Outliers explained: a quick guide to the different types of outliers

Ira Cohen

Success in business hinges on making the right decisions at the right time. You can only make smart decisions, however, if you also have the insights you need at the right time. When the right time is right now, outlier detection (aka anomaly detection) can help you chart a better course for your company as storms approach — or as the currents of business shift in your favor. In either case, quickly detecting and analyzing outliers can enable you to adjust your course in time to generate more revenue or avoid losses. And when it comes to analysis, the first step is knowing what types of outliers you're up against.

The three different types of outliers

In statistics and data science, there are three generally accepted categories which all outliers fall into:

Type 1: Global outliers (also called "point anomalies"):
A data point is considered a global outlier if its value is far outside the

entirety of the data set in which it is found (similar to how “global variables” in a computer program can be accessed by any function in the program).

Type 2: Contextual (conditional) outliers:
A data point is considered a contextual outlier if its value significantly deviates from the rest the data points in the same context. Note that this means that same value may not be considered an outlier if it occurred in a different context. If we limit our discussion to time series data, the “context” is almost always temporal, because time series data are records of a specific quantity over time. It’s no surprise then that contextual outliers are common in time series data.

Type 3: Collective outliers:
A subset of data points within a data set is considered anomalous if those values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense. In time series data, one way this can manifest is as a normal peaks and valleys occurring outside of a time frame when that seasonal sequence is normal or as a combination of time series that is in an outlier state as a group.

Think of it this way:

A fist-size meteorite impacting a house in your neighborhood is a **global** outlier because it’s a truly rare event that meteorites hit buildings. Your neighborhood getting buried in two feet of snow would be a **contextual** outlier if the snowfall happened in the middle of summer and you normally don’t get any snow outside of winter. Every one of your neighbors moving out of the neighborhood on the same day is a **collective** outlier because although it’s definitely not rare that people move from one residence to the next, it is very unusual that an entire neighborhood relocates at the same time.

This analogy is good for understanding basic differences between the three types of outliers, but how does this fit in with time series data of business metrics?

Let's move on to examples which are more specific to business:

A banking customer who normally deposits no more than $1000 a month in checks at a local ATM suddenly makes two cash deposits of $5000 each in the span of two weeks is **a global anomaly** because this event has never before occurred in this customer's history. The time series data of his weekly deposits would show an abrupt recent spike. Such a drastic change would raise alarms as these large deposits could be due to illicit commerce or money laundering.

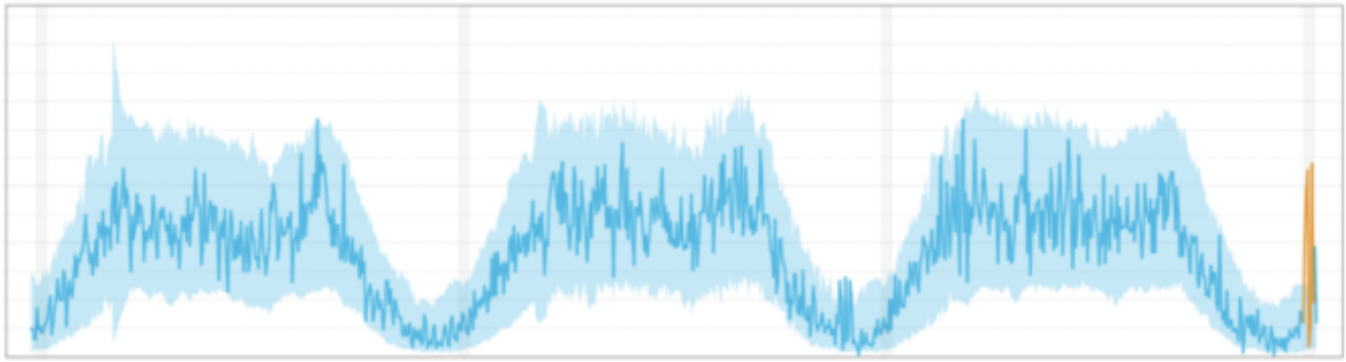
A sudden surge in order volume at an ecommerce company, as seen in that company's hourly total orders for example, could be a **contextual outlier** if this high volume occurs outside of a known promotional discount or high volume period like Black Friday. Could this stampede be due to a pricing glitch which is allowing customers to pay pennies on the dollar for a product?

A publicly traded company's stock is never a static thing, even when prices are relatively stable and there isn't an overall trend, and there are minute fluctuations over time. If the stock price remained at exactly the same price (to the penny) for an extended period of time, then that would be a **collective outlier**. In fact, this very thing occurred to not one, but several tech companies on July 3, 2017 on the Nasdaq exchange when the listed stock prices for several companies — including tech giants Apple and Microsoft — were listed as $123.45.

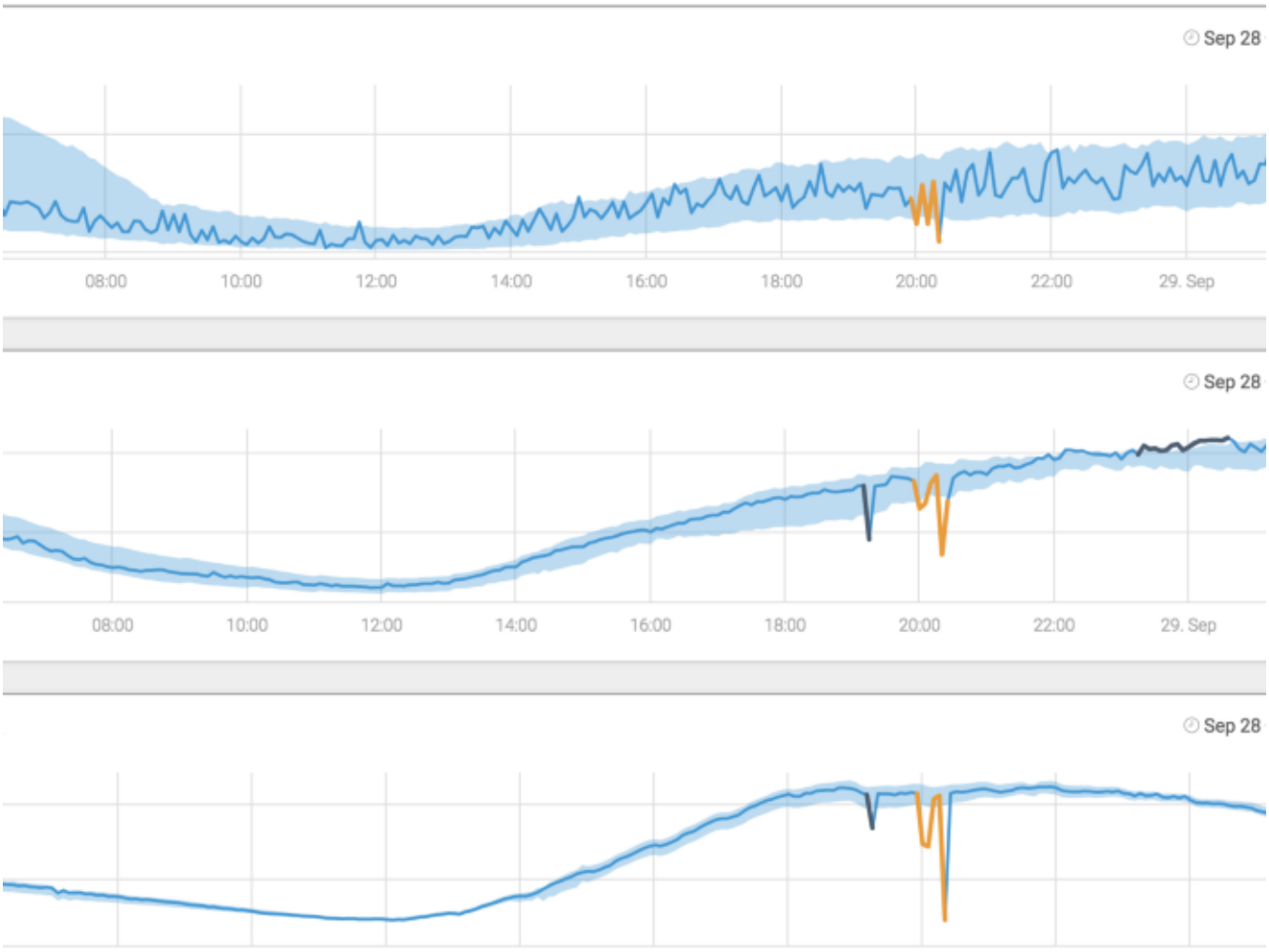
How do these anomalies look like? Below are a few examples.

Global anomaly: A spike in number of bounces of a homepage is visible as the anomalous values are clearly outside the normal global range.

Contextual anomaly: App crashes happen all the time and have a seasonal pattern (more users = more crashes). However, the number of app crashes in this anomaly are not outside the normal global range, but are abnormal compared to the seasonal pattern.



Collective anomaly: In the example, the anomalous drop in the number of successful purchases for three different product categories were discovered to be related to each other and are combined into a single anomaly. For each time series the individual behavior does not deviate significantly from the normal range, but the combined anomaly indicated a bigger issue with payments.



The three key steps to detect all types of outliers

Regardless of industry, no matter the data source, an outlier detection system should find all types of outliers in time series data, in real time, and at the scale of millions of metrics.

Outlier and Anomaly detection algorithms have been researched in academia and lately have started becoming available as commercial services as well as open source software. All rely on statistical and machine learning algorithms, based on methods such as ARIMA, Holt-Winters, Dynamic state-space models (HMM), PCA analysis, LSTMs and RNNs, and more. Beyond the base algorithms, there are many additional considerations in building such a system.

A comprehensive guide to how to build such a system is outlined in the 3-part whitepaper on anomaly detection. The key steps, applicable to all base outlier detection algorithms, that help detect the various types of outliers are:

1. Choosing the most appropriate model and distribution for each time series: This is a critical step to detect any outlier because time series can behave in various ways (stationary, non-stationary, irregularly sampled, discrete, etc), each requiring a different model of normal behavior with a different underlying distribution.
2. Accounting for seasonal and trend patterns: contextual and collective outliers cannot be detected if seasonality and trend are not accounted for in the models describing normal behavior. Detecting both automatically is crucial for an automated anomaly detection system as the two cannot be manually defined for all data.
3. Detecting collective anomalies involves understanding the relationships between different time series, and accounting for those for detecting and investigating anomalies.

Outliers are often visible symptoms of underlying problems that you need to fix fast. However, those symptoms are only as visible as your outlier detection system makes them to be.

For additional visuals, see the our short explanation video.

Originally published in Nov 2019 here.