Workshop Report:

Paving the Road to Future Automotive Research Datasets: Challenges and Opportunities

February 2022

Draft report for the workshop held virtually on November 18-19, 2021

Table of Contents

Preface

This report summarizes the presentations, discussions, and key findings from the workshop, “Paving the Road to Future Automotive Research Datasets: Challenges and Opportunities,” held virtually November 18-19, 2021. Workshop participants attended virtually via Zoom. 69 people from 31 organizations across academia, industry, and government in the United States as well as from South Korea, the United Kingdom, and Sweden attended the workshop.

The workshop served as a forum for learning and understanding the wide variety of automotive research datasets available to support automotive cybersecurity and other research, as well as the broad range of applications that could benefit from research and development supported by such datasets. The workshop benefited from the degree of openness and interaction displayed by the participants while discussing the challenges and opportunities of automotive datasets they produce or use in their research. By openly sharing their experiences and knowledge, insights were gained which are documented in this report and which should provide value to all the participants and their respective organizations. Furthermore, the workshop was the first in what will hopefully be an ongoing series of discussions and interactions and the start of a new community centered around the production and use of automotive research datasets.

Workshop Organizers

David Balenson
Senior Computer Scientist, Computer Science Laboratory, SRI International
david.balenson@sri.com

Christos Papadopoulos, Ph.D.
Professor and Sparks Family Chair of Excellence in Global Research Leadership, Computer Science Department, University of Memphis
christos.papadopoulos@memphis.edu

Glenn Atkinson, P. Eng.
Vice President, Product Safety, Geotab, Inc.
glennatkinson@geotab.com

Ted Guild
Connectivity Standards Lead, Geotab, Inc.
tedguild@geotab.com

Stacy J. Prowell, Ph.D.
Chief Cyber Security Research Scientist, Computational Sciences and Engineering Division, Oak Ridge National Laboratory
prowellsj@ornl.gov

Samuel C. Hollifield
Cybersecurity Researcher, Cyber Resilience and Intelligence Division, Oak Ridge National Laboratory
hollifieldsc@ornl.gov

Executive Summary

A two-day workshop entitled “Paving the Road to Future Automotive Research Datasets: Challenges and Opportunities,” was held virtually on November 18-19, 2021. This report summarizes the presentations, discussions, and key findings from the workshop.

As vehicles are becoming more connected and autonomous, telematics and other data from such vehicles is critical to support research and building applications for the vehicles themselves as well as their environment. Such datasets are scarce and limited at best, partly due to the difficulty in collecting them and the privacy considerations that accompany them. Unlocking the vast potential of vehicle applications requires open availability of diverse, high-quality datasets.

The overall goal of the workshop was to initiate a coordinated effort to bring together a community around the development and sharing of robust automotive datasets to foster and support new, open research in areas with strong societal impact such as smart and connected communities and development of cybersecurity and privacy protections for automotive applications. Over the course of the two-days, participants not only learned about current automotive research datasets available to support automotive cybersecurity, but also explored future needs for automotive datasets as well as applications that could benefit from research and development supported by such datasets.

A broad group of 69 scientists, engineers, and technologists from 31 organizations across academia, industry, and government gathered to discuss their experiences and interests in producing and using automotive datasets in their research. Participants included researchers producing and/or using automotive datasets in their research, commercial vehicle telematics providers who are willing to share data with researchers, other industry representatives interested in collaborating with researchers to support and benefit from their work, and leaders from standards organizations interested in developing common interfaces and data formats needed to support automotive datasets and their application. There were also representatives from funding agencies and other government organizations interested in a robust research ecosystem to develop new, innovative automotive applications that benefit society.

Presentations highlighted research and other efforts to produce automotive datasets and/or use datasets in support of the research; commercial vehicle telematics providers willing to share data with researchers; and standardization efforts. A virtual tour of the National Transportation Research Center at Oak Ridge National Laboratory highlighted a number of key facilities that have been instrumental in their transportation technology research. The working sessions identified future dataset needs, including desired characteristics, dataset quality and trustworthiness, and potential sources of datasets; and potential applications of automotive datasets, including analytic tools, sharing infrastructure and frameworks, privacy considerations, and industry participation.

The workshop produced key findings regarding the critical need for data to drive research, applications with high societal impact, the importance of working with industry, and the benefits of working as a community.

While there is a critical and increasing need for automotive data, the current lack of data is a serious impediment to future research. Examples include in-vehicle data for cybersecurity, extra-vehicle data for intelligent transportation and smart communities, data for AI/ML research, EV data to design and develop the EV ecosystem, and safety and convenience data to benefit vehicle occupants. To improve data utility, the community needs standardization, tools, and best practices in data collection, data handling, cybersecurity, and privacy.

Cybersecurity, advanced driver-assistance systems, and safety were identified as applications of high importance. Forensics, comfort, and convenience applications were also identified. Secure V2V and V2I communication is the bedrock of extra-vehicle applications such as intelligent transportation and smart communities. Emerging personal vehicle and fleet applications such as vehicle repair and maintenance, cost comparisons between EVs and internal combustion engine vehicles, and new insurance models are becoming pervasive. Along with transportation, applications such as a vehicle AirBnB, mapping road conditions, and hyperlocal weather are new enablers to many smart and connected community applications.

The nature of vehicle and transportation research requires working with industry (OEMs, telematics providers, Tier 1 suppliers, insurance, etc.) to create collaborations that allow research and data to flow each way. To respect privacy and intellectual property (IP) we need collaboration blueprints, agreements, and tools.

Establishing a community is beneficial at many levels to foster connected vehicle and intelligent transportation research. Community activities should be supported by a collaboration platform to act as a catalyst and establish a data ecosystem to bring together providers and consumers, identify common data needs, requirements, and tools, develop common data naming formats, and address privacy and IP concerns.

The robust workshop participation, wide-ranging presentations, and productive working session discussions demonstrated a strong need for a coordinated effort to bring together a community around automotive datasets. To help support such a community, several workshop organizers plan to stand up an open collaboration platform at the University of Memphis. The platform will provide a central catalog and clearinghouse for current and anticipated future automotive research datasets, as well as provide other services that will help catalyze our efforts and provide a means for community engagement and interaction.

The workshop was the first in what will hopefully be an ongoing series of discussions and interactions and the start of a new, vibrant community centered around the production and use of automotive research datasets. We plan to hold additional workshops, likely on an annual basis. We will target holding the next workshop in November 2022 and hope to learn about new datasets, facilities, research initiatives, and applications.

Introduction

As vehicles are becoming more connected and autonomous, telematics and other data from such vehicles is critical to support research and building applications for the vehicles themselves as well as their environment. Such datasets are scarce and limited at best, partly due to the difficulty in collecting them and the privacy considerations that accompany them. Unlocking the vast potential of vehicle applications requires open availability of diverse, high-quality datasets.

A number of researchers at universities and laboratories around the world have collected data from cars and trucks in order to support their research into cybersecurity and other applications. Many researchers have shared their datasets and made them openly available to other researchers to use in their own research. Some known examples of openly available datasets include:

- Korea University Hacking and Countermeasures Research Lab (HCRL) Datasets: several automotive datasets generated by HCRL between 2014-2019 using real commercial cars, including a car-hacking dataset, and CAN intrusion dataset (OTIDS), a survival analysis dataset for automobile IDS, and an automotive ethernet intrusion dataset [HCRL-Datasets].
- Cephas Baretto Dataset: dataset collected from the OBD-II port of an automobile by a Portuguese student to develop a machine learning model for his Masters research project in 2018 [Baretto, Baretto2018].
- TU Eindhoven Automotive CAN Bus Intrusion Dataset v2: data collected by researchers at TU Eindhoven in 2019 from two cars and a self-built CAN bus prototype to evaluate a network IDS [Guillaume2019, TUEindhoven].
- CrySyS Lab CAN-Log Infector and Ambient CAN Traces: CAN messages and GPS location data collected from a vehicle by researchers with the Cryptography and System Security (CrySyS Lab) at the Budapest University of Technology in 2020 to support research into intrusion detection, driver identification, location tracking [CrySyS].
- ETAS/Bosch SynCAN Dataset (Synthetic CAN Bus Data): synthetic signal-level data created by researchers at ETAS/Bosch in the 2020 timeframe as supplementary material for their CANet intrusion detection system [Hanselm2020, SynCAN].
- Real ORNL Automotive Dynamometer (ROAD) CAN Intrusion Dataset: anonymized CAN data recorded by ORNL in 2020 via OBD-II port of a vehicle to support anomaly detection and IDS research [Verma2020, ORNL-ROAD].
- Colorado State University Heavy Vehicle CAN Data: CAN and J1939 data collected from various trucks and operations in support of research into heavy vehicle event data recording and analysis of hard braking events and crash forensics [Daily].

One of the underlying objectives of the workshop was to learn about these and other related datasets and to explore what future automotive datasets are needed to support additional research.

Additionally, industry may make vehicle data available. For example, Geotab Inc., a company in Canada, is a leading provider of telematics units for automotive fleets, including trucks, cars, electric vehicles, and even off road equipment [Geotab]. Their technology connects commercial vehicles to the Internet and provides web-based analytics for fleet management solutions. The units capture a vehicle's location, speed, accelerometer data (anytime the vehicle undergoes acceleration forward/backward, side-to-side, or up and down), and other data from a vehicle's computer including seat belt usage, detailed engine diagnostics, and more. Geotab collects this data and applies data analytics and machine learning to help customers improve productivity, optimize fleets through the reduction of fuel consumption, enhance driver safety, and achieve regulatory compliance. Geotab provides customers and others free open access to data aggregated from hundreds of thousands of vehicles through their telematics devices worldwide via the Geotab Ignition platform [Geotab-Ignition]. The data is categorized in relation to urban infrastructure, weather, and location analytics as a demonstration of the powerful insights possible. They have leveraged this data in creating insights to help responses to natural disasters and other greater good uses. The datasets are fully anonymized to prevent disclosure of personal, customer, or driver information. Another objective of the workshop was to learn about Geotab's telematics devices and their use to support a variety of current and potential future applications.

The U.S. government also makes transportation data available via online portals. For example, the U.S. Department of Transportation (U.S. DOT) Public Data Portal [USDOT-Portal] makes an extensive catalog of data available related to railroads, roadways and bridges, pipelines and HAZMAT, trucking and motorcoaches, aviation, public transportation, automobiles, maritime and waterways, research and statistics, and bicycles & pedestrians. The automobile data covers a diverse set of topics such as affected population / repaired air bags over time from airbag recalls, detailed vehicle trajectory data (vehicle lane positions and locations relative to other vehicles) collected on U.S. highways to support simulation programs, freeway car-following behavior (such as velocity, acceleration, and relative position) for the car-following instances observed on actual roads to develop micro simulation models for improved work zone planning, data collected during safety pilot model deployment including basic safety messages (BSM), vehicle trajectories, various driver-vehicle interactions data, contextual data, and much more.

The USDOT Intelligent Transportation Systems (ITS) Connected Vehicle Pilot Deployment Program integrates connected vehicle research concepts into practical and effective elements to enhance existing operational capabilities [USDOT-CVP]. Data were collected throughout each pilot to facilitate independent evaluations of the use of connected vehicle technology on real roadways. USDOT makes data from these pilots publicly available to encourage additional study and reuse of the data [USDOT-Connected]. The pilots include the Wyoming DOT (WYDOT) Pilot [USDOT-WYDOT], Tampa-Hillsborough Expressway Authority (THEA) Pilot [USDOT-THEA,THEA], and the New York City DOT (NYCDOT) Pilot [USDOT-NYCDOT]. Each pilot site makes sanitized and anonymized data from these projects available to the public along with various tools for interacting with these data. The data includes Basic Safety Messages

(BSM), Traveler Information Messages (TIM), Signal Phase and Timing (SPaT), and Event Logs (EVENT).

There is a vast array of potential applications of high-quality automotive datasets. An important area is vehicle safety and cybersecurity, including CAN bus anomaly detection and intrusion detection, sensor security, AI security, and multi-sensor fusion. Other vehicle-oriented applications include safety and security of Advanced Driver Assist Systems (ADAS), Connected and autonomous vehicles (CAVs), Electric vehicles (EVs), heavy trucks, as well as system monitoring and optimization, in-vehicle infotainment, predictive maintenance, and route and trip planning. Transportation and fleet management applications include passenger safety, traffic management, ride sharing, multi-modal mobility, data-driven insurance. Smart city and community applications include infrastructure monitoring and management, weather sensing and mapping, and asset management, among others. There are also many other potential future applications of automotive datasets.

Geotab's telematics devices and data analytics support fleet management applications across a number of areas [Geotab]:

- Productivity: driver tracking, asset management and tracking, routing and dispatching, fleet management reports
- Optimization: keyless entry, fleet fuel management, fleet maintenance, fleet benchmarking
- Safety: driver safety reporting, driver coaching, dash cams
- Sustainability: EV fleet management, EV suitability assessment, EV battery degradation tool, temperature tool for EV range
- Compliance: DOT compliance (ELD), compliance management - Driver-Vehicle Inspection Report (DVIR), International Fuel Tax Agreement (IFTA)
- Expandability: software integration, hardware integration.

The National Science Foundation (NSF) conducts a number of research programs that support the development of new technologies and solutions involving automotive and transportation applications, including the Secure and Trustworthy Cyberspace (SaTC) [NSF-SaTC], Smart and Connected Communities (S&CC) [NSF-SCC], Cyber-Physical Systems (CPS) [NSF-CPS] programs. Another underlying objective of the workshop was to learn about a few of the application areas identified above and to consider other potential applications that could be researched and developed with the support of diverse, high-quality datasets.

It's also important to note NSF's interest in big data. In 2017, NSF identified "10 Big Ideas", areas in which they sought to build a foundation through investment in pioneering research and pilot activities and to identify and support emerging opportunities for U.S. leadership that serve the Nation's future [NSF-BigIdeas]. Among the ten ideas was, "Harnessing the Data Revolution,", which sought to engage NSF's research community in the pursuit of fundamental research in data science and engineering and the development of a cohesive, federated, national-scale approach to research data infrastructure [NSF-Harnessing]. The Idea of collecting

data from cars and trucks, sharing that data, and analyzing it to provide new, innovative solutions for automotive, transportation, and smart city and community applications aligns with NSF's interest in harnessing the data revolution.

Furthermore, in 2021 NSF sought input from the community on the specific needs related to collecting, sharing, and utilizing public or private datasets for networking and computer systems research, and any challenges associated with each (NSF 21-056) [NSF-DatasetsA]. NSF was interested in assessing where research progress is slowed due to the lack of datasets, especially when such data may either already exist or can be generated using existing infrastructure (including NSF-funded infrastructure). NSF received 33 responses on the specific needs for datasets to conduct research on computer and network systems, comprising contributions from 75 named contributors from at least 39 research institutions and other organizations [NSF-DatasetsB]. Included among the responses was one from two of the workshop organizers regarding The Need for Vehicle Telematics Data to Support Broad Scale Research [Papadop2021].

The NSF community is also engaged in various other projects and other workshops around the topic of network datasets. All told, it's clear that NSF and the large research community recognizes the importance of access to data to enable new, innovative research in areas of societal importance.

Finally, another underlying objective of the workshop was to consider overarching questions regarding building a community around automotive research datasets. As they listened to the presentations and participated in the discussion, participants were encouraged to consider questions such as: Do we need a community? How to build a community? How to engage industry? What activities to pursue? And, are [they] willing to participate?

Workshop Goal, Participation, and Agenda

Workshop Goal

The overarching goal of the workshop was to initiate a coordinated effort to bring together a community around the development and sharing of robust automotive datasets to foster and support new, open research in areas with strong societal impact such as smart and connected communities and development of cybersecurity and privacy protections for automotive applications

The primary focus of the workshop was applications of vehicle telematics and other data, including in-vehicle and outside the vehicle. Some of the topics covered during the workshop included the following:

- Geotab telematics technology and datasets available to researchers
- ORNL/NTRC research and facilities including dynamometers
- Community research datasets (ORNL, HCRL, Bosch, etc.)
- Desired characteristics of datasets, including real vs. synthetic datasets, attack generation, etc.
- Sharing infrastructure and framework, including a planned NSF CISE Community Research Infrastructure (CCRI) project
- Data quality/trustworthiness issues
- Privacy considerations
- Data analytics tools
- Other industry participation

Workshop Participation

A broad set of researchers who were producing or using automotive research datasets as well as other key stakeholders across academia, industry, and government were invited to participate in the workshop. A copy of the invitation email is contained in Appendix B. The invitation was sent to over 120 people in the community.

The workshop was originally going to be held as a hybrid in-person/virtual event at ORNL's National Transportation Research Center (NTRC), located in the Oak Ridge/Knoxville, TN area. Participants were invited to attend the workshop in person or to participate virtually via teleconferencing. In order to simplify logistics, the workshop was actually held as an all virtual event via Zoom and there was not any in-person attendance.

A broad group of 69 scientists, engineers, and technologists from 31 academic, industry, and government organizations in the United States as well South Korea, the United Kingdom, and Sweden attended and participated in the virtual workshop.

- Axle Technologies
- California Polytechnic State University
- Colorado State University
- Connected Vehicle Systems Alliance
- Colorado State University
- Cybersecurity and Infrastructure Security Agency
- Department of Transportation Volpe Center
- Florida International University
- Ford
- Geotab
- HCRL, Korea University
- Inca Digital
- Lear
- National Institute of Standards and Technology
- National Science Foundation
- Oakland University
- Ohio State University
- Open Insurance
- Oak Ridge National Laboratory
- Sandia National Laboratory
- SRI International
- Stony Brook University
- Tennessee Tech University
- University of California, Irvine
- University of California, Los Angeles
- University of Colorado, Colorado Springs
- University of Memphis
- University of Michigan
- University of Texas, Arlington
- Virginia Tech
- Wayne State University

The group included not only researchers producing automotive datasets and/or interested in using such datasets to drive their research, but also commercial vehicle telematics providers who are willing to share data with researchers, other industry representatives, leaders from standards organizations, and representatives from funding agencies and other government organizations.

Workshop Agenda

The workshop featured a lineup of speakers and breakout sessions as well as a virtual tour of NTRC's vehicle systems and research facilities. Opening remarks were provided by Dr. Shaun S. Gleason, the founding Director of the Cyber Resilience and Intelligence Division at ORNL. The speakers included researchers who are either producing automotive datasets and/or are interested in using automotive datasets to drive their research; commercial vehicle telematics providers who are willing to share data with researchers; and representatives from standards organizations and funding agencies. The breakout sessions explored future automotive dataset needs and potential applications of automotive datasets.

The first day of the two-day workshop started with David Balenson from SRI International introducing the workshop goals and providing background information on automotive datasets and their applications. Dr. Gleason's opening remarks emphasize the need for automotive

cybersecurity research and the importance of data and data science in supporting such research. The first day focused primarily on current automotive research datasets. Sam Hollifield from ORNL presented their automotive research including their ROAD dataset and CAN-D data decoding tools. Professor HuyKang (Hugo) Kim from Korea University's HCRL presented their publicly available datasets and cyber security challenges. Md Hasan Shahriar, a student from VATech, presented his survey of a number of datasets available to support CAN IDS research. Paul Maida from Geotab presented Geotab's Ignition platform which makes aggregated data collated from its global array of automotive telematics devices freely available for its customers and to support research in smart cities, autonomous vehicles, and intelligent transportation. Professor Jeremy Daily from CSU presented his work on heavy truck CAN dataset collection and Ted Guild from Geotab and Gunnar Andersson from COVESA teamed up to present the W3C and COVESA joint initiative to develop a common vehicle interface to enable interoperability and innovation for the future of transportation.

The first day ended with a working session in which participants divided into a number of groups to discuss future dataset needs, including desired characteristics, dataset quality and trustworthiness, and potential sources of datasets.

The second day started off with a virtual tour of ORNL's National Transportation Research Center (NTRC) during which Sam Hollifield and a number of center leaders highlighted their research capabilities and facilities, including a couple of labs with dynamometers. The second day focused primarily on applications supported or enabled by automotive datasets.
Md Hasan Shahriar gave a presentation focusing on his research on developing and evaluating an IDS for CAN. Professor Qi (Alfred) Chen from UC Irvine presented his extensive research on vulnerabilities and defenses in the AI stack in autonomous driving and in multisensor fusion. Dr. David Corman from the National Science Foundation (NSF) gave an invited talk on NSF's Smart and Connected Communities program, which includes work on transportation and personal mobility, autonomy, and data analytics.

The second day also included a series of shorter, lightning talks giving additional participants an opportunity to highlight their work and for workshop participants to learn about a broader range of research and supporting dataset needs. Professor Qadeer Ahmed from Ohio State presented his work on model-based intrusion detection, Ruiyang Zhu and Qingzhao Zhang, two students from U. Michigan, presented their work on trajectory prediction and drivable space detection, and Professor Gedare Bloom from University of Colorado Colorado Springs presented his work on automotive cybersecurity. Professor Daily from CSU talked about the CyberAuto and CyberTruck student competitions. The lightning sessions ended with Jim Davis from Geotab and Kumar Maddali from OPIN talking about open risk models and insurance.

The second day included a second working session in which participants divided into a number of groups to discuss potential applications of automotive datasets, including analytic tools, sharing infrastructure and frameworks, privacy considerations, and industry participation. The workshop wrapped up with Professor Christos Papadopoulos from the U. Memphis shared his

observations from the past two days and his thoughts about the benefits of an organized community around automotive datasets, the activities such a community could undertake, how they could engage industry, and next steps for moving forward.

A copy of the complete workshop agenda is in Appendix C.

Summary of Presentations

The workshop featured a number of presentations on research datasets, interface standards, research programs at NSF and various universities, and insurance programs. This section provides short summaries of the presentations. Links to the slides and recordings of the presentations are listed in Appendix D.

Research Datasets

ORNL Research Efforts, including ROAD dataset, CAN-D, etc. - Sam Hollifield and Stacy Prowell (ORNL)

Oak Ridge National Laboratory (ORNL) is focusing research efforts on technologies to increase the safety of automobiles by enhancing security of automotive controller area networks (CANs). CANs are ubiquitous for intra-vehicle system communications, and while functionally reliable, the protocol lacks many basic security features. Similar to other researchers working in this subject, we have identified numerous automotive cybersecurity deficiencies and have exhibited exploits resulting in potentially grave consequences, including injury to passengers and cargo. Our particular goal has been to develop vehicle-agnostic security mechanisms that are deployed on CANs. An emphasis of our work is developing state-of-the-art network intrusion detection systems (IDSs) for in-situ use on vehicles. During development of our IDSs, we found avenues to contribute to the overall research community outside of the direct application of network monitoring.

Notably, one major impediment to current automotive cybersecurity research is the lack of comprehensible data. For security and privacy, most vehicle manufacturers hold the encodings of CAN frames proprietary; thus, it is difficult to extract meaning from CANs without considerable reverse engineering efforts. Further, CAN data encodings vary by year, make, model, and even trim—-so reverse engineering efforts must be repeated for each research vehicle. Our contribution to this problem is an algorithmic pipeline named Controller Area Network Decoder (CAN-D) [Verma2021]. CAN-D extracts signals by identifying their bit boundaries within a CAN frame, tokenizing the signals by inferring endianness (byte order) and signedness (bit-to-integer mapping). Finally, CAN-D attempts to translate the signals by linearly scaling these values to appropriate units and applying labels (e.g., signal label = speed, signal unit = km/h) by matching the tokenized time series with an external sensor time series, in particular, by using standardized On-Board Diagnostic II (OBD-II) parameter identifiers (PIDs). Overall, the output is an industry-standard database file (DBC) which can then be used to decode future CAN data from that vehicle.

There also exists a lack of peer-reviewed, verified data which includes attacks for automotive networks. Thus, one of our contributions to research is the Real ORNL Automotive

Dynamometer (ROAD) CAN Intrusion Dataset [ORNL-ROAD]. To our knowledge, this dataset includes some of the first data captured from a vehicle with real (not simulated) advanced attacks that is publicly available. This dataset consists of 33 attack captures totaling about 30 minutes, and 12 ambient captures containing roughly three hours of ambient data. To maintain manufacturer privacy, we additionally obfuscated the dataset in a systematic manner while preserving signals within the data. We then use the CAN-D technology to extract signals from the obfuscated CAN and make available both the DBC with raw CAN logs and the extracted signal time series in a comma-separated file. The dataset contains examples of attacks with varying levels of sophistication, with the goal of making multiple attack instances available for CAN IDS research.

HCRL Research and Datasets - HuyKang (Hugo) Kim (Korea University/HCRL)

Professor HuyKang (Hugo) Kim from Korea University Hacking and Countermeasure Research Lab [HCRL] presented their publicly available datasets and cyber security challenges. HCRL focuses on data-driven security based on machine learning and data mining to extract and learn useful knowledge from massive data. The lab has unique and valuable datasets collected from real-world services, including online game service data, mobile payment and e-commerce transaction data, car-driving and attack data, which are shared with the public [HCRL-Datasets]. HCRL also hosts and manages a set of cyber security challenges for students [HCRL-Challenge].

From 2014-2019, HCRL generated automotive datasets using real commercial cars, including the YF Sonata (Hyundai Motors), Soul (KIA Motors), and Spark (GM Chevrolet). The automotive datasets include a driving dataset with unique driver patterns that can be used for driver identification applications, a car-hacking dataset, and CAN intrusion dataset (OTIDS), a CAN signal extraction dataset, and a survival analysis dataset for automobile IDS. The OTIDS dataset includes various in-vehicle attacks and is used for the car hacking challenges. HCRL also developed its own automated CAN reverse engineering software tool that is openly available [HCRL-Analyzer]. HCRL's papers on IDS and related datasets are highly cited in the in-vehicle IDS research area.

HCRL's work on automotive datasets and cybersecurity R&D challenges forms a "virtuous circle" in which datasets enable AI/ML-based IDS development, the IDS development feeds competitions, and competitions facilitate development of new attack patterns and detection algorithm, engagement of government and industry, and knowledge transfer to industry. As part of the challenges, HCRL provides test cars and student teams prepare their own attacks as well as their own IDS. The attacks include an attack video, attack code, and document of attack definition and threat information. The competition attracts considerable attention from industry and government, promotes automotive cybersecurity research, and gives students invaluable experience.

From 2018-2020, HCRL has looked beyond CAN IDS to automotive Ethernet IDPS. This involves new attack identification for automotive ethernet, including AVTP protocol (frame injection attack), PTP protocol (PTP flooding attack), and CAM Table attack (MAC flooding attack). Looking into the future, HCRL plans to extend its work to EVs and other types of in-vehicle networks, including CAN-FD and automotive ethernet. They also plan to collaborate with KATECH [KATECH], an automotive technology company with an anechoic chamber, dynamometer, driving simulator, and V2X simulator. The 2021 cybersecurity challenge will focus on developing attacks and defenses for an infotainment device and the resulting datasets will be made available to the public. Together all of these activities support HCRL's mission of promoting data-driven security in automobiles.

CAN Intrusion Detection Dataset Survey - Md Hasan Shahriar (VATech)

As part of his work on developing a new AI/ML-based IDS, Md Hasan Shahriar, a Ph.D. student working with Professor Wenjing Lou at VATech, presented his survey of available CAN intrusion detection datasets. CAN IDS fall into five major categories: rule/specification-based, physical side-channel, frequency/timing-based, payload-based, and signal-based. Attack implementations used to evaluate an IDS are either real attacks or synthetic attacks. Attack categories are fabrication attacks, suspension attacks and masquerade attacks.

Shahriar examined seven CAN datasets as depicted in the following table.

Dataset	Country	Year	Implementation	Attack Categories
HCRL OTIDS	Korea	2017	Real	Fabrication
HCRL Survival Analysis	Korea	2018	Real	Fabrication
HCRL Car Hacking		2018	Real	Fabrication
TU Eindhoven CAN Bus Intrusion	Netherlands	2019	Synthetic	Fabrication, Masquerade, Suspension
CrySys Lab CAN Traces	Budapest	2020	Synthetic	Masquerade
Bosch SynCAN	Germany	2020	Synthetic	Fabrication, Masquerade, Suspension
ORNL ROAD CAN Intrusion	United States		Real, Synthetic	Fabrication, Masquerade, Suspension

Mr. Shahriar reviewed and compared the characteristics of each dataset, including source, car, year, implementation, attack types, attack categories, and context; dataset attributes and size; and benefits and drawbacks.

Mr. Shahriar selected the Bosch SynCAN dataset for his research because it is the sole signal-based CAN dataset, contains the most nuanced masquerade attacks (drift/replay), contains attacks targeting a single signal (not all 64 bits), and allows for testing a very advanced, signal-based IDS. However, it still has drawbacks, including its use of synthetic data which is an imperfect proxy for real data, real effects of the simulated attacks cannot be verified, and unavailability of CAN binary data along with the signals.

Mr. Shahriar explained that IDS research is significantly hindered by a few major issues. There is no existing comprehensive CAN IDS dataset that includes binary payload, decoded signals, driving context, ambient data, and data from other sensors (camera, lidar, radar, etc.). There are no real masquerade attacks and limited real attacks with verification of attacks' physical impacts. There are also limitations of CAN reverse engineering tools for decoding, which may not be 100% accurate and may not provide semantics.

Mr. Shahriar concluded by sharing dataset features that would greatly improve IDS research, including the payload along with the time-series signal data, physical interpretation (semantics) of the signals, ambient/external sensor data to represent the driving condition, driver's intent to represent the driving context, advanced (masquerade) attacks (both dynamometer and roads), and data from multiple vehicles for generalization.

Geotab Data Product Discovery - Paul Maida (Geotab)

During these unprecedented times, one thing that almost everyone has come to learn is just how important data is — whether for businesses trying to operate, governments trying to reopen, addressing strained supply chain issues, or for the everyday consumer trying to understand the impact of events. In fact, there's no question that data is the undercurrent of these unparalleled times, with access to it proving to be especially critical in helping governments and businesses make informed decisions. But while having access to aggregate data is beneficial to many, the ability to properly navigate the technical and visual aspect of data is not universally understood — potentially hindering users' ability to leverage data platforms properly and effectively.

To further enable all users to better analyze and digest connected vehicle insights, our data and analytics team launched Geotab Ignition [Geotab-Ignition]. Geotab Ignition aims to help advance our customers' knowledge and the research surrounding smart cities, autonomous vehicles, and intelligent transportation by providing seamless exploration of anonymous aggregate data. At Geotab we are at the forefront of ethical use of data and data anonymization. Data, and in particular, insights garnered from aggregated data can transform and disrupt entire industries.

Originally launching with 12 datasets, in just a few months' time, the team has expanded the Ignition platform to include 25 datasets to date. These aggregate datasets are divided into four different categories: Urban Infrastructure, Weather, Location Analytics and Mobility. The top five industries exploring Geotab Ignition by user volume are fleet management, education organizations, consultancies, government, and software companies.

Geotab has created an agile product discovery team within data analytics and collaborates with others (academics, research facilities, commercial businesses) to quickly develop MVP (minimum viable products), some of which can ultimately become commercial products. Researchers who want to take advantage of the Geotab Ignition platform can simply register for a free account. Once registered and logged into the platform, you can create queries to access

and visualize the available data. This will allow you to further explore the data and even embed it on your own platform. If you are interested in automatically pulling the available data from behind the scenes, contact Geotab [Geotab] to request a free service account!

Heavy Vehicle CAN Dataset Collection - Jeremy Daily (Colorado State University)

Heavy vehicles differ from passenger vehicles in they are more horizontally integrated and utilize different components from different suppliers under the same hood. For example, a Kenworth truck may have a Cummins engine or a PACCAR engine. Furthermore, truck owners often add additional networking components on the vehicle, which increases the variety of networking found on trucks. However, few (if any) public data sets were available for heavy vehicles. To this end, an NSF-sponsored project resulted in a custom CAN logger along with a corpus of CAN data from the SAE J1939 network on heavy trucks.

Some of the datasets featured in the Colorado State University featured data that came from crash tests. These were interesting because of the rich contextual data that also accompanied the data. An example was shown where a Detroit Diesel powered Freightliner tractor-trailer combination was struck in the rear by an Envoy. The event record showed a 2.5 mph speed change, but the raw CAN data showed about a 5 mph speed change. These data sets can help research crash investigations and incident response.

In the effort to collect CAN data, Dr. Daily and his student designed three versions of an open-source hardware device called the CAN Logger 3. This design leverages the ease of the Arduino IDE with the speed of an MK66 ARM-Cortex M4F microprocessor with two CAN channels. The impressive results showed fully encrypted CAN Data on multiple channels at over 6Mbps. The resources for the CAN Logger 3 and some of the datasets were linked at Dr. Daily's website [Daily].

Dr. Daily brought up three challenges for dataset from automotive systems. The first was a migration challenge that happens when a data curator moves in their career from one institution or company to another. The cited example was Dr. Daily's transition from the University of Tulsa to Colorado State University. The old files hosted at U. Tulsa were tied to his personal account and became unavailable. Standing up the new datasets may take some time. The second challenge to the community was on methods to make the data searchable and produce useful artifacts from the data. Finally, getting industry participation openly is challenging. The trucking companies that participated in the data collection efforts did not openly release their data. Instead, researchers had to login to a protected server with explicit rules against exfiltrating data.

Interface Standards

CVII Overview - Ted Guild (Geotab)

The automotive industry lacks a common vehicle interface although it sorely needs one to enable interoperability and innovation for the future of transportation. To address this need the Connected Vehicle Systems Alliance (COVESA) [COVESA] and the World Wide Web Consortium (W3C) [W3C] together have launched a joint initiative, the Common Vehicle Interface Initiative (CVII).

The cornerstone of this initiative is a common data model, the Vehicle Signal Specification (VSS) [COVESA-VSS]. Presently each automotive manufacturer does data differently on their vehicle networks and many have created proprietary application programming interfaces (API) for interacting with their vehicles. The ever-increasing number of APIs, their varying capabilities, deliberate limitations, evolution and in some cases deprecation have been a major obstacle to creating an ecosystem that could support solutions for connected vehicles.

Besides the common data model, the initiative has a common interface (API), the Vehicle Information Service Specification (VISS) [W3C-VISS], providing an abstraction layer so applications can be written toward and run across different makes of vehicles. CVII also seeks to push the benefits of a common data model for vehicle signals down the stack, into the control units that create and act upon signals information. The initiative also promotes this common data model in the cloud where it is being used for artificial intelligence and Internet of Things (IoT) interoperability. This initiative is also seeking to collaborate with other standards efforts, proposing its data model where one is absent, define ways to translate information or otherwise coordinate solutions.

Implementations for pieces of this technology stack range from proof of concept to production vehicles. Although widespread adoption is yet to be achieved, they are starting to see increased uptake from OEM, suppliers, and service and solutions providers as well as cloud providers.

Connected Vehicle Systems Alliance and CVII - Gunnar Andersson (COVESA)

Unlike W3C which produces web-based standards for several different industries, COVESA focuses solely on the automotive industry, helping develop systems necessary to support connected vehicles. COVESA is a recent relaunch of GENIVI where this joint effort started, the rename reflecting the change in scope [COVESA].

Through regular roundtables, workshops, and direct company engagements, COVESA and W3C are getting input from the automotive industry, learning of common pain points, and exploring solutions. They compare and unify approaches across manufacturers, suppliers, and other stakeholders to be able to define solutions that address a wide range of use cases from different perspectives.

The focus is on core, fundamental components but allowing for variation as having too rigid an architecture risks rejection and does not reflect the reality of independent, unique, and established if not entrenched systems. As an example, the common data model, VSS [COVESA-VSS], allows for private branches to support data unique to a given manufacturer in addition to the agreed upon standard catalog.

COVESA and W3C are seeking alignment with other connected vehicle efforts. There are ongoing efforts to engage with the various other standards efforts and trade associations in this related space. They are also creating open-source implementations of portions of the technology stack to demonstrate capabilities and facilitate adoption.

Research Programs

ORNL National Transportation Research Center Virtual Tour - Sam Hollifield (ORNL)

ORNL is also home to the National Transportation Research Center (NTRC). NTRC is the Department of Energy's only designated user facility focused on performing early-stage research and development in transportation technologies. User facilities include both full-scale hardware-in-the-loop (HIL) capabilities as well as numerous, varied dynamometers for whole vehicle integration. Some of the key facilities which have been instrumental in our data generation include:

- Vehicle Security Lab (VSL): A laboratory with a four-wheel rolling dynamometer, focusing on security assessments and prototyping security tools. The VSL is housed in a shielded environment and allows for safe, comprehensive testing of whole vehicle platforms for cybersecurity research [NTRC-VSL].

- Connected and Automated Vehicle Environment (CAVE) Laboratory: A unique environment to evaluate intelligent mobility systems using an advanced steerable chassis dynamometer with the ability for autonomous integration [NTRC-CAVE].

- Vehicle Systems Integration Laboratory (VSIL): A full-scale powertrain dynamometer with the ability to utilize HIL technologies for powertrain configurations ranging from light-duty cars to Class 8 trucks [NTRC-VSIL].

- Vehicle Research Laboratory (VRL): A laboratory with a single-axle dynamometer that is capable of performing EPA and car certification cycles, catalyst performance monitoring, and emissions sampling measurements to research dynamics of engine and emission operation [NTRC-VRL].

- Power Electronics and Electric Machinery (PEEM) Laboratory: The primary laboratory for the DOE Vehicle Technologies Office electric drive research and development. Capabilities include wireless and wired power charging, power inverters and converters, compact electric motors, and packaging. The PEEM contains environmental chambers, including temperature and humidity, numerous varied voltage sources, and high-voltage device evaluation [NTRC-PEEM].

- Fuels, Engines, and Emissions Research Center: A collection of laboratories which enable the research and development of internal combustion engine efficiency, emission controls, and fuel effects. These laboratories feature seven HIL engine dynamometers, analytical chemistry laboratories, and the ability to research exotic fuel sources [NTRC-FEERC].

CANShield: An Intrusion Detection Framework for Controller Areas Networks - Md Hasan Shahriar (VATech)

Md Hasan Shahriar, a Ph.D. student working with Professor Wenjing Lou at VATech, presented his work on CANShield, an IDS for CAN. The proliferation of ECUs communicating via CAN along with wireless external connectivity, ADAS, and infotainment has increased the cyberattack surface of modern cars. Signature-based and anomaly-based IDS are common, but none can detect advanced signal-level attacks. Bottlenecks to effective research in this area include obfuscated CAN payloads, limited CAN IDS datasets, complex data structure of CAN messages, limited advanced attack scenarios, and limited real attack scenarios.

IDS types include rule/specification-based, physical side-channel-based, frequency/timing-based, payload-based, and signal-based. CANShield falls into this last category and performs anomaly detection on signal-level multidimensional time-series data. Raw binary data collected from a vehicle's CAN bus can be decoded using a DBC file. Reverse engineering tools such as CAN-D can provide a DBC file. Raw data with missing elements is not suitable for ML-based IDS.

CANShield uses a convolutional neural network (CNN) with a data processing technique (pipeline) for the high dimensional CAN signal stream. It trains multiple CNN-based autoencoders (AEs) to detect any violations during cyberattacks. CANShield uses clustering-based signal reordering to accelerate training, a three-step analysis of the reconstruction losses, and an ensemble-based detector to boost up the overall performance.

CANShield consists of three modules: A data preprocessing module reads and decodes the binary payloads to signals, creates a data queue of time-series signal values, and creates different views with different sampling periods. A data analyzer module trains multiple AEs on each of the views and finds the reconstruction losses. An attack detection module analyzes and determines thresholds on reconstruction losses and calculates the attack probability for the "voted" model.

CANShield was evaluated on the ETAS/Bosch SynCAN dataset used with their CANet IDS [Hanselm2020]. The synthetic signal-level CAN data contains 10 CAN IDs and 20 signals in total. Attacks considered included plateau attack, continuous attack, suppress attack, flooding attack, and playback attack. Metrics considered where: attack detection - the detection of any injection or modification of any CAN message; event detection - detection of a percentage of injected messages in a collection of consecutive message injections; and detection latency - the delay between the first injected message and the first detected message.

Depending on the mapping period, CANShield detected continuous attacks (sampling period 1), plateau flooding and playback attacks (sampling period 5), and suppress attacks (sampling period 50). A “voted” model boosts up the performance against all types of attacks. CANShield outperforms CANet in flooding and suppress attacks. CANShield (“voted”) maintains high event detection rates for flooding, suppress, and plateau attacks. Detection latency varies by types of attacks and the “voted” model improves the latency. Flooding, suppress, and playback attacks are detected almost instantly.

CANShield is also robust against adaptive attacks. It is tough for the attacker to control multiple ECUs to temper the input image to an AE. Multiple AE models (different sampling periods) makes the defense more robust. The IDS works in near real-time and is robust against adversarial replay attacks.

Towards Secure & Robust AI Stack in Autonomous Driving & Beyond - Qi Alfred Chen (UC Irvine)

Qi Alfred Chen, an Assistant Professor of Computer Science at UC Irvine, presented his work on AI stack security in autonomous driving & smart transportation. Autonomous driving (AD) technology equips vehicles with various types of sensors to enable self driving, such as cameras, radar, and LiDAR, as well as GPS antenna/receiver, inertial measurement unit (IMU), and industrial PC. The AI stack in industrial-grade AD includes: controls, such as steering wheel and pedals; the physical world, including roads, other vehicles, obstacles like road cones, and traffic signs; the sensors mentioned above; perception algorithm, including obstacle avoidance, which involves object detection, object tracking, and multi-sensor fusion, lane detection, traffic light detection, and traffic sign detection; localization algorithm, including global localization and multi-sensor fusion; prediction; and planning.

Professor Chen’s research has considered the practical and fundamental attack surfaces of both sensors and the physical world and leverages spoofing and jamming attacks that have been discovered on all sensors used in AD systems. In particular, Professor Chen’s work successfully conducted security analysis on attacks and defense involving:

- 3D object detection via LiDAR spoofing combined with adversarial ML to spoof a fake front car at perception output, causing emergency brake or DoS [Cao2019, Sun2020].

- AD object tracking via stickers placed on the back of a front car and optimized bounding box position shifting to move a road-side object into the current lane, causing an emergency brake; or move a front car away, causing a crash [Jia2020].
- Multi-sensor function (MSF) for AD perception via a maliciously-shaped adversarial 3D object (e.g., a traffic cone or rock) can influence both camera pixels and LiDAR point cloud, allowing a victim to fail in detecting a front obstacle and crash into it [Cao2021, MSF-ADV].
- A production lane detection DNN model via malicious dirty road patterns and an optimization-based method to cause a victim to drive out of the current lane boundaries within 1 sec, which is far below normal driver reaction time (~2.5 sec) [Sato2020, DRP-Attack]
- AD traffic light detection via targeting the use of region-of-interest (ROI) to narrow down detection scope in raw camera input and used GPS spoofing to move the correct traffic light out of ROI, causing denial-of-service, or move the wrong traffic light into ROI, causing red light running [Tang2021, ROI-Attack].
- State-of-the-art MSF using GPS, LiDAR, and IMU for AD localization via FusionRipper, a two-stage opportunistic attack created by dynamic and non-deterministic factors such as sensor noises and algorithm inaccuracies whereby an attacker that tailgates a victim for two minutes can almost always (97% chance) find an opportunity to break sensor fusion, and cause a victim to drive off road or to the wrong way [Shen2019, Shen2020, FusionRipper].
- Connected vehicle (CV) protocols such as platooning [Hu2021].
- Infrastructure-side CV systems [Chen2018, Feng2018, Wong2019, Huang2020, Hu2020].

Professor Chen's group is actively developing the research space on AI stack security in autonomous driving and smart transportation systems [CAV-SEC]. Their current work has been mostly on the attack side and the group will mainly focus on the defense side next. Professor Chen also noted a number of community activities, including the Automotive and Autonomous Vehicle Security (AutoSec) Workshop, the IEEE SafeThings'21 Workshop, and the first AutoDriving-themed hacking competition at DEF CON.

NSF Smart & Connected Communities (S&CC) and Cyber-Physical Systems (CPS) Program Overview - David Corman (NSF)

Dr. David Corman, a Program Director in the Computer and Information Science and Engineering (CISE) Directorate at the national Science Foundation (NSF) talked about several programs of interest to workshop participants, including the Cyber-Physical Systems (CPS) [NSF-CPS] and Smart and Connected Communities (S&CC) [NSF-SCC] programs.

Challenges with security and transportation are of great interest to the CPS program. The goal of the program is to develop the core system science needed to engineer CPS upon which people can depend with high confidence. This includes research into security, autonomy ML, AI,

verification, networking, etc., which are germane to multiple applications, like individual vehicle transportation, managing transportation within a city, aerospace with airplanes, and energy with the smart grid. The work is cross cutting research, so it applies across a set of application spaces. It's a multi-agency program including the Department of Transportation, Federal Highway Administration, Department of Homeland Security, National Institute of Food and Agriculture, and National Institutes of Health.

The Smart and Connected Communities (S&CC) program, instead of just foundational research, looks to do use-inspired, community-focused research. The goal is how fundamental science and engineering impact communities to improve the quality of life for people that live in, travel through, or work in a community. The research impacts communities and communities help shape the research. The research spans the social and technical dimensions - the S&CC core is similar to CPS, but includes social sciences, sociology, psychology, economics, and privacy. Research ideas from technology and social sciences intertwine and integrate along with deep community engagement applicable across many services and domain areas such as emergency management and public safety, health and wellbeing, transportation and personal mobility, energy, and ecosystem services. Proposers include engineers, computer scientists, sociologists, urban planners, and others. Since 2016, the program has funded over $100M and over 150 awards.

The S&CC program supports university-community teams across the U.S. with over 200 committees, including small, large, and medium sized towns, and urban and rural communities. It looks for community engagement, impact, and integrated research. The program funds a wide range of research areas including AI/ML/data analytics, economics, privacy, safety, and security, and application areas such as transportation and personal mobility, urban and rural planning, and emergency management and public safety.

Dr. Corman highlighted the Civic Innovation Challenge (CIVIC) program, which is a partnership among NSF, DOE Vehicle Technology Office (Mark Smith), and DHS (David Alexander leading resilience and Elizabeth Asche from FEMA). The program has a Cooperative Agreement with Metrolab Network who help community support activities. Dr. Corman invited people to check out the program website [NSF-Civic]. The CPS and S&CC programs develop foundation research, where CPS looks for technology maybe 5-10-15 years out and S&CC probably 5-10 years out. CIVIC looks at how to take foundational research and turn it into action in the community. NSF just finished the solicitation for Civic 1.0, which included two tracks, Track A: Communities and Mobility, and Track B, Resilience to Natural Disasters. NSF is in Stage 2 of the first year of the program which is executing pilots that emerged from the first stage and is looking at transferability, scalability, and sustainability. There has been huge interest in the program, with 52 Stage 1 awards and, surprisingly unprecedented from NSF, 17 out of 51 Stage 2 awards, each about $1M to conduct a pilot that will show impact in the community. Examples of Disaster Resilience focus areas include a range of disaster types, including floods, hurricanes, and wildfires; evacuation management and planning, resource matching to optimize post-disaster response, and financial resilience.

Model-Based Intrusion Detection - Qadeer Ahmed (Ohio State)

Professor Qadeer Ahmed from Ohio State University presented his work on “Safety and Security by Design: Model-Based Gateway Intrusion Detection System”. The work argues that safety should be incorporated by design, with the help of physics-based models. In current systems errors are introduced early but are not detected until later when the cost to correct them increases substantially. Professor Ahmed noted that CAN has no inherent security and efforts to add it are costly in terms of performance and latency. He suggests that we take advantage of near future vehicle architectures that will revolve around a central gateway with various segmented network segments (CAN or Ethernet) connected to a gateway. The latter can see most traffic and implement security mechanisms. We can also map critical signals (e.g., the accelerator) between various segments. Physical system dynamics can help us determine attacks when they are outside the realm of physical properties. Professor Ahmed proceeded to present the threat model, which assumes that the gateway may be compromised enabling man-in-the-middle attacks. He then presented an analytic model of the engine and modeled the attack as an unknown input to the model observed by an “unknown input observer” who understands the physical model of the engine and filters out unknown values. Professor Ahmed then presented his implementation of the security control mechanisms. Then he presented the results of a simulated throttle attack and showed that they were able to recreate the throttle attack from the unknown signals and filter it out.

Trajectory Prediction and Drivable Space Detection - Ruiyang Zhu and Qingzhao Zhang (U. Michigan)

Ruiyang Zhu and Qingzhao Zhang, Ph.D. students working with Professor Z. Morley Mao at the University of Michigan gave a two-part presentation. Trajectory prediction is part of a pipeline, coming after object segmentation and followed by motion planning. Attacks on trajectory prediction by introducing perturbations in deep learning models can propagate and affect the outcome significantly. There are two questions: (a) can we generate adversarial examples of trajectory prediction by perturbing history trajectories; and (b) can the adversarial examples cause real-world safety issues. To answer these questions, they designed an adversarial attack on prediction. They used a model of two cars, the autonomous vehicle (AV) and the other vehicle. The goal is to perturb the behavior of the other vehicle so that the AV will reach the long trajectory prediction. They tested under two scenarios, white box and black box with three models and three datasets. The results (whose details were not presented) showed that the prediction error can be raised as high as 150% potentially triggering brakes and collisions. Then they commented on how to create better datasets by considering the worst-case performance, adding more than trajectories such as heading, semantic maps, and labeled lane areas.
The second part addressed collaborative drivable space detection (the space where a vehicle can drive based on sensor data). Currently vehicles make decisions independently, but there are benefits with collaborative drivable space detection such as eliminating blind spots. The proposed method is grid occupancy detection where 1m by 1m grids are marked as drivable/not

drivable. Preliminary results through the Carla simulator and shared LiDAR data shows an almost linear increase in both the size and accuracy of space area detection. Better results can be achieved with more comprehensive and realistic multi-vehicle datasets, better metrics, semantic (drivable space) segmentation models for multi-vehicle detection, and better techniques to cope with network transmission.

Automotive Security at UCCS - Gedare Bloom (U. Colorado Colorado Springs)

Professor Gedare Bloom runs the Embedded Systems Security Lab (ESSL) at the University of Colorado Colorado Springs. The lab focuses on the hardware/software systems side of things and applies it to security. It conducts research in a number of areas, including kernel hacking, IoT security, infrastructure security, and vehicle security. The work started in 2016 with funding from NSF and more recently from the State of Colorado. Key contributions include connections between vehicular safety and security, evaluation and data collection, and characterization of protocols suitable for in-vehicle networks (CAN). ESSL started out looking at passenger vehicles and more recently, in collaboration with Dr. Daily, they are looking at heavy trucks and other ground vehicles.

CAN datasets are a challenge. Some datasets are openly and freely available. For example, the HCRL datasets were very useful in the lab's early research. ESSL has access to its own datasets, but they could not be shared which made it harder to publish. Professor Bloom believes access to openly available datasets is critical to publishing research to enable reproducibility by other researchers. ESSL also developed its own tools for generating synthetic data, though the data has limited fidelity and quality [Bloom-SBA]. Professor Bloom believes there are opportunities to develop high fidelity protocol log generation capabilities that allow to demonstrate attacks without post processing techniques that impact fidelity. ESSL also uses benchtop setups that implement the CAN bus more realistically while controlling the behavior of the CAN or protocol they want to use. And the lab also runs things on real cars. Professor Bloom and his team have a lot of vehicular security publications (for example [Ezeobi2020], [Olufowobi2020], and [Bloom2021]).

A next step for Professor Bloom and ESSL is access to a new facility which includes a graduate research lab, faculty office space, and a large garage bay all close together. They hope to build out advanced vehicle security infrastructure and to collect and share datasets. The ESSL team is composed of students, a post doc, and affiliated alumni. A lot of the students touch on vehicle security research, but they also work on other problems in embedded systems.

CyberAuto and CyberTruck Challenges - Jeremy Daily (CSU)

The CyberAuto Challenge was started in 2012 with a twin mission: (a) train the next generation of cybersecurity talent needed to address issues facing the industry, and (b) build and foster a community of interest from academia, industry, government, and security researchers. The same mission applies to the CyberTruck Challenge, which was started in 2017.

The presentation covered the experiences the students had and interactions they engaged in with a few photographs from the CyberTruck Challenge. The data or discoveries from the Challenge events are not released to the public, since all the participants participate under a non-disclosure agreement. However, many workshop participants were excited about the events and gave personal endorsements from their own experiences.

Insurance Applications

Collaboration Towards Open Risk Models - Jim Davis (Geotab)

The insurance industry is operating under increasingly smaller margins, especially for heavy trucks. Telematics can help. Insurance really wants to focus on driver behavior such as speeding, unsafe following distance, behavior around pedestrians and so forth. With the U.S. government 2017 mandate, Electronic Logging Devices for commercial vehicles has started opening up interest in data driven InsureTech for automotive insurance. A challenge in bringing new insurance products to market is slower moving regulatory commissioners at state/province level.

Risk models developed over years have been upended with changes in behavior resulting from COVID. Research and development needs to create compelling arguments for achieving potential attainable from telematics in order to influence regulators and demonstrate the capability to insurance companies.

Research & development with OPIN opportunities include: focus on opening aggregated commercial vehicle telematics data sets for researchers to extend the 'insurance' use case and uncover other interesting and value add use cases; collaborate with researchers, experts in safety, risk management, insurance, claims and data science; and collaborate with Open Insurance [OPIN]. It is possible to lower risk, tailor education and coaching to drivers on their behavior, reduce claims and subsequently lower costs to the consumer.

Open Insurance Project - Kumar Maddali (OPIN)

There is a need to translate data into usable information and from that derive actionable knowledge. Open Insurance (OPIN) [OPIN] is an insurance industry think tank comprising four hundred organizations seeking to create common data standards, open APIs and information workflows (e.g., claims) to streamline the insurance industry. Lacking those presently in the industry reduces ability to integrate solutions, stifling innovation and digital ecosystems. OPIN collaborates with other standards efforts including COVESA [COVESA] and W3C [W3C] and actively looks to engage researchers to learn the different ways telematics and other data can be leveraged. The data and APIs can be used to build profiles on driver behavior, embed logic on vehicles, enable usage-based insurance, prompt appropriate and timely predictive

maintenance and more thoroughly understand risk, leveraging artificial intelligence to analyze data en masse.

Future Dataset Needs

A working session was held to explore needs for future automotive datasets. The participants were divided into four breakout groups for this session. All of the groups were given the following suggested topics to stimulate the conversation. The actual topics addressed varied based on the members of the group and were much broader, as captured in the notes.

- Desired characteristics of rich, robust automotive datasets
- CAN, other in-vehicle networks, sensor, other data (e.g., driver cam)
- Real vs. synthetic datasets, attack generation, etc.
- Dataset quality and trustworthiness
- Sources of datasets (ad-hoc, community, industry, government, real-world pilots, reference datasets)

This section briefly summarizes the discussions around each of these topics. Detailed notes capturing the discussion from the four breakout groups are provided in Appendix E.

Desired Characteristics of Automotive Datasets

Several common themes were brought up in this session: the need for diverse datasets, a common dataset structure, detailed documentation of the underlying data collection methods with extensive metadata, and the need for high-fidelity synthetic data generation. CAN data is in high demand, but the lack of signal definition is currently the hardest problem. A CAN data steward was suggested, who will ensure datasets are normalized, and will also work with industry to get high-quality datasets. Safety is paramount when collecting automotive data, especially cybersecurity data, yet there are no safety guidelines. Similarly, there are no guidelines for preserving privacy.

CAN and Other In-Vehicle Data

Quality CAN data is very valuable but also very hard to get. IDS research requires physical level analog information (voltages) that may be vehicle specific. It also requires well-documented attacks, clear event timelines, and a detailed description of the experiment setup. Experiments may require dynamometers to ensure safety. Datasets may only be available through OEMs and other industries, who presumably generate such data at a substantial cost. Currently, there is a lack of partnerships to package and share data with researchers. Data from other in-vehicle networks and systems such as radar/LiDAR is also lacking. Finally, there is no standard naming for CAN signals, and tools employing heuristics to discover signal definitions such as CAN-D and LibreCAN [Pese2019], are not sufficient.

Real vs. Synthetic, Attack Generation, Etc.

While synthetic data is important, the fidelity of such datasets is critical; yet there are no established methodologies and guidelines to generate synthetic datasets. Attacks used by researchers thus far are ad-hoc with no organization or taxonomy. AI can play an important role in generating attacks but work in that space is at its infancy. Even generating data from real systems can be tricky; for example, one must guard against residual data from a previous experiment affecting future experiments. Crowdsourced data may produce high-quality datasets, but it needs to be standardized. Non-CAN data such as Ethernet, Flexray, and LIN, is much more difficult to find but important since vehicles are still expected to employ such networks in the near future. Aggregated datasets such as those provided by Geotab are important for specific applications.

Dataset Quality and Trustworthiness

Automotive datasets can be very large and contain redundant or superfluous data, thus we need appropriate clean-up techniques. OEMs may be a trustworthy source of data, but we need to establish channels to share. Currently, we do not have best practices to securely collect, archive, discover and distribute trustworthy data. Trustworthiness includes authenticated sources, appropriate and well-documented collection methodology, but also lack of bias. To develop innovative applications for sectors such as insurance, weather monitoring and prediction, environmental monitoring, smart cities, etc., we need data from multiple types of sensors, including microphones and cameras. We also need data that covers the entire lifecycle of the vehicle, including delivery, registration, service, accidents, etc.

Sources of Datasets

Industry is an invaluable source of automotive datasets, and the community should pursue partnerships. Researchers need to be mindful of the cost of generating datasets and any privacy/intellectual property restrictions. Datasets may not always match researcher needs and can become obsolete quickly, so the community needs a continuous feedback loop with industry. The government can also help with open datasets through pilots such as those funded by the DOT, and emerging regulatory frameworks can guide dataset needs.

Potential Applications

A second working session was held to explore potential applications of automotive datasets. The participants were divided into four breakout groups for this session. All of the groups were given the following suggested topics to stimulate the conversation. The actual topics addressed varied based on the members of the group and were much broader, as captured in the notes.

- “Killer” apps for automotive datasets
- Current and future analytic tools for processing automotive datasets
- Sharing infrastructure and frameworks for datasets
- Privacy considerations
- Industry participation

This section briefly summarizes the discussions around each of these topics. Detailed notes capturing the discussion from the four breakout groups are provided in Appendix F.

“Killer” Apps for Automotive Datasets

There are many smart city killer apps to improve transportation especially through the combination of datasets. Datasets should avoid biases due to data collected from a narrow class of vehicles (for example, expensive vehicles equipped with telematics). Lack of equity may also lead to biases. Industry collaboration was emphasized, with universities acting as catalysts in such partnerships. Advertising in apps was undesirable, but perhaps necessary. Predictive vehicle service applications were very welcome, as well as traffic management applications. Insurance would greatly benefit from automotive datasets to help devise plans that consumers would buy. Apps that enhance dataset quality were highly desirable, such as normalizing datasets, ensuring integrity, compression, and cleanup. Other useful apps include comparing ICE with EV costs, determining road conditions, and more. Apps that directly benefit vehicle owners/occupants were also discussed, e.g., turn on heated seats in cold weather, detect driver fatigue and distraction, or a vehicle equivalent to an AirBnB. Crowdsourced vehicle and smartphone data can lead to innovative apps such as weather and road conditions. Data-driven event forensics is another important application for both insurance and law enforcement. Apps that facilitate secure sharing of data between vehicles and other devices are also highly desirable. Finally, opportunities to create innovative tools were discussed, along with the use of Jupyter Notebooks to help provide context.

Tools for Processing Automotive Datasets

Datasets generated from the vehicle and through smart devices are different and there is an advantage in merging the two and investigating what applications are enabled. ML tools can be used effectively to detect anomalous behavior in vehicles used in routine tasks such as school buses. Tools are also available from other communities such as social, economic and science

communities, as well as commercial tools geared toward processing large data such as those from Splunk and Gravwell. In terms of future AI/ML tools, we should look at on-vehicle tools but beyond object recognition and LiDAR such as detecting unbalanced tires to determine if the vehicle is safe to operate. AI tools can extend cybersecurity to counter attacker actions that prevent the vehicle from moving. Federated learning across Pilots and various OEMs is currently very limited and could benefit from tools that examine multiple datasets with geographic and regional driver diversity.

Sharing Infrastructure and Frameworks

Moving data back and forth in digital twins is a challenge in terms of handling economic and privacy incentives. There are a few data sharing portals, but they are not easy to use. Portals should normalize the metadata and storage and ideally should be as easy as drag and drop. Portals can cross-correlate between different modes such as CAN, Ethernet, and other networks, automatically anonymize/obfuscate, and offer analytics capabilities. Selecting the custodian is a challenge. Data may also be available through ELD devices and the Auto-ISAC, but it has to be trusted since bad data can poison existing, carefully curated datasets. Preserving timing (and dealing with drift) is also important, especially across multimodal datasets. There are risks with centrally maintained datasets, especially when coordination is desired. Another challenge is how to maintain data in the presence of sensor degradation. Data security is also important as well as data integrity, especially AV data, and when location services are not available. Finally, there is a challenge in data sharing due to security and background checks needed for some researchers.

Privacy Considerations

The privacy process includes consent to collect data from a car, which is presumably covered when signing paperwork at purchase time. Can owners opt out? Is a privacy “button” offered and do users use it? What happens when the car is sold again? There is no privacy paperwork for used car sales. While smartphones can generate similar data, they do not have access to vehicle systems to generate it directly. Are there privacy issues for data from vehicle cameras? How about collecting data from underage drivers, who may be covered under different privacy rules? To improve privacy one can employ federated learning where the algorithm is split and sends less data for central processing. We can employ privacy lessons from other domains such as IoT and medical, but how do we leverage them? We should also investigate role-based access. We must consider international legal/regulatory issues, for example in some countries the VIN or plate numbers are PII. Laws lag technology, so we must learn how to anticipate legislation/regulation. Perfect anonymization is impossible, so we must explore the trade-off between obfuscation and utility. The community needs an investigation of a marriage between encryption and differential privacy. V2V has privacy as well as cybersecurity issues.

Industry Participation

We must work with OEMs for access to datasets. This is challenging given the closed culture of OEMs. Will it continue? Researchers may need to adapt to this culture, and we should invite more industry representatives at future workshops. Industry may be more advanced than researchers realize, which we don't know due to their secretive culture. Industry has different priorities and incentives, and researchers should take this into account. Industry is more than OEMs, there are suppliers, Tier-1s and telematics providers. Information companies such as Google are also players in this space. Ideas are currency and open standards help everyone. Researchers need to have clear and specific asks from industry, else there will be hesitation to provide anything. NDAs will most likely be necessary and research may be restricted to the pre-competitive space. We need to separate data used for sales from useful data and take advantage of open data and open APIs offered by some OEMs. Many OEMs are adding gateways to control data streaming out of cars. Heavy trucks are different from cars due to the openness of J1939, which influences the ask from industry.

Key Findings

The workshop produced key findings regarding the critical need for data to drive research, potential applications around cybersecurity as well as transportation and smart communities, the importance of working with industry, and the benefits of working as a community.

Need for Datasets

The workshop participants emphasized that there is a critical need for data to drive research and the current lack of data is an impediment. The type of data needed is driven by research objectives. For example, in-vehicle data is needed to drive cybersecurity research and other in-vehicle applications. Such data may include CAN, Ethernet, and other data from in-vehicle networks such as analog data for IDS research. AI and ML research needs sensor, camera, object segmentation and other data, frequently in its raw format. Combinations of synchronized data sources such as CAN, sensor, and both inward and outward video are also highly desirable. Finally, external data such as V2V, V2I and other transportation and community data is needed for smart transportation and smart community research.

Standardization was identified as another important factor for research data. This includes:

1) Standardization in data collection methods, including detailed documentation and metadata describing all aspects of the collection;
2) Standardization in naming to enable research and third party applications;
3) Standardization in defining realistic cyber attacks to enable evaluation and comparisons across cybersecurity and privacy research; and
4) Standardization in data handling, including best practices to securely collect, archive, discover and distribute trustworthy data.

Given the rapid proliferation of EVs there is a strong need for EV data and their ecosystem. This includes data from EVs themselves and charging station data to guide deployment depending on driver, grid, city, and community needs.

At the application level, data is needed to develop new applications that benefit the occupants of vehicles by enhancing safety and convenience, and data to drive vehicle applications such as service and maintenance.

Privacy is a critical element and could be achieved through consent, aggregation, privacy-preserving data techniques, scrubbing, and obfuscation. To effectively handle privacy, the community needs best practices guides and easy-to-use privacy tools that offer various levels of anonymization.

Finally, the community lacks good tools in various important areas. Examples include decoding tools (to decode CAN signals), robust tools for AI/ML, tools to analyze and visualize data, tools to protect data, and privacy tools.

Potential Applications

At the application space, cybersecurity took front stage. Applications include IDS and IPS, secure gateways, and techniques for authentication, encryption, and key management. These should be developed with an eye on safety, which necessitates real-time, reliable, and robust communication. Similarly important were ADAS applications to enhance safety, including applications that detect driver fatigue. Forensics was also important to aid in accident reconstruction and to assist law enforcement. These were followed by comfort and convenience applications such as proactively turning on heated seats in cold weather. For extra-vehicle applications, it was emphasized that secure V2V and V2X communication is critical for a strong foundation to build applications. Other desirable applications include vehicle repair and maintenance, cost comparisons between EVs and ICE vehicles, and new insurance models. Finally, many transportation applications are possible, such as a vehicle AirBnB, determining and mapping road conditions and hyperlocal weather. These can be combined to enable many smart and connected community applications.

Working with Industry

Working with industry is an important activity to support research and application development. Industry stakeholders include OEMs, telematics providers, Tier 1 suppliers, insurance, and more. Industry may have data that researchers need through their internal testing, and collection of telematics for fleet management. OEMs and Tier 1 suppliers are a reliable source of signal specifications. However, we need data sharing models that address IP and privacy issues. We also need standards for naming since each OEM may use different naming schemes. These needs necessitate the development of tools to sanitize data and define or standardize signal names.

Working as a Community

Establishing and working as a community around automotive research datasets offers many benefits and should be supported by a collaboration platform.

Establishing a community is beneficial at many levels. It can:

1) Bring together providers and consumers;
2) Identify common data needs, requirements and tools;
3) Develop common data naming formats; and
4) Address privacy and IP concerns.

A community will benefit greatly by a platform to act as a catalyst and establish a data ecosystem.

The platform should offer the following services:

1) A repository of community datasets to eliminate silos;
2) A home for disparate datasets offered by researchers;
3) Tools to support the production and consumption of datasets;
4) Activities to nurture and feed the ecosystem;
5) A search service with standardized, clear descriptions; and
6) A clearinghouse for datasets and tools found in the wild with a first level of vetting.

The platform would enable research in in-vehicle, extra-vehicle, transportation and smart communities applications, and enable researcher teams to pursue funding from government agencies and other sources (e.g., IRAD, industry, foundations). Such a platform could be established through a CCRI or similar grant by the NSF or funding from other government or industry organizations.

Conclusions and Next Steps

The overall goal of the automotive dataset workshop was to initiate a coordinated effort to bring together a community around the development and sharing of robust automotive datasets to foster and support new, open research in areas with strong societal impact such as smart and connected communities and development of cybersecurity and privacy protections for automotive applications. Over the course of the two-day workshop, participants not only learned about current automotive research datasets available to support automotive cybersecurity, but also explored future needs for automotive datasets as well as applications that could benefit from research and development supported by such datasets.

Given the number of scientists, engineers, and technologists from across academia, industry, and government who participated in the workshop, the wide-ranging presentations, and productive working session discussions, it was clear there is strong community interest in producing and using automotive datasets to drive future research.

Participants included researchers producing and/or using automotive datasets in their research, commercial vehicle telematics providers who are willing to share data with researchers, other industry representatives interested in collaborating with researchers to support and benefit from their work, and leaders from standards organizations interested in developing common interfaces and data formats needed to support automotive datasets and their application. There were also representatives from funding agencies and other government organizations interested in a robust research ecosystem to develop new, innovative automotive applications that benefit society.

Presentations highlighted research and other efforts to produce automotive datasets and/or use datasets in support of the research; commercial vehicle telematics providers willing to share data with researchers; and standardization efforts. The NTRC virtual tour highlighted a number of key facilities that have been instrumental in their transportation technology research. The working sessions identified future dataset needs, including desired characteristics, dataset quality and trustworthiness, and potential sources of datasets; and potential applications of automotive datasets, including analytic tools, sharing infrastructure and frameworks, privacy considerations, and industry participation.

A coordinated effort around automotive research dataset offers many benefits to the community. It can help coordinate existing isolated efforts; facilitate the exchange of knowledge and resources; encourage, nurture, and sustain ongoing conversations; and stimulate active research collaborations among users and producers of automotive datasets. The community can engage industry, including automotive manufacturers, suppliers, and other important partners. It can engage relevant standards bodies and applicable government organizations. Together, all of these stakeholders can form a robust ecosystem that works to develop, share, and exploit community resources, including automotive research datasets, facilities, and other

capabilities. This, in turn, will enable the research community to collectively address important problems, define high quality research initiatives, and develop new, innovative applications to benefit society.

A coordinated community around automotive datasets can offer its members a number of activities and services. These include dataset discovery, collection, curation, distribution; communicating datasets needs and requests; tools to normalize and apply datasets; and development of privacy best practices and other important policies, including assisting researchers with Institutional Review Boards (IRBs). The community can also facilitate education, training, and outreach to community members and others, in order to sustain and grow. This includes working with related events, such as the CyberAuto and CyberTruck challenges, to both promote the community and recruit members. It also includes interacting with the NSF research community, including researchers and others participating in the S&CC and Civic programs. Finally, the community can facilitate involvement with standards bodies such as W3C, COVESA, and others, and with government organizations, such as NSF, DOT, DHS, and others.

An especially important community activity is engaging a broad set of industry stakeholders, including OEMs; Tier 1 ECU and system suppliers; Tier 2 component, software, and connected app or service suppliers; startups; and telematics companies; as well as insurance providers and other application providers. The community would seek to engage industry stakeholders as both producers and consumers of datasets, to provide research drivers and applications, and to work collaboratively as research partners.

Next Steps

The workshop demonstrated a strong need for a coordinated effort to bring together a community around automotive datasets. To help support such a community, several workshop organizers plan to stand up an open collaboration platform at the University of Memphis. The platform will provide a central catalog and clearinghouse for current and anticipated future automotive research datasets, as well as provide other services that will help catalyze our efforts and provide a means for community engagement and interaction.

Finally, the workshop was the first in what will hopefully be an ongoing series of discussions and interactions and the start of a new, vibrant community centered around the production and use of automotive research datasets. We plan to hold additional workshops, likely on an annual basis. We will target holding the next workshop in November 2022 and hope to learn about new datasets, facilities, research initiatives, and applications. We anticipate a similar mix of scientists, engineers, and technologists from academia, industry, and government who represent various interests, including research, telematics, standards, transportation, insurance, smart communities, and others. We will build on current participants, but plan to expand participation with additional researchers, industry representatives, and government officials.

Acknowledgements

The workshop organizers thank all of the presenters who shared their research and other work regarding automotive datasets and potential application areas. The organizers also thank everyone who took the time to attend the workshop and for their open and collegial participation.

References

[Anderson2020] Anderson, R., Monitoring and Metering, in Security Engineering, A Guide to Building Dependable Distributed Systems, 3rd Edition, December 2020, Wiley, https://www.cl.cam.ac.uk/~rja14/Papers/SEv3-ch14-dec20.pdf

[Baretto] Cephas Baretto Dataset, https://www.kaggle.com/cephasax/obdii-ds3.

[Baretto2018] Barretto, Cephas Alves da Silveira. Uso de técnicas de aprendizado de máquina para identificação de perfis de uso de automóveis baseado em dados automotivos. 2018. 92f. Dissertação (Mestrado Profissional em Engenharia de Software) - Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal, 2018, https://repositorio.ufrn.br/handle/123456789/26017.

[Bloom2021] G. Bloom. WeepingCAN: A Stealthy CAN Bus-off Attack. 3rd International Workshop on Automotive and Autonomous Vehicle Security, 2021. https://dx.doi.org/10.14722/autosec.2021.23002.

[Bloom-SBA] Gedare Bloom and Sena Hounsinou, Schedule Based Bus off Attack Analysis on CAN, University of Colorado Colorado Springs, https://github.com/Embedded-Systems-Security-Lab/sba-in-can.

[Cao2021] Y. Cao et al., "Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks," 2021 IEEE Symposium on Security and Privacy (SP), 2021, pp. 176-194, https://dx.doi.org/10.1109/SP40001.2021.00076.

[Cao2019] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z. Morley Mao. “Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving”. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19). Association for Computing Machinery, https://doi.org/10.1145/3319535.3339815.

[CAV-SEC] Autonomous Driving (AD) & Connected Vehicle (CV) Systems Security (website), https://sites.google.com/view/cav-sec.

[CCPA] California Consumer Privacy Act (CCPA), https://oag.ca.gov/privacy/ccpa.

[Chen2018] Qi Alfred Chen, Yucheng Yin, Yiheng Feng, Z. Morley Mao, and Henry X. Liu, “Exposing Congestion Attack on Emerging Connected Vehicle based Traffic Signal Control”, In Proceedings Network and Distributed System Security Symposium, NDSS 2018, February 2018, San Diego, CA, Internet Society, http://dx.doi.org/10.14722/ndss.2018.23222.

[COVESA] Connected Vehicle Systems Alliance (COVESA) (website). https://covesa.global.

[COVESA-VSS] COVESA, Vehicle Signal Specification (VSS), https://github.com/COVESA/vehicle_signal_specification.

[CrySyS] CrySyS Lab CAN-Log Infector and Ambient CAN Traces, https://www.crysys.hu/research/vehicle-security/.

[DRP-Attack] Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack (DRP attack) (website), https://sites.google.com/view/cav-sec/drp-attack/.

[Daily] Jeremy Daily, Heavy Vehicle CAN Data: CAN and J1939 Data Collected from Various Vehicles and Operation (website), https://www.engr.colostate.edu/~jdaily/J1939/candata.html.

[Ezeobi2020] U. Ezeobi, H. Olufowobi, C. Young, J. Zambreno and G. Bloom, "Reverse Engineering Controller Area Network Messages Using Unsupervised Machine Learning," in IEEE Consumer Electronics Magazine, vol. 11, no. 1, pp. 50-56, 1 Jan. 2020, https://dx.doi.org/10.1109/MCE.2020.3023538.

[Feng2018] Feng, Yiheng, Shihong Huang, Qi Alfred Chen, Henry X. Liu, and Z. Morley Mao. “Vulnerability of Traffic Control System Under Cyberattacks with Falsified Data.” Transportation Research Record 2672, no. 1 (December 2018): 1–11. https://doi.org/10.1177/0361198118756885.

[FusionRipper] FusionRipper: First Attack on MSF-based AV Localization (website), https://sites.google.com/view/cav-sec/fusionripper.

[GDPR] Complete guide to GDPR compliance, https://gdpr.eu/.

[Geotab] Geotab (website), https://data.geotab.com/.

[Geotab-Ignition] Geotab Ignition (website), https://ignition.geotab.com/.

[Gravwell] Gravwell, https://gravwell.io.

[Guillaume2019] Dupont, Guillaume; Lekidis, Alexios; den Hartog, J. (Jerry); Etalle, S. (Sandro) (2019): Automotive Controller Area Network (CAN) Bus Intrusion Dataset v2. 4TU.ResearchData. Dataset. https://doi.org/10.4121/uuid:b74b4928-c377-4585-9432-2004dfa20a5d

[Hanselm2020] M. Hanselmann, T. Strauss, K. Dormann and H. Ulmer, "CANet: An Unsupervised Intrusion Detection System for High Dimensional CAN Bus Data," in IEEE Access, vol. 8, pp. 58194-58205, 2020, doi: 10.1109/ACCESS.2020.2982544.

[HCRL] HCRL, Hacking and Countermeasure Research Lab, https://ocslab.hksecurity.net/.

[HCRL-Analyzer] HCRL, Hacking and Countermeasure Research Lab Vehicle Data Analyzer, https://ocslab.hksecurity.net/software/vehicle-data-analyzer.

[HCRL-Challenge] HCRL, Cyber Security Challenge, https://sec-challenge.kr/.

[HCRL-DataSets] HCRL, Hacking and Countermeasure Research Lab Datasets, https://ocslab.hksecurity.net/Datasets.

[Hu2020] Shengtuo Hu, Qi Alfred Chen, Jiwon Joung, Can Carlak, Yiheng Feng, Z. Morley Mao, and Henry X. Liu. “CVShield: Guarding Sensor Data in Connected Vehicle with Trusted Execution Environment”. In Proceedings of the Second ACM Workshop on Automotive and Aerial Vehicle Security (AutoSec '20). Association for Computing Machinery, https://doi.org/10.1145/3375706.3380552.

[Hu2021] Shengtuo Hu, Qi Alfred Chen, Jiachen Sun, Yiheng Feng, Z. Morley Mao, and Henry X. Liu, “Automated Discovery of Denial-of-Service Vulnerabilities in Connected Vehicle Protocols” In Proceedings 30th USENIX Security Symposium, USENIX Security 21, August 2021, pp. 3219-3236, USENIX Association, https://www.usenix.org/conference/usenixsecurity21/presentation/hu-shengtuo.

[Huang2020] Shihong Huang, Wai Wong, Yiheng Feng, Qi Alfred Chen, Henry X. Liu, and Z. Morley Mao, Cyber-Vulnerability Analysis for Connected Vehicle Based Traffic Signal Control Systems, Transportation Research Board Annual Meeting (TRB) 2020.

[Hyundai-ODP] Hyundai, Open Data Platform, https://tech.hyundaimotorgroup.com/mobility-service/open-data-platform/.

[iRAP] International Road Assessment Programme (iRAP), https://irap.org/.

[Jia2020] Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei, “Fooling Detection Alone is Not Enough: Adversarial Attack against Multiple Object Tracking”, In Proceedings International Conference on Learning Representations (ICLR) 2020, https://openreview.net/forum?id=rJl31TNYPr.

[KATECH] KATECH, http://www.katech.re.kr/.

[K-Cyber] K-Cyber Security Challenge, http://datachallenge.kr/.

[Mcity] Mcity (website), https://mcity.umich.edu/.

[MSF-ADV] Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks (website), https://sites.google.com/view/cav-sec/msf-adv.

[NHTSA2020] NHTSA, Occupant Protection for Automated Driving Systems, A Proposed Rule by the National Highway Traffic Safety Administration on 03/30/2020, https://www.federalregister.gov/documents/2020/03/30/2020-05886/occupant-protection-for-automated-driving-systems

[NSF-BigIdeas] National Science Foundation, 10 Big Ideas, https://www.nsf.gov/news/special_reports/big_ideas/index.jsp.

[NSF-Civic] National Science Foundation, Civic Innovation Challenge: Powering Smart and Connected Communities (website), https://nsfcivicinnovation.org/.

[NSF-CPS] National Science Foundation, Cyber-Physical Systems (CPS), https://beta.nsf.gov/funding/opportunities/cyber-physical-systems-cps.

[NSF-DatasetsA] National Science Foundation, Dear Colleague Letter: Request for Information on the specific needs for datasets to conduct research on computer and network systems (NSF 21-056), https://www.nsf.gov/pubs/2021/nsf21056/nsf21056.jsp.

[NSF-DatasetsB] National Science Foundation, Computer and Network Systems Research Dataset Needs (website), https://www.nsf.gov/cise/cns/research_datasets/rfi_responses.jsp.

[NSF-Harnessing] National Science Foundation, Harnessing the Data Revolution, https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp.

[NSF-SaTC] National Science Foundation, Secure and Trustworthy Cyberspace (SaTC), https://beta.nsf.gov/funding/opportunities/secure-and-trustworthy-cyberspace-satc.

[NSF-SCC] National Science Foundation, Smart and Connected Communities (S&CC) Program, https://beta.nsf.gov/funding/opportunities/smart-and-connected-communities-scc.

[NTRC-CAVE] National Transportation Research Center, "Connected and Automated Vehicle Environment Laboratory," https://www.ornl.gov/content/connected-and-automated-vehicle-environment-laboratory-cave.

[NTRC-FEERC] National Transportation Research Center, "Fuels, Engines, and Emissions Research Center," https://www.ornl.gov/content/fuels-engines-and-emissions.

[NTRC-PEEM] National Transportation Research Center, "Power Electronics and Electric Machinery Laboratory," https://www.ornl.gov/division/eesrd/power-electronics.

[NTRC-VRL] National Transportation Research Center, "Vehicle Research Laboratory Virtual Tour," https://my.matterport.com/show/?m=gCLYG6qqJjW&sr=-3.02,.89&ss=140.

[NTRC-VSIL] National Transportation Research Center, "Vehicle Systems Integration Laboratory," https://www.ornl.gov/facility/ntrc/research-areas/vehicle-systems.

[NTRC-VSL] National Transportation Research Center, "Vehicle Security Laboratory," https://www.ornl.gov/content/vehicle-security-laboratory.

[NYC-DOT] New York City DOT Connected Vehicle Project, https://www.cvp.nyc/.

[Olufowobi2020] H. Olufowobi, C. Young, J. Zambreno and G. Bloom, "SAIDuCANT: Specification-Based Automotive Intrusion Detection Using Controller Area Network (CAN) Timing," in IEEE Transactions on Vehicular Technology, vol. 69, no. 2, pp. 1484-1494, Feb. 2020, https://dx.doi.org/10.1109/TVT.2019.2961344.

[OpenXC] The OpenXC Platform, http://openxcplatform.com/.

[OPIN] Open Insurance (OPIN). https://openinsurance.io/.

[ORNL-ROAD] Oak Ridge National Laboratory, Real ORNL Automotive Dynamometer (ROAD) CAN Intrusion Dataset, https://0xsam.com/road/.

[Papadop2021] Papadopoulos, C., and D. Balenson, The Need for Vehicle Telematics Data to Support Broad Scale Research, Response to NSF 21-056, Dear Colleague Letter: Request for Information on the specific needs for datasets to conduct research on computer and network systems, https://www.nsf.gov/cise/cns/research_datasets/pdf/12678256572_Papadopoulos.pdf.

[Pese2019] Mert D. Pesé, Troy Stacer, C. Andrés Campos, Eric Newberry, Dongyao Chen, and Kang G. Shin. 2019. LibreCAN: Automated CAN Message Translator. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19). Association for Computing Machinery, New York, NY, USA, 2283–2300, https://doi.org/10.1145/3319535.3363190.

[ROI-Attack] Fooling Perception via Location: A Case of Region-of-Interest Attacks on Traffic Light Detection in Autonomous Driving (website), https://sites.google.com/view/roiattack.

[Sato2020] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jack Jia, Xue Lin, and Qi Alfred Chen, “Poster: Security of Deep Learning based Lane Keeping Assistance System under Physical-World Adversarial Attack”, Network and Distributed Systems Security (NDSS) Symposium 2020, February 2020, Internet Society, https://www.ndss-symposium.org/wp-content/uploads/2020/02/NDSS2020posters_paper_15.pdf.

[Sato2021] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen, “Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Adversarial Attack”, In Proceedings 30th USENIX Security Symposium, USENIX Security 21, August 2021, pp. 3309-3326, USENIX Association, https://www.usenix.org/conference/usenixsecurity21/presentation/sato.

[Shen2019] Junjie Shen, Jun Yeon Won, Shinan Liu, Qi Alfred Chen, and Alexander Veidenbaum, “Poster: Security Analysis of Multi-Sensor Fusion-based Localization in Autonomous Vehicles”, Network and Distributed System Security (NDSS) Symposium 2019, February 2019, Internet Society, https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019posters_paper_12.pdf.

[Shen2020] Junjie Shen, Jun Yeon Won, Zeyuan Chen, and Qi Alfred Chen, “Drift with Devil: Security of Multi-Sensor Fusion based Localization in High-Level Autonomous Driving under GPS Spoofing”, In Proceedings 29th USENIX Security Symposium, USENIX Security 20, August 2020, pp. 931-948, USENIX Association, https://www.usenix.org/conference/usenixsecurity20/presentation/shen.

[SmartColumbus] SmartColumbus Datasets Curated for Visualization, https://www.smartcolumbusos.com/tools/datasets-curated-for-visualization.

[Sun2020] Jiachen Sun and Yulong Cao and Qi Alfred Chen and Z. Morley Mao, “Towards Robust LiDAR-based Perception in Autonomous Driving: General Black-box Adversarial Sensor Attack and Countermeasures”, in Proceedings 29th USENIX Security Symposium, USENIX Security 20, August 2020, pp. 877-894, USENIX Association, https://www.usenix.org/conference/usenixsecurity20/presentation/sun.

[SRTI] Safety Related Traffic Information (SRTI) Ecosystem (website), https://www.dataforroadsafety.eu/srti-ecosystem.

[SynCAN] ETAS/SynCAN, https://github.com/etas/SynCAN.

[Tang2021] Kanglan Tang, Junjie Shen, Qi Alfred Chen, “Fooling Perception via Location: A Case of Region-of-Interest Attacks on Traffic Light Detection in Autonomous Driving”, In Proceedings Third International Workshop on Automotive and Autonomous Vehicle Security (AutoSec) 2021, February 2021, Internet Society, https://dx.doi.org/10.14722/autosec.2021.23029.

[THEA] Tampa-Hillsborough Expressway Authority (THEA) Connected Vehicle Pilot, https://theacvpilot.com/.

[TUEindhoven] TU Eindhoven Lab Automotive CAN Bus Intrusion Dataset, https://data.4tu.nl/articles/dataset/Automotive_Controller_Area_Network_CAN_Bus_Intrusion_Dataset/12696950/2.

[USDOT-Connected] U.S. Department of Transportation, Using Connected Vehicle Technologies to Solve Real-World Problems, https://www.its.dot.gov/pilots/index.htm.

[USDOT-CVP] U.S. Department of Transportation, Connected Vehicle Pilot (CVP) Open Data, https://data.transportation.gov/stories/s/hr8h-ufhq.

[USDOT-Portal] U.S. Department of Transportation, Public Data Portal, https://data.transportation.gov/.

[USDOT-NYCDOT] U.S. Department of Transportation, New York City DOT (NYCDOT) Pilot, https://data.transportation.gov/stories/s/hr8h-ufhq#new-york-city-dot-(nycdot)-pilot.

[USDOT-THEA] U.S. Department of Transportation, Tampa-Hillsborough Expressway Authority (THEA) Pilot, https://data.transportation.gov/stories/s/hr8h-ufhq#tampa-hillsborough-expressway-authority-(thea)-pilot.

[USDOT-WYDOT] U.S. Department of Transportation, Wyoming DOT (WYDOT) Pilot, https://data.transportation.gov/stories/s/hr8h-ufhq#wyoming-dot-(wydot)-pilot.

[Verma2020] M. E. Verma, M. D. Iannacone, R. A. Bridges, S. C. Hollifield, B. Kay, and F. L. Combs, "ROAD: The Real ORNL Automotive Dynamometer Controller Area Network Intrusion Detection Dataset (with a comprehensive CAN IDS dataset survey & guide)," in arXiv preprint, cs.CR 2012.14600, Dec 2020.

[Verma2021] M. E. Verma, R. A. Bridges, J. J. Sosnowski, S. C. Hollifield and M. D. Iannacone, "CAN-D: A Modular Four-Step Pipeline for Comprehensively Decoding Controller Area Network Data," in IEEE Transactions on Vehicular

Technology, vol. 70, no. 10, pp. 9685-9700, Oct. 2021, https://dx.doi.org/10.1109/TVT.2021.3092354.

[Wong2019] Wong, W., Shihong Huang, Yiheng Feng, Qi Alfred Chen, Z. Morley Mao and Henry X. Liu. “Trajectory-Based Hierarchical Defense Model to Detect Cyber-Attacks on Transportation Infrastructure.” (2019). https://www.ics.uci.edu/~alfchen/wai_trb19.pdf.

[W3C] World Wide Web Consortium (website), https://www.w3.org.

[W3C-VISS] World Wide Web Consortium, Vehicle Information Service Specification, VISS version 2 - Core, First Public Working Draft W3C, 29 July 2021, https://www.w3.org/TR/2021/WD-viss2-core-20210729/.

[WYDOT] Wyoming DOT Connected Vehicle Pilot, https://wydotcvp.wyoroad.info/.

Appendix A: List of Acronyms

5G	Fifth Generation
AD	Autonomous Driving
ADAS	Advanced Driver-Assistance Systems
AI	Artificial Intelligence
Auto-ISAC	Automotive Information Sharing & Analysis Center
API	Application Programming Interface
AV	Autonomous Vehicle
AVTP	Audio Video Transport Protocol
BSM	Basic Safety Messages
CAM	Content Addressable Memory
CAN	Controller Area Network
CAN-D	CAN-Decoder
CAN-FD	Controller Area Network Flexible Data-Rate
CAV	Connected and Autonomous Vehicle
CAVE	Connected and Automated Vehicle Environment Laboratory
CISA	Cybersecurity and Infrastructure Security Agency
CPS	Cyber-Physical Systems
CV	Connected Vehicle
CVII	Common Vehicle Interface Initiative
CCPA	California Consumer Privacy Act
CCRI	CISE Community Research Infrastructure
CISE	Computing and Information Science and Engineering
COVESA	Connected Vehicle Systems Alliance
CSU	Colorado State University
DBC	CAN Bus Database
DNN	Deep Neural Network
DSRC	Dedicated Short Range Communication
DVIR	Driver-Vehicle Inspection Report
DOT	Department of Transportation
ECD	Electronic Control Device
ECU	Electronic Control Unit
ELD	Electronic Logging Device
EPA	Environmental Protection Agency
ESSL	Embedded Systems Security Lab
EU	European Union
EV	Electric Vehicle
FEMA	Federal Emergency Management Agency
GDPR	General Data Protection Regulation
GPS	Global Positioning System

HAZMAT	Hazardous Materials
HCRL	Hacking and Countermeasure Research Laboratory
HIL	Hardware-In-the-Loop
HIPAA	Health Insurance Portability and Accountability Act
ICE	Internal Combustion Engine
IDPS	Intrusion Detection and Protection System
IDS	Intrusion Detection System
IFTA	International Fuel Tax Agreement
IMU	Inertial Measurement Unit
IoT	Internet of Things
IP	Intellectual Property
iRAP	International Road Assessment Programme
IRB	Institutional Review Board
ITS	Intelligent Transportation Systems
LIN	Local Interconnect Network
MAC	Medium Access Control
ML	Machine Learning
MSF	Multi-Sensor Fusion
NDA	Non- Disclosure Agreement
NHTSA	National Highway Traffic Safety Administration
NIST	National Institute of Standards and Technology
NPRM	Notice of Proposed Rulemaking
NSF	National Science Foundation
NTP	Network Time Protocol
NTRC	National Transportation Research Center
NTS	NTP over TLS
NYC DOT	New York City Department of Transportation
OBD	On-Board Diagnostics
OBU	On-Board Unit
OEM	Original Equipment Manufacturer
OPIN	Open Insurance
ORNL	Oak Ridge National Laboratory
OTIDS	Offset Ratio and Time Interval based Intrusion Detection System
PEEM	Power Electronics and Electric Machinery Laboratory
PID	Parameter Identifier
PII	Personal Identifiable Information
PTP	Precision Time Protocol
RDT&E	Research, Development, Test, and Evaluation
ROAD	Real ORNL Automotive Dynamometer
ROI	Region-of-Interest
RtR	Right to Repair
SaTC	Secure and Trustworthy Cyberspace
S&CC	Smart and Connected Communities

SDN	Software Defined Network
SPaT	Signal Phase and Timing
SRTI	Safety Related Traffic Information
THEA	Tampa-Hillsborough Expressway Authority
TIM	Traveler Information Messages
TLS	Transport Layer Security
UCCS	University of Colorado Colorado Springs
UDS	Unified Diagnostics Service
U.S.	United States
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VIN	Vehicle Identification Number
VISS	Vehicle Information Service Specification
VRL	Vehicle Research Laboratory
VSIL	Vehicle Systems Integration Laboratory
VSL	Vehicle Security Lab
VSS	Vehicle Signal Specification
W3C	World Wide Web Consortium
WYDOT	Wyoming Department of Transportation

Appendix B: Workshop Invitation

From: David Balenson <david.balenson@sri.com>
To: David Balenson <david.balenson@sri.com>
Cc: David Balenson <david.balenson@sri.com>, Christos Papadopoulos <christos.papadopoulos@memphis.edu>, Glenn Atkinson <glennatkinson@geotab.com>, Ted Guild <tedguild@geotab.com>, Stacy Prowell <prowellsj@ornl.gov>, Sam Hollifield <hollifieldsc@ornl.gov>
Subject: INVITATION: Workshop on Future Automotive Research Datasets

Dear Colleague,

As vehicles are becoming more connected and autonomous, telematics and other data from such vehicles is critical to support research and building applications for the vehicles themselves as well as their environment. Such datasets are scarce and limited at best, partly due to the difficulty in collecting them and the privacy considerations that accompany them. Unlocking the vast potential of vehicle applications requires open availability of diverse, high-quality datasets.

As a researcher in the community who is producing or using automotive research datasets, the University of Memphis, Geotab, SRI International, and Oak Ridge National Laboratory (ORNL) invite you to participate in the workshop, **Paving the Road to Future Automotive Research Datasets: Challenges and Opportunities, to be held as a hybrid in-person/virtual event at the National Transportation Research Center (NTRC) at ORNL on Thursday-Friday, November 18-19, 2021 (10am-4pm ET)**.

The goal of the workshop is to initiate a coordinated effort to bring together a community around development and sharing of robust automotive datasets to support open research in areas with strong societal impact such as smart and connected communities and development of new, innovative cybersecurity and privacy protections for automotive applications.

The workshop will bring together researchers like yourself who are either producing automotive datasets or are interested in using automotive datasets to drive their research; commercial vehicle telematics providers who are willing to share data with researchers; and representatives from funding agencies. Some of the topics we plan to cover include:

- Applications of vehicle telematics and other data
- Geotab telematics technology and datasets available to researchers
- ORNL/NTRC research and facilities including dynamometers, that can be used to produce datasets
- Community research datasets (ORNL, HCRL, Bosch, etc.)
- Proposed efforts to build a sharing infrastructure and framework
- Privacy considerations

Additionally we will engage participants in facilitated discussions to learn more about future dataset needs, potential applications of automotive datasets, and next steps.

The workshop will be held as a hybrid in-person/virtual event at ORNL's NTRC (https://www.ornl.gov/facility/ntrc), located in the Oak Ridge/Knoxville, TN area. Participants are welcome to attend the workshop in person or to participate virtually via teleconferencing.

Please RSVP to David Balenson <david.balenson@sri.com> to let us know if you and/or others from your team can participate and whether you plan to attend in person or virtually. Additional information, including a detailed agenda, will be distributed prior to the workshop.

We greatly appreciate your consideration of this invitation and look forward to your participation in the workshop!

Sincerely,
David Balenson (SRI International)

On behalf of the workshop team:

- Christos Papadopoulos (University of Memphis)
- Glenn Atkinson (Geotab)
- Ted Guild (Geotab)
- David Balenson (SRI international)
- Stacy Prowell (ORNL)
- Sam Hollifield (ORNL)

Appendix C: Workshop Agenda

Paving the Road to
Future Automotive Research Datasets: Challenges and Opportunities
VIRTUAL WORKSHOP

November 18-19, 2021

AGENDA

Thursday, November 18, 2021 (all times EST)	
09:45-10:00	**GATHERING**
10:00-10:30	**Welcome, Introduction, and Background (Focus on Datasets)**
10:30-12:00	**Presentations - Datasets** ● ORNL Research Efforts, including ROAD dataset, CAN-D, etc. - Sam Hollifield and Stacy Prowell (ORNL) ● HCRL Research and Datasets - HuyKang (Hugo) Kim (Korea University/HCRL) ● CAN Dataset Survey - Md Hasan Shahriar and Wenjing Lou (VATech)
12:00-12:30	**BREAK**
12:30-14:00	**Presentations - Datasets and Standards** ● Geotab Data Product Discovery - Paul Maida (Geotab) ● Heavy Truck CAN Dataset Collection - Jeremy Daily (CSU) ● W3C and COVESA CVII - Ted Guild (Geotab) and Gunnar Andersson (COVESA)
14:00-14:15	**BREAK**
14:15-15:30	**Breakout Session #1: Future Dataset Needs** https://drive.google.com/drive/folders/10QeCm8ZfgjjW0REKjSRsYh3WXpv4tM2n?usp=sharing ● Desired characteristics of rich, robust automotive datasets ● CAN, other in-vehicle networks, sensor, other data (e.g., driver cam) ● Real vs. synthetic datasets, attack generation, etc. ● Dataset quality and trustworthiness ● Sources of datasets (ad-hoc, community, industry, government, real-world pilots, reference datasets)
15:30-15:45	**Breakout Session #1 Debriefs**
15:45-16:00	**Wrap-up**
16:00	**ADJOURN**

Friday, November 19, 2021 (all times EST)	
09:45-10:00	**GATHERING**
10:00-10:15	**Welcome, Recap, Overview (Focus on Applications)**
10:15-11:00	**ORNL National Transportation Research Center (NTRC) Virtual Tour**
11:00-12:00	**Presentations - Security Research** ● IDS for CAN - Md Hasan Shahriar and Wenjing Lou (VATech) ● Towards Secure & Robust AI Stack in Autonomous Driving & Beyond - Qi Alfred Chen (UC Irvine)
12:00-12:30	**BREAK**
12:30-12:45	**NSF Smart & Connected Communities (S&CC) and Cyber-Physical Systems (CPS) Program Overview - David Corman (NSF)**
12:45-13:45	**Lightning Talks** ● Model-Based Intrusion Detection - Qadeer Ahmed (Ohio State) ● Trajectory Prediction and Drivable Space Detection - Ruiyang Zhu and Qingzhao Zhang (U. Michigan) ● Automotive Security at UCCS - Gedare Bloom (U. Colorado at Colorado Springs) ● CyberAuto and CyberTruck Challenges - Jeremy Daily (CSU) ● Collaboration Towards Open Risk Models - Jim Davis (Geotab) ● Open Insurance Project - Kumar Maddali (OPIN)
13:45-14:00	**BREAK**
14:00-15:15	**Breakout Session #2: Potential Applications** https://drive.google.com/drive/folders/1XnTQRdqgiSlYpKt6xCNiFDN1rbtmGxSR?usp=sharing ● "Killer" apps for automotive datasets ● Current and future analytic tools for processing automotive datasets ● Sharing infrastructure and frameworks for datasets ● Privacy considerations ● Industry participation
15:15-15:30	**Breakout Session #2 Debriefs**
15:30-15:45	**Building a Community Around Automotive Research Datasets - Christos Papadopoulos (U. Memphis)**
15:45-16:00	**Wrap-up and Next Steps**
16:00	**ADJOURN**

Appendix D: Workshop Presentations

Introduction and Background - David Balenson (SRI International)
Slides:
https://docs.google.com/presentation/d/17BHXPjK8Ofp102XStGfA7pvy1AqGGc3lGyzgRKUzqnc/edit?usp=sharing
Video recording:
https://drive.google.com/file/d/1v55AykHEvP6w3R6pwKD8bUsFmYz5l4zv/view?usp=sharing

ORNL Research Efforts, including ROAD dataset, CAN-D, etc. - Sam Hollifield and Stacy Prowell (ORNL)
Slides:
https://drive.google.com/file/d/1hE3lH6I_PUk-anuKBPa6wzTpgNNx3pj7/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1TKpOIBbqT2PtfjdbFH0YY5-LbBW-Mjp7/view?usp=sharing

HCRL Research and Datasets - HuyKang (Hugo) Kim (Korea University/HCRL)
Slides:
https://drive.google.com/file/d/1DHFg05Vkl9YzgyLBySbRs-VpqZezRBmk/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1RsfINxZpd4QzWShr3vExjvfwErTyGKZV/view?usp=sharing

CAN Dataset Survey - Md Hasan Shahriar (VATech)
Slides:
https://drive.google.com/file/d/1Gg14B4De_5k5C4ZFmeiTP0JSv8GPUuqf/view?usp=sharing
Video recording:
https://drive.google.com/file/d/17Kt0gejEsGBc3tlhXOONMchhhVxjiwNQ/view?usp=sharing

Geotab Data Product Discovery - Paul Maida (Geotab)
Slides:
https://drive.google.com/file/d/1YgK7VuoepmRaZxZcVM7_zu0lRqYJER8S/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1l-ZrUs5mYDpnYgcYzbo1hX1QCQGl_jEa/view?usp=sharing

Heavy Truck CAN Dataset Collection - Jeremy Daily (CSU)
Slides: https://drive.google.com/file/d/1i92tU_YITiVPTZASM15jNkga2sJJL9hr/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1TK06RpDkDBvpoptS7grPsz8M4EAymhF5/view?usp=sharing

CVII Overview - Ted Guild (Geotab)
Slides:
https://drive.google.com/file/d/1NImBKJWlYwyFjcgxtlsXLAQkQmR-vi7d/view?usp=sharing

Video recording:
https://drive.google.com/file/d/1u9EMu8aAvIYBOIh9N_Bhu85FdGeyL8CH/view?usp=sharing

Connected Vehicle Systems Alliance and CVII - Gunnar Andersson (COVESA)
Slides:
https://drive.google.com/file/d/1h_6pZzXKuUXfAA5FeJqQRwTur9W3IrwG/view?usp=sharing
Recording:
https://drive.google.com/file/d/1u9EMu8aAvIYBOIh9N_Bhu85FdGeyL8CH/view?usp=sharing

Recap and Background - David Balenson (SRI International)
Slides:
https://docs.google.com/presentation/d/1p6gGZ2Es2gZ-VN11n7hOGslf7Ha-4tQsZ15pWprMj5M/edit?usp=sharing
Video recording:
https://drive.google.com/file/d/1Jd7awxHFmkeoihEACStKeGXZN_DvZ1EX/view?usp=sharing

ORNL National Transportation Research Center (NTRC) Virtual Tour
Slides: https://drive.google.com/file/d/11Dz7Q5iMTtjqQrjNaxXUifIr6u63KJRX/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1-bSYsKst2Uc9vL1goys0heLp8HGajYB4/view?usp=sharing

CANShield: An Intrusion Detection Framework for Controller Areas Networks - Md Hasan Shahriar (VATech)
Slides:
https://drive.google.com/file/d/1aEVs4W4L2DXRK0CX0e_Zn1fa0PQz7q71/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1FdQwJM80sFsSwkkMImnEZe5fQgLQbCD9/view?usp=sharing

Towards Secure & Robust AI Stack in Autonomous Driving & Beyond - Qi Alfred Chen (UC Irvine)
Slides:
https://drive.google.com/file/d/1ZxeqioHs5DCYaV_DJrgnufSyOlWq7LaH/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1rRwmNZEEBbAeFPJofuK4K56P8YArX-Jo/view?usp=sharing

NSF Smart & Connected Communities (S&CC) and Cyber-Physical Systems (CPS) Program Overview - David Corman (NSF)
Slides:
https://drive.google.com/file/d/17FH8hYpwktOkMLPMLnLhdCm-HM2WyQfw/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1uMZMLjHp2vNwea7rSejSACB99QOst4hN/view?usp=sharing

Model-Based Intrusion Detection - Qadeer Ahmed (Ohio State)
Slides:
https://drive.google.com/file/d/10ZhPyN2rjINeDzcEY25ezJxeo_POs25e/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1Nzucskkx_VH6ww0NZ1kJ2yqZZNEOlYw2/view?usp=sharing

Trajectory Prediction and Drivable Space Detection - Ruiyang Zhu and Qingzhao Zhang (U. Michigan)
Slides:
https://drive.google.com/file/d/1rYwx0wG6NKwv88HIN9yMXhIUOunCvTxE/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1nmxIdupGYAA7dil7hufswLlKe5Bf2H9b/view?usp=sharing

Automotive Security at UCCS - Gedare Bloom (U. Colorado at Colorado Springs)
Slides:
https://drive.google.com/file/d/1aoOQPWSKWqFYSOB3mrff4c5UdJfcKSEx/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1zcXeKlmWBiGUa27iwTs4wDym_uqHC33R/view?usp=sharing

CyberAuto and CyberTruck Challenges - Jeremy Daily (CSU)
Slides:
https://drive.google.com/file/d/1-dw-g4CmXhDoNEfpt1n8CtuZ1KrXd7wV/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1TIF-cJRVmwyP9uPVVd--O-4qHz7Y2hWl/view?usp=sharing

Collaboration Towards Open Risk Models - Jim Davis (Geotab)
Slides:
https://drive.google.com/file/d/1tGYVfbAWEkGoJQ8RCdTjs8RD96dTDgtG/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1-GPvbi3oHKMRedv5UoroG6u6d9n9xfVz/view?usp=sharing

Open Insurance Project - Kumar Maddali (OPIN)
Slides:
https://drive.google.com/file/d/1s5YxuWCnqfhrXPq3x8kFzQ9oNrgpBO1X/view?usp=sharing
Video recording:
https://drive.google.com/file/d/1icEpAy-aWiXhMsOTYcLHsRb9cX6KXB8B/view?usp=sharing

Building a Community Around Automotive Research Datasets - Christos Papadopoulos (U. Memphis)
Slides:
https://docs.google.com/presentation/d/1PJrUGTbyC0j4uBMPAbwi4I1RBXn5HcW1a_75eVV1sdg/edit?usp=sharing
Video recording:
https://drive.google.com/file/d/1PI2aLOe_5Ez_fP8ss8Icv2BJcEMFfyDY/view?usp=sharing

Appendix E: Breakout Sessions #1 - Future Automotive Datasets

The first working session was held to explore needs for future automotive datasets. The participants were divided into four breakout groups for this session. All of the groups were given the following suggested topics to stimulate the conversation.

- “Killer” apps for automotive datasets
- Current and future analytic tools for processing automotive datasets
- Sharing infrastructure and frameworks for datasets
- Privacy considerations
- Industry participation

The actual topics addressed varied based on the members of the group and were much broader, as captured in the notes. This appendix provides consolidated notes capturing the discussions from the four breakout groups.

Desired Characteristics of Datasets

Need for Diversity in Automotive Datasets

- Data from all types of vehicles including hybrid and Electric Vehicles (EVs)

Need for dataset structure

- Flat files
- Hierarchical
- NoSQL databases
- JSON

Sampling Methods

- Time based
- Curve logic (Geotab’s open-source summarization algorithm)
- Interrupt based (all data)
- Accuracy, precision, quality, sampling methodologies: all need to be taken into consideration

Metadata: What Should be Included?

- Data should be labeled with metadata
- Metadata should be sufficient to understand any data limitations or missing data
- The test setup should be described in great detail, and include photos, descriptions, and videos, if possible
- The source of information must be clearly described
- The collection context must be clearly described:
    - Is the source a controlled environment (testing has control over signal inputs)?
        - From a dynamometer

        - Simulated (synthetic) data
        - Benchtop communication only
        - Benchtop with signal inputs
        - Researchers should be in control of the safety environment
    - Is the source an uncontrolled environment (testers do not control signal inputs)?
        - Urban environment
        - Over the road
        - Diagnostics
        - Challenge events
        - In this case, researchers should avoid physically dangerous activity
    - Are attacks being launched during data collection?

Type of Data Needed

- Depends on use cases - for example:
    - CAN protocol attacks - bit level voltages may be needed
    - Data collection for “normal” driving requires application-level CAN data; however, the structure of CAN data frames must be kept
    - Diagnostics require transport layer messages
    - Both CAN signals and network traffic are needed
    - Aggregated data is needed for fleet utilization (this implies select signals, not all traffic)

Updating Datasets

- Datasets should not be static, but mechanisms to update and augment data are required

Privacy Preservation

- Different sources may require different privacy-preserving techniques
- Researchers must understand how to use data in the face of privacy and security controls

Replaying Data

- We must understand how to properly replay collected data
- Need a notion of actual real-time

Decoding CAN data

- The ability to decode (e.g., CAN hex) / understand the data is an immediate challenge
- Current solutions (e.g., CAN-D [ORNL-ROAD], LibreCAN [Pese2019]) are not perfect

Characterization of Datasets

- The number of data points available and various other variables can influence which AI learning models can be used
- What data is not being collected/missed that could have contributed to answering research questions?
- How appropriate is the collected data to a given problem?

Temporal Aspects

- Important states/actions that influence data might be missed if we are merely collecting data at fixed time intervals
- How do we merge disparate but related data sets?

Efficiency of Attaining or Loading/Using Datasets

- Existing datasets don't contain very sophisticated attacks
    - There are demonstrated attacks that cause dramatic effects on the car, but don't show the event timeline
    - Need to establish causality of events - what were the series of events that happened, and who caused them (driver, attacker, other)

Labeling

- Labeling is very important in the context of attacks because an attack might interfere with signals further down the road. What is the original signal in the attack?
- Should we label the driver while collecting so we can fingerprint drivers?

Imbalanced Data

- There is an imbalanced data problem: some states of the car are way more frequent (like driving forward as opposed to reverse)
- Imbalance of data availability impacts ML

Augmenting Sensor Data from External Sensors

- Good for noise cancelation

Consistency

- How do we describe CAN data consistently?
    - What signals/units?
    - Do we preserve arbitration?
    - Do we preserve signals in the dataset?

Understandability Requires Standardization

- Using a standardized method to describe data is important
- Applying standardization widely means less issues down the hierarchical structure of CAN
    - There is some layer of the process where we can translate to a standard
- Must contain environmental variables to cover all bases
    - Driving in standard mode
    - While under attack
    - We need data on mechanical failures to differentiate between failure and attack
- How to standardize autonomous Vehicle Data

Real vs. Simulated Data

- How does the simulated data compare to the real data?
- How do we gauge different environments (e.g., fully synthetic virtual CAN vs dSpace hardware-in-the-loop)?

CAN Data Steward

- A responsible steward of CAN data and translation is needed
    - Understandability: impose a standardized method to describe data
    - Partner with OEMs to work with security researchers on CAN translation to avoid future problems on consistency

Clear Copyright and Licensing Information

- If the dataset was produced with proprietary information that limits its use, this needs to be documented up front

CAN and Other In-Vehicle Data

Collection Requirements

- Correlation - time synchronization (critical for correlating data)
- Regulation-driven requirements
- Requirements around forensics and data integrity of forensic data
- Additional networks and data from system integration and upfitting

Need Datasets Explicitly for RADAR/LIDAR/Other Autonomous Features

- Datasets with intrusion detection as well or other attack vectors
- Companies such as comma.ai may provide them, but they are prohibitively expensive
- Other in-vehicle networks including sensor, and other data (e.g., driver cam)

Is There Still a Need to Convince OEM's that CAN is Vulnerable?

- Should efforts be focused on solving the CAN security issues or invent a new, secure solution?
    - E.g., communication buses that work in a secure way
    - CAN might be too entrenched/cheap

Data Through OEMs

- OEMs collect data internally for their own purposes, so what would it take to get industry and researchers to collaborate?
- There are difficulties in this process: for example, CAN databases that map CAN signals tend to change and only OEMs maintain them
- There are also IP issues that may prevent OEMs from sharing such data
- Collaborative OEM projects such as those happening in Mcity, may or may not be willing to share
- Can we work w/ OEMs/suppliers to get shareable datasets?

CAN Voltage Levels

- Existing CAN automotive datasets do not provide this information
    - Can we get sample voltage sets from OEMs?
- Such information is not normally captured in the data, yet companies want it
    - Information should include clock skew and variability between electronic components
- Is it worth capturing such data?
    - There are many difficulties, analog data can provide essentially infinite samples, so how much to capture?

Naming Standards for Signals at the CAN Level

- Needed to decipher the signals and their functionality
    - Difficult at the moment: engineers use tools to determine new signal values
    - For example, when an engineer wants a new signal, a design tool provides a new, unused value without any coordination with other OEMs
- Some signals such as diagnostics, however, have been standardized, so there is a win
- It is not clear that the introduction of Automotive Ethernet will change things since traditional CAN is still needed
- Will efforts such as Autosar bring some standardization?

Accuracy of Existing Signal Decoding Tools
- Tools such as CAN-D are not accurate enough if research requires complete signal specification

Right to Repair (RtR)
- Does RtR mean opening up signals?
- The opposite may be true
    - Consumers will get as much visibility as needed through access to service manuals that document Unified Diagnostics Services (UDS)
- Legal fight over RtR that has degenerated to a fight for the right to data
    - Colored the effort leading to distrust on all sides with many legal fights brewing

Real vs. Synthetic, Attack Generation, Etc.

General
- Are signals in a dataset generated with lab equipment (i.e. function generators)?
- Are these natural environments?
- Need a taxonomy of existing attacks
- Role of AI in attack generation
    - Echoed by the community, but few publications in this space
    - Reliability of synthetic datasets generated by AI models

Quality vs. Quantity
- Use crowdsourced data systems - smaller quantity but higher quality
- Value of aggregated data such as that provided by Geotab - utility depends on the purpose
    - can be very valuable for the appropriate questions

Unintended Results
- Reverse engineering may need resets in between evaluations
    - How do you handle unintended cached data in experiments?
    - Data may be corrupted for subsequent findings
    - What are the best practices on injecting and recording data?

System design issues
- A time series approach might be too constraining, even for identifying events
- Should we capture CAN IDs, events, and causality?

Data Sources from Industry and Research Labs
- Industry working towards a consistent way to capture the CAN status of the car (they are increasing the capacity of the on-board unit (OBU) to get snapshots of all the messages on the car) - could that data be shared?

Safety
- There is a need for a safe environment to perform attacks for real data
    - We can assume OEM's have access to this environment, so working with them under their supervision to enhance security is important
- Other dynamometer setups such as those at NTRC can provide safe environments for enhanced attack data simulation

- Need safety guidelines for data collection and injection in cybersecurity research on vehicles as some situations, replaying signals or fuzzing for example, can be quite dangerous when a vehicle is in motion

Dataset Quality and Trustworthiness

Quality

- Automotive datasets can be very large - how do we avoid duplicating uninteresting data?
- Understand what's next and what are the gaps that need to be filled
- Build a test matrix and work on filling it
- What datasets are trustworthy? Should we trust datasets from OEMs if we can get them? What is an example?
- We need a recipe for collecting and archiving, which may include:
    - Digital signatures
    - Root of Trust
- Driver data: what do we need?
- Data from additional sensors
    - Microphones can measure various vibrations in and around the vehicle, multiple sensors can be triangulated to provide directional awareness
    - Video or more appropriate inferences on video (for example the driver yawning that can signal driver fatigue) made on the vehicle instead of sending video to cloud
- Smartphone data
    - Insurance companies already augment data from smartphones. What else can we get?
- Black boxes exist for forensics purposes already - how do we take advantage of them?
- V2V communication can be also collected

Cross Industry Interests

- Insurance
- Payments (fraud mitigation)
- Traffic management
- Weather
- Road conditions/construction and maintenance needs

Vehicle Lifecycle

- Delivery (to dealer, owner)
- Registration
- Insurance
- Service
- Citations
- Accidents
- Is there a lifetime that datasets have in terms of relevance, and if yes, what is the retention time?

Intrusion Detection System

- Hard to discover the full meaning behind an attacker sending a CAN packet in some cases, such as logic not fitting existing rule base
- Should we filter out integrity check, frequency, class?
- Each ECU has its own logic to flag an event like an airbag deployment - how do we discover and document it?
- For ML purposes they have to add dimensions like production date, number of ECUs etc., to capture that diverse nature of signaling

Difficulties in Developing an Embedded IDS at Hyundai

- Using if/then rules is very complex - we must use well defined rules for two reasons:
- In trying to apply a team learning or using ML, we must first filter out bad messages through an integrity check. A common approach is to analyze CAN message arrival, frequency, and clock skew, which are good features, in real world attacks are not feasible because each ECU has a defense logic.
- When focusing on signal analysis the age of the car is important - as each electronic device gets older it adds noise and changes the statistical properties of a signal. Typically, over time, the speed gets slower.

Other Precautions

- If data is acquired by reverse engineering signals or attack emulation, one must be careful to reset every ECU, else messages remaining in an ECU may affect the next experiment
- We need to establish a standard protocol to generate datasets
- To keep the experiment safe, we need to provide minimum guidelines - e.g., don't send this specific message during driving modes

Questions on Trustworthiness

- What makes a dataset untrustworthy?
- What makes a dataset fair and trustworthy?
    - Are we getting the right bits on a lower level
    - On a higher level are we interpreting it correctly
- Removing bias from the dataset for a more representative sample - especially aggregated datasets
- How do we continuously keep up data quality
- How to build secure comms systems in first place → trust data in systems and when sent externally
    - Correct and authentic, obfuscating data in a way that preserves the meaning but shields OEM secrets
- More collaborations and more researcher eyes increase likelihood to find bad data, communicate to others
    - Benefit of community
    - Similar argument for open-source software security
- If we need to obfuscate a dataset, we must do it in way that we trust to do what it says - e.g., obfuscate but preserve original arbitration to maintain original message order
    - The Vehicle Signal Specification (VSS) can be used to remove CAN specific stuff

- Standard methods of collecting raw CAN data are needed as well

Sources of Datasets

- Sources of datasets (ad-hoc, community, industry, government, real-world pilots, reference datasets)
- Industry vs. academia vs. government vs. others: what is the cost? How much is open via partnerships, etc.? Geotab provides one dataset (ignition) free of cost
- How long are datasets usable, what is their “time constant,” how often do these datasets need to be updated, refreshed, etc.? What are the means to do so?
- Datasets depend on intended use, may not match researcher’s needs
- How important is the analysis purpose to drive the need for future data sets?
- ML/AI, RDT&E, different functions (security, safety, repair) - commercial, legal/liability?
- Government can help with automotive datasets; for example, NHTSA is involved in activities through funded efforts to MCity [Mcity], data collected through a Notice of Proposed Rulemaking (NPRM) for Advanced Driver Assistance Systems (ADAS) [NHTSA2020], and other efforts through the University of Michigan
- OpenXC platform: mostly supplies higher-level data, similar to what OBD-II diagnostic queries might provide - however, support seems to have seized [OpenXC]

Appendix F: Breakout Sessions #2 - Potential Applications

A second working session was held to explore potential applications of automotive datasets. The participants were divided into four breakout groups for this session. All of the groups were given the following suggested topics to stimulate the conversation.

- “Killer” apps for automotive datasets
- Current and future analytic tools for processing automotive datasets
- Sharing infrastructure and frameworks for datasets
- Privacy considerations
- Industry participation

The actual topics addressed varied based on the members of the group and were much broader, as captured in the notes. This appendix provides consolidated notes capturing the discussion from the four breakout groups.

“Killer” Apps for Automotive Datasets

Smart City Apps

- Applications to Improve transportation throughout cities for everyone, not just cars.
    - Pedestrians and supporting activities such as trackers
- How to increase transportation availability to improve smart cities design?
- How do you handle non-disclosures, when travelers do not wish to disclose their location
- How we can make raw data digestible
    - Quantity does not imply quality
    - There are limitations in simple car or pedestrian monitoring
- Many interesting applications are possible through the combination of multiple datasets
    - What are those datasets and how can we combine them?
    - Roadmaps and parking data would be one example, but there are many more
- New apps require that datasets are digestible by smart city planners - may require some preprocessing

Dataset Limitations and Biases

- Need to avoid drawing classist and racist conclusions because of where the collected datasets came from
- We may be mostly seeing data from luxury vehicles (until telematics propagates to all vehicle price points) and location biases from where connected vehicles initially live

Equity

- Data to assess economic equality and other aspects of diversity, equity, and inclusion

    - Are there opportunities to leverage automotive data along similar lines (combining datasets)?
- We also need rural datasets because they are different from urban datasets
    - Rural areas have unique challenges such as not enough traffic density to draw reliable conclusions and higher risk to privacy
    - The Wyoming DOT may have some data available collected through the Wyoming DOT Connected Vehicle Pilot [USDOT-WYDOT]. However, this data has limitations. It is V2X data over a stretch of the I-80 highway.
- Collaborations with social sciences and urban planners to better understand the problems around representation of people in designing datasets
- Applications relating to elder care to help them make connections

Industry Collaborations

- Mostly need-based automotive partnerships
- universities can play an important role as the catalysts

What We Don't Want

- Advertising and ads popping up as people pass by venues
- However, like in social media, advertising is an important driver and may benefit users in the form of discounts

Useful Applications

- Vehicle predictive care
    - Already an application offered by industry
    - Can direct people to services with low-wait periods
- Active applications, such as changing traffic patterns and encouraging less congestion

Insurance Apps

- Pay-as-you-drive and pay-how-you-drive, driving behavioral profiling for insurance discounts
- Accident data
- Insurance - pay as you drive based on multifactor (driver, weather, road condition, route etc.)
- Collaboration case with car insurance companies (e.g., Korea telecom, automobile company and insurance company) -- a driver who does safe/eco-driving for a long time period, his/her insurance fee can be discounted.

Apps for Datasets

- A killer app to normalize vehicle signal data to a common format
- App to verify the integrity of dataset - i.e. anomalous data detection
- Vehicle signal data compression methods to make practical lifting data from vehicle to cloud
- App to determine if "ICE" car is suitable for an EV setup (are the paths taken within charging stations, etc.)?
    - Future data capture - V2V app to optimize traffic flow - is complicated
- International Road Assessment Programme (iRAP) [iRAP] road safety rating system
- Multi-modal datasets (bicycles, scooters, pedestrians, vehicles) highly useful for planning intersections/means for people to safely be mobile

- Means to share raw vehicle data (including PII) in a privacy preserving way
    - Killer app - aggregates vehicle data to remove PII
- Using the dataset to train networks (SDN) to provide efficient V2V communications (putting them in the same multicast groups, etc.)
- app/data set to determine which vehicle signals should be request/response, and which signals should be broadcast

Apps for Vehicle Occupants / Owners

- Use vehicle driving data to determine options/subscriptions to better service the driver, i.e., if the driver is hopping into a cold car, offer heated seats.
- AV the big one promised for years
- Increase safety - Reducing driving distraction
- Incremental improvements to current state of features in vehicle
- AirBnB for auto - difficult given how expensive it is - communal owned/rideshare model might be better
    - Data needs for usage, insurance, incidents, maintenance, fuel/charge level, location

Apps from Crowdsourced Datasets

- Crowdsourced data - roving weather stations, road conditions (Waze without human input), optimize routing, the re-routing impacts pedestrians and communities (avoid school zones)
- Fusing data sources, including from pedestrians' phones
    - Vegas project[s] to reduce pedestrian accidents
    - Study pedestrian reactions to AV and vice versa
- V2V opens up range possibilities
- Predicting/scheduling maintenance
- Using captured driving dataset/CAN dataset for identification module, e.g., identifying drivers through something like continuous authentication

Apps Using Data from External Devices

- Using external devices (smartwatch, smartphone) that are sensed/connected to the car/identified by the car. E.g., is it a legit passenger or someone kidnapped?
- Personalizing automotive services, e.g., if 'Sam is driving--he likes rock music and bbq, let's suggest these as recommendations'

Forensics

- Crash forensics/vehicle forensics - 'Freeze framing' data in the event of a crash event Ability to digitally sign results from this system - can we distinguish if an attack led to the crash?
- Gather sufficient data to be able to distinguish cyberattacks from other issues, especially if there is a crash
- App that can achieve high-value data monitoring and reporting while preserving privacy - could tie into EV usage wrt. to energy grid infrastructure

Safe Data Transfer To/From Vehicle and Other Devices

- Send data from CAN to mobile phone/wireless devices - there are devices that do this, but have been proven to be vulnerable (quite literally 'killer' apps)

- LIDAR dataset captured by commercial car--ability to detect LIDAR spoofing - there is current work ongoing to develop these datasets to progress research including blind attack or creating ghost point clouds via ‘laser injection’
- Multimodal sensors to allow for better verification/context of IDS for vehicles - could also be useful for diagnosis
- With location of the car in real-time, extract traffic data in real-time, e.g., congestion, accident, etc. - give context to datasets
- Can we use sensor data to detect anomalies? E.g., if GPS is spoofed, could CAN determine that this was spoofed? Maybe false positives if going through tunnels, canyons, etc.
- Automatically synchronize time & account for clock drift in multi-modal data collection, e.g., automate ‘popping a balloon’ to synchronize different data sources

Tools

- Current and future analytic tools for processing automotive datasets
- Fuse vehicle speed data with GPS and other sensors for an overall health score.
- Python and Pandas dominate! Jupyter Notebooks help provide context
- Tableau and commercial systems are often used at the fleet level
- OTIDS data set processed in Jupyter Lab
    - It depends on the application (and data sizes) and questions that need to be addressed

Tools for Processing Automotive Datasets

General

- Dataset differences between the vehicle and the user’s smart device
    - Would it be useful to merge the two?
    - What applications are made possible through the merging?
- Machine Learning - for example, school buses have a typical behavior, and if AI is looking for anomalous driving behavior it could detect issues

Tools from Other Communities

- Research what tools other industries use to analyze large quantities of data - (Splunk/Gravwell)?
    - Reach out to other disciplines/communities (social/economic/data science/other) for ideas

AI Tools

- AI on vehicle for ADAS/AV, beyond object recognition, lidar - influence driving and recognize e.g., tire imbalance and determine from other data points the vehicle is still safe to operate or not
- AI to assess Cybersecurity on whether an attacker is trying to fool car into not driving
- Federated learning across OEM, pilot projects typically too limited and would benefit from wider range of data sets - geographic and regional driver differences
- Dynamic rerouting when more vehicles AV, influenced by city managers
    - Prevent gridlock

Sharing Infrastructure and Frameworks

General

- Sharing challenges
  - Sharing data due to security and background checks
  - Challenges for some
- How do we walk data between digital twins?
  - Have we designed datasets to be consumed by digital twins?
  - How do we handle the economic and privacy incentives crossing between digital twins and automotive datasets?

Data Portals

- The Safety Related Traffic Information (SRTI) Ecosystem (data for road safety) [SRTI]
  - EU vehicle data sharing model
- Data entry portals that condition the metadata and vehicle data for uniform storage and sharing
  - Platform to drag/drop data file, fill in information, have backend (cloud/db) to process and convert into proper file format. This would also be a “killer” app. E.g., infrastructure to ingest, process, host datasets. Could also cross-correlate time/different modes of data (CAN, Ethernet, etc.). Maybe even automatically obfuscate?
  - We could even add analytics capabilities. E.g., automatically generate notebooks based on included data (does it include OBD, UDS, DBC, etc.)
- Challenge: who is the funded responsible custodian of the data?

Data Exchange

- ELD mandate started in 2017. Can we leverage these manufacturers/information? Privacy concerns likely with this data.
- Need trustworthy data - random datasets from unknown sources can poison existing carefully curated data
- Auto-ISAC exchange protocol could be leveraged
- Has Gravwell [Gravwell] solved some of these barriers?

Timing

- Timing - satellite GPS and NTP which has security concerns, needs on area scale (traffic management).
  - NTS (NTP over TLS) offers some protection
  - Drift exists, need data to understand problem better (how much drift is there?) and influence design decisions
    - Any sources for this data? OEMs and Geotab?
  - Multiple data sources can be used to corroborate and typical drift for a given system can be tracked, major differences can be flagged as suspicious

Risks

- Any central, shared data points have same challenges and risks more so when coordination is required
  - Maintenance issue when sensors degrade

    - GENIVI AMM - spoof data for practically every sensor on vehicle, do so externally
    - Opportunity and risk of shared data security, will be more important for AV and intersections
    - Critical when there is no line-of-sight

Privacy Considerations

Owner Consent for Data Collection

- What is the consent process for collecting data from a car?
- There is significant paperwork when purchasing a car with legal language
    - Can owners opt out?
- Point of sale (vehicle) in some cases includes contract language to enable all vehicle aggregated data generated to be available to the OEM
    - Aggregate example: what is the avg time drift on a year, make and model?
- What happens when a car is sold again?
    - Most states require a rudimentary bill of sale, which does not cover any privacy issues
    - Are OEMs allowed to continue to collect data?
    - Similar issues with phones but you have to accept terms and conditions when signing in with your account

Smartphone vs. Car Data Monitoring

- Smartphones can monitor taxis, buses, bikes, walking, and other transportation infrastructure
    - But they cannot reliably monitor car systems such as brakes, suspension, etc.
    - Important if such information is needed for an investigation

Privacy, Safety and Usability

- With cars being equipped with cameras, are there privacy issues around data such as cameras collecting images/video around cars?
    - Legally different from a camera on the street?
- Collecting data from underage drivers?
    - Laws protect such drivers, how do they affect data collection?
- Federated learning algorithms - split the learning algorithm send less data to central point - anonymized in a sense
- Privacy mode; some manufacturers include button for privacy mode
- Right to remove PII?
- K-anonymity and Differential Privacy are our friends
    - Is Geotab using differential privacy?

Learning from Other Domains

- Privacy for IoT, sensor information - e.g., protecting smart meters, solar panels, health domain (HIPAA) - How do we learn from these domains that have gone through this problem? Always a problem with trade-off with privacy protection
- Common practice in smart meters - utility agreement with how to augment data

    - Want to release data without original signal or consumption value in these fields
    - Developing relationships is important (See Ross Anderson's chapter on Monitoring and Metering [Anderson2020])

Role-Based Access

- Determine roles that different users have, then implement role-based-access controls
    - Do the aggregate results derived from deleted data need to be deleted too?

Legal/Regulatory issues

- Emerging regulations such and GDPR [GDPR] and CCPA [CCPA]
- Location data is considered personal - voice commands are based on personal traces of the owners - voice recognition can cross borders
    - VIN or plate number can be treated as personal information -it depends on each country's law
- Laws have not kept up with technical needs
    - We need to try to figure out how to anticipate and guide legislation and regulation related to data sharing
- Anonymization doesn't really exist… Obfuscation can exist, but it may make aggregating data challenging
- Determine who the owners are? Do they have control to "erase" the data?
    - Different countries have different laws.

Encryption vs Privacy

- We need a "Marriage between encryption and differential privacy" paper to review it
    - Could we make this marriage happen?
    - If cryptography is used, how are keys managed?

V2V Privacy

- How do we consider V2X, V2V communications?
    - Are there privacy concerns? Yes!
    - Cybersecurity is one concern for V2X - could we use a public key infrastructure (e.g., certificates)?
    - DSRC vs. 5G is a discussion - we would need a root of trust
- Certificates require identification but privacy demands identification is impossible - how do we reconcile?

Industry Participation

Working with OEMs

- OEMs see a change in the ecosystem but come from a culture that is closed and private
    - Will that OEM culture continue?
    - Researchers need to accept that culture and work with the OEMs
    - More industry representatives should join a future workshop
    - Will industry work with the research community to exploit knowledge
- Is industry more advanced than researchers realize?
    - Don't know, industry is very secretive
    - Bridging mechanisms must take this into consideration

- Discussion points
    - Pros/cons?
        - Industry can have different priorities versus researchers when sharing data - and the result may not be as useful to researchers
    - Risks and mitigations?
    - Incentives?
    - How to get major OEMs and suppliers to share data given sensitivities?

Other Industry?

- In addition to OEMs we have:
- Suppliers
- Telematics and service providers
- Other large information companies such as Google
    - Very active through Google Automotive
- Opportunities are there: Geotab is a good example of a success story

Facilitating Sharing

- Ideas are currency; IP restricts sharing
- Benefits of open specifications/standards - these are common benefits
    - If we persist in collaborating, will they open up or shut down further?
- We need terms to make OEMs comfortable to open up, unattributed, PII protection
- Benefits - research can improve their products, make them interested
    - But make ask clear, remove uncertainty as it prompts hesitation
- Consent from individual owner (if not driver) might be required
- NDAs are often necessary. Data is likely not to be shared.
- Focus on pre-competitive research data sets.

Data Sharing Opportunities

- Often data is used to sell technologies to industry - how do we separate sales demos from open and unbiased data
- Do we take advantage of open data provided by manufacturers? For example, Hyundai's Open Data Platform [Hyundai-ODP]
- J1939 data is mostly open and leveraged by research - less for finance/driver convenience, but more for maintenance, care, and vehicle information
- Passenger cars are much different - many OEMs will add a gateway to OBD-II port to not allow streaming data out of the vehicle
- Ford used to have a data sharing program, not sure if they still do (provided an API).