Evaluating the Usefulness of Big Data to Glacier National Park Management

Interactive Qualifying Project Report

An Interactive Qualifying Project submitted to the faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the Degree of Bachelor of Science

Submitted to:

Dr. Seth Tuler

Dr. Robert Traver

Worcester Polytechnic Institute

Submitted by:

Isabella Clowes

Owen McGinley

Ethan Shaw

Aidan Syrgak Uulu

Date:

October 12th, 2022

Abstract

Glacier National Park (GNP) is interested in whether big data can improve its current data collection and analysis methods. However, GNP management does not know whether big data will be useful. Our team used a big data provider, Zartico, to examine Park management's concern. We assessed usefulness with three criteria: credibility, labor-intensiveness, and relevance. An application of the criteria indicates that GNP management's needs are not entirely addressed by big data. We review how big data might and might not be used. Finally, we suggest how GNP management can think critically about a big data provider before investing in one.

Executive Summary

The purpose of this project was to evaluate the usefulness of big data to Glacier National Park Management. We found that big data is useful to track general visitor location and movement, determine where visitors are coming from, monitor traffic congestion, and as a tool to compare against current park data. However, we found that big data has limitations: imprecise visitor tracking data, no spending data inside the Park, and not granular enough demographic data. The application of big data to wilderness areas like the Park is new. Big data is an emerging technology so it may not yet be worth the investment. Thus, we recommend a list of questions for Park staff to ask big data providers before investing in big data.

We started by identifying the informational needs of the Park. The needs were expressed in a list of thirty questions. We iterated the list with the help of GNP staff. The list was then categorized by priority and topic. This helped us better understand the Park's needs when assessing big data.

With this list of questions, we used Zartico's platform as a case study to assess big data's ability to meet the Park's informational needs. By doing this, we created graphics to assist in answering the question with data in a presentable way. Here we determined how well Zartico's data answered the questions. The obstacles that prevented us from answering the questions became the limitations of Zartico. Anything that contributed to answering the questions was a capability of Zartico's data. We compared Zartico's data to GNP's current data to determine if the data's accuracy in the Park contributed to additional limitations.

We developed criteria to assess how big data in general could be useful to Park management. Our experience using Zartico's platform helped us to formulate the initial criteria. We grouped the platform's capabilities and limitations into a few overarching themes. These themes became the initial criteria. The criteria were categorized into three sections: credibility, labor-intensity, and relevance. We then compared our criteria to the relevant literature that described the markers of useful information. We finalized our criteria by confirming with GNP staff that the criteria were relevant to Park management.

The criteria were used to evaluate Zartico's platform. Using our analyses of capabilities and limitations, we determined how well the platform meets each criterion. We also researched a previous study done by another WPI project group who used a different big data provider, Streetlight Insights. We recognized the usefulness of Streetlight Insights' data. Through our evaluation of Zartico and research of Streetlight Insights, we found our results.

Acknowledgements

Our team would like to thank Glacier National Park Deputy Superintendent, Pete Webster and Chief of Science and Resources Management, Phil Wilson. Our gratitude extends to Katherine Barrs, the Visitor Use Management Assistant for Glacier National Park, for her time and effort in looking over our work and providing us with necessary information regarding the Park. We would like to thank Katie Stadius and Bill Obreiter from Zartico for training us to use Zartico's Destination Operating System. We would also like to thank Glacier Country Tourism for providing us with access to Zartico. Lastly, we would like to thank our advisors, Dr. Robert Traver and Dr. Seth Tuler for their support and guidance throughout the duration of this project.

Authorship

All authors participated in the editing of each section.

Section	Author
1.0 Introduction	Isabella Clowes
2.0 Background	Isabella Clowes
2.1 Information Needs for Park Management	Ethan Shaw
2.2 Collecting Information	Ethan Shaw
2.2.1 Current Collection Strategies	Ethan Shaw
2.3 Big Data: A Better Way to Collect Information?	Isabella Clowes
2.3.1 What is Big Data?	Isabella Clowes
2.3.2 The Benefits of Big Data	Isabella Clowes
2.3.3 Big Data in Tourism	Isabella Clowes
2.4 Big Data Providers	Owen McGinley
2.4.1 Comparison of Zartico and Streetlight Insights as Big Data Providers	Owen McGinley
2.4.2 Potential uses of Streelight and Zartico in GNP Management Strategies	Owen McGinley
3.0 Methodology	Aidan Syrgak
3.1 List of Questions GNP Wants Answered	Aidan Syrgak
3.1.1 Generating a List of Questions	Aidan Syrgak
3.2 Data Visualization and Analysis	Owen McGinley
3.2.1 Using Zartico's Platform to Analyze Data	Owen McGinley
3.2.2 Creating Diagrams	Ethan Shaw
3.2.3 Capabilities and Limitations	Owen McGinley
3.3 Creation of Criteria to Evaluate Usefulness	Aidan Syrgak
3.3.1 Creating and Testing Criteria	Aidan Syrgak

4.0 Results	Owen McGinley
4.1 Criteria for Evaluating Usefulness of a Big Data Platform	Aidan Syrgak
4.2 Credibility of Zartico's Data	Owen McGinley
4.2.1 Analyzing Validity of Zartico's Data	Owen McGinley
4.2.2 Time Period for Which Zartico's Data is Available	Owen McGinley
4.2.3 Volume of Zartico's Data	Owen McGinley
4.3 Labor-Intensiveness of Zartico's Data	Ethan Shaw
4.3.1 Expertise Required to Use the Platform	Ethan Shaw
4.3.2 Time Spent	Ethan Shaw
4.4 Relevance of Zartico's Data to Management Decision-Making in Glacier National Park	Isabella Clowes
4.4.1 Coverage of Relevant Location	Isabella Clowes
4.4.2 Granularity	Isabella Clowes
4.4.3 Presence of Data Necessary to Answer GNP's Questions	Isabella Clowes
5.0 Recommendations and Discussion	Isabella Clowes
5.1 Questions to Ask Before Implementing Big Data	All
5.2 Where Big Data is Useful for GNP Management	Isabella Clowes
5.2.1 Visitor Location and Movement	Isabella Clowes
5.2.2 Traffic Congestion	Isabella Clowes
5.2.3 Comparison Tool	Isabella Clowes
5.3 The Issues of Big Data for GNP Management	Owen McGinley & Aidan Syrgak
5.3.1 Big Data is a Sample	Owen McGinley
5.3.2 Spending Inside and Outside the Park	Owen McGinley & Ethan Shaw
5.3.3 Risk of Project Failure	Owen McGinley

Table of Contents

List of Figures

List of Tables

1.0 Introduction

National Park service managers affect many people with their decisions (Machlis, 1996). Therefore, NPS managers need reliable information to make informed decisions. With better information, they can improve their decisions (National Research Council, 1992).

Managers rely on information taken from many sources to make decisions. Glacier National Park (GNP) management currently gathers information about park conditions and visitors in multiple ways. Their decisions to close parking lots, campgrounds, and roads have been based on information derived from traffic counters, trail cameras, and communication between rangers. Management decisions on how to limit visitation via a reservation system are also based on this information. However, GNP feels that alternate sources of information can improve informed decision-making and planning (P. Webster, Personal Communication, September 2022). An alternate source of information could be derived from big data.

Big data refers to very large and less structured data that cannot be analyzed using traditional programs like Microsoft Excel. Big data is collected with three V's: a faster velocity, higher volume, and wider variety than traditional data (Laney, 2001). These features attract organizations to use big data because "decisions of various types can be improved by adding big data" (Davenport, 2014). Previous studies in urban park systems have used big data to find where visitors are traveling from, allowing for better decisions that promoted equity of access to green spaces (Xiao et al., 2019). It is possible that GNP management can use big data to find missing information or improve the quality of information already available. Additionally, big data could make the process of collecting and analyzing data easier (Lee, 2017).

The goal of our project was to evaluate the usefulness of big data for GNP management and inform their future decisions on using big data. GNP management lacks sufficient information to know if big data would be worth investing in. We used the tools and data from a big data provider, Zartico, to explore the potential usefulness of big data in aiding GNP management to make better data-driven decisions. We then created criteria to evaluate the big data based on literature reviews and discussions with GNP employees.

2.0 Background

The following background section describes GNP management's need for information about its visitors to make effective management decisions. The Park has data collection and analysis methods to acquire information. However, big data analysis could be an improved method of acquiring information.

2.1 Information Needs for Park Management

Access to good data is crucial for organizational decision-making: "[the] largest and most successful organizations use data to their advantage when making high-impact business decisions" (Stobierski, 2019). Making decisions in business as well as in National Parks requires data. Without data about visitor behavior and park conditions, National Parks could not improve their management decisions. Data must be salient, credible, and legitimate to be effectively utilized in decision-making (Cash et al., 2002). Data analysis allows "faster identification of problems, greater understanding of cause and effects, and better insights about the prevention and management of problems" (National Research Council, 1992).

To make informed decisions, GNP management needs detailed information about their visitors. They must have information on visitor demographics, visitor flow, traffic count, and traffic congestion. Knowing how their decisions affect the local community also helps improve decisions-making (K. Barrs, personal communication, September 2022).

GNP management also has specific informational needs to understand its visitors. Park management wants to understand visitor demographics and visitor flow. They want to visualize traffic count and congestion, especially since National Park attendance rates have steadily increased in recent years (Simmonds et al, 2020). Additionally, the Park is interested in how their decisions affect the local community (K. Barrs, personal communication, September 2022).

2.2 Collecting Data

GNP uses both mechanical and digital methods of obtaining data. The Park stores the data they obtain in their IRMA (Integrated Resource Management Applications) visitor statistics. However, these data collection methods have their limitations.

2.2.1 Current Collection Strategies

There are two main data collection strategies within GNP. The first strategy is the use of traffic counters placed in strategic parts of the Park. These mechanical devices detect traffic on roads and through entrances (K. Barrs, personal communication, September 2022). When and how many cars enter the Park is tracked. The car count is used to determine total visitation. The total number of cars is multiplied by the average number of passengers in a car, determined to be 2.9 (P. Webster, personal communication, September 2022). This strategy only gives a rough estimate of the visitors that travel to the Park since not all cars have 2.9 passengers. Also, it is hard to precisely know how many traffic counters a single car will cross, so there is room for error. Another limitation of this strategy is the time spent placing traffic counters at each location. Setting them up and taking them down requires manual work. The counters also need to be removed before snowfall to avoid the destruction of equipment.

The second strategy uses the National Park Service website to digitally collect data. When a reservation is purchased, the Park receives the ZIP code on the debit/credit card used (K.

Barrs, personal communication, September 2022). This gives the Park management a rough estimate of where visitors traveled from. However, this method also has its limitations. Firstly, not every person who purchases a reservation ends up traveling to the Park. On the other hand, there are many people who travel to the Park but do not have a reservation. This leads to inconsistencies in the recorded data. This was an improvement over the previous method of having rangers record the state of license plates in parking lots.

Other strategies include trail counters and game cameras. Trail counters are a beam of infrared light that extends across a trail entrance. It counts every time an object crosses the light; this gives an estimate of the number of visitors on certain trails. Game cameras are small cameras that are primarily used to look at parking lots. It allows whoever is looking at the cameras to know when the parking lot is full, as well as get an estimate on what times they fill.

2.3 Big Data: A Better Way to Collect Information?

One solution that interests GNP management is big data. Big data can reveal "trends, patterns, and correlations … to help make data-informed decisions" (Tableau). Big data has been applied to the tourism industry, but not without limitations.

2.3.1 What is Big Data?

Big data consists of gigabytes or more of data collected across an extended period of time with numerous data points. Big data is "too big to fit on a single server, too unstructured to fit into a row-and-column database, or too continuously flowing to fit into a static data warehouse" (Davenport, 2014). It is characterized by the three V's: Volume, Velocity, and Variety. Big data arrives in large volumes, is generated and delivered at a high speed, and is available in a variety of formats (Laney, 2001). This data can be analyzed to gain helpful information.

2.3.2 The Benefits of Big Data

Benefits of big data include easy data collection and continuous, fast data analysis in many industries that outcompete the successes of traditional data.

A benefit of big data is the ease of data collection. Big data streams in from online sources collected by computers, rather than manually. Once the collection process is established, data can continuously be collected (Lee, 2017). This could help GNP save time and money during the data collection process.

Big data also allows for constant mining of information. Big data can come in at a higher velocity and volume than other data, such as GNP's current data from traffic counters as mentioned in section *2.2.1* (Current Collection Strategies). In some cases, the data can be continuously analyzed as the data streams in. "Streaming analytics involve real-time analysis to discover patterns of interest as data is being collected or generated" (Lee, 2017). Streaming analytics allow current conditions to be monitored in real-time. For GNP, a real-time feed of something like visitor movement and congestion could be useful.

Using big data has proven successful for many organizations including Netflix, Google, and even the United Nations (Davenport, 2014). Stores "such as Macy's and Target use big data to analyze shoppers' preferences and sentiments and improve their shopping experience" (Lee, 2017). Big data has also been used to determine how traffic congestion patterns change due to

different factors (Cote et al., 2021). Big data analysis has assisted many different industries including “science, engineering, healthcare, management, business”, etc. (Li et al., 2018).

2.3.3 Strengths and Weaknesses of Big Data in Tourism

Big data has been applied to the tourism industry as well. Tourist organizations can benefit from big data, but certain limitations arise when applying big data to tourism.

Big data related to tourism can be defined by user generated data, device data, and transactions or operations data (Li et al., 2018). User generated data comes from social media through images and texts. Device data includes geospatial location and data from Bluetooth devices. Operations data is collected from web searches, purchases, and bookings. A few studies showed that big data “allowed a better understanding of tourism demand, tourist behavior, tourist satisfaction, and other tourism issues” (Li et al., 2018). For example, destination-management agencies in Sydney and Melbourne, Australia, found success using GPS tracking to understand spatial behavior of tourists. They learned interesting aspects about the visitor experience such as how a lack of knowledge of the public transportation systems restricted visitors from using the transportation (Edwards & Griffin, 2013).

Applying tourism-related big data to National Parks is different from applying the data to tourist spaces like Sydney and Melbourne. National Park managers require different types of relevant data to improve their decisions. A needs analysis completed by Chaang-Iuan Ho and Yulan Yuan revealed specific needs of National Park managers: “monitor visitor flow and distribution”, “prevent illegal recreational uses”, “ease traffic conditions”, “prevent poaching and other illegal activities”, “increase environmental awareness”, “aid [in] interpretation [of] staff tasks”, “assist risk management”, “marketing and demarketing”, “collect public opinions and complaints”, and “shape planning and policy formation” (Ho & Yuan, 2017).

These needs make applying big data to National Parks a challenge because the big data to address these needs may not exist. When researchers interviewed National Park staff, some of the concerns causing hesitation to implement big data were a “lack of clarity [in] the type of data and information needed to resolve problems” and a “lack of existing data available” to complete big data analysis (Ho & Yuan, 2017). This issue is probably due to the newness of big data in tourism research. The “tourism research using big data [is] still at an early age, in terms of a short history (just beginning since 2007)” (Li et al., 2018).

Much of tourist big data uses device locations but has limitations due to privacy restrictions and bias. Device location data consists of periodic location pings because devices cannot be continuously tracked due to privacy restrictions. This means a device’s complete route cannot be tracked. It is also impossible to track how long a device stays at one location or estimate how long a device’s trip takes (B. Obreiter, Zartico Vice President, Personal Communication, August 2022).

This location data is reported by smartphone apps that track location. It is possible some apps report less location pings in areas with no cellular service, such as GNP (US DOI, 2021). Even if location pings are saved offline then reported when back online, people tend not to use devices without a connection.

Location data has bias favoring people with smartphones and apps that track location (e.g., Google, Instagram, and All Trails). Some people may not have a smartphone or apps that can be tracked by big data providers. Additionally, some people may not have a smartphone at all. Big data collection does not have the same equality that vehicle counters or trail counters

have. These issues could impact GNP's ability to use big data. Our team explored these types of issues to determine if tourism-related big data is useful to GNP.

2.4 Big Data Providers

This section details two tourism-related big data providers: Zartico and Streetlight Insights (SLI). A 2021 WPI project group used SLI in GNP. Our team used Zartico to consider the overall usefulness of big data to GNP management and how big data could be used in the future.

2.4.1 Comparison of Zartico and Streetlight as Big Data Providers

Table 1 shows a comparison between the primary forms of data that Zartico and SLI focus on.

Table 1.

Comparison Between Zartico and Streetlight Insights

Zartico	**Streetlight Insights**
Type of Data: ● Visitor Location ● Spending ● Event	Type of Data: ● Traffic ● Congestion

2.4.2 Potential uses of Streetlight and Zartico in GNP Management Strategies

A 2021 WPI project group in GNP used SLI to analyze the effects of the Going-to-the-Sun Road (GTSR) reservation system using big data (Cote et al. 2021). The project team describes SLI as "an on-demand web platform for transportation data. It tracks, collects, and analyzes big data regarding vehicular traffic for different modes of transportation across the United States and Canada" (Cote et al. 2021). Using SLI's data, the group examined the reservation system's impact on GTSR's congestion. They analyzed congestion levels and the average speed of traffic in different sections of the road. This type of big data could be useful to Park management, as managing traffic congestion remains one of their top priorities.

In contrast to SLI, Zartico does not focus on one specific topic. Instead, Zartico advertises three data streams:

1) Daily anonymized geolocation data "of over 1.6 billion mobile devices,"
2) Event data "sourced from Ticketmaster, Seat Geek, LiveNation, Eventbrite, and others," and
3) "[Spending] data from over 3,000 financial institutions…With over 10 billion transactions annually" (Zartico, 2022)

Zartico focuses on how visitors impact a surrounding community, an interest to Park management. Our project will focus on exploring Zartico's data, then using a set of criteria to evaluate the capabilities and how they relate to the Park's interest in using big data.

3.0 Methodology

The goal of this project was to evaluate the usefulness of big data to GNP's management. Our team accomplished three objectives to achieve the goal:

1. Create a list of questions to guide the assessment of big data for GNP's needs
2. Explore the capabilities and limitations of Zartico as an exemplar of big data
3. Establish the criteria to evaluate the usefulness of big data to GNP

This section describes the methods that we used to achieve these objectives. We completed each one in the order mentioned above. The list of questions in the first objective was necessary to generate responses for the second objective. These responses were then used to establish the criteria for our final objective.

3.1 List of Questions to Guide the Assessment of Big Data

Our team identified GNP's informational needs to assess the usefulness of big data to the Park. These needs were expressed as questions. We generated a list of thirty questions that we iterated with Park staff. We conducted meetings with Kat Barrs, a data analysis intern, and Pete Webster, deputy superintendent. They helped us to create a list that was categorized by high, medium, and low priority questions. Additionally, the questions were classified by topic: economic, demographic, descriptive statistics on GNP visitors, reservation system changes, and congestion. The categorizations by priority and topic helped us to understand the general nature of the Park's informational needs. Some examples of high priority questions classified by topic are shown in Table 2. The full list of questions can be seen in Appendix A.

Table 2.

High priority questions categorized by topic

	High Priority Questions
Demographic	Where are the visitors coming from? (local, out-of-state, Canada, etc.)
Descriptive Visitor Statistics	Are there areas within GNP that a higher % of locals visit compared to other areas?
Reservation System Changes	What percentage of visitors enter outside of the reservation time? (outside of 6am-4pm)

3.2 Capabilities and Limitations of Zartico as an Exemplar of Big Data

With the list of questions generated, we used Zartico's platform to produce responses to the questions. We analyzed how well our responses answered each question or determined whether the questions can be answered at all. This allowed our team to identify the capabilities and limitations of Zartico's platform.

3.2.1 Using Zartico's Platform to Analyze Data

First, we scanned through the available data in Zartico and picked the most relevant data to answer the question at hand. Second, we used Zartico's filters to narrow the data down to only within GNP. We did this by filtering data to include Glacier & Flathead Counties, and only include outdoor recreation points of interest. Finally, we applied additional filters specific to each question. For many questions, we filtered the data by date range, where devices were coming from, and if the device was classified as a non-local in Zartico's platform. The filtered data was presented in table format, as shown in Table 3.

Table 3.

Example of How Zartico Presented the Data in Table Format

POI Name	% Visitors ▼
Dusty Star Observatory	92.0%
Rocky Point	91.4%
Granite Park	91.4%
Logan Pass Visitor Center Trails	91.1%
Trail of the Cedars	90.9%
Two Med Boat Dock	90.9%
Hidden Falls	90.8%
Cut Bank Trail System	90.8%
Rising Sun	90.7%
Glacier Raft Company	90.3%

3.2.2 Creating Graphics

We used the filtered data to create graphics with the built-in graphing tool in Zartico's platform. We created many charts and graphs that addressed the questions from our list. However, we had to refine the graphics to make them more understandable and presentable. We had to use additional software to improve the graphs created by Zartico.

As a result of that, we transferred the filtered data from Zartico's graphs to Excel. We used Excel because we had more options for graph customization than in Zartico's built-in graphing tool. We also compared multiple years of data in Excel, a feature that was unavailable in Zartico's platform. To enhance our analysis of the data, we usually created many versions of graphs to provide better support for each question.

3.2.3 Identifying Capabilities and Limitations

Using the created graphics, we determined how well the data addressed the question at hand. When we were able to respond to the question, we identified the capabilities of Zartico that enabled us to produce a response. For partially answered or unanswerable questions, we considered the obstacles that prevented us from producing a response. These were identified as the limitations of the Zartico platform.

We assessed how accurate Zartico's data was to GNP's IRMA data. IRMA data has monthly visitor counts at each Park entrance since 1979 and is trusted to be accurate by the Park management. Whenever we found inaccuracies in Zartico's data, we identified this as a limitation. This second pass of data comparison allowed us to fully comprehend where Zartico could and could not be used to reliably answer questions.

3.3 Criteria to Evaluate the Usefulness of Big Data to GNP

The third objective was to develop criteria to evaluate the usefulness of big data platforms in general. The intent was to create a tool that GNP can use to determine the value a big data platform can bring to Park management. These criteria helped us formulate results and recommendations for GNP staff.

3.3.1 Creating and Testing Criteria

We conducted a coding analysis, compared our coding to the relevant literature, and talked with GNP staff to develop the criteria. First, we conducted a coding analysis using the capabilities and limitations of Zartico's platform. The capabilities and limitations were grouped into categories. Then, we organized those categories into a few overarching themes that summarized the performance of Zartico's platform. The coding analysis can be seen in Appendix B. Second, we assured that the created themes reflect the literature on useful information. We compared our themes to the markers of useful information (salience, credibility, and legitimacy) that were established by Cash et. al (2002). These comparisons allowed us to justify our themes and turn them into criteria for evaluating the usefulness of big data. Third, we talked with GNP staff to make sure that the criteria are relevant to GNP. We met with Kat Barrs, Phil Wilson (Chief of Science and Resources Management), and Pete Webster to discuss what GNP wanted from a big data analysis platform. They confirmed our initial criteria and suggested new ones.

As the last step, our team used the established criteria to evaluate Zartico's platform. Using our analyses of capabilities and limitations of Zartico, we described how well the platform meets each criterion. This helped us to assess the usefulness of Zartico to GNP management.

4.0 Results

In this chapter, we present the results of our study. We do this by identifying the criteria for evaluating the usefulness of a big data platform. Then, we evaluate each criterion and how it relates to the overall usefulness of big data.

4.1 Criteria for Evaluating Usefulness of a Big Data Platform

Our team identified three criteria to evaluate the usefulness of big data platforms to GNP management: credibility, labor-intensiveness, and relevance. Each criterion was defined by three components shown in Table 4.

Table 4.

Components of the Criteria

Credibility	Labor-Intensiveness	Relevance
● Validity of Data ● Time period for which data is available ● Volume of Data	● Cost ● Time ● Expertise	● Presence of necessary data ● Coverage of relevant location ● Granularity

Credibility encompasses three components: validity of data, time period for which data is available, and volume of data. *Validity of data* refers to accuracy of data. The data collection method needs to be reliable. Additionally, the data must match the trends and patterns taken from other sources, like IRMA visitor statistics. *Time period for which data is available* is the span of years that we can access on the platform. Having access to data for longer durations of time allows to establish historical trends and patterns. *Volume of data* refers to the size of data, and whether it can be identified as big data.

Labor-intensiveness refers to the cost, expertise, and time required to collect and analyze data. Buying access to a big data platform from a provider may have a high *cost*. Another factor is the presence of necessary personnel with the *expertise* to use the platform. Hiring an additional worker will add extra cost. Additionally, big data analysis requires *time*. It takes time and training to learn how to utilize a big data platform.

Relevance refers to the extent to which the data applies to GNP's needs. Three components define relevance: presence of necessary data, coverage of a relevant location, and granularity. *Presence of necessary data* ensures that a big data platform has the necessary types of data for GNP's needs. *Coverage of a relevant location* means that a big data platform has data within GNP. The platform might have relevant types of data, but the data is not relevant if it does not cover the area of the Park. *Granularity* is the level of detail within the data. It determines, for example, if we can view the data by hour of the day or a specific region of the Park.

4.2 Credibility of Zartico's Data

The credibility of Zartico's location data was analyzed using the three credibility components mentioned in the previous section. **We found that Zartico data is valid in the**

summer, but lacks validity in the winter, with some data not being valid at all. Additionally, Zartico does not have enough historical data to make meaningful historical comparisons and we are concerned about small sample sizes affecting results.

4.2.1 Validity of Zartico's Data

Zartico's data appeared to be more valid during the summer months than the winter months. We found and confirmed with Zartico employees that trip type and trip duration data were inaccurate. Additionally, demographic data was inaccurate, with the sum of races visiting the Park adding up to over 100%.

We assessed the validity of Zartico's data by making comparisons between total visitation numbers and IRMA visitor statistics. To ensure we were evaluating validity with the most up-to-date data collection methods, we used 2021 visitor numbers in our comparison.

To make this direct comparison, we calculated an inferred visitor count from Zartico's data. We did this by extrapolating the unique devices Zartico logged each month by the overall unique device-to-visitors ratio for all of 2021. The result of this comparison is shown in Figure 1.

Figure 1 shows that Zartico's inferred visitor count in 2021 was within 1% of the IRMA's visitor count between June and October, the peak visitation months. This means Zartico's portion of yearly devices sampled during these months was very close to the portion of yearly visitors that visited during those same months.

However, as shown in Figure 2, Zartico's data deviates in the off-season. Inferred visitors are overcounted by as much as 216% in March. We found a similar pattern in 2019 data where the off-season had overcounted inferred visitors by 200% or more. This meant that Zartico's portion of yearly devices sampled in the off-season was much higher than the portion of yearly visitors who came to the Park during the off-season.

Figure 1.

Glacier National Park Monthly Visitor Numbers and Zartico Inferred Monthly Visitor Numbers for 2021

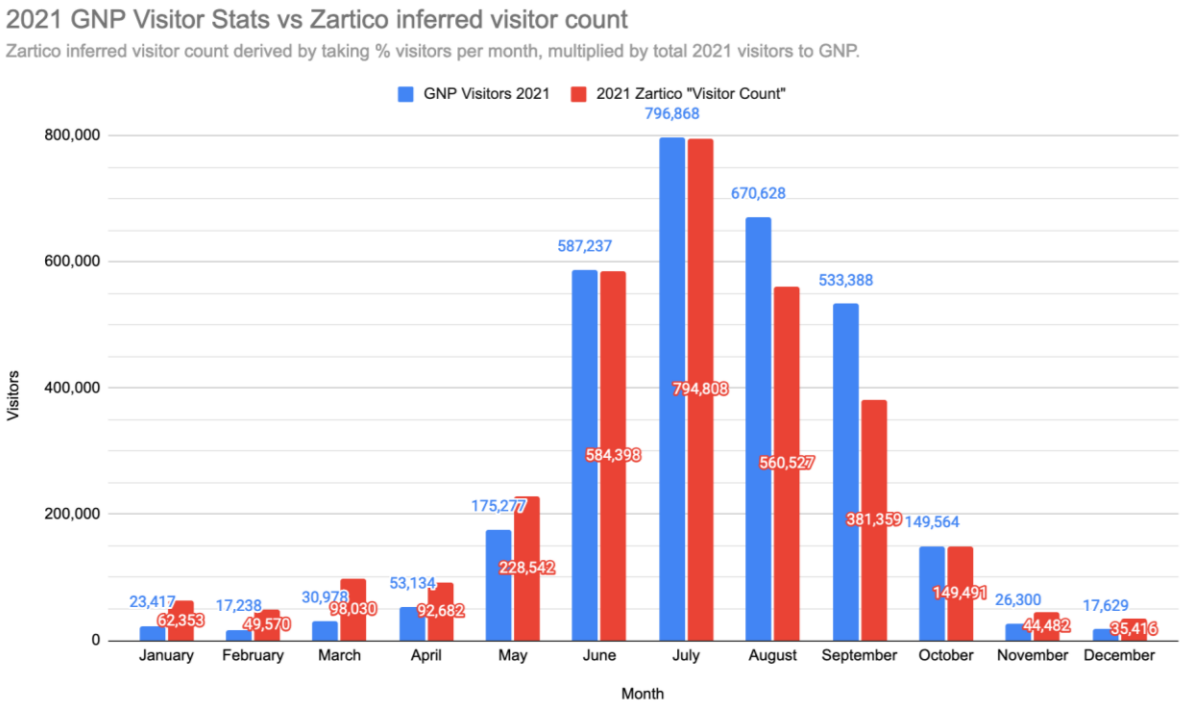
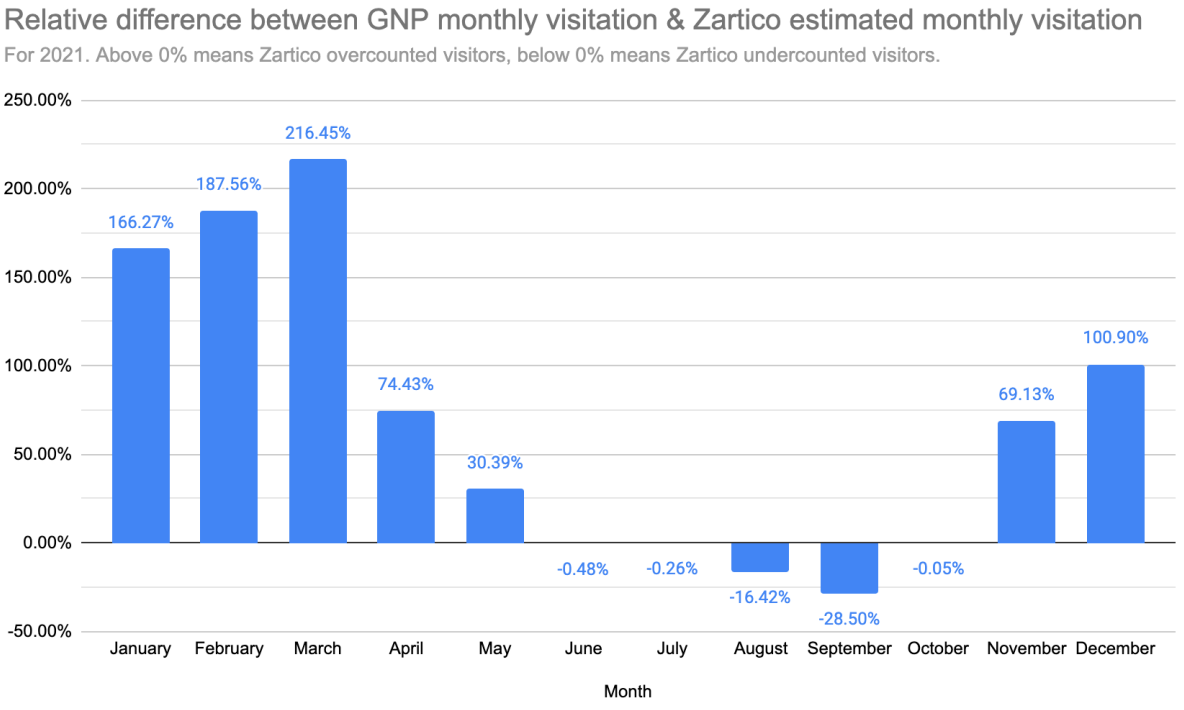


Figure 2.

Difference Between GNP Monthly Visitation and Zartico Estimated Monthly Visitation for 2021



4.2.2 Time Period for Which Zartico's Data is Available

A lack of historical data over a sufficient time period undermines the credibility of historical comparisons using Zartico's data.

Park management questioned how the implementation of the reservation system on GTSR in 2020 impacted visitation in the Park. By answering this question, we were able to evaluate Zartico's historical data availability. To do this, we analyzed how visitation changed month-by-month from 2019 to 2021. We ran the same comparisons using GNP IRMA visitation data as a baseline.

In Figure 3, Zartico's data shows visitation drops by about 20% in August and September then increase in October. The IRMA visitation numbers in Figure 4 show a drop occurred in July and August, meaning Zartico showed the drop in the wrong months. Furthermore, visitation increased 88% in October, whereas Zartico's data only shows a 48% increase in October.

Zartico only has historical data back to January 1, 2019. Due to the COVID-19 pandemic in 2020 and data being incomplete in 2022, it was impossible for us to analyze the credibility of the results in Figure 3. We would need more years of historical data to judge credibility off historical comparisons.

Figure 3.

Percent Change in Visitation by Month From 2019 to 2021 Using Zartico

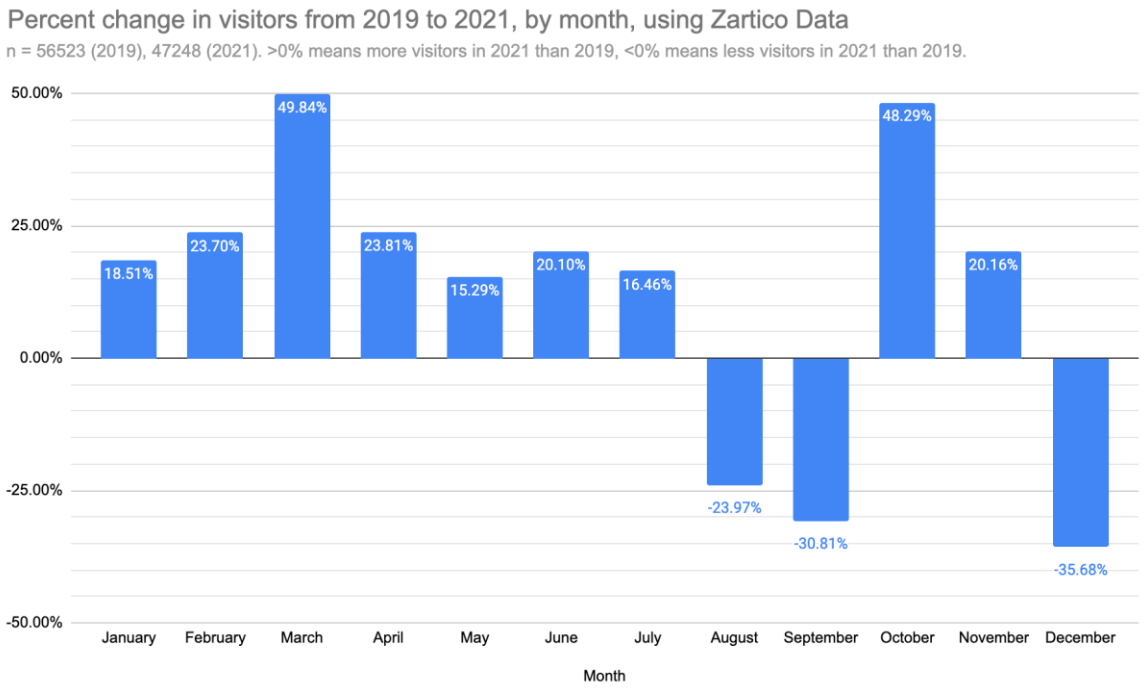
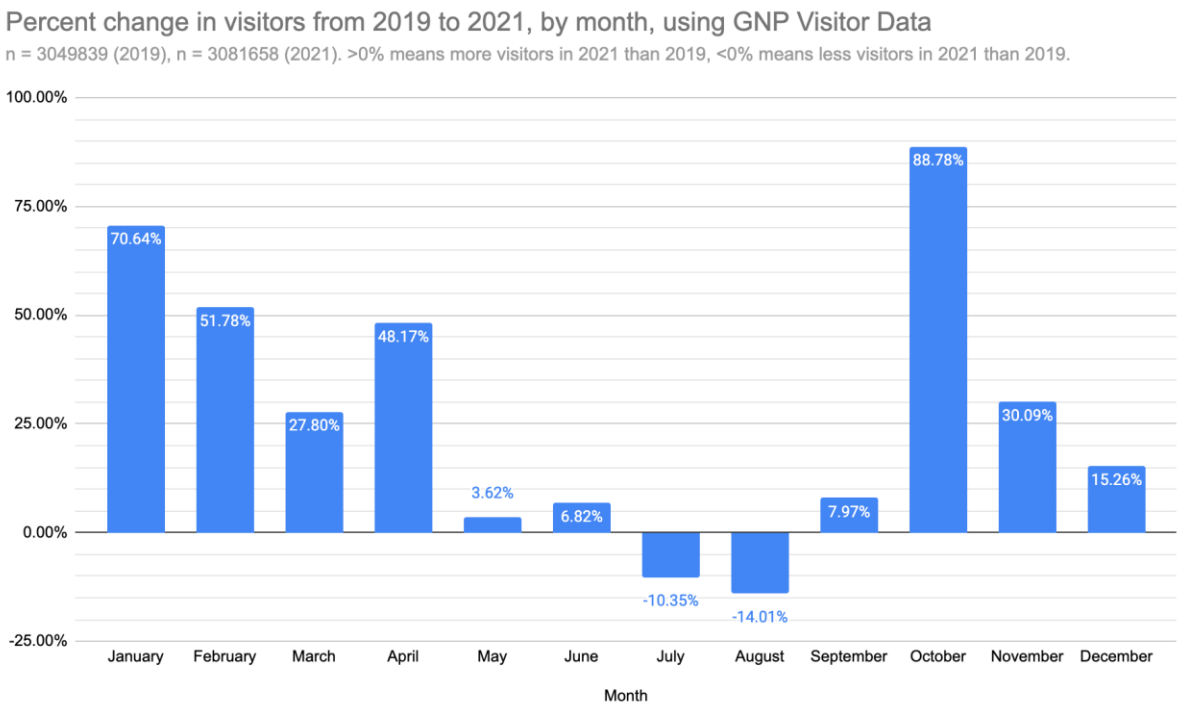


Figure 4.

Percent Change in Visitation by Month From 2019 to 2021 Using IRMA Visitation Statistics



4.2.3 Volume of Zartico's Data

We found that data is not that big inside the Park and this is a problem faced by other big data providers. During summer months (June-October), Zartico logged 2,000 to 13,000 unique devices per month at a ratio of one unique device to about 65-90 visitors. During winter months (November-May), Zartico logged 1,000 unique devices or less per month.

Zartico is not the only big data provider that faces sample size issues when used in the Park. The WPI project group evaluating SLI data ran into a similar problem with sample sizes. The group noted that "Trying to track trips ending in smaller parking lots and turn-off areas on the road did not provide a sufficient volume of data, and SLI rejected smaller zones due to anonymization concerns. This was also the case when we attempted to examine pedestrian data on [GNP] trails" (Cole, et. al, 2021). In Appendix C, we provide detailed graphs on the sample size and device to visitor ratios from Zartico's data.

4.3 Labor-Intensiveness of Zartico's Data

The presumed low labor-intensiveness of collecting data was one reason GNP management was interested in the potential of implementing big data. We evaluated the work required to understand Zartico's data and generate analyses to conclude if the platform required less work than their current data collection methods. **Overall, we found that Zartico's platform was very labor-intensive.**

4.3.1 Expertise Required to Use the Platform

The information on the Zartico platform is not easy to understand at first glance. To get accustomed to the platform, we had multiple hour-long training sessions. These training sessions helped us understand the platform, however, the platform still felt unfamiliar. It took weeks to understand some of the data and labels available.

Because of our confusion, we frequently reached out to ask questions. This increased labor and slowed down progress, since we had to rely on Zartico's team for answers and it sometimes took days to get a reply.

4.3.2 Time Spent

After we got an understanding of their platform, we used it to create graphs and charts. However, finding the correct metrics and attributes took time since there are hundreds of options available. Even after learning to create graphs and charts, when it came to creating new graphics, it did not decrease the time spent. Also, when the graph was eventually created there became issues with its readability.

Zartico's built-in graphs were hard to read due to small font sizes and unclear captions. This meant we had to move the organized data from Zartico's platform into Excel to be able to make the graphics useful. This took a lot of time since we transferred numerous data points, to create graphs only to make small adjustments. Although it takes time to get used to any new platform, the work required to use Zartico's platform is still very laborious.

4.4 Relevance of Zartico's Data to Management Decision-Making in Glacier National Park

The relevance of Zartico's data to GNP management was assessed by asking if the available data aligned with their interests and informational needs. To be relevant, the data must cover the correct location, be granular enough, and be necessary for GNP management's needs. **We found that big data is less relevant inside the Park, rather than in other tourist spaces outside the Park.**

Overall, we found that the available data could not answer or even contribute to two thirds of the questions on our list. This included at least 7 of 12 highest priority questions (Appendix A). The unanswerable high priority questions included many visitor statistics questions like "How many days did visitors spend at GNP?" and "How much time did visitors spend within the Park in a day?". There was not enough relevant data to provide the answers. The irrelevant data was either not granular enough for GNP management's needs or not applicable to the Park.

4.4.1 Coverage of Relevant Location

Some data only applied to tourist spaces outside the Park. For example: Zartico gave us the capability to explore credit card spend data. This data was granular enough with the number of cardholders, cardholder ZIP code, amount spent, and which business the transaction took place at. GNP management was interested in how spending changed before and after their reservation system was implemented (Appendix A, Questions 19 and 20). The spend data was relevant to answer spend questions related to external businesses. However, GNP staff expressed a greater interest in spending questions related to internal businesses (P. Webster, Personal

Communication, September 2022). The spend data did not cover internal businesses, therefore, we found it to not be relevant to GNP.

As seen in Table 5, we filtered the total spend data to merchants located in the 59936 and 59434 ZIP code areas, which each cover a part of the Park but include areas outside the Park as well. None of the businesses listed in the data were inside the Park. We learned from a Zartico representative that the Park reports their spending in a different way than other businesses do (K. Stadius, Personal communication, September 2022). This meant spending inside the Park could not be collected by Zartico or any other big data provider.

Table 5.

Merchants Recorded Within the 59936 and 59434 ZIP Code Areas

FILTERS ABC Merchant Zip: 59936, 59434 DRAG ABC OR HERE

Edited Description ▲	Sum of Total Spend	Sum of Cardholder Count
AVIS RENT-A-CAR	1,741.70	5.00
BUDGET RENT-A-CAR	1,219.88	6.00
Car Rental Companies (Not Listed Below)	1,429.78	1.00
Card Shops, Gift, Novelty, and Souvenir Shops	340.17	26.00
DOLLAR RENT-A-CAR	23,768.20	63.00
HERTZ RENT-A-CAR	678.84	1.00
Lodging, Hotels, Motels, Resorts, Central Reservation Services (not elsewhere classified)	3,695.30	4.00

4.4.2 Granularity

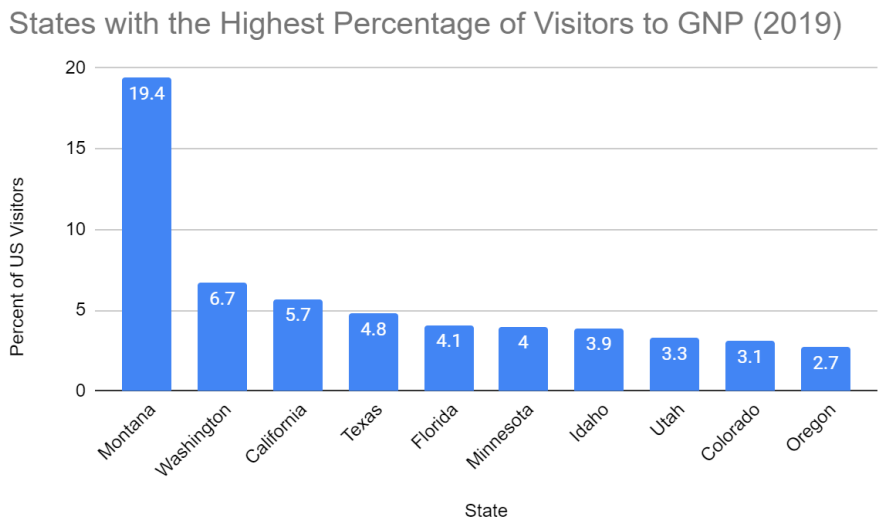
A lack of granularity made some data irrelevant to Park management. For example, management was interested in questions about the ages of visitors (Questions 12, 13, 14, Appendix A). However, the data was not granular enough to do so. The ages were not shown as a list of visitor ages but just the median age of visitors from each state. Having only median age was not relevant to answering any of the questions on the list.

4.4.3 Presence of Data Necessary to Answer GNP's Questions

On the other hand, some data was found to be relevant to GNP management's informational needs. We had access to visitor home location data, which could be separated by country, state, or city area within Zartico's platform. The home location data was relevant because it answered the high-priority question of where visitors were coming from (Question 1, Appendix A). Figure 5 shows the success of this data in answering which states visitors come from.

Figure 5.

Top Ten States Where GNP Visitors were Coming From in 2019



The home location data provided another relevant data type: visitor-to-resident ratio. This was relevant because it contributed to answering the question of whether the reservation system affected the number of locals visiting the Park (Question 11, Appendix A).

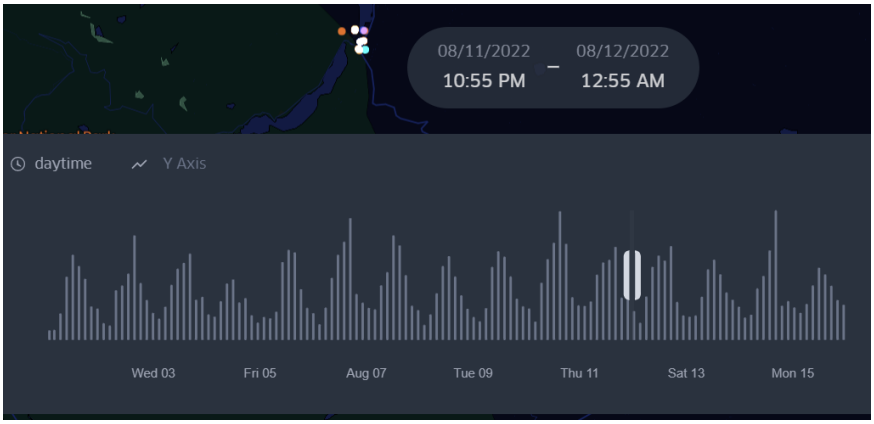
Visitor movement captured in Zartico's dynamic visualization (DV) was granular enough to be relevant and addressed where visitors stay and travel to after visiting the Park (Question 2, 15, 25, 26, Appendix A). Figure 6 shows the appearance of the DV. The dots represent devices, separated by color based on resident, in-state, and out-of-state visitor type. The DV shows visitor devices traveling in GNP and where they went after exiting the Park. The DV can be filtered down to a small window such as two hours, as shown in Figure 7. This tool is useful to "anticipate where, when and how people move and interact" within the Park (Zartico - A, 2022).

Figure 6.

Screenshot From the Zartico Dynamic Visualization of Visitor Movement

Figure 7.

Screenshot of the Timeline for the Dynamic Visualization of Visitor Movement



5.0 Recommendations and Discussion

In the following section, we provide some questions GNP management should ask to evaluate the usefulness of big data before investing in it. Then, we recommend what type of big data GNP management should or should not invest in if they decide to pursue big data.

5.1 Tool for GNP Management to Assess the Usefulness of Big Data

Our group created a list of questions that can be used to assess how well a big data platform meets each of the three criteria: credibility, labor-intensiveness, and relevance. We recommend GNP management use this list as a tool to evaluate whether a big data provider is useful to the Park. The full list is shown in Table 6.

Table 6.

Questions for big data providers

Relevance:

- To what extent does the data cover the Park area?
- To what extent is the data relevant to Park management?
- Is the data separated by time of day?
- How long would the Park have access to the big data platform?

Credibility:

- Are there sufficient points within the data that could be compared with the IRMA Visitor Statistics?
- Does the data span enough previous years to identify yearly trends?
- To what extent can the platform collect data when devices are offline, given the limited connectivity in the Park?
- To what extent does the data contain a bias against certain groups of people? (ex. cell phone provider, app usage)

Labor-intensiveness:

- To what extent will implementing this big data save labor compared to the current data collection and analysis methods?
- To what extent does the big data platform have an ability to produce graphics (graphs, tables)?
- Is there a position in the Park that would already have the skill sets to work with the data?
- To what extent is the big data platform intuitive to use? Will it require a lot of training and/or need constant customer support?
- How much would big data cost?

Additionally, we recommend that Park management ask for a trial of big data providers before purchasing data from them. This would allow Park staff to conduct an in-depth analysis of each question, allowing Park management to make a more informed purchase.

5.2 Where Big Data is Useful for GNP Management

We found that only specific big data is useful to GNP management. **We recommend that if GNP management decides to invest in big data, the data should include visitor location and movement, congestion, and traffic flow. Having the ability to compare big data to the current data available in the Park would also be useful.**

5.2.1 Visitor Location and Movement

We recommend the Park acquire visitor location data if they decide to invest in big data because it is relevant to many of GNP management's needs**.** Since GNP management was interested in how their decisions affect resident visitation, the location data proved to be relevant and credible. Out of the types of data we worked with, location data addressed the most questions in Appendix A.

5.2.2 Traffic Congestion

We recommend GNP management consider traffic and congestion big data. A previous WPI project team used the big data provider SLI. They found traffic data was useful in showing how the reservation system affected the volume of vehicles and redistributed congestion to other parts of the Park (Cote et al., 2021). This type of big data was also granular enough that the team could determine the hours in which vehicles entered the Park. This was useful to compare the distribution of arrival times before and after the reservation system was implemented (Cote et al., 2021).

5.2.3 Comparison Tool

We recommend GNP management consider acquiring big data that can be compared against current data in the Park. Some big data overlaps with data already available in the Park. Comparing the big data to other sources is a good idea to ensure the accuracy of the data. For example, SLI's traffic data could be compared to the traffic counts taken by GNP staff as well as the timing of vehicle entries. Zartico's visitor home location data could be compared to the ZIP code data of people who purchased vehicle reservation tickets. Comparison can solidify the data or help pinpoint areas where either data collection method could be improved.

5.3 The Problems of Big Data for GNP Management

Big data is an emerging technology with caveats and so it may not yet be worth investing in. The Park needs to think critically about whether implementing big data is worth it or not. **We recommend GNP avoid using big data for tracking precise visitor locations and credit card spending data. GNP should use big data to supplement, rather than replace, on-the-ground data collection.**

5.3.1 Big Data is a Sample

We recommend Park management always keep in mind that big data only provides a sample of visitor behavior. In the best-case scenario, Zartico tracked about 2% of visitors during the summertime and this may be the case for other big data providers. Conclusions from big data should not carry the same weight as conclusions from on-the-ground data sources. Rather, GNP could use big data for a birds-eye view and use on-the-ground data sources to confirm any interesting findings from big data.

Small sample sizes in big data are not unique to our project. Li describes that "For sample size, the datasets used in most existing studies were on a relatively small scale... [Small] samples are prone to selection biases and estimation biases, leading to incorrect analysis results that might be opposite to the real situations and the generalized findings" (Li, 2018). We recommend the Park avoid usage of big data outside the months of June-October, due to low sample sizes that we believe do not provide the data accuracy the Park wants.

5.3.2 Lack of Precise Visitor Tracking

We recommend Park management does not try to use big data to track precise visitor behavior. The Park should not replace any traffic counter systems, especially in areas of the Park with lower visitation numbers. As discussed, big data revolves around periodic location pings coming from a sample of devices, making it impossible to precisely track visitor behaviors.

Due to limited cell phone service in the Park, any precise location data gathered from big data may be inaccurate. We know, for example, that the location technology in Apple devices might be "unavailable, inaccurate, or incomplete" (Apple, 2022) in geographic locations that lack data services.

5.3.3 Spending Inside and Outside of Park

We recommend the Park do not purchase any spending data from a big data provider. As we discussed in section *4.4.1* (Coverage of Relevant Location), Zartico did not have any spending data inside the Park. Additionally, Park management already has monthly revenue data from concessioners operating inside the Park. Spending data from a big data provider would be redundant.

Conclusion

Data collection and analysis is vital to inform Glacier National Park management's decisions. The Park wanted to determine if big data could improve its current data collection and analysis methods.

We found that big data can fill gaps in visitor data and could be useful as a secondary data source. Big data can answer questions about general visitor behavior without the need for on-the-ground hardware. Big data can track where visitors are coming from, where they are going in the Park, and how that affects traffic congestion on Park roads.

However, using big data is a labor-intensive process due to the training and data filtering required to use the platforms. Big data falls short when trying to track precise visitor behavior and lacks historical data needed to make meaningful comparisons. Lastly, big data is still an emerging technology that might not be worth the investment.

Bibliography

Apple. About privacy and location services in iOS and iPadOS. Apple Support. (2022, February 1). Retrieved October 3, 2022, from https://support.apple.com/en-us/HT203033

Cash, D., Clark, W., Alcock, F., Dickson, N., Eckley, N., & Jäger, J. (2002, November). Salience, credibility, legitimacy and boundaries: Linking research, Assessment and Decision Making. Digital Access to Scholarship at Harvard. Retrieved September 30, 2022, from https://dash.harvard.edu/bitstream/handle/1/32067415/Salience_credibility.pdf?sequence=4

Cote, B., Showan, C., Murray, C., Bonofiglio, K., & Laber, J. (2021). *Evaluating Glacier National Park's 2021 Ticketed Entry System.* [Interactive Qualify Project, Worcester Polytechnic Institute]. Digital WPI. https://digital.wpi.edu/pdfviewer/79408140t

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities.* Harvard Business Review Press.

Edwards, D., Griffin, T. (2013). Understanding tourists' spatial behaviour: GPS tracking as an aid to sustainable destination management. *Journal of Sustainable Tourism, 21*(4), 580-595.

Ho, C.I. Yuan, Y. (2017). Needs analysis of national parks for applying big data solutions in tourism management. *International Journal of Advanced and Applied Sciences, 4*(12), 26-30.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), 1.

Lee, I. (2017). Big data: dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*(4), 293-303. https://www.sciencedirect.com/science/article/pii/S0007681317300046?casa_token=57sQ4OEitrAAAAAA:Ex941uCtzA8pRq8QzgG8UJFX1xtRt0uyTUD58Ye2neAsULupv2g-iGZVPBwig4bNq_cAxVyxwDU#bib0040

Li, J., Xu, L., Tang, L., Wang, S., Li, L. (2018). Big data in tourism research: a literature review. *Tourism Management, 68,* 301-323. https://www.sciencedirect.com/science/article/pii/S0261517718300591?casa_token=rk8TEAijK9EAAAAA:5GGxkoRLDBxEAnhQOuG5ECx6ekid12nJO7x8sIFeWY_OiEUcxocw_gU8RMJXCEfkvMuD4grl6qo#!

Machlis, G. (1996). Usable knowledge, a plan for furthering social science and the national parks. *The National Parks Service.*

National Research Council, Committee on Improving the Science and Technology Programs of the National Park Service. (1992). *Science and the National Parks.* The National Academy Press.

Simmonds, C., McGivney A., Reilly, P., Maffly, B., Wilkinson T., Canon, G., Wright, M., Whaley, M. (2018, November 20). Crisis in our national parks: How tourists are loving nature to death. The Guardian. Retrieved September 30, 2022, from https://www.theguardian.com/environment/2018/nov/20/national-parks-america-overcrowding-crisis-tourism-visitation-solutions

Stobierski, T. (2019). The advantages of data-driven decision-making. *Harvard Business School.* https://online.hbs.edu/blog/post/data-driven-decision-making

U.S. Department of the Interior. (2021, December 2). Cell and WIFI connectivity. National Parks Service. Retrieved October 6, 2022, from https://www.nps.gov/glac/planyourvisit/connectivity.htm

Xiao, Y., Wang, D., & Fang, J. (2019). Exploring the disparities in park access through mobile phone data: Evidence from Shanghai, China. Landscape and Urban Planning, 181, 80–91.

Zartico. (2022). ZDOS. *Zdos*. http://www.zartico.com/zdos

Zartico - A. (2022). Zartico Dynamic Visualizations. *Zartico.* https://www.zartico.com/zartico-dynamic-visualizations

Appendix A: Questions Demonstrating GNP's Informational Needs

This appendix contains our thirty questions from Appendix A organized by priority (to what extent GNP is interested in the question) and by category. The following color key indicates if the questions can be answered with the available data or not:

(No Color) Yes, the question can be answered with Zartico data, either partially or fully

No, this question cannot be answered with Zartico data.

The questions are also categorized by type of question into the following categories: Economic, Demographic, Visitor Statistics, Reservation System Change, and Congestion.

Highest Priority Questions

1. (Demographic) Where are the visitors coming from? (local, out-of-state, Canada, etc.)
2. (Visitor Stats) What points of interest did visitors frequent after exiting the Park? (other hiking trails close by, fishing spots, etc.)
3. (Visitor Stats) How many days did visitors spend at GNP?
4. (Visitor Stats) How much time did visitors spend within the Park in a day?
5. (Visitor Stats) Are there areas within GNP that a higher % of locals visit compared to other areas?
6. (Visitor Stats) How has foot traffic changed on smaller/less popular trails as the number of visitors increased?
7. (Visitor Stats) What time of day has the peak number of visitors?
8. (Visitor Stats) What percentage of visitors enter outside of the reservation time? (outside of 6am-4pm)
9. (Visitor Stats) What percentage of visitors enter within Managed Access? (6 am to 4 pm in 2022, 6 am to 5 pm in 2021
10. (Reservation System Change) How has visitation timing changed comparing before and after the reservation system? (spreading out into early/late season, shoulder season)
11. (Reservation System Change) Did implementation of the reservation system affect the number of locals visiting the Park?
12. (Reservation System Change) Is there a demographic change in visitors comparing before and after the reservation system?
    a. age, race, instate vs out state

Medium Priority Questions

13. (Demographic) What is the age breakdown of visitors?

14. (Demographic) Are there differences in which activities certain age groups participate in within the Park?

15. (Demographic) How does the visitor's place of origin affect the activities they partake in?

16. (Visitor Stats) How many people were in a party, on average?

17. (Visitor Stats) How long did visitors spend at specific attractions in GNP?

18. (Visitor Stats) How many people were on a specific trail?

19. (Reservation System Change) How did the reservation system change spending at businesses in the communities surrounding GNP?

20. (Reservation System Change) How did the reservation system change the spending at Apgar village businesses?

b. Compare West Glacier Businesses maybe
c. Contact Commercial Services Division

21. (Demographic) Is there a demographic change in visitors between pre- and post-covid?

22. (Congestion) Are there large numbers of people that clump around at shuttle transfer spots?

23. (Congestion) Are there certain Park resources that are used more than others?

24. (Congestion) Are there any ecologically vulnerable areas in GNP that overlap/are close to overcrowded areas?

Low Priority Questions

25. (Economic) Which businesses did visitors frequent after exiting the Park?

26. (Economic) Where did visitors stay during their trip? (hotels, etc.)

27. (Economic) Which specific points of interest do visitors spend the most money on?

28. (Visitor Stats) What times do people go to certain attractions?

29. (Visitor Stats) Does the number of facilities in an area affect the number of visitors? (parking spots, bathrooms, etc.)

30. (Visitor Stats) How long were people on a specific trail?

Appendix B: Coding Analysis of Responses to the Questions

The goal of this coding exercise was to establish some of the criteria that would help us to assess the usefulness of a big data platform. We created seven categories which summarized our analyses of the question list in Appendix A. The categories were further grouped into broad themes. The themes "relevance" and "credibility" became the criteria. "Further analysis" responses were examined further to generate graphics and evaluate the capabilities of Zartico's platform.

	Question #																													
Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Total
Granularity				x		x	x	x	x								x	x				x	x	x				x	x	x	13
Presence of a necessary dataset		x	x	x		x							x	x		x				x	x				x	x				x	12
Presence of a necessary location												x																			1
Validity of Data				x													x														2
Volume of Data												x																			1
Time period for which data is available										x										x									x		3
Further analysis	x				x					x	x				x						x										6

	#Categories	#Q's Total
Relevance	3	23
Credibility	3	5
Further analysis	1	6

Appendix C: Total Unique Devices per month inside GNP, 2019 and 2021

Figure D1.
Total Unique Devices per month inside GNP during 2019, by month.

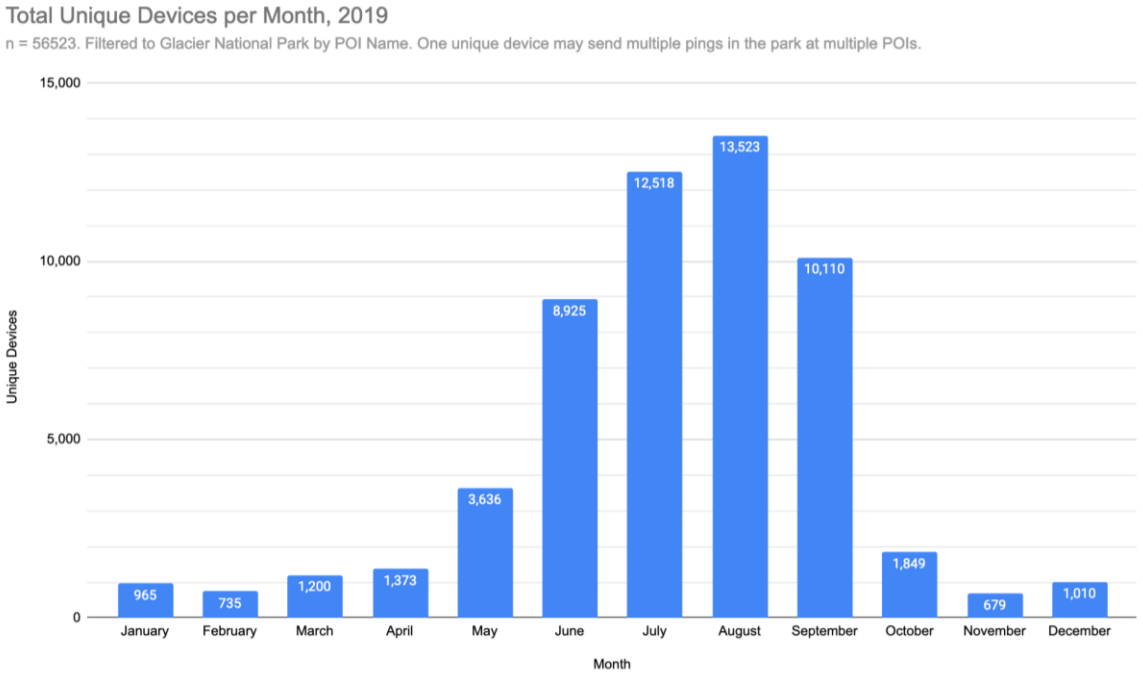


Figure D2.
Total Unique Devices per month inside GNP during 2021, by month.

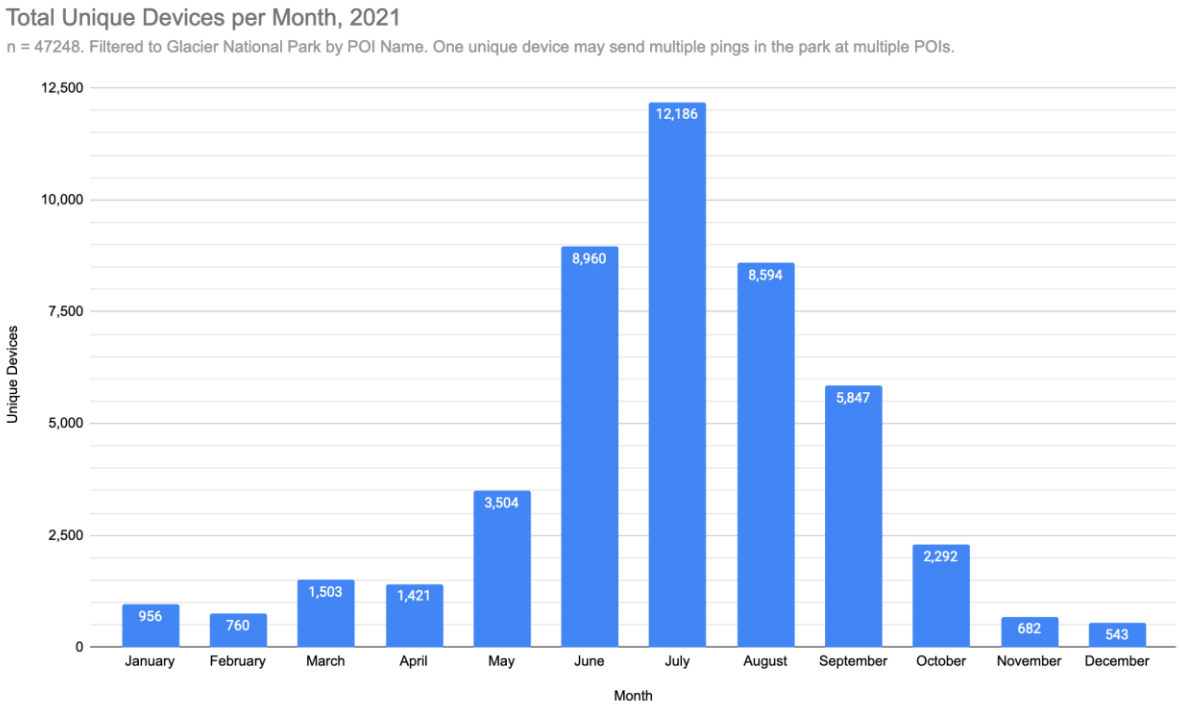


Figure D3.
Visitors to Unique Device Ratio for Zartico data, by month, in 2019.

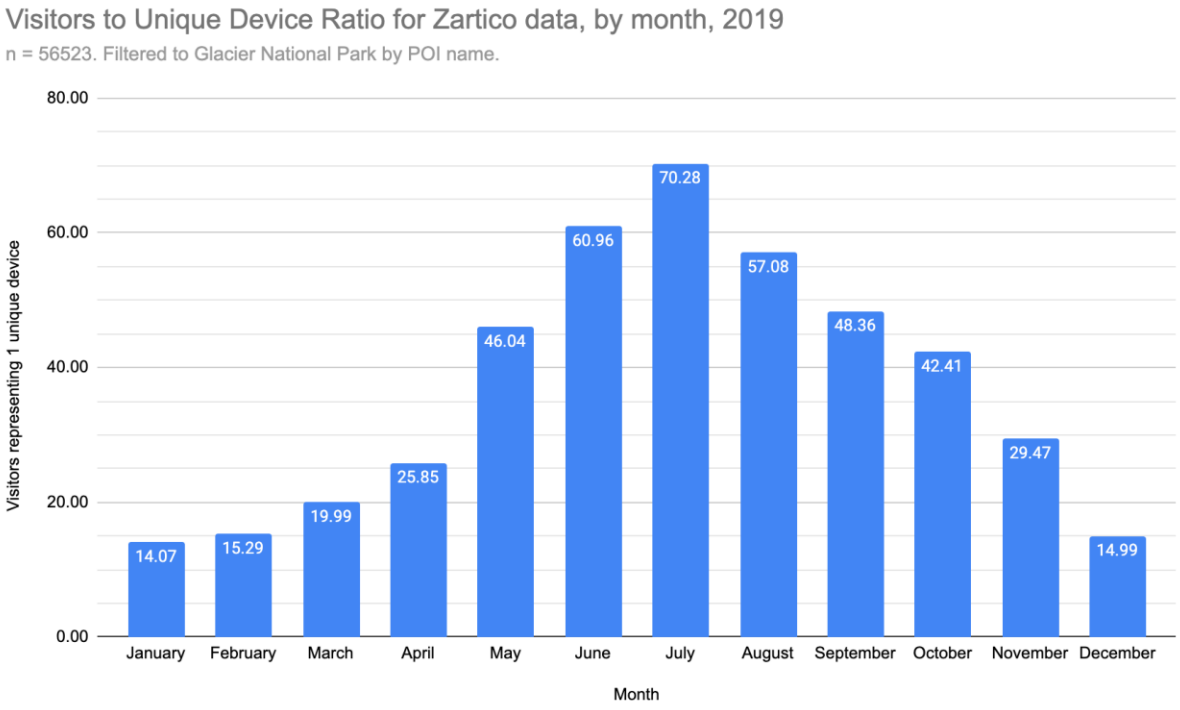


Figure D4.
Visitors to Unique Device Ratio for Zartico data, by month, in 2021.