THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

EXPLORING THE MENTAL REPRESENTATION OF HOW PEOPLE INTERPRET VISUAL ANALYTIC TASKS, BASED ON COGNITIVE FOUNDATION

FANGYUAN WEN
Spring 2019

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Information Science and Technology
with honors in Information Science and Technology

Reviewed and approved* by the following:

Xiaolong Zhang
Associate Professor of Information Sciences and Technology
Thesis Supervisor

David Fusco
Assistant Teaching Professor of Information Sciences and Technology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Background: The technology of visualization is developed at a fast pace as it is highly demanded in this big data era. However, current visual analytic tools largely focus on providing visualization tools to target for specific data types and often overlook the processes and activities involved in visual analytics. Visual analytics often involve high-level cognitive activities, and the results of visual analytics can be directly influenced by the cognition of analysts in data analysis. One of important cognitive factors in visual analytics is cognitive bias, which may mislead the analysis in various ways.

Research Questions: How do cognitive biases affect the results of visual analytics?

Research Method and Result: The research will use a qualitative method to analyze the solutions submitted to the 2017 IEEE VAST Challenge. By comparing some solutions with the ground truth of the challenge, this research identifies the gaps between the correct answers and the submitted answers and from the perspective of cognitive biases explore the causes of these differences. The analysis of this research focuses on two common cognitive biases: anchoring bias and availability bias.

Conclusion: The research confirms the existence of cognitive biases in data analysis supported by visual analytics tools. The results also provide some guidance to mitigate the impacts of cognitive biases in visual analytics. This work can help analysts understand the potential biases they may have in their work, and help the designers of visual analytics systems understand what can be done to improve their systems.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1
Introduction

Visual analytics is a discipline that aims at facilitating analytical reasoning through interactive visualization interfaces. It is used in the problems which requires human and machine participation in analysis. The article *Foundations and Frontiers of Visual Analytics* (Kielman, Thomas, & May, 2009) indicates that "the study of visual analytics promotes technical skills in analytical reasoning, user interaction, data transformations and representations for the purposes of computation and visualization" , etc. Visual analytics is a discipline that integrates such disciplines as computer science, information visualization, cognitive sciences, and social sciences. Visual analytics involves cognitive activities, such as reasoning, sensemaking, and decision-making. It is important to understand the relationships between the characteristics of cognition and visual analytics. The focus of this thesis is cognitive biases in visual analytics.

Cognitive bias is a systematic error from norm in judgement (Haselton, Nettle & Andrews, 2005). The study of cognitive bias starts from psychology, and the existing research has summed up a large number of biases from the area of cognitive science, social psychology, and behavioral economics. Research on these cognitive biases provides good understanding of the systematic errors, from the perspective of perception and judgment, in reasoning processes (Kahneman & Tversky, 1996).

Cognitive bias can lead to wrong judgement and decision. Recent research such as Wall at al. (2017) argued that cognitive biases could lead to negative consequences in visual analytics, such as inaccurate knowledge and wrong solutions in data-driven decision-making processes.

However, research on the analysis of what specific types of biases exist in visual analytics is rare. This thesis aims at analyzing how cognitive biases may affect real-world analytical work and exploring what could be a good way to prevent these biases based on existing theories.

This thesis is organized as follows. Chapter 2 reviews relevant literature on visual analytics, cognitive biases, and cognitive biases in visual analytics. In Chapter 3, the discussion will focus on the concept of heuristics and representativeness heuristics. Chapter 4 and 5 analyze two types of biases: anchoring bias and availability bias. Chapter 6 discusses the results and implications of this research.

Chapter 2

Literature review

This research concerns cognitive biases in visual analytics. Thus, literature review here focuses on basic concepts in visual analytics, cognitive biases, and cognitive biases in visual analytics.

2.1 Visual Analytics: Concepts, Methods and Current Challenges

The computer-based visualization system is the focus of this paper. A computer-based visualization system can help people deal with data and information more effectively through the visual representation of datasets. Visualization usually support tasks that requires human involvement, rather than providing an environment where all decision-making is done by computers. A good visualization system should consider both the creation of appropriate visualization tools and user interaction with them. There are three kinds of resource constraints in the visualization designing area: those concerning machines, human users, and displays. According *Visualization Analysis and Design* (Munzner, 2016), the usage of visualization is "analyzed in terms of why the user needs it, what data is shown, and how the idiom is designed." In this paper, various visual analytic tools will be compared and analyzed in the case study to explore their performances in visual analytical tasks.

Card et al. (1999) offered a good guidance on the development of information visualization research. Early on, research on visualization was about scientific visualization to help scientists understand scientific data and confirm hypothesis. Information visualization research emphasizes the use of visualization tools to understand the relationship among complex data and form hypothesis. Visual analytics, the latest member in the visualization family, considers more heterogenous datasets and aims at helping analysts gain new knowledge and insight from data.

Research on visual analytics takes complex approaches. In addition to technical design, researchers also consider the nature of data, algorithm, and more importantly, the cognitive characteristics of analytical behaviors (Wong & Zhang, 2018). Often, good design is a result of in-depth analysis and solid understanding of user cognitive activities.

2.2 Cognitive biases

Visual analytics often involves decision making. Decision-making includes subjective judgments made by human beings. Mistakes, misunderstand, distrust, and conflict often causes ineffective in decision-making processes. Nowadays, as the psychologists' study in cognitive science, they have found the relationship between human's subjective judgments and heuristics. Heuristics is a kind of judgmental rules in people's daily lives. They are useful but also risky. Good employment of heuristics can reduce the workload of mental tasks; however, in some circumstances, they may lead to unwanted disturbs, which called cognitive biases.

Cognitive bias is a term to describe a human tendency to make wrong judgement. According to Dr. Haselton's psychology handbook *The Evolution of Cognitive Bias*, Cognitive bias is defined as "*a systematic pattern of deviation from norm or rationality in judgment*" (Haselton, Nettle & Murray, 2015). A very common kickback of cognitive bias is making inaccurate judgments in decision-making process. As mentioned before, a heuristic is like a mental shortcut that can speed up decision-making process. Cognitive bias can be regarded as a by-product of heuristics, which makes human reasoning under the risk of lack of appropriate mental mechanisms.

Tversky and Kahneman, the pioneers in cognitive science, has categorized heuristics and biases into many different types (Tversky & Kahneman, 1974). The first one is called availability, a heuristic that is relevant to risk perception. People judge events as occurred frequently, when instances of this event can be recalled easily and quickly. It seems reasonable that frequently occurring events are easy to recall in mind, but in the real world, availability is also affected by factors other than frequency. Representativeness is another common cognitive bias in which similarity in outcomes means one originates from the other. An interesting example of Kahneman and Tversky gave is throwing coins. People usually think the next one is tail after five times of heads. However, the result of each throw is independent and each side has a 50% chance.

This research specifically focuses on explaining how those heuristics and cognitive biases work and present in visualization analytics tasks.

2.3 Cognitive biases presented in visual analytics

In the big data era, visual analytics techniques have been widely studied and used to making sense of big data for business needs, medical needs, and even political needs. This technology can provide highly interactive visual information for human to address the problem of information overload faced by the decision layer. The development of visual analytics includes two parts, human reasoning ability and data processing ability of computer. In the process of visual analysis, human analysts gain insight in complex real-world problems from visual mapping, model-based analysis, and human-machine interactions. Thus, the visual analytic tools play a key role in such environment. However, when the data becomes more complicated, analytics could become increasingly difficult, and the designing of good visual analytic tools is very significant in this stage. In today's world, many existing visual analytic tools are hard to use because there is a huge cognitive gap between the data analysts' mind and the support of the tools in the tasks. One of the difficulties is the cognitive biases people meet during visualization tasks. To understand how the problem performs during visualization process, we can gain a perception on how to improve visual analytic tools in the future tool designing.

After we understand what visual analytics and cognitive bias are, we need to learn how cognitive bias behaves in the field of visual analytics. Wall et al. (2018) describes cognitive bias in four perspectives: Bias as a cognitive processing error, bias as a filter for information, bias as a preconception, and bias as a model mechanism. Early research (Wall et al., 2017) introduced some approaches to minimize the present of cognitive biases by using some metrics.

In this paper, the focus will be on the presentation of cognitive biases in visual analytics tasks. Thus, different kinds of cognitive biases in the following case studies will be shown as well as the different performance and results they produced in the human-machine cooperation tasks.

Chapter 3

Analysis of Heuristics

This chapter discusses heuristics and representation heuristics, two concepts that are essential to the analysis in later chapters.

3.1 Heuristics

Heuristics, as a problem-solving method, is known for its feature of judgment shortcut (Gilovish & Savitsky, 2002). Heuristics helps us make decisions quickly, however, under the risk of cognitive biases.

According to Wall et al. (2018), the cognitive biases can act as a filter of information in human mind. Once people get information overloaded in their decision-making process, they may filter unimportant information and leaves significant ones. This step does help to speed up decision making, but also leads to bad consequence, such as that people actually filter out some key information.

3.2 Representativeness Heuristics

Representativeness heuristic is one of the heuristics that mentioned in Amos Tversky and Daniel Kahneman's *Judgment under Uncertainty: Heuristics and Biases* (1982), in which representativeness heuristic is defined as a way people evaluate the probability that event A originates from event B. If event A is highly representative of event B by degree, A is considered highly possible to originates from B, or vice versa. Representativeness biases occur when event

A has no relationship with event B which is the truth, but people mistakenly regard higher similarity of the A and B as higher possibility that A originates from B (Tversky & Kahnema, 1974).

Chapter 4

Anchoring heuristics

This chapter focuses on the analysis of anchoring heuristics. First, we provide a definition of anchoring heuristics, and then conduct a case analysis of anchoring biases. We also discuss our findings and the implications.

4.1 Definition of anchoring heuristics

Anchoring bias, according to Tversky and Kahneman (1982), involves starting from a readily available number—the "anchor"—and shifting either up or down to reach the final answer. In other words, anchoring heuristic is just an adjustment on an initial value. This initial value can be a guess of the problem answer or any pre-knowledge. By using anchoring heuristics, people can save time on decision making, but as it is easy to think about, different initial values will lead to multiple final answers. The anchoring biases generated while the adjustment people make are insufficient to get the correct final answer, and the wrong answer is called biased toward the anchor. Next, a case analysis will be used to elaborate the anchoring biases in visual analytics.

4.2 Case analysis of anchoring biases

The 2017 IEEE VAST challenge submissions have been taken as cases to identify the biases in visual analytics. In order to control variables, two groups, Singapore Management

University – Shridhar[1], which will be referred as Shridhar solution in the rest of the paper, and Singapore Management University – Yifei, referred as Yifei solution, have been selected for a comparative study. Both of the groups came from Singapore Management University and they used the same visual analytic tools, Tableau and SAS JMP. This control can minimize the difference between those participants in either the pre-knowledge received in college study or the understanding of visual analytic tools. Thus, their way of thinking and results should be close in anticipation. However, one of the groups, the Yifei's group has received an award: Actionable and Detailed Analysis, but the other did not. According to the Ground truth of the 2017 IEEE VAST challenge, the Yifei solution has hit the correct answer in their mini challenge 1, while the other was not. The reason for why one group's results were less actionable and less detailed will be analyzed below with a finding of anchoring bias.

The 2017 VAST challenge has 3 mini challenges, and every mini challenge contains several questions that each of them is a step toward the final answer. The first comparative study is based on the first mini challenge. The background of this challenge is to use given data to identify the possible reasons why the number of Blue Pipit, a bird, is decreasing in Mistford. Though the two groups got their answers slightly different, their reasoning processes determined the different achievements of the two groups.

The first question in mini challenge 1 is stated as the following:

> *Patterns of Life analyses depend on recognizing repeating patterns of activities by individuals or groups. Describe up to six daily patterns of life by vehicles traveling through and within the park. Characterize the patterns by*

---

[1] The answers to the challenge by these two groups were obtained from Visual Analytics Benchmark Repository: https://www.cs.umd.edu/hcil/varepository/benchmarks.php

describing the kinds of vehicles participating, their spatial activities (where do they go?), their temporal activities (when does the pattern happen?), and provide a hypothesis of what the pattern represents.

The Yifei solution first classified visitors as day campers, extended campers, rangers and sightseeing coaches/visitors. Then they found the activity routes as well as activity time followed by hypothesis one by one. Nevertheless, the Shridhar solution did this question a little bit differently. Their findings were randomly selected. Compared with the Yifei solution, the Shridhar solution did not classify visitors, and did not follow a pattern for data analysis, such as data type, spatial activities, and temporal activities. Some of their answers lacked hypothesis and some of their answers were not describing specific kind of vehicles. The Shridhar solution showed a typical trend which does not exist in the other group, that their findings #3 and #6 inferred to cars were in high speed and findings #5 and #6 have a conclusion of "high-speed cars disturb the bird." This trend has been enhanced in the second questions they answered, and also caused biases, as shown in their answer:

Abstract of the findings #3, #5 and #6 of the Shridhar group's question 1

"Finding #3: The speeding patterns over a day illustrate vehicles travel at higher speeds from midnight to early morning.

Finding #5: The diametrically opposite location of the two popular camps shows that there will be frequent movement in both the top-down as well as the right-left (lateral) directions, leading to frequent disturbances for the Pipit.

Finding #6: The car types identified in Finding #2 have smaller total times spent in the preserve, but as a result, they tend to move through paths with higher speeds."

The mini challenge has another question, Question 2:

Patterns of Life analyses may also depend on understanding what patterns appear over longer periods of time (in this case, over multiple days). Describe up to six patterns of life that occur over multiple days (including across the entire data set) by vehicles traveling through and within the park. Characterize the patterns by describing the kinds of vehicles participating, their spatial activities (where do they go?), their temporal activities (when does the pattern happen?), and provide a hypothesis of what the pattern represents.

This question is similar to the first question, and the only difference is that findings are based on long periods. The Yifei solution answered the question in order and comprehension like what they did in the first question. As for the Shridhar group, their answer showed the same trend as they did in former question. Their first finding results in *"passing cars disturbed the bird*," which is exactly anchored on the answer to the previous question. Even this answer is a finding in a long period which is different than short period findings in question 1, they started to show a type of bias called anchoring bias. In the Shridhar solution, their anchor was the result from previous one: *the traffic disturbed the bird and caused a decrease in the number of it*. More evidence can be seen in their following findings. The #2 finding talked about how high traffic disturbed bird in holidays, and the #4 and #6 findings are about the observation of traffic speed.

This trend in an intensified investigation on traffic did not appear in the Yifei solution. The bias toward the wrong anchor made the Shridhar group get a wrong conclusion in the final answer:

Abstract of the findings #2, #4 and #6 of the Shridhar group's question 2

"Finding #2: Peak activity is found in the month of July with as much as 193 cars entering from 2pm to 3 pm during the week of 6-12 July 2015, with the peak occurring on the 11th of July 2015. This might have coincided with the mating season of the blue pipit, and increased traffic might have scared the birds away.

Finding #4: The speed of vehicles begins to increase post the peak season.

Finding #6: Post the peak season, there is a steep drop in the traffic influx into the reserve."

In the final question, "What are the top 3 patterns you discovered that you suspect could be most impactful to bird life in the nature preserve? " The two group has the following answers:

The Yifei solution:

"1) The trespassing service trucks we identified in 12.1. As explained, the movements by the trucks were the most suspicious among all the anomalies discovered. Further investigations are recommended to be conducted to find out what were they transporting in or out of the preserve – whatever it was, it would not be good for the preserve. Even the truck drivers were not hunting the birds directly, they might be transporting certain things that are harmful to the natural environments in the preserve, which in turn would affect the birds' habitats.

2) The long stay duration car ID 20154519024544-322. This car recorded 281 visits to 10 gates, which shows a very suspicious loitering behavior. It could be moving around the preserve searching for something, could be the birds!

3) The movement of service trucks and sightseeing guests/coaches all night long. We observed that unlike the campers, the service trucks and sightseeing coaches move along the connecting paths all night long, going from one side of the preserve to another. The constant traffic at night might be a nuisance to the birds' life, and caused the birds to migrate to other quieter places outside the preserve.

Besides the possible caused identified above, there is another area of concern – the fixed patrolling routine of the rangers. In our path analysis, we noticed that the rangers are working on fixed patrolling routes. It would be relatively easier for people with evil intentions to observe the patrolling routes and schedule, and plan their activities accordingly to avoid being caught. This might be a potential threat to the natural life in the preserve."

The Shridhar solution:

"1) The contrasting patters of traffic and speed across the observation period illustrate that when the roads are clogged in the holiday season, the traffic is affecting birdlife directly. When the traffic eases, the speeds by all vehicle types begin to increase, which causes the natural rhythm of the bird habitat to be

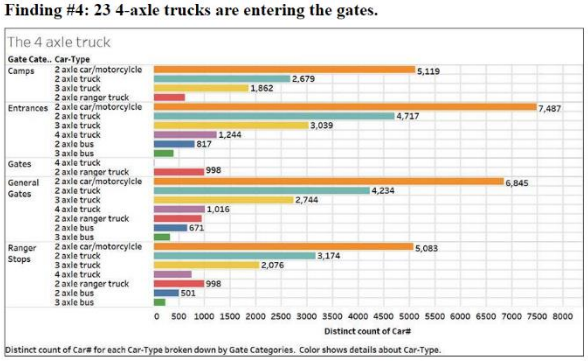
disturbed. The two-fold effect of traffic, whether it is high or low, is thereby suspected to have an impact.

2) *From the traffic patterns on corridors, it can be inferred that the middle corridors between General gate-5- general-gate 2 and the top belt from general gate 2-Ranger stop 0, Ranger stop 2, General gate 1 are being used the most by vehicles and that the diametrically opposite camps- camping 8 and camping 5 are being thronged by people almost all throughout the 13-month period. What this means is that even if the birds decide to migrate within the preserve, they might not be spared as there is constant footfall on either side of the preserve.*

3) *The persistent traffic in gates, with their fair share of heavy equipment which might presumably be used for road construction works and inconsistent patters such as ranger vehicles being found at night, 4 axle trucks moving through gates present patters that are not consistent. There is a case that the security by the patrolling is either weak, where they allow 4 axle trucks to go into the gates. This is because according to the park rules, no other vehicle is allowed access. In addition, repeat visitors who are familiar with the camp such as the vehicle ID 20154519024544-322 might now begin to trespass to new areas, since they become familiar with the terrain every time they enter and hop across camps, etc. A stricter watch on these kind of anomalous behaviours should hopefully save the rose crested Blue Pipits and allow people coming to the preserve to enjoy nature's pristine beauty!"*

The dumping operation in the Preserve by a truck in a restricted area is the ground truth, and the first point of the Yifei's solution exactly matches it. Their reasoning and evidences were well explained in their question 3. Although the Shridhar solution also mentioned the trespassing truck in the Preserve both in their finding #4 of question 3 and the third point of their final answer, as shown in their answer above, this group mentioned the record of trespassing trucks just in order to state the opinion that the security by the patrolling is weak. Besides, in their finding #4 of question 3, they only presented the record as Figure 1 below. They have the 23 times of records of trespassing, but still lack an actionable and detailed analysis on those data.

Finding #4: 23 4-axle trucks are entering the gates.



Intrusion by 4 axle trucks, which are considered general traffic is not allowed. So, this is a potential anomaly arising out of the movement of the **4 axle trucks in restricted areas!**

Figure 1 Finding #4 of question 3 of the Shridhar group

On the contrast, the Yifei solution is more detailed. They presented the paths of those trespassing trucks, and found that it may be a planned act as the trespassing paths are exactly the same. Because of all of the records took place after midnight, they assumed that the trucks were doing illegal things. Because 4-axle trucks were heavy trucks which were used to transport materials, this group believed those trucks might be transporting illegal materials which could harm the birds. Their reasoning and evidence were exactly the ground truth, leading them to win the award of “actionable and detailed” according to the judge.

b. 4 vehicle entered from entrance 3 trespassing restricted areas (gate5,gate6,rangerstop6,gate3 & ranger stop 3) 23 times from 2am to 5 am, only observed on Tuesdays and Thursdays. The trespassing cars followed almost exactly the same paths. This looks like some planned acts which were only performed under the masks of the dark night. Type 4 vehicles are the heavy trucks; they could be transporting some illegal materials in or out of the preserve repeatedly.

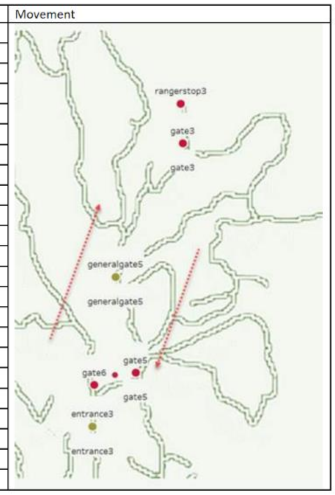
Car ID	Movement
20150104020118-228	
20150416040441-902	
20150505020522-625	
20150920030917-854	
20151112031119-409	
20151201031245-77	
20151414041406-386	
20151415031450-923	
20151520021556-881	
20151521021518-235	
20152824032830-251	
20152925022919-735	
20153923043910-954	
20154702044723-914	
20154901044910-777	
20154907044911-419	
20155201025245-696	
20162219032229-226	
20162401032410-101	
20162419042411-322	
20163016033037-38	
20164531024545-131	
20165003035005-470	

Figure 2 Evidences from the Yifei group

The failure of the Shridhar group can be regarded as the result of cognitive biases. As we can see from the analysis above, the Shridhar solution was heavily anchoring to one of their conclusion "the traffic disturbed the bird". The conclusion was supported “strongly” by various evidence (almost 7 or more findings were provided to support this argument), but this unbalanced evidence collection definitely made their answer weak in other two conclusions. The evidence for "long period stay" were only finding #5 in question 1 and finding #1 in question 2 (see as below), and the evidence for "trespassing" was only finding #4 in question 3 (see as below). That will be a reason why the judge thinks their group is not actionable and detailed enough.

Abstract of findings of Shridhar group

"Finding #1 in question 2: There is constant traffic at the gates as compared to the other zones across the 13 months of observation.

Finding #4 in question 3: 23 4-axle trucks are entering the gates. Intrusion by 4 axle trucks, which are considered general traffic is not allowed. So, this is a potential anomaly arising out of the movement of the 4 axle trucks in restricted areas!"

How the Shridhar group got bias is hard to track. One possible explanation would be they found the traffic evidence at the very beginning of the data analysis work. They thought it could be a good answer for the final question in this mini challenge; thus, they tended to find more of the same evidences subconsciously. This type of bias can be controlled with the observation on the Yifei group. Their group showed almost no anchoring bias because they did data classification first and all their following findings were put in an order. The artificial template they made for themselves helped prevent the anchoring bias. Once people do not select information randomly, their biases could be controlled.

4.3 Discussion in anchoring heuristics

The failure of the Shridhar solution is certainly, not just the fault of human beings. Switch to the visual analytic tool part, the tool they used, Tableau, does not provide a warning or alert for its user to notice their anchor at data searching process. Tableau sure doesn't have this function and together with the cognitive bias, the Shridhar group thus got a less perfect result in the end. For the future designing of the visual analytic tool, like Tableau and alike, specific features like cognitive bias warning could be a good improvement. For example, the software will return a warning after each sorting of data to ask the users whether they ignored group A and group B data since they have been searching a lot in one group of specific data, group C data. The software could also equip with a more advanced feature that can ask what kind of data the user are interested in before starting a job, so that the machine can pre-process data by itself instead of letting users look through data by themselves. Doing so could potentially decrease the chance of getting into an anchoring bias.

Chapter 5

Availability heuristics

For the study of availability bias, the same case of anchoring bias has been chosen, two groups' visual analytical results in 2017 VAST challenge mini challenge 1.

5.1 Definition of Availability heuristics

Availability, as one of the powerful heuristic method, helps people judge the frequency and probability of events by whether some related examples can be recalled easily. Thus, availability heuristic is sometimes called the mental shortcuts. When a person sees several theft alerts in her quarter, she will think the theft rate is higher in this place so she will be more aware of her private property. This is an example of how availability heuristic helps people make decisions. However, according to Tversky and Kahneman (1974), cognitive biases usually come together with heuristics. The availability bias is generated by people's reliance on availability heuristics. In other words, when people recall similar events in their memory, they tend to give greater credence to the recalled memory and thus may be biased to the true situation. Availability helps people make decision quickly, but thoughtlessly. In the following paragraphs, a case study will be analyzed to elaborate how this type of cognitive bias is represented in visual analytics.

5.1.1 Definition of Retrievability of the instances

Tversky and Kahneman (1982) stated several different causes of availability bias. One of them is called *Retrievability of the instances*. It can be simply understood as the easiness of getting access to a certain information. If a person can get to a certain information more easily, she will believe that piece of information appears more frequently than others, even if it does not in reality. This is called the availability bias due to the retrievability of the instances.

5.1.2 Retrievability of the instances case analysis

In the case study, the Yifei solution won the award: Actionable and Detailed Analysis, which means the group covered some details and actionable findings in their work. The Shridhar solution did not receive any award. From sub-question 1, the details of which have been mentioned above, the finding #1 result of the Shridhar solution is "*No movement inside camps and gates at silent hours*" (Figure 3). Their finding #1 for sub-question 2 is that there is constant traffic at gates in long time period (Figure 4). In addition, the finding #4 in sub-question 3 is that 23 4-axle trucks enter the gate (Figure 5). The repeated traffic observation on the gates shows a trend that this group typically payed more attention on the data of car movement at the gate and they may not be aware of that. As a result, the solution did not show any trends while presenting information because they categorized their findings in car types, and all their answers are rigorously formatted. For example, in sub-question 1, they first stated when and where this type of car moved, followed by what was the most popular place for those cars. We can link the behavior of the Shridhar solution to availability bias due to the *Retrievability of the instances*.

The bias started when they answered the second sub-question. The difference between sub-question 1 and sub-question 2 was only the time period. The first question was asking about daily pattern while the second question was about the long-term pattern. It is highly possible the Shridhar group traced back to their answers in question 1, while finding the answer for question 2. Because of the similarity between the two questions, this group very likely recalled the findings in question 1 as a reference. In terms of retrievability, the memory of findings in question 1 is a type of similar recent example here, and the analysts may subconsciously count this example as important while working on answering question 2. In addition, the questions asking for 6 findings from the given data, and from their work we can know that the Shridhar group selected data and findings randomly from the observed data, unlike the other group that selected its data by category, thereby proving well organized answers.

While using Tableau to visualize the data, this group typically looked at the information that related to the category of car movement at gates. This can also be explained that the word “gates” was easily recalled in their mind (because the high occurrence number of the idea “gates”) than any other longer or complex ideas like “Ranger Stop” or “Camps”. Thus, their findings were unconsciously presenting more information on the findings that related to the gates. This availability bias, however, shows no harm up to here. In the following explanations, more of it will be explored.

Finding #1: No movement inside camps and gates at Silent Hours?

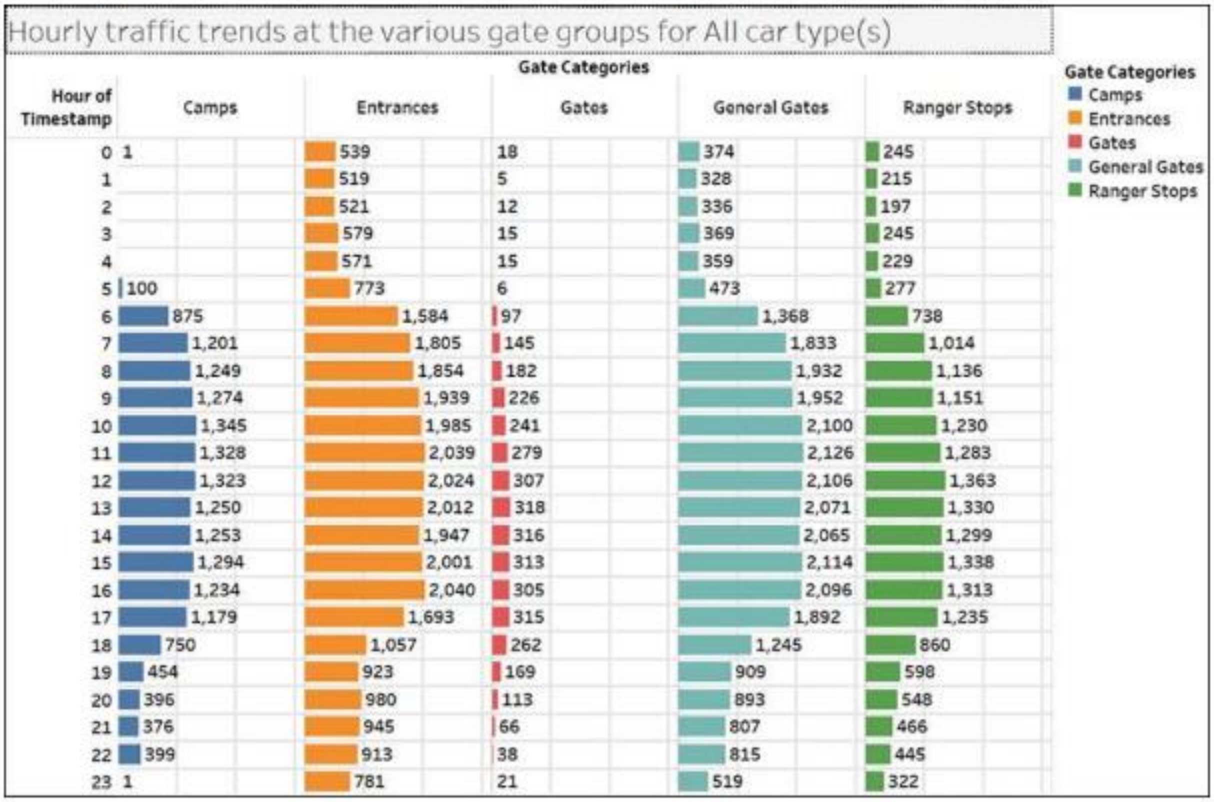


Figure 3 The finding #1 of question 1 of Shridhar's group

Finding #1: There is constant traffic at the gates as compared to the other zones across the 13 months of observation.

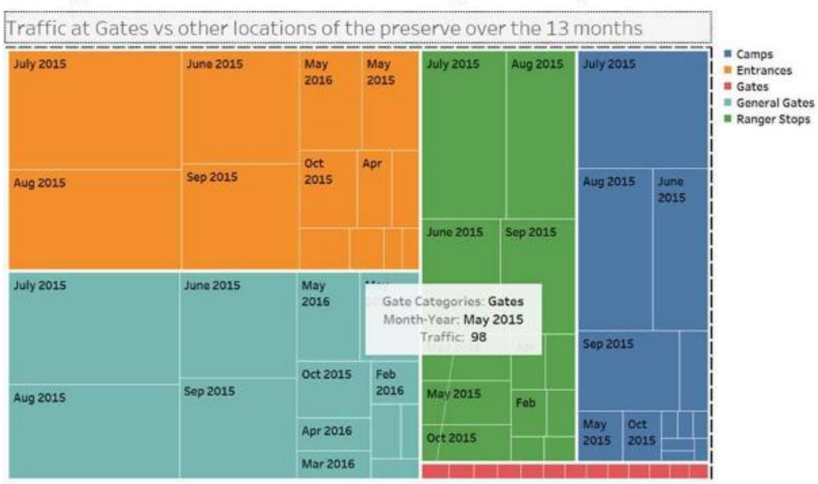


Figure 4 The finding #1 of question 2 of Shridhar's group

Finding #4: 23 4-axle trucks are entering the gates.

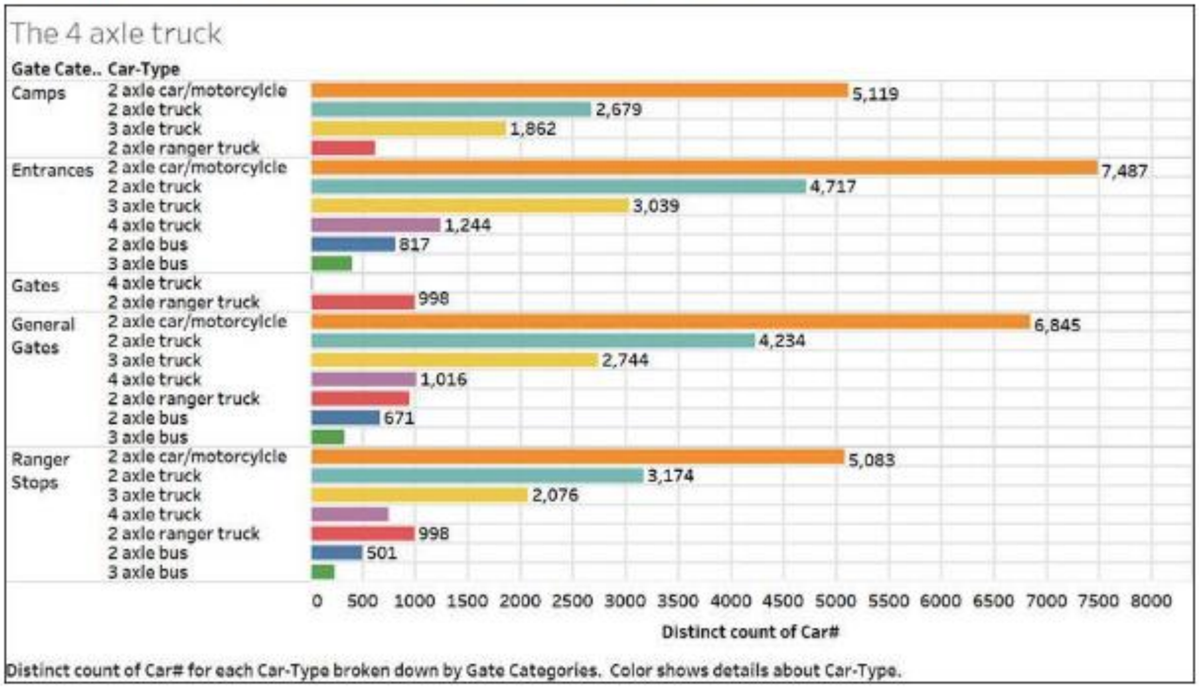


Figure 5 The finding #4 of question 3 of Shridhar's group

5.1.3 Definition of Illusory correlation

According to Tversky and Kahneman (1974), there is a concept called *Illusory correlation*. In a simple sentence, illusory correlation is just a situation in which people think two events have a certain relationship, even if it does not exist in the real world. This concept is based on the study of frequency of paired items. People used the same strategies to evaluate the frequency of a single item as the frequency of paired items. Tversky and Kahneman indicate that in human mind, the repetition of a pair enhances the association among its members. If a paired item is frequently repeated in a person's recent experience, the person will possibly think there is a strong correlation between the two items. In the case analysis below, the "paired item" in the analysts' mind is the "car movement at gates" and "effect on blue pipit".

5.1.4 Illusory correlation case analysis

Comparing the two group's answers for question 4, the final conclusion of what findings are affecting the blue pipit. Here are two group's answers:

The Yifei solution:

"1) The *trespassing service trucks we identified in 12.1. As explained, the movements by the trucks were the most suspicious among all the anomalies discovered. Further investigations are recommended to be conducted to find out what were they transporting in or out of the preserve – whatever it was, it would not be good for the preserve. Even the truck drivers were not hunting the birds directly, they might be transporting certain things that are harmful to the natural environments in the preserve, which in turn would affect the birds' habitats.*

2) The long stay duration car ID 20154519024544-322. This car recorded 281 visits to 10 gates, which shows *a very suspicious loitering behavior. It could be moving around the preserve searching for something, could be the birds!*

3) The movement of service trucks and sightseeing guests/coaches all night long. We observed that unlike the campers, the service trucks and sightseeing coaches move along the connecting paths all night long, going from one side of the preserve to another. The constant traffic at night might be a nuisance to the birds' life, and caused the birds to migrate to other quieter places outside the preserve.

Besides the possible caused identified above, there is another area of concern – the fixed patrolling routine of the rangers. In our path analysis, we noticed that the rangers are working on fixed patrolling routes. It would be relatively easier for people with evil intentions to observe the patrolling routes and schedule, and plan their activities accordingly to avoid being caught. This might be a potential threat to the natural life in the preserve."

The Shridhar solution:

"1) The contrasting patters of traffic and speed across the observation period illustrate that when the roads are clogged in the holiday season, the traffic is affecting birdlife directly. When the traffic eases, the speeds by all vehicle types begin to increase, which causes the natural rhythm of the bird habitat to be disturbed. The two-fold effect of traffic, whether it is high or low, is thereby suspected to have an impact.

2) From the traffic patterns on corridors, it can be inferred that the middle corridors between General gate-5- general-gate 2 and the top belt from general gate 2-Ranger stop 0, Ranger stop 2, General gate 1 are being used the most by vehicles and that the diametrically opposite camps- *camping 8 and camping 5 are being thronged by people almost all throughout the 13-month period. What this means is that even if the birds decide to migrate within the preserve, they might not be spared as there is constant footfall on either side of the preserve.*

3) The *persistent traffic in gates, with their fair share of heavy equipment which might presumably be used for road construction works and inconsistent patters such as ranger vehicles being found at night, 4 axle trucks moving through gates present patters that are not consistent. There is a case that the security by the patrolling is either weak, where they allow 4 axle trucks to go into the gates. This is because according to the park rules, no other vehicle is allowed access. In addition, repeat visitors who are familiar with the camp such as the vehicle ID 20154519024544-322 might now begin to trespass to new areas, since they become familiar with the terrain every time they enter and hop across camps, etc. A stricter watch on these kind of anomalous behaviours should hopefully save the rose crested Blue Pipits and allow people coming to the preserve to enjoy nature's pristine beauty!*

Some details in their answer show the difference between the thinking of two groups. In the Yifei group, the three short paragraphs are used to state the points of view respectively, and each of it has around two pieces of information as evidence in the previous findings. The Shridhar group also has three points, but the evidences of the points are not evenly listed in the previous explanation in their work. Like what has been mentioned in Chapter 4, they have around seven related evidences for the "traffic" point, two pieces of related information for the "long turn crowd" and only one piece of evidence for the last point, trespassing. Actually, their last point shows another presentation of availability bias. Their last point is based on the findings of the car movement at gate. They thought the heavy traffic at the gate is the most important information. The trespassing record is an evidence of weak security. However, the ground truth is that the trespassing trucks are dumping harmful substance. This groups overvalued on the traffic at gate, which finally causes their ignorance of the key information. They did find the evidence of the trespassing trucks, but they did not explore deeply on this finding.

Illusory correlation can be caused through salient data like what the Shridhar group had found in the car movement events at gates. The data is very fluctuating, which presses in the human mind and makes them think it is important than other evidences falsely. Looking back to the Yifei group, they have no related information of "gates" appears in their final conclusion, even some of their previous findings do mentions "gates" related data. This means, for the Yifei group, they considered the word "gates" as a non-related idea in this mini challenge. Their judgement is not biased.

5.1.5 Illusory correlation case analysis 2

In the following contrastive analysis, the results of the two groups above and another group, the Eindhoven University of Technology Group[2] will be used (the solution will be referred as the Eindhoven solution). This group got the award of Elegant Support for Hypothesis Generation and Testing. How the visual analytic tools used by the group helped them win the prize, and why the group won this award, compared to the two groups above, will be explored in the lenses of cognitive bias.

The Eindhoven solution is based a different visual analytic tool, EventPad, which is developed by their group member. The Eventpad groups events by their attribute of interest, and shows those events in glyph sequence (Figure 6).

[2] The solution by the Eindhoven University of Technology Group was also obtained from the Visual Analytics Benchmark Repository: https://www.cs.umd.edu/hcil/varepository/benchmarks.php

Figure 6 Eventpad

The interest here is to understand why the Eventpad solution won the prize of Elegant Support for Hypothesis Generation and Testing, in comparison with the Shridhar solution. In sub-question 1, the Eindhoven solution categorized their findings into 4 patterns, “General patterns”, “Camping patterns”, ”Day trips and traffic speed”, and “Ranger shifts”. Due to the feature of Eventpad that can find and group events by attribute of interest easily, this group listed more information than the Shridhar solution. For example, below are their descriptions on data of ranger vehicles. The Shridhar group only gave time schedule of the ranger vehicles, while the other group provided not only time schedules but also the frequency of the shifts and the accessed ranger stops and the properties of the longest shift.

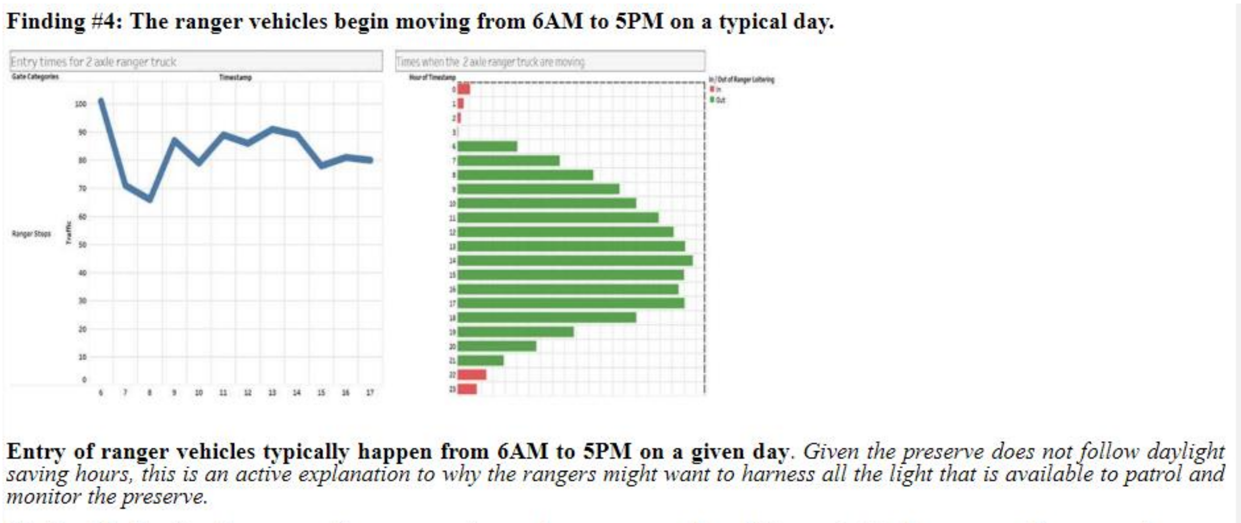


Figure 7 The shridhar group

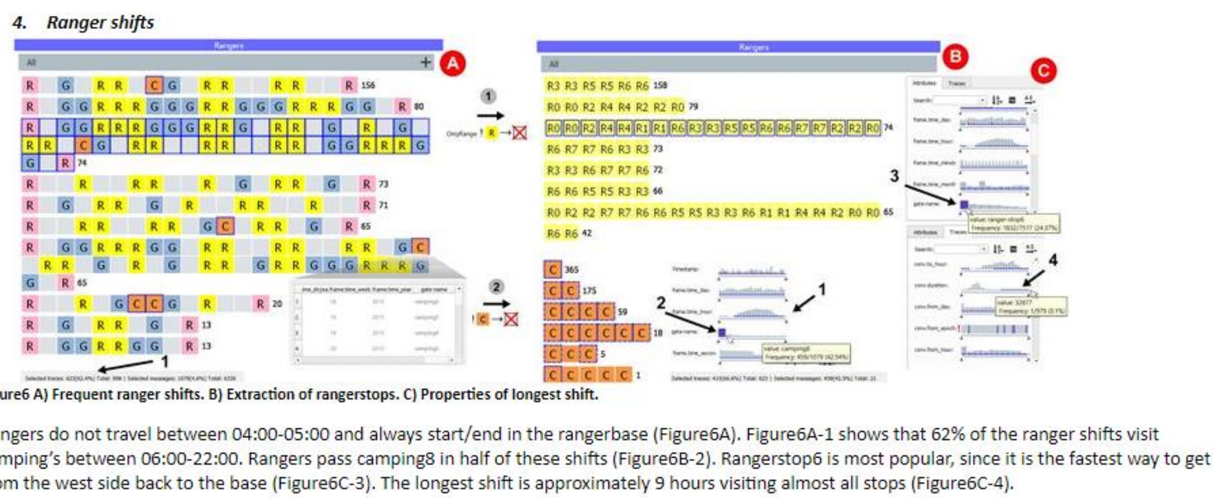


Figure 8 The Eindhoven University of Technology group

Although both the Eindhoven solution and the Yifei solution won awards in this challenge, the final answer given by the former in mini challenge 1 is not as good as that by the latter. One flaw of the Eindhoven solution is the lack of the exact ground truth in question 4, though they had enough evidence to support and infer to the ground truth in their question 3. Compare Figure 9 below and the findings of the Yifei solution in question 3, we can easily see that they almost find the same collection of abnormal data. What this group missed is to make a reasonable reasoning according to the abnormal information they found.

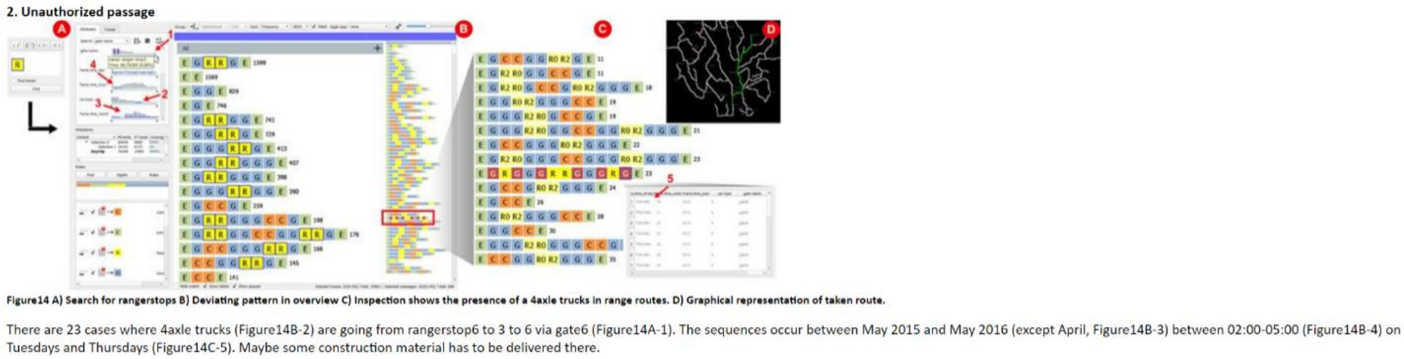


Figure 9 Finding of the Eindhoven University of Technology Group

"There are 23 cases where 4axle trucks (Figure14B-2) are going from rangerstop6 to 3 to 6 via gate6 (Figure14A-1). The sequences occur between May 2015 and May 2016 (except April, Figure14B-3) between 02:00-05:00 (Figure14B-4) on Tuesdays and Thursdays (Figure14C-5). Maybe some construction material has to be delivered there."

Part of the reason that the Eindhoven solution failed to get the award of "Actionable and Detailed Analysis" can be attributed to cognitive biases. Looking at their results below, we can find that they got a very brief answer compared with the above two groups.

Result of the Eindhoven University of Technology Group

"4 — What are the top 3 patterns you discovered that you suspect could be most impactful to bird life in the nature preserve? (Short text answer)

1. Vehicles on the road between entrance 0 and 3 drive too fast. The noise can disturb the wildlife.

2. The repeated access of vehicles to unauthorized locations (in the middle of the night) and the presence of systematic travel activity across the entire reservoir (e.g., tourist busses) during high-season can prevent wildlife from establishing a proper breeding place.

3. The continuous nightly activity of vehicles such busses and trucks over the entire year can disturb the wildlife."

Once they made analysis on their #2 conclusion, they had a very high chance to hit the ground truth. In terms of cognitive biases, this is called the "biases as the filter of information" (Wall et al., 2018). Once information get overload in human brain, they tend to filter it and only left what they thought is important. In this situation, the Eindhoven team may have too much detailed information and possible answers in their former findings. Thus, they omitted the real key information. In Figure 9 above, we can find that they put the "trespassing truck event" and the "traveling buses at the reservoir" together and got a result of "prevent bird from establishing breeding place." The truth is the trespassing truck dumping harmful material and it has no relationship with the travels at the reservoir events. According to the definition provided above, this case can be regard as a type of illusory correlation. The Eindhoven solution mistakenly treated the two event as being related and got a wrong conclusion.

5.2 Discussion in availability heuristics

There are also similarities shared between bias to bias. For instance, the anchoring bias mentioned in last chapter could also be used to explain availability bias in this chapter. The repeating information related to "traffic disturbed the birds" that is the evidence of anchoring, and meanwhile, it is also an evidence of availability. Once the idea of traffic has become an important piece of information in those analysts' minds, people may recall it more frequently during their work and think it is a more important information than others.

How to avoid biases while using availability as a problem-solving method is a valuable question in this process. By understanding that availability biases happen when people give more credence to the frequent recalled solution or evidence while solving a problem or making a decision, we can try several strategies to avoid availability bias. One of them is to add one more step to our thinking process. We can ask ourselves, *did I get a wrong answer because missing something?* By questioning ourselves, we will have a high probability of finding that they have just made a wrong decision. We will realize that we may think too sloppy. After that, it is highly possible that we start making deliberation choice that will finally bring them to the real correct decisions.

In addition, designers should also consider tools to help people deal availability bias. For instance, if the visual analytic tool the analysts have has the capability to return messages to users to notice them the similarity of results between different problems, then users may be able to realize that there might be something wrong in their thinking processes.

Chapter 6
Discussion and future work

Based on the research results of Tversky and Kahneman (1974, 1982, 2000) in cognitive science, a serious of questions have been raised in this research: *Do cognitive biases exist in the visualization area? If so, how the biases affect the visual analysis work?* This research used 2017 VAST challenge results as the research object, and showed the existence of cognitive biases in visual analysis and the consequences of such biases on analytical results.

This research can benefit analysts and visual analytics designers in various ways. On one hand, analysts can benefit from our research by understanding the presences and characteristic of cognitive biases in visual analytics and take precautions in their work. Our results identified some pitfalls that analyst may fall in when they conduct visual analysis. Being consciously aware of the danger of cognitive biases, they may be able to recognize the potential traps and take actions to reduce their damages. For example, knowing the existence of anchoring bias, analysts can go beyond their comfort zones by having a hypothesis very different from what they usually have and see how results may differ.

The results of this research can also benefit the designers of visual analytics systems. Visualization technology is still in development, and there is a huge gap between the existing visual analytic tools and user cognitive activities. This gap makes the visual analytic tools hard to learn and use in many situations. The cognitive bias is one of the problems. Through this thesis, the designers can learn more about what tools people may need to overcome cognitive biases and how to design them in the future. For instance, the method mentioned in Chapter 4 introduces a solution to warning users about potential anchoring bias. After gaining better designing idea and correct orientation, designers can work better to serve the users, and make a greater contribution to visual analytics.

This research could be extended in several directions. First, more efforts can be made to identify other cognitive biases and their specific impacts on visual analytics. Another direction is to develop design guidelines to address the cognitive biases discussed in this research and to use these guideline for system development and user evaluation.

BIBLIOGRAPHY

Card, S., Mackinlay, J., & Shneiderman, B. (1999). Readings in Information Visualization: Using Vision To Think.Gilovich, T., & Savitsky, K. (n.d.). Like Goes with Like: The Role of Representativeness in Erroneous and Pseudo-Scientific Beliefs. Heuristics and Biases, 617-624. doi:10.1017/cbo9780511808098.036

Kielman, J., Thomas, J., & May, R. (2009). Foundations and Frontiers in Visual Analytics. Information Visualization, 8(4), 239-246. doi:10.1057/ivs.2009.25

Haselton, M. G.; Nettle, D. & Andrews, P. W. (2005). The evolution of cognitive bias. In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology: Hoboken, NJ, US: John Wiley & Sons Inc. pp. 724–746.

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The Evolution of Cognitive Bias. The Handbook of Evolutionary Psychology, 1-20. doi:10.1002/9781119125563.evpsych241

Tversky, A. & Kahneman, D. (1974). "Judgement under uncertainty: Heuristics and biases". Science. 185 (4157): 1124–1131. Bibcode:1974Sci...185.1124T. doi:10.1126/science.185.4157.1124. PMID 17835457.

Kahneman, D.; Tversky, A. (1972). "Subjective probability: A judgment of representativeness". Cognitive Psychology. 3 (3): 430–454. doi:10.1016/0010-0285(72)90016-3.

Kahneman, D. & Tversky, A. (1996). "On the reality of cognitive illusions". Psychological Review. 103 (3): 582–591. CiteSeerX 10.1.1.174.5117. doi:10.1037/0033-295X.103.3.582. PMID 8759048.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases (1st ed.). Cambridge University Press.

Tversky, A. & Kahneman, D. (1983). "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement". Psychological Review. 90 (4): 293–315. doi:10.1037/0033-295X.90.4.293.

Tversky, Amos; Daniel Kahneman (September 27, 1974). "Judgment under Uncertainty: Heuristics and Biases". Science. 185 (4157): 1124–1131. Bibcode:1974Sci...185.1124T. doi:10.1126/science.185.4157.1124. PMID 17835457.

Wall, E., Blaha, L. M., Franklin, L., & Endert, A. (2017). Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). doi:10.1109/vast.2017.8585669

Wall, E., Blaha, L. M., Paul, C. L., Cook, K., & Endert, A. (2018). Four Perspectives on Human Bias in Visual Analytics. Cognitive Biases in Visualizations, 29-42. doi:10.1007/978-3-319-95831-6_3

Wong, J., and Zhang X. (2018) "MessageLens: A Visual Analytics System to Support Multifaceted Exploration of MOOC Forum Discussions," In Proceedings of PacificVAST Conference 2018 (Kobe, Japan, April 10-13, 2018).

Tversky, A., & Kahnema, D. (1971). Availability: A heuristic for judging frequency and probability. PsycEXTRA Dataset. doi:10.1037/e301722005-001

Tversky, A., & Kahneman, D. (2000). Loss Aversion in Riskless Choice: A Reference-Dependent Model. Choices, Values, and Frames, 143-158. doi:10.1017/cbo9780511803475.008

Munzner, T. (2016). Keynote speaker: Visualization analysis and design. 2016 IEEE Pacific Visualization Symposium (PacificVis). doi:10.1109/pacificvis.2016.7465242

Wall, E., Blaha, L. M., Paul, C. L., Cook, K., & Endert, A. (2018). Four Perspectives on Human Bias in Visual Analytics. Cognitive Biases in Visualizations, 29-42. doi:10.1007/978-3-319-95831-6_3

Wall, E., Blaha, L. M., Franklin, L., & Endert, A. (2017). Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). doi:10.1109/vast.2017.8585669

Gilovich, T., & Savitsky, K. (n.d.). Like Goes with Like: The Role of Representativeness in Erroneous and Pseudo-Scientific Beliefs. Heuristics and Biases, 617-624. doi:10.1017/cbo9780511808098.036

Academic Vita of Fangyuan Wen

fzw5071@psu.edu

EDUCATION

Pennsylvania State University 2015-2019

- Schreyer Honors College
- Major: Information Science and Technology

WORK EXPERIENCE

Undergraduate research assistant,
College of Information Sciences and Technology, Penn State 2019

- Schreyer Honors College conducting research on visual analytics and cognitive biases supervised by Dr. Xiaolong (Luke) Zhang

SKILLS

- Language Proficiency: Native in Chinese, proficient in English, intermediate in Japanese