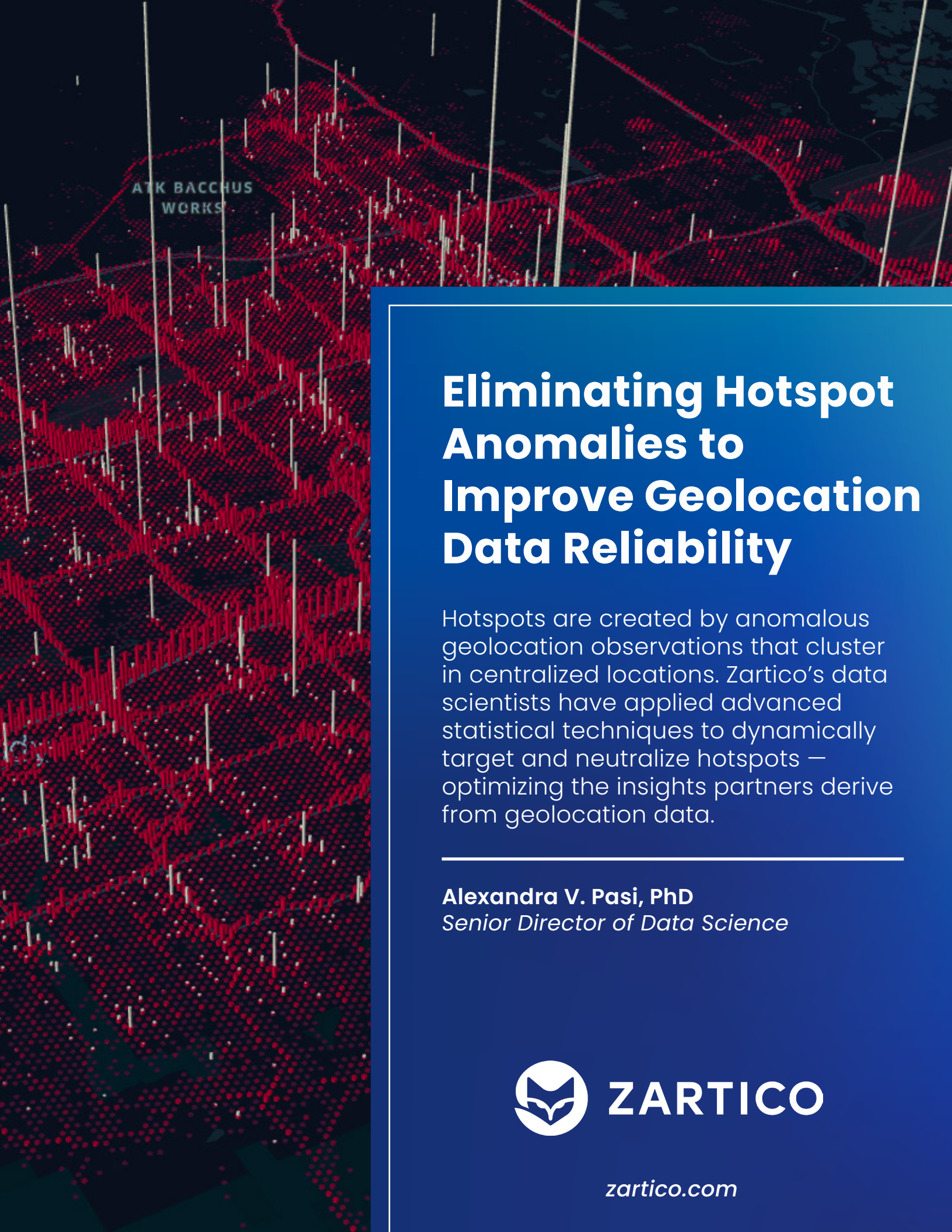




ATK BACCHUS
WORKS

Eliminating Hotspot Anomalies to Improve Geolocation Data Reliability

Hotspots are created by anomalous geolocation observations that cluster in centralized locations. Zartico's data scientists have applied advanced statistical techniques to dynamically target and neutralize hotspots — optimizing the insights partners derive from geolocation data.

Alexandra V. Pasi, PhD
Senior Director of Data Science



ZARTICO

zartico.com



Alexandra V. Pasi, PhD
Senior Director of Data Science

Alexandra V. Pasi holds a PhD in Mathematics, and has broad industry and academic experience in machine learning, statistics, data science, and quantitative and qualitative modeling and forecasting with applications to a variety of fields including finance, genomics, and geopolitics. She is passionate about communicating the fundamentals of good mathematical and scientific practice to decision-makers looking to make data-informed decisions.

Executive Summary:

- 1) Hotspots are a type of geolocation data anomaly created by approximate or anomalous location observations clustered in centralized locations. These clusters artificially inflate visitation counts to certain places or categories of places, generating point-of-interest counts or ranks that don't reflect actual visitation.
- 2) Standard solutions for addressing hotspots risk removing too much of the precise movement data or fail to address the fluctuating nature of hotspots.
- 3) Zartico's dedicated data science team applied advanced statistical processes to create a dynamic hotspot filter that selectively excises the approximate observations while maintaining the precise geolocation signal.

What is geolocation data?

Geolocation data — sometimes called mobile location data — is a set of latitude and longitude coordinates that correspond to the location of a mobile device. Smartphone users opt into a wide variety of mobile apps that collect and share this data, which is paired with an anonymized device ID and a timestamp.

A device's location is only captured when the user has activated one of these apps, so it doesn't follow individuals around 24/7. Some of the geolocation observations are highly precise, while others may be rounded to a central location within the city or neighborhood.

Eliminating Hotspot Anomalies to Improve Geolocation Data Reliability

Geolocation data is a powerful source of information helpful for understanding how people move around a place, both in their day-to-day routines, and when they travel away from home.

As powerful as it is, geolocation data is also imperfect. Inherent in this valuable data stream are imprecise observations that can interfere with the information and even generate counterintuitive results. As with any source of big data, geolocation data must be cleaned and balanced before it can provide reliable insights for analysis or decision-making.

Hotspots are one of the more perplexing issues found in geolocation data sets. All geolocation data sets contain a mixture of highly precise and approximate location observations. While the

precise observations mirror the expected patterns of human movement, the approximate observations tend to cluster in centralized locations, which we call hotspots.

To understand the difference between precise and approximate observations — a typical cause of certain types of hotspots — imagine this interaction with a mapping app:

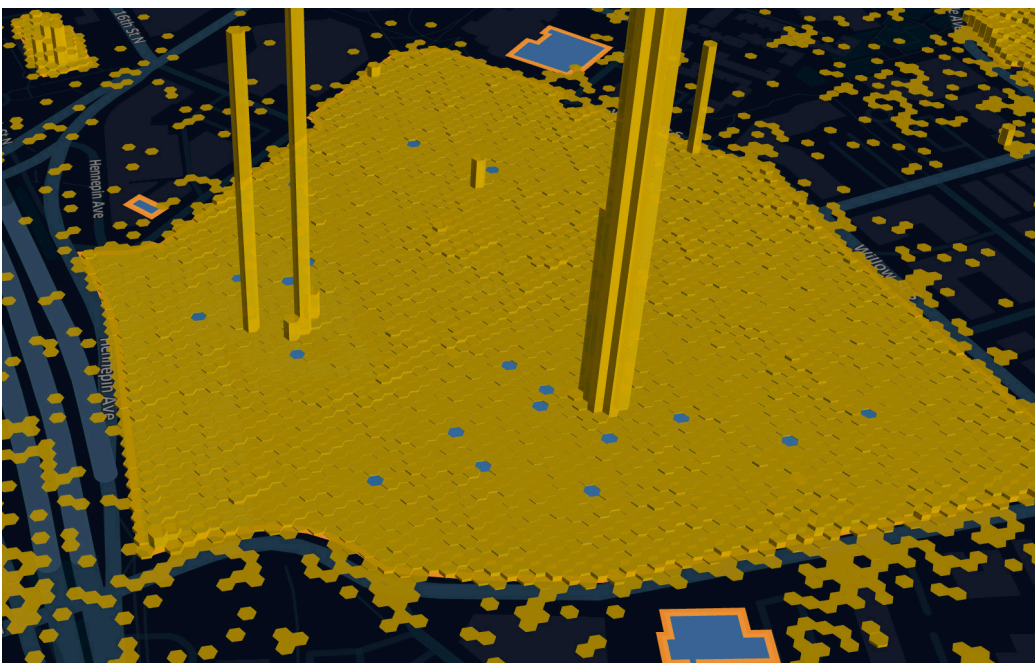
- In one instance you search for the location of Salt Lake City. Your map result will display the entire city and may drop a location marker in the approximate center of the city limits.
- Alternatively, if you search for the location of the eastbound light rail station on the north side of the Vivint Arena in Salt Lake City, the location marker will fall precisely on that place.

Now let's say that you're standing in this light rail station, using your mapping app to plan your route home after a concert. If the app records your precise location, then the geolocation observation would match the latitude and longitude coordinates of the station. But if it instead records an approximate location, it may position you somewhere in the "middle" of Salt Lake, potentially miles away from your actual location.

In short, hotspots interfere with the organic geolocation signal by locating a large number of devices in a single point in space, when in reality those devices were distributed throughout the area.

Although the exact cause of hotspots is unknown, we have clearly identified the resulting behavior, allowing us to confidently remove them from the data set. Furthermore, Zartico's statistical approach to hotspot filtering allows us to remove only the approximate observations, while retaining precise observations from the same device.

“Hotspots interfere with the organic geolocation signal by locating a large number of devices in a single point in space, when in reality those devices were distributed throughout the area.”



This visualization illustrates hotspots in Loring Park, a 34-acre park in Minneapolis, Minn. Here you can see how the hotspots concentrate approximate geolocation observations into very small areas. Each of the hexagons in this image covers an area roughly 25 feet wide.

Spotting the Spikes

We first noticed the effects of hotspots in some of our partner destinations when certain points of interest (or POIs) began to show way more visitor traffic than would make sense. For example, small city parks or municipal golf courses were showing inflated visitation numbers over and above marquee attractions or high-traffic convention centers.

When we visualized these destinations, plotting the location observations on a map like a three-dimensional bar chart, we noticed enormous spikes — we could actually see the hotspots. It was immediately apparent that these were not representative of human movement as it is naturally dispersed throughout the region.

When these hotspots manifest within the boundaries of a designated POI, such as a park or golf course, they don't only skew the device count observed at those locations. Hotspots can also throw off trends related to categories of POIs, origin markets visiting these POI categories, or cross-visitation related to the affected POIs.

Hotspots are not fixed. They can appear and disappear over time, or become more or less pronounced depending on several factors, including the collection of apps that are delivering approximate observations and how those apps are being used by destination visitors. This means that hotspots can also interfere with understanding changing trends related to a POI or category of POIs over time.



In this visualization of Salt Lake City, the white spikes denote the locations of hotspots. After the filter is applied, the remaining red spikes reflect the natural movement of people throughout the city, showing higher concentrations of individuals along major roadways and in downtown districts.

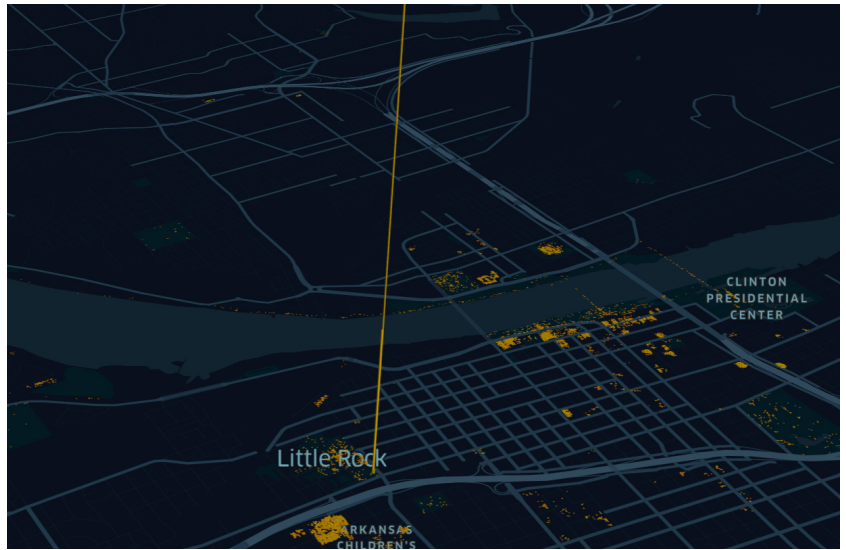
Solutions That Fall Short

In any given area, there are often two or three massive hotspots that dwarf the remaining geolocation signal. If you were looking at these plotted on a map, you may say that these are the hotspots that could be seen from space (see image). Because there are relatively few of these and because they are the most persistent, they are also the most easily removed.

But once the extreme outliers are eliminated, the remaining hotspots come into focus. These have been the most difficult to address at scale. Data teams at other companies have tried a number of different methods, but their techniques face some significant challenges.

First, a blanket hotspot removal risks being too aggressive, resulting in the loss of too much precise traffic. These aggressive filters may mistakenly target areas of naturally high density — such as airports or high-attendance events — where our destination partners actually need to see the clearest geolocation signal.

The second method is to create a long list of exclusionary criteria — a laundry list of the ways and places that hotspots manifest based on observed device behavior. However, this method does not account for new causes of hotspots or hotspots behaving in new ways outside the parameters of the established criteria.



In addition to dwarfing the genuine data, the scale of these massive hotspots also hides many more hotspots that incorrectly inflate observation counts.

“We are able to **selectively excise** approximate observations while maintaining **95% of the precise geolocation observations** — isolating and amplifying the good signal beneath the noise.”

Building a Dynamic Filter

Our approach on the Zartico data science team instead looked for ways to stay ahead of the curve — to understand all of the ways that hotspots could behave and to proactively predict and neutralize them.

So how do you target an anomaly with an unknown cause that also shifts and changes over time?

While the behavior and location of hotspots may not follow a particular pattern, they do break another established pattern — the pattern of how people naturally move and distribute themselves within a space.

Our solution was to create an advanced mathematical procedure to do exactly what the human eye does when it looks at hotspots on a map — pick out the places where the data “looks” unnatural, where it doesn’t follow the rules of human movement.

This statistical approach can be applied across any given map and is not limited by time or geography, so it can dynamically identify hotspots, even as their behavior changes. In this way, we are able to selectively excise approximate observations while maintaining 95% of the precise geolocation observations — isolating and amplifying the good signal beneath the noise.

One way we retain the good signal is by filtering out approximate observations while keeping precise observations from the same device. Geolocation data is captured by a variety of different apps and sources, so it’s possible that one app on your phone is recording approximate locations while another is recording precise ones. Retaining the precise observations from these devices allows us to maximize the strength and accuracy of the signal.

Mean-field Game Theory

This emerging area of study looks at the ways that individually motivated real-time decisions inform the movement of a large group within space and time. If you imagine a time-lapse video of a large crowd of people exiting a venue, for example, you would see how the crowd flows through the space almost like a liquid — not dissimilar from the way a school of fish or a flock of starlings might move.

The theory behind these movements involves each individual in a dense population making tiny choices or adjustments in reaction to the individuals around them. If you were exiting that same venue, you would be constantly making small optimizations to find your path to the door, shifting each moment based on the people around you and their movements.

In short, if you have enough players making independent decisions in a space, they will behave like fluids, distributing in a predictable way. When it comes to hotspots, we can use this law of motion to see where a crowd is behaving in ways that are organic, or where the anomalous behavior looks unphysical.



What We've Learned

This is no simple task.

Hotspots in geolocation data streams present a difficult obstacle to isolating and understanding the significance of the sound geolocation signal. And combating these anomalies requires enormous amounts of data science work. Even the most exhaustive heuristic or technical solution cannot account for all of the ways hotspots manifest or where they may pop up tomorrow.

A data science approach to these challenges instead operates on first principles that apply universally, ensuring that no matter what changes in the future, we can still build tools that will be robust and adaptable to any uncertainties to come.



ZARTICO

zartico.com