Emerging Legal Issues for Transportation Researchers Using Passively Collected Data Sets

AUGUST 2019 | 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. TTI-03-01	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Emerging Legal Issues for Transportation Researchers Using Passively Collected Data Sets		5. Report Date August 2019 6. Performing Organization Code:
7. Author(s) Gretchen Stoeltje Maarit Moran Johanna Zmud Nijm Ramsey Jayson Stibbe		8. Performing Organization Report No. Report TTI-03-01
9. Performing Organization Name and Address: Safe-D National UTC Texas A&M Transportation Institute		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/Project TTI-03-01
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT)		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

With the advent of new technologies to gather and process data, large data sets are being collected that are of interest to transportation researchers. However, legal and ethical questions around data ownership and protection in the context of emerging technologies, especially with regard to emerging automated and connected vehicle technologies, are still being formulated and addressed, but are not settled. This research compares the uses of primary and secondary, passively collected data sources to identify legal considerations affecting access to these data for transportation researchers. With privately sourced data becoming more prevalent, researchers are faced with additional duties and changing practices. This exploratory research aims to provide guidance to transportation researchers on the legal and ethical requirements for data protection.

17. Key Words Data, Ownership, Protection, Technology, Commercial, Data sets, Research, Legal, Human Subjects, Ethical, Responsibility, Right, Duty, Agreement	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 40	22. Price $0

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

Abstract

With the advent of new technologies to gather and process data, large data sets are being collected that are of interest to transportation researchers. However, legal and ethical questions around data ownership and protection in the context of emerging technologies, especially with regard to automated and connected vehicle technologies, are still being formulated and addressed but are not settled. This research compares the uses of primary and secondary, passively collected data sources to identify legal considerations affecting access to these data for transportation researchers. With privately sourced data becoming more prevalent, researchers are faced with additional duties and changing practices. This exploratory research aims to provide guidance to transportation researchers on the legal and ethical requirements for data protection.

Acknowledgements

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.

The project team is very grateful for the direction of Dr. Susan Chrysler, Senior Research Scientist at the Texas A&M Transportation Institute. We also thank Dr. Eva Shipp, Research Scientist at the Texas A&M Transportation Institute, for her review of the report.

Table of Contents

List of Figures

List of Tables

Introduction

Issue

With the advent of new technologies used to gather and process data, large data sets are being collected that are of interest to transportation researchers. However, new legal and ethical questions regarding data ownership and privacy protection have surfaced, especially with automated and connected vehicle technologies. Such questions are still being formulated, and are far from settled. This research examines the legal considerations regarding researcher use of large, passively collected data sets.

Research Objective and Scope

The objective of this exploratory research is to provide guidance to transportation researchers on the legal and ethical requirements for privacy and data protection. The project meets this objective by answering the following two research questions:

- What are the legal and ethical requirements for primary human subjects research?
- What are the implications for those legal and ethical requirements now that researchers may be more heavily dependent on privately sourced, passively collected data?

The inquiry focused on the privacy and consent requirements for actively and passively acquired data. The shift toward working with privately sourced data means that practices to meet requirements may need to change, or that new requirements may arise that may not have been present before. This project addresses both of these conditions.

Structure of Report

The report begins with background information on the rationale for the study, supported by the findings of the literature review. Next, it describes the methods used to identify the current knowledge and state of practice for collecting and using data sets gathered either actively by a researcher or passively by another entity. This section is followed by a summary of those results. The discussion section following interprets the findings, identifies research needs and concerns, and provides use cases illustrating data acquisition paths and the necessary legal agreements for using the data. The report ends with conclusions and recommendations for further research.

Background

This study was motivated by a shift in the way data are gathered, manipulated, and disseminated [1]. Standards and taxonomies for data collection and analysis have rapidly changed over the past two decades as information technology has been integrated into data-gathering and data-management activities. Specifically, university-based researchers are increasingly relying on existing data sources rather than using primary data that they themselves collect for a specific research objective. Primary data give researchers direct access to the raw data and also the direct

responsibility for ensuring that the necessary privacy protections are in place. Existing data, on the other hand, were originally collected by someone else and often for purposes other than research. Existing data can be collected actively (i.e., explicitly asking people for information, preferences, opinions, and behaviors) or passively (i.e., people have little awareness of the data collection effort). Speed and economy are the main advantages of using existing data since collecting primary data can be time- and cost-intensive. The main disadvantage is that researchers have little or no control over what data have been collected and how.

This research focuses on the way this shift affects transportation researchers, who are increasingly using these data sets in their work on human driver behavior. This shift in researchers' access to and use of large, passively collected data sets has raised the significance of two issues: (1) a shift from actively acquired to passively acquired data and (2) lack of clarity around data ownership and privacy interests.

Shift from Actively Acquired to Passively Acquired Data

One of the greatest challenges of transportation human factors research is getting people to agree to participate in data collection activities. The more narrow a researcher's inclusion and exclusion criteria or the bigger the sample size needed, the more challenging this process becomes; hence the appeal of passive data collection. Technologies now exist that make it easier to tap into information about people's behaviors or attitudes, and data can be collected as an incidental by-product of other activities. Technology advances that add to the proliferation of passive data collection include smart cards and scanners that collect a wide range of consumer behaviors; mobile phones that track geographic location; and sensors in vehicles that provide automated situational awareness to drivers [2,3,4]. The combination of advanced passive data collection and increasingly powerful computing analytics has become a critical item in the researcher's toolkit.

This proliferation of passive data collection has led to the emergence of "data aggregators," which are typically private companies that perform automated collection of raw data (Inrix, Air Sage, WAZE, and Google Maps are examples). Data aggregators rely on a variety of sources, including Global Positioning System (GPS) devices, Bluetooth-enabled devices, Wi-Fi-enabled devices, onboard unit transmitters from original automotive manufacturers or third parties such as insurance companies, and smartphone applications. The aggregators then fuse that data with other data, and archive the resulting database for different applications. Often data are leased from data aggregators. Such data are typically timely, detailed, and scalable, but the lessee lacks control over the data [5].

Data aggregators may be either public or private, with subtle differences in their perceptions of data access and use. Private data aggregators generally rely on a chain of legally binding agreements to access source data and to package and resell it to buyers for limited uses and functions. Public agencies who aggregate data often have the perception that data may be collected, owned, and put to use by anyone capable of gathering the data, as long as the data are scrubbed of personally identifiable information (PII) [6].

Lack of Clarity Around Data Ownership and Privacy Interests

Technological advances are enabling new levels of ease for data creation, proliferation, and access. This situation has greatly diminished the clarity of data ownership, which is an important question because ownership of data is tantamount to control. Determining ownership defines who can collect, process, use, and disseminate data [6]. Ownership implies who can profit from what is owned, but it also reflects who is responsible for ensuring privacy protection and access security. Resolving the question of who owns a particular data set is complex [7]. As data are combined from greater numbers of sources, data provenance and ownership become murkier.

A dilemma for researchers in accessing and using existing, passive data is that often the data owners are unfamiliar with study design, research ethical norms, and research methods for studies with human subjects [8]. Guidelines for conducting scientifically sound and ethically acceptable research do not always directly transfer to the context of existing, passively collected data. Collecting data through experiments or surveys typically, but not always, requires academic researchers to obtain approval by an institutional review board (IRB) before carrying out a study that involves human subjects. The IRB is a university-level committee designated to approve, monitor, and review biomedical and behavioral research involving humans. In the United States, any university or body that receives federal funds is required to have an IRB, which is governed by the Common Rule—a rule of ethics regarding biomedical and behavioral research involving human subjects. Yet, data aggregators or other entities that license and sell data are not governed by such guidelines, and as mentioned previously, a disadvantage to using existing, passive data is that the researcher has little or no control over what data have been collected and how.

Many data source owners such as cellular networks and data aggregators de-identify the data using anonymization processes that remove PII as a privacy protection strategy. PII is any data that could potentially identify a specific individual, including any information that could be used to distinguish one person from another or that could be used for de-anonymizing anonymous data [9]. There is no one list of what constitutes PII. A single piece of data can be PII, such as a social security number. Likewise, multiple pieces of data when merged can be PII, even when the individual pieces would not be. Such associations can happen through consumers' smartphones, their use of in-car telematics systems, or some connected vehicle applications. The shift to the access and use of existing passive data has weakened traditional means of protecting individuals' privacy, leading to increasing risks associated with misuse of PII. One famous example of disclosure of PII from de-identified data sets was the release of New York City taxi trip data. Though the data had been scrubbed of vehicle and medallion owners' names, when combined with photographs of celebrities getting into or out of taxis, the trip data revealed not only the location of that individual person at a particular time but also their fare and tip amounts [10]. Given such scenarios, data privacy has the potential to become a challenging issue for university-based researchers who draw on existing, passively collected data. Harmonized policies and practices regarding legal and ethical requirements for data protection need to be identified, documented, and applied.

Method

Literature Review

The literature review covered academic writings on the emerging uses of commercially available data sets, the change in pathways for researchers accessing those and other data sets, and the understanding of data ownership and privacy interests associated with newly available data sets.

Desktop Legal Research

Desktop legal research consisted of a review of the legal and regulatory environments for both human subjects research and private sector data collection. The team reviewed federal regulations mandating privacy protection practices in publicly funded research, federal and state privacy protections, the federal guidelines shaping much privately conducted data collection, the new European law that may affect U.S. corporate and public research practices, and emerging federal legislation that would regulate car-collected data collection. This information was available online.

The research team also researched and reviewed legal agreements articulating the terms of privacy protection and data use between consumers/participants, public researchers, and private data aggregators. Researchers reviewed a total of 23 documents that were available online, in downloaded apps, or from the university researchers who were interviewed and who worked with the documents. These documents include consent and licensing agreements, non-disclosure agreements, commercial terms of use, terms and conditions agreements, and others.

Semi-structured Discussions

The purpose for the semi-structured discussions was to identify the current state of knowledge and practices regarding data protection, as well as emerging questions from the actual or anticipated experience of individuals working with privately and passively collected data sets. Using an interview guide, the project team spoke with 10 public university transportation researchers, one research project manager for an automaker, and one government official. Researchers were chosen based on their experience working with primary, actively collected human subjects or secondary data, collected either commercially or by other public entities or university researchers. Though some had extensive experience working with passively collected, secondary data, most were more familiar with the processes of collecting their own data with IRB oversight. These discussions also suggested the need to develop use cases to demonstrate the various data acquisition models that researchers are now engaged in. The discussion identified changes that researchers noted in their practice related to data acquisition and protection, and what they may need to know going forward.

Results

Legal and Ethical Requirements for Data Acquisition

This section describes the legal protection of privacy and the regulatory framework for human subjects research.

The Legal Environment for Protection of Privacy

Right to Privacy and Its Protections

Privacy is "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [11]. In our current age of information and the flow of that information, the general public has a reasonable expectation that their daily activity generates many forms of information that are stored by others. However, some forms of that information are much more personal than others, such as PII.

Though the majority of Americans believe it is important that their everyday activities remain private [12], the United States has not harmonized its privacy laws nationwide. Instead what exists is a patchwork "system of federal and state laws and regulations that can sometimes overlap, dovetail and contradict one another" [13]. This lack of legal jurisdictional clarity raises questions as to where privacy protections are found, and to whom they apply. And while the U.S. Constitution does not expressly protect the right to privacy, several U.S. Supreme Court cases have narrowly construed the Constitutional meaning of privacy to the right to be free from unwarranted governmental search and seizure [14], the use of contraception [15], the right to an abortion [16], and the right to the sexual conduct of persons of the same sex [17].

Federal and State Regulations

Federal statutes have further expanded an individual's right to privacy beyond the Supreme Court's rulings, but they tend to be narrowly tailored to areas and practices of particular industries. Some examples of this type of federal privacy regulation include requirements pertaining to medical records [18], financial records [19], telemarketing activity [20], pornographic email solicitation [21], online collection of children's data [22], and consumer credit information [23], to name a few. For federally regulated data privacy matters that do not fall under the industry-specific legislation, the Federal Trade Commission Act (FTC Act) is "used to prohibit unfair or deceptive practices involving the collection, use, processing, protection, and disclosure of personal information" [13]. Zmud et al. point out that "in general, FTC enforcement has been mostly procedural, focusing on companies' notice and consent actions, such as ensuring that online companies have privacy policies, that the policies are not hidden in obscure places on company websites, etc." [6]. Though the FTC Act prohibits unfair and deceptive data collection practices, it does not govern the use, collection, and sale of personal information by private companies [24].

Notably, at the state legislative level, California and Massachusetts have taken regulatory steps beyond most other states. California additionally regulates online privacy [25], requires disclosure

of information sharing practices [26], reasonable data security [27], and data breach notifications [28]. Massachusetts has implemented rigorous data security requirements that cover its residents whether they are in or out of the state [29].

Relevant European Law

The European Union (EU) has taken significant steps in regulating data privacy with the implementation of the General Data Protection Regulation (GDPR) [30]. The GDPR replaces the 1995 Data Protection Directive by strengthening individual control over collected data, harmonizing regulation across the EU, and issuing severe penalties for a company's lack of compliance. The GDPR applies to all organizations that collect and process data on EU citizens, regardless of whether or not that organization is located in the EU [31]. This means that U.S. data collecting companies, especially online businesses that collect the personal information of EU subjects, must comply with the GDPR [32]. Thus, many Web-based businesses have brought their practices in all markets into GDPR compliance [33].

The Secretary's Advisory Committee on Human Research Protections of the U.S. Department of Health and Human Services (HHS) has raised concerns about the impact of the GDPR on human subjects research, specifically:

> (i) issues with the GDRP's potential application to U.S.-based research and (ii) two problem areas resulting from EU officials' interpretations of consent as a basis for data processing under the GDPR: (a) the ability to obtain, at the time personal data are collected, consent to future research uses and (b) the need for continued use of personal data to satisfy legal obligations following subjects' withdrawal of consent for the processing of their data, such as those imposed by the U.S. Food and Drug Administration. [34]

The advisory committee urges HHS to coordinate with U.S. officials and their EU counterparts to harmonize cross-border laws that would better facilitate multinational scientific research.

Regulatory Framework for Human Subjects Research

Purpose for Regulating Human Subjects Research

Outrage about the scientific studies performed on human subjects by the Nazis during World War II led to international acceptance of the suggested regulatory framework of the Nuremberg Code and later of the Declaration of Helsinki [35]. The Nuremberg Code is a 10-point list stipulating that participation in research must be voluntary and identifying suggestions for carrying out ethical human subject experimentation [36]. Whereas the Nuremberg Code focused on the responsibility of the individual scientist, the Declaration of Helsinki provided a model of legally enforceable regulation and an independent committee, which led to the development of the IRB [35].

Unethical research practices were not confined to Germany. In the United States, notable severe violations of ethical research principles include the intentional lack of treatment of syphilis patients for 40 years [37], radiation injection experiments [38], purposely infecting mentally disabled children with the hepatitis virus [39], a misleading oral contraceptive study on poor women that

led to many unintended pregnancies [35], and many poorly vetted studies that resulted in a subject's death [35]. As a result of revelations about such unethical experiments, the U.S. Congress enacted the National Research Act of 1974 (The Act) [40]. The Act outlined the primary ethical principles to be used in reviewing human subjects research and led to the Belmont Report in 1979, which "concluded that the primary principles underlying ethical research with human beings are respect for persons, beneficence, and justice. The methods used to recognize these principles are informed consent, risk/benefit analysis, and appropriate selection of patients" [40].

Creation and Governance of Institutional Review Boards

In 1991 the Federal Policy for the Protection of Human Subjects, known as The Common Rule, was published and codified by 15 federal departments and agencies governing federally connected human subjects research [41]. It was updated in July 2018 as the Final Rule. (Note: the research in this report was based on the pre-2018 rule updates and does not reflect potential changes that may address some of the concerns raised here.) The Common Rule introduced the concept of the IRB, a committee formed at research institutions tasked with ensuring that human subjects research protocols meet ethical standards. Each IRB may only approve research that satisfies the conditions under the Common Rule and must oversee the research it approves [42].

> Each IRB shall have at least five members, with varying backgrounds to promote complete and adequate review of research activities commonly conducted by the institution. The IRB shall be sufficiently qualified through the experience and expertise of its members (professional competence), and the diversity of its members, including race, gender, and cultural backgrounds and sensitivity to such issues as community attitudes, to promote respect for its advice and counsel in safeguarding the rights and welfare of human subjects.

Though there are exemptions to IRB oversight, a researcher must submit a research protocol to an IRB so that the IRB may determine if an exemption applies [43]. For example, an IRB may exempt secondary research analysis of data collected under an original broad consent from the subjects if the data have been de-identified.

Current State of Knowledge and Practice

Respondents in the semi-structured conversations described their current research practices and the procedures and protocols developed to protect the privacy of study participants. They also discussed emerging and more complex data acquisition or sharing arrangements that they were engaged in or anticipated engaging in. Regarding this work, they shared questions and challenges emerging from the changing data acquisition models.

Data and Use Methods

Existing Data Types and Collection Methods

The respondents reported using a wide range of data types. Many used video data that they collected themselves either by instrumenting participant vehicles with university-owned equipment, or by running the studies in-house in driving simulators. Many used GPS data to track

location and speed. Several also reported using crash data, health data, hospital data, biometric data, transportation system data (such as signal timing or infrastructure data), vehicle performance data, weather and mapping app data, and eye tracking data.

Most respondents reported that they collected primary data themselves with the oversight of their IRB. Several reported using secondary data acquired from other universities or public entities or from private or commercial sources, and sharing data with other universities. One reported working exclusively with commercially collected data, and one reported exclusively working with aggregated data collected and analyzed by university researchers.

Recent Technological Changes to Data and Uses

Interview respondents have observed a variety of technological changes to data collection methods in recent years, as well as to the data itself. These changes have enriched research while at the same time introducing new technical and ethical challenges. Data sets have increased in size and comprehensiveness, as well as in quality and accuracy. The variety of data sets has also increased, such as improved onboard sensor technology that may enable richer data collection for naturalistic studies. One new development mentioned by respondents is a program that automatically replaces video imagery of faces with avatars.

These improvements mean that more resources are needed to manage, analyze, and store the data. For example, data storage has become cheaper but file size has also increased, necessitating more storage capacity. In addition, some options may be more vulnerable to unauthorized access than earlier options, requiring more stringent protection methods. For instance, higher levels of video image quality provide more accurate data but may make PII, such as the facial identities of unconsented passengers or the license plate numbers of nearby vehicles, easier to discern. This increase in detail requires researchers to spend more resources to de-identify data.

Policies and Procedures for Data and Privacy Protection

Existing Policies and Practices for Working with Primary Data

All researchers noted that policies and procedures for university human subjects research are set and guided by IRBs. That oversight can take the form of requiring and providing data management plans, data storage plans, and informed consent forms, among others. IRBs vet each proposed research project and work with researchers to tailor consent forms and data sharing agreements to each project, as well as interface with the federal regulations and ensure researcher compliance.

Respondents vary their practices for technically protecting data depending on the data type and use. These practices include de-identifying data, encryption involving multiple passcodes, changing passcodes often, using secure servers, coding or "reducing" information from video footage to spreadsheet data for sharing purposes, destruction of original data after an agreed-upon time period, and using portable hard drives requiring thumbprint recognition for access. One very robust program has created a highly secure data enclave that researchers or sponsors may access in person only under very strict conditions.

In some cases it is possible to work with third parties like data warehouses that store, analyze, protect, and de-identify the data. In the health care industry, for example, these entities function under agreements (called business associates agreements) with the data owners to manage the data and have access to identifiers. These data warehouses have the capability to de-identify data so that it can be shared with researchers. By working with these institutions, researchers minimize the risk of themselves working with PII and build trust with the data owners.

Another protection available to some researchers is a Certificate of Confidentiality from the National Institutes of Health (NIH). This is a legally binding document between the NIH, acting with Congressional authority, exempting the research institution from the subpoena power of a party to a lawsuit and releasing it from the legal duty to turn over data in a U.S. legal proceeding.

Several researchers noted that human subjects research funding is often subject to the condition that it be made public or shared in some way. In the past, researchers reported that consent forms would indicate that data would be destroyed after three years. In recent years, however, the terms more commonly state that the data will be kept indefinitely and may be accessed by other researchers. Researchers whose work was so conditioned reported that they had to consider not only the protection of the data while it was in their custody and managed for their own use, but also how it would most safely be prepared and maintained for use by others in the future.

Existing Policies and Practices for Working with Passively Collected, Secondary Data

Interview respondents who have worked with passively collected, commercially sourced data reported that protecting PII was much less of a concern since the data should have already been de-identified, aggregated, and encrypted. In these cases, the complexity of the data set and the difficulty working with it are itself considered protection. Researchers are responsible for protecting the data while it is in their possession, and unless data sharing is allowed by the terms of the use license, they must also not share it with other researchers or put it into the open market in any form. These researchers have also been limited, in the licensing agreements, to use of the data for the specific research question being explored, meaning that researchers may not use data already in their possession to explore research questions not stated in the agreement.

Researchers reported being legally responsible for protecting a business interest, such as a reputational interest or proprietary business information, rather than the PII of a subject. This might mean that researchers may not publish comparisons with other data sets or even conduct them, or publish results that are not favorable to the data sources. One researcher reported that in these arrangements, signing a non-disclosure agreement (NDA) is often required, as well as a data sharing license. In these cases, and in others where one entity's data security is at the mercy of a researcher's integrity and competence, even with legal agreements in place, researchers reported that trust is a very important factor in the relationship with the data owners. As one said, "Just because you sign the agreement doesn't mean you will follow it. People want to feel that you will be a good steward of the data."

Engagement with Legal Requirements

Most human subjects researchers expressed awareness of the legal requirements that govern their work, but generally relied on their IRB to know the law and to act as internal regulators of their practice. Several respondents who have served on an IRB were more familiar with the legal background and reasons for human subjects policies. Nevertheless, all respondents understood that informed consent forms are binding and that they determine who owns the data, how a researcher can use the data, including sharing, for how long, what is done with the data at the end of the study period, and any known risks to participants as a result of participating in the study. They also know that if the data need to be used in some way that was not initially considered in the informed consent form, they may have to go back to the participant to request that use.

Respondents were less clear on data ownership for passively collected and commercially sourced data. Some who work regularly with data aggregators indicated that the aggregators own or have permission to use the data by way of the terms and conditions agreement that end users agree to, and that they (the researchers) typically verify that the data were properly collected by reviewing those agreements. Others indicated that in the case of automated vehicle data, data ownership has not yet been defined and thus pathways to access, ethical responsibilities, and legal liabilities for using the data are also unclear.

The few researchers interviewed who regularly use commercial data reported that there is no internal regulator, comparable to an IRB, to guide them through questions they may have about protecting PII or understanding their use rights in a license with a private entity. They may work with their IRB to determine whether the research is human subjects research, but if it is determined that it is not, then IRB involvement ends. These researchers rely on institutional knowledge passed on by senior project managers and mentors to learn what they can and cannot do under these contracts. Contracts are negotiated by senior researchers with the guidance of the institution's lawyers, but the researchers doing the actual work may never see the contracts. Contract terms tend to be conveyed verbally between researchers and sometimes amended verbally between researcher and data seller or source. While no researcher could recall any dispute regarding data security or legal duty, several indicated that it was very important to have a solid understanding of their legal rights and duties under these contracts. At stake were both the protection of individual privacy interests and the business interest of the commercial or other entity.

Emerging Challenges

Uneven Levels of Familiarity with Ethical Rules

Several respondents have observed an uneven level of understanding among researchers about data protection requirements and legal and ethical duties for human subjects research. One acknowledged that the field was still evolving and that, a decade from now, today's best practices will seem insufficient. Nevertheless, several expressed concern over a lack of awareness about data protection among younger researchers and researchers who come from backgrounds that place less emphasis on privacy protection, notably engineering.

Several also reported differences between IRBs in their support of research projects. One respondent reported that the IRB did not require or offer data management plans for some projects, and that consequently students were using insecure data storage methods like Google Drive or portable hard drives that they carried between the research site and their homes. Another noted that the IRB at their institution seemed unaware of the risk of re-identification of private data when de-identified data sets are combined. One researcher recommended that IRB protocol be improved to include more information on data security, sharing, and collaboration.

Risk of Re-identification

A major concern raised by several researchers is the increased risk of the re-identification of private data when combined with other data sets. As more data sets are created and made available, the combination of two or more de-identified data sets can lead to re-identifying the private data in the individual sets. An example of this scenario given by several researchers was crash data. If de-identified data used in research could be paired with press reporting on crashes, it might be possible to link research findings to specific individuals who are named in public sources. Interview respondents were unsure of their legal liability in such a scenario.

Interviewees were also concerned with their ethical duty to protect subject privacy. Several expressed the need for heightened ethical awareness among researchers when collecting, analyzing, and especially sharing or reporting on data. They called for a very rigorous thought process that would have researchers anticipate whether an event or pieces of information in the data would ever show up in other data sets or information streams that could enable identification of individual people. One cited a process in place at their institution called a Re-identification Risk Assessment that systematizes this thought process.

Respondents also commented that the entry of technology companies into the private automotive space sometimes creates a pressure to move quickly without solid protocols and processes in place for protecting human data. However, one researcher noted that due to changes in European law, data aggregators are introducing stronger protection of PII. Another noted that the risk of re-identification is being addressed by IRBs at the national level to determine what protocols and protections are appropriate to manage this risk.

Data Breaches

Large data breaches raise questions for one researcher about how IRBs will respond, specifically regarding data storage. While cloud storage is a suitable option for the scale of data that some researchers now collect, the concern is whether IRBs will approve cloud storage as being sufficiently secure. Cloud storage also presents additional, inter-territorial jurisdictional concerns regarding a nation's subpoena power. Some respondents said that their cloud storage agreements dictate that research information only be held on U.S.-based servers in order not to cross legal boundaries and raise governance questions or create vulnerabilities to non-U.S. law.

Another respondent, who observed that their research institution had not yet evolved policies and practices for working with large, existing data sets, noted, "At the highest level, I often wonder if eventually big data leaks are going to happen. Are we just trying to fight the inevitable, or will a policy we use eliminate the risk?"

Transparency and Replicability

Working with secondary, interpreted, or summarized data poses challenges to scientific principles for the researchers with extensive experience with these data. The project manager who oversaw university research noted that because they do not have access to the original data, they are unable to "ground truth" or verify the data and must trust in the competence and integrity of the researchers to produce accurate interpretation and analyses. This respondent also noted an emerging variation on this issue about the use of data whose collection method is not disclosed: "If you read something and it is not clear exactly how the data was collected and by whom, or if the work was published without peer review, the result is always suspect." This respondent foresaw more of this type of research as a result of using commercially collected data sets.

Another researcher, who has recently begun working with commercially, passively collected data sets acquired from aggregators voiced similar concerns about the effect of working with data collected by another party whose exact methods are not disclosed or accessible: "Working with these data sets is throwing the scientific principles of transparency and replicability into question."

Use Cases: Five Sample Data Acquisition Arrangements

From the semi-structured conversations, several use cases emerged that describe data acquisition arrangements that researchers are currently engaged in. These use cases trace the pathways, both of the data and the consent, from participant or consumer to researcher. These use cases range in complexity from the most direct relationship between participant and researcher, where the researcher is collecting the data, to the most complex relationship, where the researcher is acquiring data from another research entity, a private data aggregator, or a public agency. The terms of these relationships are expressed in legal agreements that define and describe data ownership, the granting of participant or consumer consent, the use rights of a researcher obtaining the data, the promise to protect the data, the future uses to which the data may be put, and the limits on liability of each party. These legal agreements are briefly described in Table 1 at the end of the section and are discussed in the context of the five different use cases.

Use Case 1: Primary Collector

In the primary collector use case, a researcher directly collects the primary data. The researcher first works with the IRB to determine whether the research is human subjects research. If the IRB determines it is, the researcher and IRB collaborate on drafting an informed consent form that the subject must sign before participating in the study. The researcher may then collect, use, and protect the data according to the terms of the agreement and following the guidelines and protocols developed by the IRB.

Figure 1. Flowchart. Consent and data pathways for primary collector use case.

Use Case 2: Sharing Primary Collector

In the sharing primary collector use case, researchers at public institutions share primary data. The first researcher collects data under an informed consent agreement. The second researcher acquires data, which may or may not be anonymized, from the first through a data sharing agreement. Because both research entities are public, if the shared data are not anonymized, the second researcher's study will also be governed by an IRB. If the data are anonymized, the second researcher's institution may require legal documentation confirming that a full IRB application is not required. It is also possible that the initial data were collected under a Certificate of Confidentiality, protecting the data from compelled disclosure in a U.S. legal proceeding.

Figure 2. Flowchart. Consent and data pathways for sharing primary collector use case.

Use Case 3: Public Agency Sublicensing

In the public agency sublicensing use case, a private data collector acquires data from an end user of a mobile app or other device through a series of agreements embodied in a terms and conditions agreement. The private data collector then shares de-identified data with a government entity through a data licensing agreement which allows for sublicensing to associated public researchers through an additional data sublicensing agreement. The sublicensing researcher may be bound by the terms of the initial data licensing agreement.

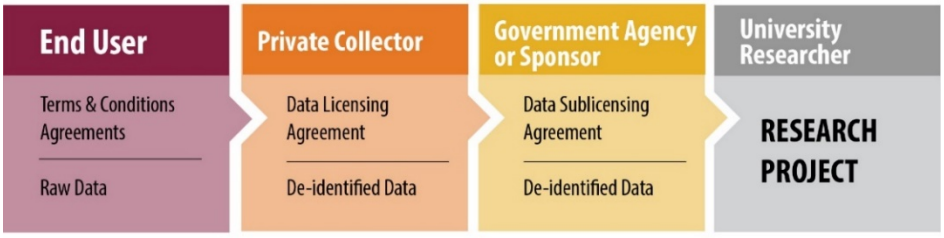

Figure 3. Flowchart. Consent and data pathways for public agency sublicensing use case.

Use Case 4: Sharing Data Aggregator

In the sharing data aggregator use case, researchers acquire passively collected data from a private entity. First, a private data collector, such as a mobile app developer, acquires data from an end user under the terms of one or more digital agreements embodied in a terms and conditions agreement. The private data collector then sells or shares the data with a data aggregator through a data licensing agreement. The data may or may not be de-identified. The data aggregator may then combine data from multiple private and/or public data sources, depending on the researcher's need. That aggregated data are de-identified. The data aggregator then sells or shares that data with the public researcher through a data licensing agreement. The aggregator may also require the researcher to sign an NDA that protects the company's proprietary business information.

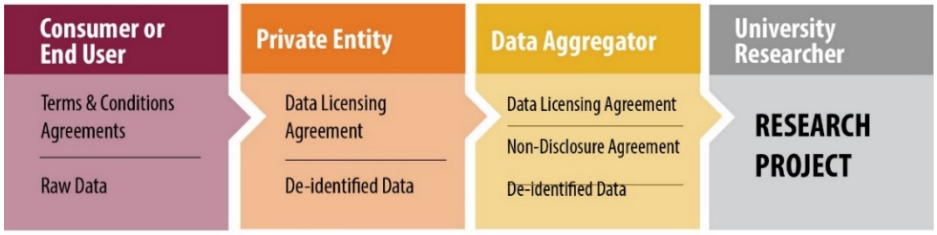

Figure 4. Flowchart. Consent and data pathways for sharing data aggregator use case.

Use Case 5: Data Warehouse Intermediary

In the data warehouse use case, a health center collects data from patients and secures the legal collection and use of patient data through a medical consent form. The health center sends the data to a data warehouse using a business associates agreement. If the data contain information regarded as sensitive by the Health Insurance Portability and Accountability Act (HIPAA), the entirety of the collected data is stored at a data warehouse. The researcher then arranges for use of a data set, in this case a limited data set as defined under federal regulations, using a data use agreement with the health center and the data warehouse. The researcher must also obtain IRB approval and, if the data are not de-identified, obtain a waiver of the informed consent requirement.

Figure 5. Flowchart. Consent and data pathways for data warehouse intermediary use case.

Table 1. Description of Relevant Legal Agreements

Legal Agreements	Description
Informed Consent Form	Agreement between primary data collector and human subject containing comprehensible information to be shared with a potential subject that allows voluntary participation [44]. This information describes the research procedure, purpose, risks, and anticipated benefits, and grants the subject the opportunity to ask questions and to withdraw from the research project at any time [45,46,47].
Business Associates Agreement	Agreement between collectors of personal health information protected by the HIPAA Privacy Rule and other entities handling that information. The agreement stipulates how those other entities will interact with the data [47].
Certificate of Confidentiality	Agreement between research universities and the NIH and other HHS agencies that help researchers protect the privacy of sensitive health-related research. Certificates protect against compulsory U.S. legal demands, such as court orders and subpoenas, for identifying information or identifying characteristics of a research participant [48].
Data Sharing Agreement	Agreement between researchers at public agencies that clarifies what data are being shared and how it can be used. It protects the agency providing the data by ensuring that the data are not misused, and prevents miscommunication between the agencies. This agreement serves to encourage accountability and transparency, enabling researchers to validate one another's findings and discourages duplication of effort in data collection [49].
Data Licensing or Data Use Agreement	Agreement between researcher and data owners (either nonprofit, government, or private) that stipulates how data will be used and what restrictions are placed on that use [50]. The agreement should clearly describe each party's relationship, access, and obligations to the data [51].
Data Sublicensing Agreement	Agreement between researchers and public agencies or sponsors that allows researchers to use data licensed to the public agency [52]. The original license determines the terms of the sub-licensing agreement, clarifies permitted uses and restrictions, and articulates limitations on liability [53].
Terms and Conditions Agreement	Agreement between end user and data collector allowing passive collection of user data. End user permission may be granted from a set of licenses and notices that are either bundled into one agreement or closely situated on an app or company website. Together, these instruments inform the user of a company's privacy policies, stipulate permitted and restricted uses of the company's product (software, for example), and grant the company the right to use the user's data, among other things [54]. These documents include privacy notices [55,56], terms of use or service agreements [57], and end user license agreements [58].
Non-Disclosure Agreement	Agreement between researcher and private data owner or aggregator that secures the researcher's promise to treat specific information as trade secrets and not disclose it to others without authorization [59].

Discussion: Observations About Results

New Legal Duties for Researchers

Researcher rights and duties regarding newer data sets can be very different from those governed by human subjects research regulations and overseen by IRBs. It is important for researchers to be as well-informed about the legal and ethical requirements for the former as they are about the latter.

One fundamental difference is that one of the researcher's primary duties in traditional human subjects research, expressed in legally binding informed consent agreements, is to protect the privacy of the participant. In contrast, when using commercially acquired data sets, the researcher may also be responsible for the protection of the business interest of the private data source or seller. This duty requires protection protocols different from those required for primary, human subjects data collected by the researcher. These duties may be expressed in a use license or NDA, or they may be conveyed verbally by senior researchers, the data sellers, or the sources themselves.

At the moment, researchers who do work with such data have evolved processes and guidelines on an ad hoc basis within their organizations. In one case, senior researchers negotiate the contracts, often in collaboration with institutional lawyers, but researchers themselves may never see the contract. They rely on senior researchers or mentors to pass on best practices and to provide and interpret the terms or the contracts verbally. Those terms may also be amended verbally throughout the course of the project.

IRBs function to establish best practices and internal regulation for researchers who collect primary data. There is no analogous internal regulator to guide researchers working with passively collected, commercially sourced data. This leaves researchers without a systematic approach to data protection or a consistent working knowledge of their legal duties under these contracts.

Third-party Intermediaries

The rise of commercial data collectors, aggregators, and other owners brings with it a new set of interests to protect: corporate proprietary processes and information. With few exceptions, researchers have little official training or support in this area. The data warehouse intermediary use case may present a useful model for engaging a trusted third party who can provide systematic data management protocols and protections. In this arrangement, the data warehouse creates legal relationships with data owners for housing and managing data, and researchers then contract with the data warehouse for licensing the data.

This model evolved from health and medicine but is being adopted for use in transportation. The University of Washington has launched the Transportation Data Collaborative to house sensitive data from both public and private transportation providers. The goal is facilitate public and private data sharing through a neutral, experienced third party who can protect, house, and manage data with policies and protocols that address data ownership, access, use, and related privacy and ethics

interests [60]. Similarly, the National Association of City Transportation Officials has launched Shared Streets, a non-profit digital commons designed to facilitate public-private data sharing while also standardizing those practices and maintaining individual and corporate privacy [61].

These initiatives could bridge the knowledge and process gaps identified by researchers who find themselves facing new legal and ethical duties, as well as new data protection challenges, without regular institutional support. Through standardized protocols and practices, they may also lower the risk of data re-identification. Through contractual agreements, they could also reduce researcher liability for that risk.

Changes to International Law Could Change Consenting Processes

The recent changes in European law could have implications for how consent is granted in the passive data collection process. Currently in the commercial context, customers and users grant consent to use their data when they agree to a terms and conditions agreement. Some may read those terms, but many do not. Generally, users must agree to the terms and, in the United States, few, if any, opportunities are offered to negotiate those terms or control the use of their data.

The EU's newly enacted GDPR cedes significant control over an individual's data to that individual, including how the data may be used and whether they should be destroyed or withdrawn from the corporate holdings. Because of the broad scope of the GDPR, many companies are bringing their practices into compliance with the regulation in all markets. More pointedly, in the United States, HHS is being encouraged by its advisory committee to consider harmonizing its human research regulations with their EU counterparts. In addition, state legislation in California and Massachusetts returns significant control over their data to end users and data subjects.

These legal shifts could create new data collection and protection requirements for private sector entities, and could also provide closer alignment between current human subjects research protocols and passively collected data use and management. For example, processes could be established for obtaining and managing informed consent to future research uses of the data, or for mitigating the legal effect on a research institution of a subject's withdrawal of consent.

Conclusions and Recommendations

Are Current Requirements Meaningful in this Transition?

With the shift toward working with passively collected and commercially controlled data sets, researchers face the challenge of developing practices and protocols without the institutional support they enjoy from IRBs. IRBs are governed by laws designed to protect human subjects data and guided by protection methods that researchers can design and control. Where consent is granted as a part of passively and commercially collected data collection, the IRB processes governing consent likely will not apply, although researchers planning to use passively collected data must still apply to the IRB for review and approval of the project. If the project is deemed not

to be human subjects research, however, the IRB will not stay involved, and researchers are left to develop legal relationships and protection protocols on their own.

Laws around data protection are also changing, both in the United States and elsewhere. New regulations in Europe and the United States guarantee a much greater degree of control over their data to consumers and end users than they previously enjoyed. These guarantees also require that data collectors change processes to accommodate that ceded control. These changes may affect researcher use of data gathered under these new regulations, and may also result in the revision of IRB policies if they are required to be harmonized with the new laws.

What Are the Implications of the Results?

Just as technology has enabled great technical improvements in the practices of transportation researchers, it has also presented new challenges and risks, including ethical and legal ones. When collecting primary data under IRB supervision, researchers are typically engaged in drafting the consent forms that expressly grant consent and use rights, so they tend to understand what participants have granted, and what the data may and may not be used for. They understand that they are bound by ethical and legal duties to protect the privacy of participants.

New data sets arrive with the additional legal duty to protect the private business interests of the commercial data owner. However, most researchers are not trained in what those interests are, how to protect them technically, or how to clearly understand their own legal duties under those contracts. Managing and handling proprietary business data without institutional support or consistent protocols could leave researchers vulnerable to inadvertently breaching contract or damaging relationships.

The increased risk of PII being re-identified may also create another legal question for researchers: might they be liable for pursuing research that could contribute to the risk of re-identification? When collecting data from subjects directly, a researcher has the opportunity to convey the risks of the study to the would-be subject in the informed consent form. This puts the subject on notice of the risk and allows them to assume that risk, thereby eliminating or at least reducing the researcher's liability. When using passively collected data sets collected by someone else, researchers do not have access to the subjects and cannot directly communicate the risk. And although the researcher may be acting within the rights granted by the contract, it is also possible that if the risk of re-identification was foreseeable, the researcher could be liable if it can be shown that the risk could have been reasonably anticipated [62].

Public agencies who are in possession of passively collected data that they share with researchers may also be liable for the risk of any re-identification flowing from the research for which the data were shared.

Suggestions for Further Research

Investigate Possibilities for Increased Internal Guidance

One area for further research would be the possibility of expanding the scope and oversight of the IRB so that it better supports researchers working with passively collected, commercially sourced data sets. Changes to the Common Rule that went into effect after the completion of this research may begin to address researcher use of data sets that may not have previously been considered human subjects research. New considerations of whether these data sets contain identifiable data could place some of this research back into the purview of the IRB, thus theoretically providing institutional support of protocol development and protection of data. Analysis of the implication of these changes would provide further clarity.

Another possibility to explore would be the creation of a separate, internal regulating entity to support researchers in working with these data sets. This entity could be created through a collaboration between the IRB and the Office of General Council so that both scientists and attorneys could work together to develop new protocols that encompass protections of both individual privacy and corporate business interests.

Clarify Researcher Liability for Re-identification Claims

A second recommendation would be to analyze personal and institutional liability for data re-identification claims. Because of the increasing ease with which PII can be re-identified when paired with a separate data set not handled by the researcher, it has become more important for researchers and institutions to understand that risk and develop protocols for assessing and mitigating it in their research practices.

Understand Private Data Owner Interest in Sharing Data with Researchers

A third recommendation would be to query the community of commercial data owners to understand their concerns and interests regarding sharing data for research purposes. Identifying areas of commonality, as well as gaps or conflicting interests, would offer insight into whether and how heartier partnerships could put rich data sets to mutually beneficial use.

Profile and Analyze Emerging New Models for Data Repositories

A final recommendation would be to investigate the efficacy or desirability of trusted data repositories to act as secure intermediaries in brokering public/private data sharing relationships for research purposes. Developing case studies based on existing arrangements and neutral repositories could help to determine the best way to evolve these relationships.

Additional Products

The Education and Workforce Development (EWD) and Technology Transfer (T2) products created as part of this project can be downloaded from the project page on the Safe-D website. The final project data set is located on the Safe-D Collection of the VTTI Dataverse, as described below.

Education and Workforce Development Products

The student(s) working on this project provided an impact statement describing what the project allowed them to learn/do/practice and how it benefited their education. In addition, a PowerPoint presentation was used in an educational setting to introduce students and researchers to the project, and may be used for future presentations.

Technology Transfer Products

The research team conducted a webinar that reviewed exploratory research to identify legal considerations affecting researchers' access to and use of data from both commercial, passively collected sources and human subjects research. The aim was to provide guidance to transportation researchers on the legal and ethical requirements for data protection. This webinar is available on the project page of the Safe-D website.

Data Products

The data set for this project is available on the Safe-D Collection of the VTTI Dataverse. It consists of a summary of findings from guided discussions with researchers and the discussion guide.

Using the discussion guide, the project team spoke with 12 public university transportation researchers, one research project manager who worked for an automaker, and one government official. Researchers were chosen based on their experience working with human subjects or commercially collected data. Though some had extensive experience working with existing data, most were more familiar with the processes of collecting primary data with IRB oversight. These discussions also suggested the need to develop use cases to demonstrate the various data acquisition models that researchers are now engaged in. The discussion identified the changes that researchers have noted in their practice related to data acquisition and protection, and what they may need to know going forward.

Some of the topics covered in the interview guide included:

- Types of transportation data used and collection methods, with a focus on human subjects data;
- Policies and procedures for data and privacy protection;
- Engagement with legal requirements for data collection and use;
- Current and emerging challenges in using transportation data, however owned, collected, or acquired.

References

1. Panel on Developing Science, Technology, and Innovation Indicators for the Future. *Capturing Change in Science, Technology, and Innovation: Improving Indicators to Inform Policy* (R. Litan, A. Wyckoff, and K. Fealing, Eds.). National Academies Press, Washington, DC, 2014.

2. Chen, C., L. Bian, and J. Ma. From Sightings to Activity Locations: How Well Can We Guess the Locations Visited from Mobile Phone Sightings. *Transportation Research Part C*, Vol. 46, No. 10, 2014, pp. 326–337.

3. Chung, E., and M. Kuwahara. Mapping Personal Trip OD from Probe Data." *Int J ITS Res*., Vol. 5, No. 1, 2007, pp. 1–6.

4. Pelletier, M. P., M. Trepanier, and C. Morency. Smart Card Data Use in Public Transit: A Literature Review. *Transportation Research Part C*, Vol. 19, No. 4, 2011, pp. 557–568.

5. Transportation Research Board. *How We Travel: A Sustainable National Program for Travel Data.* National Academies Press, Washington, DC, 2011.

6. Zmud, J., et al. *Data Ownership and Use Issues with Connected Vehicle (CV) Applications*. 2016.

7. Loshin, D. Rule-based Data Quality. *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management*, November 4-9, McLean, VA, USA, 2002, pp. 614–616.

8. Shmueli, G. (2017). Research Dilemmas with Behavioral Big Data. *Big Data*, Vol. 5, No. 2 (2017). Retrieved October 11, 2018, from https://www.liebertpub.com/doi/full/10.1089/big.2016.0043.

9. Jolly, L. Data Protection in the United States: Overview. *Practical Law*. Multi-Jurisdictional Guide 2014/15. Association of Corporate Counsel, 2014. Retrieved October 11, 2018, from http://us.practicallaw.com/6-502-04671.

10. Douriez, M., H. Doraiswamy, J. Freire, and C. T. Silva. Anonymizing NYC Taxi Data: Does It Matter? *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 140–148.

11. Westin, Alan. *Privacy and Freedom*. Atheneum, New York, 1967, p. 7.

12. Madden, M., and L. Rainie. Americans' Attitudes About Privacy, Security and Surveillance. 2015. Retrieved September 5, 2018, from

http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/.

13. Jolly, L. US Privacy and Data Security Law: Overview (Practical Law Practice Note Overview 6-501-4555). Retrieved September 6, 2018 from https://us.practicallaw.thomsonreuters.com/6-501-4555.

14. Katz v. United States, 389 U.S. 347, 357, 88 S. Ct. 507, 514 (1967).

15. Eisenstadt v. Baird, 405 U.S. 438, 446-47, 92 S. Ct. 1029, 1035 (1972).

16. Roe v. Wade, 410 U.S. 113, 152-54, 93 S. Ct. 705, 726-27 (1973).

17. Lawrence v. Texas, 539 U.S. 558, 565-66, 123 S. Ct. 2472, 2477 (2003).

18. Health Insurance Portability and Accountability Act of 1996, 1996 Enacted H.R. 3103, 104 Enacted H.R. 3103, 110 Stat. 1936.

19. Gramm-Leach-Bliley Act, 1999 Enacted S. 900, 106 Enacted S. 900, 113 Stat. 1338.

20. Telephone Consumer Protection Act of 1991, 1991 Enacted S. 1462, 102 Enacted S. 1462, 105 Stat. 2394.

21. Controlling the Assault of Non-Solicited Pornography and Marketing Act of 2003; CAN-SPAM ACT OF 2003, 108 P.L. 187, 117 Stat. 2699.

22. 1998 Enacted H.R. 4328, 105 Enacted H.R. 4328, Part 1 of 3, 112 Stat. 2681.

23. 15 U.S.C.S. § 1681.

24. GAO, *Vehicle Data Privacy*, GAO-17-656, July 2017, https://www.gao.gov/products/GAO-17-656, p. 9.

25. 2003 Cal ALS 829, 2003 Cal AB 68, 2003 Cal Stats. ch. 829.

26. Cal Civ Code § 1798.83.

27. Cal Civ Code § 1798.100.

28. Cal Civ Code § 1798.29, Cal Civ Code § 1798.82.

29. 201 Mass. Code Regs. §§ 17.01-17.05; Jolly, L. US Privacy and Data Security Law: Overview, Practical Law Practice Note Overview 6-501-4555 Retrieved September 6, 2018 from https://us.practicallaw.thomsonreuters.com/6-501-4555.

30. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. The European Parliament and The Council Of The European Union, General Data Protection Regulation, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri =CELEX:32016R0679.

31. EUGDPR.org. GDPR FAQs. Retrieved from https://eugdpr.org/the-regulation/gdpr-faqs/.

32. International Association of Privacy Professionals. Series: Countdown to GDPR: Part 2 – Opportunities and Challenges. Accessed May 23, 2018 from https://iapp.org/news/a/countdown-to-gdpr-part-2-opportunities-and-challenges/.

33. How Europe's New Privacy Law Will Change the Web, and More. *Wired*, Mar. 19, 2018, Accessed Oct. 30, 2018, from https://www.wired.com/story/europes-new-privacy-law-will-change-the-web-and-more/.

34. Department of Health and Human Services. Attachment B - European Union's General Data Protection. Accessed Oct. 30, 2018, from https://www.hhs.gov/ohrp/sachrp-committee/recommendations/attachment-b-implementation-of-the-european-unions-general-data-protection-regulation-and-its-impact-on-human-subjects-research/index.html.

35. Kim, W. O. Institutional Review Board (IRB) and Ethical Issues in Clinical Research. *Korean Journal of Anesthesiology*, Vol. 62, No. 1, 2012, pp. 3–12. http://doi.org/10.4097/kjae.2012.62.1.3

36. Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10, Vol. 2, pp. 181–182. U.S. Government Printing Office, Washington, D.C., 1949.

37. Brandt, A. M. Racism and Research: The Case of the Tuskegee Syphilis Study. *Hastings Center Report*, Vol. 8, No. 6, 1978, pp. 21–29.

38. Rosenburg, H. Informed Consent. *Mother Jones*, September-October, 1981, pp. 21–44.

39. Beecher, H. K. Ethics and Clinical Research. *New England Journal of Medicine*, Vol. 274, 1966, pp. 1354–1360.

40. Department of Health, Education and Welfare, Office of the Secretary. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*.

1979. Accessed Mar. 28, 2019, from https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html#xethical

41. 45 C.F.R. § 46.

42. 45 C.F.R. § 46.107(a).

43. 45 C.F.R. § 46.104.

44. Fischer, B. A. A Summary of Important Documents in the Field of Research Ethics. *Schizophrenia Bulletin*, Vol. 32, No. 1, 2006, pp. 69–80. doi:10.1093/schbul/sbj005

45. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (aka the Belmont Report). 1979. Retrieved from http:// ohsr.od.nih.gov/guidelines/belmont.html.

46. The Importance of Business Associate Agreements (BAAs). *Datica Blog,* Dec. 11, 2014. Accessed 2 Oct. 2018, from https://datica.com/blog/the-importance-of-business-associate-agreements-baas/.

47. 3 Common Misconceptions about Business Associate Agreements. Accessed Oct. 1, 2018, from https://datica.com/blog/3-common-misconceptions-about-business-associate-agreements/.

48. Texas A&M University. Investigator Manual (HRP-103). Rev. Apr. 20, 2018. Accessed Oct. 30, 2018, from https://rcb.tamu.edu/humansubjects/resources/HRP103INVESTIGATOR MANUAL.6.15.2017.pdf.

49. University Research Administration, The University of Chicago. Data-sharing Agreements. Accessed Oct. 3, 2018, from https://ura.uchicago. edu/page/data-sharing-agreements.

50. Office of Research, University of Pittsburgh. Data Use Agreements. Accessed Sep. 27, 2018, from http://www.research. pitt.edu/ccc-data-use-agreements.

51. Database Access Agreement Drafting. LexisNexis Practical. Updated April 5, 2018.

52. State of Oregon Intergovernmental Agreement INRIX Data Transfer, Misc. Contracts and Agreements No. 28134, EXHIBIT C 2.1.

53. Data as IP and Data License Agreements, Practical Law Practice Note 4-532-4243.

54. Bayley, E. The Clicks That Bind: Ways Users "Agree" to Online Terms of Service. Electronic Frontier Foundation. Accessed Oct. 29, 2018, from https://www.eff.org/wp/clicks-bind-ways-users-agree-online-terms-service.

55. Lexis Practice Advisor, Privacy Considerations for Mobile Apps, Mark W. Brennan, Reviewed 10/12/2018.

56. Drafting Privacy Notices, Practical Law Practice Note w-000-9621.

57. Lexis Practice Advisor, Electronic Contracts, Tim Murray, Review on 07/20/2018.

58. Mobile App End User License Agreement (EULA), Practical Law Standard Document 6-525-6926

59. Legal Information Institute, Cornell Law School. Nondisclosure Agreement. Accessed Oct. 29, 2018, from https://www. law.cornell.edu/wex/nondisclosure_agreement.

60. University of Washington Transportation Data Collaborative. What Is the Transportation Data Collaborative? Accessed Oct. 23, 2018, from https://www.uwtdc.org/about/.

61. National Association of City Transportation Officials. Shared Streets. Accessed Oct. 23, 2018, from http://sharedstreets.io/.

62. Tikriti, A. Foreseeability and Proximate Cause in an Injury Case. All Law. Accessed Oct. 26, 2018, from https://www. alllaw.com/articles/nolo/personal-injury/foreseeability-proximate-cause.html.

Appendix A

Researcher Interviews: Findings

The purpose for the semi-structured discussions was to identify the current state of knowledge and practice around data protection, as well as emerging questions from the actual or anticipated experience of working with existing, passively collected data sets. Using an interview guide, the project team spoke with 12 public university transportation researchers, one research project manager who worked for an automaker, and one government official. Researchers were chosen based on their experience working with human subjects or commercially collected data. Though some had extensive experience working with existing data, most were more familiar with the processes of collecting primary data with Institutional Review Board oversight. These discussions also suggested the need to develop use cases to demonstrate the various data acquisition models that researchers are now engaged in. The discussion identified the changes that researchers have noted in their practice related to data acquisition and protection, and what they may need to know going forward.

Some of the topics covered in the interview guide included:

- Types of transportation data used and collection methods, with a focus on human subjects data;
- Policies and procedures for data and privacy protection;
- Engagement with legal requirements for data collection and use;
- Current and emerging challenges in using transportation data, however owned, collected, or acquired.

Researcher Interview Responses

The researchers interviewed 10 university researchers, one non-U.S.-government researcher, and one private sector research project manager. The interview questions (bolded) and answers from each of the participants are provided below.

Types of Transportation Data Used and the Collection Methods, with a Focus on Human Subjects Data

Primary Transportation Data Collected, Developed, and Maintained by Researchers

- Researcher #1: They collect transportation data, such as GPS data that tracks speed and location of participants, video recording data, driver behavior data, and eye-tracking data. Although their organization collects and shares data, they do not typically access human subjects data from either research organizations or commercial/private data sets.
- Researcher #2: They collect human-centered and vehicle data; performance data, which includes driver input; usability data, which is survey data for participants

asking about their experience; workload data, which includes physiological state changes like eye tracking performance measures, sweat response, and heartrate data; weather data. Approximately 90-95 percent of data used is generated internally.

- Researcher #3: They primarily collect data. Anytime they do a "secondary data analysis", where they use data supplied by an aggregator, it is assumed that the sharing of data was approved in the terms of use. They collect vehicle-generated data, event data recorder (EDR) data, naturalistic data, and video data. Most trials are minimal risk. They occasionally use human subjects data collected or owned by other organizations.
- Researcher #4: They collect, develop, and maintain transportation data. Additionally, they have bidirectional data sharing arrangements with third party commercial providers. They collect video data, vehicle-generated data, signal timing data, GPS, bluetooth, and infrastructure data.
- Researcher #5: They collect some human subjects data and use large, existing data sets that are pretty anonymized. They keep the data mostly to themselves, but have shared numbers in excel files with government entities. They collect video data, driver behavior data, eye-tracking data, vehicle simulator data, biological data and hospital records.
- Researcher #6: They have partnered with hospitals or a state DOT to get data, including hospital data. There is always a data use agreement that governs what can and cannot be shared.
- Researcher #7: They perform operational tests that collect video data and vehicle data.
- Researcher #8: They generally collect "driving performance data" using their own, in-house developed data acquisition systems. They instrument vehicles, including video cameras, collecting continuous data, multiple camera views. They collect a lot of PII, such as continuous videos of driver faces and driver hands. The video data can also capture behavior of other drivers, and potentially pedestrians. They also collect GPS and location data, as well as some vehicle network data.
- Researcher #9: Naturalistic Driving data, along with video data, and driver behavior data were all collected. They also have licensed data from other universities and commercial entities if the data was collected with IRB approval. In these cases, they still use a data use agreement.
- Researcher #10: They collect naturalistic data, including audio and video recording in participants' vehicles. They also use aggregated data from public data sets. They access data from private data sets and perform secondary data analysis on the data.
- Researcher #11: They collect data from private sources, including companies, aggregators, and data providers. They have also collected data from publicly owned data owners. A lot of the data is aggregate GPS data. They have also collected toll tag data.
- Researcher #12: They acquire data through an aggregator. The aggregator uses location-based services data—data that is collected from apps such as weather and mapping apps. The data includes trip data from the app users.

Recent Changes in Data Collection and Use Methods

- Researcher #1: The use of Bluetooth is now common for traffic volume and speed measurement. Bluetooth readers are all over Houston pinging cars anonymously. Video data is an increasing source of data with its own issues and risks.
- Researcher #2: They now use multidisciplinary teams that look at new variables and give new information. Vehicles dynamists and psychologist have become additions to teams as data sets grow. Figuring out how to control the expanding data sets is a real challenge. Additionally, many of the engineers have never been exposed to protecting human-based data before. Now, they need a data protection and data monitoring plan.
- Researcher #3: The quantity of data available is greater now.
- Researcher #4: With the emergence of AV/CV technologies, a huge part of the job is just cleaning data. This surprises many people.
- Researcher #5: A big difference with AV/CV data is that the use is widely distributed.
- Researcher #7: The overall data sets have increased in size and comprehensiveness. Now, their studies can include personal information via questionnaires and cognitive function testing, vehicle and car performance data, as well as video data.
- Researcher #8: Sensor technology improvements have made it possible to collect data on-board vehicles. Data storage has become much cheaper. Technology will continue to improve, but privacy concerns will stay. Cloud storage is a great option, but it needs protections. With AV/CV data, there are many more questions. Who owns that data? Is it accessible? Can it be subpoenaed? Who is responsible in a crash?
- Researcher #9: They used new technology in a long study to obtain naturalistic driving data. They are also able to obtain data collected from app developers.
- Researcher #10: With new technologies such as AV/CV, there is sometimes a rush where private companies will move too fast and loose, without solid processes.
- Researcher #11: Due to changes in European laws, the data aggregator has changed the way they provide data, such as requiring a threshold of people to give data on a given link to reduce the risk of personally identifying information being exposed.

Need for Additional Transportation Data

- Researcher #1: There is a need for physiological data including heart rate, skin response, blood sugars levels. Some research is done on blood-alcohol content and drug effects but this is in controlled environments. Also, passenger data and cellphone use data in naturalistic studies would be useful.
- Researcher #2: There is a need for recognition data that can detect stress levels more easily.
- Researcher #4: The problem is with the quantity of data, specifically with more expensive equipment. It is challenging to get enough people to participate in a study.
- Researcher #7: It would be good to have more psycho/social information on participants, but it is often not collected.
- Researcher #8: Data on how many vehicle miles are traveled. They do not know of any system that collect both primary driver and mileage. VMT could be important to understanding crash rates.

- Researcher #11: There is hope that AV/CV technology is going to start pouring data into government data bases in the next five or ten years. Connected infrastructure would help the public sector collect much more data. Drowsy driving monitoring would be useful within the next few years. More data from dockless bike sharing companies would be valuable.

Policies and Procedures for Data and Privacy Protection

General Policies for Data Protection

- Researcher #1: They use subject numbers instead of names, then destroy the link between the subject names and numbers. For demographic information, they work to obscure the exact home/work locations through the use of geocoding.
- Researcher #2: They only recently developed a system, which follows this general pattern: everyone does IRB training at the start of a project, because it is likely that everyone will come in to contact with the data in some way. The principal investigator handles the data analysis and publishing and reminds team members to be careful with the data. There is not a formal plan other than this.
- Researcher #3: As a primarily Human Subjects operation, there is a full-time staff who looks at privacy protection, consent forms, etc. Consent forms outline specifically what is being recorded. Before using recordings for public use, the participant is contacted and signs another waiver indicating their approval of the use. For data storage, it is encrypted with multiple passcodes, offloaded onto a secure server (this is typical in the industry). We want the same security for our participants that hospitals use (HIPAA level protection). We talk to our IRB about any legal issues.
- Researcher #4: They submit an IRB application that includes the protocol, effort, compensation, and the specifics of how data will be handled, stored, protected, and used. For naturalistic data, the endpoints of a trip are removed.
- Researcher #5: The people handling data must be well trained and understand that participants are giving their time. Because they are here to help, it is important to treat participants professionally and take their informed consent seriously. They also do not associate names with data. Data storage is important—it is kept under a double lock and key system. We are extra cautious with our data because it becomes pretty easy to start piecing data together to develop PII.
- Researcher #6: For data collection of farm workers, there were several partners, and so there were many steps. They partnered with a data warehouse company with HIPAA experience. They developed a Business Associates Agreement between the data warehouse and the health center. Then, individual patient visits linked and data was scrubbed of PII. A limited data set was created containing some geographic information (not completely de-identified). IRB approval was obtained for the consortium and other researchers needed to agree to the data use agreements. It is a team process to design a study plan. The IRB process forces researchers to document and think through what they are doing or proposing.

- Researcher #7: All partners are required to do IRB training. They meet several times to work collaboratively on experimental design of the project. The PI is responsible to take care of the human subjects.
- Researcher #8: They implement the federally-required IRB protections. After IRB approval, they have used certificates of confidentiality, given by the NIH to protect data from being subpoenaed. The data that are collected on board their vehicles is encrypted, so even if it was stolen, it could not be unencrypted unless it was back in their lab. There, IRB training is required to access the data, and other protections are put on to avoid the data being shared outside of the "data enclave". Any crash data, where crash details could be found on the internet, would be harder to access. GPS data was considered PII and protected as such, as location, time, and date can be identifying.
- Researcher #9: Due to the scale of the project, they began planning on the IRB side very early, knowing that it would take a while. They wanted to obtain certificates of confidentiality for the participants as they knew this would be important to the IRBs when dealing with crash data. At that time, they drafted consent agreements and recruitment plans. They needed a video and other material that could adequately describe what the participants' role would be. Additionally, they needed parental permission forms for minors. They also needed secondary driver consent forms and a way to identify and eliminate data when it was from a secondary driver that did not consent. Over the course of the study, they added 17 amendments to the consent forms due to adverse events. They ran a pilot IRB test before submitting the full IRB request. Recently, they hired a data security and HIPAA expert to help with data security and privacy protection. They note that as far as IRB protocol goes, many people are behind. Medical researchers do well, in large part because of HIPAA, but in other fields there are many people who are clueless and who have never thought about data rights or data security plans. They recommend changing IRB protocol to be more specific and include more info on data security, sharing, and collaboration while trying to educate people on protocol.
- Researcher #10: They have a full-time staff who look at protection of privacy. They use informed consent forms assessed by the IRB and have internal protections as well. They do not release PII data. Typically, they do not need IRB approval for secondary data analysis. Instead, they operate under the assumption that subjects were properly notified when the data was collected. They typically verify that the data was collected properly. When dealing with private data sets, they look over the Terms of Use for any collected video data. If they want to use the data publicly, like at a conference, they will contact the participant again and get another waiver to ensure their comfort with the new use. To protect data, they encrypt it with multiple passcodes and offload in onto a secure server.
- Researcher #12: They receive de-identified data in a comma delimited file (CSV). They store the data on company hard drives that are protected by the institution's IT protocols. Sometimes the data is stored on cloud storage services.

Engagement with Legal Requirements for Data Collection and Use

Ownership of Data

- Researcher #7: In their collaboration with a university, the university partners owned all data. They received models based on the data, but did not own or have access to the data files.
- Researcher #11: The private data collector/aggregator owns the data.
- Researcher #12: They have a use agreement with the aggregator. The aggregator retains full rights.

Agreements and Contracts

- Researcher #1: They use informed consent forms. Contractors sometimes require proof of IRB approval and some have their own IRB process. One contractor required that the raw data collected be shared with them.
- Researcher #2: They use informed consent forms. They are not familiar with state or federal regulations governing these agreements. They are careful when going to the IRB to know what the sponsor's requirements are for their contracts. In any contract, they push for the ability to publish articles and data.
- Researcher #3: They use informed consent forms assessed during the IRB process. These forms outline the protection of data. They view their informed consent forms to be more of institutional issues and ethical considerations than state or federal issues.
- Researcher #4: They use informed consent forms. While they do not typically grant their organization the permission to distribute or license the collected data to other organizations, their agreements can be made to allow it. They have not done this but recognize that some companies do. Project sponsors will often review the agreements, but they tend to leave it to the IRB to approve without adding additional scrutiny to the process. They have not experienced conflict with sponsor contracts, informed agreements, and state and federal regulations, citing their template with many terms that must be on all forms.
- Researcher #5: They always use informed consent forms. These forms do not allow the organization permission to license and distribute the data to other parties. They acknowledge that there are differences in privacy protection requirements between the ethics boards of different countries. However, they have not experienced conflicts between the different requirements. Despite keeping the data mostly to themselves, they have licensed the data to other institutions in the past.
- Researcher #6: When they have collected data for state DOTs, the DOT owns the data and requires the collecting organization to obtain permission to use the data in certain ways. With existing data, the IRB will ask who owns the data and what the risk level is with a new use.
- Researcher #7: Participants sign consent forms, some of which authorize the use of the data for other purposes. They rely on the project PI to take care of human subjects and IRB.
- Researcher #8: They use informed consents. In these, the participants have to agree to who can use the data and what it is for. Sometimes they get more information with OEM approval, although the interviewee did not have much experience with those

agreements. In those cases, the data they collect is considered proprietary data from the car company.

- Researcher #9: They used informed consent forms and amended them over the course of the study as adverse event occurred. They also used certificates of confidentiality to avoid subpoena. They used parental permission forms to collect data on teenage drivers, and they used secondary driver consent forms to gain consent from secondary drivers.
- Researcher #10: They do use informed consent forms as per the IRB requirements. The consent forms outline specifically what is being recorded. The consent form will tell participants that their data will be handed over to the sponsor when applicable. For some of their secondary data analyses, they will refer to the terms of use, which lays out to the users the future of data use.
- Researcher #11: The users of the app accept the terms of use, often without looking at it. This gives the aggregator ownership of their data. The aggregator sells data to 3rd parties and has its own set of terms and agreements. There is a data licensing agreement that spells out the terms of how the data can be used, who owns the data, who owns derivative data, etc.
- Researcher #12: The consumer downloads the app and accepts the terms of use. There are agreements between the app developers and the data aggregator. The aggregator use very specific contracts. The terms of use are communicated implicitly. They have a team for negotiating and completing such contracts.

Legal Barriers in Human Subjects Research

- Researcher #1: Getting IRB agreement among different IRB organizations can be challenging. Things such as access to cell phone records and traffic violations records is often limited. Medical records are hard to acquire due to HIPAA restrictions. International data and studies can raise legal and ethical issues as there are different standards and requirements in other countries. SHRP2 exposed researchers to a lot of questions about data access and privacy. In a study involving teenage drivers, there was a requirement to share information about related to certain crimes with the subjects' parents. They found that a "certificate of confidentiality" could be used to avoid being subpoenaed. When considering internationally-collected data, there is a question as to what extent US researchers can access and use that data. International laws concerning data use and protection vary.
- Researcher #2: They use certificates of confidentiality to protect young people from being subpoenaed. They have mandated that data should only be shared on a hard drive, while recognizing that even this precaution may not make the data that secure still. They have taken actions such as sending the hard drive with the protected data through the mail. They do not license the data they collect to organizations, but instead make de-identified data public.
- Researcher #3: The legal barriers for human subjects data use have been around for decades without changing much. IRBs are part of a federal program with a bureaucratic system in place. There is a privacy issue when armed forces access and use the data. The data aggregator had a huge PR problem over this and has changed

some of its settings in reaction to this, almost simultaneous to the new GDPR regulations.

- Researcher #4: One big legal barrier for dealing with human subjects data has been a discoverability issue. For example, If commercial fleets are monitored and then they can see that they have a distracted driver, are they now ratifying such behavior by not addressing it? Could the data be used to prove they did not do enough? Participants absolutely do not want the data to be used against them. Another issues is of accessibility. When PII is reviewed, it must be done under high security in a "Data enclave." This is a highly controlled environment where the researchers must physically be present. Another barrier is that when a private company collects useful data, particularly when their goal is not primarily to sell it, it is difficult to figure out a procedure that does allow data-sharing. They have to check their data agreements to see what they can do and if they are allowed to share it.
- Researcher #5: A legal barrier facing researchers is the issue of incidental findings, specifically regarding the duty of researchers to report incidental findings. It has been more challenging in the US to collect audio recording data in addition to video recording. Also, the process is different for working with minors. Legal barriers for human subjects data use may not have changed much in recent years, but there is not heightened awareness over it. GDPR in the EU is one example of change. Participants are becoming more aware of the value of data and how easy it is to share it, so they are sometimes more cautious.
- Researcher #7: It used to be easier to collect both video and audio data from inside the cabin. Now, audio data is more difficult to collect because of the issue of passengers who have not given consent. There is so much data and so many data sources, it is important to figure out how, and by who, a data set can be used. Are researchers able to go back to participants? Would that be added to the consent form? These are difficult questions that need more discussion.
- Researcher #8: The rules for data access are changing. There is a push toward not needing IRB approval for data that already exists.
- Researcher #9: International collaborations are difficult with naturalistic driving studies. For example, the Certificates of Confidentiality given to participants are governed only by US law. They do not cover data collection efforts occurring outside of the US and will not be upheld outside of US jurisdiction. IRB is adding data security plans to the required data management plans.
- Researcher #10: They installed camera units in the vehicles of 16 year olds. As an organization, they were among the first to develop IRB-procedures, or trigger-points for when to involve parents and the police over illegal behavior on the recordings. Even after signing the acknowledgement forms and understanding how the procedures works, participants would still engage in illegal activities. They note that there are very few legal scholars in this area. Typically, all regulations governing agreements are institutional, resulting in ethical issues rather than legal issues.
- Researcher #11: Researchers need to know what to ask for when reading licensing agreements.
- Researcher #12: When purchasing data from any collector or aggregator, they maintain significant control and ownership of data. There may be significant restriction on what researchers can do with the data. In some cases, they may not

want researchers to compare the data to other data sets. In other cases, if the results of their data do not show positive results, they may not want researchers to show that. It is necessary for researchers to understand what they are buying and what limitations they will have. They had to get IRB clearance to study a billion persons due to the size of the aggregated data set, although this was a fairly lenient process due to the nature of the data being collected and the fact that they were not working directly with the participants.

Current and Emerging Challenges in Using Transportation Data, However Owned, Collected, or Acquired

Common Challenges in Human Subjects' Transportation Data

- Researcher #1: They recognize that data is messy. It needs to be cleaned before it can be used. There are growing privacy concerns with GPS data and the location information it can give. The IRB does not always understand the privacy concerns with new data.
- Researcher #2: For data collection, it is a challenge to keep the data clean and know what is in the data. For privacy and data security, finding an approach to keeping the raw data set secure can be challenging, especially when multiple stakeholders need to access the data.
- Researcher #3: They have not had any issue with data access. They have found a bigger hurdle to be interpreting heavily encrypted data from non-traditional car and computer companies. They have had to back off minute-level detail because of privacy concerns.
- Researcher #4: They are often asked to make data available for many purposes but those purposes often fall outside the scope of the informed agreements. They have to scrub the data of PII to do this, which can limit its value for customers. Some people can also be overwhelmed by the size of the data sets provided.
- Researcher #5: The question of a "duty to report" adverse behavior is a challenge; partnering with other organizations [and sharing data] can be complex.
- Researcher #6: Partnerships and building trust is difficult. Even with tech to process data, a lot of times it comes down to having good relationships with the companies housing the data—they even sometimes give discounts. It can be a tricky task to remove data identifiers sufficiently while maintaining a useful data set.
- Researcher #7: Cost is an important challenge. It is expensive to create or buy data. Quality is another. It is hard to know what the ground truth of data is. It is hard to verify that the data is accurate, especially if you did not collect it yourself. Data sets are so large, that a lot of times you have to hope that they give you what you need. It is very difficult to anonymize video data. This job is the responsibility of the data collectors. The ability to transfer data via the cloud is huge, but it does not always work with large data sets.
- Researcher #8: The biggest challenge they commonly face is getting the resources necessary to answer questions. Not many funders are interested in some of the answers they think they can provide with their data set.

- Researcher #9: It was a balancing act to ensure that data could be shared. They needed to use strong language while being vague enough to allow for flexibility of data use, especially as technology changes.
- Researcher #10: They worry that some smaller consulting companies without proper protocols in place will wrongly share data. To combat this, they went overboard in designing procedures. IRB has all of the relevant information on their website and are pretty standard for all research institutes.
- Researcher #11: In general, the trade-off between specificity of data and privacy protection is challenging. Even with aggregate trip data, the movements may be so precise so as to reveal too much context.
- Researcher #12: Although the data was de-identified, with the right tools, there is a possibility that it could be re-identified.

Technological Barriers in Human Subjects' Transportation Data

- Researcher #1: File size, data handling and data storage are issues that researchers have faced in the past. In recent years, as data acquisition gets easier, data gets bigger. Where do you store it? Some agencies have had to make very big storage facilities.
- Researcher #2: There are not technological tools that have created a greater access to data. Many times teams do not have protocols or policies in place either. Google Drive and One Drive are only password protected, so they are not a good place to maintain data or PII (although researchers and students will still sometimes use this). One project used Subject Book, where you input raw data and give access level to users. It allows anyone to see the aggregate data but limits the PII data to those with access.
- Researcher #3: Some agencies participate in "Goaltending", which makes accessing data financially prohibitive. So in addition to the technological encryption, there is a financial barrier to accessing data.
- Researcher #4: It has been hard dealing with video data when there are unconsented passengers in a vehicle. They have to be scrubbed out. Is it even okay to collect this data in the first place? Can it be collected and then thrown away? With video recording data, there are even issues with vehicles on the outside of the subject car if license plates are visible on the recordings. To really remove PII you have to go through great lengths. The technological barriers have absolutely changed in the recent years. Video data quality has gone way up, which has made data collection easier and more affordable. This has led to much bigger data sets. One tool that has made it easier to use data without threat of compromising subject anonymity is the automated process where faces are replaced with avatars, anonymizing the participant faces. When using GPS data, it is possible now to convert from absolute to off-set GPS. This removes the real-world, identifying background.
- Researcher #5: In many ways technological barriers in the collection of human subjects' data have gotten a lot easier. Both the ease of data collection and its accuracy have improved. For example, eye-tracking, vehicle outfitting, and body temperature testing are all easier. Recently, avatars used to mask people's faces have

improved anonymity for video recording data. There is still more to do on this front—clothing, tattoos, etc. can be PII as well.

- Researcher #6: Improvements in technology have helped researchers develop better data safety protocols. This can also help researchers safely access data in more locations.
- Researcher #8: Vehicle sensor data was not as reliable in the past as it is today. It continues to get more efficient. Storage was more costly in the past as well.
- Researcher #9: Collecting data from web apps can be problematic. They require that the privacy agreement is included in the terms of service for the app, and cannot use the material otherwise. IRBs are becoming savvier about apps that are used to collect data.
- Researcher #10: One of the bigger challenges is when non-traditional car and computer companies are doing driving research and their data is so encrypted no one else can access or interpret it. This can also mean that when partnering with such a company, they only let you access the data that they want, or the data that paints them in a positive light.

Additional Comments:

- Researcher #1: In engineering fields, the level of ignorance about the IRB and risk management can be astounding.
- Researcher #2: It takes a lot of effort to separate data from the person, but it is necessary. With increasing team size, there is risk that data will get out eventually. Engineers are not always sensitive to privacy, so data may not be safeguarded well. They often want to know more information on demographics which leads to more identifying information. They also will give their data to graduate students who do not thoroughly understand privacy protection. The lack of a data management plan or sloppy handling of data can lead to it getting out. Can policies eliminate the risk or are big data leaks inevitable?
- Researcher #3: Working with private companies can be difficult. Often, they will not share data with public entities. Newer car companies sometimes do not cooperate with more traditional car companies, breaking traditions and controlling whatever information they can.
- Researcher #5: They are amazed at people's lack of concern over data privacy with the emergence of CV/AV technology.
- Researcher #6: Private companies are able to do interesting things without following IRB guidelines. They can essentially do human subjects research without oversight or monitoring. Although data protection is extremely important, data is never really fully protected. Someone who knows what they are doing could find a way to get the data. It is important to think about what they could do with the data once they get it. The researcher should be honest with themselves and with the IRB about what could happen with a data breach. In working with engineers, they found that many just did not know they had to go through the IRB.
- Researcher #7: There are times when a researcher can read something and not know how the data was collected. This should raise suspicion. There are so many more data

sources and so much more data that it will likely only get less clear. There are many important issues in data use to consider, such as how long a database can be and by whom, as well as what future uses are permissible and whether the participants can be contacted again at a later date.

- Researcher #9: A lot of people in university departments are behind as far as data protection requirements go. Medical researchers do fairly well because of HIPAA, but other researchers may not think about data rights or about who has the right to share data.
- Researcher #11: Although much more AV/CV data is being collected, much of this data will be collected through the private sector and sold for profit. For example, they will be able to record engine diagnostics—some companies what this information.

Appendix B

Researcher Interviews: Interview Guides

Questions on Data Collection, Usage, and Ownership

1. Does your organization directly collect, develop, or maintain any transportation-related data?
2. Which of these types of data does your organization use or collect?
   a. Crash records
   b. Vehicle-generated data
   c. Video data from inside or outside a vehicle
   d. Infrastructure data
   e. GPS/Bluetooth/wireless device data
   f. Driver behavior data measuring driver input to steering, brake, or throttle
   g. Hospital records
   h. Insurance records
   i. Other? Please list any other sources of data collected related to transportation?

   _____

3. Does your organization collect or has it collected human subjects data?
   a. If yes, what is the nature of the data?
      i. *Probe on type: biometric? Geographic? Personal Preference? Other?*
      ii. *Probe on how the data is collected (digitally? electronically? Photographically? Sonically?)*
   b. If no, why not?
4. Does your organization use human subjects data sets that are existing and/or owned by other organizations?
   a. If yes, who or what are typical owners of existing human subjects data sets?
   b. If yes, what kind of data sets do they own (data type, collection method, etc.)
5. Do you have a need for transportation-related data that are not currently being collected?
   a. If yes, what types and for what purpose?
6. Do research organizations typically require a use license agreement to use their data?
   a. If yes, do they charge for the license?
      i. If yes, how much might they charge for the license?

Questions on Policies and Procedures for Data and Privacy Protection

7. In relation to human subjects data, what policies and procedures are in place to protect subjects privacy?

Questions on Engagement with Legal Requirements for Data Collection and Use

8. Have your organization's data collection or use procedures changed with the emergence of automated and connected vehicle technologies?
9. Who or what are typical owners of commercial data sets of private data?
10. What kind of data sets do they own?
11. By what technical means have some of these owners collected data from private individuals?
12. What ownership interests have those data set owners typically asserted over the data they collect from private individuals?
13. What legally binding instrument grants that ownership right to those corporate or natural persons that claim ownership?
    a. Can you share examples of end user license agreements or other legal instruments granting these ownership interests?
14. Does the organization use informed consent agreements?
    a. If yes, can you provide a template or copy of the language?
    b. Do these agreements typically grant your organization the right to distribute or license the data to other parties?
15. Are there state or federal regulations that govern your organization's agreements with the individual participating in the research?
16. How do sponsor contracts influence those agreements?
17. Are these documents, regulations and practices ever in conflict with one another?
18. In collecting new or accessing existing human subjects data for research purposes, what legal barriers have restricted or blocked access in the past?
19. Have those legal barriers changed in recent years?
    a. [Probe] Are there new legal terms that can be included in informed consent agreements that allow greater access to data while still protecting the privacy of the human subjects?
    b. [Probe] Are there new legal terms that can be included in use license agreements that allow greater access to data while still protecting the privacy of the human subjects?
20. Are there strategies, techniques or tools that have been used to overcome these barriers in the past?
21. Does your organization license the data it collects to other organizations or institutions?
    a. If yes, does it require parties to enter into a use license agreement?
        i. If yes, how much does it charge for the license?
22. What privacy rights have individuals either retained or waived regarding their data?
23. Do private entities require other entities to enter into use license agreements in order to use their data?
    a. Can you share examples of a use license agreement or other legal instruments granting these use rights?
24. Are there newly available data or new collection methods that necessitate new terms to define those ownership interests?

25. Are there newly available data or new collection methods that necessitate new terms to define those use rights?

Questions on Current and Emerging Challenges in Using Transportation Data

26. What challenges are most frequently reported/communicated by your staff who use transportation data?
27. In the collection of new human subjects data for research purposes, what technological barriers have restricted access to that data in the past?
    a. Probe: Do any of these barriers specifically restrict efforts to de-anonymize or aggregate the data?
28. Have technological barriers changed in recent years?
29. Are there new technological tools that have created or could create greater access to data?
    a. [Probe] Are there new tools for anonymizing or aggregating data that make it easier to use data without threat of compromising subject anonymity?