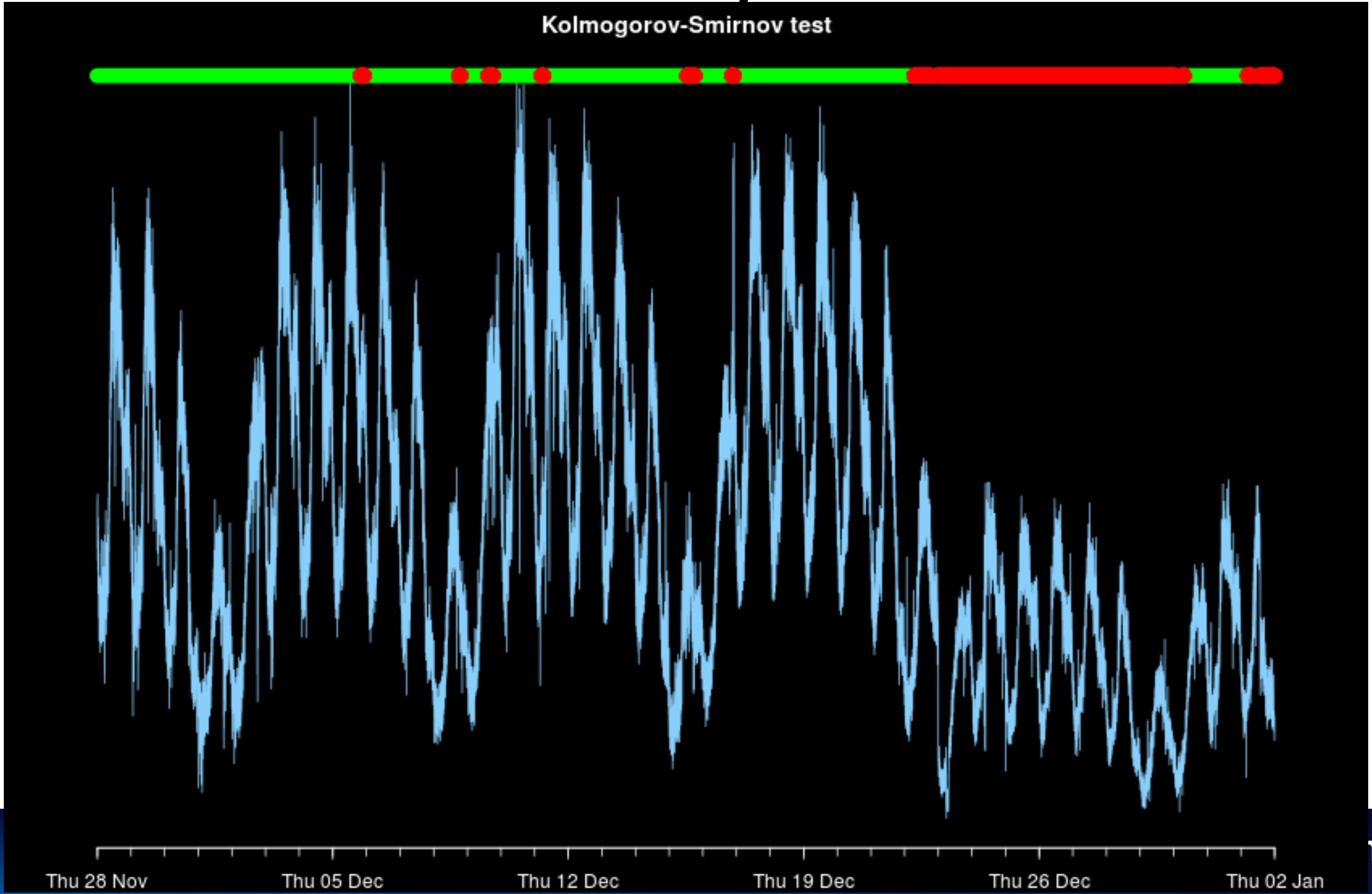
Cumulative distribution for those windows



Sliding window of KS scores

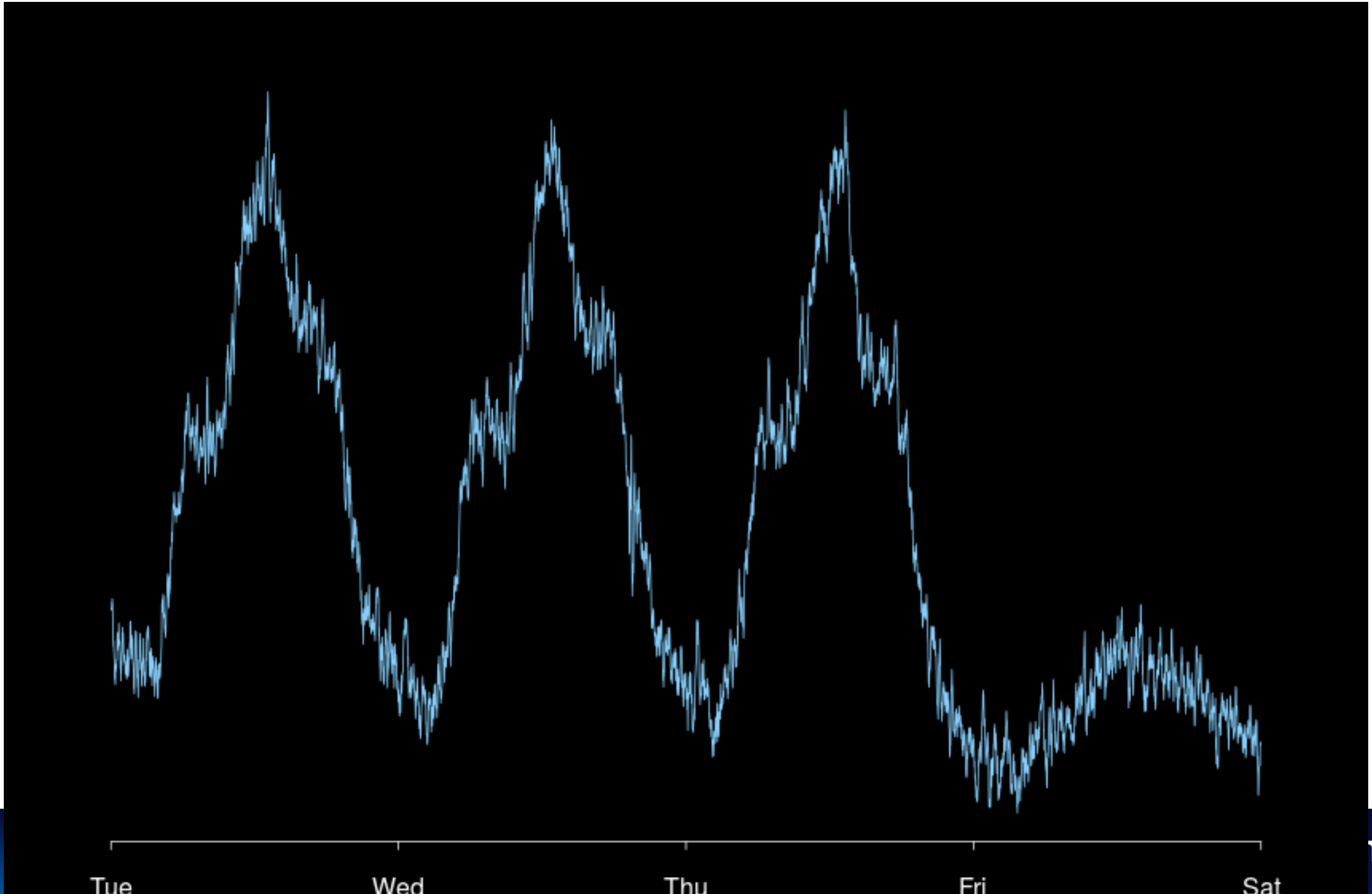


KS anomaly results

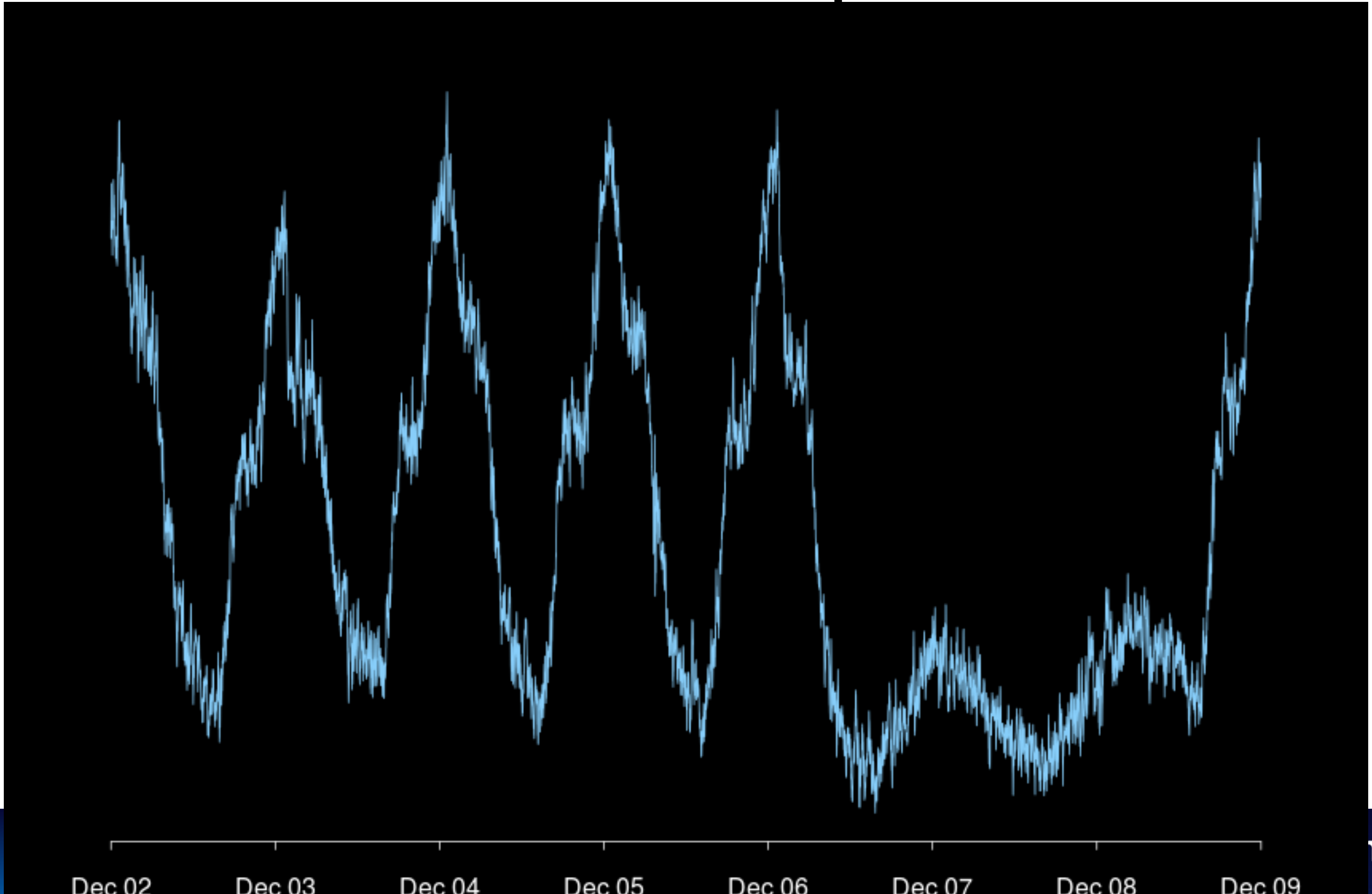


Thing 3:
Take Scope and Context into account!

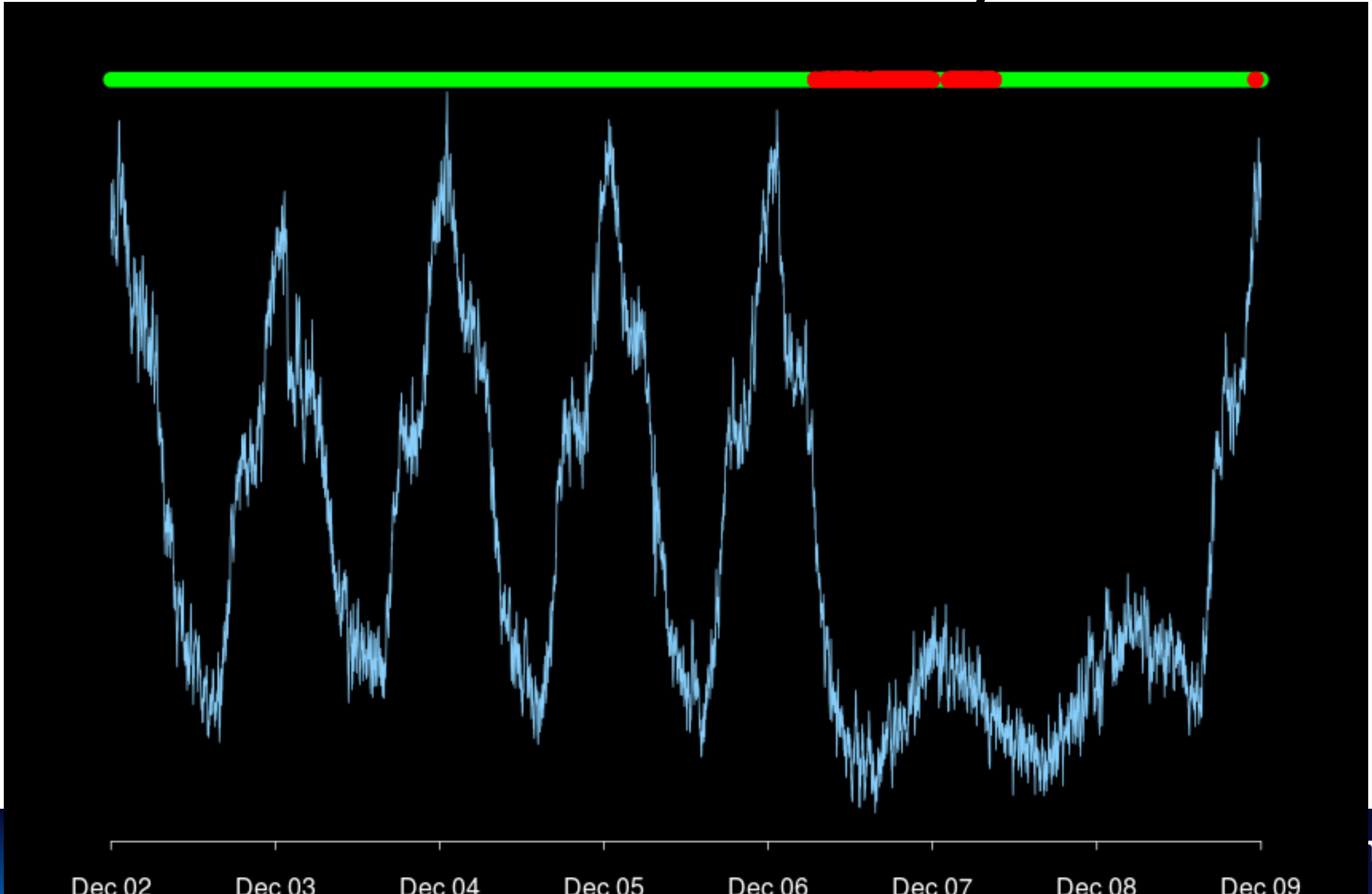
Some data – is that normal?



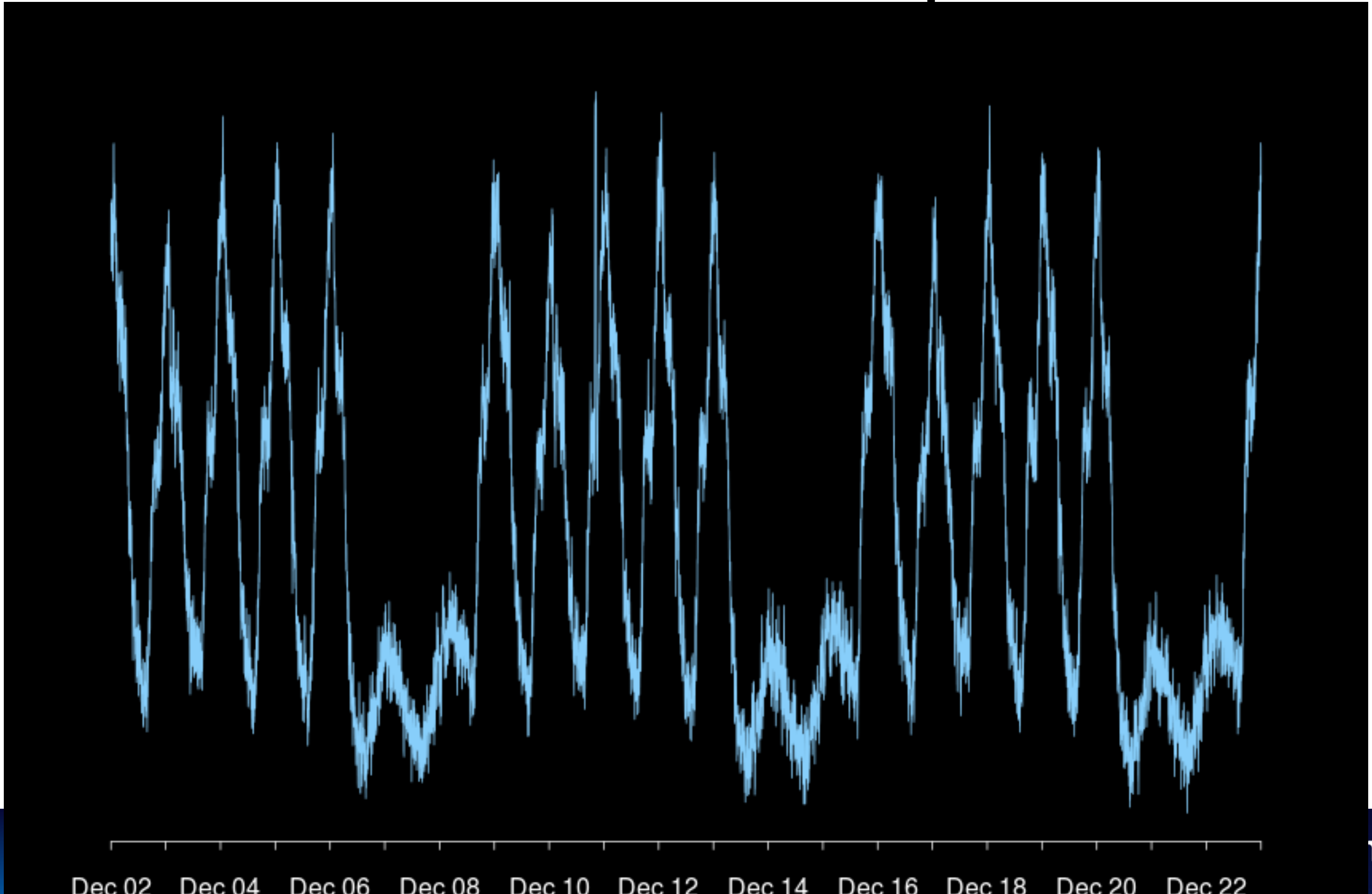
Wider scope



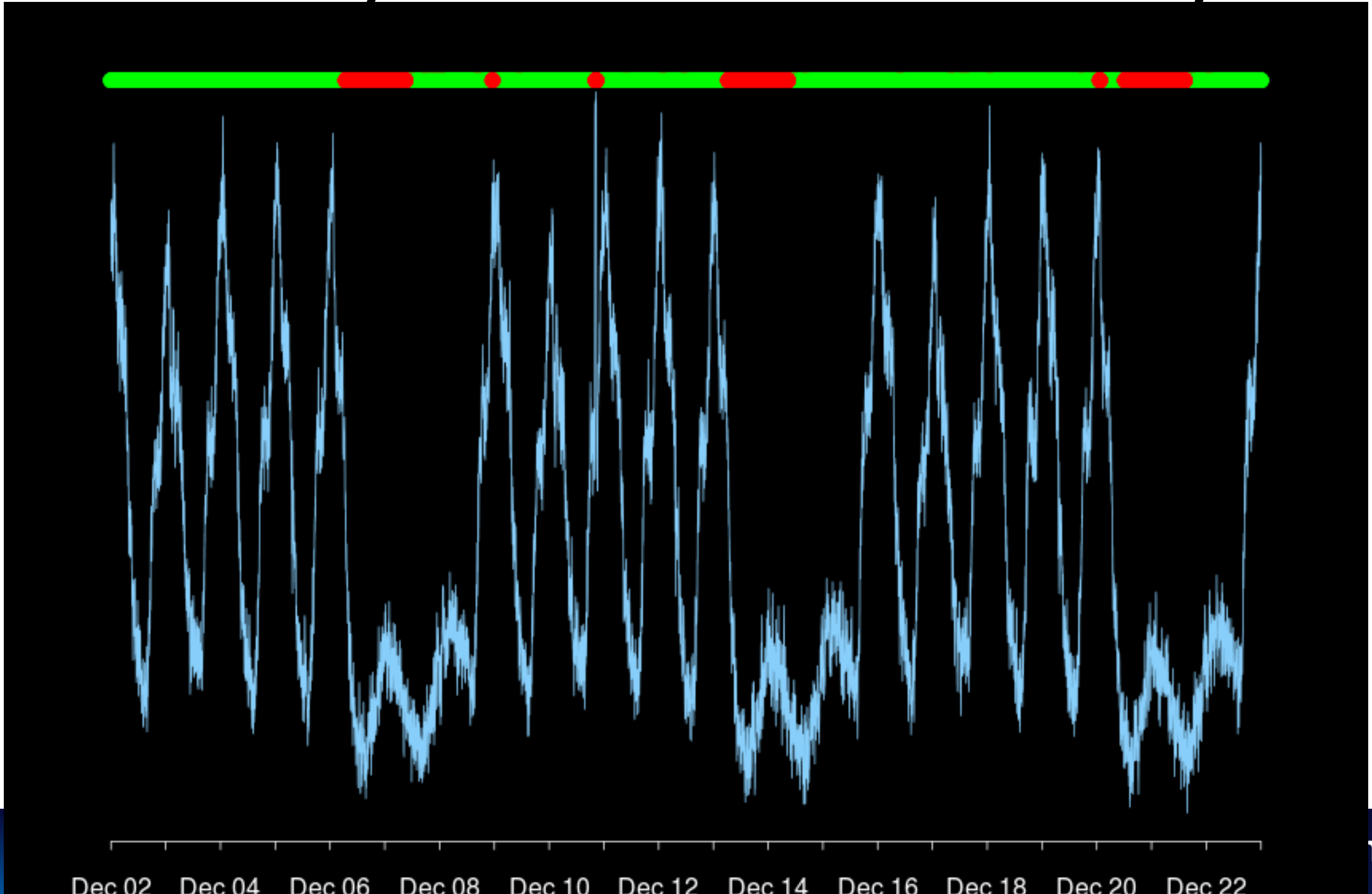
Is this an anomlay?



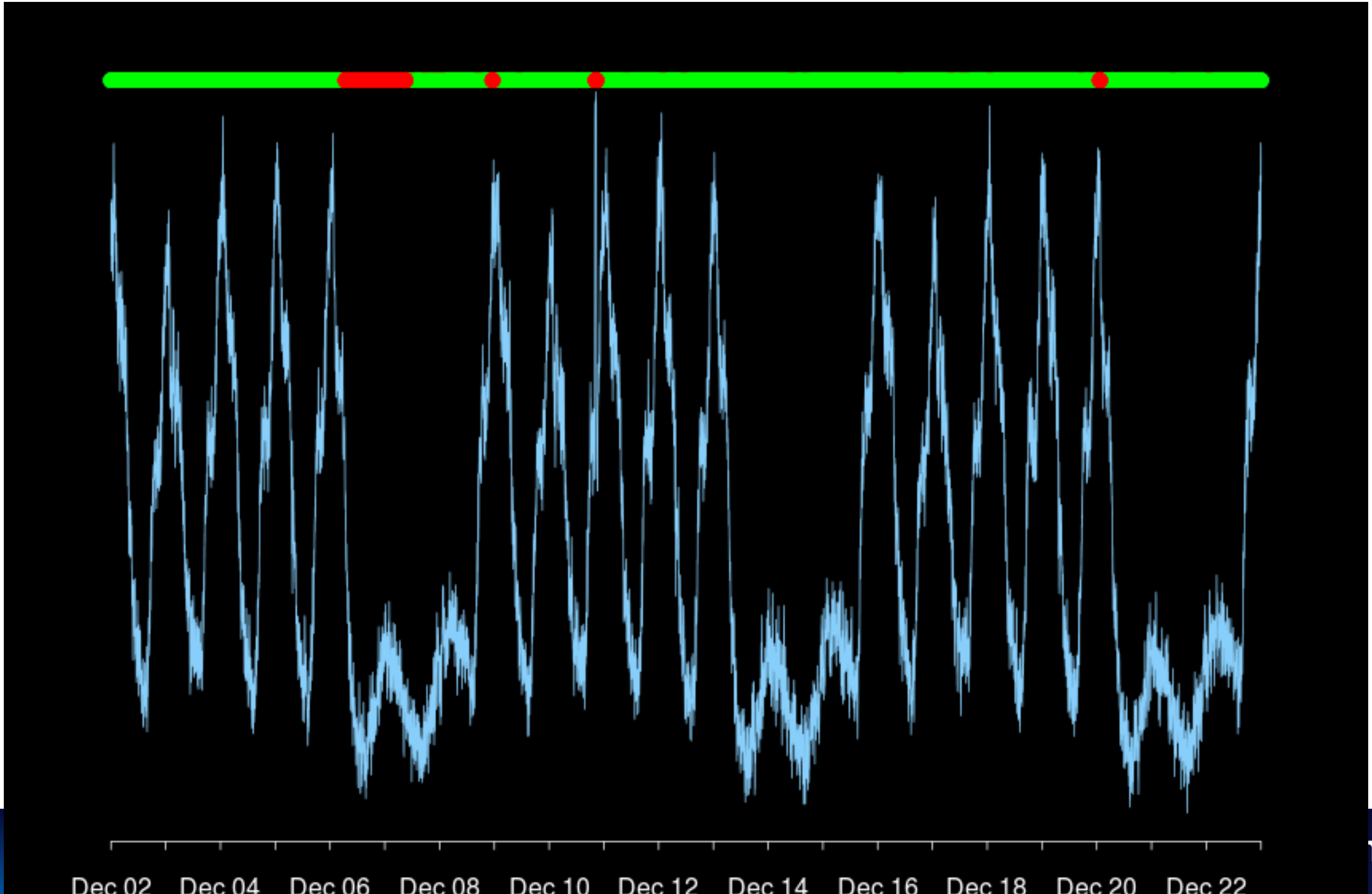
Even wider scope



Is every weekend an anomaly?



Would this be more accurate?



Use domain knowledge!

- Domain knowledge is NOT a bad thing!
  - There is no algorithm that will work on everything
  - Know your data and it general patterns
    - Periodicity/Seasonality
    - Known events (maintenance, backups, etc)
  - Apply the appropriate algorithms, taking into account enough scope for any inherent periodicity to appear
  - Customize your alerts to take into accounts known events

How does ML fit within ITOA?

What is IT Operations Analytics (ITOA)?

"IT operations analytics builds on Big Data processing capabilities to provide IT log management, log search and analysis, and related historical and predictive performance, capacity, and root cause analytics" – IDC*

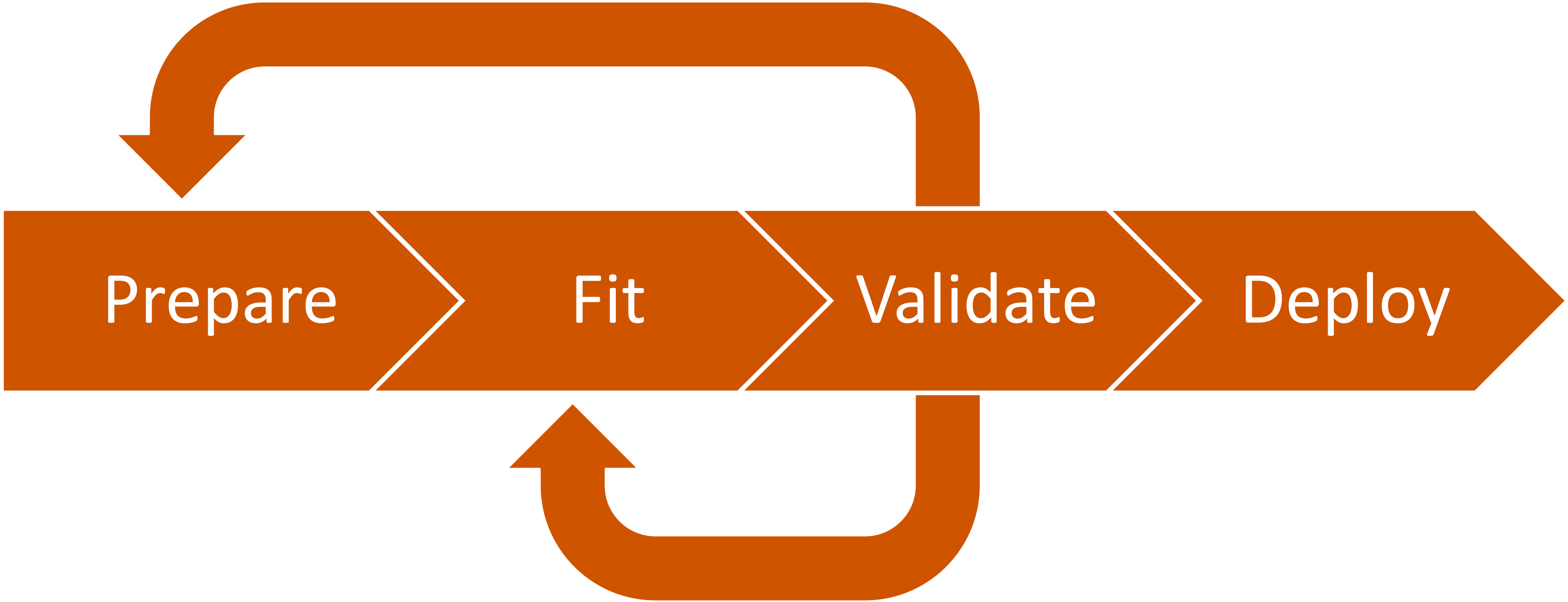
* IDC's Worldwide IT Operations Analytics Taxonomy Special Study, 2015

Principal benefits of ITOA*

- Avoidance of service interruptions, slowdowns, and outages
- Faster root cause analysis and problem recovery times
- Enhanced system and application performance
- Improved end-user experience
- Increased operational efficiency
- Improved compute resource utilization

* IDC's Worldwide IT Operations Analytics Taxonomy Special Study, 2015

Appling the ML Process to ITOA



Splunk ML Algorithms

Unsupervised

Continuous

Clustering:

- kmeans, cluster
- K-means
- DBSCAN
- Birch
- Spectral Clustering

Dimensionality reduction:

- PCA
- KernelPCA

Categorical

Association Analysis

- Apriori
- FP-Growth
- Hidden Markov Model

Vectorization:

- TFIDF

SPL command

ML Toolkit App v1.3

Supervised

Regression:

- Linear Regression
- Polynomial Regression
- ElasticNet
- Ridge
- Lasso
- RandomForestRegr.
- Decision Trees

predict
outliers
anomalies
anomalydetection

Classification:

- Logistic Regression
- Support Vector Machine
- Naïve-Bayes (Gaussian, Bernoulli)
- RandomForestClassifier
- KNN, Trees

… plus 300+ algos from Python

Machine Learning in IT Service Intelligence

Anomaly Detection

- Employ machine learning to baseline normal operations and alert on anomalous conditions
- Identify abnormal trends and patterns in KPI data
- Catch issues that thresholds cannot

Machine Learning in IT Service Intelligence

Adaptive Thresholds

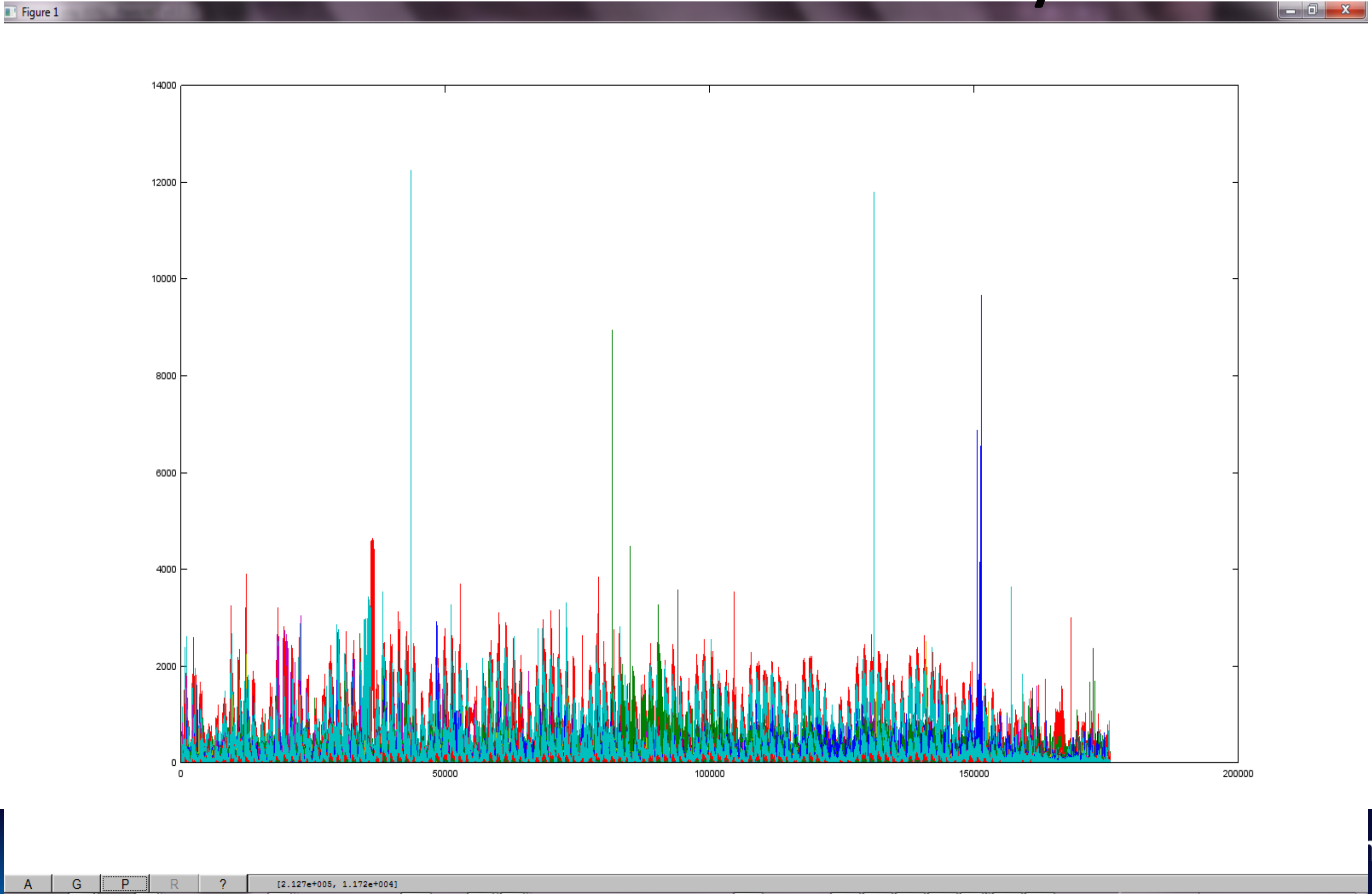
- Baseline normal activity and use stats to dynamically adapt KPI thresholds by time
- Easily create and set thresholds on KPIs
- Easily manage and maintain KPIs

Machine Learning in IT Service Intelligence

Event Correlation

- Reduce event clutter, false positives and extensive rules maintenance
- Events are auto-grouped together (supressed, de-duped)
- Easily provide feedback on auto-grouping of events & alerts

About that anomaly



Look closer

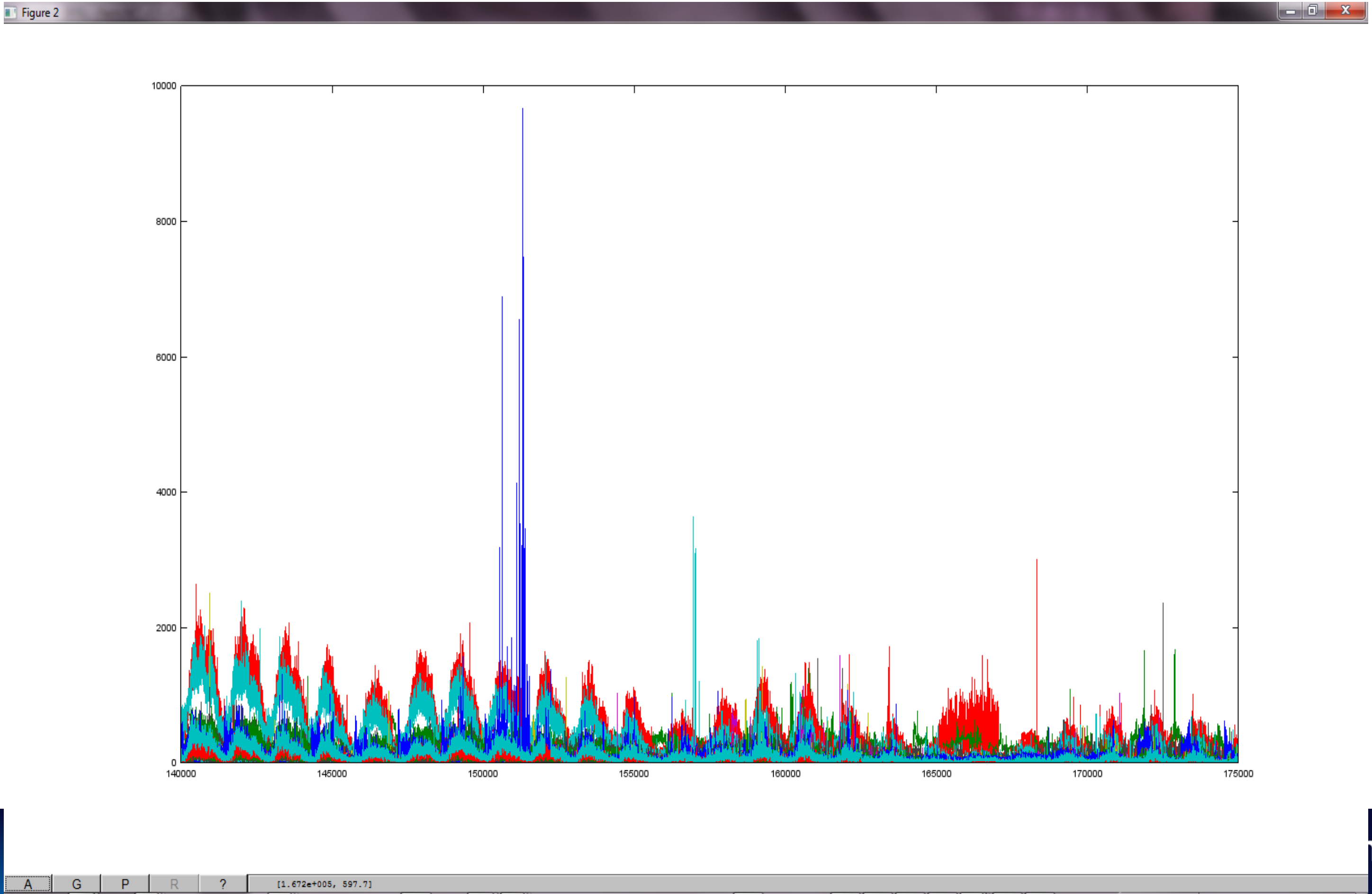
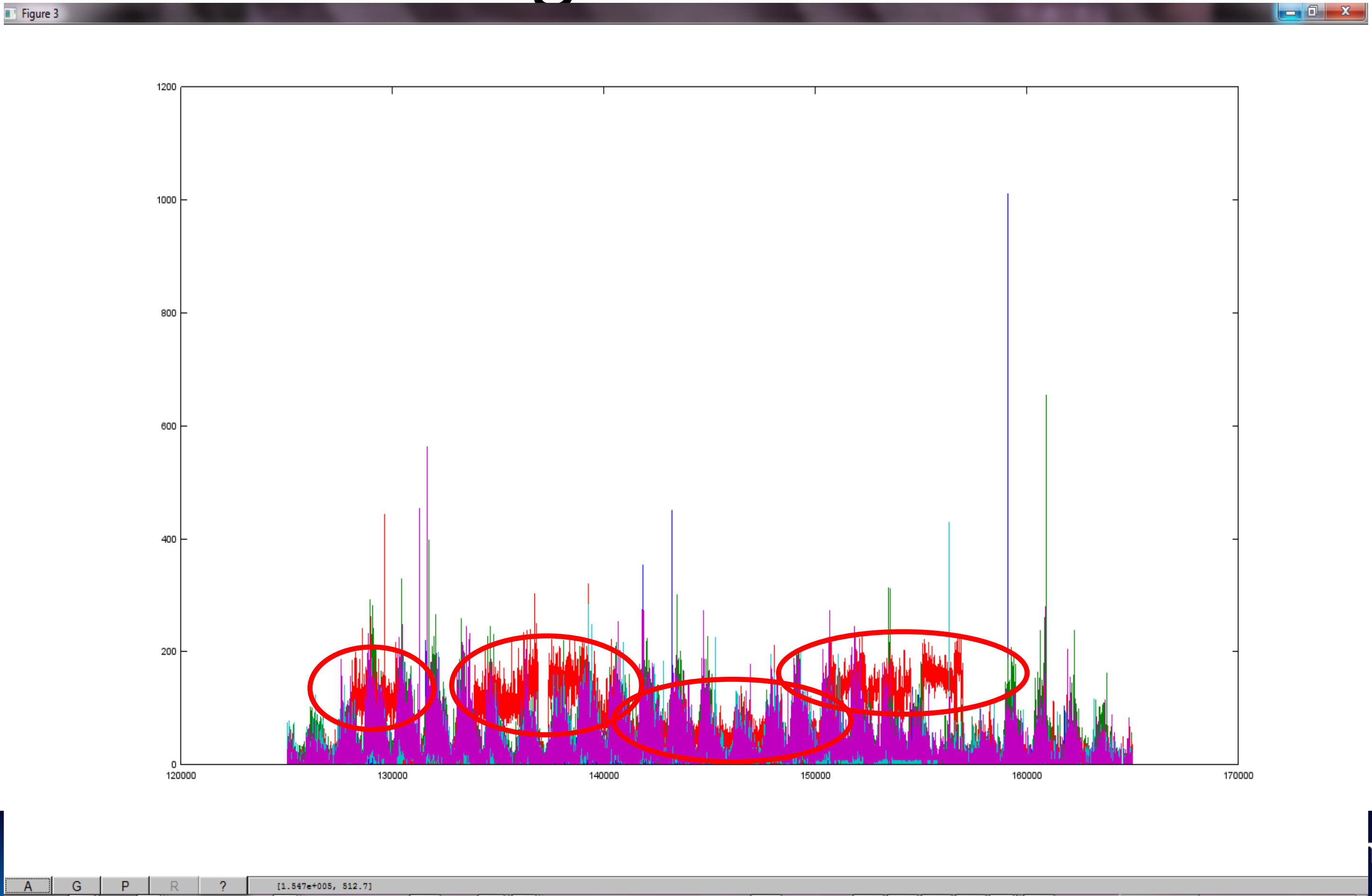


Figure 2
10000
8000
6000
4000
2000
0
140000
145000
150000
155000
160000
165000
170000
175000
A
G
P
R
?
[1.672e+005, 597.7]

Hiding in the noise



Key Takeaways

- Machine Learning is an evolution in the tools available to us
- ML is not one thing, it's many different types of things that can be applied to different types of problems
- ML applications and techniques vary so like any other tool, it helps to use the right tool for the right problem space
- When it comes to statistical learning
  – Your data is probably (heh) not Gaussian
  – You should try and say Komogorov-Smirnov
  – Take context into account when leveraging ML tools

If interested, go see this

- **Advanced Machine Learning in SPL with the Machine Learning Toolkit**
- **Thursday, September 29, 2016 | 12:25 PM-1:10 PM**
- ADVANCED | **Products:** Splunk Enterprise, Other | **Role:** Data Scientist/Analyst, Splunk Technical Champion | **Track:** Splunk Foundations | **Session Focus:** Search Language | **Other Topics:** Machine Learning
- **Speaker: Jacob Leverich**, Director of Engineering, Splunk Inc.

References and sources

- http://www.gartner.com/newsroom/id/3114217
- https://www.linkedin.com/pulse/gartner-2015-hype-cycle-big-data-out-machine-learning-sherif-fathy
- http://www.slideshare.net/tboubez/simple-math-for-anomaly-detection-toufic-boubez-metafor-software-monitorama-pdx-20140505
- http://sebastianraschka.com/Articles/2015_singlelayer_neurons.html
- http://cs231n.github.io/neural-networks-1/
- http://www.datarobot.com/blog/a-primer-on-deep-learning/
- http://projecteuclid.org/download/pdf_1/euclid.ss/1009213726
- http://www.webpages.ttu.edu/dleverin/neural_network/neural_networks.html
- http://www.ebtic.org/pages/ebtic-view/ebtic-view-details/machine-learning-on-big-data-d/687

THANK YOU