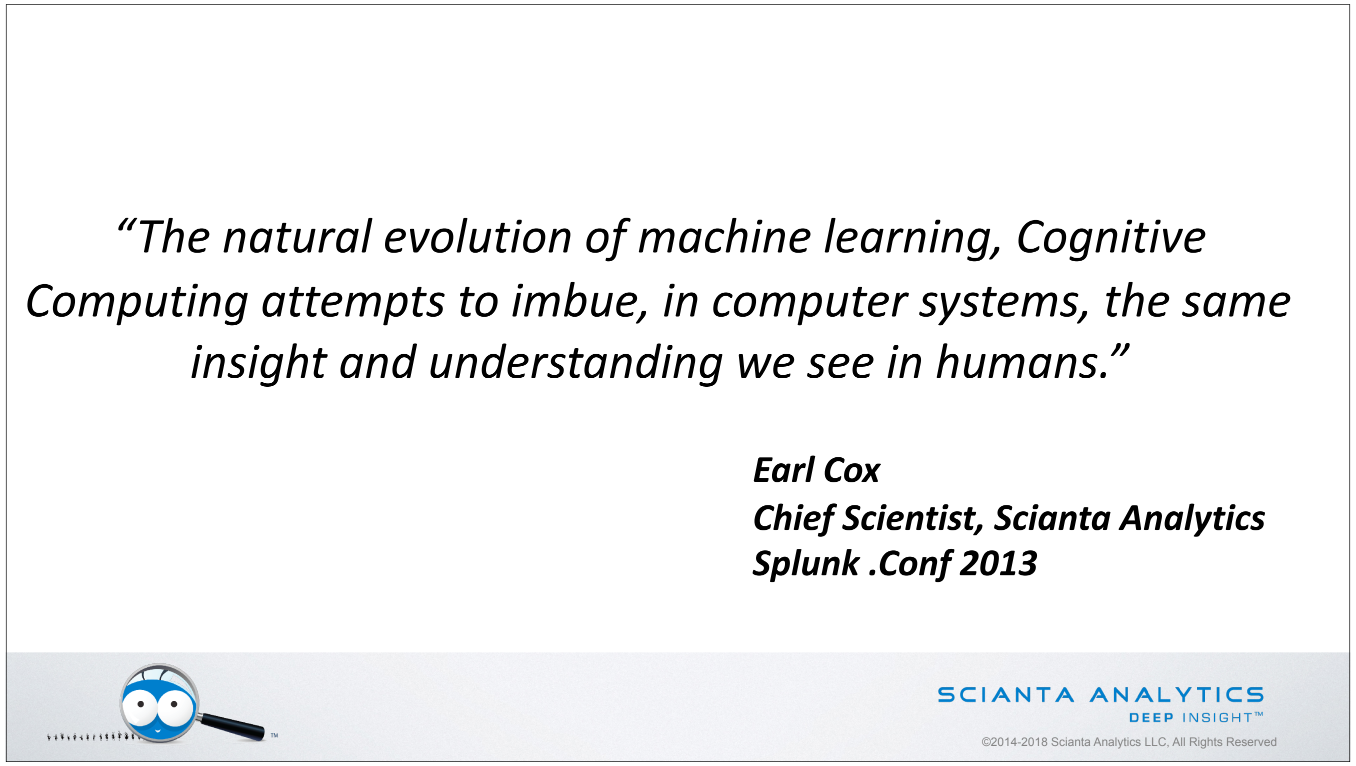
ANOMALY DETECTION

TM

SCIANTA ANALYTICS
DEEP INSIGHT™



“The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans.”

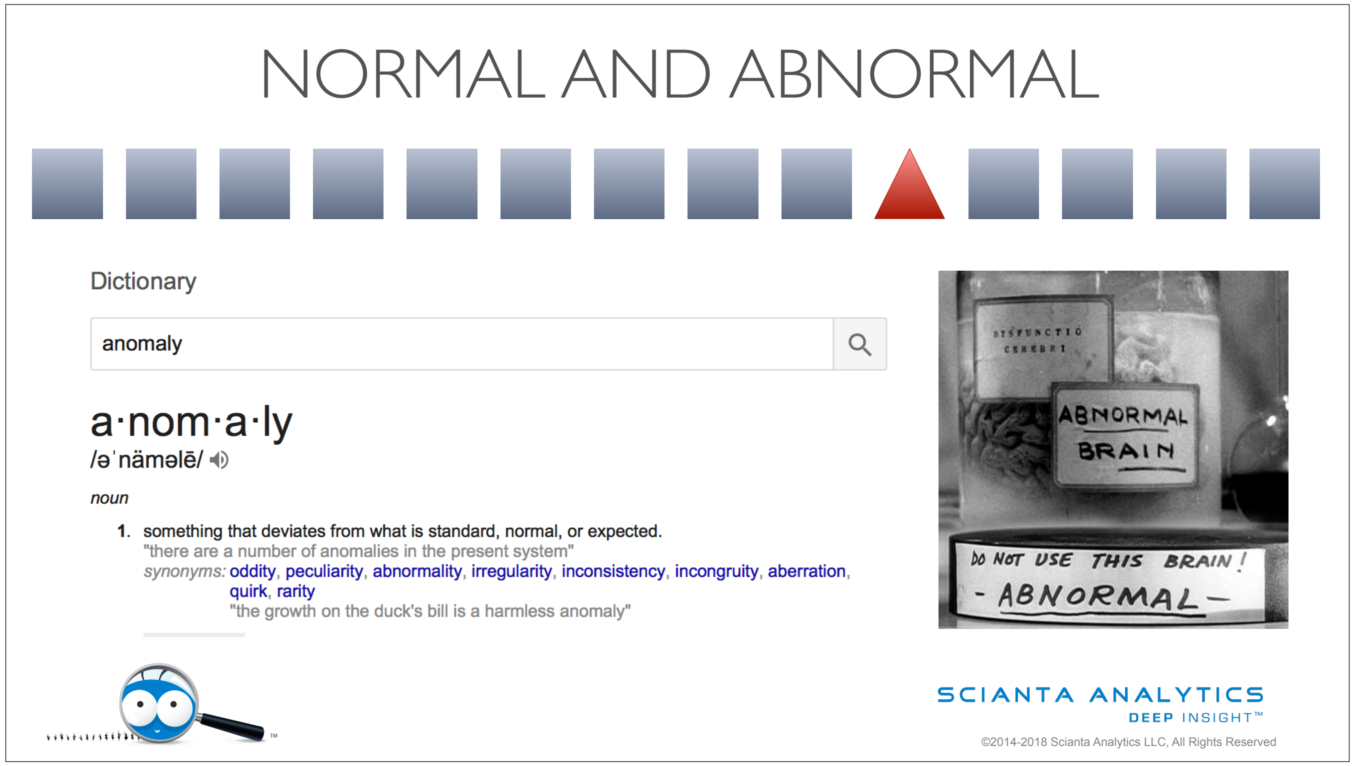
Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013

©2014-2018 Scianta Analytics LLC, All Rights Reserved

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness

©2014-2018 Scianta Analytics LLC, All Rights Reserved



NORMAL AND ABNORMAL

Dictionary

anomaly

a·nom·a·ly
/əˈnäməlē/

noun

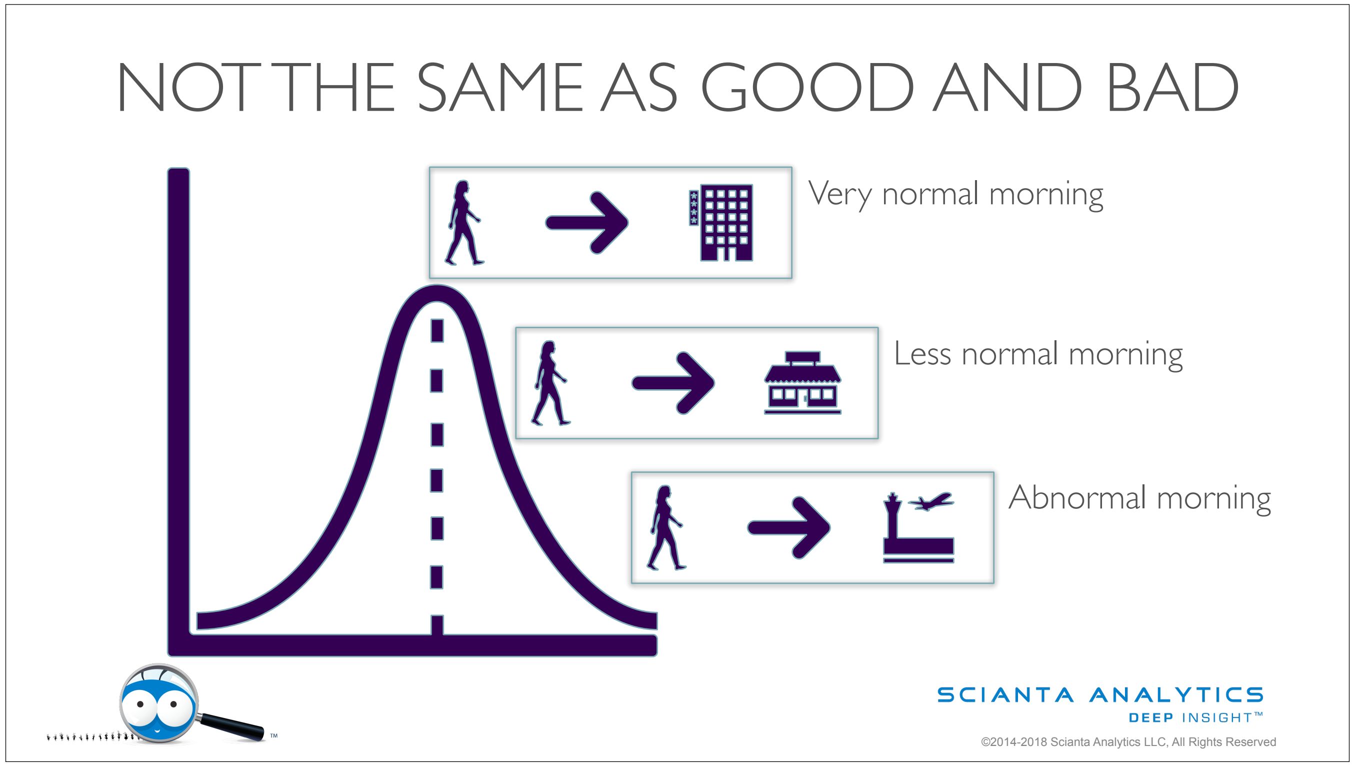
1. something that deviates from what is standard, normal, or expected.
   "there are a number of anomalies in the present system"
   synonyms: oddity, peculiarity, abnormality, irregularity, inconsistency, incongruity, aberration, quirk, rarity
   "the growth on the duck's bill is a harmless anomaly"

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

So what is an anomaly, anyway? According to the dictionary it's something that is not normal. Great, so we just have to define what normal is! That may or may not be easy, it depends on the data that's been collected. What does that data tell us about what is normal and abnormal? If there's a clear pattern of normal behavior in the data, then our cognitive computing system can use machine learning algorithms to determine which events are abnormal. Some data sets are very good for this technique, things that have a narrow band of normal behavior and produce lots of data. For instance, if you measure the speed of a grinder in a machine shop, it's probably around the same speed all the time, and if it slows down by 50% you've got an abnormality. Time is also valuable to consider. 70 degrees Fahrenheit is a normal temperature for Minneapolis in June, but abnormal in January.

Of course, if there isn't a normal, then we don't know what's abnormal, all we can do is report how different the current value is from what came before. What's the normal price for a Bitcoin?



NOT THE SAME AS GOOD AND BAD

Very normal morning

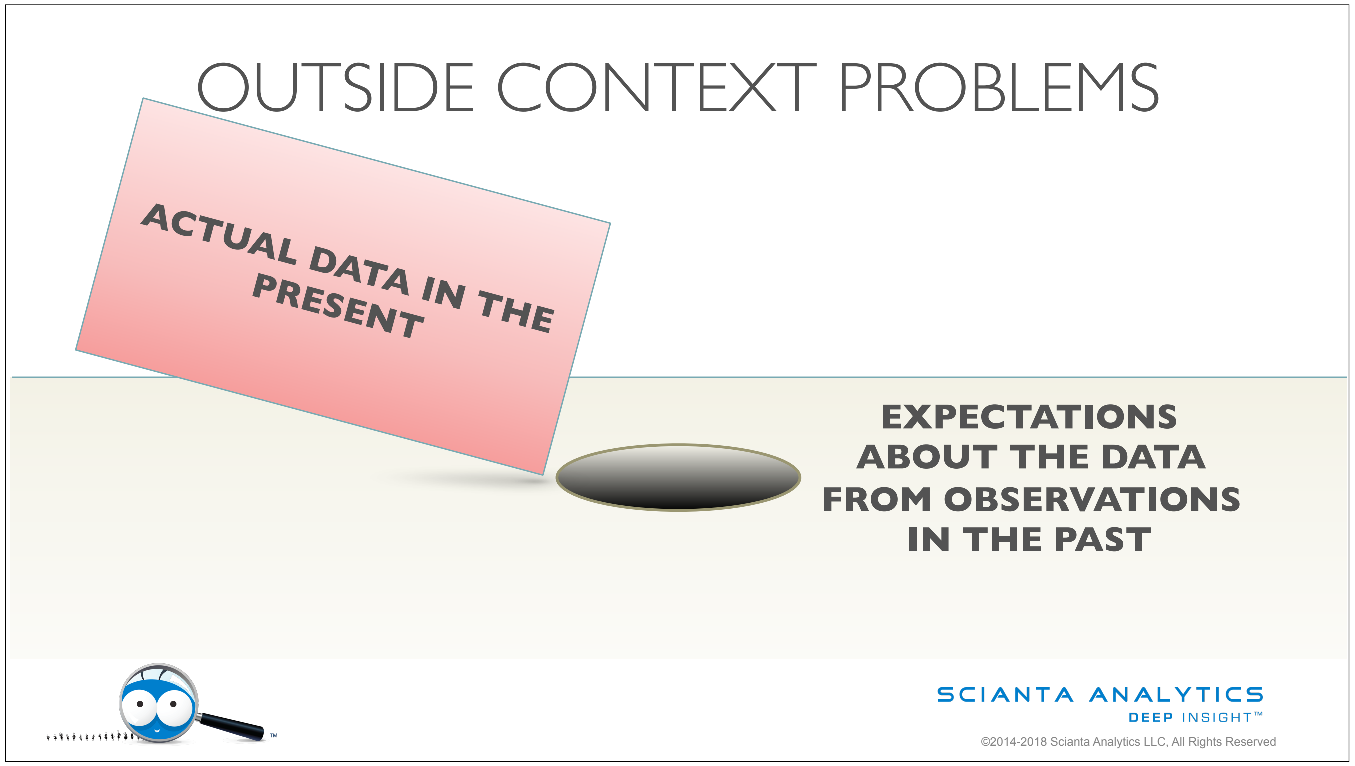
Less normal morning

Abnormal morning

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

The next key to understanding anomalies is that they aren't always good or bad. Continuing to think about the cost of a cryptocurrency like Bitcoin… a high value is good for sellers and bad for buyers, but our cognitive computing system isn't either one of those. All we can do is measure how normal the price is. An even clearer example is to take a person's morning commute as data. If they go to the office 80% of the time that they leave their home, that's a strong normal signal. If they go to the nearest cafe 10% of the time, that's a weak signal. But the morning that they go to the airport instead of either of those destinations is not a signal of something bad happening, it's just a signal that something different happened.



OUTSIDE CONTEXT PROBLEMS

ACTUAL DATA IN THE PRESENT

EXPECTATIONS ABOUT THE DATA FROM OBSERVATIONS IN THE PAST

SCIANTA ANALYTICS
DEEP INSIGHT™

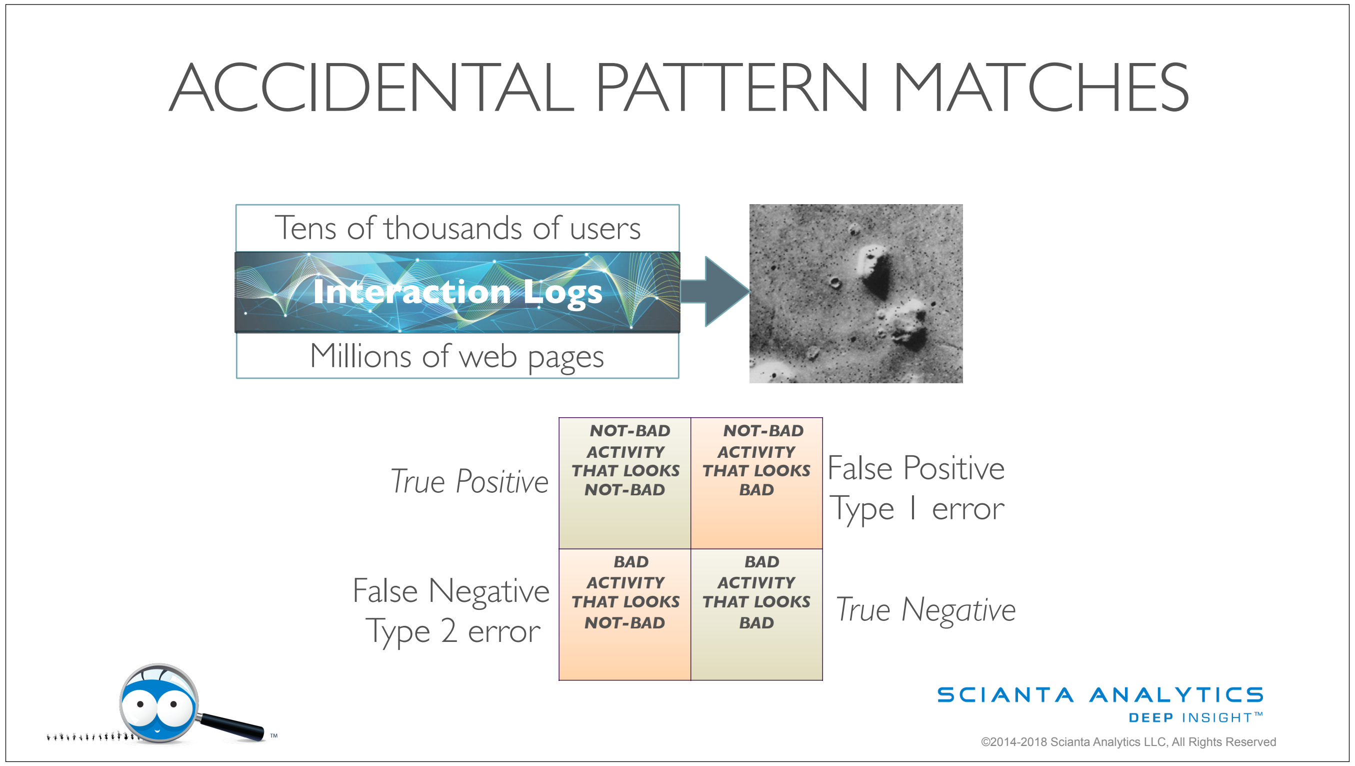
©2014-2018 Scianta Analytics LLC, All Rights Reserved

The OCP is a very simple problem to explain with very surprising effects. You're measuring within the context you understand, but the situation changes in an unpredictable way. Data is outside of the expected context, and the past is no longer an accurate predictor of the future. There are models that self-train from observed data, but please realize that they still assume past predicts future; they're just redefining the past as they go. A self-training model can still fail on unusual data.

So, you've got two techniques for dealing with outside context problems:

First and best… don't use cognitive computing concepts on datasets where the past is not predictive of the future. The price of a stock or crypto-currency is a good example of this; the pricing data set on its own does not give you sufficient data to make predictions of future performance. You can do some lower value predictions by severely limiting the time scope, but it's unlikely to predict the next market recession. If you want to learn more about limited scope prediction, the domain of High Frequency Trading (HFT) is worth researching.

Second, your cognitive computing system's alerts should go to a person, not an automated response. In theory, a complete context could enable a machine to produce perfect answers; but in practice, sufficient information is rarely available. Humans have to stay involved, because machines will never have complete context, so they'll produce weird alerts that need interpreting. Note that automated response systems can be very valuable for preparing enhanced information for the human to review.



ACCIDENTAL PATTERN MATCHES

Tens of thousands of users

Interaction Logs

Millions of web pages

True Positive	NOT-BAD ACTIVITY THAT LOOKS NOT-BAD	NOT-BAD ACTIVITY THAT LOOKS BAD	False Positive Type I error
False Negative Type 2 error	BAD ACTIVITY THAT LOOKS NOT-BAD	BAD ACTIVITY THAT LOOKS BAD	*True Negative*

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

Humans are the best pattern matching machines on the planet, but we're gradually teaching computers to do it like we do. That means they can make the same mistakes that we do! Far enough down that road we hit stereotyping, prejudice, and all sorts of bad outcomes, but we'll deal with that at a later date. For this introduction, let's focus on accidental match errors.

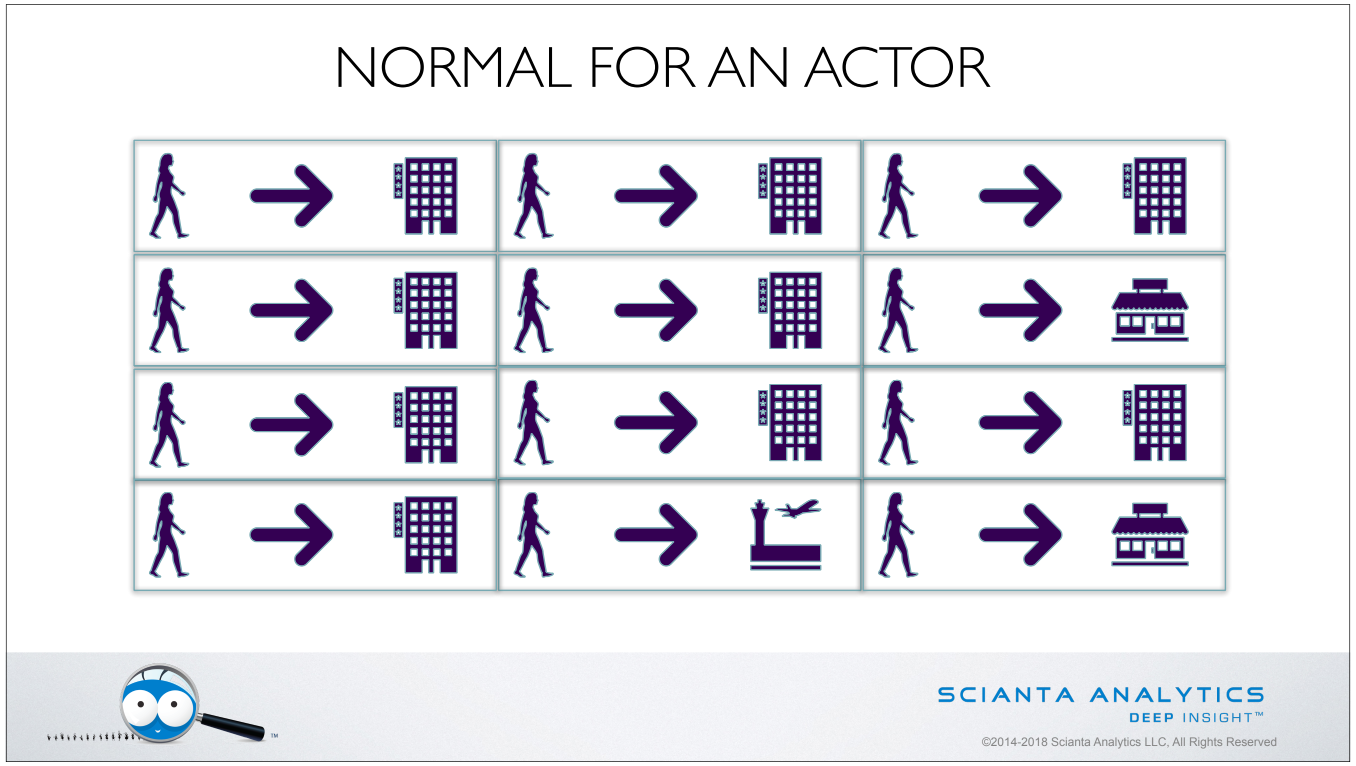
If you use a broad and varied data set and your definition of "Bad" is fuzzy enough, you can easily get two types of errors in your alerting. The first is a false positive, in which good behavior is considered bad. If you've ever had to fish a legitimate email out of your spam folder, you've seen this type of error. The second is a false negative, in which bad behavior is considered good. These are potentially scarier since there are people in the world trying to induce them for gain, so all alerting systems tend to bias towards more type 1 than type 2. Caution is required though, because the human operators of a system are very good at pattern matching. If they determine that the system is often incorrect, they'll apply that decision by ignoring its alerts.

AGENDA

Introduction to Machine Intelligence	Data Handling 1	Data Handling 2	Anomaly Detection	Transactional Behavior	Impact Analysis
Academic Concepts	Collection	Retention	Anomaly Definition	Defining Transactions	Organizational Visibility
Data Systems	Storage	Format	Measuring Normality	Transaction Relationships	Types of Impact
Maturity Curve	Security	Labeling		Probability Measurement	Responsiveness

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

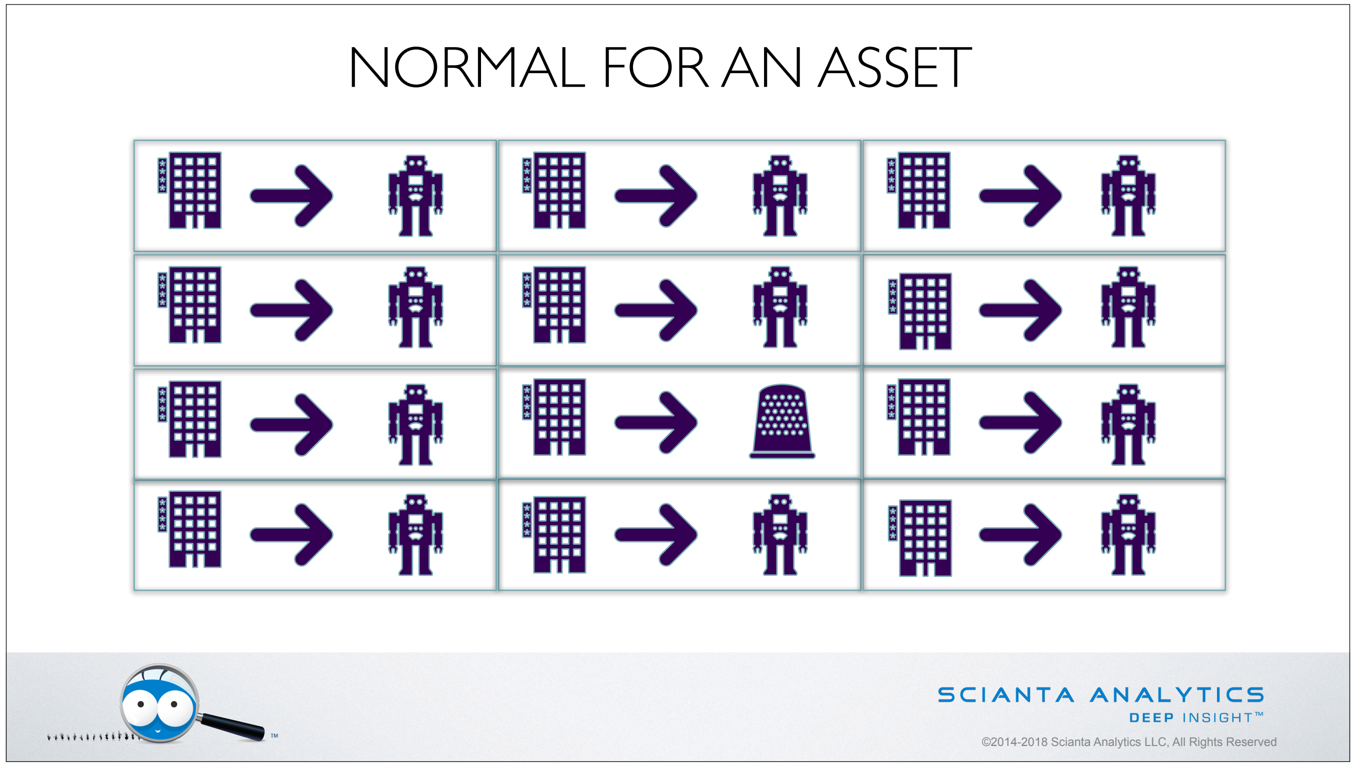


NORMAL FOR AN ACTOR

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

The typical way to work around false pattern matches is to determine our pattern by reviewing past behavior. So let's return to our commuter: we've got 12 commutes. 9 to the office, 2 to a cafe, and 1 to an airport. This actor's behavior is 75% normal, and 25% abnormal. That's a fairly weak signal of normalcy, but it's better than nothing. Still, we'd prefer to see 95% normal and 5% abnormal. 95th percentile is a standard measurement of normal versus abnormal, which comes from using statistical evaluation to find the third standard deviation on a bell curve. If you can get a signal that strong, few will question the result.



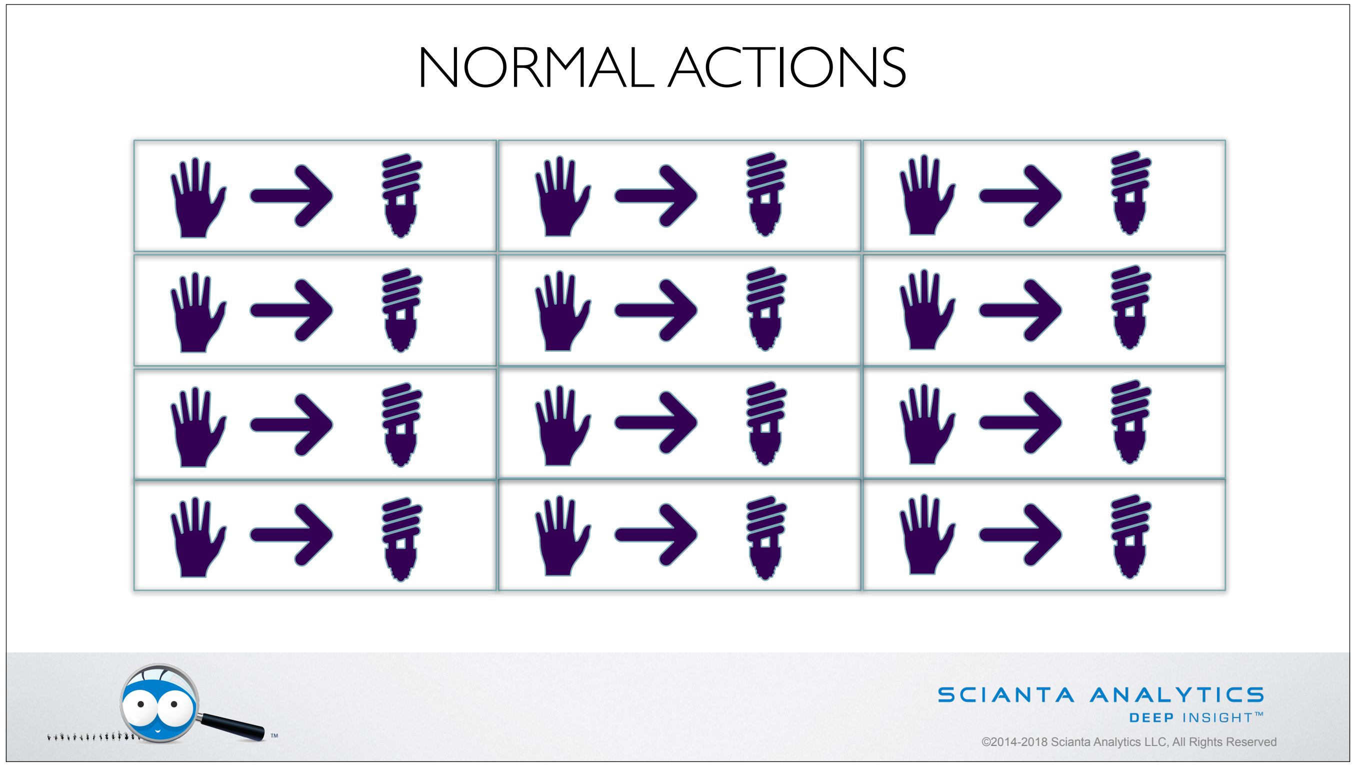
NORMAL FOR AN ASSET

SCIANTA ANALYTICS

DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

So let’s say that our commuter operates a machine at the office, and that machine makes robots. Machines tend to have really good normalcy; they do what they do, and when they stop doing that it’s dramatically abnormal. A pretty simple test is usually sufficient to accurately determine abnormal behavior. The classic example of this test is called a Shewhart control chart; you measure the median output, find a standard deviation, and stop the machine if its output is three standard deviations from prior mean. That’s 95% wrong, after all!

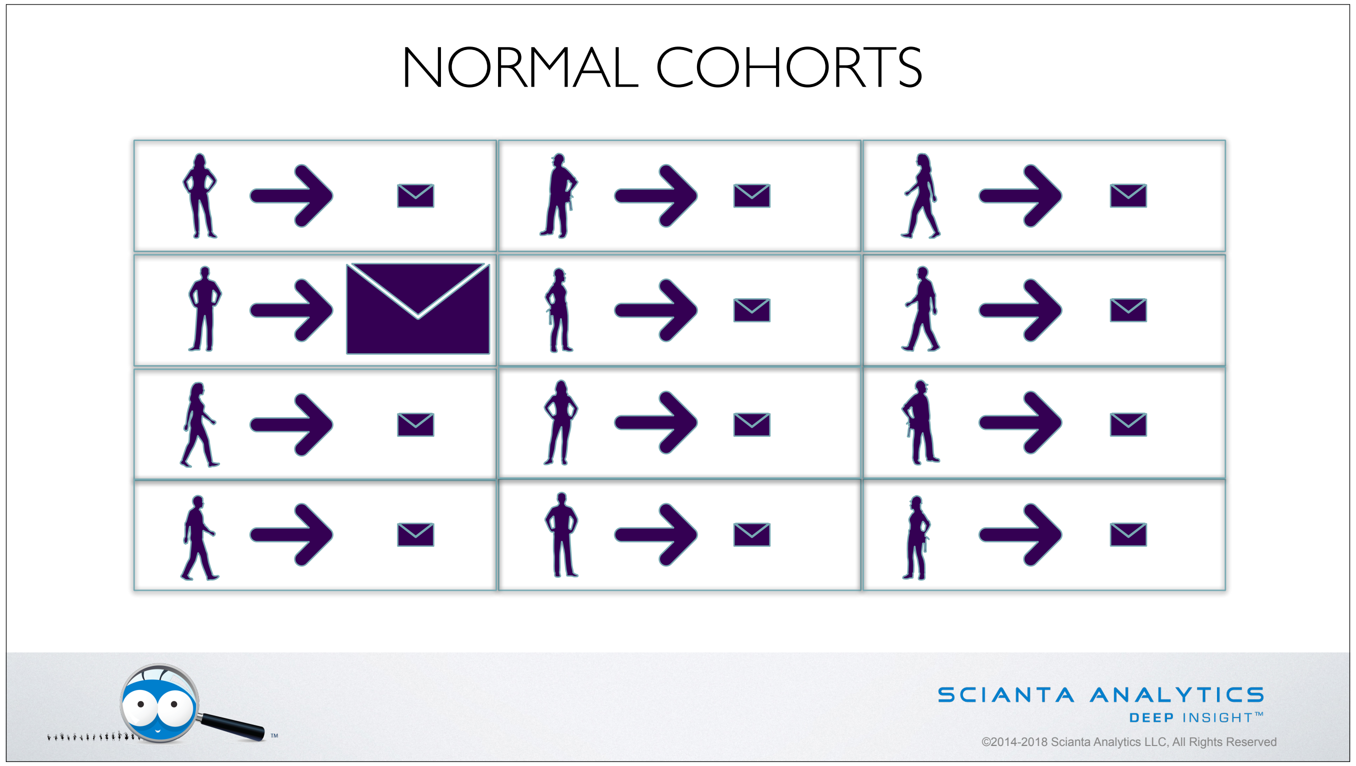


NORMAL ACTIONS

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

That approach can work for actions between actors and assets as well, if the set of possible actions is constrained. For instance, a light can be turned on or off, and that means the actions of switching it on and off can be treated as normal or abnormal. It's a binary universe of discourse. A harder action problem might be the size of emails that a user sends… how big is normal, and how bad is it that an abnormal email is sent? We can still find normal and abnormal, but our signal strength is lower. We are less sure that we've found a problem when we find an abnormal mail size.

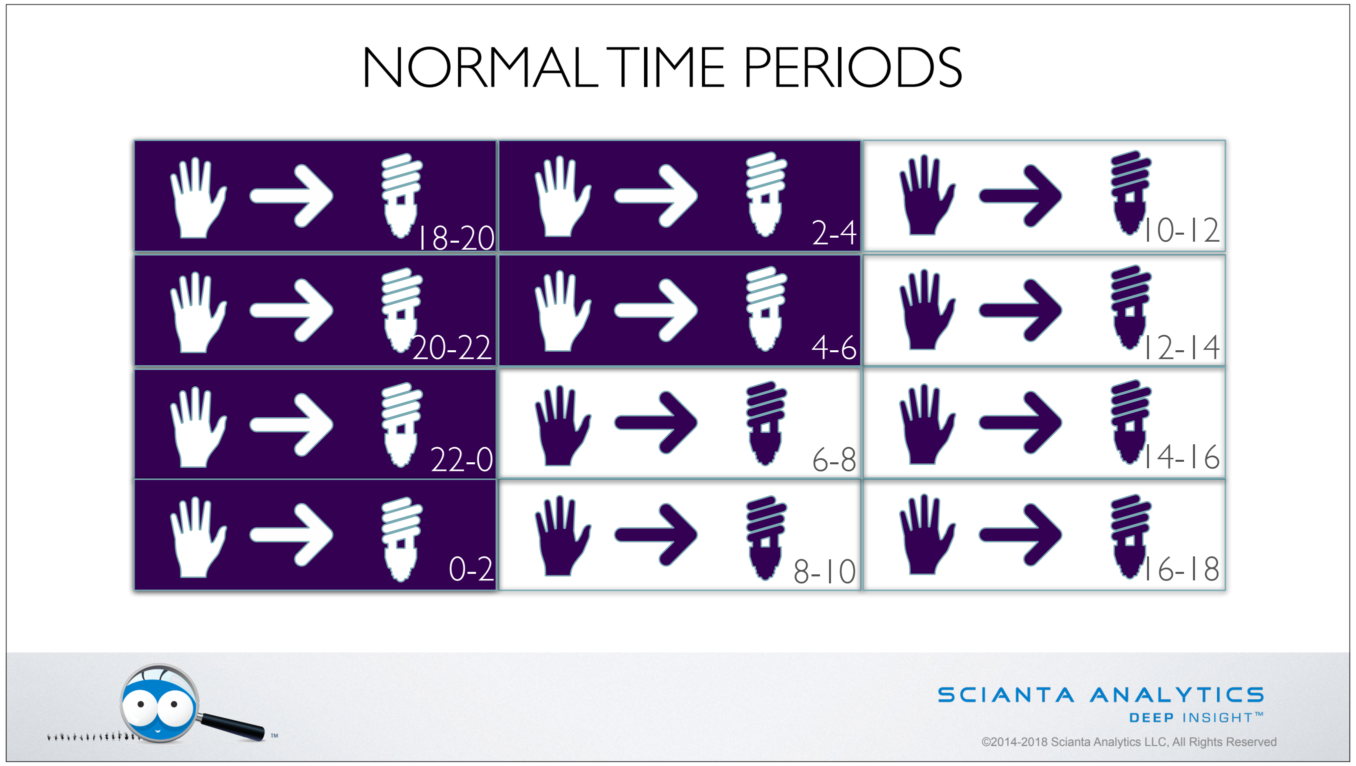


NORMAL COHORTS

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

We can increase our level of confidence in what is normal and abnormal by looking at groups of actors though. For instance, any one user can send big emails, and we're still not sure whether that is bad or not. But we can look at the emails sent by everyone on a team, and signal when a single actor's email is larger than what everyone else in that team sends. That can help us determine two things: is this actor consistent with their team, and is this actor consistent with themselves? If the answer to both is "no", we've found a fairly strong sign of abnormality. It's still not a sign of "badness", but it's strongly abnormal.



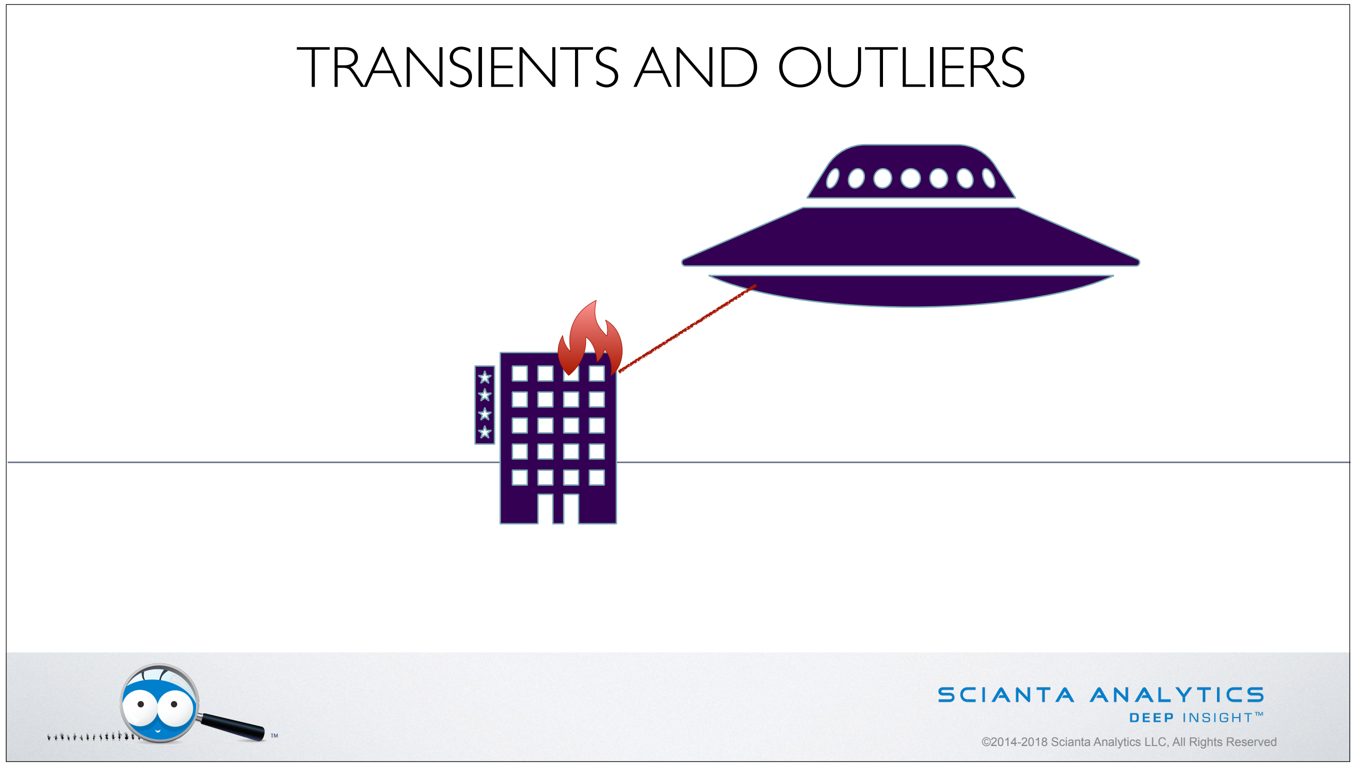
NORMAL TIME PERIODS

18-20	2-4	10-12
20-22	4-6	12-14
22-0	6-8	14-16
0-2	8-10	16-18

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

Another handy way to consider normalcy is to look at the time. It's very normal for a light to be on when it's dark outside, and it's less normal for the light to be on when it's light outside. This means that we can divide our data set into time blocks and measure normalcy by block. Given the data set shown here, we cannot say if the light being on is normal or not; but we can say that we expect it to be on from 2 to 4 in the morning.



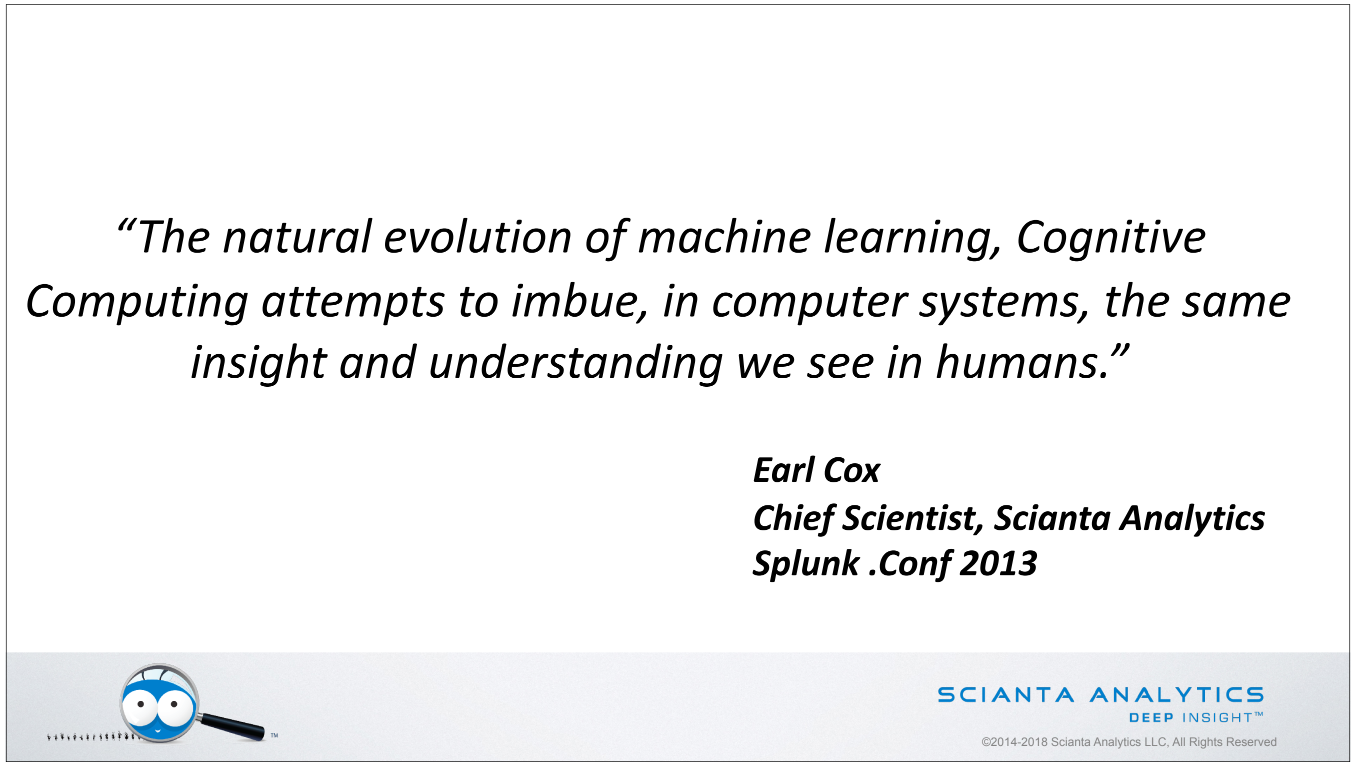
TRANSIENTS AND OUTLIERS

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

One last gotcha though; those outside context problems are still there. No matter how strongly we believe in our picture of what's normal in our data, the outside world can produce unexpected data. Perhaps our data collection system stops working, or perhaps there's a natural disaster that changes our priorities. As a data scientist or engineer, you've got a decision to make; do you remove this outlier data from your data set, or leave it in? If you leave it in, it will produce abnormality signals; do you suppress them, or let them go?

The benefit of allowing outliers to affect your signals is that you provide feedback to your users; it confirms for them that your system is also aware of the strange context, and helps them trust it. On the other hand, it's also producing alerts for them to consider, and that can be annoying. This means that the best decision depends on your context, particularly the amount of concern your users have about errors.



“The natural evolution of machine learning, Cognitive Computing attempts to imbue, in computer systems, the same insight and understanding we see in humans.”

Earl Cox
Chief Scientist, Scianta Analytics
Splunk .Conf 2013

SCIANTA ANALYTICS
DEEP INSIGHT™

©2014-2018 Scianta Analytics LLC, All Rights Reserved

Computers are force multipliers, not analysts; but they can help an analyst be more productive and successful. I hope this has been a useful overview of anomaly detection; next, we will discuss multiple step transactions. Thank you!

Thank you!