**Overview:**

High quality, real-life vehicle network datasets are needed by CISE researchers who are advancing the state of the art in automotive and related systems. However, when it comes to passenger cars and heavy vehicles, such datasets are hard to obtain, which prevents the research community from growing the discipline. The vision for PIVOT (Platform for Innovative Use of Vehicle Open Telematics) is to transform the ad-hoc, small-group endeavors for vehicle data curation into a scientific body of work done by a larger synergistic community.

The phrase “vehicle open telematics”’ indicates telemetry data transmitted between electronic control units both within and external to a vehicle. The dominant form of in-vehicle networking utilizes the controller area network (CAN). Telemetry data available on the CAN bus can also be transformed and transmitted over cellular networks, which has sprouted a rapidly growing industry for fleet and vehicle management. Since this aggregated telemetry data is of interest to researchers, the PIVOT project, through a collaboration with Geotab (a telematics service provider), will provide free access to high-level aggregated datasets.

The word “platform” in PIVOT means the project will build the infrastructure to support the needs of researchers. This infrastructure utilizes five pillars to realize the vision for PIVOT: (a) dedicated platform, (b) curated data, (c) user tools, (d) researcher services, and (e) community outreach and engagement. The incorporation of these aspects of the infrastructure creates complexity that will be managed through a disciplined systems engineering approach that utilizes community feedback mechanisms to continuously improve the PIVOT platform.

PIVOT will be built utilizing existing successful implementations. For example, access to Geotab’s raw telematics data will use the existing Spindle program where Geotab has supplied telematics devices toa mini fleet of volunteer researchers willing to share the data from their vehicles. Also, in-vehicle data collection efforts will be crowdsourced by providing effective open-source data collection devices (e.g., the CAN Logger 3 or a Raspberry Pi) to participants for community data collection efforts.

For PIVOT to be successful, not only will the data need to be collected, but it will need to be measurably useful for the research community. A searchable index of vehicle data and software-based tools will provide utilization insights for PIVOT resources. Community outreach and engagement effort will elicit actionable and measurable feedback to utilize as inputs for requirements utilized in system development.

**Keywords:** CAN bus data; vehicle telematics data; vehicle cybersecurity; intelligent transportation; smart cities and communities

**Intellectual Merit:**

The PIVOT system contains five pillars of merit: (a) robust and reliable hardware/software platform upon which the system runs, (b) the curation and sharing of the data and contextual information, (c) researcher-centric services for sharing, securing, and evaluating datasets, (d) common software-based tooling to collect, transform, combine, filter, and visualize the data, and (e) extensive community outreach and engagement to improve the data utility using design feedback mechanisms. These pillars focus on satisfying the needs of CISE researchers consuming and producing vehicle data as they pursue research in fields from cybersecurity, intelligent transportation, automation, and smart and connected cities. The community will benefit from access to new, hard-to-get CAN and telematics datasets, new tools and tool add-ons to enhance researcher capabilities, and telematics from millions of vehicles through our commercial collaborator. The project will also strengthen the community by providing a forum to exchange ideas and resources, and help researchers form and expand collaboration teams.

**Broader Impacts:**

Successful execution of this project will result in new datasets and tools available to the CISE community that will enable new, innovative research in automotive and transportation-related areas, and strengthen the research community through collaborations built around common datasets, toolsand industry collaborations. PIVOT will provide artifacts to educate the next generation of automotive cyber engineers through classes in computer science (networking, security, machine learning, digital forensics) as well as classes in transportation and smart and connected communities. The project will emphasize diversity through efforts targeting minority institutions and underrepresented groups and by reaching outto students participating in the industry sponsored CyberAuto and CyberTruck Challenge events. Core concepts in PIVOT came from the results of an initial community workshop. PIVOT will continue with annualworkshops to build and enhance community and support strong advances in automotive security, smart transportation, smart cities and communities, security, safety, privacy, sustainability, and energy.

# 1 Introduction and Motivation

CISE researchers interested in developing new results based on data obtained from in-vehicle networks have explicitly expressed needs for a corpus of reliable and meaningful datasets, tools, and services. To this end, we propose an infrastructure called PIVOT (Platform for Innovative use of Vehicle Open Telematics). The core vision for PIVOT is graphically depicted in Figure 1. At its core, PIVOT focuses on the needs of researchers pursing research in cybersecurity, intelligent transportation, and smart and connected communities. The platform component of PIVOT is the hardware and software infrastructure needed to host the data, tools, and services reliably at multiple locations. The community component of PIVOT focuses on engagement, outreach, and feedback to create a synergistic system to support research efforts. The stakeholders benefiting from the research enabled by PIVOT are depicted around the outside of the ring in Figure 1. The figure represents both the core focus of the researchers and the broad impact enabled by the PIVOT resources.

**Understanding Researcher Needs based on Workshop Results.** Recognizing the potential need to initiate a coordinated effort to bring together a community around the development and sharing of robust automotive datasets, PIs Balenson and Papadopoulos and collaborators from Geotab Inc. and Oak Ridge National Laboratory (ORNL) organized a community workshop; PI Daily was a speaker and substantive contributor with one presentation on research and one presentation on outreach. The workshop, entitled "Paving *the Road to Future Automotive Research Datasets: Challenges and Opportunities*," was held virtually on November 18-19, 2021. A report that summarizes the presentations, discussions, and key findings fromthe workshop is available online [10].

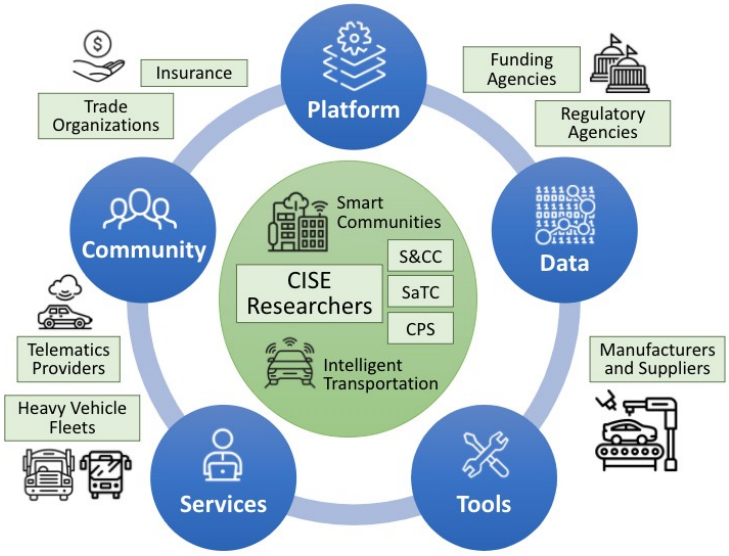


Figure 1: PIVOT is the ring of resources for CISE researchers to access and probe vehicle and infrastructure data sets to enable research into new, innovative applications for cybersecurity, intelligent transportation, and smart and connected communities

The workshop was attended by a broad group of 68 researchers from 30 organizations across academia, industry, and government. In fact, most of the researchers who provided letters of collaboration are CISE researchers who participated in the workshop. It served as a forum for learning and understanding the wide variety of automotive research datasets needed to support automotive cybersecurity and otherresearch, as well as the broad range of applications that could benefit from research and development supported by such datasets. Participants discussed numerous, broad applications enabled by sharing the continuous and diverse information collected from and by connected vehicles.

The PIVOT platform will serve the following core CISE communities: CCF, CNS, and IIS, plus the SaTC, CPS,and S&CC programs where researchers work on applications in automotive cybersecurity, vehicle service and maintenance, vehicle occupant services, connected and autonomous vehicles, electric vehicles and the charging infrastructure, transportation, smart and connected communities, and more. The level of participation and degree of interactions in the workshop clearly demonstrated a strong desire for a community around automotive datasets and their use in research applications. The proposed PIVOT project is informed by, and intended to coordinate, support, and grow this community.

The following needs were identified during the workshop. These provide the motivation for establishing the PIVOT system.

**The Need for In-Vehicle Datasets.** While there are many communication systems in the automotive domain, researchers attending the workshop emphasized the need for Controller Area Network (CAN) data. With these datasets available, researchers expressed they could conduct the following types of research:

(a) intrusion and anomaly detection based on analog signals seen on the CAN bus wires, (b) CAN error frame detection and schedulabilty studies based on time histories of digital transitions for CAN bits, (c) reverse engineering the meaning of the network data to further populate database (DBC) files, (d) protocol analysis research to examine security based on assembling messages into protocol data units (PDUs), and (e) research in anomaly detection and situational awareness by using datasets with correlated measurements of phenomena outside the vehicle network.

**The Need for Telematics Data.** Automotive telematics service providers use on-board devices connected to the vehicle network to summarize data in real-time and send it to the cloud. Through our collaboration with Geotab, a major telematics service provider, CISE researchers will gain access to telematics data from over 2.5M cars and trucks. Such data enables research related to predictive service, maintenance, cybersecurity, traffic management, infrastructure evaluation and planning, parking, and localized weather applications.

**The Need for an Open-Source Data Repository.** The workshop revealed that while there exists open-source dataon the web, it is not well organized, and a web search may miss much of it or require work from researchers to understand its relevance. The community needs an extra level of review of open-source data to collect, organize, index, and apply a uniform description to augment search results. The extra review of open-source data provides researchers with additional confidence in the data they rely on for their research.

**The Need for New Open-Source User Tools.** There are capable commercial tools available to process CAN data, but they are costly and proprietary, which does not serve the broader research community well. Fortunately, there are several open-source low-level CAN tool suites such as can-utils [7] and some higher-level tools such as LibreCan [43], but they are not as usable and capable as commercial tools. Workshop participants expressed a desire for an open-source ecosystem that provides software tools to enhance usability and add capabilities to existing open-source tools.

**The Need for User Services.** User services must support common user activities and provide a productive, friendly, and easy to use environment for interactions with the portal and each other. Researchersattending the workshop with experience in working with data indicated they would benefit from services thataccommodate: (a) maintenance of user accounts and identity, (b) data uploading mechanisms like drag anddrop infrastructure, (c) data and metadata review for consistency and quality, (d) keyword search results,(e) crowdsourced data collection activity descriptions, (f) access and privacy controls for archived data, and (g) identification and removal of personally identifiable information (PII). It was also apparent that the workshop participants have fantastic ideas for common services, so PIVOT specifically focuses on incorporatinguser feedback to build and improve services.

**The Need for Privacy.** Automotive datasets may contain sensitive data such as vehicle identification numbers or GPS coordinates, and that information needs to be carefully managed to avoid violatinguser privacy. At the same time, researchers at the workshop identified a trade-off between usability and privacy when anonymization and obfuscation are employed. PIVOT will employ several privacy controls thatalready exist and invite researchers to further advance approaches to privacy using PIVOT datasets.

**Intellectual Merit**

The PIVOT system contains five pillars of merit: (a) a robust and reliable hardware/software platform upon which the system runs, (b) the curation and sharing of the data and contextual information, (c) researcher-centric services for sharing, securing, and evaluating datasets, (d) common software-based tooling to collect, transform, combine, filter, and visualize data, and (e) extensive community outreach and engagement to improve the data utility using design feedback mechanisms. These pillars focus on satisfying the needs of CISE researchers consuming and producing vehicle data as they pursue research in fields from cybersecurity, intelligent transportation, automation, and smart and connected communities. The community will benefit from access to new, hard-to-get CAN and telematics datasets, new tools, and tool add-ons to enhance researcher capabilities, and telematics from millions of vehicles through our commercial collaborator. The project will also strengthen the community by providing a forum to exchange ideas and resources, and help researchers form and expand collaboration teams.

**Broader Impacts**

Successful execution of this project will result in new datasets and tools available to the CISE community that will enable new, innovative research in automotive and transportation-related areas, and strengthen the research community through collaborations built around common datasets, tools, and industry collaborations. PIVOT will provide artifacts to educate the next generation of automotive cyber-engineers

through classes in computer science (networking, security, machine learning, digital forensics) as well as classes in transportation and smart and connected communities. The project will emphasize diversity through efforts targeting minority institutions and underrepresented groups and by reaching outto students participating in the industry sponsored CyberAuto and CyberTruck Challenge events. While the core concepts in PIVOT came from the results of our community workshop, PIVOT will continue with annual workshops to build and enhance the community, and support strong advances in automotive security, smart transportation, smart cities and communities, safety, privacy, sustainability, and energy.

## 2 Infrastructure Description

In this section, we discuss the fundamental infrastructure and integrated tools we plan to create, vehicle telematics data we will import or link to, user services for the infrastructure, how we will actively engage the CISE community in the design of our work, and our community outreach efforts.

### 2.1 Fundamental infrastructure

We will implement a scalable, interactive platform to provide user services and access to data. The platform will host a web server, database, and microservices. It will implement robust security, have a firewall, and will be mirrored at partner institutions for backup, redundancy, and seamless recovery.

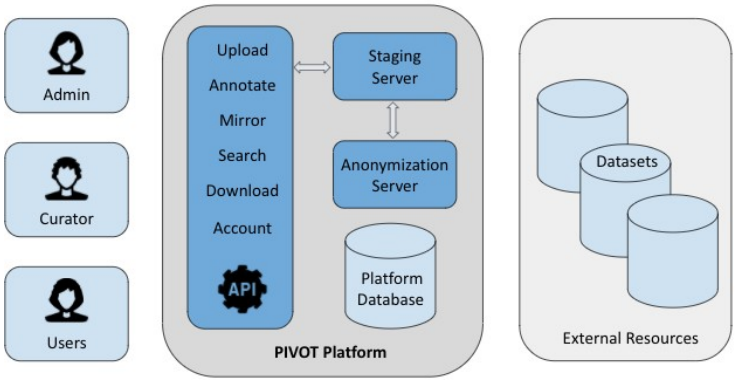


Figure 2: High-level design components

**Service Description.** User account creation and maintenance will be responsible for all account related tasks. It will utilize a modern Single Sign On identity provider (OAuth/OpenID) and it will include a rate limiting capability via a CAPTCHA to prevent brute forcing attempts. User accounts will range in privilege levels, and will include admins, curators, and users.

**PIVOT Service Architecture.** The functionality of the platform is reflected in the verbs which will be accessible viaan HTTP RESTful interface. Logic will be implemented as a group of microservices that will be responsible to implement these verbs. This design will exhibit good scaling properties under the assumption of mostly read intensive workloads. Every dataset offered will be identified with a unique identifier and stored in the SQL database alongside with the setof files that comprise it and their checksums. In addition, metadata about each dataset will exist both in the database and in special metadata files in the filesystem. Users will have unique IDs and they will be associated with datasets they own and comments and annotations they have left on existing datasets.

**Database Design.** We will implement persistence in the platform through a relational database. Tables will hold (a) user information like username, privilege level, authentication details; (b) datasets and their metadata like number of messages, size, files, hashes, curator, co-curators, vehicle id, etc.; (c) comments and annotations, the IDs of users and datasets they relate to, and copies to blobs that can be internal data of a dataset; and (d) mirrors that will be remote filesystem paths where datasets are hosted.

**Filesystem.** We will use a RAID-backed filesystem that will hold the bulk of the content and be accessible via HTTP/FTP. It will serve both open and restricted content. For the latter data will be placed behind a protected access interface that utilizes a platform authid token. Mirroring operations on the data will happen through a standard filesystem tool (rsync over secure shell).

**Microservice-Oriented Design.** We will implement communication between the Database and the Filesystem through microservices (read only or read/write). Each microservice service will be responsible for a verb offered by the API. This approach offers a fast path to a working prototype, incremental development capabilities, as well as a quick turnaround time of incorporating new features.

**Web Servers.** Web servers will perform load balancing, rate limiting (authentication, CAPTCHAs), serve static content, and caching. Web servers will communicate with the appropriate microservice endpoints via gRPC. This allows for efficient data interchange even in the case of streaming requests.

**Scalability and Redundancy.** The addition of new storage/mirroring resources should be straightforward and need minimal cooperation with the core of the platform. On the other hand, the bulk of the platform's functionality is persisted on the central database that will be periodically backed up/replicated atpartner institutions. The microservices and web servers will be offered as containerized applications along with configuration scripts to instantiate them, so even if the primary location (U. of Memphis)goes offline, a secondary location (Colorado State University) can execute the application containers and using the replicated database, have a working copy of the whole platform within minutes. DNSfailover will make the transition seamless, with only resources that were not mirrored be affected.

## 2.2 Tools, Resources, and Datasets

In this section we describe the datasets and tools we will create, collect, and provide to the community through our portal. We will collect, curate, or index four types of datasets: (a) CAN data and its layers, (b) telematics from our industry collaborator, (c) raw telematics from our volunteer-based mini-fleet, and (d) open-source data from the web. We also describe the data access methods we will offer along with some common tools to help put the data in a usable format.

### 2.2.1 CAN Data Collection Devices

A key roadblock for research using CAN datasets is the interpretation of CAN signals. CAN signals carry information between various vehicle components. Capturing CAN traffic traversing over the CAN network means researchers have visibility to the communication between the various ECUs and sensors, which is critical in understanding vehicle operation. Signals are encoded in CAN frames, which can be captured with CAN loggers such as those we propose to build. However, except for standardized signals, the encoding of CAN data is proprietary to each OEM, which means researchers cannot interpret the signals. Even heavy vehicles that utilize J1939 for encoding still utilize proprietary CAN frames. This has initiated several community efforts to reverse engineer the signals by building CAN signal decoders such as ORNL'sCAN-D and University of Michigan's LibreCAN. These decoders often use statistical techniques to decodesignals, resulting in signal definitions that are not deterministic. The accuracy of these tools can be improved with access to more CAN datasets. The rules for decoding CAN signals are typically captured in a *CAN database file (DBC)*. Such files are valuable for CAN researchers and any improvement in the process of deriving them would greatly enhance automotive research.

**Why Decoding CAN Signals is Hard.** We briefly describe why decoding CAN signals is hard. There are several more detailed explanations available online, but the following should be sufficient to understand the problem. The CAN frame payload is typically eight bytes (64 in CAN-FD). Each CAN frame is identified by an arbitration ID, which determines the transmission priority and meaning of the frame (lower ID means higher priority). The payload of each frame typically encodes the values of several signals keyed on the frame ID. The problem with decoding these signals is that the encoding is proprietary: OEMs define their own mapping between frame IDs and signals and the layout of signals inside a given frame ID. Encoding defines values for several variables including the offset (where in the payload each signal starts), the number of bits used for the signal, and the endianness and signedness (bit-to-integer encoding). Finally, even if we can accurately reverse engineer the format of each signal inside a given frame, we still need to determine its function and units (e.g., speed in km/h). The situation is worse in real life since thesame OEM may use a different DBC (frame IDs and signal mapping) for different vehicle models and years.An OEM engineer at the workshop pointed out they use an internal tool to define new CAN frame IDs when engineers need them. This means that coordination across OEMs for DBCs would be very hard,given current practice, so researchers need to continue to be innovative with tools and methods to decodeDBC. While existing tools such as CAN-D and LibreCan help, the creators of these tools indicated to us that access to more CAN datasets would help significantly in improving their accuracy.

**Community CAN Data Collection.** We will streamline and standardize CAN dataset collection through our CAN crowdsourcing program. The program will provide CAN Loggers and collect many CAN traces from volunteers who are willing to install these loggers in their cars or trucks. This task will createnew data for researchers to improve the accuracy of existing DBCs, which will provide accurate visibility to communications and open a wide field for more innovative research.

We will provide two open-source solutions for CAN data collection devices. This keeps the cost down and more importantly it gives us complete control of the devices. One is based on a Raspberry Pi running

embedded Linux and open-source software. The other is based on the open CAN Logger 3 device developed by PI Daily and his student Duy Van [21]. The cost of either device is low at about $200 each.

The CAN Logger 3 devices are custom, yet open source, and can be programmed either withthe Arduino development environment over USB or a more traditional development environment using a the JTAG port for its ARM-Cortex M4F processor. Since the CAN Logger 3 uses an NXP K66 processor at180 MHz, it is capable of logging at higher speeds (up to 6Mbit/s) than the Pi and it can provide bit-level information that the Raspberry Pis cannot. Bit level information is important when detecting CAN errors that may happen during component malfunction or an attack. This feature will provide researchers additional data when investigating anomaly detection systems and security approaches.

We will build and distribute a limited amount of both CAN Logger 3 devices and Raspberry Pis with CAN hats to use for CAN data collection. Both devices already exist and have been validatedas capable, low-cost hardware solutions to collect vehicle data. These devices store CAN data on local SDCards during collection. After the data collection activity is finished, the volunteers will upload the collectedlogs to our servers periodically. During this project, we will improve the software on these devices to enablethe existing Wi-Fi hardware capabilities and automatically upload data.

These collection efforts and hardware systems will work for all vehicles of interest, including passenger cars and commercial vehicles. The devices also have CAN-FD capabilities. As mentioned in Section 2.3, some of these tools will be tested/reviewed to help CISE researchers who want to acquire their own devices.

**Expected Volunteer Community.** We expect to recruit about 50 volunteers to install Geotab telematics devices to their vehicles (Geotab has committed at least 50 devices). We expect to have about 50volunteers to install our CAN loggers (some of these may install both Geotab devices and CAN loggers). All 50 volunteers with Geotab devices will be invited to join the Spindle mini-fleet program. We will recruit volunteers from the list of over 100 participants and invitees from our November 2021 workshop, at conferences and workshops such as AutoSec, ESCAR, and VNC, as well as participants of the CyberAuto and CyberTruck Challenge events, which train approximately 30-40 students each year. We will also advertisevia word of mouth through other faculty and make a concerted effort to recruit underrepresented groups in the pool of volunteers, until we have exhausted our available devices.

### 2.2.2 Telematics from our Commercial Collaborator

We are collaborating with Geotab, a large telematics provider, who has committed to facilitate data access to CISE researchers. Telematics providers offer services to fleets. Data is typically collected by devices connected to the diagnostics port of a vehicle, which transmits summarized data in real-time to the provider's cloud, where it is further processed into services and applications for fleet managers. Geotab agreed to make telematics data from over 2.5M cars and trucks available to the CISE researcher community (see letter of collaboration). Geotab already makes aggregated data available through a free account at their Ignition website [25]. The purpose of Ignition is twofold: (a) to demonstrate the capabilities Geotab provides to its current and future customers, and (b) to nurture innovative uses of the data. Through our collaboration, Geotab will make additional services available to CISE researchers, including training material, access to Geotab engineers, and access to data currently not offered through the free account. In addition, we will facilitate dialog between Geotab and CISE researchers.

### 2.2.3 The PIVOT Spindle Program

*PIVOT Spindle* is a mini-fleet of volunteers who will install Geotab telematics devices in their own vehicles. Spindle is currently active and is managed by PIs Papadopoulos and Balenson. The main differencebetween Spindle and Geotab's standard services is that Spindle users have access to the raw data collectedby the telematics devices. Volunteers who want to participate in the Spindle program will sign a consent form and receive a device with instructions on how to install in their vehicles. Devices are transferable betweenvehicles – a user can simply archive the database of the current vehicle and start a new one. There are currently a handful of users participating at Spindle that include standard internal combustion engine (ICE)cars, electric vehicles (EVs). and hybrids. It is worth noting that Geotab's Glenn Atkinson (who wrote our collaboration letter) is a Spindle participant. Geotab offers an international University R&D Program wherefaculty and students are offered up to two telematics devices [26]. There are currently several participantsin this program who will be invited to join Spindle.

**Combined Collection with a CAN Logger and Spindle.** Users who opt to install both a Geotab device

and a CAN logger will produce richer datasets through the combination of complete CAN data and targeted telematics. The CAN logger will provide a detailed view of vehicle communications while telematics will provide a richer context to vehicle communication. We will provide the required adapter connect both devices and guide users to register the Geotab device with the Spindle project. Data collection will be separate, with the Geotab device uploading data to the cloud and the CAN Logger collecting data in the SD card. We will handle the synchronization of the datasets after the fact (there is also an opportunity to use a Wi-Fi connection between the two devices that we plan to explore). These enriched datasets will beuseful to researchers as they can augment telematics with CAN data. For example, GPS information fromthe telematics device will provide location information for the CAN datasets if the CAN logger does not haveGPS. Adding context to CAN traces is also useful for cybersecurity applications by providing driving contextsuch as matching sensor data with the driving terrain.

### 2.2.4 Open-Source Datasets and Tools

There is significant automotive and transportation open-source data available on the web. Examples include the ETAS/Bosch SynCAN dataset [32], ORNL ROAD dataset [51], and USDOT Public Data Portal [41]. These datasets, however, are not organized and a standard web search may not return appropriate results or may miss many of them, resulting in missed opportunity for CISE researchers. We have been collecting pointers to these datasets for the past year. We will index them with keywords and make them available through a search facility in the PIVOT portal, after a light review to ensure appropriateness and link liveness. Similarly, there are many tools, predominantly CAN tools, available on the web [24, 27, 36, 43], but they also lack organization and indexing. With the help of the community, we will find such tools, index them, provide keywords, a uniform description for each, and mark those that are still active. Our portal will also provide a capability for the community to rate open-source datasets and tools.

### 2.2.5 User Software Tools and Utilities

In addition to the hardware systems comprising the Raspberry Pi CAN Logger and the CAN Logger 3 described above, we will offer the following software-based user tools.

**Tool 1: CAN Log Format Converters.** There are many tools available for CAN data collection including: Raspberry Pi with a CAN hat, BeagleBone Black with a CAN Cape, our own CAN Logger 3, DG Technologies Beacon, Intrepid Control Systems Value CAN and NeoVI, Peak CAN USB Adapters, and Vector CAN Case XL Log, just to name a few. Many hardware devices support the SocketCAN interface, which is preferred by the research community. However, some devices use Windows-based drivers and software, which is a different format than SocketCAN. To include researchers producing data sets with other devices, *we develop tooling for easy and fast conversions between data formats*. The standard data format for PIVOT is based on the open source can-utils' candump format.

**Tool 2: Convert Raw CAN into Protocol Data Units.** In application layer protocols such as SAE J1939, additional definitions are attributed to the CAN frames [35]. For example, in J1939 the extended 29-bit CAN ID is broken down into four parts: the message priority, parameter group number (PGN), destinationaddress (DA), and source address (SA). The combination of these fields and the data comprise the ProtocolData Unit (PDU). Often, researchers would rather focus on the assembled protocol data units for these communication systems as opposed to the raw frames. Therefore, *we will engineer user tooling to extract PDUs from raw CAN logs*. This will also include transport protocols for messages following either J1939 or ISO 15765 formats.

**Tool 3: Data Decoding.** Once a PDU or data frame is available, it needs to be decoded. Some standards, like J1939 for heavy trucks, explain how data is encoded within a PDU [35], but many times the decoding information is proprietary and needs to be inferred or obtained under agreement with the OEMs. Telematics providers and diagnostic tool makers are often privy to the encoding and decoding schemes. We will provide tooling to store decoding data in both DBC and JSON data formats. PIVOT will provide tooling for a translator to convert DBC (and other decoding data) into a serialized JSON file format that can be loaded directly into a researcher's application. This enables a rapid decoding engine to interpret data into engineering values. We will extend the work done with NMFTA's pretty-J1939 tool and apply it for the 11-bit IDs used in passenger cars [9]. The result will be a fast, lightweight extensible software *tool to convert raw data into engineering signals*.

**Tool 4: CAN Data Log Slicing and Filtering.** Often data collection is done in ways where the duration

of the data files does not align with the research questions and researchers need to be able to separate, join, concatenate, filter, and combine data files. We will *engineer efficient tooling for users manipulate log files and create logs that are specific to community needs*. The integrity of the messages and order will be maintained, uninteresting messages can be discarded, and multiple files can be concatenated. This will likely use Python Pandas data frames and leverage the built-in data manipulation features of that package.
**Tool 5: Data Visualization.** Users benefit from being able to visualize the data by plotting time series data of both the raw data values as well as the decoded values. *We will build a tool to leverage plotting packages, like matplotlib, to render images that represent various time-series data*. The utility will be written with a GUI framework, like PyQT, and plot data based on message selection and data range definition. Furthermore, if a decoding engine is available, then the data can be plotted based on the engineering values. An example prototype of this tool was produced under another NSF project by PI Daily [19], but it needs robust software engineering processes applied to it to make it usable by the community.

## 2.3 User Services

Services running on the PIVOT platform are needed to create a simple, friendly, and useful user experience, and seamless operation for working with research enabling data. These services are built and maintained by the PIVOT platform providers and are intended to be “unnoticed” by the user. As such, the technologies employed for user services are based on existing tools and best practices. These include core services, privacy, and data access policies.

### 2.3.1 Core Services

Core services are services that manage user information or that users use directly.

**Identity Management.** We will use standard user authentication procedures including SSO, passwords, CAPTCHAs, and verified emails. Portal admins will use stricter processes such as two-factor authentication.

**Contributing Data.** The core service of the PIVOT system is to accept user contributed vehicle data. Each dataset that is offered to PIVOT will also have accompanying metadata. These metadata fields with have automatically generated information, like time, message summaries, and uploader identity. Also, the user will have the ability to add additional contextual information. For example, a video showing the experiment in which the data was collected. The service will need to scan uploaded files for potential security threats and unset any executable flag on the upload. We will work with the community to provide several upload options (URL, FTP, cloud-to-cloud, etc.), and the user will be asked to fill out a form describing whatthey are uploading. We will provide options for the most common types of uploads with information in the form adjusted automatically.

**Retrieving Data.** A typical retrieval involves a user search producing a list of candidate datasets. Once the user selects the desired downloads, we will generate limited lifetime clickable download links tied to the UUIDs of the files and email them to the user after approval. Similar to uploads, we will work with the community to provide several download options (URL, FTP, cloud-to-cloud, etc.).

**Search Service.** We will try to provide a search service on basic search terms that match the dataset name, provider, keywords, or words in the description. We will sort results by popularity, based on community feedback. The search domain will also cover tags such as software tools, documentation, decodingengines, and research products from the community. The search engine will continue to evolve based on feedback from the community.

### 2.3.2 Services to Nurture Community

In addition to data upload, download, and search, the following services will also create value for the research community. *News and Announcements*: A community powered blog service related to vehicle datasets. This would include publishing opportunities, new vehicle features, technology news, security incidents, or other items of interest to the PIVOT community. This content will be moderated to keep the content focused on the stories related to vehicle data. Decorum rules will be enforced. *Bulletin Boards and Forums*: These provide users a way to ask and answer questions regarding vehicle data and tools. It will facilitate the community engagement and provide users an opportunity to contribute to PIVOT without needing to be an administrator. *Schedule of Events*: Upcoming conferences, workshops, Cyber Challenges, and other events of interest will be maintained on a shared calendar for PIVOT. The fosters community and creates awareness of for the researchers producing and consuming PIVOT data. *Technology Reviews*:

Quarterly articles with succinct lessons and practices for utilizing the technologies associated with PIVOT data collection and utilization. Methods and techniques for data collection, tool use, and presentation of results are all topics of interest. *Badges of Honor*: Users will earn badges based on their communitycontributions (uploads, community event organizer, student mentor, etc.)

### 2.3.3 Policy and Privacy

Policy and privacy are at the center of data sharing. Data providers must not jeopardize the privacy of their customers and researchers do not want to knowingly violate policies. Bad policies or restrictive privacy controls may make research and collaboration very hard or impossible, so both providers and researchers require a robust policy and privacy framework. Our policies will strive to encourage participation; our privacy approach will be guided by the Menlo Report [39] and its principals of respect for persons, beneficence, justice, and respect for law and public interest. We will attempt to balance risks with benefits of research (beneficence) and provide auditability and disclosures (respecting law and public interest).

Our portal will provide the following policy and privacy services.

**User Vetting**. We will vet users by confirming they are part of an institution with an established reputation. Users will also be asked to create accounts in the portal and provide their name, affiliation, institutionale-mail address, prior research (or research advisor, for students), a brief description of their research usingour infrastructure. Gathering this information is common in data sharing communities, allows us to verify affiliation, and to confirm the requested data is plausibly relevant to the research.

**Open-Source Data.** This is existing data scraped from the Web or Geotab's Ignition public data. Open-source data does not have privacy issues. We will index scraped data from the Web, so it is searchable and discoverable, and point to Ignition for Geotab's public data. We will make results available to researchers through a hold-harmless click-through agreement.

**User Protections.** We will solicit volunteers who are willing to openly share their data and are willing to sign a consent form to donate all their data in raw form. However, to expand our volunteer base, we will also offer various levels of anonymization of the contributed data. We will provide multiple privacy controls for volunteers. Beyond the ability to mask information that may be deemed PII (VIN, location data), we will scramble the same dataset with different values. For example, the VIN number for one researcher will be different than another. This may provide a rudimentary level of watermarking to guard against accidental release.

**Community Data Access Controls.** We will couple technical (anonymization, selection, separation), policy (MOUs) and user vetting methods. Technical methods include anonymizing information such as adding noise, obscuring the VIN number, GPS, and other user identifiable information from datasets. With selection we will remove privacy sensitive information from the datasets before they are distributed. With separation we will ensure that a researcher does not get versions of datasets that would break privacy if they were joined. Note, however, that these are privacy controls applied to other datasets (network traffic for example) and they are a reasonable place to start. Since there are no established privacy controls for automotive datasets yet, we will conduct a privacy review with the help of the community and the advisory board to determine which privacy method is best for a given dataset type. We expect our datasets and tools will facilitate privacy research in automotive datasets.

**Legal and Policy Controls.** During the application process users will be asked to agree to our data use policies. For sensitive data we will employ a signed Data Use Agreement for more sensitive data. These agreements outline specific researchers' responsibilities and acceptable data uses, including prohibiting resharing of data, sharing of login credentials, and deanonymization the data.

**Proprietary Information Disclosure Controls.** Our infrastructure will need safeguards to avoid disclosure of OEM proprietary information. Such information may be inadvertently (or deliberately) be captured in CAN traces for example. A CAN trace may capture the firmware if taken during a firmware update oran ECU reflash, or fuel maps, diagnostic messages, key exchanges, and interactions between the vehicleand proprietary diagnostic tools. Our infrastructure will employ tools to sanitize such information if needed.When we detect datasets with potentially sensitive information (CAN datasets or other) we will consult withthe OEM, the community, and the advisory board to ensure we remain within the law and whatever sanitization techniques we use will still leave a useful dataset. Each provider may have different requirements, so it is not clear at this moment that a common sanitization methodology exists. For this project we will restrict ourselves to simple techniques such as removing data, adding noise, obfuscating,

and filtering. The problem of sanitizing datasets is an important one for the community to solve, and our datasets, along thedialog we will facilitate with OEMs, will stimulate new ideas.

PIs Papadopoulos and Balenson were involved with the DHS IMPACT program [33], which developed similar processes and agreements for network data, and will bring that expertise to the project. All PIs will work with the community and the PIVOT advisory board to define acceptable policy and privacy controls.

### 2.3.4 Data Access Services

Once researchers are approved for data access, we will make data available to them through a method that depends on the associated controls. We list these methods below.

**Open-Source Data.** For open-source data we will provide links to the websites that host them. This preserves any analytics the websites are running, while still enabling our infrastructure to index, link and enable searching for the data. We may mirror some of the data for ease of access to PIVOT users after consulting with the data owners. We will assign keywords for the search, which will be continuously refined by the community.

**Geotab Ignition Data.** We will provide a description of the Geotab Ignition website on our platform to help researchers find it and create free accounts. Researchers can download the data directly from Ignition. We will coordinate with Geotab to obtain analytics they collect about our users (we will provide a list of users) so we can add them to our analytics. Researchers will benefit using PIVOT for Ignition data because they will also gain access to other resources such as consultation with Geotab personnel and access to training videos.

**Geotab Cloud.** The Geotab offers cloud processing through their own tools to researchers who use the Ignition platform. Researchers get access to these resources as part of registering their free account.

**Direct Download.** This would be our preferred way of providing data with few restrictions. We will create a password-protected download link and provide it to the researcher. The link will expire after some reasonable time (a few days). This method incurs very little work for both us and the researchers.

**Cloud-Based.** When data is provided to us through the Cloud, we will use a cloud-based access method. We may also upload data to the cloud if a researcher requests it. Accessing data through the cloud allows researchers to use any cloud credits or cloud resources they have (e.g., Amazon S3 or CloudBank). With this approach the compute power available to researchers is limited only by their cloud budget. If needed, we will encrypt cloud data.

**Hosted Datasets.** We will provide access to datasets to researchers through our own platform. This approach will be used for sensitive data that the provider prefers to contain on our platform. Researchers will get accounts on our platform and perform their computations there. We will leverage existing infrastructureat the University of Memphis used to provide network data access to the networking community. This infrastructure contains over 20 1U servers and 0.5PB of storage and has been available to the community for over 10 years.

**Code-to-Data.** This is the most restrictive access and will be done at the request of the provider. Researchers will not directly access the data, but will provide their code to our administrators, and allow us to audit the output before we hand it over to the researcher. Auditing includes checking for PII, side channels, and other forbidden behavior. Output auditing is not foolproof, but along with the other mechanisms and MOUs it should result in reasonably strong privacy protection. Code-to-data is appropriate for very sensitive data, and we expect to offer it in rare circumstances, upon researcher request and approval by the provider.

## 2.4 Community Engagement

One of the most effective means for building an active community is to instill a sense of ownership. We are planning a series of project activities to do just that. Activities are planned around key issues in building an active and enduring user base: awareness, usability, usefulness, and cost. Our community outreach efforts (Section 2.5) will focus on awareness. Usability and usefulness will be addressed by the requirements, design, testing, and evaluation phases of our effort. Cost will be addressed by our sustainability plan (Section 6). Here we discuss the details of our community engagement plan.

**Advisory Board.** We will form and engage an advisory board for the PIVOT platform project. The PIVOT Advisory Board (PIVOT-AB) will include 7-8 academic, industry, and government senior leaders and visionaries from diverse communities that are relevant to automotive, smart transportation applications, smart and connected communities applications, vehicle data and telematics, big data and data analytics,

and data security and privacy. We will seek advice and support from the advisory group throughout the lifeof the project. Key tasks include: (a) review and help refine and mature our overall plans; (b) identify and recruit influential researchers and other community members to participate in PIVOT; (c) review and provide feedback on requirements for the PIVOT platform, datasets, and analytical tools, and on research drivers and applications that can benefit from datasets and tools; (d) review and help develop plans for community engagement and outreach, including annual workshops; and (e) provide advice on the identification and pursuit of long-term PIVOT platform sustainability options. A significant part of the advisory group role will be ensuring that the proposed effort engages a wide swath of CISE researcher participants. The PIVOT Advisory Board members will be drawn from key collaborators and supporters, among others, from across the data provider, data analytics, and researcher communities. We plan to engage the advisory board through a virtual kick-off meeting, quarterly virtual meetings and annual in-person meetings.

**Community Recruitment.** In addition to contacts provided by the PIVOT-AB, we will use our community outreach efforts to establish contact with and recruit CISE community members to participate in all phases of our effort. Section 2.5 provides details in how we will recruit community members.

**Broad Requirements Elicitation and Review.** Understanding that not all interested community members will be able to attend a workshop and that we want to engage as many participants as possible, we willreach out individually and conduct interviews of other community members to understand how they preferto collaborate and what datasets and tools they would find the most useful.

**Design Review and Feedback.** Participating PIVOT community members will be invited to review and provide feedback on our designs. We will also post the designs for open comment on the PIVOT website with a notice to CCRI-VO. We will refine the designs using community feedback.

**Community as Testers and Early Adopters.** We will invite a set of community members to help with early testing of the features, capabilities, and content of the PIVOT platform through alpha and beta testing programs. Testers will be recruited from CISE communities that align with the different use cases to ensure workflow for each use case is exercised. Some of these testers may also come from our own organizations. For each year of the project, we will host a community workshop to demonstrate and promote the PIVOT platform, recruit users, and provide a venue for data contributors to discuss their collections. We will also reach out to other community members and extend an opportunity to participate.

We will develop and provide documentation and tutorials to help prospective users rapidly make effective use of the platform.

**Data and Tool Contributions.** Data and tool contribution is of paramount importance: a great design with no content will not attract and sustain a user community. As noted earlier, we will start with existing datasets, Geotab data, and other publicly available data and with existing tools and other data analytics tools. The PIVOT Spindle and CAN logger efforts will directly engage researchers and other in the crowdsourced collection of CAN and telematics data from across a diverse set of cars and trucks. As part of community outreach, we will recruit other data and tool providers from across the globe and encourage contributions. We will use reward-based “hackathons” and competitions among researchers and other community members to encourage more and better data contribution. We will also use these exercises to obtain feedback on different parts of the platform. We will work with industry, an important partner, to address their interests and seek ways to share data from their vehicles and components.

**Industry Engagement.** We will reach out to and engage a broad set of industry stakeholders, including manufacturers, suppliers, startups, telematics, insurance, and many other applications of automotive datasets. We will leverage our relationships with members of the Automotive Cybersecurity Industry Consortium (ACIC) [1], and the Automotive Information Sharing and Analysis Center (Auto-ISAC) [2], NationalMotor Freight Traffic Association (NMFTA) [5], and others. We will seek to engage industry as both a producer and a consumer of automotive and transportation-related datasets and analytical tools. We will alsoengage industry as a research partner to help drive requirements for datasets and tools and to provide research drivers and applications that can benefit from datasets and tools.

## 2.5 Community Outreach

The overall goal of the PIVOT community outreach effort is to develop a diverse and vibrant community of CISE researchers contributing to and leveraging the shared resources of the PIVOT platform. We will reach out to researchers across the entire CISE ecosystem and beyond to raise awareness of the PIVOT

platform and promote its use. Our outreach efforts will leverage and build on the automotive cybersecurity, vehicle data and telematics, and CISE research communities with which we already engage. This section summarizes our community outreach plans. Additional details are contained in the supplementary Community Outreach Plan.

**Diversity.** PIVOT will make automotive datasets and tools available to those who may not otherwise have access to such specialized resources. We will plan activities targeting faculty, researchers, and students at minority institutions and underrepresented groups, e.g., we will train students, including undergraduates, to assemble and test CAN loggers, as well as train other students in assembly and use.

**Publications.** We will prepare and publish newsletters, blogs, videos, and other news items to promote the PIVOT effort and engagement activities across the broader CISE community.

**Technology Reviews.** We will prepare quarterly technology review articles with succinct lessons and practices for utilizing the technologies associated with PIVOT data collection and utilization.

**Social Media.** We will maintain an active presence on social media sites to promote the PIVOT effort and engagement activities across the broader CISE community.

**Webinars.** We will conduct quarterly webinar presentations to promote the PIVOT effort and engagement activities across the broader CISE community.

**Conferences and Workshops.** We will attend relevant conferences and workshops to present posters, conduct Birds-of-a-Feather sessions, etc. to promote the PIVOT effort and engagement activities across the broader CISE community.

**PIVOT Community Workshops.** We will organize and hold annual Community Outreach and Engagement Workshops to promote and develop the PIVOT effort and conduct in-person engagement activities with the broader CISE Community.

**Cyber Challenge Events.** We will promote PIVOT and recruit users at the annual CyberTruck and CyberAuto Challenge events [4, 34] that PI Daily coordinates.

**Customer Relationship Management (CRM).** We will install and maintain a CRM system to track and manage relationships with CISE community members and beyond.

**CCRI PI Meetings.** The PIVOT team commits to attend and actively participate in the annual CCRI PI meetings to share PIVOT status and plans with the CCRI community.

**CCRI Virtual Organization.** As part of its Community Outreach efforts, the PIVOT team commits to actively work to supply and keep up-to-date information about the PIVOT platform, its resources, and community outreach activities on the CCRI-VO web site.

Our community outreach will be spearheaded by PI Balenson, but all PIVOT team members will contribute to outreach and work as ambassadors for the PIVOT platform. PI Daily will lead interactions with Cyber Challenge events. We will also leverage PIVOT-AB members and other collaborators.

## 3 Evaluation

PIVOT will rely on a feedback loop and an iterative process to improve the offered datasets, tools, and services. PIVOT is positioned at the lower part of the knowledge triangle by converting *raw data to information* (for example, CAN frames to PDUs and signals, purpose-specific datasets, and visualizations), as well as facilitating the creation of DBCs that are the bedrock of comprehending in-vehicle communications. Theresulting information will be moved up the triangle by the research community who will turn it into *knowledge*and create applications about the vehicle (service, maintenance, etc.) and around the vehicle (traffic management, parking, transportation infrastructure, etc.) PIVOT also serves the community by encouraging and nurturing community expansion and industry collaborations.

We will investigate the metrics listed in Table 1 to evaluate our platform. We expect the metrics to change or be adjusted through community feedback.

We will carry out our evaluation as part of our engagement and outreach plans and we will target both providers and consumers as follows.

**Providers.** Data providers include Geotab and other industry partners, as well as researchers who will share their vehicle telematics data (Section 2.3). Users may create derivative datasets and contribute them back. We plan to measure contributor satisfaction by periodically collecting: (a) the number of contributions to the platform (original or revisions), (b) the number of unique contributors, and (c) the demographics of contributors. In addition to these quantitative measures, we will also collect qualitative measures through surveys to evaluate unmet needs of contributors.

| **Consumers.** We will quantitatively measure consumer satisfaction by periodically collecting: (a) the **Metric** | **Year 1** | **Year 2** | **Year 3** |
|---|---|---|---|
| Data quality markers as assessed with user feedback | 75% of users find datasets useful | | |
| Data diversity as determined by vehicle type | 20% or greater for heavy vehicles | | |
| Number of CAN Loggers distributed | 10 | 25 | 50 |
| Number of Spindle devices distributed | 10 | 25 | 50 |
| Number of unique VINs from which data is collected | 20 | 50 | 200 |
| Number of CAN data messages | 1e10 | 2e10 | 10e10 |
| Number of registered PIVOT Platform Users | 10 | 20 | 100 |
| Fraction of registered users from underrepresented groups | 10% | 15% | 20% |
| Time needed to stand up the PIVOT Platform in a new location | less than 8 hrs | | |
| Number of researcher/industry collaborations | 3-5 | 6-10 | 15-20 |
| Publications using PIVOT datasets and tools | 5 | 10 | 20+ |
| Research proposals relying on and/or supporting PIVOT | 2 | 5 | 10 |

Table 1: Proposed evaluation metrics for the PIVOT project

**Consumers**. We will quantitatively measure consumer satisfaction by periodically collecting: (a) the number of unique visitors and visits to our platform, (b) the number of unique visitors who consumed data and the number of unique visits that resulted in consumption, (c) the total count of visitors who engaged with the platform (left rating, comment, etc.), (d) the number of bug reports/feature requests and the number of resolutions, and (e) the demographics of contributors. We will further collect the number of publications that reference our datasets. We will also collect qualitative measures through surveys of current and potential consumers, and through reading of the consumer feedback on the PIVOT platform.

## 4 Team Qualifications

**Christos Papadopoulos**, the lead PI, is a Professor and Sparks Family Chair of Excellence in Global Research Leadership at the University of Memphis. Between 2018-2020 he was a program manager at the Department of Homeland Security Science and Technology Directorate, working on Cyber-Physical Systems Security with an emphasis on vehicle cybersecurity. At DHS he managed the public/private Automotive Cybersecurity Industry Consortium, where he forged close ties with cybersecurity experts at severalOEMs. He conducted and managed several academic research projects in the CISE community such as network measurements and security, NDN and applications of NDN in science. He has experience providing sensitive datasets to the networking community through the DHS Predict and IMPACT projects andparticipated in the work that resulted in the Menlo Report [39]. He is currently teaching network forensics.

**Jeremy Daily**, the PI for Colorado State University, is an Associate Professor of Systems Engineering. He works on heavy vehicle cybersecurity and digital forensics. Dr. Daily is a veteran of the U.S. Air Force where he worked with flightline electronics. He comes out of a formal background in Mechanical Engineering. In 2014, he raised venture capital to start a technology company, Synercon Technologies, LLC, tocreate tools and services for heavy vehicle digital forensics and crash investigations. This company wassold in 2018 to the Dearborn Group. Dr. Daily also co-founded the CyberTruck Challenge in 2017 with KarlHeimer. Currently he is the treasurer of the CyberAuto Challenge and CyberTruck Challenge, sits onthe board of directors for the CyberBoat Challenge, and is an active instructor for the CyberAg Challenge.Dr. Daily is a Co-PI on the evaluation and integration team for a DARPA funded project on Assured Micropatching and has served as PI on multiple NSF awards related to vehicle data. He teaches courses on secure vehicle and industrial communications, design of experiments, and systems engineering processes.

**David Balenson**, the SRI International PI, is a Senior Computer Scientist in the Computer Science Laboratory. He is SRI's Co-PI and spearheads community outreach for the collaborative NSF-funded SEARCCH project with USC/ISI, U. Illinois, and U. Utah, which is developing an open collaboration platform to help researchers package, import, locate, understand, and reuse cybersecurity experiment artifacts [6]. He was SRI's Co-PI for the NSF-funded Cybersecurity Experimentation of the Future project, acommunity-based effort to study current and expected cybersecurity experimentation infrastructure and

produce a strategic plan and roadmap for developing infrastructure to support tomorrow's research [11]. Balenson has over 35 years of technical experience working to research, develop, test, evaluate, and transitioninnovative cybersecurity technologies. Since 2012 he has provided technical support to the DHS S&T cyber-security R&D program [48], where he currently supports the Commercialization Accelerator Program [46] and previously supported the Cyber-Physical Systems Security [47], Smart Cities [49], and IMPACT [33] programs. Balenson helped establish and supports the Automotive Cybersecurity Industry Consortium (ACIC), a public-private partnership between automotive manufacturers ad DHS [1]. He also helps drive community activities such as the Annual Computer Security Applications Conference, the ISOC Network and Distributed System Security Symposium, and the Automotive and Autonomous Vehicle Security Workshop.

## 5 Project Management Plan

Project management for a community research infrastructure goes beyond team management: it additionally requires managing continuous dialog, contributions, design evaluations, and community building. Collectively we bring decades of managing community infrastructure and building communities.

All project PIs have collaborated before. For this project, we will meet bi-weekly using Zoom to report on progress and plans. We will operate and use hosted collaboration tools (git, Google drive, issue tracking, project wiki, etc.), to document project progress and support collaboration.

We propose a three-year project. The first year will focus on building the PIVOT platform, prepopulating it with initial datasets and tools, and reaching out to and engaging the community to elicit their input and encourage their contributions to and use of the platform. The second and third years will focus on updating the PIVOT platform, populating it with additional datasets and tools, andfurther reaching out to and engaging the community to elicit their feedback and encourage their contributions to and useof the platform. All three years will include pursuit of plans to provide long-term sustainability of the infrastructure.

Figure 3 outlines the major tasks, timeline, and responsible parties. Specific activities are discussed in more detail in the supplementary documents on Project Roles and Responsibilities and Community Outreach Plan.

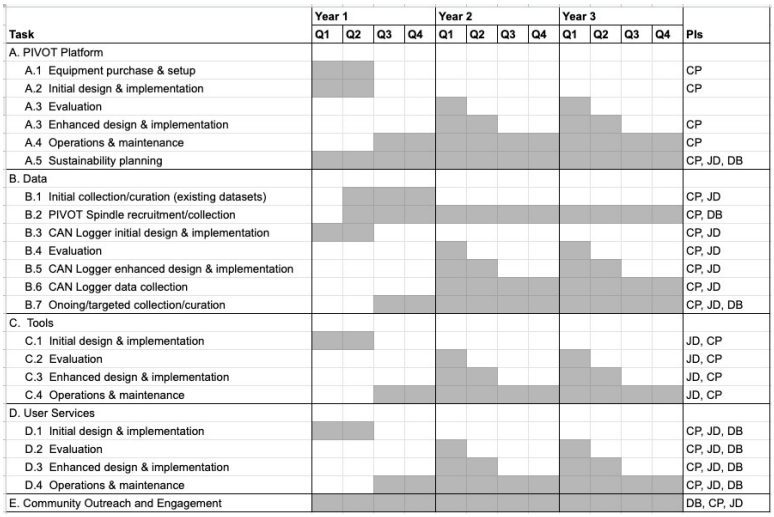


Figure 3: Proposed tasks and timeline. The last column denotes PIs working on the task. The first PI is the lead for the task. CP=Papadopoulos, JD=Daily, DB=Balenson.

## 6 Sustainability

Sustainability is a system design feature to keep the infrastructure available long after the initial funding expires. It is a long and continuous pursuit, and we will engage in sustainability efforts from the beginning of the project. It is important to have plans for technical transition, build community, and help stakeholders understand the project's value, which should encourage them to sustain the project. Weplan to use an evaluation and adaptation loop for sustainability similar to systems development. These sustainability ideas will be continuously refined and tested through feedback from the community and key stakeholders.

**Platform-in-a-Box.** As described in the Data Management Plan, all hardware specifications, software, scripts, policies, configurations, and other information needed to replicate the portal, tools, services, and data access will be documented as a step-by-step deployment guide, so that other institutions or non-profits that wish to mirror the platform will be able to do so quickly and easily. CSU will test and refine the guide by deploying a redundant mirror service that also supports failover. The time it takes to stand up a new deployment of PIVOT will be tracked and minimized.

In addition to supporting redundancy, backup, failover, and recovery, this approach enables any partner

with available resources and desire to sustain the PIVOT platform. In addition to sharing technical documentation and expertise, we will also share any policies and legal documents signed by PIVOT users so they can be resigned with the new partner.

**Continued Operations.** PI Papadopoulos maintains a research lab at the university of Memphis and is committed to continue to host the PIVOT infrastructure for as long as the lab is in existence. The lab is currently supported by other funds. Similarly, PI Daily will commit to maintaining the mirror at Colorado State University while the data center exists.

**Private Sources.** Industry partners may want to support the platform to gain access to the community, students, and faculty involved in the various projects. We will engage with several private entities to seek funds and/or mirroring for continued support of the platform. Potential funding sources include the Industry Advisory Boards (IABs) at Memphis and CSU, and gift funds from industry, including industry participants in the Auto-ISAC, CyberTruck and CyberAuto Challenges. All PIs have extensive connections with industry at various levels, including OEMs, suppliers, and others.

**Sustainability Meetings.** We will include sustainability as an agenda topic in our PIVOT-AB meetings as well as our annual workshops, which will include industry representatives, funding agencies, and other community members. We will also leverage other meeting opportunities such as PI meetings and conferences. These meetings will serve as a source of ideas and opportunities for sustainability. We will avoid any sort of subscription models for sustained funding.

**Continuous Promotion.** We will leverage community outreach activities to promote the platform to stakeholders. This will create awareness and appreciation as well as encourage financial and/or mirroring support past the current funding. It will also create an opportunity for interaction with potential donors to ensure the platform is responsive to their needs.

## 7 Related Work

We are not aware of any projects to coordinate wide scale collection and sharing of automotive and transportation datasets and tools. A number of **existing automotive datasets** have been developed by various labs and researchers and are openly available to the community: HCRL Datasets [28–31], Cephas Baretto Dataset [12], TU Eindhoven Automotive CAN Bus Intrusion Dataset v2 [23], CrySyS Lab CAN-Log Infector and Ambient CAN Traces [17], ETAS/Bosch SynCAN Dataset (Synthetic CAN Bus Data) [32],Real ORNL Automotive Dynamometer (ROAD) CAN Intrusion Dataset [51], and CSU Heavy Vehicle CANData [18]. **Geotab**, collects data from cars and trucks around the globe and makes aggregated data publicly available to its customers and others. ABI Research ranked Geotab as the top telematics company in the world in both 2019 [50] and 2020 [44]. However, Geotab's data is not well known to the CISE community. We will develop extensive automotive datasets via CAN loggers and the Spindle program. We will facilitate access to Geotab and other services. The PIVOT platform will provide a searchable catalog for these datasets and other transportation data for easy access and use by CISE researchers.

There are several projects generating and sharing cybersecurity and/or networking data. **CAIDA (Center for Applied Internet Data Analysis)** has a long history of sharing network data [38]. CAIDA excels in collecting basic data and providing it to researchers, while also developing new measurement techniques. They also have a strong record of community building and outreach through workshops. We will draw on CAIDA's experiences with data collection and sharing. We also plan to build strong community outreach through our workshops and other outreach activities. The **DHS S&T IMPACT program** grew into a 10-year, multi-organization data sharing collaboration, ending in Dec. 2020 [33]. Thesuccesses and challenges of IMPACT's web portal will inform our design on technical and legal agreementsfor access control and privacy. Its ethical guidelines in the Menlo Report [39] will be reflected in PIVOT. **CRAWDAD (Community Resource for ArchivingWireless Data At Dartmouth)** supports dataset archive and distribution for the wireless community [14]. Its success is based on a well-defined community that contributes datasets. We hope to achieve CRAWDAD's success in the automotive space.

The **U.S. Department of Transportation (USDOT) Public Data Portal** makes available a catalog of data related to transportation infrastructure, including automobiles [41]. USDOT also makes available data collected through its **Intelligent Transportation Systems (ITS) Connected Vehicle Pilot Deployment Program** [40]. We plan to index both transportation data sources in the PIVOT platform.

There are several projects collecting and sharing AI/ML data for connected and autonomous vehicles, such as the **Berkeley Deep Drive BDD100K dataset** [53], **comma.ai driving dataset** [45], and **Ford**

**Campus Vision and Lidar dataset** [42]. Younggun Cho maintains a website, **Awesome SLAM Datasets**, with a Google spreadsheet listing all SLAM (Simultaneous Localization and Mapping) datasets along with links, primary characteristics, and instructions on how to filter the spreadsheet [13]. PIVOT will go well beyond a simple spreadsheet and develop a comprehensive, searchable index of automotive datasets and tools and a vibrant community contributing to and benefiting from the platform.

The **World Wide Web Consortium (W3C) Automotive Working Group (AWG)** [52] and **Connected Vehicle Systems Alliance (COVESA)** (formerly GENIVI) [3] are working jointly on the Common Vehicle Interface Initiative (CVII), which seeks to align fragmented data modeling and service-oriented architecture approaches to a common industry standard [8]. CVII is developing a common data model, the *Vehicle Signal Specification (VSS)* [16], as well as a common interface or API, the *Vehicle Information Service Specification (VISS)* [15]. PIs Papadopoulos, Daily, and Balenson have close connections with key leaders in W3C and we plan to leverage these relations to align our work with these emerging standards. Ted Guild, a collaborator from Geotab, co-chairs the W3C AWG and has agreed to share information with and elicit feedback from the PIVOT community.

There are several industry groups working in the automotive cybersecurity and transportation spaces. The **Automotive Cybersecurity Industry Consortium (ACIC)** is a partnership between automotive manufacturers and DHS to conduct pre-competitive research, development, testing, and evaluation proceduresto improve cybersecurity in automotive vehicles [1]. The **Automotive Information Sharing and AnalysisCenter (Auto-ISAC)** is an industry-driven community to share and analyze intelligence about emerging cybersecurity risks to vehicles, and to collectively enhance vehicle cybersecurity capabilities across theglobal automotive industry [2]. The **National Motor Freight Trade Association (NMFTA)** is a nonprofit membership organization comprised of motor carriers operating in interstate, intrastate and foreign commerce that supportsits members in a number of areas, including “research, analyze and distribute information and aggregate data that will be of benefit to the motor carrier industry in the conduct of transportation operations.” [5] PIs Papadopoulos, Daily, and Balenson all participate in and have close connections with key leaders in the aforementioned groups and plan to leverage these relationships to help build and promote PIVOT.

## 8 Results from Prior NSF Support

PI **Christos Papadopoulos** was a member of the core Named Data Networking (NDN) team since 2010. He was PI/Co-PI on two past NSF awards: #1659403 and #1340999. *Intellectual Merit:* These projects explored the application of NDN in High-Energy Physics and Climate Data. *Broader Impacts:* the projects provided new data management tools for both scientific communities.

PI **Jeremy Daily** is the PI for award #1951224 *SaTC: CORE: Small: Collaborative: GOALI: Detecting and Reconstructing Network Anomalies and Intrusions in Heavy Duty Vehicles*, 08/01/2017–12/31/2021, $240,000. *Intellectual Merit*: Developed data collection devices and gathered live network traffic for heavy vehicles as they were deployed to understand the actual data on the heavy vehicle networks. *Broader Impacts*: Supported two graduate students who have transitioned into the cybersecurity workforce. The effortled to techniques and tools utilized at the CyberTruck Challenge, which helps produce talent to address cybersecurity for heavy vehicles and establish a community of interest. *Publications*: Produced two MS theses and the following papers [20–22, 37]. *Research Products*: Open-source software for the CAN Logger and visualization tools are in GitHub [19] and datasets are on the PI’s website [18].

PI **David Balenson** is Co-PI and spearheads community outreach activities for the SEARCCH project (CNS-1925616, “SaTC-CCRI: Collaborative Research: Sharing Expertise and Artifacts for Reuse through Cybersecurity Community Hub (SEARCCH),” 10/2019-9/2022, $474,949). *Intellectual Merit:* Open collaboration platform to help researchers package, import, locate, understand, and reuse cybersecurity experiment artifacts. *Broader Impacts:* Enables new cybersecurity research in CISE disciplines through the ready and increased sharing and reuse of cybersecurity experiment expertise and artifacts.

## References


[1] Automotive Cybersecurity Industry Consortium (ACIC) (website). https://www.acic-auto.org/.

[2] Automotive Information Sharing and Analysis Center (Auto-ISAC) (website). https://automotiveisac.com/.

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

[4] CyberTruck Challenge (website). https://www.cybertruckchallenge.org/.

[5] National Motor Freight Trade Association (NMFTA) (website). http://www.nmfta.org/.

[6] Sharing Expertise and Artifacts for Reuse through Cybersecurity Community Hub (SEARCCH) (website). https://searcch.cyberexperimentation.org/.

[7] SocketCAN userspace utilities and tools (linux-can / can-utils). https://github.com/linux-can/can-utils.

[8] Gunnar Andersson. Common Vehicle Interface Initiative – Home Page. https://wiki.covesa.global/display/WIK4/Common+Vehicle+Interface+Initiative+-+Home+Page.

[9] National Motor Freight Traffic Association. nmfta-repo / pretty_j1939. https://github.com/nmfta-repo/nmfta-opentelematics-prototype.

[10] D. Balenson, C. Papadopoulos, G. Atkinson, T. Guild, S. Prowell, and S. Hollifield. Workshop Report: Paving the Road to Future Automotive Research Datasets: Challenges and Opportunities. https://bit.ly/3t85Kx2, January 2022.

[11] David Balenson, Laura Tinnel, and Terry Benzel. Cybersecurity Experimentation of the Future (CEF): Catalyzing a New Generation of Experimental Cybersecurity Research. http://www.cyberexperimentation.org/files/2114/5027/2222/CEF_Final_Report_Bound_20150922.pdf, July 2015.

[12] Cephas Baretto. OBD-II datasets: Datasets from my Master Degree research. https://www.kaggle.com/cephasax/obdii-ds3.

[13] Younggun Cho. Awesome SLAM Datasets. https://sites.google.com/view/awesome-slam-datasets/home.

[14] Dartmouth College. CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. https://crawdad.org/.

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

[16] COVESA. Vehicle Signal Specification. https://github.com/COVESA/vehicle_signal_specification.

[17] CrySys. Vehicle Security Research. https://www.crysys.hu/research/vehicle-security/.

[18] Jeremy Daily. Heavy Vehicle CAN Data: CAN and J1939 Data Collected from Various Vehicles and Operations. https://www.engr.colostate.edu/~jdaily/J1939/candata.html.

[19] Jeremy Daily. SystemsCyber / CANLoggerFileDecodingGUI. https://github.com/SystemsCyber/CANLoggerFileDecodingGUI.

[20] Jeremy Daily and Ben Gardiner. Cybersecurity considerations for heavy vehicle event data recorders. *SAE International Journal of Transportation Cybersecurity and Privacy*, 1(2):113–143, dec 2018.

[21] Jeremy Daily and Duy Van. Secure controller area network logging. In *SAE WCX Digital Summit*. SAE International, apr 2021.

[22] Jeremy Daily, Duy Van, and Matthew DiSogra. Chip and board level digital forensics of Cummins heavy vehicle event data recorders. *SAE International Journal of Advances and Current Practices in Mobility*, 2(4):2374–2388, apr 2020.

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

[24] The Linux Foundation. Automotive Grade Linux. https://automotiveisac.com/.

[25] Geotab. Ignition Platform (website). https://ignition.geotab.com/.

[26] Geotab. University R&D Program (website). https://ignition.geotab.com/.

[27] GolangRepo. Controller Area Network (CAN) SDK for Go. https://golangrepo.com/repo/einride-can-go.

[28] Korea University Hacking and Countermeasure Research Laboratory. Automotive Ethernet Intrusion Dataset. https://ocslab.hksecurity.net/Datasets/automotive-ethernet-intrusion-dataset.

[29] Korea University Hacking and Countermeasure Research Laboratory. CAN Dataset for intrusion detection (OTIDS). https://ocslab.hksecurity.net/Dataset/CAN-intrusion-dataset.

[30] Korea University Hacking and Countermeasure Research Laboratory. Car-Hacking Dataset for the intrusion detection. https://ocslab.hksecurity.net/Datasets/CAN-intrusion-dataset.

[31] Korea University Hacking and Countermeasure Research Laboratory. Survival Analysis Dataset for automobile IDS. https://ocslab.hksecurity.net/Datasets/survival-ids.

[32] Markus Hanselmann, Thilo Strauss, Katharina Dormann, and Holger Ulmer. CANet: An unsupervised intrusion detection system for high dimensional CAN bus data. *CoRR*, abs/1906.02492, 2019.

[33] ImpactCybertrust.org. IMPACT: Information Marketplace for Policy and Analysis of Cyber-Risk & Trust. https://impactcybertrust.org/.

[34] SAE International. SAE CyberAuto Challenge (website). https://www.sae.org/attend/cyberauto/.

[35] SAE International. SAE J1939 Standards Collection. https://www.sae.org/standardsdev/groundvehicle/j1939a.htm.

[36] The kernel development community. SocketCAN - Controller Area Network. https://www.kernel.org/doc/html/latest/networking/can.html.

[37] Subhojeet Mukherjee, Jacob Walkery, Indrakshi Rayz, and Jeremy Daily. A precedence graph-based approach to detect message injection attacks in J1939 based networks. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pages 67–6709, 2017.

[38] University of California San Diego Supercomputer Center. CAIDA: Center for Applied Internet Data Analysis. https://www.caida.org/.

[39] Department of Homeland Security Science and Technology Directorate. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. Technical report, August 2012.

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

[41] U.S. Department of Transportation. U.S. Department of Transportation's public data portal (website). https://data.transportation.gov/.

[42] Gaurav Pandey, James McBride, and Ryan Eustice. Ford Campus vision and lidar data set. *I. J. Robotic Res.*, 30:1543–1552, 10 2011.

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

[44] ABI Research. Geotab and Verizon Hold Lead Again in ABI Research's Commercial Telematics Competitive Ranking. https://www.abiresearch.com/press/geotab-and-verizon-hold-lead-again-in-abi-researchs-commercial-telematics-competitive-ranking/, September 2020.

[45] Eder Santana and George Hotz. Learning a Driving Simulator. *CoRR*, abs/1608.01230, 2016.

[46] DHS S&T. Commercialization Accelerator Program (CAP) (website). https://www.dhs.gov/publication/st-commercialization-accelerator-program.

[47] DHS S&T. Cyber-Physical Systems Security (CPSSEC) (website). https://www.dhs.gov/science-and-technology/cpssec.

[48] DHS S&T. DHS S&T Cybersecurity Programs (website). http://www.cyber.st.dhs.gov/.

[49] DHS S&T. Smart Cities (website). https://www.dhs.gov/science-and-technology/csd-smart-cities.

[50] Automotive Fleet Staff. Telematics Companies Ranked by ABI Research. *Automotive Fleet*, June 2019.

[51] Miki E. Verma, Michael D. Iannacone, Robert A. Bridges, Samuel C. Hollifield, Bill Kay, and Frank L. Combs. ROAD: the real ORNL automotive dynamometer controller area network intrusion detection dataset (with a comprehensive CAN IDS dataset survey & guide). *CoRR*, abs/2012.14600, 2020.

[52] W3C. Automotive Working Group (website). https://www.w3.org/auto/wg/.

[53] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018.