Sources and Mitigation of Bias in Big Data for Transportation Safety

November 2018 | Final Report

SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VT | TRANSPORTATION INSTITUTE
VIRGINIA TECH

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 02-026	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Sources and Mitigation of Bias in Big Data for Transportation Safety		5. Report Date November 2018 6. Performing Organization Code:
7. Author(s) Greg P. Griffin Meg Mulhall Chris Simek		8. Performing Organization Report No. Report 02-026
9. Performing Organization Name and Address: Safe-D National UTC Texas A&M Transportation Institute The Texas A&M University System College Station, Texas 77843-3135		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/Project 02-026
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT) State of Texas		13. Type of Report and Period Final Research Report 14. Sponsoring Agency Code
15. Supplementary Notes This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas.		
16. Abstract Emerging big data resources and practices provide opportunities to improve transportation safety planning and outcomes. However, researchers and practitioners recognize that big data includes biases in who the data represents and accuracy related to transportation safety statistics. This study systematically reviews both the sources of bias and approaches to mitigate bias through review of published studies and interviews with experts. The study includes quantified analysis of topic frequency and evaluation of the reliability of concepts by using two independent trained coders. Results show a need to keep transportation experts and the public central in determining the right goals and metrics to evaluate transportation safety, in the development of new methods to relate big data to the total population's transportation safety needs, in the use of big data to solve difficult problems, and to work ahead of emerging trends and technologies.		
17. Key Words big data, bias, transportation, safety	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.	

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 26	22. Price $0

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Abstract

Emerging big data resources and practices provide opportunities to improve transportation safety planning and outcomes. However, researchers and practitioners recognize that big data includes biases in who the data represents and accuracy related to transportation safety statistics. This study systematically reviews both the sources of bias and approaches to mitigate bias through review of published studies and interviews with experts. The study includes quantified analysis of topic frequency and evaluation of the reliability of concepts by using two independent trained coders. Results show a need to keep transportation experts and the public central in determining the right goals and metrics to evaluate transportation safety, in the development of new methods to relate big data to the total population's transportation safety needs, in the use of big data to solve difficult problems, and to work ahead of emerging trends and technologies.

Acknowledgements

The authors appreciate interview coding support by TTI researcher Boya Dai, AICP, and expert review by Ed Chow, PhD.

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas.

Table of Contents

List of Figures

List of Tables

Introduction to the Problem of Bias in Big Data

Historically, high-quality data for transportation safety planning has been expensive and slow to obtain. Recently, new big data sources have allowed for more detailed analysis of vehicle, transit, bicycle, and pedestrian trips than ever before. However, big data generally represents transactions, such as rail transit payments, rather than trips, which could also include a walking portion of the journey, which means they inherently include a range of biases related to representation. Big data sources offer both prospects and problems for transportation planning in terms of how well they reflect the broad population of transportation system users or individual markets subject to digital divide and other representation biases. Research has identified far-reaching bias issues in big data sources; this study will focus on those with an impact on transportation safety planning. After conducting a synthetic literature review and interviews with expert practitioners, results suggest implications for transportation safety research and practices to identify and mitigate bias in big data.

Using a synthesis of literature and interviews with expert practitioners, this project addresses two critical questions:

1. What are the sources of bias in big data for transportation safety planning?
2. What are approaches to mitigating bias in big data for passenger vehicles, transit, bicycling, and pedestrians?

One practical definition of big data in a planning context is "when [the data] is too large and too complex to be stored, transferred, shared, curated, queried, and analyzed by traditional processing applications. There is no specific size that is assigned to big data, as it is always growing" (*1*). Other definitions include dimensions of the *volume* of data starting in the terabytes; *velocity* ranging from very recent to real-time; and *variety*, including multiple data formats that may be highly structured or informal, such as social media. Some suggest "the volume of data continues to double every three years as information pours in from digital platforms, wireless sensors, and billions of mobile phones" (*2*).

Big data is changing the way that transportation planners work, leading to questions and challenges of justice in how that data is used, with recent scholarship suggesting substantial changes may be needed to mitigate potential problems (*3*). For example, fitness app data is being used to gain a better understanding of cycling routes, but studies have shown that big data sources represent only a segment of the population, varying significantly from survey data using traditional sampling methods (*4*). Similarly, comparisons of multiple big data sources on the same routes show significant differences and represent a small fraction of total travel (*5*). Despite these and other challenges, big data "can reveal new dynamics, can allow for the study of certain processes in real time and can highlight relationships and correlations that may pass unnoticed using classical methods and data" (*6*). Leveraging new research on the veracity of big data in transportation and

smart cities, this study will guide both research into and practices of transportation planning. This project involves two phases: 1) a synthesis of literature on bias in big data and 2) interviews with leading practitioners of transportation safety planning with big data.

The next section briefly provides background on recent developments in big data for transportation and documented research needs. Following the background, we describe this study's methods, results, and conclusions.

Background of Big Data in Transportation and Research Needs

Connecting Big Data and Transportation Safety

Big data, collected by sensors that are part of transportation infrastructure, vehicles, cargo, or people, are promising for helping to answer questions about transportation safety that could lead to lives being saved. One of the key potential benefits of big data for transportation safety is understanding *exposure* to risk, which is often expressed as a measure of travel activities not captured by traditional data, such as walking trips (*7*). Some sensors, such as 24-hour cameras, provide non-biased data consistently if the stream can be parsed into useful information (*8*). However, most big data resources represent *items* rather than *people*—such as location tracking via smartphones or vehicles, or selectively posted social media. Big data categories and variables change with product cycles. A product update or data policy change impacts data used for safety analysis or research. Because of this, researchers using big data may have to alter their conceptions of transportation safety in ways that they do not when using the age-tested concepts of driving data collection. The fundamental issue is that "however big the data, Big Data are not about society, but about users and markets" (*9*). However, for transportation safety, the advantage of big data is that it allows analysis of changes nearly instantly. Further, the data may be available at spatial and temporal levels that reveal new relationships that are undiscoverable by traditional methods.

Both big data and traditional sources of data for transportation safety may include different forms of bias that might distort analyses. In some cases, these biases can impact real outcomes, such as where transportation funding is needed for safety improvements. Big data may include more observations about a transportation safety phenomena, but like surveys, still may not include all occurrences of those phenomena. Though survey research involves the development of a sampling frame that describes differences from the total population, the variance of big data observations from the entire population is not always available, creating challenges with measuring error and describing population coverage (*10*).

However, big data is in some cases the only data covering an issue of interest. Transportation agencies collect bicycling and pedestrian volume data sparingly, if at all, creating a challenge for understanding relative safety risks (*11*). New big datasets such as Google Street View imagery can

potentially be used for collecting pedestrian volumes, but are limited to available images, which do not currently include time-of-day metadata (*12*).

Emerging transportation technologies, such as automated vehicles (AVs), and vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) might leverage big data to improve safety and other outcomes (*13*). Big data also supports transportation network companies, helping them predict demand and allocate drivers and vehicles appropriately. Increases in big data-driven mobility services could "improve road safety by creating a more viable option that keeps people from getting behind the wheel when they have been drinking, they are excessively tired, or they have other impairments (such as difficulties with night vision)" (*2*). However, the way big data is used for transportation safety can also be a concern. False matches or ambiguities "will provide fertile ground for speculation, innuendo, and the exercise of preexisting biases for, and particularly against, racial, ethnic, religious, and socioeconomic stereotypes" (*14*). For instance, algorithmic selection of individuals for no-fly lists can inadvertently restrict personal rights (*9*). Moreover, peoples' awareness of privacy issues may impact the information they provide or their willingness to use a particular transportation technology, further complicating the availability of data and sampling biases. Depending on how agencies deploy big data analytics, adopted approaches could "exacerbate, magnify and accelerate the problem" (*9*).

As early as 2013, researchers shared the problem of biases in big data with broader audiences, including business leaders (*15*). Many already understand the potential for bias, but biases may be challenging to identify. As Crawford notes, "Hidden biases in both the collection and analysis stages present considerable risks and are as important to the big-data equation as the numbers themselves" (*15*). Articulation of needed research in this area can help focus what this means for the field of transportation.

Research Needs

Use of big data for transportation analysis calls for new quantitative approaches. Location-aware sensors in transportation networks, such as in-vehicle GPS and engine sensors, provide data that could potentially be used to improve transportation efficiency and safety (*16*). Next-generation research should take advantage of the high spatial and temporal resolution of big data, but also must develop new approaches to manage computational complexity and analysis times (*16*). The rise in the volume and velocity of data can potentially support predictive analytics—the use of big data to anticipate problems such as transportation safety issues—before they happen. However, little research to date explores how to implement or evaluate predictive analytics (*17*). Furthermore, new approaches to describing uncertainty and error propagation are needed for big data (*18*).

Beyond these needs, the transportation field needs to be able to leverage the best characteristics of both big data and traditional sources while balancing cost, and also must ensure professional training. Data fusion methods are needed to reap the advantages of big data, while still controlling for various biases through combination with datasets of known populations. Surveys are a key

opportunity for data fusion with big data, potentially supporting adjustment of big data as estimated values through controlling for population characteristics (*10*). Cost-effectiveness of using big data for transportation safety analysis needs further research, as agencies typically "do not calculate the costs associated with collecting, cleaning, managing, and updating Big Data" (*17*). Education content and approaches should be re-visited in light of big data's impact on the industry but should be supported by research evaluating its effectiveness. Initial research suggests planners are still trained for "a data poor environment," whereas they should be receiving education on how to handle big data through data mining, machine learning, simulation, and visualization (*19*).

Method

Synthesis of Literature

Research on big data has advanced quickly since 2010, and is influencing transportation and safety scholarship and practice. We searched three leading scholarly databases—Scopus, TRID, and Science Direct—to identify research related to sources and mitigation of bias in transportation research. Search terms in these databases included "big data," "transportation," and "safety," resulting in 135 publications addressing this study's research questions between the years 2010 and 2017. Among these, the project's graduate student found 75 studies that addressed the research questions directly enough to be included for analysis of method, type of data, sources of bias, mitigation techniques, topics (transportation planning, planning in general, safety, and smart cities), and surface mode (transit, surface freight, automobile, bicycle, and pedestrian). The principal investigator reviewed initial categorizations for each study and revised four.

We used this categorization of literature to form a basis for synthesizing findings and recommendations across fields on the topics of identifying and mitigating bias in big data.

Expert Practitioner Interviews

To further explore the issues associated with bias when using big data in transportation planning practice, the second part of this report is informed by interviews with practitioners in the field.

Developing an Interview Guide

We constructed a semi-structured interview guide to focus on insights from expert practitioners of big data in transportation, intended to capture ideas on our research questions beyond already published research. We expected expert practitioners to come from a range of backgrounds and employment, with participants including employees from public departments of transportation, private sector consultants and data providers, and researchers. The semi-structured interview approach allowed the interviewer to focus the discussion on topics of interest while encouraging the interviewee to emphasize particular areas of interest and expertise (*20*). We included "probes" as bulleted topic items, which were related to our research questions, on the interview guide. Some probes were expected to fit some interviewees more than others, providing additional interview flexibility. The interview guide was reviewed by the human subjects protection programs at both

Texas A&M and Virginia Tech and is available in this report as Appendix A: Semi-Structured Interview Guide: Bias in Big Data for Transportation Safety.

Identifying Big Data Experts

Interview candidates were identified by using the search term "big data" in conference agendas for two recent planning and transportation conferences: the 2017 and 2018 Transportation Research Board (TRB) Annual Meetings, and the 2017 American Planning Association (APA) National Planning Conference. From these search results, interview candidates were chosen according to the following criteria: a) they were presenting research or participating in round table discussion of big data use at the named conferences, b) they held a position within their organization that gave them substantial knowledge of the strengths and limitations of big datasets, and c) their current contact information was readily accessible through conference agenda(s) or online.

The second criterion for interview candidate selection—that the practitioner hold a position within their organization that gives them substantial knowledge on the strengths and weaknesses of big datasets—is intentionally broad. While we hoped to reach interviewees with knowledge of and experience in applying the techniques necessary to mitigate bias encountered in big datasets, this broad scope of inclusion allowed us to capture organizational representatives with more diverse career backgrounds and industry affiliations. For example, this criterion allowed us to capture interview candidates such as a travel demand modeler at a state department of transportation as well as a salesperson at a company that compiles and sells big datasets for transportation planning purposes.

We requested interviews via email with 39 experts in the use or analysis of big data for transportation, and 10 respondents completed interviews, resulting in a response rate of 26%. Figure 1 shows that most interviewees were located across the United States, with one respondent located in Toronto, Canada. Interviewees were engaged in four categories of work: universities (N=4), private sector (N=3), state departments of transportation (N=2), and a city transportation department (N=1). To encourage candid responses, we offered anonymity.

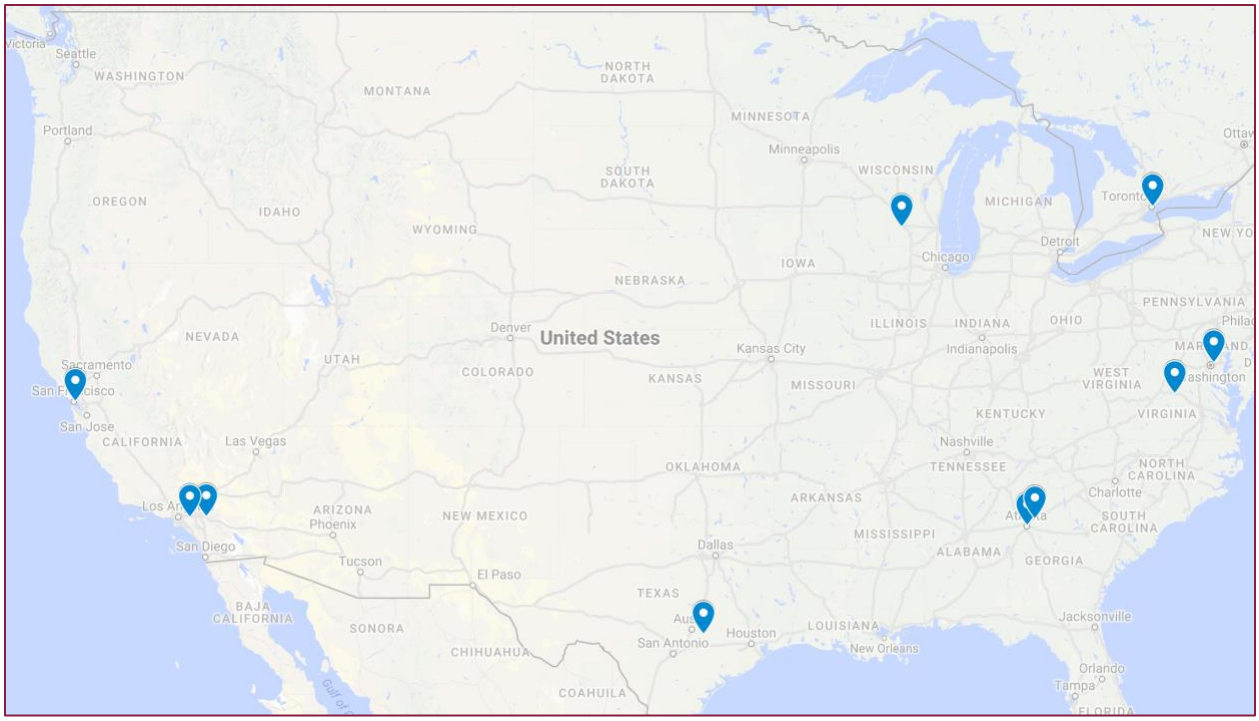

Figure 1. Map of interviewee locations.

Conducting Interviews

We completed interviews remotely via the chat function in online conferencing software. We requested interviews via text-only (no audio or video). One participant preferred to use full audio and video. There are advantages and disadvantages to conducting interviews online and the using text-only approach in particular (*21*). The first advantage is pragmatic—given the limited time and budget of the project, text-based interviewing requires no transcription. The software provides a complete transcript of text between all parties in the interview, including time stamps. The second advantage is that there is no loss of data through mis-transcription of audio, or note-taking during the meeting. Third, online interviewing was convenient for our study population of big data experts, who have access to broadband internet and may be better able to accommodate an interview in their workspace with little interruption or disruption of schedule. Online text-based interviews are synchronous—allowing real-time interaction—yet they allow hesitations for reflection that might be awkward in person. Disadvantages of online text-based interviewing include the lack of non-verbal communication and nuances in speech, and limits on the speed of information flow due to typing, as compared with real-time speech. We typed notes during the interview with the single interviewee who preferred full audio and video to typing. This interview process resulted in a qualitative dataset useful for original insights, but we caution generalization of the responses to others due to the small number of interviewees.

Coding

Despite the relatively small dataset, coding of key issues in the interview corpus allows review of content in a consistent method focusing on crucial themes (*22*). Coding of qualitative data, such

as interviews, requires observers or readers to categorize, scale, or measure each of a given set of predefined units of analysis, in effect characterizing them by one value from each variable of analytical interest. However, there are many occasions in which texts…have multiple interpretations" (*23*). This study employs an approach developed in 2016, in which two analysts evaluate the reliability of multiple codes assigned to each interview question, resulting in an overall reliability coefficient (*23*).

Interview coding involved three steps: 1) developing a codebook to guide analysis of interviews, 2) coding of the interview data itself, and 3) preparation for reliability analysis. The principal investigator developed a codebook to enable evaluation of the interview content by other researchers. Included in this report as Appendix B: Interview Coding Instruction, the codebook focuses only on key concepts related to this study's research questions, described in qualitative analysis research as structural code analysis (*24*). The principal investigator then formatted interview content into de-identified spreadsheets, with responses to each question on a new row for coder review. Two researchers then independently coded the interview corpus, reviewing each interview response for each question for the occurrence of any of the seven codes. This process mirrors a double-blind review since neither the coders nor interviewees knew each other. We went a step further and worked with two researchers, who were not involved in the interviews in any way, to code data. Since assigning qualitative data code to characterize interview content is inherently subjective, responses from both researchers' codes required comparison to evaluate the reliability of the codes as constructs describing big data for transportation safety. Researchers did not iteratively discuss or re-evaluate their coding based on each other's preliminary work. Codes were then re-formatted for reliability evaluation using open-source Multiple Valued Nominal Alpha software, available from https://github.com/rcraggs/mvna/releases and http://www.asc.upenn.edu/mvnAlpha.

Reliability Through Multi-Valued Nominal Agreement

Two measures of coding reliability are common: simple percentage agreement of codes between coders, and a coefficient ranging from 0 = complete disagreement to 1 = perfect agreement, which includes calculation of the random likelihood of codes being the same. This study includes both, with the latter including advancement to support multiple values (codes) for a single unit of data (interview response to a question).

Reliability coefficients are evaluated as the relationship between the observed and expected disagreements between coders. The equation below shows a general form of this calculation, with the multi-valued alpha coefficient ${}_{mv}\alpha$ expressed as one minus the ratio of observed multi-valued codes across the data corpus over the expected codes:

$$ {}_{mv}\alpha = 1 - \frac{{}_{mv}D_o}{{}_{mv}D_e} $$

Multi-valued nominal agreement extends this formula to account for the use of multiple codes per analysis unit, no applicable codes per unit, missing data, and any combination of these issues, in addition to bootstrapping and other techniques to evaluate the probability of codes reflecting agreed constructs, rather than agreement by chance (*23*). Coders are not required to choose one code that "best" characterizes the data unit—they can choose to use no codes, one, many codes, or skip the entry. In this way, we can evaluate the reliability of the key constructs of bias and mitigation of big data bias in the interview corpus.

Results

Overview of Topics in Literature and Interviews

Before detailing sources and mitigation of bias in big data, this section provides a broad characterization of our findings from the literature search and interviews.

Literature

Review of the big data literature by topic and mode in this sample reveals an emphasis on studies of transit. Ten of the transit studies exploited social media data. Social media included both geo-tagged posts to analyze location and time, and textual analysis to study perceptions. The emphasis on social media may be particular to the period of analysis, spanning the origin and rise of social media in contemporary life, and riders' use of social media. We also found safety analysis within other travel modes, including surface freight, automobile use, bicycling, and walking. In addition to safety, we found applicable studies in transportation planning, planning in general, and smart cities that addressed issues of bias in big data.

Table 1. Co-occurrence of Topic and Travel Mode in Big Data in Transportation Articles, 2010–2017 (N = 75)

Research Topic	Transit	Freight	Automobile	Bicycle	Pedestrian
Transportation Planning	4	1	4	1	2
Planning (general or field other than transportation)	1	0	1	0	0
Safety	14	2	3	1	1
Smart Cities	6	2	2	2	4

Note: Some articles included only one topic (no co-occurrence), while others included multiple listed topics and modes, so total co-occurrences do not match the number of articles.

Interviews

Analysis of interview data focused on the codes developed to address the research questions regarding sources and mitigation of bias in big data. In addition to codes for bias and mitigation, we included the same travel mode classifications as in the literature review: transit, freight,

automobile, bicycle, and pedestrian. We added an additional code, *interpreting*, to help identify how interviewees understood big data as a concept or technology.

Big data experts responded to our interview prompts with insights ranging between two- and more than 200-word responses to individual questions, some including several of our concepts for each of the seven items. Our interview coding system allowed for analysis of both individual and combinations of topical codes in a single response. Co-occurrence analysis of topics similar to Table 1 resulted in 77 different combinations of codes, including each of the codes and combinations of codes found in each response. Overall, the two coders agreed on 76% of individual responses, resulting in a multi-value nominal alpha coefficient ($_{mvn}\alpha$) of 0.544. This means that coders interpreted individual interview responses the same more than half of the time after eliminating the probability of agreement on any code due to chance.

Source (of bias) was the most common code in the dataset, identified 23 times by coders. For instance, a respondent employed as a researcher noted that, "big data from mobile devices probably under-represents older segments of the population and maybe lower income populations. However, it also does a better job of representing under-represented road users like bicyclists and pedestrians." Other respondents similarly showed nuance in how bias is prevalent in big data and used in practice—the presence of bias did not preclude thoughtful implementation for our respondents.

The frequency of codes relating to travel modes suggests interviewees' interests and experience. The most common modal code was vehicles (mentioned nine times), then bicyclists (mentioned eight times), and transit (mentioned three times) Interviewees did not explicitly discuss pedestrian travel in the context of big data.

Mitigation (of bias) was coded in nine interviews instances. One researcher described a range of ways that big data users can mitigate bias, including the following suggestions:

> Combine big data with "traditional data" like surveys. Interpolate big data or use it to create synthetic populations. Wait for larger sample sizes. Skilled experts can work with raw data using data mining and machine learning techniques. Less-skilled users can acquire data from intermediaries that make sense of it for them. Even then, they probably need basic data management skills and often GIS capability.

This type of response suggests that productive mitigation tactics are, in many cases, still on the horizon of practice. Review of the literature shows that data fusion techniques to mitigate bias are of interest, but few big data providers or practitioners currently use these methods—mitigation of bias is limited by skills.

Interpretation (of how big data is understood as a concept or technology) was coded 11 times. A researcher described big data as potentially problematic regarding high volumes that may contribute little to understanding, noting, "our cities are getting data-obese. We need to learn how

to cure them. I think data quality is an important factor to consider. Data bias is a concerning sign of sickness."

Sources of Bias in Big Data

Review of literature and interviews suggests three broad categories of sources of big data, and five types of bias found in big data. This section reviews sources of bias in big data from mobile phones, social media, and travel observation, and describes *aggregation bias, coverage bias, non-response bias, sampling (or demographic) bias, selection bias,* and *social desirability bias* through examples from research and interviews.

Mobile Phone Data

Mobile phone data is one of the most widely used big datasets in transportation planning (*25*). One interviewee described big data as mainly "being generated through cell phones or non-engagement [passive] methods." There are generally two types of data passively collected from mobile phones that are available to transportation planners: call data records (CDR) and sightings data. CDR are records of interactions; information such as the caller, call recipient, duration of the call, and the location of the tower routing the call are preserved each time a mobile phone user places a call. Sightings data are recorded each time a mobile phone is positioned and, therefore, are likely to have higher temporal and spatial resolution than CDR, which record one entry per call and record one tower location for that call (*25*).

However, because these data are generated passively, recorded each time a mobile phone connects with a tower in the cellular network, integrating this information into the transportation planning process requires a great deal of inference. Because mobile phone data, like CDR or sightings data, are not collected for transportation planning purposes, the data often do not answer typical transportation safety research questions (*26*). For this reason, bias can enter the transportation planning process anywhere planners are making inferences from cell phone-based big datasets (*25*).

The biases arising from inference can be exacerbated in situations where researchers lack a contextual understanding of what the data mean on the ground. The current gaps in access and capacity between research disciplines have led to a scenario where big data research is often undertaken remotely by computer and data scientists, rather than in the study community by social scientists (*27*). Proprietary datasets held by corporations don't allow access to certain big datasets, and the capacity of different research disciplines to employ technical skills in advanced statistics and computer science has led to a noticeable gap between those able to analyze big data and those able to understand the context of the data in terms of the study community (*27*).

Bias can also creep into the transportation planning process if researchers are not clear about the provenance of their mobile phone big data. Although mobile phone data are one of the most applied and researched in the realm of transportation planning, some researchers have expressed uncertainty about the way in which the data they access were collected (*26*). Understanding, for

example, the different "levels of geographic specificity" between call data records and sightings data as discussed above is a key prerequisite for integrating such big data into transportation planning processes (*25*, *27*).

Once transportation planners and researchers have a clear understanding of how their big data were collected, and if they can find ways to contextualize the data in their study communities, they can then begin to focus on mitigating *sampling-related biases* that make mobile phone data unrepresentative. First, issues like mobile phone ownership rates fundamentally affect who is covered in a sampling frame based on mobile phone data (*25–27*). Because less than 100% of the transportation planner's target population uses a mobile phone, let alone a mobile phone served by the same carrier, the sampling frame is subject to *coverage bias* (*25–28*). Furthermore, if those not present in the data sample (those who do not own or use a mobile phone) differ systematically from those who are present in the sample (mobile phone users) in terms of key demographics or other inputs, the sample is subjected to *non-response bias* (*25–28*).

Beyond market penetration by mobile phone carriers and personal mobile phone ownership rates, the differences in personal use of mobile phones can inject bias into transportation planning based on mobile phone data. As explored in the definitions of call data records and sightings data, data is only recorded when a mobile phone is in use and connected to the cellular network, resulting in irregular sampling frequencies (*26*). Transportation planners using mobile phone data in their work should recognize that, due the way data is recorded, differences in mobile phone use from user to user may translate to some users being underrepresented and some being overrepresented in a given sample (*27*).

Finally, mobile phone data is subject to *measurement bias* because the records in the dataset may not correctly or precisely describe the indicators of interest to the researcher (*28*). For example, transportation planning researchers cannot assume that one SIM card record in their dataset represents just one targeted user (*27*). Multiple users may be using one shared SIM card or phone or, on the other hand, one user may have multiple mobile phones or SIM cards (*25*, *27*). Furthermore, the proximity of mobile phones in the study area may make it difficult for transportation planning researchers to parse out individual mobile phone users to make inferences about subjects such as travel trajectory (*25*).

Social Media Data

Despite the findings present in the literature of frequent use of social media as a source of big data for research, our interviewees did not mention this connection. Several studies produced particular insights on the topic of bias. Tass and Hong identified three kinds of sampling bias from the use of geotagged social media data to understand urban dynamics (*29*). First, the use of Twitter and Foursquare represent participants' voluntary actions, rather than strict records of urban movement. For example, the authors note that "the Museum of Modern Art in New York has more check-ins than Atlanta's airport, even though the airport had almost three times as many visitors in the period that was studied" (*29*). Second, social media users who provide location data do so to show where

they want to be seen, rather than where they actually travel. Researchers describe this problem as *social desirability bias* (*10*, *30*), in which people provide information to confer status, rather than to accurately depict their activity. Third, urbanites in general, and "young, male, technology-savvy people" use social media at higher rates than other groups, creating a *sampling bias* for analysts wishing to understand transportation system users in general (*31*).

Travel Observation

A literature review revealed examples of big data research dealing with bias in each major travel mode. A private-sector interviewee described their work with big data for travel observation as including "a lot of multi-agency systems and traveler information systems, [integrating] data from a variety of sources … for better information." This interviewee noted that, "local planners have come to us once we have started data warehouses to get data for their needs," suggesting this area is an emerging field that depends on finding ways to make big data resources more refined and accessible.

Automobile data collection can include sensors that are part of the vehicle itself, such as GPS and toll tags, as well as sensors carried by the driver, including smartphones. As previously suggested, a *sampling (or demographic) bias* occurs when the people buying the products tracked, like a car or phone providing travel observation data, do not represent the total population. GPS-based travel surveys may be more accurate than traditional travel diaries in terms of correctly logging the time and routing of trips, but can introduce problems with correct identification of travel mode and trip purpose (*32*), which are key inputs for travel modeling. However, big data can be particularly useful for tracking complex travel behaviors such as ridesplitting (*33*). This area is likely to advance quickly as more data from transportation network companies such as Uber and Lyft become available.

Bicycle data collection incorporates sensors that may be part of a dedicated bicycle computer (combined speedometer, GPS map, and often heart rate or power calculations) as well as mobile phone apps. Garmin and Strava, for instance, both offer sports-oriented platforms that can record the same travel observation data on either a dedicated bicycle computer or a smartphone app (*34*, *35*). These datasets provide a new opportunity to understand bicycling, but the use of the equipment and apps (often expensive) create a significant *sampling (or demographic) bias* (*36*). Review of demographics of Strava users in Travis County, Texas, and elsewhere, shows that the users of these apps reflect more of the male, and young to a middle-aged segment of the population (*5*, *37*, *38*). Further, bicyclists may choose not to record all of their trips with many of these apps that also provide online social sharing (*39*)—again connected to a *social desirability bias*, where participants may only log bicycle trips that are sufficiently fast or long to record and share as an accomplishment. Bicycle sharing systems and emerging scooter platforms also generate trip data, which still only reflect the customers of each system.

Pedestrian travel observation through big data, beyond simple counts using automatic detectors, is in its infancy (*11*). One researcher interviewee confirmed that the problem with big data for

pedestrian travel "is that there is some imbalance from those represented in big data versus all pedestrians…this could bias against older people who do not use a cell phone." Mobile apps that track running have the same bias problems as bicycling data (*36*), but may be even more exacerbated based on use—walking trips may be too numerous to track using an app that requires activation for each trip. However, new approaches that can track pedestrian trips using accelerometers or other sensors may support more broad representation of a pedestrian community. Ride Report is one app developed for bicycling that attempts a classification of walking trips and may be useful for studies of transportation and health (*40*). The app also supports the ability for individuals to re-classify their travel mode for each trip, in case the algorithm detects it incorrectly. As of this writing, Ride Report does not have a pedestrian data product, but is part of a rapidly-advancing developer community that may provide additional pedestrian travel observation data.

Surface freight is a particularly challenging mode to track, which is perhaps associated with the proprietary nature of competitive businesses. To protect business interests, big data providers must aggregate individual trips and shipments at a level tolerable to the freighters. Aggregated data may yield different analysis results from individual data, resulting in an *aggregation bias* inherent in analysis products (*41*).

Transit data ranges widely from automatic counter systems and smartcards that record essentially all users of the system—and therefore have little inherent bias—to social media data that fails to represent the entire population of users. One analysis of transit reviews on social media found that the demographic of social contributors did not match the target of transit users—the demographic skewed young and affluent (*42*). Also, the social reviews found online did not always focus on how transit systems could improve, limiting the usefulness of this information for system planning and operation.

Mitigating Bias in Big Data

Overall Methodological Approaches

Rigorous research designs, such as randomized control trials and mixed-methods approaches, may identify specific biases in transportation data. Randomized control trials involve assigning individuals or groups randomly to categories, such as in a transportation analysis process, and then evaluating impacts to each category. The differences may reveal biases in transportation planning or operational data that could result in unfair and/or dangerous conditions. Mixing methods, such as incorporating qualitative data like interviews with big quantitative data may help contextualize data for better understanding of biases in big data, and approaches to mitigating those biases (*1*).

Mitigating Bias in Automobile, Bicycling, and Walking Data

Five different studies pointed out methods to mitigate bias in big data relating to motorized vehicles, but the methods are broadly applicable to data from any transportation mode. The first step of mitigating bias is to filter out unreasonable data points; this could include speed-based or pattern-based filtering methods to flag likely incorrect or inapplicable data for a given

circumstance (*25*, *43*). One interviewee from a department of transportation noted the importance of treating vehicular customers as equals, describing their operational intent as follows:

> …all vehicles are equal and so we work to manage traffic as equally as possible… whether the [data-providing] vehicle is a traditional personally owned vehicle, an Uber in route to a pickup, or an Uber with a customer.

Regardless of the apparent focus on vehicles rather than people as customers, this suggests a need to evaluate how big data represents actual users of the system. Data fusion, such as integrating big data with census data, can help control for *sampling bias*, making the combined product more representative of the population (*26*). Another approach to improve representation includes over-or-under sampling of data, such as collecting more data from under-represented users (*42*). Developing and including high-quality metadata that describes the development and refining of any big data resource helps users evaluate and fix problems with bias (*44*). Finally, it would be useful to provide ways for data users or consumers to share data cleaning and analysis methods somehow, such as in an online forum (*44*). This final method could be particularly helpful for new markets and applications of big data, where the users represent the subject matter experts for how the data should be interpreted and deployed in practice. These techniques are broadly applicable to mitigating bias in big data, including non-motorized modes.

Mitigating Bias in Freight Data

Surface freight can be challenging to understand as a whole, limiting opportunities to mitigate biases in big data. This report mentioned *aggregation bias* in freight data resulting from groupings of individual data. Mehmood and others describe a Markovian approach to correct for *aggregation bias* (*41*), which may be considered along with other techniques such as temporal or spatial models that incorporate a random selection of individual data to detect bias and adjust fit to the aggregated data.

Mitigating Bias in Transit Data

Operational characteristics of transit supports several approaches to mitigate bias in big data. Monitoring can occur at the point of fare purchases, vehicle boarding, and vehicle alighting. In addition, requirements for planning and reporting to funding agencies provide aggregate statistics to evaluate other data sources. For instance, big data can be validated against traditional travel survey methods like on-board counts and surveys, in addition to gaining an understanding through interviews (*45*). Combination of qualitative data, such as interviews, with big quantitative data such as alighting counts, shows a mixed-methods approach to validating big data (*46*). Mixed-methods approaches can help answer questions relating to *how* and *why* big data should be viewed skeptically or mitigated in specific contexts.

Discussion of Implications for Research and Practice

Big Data as New Data

Governments and private sector organizations have developed data gathering methods appropriate for transportation safety over decades, if not centuries. Big data does not automatically replace these efforts; in fact, in some cases, big data may be less appropriate. However, some sources of big data also constitute *new* data, capable of addressing problems that traditional approaches have missed. For instance, use of crowdsourced bicycling data provides travel volumes over a broad area and fine time scale, which was not previously captured (*5*), enabling analysis of safety considering relative risk similar to motorized modes. Similarly, vehicle-tracking data such as that provided by ridehailing companies like Lyft enable analysis of detailed origin-destination trips at a fine time scale (*47*). Similar to the new bicycling data, this resource could support thorough understanding of modes not captured by traditional data collection techniques. However, neither approach is widely used for transportation safety at the time of this report. Agencies should explore how to test and implement these approaches to save lives while improving mobility.

Disparate Impact

Advocates and researchers have identified ways in which software algorithms may leverage data so as to result in a discriminatory policy. As Desouza and Smith note, "Disparate impact is the idea that a policy is discriminatory if it has an adverse impact on any group based on race, gender, sexual orientation, religion, or any other protected status" (*1*). Statistical analysis of housing data in Texas resulted in a 2015 Supreme Court finding of disparate impact that the Texas Department of Housing and Community Affairs had perpetuated "segregated housing patterns by allocating too many tax credits to housing in predominantly black inner-city areas and too few in predominantly white suburban neighborhoods" (*48*). Suresh Venkatasubramanian at the University of Utah's School of Computing conducts test on datasets that include no race or gender information through modeling to correctly predict race or gender. He suggests these datasets can be re-arranged in ways that do not harm the data, but prevent them from identifying groups that could lead to disparate impacts (*1*). Though big data is only beginning to impact research and practice for transportation safety, the potential for inequitable impact is logical. As noted in the aforementioned Supreme Court finding, "Big data continues to present blind spots and problems of representativeness, precisely because it cannot account for those who participate in the social world in ways that do not register as digital signals" (*49*). Proactive organizations can analyze big data related to transportation safety to identify potential disparate impacts related to transportation safety.

Open Data

Review of literature and interviews show the importance of verifying big datasets against other resources. Open data can refer to an organizational policy of providing data to the public, usually at no cost, as a standard practice rather than only releasing certain datasets. Researchers have shown the importance of open data in terms of planning (*50*), transit system operation and

management (*51*), and megaregional governance (*52*). Open data is an ethical issue for some, who note the potential of big data systems to "consolidate power in the hands of experts and large private firms to the exclusion of citizens and small, independent firms" (*50*). However, merely providing public access to data does not mitigate problems of bias in big data alone—all of the methods discussed in this study still apply as does the need for agencies to simplify access to, and interpretation of, open data.

Conclusions and Recommendations

Our review of current literature and a limited set of interviews show both challenges and opportunities for the use of big data for transportation safety. The field is still emerging, and researchers to date have focused on isolated projects with little synthesis of understanding for improving transportation safety analysis. We used traditional techniques of literature review combined with recent analytical improvements in qualitative coding to evaluate the key issues in bias and how to mitigate those issues in practice and research. This report collects the existing research on the topic to date, and contextualizes current insights through interviews with experts.

Transportation organizations have four issues of key concern when thinking about biases in big data and how to mitigate those biases in transportation safety:

1. **Keep transportation experts and the public central in determining the right goals and metrics to evaluate transportation safety.** Big data may provide detail on how to address specific issues, but research and expert interviews suggest the existence of a particular dataset should not drive the prioritization of goals and metrics.
2. **Develop new methods to relate big data to the total population needed for transportation safety.** Integrating traditional census data, surveys, and traffic counts with new big data sources may help users understand who, what, where, and how many crashes occur at fine locational and time scales. These approaches represent the cutting-edge in current research on big data for transportation safety; effective use of data fusion to make big data more representative should be a key concern.
3. **Leverage big data to answer difficult questions.** Assumptions and expectations about what transportation safety problems are answerable using traditional methods may not hold when considering analytic opportunities presented by big data. Bicycling, ridehailing, and pedestrian trips represent emerging areas for which agencies have not traditionally held quality data for evaluating safety.
4. **Work ahead to transfer emerging knowledge to future problems.** The transportation safety implications of automated vehicles are enormous (*53–56*). However, many studies to date have used simulated data and conjecture to forecast potential safety impacts. Current big data, such as ridehailing data, could be extremely useful as empirical proxies for technologies that are not currently measurable on a large scale, such as autonomous vehicles.

All of these approaches suggest needs for a tight relationship between transportation safety researchers and practitioners. Researchers should focus on the real and ethical implications of their big data research on transportation safety, listening to practitioners for key issues that should drive outcomes and lead to lives saved. Practitioners can turn to researchers to help evaluate new big data resources, remaining cognizant that most datasets do not represent the total population, and few methods exist to reasonably adjust them to compensate. The private sector is rapidly developing new big data resources that can potentially save lives with thoughtful application, but their advancements should be critically reviewed before implementation in transportation safety practice.

This study suggests more opportunities for further research than answers to bias problems to date. Non-motorized modes present a major opportunity to improve transportation safety planning, operation, and outcomes. Emerging travel modes such as neighborhood electric vehicles and autonomous travel are also principal opportunities to leverage existing big data for present and future endeavors.

Additional Products

Education and Workforce Development Products

Findings from this study were presented to a national webinar audience, through the Transportation Research Board Bicycle and Pedestrian Data Subcommittee. Held June 26, 2018, project PI Greg P. Griffin presented the webinar under the title: "What do the Experts Do? Insights from Interviews & Literature to Deal with Bias in Big Data." Two other presenters contributed to the webinar event called "Conversations about Counting: Big Data – Implications for Bicycle and Pedestrian Traffic Analysis."

The graduate student involved in this project, Meg Mulhall, is also finalizing her master's report in Community and Regional Planning at The University of Texas at Austin, based on this study.

Technology Transfer Products

The project report will be the primary product of this study, which will be posted on at least three platforms: the Safe-D website, the TTI publications catalog, and the TRID database.

A summary research paper of these findings was accepted for presentation at the 2019 Transportation Research Board Annual Meeting, paper number 19-03196.

Researchers are revising the paper for submission to a double-blind peer-reviewed journal.

A practitioner-oriented article will be developed for a trade magazine or blog.

Data Products

Complete transcripts are available on the Safe-D Dataverse in a single text file. The transcripts are organized by interview question, from 1 through 7. Transcripts do not include personal identifiers.

References

1. Desouza, K. C., and K. L. Smith. *PAS Report 585 Big Data and Planning*. American Planning Association, Chicago: IL, 2016.

2. Henke, N., J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, and G. Sethupathy. *The Age of Analytics : Competing in a Data-Driven World*. 2016.

3. Schweitzer, L. A., and N. Afzalan. Four Reasons Why AICP Needs an Open Data Ethic. *Journal of the American Planning Association*, Vol. 83, No. 2, 2017, pp. 161–167. https://doi.org/10.1080/01944363.2017.1290495.

4. Bergman, C., and J. Oksanen. Estimating the Biasing Effect of Behavioural Patterns on Mobile Fitness App Data by Density-Based Clustering. In *Geospatial Data in a Changing World* (T. Sarjakoski, M. Y. Santos, and L. T. Sarjakoski, eds.), Springer, Cham, pp. 199–218.

5. Griffin, G. P., and J. Jiao. Crowdsourcing Bicycle Volumes: Exploring the Role of Volunteered Geographic Information and Established Monitoring Methods. *URISA Journal*, Vol. 27, No. 1, 2015, pp. 57–66.

6. Shearmur, R. Dazzled by Data: Big Data, the Census and Urban Geography. *Urban Geography*, Vol. 36, No. 7, 2015, pp. 965–968. https://doi.org/10.1080/02723638.2015.1050922.

7. Jiao, J., A. V. Moudon, and Y. Li. Locations with Frequent Pedestrian-Vehicle Collisions: Their Transportation and Neighborhood Environment Characteristics in Seattle and King County, Washington. In *Planning Support Systems for Sustainable Urban Development* (S. Geertman, F. Toppen, and J. Stillwell, eds.), Springer Berlin Heidelberg, Berlin, pp. 281–296.

8. Hipp, J. A., D. Adlakha, A. A. Eyler, R. Gernes, A. Kargol, A. H. Stylianou, and R. Pless. Learning from Outdoor Webcams: Surveillance of Physical Activity Across Environments. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 471–490.

9. Shearmur, R. Dazzled by Data: Big Data, the Census and Urban Geography. *Urban Geography*, No. August 2015, 2015, pp. 1–4. https://doi.org/10.1080/02723638.2015.1050922.

10. Johnson, T. P., and T. W. Smith. Big Data and Survey Research: Supplement or Substitute? In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 113–125.

11. Griffin, G., K. Nordback, T. Götschi, E. Stolz, and S. Kothuri. *Monitoring Bicyclist and Pedestrian Travel and Behavior*. Transportation Research Board, Washington, D.C., 2014.

12. Yin, L., Q. Cheng, Z. Shao, Z. Wang, and L. Wu. 'Big Data': Pedestrian Volume Using Google Street View Images. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N.

Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 461–469.

13. Krishnamurthy, R., K. L. Smith, and K. C. Desouza. Urban Informatics: Critical Data and Technology Considerations. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 163–188.

14. Wigan, M. R., and R. Clarke. Big Data's Big Unintended Consequences. *Computer*, Vol. 46, No. 6, 2013, pp. 46–53. https://doi.org/10.1109/MC.2013.195.

15. Crawford, K. The Hidden Biases in Big Data. *Harvard Business Review*. 9–10. https://hbr.org/2013/04/the-hidden-biases-in-big-data. Accessed Aug. 4, 2017.

16. Gunturi, V. M. V, and S. Shekhar. Big Spatio-Temporal Network Data Analytics for Smart Cities: Research Needs. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 127–140.

17. Nguyen, M. T., and E. Boundy. Big Data and Smart (Equitable) Cities. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 517–542.

18. Thakuriah, P., N. Y. Tilahun, and M. Zellner. Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 11–45.

19. French, S. P., C. Barchers, and W. Zhang. How Should Urban Planners Be Trained to Handle Big Data? In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 209–217.

20. Adams, W. C. Conducting Semi-Structured Interviews. In *Handbook of Practical Program Evaluation*, John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 492–505.

21. O'Connor, H., C. Madge, R. Shaw, and J. Wellens. Internet-Based Interviewing. In *The SAGE Handbook of Online Research Methods* (N. Fielding, R. M. Lee, and G. Blank, eds.), SAGE Publications, Ltd, London, pp. 271–289.

22. Ose, S. O. Using Excel and Word to Structure Qualitative Data. *Journal of Applied Social Science*, Vol. 10, No. 2, 2016, pp. 147–162. https://doi.org/10.1177/1936724416664948.

23. Krippendorff, K., and R. Craggs. The Reliability of Multi-Valued Coding of Data. *Communication Methods and Measures*, Vol. 10, No. 4, 2016, pp. 181–198. https://doi.org/10.1080/19312458.2016.1228863.

24. Saldaña, J. *The Coding Manual for Qualitative Researchers*. SAGE Publications, Thousand Oaks, CA, 2016.

25. Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis. *Transportation Research Part C: Emerging Technologies*, Vol. 68, 2016, pp. 285–299. https://doi.org/10.1016/j.trc.2016.04.005.

26. Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González. The Path Most Traveled: Travel Demand Estimation Using Big Data Resources. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 162–177. https://doi.org/10.1016/j.trc.2015.04.022.

27. Taylor, L. No Place to Hide? The Ethics and Analytics of Tracking Mobility Using Mobile Phone Data. *Environment and Planning D: Society and Space*, Vol. 34, No. 2, 2016, pp. 319–336. https://doi.org/10.1177/0263775815608851.

28. Bonnel, P., C. Bayart, and B. Smith. ScienceDirect Workshop Synthesis: Comparing and Combining Survey Modes. *Transportation Research Procedia*, Vol. 11, 2015, pp. 108–117. https://doi.org/10.1016/j.trpro.2015.12.010.

29. Tasse, D., and J. I. Hong. Using User-Generated Content to Understand Cities. In *Seeing Cities Through Big Data* (P. (Vonu) Hakuriah, N. Tilahun, and M. Zellner, eds.), Springer Geography, Cham, Switzerland, pp. 49–64.

30. Beecham, R., and J. Wood. Exploring Gendered Cycling Behaviours within a Large-Scale Behavioural Data-Set. *Transportation Planning and Technology*, Vol. 37, No. 1, 2013, pp. 83–97. https://doi.org/10.1080/03081060.2013.844903.

31. Murphy, J., M. W. Link, J. H. Childs, C. L. Tesfaye, E. Dean, M. Stern, J. Cohen, M. Callegaro, and P. Harwood. *Social Media in Public Opinion Research : Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research*. Deerfield, IL, 2014.

32. Vij, A., and K. Shankari. When Is Big Data Big Enough? Implications of Using GPS-Based Surveys for Travel Demand Analysis. *Transportation Research Part C*, Vol. 56, 2015, pp. 446–462. https://doi.org/10.1016/j.trc.2015.04.025.

33. Chen, X., M. Zahiri, and S. Zhang. Understanding Ridesplitting Behavior of On-Demand Ride Services: An Ensemble Learning Approach. *Transportation Research Part C*, Vol. 76, 2017, pp. 51–70. https://doi.org/10.1016/j.trc.2016.12.018.

34. Kitchel, D., and B. Riordan. *Strava Metro Product Documentation*. Hanover, NH, 2014.

35. Garmin. Garmin Connect. https://connect.garmin.com/en-US/. Accessed May 31, 2018.

36. Bergman, C., and J. Oksanen. Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing. *Transactions in GIS*, Vol. 20, No. 6, 2016, pp. 848–868. https://doi.org/10.1111/tgis.12192.

37. Griffin, G. P., and J. Jiao. Where Does Bicycling for Health Happen? Analysing Volunteered Geographic Information through Place and Plexus. *Journal of Transport & Health*, Vol. 2, No. 2, 2015, pp. 238–247. https://doi.org/10.1016/j.jth.2014.12.001.

38. Boss, D., T. Nelson, M. Winters, and C. J. Ferster. Using Crowdsourced Data to Monitor Change in Spatial Patterns of Bicycle Ridership. *Journal of Transport & Health*, 2018. https://doi.org/10.1016/j.jth.2018.02.008.

39. Smith, W. R. Communication, Sportsmanship, and Negotiating Ethical Conduct on the Digital Playing Field. *Communication & Sport*, Vol. 5, No. 2, 2017, pp. 160–185. https://doi.org/10.1177/2167479515600199.

40. Porter, A. K., and M. Schwartz. Ride Report: Mobile App User Guide. *British Journal of Sports Medicine*, No. figure 1, 2017, p. bjsports-2017-098364. https://doi.org/10.1136/bjsports-2017-098364.

41. Mehmood, R., R. Meriton, G. Graham, P. Hennelly, and M. Kumar. Exploring the Influence of Big Data on City Transport Operations: A Markovian Approach. *International Journal of Operations & Production Management*, Vol. 37, No. 1, 2017, pp. 75–104. https://doi.org/10.1108/IJOPM-03-2015-0179.

42. Mondschein, A. Five-Star Transportation: Using Online Activity Reviews to Examine Mode Choice to Non-Work Destinations. *Transportation*, Vol. 42, No. 4, 2015, pp. 707–722. https://doi.org/10.1007/s11116-015-9600-7.

43. Bao, J., P. Liu, H. Yu, and C. Xu. Incorporating Twitter-Based Human Activity Information in Spatial Analysis of Crashes in Urban Areas. *Accident Analysis and Prevention*, Vol. 106, 2017, pp. 358–369. https://doi.org/10.1016/j.aap.2017.06.012.

44. Mcardle, G., and R. Kitchin. Improving the Veracity of Open and Real-Time Urban Data. *Built Environment*, Vol. 42, No. 3, 2016, pp. 457–473. https://doi.org/10.2148/benv.42.3.457.

45. Gschwender, A., M. Munizaga, and C. Simonetti. Using Smart Card and GPS Data for Policy and Planning: The Case of Transantiago. *Research in Transportation Economics*, Vol. 59, 2016, pp. 242–249. https://doi.org/10.1016/j.retrec.2016.05.004.

46. Teddlie, C., and A. Tashakkori. Overview of Contemporary Issues in Mixed Methods Research. In *SAGE Handbook of Mixed Methods in Social & Behavioral Research*, SAGE Publications, Inc., Thousand Oaks, CA, pp. 1–42.

47. Rayle, L., D. Dai, N. Chan, R. Cervero, and S. Shaheen. Just a Better Taxi? A Survey-Based Comparison of Taxis, Transit, and Ridesourcing Services in San Francisco. *Transport Policy*, Vol. 45, No. 2016, 2016, pp. 168–178. https://doi.org/10.1016/j.tranpol.2015.10.004.

48. Supreme Court of the United States. *Texas Dept. of Housing and Community Affairs et Al., v. Inclusive Communities Project, Inc., et Al.* 2015.

49. Crawford, K., K. Miltner, and M. L. Gray. Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication*, Vol. 8, 2014, pp. 1663–1672.

50. Schweitzer, L. A., and N. Afzalan. Four Reasons Why AICP Needs an Open Data Ethic. *Journal of the American Planning Association*, Vol. 83, No. 2, 2017, pp. 161–167. https://doi.org/10.1080/01944363.2017.1290495.

51. Williams, S., A. White, P. Waiganjo, D. Orwa, and J. Klopp. The Digital Matatu Project: Using Cell Phones to Create an Open Source Data for Nairobi's Semi-Formal Bus System.

Journal of Transport Geography, Vol. 49, 2015, pp. 39–51. https://doi.org/10.1016/j.jtrangeo.2015.10.005.

52. Curtin, G. G. Free the Data!: E-Governance for Megaregions. *Public Works Management and Policy*, Vol. 14, No. 3, 2010, pp. 307–326. https://doi.org/10.1177/1087724X09359352.

53. Crayton, T. J., and B. M. Meier. Autonomous Vehicles: Developing a Public Health Research Agenda to Frame the Future of Transportation Policy. *Journal of Transport & Health*, Vol. 6, No. April, 2017, pp. 245–252. https://doi.org/10.1016/j.jth.2017.04.004.

54. Buehler, R. Can Public Transportation Compete with Automated and Connected Cars? *Journal of Public Transportation*, Vol. 21, No. 1, 2018, pp. 7–18. https://doi.org/10.5038/2375-0901.21.1.2.

55. Guerra, E., and E. A. Morris. Cities, Automation, and the Self-Parking Elephant in the Room. *Planning Theory and Practice*, Vol. 9357, 2018, pp. 1–7. https://doi.org/10.1080/14649357.2017.1416776.

56. Fagnant, D. J., and K. M. Kockelman. The Travel and Environmental Implications of Shared Autonomous Vehicles, Using Agent-Based Model Scenarios. *Transportation Research Part C: Emerging Technologies*, Vol. 40, No. 2014, 2014, pp. 1–13. https://doi.org/10.1016/j.trc.2013.12.001.

Appendices

Appendix A: Semi-Structured Interview Guide: Bias in Big Data for Transportation Safety

Noted: bulleted items are optional probes for further questioning, depending on the interviewee's responses.

Interviewer:

Interviewee:

Position of Interviewee:

Place:

Date:

Time of interview: Start End

Optional Introductory Text: **We want the interview to flow as much as possible and for you to feel that you can contribute exactly what you want to the discussion – almost as if we were having a conversation. However, we think it might be worth mentioning a few guidelines prior to starting the discussion.** **As this is an 'interview', we do have some topics that we would like to cover and we will probably use these to guide the discussion. However, please feel free to ask questions yourself and to raise any topics that you think are relevant that we have not mentioned – but do try and stick as much as possible to the theme of biases in big data.** **Do you have any questions before we start?**
Questions: **1. When did you start working with big data in transportation?** • **Related to transportation safety?** • **Does this differ from others in your organization?**

2. Why did your organization decide to use new sources of big data? • **Champion** • **Organizational leaders** • **Planners** • **Public**
3. Has using big data helped improve transportation planning? • **How do decision makers (clients) interpret your big data applications and insights?**
4. Are there ways that the data does not represent the entire population of interest for transportation planning?
5. How do you mitigate the impact of big data not representing the population? • **What skills are required in order to work with big data in this way?**
6. Overall, has using big data has improved planning for transportation safety in your applications? • **Process** • **Geographic scale** • **Outcomes**
7. Is there anything else I haven't asked about that you would like to add?

Appendix B: Interview Coding Instruction

Instructions for Qualitative Coding in the Excel Workbook

1. Go to the first question worksheet "Q1", and read the entire entry for the first interviewee.

2. Review the codes in columns to the right of the text entry, and code a "1" if the code topic is present in the response, and "0" (zero, not O) if the code is not present in the response. The column code will include each coder's name, for later comparisons.

Code Descriptions:

SOURCE - This code indicates that the response includes description of something that causes some type of biased understanding of transportation; it does not represent the full population in some way. SOURCE could relate to the users of Waze more likely being younger than the population, for instance.

MITIGATE - This code represents some way that the respondent mentions dealing with issues of bias in big data. MITIGATE could include surveying Waze users to weight responses by community demographics, for instance.

VEHICLES - This code designates the respondent discussing passenger vehicles; this could include motorcycling, Uber, or electric cars, but not freight or other modes.

TRANSIT - This code includes discussion of buses, streetcar, commuter rail and related passenger transit, but not the use of other modes to get to transit.

BICYCLING - This code is for discussion of bike riding for any trip purpose, including electric-assist bikes or bike sharing.

WALKING - This code is for any pedestrian issues for any trip purpose.

INTERPRETING - This code includes an interviewee describing how they understand big data as a concept or technology.

3. Repeat this process for the Q2 through Q7.

4. Email g-griffin@tti.tamu.edu to notify when you've finished the coding assignment.